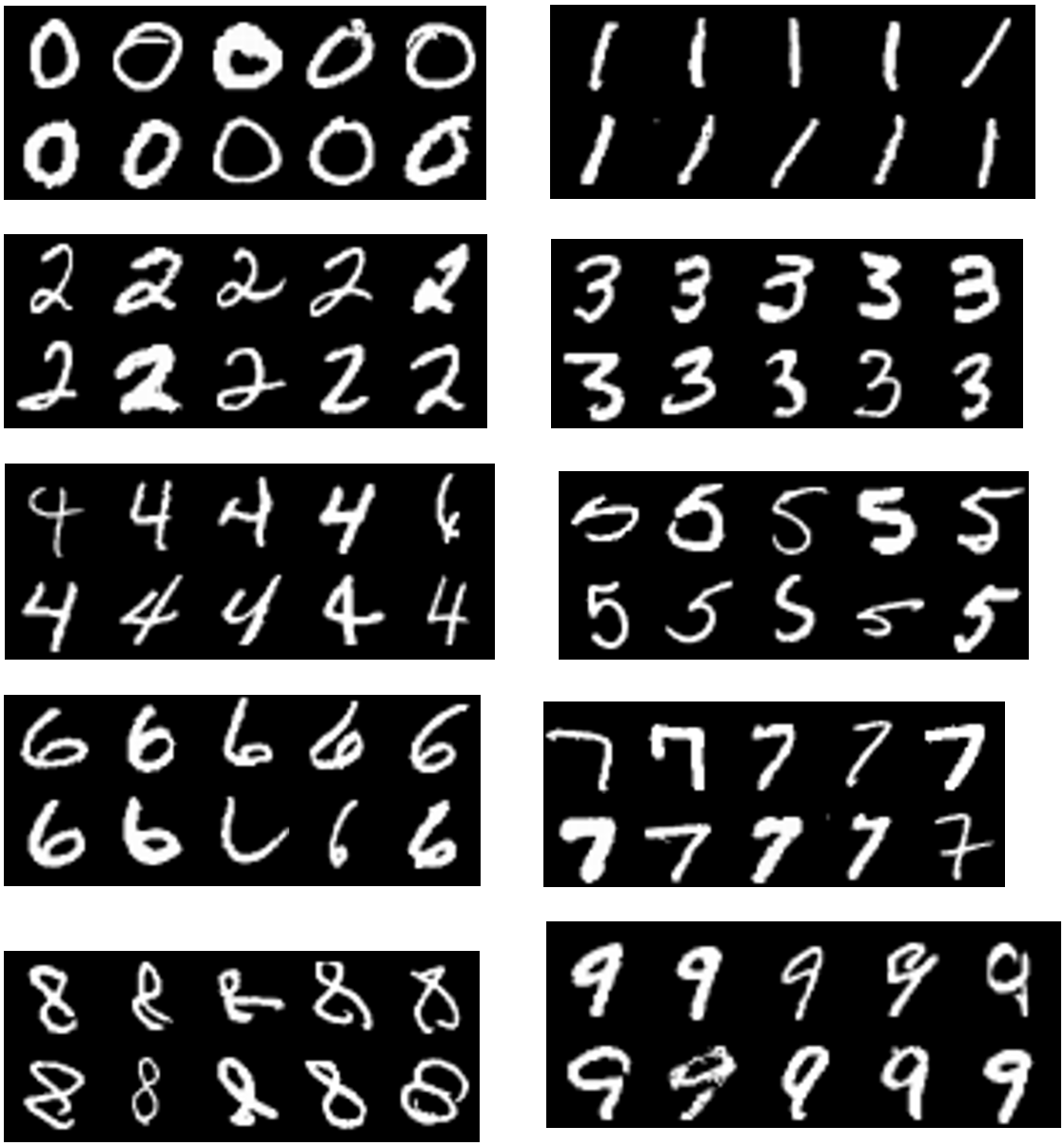
Report Task1 - PCA and Clustering

Task 1.1



Figures 1-10 (counting from left) : shows the images of the first ten samples in the training data (Xtrn) for Class k, where k ={1,..,10).

Task 1.2

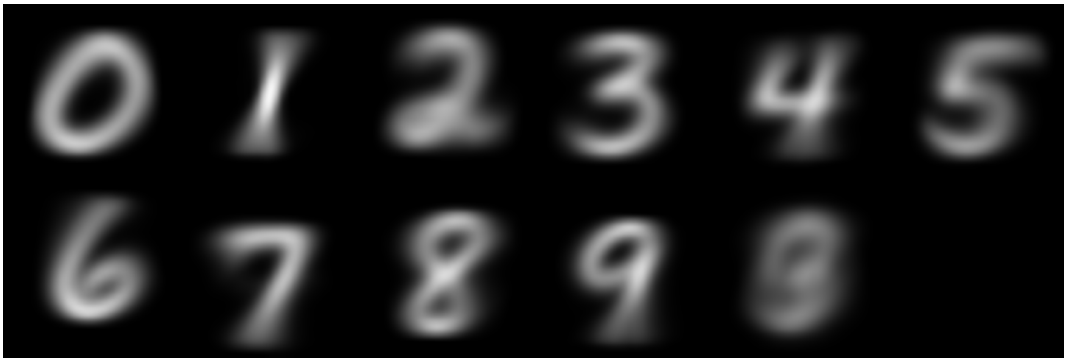


Figure 11: the image of the mean vector each class Class , and image 11th is the image of the mean vector for all the classes.

Task 1.3

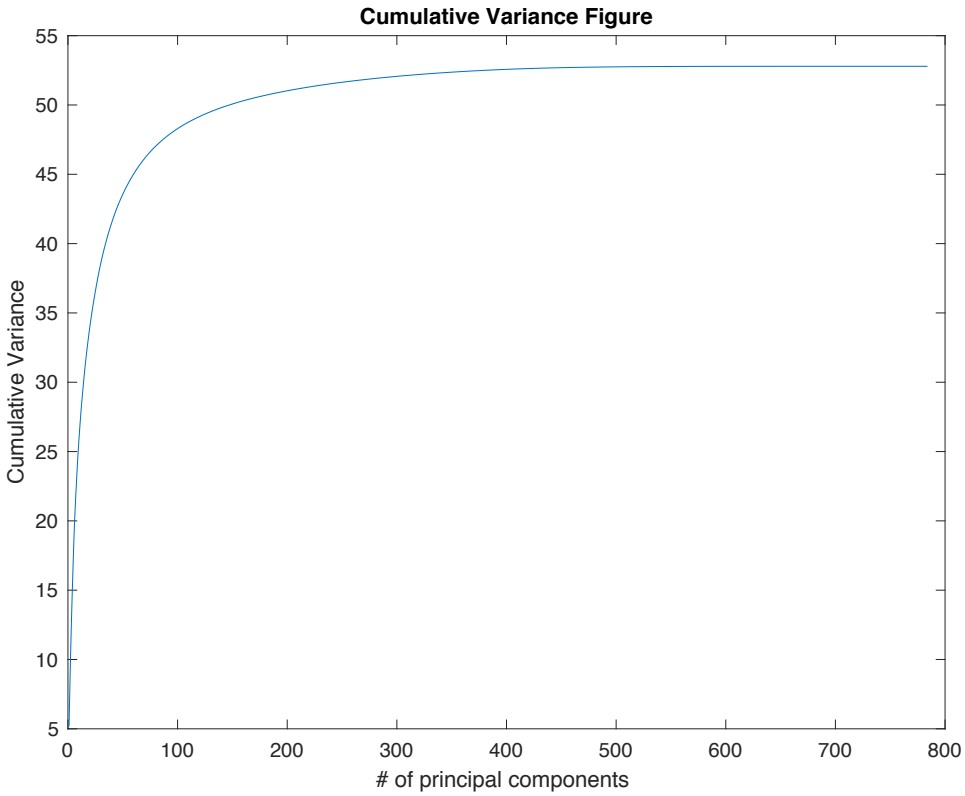


Figure 12: cumulative variance of PCA

Percentage	Minimum Dimension
70%	26
80%	44
90%	87
95%	154

Table 1: Minimum Dimension of required to cover certain percentages of dataset using PCA

Task 1.4

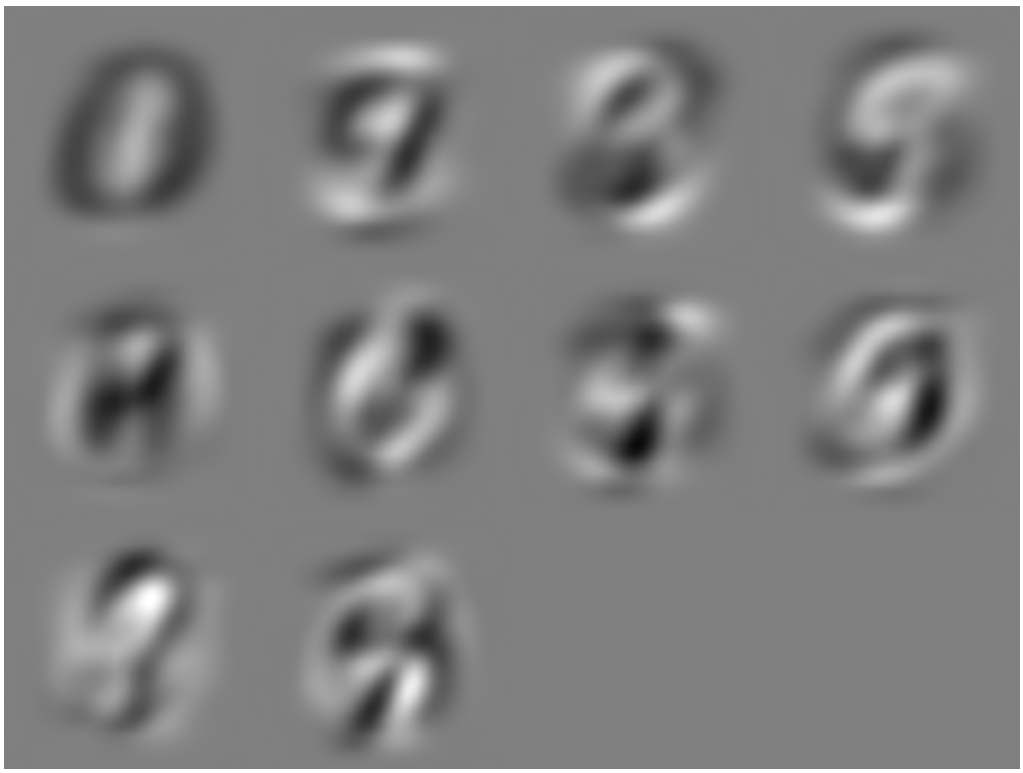


Figure 13 first ten principal components, where image i (i=1,…,10) is the image of i-th principal component.

Task 1.5

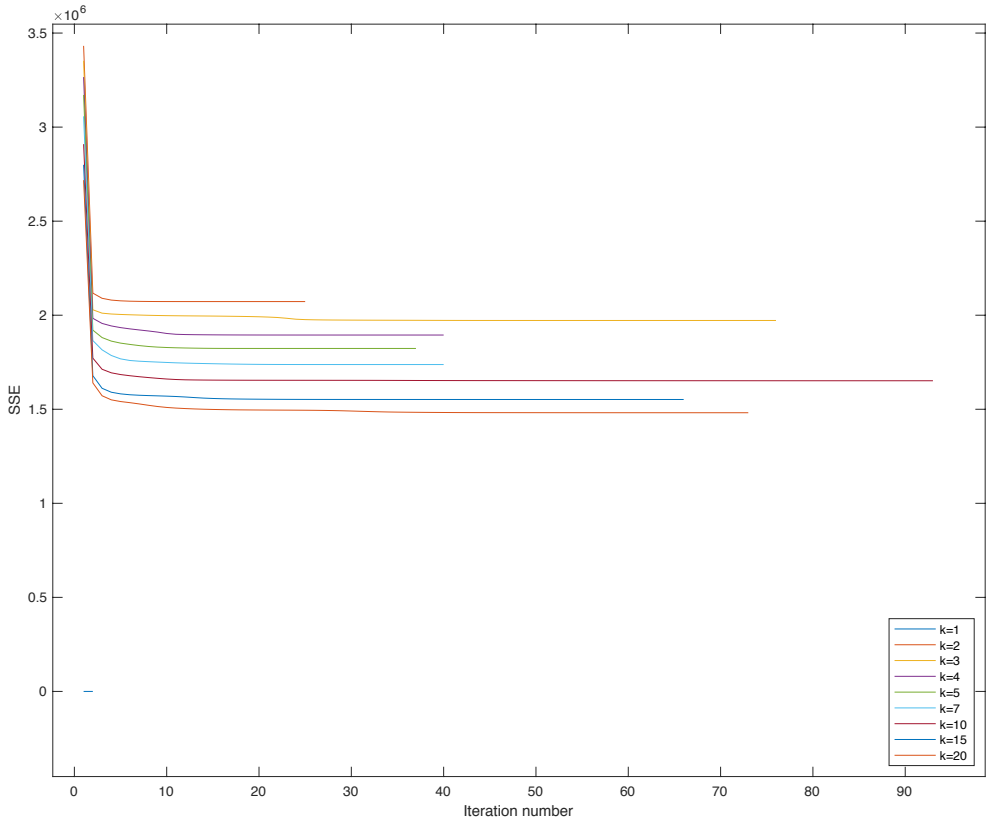
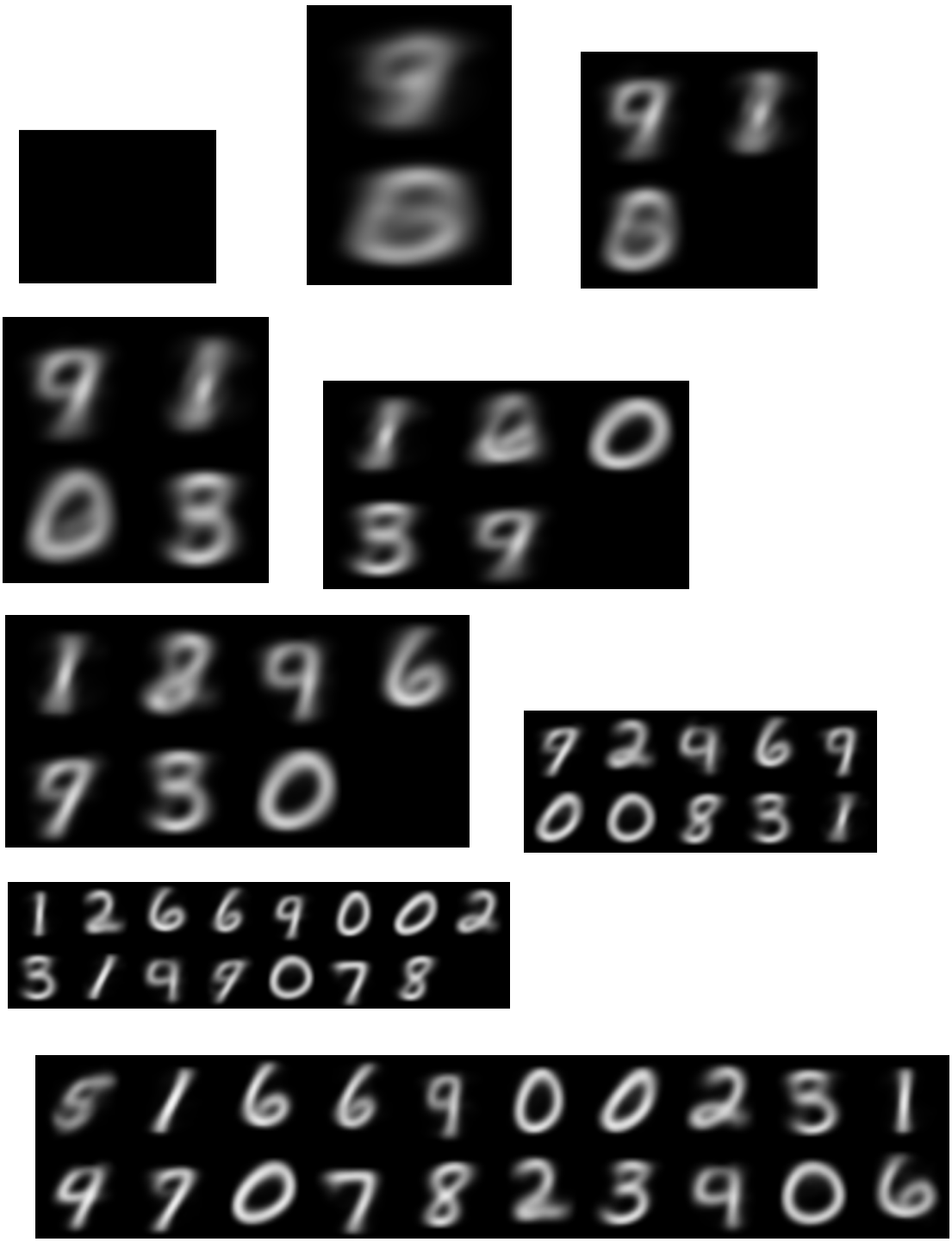


Figure 14: A figure that displays a graph of SSE vs k.

K	Time taken to converge(seconds)
1	0.5068
2	10.8402
3	44.2227
4	29.3311
5	32.5200
7	46.0005
10	144.7062
15	151.4977
20	218.4089

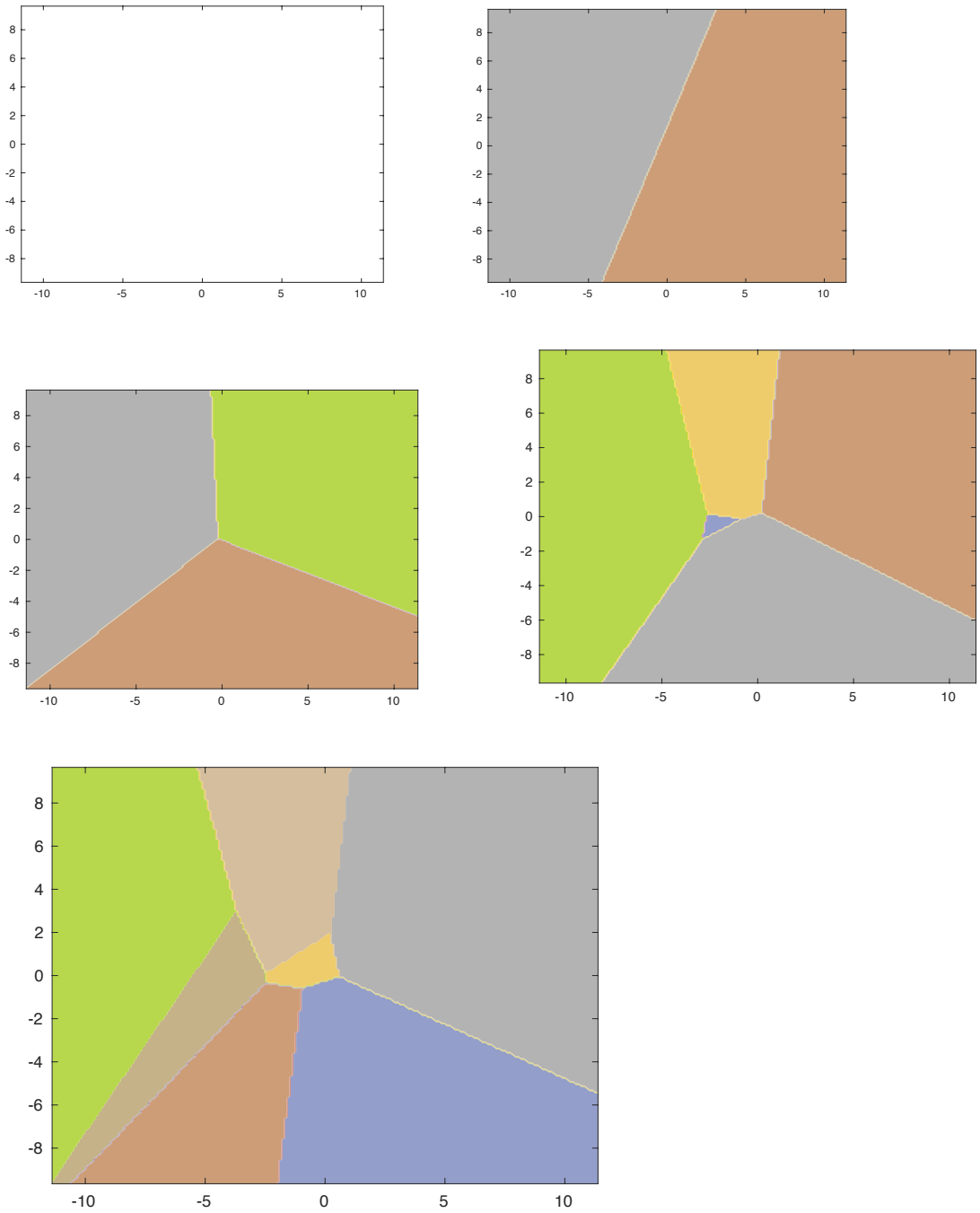
Table 2: it displays the time taken for each k to converse using k-means clustering.

Task 1.6



Figures 15-29(counting from left): display the images of cluster centres obtained with k-Mean clustering, k = number of numbers displayed on each . The i-th element in [1,2,3,4,5,7,10,15,20].

Task 1.7



Figures 30-35 : Figure i (i=1,...,5) displays the cluster regions for k, where k is the i-th element in [1,2,3,5,10]

Task 1.8

My Mini Research

Summary: This is a mini research project investigate the k-means clustering in terms of initial cluster centres, i.e how different initial cluster centres result in different cluster centres, for which I employ SSE to measure the clustering performance.

Approaches:

1. Fixed initialization of centres : chooses k number of centres from the start, middle, or at the end of dataset. 33.5200
2. Random initializing (uniform randomization): choose k number of centres randomly from dataset. 31.87 seconds, 33.46
3. Splitting dataset into k equal chunks(portions) and choose centres from each start or end of each chunk(or partition). 71.1212, 41.82
4. Using random initialization multiple times and put a point depending to majority of votes from all clustering.

Results:

This result are for k=5

- Fixed initialization average time = 33.5200 seconds
- Random initializing average time =31.87
- Kth chunks(3) method was as follow:
    - Picking first centres from 0^{th} index of dataset average time =41.82
    - Picking from kth index of a portion up to the end average = 71.1212

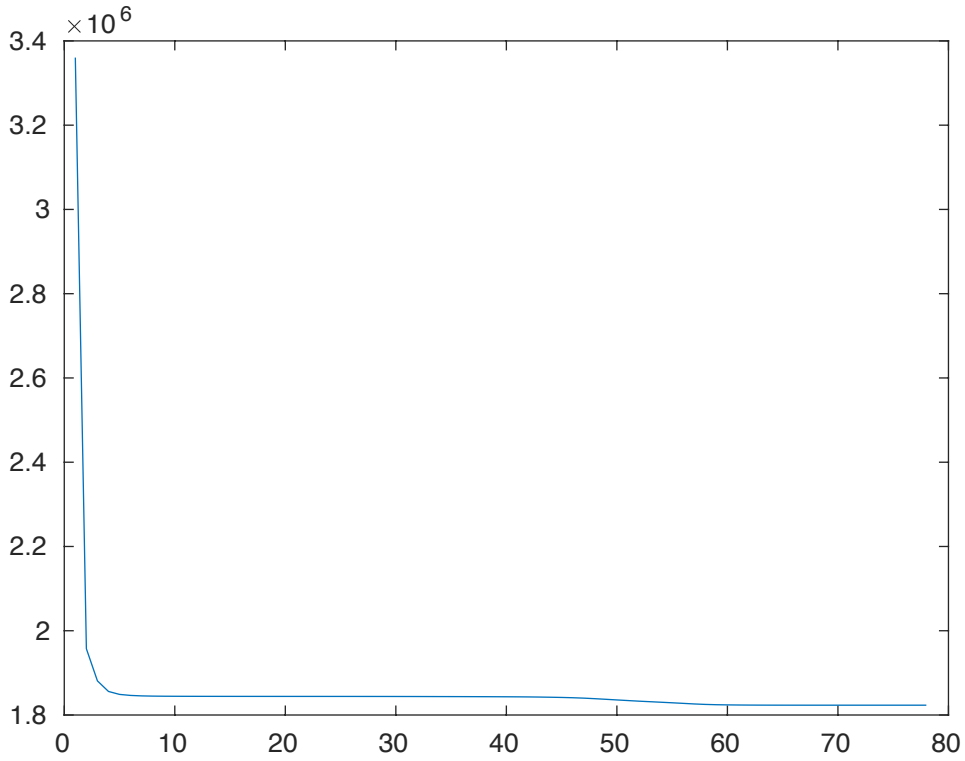


Figure SEE vs number iteration Choosing chunks from kth portions of equal size of dataset

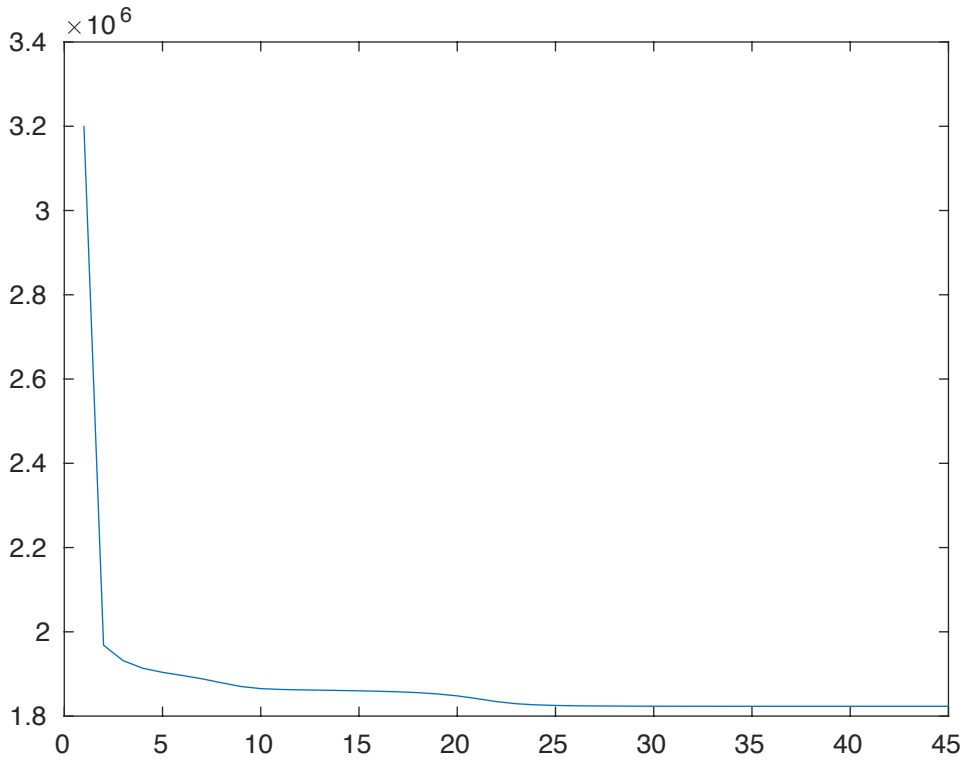


Figure 2 SEE vs number iteration of Choosing chunks kth portions of equal size of dataset

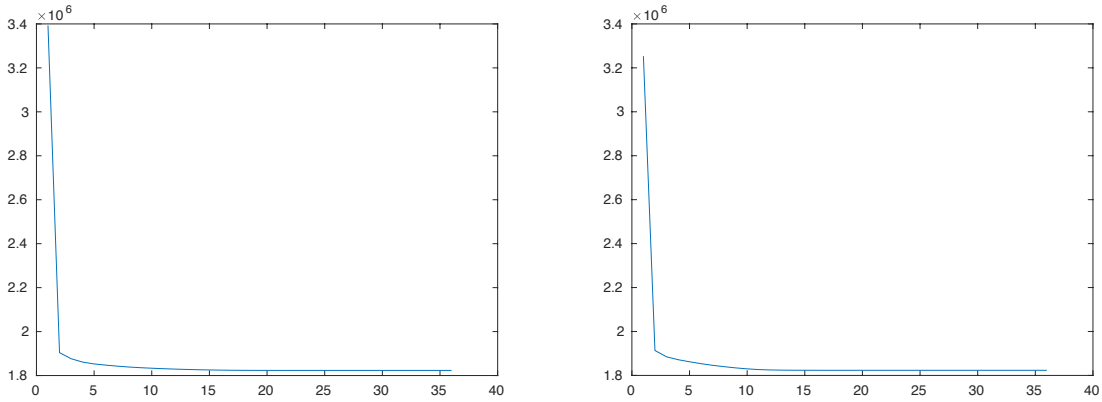


Figure : SEE vs number iteration for random initialization

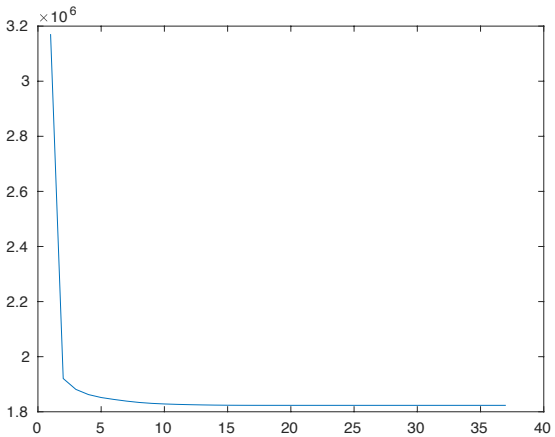


Figure SEE vs number iteration for fixed. Initialization

Discussion:

From the above results, it's clear that random initialization outperformed all other methods. Although my fixed initialization method outperformed my chunking method, it depended mainly on the distribution the dataset. However, without knowing a distribution of your dataset, the chunking method would be better than fixed one as it would better avoid falling in some local minima. In addition, it would perform better -- even better than random initialization -- as a number of k (clusters) increases because the probability for a proximity of initial centres to the converging centre would increase.

Lastly, I did get time to develop the majority voting method from multiple initialization. However, I think it would improve the performance in terms of SSE and accurate clustering depending on how many clustering you make. Hence it would perform better than all those other methods. Except that it could take a lot amount of time if the dataset is too huge.