

A Bayesian Nonparametric Approach to Ecological Risk Assessment

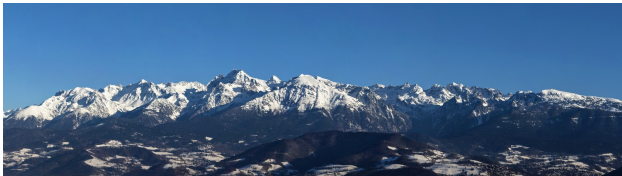
Julyan Arbel¹, Guillaume Kon Kam King³, Igor Prünster³

¹Inria Grenoble - Rhône-Alpes & Université Grenoble Alpes, France

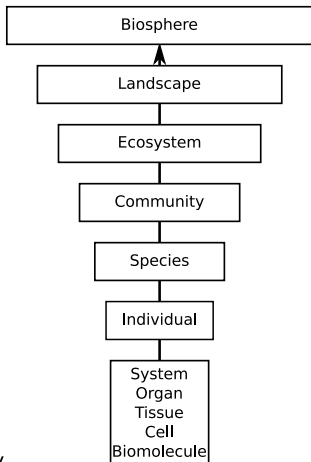
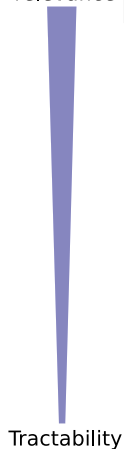
²Collegio Carlo Alberto, Moncalieri & University of Torino, Italy

³Bocconi University, Milan, Italy

Workshop on Statistical Methods for Post Genomic Data,
Université de Montpellier, France, 11-12 January 2018



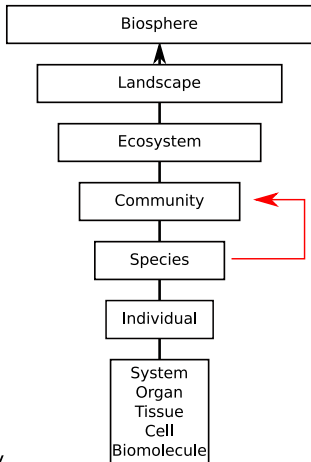
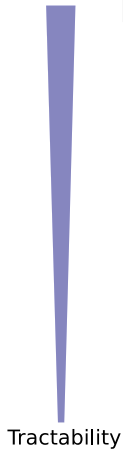
Environmental
relevance



Studying the effect of
contaminants
on ecosystems

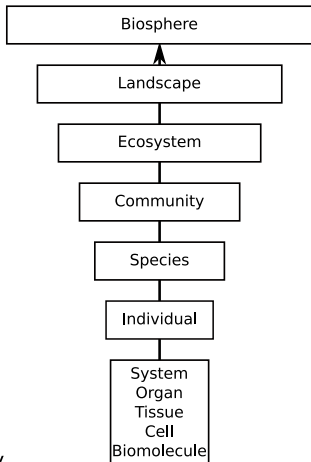
- pesticides
- hospital waste (effluent)
- heavy metals
- ...

Environmental
relevance



Extrapolation of effects

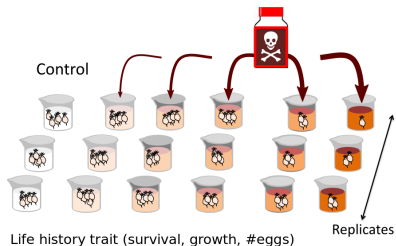
Environmental
relevance



Tractability

Experiments

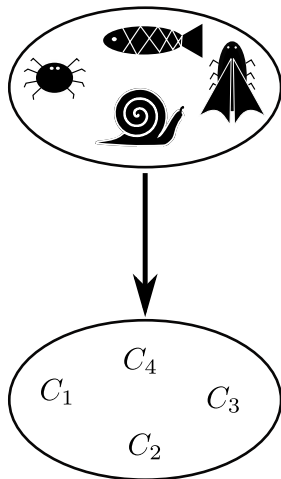
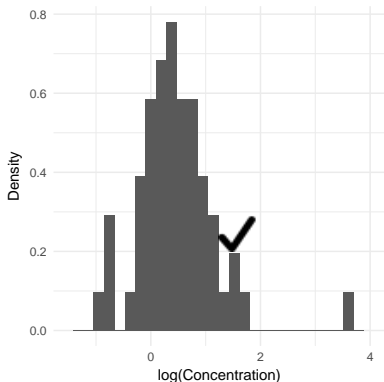
Increasing concentrations of contaminant/stressor



High costs for data acquisition

Typical sample size $\in [10, 15]$

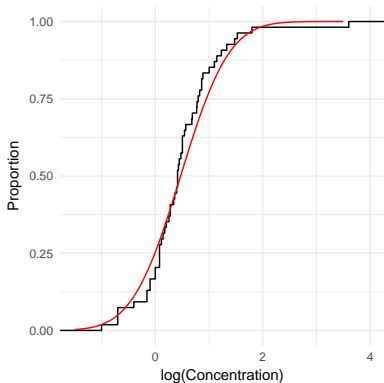
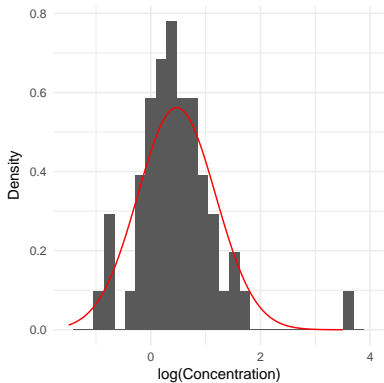
Schematic representation of the classical method



SSD = Species Sensitivity Distribution

Schematic representation of the classical method

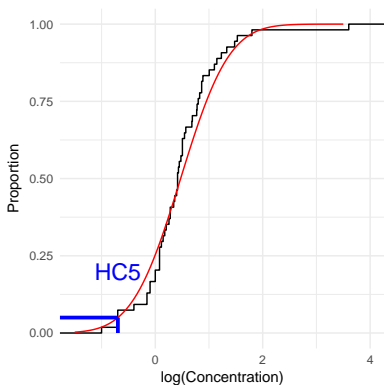
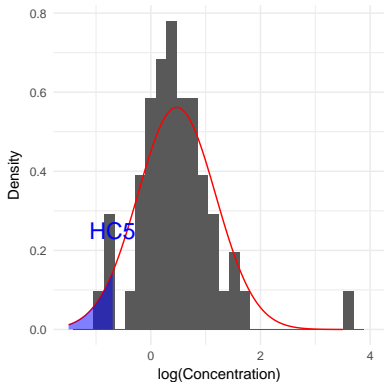
Class. method = normal SSD



SSD = Species Sensitivity Distribution

Schematic representation of the classical method

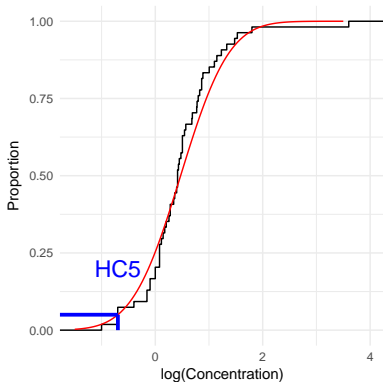
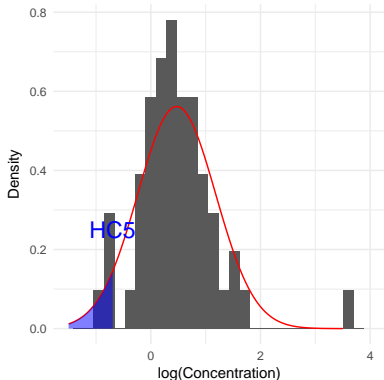
Normal SSD and HC_5



$HC_5 = 5\text{th percentile}$

Schematic representation of the classical method

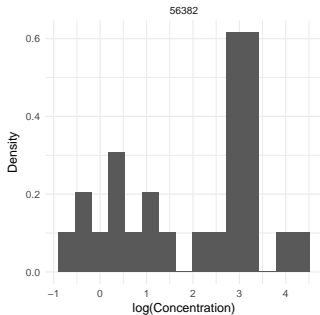
Normal SSD and HC_5



Seems rudimentary ?

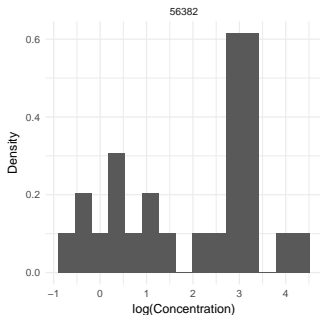
- Normal SSD is the reference method
- Widely used (EU, US, China, Australia, South Africa, etc.)

Normal SSD may be inappropriate



- Pesticides often target specific species
- Species naturally separate into groups

Normal SSD may be inappropriate



Existing solutions to deal with non-normal data:

- Finite normal mixture model: arbitrary
- Distribution-free approaches using order statistics: Need large datasets, which are uncommon
- Kernel Density Estimate with asymptotically optimal bandwidth: Most recent proposal¹, but we can do better

1

¹Wang, Y., et al. (2015). Non-parametric kernel density estimation of species sensitivity distributions in developing water quality criteria of metals. Environmental Science and Pollution Research, 22(18)

A Bayesian nonparametric approach

We propose a BNP mixture model:

Data : (x_1, \dots, x_n) , param $\theta = (\mu, \sigma)$

$$G \sim \text{NRM}$$

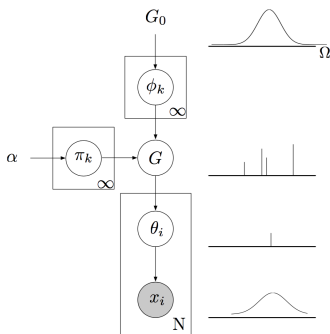
$$\theta_i = (\mu_i, \sigma_i) | G \sim G$$

$$x_i | \mu_i, \sigma_i \sim \mathcal{N}(\mu_i, \sigma_i)$$

where G is discrete and induces ties.

Generalisation of the Dirichlet Process Mixture (DPM) model: *Normalised Random Measures*. [More flexible](#).

Inference via Ferguson & Klass algorithm with a mix of Gibbs and MCMC (R package `BNPdensity`²)



²Barrios, E., Lijoi, A., Nieto-Barajas, L. E., & Prünster, I. (2013). Modeling with Normalized Random Measure Mixture Models. *Statistical Science*, 28(3)

Dirichlet, Pitman–Yor, Normalized Random Measure

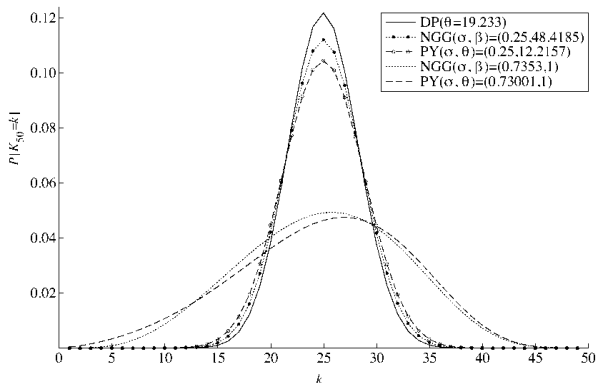
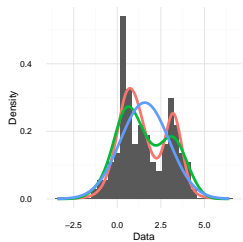
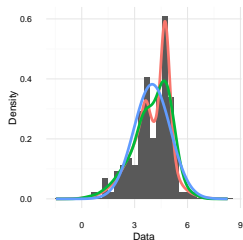


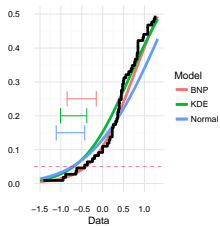
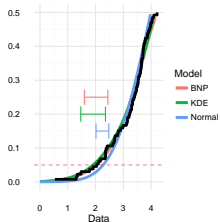
Figure : Prior distributions on the number of clusters corresponding to the Dirichlet (DP), the Pitman–Yor (PY) and the normalized generalized gamma (NGG) processes. The values of the parameters are set in such a way that $E(K_{50}) = 25$.

Example on real data

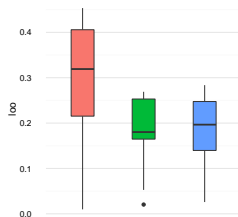
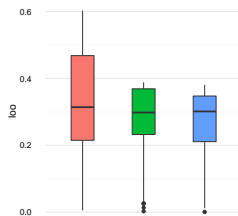
Density



CDF (zoom)

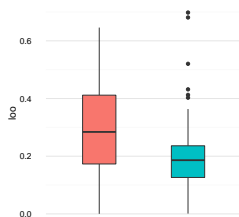
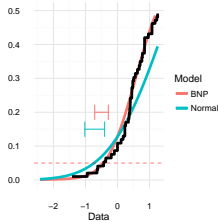
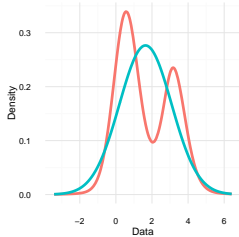
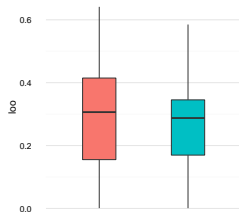
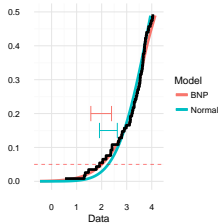
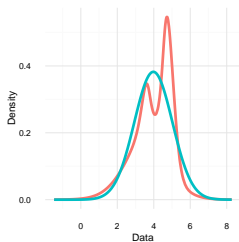


Leave-One-Out

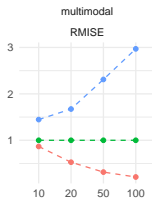
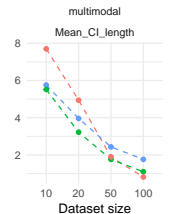
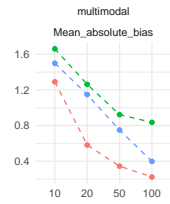
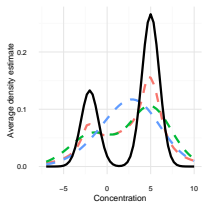
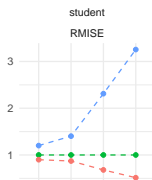
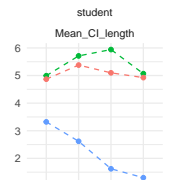
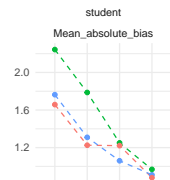
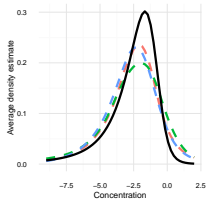
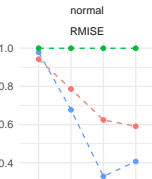
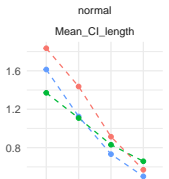
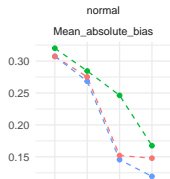
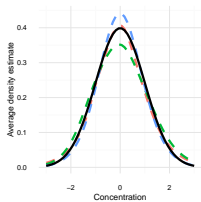


Extension to censored data data

Censored data are common in ecotoxicology.



Systematic test on simulated data



Model
- BNP
- KDE
- Normal

Added value of the BNP-SSD:

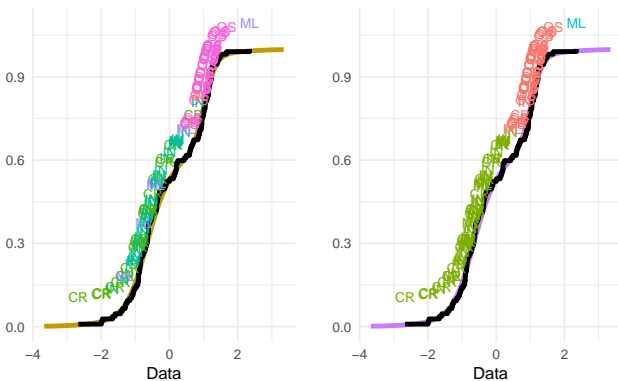
- The BNP-SSD is **more flexible** than the KDE SSD, but no less robust.
- The BNP-SSD can work well with **small samples**.
- The BNP-SSD can be extended to **censored data**.

Moreover:

- a normal mixture model induces a **clustering of the data**.
- what do these cluster represent ?
- are they biologically meaningful ?

Comparing the clustering with meta data

Fenthion

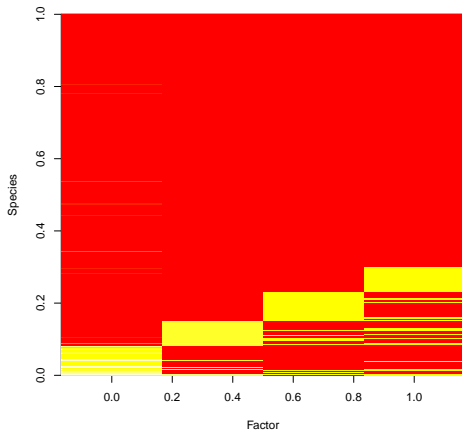


Left: Colored by **major taxon**
(fish, insect, ...)

Right: Colored by **cluster**

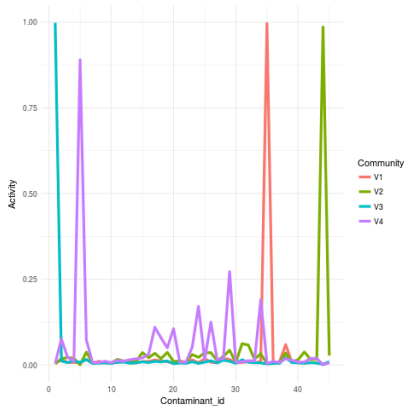
Community detection via non negative tensor factorisation

Quick empirical analysis of the groups
Feature extraction reveal 4 structures



Activity pattern of the extracted structure

These structure are active only for certain contaminants



- Further study of the clusters: additional meta data on the contaminants, species
- Better than ad-hoc feature extraction: [Hierarchical BNP model](#)
- Clustering in higher dimensions: [Identify species by more than one value by using raw data](#)

Bayes in Grenoble

- BIG seminar:

<https://sites.google.com/view/bigseminar/>

Thank you for your attention !

<http://www.julyanarbel.com/speaking>

$$C_i | \mu_i, \sigma_i \sim \mathcal{N}(\mu_i, \sigma_i)$$

$$(\mu_i, \sigma_i) | \tilde{P} \sim \tilde{P}$$

$$\tilde{P} \sim \text{NRMI}$$

$$L(\theta) = \prod_{i=1}^{N_{nc}} f(C_i | \theta)$$

$$\times \prod_{j=1}^{N_{lc}} \left(F(C_j^{up} | \theta) \right)$$

$$\times \prod_{k=1}^{N_{rc}} \left(1 - F(C_k^{low} | \theta) \right)$$

$$\times \prod_{l=1}^{N_{ic}} \left(F(C_l^{up} | \theta) - F(C_l^{low} | \theta) \right)$$