Bayesian Modelling of Sequential Discoveries

Alessandro Zito[1], Tommaso Rigon[2], Otso Ovaskainen[3,4,5], and David B. Dunson[1]

[1]Department of Statistical Science, Duke University, Durham, N.C., U.S.A
[2]Department of Economics, Management and Statistics, University of Milano–Bicocca, Milan, Italy
[3]Department of Biological and Environmental Science, University of Jyväskylä, Jyväskylä, Finland
[4]Organismal and Evolutionary Biology Research Programme, University of Helsinki, Helsinki, Finland
[5]Centre for Biodiversity Dynamics, Department of Biology, Norwegian University of Science and Technology, Trondheim, Norway

March 28, 2022

Abstract

We aim at modelling the appearance of distinct tags in a sequence of labelled objects. Common examples of this type of data include words in a corpus or distinct species in a sample. These sequential discoveries are often summarised via accumulation curves, which count the number of distinct entities observed in an increasingly large set of objects. We propose a novel Bayesian method for species sampling modelling by directly specifying the probability of a new discovery, therefore allowing for flexible specifications. The asymptotic behavior and finite sample properties of such an approach are extensively studied. Interestingly, our enlarged class of sequential processes includes highly tractable special cases. We present a subclass of models characterized by appealing theoretical and computational properties, including one that shares the same discovery probability with the Dirichlet process. Moreover, due to strong connections with logistic regression models, the latter subclass can naturally account for covariates. We finally test our proposal on both synthetic and real data, with special emphasis on a large fungal biodiversity study in Finland.

Keywords: Accumulation curves; Dirichlet process; Logistic regression; Poisson-binomial distribution; Species sampling models.

1 Introduction

Our goal is to develop a flexible procedure for modelling the appearance of previously unobserved objects in a sequence. The sequential recording of distinct entities can be represented through an *accumulation curve*, namely the cumulative number of distinct entities K_n within a collection of n objects (Christen and Nakamura, 2000; Gotelli and Colwell, 2001). These entities can be of various nature, including biological species (Good, 1953; Good and Toulmin, 1956), words (Efron and Thisted, 1976; Thisted and Efron, 1987), genes (Ionita-Laza et al., 2009), bacteria (Hughes et al., 2001; Gao et al., 2007) and cell types (Camerlenghi et al., 2020). The analysis of accumulation curves has a rich history in statistics, as testified by the early contributions of Fisher et al. (1943), Good (1953), and Good and Toulmin (1956). We refer to Bunge and Fitzpatrick (1993); Gotelli and Colwell (2001) for a historical account. Several nonparametric approaches have been developed, aiming at i) predicting the number of unseen entities (e.g. Shen et al., 2003), or ii) estimating the probability of a new discovery (e.g. Chao and Shen, 2004; Mao, 2004; Favaro et al., 2012). Similar tasks have also been dealt with in parametric ways (e.g. Arrhenius, 1921; Soberon and Llorente, 1993; Flather, 1996; Diaz-Frances and Gorostiza, 2002).

Our work is inspired by the class of Bayesian nonparametric methods called *species sampling models* (Pitman, 1996). In one of our motivating applications, we aim to assess how many of the species present in a sample are missed when a given number of DNA barcode sequences are obtained. Let $(X_n)_{n\geq 1}$ be a sequence of objects, such as fungal DNA sequences in a single soil or air sample (Abrego et al., 2020), taking values in $\mathbb{X}$, which is the space of fungal species. Among the first n observed objects $X_1, \ldots, X_n$, there will be $K_n \leq n$ distinct entities, or species, representing the nth value of the accumulation curve. The values $(X_n)_{n\geq 1}$ are randomly generated in a sequential manner, so that the tag X_{n+1} is either new or equal to one of the previously observed objects. For instance, in the Dirichlet process case, the sequential allocation

mechanism for any $n \geq 1$ proceeds as:

$$(X_{n+1} \mid X_1, \ldots, X_n) = \begin{cases} \text{“new”}, & \text{with probability} \quad \alpha/(\alpha+n), \\ X_i, & \text{with probability} \quad 1/(\alpha+n), \quad i = 1, \ldots, n, \end{cases} \tag{1}$$

where $\alpha > 0$ controls the rate of new discoveries; see also Blackwell and MacQueen (1973). We refer to the quantity $\alpha/(\alpha+n)$ as the *discovery probability* for the Dirichlet process.

The predictive scheme in (1) is restrictive in depending on a single parameter and in inducing a logarithmic growth for the accumulation curve $(K_n)_{n\geq 1}$. These limitations motivated the development of random processes with more flexible growth rates. Notorious examples include the two parameter Poisson–Dirichlet process of Perman et al. (1992), often called the Pitman–Yor process when the number of species is assumed to be infinite or the Dirichlet-multinomial process in the finite case (Pitman and Yor, 1997), and the general class of Gibbs-type priors (Gnedin and Pitman, 2005). Under these models, the labels $(X_n)_{n\geq 1}$ are *exchangeable*, meaning their order of appearance is irrelevant for inferential purposes. While convenient, exchangeability can be restrictive to obtain (Lee et al., 2013). For this reason, generalizations of species sampling models that go beyond exchangeablility have been proposed (Berti et al., 2004; Bassetti et al., 2010; Fortini et al., 2018; Cassese et al., 2019; Ascolani et al., 2021), often admitting (1) as a special case. One flexible model is the BETA-GOS process (Airoldi et al., 2014), where the allocation probabilities are functions of independent beta random variables.

Bayesian species sampling models induce a distribution for K_n at every n, which arises from a pure-birth inhomogeneous Markov process governed by the discovery probabilities. As such, they are naturally endowed with in- and out-of-sample estimators for the accumulation curve, $\mathbb{E}(K_n)$ and $\mathbb{E}(K_{n+m} \mid K_n = k)$, $m \geq 1$. In line with the ecological literature (e.g. Gotelli and Colwell, 2001), we refer to these as model-based *rarefaction* and *extrapolation* estimators, respectively. For Pitman–Yor and general Gibbs-type priors, extrapolations are available in closed form (Lijoi et al., 2007; Favaro et al., 2009). However, such models are often too restrictive, as is evident

from Figure 1, which shows in- and out-of-sample performance in estimating the number of distinct fungi species in a given number of fungal DNA-barcode sequences[1]. The Dirichlet process performs poorly in sample, while the Pitman–Yor has good in-sample fit but inadequate out-of-sample predictive accuracy. This is not surprising, as the Pitman–Yor process depends on only two parameters and assumes that $K_n \to \infty$ almost surely as $n \to \infty$. As there are finitely many fungi species, K_n should more realistically converge to a finite constant. Such is the case for the Dirichlet-multinomial process, for which $\lim_{n\to\infty} K_n = K_\infty$. However, its trajectory has similar lack of fit as the Dirichlet process. The BETA-GOS process admits both $K_\infty = \infty$ and $K_\infty < \infty$ depending on the values of its parameters. Nonetheless, it often shows similar out-of-sample behavior as the Pitman-Yor process. Potentially one could use a predictive scheme that is more flexible than the Pitman–Yor, while also allowing finite K_∞; recent examples include Camerlenghi et al. (2018); Lijoi et al. (2020). However, such specifications involve cumbersome combinatorial structures in the sampling mechanism, effectively preventing their application in the types of large datasets that are routinely collected in our motivating application areas. For example, in fungi biodiversity studies, it is common to obtain DNA barcodes for millions of sequences from 10,000s of species (e.g. Ovaskainen et al., 2020).

We address the above limitations through a novel modelling framework, which is highly flexible, analytically tractable, and computationally efficient. The key distinction compared to species sampling models, such as (1), is that we directly specify a model for the accumulation curve $(K_n)_{n\geq 1}$, whereas the tags $(X_n)_{n\geq 1}$ are regarded as nuisance parameters. Specifically, we consider a collection of Bernoulli random variables $(D_n)_{n\geq 1}$ representing whether at the $(n+1)$th step a new entity has been discovered or not, namely

$$\mathbb{P}(D_{n+1} = 1) = \mathbb{P}(X_{n+1} = \text{“new”} \mid X_1, \ldots, X_n), \qquad n \geq 1,$$

[1]Species are defined in this article based on genetic sequences being sufficiently distinct, but the terminology used by ecologists is “operational taxonomic units” as determining species requires additional verification.

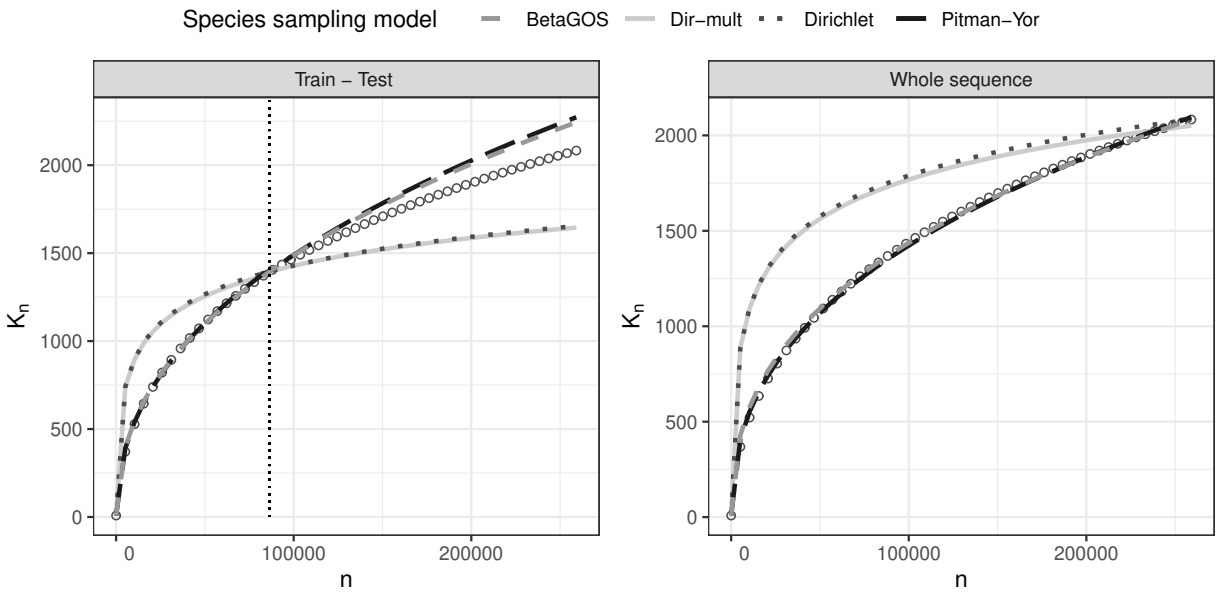

Figure 1: Empirical and estimated accumulation curve in one air fungal DNA-barcoding sample from Finland. White dots indicate observed values. Left panel: the vertical line is the training-test set cutoff, set to 1/3 of the total number of genetic sequences. The parameters of the Dirichlet, Pitman–Yor and Dirichlet-multinomial are estimated on the training set via empirical Bayes, while estimation for BETAGOS relies on method of moments. Right panel: the curves are estimated using the full data. See the Supplementary Material for further details on model parametrizations.

having set $D_1 = 1$. The accumulation curve is obtained by summing over these binary indicators: $K_n = \sum_{i=1}^{n} D_i, n \geq 1$. Differently from general species sampling models, in our framework, the Bernoulli indicators $(D_n)_{n\geq 1}$ are assumed to be *independent*, albeit not identically distributed. Hence, we aim at developing suitable formulations for the probabilities $(\pi_n)_{n\geq 1}$, with $\pi_n = \mathbb{P}(D_n = 1)$, for any $n \geq 1$. It is natural to require these probabilities to be decreasing over n, so that the discovery of a new entity is increasingly difficult the more data we collect. Moreover, $\pi_1 = \mathbb{P}(D_1 = 1) = 1$, since the first entity of the sequence is necessarily new. Both requirements are satisfied by the Dirichlet process, where $\pi_n = \alpha/(\alpha + n - 1)$. We propose a general strategy for the specification of $(\pi_n)_{n\geq 1}$, relying on the notion of survival functions, and study the impact of specific choices on the asymptotic behavior of K_n.

A specific subclass of our framework is particularly appealing in terms of analytic and computational simplicity, due to connections with logistic regression. This subclass includes the Dirichlet process and naturally leads to covariate-dependent extensions. Existing covariate-dependent species sampling models are typically complex to implement; refer to Quintana et al. (2022) for an overview. In contrast, our approach simply involves implementing a constrained logistic regression. We illustrate the flexibility and computational tractability through application to data on copepod and fungi biodiversity.

The paper is organized as follows. Sections 2-3 introduce our modeling framework, investigate the theoretical properties and describe a subclass of models connected with logistic regression. Inferential strategies together with a solution to order dependence are presented in Section 4. In Section 5 we test our model on simulated scenarios. Section 6 details the applications to real datasets. Concluding remarks are given in Section 7.

2 A general modelling framework for accumulation curves

2.1 Background on species sampling models

In this Section we review key concepts about species sampling models that will be used throughout the paper. For a broader overview, refer to Pitman (1996) and De Blasi et al. (2015). For generalizations that go beyond exchangeability, see Berti et al. (2021).

Let $(X_n)_{n\geq 1}$ be a sequence of objects. Given the discrete nature of the data, there will be ties among $X_1, \ldots, X_n$, comprising a total of $K_n = k$ distinct entities $X_1^*, \ldots, X_k^*$, having frequencies $n_1, \ldots, n_k$, with $\sum_{j=1}^k n_j = n$. Frequencies $n_1, \ldots, n_k$ are referred to as *abundances* in the ecological literature (Gotelli and Colwell, 2001). One generalization of the sequential

allocation scheme of the Dirichlet process in (1) is given by

$$(X_{n+1} \mid X_1, \ldots, X_n) = \begin{cases} \text{“new”}, & \text{with probability} \quad \pi_{n+1}, \\ X_i, & \text{with probability} \quad p_{i,n+1}, \quad i = 1, \ldots, n, \end{cases} \tag{2}$$

for $n \geq 1$, suitable probabilities $\sum_{i=1}^{n} p_{i,n+1} = 1 - \pi_{n+1}$ and $X_1 =$ “new”. For Gibbs-type processes (Gnedin and Pitman, 2005), π_{n+1} and $p_{i,n+1}$ depend on previous values only through k and the frequencies $n_1, \ldots, n_k$, respectively. One example is the Pitman–Yor process, where $\pi_{n+1} = (\alpha + \sigma k)/(\alpha + n)$, $p_{i,n+1} = (1 - \sigma \bar{n}_i^{-1})/(\alpha + n)$, for $i = 1, \ldots, n$, $\sigma \in [0, 1)$ and $\alpha > -\sigma$, where $\bar{n}_i$ is the frequency of the associated tag X_i within the sample; the Dirichlet process is recovered with $\sigma = 0$. Another is the Dirichlet-multinomial, which has the same sampling scheme of the Pitman–Yor but with $\sigma < 0$ and $\alpha = H|\sigma|$, with $H \in \mathbb{N}$ the total number of species. For the above examples, the law of $(X_n)_{n \geq 1}$ is *exchangeable*, i.e. invariant to reordering of the sequence, requiring strict conditions on $p_{i,n+1}$ and π_{n+1} (Lee et al., 2013).

To relax exchangeability while maintaining certain desirable properties, Berti et al. (2004) proposed *conditionally identically distributed* (CID) sequences. For CID sequences, the labels X_{n+m} are identically distributed conditioned on $X_1, \ldots, X_n$ for $n, m \geq 1$. Examples include *generalized Poisson–Dirichlet* and *generalized Ottawa sequences* (GOS) (Bassetti et al., 2010), and GOS sequences with latent beta reinforcements (BETA-GOS) (Airoldi et al., 2014). For BETA-GOS, the random allocation probabilities are $\pi_{n+1} = \prod_{i=1}^{n} W_i$ and $p_{i,n+1} = (1 - W_i) \prod_{j=i+1}^{n} W_j$, where $W_n \sim \text{BETA}(a_n, b_n)$ are independent beta random variables for $n \geq 1$. As we describe in Section 5, the values for a_n and b_n determine the asymptotic behavior of the sequence. The sequential mechanism in (2) induces a law for the accumulation curve $(K_n)_{n \geq 1}$.

Let $K_m^{(n)}$ denote the number of new entities in a future sample of size m conditioning on training data $X_1, \ldots, X_n$. Under a Dirichlet process, both the prior mean for the accumulation curve K_m and the posterior mean for $K_m^{(n)}$ have simple expressions:

$$\mathbb{E}(K_n) = \sum_{i=1}^{n} \frac{\alpha}{\alpha + i - 1}, \quad \mathbb{E}(K_m^{(n)} \mid X_1, \ldots, X_n) = \sum_{i=1}^{m} \frac{\alpha}{\alpha + n + i - 1}. \tag{3}$$

The Dirichlet process is the only exchangeable species sampling model for which such a simplification occurs (Lijoi et al., 2007). For BETA-GOS priors, the prior expected accumulation curve also has a simple form: $\mathbb{E}(K_n) = 1 + \sum_{i=1}^{n-1} \prod_{j=1}^{i} a_j (a_j + b_j)^{-1}$, $n \geq 2$. However, beyond the Dirichlet process, the posterior expectation of $K_m^{(n)}$ is typically complex.

2.2 The model

In species sampling models, the distribution of the accumulation curve $(K_n)_{n\geq 1}$ is essentially a byproduct of the specification for the values $(X_n)_{n\geq 1}$. We propose a more direct formulation for $(K_n)_{n\geq 1}$ which avoids modelling of the sequence $(X_n)_{n\geq 1}$.

Let $(D_n)_{n\geq 1}$ be a collection of *independent* binary indicators, denoting the discoveries, with probabilities $(\pi_n)_{n\geq 1}$. Moreover, let $K_n = \sum_{i=1}^{n} D_i$ for any $n \geq 1$ be the accumulation curve. By being the sum of independent but not necessarily identically distributed Bernoulli trials, K_n follows a Poisson-binomial distribution with parameters $\pi_1, \ldots, \pi_n$. We denote it as $K_n \sim \text{PB}(\pi_1, \ldots, \pi_n)$. The Poisson-binomial, often denoted as the Pólya frequency distribution or as a convolution of heterogeneous Bernoulli, has been extensively studied in the literature, with early contributions from Le Cam (1960); Hoeffding (1956) and Darroch (1964). See also Gleser (1975); Pitman (1997); Xu and Balakrishnan (2011). When the probabilities $(\pi_n)_{n\geq 1}$ are all equal, K_n has a binomial distribution. In our setting, $\pi_n > \pi_{n+1}$ for every $n \geq 1$ with $\pi_1 = 1$. In addition, $\lim_{n\to\infty} \pi_n = 0$, so the probability of making a new discovery eventually approaches zero. A general strategy for constructing such a set of probabilities is described as follows.

Definition 1. Let T be a random variable on $(0, \infty)$ with strictly increasing cumulative distribution function $F(t; \boldsymbol{\theta})$ indexed by $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$. Moreover, let $S(t; \boldsymbol{\theta}) = 1 - F(t; \boldsymbol{\theta})$ be its

survival function. The set of probabilities $(\pi_n)_{n\geq 1}$ are said to be directed by $S(t;\boldsymbol{\theta})$ if

$$\pi_n = \mathbb{P}(T_n > n-1) = S(n-1;\boldsymbol{\theta}), \qquad \text{for any } n \geq 1, \tag{4}$$

where $(T_n)_{n\geq 1}$ are independent and identically distributed random variables following $F(t;\boldsymbol{\theta})$.

It is easy to check that a set of probabilities $(\pi_n)_{n\geq 1}$ directed by $S(t;\boldsymbol{\theta})$ satisfies the aforementioned requirements. Indeed, one has that $\pi_1 = S(0;\boldsymbol{\theta}) = 1$ for any $\boldsymbol{\theta} \in \Theta$, since T is supported on $(0,\infty)$. Moreover, $\pi_n = S(n-1;\boldsymbol{\theta}) > S(n;\boldsymbol{\theta}) = \pi_{n+1}$, because by assumption $S(t;\boldsymbol{\theta})$ is strictly decreasing. Furthermore, one has that $\lim_{n\to\infty} \pi_n = \lim_{n\to\infty} S(n-1;\boldsymbol{\theta}) = 0$, as desired, since $S(t;\boldsymbol{\theta})$ is a survival function. Each binary random variable D_n may be represented as $D_n = \mathbb{1}(T_n > n-1)$, with $\mathbb{1}(\cdot)$ denoting the indicator function.

The discovery indicators can be alternatively viewed as the difference of two consecutive points in the curve, namely $D_n = K_n - K_{n-1}$ for any $n \geq 2$ with $D_1 = 1$. Hence, the discoveries $(D_n)_{n\geq 1}$ and the accumulation curve $(K_n)_{n\geq 1}$ carry the same information, having a one-to-one relationship. Then, if the probabilities $(\pi_n)_{n\geq 1}$ are directed by $S(t;\boldsymbol{\theta})$, inferential statements about the parameter vector $\boldsymbol{\theta} \in \Theta$ can be based on the likelihood function $\mathscr{L}(\boldsymbol{\theta} \mid D_1,\ldots,D_n)$ or, equivalently, on $\mathscr{L}(\boldsymbol{\theta} \mid K_1,\ldots,K_n)$. The former is readily available as

$$\mathscr{L}(\boldsymbol{\theta} \mid D_1,\ldots,D_n) \propto \prod_{i=2}^{n} S(i-1;\boldsymbol{\theta})^{D_i}\{1 - S(i-1;\boldsymbol{\theta})\}^{1-D_i}, \tag{5}$$

having excluded the degenerate term $D_1 = 1$. A similar one-to-one relationship between $D_1,\ldots D_n$ and the set of labels $X_1,\ldots,X_n$ is generally not true in species sampling models and their generalizations; $\mathscr{L}(\boldsymbol{\theta} \mid X_1,\ldots,X_n)$ can be more informative than $\mathscr{L}(\boldsymbol{\theta} \mid D_1,\ldots,D_n)$ in such cases. A notable exception (see below Theorem) is the Dirichlet process, where $\mathscr{L}(\boldsymbol{\theta} \mid D_1,\ldots,D_n)$ is retrieved in our setting by assuming $S(t;\boldsymbol{\theta}) = \alpha/(\alpha+t)$ with $\boldsymbol{\theta} = \alpha > 0$.

Theorem 1. *Let $(X_n)_{n\geq 1}$ be a sequence of objects directed by a Dirichlet process as in (1) and let $(D_n)_{n\geq 1}$ be the associated discovery indicators. Then for a sample $X_1,\ldots,X_n$ with*

$K_n = k$ distinct values one has $\mathscr{L}(\alpha \mid D_1, \ldots, D_n) \propto \mathscr{L}(\alpha \mid X_1, \ldots, X_n) \propto \alpha^k/(\alpha)_n$, with $(a)_n = a(a+1)\cdots(a+n-1)$ denoting the Pochhammer symbol, for any $a > 0$ and $n \geq 1$.

Hence, it is equivalent to base inferences on the Dirichlet process parameter α on the likelihood (5) for the discovery indicators instead of the usual likelihood for $X_1, \ldots, X_n$. This occurs because $K_n = \sum_{i=1}^n D_i$ is the minimal sufficient statistic for α in the Dirichlet process; see Lijoi et al. (2007) for similar considerations. An implication is that the empirical Bayes estimate of α, obtained by maximizing $\alpha^k/(\alpha)_n$, coincides with the maximizer of (5).

Remark 1. If a sequence of discoveries $(D)_{n\geq 1}$ is directed by $S(t;\boldsymbol{\theta})$, the general predictive scheme in equation (2) may be specified as

$$(X_{n+1} \mid X_1, \ldots, X_n) = \begin{cases} \text{“new”}, & \text{with probability} \quad S(n;\boldsymbol{\theta}), \\ X_i, & \text{with probability} \quad p_i(n;\boldsymbol{\theta}), \quad i = 1, \ldots, n, \end{cases}$$

with $\sum_{i=1}^n p_i(n;\boldsymbol{\theta}) = 1 - S(n;\boldsymbol{\theta}) = F(n,\boldsymbol{\theta})$. As long as probabilities $p_i(n;\boldsymbol{\theta})$ sum to the cumulative distribution function of T, any choice for their functional form is valid. Hence, the function $S(t;\boldsymbol{\theta})$ does not uniquely identify a sampling model for $X_1, \ldots, X_n$. Careful choices of $S(t;\boldsymbol{\theta})$ and $p_i(n;\boldsymbol{\theta})$ can lead to exchangeability (Lee et al., 2013) or conditional identity in distribution (Berti et al., 2004). For example, when $S(n;\boldsymbol{\theta}) = \alpha(\alpha + n^{1-\sigma})^{-1}$, with $\sigma \in [0,1)$ and $\alpha > 0$, letting $p_i(n;\boldsymbol{\theta}) = (i^{1-\sigma} - (i-1)^{1-\sigma})/(\alpha + n^{1-\sigma})$ generates a CID sequence in the family of generalized Ottawa sequences (Bassetti et al., 2010). However, the resulting likelihood function lacks a simple analytical form. Given our focus on the sequence of discoveries, we focus on likelihood (5), treating the labels $(X_n)_{n\geq 1}$ as nuisance parameters.

2.3 Smoothing, prediction and posterior representations

In this Section, we present prior and posterior properties of K_n, which may be useful for both smoothing and prediction. Supposing $(\pi_n)_{n\geq 1}$ is directed by $S(t;\boldsymbol{\theta})$, it immediately follows that

$K_n \sim \text{PB}\{1, S(1;\boldsymbol{\theta}), \ldots, S(n-1;\boldsymbol{\theta})\}$. The probability mass function $\mathbb{P}(K_n = k)$ of the Poisson-binomial is cumbersome to evaluate, especially for large n and large k; certain choices of $S(t;\boldsymbol{\theta})$ greatly simplify $\mathbb{P}(K_n = k)$, as we clarify in Section 3.2.

However, moments are easily specified, with prior mean and variance equal to

$$\mathbb{E}(K_n) = \sum_{i=1}^{n} S(i-1;\boldsymbol{\theta}), \quad \text{var}(K_n) = \sum_{i=1}^{n} S(i-1;\boldsymbol{\theta})\{1 - S(i-1;\boldsymbol{\theta})\}, \qquad n \geq 1.$$

These formulas may be useful in choosing the parametric form of $S(t;\boldsymbol{\theta})$ and for prior elicitation for $\boldsymbol{\theta}$. We refer to $\mathbb{E}(K_n) = \sum_{i=1}^{n} \mathbb{P}(D_i = 1)$ as the *rarefaction* estimator for the accumulation curve; this amounts to smoothing of the $K_1, \ldots, K_n$ values observed in the training samples. This expectation does not depend on the ordering of the data, at least for any fixed value of $\boldsymbol{\theta}$.

Similar considerations can be made for *extrapolation*. Suppose we are given a sample of $D_1, \ldots, D_n$ discoveries displaying $K_n = k$ distinct entities and that we are interested in predicting future values of the accumulation curve $K_{n+1}, \ldots, K_{n+m}$ or in predicting the number of new entities within a future sample of size m, $K_m^{(n)} = K_{n+m} - K_n = \sum_{i=n+1}^{n+m} D_i$. The posterior distribution of $(K_m^{(n)} \mid D_1, \ldots, D_n)$ is available in closed form, namely

$$(K_m^{(n)} \mid D_1, \ldots, D_n) \sim \text{PB}\{S(n;\boldsymbol{\theta}), \ldots, S(n+m-1;\boldsymbol{\theta})\}.$$

Hence, $\mathbb{E}(K_m^{(n)} \mid D_1, \ldots, D_n) = \sum_{i=n+1}^{n+m} \mathbb{P}(D_i = 1) = \sum_{j=1}^{m} S(j+n-1;\boldsymbol{\theta})$, so the posterior distribution of $K_m^{(n)}$ given the discoveries $D_1, \ldots, D_n$ is conjugate, being a Poisson-binomial with updated parameters. The distribution of $(K_{n+m} \mid D_1, \ldots, D_n) = K + K_m^{(n)}$ is then a shifted Poisson-binomial, and we have the out-of-sample extrapolation estimator as

$$\mathbb{E}(K_{n+m} \mid D_1, \ldots, D_n) = k + \mathbb{E}(K_m^{(n)} \mid D_1, \ldots, D_n) = k + \sum_{j=1}^{m} S(j+n-1;\boldsymbol{\theta}),$$

which can be interpreted as the sum of discovery probabilities.

2.4 Asymptotic behavior of K_n

The limit of K_n as $n \to \infty$ is often of inferential interest, representing the random number of entities one would eventually discover. Depending on the choice of $S(t;\boldsymbol{\theta})$, two scenarios can occur: i) the number of distinct entities diverges, as in the Dirichlet process case, so that $K_n \to \infty$ almost surely as $n \to \infty$. In this regime, it is useful to study the growth rate of K_n. Alternatively, we could find that ii) the number of distinct species converges to some non-degenerate random variable $K_n \to K_\infty$, almost surely, as $n \to \infty$. Within ecology the random variable K_∞ is called the *species richness* (e.g. Colwell, 2009).

The asymptotic behaviour of K_n is controlled by the structure of the chosen survival function $S(t;\boldsymbol{\theta})$. Before stating our first result, let us define $\mathbb{E}(T) = \int_0^\infty \mathbb{P}(T > t)\mathrm{d}t = \int_0^\infty S(t;\boldsymbol{\theta})\mathrm{d}t$, that is, the expectation of the latent variables in Definition 1.

Proposition 1. *Let* $K_n \sim \text{PB}\{1, S(1;\boldsymbol{\theta}), \ldots, S(n-1;\boldsymbol{\theta})\}$. *Then, there exists a possibly infinite random variable* K_∞ *such that* $\lim_{n\to\infty} K_n \to K_\infty$, *almost surely, with* $\mathbb{E}(K_\infty) = \sum_{i=0}^\infty S(i;\boldsymbol{\theta})$. *Moreover,*

$$\mathbb{E}(T) \leq \mathbb{E}(K_\infty) \leq \mathbb{E}(T) + 1. \tag{6}$$

Equation (6) provides lower and upper bounds for the asymptotic mean, which can be used to summarize the species richness. The expected value of $\mathbb{E}(T)$ represents a simple tool to determine whether the accumulation curve diverges or not, as the following clarifies.

Corollary 1. *Under the conditions of Proposition 1,* $K_\infty = \infty$ *almost surely if and only if* $\mathbb{E}(T) = \infty$.

Let us consider the first asymptotic regime, corresponding to the $K_\infty = \infty$ case. In this case, the rate of growth is controlled by $S(t;\boldsymbol{\theta})$, as clarified in the following Theorem, which also presents a central limit approximation.

Theorem 2. *Let* $K_n \sim \text{PB}\{1, S(1;\boldsymbol{\theta}), \ldots, S(n-1;\boldsymbol{\theta})\}$ *and suppose* $K_\infty = \infty$ *almost surely. Then, as* $n \to \infty$, $K_n/s_n \to 1$ *almost surely, for* $s_n = \int_1^n S(t-1;\boldsymbol{\theta})\mathrm{d}t$. *In addition,*

$$\frac{K_n - \mathbb{E}(K_n)}{\text{var}(K_n)^{1/2}} \to N(0,1), \qquad n \to \infty, \qquad \textit{in distribution.}$$

Theorem 2 implies that the growth rate of K_n corresponds to $s_n = \int_1^n S(t-1;\boldsymbol{\theta})\mathrm{d}t$. In the Dirichlet process case, $s_n = \alpha \log(\alpha + n - 1) - \alpha \log \alpha$, corresponding to the well-known growth rate $\alpha \log n$ (Korwar and Hollander, 1973). The $N(0,1)$ limiting distribution allows one to assess uncertainty in K_n for large n. For similar results in generalized species sampling models settings, see Bassetti et al. (2010).

Consider now the second asymptotic regime: $K_\infty < \infty$. Although the distribution of K_∞ is generally not available in closed form, the first two moments are well defined.

Corollary 2. *Under the conditions of Proposition 1, if* $K_\infty < \infty$ *almost surely, then* $\mathbb{E}(K_\infty) = \sum_{i=1}^\infty S(i-1;\boldsymbol{\theta}) < \infty$ *and* $\text{var}(K_\infty) = \sum_{i=1}^\infty S(i-1;\boldsymbol{\theta})\{1 - S(i-1;\boldsymbol{\theta})\} < \infty$.

Hence, a natural estimator for the species richness is $\mathbb{E}(K_\infty)$, which may be numerically approximated; for instance by truncating the infinite summation $\mathbb{E}(K_\infty) = \sum_{i=0}^\infty S(i;\boldsymbol{\theta})$. Alternatively, one could exploit equation (6) and consider the arithmetic mean of the bounds, obtaining the approximation $\mathbb{E}(K_\infty) \approx \mathbb{E}(T) + 1/2$, which is highly accurate when the number of species is not small. Poisson-binomial conjugacy leads to a related estimator for the posterior species richness, namely $\mathbb{E}(K_\infty \mid D_1, \ldots, D_n)$. Consider $\mathbb{E}(K_{m+n} \mid D_1, \ldots, D_n)$ and let $m \to \infty$. Then, it is straightforward to see that $\mathbb{E}(K_\infty \mid D_1, \ldots, D_n) = k + \mathbb{E}(K_\infty^{(n)} \mid D_1, \ldots, D_n)$, where $\mathbb{E}(K_\infty^{(n)} \mid D_1, \ldots, D_n) = \sum_{j=1}^\infty S(j+n-1;\boldsymbol{\theta})$.

3 Logistic models

3.1 The log-logistic distribution

The framework in the previous Section requires elicitation of $S(t;\boldsymbol{\theta})$. In this Section, we focus on a class of survival functions, which lead to a generalization of the Dirichlet process, enjoy appealing analytical and computational properties and result in natural covariate-dependent extensions, as described in Section 3.3. In particular, we first consider a two parameter case

$$S(t;\alpha,\sigma)=\frac{\alpha}{\alpha+t^{1-\sigma}}, \qquad t\geq 0, \tag{7}$$

where $\alpha>0$ and $\sigma<1$. The survival function $S(t;\alpha,\sigma)$ characterizes a two-parameter log-logistic distribution, and therefore we will write $T\sim \textsc{ll}(\alpha,\sigma)$. Clearly, when $\sigma=0$, $S(t;\alpha,0)$ reduces to the Dirichlet process case. The parameter σ plays a similar role to the discount parameter of the Pitman–Yor process and general Gibbs-type priors. For any $\sigma<0$, one has

$$\mathbb{E}(T)=\frac{\alpha^{1/(1-\sigma)}\pi}{(1-\sigma)\sin\{\pi/(1-\sigma)\}},$$

implying that when $\sigma<0$ the limiting distribution $K_\infty<\infty$ is non-degenerate, thanks to Corollary 1. Conversely, when $0\leq\sigma<1$, one has that both $K_\infty=\infty$ and $\mathbb{E}(T)=\infty$. The rate at which this occurs is logarithmic in the Dirichlet process case in which $\sigma=0$. In contrast, for $\sigma>0$, one can show that the growth of K_n is polynomial, so that in the notation of Theorem 2 one has $s_n=\int_1^n S(t;\alpha,\sigma)\mathrm{d}t=\mathcal{O}(n^\sigma)$. These considerations reinforce the parallelism with Gibbs-type priors; see Gnedin and Pitman (2005) and De Blasi et al. (2015) for details.

In the next Section, we describe a three-parameter extension of the log-logistic distribution and derive combinatorial tools and distributional properties that also apply to $S(t;\alpha,\sigma)$ in (7).

3.2 A three parameter log-logistic distribution

In this Section we extend the log-logistic specification by including an additional parameter, denoted as ϕ, which forces K_n to converge to a non-degenerate distribution. This allows us to

restrict focus to the second asymptotic regime. In particular, we let $\boldsymbol{\theta} = (\alpha, \sigma, \phi)$ and

$$S(t; \alpha, \sigma, \phi) = \frac{\alpha \phi^t}{\alpha \phi^t + t^{1-\sigma}}, \qquad t \geq 0, \tag{8}$$

with $\alpha > 0$, $\sigma < 1$ and $0 < \phi \leq 1$. The two parameter specification is recovered when $\phi = 1$. We call the distribution of $S(t; \alpha, \sigma, \phi)$ a three-parameter log-logistic, written $T \sim \text{LL}(\alpha, \sigma, \phi)$.

Proposition 2. *Let $K_n \sim \text{PB}\{1, S(1; \boldsymbol{\theta}), \ldots, S(n-1; \boldsymbol{\theta})\}$, with $S(t; \boldsymbol{\theta})$ defined as in equation (8). Then for any $0 < \phi < 1$ it holds that $K_n \to K_\infty < \infty$ almost surely as $n \to \infty$.*

Proposition 2 ensures that for $0 < \phi < 1$ the species richness is always finite. For the remainder of the Section, we discuss some combinatorial properties related to the law of K_n. While having their own theoretical relevance, our results facilitate computation of the probability mass function of K_n and draw further parallels with Gibbs-type priors.

Definition 2. Let $\alpha > 0$, $\sigma < 1$ and $0 < \phi \leq 1$. Then for any $n \geq 1$ and $0 \leq k \leq n$ we define $\mathscr{C}_{n,k}(\sigma, \phi)$ as the coefficients of the polynomial expansion $\prod_{k=0}^{n-1}(\alpha + k^{1-\sigma}\phi^{-k}) = \sum_{k=0}^{n} \alpha^k \, \mathscr{C}_{n,k}(\sigma, \phi)$, having set $\mathscr{C}_{0,0}(\sigma, \phi) = 1$.

In the special case $\phi = 1$ and $\sigma = 0$ one recovers the definition of the signless Stirling numbers of the first kind, namely $\mathscr{C}_{n,k}(0, 1) = |s(n, k)|$; see Charalambides (2005). In addition, the coefficients $\mathscr{C}_{n,k}(\sigma, \phi)$ can be conveniently computed through recursive formulas.

Theorem 3. *The coefficients $\mathscr{C}_{n,k}(\sigma, \phi)$ of Definition 2 satisfy the triangular recurrence*

$$\mathscr{C}_{n+1,k}(\sigma, \phi) = \mathscr{C}_{n,k-1}(\sigma, \phi) + n^{1-\sigma}\phi^{-n}\mathscr{C}_{n,k}(\sigma, \phi),$$

for any $n \geq 0$ and $1 \leq k \leq n+1$, with initial conditions $\mathscr{C}_{0,0}(\sigma, \phi) = 1$, $\mathscr{C}_{n,0}(\sigma, \phi) = 0$, $n \geq 1$, $\mathscr{C}_{n,k}(\sigma, \phi) = 0$, $k > n$. Moreover, for any $1 \leq k \leq n$ and $n \geq 2$, one has

$$\mathscr{C}_{n,k}(\sigma, \phi) = \sum_{(i_1, \ldots, i_{n-k})} \prod_{j=1}^{n-k} i_j^{1-\sigma} \phi^{-i_j},$$

where the sum runs over the $(n-k)$*-combinations of integers* $(i_1, \ldots, i_{n-k})$ *in* $\{1, \ldots, n-1\}$.

We can now state the main theoretical result, namely the probability mass function of K_n, which can be expressed in terms of the coefficients $\mathscr{C}_{n,k}(\sigma, \phi)$.

Theorem 4. *Let* $K_n \sim \text{PB}\{1, S(1;\alpha,\sigma,\phi), \ldots, S(n-1;\alpha,\sigma,\phi)\}$ *for every* $n \geq 1$. *Then,*

$$\mathbb{P}(K_n = k) = \frac{\alpha^k}{\prod_{i=0}^{n-1}(\alpha + i^{1-\sigma}\phi^{-i})}\mathscr{C}_{n,k}(\sigma,\phi).$$

Theorem 4 reduces to the distribution obtained by Antoniak (1974) when $\sigma = 0$ and $\phi = 1$. Gibbs-type priors enjoy a similar structure for the distribution of K_n, replacing $\mathscr{C}_{n,k}(\sigma,\phi)$ with generalized factorial coefficients; see Gnedin and Pitman (2005); De Blasi et al. (2015).

3.3 Covariate-dependent models

Under the three parameter log-logistic specification, the discovery probabilities are $\pi_{n+1} = \mathbb{P}(D_{n+1} = 1) = \alpha\phi^n(\alpha\phi^n + n^{1-\sigma})^{-1}$ for $n \geq 1$ with $\pi_1 = 1$. An interesting and practically useful property of our model is the following representation

$$\log\frac{\pi_{n+1}}{1-\pi_{n+1}} = \log\alpha - (1-\sigma)\log n + (\log\phi)n = \beta_0 + \beta_1\log n + \beta_2 n, \qquad n \geq 1, \tag{9}$$

having set $\beta_0 = \log\alpha$, $\beta_1 = \sigma - 1 < 0$ and $\beta_3 = \log\phi \leq 0$. Equation (9) has the form of a logistic regression for the binary indicators $D_2, \ldots, D_n$, with coefficients β_2 and β_3 constrained to be negative. By letting $\beta_1 = -1$ and $\beta_2 = 0$ one recovers the discovery probability of the Dirichlet process.

The logistic regression representation in (9) facilitates extensions to include covariates. Suppose we are given a collection of L accumulation curves, $(K_{1n})_{n\geq 1}, \ldots, (K_{Ln})_{n\geq 1}$, representing sequential discoveries at different sampling locations. Each location is associated with covariates $\mathbf{z}_\ell^{\mathrm{T}} = (z_{\ell 1}, \ldots, z_{\ell p}) \in \mathbb{R}^p$ for $\ell = 1, \ldots, L$. Let $(D_{\ell n})_{n\geq 1}$ be the sequence of discovery indicators for

the ℓth location, with probabilities $(\pi_{\ell n})_{n\geq 1}$. The most flexible specification for $K_{\ell n}$ corresponds to the case in which all the parameters are location-specific, so that for any $n \geq 1$,

$$\log \frac{\pi_{\ell n+1}}{1-\pi_{\ell n+1}} = \beta_{\ell 0} + \beta_{\ell 1} \log n + \beta_{\ell 2} n, \qquad (\ell = 1, \ldots, L).$$

This specification can borrow information across locations via a hierarchical model on $\boldsymbol{\beta}_l = (\beta_{l0}, \beta_{l1}, \beta_{l2})^{\mathrm{T}}$ or by fixing certain parameters. Alternatively, systematic variation across locations can be modeled through including covariates $\mathbf{z}_\ell$ via

$$\log \frac{\pi_{\ell n+1}}{1-\pi_{\ell n+1}} = \beta_{\ell 0} + \beta_{\ell 1} \log n + \beta_{\ell 2} n = \mathbf{z}_\ell^{\mathrm{T}} \boldsymbol{\gamma}_0 + (\mathbf{z}_\ell^{\mathrm{T}} \boldsymbol{\gamma}_1) \log n + (\mathbf{z}_\ell^{\mathrm{T}} \boldsymbol{\gamma}_2) n, \tag{10}$$

for $\ell = 1, \ldots, L$, with $\boldsymbol{\gamma}_0, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2 \in \mathbb{R}^p$ being vectors of coefficients such that $\mathbf{z}_\ell^{\mathrm{T}} \boldsymbol{\gamma}_2 < 0$ and $\mathbf{z}_\ell^{\mathrm{T}} \boldsymbol{\gamma}_2 \leq 0$. This specification is still in the form of a logistic regression and therefore inference on the parameters $\boldsymbol{\gamma}_0, \boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ can be conducted through straightforward modifications of standard algorithms.

4 Posterior computation

4.1 Estimation procedures

Consider the model in equation (9). The parameters $\boldsymbol{\theta} = (\alpha, \sigma, \phi)$ can be estimated by maximizing the likelihood in equation (5), with $S(t; \boldsymbol{\theta}) = S(t; \alpha, \sigma, \phi)$, $\beta_1 < 0$ and $\beta_2 \leq 0$. In practice, it may suffice to ignore these restrictions and apply routine algorithms for fitting logistic regression, as the maximum likelihood estimates typically satisfy the constraints. In this case, the resulting estimate $\hat{\boldsymbol{\theta}}$ has the following appealing property.

Proposition 3. *Let $\hat{\boldsymbol{\theta}} = (\hat{\alpha}, \hat{\sigma}, \hat{\phi})$ be the unconstrained maximizer of equation (5) under the three-parameter specification in (8), if it exists. If $K_n = k$ is the number of discoveries within the data $D_1, \ldots, D_n$, then the expectation $\mathbb{E}(K_n)$, evaluated at $\hat{\boldsymbol{\theta}}$, equals k.*

Hence, $\mathbb{E}(K_n)$ matches the total number of distinct labels observed in the sequence when the parameters are estimated through unconstrained maximum likelihood. Although we can obtain confidence intervals and standard errors for the parameters via maximum likelihood, conducting inferences in this manner ignores the parameter constraints. In contrast, a fully Bayesian approach can easily incorporate them through a prior, such as $\boldsymbol{\beta} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})\mathbb{1}(\beta_1 < 0; \beta_2 \leq 0)$. The covariate-dependent regression in equation (10) can be implemented in a similar manner; for details see the Supplementary Materials.

4.2 Removing order dependence

The construction of accumulation curves is inherently order-dependent (Gotelli and Colwell, 2001). As such, inference on the parameter $\boldsymbol{\theta} \in \Theta$ depends on the order of the observations. This can be problematic when only the frequencies $n_1, \ldots, n_k$ are available, as there are $(n-1)!/\{(n-k)!(k-1)!\}$ curves that are consistent with these frequencies. This has motivated the derivation of the *individual-based rarefaction curve* (Smith and Grassle, 1977; Colwell et al., 2012),

$$\bar{K}_i = k - \binom{n}{i}^{-1} \sum_{j=1}^{k} \binom{n - n_j}{i}, \quad i = 1, \ldots, n, \tag{11}$$

with $K_n = k$, where (11) represents the average accumulation curve over all the possible orderings of the discoveries, each having the same probability. This proves useful in our case, as we can effortlessly apply our method relying on (11). Specifically, consider the auxiliary random variables $\bar{D}_i = \bar{K}_{i+1} - \bar{K}_i$ with $\bar{D}_1 = 1$. We can estimate $\boldsymbol{\theta}$ through the likelihood

$$\mathscr{L}(\boldsymbol{\theta} \mid \bar{D}_1, \ldots, \bar{D}_n) = \prod_{i=2}^{n} S(i-1; \boldsymbol{\theta})^{\bar{D}_i} \big\{1 - S(i-1; \boldsymbol{\theta})\big\}^{1-\bar{D}_i}, \tag{12}$$

in place of equation (5). Inference about $\boldsymbol{\theta}$ based on (12) will refer to the average accumulation curve. This procedure can be regarded as the approximation of a suitable marginal likelihood

$\mathbb{E}\{\mathscr{L}(\boldsymbol{\theta} \mid D_1, \ldots, D_n)\}$, representing the average likelihood over all the possible orderings of the discoveries. Thus, by interchanging the expectation operator inside the likelihood function, we obtain the approximation $\mathbb{E}\{\mathscr{L}(\boldsymbol{\theta} \mid D_1, \ldots, D_n)\} \approx \mathscr{L}(\boldsymbol{\theta} \mid \bar{D}_1, \ldots, \bar{D}_n)$.

5 Simulations

We test our log-logistic model on synthetic sequences generated from different asymptotic regimes. In each simulation, we randomly generate one sequence of labels from a given model and take the first $n = 10,000$ observations as a training set. The remaining $m = 20,000$ observations are used as a test set. We compare in- and out-of-sample performances of seven different models: our one-, two- and three-parameter log-logistic models, labelled as LL1, LL2, and LL3 henceforth, the two versions of the BETA-GOS detailed in Proposition 1 in Airoldi et al. (2014), the Pitman–Yor model and the Dirichlet-multinomial model. Our LL1 coincides with the Dirichlet process by Theorem 1.

When possible, model estimation proceeds via empirical Bayes on the training set. For the log-logistic model we rely on the constrained logistic regression representation (9). Parameters in the Pitman–Yor and Dirichlet-multinomial are obtained via maximization of the exchangeable partition probability function (Pitman, 1996), setting an arbitrarily high upper bound on H in the Dirichlet-multinomial equal to $k_n + 10,000$, with k_n being the number of distinct species observed in the training set at $n = 10,000$. Lacking a tractable likelihood, BETA-GOS processes are estimated via method of moments. Recalling that the discovery probability in BETA-GOS is $\pi_{n+1} = \prod_{i=1}^{n} W_i$ with independent $W_n \sim \text{BETA}(a_n, b_n)$, we employ two versions of the process. The first, BG-1(a, b), lets $a_n = a > 0$ and $b_n = b > 0$. In this case, the estimator for the accumulation curve is $\mathbb{E}(K_n) = (1 - \rho^n)/(1 - \rho)$, with $\rho = a/(a + b)$ and thus $K_\infty < \infty$ almost surely. We can estimate ρ by solving the equation $\mathbb{E}(K_n) = k_n$, with k_n defined as above. In the second version, BG-2(θ, β), we let $a_n = \theta + n - 1$ and $b_n = \beta$, with $\theta > 0$ and $\beta > 0$. The associated rarefaction is $\mathbb{E}(K_n) = \sum_{i=1}^{n} (\theta)_\beta/(\theta + i)_\beta$. This case admits both a finite and an

	DIR-MULT $H = 500, \sigma = -1$		BETA-GOS-2 $\theta = 500, \beta = 1.5$		DIRICHLET $\alpha = 10$		PITMAN–YOR $\alpha = 10, \sigma = 0.5$	
MODEL	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
DP (LL1)	$3,011.3$	$3,954.2$	$2,500.5$	$6,313.1$	6.7	7.6	$7,358.2$	3.3×10^4
PY	$3,011.3$	$3,953.6$	$2,500.5$	$6,311.6$	5.5	8.9	109.6	544.0
DIR-MULT	20.5	14.6	$1,266.7$	$2,908.6$	5.9	9.1	$6,857.1$	3.5×10^4
BG-1(a, b)	$1,633.4$	138.5	$9,345.8$	$3,905.5$	171.1	58.6	4.1×10^4	8.3×10^4
BG-2(θ, β)	18.3	23.6	38.4	152.3	4.4	14.6	51.4	982.6
LL2	71.8	141.0	77.7	428.1	3.9	11.3	70.0	$1,087.4$
LL3	11.8	50.1	22.5	452.2	3.1	16.1	67.3	$1,640.4$

Table 1: Models performance for curves simulated from Bayesian nonparametric predictive schemes. Values report average mean square error across 500 simulations of each scenario, with curves of length $30,000$. Training set consists of the first $10,000$ observations.

infinite species richness, as $K_\infty < \infty$ when $\beta > 1$ and $K_\infty = \infty$ when $\beta \in (0, 1]$. For further details, see Airoldi et al. (2014). Method of moment estimates for θ and β can be derived as a solution of the equations $\mathbb{E}(K_n) = k_n$ and $\mathbb{E}(K_{n/3}) = k_{n/3}$.

Table 1 reports the average mean square error across 500 accumulation curves simulated via Bayesian nonparametric predictive schemes. The first two scenarios, the Dirichlet-multinomial and BETA-GOS, feature a finite species richness. The other two assume a divergent accumulation curve. The purpose of our analysis is to compare the performance of our logistic models over species sampling sequences with the true generating model as a competitor. The in-sample average mean square error of LL3 is generally lower than other models, except in the Pitman–Yor case, where BG-2(θ, β) performs better. This reconfirms the strong similarity between the trajectories of the Pitman–Yor and BETA-GOS highlighted in the Introduction. Not surprisingly, the best model is always the true generating one in the test set. In almost every case, however, differences between the log-logistic specifications and the true model are small.

Following the same structure as above, Table 2 investigates the predictive performance of the models in the misspecified case in which the species probabilities follow geometric or Zipf distributions with or without truncation to finite support. We mirror the structure in Table 1,

	FINITE GEOM. $H = 100, \eta = 0.95$		FINITE ZIPF $H = 3000, \eta = 0.25$		GEOMETRIC $\eta = 0.1$		ZIPF $\eta = 2$	
MODEL	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
DP (LL1)	117.6	116.6	4.5×10^4	5.8×10^4	1.7	2.1	760.6	2,165.5
PY	117.6	116.6	4.5×10^4	5.8×10^4	1.7	2.1	405.1	105.7
DIR-MULT	5.1	4.3	192.1	1,156.2	1.5	2.9	747.2	2,181.8
BG-1(a, b)	85.4	0.2	803.9	1,259.0	23.3	8.1	2,847.9	3,895.9
BG-2(θ, β)	9.0	2.4	140.3	1,539.9	1.6	4.4	12.6	187.8
LL2	7.3	12.0	2062.0	8.3×10^4	1.2	2.8	11.6	125.0
LL3	1.4	0.6	62.9	849.9	1.0	3.8	9.7	293.8

Table 2: Performance for curves simulated via independent samples from finite and infinite support distributions. Values report average mean square error across 500 simulations of each scenario, with curves of length 30,000. Training set consists of the first 10,000 observations.

with the first two models having $K_\infty < \infty$ and the last two $K_\infty = \infty$. Details are provided in the Supplementary Material. LL3 achieves the best in-sample performance, and log-logistic models perform particularly well in finite (truncated) cases. In the infinite cases, the Pitman-Yor had good predictive performance, likely due to similar tail behavior between PY and these two distributions, but failed badly in-sample for the Zipf.

The values for the parameters of the generating models we have chosen in this Section are intended to simulate representative trajectories for the accumulation curves, both in converging and diverging cases. For an extended analysis on more scenarios and varying parameters, including plots of the generated curves, refer to the Supplementary Material.

6 Applications

6.1 Copepod species counts

We test our model on a dataset of abundances of distinct copopod species from the Southampton National Oceanography Centre, available in the `R` package `untb` (Hankin, 2007). The data

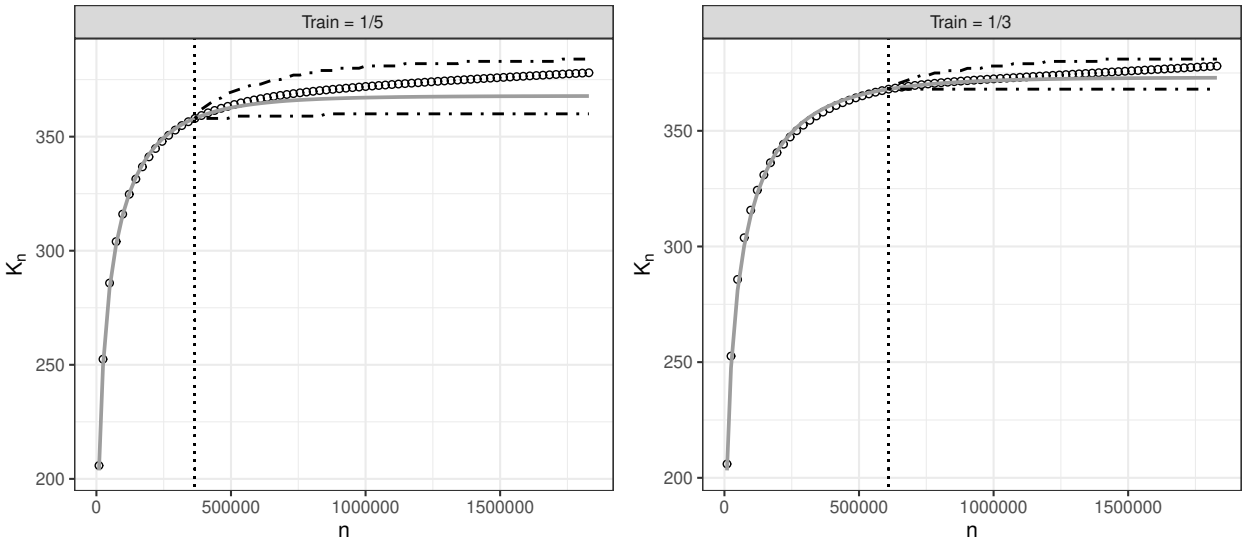

Figure 2: Performance of LL3 on the copepod species counts data. Circles: individual-based rarefaction curve. Grey line: predicted in- and out-of-sample accumulation curve computed by averaging over posterior samples of $\mathbb{E}(K_n)$ and $\mathbb{E}(K_n \mid D_1, \ldots, D_n)$, respectively. Black dashed lines indicate the posterior 95% posterior predictive credible interval, obtain by simulating one posterior trajectory for each sample. The black vertical line indicates the training-test cutoff.

consist of $n = 1,829,767$ observations divided into 378 species, with 10 appearing only once, 3 appearing twice and the most abundant species appearing $503,319$ times. As depicted by the circles in Figure 2, the individual-based rarefaction curve seems close to convergence, facilitating assessments of model performance that attempt to predict the later part of the curve and species richness based on an initial part of the curve.

We compare the models of Section 5 by considering two training-test settings, taking random subsets of one-fifth and one-third of the data as training sets. We extrapolate the fitted curves for the remaining samples. Model fitting proceeds with a fully Bayesian approach when possible, initializing the chain at the maximum likelihood estimate and performing $10,000$ iterations after a $5,000$ burn-in. For the Pitman–Yor process we adopt normal priors centered at 0 with a standard deviation of 10 for $\gamma_1 = \log(\alpha + \sigma)$ and $\gamma_2 = \log\{\sigma(1-\sigma)^{-1}\}$, and apply Adaptive Metropolis (Haario et al., 2001) keeping one sample every 10 iterations. A similar procedure is

applied to the Dirichlet process with $\beta_0 = \log \alpha \sim N(0, 10)$, but saving every iteration. For the log-logistic models we use equation (9), and impose the constraints with truncated normal priors as in Section 4. Posterior samples for LL2 and LL3 are obtained via the Metropolis adjusted Langevin algorithm (Roberts and Rosenthal, 1998) with the proposal covariance equal to $\epsilon^2\hat{\Sigma}$, where $\hat{\Sigma}$ is the inverse of the Hessian of the model evaluated at the maximum likelihood estimate and ϵ^2 is a scaling parameter iteratively tuned to reach an acceptance rate of 0.576.

All the samplers had effective sample sizes between $2,000$ and $6,000$. Finally, we sample from the posterior of the Dirichlet-multinomial by discretizing σ into $5,000$ equally spaced values between -0.005 and -3, fixing an upper bound on H equal to $5,000$ plus the observed K_n and setting a discrete uniform prior over each interval. For the BETA-GOS models, the absence of a simple form for the likelihood limits the availability of posterior samplers. Thus, we estimate the parameters via method of moments by solving the linear systems described in Section 5, taking k_n to be the nth value of the individual-based rarefaction in equation (11).

Table 3 compares in- and out-of-sample performance. The MSE columns report the mean square error between the individual-based rarefaction curve $\bar{K}_{n+m}$ and the model-based rarefaction estimator, obtained by averaging $\mathbb{E}(K_n)$ over posterior samples. In both cases, LL3 shows the best in-sample performance. To test out-of-sample performance, we first compute the individual-based rarefaction curve for the test set, $\bar{K}_{n+m}$, by averaging across $5,000$ randomly sampled orders of appearance in the test set. Then, we extrapolate by simulating one trajectory $K_{n+m} \mid D_1, \ldots, D_n$, $m \geq 1$, for each sample drawn from the posterior distribution of the parameters. This is straightforward for the species sampling and log-logistic models, but problematic for BETA-GOS due to prohibitive computational cost for large n. Fortunately, for large n, the variance of the discovery probability is typically small and goes to 0 as $n \to \infty$ when $\beta \geq 1$. This implies that fixing the discovery probabilities to their average values when sampling one accumulation curve induces only a minor reduction in uncertainty. For more details, see the Supplementary Material. In both cases, the 95% posterior predictive credible interval for $K_{n+m} \mid D_1, \ldots, D_n$ for LL3 contains the true value $\bar{K}_{n+m}$, and the posterior predictive mean

	TRAIN=1/5				TRAIN=1/3			
		$n = 365,953$; $K_n = 358$				$n = 609,922$; $K_n = 368$		
	MSE	$m = n/2$	$m = n$	$m = 4n$	MSE	$m = n/4$	$m = n$	$m = 2n$
$\bar{K}_{n+m}$		365.16	368.87	378		370.32	373.97	378
DP (LL1)	50.41	373.93 (367, 382)	385.21 (375, 397)	421.15 (405, 439)	130.99	376.56 (371, 383)	394.47 (385, 406)	409.87 (397, 424)
PY	60.91	374.56 (367, 384)	386.35 (376, 399)	424.03 (406, 446)	131.21	376.66 (371, 383)	394.83 (385, 406)	410.45 (397, 425)
DIR-MULT	41.60	373.01 (366, 381)	383.47 (374, 394)	416.95 (401, 434)	90.85	375.97 (371, 382)	392.54 (383, 403)	406.69 (394, 420)
BG-1(a, b)	2703.71	358 (358, 358)	358 (358, 358)	358 (358, 358)	2101.28	368 (368, 368)	368 (368, 368)	368 (368, 368)
BG-2(θ, β)	73.8	369.72 (364, 377)	377.79 (370, 387)	402.40 (390, 416)	95.25	372.97 (369, 378)	382.94 (376, 391)	391.04 (382, 401)
LL2	125.22	376.48 (368, 386)	389.73 (378, 408)	432.85 (410, 459)	178.23	377.04 (372, 384)	396.08 (385, 409)	412.58 (397, 430)
LL3	1.59	363.70 (359, 371)	365.91 (359, 377)	367.80 (360, 384)	3.52	370.17 (368, 374)	372.40 (368, 380)	372.93 (368, 381)

Table 3: Model performances and out-of-sample predictions on the copepod species counts data. The columns under MSE report in-sample the mean square error of $\mathbb{E}(K_n)$. Values in brackets report the 95% posterior predictive credible interval for the extrapolation estimator.

$\mathbb{E}(K_{n+m} \mid D_1, \ldots, D_n)$ slightly underestimates the truth. This is further confirmed by looking at the whole trajectory, as depicted in Figure 2. The species sampling models do not correctly capture the average out-of-sample trajectory of the test set. This is expected in the Dirichlet and Pitman-Yor processes, as both assume a divergent K_n. However, the Dirichlet-multinomial also performs badly, likely due to the behavior resembling the Dirichlet process for values of σ close to 0 and large values of H. For BETA-GOS-2, the out-of-sample trajectory is captured only for values close to the training-test cutoff. Finally, BETA-GOS-1 performs poorly due to the lack of flexibility of the underlying exponential behavior of the model. For more results on the data, including plots, posterior estimates of the parameters and additional training-test splits, refer to the Supplementary Material.

6.2 Fungal biodiversity

We analyze data from a fungi biodiversity study in Finland (Abrego et al., 2020). Each sample contains a large number of fungal DNA barcode sequences obtained either from air samples or soil samples. As it is too expensive to barcode all the fungi spores in a sample, it is important to be able to predict how many species are missed when sequencing a particular amount. The goal of our analysis is to answer this question.

The data consist of 174 different samples from different sites across five cities in Finland. For each site, fungi samples are collected on the same dates at two urban areas, one at the core and one at the edge of the city, and two nearby natural areas, again with one at the core and one at the edge. Two different sampling methods were used: i) through air, via a cyclone trap and continuously for 24 hours, and ii) through soil, gathering a small portion of soil close to the air trap. We exclude samples with less than $10,000$ sequences, as in such cases the samples lacked sufficient numbers of spores for more comprehensive barcoding. This leaves us with a total of 150 samples. An issue in pre-processing the data is reliable identification of singletons, OTUs that have been identified only once within a given sample. Ecologists often discard such singletons from the analysis, leading to significant bias. In the Supplementary Materials we instead propose a simple imputation approach.

The average number of barcoded DNA sequences per sample is $124,271$ and the average number of species discovered is $2,161$. As a first step, we compare the in-sample performances of four different models: Pitman-Yor, BETA-GOS-2 and two- and three- parameter log-logistic models. We exclude BETA-GOS-1, Dirichlet-multinomial and Dirichlet/one-parameter log-logistic, as they showed very poor performance. Model fitting and prediction proceeded exactly as in Section 6.1.

Figure 3 displays in-sample performance of the models across the 150 samples. Each point represents the percentage absolute error between $\mathbb{E}(K_n)$, obtained by averaging the model rarefaction across the posterior samples, and $\bar{K}_n$ at a given fraction of a curve. All models perform

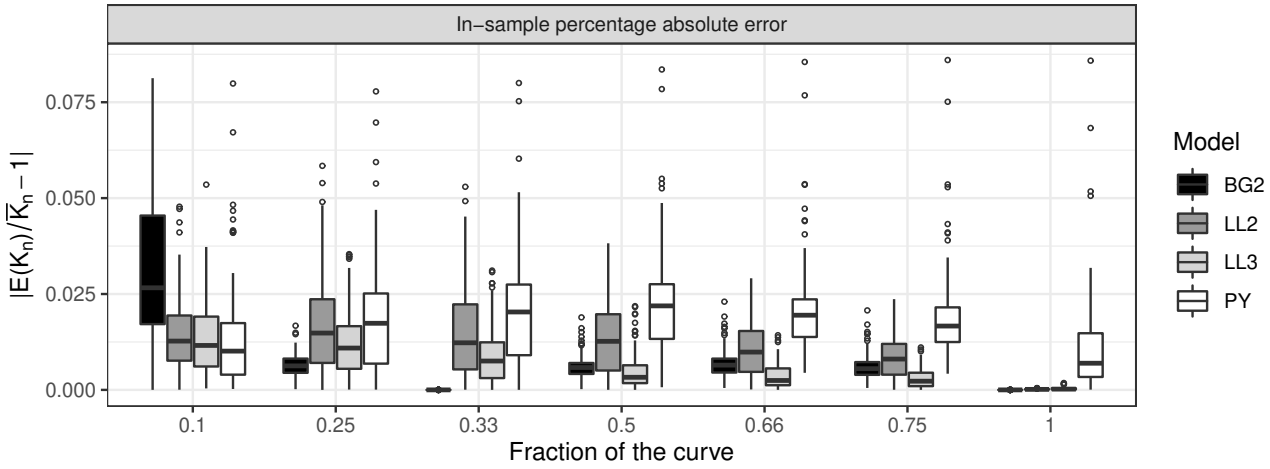

Figure 3: In-sample performance in the Finnish fungal biodiversity data. The in-sample estimator $\mathbb{E}(K_n)$ is computed by averaging model rarefaction across posterior samples. The value of $\bar{K}_n$ indicates the individual-based rarefaction curve at n.

well overall, deviating from the true values of $\bar{K}_n$ by less than 1%. The Pitman–Yor is the least flexible in-sample. BETA-GOS-2 yields perfect fit at fractions 0.33 and 1 due to the estimates of θ and β being the solution of $\mathbb{E}(K_{n/3}) = \bar{k}_{n/3}$ and $\mathbb{E}(K_n) = \bar{k}_n$, with n the total length of a given curve. The consequence of this choice is that the beginning of the curve, namely fraction 0.1, shows more error variability. For the log-logistic models, the high accuracy at fraction 1 is an indirect consequence of Proposition 3 and vague priors over the regression coefficients.

Although the above models fit well, only LL3 estimated a convergent K_∞. For BETA-GOS2 all curves estimated $\beta < 1$, implying $K_\infty = \infty$. Thus, we rely on LL3 in performing inferences on i) the sample species richness, which is the total number of species that can be detected through barcoding within a sample, and ii) whether DNA barcoding has reached *saturation* at different sites, meaning that only very few species are missed. To address i), we estimate the posterior mean $\mathbb{E}(K_\infty \mid D_1, \ldots, D_n)$ for each individual sample, which is guaranteed to be finite. The results are reported in the left panel of Figure 4, which displays the expected sample species richness for each of the 150 samples across site characteristics. Air samples tend to contain

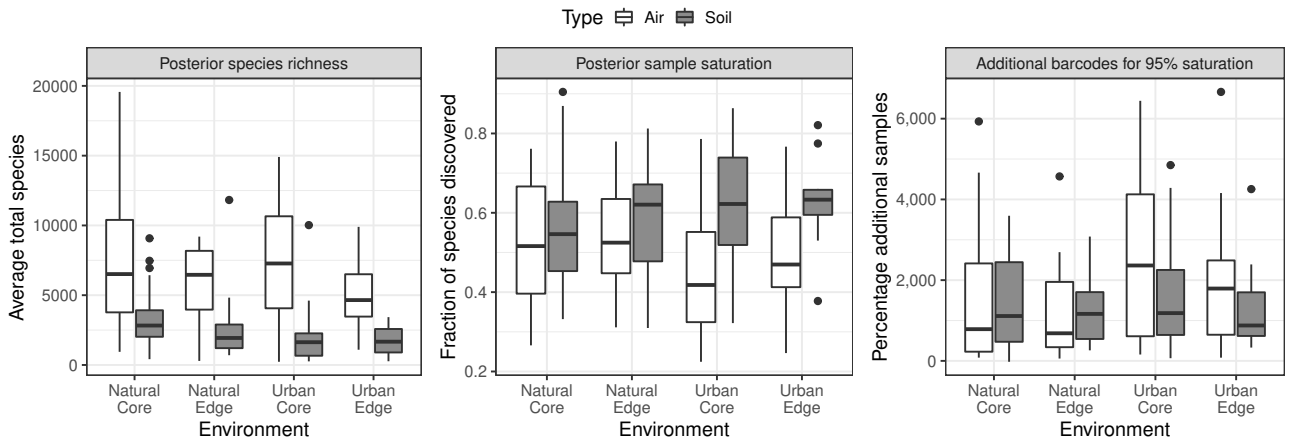

Figure 4: Left panel: distribution of the posterior mean species richness for the 150 samples. Center panel: distribution of the posterior mean sample saturation for the 150 samples. Right panel: additional number of samples (in percentage) required to reach a target posterior saturation of 0.95.

more species, and there is some evidence of greater species richness in natural environments, as reported by Abrego et al. (2020).

For task ii), let $C_n = K_n/K_\infty \leq 1$ represent the saturation level of a given sample after n barcoded sequences. Differences across sites can be evaluated via $\mathbb{E}(C_n \mid D_1, \ldots, D_n)$, which represents the posterior expected saturation level of a sample. Figure 4, right panel, summarizes posterior mean saturation stratified by sampling site characteristics. While there is some variability across sites, most of them have a ratio around 0.5. The results suggest that if additional DNA sequences are barcoded there is the opportunity to detect approximately $20 - 50\%$ more species in each sample. Urban soil samples seem to have a systematically higher saturation than their Air counterparts. Finally, we can estimate the number m of additional sequences that would need to be barcoded to reach a desired saturation level C_{n+m}. This is reported in the right panel of Figure 4, where the target saturation level is 95%. This confirms the fact that generally all samples require a high barcoding effort to detect almost all the species.

7 Discussion

In this paper we proposed a novel method for predicting the appearance of previously unobserved objects in a sequence. We showed that our procedure generalizes the discovery probability of the Dirichlet process. Finite sample and asymptotic properties of the number of distinct species K_n were extensively studied. In addition, we showed that a subclass of models is linked to a logistic regression with constrained coefficients. This has major computational advantages compared to existing Bayesian nonparametric procedures, which allowed us to implement our modelling strategies in large datasets. All of our estimators are based on moments of the Poisson-binomial distribution. Despite its rather complex shape (Chen, 1975), this distribution admits several approximations (e.g. Goldstein, 2010; Hong, 2013). These may be useful in obtaining approximations to the distribution of K_∞.

From a Bayesian nonparametric perspective, our species discovery framework enriches the increasingly large literature on models beyond exchangeability (e.g. Berti et al., 2004; Airoldi et al., 2014; Fortini et al., 2018; Ascolani et al., 2021; Berti et al., 2021). Indeed, the construction of an *accumulation curve* is intrinsically a non-exchangeable procedure, because the sequential discoveries necessarily depend on the chosen ordering (Gotelli and Colwell, 2001). We solved the order dependence by applying our framework to the individual-based rarefaction curve, which is the *average* accumulation curve for given abundances (Smith and Grassle, 1977). As detailed in Remark 1, choices for the allocation probabilities under a sequential discovery model without exchangeability may still retain certain convenient properties (Bassetti et al., 2010). Urn-based non-exchangeable models are particularly promising for sequential and dynamic data.

Instead of taking a Bayesian nonparametric perspective, similar methodology and theoretical conclusions could have been achieved by modelling the trajectory in the distinct species as the output of a discrete time pure-birth in-homogeneous Markov process with birth probability $S(t;\boldsymbol{\theta})$. The link between pure-birth processes and accumulation curves has long been known (Soberon and Llorente, 1993; Diaz-Frances and Gorostiza, 2002). We chose to focus on the

Bayesian nonparametric viewpoint due to the rich statistical literature on species sampling taking this perspective.

We extensively investigated in Section 3 the logistic subclass of models, which has appealing theoretical and computational properties. However, different survival functions $S(t;\boldsymbol{\theta})$ may be considered (e.g. exponential, Weibull, Gompertz) to accommodate different shapes and growth rates. For example, one can impose $S(t;\boldsymbol{\theta})$ to be equal to the average discovery probability of the BETA-GOS-2 with $\beta \geq 1$. Indeed, our results of Section 2 are fully general and can be readily specialized to any survival function. This is an interesting research direction.

SUPPLEMENTARY MATERIAL

The Supplementary Material includes the proofs of the theorems and the propositions stated above, some extended simulations and additional details on the data preprocessing and on the general results on the application section.

Acknowledgments

This project has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement No 856506).

References

Abrego, N., B. Crosier, P. Somervuo, N. Ivanova, A. Abrahamyan, A. Abdi, K. Hämäläinen, K. Junninen, M. Maunula, J. Purhonen, and O. Ovaskainen (2020). Fungal communities decline with urbanization – more in air than soil. *The ISME Journal 14*, 2806–2815.

Airoldi, E. M., T. Costa, F. Bassetti, F. Leisen, and M. Guindani (2014). Generalized species sampling priors with latent beta reinforcements. *Journal of the American Statistical Association 109*(508), 1466–1480.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics 2*(6), 1152–1174.

Arrhenius, O. (1921). Species and area. *Journal of Ecology 9*(1), 95–99.

Ascolani, F., A. Lijoi, and M. Ruggiero (2021). Predictive inference with Fleming–Viot-driven dependent Dirichlet processes. *Bayesian Analysis 16*(2), 371–395.

Bassetti, F., I. Crimaldi, and F. Leisen (2010). Conditionally identically distributed species sampling sequences. *Advances in Applied Probability 42*(2), 433–459.

Berti, P., E. Dreassi, L. Pratelli, and P. Rigo (2021). A class of models for Bayesian predictive inference. *Bernoulli 27*(1), 702–726.

Berti, P., L. Pratelli, and P. Rigo (2004). Limit theorems for a class of identically distributed random variables. *Annals of Probability 32*(3), 2029–2052.

Blackwell, D. and J. B. MacQueen (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics 1*(2), 353–355.

Bunge, J. and M. Fitzpatrick (1993). Estimating the number of species: a review. *Journal of the American Statistical Association 88*(421), 364–373.

Camerlenghi, F., B. Dumitrascu, F. Ferrari, B. E. Engelhardt, and S. Favaro (2020). Nonparametric Bayesian multi-armed bandits for single cell experiment design. *The Annals of Applied Statistics 14*(4), 2003–2019.

Camerlenghi, F., A. Lijoi, and I. Prünster (2018). Bayesian nonparametric inference beyond the Gibbs-type framework. *Scandinavian Journal of Statistic 45*, 1062–1091.

Cassese, A., W. Zhu, M. Guindani, and M. Vannucci (2019). A Bayesian nonparametric spiked process prior for dynamic model selection. *Bayesian Analysis 14*(2), 553–572.

Chao, A. and T.-J. Shen (2004). Nonparametric prediction in species sampling. *Journal of Agricultural, Biological and Environmental Statistics 9*, 253–269.

Charalambides, C. A. (2005). *Combinatorial Methods in Discrete Distributions*. Hoboken, NJ: Wiley.

Chen, L. H. Y. (1975). Poisson approximation for dependent trials. *The Annals of Probability 3*(3), 534–545.

Christen, J. A. and M. Nakamura (2000). On the analysis of accumulation curves. *Biometrics 56*(3), 748–754.

Colwell, R. K. (2009). Biodiversity: concepts, patterns, and measurement. In *Levin SA. (ed.). The Princeton Guide to Ecology*, pp. 257–263. Princeton: Princeton University Press.

Colwell, R. K., A. Chao, N. J. Gotelli, S.-Y. Lin, C. X. Mao, R. L. Chazdon, and J. T. Longino (2012). Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology 5*(1), 3–21.

Darroch, J. N. (1964). On the Distribution of the Number of Successes in Independent Trials. *The Annals of Mathematical Statistics 35*(3), 1317–1321.

De Blasi, P., S. Favaro, A. Lijoi, R. H. Mena, I. Prünster, and M. Ruggiero (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Transactions on Pattern Analysis and Machine Intelligence 37*(2), 212–229.

Diaz-Frances, E. and L. G. Gorostiza (2002). Inference and model comparison for species accumulation functions using approximating pure birth processes. *Journal of Agricultural, Biological, and Environmental Statistics 7*(3), 335–349.

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

Favaro, S., A. Lijoi, R. H. Mena, and I. Prünster (2009). Bayesian non-parametric inference for species variety with a two-parameter Poisson–Dirichlet process prior. *Journal of the Royal Statistical Society, Series B 71*(5), 993–1008.

Favaro, S., A. Lijoi, and I. Prünster (2012). A new estimator of the discovery probability. *Biometrics 68*(4), 1188–1196.

Fisher, R. A., A. S. Corbet, and C. B. Williams (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*(1), 42–58.

Flather, C. (1996). Fitting species-accumulation functions and assessing regional land use impacts on avian diversity. *Journal of Biogeography 23*(2), 155–168.

Fortini, S., S. Petrone, and P. Sporysheva (2018). On a notion of partially conditionally identically distributed sequences. *Stochastic Processes and their Applications 128*(3), 819–846.

Gao, Z., C.-H. Tseng, Z. Pei, and M. Blaser (2007). Molecular analysis of human forearm superfical skin bacterial biota. *Proceedings of the National Academy of Sciences of the United States of America 104*, 2927–2932.

Gleser, L. J. (1975). On the distribution of the number of successes in independent trials. *The Annals of Probability 3*(1), 182–188.

Gnedin, A. and J. Pitman (2005). Exchangeable Gibbs partitions and Stirling triangles. *Zapiski Nauchnykh Seminarov, POMI 325*, 83–102.

Goldstein, L. (2010). Bounds on the constant in the mean central limit theorem. *The Annals of Probability 38*(4), 1672–1689.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika 40*(3-4), 237–264.

Good, I. J. and G. H. Toulmin (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika 43*(1-2), 45–63.

Gotelli, N. J. and R. K. Colwell (2001). Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters 4*, 379–391.

Haario, H., E. Saksman, and J. Tamminen (2001). An adaptive Metropolis algorithm. *Bernoulli 7*(2), 223–242.

Hankin, R. K. S. (2007). Introducing `untb`, an `R` package for simulating ecological drift under the unified neutral theory of biodiversity. *Journal of Statistical Software 22*(12), 1–15.

Hoeffding, W. (1956). On the Distribution of the Number of Successes in Independent Trials. *The Annals of Mathematical Statistics 27*(3), 713–721.

Hong, Y. (2013). On computing the distribution function for the Poisson binomial distribution. *Computational Statistics & Data Analysis 59*, 41–51.

Hughes, J. B., J. J. Hellmann, T. H. Ricketts, and B. J. M. Bohannan (2001). Counting the uncountable: statistical approaches to estimating microbial diversity. *Applied and Environmental Microbiology 67*(10), 4399–4406.

Ionita-Laza, I., C. Lange, and N. M. Laird (2009). Estimating the number of unseen variants in the human genome. *Proceedings of the National Academy of Sciences of the United States of America 106*(13), 5008–5013.

Korwar, R. M. and M. Hollander (1973). Contributions to the theory of Dirichlet processes. *The Annals of Probability 1*(4), 705–711.

Le Cam, L. (1960). An approximation theorem for the poisson-binomial distribution. *Pacific Journal of Mathematics 10*(4), 1181–1197.

Lee, J., F. A. Quintana, P. Müller, and L. Trippa (2013). Defining predictive probability functions for species sampling models. *Statistical Science 28*(2), 209–222.

Lijoi, A., R. H. Mena, and I. Prünster (2007). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika 94*(4), 769–786.

Lijoi, A., I. Prünster, and T. Rigon (2020). The Pitman–Yor multinomial process for mixture modeling. *Biometrika 107*(4), 891–906.

Mao, C. X. (2004). Predicting the conditional probability of discovering a new class. *Journal of the American Statistical Association 99*(468), 1108–1118.

Ovaskainen, O., N. Abrego, P. Somervuo, I. Palorinne, B. Hardwick, J.-M. Pitkänen, N. R. Andrew, P. A. Niklaus, N. M. Schmidt, S. Seibold, J. Vogt, E. V. Zakharov, P. D. N. Hebert, T. Roslin, and N. V. Ivanova (2020). Monitoring fungal communities with the global spore sampling project. *Frontiers in Ecology and Evolution 7*, 511.

Perman, M., J. Pitman, and M. Yor (1992). Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields 92*(1), 21–39.

Pitman, J. (1996). Some developments of the Blackwell-Macqueen urn scheme. In T. S. Ferguson, L. S. Shapley, and J. B. MacQueen (Eds.), *Statistics, Probability and Game Theory. Papers in honor of David Blackwell*, Volume 30 of *IMS Lecture notes, Monograph Series*, pp. 245–267. Hayward: Institute of Mathematical Statistics.

Pitman, J. (1997). Probabilistic bounds on the coefficients of polynomials with only real zeros. *Journal of Combinatorial Theory, Series A 77*, 279–303.

Pitman, J. and M. Yor (1997). The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator. *The Annals of Probability 25*(2), 855–900.

Quintana, F. A., P. Müller, A. Jara, and S. N. MacEachern (2022). The Dependent Dirichlet Process and Related Models. *Statistical Science 37*(1), 24–41.

Roberts, G. O. and J. S. Rosenthal (1998). Optimal scaling of discrete approximations to langevin diffusions. *Journal of the Royal Statistical Society: Series B 60*(1), 255–268.

Shen, T.-J., A. Chao, and C.-F. Lin (2003). Predicting the number of new species in further taxonomic sampling. *Ecology 84*(3), 798–804.

Smith, W. and F. Grassle (1977). Sampling properties of a family of diversity measures. *Biometrics 33*(2), 283–292.

Soberon, J. and J. Llorente (1993). The use of species accumulation functions for the prediction of species richness. *Conservation Biology 7*, 480–488.

Thisted, R. and B. Efron (1987). Did Shakespeare write a newly-discovered poem? *Biometrika 74*(3), 445–455.

Xu, M. and N. Balakrishnan (2011). On the convolution of heterogeneous Bernoulli random variables. *Journal of Applied Probability 48*(3), 877–884.