POLITECNICO DI MILANO

Scuola di Ingegneria Industriale e dell'Informazione
Corso di Laurea in Ingegneria Matematica

Joint Species Distribution Models: Review and Methodological development

Giovanni Poggiato

Advisor:
Prof. Alessandra Guglielmi
Coadvisors:
Julyan Arbel
Wilfried Thuiller

Abstract

The modelling of species distribution plays an important role in both theoretical and applied ecology. Given a set of species occurrences, the aim is to infer its spatial distribution over a given territory. Joint Species Distribution models (JSDM) study joint species occurrences or abundances at different locations. In the Bayesian framework, this is often done by modelling a continuous latent variable in a hierarchical generalised mixed linear model framework. The regression term models the effect of the environmental conditions on the species, to account for their habitat, while the covariance structure of the residuals of the regression models the interactions between species.

We have reviewed and compared different JSDMs on a simulated dataset, in order to understand their prediction ability as well as the use of these models to infer the interactions between species.

In order to reduce the dimension of the parameters space of these models, as a second goal of this thesis, we have proposed four models that combine latent factors and a Bayesian nonparametric prior to cluster and further reduce the effective number of rows of the matrix representing the random effect induced by the covariance matrix. Such models, that are extensions of a model which has already appeared in the literature , allow the underlying clustering process to be more flexible and to take into account an ecological prior knowledge on the number of clusters. In two of these extensions we used a Pitman–Yor (PY) process prior, that we have approximated using a new truncation method that could satisfy our need for a fast sampling scheme. We have implemented these new models in R and we have tested their predictive performance on both simulated and real datasets.

Estratto

La modellizzazione delle specie gioca un ruolo importante sia nell'ecologia teoretica che in quella sperimentale. Dato un insieme di osservazioni di una specie, l'obiettivo è quello di fare inferenza sulla distribuzione spaziale della specie in un certo territorio. I Joint Species Distribution models (JSDM) studiano la distribuzione di più specie contemporaneamente. In un contesto Bayesiano, tale compito viene svolto modellizando una variabile latente continua all'interno di un modello gerarchico lineare misto generalizzato. Il termine di regressione modellizza l'effetto delle condizioni ambientali delle specie, per tenere conto del loro habitat, uno dei principali fattori per la presenza di una specie. Un altro fattore importante, le interazioni tra le specie, sono modellizzate dalla struttura di covarianza dei residui della regressione.

Abbiamo recensito e paragonato alcuni JSDMs su un dataset simulato, per capire la capacità predittiva di tali modelli e la loro abilità nel ricostruire le interazioni tra le specie.

Per ridurre la dimensione dello spazio dei parametri di tali modelli, come secondo obiettivo della tesi, abbiamo proposto quattro modelli che combinano i modelli a fattori latenti con un prior Bayesiano non parametrico per raggruppare e ridurre il numero di righe della matrice che rappresenta gli effetti aleatori indotti dalla matrice di covarianza. Tali modelli, che sono estensioni di un modello già esistente in letteratura, permettono di incorporare una conoscenza ecologica a priori sul numero di gruppi. In due di queste estensioni abbiamo usato come prior un processo di Pitman–Yor (PY), e abbiamo introdotto un nuovo metodo di troncamento per consentire una simulazione efficace della posterior. Abbiamo implementato questi nuovi modelli in R a ne abbiamo le capacità predittive su dataset sia simulati che reali.

Contents

Introduction

Understanding how species are distributed across space has been one of the main goals of ecology. In particular, investigating which environmental factors drive species distribution within communities, across regions or along environmental gradients can improve our understanding of fundamental ecological processes underlying such patterns, as well as our ability to anticipate future biodiversity changes (Guisan et al., 2017; Thuiller et al., 2013). When building models to explain and predict the distribution of organisms we necessarily need to ask the same questions as the early biogeographers. It is now clear that three main conditions need to be met for a species to occupy a site and maintain populations (see Figure 1.1, Pulliam, 2000; Lortie et al., 2004; Soberón, 2007) :

- the species has to physically reach the site, i.e. to access the region (Barve et al., 2011);
- the abiotic environmental conditions (i.e. temperature, precipitation...) must be physiologically suitable for the species ;
- the biotic environment (interactions with other species) must be suitable for the species.

The first condition is a matter of species **dispersal** capacity from those areas previously occupied by the species. It includes the biogeographic history of the species, and thus all factors limiting its distribution from the place where it first originated, such as barriers to migration, biotic and abiotic dispersal vectors, etc.
The second condition is the matter of abiotic **habitat suitability** for the target species, which means that the combination of abiotic environmental variables at the site – often referred to as environmental suitability - are within the range of environmental conditions that the species requires to grow and maintain viable populations. These suitable environmental conditions are what ecologists call the *environmental niche* (Hutchinson, 1957).

The third condition concerns **biotic interactions**, i.e. interactions with other organisms (see Figure 1.2), either positive (commensalism, mutualism) or negative (competition, predation), which themselves are dictated by the environment through their influence on all organisms in the local community.

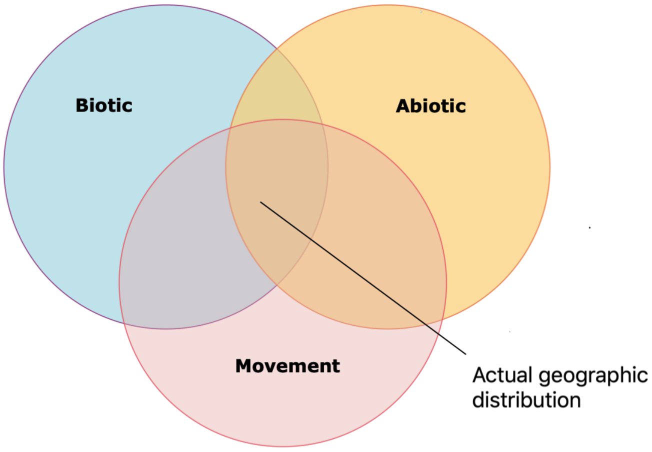

Figure 1.1 – *The three factors that determine the actual distribution of a species (Soberon and Peterson, 2005)*

From a statistical point of view, the most common tools to model how species are distributed across space are species distribution models (SDMs). There are a variety of SDMs that differ in statistical methods or flexibility (Guisan and Thuiller, 2005; Merow et al., 2014; Guisan et al., 2017), but they all relate the presence or abundance, and sometimes the absence, of a species to a set of environmental variables and project this relationship in space and/or time. While SDMs have proven to be very useful and reliable in many different areas and fields (see Yates et al., 2018; Guisan et al., 2017, for reviews), they also have well-known limitations and assumptions that run counter to ecological niche theory (Guisan and Zimmermann, 2000) and that may question the robustness of their predictions. A first major criticism of SDMs is that they model species independently, making the assumption that species respond individualistically to the environment. However, species interactions, in the same way as environmental filtering, are known to be a major factor shaping ecological communities and the abundance of species (Alexander et al., 2015). The importance of species interactions at the scale typically used in SDM studies is up for debate (Soberón, 2007; Godsoe et al., 2017), but given it remains unknown for the vast majority of species, it is clear that missing these interactions could decrease the predictive power of SDMs.

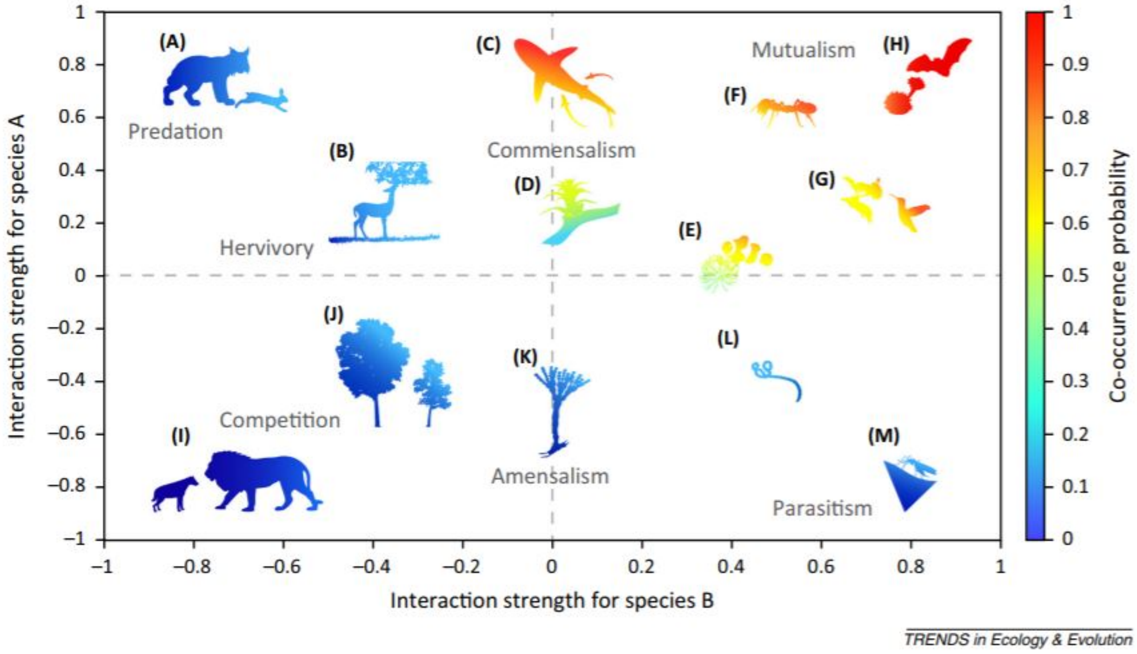

Figure 1.2 – *Interaction strength and co-occurrence probability (Morales-Castilla et al., 2015)*

Recent advances in statistical methodologies and computing power have enabled new models that have begun to address these limitations. Now species can be modelled simultaneously with joint species distribution models (JSDMs) that combine two major recent advances. The first advance is the ability to model species hierarchically, which allows for estimates of species- and group-level responses to predictor variables (Gelfand et al., 2005; Ovaskainen and Soininen, 2011; Pollock et al., 2012). Second, JSDMs estimate associations between species through their residual correlations (Clark et al., 2017; Ovaskainen et al., 2010; Pollock et al., 2014). In a Bayesian framework, this is done by modelling a multivariate continuous latent variable in a hierarchical generalised mixed linear model framework. Species co-occurrence are modelled as the covariance structure of the residuals of the regression (see Chib and Greenberg (1998) for presence/absence data). Several JSDM implementations have been proposed in the recent literature (Ovaskainen et al., 2010; Kissling et al., 2012; Clark et al., 2014, 2017; Pollock et al., 2014; Warton et al., 2015; Golding et al., 2015; Letten et al., 2015; Harris, 2015; Thorson et al., 2016; Ovaskainen et al., 2016a; Nieto-Lugilde et al., 2018), and recent studies compared these different models (Norberg et al., 2019; Wilkinson et al., 2019).

However, as promising as they are, the large and extensive use of JSDMs is still hampered by a number of limitations. From an ecological point of view, JSDMs provide estimates of correlation between species after accounting for influential environmental effects. But do these correlations indicate true interactions between species? And if so, are some types of biotic interactions better captured than others, e.g. can we identify symmetric vs. asymmetric competition and negative vs. positive interactions? At the Laboratory of Alpine Ecology in Grenoble, Pollock et al. (2019) tested one of these models (Pollock et al., 2014) on a dataset that simulates an ecological process where species distributions are filtered from both biotic and abiotic filtering. On the suggestion of Wilfried Thuiller, we have enlarged this study by comparing the results of two other important models (GJAM, HMSC) in order to have a better understanding of the differences between the models, and to investigate whether species interactions can be captured by JSDM in the residual correlation matrix. Thanks to our contribution to the work, we will be among the authours of Pollock et al. (2019). The ecological interpretation of the results was carried together with Wilfried Thuiller and Laura Pollock.

From a mathematical point of view, JSDMs suffer from the curse of dimensionality, since we deal with the challenge of joint modeling for a large number of species. To appreciate the challenge in the simplest way, with just presence/absence (binary) response and say, S species, we have an S-way contingency table with 2^S cell probabilities. Even if S is as small as 100 this is an enormous table, unfeasible to work with without some structure to reduce dimension.

To address this challenge, Taylor-Rodriguez et al. (2017) proposed to reduce the dimension of the residual variance-covariance matrix using Bayesian non-parametric priors. In particular the authors use a Dirichlet Process (DP), which is parameterized by a so called concentration parameter $\alpha > 0$, in order to cluster the species that share the same residuals correlations with respect to other species. In the authors application the concentration parameter α is fixed equal to the number of modelled species, even if such a parameter has a strong effect on the clustering properties of the model (De Blasi et al., 2015; Murphy et al., 2017). Moreover a useful information that ecologists have and that was not used in the study is the number of groups of species in the network (i.e. the number of cluster of species that share the same interactions with respect to other species). At best, this could be known by applying a stochastic block model (Lee and Wilkinson, 2019) on the interactions graph if it is known. However, we can also retrieve this information from other sources, for example the functional traits. Our idea is to modify the model to take into account this prior knowledge on the number of groups in the interaction networks. Even if the groups in the residual correlation matrix are not necessarily the ones of the interaction networks, we believe that this information could however improve the abilities of the model by giving more flexibility and sparsity.

The contributions of our thesis are the following. First, we have generalized Taylor-Rodriguez et al. (2017) method by adding a hierarchical layer to the the DP, where the prior distribution for the concentration parameter α was chosen in order to match a prior knowledge on the number of groups in species interaction network. Then, we have extended the DP to a Pitman–Yor process (PY), the natural extension of the DP that allows for a more flexible clustering. Due to the big dimension of the problem, the already implemented truncation techniques for PY were not feasible for our problem. As an alternative, we have decided to use another truncation technique than those proposed by Arbel et al. (2019) and the application of this method will be described and justified in this manuscript. As for DP, we added a hierarchical layer for the concentration parameter α and the discount parameter σ of the PY process, by fixing their prior distribution in order to match the ecological prior knowledge of the number of groups in the interaction network. Julyan Arbel suggested to extend the DP to a PY process, while the idea of using an ecological prior knowledge for the residual variance covariance was born during a discussion together with Julyan Arbel and Wilfried Thuiller.

We implemented our new models and tested them on a simulated dataset, to test the abilities of the models in retrieving the true number of clusters. Since these simulations are simple, the original model from Taylor-Rodriguez et al. (2017) behaves well and could retrieve the original number of clusters in most cases. However in the cases where the wrong choice of the authors to fix the concentration parameter led to a prior number of clusters really far from the true one, the original model was not able to retrieve the true number of clusters, while our models did. We also applied our method to a real dataset, that collects presences and absences of plants in the Bauges Natural Regional Park, where we fixed the prior number of clusters to be equal to the number of Plant Functional Groups (PFGs). The choice of the prior on the hyperparameters strongly influenced the posteriors of the number of clusters, even if there was no improvement with respect to the original model in terms of prediction. Since for our models the posterior of the number of clusters did not move from the prior suggests that PFGs are a good proxy for interactions between plants, which is an interesting ecological results to deepen.

The posterior of the number of clusters of the original model moved from its prior distribution towards the number of Plant Functional Groups (PFGs), but could not reach it due to the peakiness of the prior distribution of the number of clusters. Instead, the posterior of the number of clusters of our extensions did not move from the prior, suggesting that PFGs are a good proxy for interactions between plants, which is an interesting ecological results to deepen. This work has been submitted as a poster to the 12th International Conference on Bayesian Nonparametrics in Oxford, 24-29 June 2019.

The setup of the thesis is as follows. Chapter 1 gives a q uick introduction to Bayesian statistics. Chapter 2 reviews the esisting Joint Species Distribution Models and compare them on a simulated dataset. Chapter 3 introduces Bayesian non-parametric priors and the way they are used in JSDM, and describes our extensions to GJAM with the related computations of the full conditionals. Chapter 4 describes the application of the models to a simulated and a real dataset and their comparison. In Chapter 5 we resume our work and discuss further developments and extensions. This manuscript is a thesis (master level) in Mathematical engineering; for this reason it is addressed to an audience of mathematicians than of ecologists.

Chapter 1

An introduction to Bayesian statistics

All the Joint Species Distribution Models (JSDM) that try to solve the ecological problems discussed in the introduction are Bayesian models. In order to allow every reader to understand this manuscript, this chapter gives a quick introduction to Bayesian statistics, without claiming to be exhaustive. To a formal and complete introduction to Bayesian statistics, see Christensen et al. (2011), for instance.

1.1 Bayes' theorem

Classical statistics is based on a framework where observations $Y_1, Y_2 \dots$ are assumed to be independent and identically distributed (i.i.d.) from an unknown probability distribution p_θ that is a member from a family of distribution $\mathscr{P} = \{p_\theta : \theta \in \Theta\}$. A classical example is the case where $\mathscr{P} = N(\mu, \sigma^2)$, and thus $\theta = (\mu, \sigma^2)$. The aim of statistical inference is to provide an estimation on the value of θ. Bayesian statistics gives a prior distribution on the space of parameters $\pi(\theta)$ and uses Bayes' Theorem to compute the posterior distribution of θ given the observed data $\mathbf{y}$. We introduce:

- $\pi(\theta)$ the prior distribution (tipically a density) of the parameters, that reflects our information on the parameters themselves,
- $p(\mathbf{y}|\theta)$ the likelihood of the data, which describes our subjective belief that $\mathbf{y}$ is the outcome when θ is the true parameter value,
- $\pi(\theta|\mathbf{y})$ is the posterior probability of the parameters given the data, that will be used for the inference.

Bayes' Theorem describes how our probability judgment on the parameters is updated in the light of data to get the posterior distribution:

$$\pi(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)\pi(\theta)}{\int_\Theta p(\mathbf{y}|t)\pi(t)dt} \tag{1.1}$$

Since $m(\mathbf{y}) = \int_\Theta p(\mathbf{y}|t)\pi(t)dt$ is a constant w.r.t the parameters, it is common to write:

$$\pi(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)\pi(\theta) \tag{1.2}$$

Once the posterior distribution is determined, it is possible to compute the probability distribution of a new sample y_{n+1} given the previously observed data $\mathbf{y} = y_1, \ldots, y_n$: the posterior predictive distribution. Indeed one has:

$$\pi(y_{n+1}|y_1, \ldots, y_n) = \int_\Theta p(\mathbf{y_{n+1}}|\theta)\pi(\theta|\mathbf{y})d\theta \tag{1.3}$$

1.1.1 Conjugate priors

The posterior distribution $\pi(\theta|\mathbf{y})$ *is* a probability distribution, but it is not always easy to determine it in closed form or even to sample from it.

The easiest case to deal with is when the prior distribution and the posterior distribution are in the same probability distribution family. The prior is then called a conjugate prior for the likelihood function. For example, the Gaussian family is conjugate to itself with respect to a Gaussian likelihood function. Suppose to have observed $y_1, \ldots y_n$ such that:

$$y_i|\mu, \sigma^2 \overset{iid}{\sim} N(\mu, \sigma^2), \tag{1.4}$$

where the variance σ^2 is known. Choosing a Gaussian prior for the mean $\mu \sim N(\mu; \mu_0, \sigma_0^2)$ will ensure that the posterior distribution is also Gaussian. Indeed:

$$\pi(\mu|\mathbf{y}) \propto N(\mathbf{y}; \mu, \sigma^2)N(\mu; \mu_0, \sigma_0^2), \tag{1.5}$$

and thus:

$$\pi(\mu|\mathbf{y}) = N(\mu; \mu_N, \sigma_N^2), \tag{1.6}$$

with $\mu_N = \sigma_N^2 \left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_i y_i}{\sigma_2}\right)$ and $\sigma_N^2 = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma_2}\right)^{-1}$.
A conjugate prior is an algebraic convenience, giving a closed-form expression for the posterior.

1.1.2 Montecarlo integration

Even when the posterior distribution is available in closed form, the calculation of the integrals to compute its moments or the posterior predictive distribution can be hard or even impossible. Historically, the application of Bayesian methods was limited by one's ability to perform these integrations. Modern Bayesian statistics relies on computer simulations to approximate the values of integrals using Monte Carlo approximation. Perhaps the simplest examples of Monte Carlo integration involve computing the mean of a random variable. For instance, if we are interested in computing the mean of the posterior distribution $\pi(\theta|\mathbf{y})$, i.e. $\mathbb{E}_{[}\theta|\mathbf{y}] = \int_{\Theta} \theta\pi(\theta|\mathbf{y})d\theta$, we can use the following Monte Carlo algorithm:

Algorithm 1: Monte Carlo estimation of the mean

```
for $t$ =1 to $T$ do
    sample $\theta^{(t)}$ from $\pi(\theta|\mathbf{y})$
end
Estimate $\mathbb{E}[\pi(\theta|\mathbf{y})]$ with $T^{-1}\sum_{t=1}^{T}\theta^{(t)}$
```

This method is based on the law of large numbers (LLN). Given a set of i.i.d. samples $x_1, x_2, \ldots, x_n$ from a random variable X with expected value μ, the LLN guarantees that the empirical mean $\bar{X} = \frac{1}{N}\sum_{i=1}^{n} x_i$ converges to the true expected value μ:

$$\bar{X} \to \mu \quad \text{for} \quad n \to \infty \tag{1.7}$$

There exists many useful applications and extension of the Monte Carlo principle (e.g. importance sampling, the method of composition...) see Jackman (2009) for a thorough review.

1.1.3 MCMC

Modern Bayesian statistics is based on the use of Monte Carlo integration, that requires to sample from the posterior distribution $\pi(\theta|\mathbf{y})$. However, i.i.d.-sampling from a multivariate distribution is not always straightforward. A key mathematical tool to do so are Markov chain. The combination of Markov chains and Monte Carlo integration gives birth to one of the most famous techniques in modern Bayesian statistics: Markov Chain Monte Carlo (MCMC). MCMC is a very broad topic, but we will try to introduce it quickly in this paragraph. The idea of Markov chain Monte Carlo is to build a Markov chain on the parameter space Θ whose invariant distribution approaches the posterior density $\pi(\theta|y)$. We can store the ‘path’, ‘trajectory’ or ‘iterative history’ of the chain, treating these values as a series of samples from the posterior density of interest. We can then use the Monte Carlo principle considered before to compute any information that we want to obtain from the posterior. The consistency of this method is guaranteed by the Ergodic theorem:

Theorem 1 (Ergodic Theorem)**.** Let $\{\theta^{(t)}\}$ be an ergodic Markov chain see (see Jackman, 2009, for the definition of a ergodicity) on the parameter space Θ with an invariant distribution π. Consider a measurable function $h : \Theta \rightarrow \mathbb{R}$ such that $\int_{\Theta} |h(\theta)|\pi(\theta)d\theta < \infty$. Then

$$\lim_{T \to \infty} T^{-1} \sum_{t=1}^{T} h(\theta^{(t)}) = \int_{\Theta} h(\theta)\pi(\theta)d\theta, \tag{1.8}$$

where in our application $\pi(\theta)$ will always be equal to the posterior density $\pi(\theta|\mathbf{y})$. Such a theorem justify to estimate $h(\theta)$ with the following method. If we can construct a Markov chain the "right way", then:

- the Markov chain will have a unique, limiting distribution, a posterior density that we happen to be interested in, $\pi = \pi(\theta|\mathbf{y})$;
- no matter where we start the Markov chain, if we let it run long enough, it will eventually end up visiting sites in the parameter space $\mathscr{A} \in \Theta$ with relative frequency proportional to $\int_{\mathscr{A}} \pi(\theta|\mathbf{y})d\theta$;
- the ergodic theorem means that averages $\bar{h} = T^{-1}\sum_{t=1}^{T} h(\theta(t))$ taken over the Markov chain output are simulation-consistent estimates of

$$\mathbb{E}[h(\theta)|\mathbf{y}] = \int_{\Theta} h(\theta)\pi(\theta|\mathbf{y})d\theta.$$

The construction of a Markov Chain with a given stationary distribution $\pi(\theta|\mathbf{y})$ was made possible thanks to the Metropolis–Hastings algorithm (1970). The Metropolis-Hastings algorithm defines a set of acceptance/rejection steps that generate a Markov chain on Θ, the support of $\pi(\theta|\mathbf{y})$. At the start of iteration t, we have $\theta^{(t-1)}$ and we make the transition to $\theta^{(t)}$ as follows:

Algorithm 2: Metropolis–Hastings

```
sample the candidate $\theta^{\star}$ from a proposal distribution $q(\theta^{\star}, \theta^{(t-1)})$ ;
$r \leftarrow \frac{q(\theta^{\star}, \theta^{(t-1)} p(\theta^{\star}|\mathbf{y}))}{q(\theta^{(t-1)}, \theta^{\star}) p(\theta^{(t-1)}|\mathbf{y})}$ ;
$\alpha \leftarrow min\{r, 1\}$ ;
sample $U \sim \text{Unif}(0, 1)$ ;
if $U \leq \alpha$ then
    $\theta^{(t)} \leftarrow \theta^{\star}$
else
    $\theta^{(t)} \leftarrow \theta^{(t-1)}$
end
```

The quantity r is an acceptance ratio, assessing the plausibility of the candidate point $\theta^{\star}$ relative to the current value $\theta^{(t-1)}$. This scheme means that if $r \geq 1$, then the algorithm makes the transition $\theta^{(t)} \leftarrow \theta^{\star}$ with probability 1; otherwise we make that transition with probability r. With probability $1 - r$ the algorithm does not move at iteration t , setting $\theta^{(t)} \leftarrow \theta^{(t-1)}$. The candidate density $q(x, y)$ is the key to the algorithm, and an accurate choice of such a function will be crucial for a good behaviour of the algorithm.

When θ is high dimensional, as if often the case in many statistical models, sampling from the posterior density $\pi(\theta|\mathbf{y})$ is simply too hard for the Metropolis-Hastings algorithm. Rather than sample from the p3ossibly high-dimensional density $\pi(\theta|\mathbf{y})$, we will rely on algorithms that sample from the lower-dimensional *conditional densities* that together characterize the joint density. This idea drives one of the most widely used MCMC algorithms, the Gibbs sampler. Consider partitioning the parameter vector θ into d blocks or sub-vectors (possibly scalars), $\theta = (\theta_1, \theta_2, \ldots, \theta_d)$. Then the Gibbs sampler works as follows, with t indexing iterations:

Algorithm 3: Gibbs sampler

```
for t=1 to T do
    sample $\theta_1^{(t+1)}$ from $g_1(\theta_1|\theta_2^{(t)}, \theta_3^{(t)}, \ldots, \theta_d^{(t)}, \mathbf{y})$ ;
    sample $\theta_2^{(t+1)}$ from $g_2(\theta_2|\theta_1^{(t)}, \theta_3^{(t)}, \ldots, \theta_d^{(t)}, \mathbf{y})$ ;
    ...
    sample $\theta_d^{(t+1)}$ from $g_d(\theta_d|\theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_{d-1}^{(t)}, \mathbf{y})$ ;
    $\theta^{(t+1)} \leftarrow (\theta_1^{(t+1)}, \theta_2^{(t+1)}, \ldots, \theta_d^{(t+1)})$
end
```

These algorithms aims to create a Markov chain that converges to its invariant distribution that is equal to $\pi(\theta|\mathbf{y})$. It is thus very important to check for the convergence of these algorithms, using diagnostics like traceplots, the autocorrelation function and many others that are fully described in Christensen et al. (2011).

In this chapter we have tried to give an introduction to Bayesian statistics and a quick description of the tools that we used in this manuscript. This chapter is of course not an exhaustive description of these topics, but will hopefully allow the readers of these manuscript to have a better comprehension of it.

Chapter 2

Review and comparison of Joint Species Distribution Models

The chapter is focused on the description and comparison of three Joint Species Distribution Models (JSDM) to predict species distribution, their correct response to environmental gradients and their inferred associations between them. We discussed how we tested the different models on a simulated dataset to compare the results. Thanks to the simple properties of this dataset, we were able to fully understand the real properties of JSDM, in particular concerning their abilities in retrieving the environmental niche and the interactions with other species.

Very few papers have so far compared JSDMs on simulated and real dataset, and none of them really focused on understanding the reasons of their results. Here, we tested the models on one simple dataset, where we analyzed the results deeply in order to understand the limits and the potential of JSDM.

2.1 Models description

The three JSDMs that we will study are extensions of the generalised linear modelling (GLM) framework, which is widely used for modelling species distribution data (Gelfand et al., 2006). The statistical models are defined below using a common notation.

The following terms are consistent across all models:

- Subscript notation for sites is $i = 1, \dots, n$, for species $j = 1, \dots, J$, and for covariates $k = 1, \dots, K$;
- $\mathbf{y}$ is the response variable, a $n \times J$ matrix where $y_{i,j}$ is 1 if species j is present at site i, and 0 viceversa;

- $\mathbf{z}$, a normally distributed latent variable, which is also a $n \times J$ matrix;
- μ, the linear predictor for the measured covariates;
- $\mathbf{X}$, the $n \times K$ matrix of measured covariates;
- $\mathbf{I}$, the identity matrix.

2.1.1 The core model

One of the first JSDMs was proposed by Pollock et al. (2014), and we will refer to it as the core model (**CM**). This model is built on the multivariate probit regression model (Chib and Greenberg, 1998), by using a latent variable parametrisation of a probit model rather than the probit link directly.

The probability of species presence is modelled as the probability of a latent multivariate normally distributed variable exceeding a threshold, such that $y_{ij} = 1$ if $z_{ij} > 0$, and $y_{ij} = 0$ otherwise. The community (i.e. the set of species) present at site i is thus characterized by the multidimensional latent variable $z_{i,.}$.

$$\begin{aligned} y_{ij} &= 1(z_{ij} > 0) \\ z_{ij} &= B_{.j} X_{i.} + e_{ij} \\ B_{jk} &\overset{ind}{\sim} N(\omega_k, \sigma_k) \\ e_{i.} &\overset{iid}{\sim} MVN(0, R), \end{aligned} \tag{2.1}$$

where the dot notation $B_{.j}$ represents the j-th column of B, and $X_{i.}$ is the i-th row of X.

The linear predictor $\mu_{.j} = B_{.j} X$ represents the environmental filtering over the distribution of species j. The suitable environmental conditions for each species is what is commonly called *environmental niche* in ecology. Correlations in the residual error at each site e_i are captured by R, a symmetric and positive-definite matrix. Its diagonal elements are 1 and its off-diagonal elements are restricted between -1 and 1, as imposed by the multivariate probit regression. The elements of R reflect species co-occurrence patterns not described by the environmental predictor: species interactions, or missing predictors. By considering the interactions between species and the environmental niche, JSDMs take into account two of the main drivers of species distributions.

We complete the Bayesian model above assuming

$$\omega_k \overset{iid}{\sim} N(0, 100),$$

while the standard deviations have uniform prior

$$\sigma_k \overset{iid}{\sim} U(0,100),$$

and correlation coefficients have an inverse Wishart prior

$$R \sim IW(J+1,\mathbf{I}),$$

where $J+1$ is the degree of freedom and I the scale matrix (this notation is consistent throughout the manuscript). All hyperpriors are chosen to be vague since no prior information of these parameters is available.
The posterior distribution is sampled via MCMC using Gibbs sampling in JAGS (Plummer, 2014).

2.1.2 Hierarchical Model of Species Communities (HMSC)

Hierarchical Model of Species Communities (HMSC) is a model appeared in a sequence of papers (Ovaskainen et al., 2017a,b; Tikhonov et al., 2017) that aim to give a very complete framework that takes into account all possible information about species in one single hierarchical model. Since the data we used for the comparison of the models do not include informations on functional traits and phylogeny, we are not going to present these features in details (see the above cited papers for a complete information on them).

HMSC is a very similar to (2.1), but allows the regression coefficients to be correlated:

$$\begin{aligned}
y_{ij} &= 1(z_{ij} > 0) \\
z_{ij} &= B_{.j} X_{i.} + e_{ij} \\
B_{.j} &\sim MVN(\omega, V) \\
e_{ij} &= \nu_{ij} + \varepsilon_{ij} \\
\nu_{ij} &= \eta_{i,.} \lambda_{j,.} \\
\varepsilon_{ij} &\overset{iid}{\sim} MVN(0,1) \\
\eta_{,i} &\overset{iid}{\sim} MVN(0, I_{n_f})
\end{aligned} \tag{2.2}$$

where $\omega = (\omega_k)_{k=1,\dots K}$ and each element of the vector has a vague prior

$$\omega_k \overset{iid}{\sim} N(0,100).$$

The variance-covariance matrix of the regression coefficients of each species is V, with

$$V \sim IW(5, I).$$

As above, the correlation coefficients have an inverse Wishart prior

$$R \sim IW(J+1, \mathbf{I}).$$

HMSC biggest improvement concern the different representation of the error e_{ij}. In the CM, the full rank matrix R represents the variation in species occurrences and co-occurrences that cannot be attributed to the responses of the species to the measured covariates. With J species, each covariance matrix R as bijective mapping to a space of $\frac{J(J+1)}{2}$ unrestricted parameters (Lewandowski et al., 2009), making their estimation numerically challenging. In order to reduce the parameter space, HMSC uses latent factors and latent loadings. Under the classic assumption made in factor models that the latent factors marginally follow multivariate normal distribution $\eta_{,i} \overset{iid}{\sim} MVN(0, I_{n_f})$, the latent loadings provide then a parametrisation of R as $R = \Lambda^T \Lambda + I_J$, where $\Lambda = \{\lambda_{ij}\}$ is the $J \times n_f$ matrix containing all latent loadings. The utility of the latent factor approach comes from the dimension-reduced parametrization of R in case where $n_f << J$. Instead of fixing the number of latent factors n_f, HMSC treats n_f as an unknown parameter through the shrinkage approach of Bhattacharya et al. (2013), see this paper for a complete definition of the hierachical model and the hyperpriors definition. This variance decomposition could be considered similar to a linear regression where the latent loadings $\lambda_{j,q}$ are the parameters of the regression, and the latent factors are interpreted to model some missing covariates, which have an impact on the species occurrences and are not represented in the matrix. For more detailed treatment of this interpretation see Warton et al. (2015).

The posterior distribution was sampled via MCMC using a Gibbs sampler, implemented in R (R Core Team, 2013) in the HMSC package (Ovaskainen et al., 2016b).

2.1.3 Generalized Joint Additive Model (GJAM)

GJAM is a Joint Species Distribution Model that aims to fit all type of response data, using a latent variable. This is an important feature: since in ecology the collection of data can be very heterogeneous, it is suitable to have a single model to deal with multifarious data.

For presence-absence data it is a multivariate probit regression model that takes on two different forms depending on J, the number of species to be modeled, for the same reasons we discussed above: when the number of species J is big, the model suffers from the "curse of dimensionality". The small dataset form (i.e. when J is small) is equivalent to the core model (2.1), but the regression coefficients B_{jk} are indipendent and vague:

$$\begin{aligned} y_{ij} &= 1(z_{ij} > 0) \\ z_{ij} &= B_{.j}X_{i.} + e_{ij} \\ B_{j.} &\overset{iid}{\sim} N(0, 100I) \\ e_{i.} &\overset{iid}{\sim} MVN(0, R) \\ R &\sim IW(J+1, \mathbf{I}), \end{aligned} \tag{2.3}$$

The big dataset form (i.e. when J is big and dimension reduction is needed) proposes a latent factor approach similar to HMSC.

$$\begin{aligned} y_{ij} &= 1(z_{ij} > 0) \\ z_{i,.} &= BX_{i,.} + \Lambda\eta_{i,.} + \varepsilon_i \\ \varepsilon_i &\overset{iid}{\sim} N(0, \sigma_\varepsilon^2 I_J) \\ B_{j.} &\overset{iid}{\sim} N(0, 100I), \end{aligned} \tag{2.4}$$

where the random vector η_i are i.i.d. with $\eta_i \overset{iid}{\sim} N(0, I_{n_f})$. The variance of the error ε_i has an inverse gamma prior

$$\sigma_\varepsilon^2 \sim IG(0.01, 0.01).$$

Here the number of factor n_f is fixed, and can be chosen in order to maximize some goodness-of-fit/prediction metrics like DIC, BIC, or LPML (see Gelman et al., 2004, for a complete description of these metrics).

A further reduction in the number of parameters can be attained by finding common rows in Λ, using a Dirichlet process (DP). We are going to give only a brief description of the Bayesian non parametric dimension reduction here, because it will be fully analyzed in the following chapter. GJAM exploits the clustering properties of the Dirichlet Process to find groups of species in the rows of Λ. Species in the same group will have the share response to the unmeasured variables, meaning that in the variance covariance matrix $R = \Lambda^T\Lambda + \sigma_\varepsilon^2 \times I_J$ these species will share the same behaviour with respect to other species.

GJAM uses the finite approximation of the DP proposed by Ishwaran and Zarepour

(2000), which facilitates sampling with the blocked Gibbs sampler of Ishwaran and James (2001). Given the truncation level N, the N vectors $Z_j \in \mathbb{R}^{n_f}$ denote the $N \times n_f$ matrix whose rows make up all potential atoms (i.e., vector values that the rows of Λ may take). The important thing to keep in mind, is that we only need to estimate a $N \times n_f$ matrix instead of a $J \times n_f$ one. GJAM take $N = min\{150, J\}$ that gives a good dimension reduction. The full hierarchical model proposed in GJAM will be fully specified in the next chapter. GJAM is implemented in R (R Core Team, 2013) in the package GJAM (Clark, 2017), where the posterior distribution of the parameters is sampled using a Gibbs sampler.

2.2 Models comparison

2.2.1 Introduction

In order to fully understand the models and their ability to fit the data and make prediction, we decided to test them. Until recently, it was still unclear which models perform best for interpolation or extrapolation of existing data sets, particularly when one is concerned with species assemblages. To evaluate these differences between the models, some recent papers have tested many JSDM on different datasets and compared the results with stacked SDM (Norberg et al., 2019; Wilkinson et al., 2019). These works show that there is no over performing model on all kind of datasets, and suggest that researchers should choose the suitable model by evaluating the performances on their own datasets. Moreover, this literature shows that JSDM do not lead to overwhelming better results than training an indipendent SDM for each species.

To complete the existing literature, we decided to test the models on a dataset based on a simulated ecological process, following the work of Pollock et al. (2019) that fitted the core model on these data. The interest of this approach is to understand not only the ability to fit and predict, but also to understand whether the residual correlation matrix is able to retrieve the network of interactions between species (a task that SDM cannot achieve, since each species is supposed to be independent from the others). The inference of the interaction network is a very complex topic in ecology and represents an important feature that JSDM are supposed to achieve (according to the authors of the models), and we think that it is important to test the power of the different models on this task.

While almost every paper concerning JSDM evaluate how species attract/repel by considering the residual covariance matrix, we also took into account partial correlations, that are expected to be more informative about causal links among the species than the raw correlations.

JSDM have been recently shown to assign positive and negative interaction in a homogeneous environment (Zurell et al., 2018), while other studies have shown that species interactions can result in indirect effects that are not easy to measure from co-occurrence data (Cazelles et al., 2016) and that strong species interactions can be obscured by environmental variables in SDMs (and presumably in JSDMs) if the two are correlated (Godsoe et al., 2017). Pollock et al. (2019) extended these tests on the ability of JSDM to capture species interactions to more realistic situations where species interactions need to be separated from the environmental context. In their work, the authors created simulated communities where there is no residual correlation due to unmeasured covariates. Therefore the covariance matrix is only related to interactions, and this allows to really test whether the covariance matrix is a good tool to retrieve the interactions between species. Our work in this section contributed to Pollock's paper and is waiting for submission.

2.2.2 Simulations of the ecological process

Ecological communities were simulated with a process-based, stochastic model that simulates the assemblage of individuals into communities (a community is given by the set of individuals living in one site) from a regional species pool (Münkemüller and Gallien, 2015). These simulations take as input (see Figure 2.1 for illustration):

- n, the number of communities that we want to simulate (one community in each site). We consider $n = 800$;
- One environmental covariate (that we call the environment) at each site (randomly selected between 0 and 100);
- J, the number of species in the species pool. We fix $J = 5, 10, 20$;
- The environmental preferences (i.e. the environmental niche) of each species, which is represented as a Gaussian curve, with a given optima and variance. For simplicity, the optima of the environmental niches for each species was taken on a regular grid between 0 and 100, and the variances were set to 20 for all species;

- A network of interactions between species, in the form of a matrix. The interactions can be all positive (facilitation only), all negative (competition only) or both (facilitation and competition). We also simulated the case with no interactions (i.e. the matrix will be empty). The interactions can be either sparse (one species can interact with at most 1 one species) or dense (one species can interact with multiple species). However, even in the dense scenario, the matrix remained highly sparse (we will never use matrices with more that 50% of non-zero elements). We considered only symmetric matrices as input.

- A carrying capacity C, the maximum number of individuals that can live in one site. We will took C=40.

Therefore, all the possible combinations of simulations parameters are 21 (Figure 2.1 for illustration).

The community assembly process is randomly initialized with a set of individuals that are randomly chosen from the species pool until the carrying capacity, C, is reached. At each time step, the probability of an individual from species j to replace a random individual of the community i is $W_{j,i}$. This probability depends on how far the environmental conditions at site i are suitable for species j (environmental filter), on the number of individuals present in community i that interacts with species j (facilitation and competition filter), and on the number of individuals of species j that are already present in the community (reproduction filter).
More precisely, we have:

$$\begin{aligned} W_{j,i} = exp(&B_{env} \times log(P_{env,j,i}) + \\ &B_{comp} \times log(P_{comp,j,i}) + \\ &B_{fac} \times log(P_{fac,j,i}) + \\ &B_{abund} \times log(P_{abund,j,i})) \end{aligned} \tag{2.5}$$

This equation defines the relative importance of environmental vs. competition vs. facilitation vs. reproduction filters. Probability weights were then normalized to sum to 1 over all species in the species pool to obtain a probability of replacement for each species.
$P_{env,j,i}$ accounts for the environmental filtering and is the normalized Gaussian density with the optimum of species j as mean and variance 20, evaluated at at the environmental value of community i. The more suitable suitable the environmental conditions are in community i, the higher the probability for species j to enter the community.

Concerning the terms related to the interactions we have:

$$\begin{aligned} P_{comp,j,i} &= 1 - \frac{1}{C} \sum_{l\,:\,l \text{ competes with } j} N_{l,i} \\ P_{fac,j,i} &= \frac{1}{C} \sum_{l\,:\,l \text{ facilitates with } \ j} N_{l,i}, \end{aligned} \tag{2.6}$$

where $N_{l,i}$ is the number of individuals of species l in community i. When in community i there are many individuals of species that facilitate species j, then its probability to enter the community is higher. Viceversa for competition.

The reproduction filter $P_{abund,j,i} = N_{j,i}/C$ describes the probability of an individual entering the community through the reproduction of conspecifics already present. The more abundant the species in the community, the higher its probability of entering.

The coefficients $B_{env}, B_{comp}, B_{fac}, B_{abund}$ weight the importance of the different filters. We took all coefficients equal to 1.

At each time step the algorithm updates the communities randomly using the probabilities defined above. The simulations are thus stochastic. The algorithm keeps iterating until an equilibrium is reached (no changes in the communities for a fixed number of time steps). When the algorithm stops, the communities are transformed into an $n \times J$ matrix, whose element is 1 in position (i, j) if species j is present in the $i-$th community. We fitted our models on 500 hundreds sites (our training set), keeping the other 300 sites for validation.

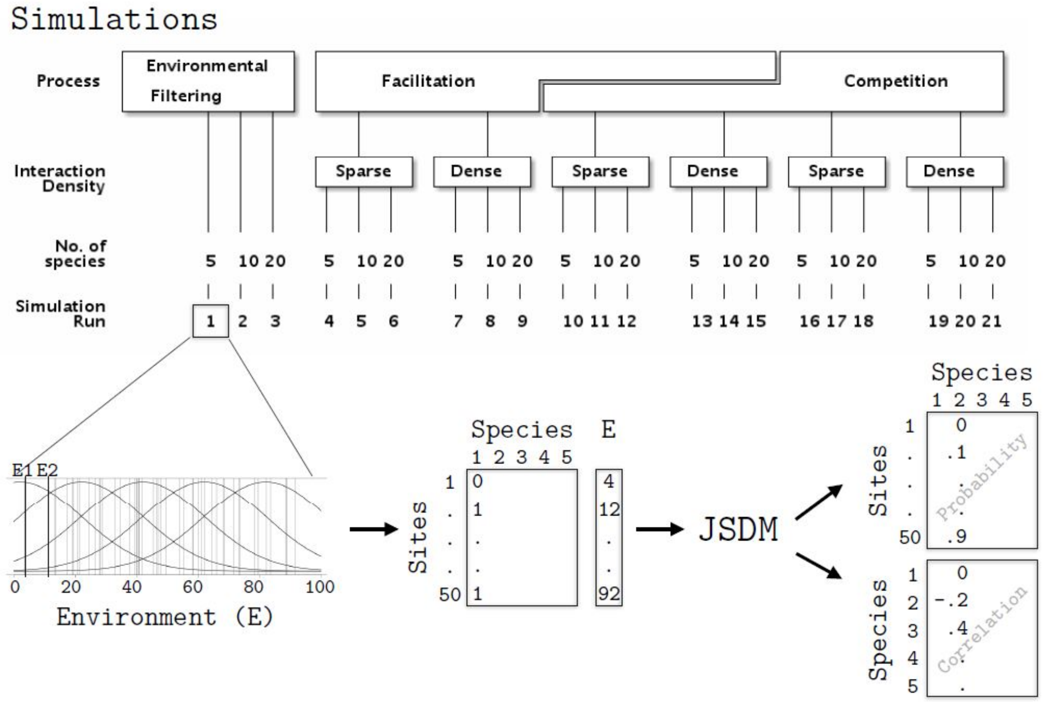

Figure 2.1 – *Scheme of the simulations workflow*

2.2.3 Model fitting

The simulations gave as output a presence-absence dataset of 800 communities, each with an environmental value, that we used to fit the different models described in Section 2.1. On each of the 21 simulated datasets we fitted the CM model, HMSC and GJAM. We fitted GJAM both with and without dimensions reduction (DR-GJAM and full GJAM from now on), leading to a total of 4 models to be fitted for each of the 21 datasets. All models used as covariate the orthogonal polynomial of grade two of the environment.

The CM was fit using the MCMC sampling software JAGS via the R language interface “jagsUI” with 200,000 iterations and 20,000 as burn-in, 5 chains, thinned to keep 1 every 10 samples, for a final sample size of 100,000.

HMSC was trained in R using the HMSC package, using 100000 iterations and 10,000 as burn-in, 2 chains, thinned to keep 1 every 10 samples, for a final sample size of 18,000. Since we had no data concerning traits and philogeny, we did not take them into account.

GJAM was fit in R using the package GJAM. For the full model without dimension reduction we used 60,000 iterations with 10,000 burn-in for two chains, for a final sample size of 100,000. The GJAM package does not allow to thin inside the Gibbs sampler, we then decided not to thin the chains a posteriori, motivated by the recent beliefs that thinning of chains is not usually appropriate when the goal is precision of estimates from an MCMC sample, since the only advantages are memory or time constraints (Link and Eaton, 2012). For GJAM with dimension reduction we fixed the number of latent factors in order to minimize the DIC, to favour a good fit while penalizing the number of parameters. As for the full GJAM, we used 60,000 iterations with 10,000 burn-in for two chains, without thinning, for a final sample size of 100,000.

We fixed the hyperpriors distribution in a vague and consistent way across the different models. Convergence was assessed by visual inspection of trace plots, and by considering for each model the R_{hat} values and the number of effective sample size (nESS).

2.2.4 Methods of comparison

The aim of the study was to quantify how good the different models are in both prediction and inference of the parameters. We thus repeated some of the analyses described in (Pollock et al., 2019), and added a few different checks.

We assessed the predictive abilities of the models both for the in-sample and the out-of-sample prediction. For the in-sample prediction we calculated the area under the receiving operating characteristic curve (AUC) by comparing the predicted expected probability of occurrence from the different models to the observed presence-absences over all sites. Concerning the out-of-sample prediction we used the models fitted on the 500 sites to predict on the test dataset of 300 sites, whose environment conditions are in the same range as in the training dataset. We then calculated the AUC to compare the prediction versus the out-of-sample presence/absences. For each model and for each kind of interaction, we averaged the AUC over all species and simulations.

We also checked for the differences between the inferred regression coefficients by representing the predicted responses, calculated for each species j as $B_{.,j}X$, along the environmental gradient. We then compared this curve with the environmental niche of each species (the Gaussian distribution used as input to the simulation models) and the presence-absences data. We also fitted a univariate GLM where each species was considered indipendently (SDM), to understand whether JSDM improved the detection of the fundamental niche compared to SDM.

In order to evaluate how well the models detect interacting species pairs, we calculated a success rate for competing, facilitating and non interacting species.

Pollock defined the success rate as the number of pairs correctly identified out of the total number of pairs interacting (or not) in the simulations. The pair of competing (respectively facilitating) species i, j was correctly identified if the 95% credibility interval of the corresponding element of the residual correlation matrix $R_{i,j}$ was entirely below zero (respectively above zero). We called this pair of correctly identified species a true positive. The pair of non-interacting species i, j was correctly identified if the 95% credibility interval of the corresponding element of the residual corretion matrix $R_{i,j}$ overlapped zero. We called this pair of correctly identified species a true negative. By considering the ratio of correctly identified interacting species, we detected the sensitivy, while we determined specificity by examining the ratio of correctly identified non-interacting species. There were no case of totally misclassified species (facilitating species assigned to competion and viceversa).

We then focused on the differences between correlation and partial correlation matrix. Starting from the covariance matrix given by the models, one can easily compute the partial correlation matrix as the scaled opposite of the inverse of the covariance matrix. Indeed, for all mentioned models the underlying variable that represent species has a multivariate normal distribution, and for such distribution the partial correlation matrix coincides with conditional correlation, meaning that the 0 value for the partial correlation coefficient implies that the two random variables are conditionally independent.

The partial correlation could be more suitable for inferring the species interactions as partial correlation represents direct associations, while correlation coefficients represent marginal associations, meaning that both direct and indirect association are represented in this matrix. Even if this is well known in ecology, only Ovaskainen et al. (2016a) considered both partial correlation and correlation for studying the interactions in the community. However, in this study the authors showed that there were not a lot of differences between partial correlation and correlation.

We computed the partial correlation matrices and we measured the difference between the correlation and partial correlation matrix as : $K_{diff} = \frac{|R-R_p|}{2}$ where R, R_p were the matrices obtained from the correlation and partial correlation matrices as $R = \text{sign}(R)$, $R_p = \text{sign}(R_p)$, so that we compared the position and the sign of non-zero. We then calculated the mean value of K_{diff} across the different simulations, to see the differences across models. Moreover, we repeated the same analysis for the correlation matrix described above also for the partial correlation matrix.

We repeated Pollock's study in assessing the pattern of species co-occurrences from the simulated communities (the 21 site-by-species matrices) to compare and qualify the JSDM results with a method that does not account for the environmental effect. We quantified co-occurrence between all species pairs using the probabilistic co-occurrence method (Veech, 2013), which has a relatively straightforward interpretation with a simple metric (the number of shared sites between two species), and it corrects type I and type II errors that arise from using randomization procedure to produce a test statistic. We used the standardized effect size, which is different between observed and expected co-occurrence relative to the number of sampling sites. This produces a metric between -1 and 1 that matches the range of the JSDM residual correlation (although with a different specific interpretation). Significance was assessed at alpha=0.05.

2.2.5 Results

Convergence

We first analyzed the convergence of the models. For convergence assessment we used R_{hat} characteristics, effective size and visual assessment of the trace plots, as none of the methods is reliable on its own. For R_{hat} parameter the desired values are below 1.1, for the effective-size, the numbers are supposed to be close to the number of samples from posterior, and the trace-plots should be stationary after some period. Considering this three characteristics, we could see that HMSC and CM converged fairly well, with an acceptable nESS and the R_{hat} below 1.1 for almost every parameter (Figure 2.2). The time for running HMSC was significantly smaller then for CM. It never took more than 1 hour minutes to complete, while the core model was much slower, and needed up to 30 hours to achieve completion. Since the softwares to fit the two models are different, this comparison is not really fair. However, the slowness of the sampling from the CM is a problem that has to be highlighted.

The full GJAM was not able to reach convergence, both for the beta coefficients and the residual correlation matrix. The traceplots and the autocorrelation of the chains were particularly bad for the off-diagonal terms of the covariance matrix. Most of the parameters had a R_{hat} greater than 1.5, and the nESS were extremely small (Figure 2.2), especially for the elements of the residual correlation matrix.
Our opinion is that this is due to the fact that the full GJAM and the core model are really similar, but their implementation is different.
As highlighted above, the two models are the same except for the prior specification of

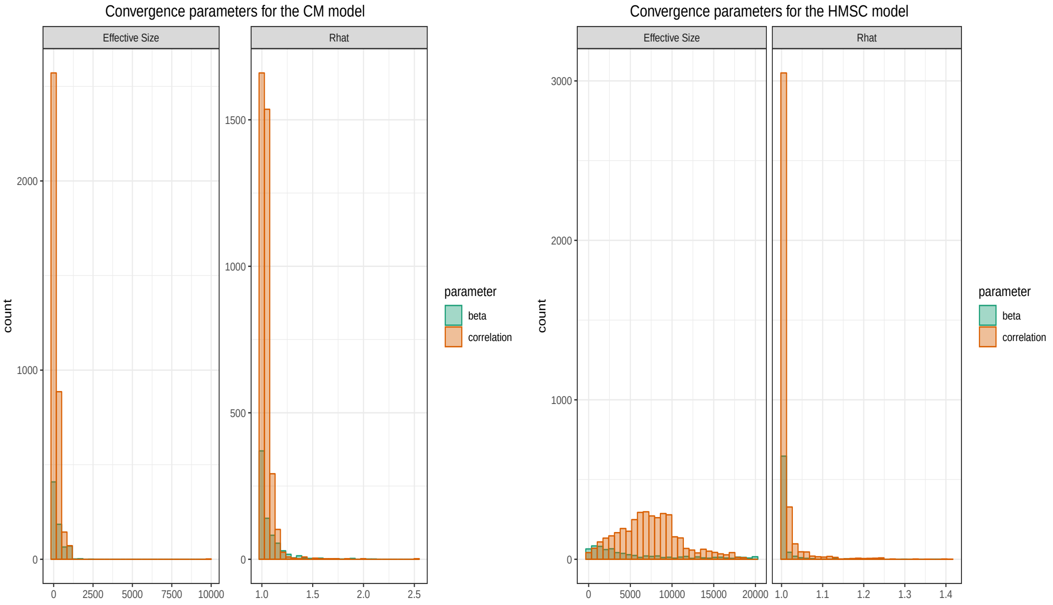

Figure 2.2 – *Convergence assessment for CM model (left) and HMSC model (right). For each model, we represent the histogram of the nESS and the R_{hat} for all the 21 simulations. We represent in green the regression coefficients, and in red the elements of the residual correlation matrix.*

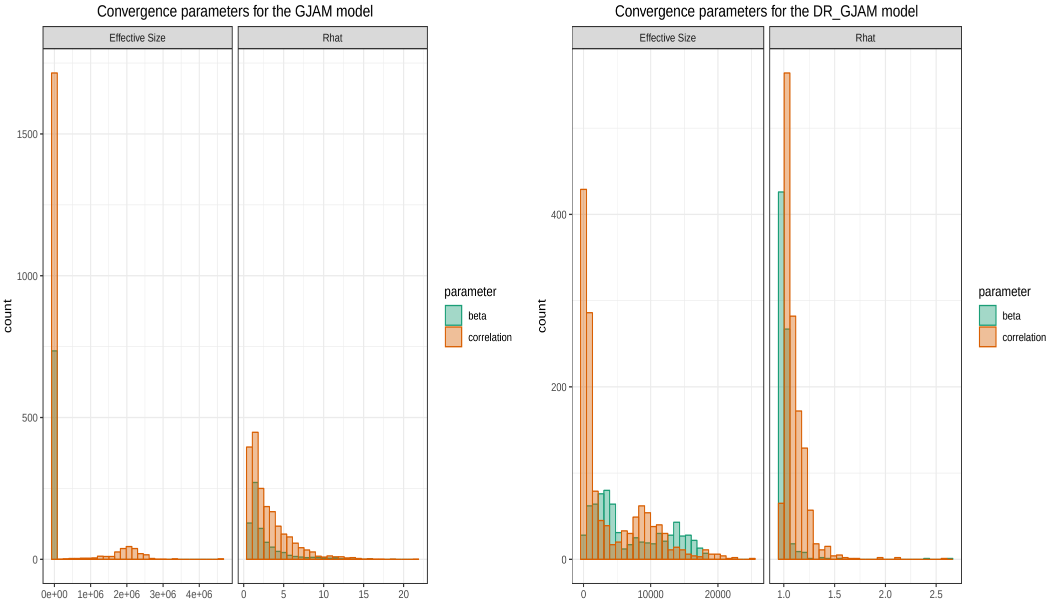

Figure 2.3 – *Convergence assessment for GJAM model (left) and DR-GJAM model (right). For each model, we represent the histogram of the nESS and the R_{hat} for all the 21 simulations. We represent in green the regression coefficients, and in red the elements of the residual correlation matrix.*

the regression coefficients. However, the CM takes advantages in term of optimization of the posterior sampling by its implementation on JAGS, and was able to run for a really high number of iterations.On the other hand, GJAM's Gibbs sampler is directly implemented in a R package, leading to a much less optimized sampling. Moreover, it was not very convenient to run GJAM for the same number of iterations of the core model, since the output object of the functions was to heavy to deal with.
DR-GJAM had a much better convergence then the full mode. The minimization of the DIC generally gave a small number of latent factors n_f, leading to a lighter model, in term of number of parameters. Convergence is now almost reached, especially for the regression coefficients (Figure 2.3).

In-sample prediction

In terms of the in-sample AUC, models' performances were all extremely close to each others (Figure 2.4), and the values where globally high, between 0.8 and 0.95. However, competition led to a decrease in terms of in-sample prediction power, especially when interactions were dense.

Out-of-sample prediction

The AUC concerning the out-of-sample prediction reflected the same results of the in-sample prediction (Figure 2.5). There were no clear differences across the models, and, consistently, the AUC values were slightly smaller then the in-sample AUC, since we were testing the goodness of the models on data that were not used to train the model.
Again, the models had worse AUC values on the simulations with competition between species.

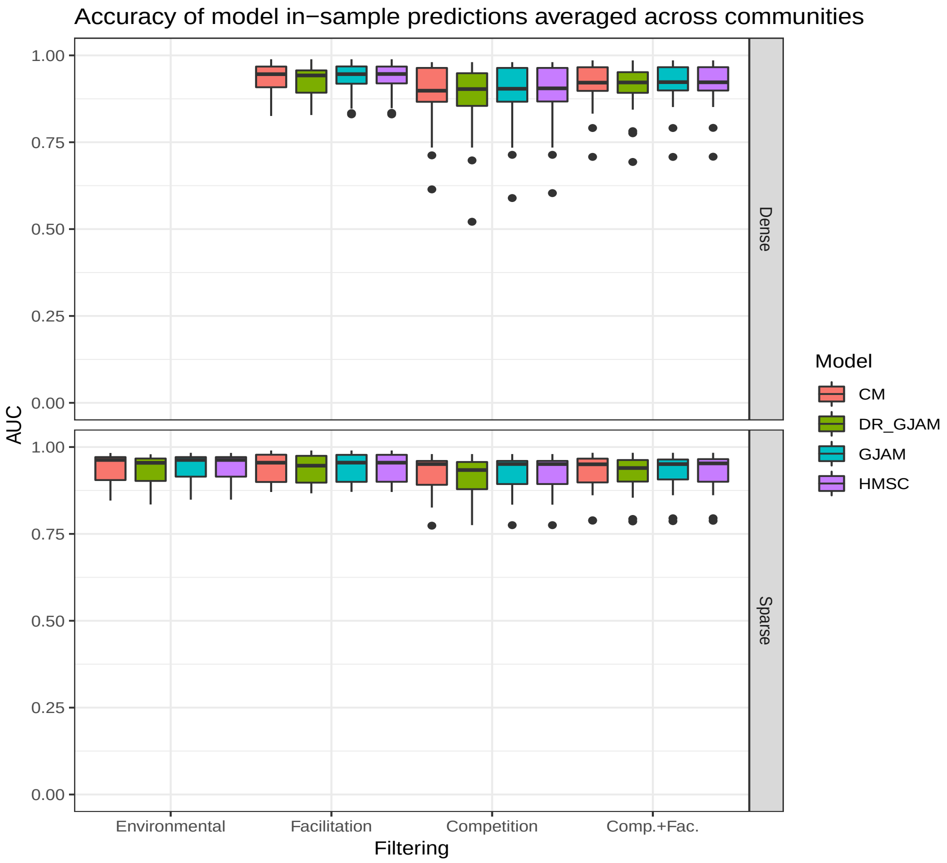

Figure 2.4 – *Accuracy of JSDM predictions on the training dataset, averaged across entire communities, represented as mean AUC and quantiles of order 2.5% and 97.5%. AUC values are averaged across all species within communities simulated with environmental filtering, competition, facilitation or both competition and facilitation.*

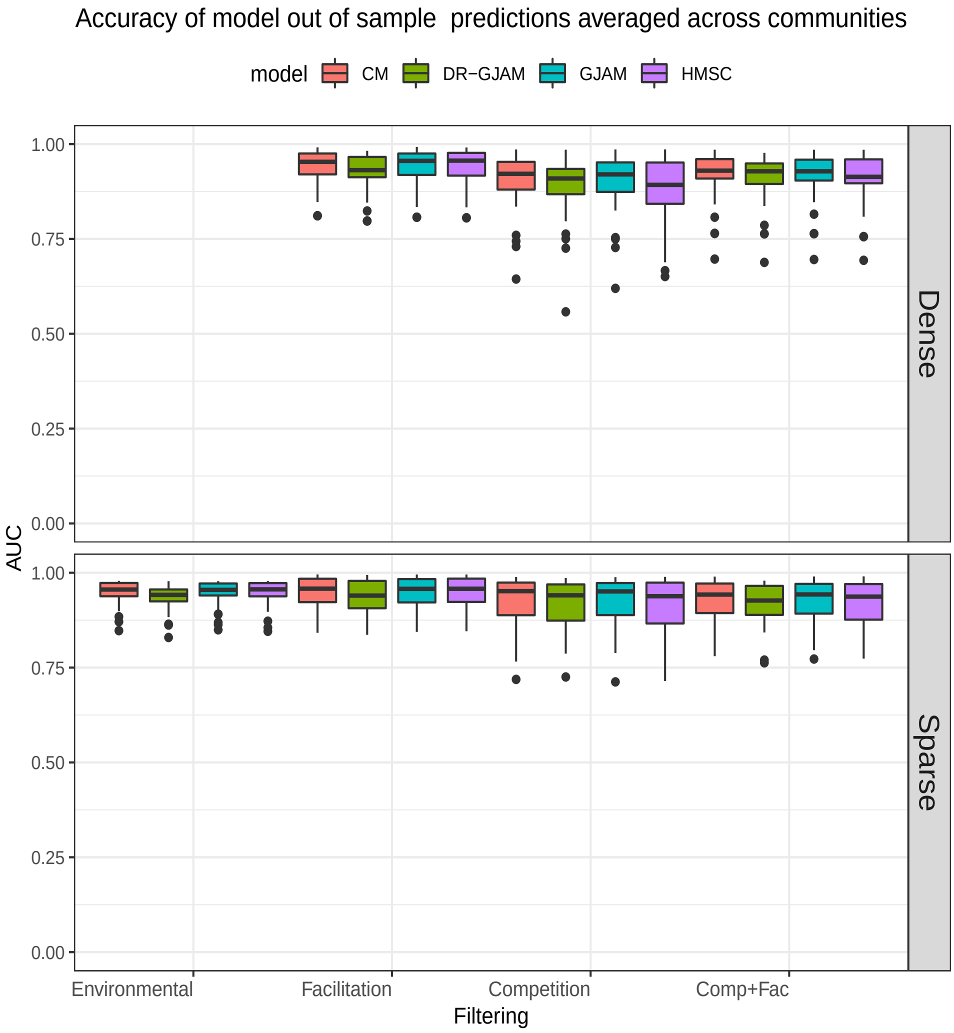

Figure 2.5 – *Accuracy of JSDM predictions on an out-of-sample dataset, averaged across entire communities and represented as mean AUC and 95% confidence intervals. AUC values are averaged across all species within communities simulated with environmental filtering (Enviro.), competition (Comp.), facilitation (Fac.) or both competition and facilitation (C+F).*

Regression coefficients

Concerning the case with environmental filtering only, for the core model, HMSC and full GJAM all the beta coefficients were really close and their credibility intervals overlapped, meaning that there was no evidence to say that the elements of the $\boldsymbol{B}$ matrix are different across models (for example see the coeffiecients and the predicted niche of species 10 in the case of a simulation with 20 species and environmental filter only, Figure 2.6). By consequence, also the predicted environmental niches were similar, and were close to the fundamental niche, confirming the goodness of fit of the models. We did not find any difference in the predicted environmental niche for the independent model (SDM). For DR-GJAM, the results were slightly different for the regression coefficients, but we couldn't assess a real trend in terms of departure from the other models. Moreover, the predicted niche was never far from the ones of the other JSDM. In particular, the intersection with the line of probability 0.5 always happened almost at the same point along the environmental gradient for all models, meaning that the prediction of presences and absences was consistent across all models.

In the facilitation case, the retrieved niche of a species that facilitates was skewed towards the species it facilitated with, leading to an overestimation of the fundamental niche. In Figure 2.7 we showed a comparison of the predicted niche of species 6 in the case of 10 species, with and without facilitation. Species 6 facilitated with 4,8,9 (whose environemental optima were to the left of the one of species 6 for species 4, and to the right for species 8 and 9) and we noticed that the species completely colonized the gradient for environmental values above 25 (dots in the Figure). The predicted environmental niche was consistent across models, and quite far from the fundamental niche, being strongly skewed both to the left and to the right. All sites where the environment value was above 25 were considered suitable.

Instead, in the case where a species competed, the predicted niche was less skewed, but was decreased in magnitude with respect to the fundamental niche. Figure 2.8 shows the effect of competition in the case of a simulation with 10 species. Species 7 competed with species 2,5 and 10, and the results of competition was that we had many absences at sites where the environment was suitable for the species. The predicted environmental niche was slightly skewed but was strongly lower in magnitude than the fundamental niche.

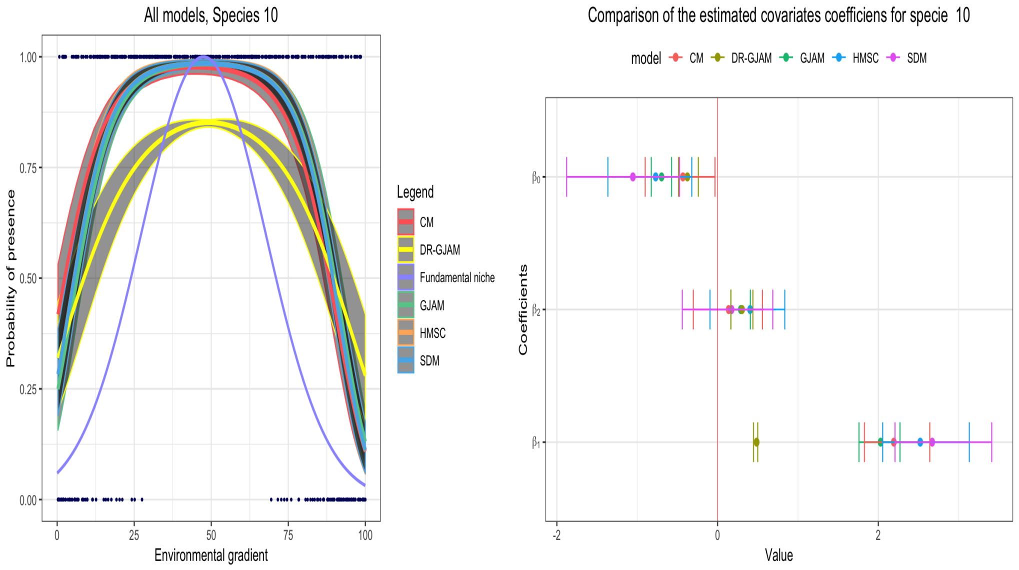

Figure 2.6 – *Left: Predicted environmental niche for the different models for species 10 (in the simulation case of 20 species) and sparse competition and facilitation. The continuous line represents the prediction mean, while the ribbons are the 95% credibility intervals. Right: The mean and the 95% credibility intervals of the three regression coefficients of species 10 for the different models, in the same case study as above.*

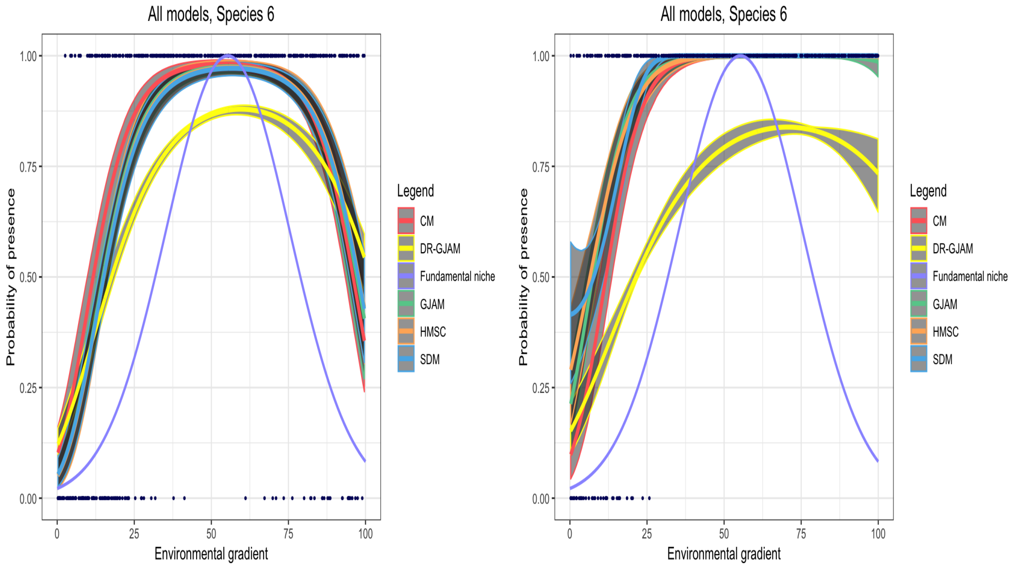

Figure 2.7 – *Predicted environmental niche for the different models for species 6 (in the simulation case of 10 species) with dense facilitation where species 6 facilitates with species 4,8,9 (left) and environment only (right). The continuous line represents the prediction mean, while the ribbons are the 95% credibility intervals. Dots are the presences/absences in the dataset and the violet curve is the fundamental niche (the Gaussian curve given as input to the model)*

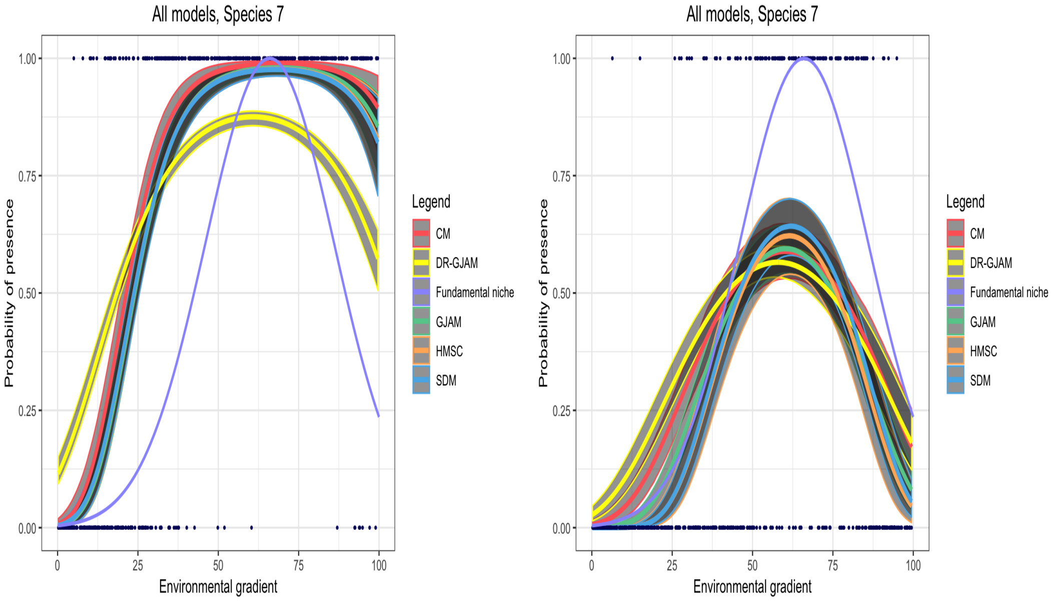

Figure 2.8 – *The predicted environmental niche for the different models for species 7 in the simulation case of 10 species, with dense competition,where species 7 competes with species 2,5,7 (right) and environment only (left). The continuous line represents the prediction mean, while the ribbons are the 95% credibility intervals. Dots are the presences/absences in the dataset and the violet curve is the fundamental niche (the Gaussian curve given as input to the model)*

Residual covariance matrix

HMSC and the CM were really close in terms of ability in retrieving the interactions, while GJAM and DR-GJAM had a different behaviour compared to the previous models, but a similar behaviour between them (Figure 2.9 and 2.10). HMSC and CM could not retrieve facilitation. In particular, HMSC tended to identify less interacting species then the CM, leading to a lower sensitivity but a higher specificity. Instead, both GJAM and DR-GJAM retrieved many more non-zero coefficients in the residual correlation matrix, leading to some well identified facilitating pairs (higher sensitivity), but also to a lot of misclassified non interacting species (lower specificity). Between the two models, DR-GJAM had higher sensitivity but lower specificity, with a lot of misclassified species. All models tended to correctly identify more species when the species pool was smaller. This was not true for DR-GJAM, that globally had unchanged performances when the species pool grows.
However, the number of correctly identified facilitation species was globally low.

Concerning competition, all models had better results than facilitation. In particular both HMSC and JSDM were able to increase a lot in terms of sensitivity, but managed

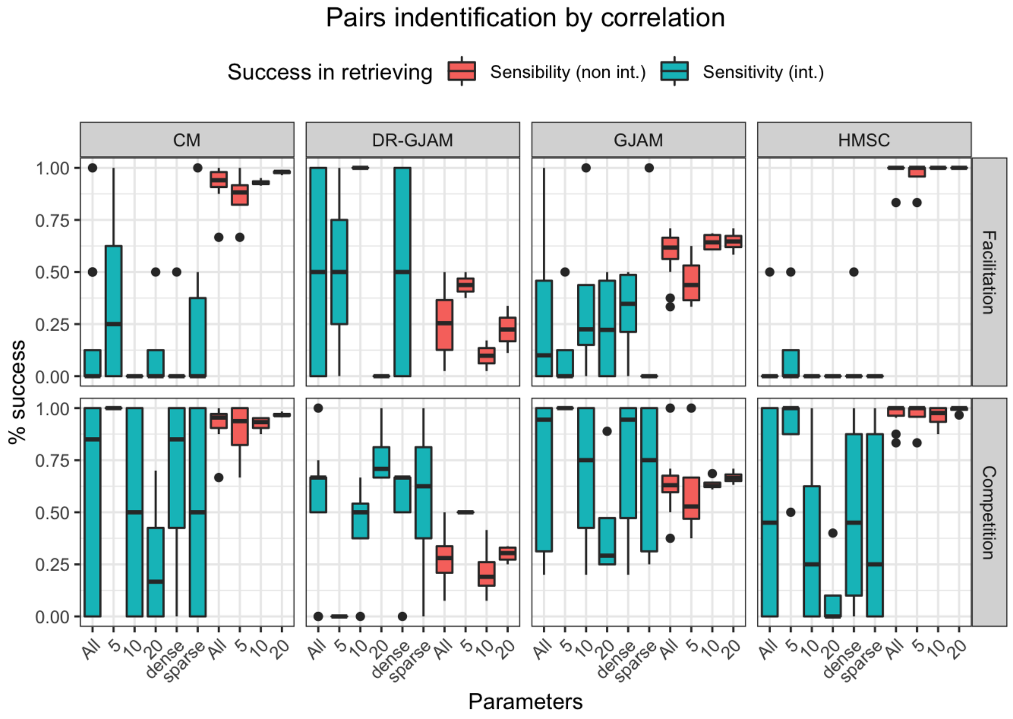

Figure 2.9 – *Success rate calculated on the covariance matrix of JSDMs for interacting (competitors or facilitators, in green) and non-interacting species (in red) in communities simulated with all possible interactions scenarios for the four different models. Bars represent the following groups: all species pairs, species pools (5, 10, or 20 species) and the density of interactions.*

to keep the same very high level of specificity that they had in retrieving facilitation. Again, GJAM and DR-GJAM had slightly different results compared to the two previous models, for the same reason as above. GJAM improved even more its sensitiviy, being able to almost always correctly indentify at least half of the competing species, without loosing power in terms of specificity. DR-GJAM had worse results, with a lot of variations.

Concerning the results on the precision matrix, the differences compared to the correlation matrix were not stunning. All the models showed globally the same behaviour both in competition and facilitation (Figures A.1, A.2), but the element of the precision matrix were more shrunk towards zero, leading to a detection of less non-zero interactions.

For probabilistic co-occurrence analyses that ignored the environmental influence, facilitating species pairs tended to co-occur more than non-interacting species, especially in communities with sparse interactions (Figure 2.10). Unexpectedly, competing species pairs were indistinguishable from non-interacting species in all cases except

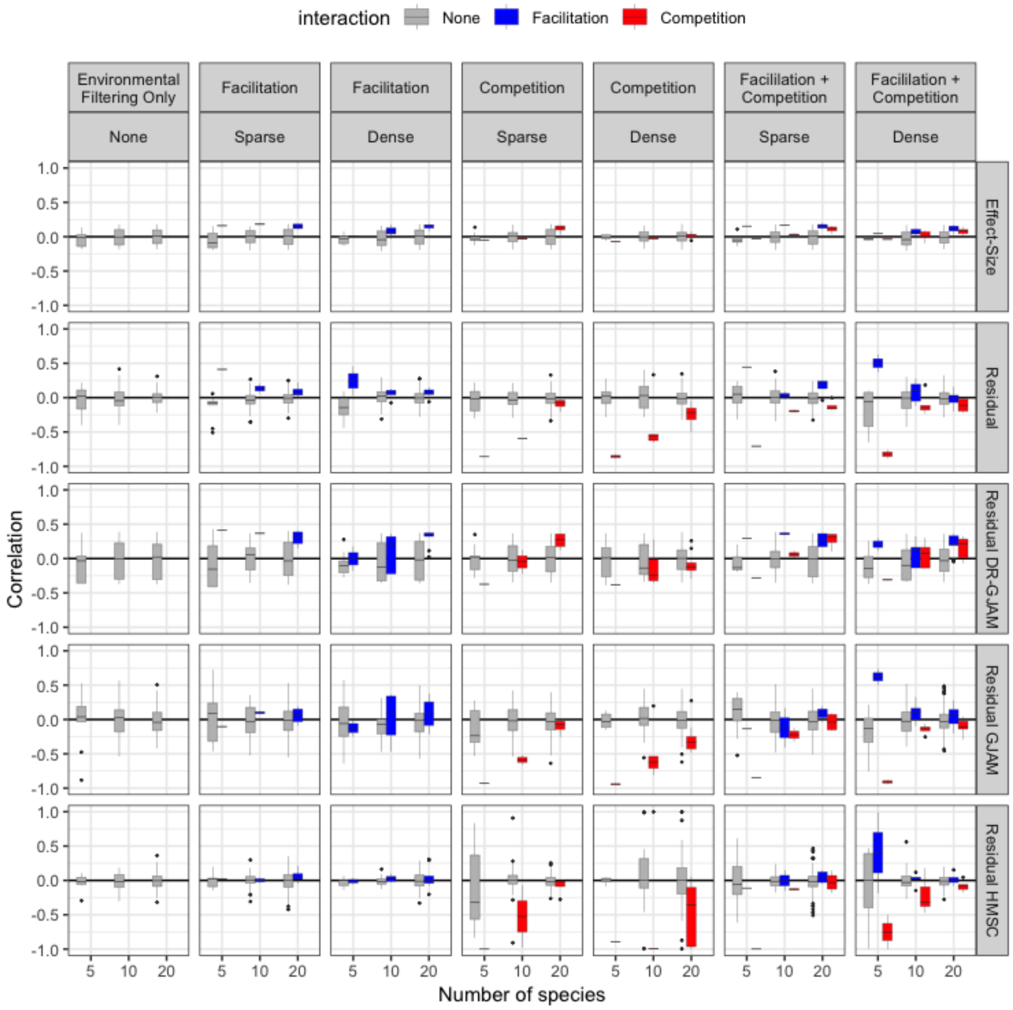

Figure 2.10 – *Barplots representing the values of the probabilistic co-occurrence index (top row) and of the residual correlation matrix for each model (the other rows). For each scenario, species pairs are grouped into bars representing those that do not interact (grey), that compete (red) or that facilitate (blue).*

the 'dense' interaction, 5-species pool (Figure 2.10). Further, nearly all pairs were deemed to be 'significantly' co-occurring more or less than expected, so this $\alpha = 0.05$ level was not useful for assigning interacting species pairs in this study.

Partial correlation

By comparing the Partial correlation matrix, we observed similar result as in Ovaskainen et al. (2016a). From Figure 2.11 we could see that the for CM and HMSC models the partial correlation and the variance-covariance matrix were really close in the norm that we defined. For DR-GJAM both the correlation and partial correlation had a lot

of noise, and therefore there were more differences between the two matrices. The GJAM model didn't fully converge and for this reason the matrices contained a lot of noise, and there was thus a bigger difference between the two matrices.
However,the differences between the two models didn't depend on the type of interactions or the number of species. (Figures A.1 and A.2).

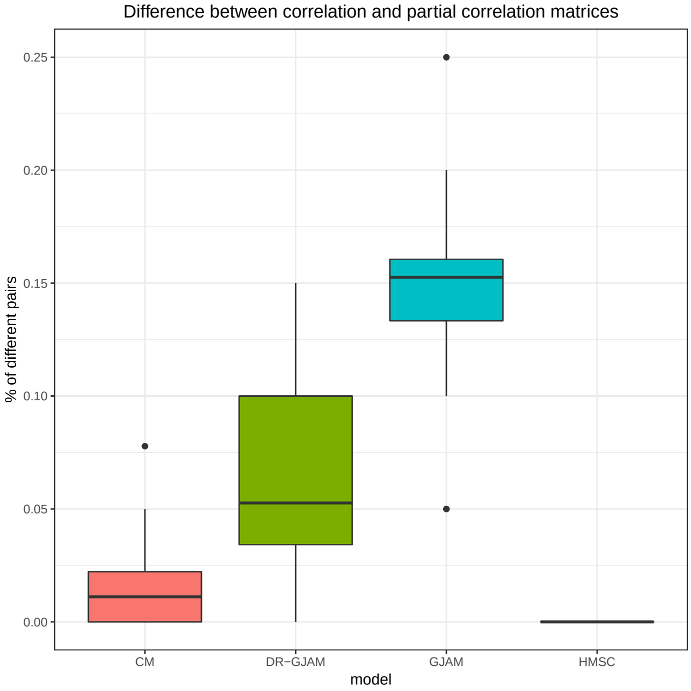

Figure 2.11 – *Percentage of the number of different cells in the correlation and partial correlation matrix for each of the 4 models across the 21 different scenarios*

2.3 Discussion

We have compared four models on simulated communities with different ecological processing, with different types and strength of interactions. For GJAM and CM the likelihood is the same, but the prior is different, as well as the software used to sample from the posterior. While HMSC and DR-GJAM are two types of models that use different approaches to reduce the dimensionality of the problem.

The datasets we used for this simulation had a relatively low/moderate dimension, but even for this data size the CM and GJAM had difficulties with convergence and needed many iterations to obtain reliable estimates, although CM was finally able to converge due to his efficient sampling implemented in JAGS. In higher dimension we would expect even more difficulties for both convergence and parameters estimation, due to the quadratic growth of the number of parameters for these models.

For this reasons we are particularly interested in models that use dimension reduction such as HMSC and DR-GJAM.

We showed that the models were often able to retrieve competition, while they struggled in retrieving facilitation. This was due to the way JSDM fitted the data, and in the difference in the simulated data related to facilitation and competition.

When a species facilitated with other species, the simulations led to a complete colonisation of the gradient towards the species it facilitated with. There is thus a neat division between the suitable and non-suitable environment. Since JSDM fitted the data, the environment was predicted to be suitable at all the sites colonized by the species. The fit was globally very good (high AUC values), and the residuals were thus really low and do not contain any information about facilitation, and thus the variance-covariance matrix Σ cannot retrieve significant interactions.

On the other one hand, competition did not lead to complete extinction, but causes a lot of absences at the sites where the environment should be highly suitable. By consequence, since the models fitted the data, the predicted probability of presence was much lower then 1 even at the optima of the fundamental environmental niche. This led to a worse fit (lower AUC values) and to more significant residuals then in the facilitation case. Since these residuals were only due to competition, their variance-covariance matrix Σ could correctly infer the interactions.

We thus cannot state that JSDM could retrieve competition but not facilitation. Instead, we could affirm that JSDM do not disentangle the effect of the environment from the one of biotic interactions, and thus parameter estimation strongly depends on the way biotic interactions impact species distribution. In particular, when interactions caused complete extinction/colonisation the model related this colonisation to the environment and not to the interactions. However, JSDM could retrieve interactions in less neat cases, where the regression on the environmental covariates could not adequately fit the data. An interesting future work could study the results of these models when considering abundances (counts) instead of presence-absences. Indeed, when we transformed the simulated communities in presence-absences we lost a lot of information, especially in the facilitation case, where presence-absence data show a too neat separation of the environmental gradient.

The models had similar results across all the different tasks, except for GJAM and DR-GJAM that showed a lot of noise in the covariance matrix, due to convergence issues for the former and to the approximation hypotheses. For DR-GJAM we could see that the approximations led to a a change for the predicted environmental niche, and thus this also impacts the variance-covariance matrix σ. However, DR-GJAM is suitable for larger number of species compared to the ones we considered here, and by the way perform globally well, being the only model that did not worsen his retrieving ability when the number of species increased.

We were interested to test if the partial correlation would provide better estimates for interactions in general, and for facilitation in particular, but the results were very similar and didn't provide additional information, since partial correlation and correlation matrices were very close. This is coherent with some observations concerning sparse covariance matrices. For example Sojoudi (2016) stated that when the correlation matrix is sparse, there is almost no different between correlation and partial correlation. In our case we know that the true number of interactions is small, and this sparsity is also found in the inferred covariance matrices. Informally, in the sparse case the number of indirect interactions is limited, and hence correlation and partial correlation matrices are close.

In this chapter we repeated and deepened the work of Pollock et al. (2019), by enlarging the study to multiple models, by analysing the partial correlation matrix. We worked together with Laura Pollock, Wilfried Thuiller and Tamara M ünkem üller to interpret the results of their study in the different and more complete way described above. Our work will thus contribute to Pollock et al. (2019).

Chapter 3

GJAM using Bayesian nonparametric priors

In this chapter we worked on GJAM and its dimension reduction approach that makes use of the Dirichlet Process (DP). We highlighted the limits of the proposed implementation, and extended the model in four different ways that should better behave in terms of clustering properties, and that allow to give to the model a prior knowledge on th expected number of clusters. In two of these extensions we replace the DP with a Pitman–Yor (PY) process, and, based on the results of Arbel et al. (2019) we proposed a novel approximation of the truncation error of stick-breaking representation of the PY process, that allows to a fast approximated sampling from its posterior.

3.1 The Dirichlet process and its application to GJAM

Generalized Joint Additive Model showed its limits in term of applicability due to the high dimension of the full residual correlation matrix when the number of species grows. The dimension reduction proposed by Taylor-Rodriguez et al. (2017) is based on the Dirichlet process, a Bayesian nonparametric prior that is widely used in many different fields, and gave great advantages to GJAM in terms of convergence and computational time. A brief introduction to the Dirichlet process is given in the next section. For a complete introduction we refer to the classical literature in Bayesian nonparametrics such as Ghosal and Van der Vaart (2017); Hjort et al. (2010). From now on, we will focus on GJAM model and its dimension reduction approach, first of all to give it a full description that completes the brief one that we gave in the previous chapter. We also generalize the model to extend it to a more natural framework.

3.1.1 The Dirichlet process: background theory

The Dirichlet process (DP) is a distribution over distributions on some space Θ. The Dirichlet Process was first defined by Ferguson (1973).

Definition 1 (Dirichlet Process)**.** Let H be a distribution over Θ and α be a positive real number. We say that G is a Dirichlet Process, namely $G \sim DP(\alpha H)$ if for any finite measurable partition $\{A_1,\ldots,A_r\}$ of Θ, we have:

$$(G(A_1),\ldots,G(A_r)) \sim Dir(\alpha H(A_1),\ldots,\alpha H(A_r)), \tag{3.1}$$

where $Dir(\alpha H(A_1),\ldots,\alpha H(A_r))$ with positive $A_1,...,A_r$ denotes the Dirichlet distribution. H is called base distribution, and α the concentration parameter.

The formal definition is although not very informative about the properties of the Dirichlet Process.

Another feature is the Pólya Urn representation of the Dirichlet Process (Blackwell and MacQueen, 1973). Let $\theta_1,\ldots,\theta_n$ be a sample from a Dirichlet process G, i.e.:

$$\begin{aligned} \theta_1,\ldots,\theta_n \mid G &\overset{iid}{\sim} G, \\ G &\sim DP(\alpha H). \end{aligned} \tag{3.2}$$

The marginal distribution of the realisations $\theta_1,\ldots,\theta_n$ can be described as follows:

$$\begin{aligned} \theta_1 &\sim H \\ \theta_j|\theta_{j-1},\ldots,\theta_1 &\sim \frac{\alpha H + \sum_{i=1}^{j-1}\delta_{\theta_i}}{\alpha+j-1}, \qquad \text{for} j=2,...,n, \end{aligned} \tag{3.3}$$

so that :

$$(\theta_1,\ldots,\theta_n) \sim H(\theta_1)\prod_{i=2}^{n}\frac{\alpha H(\theta_i)+\sum_{j=1}^{i-1}\delta_{\theta_j}}{\alpha+i-1}. \tag{3.4}$$

We are going to give a quick description of the generalized Pólya urn sampling scheme in order to understand what is a sample from a DP. Suppose each value in Θ is a unique color, and H is a distribution over the colors. The draws $\theta \sim H$ are thus balls with the drawn value being the color of the ball. In addition we have an urn containing previously seen balls. In the beginning there are no balls in the urn, and we pick a color drawn from H, i.e. draw $\theta_1 \sim H$, paint a ball with that color, and drop it into the urn. In subsequent steps, say the $(n+1)$st, we will either, with probability $\frac{\alpha}{\alpha+n}$, pick a new color (draw $\theta_{n+1}H$), paint a ball with that color and drop the ball into the urn, or, with probability $\frac{n}{\alpha+n}$, reach into the urn to pick a random ball out (draw $_{n+1}$ from the empirical distribution), paint a new ball with the same color and drop both balls back into the urn. The predicted law for the $j-$th sample ($j > 1$) from the urn is thus:

$$\theta_j|\theta_{j-1},\ldots,\theta_1 \sim \frac{\alpha H + \sum_{i=1}^{j-1}\delta_{\theta_i}}{\alpha + j - 1}. \tag{3.5}$$

The resulting distribution over the colours with the sampling scheme just derived above is the same as the distribution over values in a Dirichlet process. Since the values of draws (ball colors) θ_k are repeated, let $\theta_1^\star,\ldots,\theta_m^\star$ be the unique values among $\theta_1,\ldots,\theta_n$, and n_k be the number of repeats of $\theta_k^\star$. Then the predictive distribution can be equivalently rewritten as:

$$\theta_n|\theta_{n-1},\ldots,\theta_1 \sim \frac{\alpha H + \sum_{j=1}^{m} n_j \delta_{\theta_j}^\star}{\alpha + n - 1}. \tag{3.6}$$

Notice that θ_n will be equal to $\theta_k^\star$ with probability proportional to n_k, the number of times it has already been observed. The larger n_k is, the higher the probability that it will grow. This is a rich-gets-richer phenomenon, where large clusters (a set of θ_i's with identical values $\theta_k^\star$ being considered a cluster) grow larger and larger. This leads to the fact that the draws from a DP(α, H) are discrete with probability 1, the celebrated clustering property of the Dirichlet Process.

The polya urn representation of the DP can also be reinterpreted as a Chinese restaurant process (CRP). The CRP(α) is a single parameter distribution over partitions of integers. The idea of this representation is the following: suppose a restaurant have infinite number of tables and sequence of customers labeled by $\{1,\ldots n\}$, the first customer sits at the first table and then each new customer joins a table populated by n_j customers with probability $\frac{n_j}{\alpha+n}$, where n is the overall number of customers that already entered the restaurant, or can sit at a new table with probability $\frac{\alpha}{\alpha+n}$. The CRP is a random partition induced by DP. With the given description it is easy to see the similarity with Polya Urn representation, where taking a particular occupied table is similar to choose an existing colour. Sethuraman (1994) proposed a constructive representation of the Dirichlet process: the stick-breaking construction. Consider two independent families $\{V_k\}$ and $\{Z_k\}$ of random variables:

$$V_k \overset{iid}{\sim} Be(1,\alpha) \qquad Z_k \overset{iid}{\sim} H \qquad k = 1,2,\ldots. \tag{3.7}$$

We can now define the random weights as:

$$\begin{aligned} p_1 &= V_1, \\ p_k &= V_k \prod_{j=1}^{k-1}(1-V_j), \qquad k = 2, \dots \end{aligned} \tag{3.8}$$

The construction of p can be understood metaphorically as follows. Starting with a stick of length 1, we break it at V_1, assigning p_1 to be the length of stick we just broke off. Now we recursively break the other portion proportionally to V_2, V_3, to obtain p_2, p_3 and so forth. Notice that $\sum_{k=1}^{\infty} p_k = 1$ *a.s.*.
Then the stick-breaking representation of the Dirichlet process can be written as:

$$G := \sum_{k=1}^{\infty} p_k \delta_{Z_k}. \tag{3.9}$$

The three different representations of DP can be united using the de Finetti theorem. This theorem is based on the notion of exchangeability.

Definition 2 (Exchangeability of random sequence)**.** A random process $(\theta_1, \theta_2, \dots)$ is called infinitely exchangeable if for any finite $n \in \mathbb{N}$ and any permutation σ on $1, \dots, n$, the joint probability of $(\theta_1, \theta_2, \dots, \theta_n)$ is equal to the one of the permuted vector $(\theta_{(1)}, , \dots, \theta_{\sigma(n)})$:

$$P(\theta_1, \dots, \theta_n) = P(\theta_{\theta_{\sigma(1)}, \dots, \theta_{\sigma(n)}}). \tag{3.10}$$

Theorem 2 (de Finetti's Theorem)**.** If a random process $(\theta_1, \theta_2, \dots)$ is infinitely exchangeable, then the joint probability $p(\theta_1, \theta_2, \dots, \theta_n)$ could be written as:

$$P(\theta_1, \dots, \theta_n) = \int \left\{ \prod_{i=1}^{n} G(\theta_i) \right\} dP(G), \quad \text{for any } n \tag{3.11}$$

for some random variable G.

By considering the Polya Urn representation of the DP that we have described before, it can be proved that the random sequences generated by these processes are exchangeable, and that the underlying distribution is the Dirichlet Process.

3.1.2 Dirichlet process approximation

The infinite dimension of the stick-breaking representation comes with a cost: we can not directly sample from it. There are two possible ways to solve this problem.
The first one is to marginalize out the base random measure, using the Pólya Urn sampling scheme. Indeed thanks to the Blackwell and MacQueen (1973) Pólya Urn characterisation the random variables $Y_1, \ldots, Y_n$ are exchangeable and it holds that:

$$p(Y_i|\boldsymbol{Y}_{-i}) = \frac{\alpha}{\alpha+n-1}G_0 + \frac{1}{\alpha+n-1}\sum_{j=1}^{n-1}\delta_{Y_j} \qquad \forall i = 1, \ldots, n. \tag{3.12}$$

This possible solution is easily implementable in a Gibbs sampler, but has several drawbacks. The most important one is the slow mixing of the chains, and the tendency of the algorithm to get stuck for several iterations. When this occurs, the sampler can get stuck at the current unique values $Y_1^\star ... Y_m^\star$ of $\boldsymbol{Y}$ and it may take many iterations before any new Y values are generated (West and Escobar, 1993). The second solution is to use the stick-breaking representation of the Dirichlet process, and to truncate the infinite sum at some truncation level N.

Ishwaran and James (2001) defined $\mathscr{P}_{\mathscr{N}}(\boldsymbol{a}, \boldsymbol{b})$ as stick-breaking random measure if

$$\mathscr{P}(.) = \sum_{k=1}^{N} p_k \delta_{Z_k}(.) \tag{3.13}$$

and $p_1 = V_1$, $p_k = (1-V_1)(1-V_2)..(1-V_{k-1})V_k, k \geq 2$, where $V_i \overset{\perp\!\!\!\perp}{\sim} \text{Beta}(a_k, b_k)$, $a_k, b_k > 0$, $\boldsymbol{a} = (a_1, \ldots), \boldsymbol{b} = (b_1, \ldots)$.
The Dirichlet process is a special case of stick-breaking random measure where $N = \infty$ and the parameters a_k and b_k are defined as $a_k = 1$, $b_k = \alpha$ $\forall k$. By consequence the random variables V_k are i.i.d. If $N < \infty$ then $\mathscr{P}_{\mathscr{N}}(\mathbf{a,b})$ is a finite dimensional prior. In this case, in order to have $\sum_k p_k = 1$, one needs to have V_N=1.
The weights $p = (p_1, ., p_N)$ defined with the stick-breaking method have a Generalized Dirichlet distribution (Connor and Mosimann, 1969) :

$$p = (p_1, ., p_N) \sim \mathscr{G}\mathscr{D}(\mathbf{a,b}). \tag{3.14}$$

Where the density for $\mathscr{GD}(\mathbf{a,b})$ is :

$$\prod_{k=1}^{N-1} \frac{\Gamma(a_k+b_k)}{\Gamma(a_k)\Gamma(b_k)} p_1^{a_1-1} \cdots p_{N-1}^{a_{N-1}-1} p_N^{b_{N-1}-1} (1-P_1)^{b_1-(a_2+b_2)} \dots (1-P_{N-2})^{b_{N-2}-(a_{N-1}+b_{N-1})} \tag{3.15}$$

where $P_k = p_1 + \ldots + p_k$.

An important property of this distribution is that if a vector of random variables follow a Dirichlet distribution $p = (p_1, \ldots, p_N) \sim \text{Dir}(a_1, \ldots, a_N)$, then p has a $\mathscr{GD}(\mathbf{a}, \mathbf{b})$ distribution with $\mathbf{a} = (a_1, \ldots, a_{N-1})$ and $\mathbf{b} = (\sum_{k=2}^{N} a_k, \cdots, a_N)$.
This means that if $\mathscr{P} = \sum_{k=1}^{N} p_k \delta_{Z_k}(.)$ is a random measure where the weights follows a Dirichlet distribution

$$p = (p_1, \ldots, p_N) \sim \text{Dir}(a_1, \ldots, a_N), \tag{3.16}$$

then and $\mathscr{P}$ is a $\mathscr{P}_N(a,b)$ stick-breaking measure with $V_k \sim \text{Beta}(a_k, \sum_{s=k+1}^{N} a_s)$. Moreover, as the Dirichlet distribution, the Generalized Dirichlet distribution is conjugate with the multinomial distribution.
Working with a Dirichlet distribution is very convenient thanks to its conjugacy with the multinomial density. By theorem 4.19 in Ghosal and Van der Vaart (2017) if $\mathscr{P}_N(a,b)$ is a stick-breaking measure with Dirichlet weights $p = (p_1, \ldots, p_N) \sim \text{Dir}(\frac{\alpha}{N}, \ldots, \frac{\alpha}{N})$, then:

$$\mathscr{L}\{\mathscr{P}_N(g)\} \to \mathscr{L}\{\mathscr{P}_\infty(g)\}, \tag{3.17}$$

where $\mathscr{P}_\infty$ is a $\text{DP}(\alpha H)$. That is, $\mathscr{P}_N$ converges weakly in distribution to a DP, and it is thus a good approximation of a Dirichlet process. Due to this property, we will call this approximation of the DP the weak limit representation. As suggested in Ishwaran and Zarepour (2000), even for a very large number of elements, a moderate level of truncation should suffice to approximate a $DP(\alpha H)$.

We are now interested in sampling from the posterior distribution of the following hierarchical model with a Bayesian nonparametric prior:

$$\begin{aligned} (X_i \mid Y_i) &\overset{iid}{\sim} p(X_i \mid Y_i) \\ (Y_i \mid G) &\overset{iid}{\sim} G \\ G &\sim \mathscr{P}. \end{aligned} \tag{3.18}$$

In GJAM model the X_i are the rows of a matrix, and the base measure H is a multivariate Gaussian distribution, but we will fully discuss about this later in Section .

By rewriting the model under the approximation described above, we have:

$$\begin{aligned}(X_i \mid Z, K) &\overset{iid}{\sim} \pi(X_i \mid Z_{K_i})\\ (K_i \mid p) &\overset{iid}{\sim} \sum_{k=1}^{N} p_k \delta_k(.)\\ (p, Z) &\sim \pi(p)\pi(Z),\end{aligned} \tag{3.19}$$

Where $K = (K_1, \dots, K_n)$, are the conditionally independent classification variable. As we have seen before, p follows a Dirichlet distribution, $p \sim Dir(\frac{\alpha}{N}, \dots, \frac{\alpha}{N})$, that can be rewritten in terms of a $\mathscr{GD}(\mathbf{a}, \mathbf{b})$ distribution, where $a_k = \frac{\alpha}{N}$ and $b_k = \frac{\alpha(N-k)}{N}$ $k = 1, \dots, N-1$. We exploit the conjugacy of the latter distribution with the multinomial distribution to sample from the full conditional of p:

$$f(p \mid K) \propto f(K \mid p) f(p) \sim \mathscr{GD}(\mathbf{a'}, \mathbf{b'}), \tag{3.20}$$

where $\mathbf{a'} = (\frac{\alpha}{N} + m_1, \dots, \frac{\alpha}{N} + m_{N-1})$ and $\mathbf{b'} = (\frac{\alpha(N-1)}{N} + \sum_{i=2}^{N} m_i, \dots)$ where $m_k = card\{K_i = k\}$.
By consequence (the detailed computations of the full conditional can be seen in Ishwaran and Zarepour (2000)), the sampling steps will be the following:

- $k_i \overset{iid}{\sim} \sum_{j=1}^{N} p_{j,i} \delta_i(k_i)$
  $p_{j,i} \propto p_j \times p(x_i \mid Z_j)$

- $p_1 = V_1,\, p_k = V_k \prod_{l=1}^{k-1}(1 - V_l),\; p_N = 1 - p_1 - \dots - p_{N-1}$
  $V_k \overset{iid}{\sim} \text{Beta}(\frac{\alpha}{N} + m_k, \frac{\alpha(N-K)}{N} + \sum_{j=k+1}^{N} m_j)$, where $m_k = card\{i : k_i = k\}$

- $Z_j \overset{iid}{\sim} H,\, j \notin \boldsymbol{k}$
  $Z_j \overset{iid}{\sim} H(dZ_j) \prod_{i:k_i=j} p(x_i \mid Z_j)$, $j \in \boldsymbol{k}$

In the next paragraph we are going to describe the DP application to GJAM, and we will describe the related sampling scheme. Since Taylor-Rodriguez et al. (2017) used this weak limit approximation truncation method, and the description of its Gibbs sampler will be an example of the sampling scheme described above.

3.1.3 The DP application to GJAM

Starting from the full GJAM model described in Section 2.1, we are going to build the dimension reduction model proposed by Taylor-Rodriguez et al. (2017). We want to model a latent variable $V_i \in \mathbb{R}^S$ (previously called $z_i,.$) where S is the number of species (previously called J) in every site $i = 1, \dots, n$. As in (2.3), we model V_i as:

$$V_i = Bx_i + e_i \qquad \text{with } \varepsilon_i \sim MVN(0, \Sigma), \tag{3.21}$$

where B is a $N \times K$ matrix that contains the regression coefficients and Σ is the $S \times S$ matrix of the residual correlations (previously called R).

Since the size of Σ grows quadratically with S, we rewrite the model using latent factors:

$$V_i = Bx_i + Aw_i + \varepsilon_i, \tag{3.22}$$

where the random vectors w_i are iid with $w_i \sim MVN(0, I_r)$ and $\varepsilon_i \sim MVN(0, \sigma_\varepsilon^2 I_s)$. Notice that in Section 2.3 we called Λ the matrix of the latent loadings A, and η_i the latend loadings w_i. The number of latent factors r gives the size of the $S \times r$ matrix A. The model is actually the same as above, but the residual correlation Σ is approximated with $\Sigma = AA^T + \sigma_\varepsilon^2 I$.

Taylor-Rodriguez et al. (2017) made use of the clustering property of the Dirichlet process to allow some rows of A to be common, which corresponds to clustering species in their dependence behavior. Using the stick-breaking representation and the truncation defined in Section 3.1.2, we have:

$$\begin{gathered} DP_N(\alpha H) = \sum_{j=1}^{N} p_j \delta_{Z_j}, \\ \text{with } p \sim \mathscr{GD}(a_\alpha, b_\alpha), a_\alpha = (\tfrac{\alpha}{N}, \dots, \tfrac{\alpha}{N}) \text{ and } b_\alpha = (\tfrac{\alpha(N-1)}{N}, \tfrac{\alpha(N-1)}{N}, \dots, \tfrac{\alpha}{N}) \end{gathered} \tag{3.23}$$

The atoms of the stick-breaking representation are $(Z_j^T)_{J=1}^N$ (with $Z_j \overset{iid}{\sim} H$) that form the lines of the $N \times r$ matrix $\boldsymbol{Z} = (Z_j^T)_{J=1}^N$ representing all the vector values that the lines of A may take.

The base measure H is such that $Z_j | D_z \overset{iid}{\sim} MVN(0, D_z)$. The prior specification for the $r \times r$ matrix D_z will be given below and it follows the noninformative strategy to sample covariance matrices described in Huang and Wand (2013).

In this setup, we need a vector of grouping labels $k = (k_1, \dots, k_s)(1 \le k_l \le N)$ such that $a_l = Z_{k_l}$. Now A can be represented as $A = Q(k)\boldsymbol{Z}$ where $Q(k) = (\boldsymbol{e}_{k_1}, \dots, \boldsymbol{e}_{k_s})^T$ is a $S \times N$ matrix where $\boldsymbol{e}_{k_l}$ is the N-dimensional vector with a 1 in position k_l and 0's

elsewhere.
Using this notation, the model is:

$$
\begin{aligned}
V_i|k,\boldsymbol{Z},w_i,B,\sigma_\varepsilon^2 &\overset{iid}{\sim} MVN(Bx_i+Q(k)\boldsymbol{Z}w_i,\sigma_\varepsilon^2 I_s, \qquad \text{for } i=1,\dots,n,\\
(B,\sigma_\varepsilon^2) &\propto \frac{1}{\sigma_\varepsilon^2},\\
w_i &\overset{iid}{\sim} MVN(0,I_r),\\
k_l|\boldsymbol{p} &\overset{iid}{\sim} \sum_{j=1}^{N} p_j\delta_{j(k_l)}, \qquad for\ j=1,\dots,S,\\
Z_j|D_z &\overset{iid}{\sim} MVN(0,D_z), \qquad for\ j=1,\dots,N,\\
\boldsymbol{p} &\sim \mathscr{GD}_N(a_\alpha,b_\alpha), \qquad \text{with } \ a_\alpha=(\tfrac{\alpha}{N},\dots,\tfrac{\alpha}{N})\\
&\qquad\qquad\qquad\qquad \text{and } \ b_\alpha=(\tfrac{\alpha(N-1)}{N},\tfrac{\alpha(N-1)}{N},\dots,\tfrac{\alpha}{N}),\\
D_z &\sim IW(2+r-1,4\mathrm{diag}(1/\eta_1,\dots,1/\eta_r)),\\
\eta_h &\overset{iid}{\sim} IG(1/2,1/10^4), \text{ for } h=1,\dots,r.
\end{aligned}
\tag{3.24}
$$

It is important to notice that both the regression coefficients B and the standard deviations σ_ε^2 have an improper prior (whose posterior is proper). To sample from the posterior of the parameters of the models, the authors use a Gibbs Sampler described in the Appendix A of Taylor-Rodriguez et al. (2017).

3.2 Motivations and improvement directions

We have already discussed the advantages of the dimension reduction method proposed by Taylor-Rodriguez et al. (2017). But what are its limits?

First of all, the concentration parameter α is treated as a constant, equal to the number of species S, but α has an important impact on the model. In the stick-breaking representation of the DP, the concentration parameter rules the distribution on the random weights. When α is small, the Beta-distributed variables V_i tend to be closer to 1 than to 0, implying that the random weights decrease (in expectation) really fast. Viceversa, when α is large, the random weights decrease really slow in expectation. Therefore, the realisations will be more concentrated in few clusters when alpha is small, and the number of clusters increases with α. This is confirmed by the Pólya Urn representation. Indeed, given the previous n observations, a new sample will be assigned to a new cluster with probability $\alpha/\alpha+n-1$. The bigger α, the higher the probability of creating new clusters.

In GJAM, the clusters are the species that share the same behaviour with respect to the other species in the residual covariance matrix. In many real applications it is possible to have a prior knowledge on the species interaction network, even if this information is often qualitative, or very weak. A useful information that ecologists have is the number of groups of species in the network. At best, this could be known by applying a stochastic block model (Lee and Wilkinson, 2019) on the interactions graph. However, we can also retrieve this information from other sources, for example the functional traits. Indeed, it is well known that species with similar traits tend to compete. Under this hypothesis, one could cluster the species depending on their functional traits, and use the obtained number of clusters as a prior information for the expected number of groups of species in GJAM.

Our aim is to exploit the influence of α on the clustering properties of the DP to improve the model, by fixing alpha such that the prior expectation on the number of clusters matches our prior knowledge on the number of groups of species that share the same behaviour with respect to other species. Even if the groups in the residual correlation matrix are not necessarily the ones of the interaction networks, we believe that this information could however improve the abilities of the model.

Another further improvement concerns the hierarchical structure of the model. One of the advantages of Bayesian hierarchical modelling is the possibility of assigning a prior distribution to the hyperparameters, treating them as random variables. Instead of fixing α, we can leave it to be learnt from the data. We can work on the prior distribution for α, in order to have that its prior mean matches the concentration parameter determined above. That is, we will treat α as a random parameter, but the prior expected number of groups will still match the prior knowledge on the number of clusters in the species interaction network.

Another possible improvement of the non parametric method of Taylor-Rodriguez is the extension of the Dirichlet process to the Pitman–Yor process (PY) (Ferguson, 1973), that is the natural extension of the DP. We are going to describe the PY process later, but for now what is important to say is that PY is a stick-breaking random measure where the Generalized Dirichlet distribution of the weights depends on two parameters: the concentration (α) and the discount (σ) parameters. The distribution of the number of distinct clusters has a heavy-tailed power law with the number of observations, meaning that, compared to the DP (where the number of distinct clusters grows logarithmically with n), the PY process leads to a higher number of distinct clusters.

Since GJAM suffers from slow convergence, we want to have an efficient sampling scheme. As for DP, PY can be implemented with a generalized Pólya urn sampling scheme, but this leads to slow mixing. We implemented a truncated PY process, based on the results of Arbel et al. (2019), but by fixing the truncation level N in a novel approximating way that leads to a faster sampling scheme. Thanks to recent developments in the statistical literature De Blasi et al. (2015), the mean of the prior number of clusters for fixed α and σ is available in closed form. We want to fix the hyperparameters α and σ such that the prior mean of the number of clusters is fixed to the number of clusters known a priori.

For the same motivations as before, we added a hierachical layer to the model, by letting α and σ depend on data. Again, we want to choose their prior distributions such that the prior mean of α and σ will match the values found before.

In the following sections we will give a full description of the extensions that we gave to the model, as well as their implementation in the Gibbs sampler.

3.3 GJAM improvements related to DP

The first improvement of the method proposed by Taylor-Rodriguez et al. (2017), is to carefully choose the concentration parameter α.

Suppose from now on that we have a prior knowledge on the number of groups of species that share the same behaviour with respect to the other species, and call this number $\bar{K}$.

Call $K_{n,\alpha}$ the discrete random variables that represents the number of distinct values of n realizations. The law of $K_{n,\alpha}$ is given by (Antoniak, 1974) and depends on α:

$$P(K_{n,\alpha} = k) = \frac{\alpha^k \Gamma(\alpha)}{\Gamma(\alpha + n)} |s(n,k)|, \tag{3.25}$$

where $|s(n,k)|$ is the Stirling number of the first type. When n is large, $\mathbb{E}[K_{n,\alpha}]$ behaves as $\alpha \log(n)$.

As it was shown in De Blasi et al. (2015) (see Section 3), this distribution is highly peaked, therefore the specification of α should be done carefully, as the posterior should not be able to move far away from a bad defined prior distribution, espacially when the data sample size is small. Concerning the mean of this distribution, it can be easily calculated thanks to the Pólya urn representation of the DP. Indeed, in (3.6) it is clear that each new realisation, given the previous $k-1$, can take a new value (i.e. a new cluster) with probability $\frac{\alpha}{\alpha+k-1}$.

The mean of the prior number of distinct values among n realisations of a DP is then just the sum over this value:

$$\mathbb{E}[K_{n,\alpha}] = \sum_{k=1}^{n} \frac{\alpha}{\alpha + k - 1} \tag{3.26}$$

The illustration of mean of prior number of distinct values for different values of α and n could be seen in the appendix Section (A.2.6), on the Figure A.5 Our aim is to fix $\alpha = \bar{\alpha}$, where $\bar{\alpha}$ is such that $\mathbb{E}[K_{n,\bar{\alpha}}] = \bar{K}$.
Since it is not possible to analytically retrieve $\bar{\alpha}$ from (3.26), we will approximate it numerically, using the bisection algorithm, that proved to work really well due to the monotonicity of (3.26).

We believe in the importance of α on the results of our model, and we thus want it to learn from data. Therefore, a hierarchical layer is added by putting a prior distribution on the concentration parameters α of the DP, and by fixing its related hyperparameters so as to let the expected prior number of clusters match our ecological prior knowledge on the number of clusters (defined as $\bar{K}$ above).
The first strategy is to use the same DP truncation used in GJAM, the weak limit representation described in Section 3.1.2. Then, we will describe another approximation (the finite DP introduced by Ishwaran and Zarepour, 2000) that will lead to an easier sampling scheme for the posterior of α.

3.3.1 Prion on α using the weak limit representation of the DP

The description of the weak limit representation given in Section 3.1.2 is characterised by the Dirichlet distributed weigths:

$$p|\alpha \sim Dir(\tfrac{\alpha}{N}, \dots, \tfrac{\alpha}{N}). \tag{3.27}$$

The conditional density of $\alpha|p$ is then:

$$\pi(\alpha|p) \propto \frac{\Gamma(\alpha)}{\Gamma(\alpha/N)^{N}} p_1^{\alpha/N-1} \dots p_N^{\alpha/N-1} \pi(\alpha). \tag{3.28}$$

In the literature the gamma distribution is common choice as a prior for the concentration parameter α (Ishwaran et Zarepour 2000):

$$\alpha \sim Ga(\nu_1, \nu_2) \tag{3.29}$$

The Gibbs sampler for this model is the same one of (3.24) with an additional step to sample from the full conditional of α :

$$\alpha|p \propto \frac{\Gamma(\alpha)}{\Gamma(\alpha/N)^N} p_1^{\alpha/N-1} \dots p_N^{\alpha/N-1} \alpha^{\nu_1-1} e^{-\nu_2\alpha} \tag{3.30}$$

To do this, we need a Metropolis–Hasting step in our Gibbs sampler. We modified the Gibbs sampler of (3.24) implemented in the GJAM package and added a Random Walk Metropolis–Hasting step for the concentration parameter α, with a truncated normal as proposal, since our target distribution has support on $\mathbb{R}^+$. The variance of the proposal density will be adjusted a posteriori by checking the convergence metrics.

We now want to fix the hyperparameters of the prior distribution of α in order to have the desired expected prior number of clusters. We have:

$$\alpha \sim Ga(\nu_1, \nu_2), \tag{3.31}$$

and we want:

$$\mathbb{E}[\alpha] = \frac{\nu_1}{\nu_2} = \bar{\alpha}, \tag{3.32}$$

where $\bar{\alpha}$ is the one found above.

We want to have a quite large variance in order to let α move away from this prior specification if indicated by the data.

Therefore, we set:

$$Var[\alpha] = \frac{\nu_1}{\nu_2^2} = \frac{\bar{\alpha}}{\nu_2} = 20, \tag{3.33}$$

(we can eventually modify the number 20)

From which it follows:

$$\begin{cases} \mathbb{E}[\alpha] = \frac{\nu_1}{\nu_2} = \bar{\alpha} \\ Var[\alpha] = \frac{\nu_1}{\nu_2^2} = \frac{\bar{\alpha}}{\nu_2} = 20 \end{cases} \implies \begin{cases} \nu_1 = \frac{\bar{\alpha}^2}{20} \\ \nu_2 = \frac{\bar{\alpha}}{20} \end{cases} \tag{3.34}$$

Concerning the truncation level N, since we are not changing the way we approximate the DP, we are going to stick to the indication of the author and take $N = min\{S, 150\}$.

3.3.2 Prion on α using the almost sure truncation of a DP

Ishwaran and Zarepour (2000) proposed another approximation of the Dirichlet process, called the almost sure truncation.

They proposed to consider equation (3.13) and to take all $V_i \sim Be(1,\alpha)$ and $V_N = 1$, to guarantee that the weights sum to one. Which is equivalent of setting $p_N = 1 - p_1 - \ldots - p_{N-1}$.

By consequence:

$$p \sim \mathscr{GD}(\boldsymbol{a},\boldsymbol{b}), \text{ with } a_i = 1, b_i = \alpha \ \forall i = 1,\ldots,N \tag{3.35}$$

Ishwaran and James (2001) established an upper bound of N by considering the Bayesian marginal density. Indeed, if the random measure $\mathscr{P}_N(\boldsymbol{a},\boldsymbol{b})$ is applied in a Bayesian hierarchical model as a prior, then an appropriate method for selecting N is to choose a value that yields a Bayesian marginal density that is almost indistinguishable from its limit .

Being $\boldsymbol{\mu}_N$ the marginal density of the observed data under the $\mathscr{P}_N(\boldsymbol{a},\boldsymbol{b})$ prior and $\boldsymbol{\mu}_\infty$ the marginal density of the observed data under DP(α H), the $\mathscr{L}_1$ norm of the difference between the two marginal densities can be approximated as (theorem 2 in Ishwaran and James (2001)):

$$\|\boldsymbol{\mu}_N - \boldsymbol{\mu}_\infty\|_1 \sim 4n\, exp(-(N-1)/\alpha), \tag{3.36}$$

that shows that the sample size n has almost no effect compared to the truncation threshold N. In particular, the authors suggest to take $N = 150$ to guarantee a good convergence for any n and α.

An important advantage of this approximation is that when weights p are distributed as in (3.35), then the prior Gamma distribution for α is conjugate. Indeed:

$$\begin{aligned} &\pi(p \mid \alpha) \propto \alpha^{N-1} p_N^{\alpha-1} \qquad \text{implies :} \\ &\pi(\alpha \mid p) \propto \pi(p \mid \alpha)\pi(\alpha) \propto \alpha^{N-1} p_N^{\alpha-1} \alpha^{\nu_1-1} e^{-\nu_2\alpha} \propto \text{Ga}(N+\nu_1-1, \nu_2 - \log p_N), \end{aligned} \tag{3.37}$$

where ν_1 and ν_2 are defined as in (3.34) for the same reasons.

We implemented this model in R by modifying the Gibbs sampler of (3.24) implemented in the GJAM package (Clark, 2017). In this new Gibbs sampler we sampled from α using (3.37), and the weights using:

$$\begin{aligned} p_1 = V_1,\ p_k = V_k \prod_{l=1}^{k-1}(1-V_l),\ p_N = 1 - p_1 - \ldots - p_{N-1}, \\ V_k \sim \text{Beta}(1+m_k, \alpha + \sum_{j=k+1}^{N} m_j), \text{where } m_k = card\{i : k_i = k\}. \end{aligned} \tag{3.38}$$

Even if this truncation has good approximation properties only for big N, compared to the weak limit representation, it comes with the great advantage of the conjugacy for α.

3.4 GJAM improvements related to the Pitman–Yor process

3.4.1 Pitman–Yor description

After having explored all the possible ways to increase the flexibility of the clustering properties of GJAM, we decided to replace the Dirichlet Process with the Pitman–Yor (PY) process. The Pitman–Yor process is a generalization of the DP process, and is also a special case of a larger class of priors called Gibbs-type priors, which were introduced in the seminal works of Pitman and Gnedin Pitman (2003); Gnedin and Pitman (2006), see De Blasi et al. (2015) for a review.
The Pitman–Yor process was firstly defined by Pitman:

Definition 3 (Pitman–Yor process)**.** The random measure $G = \sum_{k=1}^{\infty} p_k \delta_{Z_k}$ defined by the weights:

$$\begin{aligned} & V_k \overset{ind}{\sim} \text{Beta}(1-\sigma, \alpha + k\sigma) \\ \text{with} \quad & p_1 := V_1, p_k := V_k \prod_{j=1}^{k-1}(1-V_j) \\ \text{and} \quad & Z_1, Z_2, \ldots, \overset{iid}{\sim} H \end{aligned} \tag{3.39}$$

is called Pitman-Yor process with concentration parameter α, discount (or diversity) parameter σ (with $0 \le \sigma \le 1$) and base measure H.

Notice that when $\sigma = 0$ it holds: $PY(\alpha, 0, H) = DP(\alpha H)$.
The difference in terms of the stick-breaking representation is in the sampling of the beta distributed variables V_k that now depend on two parameters. These variables are no more identically distributed, and when k increases, the second parameter increases,

leading to a decrease in the mean value of V_k. The V_k decrease (in mean) when k grows, meaning that the weights p_k will decrease (in mean), slower than in the DP. That implies that the PY process have a greater number of distinct clusters.
Indeed, when the data are modelled with a Dirichlet process, the number of obtained distinct clusters K_n grows logaritmically with n, while for the Pitman–Yor process K_n grows as a power law with n. This power law property is a more natural assumption for many applications, since it means that we have a small number of clusters with a high number of observations, and a large number of clusters with only few observations (a good example is the twitter users and their number of followers).

The Pitman–Yor process can also be described in terms of the generalized Pólya urn representation.
Let G be a sample from a PY process, and let $\theta_1, \dots, \theta_n$ be realisations from G. That is:

$$\begin{aligned} \theta_1, \dots, \theta_n | G &\overset{iid}{\sim} G \\ G &\sim PY(\alpha, \sigma, H). \end{aligned} \tag{3.40}$$

By marginalizing out the base measure G, the predictive law of a new realisation given the previous n realisation is:

$$\theta_{n+1} | \theta_n, \dots, \theta_1 \sim \frac{(\alpha + \sigma K_n) H + \sum_{j=1}^{n} (\mathcal{N}_j^n - \sigma) \delta_{\theta_j}}{\alpha + n}, \tag{3.41}$$

Where K_i is the number of distinct clusters between the first n observations, and $\mathcal{N}_j^n$ is the number of elements in the j-th clusters after n realisations. The probability that a new realisation, given the previous n, enters in a new cluster is $\frac{\alpha+\sigma K_n}{\alpha+n}$, which is greater than the one given by a DP base measure, confirming what stated above.
We are now going to describe how we can approximate a PY process to be able to sample from it.

3.4.2 Truncation of a Pitman–Yor process

As for the Dirichlet Process, the Pitman–Yor process has infinite parameters, and thus we can not directly sample from it.
A possible solution is to marginalize out the base random measure, using the Pólya Urn sampling scheme and exploiting the exchangeability of the realizations. However, as for the DP, this leads to slow mixing (West and Escobar, 1993).
The second solution is to use the stick-breaking representation of the Pitman–Yor

process, and to truncate the infinite sum at some truncation level N, that is:

$$\mathscr{P}_N(.) = \sum_{k=1}^{N} p_k \delta_{Z_k}(.), \tag{3.42}$$

and $p_1 = V_1$, $p_k = (1-V_1)(1-V_2)\dots,(1-V_{k-1})V_k, k \geq 2$, where $V_k \overset{\perp\!\!\!\perp}{\sim} \text{Beta}(1-\sigma, \alpha + k\sigma)$ and $V_N = 1$.

Once the truncation level is fixed, the sampling scheme is the same described in (3.38),thanks to the conjugacy of the $\mathscr{GD}(\boldsymbol{a}, \boldsymbol{b})$ distribution. However, instead of having

$$a_k = 1, b_k = \alpha \;\; \forall k = 1, \dots, N, \tag{3.43}$$

the parameters will be:

$$a_k = 1-\sigma, b_k = \alpha + k\sigma \;\; \forall k = 1, \dots, N. \tag{3.44}$$

Therefore, the step to sample the weights in the Gibbs sampler of (3.24) is changed to:

$$\begin{gathered} p_1 = V_1,\ p_k = V_k \prod_{l=1}^{k-1}(1-V_l),\ p_N = 1 - p_1 - \ldots - p_{N-1}, \\ V_k \sim \text{Beta}(1-\sigma+m_k, \alpha + k\sigma + \sum_{j=k+1}^{N} m_j),\ \text{where } m_k = card\{i : k_i = k\}. \end{gathered} \tag{3.45}$$

Nevertheless, the truncation error for PY is harder to bound than DP, and there's no result about a weakly convergence in limit as for the DP.

The key quantity to consider is thus the approximation error

$$R_n = \sum_{i>n} p_i = \prod_{j \leq n}(1-V_j), \tag{3.46}$$

since when R_n is small the resulting truncated process $\mathscr{P}_N$ will be close to PY according to $|PY(A) - \mathscr{P}_N(A)| \leq R_n$ for any measurable set A. Ishwaran and James (2001) proposed to determine the truncation level based on the moments of R_n. Instead, Arbel et al. (2019) investigate a random truncation by setting n such that R_n is smaller than a predetermined value $\varepsilon \in (0,1)$ with probability one. The authors define the ε-PY process as the Pitman–Yor process truncated at $n = \tau(\varepsilon)$, where $\tau(\varepsilon) = \min\{n \geq 1 : R_n < \varepsilon\}$ and R_n is the truncation error

$$R_n = \sum_{i>n} p_i. \tag{3.47}$$

Even if for studying the asymptotic behaviour of $\tau(\varepsilon)$ the common notations for the precision parameter is θ and for discount it is α, we decided to stick to our previous notation and to call α the precision parameter and σ the discount. The authors proved that for the Pitman–Yor process with parameters α, σ the asymptotic distribution of $\tau(\varepsilon)$ when $\varepsilon \to 0$ is defined as (see Theorem 2 in Arbel et al., 2019):

$$\tau(\varepsilon) - 1 \sim_{a.s} (\varepsilon T_{\sigma,\alpha}/\sigma)^{-\frac{\sigma}{(1-\sigma)}} \text{ as } \varepsilon \to 0, \tag{3.48}$$

where $T_{\sigma,\alpha}$ is a polynomially tilted random variable with density :

$$\frac{\Gamma(1+\alpha)}{\Gamma(1+\alpha/\sigma)} x^{-\alpha} f_\sigma(x), \tag{3.49}$$

where $f_\sigma(t)$ is the density of a positive stable random variable S_σ with exponent σ, which satisfies $\mathbb{E}(e^{-sT\sigma}) = e^{-s\sigma}$.

The authors suggested two algorithms to sample from a $\varepsilon - PY$ process. However, even the fastest algorithm (Algorithm 2) still needs a supplementary sampling step inside the Gibbs sampler, to sample from $T_{\sigma,\alpha}$.

Our approach was to fix the truncation level N to a value that guaranteed a bound to the probability of $\tau(\varepsilon)$ to exceed N. Because of the properties of its distribution, it was not easy to fix this bound.

We first wanted to use the mean of $\tau(\varepsilon)$ as estimator for the truncation number. However, the shape of distribution $(\varepsilon T_{\sigma,\alpha}/\sigma)^{-\frac{\sigma}{(1-\sigma)}}$ strongly depends on the parameters α, σ, and can become very skewed. Figure 3.1 shows this variation, and how skewness and kurtosis grow when σ increases. More precisely, the graphs in Appendix in Section A.2.5 , Figure (A.4) show the skewness and kurtosis on grid for α, σ.

Because of the shape of the distribution, the mean value could not be the optimal estimator. Indeed $\tau(\varepsilon)$ can take values much greater then the mean with a considerable probability, inducing an important error in truncation.

We thus decided to approximate N with the 95%-quantile of $\tau(\varepsilon)$ to bound the probability that the truncation error exceeds a desired threshold.

To work this out, we studied the moments of $\tau(\varepsilon)$: if we could find the analytical expression for the moments, we could find a desirable tractable expression which would be close to the 95% quantile, in order to have $P(\tau(\varepsilon)) > N) < 0.05$.

The distribution of the truncation number $\tau(\varepsilon)$ depends on the parameters $\varepsilon, \sigma, \alpha$. We defined the moments of the truncation number as $M^k_{\varepsilon,\sigma,\alpha} = \mathbb{E}[\tau^k_{\varepsilon,\sigma,\alpha}] = \mathbb{E}[\tau(\varepsilon)^k]$.

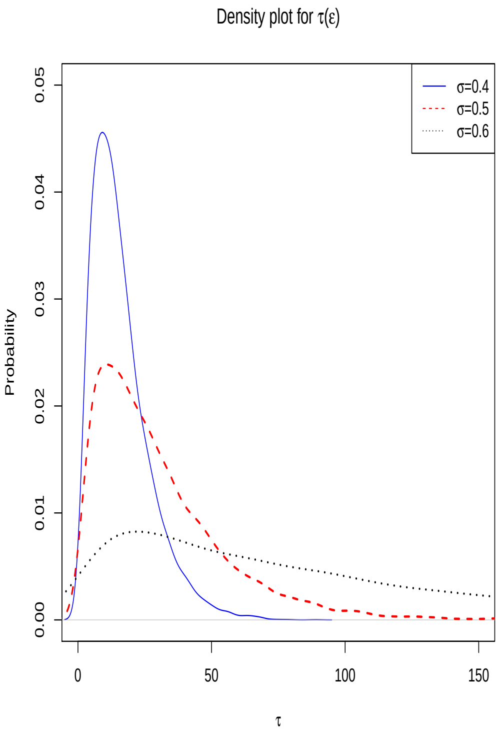

Figure 3.1 – *Density plot for approximation of $\tau(\varepsilon)$ for values $\sigma = \{0.4, 0.5, 0.6\}$, $\alpha = 1$ and $\varepsilon = 0.1$, computed for $n = 10^4$.*

For a fixed precision ε and fixed parameters σ, α, we computed $M^k_{\varepsilon,\sigma,\alpha}$ using (3.48):

$$\mathbb{E}[\tau(\varepsilon)^k] \approx \mathbb{E}[((\varepsilon T_{\sigma,\alpha}/\sigma)^{-\sigma/(1-\sigma)})^k] = (\varepsilon/\sigma)^{-k\sigma/(1-\sigma)}\mathbb{E}[T_{\sigma,\alpha}^{-k\sigma/(1-\sigma)}] \text{ as } \varepsilon \to 0 \tag{3.50}$$

Then, using the fact that $E[S_\sigma^{-r}] = \frac{\Gamma(1+r/\sigma)}{\Gamma(1+r)}$:

$$\begin{aligned}\mathbb{E}[T_{\sigma,\alpha}^{\frac{-k\sigma}{(1-\sigma)}}] &= \int \frac{\Gamma(1+\alpha)}{\Gamma(1+\alpha/\sigma)} x^{\frac{-k\sigma}{(1-\sigma)}} x^{-\alpha} f_\sigma(x)dx = \frac{\Gamma(1+\alpha)}{\Gamma(1+\alpha/\sigma)}\mathbb{E}[S_\sigma^{-(\alpha+\frac{k\sigma}{(1-\sigma)})}] \\ &= \frac{\Gamma(1+\alpha)\Gamma(1+\alpha/\sigma+k/(1-\sigma)))}{\Gamma(1+\alpha/\sigma)\Gamma(1+k\sigma/(1-\sigma)+\alpha)}\end{aligned} \tag{3.51}$$

In particular for $k = 1$:

$$M_{\varepsilon,\sigma,\alpha} \approx (\varepsilon/\sigma)^{\frac{-\sigma}{(1-\sigma)}}\mathbb{E}[(T_{\sigma,\alpha})^{\frac{-\sigma}{(1-\sigma)}}] = (\varepsilon/\sigma)^{\frac{-\sigma}{(1-\sigma)}} \frac{\Gamma(1+\alpha)\Gamma(1+\alpha/\sigma+1/(1-\sigma)))}{\Gamma(1+\alpha/\sigma)\Gamma(1+\sigma/(1-\sigma)+\alpha)} \tag{3.52}$$

Thanks to the well-known property of the Gamma function:

$$\Gamma(x+1) = x\Gamma(x+1), \tag{3.53}$$

we simplified (3.51) using (3.53):

$$M^k_{\varepsilon,\sigma,\alpha} \approx (\varepsilon/\sigma)^{-k\sigma/(1-\sigma)} \frac{\Gamma(\alpha)\Gamma(\alpha/\sigma + k/(1-\sigma))}{\Gamma(\alpha/\sigma)\Gamma(\alpha + k\sigma/(1-\sigma))} \tag{3.54}$$

We then wanted to study the asymptotic behaviour for $M^k_{\varepsilon,\sigma,\alpha}$ when $\sigma \to 1$ and $\alpha \to \infty$ since we were interested to know the growth rate for the truncation number with the change of parameters.
Firstly, consider the case where $\sigma \to 1$

Using Stirling formula we obtained (see details in the Appendix A.2.1) the following:

$$M^k_{\varepsilon,\sigma,\alpha} \approx \varepsilon^{\frac{-k\sigma}{(1-\sigma)}} e^{-k} \left(\alpha + \frac{k\sigma}{(1-\sigma)}\right)^k c_k(\alpha,\sigma) \tag{3.55}$$

where $c_k(\alpha,\sigma) \to 1$ as $\sigma \to 1$.

Hence, we got

$$M^k_{\varepsilon,\sigma,\alpha} \approx \varepsilon^{\frac{-k\sigma}{(1-\sigma)}} \left(\frac{k}{1-\sigma}\right)^k \text{ with } \sigma \to 1. \tag{3.56}$$

We could see that for fixed value of ε, α we had an exponential growth of the moments of $\tau(\varepsilon)$ as σ approaches 1. Even if σ cannot be 1 by definition, we can intuitively think that in the limiting case $\sigma \to 1$, the probability of joining the existing cluster would to 0, and so each new realisation would generate a new cluster, and we thus could not truncate the infinite sum at any level. However, the case where $\sigma \to 1$ is out of our interest, since we are using the PY process because of its clustering properties.
Now we consider the case where $\alpha \to \infty$

Similarly to previous case using the Stirling formula and substituting in (3.54) (see details in Appendix Section A.2.2) in case where $\alpha \to \infty$ we got:

$$M^k_{\varepsilon,\alpha,\sigma} \approx \left(\alpha + \frac{k\sigma}{(1-\sigma)}\right)^k \approx \alpha^k \text{ with } \alpha \to \infty. \tag{3.57}$$

So, with two above results we obtained that we had an exponential growth when σ approaches 1 and polynomial growth with growth of α. In Appendix A.2.4 we computed the first moment for different $\varepsilon = 0.1$ on a grid for α, σ, we could observe the different behaviour along the two axes (Figure A.3) .
We decided to approximate the 95%-quantile by taking $N = \mathbb{E}[\tau(\varepsilon)] + 2\sqrt{Var[\tau(\varepsilon)]}$.

As we obtained an analytical expression with Gamma function for mean and moments, we could easily compute the mean and standard deviation, but the question was how well it would approximate the quantile?

To answer this question we could use concentration inequalities, which assess deviation ("concentration") of random variable around its mean. We had a distribution for which we knew all the moments analytically and we know also that they are finite. By calculating the asymptotic behaviour for the k^{th} moment, we could classify the distribution by its tail behaviour.
In our case we could easily see (see details in Appendix Section A.2.3) that:

$$\mathbb{E}[X^k] \approx k^k. \tag{3.58}$$

Hence, using Theorem 2.1 from Vladimirova and Arbel (2019), there exists a constant K_1, $K_1 > 0$ such that:

$$P\{X - \mathbb{E}[X] \geq t\} \leq \exp(-t/K_1) \text{ for all } t \geq 0, \tag{3.59}$$

where for us $t = 2\sqrt{Var[X]}$.
Based on this bound on the tail behaviour, we could use $N = \mathbb{E}[\tau(\varepsilon)] + 2\sqrt{Var[\tau(\varepsilon)]}$ as an approximation of the 95% quantile.
We have tested our approximation by calculating the 95% quantile by sampling from the distribution of $\tau(\varepsilon)$, using the algorithm proposed by (Arbel et al., 2019). We compared the results with our approximation $N_{est} = \mathbb{E}[\tau(\varepsilon)] + 2\sqrt{Var[\tau(\varepsilon)]}$. Figure 3.2 shows the difference between the two, for different values of ε,α and σ.
We could see that the N_{est} is larger for all the values of the parameters, meaning that N_est was conservative in estimation of the 95%-th quantile and thus it always guaranteed that:

$$P(\tau(\varepsilon)) > N) < 0.05 \tag{3.60}$$

Moreover, when σ< 0.5 and α<10 the difference between the approximation and quantile was small. This implied that we don't add any computational effort by doing this approximation of the quantile. The gap between the true quantile and its approximation grows with $\varepsilon \to 0$, and therefore we overestimate the quantile more when ε is smaller.

Since the truncation approximation error ε is the sum of the weights after the truncation number, by choosing $\varepsilon = 0.1$ we obtain a bound on the largest p_k that we

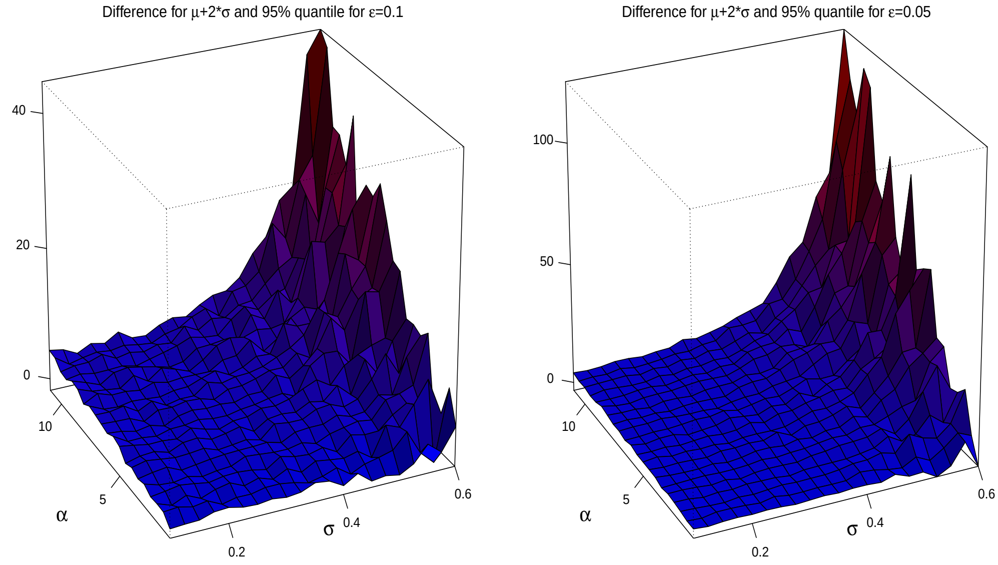

Figure 3.2 – *Difference between 95% quantile for* $\tau(\varepsilon)$ *and* $\mu+2SD$ *for values* $\alpha \in [0.1, 0.6], \theta \in [1, 10]$ *and* $\varepsilon = 0.1$ *(left) and and* $\varepsilon = 0.05$ *(right).*

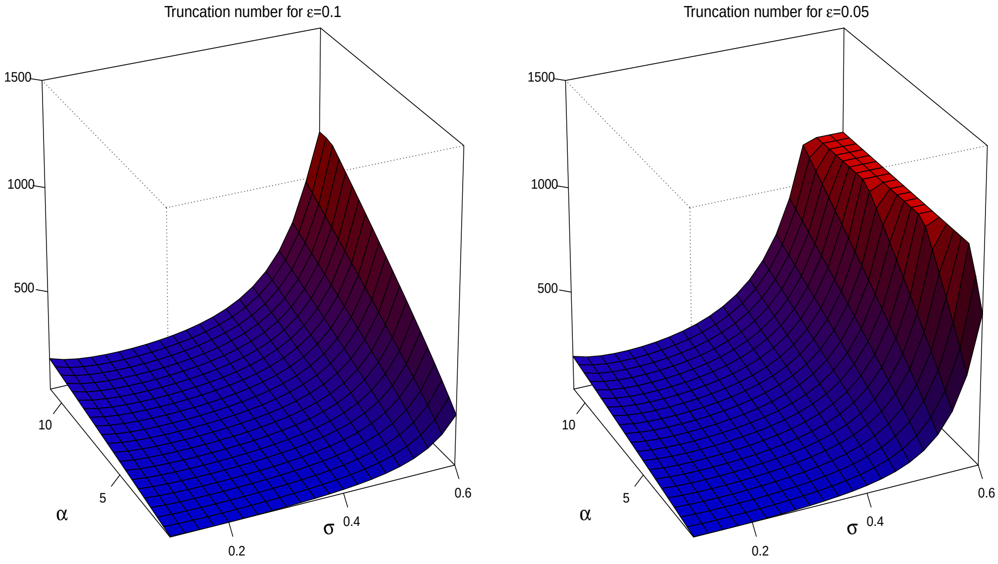

Figure 3.3 – *Truncation number* $N = \min\{\mu + 2SD, 1000\}$ *computed for different values of* ε*:* $\varepsilon = 0.1$ *(left) and* $\varepsilon = 0.05$ *(right).*

discarded, since $p_k < 0.1 \ \forall k \geq N$.

We justified both analytically and numerically our approximation of the 95% quantile using $N = \mu + 2SD$. To sample from a PY process we can thus fix N a priori, without adding an unbounded error, and then follow the blocked Gibbs sampler described above. This error bound is valid only a priori, and we will check the posterior of the truncation error in order to validate our approximation.

3.4.3 Fixing α and σ for Pitman–Yor process: properties of the prior number of clusters

In the previous sections we have described the Pitman–Yor process, listed some of its properties, and discussed how to sample from its posterior.
It is now time to elicit the values of the discount σ and concentration α parameters. In order to emphasize the motivations of our approach, we are going to investigate the distribution of the prior number of distinct values of the realisations from a PY process, and to compare it to the one of the realisations from a DP.
The distribution of the prior number of clusters of the realisations from a PY process has been studied by De Blasi et al. (2015), and is given by:

$$P(K_{n,\alpha,\sigma}) = \frac{V_{n,k}}{\sigma^k}\mathscr{C}(n,k;\sigma), \tag{3.61}$$

with:

$$\begin{aligned} \mathscr{C}(n,k;\sigma) &= \frac{1}{k!}\sum_{i=0}^{k}(-1)^i\binom{k}{i}(-i\sigma)_n \\ V_{n,k} &= \frac{\prod_{i=0}^{k-1}(\alpha+i\sigma)}{(\alpha+1)_{n-1}}. \end{aligned} \tag{3.62}$$

By taking $\sigma \to 0$ one founds the distribution of the prior number of clusters in a DP, given in (3.25).
In their work, the authors highlight the properties of this distribution (and its differences with respect to a DP) by a graphical display (Figure 3.4).

They fix $n = 50$ and consider different combinations of α and σ. For the Dirichlet process, as we discussed in Section 3.2, the total mass parameter α controls the location of the distribution of K_{50}: larger values of α lead to a right-shift of the distribution implying an (a priori) larger number of components (Figure 3.4). Moreover, the distribution is highly peaked.
From Figure 3.4 it is evident that the addition of σ allows to control the flatness, or the variability, of the distribution of K_{50} thus yielding a higher degree of flexibility for the model. Indeed, as σ increases the distribution becomes flatter, less informative. Elicitate in a correct way the parameters of a PY process is thus important, and needs to be done carefully. We use the same approach for the DP (Section 3.3), and fix α and σ in order to let the expected prior number of clusters match an ecological prior knowledge on the number of clusters.

The expected prior number of clusters from a PY process comes from the Pólya

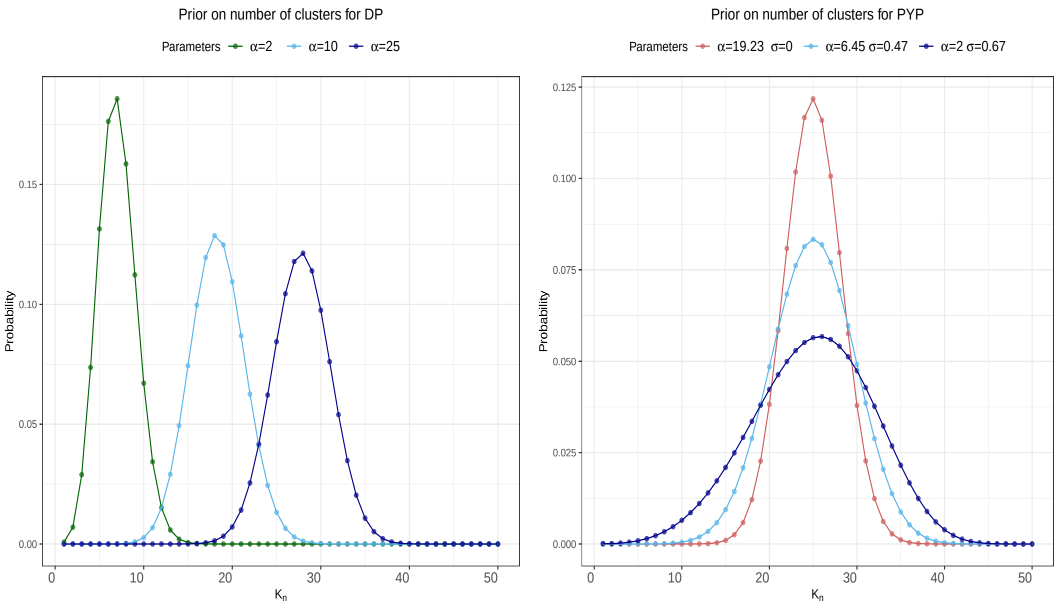

Figure 3.4 – *DP (left) and PYP (right) priors when N = 50, with different concentration and discount (here $\sigma = d$) parameter settings.*

urn representation described in (3.41), where the probability that a new realisation, given the previous n, enters in a new cluster is $\frac{\alpha+\sigma K_n}{\alpha+n}$. This last observation is useful to compute the expected number of clusters.
Indeed, from above it easily follows that:

$$\mathbb{E}[K_{n+1,\alpha,\sigma}] = \mathbb{E}[K_{n,\alpha,\sigma}] + \frac{\alpha + \mathbb{E}[K_{n,\alpha,\sigma}]\sigma}{\alpha + n}, \tag{3.63}$$

which implies:

$$\mathbb{E}[K_{n,\alpha,\sigma}] = \frac{\alpha}{\sigma}\left\{\prod_{j=1}^{n} \frac{\alpha+\sigma+j-1}{\alpha+j-1} - 1\right\}, \tag{3.64}$$

by induction on n. Stirling's approximation (see Pitman (2002) Chapter 3) then implies:

$$\mathbb{E}[K_{n,\alpha,\sigma}] \asymp \frac{\Gamma(\alpha+1)}{\sigma\Gamma(\alpha+\sigma)} n^{\sigma}, \tag{3.65}$$

It is interesting to notice that, with the same computations as in Section 3.4.2, one can also show the consistency of the Stirling approximation of the expected (prior) number of clusters. Indeed, Pitman (2002) gives the asymptotic: where $T_{\alpha,\theta}$ is a polynomially tilted random variable. Hence:

$$\mathbb{E}[K_{n,\alpha,\sigma}] = n^{\sigma}\mathbb{E}[(T_{\sigma,\alpha})^{-\sigma}] = n^{\sigma}\frac{\Gamma(\alpha+1)}{\Gamma(\sigma+\alpha)\sigma}, \tag{3.66}$$

which is consistent with (3.65), obtained via the Pólya urn representation.

Our aim is to find the pair:

$$(\bar{\alpha}, \bar{\sigma}) \ \text{ s.t. } \ \mathbb{E}[K_{n,\bar{\alpha},\bar{\sigma}}] = \bar{K}, \tag{3.67}$$

where $\bar{K}$ is the prior knowledge that we have on the expected number of groups. As in Section 3.3 we are not able to analytically find $(\bar{\alpha}, \bar{\sigma})$, but we can however approximate them numerically. Here we potentially have an infinite number of parameters satisfying (3.67), since we are working on a surface (see for example Figure A.5). Since $\mathbb{E}[K_{n,\bar{\alpha},\bar{\sigma}}]$ is monotone with respect to both α and σ, we can fix one of the two parameters on a grid (let's say σ fixed to $\sigma_1, \ldots, \sigma_k$) and then for all σ_i use the bisection method to find the value α_i that satifies:

$$(\alpha_i, \sigma_i) \ \text{ s.t. } \ \mathbb{E}[K_{n,\alpha_i,\sigma_i}] = \bar{K}, \tag{3.68}$$

obtaining thus a set of pairs $\{(\alpha_i, \sigma_i)\}_{i=1,\ldots,k}$ among the ones we will choose $(\bar{\alpha}, \bar{\sigma})$. Since the choice of $(\bar{\alpha}, \bar{\sigma})$ also drives the truncation level N and the vagueness of the distribution of the prior number of clusters, we will them among all the pairs $\{(\alpha_i, \sigma_i)\}_{i=1,\ldots,k}$ by finding a trade-off between a high vagueness of the prior distribution of the number of clusters (that increases with σ) and a low truncation level N (that increases with σ and α). See Figure A.7 in the Appendix to see graphically the relationship between $\alpha, \sigma, \mathbb{E}[K_{n,\alpha,\sigma}]$ and N, and a possible choice for $(\bar{\alpha}, \bar{\sigma})$. The truncation level N will thus fixed depending on $(\bar{\alpha}, \bar{\sigma})$ with the formula determined in Section 3.4.2. We implemented this model in R by modifying the Gibbs sampler of (3.24) implemented in the GJAM package (Clark, 2017). In this new Gibbs sampler we sampled the weights as described in (3.45).

3.4.4 Priors for the parameters α and σ

In the Bayesian framework when hyperparameters have a great influence on the results of a model, it is important to consider them as a random variable and to give them a prior knowledge that reflects some prior knowledge. We will thus repeat what we have done in Section 3.3 for the Dirichlet Process, by giving a prior distribution to the concentration parameter α and the discount parameter σ of the Pitman–Yor process, and by elicitating the corresponding hyperparameters in order to guarantee that the expected prior number of clusters matches an ecological prior knowledge.

This natural improvement comes however with two important drawbacks. First of all, as described in Section 3.4.2, the truncation level N depends on both α and σ. This means that we should evaluate the complex formula that give $N = \mu + 2SD$ at each iteration, as a function on α and σ. Moreover, the truncation level N grows when α and σ increase, and becomes dramatically big when σ is greater than 0.5.

Secondly, since the $\mathscr{GD}$ of $(p|\alpha,\sigma)$ depends now not only on α but also on σ, we can not find a conjugate prior. Indeed, by replacing $a_k = 1-\sigma$ and $b_k = \alpha + k\sigma$ in (3.15) we obtain:

$$\pi(p|\alpha,\sigma) \propto \left(\prod_{k=1}^{N-1} \frac{\Gamma(1+\alpha+(k-1)\sigma)}{\Gamma(1-\sigma)\Gamma(\alpha+k\sigma)}\right) p_1^{-\sigma} \cdots p_{N-1}^{-\sigma} p_N^{\sigma(N-1)+\alpha-1} \tag{3.69}$$

We are going to solve the first problem by giving a truncated prior distribution to σ, so that a posteriori it will never overcome a certain threshold with probability 1. We are not going to truncate the support of α, but we are going to elicitate his hyperparameter in order to fix most of the probability mass below a certain threshold.

We are then going to fix $\bar{N} = N_{\alpha_{max},\sigma_{max}}$, where α_{max} and σ_{max} will be the maximum values that the parameters could take. Since N is monotone with α and σ, we will guarantee that $\bar{N} = N_{\alpha,\sigma}$ for all α and σ that have a non zero probability a posteriori (and thus they almost never occur in the Gibbs sampler).

To simplify the problem, we decided to consider α and σ to be independent, so that:

$$\pi(\alpha,\sigma) = \pi(\alpha)\pi(\sigma) \tag{3.70}$$

As we did for the Dirichlet process, we are going to give a gamma prior distribution to the concentration parameter α:

$$\begin{gathered} \alpha \sim Ga(\nu_1,\nu_2) \\ \mathbb{E}[\alpha] = \frac{\nu_1}{\nu_2} = \bar{\alpha}, \end{gathered} \tag{3.71}$$

where $\bar{\alpha}$ is the one found at the end of Section 3.3.

We do not want α to be too big, since the truncation level N grows linearly with alpha. We will thus set:

$$Var[\alpha] = \frac{\nu_1}{\nu_2^2} = \frac{\bar{\alpha}}{\nu_2} = 10, \tag{3.72}$$

(we can eventually modify the number 10)

From which it follows:

$$\begin{cases} \mathbb{E}[\alpha] = \frac{\nu_1}{\nu_2} = \bar{\alpha} \\ Var[\alpha] = \frac{\nu_1}{\nu_2^2} = \frac{\bar{\alpha}}{\nu_2} = 10 \end{cases} \implies \begin{cases} \nu_1 = \frac{\bar{\alpha}^2}{10} \\ \nu_2 = \frac{\bar{\alpha}}{10} \end{cases} \tag{3.73}$$

Concerning the prior distribution for σ, we decided to use as a prior a mixture of a point mass at zero and a continous distribution as in Carmona et al. (2019), (what is called a spike and slab prior) the first component is spike concentrated at zero and second component is comparable flat slab. For our prior distribution we used uniform distribution as slab:

$$\pi(\sigma) = \rho\delta_0 + (1-\rho)\mathscr{U}_{(0,0.5)} \tag{3.74}$$

where $\mathscr{U}_{(0,0.5)}$ is the uniform distribution with support between 0 and 0.5.
Our prior is thus a mixture of two components. With probability ρ, $\sigma = 0$, making thus the Pitman–Yor process become a Dirichlet process, while with probability $1-\rho$, the discount parameter σ will be uniformly distributed around 0 and 0.5.
This choice leads to two important advantages. Firstly, by looking at the posterior distribution of σ we will be able to see how often the Pitman–Yor recovers the Dirichlet Process case, and this will be a strong justification of the goodness (or not) of our choice to extent the DP to a PY process.
Secondly, the support of σ is bounded between 0 and 0.5, and thus we avoid unfeasible values of the truncation level N.
Concerning the value of ρ, we want:

$$\mathbb{E}[\sigma] = \bar{\sigma}, \tag{3.75}$$

where $\bar{\sigma}$ is the one found at the end of Section 3.3.
Hence: to fix the value of ρ, we impose:

$$\mathbb{E}[\sigma] = \frac{(1-\rho)}{2} = \bar{\sigma} \tag{3.76}$$

and thus:

$$\rho = 1 - 2\bar{\sigma} \tag{3.77}$$

which requires that $\bar{\sigma} \leq \frac{1}{2}$. We now want to fix the truncation level $\bar{N} = N_{\alpha_{max},\sigma_{max}}$.

The maximum value that σ could attain a posteriori is $\sigma_{max} = 0.5$, due to his truncated support.
For α the problem is slightly more complicated, since his support is bounded. How-

ever, we are going to take α_{max} as the 95%-th quantile of its prior distribution. This is of course an approximation, since we could have $\alpha \geq \alpha_{max}$, but with a low probability. We are going to check a posteriori that this last hypotesis has worked out.

The full conditionals of α and σ are thus:

$$\pi(\alpha|\sigma,p) \propto \left(\prod_{k=1}^{N-1} \frac{\Gamma(\alpha+1+\sigma(k-1))}{\Gamma(\sigma k+\alpha)} \right) p_N^{\alpha} \alpha^{\nu_1 - 1} e^{-\nu_2 \alpha} \tag{3.78}$$

$$\pi(\sigma|\alpha,p) \propto \frac{1}{\Gamma(1-\sigma)^N} \left(\prod_{k=1}^{N-1} \frac{\Gamma(\alpha+1+\sigma(k-1))}{\Gamma(\sigma k+\alpha)} \right) p_1^{-\sigma} p_2^{-\sigma} \dots p_N^{\sigma(N-1)} [\rho \delta_0(\sigma) + 2(1-\rho) \mathbb{1}_{[0,0.5]}(\sigma)] \tag{3.79}$$

We implemented this model in R by modifying the Gibbs sampler of (3.24) implemented in the GJAM package (Clark, 2017). In this new Gibbs sampler we sampled the weights as described in (3.45). To sample from (3.78) we used a random walk Metropolis–Hasting step with a truncated normal proposal distribution, for the same reason as described in Section 3.3.1. Again, the variance of the proposal density was adjusted a posteriori by checking the convergence metrics.

Concerning the conditional distributions of σ we implemented a MH step with independent proposal distribution $\pi(\sigma) = \frac{1}{2}\delta_0 + \frac{1}{2}\mathscr{U}_{[0,0.5]}$ as suggested in Carmona et al. (2019), but using a Uniform distribution instead of a Beta to constrain the proposal between 0 and 0.5.

Chapter 4

Tests and applications

In this chapter we have applied our models (described in Sections **??** 3.3.2 3.4.3 3.4.4) as well as the original GJAM model (Section 3.1.3). We have first used a very simple simulation case, where the models were tested on their ability to retrieve the true number of clusters in the covariance matrix that generated the data. Being the task really easy, all models behaved well and we could not see any difference across the model for most simulations scenarios. However, the original GJAM struggled to retrieve the true number of clusters in the scenario with a lot of species and a low number of clusters. Even if this is a quite extreme case, it has encouraged us to test all the models on a more complex data. We have also applied our models on the real dataset of the plants in the Bauges Regional park (BRP), using Plant Functional Groups (PFG) to fix the a priori number of clusters. Our extensions (in Chapter 3) showed to outperform the original GJAM, even if the difference was not huge. The posterior of the unique number of clusters showed the importance of carefully choosing the prior, and suggests that the PFG partition reflects the way species interact among them.

4.1 Simulation for continuous data

4.1.1 Motivation for simulation study

The aim of this simulation study was to see how the prior choice of α for DP and α and σ for PY processes affected the clustering in our model. We know that the DP and PY mixture models are consistent in density estimation, but inconsistent in the identification of the true number of clusters (Miller and Harrison, 2014). But by comparing the number of estimated clusters with the true one across the different models, we were interested in seeing the difference in the models behaviour. We tested how the models we implemented in Chapter 3 influences the posterior number of clusters and if this leads to better results in model fitting, which we evaluated as the error in the estimation of the covariance matrix. While analyzing the posterior of the models, we also discussed the approximation error for the truncation of stick-breaking random measure by considering the posterior mean for the last weight $p_N = 1 - \sum_{k=1}^{N-1} p_k$.

4.1.2 Description of the data

The first type of simulation was made for continuous data. Following the framework of Taylor-Rodriguez et al. (2017) we constructed the simulation for the latent multivariate variable V, which is described in model definition for GJAM (3.21), to see if we could recover the specified multivariate dependence structure. For simplicity, we took the environmental response equal to zero, so for each plot i the response vector $y_i \sim N_S(0_S, \Sigma)$. We constructed the covariance matrix Σ of size $S \times S$ in the following way:

The structure of Σ in the model defined as $\Sigma = AA^T + \tilde{\sigma} I_S$, where A was matrix of size $S \times q$ and $q << S$. To have a clustered covariance structure with K_{true} number of groups, A had K_{true} repeated rows. Let $a_l,\ l = 1, ...S$ be the rows of A. Then:

$$a_l = v_k, \text{ where } v_k \sim N(0, \sigma_A I_q), \qquad k \in \{1, \dots K_{true}\},\ l \in \{1 \dots S\}. \tag{4.1}$$

Hence, we tested the model ability to recover the true number of clusters K_{true}. We investigated how well the models retrieved the good number of clusters using the posterior distribution of the unique numbers of rows of A. To measure the error between the true covariance matrix Σ obtained with true A matrix as described before and the estimated covariance matrix $\hat{\Sigma}$ (the posterior mean of Σ) we calculated the normalized Frobenius norm on the difference $\Sigma - \hat{\Sigma}$, where the Frobenius norm is defined as $||A||_F = \sqrt{\sum_{i=1}^{S} \sum_{j=1}^{S} |a_{ij}|^2}$, if A is a $S \times S$ matrix, and we then normalized it on the number of elements of A.

We fitted the models on the datasets simulated by using $\sigma_A = 3$, $\tilde{\sigma} = 0.1$, $q = 20$. Since we wanted to understand what is the effect of the number of plots n, the number of species S and the number of true clusters K_{true}, we varied these parameters. For each combination of r, S, K_{true}, n we generated five datasets and all the models were fitted individually on each dataset, in order to reduce the stochasticity related to data generation.

4.1.3 Models parametrization

We varied all the parameters of the simulations in order to have a complete understanding of the results and of the effect of the parameters on them.
We considered the number of species $S \in \{100, 300, 500, 1000\}$ and the true number of clusters $K_{true} \in \{4, 8, 10\}$. We also changed the sampling size n, with $n \in \{10, 50, 100, 500\}$ in order to study the cases whether the sampling size is small and thus the prior information had more power, which is one of the case where our model should work better.
GJAM also allows to choose the parameter r, which is the number of columns in the A matrix. We know that the true number of columns is $q = 20$, and we tested different values of r to assess for the impact of the latent factor approximation on the simulations. In general, the choice of r could be done by model selection, based on DIC or out-of-sample error (as we did in chapter 2), however for this simulation we were not interested in selecting the best r, but in testing the behaviour of the models for different values of $r \in \{5, 10, 15, 20\}$. Concerning the description of each model and their truncation level we refer to sections 3.3 and 3.4. The choice of the hyperparameters is described below:

- *GJAM* refers to the original model proposed by Taylor-Rodriguez et al. (2017). We follow authors indication by setting $N = min\{S, 150\}$ as truncation level and α is automatically set to $\alpha = S$. We modified the package only to have the possibility to have the chains for the stick-breaking weights p_k.

- *GJAM1* is the almost sure truncation version (see Section 3.3.2), where we set truncation $N = 150$, and we use the bisection method on (3.26) to find $\bar{\alpha}$ such that the expected prior number of clusters is K_{true}. We then use (3.34) to fix the hyperparameters of the prior distribution ν_1, ν_2.

- *GJAM2* is the weak limit version (see Section 3.3.1) where we set $N = min\{S, 150\}$ $\alpha = \bar{\alpha}$. For the hyperparameters of the prior of α we use the same method as in *GJAM1*.

- *GJAM3* is the PY version with fixed α,σ (see Section 3.4.3), we use the method described in the section to define $\bar{\alpha}$ and $\bar{\sigma}$. We set the truncation error as $\varepsilon = 0.05$ and truncation number using $N = N_{eps}$ as described in the section.

- *GJAM4* is the PY version with prior distribution for parameters α, σ (see Section 3.4.4). We take the same $\bar{\alpha}$,$\bar{\sigma}$ of GJAM3, and fix the hyperparameters

ν_1, ν_2, ρ in the way described in the section. We choose $\varepsilon = 0.1$ as we use more conservative upper bound for $N = N_{eps}$.

All models were run for 10,000 iterations and 3000 as burn-in, and thinned to keep 1 every 10 samples, for a final sample size of 1000. For *GJAM* the posterior was simulated using the GJAM package (Clark, 2017) in R. To sample from the posterior of the other models we used the functions that we implemented in R.

4.1.4 Posterior inference

The MCMC of all models typically proved to converge for all possible combinations of the simulation parameters. However, we had poor mixing regarding the number of unique values of the DP realisations (Figure 4.1), for all models and all combinations of the simulation parameters.

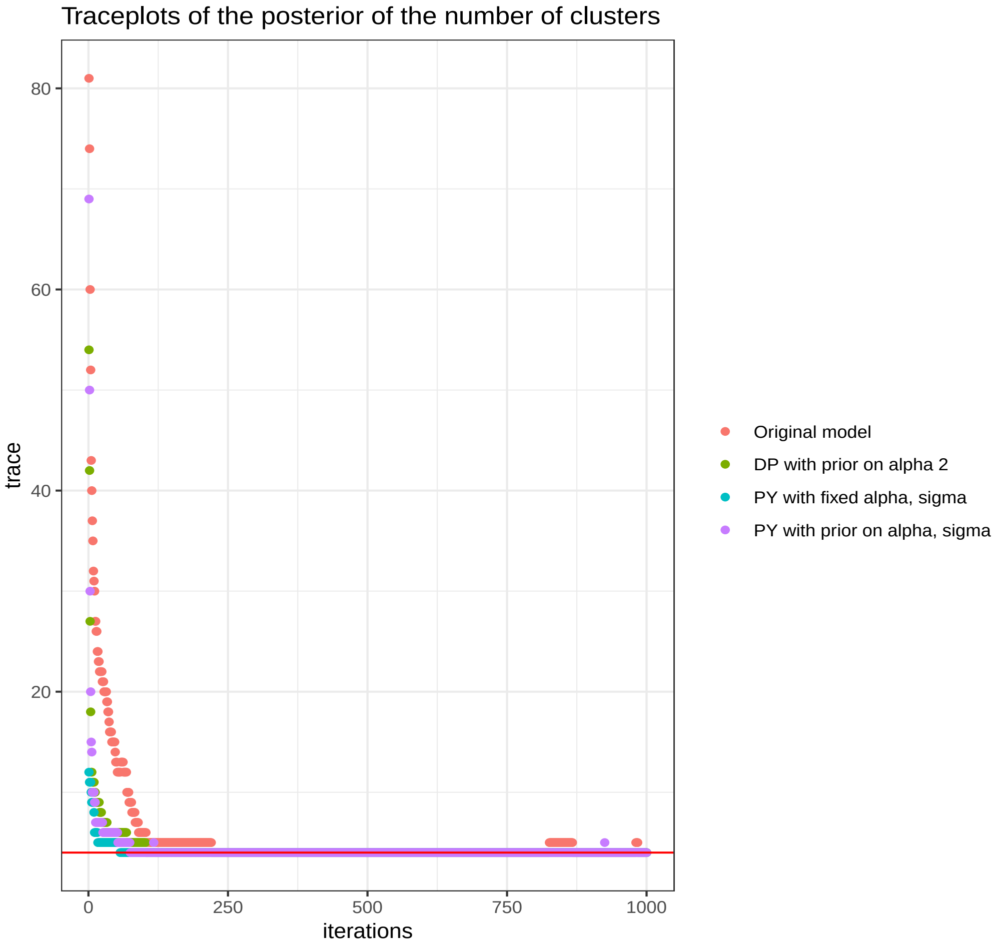

Figure 4.1 – *Traceplots of the number of clusters under the different models, for $S = 300$ and $r = 5$, $K_{true} = 10$, $n = 500$.*

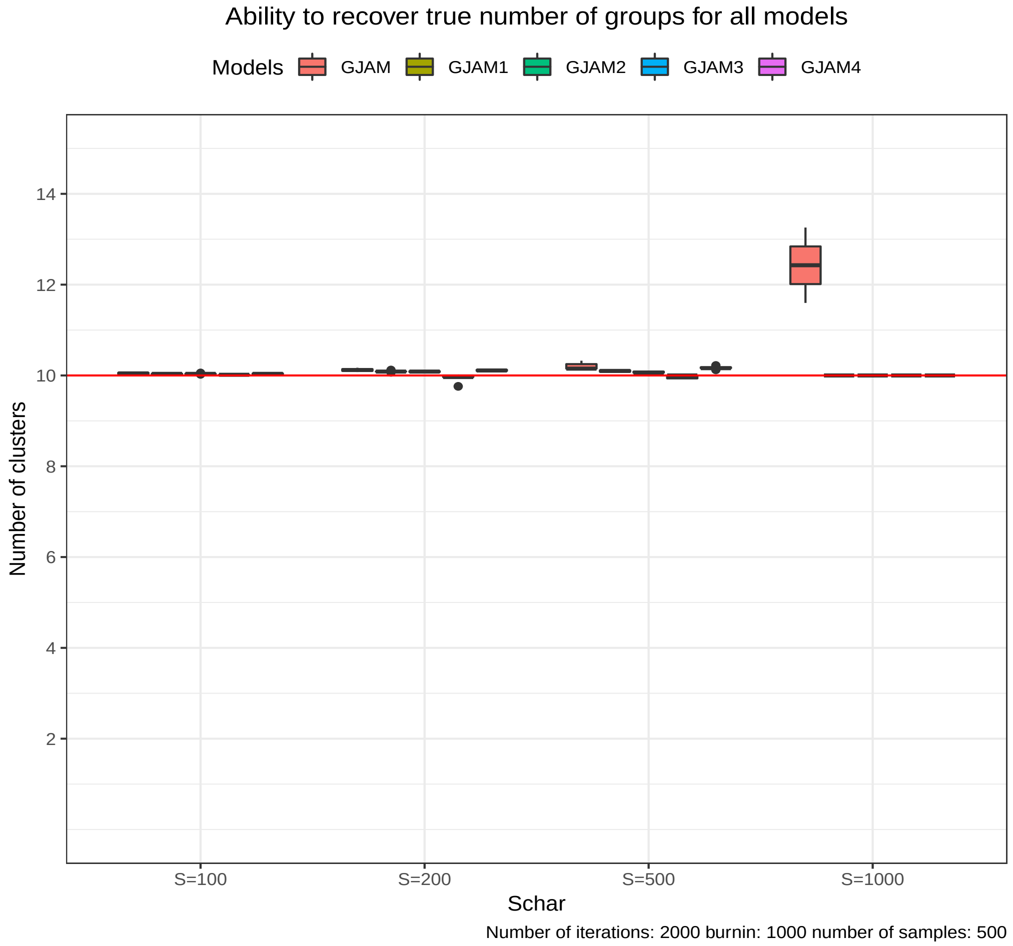

Figure 4.2 – *Boxplots of the posterior means of the unique values of the rows of A for all the models, a varying number of species S and $r = 5$, $K_{true} = 10$, $n = 500$.*

We first analysed the influence of the number of species S in the simulations. To represents easily the results, we fixed the true number of clusters $K_{true} = 10$ and the number of samples $n = 500$ and $r = 5$. We varied S by taking $S \in \{100, 300, 500, 1000\}$ as described above and for each S we fitted the models on all the 5 simulated datasets. For every tested S, except $S = 1000$ there were no differences across all models. They were all able to retrieve the correct number of clusters (Figure 4.2), and the error in retrieving the true values of Σ was the same across models and grew with the number of species (Figure 4.3). When the number of species was big, in particular in the case $S = 1000$, the original GJAM got worse both in the posterior number of clusters and in the posterior of Σ, while all the other models kept doing well, with no substantial difference across models. We had consistent results also with other values of K_{true}, n, r, and in particular the case of $K_{true} = 4$ is presented in Appendix A.2.9.

We analyzed the effect of the number of random factor r by seeing whether the error in retrieving the true matrix Σ depended on r, by varying r and S and leaving other parameters unchanged. Figure 4.4 shows that for these settings of the parameters all the models had the same ability in retrieving the true matrix: the variation of r did not lead to a substantial change of models error for any value of S.

We also checked for the goodness of the truncation approximations in the models by inspecting the posterior distribution of the last weight of the stick-breaking representation (see Section 3.4.2). For all models the posterior distribution of the error was globally low, with the approximation error of the DP model well below 0.1 (Figure 4.5 shows the posterior of the last weight p_N for some simulation scenarios). The PY approximations was worse than the DP approximation, as expected. However, the posterior mean was always below the maximum error ε that a priori we fixed to be 0.05 for *GJAM3* and 0.1 for *GJAM4*.

Table 4.1 gives the behaviour of N_ε when K_{true} and S varied and shows that by choosing the hyperparameters as described in section 3.4.3 the truncation level N_ε did not grew too much compared to the original case where $N = min\{S, 150\}$.

We were interested in checking for the posterior density of the hyperparameters when we added a hierachical level to the model.

For *GJAM1* and *GJAM2* the concentration parameter α had the same prior distribution, that was updated with a supplementary Metropolis–Hastings step for *GJAM2*, while *GJAM1* exploits conjugacy to easily update α in the Gibbs sampler (Sections 3.3.1 and 3.3.2). Since the prior distribution of α was such that the prior expected number of clusters was equal to K_{true}, we do not expect the posterior to move away from the prior. This not what happened in *GJAM1* where the posterior distribution of α went far away from its prior distribution setting to values between 25 and 45 instead of 2.5 (Figure 4.6). Instead in *GJAM2* the posterior of α stayed really close to its prior distribution as expected. (Figure 4.6).

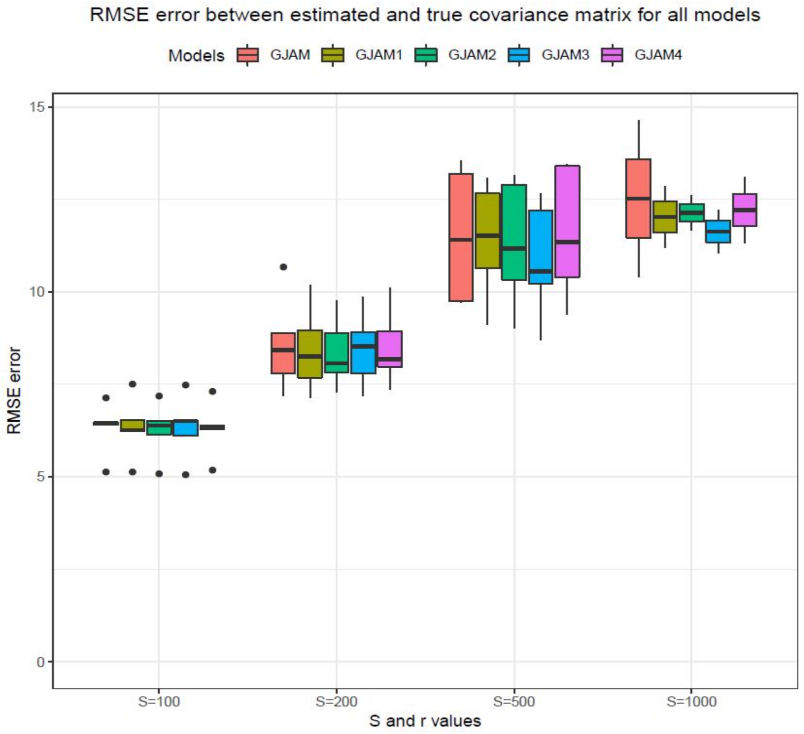

Figure 4.3 – *Error between the true matrix* Σ *and the inferred posterior mean* $\hat{\Sigma}$ *in the normalized Frobenius norm with different values of S for different models. Here* $n = 500$*,* $r = 5$ *and* $K_{true} = 10$ *and S varies.*

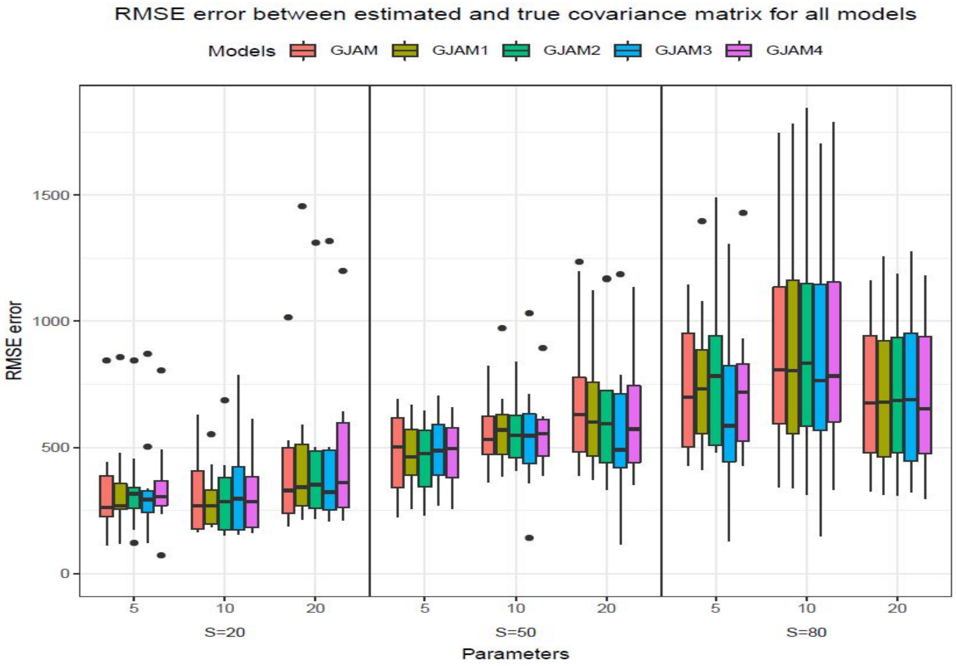

Figure 4.4 – *Error between the true matrix* Σ *and the inferred posterior mean* $\hat{\Sigma}$ *in the normalized Frobenius norm for different values of r and S (for small values of S), and* $n = 500, K_{true} = 10$

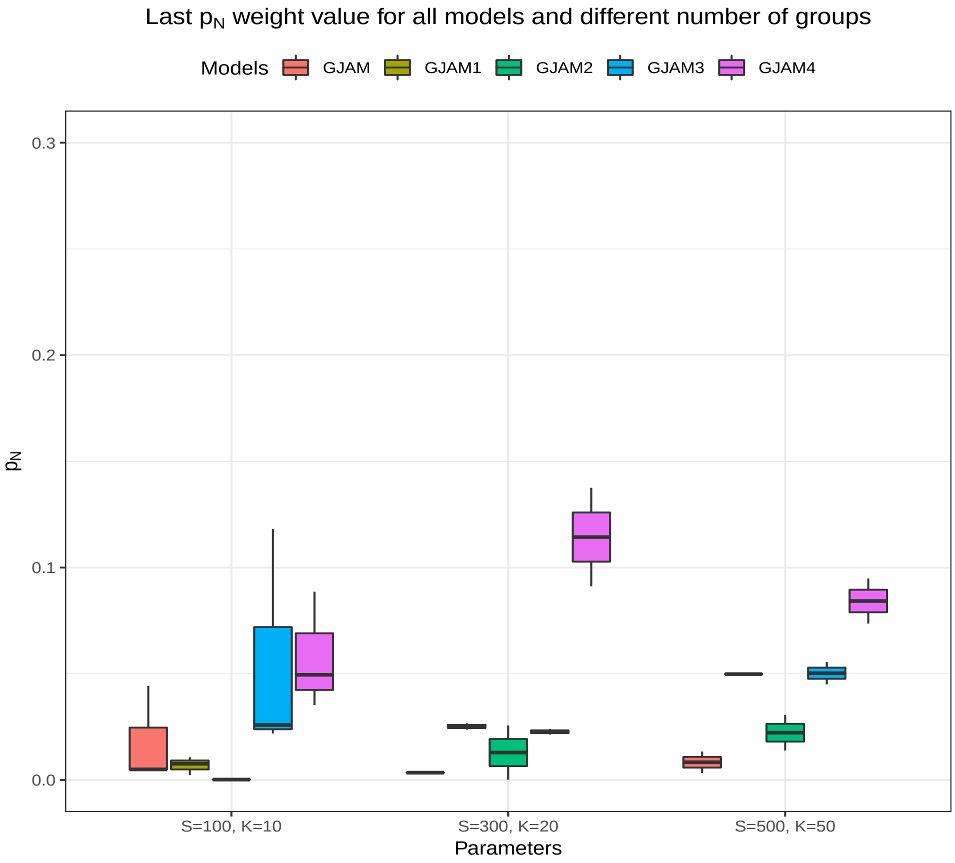

Figure 4.5 – *Posterior distribution of the last weight p_N for different values of K_{true} and S.*

	Model	S	K	N_{eps}
	GJAM 3	100	10	26
	GJAM 4	100	10	203
	GJAM 3	300	20	38
	GJAM 4	300	20	203
	GJAM 4	500	50	98
	GJAM 4	500	50	203

Table 4.1 – *Truncation number N_{eps} for models GJAM3 and GJAM4 for different values of S and K .*

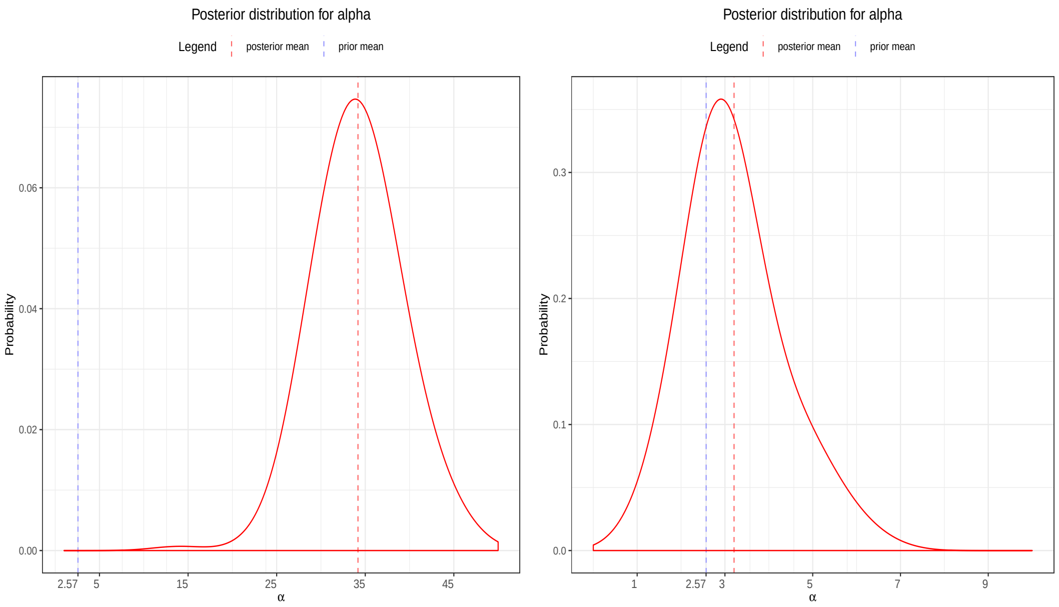

Figure 4.6 – *Posterior distribution for α in models GJAM1 (left) and GJAM2 (right) for case when $K_{true} = 10$, $n = 500$, $S = 100$, $r = 5$. The blue dashed line is the prior mean (equal for both distributions) while the red dashed line is the posterior mean.*

4.1.5 Discussion

The main goal of these simulations was to test the clustering properties of the original GJAM version and to compare it to those of the models that we implemented. We decided to simulate the data in a very simple way, by generating them from a multivariate normal with zero mean, and whose variance-covariance matrix is such that its Cholesky decomposition contains repeated rows. This is a very simple case because the data generating process has the same form of the model that we used to fit the data. Since the original GJAM model always sets $\alpha = S$, the prior expected number only depends on S due to (3.26) (n in the formula is the number of realisations of the DP, and is thus the number of species S). Moreover, due to the peaky shape of the prior distribution of the number of clusters in a DP, we expected that when the true number of clusters in our data generating process is far from the expected prior number of clusters given by (3.26), the posterior would be far from the true value.

The unexpected result was that the original GJAM model can retrieve the true number clusters, also when the prior was very far from it. For example, in Figure 4.2 we see that GJAM was always able to retrieve the true number of cluster $K_{true} = 10$ for $S = 100, 200, 500$, even if the expected prior number of clusters was, respectively, $\mathbb{E}[K_{S,\alpha}] = 69, 138, 346$. Moreover, also the values of the variance-covariance matrix where retrieved with the same precision of the other models.

This unexpected ability of the original model contradicts our prior expectation based on the results of De Blasi et al. (2015). However, we think that this can be due to the easy structure of the data generating process, that leads to a very strong likelihood that compensates the effect of the wrong prior. Since the number of data-points n is different from the number of realizations of the DP (that is the number of species S), we think that data can help the model in retrieving the good numbers of clusters. We also noticed the chain of the unique number of clusters goes really fast to a value and does not mix anymore (Figure 4.1). Therefore, we think that the differences between the models might arise on more complex datasets.

However, when the number of species reached very high values, for example $S = 1000$, the original GJAM encountered some difficulties in retrieving the true number of clusters and the true values of Σ, since in this case the expected prior number of clusters was very big (around 690) and far from K_{true}. Our models could instead easily retrieve the true number of clusters. Since the dimension reduction proposed in Taylor-Rodriguez et al. (2017) is particularly addressed to datasets were the number of species is really high (it is common to have $S > 1000$ in microbiome dasets, for example) our extensions can be considered as an improvement of GJAM and this result encourages us for testing the models on more complex datasets.

These simulations also allowed to test and compare the different models that we implemented. In Chapter 3 we discussed the a priori truncation error of the BNP priors, and we were interested in observing its behaviour a posteriori. The posterior distribution of this approximation error, given by the last weight p_N, was globally low, meaning that all the truncations discussed before were a fair approximation. In particular, for the models based on PY process the bound on the error that we put a priori is confirmed a posteriori, showing that our approximation method worked correctly. Not only the bound on the error was respected a posteriori, but the truncation level N_ε remained at feasible levels and we could not see any substantial difference concerning the computation times of the different models. Our approximation proved to be a good tool, providing a new possible algorithm to sample from the posterior of a PY process.

We were also interested in checking the behaviour of the posterior distribution of the hyperparameters for the models where we added a hierachical layer. Concerning *GJAM2*, based on the weak limit approximation where we sample α using a MH step, we observed that the posterior distribution of α does not go far from the prior distribution (whose mean is set to the value of α that gives a prior the true number of clusters K_{true}), meaning that our MH step was well implemented, and that this model worked as expected. However, this was not the case for *GJAM1*, that uses the finite DP truncation, where the posterior of α moved far away from the prior. For example in the case where $K_{true}, S = 100$, the correct value of α that gives an expected number of clusters equal to K_{true} was $\alpha \simeq 2.5$, but the posterior moved far away with a posterior mean around 35. Looking at the full conditional of α in *GJAM1*, given by (3.37), we see that the only parameter that change along the iterations is the truncation error p_N. In particular in order to have small values in the posterior of α, we need to have very small values of p_N. In our case if we wanted be able to move to very small values of α we need to have extremely low values of p_N, that are not easily reachable. This means that *GJAM1* struggles when it has to fit data that require a very small value of α. However, due to simplicity of the datasets, *GJAM1* was still able to retrieve the true number of clusters, since the posterior of α still set around more reasonable values then the S that is taken by GJAM.

4.2 Application to plants data in the Bauges Natural Regional Park

4.2.1 The dataset

We applied our models to a dataset containing the presences and absences of over 1.500 species over 17,351 plots (i.e. *plot* is the common term for *site* in ecology) in the Bauges Natural Regional Park (BNRP) available from the Alpine Botanical Conservatory, CBNA, Figure 4.7. BNRP is a typical subalpine massif of 90,000ha located in the northern French Alps, with an elevation ranging from 250m to 2,217m. More than 70% of the BNRP is covered by forests up to 1,500 m, and the remaining areas are covered by open pasture and cliffs. See Thuiller et al. (2018) for a thourough description of the dataset.

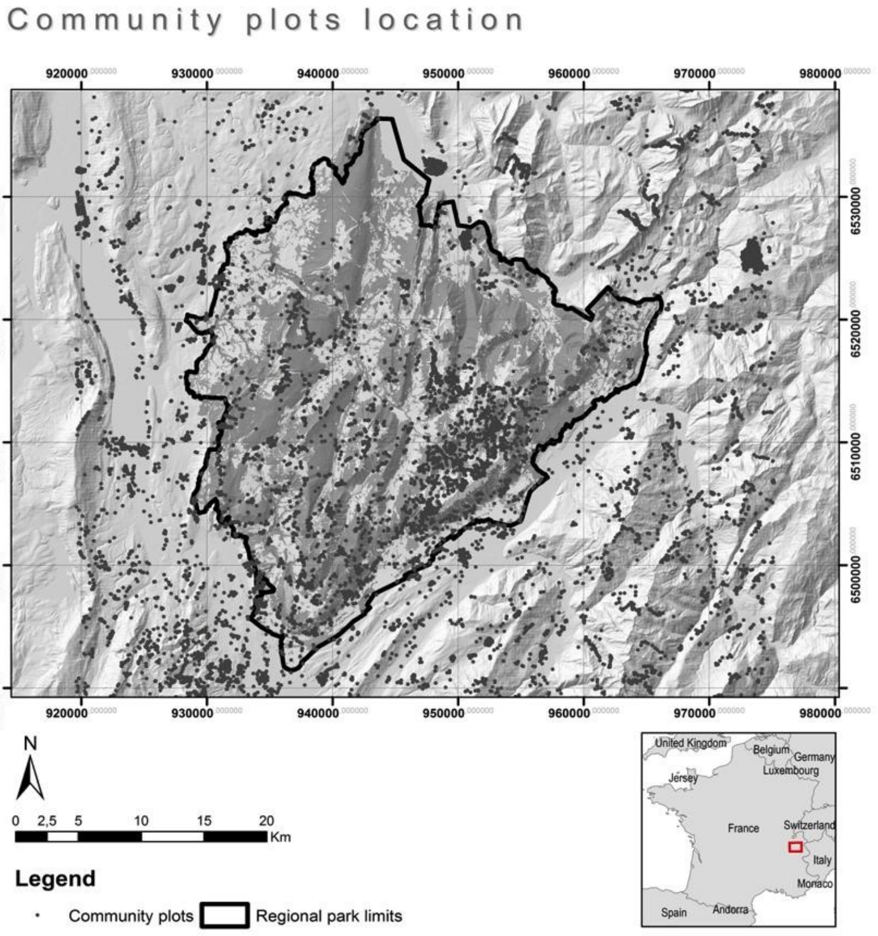

Figure 4.7 – *Spatial distribution of the vegetation plots used to select the dominant species list and to model the habitat suitability of the plant functional groups.*

We considered only significant species, by selecting species whose presence counts in vegetation plots were within the 95% quantile among all species, and species that are characteristic of each habitat of the park and thus occur in at least 25% of the vegetation plots within those habitats. The habitat classification and mapping were extracted from the CBNA data at a 1:5 000 resolution. We ended up having 136 dominant species (to select the dominant species we followed the same procedure of Thuiller et al., 2018).

We considered the same Plant Functional Groups (PFGs) that were built on the same dataset in Thuiller et al. (2018). PFGs are groups of species that share the same characteristics, in particular:

- their tolerance of abiotic conditions (e.g. temperature, precipitation...),
- their response to competition for light (whether they germinate and grow under specific light conditions),
- their "physical" characteristics (e.g. heights, type of leaf...)
- their demographic characteristics (e.g. age of maturity, longevity...) .

Using this variables, the authors run a hierarchical clustering that gave 16 Plant functional groups. See Thuiller et al. (2018) for a complete description of these PFGs and the way they were built. We considered such a number, $\bar{K} = 16$, to be the a priori number of groups of species that share the same behaviour with respect to other species, and used this information to fix the models hyperparameters as described in Chapter 3.

We extrapolated the climatic covariates for GJAM from the WorldClim daset (https://www.worldclim.org/), a set of global climate layers with a spatial resolution of about 1km^2. Such a dataset is composed of 19 bioclimatic variables (see https://www.worldclim.org/bioclim for a description), that we found out to be highly correlated in our region of interest. Because of this, we only considered the least correlated variables (mean annual precipitation and slope), and their interaction and quadratic terms as covariates for GJAM.

4.2.2 Models parametrization

Since the model with the Dirichlet Process prior and the almost sure truncation ($GJAM1$) seemed to have some troubles with the truncation error p_N and the posterior of the concentration parameter α, we decided not to consider this model. We thus fitted

$GJAM, GJAM2, GJAM3$ and $GJAM4$. We chose the hyperparameters of each model as described in Sections 3.3.1, 3.4.3 and 3.4.4 respectively, with $\mathbb{E}[K_n] = \bar{K} = 16$. In particular this gave:

- *GJAM2* $\nu_1 = 1.12$, $\nu_2 = 0.23$
- *GJAM3* $\sigma = 0.25, \alpha = 2.13$
- *GJAM4* $\rho = 0.2, \nu_1 = 0.08$, $\nu_2 = 0.09$

The choice of the number of latent factors r was done by fitting the models for different values of r, and choosing the value of r that gave the smallest DIC. We found $r = 5$ to be the optimum number of latent factors, coherently with the other applications of GJAM (Chapter 2) where the values of r were always quite small. We randomly split the plots in training (70% of the plots) and test (30% of the plots) to test compare the prediction abilities of the models.

All models were run for 10k iterations, 5k burnin, thinned to keep 1 every 10 samples, for a sample size of 1000.

4.2.3 Posterior inference and discussion

All the four models globally converged. In particular the mixing of the number of clusters was way better then in Section 4.1 (Figure 4.8). This is probably due to the simplicity of the previous simulated dataset, where the true unique number of values was too obvious and identifiable for the model. Real data come of course with a lot of noise and the structure of data is not so clear. However, the variance of the posterior of our models was much smaller then the original model and the posterior mean was close to the prior mean, meaning that the prior number of clusters is well specified (Figure 4.8). Instead the original model, as we expected, was not able to converge to the same value of the other models, due to the wrong specification of the concentration parameter α. Here with 136 species and thus $\alpha = 136$ the prior number of clusters for the original model was 94: the posterior moved away from the prior, but was not able to go further and reach the same value of the other models.

The fact that the number of PFGs is also the posterior mean of the unique number of clusters in the variance covariance matrix is very interesting from an ecological point of view. PFGs are defined as groups of species sharing the same characteristics, but the implication that these species also share the same behaviour with respect to other species is not straightforward. The fact that the posterior number of cluster is the same as the number of PFGs is a hint in this sense. We did not have the time to analyze the cluster assignment, but if the partition of the rows of the variance-covariance matrix Σ was the same of the PFGs, it would be a nice ecological discovery.

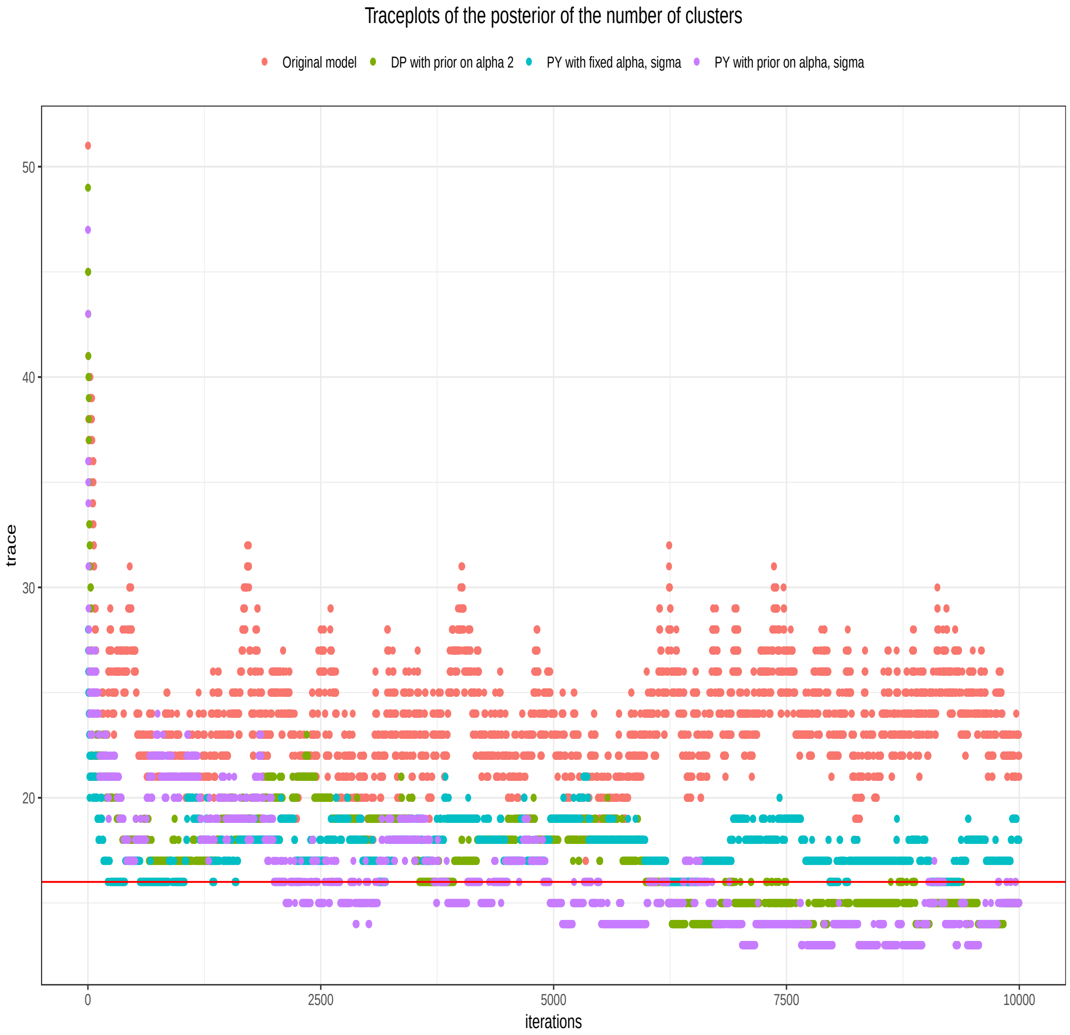

Figure 4.8 – *Traceplots of the posterior of the unique number of clusters for the four models. The red line shows the prior mean* $\bar{K} = 16$

We tested the ability of the models in prediction by computing for each species the AUC (Area Under the Curve of the Receiver Operating Characteristic) and the Tjur R^2. Tjur coefficient (Tjur, 2009) compares the estimated probabilities of presence for observed presences and observed absences. Using the notations from the (Taylor-Rodriguez et al., 2017) $\overline{TR} = (\hat{\pi}^1 - \hat{\pi}^0)$,where $\hat{\pi}^1$ and $\hat{\pi}^0$ are average probabilities of presence for observed ones and zeros respectively. This coefficient measures the models ability to discriminate between presences and absences. We then computed the average AUC and Tjur R^2 across all species. Both indices showed that our models are not worse than the original one (Table 4.2). These values are globally very small, but this is due to the complexity of modelling more then one hundred species over a very large area such as the BNRP. Also the distribution of the AUCs was really similar across the models, even if the *GJAM3* and *GJAM4* had less species with a very low AUC (Figure 4.9).

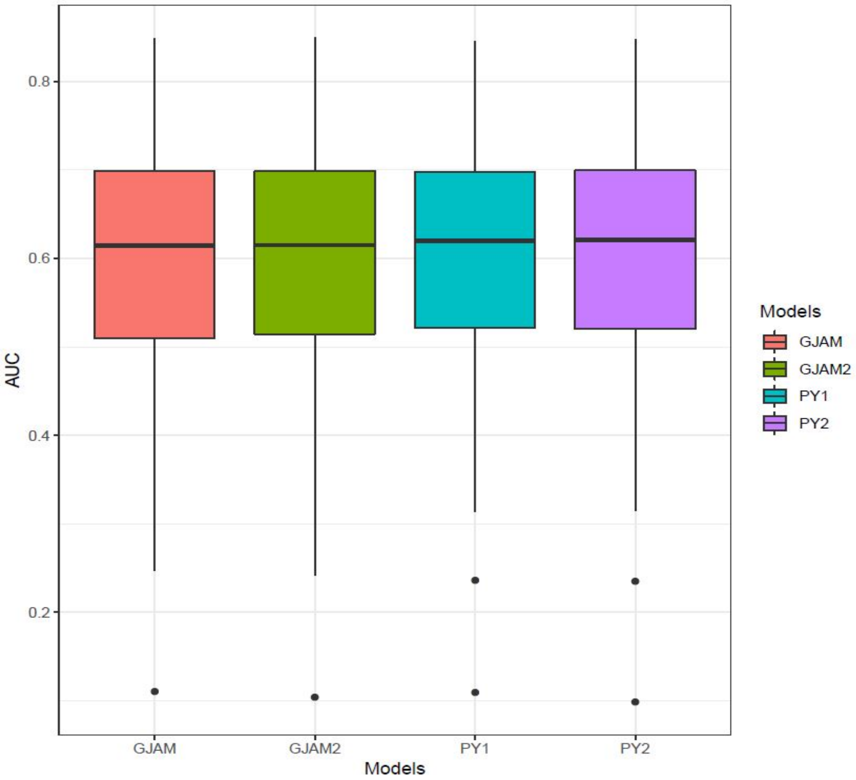

Figure 4.9 – *Accuracy of JSDM predictions represented as the species AUC. Each boxplot shows the values of the AUC of all species for a given model.*

Model	Mean AUC	Mean Tjur R^2
GJAM	0.595	0.035
GJAM 2	0.596	0.036
GJAM 3	0.601	0.036
GJAM 4	0.599	0.036

Table 4.2 – *Accuracy of JSDM predictions averaged across communities, represented by average AUC and average Tjur R^2. Both indices are averaged across species.*

The posterior of the concentration parameter in *GJAM2* did not move from the prior distribution, consistently with the fact the number of values (Figure 4.10). Choosing such a prior distribution for α was the right way to allow the number of clusters to be the good one. Instead the posterior of α for *GJAM2* shifted to the right of the prior, while the posterior of σ moved to the left of the prior. Nonetheless, if we replace the posterior mean of α and σ in the formula to compute the expected number of clusters in a PY process, we still get $\bar{K} = 16$. The model prefers to settle on a different combination of such parameters, but without changing the expected number of clusters, coherently with Figure 4.8. As for the simulation case, the spike of the prior distribution of σ did not affect the posterior distribution, and $\sigma = 0$ was never visited by the chain, confirming that the PY process is better than DP for this problem.

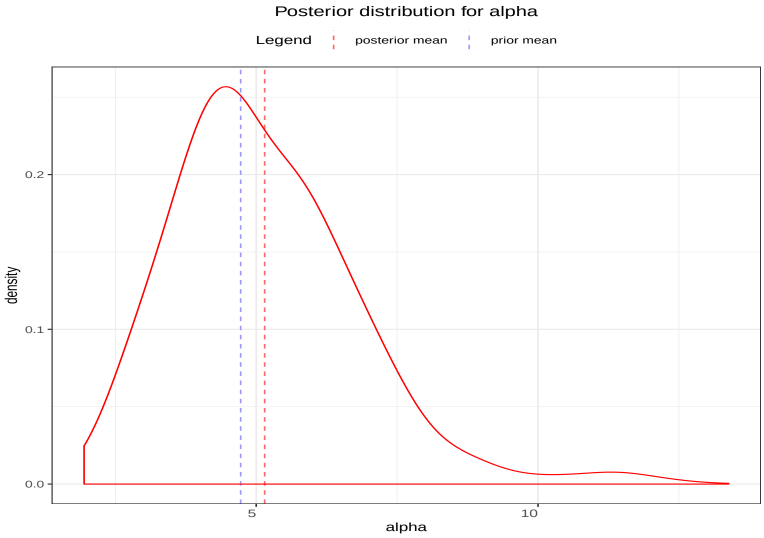

Figure 4.10 – *Posterior distribution of the concentration parameter α in GJAM2 The blue dashed line is the prior mean while the red dashed line is the posterior mean.*

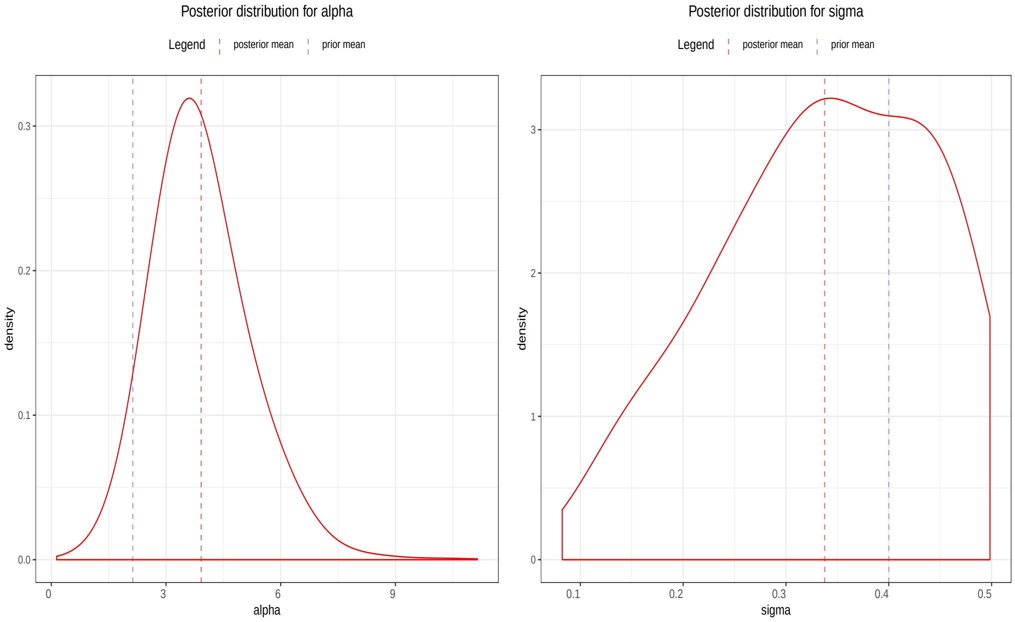

Figure 4.11 – *Posterior distribution of α (left) and σ (right) in GJAM4. The blue dashed line is the prior mean while the red dashed line is the posterior mean.*

Chapter 5

Conclusions

Understanding species distribution is one of the main goals of ecological research. Among the other possible methods, JSDMs allow to simultaneously estimate the species-environment relationship and the residual correlation between those species. In the recent literature there have been a few comparison studies that showed the similarities in the estimation of environmental coefficients between JSDM and SDM, as well as between different JSDMs.

Pollock et al. (2019) applied one family of JSDMs to a simple process-based simulation data, to understand whether JSDMs could detect species interactions, showing that JSDMs were better in retrieving the negative interactions. To better understand JSDMs, we have extended their work to two other state-of-the-art models (HMSC, GJAM). We have confirmed the previous findings and gave a better interpretation of the inference, showing that, in many cases, JSDMs can not disentangle the effect of the environment from the effect of species interactions, and thus the inference from these models should be interpreted with caution.

Due to estimation of residual covariance matrix, the number parameters grows as $O(S^2)$ and these models suffer from the curse of dimensionality. We have studied a particular dimension reduction approach used in GJAM, and we have proposed few extensions Bayesian nonparametric extensions, that allow the underlying clustering process to be more flexible and to take into account prior knowledge on the number of clusters. We have implemented four different BNP models and tested them on simulated data. On our very simple simulated dataset, the original GJAM could work very well, and was outperformed by our models only for scenarios with a high number of species, where the prior number of clusters for the original model was far away from the true one. We also tested our model on a real dataset (plants in the Bauges Regional park), using Plant Functional Groups (PFG) to fix the a priori number of clusters. Even if there was no improvement with respect to the original model in terms of prediction, the lack of robustness of the posterior of the number of cluster is what shows the importance of carefully choose the prior, and suggests that the PFG partition could reflect the way species interact among them.

This is an ongoing work that I will continue during my Phd. Concerning the first part of the work, I would like to test how JSDMs behave on the simulated datasets with abundances, without clumping the simulated communities to presence-absences that can lead to a loss of information. Regarding the second part, I would like to analyze more carefully the real data application, that I did not have the time to deepen. I would also like to check the posterior of the clusters assignments, to check if the clusters of the variance covariance matrix match the PFGs, which would be an interesting ecological statement. I would also like to introduce the polynomial PY instead of the truncated PY that we implemented.

5.1 Acknowledgments

I have worked on my Master thesis together with Daria Bistrova, a master student in Mathematics in Grenoble. We worked under the supervision of Julyan Arbel, from Inria Grenoble, and Wilfried Thuiller, from Laboratoire d'Ecologie Alpine (LECA) in Grenoble. Professor Alessandra Guglielmi was my supervisor in Politecnico di Milano.

Appendix A

Appendix

The supplementary material.

- A.1 section corresponds to part 1: comparisons of JSDM models.
- A.2 section corresponds to part 2: GJAM using Bayesian nonparametric priors.

A.1 Comparison of JSDM models

A.1.1 Partial correlation

In the chapter 3 we used the plots illustrating the similarity between the correlation matrix and matrix of true interactions, used as input for the simulation. In this section we provide similar plot, where the similarity was considered between the partial correlation matrix and true interaction matrix.

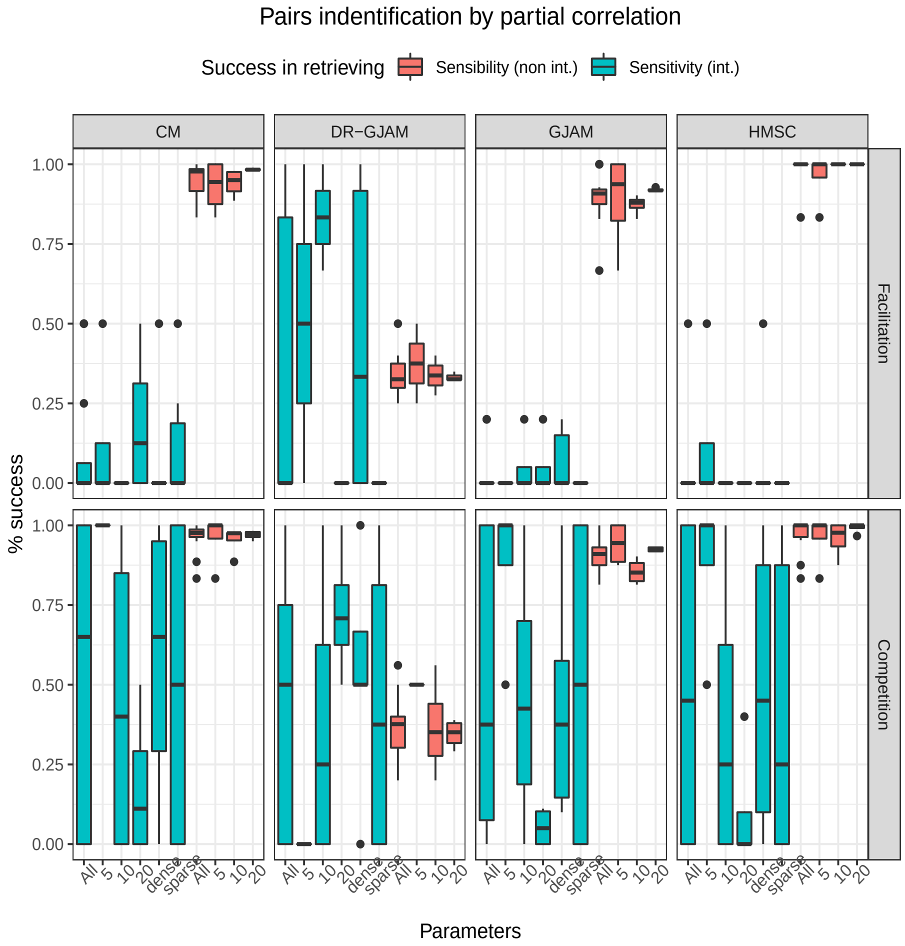

Figure A.1 – *Success rate calculated on the partial correlation matrix of JSDMs for interacting (competitors or facilitators, in green) and non-interacting species (in red) in communities simulated with all possible interactions scenarios for the four different models. Bars represent the following groups: all species pairs, species pools (5, 10, or 20 species) and the density of interactions.*

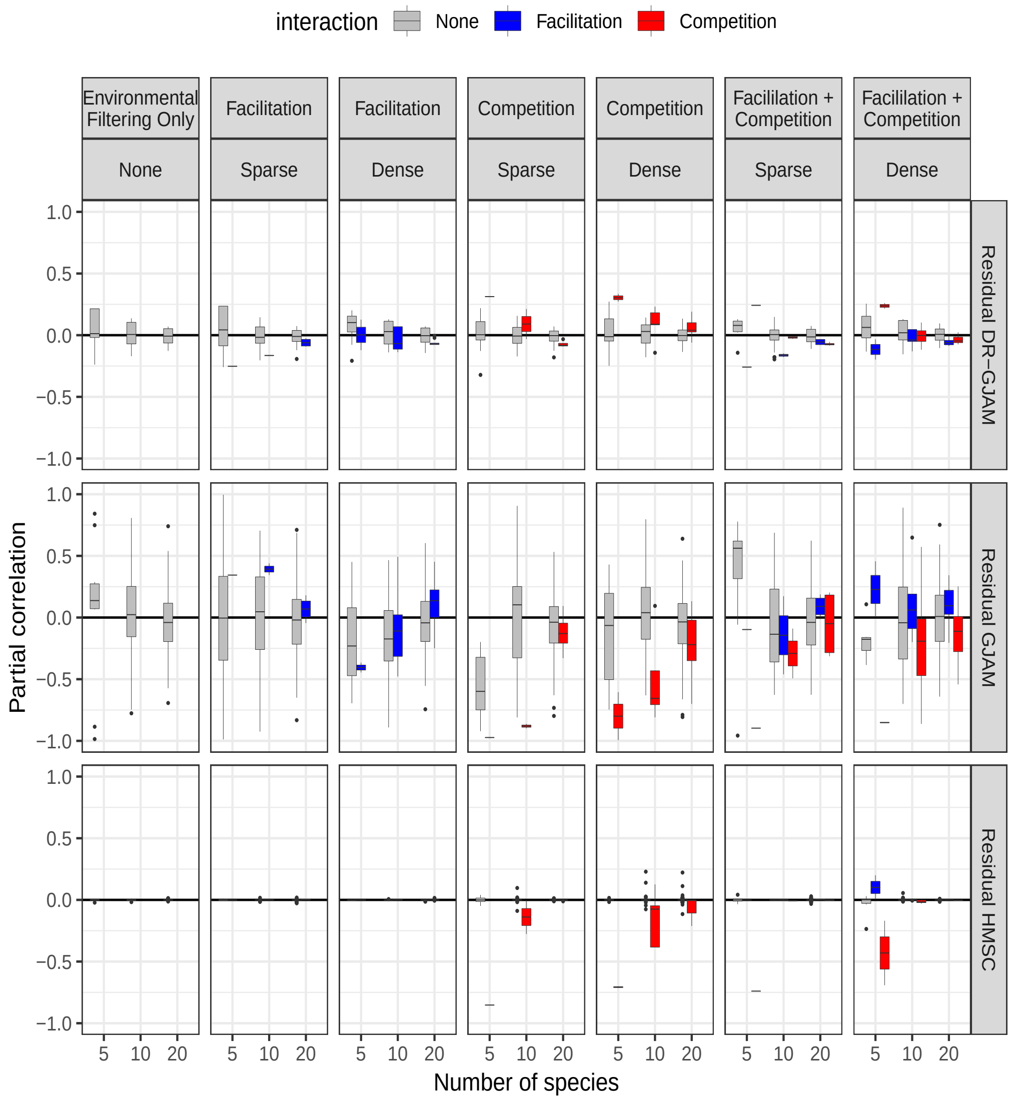

Figure A.2 – *Barplots representing the values of the probabilistic co-occurrence index (top row) and of the partial correlation matrix for each model (the other rows). For each scenario, species pairs are grouped into bars representing those that do not interact (grey), that compete (red) or that facilitate (blue).*

A.2 GJAM using Bayesian nonparametric priors

A.2.1 Asymptotic behaviour of the k^{th} moment for $\sigma \to 1$

Firstly, consider the case where $\sigma \to 1$

Using the following expression for the Gamma function we obtain:

$$\Gamma(x) = \sqrt{\frac{2\pi}{x}} \left(\frac{x}{e}\right)^x \left(1 + \frac{1}{12x} + o\left(\frac{1}{x}\right)\right). \tag{A.1}$$

Using (A.1) for (3.54):

$$\frac{\Gamma((\theta/\sigma + k/(1-\sigma)))}{\Gamma((\theta + k\sigma/(1-\sigma))} = \frac{\sqrt{2\pi}\sqrt{\sigma(1-\sigma)}(\theta(1-\sigma)+k\sigma)^{\theta/\sigma+k/(1-\sigma)}\sqrt{\theta(1-\sigma)+k\sigma}}{\sqrt{\theta(1-\sigma)+k\sigma}(\sigma(1-\sigma)e)^{\theta/\sigma+k/(1-\sigma)}\sqrt{2\pi}\sqrt{1-\sigma}} \times$$
$$\times \frac{(e(1-\sigma))^{\theta+k\sigma/(1-\sigma)}12(\theta(1-\sigma)+k\sigma)+\sigma(1-\sigma))+o((1-\sigma))}{(\theta(1-\sigma)+k\sigma)^{\theta+k\sigma/(1-\sigma)}(12(\theta(1-\sigma)+k\sigma)+(1-\sigma)+o(1-\sigma))}$$

Simplifying this we get

$$\frac{\Gamma((\alpha/\sigma + k/(1-\sigma)))}{\Gamma((\alpha + k\sigma/(1-\sigma))} = \sigma^{1/2-\alpha/\sigma-k/(1-\sigma)} \cdot e^{\alpha-\alpha/\sigma-k} \cdot \left(\alpha + \frac{k\sigma}{(1-\sigma)}\right)^{\alpha/\sigma-\alpha+k} \cdot (1+o(1-\sigma))$$

Substituting this in (3.54):

$$M^k_{\varepsilon,\sigma,\alpha} = \varepsilon^{\frac{-k\sigma}{(1-\sigma)}} \sigma^{1/2-\alpha/\sigma-k} e^{\alpha-\alpha/\sigma-k} \frac{\Gamma(\alpha)}{\Gamma(\alpha/\sigma)} \left(\alpha + \frac{k\sigma}{(1-\sigma)}\right)^{\alpha/\sigma-\alpha+k} \cdot (1+o(1-\sigma)) =$$
$$= \varepsilon^{\frac{-k\sigma}{(1-\sigma)}} e^{-k} \left(\alpha + \frac{k\sigma}{(1-\sigma)}\right)^k \cdot c_k$$

where $c_k = \sigma^{1/2-\alpha/\sigma-k} \cdot \frac{\Gamma(\alpha)}{\Gamma(\alpha/\sigma)} \left(\alpha + \frac{k\sigma}{(1-\sigma)}\right)^{\alpha/\sigma-\alpha} \cdot (1+o(1-\sigma))$ and $c_k \to 1$ as $\sigma \to 1$.

In particular:

$$\lim_{\sigma\to 1}\left(\alpha+\frac{k\sigma}{(1-\sigma)}\right)^{\alpha/\sigma-\alpha} = \lim_{\sigma\to 1}\exp\left((\alpha/\sigma-\alpha)\ln\left(\alpha+\frac{k\sigma}{(1-\sigma)}\right)\right) =$$

$$= \lim_{\sigma\to 1}\exp\left(\frac{\ln\left(\alpha+\frac{k\sigma}{(1-\sigma)}\right)}{1/(\alpha/\sigma-\alpha)}\right) = \exp\lim_{\sigma\to 1}\left(\frac{\ln\left(\alpha+\frac{k\sigma}{(1-\sigma)}\right)}{1/(\alpha/\sigma-\alpha)}\right) = \exp(0) = 1$$

Hence,we get

$$M^k_{\varepsilon,\sigma,\alpha} \approx \varepsilon^{\frac{-k\sigma}{(1-\sigma)}}\left(\alpha+\frac{k\sigma}{(1-\sigma)}\right)^k \approx e^{\frac{-k\ln\varepsilon\sigma}{(1-\sigma)}}\left(\frac{k}{1-\sigma}\right)^k \text{ with } \sigma\to 1. \tag{A.2}$$

where for $0\le\varepsilon\le 1$, $-\ln\varepsilon>0$

A.2.2 Asymptotic behaviour of the k^{th} moment for $\alpha\to\infty$

Now we consider the case where $\alpha\to\infty$.
Similarly to previous case substituting (A.1) in (3.54) in case where $\alpha\to\infty$:

$$\frac{\Gamma(\alpha)}{\Gamma(\alpha/\sigma)} = \alpha^{\alpha-\alpha/\sigma}e^{\alpha/\sigma-\alpha}\sigma^{\alpha/\sigma-1/2}\left(1+\frac{1-\sigma}{12\alpha}+o\left(\frac{1}{\alpha}\right)\right)$$

And

$$\frac{\Gamma((\alpha/\sigma+k/(1-\sigma)))}{\Gamma((\alpha+k\sigma/(1-\sigma))} = \sigma^{1/2-\alpha/\sigma-\frac{k}{(1-\sigma}}\left(\alpha+\frac{k\sigma}{(1-\sigma)}\right)^{\alpha/\sigma+k-\alpha} e^{\alpha-\alpha/\sigma-k}\times$$
$$\times\left(1-\frac{(\sigma-1)^2}{12(\alpha(1-\sigma)+k\sigma)}+o\left(\frac{1}{\alpha}\right)\right)$$

Using both above equations and substituting in (3.54)

$$M^k_{\varepsilon,\sigma,\alpha}(\varepsilon/\sigma)^{-k\sigma/(1-\sigma)}e^{-k}\sigma^{-k/(1-\sigma)}\left(\alpha+\frac{k\sigma}{(1-\sigma)}\right)^{\alpha/\sigma-\alpha+k}\alpha^{\alpha-\alpha/\sigma}\left(1+o\left(\frac{1}{\alpha}\right)\right) =$$
$$= \varepsilon^{-k\sigma/(1-\sigma)}\sigma^{-k}\left(\alpha+\frac{k\sigma}{(1-\sigma)}\right)^k c_k$$

where $c_k = \left(1+\frac{k\sigma}{\alpha(1-\sigma)}\right)^{\alpha/\sigma-\alpha}e^{-k}\left(1+o\left(\frac{1}{\alpha}\right)\right)$ and $c_k\to 1$ as $\alpha\to\infty$.
Hence,we get

$$M^k_{\varepsilon,\sigma,\alpha} \approx \left(\alpha+\frac{k\sigma}{(1-\sigma)}\right)^k \approx \alpha^k \text{ with } \alpha\to\infty. \tag{A.3}$$

A.2.3 Asymptotic behaviour for the k^{th} moment as $k \to \infty$

Asymptotic behaviour for the k^{th} moment as $k \to \infty$

Using the equation (3.54) and applying (A.1) formula for the part that depends on k we have:

$$\frac{\Gamma(\alpha/\sigma + k/(1-\sigma))}{\Gamma(\alpha + k\sigma/(1-\sigma))} = \sigma^{\frac{1}{2}+\alpha/\sigma+k/(1-\sigma)} \left(\alpha e^{-1} + \frac{k\sigma}{e(1-\sigma)}\right)^{\alpha/\sigma+k-\alpha} \left(1+o\left(\frac{1}{k}\right)\right)$$

Hence, for $M^k_{\varepsilon,\sigma,\alpha}$ we have

$$M^k_{\varepsilon,\sigma,\alpha} \approx K_1 \sigma^k \varepsilon^{k\sigma/(1-\sigma)} e^{-k} \left(\alpha e^{-1} + \frac{k\sigma}{e(1-\sigma)}\right)^{\alpha/\sigma+k-\alpha} \left(1+o\left(\frac{1}{k}\right)\right) =$$
$$= K_1 (c_1)^k (c_2+k)^{\alpha/\sigma+k-\alpha} \left(1+o\left(\frac{1}{k}\right)\right) \approx K c^k k^{k+\theta}$$

where K,c,θ are constants that doesn't depend on k.

We also know that $x^{1/x} \to 1$ as $x \to \infty$, because $x^{1/x} = e^{ln(x)/x}$ and we know that $ln(x)/x \to 0$ as $x \to \infty$, hence:

$$(E[|X|^k])^{\frac{1}{k}} = (M^k_{\varepsilon,\sigma,\alpha})^{\frac{1}{k}} \approx k$$

A.2.4 First moment of truncation number

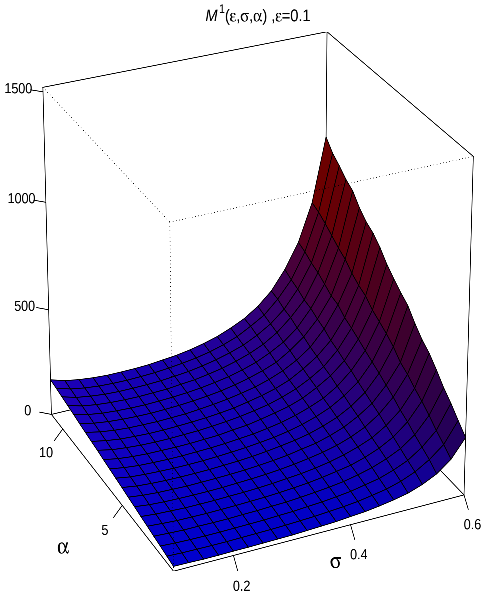

Figure A.3 – *First moment for $\tau(\varepsilon)$ for values $\alpha \in$ [0.1, 0.6],$\alpha \in$ [1,10] and $\varepsilon = 0.1$*

A.2.5 Skewness and kurtosis

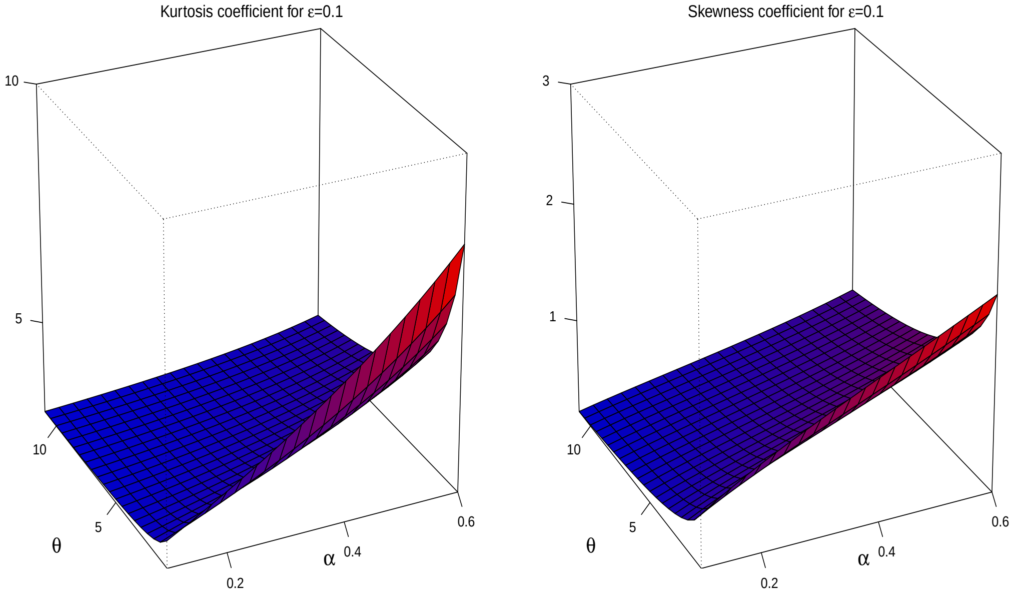

Figure A.4 – *Skewness parameter for values $\alpha \in$ [0.1, 0.6],$\theta \in$ [1,10] and $\varepsilon = 0.1$ (rigth) and kurtosis for the same values (left).*

A.2.6 Expected number of clusters for DP and PY

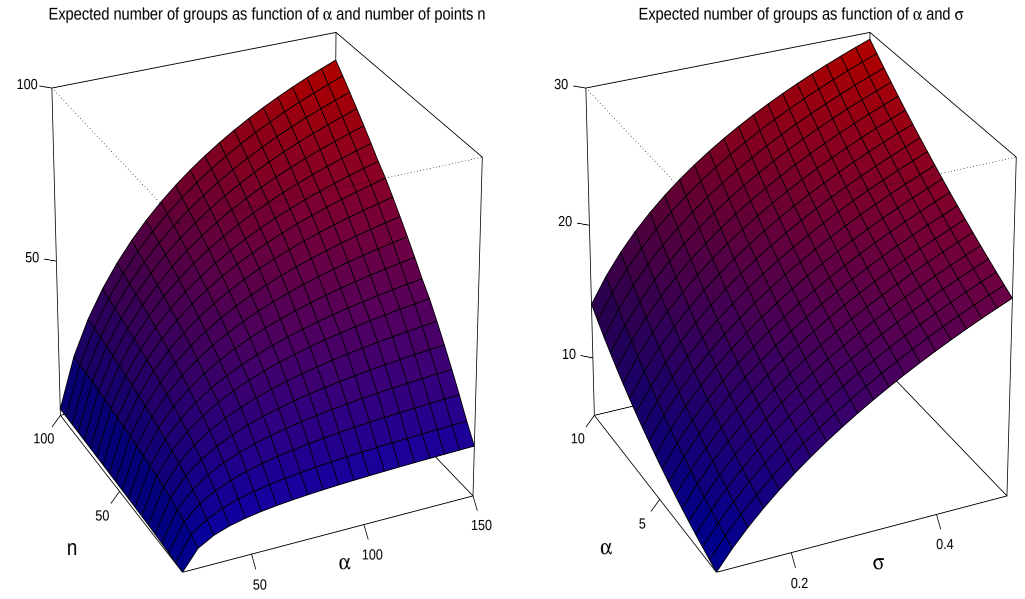

Figure A.5 – *Expected number of clusters for DP process and values of $\alpha \in$ [1, 100], and different number of points $n \in$ [20,150] (left) and Expected number of clusters for PY process for values of $\alpha \in$ [1, 10] and $\sigma \in [0.1, 0.5]$ (right).*

A.2.7 Expected number of clusters for DP

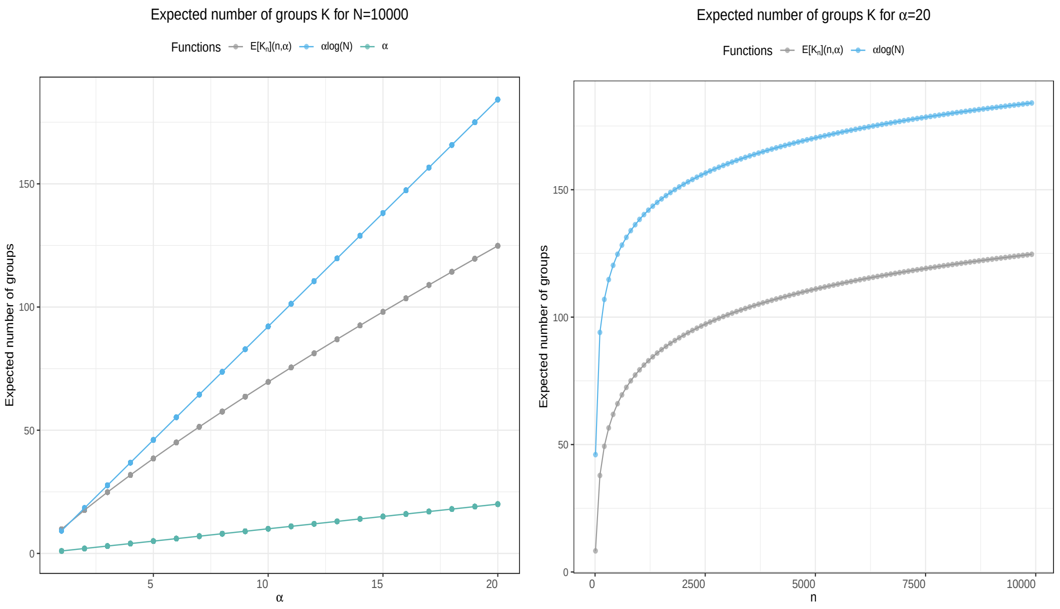

Figure A.6 – *Expected number of clusters for DP process as function of α for fixed $N = 10^4$ (left) and function of N for $\alpha = 20$*

A.2.8 Relations between α and σ for particular expectation on prior number of clusters

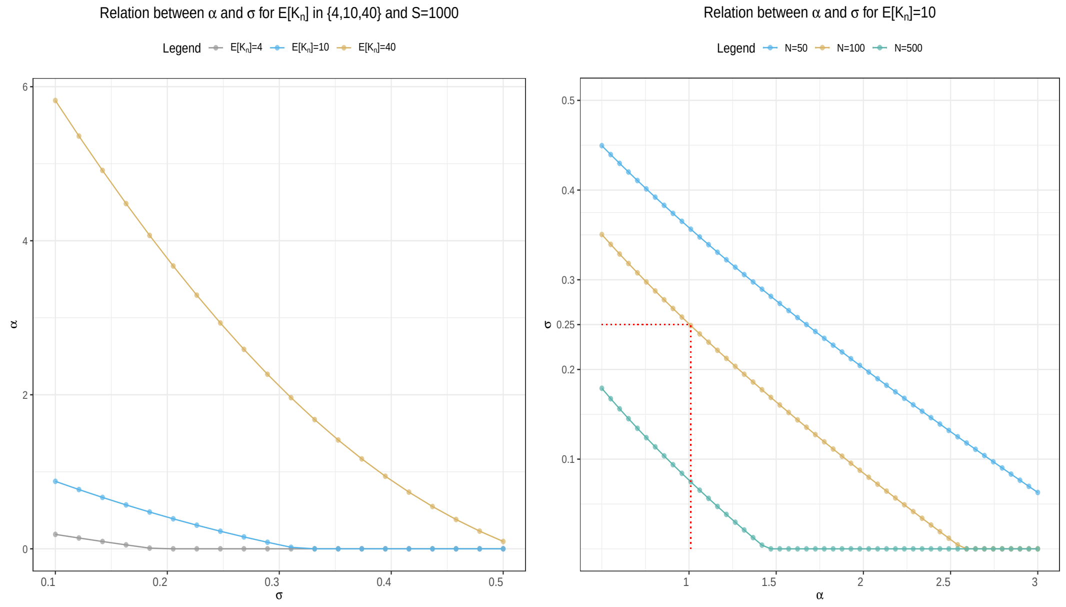

Figure A.7 – *A priori expected number of unique values from a sample of size $n = 10^4$ (left) and as a function of n for $\alpha = 20$ (right). Once we have the sample size, we can then decide a suitable combination of α and σ that guarantees the needed $\mathbb{E}[K_{n,\alpha,\sigma}]$.*

A.2.9 Posterior number of groups for $K = 4$

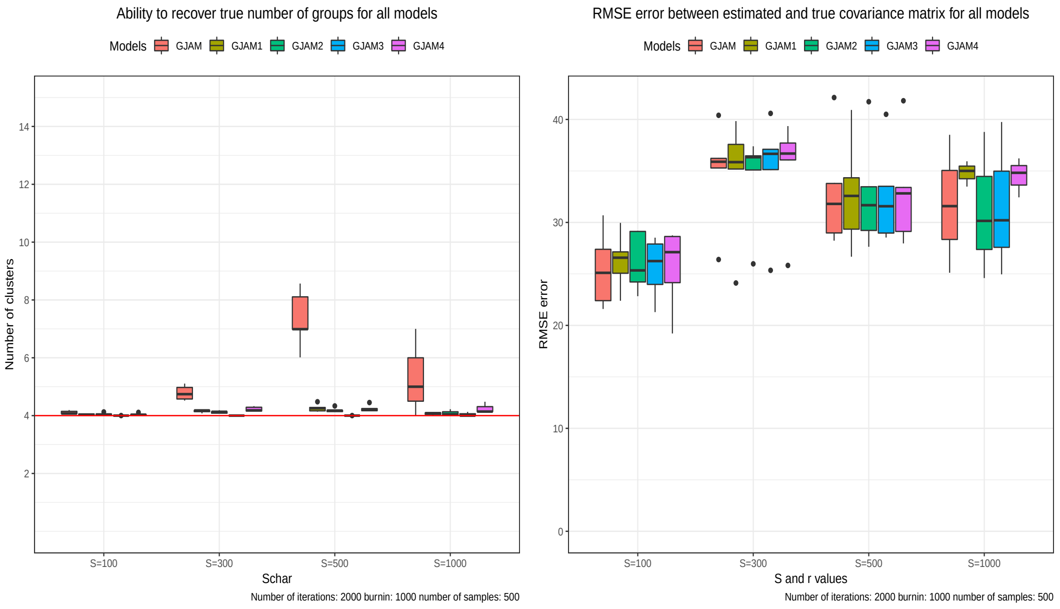

Figure A.8 – *Posterior mean of number of clusters (left) and RMSE error between estimated and true covariance matrix for all models (right) for a number of species* $S \in \{100, 300, 500, 1000\}$,*$n = 500$ and $K_{true} = 4$.*

	S	GJAM	GJAM 1	GJAM 2	GJAM 3	GJAM4
1	100	0.005	0.008	<0.001	0.071	0.096
2	200	0.015	0.006	0.012	0.049	0.062
3	500	0.003	0.046	<0.001	0.045	0.086
4	1000	0.003	0.004	<0.001	0.061	0.058

Table A.1 – *Posterior mean for the last weight p_N averaged across all the simulations for the same set of parameters as in A.8, and for the different models.*

Bibliography

Alexander, J. M., Diez, J. M., and Levine, J. M. (2015). Novel competitors shape species' responses to climate change. *Nature*, 525:515–518.

Antoniak, C. E. (1974). Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174.

Arbel, J., De Blasi, P., and Prünster, I. (2019). Stochastic approximations to the Pitman–Yor process. *Bayesian Analysis*.

Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S. P., Peterson, A. T., Soberón, J., and Villalobos, F. (2011). The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecological Modelling*, 222(11):1810 – 1819.

Bhattacharya, A., Pati, D., and Dunson, D. B. (2013). Anisotropic function estimation using multi-bandwidth Gaussian processes. Technical report, Duke University.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via polya urn schemes. *The Annals of Statistics*, 1(2):353–355.

Carmona, C., Nieto-Barajas, L., and Canale, A. (2019). Model-based approach for household clustering with mixed scale variables. *Advances in Data Analysis and Classification*, 13(2):559–583.

Cazelles, K., Araújo, M. B., Mouquet, N., and Gravel, D. (2016). A theory for species co-occurrence in interaction networks. *Theoretical Ecology*, 9(1):39–48.

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, 85:347–361.

Christensen, R., Johnson, W., Branscum, A., and E. Hanson, T. (2011). Bayesian ideas and data analysis. an introduction for scientists and statisticians.

Clark, J. S. (2017). *gjam: Generalized Joint Attribute Modeling*. R package version 2.2.7.

Clark, J. S., Gelfand, A. E., Woodall, C. W., and Zhu, K. (2014). More than the sum of the parts: forest climate response from joint species distribution models. *Ecological Applications*, 24(5):990–999.

Clark, J. S., Nemergut, D., Seyednasrollah, B., Turner, P. J., and Zhang, S. (2017). Generalized joint attribute modeling for biodiversity analysis: median-zero, multivariate, multifarious data. *Ecological Monographs*, 87(1):34–56.

Connor, R. J. and Mosimann, J. E. (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association*, 64(325):194–206.

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and Ruggiero, M. (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(2):212–229.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230.

Gelfand, A., Silander, J., Wu, S.-s., Latimer, A., Lewis, P., Rebelo, A., and Holder, M. (2006). Explaining species distribution patterns through hierarchical modeling. *Bayesian Analysis*, 1.

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association*, 100(471):1021–1035.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Ghosal, S. and Van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*. Cambridge University Press.

Gnedin, A. and Pitman, J. (2006). Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical Sciences*, 138(3):5674–5685.

Godsoe, W., Franklin, J., and Blanchet, F. G. (2017). Effects of biotic interactions on modeled species' distribution can be masked by environmental gradients. *Ecology and Evolution*, 7(2):654–664.

Golding, N., Nunn, M., and Purse, B. (2015). Identifying biotic interactions which drive the spatial distribution of a mosquito community. *Parasites vectors*, 8:367.

Guisan, A. and Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, 8(9):993–1009.

Guisan, A., Thuiller, W., and Zimmermann, N. E. (2017). *Habitat Suitability and Distribution Models: With Applications in R*. Cambridge University Press.

Guisan, A. and Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, 135(2):147 – 186.

Harris, D. J. (2015). Generating realistic assemblages with a joint species distribution model. *Methods in Ecology and Evolution*, 6(4):465–473.

Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (2010). *Bayesian nonparametrics*. Cambridge University Press.

Huang, A. and Wand, M. P. (2013). Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Anal.*, 8(2):439–452.

Hutchinson (1957). Population studies: Animal ecology and demography. *Bulletin of Mathematical Biology*, 53(1):193 – 213.

Ishwaran, H. and James, L. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173.

Ishwaran, H. and Zarepour, M. (2000). Markov Chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika*, 87(2):371–390.

Jackman, S. (2009). *Bayesian Analysis for the Social Sciences*. Wiley.

Kissling, W. D., Dormann, C. F., Groeneveld, J., Hickler, T., Kühn, I., McInerny, G. J., Montoya, J. M., Römermann, C., Schiffers, K., Schurr, F. M., Singer, A., Svenning, J.-C., Zimmermann, N. E., and O'Hara, R. B. (2012). Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. *Journal of Biogeography*, 39(12):2163–2178.

Lee, C. and Wilkinson, D. J. (2019). A review of stochastic block models and extensions for graph clustering. *arXiv preprint arXiv:1903.00114*.

Letten, A. D., Keith, D. A., Tozer, M. G., and Hui, F. K. (2015). Fine-scale hydrological niche differentiation through the lens of multi-species co-occurrence models. *Journal of Ecology*, 103(5):1264–1275.

Lewandowski, D., Kurowicka, D., and Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100(9):1989–2001.

Link, W. A. and Eaton, M. J. (2012). On thinning of chains in mcmc. *Methods in Ecology and Evolution*, 3(1):112–115.

Lortie, C. J., Brooker, R. W., Choler, P., Kikvidze, Z., Michalet, R., Pugnaire, F. I., and Callaway, R. M. (2004). Rethinking plant community theory. *Oikos*, 107(2):433–438.

Merow, C., Smith, M. J., Edwards Jr, T. C., Guisan, A., McMahon, S. M., Normand, S., Thuiller, W., Wüest, R. O., Zimmermann, N. E., and Elith, J. (2014). What do we gain from simplicity versus complexity in species distribution models? *Ecography*, 37(12):1267–1281.

Miller, J. W. and Harrison, M. T. (2014). Inconsistency of Pitman-Yor process mixtures for the number of components. *The Journal of Machine Learning Research*, 15(1):3333–3370.

Morales-Castilla, I., Matias, M., Gravel, D., and Bastos, M. (2015). Inferring biotic interactions from proxies. *Trends in Ecology Evolution*, 30(6):347–356.

Murphy, K., Viroli, C., and Gormley, I. C. (2017). Infinite mixtures of infinite factor analysers.

Münkemüller, T. and Gallien, L. (2015). Virtualcom: a simulation model for eco-evolutionary community assembly and invasion. *Methods in Ecology and Evolution*, 6(6):735–743.

Nieto-Lugilde, D., Maguire, K. C., Blois, J. L., Williams, J. W., and Fitzpatrick, M. C. (2018). Multiresponse algorithms for community-level modelling: Review of theory, applications, and comparison to species distribution models. *Methods in Ecology and Evolution*, 9(4):834–848.

Norberg, A., Abrego, N., Blanchet, F. G., Adler, F. R., Anderson, B. J., Anttila, J., Araújo, M. B., Dallas, T., Dunson, D., Elith, J., et al. (2019). A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological Monographs*, 89(3):834– 848.

Ovaskainen, O., Abrego, N., Halme, P., and Dunson, D. (2016a). Using latent variable models to identify large networks of species-to-species associations at different spatial scales. *Methods in Ecology and Evolution*, 7(5):549–555.

Ovaskainen, O., Hottola, J., and Siitonen, J. (2010). Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. *Ecology*, 91(9):2514–2521.

Ovaskainen, O. and Soininen, J. (2011). Making more out of sparse data: hierarchical modeling of species communities. *Ecology*, 92(2):289–295.

Ovaskainen, O., Tikhonov, G., Dunson, D., Grøtan, V., Engen, S., Saether, B.-E., and Abrego, N. (2017a). How are species interactions structured in species-rich communities? a new method for analysing time-series data. *Proceedings of the Royal Society B: Biological Sciences*, 284.

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., Roslin, T., and Abrego, N. (2016b). *HMSC: Hierarchical modelling of species community*. R package version 2.2.7.

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., Roslin, T., and Abrego, N. (2017b). How to make more out of community data? a conceptual framework and its implementation as models and software. *Ecology Letters*, 20(5):561–576.

Pitman, J. (2002). *Combinatorial stochastic processes*, volume 1875. Ecole d'Eté de Probabilités de Saint-Flour XXXII.

Pitman, J. (2003). *Poisson-Kingman partitions*, volume Volume 40 of *Lecture Notes–Monograph Series*. Institute of Mathematical Statistics, Beachwood, OH.

Plummer, M. (2014). *rjags: Bayesian graphical models using MCMC. R package*.

Pollock, L. J., Morris, W., Münkemüller, T., Vesk, P. A., Zurell, D., Bistrova, D., Poggiato, G., and Thuiller, W. (2019). Detecting negative (but not positive) interactions from co-occurrence and environment data with joint species distribution models. In preparation.

Pollock, L. J., Morris, W. K., and Vesk, P. A. (2012). The role of functional traits in species distributions revealed through a hierarchical model. *Ecography*, 35(8):716–725.

Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R. B., Parris, K. M., Vesk, P. A., and McCarthy, M. A. (2014). Understanding co-occurrence by modelling species simultaneously with a joint species distribution model. *Methods in Ecology and Evolution*, 5(5):397–406.

Pulliam, H. (2000). On the relationship between niche and distribution. *Ecology Letters*, 3(4):349–361.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650.

Soberon, J. and Peterson, A. (2005). Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodiversity informatics*, 2.

Soberón, J. (2007). Grinnellian and eltonian niches and geographic distributions of species. *Ecology Letters*, 10(12):1115–1123.

Sojoudi, S. (2016). Equivalence of graphical lasso and thresholding for sparse graphs. *Journal of Machine Learning Research*, 17(115):1–21.

Taylor-Rodriguez, D., Kaufeld, K., Schliep, E. M., Clark, J. S., and Gelfand, A. E. (2017). Joint species distribution modeling: dimension reduction using Dirichlet processes. *Bayesian Analysis*, 12(4):939–967.

Thorson, J. T., Ianelli, J. N., Larsen, E. A., Ries, L., Scheuerell, M. D., Szuwalski, C., and Zipkin, E. F. (2016). Joint dynamic species distribution models: a tool for community ordination and spatio-temporal monitoring. *Global Ecology and Biogeography*, 25(9):1144–1158.

Thuiller, W., Guéguen, M., Bison, M., Duparc, A., Garel, M., Loison, A., Renaud, J., and Poggiato, G. (2018). Combining point-process and landscape vegetation models to predict large herbivore distributions in space and time—a case study of rupicapra rupicapra. *Diversity and Distributions*, 24(3):352–362.

Thuiller, W., Münkemüller, T., Lavergne, S., Mouillot, D., Mouquet, N., Schiffers, K., and Gravel, D. (2013). A road map for integrating eco-evolutionary processes into biodiversity models. *Ecology Letters*, 16(s1):94–105.

Tikhonov, G., Abrego, N., Dunson, D., and Ovaskainen, O. (2017). Using joint species distribution models for evaluating how species-to-species associations depend on the environmental context. *Methods in Ecology and Evolution*, 8(4):443–452.

Tjur, T. (2009). Coefficients of determination in logistic regression models—a new proposal: The coefficient of discrimination. *The American Statistician*, 63(4):366–372.

Veech, J. (2013). A probabilistic model for analysing species co-occurrence. *Global Ecology and Biogeography*, 22(2):252–260.

Vladimirova, M. and Arbel, J. (2019). Sub-Weibull distributions: generalizing sub-Gaussian and sub-Exponential properties to heavier-tailed distributions. *Preprint*.

Warton, D., D. Foster, S., De'ath, G., Stoklosa, J., and K. Dunstan, P. (2015). Model-based thinking for community ecology. *Plant Ecology*, 216:669–682.

West, M. A. and Escobar, M. A. R. (1993). Hierarchical priors and mixture models, with applications in regression and density estimation. *Institute of Statistics and Decision Sciences, Duke University*.

Wilkinson, D. P., Golding, N., Guillera-Arroita, G., Tingley, R., and McCarthy, M. A. (2019). A comparison of joint species distribution models for presence–absence data. *Methods in Ecology and Evolution*, 10(2):198–211.

Yates, K., Bouchet, P., Caley, M., and K, e. a. M. (2018). Outstanding challenges in the transferability of ecological models. *Trends in Ecology & Evolution*, 33(10):790–802.

Zurell, D., Pollock, L. J., and Thuiller, W. (2018). Do joint species distribution models reliably detect interspecific interactions from co-occurrence data in homogenous environments? *Ecography*, 41(11):1812–1819.