Homework 4 - Question 3

Sahil Shah and Preey Shah

October 2017

1 Question 3

As asked, we have submitted the graphs of the eigenvalues for each digit in the files eigen<value of the digit>.png and the three images for each digit can be found in the files image<value of digit>.png.
The file generateData.m, we have written the function used to obtain the arrays from the data that we found on the link given. In the file p3.m, we find the actual code used to compute all the covariances, find the eigenvalues and eigenvectors, means, etc for each of the images given in the dataset.

On sorting the eigenvalues in ascending order and plotting them, we observed that the graph has a very large value at the beginning and then rapidly falls, indicating that the first few eigenvalues (for every digit) are large and the values after that are relatively negligible. This is justified because we have taken 28*28 dimensions and the variance along most of them will be exceedingly small. This very example was given in class as an example of how Principle Component Analysis may have real world applications.
The reason that the number of useful dimensions are far lesser than the number of pixels is that when people write digits most of the pixels have the same values (like the ones on the edges will always be black.
The significant modes of variation are far lesser than 28*28 for each digit the (approximate) number of modes of variation are given as follows (from what we observe from the graphs obtained for each digit):

0	25
1	20
2	30
3	25
4	30
5	25
6	25
7	20
8	30
9	30

v_1 denotes the vector along the direction of maximum variation. Hence, the two plots represent how much the digit varies around the mean. We check the values at a distance one standard deviation away from the mean along the direction of principle variation (on both sides of the mean). Here, we note that the eigenvalue represents the variance and hence, its root represents the standard deviation.
We would expect that most of the other images of the same digits would lie "between" the two extremes that we computed.
This is analogous to considering a gaussian and expecting most values drawn from it to lie within a range of 2σ centred around the mean μ.