CS 551: PATTERN RECOGNITION

ASSIGNMENT #3:

FEATURE REDUCTION

OMER GOZUACIK

21301328

Question 1:

Question 2:

In this question, I used PCA to reduce the dimension of the data to lower dimensions. PCA is applied to whole dataset then the specific amount of eigenvectors found from it are multiplied with training and test data in order to map data to lower dimension. I implemented my own code in this part.

1) Plot for eigenvalues

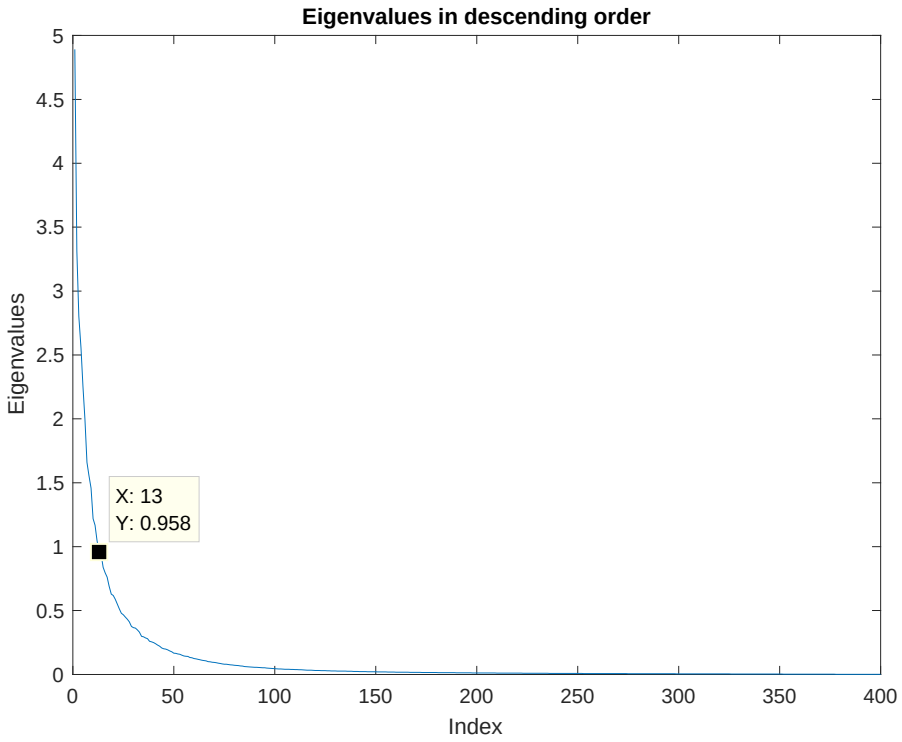
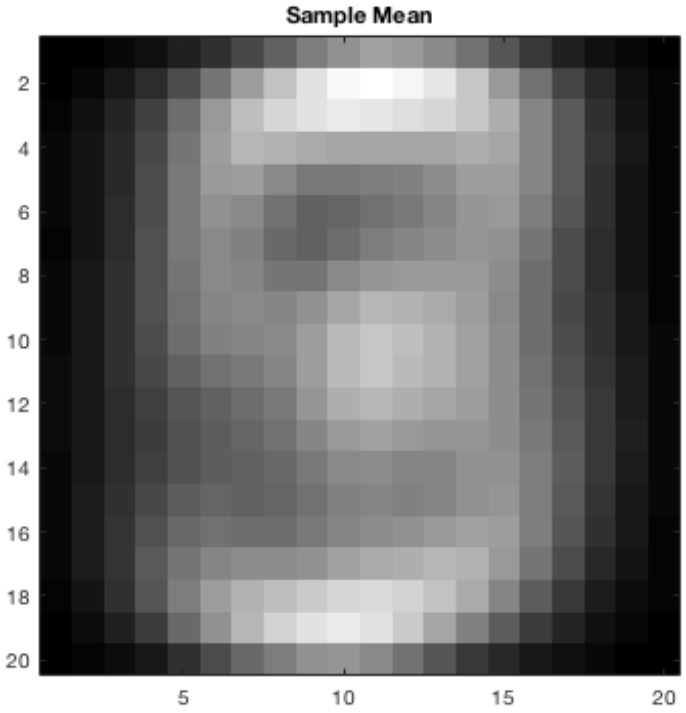
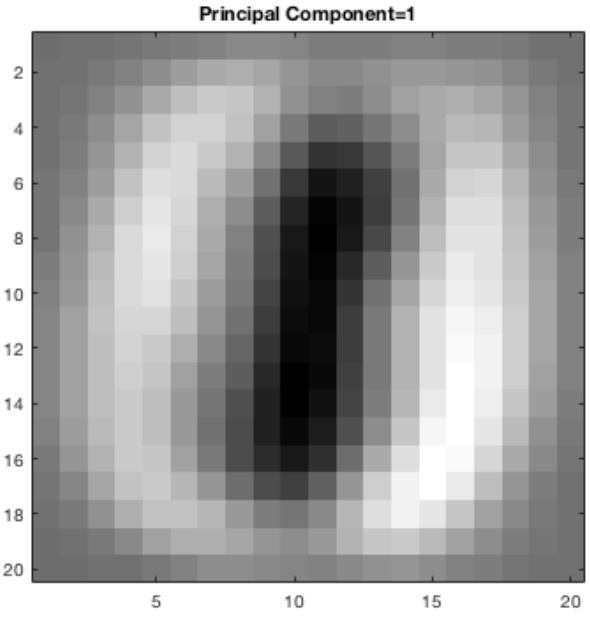


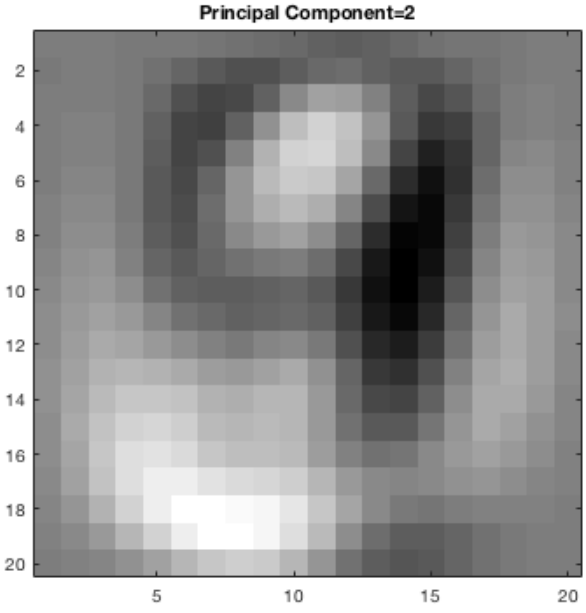
Figure 1: Eigenvalues of data in descending order

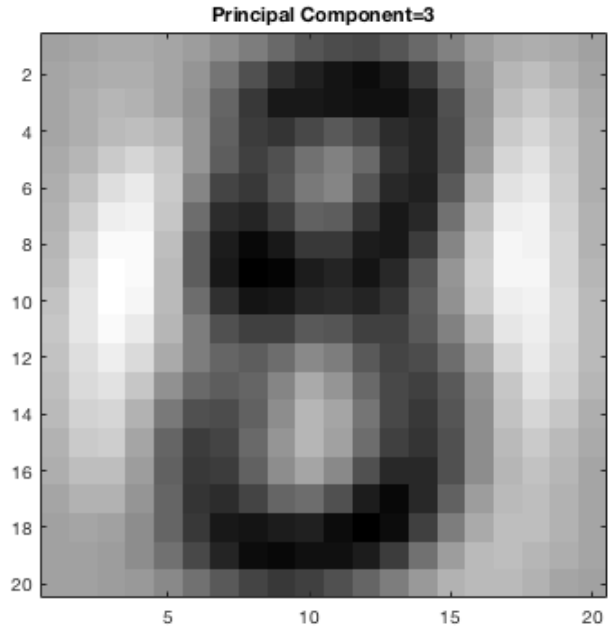
I used Kaiser rule [1] to determine the number of components. According to this rule, eigenvalues larger than 1 should be selected as bases. By looking to the graph above, we can see that 12 components are enough to represent the data.

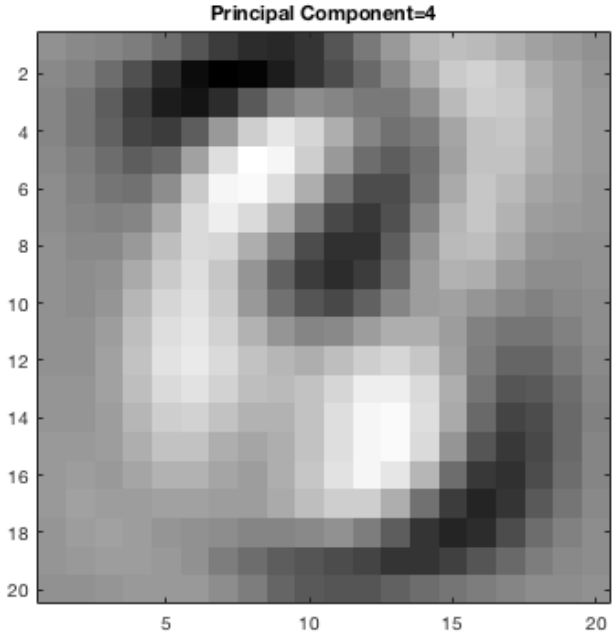
2) Images of sample mean and selected bases

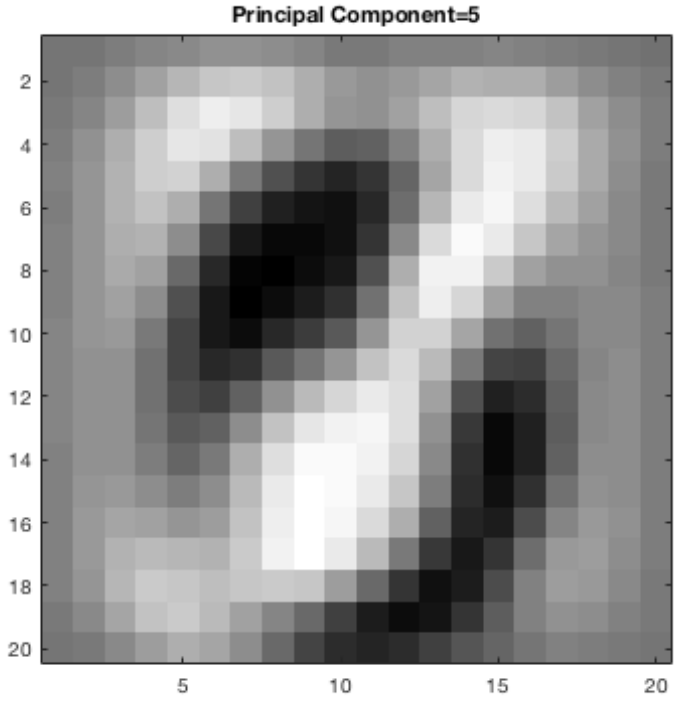


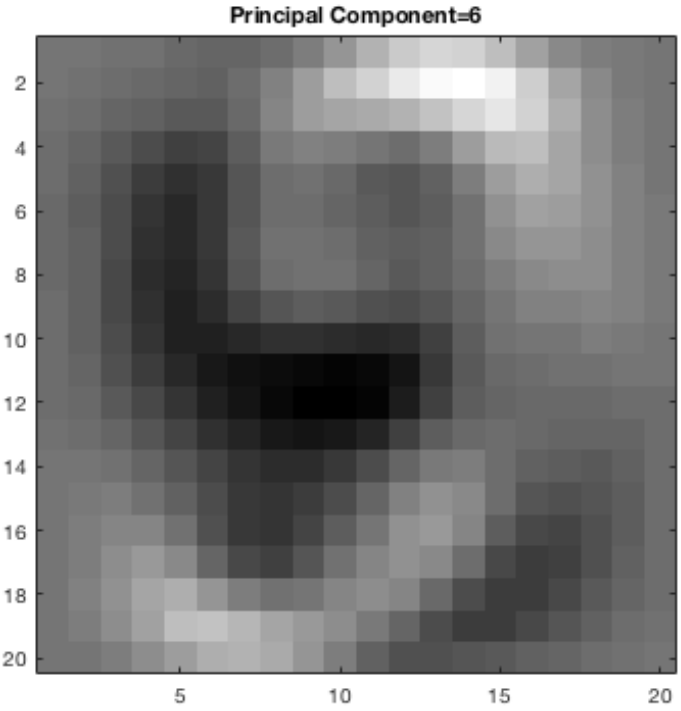

Principal Component=1
2
4
6
8
10
12
14
16
18
20
5
10
15
20

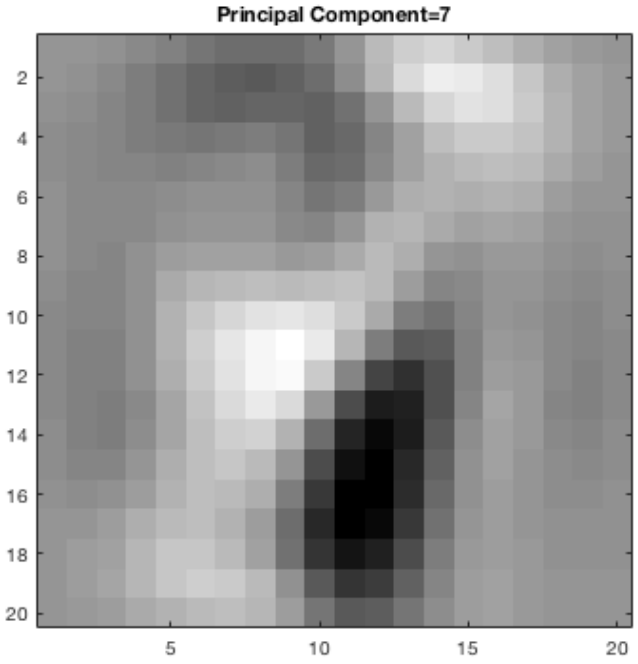

Principal Component=2
2
4
6
8
10
12
14
16
18
20
5
10
15
20

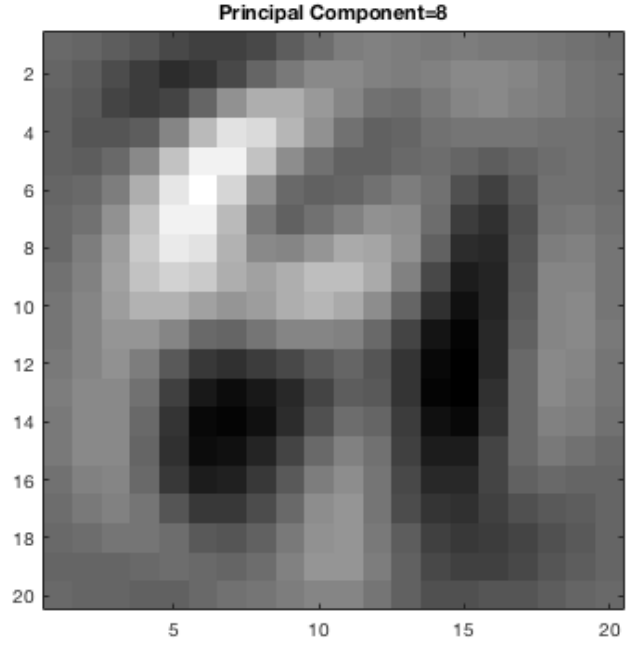

Principal Component=3
2
4
6
8
10
12
14
16
18
20
5
10
15
20

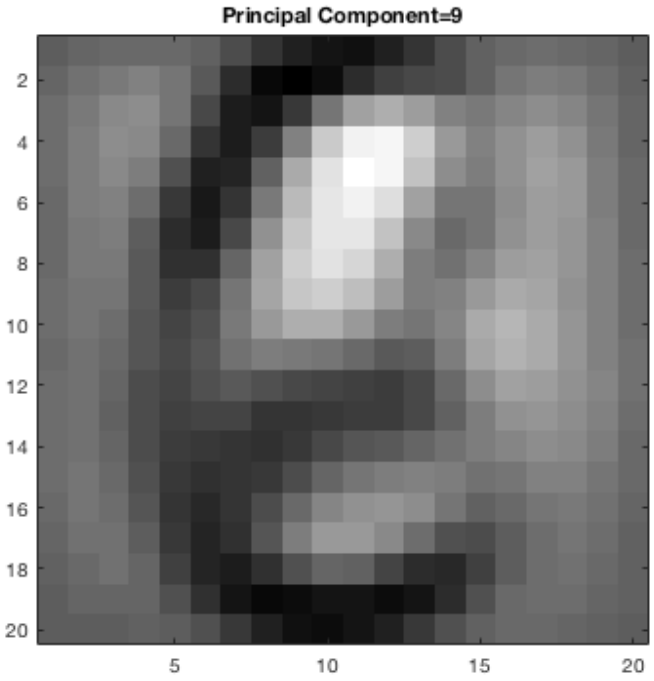

Principal Component=4
2
4
6
8
10
12
14
16
18
20
5
10
15
20

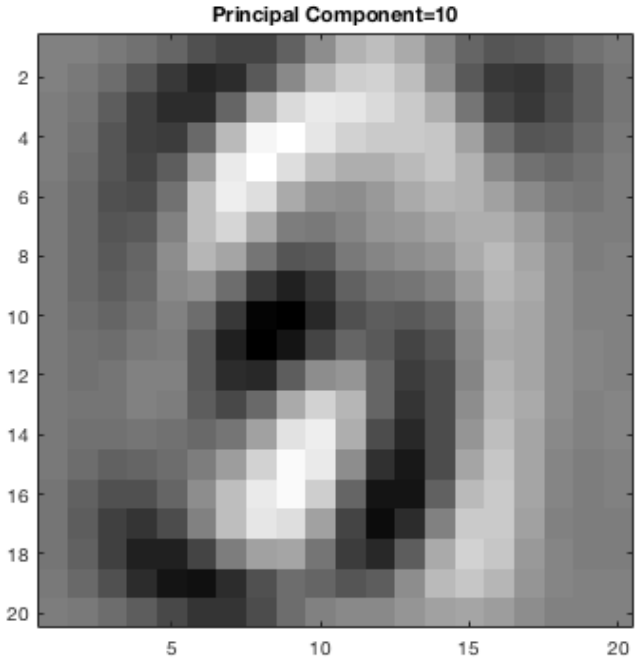

Principal Component=5
2
4
6
8
10
12
14
16
18
20
5
10
15
20

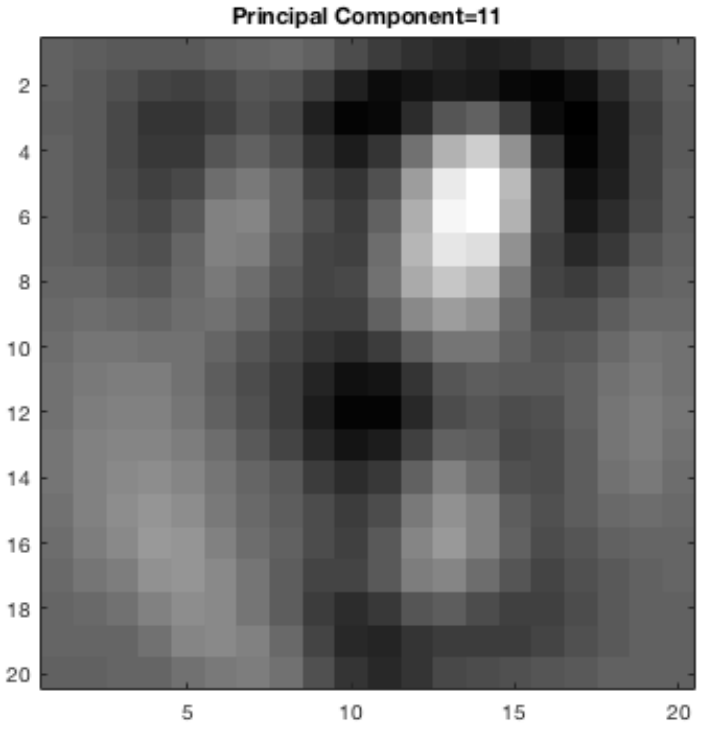

Principal Component=6
2
4
6
8
10
12
14
16
18
20
5
10
15
20

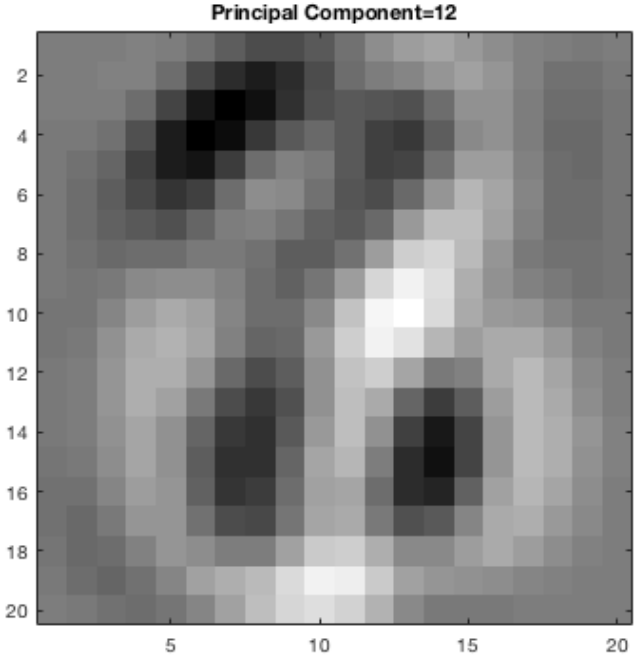

Principal Component=7
2
4
6
8
10
12
14
16
18
20
5
10
15
20


Principal Component=8
2
4
6
8
10
12
14
16
18
20
5
10
15
20


Principal Component=9
2
4
6
8
10
12
14
16
18
20
5
10
15
20


Principal Component=10
2
4
6
8
10
12
14
16
18
20
5
10
15
20





As expected, the some of the resulting bases looks like the digits that we want to classify and some look like their mixture. Their mean almost looks like 8 as most of thee digits have a hole in their upper or lower portion.

3) Training and testing Gaussian classifiers for different subspaces with different dimensions

Gaussian models for each class is trained for different subspace dimensions. MLE is used to estimate the parameters. Then these models are combined to create a classifier. If a point X has larger pdf in a class than the others, then X is classified as that class.

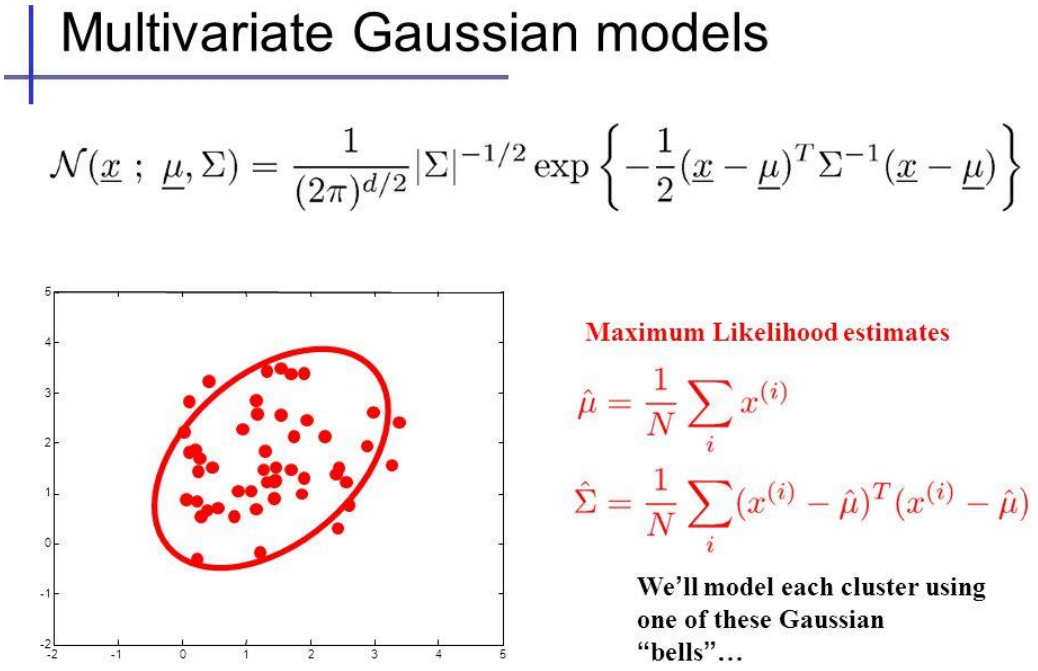


Figure 2: Multivariate Gaussian Formula and MLE estimates [2]

4) Classification error vs. the number of components used for each subspace

Gaussian classifiers are trained for different subspaces with different dimensions. Afterwards, they are used to classify both training and test data. Accuracies with respect to # of dimensions are found below:

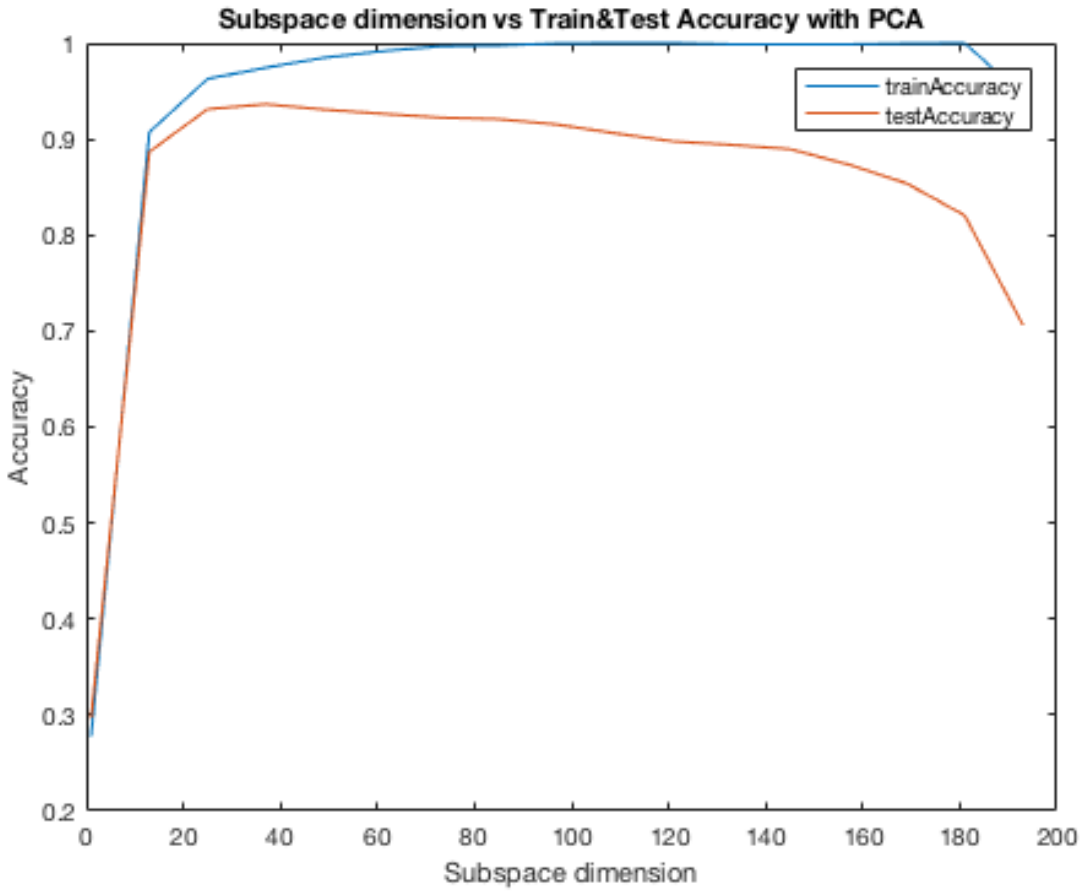


Figure 3: Number of subspace dimensions' vs test and training accuracies

In the assignment it was asking me to run the dimensions in between 1-400 but I stopped at 200 as my computer is slow and the general pattern can be seen in this range as well. According to the Kaiser rule, 12 dimensions seemed to be enough for good accuracy but, the model performs best when the number of dimensions is around 30+-5. Training accuracy increases as the number of component increases. However, test accuracy decreases when the number is over 40 as model **overfits** the training data starting on that point. When I increased the dimension over 200, training accuracy also dropped due to underflows (my code is vulnerable to underflows).

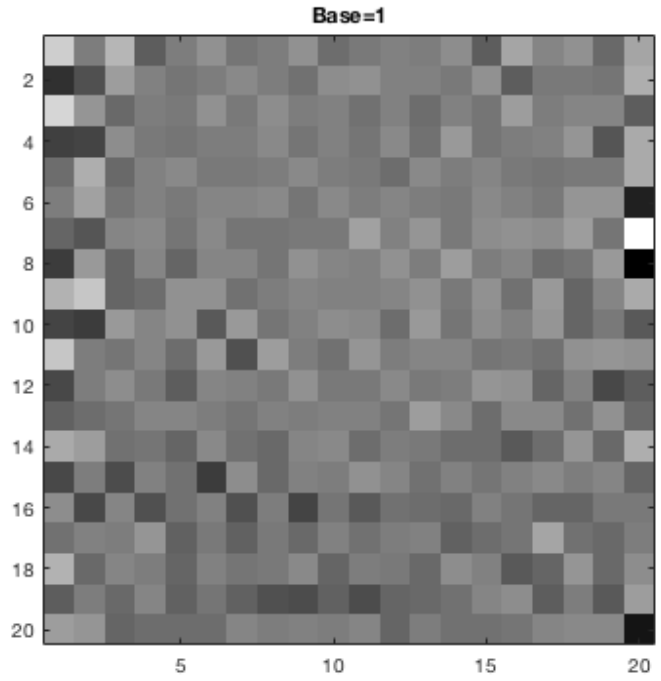
Question 3:

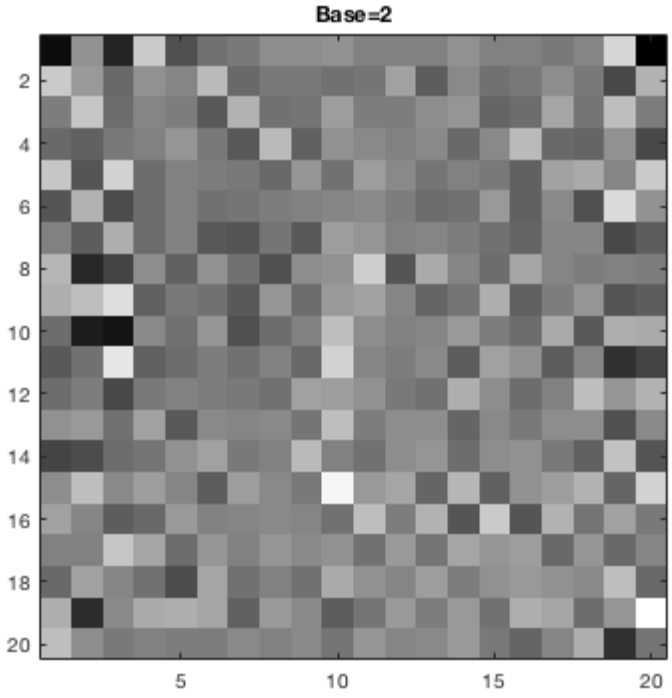
In this question, I used LDA to reduce the dimension of the data to lower dimensions. LDA is applied to whole dataset then the bases found from it are multiplied with training and test

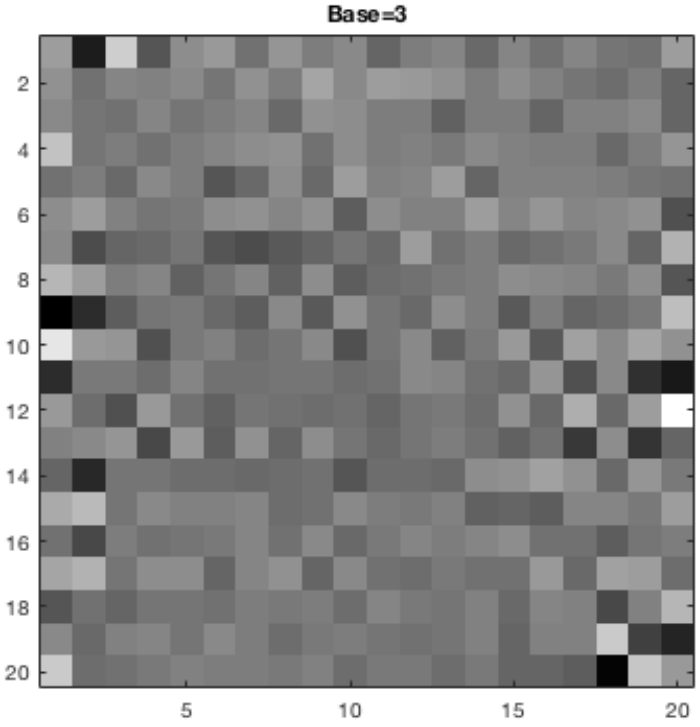
data in order to map data to lower dimension. I used the toolbox provided by Laurens van der Maaten [3].

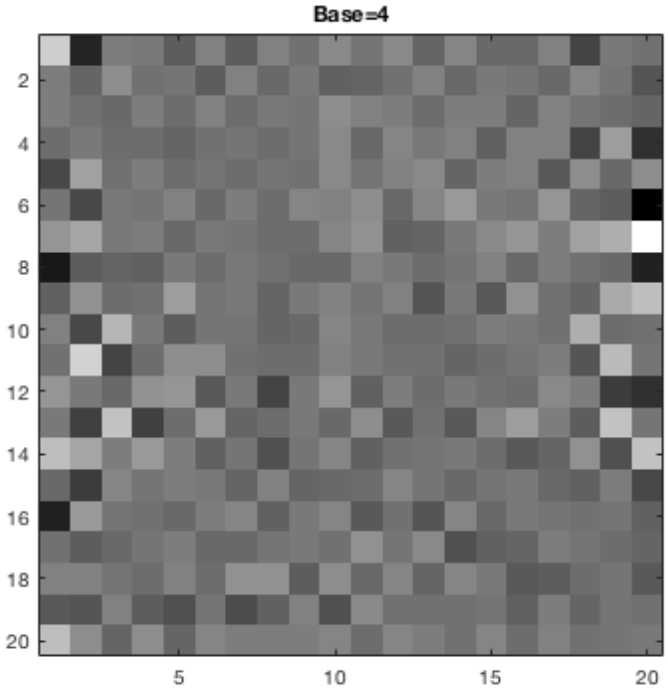
1) Obtaining new bases with LDA

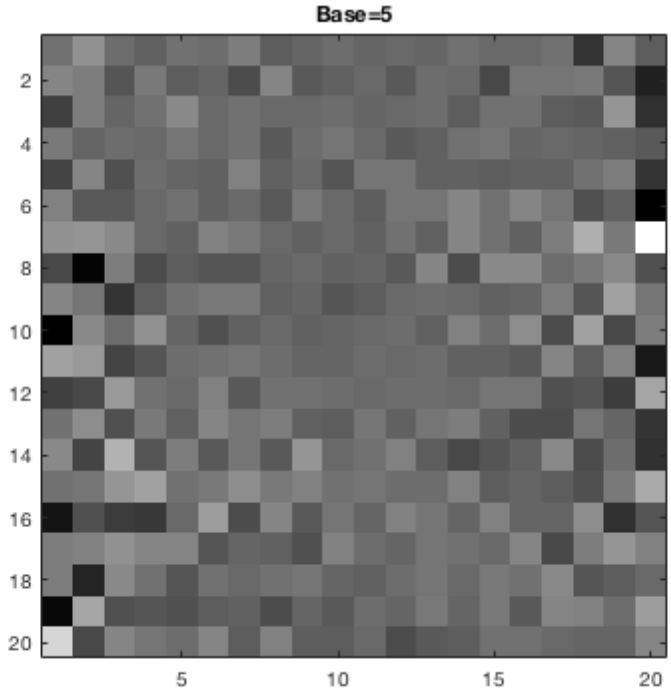
When LDA is used on to the whole data, as there are 10 classes, 9 new bases are found.

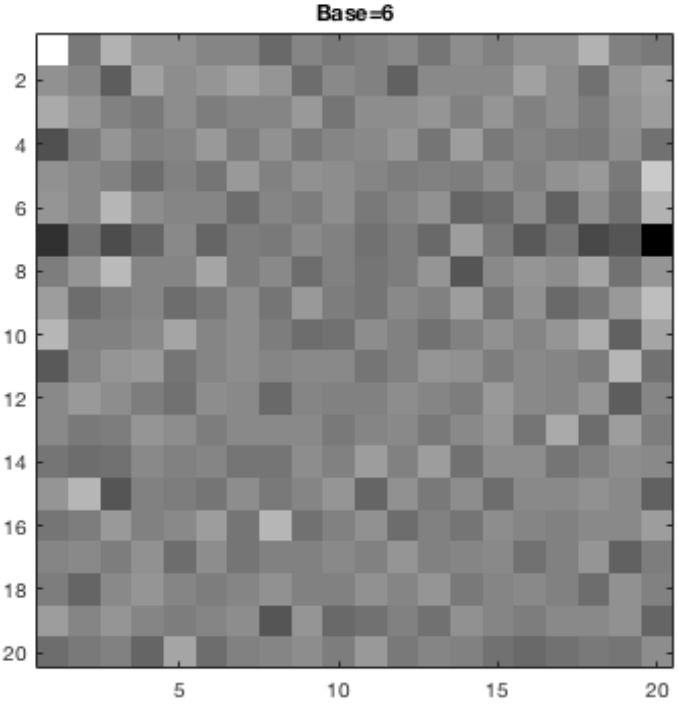

Base=1
2
4
6
8
10
12
14
16
18
20
5
10
15
20

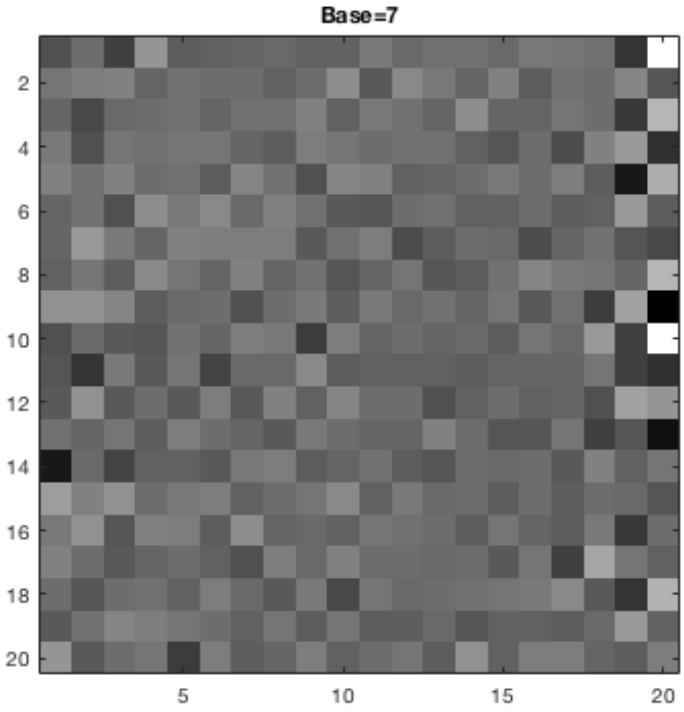

Base=2
2
4
6
8
10
12
14
16
18
20
5
10
15
20

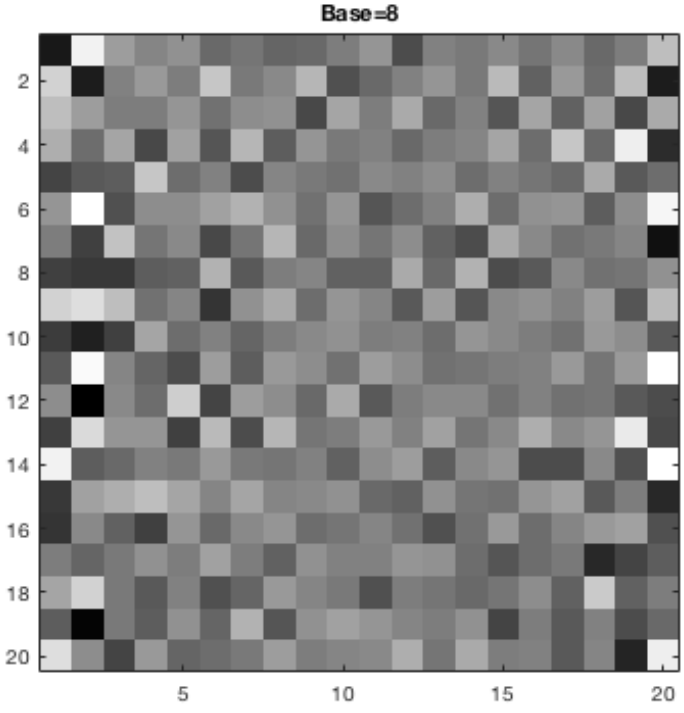

Base=3
2
4
6
8
10
12
14
16
18
20
5
10
15
20

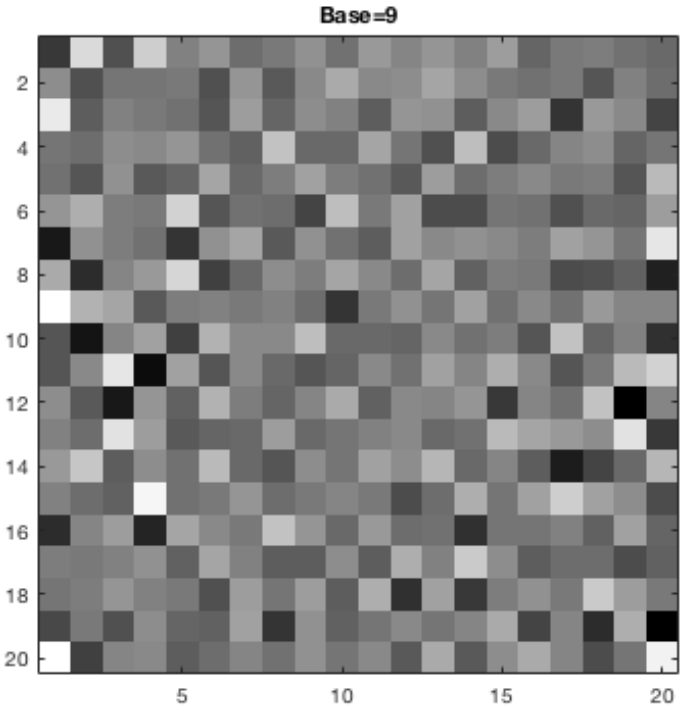

Base=4
2
4
6
8
10
12
14
16
18
20
5
10
15
20


Base=5
2
4
6
8
10
12
14
16
18
20
5
10
15
20


Base=6
2
4
6
8
10
12
14
16
18
20
5
10
15
20


Base=7
2
4
6
8
10
12
14
16
18
20
5
10
15
20


Base=8
2
4
6
8
10
12
14
16
18
20
5
10
15
20



As expected, these new bases are not descriptive as the bases found with PCA. In PCA, the bases looked like digits as the algorithm focuses on data without knowing labels. However, LDA focusses on differences between labels, that's why bases found by it do not represent anything similar to a digit.

2) Training and testing Gaussian classifiers with subspaces of dimension 1 to 9

As if in the previous question, Gaussian models for each class is trained for different subspace dimensions. MLE is used to estimate the parameters. Then these models are combined to create a classifier. If a point X has larger pdf in a class than the others, then X is classified as that class.

3) Classification error vs. the number of components used for each subspace

Gaussian classifiers are trained for different subspaces with different dimensions. Afterwards, they are used to classify both training and test data. Accuracies with respect to # of dimensions are found below:

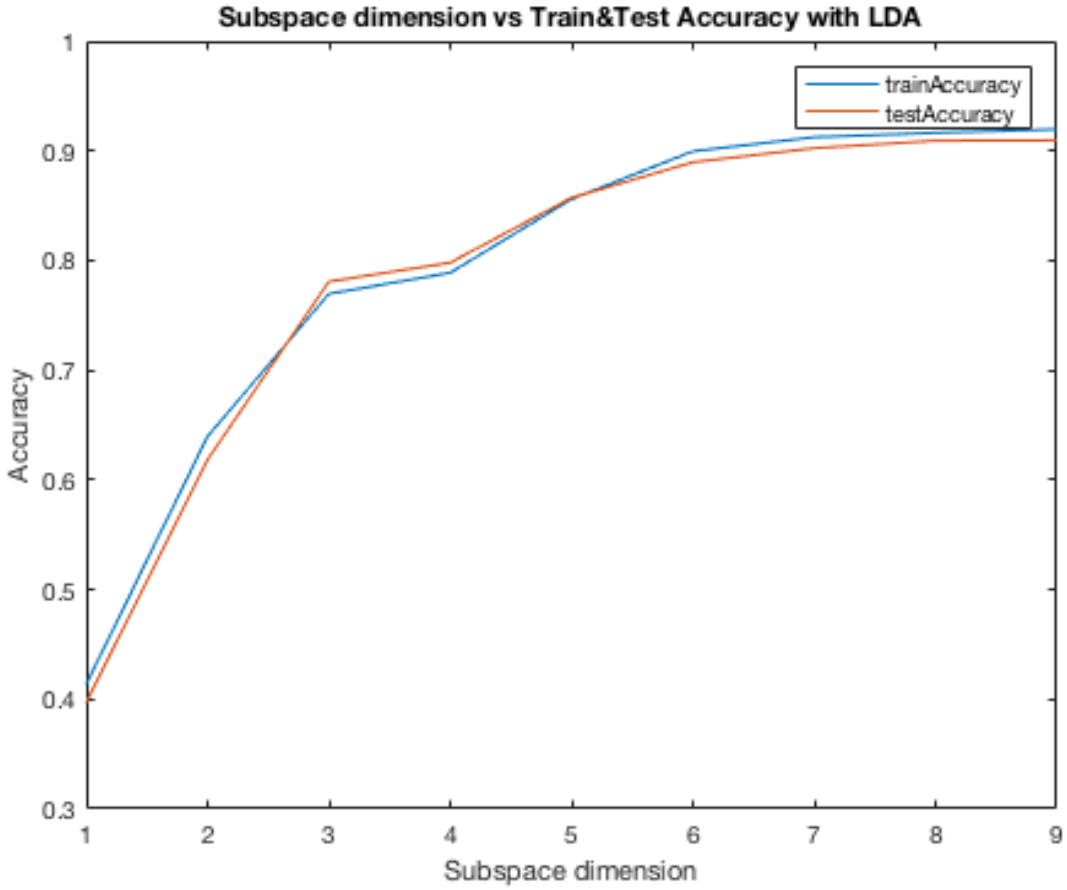


Figure 4: Number of subspace dimensions' vs test and training accuracies

As in the previous case, training&test accuracy increases as the subspace dimension increases. As dimension of the data can be at most 9, the Gaussian models can not overfit. That's why, accuracy for both training and test data are similar.