Task 1.1 S1891130

figure 1.1.1: Class1

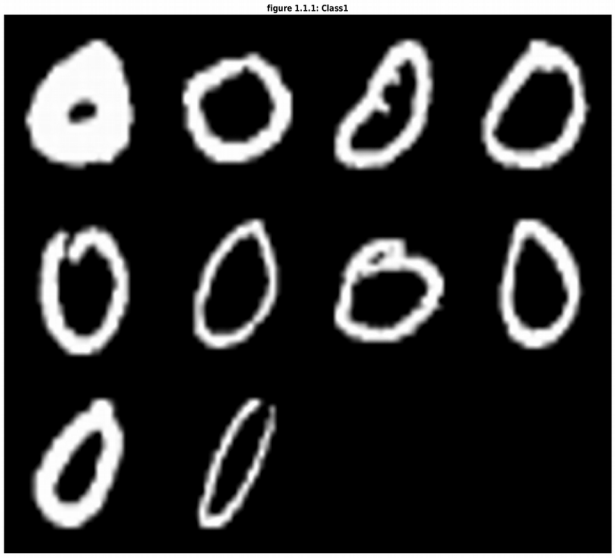


figure 1.1.2: Class2

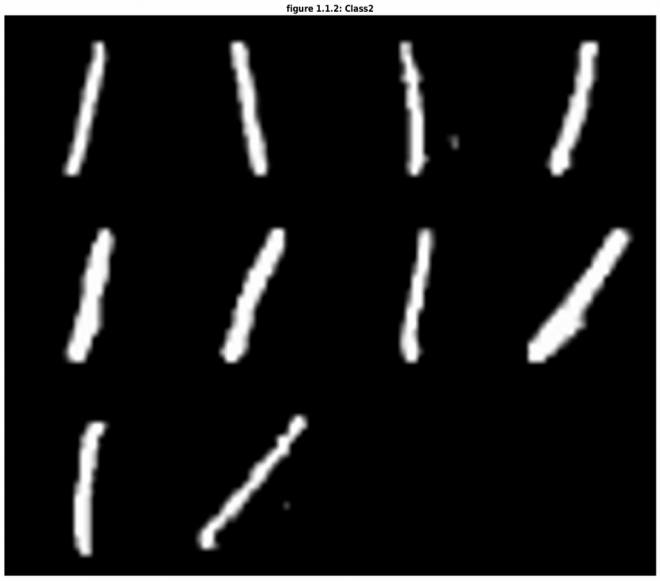


figure 1.1.3: Class3

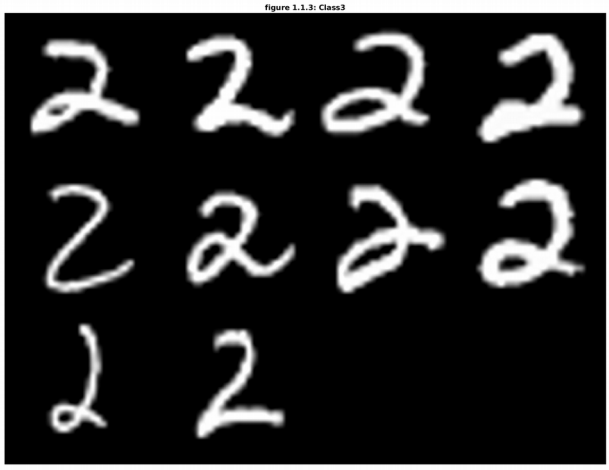


figure 1.1.4: Class4

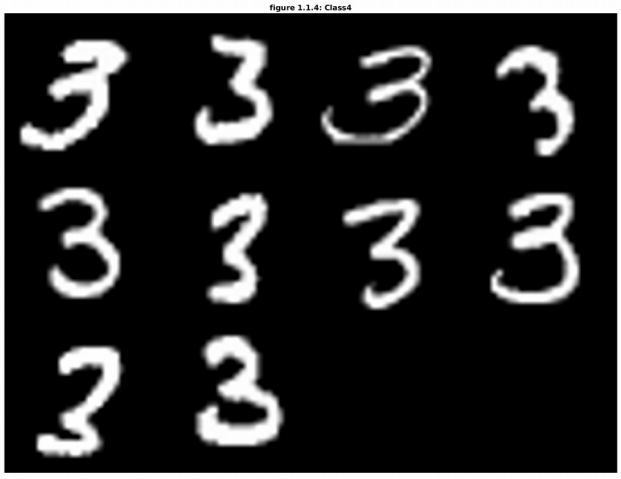


figure 1.1.5: Class5

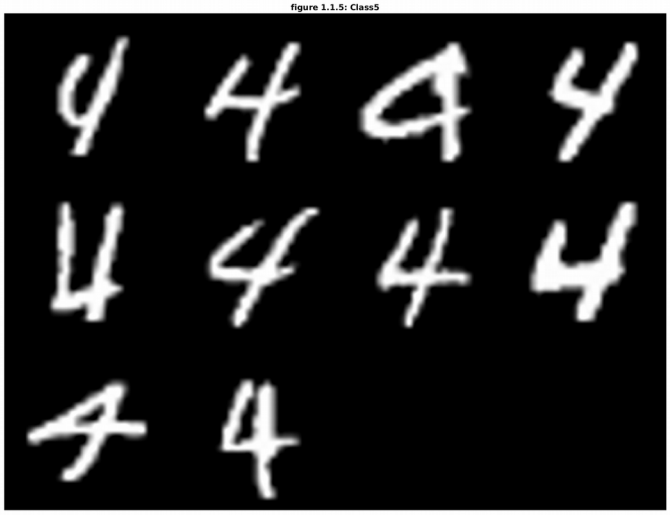


figure 1.1.6: Class6

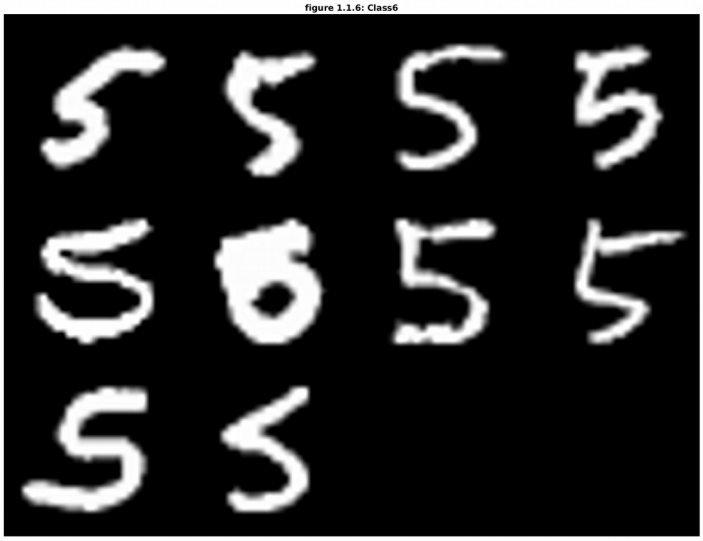


figure 1.1.7: Class7

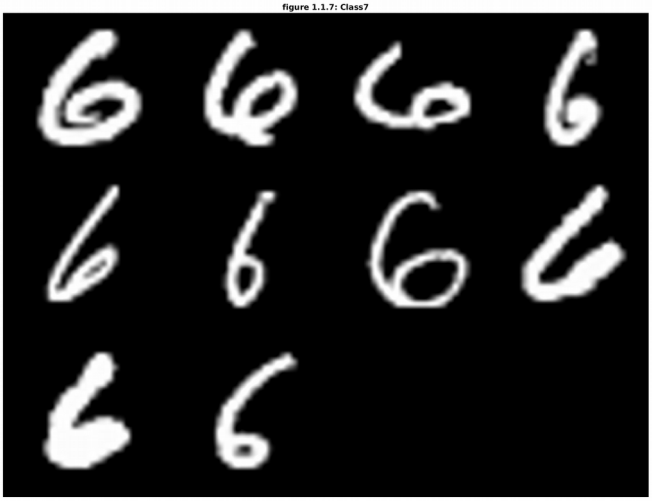


figure 1.1.8: Class8

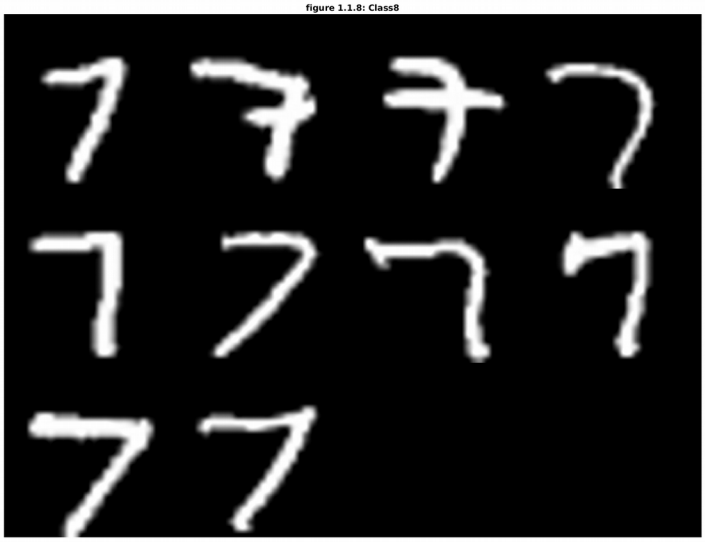


figure 1.1.9: Class9

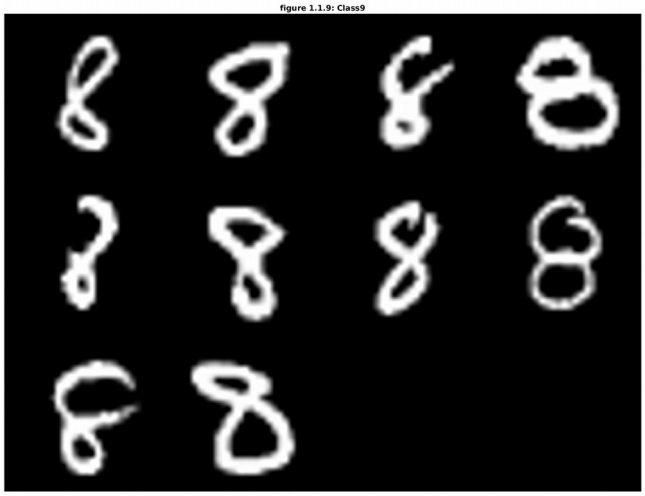
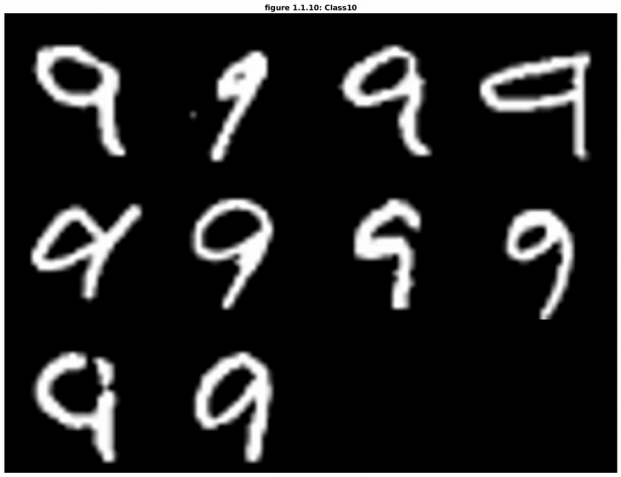
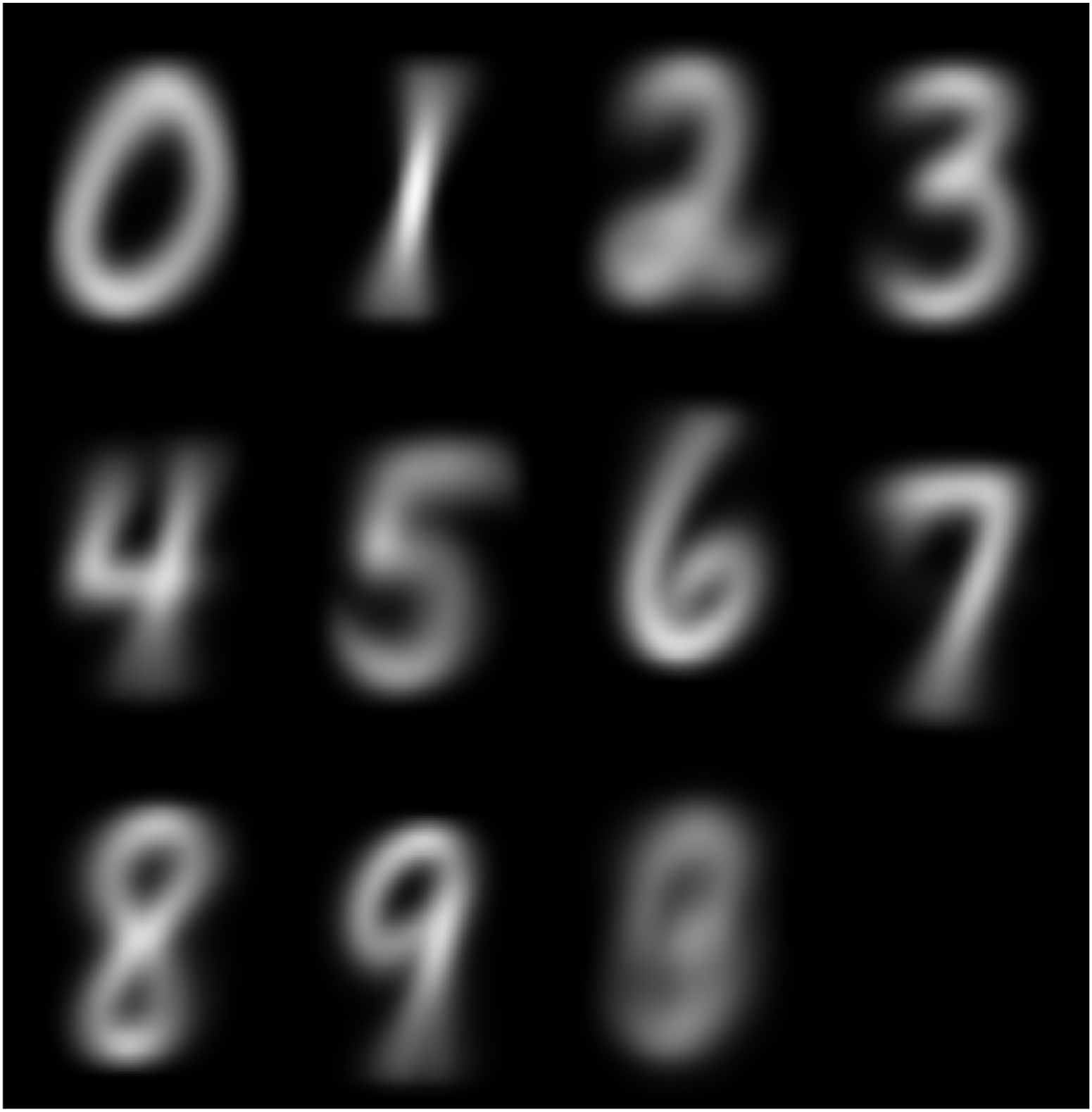


figure 1.1.10: Class10



Task 1.2

figure1.2:mean vectors



task1.3

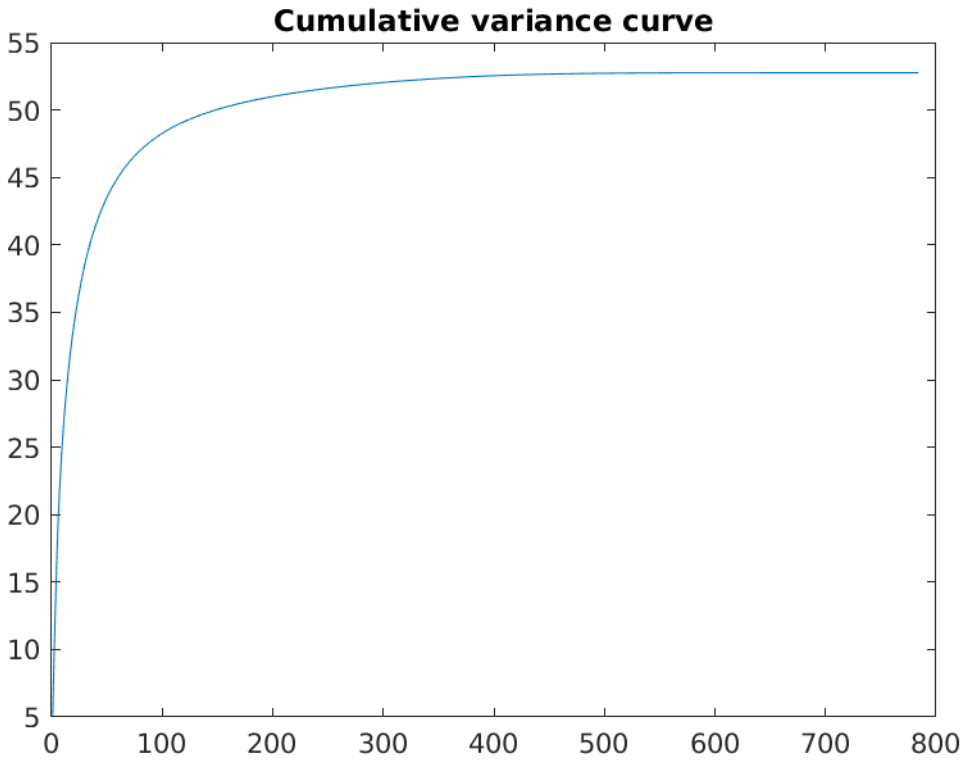


Figure1.3

Covered variance	
70%	26
80%	44
90%	87
95%	154

Task1.4

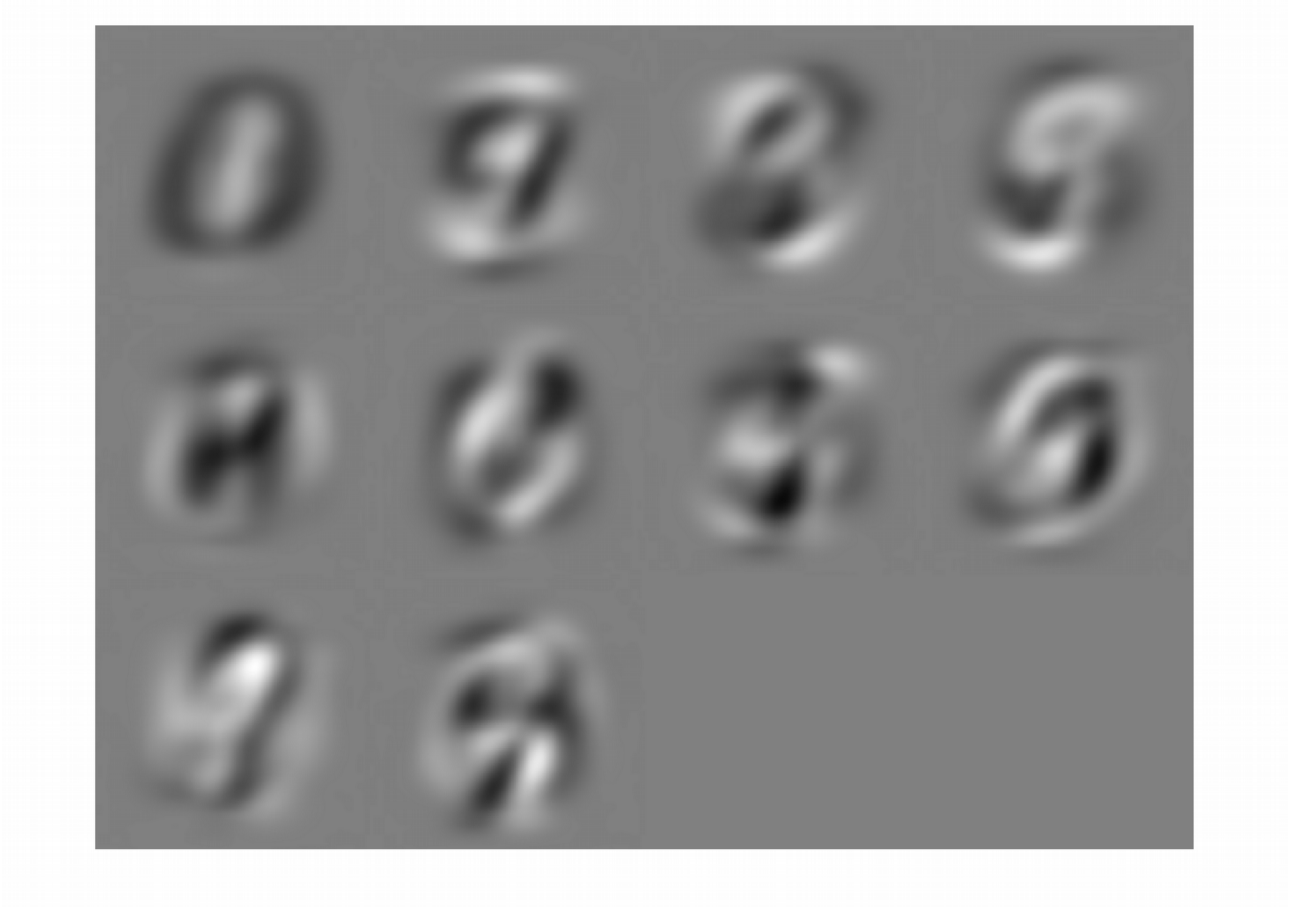


Figure 1.4 First 10 principal components

Task1.5

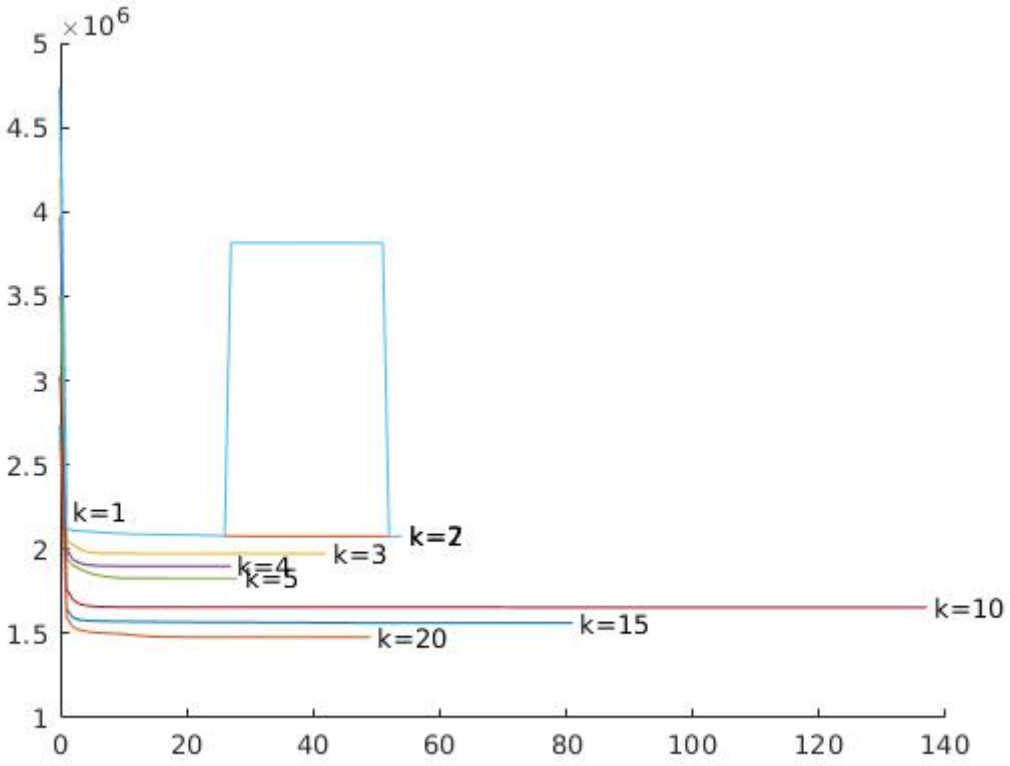


Figure 1.5 SSE vs iteration plot

k	1	2	3	4	5	7	10	15	20
T/s	1.520	40.69	33.23	22.44	24.21	97.36	131.7	79.51	51.77

#The time are calculated using ‘tic-toc’ function of matlab

Task 1.6

Figure 1.6.1Image of cluster centres for k = 1

Figure 1.6.3Image of cluster centres for k = 3

K = 1

K = 3

Figure 1.0.2Image of cluster centres for K = 2

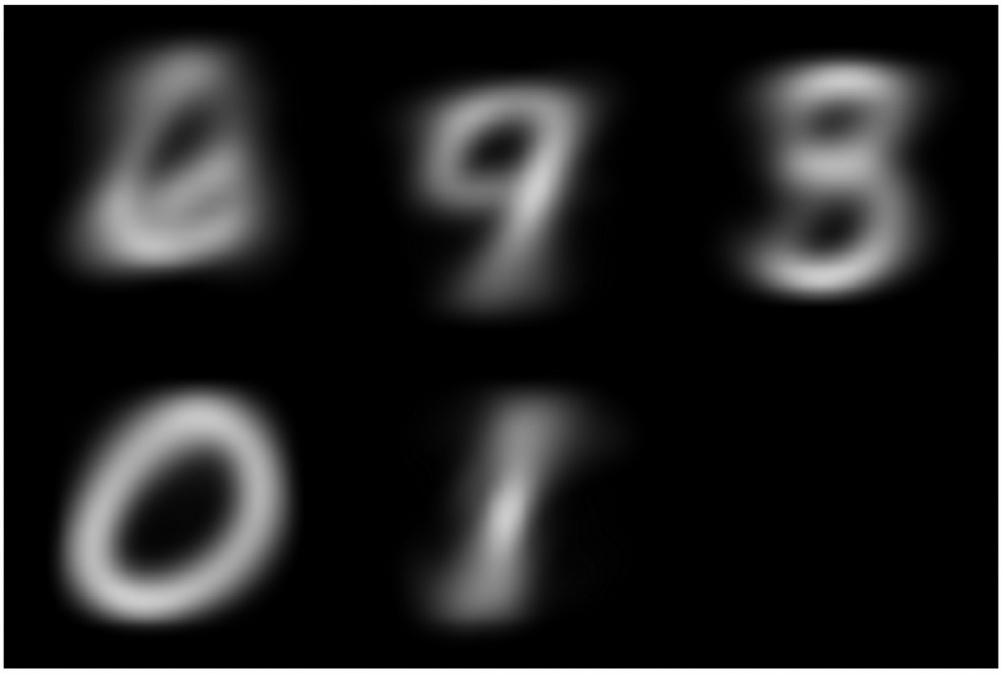
K = 2

Figure 1.0.7Image of cluster centres for K =

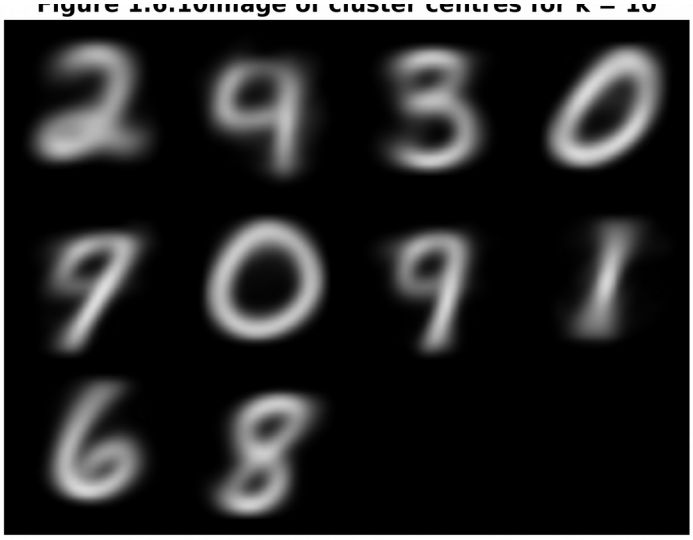
Figure 1.0.4Image of cluster centres for K =

K = 4

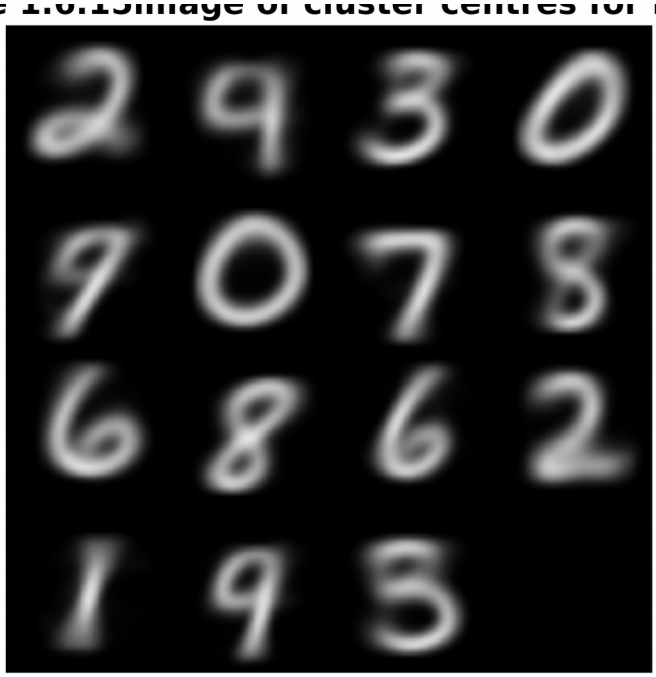
K = 7



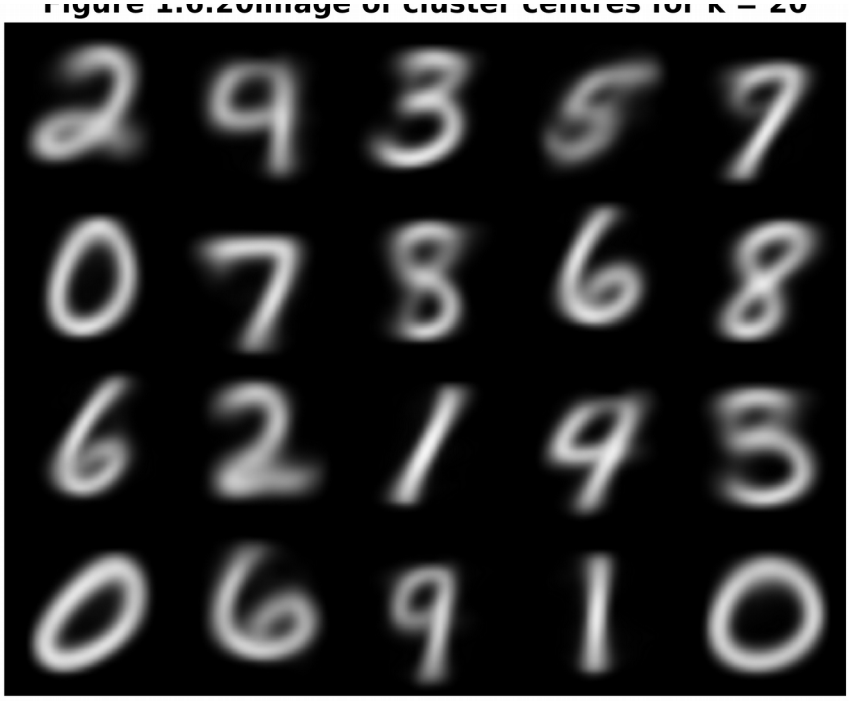
K = 5



K = 10



K = 15



K = 20

Task 1.7

Figure 1.7.1cluster region of k=1

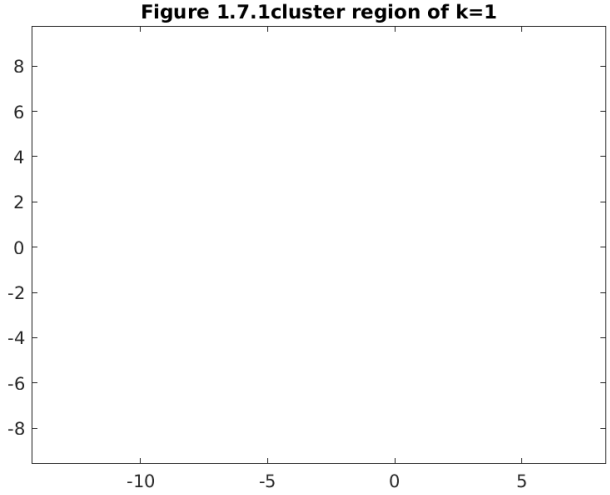


Figure 1.7.2cluster region of k=2

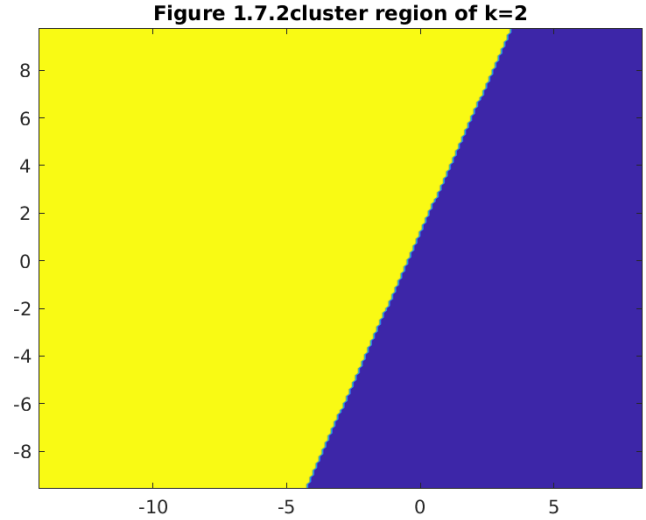


Figure 1.7.3cluster region of k=3

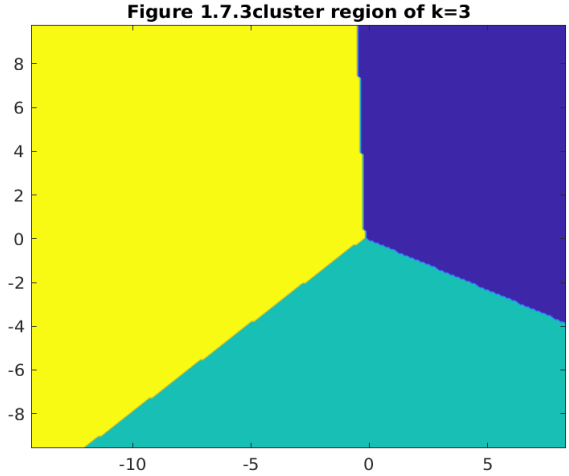


Figure 1.7.5cluster region of k=5

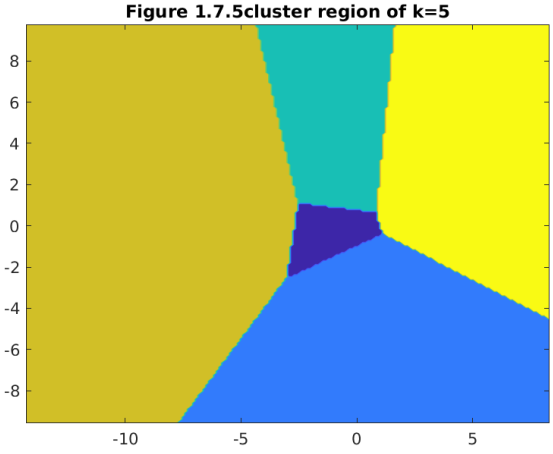
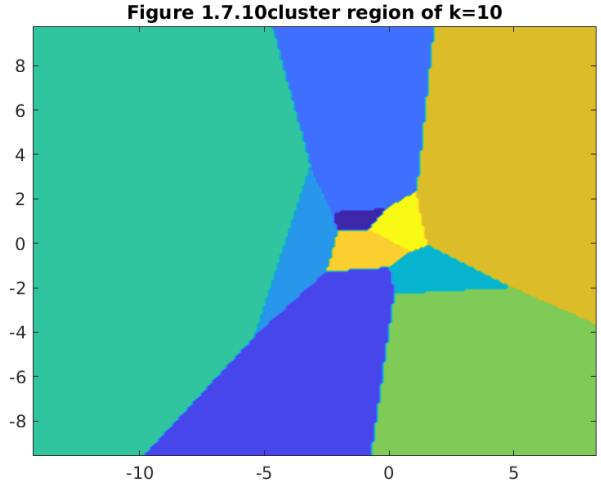


Figure 1.7.10cluster region of k=10



description of methods:

Let the original coordinate system be X, and the coordinate system derived from the 2D plane be Y.
1. generate the 2D plane ‘grid2d’;
2. applying coodinate conversion to get projection of the cluster centres on the 2D plane using the formula:
$y1 = v_1^T (x - p)$
$y2 = v_2^T (x - p)$
3. apply k-means clustering on the 2D plane for only one time(i.e. iteration = 0, just assign each data to the nearest initial centre).
4.reshape and plot the ‘idx’ variable got from the clustering process in part 3.

Task 1.8

In this task, 4 methods of choosing the initial cluster centroid positions are implemented:

1. Unform: select k points uniformly distributed from the range of X.
2. Plus: generate k seeds by invoke the k-means++ algorithm.
3. Sample: Choose k observations from the training data X at random.
4.Cluster: Process a previous clustering on a random subset of X*ratio, the preliminary phase is initialized by ‘sample’ itself(In this task, ratio = 0.1).

Implementation: For each k, the k-means clustering with each method of initialization will be executed for 5 times, and the data of mean running time as well as the sum of final SSE of each deminsion will be recorded.

Performance:

For k =3:

Method	uniform	plus	sample	cluster
Time (s)	10.26	11.53	19.30	17.74
SSE(10^6)	1.973	1.978	1.973	1.973

For k = 5:

Method	uniform	plus	sample	cluster
Time (s)	7.5986	9.9433	17.06	13.69
SSE(10^6)	18.27	18.27	18.31	18.31

For k = 7:

Method	uniform	plus	sample	cluster
Time (s)	13.33	11.25	23.67	16.29
SSE(10^6)	1.740	1.739	1.739	1.739

These results indicates that ‘uniform’ and ‘plus’ methods are significantly more efficient than the other two methods. However, the accuracy of the four methods are very close. Further test can be done using bigger K and different data.