Problem set 8: Statistical dependence

BIO-369

Prof. Anne-Florence Bitbol
EPFL

1 Some examples

In this problem, we will consider various examples of sets of two random variables X and Y, and measure their statistical dependence with correlation and mutual information. While we usually focus on discrete random variables, here we will consider continuous ones, but then, in our estimates of mutual information, we will always discretize them into 10 bins, so that they effectively become discrete. Note that choosing the number of bins is an interesting problem that we will not tackle here – if there are too many bins for a given number of draws of X and Y, then frequencies are bad estimates of probabilities, and if there are too few then we lose resolution.

a) Consider that X has a normal (Gaussian) distribution with mean 0 and standard deviation 1, and that $X = Y$. Using Python, draw $N = 1000$ samples x of X. You can use the `numpy` function `random.normal` for this. What are the values y of Y for these draws? For each draw, plot the value y of Y versus that (x) of X, yielding a scatter plot of y versus x. Also plot the histogram of the values x (with 10 bins).

b) What do you expect for the correlation coefficient between X and Y? Estimate it from the samples drawn above, e.g. using `pearsonr` from `scipy.stats`.

c) Write a function that takes as input the number N of samples, the arrays of values of X and Y and the number B of bins, and that returns an array containing the estimates of the entropy $H(X)$ of X, of the entropy $H(Y)$ of Y, of the joint entropy $H(X, Y)$ of X and Y, of their mutual information $I(X;Y)$, and of the conditional entropies $H(X|Y)$ and $H(Y|X)$. Note that `histogram` (resp. `histogram2d`) from `numpy` can help you to estimate frequencies (resp. joint frequencies), and recall that frequencies should sum to one.

d) Using the function you just wrote and taking the previously drawn values of X and Y (from question a), compute all these entropies for these random variables, with $B = 10$ and $N = 1000$. How does $H(X)$ compare to $H(Y)$? How does $H(X, Y)$ compare to $H(X)$? Why? Comment on the values you find for $I(X;Y)$, $H(X|Y)$ and $H(Y|X)$.

e) Now consider that X still has a normal (Gaussian) distribution with mean 0 and standard deviation 1, but that Y is equal to X plus another normal (Gaussian) distribution with mean 0 and standard deviation 0.5. Make a scatter plot of y versus x. Also plot the histogram of the values x and another one of the values y (with 10 bins). Compute the correlation coefficient of X and Y and the various entropies using the function you wrote in question c, with $B = 10$ and $N = 1000$. Comment on the results obtained, and compare them to those from question d.

f) Perform the same analysis for X having a normal (Gaussian) distribution with mean 0 and standard deviation 1, and $Y = X^2$. Comment on the results obtained. In particular, compare the entropies of X and Y, and discuss the values of the correlation coefficient and of the mutual information.

g) Now consider that X still has a normal (Gaussian) distribution with mean 0 and standard deviation 1, but that Y is equal to X^2 plus another normal (Gaussian) distribution with mean 0 and standard deviation 1. Perform the same analysis again and comment on the results obtained, especially comparing to those of the previous question.

2 Coevolving sites in interacting proteins

Two-component systems constitute a major class of signaling pathways that enable prokaryotes to sense and respond to environment stimuli. In these pathways, a histidine kinase (HK) interacts specifically with its cognate response regulator (RR), which induces a cellular response to the initial stimulus. The HK-RR interaction is highly specific. Here we study how this specificity is encoded in HK-RR sequences, starting from a multiple sequence alignment of concatenated sequences of cognate HK-RR pairs. These concatenated sequences are each made of the sequence of an HK followed by that of its cognate RR. Denoting by i the indices of the columns of the multiple sequence alignment, i ranges from 0 to $L_{\mathrm{HK}} + L_{\mathrm{RR}}$ where L_{HK} is the number of amino acid sites (columns) in the HK and L_{RR} is the number of amino acid sites (columns) in the RR.

a) Using Python, load the sequence data in the file `Data9.npy`. It has already been converted to a `numpy` array of integers. You can also look at the raw sequence data file `Data9.fasta` and use it. How many possible states are there in this data? In these files, columns 0 to 63 correspond to the HK while columns 64 to 175 correspond to the RR. In other words, $L_{\mathrm{HK}} = 64$ and $L_{\mathrm{RR}} = 112$.

b) Produce three `Numpy arrays` that contain respectively the entropy of each column in the HK, the entropy of each column in the RR, and the joint entropy of each column i in the HK and each column j in the RR.

c) Produce a `Numpy array` structured as a matrix with L_{HK} rows and L_{RR} columns that contains the mutual information between each column i in the HK and each column j in the RR. Plot the resulting matrix in colorscale using `matshow` from `matplotlib.pyplot`. Also plot the histogram of the mutual informations. Comment on these results.

d) In Ref. [1], the pairs of columns with mutual information larger than $0.35/\ln(2)$ were considered as having high mutual information. Produce lists of the indices i and j of these pairs as well as of their mutual information. Note that in the end, j should range between 64 and 175.

e) Load the data in the file `Data9b.npy`: it contains a list of the indices i and j of the pairs of sites (each site corresponding to a column in the alignment) that are in contact at the interface between the HK and the RR in the three-dimensional structure of the HK-RR complex. Note that the file `Data9b.npy` was obtained from the PDB file `3DGE.pdb`, reporting the X-ray cristallography structure of an HK-RR complex. Construct a matrix containing 0 for all (i, j) that are not in contact, and 1 for the pairs of sites (i, j) that are in contact, and plot it in colorscale using `matshow` from `matplotlib.pyplot`. Compare it visually to the colorscale representation of the matrix of mutual informations you produced before, and comment.

f) Determine whether each of the pairs of sites with high mutual information corresponds to a pair of sites that are in contact in the three-dimensional structure of the HK-RR complex or not. Comment on the results obtained.

g) In Ref. [1], a kinase where sites number 17, 21, 22 and 39 (in our numbering of alignment columns) were mutated was produced. Specifically, the four amino acids of the HK EnvZ corresponding to these sites were substituted to match those observed in another HK, RtsB. Are these sites among those that are involved in the high-mutual information pairs in your analysis? In the experimental study, it was observed that the mutated EnvZ phosphorylates RtsA (the cognate RR of RtsB) much faster than OmpR (the one of EnvZ).

References

[1] J. M. Skerker, B. S. Perchuk, A. Siryaporn, E. A. Lubin, O. Ashenberg, M. Goulian, and M. T. Laub. Rewiring the specificity of two-component signal transduction systems. *Cell*, 133(6):1043–1054, Jun 2008.