Ecological modeling via Bayesian nonparametric species sampling priors

Alessandro Zito

Ph.D. dissertation in Statistical Science

Advisor: David B. Dunson

Acknowledgments

Tommaso Rigon

David Dunson

Acknowledgments

The committee


David B. Dunson


Peter Hoff


Mike West


James Clark

Statistical Science department

Lifeplan research group

LIFEPLAN
A PLANETARY INVENTORY OF LIFE

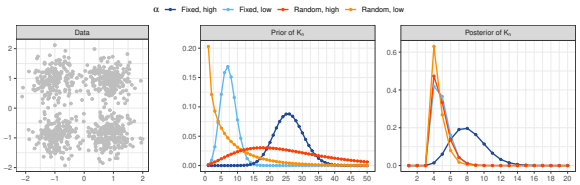
Introduction

- Species sampling models (Pitman, 1996) are class of discrete Bayesian nonparametric priors that model the **sequential appearence** of **distinct tags** in a sequence of **labelled objects**
- The tags are metaphorically called **distinct species**, and can be also interpreted as **clusters**. Thus, very useful to model **species novelty**
- The field dates back to 50 years ago, when Ferguson (1973) introduced the **Dirichlet process**. Since then...
- ... rich theoretical and methodological development in **mixture modeling settings**, such as clustering, density estimation, community detection, species discovery and more
- However, these models have found limited application among **ecologists**, whose primary aim often involves the modeling of *actual* species

Dissertation goal

- Our **goal** is to open a path towards a broader use of species sampling model-based methods, especially in applied ecological settings

Theory

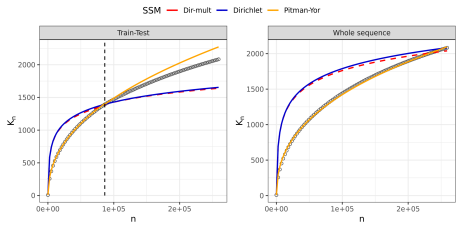


Can we **robustify inference when clustering** via Dirichlet process mixture models?

Bayesian nonparametric modeling of latent partitions via Stirling-gamma priors

R software: `ConjugateDP`

Methods

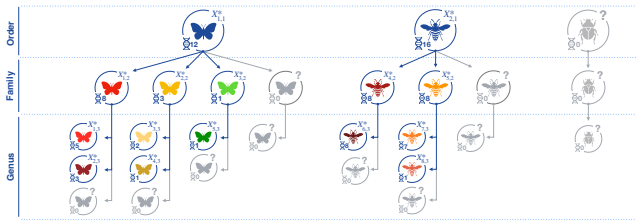


Can we rely on species sampling models to **infer the species richness** in a location?

Bayesian modeling of sequential discoveries

R software: `BNPvegan`

Application



Can species sampling models be helpful in **taxonomic classification** of DNA sequences?

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

R software: `BayesANT`

Overview of species sampling models

Species sampling priors

- A species sampling model is a random probability measure $\tilde{p}$ defined as

$$\tilde{p} = \sum_{j=1}^{\infty} \pi_j \delta_{\theta_j}, \qquad \theta_j \overset{\text{iid}}{\sim} P_0, \quad \sum_{j=1}^{\infty} \pi_j = 1,$$

  where π_j are random weights and θ_j are **atoms** from a (diffuse) baseline distribution P_0

- When some exchangeable random variables $(X_n)_{n \geq 1}$ are from $\tilde{p}$, namely

$$X_1, \ldots, X_n \mid \tilde{p} \overset{\text{iid}}{\sim} \tilde{p}, \quad n \geq 1,$$

  the **discreteness** makes the X_is take on $K_n = k$ **distinct species**, called $X_1^*, \ldots, X_k^*$, with frequencies $n_1, \ldots, n_k$

- Under $\tilde{p}$, the units $\{1, \ldots, n\}$ are partitioned into **clusters** $C_1, \ldots, C_k$, with $C_j = \{i : X_i = X_j^*\}$, and $n_j = |C_j|$

Famous example: the Dirichlet process

- The **Dirichlet process** $\tilde{p} \sim \text{DP}(\alpha P_0)$ with **precision parameter** $\alpha > 0$ is

$$\tilde{p} = \sum_{j=1}^{\infty} \pi_j \delta_{\theta_j}, \qquad \pi_j = v_j \prod_{h=1}^{j-1} (1 - v_h), \quad v_j \overset{\text{iid}}{\sim} \text{Be}(1, \alpha),$$

- The resulting **exchangeble partition probability function** (EPPF) is

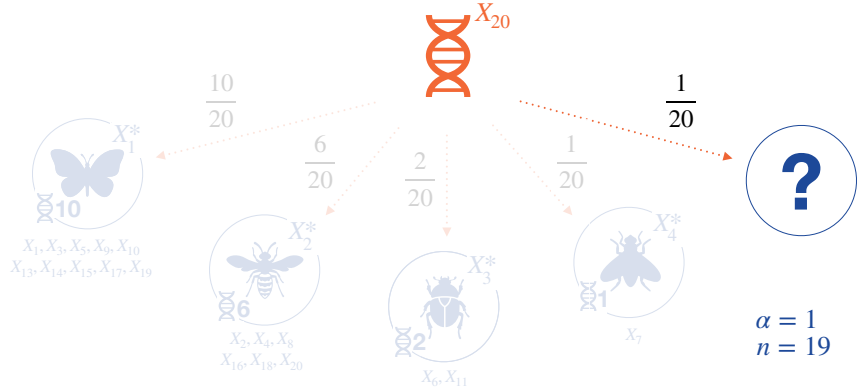
$$\mathbb{P}(\Pi_n = \{C_1, \ldots, C_k\} \mid \alpha) = \frac{\alpha^k}{(\alpha)_n} \prod_{j=1}^{k} (n_j - 1)!$$

where $(\alpha)_n = \Gamma(\alpha + n)/\Gamma(\alpha)$ is the ascending factorial

- The random partition is generated with an **urn scheme**

$$\mathbb{P}(X_{n+1} \in A \mid X_1, \ldots, X_n) =$$

$$= \frac{\alpha}{\alpha + n} P_0(A) + \sum_{j=1}^{k} \frac{n_j}{\alpha + n} \delta_{X_j^*}(A)$$



Famous example: the Dirichlet process

- The **Dirichlet process** $\tilde{p} \sim \text{DP}(\alpha P_0)$ with **precision parameter** $\alpha > 0$ is

$$\tilde{p} = \sum_{j=1}^{\infty} \pi_j \delta_{\theta_j}, \qquad \pi_j = v_j \prod_{h=1}^{j-1}(1 - v_h), \quad v_j \overset{\text{iid}}{\sim} \text{Be}(1, \alpha),$$

- The resulting **exchangeble partition probability function** (EPPF) is

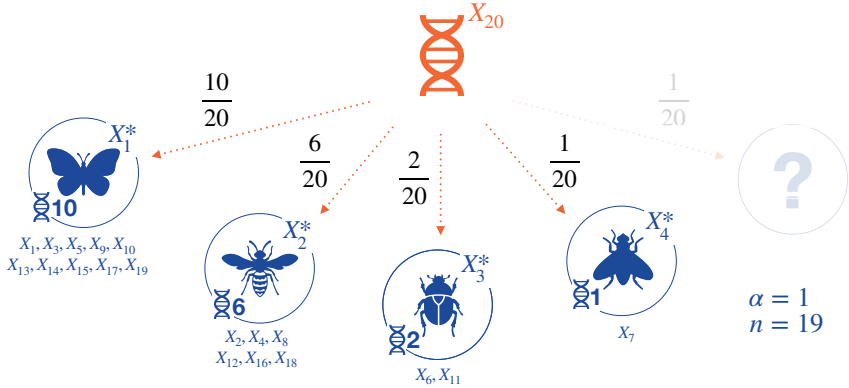
$$\mathbb{P}(\Pi_n = \{C_1, \ldots, C_k\} \mid \alpha) = \frac{\alpha^k}{(\alpha)_n} \prod_{j=1}^{k}(n_j - 1)!$$

where $(\alpha)_n = \Gamma(\alpha + n)/\Gamma(\alpha)$ is the ascending factorial

- The random partition is generated with an **urn scheme**

$$\mathbb{P}(X_{n+1} \in A \mid X_1, \ldots, X_n) =$$

$$= \frac{\alpha}{\alpha + n} P_0(A) + \sum_{j=1}^{k} \frac{n_j}{\alpha + n} \delta_{X_j^*}(A)$$



Beyond the Dirichlet process: Gibbs-type priors

- A **Gibbs-type prior** (Gnedin and Pitman, 2005; De Blasi et al., 2015) is a species sampling model where the EPPF is

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

- The coefficients satisfy the **forward recursion**

$$V_{n,k} = (n - \sigma) V_{n+1,k} + V_{n+1,k+1},$$

for any $k = 1, \ldots, n$ and $n \geq 1$, with $V_{1,1} = 1$

Dirichlet process, $\sigma = 0$

$$V_{n,k} = \frac{\alpha^k}{(\alpha)_n}$$

Pitman–Yor process, $\sigma \in (0,1)$

$$V_{n,k} = \frac{\prod_{i=1}^{k-1} (\alpha + i\sigma)}{(\alpha+1)_{n-1}}$$

Dirichlet-multinomial, $\sigma < 0, H \in \mathbb{N}$

$$V_{n,k} = \frac{|\sigma|^{k-1} \prod_{i=1}^{k-1} (H - i)}{(|\sigma| H + 1)_{n-1}}$$

Sequential predictive rule and number of clusters

- The **predictive rule** for the species of X_{n+1} given a sample $X_1, \ldots, X_n$ under a Gibbs-type process has a simple form:

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

- The distribution of the resulting **number of clusters** is

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

where $\mathscr{C}(n, k; \sigma)$ is the generalized factorial coefficient

Dirichlet process, $\sigma = 0$

$K_n \sim \alpha \log n$

Pitman–Yor process, $\sigma \in (0, 1)$

$K_n \sim n^{\sigma}$

Dirichlet-multinomial,
$\sigma < 0, H \in \mathbb{N}$

$K_n \to H$

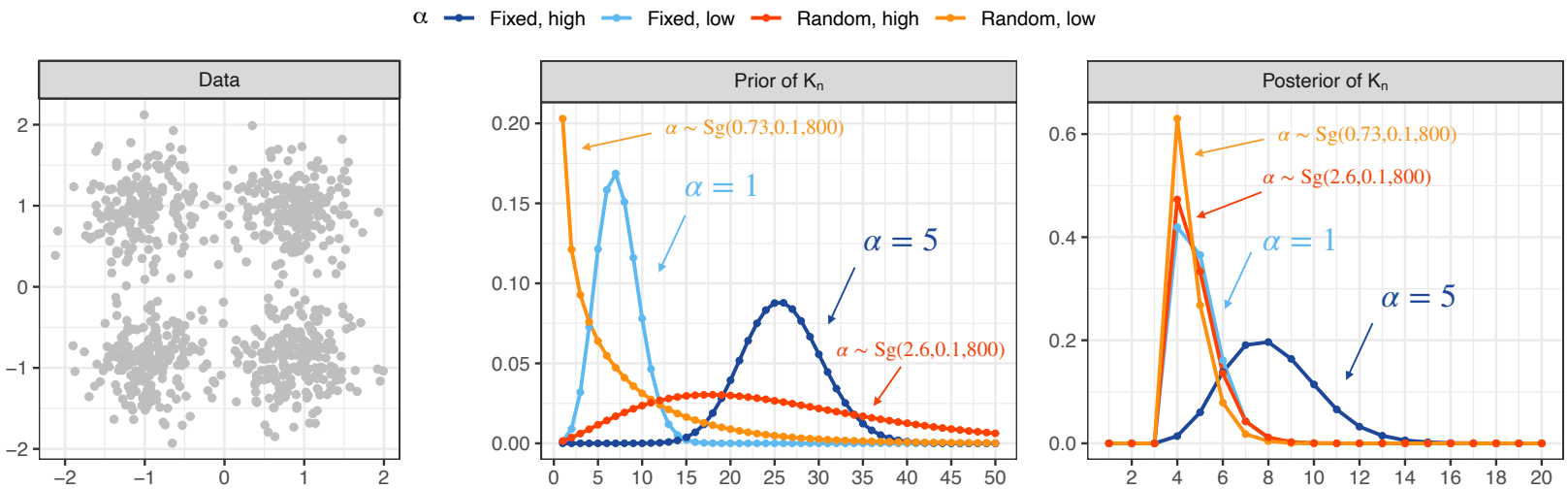
Bayesian nonparametric modeling of latent partitions via Stirling-gamma priors

Dirichlet process mixture models

- A **Dirichlet process mixture** models observations $Y_1, \ldots, Y_n$ as

$$Y_i \mid X_i \overset{\text{ind}}{\sim} f(y \mid X_i), \quad X_i \mid \tilde{p} \overset{\text{iid}}{\sim} \tilde{p}, \quad \tilde{p} \sim \text{DP}(\alpha P_0), \quad (i = 1, \ldots, n)$$

- The discreteness allows us to find K_n **clusters** in the data via ties among $X_1, \ldots, X_n$. However, fixing the precision α is a **highly informative choice**



- Letting $\alpha \sim \pi(\alpha)$ **robustifies the analysis**. However, why is it the case? And what prior should we choose?

The random precision parameter

- When $\alpha \sim \pi(\alpha)$ in a Dirichlet process, we have a Gibbs-type partition with EPPF

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

  All Gibbs-type priors with $\sigma = 0$ have this representation (Gnedin and Pitman, 2005)

- Common choice is the **gamma prior** $\alpha \sim \mathrm{Ga}(a, b)$ as in Escobar and West (1995)

- The induced prior on the number of clusters is

$$\mathbb{P}(K_n = k) = V_{n,k} |s(n, k)|$$

  where $s(n, k)$ are called Stirling-numbers of the first kind, but **does not have an analytic form**

- This **complicates prior elicitation**. For example, $\mathbb{E}(K_n) =$?

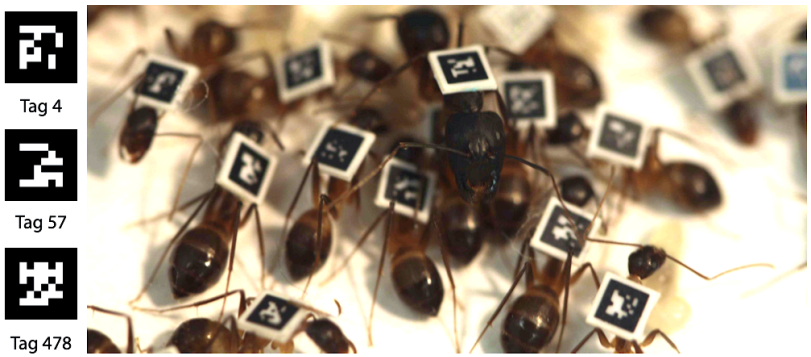
Application: community detection in a colony of worker ants

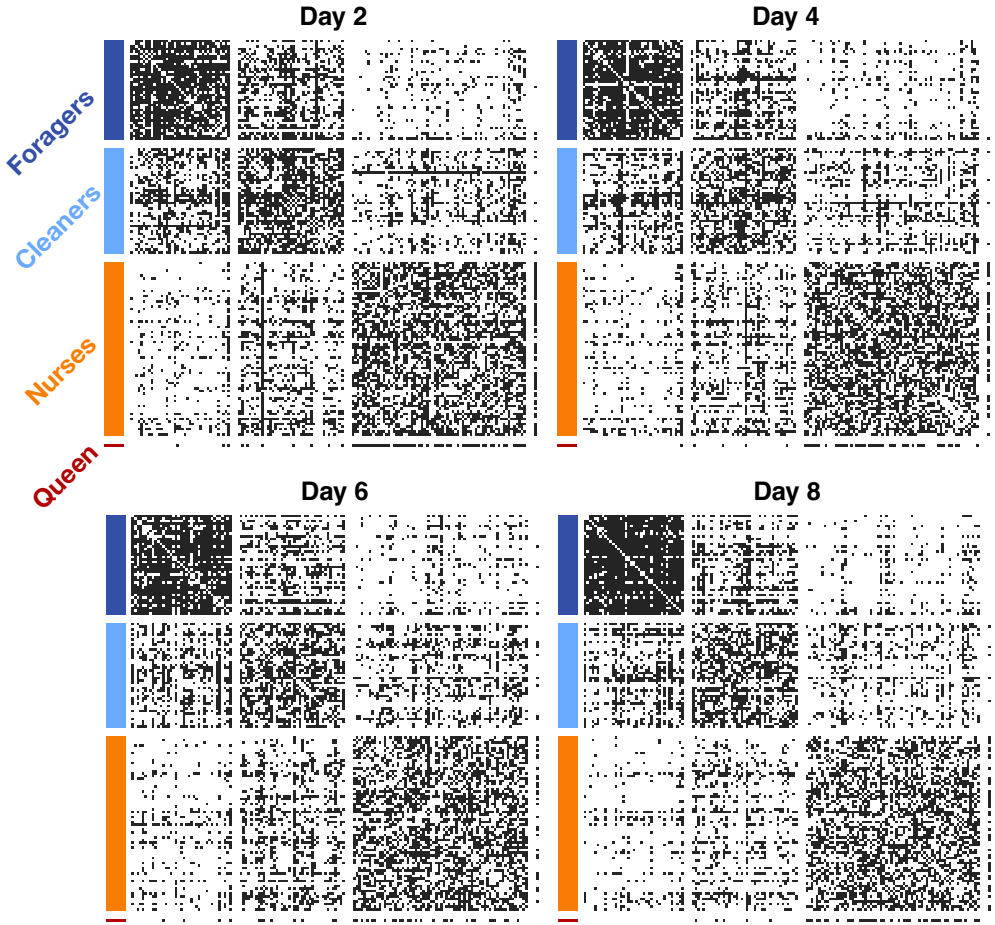
Tracking Individuals Shows Spatial Fidelity Is a Key Regulator of Ant Social Organization

Danielle P. Mersch,[1]* Alessandro Crespi,[2] Laurent Keller[1]*

Ants live in organized societies with a marked division of labor among workers, but little is known about how this division of labor is generated. We used a tracking system to continuously monitor individually tagged workers in six colonies of the ant *Camponotus fellah* over 41 days. Network analyses of more than 9 million interactions revealed three distinct groups that differ in behavioral repertoires. Each group represents a functional behavioral unit with workers moving from one group to the next as they age. The rate of interactions was much higher within groups than between groups. The precise information on spatial and temporal distribution of all individuals allowed us to calculate the expected rates of within- and between-group interactions. These values suggest that the network of interaction within colonies is primarily mediated by age-induced changes in the spatial location of workers.

Mersch et al (2013)





- Three groups of ant workers: foragers, cleaners, and nurses. We want to incorporate this into out model while ensuring robustness

The Stirling-gamma distribution

Definition

A positive random variable follows a Stirling-gamma distribution $\alpha \sim \mathrm{Sg}(a, b, m)$ with parameters $a, b > 0$ and $m \in \mathbb{N}$ satisfying $1 < a/b < m$, if its density function is

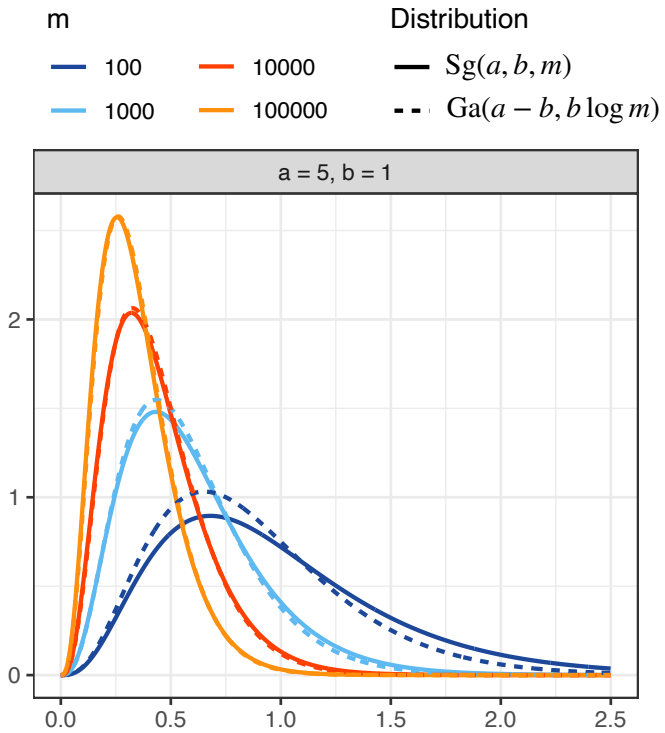
$$p(\alpha) = \frac{1}{\mathcal{S}_{a,b,m}} \frac{\alpha^{a-1}}{\{(\alpha)_m\}^b}, \qquad \mathcal{S}_{a,b,m} = \int_{\mathbb{R}_+} \frac{\alpha^{a-1}}{\{(\alpha)_m\}^b} \, \mathrm{d}\alpha.$$

- Heavy-tailed distribution
- $\mathcal{S}_{a,b,m} < \infty$ for appropriate choice of a, b and m. If these are all integers, $\mathcal{S}_{a,b,m}$ has a closed-form expression

Proposition

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

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



The Stirling-gamma process

- When $\alpha \sim \mathrm{Sg}(a, b, m)$, we have a Stirling-gamma process, whose **Gibbs-type coefficients** are

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

Theorem

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 $X_1, \ldots, X_m$ is distributed as

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

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

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

- Interpretation: a/b is a **location**, b is a **precision** and m is a **reference sample size**
- We can show that $\mathcal{D}_{a,b,m} \approx 1$. This is very useful for elicitation!

Asymptotic behavior of the number of clusters

Theorem

The following convergence in distribution holds for the number of clusters at m:

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

- Notice that m is a fixed quantity. According to Pitman (1996), we still have that $K_n / \log n \to \alpha \sim \text{Sg}(a, b, m)$
- Roughly speaking, the logarithmic convergence to zero of the Stirling-gamma **counterbalances the divergence** of the number of clusters K_n
- In contrast, the **Dirichlet process has a Poisson-type behavior**: letting $\alpha = \lambda / \log m$ for some $\lambda > 0$, then $K_m \to K_\infty$, $K_\infty \sim 1 + \text{Po}(\lambda)$ for $m \to \infty$.
- A random α grants **additional robustness!**

The conjugate Stirling-gamma prior

- A simplification occurs when $m = n$, i.e. when the prior **depends on the sample size**
- The EPPF of a Dirichlet process is an exponential family after writing $\xi = \log \alpha$

$$\mathbb{P}(\Pi_n = \{C_1, \ldots, C_k\} \mid \xi) \propto \exp\{k\xi - \mathcal{K}(\xi, n)\}$$

  where $\mathcal{K}(\xi, n) = \log \Gamma(e^\xi + n) - \log \Gamma(e^\xi)$ is the cumulant generating function

- Diaconis and Ylvisaker (1979): **every exponential family admits a conjugate prior**

Proposition

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

- This follows from a simple Bayesian update

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

Projectivity and the population of partition framework

- The dependency on the sample size is useful, since $\mathbb{E}(K_n) = a/b$. The **Gibbs-type recursion** characterizing the coefficients $V_{n,k}$ **no longer holds**

$$V_{n,k} \neq kV_{n+1,k} + V_{n+1,k+1}$$

- This **breaks the projectivity** of the species sampling model. Problematic when extrapolating from the sample to the general population, but less so when clustering

- **Population of partition framework**: we observe N partitions of the same units $\{1, \ldots, n\}$, namely $\boldsymbol{\Pi}_{n,N} = (\Pi_{n,1}, \ldots, \Pi_{n,N})$, from a Dirichlet process with shared α

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

- If $\alpha \sim \mathrm{Sg}(a, b, n)$, then $(\alpha \mid \boldsymbol{\Pi}_{n,N}) \sim \mathrm{Sg}\left(a + \sum_{s=1}^{N} k_s, b + N, n\right)$

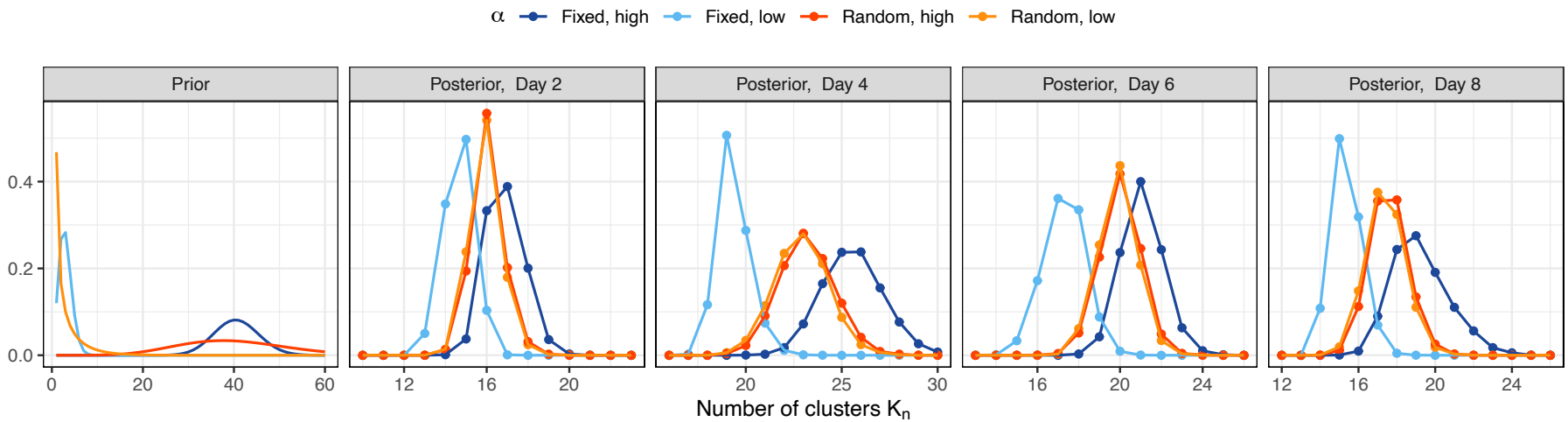
Back to ant community detection

- We detect ant communities in each day via a **stochastic block model** where edge probabilities are

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

with $Z_{i,s} = h$ if node $i \in C_{h,s}$ in network s, whose partition is $\Pi_{n,s} = \{C_{1,s}, \ldots, C_{k_s,s}\}$

- The quantity $\nu_{h,h',s}$ is the edge probability in the block identified by $C_{h,s}$ and $C_{h',s}$



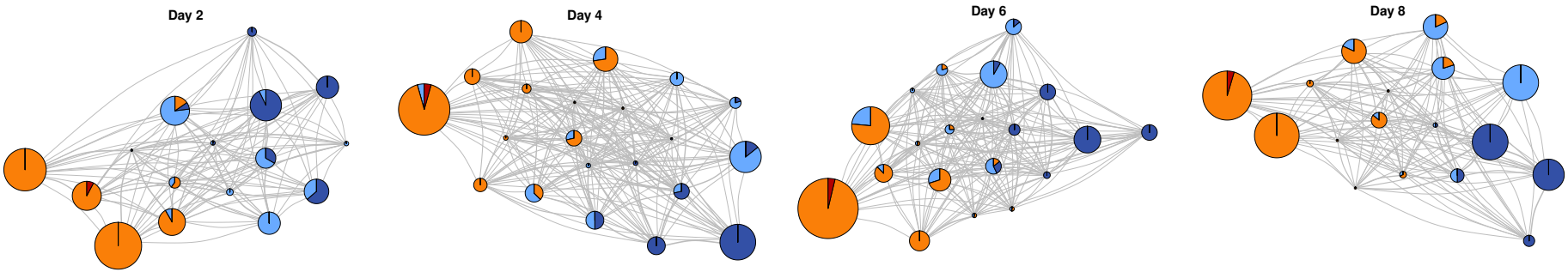
Back to ant community detection

- We detect ant communities in each day via a **stochastic block model** where edge probabilities are

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

with $Z_{i,s} = h$ if node $i \in C_{h,s}$ in network s, whose partition is $\Pi_{n,s} = \{C_{1,s}, \ldots, C_{k_s,s}\}$

- The quantity $\nu_{h,h',s}$ is the edge probability in the block identified by $C_{h,s}$ and $C_{h',s}$



Bayesian modeling of sequential discoveries

Sequential discoveries

- We re-frame the number of new species $(K_n)_{n\geq 1}$, called **accumulation curve**, via some discovery indicators $(D_n)_{n\geq 1}$

$$K_n = \sum_{i=1}^{n} D_i, \quad D_i = \mathbb{1}\{X_i = \text{“new”}\}$$

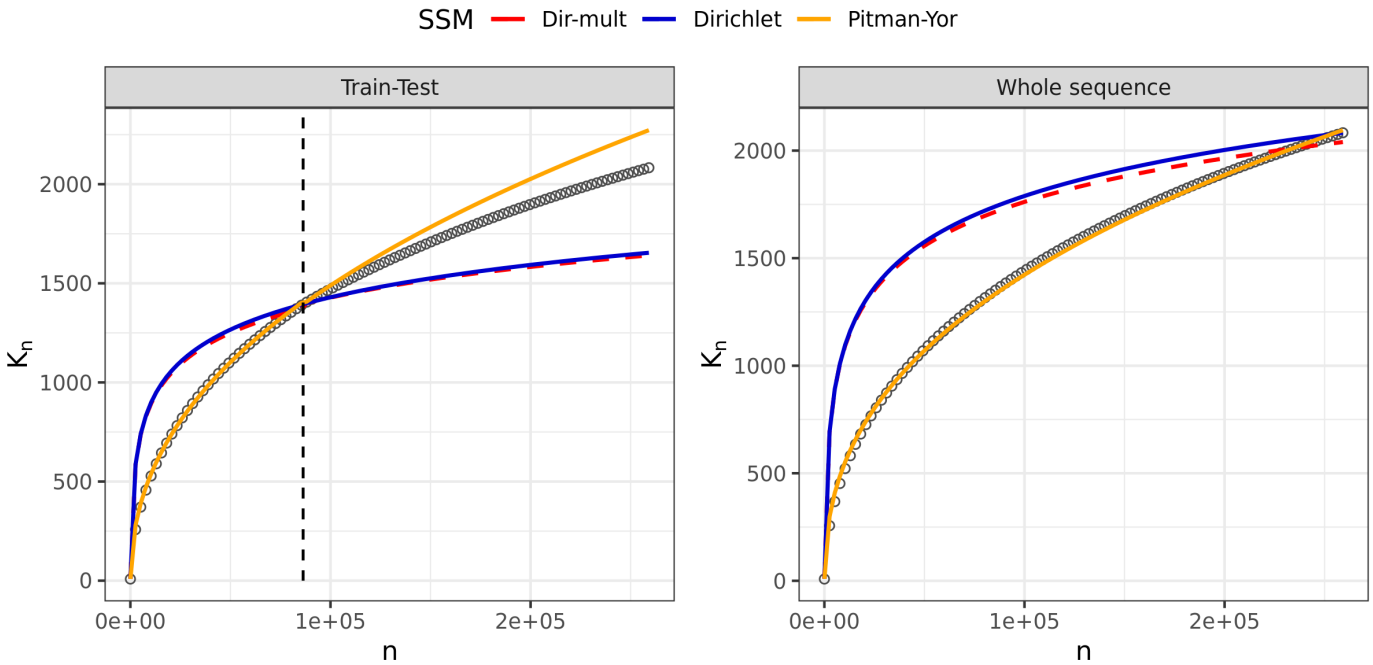
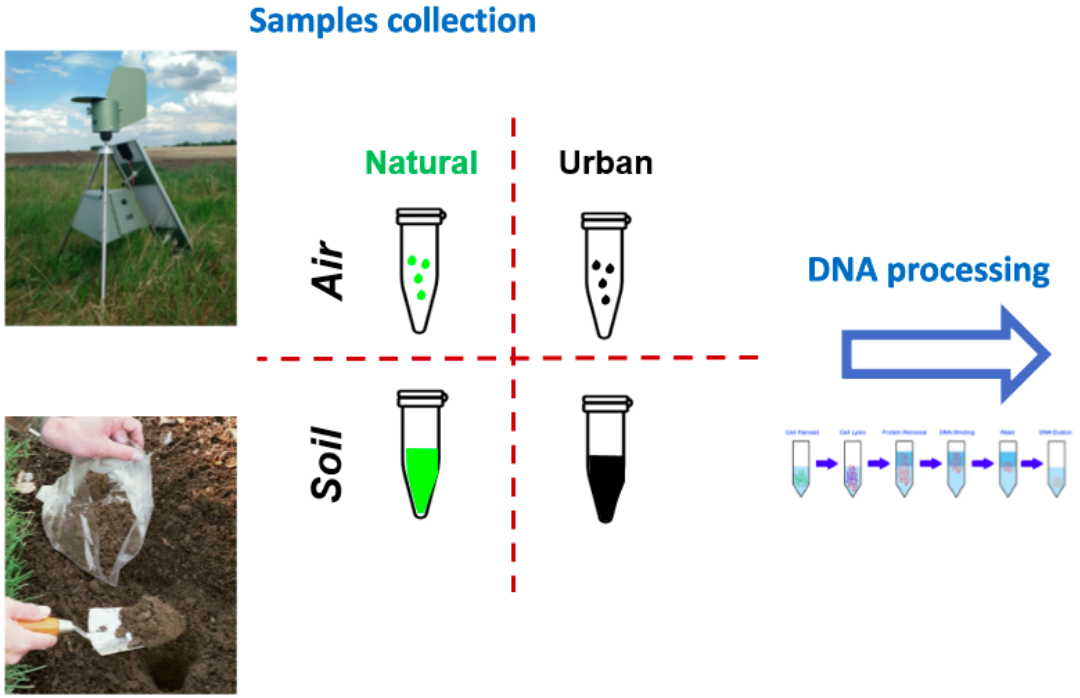
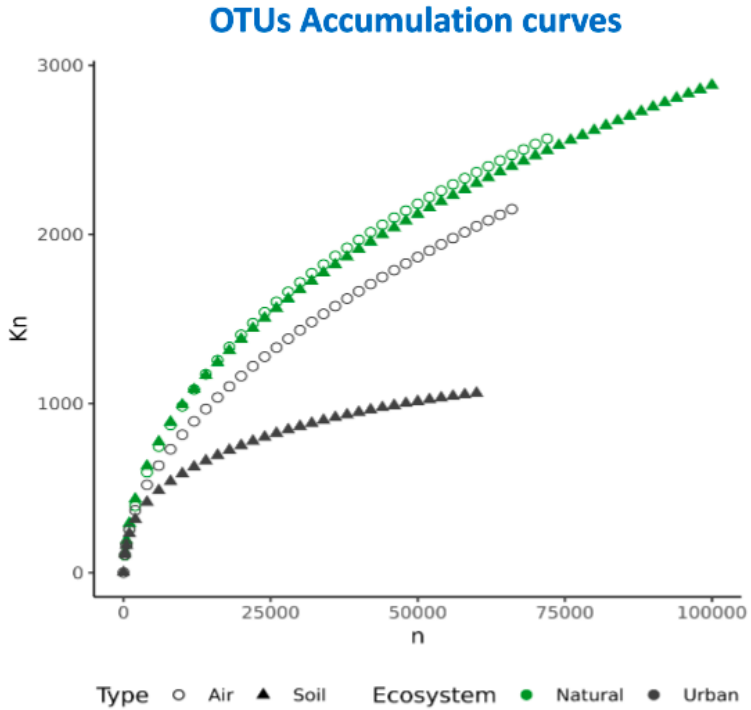


Figure: The classic species sampling models are sometimes not flexible enough to capture the in- and out-of-sample accumulation curve trajectories. Moreover, $K_n \to \infty$ in the Dirichlet and in the Pitman–Yor, but we may want a **finite** species richness $K_\infty < \infty$

Application: fungal biodiversity

- When doing high-throughput sequencing of DNA, the resulting $(X_i)_{i=1}^n$ are called **operational taxonomic units** (OTUs) - a proxy for species based on DNA similarity.





- **How far is each curve from saturation? How much should be keep on sequencing our samples?**

Sequential Discovery Framework

- Let T be a continuous **latent variable** on $(0, \infty)$ with strictly decreasing survival function $S(t; \theta)$ and $\theta \in \Theta \subset \mathbb{R}^p$
- The **discovery probability** at $n \geq 1$ is equal to

$$\pi_n = \mathbb{P}(D_n = 1) = \mathbb{P}(T_n > n - 1) = S(n - 1; \theta)$$

  where $(T_n)_{n \geq 1}$ are iid distributed as T.

- **Discoveries** $(D_i)_{i=1}^n$ are **independent** and K_n is a **Poisson-Binomial**:

$$K_n = \sum_{i=1}^{n} D_i \sim \mathrm{Pb}\{1, S(1; \theta), \ldots, S(n - 1; \theta)\}.$$

- Any T work as long as $\pi_1 = S(0; \theta) = 1$, $S(n - 1; \theta) > S(n; \theta)$ and $S(n; \theta) \to 0$ as $n \to \infty$

Sequential Discovery Framework

- The properties of the Poisson-Binomial allow to naturally fulfill our goals:
  1. The **in-sample trajectory estimator** is the expectation of K_n:

$$\mathbb{E}(K_n) = \sum_{i=1}^{n} S(i-1;\theta).$$

  2. The **out-of-sample estimator** is a posterior expectation:

$$\mathbb{E}(K_{m+n} \mid K_n = k) = k + \sum_{j=1}^{m} S(j+n-1;\theta).$$

  3. The latent variables control the **asymptotic behavior**:

Proposition

Under the latent structure setting, $\mathbb{E}(K_\infty) = \sum_{i=1}^{\infty} S(i-1;\theta)$ is such that

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

with $\mathbb{E}(T) = \int_0^\infty S(t;\theta)dt$. Moreover, $K_\infty = \infty$ almost surely if and only if $\mathbb{E}(T) = \infty$.

The log-logistic model

- Our choice for the shape of T is a **three parameter log-logistic**. If $T_n \overset{\text{iid}}{\sim} \text{LL}(\alpha, \sigma, \phi)$, then

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

with $\alpha > 0$, $\sigma < 1$ and $\phi \in (0, 1]$.

MODEL	PARAMETERS	K_n BEHAVIOR	K_∞	SSM COUNTERPART
LL1	$\sigma = 0, \quad \phi = 1$	$\mathcal{O}(\alpha \log n)$	∞	**Dirichlet**
LL2	$\sigma \in (0, 1), \quad \phi = 1$	$\mathcal{O}(n^\sigma)$	∞	$\approx$ **Pitman–Yor**
LL2	$\sigma < 0, \quad \phi = 1$	K_n converges	$\approx \mathbb{E}(T)$	$\approx$ **Dir-multinomial**
LL3	$\phi < 1$	K_n converges	$\approx \mathbb{E}(T)$	-

- Estimation via **constrained logistic regression** using **truncated normal priors**

$$\log \frac{\pi_{n+1}}{1 - \pi_{n+1}} = \log \alpha - (1 - \sigma) \log n + (\log \phi) n$$

$$(\log \alpha) \sim N(0, 10^2), \quad (\sigma - 1) \sim N_{(-\infty, 0)}(0, 10^2), \quad (\log \phi) \sim N_{(-\infty, 0)}(0, 10^2).$$

Results

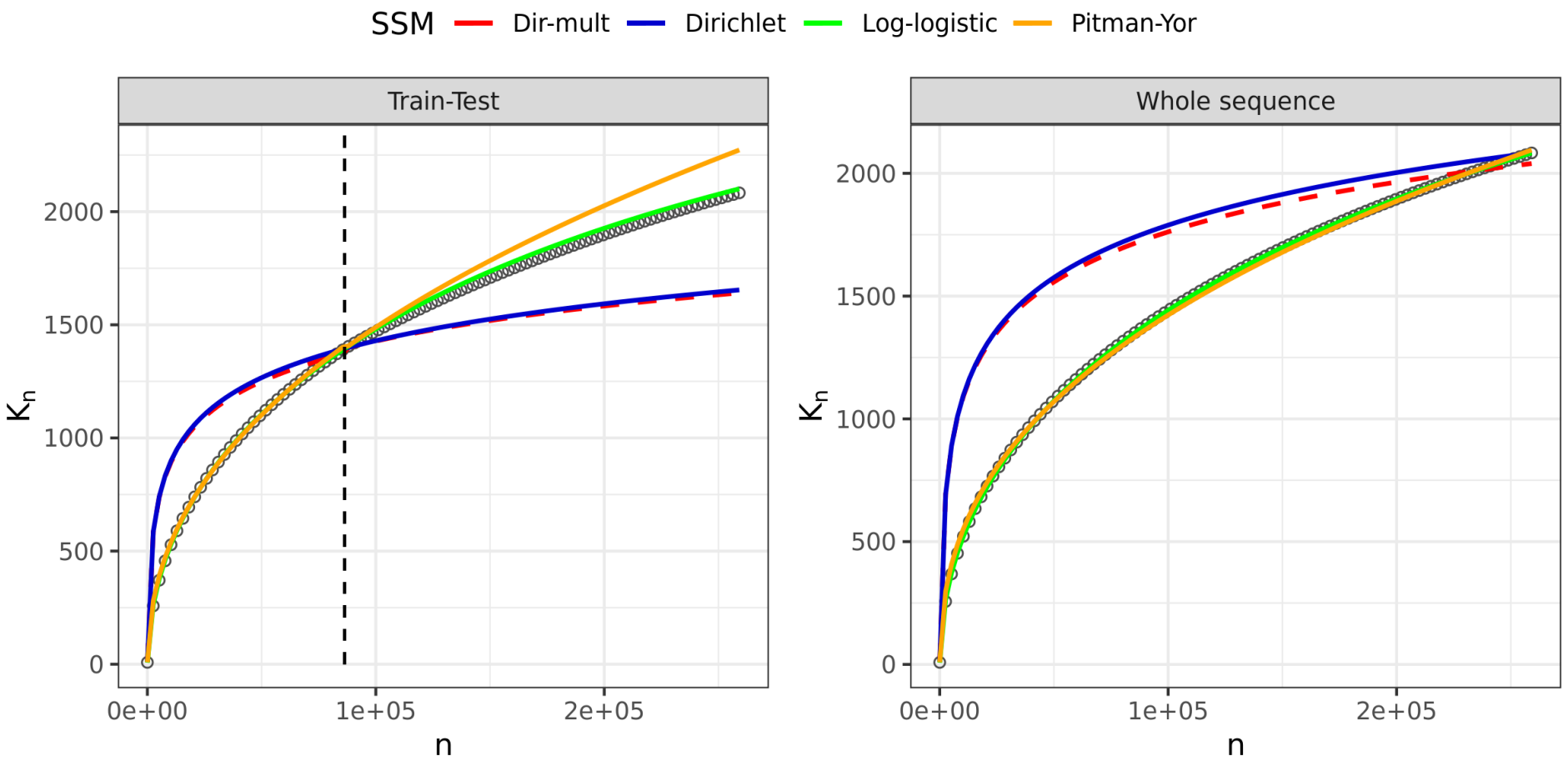
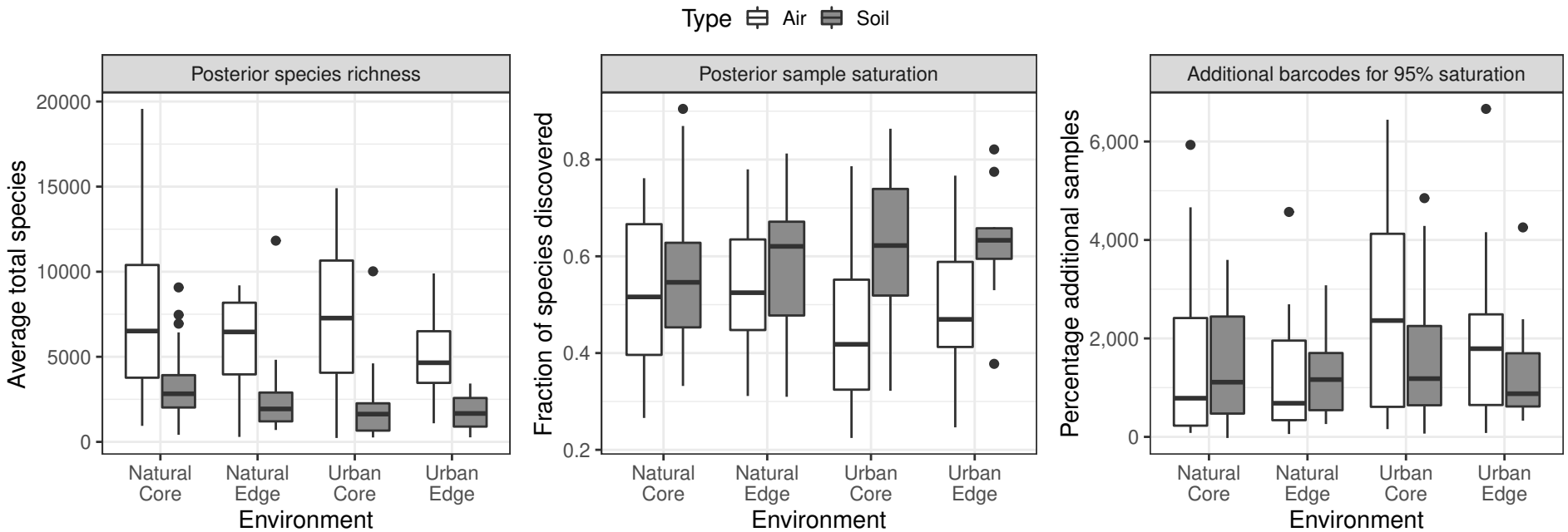


Figure: Performance of the three-parameter log-logistic against other BNP sampling schemes

Species richness and saturation in the Finnish fungal study

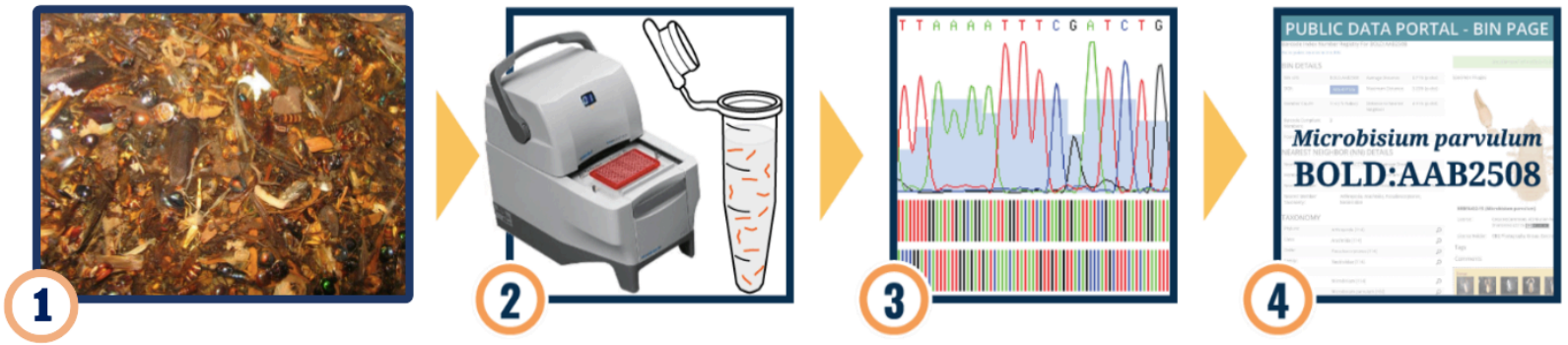
- For the 150 samples of fungal spores collected in Finland, we aim at calculating the **species richness** K_∞, the **sample saturation** $C_n = K_n/K_\infty$ and the additional number of samples needed to get the **desired saturation**



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

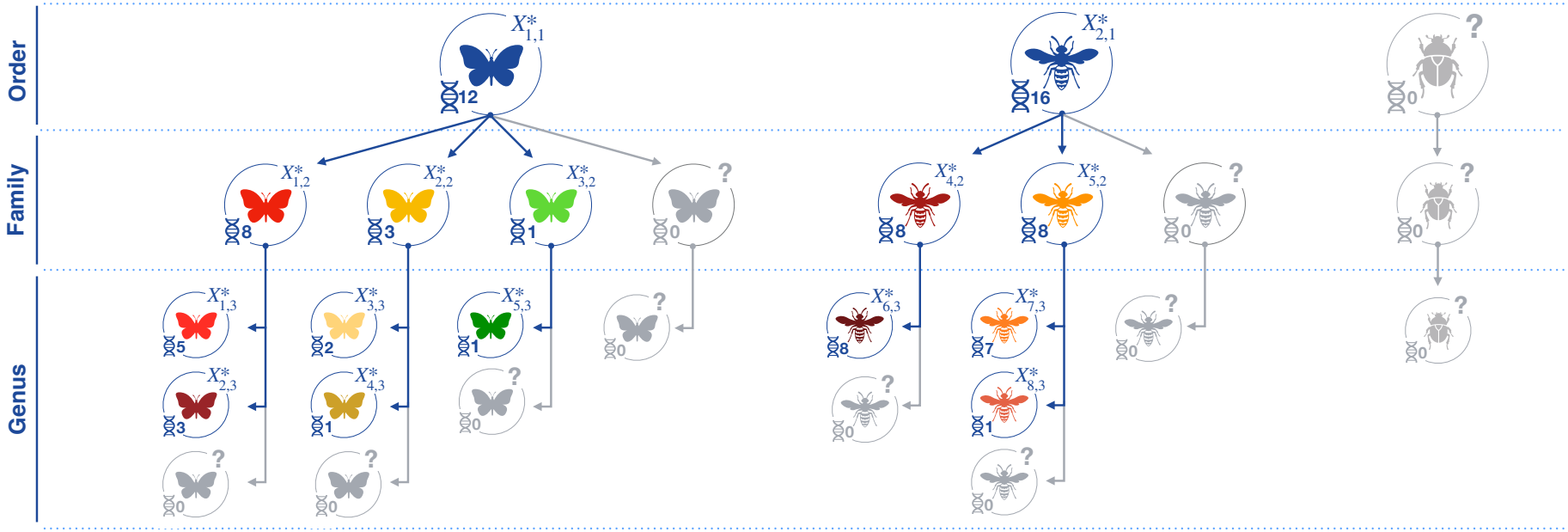
DNA barcoding

- DNA barcoding is the practice of placing DNA sequences within a Linnean taxonomy (eg. phylum, class, order, genus, species). Insects are captured via Malaise traps



Taxonomic trees may be incomplete!

- Libraries of labeled DNA (reference libraries) are often **incomplete**. For example, many species do not have a reference barcode or are still unknown to science



- When doing classification, we also need to account for the **potential novel branches**. We do this by relying again on **species sampling models**

Overview of BayesANT

- The taxonomic library of L levels is $\mathscr{D}_n = (\mathbf{X}_i, \mathbf{Y}_i)_{i=1}^n$, where $\mathbf{Y}_i$ are **DNA sequences** and $\mathbf{X}_i = (X_{i,1}, ..., X_{i,L})$ their **annotations**, such as

$$X_{i,1} = \text{“Insecta”}, \quad X_{i,2} = \text{“Diptera”}, \quad X_{i,3} = \text{“Tephritidae”}, \quad \text{etc.}$$

- Given $\mathscr{D}_n = (\mathbf{X}_i, \mathbf{Y}_i)_{i=1}^n$ and a new DNA sequence $\mathbf{Y}_{n+1}$, we **classify** the corresponding taxonomic labels $\mathbf{X}_{n+1}$ as

$$p(\mathbf{X}_{n+1} \mid \mathbf{Y}_{n+1}, \mathscr{D}_n) \propto \underbrace{p(\mathbf{X}_{n+1} \mid \mathbf{X}^{(n)})}_{\text{species sampling prior}} \times \underbrace{p(\mathbf{Y}^{(n+1)} \mid \mathbf{X}^{(n+1)})}_{\text{DNA sequence likelihood}},$$

  where $\mathbf{X}^{(n+1)} = (\mathbf{X}_i)_{i=1}^{n+1}$ and $\mathbf{Y}^{(n+1)} = (\mathbf{Y}_i)_{i=1}^{n+1}$.

- We call our method `BayesANT` - Bayesian nonparametric taxonomic classifier

The model in details

- **Taxonomic prior**: **enriched Pitman–Yor process** across L levels

$$(X_{n+1,\ell} \mid X_{n+1,\ell-1} = x, \mathbf{X}^{(n)}_{\cdot,\ell}) = \begin{cases} \text{“new”} & \text{w.p.} \quad \{\alpha_\ell + \sigma_\ell K(x)\}/\{\alpha_\ell + n(x)\}, \\ X^*_{i,\ell} & \text{w.p.} \quad \{n(X^*_{i,\ell}) - \sigma_\ell\}/\{\alpha_\ell + n(x)\}, \end{cases}$$

  where $\mathbf{X}^{(n)}_{\cdot,\ell} = (X_{i,\ell})_{i=1}^n$ and where $n(x)$ and $K(x)$ are the number of sequences and the distinct nodes linked to x.

- **DNA likelihood**: call $\boldsymbol{\theta}_{x_L}$ the **leaf-specific parameters** and $\mathcal{K}$ a generic **kernel**. Then,

$$(\mathbf{Y}_i \mid \mathbf{X}_i = (x_1, \ldots, x_L), \boldsymbol{\theta}_{x_L}) \overset{\text{ind}}{\sim} \mathcal{K}(y; \boldsymbol{\theta}_{x_L})$$

- If the sequences are **globally aligned** of the same length p, namely $\mathbf{Y}_i = (Y_{ij})_{j=1}^p$ with $Y_{ij} \in \{\text{A}, \text{C}, \text{G}, \text{T}\}$, we assume a **product-multinomial kernel**

$$\mathcal{K}(y; \boldsymbol{\theta}_{x_L}) = \prod_{j=1}^{p} \prod_{g \in \{\text{A},\text{C},\text{G},\text{T}\}} \theta_{x_L,j,g}^{\mathbb{1}\{y_j = g\}},$$

$$\theta_{x_L,j} \sim \text{Dir}(\xi_{x_L,j,\text{A}}, \ldots, \xi_{v_L,j,\text{T}})$$

```
width seq
  658 -ACTTTGTATTTTGTTTTTGGGGCTTGGGCTGCTA..
  658 -ACTTTATATTTTATTTTCGGTGCTTGATCAGGCA..
  658 -ACTTTATATTTTATTTTCGGTGCTTGATCAGGCA..
  658 -ACTTTATATTTCATTTTTGGTGCTTGATCTGGTA..
  658 -ACTTTATATTTCATTTTTGGTGCTTGATCTGGTA..
  ... ...
  658 -ACTTTATATTTTATATTTGGAATTTGATCTGGAC..
  658 -ACTTTATATTTTATCCTTGGGGCTTGGGCAGGGA..
  658 ----------------------------------..
  658 -ACATTATATTTTATTTTTGGGGCTTGGGCAGGAA..
  658 -ACTCTATATTTCATTTTTGGTACTTGAGCAGGAA..
```

One-step-ahead prediction

- The prior probabilities of a future taxonomic label are obtained via chain rule

$$\mathbb{P}(\mathbf{X}_{n+1} = \mathbf{x} \mid \mathbf{X}^{(n)}) = \mathbb{P}(X_{n+1,1} = x_1 \mid \mathbf{X}^{(n)}) \prod_{\ell=2}^{L} \mathbb{P}(X_{n+1,\ell} = x_\ell \mid X_{n+1,\ell-1} = x_{\ell-1}, \mathbf{X}^{(n)})$$

- The resulting **one-step ahead prediction rule** for the taxonomic labels $\mathbf{X}^{(n+1)}$ becomes

$$\mathbb{P}(\mathbf{X}_{n+1} = \mathbf{x} \mid \mathbf{Y}_{n+1}, \mathbf{X}^{(n)}) \propto \mathbb{P}(\mathbf{X}_{n+1} = \mathbf{x} \mid \mathbf{X}^{(n)}) \int \mathcal{K}(\mathbf{Y}_{n+1}; \boldsymbol{\theta}_{x_L}) p(\boldsymbol{\theta}_{x_L} \mid \mathscr{D}_n) \mathrm{d}\boldsymbol{\theta}_{x_L},$$

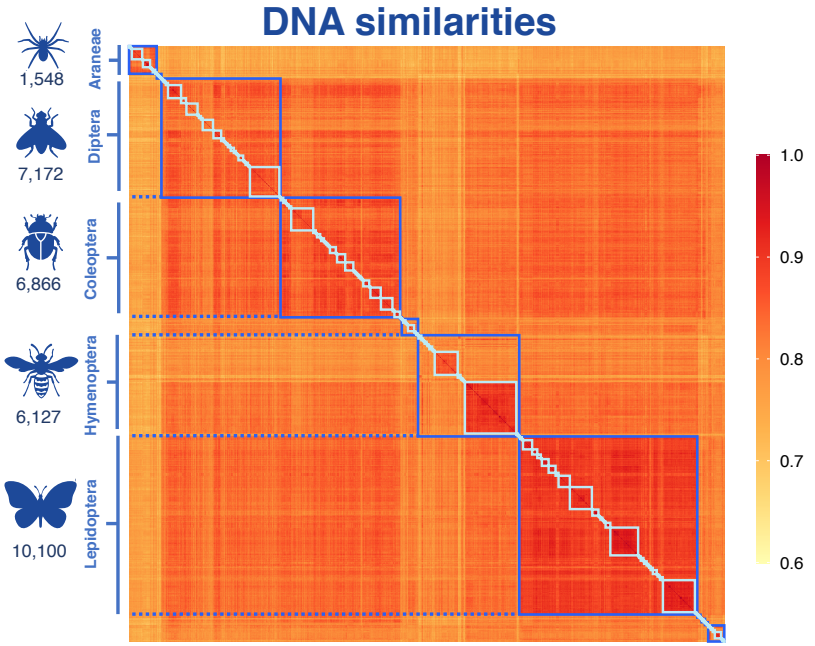
  where $p(\boldsymbol{\theta}_{x_L} \mid \mathscr{D}_n) = p(\boldsymbol{\theta}_{x_L})$ if x_L is "new"

- We tune the hyperparameters $\boldsymbol{\xi}_x$ via method of moments, and we account for model **misspecification** by **recalibrating** the probabilities, raising them to a power $\rho \in (0, 1)$

- **Classification rule:** iteratively select the taxon having the highest probability given the previously selected branch so that a meaningful taxonomic structure is preserved.

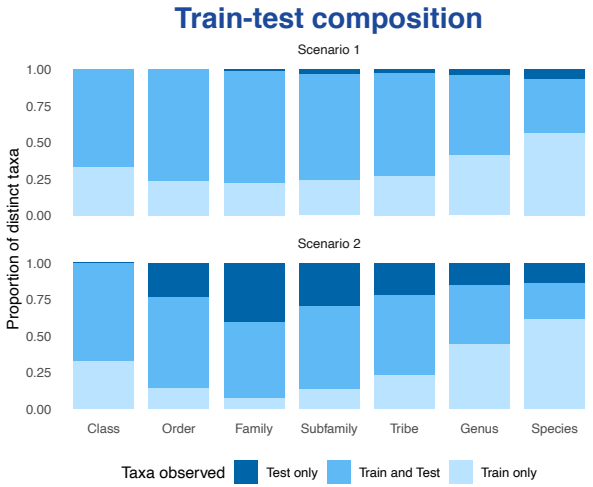
The FinBOL library and classification performance

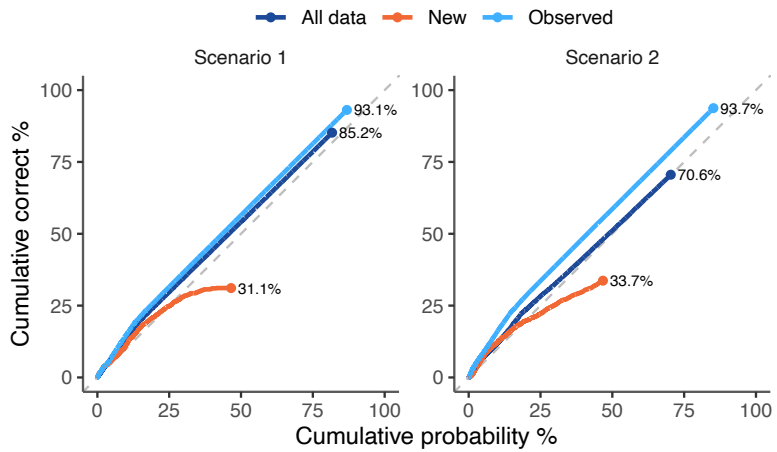
- The FinBOL library Roslin et al. (2022) contains 34624 labeled across seven levels: Class, Order, Family, Subfamily, Tribe, Genus, Species - 10985 distinct Species.



MODEL	CLASS	ORDER	FAMILY	SUBFAMILY	TRIBE	GENUS	SPECIES
M-1	100.0	99.9	98.6	97.5	96.0	92.1	85.2
	(1)	(1)	(.98)	(.96)	(.94)	(.91)	(.82)
M-2	100.0	99.9	98.4	97.2	95.8	92.4	85.4
	(1)	(1)	(.98)	(.97)	(.95)	(.93)	(.86)

Percentage o DNA sequences correctly labeled and average prediction probabilities





Conclusions

Concluding remarks

- Species sampling priors offer a rich framework to model ecological problems, from community detection to species richness estimation
- To facilitate use, all methods are made available in `R` packages:
  1. `ConjugateDP` to sample from the Stirling-gamma
  2. `BNPvegan` to estimate the sequential discovery model
  3. `BayesANT` to predict DNA sequence
- We hope that these will prove useful in the years to come!

Next steps

- In the **Stirling-gamma process**, the we have $K_\infty < \infty$ under a limiting argument. We can draw a parallel with **mixtures-of-finite-mixture**, i.e. mixture models with a prior on the number of components (Miller and Harrison, 2018)

- The **sequential discovery framework** deals with each location separately. We can extend the framework to model abundance data from multiple location, in the same spirit of **indian buffet and feature sampling models** (Griffiths and Ghahramani, 2011; Battiston et al., 2018; Masoero et al., 2021)

- There are many ways in which **BayesANT** can be extended. For example, we can choose a more flexible kernel. However, the general consensus is that we need better training libraries

Thank you!

Main references I

Battiston, M., S. Favaro, D. M. Roy, and Y. W. Teh (2018). A characterization of product-form exchangeable feature probability functions. *The Annals of Applied Probability 28*(3), 1423 – 1448.

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

Diaconis, P. and D. Ylvisaker (1979). Conjugate Priors for Exponential Families. *The Annals of Statistics 7*(2), 269 – 281.

Escobar, M. D. and M. West (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association 90*(430), 577–588.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics 1*(2), 209–230.

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

Griffiths, T. L. and Z. Ghahramani (2011). The indian buffet process: An introduction and review. *Journal of Machine Learning Research 12*(32), 1185–1224.

Main references II

Masoero, L., F. Camerlenghi, S. Favaro, and T. Broderick (2021, 02). More for less: predicting and maximizing genomic variant discovery via Bayesian nonparametrics. *Biometrika 109*(1), 17–32.

Miller, J. W. and M. T. Harrison (2018). Mixture models with a prior on the number of components. *Journal of the American Statistical Association 113*(521), 340–356. PMID: 29983475.

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

Roslin, T., P. Somervuo, M. Pentinsaari, P. D. N. Hebert, J. Agda, P. Ahlroth, P. Anttonen, J. Aspi, G. Blagoev, S. Blanco, D. Chan, T. Clayhills, J. deWaard, S. deWaard, T. Elliot, R. Elo, S. Haapala, E. Helve, J. Ilmonen, ..., and M. Mutanen (2022). A molecular-based identification resource for the arthropods of Finland. *Molecular Ecology Resources 22*(2), 803–822.