**Data visualization using the distance correlation**

---

Visualization is one of the primary tools to achieve better understanding of a data set. Within the data visualization domain, different types of problems arise : how to display datasets with a large number of observations ? How to display uncertainty ? How to display non numerical data ? The focus of this project is how to display a high dimensional data set.

The problem of high dimensional data visualization can be seen as a problem of projection, either linear or non-linear, from the original space to a lower dimensional space (with two or three dimensions) where data can be visualized. The main technical problems behind it are the following: 1) to define a criterion to find the best projection; 2) to efficiently optimize this criterion with respect to the projected data points. Main existing criteria are based on retaining the variance of the data points (principal component analysis [1]) or the distances between data points (multidimensional scaling [2]) or neighborhood information (stochastic neighbor embedding [3]).

The aim of this project is to test a projection criterion based on the statistical dependence between the original data set and the projected data set. For doing so, we will use the empirical version of the distance correlation criterion proposed in [4].

The students will first briefly study and apply the previously cited techniques of dimensionality reduction to a publicly available high dimensional dataset, for example, the MNIST dataset, available at http://yann.lecun.com/exdb/mnist/. The tests can be done either in *python* with *scikit-learn* or in *matlab*. The students will then study and code the distance correlation criterion. A first attempt on maximizing it will be carried out by simple search on a predefined grid of projected values. If this approach gives comparable or superior performance to the existing methods, the students will be asked to test an efficient method for its computation [5] and to propose a more efficient method for its maximization.

In the case the students use *python* for coding, they will be asked to deliver a *python notebook* and its *pdf* version containing the code, a brief presentation of all methods and comments on the results. In the case the students use *matlab*, they will be asked to deliver the matlab code along with a separate report presenting the methods and the results.

---

References :


[1] Jolliffe, I. (2011). Principal component analysis. In *International encyclopedia of statistical science* (pp. 1094-1096). Springer, Berlin, Heidelberg.
[2] Cox, T. F., & Cox, M. A. (2000). *Multidimensional scaling*. CRC press.
[3] Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(Nov), 2579-2605.
[4] Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The annals of statistics*, *35*(6), 2769-2794.
[5] Huo, X., & Székely, G. J. (2016). Fast computing for distance covariance. *Technometrics*, *58*(4), 435-447.



Project supervisor : Rodrigo Cabral Farias
Assistant professor, Polytech MAM, I3S laboratory