Application of MaxEnt & Network Analysis

- Barro Colorado Forest -

Paolo Frazzetto

Statistical Mechanics of Complex Systems - 7/4/2019

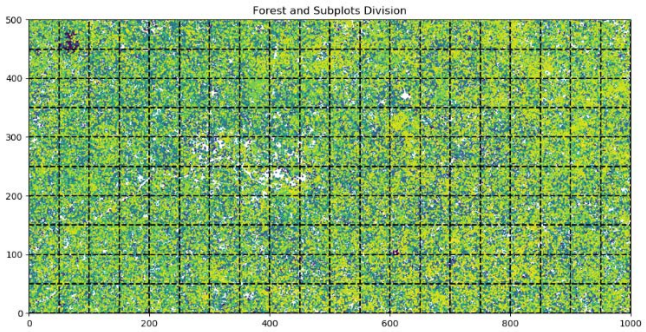
TASK 1

In this assignment we were given a dataset containing the census data of a forest located at Barro Colorado Island, Panama. The whole data analysis has been carried out by means of a Python Jupiter notebook, attached to this report.

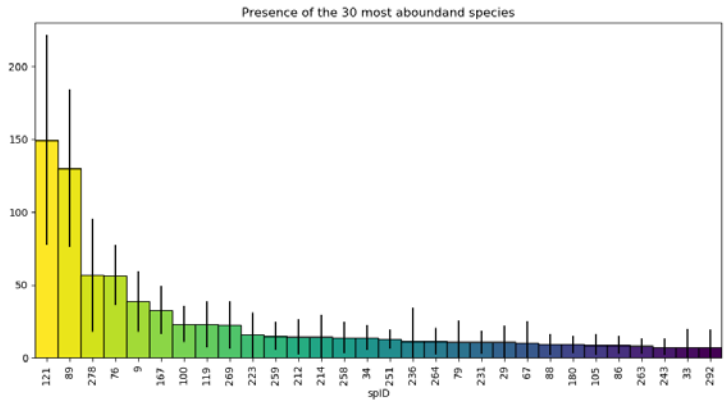
First, the raw dataset has been imported from the `.csv` file and it has been pre-processed by discarding all the dead trees, resulting in data for $\approx 200\,000$ alive trees spread over a 50 hectares region. The whole population is made up of $S = 299$ different species. Each species has a given label, but for simplicity of the further analysis I encoded them in numerical categories, called `'spID'`.

TASK 2

In order to have more sample data to compute statistics on, the forest has been split in 200 subplots that will be treated as i.i.d., as it is pictorially shown in the next image:

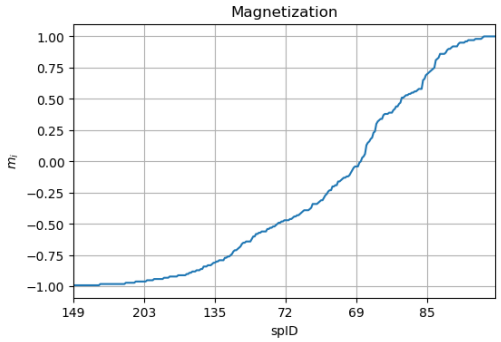


For each subplot, the count of every species has been computed and gathered in the *abundances vector,* that can be fully viewed via the notebook. By averaging over the 200 subplots, we get an estimation of the average *presence* of each species p_i. In the following histogram the 30 most common species presence with standard deviation error bars are reported:



TASK 3

We can now build the first Maximum Entropy model: from the abundances vector I replaced every present species with $+1$ and every absent one with -1. In this way I defined a new vector σ for each subplot, with elements $\sigma_i = \{\pm 1\},\ i = 1, \dots, S$, and this problem can be easily translated into the Ising Model. The only constrain is that $\langle \sigma_i \rangle_{emp} = m_i$ where $p_i = (m_i + 1)/2$, thus we are requiring that the for each species the "magnetization" matches the one obtained from the data:



We are looking for the Lagrangian parameters λ_i such that:

$$P(\sigma) = \frac{1}{Z} e^{-\sum_{i=1}^{S} \lambda_i \sigma_i}$$

With similar steps to the computation of the partition function of the Ising model, it results that

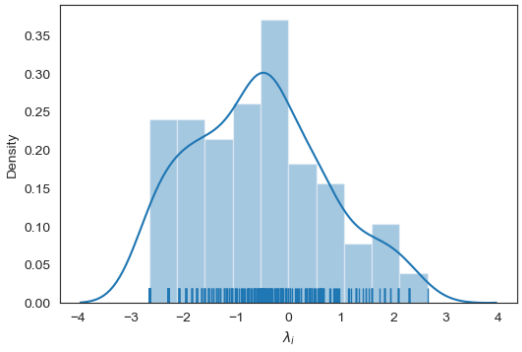
$$Z = 2^S \prod_{i=1}^{S} cosh(\lambda_i)$$

So eventually

$$\frac{\partial log(Z)}{\partial \lambda_j} = tanh(\lambda_j) = m_j = \langle \sigma_j \rangle_{emp}$$

$$\Longrightarrow \lambda_i = tanh^{-1}(m_i)$$

This analytic expression tells us that the parameters are closer to zero for those species that are averagely present everywhere in the forest $(p_i \approx 0.5)$, whereas for those species that are the rarest or most abundant the parameters assume larger values, as one could easily expect from the definition of σ. In fact, notice that since there are 9 species that are present in every subplot, their λ_is diverge. Therefore, these species have been discarded in the following plot, that shows the distribution of the parameters together with its Kernel Density Estimation:



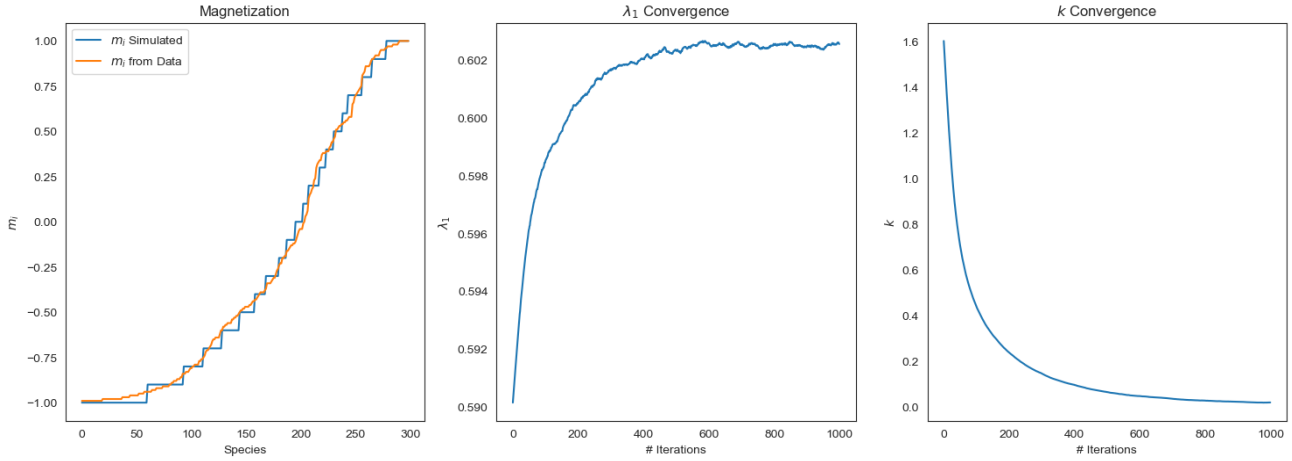
TASK 4

A more advanced model would require also that the average number of species present (S_+) or absent (S_-) in a plot matches the real-world data. In this case the Hamiltonian is:

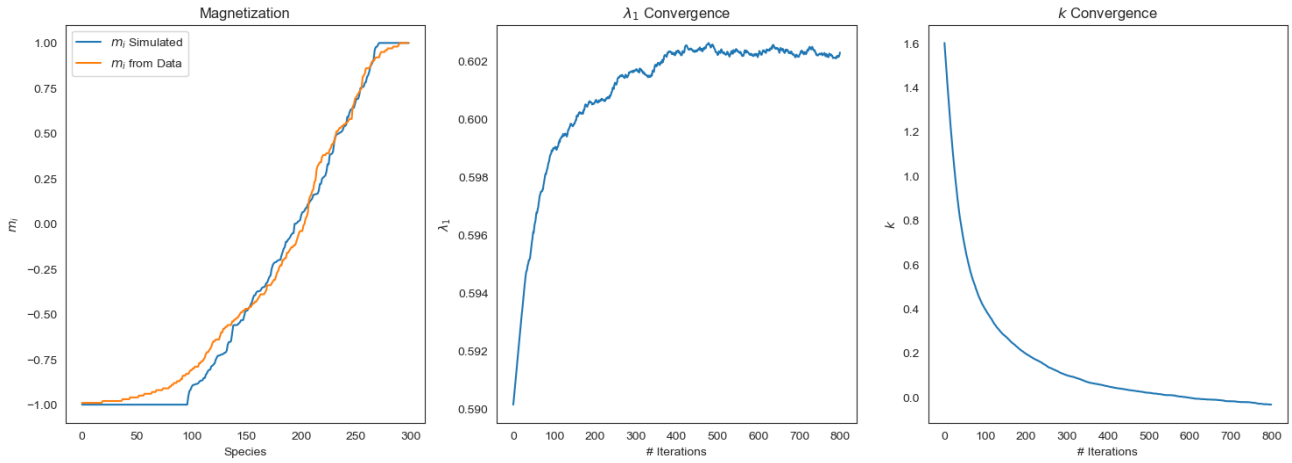
$$H(\sigma) = -\sum_{i=1}^{S} \lambda_i \sigma_i + \frac{k}{S}\left(\sum_{i=1}^{S} \sigma_i\right)^2$$

With Lagrangian parameters λ_i, k such that $p_i = (m_i + 1)/2$ and $\langle\left(\sum_{i=1}^{S} \sigma_i\right)^2\rangle_{emp} = \langle(S_+ - S_-)^2\rangle_{emp}$.

There is no analytical expression to compute them, so the parameters can only be estimate by some approximations or by means of a Monte Carlo simulation. I chose to follow the latter way with two different approaches. In the first one, I selected the coefficients of Task 3 as starting points, then by means of the Metropolis algorithm I begin with 20 random spin configurations and I store their last values after 2000 energy minimization iterations. With these configurations I compute the new constrains and I update the λs with a standard Gradient Descend with hyperparameter $\eta = 0.0001$ and looped for 1000 runs. The results are shown in the following plot, comparing the magnetization with the data, the convergence of λ_1 and k:



In the second approach, I start with one random spin configuration, but I store the last 20% of 10000 Metropolis iterations. It provides slightly worse results but six times faster than the first method:



These simulations were quite time demanding for my laptop (the first one takes more than 30 minutes) but the results are satisfactory: if we compare the values of the quadratic constrain, $m^{(2)}$, computed with the

last iteration values of the Lagrangian parameters, it holds that $m^{(2)}_{data} = 21.33$, $m^{(2)}_{1st\ appr.} = 21.26$, $m^{(2)}_{2nd\ appr.} = 25.17$. In conclusion, I am confident that with more computational resources, a better tuning of the hyperparameters and more sophisticated algorithms than the Gradient Descent and Metropolis, it is doable to perform simulations that agree well with the data. For reference, all values of the parameters are displayed in the notebook.

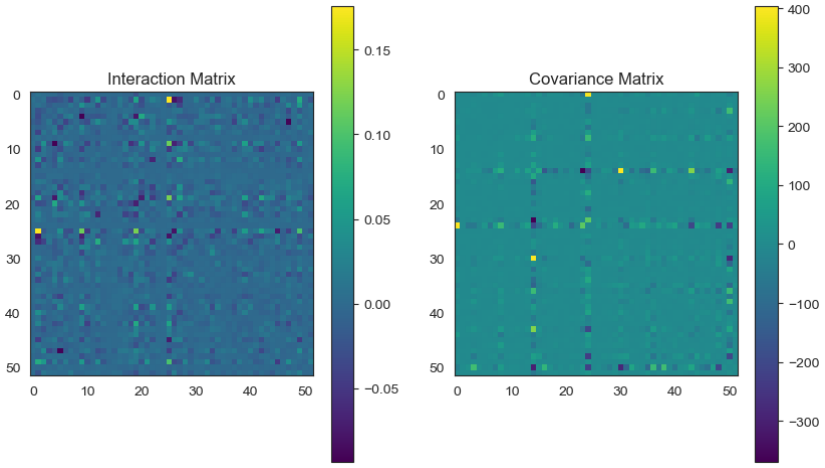
TASK 6

We can compute the probability distributions for the most abundant species thanks to the Gaussian approximation, with the constrains $\langle x_i \rangle_{emp} = \langle x_i \rangle_{model}$ and $\langle x_i x_j \rangle_{emp} = \langle x_i x_j \rangle_{model}$. The definition of "most abundant species" is determined by $\langle x_i \rangle_{emp} > \sigma(x_i)$, that is satisfied by 52 species.

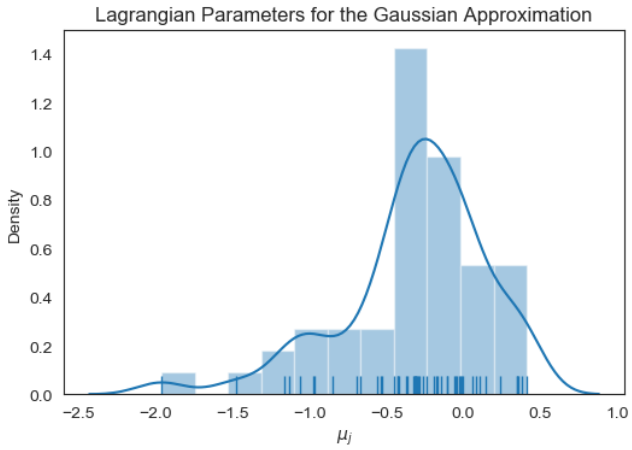
As we have seen in the lectures, it holds that

$$\mu_i = -M\langle x_i \rangle_{emp} \qquad Cov(x_i, x_j) = M^{-1}_{ij}$$

The covariance matrix and M are shown in the following image. Notice that the diagonal elements have been set to zero, to ignore the standard deviation and self-interactions.

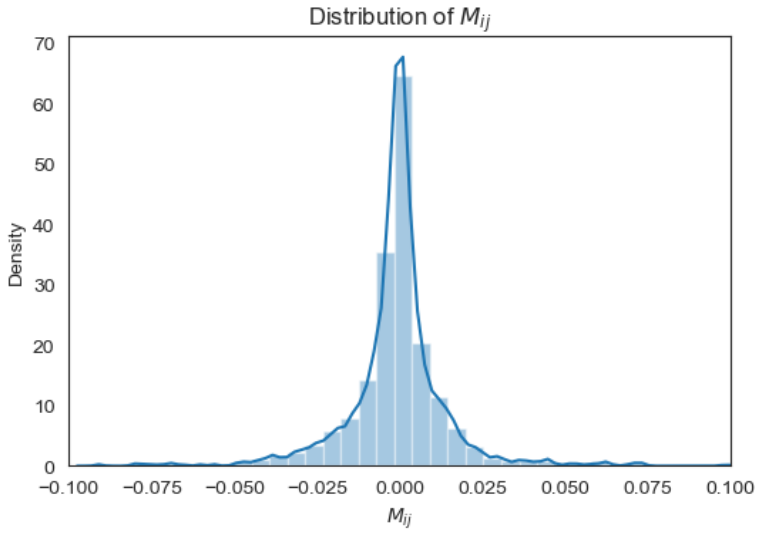


Eventually, the Lagrangian parameters μ_i that model the probability distribution of the most important species are distributed as follows:

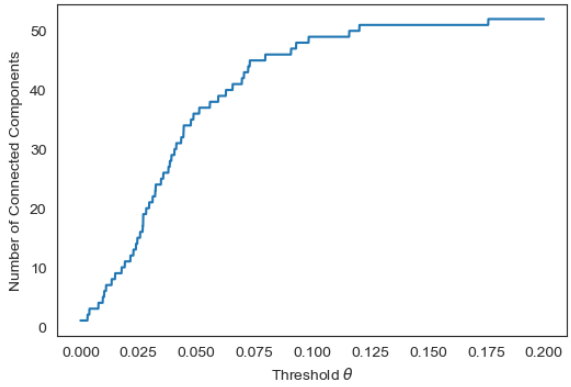


TASK 7

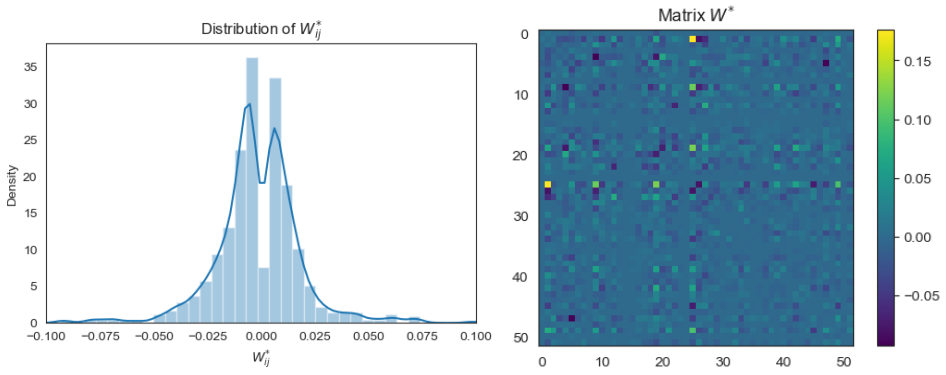
The distribution of the $M_{ij}s$ and its KDE are reported in the next figure. It is peaked with mean -0.00056 and standard deviation of 0.017, so generally speaking the interactions among species are weak - close to zero:



We would like to set to zero all the interactions smaller than a certain threshold θ^*, so this new trimmed matrix can be considered the weighted adjacency graph W of the species interactions networks. The number of connected components in function of θ is shown below:

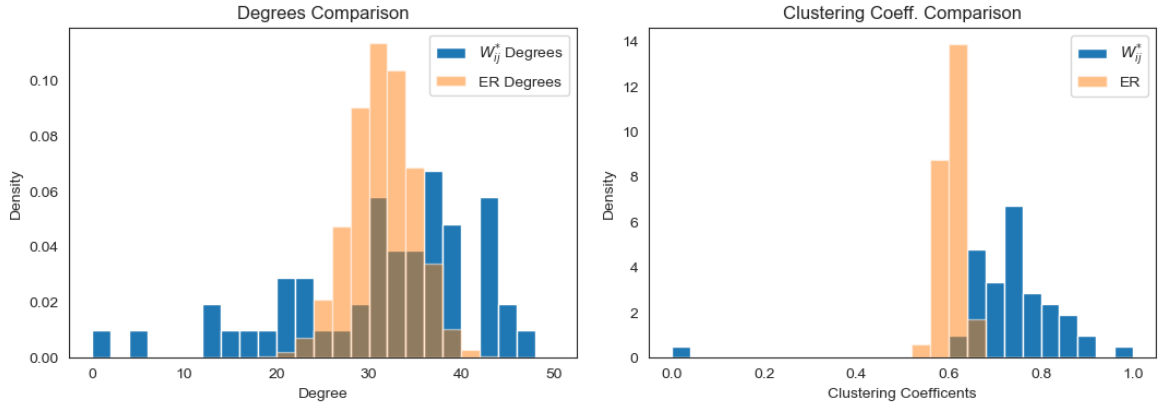


As suggested, I chose θ^* such that for any $\theta > \theta^*$ the graph is not made by one single connected component. It results that $\theta^* = 0.0030$ and so the W_{ij}^* are distributed as:



Finally, we wish to study this network of strongly interacting species and compare it with random Erdos-Renyi graphs. I generated 200 ER graphs with parameter $p = \frac{\langle k \rangle}{n-1}$ to compute some averaged structural properties and compare them with the trees network:

Properties	*Barro Colorado*	*Erdos-Reniy*
Degree k	52	30.9
Diameter	3	2
Mean Clustering Coeff.	0.73	0.61
Mean Degree Assortativity Coeff.	-0.023	-0.039
Mean Betweenness Centrality	0.00799	0.00786



At a first glance, we can tell that the forest network does not (luckily) follow a random structure.