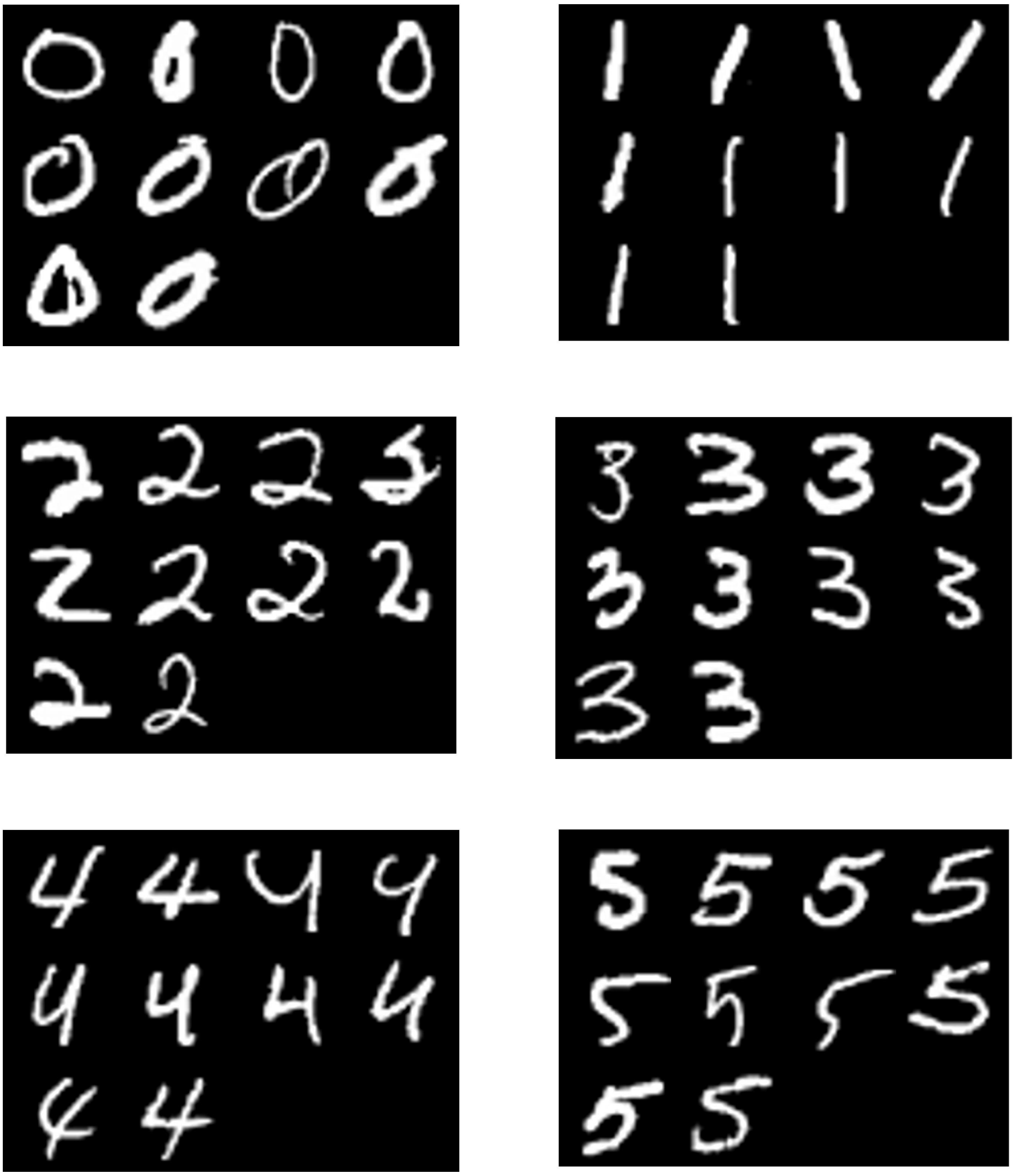
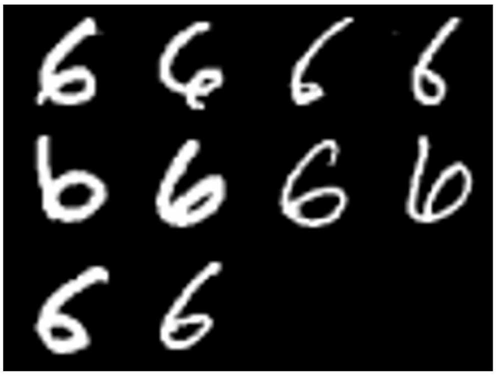
TASK 1 - PCA and Clustering

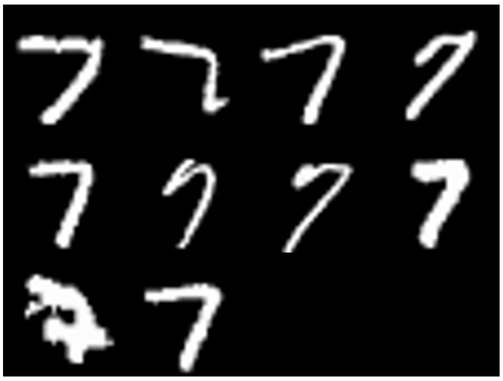
S1701688

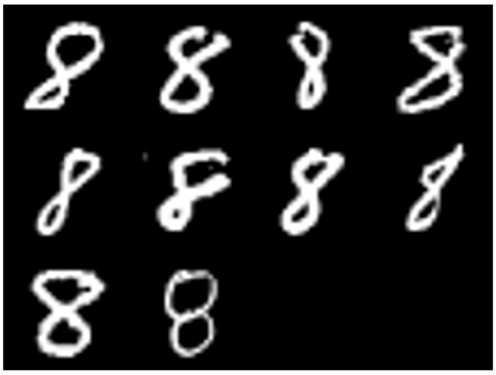
Task 1.1

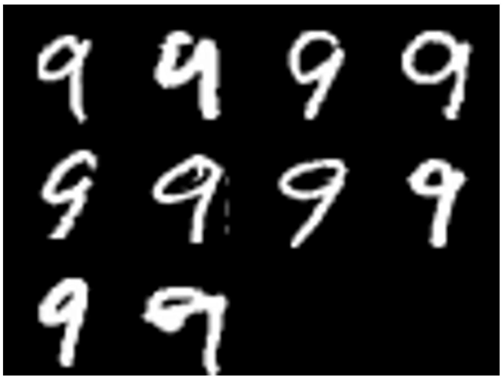
Images of the ten first samples of each class.





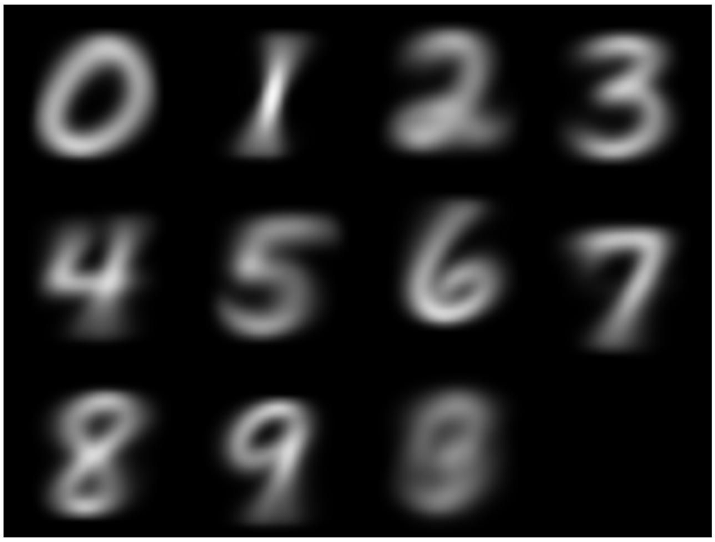






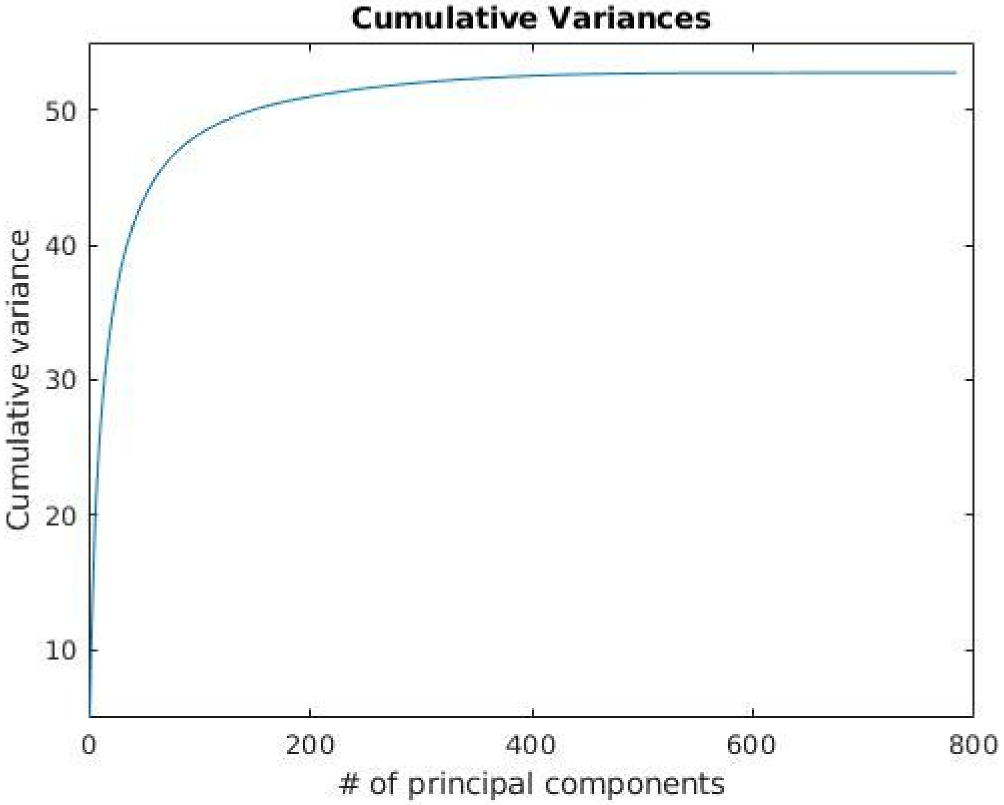
Task 1.2

Image of the 11 mean vectors, where the 11th is the mean of all the classes.



Task 1.3

Cumulative variance figure

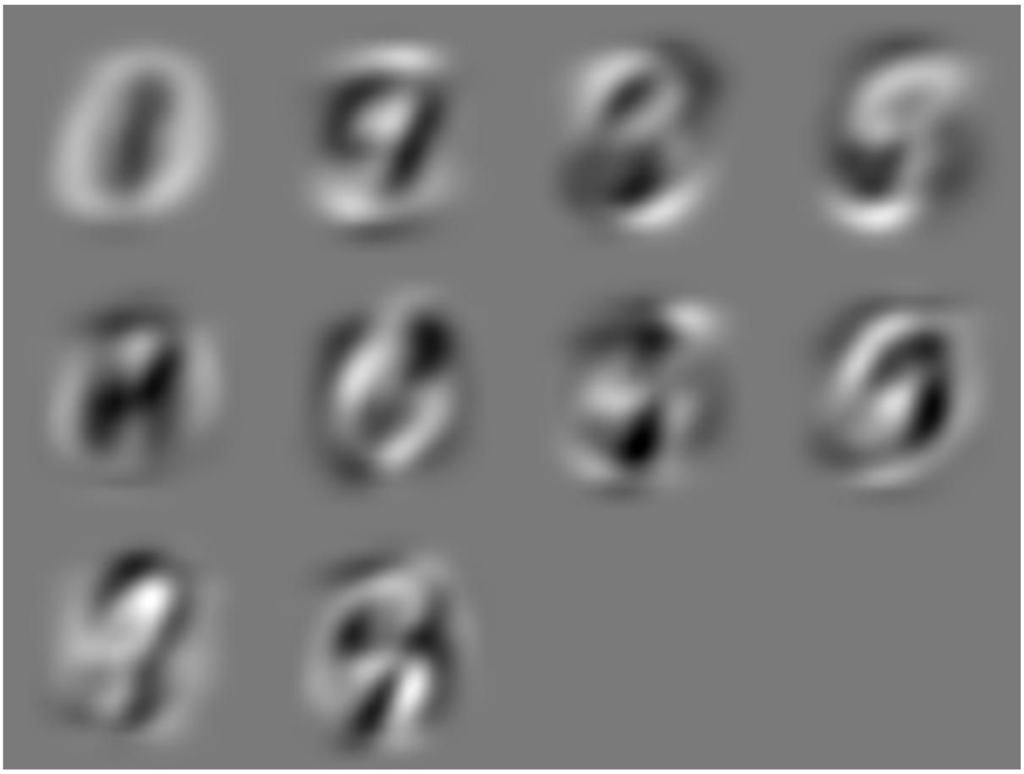


Values of MinDims

43
86
153
784

Task 1.4

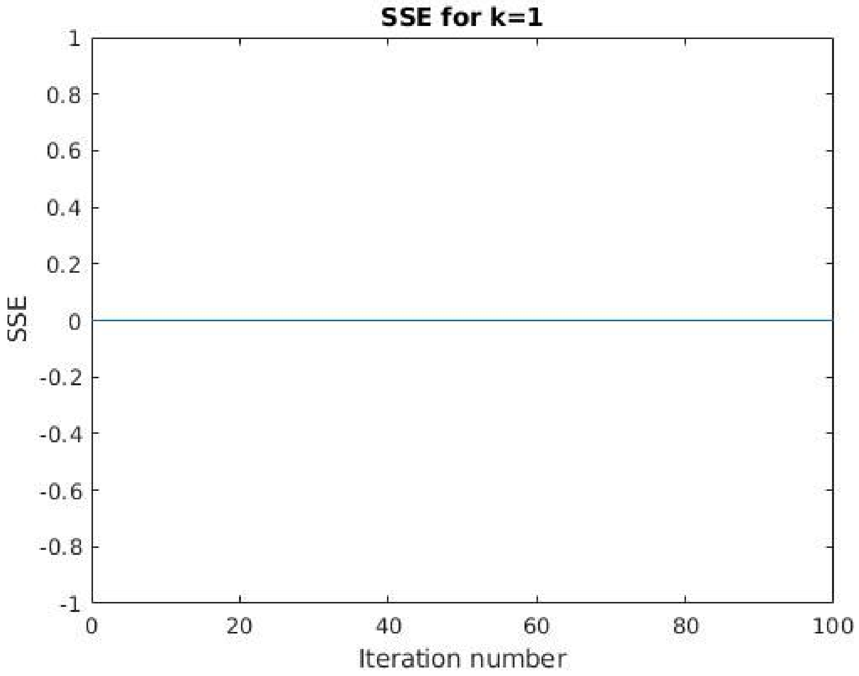
Images of first ten principal axes of PCA



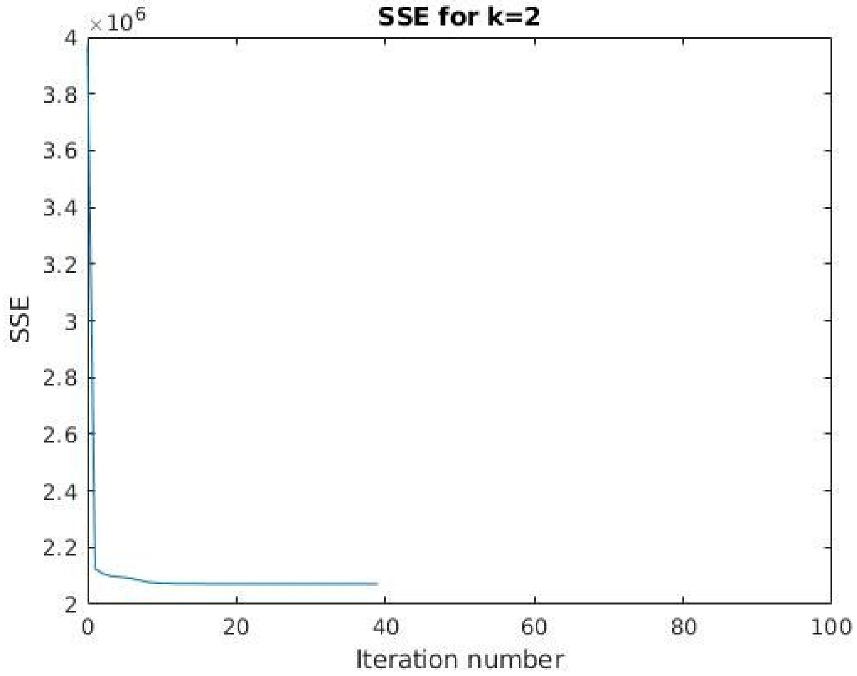
Task 1.5

K-means clustering

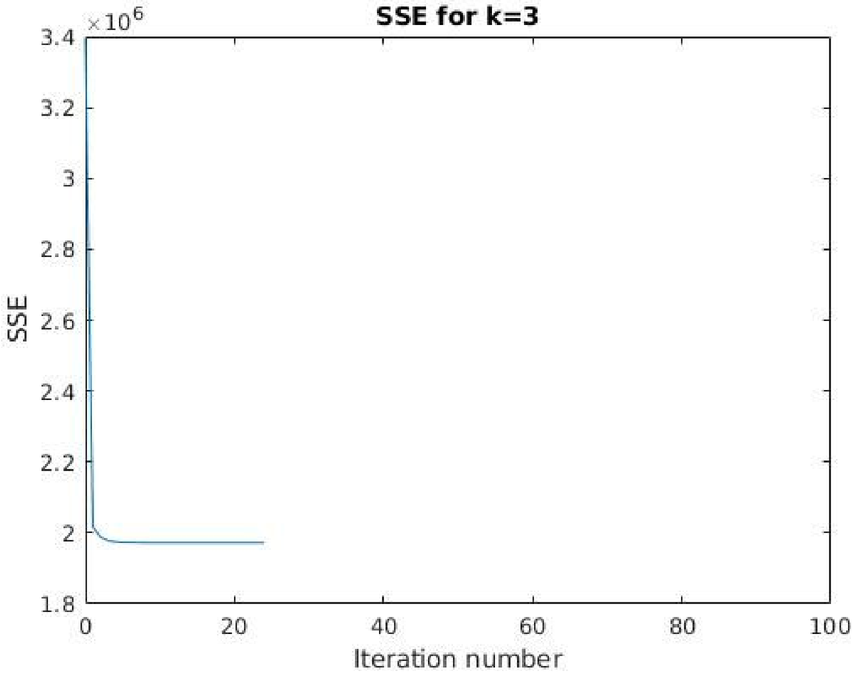
k = 1 0.062663 seconds



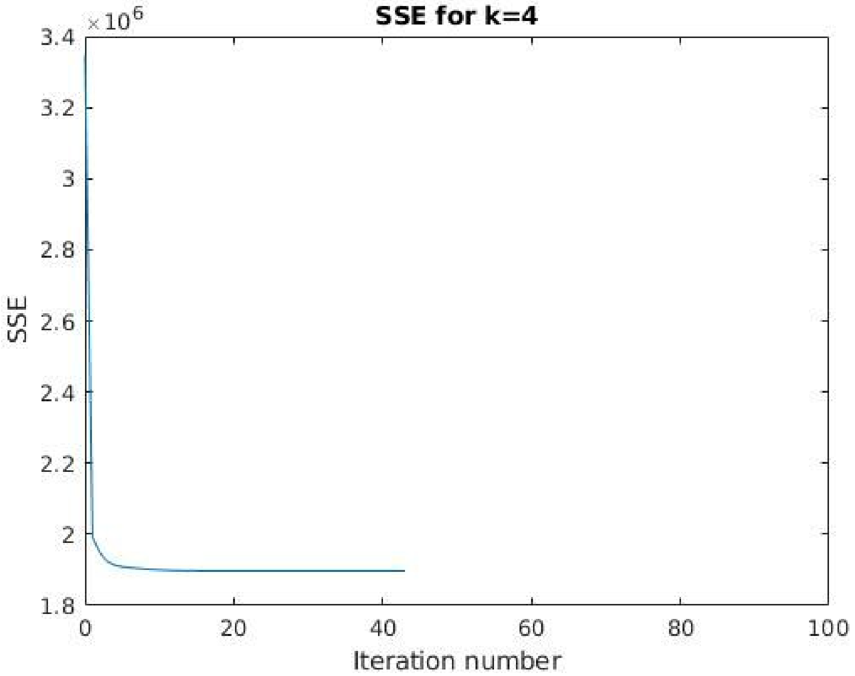
k = 2 13.810594 seconds



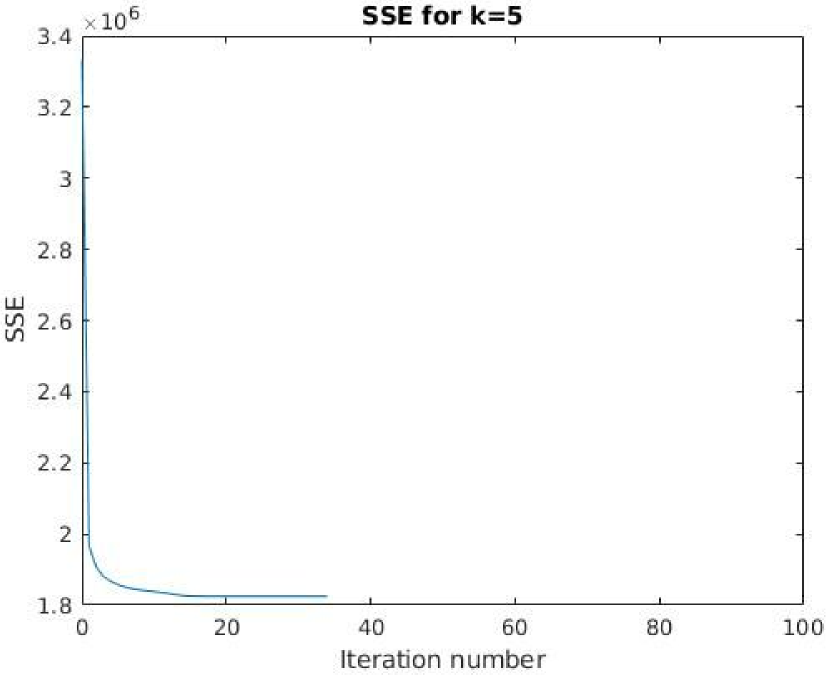
k = 3 11.823958 seconds



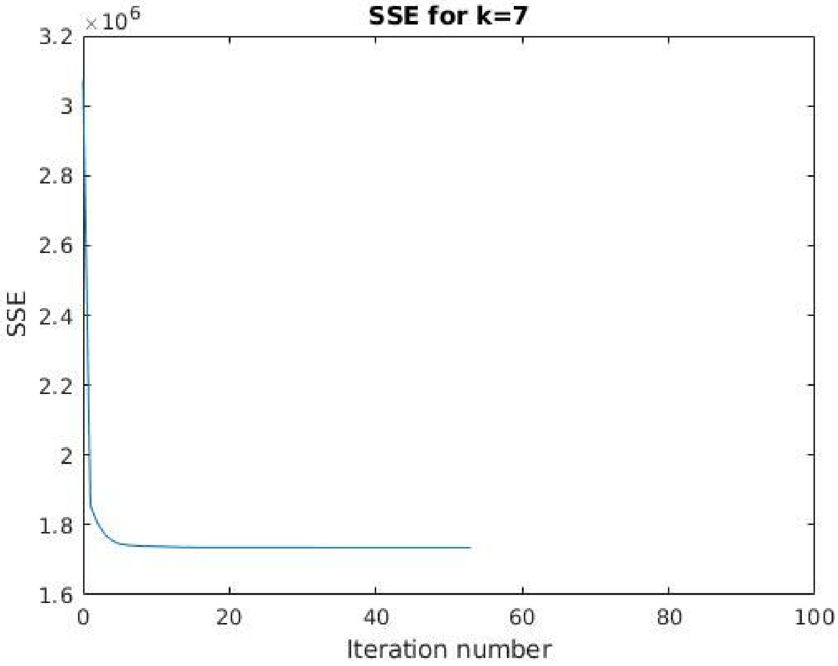
k = 4 25.463011 seconds



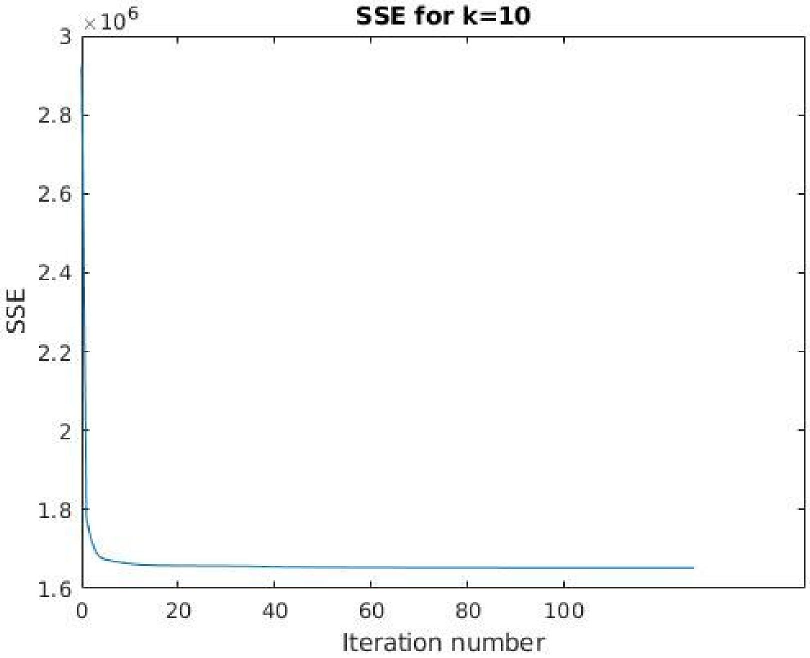
k = 5 23.689972 seconds



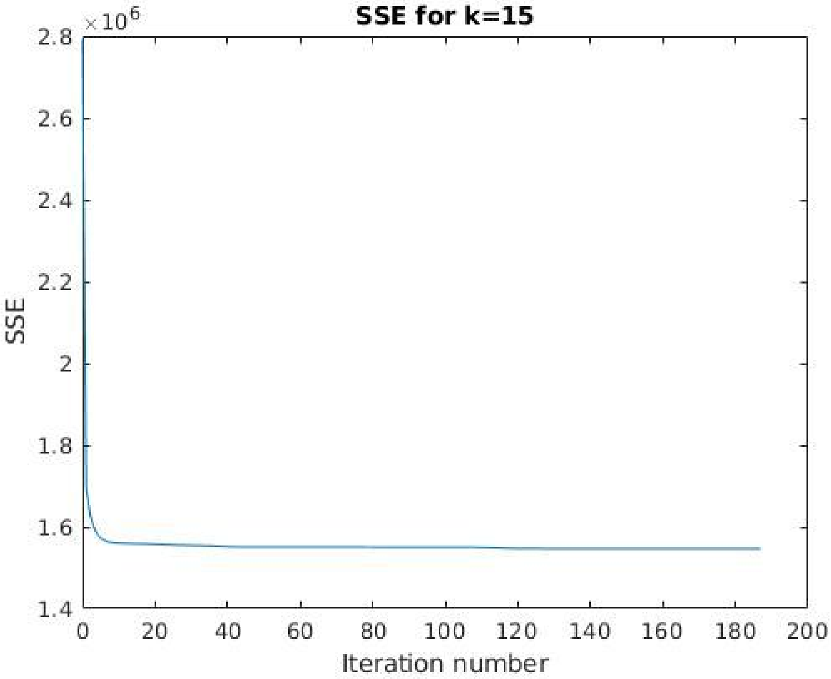
k = 7 47.379904 seconds



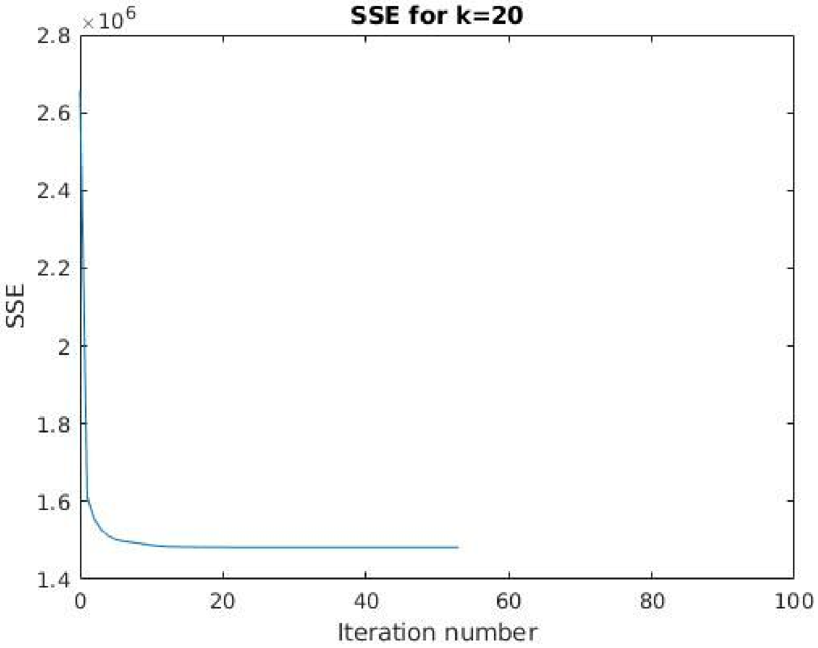
k = 10 152.596886 seconds



k = 15 298.894207 seconds



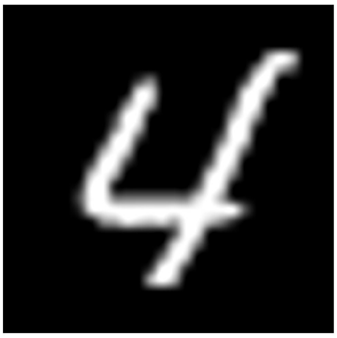
k = 20 121.640869 seconds



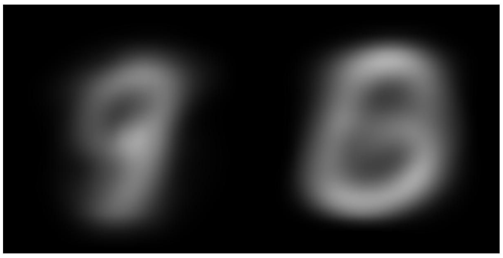
Task 1.6

Image of each cluster centre

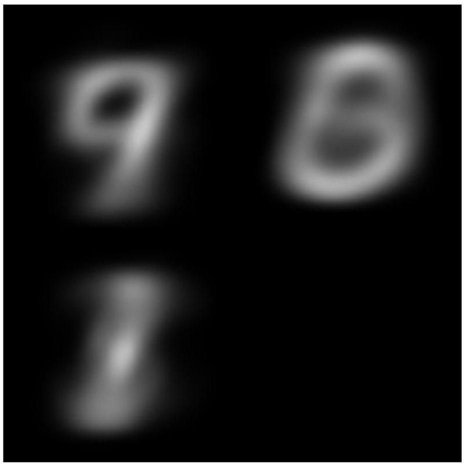
k = 1



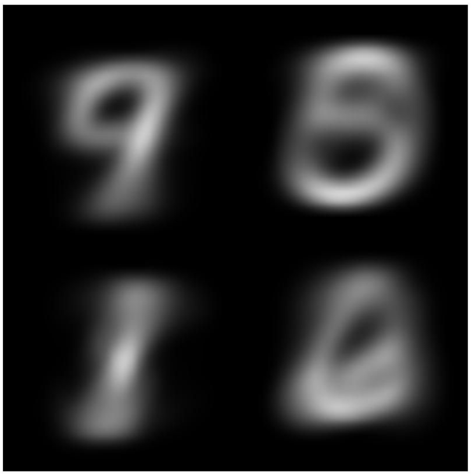
k = 2



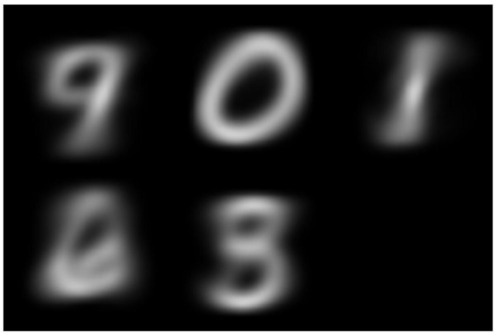
k = 3



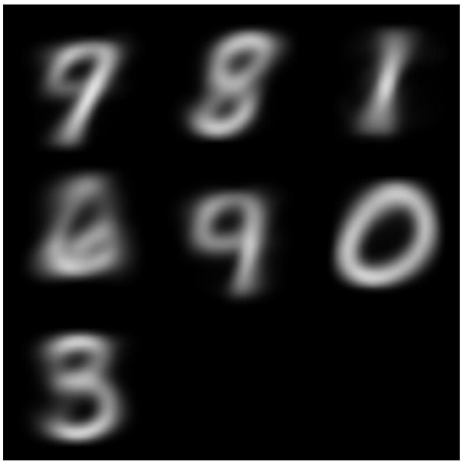
k = 4



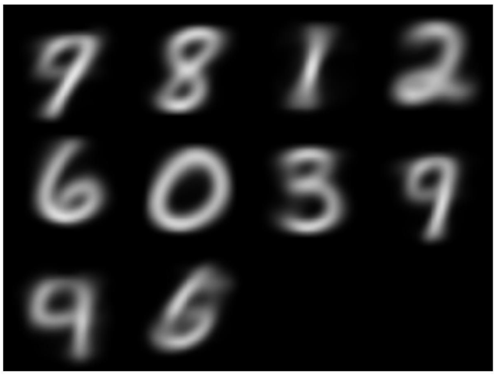
k = 5



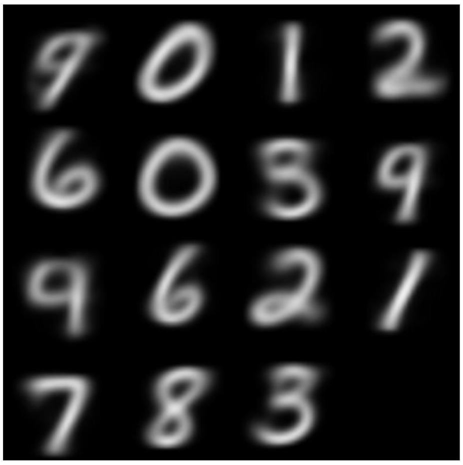
k = 7



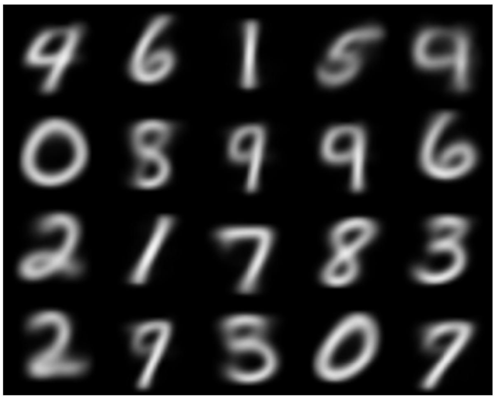
k = 10



k = 15



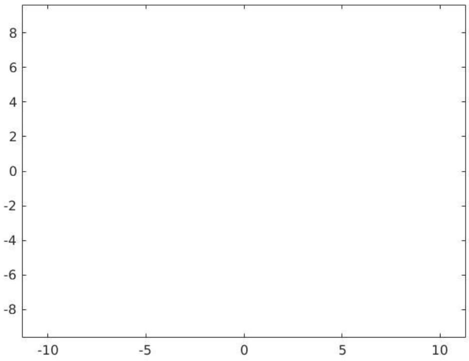
k = 20



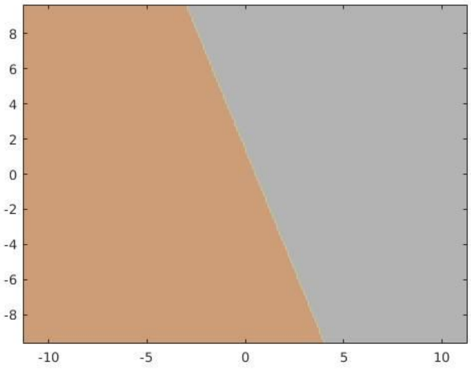
Task 1.7

Cross section images of cluster regions with 2D PCA plane for each class k. In order to visualize the regions, I transformed my D dimensional data into 2 dimensional and graphed a cross section of the D dimensional figure on the 2 dimensional graph. The regions can clearly be differentiated on the figures shown, where each different colour defines a different region. In the case of k = 1, where there is only one cluster centre, it is obvious that all the data points will belong to that cluster centre, which is why the whole cross section is the same colour.

k = 1

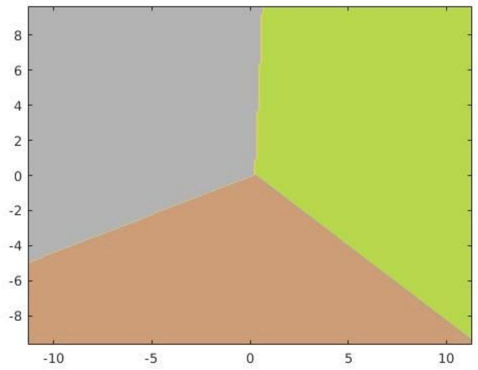


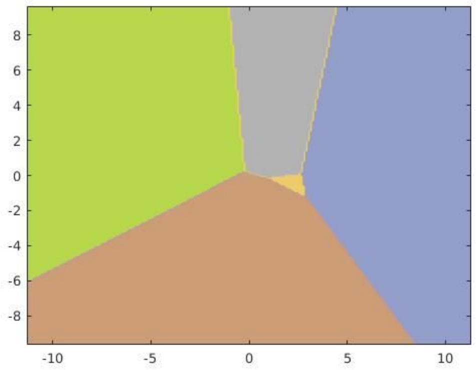
k = 2



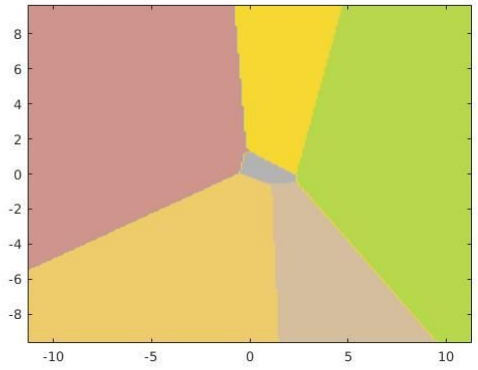
k = 3

k = 5





k = 10



Task 1.8

In general, SSE tends to 0 as the number of clusters k increases.
This can be visualized in the graph 'task1_5_graph.pdf'. Such an observation means that SSE = 0 when the number of clusters k is the same as the number of samples (data points) we have in the dataset, given that in such a situation, each data point is its own cluster and there is no error between said data point and the center of its cluster.

There are various methods of initializing the cluster centres in k-means clustering, where each method affects the overall performance of the clustering.

- Random selection of initial cluster centre points → since the initial cluster centres are randomly allocated, the SSE is very likely to be high. However, at each iteration, where the cluster centres are recalculated using the mean, the

SSE decreases. In general, when using k-means clustering, minimizing the SSE is the aim.

- Group representative cluster centre points → select the k most representative centres from the dataset, where the first cluster centre is the closest to the dataset centroid. The rest of the (k-1) centres are selected by considering if said centres are closer to a set of points than each of these is to any of the already existing cluster centres of the dataset.
- Farthest points → first cluster centre is selected randomly from the provided dataset. Second centre is selected as the point that is the most distant from that first cluster centre. The third centre is a point that is far from both of the previous centres, and so on.
- Random farthest points / k-means++ → first cluster centre is selected randomly from the dataset. n-th centre is selected randomly, but the probability of selection is proportional to the distance to the nearest (n-1) centres.