Bayesian Analysis (2024) **TBA**, Number TBA, pp. 1–26

A Bayesian Nonparametric Approach to Species Sampling Problems with Ordering

Cecilia Balocchi[*] Federico Camerlenghi[†] , and Stefano Favaro[‡]

Abstract. Species-sampling problems (SSPs) refer to a vast class of statistical problems calling for the estimation of (discrete) functionals of the unknown species composition of an unobservable population. A common feature of SSPs is their invariance with respect to species labeling, which is at the core of the Bayesian nonparametric (BNP) approach to SSPs under the popular Pitman-Yor process (PYP) prior. In this paper, we develop a BNP approach to SSPs that are not "invariant" to species labeling, in the sense that an ordering or ranking is assigned to species' labels. Inspired by the population genetics literature on age-ordered alleles' compositions, we study the following SSP with ordering: given an observable sample from an unknown population of individuals belonging to species (alleles), with species' labels being ordered according to weights (ages), estimate the frequencies of the first r order species' labels in an enlarged sample obtained by including additional unobservable samples. By relying on an ordered PYP prior, we obtain an explicit posterior distribution of the first r order frequencies, with estimates being of easy implementation and computationally efficient. We apply our approach to the analysis of genetic variation, showing its effectiveness in estimating the frequency of the oldest allele, and then we discuss other potential applications.

Keywords: Bayesian nonparametrics, exchangeable partition probability function, first r order frequency, ordered Pitman-Yor process prior, species sampling problems, population genetics.

1 Introduction

Species sampling problems (SSPs) refer to a vast class of statistical problems, of which the estimation of the number of unseen species is arguably the most popular example (Good and Toulmin, 1956; Efron and Thisted, 1976; Lijoi et al., 2007; Orlitsky et al., 2016). Consider $n \geq 1$ observable samples from a generic population of individuals, with each individual taking a value in a (possibly infinite) discrete space of symbols or species' labels. The unseen-species problem assumes that observable samples are modeled as a random sample $(X_1, \ldots, X_n)$ from an unknown discrete distribution p, and calls for estimating

$$K_m^{(n)} = |\{X_{n+1}, \ldots, X_{n+m}\} \setminus \{X_1, \ldots, X_n\}|, \tag{1}$$

arXiv: 2203.07342

[*]School of Mathematics, University of Edinburgh, cecilia.balocchi@ed.ac.uk

[†]Department of Economics, Management and Statistics, University of Milano - Bicocca, federico.camerlenghi@unimib.it

[‡]Department of Economics, Social Studies, Applied Mathematics and Statistics, University of Torino and Collegio Carlo Alberto, stefano.favaro@unito.it

 https://doi.org/10.1214/24-BA1418

namely the number of hitherto unseen symbols that would be observed if $m \geq 1$ additional samples $(X_{n+1}, \ldots, X_{n+m})$ were collected from the same distribution. SSPs comprise of generalizations or refinements of the unseen-species problem, calling for the estimation of (discrete) functionals of the species' composition of unobservable samples, e.g. missing mass, discovery probabilities, unseen species with prevalences and coverages of prevalence. We refer to Deng et al. (2019) and Balocchi et al. (2022) for reviews of SSPs, both in methods and applications, mostly in the field of biological sciences but also in machine learning, electrical engineering, computer science and information theory.

A common feature of SSPs is that species' labels identifying the X_i's are immaterial in the definition of the functional of interest, as for instance in (1), thus making SSPs "invariant" to species labeling. Such a feature is at the core of the Bayesian nonparametric (BNP) approach to SSPs (Lijoi et al., 2007, 2008; Favaro et al., 2009, 2013), which relies on the specification of a (nonparametric) prior $\mathscr{P}$ for the unknown distribution p, i.e.

$$\begin{aligned} X_i \,|\, P &\overset{\text{iid}}{\sim} P \qquad i = 1, \ldots, n, \\ P &\sim \mathscr{P}. \end{aligned} \tag{2}$$

Species sampling models (SSMs) (Pitman, 1996) provide a natural choice for the prior distribution $\mathscr{P}$, including the celebrated Dirichlet process (DP) prior (Ferguson, 1973) and the Pitman-Yor process (PYP) prior (Pitman and Yor, 1997). Under the BNP model (2) with a SSM for P, the random sample $(X_1, \ldots, X_n)$ induces a random partition $\tilde{\pi}_n$ of $[n] = \{1, \ldots, n\}$ whose blocks correspond to the (equivalence) classes induced by the equivalence relation $i \sim j \iff X_i = X_j$ almost surely. In particular, $\tilde{\pi}_n$ is exchangeable (Pitman, 2006, Chapter 2), namely its distribution is such that the probability of any partition of $[n]$ with k blocks of frequencies $(n_1, \ldots, n_k)$ is a symmetric function of compositions $(n_1, \ldots, n_k)$ of $[n]$. The exchangeability of $\tilde{\pi}_n$ implies that blocks' labels are immaterial, and therefore it legitimates the BNP approach to SSPs under the class of SSMs.

1.1 Our contributions

In this paper, we consider SSPs that are not "invariant" with respect to species labeling, in the sense that an ordering or ranking is assigned to species' labels, and we develop a BNP approach to such problems. Under the infinitely-many neutral alleles model for the evolution of genetic populations (Ewens, 1972; Kingman, 1975; Watterson and Guess, 1977; Griffiths, 1979), the work of Donnelly and Tavaré (1986) first investigated the alleles' composition of a random sample from the population by also taking into account the ages of alleles, namely the times elapsed since the first time each allele first appeared in the sample. This study led to the introduction of an age-ordered version of the random partition induced by the DP prior, where species are alleles and species' labels are ordered according to the age of alleles in such a way that the smaller the order the older the allele. Besides providing distributional properties of the age-ordered random partition, Donnelly and Tavaré (1986) applied such a model to answer

a critical question raised in Crow (1972): "Is the most frequent allele the oldest?". Under the infinitely-many neutral alleles model, Donnelly and Tavaré (1986) came up with a positive answer to such a question, showing that the probability that an allele represented i times in sample of size n is the oldest is i/n (Kelly, 1977; Watterson and Guess, 1977). The question of Crow (1972) is to some extent a SSP with ordering, as the object of interest involves a species' label with a precise order, namely the species' label of order 1 that corresponds to the oldest allele.

Inspired by the seminal work of Donnelly and Tavaré (1986), we study the following SSP with ordering: assuming n observable samples to be modeled as a random sample $((T_1, X_1), \ldots, (T_n, X_n))$ from an unknown discrete distribution q on $\mathbb{R}_+ \times \mathbb{X}$, with the positive T_i's being considered as weights inducing an ordering among species' labels identifying the X_i's, on a general (measurable) space $\mathbb{X}$, we estimate the frequencies of the first r order species' labels in an enlarged sample obtained by including m additional unobservable samples $((T_{n+1}, X_{n+1}), \ldots, (T_{n+m}, X_{n+m}))$ from the same unknown distribution. Within the population genetic setting of Donnelly and Tavaré (1986), this ordered SSP corresponds to the estimation of the frequency of the r oldest alleles in a sample of size $(n + m)$ based on n observable samples. We introduce a BNP approach to estimate the first r order frequencies, which relies on the use of the class of spatial neutral to the right SSMs (James, 2006a), or ordered SSMs, as prior distributions for q. The most popular ordered SSM is the ordered DP prior, which is known to induce the age-ordered random partition of Donnelly and Tavaré (1986). See also Gnedin and Pitman (2005) and references therein. Here, we consider the more general ordered PYP prior (Gnedin and Pitman, 2005; James, 2006a), and we determine the posterior distribution of the first r order frequencies; then, a BNP estimator is proposed in terms of the posterior mean, whose closed-form expression results to be of easy implementation and also computationally efficient. Of special interest is the case $r = 1$ that, in the original setting of Donnelly and Tavaré (1986), leads to an estimate of the frequency of the oldest allele.

We present an empirical validation of the effectiveness of our BNP approach, both on synthetic data and real data. It is natural to focus on applications to genetic data, for which the weights T_i's have an interpretation as the ages of the alleles X_i's. The problem of modeling the interplay between the alleles' composition of a genetic population and the age of alleles dates back to the 1970s and the 1980s, and nowadays the genealogical structure of alleles is well recognized as a fundamental aspect in many inferential (decision) processes in the field of population genetics. In particular, investigating genetic variation while incorporating the information on the variants' age enhances the investigation of several problems, such as analyzing and comparing populations structure, detecting which samples are related, studying demographic history, and learning about genetic susceptibility to disease (Mathieson and McVean, 2014). For example, it enables researchers to use variants' age distribution to compare populations, to differentiate age distributions in pathogenic and benign variants, and to learn about genealogical history (Albers and McVean, 2020). Here, we apply our BNP methodology to the problem of estimating the frequency of the oldest allele, using genetic variation data from the 1000 Genomes Project (1000 Genomes Project Consortium, 2015) and variants' age estimates from the Human Genome Dating Project (Albers and McVean, 2020). Thanks to our

posterior estimator, we can not only answer inferential questions on the frequency of the oldest allele in an observed sample, but also make predictions by analyzing an enlarged sample. By studying the trajectory of this frequency as a function of the enlarged sample size, we can enhance our understanding of the population distribution, thereby addressing the investigation of the aforementioned issues more effectively.

Besides population genetics, SSPs with ordering arise in at least other two contexts: i) citations to academic articles and ii) online purchases of items. In the context of citations to academic articles, with articles being ordered according to their publication's dates, one may be interested in the frequency of citations to the oldest paper. Citation data are often analyzed in the framework of citation networks to study the movement of ideas in academic fields or examine scholars' influence (Portenoy et al., 2017). Incorporating knowledge of articles' age permits the investigation of the effects of time on the number of citations (Hajra and Sen, 2005) and to answer questions such as "are older papers more frequently cited than newer papers?". Assuming that each citation represents an observation, that the article cited represents a species, and that the order of the cited article is determined by its publication's date, we may apply our BNP approach in order to predict the frequency of citations to the oldest article in future observations. In the context of online purchases of items, species' labels are represented by the items purchased, with items being ordered according to their costs, as well as by a generic independent measure of popularity. In such a context, we may apply our BNP approach to study the distribution of the most popular item in future purchases, which is particularly relevant in order to plan suitable changes in the current marketing strategies.

1.2 Organization of the paper

The paper is structured as follows. In Section 2 we present the ordered PYP prior and review its sampling structure in terms of sampling formulae and predictive distributions. In Section 3 we provide the posterior distribution of the first r order frequencies, with emphasis on the special case $r = 1$, and obtain corresponding estimators. Section 4 contains numerical illustrations of our BNP approach, both on synthetic and real data, whereas in Section 5 we discuss our work and some directions for future research. Additional numerical illustrations on genetic data, an illustration in the context of citations to academic articles, and the proofs of our results are deferred to the Supplementary Material (Balocchi et al., 2024).

2 The ordered PYP

To introduce the ordered PYP, it is useful to recall the PYP and its sampling structure. Let P be a PYP with parameter $\alpha \in [0, 1)$ and $\theta > -\alpha$ on a measurable space $\mathbb{X}$. That is $P = \sum_{i \geq 1} P_i \delta_{S_i}$, where: i) $P_1 = V_1$ and $P_i = V_i \prod_{1 \leq j \leq i-1}(1 - V_j)$ with $(V_i)_{i \geq 1}$ being independent Beta random variables with parameter $(1 - \alpha, \theta + i\alpha)$; ii) $(S_i)_{i \geq 1}$ be random variables, independent of the V_i's, and independent and identically distributed according to a non-atomic distribution ν on $\mathbb{X}$ (Perman et al., 1992; Pitman, 1995).

Because of the (almost sure) discreteness of P, a random sample $(X_1, \ldots, X_n)$ from P induces a random partition $\tilde{\pi}_n$ of $[n]$ into $K_n \leq n$ blocks, labelled by $\{X_1^*, \ldots, X_{K_n}^*\}$, with frequencies $N_{j,n} = |i \in [n] : X_i = X_j^*|$ for $j = 1, \ldots, K_n$ and such that $N_{j,n} \geq 1$ and $\sum_{1 \leq j \leq K_n} N_{j,n} = n$. In particular, if we set $(a)_{(r)} = \prod_{0 \leq i \leq r-1}(a+i)$ for any $a \geq 0$ and $r \in \mathbb{N}_0$, then the probability of any partition of $[n]$ with k blocks of frequencies $(n_1, \ldots, n_k)$ is

$$\Pi_k^{(n)}(n_1, \ldots, n_k) = \frac{\prod_{i=1}^k (\theta + (i-1)\alpha)}{(\theta)_{(n)}} \prod_{i=1}^k (1-\alpha)_{(n_i-1)}. \tag{3}$$

Equation (3) is referred to as the exchangeable partition probability function (EPPF), a concept introduced in Pitman (1995) as a development of results in Kingman (1978). For $\alpha = 0$ the PYP reduces to DP, and hence (3) reduces to the Ewens sampling formula (Ewens, 1972). See Pitman (2006, Chapter 3 and Chapter 4) for a detailed account of EPPFs.

The predictive distribution of the PYP provides a generative scheme for the random partition $\tilde{\pi}_n$. This is typically stated in terms of the Chinese Restaurant Process (Pitman, 2006, Chapter 3), which is a sequential construction of $\tilde{\pi}_n$ through the metaphor of customers (observations) sitting at tables (species) of a restaurant. Under Chinese Restaurant Process, the first customer X_1 arrives and is assigned to a table. After n customers $(X_1, \ldots, X_n)$ have arrived and have been assigned to k tables $\{X_1^*, \ldots, X_k^*\}$, with n_i being the number of customers at table $i = 1, \ldots, k$, the customer X_{n+1} arrives and

i) she will sit at a ("new") table X^*, that is a table not already occupied, with a probability
$$p^{(new)} = \frac{\theta + k\alpha}{\theta + n}; \tag{4}$$

ii) she will sit at a table X_j^* that has been already occupied, for $j = 1, \ldots, k$, with a probability
$$p_j^{(old)} = \frac{n_j - \alpha}{\theta + n}. \tag{5}$$

We refer to Pitman (2006, Chapter 3 and Chapter 4) for a detailed account of Chinese Restaurant Process and its generalizations to SSMs. In particular, the PYP is characterized as the sole SSM for which $p^{(new)}$ depends only on (n, k) and $p_j^{(old)}$ depends only on (n, n_j) (Zabell et al. (1997); see Bacallado et al. (2017) for more general sufficiency postulates).

Equation (3) is a symmetric function of compositions $(n_1, \ldots, n_k)$ of $[n]$, that is the random partition $\tilde{\pi}_n$ induced by the PYP is an exchangeable random partition (Pitman, 2006, Chapter 2). The ordered PYP is a discrete random probability measure generalizing the PYP, in the sense that random sampling from the ordered PYP allows to couple each species' label with a corresponding order (Gnedin and Pitman, 2005; James, 2006a). An ordered PYP Q is an almost surely discrete random probability

measure with parameters $\alpha \in [0,1)$ and $\theta > 0$, that can be defined relying on the de Finetti theorem. Indeed, if $\{(T_i, X_i)\}_{i\geq 1}$ is an exchangeable sequence of observations whose directing measure is an ordered PYP Q, i.e., $(T_i, X_i) \,|\, Q \overset{\text{iid}}{\sim} Q$ as $i \geq 1$, we can characterize Q by assigning the predictive distributions of the associated exchangeable sequence. In order to do this, consider the random sample $((T_1, X_1), \ldots, (T_n, X_n))$ from Q, with the T_i's being viewed as weights that induce an ordering among species' labels identifying the X_i's. Because of the (almost sure) discreteness of Q, the random sample $((T_1, X_1), \ldots, (T_n, X_n))$ from Q induces a random partition of $[n]$ into $K_n \leq n$ blocks, labelled by a K_n-tuple $((T^*_1, X^*_1), \ldots, (T^*_{K_n}, X^*_{K_n}))$ that is ordered according to the T^*_j's in such a way that $T^*_1 > \cdots > T^*_{K_n}$, with corresponding ordered frequencies $M_{j,n} = |i \in [n] : (T_i, X_i) = (T^*_j, X^*_j)|$ for $j = 1, \ldots, K_n$ and such that $M_{j,n} \geq 1$ and $\sum_{1\leq j\leq K_n} M_{j,n} = n$. Species' labels X^*_j's are thus ordered with respect to the decreasing ordering of the weights T^*_j's, namely the larger the weight the smaller the order, such that $M_{j,n}$ is the frequency of the species' label of order j that corresponds to the j-th largest weight T^*_j. An analogous construction follows for ordered SSMs (James, 2006a).

In analogy with the Chinese Restaurant Process, the predictive distribution of the ordered PYP Q may be stated as an ordered version of the Chinese Restaurant Process, with tables ordered according to weights (James, 2006a). In particular, under the ordered Chinese Restaurant Process, the first n customers $((T_1, X_1), \ldots, (T_n, X_n))$ arrive and they are assigned to the k ordered tables $((T^*_1, X^*_1), \ldots, (T^*_k, X^*_k))$, with the table of order j corresponding to the j-th largest weight T^*_j. Hereinafter, we denote by m_j the number of customers seated at the table of order j for $j = 1, \ldots, k$, and we set $r_j = m_j + \cdots + m_k$, for any $j = 1, \ldots, k$, and $r_{k+1} = 0$. Then, the customer (T_{n+1}, X_{n+1}) arrives and

i) she will sit at a "new" table (T^*, X^*) of order $j = 1, \ldots, k+1$, that is a table not already occupied and whose order j is determined through the weight T^*, with a probability

$$q_j^{(new)} = \frac{\theta + \alpha r_j}{(1 + r_j)(\theta + n)} \prod_{i=1}^{j-1} \frac{r_i(\alpha r_{i+1} + \alpha + \theta m_i)}{(r_i + 1)(\alpha r_{i+1} + \theta m_i)}, \tag{6}$$

where T^* and X^* are generated from two non-atomic distributions on $\mathbb{R}_+$ and $\mathbb{X}$ respectively;

ii) she will sit at a table (T^*_j, X^*_j) that has been already occupied, for $j = 1, \ldots, k$, with a probability

$$q_j^{(old)} = \frac{r_j(m_j - \alpha)(\alpha r_{j+1} + \theta m_j + \theta)}{(1 + r_j)(\theta + n)(\alpha r_{j+1} + \theta m_j)} \prod_{i=1}^{j-1} \frac{r_i(\alpha r_{i+1} + \alpha + \theta m_i)}{(r_i + 1)(\alpha r_{i+1} + \theta m_i)}. \tag{7}$$

Given that the T^*_j only affect the distribution of an ordered partition through the ordering induced on the clusters, we avoid using specific notation for its distribution, as it is immaterial. We refer to Gnedin and Pitman (2005) and James (2006a) for a

detailed account of (6) and (7). The predictive distribution of the ordered DP arises from (6) and (7) by setting $\alpha = 0$.

Equation (6) and Equation (7) provide a generative scheme for the random partition of $[n]$ induced by the ordered PYP Q. That is: i) if the $(n+1)$-th customer (T_{n+1}, X_{n+1}) sits at the new table (T^*, X^*), which happens with probability (6), then the order of such a table with respect to the ordering of the already occupied tables $((T_1^*, X_1^*), \ldots, (T_k^*, X_k^*))$ is determined by T^*, thus possibly changing the ordering of occupied tables by shifting the order of tables (T_j^*, X_j^*)'s with weights smaller than T^*; ii) if the $(n+1)$-th customer (T_{n+1}, X_{n+1}) sits at a table (T_j^*, X_j^*) that is already occupied, which happens with probability (7), then the order of such a table with respect to the ordering of the already occupied tables is determined by T_j^*, thus not changing the ordering of occupied tables. In other terms, a new customer sitting at a new table may determine a change in the ordering of the occupied tables, whereas a new customer sitting at a table occupied does not determine a change in the ordering of the already occupied tables.

Gnedin and Pitman (2005) and James (2006a) first investigated properties of the random partition induced by the ordered PYP, and introduced the notion of ordered EPPF. Generalizing the definition of EPPF, the ordered EPPF is defined as the probability of any ordered partition of $[n]$ with k blocks of frequencies $(m_1, \ldots, m_k)$. Here the term *ordered partition* refers to a partition of $[n]$, where the blocks are ordered in accordance with the weights T_j. Gnedin and Pitman (2005) showed that the ordered PYP induces a random partition whose ordered EPPF is

$$\Phi_k^{(n)}(m_1, \ldots, m_k) = \frac{\prod_{i=1}^k \frac{\theta m_i + \alpha r_{i+1}}{r_i}}{(\theta)_{(n)}} \prod_{i=1}^k (1-\alpha)_{(m_i-1)}. \tag{8}$$

See also Gnedin (2010), and references therein, for a comprehensive account on ordered EPPFs and generalizations thereof. Note that the EPPF (3) can be recovered from the ordered EPPF (8) by summing over the set S_k of all possible permutations of the k blocks, that is

$$\Pi_n^{(k)}(m_1, \ldots, m_k) = \sum_{\pi \in S_k} \Phi_k^{(n)}(m_{\pi(1)}, \ldots, m_{\pi(k)}). \tag{9}$$

See Section S1.1 of the Supplementary Material for details on Equation (9). The distribution of the age-ordered partition of Donnelly and Tavaré (1986) arises from (8) by setting $\alpha = 0$, where species' labels are ordered according to weights T_i's that are interpreted as the ages of alleles. Another special case of the ordered EPPF (8) is obtained by setting $\alpha \in (0,1)$ and $\theta = 0$. See Favaro and James (2016) and references therein for details.

By applying the ordered EPPF (8), one may compute the probability $P_n(i; \alpha, \theta)$ that a species with frequency i has species' label of order 1, i.e. the species' label corresponding to the largest weight T_1^*. For $\alpha = 0$, Donnelly and Tavaré (1986) computed such a probability, showing that it is independent of θ and also an increasing (linear) function of i, i.e.

$$P_n(i; 0, \theta) = \frac{i}{n}. \tag{10}$$

Within the population genetic setting of Donnelly and Tavaré (1986), Equation (10) shows that the most frequent allele is the oldest allele. In general, for any $\alpha \in [0, 1)$ and $\theta > 0$ it holds

$$P_n(i; \alpha, \theta) = \frac{\alpha n + i(\theta - \alpha)}{n} \mathbb{E}\left[\frac{1}{\theta + \alpha K_{n-i}}\right], \tag{11}$$

where K_{n-i} is the number of distinct species in $(n-i)$ random samples for the ordered PYP Q, with the proviso $K_0 = 0$ (Pitman, 2006, Chapter 3). See Section S1.2 of the Supplementary Material for the proof of Equation (11). It is easy to show that (11) reduces to the probability (10) for $\alpha = 0$. The comparison between (11) and (10) is critical, as it highlights the increased flexibility of the ordered PYP compared to the ordered DP ($\alpha = 0$). In fact, differently from the probability (10), the probability (11) depends on (α, θ) and, most importantly, it is no more an increasing (linear) function of i. Figure 1 shows that the probability (11) may increase or decrease in i according to the value of (α, θ); for instance, for $\alpha \in (0, 1)$ and $\theta = 0$ the probability (11) is the product of a term decreasing in i and one increasing in i. The non-increasing behavior of $P_n(i; \alpha, \theta)$ for $\alpha > \theta$ and for $\theta = 0$ is depicted in Figure S1 of the Supplementary Material.

To conclude, it is worth mentioning the construction of the order PYP Q that induces the predictive distributions (6) and (7), though we will not make use of such a construction in the paper. The ordered PYP belongs to the class of spatial neutral to the right SSMs defined by James (2006a), and hence it is defined as a discrete random probability measure on the product space $\mathscr{S} = \mathbb{R}^+ \times \mathbb{X}$, from which the observations (T_i, X_i), as $i = 1, \ldots, n$, are sampled. To formalize such a definition, we consider a marked Poisson process N (Kingman, 1993) on the space $[0, 1] \times \mathscr{S}$ with mean intensity given by

$$\nu(\mathrm{d}u, \mathrm{d}s, \mathrm{d}x) := \rho(\mathrm{d}u|s)\Lambda_0(\mathrm{d}s, \mathrm{d}x)$$

where ρ is a Lévy density, while Λ_0 is a hazard measure on the space $\mathscr{S}$. Thus, one may define a functional of the Poisson process N as follows $\Lambda(\mathrm{d}s, \mathrm{d}x) = \int_0^1 uN(\mathrm{d}u, \mathrm{d}s, \mathrm{d}x)$, which turns out to be a completely random measure (Daley and Vere-Jones, 2008); Λ represents a hazard measure in the framework of survival analysis. Now, define the survival function associated with Λ as $-\log(S(t-)) := \int_{[0,1]\times\mathscr{S}}[-\mathbb{1}_{\{s<t\}} \log(1 - u)]N(\mathrm{d}u, \mathrm{d}s, \mathrm{d}x)$. Then, a spatial neutral to the right random probability measure equals $Q(\mathrm{d}t, \mathrm{d}x) := S(t-)\Lambda(\mathrm{d}t, \mathrm{d}x)$. By choosing ρ as in (James, 2006a, Section 6.2), the law of the resulting Q is the de Finetti measure associated with the prediction rules (6)–(7). Note that in this construction the T_i's are considered as times, but they can be seen more generally as weights inducing an order, making the model more widely applicable. See James (2006a) for general properties of the ordered PYP Q, including the posterior distribution.

3 BNP inference for the first r order frequencies

In analogy with SSPs, SSPs with ordering assume $n \geq 1$ observable samples from a population of individuals, with each individual taking a value in a (possibly infinite) discrete space of symbols, and then consider $m \geq 1$ additional unobservable sample from

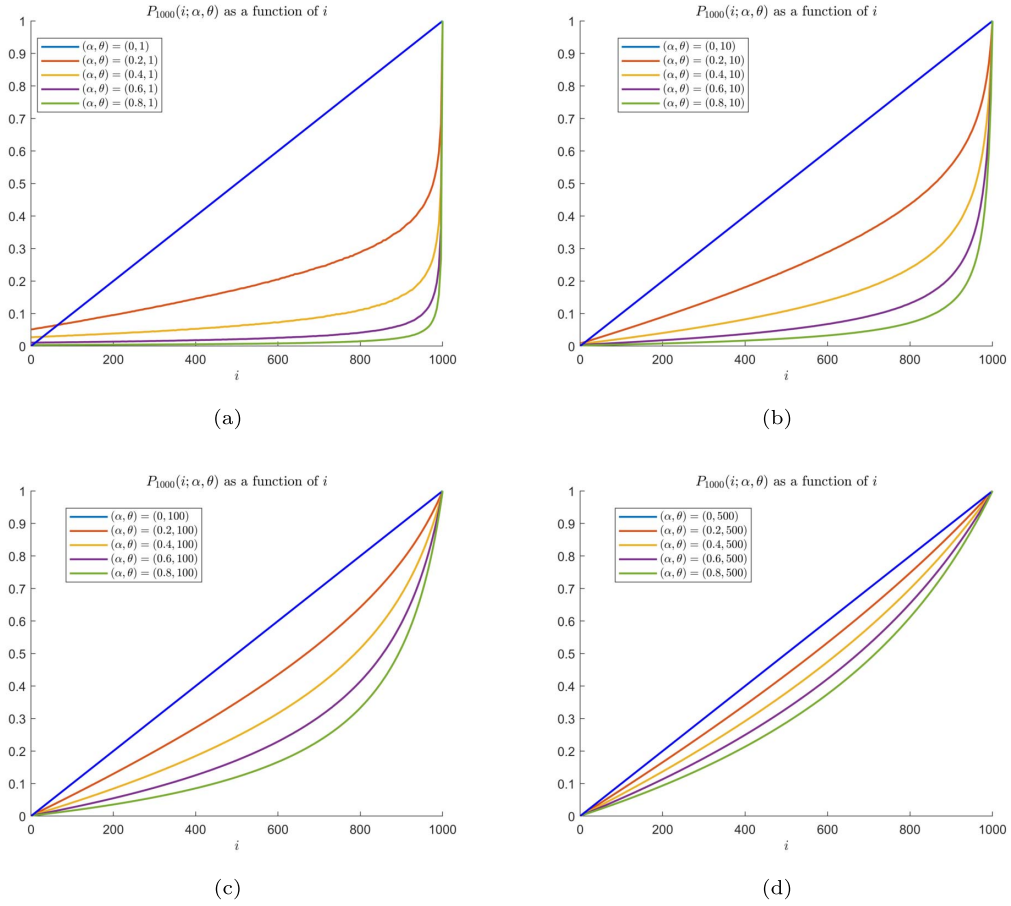

Figure 1: The probability $P_n(i;\alpha,\theta)$, where $n = 1000$, as a function of the frequency i, for different values of α and θ. Each panel corresponds to a different θ, from top left to bottom right, we have: $\theta = 1, 10, 100, 500$.

the same population. The critical difference between SSPs and SSPs with ordering lies in the definition of the (discrete) functional of interest: while in SSPs such a functional is "invariant" with respect to species ordering and deals with the species' composition of the additional samples, in SSPs with ordering the functional is not "invariant" with respect to species ordering and deals with the species' composition of both the additional samples and the enlarged sample. The estimation of the first r order frequencies is arguably the most natural example of SSPs with ordering. Assuming n observable samples to be modeled as a random sample $((T_1, X_1), \ldots, (T_n, X_n))$ from the ordered PYP Q:

$$(T_i, X_i) \,|\, Q \overset{\text{iid}}{\sim} Q \qquad i = 1, \ldots, n, \tag{12}$$

i.e., the observations are updated according to the predictive laws (6)-(7). We introduce a BNP approach to estimate the frequencies of the first r order species in an enlarged sample obtained by collecting m additional samples $((T_{n+1}, X_{n+1}), \ldots, (T_{n+m}, X_{n+m}))$ from the same Q.

3.1 Posterior distributions for the first r order frequencies

We start by introducing a marginal distribution related to the random partition induced by an ordered PYP Q, with parameters (α, θ). For any $n \geq 1$ let $((T_1, X_1), \ldots, (T_n, X_n))$ be a random sample under the BNP model (12), such that the sample features $K_n = k$ distinct species with ordered frequencies $\mathbf{M}_n = \mathbf{m}$. Hereinafter, for the sake of simplicity in notation, we denote by $|\mathbf{m}|_{1:r}$ the sum of the first r elements of $\mathbf{m}$, i.e. $|\mathbf{m}|_{1:r} = \sum_{1 \leq j \leq r} m_j$, with the proviso $|\mathbf{m}|_{1:0} = 0$. For any index $r \in \{1, \ldots, n\}$ such that $r \leq |\mathbf{m}|_{1:r} \leq n - k + r$, if we set

$$C_{r,n}(\alpha, \theta, \mathbf{m}) = \prod_{j=1}^{r} \frac{[\alpha(n - |\mathbf{m}|_{1:j}) + \theta m_j]}{n - |\mathbf{m}|_{1:j-1}} (1 - \alpha)_{(m_j - 1)}$$

then

$$\begin{aligned} &\Pr[M_{1,n} = m_1, \ldots, M_{r,n} = m_r, K_n \geq r] \\ &= \binom{n}{m_1, \ldots, m_r, n - |\mathbf{m}_{1:r}|} \frac{C_{r,n}(\alpha, \theta, \mathbf{m})}{(\theta + n - |\mathbf{m}|_{1:r})_{|\mathbf{m}|_{1:r}}}. \end{aligned} \tag{13}$$

See Section S1.4 of the Supplementary Material for the proof of Equation (13). Equation (13) generalizes Donnelly and Tavaré (1986, Proposition 6.1), which is recovered from (13) by letting $\alpha \to 0$. For $r = 1$, Equation (13) provides the distribution of first order frequency, i.e.

$$\Pr[M_{1,n} = m_1] = \binom{n}{m_1} \frac{\alpha(n - m_1) + \theta m_1}{n(\theta + n - m_1)_{(m_1)}} (1 - \alpha)_{(m_1 - 1)}. \tag{14}$$

Within the population genetic setting of Donnelly and Tavaré (1986), Equation (14) with $\alpha = 0$ provides the distribution of the frequency of the oldest allele (Kelly, 1977; Watterson and Guess, 1977).

Now, we can state our main results on the posterior distribution of the first r order frequencies. Let $((T_1, X_1), \ldots, (T_n, X_n))$ be a random sample from the ordered PYP Q, and let $((T_{n+1}, X_{n+1}), \ldots, (T_{n+m}, X_{n+m}))$ be an additional random sample from the same ordered PYP Q. Moreover, we denote by $K_m^{(n)}$ the number of distinct species in the sample $((T_{n+1}, X_{n+1}), \ldots, (T_{n+m}, X_{n+m}))$ that are not in $((T_1, X_1), \ldots, (T_n, X_n))$, i.e. $K_m^{(n)} = K_{n+m} - K_n$, and we denote by $W_{i,n+m}$ the frequency of the specie's label of order i in the enlarged sample $((T_1, X_1), \ldots, (T_{n+m}, X_{n+m}))$, for $i = 1, \ldots, K_{n+m}$. To determine the distribution of the ordered frequencies $\mathbf{W}_{n+m} = (W_{1,n+m}, \ldots, W_{K_{n+m},n+m})$, it is useful to set

$$A_r = \{\text{species' labels with order } 1, \ldots, r \text{ are new}\}, \tag{15}$$

i.e., the event that the species' labels with higher weights have not been recorded in the first sample, and

$$B_r = \{\text{species' labels with order } 1, \dots, r \text{ are old}\}, \tag{16}$$

i.e. the event that the observations with higher weights have been recorded in the initial sample.

Theorem 1. *Let $((T_1, X_1), \dots, (T_n, X_n))$ be a random sample under the BNP model (12), such that the sample features $K_n = k$ distinct species with ordered frequencies $\mathbf{M}_n = \mathbf{m}$. Let $((T_{n+1}, X_{n+1}), \dots, (T_{n+m}, X_{n+m}))$ be an additional random sample under the same BNP model (12) such that the enlarged sample $((T_1, X_1), \dots, (T_{n+m}, X_{n+m}))$ features K_{m+n} distinct species with corresponding ordered frequencies $\mathbf{W}_{n+m}$, and set $K_m^{(n)} = K_{n+m} - K_n$. If A_r and B_r are the events defined in (15) and (16), respectively, then it holds:*

i) for $r \in \{1, \dots, n+m\}$ such that $r \leq |\mathbf{w}|_{1:r} \leq m$,

$$\begin{aligned} &Pr[A_r, W_{1,n+m} = w_1, \dots, W_{r,n+m} = w_r, K_m^{(n)} \geq r \,|\, K_n = k, \mathbf{M}_n = \mathbf{m}] \\ &= \binom{m}{w_1, \dots, w_r, m - |\mathbf{w}|_{1:r}} \frac{C_{r,n+m}(\alpha, \theta, \mathbf{w})}{(\theta + n + m - |\mathbf{w}|_{1:r})_{(|\mathbf{w}|_{1:r})}}; \end{aligned} \tag{17}$$

ii) for $r \in \{1, \dots, k\}$ such that $0 \leq |\mathbf{w}|_{1:r} \leq m$,

$$\begin{aligned} &Pr[B_r, W_{1,n+m} = w_1 + m_1, \dots, W_{r,n+m} = w_r + m_r \,|\, K_n = k, \mathbf{M}_n = \mathbf{m}] \\ &\times \binom{m}{w_1, \dots, w_r, m - |\mathbf{w}|_{1:r}} \frac{\frac{C_{r,n+m}(\alpha, \theta, \mathbf{w}+\mathbf{m})}{(\theta+n+m-|\mathbf{w}+\mathbf{m}|_{1:r})_{(|\mathbf{w}+\mathbf{m}|_{1:r})}}}{\frac{C_{r,n}(\alpha, \theta, \mathbf{m})}{(\theta+n-|\mathbf{m}|_{1:r})_{(|\mathbf{m}|_{1:r})}}} \end{aligned} \tag{18}$$

See Section S1.5 of the Supplementary Material for the proof of Theorem 1. Theorem 1 may be viewed as the posterior counterpart of Equation (13), with respect to an initial observable sample $((T_1, X_1), \dots, (T_n, X_n))$. In particular, Equation (17) and Equation (18) provide two posterior distributions of the first r order frequencies under the events A_r and B_r, respectively, for $r \geq 1$. Equation (17) provides the posterior distribution of the first r order frequencies having species' labels not belonging to the additional observable samples; that is the ordering of species' labels in the initial sample $((T_1, X_1), \dots, (T_n, X_n))$ is changed according to the additional sample $((T_{n+1}, X_{n+1}), \dots, (T_{n+m}, X_{n+m}))$. Equation (18) provides the posterior distribution of the first r order frequencies having species' labels belonging to the additional observable samples; that is the ordering of species' labels in the initial sample $((T_1, X_1), \dots, (T_n, X_n))$ is not changed according to the additional sample $((T_{n+1}, X_{n+1}), \dots, (T_{n+m}, X_{n+m}))$. BNP estimators of the first r order frequencies, with respect to a squared loss function, are obtained in terms of posterior expectations, i.e. the vectors of expected values with respect to the posterior distributions (17) and (18). As a corollary of Theorem 1, we obtain the posterior distributions of the frequency of order 1.

Corollary 1. *Let* $((T_1, X_1), \ldots, (T_n, X_n))$ *be a random sample under the BNP model* (12)*, such that the sample features* $K_n = k$ *distinct species with ordered frequencies* $\mathbf{M}_n = \mathbf{m}$*. Let* $((T_{n+1}, X_{n+1}), \ldots, (T_{n+m}, X_{n+m}))$ *be an additional random sample under the same BNP model* (12) *such that the enlarged sample* $((T_1, X_1), \ldots, (T_{n+m}, X_{n+m}))$ *features* K_{m+n} *distinct species with corresponding ordered frequencies* $\mathbf{W}_{n+m}$*, and set* $K_m^{(n)} = K_{n+m} - K_n$*. If* A_1 *and* B_1 *are the events defined in* (15) *and* (16)*, respectively, then it holds:*

i)

$$Pr[A_1, W_{1,n+m} = w_1, K_m^{(n)} \geq 1 \,|\, K_n = k, \mathbf{M}_n = \mathbf{m}] \tag{19}$$
$$= \binom{m}{w_1} \frac{\alpha(n+m-w_1)+\theta w_1}{(n+m)(\theta+n+m-w_1)_{(w_1)}} (1-\alpha)_{(w_1-1)};$$

ii)

$$Pr[B_1, W_{1,n+m} = w_1 + m_1 \,|\, K_n = k, \mathbf{M}_n = \mathbf{m}] \tag{20}$$
$$= \binom{m}{w_1} \frac{\frac{\alpha(n+m-w_1-m_1)+\theta(w_1+m_1)}{(n+m)(\theta+n+m-w_1-m_1)_{w_1+m_1}}}{\frac{\alpha(n-m_1)+\theta m_1}{n(\theta+n-m_1)_{(m_1)}}} (m_1-\alpha)_{(w_1)};$$

iii)

$$Pr[W_{1,n+m} = w \,|\, K_n = k, \mathbf{M}_n = \mathbf{m}] \tag{21}$$
$$= \frac{[\alpha(n+m-w)+\theta w](1-\alpha)_{(w-1)}}{(n+m)(\theta+n+m-w)_{(w)}}$$
$$\times \left[\mathbb{1}_{\{1,\ldots,m\}}(w) \binom{m}{w} \right.$$
$$\left. + \mathbb{1}_{\{m_1,\ldots,m_1+m\}}(w) \binom{m}{w-m_1} \frac{n(\theta+n-m_1)_{(m_1)}}{[\alpha(n-m_1)+\theta m_1](1-\alpha)_{(m_1-1)}} \right].$$

The proofs of Equation (19) and Equation (20) follow directly from Theorem 1 by setting $r = 1$, whereas Equation (21) follows by combining (19) and (20). Within the population genetic setting of Donnelly and Tavaré (1986), Equation (21) with $\alpha = 0$ provides the posterior distribution of the frequency of the oldest alleles. By exploiting Corollary 1, BNP estimators of the frequency of order 1, with respect to a squared loss function, are obtained in terms of the expected values of the posterior distributions (19), (20) and (21). Here, we report the BNP estimator with respect to the posterior distribution (21) and we refer to Section S1.6 of the Supplementary Material for the BNP estimators with respect to the posterior distributions (19) and (20). In particular, if we set

$$C(\alpha, \theta, n, m, m_1)$$
$$= [\alpha(n+m-m_1)+\theta m_1] \left[m_1 + m \frac{m_1-\alpha}{\theta+n-\alpha} \right]$$

$$+\left[m(\theta-\alpha)\frac{m_1-\alpha}{\theta+n-\alpha}\right]\left[(m_1+1)+(m-1)\frac{m_1+1-\alpha}{\theta+n+1-\alpha}\right]$$

then

$$\begin{aligned}&\mathbb{E}[W_{1,n+m}\,|\,K_n=k,\mathbf{M}_n=\mathbf{m}] \\ &=\frac{m(\theta+n+1-\alpha)_{(m)}}{(n+m)(\theta+n)_{(m)}}\frac{\theta+n\alpha}{\theta+n+1-\alpha}+\frac{n(\theta+n-\alpha)_{(m)}}{(n+m)(\theta+n)_{(m)}}\frac{C(\alpha,\theta,n,m,m_1)}{\alpha(n-m_1)+\theta m_1}.\end{aligned} \tag{22}$$

We refer to Section S1.7 of the Supplementary Material for explicit expressions of the posterior probabilities of the events A_1 and B_1, which complete the main result of Corollary 1.

3.2 Estimation of prior's parameter (α, θ)

The closed-form expressions of our results facilitate posterior inferences. In particular, if the prior's parameters (α, θ) are estimated with an empirical Bayes approach and fixed, then the inferential procedure becomes straightforward and efficient. Here we consider both an empirical Bayes approach and a fully Bayes approach for the estimation of (α, θ), the latter considering the specification of a prior distribution on (α, θ). While the empirical Bayes approach takes advantage of closed-form formulae, the fully Bayes approach may sometimes be preferable. In both cases, the inference is based on an initial sample of n observations, which are then used to make predictions on a second set of m data points.

Within the empirical Bayes approach, we consider methods relying on maximum likelihood estimation and methods relying on moment-based estimation. With regards to maximum likelihood estimation, the problem consists in finding the values of α and θ that maximize the (marginal) likelihood function, which in this context is equal to the EPPF. Under the ordered PYP, this coincides with (8), and the parameters found by solving:

$$\max_{\alpha,\theta}\Phi_k^{(n)}(m_1,\ldots,m_K;\alpha,\theta).$$

As a term of comparison, we also consider the performance of estimating the prior's parameters when the model is misspecified, specifically assuming an ordered DP prior (i.e. fixing $\alpha = 0$), or ignoring the ordering structure or the model, i.e. maximizing the EPPF of the standard PYP (3). The sets of prior's parameters obtained by optimizing these EPPFs are respectively denoted with `ordPYP`, `ordDP` and `stdPYP`. Note that the approaches based on the "misspecified" likelihood (`ordDP` and `stdPYP`) do not take full advantage of the increased complexity of the model, and are considered only as reference. When the model is correctly specified, these methods do not estimate the correct parameters.

With regard to moment-based estimation, we consider a statistic of interest and then match its population first moment with the corresponding observed sample statistic. Rather than doing this for the full initial sample, we consider a collection of samples of increasing size and match the statistics of interest's trajectory given by increasing

sample sizes, using a least squares method. Specifically, we consider a grid of sample size values $1 \leq n_1 \leq \ldots \leq n_d = n$; for each n_i, we compute the discrepancy between the first moment and the observed statistics for the first n_i observations; finally, we minimize the sum of the squared discrepancies, where the sum is for $i = 1, \ldots, d$. The statistics we consider are the frequency of the first ordered species $M_{1,n}$ and the number of distinct species K_n; the parameters obtained are respectively denoted as `lsM1` and `lsK`:

$$(\hat{\alpha}, \hat{\theta})_{\texttt{lsM1}} = \arg\min_{\alpha,\theta} \sum_{i=1}^{d} \left(\mathbb{E}[M_{1,n_i}; \alpha, \theta] - M_{1,n_i}\right)^2,$$

$$(\hat{\alpha}, \hat{\theta})_{\texttt{lsK}} = \arg\min_{\alpha,\theta} \sum_{i=1}^{d} \left(\mathbb{E}[K_{n_i}; \alpha, \theta] - K_{n_i}\right)^2,$$

where M_{1,n_i} and K_{n_i} are the frequency of the first ordered species and the number of distinct species in the first n_i samples, respectively. In Section S2.1 of the Supplementary Material, we provide additional details, as well as report the pseudocode algorithms, to obtain these parameter estimates. Moment-based estimation allows to focus the estimation problem on a specific feature or property of the data, by choosing the summary statistics of interest. Moreover, by considering a grid of sample size values, the method learns the growth curve over n, ideally being more robust compared to methods that only look at one "snapshot" given by the full dataset. Additionally, because they do not rely on the full likelihood, these methods could be more robust in the case of model misspecification.

We also consider a fully Bayes approach (`FB`), by specifying a suitable prior distribution for the parameters (α, θ) and focusing the inference on the posterior distribution. This can be implemented using standard MCMC algorithms. Here, we consider independent non-informative prior distributions, setting $p(\theta) = G(0.1, 0.1)$ and $p(\alpha) = Unif(0, 1)$.

4 Numerical illustrations

We empirically study the ordered PYP and assess the performance of our BNP approach for estimating the frequency of the first ordered species, using synthetic and real data. Moreover, we empirically study distributional properties of the ordered random partitions induced by the model, in particular focusing on the species' ordering distribution. In the synthetic data, we compare the performance when the data is generated from the model, i.e. the PYP prior is correctly specified, and when the data is generated from different distributions, thus under model misspecification. For the application of our model to real data, we analyze genetic variation using samples from the 1000 Genome Project (1000 Genomes Project Consortium, 2015), which we combine with variants' age estimates obtained from the Human Genome Dating Project (Albers and McVean, 2020). Studying genetic variation is of great importance to investigate population structure, to detect related samples, to investigate demographic history, and to learn about the risk of diseases and different quantitative traits. By also incorporating information on the

variants' age, similar and further issues can be assessed, such as using the variants' age distribution to compare populations, differentiating age distributions in pathogenic and benign variants, and learning about genealogical history (Albers and McVean, 2020). Here we assess the performance of our method on predicting frequencies of variants ordered by their age, focusing in particular on the oldest variant. Code is available at https://github.com/cecilia-balocchi/OrderedSSP.

4.1 Preliminaries

While the distributional properties of the random partition induced by the PYP are well-known in the BNP literature, the properties induced by the ordered PYP are less understood. Because of the marginality property (9), some distributional properties of the ordered PYP are equivalent to the ones of the PYP. For example, the distribution of the number of distinct species (or clusters) induced by the ordered PYP is equal to the one induced by the PYP, reported in equation (S3). However, other properties that relate to the ordering on the species are not as well understood. In particular, we are interested in learning the behavior of the distribution that assigns the order to a new cluster. We aim to characterize it using simple descriptive features. We achieve this goal by studying the predictive distribution (6) that assigns the $n+1$ observation to a *new* species of order j, and to examine its behavior marginalizing on all the configurations of partitions of n observations into K_n species. In other words, we empirically study the marginal distribution $\Pr(\text{order}(n+1) = j | (T_{n+1}, X_{n+1}) = (T^*_{\text{new}}, X^*_{\text{new}}), K_n = k)$ that a new species after n observations is assigned order j, given that the previous n observations are partitioned into $K_n = k$ species, for $j = 1, \dots, K_n+1$. This distribution is "marginal" compared to (6), because it does not condition on the frequencies of the ordered species, $\mathbf{M}_n = (m_1, \dots, m_k)$, and it is obtained by marginalizing over all possible configurations $(m_1, \dots, m_k)$ of partitions of n into $K_n = k$ species.

In Figure 2 we depict the marginal ordering distribution for a new species given an observed sample of size $n = 10$. The solid color lines represent the ordering distributions given the number of previously observed clusters, marginally on the partition configuration (different colors correspond to different numbers of clusters), $\Pr(\text{order}(n+1) = j | (T_{n+1}, X_{n+1}) = (T^*_{\text{new}}, X^*_{\text{new}}), K_n = k)$. The colored points instead represent the realizations of the ordering probabilities conditional on individual partition configurations $\mathbf{M}_n$, $\Pr(\text{order}(n+1) = j | (T_{n+1}, X_{n+1}) = (T^*_{\text{new}}, X^*_{\text{new}}), K_n = k, \mathbf{M}_n = (m_1, \dots, m_k))$, for different values of $(m_1, \dots, m_k)$; note that this conditional ordering probability can be found from (6) as $q_j^{(new)} / \sum_{i=1}^{k+1} q_i^{(new)}$. The ordering distribution has been sketched for different configurations (parameters) of the PYP prior. In particular, from left to right, we represent the distribution under the ordered DP ($\theta > 0, \alpha = 0$, first panel), the ordered PYP with $\alpha < \theta$, $\alpha = \theta$, and $\alpha > \theta$ and the ordered α-stable process ($\theta = 0, \alpha > 0$, last panel). Figure 2 shows that the ordering distribution changes depending on the parameters θ and α: for $\theta > \alpha$ the probability that a new cluster is assigned to order j increases with j for each previous number of species K_n (first and second panels from the left) and the increasing trend in stronger when the difference $\theta - \alpha$ is large. For $\theta = \alpha$ (third panel from the left) the trend is constant over j, for all K_n. For $\theta < \alpha$ we see instead that the trend is decreasing with j, for each K_n (fourth

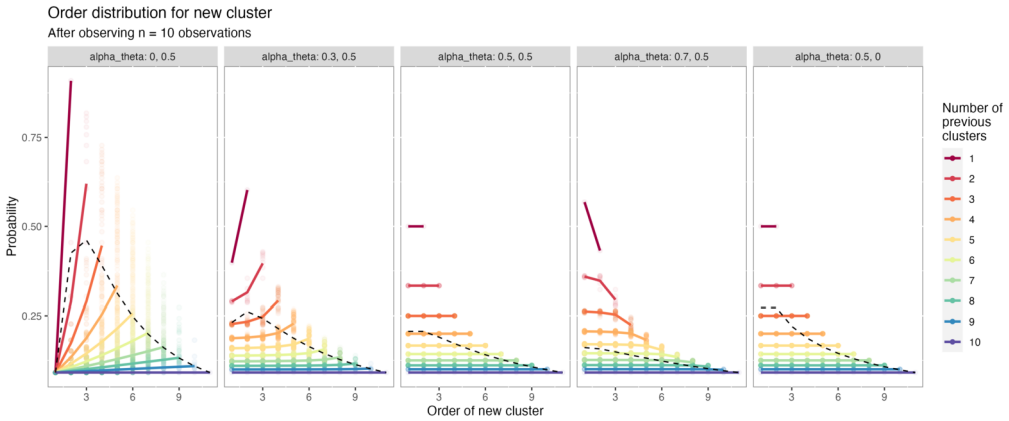

Figure 2: Order distribution for a new species, given a sample of $n = 10$ observations, divided into $K_n = 1, \dots, 10$ species (each color represents a different value of K_n). Solid colored curves represent the ordering distributions marginally on the partition configuration $\mathbf{M}_n$, $\Pr(\text{order}(n+1) = j | (T_{n+1}, X_{n+1}) = (T^*_{\text{new}}, X^*_{\text{new}}), K_n = k)$; the colored points instead show the variation of the ordering distribution conditional on $\mathbf{M}_n$, across different partition configurations $\mathbf{M}_n = (m_1, \dots, m_k)$. The dashed black line represents the order distribution for a new species, marginally on the number of previous species K_n.

panel). The last panel shows that in the case of the α-stable process ($\theta = 0 < \alpha$) the trend is again constant. These intuitions are useful for constructing a distribution for ordered partitions that has properties similar to those induced by the ordered PYP prior. Moreover, Figure 2 emphasizes an additional aspect of the ordered PYP's improved flexibility compared to the ordered DP ($\alpha = 0$).

4.2 Analysis of synthetic data

We consider the problem of making inference on some quantities of interest with our BNP approach, in the context of synthetic data. In particular, we study the performance on an additional sample of size m of the posterior mean predictors for the total number of species K_{n+m}, the frequency of the oldest cluster $W_{1,n+m}$, and the frequency $W_{1,n+m}$ conditionally on the knowledge that either the event A_1 or B_1 happened ($W_{1,n+m}|A_1$ and $W_{1,n+m}|B_1$). For notational simplicity, we will remove the dependence on $n+m$ in the notation. All of the estimators for these quantities have closed-form expressions, thanks to the results in Section 3. In particular, we estimate K_{n+m} using the posterior mean of the number of unseen species $K_m^{(n)}$ (see Section S1.3 of the Supplementary Materials), while $W_{1,n+m}$ is estimated using (22), and $W_{1,n+m}|A_1$ and $W_{1,n+m}|B_1$ using respectively combining formulas (S14) with (S20), and (S19) with (S21) from the Supplementary Materials. We compare the predictive performance of the ordered PYP model, under a full Bayes approach, and when the prior's parameters are estimated with the empirical Bayes methods described in Section 3.2. We generate the synthetic

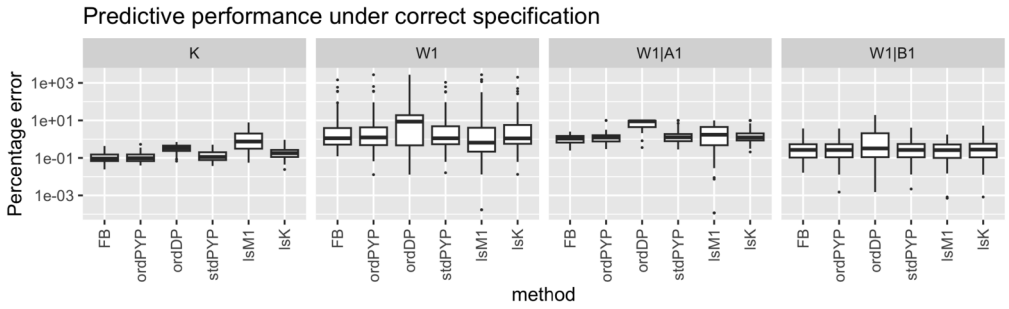

Figure 3: Predictive performance for our BNP approach for the total number of distinct species K, the frequency of the first ordered cluster W_1, and the conditional frequencies $W_1|A_1$ and $W_1|B_1$ (shown in different columns). The boxplots display the median absolute percentage error between the predicted posterior mean and the true value, across several datasets simulated from the ordered PYP model. The full Bayes approach and different parameter-estimating methods are compared.

data under different scenarios. We first consider a framework where the model is correctly specified, i.e. the data is generated from the ordered PYP prior. We then focus on a framework where the model is misspecified, as the data is generated from different distributions.

Inference of the first ordered frequency under correct specification

We first generate the data from the ordered PYP prior, i.e. under correct model specification. Specifically, we consider 100 datasets of size $n = 500$, generated from the ordered PYP prior with randomly sampled parameters (θ, α); for each dataset, we consider 25 additional datasets of size $m = 5000$, and compute the median prediction error across the 25 additional datasets (absolute percentage error is computed for the four quantities of interest K, W_1, $W_1|A_1$, $W_1|B_1$).

Figure 3 compares the performance of the full Bayes approach with the different empirical Bayes approaches, for the four quantities of interests. Overall we notice that the performance of the full Bayes approach (`FB`) is almost identical to the one of the approach based on the EPPF of the ordered PYP (`ordPYP`). These two approaches tend to have the best prediction error for all the quantities of interest, except for the estimation of the frequency of the first ordered species W_1, where the best predictive performance is achieved by `lsM1`. The empirical Bayes approach based on the standard PYP likelihood `stdPYP` has a similar but slightly worse performance to `ordPYP` and `FB`. `ordDP` instead shows poor predictive performance for all quantities of interests.

Inference of the first ordered frequency under misspecification

We then consider a synthetic data framework where the data is not generated from the ordered PYP prior, i.e. under model misspecification. We aim at evaluating how our

method performs in such adverse conditions, which are in general also to be expected from real data. We consider several different data-generating processes. In all of them, we generate the species order independently from the observations' species (or cluster) assignment. For the clustering distribution, we consider: (a) the standard `DP`, (b) the standard `PYP`, and (c) the (infinite support) `Zipf` distribution (also known as the zeta distribution). In the latter, each observation is associated with an integer (sampled from the Zipf distribution) and clusters are formed by aggregating observations mapped to the same integer. Given that these distributions do not induce an order on the species, we additionally consider an ordering distribution for each new species. For the ordering distribution, we consider: (1) the `alpha-stable` distribution (induced by the ordered α-stable process, a special case of the ordered PYP with $\theta = 0$ where the predictive distribution (6) simplifies significantly), and (2) what we call the `arrival-weighted` distribution. The latter considers the ordering induced by the cluster arrival and introduces a random component by sampling for each new cluster an exponentially distributed weight with mean given by the arrival order, and ordering the species according to the weight. We generate 100 datasets of size $n = 500$ using randomly sampled parameters, and for each of them consider 25 additional datasets of size $m = 5000$; we consider the median absolute percentage errors across the 25 additional datasets.

Figure 4 displays the predictive performance results for the different measures of interests (K, W_1, $W_1|A_1$ and $W_1|B_1$) across different rows, and for different data-generating distributions (clustering and ordering distributions) across different columns. We compare the performance of the full Bayes and of the different empirical Bayes approaches. The first row of Figure 4 focuses on the prediction of the number of distinct species K. Overall, the best prediction error is achieved by the empirical Bayes method maximizing the likelihood of the non-ordered (standard) PYP (`stdPYP`). This is not surprising given that the ordered PYP is not correctly specified, and the number of distinct species has the same behavior under the ordered and the standard PYP. The second best performance is achieved by `lsK`, and this is consistent with the fact that it was designed to be more robust and learn this "feature" (K) even under model misspecification. The performance of the full Bayes approach `FB` and the empirical Bayes approach based on the ordered PYP likelihood `ordPYP` are similar and slightly worse when the clustering distribution is the PYP or the Zipf distribution, but it gets considerably worse when the clustering distribution is the DP, suggesting that these methods are not able to learn that α is equal to zero in such simulated datasets. The performance of `ordDP` is good under the DP clustering distribution, but quite bad otherwise.

The second row of Figure 4 shows the error committed when predicting the frequency of the first ordered species W_1. The best performance is often achieved by `lsM1`, followed by `ordPYP` and `FB`. When the clustering distribution is the Zipf distribution, the difference between `lsM1` and the other two is considerable, with the former being more robust. Sometimes, `stdPYP` achieves comparable results. Overall, the distribution of the errors is much more spread out when the ordering is generated from the `arrival-weighted` distribution, compared to the `alpha-stable` distribution, meaning that the behavior induced by the former is more different from the one described by 22. Finally, the third and the fourth rows focus on the predictive performance for the conditional frequencies

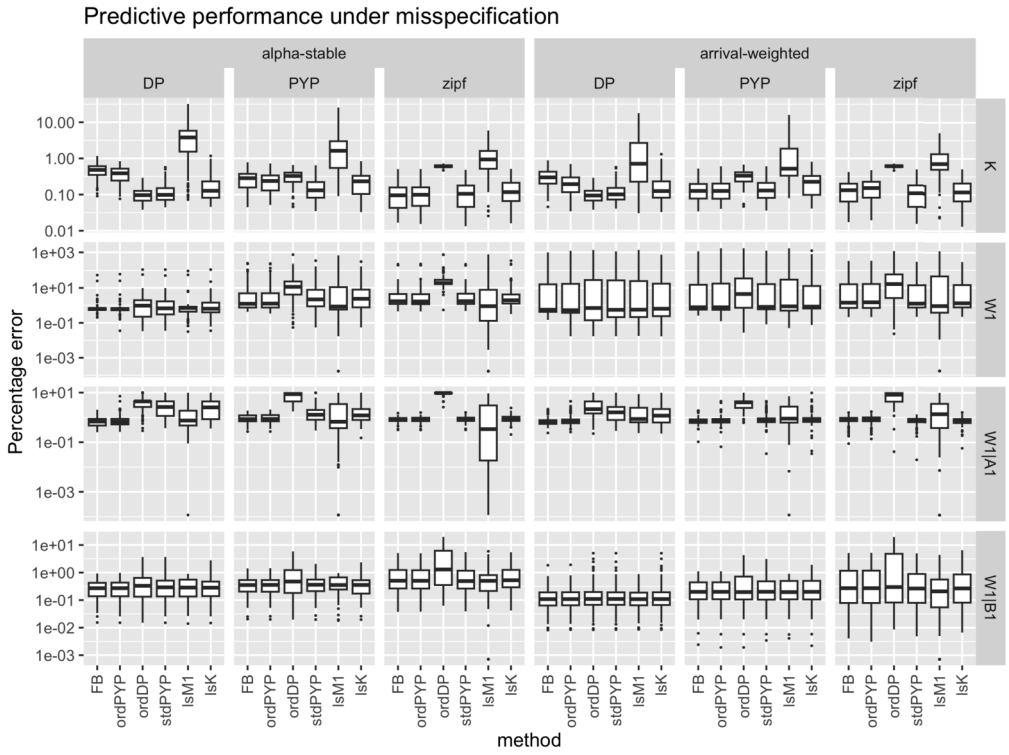

Figure 4: Predictive performance for our BNP approach, for the total number of clusters K, the frequency of the first ordered cluster W_1, and the conditional frequencies $W_1|A_1$ and $W_1|B_1$ (shown in different rows). The boxplots display the median absolute percentage error between the predicted posterior mean and the true value, across several datasets simulated under different distributions (shown in different columns). We compare various parameter-estimating methods.

$W_1|A_1$ and $W_1|B_1$. In both cases, the best performance is achieved by `ordPYP`, `FB`, and `lsM1`.

Overall, the full Bayes approach achieves results very similar to the ones of the empirical Bayes approach based on the likelihood of the ordered PYP (`ordPYP`). When the model is misspecified, `lsM1` seems to be more robust for the estimation of both the marginal distribution and the conditional distribution of W_1, but it performs quite poorly for the prediction of K. While the parameter-estimation method `ordPYP` and the full Bayes approach `FB` do not always produce the best predictions, they seem to provide a good balance between learning the number of species K and the frequency of the first order species W_1. Alternatively, the parameter estimation method could be selected based on the quantity of interest.

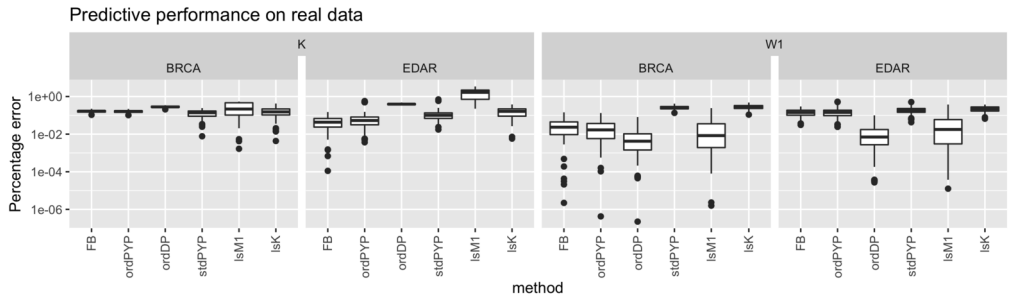

Figure 5: Prediction performance for the number of distinct species K and the frequency of the oldest cluster W_1 across several training-testing sets for BRCA2 (panels one and three) and EDAR (panels two and four).

4.3 Analysis of genetic variation data

We employ our method to analyze genetic variation data from the 1000 Genomes Project (1000 Genomes Project Consortium, 2015). We examine single nucleotide polymorphisms (SNPs) corresponding to certain genes for a sample of 2548 individuals. In particular, we consider a variant to be present for a given individual and SNP locus if the present DNA base is different from the reference genome in either allele. We combine the variants from all individuals and all SNPs corresponding to a certain gene, to create our basic sample, where the variant location (i.e. the SNP) represents the species each sample unit belongs to. To associate an ordering to each species, we used SNPs age estimates from Human Genome Dating project (Albers and McVean, 2020), and discarded any variant for which no age information is available. We study two datasets, formed by the genetic variants corresponding respectively to the BRCA2 and the EDAR genes, for which we analyze respectively 1482 and 1073 unique SNPs. We focus on the predictive performance for the number of distinct variants K and the frequency of the oldest variant W_1, using the different parameter estimation methods for our BNP approach. We repeat our analysis for 100 different training and testing sets, randomly sampled so that the testing set size is approximately 20 times larger than the training set size.

In Figure 5 we report the percentage errors for predicting K and W_1 for the datasets corresponding to the BRCA2 and the EDAR genes. For the prediction of the number of species, we find the method `stdPYP` provides the best results for the BRCA2 dataset, but that `ordPYP` and `FB` have the best prediction for the EDAR dataset. In terms of predicting the frequency of the oldest variant, `ordDP` and `lsM1` perform better for both. We note that the performances of `ordPYP` and `FB` are comparable to that of `ordDP` in the BRCA2 dataset, as the former often estimated values of α close to zero; however, the two methods achieved different performance in the EDAR dataset, due to the fact that `ordPYP` and `FB` estimated positive values of α. Under further inspection, we note that `lsM1` has performance comparable to `ordDP` in the prediction of W_1 for the EDAR dataset because it often estimates values of α very close to zero. It is also surprising to see that these datasets seem to have power-law behavior in terms of the number of species, for which values of α greater than 0 provide more accurate predictions, but the

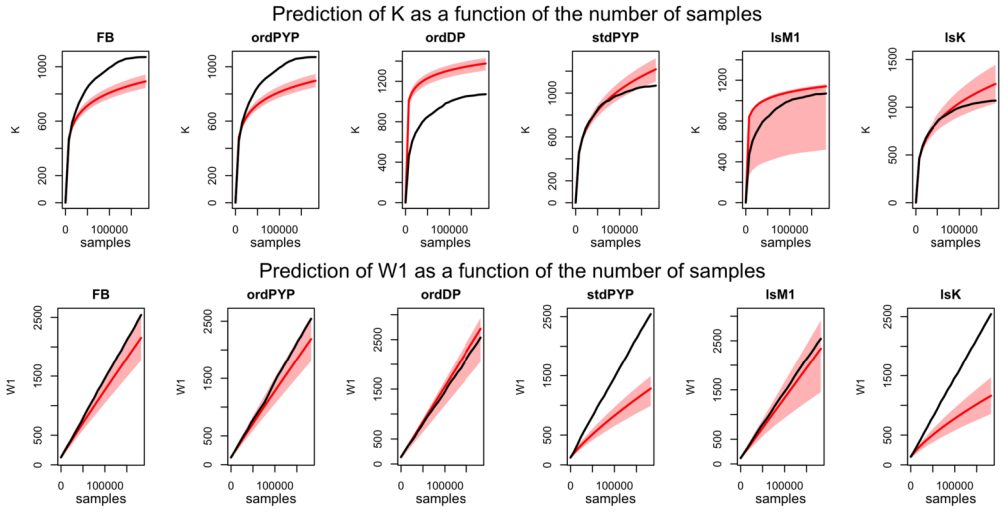

Figure 6: Prediction of the curve of K (top panels) and W_1 (bottom panels) as a function of sample size, for the BRCA2 dataset. The curves represented correspond to the training-testing split that resulted in the median error, with the observed curve depicted in black, the predicted curve in red, and the red confidence bands corresponding to 95% empirical quantiles.

best predictions for the frequency of the oldest variant W_1 are produced by methods that estimate $\alpha = 0$. This is probably due to the model being not correctly specified for these data. Thus we recommend analyzing the two prediction problems separately.

Our method can also be used to analyze the predictions as functions of the number of samples, rather than simply make a prediction for the entire test set. Using the multiple training and testing splits method, we also looked at the performance in the prediction of the curve of K and W_1. Figure 6 reports the results for each parameter-estimating method, applied to the BRCA2 dataset. Similar plots are reported in Section S2.2 of the Supplementary Material for the EDAR gene dataset. For simplicity of visualization, we depict in black the actual curve of K (or respectively W_1) observed in the training-testing split that most closely represents the median error achieved by each method. The prediction curve (depicted in red) is also the one corresponding to the "median" training-testing split, while the red bands represent the 95% confidence bands. We report the empirical confidence bands, computed using the empirical quantiles of the curve estimates across the various training-testing splits. For the estimation of the number of distinct species K, the top panel of Figure 6 confirms that `stdPYP` and `lsK` have the best performance, with quite accurate prediction and with the represented curve falling within the 95% bands. In the other parameter-estimating methods the prediction is much worse, and often the curve does not fall within the confidence bands. In terms of predictions for W_1, the results are consistent with the expectations from the predictions on the whole test set, with `ordDP`, `ordPYP`, `FB` and `lsM1` performing better, with very accurate prediction of W_1 as a function of the number of additional samples.

5 Discussion

A common feature of SSPs is the invariance with respect to species labeling, i.e. species' labels are immaterial in the definition of the functional of interest, which is at the core of the development of the BNP approach to SSPs under the popular PYP prior (Lijoi et al., 2007, 2008; Favaro et al., 2009, 2013). In this paper, we considered SSPs that are not "invariant" to species labeling, in the sense that an ordering or ranking is assigned to species' labels, and we developed a BNP approach to such problems. In particular, inspired by the seminal work of Donnelly and Tavaré (1986) on age-ordered alleles' compositions, with a renowned interest in the frequency of the oldest allele (Crow, 1972; Kelly, 1977; Watterson and Guess, 1977), we studied the following SSP with ordering: given an observable sample from an unknown population of individuals belonging to some species (alleles), with species' labels being ordered according to some weights (ages), estimate the frequencies of the first r order species' labels in an enlarged sample obtained by including additional unobservable samples from the same population. Our BNP approach relies on the ordered PYP prior, which leads to an explicit posterior distribution of the first r order frequencies, with corresponding estimates being simple and also computationally efficient. We presented an empirical validation of our approach on both synthetic and real data. The proposed methodology has been applied to the analysis of the genetic variation, showing its effectiveness in the estimation of the frequency of the oldest allele.

The sampling structure of the ordered PYP has been first presented in Gnedin and Pitman (2005) and James (2006a), and since then no other works have further investigated such a sampling structure in BNPs. To date, the sampling formulae of ordered PYP prior appear to be largely unknown and unexplored in the BNP literature. Our work highlights the great potential of the ordered PYP in BNP inference for SSPs with ordering, paving the way for future research. First, in our work we considered the problem of estimating the frequency of the first r order species, which is arguably the most natural SSP with ordering; other (discrete) functionals of the ordered species' composition of unobservable samples may be of interest, e.g. the number of unseen species with order less than r that would be observed in additional samples, and the number of unseen species observed with a certain prevalence and order less than r in the enlarged sample. Second, while SSPs with ordering have a natural motivation in population genetics, they may be of interest in different fields; in Section S2.3 of the Supplementary Material, we present an application in the context of citations to academic articles, where each cited article is a species whose order is determined by the publication date; another application is in the context of online purchasing of items, with order being determined by a suitable measure of popularity assigned to items. When analyzing these applications in terms of SSMs, some simplifications need to be considered, which might impact certain aspects of the analysis. For example, our model cannot consider the case of different species having the same ranking (such as different papers published on the same date). However, we argue that it is still interesting to understand if the ordered PYP is an adequate model to describe the behaviors observed in such applied contexts.

We believe that interest in ordered SSMs is not limited to BNP inference for SSPs with ordering. Ordered SSMs, and in particular the ordered PYP prior, may be also

applied to a setting in which species are not explicitly observed, but need to be inferred. This is the case of Bayesian mixture models (Frühwirth-Schnatter et al., 2019), where each observation is assumed to be assigned to a latent component (species' label), which characterizes some features of the distribution of the observations. In such a setting, ordered SSMs may be used to specify the distribution of the mixture's probabilities of the latent components, with such components being ordered according to some weights (James, 2006b). Ordered SSMs admit a natural extension to the features sampling framework, which generalizes the species sampling framework by allowing each observation to belong to multiple species, now called features. Feature sampling problems first appeared in ecology for modeling the presence or absence of an animal in a trap, and their importance has grown dramatically in recent years driven by numerous applications in biological and physical sciences (Camerlenghi et al., 2022; Masoero et al., 2022). In such a context, the Beta process prior (Broderick et al., 2013) is the most popular nonparametric prior for modeling the unknown feature's composition of the population. The definition of an ordered version of the Beta process prior, and generalizations thereof, would be the starting point to introduce and investigate feature sampling problems with ordering.

Acknowledgments

The authors are grateful to the Associate Editor and two anonymous Referees for their comments and corrections that allow them to improve remarkably the paper. The authors wish to thank Paul Jenkins for useful discussions on the use of age-ordered random partitions in population genetics. Federico Camerlenghi is a member of the *Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni* (GNAMPA) of the *Istituto Nazionale di Alta Matematica* (INdAM). Stefano Favaro is also affiliated to IMATI-CNR "Enrico Magenes" (Milan, Italy).

Funding

The first and third authors gratefully acknowledge funding from the European Research Council (ERC), under the European Union's Horizon 2020 research and innovation programme, Grant agreement No. 817257. The second and third authors gratefully acknowledge the support from the Italian Ministry of University and Research (MUR), "Dipartimenti di Eccellenza" grant 2023-2027. The second author was supported by the European Union – Next Generation EU funds, component M4C2, investment 1.1., PRIN-PNRR 2022 (P2022H5WZ9).

Supplementary Material

Supplementary Material to "A Bayesian Nonparametric Approach to Species Sampling Problems with Ordering"
(DOI: 10.1214/24-BA1418SUPP; .pdf). A manuscript containing detailed proofs of all results presented in this paper, and additional results regarding synthetic and real data evaluation.

References

1000 Genomes Project Consortium (2015). "A global reference for human genetic variation." *Nature*, 526(7571): 68. 3, 14, 20

Albers, P. K. and McVean, G. (2020). "Dating genomic variants and shared ancestry in population-scale sequencing data." *PLoS biology*, 18(1): e3000586. 3, 14, 15, 20

Bacallado, S., Battiston, M., Favaro, S., and Trippa, L. (2017). "Sufficientness postulates for Gibbs-type priors and hierarchical generalizations." *Statistical Science*, 487–500. MR3730518. doi: https://doi.org/10.1214/17-STS619. 5

Balocchi, C., Favaro, S., and Naulet, Z. (2022). "Bayesian nonparametric inference for "species-sampling" problems." *arXiv preprint arXiv:2203.06076*. 2

Balocchi, C., Camerlenghi, F., and Favaro, S. (2024). "Supplementary Material to "A Bayesian Nonparametric Approach to Species Sampling Problems with Ordering"" *Bayesian Analysis*. doi: https://doi.org/10.1214/24-BA1418SUPP. 4

Broderick, T., Jordan, M. I., and Pitman, J. (2013). "Cluster and feature modeling from combinatorial stochastic processes." *Statistical Science*, 28(3): 289 – 312. MR3135534. doi: https://doi.org/10.1214/13-STS434. 23

Camerlenghi, F., Favaro, S., Masoero, L., and Broderick, T. (2022). "Scaled process priors for Bayesian nonparametric estimation of the unseen genetic variation." *Journal of the American Statistical Association*, 1–12. 23

Crow, J. F. (1972). "The dilemma of nearly neutral mutations: how important are they for evolution and human welfare?" *Journal of Heredity*, 63(6): 306–316. 3, 22

Daley, D. J. and Vere-Jones, D. (2008). *An introduction to the theory of point processes: volume II: general theory and structure, second edition*. Springer, New York. MR2371524. doi: https://doi.org/10.1007/978-0-387-49835-5. 8

Deng, C., Daley, T., De Sena Brandine, G., and Smith, A. D. (2019). "Molecular Heterogeneity in Large-Scale Biological Data: Techniques and Applications." *Annual Review of Biomedical Data Science*, 2: 39–67. 2

Donnelly, P. and Tavaré, S. (1986). "The ages of alleles and a coalescent." *Advances in Applied Probability*, 18(1): 1–19. MR0827330. doi: https://doi.org/10.2307/1427237. 2, 3, 7, 8, 10, 12, 22

Efron, B. and Thisted, R. (1976). "Estimating the number of unseen species: How many words did Shakespeare know?" *Biometrika*, 63(3): 435–447. 1

Ewens, W. J. (1972). "The sampling theory of selectively neutral alleles." *Theoretical population biology*, 3(1): 87–112. MR0325177. doi: https://doi.org/10.1016/0040-5809(72)90035-4. 2, 5

Favaro, S. and James, L. F. (2016). "Relatives of the Ewens sampling formula in Bayesian nonparametrics." *Statistical Science*, 31(1): 30 – 33. MR3458589. doi: https://doi.org/10.1214/15-STS538. 7

Favaro, S., Lijoi, A., Mena, R. H., and Prünster, I. (2009). "Bayesian non-parametric inference for species variety with a two-parameter Poisson–Dirichlet process prior." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(5): 993–1008. MR2750254. doi: https://doi.org/10.1111/j.1467-9868.2009.00717.x. 2, 22

Favaro, S., Lijoi, A., and Prünster, I. (2013). "Conditional formulae for Gibbs-type exchangeable random partitions." *The Annals of Applied Probability*, 23(5): 1721 – 1754. MR3114915. doi: https://doi.org/10.1214/12-AAP843. 2, 22

Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems." *The annals of statistics*, 209–230. MR0350949. 2

Frühwirth-Schnatter, S., Celeux, G., and Robert, C. P. (2019). *Handbook of mixture analysis*. CRC press. MR3889980. 23

Gnedin, A. and Pitman, J. (2005). "Regenerative composition structures." *The Annals of Probability*, 33(2): 445–479. MR2122798. doi: https://doi.org/10.1214/009117904000000801. 3, 5, 6, 7, 22

Gnedin, A. V. (2010). "Regeneration in random combinatorial structures." *Probability Surveys*, 7: 105–156. MR2684164. doi: https://doi.org/10.1214/10-PS163. 7

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(1-2): 45–63. MR0077039. doi: https://doi.org/10.1093/biomet/43.1-2.45. 1

Griffiths, R. (1979). "Exact sampling distributions from the infinite neutral alleles model." *Advances in Applied Probability*, 11(2): 326–354. MR0526416. doi: https://doi.org/10.2307/1426843. 2

Hajra, K. B. and Sen, P. (2005). "Aging in citation networks." *Physica A: Statistical Mechanics and its Applications*, 346(1-2): 44–48. 4

James, L. F. (2006a). "Poisson calculus for spatial neutral to the right processes." *The Annals of Statistics*, 34(1): 416–440. MR2275248. doi: https://doi.org/10.1214/009053605000000732. 3, 5, 6, 7, 8, 22

James, L. F. (2006b). "Spatial neutral to the right species sampling mixture models." *arXiv preprint math/0604266*. MR2416127. doi: https://doi.org/10.1142/9789812708298_0021. 23

Kelly, F. (1977). "Exact results for the Moran neutral allele model." *Advances in Applied Probability*, 9(2): 197–201. 3, 10, 22

Kingman, J. F. (1975). "Random discrete distributions." *Journal of the Royal Statistical Society: Series B (Methodological)*, 37(1): 1–15. MR0368264. 2

Kingman, J. F.(1978). "The representation of partition structures." *Journal of the London Mathematical Society*, 2(2): 374–380. MR0509954. doi: https://doi.org/10.1112/jlms/s2-18.2.374. 5

Kingman, J. F. C. (1993). *Poisson processes*, volume 3. Clarendon Press. MR1207584. 8

Lijoi, A., Mena, R. H., and Prünster, I. (2007). "Bayesian nonparametric estimation of the probability of discovering new species." *Biometrika*, 94(4): 769–786. MR2416792. doi: https://doi.org/10.1093/biomet/asm061. 1, 2, 22

Lijoi, A., Prünster, I., and Walker, S. G. (2008). "Bayesian nonparametric estimators derived from conditional Gibbs structures." *The Annals of Applied Probability*, 18(4): 1519–1547. MR2434179. doi: https://doi.org/10.1214/07-AAP495. 2, 22

Masoero, L., Camerlenghi, F., Favaro, S., and Broderick, T. (2022). "More for less: predicting and maximizing genomic variant discovery via Bayesian nonparametrics." *Biometrika*, 109(1): 17–32. MR4374638. doi: https://doi.org/10.1093/biomet/asab012. 23

Mathieson, I. and McVean, G. (2014). "Demography and the age of rare variants." *PLoS genetics*, 10(8): e1004528. 3

Orlitsky, A., Suresh, A. T., and Wu, Y. (2016). "Optimal prediction of the number of unseen species." *Proceedings of the National Academy of Sciences*, 113(47): 13283–13288. MR3582444. doi: https://doi.org/10.1073/pnas.1607774113. 1

Perman, M., Pitman, J., and Yor, M. (1992). "Size-biased sampling of Poisson point processes and excursions." *Probability Theory and Related Fields*, 92(1): 21–39. MR1156448. doi: https://doi.org/10.1007/BF01205234. 4

Pitman, J. (1995). "Exchangeable and partially exchangeable random partitions." *Probability theory and related fields*, 102(2): 145–158. MR1337249. doi: https://doi.org/10.1007/BF01213386. 4, 5

Pitman, J. (1996). "Some developments of the Blackwell-MacQueen urn scheme." *Lecture Notes-Monograph Series*, 245–267. MR1481784. doi: https://doi.org/10.1214/lnms/1215453576. 2

Pitman, J. (2006). *Combinatorial stochastic processes*. Springer. MR2245368. 2, 5, 8

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

Portenoy, J., Hullman, J., and West, J. D. (2017). "Leveraging citation networks to visualize scholarly influence over time." *Frontiers in Research Metrics and Analytics*, 2: 8. 4

Watterson, G. and Guess, H. A. (1977). "Is the most frequent allele the oldest?" *Theoretical population biology*, 11(2): 141–160. 2, 3, 10, 22

Zabell, S. L., Earman, J., and Norton, J. (1997). "The continuum of inductive methods revisited." *The cosmos of science, Pittsburgh-Konstanz Series in the philosophy and history of science*, 351–385. 5