Università degli Studi di Padova

Dipartimento di Fisica e Astronomia “Galileo Galilei”

Corso di Laurea in Physics of Data

EX 1 STATISTICAL MECHANICS OF COMPLEX SYSTEM

OF

MARCO AGNOLON, TOMMASO TABARELLI

Anno accademico 2018-2019

The system

Our task was to analyze the trees composition of the forest of the Barro Colorado Island. The database contains many information about every tree in the selected fifty hectares, among them we considered only the positions and the species of the trees. The position are given by a couple of number which represent the two-dimensional coordinates of the tree. We set a number for each species to have a one to one identification of the species with a specific number up to 299, i.e. the total number of species.
In order to build a statistic we divided the fifty hectares in 200 subplots of 0.25 hectares each. Then we mapped each tree in the correspondent subplot. In this way every subplot was made of different types of species and we could define a variable p_i that represents the average presence of a species among all the subplots: if $p_i = 1$ a species was always present, vice versa if ($p_i = 0$) was missing in every subplot (i is representing the species number).
To use a paradigm similar to the Ising model we need quantities that goes from -1 to $+1$. Hence, we operated a change of variables: $m_i = 2p_i - 1$, where m_i is defined to be the average magnetization over the subplots, and so it is obtainable from the empirical data.
The local variables of our model are σ_i that can assume values ± 1 and represent the presence or the absence of a species; they also correspond to the spins of an Ising model. A set of 299 σ_i is called *configuration*; it has a energy given by an appropriate Hamiltonian.
Our goal was, after having choose an appropriate Hamiltonian, to find the Lagrangian parameters that better fit the original data.

Max entropy model 1

The Hamiltonian assumed for this model is:

$$H = -\sum_{i=1}^{S} \lambda_i \sigma_i$$

It is important to notice that this kind of model does not take into account any interaction between species and so it considers them independent. This is a very poor approximation but it is useful because it will lead to an analytically solvable model and hence to very simple results. The probability assigned by the max-entropy principle is given by:

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

Then, the partition function is the following:

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

The theoretical average is:

$$\langle \sigma_i \rangle_{model} = -\frac{\partial \log(Z)}{\partial \lambda_i} = -\tanh(\lambda_i) \implies \lambda_i = -\tanh^{-1}(m_i)$$

The corresponding empirical quantity $\langle \sigma_i \rangle_{emp}$ can be chosen to be the average presence of the different species: the presence is "1" if the species is present in a subplot and "0" otherwise; the average presence is the sum of the presences of a species on all subplots divided by the number of subplots.

To find λs we imposed:

$$\langle \sigma_i \rangle_{model} = \langle \sigma_i \rangle_{emp} = m_i$$

and then we took the $-\tanh^{-1}$ of m_i. The results are shown in Figure 1. One can notice that the distribution of the λ_i is approximately around zero.

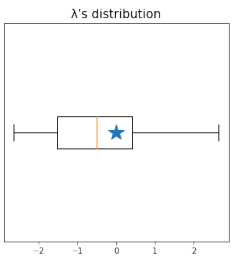

Figure 1: Distribution of λ_i (boxplot) compared with 0 (represented by the star).

Max entropy model 2

If we want to build a more realistic model, we have to use another Hamiltonian, in which the interactions between the species are considered:

$$H_{\lambda_i,K}(\sigma_i) = -\sum_{i=1}^{S} \lambda_i \sigma_i - \frac{K}{S}\left(\sum_{j=1}^{S} \sigma_j\right)^2$$

From the data we can evaluate the following constraint:

$$\langle \left(\sum_{i=1}^{S} \sigma_i\right)^2 \rangle_{data} = \langle (S_+ - S_-)^2 \rangle_{emp}$$

where S_+ and S_- are respectively the number of species present and absent (in a subplot), and then the average is taken over all subplots.

This kind of model can not be solved analytically; our approach was the following:

1. to simulate the data via metropolis importance sampling, starting from arbitrary choices of the initial parameters (beacuase of the convexity of the problem; we explain this later);
2. to update them using a gradient descent algotihm and constraints (evaluated after each iteration);
3. to repeat the 2 previous steps untill convergence or after a certain number of iterations.

Step 1 We needed simulations because, as already mentioned, the model was not analitycally solvable. Furthermore, the number of possible configurations was huge and it was not possible to evaluate the partition function, hence constraints were not explicitly avaiable. In order to be able to estimate the constraints, we generated a certain number of 1-d configurations made by 299 spins that could assume only ± 1 values (in our notebook the number of configurations generated is 30k and we keep only the last fifth of them) and averaged the respective quantities over them.
As initial condition for λs we chose to adapt the "Max Entropy Model 1" results. For this purpouse we modified the coefficients found there, discarding the coefficients whose values diverged to infinity and substitute them with an arbitrary large number (in our case we chose 10^5). For the coefficient K we chose, after some attempts, to set it to 0.5 because we noticed that it was a value for which the algorithm converged faster than others choices.
Metropolis importance sampling needed a certain number of steps to assure oscillations of configurations's energy became "small" (see Figure 2). To avoid problems due to *transient configurations*, we generated a lot of them and we kept only the last ones.

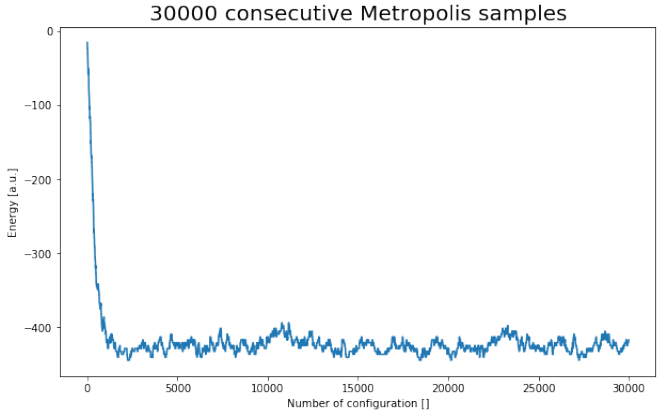

Figure 2: 30000 consecutive metropolis samples generated by the algorithm: one can notice that after a transient period, the energies start to oscillate in proximity of a fixed value.

Step 2 Thanks to the fact the function is convex with respect to the parameters we can use a gradient descent rule to update them.
The updating rules turned out to be:

$$\lambda_i \Leftarrow \lambda_i + \eta(\langle \sigma_i \rangle_{Data} - \langle \sigma_i \rangle_{model})$$

$$K \Leftarrow K + \frac{\eta}{S}\left(\langle \left(\sum_{j=1}^{S} \sigma_j \right)^2 \rangle_{Data} - \langle \left(\sum_{j=1}^{S} \sigma_j \right)^2 \rangle_{model} \right)$$

In our code we decided to implement a more advanced gradient descent method called A.D.A.M.S. because it is faster and it auto-adjust its learning rate (see Pankaj Mehta, Ching-Hao Wang, Alexandre G. R. Day, and Clint Richardson, *A high-bias, low-variance introduction to Machine Learning for physicists* for details).

Step 3 We decided to iterate the algorithm 2000 times to be sure that it arrived to convergence (the update rules take a while to establish parameter values in a convergerce regime). Moreover, it is not excluded that the set of configurations we select from metropolis sampling does not represent properly the energy landscape; if so, parameter updatings would end up in different minima due to the stochasticity of configurations sampling even if the problem is convex. The results can be found in the section "Task 4" of the jupyter notebook.

Comparing results with RFIM

After this simulation we plotted the resulting point (σ, K) (where σ is the standard deviation of the λ_i) in the K *vs* σ plane, comparing it with the expected results for a random field Ising model with random fields gaussian distributed. The results are showed in Figure 3.

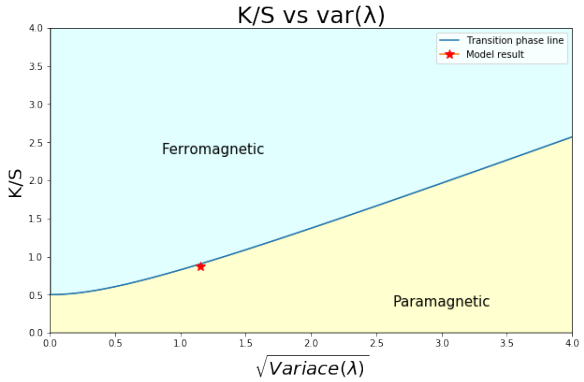

Figure 3: Phase diagram of a RFIM with $\beta = 1$ and results of the algorithm plotted in the same axis.

As one can see, the results are located exactly on the ferromagnetic-paramagnetic transition line; this leads us to think that the correlation length, for what conecerns this specific realization of the Hamiltonian, is aprroximately infinite (as indeed it is in a critical point).
This can also be deduced looking at the Hamiltonian we used in this model: in fact the intensity of all possible interactions, represented by the parameter K, among all species is assumed to be the same; in this way there is no typical finite correlation length.
The concrete consequence of this fact is that the different tree species are stronlgy dependent one by the other, as we now explain:

- they are not randomly distributed, indeed this would lead to a paramagnetic system in which all sites were independent;
- on the other hand, they are also not identically distributed as they would be in a ferromagnetic phase, otherwise if 1 species was present, then all the others would be present as well (and this fact is not compatible with the observations).

In order to check that the results were compatible with the RFIM, we plotted the average magnetization in the following way: using last updated parameters, we generated another metropolis simulation in the same way as before and evaluated average magnetization comparing them with those of data.

This comparison plot shows that configurations generated with parameters estimated via the algorithm are compatibles with experimental ones.

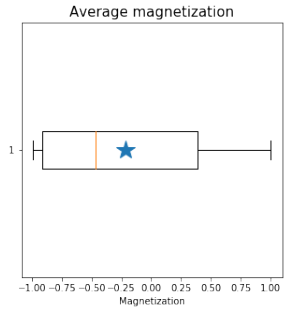

Figure 4: Boxplot of the experimental magnetization (evaluated from the 200 initial configurations) compared with the average magnetization evaluated from simunlation (represented by the star).

Max entropy model 3

For what concerns this point we took into account the abundances (namely, the number of individuals in each subplot) of the different species instead of the presences. In this way, the corresponding Hamiltonian in the gaussian aprroximation framework turned out to be:

$$H_{\mu_i,M_{ij}}(x_i) = -\sum_{i=1}^{S} \mu_i x_i - \frac{1}{2}\sum_{i\neq j} x_i M_{ij} x_j + o(\vec{x}^3)$$

In this way, the parameters can be estimated using the data as follows:

- first we need to estimate M:

$$(M^{-1})_{ij} = \mathrm{Cov}(x_i, x_j) = \langle (x_i - \langle x_i \rangle_{data})(x_j - \langle x_j \rangle_{data}) \rangle$$

- then we can evaluate μ_i:

$$\vec{\mu} = -M \cdot \langle \vec{x} \rangle_{data}$$

In order to consider only the most abundant species we kept those ones for which $\langle x_i \rangle_{data} - \sigma_{x,i} > 0$. In this way, in the abundance matrix there were only 52 different species left. To find matrix M we evaluated the *covariance matrix* of abundance matrix and then inverted it. After this, we set manually all diagonal coefficients to 0 in order to have no self-interactions terms. The results for the M matrix and μ parameters can be found in the section "Task 6" of the notebook (the matrix code has to be uncommented to save the matrix in a file if the reader wants to see it).

Results analysis and comparison with E-R graph

The left plot of Figure 5 is the histogram of the distribution of the M coefficients we found (M is a 52x52 matrix, as explained in the just above section). The distribution of the coefficients of the matrix M has a peak near 0. This means that the majority of the interacion terms are very low.

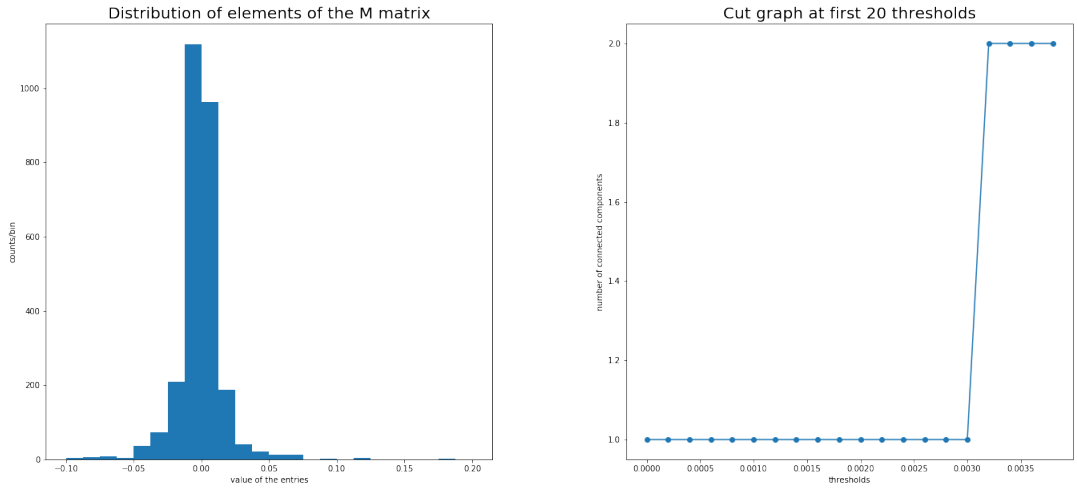

Figure 5: Figure on the left is the histogram of the M entries. The plot on the right represents the number of connected components in the graph generated with M depending on the value of the threshold.

Then, we selected the matrix coefficients holding only those values greater than an increasing threshold. This was done in order to find the value for which the number of connected components, in a undirected graph W^* (the matrix M is symmetric) represented by M, is greater than 1. This threshold turned out to be 0.003 (in our specific case; see notebook for details).

Then, we analyzed the graph built from the matrix thresholded with 0.003. Important quantities to analyze were: diameter of the graph, which represent the maximum distance between 2 nodes in a graph; clustering coefficient, which is a measure of the transitivity among nodes in the network (in other words, it counts the fraction of triangles with respect to all possible fulfilled links involving 3 nodes); the assortativity coefficient, which represents the trand of the nodes to be linked to nodes of same degree (it is bounded between -1 and 1); the betweenes centrality, which is a measure of the "importance" of the nodes (the more the possible shortest paths passing through a node, the greater its betweenes centrality is; it is bounded between 0 and 1). We found, for W^*:

- the diameter of the graph is: 3
- the clustering coefficient is: 0.71
- the assortativity coefficient is: -0.23
- the average betweenes centrality is: 0.008 (see "Task 7" of notebook for details)
- the degree distribution is:

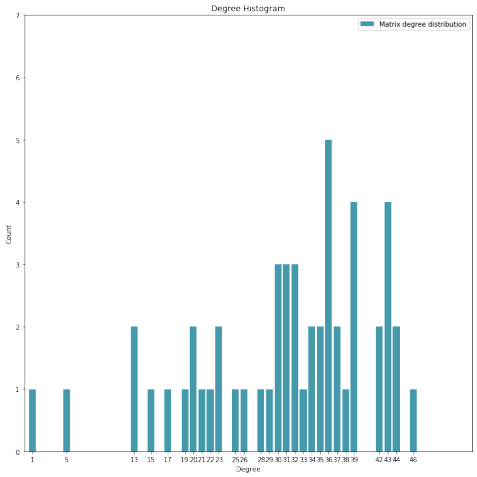

Figure 6: Degree distribution of W^*.

The next step was to compare obtained results with those of an ensemble of Erdos-Renyi graphs. This ensemble was generated fixing the probability of having an edge to be the same probability of W^*. We chose to create 100 E-R realizations. The results comparison was done for all parameters aforementioned (see Figure 7 and Figure 8):

- the average diameter: 2.0 ± 0.0 (all generated graphs have diameter = 2)
- the average clustering coefficient: 0.61 ± 0.01
- the average assortativity coefficient: -0.04 $\pm$ 0.02
- the betweeness centrality: 0.008
- the expected degree distribution for an E-R graph is:

$$P(k) = \binom{n-1}{k} \cdot p^k (1-p)^{n-1-k}$$

where: n is number of total possible edges, k is the number of present edges, p is the probability to have an edge, which is the same for every possible edge position. In our case, $n = 52$, $p \simeq 0.61$ (see notebook for details). The degree distributions of both data and simulated graphs (plotted in the same figure) are the following.

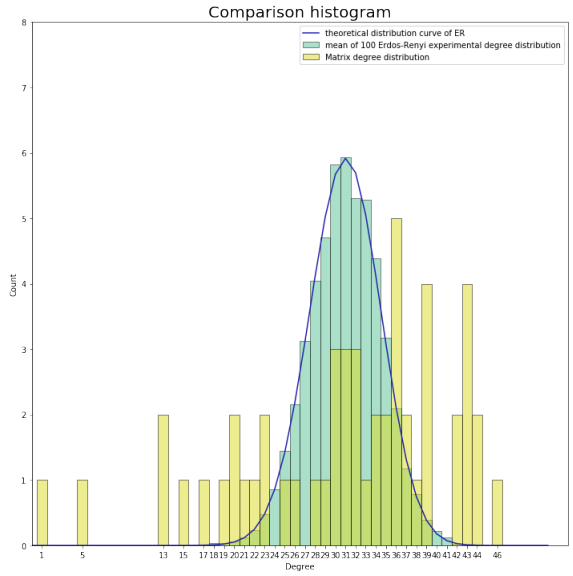

Figure 7: Comparison between the degree distributions of W^* and that of the E-R ensemble.

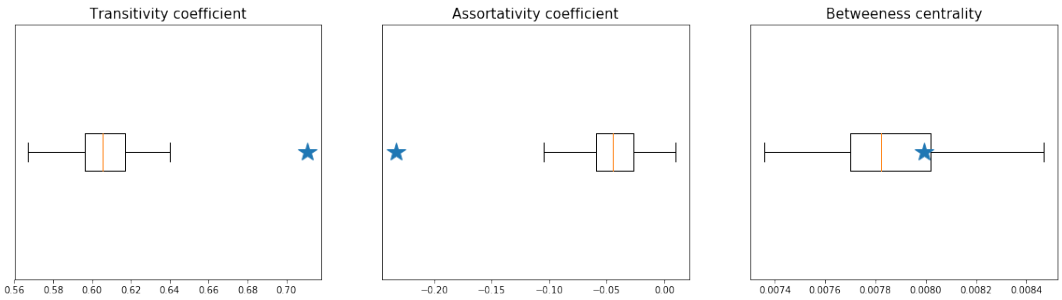

Figure 8: Comparison between characteristic of W^* (plotted as a star) and those of the E-R ensemble (boxplot).

Characterisitcs of W^* are different from that of a E-R graph (except for the *betweeness centrality*). This means that our tree configurations are not compatible with random configurations. Moreover, this was already noticed in section "Max Entropy Model 2", when we found the system is well-described by a critical behaviour and not by a paramagnetic behaviour as a random graph would suggest.