Ecological Modeling via Bayesian Nonparametric Species Sampling Priors

by

Alessandro Zito

Department of Statistical Science
Duke University

Date: __________________________
Approved:

____
David B. Dunson, Supervisor

____
Peter Hoff

____
Mike West

____
James Clark

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2023

Abstract

Ecological Modeling via Bayesian Nonparametric Species Sampling Priors

by

Alessandro Zito

Department of Statistical Science
Duke University

Date: ______________________
Approved:

____
David B. Dunson, Supervisor

____
Peter Hoff

____
Mike West

____
James Clark

An abstract of a dissertation submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2023

Abstract

Species sampling models are a broad class of discrete Bayesian nonparametric priors that model the sequential appearance of distinct tags, called species or clusters, in a sequence of labeled objects. Over the last 50 years, species sampling priors have found much success in a variety of settings, including clustering and density estimation. However, despite the rich theoretical and methodological developments, these models have rarely been used as tools by applied ecologists, even though their primary investigation often involves the modeling of actual species. This dissertation aims at partially filling this gap by elucidating how species sampling models can be useful to scientists and practitioners in the ecological field. Our emphasis is on clustering and on species discovery properties linked to species sampling models. In particular, Chapter 2 illustrates how a Dirichlet process mixture model with a random precision parameter leads to greater robustness when inferring the number of clusters, or communities, in a given population. We specifically introduce a novel prior for the precision, called Stirling-gamma distribution, which allows for transparent elicitation supported by theoretical findings. We illustrate its advantages when detecting communities in a colony of ant workers. Chapter 3 presents a general Bayesian framework to model accumulation curves, which summarize the sequential discoveries of distinct species over time. This work is inspired by traditional species sampling models such as the Dirichlet process and the Pitman–Yor process. By modeling the discovery probability as a survival function of some latent variables, a flexible specification that can account for both finite and infinite species richness is developed. We apply our model to a large fungal biodiversity study from Finland. Finally, Chapter 4 presents a novel Bayesian nonparametric taxonomic classifier called BayesANT. Here, the goal is to predict the taxonomy of DNA sequences sampled from the environment. The difficulty of such a task is that the vast majority of species do not have a reference barcode or are yet unknown to science. Hence, species novelty needs to be accounted for when doing classification. BayesANT builds upon Dirichlet-multinomial kernels to model DNA sequences, and upon species sampling models to account for such potential novelty. We show how it attains excellent classification performances, especially when the true taxa of the test sequences are not observed in the training set. All methods presented in this dissertation are freely available as `R` packages. Our hope is that these contributions will pave the way for future utilization of Bayesian nonparametric methods in applied ecological analyses.

Dedication

Alla Pucci,
per tutti i giorni trascorsi lontani.

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Symbols

α Dirichlet process precision parameter.

σ Pitman–Yor strength parameter.

$C_1, \dots C_k$ A partition of the units $\{1, \dots, n\}$ into k clusters.

n_j The number of elements in the jth cluster C_j from a species sampling sequence.

K_n The number of species, or clusters, in the species sampling sequence of size n.

K_∞ The species richness, or the total number of species, or clusters, in the population.

X_i The ith realization from a species sampling sequence.

X_j^* The jth distinct value from a species sampling sequence.

$(x)_n$ Ascending factorial, equal to $(x)_n = \prod_{i=0}^{n-1}(x+i)$.

$|s(n,k)|$ Signless Stirling-number of the first kind.

$B_p(x_1, \dots, x_p)$ Complete exponential Bell polynomial of order p, evaluated at $x_1, \dots, x_p$.

$\Gamma(x)$ Gamma function evaluated in x.

$\mathcal{S}_{a,b,m}$ Normalizing constant of the Stirling-gamma distribution with with parameters a, b and m.

$V_{n,k}$ Coefficients of a Gibbs-type process under n observations and k species.

$\mathrm{Dir}(\alpha_1, \dots, \alpha_p)$ The Dirichlet distribution with parameters $\alpha_1, \dots, \alpha_p$.

$\mathrm{Ga}(a, b)$ The Gamma distribution with shape a and rate b.

$\mathrm{LL}(\alpha, \sigma, \phi)$ The three-parameter log-logistic distribution, with parameters α, σ and ϕ.

$\mathrm{Mult}(n; \pi_1, \dots, \pi_S)$ The multinomial distribution with n trials and success probabilities $\pi_1, \dots, \pi_S$ that lie on the S-simplex.

$\mathrm{N}(\mu, \sigma^2)$ The normal distribution with mean μ and variance σ^2.

$\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ The p-variate normal distribution with mean $\boldsymbol{\mu}$ and variance-covariance $\boldsymbol{\Sigma}$.

$\mathrm{Negbin}(r, q)$ The negative binomial distribution with mean $r(1-q)/q$.

$\mathrm{Pb}(q_1, \dots, q_p)$ The Poisson binomial distribution with success probabilities q_1 ,. . ., q_p.

$\mathrm{Po}(\lambda)$ The Poisson distribution with rate parameter λ.

$\mathrm{Sg}(a, b, m)$ The Stirling-gamma distribution with location a, precision b and reference sample size m.

Abbreviations

BayesANT Bayesian Nonparametric taxonomic classifier.

BETA-GOS Generalized Ottawa sequence with latent beta reinforcements.

BOLD Barcode of Life dataset.

DP Dirichlet process.

FinBOL Finnish Barcode of Life dataset.

LL-1 One parameter log-logistic model (Dirichlet process).

LL-2 Two-parameter log-logistic model.

LL-3 Three-parameter log-logistic model.

OTU Operational Taxonomic Unit.

PY Pitman–Yor process.

Acknowledgements

Caro piccolo insetto
che chiamavano mosca non so perché,
stasera quasi al buio
mentre leggevo il Deuteroisaia
sei ricomparsa accanto a me,
ma non avevi occhiali,
non potevi vedermi
né potevo io senza quel luccichìo
riconoscere te nella foschia.[1]

Eugenio Montale - Xenia I, Satura (1971)

To a great extent, "recognizing insects in the haze" has been one of the primary goals of my research. As Montale mentions in one of his later poems, however, this recognition is only possible given the "spark" that makes the "fly's glasses" glow in the dark. Without it, nothing can be seen, no matter how close the little insect is or how deeply we read the "Deutero-Isaiah". Thankfully, all these years, I have had the pleasure to encounter, collaborate with, and be guided by many wonderful people, each of whom ignited the many sparks that led to this dissertation.

First and foremost, the largest fire was lit by my advisor David Dunson, without whom I would not be the researcher I am today. It has been an honor to work alongside him. I will forever be grateful for all his support, his trust, and all the profound

[1] Dear little insect | who they called - I don't know why - fly, | this evening on the brink of dark | while I was reading Deutero-Isaiah | you reappeared, right here, beside, | but you had no glasses, | you couldn't see me | nor could I, without their spark | recognize you in the haze.

insights he has provided all these years. An equally heartfelt acknowledgment goes to Tommaso Rigon, from which I learned so much about how to think, connect the dots, and later write and present ideas. He never ceased to believe in our work and has pushed me to strive for clarity and excellence. When the light was dim, David and Tommi never were short of new and glowing sparks. A sincere thank you to both of you for everything.

I would then like to extend my gratitude to the whole Department of Statistical Science at Duke, which has been more than a workplace these past four years. I am especially grateful to Amy Herring for her support and to Mike West, Peter Hoff, and Jim Clark, who kindly agreed to be on my defense committee. Thanks also to the Bocconi professors Daniele Durante, Igor Prünster, and Antonio Lijoi, who first lit my passion for statistics and later paved my way to Duke.

When I started my Ph.D., I was definitely not expecting that ecology would be my main focus. Thankfully, I had the pleasure of receiving constant help from all the researchers collaborating on the Lifeplan project[2]. My deep gratitude goes to all of them, in particular to Otso Ovaskainen, Tomas Roslin, and Panu Somervuo, whose insightful suggestions lead to the contributions in Chapters 3 and 4.

I also would like to extend my gratitude to my former managers Lee Richardson and Jacopo Soriano, and then to Chris Haulk and the whole YouTube Data Science team at Google, who have undoubtedly strengthened my confidence as a statistician at large.

One further important acknowledgment goes to the many people, friends, and colleagues, I encountered these years who made my stay in the United States a wonderful journey. Thanks in particular to Vittorio, Federico, Michele Caprio, Lorenzo, Jordan, Michele Peruzzi, Riccardo, Niccolò, Federica and Davide, Heather, Evan,

[2] Lifeplan is funded by the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement No 856506).

Betsy, Joseph, David, Michael, Shounak, Andy, Graham, Alex, Jennifer, Andrea, Rick, Emily, and Zeki.

My deepest thank you goes to my family at home in Italy, who sustained me throughout difficult times. Thanks to Mom, Dad, Elena, and then to Zia Grazia, Zio Nicola, Michela, Nonna Lina e Nonno Pietro.

Finally, I would like to thank my girlfriend Laura, to whom I dedicate this thesis, for all the support and love over these years, which proved stronger than the different time zones and the different continents.

1

Introduction

Species sampling models (Pitman, 1996) are a broad class of Bayesian discrete nonparametric priors that model the sequential appearance of distinct tags, called *clusters* or *distinct species*, in a sequence of labeled objects. Since the introduction of their arguably most famous member, the Dirichlet process, around 50 years ago (Ferguson, 1973), species sampling priors have been thoroughly investigated from a theoretical standpoint as models for random partitions and have found much success in a variety of applied mixture modeling settings, like regression analysis, hierarchical modeling, density estimation, clustering, and community detection to name a few. These rich theoretical and applied developments, however, have rarely appealed to ecologists, whose primary scientific investigation often revolves around the modeling of *actual* species and their behavior in nature at large. In this dissertation, we describe some theoretical, methodological, and applied Bayesian nonparametric contributions that we hope will both be of independent interest to statisticians in the field, and also open a path towards a broader use of species sampling model-based methods among practitioners in applied ecological settings.

Our first contribution is presented in Chapter 2, which focuses on species sampling priors from a mixture modeling perspective. In particular, we study how Dirichlet process mixtures are crucially sensitive to the choice of the so-called precision parameter. Our goal is to show how randomization of the precision through the use of a prior leads to greater robustness in inferential procedures, especially in terms of the posterior number of clusters estimated from the data. However, common choices of priors, such as the gamma distribution (Escobar and West, 1995), do not allow for transparent elicitation due to a lack of analytical results. For this reason, we introduce the novel *Stirling-gamma* prior, which makes the distribution of the number of clusters analytically tractable. Our theoretical investigation clarifies the reasons for the improved robustness. For instance, the number of clusters in a Dirichlet process with a Stirling-gamma prior on the precision follows approximately a negative binomial distribution, whereas a fixed precision leads to a Poisson-type behavior instead. Under specific choices of its hyperparameters, the Stirling-gamma has the important property of being conjugate to the law of the random partition of a Dirichlet process. This is particularly useful in applied settings where inference on the partition is of primary interest. We illustrate the above advantages in a community detection problem, where our goal is to infer the number of communities within a colony of ants from networks of individual ant-to-ant interactions.

In Chapter 3, we present a novel Bayesian methodology to model accumulation curves, which express the count of “new” species discovered as a function of the number of individuals observed. Specifically, it is of great interest for ecologists to both correctly fit the “in-sample” behavior of the curve, or *rarefaction* and to predict the “out-of-sample” trajectory, or *extrapolation*, which amounts to estimating the number of new additional species that would be observed if more samples were collected. Extrapolating the curve to infinity yields an estimate of the species richness, namely the total number of distinct species that live in a location. Our method is inspired by traditional species sampling models, such as the Dirichlet (Ferguson, 1973), the

Pitman–Yor (Pitman and Yor, 1997) and the Dirichlet-multinomial processes (Perman et al., 1992). Unfortunately, these models are either not sufficiently flexible for rarefactions and extrapolations or only admit an infinite number of species asymptotically. To deal with these issues, we introduce a Bayesian framework where the probability of discovering a new species at any given time is modeled through a survival function of a chosen latent random variable. The resulting law for the accumulation curve is a Poisson-binomial distribution, which allows for simple in- and out-of sample estimators, and both finite and infinite species richness depending on the shape of the survival function. We specifically focus on a three-parameter log-logistic class of survival functions, which includes the Dirichlet process discovery probability as a special case and whose parameters can be estimated via a constrained logistic regression. We test our proposal on data collected from a large fungal biodiversity study in Finland (Abrego et al., 2020).

Chapter 4 illustrates how species sampling models can be used as a building block to develop powerful classification tools when the true labels in the test set are not observed in training. Such a problem frequently arises in DNA barcoding tasks (Somervuo et al., 2017), which are often used to quantify biodiversity in a given area. Indeed, modern taxonomic identification methods leverage upon existing libraries of DNA barcodes to automatically annotate DNA sequences collected from field experiments. However, these libraries are often incomplete, as many species are unknown to science or do not have a reference barcode. Thus, the taxonomic novelty of a sequence needs to be accounted for when doing classification. To solve the issue, we develop BayesANT, a *Bayes*i*a*n *N*onparametric *t*axonomic classifier, which uses species sampling model priors to allow new taxa to be discovered at each taxonomic rank. Using a simple product multinomial likelihood with conjugate Dirichlet priors at the lowest rank, a highly efficient algorithm is developed to provide a probabilistic prediction of the taxa placement of each sequence at each rank. BayesANT is shown to have excellent performance when many sequences in the test set belong to unobserved taxa.

All the research presented in this dissertation has received funding from Project Lifeplan[1], whose aim is to map the current state of biodiversity across the globe. Every Chapter is co-authored with David B. Dunson and Tommaso Rigon, with the addition of Otso Ovaskainen for Chapter 3. As our overarching goal is to broaden the application of Bayesian nonparametric tools in ecology, much software was developed in `R` to facilitate usage[2]. A sampler for the Stirling-gamma distribution of Chapter 2 is available in the `ConjugateDP` package. Codes to perform all models described in Chapter 3 can be found in the `BNPvegan` package. The taxonomic classification algorithm of Chapter 4 is available in the `BayesANT` package. Proofs for the statements and additional simulation studies are presented in Appendix A, B and C.

[1] Lifeplan is funded by the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement No 856506).

[2] All code is publicly available in GitHub at `https://github.com/alessandrozito`

2

Bayesian nonparametric modeling of latent partitions via Stirling-gamma priors

2.1 Introduction

Discrete Bayesian nonparametric priors have been thoroughly investigated in recent decades motivated by their wide applicability in model-based clustering and density estimation problems. Suppose $Y_1, \ldots, Y_n$ are n observations taking values on $\mathbb{Y}$ and $f(y \mid x)$ is a density function on the same space, indexed by x. Then, a Bayesian nonparametric mixture model is defined through the following hierarchical representation

$$Y_i \mid X_i \overset{\text{ind}}{\sim} f(\cdot \mid X_i), \qquad X_i \mid \tilde{p} \overset{\text{iid}}{\sim} \tilde{p}, \qquad \tilde{p} \sim \mathscr{Q}, \qquad (i = 1, \ldots, n), \tag{2.1}$$

where $X_1, \ldots, X_n$ are latent random variables in $\mathbb{X}$, $\tilde{p}$ is a discrete random probability measure and $\mathscr{Q}$ represents its prior. Some notable instances of prior laws $\mathscr{Q}$ include the Pitman–Yor process (Perman et al., 1992; Pitman and Yor, 1997), Gibbs-type priors (Gnedin and Pitman, 2005; De Blasi et al., 2015), and normalized random measures with independent increments (Regazzini et al., 2003). Arguably, the most popular and widely employed discrete nonparametric prior is the Dirichlet process introduced by Ferguson (1973), due to its simplicity and analytical tractability.

The discreteness of $\tilde{p}$ induces a clustering of the observations by generating ties among the latent variables. More precisely, there will be $K_n = k$ distinct values among $X_1, \dots, X_n$, which partitions the statistical units $\{1, \dots, n\}$ into k clusters, say $C_1, \dots, C_k$. Hence, two statistical units i and i' belong to the same cluster, say the jth, if $i, i' \in C_j$ or, equivalently, if $X_i = X_{i'}$. Moreover, we will say that $\Pi_n = \{C_1, \dots, C_k\}$ is the random partition induced by $\tilde{p}$. In a Dirichlet process mixture model, the law of such a random partition Π_n is

$$\mathbb{P}(\Pi_n = \{C_1, \dots, C_k\} \mid \alpha) = \frac{\alpha^k}{(\alpha)_n} \prod_{j=1}^{k} (n_j - 1)!, \tag{2.2}$$

where $\alpha > 0$, with $(\alpha)_n = \alpha(\alpha + 1) \cdots (\alpha + n - 1)$ being the ascending factorial, and with $n_j = |C_j|$ being the number of elements in cluster C_j, so that $\sum_{j=1}^{k} n_j = n$. The quantity α is called *precision parameter* and, together with the sample size n, governs the law of the partition and the number of clusters K_n. In our motivating application, we rely on such a random partition mechanism to infer the latent communities in a colony of ant workers. Specifically, we model individual ant-to-ant interaction networks via stochastic block models (Nowicki and Snijders, 2001), which are a variant of the mixture model in (2.1). See Kemp et al. (2006); Geng et al. (2019); Legramanti et al. (2022) for other applications of discrete nonparametric priors in community detection tasks.

It has been pointed out by several scholars that Dirichlet process mixtures are particularly sensitive to the precision parameter (Escobar, 1994; Lijoi et al., 2007b; Booth et al., 2008). For instance, different values of α can lead to dramatically different posterior distributions of K_n, even when sufficient cluster separation is present in the data. Such a lack of robustness is problematic when the posterior partition is of inferential interest, such as in clustering and community detection.

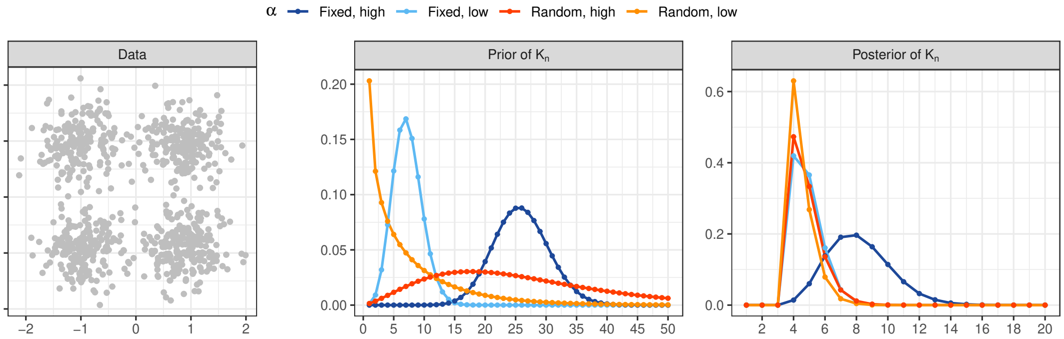

Figure 2.1: Left panel: 800 data points from a four-component mixture of normals. Center panel: probability mass function of the prior distribution of K_n under different choices of α. Parameters were set to have $\mathbb{E}(K_n) = 7.26$ and $\mathbb{E}(K_n) = 25.9$ in low and high cases, respectively, with $\alpha = 1$ and $\alpha = 5$ in the fixed cases. Right panel: posterior distribution of K_n estimated from the data via a Dirichlet process mixture. See Addendum II for details.

As we illustrate in Figure 2.1 with a simulated example, randomizing the precision through the use of a prior $\pi(\alpha)$ can attenuate this unpleasant behavior. Here, fixing $\alpha = 1$ as opposed to $\alpha = 5$ causes the posterior mode of K_n to shift from four to eight clusters, even if the data are generated from a mixture with four well-separated components. On the contrary, allowing α to be random induces more flexibility in the prior for K_n, and in turn, yields two posterior distributions that are similar to each other even when the means in the priors for α are far apart.

The most common choice of $\pi(\alpha)$ is the gamma distribution (Escobar and West, 1995). However, gamma priors lead to an analytically intractable prior over the partition. This prevents transparent elicitation of the prior hyperparameters and complicates the inclusion of available prior information on the clustering structure of the data. Moreover, while it has been shown that the distribution of K_n arising from a Dirichlet process can be approximated with a Poisson distribution when α is fixed, no such approximation is available for the random α case. We aim at filling this gap by introducing a novel prior over the Dirichlet process precision that (i) is simple and easily sampled from, (ii) makes the induced prior on K_n analytically tractable and

(iii) leads to an approximate negative binomial prior on the number of clusters. Our proposed prior for α has a novel distribution, which we refer to as *Stirling-gamma*, due to its connection with Stirling numbers and the gamma distribution. Under an appropriate logarithmic rescaling, the Stirling-gamma is equivalent to the gamma in a limiting case.

When α follows a Stirling-gamma prior, we will say that the random partition is from a *Stirling-gamma process*. This belongs to the larger class of Gibbs-type partition models, which are discrete nonparametric priors that enjoy several appealing theoretical properties. See for instance De Blasi et al. (2015). We provide several distributional results for the Stirling-gamma process. In particular, we show that the hyperparameters have an interpretable link with the induced law for the partition and the associated number of clusters. The resulting negative binomial-type behavior of the Stirling-gamma process, as opposed to the Poisson-type one of the Dirichlet process, helps explain the greater robustness of mixture models with random α.

The Stirling-gamma has the further fundamental advantage of being the conjugate prior to the law of the random partition of the Dirichlet process if one of its hyperparameters equals n. This happens because the distribution in equation (2.2) belongs to the class of natural exponential families, which always admit a conjugate prior (Diaconis and Ylvisaker, 1979). We illustrate how this conjugacy result further facilitates both posterior inferences on α and prior elicitation. The consequences of the prior dependency on n are thoroughly discussed. In particular, we show how the Stirling-gamma can be a useful prior when modeling independently repeated partitions of the same n statistical units, such as the ant worker interaction networks of our illustrative application.

The Chapter is organized as follows. Section 2.2 formally introduces the species sampling framework and presents the Stirling-gamma process. Section 2.3 focuses on the conjugate Stirling-gamma prior, while Section 2.4 shows the community detection application. Section 2.5 contains concluding remarks. The two additional Addendum Sections contain results on the Stirling-gamma coefficients, and details on the simulation in Figure 2.1. Proofs for the statements and details on the sampler for the Stirling-gamma are presented in Appendix A, which also includes an additional simulation study. To sample from the Stirling-gamma distribution, refer to the function `rSg` in the `ConjugateDP` package.

2.2 Distribution theory for Stirling-gamma processes

2.2.1 Background

Before introducing the Stirling-gamma distribution and the related process, we provide a probabilistic background on partition models that will be useful throughout the paper. Suppose that the latent variables X_i in model (2.1) belong to an infinite exchangeable sequence $(X_n)_{n \geqslant 1}$ and that they live in a complete and separable metric space $\mathbb{X}$ endowed with a Borel sigma-algebra $\mathscr{B}(\mathbb{X})$. The *species sampling models* introduced by Pitman (1996) provide a broad class of discrete nonparametric priors, defined as $\tilde{p} = \sum_{j=1}^{\infty} \tilde{p}_j \delta_{\xi_j}$ with $\sum_{j=1}^{\infty} \tilde{p}_j = 1$. Here, δ_x is the Dirac measure at x, while the ξ_js are drawn independently from a non-atomic *baseline distribution* P_0 on $\mathscr{B}(\mathbb{X})$ and are also independent from the random weights $\tilde{p}_j$. Since the realizations of a species sampling model are almost surely discrete, we have $\mathbb{P}(X_i = X_{i'}) > 0$ for any $i \neq i'$. As such, the latent variables $X_1, \dots, X_n$ will take on $K_n = k$ distinct values, called $X_1^*, \dots, X_k^*$, with frequencies $n_1, \dots, n_k$ and $\sum_{j=1}^{k} n_j = n$. This induces a random partition of the statistical units $\{1, \dots, n\}$ into groups $C_1, \dots, C_k$, where $C_j = \{i : X_i = X_j^*\}$ for $j = 1, \dots, k$. Traditionally, $X_1^*, \dots, X_k^*$ are also referred to as *distinct species* thanks to the metaphor discussed in Pitman (1996). Hence, we talk about *species sampling models.*

There exists a rich variety of exchangeable priors to model the random partition mechanism generating the clusters $C_1, \dots, C_k$. See Ghosal and van der Vaart (2017) for an extensive account. Among them, *Gibbs-type* processes (Gnedin and Pitman, 2005; De Blasi et al., 2015) form a particularly rich class. We say that the law of $\tilde{p}$ is of Gibbs-type if

$$\mathbb{P}(\Pi_n = \{C_1, \ldots, C_k\}) = V_{n,k} \prod_{j=1}^{k} (1-\sigma)_{n_j - 1}, \tag{2.3}$$

where $\sigma < 1$ and the coefficients $V_{n,k}$ satisfy the forward recursion $V_{n,k} = (n - \sigma k)V_{n+1,k} + V_{n+1,k+1}$ for all $k = 1, \ldots, n$ and $n \geqslant 1$, with $V_{1,1} = 1$. Equation (2.3) is the so-called *exchangeable partition probability function* of the process (Pitman, 1996). This depends on the cluster frequencies through a product structure, which implies that Gibbs-type priors are a special instance of product partition models (Hartigan, 1990; Barry and Hartigan, 1992; Quintana and Iglesias, 2003). A detailed list of examples of models that follow equation (2.3) is presented in Appendix B. The coefficients $V_{n,k}$ determine the system of predictive equations of the random partition Π_n, that is

$$\mathbb{P}(X_{n+1} \in A \mid X_1, \ldots, X_n) = \frac{V_{n+1,k+1}}{V_{n,k}} P_0(A) + \frac{V_{n+1,k}}{V_{n,k}} \sum_{j=1}^{k} (n_j - \sigma) \delta_{X_j^*}(A), \tag{2.4}$$

for $n \geqslant 1$ and every $A \in \mathscr{B}(\mathbb{X})$. The $(n+1)$st latent parameter X_{n+1} is drawn from the baseline P_0 with probability $V_{n+1,k+1}/V_{n,k}$, and is equal to one of the previous X_j^* with probability $V_{n+1,k+1}(n_j - \sigma)/V_{n,k}$. Specifically, sampling X_{n+1} from the baseline automatically generates a new cluster, or a *new species*, due to the diffuse nature of P_0. Refer to De Blasi et al. (2015) for an overview, and to Chapter 3 for an interpretation of such mechanism for an applied ecological perspective.

When $\sigma = 0$ and $V_{n,k} = \alpha^k/(\alpha)_n$ in equation (2.3), one recovers the exchangeable partition probability function of a Dirichlet process in equation (2.2). A more robust specification can be obtained by introducing a prior for α. In this case, the resulting distribution is

$$\mathbb{P}(\Pi_n = \{C_1, \ldots, C_k\}) = V_{n,k} \prod_{j=1}^{k} (n_j - 1)!, \quad V_{n,k} = \int_{\mathbb{R}_+} \frac{\alpha^k}{(\alpha)_n} \pi(\alpha) \mathrm{d}\alpha, \tag{2.5}$$

which has more flexibility through varying the hyperparameters of $\pi(\alpha)$. Gnedin and Pitman (2005) show that every Gibbs-type prior with $\sigma = 0$ is uniquely characterized by equation (2.5). Commonly adopted priors $\pi(\alpha)$, such as the gamma distribution proposed by Escobar and West (1995), do not lead to an analytically tractable form for $V_{n,k}$. This is a crucial point because $V_{n,k}$ are the key quantities that determine the distribution of the number of clusters, that is

$$\mathbb{P}(K_n = k) = V_{n,k} |s(n,k)|, \qquad (k = 1, \ldots, n), \tag{2.6}$$

where $|s(n,k)|$ are the signless Stirling number of the first kind (Charalambides, 2005). Refer to Antoniak (1974) and Gnedin and Pitman (2005) for derivations. Thus, our goal is to develop a prior whose hyperparameters have a clear and interpretable link with the distribution of K_n in equation (2.6). In what follows, we show how this can be achieved using a Stirling-gamma prior.

2.2.2 The Stirling-gamma distribution

In this Section, we introduce the Stirling-gamma distribution and describe its properties.

Definition 1. *A positive random variable follows a Stirling-gamma distribution with parameters $a, b > 0$ and $m \in \mathbb{N}$ satisfying $1 < a/b < m$, if its density function is*

$$p(\alpha) = \frac{1}{\mathcal{S}_{a,b,m}} \frac{\alpha^{a-1}}{\{(\alpha)_m\}^b}, \qquad \mathcal{S}_{a,b,m} = \int_{\mathbb{R}_+} \frac{\alpha^{a-1}}{\{(\alpha)_m\}^b} \, \mathrm{d}\alpha.$$

We will write $\alpha \sim \mathrm{Sg}(a, b, m)$.

The name of the Stirling-gamma distribution stems from the presence of the ascending factorial in the density function, whose polynomial expansion defines Stirling numbers of the first kind (Charalambides, 2005), and the following connection with the gamma distribution.

Proposition 2. *Let* $\alpha \sim \mathrm{Sg}(a, b, m)$*. Then, the following convergence in distribution holds:*

$$\alpha \log m \to \gamma, \quad \gamma \sim \mathrm{Ga}(a-b, b), \quad m \to \infty.$$

In the above statement, $\mathrm{Ga}(a_0, b_0)$ denotes the gamma distribution with mean a_0/b_0 and variance a_0/b_0^2. Proposition 2 has two fundamental implications. The first is that the density of the Stirling-gamma $\mathrm{Sg}(a, b, m)$ progressively resembles that of $\mathrm{Ga}(a-b, b\log m)$ as m becomes larger. The second is that $\alpha \to 0$ in probability as $m \to \infty$ with a logarithmic rate of convergence via a direct application of Slutzky's theorem. Both properties are illustrated in Figure 2.2, which displays the probability density function of the two distributions for varying values of m and b when $a = 5$. In particular, high values for a/b require a larger m to make the two densities indistinguishable. Both distributions progressively shift towards zero as m increases. However, the right tail of the Stirling-gamma is heavier than the one of the gamma because it is a heavy-tailed distribution. We provide a formal proof in Appendix A.

The density function of a Stirling-gamma is proper, namely $\mathcal{S}_{a,b,m} < \infty$, if only if $1 < a/b < m$, as shown in Appendix A. Interestingly, the normalizing constant $\mathcal{S}_{a,b,m}$ is the key to calculating the moments of the distribution, which are obtained as follows.

Proposition 3. *Let* $\alpha \sim \mathrm{Sg}(a, b, m)$ *and suppose that* $0 < s < mb - a$*. Then*

$$\mathbb{E}(\alpha^s) = \frac{\mathcal{S}_{a+s,b,m}}{\mathcal{S}_{a,b,m}}.$$

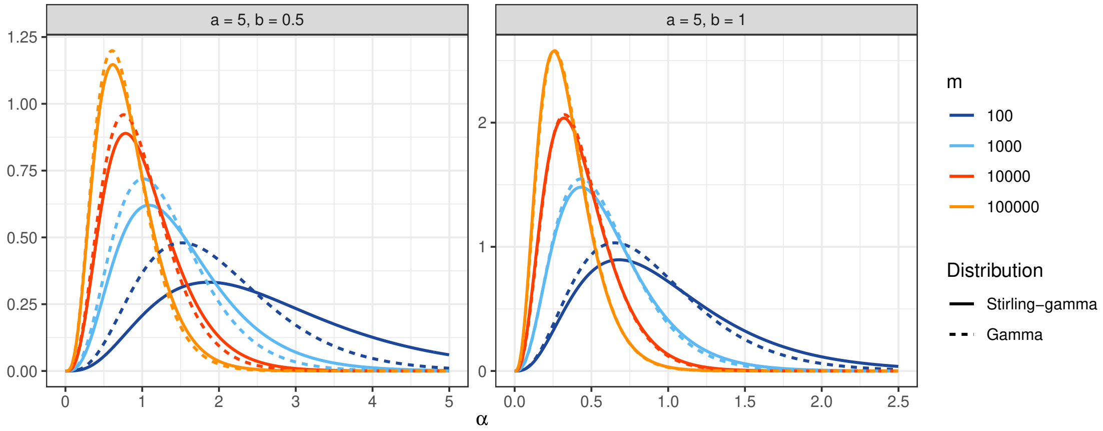

FIGURE 2.2: Probability density function of a Stirling-gamma $\mathrm{Sg}(a, b, m)$, depicted by the solid lines, and a $\mathrm{Ga}(a - b, b \log m)$, indicated by the dashed lines, for varying values of m and b, and $a = 5$.

When $s > mb - a$, instead, then one has $\mathbb{E}(\alpha^s) = \infty$. In general, explicit analytic expressions for the moments are not available. One possibility is to approximate $\mathcal{S}_{a,b,m}$ and, consequently, $\mathbb{E}(\alpha^s)$ via Monte Carlo integration since samples from the Stirling-gamma can be drawn efficiently; see the Supplementary material. Alternatively, when m is large, we have that $\mathbb{E}(\alpha) = \mathcal{S}_{a+1,b,m}/\mathcal{S}_{a,b,m}$ is roughly equal to $(a/b - 1)/\log m$ and that $\mathcal{S}_{a,b,m} \approx (b \log m)^{a-b}/\Gamma(a - b)$ by means of Proposition 2. Also, in the special instance where $a, b \in \mathbb{N}$, we can express $\mathcal{S}_{a,b,m}$ analytically as an alternating sum of logarithms, as shown in Theorem 12 in the Appendix.

2.2.3 Random partitions via Stirling-gamma priors

When the precision parameter of a Dirichlet process follows a Stirling-gamma distribution $\alpha \sim \mathrm{Sg}(a, b, m)$, we have a *Stirling-gamma process*. As described in Section 2.2.1, this is a member of the Gibbs-type family with $\sigma = 0$. Thus, the associated exchangeable partition probability function is readily available from the results of Gnedin and Pitman (2005).

Theorem 4. *The exchangeable partition probability function of a Stirling-gamma* *process with* $\alpha \sim \mathrm{Sg}(a, b, m)$ *is*

$$\mathbb{P}(\Pi_n = \{C_1, \ldots, C_k\}) = \frac{\mathscr{V}_{a,b,m}(n,k)}{\mathscr{V}_{a,b,m}(1,1)} \prod_{j=1}^{k} (n_j - 1)!,$$

where the coefficients are equal to

$$\mathscr{V}_{a,b,m}(n,k) = \int_{\mathbb{R}_+} \frac{\alpha^{a+k-1}}{\{(\alpha)_m\}^b (\alpha)_n} \mathrm{d}\alpha.$$

It is easy to see that the Gibbs-type coefficients are $V_{n,k} = \mathscr{V}_{a,b,m}(n,k)/\mathscr{V}_{a,b,m}(1,1)$, for $k = 1, \ldots, n$ and $n \geqslant 1$, with $V_{1,1} = 1$, and that the forward recursion is satisfied since $\mathscr{V}_{a,b,n}(n,k) = n\mathscr{V}_{a,b,n}(n+1,k) + \mathscr{V}_{a,b,n}(n+1,k+1)$. Relying on similar reasoning as the one used for prior coefficients in Definition 1, we can also derive an explicit analytical form for $\mathscr{V}_{a,b,m}(n,k)$ when $a, b \in \mathbb{N}$. We show this in Theorem 14 in Addendum I, which, combined with Theorem 12, implies that $V_{n,k}$ can be expressed as ratios of alternating sums of logarithms after noticing that $\mathscr{V}_{a,b,m}(1,1) = \mathcal{S}_{a,b,m}$.

By being a genuine Gibbs-type prior, the Stirling-gamma process admits an urn scheme representation of the form in equation (2.4). In particular, the latent variables $(X_n)_{n\geqslant 1}$ abide the following generative mechanism:

$$\mathbb{P}(X_{n+1} \in A \mid X_1, \ldots, X_n) = \frac{\mathscr{V}_{a,b,m}(n+1,k+1)}{\mathscr{V}_{a,b,m}(n,k)} P_0(A) + \frac{\mathscr{V}_{a,b,m}(n+1,k)}{\mathscr{V}_{a,b,m}(n,k)} \sum_{j=1}^{k} n_j \delta_{X_j^*}(A), \tag{2.7}$$

for $n \geqslant 1$ and for every $A \in \mathscr{B}(\mathbb{X})$. The fundamental difference between the predictive scheme in equation (2.7) and the one arising from the generic distribution in (2.5) lies in the fact that the hyperparameters of the Stirling-gamma prior are interpretable in terms of the induced number of clusters in the latent partition. We elucidate this with the following key result.

Theorem 5. *Let $\alpha \sim \mathrm{Sg}(a, b, m)$ and $\mathcal{D}_{a,b,m} = \mathbb{E}\{\sum_{i=0}^{m-1} \alpha^2/(\alpha+i)^2\}$. The number of clusters K_m obtained from the first m random variables $X_1, \ldots, X_m$ generated from the predictive scheme in equation* (2.7) *is distributed as*

$$\mathbb{P}(K_m = k) = \frac{\mathscr{V}_{a,b,m}(m,k)}{\mathscr{V}_{a,b,m}(1,1)} |s(m,k)|, \tag{2.8}$$

for $k = 1, \ldots, m$, with mean and variance equal to

$$\mathbb{E}(K_m) = \frac{a}{b}, \qquad \mathrm{var}(K_m) = \frac{b+1}{b}\left(\frac{a}{b} - \mathcal{D}_{a,b,m}\right).$$

In Section 2.2.4 we further show that $\mathcal{D}_{a,b,m} \approx 1$ for m large enough. The above statement suggests that hyperparameters a, b and m have an important meaning: when $\alpha \sim \mathrm{Sg}(a, b, m)$, the first m statistical units $\{1, \ldots, m\}$ arising from the Stirling-gamma process identify a/b clusters on average, with variance inversely related to b. For this reason, we can refer to m as a hypothetical *reference sample size*, a/b as a *location*, and b as a *precision*. Theorem 5 also provides an explicit motivation for why the hyperparameters of the Stirling-gamma must satisfy $1 < a/b < m$ as in Definition 1: having $a/b = 1$ is equivalent to having $\mathbb{E}(K_m) = 1$, which corresponds to a Dirichlet process where $\alpha \to 0$. On the contrary, setting $a/b = m$ leads to $\mathbb{E}(K_m) = m$, meaning that every observation identifies a new cluster. This is the case of a Dirichlet process where $\alpha \to \infty$. Setting $1 < a/b < m$ avoids both degenerate behaviors.

The results in Theorem 5 hold exclusively at the mth sample from the Stirling-gamma process. For arbitrary *fixed* values of a, b and m the distribution of the number of cluster K_n is given in equation (2.6), whose moments are not available in closed form. It is well known that the number of clusters K_n arising at a generic nth draw from the predictive scheme in equation (2.7) maintains the logarithmic divergence typical of the Gibbs-type processes with $\sigma = 0$. This is because

$K_n/\log n \to \alpha \sim \mathrm{Sg}(a, b, m)$ in distribution as $n \to \infty$, as discussed in Pitman (1996). On the other hand, one key aspect of Theorem 5 is that the expectation of the number of clusters among $X_1, \dots, X_m$, obtained from equation (2.7), is *independent* of m. Indeed, it will be shown in Section 2.2.4 that the distribution of K_m provided in equation (2.8) converges to a finite discrete random variable as $m \to \infty$. This is a consequence of Proposition 2 and the diverging nature of K_n discussed above: while K_n diverges at a logarithmic rate in n, the Stirling-gamma prior makes α approach zero logarithmically in m. The two rates perfectly compensate each other when the *observed* sample size n reaches the *reference* sample size m.

2.2.4 Robustness properties

In this Section, we investigate the behavior of the number of clusters of the Stirling-gamma process under a large reference sample size. Interestingly, if m itself is chosen large, we are able to show that K_m approaches a well-known distribution.

Theorem 6. *Under the same assumptions of Theorem 5, the following convergence in distribution holds:*

$$K_m \to K_\infty, \quad K_\infty \sim 1 + \mathrm{Negbin}\left(a - b, \frac{b}{b+1}\right), \quad m \to \infty.$$

The negative binomial distribution in Theorem 6 is parametrized so that

$$\mathbb{E}(K_\infty) = \frac{a}{b}, \qquad \mathrm{var}(K_\infty) = \frac{b+1}{b}\left(\frac{a}{b} - 1\right).$$

Hence, the quantity $\mathcal{D}_{a,b,m}$ defined in Theorem 5 converges to one when $m \to \infty$. Thus, Theorem 6 provides a reliable approximation for the prior distribution of the number of clusters. The same result is maintained when $\alpha \sim \mathrm{Ga}(a - b, b \log m)$. This should not come as a surprise considering the asymptotic equivalence discussed in Proposition 2.

In view of the above Theorem, it is natural to draw a comparison between the Stirling-gamma process and the Dirichlet process. To mimic the behavior of α under a Stirling-gamma prior, we study the number of clusters from a Dirichlet process at the reference sample size m when $\alpha = \lambda/\log m$, with $\lambda > 0$ being a positive constant. The large m behavior is illustrated in the next Proposition, where $\text{Po}(\lambda)$ denotes the Poisson distribution with mean λ.

Proposition 7. *Let $X_1, \dots, X_m$ be the first m realizations from a Dirichlet process, obtained by setting $V_{n,k} = \alpha^k/(\alpha)_n$ and $\sigma = 0$ in equation* (2.4)*. If $\alpha = \lambda/\log m$ for some $\lambda > 0$, then the following convergence in distribution holds:*

$$K_m \to K_\infty, \qquad K_\infty \sim 1 + \text{Po}(\lambda), \qquad m \to \infty.$$

Similar Poisson-type behaviors for the number of clusters in the Dirichlet process have already been shown in the literature. See for example Proposition 4.8 in Ghosal and van der Vaart (2017). Theorem 6 and Proposition 7 suggest a theoretical reason for why a Dirichlet process with random precision is more flexible than the fixed precision counterpart. When α is kept fixed and sufficiently small, the number of clusters is approximately distributed as a Poisson, whose mean and variance are uniquely controlled by one parameter. On the contrary, choosing a Stirling-gamma prior with large m induces an approximately negative-binomial prior for K_n, leading to much greater robustness to the prior expectation for K_n, as illustrated in Figure 2.1.

2.3 Conjugate inference under Stirling-gamma priors

In this Section, we illustrate how the Stirling-gamma distribution has the further important property of being *conjugate* to the law of the partition of the Dirichlet process. As we show in Proposition 8, this happens when the reference sample size m is set equal to the number of data points n in equation (2.2).

Proposition 8. *Suppose we observe a partition* Π_n *distributed according to the Dirichlet process in* (2.2) *and let* $\alpha \sim \mathrm{Sg}(a, b, n)$. *Then,*

$$(\alpha \mid \Pi_n = \{C_1, \ldots, C_k\}) \sim \mathrm{Sg}(a + k, b + 1, n).$$

The same result can be derived by conditioning on $K_n = k$ alone as in Escobar and West (1995) because of its sufficiency for α. The above conjugacy simplifies computations when sampling from the posterior distribution in a Dirichlet process mixture model with random precision, which in the case of the gamma prior requires a data augmentation step. Under the conjugate Stirling-gamma prior, elicitation is straightforward by virtue of Theorems 5 and 6. Thus, one can transparently tune the Stirling-gamma prior by leveraging upon information available on the clustering structure of the n observations through choices of a and b.

The existence of the conjugate Stirling-gamma prior follows directly from the results of Diaconis and Ylvisaker (1979) for natural exponential families, which the partition law of the Dirichlet process is a member of. Nevertheless, the prior dependency on n has some important consequences on the process, which must be handled with care. In particular, while the distribution in Theorem 4 remains the one of a finitely exchangeable product partition model, the Gibbs-type recursion characterizing the coefficients $V_{n,k}$ no longer holds. Namely, $V_{n,k} \neq nV_{n+1,k} + V_{n+1,k+1}$. This breaks the predictive scheme of equation (2.7), causing the sequence to lose the *projectivity* property typical of species sampling models (Lee et al., 2013). In other terms, the distribution in Theorem 4 under n observations does not coincide with the one obtained by marginalizing out the $(n + 1)$th sample from the same distribution under $n + 1$ data points. This is a limitation when one is interested in extrapolating inferences from a sample to the general population, but less so on clustering problems where out-of-sample predictions are not the main focus (Betancourt et al., 2020).

The lack of projectivity of the sequence under $m = n$, however, is less relevant in settings where n plays the role of the dimension of the data rather than the number of observed data points. We illustrate this by introducing the following *population of partitions* framework. Let $\Pi_{n,1}, \ldots, \Pi_{n,N}$ denote N independent and identically distributed realizations of a random partition of the same units $\{1, \ldots, n\}$ from an exchangeable partition probability function. If each partition is from a Dirichlet process with precision α, then we have

$$\mathbb{P}(\Pi_{n,s} = \{C_{1,s}, \ldots, C_{k_s,s}\} \mid \alpha) = \frac{\alpha^{k_s}}{(\alpha)_n} \prod_{j=1}^{k_s} (n_{j,s} - 1)!, \quad (s = 1, \ldots, N), \tag{2.9}$$

where $n_{j,s} = |C_{j,s}|$ is the number of elements in the jth cluster $C_{j,s}$ within the sth partition, and k_s is the associated number of clusters. The model in equation (2.9) is suitable for instances where, for example, we measure the interactions among the same n nodes of a network multiple times. Similar data often occur in neuroscience studies, where the same n brain regions are scanned for N different individuals (Durante et al., 2017), or in ecology, where the interactions among n species are recorded for N days (Mersch et al., 2013). The inferential goal of model (2.9) is to retrieve the network-specific partition through a shared Dirichlet process precision parameter. Then, the following Theorem holds.

Theorem 9. *Let $\Pi_{n,1}, \ldots, \Pi_{n,N}$ be independent and identically distributed realizations from equation* (2.9). *If $\alpha \sim \mathrm{Sg}(a, b, n)$, then*

$$(\alpha \mid \Pi_{n,1}, \ldots, \Pi_{n,N}) \sim \mathrm{Sg}\left(a + \sum_{s=1}^{N} k_s, b + N, n\right).$$

It is straightforward to notice that Proposition 8 is retrieved by letting $N = 1$ in the above. In light of Theorem 9, we can also derive the classic Bayesian decomposition of the posterior mean as a weighted average between the observed data and the prior. Recall that $\mathbb{E}(K_n \mid \alpha) = \sum_{i=0}^{n-1} \alpha/(\alpha + i)$ is the conditional mean for the number of clusters generated by a Dirichlet process over partitions of the units $\{1,\dots,n\}$, and that $\mathbb{E}(K_n) = \mathbb{E}\{\mathbb{E}(K_n \mid \alpha)\} = a/b$ thanks to the law of the iterated expectation. Then, the next Proposition holds.

Proposition 10. *Under the same setting of Theorem 9, we have*

$$\mathbb{E}\left(\sum_{i=0}^{n-1} \frac{\alpha}{\alpha+i} \mid \Pi_{n,1},\dots,\Pi_{n,N}\right) = \frac{b}{b+N}\frac{a}{b} + \frac{N}{b+N}\bar{k},$$

where $\bar{k} = N^{-1}\sum_{s=1}^{N} k_s$ *is the average number of clusters observed across the partitions.*

The above statement is a direct consequence of the conjugacy of the Stirling-gamma prior under $m = n$. See Diaconis and Ylvisaker (1979) and the Supplementary material for details.

Remark 11. *There exists a rich variety of algorithms to sample from the posterior distribution of the mixture model in* (2.1)*. For Gibbs-type processes, one popular approach lies in the class of marginal samplers, which rely on the sequential predictive scheme of equation* (2.4)*. See Escobar and West (1995); Neal (2000) for examples. In light of its hierarchical construction, inference under the Stirling-gamma process mixture model can be performed under the same marginal scheme of the Dirichlet process, with an additional sampling step for* α*. Such a step is provided by Proposition 8 or Theorem 9 depending on the setting. In both cases, the conditioning is with respect to the last sampled partition at the given iteration.*

2.4 Inferring communities in ant interaction networks

We now illustrate how modeling the precision parameter α via a Stirling-gamma prior in a Dirichlet process mixture as opposed to keeping it fixed yields a more robust estimate of the posterior partition. We specifically consider the problem of

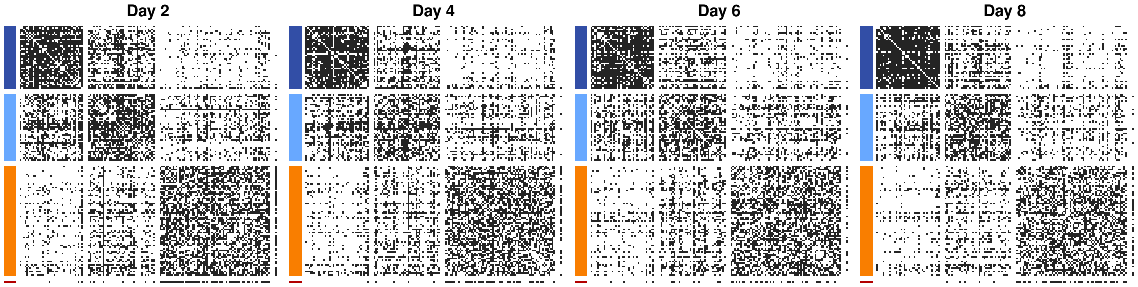

FIGURE 2.3: Binary ant-to-ant interaction networks in a colony *Camponotus fellah* observed in four different days. Each node is an ant, and black points denote edges. The colors on the left indicate the three groups of workers, namely foragers (dark blue), cleaners (light blue), and nurses (orange). The bottom red node indicates the queen.

detecting community structures in a colony of ant workers by modeling daily ant-to-ant interaction networks via stochastic block models (Nowicki and Snijders, 2001). The data were collected by Mersch et al. (2013) by continuously monitoring six colonies of the ant *Camponotus fellah* through an automated video tracking system over a period of 41 days. Each day yielded a weighted adjacency matrix whose edges contain the number of individual interactions between workers. In this analysis, we model a binary version of the data from days two, four, six, and eight for the fifth colony ($n = 149$), where $Y_{i,j,s}$ equals one if ant i and ant j in network s interacted more than five times, and zero otherwise. In line with the setting proposed by Theorem 9, we use the Stirling-gamma process to independently model the $N = 4$ latent partitions of the same $n = 149$ ants. Figure 2.3 reports the binary adjacency matrices recording ant interactions. Rows and columns have been sorted according to the three social organization groups retrieved by Mersch et al. (2013), namely foragers, cleaners, and nurses. The last group is composed of younger individuals and forms a stronger connection with the queen.

In order to perform community detection on each of the four networks, we rely on a stochastic block model formulation. This is a variant of the mixture of equation (2.1) of the Introduction, which is best rewritten with the help of auxiliary variables representing cluster assignment as follows. Given a random partition of the nodes $\Pi_{n,s} = \{C_{1,s}, \ldots, C_{k_s,s}\}$ in s, call $Z_{i,s}$ an auxiliary variable so that $Z_{i,s} = h$ if the node $i \in C_{h,s}$, for $i = 1, \ldots, n$. The probability of detecting an edge between nodes i and j in network s is specified as

$$\mathbb{P}(Y_{i,j,s} = 1 \mid Z_{i,s} = h, Z_{j,s} = h', \nu) = \nu_{h,h',s}, \quad \nu_{h,h',s} \sim \mathrm{Be}(1, 1). \tag{2.10}$$

Here, $\nu_{h,h',s} \in \nu = (\nu_{1,1,1}, \ldots,)$ is the edge probability in the block identified by clusters $C_{h,s}$ and $C_{h',s}$, and $\mathrm{Be}(a_0, b_0)$ is the Beta distribution with mean $a_0/(a_0 + b_0)$. We assume no node self-relation, thus ignoring the diagonal entries $Y_{i,i,s}$. By modeling the latent partition via the Dirichlet process prior $\mathbb{P}(\Pi_{n,s} \mid \alpha)$ as in equation (2.2), we can flexibly find a grouping of the nodes with a similar edge distribution and thus infer the number of ant worker communities without pre-specifying an upper bound to the number of clusters. See Legramanti et al. (2022) and references therein for a description of the posterior sampling algorithm.

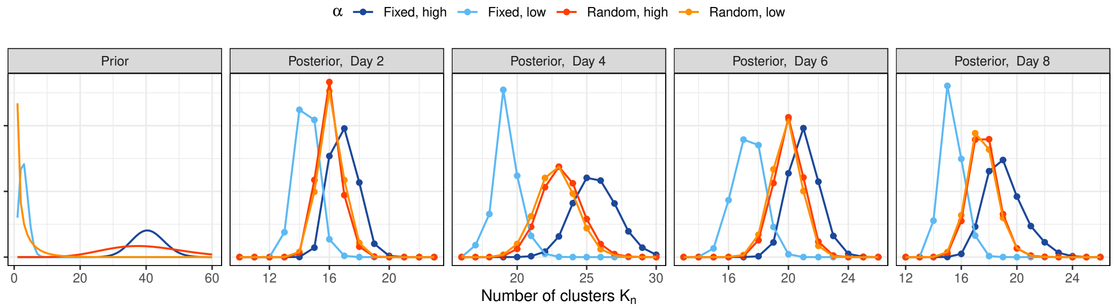

Figure 2.4: Prior and posterior distribution of the number of clusters detected in the ant-to-ant binary interaction networks. Different colors refer to the four models tested. "Fixed, low" (light blue) refer to case (i) when $\alpha = 0.4$. "Fixed, high" (dark blue) is case (ii) when $\alpha = 18$. "Random, low" (light orange) is case (iii) with $\alpha \sim \mathrm{Sg}(0.6, 0.2, 149)$, and "Random, high" refer to case (iv) with $\alpha \sim \mathrm{Sg}(8, 0.2, 149)$.

Our intent is to investigate the impact that different choices of α have on each posterior partition from the model in equation (2.10). In particular, we compare a Dirichlet process mixture with (i) $\alpha = 0.4$ and (ii) $\alpha = 18$, against Stirling-gamma processes with (iii) $\alpha \sim \mathrm{Sg}(0.6, 0.2, 149)$, and (iv) $\alpha \sim \mathrm{Sg}(8, 0.2, 149)$. The hyperparameters in models (i) and (iii) are chosen such that $\mathbb{E}(K_n) = 3$ so as to incorporate the *a priori* knowledge of the three groups described by Mersch et al. (2013). To check for posterior robustness, cases (ii) and (iv), instead, have $\mathbb{E}(K_n) = 40$. As is evident from the leftmost panel of Figure 2.4, the Stirling-gamma prior enables additional vagueness to K_n as a direct consequence of choosing $b = 0.2$. We obtain the posterior partition in each model by running a collapsed Gibbs sampler as in Legramanti et al. (2022) for 40,000 iterations, treating the first 10,000 as burn-in. The full conditional for α in both Stirling-gamma processes is provided by Theorem 9, setting $N = 4$ and $n = 149$. The resulting effective sample size for α in cases (iii) and (iv) is $11,932.22$ and $19,422.59$, indicating good mixing. Figure 2.4 plots the posterior distribution for the number of retrieved clusters K_n in each network. We see that there exists a non-negligible difference between the two posteriors where α is kept fixed, leading to under- and over-clustering in cases (i) and (ii), respectively. On the contrary, making α random with a sufficiently vague Stirling-gamma prior retrieves a virtually identical posterior irrespective of the induced prior on K_n. This is due to the additional flexibility granted by the Stirling-gamma, which enables the model to infer α from the data.

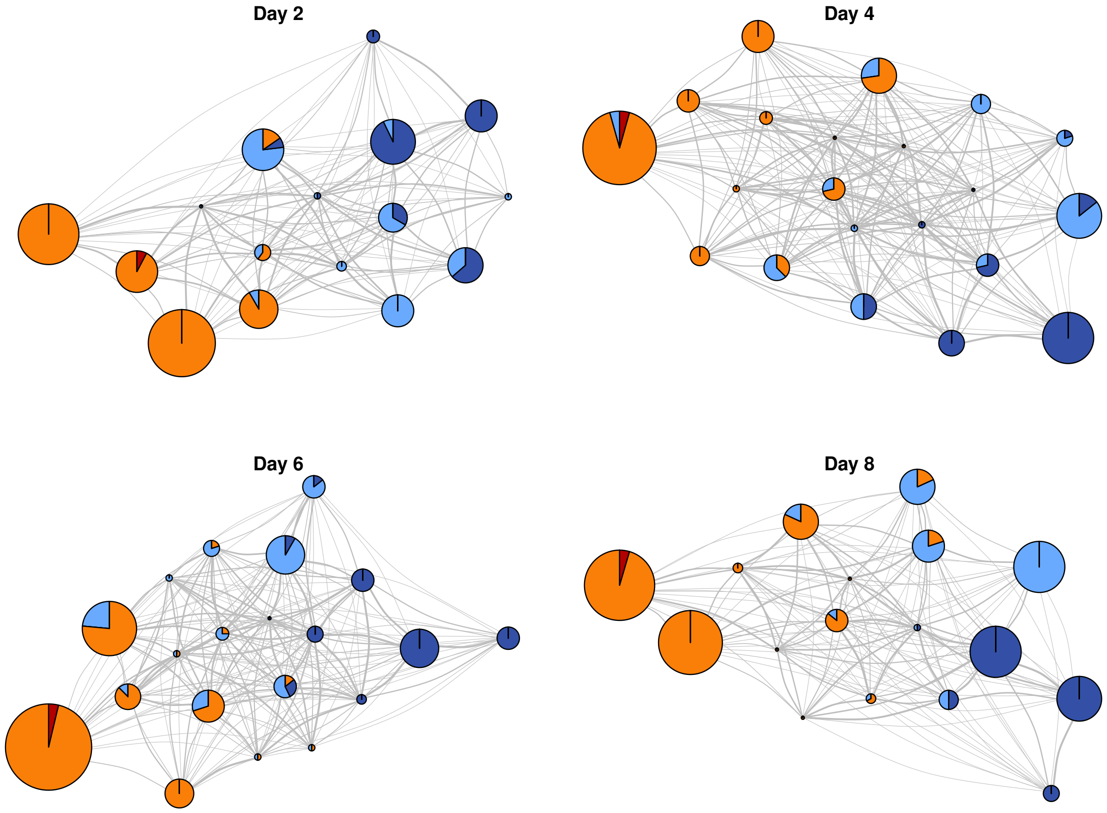

Figure 2.5: Network representation of the inferred partition in the four networks displayed in Figure (2.3). Nodes represent the retrieved clusters, with size determined by the number of ants they contain. Colors reflect the composition of each cluster according to the groups identified by Mersch et al. (2013): foragers (dark blue), cleaners (light blue), and nurses (orange). The queen is indicated in red. We obtain the node positions through force-directed placement (Fruchterman and Reingold, 1991). The width of the connections is determined by the posterior mean for the estimated block probabilities $\nu_{h,h',s}$, ignoring the ones below 0.1 for aesthetic reasons.

To further investigate the communities retrieved by our model, we look at the posterior obtained from the Stirling-gamma process in model (iii). As we can see from Figure 2.4, the average number of clusters detected in each day is much larger than the original grouping suggested by Mersch et al. (2013). As such, the stochastic block model in equation (2.10) recovers a more complex ant organization than the one originally proposed, effectively detecting worker sub-communities. Figure 2.5

displays the posterior partitions obtained by minimizing the variation of information metric leveraging on the approach of Wade and Ghahramani (2018). We observe 15, 21, 20, and 17 clusters on days 2, 4, 6 and 8, respectively. Such differences are due to the within-day variability in worker interactions. However, the core structure of the social organization remains similar across days, with the detection of sub-communities uniquely characterized by nurses (in orange) and foragers (in dark blue). Cleaners (in light blue) are instead co-clustered with the other two groups as they play a fundamental role in handling the passage of information within the colony. Finally, in each day we are able to detect the sub-community of nurses which interacts the most with the queen.

2.5 Discussion

Our proposed Stirling-gamma prior was motivated by improving robustness to prior choice and transparency in prior elicitation in Dirichlet process mixture models. Fixing the precision parameter is a poor choice in most applications, since it implies a highly informative prior for the induced number of clusters. While the usual gamma prior can improve robustness to one's prior guess for the number of clusters, the implications of the gamma choice are unclear due to the lack of an analytically tractable form for the induced partition prior. The Stirling-gamma has the dual advantages of being heavier-tailed than the gamma, leading to improved robustness to prior elicitation, while also being much more transparent in terms of the induced clustering prior.

More broadly, the Stirling-gamma is of interest as a new heavy-tailed distribution having positive support. There are multiple other application areas in which this new distribution may be useful. For example, the Stirling-gamma could be used as the choice of exponential family distribution within a generalized linear model framework when the common log-normal or gamma choices lack sufficiently heavy tails for the data at hand. Alternatively, noting that the Dirichlet distribution arises by normalizing independent gamma random variables, one could obtain an alternative distribution on the probability simplex by normalizing Stirling-gamma random variables. This new distribution may be ideal at characterizing the case in which there are a small proportion of large probabilities with the remaining concentrated near zero; a common desirable behavior for shrinkage priors on the simplex.

When the reference sample size m of a Stirling-gamma prior $\alpha \sim \mathrm{Sg}(a, b, m)$ diverges, the total number of clusters generated from a Stirling-gamma process as $n \to \infty$ is negative binomial-distributed. In future work, it will be interesting to study the relationship with models that choose a negative binomial prior directly for the number of components in a finite mixture; refer, for example, to Miller and Harrison (2018) and to the literature on Gibbs-type processes in which the number of clusters is a finite random quantity (Gnedin and Pitman, 2005; De Blasi et al., 2015).

2.6 Addendum I: Closed-form expressions for Stirling-gamma coefficients

We hereby show how the coefficients $\mathcal{S}_{a,b,m}$ and $\mathscr{V}_{a,b,m}(n,k)$ introduced in Definition 1 and Theorem 4 admit an explicit form. These depend on complete exponential Bell polynomials, which are defined as follows. Given the variables $x_1, \ldots, x_s$ for $s \geqslant 1$, the sth complete exponential Bell polynomial is

$$B_s(x_1, \ldots, x_s) = \sum_{(i_1,\ldots,i_s)\in I_s} \frac{s!}{i_1! i_2! \cdots i_s!} \left(\frac{x_1}{1!}\right)^{i_1} \left(\frac{x_2}{2!}\right)^{i_2} \cdots \left(\frac{x_s}{s!}\right)^{i_s}, \tag{2.11}$$

where I_s is the set of all non-negative integers $\{i_1, \ldots, i_s\}$ that satisfy the equality constraint $i_1 + 2i_2 + \ldots + s i_s = s$. See Charalambides (2005) for details. To simplify readability, we also introduce the polynomials

$$\mathscr{S}_{b,j}(x_1,\dots,x_b) = \sum_{s=1}^{b} \frac{B_{b-s}(x_1,\dots,x_{b-s})}{(b-s)!}\phi_s(j), \quad \phi_s(j) = \begin{cases} -\log j, & s=1, \\ j^{1-s}/(s-1) & s=2,3,\dots, \end{cases} \tag{2.12}$$

where $B_0(x_0)=1$ and $B_s(x_1,\dots,x_s)$ is defined in equation (2.11). Then, the following result holds.

Theorem 12. *If $a,b\in\mathbb{N}$, then*

$$\mathcal{S}_{a,b,m} = \sum_{j=1}^{m-1}(-1)^{\bar{c}+bj}\frac{j^{\bar{c}}}{\{\Gamma(j)\Gamma(m-j)\}^b}\mathscr{S}_{b,j}(h_{j,1},\dots,h_{j,b}),$$

where $\bar{c}=a-b-1$, $\mathscr{S}_{b,j}$ is defined in equation (2.12) and

$$h_{j,s} = -(a-1)\frac{(s-1)!}{j^s} - b(s-1)!(H_{m-j-1,s}-H_{j,s}),$$

with $H_{j,s}=\sum_{i=1}^{j}1/i^s$ being the jth generalized harmonic number of order s.

Notice that the expression above can be simplified when $b=1$, as in the following Corollary.

Corollary 13. *The normalizing constant when $\alpha\sim\mathrm{Sg}(a,1,m)$ and $a\in\mathbb{N}$ and $m\geqslant 3$ is*

$$\mathcal{S}_{a,1,m} = \sum_{j=1}^{m-1}(-1)^{a+j}\frac{j^{a-2}\log j}{\Gamma(j)\Gamma(m-j)}.$$

By similar reasoning, we can derive an analytical expression also for the posterior coefficients.

Theorem 14. *Let $a, b \in \mathbb{N}$ and $m \geqslant 2$, and call $M = \min\{n, m\}$ and $\ell = |n - m|$. Then it holds that*

$$\mathscr{V}_{a,b,m}(n,k) = \sum_{j=1}^{M-1} (-1)^{\bar{k}-j(b+1)} \frac{j^{\bar{k}}}{\{\Gamma(j)\Gamma(M-j)\}^{b+1}(M-j)_\ell} \mathscr{S}_{b+1,j}(g_{j,1}, \ldots, g_{j,b+1})$$

$$+ \sum_{i=0}^{\ell-1} (-1)^{\bar{k}+i} \frac{(M+i)^{\bar{k}}}{\Gamma(i+1)\Gamma(\ell-i)\{(i+1)_{M-1}\}^{b+1}} \log(M+i),$$

where $\bar{k} = a + k - b - 2$, and

$$g_{j,s} = -(a+k-1)\frac{(s-1)!}{j^s} - (s-1)!\{bH_{M-j-1,s} - (b+1)H_{j,s} + bH_{M-j-\ell-1,s}\}.$$

2.7 Addendum II: Details of the simulation in Figure 2.1

The data in the left panel of Figure 2.1 consists of $n = 800$ observations generated independently from a mixture of four equally weighted bivariate normal distributions with variance-covariance matrix equal to $\text{diag}\{0.15, 0.15\}$ and means equal to $(-1,-1)$, $(1,-1)$, $(-1,1)$ and $(1,1)$, respectively. We let $X_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ and $f(y \mid X_i) = N_2(y; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ in equation (2.1), with $N_2(y, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denoting a normal distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^2$ and variance-covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{2\times 2}$. Our prior $\mathscr{Q}$ is a Dirichlet process with precision parameter α and normal-inverse Wishart baseline distribution $N(\mu; \mathbf{0}, \boldsymbol{\Sigma}/\kappa_0)IW(\boldsymbol{\Sigma}; \nu_0, \mathbf{I})$ with $\kappa_0 = \nu_0 = 2$. Four different scenarios are considered with respect to α: “Fixed, low” sets $\alpha = 1$, “Random, high” sets $\alpha = 5$, “Random, low” lets $\alpha \sim \text{Sg}(0.73, 0.1)$ and “Random, high” lets $\alpha \sim \text{Sg}(2.6, 0.1)$. The induced distribution on K_n has mean $\mathbb{E}(K_n) = 7.26$ in low cases and $\mathbb{E}(K_n) = 26$ in high ones. Inference on the number of clusters K_n in each scenario is performed by running a marginal Gibbs sampler as in Algorithm 3 in Neal (2000) for 20,000 iterations, discarding the first 5,000 as burn-in.

3

Bayesian modeling of sequential discoveries

3.1 Introduction

Our goal is to develop a flexible procedure for modeling 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), which were introduced in Chapter 2. 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 $(X_n)_{n \geqslant 1}$ 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 $K_n \leqslant n$ distinct entities, or species, representing the nth value of the accumulation curve. The values $(X_n)_{n \geqslant 1}$ 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 $n \geqslant 1$ proceeds as:

$$(X_{n+1} \mid X_1, \ldots, X_n) = \begin{cases} \text{``new''}, & \text{with probability} \quad \alpha/(\alpha+n), \\ X_i, & \text{with probability} \quad 1/(\alpha+n), \quad (i = 1, \ldots, n), \end{cases} \tag{3.1}$$

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. See equation (2.2) for the distribution of the induced partition of $\{1, \ldots, n\}$.

The predictive scheme in (3.1) is restrictive in depending on a single parameter and in inducing a logarithmic growth for the accumulation curve $(K_n)_{n \geqslant 1}$. 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 $(X_n)_{n\geqslant 1}$ 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 (3.1) 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 $\mathbb{E}(K_{n+m} \mid K_n = k)$, $m \geqslant 1$. 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., 2007a; Favaro et al., 2009). However, such models are often too restrictive, as is evident from Figure 3.1, 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

[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.

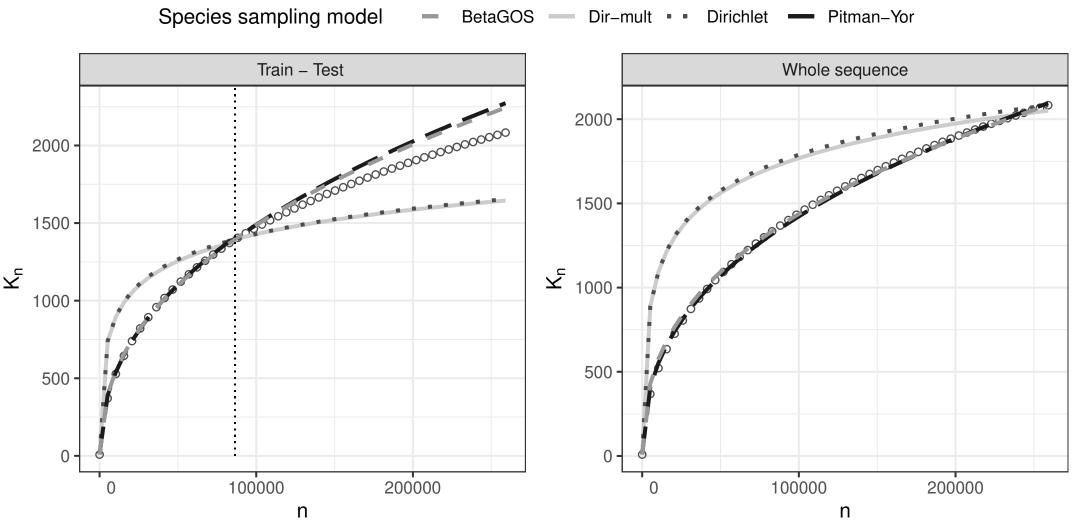

FIGURE 3.1: Empirical and estimated accumulation curve in one air fungal DNA-barcoding sample from Finland. White dots indicate observed values. Left panel: the vertical line is the training-test set cutoff, set to 1/3 of the total number of genetic sequences. The parameters of the Dirichlet, Pitman–Yor and Dirichlet-multinomial are estimated on the training set via empirical Bayes, while estimation for BETA-GOS 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.

has a 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 modeling framework, which is highly flexible, analytically tractable, and computationally efficient. The key distinction compared to species sampling models, such as (3.1), is that we directly specify a model for the accumulation curve $(K_n)_{n \geqslant 1}$, whereas the tags $(X_n)_{n \geqslant 1}$ are regarded as nuisance parameters. Specifically, we consider a collection of Bernoulli random variables $(D_n)_{n \geqslant 1}$ representing whether at the $(n+1)$th step a new entity has been discovered or not, namely

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

for $n \geqslant 1$, having set $D_1 = 1$. The accumulation curve is obtained by summing over these binary indicators: $K_n = \sum_{i=1}^{n} D_i, n \geqslant 1$. Differently from general species sampling models, in our framework, the Bernoulli indicators $(D_n)_{n \geqslant 1}$ are assumed to be *independent*, albeit not identically distributed. Hence, we aim at developing suitable formulations for the probabilities $(\pi_n)_{n \geqslant 1}$, with $\pi_n = \mathbb{P}(D_n = 1)$, for any $n \geqslant 1$. 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 $(\pi_n)_{n \geqslant 1}$, 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 Chapter is organized as follows. Sections 3.2-3.3 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 3.4. In Section 3.5 we test our model on simulated scenarios. Section 3.6 details the applications to real datasets. Concluding remarks are given in Section 3.7. The proofs of the statement and additional simulations are reported in Appendix B.

3.2 A general modeling framework for accumulation curves

3.2.1 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 $(X_n)_{n \geqslant 1}$ 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 (3.1) is given by

$$(X_{n+1} \mid X_1, \ldots, X_n) = \begin{cases} \text{``new''}, & \text{with probability} \quad \pi_{n+1}, \\ X_i, & \text{with probability} \quad q_{i,n+1}, \quad (i = 1, \ldots, n), \end{cases} \tag{3.2}$$

for $n \geqslant 1$, suitable probabilities $\sum_{i=1}^n q_{i,n+1} = 1 - \pi_{n+1}$ and $X_1 =$ "new". For Gibbs-type processes (Gnedin and Pitman, 2005), π_{n+1} and $q_{i,n+1}$ depend on previous values only through k and the frequencies $n_1, \ldots, n_k$, respectively. One example is the Pitman–Yor process, where $\pi_{n+1} = (\alpha + \sigma k)/(\alpha + n)$, $q_{i,n+1} = (1 - \sigma \bar{n}_i^{-1})/(\alpha + n)$, 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 $(X_n)_{n \geqslant 1}$ is *exchangeable*, i.e. invariant to reordering of the sequence, requiring strict conditions on $q_{i,n+1}$ 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 $n, m \geqslant 1$. 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 $q_{i,n+1} = (1 - W_i) \prod_{j=i+1}^{n} W_j$, where $W_n \sim \text{BETA}(a_n, b_n)$ are independent beta random variables for $n \geqslant 1$. As we describe in Section 3.5, the values for a_n and b_n determine the asymptotic behavior of the sequence. The sequential mechanism in (3.2) induces a law for the accumulation curve $(K_n)_{n \geqslant 1}$.

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:

$$\mathbb{E}(K_n) = \sum_{i=1}^{n} \frac{\alpha}{\alpha + i - 1}, \quad \mathbb{E}(K_m^{(n)} \mid X_1, \ldots, X_n) = \sum_{i=1}^{m} \frac{\alpha}{\alpha + n + i - 1}. \tag{3.3}$$

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

$n \geqslant 2$. However, beyond the Dirichlet process, the posterior expectation of $K_m^{(n)}$ is typically complex.

3.2.2 The model

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

Let $(D_n)_{n\geqslant 1}$ be a collection of *independent* binary indicators, denoting the discoveries, with probabilities $(\pi_n)_{n\geqslant 1}$. Moreover, let $K_n = \sum_{i=1}^n D_i$ for any $n \geqslant 1$ 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 $\pi_1, \dots, \pi_n$. We denote it as $K_n \sim \text{Pb}(\pi_1, \dots, \pi_n)$. 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 $(\pi_n)_{n\geqslant 1}$ are all equal, K_n has a binomial distribution. In our setting, $\pi_n > \pi_{n+1}$ for every $n \geqslant 1$ 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 15. *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 $(\pi_n)_{n\geqslant 1}$ are said to be directed by $S(t; \boldsymbol{\theta})$ if*

$$\pi_n = \mathbb{P}(T_n > n - 1) = S(n - 1; \boldsymbol{\theta}), \tag{3.4}$$

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

It is easy to check that a set of probabilities $(\pi_n)_{n\geqslant 1}$ 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 $n \geqslant 2$ with $D_1 = 1$. Hence, the discoveries $(D_n)_{n\geqslant 1}$ and the accumulation curve $(K_n)_{n\geqslant 1}$ carry the same information, having a one-to-one relationship. Then, if the probabilities $(\pi_n)_{n\geqslant 1}$ 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

$$\mathscr{L}(\boldsymbol{\theta} \mid D_1,\ldots,D_n) \propto \prod_{i=2}^{n} S(i-1;\boldsymbol{\theta})^{D_i}\{1 - S(i-1;\boldsymbol{\theta})\}^{1-D_i}, \tag{3.5}$$

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 16. *Let* $(X_n)_{n\geqslant 1}$ *be a sequence of objects directed by a Dirichlet process as in* (3.1) *and let* $(D_n)_{n\geqslant 1}$ *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 $n \geqslant 1$.

Hence, it is equivalent to base inferences on the Dirichlet process parameter α on the likelihood (3.5) 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. (2007a) 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 (3.5).

Remark 17. *If a sequence of discoveries $(D)_{n\geqslant 1}$ is directed by $S(t;\boldsymbol{\theta})$, the general predictive scheme in equation* (3.2) *may be specified as*

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

with $\sum_{i=1}^n q_i(n;\boldsymbol{\theta}) = 1 - S(n;\boldsymbol{\theta}) = F(n,\boldsymbol{\theta})$. As long as probabilities $q_i(n;\boldsymbol{\theta})$ 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 $q_i(n;\boldsymbol{\theta})$ 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 $q_i(n;\boldsymbol{\theta}) = (i^{1-\sigma} - (i-1)^{1-\sigma})/(\alpha + n^{1-\sigma})$ 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* (3.5)*, treating the labels $(X_n)_{n\geqslant 1}$ as nuisance parameters.*

3.2.3 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 $(\pi_n)_{n\geqslant 1}$ is directed by $S(t;\boldsymbol{\theta})$, it immediately follows that $K_n \sim \mathrm{Pb}\{1, S(1;\boldsymbol{\theta}), \ldots, S(n-1;\boldsymbol{\theta})\}$. 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 3.3.2.

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

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

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

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

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.

3.2.4 Asymptotic behavior of the number of distinct species

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 15.

Proposition 18. *Let* $K_n \sim \mathrm{Pb}\{1, S(1;\boldsymbol{\theta}), \dots, S(n-1;\boldsymbol{\theta})\}$. *Then, there exists a possibly infinite random variable* K_∞ *such that* $\lim_{n\to\infty} K_n = K_\infty$, *almost surely, with* $\mathbb{E}(K_\infty) = \sum_{i=0}^\infty S(i;\boldsymbol{\theta})$. *Moreover,*

$$\mathbb{E}(T) \leqslant \mathbb{E}(K_\infty) \leqslant \mathbb{E}(T) + 1. \tag{3.6}$$

Equation (3.6) 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 19. *Under the conditions of Proposition 18,* $K_\infty = \infty$ *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 20. *Let $K_n \sim \mathrm{Pb}\{1, S(1;\boldsymbol{\theta}), \dots, S(n-1;\boldsymbol{\theta})\}$ 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,*

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

Theorem 20 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 21. *Under the conditions of Proposition 18, if $K_\infty < \infty$ almost surely, then $\mathbb{E}(K_\infty) = \sum_{i=1}^\infty S(i-1;\boldsymbol{\theta}) < \infty$ and $\mathrm{var}(K_\infty) = \sum_{i=1}^\infty S(i-1;\boldsymbol{\theta})\{1 - S(i-1;\boldsymbol{\theta})\} < \infty$.*

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 (3.6) 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 $\mathbb{E}(K_\infty \mid D_1, \dots, D_n)$. Consider $\mathbb{E}(K_{m+n} \mid D_1, \dots, D_n)$ and let $m \to \infty$. Then, it is straightforward to see that $\mathbb{E}(K_\infty \mid D_1, \dots, D_n) = k + \mathbb{E}(K_\infty^{(n)} \mid D_1, \dots, D_n)$, where $\mathbb{E}(K_\infty^{(n)} \mid D_1, \dots, D_n) = \sum_{j=1}^\infty S(j+n-1;\boldsymbol{\theta})$.

3.3 Logistic models

3.3.1 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 3.3.3. In particular, we first consider a two parameter case

$$S(t;\alpha,\sigma) = \frac{\alpha}{\alpha + t^{1-\sigma}}, \qquad t \geqslant 0, \tag{3.7}$$

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 19. Conversely, when $0 \leqslant \sigma < 1$, 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 20 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 (3.7).

3.3.2 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

$$S(t; \alpha, \sigma, \phi) = \frac{\alpha\phi^t}{\alpha\phi^t + t^{1-\sigma}}, \qquad t \geqslant 0, \tag{3.8}$$

with $\alpha > 0$, $\sigma < 1$ and $0 < \phi \leqslant 1$. 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 $T \sim \textsc{ll}(\alpha, \sigma, \phi)$.

Proposition 22. *Let $K_n \sim \mathrm{Pb}\{1, S(1; \boldsymbol{\theta}), \dots, S(n-1; \boldsymbol{\theta})\}$, with $S(t; \boldsymbol{\theta})$ defined as in equation* (3.8)*. Then for any $0 < \phi < 1$ it holds that $K_n \to K_\infty < \infty$ almost surely as $n \to \infty$.*

Proposition 22 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 23. *Let $\alpha > 0$, $\sigma < 1$ and $0 < \phi \leqslant 1$. Then for any $n \geqslant 1$ and $0 \leqslant k \leqslant n$ 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 24. *The coefficients $\mathscr{C}_{n,k}(\sigma, \phi)$ of Definition 23 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 $n \geqslant 0$ and $1 \leqslant k \leqslant n+1$, with initial conditions $\mathscr{C}_{0,0}(\sigma, \phi) = 1$, $\mathscr{C}_{n,0}(\sigma, \phi) = 0$, $n \geqslant 1$, $\mathscr{C}_{n,k}(\sigma, \phi) = 0$, $k > n$. Moreover, for any $1 \leqslant k \leqslant n$ and $n \geqslant 2$, one has

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

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

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 25. *Let $K_n \sim \mathrm{Pb}\{1, S(1; \alpha, \sigma, \phi), \dots, S(n-1; \alpha, \sigma, \phi)\}$ for every $n \geqslant 1$. 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 25 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).

3.3.3 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 $n \geqslant 1$ with $\pi_1 = 1$. An interesting and practically useful property of our model is the following representation

$$\log \frac{\pi_{n+1}}{1 - \pi_{n+1}} = \log \alpha - (1-\sigma)\log n + (\log \phi)n = \beta_0 + \beta_1 \log n + \beta_2 n, \qquad n \geqslant 1, \quad (3.9)$$

having set $\beta_0 = \log \alpha$, $\beta_1 = \sigma - 1 < 0$ and $\beta_3 = \log \phi \leqslant 0$. Equation (3.9) 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 (3.9) facilitates extensions to include covariates. Suppose we are given a collection $(K_{1,n})_{n\geqslant 1}, \ldots, (K_{L,n})_{n\geqslant 1}$ of L accumulation curves, 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 $(D_{\ell,n})_{n\geqslant 1}$ be the sequence of discovery indicators for the ℓth location, with probabilities $(\pi_{\ell,n})_{n\geqslant 1}$. The most flexible specification for $K_{\ell,n}$ corresponds to the case in which all the parameters are location-specific so that for any $n \geqslant 1$,

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

This specification can borrow information across locations via a hierarchical model on $\boldsymbol{\beta}_l = (\beta_{l,0}, \beta_{l,1}, \beta_{l,2})^{\mathrm{T}}$ or by fixing certain parameters. Alternatively, systematic variation across locations can be modeled by including covariates $\mathbf{z}_\ell$ via

$$\log \frac{\pi_{\ell n+1}}{1 - \pi_{\ell,n+1}} = \beta_{\ell,0} + \beta_{\ell,1} \log n + \beta_{\ell,2} n = \mathbf{z}_\ell^{\mathrm{T}} \boldsymbol{\gamma}_0 + (\mathbf{z}_\ell^{\mathrm{T}} \boldsymbol{\gamma}_1) \log n + (\mathbf{z}_\ell^{\mathrm{T}} \boldsymbol{\gamma}_2) n, \quad (3.10)$$

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 $\mathbf{z}_\ell^{\mathrm{T}} \boldsymbol{\gamma}_2 \leqslant 0$. 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.

3.4 Posterior computation

3.4.1 Estimation procedures

Consider the model in equation (3.9). The parameters $\boldsymbol{\theta} = (\alpha, \sigma, \phi)$ can be estimated by maximizing the likelihood in equation (3.5), with $S(t; \boldsymbol{\theta}) = S(t; \alpha, \sigma, \phi)$, $\beta_1 < 0$

and $\beta_2 \leqslant 0$. 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 26. *Let $\hat{\boldsymbol{\theta}} = (\hat{\alpha}, \hat{\sigma}, \hat{\phi})$ be the unconstrained maximizer of equation* (3.5) *under the three-parameter specification in* (3.8)*, if it exists. If $K_n = k$ is the number of discoveries within the data $D_1, \dots, D_n$, 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 $\boldsymbol{\beta} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})\mathbb{1}(\beta_1 < 0; \beta_2 \leqslant 0)$. The covariate-dependent regression in equation (3.10) can be implemented in a similar manner; for details, see Appendix B.

3.4.2 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 $n_1, \dots, n_k$ 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),

$$\bar{K}_i = k - \binom{n}{i}^{-1} \sum_{j=1}^{k} \binom{n - n_j}{i}, \quad i = 1, \dots, n, \tag{3.11}$$

with $K_n = k$, where (3.11) 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 (3.11). 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

$$\mathscr{L}(\boldsymbol{\theta} \mid \bar{D}_1, \ldots, \bar{D}_n) = \prod_{i=2}^{n} S(i-1; \boldsymbol{\theta})^{\bar{D}_i} \{1 - S(i-1; \boldsymbol{\theta})\}^{1-\bar{D}_i}, \tag{3.12}$$

in place of equation (3.5). Inference about $\boldsymbol{\theta}$ based on (3.12) 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 $\mathbb{E}\{\mathscr{L}(\boldsymbol{\theta} \mid D_1, \ldots, D_n)\} \approx \mathscr{L}(\boldsymbol{\theta} \mid \bar{D}_1, \ldots, \bar{D}_n)$.

3.5 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 16.

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

(3.9). 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-$2(\theta, \beta)$, 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 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 3.1: 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.

	DIR-MULT		BETA-GOS-2		DIRICHLET		PITMAN–YOR	
	$H = 500, \sigma = -1$		$\theta = 500, \beta = 1.5$		$\alpha = 10$		$\alpha = 10, \sigma = 0.5$	
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-$2(\theta, \beta)$	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 3.1 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.

Table 3.2: 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.

	FINITE GEOM.		FINITE ZIPF		GEOMETRIC		ZIPF	
	$H = 100, \eta = 0.95$		$H = 3000, \eta = 0.25$		$\eta = 0.1$		$\eta = 2$	
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

Following the same structure as above, Table 3.2 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 3.1, with the first two models having $K_\infty < \infty$ and the last two $K_\infty = \infty$. Details are provided in Appendix B. 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 Appendix B.

3.6 Applications

3.6.1 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 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 3.2, 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 3.5 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

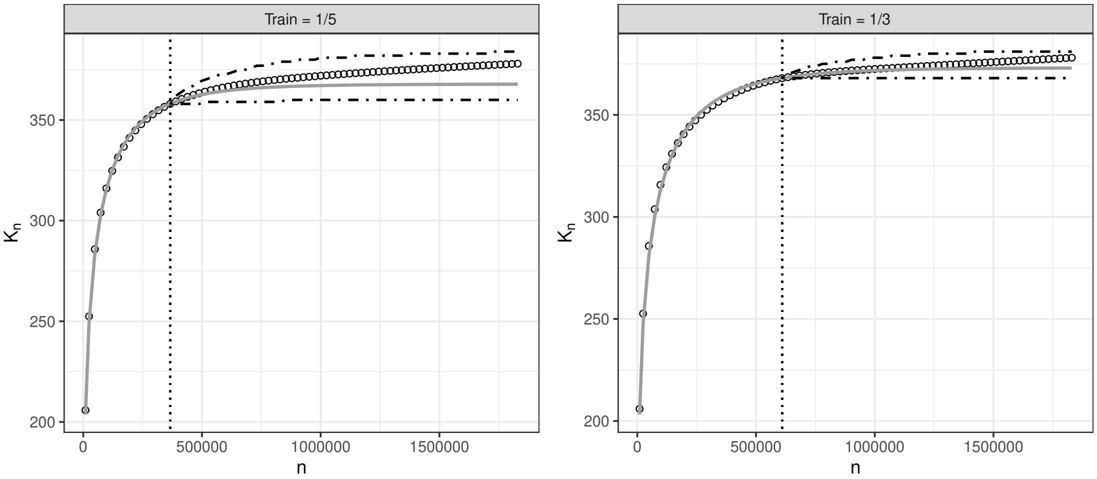

FIGURE 3.2: Performance of LL3 on the copepod species counts data. Circles: individual-based rarefaction curve. Grey line: predicted in- and out-of-sample accumulation curve computed by averaging over posterior samples of $\mathbb{E}(K_n)$ and $\mathbb{E}(K_n \mid D_1, \ldots, D_n)$, respectively. Black 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.

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 (3.9), and impose the constraints with truncated normal priors as in Section 3.4. Posterior samples for LL2 and LL3 are obtained via the Metropolis adjusted Langevin algorithm (Roberts and Rosenthal, 1998) with the proposal covariance equal to $\epsilon^2\hat{\mathbf{\Sigma}}$, where $\hat{\mathbf{\Sigma}}$ 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 3.5, taking k_n to be the

nth value of the individual-based rarefaction in equation (3.11).

Table 3.3: 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.

	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 3.3 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 $K_{n+m} \mid D_1, \ldots, D_n$, $m \geqslant 1$, 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 $\beta \geqslant 1$. 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 $K_{n+m} \mid D_1, \dots, D_n$ for LL3 contains the true value $\bar{K}_{n+m}$, and the posterior predictive mean $\mathbb{E}(K_{n+m} \mid D_1, \dots, D_n)$ slightly underestimates the truth. This is further confirmed by looking at the whole trajectory, as depicted in Figure 3.2. 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 Appendix B.

3.6.2 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

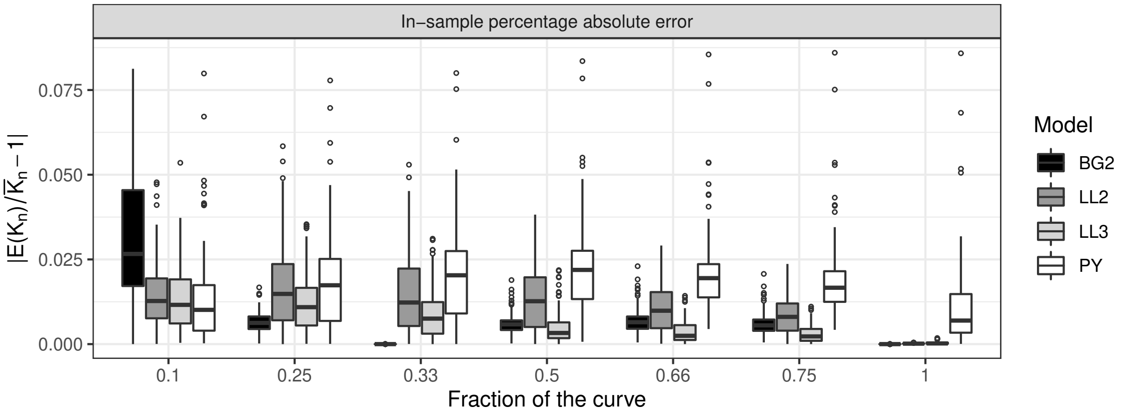

Figure 3.3: In-sample performance in the Finnish fungal biodiversity data. The in-sample estimator $\mathbb{E}(K_n)$ is computed by averaging model rarefaction across posterior samples. The value of $\bar{K}_n$ indicates the individual-based rarefaction curve at n.

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 3.6.1.

Figure 3.3 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

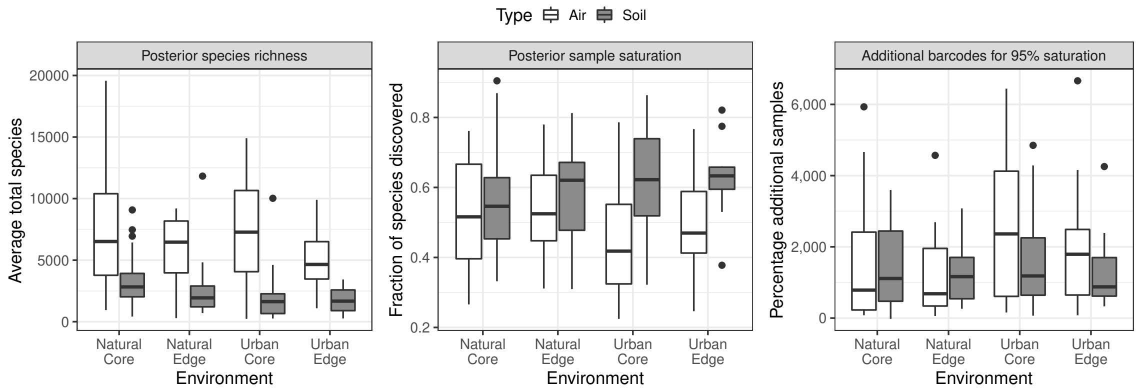

FIGURE 3.4: Left panel: distribution of the posterior mean species richness for the 150 samples. Center panel: distribution of the posterior mean sample saturation for the 150 samples. Right panel: additional number of samples (in percentage) required to reach a target posterior saturation of 0.95.

averaging the model rarefaction across the posterior samples, and $\bar{K}_n$ at a given fraction of a curve. All models perform 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 26 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 3.4, which displays the expected sample species richness for each of the 150 samples across site characteristics. Air samples tend to contain 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 $C_n = K_n/K_\infty \leqslant 1$ represent the saturation level of a given sample after n barcoded sequences. Differences across sites can be evaluated via $\mathbb{E}(C_n \mid D_1, \dots, D_n)$, which represents the posterior expected saturation level of a sample. Figure 3.4, 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 3.4, 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.

3.7 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 modeling 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 17, 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 modeling 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 3.3 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 $\beta \geqslant 1$.

Indeed, our results of Section 3.2 are fully general and can be readily specialized to any survival function. This is an interesting research direction.

4

Inferring taxonomic placement from DNA barcoding aiding in discovery of new taxa

4.1 Introduction

DNA barcoding refers to the practice of identifying the taxonomic affiliation of unknown specimens through short fragments of their DNA molecular sequence called barcoding genes (Hebert et al., 2003). Typically, this assessment is performed by comparing the DNA obtained from the high-throughput sequencing of a bulk sample to libraries of genes whose Linnean taxonomy is well-established. Examples of these collections are numerous, with the Barcode of Life project (BOLD; Sarkar and Trizna, 2011) and GenBank (Benson et al., 2012) being particularly notable cases. For the identification to be reliable, reference DNA sequences should be characterized by limited intra-species and high inter-species gene variation and should be sufficiently simple to align and compare (Hebert et al., 2003). In the animal kingdom and insects especially, these characteristics have been found in a region of approximately 650-base-pairs near the 5th end of the mitochondrial cytochrome c oxidase sub-unit I, or COI, gene (Janzen et al., 2005). This region has become routinely used in animal species identification. In particular, libraries in BOLD are formed by clustering similar COI sequences under a common Barcode Index Number, or BIN, which identifies a given species (Ratnasingham and Hebert, 2013).

The impact of DNA barcoding in biodiversity assessment has been dramatic. It took more than 200 years to describe approximately 1 million species of insects through morphological inspection, whereas nearly 400,000 BINs have been categorized within just 10 years (Wilson et al., 2017). DNA barcoding offers a way to categorize large quantities of specimens collected by modern automatic sampling methods. For example, flying insects are routinely captured with Malaise traps (Malaise, 1937), which collect the sampled insects together in a preservative within a storage cylinder. While this method often causes deterioration of the captured animals, making them morphologically unrecognizable, the biological material can be processed relatively cheaply (Shokralla et al., 2014) through a practice called DNA metabarcoding (Yu et al., 2012), which groups similar sequences detected in the samples into *operational taxonomic units*, or OTUs. These provide initial hypothesized species labels for the animals in the sample, and assessing their taxonomic placement is the final key stage of a bioinformatics pipeline.

Despite the advantages described above, taxonomic assessment of OTUs presents its own challenges, especially at lower level ranks. While it is relatively easy to accurately place a DNA sequence to a phylum, a class, or an order (Yu et al., 2012), the information obtainable via high-throughput methods is limited by the short length of the sequences extracted. This makes the identification at the family, at the genus, and at the species level subject to higher uncertainty (Pentinsaari et al., 2020). Moreover, DNA metabarcoding may be prone to sequencing and clustering errors. Consequently, it can either split biologic material from the same species into two different clusters or merge different species into a single cluster (Somervuo et al., 2017). Finally, reference sequence libraries can be subject to mislabelling errors

(Somervuo et al., 2016) and can be incomplete (Virgilio et al., 2012; Wilkinson et al., 2017; Weigand et al., 2019). This leads to the necessity of developing classification methods that provide a reliable characterization of uncertainty when taxonomically annotating the collected OTUs, accounting for the potential lack of information and therefore barcode novelty in the library. Ultimately, such methodologies allow one to quantify and track the biodiversity of a given sampling region only if the classification probabilities are reliable and OTUs are obtained consistently across time and space.

Much software for taxonomic recognition has been developed, relying on different prediction methods. One approach labels a query DNA with the taxon of the reference sequence having the highest similarity (Huson et al., 2007; Nguyen et al., 2014). This requires applying local or global alignment procedures to the sequences in the library, such as the BLAST - Basic Local Alignment Search Tool - similarity score (Altschul et al., 1990). When alignment is undesirable due to computational costs, fast algorithms that exploit a κ-mer representation of the sequences can be adopted. Widely used examples are the naïve Bayes RDP classifier (Wang et al., 2007) and its non-Bayesian heuristic alternatives (e.g. SINTAX; Edgar, 2013). More recent methods use modern machine learning and deep learning techniques including tree-based classification algorithms (IDTAXA; Murali et al., 2018) and convolutional neural networks (Vu et al., 2020).

While these approaches can provide good classification results when the training data are sufficiently informative of the biodiversity of the environmental sample (Bazinet and Cummings, 2012), they can lead to unreliable matches when the reference sequence set is incomplete, as is often case (Wilkinson et al., 2017; Murali et al., 2018). Thus, algorithms must coherently account for potential taxonomic novelty when doing classification (Somervuo et al., 2017). Specifically, sequences are regarded as "new" if their true taxonomic annotation is unobserved in the training library. This does not necessarily imply that the specimen from which DNA has been sequenced identifies a taxon new to science. Instead, novelty may be driven by a lack of reference sequencing data for known taxa, limited training libraries, low quality and gaps in barcodes, and sequencing errors in queries. All these factors can potentially lead to false positives, labelling a sequence as “new” when it is not, or false negatives, predicting a known taxon when a new one should be identified. This common issue has been addressed in the literature in various ways (Lanzén et al., 2012; Lan et al., 2012; Edgar, 2013; Bokulich et al., 2018). The widely adopted solution is to select a confidence probability cutoff and regard the classification as unreliable if the predicted taxon has a probability below that threshold (Wang et al., 2007). For example, the default RDP classifier does not report the predicted genus of a query if the most likely genus has a prediction probability lower than 0.8. This cutoff depends on the specific algorithm and often requires appropriate tuning (Lan et al., 2012). Moreover, confidence thresholds might be species-dependent due to differences in genetic variability between and within taxa. A second possibility is to explicitly allow the algorithm to signal if the queries are likely from previously unobserved taxa, as is done by PROTAX - PRObabilistic TAXonomic placement (Somervuo et al., 2016). PROTAX classifies DNA sequences by training a multinomial regression model on a sub-sample of the reference library reflecting prior knowledge of the existing taxonomy. The algorithm can lead to over- or under-detection of new taxa at any rank if the training dataset is not representative. With this approach, novel nodes in the taxonomic tree are explicitly treated as separate classes to be modelled, and they are assigned a prediction probability when classifying queries.

In this paper, we follow the latter approach and develop an off-the-shelf Bayesian nonparametric model for DNA barcode data that explicitly accounts for novelty by modelling the potential undetected nodes at every unlabelled taxonomic level. As our application primarily focuses on insects, we name our method BayesANT, short for BAYESiAn Nonparametric Taxonomic classifier. BayesANT is a supervised prediction algorithm that is trained on a set of sequences whose taxonomic affiliation is known and later annotates unlabelled DNA barcoding sequences in a probabilistic manner. In particular, it computes taxon-assignment probabilities at all unlabelled ranks by combining a prior distribution for the taxonomic tree with a kernel-based approach to modelling the distribution of the nucleotide sequences conditioned on their full taxonomic affiliation. Taxon novelty is incorporated through a Pitman–Yor process prior (Pitman and Yor, 1997), which is a species sampling model urn scheme (Blackwell and MacQueen, 1973; Pitman, 1996) that automatically specifies probabilities for the appearance of undiscovered species (Lijoi et al., 2007a; Favaro et al., 2009) in a coherent way. For aligned sequences, we use a Dirichlet-multinomial product kernel over nucleotides, while, for unaligned sequences, we use a multinomial kernel over κ-mer counts. The resulting model facilitates fast computation of a probabilistic classifier, which provides careful uncertainty assessments in taxonomic annotations. Unlike the other methods described above, our method avoids using an arbitrary threshold to annotate a sequence as being from a clade unobserved in training. In particular, taxonomic novelty in BayesANT can be aided through the choice of the Pitman–Yor prior hyperparameters, which can be either fixed ex-ante based on prior knowledge or estimated from the data. We test BayesANT on a library of arthropod DNA sequences collected in Finland (Roslin et al., 2022).

The Chapter is organized as follows. Section 4.2 presents the model in great detail. Section 4.3 presents the analysis of the Finnish DNA barcoding library. Concluding remarks are discussed in Section 4.4. Additional simulations and results are reported in Appendix C.

4.2 Materials and Methods

BayesANT evaluates the probabilities that a given DNA query sequence belongs to each of the nodes of the observed taxonomy, allowing for unobserved nodes in the taxonomic tree to be discovered. These probabilities are derived via Bayes rule, while taxonomic novelty arises through Pitman–Yor process priors. Let $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,L})$ be the taxonomic labels of the ith sequence in a library of L ranks, and $\mathbf{Y}_i$ the associated nucleotide sequence from any barcoding gene, such as COI for insects or ITS2 for fungi. We indicate taxonomic library of n sequences as $\mathscr{D}_n = (\mathbf{X}_i, \mathbf{Y}_i)_{i=1}^n$. See Sections 4.2.2 and 4.2.4 for more details on how the data are structured. The goal of BayesANT is to predict $\mathbf{X}_{n+1}$, the labels for $(n+1)$th sequence, treating the DNA $\mathbf{Y}_{n+1}$ as covariate. We perform this by paralleling the construction behind naïve Bayes classifiers and linear discriminant analysis: the probability that the $(n+1)$th query belongs to the taxonomic branch $\mathbf{x} = (x_1, \ldots, x_L)$ conditioned on library $\mathscr{D}_n$ and sequence $\mathbf{Y}_{n+1}$ is

$$p(\mathbf{X}_{n+1} = \mathbf{x} \mid \mathbf{Y}_{n+1}, \mathscr{D}_n) \propto p(\mathbf{X}_{n+1} = \mathbf{x} \mid \mathbf{X}^{(n)}) \times p(\mathbf{Y}_{n+1} \mid \mathbf{X}_{n+1} = \mathbf{x}, \mathscr{D}_n), \quad (4.1)$$

where $\mathbf{X}^{(n)} = (\mathbf{X}_i)_{i=1}^n$ are the observed taxonomic labels, $p(\mathbf{X}_{n+1} = \mathbf{x} \mid \mathbf{X}^{(n)})$ is the prior probability of branch $\mathbf{x}$ and $p(\mathbf{Y}_{n+1} \mid \mathbf{X}_{n+1} = \mathbf{x}, \mathscr{D}_n)$ is the distribution of the DNA sequence conditioned on $\mathbf{x}$ being its assigned branch. Refer to the Supporting Information for a step-by-step derivation of equation (4.1). In what follows, we carefully specify how each component is determined.

4.2.1 Preliminaries: the Pitman–Yor process

The Pitman–Yor (Pitman and Yor, 1997) is a sequential process for label assignment whose allocation probabilities depend on a precision parameter α, on a discount parameter σ, and on the size of the clusters previously detected. While we have introduced the sequential allocation scheme in the previous two Chapters, we re-state it here to ease readability. Suppose that $X_1, \ldots, X_n$ are the taxon assignments for the DNA sequences in our library of barcodes at a given rank (such as phylum or class). Specifically, these sequences identify $K_n = k$ distinct taxa, named $X_1^*, \ldots, X_k^*$, with frequencies $n_1, \dots, n_k$ and $\sum_{j=1}^{k} n_j = n$. Then, the probability that the $(n+1)$th sequence belongs to the jth of the known taxa is

$$p(X_{n+1} = X_j^* \mid X_1, \dots, X_n) = \frac{n_j - \sigma}{\alpha + n}, \tag{4.2}$$

for $j = 1, \dots, k$, while the probability of observing a new taxon is

$$p(X_{n+1} = \text{“new”} \mid X_1, \dots, X_n) = \frac{\alpha + \sigma k}{\alpha + n}, \tag{4.3}$$

where $\alpha > -\sigma$ and $\sigma \in [0, 1)$. Figure 4.1 sketches the mechanism when $n = 19$ sequences and $k = 4$ different groups are observed. High values of α or values of σ close to 1 lead to a high probability of discovering a new taxon. The probability that a sequence is assigned to taxon label X_j^* increases with its abundance n_j. This process allows barcodes to be clustered together a priori by being assigned to the same existing or newly detected taxa. Both parameters can be easily estimated from the data via empirical Bayes if taxonomic frequencies $n_1, \dots, n_k$ are observed. Refer to the Supporting Information for details, and to Favaro et al. (2009) and De Blasi et al. (2015) for a general overview.

4.2.2 Notation and taxonomic structure

A taxonomic library can be represented as a tree with branches of length $L \geqslant 2$, where DNA sequences are uniquely associated with one leaf. We denote such a library as $\mathscr{D}_n = (\mathbf{X}_i, \mathbf{Y}_i)_{i=1}^n$, where n is the number of sequences, $\mathbf{X}_i = (X_{i,\ell})_{\ell=1}^L$ indicates the taxonomic labels of the ith sequence and $\mathbf{Y}_i$ is a representation of the associated DNA. For example, the library we use in our application is fully annotated up to rank $L = 7$, where L represents the species level. Figure 4.2 displays an example of a taxonomic tree where sequences are classified into order, family and genus. Blue circles indicate nodes associated with at least one DNA sequence, while undiscovered branches are coloured in grey. The labels at a given level ℓ, namely $X_{1,\ell}, \dots, X_{n,\ell}$,

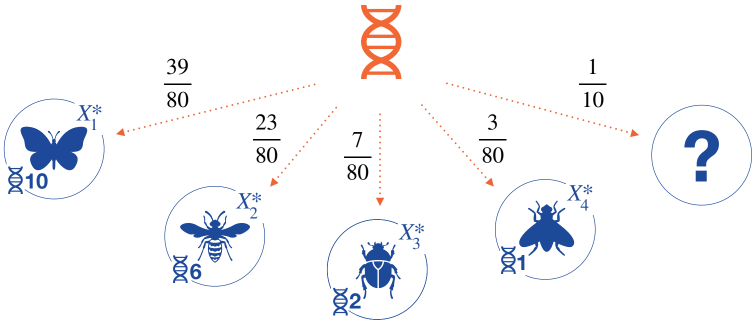

FIGURE 4.1: Example of a Pitman–Yor process with $n = 19$, $\alpha = 1$, $\sigma = 0.25$ and $K_n = 4$. Taxon names are reported on top of the circles, and frequencies of appearance are written on the right to the blue DNA sequences, respectively. Fractions in black denote the taxon probabilities for the orange DNA sequence. For example, the probability of observing the butterfly-shaped taxon X_1^* is $(n_1 - \sigma)/(\alpha + n) = (10 - 0.25)/(19 + 1) = 39/80$. The probability for the unknown question mark taxon is $(\alpha + \sigma k)/(\alpha + n) = (1 + 4 \times 0.25)/(19 + 1) = 1/10$.

take values in the space $\mathbb{X}_\ell$ of distinct taxa. Given their discrete nature, multiple $X_{i,\ell}$ can be associated with the same taxon. These realizations, which we denote as $X^*_{1,\ell}, \ldots, X^*_{k_\ell,\ell}$, are the nodes in our hierarchical taxonomy, with k_ℓ being their total observed number at level ℓ. For example, the 28 sequences in Figure 4.2 identify two taxa at the order level: one that has a butterfly-type morphological trait, $X^*_{1,1}$, and one with a bee-type trait, $X^*_{2,1}$. Thus, $k_1 = 2$. The beetle-shaped insect node instead represents a potential order unobserved in the library.

Due to the tree structure of the taxonomy, each generic node x_ℓ at level ℓ has a unique parent at level $\ell - 1$, denoted as $\mathrm{pa}(x_\ell)$. In Figure 4.2, for instance, $\mathrm{pa}(X^*_{1,2}) = X^*_{1,1}$ and $\mathrm{pa}(X^*_{1,3}) = \mathrm{pa}(X^*_{2,3}) = X^*_{1,2}$. For coherence, assume that the tree is rooted, namely $\mathrm{pa}(x_1) = x_0$ for any $x_1 \in \mathbb{X}_1$. Each node in the tree is linked to multiple taxa at lower ranks. Let $\rho_n(x_\ell)$ be the set of observed nodes $x_{\ell+1}$ for which $\mathrm{pa}(x_{\ell+1}) = x_\ell$ when n sequences are observed, $K_n(x_\ell) = |\rho_n(x_\ell)|$ be its cardinality and $N_n(x_\ell)$ be the number of DNA sequences belonging to x_ℓ. In Figure 4.2, $\rho_n(X^*_{1,2}) = \{X^*_{1,3}, X^*_{2,3}\}$ and $K_n(X^*_{1,2}) = 2$, while $\rho_n(x_0) = \{X^*_{1,1}, X^*_{2,1}\}$ and $K_n(x_0) = 2$ for the order level. Finally, the size of a node in our representation is determined as a sum of the number of

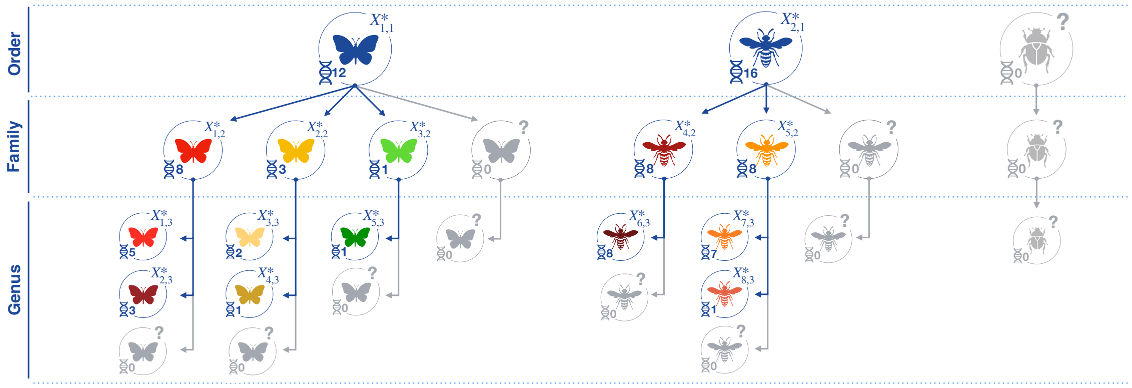

FIGURE 4.2: Example of a three-level taxonomic library under our model. On the bottom-left corner of every node, we report the number of DNA sequences linked to it. The total sample size of this example is $n = 28$. Circles in blue indicate nodes linked to leaves with observed DNA sequences, while grey circles show all the possible missing or undiscovered branches, labelled with a question mark on the top-right corner. Variation in insect colour along each branch and across branches indicate DNA and morphological similarities and differences, respectively.

sequences associated with all leaves connected to it. For example, $N_n(X^*_{1,2}) = 8$, and $N_n(X^*_{1,1}) = 12$. The quantities $\mathrm{pa}(\cdot)$, $\rho_n(\cdot)$, $K_n(\cdot)$ and $N_n(\cdot)$ are the key ingredients upon which we build our taxonomic prior in equation (4.1).

4.2.3 Taxonomic prior

The first step in our analysis consists of specifying a flexible prior for the frequencies of occurrence of different types of organisms at each taxonomic rank ℓ, including organisms of "new" types. In particular, we incorporate the Pitman–Yor process allocation probabilities in equations (4.2) and (4.3) into the tree structure. Let α_ℓ and σ_ℓ denote the allocation parameters for level ℓ, with $\alpha_\ell > -\sigma_\ell$ and $\sigma_\ell \in [0, 1)$. Write $\mathbf{X}^{(n)}_{\cdot,\ell} = (X_{i,\ell})_{i=1}^n$ as the sequence of taxonomic labels observed at level ℓ. Then, the taxon of sequence $n+1$ at level ℓ, conditioned on it being allocated to node $x_{\ell-1}$ at level $\ell - 1$, has probabilities

$$p(X_{n+1,\ell} = X^*_{j,\ell} \mid X_{n+1,\ell-1} = x_{\ell-1}, \mathbf{X}^{(n)}_{\cdot,\ell}) = \frac{N_n(X^*_{j,\ell}) - \sigma_\ell}{\alpha_\ell + N_n(x_{\ell-1})}, \tag{4.4}$$

if the node $X^*_{j,\ell}$ is such that $\text{pa}(X^*_{j,\ell}) = x_{\ell-1}$, and

$$p(X_{n+1,\ell} = \text{“new”} \,|\, X_{n+1,\ell-1} = x_{\ell-1}, \mathbf{X}^{(n)}_{\cdot,\ell}) = \frac{\alpha_\ell + \sigma_\ell K_n(x_{\ell-1})}{\alpha_\ell + N_n(x_{\ell-1})}, \tag{4.5}$$

if the node is new and originates from $x_{\ell-1}$. The structure of equations (4.4) and (4.5) is the same as the one in equations (4.2) and (4.3), with the only difference being that nodes at ℓ are generated from their parent-specific process. The level-specific parameters α_ℓ and σ_ℓ are important in allowing diversity to vary with taxonomic rank. Similarly to the one-level case discussed in Section 4.2.1, these parameters will be estimated based on the data. See Appendix C.

4.2.4 DNA sequence likelihood

The second step to build the predictive rule in equation (4.1) is to specify a distribution for the DNA sequences. We do this by adopting a kernel-based approach that flexibly accommodates different DNA representations.

As depicted in Figure 4.2, a query sequence $\mathbf{X}_i$ is uniquely associated with one leaf of the taxonomic tree. Recalling that $\mathbf{x} = (x_1, \ldots, x_L)$ denotes a taxonomic branch whose leaf is $x_L \in \mathbb{X}_L$, we let

$$(\mathbf{Y}_i \mid \mathbf{X}_i = \mathbf{x}, \boldsymbol{\theta}_{x_L}) \overset{\text{ind}}{\sim} \mathcal{K}(\mathbf{Y}_i; \boldsymbol{\theta}_{x_L}), \tag{4.6}$$

for every sequence $i = 1, \ldots, n$, where $\mathcal{K}(\mathbf{Y}_i; \boldsymbol{\theta})$ is a kernel depending on parameters $\boldsymbol{\theta}$ representing the likelihood of sequence data $\mathbf{Y}_i$, and $\boldsymbol{\theta}_{x_L}$ is a collection of leaf-specific parameters. Implicitly, we assume that all DNA sequences associated with leaf x_L are independent and identically distributed as $\mathcal{K}(\cdot; \boldsymbol{\theta}_{x_L})$. Table 4.1 provides three examples of multinomial-type kernels when sequences are aligned and when they are not. Here, *alignment* implies that all the sequences are pre-processed to have the same length p so that the nucleotides at each position $s = 1, \ldots, p$ are meaningfully

Table 4.1: Examples of multinomial kernels for the DNA sequences. The column SEQUENCES specifies whether the sequences in the library are aligned or not. Column KERNEL TYPE is the type of kernel chosen to model the DNA. Columns LIKELIHOOD and PRIOR FOR $\boldsymbol{\theta}_{x_L}$ are the likelihood ad the prior in each model, with DIR indicating the probability density function of the Dirichlet distribution. $\mathcal{N}_\kappa$ is the set of all κ-mers on which the sequence is decomposed. In the aligned case, this is a set of monomers $\mathcal{N}_1 = \{\text{A,C,G,T}\}$. The quantity $\mathbb{1}\{Y_{i,s} = g\}$ is an indicator equal to one if $Y_{i,s} = g$ and zero otherwise.

SEQUENCES	KERNEL TYPE	LIKELIHOOD	PRIOR FOR $\boldsymbol{\theta}_{x_L}$
Not aligned	κ-mers	$\prod_{g\in\mathcal{N}_\kappa} \theta_{x_L,g}^{n_{i,g}}$	$\text{DIR}(\boldsymbol{\xi}_{x_L})$
Aligned	Product	$\prod_{s=1}^{p}\prod_{g\in\mathcal{N}_1} \theta_{x_L,s,g}^{\mathbf{1}\{Y_{i,s}=g\}}$	$\prod_s \text{DIR}(\boldsymbol{\xi}_{x_L,s})$
Aligned	κ-Product	$\prod_{s=1}^{p}\prod_{g\in\mathcal{N}_\kappa} \theta_{x_L,s,g}^{\mathbf{1}\{Y_{i,s}=g\}}$	$\prod_s \text{DIR}(\boldsymbol{\xi}_{x_L,s})$

comparable. Then, $X_{i,s}$ is the nucleotide in the sth position of the ith query sequence, and $\theta_{x_L,s,g}$ is the probability that nucleotide $g \in \mathcal{N}_1 = \{\text{A, C, G, T}\}$ is seen at s for taxon x_L. Assuming independence across locations s as a simplifying assumption to improve computational efficiency in constructing a probabilistic classifier, the resulting kernel is a product of multinomials with location-specific parameters.

When sequences are not aligned, each has its own length p_i. A viable option is to use a κ-mer decomposition. This amounts to counting the number of times all possible 4^κ substrings of length κ appear within the sequence. We denote as $\mathcal{N}_\kappa$ the set of all κ-mers of length κ. For instance, 3-mers live in $\mathcal{N}_3 = \{\text{AAA, ACG, AGT}\ldots\}$, with a total of $4^3 = 64$ substrings. In Table 4.1, $n_{i,g} = \sum_{s=1}^{t_i} \mathbb{1}\{Y_{i,s} = g\}$ denotes the number of times a κ-mer $g \in \mathcal{N}_\kappa$ appears in the ith sequence, with $t_i = p_i - \kappa + 1$ being the total number of κ-mers observed when the length is p_i. We model these counts as the output of a multinomial distribution, where $\theta_{x_L,g}$ is the probability of κ-mer g at taxon x_L. The κ-mer length parameter κ is chosen a priori as a modeling choice and usually requires adequate tuning. Finally, if sequences are aligned, it is also possible to combine the two kernels by considering a κ-mer/location-specific multinomial distribution. For example, choosing a 2-Product kernel for a sequence

AATGTA means that the realizations of the multinomial are AA in the first location, AT in the second, TG in the third, and so on. This approach allows to better capture site dependencies but bears heavy computational costs for values of κ greater than 2.

The choice of kernel depends on the application and the data. For example, insect DNA sequences can be easily aligned via Hidden Markov models (Eddy, 1995), while fungal sequences often come without alignment due to their higher intrinsic variability. Irrespective of the structure of the data, our proposed multinomial kernels have the advantage of simplicity in computation, with the posterior distribution for θ_{x_L} obtained in analytic form by adopting conjugate Dirichlet priors as in Table 4.1. Computational efficiency is a critical issue both in training and in classifying very large numbers of sequences, making it intractable to consider elaborate likelihoods derived from realistic generative models of nucleotide sequences.

4.2.5 Prediction rule

The prior on the tree and the DNA sequence likelihood defined so far allow us to predict the set of labels $\mathbf{X}_{n+1}$ for the query sequence $\mathbf{Y}_{n+1}$. BayesANT does this in *bottom-up* and *top-down* steps. In the *bottom-up* step, we use equations (4.4), (4.5) and (4.6) to determine the posterior probability that $\mathbf{Y}_{n+1}$ belongs to *any* leaf in the tree. These include both the observed and the new taxa at the lowest level, as illustrated in Figure 4.2[1]. Then, probabilities of higher nodes are computed aggregating upward. In the *top-down* step, instead, BayesANT predicts a branch by iteratively choosing the child node with the highest probability at each level, starting from the root.

Let $\pi_{n+1}(\mathbf{x}) = p(\mathbf{X}_{n+1} = \mathbf{x} \mid \mathbf{X}^{(n)})$ be the prior probability for branch $\mathbf{x}$ after

[1] Under the assumption that a new node at level ℓ automatically creates a new node at all levels $\ell + 1, \dots, L$ below, the total number of potentially unobserved leaves is equal to the number of nodes up to $L - 1$ plus 1

having observed all labels $\mathbf{X}^{(n)}$. By the chain rule, this is equal to the product of the prior conditional probabilities in equations (4.4) and (4.5) of all nodes in the branch, which is

$$\pi_{n+1}(\mathbf{x}) = p(X_{n+1,1} = x_1 \mid \mathbf{X}^{(n)}_{\cdot,1}) \prod_{\ell=2}^{L} p(X_{n+1,\ell} = x_\ell \mid X_{n+1,\ell-1} = x_{\ell-1}, \mathbf{X}^{(n)}_{\cdot,\ell}). \tag{4.7}$$

Equation (4.7) corresponds to the prior taxon probability in equation (4.1). Notice that if $x_\ell =$ "new" at some ℓ, the conditional probabilities at lower nodes are equal to 1. But then, the probability that $\mathbf{X}_{n+1}$ is associated to branch $\mathbf{x}$ *conditioned on the DNA sequence* $\mathbf{Y}_{n+1}$ *and* $\mathscr{D}_n$, namely equation (4.1), is

$$p_{n+1}(\mathbf{x}) = p(\mathbf{X}_{n+1} = \mathbf{x} \mid \mathbf{Y}_{n+1}, \mathscr{D}_n) \propto \pi_{n+1}(\mathbf{x}) \int \mathcal{K}(\mathbf{Y}_{n+1}; \boldsymbol{\theta}_{x_L}) p(\boldsymbol{\theta}_{x_L} \,|\, \mathscr{D}_n) \mathrm{d}\boldsymbol{\theta}_{x_L}. \tag{4.8}$$

The integral in equation (4.8) is the posterior predictive distribution of DNA sequence $\mathbf{Y}_{n+1}$ with respect to the posterior of $\boldsymbol{\theta}_{x_L}$. When $x_L =$ "new", this posterior is equal to the prior, i.e. $p(\boldsymbol{\theta}_{x_L} \,|\, \mathscr{D}_n) = p(\boldsymbol{\theta}_{x_L})$, since no sequence for x_L is observed. The convenient property of the models in Table 4.1 is that both the prior and the posterior predictive distribution have simple and easy-to-compute analytic forms. Once equation (4.8) has been evaluated for all leaves, the probabilities of higher nodes in the taxonomy can be easily derived via upward aggregation. Then, we predict the taxa by starting from the root of the tree and recursively selecting the child node with the highest probability. Specifically, the predicted sequence of taxa $\mathbf{x}^* = (x^*_\ell)_{\ell=1}^L$ for the DNA sequence at $n+1$ satisfies

$$x^*_\ell = \arg\max_{x_\ell \in \rho_n(x^*_{\ell-1})} \sum_{\mathbf{x}:\ x_L \in \mathcal{L}_n(x_\ell)} p_{n+1}(\mathbf{x}), \tag{4.9}$$

where $\{\mathbf{x} :\ x_L \in \mathcal{L}_n(x_\ell)\}$ is the set of all branches $\mathbf{x} = (v_1, \ldots, x_L)$ whose leaves x_L are linked to node x_ℓ in a library of n DNA sequences.

4.2.6 Hyperparameter tuning

The hyperparameters $\boldsymbol{\xi}_{x_L}$ of the multinomial kernel play a fundamental role in novel species recognition. As detailed above, when $x_L =$ "new", then equation (4.8) is a prior predictive probability, since no sequence is observed for x_L and thus $p(\boldsymbol{\theta}_{x_L} \mid \mathscr{D}_n) = p(\boldsymbol{\theta}_{x_L})$. In such cases, prior hyperparameters should contain information regarding the taxonomic branch and level where novelty appears. Uniform priors may be unreasonably vague, leading to underestimation of the prior predictive probability of novel taxa relative to the true proportion. Thus, we tune each $\boldsymbol{\xi}_{x_L}$ as follows. Consider a taxon x_{L-1} at level $L-1$. If x_{L-1} is not "new", the hyperparameters $\boldsymbol{\xi}_{x_L}$ of all the leaves $x_L \in \mathcal{L}_n(x_{L-1})$ linked to it - including the new one - are all equal, and they are obtained via method of the moments from the DNA sequences $\mathbf{Y}_i$ with $X_{i,L-1} = x_{L-1}$. If instead x_{L-1} is a "new" node and the last not novel node in its branch is x_ℓ at level $\ell \leqslant L-1$, the method of the moments is applied on the set of sequences $\mathbf{Y}_i$ such that $X_{i,\ell} = x_\ell$. This ensures borrowing of information between the branches when the novelty appears at higher levels in the taxonomy. Moreover, this approach tailors the prior predictive distribution of a node to the intrinsic location-specific nucleotide variability of the sequences linked to it. Thus, novelty probability is high in a node if the query is coherent with the observed variability at that node but is not sufficiently similar to any of the training sequences linked to the children nodes in terms of the kernel. For mathematical details on the method of the moments applied to the multinomial kernels of Table 4.1, see Appendix C.

4.2.7 Calibration of prediction probabilities

Misspecification of a Bayesian model, due to inaccuracies in the prior and/or likelihood function, may lead to predictive probabilities that are not sufficiently well calibrated to accurately capture predictive uncertainties (Grünwald and van Ommen, 2017; Miller and Dunson, 2019). Given the complexity of the true data-generating likelihood underlying DNA barcoding data, and the necessity of using a simple likelihood for computational tractability, some degree of misspecification is inevitable. We apply a simple re-calibration approach to adjust the predictive probabilities used in equation (4.9) for misspecification.

In particular, we post-process the prediction probabilities in equation (4.8) by exponentiating them by a coefficient $\rho \in (0, 1]$ and later renormalizing. Then, the new probabilities for the $(n + 1)$th sequence are

$$\tilde{p}_{n+1}(\mathbf{x}) = \frac{p_{n+1}(\mathbf{x})^{\rho}}{\sum_{\mathbf{x}'} p_{n+1}(\mathbf{x}')^{\rho}}, \tag{4.10}$$

and can be used in place of $p_{n+1}(\mathbf{x})$ in equation (4.9). Such a strategy does not alter the ranking of the original probabilities since the transformation is monotonic. Moreover, if $p_{n+1}(\mathbf{x}) = 1$, then also $\tilde{p}_{n+1}(\mathbf{x}) = 1$. This implies that we do not substantially alter the prediction whenever the BayesANT is certain about a taxon. Choices for ρ can be adopted via cross-validation on a hold-out subset of the training library following strategies such as the ones described in (Guo et al., 2017). Specifically, prediction probabilities are calibrated if the average probability for the predicted nodes is equal to the classification accuracy (Somervuo et al., 2016). For example, if 90% of the sequences are correctly classified, ideally the average classification probability is approximately 0.9. An average value of 0.5 and 0.99, instead, means that the algorithm is too conservative when right and too confident when wrong, respectively. In the application discussed in this paper, we select $\rho = 0.1$ and $\rho = 0.06$ depending on the testing scenario.

4.3 Results

4.3.1 The FinBOL library

The Finland Barcode of Life initiative[2] (FinBOL) is a DNA barcoding library that contains reference sequences with highly reliable taxonomic annotations for the arthropod species of Finland. The data have been constructed placing substantial effort on barcode quality thanks to the collective effort of about 150 taxonomists. Biologic material was collected from previously identified specimens conserved in museums or private collections, and later processed via PCR sequencing. For a thorough description of how the library was assembled and later tested, refer to Roslin et al. (2022).

The version of the data we consider contains a total of $34,624$ DNA sequences annotated across seven taxonomic levels, namely class, order, family, subfamily, tribe, genus and species. Reference annotations are based on the national checklist of Finnish species (FinBIF, 2020) with the inclusion of dummy taxa whenever subfamily and tribe were missing. The library has been globally aligned via Hidden Markov Models using the HMMER software (Eddy, 1995). As a result, each sequence has a length of 658 base pairs, consisting of nucleotides "A", "C", "G" and "T" and alignment gaps "-". Other infrequent special characters are ignored and treated as missing values for simplicity. Taxonomic labels in the data comprise 3 classes: *Arachnida*, *Insecta* and *Malacostraca*, appearing 1,842 and 32,781 and 1 times, respectively. The sequences are further divided into 21 orders, 476 families, 896 subfamilies, 1,355 tribes, 3,855 genera and 10,985 species, 3,025 of which have a single reference sequence associated with them.

Figure 4.3 depicts the pairwise raw DNA similarities, calculated as the fraction of locations with identical nucleotides, between 3,000 sequences randomly sampled

[2] https://en.finbol.org/

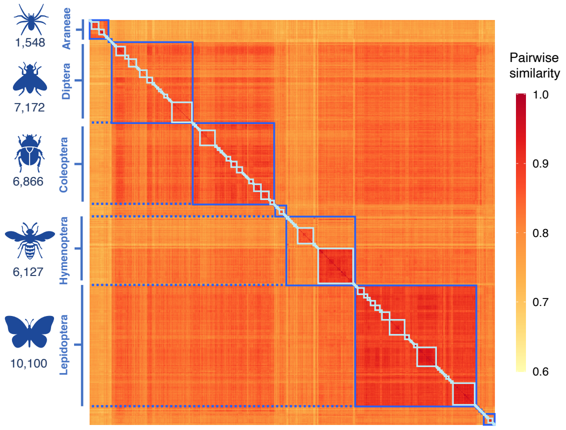

Figure 4.3: Pairwise DNA similarities between 3,000 randomly sampled sequences from the FinBOL library. The blue and light blue boxes along the main diagonal identify the orders and the families, respectively. Numbers on the left side represent the frequencies for the five largest orders in the data. Darker tones of red indicate higher similarity.

without replacement from the library. Each row/column represents the DNA similarity between one sequence and all the other sampled ones, with darker tones indicating higher similarities. Sequences are sorted alphabetically first by order and then by family to ensure cluster separation. In particular, boxes in dark blue along the main diagonal highlight the cross similarities within the orders, while boxes in light blue refer to the families. On the left side of the Figure we report the name and the sizes of the 5 most frequent orders, namely *Araneae*, *Diptera*, *Coleoptera*, *Hymenotptera* and *Lepidoptera*. In an ideal setting, the within-taxon similarities along the main diagonal should be higher than the cross-taxa ones. However, this is only true for *Lepidoptera* and for the two largest families - *Ichneumonidae* and *Tenthredinidae* - in *Hymenoptera*. Indeed, *Diptera* and *Coleoptera* are virtually indistinguishable, as they show a similar within- and between-order similarity. Moreover, these two taxa show a high cross-similarity with *Lepidoptera*, as indicated by the off-diagonal orange rectangles. Overall, the average DNA similarity in the library is around 0.81, with a standard deviation of 0.04, indicating that the sequences are highly homogeneous.

Testing scenarios

We aim to evaluate the performance of the predictive taxa classification probabilities produced by BayesANT. These probabilities should reflect whether the true taxa of a test sequence are observed in training or not. In the first case, the ideal output assigns a high or close-to-one probability to the true branch at every level, and a near-zero one to all the other branches in the tree. In the second case, instead, if the true affiliation of a sequence is observed for levels $1, \dots, \ell$ and unobserved for levels $\ell+1, \dots, L$, we would like BayesANT to output a high probability for the true nodes up to ℓ and the highest conditional probability to the "new" clade at level $\ell + 1$. To test our algorithm, we train the classifiers on a random subset of 80% of the FinBOL data and predict the taxonomic affiliation for the remaining 20% of the sequences. By construction, this procedure makes some taxa present in the training set only, others in both the training and the test, and some solely in the test set. We refer to this last category as to the "new", the "novel" or the "unobserved" taxa, treating the three terms as interchangeable synonyms.

We consider two testing scenarios summarized in panels (A) and (B) in Figure 4.4. In the first, each sequence in the library has equal probability of being allocated to the test set. This makes the taxonomic composition of the training and test set similar. As a result, only a relatively small fraction of the taxa will be unobserved in training, as is evident from both plots at the top of panels (A) and (B). In the second scenario, we create the test set by stratified sampling: for each test observation, we first sample the family, and then draw one sequence within that family. This assigns each family an equal probability of being selected, irrespective of its frequency of appearance in the data. Such a procedure yields a different composition between training and test, resulting in many more test taxa unobserved in training. In total, the number of barcodes whose true branch has at least one node unobserved in training is 884 in

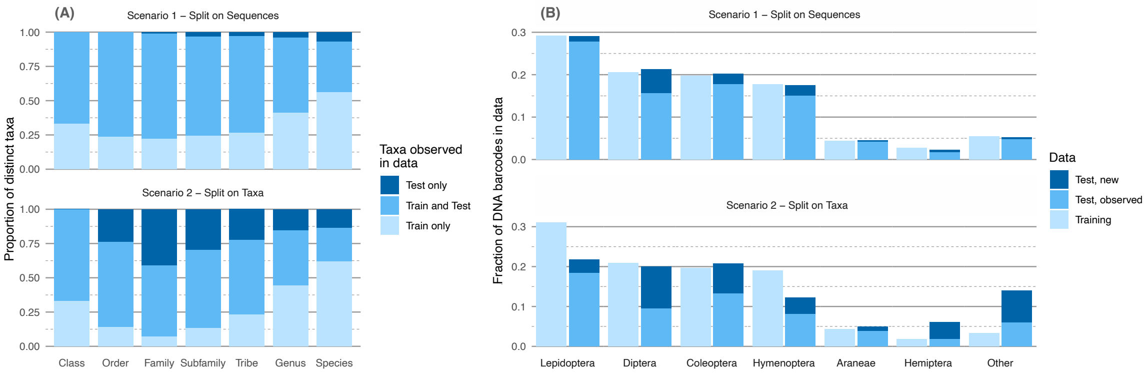

FIGURE 4.4: Panel (A): taxonomic composition of the training and the test libraries in the two splitting scenarios. Panel (B): proportion of DNA barcoding sequences pertaining to the larger orders in the data in both scenarios. The fractions highlighted in dark blue refer to the barcodes which truly belong to the mentioned order but whose true species is unobserved in training. The total number of sequences in each scenario is 27, 699 in the training library and 6, 925 in the test.

scenario 1 and 2,672 in scenario 2, while the total number of query test sequences is 6,924 in each case. Furthermore, the proportion of test DNA sequences associated with the most frequent orders differs from their training counterpart. For example, 30% of the sequences in the training library in scenario 2 are *Lepidoptera* and only 2.5% pertain to *Hemiptera*; in the test set, however, these fractions become 20% and 5%, respectively, with a much larger proportion of unknowns than is scenario 1. See the bottom of Figure 4.4, panel (B).

4.3.2 Test results

BayesANT computes the probability of every node in the taxonomic tree, including potential novel ones, for every test DNA sequence. The predicted annotation is the taxonomic branch with the highest probability at every rank. These probabilities express the uncertainty of the classification, and need to be well calibrated to be reliable: for instance, if 90% of the sequences are correctly classified, then the average probability with which they are classified should be around 0.9. Ideally, we would like to limit cases in which the algorithm is too confident when wrong and too conservative when right; see the Materials and Methods section. Moreover, evaluating the performance of BayesANT requires a clear definition of correctness of the classification under novel taxa. Suppose the true annotation of a test sequence shows a taxon that is unobserved in training. In that case, the prediction outcome may be the correct novel taxonomic leaf, or a new taxon but in an incorrect branch, or a taxon observed in training. We consider the classification correct in the first case and wrong otherwise. For example, if the true annotation of the test sequence is

```
Insecta -> Diptera-> Tephritidae -> Trypetinae -> Trypetini -> Acidia
    -> Acidia cognata
```

but `Acidia` is a genus not observed in the training set, then the correct classification up to the species rank is

```
Insecta -> Diptera -> Tephritidae -> Trypetinae -> Trypetini
        -> New Genus in Trypetini
        -> New Species in New Genus in Trypetini
```

since the novelty produces a new genus and automatically a new species linked to it. As `Acidia` is not observed, necessarily also the species `Acidia cognata` is unseen and the classification at the species level is correct only if BayesANT recognizes the novel genus. An outcome such as

```
Insecta -> Diptera -> Tephritidae -> Trypetinae -> Trypetini -> Trypeta
        -> New Species in Trypeta
```

is wrong but recognizes a novel leaf, while

```
Insecta -> Diptera -> Tephritidae -> Trypetinae -> Trypetini -> Trypeta
-> Trypeta zoe
```

is wrong since it predicts an observed species. When computing accuracy, the first example is correct at the genus and species level, while the other two are not. Unlike other approaches (e.g. see Edgar, 2018), this over-penalizes the cases when the algorithm fails at predicting the correct novel clade.

Figure 4.5 displays the prediction probabilities of BayesANT in both FinBOL scenarios by plotting the relationship between the % cumulative probability and the % cumulative accuracy at the species level. As the library is globally aligned, we adopt a simple product-multinomial kernel in which the probabilities of nucleotides “A”, “C”, “G” and “T” vary by loci and species. We treat the alignment gap “-” as a missing value and ignore the likelihood contribution of the locations where it appears. For an assessment of how these missing values affect the classification, see the Supporting Information. The rank-specific parameters α_ℓ and σ_ℓ are estimated from the data and we report their values in Table 4.2. Operations were performed on an AMD Ryzen 3900-based dedicated server with 128GB of memory on Ubuntu 20.04, R version 4.1.1 linked to Intel MKL 2019.5-075. Training the algorithm on 27,699 sequences took 1.7 minutes in scenario 1 (10,422 species) and 1.4 minutes in scenario 2 (9,490 species), while predicting the remaining 6,924 test queries took 10.2 minutes on a single thread and 1.4 minutes on 24 separate threads in each scenario. See the Supporting Information for additional details on computational time. In Figure 4.5, the dashed diagonal indicates a perfectly calibrated output, while trajectories below and above it imply over- and under-confidence, respectively. The dark blue lines show that BayesANT produces well-calibrated predictive probabilities on the test data, with a prediction accuracy equal to 85.2% and 70.6% from the test data and an average prediction probability of 0.82 and 0.70 in Scenarios 1 and 2, respectively. Results in scenarios 1 and 2 below are based on adjusting initial probabilities with a temperature parameter $\rho = 0.1$ and $\rho = 0.06$, respectively. Both values were chosen via standard cross-validation methods as follows. For a given training-test split, we first randomly assign 20% of the training sequences to a hold-out validation set. Then, we train the model on the remaining 80% of the training library and evaluate the prediction probabilities for the validation sequences against a set of pre-determined values for ρ. As a final step, we re-train the model on the full training set

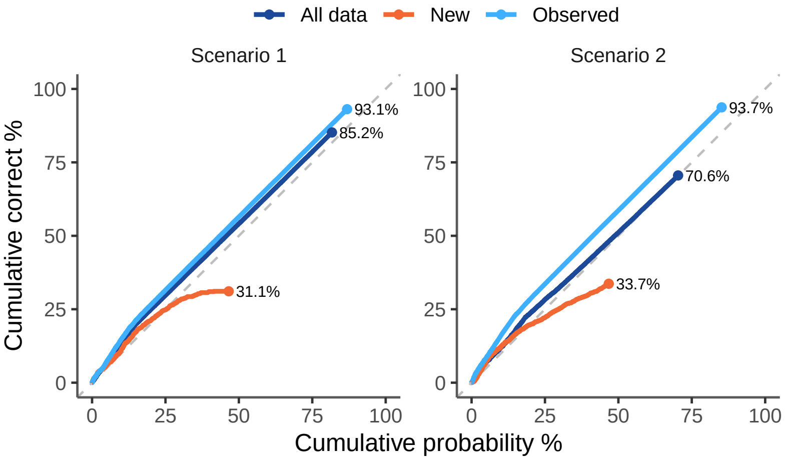

FIGURE 4.5: Calibration plot for the prediction of BayesANT at the species level under both scenarios. The dashed diagonal line indicates perfect calibration, while the percentages next to the points are the species accuracies in the test sets. Notice that "All data" includes all the 6,925 query sequences in the test set, "New" refers to those whose true taxon is not observed in training at some rank (884 in scenario 1, $2,642$ in scenario 2), while "Observed" restricts to the cases where the true taxonomy is fully observed.

Table 4.2: Estimated Pitman-Yor parameters for each level in the FinBOL taxonomic tree.

SCENARIO	PARAM.	CLASS	ORDER	FAMILY	SUBFAM.	TRIBE	GENUS	SPECIES
1	α_ℓ	0.19	1.17	4.16	1.04	1.16	1.86	7.11
	σ_ℓ	0.00	0.01	0.12	0.00	0.00	0.05	0.00
2	α_ℓ	0.19	0.76	2.66	1.08	1.12	1.85	6.74
	σ_ℓ	0.00	0.03	0.13	0.00	0.00	0.07	0.00

and predict the test sequences using the value of ρ that yielded the best calibration in the hold-out.

For the novel cases, the number of sequences predicted to belong to a "new" leaf in Scenario 1 is 958, while their true number is 884. Of these 884 queries, 77.9% are correctly recognized as novel, and 31.1% are effectively correct up to the species level included, with average probability equal to 0.44, as depicted by the orange line in the left panel. This implies that, while the exact "new" leaf in the taxonomy is generally challenging to retrieve due to insufficient signal in the dataset, BayesANT recognizes fairly well the potential novelty of the taxon of a sequence. Similar results are obtained in Scenario 2. While accuracy is lower due to a higher number of sequences with unobserved taxa, the predicted novel leaves are $2,736$ against $2,672$ truly "new". Here, 93.8% are recognized novel, and 33.7% are placed in the correct novel clade in the taxonomic tree. Verifying the effective novelty of the predicted "new" branches requires carefulness and further investigation - for example, by morphological assessment and more comprehensive reference barcode sequencing of new samples collected at the same geographic location. For instance, training BayesANT on a library of insects collected in Finland and using it to predict queries collected from South Africa might lead to an overwhelming number of barcodes labelled as "new" simply due to structural differences between the data, even if the latter has a well-established taxonomy.

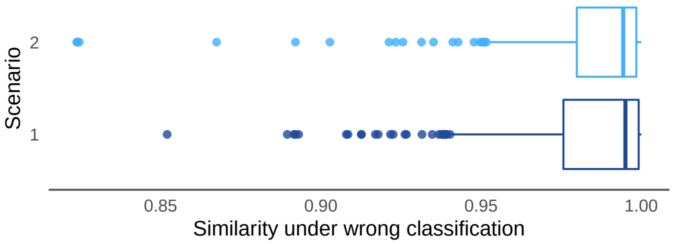

Figure 4.6: Average DNA similarity between the test query sequences and the predicted taxa when BayesANT incorrectly predicts a taxon observed in training.

In considering these taxonomic classification results, it is important to keep in mind the limited information provided by the available nucleotide sequencing in the COI gene. This information can be insufficient to assign certain query sequences to the correct taxon accurately. As we described in Figure 4.3, for example, orders *Coleoptera* and *Diptera* show a high cross-similarity. Indeed, these are orders who appear to be harder to classify: in Scenario 1, 36.2% of the incorrectly classified sequences at the species level are *Diptera*, followed by *Hymenoptera* (23.5 %), *Coleoptera* (16.7 %) and *Lepidoptera* (12.2 %). This is even more evident in Scenario

2, with 29.9% of wrong prediction for *Diptera* and 19.6% for *Coleoptera*. Notoriously, these are the most prone to barcoding mislabelling (Meier et al., 2006). For a full breakdown of the accuracies across orders, refer to the Supporting Information.

To investigate whether this lack of information is a primary cause when BayesANT produces incorrect classifications, we measured the average similarity between the query test sequence and the sequences in the training, which are annotated with the predicted taxa when the classification is wrong. Figure 4.6 shows the distribution of these average similarities. Indeed, these are generally high, with an average of 0.983 under both Scenarios. This suggests that misclassification tends to be due to insufficient information to distinguish between the true taxon and an incorrect taxon that is extremely close in the COI region to the query sequence, which sometimes even leads to small discrepancies between barcode similarities and true taxonomic affiliation. An example can be seen in Figure 4.7, which reports the pairwise DNA similarity between the query test sequence FISYO1282-18 and the training barcodes whose species are labelled as *Allantus calceatus* and *Allantus basalis*. In FinBOL, the true species of the query sequence is *Al. basalis*, while BayesANT wrongly suggests that the most likely species is *Al. calceatus* with a prediction probability equal to 0.942. The resemblance between the orange picture and those referring to FISYO2086-18 and FISYO270-18 should be evident. However, the DNA barcodes suggest the opposite: the average similarity between the query and *Al. calceatus* is higher than the one with *Al. basalis*, thus explaining the incorrect prediction. Indeed, such discrepancies have led to the introduction of the Barcode Index Number system (BIN) to cluster similar COI barcodes into OTUs. For example, all the sequences in Figure 4.7 fall into the same BIN called BOLD:ABZ8200. For an extensive discussion on the topic, refer to Ratnasingham and Hebert (2013); Phillips et al. (2019).

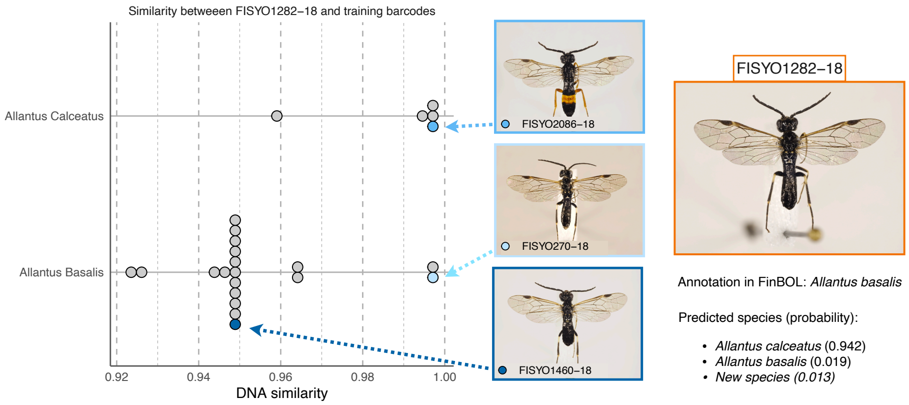

Figure 4.7: Pairwise DNA similarity between the test sequence FISYO1282-18 and the training barcodes belonging to the species *Allantus basalis* (19 sequences) and *Allantus calceatus* (5 sequences) in scenario 1. Each dot represents a pairwise similarity, with stacked dots indicating equality. The pictures in the blue boxes depict the specimen from which the training barcodes associated with the blue points have been sequenced. The orange box is the specimen of the test query, which is annotated as *Allantus basalis* in FinBOL. The bottom-right corner reports the predicted species probabilities returned by BayesANT. All pictures are publicly available at `https://www.boldsystems.org/` and licensed under CC BY-NC 3.0. License holder: Marko Mutanen, University of Oulu.

4.3.3 Benchmarking

As the last step in our analysis, we benchmark the performance of BayesANT on the FinBOL library against several alternatives in terms of accuracy. Table 4.3 reports the results under both Scenarios. m-1 refers to the single location multinomial kernel we adopted in our analysis above. While this is our method of reference due to its simplicity and flexibility, it treats loci as independent. Dependence can be introduced by adopting a 2-mer location kernel, m-2, where the support of the multinomial is in $\{\text{AA}, \text{AC}, \text{AT}, \dots, \text{TT}\}$ and 2-mers are overlapping. To assess the advantage of adopting a Pitman–Yor prior over the taxonomic tree, we also compare with an analysis that lets $\alpha_\ell = \sigma_\ell = 0$ at every level ℓ. This does not allow new species, and the prior is the proportion with which each taxon appears in the library at every rank. These methods are labelled as m-1, no new and m-2, no new in Table 4.3.

Although sequences are aligned, we also test the performance of BayesANT under the multinomial kernel over the κ-mer decomposition. In particular, K-5 and K-6 report accuracies and average prediction probabilities when fixing $\kappa = 5$ and $\kappa = 6$, respectively. Finally, we benchmark all these alternatives against the popular RDP classifier (Wang et al., 2007, version 2.13, 2020). While the number of taxonomic classifiers is rather vast, we focus on RDP as it is a longstanding method that, similarly to ours, relies on Naïve Bayes classification strategies and provides a minimum standard for accuracy. In particular, we do not set a confidence cutoff for RDP, but we consider its full classifications up to the species rank. This allows to benchmark its calibration under wrong predictions, which necessarily happen when the sequences are novel.

We first notice that no method is uniformly better or worse than the others, except for the κ-mer kernels. This is likely because the library has been reliably aligned. Aside from K-5 and K-6, performances in Scenario 1 are approximately similar both in terms of prediction probabilities and accuracy. Minor differences are found at the species level, where the inclusion of novel taxa leads to higher accuracy for both M-1 and M-2. When new taxa are in the data, all methods other than BayesANT have a lower percentage of correctly identified sequences in both scenarios. Moreover, the algorithms show a similar behaviour in Scenario 2, which features a much higher proportion of unobserved taxa in training, except the species level. Here, BayesANT shows its advantage, as it attains a prediction that is 10% more accurate than the RDP classifier. When we restrict to the species observed in training, however, model M-1 shows an accuracy of 93.1% in Scenario 1 and 93.7% in Scenario 2, while RDP shows 95.3% and 95.9%, respectively. The better performance of RDP over BayesANT under observed species can be explained by the latter having to account for taxonomic novelty as well, which translates into evaluating probabilities for a larger taxonomic tree. As such, BayesANT pays a price in terms of accuracy under observed taxa in favour of a much higher gain overall. Indeed, if we neglect novelty in BayesANT by fixing α and σ to 0, the accuracies of M-1, NO NEW on the observed species are 95.5% and 96.7%. See the Supporting Information for computational times and additional results on prediction accuracies, including a further benchmarking of the algorithms when the size of the training library is progressively lower.

Table 4.3: Overall predictive performances of DNA barcoding algorithms on the FinBOL library under the two testing scenarios. Values report the % of DNA sequences correctly labelled, while values in parenthesis denote the average prediction probabilities in the whole test set. Underlined values indicate the best performances.

	SCENARIO 1 - PURE RANDOM SPLIT							SCENARIO 2 - STRATIFIED RANDOM SPLIT						
MODEL	CLASS	ORDER	FAMILY	SUBFAMILY	TRIBE	GENUS	SPECIES	CLASS	ORDER	FAMILY	SUBFAMILY	TRIBE	GENUS	SPECIES
M-1	100.0	99.9	98.6	97.5	96.0	92.1	85.2	99.8	97.0	82.6	80.8	79.7	75.3	70.6
	(1)	(1)	(.98)	(.96)	(.94)	(.91)	(.82)	(.99)	(.97)	(.87)	(.83)	(.8)	(.77)	(.7)
M-2	100.0	99.9	98.4	97.2	95.8	92.4	85.4	99.8	97.1	82.1	80.3	79.1	75.7	69.8
	(1)	(1)	(.98)	(.97)	(.95)	(.93)	(.86)	(.98)	(.96)	(.88)	(.84)	(.81)	(.8)	(.74)
M-1, NO NEW	100.0	99.5	98.0	97.2	96.7	94.3	83.3	98.7	91.8	75.8	75.3	74.6	72.1	59.4
	(1)	(1)	(.99)	(.98)	(.98)	(.98)	(.92)	(1)	(.98)	(.91)	(.89)	(.89)	(.88)	(.78)
M-2, NO NEW	100.0	99.5	97.5	96.7	96.2	93.8	83.2	96.8	89.3	73.8	73.3	72.8	70.8	59.1
	(1)	(1)	(.99)	(.98)	(.98)	(.98)	(.91)	(1)	(.98)	(.91)	(.89)	(.89)	(.88)	(.74)
K-5	99.5	96.4	92.8	91.6	91.1	89.4	79.8	96.1	80.6	66.3	65.9	65.7	65.0	57.3
	(1)	(.95)	(.92)	(.91)	(.91)	(.9)	(.87)	(.99)	(.8)	(.70)	(.69)	(.68)	(.67)	(.64)
K-6	99.4	94.9	92.0	91.0	90.7	89.6	80.3	95.9	77.2	66.8	66.4	66.2	65.6	57.5
	(1)	(.96)	(.94)	(.94)	(.94)	(.93)	(.91)	(.98)	(.80)	(.73)	(.72)	(.71)	(.71)	(.68)
RDP	100.0	99.6	97.9	97.1	96.7	94.2	83.1	99.6	95.1	77.8	76.9	76.1	72.9	58.9
	(1)	(.99)	(.97)	(.96)	(.95)	(.94)	(.92)	(.99)	(.92)	(.79)	(.78)	(.77)	(.75)	(.73)

4.4 Discussion

This article has proposed a new probabilistic taxonomic classifier for DNA barcoding sequences, BayesANT, which has the key property of allowing one to build on an existing taxonomic library probabilistically. This is motivated by the fact that existing arthropod libraries are incomplete, containing reference DNA sequences for a subset of the nodes of the taxonomic tree. The potential reasons for this lack of reference barcodes are: insufficient sequencing, mislabelling and, in some extreme cases, novelty for science itself. For example, it is estimated that approximately 1.5 million, 5.5 million, and 7 million species of beetles, insects, and terrestrial arthropods, respectively, are either awaiting a proper description, do not have a reference sequence yet or are simply undiscovered (Stork, 2018), with estimates varying every year. BayesANT uses species sampling priors (Pitman, 1996) to allow for the discovery of previously unobserved branches of the tree. As such, it avoids arbitrary thresholds for novelty of other algorithms, thus characterizing uncertainty in all aspects of taxonomic classification. Probabilistic forecasts providing accurate characterizations of predictive uncertainty are said to be well calibrated (Somervuo et al., 2016). BayesANT guarantees well-calibrated predictions through a cross-validation approach.

Our method builds on a popular species sampling prior known as the Pitman-Yor process (Pitman and Yor, 1997). In its standard formulation, this prior does not take into account the taxonomic tree structure and instead treats all species as exchangeable. However, by specifying a Pitman-Yor at each level of the tree, with different parameters for each taxonomic rank, we obtain a highly flexible generative probabilistic process that can predict the probability of a query sequence to belong to different and potentially novel taxa at each level of the tree. By estimating the Pitman-Yor parameters based on the training data, we allow the process to adapt to existing knowledge about the level of diversity at each taxonomic rank.

Since taxonomic classification in ecology studies typically relies on sequencing of a relatively short region of the genome, there is necessarily substantial uncertainty in classification (Pentinsaari et al., 2020). For instance, different species often have indistinguishable nucleotide sequences in the region being sequenced, making it impossible to reliably distinguish sequences from such species relying on DNA metabarcoding alone without supplemental morphological data. This can be seen in Figure 4.7, which shows an example of both morphological and genetic variability of OTU-based clusters. In this respect, the recent development of Amplicon Sequence Variants (ASVs; Callahan et al., 2017; Bokulich et al., 2018) appears to be a promising direction to resolve these issues. ASV methods avoid the clustering step typical of OTUs, and provide better characterization of the biological variations in a dataset. In turn, this leads to an increased number of unique sequences, which BayesANT can easily handle through adequate tuning of the parameters α and σ in the Pitman–Yor prior. For instance, a large number of singletons sequences at the lowest level would lead to large values for σ_L and α_L, thus favouring the prediction of novel clades. Exploring the performance of our method on ASV datasets is an interesting potential future direction. Another possibility would be to explore ways to incorporate priors derived from phylogenetic analysis into the proposed structure. This could better resolve the ambiguities in the data and add a further clustering layer to the method.

The modelling choices made in building BayesANT reflect a balance between flexibility and pragmatism in developing an efficient off-the-shelf algorithm that can easily handle the classification of a large number of sequences. This is needed in our motivating applications to biodiversity monitoring studies that routinely collect and metabarcode samples from many different sites and multiple time points for each site. In future research, it may be helpful to consider other modelling choices which modify the Pitman-Yor structure and/or choices of kernels considered here. For example, instead of the simple multinomial kernels, it may be useful to explore pairwise similarity and latent variable-based likelihoods, for example, using the projected κ-mer decomposition of a sequence into a lower dimensional feature space. Another alternative is to specify multinomial kernels that better account for nucleotide dependencies along the sequences without excessively burdening time and memory requirements. These include, for example, mixture models as in Dunson and Xing (2009).

Taxonomic novelty due to missing branches in the reference libraries is discussed in the literature (Lan et al., 2012; Edgar, 2013; Somervuo et al., 2017). Interpretation of the detected “new”, however, is fairly delicate and context-dependent, and it requires further analyses on the sequenced DNA, such as the investigation on potential sequencing errors. Moreover, novelty is inherently related to the tree structure of the annotations in the library, which sometimes does not reflect the genetic distances between the barcodes in the nodes at a rank. The within- and the cross-taxa similarities of *Diptera*, *Lepidoptera* and *Coleoptera* depicted in Figure 4.3 are an example. In BayesANT, these distances are indirectly taken into account by the choice of kernel, which, under sufficient flexibility, can correctly discriminate between taxa. However, the creation of new clades is still biased toward the nodes that show a higher within-genetic variability (e.g. *Diptera*) than those that are more similar (like *Lepidoptera*). This is an issue shared by all taxonomic classifiers due to the current taxonomic system, and adjusting for this bias would require additional information e.g. from morphology. One potential solution in BayesANT is to specify node-specific Pitman–Yor prior parameters to counter the low/high generic variability with higher/lower prior probabilities for novel clades.

5

Conclusion

Since the introduction of the Dirichlet process 50 years ago (Ferguson, 1973), Bayesian nonparametric species sampling models have experienced rich theoretical and methodological developments in a variety of settings. In this dissertation, we have presented three different contributions to the Bayesian nonparametric field by emphasizing the usefulness of species sampling model-based frameworks in ecological applications. Our overarching goal was to increase the appeal of Bayesian nonparametric methods in applied settings, especially when the number of clusters, or novel species, is of primary interest. To this extent, there are several possible extensions to the approaches presented, both theoretical and practical, which we now summarize.

In Chapter 2, we have introduced the Stirling-gamma distribution, and we have illustrated how its adoption as a prior for the precision parameter in Dirichlet process mixtures leads to greater posterior robustness. This is desirable when one is interested in the posterior partition, like the ant sub-communities we have described in our illustrative application. There exists a rich literature on stochastic block models and their nonparametric extensions. In particular, Legramanti et al. (2022) recently introduce a general *extended* stochastic block model framework, where community detection can be further refined through the incorporation of node covariates in the exchangeable partition probability function of a species sampling model prior. Interestingly, in the covariate-dependent Dirichlet process case, the Stirling-gamma retains the conjugacy property discussed in Proposition 8. This suggests that the additional robustness granted by the Stirling-gamma can also be included in more complex dependent Dirichlet process mixtures model settings.

One important debate in recent years revolves around the consistency of Gibbs-type process mixture models in retrieving the "true" number of components from which the data are generated. In certain specific settings, Dirichlet process mixtures with a fixed precision parameter α have been shown to lack such a property (Miller and Harrison, 2014), whereas a random α can lead to consistency (Ascolani et al., 2022). Exploring the behavior of the Stirling-gamma process in such a context is an interesting future direction, especially in light of its potential connection with mixture models with a prior on the number of components discussed in Section 2.5.

In Chapter 3, we have presented a general sequential discovery framework to model accumulation curves, which is inspired by species sampling models. Its advantage lies in the fact that it is a simple and flexible framework that can capture a wide variety of accumulation curve trajectories, both allowing for finite and infinite species richness. This is especially useful, as we have shown to determine the number of potential novel OTUs one may find after having reached a certain sequencing depth. Similar tasks have also been performed within a more general *feature sampling model framework* (Masoero et al., 2021; Camerlenghi et al., 2022). One interesting extension is to consider a multivariate sequential discovery framework to model multiple locations, relying on, for example, Indian buffet-type constructions (Griffiths and Ghahramani, 2011). See Battiston et al. (2018) for a description of feature sampling models from a species sampling perspective.

Finally, Chapter 4 discusses how species sampling models can lay the foundations of efficient classification tools that account for taxonomic novelty. This is a crucial factor to account for when classifying DNA sequences since many species do not have a reference DNA barcode yet or are simply unknown to science. In this respect, BayesANT, our Bayesian nonparametric taxonomic classifier, is able to perform accurate predictions, especially when the taxa of the test sequences are unobserved in training. While BayesANT scales surprisingly well, it models DNA sequences using a relatively simple Dirichlet-multinomial kernel. Exploring more advanced approaches is an interesting and useful direction one can follow.

Appendix A

Supplementary material for Chapter 2 - Bayesian nonparametric modeling of latent partitions via Stirling-gamma priors

This Appendix contains the proofs for all the statements in Chapter 2, additional simulations, and a rejection algorithm to sample from the Stirling-gamma distribution. It is divided as follows. Section A.1 contains some preliminary lemmas that are useful for the main proofs. Section A.2 reports the proofs of the statements. Section A.3 presents additional useful theoretical results. Section A.4 presents the proofs for the formulas of the normalizing constants in Addendum I. Section A.5 describes an algorithm to generate random samples from the Stirling-Gamma distribution. Section A.6 presents an additional simulation study on the *population of partition* framework. Throughout the Chapter, we will write that $a(n) \sim b(n)$ as $n \to \infty$ to indicate that $\lim_{n\to\infty} a(n)/b(n) = 1$. While $\sim$ indicates "is distributed as" in the main document, here we employ such a slight abuse of notation to ease the readability of the statements below.

A.1 Preliminary lemmas

This Section reports some preliminary lemmas that will be useful for the proofs presented in Section A.2. We begin by recalling two asymptotic approximations for the gamma function:

$$\Gamma(m) \sim \sqrt{2\pi} e^{-m} m^{m-1/2}, \qquad m \to \infty; \tag{A.1}$$

$$\Gamma(z) \sim 1/z, \qquad z \to 0. \tag{A.2}$$

Equation (A.1) is the famous Stirling approximation (Abramowitz and Stegun, 1972, equation 6.1.37, p257). Equation (A.2) instead comes from taking the limit for $z \to 0$ to the product formula for the gamma function, since $\lim_{z\to 0} \Gamma(z) = \lim_{z\to 0} \Gamma(z+1)/z = \lim_{z\to 0} 1/z$. Then, the following lemmas hold.

Lemma 27. *For any $a \in \mathbb{R}$, we have*

$$\lim_{m\to\infty} \left(\frac{a}{\log m} + m\right)^{\frac{a}{\log m}} = e^a$$

Proof. By letting $x = \log m$ and collecting the term e^x, the limit simplifies as $\lim_{x\to\infty} e^a (a e^{-x}/x + 1)^{a/x} = e^a$. □

Lemma 28. *For any $x, z > 0$, we have*

$$\lim_{m\to\infty} \frac{(xz/\log m)_m}{(x/\log m)_m} = z e^{zx - x}$$

Proof. Recall that the ascending factorial can be defined as $(x)_a = \Gamma(x+a)/\Gamma(x)$. We can then rewrite the limit as

$$\lim_{m\to\infty} \frac{(xz/\log m)_m}{(x/\log m)_m} = \lim_{m\to\infty} \frac{\Gamma(x/\log m)}{\Gamma(xz/\log m)} \times \frac{\Gamma(xz/\log m + m)}{\Gamma(x/\log m + m)}.$$

We study each fraction separately. By relying on the approximation in equation (A.2), we have

$$\Gamma\left(\frac{a}{\log m}\right) \sim \frac{\log m}{a}, \qquad m \to \infty, \tag{A.3}$$

for any $a > 0$. Thus, the limit of the first fraction is equal to

$$\lim_{m\to\infty} \frac{\Gamma(x/\log m)}{\Gamma(xz/\log m)} = \lim_{m\to\infty} \frac{\log m}{x} \frac{xz}{\log m} = z.$$

From equation (A.2), we can also write that

$$\Gamma\left(\frac{a}{\log m} + m\right) \sim \sqrt{2\pi} e^{-\frac{a}{\log m}} \left(\frac{a}{\log m} + m\right)^{\frac{a}{\log m} + m - \frac{1}{2}}, \qquad m \to \infty, \tag{A.4}$$

for any $a > 0$. But then, thanks to the result in Lemma 27, the second fraction simplifies as

$$\begin{aligned}
&\lim_{m\to\infty} \frac{\Gamma(xz/\log m + m)}{\Gamma(x/\log m + m)} \\
&\quad = \lim_{m\to\infty} e^{\frac{-xz+x}{\log m}} \left(\frac{xz}{\log m} + m\right)^{\frac{xz}{\log m} + m - \frac{1}{2}} \left(\frac{x}{\log m} + m\right)^{-\frac{x}{\log m} - m + \frac{1}{2}} \\
&\quad = \lim_{m\to\infty} \left(\frac{xz}{\log m} + m\right)^{\frac{xz}{\log m}} \left(\frac{x}{\log m} + m\right)^{-\frac{x}{\log m}} \left(\frac{xz + m\log m}{x + m\log m}\right)^{m - \frac{1}{2}} \\
&\quad = \lim_{m\to\infty} \left(\frac{xz}{\log m} + m\right)^{\frac{xz}{\log m}} \left(\frac{x}{\log m} + m\right)^{-\frac{x}{\log m}} \\
&\quad = e^{zx - x}
\end{aligned}$$

This completes the proof. □

Lemma 28 will be useful when proving the convergence of K_m to the negative binomial distribution under the Stirling-gamma process. The next two lemmas characterize the asymptotic behavior of the normalizing constant of the Stirling-gamma distribution.

Lemma 29. *The following asymptotic approximation holds for any $x > 0$:*

$$\frac{1}{(\log m)^a} \frac{x^{a-1}}{\{(x/\log m)_m\}^b} \sim g(m, a, b) x^{a-b-1} e^{-bx}, \qquad m \to \infty,$$

where $g(m, a, b) = (2\pi)^{-b/2} (\log m)^{b-a} e^{bm} m^{-bm+b/2}$.

Proof. This asymptotic behavior follows from equations (A.3) and (A.4). In particular, by expressing the ascending factorial in the denominator as a ratio of gamma functions, we have

$$\begin{aligned}
&\lim_{m\to\infty} \frac{1}{(\log m)^a} \frac{x^{a-1}}{\{(x/\log m)_m\}^b} \\
&\quad = \lim_{m\to\infty} (2\pi)^{-\frac{b}{2}} (\log m)^{b-a}\ e^{\frac{bx}{\log m}+bm} \left(\frac{bx}{\log m} + m\right)^{-\frac{bx}{\log m} - bm + \frac{b}{2}} x^{a-b-1} \\
&\quad = \lim_{m\to\infty} (2\pi)^{-\frac{b}{2}} (\log m)^{b-a}\ e^{bm} m^{-bm+\frac{b}{2}}\ x^{a-b-1} e^{-bx},
\end{aligned}$$

where the simplifications follow from $bx/\log m \to 0$ and the limit in Lemma 27. We complete the proof by calling $g(m, b) = (2\pi)^{-b/2}(\log m)^{b-a} e^{bm} m^{-bn+b/2}$ the part that depends on m and b and not on x. □

Lemma 30. *When $a, b > 0$ and $1 < a/b < m$, the following limit holds for the normalizing constant of a Stirling-gamma distribution:*

$$\lim_{m\to\infty} \frac{g(m, a, b)}{\mathcal{S}_{a,b,m}} = \frac{b^{a-b}}{\Gamma(a-b)},$$

where $g(m, a, b)$ is defined in Lemma 29.

Proof. The proof is a direct consequence of Lemma 29 and of the monotone convergence theorem. Consider the change of variable $\alpha = x/\log m$. Then, the normalizing constant can be rewritten as

$$\mathcal{S}_{a,b,m} = \int_{\mathbb{R}_+} \frac{\alpha^{a-1}}{\{(\alpha)_n\}^b} \mathrm{d}\alpha = \int_{\mathbb{R}_+} \frac{1}{(\log m)^a} \frac{x^{a-1}}{\{(x/\log m)_n\}^b} \mathrm{d}x.$$

Provably, both integrands are monotonically decreasing in m. By monotone convergence theorem, this ensures that the limit and the integral can be interchanged. Invoking the approximation of Lemma 29, we have

$$\lim_{m\to\infty} \frac{g(m, a, b)}{\mathcal{S}_{a,b,m}} = \lim_{m\to\infty} \frac{g(m, a, b)}{g(m, a, b) \int_{\mathbb{R}_+} x^{a-b-1} e^{-bx} \mathrm{d}x} = \frac{b^{a-b}}{\Gamma(a-b)}.$$

Notice that $\mathcal{S}_{a,b,m} < \infty$ when $1 < a/b < m$, as we show in Proposition 32 below. □

We conclude the section by providing an expression for the Laplace transform of the distribution for K_m conditional on α.

Lemma 31. *Let $\theta_1, \dots, \theta_m$ be a sample from a Dirichlet process with precision parameter α. The conditional Laplace transform of the number of clusters K_m is*

$$\mathbb{E}(e^{-tK_m} \mid \alpha) = \frac{(\alpha e^{-t})_m}{(\alpha)_m}, \qquad t \geqslant 0.$$

Proof. The result follows directly from the relationship between the ascending factorial and the signless Stirling numbers of the first kind detailed in Charalambides (2005):

$$\mathbb{E}(e^{-tK_m} \mid \alpha) = \sum_{k=1}^{m} e^{-tk} \frac{\alpha^k}{(\alpha)_m} |s(m,k)| = \frac{1}{(\alpha)_m} \sum_{k=1}^{n} (\alpha e^{-t})^k |s(m,k)| = \frac{(\alpha e^{-t})_m}{(\alpha)_m},$$

for any $t > 0$. □

A.2 Main proofs

A.2.1 Proof of Proposition 1

Proof. To prove the convergence in distribution, it is sufficient to show that the Laplace transform of the quantity $\alpha \log m$ converges to that of a gamma distribution. In particular, for any $t > 0$ we have

$$\mathbb{E}(e^{-t\alpha \log m}) = \int_{\mathbb{R}_+} \frac{1}{\mathcal{S}_{a,b,m}} e^{-t\alpha \log m} \frac{\alpha^{a-1}}{\{(\alpha)_m\}^b} \mathrm{d}\alpha = \int_{\mathbb{R}_+} \frac{1}{\mathcal{S}_{a,b,m}} \frac{e^{-tx}}{(\log m)^a} \frac{x^{a-1}}{\{(\alpha)_m\}^b} \mathrm{d}x,$$

where the second equality comes from changing the integration variable to $x = \alpha \log m$. By relying on the bounded convergence Theorem, we can interchange the integral and the limit when $m \to \infty$. Thus, from Lemma 29 and 30, we can write that

$$\lim_{m\to\infty}\mathbb{E}(e^{-t\alpha\log m}) = \lim_{m\to\infty}\frac{g(m,a,b)}{\mathcal{S}_{a,b,m}}\int_{\mathbb{R}_+} x^{a-b-1}e^{-(b+t)x}\mathrm{d}x$$

$$= \frac{b^{a-b}}{\Gamma(a-b)}\int_{\mathbb{R}_+} x^{a-b-1}e^{-(b+t)x}\mathrm{d}x = \left(\frac{b}{b+t}\right)^{a-b},$$

which is the Laplace transform of a $\mathrm{Ga}(a-b,b)$ random variable. □

A.2.2 Proof of Proposition 2

Proof. By definition of expected value, we have

$$\mathbb{E}(\alpha^s) = \frac{1}{\mathcal{S}_{a,b,m}}\int_{\mathbb{R}_+}\frac{\alpha^{a+s-1}}{\{(\alpha)_m\}^b}\mathrm{d}\alpha = \frac{\mathcal{S}_{a+s,b,m}}{\mathcal{S}_{a,b,m}},$$

where the integral at the numerator is finite if and only if $0 < s < mb - a$. See Proposition 32 for a proof. □

A.2.3 Proof of Theorem 1

Proof. Let $\alpha \sim \mathrm{Sg}(a,b,m)$ in equation (1). Then,

$$\mathbb{P}(\Pi_n = \{C_1,\ldots,C_k\}) = \frac{1}{\mathcal{S}_{a,b,m}}\left\{\int_{\mathbb{R}_+}\frac{\alpha^{a+k-1}}{\{(\alpha)_m\}^b(\alpha)_n}\mathrm{d}\alpha\right\}\prod_{j=1}^k(n_j-1)!$$

Calling $\mathscr{V}_{a,b,m}(n,k) = \int_{\mathbb{R}_+}\alpha^{a+k-1}/[\{(\alpha)_m\}^b(\alpha)_n]^{-1}\mathrm{d}\alpha$ and $\mathscr{V}_{a,b,m}(n,k) = \mathcal{S}_{a,b,m}$ completes the proof. □

A.2.4 Proof of Theorem 2

Proof. We begin by showing the shape of the probability mass function for K_m. This follows directly from the formula in Antoniak (1974). In particular,

$$\mathbb{P}(K_m = k) = \int_{\mathbb{R}_+}\mathbb{P}(K_m = k \mid \alpha)p(\alpha)\mathrm{d}\alpha = \frac{1}{\mathcal{S}_{a,b,n}}\left[\int_{\mathbb{R}_+}\frac{\alpha^{a+k-1}}{\{(\alpha)_m\}^b(\alpha)_n}\mathrm{d}\alpha\right]|s(m,k)|$$

$$= \frac{\mathscr{V}_{a,b,m}(m,k)}{\mathscr{V}_{a,b,m}(1,1)}|s(m,k)|,$$

for any $k = 1, \ldots, m$. We now provide a formula for the mean and the variance of the distribution above. First, recall that the expected value and the variance of K_m conditional on α are

$$\mathbb{E}(K_m \mid \alpha) = \alpha\{\psi(\alpha + m) - \psi(\alpha)\} \tag{A.5}$$

$$\mathrm{var}(K_m \mid \alpha) = \alpha\{\psi(\alpha + m) - \psi(\alpha)\} + \alpha^2\{\psi'(\alpha + m) - \psi'(\alpha)\}, \tag{A.6}$$

where $\psi(x) = \Gamma'(x)/\Gamma(x)$ is the digamma function and $\psi'(x)$ is its derivative, called trigamma function. The property that $\mathbb{E}(K_m) = a/b$ follows immediately from Diaconis and Ylvisaker (1979): writing $\eta = \log \alpha$, we have

$$\mathbb{P}(K_m = k \mid \eta) \propto \exp\{k\eta - \mathcal{K}(\eta, m)\},$$

with $\mathcal{K}(\eta, m) = \log \Gamma(e^\eta + m) - \log \Gamma(e^\eta)$. Thus, the associated conjugate prior in the natural parametrization is $p(\eta) \propto \exp\{\tau_0 k_0 \eta - \tau_0 \mathcal{K}(\eta, m)\}$. Moreover, we have that $\mathrm{d}\mathcal{K}(\eta, n)/\mathrm{d}\eta = e^\eta\{\psi(e^\eta + n) - \psi(e^\eta)\}$. From Theorem 2 in Diaconis and Ylvisaker (1979), we have

$$\mathbb{E}(K_m) = \mathbb{E}\{\mathbb{E}(K_m \mid \eta)\} = \mathbb{E}\left\{\frac{\mathrm{d}}{\mathrm{d}\eta}\mathcal{K}(\eta, m)\right\} = E[e^\eta\{\psi(e^\eta + m) - \psi(e^\eta)\}] = k_0.$$

Substituting again $\alpha = e^\eta$ and calling $k_0 = a/b$ and $\tau_0 = b$ in the above proves the statement.

To prove the expression for the variance, we rely on the law of iterated variances, that is

$$\mathrm{var}(K_m) = \mathbb{E}\{\mathrm{var}(K_m \mid \alpha)\} + \mathrm{var}\{\mathbb{E}(K_m \mid \alpha)\}.$$

Both terms can be expressed as a function of the quantity $\mathcal{D}_{a,b,m} = \mathbb{E}[\alpha^2\{\psi'(\alpha) - \psi'(\alpha+m)\}] = \mathbb{E}\{\sum_{i=0}^{m-1} \alpha^2/(\alpha+i)\}$. To simplify the expression, we rely on integration by parts. Consider the following functions:

$$I(\alpha) = \frac{\alpha^{a+1}}{\{(\alpha)_m\}^b}, \qquad M(\alpha) = \psi(\alpha + m) - \psi(\alpha),$$

whose derivatives with respect to α are equal to

$$I'(\alpha) = \frac{(a+1)}{\alpha} I(\alpha) - bM(\alpha)I(\alpha), \qquad M'(\alpha) = \psi'(\alpha+m) - \psi'(\alpha).$$

Then, the integral simplifies as

$$\begin{aligned}
\mathcal{D}_{a,b,m} &= -\frac{1}{\mathcal{S}_{a,b,m}} \int_{\mathbb{R}_+} M'(\alpha)I(\alpha)\mathrm{d}\alpha \\
&= -\frac{1}{\mathcal{S}_{a,b,m}} \left|M(\alpha)I(\alpha)\right|_{\alpha=0}^{\infty} - \frac{b}{\mathcal{S}_{a,b,m}} \int_{\mathbb{R}_+} M(\alpha)^2 I(\alpha)\mathrm{d}\alpha + \frac{a+1}{\mathcal{S}_{a,b,m}} \int_{\mathbb{R}_+} \frac{M(\alpha)}{\alpha} I(\alpha)\mathrm{d}\alpha \\
&= -b\mathbb{E}\{\alpha^2 M(\alpha)^2\} + (a+1)\mathbb{E}\{\alpha M(\alpha)\} \\
&= -b\mathbb{E}\{\alpha^2 M(\alpha)^2\} + \frac{a(a+1)}{b} \\
&= -b\left[\mathbb{E}\{\alpha^2 M(\alpha)^2\} - \frac{a^2}{b^2}\right] + \frac{a}{b}.
\end{aligned}$$

Moreover, from equation (A.5), we can write

$$\begin{aligned}
\mathrm{var}\{\mathbb{E}(K_m \mid \alpha)\} = \mathrm{var}\{\alpha M(\alpha)\} &= \mathbb{E}\{\alpha^2 M(\alpha)^2\} - [\mathbb{E}\{\alpha M(\alpha)\}]^2 \\
&= \mathbb{E}\{\alpha^2 M(\alpha)^2\} - \frac{a^2}{b^2} \\
&= \frac{a}{b^2} - \frac{\mathcal{D}_{a,b,m}}{b},
\end{aligned}$$

where the last equality comes from plugging in the new expression for $\mathcal{D}_{a,b,m}$. Finally, from equation (A.6), we also have

$$\mathbb{E}\{\mathrm{var}(K_m \mid \alpha)\} = \mathbb{E}\{\alpha M(\alpha)\} + \mathbb{E}\{\alpha^2 M'(\alpha)\} = \frac{a}{b} - \mathcal{D}_{a,b,m}.$$

Combining the last two equalities in the law of iterated variance proves the result. □

A.2.5 Proof of Theorem 3

Proof. The proof of convergence in distribution to the negative binomial relies on Lemmas 28, 30 and 31. Substituting $\alpha = x/\log m$ in the integral, the marginal

Laplace transform of K_m is

$$\mathbb{E}(e^{-tK_m}) = \mathbb{E}\{\mathbb{E}(e^{-tK_m} \mid \alpha)\} = \frac{1}{\mathcal{S}_{a,b,m}} \int_{\mathbb{R}_+} \frac{(\alpha e^{-t})_m}{(\alpha)_m} \frac{\alpha^{a-1}}{\{(\alpha)_m\}^b} \mathrm{d}\alpha$$

$$= \frac{1}{\mathcal{S}_{a,b,m}} \int_{\mathbb{R}_+} \frac{(xe^{-t}/\log m)_m}{(x/\log m)_m} \frac{1}{(\log m)^a} \frac{x^{a-1}}{\{(x/\log m)_m\}^b} \mathrm{d}x$$

Then, the limit is

$$\lim_{m\to\infty} \mathbb{E}(e^{-tK_m}) = \lim_{m\to\infty} \frac{1}{\mathcal{S}_{a,b,m}} \int_{\mathbb{R}_+} \frac{(xe^{-t}/\log m)_m}{(x/\log m)_m} \frac{1}{(\log m)^a} \frac{x^{a-1}}{\{(x/\log m)_m\}^b} \mathrm{d}x$$

$$= \lim_{m\to\infty} \frac{g(a,b,m)}{\mathcal{S}_{a,b,m}} \int_{\mathbb{R}_+} e^{e^{-t}x - x - t} x^{a-b-1} e^{-bx} \mathrm{d}x$$

$$= \frac{e^{-t} b^{a-b}}{\Gamma(a-b)} \int_{\mathbb{R}_+} x^{a-b-1} e^{-(1+b-e^{-t})x} \mathrm{d}x$$

$$= e^{-t} \left(\frac{b}{1+b-e^{-t}} \right)^{a-b},$$

which is the Laplace transform of $1 + \mathrm{Negbin}(b/(b+1), a-b)$, whose probability mass function is

$$\mathbb{P}(K_\infty = k) = \frac{\Gamma(a-b+k-1)}{(k-1)!\Gamma(a-b)} \left(\frac{1}{b+1} \right)^{k-1} \left(\frac{b}{b+1} \right)^{a-b}, \quad k = 1, 2, \ldots$$

The limit and integral can be interchanged thanks to the bounded convergence theorem, since the Laplace transform is always bounded by 1. □

A.2.6 Proof of Proposition 3

Proof. The proof follows directly from Lemma 28 and Lemma 31. In particular, substituting $\alpha = \lambda/\log m$ in the limit, we have

$$\lim_{m\to\infty} \mathbb{E}(e^{-tK_m}) = \lim_{m\to\infty} \frac{(\alpha e^{-t})_m}{(\alpha)_m} = \lim_{m\to\infty} \frac{(\lambda e^{-t}/\log m)_m}{(\lambda/\log m)_m} = e^{e^{-t}\lambda - \lambda - t},$$

which is the Laplace transform of $1 + \mathrm{Po}(\lambda)$. □

A.2.7 Proof of Proposition 4

Proof. The statement follows trivially from Bayes theorem:

$$p(\alpha \mid \Pi_n = \{C_1, \dots, C_k\}) \propto p(\alpha)\ \mathbb{P}(\Pi_n = \{C_1, \dots, C_k\} \mid \alpha) \propto \frac{\alpha^{a-1}}{\{(\alpha)_n\}^b} \frac{\alpha^k}{(\alpha)_n}.$$

Therefore, $(\alpha \mid \Pi_n = \{C_1, \dots, C_k\}) \sim \mathrm{Sg}(a + k, b + 1, n)$. Notice that if $a, b > 0$ and $1 < a/b < n$, then also $1 < (a + k)/(b + 1) < n$, since $1 \leqslant k \leqslant n$. Thus, a proper prior implies automatically a proper posterior. □

A.2.8 Proof of Theorem 4

Proof. The proof is similar to the one of Proposition 4. In particular, we have that

$$p(\alpha \mid \Pi_n = \{C_1, \dots, C_k\}) \propto p(\alpha) \prod_{s=1}^{N} \mathbb{P}(\Pi_{n,s} = \{C_{1,s}, \dots, C_{k_s,s}\} \mid \alpha) \propto \frac{\alpha^{a-1}}{\{(\alpha)_n\}^b} \frac{\alpha^{\sum_{s=1}^{N} k_s}}{\{(\alpha)_n\}^N},$$

which proves the statement. □

A.2.9 Proof of Proposition 4

Proof. The proof naturally follows from equation (2.10) in Diaconis and Ylvisaker (1979). Alternatively, we can derive the same result via Theorem 4, integrating over $\alpha \sim \mathrm{Sg}(a + N\bar{k}, b + N, m)$. □

A.3 Additional results

A.3.1 Finiteness of the normalizing constant

Proposition 32. *The normalizing constant of a Stirling-Gamma distribution, namely*

$$\mathcal{S}_{a,b,m} = \int_{\mathbb{R}_+} \frac{\alpha^{a-1}}{\{(\alpha)_m\}^b} \mathrm{d}\alpha,$$

is finite if and only if $a, b > 0$ and $m \in \mathbb{N}$ with $1 < a/b < m$.

Proof. Let $\mathcal{S}_{a,b,m} = \mathcal{S}^{0,1}_{a,b,m} + \mathcal{S}^{1,\infty}_{a,b,m}$, with $\mathcal{S}^{\ell,u}_{a,b,m} = \int_\ell^u \alpha^{a-1}\{(\alpha)_m\}^{-b}\mathrm{d}\alpha$. To prove the statement, it is sufficient to show that both integrals are finite. For $\mathcal{S}^{0,1}_{a,b,m}$, recall that for $0 < \alpha < 1$ and $b > 0$, we have $\{(\alpha)_m\}^b = \alpha^b\{\prod_{j=1}^{m-1}(\alpha + j)\}^b > \alpha^b$, since $\prod_{j=1}^{m-1}(\alpha + j) \geqslant (m-1)! \geqslant 1$. This implies that $\mathcal{S}^{0,1}_{a,b,m} \leqslant \int_0^1 \alpha^{a-b-1}\mathrm{d}\alpha$, which is finite if and only if $a/b > 1$. For $\mathcal{S}^{1,\infty}_{a,b,m}$, instead, recall that $\lim_{\alpha\to\infty} \alpha^m\Gamma(\alpha)/\Gamma(\alpha + m) = 1$ for any $m \geqslant 1$. But then, letting $q_1(\alpha) = \alpha^{a-1}\{(\alpha)_m\}^{-b}$ and $q_2(\alpha) = \alpha^{a-1-mb}$, we have that $\lim_{\alpha\to\infty} q_1(x)/q_2(x) = \lim_{\alpha\to\infty}\{\alpha^m\Gamma(\alpha)/\Gamma(\alpha+m)\}^b = 1$, which implies that $\mathcal{S}^{1,\infty}_{a,b,m} = \int_1^\infty q_1(\alpha)\mathrm{d}\alpha < \infty$ if and only if $\int_1^\infty q_2(\alpha)\mathrm{d}\alpha < \infty$ by the limit comparison test. The latter integral is finite if and only if $a/b < m$ and $b > 0$. Both sides require $b > 0$, which in turn implies that also $a > 0$. This completes the proof. □

A.3.2 Theorem: convergence to a negative binomial via gamma prior

Theorem 33. *In the same setting of Theorem 3, let* $\alpha \sim \mathrm{Ga}(a - b, b\log m)$*. Then, the following convergence in distribution holds:*

$$K_m \to K_\infty, \quad K_\infty \sim 1 + \mathrm{Negbin}\left(a - b, \frac{b}{b+1}\right), \quad m \to \infty.$$

Proof. The proof follows similarly to the proof of Theorem 4 in the previous section, substituting again $\alpha = x/\log m$ in the integral:

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{E}(e^{-tK_m}) &= \lim_{m\to\infty}\int_{\mathbb{R}_+} \frac{(b\log m)^{a-b}}{\Gamma(a-b)}\frac{(\alpha e^{-t})_m}{(\alpha)_m}\alpha^{a-b-1}e^{-\alpha b\log m}\mathrm{d}\alpha \\
&= \lim_{m\to\infty}\frac{b^{a-b}}{\Gamma(a-b)}\int_{\mathbb{R}_+}\frac{(xe^{-t}/\log m)_m}{(x/\log m)_m}x^{a-b-1}e^{-bx}\mathrm{d}x \\
&= \frac{e^{-t}b^{a-b}}{\Gamma(a-b)}\int_{\mathbb{R}_+} x^{a-b-1}e^{-(1+b-e^{-t})x}\mathrm{d}x = e^{-t}\left(\frac{b}{1+b-e^{-t}}\right)^{a-b},
\end{aligned}$$

which is again the Laplace transform of $1 + \mathrm{Negbin}(b/(b+1), a-b)$. □

A.3.3 Proposition: tail of the Stirling-gamma distribution

Proposition 34. *The Stirling-gamma distribution* $\alpha \sim \mathrm{Sg}(a, b, m)$ *is heavy-tailed, namely*

$$\lim_{x\to\infty} e^{tx}\mathbb{P}(\alpha > x) = \infty.$$

Proof. The proof relies on the Stirling approximation of the Gamma function applied to the ascending factorial. Following equation (A.1), we have

$$\frac{1}{(x)_m} = \frac{\Gamma(x)}{\Gamma(x+m)} \sim \frac{e^{-x}x^{x-1/2}}{e^{-x-m}(x+m)^{x+m-1/2}} \sim \frac{1}{(x+m)^m}, \qquad x \to \infty,$$

since $\lim_{x\to\infty}\{x/(x+m)\}^{x-1/2} = e^{-m}$. This implies that

$$\frac{x^{a-1}}{\{(x)_m\}^b} \sim \frac{x^{a-1}}{(x+m)^{mb}} \sim \frac{1}{x^{mb-a+1}}, \qquad x \to \infty,$$

because $mb > a$ by definition. But then, we write

$$\lim_{x\to\infty} e^{tx}\mathbb{P}(\alpha > x) = \lim_{x\to\infty} \frac{1}{\mathcal{S}_{a,b,m}} \frac{\int_x^\infty \alpha^{a-1}\{(\alpha)_m\}^{-b}\mathrm{d}\alpha}{e^{-tx}} = \lim_{x\to\infty} \frac{1}{\mathcal{S}_{a,b,m}} \frac{x^{a-1}\{(x)_m\}^{-b}}{te^{-tx}}$$
$$= \lim_{x\to\infty} \frac{1}{\mathcal{S}_{a,b,m}} \frac{e^{tx}}{tx^{mb-a+1}} = \infty,$$

where the second equality follows by looking at the ratio of the derivatives with respect to x using L'Hôpital's rule. □

Corollary 35. *Let* $p_{\mathrm{Sg}}(\alpha)$ *denote the density of* $\alpha \sim \mathrm{Sg}(a, b, m)$*, and* $p_{\mathrm{Ga}}(\alpha)$ *the density of* $\alpha \sim \mathrm{Ga}(a - b, b\log m)$*. The following limit holds:*

$$\lim_{\alpha\to\infty} \frac{p_{\mathrm{Sg}}(\alpha)}{p_{\mathrm{Ga}}(\alpha)} = \infty.$$

Hence, a Stirling-gamma has a heavier right tail than the gamma distribution.

Proof. Let $\mathcal{C} = \Gamma(a-b)/\{\mathcal{S}_{a,b,m}(b\log m)^{a-b}\}$ denote the ratio of the normalizing constants. Then, the limit of the ratio of the densities is equal to

$$\lim_{\alpha\to\infty}\frac{p_{\mathrm{Sg}}(\alpha)}{p_{\mathrm{Ga}}(\alpha)} = \lim_{\alpha\to\infty}\mathcal{C}\frac{\alpha^{a-1}\{(\alpha)_m\}^{-b}}{\alpha^{a-b-1}e^{-\alpha b\log m}} = \lim_{\alpha\to\infty}\mathcal{C}\frac{\alpha^b e^{\alpha b\log m}}{(\alpha+m)^{bm}} = \infty.$$

This implies that the tail of the gamma distribution decays faster than that of the Stirling-gamma. □

A.4 Proofs of the results in Addendum I

A.4.1 Proof of Theorem 12

We break down the proof of Theorem 12 into three steps to ease readability. First, we prove that the quantity $\alpha^{a-1}/\{(\alpha)_m\}^b$ can be rewritten as a sum of partial fractions with Lemma 36. Second, we illustrate how this decomposition is useful to evaluate the normalizing constant integral via Lemma 37. Then, the proof of the statement follows using Faà di Bruno's formula.

Lemma 36. *Let a and b be integers. Then, we can write*

$$\frac{\alpha^{a-1}}{\{(\alpha)_m\}^b} = \sum_{j=1}^{m-1}\sum_{s=1}^{b}\frac{A_{s,j}}{(\alpha+j)^s}, \tag{A.7}$$

where $A_{s,j} = \rho_j^{(b-s)}(-j)/(b-s)!$ and $\rho_j^{(d)}(\alpha)$ is the d^{th} derivative of the function $\rho_j(\alpha) = \alpha^{a-b-1}/\{(\alpha+1)_{j-1}(\alpha+j+1)_{m-j-1}\}^b$.

Proof. From the definition of ascending factorial, we can write

$$\frac{\alpha^{a-1}}{\{(\alpha)_m\}^b} = \frac{\alpha^{a-b-1}}{\prod_{i=1}^{m-1}(\alpha+i)^b},$$

which is ratio of polynomials whose roots for the denominators are $-1,\dots,-m+1$. Following the algorithm of Section 2.102, page 66 in Gradshteyn and Ryzhik (2007), we can rewrite the above as the sum of partial fractions in equation (A.7), where the coefficients $A_{s,j}$ of the expansion depend on the derivatives of the function

$$\rho_j(\alpha) = \frac{\alpha^{a-b-1}}{\prod_{i=1}^{m-1}(\alpha+i)^b}(\alpha+j)^b, \qquad (j = 1, \dots, m-1),$$

evaluated at the solution $\alpha = -j$. In particular, we have $A_{s,j} = \rho_j^{(b-s)}(-j)/(b-s)!$ We calculate their exact values below, after noticing that

$$\rho_j(\alpha) = \frac{\alpha^{a-b-1}}{\prod_{i=1}^{j-1}(\alpha+i)^b \prod_{i=j+1}^{m-1}(\alpha+i)^b} = \frac{\alpha^{a-b-1}}{\{(\alpha+1)_{j-1}(\alpha+j+1)_{m-j-1}\}^b}. \tag{A.8}$$

□

Lemma 37. *The normalizing constant of the Stirling-gamma $\alpha \sim \mathrm{Sg}(a, b, m)$ where $a, b \in \mathbb{N}$ can be expressed as*

$$\mathcal{S}_{a,b,m} = \sum_{j=1}^{m-1}\sum_{s=1}^{b} A_{s,j}\phi_s(j), \qquad \phi_s(j) = \begin{cases} -\log j, & s = 1, \\ j^{1-s}/(s-1) & s = 2, 3, \dots \end{cases}$$

where $A_{s,j}$ are defined in Lemma 36.

Proof. Recall that $\int_{\mathbb{R}_+} 1/(\alpha+j)^s \mathrm{d}\alpha = 1/\{(s-1)j^{s-1}\}$ for $s > 1$, while

$$\int_{\mathbb{R}_+} \frac{1}{\alpha+j}\mathrm{d}\alpha = \lim_{\alpha\to\infty} \log(\alpha+j) - \log j.$$

Define the functions

$$\phi_1(j) = -\log j, \qquad \phi_s(j) = \frac{1}{(s-1)j^{s-1}}, \quad s = 2, 3, \dots$$

From Lemma 36, we have

$$\mathcal{S}_{a,b,m} = \int_{\mathbb{R}_+} \frac{\alpha^{a-1}}{\{(\alpha)_m\}^b}\mathrm{d}\alpha = \int_{\mathbb{R}_+} \sum_{j=1}^{m-1}\sum_{s=1}^{b} \frac{A_{s,j}}{(\alpha+j)^s}\mathrm{d}\alpha = \sum_{j=1}^{m-1}\sum_{s=1}^{b} A_{s,j}\phi_s(j)$$

The last equality holds because $\sum_{j=1}^{m-1} A_{1,j} = 0$, which makes the limit of the sum of logarithms necessarily equal to zero. This can be shown by contradiction. If $\sum_{j=1}^{m-1} A_{1,j} \neq 0$, then necessarily $\left|\sum_{j=1}^{m-1} A_{1,j} \lim_{\alpha\to\infty} \log(\alpha + j)\right| = \infty$, which implies that $|\mathcal{S}_{a,b,m}| = \infty$. However, this contradicts Proposition 32, which states that $0 < \mathcal{S}_{a,b,m} < \infty$ for appropriate choices of a, b and m. Hence, the divergence of each logarithmic term is compensated by the alternating sum. For an alternative proof of why this happens, refer to Zhu and Luo (2021) and references therein. □

Proof of Theorem 12. Lemma 36 and Lemma 37 show that we can write the normalizing constant as a sum of logarithms. It remains to calculate the values for the coefficients $A_{s,j} = \rho_j^{(b-s)}(-j)/(b-s)!$ We start by rewriting equation (A.8) as

$$\rho_j(\alpha) = \exp\left[(a-b-1)\log\alpha - b\left\{\log(\alpha+1)_{j-1} + \log(\alpha+j+1)_{m-j-1}\right\}\right].$$

Recalling that $\frac{\mathrm{d}}{\mathrm{d}x}\log(x)_n = \psi(x+n) - \psi(x)$ and that $\psi(x+1) = \psi(x) + 1/x$, we have that

$$\rho_j'(\alpha) = \rho_j(\alpha)h_j(\alpha), \quad h_j(\alpha) = \frac{a-1}{\alpha} - b\{\psi(\alpha+m) - \psi(\alpha) - \psi(\alpha+j+1) + \psi(\alpha+j)\}.$$

Then, the s^{th} derivative of $\rho_j(\alpha)$ can be expressed via Faà di Bruno's formula as

$$\rho_j^{(s)}(\alpha) = \rho_j(\alpha)B_s\{h_j(\alpha), h_j'(\alpha), \ldots, h_j^{(s-1)}(\alpha)\}, \tag{A.9}$$

where

$$h_j^{(d)}(\alpha) = (-1)^d d!\frac{a-1}{\alpha^{d+1}} - b\{\psi^{(d)}(\alpha+m) - \psi^{(d)}(\alpha) - \psi^{(d)}(\alpha+j+1) + \psi^{(d)}(\alpha+j)\} \tag{A.10}$$

for $d = 0, \ldots, s-1$, and

$$B_s(x_1, \ldots, x_s) = \sum_{(j_1,\ldots,j_s)\in I_s} \frac{s!}{j_1!j_2!\cdots j_s!}\left(\frac{x_1}{1!}\right)^{j_1}\left(\frac{x_2}{2!}\right)^{j_2}\cdots\left(\frac{x_s}{s!}\right)^{j_s} \tag{A.11}$$

is the complete exponential Bell polynomial of order s and I_s is the set of all nonnegative integers $(j_1, \ldots, j_s)$ that satisfy $j_1 + 2j_2 + \ldots + sj_s = s$. It remains to evaluate the function $\rho_j^{(s)}(\alpha)$ when $\alpha = -j$. We start by noticing that

$$\prod_{i=1}^{j-1}(-j+i)^b = \{(1-j)(2-j)\cdots(-2)(-1)\}^b = (-1)^{bj}\Gamma(j)^b,$$

$$\prod_{i=j+1}^{m-1}(-j+i)^b = \{1 \cdot 2 \cdot 3 \cdots (m-j-1)\}^b = \Gamma(m-j)^b.$$

Plugging in the above in equation (A.8), we have

$$\rho_j(-j) = \frac{(-j)^{a-b-1}}{\prod_{i=1}^{j-1}(-j+i)^b \prod_{i=j+1}^{m-1}(-j+i)^b} = \frac{(-1)^{a-b(j+1)-1} j^{a-b-1}}{\Gamma(j)^b \Gamma(m-j)^b}. \tag{A.12}$$

Moreover, we have that the derivatives in equation (A.10) can be rewritten as

$$h_j^{(d)}(\alpha) = (-1)^d d! \left[\frac{a-1}{\alpha^{d+1}} - b \left\{ \sum_{i=0}^{j-1} \frac{1}{(\alpha+i)^{d+1}} + \sum_{i=j+1}^{m-1} \frac{1}{(\alpha+i)^{d+1}} \right\} \right]. \tag{A.13}$$

Calling $h_{j,d} = h_j^{(d)}(-j)$, we then have that

$$\frac{h_{j,d+1}}{d!} = -\frac{(a-1)}{j^{d+1}} - b(H_{m-j-1,d+1} - H_{j,d+1}), \tag{A.14}$$

with $H_{j,s} = \sum_{i=1}^{j} 1/i^s$ the j^{th} generalized harmonic number of order s. Plugging equations (A.12) and (A.14) into (A.9) and recalling that $A_{s,j} = \rho_j^{(b-s)}(-j)/(b-s)!$, we write the partial fraction decomposition coefficients as

$$A_{s,j} = \frac{1}{(b-s)!} \frac{(-1)^{a-b(j+1)-1} j^{a-b-1}}{\{\Gamma(j)\Gamma(m-j)\}^b} B_{b-s}(h_{j,1}, h_{j,2}, \ldots, h_{j,b-s}).$$

Combining this expression with Lemma 37 yields the final form

$$\begin{aligned}\mathcal{S}_{a,b,m} &= \sum_{j=1}^{m-1} \sum_{s=1}^{b} A_{s,j} \phi_s(j) \\ &= \sum_{j=1}^{m-1} \sum_{s=1}^{b} \frac{(-1)^{a-b(j+1)-1} j^{a-b-1}}{\{\Gamma(j)\Gamma(m-j)\}^b} \frac{B_{b-s}(h_{j,1}, h_{j,2}, \ldots, h_{j,b-s})}{(b-s)!} \phi_s(j).\end{aligned}$$

Rearranging the terms and collecting the Bell polynomials into the quantity

$$\mathscr{S}_{b,j}(x_1, \ldots, x_b) = \sum_{s=1}^{b} \frac{B_{b-s}(x_1, \ldots, x_{b-s})}{(b-s)!} \phi_s(j),$$

yields the desired result. □

A.4.2 Proof of Corollary 13

Proof. Setting $b = 1$ in the above statement proves the statement. To see why, recall that $A_{s,j} = \rho_j^{(b-s)}(-j)/(b-s)!$ Since $b = 1$, we only have $A_{1,j} = \rho_j(-j)$, whose general formula is provided in equation (A.12). □

A.4.3 Proof of Theorem 14

The proof of Theorem 14 follows the same reasoning as the one of Theorem A1. We discuss the highlights with the help of the following statements.

Lemma 38. *Let $a, b, m, n, k \in \mathbb{N}$ with $1 < a/b < m$, and $1 \geqslant k \geqslant n$, and call $M = \min\{n, m\}$ and $\ell = |n - m|$. Then, we can write*

$$\frac{\alpha^{a+k-1}}{\{(\alpha)_m\}^b (\alpha)_n} = \sum_{j=1}^{M-1} \sum_{s=1}^{b+1} \frac{T_{s,j}}{(\alpha + j)^s} + \sum_{i=0}^{\ell-1} \frac{U_i}{\alpha + M + i}, \tag{A.15}$$

where $T_{s,j} = \tau_j^{(b+1-s)}(-j)/(b+1-s)!$ and $\tau_j^{(d)}(\alpha)$ is the d^{th} derivative of the function

$$\tau_j(\alpha) = \frac{\alpha^{a+k-b-2}}{\{(\alpha+1)_{j-1}(\alpha+j+1)_{M-j-1}\}^{b+1}(\alpha+M)_\ell}, \tag{A.16}$$

$U_i = u_i(-M - i)$ and

$$u_i(\alpha) = \frac{\alpha^{a+k-b-2}}{\prod_{j=1}^{M-1}(\alpha+j)^{b+1} \prod_{t=0}^{i-1}(\alpha+M+t) \prod_{v=i+1}^{\ell-1}(\alpha+M+v)} \tag{A.17}$$

Proof. The proof follows from the same reasoning discussed in Lemma 36, which is a direct consequence of the algorithm of Section 2.102, page 66 in Gradshteyn and Ryzhik (2007), after writing that

$$\frac{\alpha^{a+k-1}}{\{(\alpha)_m\}^b(\alpha)_n} = \frac{\alpha^{a+k-1}}{\{(\alpha)_M\}^{b+1}(\alpha+M)_\ell} = \frac{\alpha^{a-b+k-2}}{\prod_{j=1}^{M-1}(\alpha+j)^{b+1}\prod_{i=0}^{\ell-1}(\alpha+M+i)}. \quad \text{(A.18)}$$

The quantities $\tau_j(\alpha)$ and $u_j(\alpha)$ are obtained by multiplying equation (A.18) by $(\alpha + j)^{b+1}$ and $(\alpha + M + i)$, respectively, and simplifying appropriately. □

Lemma 39. *The coefficients $\mathscr{V}_{a,b,m}(n,k)$ when $a, b \in \mathbb{N}$ are expressed as follows:*

$$\mathscr{V}_{a,b,m}(n,k) = \sum_{j=1}^{M-1}\sum_{s=1}^{b+1} T_{s,j}\phi_s(j) - \sum_{i=0}^{\ell-1} U_i \log{(M+i)},$$

where $T_{s,j}$ and U_i are defined in Lemma 38 and the coefficients $\phi_s(j)$ are defined in Lemma 37.

Proof. Follow the same line of reasoning as the proof of Lemma 37, since we have rational functions as integrands. In particular, we point out that the coefficients multiplying the logarithms are such that $\sum_{j=1}^{M-1} T_{1,j} + \sum_{i=0}^{\ell-1} U_i = 0$. Again, this must happen because the diverging logarithms resulting from the integration must cancel each other out because $\mathscr{V}_{a,b,m}(n,k) < \infty$ by definition. □

We are now ready to provide proof of the final statement.

Proof of Theorem 14. First of all, we provide a simpler expression for U_i. This is equal to $u_i(-M-i)$ in equation (A.17). In particular, the quantities at the denominator simplify as

$$\prod_{j=1}^{M-1}(-M-i+j)^{b+1} = \{(-M-i+1)(-M-i+2)\cdots(-i-1)\}^{b+1} = \{(i+1)_{M-1}\}^{b+1}$$

$$\prod_{t=0}^{i-1}(-M-i+t) = (-i)(-i+1)\cdots(-2)(-1) = (-1)^i\Gamma(i+1)$$

$$\prod_{v=i+1}^{\ell-1}(-M-i+v) = (1)(2)\cdots(\ell-1+i) = \Gamma(\ell-i).$$

This implies that

$$U_i = (-1)^{a-b+k-2+i}\frac{(M+i)^{a-b+k-2}}{\{(i+1)_{M-1}\}^{b+1}\Gamma(i+1)\Gamma(\ell-i)},$$

which is the generic coefficient in the second sum of logarithms. As for $T_{s,j}$, we rely on a similar argument as the proof of Theorem A1, which depends on Faà di Bruno's formula. Rewriting $\tau_j(\alpha)$ as

$$\begin{aligned}\tau_j(\alpha) &= \exp\{\log\tau_j(\alpha)\}\\ &= \exp[(a+k-b-2)\log\alpha-(b+1)\{\log(\alpha+1)_{j-1}+\log(\alpha+j+1)_{M-j-1}\}\\ &\qquad -\log(\alpha+M)_\ell],\end{aligned}$$

we have that $\tau_j'(\alpha) = \tau_j(\alpha)g_j(\alpha)$ with

$$\begin{aligned}g_j(\alpha) = \frac{a+k-1}{\alpha}&-(b+1)\{\psi(\alpha+j)-\psi(\alpha)+\psi(\alpha+M)-\psi(\alpha+j+1)\}\\ &-\psi(\alpha+M+\ell)+\psi(\alpha+M)\end{aligned}$$

whose dth derivative is equal to

$$\begin{aligned}g_j^{(d)}(\alpha) &= (-1)^d d!\frac{a+k-1}{\alpha^{d+1}}-(b+1)\{\psi^{(d)}(\alpha+j)-\psi^{(d)}(\alpha)+\psi^{(d)}(\alpha+M)\\ &\qquad -\psi^{(d)}(\alpha+j+1)\}-\psi^{(d)}(\alpha+M+\ell)+\psi^{(d)}(\alpha+M)\\ &= (-1)^d d!\left[\frac{a+k-1}{\alpha^{d+1}}-(b+1)\left\{\sum_{i=0}^{j-1}\frac{1}{(\alpha+i)^{d+1}}+\sum_{i=j+1}^{M-1}\frac{1}{(\alpha+i)^{d+1}}\right\}+\sum_{i=0}^{\ell-1}\frac{1}{\alpha+M+i}\right].\end{aligned}$$

Using Faà di Bruno's formula and recalling the complete exponential Bell polynomial in equation (A.11), we obtain that

$$\tau_j^{(d)}(\alpha) = \tau_j(\alpha) B_d\left(g_j(\alpha), g_j'(\alpha), \ldots, g_j^{(d-1)}(\alpha)\right). \tag{A.19}$$

We are finally ready to calculate $T_{s,j} = \tau_j^{(b+1-s)}(-j)/(b+1-s)!$. Calling $g_{j,d+1} = g_j^{(d)}(-j)$, we can write that

$$\frac{g_{j,d+1}}{d!} = -\frac{a+k-1}{j^{d+1}} + bH_{M-j-1,d+1} - (b+1)H_{j,d+1} + H_{M-j+\ell-1,d+1}, \tag{A.20}$$

since $\sum_{i=0}^{\ell-1} 1/(M+i-j)^{d+1} = H_{M-j+\ell-1,d+1} - H_{M-j-1,d+1}$, and $H_{j,s} = \sum_{i=1}^{j} 1/i^s$ is again the jth generalized harmonic number of order s. With similar calculations as the one for equation (A.12), we also have that

$$\begin{aligned}\tau_j(-j) &= \frac{(-j)^{a-b+k-2}}{\{\prod_{i=1}^{j-1}(-j+i)\prod_{i=j+1}^{m-1}(-j+i)\}^{b+1}(M-j)_\ell} \\ &= (-1)^{a-b+k-2+b(j+1)}\frac{j^{a-b+k-2}}{\{\Gamma(j)\Gamma(m-j)\}^{b+1}(M-j)_\ell}.\end{aligned} \tag{A.21}$$

Plugging equations (A.21), (A.20) and (A.19) into the formula for $T_{s,j}$ yiels

$$T_{s,j} = \frac{(-1)^{a-b+k-2+b(j+1)}j^{a-b+k-2}}{\{\Gamma(j)\Gamma(m-j)\}^{b+1}(M-j)_\ell}\frac{B_{b+1-s}(g_{j,1}, g_{j,2}, \ldots, g_{j,b+1-s})}{(b+1-s)!}.$$

The rest of the proof follows by regrouping the coefficients in a similar manner as the one in the proof for Theorem 12. □

A.5 Random sample generation for the Stirling-gamma distribution

In this Section, we illustrate a strategy to draw random samples from the Stirling-gamma distribution. Unfortunately, a rejection sampler using a gamma Ga$(a-b, b\log m)$ as the proposal is not feasible because the Stirling-gamma has heavier tails; see Corollary 35 above. As such, an ideal proposal must be itself a heavy-tailed distribution from which sampling is relatively easy. Another valid alternative is to use rejection sampling via the ratio of uniforms method, which can be adapted to any distribution for which the density $f(x)$ and the function $x^2 f(x)$ can be maximized; see (Devroye, 1986a) for details. Unfortunately, the density of the Stirling-gamma is not always bounded. To see this, let

$$G(\alpha) = \frac{\alpha^{a-1}}{\{(\alpha)_m\}^b} = \frac{\alpha^{a-b-1}}{\prod_{i=1}^{m-1}(\alpha+i)^b} \tag{A.22}$$

with $\alpha \geqslant 0$ denote the unnormalized density function of the Stirling-gamma. Then, as $\alpha \to 0$, we have that

$$\lim_{\alpha \to 0} G(\alpha) = \begin{cases} \infty, & \text{if} \quad a-b<1, \\ \Gamma(m)^b, & \text{if} \quad a-b=1, \\ 0, & \text{if} \quad a-b>1, \end{cases}$$

which means that $G(\alpha)$ admits a maximum if and only if $a-b \geqslant 1$. Due to such behavior, we must consider two different sampling strategies depending on the value of $a-b$, one where $a-b \geqslant 1$, and another where $a-b<1$. In the first case, we rely on the ratio of uniforms method, which we illustrate in Algorithm 1 below.

Algorithm 1: Rejection sampler for the Stirling-gamma distribution when $a-b \geqslant 1$

```
1 Let $M_u = \max_{\alpha \geqslant 0} G(\alpha)$ and $M_v = \max_{\alpha \geqslant 0} \alpha^2 S(\alpha)$, with $G(\alpha)$ as in (A.22);
2 Sample $u$ uniformly in $[0, M_u^{1/2}]$;
3 Sample $v$ uniformly in $[0, M_v^{1/2}]$;
4 If $2 \log u \leqslant \log G(v/u)$, set $\alpha = v/u$. Otherwise, return to 2;
5 Output: a sample from $\alpha \sim \mathrm{Sg}(a, b, m)$.
```

The above strategy yields valid samples from the Stirling-gamma distribution as long as $a-b \geqslant 1$. When $a-b<1$, we instead rely on a rejection sampler with a *generalized beta prime distribution* $\alpha \sim \mathrm{BeP}(a_0, b_0, r)$ as proposal, whose density is

$$p_{\mathrm{BeP}}(\alpha) = \frac{(\alpha/r)^{a_0-1}(1+\alpha/r)^{-a_0-b_0}}{r\mathcal{B}(a_0, b_0)}$$

with $\alpha > 0$ and $\mathcal{B}(a_0, b_0) = \Gamma(a_0)\Gamma(b_0)/\Gamma(a_0+b_0)$ denoting the Beta function. Sampling from $\alpha \sim \mathrm{BeP}(a_0, b_0, r)$ can be performed by letting $\alpha = rx/(1-x)$ with $x \sim \mathrm{Be}(a_0, b_0)$. Thus, our goal is to find appropriate choices for a_0, b_0, and r such that the generalized beta prime (i) has an asymptote as $\alpha \to \infty$, (ii) is heavy-tailed and (iii) "covers" the Stirling-gamma density, that is $p_{\mathrm{Sg}}(\alpha)/p_{\mathrm{BeP}}(\alpha) \leqslant M < \infty$, with a sufficiently small M. All these can be obtained by finding an appropriate bound for $G(\alpha)$ in equation (A.22), as we now show. Let $r(\alpha)$ be the function of $\alpha \geqslant 0$ defined as

$$r(\alpha) = \left\{\prod_{i=1}^{m-1}(\alpha+i)\right\}^{1/(m-1)} - \alpha.$$

It is easy to see that $r(\alpha)$ is a monotonically increasing function of α whose minimum value is $r = r(0) = \Gamma(m)^{1/(m-1)}$. Since $(\alpha + r(\alpha))^{m-1} = \prod_{i=1}^{m-1}(\alpha+i)$, then

$$(\alpha + r)^{m-1} \leqslant \prod_{i=1}^{m-1}(\alpha+i), \qquad r = \Gamma(m)^{1/(m-1)}, \tag{A.23}$$

for every $\alpha > 0$. But then, we can bound the unnormalized density function in equation (A.22) as follows

$$G(\alpha) = \frac{\alpha^{a-b-1}}{\prod_{i=1}^{m-1}(\alpha+i)^b} \leqslant \frac{\alpha^{a-b-1}}{(\alpha+r)^{b(m-1)}} = r^{a-mb-1}\frac{(\alpha/r)^{a-b-1}}{(1+\alpha/r)^{b(m-1)}} = r^{a-mb-1}Q(\alpha).$$

The $Q(\alpha)$ defined on the right-hand side of the above inequality is the kernel of a generalized beta prime distribution $\alpha \sim \mathrm{BeP}(a-b, mb-a, r)$, with $r = \Gamma(m)^{1/(m-1)}$.

Under such a proposal, we have

$$\begin{aligned}\frac{p_{\mathrm{Sg}}(\alpha)}{p_{\mathrm{BeP}}(\alpha)} &= \frac{r\beta(a-b, mb-a)}{\mathcal{S}_{a,b,m}} \frac{\alpha^{a-b-1}}{\prod_{i=1}^{m-1}(\alpha+i)^b} \frac{(1+\alpha/r)^{b(m-1)}}{(\alpha/r)^{a-b-1}} \\ &= \frac{(\alpha+r)^{b(m-1)}}{(\alpha+1)_{m-1}^{b}} \frac{\mathcal{B}(a-b, mb-a)}{r^{mb-a}\mathcal{S}_{a,b,m}} \\ &\leqslant \frac{\mathcal{B}(a-b, mb-a)}{r^{mb-a}\mathcal{S}_{a,b,m}} = M < \infty,\end{aligned}$$

where the inequality follows from equation (A.23) after writing $\prod_{i=1}^{m-1}(\alpha+i) = (\alpha+1)_{m-1}$. This makes the acceptance function in an accept-reject algorithm equal to

$$A(\alpha) = \frac{p_{\mathrm{Sg}}(\alpha)}{Mp_{\mathrm{BeP}}(\alpha)} = \frac{(\alpha+r)^{b(m-1)}}{(\alpha+1)_{m-1}^{b}}, \qquad r = \Gamma(m)^{1/(m-1)}. \tag{A.24}$$

The resulting sampling procedure is detailed in Algorithm 2 below.

```
Algorithm 2: Rejection sampler for the Stirling-gamma distribution when
$a - b < 1$
1 Let $r = \Gamma(m)^{1/(m-1)}$ and $A(\alpha)$ be as in equation (A.24);
2 Sample $x \sim \mathrm{Be}(a-b, mb-a)$ and set $y = rx/(1-x)$;
3 Sample $u$ uniformly in $[0, 1]$;
4 If $\log u \leqslant \log A(y)$, set $\alpha = y$. Otherwise, return to 2;
5 Output: a sample from $\alpha \sim \mathrm{Sg}(a, b, m)$.
```

To ease reproducibility, we implement both Algorithm 1 and Algorithm 2 in the `R` package `ConjugateDP` via the function `rSg`. Table A.1 and Table A.2 report the acceptance rates for the samplers above for selected values of a, b, and m. We see that the acceptance rates for Algorithm 1 range between 0.3 and 0.7, which is fairly large considering that the ratio of uniforms is a method that is not tailored specifically to the Stirling-gamma. As for Algorithm 2, we see that the rates are much larger and range between 0.4 and 0.95. Indeed, this is due to the high similarity between the generalized beta prime distribution and the Stirling-gamma when $a - b < 1$.

In principle, one could use Algorithm 2 to draw samples from $\alpha \sim \mathrm{Sg}(a, b, m)$ for any choice of a and b, since the acceptance function in equation (A.24) is always a valid one. However, we noticed that when $a - b \geqslant 1$, the acceptance probability $1/M = r^{mb-a}\mathcal{S}_{a,b,m}/\mathcal{B}(a - b, mb - a)$ with the generalized beta prime proposal is particularly low. Therefore, we still rely on the ratio of uniforms for the general case when $a - b \geqslant 1$.

Table A.1: Acceptance probabilities for Algorithm 1 under varying a, b and m, when $a - b \geqslant 1$. Values are obtained by averaging the acceptance rate obtained in 1000 trials of Algorithm 1 under 100 replicates. Standard deviations were all around 0.01 and therefore are omitted from the table. Empty cells indicate when $1 < a/b < m$ and $a - b \geqslant 1$ are violated.

	$m = 100$				$m = 1000$			
	$a = 2$	$a = 3$	$a = 10$	$a = 15$	$a = 2$	$a = 3$	$a = 10$	$a = 15$
$b = 0.2$	0.756	0.701	0.544	0.594	0.742	0.668	0.425	0.358
$b = 1$	0.679	0.724	0.445	0.377	0.680	0.717	0.419	0.346
$b = 1.5$		0.754	0.446	0.372		0.752	0.427	0.349
$b = 5$			0.528	0.394			0.523	0.386

Table A.2: Acceptance probabilities for Algorithm 2 under varying a, b and m, when $a - b < 1$. Values are obtained by averaging the acceptance rate obtained in 1000 trials of Algorithm 2 under 100 replicates. Standard deviations were all around 0.01 and therefore are omitted from the table.

	$m = 100$				$m = 1000$			
	$a = 0.2$	$a = 0.6$	$a = 0.7$	$a = 1$	$a = 0.2$	$a = 0.6$	$a = 0.7$	$a = 1$
$b = 0.1$	0.949	0.788	0.760	0.678	0.911	0.638	0.593	0.458
$b = 0.2$		0.799	0.757	0.655		0.683	0.622	0.476
$b = 0.5$		0.940	0.883	0.733		0.907	0.822	0.609
$b = 0.6$			0.938	0.775			0.905	0.670

A.6 Simulation study

In this Section, we present a simulation study within the same *population of partition* framework introduced in Section 3 of the main paper. In particular, our goal is to show how the conjugate Stirling-gamma prior can lead to borrowing of information when inferring the latent partition across multiple networks, thus reducing uncertainty. Consider the same stochastic block model setting of Section 4, namely

$$\mathbb{P}(Y_{i,j,s} = 1 \mid Z_{i,s} = h, Z_{j,s} = h', \nu) = \nu_{h,h',s}, \quad \nu_{h,h',s} \sim \mathrm{Be}(1,1), \tag{A.25}$$

where $Y_{i,j,s}$ is a binary random variable indicating an edge between nodes i and j in networks $s = 1, \dots N$. The variables $Z_{i,s}$ denote cluster assignment, with $Z_{i,s} = h$ if and only if $i \in C_{h,s}$ in network s, and $\nu_{h,h',s}$ denotes the edge probabilities in the block identified by clusters $C_{h,s}$ and $C_{h',s}$. We model the latent partition in each network independently as follows:

$$\mathbb{P}(\Pi_{n,s} = \{C_{1,s}, \dots, C_{k_s,s}\} \mid \alpha_s) = \frac{\alpha_s^{k_s}}{(\alpha_s)_n} \prod_{j=1}^{k_s} (n_{j,s} - 1)! \quad (s = 1, \dots, N),$$

where α_s is the precision parameter specific to partition α_s.

Within this framework, we are interested in investigating the impact of different choices of precision parameters $\alpha_1, \dots, \alpha_N$ on the inferred latent partition. We consider three priors: α_s is fixed and equal across networks, α_s is random with $\alpha_s \sim \mathrm{Sg}(a, b, n)$ separately for each network, and the precision is pooled across networks, namely $\alpha_1 = \dots = \alpha_N = \alpha \sim \mathrm{Sg}(a, b, n)$. In the third case, the shared α induces borrowing of information since the number of clusters k_s in every network contributes to the posterior distribution in Theorem 4.

We simulate $N = 6$ networks of $n = 100$ nodes from the stochastic block model in equation (A.25). The true partition is generated by randomly dividing the nodes between six clusters with assignment probabilities drawn from a Dirichlet distribution $\mathrm{Dir}(10, 10, 10, 10, 10, 10)$. Binary edges are independently simulated with probabilities $(\nu_{h,h,1}, \dots, \nu_{h,h,N}) = (0.95, 0.90, 0.85, 0.80, 0.75, 0.70)$ for nodes within the same cluster, and $(\nu_{h,h',1}, \dots, \nu_{h,h,N}) = (0.05, 0.10, 0.10, 0.15, 0.15, 0.30)$ for any $h \neq h'$.

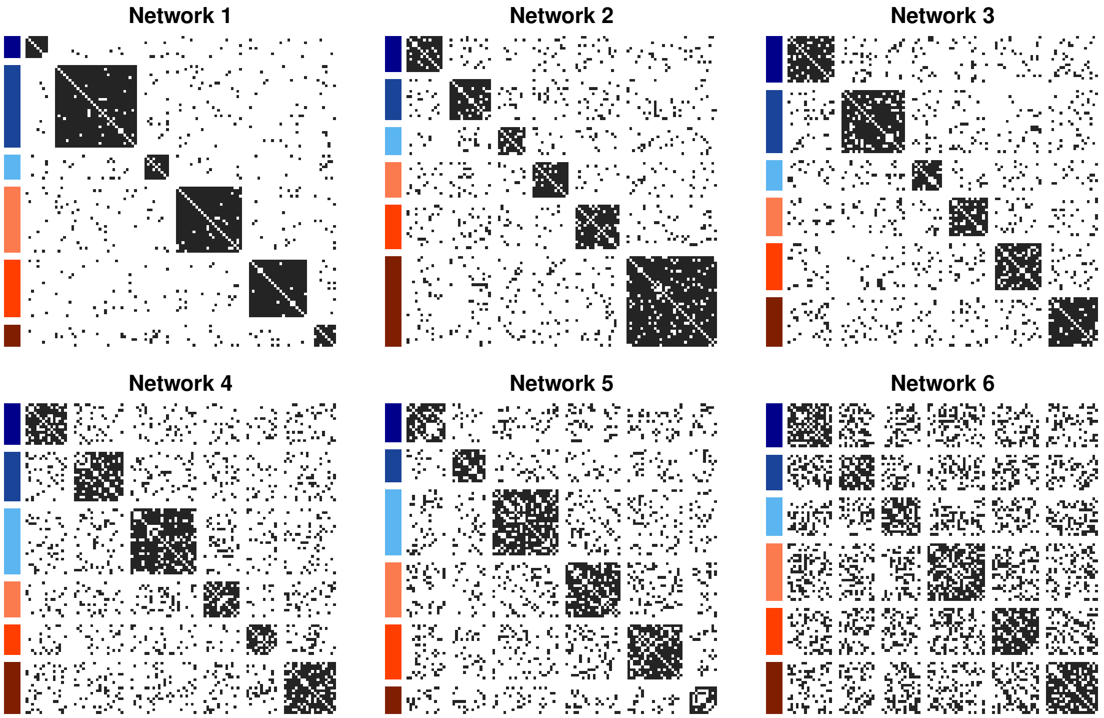

FIGURE A.1: Simulated networks of size $n = 100$ nodes. Columns and rows represent nodes of each network, and black dots indicate the edges. Nodes are sorted according to the true cluster assignment, highlighted by the different colors on the left of each plot.

This allows each network to have a different block structure with decreasing signal-to-noise ratios. As such, we expect to infer the true communities in Networks 5 and 6 with a higher uncertainty than in Networks 1 and 2. Figure A.1 displays the six generated datasets. Black points indicate an edge between each pair of nodes. Rows and columns have been sorted according to the true cluster assignment for better visualization. We set $\alpha_s = 7.5$ in the fixed case and $a = 6$ and $b = 0.3$ in random and pooled cases, so that $\mathbb{E}(K_n) = 20$ in all priors. Inference is performed by running a collapsed Gibbs sampler as in Legramanti et al. (2022) for 10,000 iterations, treating the first 2,000 as burn-in. The full conditional for α_s in the random case and for α in the pooled case are reported in Proposition 4 and Theorem 4 in the main manuscript, respectively.

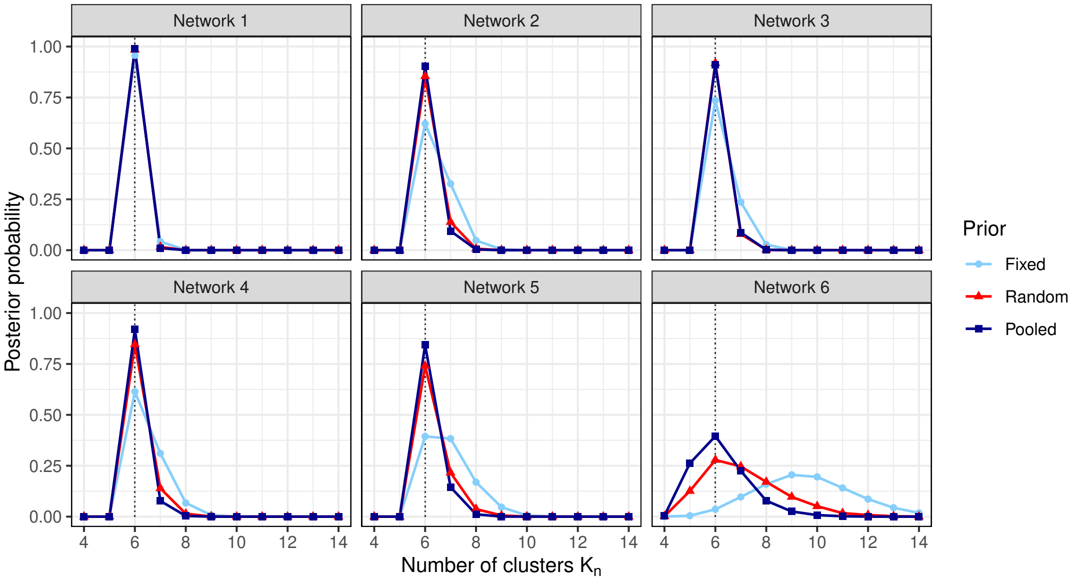

FIGURE A.2: Posterior distribution of the number of clusters K_n detected in each simulated network in the three cases: $\alpha_s = 7.5$ (light blue), $\alpha_s \sim \text{Sg}(6, 0.3, 100)$ independently in each network (red), and $\alpha_s = \alpha \sim \text{Sg}(6, 0.3, 100)$ (blue). The dotted vertical line highlights the true number of communities.

Figure A.2 displays the posterior distribution of the number of detected clusters K_n in each dataset for the three choices of precision parameter. Except for Network 6, the posterior mode of K_n coincides with the truth in each model. However, the pooled case shows lower uncertainty than the random one, thanks to the borrowing of information granted by the common α. In the fixed cases, instead, K_n explodes as the signal-to-noise ratio decreases. This is particularly evident in Network 6, which confirms the lack of robustness of Dirichlet process mixtures with fixed α. To further highlight these differences, we calculate the average adjusted Rand index for the posterior partition retrieved by the three models with respect to the truth. This equals 0.943 for the pooled case, 0.940 for the random, and 0.929 for the fixed, indicating that pooling α yields a better estimate across networks.

Appendix B

Supplementary material for Chapter 3 - Bayesian modeling of sequential discoveries

This Appendix contains the proofs for the statements in Chapter B, and additional results and simulations. It is organized as follows. Section B.1 contains the proofs of the statements in the main paper. Section B.2 and B.3 show how to perform posterior inference on the single-site accumulation curve and in the covariate-dependent extension, respectively. Section B.4 presents details of each model and shows additional simulation results. Section B.5 extends the discussion on the Copepod dataset. Finally, Section B.6 describes the singletons imputation strategy for the finnish fungal biodiversity study.

B.1 Proofs

Proof of Theorem 16. We first discuss the likelihood $\mathscr{L}(\alpha \mid X_1, \dots, X_n)$. Given the sequence of tags $X_1, \dots, X_n$, any exchangeable prediction scheme defines a random partition Π_n of the integers $\{1, \dots, n\}$ such that i and j belong to the same set in Π_n if and only if $X_i = X_j$. Let $\{C_1, \dots, C_k\}$ be a partition of n into k

groups, with $n_j = \mathrm{card}(C_j)$, $(j = 1, \dots, k)$ being the cardinality of C_j. The resulting law of the random partition Π_n in the Dirichlet process case equals $\mathbb{P}(\Pi_n = \{C_1, \dots, C_k\}) = \alpha^k/(\alpha)_n \prod_{j=1}^k (n_j - 1)!$. This is the likelihood function for α, so that $\mathscr{L}(\alpha \mid X_1, \dots, X_n) \propto \alpha^k/(\alpha)_n$, which only depends on $K_n = k$ and not on $n_1, \dots, n_k$. By letting $(\alpha)_n = \alpha \prod_{i=2}^n (\alpha + i - 1)$, we get

$$\log \mathscr{L}(\alpha \mid X_1, \dots, X_n) = (k-1)\log\alpha - \sum_{i=2}^n \log(\alpha + i - 1) + c_X, \tag{B.1}$$

where c_X is a constant not depending on α. On the other hand, the logarithm of the likelihood induced by the discovery indicators is equal to

$$\log \mathscr{L}(\alpha \mid D_1, \dots, D_n) = \log\alpha \sum_{i=2}^n D_i - \sum_{i=2}^n \log(\alpha + i - 1) + c_D, \tag{B.2}$$

with c_D being a constant not depending on α. Since $\sum_{i=2}^n D_i = k - 1$, one has that equation (B.1) and (B.2) are equal up to an additive constant. Thus, the result follows.

Proof of Proposition 18. Recall that $K_n = \sum_{i=1}^n D_i$ is non-decreasing in n. Taking the limit as $n \to \infty$, we have that $K_n \to K_\infty = \sum_{i=1}^\infty D_i$ almost surely. Then, $\mathbb{E}(K_\infty) = \sum_{i=1}^\infty S(i-1; \boldsymbol{\theta})$, as a consequence of the monotone convergence theorem. Moreover, equation (3.6) follows from the remainder estimate for the integral test for convergence of an infinite series. In particular, being $S(t; \boldsymbol{\theta})$ positive, continuous and strictly decreasing in t, we have that

$$\int_2^\infty S(t-1; \boldsymbol{\theta}) \mathrm{d}t \leqslant \sum_{i=2}^\infty S(i-1; \boldsymbol{\theta}) \leqslant \int_1^\infty S(t-1; \boldsymbol{\theta}) \mathrm{d}t.$$

Recalling that $\sum_{i=2}^\infty S(i-1; \boldsymbol{\theta}) = \mathbb{E}(K_\infty) - S(0; \boldsymbol{\theta}) = \mathbb{E}(K_\infty) - 1$ and that $\mathbb{E}(T) = \int_0^\infty S(t; \boldsymbol{\theta}) \mathrm{d}t = \int_1^\infty S(t-1; \boldsymbol{\theta}) \mathrm{d}t$, we have that

$$\mathbb{E}(T) - \int_0^1 S(t; \boldsymbol{\theta}) \mathrm{d}t + 1 \leqslant \mathbb{E}(K_\infty) \leqslant \mathbb{E}(T) + 1.$$

Finally, as $\int_0^1 S(t;\boldsymbol{\theta})\mathrm{d}t \leqslant \int_0^1 S(0;\boldsymbol{\theta})\mathrm{d}t = 1$, the result follows.

Proof of Corollary 19. We begin by proving that $K_\infty = \infty$ if and only if $\mathbb{E}(K_\infty) = \infty$. One side follows from the monotone convergence theorem: if $K_\infty = \infty$, then necessarily $\lim_{n\to\infty} \mathbb{E}(K_n) = \mathbb{E}(K_\infty) = \infty$ by the same argument in the proof of Proposition 18. The other direction can be proved by contrapposition: suppose that $K_\infty < \infty$. Then, there exists a positive constant $M < \infty$ such that $K_\infty < M$ almost surely. This means that $\mathbb{E}(K_\infty) < \mathbb{E}(M) = M < \infty$. The rest of the claim naturally follows from the inequality in equation (3.6) of Proposition 18.

Proof of Theorem 20. The first part of the theorem is a consequence of the strong law of large numbers for the sum of independent random variables. In particular, let $s_n = \int_1^n S(t-1;\boldsymbol{\theta})\mathrm{d}t$. Then, $s_n < s_{n+1}$ for every n, and $s_n \to \mathbb{E}(T) = \infty$ as $n \to \infty$. Since by assumption $S(n-1;\boldsymbol{\theta}) > S(n;\boldsymbol{\theta})$ and $s_n^2 < s_{n+1}^2$ for every n, we have that $\sum_{n=1}^\infty \mathrm{var}(D_n)/s_n^2 < \infty$, which holds by the series convergence test, because

$$\begin{aligned}\lim_{n\to\infty} \frac{\mathrm{var}(D_{n+1})}{s_{n+1}^2}\frac{s_n^2}{\mathrm{var}(D_n)} &= \lim_{n\to\infty} \frac{S(n;\boldsymbol{\theta})\{1-S(n;\boldsymbol{\theta})\}}{S(n-1;\boldsymbol{\theta})\{1-S(n-1;\boldsymbol{\theta})\}}\frac{s_n^2}{s_{n+1}^2}\\ &< \lim_{n\to\infty}\frac{1-S(n;\boldsymbol{\theta})}{1-S(n-1;\boldsymbol{\theta})} = 1.\end{aligned}$$

Hence, the above condition ensures that $\{K_n - \mathbb{E}(K_n)\}/s_n \to 0$ almost surely as $n \to \infty$ by the strong law of large numbers. This means that $\lim_{n\to\infty} K_n/s_n = \lim_{n\to\infty} \mathbb{E}(K_n)/s_n = 1$, almost surely, as a consequence of Proposition 18.

The second part of the claim follows from Lyapunov's central limit theorem. Define $\sigma_n^2 = \mathrm{var}(K_n)$ for every n. As the discovery indicators $(D_n)_{n\geqslant 1}$ are all independent, we can prove the central limit theorem for K_n by showing that there exists a $\delta > 0$ such that

$$\lim_{n\to\infty} 1/\sigma_n^{2+\delta}\sum_{i=1}^n \mathbb{E}(|D_i - \pi_i|^{2+\delta}) = 0,$$

where $\pi_n = S(n-1;\boldsymbol{\theta})$ is the discovery probability at every n. Fix $\delta = 2$. From the proofs of Corollaries 19 and 21, we have that $K_\infty = \infty$ implies that $\lim_{n\to\infty} \sigma_n^2 = \infty$. Moreover, by looking at the fourth centered moment of a Bernoulli distribution we have that

$$\sum_{i=1}^{n} \mathbb{E}(|D_i - \pi_i|^4) = \sum_{i=1}^{n} \pi_i(1-\pi_i)\{1 - 3\pi_i(1-\pi_i)\} \leqslant \sum_{i=1}^{n} \pi_i(1-\pi_i) = \sigma_n^2,$$

which leads to $0 \leqslant \lim_{n\to\infty} 1/\sigma_n^4 \sum_{i=1}^{n} \mathbb{E}(|D_i - \pi_i|^4) \leqslant \lim_{n\to\infty} 1/\sigma_n^2 = 0$, concluding the proof.

Proof of Corollary 21. To prove this claim, we rely on the limit comparison test for the ratio of two series. In particular,

$$\lim_{n\to\infty} \frac{S(n-1;\boldsymbol{\theta})\{1 - S(n-1;\boldsymbol{\theta})\}}{S(n-1;\boldsymbol{\theta})} = 1 - \lim_{n\to\infty} S(n-1;\boldsymbol{\theta}) = 1.$$

This implies that $\mathrm{var}(K_\infty) = \sum_{i=1}^{\infty} S(i-1;\boldsymbol{\theta})\{1 - S(i-1;\boldsymbol{\theta})\}$ diverges if and only if $\mathbb{E}(K_\infty) = \sum_{i=1}^{\infty} S(i-1;\boldsymbol{\theta})$ diverges. Following the same argument in the proof of Corollary 19, having $K_\infty < \infty$ almost surely implies that $\mathbb{E}(K_\infty) < \infty$, and in turn $\mathrm{var}(K_\infty) < \infty$.

Proof of Proposition 22 This can be proved by means of the series convergence test. By the fact that

$$\lim_{n\to\infty} \frac{S(n;\alpha,\sigma,\phi)}{S(n-1;\alpha,\sigma,\phi)} = \lim_{n\to\infty} \frac{\alpha\phi^n}{\alpha\phi^n + n^{1-\sigma}} \frac{\alpha\phi^{n-1} + (n-1)^{1-\sigma}}{\alpha\phi^{n-1}} = \phi,$$

having $\phi < 1$ implies that $\mathbb{E}(K_\infty) = \sum_{i=1}^{\infty} S(i-1;\alpha,\sigma,\phi) < \infty$ almost surely. But then, $K_\infty < \infty$ as well by the proof of Corollary 19.

Proof of Theorems 24 and 25 The proofs of Theorems 24 and 25 are presented together. The arguments we use follow a similar line of reasoning as in Charalambides (2005). As a first step, we prove the triangular recurrence in Theorem 24. Following Definition 23, we can write $\prod_{k=0}^{n}(\alpha + k^{1-\sigma}\phi^{-k}) = (\alpha + n^{1-\sigma}\phi^{-n}) \prod_{k=0}^{n-1}(\alpha + k^{1-\sigma}\phi^{-k})$, for any $n \geqslant 1$, from which it follows that

$$\sum_{k=0}^{n+1} \alpha^k \mathscr{C}_{n+1,k}(\sigma, \phi) = \sum_{k=1}^{n+1} \alpha^k \mathscr{C}_{n,k-1}(\sigma, \phi) + \sum_{k=0}^{n} \alpha^k n^{1-\sigma} \phi^{-n} \mathscr{C}_{n,k}(\sigma, \phi).$$

Hence, all the coefficients associated to each α^k must coincide under both sides of the above equation. This means that $\mathscr{C}_{n+1,k}(\sigma, \phi) = \mathscr{C}_{n,k-1}(\sigma, \phi) + n^{1-\sigma}\phi^{-n}\mathscr{C}_{n,k}(\sigma, \phi)$. As for the initial conditions, it is easy to check that they naturally follow from Definition 23.

To prove the second part of Theorem 24, we start by considering $\mathbb{P}(K_n = k)$. Call $j_1, \dots, j_n$ a sequence of indexes such that $D_{j_s} = 1$ for $s = 1, \dots, k$, and $D_{j_s} = 0$ for $s = k+1, \dots, n$. By independence of the indicators, the probability of such a configuration is

$$\mathbb{P}(D_{j_1} = 0, \dots, D_{j_k} = 1, D_{j_{k+1}} = 0, \dots, D_{j_n} = 0) =$$

$$= \prod_{s=1}^{k} S(j_s - 1; \alpha, \sigma, \phi) \prod_{s=k+1}^{n} \{1 - S(j_s - 1; \alpha, \sigma, \phi)\}$$

$$= \frac{\alpha^k}{\prod_{i=0}^{n-1}(\alpha + i^{1-\sigma}\phi^{-i})} \prod_{j=1}^{n-k} i_j^{1-\sigma} \phi^{-i_j},$$

where the product in the last equality follows from relabeling the indexes as $i_1 = j_{k+1} - 1, \dots, i_{n-k} = j_n - 1$. Moreover, note that $\{i_1, \dots, i_{n-k}\}$ is one of the $n-k$ possible combinations of the $n-1$ positive integers $\{1, \dots, n-1\}$ for which we obtain precisely k discoveries, with $1 \leqslant k \leqslant n$ and $n \geqslant 2$. Hence, summing over all the possible combinations of $\{i_1, \dots, i_{n-k}\}$ leads us to the probability

$$\mathbb{P}(K_n = k) = \frac{\alpha^k}{\prod_{i=0}^{n-1}(\alpha + i^{1-\sigma}\phi^{-i})} \sum_{(i_1, \dots, i_{n-k})} \prod_{j=1}^{n-k} i_j^{1-\sigma} \phi^{-i_j}, \tag{B.3}$$

for $1 \leqslant k \leqslant n$ and $n \geqslant 2$. The object in equation (B.3) is a probability mass function.

This means that

$$\sum_{k=0}^{n} \mathbb{P}(K_n = k) = \sum_{k=0}^{n} \frac{\alpha^k}{\prod_{i=0}^{n-1}(\alpha + i^{1-\sigma}\phi^{-i})} \sum_{(i_1,\ldots,i_{n-k})} \prod_{j=1}^{n-k} i_j^{1-\sigma}\phi^{-i_j} = 1,$$

recalling that $\mathbb{P}(K_n = 0) = 0$. Rearranging the equality, one has that

$$\prod_{i=0}^{n-1}(\alpha + i^{1-\sigma}\phi^{-i}) = \sum_{k=0}^{n} \alpha^k \sum_{(i_1,\ldots,i_{n-k})} \prod_{j=1}^{n-k} i_j^{1-\sigma}\phi^{-i_j},$$

which is the same polynomial expansion proposed in Definition 23. Hence, it must be that

$$\mathscr{C}_{n,k}(\sigma, \phi) = \sum_{(i_1,\ldots,i_{n-k})} \prod_{j=1}^{n-k} i_j^{1-\sigma}\phi^{-i_j}, \tag{B.4}$$

again for $1 \leqslant k \leqslant n$ and $n \geqslant 2$. This last equality proves the second part of Theorem 24. Finally, Theorem 25 naturally follows by plugging equation (B.4) into (B.3).

Proof of Proposition 26 To find the maximizer $\hat{\boldsymbol{\theta}} = (\hat{\alpha}, \hat{\sigma}, \hat{\phi})$ of the likelihood in equation (3.5) with the the three-parameter log-logistic specification we rely on the first order condition with respect to α. In particular, the logarithm of the likelihood becomes

$$\log \mathscr{L}(\alpha, \sigma, \phi \mid D_1, \ldots D_n) = \log \alpha \sum_{i=2}^{n} D_i - \sum_{i=2}^{n} \log\{\alpha\phi^{i-1} + (i-1)^{1-\sigma}\} + c_{\sigma,\phi},$$

where $c_{\sigma,\phi}$ is a constant not dependent on α. Hence, the first order condition with respect to α leads to

$$\sum_{i=1}^{n} \frac{\alpha\phi^{i-1}}{\alpha\phi^{i-1} + (i-1)^{1-\sigma}} = \sum_{i=1}^{n} D_i = k = \mathbb{E}(K_n).$$

This equality must be maintained at the solution $\hat{\boldsymbol{\theta}}$.

B.2 Posterior inference for a single accumulation curve

In the following we describe the estimation of the parameters $\theta = (\alpha, \sigma, \phi)$ under the three-parameter log-logistic specification and using Markov Chain Monte Carlo. Let $(D_n)_{n \geqslant 1}$ be a sequence of discovery indicators with $D_1 = 1$ and

$$\pi_{n+1} = \mathbb{P}(D_{n+1} = 1 \mid D_1, \dots, D_n) = \frac{\alpha\phi^n}{\alpha\phi^n + n^{1-\sigma}}, \qquad n \geqslant 1,$$

for $\alpha > 0$, $\sigma < 1$, $0 < \phi \leqslant 1$ and $\pi_1 = 1$. As discussed in the manuscript, this implies that

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

with $\beta_0 = \log \alpha$, $\beta_1 = \sigma - 1 < 0$ and $\beta_3 = \log \phi \leqslant 0$. These constraints are imposed through a truncated normal prior, namely $\boldsymbol{\beta} \sim N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})\mathbb{1}(\beta_1 < 0; \beta_2 \leqslant 0)$.

Samples from the posterior can be easily obtained via the Pólya-gamma data-augmentation strategy introduced in Polson et al. (2013). This procedure introduces Pólya-gamma distributed positive latent variables $\boldsymbol{\omega} = (\omega_2, \dots, \omega_n)^{\mathrm{T}}$. The resulting full conditional distributions for $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ are available in closed form. Let $\mathbf{d} = (d_2, \dots, d_n)^{\mathrm{T}}$ be the observed values for the discovery indicators $D_2, \dots, D_n$ and let $\mathbf{V}$ be the design matrix, with $n-1$ rows and 3 columns and entries $\mathbf{v}_i = (1, \log i, i)^{\mathrm{T}}$, for $i = 1, \dots, n-1$. Then, the full conditional for $(\boldsymbol{\beta} \mid \boldsymbol{\omega}, \mathbf{d})$ is a multivariate truncated normal distribution with parameters $\boldsymbol{\mu}_{\boldsymbol{\omega}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\omega}}$ equal to

$$\boldsymbol{\Sigma}_{\boldsymbol{\omega}} = (\mathbf{V}^{\mathrm{T}}\boldsymbol{\Omega}\mathbf{V} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^{-1}, \quad \boldsymbol{\mu}_{\boldsymbol{\omega}} = \boldsymbol{\Sigma}_{\boldsymbol{\omega}}(\mathbf{V}^{\mathrm{T}}\boldsymbol{\kappa} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}), \tag{B.5}$$

with $\boldsymbol{\kappa} = (d_2 - 1/2, \dots, d_n - 1/2)^{\mathrm{T}}$ and $\boldsymbol{\Omega} = \operatorname{diag}(\omega_2, \dots, \omega_n)$. The algorithm below outlines the sampling procedure.

In our work we obtain samples from the multivariate truncated normal through the efficient algorithm proposed in Botev (2017).

Algorithm 3: Pólya-Gamma Gibbs sampler for single site accumulation curve

```
Set an initial value $\boldsymbol{\beta}$ and set number of samples $R$;
for $r = 1$ to $r = R$ do
    for $i = 1$ to $i = n - 1$ do
        Sample $(\omega_i \mid \boldsymbol{\beta}) \sim \mathrm{PolyaGamma}(1, \mathbf{v}_i^{\mathrm{T}}\boldsymbol{\beta})$;
    end
    Sample $(\boldsymbol{\beta} \mid \boldsymbol{\omega}, \boldsymbol{d}) \sim N_3(\boldsymbol{\mu}_{\boldsymbol{\omega}}, \boldsymbol{\Sigma}_{\boldsymbol{\omega}})\mathbb{1}(\beta_1 < 0, \beta_2 \leqslant 0)$, with $\boldsymbol{\mu}_{\boldsymbol{\omega}}$, $\boldsymbol{\Sigma}_{\boldsymbol{\omega}}$ in (B.5);
end
Output: collection of $R$ samples for $\boldsymbol{\beta}$
```

B.3 Posterior sampling for multi-site data

We now describe a Markov Chain Monte Carlo algorithm for Bayesian inference for the covariate-dependent model described in the manuscript. Recall that we are given a collection of L accumulation curves $(K_{1n})_{n\geqslant 1}, \dots, (K_{Ln})_{n\geqslant 1}$ observed up to the terms $n_1, \dots, n_L$. Each curve is associated to a set of covariates $\mathbf{z}_\ell = (z_{\ell 1}, \dots, z_{\ell p})^{\mathrm{T}}$ for $\ell = 1, \dots, L$. Future observations correspond to new discoveries within the set of the considered L curves, so that new covariates values are not expected.

Let $(D_{\ell n})_{n\geqslant 1}$ be the sequence of discovery indicators for the ℓth location, with probabilities $(\pi_{\ell n})_{n\geqslant 1}$. Hence, we get

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

with $\boldsymbol{\gamma}_0, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2 \in \mathbb{R}^p$ coefficient vectors such that $\mathbf{z}_\ell^{\mathrm{T}}\boldsymbol{\gamma}_2 < 0$ and $\mathbf{z}_\ell^{\mathrm{T}}\boldsymbol{\gamma}_2 \leqslant 0$ for every $\ell = 1, \dots, L$. The above specification is a logistic regression and therefore inference on the parameters $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_0, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)^{\mathrm{T}}$ may be conducted through a simple modification of Algorithm 3.

Let $N = \sum_{\ell=1}^{L}(n_\ell - 1)$ and let $\mathbf{V}$ be a design matrix with N rows and $3p$ columns, with rows $\mathbf{v}_{(\ell i)} = (\mathbf{z}_\ell^{\mathrm{T}}, \mathbf{z}_\ell^{\mathrm{T}}\log i, \mathbf{z}_\ell^{\mathrm{T}} i)^{\mathrm{T}}$ for $i = 1, \dots, n_\ell - 1$ and $\ell = 1, \dots, L$. Moreover, call $\mathbf{d} = (\mathbf{d}_1^{\mathrm{T}}, \dots, \mathbf{d}_L^{\mathrm{T}})^{\mathrm{T}}$ the realized discoveries, with $\mathbf{d}_\ell = (d_{\ell 2}, \dots, d_{\ell n_\ell})^{\mathrm{T}}$ be the observed values for $D_{\ell 1}, \dots, D_{\ell n}$, for every $\ell = 1, \dots, L$. As before, we can incorporate the constraints $\mathbf{z}_\ell^{\mathrm{T}}\boldsymbol{\gamma}_1 < 0$ and $\mathbf{z}_\ell^{\mathrm{T}}\boldsymbol{\gamma}_2 \leqslant 0$ by assigning $\boldsymbol{\gamma}$ a multivariate truncated normal prior,

$$\boldsymbol{\gamma} \sim N_{3p}(\boldsymbol{\mu}, \boldsymbol{\Sigma})\mathbb{1}(\mathbf{z}_\ell^{\mathrm{T}}\boldsymbol{\gamma}_1 < 0; \mathbf{z}_\ell^{\mathrm{T}}\boldsymbol{\gamma}_2 \leqslant 0; \ell = 1, \dots, L).$$

Let $\boldsymbol{\omega}$ be a N-dimensional vector of Pólya-gamma latent variables. Then, the full conditional for $(\boldsymbol{\gamma} \mid \boldsymbol{\omega}, \mathbf{d})$ is a multivariate truncated normal distribution with mean $\boldsymbol{\mu}_{\boldsymbol{\omega}}$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\omega}}$ equal to equation (B.5), whereas $\boldsymbol{\Omega}$ is a diagonal matrix whose diagonal elements are those of the vector $\boldsymbol{\omega}$.

In Algorithm 4 we employ a vanilla acceptance rejection sampler for the full conditional $(\boldsymbol{\gamma} \mid \boldsymbol{\omega}, \mathbf{d})$. This is indeed a reasonable approach in most practical settings, as the data usually support the required constraints, leading to very high acceptance rates. If needed, suitable adaptations of the ideas of Botev (2017) may be alternatively considered.

Algorithm 4: Pólya-Gamma Gibbs sampler for covariate-dependent accumulation curves

```
Set an initial value $\boldsymbol{\gamma}$ and set the number of samples $R$;
for $r = 1$ to $r = R$ do
    for $\ell = 1$ to $\ell = L$ do
        Sample $(\omega_{(\ell i)} \mid \boldsymbol{\gamma}) \sim \mathrm{PolyaGamma}(1, \mathbf{v}_{(\ell i)}^{\mathrm{T}}\boldsymbol{\gamma}), \quad i = 1, \dots, n_\ell - 1$;
    end
    Sample $(\boldsymbol{\gamma} \mid \boldsymbol{\omega}, \mathbf{d}) \sim N_{3p}(\boldsymbol{\mu}_{\boldsymbol{\omega}}, \boldsymbol{\Sigma}_{\boldsymbol{\omega}})$, with $\boldsymbol{\mu}_{\boldsymbol{\omega}}$, $\boldsymbol{\Sigma}_{\boldsymbol{\omega}}$ as in (B.5) until $\boldsymbol{\gamma}$
      satisfies $\mathbf{z}_\ell^{\mathrm{T}}\boldsymbol{\gamma}_1 < 0$ and $\mathbf{z}_\ell^{\mathrm{T}}\boldsymbol{\gamma}_2 \leqslant 0$ for every $\ell = 1, \dots, L$;
end
Output: collection of $R$ samples for $\boldsymbol{\gamma}$
```

B.4 Simulation

In this Section, we extend the simulation studies performed in Section 5 in the main paper. Our purpose is to evaluate the performance of the models under different scenarios, with varying hyperparameters. Details of the sampling distributions are provided in each setting. For aesthetic reasons, all plots and tables are reported at the end of the section.

B.4.1 Dirichlet process

The Dirichlet process (Ferguson, 1973) is the first example of species sampling model we consider. Given an exchangeable sequence $(X_n)_{n\geqslant 1}$, let $X_1^*, \dots, X_k^*$ denote the labels of the $K_n = k$ distinct species appeared up to n, with frequencies $n_1, \dots, n_k$ and $\sum_{j=1}^k n_j = n$. As described in Blackwell and MacQueen (1973), the species of the $n+1$ individual is determined by the allocation scheme

$$X_{n+1} \mid X_1, \dots, X_n = \begin{cases} \text{“new”}, & \text{with probability} \quad \alpha/(\alpha+n) \\ X_j^*, & \text{with probability} \quad n_j/(\alpha+n), \quad (j = 1, \dots, k) \end{cases} \tag{B.6}$$

with $\alpha > 0$. As α increases, the probability of observing new species increases; the higher α, the higher K_n is for given n, with $K_n \sim \alpha \log n$ asymptotically. Table B.1 reports the performance of the different models in terms of average mean square error when the data are simulated from a Dirichlet process. Each simulation is repeated 500 times in each scenario in the same way as described in the main paper. Generally, the three-parameter log-logistic performs well in sample, while the true model achieves the best performance in the test set. Figure B.1 depicts one example of a simulated accumulation curve for each value chosen for α. The performance of each model is roughly comparable both in- and out-of sample, with the exception of the BETA-GOS-1. The reason is that almost all models admit the Dirichlet process as a special case.

B.4.2 Pitman–Yor process

The Pitman-Yor process (Perman et al., 1992) generalizes the Dirichlet process in equation (B.6) by the inclusion of an additional parameter, σ, which controls the tail behavior of the allocation scheme. For an exchangeable sequence $(X_n)_{n\geqslant 1}$, given $K_n = k$ distinct species detected at n, labelled as $X_1^*, \ldots, X_k^*$ with frequencies $n_1, \ldots, n_k$, the species of the $n+1$ individual follows

$$X_{n+1} \mid X_1, \ldots, X_n = \begin{cases} \text{“new”}, & \text{with probability} \quad (\alpha + \sigma k)/(\alpha + n) \\ X_j^*, & \text{with probability} \quad (n_j - \sigma)/(\alpha + n), \quad (j = 1, \ldots, k), \end{cases} \tag{B.7}$$

with $\alpha > -\sigma$ and $\sigma \in [0, 1)$. When $\sigma = 0$, Equation (B.7) reduces to (B.6), while, when σ increases, the probability of detecting new species increases as well. This implies a steeper accumulation curve, as depicted in Figure B.2. The Pitman–Yor admits a closed form law for the random partition $\Pi_n = \{C_1, \ldots, C_k\}$ with $n_j = \text{card}(C_j)$ generated by equation (B.7). This is known as an exchangeable partition probability function, and for $K_n = k$ is equal to

$$\mathbb{P}(\Pi_n = \{C_1, \ldots, C_k\}) = \frac{\prod_{i=1}^{k-1}(\alpha + i\sigma)}{(\alpha + 1)_{n-1}} \prod_{i=1}^{k} (1 - \sigma)_{n_i - 1}. \tag{B.8}$$

Equation (B.8) can be used as a likelihood when estimating the parameters α and σ. In this section and in Section 5 of the man paper, this is done by empirical Bayes through standard maximization routines. Table B.2 reports the average mean square error of the models across 500 accumulation curves simulated from the Pitman–Yor process with varying values for σ. The results further confirm the similarities between the Pitman–Yor process and the BETA-GOS-2 in terms of accumulation curves. This similarity, however, does not hold when considering clustering, as detailed in Airoldi et al. (2014).

B.4.3 Dirichlet-multinomial

The Dirichlet-multinomial process is a special case of the urn scheme in equation (B.7) with parameters $\sigma < 0$ and $\alpha = -\sigma H$, where H is an integer representing the total number of distinct species tags in the population. In this case, estimation can proceed by selecting a grid of values for H, say $H = k, \dots, k + h_m$ with $h_m \in \mathbb{N}$, and then estimating σ by maximizing equation (B.8) conditional on each H. The optimal value for (H, σ) is the pair that has the highest likelihood. Notice that σ controls the steepness with which the accumulation curve reaches H, as depicted in Figure B.3. The majority of the models, with the exception of the Pitman–Yor and the Dirichlet process, perform well in that they correctly guess the value of H. Table B.3 confirms this behavior. Not surprisingly, the best out-of-sample performance is achieved by the Dirichet-multinomial, while the three-parameter log-logistic performs well in-sample. Good performances are also obtained by the BETA-GOS-2.

B.4.4 BETA-GOS

The BETA-GOS process (Airoldi et al., 2014) is a generalization of the urn schemes in equation (B.7) that relaxes the exchangeability assumption. In particular, given the sequence of tags $X_1, \dots, X_n$, the species of the $n+1$ individual follows

$$X_{n+1} \mid X_1, \dots, X_n = \begin{cases} \text{“new”}, & \text{with probability } \prod_{i=1}^{n} W_i, \\ X_i, & \text{with probability } (1 - W_i) \prod_{j=i+1}^{n} W_j, \ (i = 1, \dots, n), \end{cases} \tag{B.9}$$

with $W_i \sim \text{BETA}(a_i, b_i)$ for each $i = 1, \dots, n$. Hence, the allocation probabilities are random instead of fixed. The freedom in choosing a_i and b_i allow for a flexible model. We consider the case when $a_i = \theta + i - 1$ with $\theta > 0$ and $b_i = \beta > 0$. This is a slightly broader scenario than the one in Airoldi et al. (2014), where instead $\beta \geqslant 1$. We adopt this choice to mirror the same behavior as the one in our two-parameter log-logistic distribution, thus making the BETA-GOS a valid competitor. Indeed, it is possible to prove that when $\beta > 1$ then $\mathbb{E}(K_n) < \infty$, while if $\beta \leqslant 1$ we have that $\mathbb{E}(K_n) = \infty$. Estimation of the parameters β and θ can be carried via method of moments as described in the main paper. Figure B.4 shows the shape of the accumulation curve generated from the BETA-GOS-2 with varying values from β and $\theta = 500$. Notice that the curves are diverging for values of $\beta \leqslant 1$. Overall, the best performance when $\beta = 0.25$ and $\beta = 0.75$ is achieved by the BETA-GOS-2. In the other two cases, the log-logistic model produces a good fit as well. Table B.4 further confirms this fact.

B.4.5 Finite Geometric

In the following subsections, we replicate the analysis described so far on accumulation curves that are generated by taking independent and identically distributed samples from discrete distributions. Unlike previous scenarios, in this case none of the competing model is the data generating process. In particular, let $Y \sim p(y)$ denote a discrete random variable, and let $Y_1, \dots, Y_n \overset{iid}{\sim} p(y)$. Then, we construct the species sequence $(X_i)_{i=1}^n$ by setting $X_i = y_i$ for each $i = 1, \dots, n$, where y_i is the realized value of Y_i. As these draws are discrete, the sequence $(X_i)_{i=1}^n$ will show ties among the associated labels. This allows for the construction of a valid accumulation curve.

The first distribution we consider is the finite geometric, where

$$p(y) \propto \eta^y, \quad y = 1, \dots, H,$$

with $\eta \in (0, 1)$. Here, the truncation point H and can be interpreted as the species richness, while the parameter η governs the trajectory of the generated curve. The closer η is to 1, the higher the probability of observing values for y close to H is. See Figure B.5 for an example. Table B.5 reports the performances of the models when tested on curves generated with varying values for η, fixing $H = 100$ for simplicity. Generally, all models perform well, with relatively small differences in terms of average mean square error.

B.4.6 Finite Zipf

The second example we consider is when Y follows a Zipf distribution with finite support, for which

$$p(y) \propto y^{-\eta}, \quad y = 1, \dots, H$$

with $\eta > 0$ and H being the species richness. As seen in Figure B.6, smaller values for η make the curve quickly reach the asymptote H. Model performances are reported in Table B.6. Interestingly, the Dirichlet-multinomial model becomes unstable when η is low. In this simulation setting, our three-parameter log-logistic performs particularly well.

B.4.7 Geometric

We consider now the cases where Y is a distribution with unbounded domain. In this case, the number of species observable is infinite. When Y follows a geometric distribution, the probability mass function is

$$p(y) = (1 - \eta)^{y-1}\eta, \quad y = 1, 2, \dots$$

with $\eta \in (0, 1)$ regulating the shape of trajectory. The closer η is to 0, the higher the number of distinct species appearing is. On the other hand, if η is close to 1, the number of distinct species is generally low. See Figure B.7. Table B.7 summarises model performances. We notice that the three-parameter log-logistic performs well in-sample, while the best performances out-of-sample are retained by Dirichlet and Pitman–Yor.

B.4.8 Zipf

As a last case, we consider curves generated from a Zipf distribution with unbounded domain, where

$$p(y) = \frac{1}{y^{\eta}\zeta(\eta)}, \quad y = 1, 2, \dots$$

where $\eta > 1$ and $\zeta(\eta) = \sum_{y=1}^{\infty} y^{-\eta}$ is the Riemann zeta function. To sample from the distribution, refer to Devroye (1986b), Chapter 10.6.1. As shown in Figure B.8, the closer η is to 1, the steeper the associated accumulation curve is. Model performances are reported in Table B.8. The BETA-GOS-2 and the two- and three-parameter log-logistic all perform well. The Pitman–Yor shows some instability in-sample as η increases, but predicts well out-of-sample. Finally, the Dirichlet, the Dirichlet-multinomial and the BETA-GOS-1 lack flexibility.

Table B.1: Performance for curves simulated via independent samples from the Dirichlet process. 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.

	DIRICHLET PROCESS							
	$\alpha = 5$		$\alpha = 10$		$\alpha = 100$		$\alpha = 500$	
MODEL	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
DP (LL1)	3.6	3.6	6.5	7.6	47.9	72.9	161.4	365.9
PY	3.1	4.2	5.5	8.8	41.9	84.6	144.7	481.8
DM	3.0	4.2	5.5	9.4	45.0	90.3	211.9	807.6
BG-1(a, b)	47.7	16.5	166.7	58.7	8,388.3	5,176.4	49,896.8	118,940.6
BG-2(θ, β)	2.4	7.1	4.4	14.7	29.6	157.8	82.9	1191.0
LL2	2.1	4.9	3.8	11.3	26.8	108.1	78.5	706.4
LL3	1.5	7.2	3.0	16.8	21.1	279.3	61.7	2283.7

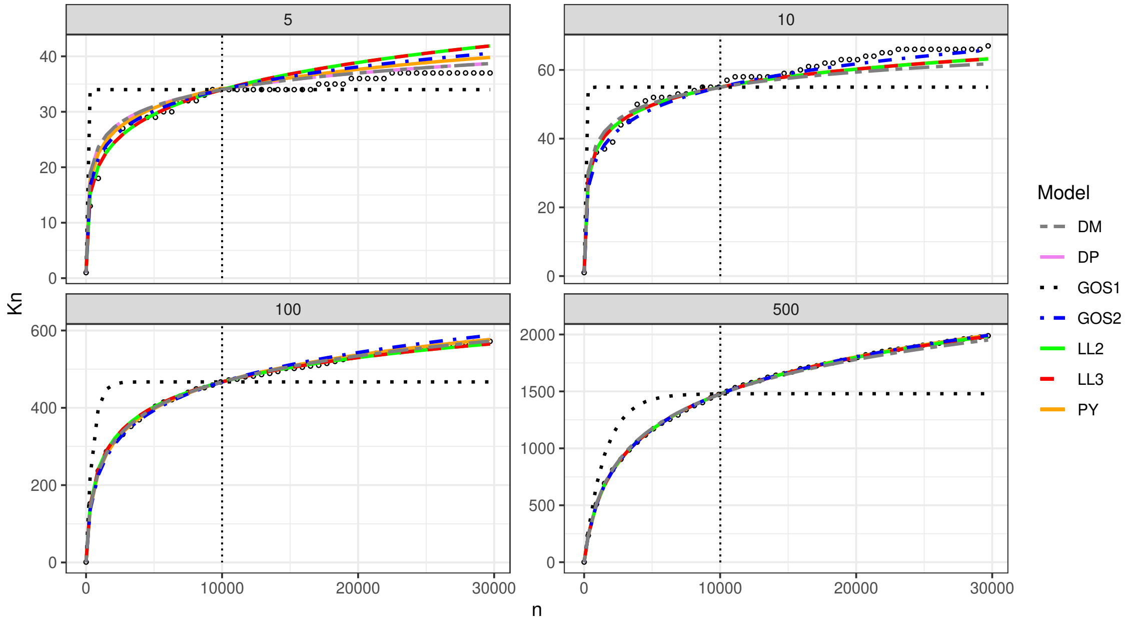

FIGURE B.1: In- and out-of-sample predictions for the different models on four randomly simulated curves from the Dirichlet process with varying α. White dots indicate the true values. Vertical line is the train-test cutoff. Curve simulation and model estimation proceeded as in Table B.1

Table B.2: Performance for curves simulated via independent samples from the Pitman–Yor process. See Table B.1 for further details

	PITMAN–YOR PROCESS							
	$\alpha = 10, \sigma = 0.1$		$\alpha = 10, \sigma = 0.3$		$\alpha = 10, \sigma = 0.7$		$\alpha = 10, \sigma = 0.9$	
MODEL	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
DP (LL1)	22.9	31.8	552.8	1279.8	5.9×10^4	6.6×10^5	1.4×10^5	7.3×10^6
PY	7.9	21.8	28.2	123.6	378.2	2300.9	2002.9	6404.4
DM	22.9	32.9	543.1	1312.4	4.6×10^4	7.7×10^5	9.6×10^4	1.4×10^7
BG-1(a, b)	510.1	242.4	4784.5	4367.2	2.4×10^5	1.6×10^6	2.4×10^5	1.6×10^7
BG-2(θ, β)	7.2	31.7	19.6	193.5	135.9	4599.1	247.4	14983.9
LL2	6.3	23.0	20.8	149.2	229.9	10775.0	406.4	56110.5
LL3	5.1	39.5	19.0	248.4	227.4	12461.9	400.0	68663.7

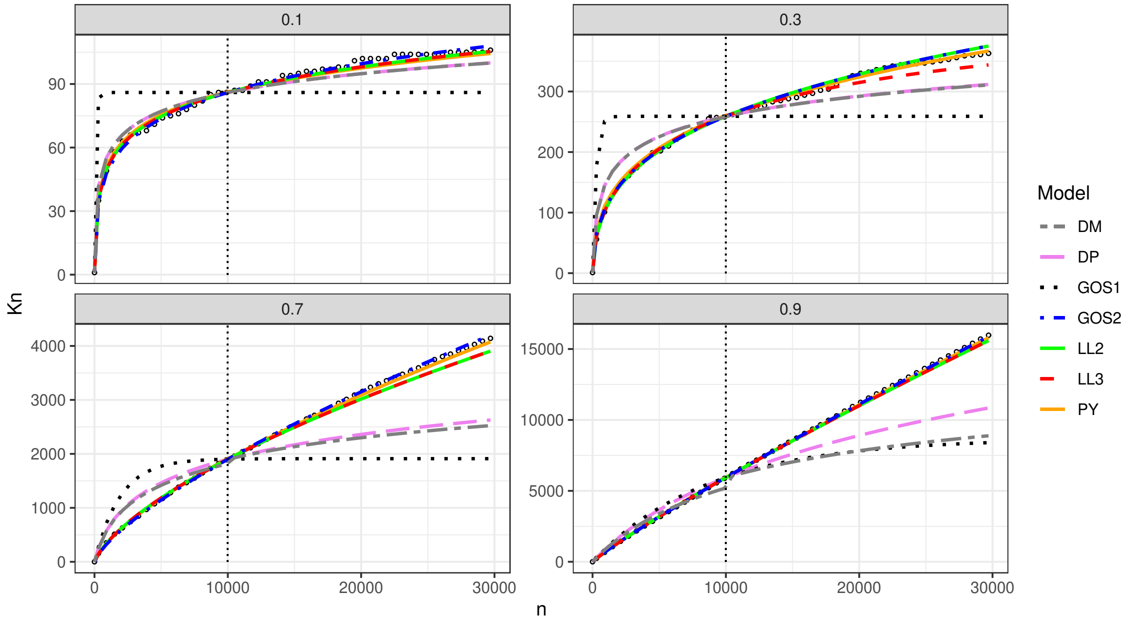

FIGURE B.2: In- and out-of-sample predictions for the different models on four randomly simulated curves from the Pitman–Yor process with varying σ and $\alpha = 10$. White dots indicate the true values. Vertical line is the train-test cutoff. Curve simulation and model estimation proceeded as in Table B.2

Table B.3: Performance for curves simulated via independent samples from the Dirichlet-multinomial. See Table B.1 for further details

	DIRICHLET-MULTINOMIAL							
	$H = 1000, \sigma = -0.1$		$H = 1000, \sigma = -0.5$		$H = 1000, \sigma = -1.5$		$H = 1000, \sigma = -3$	
MODEL	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
DP (LL1)	97.1	168.0	2409.9	6240.5	10633.5	24150.6	17780.4	33240.5
PY	97.0	168.4	2409.9	6239.0	10633.5	24146.9	17780.4	33235.9
DM	25.9	65.4	51.1	82.9	48.0	35.0	40.0	9.8
BG-1(a, b)	4390.6	1862.8	9466.2	3880.3	4489.4	776.4	1699.0	95.0
BG-2(θ, β)	22.2	110.3	38.6	161.2	34.9	65.3	30.9	17.7
LL2	20.8	79.7	78.0	442.0	300.7	1240.3	528.0	1235.1
LL3	14.2	178.9	23.4	462.8	26.8	187.3	22.5	33.3

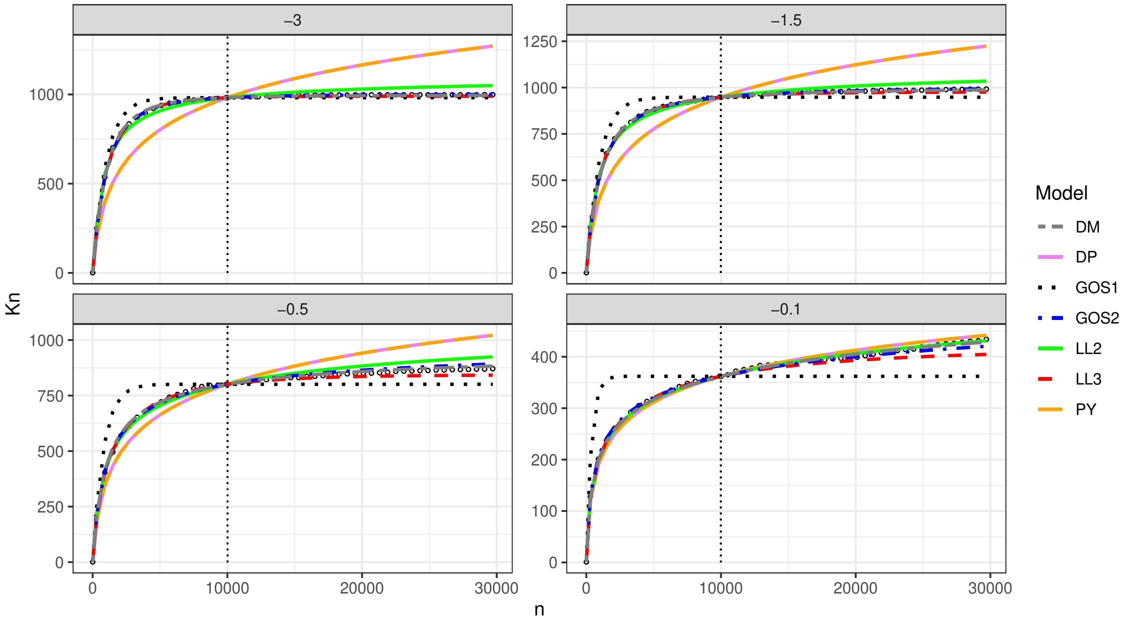

FIGURE B.3: In- and out-of-sample predictions for the different models on four randomly simulated curves from the Dirichlet-multinomial with varying σ and $H = 1000$. White dots indicate the true values. Vertical line is the train-test cutoff. Curve simulation and model estimation proceeded as in Table B.3

Table B.4: Performance for curves simulated via independent samples from the BETA-GOS-2 process. See Table B.1 for further details

	BETA-GOS							
	$\theta = 500, \beta = 0.25$		$\theta = 500, \beta = 0.75$		$\theta = 500, \beta = 1$		$\theta = 500, \beta = 1.5$	
MODEL	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
DP (LL1)	37877.0	2.9×10^6	3673.8	42109.9	162.6	426.3	2492.5	6312.5
PY	20823.0	1.4×10^6	1581.5	18097.5	148.1	529.7	2492.5	6311.1
DM	29561.1	5.8×10^6	4696.9	69629.6	200.8	921.8	1264.9	2927.3
BG-1(a, b)	87671.7	9.5×10^6	91985.8	640285.0	49465.2	119401.4	9341.8	3882.7
BG-2(θ, β)	220.7	17177.4	132.6	3384.9	92.6	1263.8	38.4	159.0
LL2	571.7	177321.3	165.9	4827.2	85.9	765.5	79.8	438.9
LL3	571.5	177631.4	162.6	6358.2	68.3	2636.6	22.8	458.3

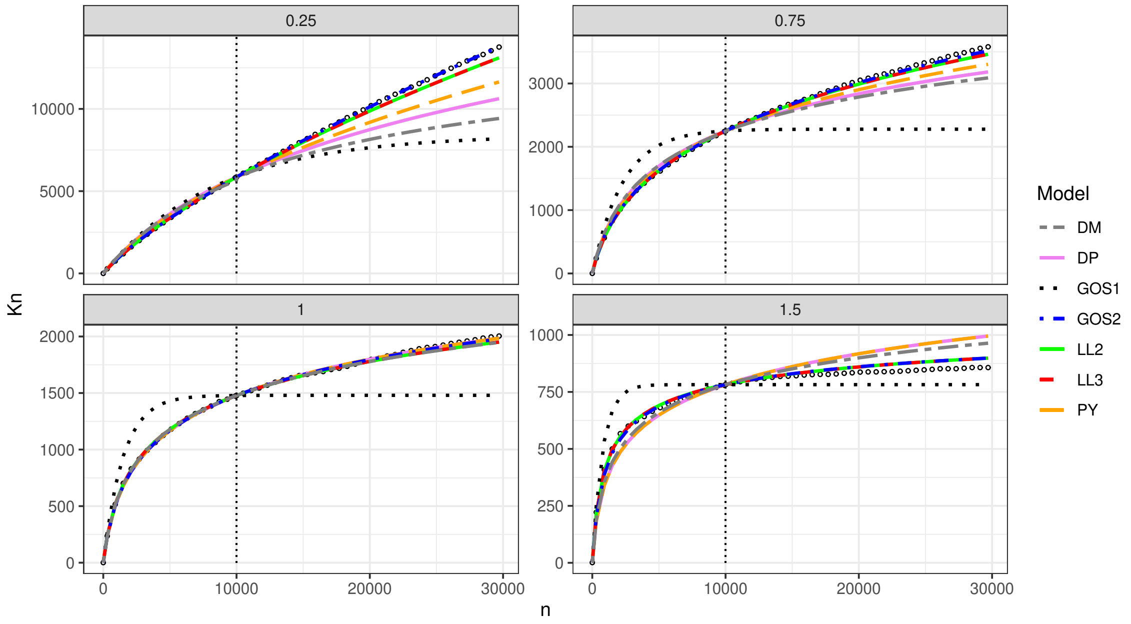

FIGURE B.4: In- and out-of-sample predictions for the different models on four randomly simulated curves from the BETA-GOS with $\theta = 500$ and varying β. White dots indicate the true values. Vertical line is the train-test cutoff. Curve simulation and model estimation proceeded as in Table B.4

Table B.5: Performance for curves simulated via independent samples from a finite Geometric distribution. See Table B.1 for further details

	FINITE GEOMETRIC							
	$\eta = 0.3, H = 100$		$\eta = 0.5, H = 100$		$\eta = 0.75, H = 100$		$\eta = 0.9, H = 100$	
MODEL	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
DP (LL1)	0.5	0.4	0.8	0.8	1.9	2.1	5.9	6.9
PY	0.5	0.4	0.7	0.8	1.8	2.1	5.9	6.9
DM	0.5	0.6	0.7	1.0	1.6	2.8	4.3	8.6
BG-1(a, b)	1.9	0.7	4.8	1.7	23.9	7.4	147.6	45.5
BG-2(θ, β)	0.5	1.1	0.8	1.8	1.7	4.9	4.2	13.5
LL2	0.3	0.6	0.5	1.1	1.3	2.9	3.4	8.5
LL3	0.2	0.7	0.4	1.3	1.1	3.8	3.1	12.5

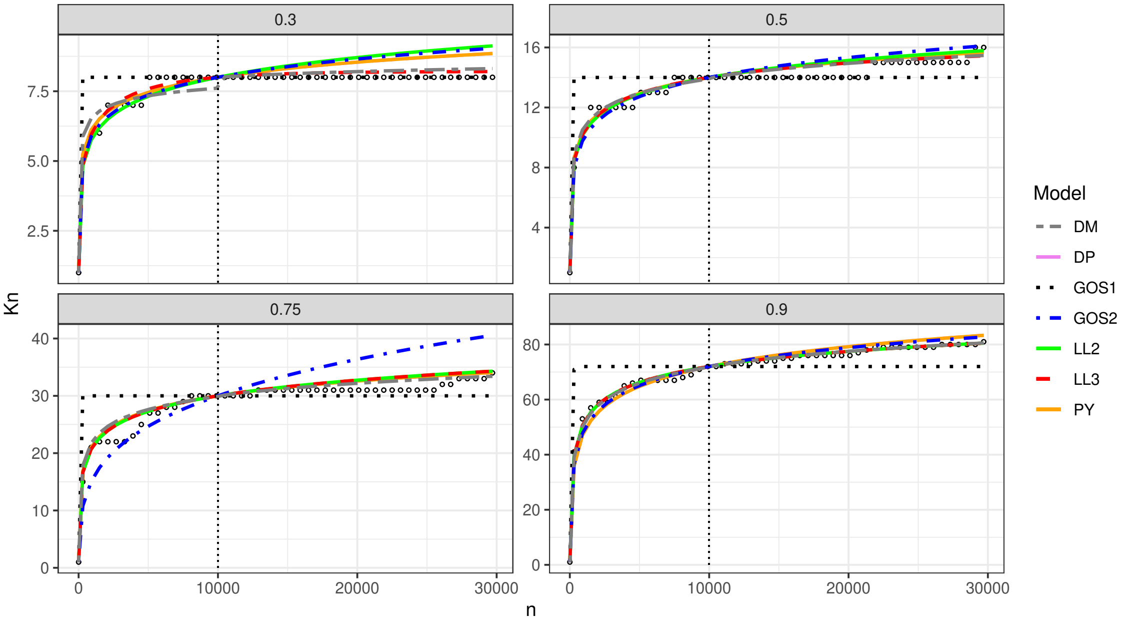

FIGURE B.5: In- and out-of-sample predictions for the different models on four randomly simulated curves from the Finite Geometric for varying η and $H = 100$. White dots indicate the true values. Vertical line is the train-test cutoff. Curve simulation and model estimation proceeded as in Table B.5

Table B.6: Performance for curves simulated via independent samples from a Zipf with finite support. See Table B.1 for further details.

	FINITE ZIPF							
	$\eta = 0.1, H = 3000$		$\eta = 0.5, H = 3000$		$\eta = 1, H = 3000$		$\eta = 1.5, H = 3000$	
MODEL	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
DP (LL1)	5.4	638157.2	16557.5	372939.0	18691.3	17251.5	9045.9	34708.9
PY	54380.8	638070.3	16557.5	372877.9	671.0	84226.4	414.9	1033.0
DM	.	.	656.2	37998.4	16059.1	31774.8	8457.8	35998.8
BG-2(a, b)	182.0	142.6	12946.8	21034.5	136767.4	319327.3	38438.3	72275.6
BG-2(θ, β)	147.5	569.4	432.1	29965.3	610.5	92544.8	49.8	1443.2
LL2	2389.0	80017.1	1767.5	117674.8	627.0	91648.5	42.0	901.1
LL3	67.2	652.6	71.6	1470.8	67.9	7858.0	31.1	1799.6

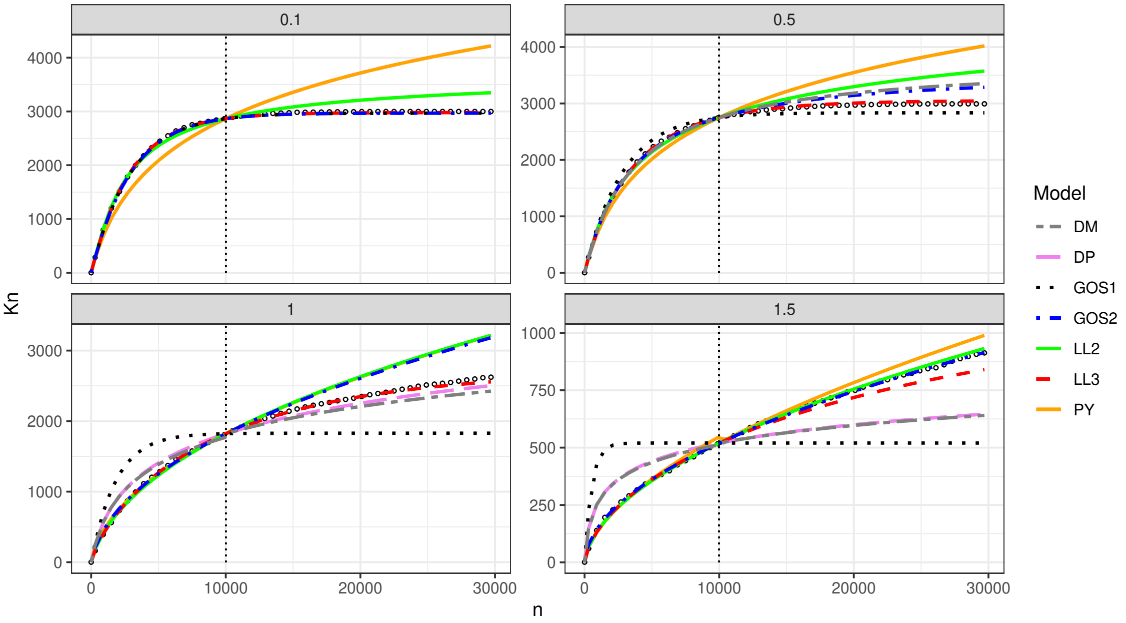

FIGURE B.6: In- and out-of-sample predictions for the different models on four randomly simulated curves from the Finite Zipf for varying η and $H = 3000$. White dots indicate the true values. Vertical line is the train-test cutoff. Curve simulation and model estimation proceeded as in Table B.6

Table B.7: Performance for curves simulated via independent samples from the Geometric distribution. See Table B.1 for further details

	GEOMETRIC							
	$\eta = 0.001$		$\eta = 0.01$		$\eta = 0.25$		$\eta = 0.5$	
MODEL	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
DP (LL1)	5445.5	31611.5	221.4	181.6	1.7	2.0	0.8	0.9
PY	5445.5	31592.5	221.4	181.5	1.7	2.0	0.7	0.9
DM	289.8	19764.1	56.8	267.1	1.5	2.8	0.7	1.2
BG-1(a, b)	27689.4	3.9×10^5	7798.4	5127.3	23.1	8.0	4.8	1.9
BG-2(θ, β)	233.8	21554.7	43.1	199.5	1.6	4.3	0.8	2.2
LL2	140.1	2737.7	37.5	195.2	1.2	2.7	0.5	1.2
LL3	98.3	16803.9	36.8	248.2	1.0	3.9	0.4	1.5

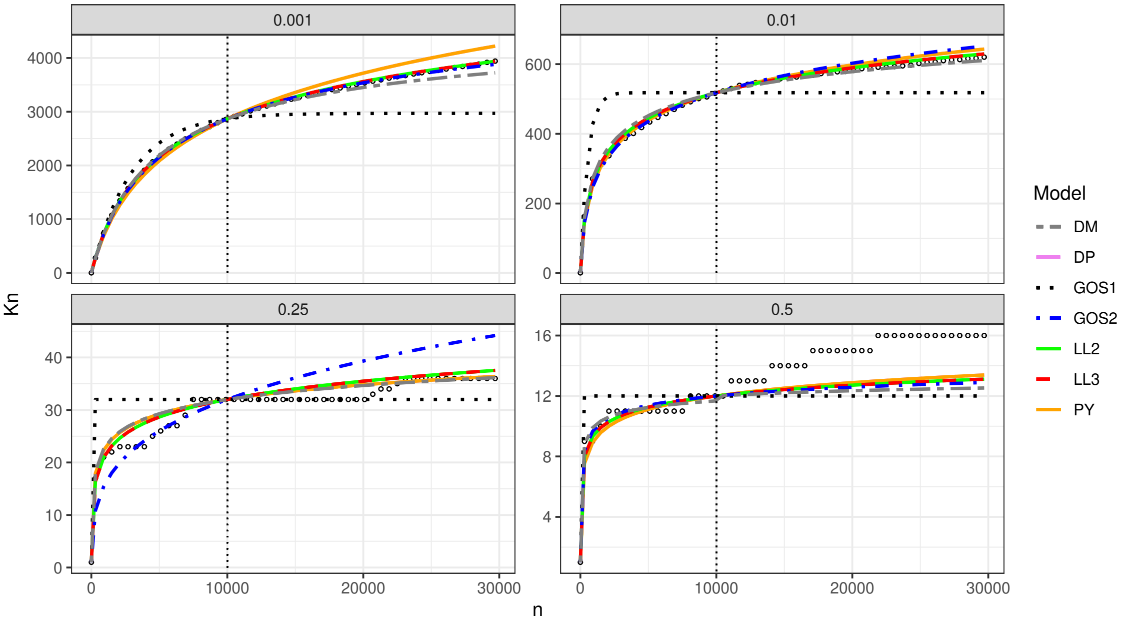

Figure B.7: In- and out-of-sample predictions for the different models on four randomly simulated curves from the geometric distribution. White dots indicate the true values. Vertical line is the train-test cutoff. Curve simulation and model estimation proceeded as in Table B.7

Table B.8: Performance for curves simulated via independent samples from the Zipf distribution. See Table B.1 for further details

	ZIPF							
	$\eta = 1.25$		$\eta = 1.75$		$\eta = 2.5$		$\eta = 3$	
MODEL	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
DP (LL1)	1.1×10^5	1.5×10^6	3132.3	11702.0	95.6	184.0	21.8	31.8
PY	3945.7	3178.0	942.8	243.3	116.7	29.2	47.3	11.6
DM	78659.6	1.8×10^6	3022.7	11878.1	95.0	184.6	21.7	31.9
BG-1(a, b)	3.2×10^5	3.2×10^6	11842.0	20908.0	347.3	341.3	78.4	60.9
BG-2(θ, β)	165.0	6673.2	24.2	421.7	4.7	53.4	2.2	20.0
LL2	201.1	10458.4	24.1	290.5	4.2	32.2	2.0	12.1
LL3	194.4	17044.0	21.4	713.4	3.3	65.2	1.5	20.3

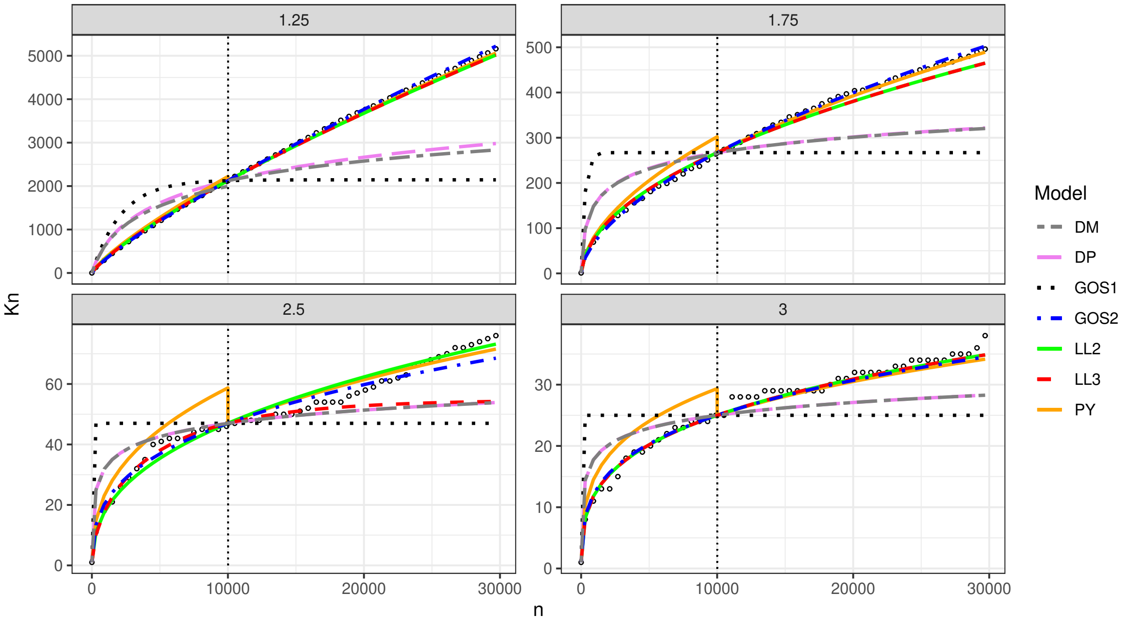

Figure B.8: In- and out-of-sample predictions for the different models on four randomly simulated curves from the Zipf distribution. White dots indicate the true values. Vertical line is the train-test cutoff. Curve simulation and model estimation proceeded as in Table B.8

B.5 Application: Copepod species counts

In this Section, we extend the discussion about the copepod species counts application of Section 5. In particular, we consider further train-test splits, we include plots for the model performances and estimates for the parameters.

In this analysis, we adopt a fully Bayesian approach to estimation when possible. This allows to compute 95% posterior credible intervals for the out-of-sample predictions, which we construct by simulating one posterior trajectory $K_{n+m} \mid D_1, \dots, D_n$ with $m \geqslant 1$ for each posterior sample. For the specific prior structure in each model, refer to the main paper. This is relatively easy for our log-logistic models and for the species sampling sequences. On the other hand, the latent beta reinforcements in the BETA-GOS processes make both the parameter likelihood intractable and posterior predictive checks costly. For example, drawing one posterior trajectory $K_{n+m} \mid D_1, \dots, D_n$ requires to sample $n + m$ independent beta random variables at each m, with an associated computational cost of $\mathcal{O}(m^2 + nm)$. This is infeasible for long accumulation curves like the one in the copepod data. Similarly to the simulation Section, we solve the first issue by doing estimation via method of moments. As for the second issue, we decide to simulate trajectories by fixing the values of each W_i to their averages. Here, the associated loss in uncertainty quantification is minor. To see this, it is sufficient to look at the variance of the discovery probability $r_n = \prod_{i=1}^n W_i$ at $n + 1$, which is equal to

$$\mathrm{var}(r_n) = \left\{\frac{\theta + \beta}{\theta}\frac{\theta + n}{\theta + \beta + n} - 1\right\}\left\{\frac{(\theta)_\beta}{(\theta + n)_\beta}\right\}^2. \tag{B.10}$$

This follows from the fact that the W_i's are independent beta random variables, so that $\mathrm{var}(r_n) = \prod_{i=1}^n \mathbb{E}(W_i^2) - \left\{\prod_{i=1}^n \mathbb{E}(W_i)\right\}^2$, with

$$\mathbb{E}(W_i) = \frac{i + \theta - 1}{i + \theta - 1 + \beta}, \quad \mathbb{E}(W_i^2) = \frac{i + \theta - 1}{i + \theta - 1 + \beta}\frac{i + \theta}{i + \theta + \beta}.$$

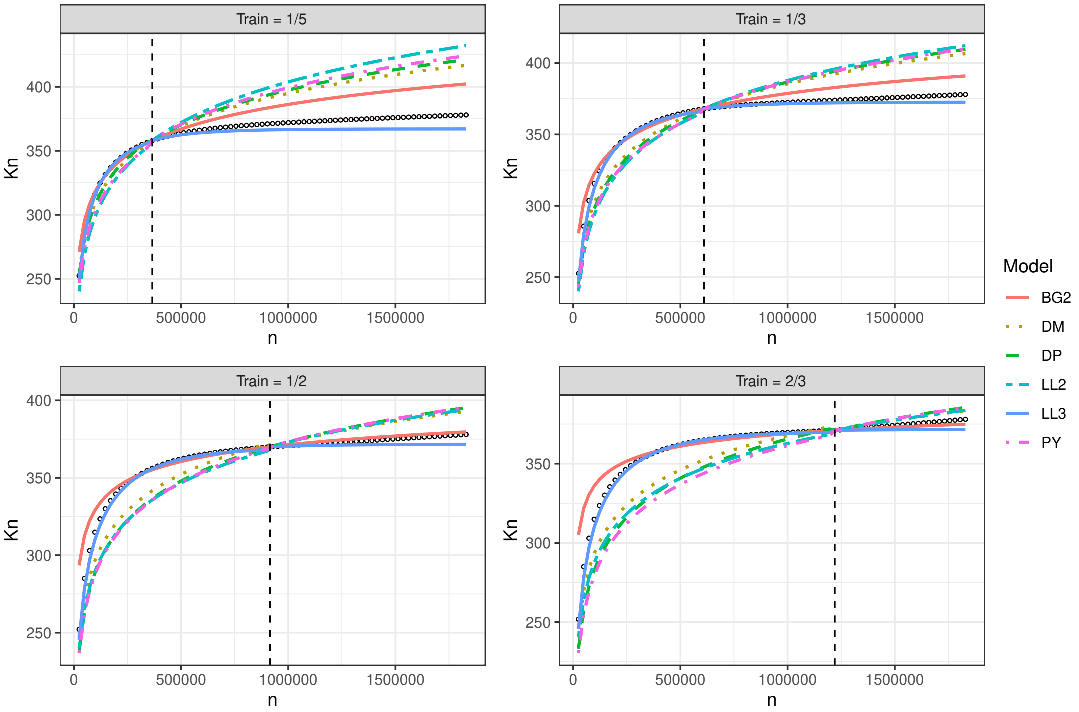

FIGURE B.9: In- and out-of-sample predictions over different random splits in the copepod dataset. White dots indicate true values. Lines obtained by averaging $\mathbb{E}(K_n)$ and $\mathbb{E}(K_{n+m} \mid D_1, \ldots, D_n)$ for each posterior sample.

For large values of θ and n, the quantities $(\theta + \beta)/\theta$ and $(\theta + n)/(\theta + \beta + n)$ in equation (B.10) are approximately one. Moreover, $(\theta + n)_\beta \to \infty$ as $n \to \infty$ for $\beta \geqslant 1$. Overall, this implies that the discovery probability r_n is heavily concentrated around its mean when n is large and $\beta \geqslant 1$. In this case, $\mathrm{var}(r_n) \to 0$ as $n \to \infty$.

Figure B.9 displays the in- and out-of-sample predictions of all the competitor models against different random splits of the data. The lines reported are a juxtaposition of the average values of $\mathbb{E}(K_n)$ and $\mathbb{E}(K_n \mid D_1 \ldots, D_n)$ computed for each posterior sample. We omit reporting the BETA-GOS-1 due to its lack of flexibility. The following are worth mentioning. First, the three-parameter log-logistic performs generally well in-sample, and slightly under-predicts the true number of species out-of-sample. Second, the closer is the train-test cutoff to the true length of the curve, the better the performance of the BETA-GOS-2 appears. This confirms the flexibility of the process already noted in the simulation study. On the other hand, all the other models estimate a wrong trajectory, irrespective of the cutoff. This is expected in the Pitman–Yor and in the Dirichlet process, as they assume an infinite species richness.

Tables B.9, B.10, B.11 and B.12 report the in-sample mean square error, the out-of-sample predictions and the average posterior estimates of the parameters for each model in the same setting of Figure B.9. True values $\bar{K}_{n+m}$ for the rarefaction curve are indicated in the first row. It is worth noticing the following. First, in the cases when the train-test cutoff is 1/5 and 1/3, the three-parameter log-logistic is almost the unique model that covers the true values of the curve. All the other models instead tend to overestimate the number of distinct species out-of-sample, with the exception of the BETA-GOS-1, which however shows a severe lack of flexibility. On the other hand, the 1/2 and the 2/3 scenarios seem to favor the BETA-GOS-2 as the most accurate model, whereas the three-parameter log-logistic converges too quickly.

Table B.9: Model performance and estimated parameters over the 1/5 train-test split.

TRAIN=1/5			$n = 365,953,\ K_n = 358$			
MODEL	PARAMETERS	MSE	$m = n/2$	$m = n$	$m = 2n$	$m = 4n$
$\bar{K}_{n+m}$			365.16	368.87	372.86	378
DP (LL1)	$\alpha = 39.1$	50.41	373.93	385.21	401.08	421.15
			(367, 382)	(375, 397)	(388, 415)	(405, 439)
PY	(α, σ) = (37.43, 0.01)	60.91	374.56	386.35	402.94	424.03
			(367, 384)	(376, 399)	(389, 419)	(406, 446)
DM	(σ, H) = (-0.01, 3436)	41.60	373.01	383.58	398.36	416.95
			(366, 381)	(374, 394)	(386, 412)	(401, 434)
BG-1(a, b)	$\rho > 0.99$	2703.71	358	358	358	358
			(358, 358)	(358, 358)	(358, 358)	(358, 358)
BG-2(θ, β)	(θ, β) = (54.75, 1.07)	73.8	369.72	377.79	388.87	402.4
			(364, 377)	(370, 387)	(379, 400)	(390, 416)
LL2	(α, σ) = (13, 16.71)	125.22	376.48	389.73	408.56	432.85
			(368, 386)	(378, 404)	(392, 428)	(410, 459)
LL3	(α, σ, ϕ) = (17.45, 0.12, ¿0.99)	1.59	363.7	365.91	367.32	367.8
			(359, 371)	(359, 377)	(360, 381)	(360, 384)

Table B.10: Model performance and estimated parameters over the 1/3 train-test split.

TRAIN=1/3			$n = 609,922,\ K_n = 368$			
MODEL	PARAMETERS	MSE	$m = n/4$	$m = n/2$	$m = n$	$m = 2n$
$\bar{K}_{n+m}$			370.32	371.8	373.97	378
DP (LL1)	$\alpha = 37.95$	130.99	376.56	383.47	394.47	409.87
			(371, 383)	(376, 392)	(384, 406)	(397, 424)
PY	$(\alpha, \sigma) = (37.48, 0)$	131.21	376.66	383.67	394.83	410.45
			(371, 383)	(376, 392)	(385, 406)	(397, 425)
DM	$(\sigma, H) = (-0.02, 3084)$	90.85	375.97	382.38	392.54	406.69
			(371, 382)	(375, 390)	(383, 403)	(394, 420)
BG-1(a, b)	$\rho > 0.99$	2101.28	368	368	368	368
			(368, 368)	(368, 368)	(368, 368)	(368, 368)
BG-2(θ, β)	$(\theta, \beta) = (65.84, 1.12)$	95.25	372.97	376.87	382.94	391.04
			(369, 378)	(371, 383)	(376, 391)	(382, 401)
LL2	$(\alpha, \sigma) = (15.22, 19.37)$	178.23	377.04	384.38	396.08	412.58
			(372, 384)	(376, 394)	(385, 409)	(397, 430)
LL3	(α, σ, ϕ) = (18.88, 0.1, ¿0.99)	3.52	370.17	371.33	372.4	372.93
			(368, 374)	(368, 377)	(368, 380)	(368, 381)

Table B.11: Model performance and estimated parameters over the 1/2 train-test split.

TRAIN=1/2			$n = 914,884,\ K_n = 370$			
MODEL	PARAMETERS	MSE	$m = n/5$	$m = n/3$	$m = n/2$	$m = n$
$\bar{K}_{n+m}$			371.63	372.68	373.99	378
DP (LL1)	$\alpha = 36.49$	288.2	376.69 (372, 382)	380.54 (375, 387)	384.87 (378, 393)	395.39 (386, 406)
PY	$(\alpha, \sigma) = (36.26, 0)$	289.62	376.74 (372, 382)	380.62 (375, 388)	384.98 (378, 393)	395.59 (386, 406)
DM	$(\sigma, H) = (-0.02, 2406)$	181.41	376.03 (372, 381)	379.51 (374, 386)	383.4 (377, 391)	392.81 (383, 403)
BG-1(a, b)	$\rho > 0.99$	1517.17	370 (370, 370)	370 (370, 370)	370 (370, 370)	370 (370, 370)
BG-2(θ, β)	$(\theta, \beta) = (83.39, 1.19)$	140.17	372.67 (370, 376)	374.18 (371, 379)	375.83 (372, 381)	379.68 (374, 386)
LL2	$(\alpha, \sigma) = (17.86,\ 22.58)$	246.35	376.39 (372, 382)	380.08 (374, 387)	384.2 (377, 393)	394.22 (384, 406)
LL3	$(\alpha, \sigma, \phi) = (19.75, 0.1, > 0.99)$	3.87	370.84 (370, 373)	371.17 (370, 374)	371.44 (370, 375)	371.76 (370, 376)

Table B.12: Model performance and estimated parameters over the 2/3 train-test split.

TRAIN=2/3			$n = 1,219,845,\ K_n = 371$			
MODEL	PARAMETERS	MSE	$m = n/10$	$m = n/5$	$m = n/3$	$m = n$
$\bar{K}_{n+m}$			372.4	373.76	375.63	378
DP (LL1)	$\alpha = 35.48$	441.12	374.4 (371, 378)	377.45 (373, 383)	381.21 (375, 388)	385.44 (378, 394)
PY	$(\alpha, \sigma) = (35.34, 0)$	436.06	374.42 (371, 378)	377.49 (373, 383)	381.27 (375, 388)	385.53 (378, 394)
DM	$(\sigma, H) = (-0.03, 1759)$	251.61	373.93 (371, 378)	376.55 (372, 382)	379.78 (374, 386)	383.39 (377, 391)
BG-1(a, b)	$\rho > 0.99$	1174.38	371 (371, 371)	371 (371, 371)	371 (371, 371)	371 (371, 371)
BG-2(θ, β)	$(\theta, \beta) = (100.95, 1.25)$	182.05	371.97 (371, 374)	372.81 (371, 376)	373.82 (371, 378)	374.91 (372, 379)
LL2	$(\alpha, \sigma) = (19.94, 25.1)$	302.53	374.02 (371, 378)	376.72 (372, 382)	380.04 (374, 387)	383.76 (377, 392)
LL3	$(\alpha, \sigma, \phi) = (20.01, 0.09, > 0.99)$	3.53	371.23 (371, 373)	371.38 (371, 373)	371.51 (371, 374)	371.61 (371, 374)

B.6 Finnish fungal biodiversity imputation strategy

In this Section, we comment on the imputation strategy of the finnish fungal biodiversity study. The original version of the data consist of 174 samples from different 5 cities in Finland (Helsinki, Lathi, Joensuu, Tampere and Jyväskylä), collected either from air or soil in an urban or a rural environment. Each sample reports the abundances (i.e. the frequencies) of the different fungal *operational taxonomic units* - OTUs, a proxy for species - identified via high-throughput sequencing of the collected DNA. Such identification pipeline, however, is potentially subject to sequencing error. To cope for the issue, ecologists often remove all the OTUs appearing only once in the a given sample. See (Abrego et al., 2020) for further details.

The presence of true singletons in the data however has a large impact on the steepness of the accumulation curves and in the subsequent estimators for the species richness (e.g Chao, 1984). For this reason, we decide to adopt the following imputation strategy. Let $j = 1, 2, 3 \dots$ be the count index, and $\bar{n}_j$ be the number of species in a given sample having frequency j. In our data, $\bar{n}_1 = 0$. To impute it, we run the following linear model:

$$\log \bar{n}_j = \delta_0 + \delta_1 \log j + \epsilon_j,$$

for $j = 2, \dots, J$. The parameters δ_0 and δ_1 are estimated through ordinary least squares, rounded to the closes integer. Then, the estimate is for $\bar{n}_1$ is exactly $\exp\{\hat{\delta}_0\}$, namely the exponential of the intercept term. Notice that the choice for the truncation frequency J has to be carefully selected, as typically $\bar{n}_j$ is equal to 1 for large values of j. This is because highly frequent species do not appear the same exact number of times in large samples. If we where to include all the possible values for j, this would result in a very low estimate for the intercept. Our choice for J is the first j for which $\sum_{s=2}^{j} \bar{n}_s / k \geqslant 0.75$, where k is the total number of species in the samples excluding the singletons.

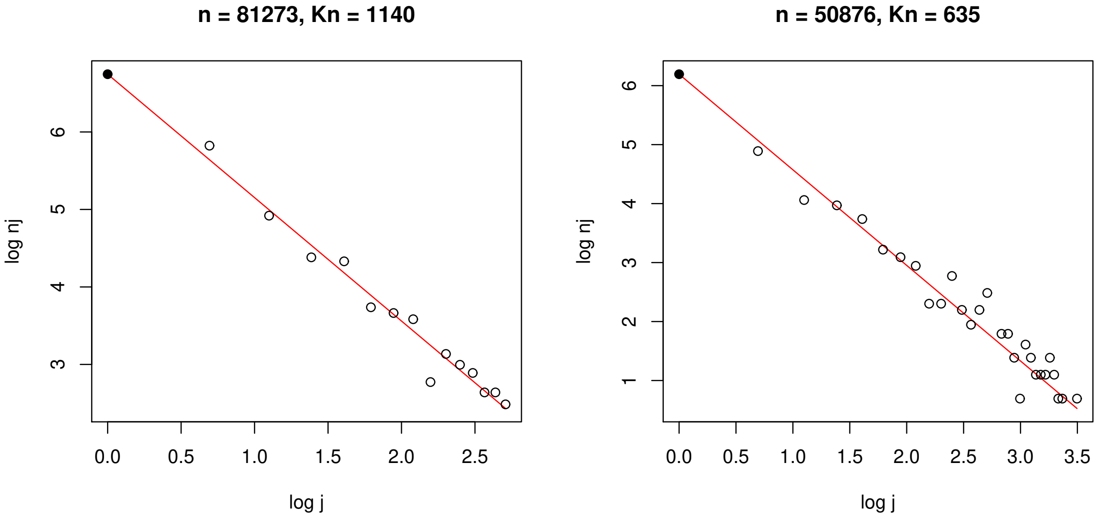

Figure B.10: Linear imputation of the singletons in two samples from the fungal biodiversity study. The red line is the least squares regression line, the circles indicate the observed frequencies and the black dot indicates the imputed value for $\log \bar{n}_1$

Figure B.10 reports two examples of the our linear model imputation strategy. For further details on the data, refer to Chapter 3.

Appendix C

Supplementary material for Chapter 4 - Inferring taxonomic placement from DNA barcoding aiding in discovery of new taxa

This Appendix contains mathematical details for the BayesANT algorithm - BAYESiAn Nonparametric Taxonomic classifier - described in Chapter 4. Emphasis is on the explicit formulas for the taxonomic annotation probabilities and the associated estimation method for the model parameters. Moreover, it reports additional simulations 1) to test the algorithm under the presence of alignment gaps "-" in both the training and the test sequences, and 2) to test the method when the training library is significantly smaller than the test one. The Appendix is divided as follows. Section C.1 contains details on the prior taxonomic probabilities, and Section C.2 on the associated posterior ones. Section C.3 contains additional analyses on the FinBOL data. Section C.4 assesses the impact of the alignment gaps on the prediction. Finally, Section C.5 evaluates the performance of BayesANT under varying size of the training library.

C.1 Prior taxonomic probabilities

In this Section, we provide details on the prior probabilities over the nodes in the taxonomic tree and describe the estimation procedure for the associated hyperparameters.

C.1.1 The Pitman–Yor process and the exchangeable partition probability function

BayesANT models taxonomic novelty via Pitman–Yor process priors (Pitman and Yor, 1997). As already detailed in the main paper, the process works as follows. Let $X_1, \dots, X_n$ be a sequence of taxon assignments for the DNA sequences in the training library, comprising of a total of $K_n = k$ distinct labels denoted as $X_1^*, \dots, X_k^*$ and appearing with frequencies $n_1, \dots, n_k$. Then, the taxon of the $(n+1)$th observation is determined via the following allocation scheme:

$$(X_{n+1} \mid X_1, \dots, X_n) = \begin{cases} \text{“new”}, & \text{with prob.} \quad (\alpha + \sigma k)/(\alpha + n), \\ X_j^*, & \text{with prob.} \quad (n_j - \sigma)/(\alpha + n), \ (j = 1, \dots, k), \end{cases} \tag{C.1}$$

where $\sigma \in [0, 1)$ is a discount parameter governing the tail of the process and $\alpha > -\sigma$ is a precision parameter. High values for α and σ lead to a high number of distinct labels K_n. Moreover, high values for n_j lead to a high probability that taxon X_j^* will be observed in the future. See Figure 1 in the main paper for a practical illustration. Estimation of both parameters can be performed via an empirical Bayes procedure (Favaro et al., 2009) through maximization of the quantity known as *exchangeable partition probability function* (EPPF, Pitman, 1996). Let N_j denote the random variable corresponding to the frequency of appearance of taxon X_j^*, with n_j the realization of this random variable for $K_n = k$. The EPPF is defined as

$$p(K_n = k, N_1 = n_1, \ldots, N_k = n_k) = \frac{\prod_{i=1}^{k-1}(\alpha + i\sigma)}{(\alpha+1)_{n-1}} \prod_{j=1}^{k} (1-\sigma)_{n_j - 1}, \tag{C.2}$$

where $(x)_a = \Gamma(x + a)/\Gamma(x)$ is the Pochhammer factorial and $\Gamma(x)$ is the gamma function. The quantity in equation (C.2) can be interpreted as a likelihood function arising from the process in equation (C.1). Then, one can simply apply maximum likelihood estimation as

$$(\hat{\alpha}, \hat{\sigma}) = \arg\max_{\alpha,\sigma} \left\{ \frac{\prod_{i=1}^{k-1}(\alpha + i\sigma)}{(\alpha+1)_{n-1}} \prod_{j=1}^{k} (1-\sigma)_{n_j - 1} \right\}, \quad \sigma \in [0, 1), \alpha > -\sigma.$$

In what follows, we apply this procedure to estimate the parameters α_ℓ and σ_ℓ for all levels $\ell = 1, \ldots, L$ in the taxonomic tree.

C.1.2 Level-specific Pitman–Yor priors

Consider a taxonomic library $\mathscr{D}_n = (\mathbf{X}_i, \mathbf{Y}_i)_{i=1}^n$ of size n and of $L \geqslant 2$ levels, where $\mathbf{X}_i = (X_{i,\ell})_{\ell=1}^L$ are the taxonomic annotations for DNA sequence $\mathbf{Y}_i$. Following the notation in the main paper, we let $X^*_{j,\ell}$ be the jth taxon and $\mathbf{X}^{(n)}_{\cdot,\ell} = (X_{i,\ell})_{i=1}^n$ be the sequence of taxa observed for level ℓ. To construct the taxonomic tree, we introduce the following quantities. For a generic taxon x_ℓ at level ℓ, we define $\mathrm{pa}(x_\ell)$ as the unique parent node of x_ℓ at level $\ell - 1$ and $\rho_n(x_\ell)$ as the set of nodes $x_{\ell+1}$ at level $\ell + 1$ such that $\mathrm{pa}(x_{\ell+1}) = x_\ell$. We also let $K_n(x_\ell) = |\rho_n(x_\ell)|$ be the number of nodes linked to x_ℓ at level $\ell + 1$ and $N_n(x_\ell)$ be the size of the taxon, namely the number of DNA sequences linked to x_ℓ. Then, BayesANT follows the prediction scheme in equation (C.1) by letting

$$(X_{n+1,\ell} \mid X_{n+1,\ell-1} = x_{\ell-1}, \mathbf{X}^{(n)}_{\cdot,\ell}) = \begin{cases} \text{“new”}, & \text{with probability} \quad \frac{\alpha_\ell + \sigma_\ell K_n(x_{\ell-1})}{\alpha_\ell + N_n(x_{\ell-1})}, \\ X^*_{j,\ell}, & \text{with probability} \quad \frac{N_n(X^*_{j,\ell}) - \sigma_\ell}{\alpha_\ell + N_n(x_{\ell-1})}, \end{cases} \tag{C.3}$$

for $j : \mathrm{pa}(X^*_{j,\ell}) = x_{\ell-1}$, where $\sigma_\ell \in [0,1)$ and $\alpha_\ell > -\sigma_\ell$ are rank-specific parameters. Equation (C.3) holds independently for all the observed nodes $x_{\ell-1}$ at level $\ell - 1$. Specifically, we model all the separate sets of taxa $\rho_n(x_{\ell-1})$ at a given rank ℓ as realizations from independent Pitman–Yor processes. In estimating parameters α_ℓ, σ_ℓ, we borrow information across branches. The level-specific EPPF is a product of EPPFs, and the estimates for α_ℓ and σ_ℓ are obtained as

$$(\hat{\alpha}_\ell, \hat{\sigma}_\ell) = \arg\max_{\alpha_\ell,\sigma_\ell} \left\{ \prod_{x \in \mathbb{X}_{\ell-1}} \frac{\prod_{i=1}^{K_n(x)-1} (\alpha_\ell + i\sigma_\ell)}{(\alpha_\ell + 1)_{N_n(x)-1}} \prod_{x_\ell \in \rho_n(x)} (1 - \sigma_\ell)_{N_n(x_\ell)-1} \right\}, \quad \text{(C.4)}$$

for $\sigma_\ell \in [0,1), \alpha_\ell > -\sigma_\ell$, where $\mathbb{X}_{\ell-1}$ denotes the set of all taxonomic nodes $X^*_{j,\ell-1}$ observed in the library at level $\ell - 1$. The maximization in equation (C.4) can easily be carried out with routine methods such as `nlminb` in `R`.

C.1.3 Taxonomic branch probabilities

After α_ℓ and σ_ℓ have been estimated from the library $\mathscr{D}_n$, BayesANT specifies the prior probability for each branch $\mathbf{x} = (x_1, \dots, x_L)$, including ones with new nodes, through the chain rule as a product of conditional Pitman–Yor probabilities. For the $(n+1)$th sequence, this is equal to

$$\begin{aligned} \pi_{n+1}(\mathbf{x}) &= p(\mathbf{X}_{n+1} = \mathbf{x} \mid \mathbf{X}^{(n)}) \\ &= p(V_{n+1,1} = x_1 \mid \mathbf{X}^{(n)}_{\cdot,1}) \times \prod_{\ell=2}^{L} p(V_{n+1,\ell} = x_\ell \mid X_{n+1,\ell-1} = x_{\ell-1}, \mathbf{X}^{(n)}_{\cdot,\ell}). \end{aligned} \quad \text{(C.5)}$$

To see this explicitly, consider the example of a $L = 4$ level library of size n and let $X^*_{1,1} \to X^*_{1,2} \to X^*_{1,3} \to X^*_{1,4}$ be a branch. This is a fully observed branch, for which $\mathrm{pa}(X^*_{1,2}) = X^*_{1,1}$, $\mathrm{pa}(X^*_{1,3}) = X^*_{1,2}$ and $\mathrm{pa}(X^*_{1,4}) = X^*_{1,3}$. Then, the prior probability of $\mathbf{x}_1 = (X^*_{1,1}, X^*_{1,2}, X^*_{1,3}, X^*_{1,4})$ is

$$\pi_{n+1}(\mathbf{x}_1) = \underbrace{\frac{N_n(X^*_{1,1}) - \sigma_1}{\alpha_1 + n}}_{\substack{\text{Prob. of choosing} \\ X^*_{1,1} \text{ at Level 1}}} \times \underbrace{\frac{N_n(X^*_{1,2}) - \sigma_2}{\alpha_2 + N_n(X^*_{1,1})}}_{\substack{\text{Prob. of choosing} \\ X^*_{1,2} \text{ at Level 2}}} \times \underbrace{\frac{N_n(X^*_{1,3}) - \sigma_3}{\alpha_3 + N_n(X^*_{1,2})}}_{\substack{\text{Prob. of choosing} \\ X^*_{1,3} \text{ at Level 3}}} \times \underbrace{\frac{N_n(X^*_{1,4}) - \sigma_4}{\alpha_4 + N_n(X^*_{1,3})}}_{\substack{\text{Prob. of choosing} \\ X^*_{1,4} \text{ at Level 4}}}.$$

Consider instead the path $X^*_{1,1} \to X^*_{1,2} \to$ "new" $\to$ "new". This identifies a new clade at level $\ell = 3$, creating a new leaf. We denote the newly created clade as $\mathbf{x}_2 = (X^*_{1,1}, X^*_{1,2}, x_3^{\text{new}}, x_4^{\text{new}})$. Then, its associated probability is

$$\pi_{n+1}(\mathbf{x}_2) = \underbrace{\frac{N_n(X^*_{1,1}) - \sigma_1}{\alpha_1 + n}}_{\substack{\text{Prob. of choosing} \\ X^*_{1,1} \text{ at Level 1}}} \times \underbrace{\frac{N_n(X^*_{1,2}) - \sigma_2}{\alpha_2 + N_n(X^*_{1,1})}}_{\substack{\text{Prob. of choosing} \\ X^*_{1,2} \text{ at Level 2}}} \times \underbrace{\frac{\alpha_3 + \sigma_3 K_n(X^*_{1,2})}{\alpha_3 + N_n(X^*_{1,2})}}_{\substack{\text{Prob. of novelty} \\ \text{at Level 3}}} \times \underbrace{1}_{\substack{\text{Prob. of novelty} \\ \text{at Level 4}}}.$$

Here, the novelty probability of the Pitman–Yor process appears at level 3. At level 4 the probability is equal to one since the node is necessarily new and $K_n(x_4^{\text{new}}) = N_n(x_4^{\text{new}}) = 0$. In a similar fashion, the probability for the branch "new" $\to$ "new" $\to$ "new" $\to$ "new", namely $\mathbf{x}_3 = (x_1^{\text{new}}, x_2^{\text{new}}, x_3^{\text{new}}, x_4^{\text{new}})$ is

$$\pi_{n+1}(\mathbf{x}_3) = \underbrace{\frac{\alpha_1 + \sigma_1 K_n(v_0)}{\alpha_1 + n}}_{\substack{\text{Prob. of novelty} \\ \text{at Level 1}}} \times \underbrace{1}_{\substack{\text{Prob. of novelty} \\ \text{at Level 2}}} \times \underbrace{1}_{\substack{\text{Prob. of novelty} \\ \text{at Level 3}}} \times \underbrace{1}_{\substack{\text{Prob. of novelty} \\ \text{at Level 4}}},$$

since the novelty is detected first at level $\ell = 1$. Finally, notice that a branch such as $X^*_{1,1} \to X^*_{1,2} \to$ "new" $\to X^*_{1,4}$ is not allowed in our representation. Under such a formulation, we are able to specify all the prior probabilities for both the observed taxa and the new ones in a coherent way.

C.2 Posterior taxonomic probabilities

BayesANT assumes DNA sequences $\mathbf{Y}_i$ associated with branch $\mathbf{x} = (x_1, \dots, x_L)$ with $x_\ell \in \mathbb{X}_\ell$ are distributed as

$$(\mathbf{Y}_i \mid \mathbf{X}_i = \mathbf{x}, \boldsymbol{\theta}_{x_L}) \overset{\text{iid}}{\sim} \mathcal{K}(\mathbf{Y}_i; \boldsymbol{\theta}_{x_L}),$$

where $\mathcal{K}(\cdot, \boldsymbol{\theta}_{x_L})$ is a distribution that depends on the leaf-specific vector of parameters $\boldsymbol{\theta}_{x_L}$. In what follows, we provide mathematical details of the single location product-multinomial model we use in the main document. Adapting the details to accommodate alternative kernels is straightforward.

C.2.1 Multinomial kernel

Let $\mathbf{Y}_i$, $i = 1, \dots, n$, indicate a collection of *aligned* DNA sequences of length p. The alignment of the sequences makes the individual loci comparable across taxa. An example for $p = 20$ loci is as follows:

LOCUS s	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
$\mathbf{Y}_1$:	A	C	C	T	C	G	G	A	A	A	T	T	T	G	G	A	A	T	C	A
$\mathbf{Y}_2$:	A	C	T	T	C	G	A	A	T	A	T	A	A	G	A	G	A	T	G	G
$\mathbf{Y}_3$:	A	T	T	C	C	G	T	A	G	G	T	T	T	G	A	G	T	T	G	A

As each locus in each sequence corresponds to a nucleotide in $\mathcal{N}_1 = \{\text{A, C, G, T}\}$, it is natural to use a multinomial likelihood,

$$(Y_{i,s} \mid \mathbf{X}_i = \mathbf{x}, \boldsymbol{\theta}_{x_L,s}) \overset{\text{iid}}{\sim} \text{Mult}(1; \theta_{x_L,s,\text{A}}, \theta_{x_L,s,\text{C}}, \theta_{x_L,s,\text{G}}, \theta_{x_L,s,\text{T}}),$$

where $\boldsymbol{\theta}_{x_L,s} = (\theta_{x_L,s,\text{A}}, \theta_{x_L,s,\text{C}}, \theta_{x_L,s,\text{G}}, \theta_{x_L,s,\text{T}})^\top$ is a vector of probabilities summing up to 1, and $\theta_{x_L,s,g}$ is the probability of observing nucleotide g in position s for leaf x_L. To simplify the analysis and ease computation, we further assume that all locations s are independent, leading to the following likelihood contribution for the ith sequence:

$$\mathcal{K}(\mathbf{Y}_i; \boldsymbol{\theta}_x) = \prod_{s=1}^{p} \prod_{g \in \mathcal{N}_1} \theta_{x,s,g}^{\mathbb{1}\{Y_{i,s}=g\}}, \tag{C.6}$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function. As a conjugate prior for the nucleotide probabilities, we choose

$$\boldsymbol{\theta}_{x,s} \sim \text{Dir}(\xi_{x,s,\text{A}}, \xi_{x,s,\text{C}}, \xi_{x,s,\text{G}}, \xi_{x,s,\text{T}}),$$

with $\boldsymbol{\xi}_{x,s} = (\xi_{x,s,\mathrm{A}}, \xi_{x,s,\mathrm{C}}, \xi_{x,s,\mathrm{G}}, \xi_{x,s,\mathrm{T}})^\top$ a vector of hyperparameters. The posterior distribution for the nucleotide probabilities at locus s under leaf x conditional on the DNA sequences assigned to x is then

$$(\boldsymbol{\theta}_{x,s} \mid \mathscr{D}_n) \sim \mathrm{Dir}(\xi_{x,s,\mathrm{A}} + n_{x,s,\mathrm{A}}, \xi_{x,s,\mathrm{C}} + n_{x,s,\mathrm{C}}, \xi_{x,s,\mathrm{G}} + n_{x,s,\mathrm{G}}, \xi_{x,s,\mathrm{T}} + n_{x,s,\mathrm{T}}),$$

where $n_{x,s,g} = \sum_{i:X_{i,L}=x} \mathbb{1}\{X_{i,s} = g\}$ indicates the number of times nucleotide $g \in \mathcal{N}_1$ is recorded at locus s for the DNA sequences linked to leaf x. The resulting posterior kernel for $\boldsymbol{\theta}_x$ is then a product of independent Dirichlet distributions, namely

$$p(\boldsymbol{\theta}_x \mid \mathscr{D}_n) \propto \prod_{s=1}^{p} \prod_{g \in \mathcal{N}_1} \theta_{x,s,g}^{\xi_{x,s,g} + n_{x,s,g}}. \tag{C.7}$$

An equivalent representation can be obtained for the 2-mer location multinomial kernel detailed in Chapter 4, but with the support of the multinomial being all pairs of nucleotides instead of singletons. When sequences are not aligned, the posterior in equation (C.7) is modified to remove the product from $s = 1, \ldots, p$ and substitute $n_{x,s,g}$ with $n_{x,g}$, which is the number of times κ-mer g is recorded in the sequence.

C.2.2 Prior and posterior predictive distribution

Once the parameters for the posterior distribution in equation (C.7) have been computed for each $\mathbf{x}$, BayesANT determines the prediction probabilities by following a similar reasoning behind naïve Bayes classifiers and linear discriminant analysis. In particular, let $\mathbf{X}^{(n)} = (\mathbf{X}_i)_{i=1}^n$ be the collection of labels and and $\mathbf{Y}^{(n)} = (\mathbf{Y}_i)_{i=1}^n$ be the associated DNA, and recall that $\mathscr{D}_n = (\mathbf{X}^{(n)}, \mathbf{X}^{(n)})$. Then, the conditional distribution for the taxa of the $(n+1)$th sequence is obtained as

$$\begin{aligned}
p(\mathbf{X}_{n+1} \mid \mathbf{Y}_{n+1}, \mathscr{D}_n) &\propto p(\mathbf{X}_{n+1}, \mathbf{Y}_{n+1}, \mathbf{Y}^{(n)}, \mathbf{X}^{(n)}) \\
&\propto p(\mathbf{X}_{n+1}, \mathbf{Y}^{(n)}) p(\mathbf{Y}_{n+1}, \mathbf{Y}^{(n)} \mid \mathbf{X}_{n+1}, \mathbf{Y}^{(n)}) \\
&\propto p(\mathbf{X}_{n+1} \mid \mathbf{Y}^{(n)}) p(\mathbf{Y}^{(n)}) p(\mathbf{Y}^{(n)} \mid \mathbf{X}_{n+1}, \mathbf{Y}^{(n)}) p(\mathbf{Y}_{n+1} \mid \mathbf{X}_{n+1}, \mathbf{Y}^{(n)}, \mathbf{Y}^{(n)}) \\
&\propto p(\mathbf{X}_{n+1} \mid \mathbf{Y}^{(n)}) p(\mathbf{Y}_{n+1} \mid \mathbf{X}_{n+1}, \mathscr{D}_n).
\end{aligned}$$

The above proportionality holds since $p(\mathbf{X}^{(n)} \mid \mathbf{X}_{n+1}, \mathbf{Y}^{(n)}) = p(\mathbf{X}^{(n)} \mid \mathbf{Y}^{(n)})$ and the joint distribution of the data $p(\mathscr{D}_n) = p(\mathbf{X}^{(n)})p(\mathbf{Y}^{(n)} \mid \mathbf{X}^{(n)})$ is independent of $\mathbf{X}_{n+1}$ and $\mathbf{Y}_{n+1}$ and thus can regarded as constant. Calling $p_{n+1}(\mathbf{x}) = p(\mathbf{X}_{n+1} = \mathbf{x} \mid \mathbf{Y}_{n+1}, \mathscr{D}_n)$, it is straightforward from the above that

$$p_{n+1}(\mathbf{x}) \propto \pi_{n+1}(\mathbf{x}) \int \mathcal{K}(\mathbf{Y}_{n+1}; \boldsymbol{\theta}_{x_L}) p(\boldsymbol{\theta}_{x_L} \mid \mathscr{D}_n) \mathrm{d}\boldsymbol{\theta}_{x_L}, \tag{C.8}$$

where $\pi_{n+1}(\mathbf{x}) = p(\mathbf{X}_{n+1} = \mathbf{x} \mid \mathbf{X}^{(n)})$ is the prior probability defined in equation (C.5), while the integral is the posterior predictive distribution for DNA sequence $\mathbf{X}_{n+1}$ under branch $\mathbf{x}$. Notice that if $\mathbf{x}$ contains a "new" taxon, equation (C.8) becomes a prior predictive distribution. The integral is explicitly available for the multinomial kernel with Dirichlet prior defined above. Specifically, it is straightforward to see that the marginal distribution for $X_{i,s}$ when the posterior follows a Dirichlet is

$$X_{n+1,s} \sim \mathrm{Mult}\Big(1; \frac{\xi_{x,s,\mathrm{A}} + n_{x,s,\mathrm{A}}}{M_{x,s}}, \frac{\xi_{x,s,\mathrm{C}} + n_{x,s,\mathrm{C}}}{M_{x,s}}, \frac{\xi_{x,s,\mathrm{G}} + n_{x,s,\mathrm{G}}}{M_{x,s}}, \frac{\xi_{x,s,\mathrm{T}} + n_{x,s,\mathrm{T}}}{M_{x,s}}\Big), \tag{C.9}$$

where $M_{x,s} = \sum_{g \in \mathcal{N}_1} (\xi_{x,s,g} + n_{x,s,g})$ is a normalizing constant for the nucleotide probabilities. The corresponding prior predictive probability is obtained by setting $n_{x,s,g} = 0$ for every $g \in \mathcal{N}_1$ and normalizing by $\xi_{x,s,0} = \sum_{g \in \mathcal{N}_1} \xi_{x,s,g}$. Then, from equation (C.9) and the location independence assumption, it can be easily shown that equation (C.8) reduces to

$$p_{n+1}(\mathbf{x}) \propto \begin{cases} \pi_{n+1}(\mathbf{x}) \prod_{s=1}^{p} (\xi_{x_L,s,g_s} + n_{x_L,s,g_s})/M_{x_L,s}, & \text{if } x_L \text{ is an observed leaf,} \\ \pi_{n+1}(\mathbf{x}) \prod_{s=1}^{p} \xi_{x_L,s,g_s}/\xi_{x,s,0}, & \text{if } x_L \text{ is a novel leaf,} \end{cases} \tag{C.10}$$

where $g_s \in \mathcal{N}_1$ is the nucleotide of sequence $\mathbf{Y}_{n+1}$ in locus $s = 1, \ldots, p$. Similar considerations hold for the κ-product multinomial kernel and the not-aligned multinomial kernel.

C.2.3 Hyperparameter tuning

From equation (C.10), it is easy to see that the hyperparameters $\boldsymbol{\xi}_{v,s}$ play an important role in defining the prediction probabilities. This is especially true for "new" leaves since no nucleotides are observed. As discussed in the main paper, uniform priors in this context may underestimate the predicted number of new taxa at the lowest level. Therefore, we need a method to tune $\boldsymbol{\xi}_{v,s}$ based on the information available in the taxonomic tree. To address this goal, we apply a method of moments estimator as detailed below.

For a node x_ℓ at level ℓ, call $\mathcal{L}_n(x_\ell)$ the set of leaves linked to it. Under the assumption that

$$\boldsymbol{\theta}_{x,s} \overset{\text{iid}}{\sim} \text{Dir}(\xi_{x_\ell,s,\text{A}}, \xi_{x_\ell,s,\text{C}}, \xi_{x_\ell,s,\text{G}}, \xi_{x_\ell,s,\text{T}}), \quad \text{for all } x \in \mathcal{L}_n(x_\ell),$$

each nucleotide probability is marginally distributed as $\theta_{x,s,g} \sim \textsc{beta}(\xi_{x_\ell,s,g}, \xi_{x_\ell,s,0} - \xi_{x_\ell,s,g})$, with $\xi_{x_\ell,s,0} = \sum_{g\in\mathcal{N}_1} \xi_{x_\ell,s,g}$ being the sum of the hyperparameters. From the moments of a beta distribution, it holds that

$$\mathbb{E}(\theta_{x,s,g}) = \frac{\xi_{x_\ell,s,g}}{\xi_{x_\ell,s,0}}, \qquad \text{and} \qquad \mathbb{E}(\theta^2_{x,s,g}) = \frac{\xi_{x_\ell,s,g}(\xi_{x_\ell,s,g}+1)}{\xi_{x_\ell,s,0}(\xi_{x_\ell,s,0}+1)},$$

for $g \in \mathcal{N}_1$. Our goal is to estimate both $\xi_{x_\ell,s,0}$ and $\xi_{x_\ell,s,g}$. This can be done as follows. Recall that $N_n(x)$ and $n_{x,s,g}$ are the number of sequences and the number of times nucleotide g is recorded at locus s for all sequences linked to leaf x, respectively. Our first method of the moments equation is

$$\hat{\theta}_{x_\ell,s,g} = \frac{1}{N_n(x_\ell)} \sum_{x\in\mathcal{L}_n(x_\ell)} \frac{n_{x,s,g}}{N_n(x)} = \frac{\xi_{x_\ell,s,g}}{\xi_{x_\ell,s,0}} = \mathbb{E}(\theta_{x,s,g}). \tag{C.11}$$

Here, $\hat{\theta}^g_{x_\ell,s}$ is an average of the observed proportion of times nucleotide g appears in the sequences linked to all leaves x connected to x_ℓ. For our second equation, we set

$$\hat{S}_{x_\ell,s} = \frac{1}{N_n(x_\ell)} \sum_{x \in \mathcal{L}_n(x_\ell)} \sum_{g \in \mathcal{N}_1} \Big(\frac{n_{x,s,g}}{N_n(v)}\Big)^2 = \sum_{g \in \mathcal{N}_1} \frac{\xi_{x_\ell,s,g}(\xi_{x_\ell,s,g}+1)}{\xi_{x_\ell,s,0}(\xi_{x_\ell,s,0}+1)} = \sum_{g \in \mathcal{N}_1} \mathbb{E}(\theta^2_{x,s,g}), \tag{C.12}$$

where $\hat{S}_{x_\ell,s}$ is the average sum of the squared nucleotide proportions for all x linked to x_ℓ. Notice that the third component in the equation can be further simplified as

$$\sum_{g \in \mathcal{N}_1} \mathbb{E}(\theta^2_{x,s,g}) = \sum_{g \in \mathcal{N}_1} \frac{\xi_{x_\ell,s,g}(\xi_{x_\ell,s,g}+1)}{\xi_{x_\ell,s,0}(\xi_{x_\ell,s,0}+1)} = \frac{1}{\xi_{x_\ell,s,0}+1} \Big\{ \xi^0_{x_\ell,s} \sum_{g \in \mathcal{N}_1} \Big(\frac{\xi_{x_\ell,s,g}}{\xi_{x_\ell,s,0}}\Big)^2 + 1 \Big\}.$$

Then, plugging equation (C.11) into (C.12) and letting $m_{v_\ell,s} = \sum_{g \in \mathcal{N}_1} \hat{\theta}^2_{x_\ell,s,g}$, one has that

$$\hat{S}_{x_\ell,s} = \frac{1}{\xi_{x_\ell,s,0}+1}(\xi_{x_\ell,s,0} m_{x_\ell,s} + 1),$$

which, combined with equation (C.11), yields

$$\xi_{x_\ell,s,0} = \frac{1 - \hat{S}_{x_\ell,s}}{\hat{S}_{x_\ell,s} - m_{x_\ell,s}}, \qquad \text{and} \qquad \xi_{x_\ell,s,g} = \xi_{x_\ell,s,0}\hat{\theta}_{x_\ell,s,g}, \quad g \in \mathcal{N}_1. \tag{C.13}$$

The quantities in equation (C.13) are the method of the moments estimators for our hyperparameters, and we can use them to borrow information across branches, as discussed in the main paper. We detail the procedure in Algorithm 5 below.

The idea behind this procedure is to incorporate taxonomic dependencies between the leaves, especially novel ones. The procedure works equally for the κ-product multinomial kernel.

C.3 Addedum to the analysis on FinBOL

In this Section, we provide additional results on the FinBOL datasets, prediction of novel sequences by method, accuracies for model M-1 divided by order, and computational times.

Algorithm 5: Hyperparameter tuning via the method of moments for the multinomial kernel

```
for leaf $x_L \in \mathbb{X}_L$ do
    if $x_L$ is a known taxon then
        Estimate $\xi_{x_{L-1},s,0}$ and $\xi_{x_{L-1},s,g}$, $g \in \mathcal{N}_1$, from equation (C.13), where
          $x_{L-1} = \mathrm{pa}(x_L)$;
        Set prior $\boldsymbol{\theta}_{x_L,s} \sim \mathrm{Dir}(\xi_{x_{L-1},s,\mathrm{A}}, \xi_{x_{L-1},s,\mathrm{C}}, \xi_{x_{L-1},s,\mathrm{G}}, \xi_{x_{L-1},s,\mathrm{G}})$;
    else
        if $x_L$ is a new taxon then
            Estimate $\xi_{x_\ell,s,0}$ and $\xi_{x_\ell,s,g}$, $g \in \mathcal{N}_1$, from equation (C.13), where $x_\ell$
              is the lowest possible known taxon along the branch leading to
              $x_L$;
            Set prior $\boldsymbol{\theta}_{x_L,s} \sim \mathrm{Dir}(\xi_{x_\ell,s,\mathrm{A}}, \xi_{x_\ell,s,\mathrm{C}}, \xi_{x_\ell,s,\mathrm{G}}, \xi_{x_\ell,s,\mathrm{G}})$;
        end
    end
end
Repeat the procedure for all locations $s = 1, \ldots, p$;
```

C.3.1 Accuracies and performance on novel taxa

We further comment on the model performances on FinBOL presented in Section 3.2 of the main paper by providing additional results. Table C.1 reports the accuracies at the species level when the test sequences are from truly novel taxa. The columns in the table under ACCURACY display the accuracies at the species levels. In particular, ALL DATA are the same values as in Table 3 in the main document, while OBSERVED and NEW are the values when the test sequences are from a truly observed taxon or a novel one, respectively. The best performance is again obtained by model M-1 under both the whole set of queries and the novel ones. RDP and M-1, NO NEW and M-2, NO NEW, instead, perform very well on the observed taxa. As mentioned in the main discussion, models K-5 and K-6 have worse performance on every subset, and they fail to predict the correct taxonomic placement of novel sequences. This is consistent across scenarios.

The columns under NEW SPECIES show the number of sequences for which the

Table C.1: Accuracies and prediction of novel taxa at the species level for BayesANT and RDP under the scenarios described in Chapter 4. ALL DATA is the accuracy on all the test query sequences. OBSERVED refers to the subset of test sequences whose true taxa are observed in sample, while NEW is the accuracy on truly novel sequences. The columns under NEW SPECIES report the behaviour of each method when the test sequences are truly new. The number in parenthesis is the actual number of queries with a new taxon at any level; PRED. is the number of sequences that each algorithm predicts as being new; T. POS. is short for "true positives" and reports the number of truly new sequences that are predicted new; F. POS. is "false positives", i.e. the number of sequences that are truly observed but predicted new; F. NEG. is "false negatives", and reports the number of sequences that are truly new but are predicted to be from an observed taxon.

	SCENARIO 1 - RANDOM SPLIT							SCENARIO 2 - STRATIFIED SPLIT						
	ACCURACY			NEW SPECIES (884)				ACCURACY			NEW SPECIES (2672)			
MODEL	ALL DATA	OBSERVED	NEW	PRED.	T. POS.	F. POS.	F. NEG.	ALL DATA	OBSERVED	NEW	PRED.	T. POS.	F. POS.	F. NEG.
M-1	85.2	93.1	31.1	958	746	212	138	70.6	93.7	33.7	2736	2566	170	106
	(.82)	(.87)	(.47)					(.7)	(.85)	(.47)				
M-2	85.4	94.6	22.9	747	649	98	235	69.8	95.1	29.6	2506	2422	84	250
	(.86)	(.91)	(.49)					(.74)	(.91)	(.48)				
K-5	79.2	90.8	0	187	88	99	796	56.8	92.5	0	550	507	43	2165
	(.79)	(.89)	(.15)					(.57)	(.89)	(.05)				
K-6	80.3	92.1	0	333	214	119	670	57.4	93.5	0	1437	1375	62	1297
	(.88)	(.97)	(.32)					(.64)	(.97)	(.12)				
M-1, NO NEW	83.3	95.5	0	0	0	0	884	59.4	96.7	0	0	0	0	2672
	(.92)	(.96)	(.62)					(.78)	(.97)	(.49)				
M-2, NO NEW	83.2	95.4	0	0	0	0	884	59.1	96.2	0	0	0	0	2672
	(.91)	(.96)	(.55)					(.74)	(.97)	(.37)				
RDP	83.1	95.3	0	0	0	0	884	58.9	95.9	0	0	0	0	2672
	(.92)	(.98)	(.47)					(.73)	(.99)	(.31)				

methods predict a new taxon. We specifically include the number of true positives, false positives and false negatives. To be counted as a true positive, for example, the sequence must be predicted new when truly new. This does not imply that the sequence is placed in the correct taxonomic clade, but simply that novelty is recognized. Similar definitions apply to false positives and false negatives. The multinomial aligned kernels predict a high number of new species, and are often able to recognize true novelty. This is not true for the κ-mer counterparts, which are less prone to predicting new sequences. Finally, M-1, NO NEW, M-2, NO NEW and RDP do not allow for novelty by construction, and therefore they always result in false negatives.

Table C.2: Accuracies of BayesANT model M-1 across orders and scenarios. The column N. SEQS. is the number of query sequences that truly belongs to the order, while N. NEW. is the number of truly novel sequences. ALL DATA is the accuracy on all the test query sequences. OBSERVED refers to the subset of test sequences whose true taxa are observed in sample, while NEW is the accuracy on truly novel sequences.

	SCENARIO 1 - RANDOM SPLIT					SCENARIO 2 - STRATIFIED SPLIT				
ORDER	N. SEQS.	N. NEW	ALL DATA	OBSERVED	NEW	N. SEQS.	N. NEW	ALL DATA	OBSERVED	NEW
ARANEAE	309	19	90.9	92.4	68.4	348	75	79.6	91.6	36.0
			(0.79)	(0.81)	(0.46)			(0.74)	(0.85)	(0.33)
COLEOPTERA	1401	169	87.8	95.5	32.0	1442	518	72.3	96.3	29.5
			(0.82	(0.87)	(0.45)			(0.67)	(0.84)	(0.37)
DIPTERA	1470	385	74.8	90.3	30.9	1392	727	56.2	90.2	25.0
			(0.74)	(0.86)	(0.41)			(0.56)	(0.86)	(0.29)
HEMIPTERA	162	37	74.1	88.8	24.3	419	288	46.8	83.2	30.2
			(0.74)	(0.79)	(0.58)			(0.67)	(0.73)	(0.64)
HYMENOPTERA	1208	164	80.0	87.8	30.5	845	282	74.8	90.8	42.9
			(0.86	(0.90)	(0.57)			(0.80)	(0.88)	(0.64)
LEPIDOPTERA	2016	82	93.8	96.5	29.3	1506	231	89.8	96.7	51.9
			(0.85)	(0.87)	(0.38)			(0.78)	(0.86)	(0.31)
OTHERS	359	28	86.6	92.1	21.4	973	551	62.0	93.1	38.1
			(0.82)	(0.83)	(0.77)			(0.76)	(0.83)	(0.70)

C.3.2 Accuracies of BayesANT M-*1 by order*

In this Section, we further report the prediction performance in the test set across orders for BayesANT model M-1. Under both scenarios, *Hemiptera* and *Diptera* are the taxa that appear to be harder to classify, followed by *Coleoptera* and *Hymenoptera*. Unsurprisingly, *Lepidoptera* appears to be the easiest to classify, especially when the true taxon of the test sequence is observed in training. As for novel taxa, sequences in scenario 1 have very similar accuracies across orders, with the positive exception of *Araneae*. In scenario 2, instead *Lepidoptera* and *Hymenoptera* have the highest accuracies on novel taxa.

C.3.3 Computational times

In this Section, we report the elapsed times for each model presented in the main paper. Operations were performed on an AMD Ryzen 3900-based dedicated server with 128GB of memory on Ubuntu 20.04, R version 4.1.1, linked to Intel MKL 2019.5-075. We split the prediction of all queries in the test data across 23 threads with

Table C.3: Computational times in minutes for BayesANT under different choices of kernel and for the RDP classifier.

	SCENARIO 1		SCENARIO 2	
MODEL	TRAINING	PREDICTION	TRAINING	PREDICTION
M-1	1.68	1.37	1.40	1.23
M-2	2.84	1.41	2.48	1.24
K-5	0.51	5.22	0.45	4.56
K-6	0.77	21.59	0.72	19.23
M-1, NO NEW	1.01	0.86	0.90	0.82
M-2, NO NEW	1.99	0.89	1.86	0.83
RDP	2.70	1.67	2.48	1.55

the R packages foreach (Microsoft and Weston, 2022) and doParallel (Corporation and Weston, 2022). The BayesANT source code is written in R and the functions to perform the prediction are built-in and Rcpp (Eddelbuettel and Balamuta, 2018) and RcppArmadillo (Eddelbuettel and Sanderson, 2014). The RDP classifier has a source code in java, and we call it from R through the use of ad-hoc functions inspired by the package rRDP (Hahsler and Nagar, 2021). Table C.3 reports the elapsed times for each method when training the sequences in scenarios 1 and 2. Notice that all BayesANT models except the K-5 and K-6 do not account for the time required to align the data. Times refer to a training library of 27,699 sequences whose taxonomy is across seven levels. The number of nodes in the last level is 10,244 in scenario 1 and 9,490 in scenario 2, while the number of test sequences is 6,925. We notice that the κ-mer models are considerably faster in training, but their prediction requires more time than their aligned counterparts. Moreover, setting BayesANT to include new branches in the taxonomy comes with a slightly higher cost in terms of prediction and the computational time required for training.

C.4 Assessing the impact of alignment gaps "-" in the sequences

In this Section, we show how alignment gaps "-" affect BayesANT predictions, both in terms of general accuracy and the rate of discovery of missing taxa. We first evaluate the effect of alignment gaps on randomly generated synthetic reference libraries of DNA sequences, and we later report on their impact on the prediction in the FINBOL library.

C.4.1 Simulation strategy for DNA barcoding libraries

We simulate synthetic DNA barcoding libraries by relying on two main components: a hierarchical taxonomy and a mutation process for the DNA sequences. We generate the first one via the level- and node-specific Pitman–Yor process described by the urn scheme in equation (C.3), fixing one α_ℓ and one σ_ℓ for each rank. In particular, we start by drawing one random sample of size $n = 10,000$ from a Pitman–Yor with parameters α_1 and σ_1 for the first level. This produces a sequence of distinct clusters $X^*_{1,1}, \dots, V^*_{K_n(v_0),1}$ with frequencies $n_{1,1}, \dots, n_{K_n(v_0),1}$. Then, for each taxon $V^*_{j,1}$ we draw another random Pitman–Yor sample of size $n_{j,1}$ with parameters α_2, σ_2 to generate the nodes in the second taxonomic level. We repeat the procedure until the desired rank L, which we set equal to 4. The advantage of this generative scheme is that it can easily control both the size of the branches and the overall number of leaves. For example, low values for α_ℓ and σ_ℓ generate few clusters with high counts, while high values tend to create many branches with few or a single observation. With this strategy, we generate three libraries of different sizes. Table C.4 reports the simulation parameters and their associated descriptive statistics. Library 1 has a low number of leaves with many reference sequences associated with them on average. Library 2 shows a higher ramification, with 5 DNA sequences associated with its leaves on average. Finally, Library 3 has a large number of leaves, resulting in only one or two reference sequences per leaf.

Table C.4: Parameters and descriptive statistics for the simulated libraries. N. TAXA refers to the number of distinct taxa in the library at a given level. AVG. SEQ. refers to the average number of reference sequences per taxa (generated taxa counts).

	Library 1				Library 2				Library 3			
LEVEL	α	σ	N. TAXA	AVG. SEQ.	α	σ	N. TAXA	AVG. SEQ.	α	σ	N. TAXA	AVG. SEQ.
1	1	0	11	909.1	1	0	9	1111.1	2.5	0	25	400
2	1	0	46	217.4	1	0.1	67	149.3	2.5	0.1	246	40.7
3	1	0	142	70.4	1	0.25	421	23.8	2.5	0.25	1255	8.0
4	1	0	374	26.7	1	0.4	1816	5.5	2.5	0.4	3744	2.7

Once the taxonomic counts are generated, we populate the tree with aligned DNA sequences through recursive mutations of an ancestral root sequence of $p = 300$ base pairs generated randomly. These mutations, ideally, should maintain a coherent DNA sequence structure such that within-taxon similarity is higher than the between-taxa ones for each rank. We obtain this by mutating sequences via an evolution model based on coalescent processes. Specifically, let $\mathbf{X}_{x_\ell} = (X_{x_\ell,j})_{j=1}^p$ be the "representative" ancestral sequence associated with a node x_ℓ at level ℓ. These are never observed in real datasets like FinBOL and are not modelled by our framework, but we introduce them for the purpose of this simulation. Then, the jth nucleotide of an ancestral sequence $\mathbf{X}_{x_{\ell+1}}$ where $\mathrm{pa}(x_{\ell+1}) = x_\ell$ mutates with probability

$$p(X_{x_{\ell+1},j} = g \mid X_{v_\ell,j} = g', t) = \begin{cases} \frac{1}{4} + \frac{3}{4}e^{-4\lambda t} & \text{if} \quad g = g', \\ \frac{1}{4} - \frac{1}{4}e^{-4\lambda t} & \text{if} \quad g \neq g', \end{cases} \tag{C.14}$$

with $g, g' \in \mathcal{N}_1 = \{\mathrm{A,C,G,T}\}$ and where λ is an instantaneous rate of substitution and t is known as coalescent time. The mutation process specified by equation (C.14) commonly known as Jukes-Cantor evolution model (Jukes and Cantor, 1969), as λ is equal for transitions - purine (A,G) ↔ purine (A,G) and pyrimidine (C,T) ↔ pyrimidine (C,T) - and transversions - purine (A,G) ↔ pyrimidine (C,T). While such an evolutionary model relies on overly simplistic assumptions, having a more complex mutation process would go beyond the purpose of our study. The steps to create the taxonomy are the following. First, we randomly generate one ancestral sequence of length $p = 300$, assigning equal probability to each nucleotide A, C, G, and T. We treat the first 25 loci as a conserved region, fixing them across the whole library. Then, we generate a coalescent process of size $K_n(v_0)$, where $K_n(v_0)$ is the number of nodes at the first rank of the taxonomic tree generated via Pitman–Yor. The process returns the mutation times $t_1, \ldots, t_{K_n(v_0)}$, which determine the mutation probabilities for the nucleotides in the ancestral sequence beyond the conserved region as in (C.14). This creates $K_n(v_0)$ new ancestral sequences, one for each taxon at the first level. Then, we repeat the process for each generic node v until the lowest level. To vary sequence similarities, we specify a different rate of mutation λ_ℓ for each level. Values for λ_ℓ close to 0 lead to very similar sequences and low mutations, while higher values lead to rapid differentiation. The advantage of this simulation strategy is that it makes certain taxa "share" a mutation history more than others. For our simulation, we chose $(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (.1, .1, .05, .02)$. Finally, we further mutate the sequences in the leaf nodes by perturbing all the loci outside the conserved region with probability 0.01 and selecting another nucleotide at random with equal probability. Coalescent times for each node and level are generated using the function `rcoal()` in the `R` software packages `ape` (Paradis and Schliep, 2019), while mutations are obtained through the function `simSeq` of the package `phangorn` (Schliep, 2011). Figure C.1 displays the sequence similarities and the cluster distribution of the three libraries. Unlike the FinBOL data, these libraries have a much clearer cluster separation, which helps in improving the classification performance.

C.4.2 Prediction performance under simulated gaps

Alignment gaps can be indirectly regarded as a measure of the overall "quality" of the DNA sequences. In the software `BayesANT`, we treat them as missing values.

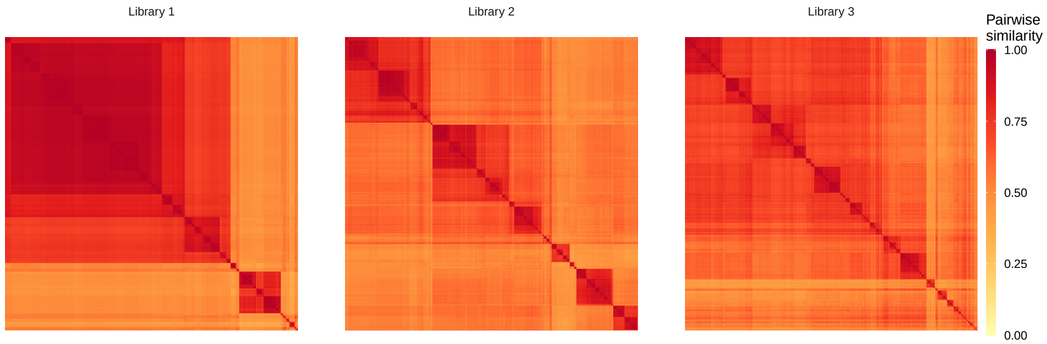

Figure C.1: Pairwise DNA similarities in each of the three generated libraries.

This means we ignore the likelihood contribution of locations with gaps when calculating the probabilities in equation (C.10). This makes the posterior prediction probabilities closer to the prior ones. In the most extreme case of a sequence composed solely of alignment gaps and no other nucleotide, BayesANT will return the branch with the leaf having the highest frequency (i.e. the highest number of DNA sequences linked to it). This property is inherited by the Pitman–Yor process, as easily seen from equation C.1. We verify this behaviour by running a simulation on the libraries generated in the previous section. As a first step, we randomly replace each nucleotide in every sequence in the libraries with a gap with probability r. If $r = 0$, no alignment gap is inserted, whereas the case when $r = 0.99$ leads to a dataset where all sequences are almost entirely made of gaps. Then, we train BayesANT with an aligned single-multinomial kernel on 5,000 randomly selected sequences, and we predict the taxonomic affiliation of the remaining 5,000. We repeat this process for values of $r \in \{0, 0.1, 0.2, \dots, 0.9, 0.99\}$. Figure C.2 depicts the test prediction accuracy at each taxonomic rank for the three libraries at varying values of r. We notice the following. First, Library 1 has the highest prediction accuracy at the lowest level when compared to the other two. This is expected since it features a larger average number of sequences per leaf. Second, the predictive performance of BayesANT deteriorates uniformly across levels and libraries, as less information is available in the training and test data due to the increasing rate of missing values. Third, when $q = 0.99$, the accuracy never quite reaches zero, and sometimes it is still high for higher levels. This happens because BayesANT predicts the taxon that appears with the highest frequency for each query test sequence. The composition of the train and the test data is approximately similar due to the random splitting mechanism, so a small fraction of sequences for which the classification will be correct. For example, in Library 2, the accuracy at Level 1 is roughly 0.5 since 1/2 of the sequences in the test set are associated with the most frequent cluster in the training set. When we look at Level 4, we notice that the accuracy reaches values close to zero due to the high number of leaves, which leads to low prior probabilities.

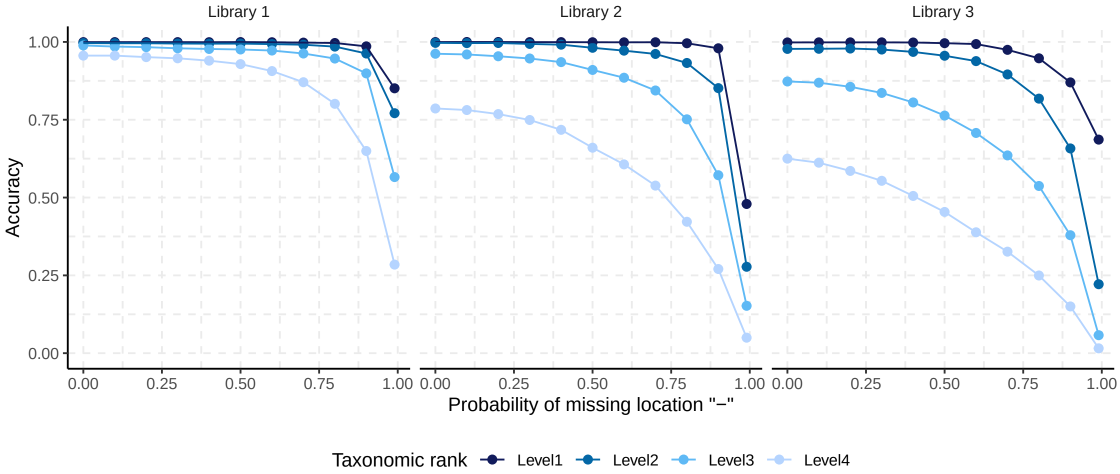

FIGURE C.2: Prediction accuracy at every taxonomic rank in each simulated library for varying missingness probability.

Figure C.3 reports the behavior of BayesANT under the simulation scenarios described above when the taxonomic affiliation of the test sequences is not observed in the training set. The true number of sequences having novel taxa is 103 in library 1, 712 in library 2 and 1673 in library 3. When there is complete absence of missing gaps, the number of sequences predicted to be new are 200, 467 and 1178, respectively. We notice the same consistent behaviour across three libraries: high missingness rates lead to fewer and fewer sequences to be predicted as novel. This is again a property of the Pitman–Yor prior, which tends to assign higher probabilities to observed taxa for small-to-moderate values for α and σ. Structurally, BayesANT will rarely label as novel any test sequence having a large number of alignment gaps.

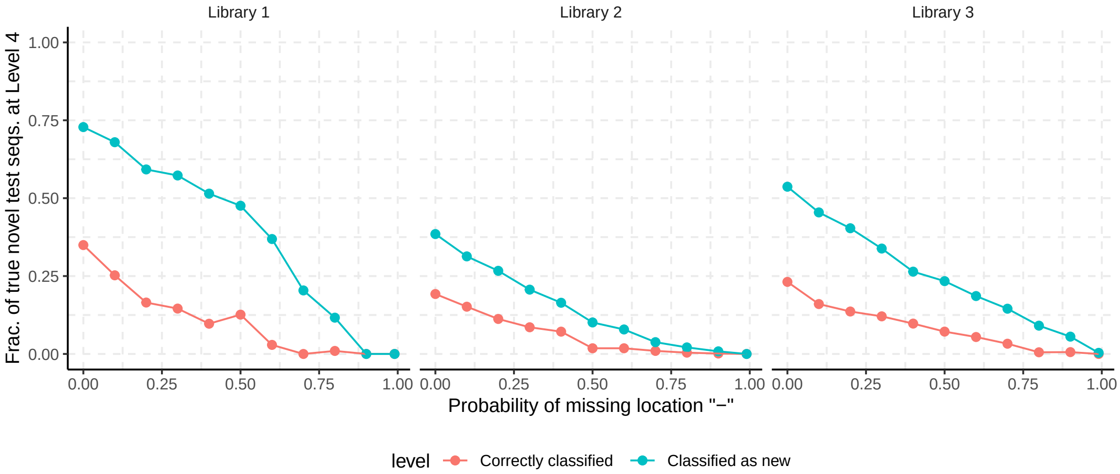

FIGURE C.3: Prediction of the novel taxonomic leaves for varying missingness probability. The red points indicate the accuracy of the algorithm, while the blue one indicates the percentage of sequences recognized as new (but not necessarily in the correct branch) as a fraction of sequences labeled with taxa that are not observed in the training set.

C.4.3 Alignment gaps "-" in FINBOL

As detailed in the main document, the FinBOL library (Roslin et al., 2022) is a dataset of DNA barcodes that have been carefully annotated through morphological inspection. The version we consider features 34,624 globally aligned sequences of length equal to 658 base pairs. This alignment process results in 15,833 (45.7%) sequences containing at least one gap "-", with median value of 5 and a maximum of 174. Notice that every sequence showed a gap in the first location, which we have thus excluded from this count. Overall, missing values are rather infrequent in the library due to the high quality of its barcodes. We expect them to have a relatively low impact on the prediction outcome. Figure C.4 displays the distribution of the prediction probabilities of the M-1 model as a function of the number of gaps in the test sequences (excluding the first location). The red boxplots refer to queries where BayesANT is wrong, while the ones in blue are those that are correctly predicted. The highest number of gaps detected in the test set is 160 in both cases. For a description of scenarios 1 and 2, refer to the main document. When BayesANT is correct, it is easy to see that the species probabilities roughly have the same distribution across scenarios as the number of gaps increases. When BayesANT is wrong, instead, the median probability increases mildly when the counts are 1-4, 5-25 and 26-50 in Scenario 1 and 1-4, 5-25 and 26-50 and 51-100 in Scenario 2, but decreases again for the remaining categories. This suggests that gaps in FinBOL have a minor impact on the prediction probabilities and that fluctuations in the distributions are due to decreasing sample sizes and choices in the partition of the gap categories. Finally, we can see that the accuracy fluctuates as the missingness increase: in Scenario 1, the fraction of correctly predicted sequences under 0 gaps is $3188/(616 + 3188) \approx 0.84$, and becomes approximately 0.79 when gaps are 1-4, 0.85 under 5-25, 0.66 for 26-50, 0.78 for 51-100 and 0.78 again for 101-161. Similar trends can be seen in Scenario 2, where the accuracy values are, respectively, 0.72, 0.70, 0.70, 0.64, 0.70 and 0.62. This again confirms that there is no clear decreasing trend in accuracy.

C.5 Assessing the effect of the size of the training library

In this Section, we evaluate the performance of BayesANT and RDP on FinBOL by varying the size of the training dataset in the two training-test splitting scenarios described in the main paper. As the training library becomes smaller, we expect to

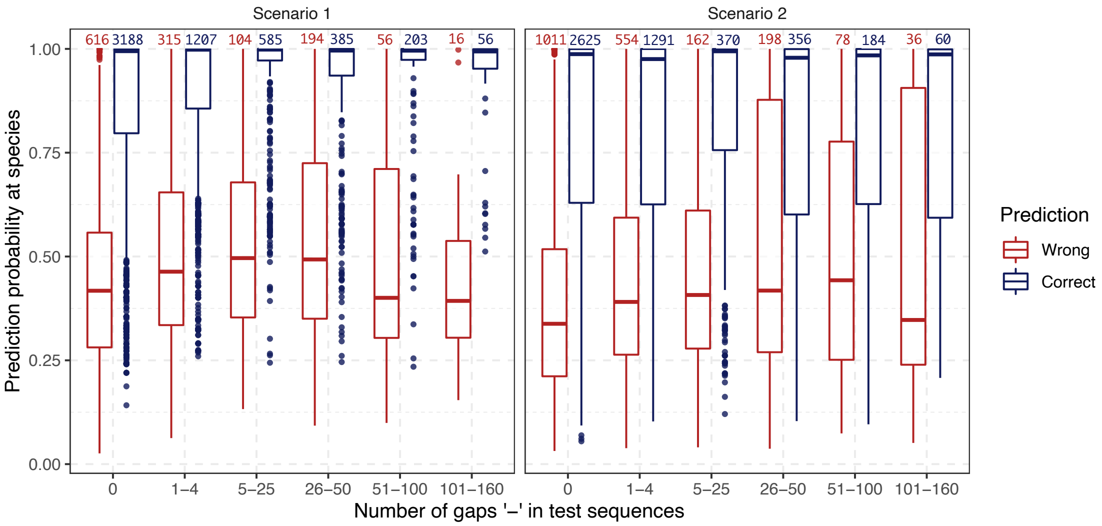

FIGURE C.4: Prediction probabilities at the species level in the test set as a function of the number of alignment gaps detected in the sequences. The number of gaps has been grouped into categories to ease visualization. Red boxplots indicate the sequences where the prediction is wrong, while blue boxplots refer to the cases where BayesANT predicts the true taxonomic branch. The numbers above each boxplot indicate the absolute count of the sequences in each gap counts group across prediction correctness. For example, there are a total of 3804 sequences in the test set with no alignment gap in Scenario 1. In these, BayesANT is correct on 3188 of them and wrong in 616.

see more query sequences whose taxonomic annotation is unobserved. This allows us to test how sensitive BayesANT is with respect to the size of the training library and whether its predictions are still calibrated. When the true taxon is observed, we expect the accuracy to be stable and independent of the size of the training dataset. When the true taxon is new, we expect a similar behaviour as well, but with a lower accuracy due to the inherited difficulty of placing a novel sequence in the correct branch. Figueres C.5, C.6 and C.7 show the results of a simulation where we randomly take a fraction q of the FinBOL library to use as training set, and use the remaining $1-q$ fraction as test. The values for q we choose are $q = \{0.1, 0.2, \dots, 0.9\}$, with $q = 0.1$ representing 10% of the library and $q = 0.8$ reporting the case described in the main paper. We notice the following. In Figure C.5, both the accuracy and

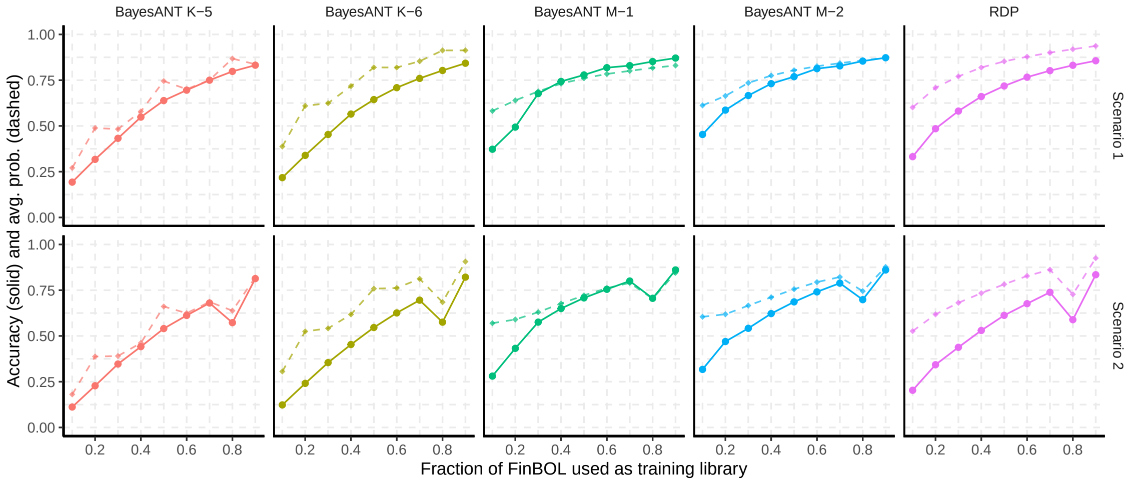

FIGURE C.5: Accuracies (solid lines) and average prediction probabilities (dashed lines) in the test sets for RDP and BayesANT under the splitting scenarios described in the main document. The size of the training library is displayed on the horizontal axis as fractions of the FinBOL library, subsetted at random.

the average prediction probability at the species level increase with the size of the library. This is expected since larger libraries contain more information and a larger number of observed species. In particular, we notice that RDP and models M-1 and M-2 have a similar accuracy for almost all fractions, with the exception of $q = 0.1$ and $q = 0.2$ for M-1. As discussed in the main document, models K-5 and K-6 have lower accuracy than their aligned counterparts consistently across fractions and scenarios. The dashed lines in the plot represent prediction probabilities. The closer the probability path is to the solid line, the more the algorithm is well calibrated. Models M-1 and M-2, whose values for ρ have been selected to 0.1 for the first and 0.06 for the second in both scenarios, appear to have a better calibration than RDP.

Figures C.6 and C.7 display the accuracies. The prediction probabilities under the same setting are described in Figure C.5 for the sequences whose true taxonomic annotations are, respectively, observed and not observed in the training library. We can see that the behaviour of M-1, M-2 and RDP is rather stable across all sizes.

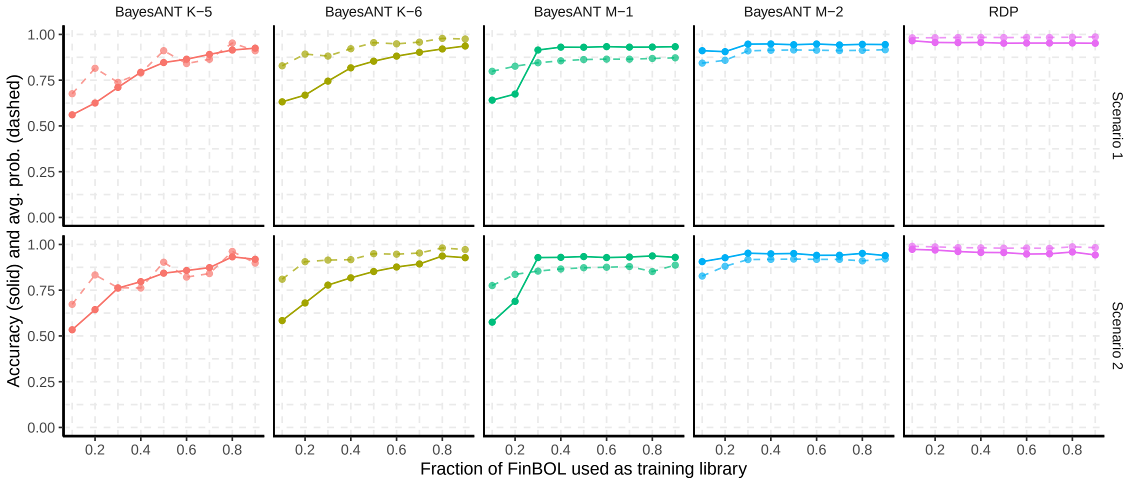

Figure C.6: Accuracies (solid lines) and average prediction probabilities (dashed lines) for the test sequences whose true taxa are observed in training. See Figure C.5.

In particular, when the true taxa are observed, we see that accuracies are always high and prediction probabilities are generally calibrated. When the taxa are new, RDP is structurally always wrong but shows a low prediction probability (around 0.4). As for BayesANT, accuracies are constant for m-1 and m-2, while they are equal to 0 for k-5 and k-6. This suggests that the multinomial kernel over the κ-mer decomposition is unsuitable for novel species prediction under this aligned library.

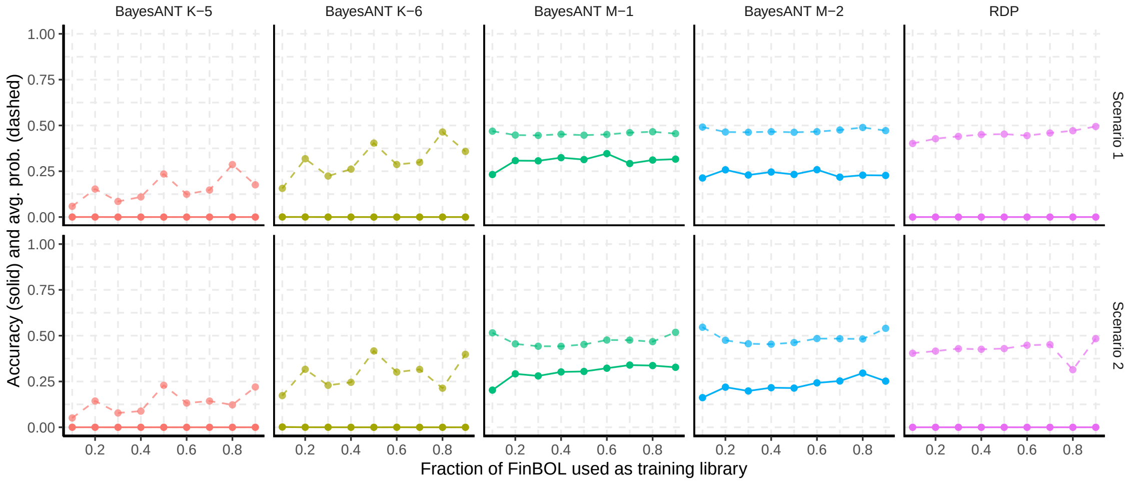

Figure C.7: Accuracies (solid lines) and average prediction probabilities (dashed lines) for the test sequences whose true taxa are not observed in training. See Figure C.5.

Bibliography

Abramowitz, M. and Stegun, I. A. (1972), Handbook of Mathematical Functions with formulas, graphs, and mathematical tables, U.S. Dept. of Commerce, National Bureau of Standards.

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

Airoldi, E. M., Costa, T., Bassetti, F., Leisen, F., and Guindani, M. (2014), "Generalized Species Sampling Priors With Latent Beta Reinforcements," Journal of the American Statistical Association, 109, 1466–1480.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990), "Basic local alignment search tool," Journal of Molecular Biology, 215, 403–410.

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

Arrhenius, O. (1921), "Species and Area," Journal of Ecology, 9, 95–99.

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

Ascolani, F., Lijoi, A., Rebaudo, G., and Zanella, G. (2022), "Clustering consistency with Dirichlet process mixtures," Biometrika, asac051.

Barry, D. and Hartigan, J. A. (1992), "Product Partition Models for Change Point Problems," The Annals of Statistics, 20, 260–279.

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

Battiston, M., Favaro, S., Roy, D. M., and Teh, Y. W. (2018), “A characterization of product-form exchangeable feature probability functions,” The Annals of Applied Probability, 28, 1423 – 1448.

Bazinet, A. L. and Cummings, M. P. (2012), “A comparative evaluation of sequence classification programs,” BMC Bioinformatics, 13, 92.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2012), “GenBank,” Nucleic Acids Research, 41, D36–D42.

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

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

Betancourt, B., Zanella, G., and Steorts, R. C. (2020), “Random Partition Models for Microclustering Tasks,” Journal of the American Statistical Association, 117, 1215–1227.

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

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., Huttley, G. A., and Gregory Caporaso, J. (2018), “Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2’s q2-feature-classifier plugin,” Microbiome, 6, 90.

Booth, J. G., Casella, G., and Hobert, J. P. (2008), “Clustering Using Objective Functions and Stochastic Search,” Journal of the Royal Statistical Society. Series B (Statistical Methodology), 70, 119–139.

Botev, Z. I. (2017), “The normal law under linear restrictions: simulation and estimation via minimax tilting,” Journal of the Royal Statistical Society, Series B, 79, 125–148.

Bunge, J. and Fitzpatrick, M. (1993), “Estimating the Number of Species: a Review,” Journal of the American Statistical Association, 88, 364–373.

Callahan, B. J., McMurdie, P. J., and Holmes, S. P. (2017), “Exact sequence variants should replace operational taxonomic units in marker-gene data analysis,” The ISME Journal, 11, 2639–2643.

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

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

Camerlenghi, F., Favaro, S., Masoero, L., and Broderick, T. (2022), “Scaled Process Priors for Bayesian Nonparametric Estimation of the Unseen Genetic Variation,” Journal of the American Statistical Association, 0, 1–12.

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

Chao, A. (1984), “Nonparametric Estimation of the Number of Classes in a Population,” Scandinavian Journal of Statistics, 11, 265–270.

Chao, A. and Shen, T.-J. (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, 534–545.

Christen, J. A. and Nakamura, M. (2000), “On the Analysis of Accumulation Curves,” Biometrics, 56, 748–754.

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

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

Corporation, M. and Weston, S. (2022), doParallel: Foreach Parallel Adaptor for the ‘parallel’ Package, R package version 1.0.17.

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

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.

Devroye, L. (1986a), Non-Uniform Random Variate Generation, Springer-Verlag, New York, NY, USA.

Devroye, L. (1986b), Non-Uniform Random Variate Generation, Springer, New York, NY.

Diaconis, P. and Ylvisaker, D. (1979), "Conjugate Priors for Exponential Families," The Annals of Statistics, 7, 269 – 281.

Diaz-Frances, E. and Gorostiza, L. G. (2002), "Inference and Model Comparison for Species Accumulation Functions Using Approximating Pure Birth Processes," Journal of Agricultural, Biological, and Environmental Statistics, 7, 335–349.

Dunson, D. B. and Xing, C. (2009), "Nonparametric Bayes Modeling of Multivariate Categorical Data," Journal of the American Statistical Association, 104, 1042–1051.

Durante, D., Dunson, D. B., and Vogelstein, J. T. (2017), "Nonparametric Bayes Modeling of Populations of Networks," Journal of the American Statistical Association, 112, 1516–1530.

Eddelbuettel, D. and Balamuta, J. J. (2018), "Extending *R* with *C++*: A Brief Introduction to *Rcpp*," The American Statistician, 72, 28–36.

Eddelbuettel, D. and Sanderson, C. (2014), "RcppArmadillo: Accelerating R with high-performance C++ linear algebra," Computational Statistics and Data Analysis, 71, 1054–1063.

Eddy, S. R. (1995), "Multiple alignment using hidden Markov models," Proceedings. International Conference on Intelligent Systems for Molecular Biology, 3, 114–120.

Edgar, R. C. (2013), "UPARSE: highly accurate OTU sequences from microbial amplicon reads," Nature Methods, 10, 996–998.

Edgar, R. C. (2018), "Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences," PeerJ, 6, e4652.

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

Escobar, M. D. (1994), "Estimating Normal Means with a Dirichlet Process Prior," Journal of the American Statistical Association, 89, 268–277.

Escobar, M. D. and West, M. (1995), "Bayesian Density Estimation and Inference Using Mixtures," Journal of the American Statistical Association, 90, 577–588.

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

Favaro, S., Lijoi, A., and Prünster, I. (2012), "A New Estimator of the Discovery Probability," Biometrics, 68, 1188–1196.

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

FinBIF (2020), The FinBIF checklist of Finnish species 2019., Finnish Biodiversity Information Facility, Finnish Museum of Natural History, University of Helsinki. Retrieved from `http://urn.fi/URN:ISSN:2490-0907`.

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.

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

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

Fruchterman, T. M. J. and Reingold, E. M. (1991), "Graph drawing by force-directed placement," Software: Practice and Experience, 21, 1129–1164.

Gao, Z., Tseng, C.-H., Pei, Z., and Blaser, M. (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.

Geng, J., Bhattacharya, A., and Pati, D. (2019), "Probabilistic Community Detection With Unknown Number of Communities," Journal of the American Statistical Association, 114, 893–905.

Ghosal, S. and van der Vaart, A. (2017), Fundamentals of Nonparametric Bayesian Inference, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

Gleser, L. J. (1975), "On the Distribution of the Number of Successes in Independent Trials," The Annals of Probability, 3, 182–188.

Gnedin, A. and Pitman, J. (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, 1672–1689.

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

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.

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

Gradshteyn, I. S. and Ryzhik, I. M. (2007), Table of integrals, series, and products, Elsevier/Academic Press, Amsterdam, seventh edn., Translated from the Russian, Translation edited and with a preface by Alan Jeffrey and Daniel Zwillinger, With one CD-ROM (Windows, Macintosh and UNIX).

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

Grünwald, P. and van Ommen, T. (2017), "Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It," Bayesian Analysis, 12, 1069–1103.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017), "On Calibration of Modern Neural Networks," in Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, pp. 1321–1330.

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

Hahsler, M. and Nagar, A. (2021), rRDP: Interface to the RDP Classifier, R package version 1.26.0.

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, 1–15.

Hartigan, J. (1990), "Partition models," Communications in Statistics - Theory and Methods, 19, 2745–2756.

Hebert, P. D. N., Cywinska, A., Ball, S., and deWaard, J. (2003), "Biological identifications through DNA barcodes." Journal of the Royal Statistical society of London B: Biological Sciences, 270, 313–321.

Hoeffding, W. (1956), "On the Distribution of the Number of Successes in Independent Trials," The Annals of Mathematical Statistics, 27, 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., Hellmann, J. J., Ricketts, T. H., and Bohannan, B. J. M. (2001), "Counting the Uncountable: statistical Approaches to Estimating Microbial Diversity," Applied and Environmental Microbiology, 67, 4399–4406.

Huson, D. H., Auch, A. F., Qi, J., and Schuster, S. C. (2007), "MEGAN analysis of metagenomic data," Genome research, 17, 377–386.

Ionita-Laza, I., Lange, C., and M. Laird, N. (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, 5008–5013.

Janzen, D. H., Hajibabaei, M., Burns, J. M., Hallwachs, W., Remigio, E., and Hebert, P. D. N. (2005), "Wedding biodiversity inventory of a large and complex Lepidoptera fauna with DNA barcoding." Philosophical Transactions of the Royal Society B, 360, 1835–1845.

Jukes, T. and Cantor, C. (1969), "Evolution of Protein Molecules." New York: Academic Press, pp. 21–132.

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., and Ueda, N. (2006), "Learning Systems of Concepts with an Infinite Relational Model," in Proceedings of the 21st National Conference on Artificial Intelligence, p. 381–388.

Korwar, R. M. and Hollander, M. (1973), "Contributions to the Theory of Dirichlet Processes," The Annals of Probability, 1, 705–711.

Lan, Y., Wang, Q., Cole, J. R., and Rosen, G. L. (2012), "Using the RDP Classifier to Predict Taxonomic Novelty and Reduce the Search Space for Finding Novel Organisms," PLoS ONE, 7, 1–15.

Lanzén, A., Jørgensen, S. L., Huson, D. H., Gorfer, M., Grindhaug, S. H., Jonassen, I., Øvreås, L., and Urich, T. (2012), "CREST – Classification Resources for Environmental Sequence Tags," PLoS ONE, 7, 1–11.

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

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

Legramanti, S., Rigon, T., Durante, D., and Dunson, D. B. (2022), "Extended stochastic block models with application to criminal networks," The Annals of Applied Statistics, 16, 2369 – 2395.

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

Lijoi, A., Mena, R. H., and Prünster, I. (2007b), "Controlling the reinforcement in Bayesian non-parametric mixture models," Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69, 715–740.

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

Malaise, R. (1937), "A new insect-trap," Entomologisk Tidskrift, 58, 148–160.

Mao, C. X. (2004), "Predicting the Conditional Probability of Discovering a New Class," Journal of the American Statistical Association, 99, 1108–1118.

Masoero, L., Camerlenghi, F., Favaro, S., and Broderick, T. (2021), "More for less: predicting and maximizing genomic variant discovery via Bayesian nonparametrics," Biometrika, 109, 17–32.

Meier, R., Shiyang, K., Vaidya, G., and Ng, P. K. L. (2006), "DNA Barcoding and Taxonomy in Diptera: A Tale of High Intraspecific Variability and Low Identification Success," Systematic Biology, 55, 715–728.

Mersch, D. P., Crespi, A., and Keller, L. (2013), "Tracking Individuals Shows Spatial Fidelity Is a Key Regulator of Ant Social Organization," Science, 340, 1090–1093.

Microsoft and Weston, S. (2022), foreach: Provides Foreach Looping Construct, R package version 1.5.2.

Miller, J. W. and Dunson, D. B. (2019), "Robust Bayesian Inference via Coarsening," Journal of the American Statistical Association, 114, 1113–1125.

Miller, J. W. and Harrison, M. T. (2014), "Inconsistency of Pitman–Yor Process Mixtures for the Number of Components," Journal of Machine Learning Research, 15, 3333–3370.

Miller, J. W. and Harrison, M. T. (2018), "Mixture Models With a Prior on the Number of Components," Journal of the American Statistical Association, 113, 340–356, PMID: 29983475.

Murali, A., Bhargava, A., and Wright, E. S. (2018), "IDTAXA: a novel approach for accurate taxonomic classification of microbiome sequences," Microbiome, 6, 140.

Neal, R. M. (2000), "Markov Chain Sampling Methods for Dirichlet Process Mixture Models," Journal of Computational and Graphical Statistics, 9, 249–265.

Nguyen, N.-P., Mirarab, S., Liu, B., Pop, M., and Warnow, T. (2014), "TIPP: taxonomic identification and phylogenetic profiling," Bioinformatics, 30, 3548–3555.

Nowicki, K. and Snijders, T. A. B. (2001), "Estimation and Prediction for Stochastic Blockstructures," Journal of the American Statistical Association, 96, 1077–1087.

Ovaskainen, O., Abrego, N., Somervuo, P., Palorinne, I., Hardwick, B., Pitkänen, J.-M., Andrew, N. R., Niklaus, P. A., Schmidt, N. M., Seibold, S., Vogt, J., Zakharov, E. V., Hebert, P. D. N., Roslin, T., and Ivanova, N. V. (2020), "Monitoring Fungal Communities With the Global Spore Sampling Project," Frontiers in Ecology and Evolution, 7, 511.

Paradis, E. and Schliep, K. (2019), "ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R," Bioinformatics, 35, 526–528.

Pentinsaari, M., Ratnasingham, S., Miller, S. E., and Hebert, P. D. N. (2020), "BOLD and GenBank revisited – Do identification errors arise in the lab or in the sequence libraries?" PLoS One, 15, 1–10.

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.

Phillips, J. D., Gillis, D. J., and Hanner, R. H. (2019), "Incomplete estimates of genetic diversity within species: Implications for DNA barcoding," Ecology and Evolution, 9, 2996–3010.

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

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

Polson, N., Scott, J., and Windle, J. (2013), "Bayesian inference for logistic models using Pólya-Gamma Latent Variables," Journal of the American Statistical Association, 108, 1339–1349.

Quintana, F. A. and Iglesias, P. L. (2003), “Bayesian clustering and product partition models,” Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65, 557–574.

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

Ratnasingham, S. and Hebert, P. D. N. (2013), “A DNA-based registry for all animal species: the Barcode Index Number (BIN) system.” PLoS ONE, 8, e66213.

Regazzini, E., Lijoi, A., and Prünster, I. (2003), “Distributional results for means of normalized random measures with independent increments,” Annals of Statistics, 31, 560 – 585.

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

Roslin, T., Somervuo, P., Pentinsaari, M., Hebert, P. D. N., Agda, J., Ahlroth, P., Anttonen, P., Aspi, J., Blagoev, G., Blanco, S., Chan, D., Clayhills, T., deWaard, J., deWaard, S., Elliot, T., Elo, R., Haapala, S., Helve, E., Ilmonen, J., ..., and Mutanen, M. (2022), “A molecular-based identification resource for the arthropods of Finland,” Molecular Ecology Resources, 22, 803–822.

Sarkar, I. and Trizna, M. (2011), “The Barcode of Life data portal: bridging the biodiversity informatics divide for DNA barcoding.” PLoS ONE, 6, e14689.

Schliep, K. P. (2011), “phangorn: phylogenetic analysis in R,” Bioinformatics, 27, 592–593.

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

Shokralla, S., Gibson, J., Nikbakht, H., Janzen, D., Hallwachs, W., and Hajibabaei, M. (2014), “Next-generation DNA barcoding: using next-generation sequencing to enhance and accelerate DNA barcode capture from single specimens.” Molecular Ecology Resources, 5, 892–901.

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

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

Somervuo, P., Koskela, S., Pennanen, J., Henrik Nilsson, R., and Ovaskainen, O. (2016), “Unbiased probabilistic taxonomic classification for DNA barcoding,” Bioinformatics, 32, 2920–2927.

Somervuo, P., Yu, D. W., Xu, C. C., Ji, Y., Hultman, J., Wirta, H., and Ovaskainen, O. (2017), "Quantifying uncertainty of taxonomic placement in DNA barcoding and metabarcoding," Methods in Ecology and Evolution, 8, 398–407.

Stork, N. E. (2018), "How Many Species of Insects and Other Terrestrial Arthropods Are There on Earth?" Annual Review of Entomology, 63, 31–45.

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

Virgilio, M., Jordaens, K., Breman, F. C., Backeljau, T., and De Meyer, M. (2012), "Identifying Insects with Incomplete DNA Barcode Libraries, African Fruit Flies (Diptera: Tephritidae) as a Test Case," PLoS ONE, 7, 1–8.

Vu, D., Groenewald, M., and Verkley, G. (2020), "Convolutional neural networks improve fungal classification." Scientific Reports, 10, 12628.

Wade, S. and Ghahramani, Z. (2018), "Bayesian Cluster Analysis: Point Estimation and Credible Balls (with Discussion)," Bayesian Analysis, 13, 559 – 626.

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007), "Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy," Applied and environmental microbiology, 73, 5261–5267.

Weigand, H., Beermann, A. J., Čiampor, F., Costa, F. O., Csabai, Z., Duarte, S., Geiger, M. F., Grabowski, M., Rimet, F., Rulik, B., Strand, M., Szucsich, N., Weigand, A. M., Willassen, E., Wyler, S. A., Bouchez, A., Borja, A., Čiamporová-Zaťovičová, Z., Ferreira, S., Dijkstra, K.-D. B., Eisendle, U., Freyhof, J., Gadawski, P., Graf, W., Haegerbaeumer, A., van der Hoorn, B. B., Japoshvili, B., Keresztes, L., Keskin, E., Leese, F., Macher, J. N., Mamos, T., Paz, G., Pešić, V., Pfannkuchen, D. M., Pfannkuchen, M. A., Price, B. W., Rinkevich, B., Teixeira, M. A., Várbíró, G., and Ekrem, T. (2019), "DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and recommendations for future work," Science of The Total Environment, 678, 499–524.

Wilkinson, M. J., Szabo, C., Ford, C. S., Yarom, Y., Croxford, A. E., Camp, A., and Gooding, P. (2017), "Replacing Sanger with Next Generation Sequencing to improve coverage and quality of reference DNA barcodes for plants," Scientific Reports, 7, 46040.

Wilson, J.-J., Sing, K.-W., Floyd, R. M., and Hebert, P. D. N. (2017), DNA Barcodes and Insect Biodiversity, chap. 17, pp. 575–592, John Wiley & Sons, Ltd.

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

Yu, D. W., Ji, Y., Emerson, B. C., Wang, X., Ye, C., Yang, C., and Ding, Z. (2012), "Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring," Methods in Ecology and Evolution, 3, 613–623.

Zhu, J.-M. and Luo, Q.-M. (2021), "A novel proof of two partial fraction decompositions," Advances in Difference Equations, p. 274.

Biography

Alessandro Zito attended Bocconi University in Milan, Italy, where he obtained a Bachelor's and a Master's degree in Economics with a final evaluation of 110 cum laude. Towards the end of his Master's degree, during which he received the Bocconi merit award scholarship, he acquired an interest in Statistics. His final thesis on bandit algorithms was completed under the supervision of Daniele Durante and discussed the use of unified skew normal priors on a probit Thompson sampling. After graduation, he interned for a year as a Data Scientist at Generali Assicurazioni. In the fall of 2019, he began his Ph.D. in Statistical Science at Duke University, which was completed in 2023 under the supervision of David B. Dunson.

During his Ph.D., his research focused on applying Bayesian nonparametric species sampling models to ecological data, collaborating with ecologists and scientists from the Lifeplan ERC project. He was awarded the Teaching Assistant of the Year for the 2020/2021 academic year and won the BEST award for student research for his work on Bayesian modeling of sequential discoveries in the 2021/2022 academic year. In the summer of 2022, he worked as a Data Scientist at Google as part of an internship program. After the completion of his Ph.D., he is joining the Department of Biostatistics at Harvard as a postdoctoral research fellow, under the supervision of Jeff Miller.