1. Research about the Spectral Clustering method, and answer the following questions:

a. ¿In which cases might it be more useful to apply?

Spectral clustering is a good option to turn to if you have a dataset without an obvious outcome variable to predict. Spectral clustering can be applied to datasets without an obvious outcome variable to identify patterns and similarities that exist across different observations.

Advantages of spectral clustering

Applicable for high dimensional datasets. One of the main advantages that spectral clustering has over other clustering algorithms is that it can be used on high-dimensional datasets with many features. This is a relatively rare quality that applies primarily to graph-based clustering methods like spectral clustering.

Not strong assumptions about cluster shape. Spectral clustering does not make strong assumptions about the shape of the clusters in the data. That means that it is appropriate to use spectral clustering even when you suspect the clusters in your data may be irregularly shaped.

Can sometimes handle categorical variables. Some implementations of spectral clustering can handle cases where you have mixed data types, such as cases where you have categorical variables in your data. This is in part because spectral clustering uses similarity metrics rather than distance metrics to determine which points have more in common.

Disadvantages of spectral clustering

Relatively slow. One disadvantage of spectral clustering is that it is relatively slow compared to other clustering algorithms like k-means clustering. If you have a dataset with many, many data points then you may be better off using a faster algorithm.

Less common. Another disadvantage of hierarchical clustering is that it is less popular and well-studied than other clustering algorithms like k-means and hierarchical clustering. That means that it will not be as easy for collaborators to provide advice on or assist on projects that use spectral clustering.

Not intuitive or easy to explain. Many clustering algorithms are intuitive and easy to explain to relatively non-technical coworkers. This is not the case with spectral clustering, which can be difficult to fully understand if you do not have a strong math background.

Sensitive to seed. Traditional spectral clustering algorithms include a step where the k-means algorithm is applied. That means that like k-means clustering, spectral clustering is sensitive to the choice of seed and initialization conditions that are used. The clustering results may change when the algorithm is run multiple times.

Need to select the number of clusters. As with many other clustering algorithms, spectral clustering requires you to select the number of clusters that should be used for your dataset. This can be difficult to do if you do not have strong intuition about the true number of clusters in the data.

b. What are the mathematical fundamentals of it?

Spectral clustering is an EDA technique that reduces complex multidimensional datasets into clusters of similar data in rarer dimensions. The main outline is to cluster the all spectrum of unorganized data points into multiple groups based upon their uniqueness 'Spectral Clustering uses the connectivity approach to clustering', wherein communities of nodes (i.e. data points) that are connected or immediately next to each other are identified in a graph. The nodes are then mapped to a low-dimensional space that can be easily segregated to form clusters. Spectral Clustering uses information from the eigenvalues (spectrum) of special matrices (i.e. Affinity Matrix, Degree Matrix and Laplacian Matrix) derived from the graph or the data set.

c. What is the algorithm to compute it?

Spectral Clustering is a growing clustering algorithm which has performed better than many traditional clustering algorithms in many cases. It treats each data point as a graph node and thus transforms the clustering problem into a graph-partitioning problem. A typical implementation consists of three fundamental steps:

Building the Similarity Graph: This step builds the Similarity Graph in the form of an adjacency matrix which is represented by A. The adjacency matrix can be built in the following manners:

> **Epsilon-neighbourhood Graph**: A parameter epsilon is fixed beforehand. Then, each point is connected to all the points which lie in it's epsilon-radius. If all the distances between any two points are similar in scale then typically the weights of the edges i.e. the distance between the two points are not stored since they do not provide any additional information. Thus, in this case, the graph built is an undirected and unweighted graph.
>
> **K-Nearest Neighbours**: A parameter k is fixed beforehand. Then, for two vertices u and v, an edge is directed from u to v only if v is among the k-nearest neighbors of u. Note that this leads to the formation of a weighted and directed graph because it is not always the case that for each u having v as one of the k-nearest neighbor,

it will be the same case for v having u among its k-nearest neighbors. To make this graph undirected, one of the following approaches is followed:

- Direct an edge from u to v and from v to u if either v is among the k-nearest neighbors of u OR u is among the k-nearest neighbors of v.
- Direct an edge from u to v and from v to u if v is among the k-nearest neighbors of u AND u is among the k-nearest neighbors of v.
- Fully-Connected Graph: To build this graph, each point is connected with an undirected edge weighted by the distance between the two points to every other point. Since this approach is used to model the local neighborhood relationships thus typically the Gaussian similarity metric is used to calculate the distance.

Projecting the data onto a lower Dimensional Space: This step is done to account for the possibility that members of the same cluster may be far away in the given dimensional space. Thus, the dimensional space is reduced so that those points are closer in the reduced dimensional space and thus can be clustered together by a traditional clustering algorithm. It is done by computing the Graph Laplacian Matrix. To compute it though first, the degree of a node needs to be defined. The degree of the ith node is given by

$$d_i = \sum_{j=1|(i,j)\epsilon E}^{n} w_{ij}$$

Note that

$$w_{ij}$$

is the edge between the nodes i and j as defined in the adjacency matrix above. The degree matrix is defined as follows:

$$D_{ij} = \begin{cases} d_i, i = j \\ 0, i \neq j \end{cases}$$

Thus, the Graph Laplacian Matrix is defined as:

$$L = D - A$$

This Matrix is then normalized for mathematical efficiency. To reduce the dimensions, first, the eigenvalues and the respective eigenvectors are calculated. If the number of clusters is k then the first eigenvalues and their eigen-vectors are taken and stacked into a matrix such that the eigen-vectors are the columns.

Clustering the Data: This process mainly involves clustering the reduced data by using any traditional clustering technique – typically K-Means Clustering. First, each node is assigned a row of the normalized of the Graph Laplacian Matrix. Then this data is clustered using any traditional technique. To transform the clustering result, the node identifier is retained.

d. Does it hold any relation to some of the concepts previously mentioned in class? Which, and how?

Spectral clustering has relation with:

- Eigenvalues: to calcule the spectrum and process with the projections in the lower dimensional space “If the number of clusters is k then the first eigenvalues and their eigen-vectors are taken and stacked into a matrix such that the eigen-vectors are the columns.”

- Dimensional reduction: The spectral clustering use dimensional reduction for the possibility that members of the same cluster may be far away in the given dimensional space.

References:

- https://www.analyticsvidhya.com/blog/2021/05/what-why-and-how-of-spectral-clustering/
- https://www.geeksforgeeks.org/ml-spectral-clustering/
- https://www.mygreatlearning.com/blog/introduction-to-spectral-clustering/
- https://crunchingthedata.com/when-to-use-spectral-clustering/#:~:text=Spectral%20clustering%20is%20a%20good,that%20exist%20across%20different%20observations.