Spectral Clustering Problem

Computational linear algebra for large scale problems

Abedal Salam Al Ashi Abou Shoushe
Politecnico di Torino
s336648

Code Repository

Date: December 16, 2024

Contents

Abstract

This report explores the application of Spectral Clustering on two datasets, Circle and Spiral, to assess its ability to identify complex, non-convex structures.

1 Introduction

Spectral Clustering is a graph-based technique used to identify clusters in data with complex, non-convex structures. In this report, we apply Spectral Clustering to two datasets (Circle and Spiral) by constructing a similarity graph, computing the Laplacian matrix, and performing spectral decomposition to obtain the key eigenvectors. These eigenvectors are clustered using the K-means algorithm, and the results are compared with Gaussian Mixture Models Clustering. The report evaluates the performance of Spectral Clustering and visualizes the resulting clusters for analysis.

2 Datasets

The datasets used in this report consist of two sets of points, *Circle* and *Spiral*, designed to test the performance of clustering methods on non-convex structures. The Circle dataset contains two columns representing the x- and y-coordinates of points forming concentric circular clusters. The Spiral dataset includes two columns for the x- and y-coordinates of points arranged in a spiral structure, along with a third column indicating the correct cluster index. These datasets are chosen to evaluate the ability of Spectral Clustering to handle complex geometries that pose challenges for traditional clustering algorithms.

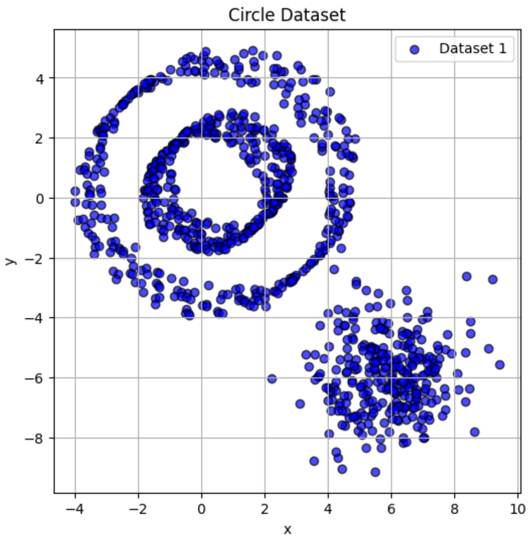

Figure 1: Circle Dataset

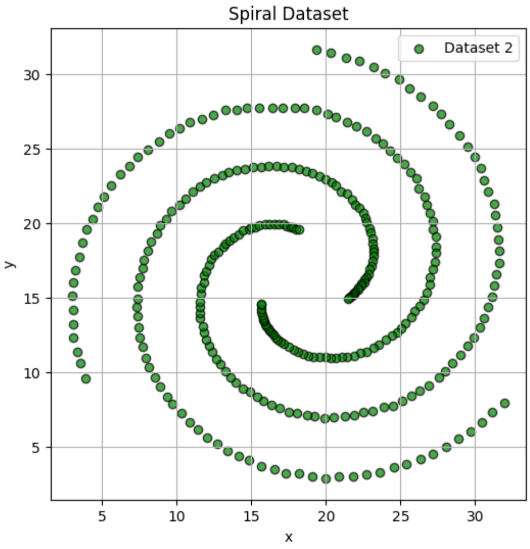

Figure 2: Spiral Dataset

3 Methodology

3.1 Similarity Graph

The similarity graph [2] is a fundamental step in Spectral Clustering, where data points are represented as vertices, and edges between vertices encode the similarity between points. The similarity between two points X_i and X_j is determined using a Gaussian kernel function, which ensures that closer points have higher similarity values. The similarity function is defined as:

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

where $\|X_i - X_j\|$ is the Euclidean distance between points X_i and X_j, and σ is a scaling parameter that controls the influence of distance on the similarity. This function results in a weighted, undirected graph that captures the local structure of the data, which is essential for the clustering process.

3.2 Degree Matrix

The degree matrix is a diagonal matrix that captures the connectivity of each vertex in the similarity graph. Specifically, the degree of a vertex v_i is defined as the sum of the weights of all edges connected to it. The degree matrix D is computed from the weighted adjacency matrix W and is defined as:

$$D_{ii} = \sum_{j=1}^{N} W_{ij}, \quad \text{and} \quad D_{ij} = 0 \text{ for } i \neq j, \tag{2}$$

where W_{ij} represents the weight of the edge between vertices v_i and v_j, and N is the total number of vertices in the graph. The degree matrix D plays a crucial role in constructing the Laplacian matrix, which is central to the Spectral Clustering process.

3.3 Laplacian Matrix

The Laplacian matrix is a crucial component in Spectral Clustering, as it encodes the structure of the graph and is used for spectral decomposition. The Laplacian matrix L is defined as:

$$L = D - W, \tag{3}$$

3.4 Sparse Format Storage

To efficiently store and manipulate large matrices, the adjacency matrix W, the degree matrix D, and the Laplacian matrix L are represented in a sparse format. Sparse storage reduces memory usage and computational overhead by storing only the non-zero elements of these matrices. Specifically, the Compressed Sparse Row (CSR) format is used due to its efficiency in matrix operations, such as summation and multiplication, which are critical for Spectral Clustering.

3.5 Connected Components

Connected components of a graph are subgraphs in which all vertices are interconnected, and no connections exist outside the subgraph. In Spectral Clustering, the number of connected components can be determined by analyzing the Laplacian matrix L. Specifically, the multiplicity of the eigenvalue $\lambda = 0$ corresponds to the number of connected components in the graph.

3.6 Number of Clusters

The number of clusters in Spectral Clustering [1] is determined by analyzing the eigenvalues of the Laplacian matrix L. The eigenvalues provide insight into the structure of the graph, and gaps between consecutive eigenvalues indicate potential clustering boundaries.

The process involves:

- Computing the smallest k eigenvalues of L using efficient methods such as the Arnoldi algorithm [4].
- Plotting the eigenvalues to visualize the gaps between consecutive values.
- Identifying the largest gap, which corresponds to the most significant division in the data and determines the optimal number of clusters.

The largest gap in the eigenvalues reflects the most natural partitioning of the data, and its position determines the number of clusters M.

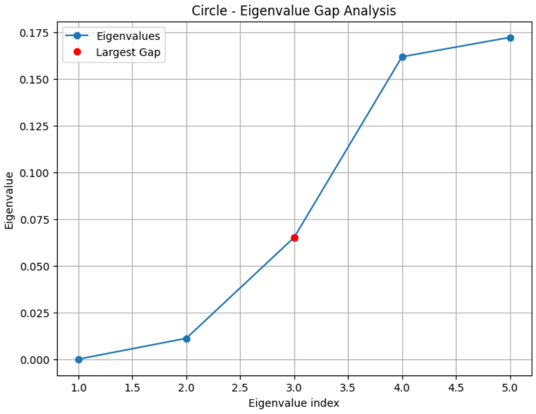

Figure 3: Circle Eigenvalues

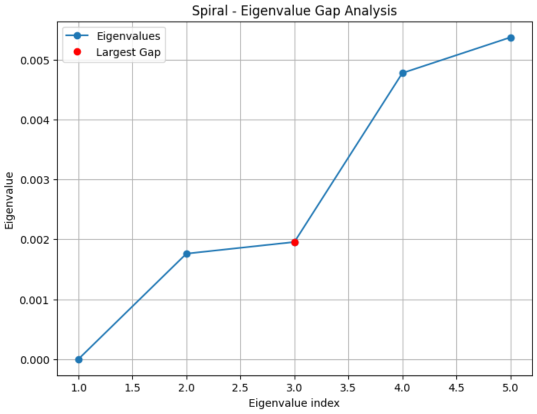

Figure 4: Spiral Eigenvalues

3.7 Applying the K-means Algorithm

After determining the number of clusters M and computing the matrix U containing the eigenvectors of the Laplacian matrix, clustering is performed in the reduced space. Each data point $y_i \in \mathbb{R}^M$ corresponds to the i-th row of U, where M is the number of clusters.

The K-means algorithm is applied to the points $\{y_1, \ldots, y_N\}$ in the M-dimensional space to assign each point to one of the M clusters $C_1, \ldots, C_M$. The algorithm minimizes the within-cluster variance, ensuring that points in the same cluster are as close as possible to their cluster center. This process generates cluster labels for all data points, which are used to map the clusters back to the original data space.

4 Results and Analysis

In this section, we present the clustering results for the *Circle* and *Spiral* datasets. Spectral Clustering is applied in combination with the K-means algorithm, whereas Gaussian Mixture Models (GMM) are applied directly to the original data.

4.1 Spectral Clustering Results

The clustering results using the K-means algorithm applied to the eigenvector representation obtained from Spectral Clustering are shown below:

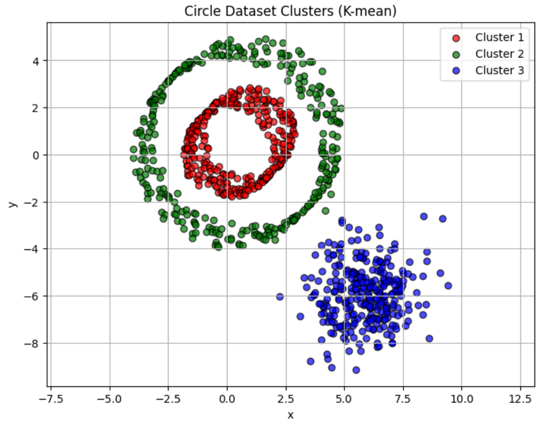

Figure 5: Spectral Clustering Results on the Circle Dataset

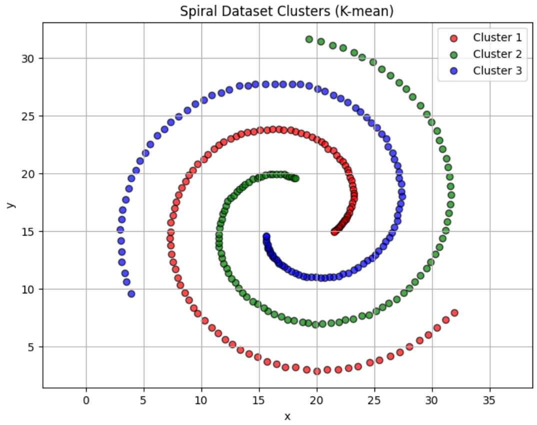

Figure 6: Spectral Clustering Results on the Spiral Dataset

The *Circle* dataset (Figure 5) shows that K-means combined with spectral clustering successfully separates the concentric circular clusters when applied in the transformed spectral space. Similarly, for the *Spiral* dataset (Figure 6), the K-means combined with spectral clustering captures the spiral-shaped clusters effectively, which would otherwise be challenging in the original space.

To evaluate the clustering performance quantitatively, we calculate the **Adjusted Rand Index (ARI)** between the ground-truth labels and the clustering results. For the *Spiral* dataset, the ARI score is:

$$\text{Adjusted Rand Index (ARI)} = 1.0 \tag{4}$$

The ARI score of 1.0 indicates perfect agreement between the predicted clusters and the ground-truth labels, demonstrating the effectiveness of Spectral Clustering combined with K-means in handling complex, non-convex cluster structures.

4.2 Gaussian Mixture Model (GMM) Results

Gaussian Mixture Models (GMM) [3] assume that the data is generated from a mixture of Gaussian distributions. In this subsection, we apply GMM directly to the original *Circle* and *Spiral* datasets to evaluate its clustering performance.

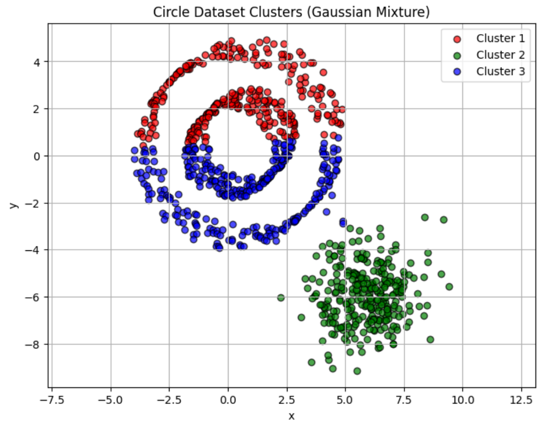

Figure 7: GMM Clustering Results on the Circle Dataset

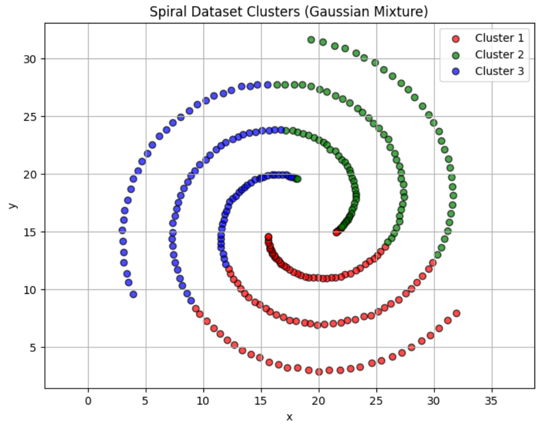

Figure 8: GMM Clustering Results on the Spiral Dataset

GMM clustering was able to identify some differences in the data, but it was unable to differentiate the clusters effectively, mixing some clusters together in both datasets. This behavior is evident in the *Circle* dataset (Figure 7), where GMM separates clusters but introduces minor overlaps. For the *Spiral* dataset (Figure 8), GMM struggles significantly due to the assumption of Gaussian distributions, failing to separate the spiral-shaped clusters.

To quantify the clustering performance, we calculate the **Adjusted Rand Index (ARI)** between the ground-truth labels and the GMM results for the *Spiral* dataset. The ARI score is:

$$\text{Adjusted Rand Index (ARI)} = -0.0051 \tag{5}$$

The low ARI score further confirms GMM's limitations in identifying non-convex cluster structures.

5 Discussion

This report evaluates the performance of Spectral Clustering combined with K-means and Gaussian Mixture Models (GMM) on the *Circle* and *Spiral* datasets, which are known for their non-convex cluster structures.

The results demonstrate the following:

- **Spectral Clustering with K-means:** Spectral Clustering outperformed traditional methods, successfully identifying the clusters in both datasets. By transforming the data into a lower-dimensional space using the eigenvectors of the Laplacian matrix, K-means was able to separate the non-linear structures effectively. This is evident from the perfect clustering achieved for the *Spiral* dataset, with an Adjusted Rand Index (ARI) of 1.0, reflecting complete agreement with the ground truth.

- **Gaussian Mixture Model (GMM):** While GMM was able to identify some differences in the data, it struggled with the non-convex shapes of the clusters. For the *Circle* dataset, GMM managed to separate the clusters but introduced minor overlaps. However, for the *Spiral* dataset, GMM performed poorly, as it failed to account for the complex spiral structure due to its assumption of Gaussian-shaped clusters. The ARI score of -0.0051 confirms that the clustering result was nearly random.

Overall, the comparison highlights the strength of Spectral Clustering in handling datasets with non-convex and complex geometries. The ability of Spectral Clustering to map the data into a transformed space where clusters become linearly separable is a key advantage over GMM, which assumes more rigid cluster shapes.

The findings suggest that Spectral Clustering, when combined with K-means, is a highly effective approach for clustering tasks where traditional methods fail, particularly in datasets with intricate structures.

References

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

[2] Chung, F.R.K. ”Spectral Graph Theory.” CBMS Regional Conference Series in Mathematics, vol. 92, 1997.

[3] Bishop, C. M. ”Pattern Recognition and Machine Learning.” Springer, 2006.

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