Exercise 01 - Application of Maximum Entropy and Network analysis

Federico Agostini, Federico Bottaro, Gianmarco Pompeo

1 Introduction

We are presented with a dataset containing information about a 2005 tree census performed at the Barro Colorado Island in Panama. The flora of this 50-hectare piece of land has been classified according to several parameters since 1982 and censused accordingly every 5 years.
The scope of this studies is to retrieve information about the dataset and to apply the Maximum Entropy method in order to be able to model different scenarioes.

2 Presentation of the data

The whole dataset is made of 368,123 rows, but we will trim it by removing all the dead plants; furthermore, we will only be considering the species classification and the coordinates of the censused trees, disregarding all the other parameters.

The total number of alive trees is 208,387 consisting of about 56.6% of the whole dataset. The total number of different species is $S = 299$. In Fig. 1 there is a visualization of how the species are distributed around the island.

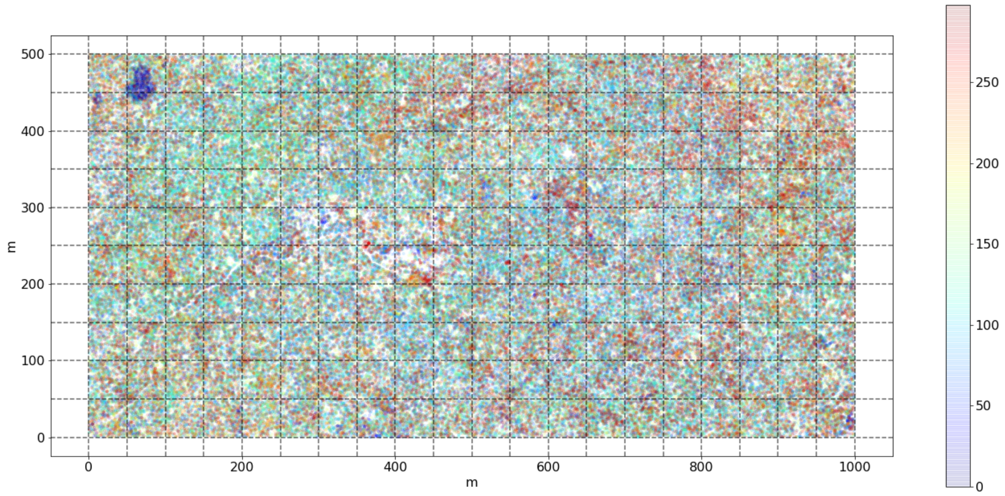
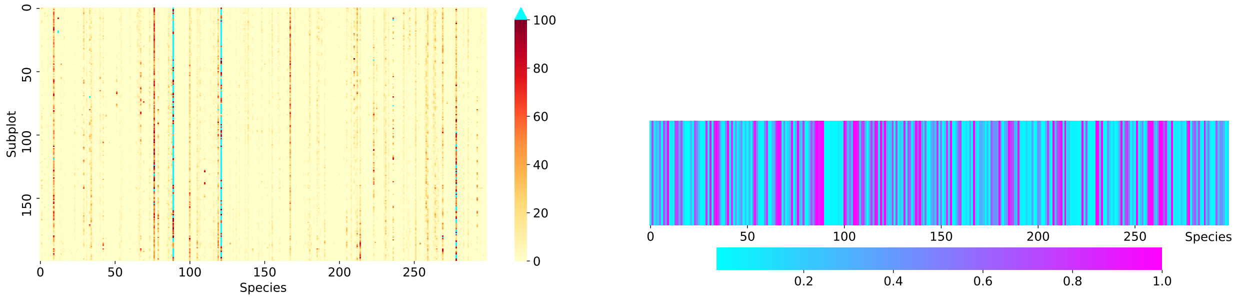


Figure 1: *Distribution of alive species around the island, where different colors indicate different species. As we can see the plot shows a big variety of species and they are well mixed; the dashed grid separates the different subplots.*

We divide such plot into regularly-shaped subplots of 0.25 hectares each, meaning we will have 200 subplots 50×50 m in size. We will assume these subplots to be statistically independent in order to be able to perform analysis on them.

We build the matrix of the abundances $\mathbf{X}$, where X_{ij} is the number of individuals of species j in subplot i; moreover, we also consider the vector of presence p_i (for $i = 1, \dots, S$) averaged over the subplots - that is, an average of 1 or 0 across each subplot depending on whether species i is present or not in the subplot itself.
Given that the size of the dataset is considerable, we represent these objects only graphically and we invite the reader to look at the Jupyter Notebook (from here, `JN`) for an explicit display of all their values.



(A) *Heatmap of the abundance matrix* $\boldsymbol{X}$.

(B) *Heatmap of the vector of presences* p_i *averaged over the subplots.*

Figure 2: *Graphical representation of the abundace matrix and average presence vector.*

In the abundance matrix we find that there are very few species that are highly dominant; to make the heatmap of $\mathbf{X}$ easier to interpret, we have decided to represent species that occur more than 100 times with the same color: this helps to discriminate the presence and the geographic distribution of the huge majority of the species which occur not quite as often (anyways, the original abundance matrix can still be found in the ⇒ `JN`).

3 Application of Maximum Entropy

We consider three different models whose parameters are computed using the Maximum Entropy method (*Max Ent* in the following).

3.1 Maximum Entropy Model 1

In the following subsection we present a first attempt at building a Max Ent 1 model on the dataset. In this first simple model the hamiltonian and the constraints are:

$$\mathcal{H}_1 = -\sum_{i=1}^{S} \lambda_i \sigma_i \qquad , \qquad \frac{1+m_i}{2} = p_i \qquad i = 1, ..., S \tag{1}$$

with $\langle \sigma_i \rangle_{emp} = m_i$. This relation serves the purpose of introducing a new average presence obtained by assigning a value of 1 to the presence of a species and a value of -1 instead of 0 to the absence: in this way it is possible to map this task into an Ising-like model.

At this stage we do not need to simulate anything to find the lagrangian parameters λ_i because it is possible to obtain analytical solutions by considering the partition function of the system $\mathcal{Z}$, as shown in the following:

$$\mathcal{Z}(\lambda_i) = \sum_{\{\sigma_i\}} e^{\sum_i \lambda_i \sigma_i} = \sum_{\{\sigma_i\}} \prod_i e^{\lambda_i \sigma_i} = \prod_i \sum_{\{\sigma_i\}} e^{\lambda_i \sigma_i} = 2^S \prod_i \cosh(\lambda_i) \tag{2}$$

where $i = 1, ..., S$ and $\sum_{\{\sigma_i\}} = \sum_{\sigma_i = -1,1}$ represents the sum over all possible configurations.

From this, we analytically compute

$$\langle \sigma_i \rangle = -\frac{\partial \ln \mathcal{Z}}{\partial \lambda_i} = -\tanh(\lambda_i) \tag{3}$$

and obtain our λ_i simply by inverting

$$\lambda_i = \tanh^{-1}(m_i) \qquad \text{i=1,...,299} \tag{4}$$

In Fig. 3 it is possible to have a visual representation of these lagrangian parameters obtained using the script.

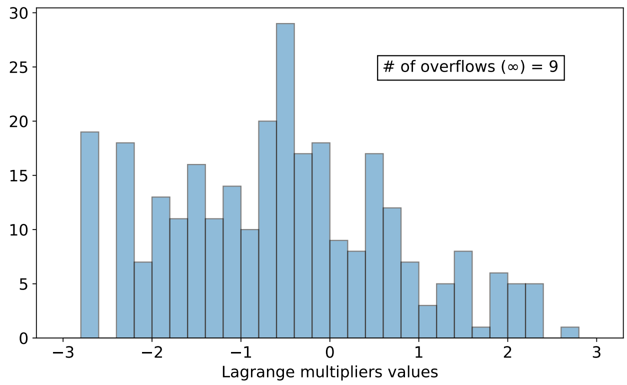


Figure 3: *Histogram of the values of λ_i for Max Ent 1. Most of them are between -3 and 3, but there are also few parameters (3%) that have an infinite value and they are not shown in this figure.*

3.2 Maximum Entropy Model 2

For this second model we consider a slightly more complicated hamiltonian

$$\mathcal{H}_2 = -\textstyle\sum_{j=1}^{S} \lambda_j \sigma_j - \frac{k}{S}\left(\sum_{j=1}^{S} \sigma_j\right)^2 \tag{5}$$

containing a self-interaction term.
In addition to the constraint of Max Ent 1 (eq. 1) we now have

$$\left\langle \left(\textstyle\sum_{j=1}^{S} \sigma_j\right) \right\rangle_{emp} = \left\langle (S_+ - S_-)^2 \right\rangle_{emp} \tag{6}$$

where S_+ and S_- are the number of species respectively present or not in a given subplot.

To estimate the lagrangian parameters and consequently build a model, we have to simulate the process because direct sampling as done previously is not feasible. For the simulation we use a *Metropolis* algorithm which starts from a randomly initialized configuration and generates more according to Eq. 5, accepting them automatically if their value of energy is lower than the previous; if that is not the case the algorithm will accept the updated configuration with probability proportional to $e^{(E_t - E_{t+1})}$. We tested the algorithm for 10,000 iterations and the system appears to have reached a minimum of the energy (apart from fluctuations) even after a couple thousands epochs; we hence decided to run the algorithm for 5,000 iterations when simulating data and then keep the final 25% configurations.
The lagrangian parameters are calculated through a *Gradient Descend* procedure, with the following update rule

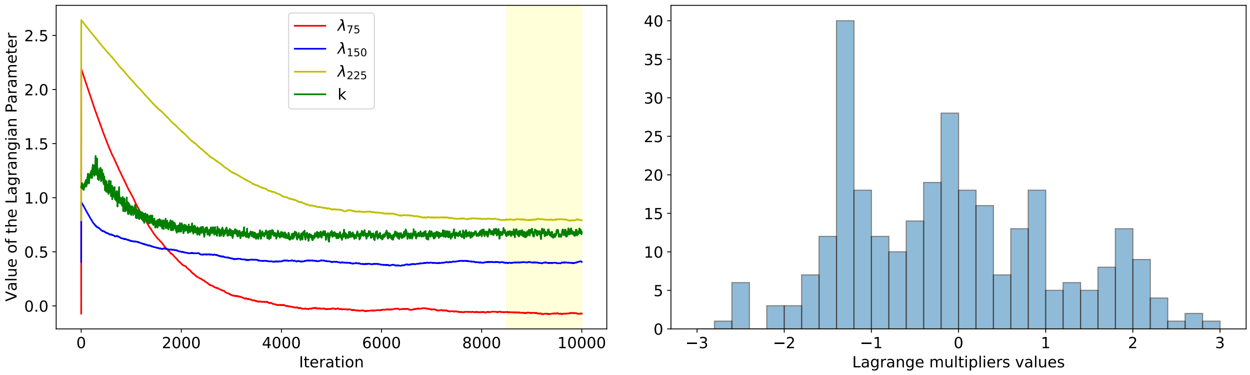
$$\lambda(t+1) = \lambda(t) + \eta(\langle \mathcal{C} \rangle_{emp} - \langle \mathcal{C} \rangle_{model}) \tag{7}$$

where η is called learning rate, set to 10^{-3}, and $\mathcal{C}$ represents the constraints computed either on the original data (*emp*) or based on the model generated by *Metropolis* (*model*). At each iteration, the compute of $\langle \mathcal{C} \rangle_{model}$ is done using the states generated by the *Metropolis* algorithm discussed above.
The final values of the parameters are obtained averaging the values from the final 15% of iterations: this way we can smooth out possible fluctuations of the *Gradient Descent* algorithm.

In the plot (Fig. 4a) we show how these parameters behave as the iteration process goes on. We end up estimating:

$$k = 0.6731 \qquad ,$$

while in the histogram to the right - Fig. 4b - we show the distribution of the final 299 values found for the λ_i (⇒ `JN` for the complete list of values).
We also introduced a Random Field model but, due to the limited space, this is not presented in this project (however, see ⇒ `JN`).



(A) *Trend of k and three randomly-chosen λ_i over the 10,000 iterations. Convergence starts to be visible after a couple thousands iterations. The final values are calculated averaging over the last 1,500 iterations (marked with the yellow background in the plot.)*

(B) *Histogram of the values of λ_i for Max Ent 2.*

Figure 4: *Results for the simulations performed for Max Ent 2.*

3.3 Maximum Entropy Model 3

The last model we developed is related to the abundances of the species instead of their presence. As a matter of fact, to build this Max Ent 3 model we are going to impose two constraints,

$$\langle x_i \rangle_{emp} = \langle x_i \rangle_{model} \tag{8}$$

where $\langle x_i \rangle$ is the average abundance of a species over the subplots, and

$$\langle x_i x_j \rangle_{emp} = \langle x_i x_j \rangle_{model} \tag{9}$$

this time related to the two-point correlation function.
Moreover we select only the species for which

$$\langle x_i \rangle_{emp} - \sigma_{x,i} > 0 \tag{10}$$

where $\sigma_{x,i}$ is the standard deviation of the species i calculated from the data (over the subplots).
Following the maximum entropy principle it is possible to obtain an analytical solution to find the lagrangian parameters, that in this case are labeled μ_j, related to the first constraint, and $\mathbf{M}$, the interaction matrix deriving from the second one.
We now consider the probability of having the configuration $\vec{x}$:

$$P(\vec{x}) = \frac{1}{\mathcal{Z}} e^{-\sum_{i=1}^{S} \mu_i x_i - \frac{1}{2}\sum_{i,j=1}^{S} M_{ij} x_i x_j} \tag{11}$$

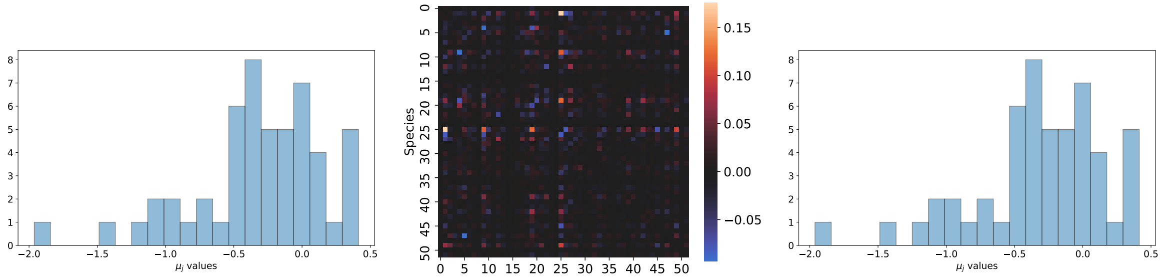
Writing the partition function of the system explicitly in matricial form we obtain

$$\mathcal{Z} = \int e^{-\frac{1}{2}\vec{x}^T \mathbf{M} \vec{x} - \vec{\mu}^T \cdot \vec{x}} \, d\vec{x} \tag{12}$$

and by introducing the Gaussian approximation we get

$$\vec{\mu} = -\mathbf{M} \langle \vec{x_i} \rangle_{emp} \qquad , \qquad M_{ij}^{-1} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle \tag{13}$$

In addition to that, we set all the diagonal values of $\mathbf{M}$ to 0: this means we will be disregarding self-interactions.



(A) *Histogram of the values of μ_j for Max Ent 3.*

(B) *Heatmap of matrix $\boldsymbol{M}$.*

(C) *Distribution of the entries of $\boldsymbol{M}$.*

Figure 5: *Graphical visualizations of lagrangian parameters for Max Ent 3.*

As usual, in Fig. 5 it is possible to visualize the distribution of the lagrangian parameters μ_j and the heatmap of the matrix $\mathbf{M}$ (obtained as the inverse of the covariance matrix) with the corresponding histogram showing the distribution of its values.

We can barely recognize any pattern in such distributions, but from the heatmap of the interaction matrix we see how very few species act as strong interacting nodes (most of the entries are dark, corresponding to a weak interaction strength).

4 Network analysis

We consider the Max Ent 3 model we just built: the distribution of the entries of M_{ij} is shown in Fig. 5c.
We can consider this matrix as a weighted adjacency graph W. If we select a threshold θ and put $M_{ij} = 0$ if $|M_{ij}| < \theta$, we can study the number of connected components as a function of θ itself.
Fig. 6 shows the behaviour of the graph as a function of this parameter: we can see there exists a critical value θ^* for which the number of single connected components becomes larger than one (the graph is not single-connected anymore).
We find

$$\theta^* = 0.0031 \quad .$$

We are now interested in studying the properties of W^*, the graph obtained with the highest value of θ that makes it single-connected (formally, $W^* \equiv W(\theta^* - \varepsilon)$ with $\varepsilon \to 0$).

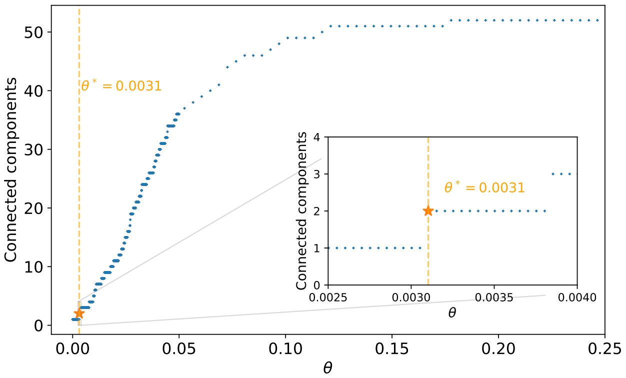
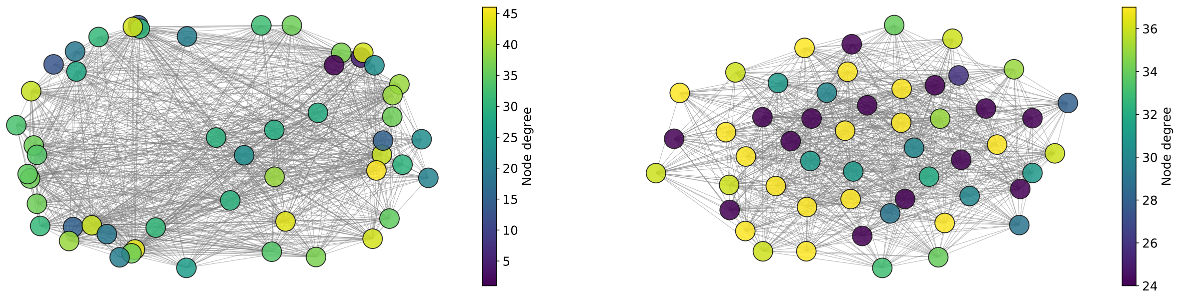


Figure 6: *Number of connected components as a function of the threshold θ; the orange star points out the value of θ^*.*

In Fig. 7a we can visualize such network. This graph has

$$\mathcal{D} = diameter = 3 \qquad \mathcal{R} = radius = 2 \qquad (Deg)_{ass} = -0.2361$$

Other properties are visually displayed in Fig. 8, together with the subsequent comparison with the ER graph. As a last study, in fact, we compare the graph W^* so obtained with a random Erdos-Rényi (ER) graph. We create that random graph using a probability of connection $p = \frac{A_d}{N-1}$ where A_d is the average degree of nodes in W^* and N is the number of nodes; the links follow a binomial distribution. In Tab. 1 and in Fig. 8 we can see the comparison between various features of the two graphs.



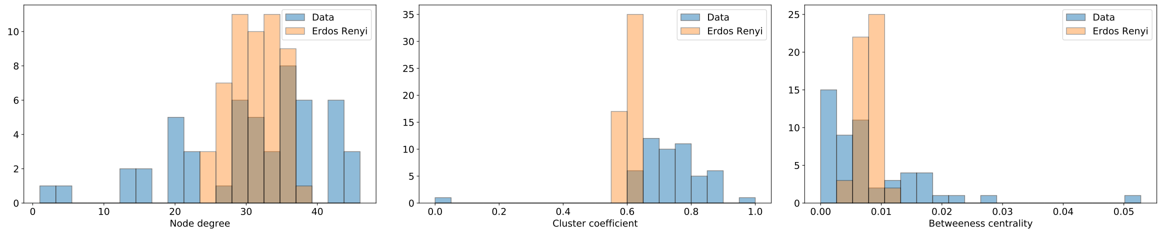
(A) *Graph of W^*.* **(B)** *Random ER graph generated for comparison.*

Figure 7: *Graphical representation of the Network W^* and the random Erdos-Rényi (ER) graph generated using a probability of connection $p = \frac{A_d}{N-1}$, where A_d is the average degree of nodes in W^* and N is the number of nodes.*

	Radius	Diameter	Density	Clustering average
W^*	2	3	0.601	0.732
ER graph	2	2	0.609	0.604

Table 1: *Comparison of the most significant quantities of W^* and the random ER graph (see Figure 8 for further visual comparisons).*

We also display other properties of W^* which are more meaningful graphically; in orange, the corresponding distributions for the random ER graph.



(A) *Histogram of the degrees of the nodes.* **(B)** *Histogram of the clustering coefficient.* **(C)** *Histogram of the betweeness centrality.*

Figure 8: *Visualization of other properties of W^*. In orange the properties of the ER graph for comparison.*

The comparison shows that the values look very little alike, which is somehow expected: the ER graph is generated randomly and it is only connected to W^* through the parameters given in input. This graph, on the other hand, represents real interactions between different plant species and it appears safe to assume that such interactions do not happen simply at random.