



**HAL**  
open science

## Bayesian Nonparametric Mixtures Why and How?

Julyan Arbel

► **To cite this version:**

Julyan Arbel. Bayesian Nonparametric Mixtures Why and How?. IFSS 2018 - 2nd Italian-French Statistics Seminar, Sep 2018, Grenoble, France. <hal-01950664>

**HAL Id: hal-01950664**

**<https://hal.science/hal-01950664v1>**

Submitted on 11 Dec 2018

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

### Introduction

#### Bayesian nonparametric framework

- Massively many parameters
- Inference on curves: pdf, cdf, hazard, link...
- Mixtures, exchangeable data  $\mathbf{X}^n = (X_1, \dots, X_n)$

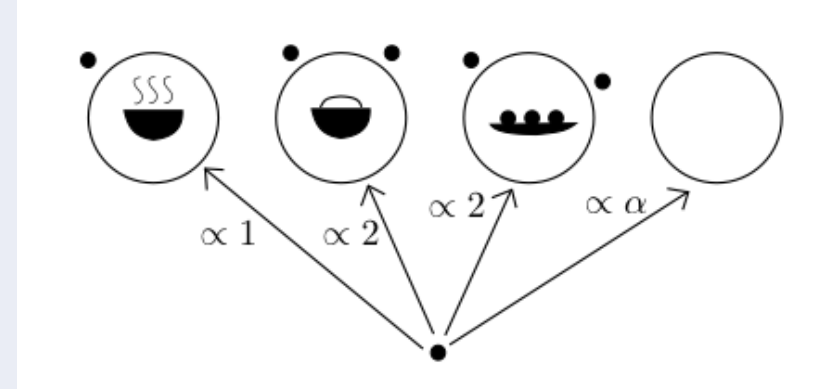
$$X_1, \dots, X_n | P \sim \begin{cases} P & \rightarrow \textcircled{1} \\ \int_{\Theta} k(\cdot | \theta) P(d\theta) & \rightarrow \textcircled{2} \textcircled{3} \end{cases}$$

- Natural uncertainty quantification
- Flexibility, avoids over-fitting by regularization (prior)
- Adapt to data complexity
- Underlying clustering
- Justify prior, expert
- Efficient posterior sampling
- Quantify truncation error

#### What prior for $P$ ?

- Learn about data through **posterior dist.**
- Discrete **random probability measure** prior
- Random weights  $(p_i)_i$  and locations  $(\theta_i)_i$

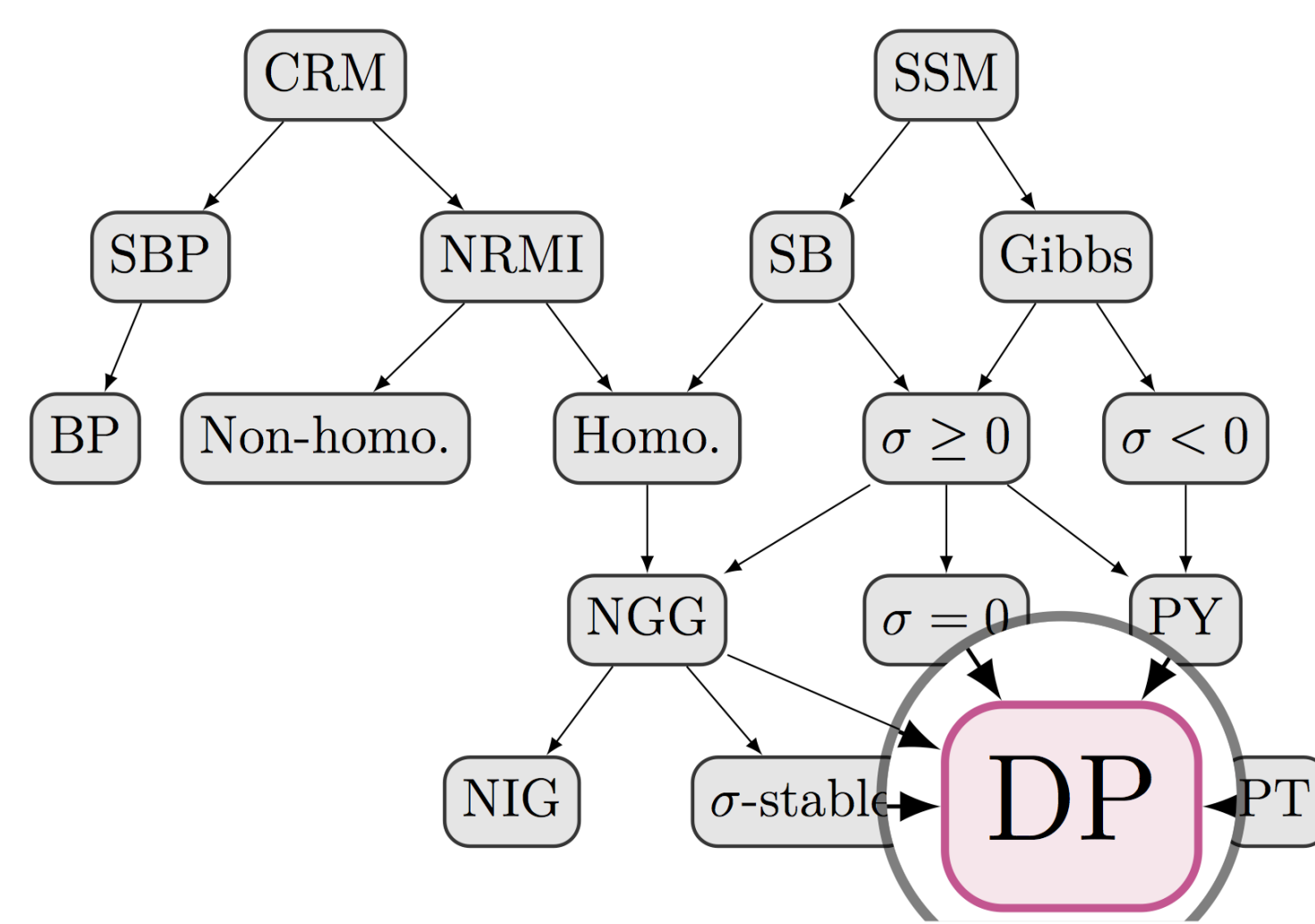
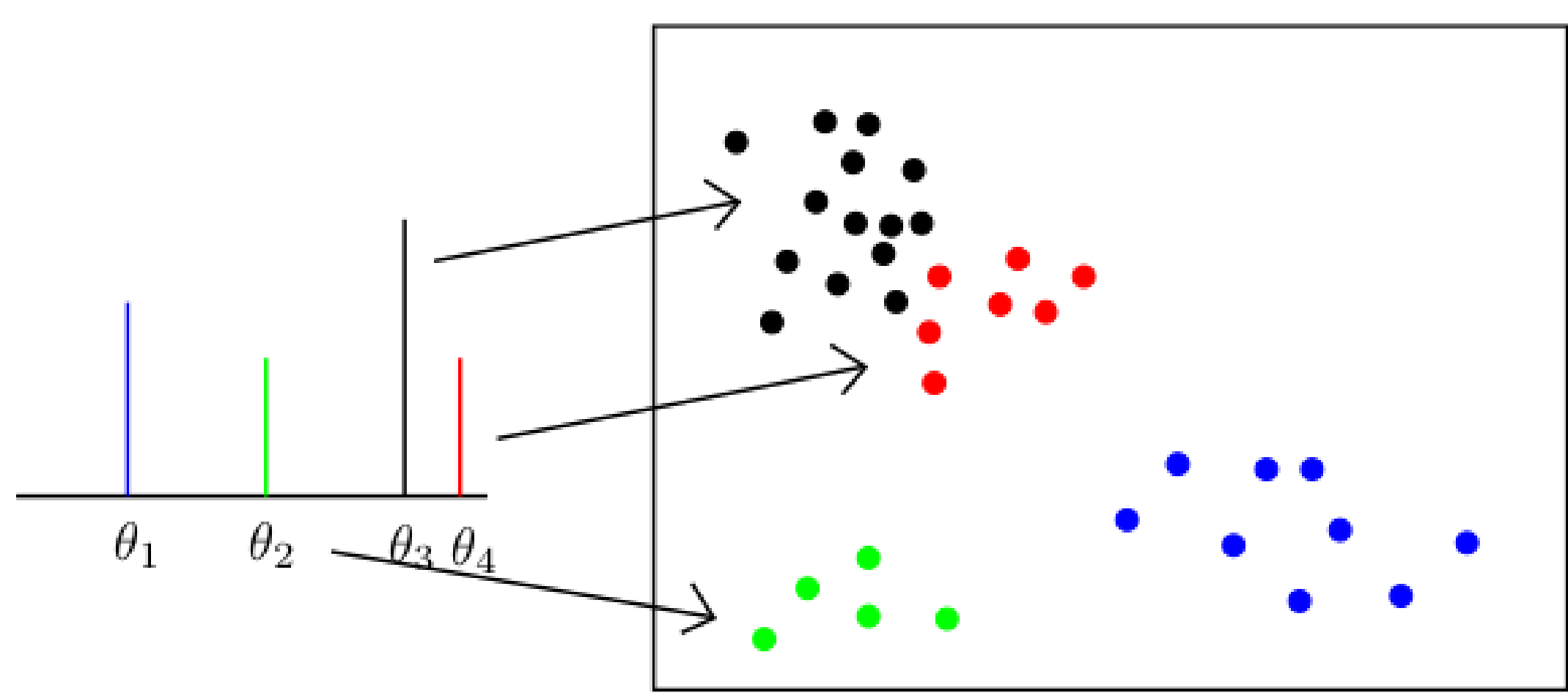
$$P = \sum_{i=1}^{\infty} p_i \delta_{\theta_i}$$



→ **Dirichlet process**  $DP(\alpha, G_0)$  (Ferguson, 1973)  
Predictive: Chinese Restaurant Process

$$\mathbb{P}(X_{n+1} \in \cdot | \mathbf{X}^n) = \frac{\alpha}{\alpha + n} G_0 + \frac{1}{\alpha + n} \sum_{j=1}^{k_n} n_j \delta_{X_j^*}$$

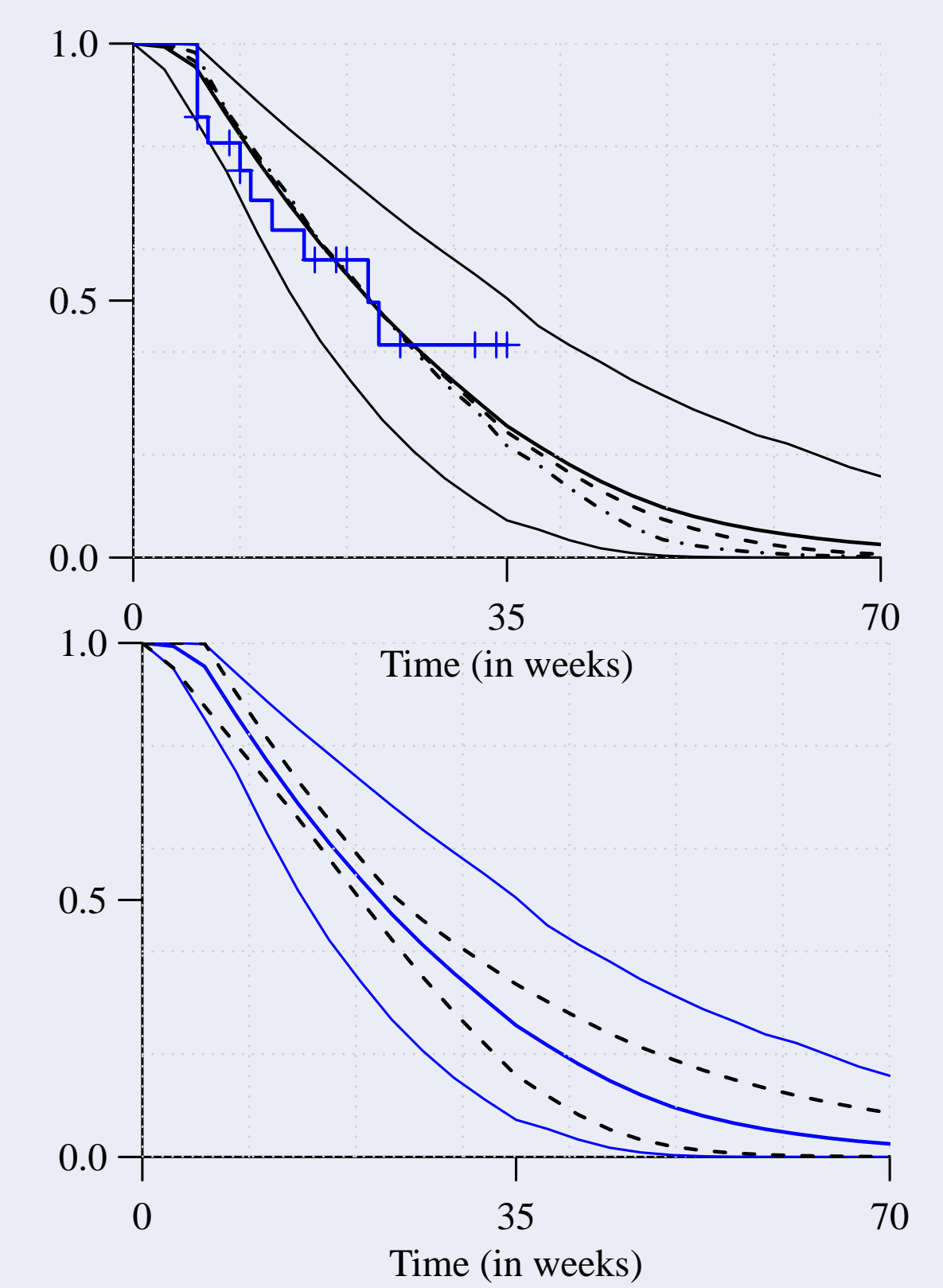
→ Or for varying  $\mathbb{P}(X_{n+1} \text{ new}) \dots \curvearrowright$



### ③ Survival Analysis

#### Bayesian hazard mixture (Arbel et al., 2016c)

- Data are (remission) times possibly censored
- Prior on **hazard rate**  $h(t)$  for every time  $t$
- Induces prior on **survival function**  $S(t)$
- Availability of **post. mean, median, mode**
- Smooth estimator VS Kaplan-Meier
- Proper uncertainty quantification



### Open Questions

- How to best use underlying **clustering**? (Wade and Ghahramani, 2015)
- Find **consistent** estimator of **number of clusters**: posterior inconsistent (Miller and Harrison, 2014), what about posterior mode?
- Devise efficient **posterior sampling**, truncation error (Arbel and Prünster, 2016)

### References & Collaborators

- Arbel, J., Favaro, S., Nipoti, B., and Teh, Y. W. (2016a). Bayesian nonparametric inference for discovery probabilities: credible intervals and large sample asymptotics. *Statistica Sinica*.
- Arbel, J., Kon Kam King, G., and Prünster, I. (2016b). Bayesian nonparametric modelling of species sampling distributions. *In preparation*.
- Arbel, J., Lijoi, A., and Nipoti, B. (2016c). Full Bayesian inference with hazard mixture models. *Computational Statistics & Data Analysis*.
- Arbel, J., Mengersen, K., Raymond, B., Winsley, T., and King, C. (2015). Application of a Bayesian nonparametric model to derive toxicity estimates based on the response of Antarctic microbial communities to fuel contaminated soil. *Ecology and Evolution*.
- Arbel, J., Mengersen, K., and Rousseau, J. (2016d). Bayesian nonparametric dependent model for partially replicated data: the influence of fuel spills on species diversity. *Annals of Applied Statistics*.
- Arbel, J. and Prünster, I. (2016). A moment-matching Ferguson & Klass algorithm. *Statistics and Computing*.
- Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*.
- Miller, J. W. and Harrison, M. T. (2014). Inconsistency of Pitman-Yor process mixtures for the number of components. *The Journal of Machine Learning Research*.
- Wade, S. and Ghahramani, Z. (2015). Bayesian cluster analysis: Point estimation and credible balls. *arXiv*.

### ① Species Modeling

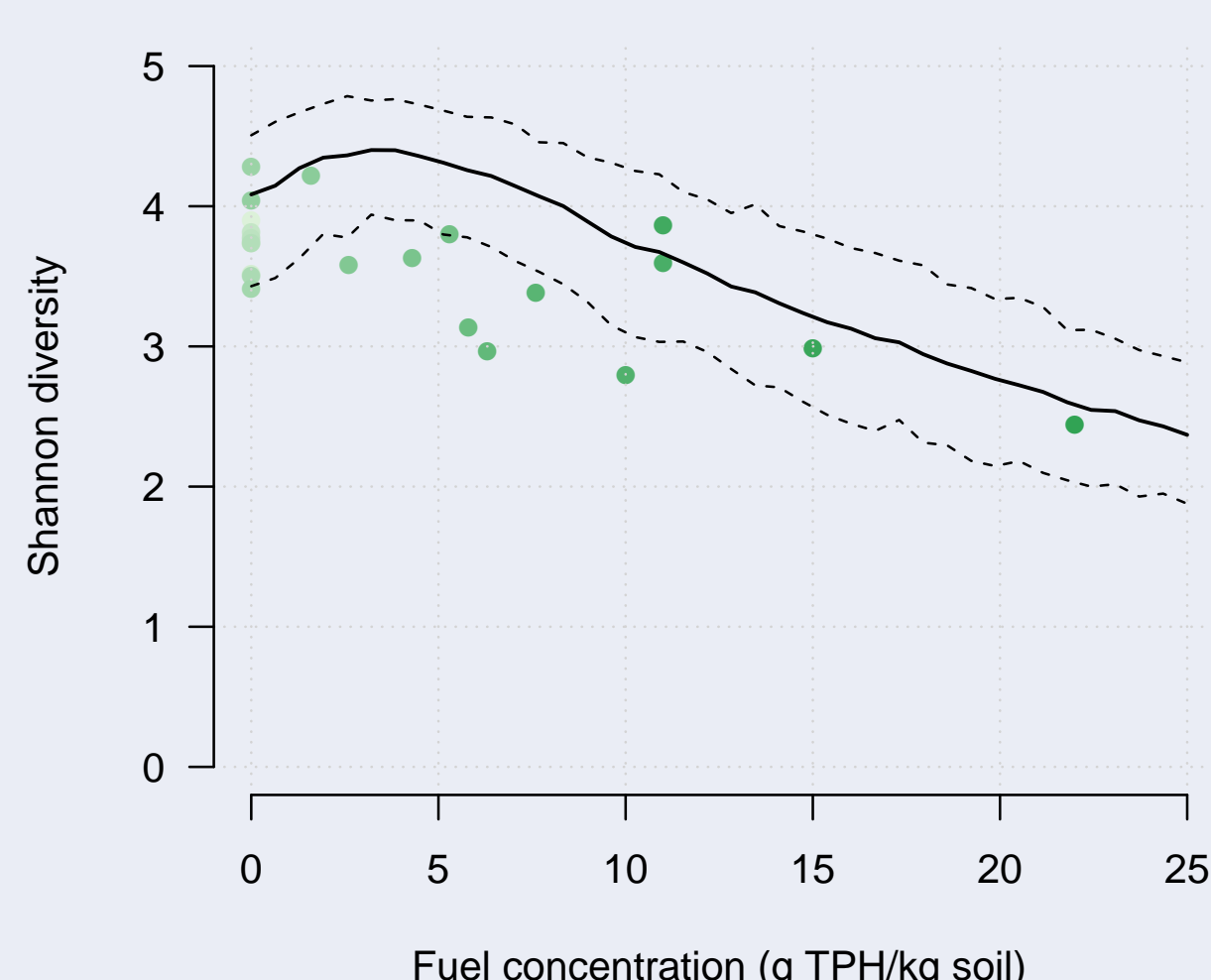
Data can be species, microbes, words, genes...

#### Discovery probabilities (Arbel et al., 2016a)

- Estimation of  **$\ell$ -discovery**
- $D_\ell = \mathbb{P}(X_{n+1} \text{ is a species seen } \ell \text{ times})$
- Comparison with Good-Turing estimator
- Closed form posterior and estimators
- Uncertainty quantif., unavailable for GT
- 2nd order (fast) approximations

#### Diversity in ecology (Arbel et al., 2015, 2016d)

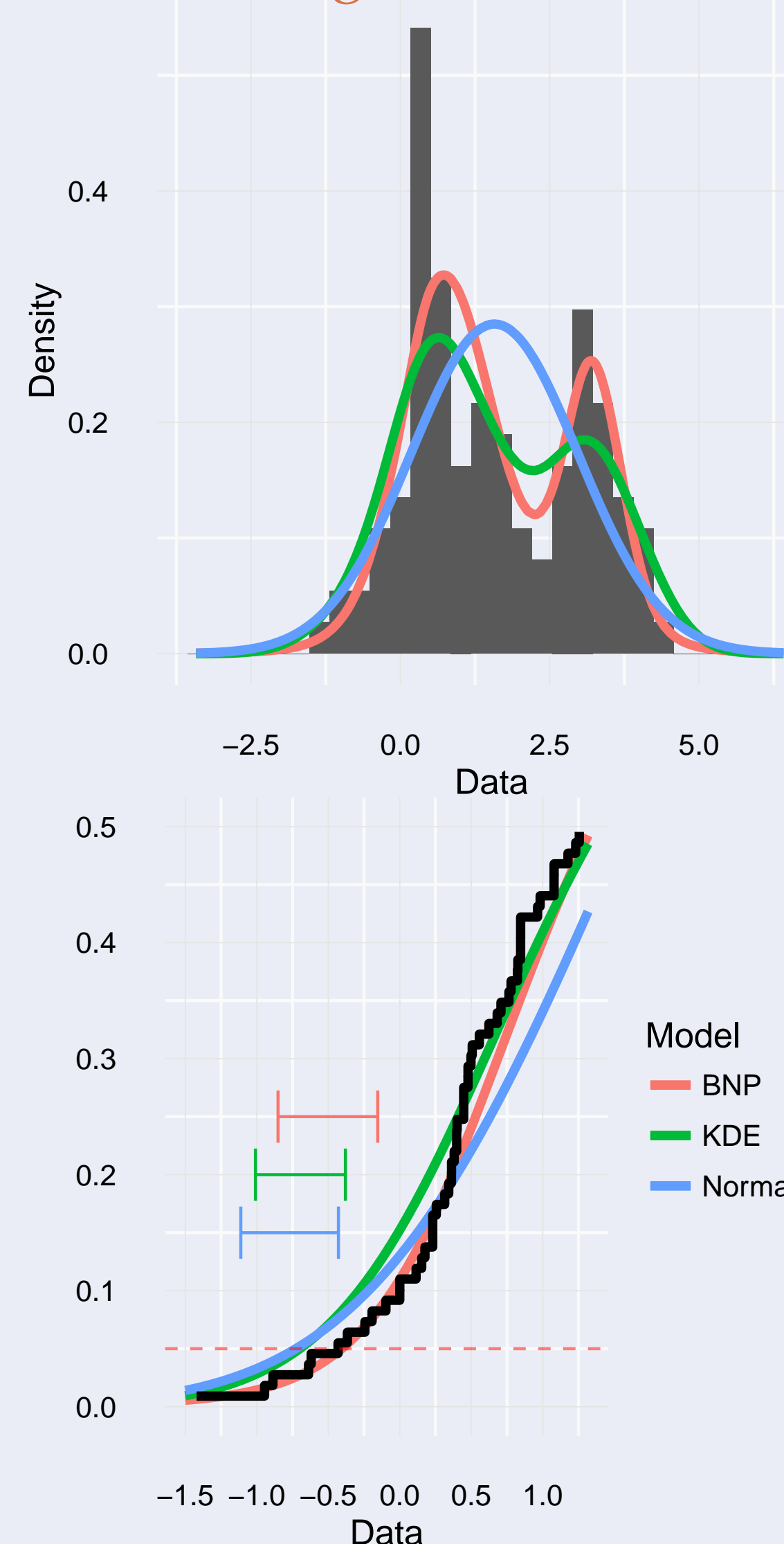
- Assess impact of pollution on microbial community via study of **diversity**
- $Div = -\sum_i p_i \log p_i$
- Model detects an **hormetic effect**
- Uncertainty quantification
- Prediction across full range of covariates



### ② Density Estimation

#### Ecological risk assessment (Arbel et al., 2016b)

- Data are species critical effect concentrations (CEC), possibly censored
- Estimation of **species sensitivity distribution (SSD)**, the density of CEC
- Safe concentration** which protects most of the species: **5th percentile of the SSD (HC<sub>5</sub>)**
- Very moderate sample sizes,  $\sim 10 - 50$
- BNP describes well **variability** of the data, without being prone to over-fitting
- Species **clustering** as an outcome



### Funding

This work is supported by the European Research Council (ERC) through StG "N-BNP" 306406.