SF2526 - Homework 2

Ville Sebastian Olsson, Silpa Soni Nallacheruvu

February 17, 2025

Problem 1

Data matrix:

$$X = \begin{bmatrix} x_1^\top \\ x_2^\top \\ x_3^\top \\ x_4^\top \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1.5 & 2 & 0 \\ -0.5 & 0 & -3 \\ 0 & -0.5 & -1 \end{bmatrix}$$

a)

The rows of the matrix

$$R = \begin{bmatrix} 3 & 3 & 3 \\ -1 & -1 & 0 \end{bmatrix}$$

contain our starting values for the two centroids, $r_1 = (3, 3, 3)$ and $r_2 = (-1, -1, 0)$.

The Euclidean distance between each of the data points of X and each of the centroids are as follows:

$$\begin{aligned} \|x_1 - r_1\|_2 &\approx 3.4641 \\ \|x_1 - r_2\|_2 &\approx 3 \\ \|x_2 - r_1\|_2 &\approx 3.5000 \\ \|x_2 - r_2\|_2 &\approx 3.9051 \\ \|x_3 - r_1\|_2 &\approx 7.5664 \\ \|x_3 - r_2\|_2 &\approx 3.2016 \\ \|x_4 - r_1\|_2 &\approx 6.1033 \\ \|x_4 - r_2\|_2 &\approx 1.5000 \end{aligned}$$

The cluster assignment of each data point in the X data matrix is represented by indicator vectors for each cluster. Let A and B denote two disjoint sets of data points which correspond to the first and second row of R, respectively. The indicator vectors can then be defined as:

$$\mathbb{1}_A = \begin{bmatrix} [\mathcal{I}_1 = 1] \\ [\mathcal{I}_2 = 1] \\ [\mathcal{I}_3 = 1] \\ [\mathcal{I}_4 = 1] \end{bmatrix} \text{ and } \mathbb{1}_B = \begin{bmatrix} [\mathcal{I}_1 = 2] \\ [\mathcal{I}_2 = 2] \\ [\mathcal{I}_3 = 2] \\ [\mathcal{I}_4 = 2] \end{bmatrix}$$

where $[\cdot]$ denotes the Iverson bracket, and

$$\mathcal{I}_i = \operatorname*{argmin}_j \|x_i - r_j\|_2$$

Based on the computed distances above, the indicator vectors are then:

$$\mathbb{1}_A = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \text{ and } \mathbb{1}_B = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

The first row of the new R matrix is defined by the mean of data point 2 and the second row of R is defined by the mean of data points 1, 3, 4. The updated value of R is then:

$$R_1 = \begin{bmatrix} (1.5)/1 & (2)/1 & (0)/1 \\ (1-0.5)/3 & (1-0.5)/3 & (1-3-1)/3 \end{bmatrix}$$

$$\Rightarrow R_1 \approx \begin{bmatrix} 1.5 & 2 & 0 \\ 0.1667 & 0.1667 & -1 \end{bmatrix}$$

Repeat the process again:

$$\begin{aligned} \|x_1 - r_1\|_2 &\approx 1.5000 \\ \|x_1 - r_2\|_2 &\approx 2.3214 \\ \|x_2 - r_1\|_2 &\approx 0 \\ \|x_2 - r_2\|_2 &\approx 2.4777 \\ \|x_3 - r_1\|_2 &\approx 4.1231 \\ \|x_3 - r_2\|_2 &\approx 2.1148 \\ \|x_4 - r_1\|_2 &\approx 3.0822 \\ \|x_4 - r_2\|_2 &\approx 0.6872 \end{aligned}$$

The new indicator vectors are then:

$$\mathbb{1}_A = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \text{ and } \mathbb{1}_B = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

The first row of the new R matrix is defined by the mean of data points 1, 2 and the second row of R is defined by the mean of data points 3, 4. The updated R matrix is then:

$$R_2 = \begin{bmatrix} (1+1.5)/2 & (1+2)/2 & (1+0)/2 \\ (-0.5)/2 & (-0.5)/2 & (-3-1)/2 \end{bmatrix}$$

$$\Rightarrow R_2 = \begin{bmatrix} 1.25 & 1.5 & 0.5 \\ -0.25 & -0.25 & -2 \end{bmatrix}$$

Repeat the process again:

$$\begin{aligned}
\|x_1 - r_1\|_2 &\approx 0.7500\\
\|x_1 - r_2\|_2 &\approx 3.4821\\
\|x_2 - r_1\|_2 &\approx 0.7500\\
\|x_2 - r_2\|_2 &\approx 3.4821\\
\|x_3 - r_1\|_2 &\approx 4.1908\\
\|x_3 - r_2\|_2 &\approx 1.0607\\
\|x_4 - r_1\|_2 &\approx 2.7951\\
\|x_4 - r_2\|_2 &\approx 1.0607
\end{aligned}$$

The new indicator vectors are then:

$$\mathbb{1}_A = \begin{bmatrix}1\\1\\0\\0\end{bmatrix} \text{ and } \mathbb{1}_B = \begin{bmatrix}0\\0\\1\\1\end{bmatrix}$$

In this iteration, we see that data points 1, 2 belong to cluster A and data points 3 and 4 belong to cluster B again. The classification is unchanged; hence, we terminate the iteration process since no further modifications would be made to the R matrix.

Finally:

$$R_2 = \begin{bmatrix}1.25 & 1.5 & 0.5\\ -0.25 & -0.25 & -2\end{bmatrix}$$

$$\begin{aligned}
A &= \{x_1, x_2\} = \{(1,1,1), (1.5,2,0)\}\\
B &= \{x_3, x_4\} = \{(-0.5,0,-3), (0,-0.5,-1)\}
\end{aligned}$$

b)

Starting centroids:

$$R = \begin{bmatrix}-0.5 & 0 & -3\\ -1 & -1 & 1\end{bmatrix}$$

Distances:

$$\begin{aligned}
\|x_1 - r_1\|_2 &\approx 4.3875\\
\|x_1 - r_2\|_2 &\approx 2.8284\\
\|x_2 - r_1\|_2 &\approx 4.1231\\
\|x_2 - r_2\|_2 &\approx 4.0311\\
\|x_3 - r_1\|_2 &\approx 0.0000\\
\|x_3 - r_2\|_2 &\approx 4.1533\\
\|x_4 - r_1\|_2 &\approx 2.1213\\
\|x_4 - r_2\|_2 &\approx 2.2913
\end{aligned}$$

Indicator vectors:

$$\mathbb{1}_A = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} \text{ and } \mathbb{1}_B = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

Updated centroids:

$$R_1 = \begin{bmatrix} (-0.5+0)/2 & (0-0.5)/2 & (-3-1)/2 \\ (1+1.5)/2 & (1+2)/2 & (1+0)/2 \end{bmatrix} = \begin{bmatrix} -0.25 & -0.25 & -2 \\ 1.25 & 1.5 & 0.5 \end{bmatrix}$$

New distances:

$$\begin{aligned} \|x_1 - r_1\|_2 &\approx 3.4821 \\ \|x_1 - r_2\|_2 &\approx 0.7500 \\ \|x_2 - r_1\|_2 &\approx 3.4821 \\ \|x_2 - r_2\|_2 &\approx 0.7500 \\ \|x_3 - r_1\|_2 &\approx 1.0607 \\ \|x_3 - r_2\|_2 &\approx 4.1908 \\ \|x_4 - r_1\|_2 &\approx 1.0607 \\ \|x_4 - r_2\|_2 &\approx 2.7951 \end{aligned}$$

Updated indicator vectors:

$$\mathbb{1}_A = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} \text{ and } \mathbb{1}_B = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

The indicator vectors did not change, so we stop iterating here. The final centroids and clusters are:

$$R_1 = \begin{bmatrix} -0.25 & -0.25 & -2 \\ 1.25 & 1.5 & 0.5 \end{bmatrix}$$

$$\begin{aligned} A &= \{x_3, x_4\} = \{(-0.5, 0, -3), (0, -0.5, -1)\} \\ B &= \{x_1, x_2\} = \{(1, 1, 1), (1.5, 2, 0)\} \end{aligned}$$

We ended up with effectively the same answer in a) and b), just with A and B swapped. In both cases, we ended up partitioning the data points of X into two clusters as follows:

$$\{x_1, x_2, x_3, x_4\} = \{x_1, x_2\} \cup \{x_3, x_4\}$$

The within-cluster sum of squares:

$$W(A, B) = \|x_1 - r_1\|^2 + \|x_2 - r_1\|^2 + \|x_3 - r_2\|^2 + \|x_4 - r_2\|^2$$

is one measure of cluster quality. However, this value is the same in a) and b) because the clusters are the same. In this respect, both solutions are equally good.

The main difference is that in a), it took two iterations for Lloyd's algorithm to converge to the answer. In b), it took only one iteration. Hence, the solution in b) can be considered marginally more efficient due to the choice of starting centroids.

Problem 2

a)

Refer to `hw2_2a.m`, to see the computation of the similarity graph using the weight and distance matrices at the cut-off $\varepsilon = 2.5$ for ε neighborhood.

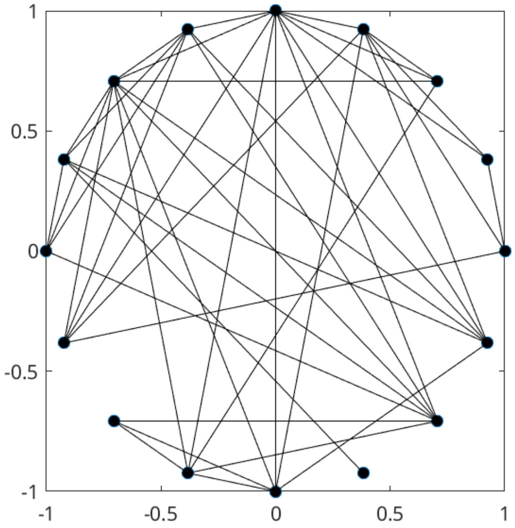


Figure 1: Similarity graph for $c = 2$.

From the similarity graph in Figure 1, we observe that there exists a path from any vertex to every other vertex. Hence, there is only one connected component, indicating that all data points belong to a single big cluster.

b)

The unnormalized Laplacian is found by subtracting the weight matrix W from the diagonal degree matrix D:

$$L = D - W$$

See `hw2_2b.m` for the computations.

```
>> hw2_2b

ans =

   -0.0000
    0.8188
    1.5428
    2.5207
    3.2215
    3.6453
    4.3127
    5.2325
    5.9444
    6.0494
    7.5214
    7.9123
    8.0162
    8.8315
   11.0997
   11.3307
```

Listing 1: Eigenvalues for c = 2, obtained by running `eig(L)`.

Based on the output in Listing 1, we see that there is only one $\lambda = 0$ eigenvalue, so the multiplicity is $k = 1$. According to Lemma 2.3.3, this implies that the graph consists of exactly $k = 1$ connected components, which is consistent with our conclusion in a).

c)

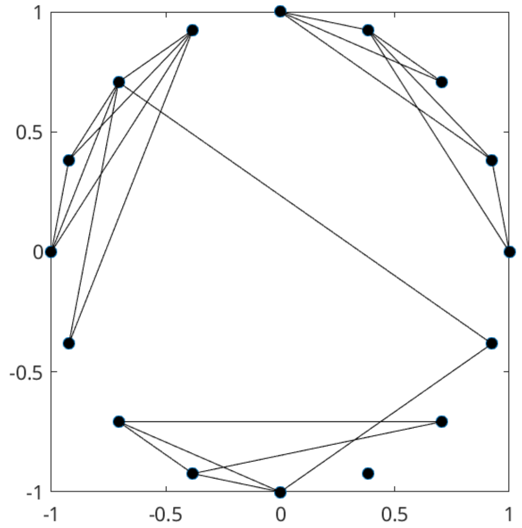


Figure 2: Similarity graph for $c = 3$.

With $c = 3$, the graph now consists of 3 connected components as shown in Figure 2. Refer to `hw2_2c.m` for computations to repeat a) and b) for $c = 3$.

```
>> hw2_2c

ans =

   -0.0000
   -0.0000
    0.0000
    0.1795
    1.4587
    1.5858
    2.0000
    2.7996
    3.0000
    4.0000
    4.0000
    4.4142
    4.4991
    5.0000
    5.0000
    6.0630
```

Listing 2: Eigenvalues for c = 3, obtained by running `eig(L)`.

The output is shown in Listing 2. This time, we obtain the eigenvalue $\lambda = 0$ with multiplicity $k = 3$. This is again consistent with Lemma 2.3.3 since the graph consists of $k = 3$ components.

d)

Refer to `hw2_2d.m` for the computation of similarity graph for $c = 3.1$ and $\varepsilon = 2.8$ and the eigenvector plot.

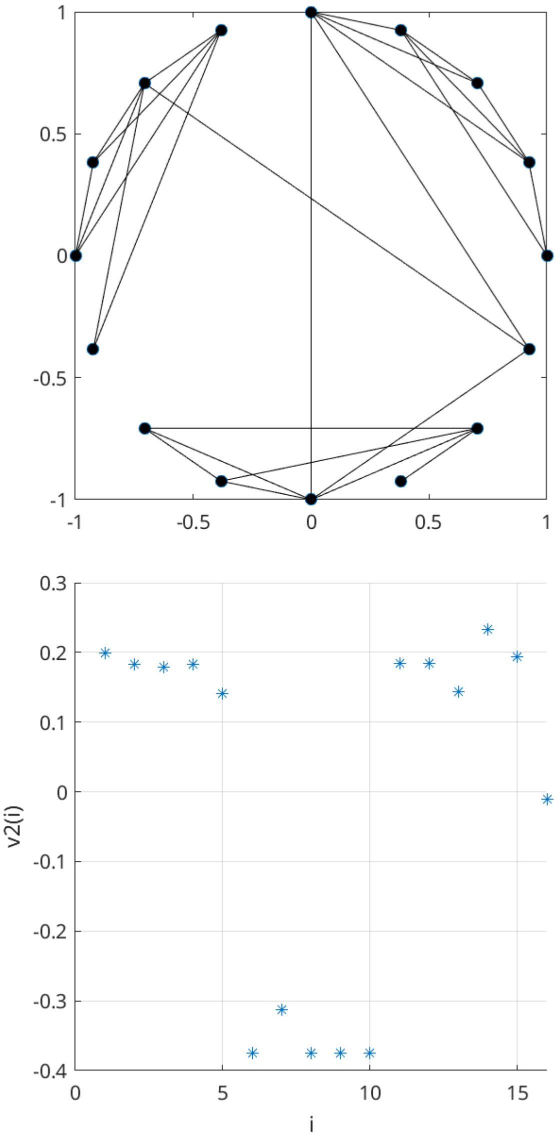


Figure 3: Top image: similarity graph for $c = 3.1$ and $\varepsilon = 2.8$. Bottom image: Plot of the values of the second eigenvector.

We know from Lemma 2.3.2 that the first eigenvector, corresponding to eigenvalue $\lambda = 0$, is:

$$v_1 = \mathbb{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

Since L is symmetric, we can apply the Rayleigh-Ritz theorem. The second eigenvector of L is:

$$v_2 = \operatorname*{argmin}_{y \in D_2(L)} r(y) = \operatorname*{argmin}_{y \in D_2(L)} \frac{y^\top L y}{\|y\|^2}$$

where

$$D_2(L) = \{y \in \mathbb{R}^{16} : \|y\| = 1, v_1^\top y = 0\} = \{y \in \mathbb{R}^{16} : \|y\| = 1, \mathbb{1}^\top y = 0\}$$

We also know that the problem of minimizing RatioCut for $k = 2$ is equivalent to the minimization problem above. The naive way of solving this minimization problem is by partitioning the data points $x_1, \dots, x_{16}$ based on the sign of each value of the second eigenvector, i.e.:

$$\begin{cases} x_i \in A \text{ if } v_{2i} \geq 0 \\ x_i \in B \text{ if } v_{2i} < 0 \end{cases}$$

Based on this rule, the points that are non-negative belongs to cluster A, and the points that are negative belongs to cluster B. Based on the plot in Figure 3, we see that values 1-5 and 11-15 of v_2 are positive and thus belong to A, while values 6-10 and 16 are negative and thus belong to B. It should be noted that point 16 is very close to zero, making it less clear whether its membership to B would really lead to a minimization of RatioCut.

Problem 3

a)

Refer to Figure 4 to visualize the measuring stations at 937 locations on the map of Bengali bay countries.

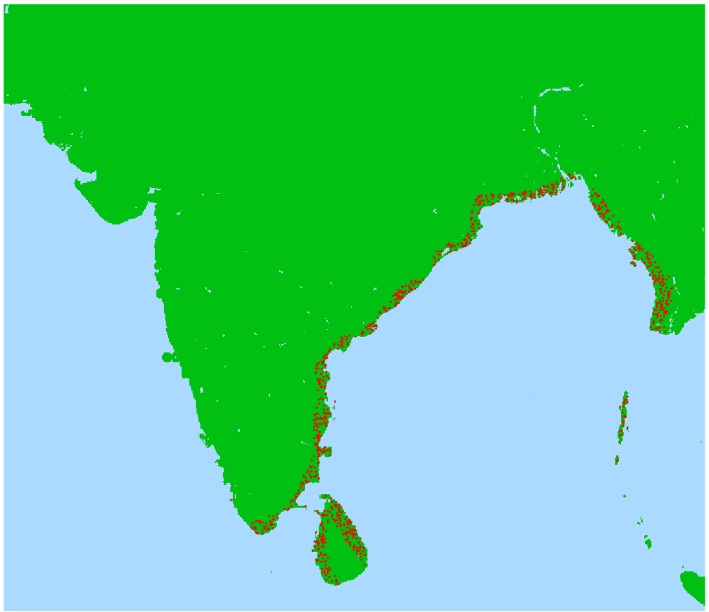


Figure 4: Map with all the measuring stations, marked in red.

b)

Refer to `hw2_3b.m` to see the construction of distance matrix from the time-series data using two-norm of the difference between the two time-series vectors. The time series distances between the individual measuring stations 102, 280 and 10 are shown in the Listing 3 below.

```
>> Dist(102, 280)

ans =

    3.3483

>> Dist(102, 10)

ans =

    4.6101

>> Dist(10, 280)

ans =

    4.8392
```

Listing 3: The time series distances between 102 and 280 and 10.

The map with the three measuring stations plotted is shown in Figure 5 below. Measuring stations 102 and 280 are located closer to Kolkata on the left, while station 10 is farther away in Burma on the right.

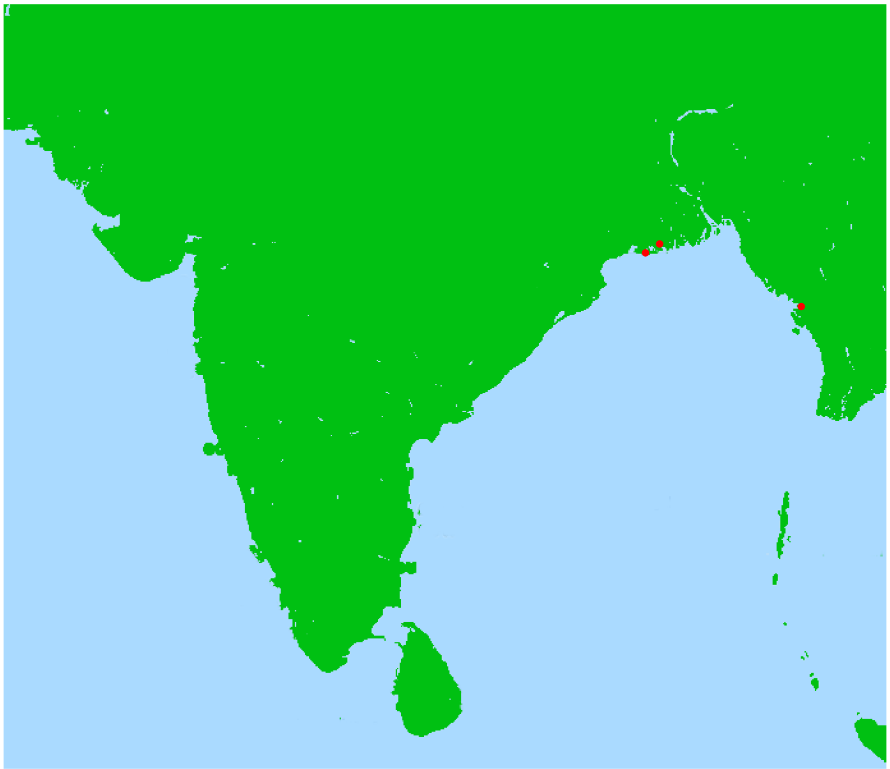


Figure 5: Map with measuring stations 102, 280 and 10, marked in red.

It makes sense that the time series distance between 102 and 10, as well as 280 and 10 are similar, as justified by their geographic locations. However, the fact that the time series distance between 102 and 280 is not significantly small could indicate a difference in plastic pollution characteristics between the two stations near the shore of Kolkata.

c)

Refer to `hw2_3c.m` for the implementation of the k-nearest neighbors using the `or` version, which is used to construct a weight matrix based on the distance matrix.

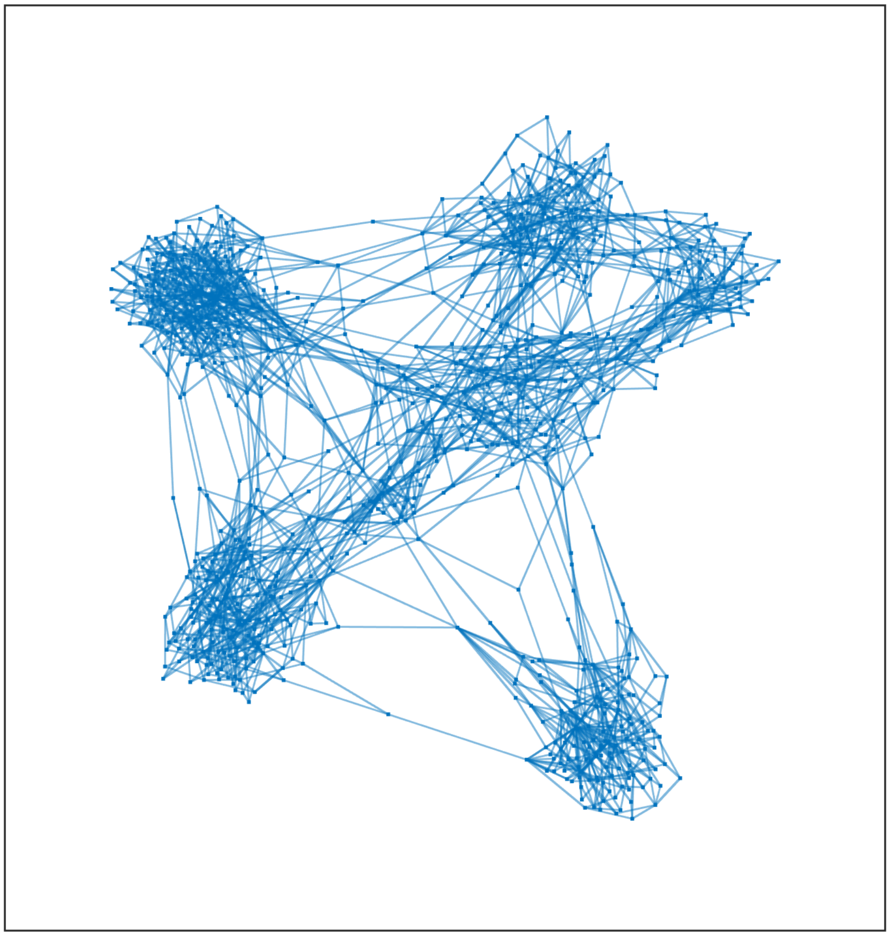


Figure 6: kNN graph with $k = 3$.

As shown in Figure 6, seven clusters can be visually identified: with two on the top, four in the middle, and one in the bottom, although the separation of the clusters in the middle is less clear.

d)

We make use of the weight matrix we constructed in task c). Together with the degree matrix D, we can now calculate the unnormalized Laplacian:

$$L = D - W$$

We now need to perform unnormalized spectral clustering with 7 clusters. Specifically, we want to find a partition $\{A_1, \ldots, A_7\}$ of the vertices of the graph such that

$$\mathrm{RatioCut}(A_1, \ldots, A_7)$$

is minimized. This is approximately equivalent to minimizing

$$\min_{H \in B} \mathrm{trace}(H^\top L H)$$

where $B = \{H \in \mathbb{R}^{n \times k} : H^\top H = I\}$. By the Rayleigh-Ritz theorem, the solution is given by $H = \begin{bmatrix} v_1 \ldots, v_7 \end{bmatrix}$ where v_i denotes the ith eigenvector of L. We then consider the rows of H to be points in $\mathbb{R}^7$, then apply k-means clustering with $k = 7$. This yields a vector of cluster indices for each data point.

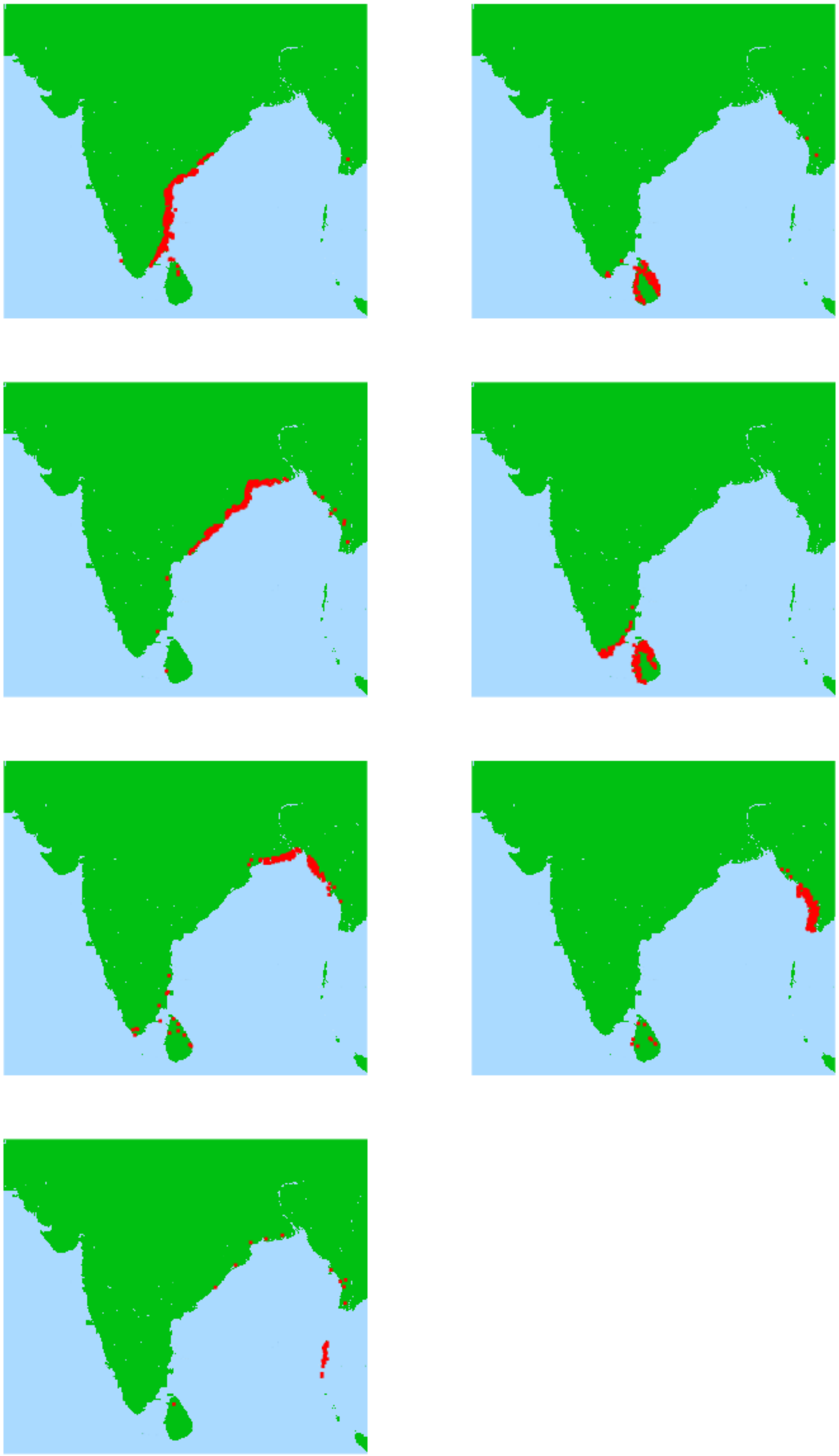


Figure 7: Plot of seven clusters on the Bengali map at $k = 3$ for kNN.

In Figure 7, we plot the (x, y) coordinates for each data point for each of the seven clusters. We have identified seven regions on the Bengali map where plastic pollution exhibits similar characteristics based on the three nearest neighbors. As seen, the cluster follows along the shores of Andaman Islands, Sri Lanka, Southern India, Bangladesh and Burma, indicating the concentration of plastic pollution.

e)

We repeat the same process for $k = 2$.

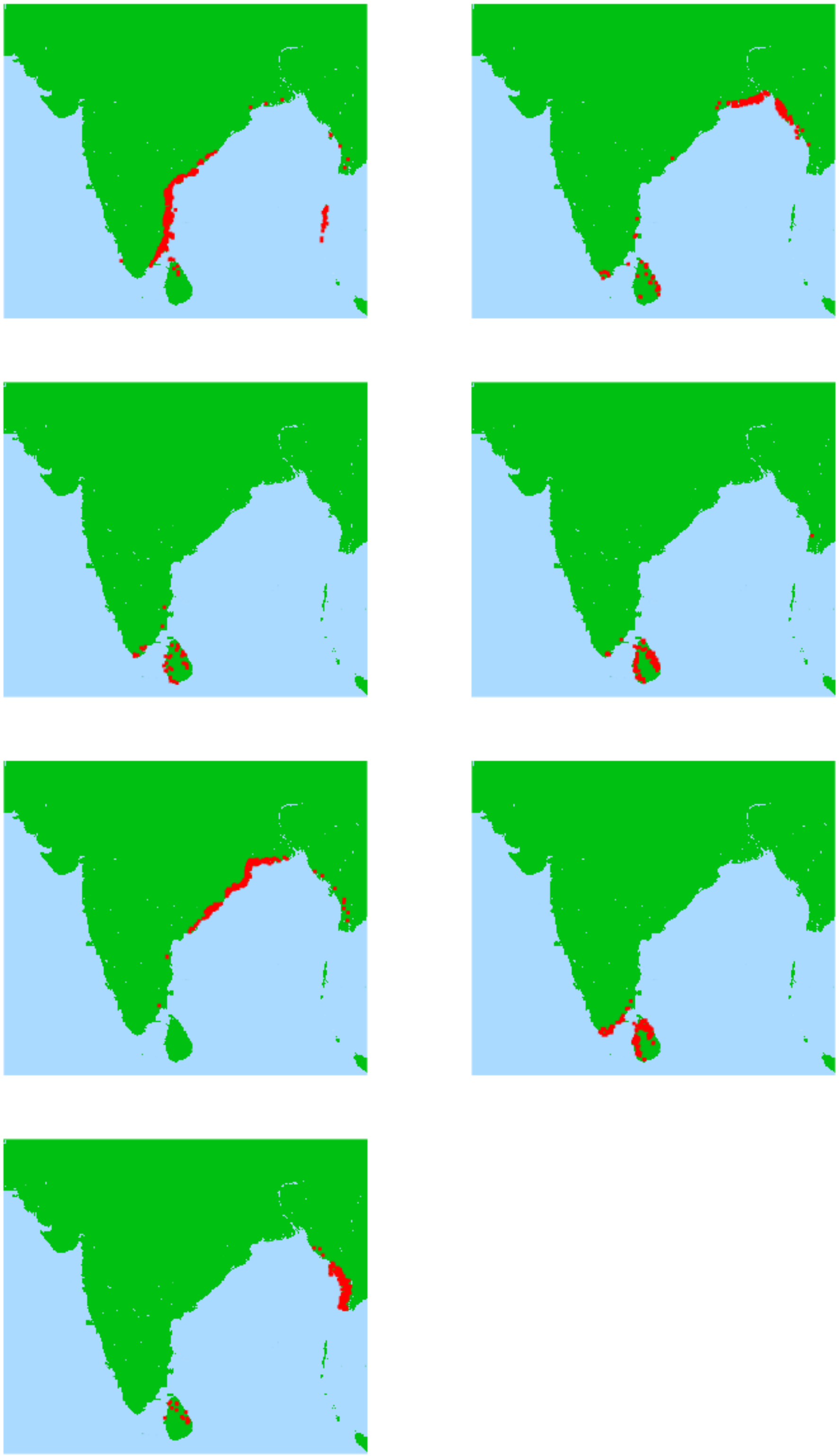


Figure 8: Plot of seven clusters on the Bengali map at $k = 2$ for kNN.

As seen in Figure 8, more localized clusters are formed at $k = 2$, focusing on smaller and immediate relationships. In contrast, at $k = 3$, we observe broader connections between regions. The general conclusion may differ by highlighting more localized sources of plastic pollution.

Problem 4

a)

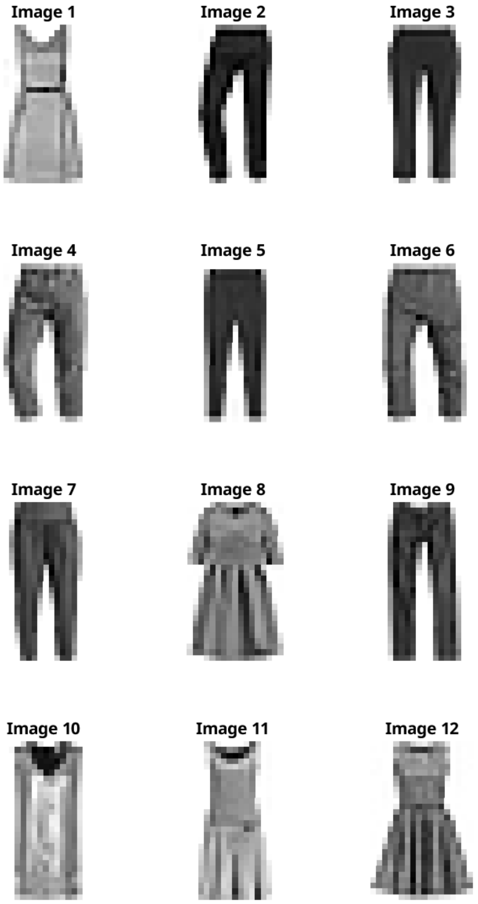


Figure 9: The first few images in row-major order.

```
>> hw2_4a
     2     1     1
     1     1     1
     1     2     1
     2     2     2
```

Listing 4: The first few values of the `correct` vector, reshaped into a matrix in row-major order.

Figure 9 shows that there are two types of items: trousers and dresses. Listing 4 suggests that our interpretation is correct. Trousers are assigned the label 1 while dresses are assigned the label 2.

b)

The four images are shown in Figure 10. We see that the two dresses have similar shapes and gray colors. Similarly, the two pairs of trousers have very similar shapes and black colors. These similarities contribute to the overall low norm when computing distances.

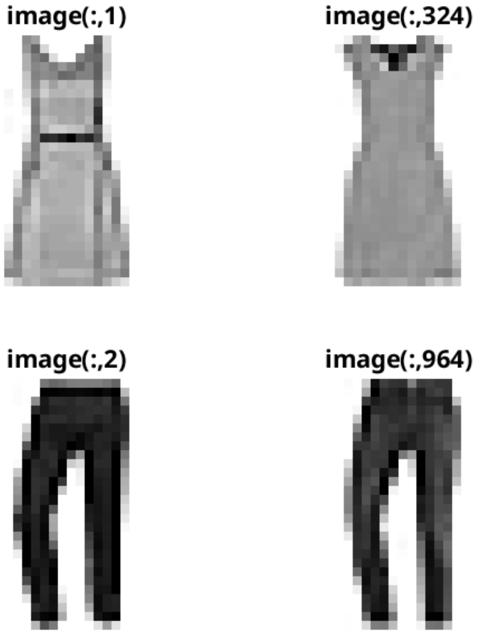


Figure 10: Top left: image 1. Top right: the nearest neighbor of image 1. Bottom left: image 2. Bottom right: the nearest neighbor of image 2.

c)

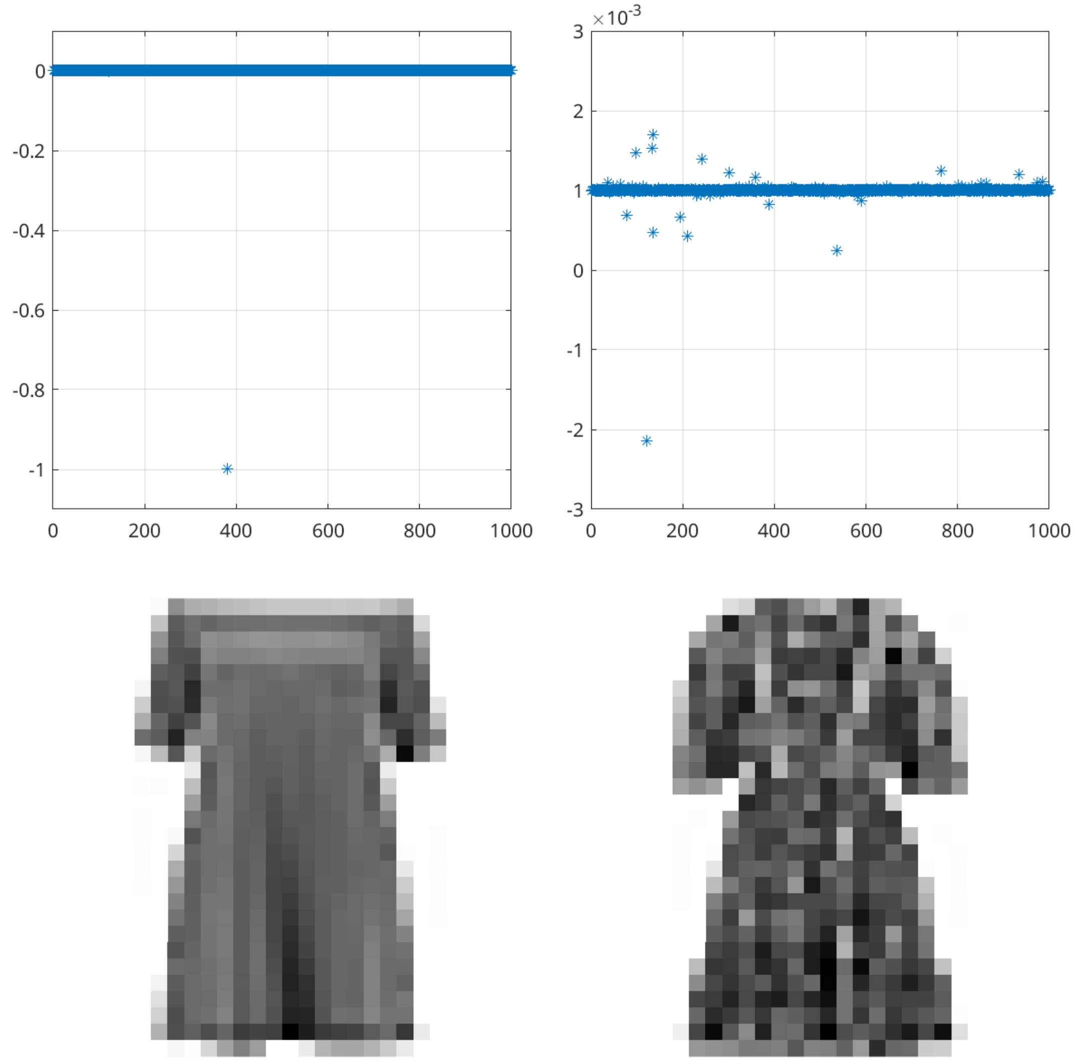


Figure 11: Top left: plot of all the $n = 1000$ values of the second eigenvector of the graph Laplacian L. Top right: zoomed in on the 999 values closest to the x-axis. Bottom: The two images corresponding to the two negative values.

We carry out spectral clustering with $k = 2$ clusters, then plot the components of the second eigenvector. In Figure 11, we see that one of the values is close to -1, while the rest of the 999 values are very close to 0. Zooming in, we also see that only two of the 1000 values are negative, while the rest of the values are positive and close to 10^{-3}.

Based on the signs alone, this indicates that the spectral clustering algorithm separates the images into two clusters: a tiny cluster containing only two images and a huge cluster containing 998 images. In addition, 999 of the values are very close to zero, so there is only weak evidence that their classification to their respective clusters is appropriate. Figure 11 reveals that the tiny cluster consists of two similar-looking dresses.

Spectral clustering is expected to work well when the values of the images are close to two discrete values. However, in this case, the vast majority of values are close to a single discrete value, namely 10^{-3}. Ideally, we would like to separate the image set into two clusters: one containing dresses and one containing trousers. Yet, we know from Figure 9 that there are more than just two dresses in the dataset. Ergo, it seems that spectral clustering does not work well in this case.

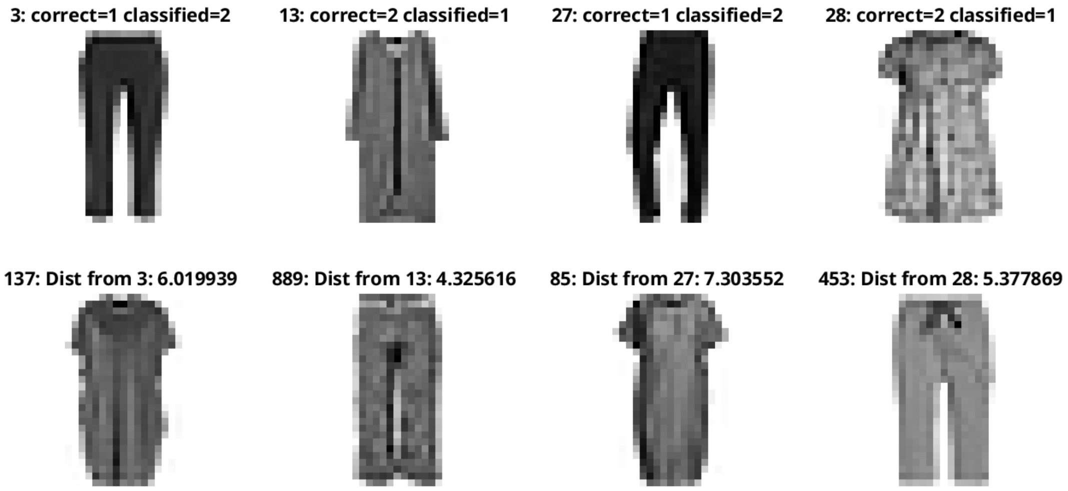
d)

We calculate τ as the median of the values in the second eigenvector, then find all values that are below τ. From this, we compute a vector of cluster indices consisting of 1s and 2s, which we then compare with the `correct` vector. The number of correct predictions is given in Listing 5.

```
>> hw2_4d
Tau: 0.001005
True positives: 818
Accuracy: 81.8 %
```

Listing 5: Number of true positives and accuracy for $\tau = 0.001005$.

e)



3: correct=1 classified=2
13: correct=2 classified=1
27: correct=1 classified=2
28: correct=2 classified=1
137: Dist from 3: 6.019939
889: Dist from 13: 4.325616
85: Dist from 27: 7.303552
453: Dist from 28: 5.377869

Figure 12: Top: four misclassified images. Below each, the image of the opposite class with the shortest distance is displayed.

In Figure 12, we see that two dresses and two pairs of trousers were misclassified. Below each pair of trousers, we display the dress that is most similar to it. Below each dress, we display the pair of trousers that is most similar to it. In this context, two images are maximally similar if their distance is minimal. Distance is given by the distance matrix computed in b).

We see that there exist some similarities between each pair of images. In particular, the pair of trousers in the second column has a similar shape as the dress above it, and a similar shade of gray.

f)

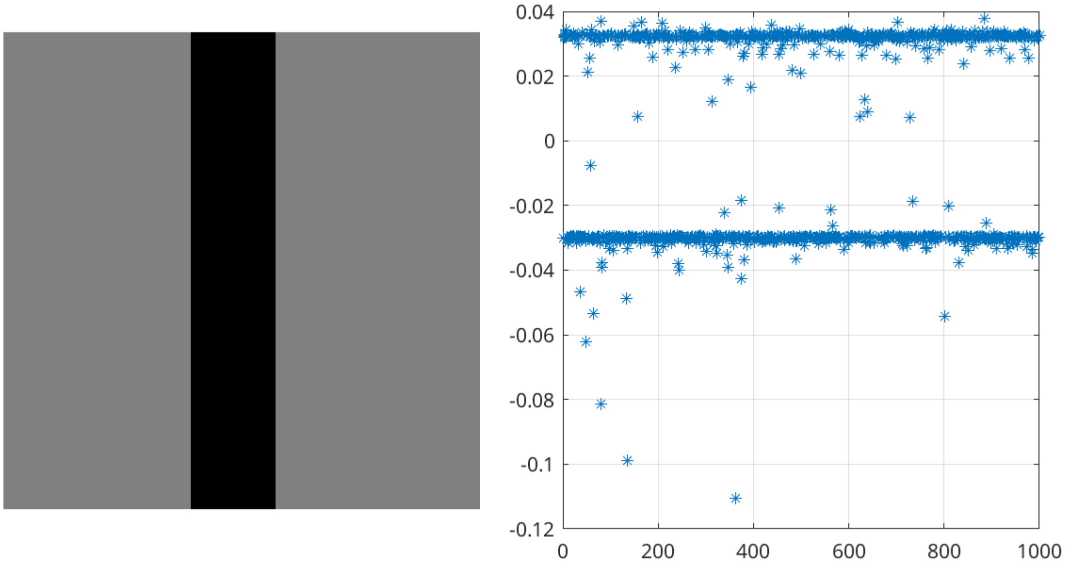


Figure 13: Left: The weight vector reshaped into an image. Right: the components of the second eigenvector when this weight vector is used.

As shown in Figure 13, the given weight vector is constructed so that the center of the image is given the most weight. This may be appropriate, as each item is consistently found in the center of the image. The edges of each image mostly consist of white background, so it may not be appropriate to treat such pixels with equal weight when computing distances. The plot of the second eigenvector now shows that the majority of values are gathered in the vicinity of two distinct values: one positive, one negative. It also looks like the two clusters are more evenly matched in terms of size.

```
>> hw2_4f
Tau: -0.023822
True positives: 994
Accuracy: 99.4 %
```

Listing 6: Number of true positives and accuracy for $\tau = -0.023822$.

Thanks to the weight, the classification accuracy is now much better. Listing 6 shows that the in-sample prediction accuracy is now 99.4 %, meaning that only 6 images have been incorrectly classified.

Problem 5

Summaries of video quizzes are given below.

Video Quiz 6: Eigenvalue multiplicity

The algebraic multiplicity of an eigenvalue λ is defined as the multiplicity of the root $s = \lambda$ of the characteristic polynomial $f(s) = \det(A - sI)$. Whereas, the geometric multiplicity is given by the number of linearly independent eigenvectors associated with λ. It is important to note that the eigenvectors are not unique, although they form a unique basis for a subspace called the eigenspace.

For symmetric matrices, orthogonal basis is often returned by MATLAB, Julia or Python.

Video Quiz 7a: Mincut

MinCut is a graph-based clustering method that partitions a graph into k subgraphs by minimizing the number or weight of edges that need to be removed. It is defined as $\mathrm{Cut}(A_1, \ldots, A_k) = \frac{1}{2}\sum_{i=1}^{k} w(A_i, \bar{A}_i)$, representing the total edge weight between the subgraphs. The goal of MinCut is to find a partition that minimizes this cut value, ensuring minimal interconnections between clusters. A lower cut value indicates better-separated clusters with fewer connections between them. However, MinCut does not capture the structure of clusters as effectively as RatioCut.

Video Quiz 7b: Rayleigh-Ritz theorem

The Rayleigh-Ritz theorem states that the eigenvalues and eigenvectors of a symmetric matrix can be obtained by minimizing the so-called Rayleigh quotient subject to a set of conditions. Specifically, the kth smallest eigenvalue of matrix A is obtained by minimizing $r(x) = \frac{x^\top Ax}{x^\top x}$ subject to $k-1$ orthogonality conditions. The orthogonality conditions state that x is orthogonal to each of the $k-1$ first eigenvectors. The minimizer x is then the kth eigenvector.

Video Quiz 8: Eigenvalue derivatives

The eigenvalue equation for a symmetric matrix is given by: $A(\varepsilon) = A_0 + \varepsilon A_1, \quad A(\varepsilon)x(\varepsilon) = \lambda(\varepsilon)x(\varepsilon)$, with the normalization constraint $\|x(\varepsilon)\| = 1$. First order pertubation of the eigenvector and eigenvalue in terms of ε, we get: $x(\varepsilon) = x_0 + \varepsilon x'(0) + \mathcal{O}(\varepsilon^2), \lambda(\varepsilon) = \lambda_0 + \varepsilon\lambda'(0) + \mathcal{O}(\varepsilon^2)$. Here, $\lambda'(0)$ represents the slope of the eigenvalue trajectory, quantifying its rate of change with respect to ε.

Video Quiz 9: Proof of the Rayleigh-Ritz theorem

The first step in proving the Rayleigh-Ritz theorem is to reduce the problem of minimizing the Rayleigh quotient of a symmetric matrix A into the problem of minimizing the Rayleigh quotient of its associated eigenvalue decomposed diagonal matrix. We then make use of the fact that the any vector in the minimization domain has $k-1$ leading zero elements. This later enables us to use the squeeze theorem in order to prove that the minimum of the quotient is equal to the kth eigenvalue.