Bayesian Nonparametric Density Estimation in Ecotoxicology

Guillaume KON KAM KING\textsuperscript{1,2}
Julyan ARBEL\textsuperscript{3}
Igor PRÜNSTER\textsuperscript{4}

\textsuperscript{1}ESOMAS Department, University of Torino, Italy
\textsuperscript{2}Collegio Carlo Alberto, Torino, Italy
\textsuperscript{3}INRIA Grenoble - Rhône-Alpes & Université Grenobles Alpes, France
\textsuperscript{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
- ...

Environmental relevance

Tractability
Our focus today:

**Extrapolation of effects from species level to community level**

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

Environmental relevance

- Biosphere
- Landscape
- Ecosystem
- Community
- Species
- Individual
- System
- Organ
- Tissue
- Cell
- Biomolecule

Tractability

Experimental data

Increasing concentrations of contaminant/stressor

Control

Life history trait (survival, growth, #eggs)

Replicates

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
$HC_5 = 5$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 proposal\(^1\), but we can do better

---

A Bayesian Non Parametric approach

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

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

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

\[ \tilde{P} \sim \text{NRMI} \]

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


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

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 $E(K_{50}) = 25$.

**NRMI:** tractable class of processes obtained by normalisation of Completely Random Measures (NGG is an NRMI)
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 $\sigma_i$: $\sigma_i \sim \mathcal{U}(0.1, 1.5)$. 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 $\mu_i$: $\mu_i \sim \mathcal{N}(\phi_1, \phi_2)$ with $(\phi_1, \phi_2)$ given conjugate hyper-priors.

---

\(^3\)Truncated normal prior gives the same results
Example on real data

Density

CDF (zoom)

Leave-One-Out

Data

Density

0.0
0.2
0.4
0.6
0
3
6
9

Data

0.0
0.2
0.4
0.6
0 1 2 3 4

Data

Model
BNP
KDE
Normal

Leave-One-Out

loo

Model
BNP
KDE
Normal

8 / 17
Censored data are common in ecotoxicology.
Systematic test on simulated data

![Graphs showing concentration vs. average density estimate for different models (normal, student, multimodal) across various dataset sizes.](image-url)
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 structures 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!
Extension to censored data data: censored likelihood

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

\[ \times \prod_{j=1}^{N_{lc}} \left( F(C_{j}^{up}|\theta) \right) \]

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

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

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

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

\[ \tilde{P} \sim \text{NRMI} \]