# Spectral Clustering Problem

Mohanna Karam Beigi S329631

**Abstract**

In machine learning and data analysis, spectral clustering is a prominent method for grouping datasets based on their inherent structures. This study explores the application of spectral clustering to two different datasets, Circle and Spiral, and compares its effectiveness against other clustering algorithms.

## 1 Introduction

Spectral clustering is a widely used unsupervised learning technique based on spectral graph theory. It efficiently clusters data points in high-dimensional spaces by projecting them onto a lower-dimensional space using the eigenvectors of a similarity graph's Laplacian matrix. This method has been successfully applied in fields such as image segmentation, community detection, and document clustering.

Clustering is a fundamental technique in unsupervised learning, where data points are grouped based on their similarities. Broadly, clustering approaches can be classified into two types: compactness-based clustering, where points in close proximity are grouped together (e.g., K-Means Clustering), and connectivity-based clustering, which clusters points based on their relationships rather than just proximity. Spectral clustering follows the connectivity-based approach, making it suitable for identifying complex structures that traditional methods struggle with.

In this study, we evaluate spectral clustering on two datasets, Circle.csv and Spiral.csv. The implementation follows a structured approach: constructing a similarity graph using a Gaussian kernel, computing the Laplacian matrix, extracting eigenvectors, and performing clustering. This process helps analyze spectral clustering's effectiveness in handling non-linearly separable data.

## 2 Datasets

This study utilizes two datasets: Circle.csv and Spiral.csv, both containing $N$ points in CSV format. Circle.csv consists of two columns representing the $x$ and $y$ coordinates of points distributed on two concentric circles with some noise. Spiral.csv contains three columns: the first two for the $x$ and $y$ coordinates, and

the third for the cluster index. This dataset consists of points forming three intertwined spirals.

Both datasets are used to generate a similarity graph, compute the Laplacian matrix, and perform spectral clustering. The results are compared with other clustering techniques, such as k-means.

# 3 Project Description

Spectral clustering consists of three key steps:

## 3.1 Constructing a Similarity Graph

- An undirected graph $G = (V, E)$ is created, where each vertex $v_i$ represents a data point.
- The adjacency matrix encodes similarities between vertices.
- An edge exists between $v_i$ and $v_j$ if their similarity $s_{ij}$ exceeds a threshold $\varepsilon$.
- The similarity graph is undirected and weighted, using the Gaussian similarity function:

$$s_{ij} = \exp\left(-\frac{||X_i - X_j||^2}{2\sigma^2}\right) \tag{1}$$

  where $\sigma$ controls the neighborhood size.

## 3.2 Computing the Laplacian Matrix and Eigenvectors

- The graph Laplacian matrix is computed as:

$$L = D - W \tag{2}$$

  where $D$ is the degree matrix and $W$ is the adjacency matrix.
- The eigenvalues and eigenvectors of $L$ are then computed:

$$Lv = \lambda v \tag{3}$$

- The smallest $k$ eigenvectors determine the number of clusters.

## 3.3 Applying K-Means on Eigenvectors

- A matrix $U \in \mathbb{R}^{N \times M}$ is constructed from the $M$ smallest eigenvectors.
- Each row of $U$ is clustered using the k-means algorithm.
- The original points are assigned to the same clusters as their corresponding rows in $U$, forming clusters $A_1, ..., A_M$.
- The clustered points are visualized using different colors.

# 4 Clustering Methods and Evaluation

Two clustering techniques are applied:

- K-Means Clustering
- Spectral Clustering

The optimal number of clusters is determined using the Elbow Method, which analyzes inertia (sum of squared distances from points to their closest cluster center) as a function of cluster count.

# 5 Conclusion

In summary, the Elbow Method suggests an optimal cluster count of three. In the Circle dataset, adjusting $k$ results in minimal variation for k-means. For spectral clustering, $K = 10$ and $K = 20$ yield similar results, whereas $K = 40$ introduces errors.

For the Spiral dataset, spectral clustering with $K = 10$ produces poorly grouped points, $K = 20$ shows minor misclassifications, and $K = 40$ results in well-separated clusters. Meanwhile, k-means fails to correctly cluster the Spiral dataset at any $k$, misclassifying most points.

These findings highlight spectral clustering's advantage in detecting complex structures, where traditional methods like k-means struggle.

# 6 References

[1] Wang, F., Zhang, C., Liu, C., Liu, Z., Zhang, X. (2020). "Spectral clustering: a review and perspectives." Frontiers of Computer Science, 14(5), 865-890.

[2] Tian, Y., Bai, X. (2021). "An Adaptive Agglomerative Clustering Framework for Multi-view Spectral Clustering." IEEE Transactions on Cybernetics, 51(2), 863-874.

[3] Baglama, J., Reichel, L. (2020). ARPACK software for large scale eigenvalue problems: Latest developments. Journal of Computational and Applied Mathematics, 375, 112594.

[4] Ankerst, Mihael, Markus M. Breunig, Hans-Peter Kriegel, and Jorg Sander. "OPTICS: ordering points to identify the clustering structure." ACM Sigmod Record 28, no. 2 (1999): 49-60. https://dl.acm.org/doi/10.1145/304181.304187

[5] Hinton, G. E., Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. Science, 313(5786), 504-507.

[6] Kohonen, T. (1990). The self-organizing map. Proceedings of the IEEE, 78(9), 1464-1480.