Spectral clustering

Brian Facundo Morales Condorpocco
Andrea Spinnicchia
Politecnico di Torino

Abstract

This report explores the use of spectral clustering on two datasets, focusing on the impact of similarity graph parameters (k) and threshold selection for eigenvalue analysis. Results highlight the importance of adaptive thresholding and parameter tuning to achieve meaningful clustering, particularly in complex geometrical datasets. The study emphasizes the role of graphical interpretation in refining clustering outcomes.

1 Introduction

Clustering is a fundamental problem in data analysis and machine learning, where the objective is to partition a dataset into meaningful groups based on similarity metrics. This report investigates spectral clustering applied to two distinct datasets, Circle and Spiral, provided in various formats (.mat and .csv).

The Circle dataset consists of two-dimensional points characterized by their x and y coordinates, while the Spiral dataset includes an additional column indicating the correct cluster index for each point.

Spectral clustering is a versatile technique used for partitioning datasets into meaningful clusters based on their underlying structure. Unlike traditional clustering methods, such as k-means, which rely on geometric distances, spectral clustering leverages the relationships between data points represented as a graph. This approach is particularly effective for datasets with non-linear or complex geometrical patterns, where traditional methods often struggle. The foundation of spectral clustering lies in the construction of a similarity graph G = (V, E). The graph is then analyzed using spectral properties of the graph Laplacian matrix, which encapsulates the connectivity and structure of the graph.

2 Application

2.1 Similarity graph

Central to this analysis is the construction of the Similarity graph, G = (V, E), where V denotes a non-empty set of vertices (data points) and E denotes the set of edges connecting pairs of vertices with a similarity measure s_{ij}. The graph is undirected and weighted, with weights determined by similarity s_{ij}, defined as the Gaussian similarity function:

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

where the parameter σ controls the width of the neighborhood.

The clustering task can therefore be reinterpreted through the similarity graph: the objective is to divide the graph into groups such that the edges connecting the vertices of different groups have very low weights (indicating dissimilarity between the points in separate clusters), while the edges within a group have high weights (indicating strong similarity between points in the same cluster) [1].

The weighted adjacency matrix of the graph is the matrix $W = (w_{ij})\ i, j = 1, ..., n$. If $w_{ij} = 0$ this means that the vertices v_i and v_j are not connected by an edge. As G is undirected, we require $w_{ij} = w_{ji}$.

For the construction of the matrix W, we construct before the k-nearest neighborhood similarity graph. For this reason, we implemented several values of k=10,20,40 and σ equal to 0.

2.2 Graph Laplacian

In a weighted graph, the degree of a vertex $v_i \in V$ is defined as the sum of the weights of all edges incident to it:

$$d_i = \sum_{j=1}^{n} w_{ij}$$

Note that, in fact, this sum only runs over all vertices adjacent to v_i, as for all other vertices v_j the weight w_{ij} is 0.

The degree matrix D is a diagonal matrix where each diagonal entry corresponds to the degree of a vertex.

$$D = diag(d_1, d_2, ..., d_n)$$

Once matrices D and W have been obtained it is possible to define the unnormalized graph Laplacian matrix as:

$$L = D - W$$

To reduce computational costs and improve efficiency, the Laplacian matrix was then converted into a dense format. This transformation ensures efficient storage and operations, especially when handling large datasets. From the visualization of the sparse matrix with k=10 in figures 1-2, it is evident that the Spiral dataset already exhibits a clear division into three regions.

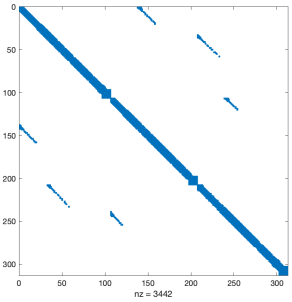

Figure 1: Laplacian Matrix for Spiral dataset

In contrast, the Circle dataset, which contains a larger and more dispersed set of points, makes it more challenging to identify distinct regions.

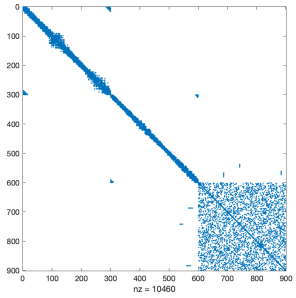

Figure 2: Laplacian Matrix for Circle dataset

2.3 Compute Eigenvalues

To proceed with clustering, the eigenvalues and eigenvectors of the Laplacian matrix were computed to analyze its spectral properties. Specifically, the eigenspace corresponding to the eigenvalue 0 was examined to determine the number of clusters(M) present in the data. A tolerance level was chosen for each different k, as eigenvalues below these threshold showed significant separation from 0. The computed eigenvalues are displayed in Figures 3-4.

The computation of eigenvalues and eigenvectors was performed using MATLAB's eigs() function. This function efficiently calculates the smallest t eigenvalues (with t = 10 in our case) using the Inverse Power Method. This approach provided accurate results for identifying the spectral characteristics of the graph.

Following the spectral analysis, the eigenvectors corresponding to the M smallest eigenvalues were utilized to construct the embedding matrix U. The rows of U were then clustered using the k-means algorithm, resulting in the identification of k distinct clusters. Each data point in the original dataset was assigned to a cluster based on this spectral embedding. The clustering results are represented in Figure 4, demonstrating the segmentation of the datasets into meaningful groups.

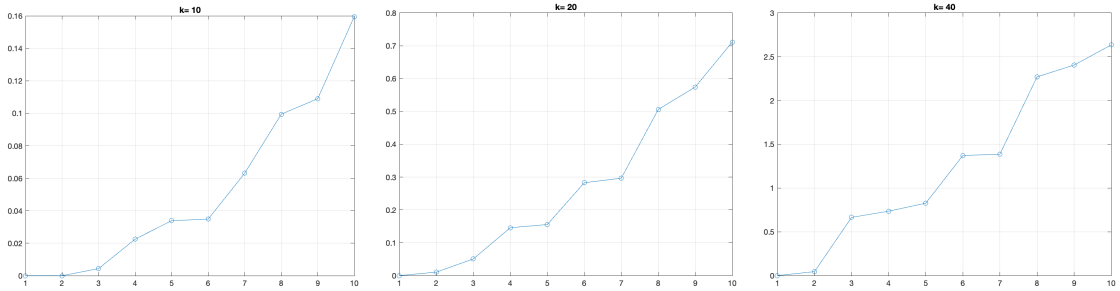

Figure 3: Eigenvalues of Circle dataset

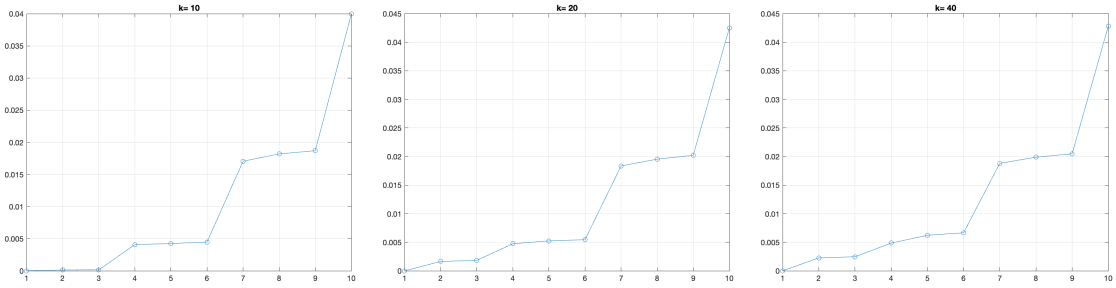

Figure 4: Eigenvalues of Spiral dataset

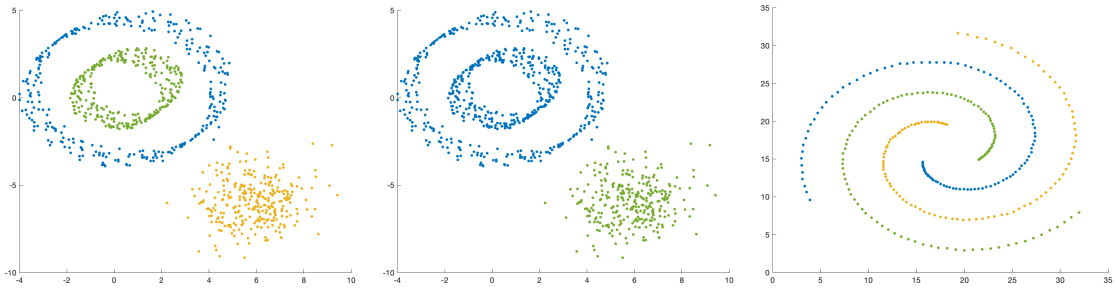

Figure 5: Cluster of Circle dataset with k=10,20(left), k=40(center). Cluster of Spiral dataset for k=10,20,40(right)

3 Result

To determine the appropriate number of clusters (M), we analyzed the eigenvalues of the Laplacian matrix, focusing particularly on the multiplicity of the eigenvalue 0. Given that the Inverse Power Method is an approximate numerical technique, it was necessary to define a threshold value to identify eigenvalues close to 0. The key findings are as follows:

- k=10: A threshold of 10^{-3} was selected, resulting in the identification of three clusters for both datasets.
- k=20: As shown in Figure 4, there is a significant gap between the third and fourth eigenvalues. Therefore choosing a threshold between this gap is possible obtaining three clusters.
- k=40: Although in the circle dataset the situation is similar at k=20; in the Spiral dataset there is a greater difference between the second and third eigenvalue, therefore in this case we opted to use two clusters.

This analysis highlights the critical role of graphical interpretation in selecting thresholds. While a consistent threshold can be reliable in many cases, for higher k-values, it may be more practical to adapt the threshold dynamically to observe potential outcomes.

4 Discussion

The results obtained using spectral clustering proved highly effective, as shown by the visualizations. In most cases, the method successfully identified the correct clusters. However, it is important to highlight that increasing the value of k in constructing the similarity graph introduced challenges in distinguishing between clusters. Specifically, the eigenvalues closest to zero tended to shift further away from zero as k increased, necessitating a reduction in the number of clusters for accurate segmentation. In extreme cases, such as k=40, , the method even produced divisions that did not align with the true structure of the data.

The results from spectral clustering were also compared with those from alternative clustering methods, such as k-means and single linkage, directly applied to the raw data. The comparative analysis revealed the following:

- The **k-means** approach was entirely ineffective, as it generated clusters based on the spatial regions where the points are located, rather than capturing the underlying spiral structure.

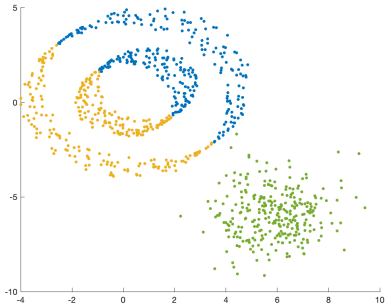

Figure 6: kmeans applied to Circle Dataset

- The **single linkage** method proved highly effective for the Spiral dataset, successfully identifying meaningful clusters. However, it fails entirely for the Circle dataset, as it cannot partition the data into any coherent groups.

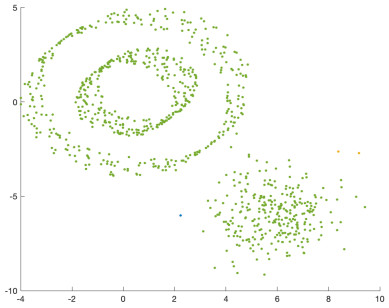

Figure 7: kmeans applied to Circle Dataset

These findings underscore the advantages of spectral clustering for datasets with complex geometrical patterns. While k-means is effective for simpler problems, its inability to detect non-linear structures demonstrates the superiority of graph-based methods like spectral clustering in such scenarios. Similarly, the single linkage method demonstrated impressive effectiveness only in certain geometries, illustrating its limitations in high density structure.

5 Bibliography

HEIN, Matthias; BÜHLER, Thomas. An inverse power method for nonlinear eigenproblems with applications in 1-spectral clustering and sparse PCA. Advances in neural information processing systems, 2010, 23.