Report Task 1

Task 1.1

Figure 1.1.1 Class 1

Figure 1.1.2 Class 2

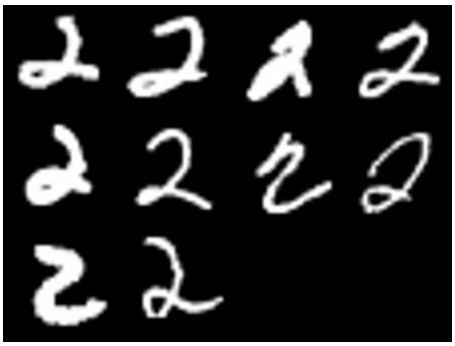


Figure 1.1.3 Class 3

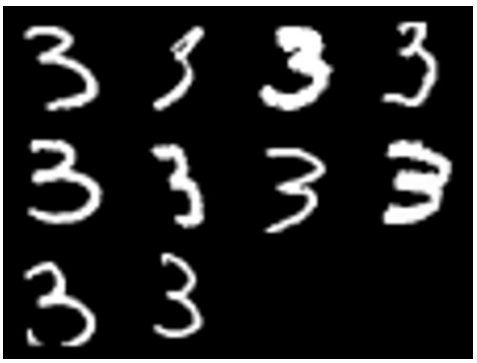


Figure 1.1.4 Class 4

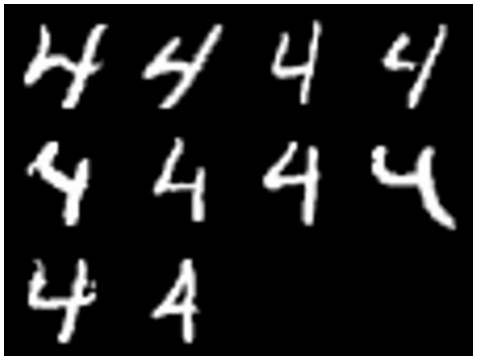


Figure 1.1.5 Class 5

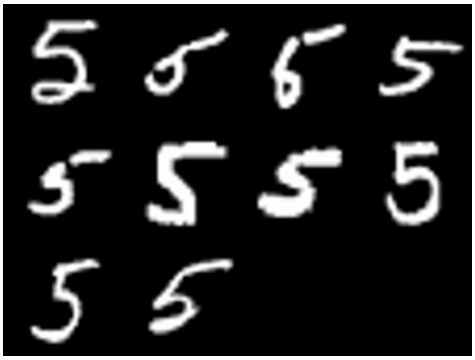


Figure 1.1.6 Class 6

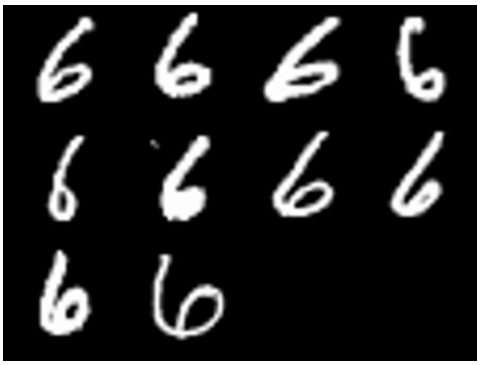


Figure 1.1.7 Class 7

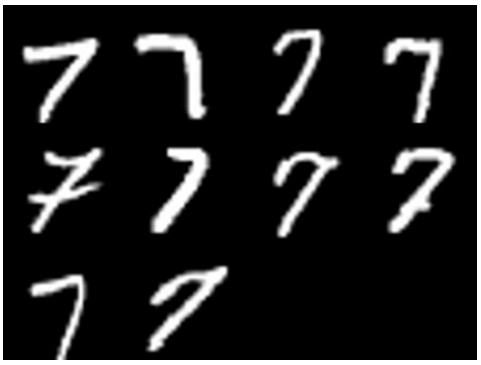


Figure 1.1.8 Class 8

Figure 1.1.9 Class 9

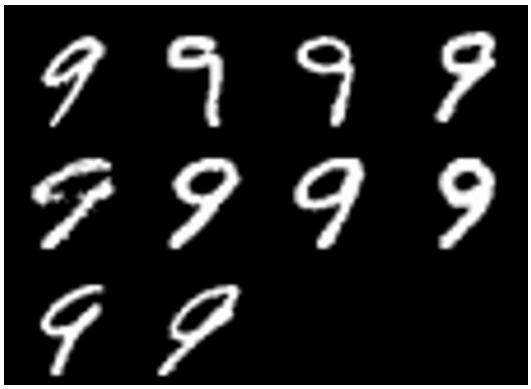


Figure 1.1.10 Class 10

Report Task 1

Task 1.2

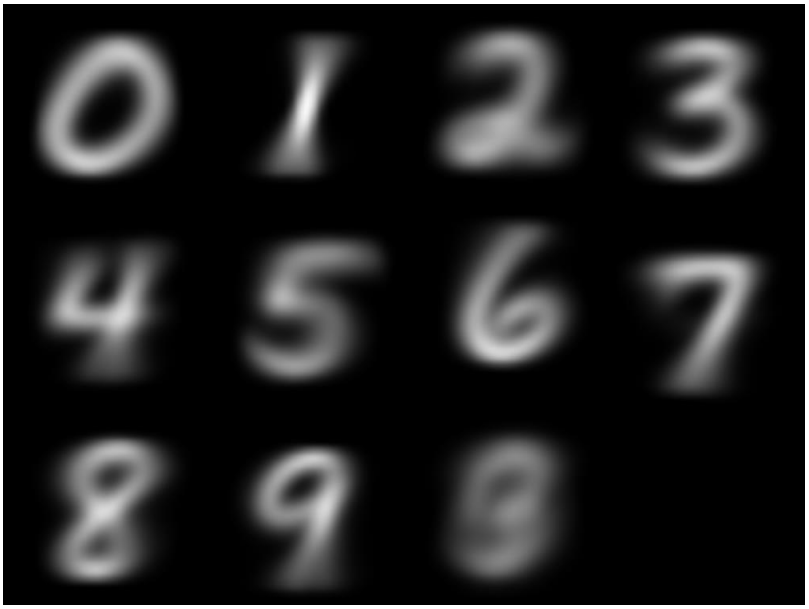


Figure 1.2: Images of the mean vectors for Class 1 to Class 10

Task 1.3

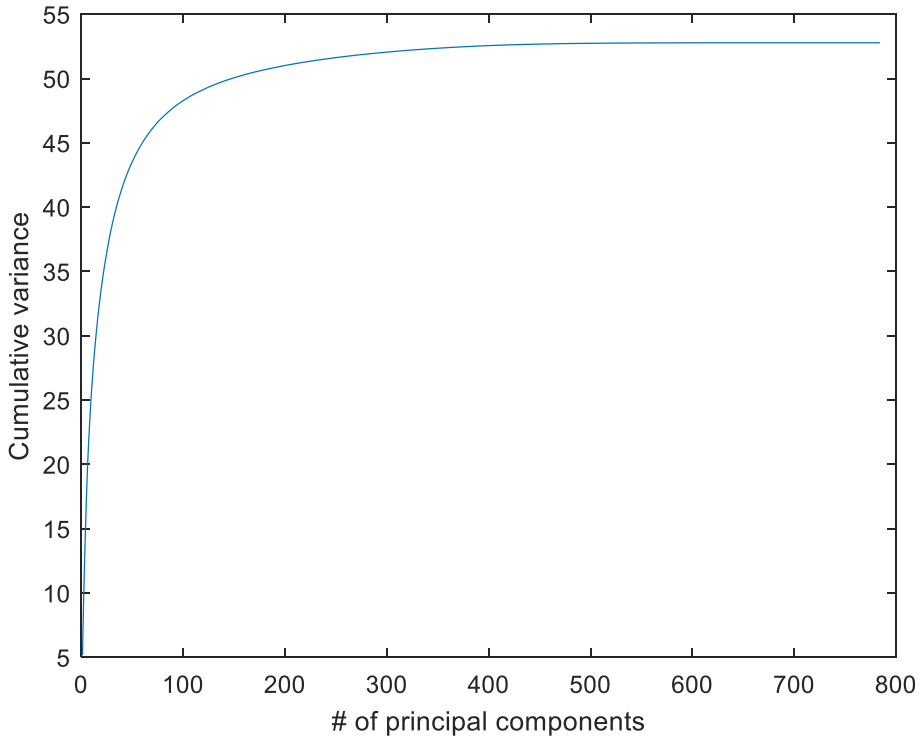


Figure 1.3: Cumulative variance

Values of MinDims (4x1) = [1; 1; 1; 1;]

Report Task 1

Task 1.4

Figure 1.4: Images of the first ten principal components

Task 1.5

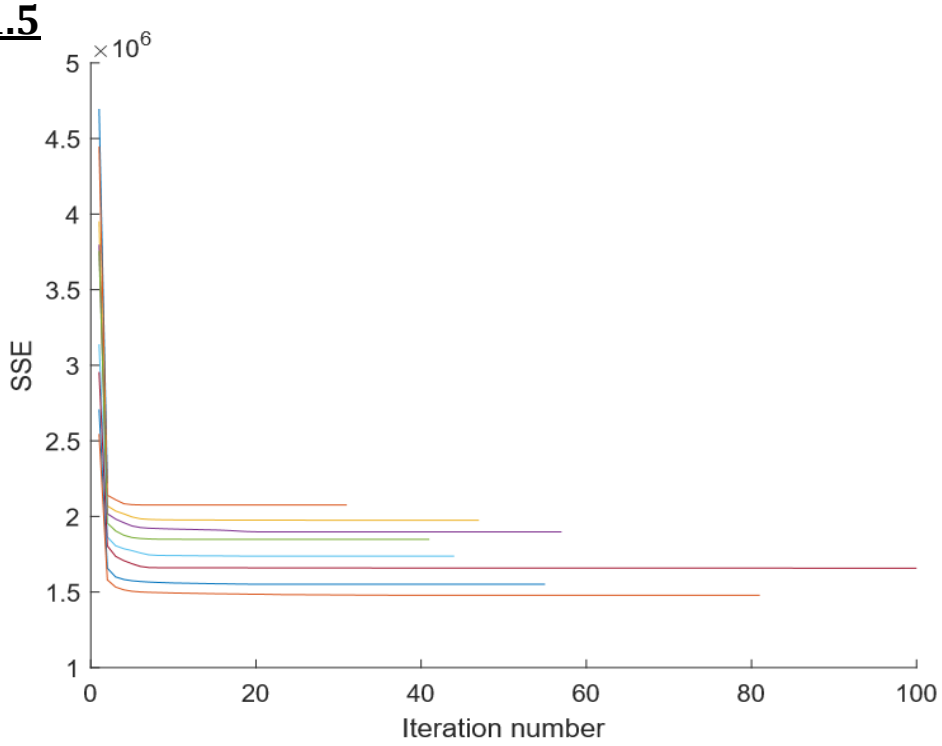


Figure 1.5: SSE vs iteration number for all k in Ks

k = 1: Elapsed time is 0.249 seconds.

k = 2: Elapsed time is 15.222 seconds.

k = 3: Elapsed time is 27.399 seconds.

k = 4: Elapsed time is 38.906 seconds.

k = 5: Elapsed time is 33.079 seconds.

k = 7: Elapsed time is 47.474 seconds.

k = 10: Elapsed time is 134.692 seconds.

k = 15: Elapsed time is 103.899 seconds.

k = 20: Elapsed time is 196.980 seconds.

Report Task 1

Task 1.6

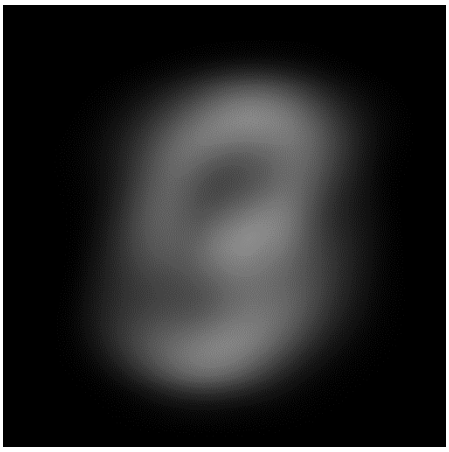


Figure 1.6.1: Cluster Centre k=1

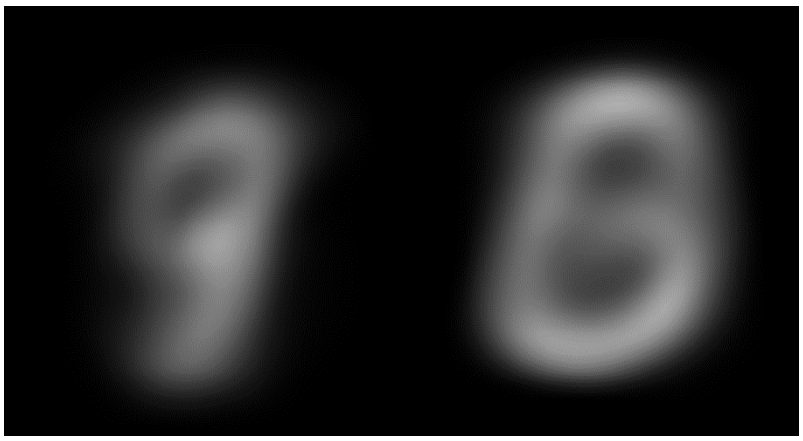


Figure 1.6.2: Cluster Centre k=2

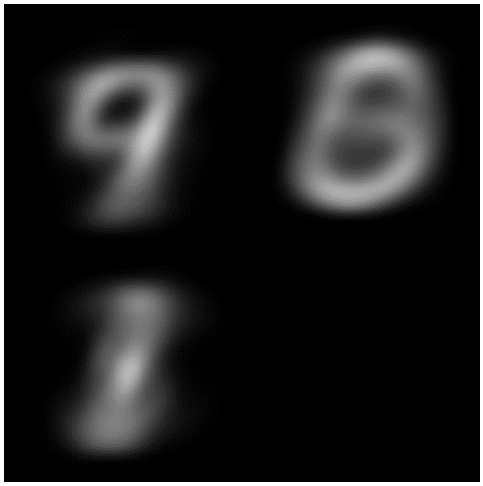


Figure 1.6.3: Cluster Centre k=3

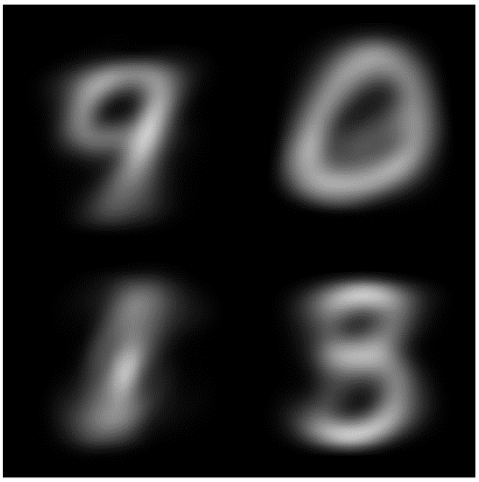


Figure 1.6.4: Cluster Centre k=4

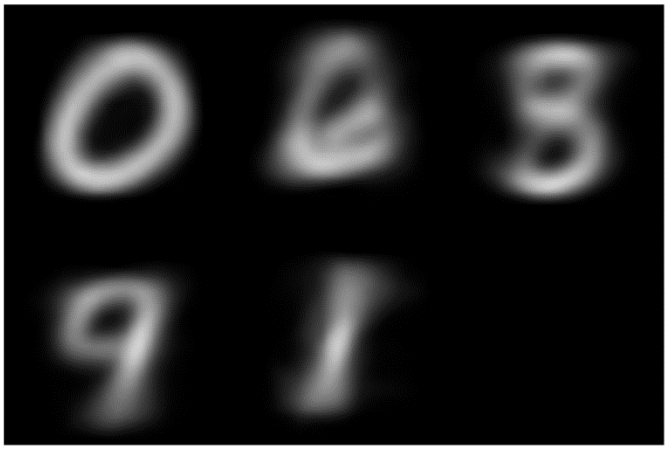


Figure 1.6.5: Cluster Centre k=5

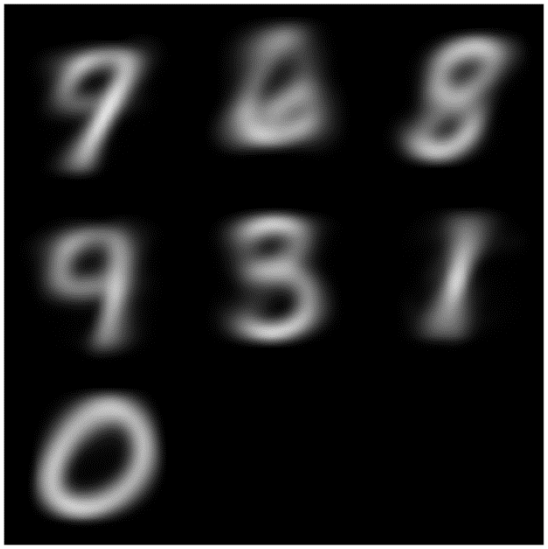


Figure 1.6.6: Cluster Centre k=7

Report Task 1

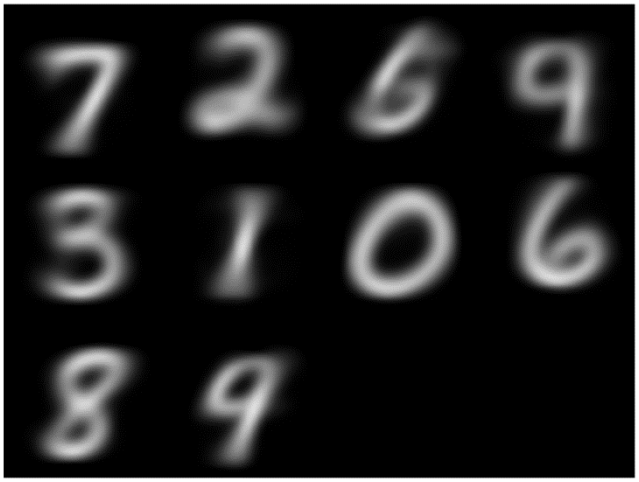


Figure 1.6.7: Cluster Centre k=10

Figure 1.6.8: Cluster Centre k=15

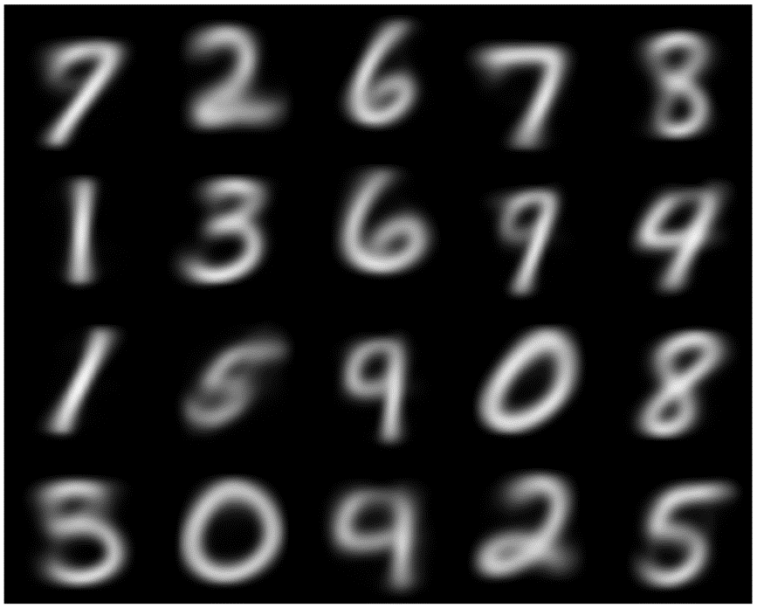


Figure 1.6.9: Cluster Centre k=20

Report Task 1

Task 1.7

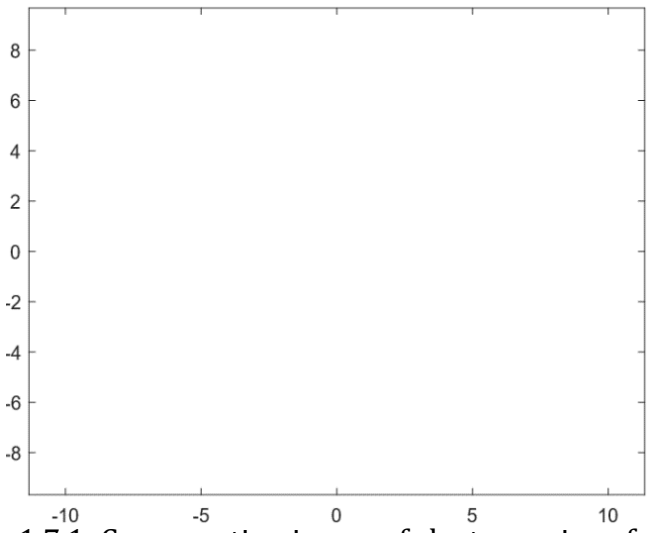


Figure 1.7.1: Cross-section image of cluster regions for k=1

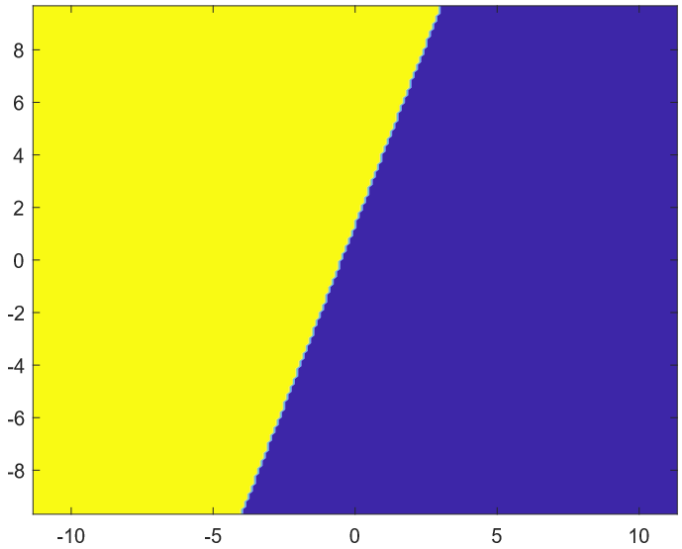


Figure 1.7.2: Cross-section image of cluster regions for k=2

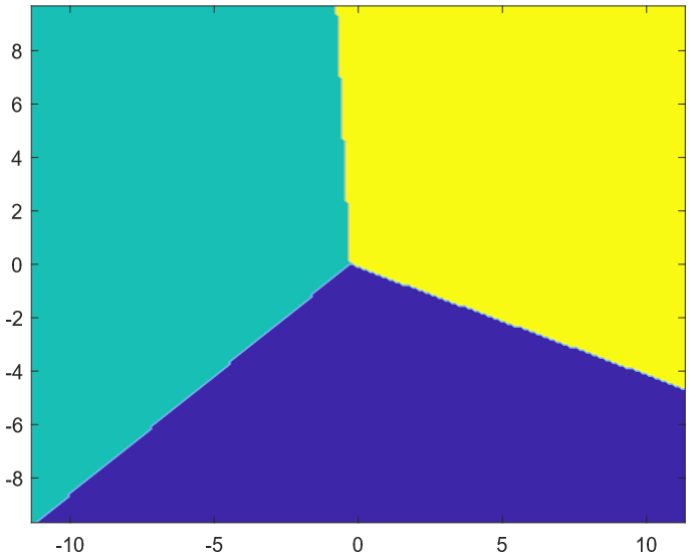


Figure 1.7.3: Cross-section image of cluster regions for k=3

Report Task 1

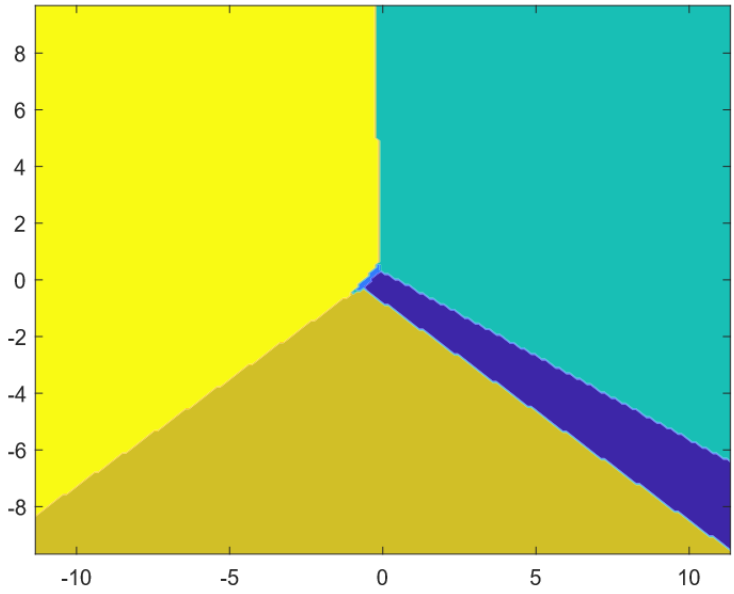


Figure 1.7.4: Cross-section image of cluster regions for k=5

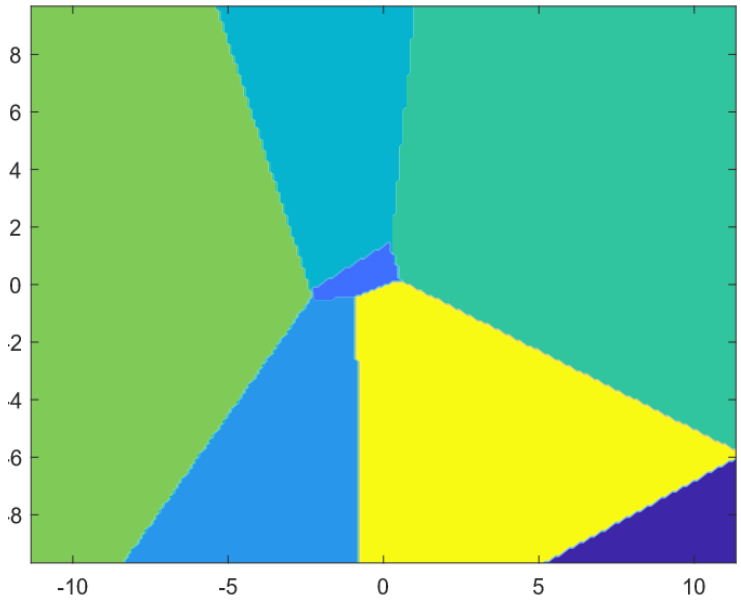


Figure 1.7.5: Cross-section image of cluster regions for k=10

These regions were visualised by using the two most significant eigenvectors from the covariance matrix (calculated by using comp_pca) as basis vectors for a plane in 2D.

The range was then set to m ± 5σ and a nbins by nbins grid was created. This plane was then projected to using the formula provided on Additional notes: Task 1.7 (solving for x). Finally k-means was applied to this plane and coloured according to corresponding clusters.

Task 1.8

This function has the definition `function task1_8(X, Ks, initialOption)`
With `initialOption` being either 0, 1 or 2. For a total of 3 different cluster initialisation methods.

Report Task 1

Figure 1.8.1: initialOption = 0

Random Initialisation: this is when k observations are sampled uniformly at random, without replacement, from the data in Xtrn.

>> task1_8(Xtrn,5,0) *k = 5* *Elapsed time is 28.861 seconds.*

>> task1_8(Xtrn,4,0) *k = 4* *Elapsed time is 21.100 seconds.*

>> task1_8(Xtrn,3,0) *k = 3* *Elapsed time is 40.227 seconds.*

Average SSE Graph for k=5

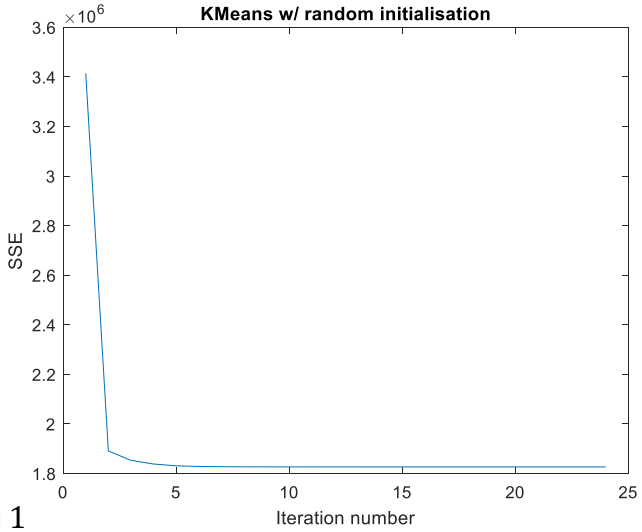


Figure 1.8.2: initialOption = 1

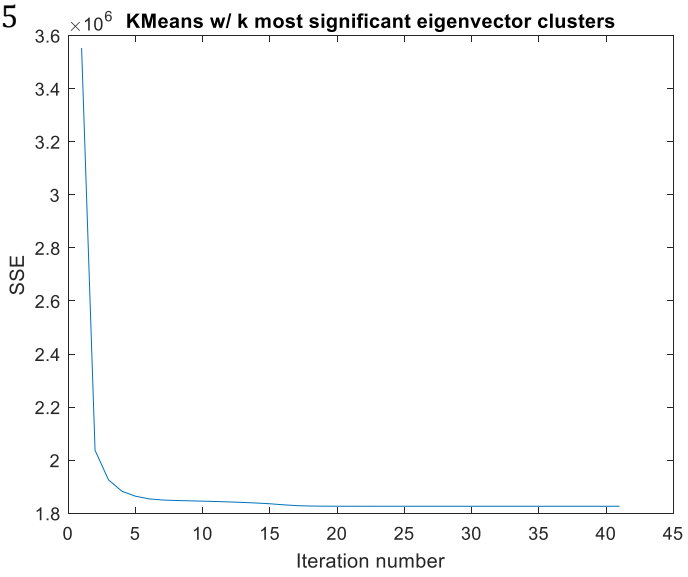
Eigen Initialisation: Computes PCA, and uses the first k most significant eigenvectors as the init' cluster centres, similar to what was done in task1_3.

>> task1_8(Xtrn,5,1) *k = 5* *Elapsed time is 65.017 seconds.*

>> task1_8(Xtrn,4,1) *k = 4* *Elapsed time is 65.017 seconds.*

>> task1_8(Xtrn,3,1) *k = 3* *Elapsed time is 38.343 seconds.*

Average SSE Graph for k=5



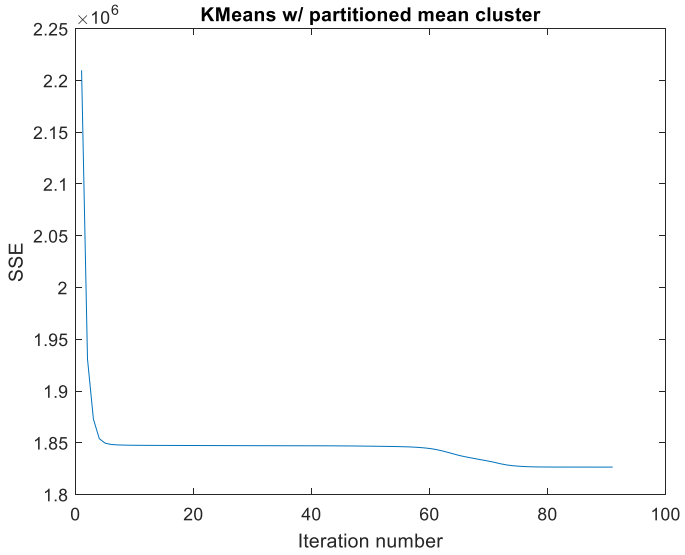
Report Task 1

Figure 1.8.3: initialOption = 2

Partition Mean Initialisation: this method involves splitting the dataset into k partitions of equal size, calculating the mean within those partition and using those as the initial centres.

>> task1_8(Xtrn,5,2)	k = 5	Elapsed time is 106.560242 seconds.
>> task1_8(Xtrn,4,2)	k = 4	Elapsed time is 77.974734 seconds.
>> task1_8(Xtrn,3,2)	k = 3	Elapsed time is 38.939352 seconds.

Average SSE Graph for k=5



Note that each of the experiments where repeated 3 times and that the elapsed time is an average (to 3 s.f.) for the 3 trials.

In conclusion,

The cluster initialisations for k-means can have an impact on the rate at which the algorithm converges, however the solution the algorithm finds does not seem to be very heavily impacted by the initial choices. This means we are better off choosing an initialisation method at results in lower runtimes, especially for higher values of k such as random (compared to partitioned means).