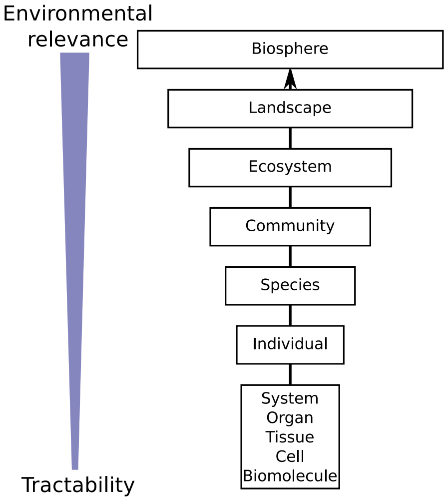
Bayesian Nonparametric Density Estimation in Ecotoxicology

Guillaume KON KAM KING[1,2]
Julyan ARBEL[3]
Igor PRÜNSTER[4]

[1]ESOMAS Department, University of Torino, Italy
[2]Collegio Carlo Alberto, Torino, Italy
[3]INRIA Grenoble - Rhône-Alpes & Université Grenobles Alpes, France
[4]Department of Decision Sciences, BIDSA and IGIER, Bocconi University, Milan, Italy

June 13th 2019, AppliBUGS

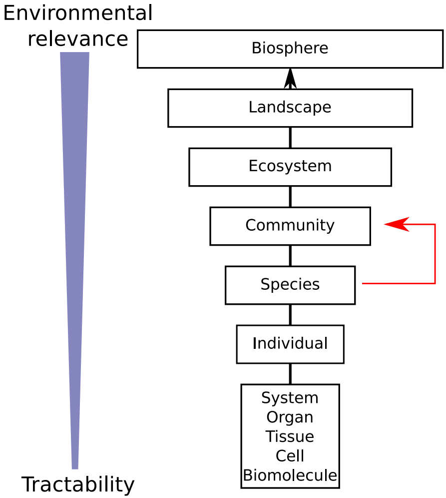
Ecotoxicology



Studying the effect of contaminants on ecosystems

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

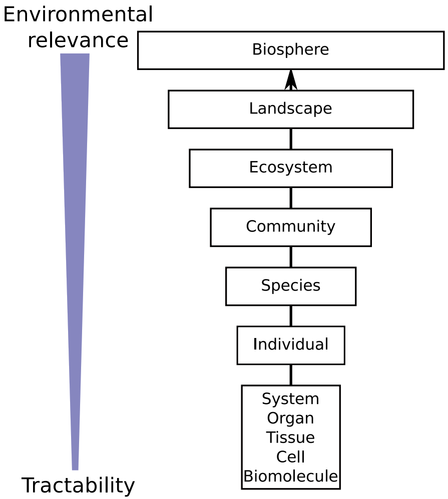
Ecotoxicology



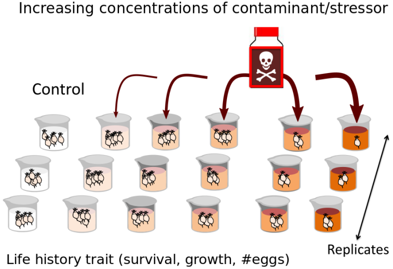
Our focus today:

Extrapolation of effects from species level to community level

Goal: find a *safe* level of contaminant (environmental regulation)



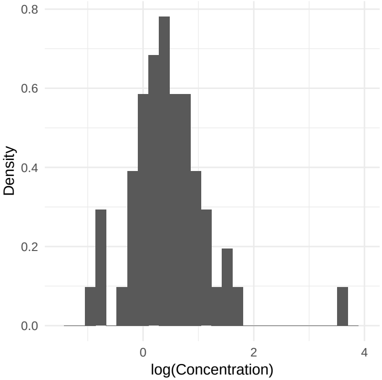
Experimental data

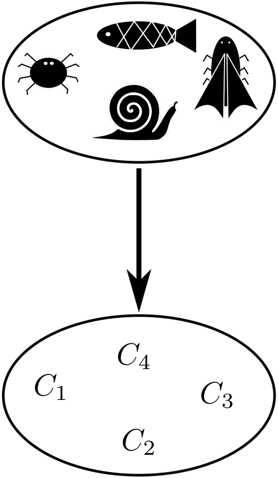


High costs for data acquisition

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

Schematic representation of the classical method

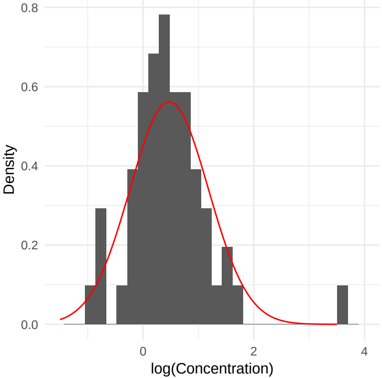


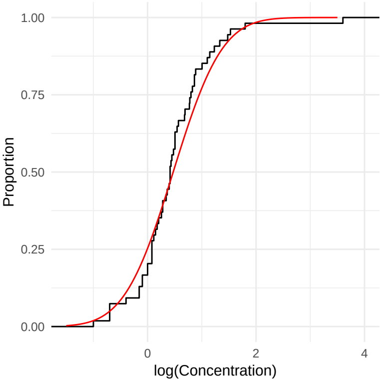


SSD = Species Sensitivity Distribution

Schematic representation of the classical method

Class. method = normal SSD

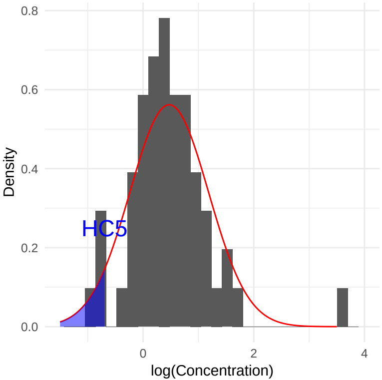


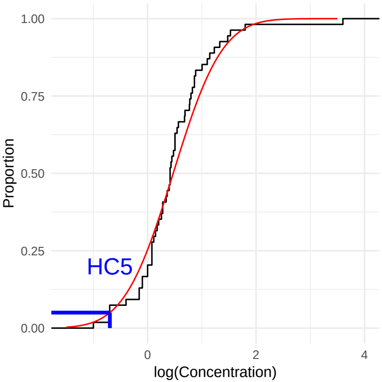


SSD = Species Sensitivity Distribution

Schematic representation of the classical method

Normal SSD and HC_5

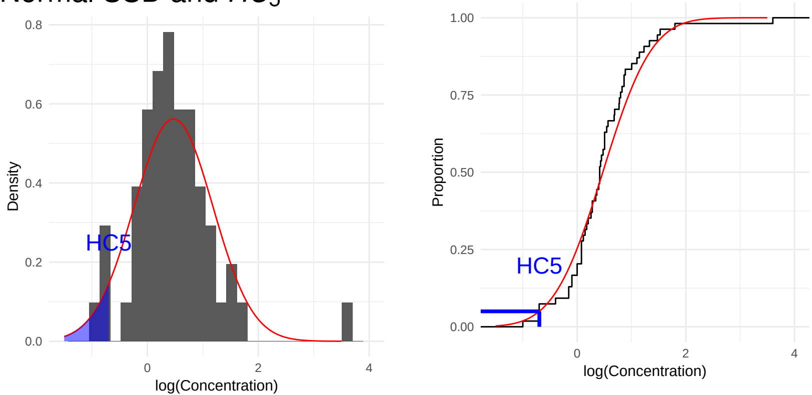




HC_5 = 5th percentile

Schematic representation of the classical method

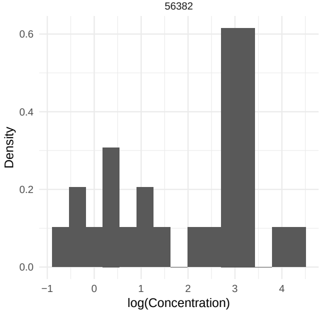
Normal SSD and HC_5



Seems rudimentary ?

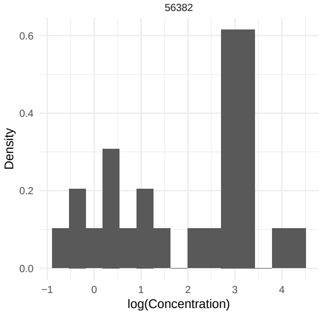
- Normal SSD is the reference method
- Widely used (EU, US, China, Australia, South Africa, etc.)

Normal SSD may be inappropriate



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

Normal SSD may be inappropriate



1

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 proposal[1], but we can do better

[1] Wang, 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 Non Parametric approach

We propose a BNP normal mixture model.
Let us denote the data as $(C_1, \dots, C_n)$

$$\begin{aligned} C_i|\mu_i, \sigma_i &\sim \mathcal{N}(\mu_i, \sigma_i) \\ (\mu_i, \sigma_i)|\tilde{P} &\sim \tilde{P} \\ \tilde{P} &\sim \text{NRMI} \end{aligned}$$

where $\tilde{P}$ is discrete and induces ties.

Generalisation of the Dirichlet Process Mixture (DPM) model: *Normalised Random Measure with Independent increments* (NRMI). More flexible than the DPM.

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

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

Dirichlet versus NRMI (and Pitman-Yor process)

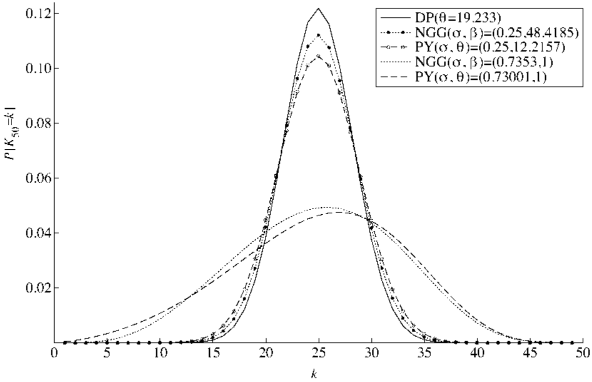


Figure : Prior distributions on the number of clusters corresponding to the Dirichlet (DP), the Pitman–Yor (PY) and the normalized generalized gamma (NGG) processes. The values of the parameters are set in such a way that $\mathrm{E}(K_{50}) = 25$.

NRMI: tractable class of processes obtained by normalisation of Completely Random Measures (NGG is an NRMI)

Prior specification

NRMI **do not require conjugacy** !

We use a Normalised-stable process with stability parameter fixed to 0.4

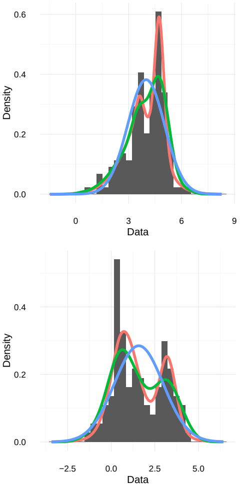
Data is log-transformed, scaled and centered.

- Uniform base measure for the σ_i: $\sigma_i \sim \mathcal{U}(0.1, 1.5)$. Because the data is scaled, $\sigma_i \leq 1$ and there is no reason to have very small clusters[3].
- Normal base measure for the μ_i: $\mu_i \sim \mathcal{N}(\phi_1, \phi_2)$ with (ϕ_1, ϕ_2) given conjugate hyper-priors.

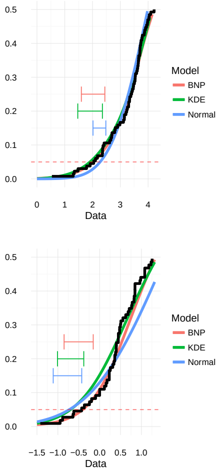
[3]Truncated normal prior gives the same results

Example on real data

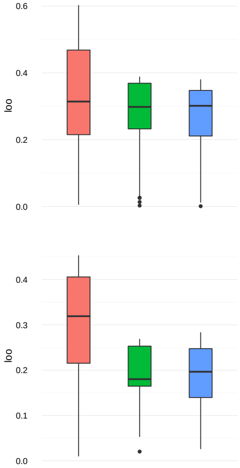
Density



CDF (zoom)

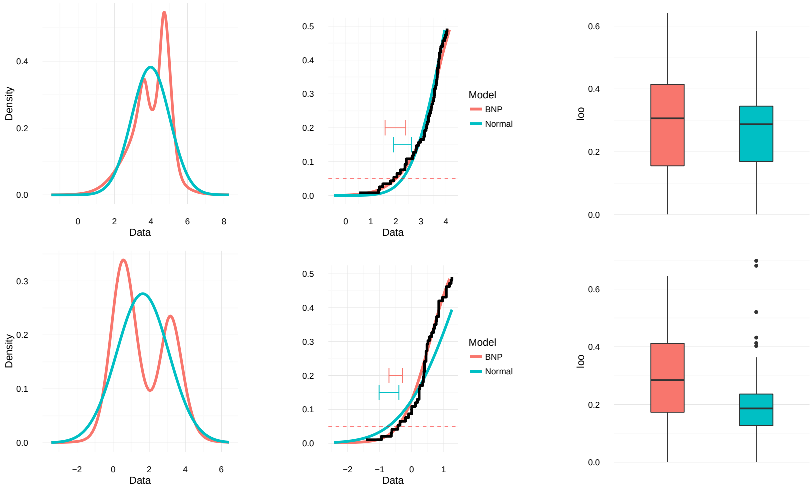


Leave-One-Out

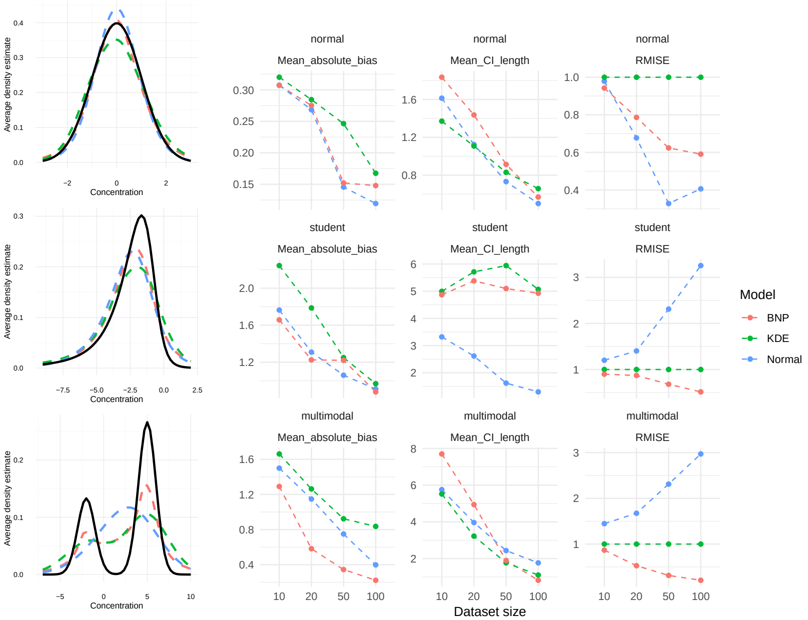


Extension to censored data data

Censored data are common in ecotoxicology.



Systematic test on simulated data



Summary for density and quantile estimation

Added value of the BNP-SSD:

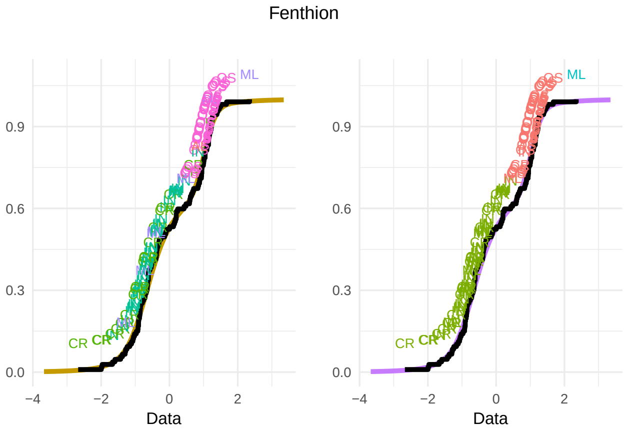
- The BNP-SSD is more flexible than the KDE SSD, but no less robust.
- The BNP-SSD can work well with small samples.
- The BNP-SSD can be extended to censored data.

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

```
## Error in readRDS(path) : unknown input format
```

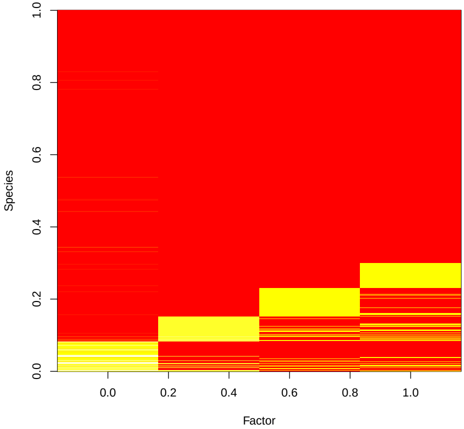


Left: Colored by major taxon (fish, insect, ...)

Right: Colored by cluster

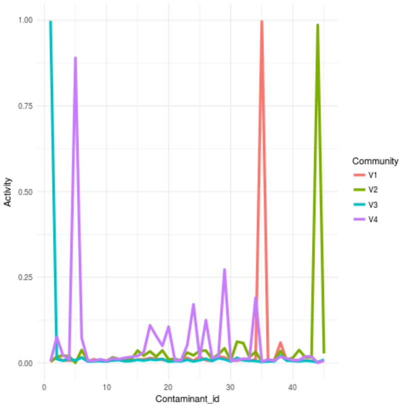
Community detection via non negative tensor factorisation

Quick empirical analysis of the groups
Feature extraction reveal 4 structures



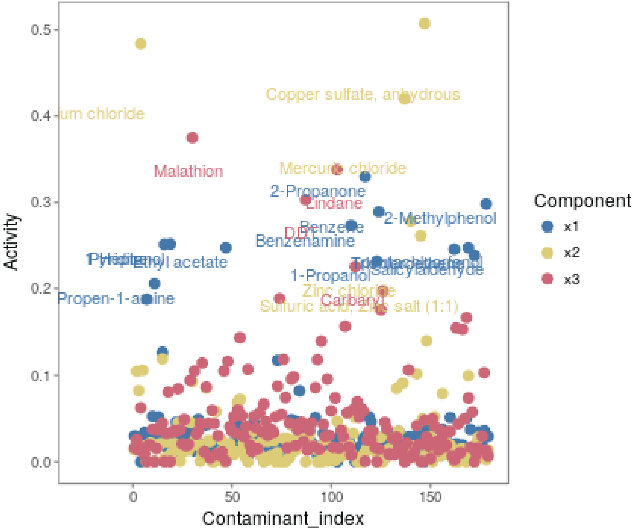
Activity pattern of the extracted structure

These structure are active only for certain contaminants



Activity pattern of the extracted structure

These structure are active only for certain contaminants



Future work

- 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

Thank you for your attention !

sites.google.com/site/guillaumekonkamking

Extension to censored data data: censored likelihood

$$
\begin{aligned}
C_i | \mu_i, \sigma_i &\sim \mathcal{N}(\mu_i, \sigma_i) \\
(\mu_i, \sigma_i) | \tilde{P} &\sim \tilde{P} \\
\tilde{P} &\sim \text{NRMI}
\end{aligned}
$$

$$
\begin{aligned}
L(\theta) = \prod_{i=1}^{N_{nc}} f(C_i|\theta) \\
\times \prod_{j=1}^{N_{lc}} \Big(F(C_j^{up}|\theta) \Big) \\
\times \prod_{k=1}^{N_{rc}} \Big(1 - F(C_k^{low}|\theta) \Big) \\
\times \prod_{l=1}^{N_{ic}} \Big(F(C_l^{up}|\theta) - F(C_l^{low}|\theta) \Big)
\end{aligned}
$$