STA 663: Project: 1st Deadline

**Project Members:** Mengshu Shao**,** Yelena Kernogitski

**Paper for Project:** On Spectral Clustering: Analysis and an algorithm (Andrew Y. Ng et al.)

**GitHub Repositories:**

https://github.com/yelenakernogitski/Spectral-Clustering-Project

https://github.com/mengshushao/Spectral-Clustering

**Project Outline:**

-**Background**

-**Methods and Algorithm**

-**Implementation in Python**

-Implementation using ideal case

-Implementation using general case

-**Implementation using a higher performance computing method**

-Implementation using ideal case

-Implementation using general case

-**Results and Comparison**

-**Conclusion**

**Abstract:**

Clustering is one of the building blocks in modern data analysis, and has been widely used in machine learning and pattern recognition. Two commonly used approaches are K-means and learning a mixture model using EM. However, these methods have some drawbacks, such as violation of the harsh simplifying assumption which the density of each cluster is Gaussian. One method that provides a possible solution in finding useful clusters is spectral clustering, which utilizes eigenvectors derived from the distance between points. The method performs dimensionality reduction in order to find clusters. In this report, First we implement a simple spectral clustering algorithm for clustering points in . Second we analyze how it works in “ideal” case in which the points are exactly far apart (i.e., when affinity matrix s strictly block diagonal), and in general case in which affinity’s off-diagonal blocks are non-zero. Then we test the algorithm by applying it to a number of challenging clustering problems. Further, we attempt to optimize the algorithm using within-Python options (such as vectorization) in addition to JIT and Cython wrapping functions. Finally, we compare the original method in Python and the latter, higher performance method by determining the efficiency of each method.