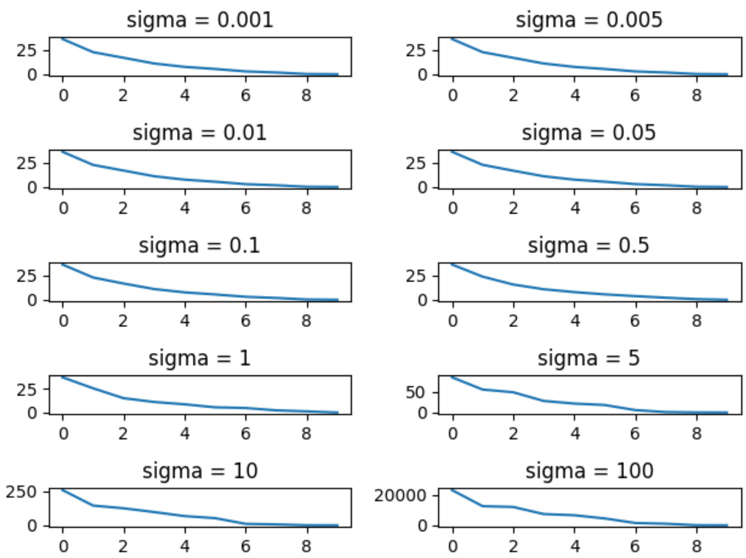
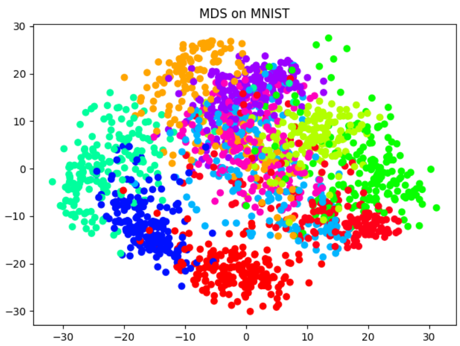
Ex 5 APML CS: Saji.Tibi ID: 207905407

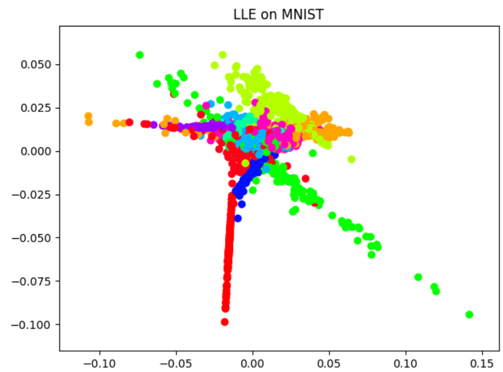
2.4.1 Scree plots for MDS:
We tried to embed random 2-d data into higher dimension using random rotation and adding gaussian noise, then we checked how this noise effects our ability to decide that the data was from 2-d using eigenvalues ( similar to elbow method in ex4) and I got this result :

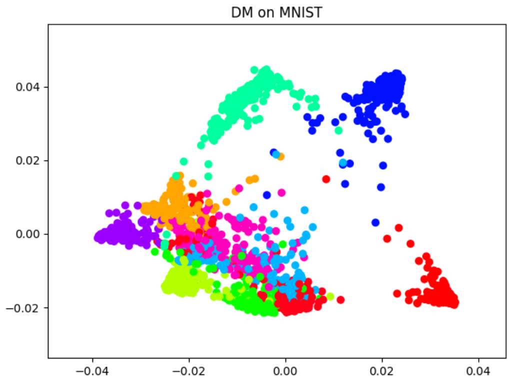



Note that on low noise we can clearly see the drop around value 2 ( if above image doesn't illustrate that enough please run the code and output each graph of noise separately this way there is less scaling in size and its more obvious) and when more noise is added the drop is around 2 is less obvious.

2.4.2 MNIST:
Running all of the algorithms on MNIST data gave good performance, although finding correct parameters for LLE and DM wasn't easy and more tuning would give betters results, note how DM clusters are best separated of all algorithms.

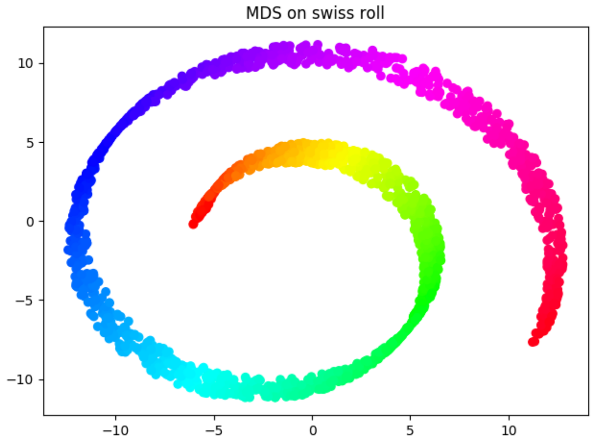


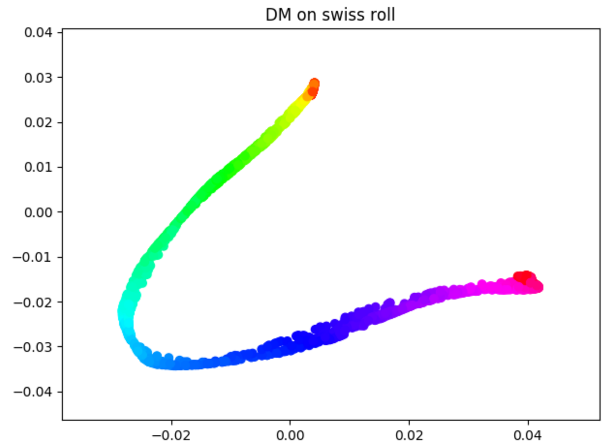



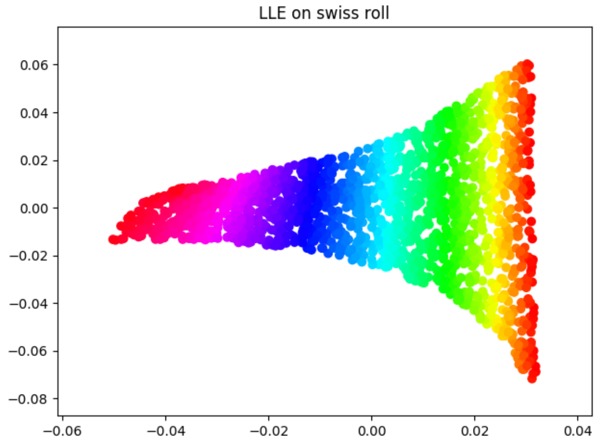



### 2.4.3 Swiss roll

Swiss roll structure isnt linear so MDS have failed as expected and gave poor results, LLE has only one parameter but its very sensitiveness while DM has two parameters and its they are less sensitive to changes, I would say LLE is easier for tuning although its senstive.
Also note that that LLE gave better results we can clearly see the clusters as if it were linear.

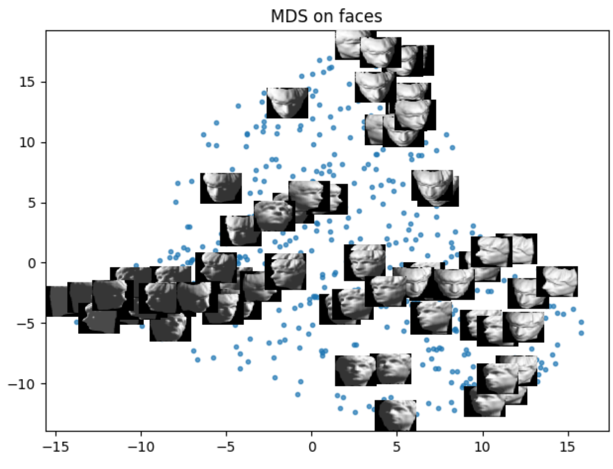


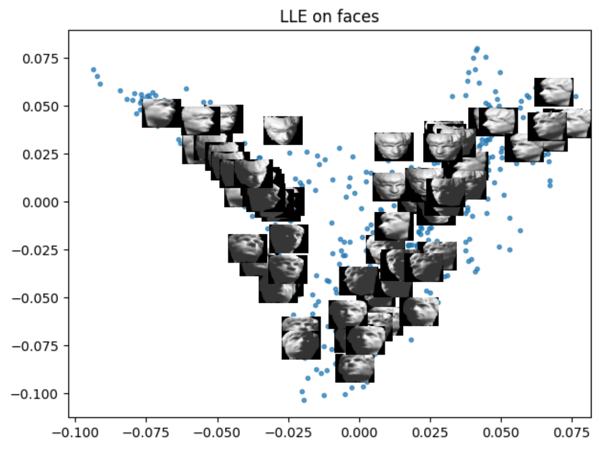



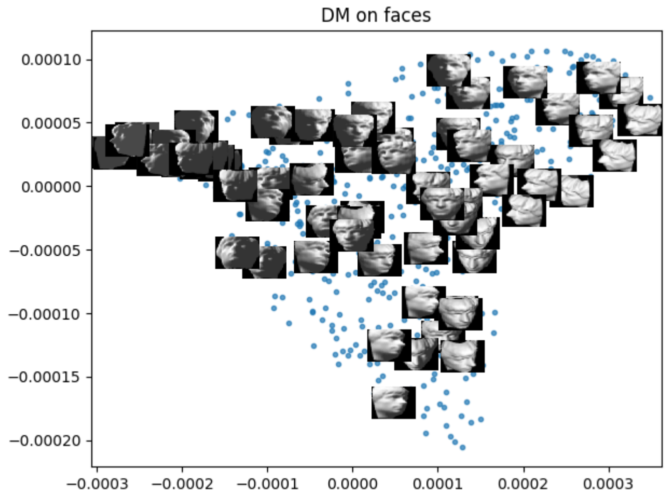




### 2.4.4 Faces

All of three algorithms could recover the structure very well, although the reduction was done from higher space to 2-d space we can see that they successfully manged to recover the directtions of faces, some of them better than other, MDS didn't have the best results but we didnt have to tone it at all, LLE and DM have really good results but I had to try multiple parameters to reach this result:



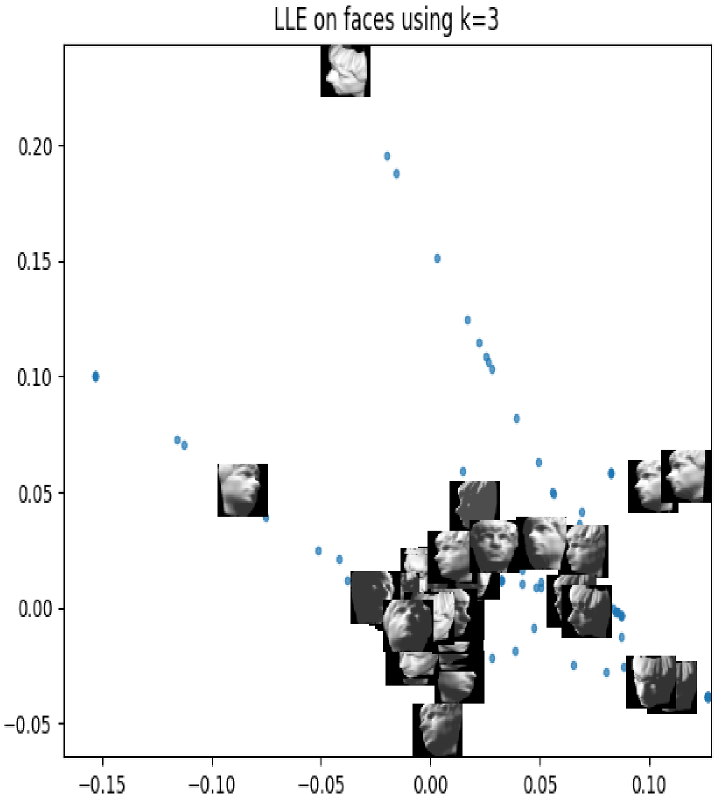



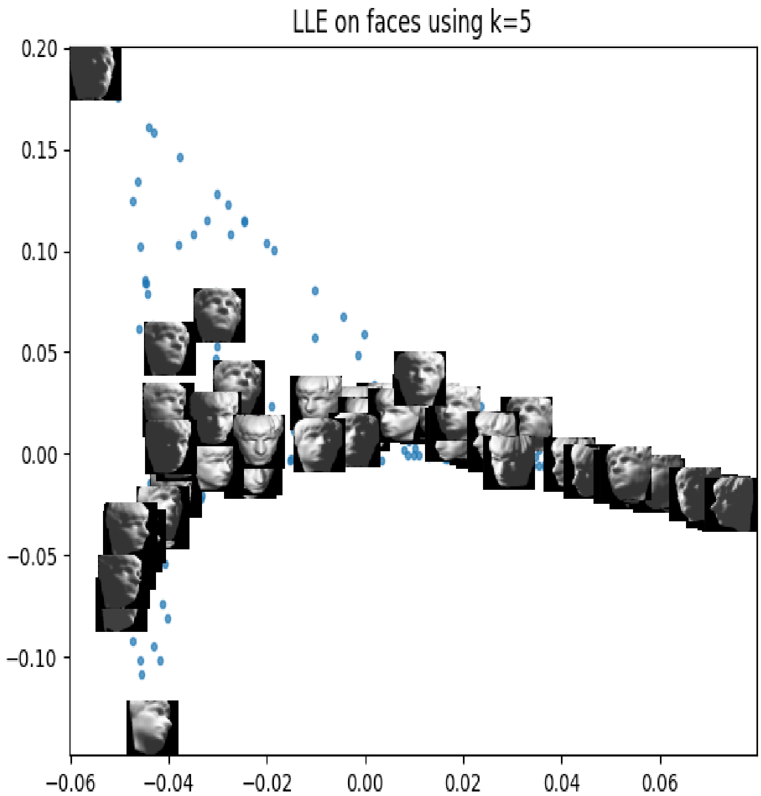


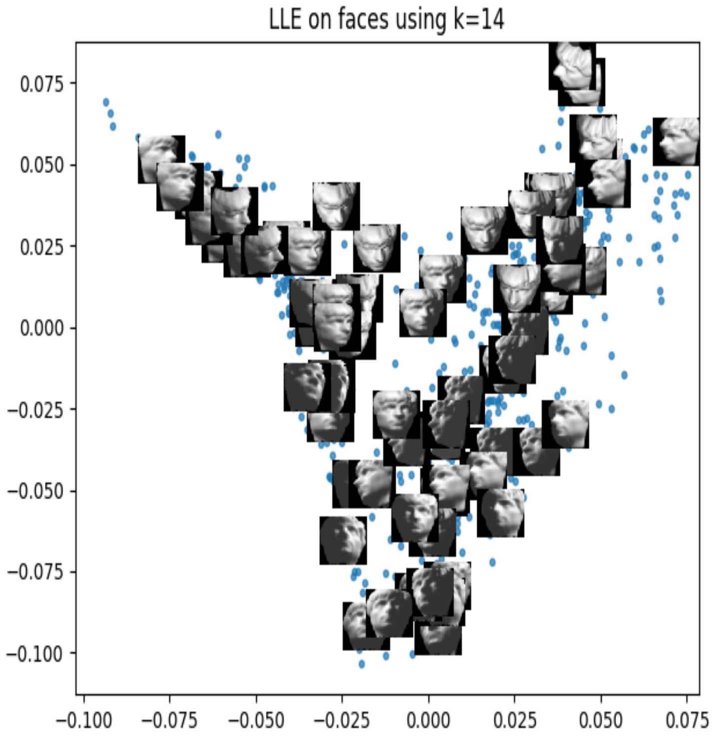

Note how on MDS some faces ( like in left side) are in wrong clusters .

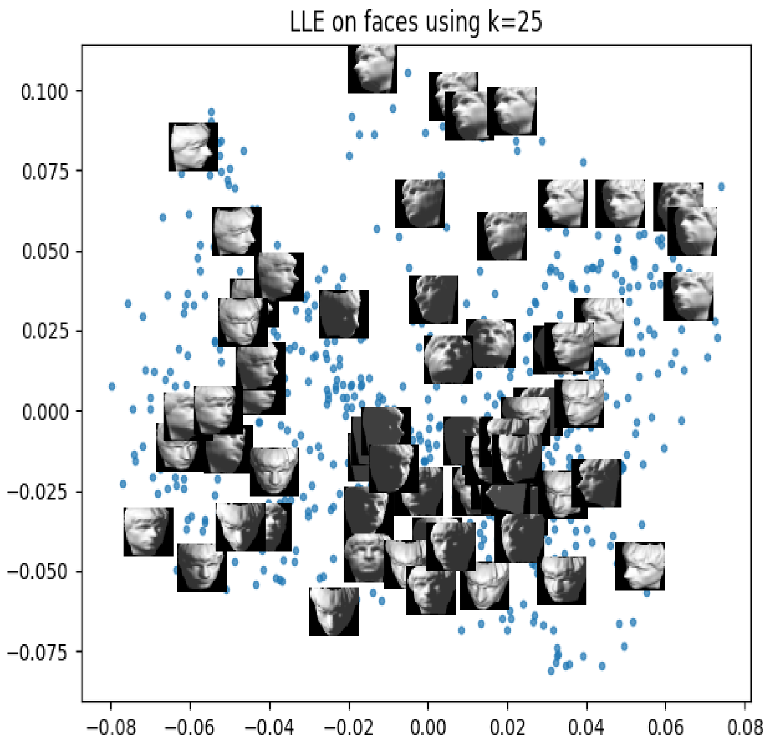
2.4.5 Parameter Tweaking

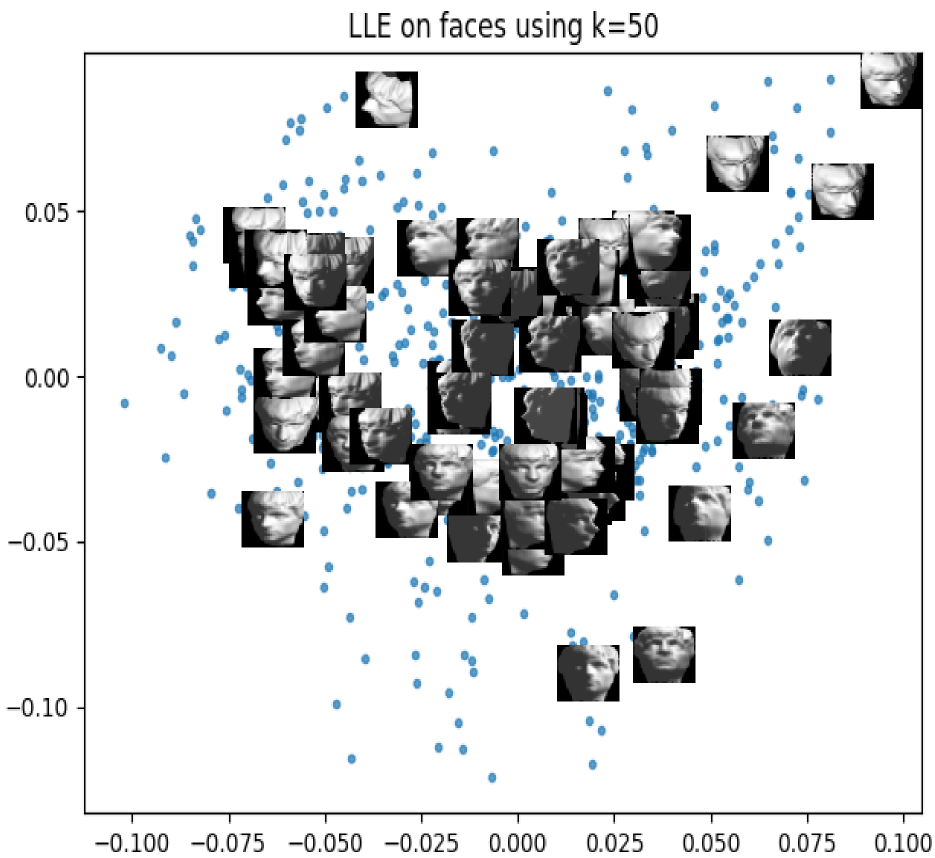
Exploring different values of k in LLE points to that small k would cause clustering very different angles in same cluster, high k causes similar angles to be clustered in separate clusters.




LLE on faces using k=14
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
−0.100
−0.100 −0.075 −0.050 −0.025 0.000 0.025 0.050 0.075


LLE on faces using k=25
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
−0.08 −0.06 −0.04 −0.02 0.00 0.02 0.04 0.06 0.08


LLE on faces using k=50
0.05
0.00
−0.05
−0.10
−0.100 −0.075 −0.050 −0.025 0.000 0.025 0.050 0.075 0.100

1.1

1) for any $z \in \mathbb{R}^n \setminus 0^n$, $zz' \geq 0$, we show that

$zSz^T \geq 0$, this means $S$ is PSD ?

$$[zSz^T]_j = [z \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T z^T]_j =$$

for each column itself

$\bar{x} = 0$

$$= [z \frac{1}{n-1} \underbrace{\sum_{i=1}^{n} x_i x_i^T}_{\text{semi positive for each } i \text{ so repeated with } x' \geq 0} z^T]_j = [\underbrace{z x' z^T}_{\text{sem positive multiplication}}]_j \geq 0$$

and this for each $j$ so sam column is semi positive and $S$ is PSD

2) since $\bar{x} = 0$, $S = \frac{1}{n-1} X^T X$, we will show that $\text{rank}(X) = \text{rank}(X^T X)$, using SVD

$X = U \Sigma V^T$, $UU^T = I$, $VV^T = I$,

$XX^T = U \Sigma^2 U^T$ from above the calcul

$\text{rank}(X) = \text{rank}(XX^T) = \text{rank}(\Sigma^2) = \text{rank}(\Sigma)$

2.1

$$[w^T G w]_j = w_j^T G w_j = w_j^T G_j w_j =$$

inner products of size K, for K neighbors of $x_j$s

$$= w_j^T \begin{bmatrix} \langle z_{i_1}^T, z_j \rangle \\ \vdots \\ \langle z_{i_K}^T, z_j \rangle \end{bmatrix} w_j = \|\sum_{j \in N(i)} w_j z_j \|^2$$

neighbors of $x_i$

2.2. ~~[illegible]~~

$$L(w_i, \lambda) = w_i^T G w_i - \lambda(1^T w - 1)$$

$$= \|\sum_{j \in N(i)} w_j z_j\|^2 - \lambda(1^T w - 1)$$

$$\frac{\partial L}{\partial w_i} = \frac{\lambda}{2} G_i^{-1} 1$$

this means $\Rightarrow$ $w = \frac{\lambda}{2} G^{-1} 1$

since this holds for each $i$

3.1)

1- ~~Bfirst~~ base case $t=1$

$A_{i,j} = P(X_1 = x_j \mid X_0 = x_i)$, this hold from given definition of $A_{i,j}$, now lets assume it holds for $t$, and prove for $t+1$

$$A^{t+1}_{i,j} = A^{t}_{i,j} A_{i,j} = P(X_t = x_j \mid X_0 = x_i)\, P(X_1 = x_j \mid X_0 = x_i)$$

$$= \frac{P(X_t = x_j \cap X_0 = x_i)}{P(X_0 = x_i)} \cdot \frac{P(X_1 = x_j \cap X_0 = x_i)}{P(X_0 = x_i)}$$

$$= \frac{P(X_t = x_j \cap X_0 = x_i)}{P(X_0 = x_i)} \frac{P(X_1 = x_j)}{P(X_0 = x_i)} \cdot \cancel{\frac{P(X_0 = x_i)}{P(X_0 = x_i)}}$$

independence

$$= \frac{P(X_{t+1} = x_j \cap X_0 = x_i)}{P(X_0 = x_i)} = P(X_{t+1} = x_j \mid X_0 = x_i)$$

3.2)

$$A\mathbb{1} = D^{-1}K = \underbrace{\text{diag}(D_1 \ldots D_n)}_{*}^{-1} K, \quad D_i = \sum_{j=1}^{n} K_{ij}$$

since * is diagonal we get that for each column in result we get

$$[A\mathbb{1}]_n = \begin{pmatrix} D_1^{-1} & \ldots & 0 & \ldots & 0 \\ & & \vdots & & \\ 0 & \ldots & D_r^{-1} & \ldots & 0 \\ 0 & \ldots & & 0 & D_n^{-1} \end{pmatrix} \begin{bmatrix} \vdots \\ \vdots \end{bmatrix} K$$

[illegible] K rows

$$= \left(\sum_{j=1}^{n} K_{ij}\right)^{-1} \cdot \begin{bmatrix} K_1 \\ \vdots \\ K_n \end{bmatrix} = \mathbb{1}$$

3.3) from definition $A v = \lambda v \Rightarrow \lambda$ is eigenvalue

since $A = D^{-1}K$ from above calculations we get each $D_i = \sum_{j=1}^{n} K_{ij} = \sum_{j=1}^{n} e^{-\|x_i - x_j\|^2/\varepsilon}$ (for some $\varepsilon$)

this means each $\|D^{i^{-1}}\| \leq 1$ (exponent property)

and $\|K_{ij}\|^2 \leq 1$ so multiplication of both is

$\|D_i^{-1} K\| \leq 1$