Bayesian nonparametric inference for "species-sampling" problems

Submitted to Statistical Science

Bayesian nonparametric inference for "species-sampling" problems

Cecilia Balocchi, Stefano Favaro[1] and Zacharie Naulet

Abstract. Given an observed sample from a population of individuals belonging to species, "species-sampling" problems (SSPs) call for estimating some features of the unknown species composition of additional unobservable samples from the same population. Within SSPs, the problems of estimating coverage probabilities, the number of unseen species and coverages of prevalences have emerged in the past three decades for being the subject of numerous methodological and applied works, mostly in biological sciences but also in statistical machine learning, electrical engineering, theoretical computer science, information theory and forensic statistics. In this paper, we focus on these popular SSPs, and present an overview of their Bayesian nonparametric (BNP) analysis under the Pitman–Yor process (PYP) prior. While reviewing the literature, we improve on computation and interpretability of existing posterior inferences, typically expressed through complicated combinatorial numbers, by establishing novel posterior representations in terms of simple compound Binomial and Hypergeometric distributions. We also consider the problem of estimating the discount and scale parameters of the PYP prior, showing a property of Bayesian consistency with respect to estimation through the hierarchical Bayes and empirical Bayes approaches, that is: the discount parameter can be estimated consistently, whereas the scale parameter cannot be estimated consistently, thus advising caution in posterior inference. We conclude our work by discussing some generalizations of SSPs, mostly in the field of biological sciences, which deal with "feature-sampling", multiple populations of individuals sharing species and classes of Markov chains.

Key words and phrases: Bayesian nonparametrics, Bayesian consistency, coverage of prevalences, coverage probabilities, empirical Bayes, hierarchical Bayes, Pitman–Yor process prior, "species-sampling" problems, unseen species.

1. INTRODUCTION

The estimation of the number of unseen species is a classical problem in statistics, dating back to the seminal work of Fisher et al. [1943]. Consider a generic population of individuals, such that each individual takes a value in a (possibly infinite) space of species' labels or symbols. Assuming $n \geq 1$ observed individuals to be modeled as a random sample $(X_1, \ldots, X_n)$ from an unknown discrete distribution p, the unseen-species problem calls for estimating

$$\mathfrak{u}_{n,m} = |\{X_{n+1}, \ldots, X_{n+m}\} \setminus \{X_1, \ldots, X_n\}|,$$

namely the number of hitherto unseen (distinct) species that would be observed if $m \geq 1$ additional samples $(X_{n+1}, \ldots, X_{n+m})$ were collected from the same distribution. The unseen-species problem may be viewed as the m-step ahead generalization of the problem of estimating the missing mass, namely the probability of discovering at the $(n+1)$-th draw a species not observed in the sample $(X_1, \ldots, X_n)$ [Goodman, 1949; Good, 1953]. One may consider several refinements or generalizations

Cecilia Balocchi is Assistant Professor, School of Mathematics, University of Edinburgh, Edinburgh, United Kingdom (e-mail: cecilia.balocchi@ed.ac.uk). Stefano Favaro is Professor, Department of Economics and Statistics, University of Torino and Collegio Carlo Alberto, Torino, Italy (e-mail: stefano.favaro@unito.it). Zacharie Naulet is Assistant Professor, Department of Mathematics, Université Paris-Saclay, Orsay, France (e-mail: zacharie.naulet@universite-paris-saclay.fr).

[1]Also affiliated to IMATI-CNR "Enrico Magenes" (Milan, Italy).

of the unseen-species problem, by defining suitable discrete functionals of the X_i's according to the features of interest on the unknown species' composition of unobserved samples [Good and Toulmin, 1956; Efron and Thisted, 1976]. We refer to these problems as "species-sampling" problems (SSPs), though SSPs may also refer to a broader class of statistical problems that deal with sampling from generic populations of species. SSPs first appeared in ecology for the estimation of the species richness or diversity of ecological populations, and their importance has grown in the most recent years driven by applications in biological and physical sciences, statistical machine learning, electrical engineering, theoretical computer science, information theory and forensic statistics.

Biological sciences are the field where SSPs have been most investigated over the past three decades, raising several challenges in both methods and applications. This is testified by the work of Deng et al. [2019], which shows how large scale genomic data provide a fertile ground for SSPs. Although sequencing technologies have advanced the understanding of genome biology, observed samples may not be perfectly representative of the molecular heterogeneity or species composition of the underlying DNA library, often providing a poor representation due to low-abundance molecules that are hard to sample. Due to the impossibility of sequencing DNA libraries up to complete saturation, it is common to make use of the observed samples, typically collected under suitable budget constraints, to infer the molecular heterogeneity of additional unobserved samples from the library, as well as of the library itself. Deng et al. [2019] identified three major questions of interest:

- Q1) what is the expected population frequency of a species with frequency $r \geq 1$ in the sample?
- Q2) how many previously unobserved species in the sample will be observed in additional samples?
- Q3) how many species with frequency $r \geq 1$ in the sample will be observed in additional samples?

These are popular examples of SSPs, with Q2) being the unseen-species problem, and they all apply to statistical analysis related to the study of sequencing complexity [Daley and Smith, 2013], design of sequencing experiments [Sims et al., 2014] and estimation of genetic diversity [Gao et al., 2007], immune receptor diversity [Robins et al., 2009] and genetic variation [Ionita-Laza et al., 2009].

Although nonparametric estimation of the number of unseen-species dates back to the 1950s, only recent works have set forth a rigorous and comprehensive treatment of such a problem [Orlitsky et al., 2016; Wu and Yang, 2019; Polyanskiy and Wu, 2020]. Besides providing estimators of $\mathfrak{u}_{n,m}$ with provable (theoretical) guarantees, these works have introduced novel tools that pave the way to investigate other SSPs [Wu and Yang, 2021]. In general, nonparametric estimation of SSPs does not rely on any assumption on the underlying distribution p of the X_i's, with provable guarantees that hold uniformly over all discrete distributions. This assumption-free framework has led to develop solid theories in their greatest generality, though worst case distributions may severely hamper the empirical performance of the proposed estimators, leading to unreliable results in applications. It is therefore useful to take into account any knowledge about the nature of data, which typically results in placing assumptions on the tail behaviour of the distribution p. That is, the assumption-free framework of SSPs may be usefully complemented through suitable prior assumptions of regularity on the tail behaviour of p. A common and flexible tail assumption is that of regular variation, which allows for p to range from a geometric tail to an heavy power-law tail behaviour [Gnedin et al., 2007]. This is well-motivated by the ubiquitous power-law type distributions, which occur in many situations of scientific interest, and nowadays have significant consequences for the understanding of numerous natural and social phenomena [Clauset et al., 2009].

A Bayesian nonparametric (BNP) approach to SSPs has been set forth in Lijoi et al. [2007], and it relies on specifying a prior on the distribution p of the X_i's. This is arguably the most natural approach to complement the assumption-free framework of SSPs. In this respect, discrete random probability measures in the form of species sampling models [Pitman, 1996] provide a broad class of nonparametric priors for p [Pitman, 2006, Chapter 3 and Chapter 4]. Among species sampling models, Lijoi et al. [2007] focussed on the Pitman–Yor process (PYP) prior [Perman et al., 1992; Pitman and Yor, 1997], whose mathematical tractability and interpretability make it the natural candidate in applications [Favaro et al., 2009, 2012]. The PYP prior is indexed by a discount parameter $\alpha \in [0, 1)$ and a scale parameter $\theta > -\alpha$, such that for $\alpha = 0$ it reduces to the Dirichlet process (DP) of Ferguson [1973]. Of special interest is the parameter α, as it admits a clear interpretation in terms of controlling the tail behaviour of the PYP prior, which ranges from geometric tails to heavy power-law tails. In particular, the larger α the heavier the tail of the PYP prior, and as a limiting case for $\alpha \to 0$ one recovers the geometric tail behaviour of the DP prior [Pitman, 2006, Chapter 3]. Such a peculiar parameterization makes the PYP a flexible prior choice, which allows for tuning of the tail behaviour of the prior with respect to the empirical distribution of the data. Among species sampling models, the PYP prior is certainly unique with respect to mathematical tractability, flexibility and interpretability [De Blasi et al., 2015; Bacallado et al., 2017].

1.1 Our contributions

In this paper, we present an overview of BNP inference for SSPs under the PYP prior. We focus on SSPs corresponding to the aforementioned questions Q1, Q2 and Q3. As for Q1, we consider the problem of estimating coverage probabilities, which include the missing mass and the coverage probability of order $r \geq 1$, namely the probability mass of species observed with frequency r in the sample. As for Q2, we consider the unseen-species problem and a generalization of it defined in terms of the estimation of the number of hitherto unseen species that would be observed with frequency $r \geq 1$ in m additional samples, here referred to as the unseen species' prevalences of order r. Finally, as for Q3, we consider the problem of estimating the coverage of prevalence of order $r \geq 1$, namely the number of species with frequency r in the sample that would be observed in m additional samples. We introduce each SSP in the assumption-free framework, and then we present its analysis in the BNP framework under the PYP prior. While reviewing posterior analyses of SSPs from the recent BNP literature, we establish some novel representations of posterior distributions. In particular, we show that posterior distributions that are typically expressed in terms of complicated combinatorial numbers, which hamper both the computation and the interpretability of posterior inferences, admit straightforward representations in terms of compound Binomial and Hypergeometric distributions. This is a step forward in the BNP approach to SSPs, especially with respect to its use in problems of practical interest, contributing to simplify and make more interpretable the posterior inferences.

Critical in the BNP approach to SSPs under the PYP prior is the estimation of the prior's parameters (α, θ). Two approaches for estimating (α, θ) are: i) the hierarchical Bayes or fully Bayes approach, which relies on the posterior distribution of (α, θ) with respect to a suitable prior specification; ii) the empirical Bayes approach, which relies on maximizing, with respect to (α, θ), the (marginal) likelihood of the observed sample. The empirical Bayes approach is the most used in practice [Lijoi et al., 2007; Favaro et al., 2009; De Blasi et al., 2015], though no provable guarantees have been established for empirical Bayes estimates of (α, θ). The lack of a theoretical understanding of the hierarchical Bayes approach and the empirical Bayes approach has precluded clear guidelines for choosing among them. Here, we investigate their properties of Bayesian consistency. Under moderate misspecification, we find that the empirical Bayes estimator of (α, θ) converges to a limit that is interpretable in terms of the "true" data generating mechanism. In particular, when the model is correct, we find that: i) both the empirical Bayes estimator and the hierarchical Bayes estimator of α are consistent; ii) θ can not be tested or estimated consistently, because of a curious anti-concentration result. Under the hierarchical Bayes approach, we characterize the large sample asymptotic behavior of the posterior distribution of the parameter (α, θ); we find that the limiting posterior distribution of θ depends on the prior, thus particular caution should be used. As for α, we prove a weak form of the Bernstein-von Mises theorem, finding its contraction rates.

1.2 Organization of the paper

The paper is structured as follows. In Section 2 we introduce the BNP framework for SSPs under the PYP prior, and review the PYP, with emphasis on the interpretation of the prior's parameters (α, θ). BNP inference for SSPs is presented in Section 3 for the coverage probabilities, in Section 4 for the number of unseen species, and in Section 5 for the coverage of prevalences. In Section 6 we consider the estimation of (α, θ), showing properties of Bayesian consistency with respect to the hierarchical Bayes and empirical Bayes approaches. In Section 7 we present some numerical illustrations for the estimation of (α, θ) and for the estimation of the unseen, using synthetic data. Section 8 contains a discussion on some emerging generalizations of SSPs, mostly in the field of biological and physical sciences, which deal with "feature-sampling" problems, multiple populations of individuals sharing species and classes of Markov chains. Additional material, technical results, and proofs of the main results are deferred to the Supplementary Material [Balocchi et al., 2024+].

2. THE BNP SPECIES SAMPLING FRAMEWORK

The assumption-free framework for SSPs, henceforth referred to as the "classical framework", assumes that $n \geq 1$ observed samples from the population are modeled as a random sample $\mathbf{X}_n = (X_1, \ldots, X_n)$ from an unknown discrete distribution p, i.e. $p = \sum_{j \geq 1} p_j \delta_{s_j}$ with p_j being the probability of species' label s_j for $j \geq 1$, such that $\sum_{j \geq 1} p_j = 1$. Here, following Lijoi et al. [2007], we consider to endow p with the PYP prior, namely we assume that

$$
\begin{aligned}
(1) \qquad X_i \,|\, P \;&\overset{\text{iid}}{\sim}\; P \qquad i = 1, \ldots, n, \\
P \;&\sim\; \text{PYP}(\alpha, \theta),
\end{aligned}
$$

with $\text{PYP}(\alpha, \theta)$ being the law of the PYP process with parameter (α, θ). We refer to (1) as the "BNP framework" for SSPs. A simple and intuitive definition of the PYP follows from its stick-breaking construction [Pitman, 1995]. For $\alpha \in [0, 1)$ and $\theta > -\alpha$ let: i) $(V_i)_{i \geq 1}$ be independent r.v.s such that V_i has Beta distribution with parameter $(1 - \alpha, \theta + i\alpha)$, for $i \geq 1$; ii) let $(S_j)_{j \geq 1}$ be r.v.s, independent of the V_i's, and i.i.d. as a non-atomic distribution ν on a measurable space $\mathbb{S}$. If we set $P_1 = V_1$ and

$P_j = V_j \prod_{1\le i\le j-1}(1-V_i)$ for $j \ge 1$, such that $P_j \in (0,1)$ for any $j \ge 1$ and $\sum_{j\ge1} P_j = 1$ almost surely, then the random probability measure $P = \sum_{j\ge1} P_j \delta_{S_j}$ is a PYP on $\mathbb{S}$ with base distribution ν, discount α and scale θ. The PYP generalizes the DP by means of the parameter $\alpha \in [0,1)$, which controls the tail behaviour of P. For any $\alpha \in (0,1)$ let $P \sim \text{PYP}(\alpha,\theta)$ and let $(P_{(j)})_{j\ge1}$ be the decreasingly ordered P_j's. Then, as $j \to +\infty$ the $P_{(j)}$'s follow a power-law distribution of exponent $c = \alpha^{-1}$ [Pitman and Yor, 1997]; that is, $\alpha \in (0,1)$ controls the tail behaviour of the PYP through the small $P_{(j)}$'s: the larger α the heavier the tail of P. As a limiting case for $\alpha \to 0$, the DP features geometric tails [Pitman, 2006, Chapter 3 and Chapter 4].

2.1 Sampling formulae for the PYP prior

The random sample $\mathbf{X}_n$ from $P \sim \text{PYP}(\alpha,\theta)$ is regarded as part of the exchangeable sequence $(X_n)_{n\ge1}$ satisfying (1), that is a sequence whose directing (de Finetti) measure is the law of the PYP. Because of the almost sure discreteness of P, $\mathbf{X}_n$ features $K_n \le n$ distinct species, labelled by $\{S^*_1,\dots,S^*_{K_n}\}$, with frequencies $\mathbf{N}_n = (N_{1,n},\dots,N_{K_n,n})$ such that $N_i \ge 1$, for $i = 1,\dots,k$, and $\sum_{1\le i\le k} N_i = n$. In other terms, $\mathbf{X}_n$ induces a random partition Π_n of $\{1,\dots,n\}$ whose blocks are the classes induced by the equivalence relation $i \sim j \iff X_i = X_j$. In particular, Π_n is an exchangeable random partition, meaning that its distribution is such that the probability of any partition of $\{1,\dots,n\}$ into k blocks with frequency $\mathbf{n} = (n_1,\dots,n_k)$ is a symmetric function $p_{n,k}$ of $\mathbf{n}$, i.e.

$$p_{n,k}(\mathbf{n}) = \frac{\left(\frac{\theta}{\alpha}\right)_{(k)}}{(\theta)_{(n)}} \prod_{i=1}^{k} \alpha(1-\alpha)_{(n_i-1)}, \tag{2}$$

where $(a)_{(u)}$ is the u-th rising factorial of a, i.e. $(a)_{(u)} = \prod_{0\le i\le u-1}(a+i)$, for $a \ge 0$ and $u \in \mathbb{N}_0$. The function $p_{n,k}$ is referred to as the exchangeable partition probability function of Π_n [Kingman, 1978; Pitman, 1995]. The sequence $(\Pi_n)_{n\ge1}$ defines an exchangeable random partition Π of $\mathbb{N}$, with exchangeability meaning that the distribution of Π is invariant under finite permutations of its elements, provided that Π_m is the restriction of Π_n to the first m elements, almost surely for all $m < n$. This implies that

$$p_{n,k}(\mathbf{n}) = p_{n+1,k+1}(\mathbf{n},1) + \sum_{i=1}^{k} p_{n+1,k}(\mathbf{n}+\mathbf{1}_i), \tag{3}$$

for all $n \ge 1$, with $\mathbf{1}_i$ being a vector of length k with 1 in the position i and 0 elsewhere [Hansen and Pitman, 2000; Nacu, 2006]. See Pitman [2006, Chapter 2] for details on exchangeable random partitions and generalizations thereof.

The construction of Π, and in particular (3), implies that its distribution is completely determined by the distribution of $(X_i)_{i\ge1}$ through the conditional probability of X_{n+1} given $\mathbf{X}_n$, i.e. the predictive probability. If $\mathbf{X}_n$ is a random sample from $P \sim \text{PYP}(\alpha,\theta)$ as described above, and $\Pr[X_1 \in \cdot] = \nu(\cdot)$, then the predictive probabilities of $(X_i)_{i\ge1}$ are

$$\Pr[X_{n+1} \in \cdot \mid \mathbf{X}_n] = \frac{\theta + k\alpha}{\theta+n}\nu(\cdot) + \sum_{i=1}^{k} \frac{n_i - \alpha}{\theta+n}\delta_{S^*_i}(\cdot), \tag{4}$$

determining the distribution of Π_{n+1} from that of Π_n, for $n \ge 1$. The probability (4) is a linear combination between: i) the probability $(\theta+k\alpha)/(\theta+n)$ that X_{n+1} is a new species, i.e. the probability of creating a new block in the random partition of $\{1,\dots,n\}$; ii) the probability $(n_i-\alpha)/(\theta+n)$ that X_{n+1} takes value S^*_i, i.e. the probability of increasing by 1 the size of the block S^*_i in the random partition of $\{1,\dots,n\}$, for $i = 1,\dots,k$. For $\theta > 0$, an intuitive description of (4) was proposed in Zabell [2005, Chapter 11]. Consider an urn containing a black ball and colored (non-black) balls, where colored balls may be interpreted as individuals with their associated species (colors). Balls are drawn from the urn, and then returned to the urn, in such a way that the probability of a ball being drawn at any stage is proportional to its weight. Initially the urn contains a black ball with weight $\theta > 0$, and at the n-th draw: i) if we pick a colored ball then the ball is returned to the urn together with a ball of the same color with weight 1; ii) if we pick a black ball, then the weight of the black ball is increased by $\alpha \in [0,1)$ and a ball of a new color with weight $1-\alpha$, i.e. any color not present in the urn, is inserted in the urn. If X_n is the color of the ball returned in the urn after the n-th draw, and such a color is generated as ν, then the conditional distribution of X_{n+1} given $\mathbf{X}_n$ is (4).

Zabell [1992, 1997] provided a characterization of the PYP through its predictive probability (4), referred to as "sufficientness postulate" [Johnson, 1932; Bacallado et al., 2017]. For $\alpha \in (0,1)$ the PYP is characterized as the sole species sampling model whose predictive probability is such that: i) the probability that X_{n+1} is of a new species depends on $\mathbf{X}_n$ only through n and K_n; ii) the probability that X_{n+1} belongs to S^*_i depends on $\mathbf{X}_n$ only through n and $N_{n,i}$. In particular, the DP is the sole species sampling model whose predictive probability is such that the probability that X_{n+1} belongs is of a new species depends on $\mathbf{X}_n$ only through n [Regazzini, 1978; Lo, 1991]. The "sufficientness postulate" and the Pólya-like urn scheme show how the parameter α drives a combined effect in terms of a reinforcement mechanism and the increase in the rate at which new species are generated according to (4). A new species S^* entering in the sample produces two effects: i) it is assigned a mass proportional

to $(1-\alpha)$ in the S^* empirical component of (4); ii) it is assigned a mass proportional to α in the probability of generating a new species. That is, the mass assigned to S^* is less than proportional to its cluster size, i.e. 1, and the remaining mass is added to the probability of generating new species. The first effect gives rise to the following reinforcement mechanism: if S^* is re-observed then the mass of S^* is increased by $1/(\theta+n+1)$, meaning that the sampling procedure tends to reinforce observed species with higher frequencies. The second effect implies that the probability of generating yet another new species, which overall still decreases in n, is increased by $\alpha/(\theta+n+1)$. The larger α the stronger the reinforcement mechanism, and the higher is the probability of generating new species. For $\alpha=0$ everything is proportional to species' frequencies, in such a way that the number of distinct species does not alter the probability of generating new species.

We conclude by recalling the sampling formula or frequency-of-frequency distribution induced by a random sample $\mathbf{X}_n$ from $P\sim\text{PYP}(\alpha,\theta)$, which follows from (2) [Pitman, 2006, Chapter 2]. Let $M_{r,n}$ be the number of distinct species with frequency r in $\mathbf{X}_n$, for $1\leq r\leq n$, i.e. $M_{r,n}=\sum_{1\leq i\leq K_n} I(N_{i,n}=r)$, such that $\sum_{1\leq r\leq n} M_{r,n}=K_n$ and $\sum_{1\leq r\leq n} rM_{r,n}=n$. The distribution of $\mathbf{M}_n=(M_{1,n},\ldots,M_{n,n})$ is defined on $\mathcal{M}_{n,k}=\{k\in\{1,\ldots,n\}$ and $(m_1,\ldots,m_n): m_i\geq 0, \sum_{1\leq i\leq n} m_i=k, \sum_{1\leq i\leq n} im_i=n\}$, and referred to as Ewens-Pitman sampling formula (EPSF). For $\mathbf{m}=(m_1,\ldots,m_n)\in\mathcal{M}_{n,k}$

(5)
$$\Pr[\mathbf{M}_n=\mathbf{m}]=n!\frac{\left(\frac{\theta}{\alpha}\right)_{(\sum_{i=1}^n m_i)}}{(\theta)_{(n)}}\prod_{i=1}^n\frac{\left(\frac{\alpha(1-\alpha)_{(i-1)}}{i!}\right)^{m_i}}{m_i!}.$$

We refer to Section S2 of the Supplementary Material [Balocchi et al., 2024+] for a representation of the EPSF in terms of a compound Poisson sampling model [Charalambides, 2005, Chapter 7]. The distribution of K_n follows from (5) [Pitman, 2006, Chapter 3]. In particular, for $a>0$, $b\geq 0$ and $u,v\in\mathbb{N}_0$ with $v\leq u$, let $\mathscr{C}(u,v;a,b)$ be the (u,v)-th non-centered generalized factorial coefficient, i.e., $\mathscr{C}(u,v;a,b)=(v!)^{-1}\sum_{0\leq i\leq v}(-1)^i\binom{v}{i}(-ia-b)_{(u)}$; see Section S1 of the Supplementary Material [Balocchi et al., 2024+]. Then, for $x\in\{1,\ldots,n\}$ it holds that

(6)
$$\Pr[K_n=x]=\frac{\left(\frac{\theta}{\alpha}\right)_{(x)}}{(\theta)_{(n)}}\mathscr{C}(n,x;\alpha,0).$$

For $\alpha=0$, i.e. under the DP, the distribution of K_n follows directly from (6) by letting $\alpha\to 0$. The resulting distribution is expressed in terms of the (u,v)-th signless Stirling number $|s(u,v)|$, which arises by means of $|s(u,v)|=\lim_{a\to 0} a^{-v}\mathscr{C}(u,v;a,0)$; see Section S1 of the Supplementary Material [Balocchi et al., 2024+] for details.

At the sampling level, the power-law tail behaviour of the PYP emerges naturally from the analysis of the large n asymptotic behaviour of K_n and $M_{r,n}$. For $\alpha\in(0,1)$ we denote by f_α the density function of a positive α-stable r.v., and for any $\theta>-\alpha$ let $S_{\alpha,\theta}$ be a r.v. with density function

(7)
$$f_{S_{\alpha,\theta}}(s)\propto s^{\frac{\theta-1}{\alpha}-1}f_\alpha(s^{-1/\alpha}),$$

that is a generalized Mittag-Leffler density function, such that $\mathbb{E}[S^r_{\alpha,\theta}]=(\theta/\alpha)_{(r)}\Gamma(\theta)/\Gamma(\theta+r\alpha)$ for $r\geq 1$, from which the mean and the variance of $S_{\alpha,\theta}$ follows immediately. Pitman [2006, Theorem 3.8] shows that, as $n\to+\infty$,

(8)
$$n^{-\alpha}K_n\to S_{\alpha,\theta}\quad\text{almost surely},$$

and

(9)
$$n^{-\alpha}M_{r,n}\to\frac{\alpha(1-\alpha)_{(r-1)}}{r!}S_{\alpha,\theta}\quad\text{almost surely}.$$

The r.v. $S_{\alpha,\theta}$ is positive and finite (almost surely), and it is typically referred to as Pitman's α-diversity [Pitman, 2003; Dolera and Favaro, 2020a,b; Bercu and Favaro, 2024]. For $\alpha=0$, we recall that as $n\to+\infty$, $K(n)/\log n\to\theta$ almost surely [Korwar and Hollander, 1973] and $M_{r,n}\to P_{\theta/r}$ almost surely [Ewens, 1972], where $P_{\theta/r}$ is a Poisson r.v. with parameter θ/r. Equation (8) shows that K_n, for large n, grows as n^α. This is precisely the growth of the number of distinct species in $n\geq 1$ random samples from a power-law distribution of exponent $c=\alpha^{-1}$. Moreover, by combining (8) and (9), it holds that $p_{\alpha,r}=\alpha(1-\alpha)_{(r-1)}/r!$ is the large n asymptotic proportion of the number of distinct species with frequency r. Therefore, $p_{\alpha,r}\simeq c_\alpha r^{-\alpha-1}$ for large r, for a constant c_α. This is precisely the distribution of the number of distinct species with frequency r in $n\geq 1$ random samples from a power-law distribution of exponent $c=\alpha^{-1}$.

3. COVERAGE PROBABILITIES

The estimation of coverage probabilities, or rare probabilities, dates back to the early work of Alan M. Turing and Irving J. Good at Bletchley Park in 1940s [Good, 1953]. Let $\mathbf{X}_n$ be a random sample under the "classical framework" for SSPs, and denote by $(N_{j,n})_{j\geq 1}$ the species' frequencies in $\mathbf{X}_n$. The coverage probability of order $r\geq 0$ is

$$\mathfrak{p}_{r,n}=\sum_{j\geq 1}p_j I(N_{j,n}=r),$$

namely the probability mass of species observed with frequency $r\geq 0$ in the sample. Of special interest is $\mathfrak{p}_{0,n}$,

namely the coverage probability of order 0, which is referred to as the missing mass. The problem of estimating $\mathfrak{p}_{r,n}$ first appeared in ecology [Fisher et al., 1943; Good, 1953; Chao and Lee, 1992; Bunge and Fitzpatrick, 1993], and over the past three decades its importance has grown in biological sciences [Kroes et al., 1999; Mao, 2004; Gao et al., 2007]. Coverage probabilities arise in DNA sequencing data in the form of sample coverage, or saturation, and frequency estimation. Sample coverage, i.e. the proportion of molecules in an (infinite) population that are observed in the sample, is related to the estimation of the population abundance of unobserved molecules (the missing mass), as it is equal to $1-\mathfrak{p}_{0,n}$. An accurate estimation of the sample coverage allows to determine if the sample is saturated, i.e. all species have been sampled. Sequencing experiments with high sample coverage are crucial to avoid sampling bias and to produce robust findings. Given that high sample coverage is often an issue in degraded DNA samples, in metagenomics, such as the study of the microbiome, and in single-cell DNA sequencing, where sample coverage varies in each cell, a correct sample coverage estimation remains critical [Deng et al., 2019, Section 4].

REMARK 1. *Besides biological and physical sciences, the problem of estimating coverage probabilities, and in particular the estimation of the missing mass, has found application in many scientific fields: i) statistical machine learning, e.g., estimation of node degrees of networks based on source-destination data [Zhang, 2005], optimal discovery with probabilistic expert advice [Bubeck et al., 2013] and frequency recovery from sketched data through random hashing [Cai et al., 2018]; ii) theoretical computer science, in the context of recovering from sketches the number of distinct species through probabilistic counting algorithms [Motwani and Vassilvitskii, 2006]; iii) information theory, in the context of universal (lossless) compression of sequences over arbitrarily large alphabets [Orlitsky et al., 2004; Ben-Hamou et al., 2018]; iv) empirical linguistics, e.g., estimation of the size of a vocabulary [Gale and Sampson, 1995] and m-gram language modeling in natural language processing [Ohannessian and Dahleh, 2012]; v) forensic DNA analysis, in the context of rare-type matching problem [Anevski et al., 2017; Cereda, 2017; Favaro and Naulet, 2024].*

3.1 A nonparametric estimator of $\mathfrak{p}_{r,n}$

Under the "classical framework", the Good-Turing estimator is the most popular estimator of $\mathfrak{p}_{r,n}$ [Good, 1953; Good and Toulmin, 1956; Robbins, 1956, 1968]. Denoting by $M_{r,n}=m_r$ the number of distinct species with frequency $r\geq 1$ in $\mathbf{X}_n$, the Good-Turing estimator is given by

$$\tilde{\mathfrak{p}}_{r,n}=(r+1)\frac{m_{r+1}}{n}. \tag{10}$$

The Good-Turing estimator is a nonparametric estimator of $\mathfrak{p}_{r,n}$, in the sense that it does not rely on any distributional assumption on the unknown p. It has a straightforward heuristic derivation, which relies on a comparison between the expectations of $\mathfrak{p}_{r,n}$ and $M_{r,n}$, for $r\geq 0$, that is

$$\mathbb{E}[\mathfrak{p}_{r,n}]=\frac{r+1}{n+1}\mathbb{E}[M_{r+1,n+1}],$$

and then the approximation of $(n+1)^{-1}\mathbb{E}[M_{r+1,n+1}]$ with the observable $n^{-1}m_r$, assuming n large enough [Good, 1953; Robbins, 1968]. The estimator $\tilde{\mathfrak{p}}_{r,n}$ also admits a nonparametric empirical Bayes interpretation [Robbins, 1956, 1964], that is $\tilde{\mathfrak{p}}_{r,n}$ may be viewed as a posterior expectation with respect to a nonparametric prior estimated from the sample [Efron and Thisted, 1976; Efron, 2003]. The Good-Turing estimator has been the subject of numerous studies, e.g., central limit theorems, local limit theorems and large deviation principles [Esty, 1982, 1983; Zhang and Zhang, 2009; Gao, 2013; Grabchak and Zhang, 2017], admissibility and concentration properties [Robbins, 1968; McAllester and Schapire, 2000; Ohannessian and Dahleh, 2012; Ben-Hamou et al., 2017; Skorski, 2020], consistency and convergence rates [McAllester and Ortiz, 2003; Mossel and Ohannessian, 2019; Ayed et al., 2018], and optimality through minimax lower bounds [Orlitsky et al., 2003; Acharya et al., 2018; Ayed et al., 2018].

3.2 BNP inference for coverage probabilities

In the "BNP framework" (1), Arbel et al. [2017] computed the posterior distribution of $\mathfrak{p}_{r,n}$, given $\mathbf{X}_n$. In particular, assume that the random sample $\mathbf{X}_n$ features $K_n=k$ distinct species with frequencies $\mathbf{N}_n=(n_1,\ldots,n_k)$, such that $M_{r,n}=m_r$ for $r\geq 1$, and let $\text{Beta}(a,b)$ be the Beta distribution with parameter (a,b), for $a,b>0$. Then,

$$\mathfrak{p}_{0,n}\,|\,\mathbf{X}_n\sim\text{Beta}(\theta+\alpha k,\, n-\alpha k), \tag{11}$$

and for $r\geq 1$

$$\mathfrak{p}_{r,n}\,|\,\mathbf{X}_n\sim\text{Beta}((r-\alpha)m_r,\, \theta+n-(r-\alpha)m_r). \tag{12}$$

According to (11) and (12), for fixed $\alpha\in(0,1)$ and $\theta>-\alpha$, $K_n=k$ is a sufficient statistic to make inference on $\mathfrak{p}_{0,n}$, whereas $M_{r,n}=m_r$ is a sufficient statistic to make inference on $\mathfrak{p}_{r,n}$. Differently, if $\alpha=0$ then n and $M_{r,n}$ are sufficient statistics to infer $\mathfrak{p}_{0,n}$ and $\mathfrak{p}_{r,n}$, respectively. Besides leading to BNP estimates of $\mathfrak{p}_{n,r}$, (11) and (12) are critical to quantify uncertainty of estimates by means of credible intervals by means of, e.g., concentration inequalities [Skorski, 2021] and Monte Carlo sampling Arbel et al. [2017]. This is possible in practice because of the simple form of the posterior distribution, which allows to exploit well known properties of the Beta distribution to

deal with BNP inference for coverage probabilities under the PYP prior.

Under the assumption of a squared loss function, BNP estimators of $\mathfrak{p}_{0,n}$ and $\mathfrak{p}_{r,n}$ follows directly from (11) and (12), respectively, by taking corresponding expected values. That is,

$$\hat{\mathfrak{p}}_{0,n} = \mathbb{E}[\mathfrak{p}_{0,n} \,|\, \mathbf{X}_n] = \frac{\theta + k\alpha}{\theta + n}, \tag{13}$$

and for $r \geq 1$

$$\hat{\mathfrak{p}}_{r,n} = \mathbb{E}[\mathfrak{p}_{r,n} \,|\, \mathbf{X}_n] = (r - \alpha)\frac{m_r}{\theta + n}. \tag{14}$$

The BNP estimators (13) and (14) first appeared in Lijoi et al. [2007] and Favaro et al. [2012], where they are obtained as means of a direct application of the predictive probability (4) of the PYP prior. See also De Blasi et al. [2015]. In particular, by combining the definition of $\mathfrak{p}_{0,n}$ and predictive probabilities, the BNP estimator of $\mathfrak{p}_{0,n}$ under a squared loss function is precisely the probability that the $(n+1)$-th draw belongs to a new species, i.e. a species not observed in the sample; this is probability $(\theta + k\alpha)/(\theta + n)$ attached to ν in (4). Similarly, by combining the definitions of $\mathfrak{p}_{r,n}$ and predictive probabilities, the BNP estimator of $\mathfrak{p}_{r,n}$ under a squared loss function is precisely the probability that the $(n+1) - th$ draw belongs to a species observed with frequency r in the sample; this is the probability $(r - \alpha)/(\theta + n)$ attached to the empirical part of (4), i.e. the probability of observing a specific species with frequency r in the sample, multiplied by the number m_r of species with frequency r in the sample.

A peculiar feature of the Good-Turing estimator (10), which arises from its heuristic derivation, is that it depends on m_{r+1}, and not on m_r as one would intuitively expect for an estimator of $\mathfrak{p}_{r,n}$. This is contrast with the BNP estimator. Such a feature, combined with the irregular behaviour of the m_r's, may lead to absurd estimates, the most common being $\tilde{\mathfrak{p}}_{r,n} = 0$ when $m_r > 0$ and $m_{r+1} = 0$. To overcome this drawback, Good [1953] suggested to smooth the estimator (10) by replacing the m_r's with more regular m'_r's, with m'_r being, e.g., a suitable parabolic function of r, a proportion of the number k of distinct species in the sample, the expectation with respect to a suitable parametric model. The BNP estimator $\hat{\mathfrak{p}}_{r,n}$ may be interpreted as a smoothed Good-Turing estimator, where the smoothing is induced by the PYP prior. In particular, let $a_n \simeq b_n$ mean that $\lim_{n\to+\infty} a_n/b_n = 1$, namely a_n and b_n are asymptotically equivalent as n tends to infinity. Favaro et al. [2016, Theorem 1] show that, as $n \to +\infty$, for $r \geq 0$

$$\hat{\mathfrak{p}}_{r,n} \simeq (r+1)\frac{m'_{r+1}}{n}, \tag{15}$$

where

$$m'_{r+1} = \frac{\alpha(1-\alpha)_{(r)}}{(r+1)!}k. \tag{16}$$

According to (15), the BNP estimator $\hat{\mathfrak{p}}_{r,n}$ is asymptotically equivalent, for large n, to a smoothed Good-Turing estimator, where the smoothed m'_r in (16) is the proportion $\alpha(1-\alpha)_{(r)}/(r+1)!$ of the number k of species in the sample. While smoothing techniques for the Good-Turing estimator were introduced as an ad-hoc tool for post-processing the m_r's in order to improve the performance of $\tilde{\mathfrak{p}}_{r,n}$, Favaro et al. [2016] show that smoothing emerges naturally from a BNP approach to estimate $\mathfrak{p}_{r,n}$. We refer to Arbel et al. [2017] for high-order refinements of (15).

4. THE NUMBER OF UNSEEN SPECIES

The unseen-species problem is an m-step ahead generalization the problem of estimating the missing mass. For $m \geq 1$ let $\mathbf{X}_{n+m} = (X_1, \ldots, X_n, X_{n+1}, \ldots, X_{n+m})$ be a random sample under the "classical framework" for SSPs, of which only the first n elements are assumed to be observed, and denote by $(N_{j,n})_{j\geq 1}$ and $(N_{j,m})_{j\geq 1}$ the species' frequencies in $\mathbf{X}_n$ and $(X_{n+1}, \ldots, X_{n+m})$, respectively. The number of unseen (distinct) species is defined as

$$\mathfrak{u}_{n,m} = \sum_{j\geq 1} I(N_{j,n} = 0)I(N_{j,m} > 0),$$

namely the number of hitherto unseen (distinct) species that would be observed if m additional samples were collected from the same population. As the missing mass, the unseen-species problem first appeared in ecology [Fisher et al., 1943; Good and Toulmin, 1956; Chao, 1984; Chao and Lee, 1992], and over the past three decades its importance has grown in biological and physical sciences. In molecular biological data, the unseen-species problem arises in the estimation of the complexity of sequencing libraries [Daley and Smith, 2013; Ionita-Laza et al., 2009]. While in low-complexity libraries a large proportion of the sample is composed by only a small number of unique molecules, high-complexity libraries usually display a large number of molecules, providing more information for a fixed level of sequencing. Hence, high-complexity libraries are often preferred by researchers. To evaluate library complexity, the complexity curve, or Species Accumulation Curve (SAC), is defined as the number of additional species that are observed as the sampling effort increases. This is the number of unseen species, interpreted as a function of m. The problem of library complexity estimation plays a crucial role to predict the benefit of additional sequencing, and optimize resource in the planning stages of experiments [Deng et al., 2019, Section 3].

REMARK 2. *As a generalization of the problem of estimating the missing mass, the unseen-species problem has appeared in some of the fields described in Remark 1,*

with interest in statistical machine learning and theoretical computer science [Haas et al., 1995; Florencio and Herley, 2007; Hao and Orlitsky, 2020], and in empirical linguistics and natural language processing [Thisted and Efron, 1987; Gale and Sampson, 1995; Ohannessian and Dahleh, 2012]. In principle, we may say that all the statistical problems described in Remark 1 admit a natural extension in which $m > 1$ additional unobserved sample are considered.

4.1 A nonparametric estimator of $\mathfrak{u}_{n,m}$

Under the "classical framework", the Good-Toulmin estimator is the most popular estimator of $\mathfrak{u}_{n,m}$ [Good and Toulmin, 1956; Efron and Thisted, 1976]. If $\lambda = m/n$ and $M_{r,n} = m_r$ denotes the number of distinct species with frequency $r \geq 1$ in $\mathbf{X}_n$, then the Good-Toulmin estimator is

$$\tilde{\mathfrak{u}}_{n,m} = \sum_{i \geq 1} (-1)^{i+1} \lambda^i m_i. \tag{17}$$

For $\lambda = n^{-1}$ the estimator (17) reduces to the Good-Turing estimator of $\mathfrak{p}_{0,n}$. The Good-Toulmin estimator (17) is a nonparametric estimator of $\mathfrak{u}_{n,m}$, as it does not rely on any distributional assumption. The estimator $\tilde{\mathfrak{u}}_{n,m}$ was first obtained by Good and Toulmin [1956] through a comparison between the expectations of $\mathfrak{u}_{n,m}$ and $M_{r,n}$, in analogy with the Good-Turing estimator in Good [1953], whereas Efron and Thisted [1976] proved that $\tilde{\mathfrak{u}}_{n,m}$ admits a nonparametric empirical Bayes derivation [Mao and Lindsay, 2002]. In particular, Efron and Thisted [1976] observed that, for $\lambda \geq 1$, the geometrically increasing magnitude of λ^i produces some wild oscillations in (17), which are undesirable. To overcome this drawback, they proposed a modification of $\tilde{\mathfrak{u}}_{n,m}$ that relies on a Euler-type random truncation of the series (17). Motivated by the increasing interest in the range $\lambda > 1$, especially in biological applications, Efron-Thisted estimator has been the subject of breakthrough studies [Orlitsky et al., 2016; Wu and Yang, 2019; Polyanskiy and Wu, 2020]. Orlitsky et al. [2016] showed that Efron-Thisted estimator provably estimates $\mathfrak{u}_{n,m}$ all of the way up to $\lambda \asymp \log n$, that such a range is the best possible, and that the estimator's mean-square error is minimax near-optimal for any λ. These theoretical guarantees do not rely on any assumption on the underlying unknown distribution p for the X_i's, and they hold uniformly over all discrete distributions, thus providing a theory in its greatest generality.

A generalization, or refinement, of the number of unseen species $\mathfrak{u}_{n,m}$ is the unseen species' prevalences. Formally, the unseen species' prevalence of order $r \geq 1$ is defined as

$$\mathfrak{u}_{r,n,m} = \sum_{j \geq 1} I(N_{j,n} = 0) I(N_{j,m} = r).$$

That is, $\mathfrak{u}_{r,n,m}$ is the number of hitherto unseen species that would be observed with frequency r if m additional samples were collected from the same population. For small values of r, the unseen species' prevalence of order r is also referred to as the number of unseen rare species. This is a critical quantity for the understanding of the species composition of the unobserved individuals. In ecology and biology, for instance, conservation of biodiversity requires a careful control of the number of species with frequency less than a certain threshold, namely rare species [Magurran, 2003; Thompson, 2004]. In genomics, rare species represent a critical issue, the reason being that species that appear only once or twice are often associated with deleterious diseases [Laird and Lange, 2010]. For $\lambda < 1$, a nonparametric estimator of $\mathfrak{u}_{r,n,m}$ may be obtained by comparing expectations of $\mathfrak{u}_{r,n,m}$ and $M_{r,n}$, which leads to

$$\tilde{\mathfrak{u}}_{r,n,m} = \sum_{i \geq 1} (-\lambda)^{i+r-1} \binom{r+i-1}{i-1} m_{i+r-1}. \tag{18}$$

For $\lambda \geq 1$, Hao and Li [2020] proposed a modification of $\tilde{\mathfrak{u}}_{r,n,m}$ in the spirit of Efron-Thisted estimator. In particular, along the same lines of te original work of Orlitsky et al. [2016], Hao and Li [2020] show that their estimator provably estimates $\mathfrak{u}_{r,n,m}$ all of the way up to $\lambda \asymp r^{-1} \log n$, that such a range is the best possible, and that the estimator's mean-square error is minimax near-optimal.

4.2 BNP inference for the number of unseen species

In the "BNP framework" (1), Lijoi et al. [2007] computed the posterior distribution of $\mathfrak{u}_{n,m}$ given $\mathbf{X}_n$. In particular, assume that the random sample $\mathbf{X}_n$ features $K_n = k$ distinct species with frequencies $\mathbf{N}_n = (n_1, \ldots, n_k)$, such that $M_{r,n} = m_r$ for $r \geq 1$. If $\alpha \in (0,1)$ then for $x \in \{0, 1, \ldots, m\}$

$$\Pr[\mathfrak{u}_{n,m} = x \,|\, \mathbf{X}_n] = \frac{\left(k + \frac{\theta}{\alpha}\right)_{(x)}}{(\theta + n)_{(m)}} \mathscr{C}(m, x; \alpha, -n + k\alpha). \tag{19}$$

Under the assumption of a squared loss function, a BNP estimator of $\mathfrak{u}_{n,m}$ [Favaro et al., 2009] is the expected value of (19), i.e.,

$$\hat{\mathfrak{u}}_{n,m} = \mathbb{E}[\mathfrak{u}_{n,m} \,|\, \mathbf{X}_n] = \left(k + \frac{\theta}{\alpha}\right) \left(\frac{(\theta + n + \alpha)_{(m)}}{(\theta + n)_{(m)}} - 1 \right). \tag{20}$$

For $\alpha = 0$, i.e. under the DP, the posterior distribution and the BNP estimator of $\mathfrak{u}_{n,m}$ are obtained from (19) and (20), respectively, by letting $\alpha \to 0$ [Lijoi et al., 2007]. In particular, for $b \geq 0$ let $|s(u, v; b)|$ be the non-centered signless Stirling number of the first type, which arises from $\mathscr{C}(u, v; a, b)$ by means of

$|s(u,v;b)| = \lim_{a\to 0} a^{-v}\mathscr{C}(u,v;a,b)$. If $\alpha = 0$, then for $x \in \{0,1,\ldots,m\}$

$$\Pr[\mathfrak{u}_{n,m} = x \,|\, \mathbf{X}_n] = \frac{\theta^k}{(\theta+n)_{(m)}}|s(m,x;n)| \tag{21}$$

and

$$\hat{\mathfrak{u}}_{n,m} = \mathbb{E}[\mathfrak{u}_{n,m} \,|\, \mathbf{X}_n] = \sum_{i=1}^{m} \frac{\theta}{\theta+n+i-1}. \tag{22}$$

For $m = 1$ the estimator (20) reduces to the estimator (13) of $\mathfrak{p}_{0,n}$. As a generalization of the estimators (20) and (22), Favaro et al. [2013] introduced a BNP estimator of the unseen species' prevalence $\mathfrak{u}_{r,n,m}$ [De Blasi et al., 2015]. However, to date no closed-form expressions are available for the posterior distribution of $\mathfrak{u}_{r,n,m}$, given the sample $\mathbf{X}_n$.

According to (19), for any fixed $\alpha \in (0,1)$ and $\theta > -\alpha$, $K_n = k$ is a sufficient statistic for estimating $\mathfrak{u}_{n,m}$. If $\alpha = 0$, then the sample size n is a sufficient statistic to infer $\mathfrak{u}_{n,m}$. The posterior distributions (19) and (21) are critical to quantify uncertainty of (20) and (22) by means of credible intervals. In practice, Monte Carlo sampling of (19) and (21) is doable for small values of n and m, and it becomes impossible for large values of n or m. This is because for large n and m, and even only for large m, the computational burden for evaluating generalized factorial coefficients and Stirling numbers becomes overwhelming. To overcome this drawback, Favaro et al. [2009] proposed large m approximations of the posterior distributions (19) and (21). Let $\mathfrak{u}_{n,m}(k)$ denote a r.v. whose distribution is (19) for $\alpha \in (0,1)$ and (21) for $\alpha = 0$. For $\alpha \in (0,1)$, Favaro et al. [2009, Proposition 2] show that, as $m \to +\infty$,

$$\frac{\mathfrak{u}_{n,m}(k)}{m^{\alpha}} \to B_{\frac{\theta}{\alpha}+k,\,\frac{n}{\alpha}-k} S_{\alpha,\theta+n} \quad \text{almost surely}, \tag{23}$$

where $B_{\theta/\alpha+k,n/\alpha-k}$ is a Beta r.v. with parameter $(\theta/\alpha+k, n/\alpha-k)$ and $S_{\alpha,\theta+n}$ is the Pitman's α-diversity with parameter $(\alpha,\theta+n)$, with $B_{\theta/\alpha+k,n/\alpha-k}$ being independent of $S_{\alpha,\theta+n}$ [Dolera and Favaro, 2020a]. For $\alpha = 0$, as $m \to +\infty$

$$\frac{\mathfrak{u}_{n,m}(k)}{\log(m)} \to (\theta+n) \quad \text{almost surely}. \tag{24}$$

The large m asymptotics (23) of $\mathfrak{u}_{n,m}(k)$ may be viewed as a posterior counterpart of the large n asymptotic behaviour of K_n in (8); the limiting r.v. (23) is referred to as posterior Pitman's α-diversity. See Favaro et al. [2013] for a generalization of (23) to the unseen species' prevalences $\mathfrak{u}_{r,n,m}$.

Favaro et al. [2009] introduced a Monte Carlo scheme for sampling the posterior α-diversity, and applied it to obtain a large m approximation of credible intervals for the BNP estimate (20). The critical step consists in sampling the Pitman's α-diversity $S_{\alpha,\theta+n}$, whose density function is (7) with $\theta+n$ in place of θ. This problem boils down to sample from a polynomially tilted positive α-stable distribution. That is, if f_α denotes the density function of a positive α-stable r.v., then the distribution of $S_{\alpha,\theta+n}^{-1/\alpha}$ has the polynomially tilted positive α-stable density function

$$f_{S^{-1/\alpha}_{\alpha,\theta+n}}(s) \propto s^{-(\theta+n)} f_\alpha(s).$$

The Monte Carlo approach of Favaro et al. [2009] is suitable for scenarios where n is not large, and m is much more large than n. This is because: i) for large n the computational burden for the sampling from a polynomially tilted positive α-stable distribution becomes overwhelming [Devroye, 2009; Hofert, 2011]; ii) the large m asymptotics (23) is of a qualitative nature, in the sense that it does not quantify the error in approximating the posterior distribution (19) with the distribution of the limiting r.v. (23). In the next proposition, we introduce a representation of the posterior distributions (19) and (21). Besides providing an intuitive interpretation of (19) in terms of compound Binomial distributions, it leads to a straightforward Monte Carlo sampling from (19) and (21), for arbitrarily large n and m. An analogous representation is introduced for the posterior distribution of $\mathfrak{u}_{r,n,m}$, thus filling a gap in the literature. We denote by Binomial(n,p) a Binomial distribution with parameter (n,p), $n \in \mathbb{N}$ and $p \in (0,1)$.

PROPOSITION 1. *Let $\mathbf{X}_n$ be a random sample from $P \sim PYP(\alpha,\theta)$, for $\alpha \in [0,1)$ and $\theta > -\alpha$, such that $\mathbf{X}_n$ features $K_n = k$ and $\mathbf{N}_n = (n_1,\ldots,n_k)$. Moreover, denote by K_m^* and $M_{r,m}^*$ the random numbers of distinct species and distinct species with frequency $r \geq 1$, respectively, in $m \geq 1$ random samples from $P \sim PYP(\alpha,\theta+n)$. Then,*

i) for $\alpha \in (0,1)$

$$\mathfrak{u}_{n,m} \,|\, \mathbf{X}_n \sim \textit{Binomial}\left(K_m^*,\, B_{\frac{\theta}{\alpha}+k,\,\frac{n}{\alpha}-k}\right), \tag{25}$$

with $B_{\theta/\alpha+k,n/\alpha-k}$ being a Beta r.v. with parameter $(\theta/\alpha+k, n/\alpha-k)$, which is independent of the r.v. K_m^*;

ii) for $\alpha = 0$

$$\mathfrak{u}_{n,m} \,|\, \mathbf{X}_n \sim \textit{Binomial}\left(K_m^*,\, \frac{\theta}{\theta+n}\right); \tag{26}$$

iii) for $\alpha \in (0,1)$

$$\mathfrak{u}_{r,n,m} \,|\, \mathbf{X}_n \sim \textit{Binomial}\left(M_{r,m}^*,\, B_{\frac{\theta}{\alpha}+k,\,\frac{n}{\alpha}-k}\right), \tag{27}$$

with $B_{\theta/\alpha+k,n/\alpha-k}$ being a Beta r.v. with parameter $(\theta/\alpha+k, n/\alpha-k)$, which is independent of the r.v. $M_{r,m}^*$;

iv) for $\alpha = 0$

$$\mathfrak{u}_{r,n,m} \,|\, \mathbf{X}_n \sim \textit{Binomial}\left(M^*_{r,m}, \frac{\theta}{\theta+n}\right). \tag{28}$$

See Section S3 of the Supplementary Material [Balocchi et al., 2024+] for the proof of Proposition 1. Proposition 1 is a consequence of the conjugacy and quasi conjugacy properties of the DP and the PYP, respectively [Ferguson, 1973; Pitman, 1996]. From (25) the posterior distribution (19) is the distribution of the number of successes in a random number K^*_m of independent Bernoulli trials, each trial with a Beta random probability $B_{\theta/\alpha+k,n/\alpha-k}$ of success. Note that the expectation of $B_{\theta/\alpha+k,n/\alpha-k}$ is the BNP estimator of the missing mass (13). We write

$$\hat{\mathfrak{u}}_{n,m} = \frac{\theta+k\alpha}{\theta+n}\mathbb{E}[K^*_m],$$

where a simple expression for $\mathbb{E}[K^*_m]$ is available in Pitman [2006, Chapter 3]. From (26), the posterior distribution (21) coincides with the distribution of the number of successes in a random number K^*_m of independent Bernoulli trials, each trial with probability $\theta/(\theta+n)$ of success, i.e. the BNP estimator of the missing mass (13). Then,

$$\hat{\mathfrak{u}}_{n,m} = \frac{\theta}{\theta+n}\mathbb{E}[K^*_m].$$

Along similar lines, from (27) and (28) we obtain an alternative representation of the BNP estimator $\mathfrak{u}_{r,n,m}$ [Favaro et al., 2013], i.e.,

$$\hat{\mathfrak{u}}_{r,n,m} = \frac{\theta+k\alpha}{\theta+n}\mathbb{E}[M^*_{r,m}],$$

where a simple closed-form expression for $\mathbb{E}[M^*_{r,m}]$ is available in Favaro et al. [2013, Proposition 1]. According to (27), for fixed $\alpha \in (0,1)$ and $\theta > -\alpha$ the number K_n of distinct species in the sample is a sufficient statistic to make inference on $\mathfrak{u}_{r,n,m}$. Moreover, if $\alpha = 0$, i.e. under the DP, then the sample size n is a sufficient statistic to infer $\mathfrak{u}_{r,n,m}$.

The representations (25) and (26) are useful for Monte Carlo sampling from the posterior distributions (19) and (21). They allow to sample from (19) and (21) for arbitrarily large values of n and m, thus avoiding the use of the large m approximation proposed in Favaro et al. [2009]. For fixed $\alpha \in (0,1)$ and $\theta > -\alpha$, Monte Carlo sampling from (19) consists of three steps: i) sample from Beta$(\theta/\alpha+k,\, n/\alpha-k)$; ii) independently of step i), sample the r.v. K^*_m under $P \sim \text{PYP}(\alpha, \theta+n)$; iii) given step i) and step ii), sample from Binomial$(K^*_m,\, B_{\theta/\alpha+k,n/\alpha-k})$. If $\alpha = 0$, i.e. under the DP, then Monte Carlo sampling of (21) consists of two steps: i) sample the r.v. K^*_m under $P \sim \text{PYP}(0, \theta+n)$; ii) given step i), sample from Binomial$(K^*_m,\, \theta/(\theta+n))$. Sampling from Beta and Binomial distributions is straightforward, for arbitrarily large n and m, and routines are available in standard software. Sampling K^*_m is also straightforward, for arbitrarily large n and m, and it exploits the predictive probabilities of the PYP (4). In particular, let Bernoulli(p) be the Bernoulli distribution with parameter p, for $p \in (0,1)$. Sampling K^*_m then reduces to sample $m-1$ Bernoulli r.v.s:

1) Set $k = 1$;
2) For $i = 1$ to $m-1$
   Set b to be a sample from Bernoulli$((\theta+n+\alpha k)/(\theta+n+i))$;
   Set $k = k+b$;
3) Return k.

Along the same lines, the representations (27) and (28) are useful for Monte Carlo sampling from the posterior distribution of $\mathfrak{u}_{r,n,m}$, given $\mathbf{X}_n$. In particular, they require to sample the r.v. $M^*_{r,m}$ rather than the r.v. K^*_m. Sampling of $M^*_{r,m}$ still exploits the predictive probabilities of the PYP although, regrettably, it does not reduce to sample Bernoulli r.v.s.

5. COVERAGES OF PREVALENCES

The estimation of coverages of prevalences, or saturations, may be viewed as the m-step ahead generalization of the problem of estimating coverage probabilities. For $m \geq 1$ let $\mathbf{X}_{n+m} = (X_1, \ldots, X_n, X_{n+1}, \ldots, X_{n+m})$ be a random sample under the "classical framework" for SSPs, of which only the first n elements are assumed to be observed, and denote by $(N_{j,n})_{j\geq 1}$ and $(N_{j,m})_{j\geq 1}$ the species' frequencies in $\mathbf{X}_n$ and $(X_{n+1}, \ldots, X_{n+m})$, respectively. The coverage of prevalence of order $r \geq 0$ is defined as

$$\mathfrak{f}_{r,n,m} = \sum_{j\geq 1} I(N_{j,n} = r) I(N_{j,m} > 0),$$

namely the number of species observed r times that would be observed if m additional samples were collected from the same population. Note that $\mathfrak{f}_{0,n,m}$, i.e. the coverage of prevalence of order 0, is the number of unseen species $\mathfrak{u}_{n,m}$. The estimation of coverages of prevalences first appeared in linguistics to answer the question "Did Shakespeare write a newly-discovered poem?" [Thisted and Efron, 1987]. The estimated coverages of prevalences was applied to test the consistency of the word usage in a previously unknown poem attributed to Shakespeare with the word usage in the entire Shakespearean canon. Similar questions, though under different vests, appear in biological sciences. In genomics data, they appear in relation to coverage depth, i.e. the average number of reads that are aligned to known reference bases [Deng et al., 2019]. Depending on the application, different levels of coverage

might be required. This leads to determine whether additional sequencing is needed, which can be motivated by a required minimum coverage threshold, or investigating rare events. The estimation of coverages of prevalences can be valuable in such a setting, as it allows researchers to estimate how many molecules observed with a given frequency would be observed again if the sampling efforts were increased.

5.1 A nonparametric estimator of $\mathfrak{f}_{r,n,m}$

Under the "classical framework", Thisted and Efron [1987] introduced an estimator of $\mathfrak{f}_{r,n,m}$. In particular, if $\lambda = m/n$ and $M_{r,n} = m_r$ denotes the number of distinct species with frequency $r \geq 1$ in $\mathbf{X}_n$, then the estimator is given by

$$\tilde{\mathfrak{f}}_{r,n,m} = \sum_{j\geq 1} (-1)^{i+1} \lambda^i \binom{r+i}{i} m_{r+i}. \tag{29}$$

For $\lambda = n^{-1}$, the estimator (29) reduces to the Good-Turing estimator of coverage probability $\mathfrak{p}_{r,n}$. The estimator (29) is a nonparametric estimator of $\mathfrak{f}_{r,n,m}$, in the sense that it does not rely on any distributional assumption on the unknown p. It was obtained in Thisted and Efron [1987] through a nonparametric empirical Bayes derivation, though an heuristic derivation by a comparison between the expectations of $\mathfrak{u}_{n,m}$ and $M_{r,n}$ is also possible. The study of provable guarantees of $\tilde{\mathfrak{f}}_{r,n,m}$ has not yet been considered in the literature. Thisted and Efron [1987] showed that $\tilde{\mathfrak{f}}_{r,n,m}$ empirically estimates $\mathfrak{f}_{r,n,m}$ for $\lambda < 1$, but without provable guarantees. For $\lambda > 1$ we expect that $\tilde{\mathfrak{f}}_{r,n,m}$ will suffer the same variance issue of $\tilde{\mathfrak{u}}_{n,m}$, that is the geometrically increasing magnitude of $\binom{r+i}{i}\lambda^i$ may produce wild oscillations as the number of terms increases. Because of the additional Binomial term $\binom{r+i}{i}$, it is natural to expect that such a variance issue worsens as r increases. As for the study of its provable guarantees, we expect that the theory developed to study minimax optimality $\tilde{\mathfrak{u}}_{n,m}$ for $\lambda \geq 1$ [Wu and Yang, 2016, 2019; Polyanskiy and Wu, 2020] is not of a direct applicability to $\tilde{\mathfrak{f}}_{r,n,m}$.

5.2 BNP inference for coverages of prevalences

In the "BNP framework" (1), the problem of estimating the coverages of prevalences is open, and here we cover this gap. Before computing the posterior distribution of $\mathfrak{f}_{r,n,m}$, given $\mathbf{X}_n$, it is useful to recall the definition of (general) hypergeometric distribution [Johnson et al., 2005, Chapter 6.2.5], as well as the definition of generalized factorial distribution [Charalambides, 2005, Chapter 2]. For any $u \in \mathbb{N}$ and $b, c > 0$, we say that a r.v. $U_{b,c,u}$ on $\{1, \ldots, u\}$ has a generalized factorial distribution if, for $x \in \{1, \ldots, u\}$,

$$\Pr[U_{b,c,u} = x] = \frac{1}{(bc)_{(u)}} \mathscr{C}(u, x; b, 0)(c)_{(x)}; \tag{30}$$

see Section S1 of the Supplementary Material [Balocchi et al., 2024+] for details. Moreover, for $u, v \in \mathbb{N}$ and $a > 0$ such that $a > u$, a r.v. $H_{a,u,v}$ on $\{0, 1, \ldots, u\}$ has a (general) hypergeometric distribution if, for $x \in \{0, 1, \ldots, u\}$ if holds

$$\Pr[H_{a,u,v} = x] = \frac{\binom{a}{x}\binom{v}{u-x}}{\binom{a+v}{u}}. \tag{31}$$

In the next proposition, we show that the posterior distribution of $\mathfrak{f}_{r,n,m}$, given $\mathbf{X}_n$, admits a representation in terms of a compound (general) hypergeometric distribution.

PROPOSITION 2. *Let $\mathbf{X}_n$ be a random sample from $P \sim PYP(\alpha, \theta)$, for $\alpha \in [0, 1)$ and $\theta > -\alpha$, such that $\mathbf{X}_n$ features $K_n = k$ and $\mathbf{N}_n = (n_1, \ldots, n_k)$. If $M_{r,n} = m_r$ is the number of distinct species with frequency $r \geq 1$ in $\mathbf{X}_n$, then*

$$\mathfrak{f}_{r,n,m} \,|\, \mathbf{X}_n \overset{d}{=} m_r - H_{\frac{\theta+n}{r-\alpha}-1, m_r, U_{r-\alpha, \frac{\theta+n}{r-\alpha}, m}}. \tag{32}$$

See Section S4 of the Supplementary Material [Balocchi et al., 2024+] for the proof of Proposition 2. From (32), for fixed $\alpha \in [0, 1)$ and $\theta > -\alpha$, $M_{r,n}$ is a sufficient statistic to make inference on $\mathfrak{f}_{r,n,m}$. Under a squared loss function, a BNP estimator of $\mathfrak{f}_{r,n,m}$ is the expected value of (32), i.e.,

$$\hat{\mathfrak{f}}_{r,n,m} = \mathbb{E}[\mathfrak{f}_{r,n,m} \,|\, \mathbf{X}_n] = m_r \left(1 - \frac{(\theta + n - r + \alpha)_{(m)}}{(\theta + n)_{(m)}}\right). \tag{33}$$

See Section S5 of the Supplementary Material [Balocchi et al., 2024+] for the proof of Equation 33. For $m = 1$ the estimator (33) reduces to the estimator (14) of $\mathfrak{p}_{r,n}$. Interestingly, the estimator (33) may be interpreted as the proportion

$$w_{r,n,m}(\alpha, \theta) = 1 - \frac{(\theta + n - r + \alpha)_{(m)}}{(\theta + n)_{(m)}} \in (0, 1),$$

of the number m_r of distinct species with frequency r. Uncertainty quantification of (33) is obtained by means of credible intervals via Monte Carlo sampling of the posterior distribution (32). Monte Carlo sampling of (32) is doable for small values of n and m, and it becomes impossible for large n or m. This is because for large n and m, and even only for large m, the computational burden for evaluating the generalized factorial coefficients in the distribution of $U_{r-\alpha,(\theta+n)/(r-\alpha),m}$. To overcome this drawback, we propose an approximation of the posterior distribution (32) for large n and m. Let $a_{n,m} \simeq b_{n,m}$ mean that $\lim_{n\to+\infty} \lim_{m\to+\infty} a_{n,m}/b_{n,m} = 1$, namely $a_{n,m}$ and $b_{n,m}$ are asymptotically equivalent as n and m

tends to infinity. Then, as $n, m \to +\infty$ with $m > 0$, for $x \in \{0, 1, \ldots, m_r\}$

(34)

$$\Pr[\mathfrak{f}_{r,n,m} = x \,|\, \mathbf{X}_n] \simeq \binom{m_r}{x} \left[1 - \left(\frac{n}{n+m}\right)^{r-\alpha}\right]^{x} \left[\left(\frac{n}{n+m}\right)^{r-\alpha}\right]^{m_r - x},$$

and hence

$$\hat{\mathfrak{f}}_{r,n,m} \simeq m_r \left[1 - \left(\frac{n}{n+m}\right)^{r-\alpha}\right].$$

See Section S6 of the Supplementary Material [Balocchi et al., 2024+] for the proof of Equation 34. From (34), for large n and m with $m > 0$, the posterior distribution (32) admits a first order local approximation in terms of a Binomial distribution with parameters $(m_r, 1 - (n/(n+m))^{r-\alpha})$, with m_r being the number of trials and $1 - (n/(n+m))^{r-\alpha}$ being the probability of success at the single trial.

5.3 Disclosure risk assessment

We conclude this section with a SSP related to the coverage of prevalence of order 1, which first appeared in the context of disclosure risk for data confidentiality [Willenborg and Waal, 2001]. Consider a microdata sample of $n \geq 1$ units (individuals) from a population of $N > n$ units, such that each unit contains identifying and sensitive information. Identifying information consists of categorical variables which might be matchable to known units of the population. A threat of disclosure results from the possibility of identifying an individual through such a matching, and hence disclose its sensitive information. To quantify disclosure risk, microdata units are partitioned according to a categorical variable formed by cross-classifying the identifying variables; that is, units are partitioned into non-empty cells containing individuals with the same combinations of values of identifying variables. Intuitively, a risk of disclosure arises from cells with frequency 1 since, assuming no errors in matching processes or data sources, for these cells the match is guaranteed to be correct [Bethlehem et al., 1990; Skinner et al., 1994; Skinner and Elliot, 2002]. Under the "classical framework" for SSPs, if microdata units are modeled as a random sample $\mathbf{X}_N$ in the classical species sampling framework, and $(N_{j,n})_{j\geq 1}$ and $(N_{j,m})_{j\geq 1}$ are the frequencies of cells (species) in $\mathbf{X}_n$ and $(X_{n+1}, \ldots, X_{n+N})$, respectively, then a popular measure of disclosure risk is defined as

$$\mathfrak{d}_{n,m} = \sum_{j\geq 1} I(N_{j,n} = 1) I(N_{j,m} = 0),$$

namely the number of cells with frequency 1 in the observed sample that are also of frequency 1 in the whole population. We refer to Camerlenghi et al. [2021] and Favaro et al. [2021b] for the estimation of $\mathfrak{d}_{n,m}$ within the "classical framework" and in the "BNP framework", respectively.

6. ESTIMATION OF PRIOR'S PARAMETERS (α, θ)

In practice, a poor assessment of the prior's parameters (α, θ) may harm any statistical inference based upon the PYP(α, θ) prior due to its rigidity. Indeed, many of the estimators derived in the previous sections critically depend on the parameter α, and in a less extent on the parameter θ. For instance, let us consider the estimator of the number of unseen species obtained in Section 4.2. As a direct consequence of (23), $\mathfrak{u}_{n,m} \sim Z m^{\alpha}$ as $m \to \infty$, for some r.v. Z. Hence, a bad choice for the parameter α may lead to a poor prediction. A natural way to add flexibility in the model and improve the inference is to estimate the prior's parameters (α, θ). Denoting by $\Phi = \{(\alpha, \theta) \in (0,1) \times \mathbb{R} : \theta > -\alpha\}$, in this section we investigate the problem of estimating (α, θ) over the set Φ.

We start by considering the case where data come from the model, the so-called well-specified case. That is, we assume the "BNP framework" (1) for a choice of the prior's parameters $(\alpha, \theta) \in \Phi$. We study the empirical Bayes approach, in which the parameter (α, θ) is estimated by maximizing the marginal likelihood function, and the hierarchical Bayes approach, where a prior distribution is placed over Φ. We show that, as $n \to +\infty$, both approaches are consistent with respect to the estimation of α. More surprising, we prove that both the approaches are inconsistent with respect to the estimation of θ. These consistency and inconsistency results are also illustrated through simulations in Section 7. Ultimately, we prove a minimax lower bound for the estimation of θ, which proves that the inconsistency issue is fundamental, and no procedure can estimate θ with vanishing maximum risk. We also consider a more general point of view than the well-specified case, thus no longer assuming the "BNP framework" for a choice of the prior's parameters. We consider instead the "classical framework", namely data are modeled as a random sample from a fixed probability measure p satisfying a regularity assumption enabling meaningful inference. In particular, we show that under this assumption, the sequence $((\hat{\alpha}_n, \hat{\theta}_n))_{n\geq 1}$ of marginal maximum likelihood estimators has an interpretable limit (α_*, θ_*). We then establish the asymptotic shape of the sequence of posterior distributions in the hierarchical Bayes model. Our result shows that the sequence of posterior distributions is consistent for α_*, but inconsistent for θ_*. Thereby, we demonstrate that inference made upon the PYP can be robust to some form of misspecification of the model.

6.1 Consistency and inconsistency in the well-specified case

For the well-specified case, we consider the "BNP framework" under the assumption that there exists a "true" parameter $(\alpha_0, \theta_0) \in \Phi$. Specifically, $\mathbf{X}_n$ is random sample from $P \sim \mathrm{PYP}(\alpha_0, \theta_0)$. The goal is to analyze the large n asymptotic behaviour of the empirical Bayes and hierarchical Bayes approaches under the assumption that $\mathbf{X}_n$ is distributed as the the prior predictive distribution, here denoted $P_n^{\alpha_0,\theta_0}(\cdot) = \int P^{\otimes n}(\cdot)\mathrm{dPYP}_{\alpha_0,\theta_0}(P)$ in the sequel. For the empirical Bayes approach, the parameter (α, θ) is estimated using a maximizer of the marginal likelihood function:

$$L_n(\alpha,\theta) = n!\frac{\left(\frac{\theta}{\alpha}\right)_{(\sum_{i=1}^n M_{i,n})}}{(\theta)_{(n)}}\prod_{i=1}^n \frac{\left(\frac{\alpha(1-\alpha)_{(i-1)}}{i!}\right)^{M_{i,n}}}{M_{i,n}!},$$

namely the EPSF (5) as a function of the parameter (α, θ) for the observed $\mathbf{M}_n = (M_{1,n}, \ldots, M_{n,n})$. A maximizer of L_n is known as the marginal maximum likelihood estimator (MMLE). The next theorem characterizes the asymptotic limit (in probability) of $((\hat{\alpha}_n, \hat{\theta}_n))_{n\geq 1}$ as $n \to \infty$.

THEOREM 1. *Assume the random sample to be such that* $\mathbf{X}_n \sim P_n^{\alpha_0,\theta_0}$. *Then, the set* $\arg\max_{(\alpha,\theta)\in\Phi} L_n(\alpha,\theta)$ *is a non-empty set with probability* $1 + o(1)$ *as* $n \to \infty$. *Furthermore,* $(\hat{\alpha}_n, \hat{\theta}_n) \in \arg\max_{(\alpha,\theta)\in\Phi} L_n(\alpha,\theta)$ *is such that*

$$\hat{\alpha}_n = \alpha_0 + o_p(1)$$

and

$$\hat{\theta}_n = Z + o_p(1),$$

where Z *is a r.v. related to* S_{α_0,θ_0} *via the relation* $S_{\alpha_0,\theta_0} = \exp\{\psi(Z/\alpha_0 + 1) - \alpha_0\psi(Z + 1)\}$. *Here,* ψ *denotes the digamma function; that is the function* ψ *is the derivative of* $\log\Gamma$.

See Section S7 of the Supplementary Material [Balocchi et al., 2024+] for the proof of Theorem 1. The theorem establishes that the sequence of MMLE is consistent to estimate α_0. The limit in probability of $(\hat{\theta}_n)_{n\geq 1}$ is, however, a r.v. This shows that θ_0 cannot be estimated consistently using the MMLE.

We continue our analysis of the well-specified case with the hierarchical Bayes approach. We put a prior distribution $G = G_\alpha \otimes G_\gamma$ over the parameter $\alpha \in (0, 1)$ and the shifted parameter $\gamma = \theta + \alpha \in (0, \infty)$. Namely, under G, the r.v. α and γ are independent with respective marginals G_α and G_γ. We denote by Π the joint distribution of $(\alpha, \gamma, P, X_1, X_2, \ldots)$. The posterior distribution of (α, γ) given $\mathbf{X}_n$ is denoted $\Pi(\cdot \mid \mathbf{X}_n)$. See that by Bayes' rule:

$$\Pi\big((\alpha,\gamma) \in A \mid \mathbf{X}_n\big) = \frac{\int_A L_n(\alpha,\gamma-\alpha)\mathrm{d}G(\alpha,\gamma)}{\int_{\mathbb{R}_+^2} L_n(\alpha,\gamma-\alpha)\mathrm{d}G(\alpha,\gamma)}.$$

THEOREM 2. *Assume the random sample to be such that* $\mathbf{X}_n \sim P_n^{\alpha_0,\theta_0}$. *Furthermore, assume that* G_α *and* G_γ *have continuous and positive densities. Then for every* $\varepsilon > 0$

$$\Pi\big(|\alpha - \alpha_0| > \varepsilon \mid \mathbf{X}_n\big) = o_p(1).$$

Moreover, there exists a r.v. $W > 0$ *such that the following holds:*

$$\Pi\big(\theta \in [\theta_0 - W, \theta_0 + W] \mid \mathbf{X}_n\big) \leq \frac{1}{2} + o_p(1).$$

See Section S8 of the Supplementary Material [Balocchi et al., 2024+] for the proof of Theorem 2. The theorem establishes the consistency of the posterior near the "true" α_0, that is the posterior will eventually put all its mass on a small neighborhood of α_0 when the sample size n gets large enough. The posterior is however inconsistent at θ_0: regardless of how large n is taken the posterior will put mass outside of a neighborhood of fixed size [see also Section 7 and Figure 1 for a numerical illustration]. Indeed, as demonstrated later in Theorem 5 the posterior distribution for γ (equivalently θ) depends on the choice of G_γ even in the asymptotic limit. Therefore, in the absence of strong prior belief on θ it is unclear if a hierarchical Bayes approach to estimate this parameter is meaningful.

Theorem 1 and Theorem 2 establish that the prior's parameter θ cannot be estimated consistently using the MMLE or the hierarchical Bayes approach. We conclude this section by demonstrating that the issue is more fundamental and no estimator can estimate θ with vanishing maximum risk.

THEOREM 3. *For all* $\alpha \in (0, 1)$, *for all* $n \geq 1$, *and for all* $t > -\alpha$

$$\inf_{\hat{\theta}_n} \sup_{-\alpha<\theta\leq t} \mathbb{E}_{(\alpha,\theta)}\big((\hat{\theta}_n(\mathbf{X}_n) - \theta)^2\big) \geq \frac{\min\big((t+\alpha)^2, t+\alpha\big)}{64}.$$

where the infimum is understood over all measurable functions of $\mathbf{X}_n$.

See Section S9 of the Supplementary Material [Balocchi et al., 2024+] for the proof of Theorem 3. The theorem shows that even with the knowledge of the prior's parameter α, it is impossible for an estimator of θ to have a vanishing maximum risk over $(-\alpha, t]$ as $n \to \infty$. It also establishes that the minimax risk over $\theta \in (-\alpha, \infty)$ is infinite.

6.2 Consistency and inconsistency in the misspecified case

Differently from the well-specified case, for the misspecified case we do not necessarily assume that $\mathbf{X}_n$ is a random sample from $P \sim \text{PYP}(\alpha_0, \theta_0)$. Instead, we consider the "classical framework", namely $\mathbf{X}_n$ is a random sample from a "true" (fixed) distribution p. To guarantee the existence of meaningful limits for our estimators, we require a moderate misspecification, as stated in the next assumption.

ASSUMPTION 1. *The distribution p is discrete, such that $p = \sum_{j\geq 1} p_j \delta_{s_j}$. Furthermore, by defining the quantity*

$$\bar{F}_p(x) = \sum_{j\geq 1} I(p_j > x),$$

there exist $L > 0$ and $\alpha_* \in (0,1)$ such that as $x \to 0$ it holds

$$\bar{F}_p(x) = L x^{-\alpha_*} + o\Big[\frac{1}{-x^{\alpha_*}\log(x)}\Big].$$

The Assumption 1 is motivated by the fact that if $P \sim \text{PYP}(\alpha, \theta)$, then results in Pitman [2003] show that $\lim_{x\to 0} x^{\alpha}\bar{F}_P(x) = S_{\alpha,\theta}/\Gamma(1-\alpha)$ almost surely, with Γ being the Gamma function [Pitman, 2006, Chapter 3 and Chapter 4]. We strengthen those results in Section S12 of the Supplementary Material [Balocchi et al., 2024+] and we establish a law of the iterated logarithm (LIL) as $x \to 0$ for $\bar{F}_P(x) - S_{\alpha,\theta}x^{-\alpha}/\Gamma(1-\alpha)$ when $P \sim \text{PYP}(\alpha, \theta)$. In particular, the LIL implies that $P \sim \text{PYP}(\alpha, \theta)$ satisfies almost-surely the Assumption 1 with $\alpha_* = \alpha$ and $L = S_{\alpha,\theta}/\Gamma(1-\alpha)$. The Assumption 1 is, however, much more general and allow for many more probability distributions. In the next theorem, we characterize the asymptotic limit of the sequence of MMLEs when the data $\mathbf{X}_n$ is independent from a distribution p satisfying Assumption 1.

THEOREM 4. *Assume $\mathbf{X}_n$ to be a random sample from p, and assume that the distribution p satisfies Assumption 1. Then the set $\arg\max_{(\alpha,\theta)\in\Phi} L_n(\alpha,\theta)$ is a non-empty set with probability $1 + o(1)$ as $n \to \infty$. Furthermore, $(\hat{\alpha}_n, \hat{\theta}_n) \in \arg\max_{(\alpha,\theta)\in\Phi} L_n(\alpha,\theta)$ is such that*

$$\hat{\alpha}_n = \alpha_* + o_p(1)$$

and

$$\hat{\theta}_n = \theta_* + o_p(1),$$

with θ_* defined through

$$L = \frac{\exp\{\psi(\theta_*/\alpha_* + 1) - \alpha_*\psi(\theta_* + 1)\}}{\Gamma(1-\alpha_*)}.$$

Here, ψ is the digamma function; that is ψ is the derivative of $\log\Gamma$.

See Section S10 of the Supplementary Material [Balocchi et al., 2024+] for the proof of Theorem 4. Interestingly, Theorem 4 establishes that if the misspecification is moderate, then (α_*, θ_*) can be directly interpreted in terms of key functionals of the "true" data generative mechanism, that is in terms of α_* and L. This also shed some lights in the meaning of the parameter (α, θ) when the model is correct. In the next theorem, we consider the hierarchical Bayes modeling and we characterize the asymptotic shape of the sequence of posterior distributions as $n \to \infty$.

THEOREM 5. *Let $G = G_\alpha \otimes G_\gamma$ be a (proper or improper) prior distribution over the parameter (α, γ) such that G_α (respectively G_γ) has a density g_α (resp. g_γ) with respect to Lebesgue measure which is positive in a neighborhood of α_* (resp. $\gamma_* = \theta_* + \alpha_*$). Furthermore assume that G_γ is such that $\int_0^\infty \frac{[L\Gamma(1-\alpha_*)]^{z/\alpha_*}\Gamma(1-\alpha_*+z)}{\Gamma(z/\alpha_*)} G_\gamma(\mathrm{d}z) < \infty$[1] and define a probability distribution H_* on $(0,\infty)$ through*

$$H_*(A) = \frac{\int_A \frac{[L\Gamma(1-\alpha_*)]^{z/\alpha_*}\Gamma(1-\alpha_*+z)}{\Gamma(z/\alpha_*)} G_\gamma(\mathrm{d}z)}{\int_0^\infty \frac{[L\Gamma(1-\alpha_*)]^{z/\alpha_*}\Gamma(1-\alpha_*+z)}{\Gamma(z/\alpha_*)} G_\gamma(\mathrm{d}z)}.$$

Then under Assumption 1, the posterior distribution of (α, γ) given $\mathbf{X}_n$, written here $\Pi(\cdot \mid \mathbf{X}_n)$, satisfies as $n \to \infty$

$$\sup_{A,B} \Big| \Pi(\hat{V}_n^{1/2}(\alpha - \hat{\alpha}_n) \in A, \gamma \in B \mid \mathbf{X}_n) - \phi(A)H_*(B) \Big| = o_p(1),$$

where the supremum is taken over all measurable sets, $\hat{V}_n = -\partial_\alpha^2 \log L_n(\hat{\alpha}_n^0, 0)$, and $\phi(A)$ is the probability that a standard normal r.v. lies in A.

See Section S11 of the Supplementary Material [Balocchi et al., 2024+] for the proof of Theorem 5. The theorem establishes that (α, γ) are asymptotically independent a posteriori. We see that the limiting marginal distribution for γ is neither converging toward a point mass, nor even Gaussian. In contrast with the MMLE, the posterior distribution for γ is difficult to interpret since it depends on the prior G_γ. For this reason, we shall be careful with posterior analysis involving γ (equivalently θ). The limiting marginal distribution for $\hat{V}_n^{1/2}(\alpha - \hat{\alpha}_n)$ is however Gaussian, which can be seen as a weak form of Bernstein-von Mises' (BvM) theorem. It is not a true BvM because the centering is taken at $\hat{\alpha}_n$ instead of α_* and the scaling is the "empirical" Fisher information instead of Fisher's information. Finally, let us mention that after writing the

[1] This is not a strong restriction since it allows to use all proper priors, as well as many improper priors.

first draft of this paper, we have been made aware that Franssen and van der Vaart [2022] have independently obtained results similar to those presented in this section. Under an assumption resembling Assumption 1, Franssen and van der Vaart [2022] are able to establish a more precise asymptotic for the estimation of α. They, however, obtain less precise results regarding the estimation of the parameter θ.

7. NUMERICAL ILLUSTRATIONS

In this section, we provide a numerical illustration of the prior parameters estimation and of the inference on the presented species sampling problems, using synthetic data.

We first compare the empirical Bayes approach and the fully Bayes approach by using several prior distributions, and we demonstrate how the effect of the estimation method changes with increasing sample sizes. In particular, we compare different prior distributions on α and θ, constructed using different combination of non-informative and (mis-specified) informative priors: (a) a non-informative prior for both parameters; (b) an informative prior for both parameters; (c) an informative prior for only θ and non-informative for α; (d) an informative prior for only α and non-informative for θ. See Section S13 of the Supplementary Materials [Balocchi et al., 2024+] for additional details. Figure 1 displays the mean error across several simulated datasets, for the different estimation methods as a function of the increasing sample size. While the mis-specified informative priors on α have a detrimental effect on the estimation, this effect gets less strong with the increasing sample size. On the contrary, the effect of mis-specified priors on θ remains large even with large sample size. Moreover, Figure 1 shows that while the relative error for α vanishes with increasing sample size for all the estimation methods, the error for θ does not. This is a clear illustration of the findings in Section 6.1. We then demonstrate the inference and prediction of the various SSP presented in this work, using synthetic data generated from both from the PYP and from a power law Zipf distribution. We compare the estimation and prediction under the Empirical Bayes (EB) approach, and the Full Bayes (FB) approach with non-informative priors on both θ and α, and informative prior for both parameters.

We first analyze the SSPs that depend only on the initial sample and on the "true" distribution P: the missing mass and the coverage probability $\mathfrak{p}_{r,n}$ for $r = 1$. Figure S1 of the Supplementary Materials [Balocchi et al., 2024+] shows the relative (percentage) error for these functionals across several synthetic datasets. The median absolute percentage errors are less than 7% for the missing mass and the coverage probability for the PY-generated data,

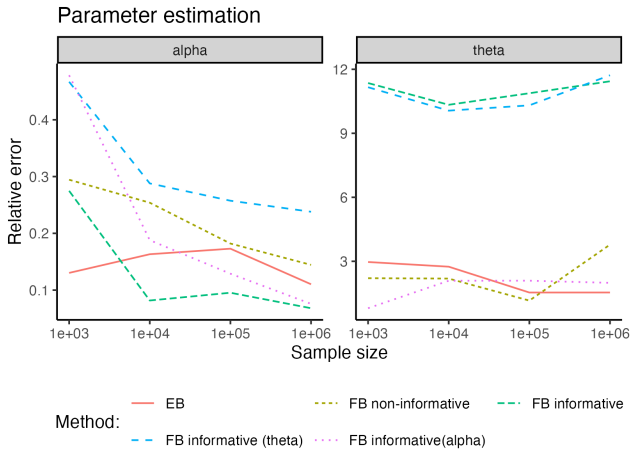

FIG 1. *Mean absolute relative (percentage) error for estimation of α and θ under the Empirical Bayes (EB) approach and the Full Bayes (FB) approach with several prior distributions.*

and less than 12% for the Zipf-generated data, demonstrating good recovery of the true functionals. We then study the "predictive" SSP (the ones that depend on an additional sample): the number of unseen species, the unseen prevalence $\mathfrak{u}_{r,n,m}$ for $r = 1$ and the coverage of prevalence $\mathfrak{f}_{r,n,m}$ for $r = 1$. The median absolute percentage error (across methods and additional sample sizes) is less then 30% for the number of unseen and the unseen prevalence, and 7% for the coverage of prevalences, when the data is generated from a PYP. Figure 2 displays the predicted and actual value for a synthetic dataset that can be thought as "representative", as it achieved the median error across the several generated datasets. For all these SSP, the prediction is somewhat accurate, and most often the credible intervals contain the "true" value of the functional of interest. When the data was generated from a power-law Zipf distributios, the predictive performance gets worse, with median errors less then 40% for the number of unseen and the unseen prevalence and 60% for the coverage of prevalences. We refer to Section S13 of the Supplementary Materials [Balocchi et al., 2024+] for more details.

8. SOME GENERALIZATIONS OF SSPS

"Feature-sampling" problems (FSPs) generalize SSPs by allowing an individual in the population to belong to more than one species, which are referred to as features. To introduce the class of FSPs, we consider a population of individuals such that each individual is endowed with a finite set of features' labels belonging to a (possibly infinite) space of features. The "classical framework" for FSPs assumes that $n \geq 1$ observed samples from the population are modeled as a random sample $(Y_1, \ldots, Y_n)$, where $Y_i = (Y_{i,j})_{j \geq 1}$ is a sequence of independent Bernoulli r.v.s with unknown feature probabilities $(p_j)_{j \geq 1}$, such that Y_r is independent of Y_s for any

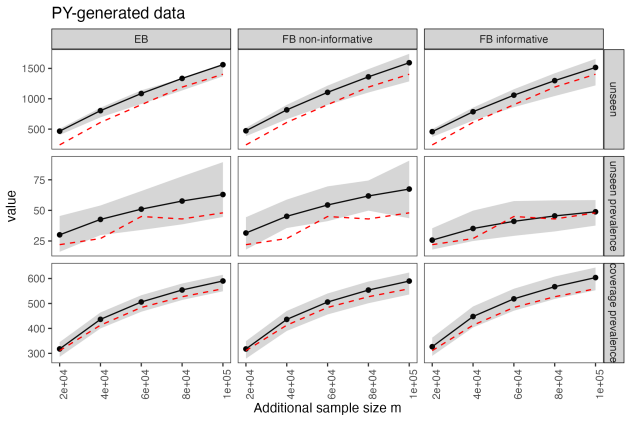

FIG 2. *Representative examples of the predicted value (black solid line) and actual value (red dashed line) for three SSP (the number of unseen $\mathfrak{u}_{n,m}$, unseen prevalence $\mathfrak{u}_{1,n,m}$, and coverage of prevalence $\mathfrak{f}_{1,n,m}$) as a function of the additional sample size m, for three parameter estimating method (empirical Bayes (EB), fully Bayes (FB) with non-informative priors, FB with informative priors). The gray bands represent the 95% credible intervals.*

$r \neq s$. In principle, each of SSPs discussed in this paper admits a corresponding feature sampling counterpart. Within the broad class of FSPs, in recent years there has been a growing interest, especially biological sciences, in the estimation of

$$\sum_{j\geq 1} I\left(\sum_{i=1}^{n} Y_{i,j} = 0\right) I\left(\sum_{i=1}^{m} Y_{n+i,j} > 0\right), \tag{35}$$

namely the number of hitherto unseen features that would be observed if m additional samples $(Y_{n+1},\ldots,Y_{n+m})$ were collected from the same $(p_j)_{j\geq 1}$. We refer to Ionita-Laza et al. [2009], Gravel [2014], Zou et al. [2016], Orlitsky et al. [2016] Chakraborty et al. [2019] for parametric and nonparametric approaches to estimate (35). The FSP (35) provides the natural feature sampling counterpart of the problem of the unseen-species problem; we refer to Ayed et al. [2021] for classical (frequentist) nonparametric inference of a feature sampling counterpart of the problem of estimating the missing mass. Recently Masoero et al. [2022] and Masoero et al. [2021] proposed a BNP approach to estimate (35), which rely on placing suitable prior distributions on the underlying probabilities $(p_j)_{j\geq 1}$. See Beraha and Favaro [2023] for a more general setting. Despite these works, BNP inference for FSPs remains still a mostly unexplored field for both methods and applications. See, e.g., Beraha et al. [2023] and Masoero et al. [2023].

SSPs may be generalized to multiple populations of individuals, in such a way that populations share species. Consider $r > 1$ populations of individuals, such that each individual is labeled by a symbol or species' label belonging to a (possibly infinite) space of symbols. That is, species' labels are shared among the populations. Following the "classical framework" for SSPs, it is assumed that r observed samples of individuals from the populations, the i-th sample being of size n_i, are modeled as a random sample $\{(X_{i,1},\ldots,X_{i,n_i})\}_{i=1,\ldots,r}$ from a collection of r unknown distributions $(p_1,\ldots,p_r)$. Then, interest is in estimating discrete functionals that encode features of additional unobserved samples from the same populations; of special interest are functionals encoding information of the number of shared species among populations. In recent years, SSPs with multiple populations have become critical in microbiome studies, i.e. microbial ecology and biology, where next generation sequencing has been applied to obtain inventories of bacteria in many different environments (populations); see Jeganathan and Holmes [2021] and references therein. Nonparametric inference for SSPs with multiple populations pose challenging mathematical hurdles to overcome [Raghunathan et al., 2017; Hao and Li, 2020]. In particular, the BNP approach requires to place a nonparametric prior on the underlying collection of distributions $(p_1,\ldots,p_r)$, in such a way to model the unknown species compositions of the populations and the dependency among these compositions. Hierarchical priors [Teh et al., 2006; Camerlenghi et al., 2019] and compound priors [Griffin and Leisen, 2017] provide a broad class of priors for $(p_1,\ldots,p_r)$, being mathematically tractable and flexible in terms of prior's parameters. However, to date, BNP inference for SSPs in multiple populations is mostly unexplored, being difficult to obtain posterior inferences that are analytically tractable and, most importantly, computationally efficient in applications.

In the "classical framework" for SSPs, the observed samples are modeled as a random sample $\mathbf{X}_n$ from an unknown distribution p. The assumption of independence among the X_i's is unrealistic in many applications, though it yields results that are interesting in themselves, and upon which more sophisticated frameworks may be built. For instance, in a natural languages the probability of appearance of a word strongly depends on the previous words, both for grammatical and semantic reasons. Likewise, the nucleotides in a DNA sequence do not form a random sample. There has been a recent interest in estimating the missing mass and coverage probabilities when observed samples are modeled as Markov chains [Asadi et al., 2014; Falahatgar et al., 2016; Hao et al., 2018; Wolfer and Kontorovich, 2019; Skorski, 2020; Cha et al., 2021; Pananjady et al., 2024]. Bacallado et al. [2013] first considered BNP analysis of SSPs under the assumption that the observed samples are modeled as a reversible Markov chain. This is motivated by the analysis of benchtop and computer experiments that produce data associated with the structural fluctuations of a protein in water, with species being protein conformational states [Pande et al., 2010]. Bacallado et al. [2013] introduced a nonparametric prior for the unknown transition

kernel of a reversible Markov chain, such that: i) the state space of the chain is uncountable; ii) the prediction for the next state visited by the chain is not solely a function of the number of transitions observed in and out of the last state, but transition probabilities out of different states share statistical strength. While the model of Bacallado et al. [2013] can be used to predict characteristics of future unobserved trajectories of reversible Markov chains, i.e. protein dynamics, the major goals of the their paper were: i) predicting how soon the chain will return to a specific state of interest; ii) predicting the number of states that the chain has not yet visited in the first (observed) transitions and that will appear in subsequent (unobserved) transitions.

ACKNOWLEDGEMENTS

The authors are grateful to the Editor (Professor Sonia Petrone), the Associate Editor and three Referees for their comments and corrections that allow to improve remarkably the paper. Stefano Favaro wishes to thank Timothy P. Daley for useful discussions on applications of species sampling problems in biological and physical sciences. Cecilia Balocchi, Stefano Favaro and Zacharie Naulet received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 817257. Stefano Favaro gratefully acknowledge the support from the Italian Ministry of Education, University and Research (MIUR), "Dipartimenti di Eccellenza" grant 2018-2022.

SUPPLEMENTARY MATERIAL

Supplemental Materials for Bayesian nonparametric inference for "species-sampling" problems
A manuscript containing detailed proofs of all results presented in this paper, and additional results regarding the numerical illustrations.

REFERENCES

ACHARYA, J., BAO, Y., KANG, Y. AND SUN, Z. (2018). Improved bounds on minimax risk of estimating missing mass. In *IEEE International Symposium on Information Theory*, 326–330.

ANEVSKI, D., GILL, R.D. AND ZOHREN, S. (2017). Estimating a probability mass function with unknown labels. *The Annals of Statistics* **45**, 2708–2735.

ARBEL, J., FAVARO, S., NIPOTI, B. AND TEH, Y.W. (2017). Bayesian nonparametric inference for discovery probabilities: credible intervals and large sample asymptotics. *Statistica Sinca* **27**, 839–858.

ASADI, M., TORGHABEH, R.P. AND SANTHANAM, N.P. (2014). Stationary and transition probabilities in slow mixing, long memory Markov processes *IEEE Transactions on Information Theory* **60**, 5682–5701.

AYED, F., BATTISTON, M., CAMERLENGHI, F. AND FAVARO, S. (2018). On consistent and rate optimal estimation of the missing mass. *Annales de l'Institut Henri Poincaré - Probabilités et Statistques* **57**, 1476–1494.

AYED, F., BATTISTON, M., CAMERLENGHI, F. AND FAVARO, S. (2021). Consistent estimation of small masses in feature sampling. *Journal of Machine Learning Research* **22**, 1–28.

BACALLADO, S., BATTISTON, M., FAVARO, S. AND TRIPPA, L. (2015). Sufficientness postulates for Gibbs-type priors and hierarchial generalizations. *Statistical Science* **32**, 487–500.

BACALLADO, S., FAVARO, S. AND TRIPPA, L. (2013). Bayesian nonparametric analysis of reversible Markov chains *The Annals of Statistics* **41**, 870–896.

BALOCCHI, C., FAVARO, S. AND NAULET, Z. (2024+). Supplemental Materials for Bayesian nonparametric inference for "species-sampling" problems.

BEN-HAMOU, A., BOUCHERON, S. AND GASSIAT, E. (2018). Pattern coding meets censoring: (almost) adaptive coding on countable alphabets. *Preprint arXiv:1608.08367*.

BEN-HAMOU, A., BOUCHERON, S. AND OHANNESSIAN, M.I. (2017). Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications. *Bernoulli* **23**, 249–287.

BERAHA, M. AND FAVARO, S. (2023). Transform-scaled process priors for trait allocations in Bayesian nonparametrics. *Preprint arXiv:2303.17844*.

BERAHA, M., MASOERO, L., FAVARO, S. AND RICHARDSON, T.S. (2023). A nonparametric Bayes approach to online activity prediction. *Preprint arXiv:2401.14722*.

BERCU, B. AND FAVARO, S. (2024). A martingale approach to Gaussian fluctuations and laws of iterated logarithm for Ewens-Pitman model. *Preprint arXiv:2404.07694*.

BETHLEHEM, J.G., KELLER, W.J. AND PANNEKOEK, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association* **85** 38–45.

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.

BUNGE, J. AND FITZPATRICK, M. (1993) Estimating the number of species: a review. *Journal of the American Statistical Association* **88**, 364-373.

CAI, D., MITZENMACHER, M. AND ADAMS, R.P. (2018). A Bayesian nonparametric view on count–min sketch. In *Advances in Neural Information Processing Systems*, **31**.

CAMERLENGHI, F., FAVARO, S., MASOERO, L. AND BRODERICK, T. (2020). Scaled process priors for Bayesian nonparametric estimation of the unseen genetic variation. *Journal of the American Statistical Association*, to appear.

CAMERLENGHI, F., LIJOI, A., ORBANZ, P. AND PRÜNSTER, I. (2019). Distribution theory for hierarchical processes. *The Annals of Statistics* **47**, 67–92.

CAMERLENGHI, F., FAVARO, S., NAULET, Z. AND PANERO, F. (2021). Optimal disclosure risk assessment. *The Annals of Statistics* **49**, 723–744.

CEREDA, G. (2017) Impact of model choice on LR assessment in case of rare haplotype match (frequentist approach). *Scandinavian Journal of Statistics* **44**, 230-248.

CHAKRABORTY, S., ARORA, A., BEGG, C,B. AND SHEN, R. (2019) Using somatic variant richness to mine signals from rare variants in the cancer genome. *Nature Communications* **10**, 5506.

CHANDRA, P., THANGARAJ, A. AND RAJARAMAN, N. (2021) Missing mass of rank-2 Markov chains. *Preprint arXiv:2102.01938*.

CHAO, A. (2017) Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics* **11**, 256-270.

CHAO, A. AND LEE, S. (1992) Estimating the number of classes via sample coverage. *Journal of the American Statistical Association* **87**, 210-217.

CHARALAMBIDES (2005) *Combinatorial methods in discrete distributions.* Wiley.

CLAUSET, A., SHALIZI, C.R. AND NEWMAN, M.E.J. (2009). Power-law distributions in empirical data. *SIAM Review* **51**, 661–703.

DALEY, T. AND SMITH, A.D. (2013). Predicting the molecular complexity of sequencing libraries. *Nature Methods* **10**, 325–327.

DALEY, T. AND SMITH, A.D. (2014). Modeling genome coverage in single-cell sequencing. *Bioinformatics* **30**, 3159–3165.

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.

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.

DEVROYE, L. (2009). Random variate generation for exponentially and polynomially tilted stable distribution. *ACM Transactions on Modeling and Computer Simulation* **19**, 4.

DOLERA, E. AND FAVARO, S. (2020a). A Berry–Esseen theorem for Pitman's α–diversity. *The Annals of Applied Probability* **30**, 847–869.

DOLERA, E. AND FAVARO, S. (2020b). Rates of convergence in de Finetti's representation theorem, and Hausdorff moment problem. *Bernoulli* **26**, 1294–1322.

EFRON, B. (2003). Robbins, empirical Bayes and micorarrays *The Annals of Statistics* **31**, 366–378.

EFRON, B. AND THISTED, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika* **63**, 435–447.

ESTY, W.W. (1982). Confidence intervals for the coverage of low coverage samples. *The Annals of Statistics* **10**, 190–196.

ESTY, W.W. (1983). A normal limit law for a nonparametric estimator of the coverage of a random sample. *The Annals of Statistics* **11**, 905–912.

EWENS, W. (1972). The sampling theory or selectively neutral alleles. *Theoretical Population Biology* **3**, 87–112.

FALAHATGAR, M., ORLITSKY, A., PICHAPATI, V. AND SURESH, A.T. (2016). Learning Markov distributions: does estimation trump compression? In *Proceedings of the IEEE International Symposium on Information Theory*, 2689–2693.

FAVARO, S., LIJOI, A. AND PRÜNSTER, I. (2012). A new estimator of the discovery probability. *Biometrics* **68**, 1188–1196.

FAVARO, S., LIJOI, A. AND PRÜNSTER, I. (2013). Conditional formulae for Gibbs-type exchangeable random partitions. *The Annals of Applied Probability* **23**, 1721–1754.

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**, 992–1008.

FAVARO, S., NIPOTI, B. AND TEH, Y.W. (2016). Rediscovery of Good-Turing estimators via Bayesian nonparametrics. *Biometrics* **72**, 136–145.

FAVARO, S., PANERO, F. AND RIGON, T. (2021b). Bayesian nonparametric disclosure risk assessment. *Electronic Journal of Statistics* **15**, 5626–5651

FAVARO, S., NAULET, Z. (2023). Near-optimal estimation of the unseen under regularly varying tail populations. *Bernoulli* **29**, 3423–3442.

FAVARO, S., NAULET, Z. (2024). Near-optimal estimation of the unseen under regularly varying tail populations. *Preprint arXiv:2306.14998*.

FERGUSON, T.S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1**, 209–230.

FISHER, R.A., CORBET, A.S. AND WILLIAMS, C.B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology* **12**, 42–58.

FLORENCIO, D. AND HERLEY, C. (2007). A large-scale study of web password habits. *Proceedings of the International Conference on World Wide Web*, 657–666.

FRANSSEN, S.E.M.P., VAN DER VAART, A.W. (2022). Empirical and Full Bayes estimation of the type of a Pitman-Yor process. *arXiv preprint* arxiv:2208:14255.

GALE, W.A. AND SAMPSON, G. (1995). Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics* **2**, 217–237.

GAO, F. (2013). Moderate deviations for a nonparametric estimator of sample coverage. *The Annals of Statistics* **41**, 641-669.

GAO, Z., TSENG, C.H., PEI, Z. AN BLASER, M.J. (2007). Molecular analysis of human forearm superficial skin bacterial biota. *Proceedings of the National Academy of Sciences of USA* **104**, 2927–2932.

GNEDIN, A., HANSEN, B. AND PITMAN, J. (2007). Notes on the occupancy problems with infinitely many boxes: general asymptotics and power law. *Probability Surveys* **4**, 146–171.

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

GOODMAN, L.A.(1949). On the estimation of the number of classes in a population. *The Annals of Mathematical Statistics* **20**, 572-579.

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.

GRABCHAK, M. AND ZHANG, Z. (2017). Asymptotic properties of Turing's formula in relative error. *Machine Learning* **106**, 1771–1785.

GRAVEL, S. (2014). Predicting discovery rates of genomic features. *Genetics* **197**, 601–610.

GRIFFIN, J. AND LEISEN, F. (2017). Compound random measures and their use in Bayesian nonparametrics. *Journal of the Royal Statistical Society Series B* **79**, 525–545.

HAAS, P.J., NAUGHTON, J.F., SESHADRI, S. AND STOKES, L. (1995). Sampling-based estimation of the number of distinct values of an attribute. *Proceedings of the Very Large Data Bases Conference*, 311–322.

HANSE, B. AND PITMAN, J. (2000). Prediction rules for exchangeable sequences related to species sampling. *Statistics and Probability Letters* **46**, 251–256.

HAO, Y. AND LI, P. (2020). Bessel smoothing and multi-distribution property estimation. In *Proceedings of the Conference on Computational Learning Theory*, **125**, 1817–1876

HAO, Y. AND LI, P. (2020). Optimal prediction of the number of unseen species with multiplicity. In *Advances in Neural Information Processing Systems*, **33**.

HAO, Y., ORLITSKY, A. AND PICHAPATI, V. (2018). On learning Markov chains. In *Advances in Neural Information Processing Systems*, **32**

HAO, Y. AND ORLITSKY, A. (2020). Profile entropy: a fundamental measure for the learnability and compressibility of distributions. In *Advances in Neural Information Processing Systems*, **34**.

HOFERT, M. (2011). Sampling exponentially tilted stable distributions. *ACM Transactions on Modeling and Computer Simulation* **31**, 3.

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

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 of USA* **106**, 5008–5013.

JEGANATHAN, P. AND HOLMES, S.P. (2021). A statistical perspective on the challenges in molecular microbial biology *Journal of Agricultural, Biological and Environmental Statistics* **26**, 131–160.

KINGMAN, J.F.C. (1978). The representation of partition structure. *Journal of the London Mathematical Society* **18**, 374–380.

KORWAR, R.M. AND HOLLANDER, M. (1973). Contributions to the theory of Dirichlet processes. *The Annals of Probability* **1**, 705–711.

KROES, I., LEPP, P.W. AND RELMAN, D.A. (1999). Bacterial diversity within the human subgingival crevice. *Proceeding of the National Academy of Sciences of USA* **96**, 14547–14552.

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

JOHNSON, N.L., KEMP, A.W. AND KOTZ, S. (2005) *Univariate discrete distributions*, Wiley Series in Probability and Statistics.

LAIRD, N.M. AND LANGE, C. (2010). The fundamentals of modern statistical genetics. *Springer*.

LIJOI, A., MENA, R.H. AND PRÜNSTER, I. (2007). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika* **94**, 769–786.

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

MAGURRAN, A.E. (2003). Measuring biological diversity. *Wiley*.

MAO, C.X. (2004). Prediction of the conditional probability of discovering a new class. *Journal of the American Statistical Association* **99**, 1108–1118.

MAO, C.X. AND LINDSAY, B.G. (2004). A Poisson model for the coverage problem with a genomic application. *Biometrika* **89**, 669–682.

MASOERO, L., BERAHA, M., RICHARDSON, T.S. AND FAVARO, S. (2023). Improved prediction of future user activity in online A/B testing. *Preprint arXiv:2402.03231*.

MASOERO, L., CAMERLENGHI, F., FAVARO, S. AND BRODERICK, T. (2020). More for less: predicting and maximizing genetic variant discovery via Bayesian nonparametrics. *Biometrika* **109**, 17–32.

MCALLESTER, D. AND ORTIZ, L. (2003). Concentration inequalities for the missing mass and for histogram rule error. *Journal of Machine Learning Research* **4**, 895–911.

MCALLESTER, D. AND SCHAPIRE, R.E. (2000). On the convergence rate of Good-Turing estimators. *Proceedings of the Conference on Computational Learning Theory*, 1–6.

MOSSEL, E. AND OHANNESSIAN, M.I. (2019). On the impossibility of learning the missing mass. *Entropy* **21**, 28.

MOTWANI, S. AND VASSILVITSKII, S. (2006) Distinct value estimators in power law distributions. In *Proceedings of the Workshop on Analytic Algorithms and Combinatorics*, 1–8.

NACU, S. (2006). Increments of random partitions. *Combinatorics, Probability and Computing* **15**, 589–595.

OHANNESSIAN, M.I. AND DAHLEH, M.A. (2012). Rare probability estimation under regularly varying heavy tails. In *Proceedings of the Conference on Learning Theory*, **23**, 2110–2124.

ORLITSKY, A., SANTHANAM, N.P. AND ZHANG, J. (2003). Always Good-Turing: asymptotically optimal probability estimation. *Science* **302**, 427-431.

ORLITSKY, A., SANTHANAM, N.P. AND ZHANG, J. (2004). Universal compression of memoryless sources over unknown alphabets. *IEEE Transaction on Information Theory* **50**, 1469–1481.

ORLITSKY, A., SURESH, A.T. AND WU, Y. (2017). Optimal prediction of the number of unseen species. *Proceeding of the National Academy of Sciences of USA* **113**, 13283–13288.

PANANJADY, A., MUTHUKUMAR, V. AND THANGARAJ, V. (2024). Just wing it: optimal estimation of missing mass in a Markovian sequence. *Preprint arXiv:2404.05819*

PANDE, V.S., BEAUCHAMP, K. AND BOWMAN, G.R. (2010). Everything you wanted to know about Markov state models but were afraid to ask. *Methods* **52**, 99–105

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.

PITMAN, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields* **102**, 145–158.

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.

PITMAN, J. (2003). Poisson-Kingman partitions. In *Science and Statistics: A Festschrift for Terry Speed*, Goldstein, D.R. Eds. Institute of Mathematical Statistics.

PITMAN, J. (2006). *Combinatorial stochastic processes*. Lecture Notes in Mathematics, Springer Verlag.

PITMAN, J. AND YOR, M. (1997). The two parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability* **25**, 855–900.

POLYANSKIY, Y. AND WU, Y. (2020). Dualizing Le Cam's method for functional estimation, with applications to estimating the unseens. *Preprint arXiv:1902.05616*.

RAGHUNATHAN, A., VALIANT, G. AND ZOU, J. (2017). Estimating the unseen from multiple populations. In *Proceedings of the International Conference on Machine Learning*, **70**. 2855–2863

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.

ROBINS, H.S., CAMPREGHER, P.V., SRIVASTAVA, S.K., WACHER, A., TURTLE, C.J., KAHSAI, O., RIDDELL, S.R., WARREN, E.H. AND CARLSON, C.S. (2009). Comprehensive assessment of T-cell receptor β-chain diversity in $\alpha\beta$ T cells. *Blood* **114**, 4099–4107.

ROBBINS, H.E. (1956). An empirical Bayes approach to statistics. *Proceedings of the Berkeley Symposium* **1**, 157–163.

ROBBINS, H.E. (1964). The empirical Bayes approach to statistical decision problems. *The Annals of Mathematical Statistics* **35**, 1–20.

ROBBINS, H.E. (1968). Estimating the total probability of the unobserved outcomes of an experiment. *The Annals of Mathematical Statistics* **39**, 256–257.

SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639–650.

SIMS, D., SUDBERY, I., ILOTT, N.,HEGER, A. AND PONTING, C. (2014). Sequencing depth and coverage: key considerations in genomic analysis. *Nature Review Genetics* **15**, 121–132.

SKINNER, AND ELLIOT, M.J. (2002). A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society Series B* **64**, 855–867.

SKINNER, C.J., MARSH, C., OPENSHAW, S. AND WYMER, C. (1994). Disclosure control for census microdata. *Journal of Official Statistics* **10**, 31–51.

SKORSKI, M. (2020). Revisiting concentration of missing mass. *Preprint arXiv:2005.10018*.

SKORSKI, M. (2020). Missing mass in Markov chains. *Preprint arXiv:2001.03603*.

SKORSKI, M. (2021). Bernstein-type bounds for Beta distribution. *Preprint arXiv:2101.02094*.

TEH, Y.W., JORDAN, M.I., BEAL, M.J. AND BLEI, D.M. (2008). Hierarchical Dirichlet processes. *Journal of the American Statistical Association* **101**, 1566–1581.

THISTED, R. AND EFRON, B. (1987). Did Shakespeare write a newly-discovered poem? *Biometrika* **74**, 445–455.

THOMPSON, W.K. (2003). Sampling rare or elusive species. *Island Press*.

ZABELL, S.L. (1992). Predicting the unpredictable. *Synthese* **90**, 205–232.

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. University of Pittsburgh Press.

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.

ZHANG, C.H. (2005). Estimation of sums of random variables: examples and information bound. *The Annals of Statistics* **33**, 2022-2041.

ZHANG, C.H. AND ZHANG, Z. (2009). Asymptotic normality of a nonparametric estimator of sample coverage. *The Annals of Statistics* **37**, 2582-2595.

WILLENBORG, L. AND DE WAAL, T. (2001). *Elements of statistical disclosure control.* Springer, New York.

WOLFER, G. AND KONTOROVICH, A. (2019). *Minimax learning of ergodic Markov chains*. In *Proceedings of the International Conference on Algorithmic Learning Theory* **98**, 904–930.

WU, Y. AND YANG, P. (2016). Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transaction on Information Theory* **62**, 3702–3720.

WU, Y. AND YANG, P. (2019). Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *The Annals of Statistics* **47**, 857–883.

WU, Y. AND YANG, P. (2021). Polynomial methods in statistical inference: theory and practice. *Foundations and Trends in Communications and Information Theory* **17**, 402–586.

ZOU, J., VALIANT, G., VALIANT, P., KARCZEWSKI, K., CHAN, S.O., SAMOCHA, K., LEK, M., SUNYAEV, S., DALY, M. AND MACARTHUR, D.G. (2016). Quantifying unobserved protein-coding variants in human populations provides a roadmap for large-scale sequencing projects. *Nature Communications* **7**, 13293.