Contaminated Model for Species Sampling
Final Report

Matteo Barin
Viviana Giorgi
Alessandro Grignani
Carlo Laurenti Argento
Fosco Eugenio Quadri
Marco Tracanella
Tutor: *Riccardo Corradin*

February 2023

Abstract

In fields such as species sampling, the choice of the prior is of paramount importance. The most widely used prior is, in this sense, the Gibbs-type prior. In this paper we introduce a new extension of such prior by taking into account contamination. The idea is to have an efficient way to deal with excesses of observations with frequency one. We first introduce this new prior and then explain its effects on a given dataset by exploiting the posterior estimates produced. We will mainly focus on the contaminated Pitman-Yor process and we will show how our "contaminated" prior will lead to much better results in terms of predicitive inference, with respect to the "classic" formulation of the Gibbs-type prior.

1 Introduction

Taking into account contamination while studying datasets coming from fields such as microbiology, genetics, or any other science which includes some sort of species sampling has proven to be fundamental to avoid making wrong conclusions at the end of the analysis. To give a clearer idea of the problems to solve, let's consider a random sample from a population (e.g. a dataset coming from zoological studies): our aim is, on the one hand, to understand how to estimate the number of species in the population and on the other to estimate the probability of discovering a new species in an additional sample. In this context, it is common that data are contaminated, for example one can think of a mistake (i.e. a typo) made by a researcher while reporting the name of a sighted species.

When this happens, of course, one has to understand how to detect contamination within the dataset. The main problem is how to distinguish between observations considered rare events and contamination produced by random human errors, since both occur in a dataset with frequency one. In order to take into account contamination and so to develop a way to have a better analysis of a dataset, we had to generalize the models commonly used in literature. We will see how we did it in the next paragraphs.

2 Models

2.1 The Gibbs-type prior

We start by introducing the Gibbs-type Prior: let $\{X_i\}$ be a sequence of exchangeable observations. We call this sequence a species sampling sequence if there exists a random probability measure $\tilde{p}$ such that $X_i \sim \tilde{p}$, where:

$$\tilde{p} = \sum_j p_j \delta_{Zj} + (1 - \sum_j p_j)P_0 \tag{1}$$

The coefficient p_j is a sequence of random weights, Z_j is a sequence of random atoms and P_0 is the contaminant (diffuse) probability measure and it therefore generates singleton blocks. We say that $\tilde{p}$ is non-proper if $\sum_j p_j < 1$.

We will see that non-proper priors are particularly suitable when dealing with contaminated observations or more generally observations with frequency one. Gibbs-type priors are the predominant priors used in species sampling problems due to advantages such as: the predictive inference structure and also the fact that it maintains their analytical tractability. In this paper we introduce a new subfamily of such priors, which will contain a contamination component. We call this subfamily: the Contaminated Gibbs-type priors. We define such prior in this way:

$$\tilde{p} = \beta\tilde{q} + (1 + \beta)P_0 \tag{2}$$

where $\tilde{q}$ is a Gibbs-type prior, P_0 is the contaminant (diffuse) probability measure and β is a weight.

2.2 Contaminated Pitman-Yor model

The first step in creating our contaminated Pitman-Yor is studying the EPPF of the contaminated model. It is worth recalling that Gibbs-type random probability measures are typically characterized in terms of the exchangeable random partition induced by the data. More precisely, given a sample $X_{1:n} := (X_1, ..., X_n)$ from a species sampling model governed by a random probability measure $\tilde{p}$, the n observations are naturally partitioned into $K_n = k$ groups of distinct values, denoted as $X_1^*, ..., X_k^*$, with corresponding frequencies $(N_{n,1}, ..., N_{n,K_n}) = (n_1, ..., n_k)$. The exchangeable partition probability function (EPPF) corresponds to the probability of observing a specific partition of the data into clusters of distinct values. The EPPF is essential for computational purposes: indeed, it is the basic building block to derive suitable sampling schemes for posterior inference, also in the mixture model case.

Gibbs-type priors are proper species sampling models $\tilde{p}$ characterized by means of their sequence of EPPFs. In the case of our model, the expression of the EPPF is as follows:

$$\Pi_k^{(n)}(n_1, \ldots, n_k) \;=\; \prod_{i=m_1+1}^{k} (1-\sigma)_{n_i-1} \sum_{\bar{m}_1}^{m_1} \binom{m_1}{\bar{m}_1} \beta^{n-\bar{m}_1}(1-\beta)^{\bar{m}_1} \frac{\sigma^{k-\bar{m}_1(\vartheta/\sigma)_{k-\bar{m}_1}}}{(\vartheta)_{n-\bar{m}_1}} \tag{3}$$

Let us analyze the four main parameters of our model and what they represent in our computation, including the law we use to initialize them a priori.

- σ: discount parameter, $\sigma \in [0, 1]$, $\mathcal{L}(\sigma) \sim Beta(1, 1)$
  affects the distribution of a random partition arising from a Gibbs-type prior: when σ grows, the distribution favours partitions with few large clusters and a considerable number of small clusters, which is our aim when we study observations with frequency one.

- θ: a priori mass, $\mathcal{L}(\vartheta) \sim Gamma(2, 0.02)$
- β: fraction of contamination, $\mathcal{L}(\beta) \sim Beta(1, 1)$
- $\bar{m}_1$: estimated number of contaminated data, $\mathcal{L}(\bar{m}_1) \sim U(0, m_1)$ where m_1 is taken from our data.

In Algorithm 1 we can observe what is the expression for the posterior of our parameters.

```
Algorithm 1: Sampling scheme for contaminated Pitman-Yor model.
  [0] Set initial values for $\vartheta$, $\bar{m}_1$
    for $r = 1, \ldots, R$ do
      [1] Update $\sigma$ from
          $\mathcal{L}(\sigma \mid X_{1:n}, \vartheta, \bar{m}_1) \propto \mathcal{L}(\sigma)\sigma^{k-\bar{m}_1} \prod_{i=m_1+1}^{k} (1-\sigma)_{n_i-1} \frac{\Gamma(\vartheta/\sigma + k - \bar{m}_1)}{\Gamma(\vartheta/\sigma)}$
      [2] Update $\vartheta$ from
          $\mathcal{L}(\vartheta \mid X_{1:n}, \sigma, \bar{m}_1) \propto \mathcal{L}(\vartheta) \frac{\Gamma(\vartheta)\Gamma(\vartheta/\sigma + k - \bar{m}_1)}{\Gamma(\vartheta/\sigma)\Gamma(\vartheta + n - \bar{m}_1)}$
      [3] Update $\beta$ from $\mathcal{L}(\beta \mid X_{1:n}, \bar{m}_1) \propto \mathcal{L}(\beta)\beta^{n-\bar{m}_1}(1-\beta)^{\bar{m}_1}$
      [4] Update $\bar{m}_1$ from
          $\mathcal{L}(\bar{m}_1 | X_{1:n}, \sigma, \vartheta, \beta) \propto \binom{m_1}{\bar{m}_1} \beta^{n-\bar{m}_1}(1-\beta)^{\bar{m}_1}\sigma^{k-\bar{m}_1} \frac{\Gamma(\vartheta/\sigma + k - \bar{m}_1)}{\Gamma(\vartheta + n - \bar{m}_1)}$
    end for
```

For both σ and θ we use a Metropolis Hastings in order to sample a random value at each iteration of the algorithm. In Appendix A.1 we reported all the useful computations (that we made "by hand") in order to simplify, by including the full prior law, the expressions of the densities.

We have also computed the ratio of our Metropolis Hastings algorithm for both parameters. Considering a new function h:

$$h(\sigma) = (\alpha - 1 + k - \bar{m}_1)log(\sigma) + (\beta - 1)log(1-\sigma) + \sum_{i=m_1+1}^{k} (log\Gamma(n_i + \sigma) - log\Gamma(1-\sigma)) + log\Gamma(\vartheta/\sigma + k - \bar{m}_1) - log\Gamma(\vartheta/\sigma)$$

then our ratio for σ is: $log(r(\sigma^*, \sigma)) = h(\sigma^*) - h(\sigma)$.

In the case of θ, we consider a new function g:

$$g(\vartheta) = (\alpha - 1)log(\vartheta) - \beta\vartheta + log\Gamma(\vartheta) + log\Gamma(\vartheta/\sigma + k - \bar{m}_1) - log\Gamma(\vartheta/\sigma) - log\Gamma(\vartheta + n - \bar{m}_1)$$

Our ratio is:

$$log(r(\vartheta^*, \vartheta)) = g(\vartheta^*) - g(\vartheta)$$

In order to stabilize our results we have considered a logarithmic transformation for both these parameters, i.e. $\psi = log(\sigma) - log(1-\sigma)$ and $\lambda = log(\vartheta)$.

2.3 Pitman-Yor model

After focusing on the contaminated Pitman-Yor model, we also built the non contaminated Pitman-Yor model so as to have a term of comparison during model evaluation. To build the Pitman-Yor model, we retraced the steps done for the construction of the contaminated model but modified the starting function, namely the EPPF:

$$\Pi_k^{(n)}(n_1, \ldots, n_k) = \frac{\prod_{i=1}^{k-1} \vartheta + i\sigma}{(\vartheta + 1)_{n-1}} \prod_{i=m_1+1}^{k} (1-\sigma)_{n_i - 1} \tag{4}$$

As one can see from the formula, the EPPF we consider in this model has only two of the four parameters seen before, namely σ and ϑ, while the parameters related to the contaminated part of the model are not present. From the function (4) we derive the full conditionals of our parameters and we update their value as shown in Algorithm 2. As before we sample from the full conditionals with the Metropolis-Hastings algorithms, see detailed calculations in section A.2 of the Appendix.

Algorithm 2: Sampling scheme for Pitman-Yor model.

```
[0] Set initial values for $\vartheta$
for $r = 1, \ldots, R$ do
    [1] Update $\sigma$ from
        $\mathcal{L}(\sigma \mid X_{1:n}, \vartheta) \propto \mathcal{L}(\sigma) \prod_{i=1}^{k-1} \vartheta + i\sigma \prod_{i=m_1+1}^{k} (1-\sigma)_{n_i - 1}$
    [2] Update $\vartheta$ from
        $\mathcal{L}(\vartheta \mid X_{1:n}, \sigma) \propto \mathcal{L}(\vartheta) \frac{\Gamma(\vartheta + 1)}{\Gamma(\vartheta + n)} \prod_{i=1}^{k-1} \vartheta + i\sigma$
end for
```

3 Synthetic Data and Validation

In this section we illustrate the last step we had to do before using our models on a real dataset. We hereby focus on validating the models introduced in the previous sections on some synthetic datasets that we generated from the Pitman-Yor process and from the contaminated Pitman-Yor Process. We focused on how our models estimated the parameters of interest.

3.1 Synthetic Data

To generated the synthetic dataset we use the predictive distribution for the contaminated Pitman-Yor model, conditional to the latent variable J_i:

$$\begin{aligned} P(X_{n+1} \in dx | X_{1:n}, J_{1:m_1}) = (1-\beta) P_0(dx) + \beta \frac{\vartheta + (k - \bar{M}_{m_1})\sigma}{\vartheta + n - \bar{M}_{m_1}} Q_0(dx) \\ + \sum_{i=1}^{m_1} J_i \beta \frac{1-\sigma}{\vartheta + n - \bar{M}_{m_1}} \delta_{X_i^*}(dx) + \sum_{i=m_1+1}^{k} \beta \frac{n_i - \sigma}{\vartheta + n - \bar{M}_{m_1}} \delta_{X_i^*}(dx) \end{aligned} \tag{5}$$

To sample from this predictive distribution we consider distribution (5) as the sum of four probabilities that represent four different types of data that we can add to the dataset we are building. The first term is the probability that the new data is a contaminated singleton while the second term is the probability that the new data is a non contaminated singleton. We distinguish these two cases thanks to the latent variable J_i which is 0 when data is contaminated and 1 when data is not contaminated. The third one is the probability that the new data belongs to a cluster with only one element, and the last one is the probability that the new data belongs to a cluster with two or more elements. To create the dataset, at each iteration we add a new element and we also update the probabilities.

To sample from the Pitman-Yor Process we start from the same predictive distribution but with $\beta = 1$ fixed and without the term $\bar{M}_{m_1}$. We obtain this predictive distribution:

$$P(X_{n+1} \in dx | X_{1:n}) = \frac{\vartheta + k\sigma}{\vartheta + n} Q_0(dx) + \sum_{i=1}^{m_1} \frac{1-\sigma}{\vartheta + n} \delta_{X_i^*}(dx) + \sum_{i=m_1+1}^{k} \frac{n_i - \sigma}{\vartheta + n} \delta_{X_i^*}(dx) \tag{6}$$

In this case also the latent variable J_i has disappeared, since we no longer need to distinguish our singleton between contaminated and non contaminated.

3.2 Model Validation

After having created the contaminated and non contaminated datasets, we evaluated them with our models. We obtained different estimates for the parameters. In the case of the non contaminated model, the parameter estimation when the dataset was contaminated is affected by error. On the other hand, as we can observe in Table 1, the parameter estimation produced by the contaminated model is good when estimating both the contaminated dataset and also the non contaminated one.

			Posterior estimates		
Data	Model	Data parameters	$\hat{\sigma}$	$\hat{\vartheta}$	$\hat{\beta}$
PY	PY	($\sigma = 0.2$, $\vartheta = 100$)	0.16	114	-
PY	cPY	($\sigma = 0.2$, $\vartheta = 100$)	0.16	119.7	1.00
cPY	PY	($\sigma = 0.2$, $\vartheta = 100$, $\beta = 0.9$)	0.73	20.5	-
cPY	cPY	($\sigma = 0.2$, $\vartheta = 100$, $\beta = 0.9$)	0.23	108.7	0.9

Table 1: Summaries of the simulation study. First column: data generating process. Second column: model used to analyze the data. Third column: parameters of the data generating process. Fourth to sixth columns: posterior mean of the main parameters of the models averaged over 20 replications.

4 Real Dataset and Prediction

4.1 The North America Ranidae dataset

After having validated our model on the synthetic data, we began to look for a real dataset. We decided to focus on datasets in the field of ecology. Our hope was to observe that our ”contaminated” model made more accurate predictions that the previously introduced ”non contaminated” models.

Looking for a dataset proved to be challenging, since many of the dataset found seemed to be very ”clean” and not contaminated. The criterion we used to choose the dataset was to select a dataset that had a a high number of singletons, and to verify this we plotted the frequencies and then looked at how many singletons appeared in the dataset. We tried to restrict our research to datasets that had between 50.000 and 100.000 observations. After many tests, we finally decided to focus on the ”The North America Ranidae dataset”, a dataset regarding a vast number of families of frogs.

4.2 The posteriors of the parameters

As a first step, we runned our validated Metroplis-Hastings algorithm on the dataset. We observe in Table 2 how the posterior parameters (θ and σ) behave in the two cases. Then, in Figure 2 we observe the posterior distributions of θ and σ.

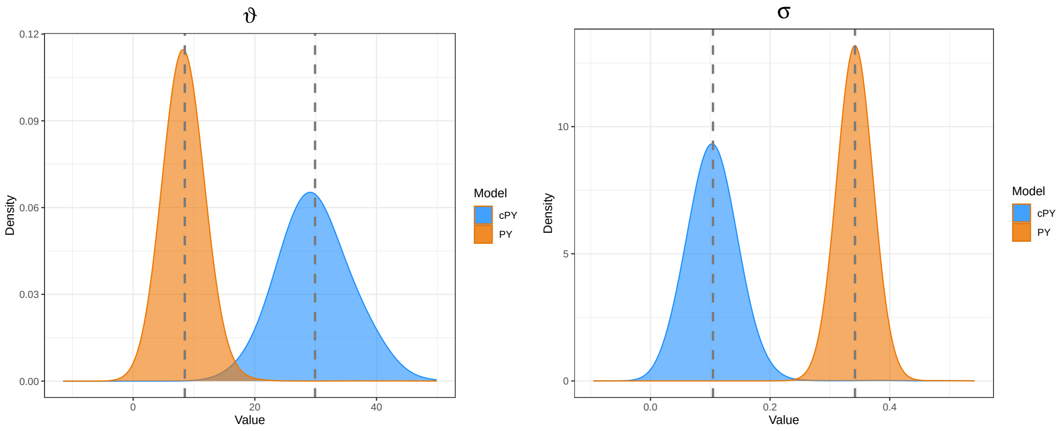

Figure 1: The posterior distribution of the parameters θ, σ

We can then observe how the posterior parameters θ and σ behave in the two cases. We plot in blue the posteriors of the ”contaminated” model and in orange the ones of the ”non contaminated” case. We notice the following differences: the ”non contaminated” model slightly underestimates the value of theta while it overestimates the value for sigma.

This could mean that: the presence of a large number of species observed only once leverages the estimation of the parameters in the ”non contaminated” model, while the use of a contamination component helps to obtain a much more suitable modeling of the data. This happens because some of the observations with frequency 1 are assigned to the ”contamination” component.

Posterior estimates		
Model	$\hat{\sigma}$	$\hat{\vartheta}$
PY	0.10	29.89
cPY	0.34	8.48

Table 2: The posterior estimates for $\hat{\sigma}$, $\hat{\vartheta}$

We then look at the frequency spectrum. On the x axis we have the number of elements of a given cluster, and on the y axis we have to number of clusters with that number of elements inside them. This clarifies how the presence of a large number of species observed only once leverages the estimation of the parameters in the Pitman-Yor model. The use of a contamination component helps to obtain a much more suitable modeling of the data. The orange band represents a 90% posterior credible interval for the PY model. The blue band represents a 90% posterior credible interval for the cPY model. The grey dashed line corresponds to the inflation induced by the "contamination".

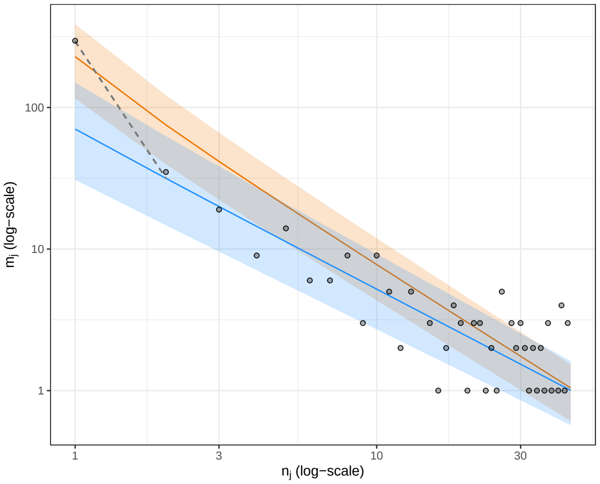

Figure 2: The frequency spectrum

4.3 Predictions on the real dataset

After having estimated the parameters, we moved to the prediction of the contaminated singletons inside the dataset. To do so, we considered an additional follow-up sample of size m of unobserved data. We focused our prediction on two different quantities:

1. $K_m^{(n-m)}$: the number of new species
2. $N_{m,1}^{(n-m)}$: the number of new species observed with frequency one

These are the formulas we used for the prediction in the two different cases. The main difference between the two formulas is (as expected) that the formulas for the cPY take into account parameters such as $\bar{M}_n$ and β, which obviously don't appear in the Pitman-Yor formulas.

The same conclusions can be drawn for both quantities of interest.

Posterior expected value of $N_{m,1}^{(n)}$ in the contaminated model:

$$E[N_{m,1}^{(n)}|X_1,\dots,X_n,\bar{M}_n] = \\ \sum_{l=0}^{m}\left[(m-l)(\vartheta+(k-\bar{M}_n)\sigma)\frac{(\vartheta+n-\bar{M}_n+\sigma)_{m-l-1}}{(\vartheta+n-\bar{M}_n)_{m-l}}+l\right]\binom{m}{l}(1-\beta)^l\beta^{m-l} \tag{7}$$

Posterior expected value of $N_{m,1}^{(n)}$ in the non contaminated model:

$$E[N_{m,1}^{(n)}|X_1,\dots,X_n] = m(\vartheta+k\sigma)\frac{(\vartheta+n+\sigma)_{m-1}}{(\vartheta+n)_m} \tag{8}$$

Posterior expected value of $K_m^{(n)}$ in the contaminated model:

$$E[K_m^{(n)}|X_1,\dots,X_n,\bar{M}_n] = \\ \sum_{l=0}^{m}\left[(k-\bar{M}_n+\vartheta/\sigma)\left(\frac{(\vartheta+n-\bar{M}_n+\sigma)_{m-l}}{(\vartheta+n-\bar{M}_n)_{m-l}}-1\right)+l\right]\binom{m}{l}(1-\beta)^l\beta^{m-l} \tag{9}$$

Posterior expected value of $K_m^{(n)}$ in the non contaminated model:

$$E[K_m^{(n)}|X_1,\dots,X_n] = (k+\vartheta/\sigma)\left(\frac{(\vartheta+n+\sigma)_m}{(\vartheta+n)_m}-1\right) \tag{10}$$

We can now observe the results of the predictions made (Tabel 3). The prediction error, for both N and K, is bigger in the case of the ”classic”, ”non contaminated” model. The error is, in both cases, almost 3 times as big. To make the prediction more accurate, we shuffled and then divided the dataset into two parts, with the first we try to predict the number of contaminated singletons in the second, and with the second we verify our predictions. We choose to divide the dataset into 80% prediction, 20% test and to run the simulation over 100 different shuffles of the dataset.

	Prediction Error	
Model	$N_{m,1}^{(n-m)}$	$K_m^{(n-m)}$
cPY	4.281	4.314
PY	11.956	11.102

Table 3: Prediction errors for the PY and cPY. Results averaged over 100 cross-validation shuffles.

We can conclude that taking into account the possible contamination of the dataset is crucial to decrease the prediction errors. We now look at the cross-validated distributions of the posterior expectation of the two predicted quantities. We compare the contaminated Pitman-Yor model with the Pitman-Yor model. It is clear that adding the contaminant measure in the model can be crucial also for its predictive properties, since the predictions made are very different. There is, infact, a remarkable difference between the posterior expectations of the same quantities under the two models.

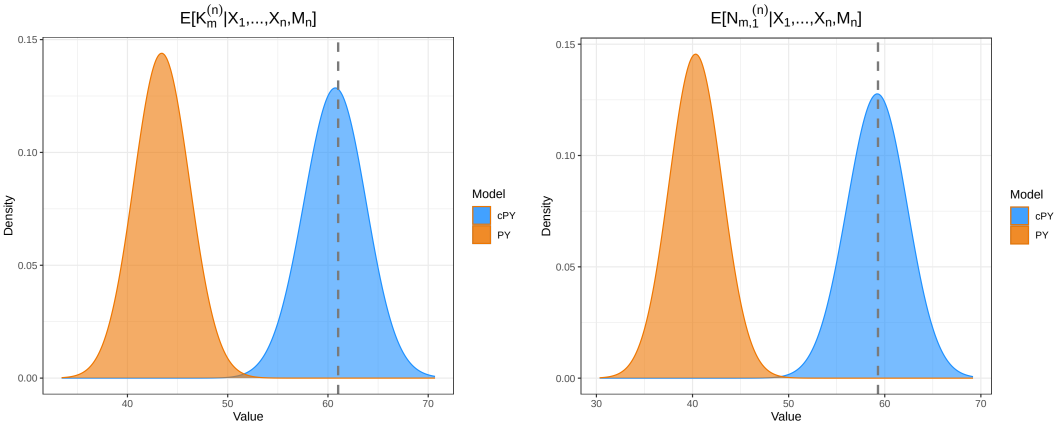

Figure 3: The predictions of $N_{m,1}^{(n-m)}$, $K_m^{(n-m)}$. The black dashed lines correspond to the true values.

These results underline how the use of this additional information is crucial to decrease the prediction errors. With the black dashed line we indicate the true values we tried to predict. We can conclude that taking into account the possible contamination of the dataset is crucial to decrease the prediction errors.

5 Conclusion

In this paper, we aimed to create a model able to efficiently take into account the contamination of a dataset. Our solution consisted in introducing a new sub-family of the most widely used priors in species sampling: the contaminated Gibbs-type prior. Our ultimate goal was to figure out a solution able to make more accurate predictions on a contaminated dataset. We made the fundamental assumption that a contamination error in a dataset is difficult to be repeated. Starting from this we focused on finding a way to address the problem of recognizing the difference between a contaminated datum and a datum representing a (rare) new occurrence.

The first thing we did was implementing the two models: the contaminated Pitman-Yor case and the non-contaminated, ”classic”, Pitman-Yor one. We decided to validate these models by testing them on synthetic data. We immediately noticed that the parameters of the contaminated model were less effected by the leverage produced by the singletons. The non-contaminated model, instead, suffered this leverage and was therefore unable to produce good estimates.

After having estimated the parameters via Metropolis-Hastings, we noticed that our model worked correctly. By taking into account the contamination of the dataset, our model returned good estimates of the parameters. We therefore showed the benefits, in terms of a better estimation of the parameters of the EPPF, of including the contamination in the model. We exploited the fact that our contaminated model performed better than the classic ”non contaminated” one.

Our model proved to be more efficient also in terms of predictions. We verified this by making predictions on an unobserved sample of a given dataset. Our goal was to verify that taking into account contamination would enrich the predictive structure of the model. We also wanted to show that a correct modeling of singletons is of paramount importance to prevent undesirable inferential conclusions. Given the results of the predictions, we can conclude that, in the scenario of possibly contaminated data, the "contaminated" model should be preferred over the "non contaminated" model.

The validation of our work led us to think of several applications in which the contaminated model can be used. In addition to cases of human observations of species in a given area, we can think of other interesting applications such as the field of microdata and the number of "hapax legomena" in a book. In this latter case, the idea is to examine the entire production of an author and discover words used only once. These words are considered peculiar to the author and thus can be treated as contaminants of the text. More generally, any application case that involves taking a dataset based on human observations of subjects that can be repeated could be tested through our model, leading to a better understanding of the dataset one is working on.

A Updating Functions

A.1 Contaminated model

A.1.1 Update sigma

Function from which we update σ

$$\mathcal{L}(\sigma \mid X_{1:n}, \vartheta, \bar{m}_1) \propto \mathcal{L}(\sigma)\sigma^{k-\bar{m}_1} \prod_{i=m_1+1}^{k} (1-\sigma)_{n_i-1} \frac{\Gamma(\vartheta/\sigma + k - \bar{m}_1)}{\Gamma(\vartheta/\sigma)}$$

We know that $\mathcal{L}(\sigma) \sim Beta(\alpha, \beta)$, so we explicit $\mathcal{L}(\sigma)$:

$$\begin{aligned}
&\mathcal{L}(\sigma)\sigma^{k-\bar{m}_1} \prod_{i\ =m_1+1}^{k} (1-\sigma)_{n_i-1} \frac{\Gamma(\vartheta/\sigma + k - \bar{m}_1)}{\Gamma(\vartheta/\sigma)} = \\
&= \frac{\sigma^{\alpha-1}(1-\sigma)^{\beta-1}}{B(\alpha,\beta)}\sigma^{k-\bar{m}_1} \prod_{i=m_1+1}^{k} (1-\sigma)_{n_i-1} \frac{\Gamma(\vartheta/\sigma + k - \bar{m}_1)}{\Gamma(\vartheta/\sigma)} = \\
&= \frac{\sigma^{\alpha-1}(1-\sigma)^{\beta-1}\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\sigma^{k-\bar{m}_1} \prod_{i=m_1+1}^{k} (1-\sigma)_{n_i-1} \frac{\Gamma(\vartheta/\sigma + k - \bar{m}_1)}{\Gamma(\vartheta/\sigma)} = \\
&= \frac{\sigma^{\alpha-1}(1-\sigma)^{\beta-1}\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\sigma^{k-\bar{m}_1} \prod_{i=m_1+1}^{k} \frac{\Gamma(n_i+\sigma)}{\Gamma(1-\sigma)} \frac{\Gamma(\vartheta/\sigma + k - \bar{m}_1)}{\Gamma(\vartheta/\sigma)} = \\
&= \frac{\sigma^{\alpha-1+k-\bar{m}_1}(1-\sigma)^{\beta-1}\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \prod_{i=m_1+1}^{k} \frac{\Gamma(n_i+\sigma)}{\Gamma(1-\sigma)} \frac{\Gamma(\vartheta/\sigma + k - \bar{m}_1)}{\Gamma(\vartheta/\sigma)} = f(\sigma)
\end{aligned}$$

For Metropolis-Hasting our ratio to evaluate a new value σ^* is:

$$\begin{aligned}
r(\sigma^*, \sigma) &= \frac{f(\sigma^*)}{f(\sigma)} \\
&= \frac{\frac{{\sigma^*}^{\alpha-1+k-\bar{m}_1}(1-\sigma^*)^{\beta-1}\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \prod_{i=m_1+1}^{k} \frac{\Gamma(n_i+\sigma^*)}{\Gamma(1-\sigma^*)} \frac{\Gamma(\vartheta/\sigma^*+k-\bar{m}_1)}{\Gamma(\vartheta/\sigma^*)}}{\frac{\sigma^{\alpha-1+k-\bar{m}_1}(1-\sigma)^{\beta-1}\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \prod_{i=m_1+1}^{k} \frac{\Gamma(n_i+\sigma)}{\Gamma(1-\sigma)} \frac{\Gamma(\vartheta/\sigma+k-\bar{m}_1)}{\Gamma(\vartheta/\sigma)}}
\end{aligned}$$

We can simplify some terms and we obtain:

$$\begin{aligned}
r(\sigma^*, \sigma) &= \frac{f(\sigma^*)}{f(\sigma)} \\
&= \\
&= \frac{{\sigma^*}^{\alpha-1+k-\bar{m}_1}(1-\sigma^*)^{\beta-1} \prod_{i=m_1+1}^{k} \frac{\Gamma(n_i+\sigma^*)}{\Gamma(1-\sigma^*)} \frac{\Gamma(\vartheta/\sigma^*+k-\bar{m}_1)}{\Gamma(\vartheta/\sigma^*)}}{\sigma^{\alpha-1+k-\bar{m}_1}(1-\sigma)^{\beta-1} \prod_{i=m_1+1}^{k} \frac{\Gamma(n_i+\sigma)}{\Gamma(1-\sigma)} \frac{\Gamma(\vartheta/\sigma+k-\bar{m}_1)}{\Gamma(\vartheta/\sigma)}}
\end{aligned}$$

If we consider the logarithm of this ratio and a new function h:

$$\begin{aligned}
h(\sigma) = {} & (\alpha - 1 + k - \bar{m}_1)log(\sigma) + (\beta-1)log(1-\sigma) + \\
& \sum_{i=m_1+1}^{k} (log\Gamma(n_i+\sigma) - log\Gamma(1-\sigma)) + log\Gamma(\vartheta/\sigma + k - \bar{m}_1) - log\Gamma(\vartheta/\sigma)
\end{aligned}$$

we find out that our ratio is: $log(r(\sigma^*, \sigma)) = h(\sigma^*) - h(\sigma)$
In Metropolis-Hasting to decide if add σ^* to our vector we compare our ratio with u, where u is

a random value generated from a uniform distribution from 0 to 1. If $u < r(\sigma^*, \sigma)$ then we add σ^* in our vector otherwise we add σ again. In this case our ratio will be $exp(h(\sigma^*) - h(\sigma))$. Before using this function instead of σ we will use $\psi = log(\sigma) - log(1 - \sigma)$, in this way we consider a variable with domain in $\mathbb{R}$ instead of $[0, 1)$. For the same reason we transform also ϑ with $\lambda = log(\vartheta)$. So our function $h(\sigma)$ became:

$$h(\psi) = (\alpha - 1 + k - \bar{m}_1)(\psi - log(1 + e^{\psi})) - (\beta - 1)log(1 + e^{\psi}) + \\ \sum_{i=m_1+1}^{k} (log\Gamma(n_i - \frac{e^{\psi}}{1 + e^{\psi}}) - log\Gamma(\frac{1}{1 + e^{\psi}})) + log\Gamma(\frac{e^{\lambda}}{e^{\psi}} + e^{\lambda} + k - \bar{m}_1) - log\Gamma(\frac{e^{\lambda}}{e^{\psi}} + e^{\lambda})$$

A.1.2 Update theta

Function from which we update ϑ

$$\mathcal{L}(\vartheta \mid X_{1:n}, \sigma, \bar{m}_1) \propto \mathcal{L}(\vartheta) \frac{\Gamma(\vartheta)\Gamma(\vartheta/\sigma + k - \bar{m}_1)}{\Gamma(\vartheta/\sigma)\Gamma(\vartheta + n - \bar{m}_1)}$$

We know that $\mathcal{L}(\vartheta) \sim Gamma(\alpha, \beta)$, so we explicit $\mathcal{L}(\vartheta)$:

$$\mathcal{L}(\vartheta) \frac{\Gamma(\vartheta)\Gamma(\vartheta/\sigma + k - \bar{m}_1)}{\Gamma(\vartheta/\sigma)\Gamma(\vartheta + n - \bar{m}_1)} = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \vartheta^{\alpha - 1} e^{-\beta\vartheta} \frac{\Gamma(\vartheta)\Gamma(\vartheta/\sigma + k - \bar{m}_1)}{\Gamma(\vartheta/\sigma)\Gamma(\vartheta + n - \bar{m}_1)} = f(\vartheta)$$

For Metropolis-Hastings our ratio to evaluate a new value ϑ^* is:

$$r(\vartheta^*, \vartheta) = \frac{f(\vartheta^*)}{f(\vartheta)} = \frac{\frac{\beta^{\alpha}}{\Gamma(\alpha)} \vartheta^{*\alpha - 1} e^{-\beta\vartheta^*} \frac{\Gamma\left(\vartheta^*\right)\Gamma\left(\vartheta^*/\sigma + k - \bar{m}_1\right)}{\Gamma\left(\vartheta^*/\sigma\right)\Gamma\left(\vartheta^* + n - \bar{m}_1\right)}}{\frac{\beta^{\alpha}}{\Gamma(\alpha)} \vartheta^{\alpha - 1} e^{-\beta\vartheta} \frac{\Gamma\left(\vartheta\right)\Gamma\left(\vartheta/\sigma + k - \bar{m}_1\right)}{\Gamma\left(\vartheta/\sigma\right)\Gamma\left(\vartheta + n - \bar{m}_1\right)}}$$

We can simplify some terms and we obtain:

$$r(\vartheta^*, \vartheta) = \frac{f(\vartheta^*)}{f(\vartheta)} = \frac{\vartheta^{*\alpha - 1} e^{-\beta\vartheta^*} \frac{\Gamma\left(\vartheta^*\right)\Gamma\left(\vartheta^*/\sigma + k - \bar{m}_1\right)}{\Gamma\left(\vartheta^*/\sigma\right)\Gamma\left(\vartheta^* + n - \bar{m}_1\right)}}{\vartheta^{\alpha - 1} e^{-\beta\vartheta} \frac{\Gamma\left(\vartheta\right)\Gamma\left(\vartheta/\sigma + k - \bar{m}_1\right)}{\Gamma\left(\vartheta/\sigma\right)\Gamma\left(\vartheta + n - \bar{m}_1\right)}}$$

If we consider the logarithm of this ratio and a new function g:

$$g(\vartheta) = (\alpha - 1)log(\vartheta) - \beta\vartheta + log\Gamma(\vartheta) + log\Gamma(\vartheta/\sigma + k - \bar{m}_1) - log\Gamma(\vartheta/\sigma) - log\Gamma(\vartheta + n - \bar{m}_1)$$

we find out that our ratio is:

$$log(r(\vartheta^*, \vartheta)) = g(\vartheta^*) - g(\vartheta)$$

Now we substitute (σ, ϑ) with (ψ, λ) and we obtain:

$$g(\lambda) = (\alpha - 1)\lambda - \beta e^{\lambda} + log\Gamma(e^{\lambda}) + log\Gamma(\frac{e^{\lambda}}{e^{\psi}} + e^{\lambda} + k - \bar{m}_1) - log\Gamma(\frac{e^{\lambda}}{e^{\psi}} + e^{\lambda}) - log\Gamma(e^{\lambda} + n - \bar{m}_1)$$

A.1.3 Update Beta

Function from which we update β

$$\mathcal{L}(\beta \mid X_{1:n}, \bar{m}_1) \propto \mathcal{L}(\beta)\beta^{n-\bar{m}_1}(1-\beta)^{\bar{m}_1}$$

In this case we have a conjugate prior so we don't need Metropolis-Hastings to find the posterior distribution.

$$\mathcal{L}(\beta) \sim Beta(\alpha, \gamma)$$

We explicit prior distribution and we find the posterior:

$$\begin{aligned}\mathcal{L}(\beta)\beta^{n-\bar{m}_1}(1-\beta)^{\bar{m}_1} &= \frac{\beta^{\alpha-1}(1-\beta)^{\gamma-1}}{B(\alpha,\gamma)}\beta^{n-\bar{m}_1}(1-\beta)^{\bar{m}_1} \\ &= \frac{1}{B(\alpha,\gamma)}\beta^{\alpha+n-\bar{m}_1-1}(1-\beta)^{\gamma+\bar{m}_1-1} \\ &= \frac{B(\alpha+n-\bar{m}_1, \gamma+\bar{m}_1)}{B(\alpha,\gamma)}\mathcal{L}(\beta_{new})\end{aligned}$$

with

$$\mathcal{L}(\beta_{new}) \sim Beta(\alpha+n-\bar{m}_1, \gamma+\bar{m}_1)$$

A.2 Non Contaminated model

A.2.1 Update sigma

Function from which we update σ

$$\mathcal{L}(\sigma \mid X_{1:n}, \vartheta) \propto \mathcal{L}(\sigma)\prod_{i=1}^{k-1}\vartheta + i\sigma \prod_{i=m_1+1}^{k}(1-\sigma)_{n_i-1}$$

We know that $\mathcal{L}(\sigma) \sim Beta(\alpha, \beta)$, so we explicit $\mathcal{L}(\sigma)$:

$$\begin{aligned}\mathcal{L}(\sigma)\prod_{i\ =1}^{k-1}\vartheta + i\sigma \prod_{i\ =m_1+1}^{k}(1-\sigma)_{n_i-1} &= \frac{\sigma^{\alpha-1}(1-\sigma)^{\beta-1}}{B(\alpha,\beta)}\prod_{i=1}^{k-1}\vartheta + i\sigma \prod_{i=m_1+1}^{k}(1-\sigma)_{n_i-1} \\ &= \frac{\sigma^{\alpha-1}(1-\sigma)^{\beta-1}\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\prod_{i=1}^{k-1}\vartheta + i\sigma \prod_{i=m_1+1}^{k}\frac{\Gamma(n_i-\sigma)}{\Gamma(1-\sigma)} \\ &= f(\sigma)\end{aligned}$$

For Metropolis-Hasting our ratio to evaluate a new value σ^* is:

$$\begin{aligned}r(\sigma^*, \sigma) &= \frac{f(\sigma^*)}{f(\sigma)} \\ &= \frac{\frac{\sigma^{*\alpha-1}(1-\sigma^*)^{\beta-1}\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\prod_{i=1}^{k-1}\vartheta + i\sigma^*\prod_{i=m_1+1}^{k}\frac{\Gamma(n_i-\sigma^*)}{\Gamma(1-\sigma^*)}}{\frac{\sigma^{\alpha-1}(1-\sigma)^{\beta-1}\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\prod_{i=1}^{k-1}\vartheta + i\sigma\prod_{i=m_1+1}^{k}\frac{\Gamma(n_i-\sigma)}{1-\sigma}}\end{aligned}$$

We can simplify some terms and we obtain:

$$r(\sigma^*,\sigma)=\frac{f(\sigma^*)}{f(\sigma)}$$
$$=\frac{\sigma^{*\alpha-1}(1-\sigma^*)^{\beta-1}\prod_{i=1}^{k-1}\vartheta+i\sigma^*\prod_{i=m_1+1}^{k}\frac{\Gamma(n_i-\sigma^*)}{\Gamma(1-\sigma^*)}}{\sigma^{\alpha-1}(1-\sigma)^{\beta-1}\prod_{i=1}^{k-1}\vartheta+i\sigma\prod_{i=m_1+1}^{k}\frac{\Gamma(n_i-\sigma)}{1-\sigma}}$$

If we consider the logarithm of this ratio and a new function h:

$$h(\sigma)=(\alpha-1)log(\sigma)+(\beta-1)log(1-\sigma)+$$
$$\sum_{i=1}^{k-1}log(\vartheta+i\sigma)+\sum_{i=m_1+1}^{k}(log\Gamma(n_i+\sigma)-log\Gamma(1-\sigma))$$

we find out that our ratio is: $log(r(\sigma^*,\sigma))=h(\sigma^*)-h(\sigma)$
As in the contaminated model we use $\psi=log(\sigma)-log(1-\sigma)$ ans $\lambda=log(\vartheta)$ so in our Metorpolis-Hastings we use this function:

$$h(\psi)=(\alpha-1)(\psi-log(1+e^{\psi}))-(\beta-1)log(1+e^{\psi})+$$
$$\sum_{i=1}^{k-1}log(e^{\lambda}+\frac{e^{\psi}}{1+e^{\psi}})+\sum_{i=m_1+1}^{k}(log\Gamma(n_i-\frac{e^{\psi}}{1+e^{\psi}}))$$

A.2.2 Update theta

Function from which we update ϑ

$$\mathcal{L}(\vartheta\mid X_{1:n},\sigma)\propto\mathcal{L}(\vartheta)\frac{\Gamma(\vartheta+1)}{\Gamma(\vartheta+n)}\prod_{i=1}^{k-1}\vartheta+i\sigma$$

We know that $\mathcal{L}(\vartheta)\sim Gamma(\alpha,\beta)$, so we explicit $\mathcal{L}(\vartheta)$:

$$\mathcal{L}(\vartheta)\frac{\Gamma(\vartheta+1)}{\Gamma(\vartheta+n)}\prod_{i\ =1}^{k-1}\vartheta+i\sigma=\frac{\beta^{\alpha}}{\Gamma(\alpha)}\vartheta^{\alpha-1}e^{-\beta\vartheta}(\vartheta)\frac{\Gamma(\vartheta+1)}{\Gamma(\vartheta+n)}\prod_{i=1}^{k-1}\vartheta+i\sigma$$
$$=f(\vartheta)$$

For Metropolis-Hastings our ratio to evaluate a new value ϑ^* is:

$$r(\vartheta^*,\vartheta)=\frac{f(\vartheta^*)}{f(\vartheta)}$$
$$=\frac{\frac{\beta^{\alpha}}{\Gamma(\alpha)}\vartheta^{*\alpha-1}e^{-\beta\vartheta^*}\frac{\Gamma(\vartheta^*+1)}{\Gamma(\vartheta^*+n)}\prod_{i=1}^{k-1}\vartheta^*+i\sigma}{\frac{\beta^{\alpha}}{\Gamma(\alpha)}\vartheta^{\alpha-1}e^{-\beta\vartheta}\frac{\Gamma(\vartheta+1)}{\Gamma(\vartheta+n)}\prod_{i=1}^{k-1}\vartheta+i\sigma}$$

We can simplify some terms and we obtain:

$$r(\vartheta^*,\vartheta)=\frac{f(\vartheta^*)}{f(\vartheta)}$$
$$=\frac{\vartheta^{*\alpha-1}e^{-\beta\vartheta^*}\frac{\Gamma(\vartheta^*+1)}{\Gamma(\vartheta^*+n)}\prod_{i=1}^{k-1}\vartheta^*+i\sigma}{\vartheta^{\alpha-1}e^{-\beta\vartheta}\frac{\Gamma(\vartheta+1)}{\Gamma(\vartheta+n)}\prod_{i=1}^{k-1}\vartheta+i\sigma}$$

If we consider the logarithm of this ratio and a new function g:

$$g(\vartheta) = (\alpha - 1)log(\vartheta) - \beta\vartheta + log\Gamma(\vartheta + 1) - log\Gamma(\vartheta + n) + \sum_{i=1}^{k-1} log(\vartheta + i\sigma)$$

we find out that our ratio is:

$$log(r(\vartheta^*, \vartheta)) = g(\vartheta^*) - g(\vartheta)$$

Now we substitute (σ, ϑ) with (ψ, λ) and we obtain:

$$g(\lambda) = (\alpha - 1)\lambda - \beta e^{\lambda} + log\Gamma(e^{\lambda} + 1) - log\Gamma(e^{\lambda} + n) + \sum_{i=1}^{k-1} log(e^{\lambda} + i\frac{e^{\psi}}{1 + e^{\psi}})$$

References

[1] Federico Camerlenghi, Riccardo Corradin, Andrea Ongaro (2021). *"Contaminated Gibbs-type priors"*