Contaminated Model for Species Sampling

M. Barin *V. Giorgi* *A. Grignani*
C. Laurenti Argento *F. E. Quadri* *M. Tracanella*

Politecnico di Milano

Tutor: Riccardo Corradin

November 11, 2022

Species Sampling

1. IDEA
   - A random sample taken from a population.

2. WHAT ARE THE PROBLEMS TO SOLVE
   - how do we estimate the number of species in the population?
   - how do we estimate the probability of discovering a new species in one additional sample?

Remarkable Examples

- Microbiological studies
- Linguistic
- Genetics
- Observational ecological studies

It is common that data are contaminated

An additional problem: Contamination

Contaminated Model

1. Contamination derives from human errors: frequency one.
2. Rare Events can also show up with frequency one.
3. Our goal is distinguish between contamination and events.

Accounting for the contamination in the model requires to generalize the models commonly used in literature

Solution

Gibbs-type Prior

Let $\{X_i\}_{i\geq 1}$ be a sequence of exchangeable observations. We call this sequence a species sampling sequence if there exists a random probability measure $\widetilde{p}$ such that $X_i \sim \widetilde{p}$, where:

$$\widetilde{p} = \sum_{j\geq 1} p_j \delta_{Z_j} + \left(1 - \sum_{j\geq 1} p_j\right) P_0$$

- $\{p_j\}_{i\geq 1}$: sequence of random weights
- $\{Z_j\}_{i\geq 1}$: sequence of random atoms
- P_0: contaminant (diffuse) probability measure

We define $\widetilde{p}$ a non-proper prior when: $\sum_{j\geq 1} p_j < 1$

Solution

The Non-proper Prior

1. We will show how non-proper models are particularly suited to take into account contaminated observations or more generally observations with frequency one.
2. Remember: the component P_0 generates singleton blocks, which are called dust in the probabilistic literature on random partitions.

Benefits

1. Takes into account observations with frequency 1.
2. Their predictive structure (w.r.t. the non contaminated prior).
3. Maintains the analytical tractability of non contaminated priors.

Solution

Contaminated Gibbs-type Priors

1. Gibbs-type priors are predominant priors in species sampling problems.
2. We introduce a new subfamily by including a contaminant component.
3. We will call this family: Contaminated Gibbs-type priors.

Solution

Definition

$$\widetilde{p} = \beta\widetilde{q} + (1-\beta)P_0 \qquad \beta \in [0,1]$$

- $\widetilde{p}$: contaminated Gibbs-type prior
- $\widetilde{q}$: Gibbs-type prior
- P_0: contaminant probability measure
- β: weight

The Project

Our Path

1. We hope to show the advantages of using these types of priors compared to the non contaminated ones.
2. One of the main advantages is that: it will become less probable to make wrong conclusions.
3. To validate our analytical conclusion regarding contaminated Gibbs priors, we will use the Global Biodiversity Information Facility dataset.