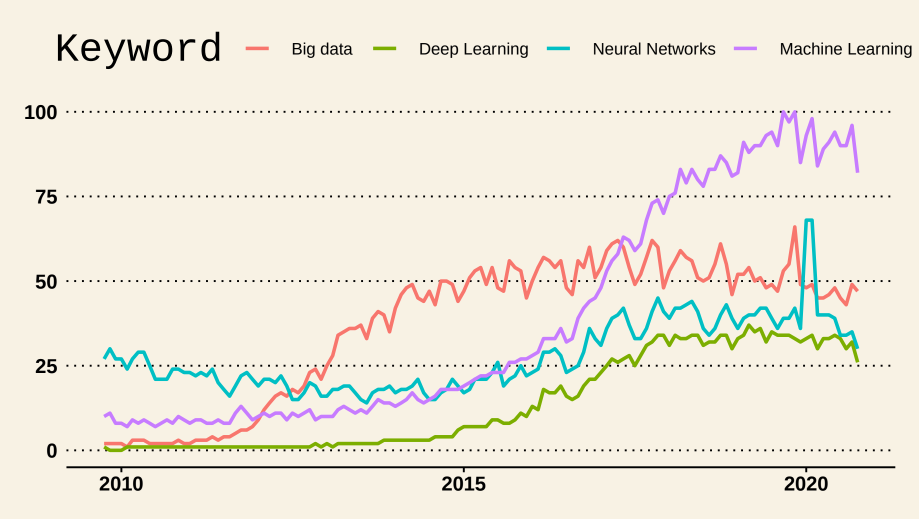
Statistical learning via Bayesian nonparametrics

Tommaso Rigon

Università degli studi di Milano-Bicocca

Milan, 2020-10-21

Google trends



Failures of the machines

- There is vast interest in **automated methods** for complex data analysis such as deep learning. However, there is a lack of consideration of the following phenomena:

- **Interpretability**. Why things work? Models vs black-box algorithms.

- **Uncertainty quantification**. A.k.a. inferential statistics: interval estimation and testing.

- **Applications with limited training data**. Data are complex but the sample size might still be very low (i.e. in neuroscience).

- **Selection bias**. If data are badly selected, having tons of data points only reduces the uncertainty in estimating the wrong quantity.

Related paper

Dunson, D. B. (2018). Statistics in the big data era: failures of the machine. *Statistics & Probability Letters*, **136**, 4–9.

Models and algorithms

- The "model vs algorithm" dispute is certainly not novel.
- Usually the following "equations" are assumed to be true:

$$\text{Machine learning} = \text{prediction}, \qquad \text{Statistics} = \text{inference}.$$

- However, modern statistics (=data science?) is **both** inference and prediction.
- "Classical" **statistical modeling** can be helpful also in **prediction tasks**: they are not complementary e.g. to random forests.

(Well-known) related paper

Breiman, L. (2001). Statistical modeling: the two cultures. *Statistical Science*, **16**(3), 199–231.

Parametric and nonparametric approaches

- However, it is certainly true **data** are becoming increasingly **complex**.
- Data may have unusual structures (networks, functions, tensors), huge dimensionality (i.e. when $p > n$), be highly non-linear, etc.
- **The statistical challenge** is researching new **flexible** modeling tools that are nonetheless interpretable and possibly scalable to large dataset.
- **Example**. In the context of regression, this means moving away e.g. from the linear model, in favor of more flexible **nonparametric specifications**, i.e.

$$\begin{aligned} \text{Parametric model}: &\qquad y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \\ \text{Nonparametric model}: &\qquad y_i = f(x_i) + \epsilon_i, \end{aligned}$$

with $f(\cdot)$ belonging to some flexible class of functions.

Bayesians & frequentists

- There are two main inferential paradigms: the **frequentist** and the **Bayesian**.
- The "frequentist vs Bayesian" discussion has been a real **ideological battlefield**.
- Before the MCMC revolution, Bayesian statistics was mainly regarded as an (elegant?) mathematical framework for inference rather than a practical tool.
- **The pragmatic Bayesian** is the statistician who makes use of Bayesian statistics because it is naturally suited for the modeling of many complex data.
- **Key idea**: incorporate in the modeling **context information** if available. This can be done both by frequentists and Bayesians, the latter disposing of a wider framework.

Related paper

Gelman, A. and Robert, C. P. (2013). "Not only defended but also applied": the perceived absurdity of Bayesian inference. *The American Statistician*, **67**, 1–5.

Bayesian nonparametrics

- **Bayesian nonparametrics** (BNP) is obviously = Bayes + nonparametric statistics.
- Its **theoretical** development began much later that parametric Bayes, after the seminal 1973 *Annals of Statistics* paper by Ferguson on the Dirichlet process.
- The availability of algorithms for posterior inference opened new directions for BNP modeling in applications, especially in the '00s and '10s.
- BNP is nowadays a mature and lively research field.
- This talk is a "mixture" of 3 separate projects involving BNP approaches in presence of complex data, for testing hypotheses, summarizing the data, and making predictions.

Bayesian testing for network partitions

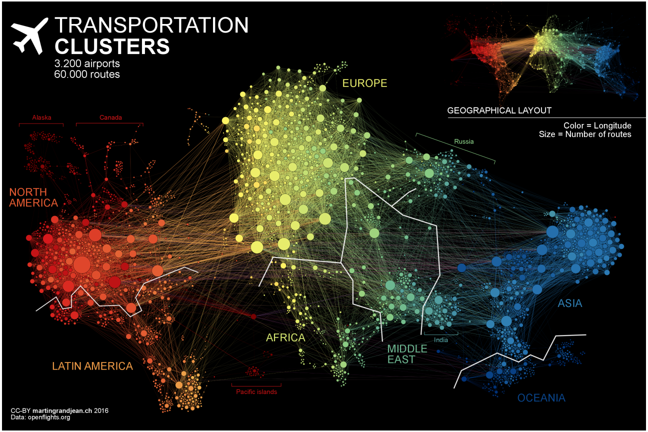
Joint work with:

Daniele Durante
(Bocconi University)

Sirio Legramanti
(Bocconi University)

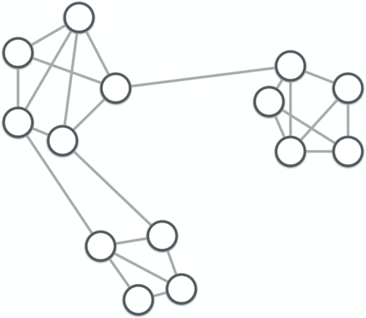
Network data

- Sometimes **relations** are more informative than **individual** characteristics

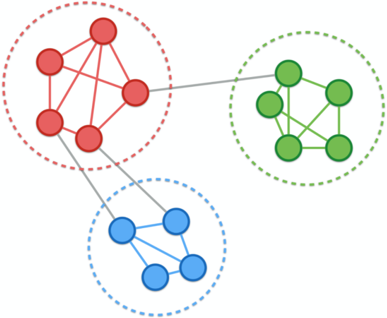


Community detection

From **network** data

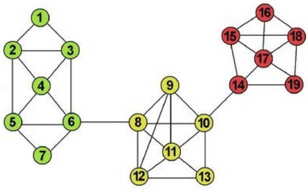


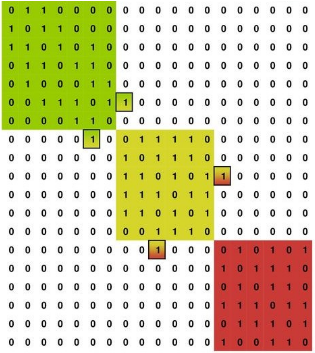
Infer the **partition** of the nodes



Network data as a binary matrix

- Networks (graphs) can be represented via their **adjacency matrix** $\mathbf{Y}$.
- Rearranging rows/columns according to the partition, $\mathbf{Y}$ may exhibit a **block structure**.





0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	1	1	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	1	0	1	0	1	1	0	0	0	0	0
0	0	0	0	0	0	0	1	1	1	0	1	1	0	0	0	0	0	0
0	0	0	0	0	0	0	1	1	0	1	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	1	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0
0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0

Stochastic block models (SBM)

- The entries of **adjacency** matrix $\mathbf{Y} = [y_{vu}]$ are defined as

$$y_{vu} = \mathbb{1}\{v \longleftrightarrow u\}, \quad v, u = 1, \dots, V.$$

- We consider **undirected network** ($\Longrightarrow y_{vu} = y_{uv}$), with no **no self-loops** ($\Longrightarrow y_{vv} = 0$).

Stochastic block models

- Let $z_v \in \{1, \dots, H\}$ be the **cluster membership** of node v
- let $\theta_{hk} \in (0, 1)$ be the **probability of an edge** between clusters h and k.
- The **likelihood** of the adjacency matrix is

$$p(\mathbf{Y} \mid \boldsymbol{\Theta}, \boldsymbol{z}) = \prod_{1 \le u < v \le n} p(y_{uv} \mid z_u, z_v, \boldsymbol{\Theta}) = \prod_{1 \le u < v \le n} \text{Bern}(y_{uv} \mid \theta_{z_u z_v}).$$

- In other words, **within clusters** the edges are iid Bernoulli random variables.

Bayesian stochastic block models

- **Edge probabilities** are given independent $\text{Beta}(a, b)$ priors, which are **conjugate**.
- The focus is on the clustering $\boldsymbol{z}$, implying that $\boldsymbol{\Theta}$ is a nuisance parameter and can be marginalized out:

$$p(\mathbf{Y} \mid \mathbf{z}) = \int p(\mathbf{Y} \mid \mathbf{z}, \boldsymbol{\Theta}) p(\boldsymbol{\Theta}) \,\mathrm{d}\boldsymbol{\Theta} = \prod_{h=1}^{H} \prod_{k=1}^{h} \frac{\mathrm{B}(a + m_{hk}, b + \overline{m}_{hk})}{\mathrm{B}(a, b)}.$$

- The integers m_{hk} are **# of edges** between clusters h and k.
- The integers $\overline{m}_{hk}$ are the **# of non-edges** between clusters h and k.
- What prior should we choose for $p(\mathbf{z})$?

Bayesian SBMs

- A simple choice is

$$\text{pr}(z_v = h \mid \boldsymbol{\pi}) = \pi_h, \qquad \boldsymbol{\pi} = (\pi_1, \dots, \pi_H) \sim \text{Dir}(\boldsymbol{\alpha}),$$

resulting in a **Dirichlet–multinomial** prior with H components.

- The number H is **fixed and finite**. How do we estimate it? Usual approaches (AIC, BIC, etc.) seem inappropriate here.

The BNP prior

- Instead of choosing it, we let $H \to \infty$. Hence, we are considering an **infinite relational model**. An alternative would be a **sparse** Dirichlet multinomial.

- The corresponding BNP prior is $\boldsymbol{z} \sim$ Chinese Restaurant Process, so that

$$\text{pr}(z_v = h \mid \mathbf{z}_{-v}) \propto \begin{cases} n_{h,-v} & \text{if } \; h = 1, \dots, \bar{H}_{-v}, \\ \alpha & \text{if } \; h = \bar{H}_{-v} + 1. \end{cases}$$

Bayesian testing of exogenous partitions

- Consider the competing models

$$\begin{aligned} \mathcal{M} &: \mathbf{z} \sim \text{Infinite Relational Model}, \\ \mathcal{M}^* &: \mathbf{z} = \mathbf{z}^* \quad \text{(exogenous assignment)}. \end{aligned}$$

- We assume that a priori

$$p(\mathcal{M}) = p(\mathcal{M}^*).$$

- Then, we test $\mathcal{M}$ vs $\mathcal{M}^*$ through the **Bayes factor**

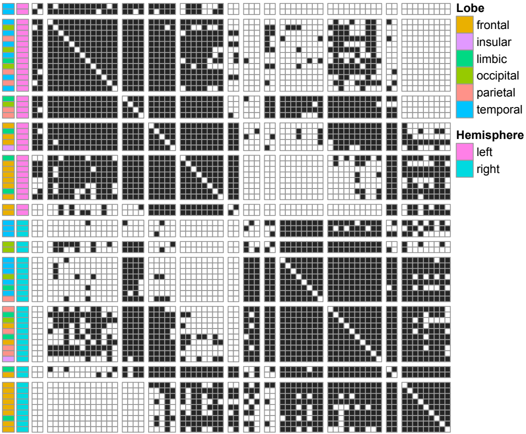
$$\mathcal{B}_{\mathcal{M},\mathcal{M}^*} = \frac{p(\mathbf{Y} \mid \mathcal{M})}{p(\mathbf{Y} \mid \mathcal{M}^*)} = \frac{\sum_{\mathbf{z} \in \mathcal{Z}} p(\mathbf{Y} \mid \mathbf{z}) p(\mathbf{z})}{p(\mathbf{Y} \mid \mathbf{z}^*)},$$

which coincides with the posterior odds

$$\frac{p(\mathcal{M} \mid \mathbf{Y})}{p(\mathcal{M}^* \mid \mathbf{Y})}.$$

- Bayes factors are computed using suitable MCMC algorithms.

Alzheimer's brain network data



- Presence of **white matter fibers** among **68 anatomical regions** in a representative Alzheimer's brain network, split according to the estimated endogenous assignments

Testing exogenous partitions

- Do hemispheres or lobes capture endogenous blocks? **No**, at least according to Bayes factors.

- Recall that $2 \log \hat{\mathcal{B}}_{\mathcal{M},\mathcal{M}^*} \gg 0$ supports the choice $\mathcal{M} =$ Infinite Relational Model.

	Hemisphere	Lobes
$2 \log \hat{\mathcal{B}}_{\mathcal{M},\mathcal{M}^*}$	712.33	1290.50

- **Explanation:** there exist sub–blocks (groups) within hemispheres, comprising regions in different lobes.

Testing exogenous partitions

- Our network data is a representative brain with Alzheimer's disease.

- We let $\mathbf{z}^*$ be the estimated partition from a representative brains of individuals characterized by normal aging, early and late cognitive decline

$$
\begin{aligned}
\mathcal{M}: &\quad \mathbf{z} \sim \text{Infinite Relational Model} \\
\mathcal{M}^*: &\quad \mathbf{z} = \mathbf{z}^*
\end{aligned}
$$

	Normal Aging	Early Decline	Late Decline
$2 \log \hat{\mathcal{B}}_{\mathcal{M},\mathcal{M}^*}$	155.01	100.21	39.88

- $\mathcal{M}^*$ is always rejected, **BUT** evidence against $\mathcal{M}^*$ decreases moving towards the disease state $\implies$ inferred partitions as diagnostics for the disease progress?

Extended stochastic block models

- The Chinese restaurant process prior for $\mathbf{z}$ is a simple yet (sometimes) insufficiently flexible prior.
- In more recent works (Legramanti et al., 2020), we make use of **Gibbs-type prior** for $p(\mathbf{z})$ rather than implicitly relying on the Dirichlet process.
- We called this class **extended stochastic block models** (ESBM).
- The so-called **Gnedin process** prior seems to have better empirical performance in sumulations and applications while remaining computationally tractable.
- Interestingly, in ESBM we can choose whether H is finite, random or infinite within the same unified framework.

Enriched processes & flight-route segmentation

Joint work with:

Sonia Petrone
(Bocconi University)

Bruno Scarpa
(University of Padova)

Outline of the project

- We aim at clustering **functional observations** via nonparametric Bayes.

- In this motivating application, each statistical unit is a **flight route**.

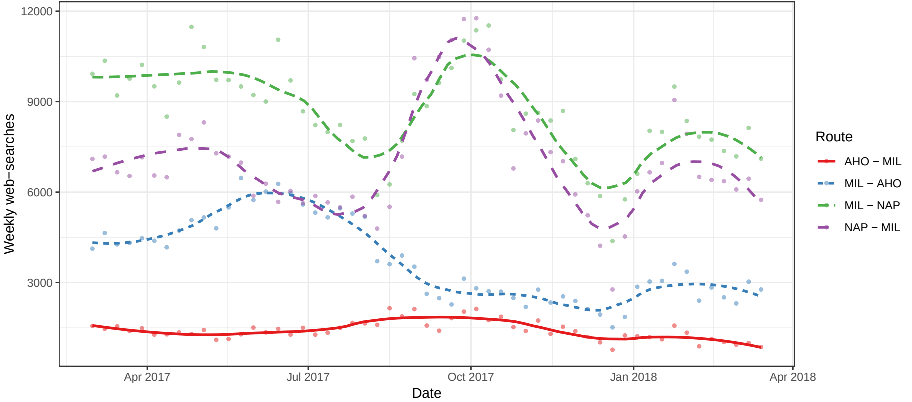
- In particular, we consider the number of times that a specific route has been searched on the website of an **e-commerce** company.

Statistical challenges

- **Bounding the complexity**. Infinite-dimensional BNP priors often lead to overly complex cluster solutions.

- **Functional constraints**. Prior knowledge about the functional shapes is available, but it is not easy to incorporate.

E-commerce dataset



- The total number of flight routes is $n = 214$.
- Each trajectory is observed over a **weekly time grid** $\boldsymbol{t}_i = (1, \dots, 55)$. Hence, the dataset can be represented as a 214×55 matrix with 11770 entries.

General considerations

- Could we consider different **metrics**?

- **Yes, but** private companies are (rightly!) worried about disclosing their data to the public. In principle, other metrics might include:
  - Route prices;
  - Route marginal earnings;
  - Route-specific customer satisfaction;
  - Conversion rates;
  - ...

- A very crude but operative summary of each time series is its **average**.

- Missing part of the story: **clustering shapes** and not average levels.

Model formulation

- Functional observations are **standardized**, i.e. they have zero mean and unit variance. Moreover, let

$$y_i(t) = f_i(t) + \epsilon_i(t), \qquad (f_i \mid \tilde{p}) \overset{\text{iid}}{\sim} \tilde{p}, \qquad i = 1, \dots, n,$$

where $\epsilon_i(t)$ is a Gaussian error and $t \in \mathbb{R}^+$.

- Clustering is induced through a **discrete prior** $\tilde{p}$, whose choice is critical.

- The **functional** DP (Bigelow and Dunson, 2009; Dunson, Herring and Siega-Riz, 2008) would **fail** in bounding the complexity and incorporating functional constraints.

An enriched discrete prior

- The proposed process is a **mixture of random probability measures**:

$$\tilde{p} = \sum_{\ell=1}^{L} \Pi_\ell \tilde{p}_\ell = \sum_{\ell=1}^{L} \Pi_\ell \sum_{h=1}^{H_\ell} \pi_{\ell h} \delta_{\theta_{\ell h}(t)}, \quad \theta_{\ell h}(t) \stackrel{\text{ind}}{\sim} P_\ell,$$

for $h = 1, \dots, H_\ell$ and $\ell = 1, \dots, L$.

- Each P_ℓ is a **diffuse** probability measure taking values on a given **functional class** (monotone, cyclical, linear, S-shaped functions, etc).

- Closely related to the **enriched processes** of Wade et al. (2011) and Scarpa and Dunson (2014), but the number of clusters is **bounded**.

Clustering allocation process

- $G_i \in (\ell, h)$ is a latent **cluster** indicator, so that $f_i(t)$ and $f_j(t)$ belong to the same group if $G_i = G_j$.

- $F_i \in \{1, \dots, L\}$ is a latent **functional class** indicator.

- Functional class allocation: $\mathbb{P}(F_i = \ell) = \Pi_\ell$,
- Within-class allocation: $\mathbb{P}(G_i = (\ell, h) \mid F_i = \ell) = \pi_{\ell h}$,
- Cluster allocation: $\mathbb{P}(G_i = (\ell, h)) = \Pi_\ell \pi_{\ell h}$.

- **Sparsity** can be induced as in Rousseau and Mengersen (2011).
- Functional class prior: $(\Pi_1, \dots, \Pi_{L-1}) \sim \text{DIRICHLET}(\alpha_1, \dots, \alpha_L)$.
- Shrinkage prior: $(\pi_{\ell 1}, \dots, \pi_{\ell H_\ell - 1}) \overset{\text{ind}}{\sim} \text{DIRICHLET}\,(c_\ell / H_\ell, \dots, c_\ell / H_\ell)$.

Baseline measure specification

- Each P_ℓ can be interpreted as a functional prior guess, since

$$\mathbb{E}\{\tilde{p}(\cdot)\} = \sum_{\ell=1}^{L} \mathbb{E}(\Pi_\ell) P_\ell(\cdot) = \frac{1}{\alpha}\sum_{\ell=1}^{L} \alpha_\ell P_\ell(\cdot), \qquad \alpha = \sum_{\ell=1}^{L} \alpha_\ell.$$

- We assume that $\theta_{\ell h}(t)$ is **linear in the parameters**:

$$\theta_{\ell h}(t) = \sum_{m=1}^{M_\ell} \mathcal{B}_{m\ell}(t)\beta_{m\ell h},$$

where each $\mathcal{B}_{1\ell}(t), \dots, \mathcal{B}_{M_\ell \ell}(t)$ for $\ell = 1, \dots, L$ is a set of **pre-specified basis functions**.

- Moreover, we assume $(\beta_{1\ell h}, \dots, \beta_{M_\ell \ell h})^\intercal$ have Gaussian priors.

Baseline measure specification

- The first functional class ($\ell = 1$) captures yearly **cyclical patterns** and characterizes the routes having e.g. a peak of web-searches during either the summer or the winter.

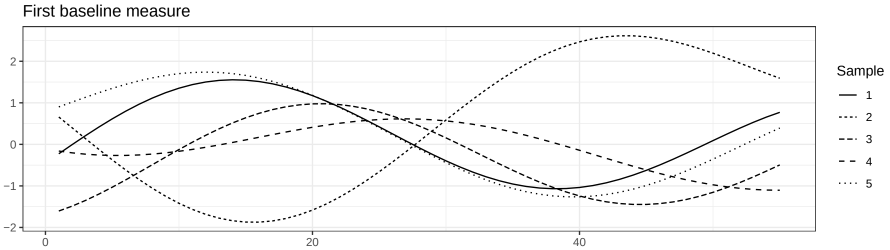
$$\theta_{1h}(t) = \sum_{m=1}^{4} \beta_{m1h}\mathcal{S}_m(t) + \beta_{51h}\cos\left(2\pi\frac{7}{365}t\right) + \beta_{61h}\sin\left(2\pi\frac{7}{365}t\right),$$

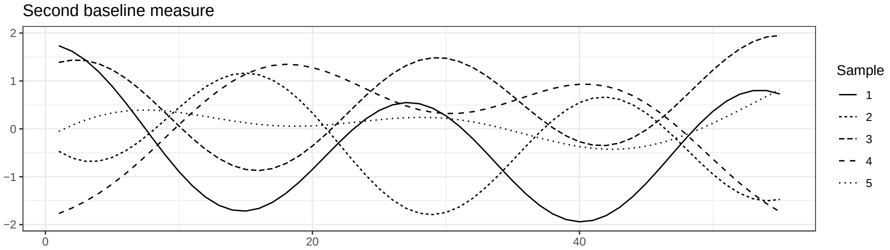
where $\mathcal{S}_1(t), \dots, \mathcal{S}_4(t)$ are deterministic cubic spline basis functions.

- The second functional class ($\ell = 2$) characterizes functions having **two peaks per year**, which amounts to let

$$\theta_{2h}(t) = \sum_{m=1}^{4} \beta_{m2h}\mathcal{S}_m(t) + \beta_{52h}\cos\left(2\pi\frac{14}{365}t\right) + \beta_{62h}\sin\left(2\pi\frac{14}{365}t\right).$$

Baseline measure specification

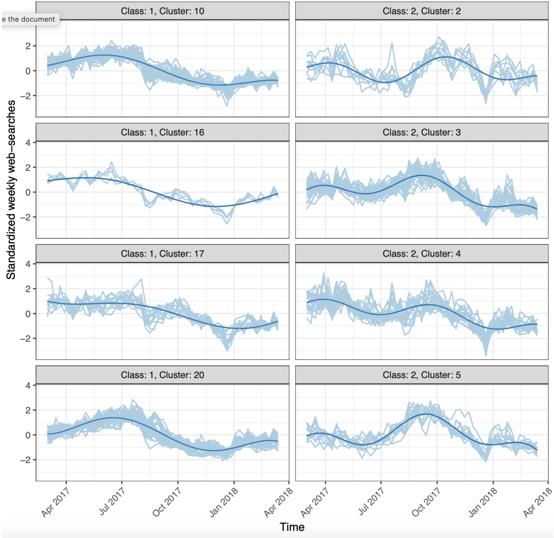




On the selection of the upper bounds

- The number of clusters is bounded by $H = \sum_{\ell=1}^{L} H_\ell$. We consider a large H and employ a **sparse prior**, following Rousseau and Mengersen (2011).
- In practice, we let $H = \sum_{\ell=1}^{L} H_\ell$ be **the largest value for which the resulting clustering solution is still useful** in practice.
- Such a value is evidently quite subjective and it depends on the specific statistical problem.
- In our e-commerce application we let the upper bounds $H_1 = 20$ and $H_2 = 5$.

Clustering solution



Macro clusters A and B

		Arrival		
		North	Center	South & Islands
Departure	North	0	2	49
	Center	0	0	24
	South & Islands	6	3	12

		Arrival		
		North	Center	South & Islands
Departure	North	0	7	6
	Center	10	0	0
	South & Islands	47	21	7

Appendix: theoretical developments

- Define independently among themselves

$$\tilde{\mu}_x \sim \mathrm{GAP}(\alpha P_x), \qquad \tilde{p}_{y|x}(\cdot \mid x) \overset{\text{ind}}{\sim} \mathrm{PY}\{\sigma(x), \beta(x) P_{y|x}(\cdot \mid x)\}, \qquad x \in \mathbb{X}.$$

  A **gamma and Pitman–Yor** (GA-PY) random measure $\tilde{\mu}$ is defined as

$$\tilde{\mu}(A \times B) = \int_A \tilde{p}_{y|x}(B \mid x)\tilde{\mu}_x(\mathrm{d}x), \qquad A \subseteq \mathbb{X}, \quad B \subseteq \mathbb{Y}.$$

- Then $\tilde{p}$ is called **enriched Pitman–Yor** process (EPY) if

$$\tilde{p} = \frac{\tilde{\mu}}{\tilde{\mu}(\mathbb{X} \times \mathbb{Y})}.$$

Appendix: theoretical characterizations

Theorem

Let $\tilde{\mu} \sim \text{GA-PY}(\alpha P_x, \sigma(x), \beta(x) P_{y|x})$ and let $\alpha P_x(\cdot) = \sum_{\ell=1}^{L} \alpha_\ell \, \delta_{x_\ell}(\cdot)$ be a discrete measure. Then,

$$\mathbb{E}\left\{e^{-\tilde{\mu}(g)}\right\} = \prod_{\ell=1}^{L} \mathbb{E}\left[\left\{1 + \tilde{p}_{y|x}(g \mid x_\ell)\right\}^{-\alpha_\ell}\right],$$

where $\tilde{p}_{y|x}(f \mid x) = \int_{\mathbb{Y}} g(x,y) \tilde{p}_{y|x}(\mathrm{d}y \mid x)$ for any $x \in \mathbb{X}$ and for any measurable function $g : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}^+$ such that $\tilde{\mu}(g) < \infty$ almost surely.

- The expectation appearing in the right hand side of the above Laplace functional is a generalized **Cauchy-Stieltjes transform**.
- Further simplifications occurs in the EDP case thanks to the **Cifarelli-Regazzini** identity.

A unified class of enriched priors

- Our model, the enriched process of Wade et al. (2011), Dunson and Scarpa (2014) belong to this general class of enriched priors.
- **Mixture of mixture** models by Malsiner-Walli et al. (2017) can be also viewed as EPY processes.
- **Spike-and-slab** Dirichlet priors (Dunson, Herring and Hengel, 2008; Guindani, Müller and Zhang, 2009) can be also regarded as EPYs.
- Further connections with the **dependent Dirichlet processes** of Müller, Quintana and Rosner (2004), Lijoi, Nipoti and Prünster (2014).
- These aspects are explored in Rigon, Scarpa and Petrone (2020+).

BNP modeling of sequential discoveries

Joint work with:

David Dunson
(Duke University)

Otso Ovaskainen
(University of Finland)

Alessandro Zito
(Duke University)

Outline of the project

- The sequential modelling of the appearance of distinct species is a widely studied problem. Famous example: **Fisher et al. (1943)**

- How many more **new species** exists in a community I did not observed yet?

- Suppose that for a given location we observe a sequence of species.

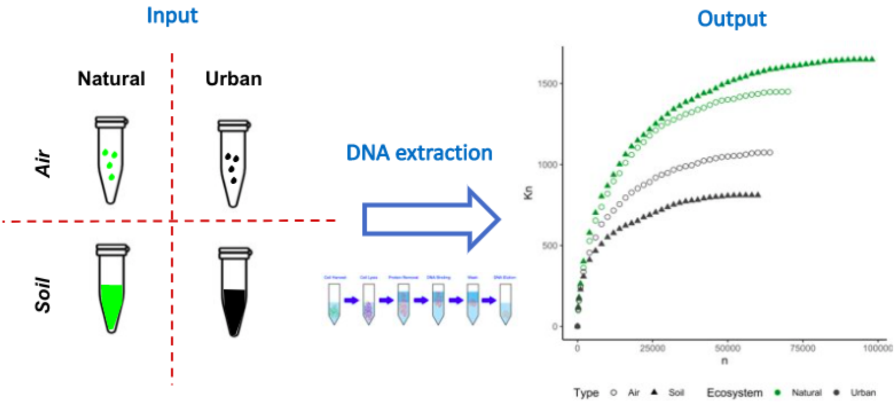
$$X_1 = \text{“squirrel”},\ X_2 = \text{“dog”},\ X_3 = \text{“cat”},\ X_4 = \text{“squirrel”},\ \text{etc.}$$

- A **species accumulation curve** is the **trajectory** of the total number of new species K_n as a function of n, where

$$K_n = \sum_{i=1}^{n} \mathbb{1}(X_i = \text{“new”}), \quad n \geq 1.$$

- In modern experiments however, species may be detected from their DNA.

DNA sequencing



Species sampling modeling

- **Issue**: DNA process is costly, and no guarantee that it captures all the species!
- We need a method to assess whether we have detected *all* the species trapped (**sample saturation**).
- Given an accumulation curve

$$K_n = \sum_{i=1}^{n} D_i, \qquad D_i = \mathbb{1}(X_i = \text{``new''}),$$

our method should be able to:

  - Smooth the **in-sample** trajectory $K_1, \dots, K_n$.
  - Predict the **out-of-sample** trajectory of $K_{n+1}, \dots, K_{n+m}$ for any $m \geq 1$.
  - Study the behavior of $K_\infty = \lim_{n\to\infty} K_n$ (i.e. saturation level).

BNP species sampling models

- These requirements can be answered by **Bayesian nonparametric species sampling models**.

- A common **exchangeable** model for $(X_n)_{n \geq 1}$ is the Pitman–Yor, in which:

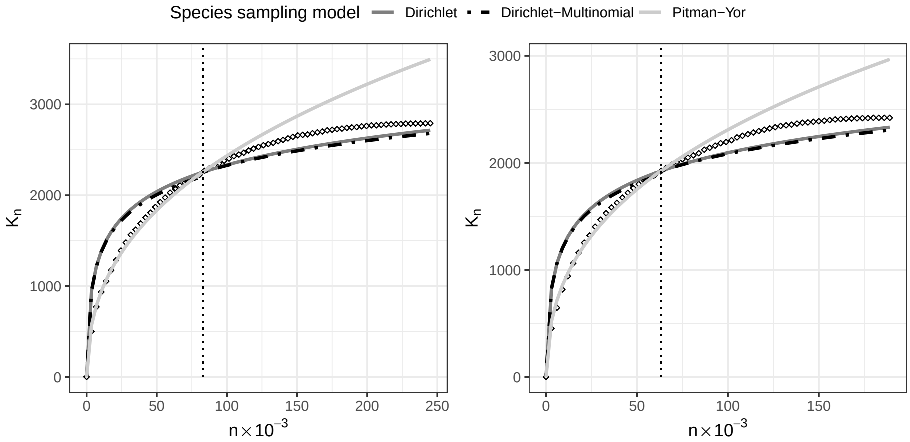
$$\mathrm{pr}(X_{n+1} = \text{“new”} \mid X_1, \ldots, X_n) = \frac{\alpha + \sigma K_n}{\alpha + n}$$

- The σ parameters controls the asymptotic behavior of K_n.

  - If $\sigma = 0$ and $\alpha > 0 \Rightarrow$ **Dirichlet process**
    - $K_n \sim \alpha \log n$ and $K_\infty = \infty$ a.s.
  - If $\sigma \in (0, 1)$ and $\alpha > -\sigma \Rightarrow$ **Pitman-Yor process**
    - $K_n \sim \mathcal{O}(n^\sigma)$ and $K_\infty = \infty$ a.s.
  - If $\sigma < 0$ and $\alpha = H|\sigma|$, with $H \in \mathbb{N} \Rightarrow$ **Dirichlet-Multinomial**
    - $K_\infty = m < \infty$ a.s, but very hard to estimate

- All three models have closed-form estimators for the predictions. **However...**

Bayesian nonparametric species sampling models



- Failures of common species sampling models. Dots are the **observed values** for K_n.

The model

- We assume $(D_n)_{n\geq 1}$ are **independent Bernoullies** with probabilities $(\pi_n)_{n\geq 1}$.

- For any $n \geq 1$, the **discovery probability** is equal to

$$\pi_n = \mathrm{pr}(D_n = 1) = \mathrm{pr}(T_n > n-1) = S(n-1;\theta), \quad \theta \in \Theta \subseteq \mathbb{R}^p,$$

  where $(T_n)_{n\geq 1}$ are iid continuous latent variables defined in $(0,\infty)$ with strictly decreasing **survival function** $S(t;\theta)$.

- The resulting distribution for K_n follows a Poisson-Binomial distribution:

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

- The **likelihood** is readily available as

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

Basic properties

- A **in-sample prediction** of the trajectory is the expectation of K_n, namely

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

- A **out-of-sample prediction** of the trajectory is a posterior expectation:

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

- The latent variable T controls the **asymptotic behavior**.

Proposition

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

$$E(T) \leq E(K_\infty) \leq E(T) + 1.$$

Moreover, $K_\infty = \infty$ almost surely if and only if $E(T) = \infty$.

The Dirichlet process

- Our starting point is the **Dirichlet process**

$$\pi_{n+1} = S(n;\alpha) = \frac{\alpha}{\alpha + n}, \quad \alpha > 0.$$

- **Interesting fact**: $S(t;\alpha)$ is the survival function of a **log–logistic distribution**.

Theorem

If $(X_n)_{n\geq 1}$ is directed by a Dirichlet process, then for a sample of $X_1,\dots,X_n$ with $K_n = k$ it holds that

$$\mathscr{L}(\alpha \mid X_1,\dots,X_n) \propto \mathscr{L}(\alpha \mid D_1,\dots,D_n) \propto \frac{\alpha^k}{(\alpha)_n},$$

where $(\alpha)_n = \alpha(\alpha+1)\cdots(\alpha+n-1)$.

- **Advantage**: Same likelihood $\implies$ same inference.

- **Disadvantage**: K_n diverges to ∞ at the rate $\alpha \log n \implies$ too strict!

A three parameters log-logistic

- A possibile choice for the distribution of T is a **three parameter log-logistic**.

- Hence, when $T \sim \text{LL}(\alpha, \sigma, \phi)$ we obtain

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

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

- This embeds several behaviors:
  - For $\phi = 1$ and $\sigma = 0 \Rightarrow$ **Dirichlet process**
    - $K_n \sim \alpha \log n$ and $K_\infty = \infty$ a.s.
  - For $\phi = 1$ and $\sigma < 0 \Rightarrow$ *similar* to **Dirichlet-Multinomial**
    - $K_\infty < \infty$ a.s., $E(T)$ is in closed form
  - For $\phi = 1$ and $\sigma \in [0, 1) \Rightarrow$ *similar* to **Pitman-Yor**
    - $K_n \sim \mathcal{O}(n^\sigma)$ and $K_\infty = \infty$ a.s.
  - For $\phi < 1 \Rightarrow$ **convergence always ensured**
    - $K_\infty < \infty$ a.s., $E(T)$ needs approximation.

Logistic regression representation

- Under $T \sim \mathrm{LL}(\alpha, \sigma, \phi)$, the pdf of K_n is available.

- The estimation of the parameters can be easily carried due to link with **logistic regression**:

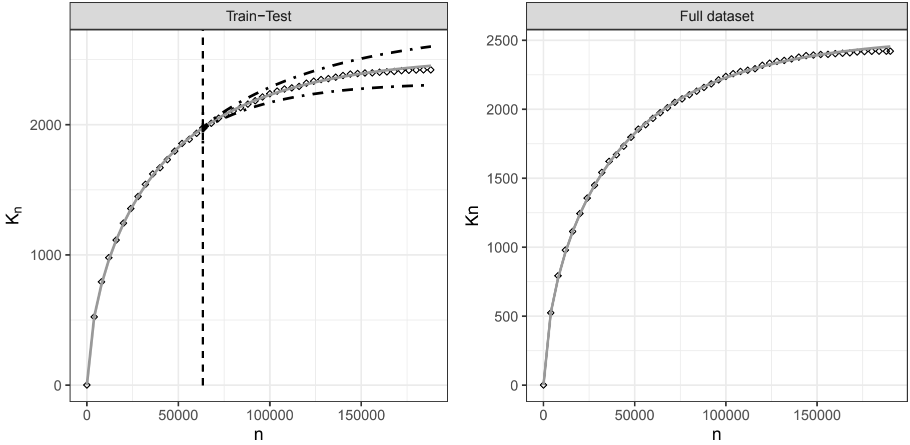
$$\log \frac{\pi_{n+1}}{1 - \pi_{n+1}} = \log \alpha - (1 - \sigma) \log n + (\log \phi) n = \beta_0 + \beta_1 \log n + \beta_2 n,$$

with $\beta_0 = \log \alpha$, $\beta_1 = \sigma - 1 < 0$ and $\beta_2 = \log \phi \leq 0$ for every $n \geq 1$.

- If **truncated normals** are employed for β_1 and β_2 $\implies$ posterior can be obtained via Pólya-Gamma data augmentation (Polson et al. 2013).

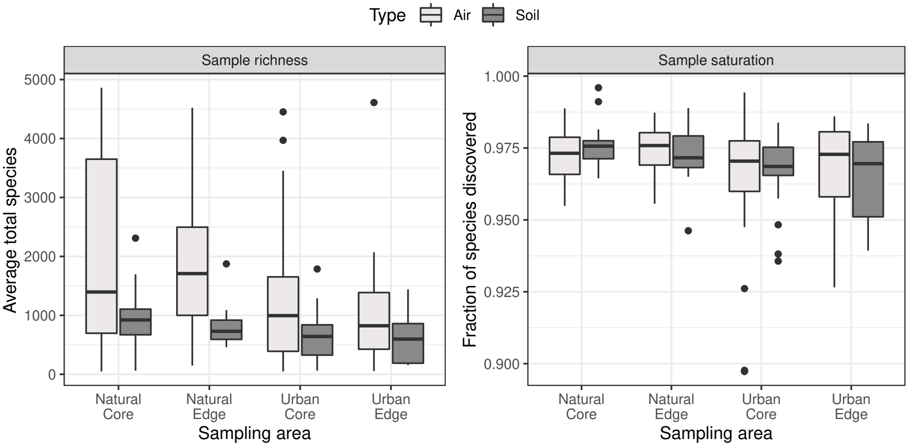
- **Summary**. We trade exchangeability in favor of
  - A model easier to fit;
  - A wider and more flexible class of trajectories;
  - A model in which K_∞ is always finite.

The LIFEPLAN data



- Performance of the three-parameter log-logistic performance.

The LIFEPLAN data



Summary

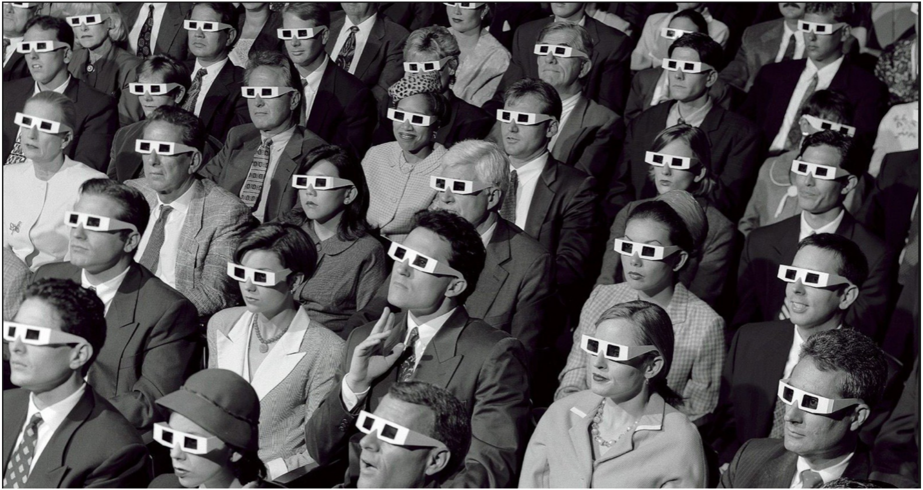
Pros

- The Poisson–Binomial offers an alternative general framework to model the sequential discovery of new species.
- The three-parameter log-logistic is simple generalization of the Dirichlet process which increases the flexibility and allowing for different asymptotic regimes.
- The model is easy and fast to estimate both in terms of empirical Bayes and with fully Bayesian approach.

Side effect

- The resulting model drops the exchangeability assumption. This means that the results will be always sequence-dependent.

Thanks!



Appendix for BNP species sampling model

Theorem

Let $K_n \sim \mathrm{PB}\{1, S(1;\theta), \ldots, S(n-1;\theta)\}$ *and suppose that* $K_\infty = \infty$ *almost surely. Then*

$$\frac{K_n}{b_n} \to 1, \qquad b_n = \int_1^n S(t-1;\theta)\mathrm{d}t, \qquad n \to \infty,$$

almost surely. In addition, it holds that

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

in distribution.

Appendix for BNP species sampling model

Theorem

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

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

where for any $1 \leq k \leq n$ *and* $n \geq 2$ *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* $\{1,\dots,n-1\}$.

- **Recursion**: $\mathscr{C}_{n+1,k}(\sigma,\phi) = \mathscr{C}_{n,k-1}(\sigma,\phi) + n^{1-\sigma}\phi^{-n}\mathscr{C}_{n,k}(\sigma,\phi)$, for any $n \geq 0$ and $1 \leq k \leq n+1$.
- **Initial conditions**: $\mathscr{C}_{0,0}(\sigma,\phi) = 1$, $\mathscr{C}_{n,0}(\sigma,\phi) = 0$ for $n \geq 1$, $\mathscr{C}_{n,k}(\sigma,\phi) = 0$ for $k > n$.