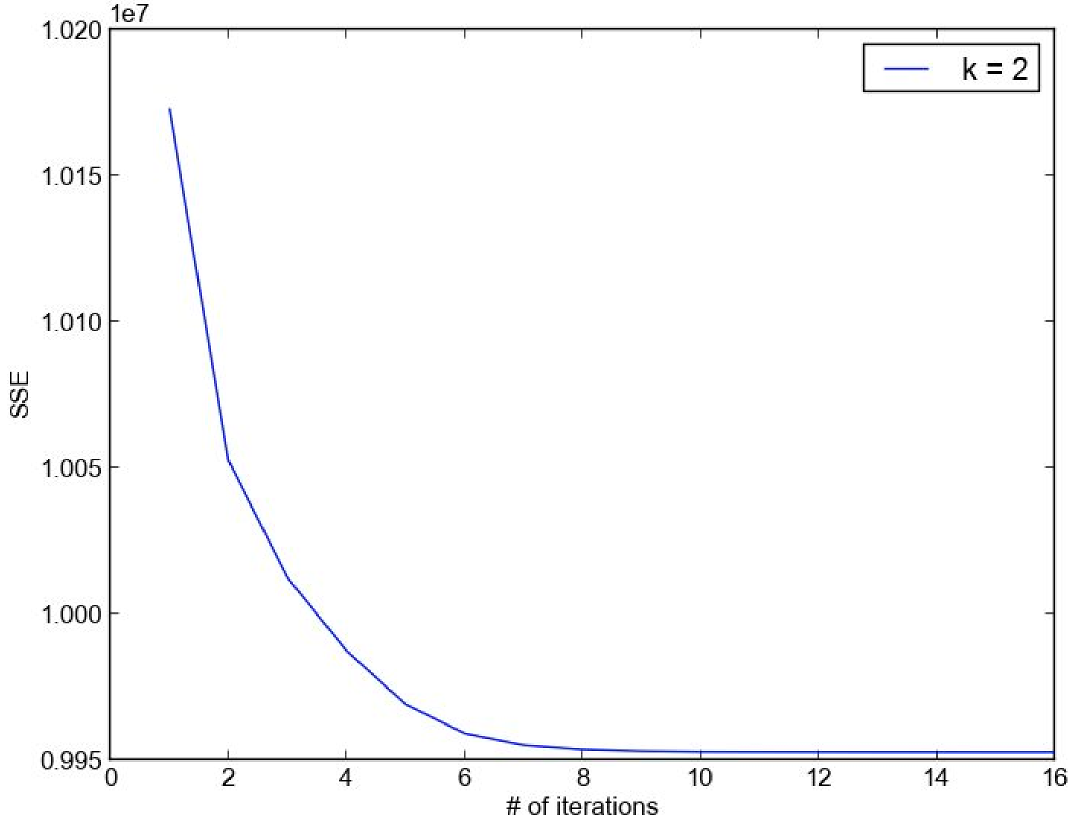
CS434 Assignment #4 Report

Name: Tsewei Peng

Part 2. Non-Hierarchical Clustering - K-means Algorithm

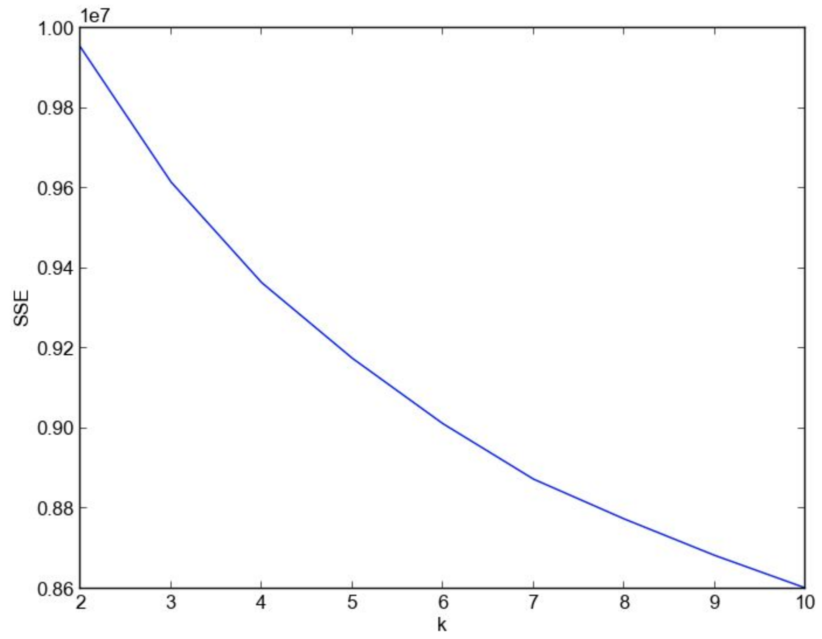
Problem 1.

Below is the graph of SSE for each iteration when running K-means algorithm on the data when
K = 2.



Problem 2.

Below is the graph of lowest SSE for each k value, ranging from 2 to 10. For each k value, the algorithm is repeated 10 times. According to the curve on the graph, the best k value would be 10, since it creates the lowest SSE, and the SSE decreases with a very steady slope without increasing from k = 2 to k = 10.



Part 3. Principal Component Analysis

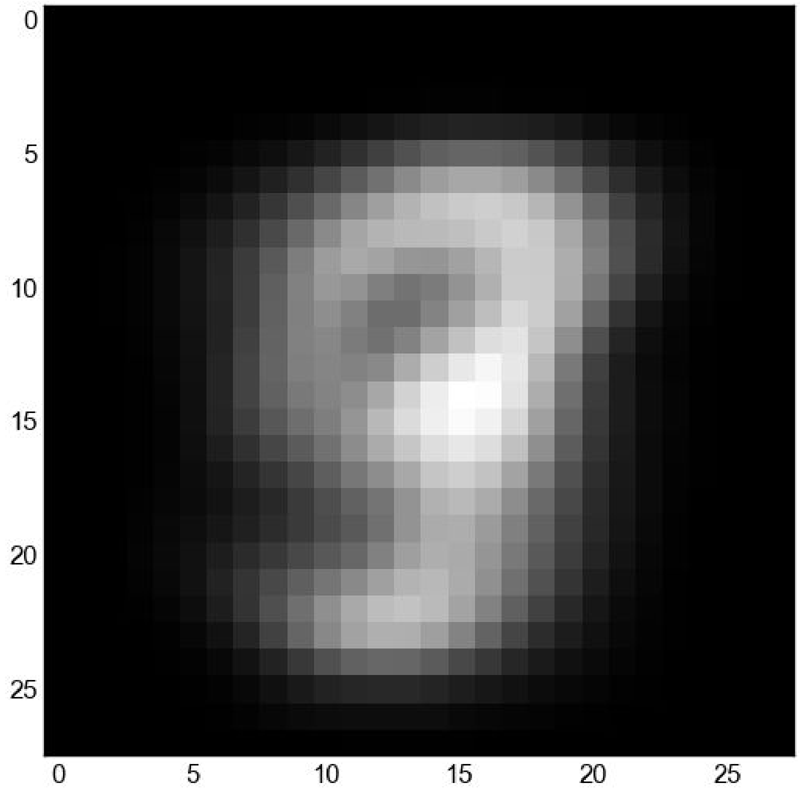
Problem 1.

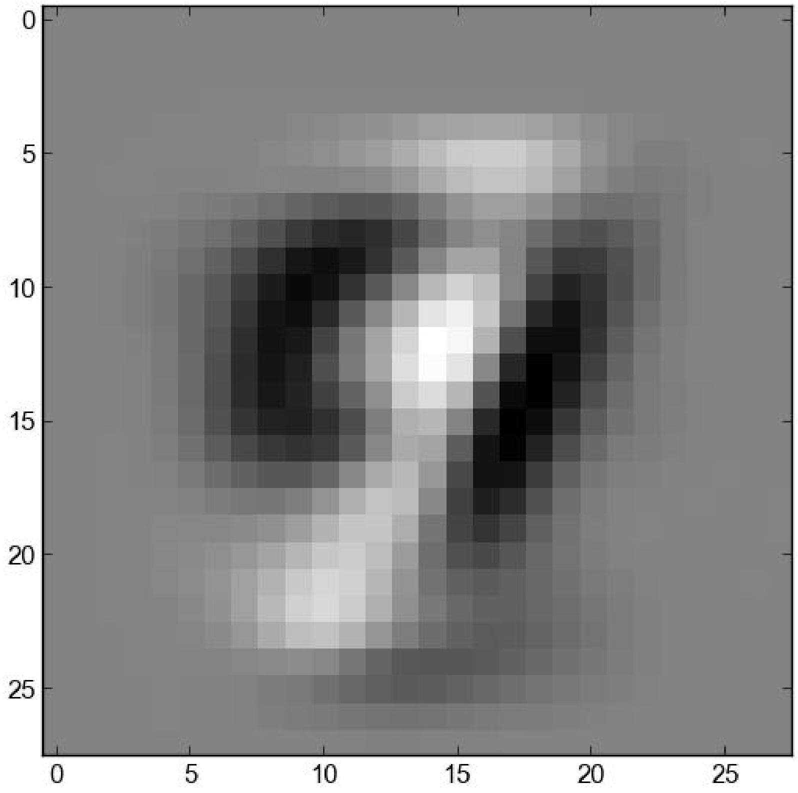
10 largest eigen values for covariance matrix:

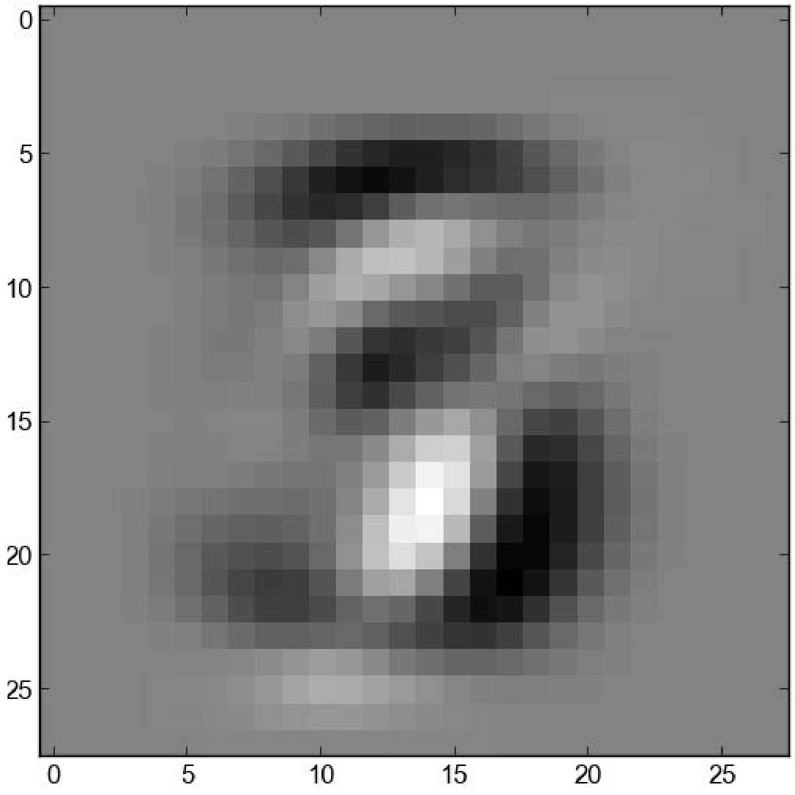
1	352868.69125592
2	267895.86687052
3	227632.6992441
4	174703.49025817
5	130486.76236012
6	115542.50268179
7	99726.43672642
8	90576.05787847
9	85326.53680823
10	71547.9660124

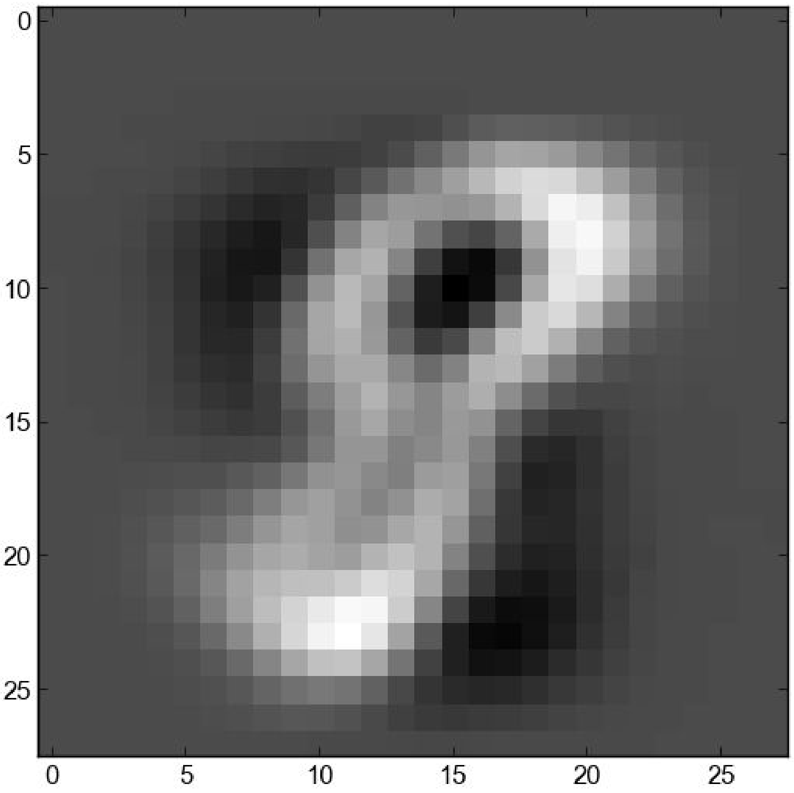
Problem 2.

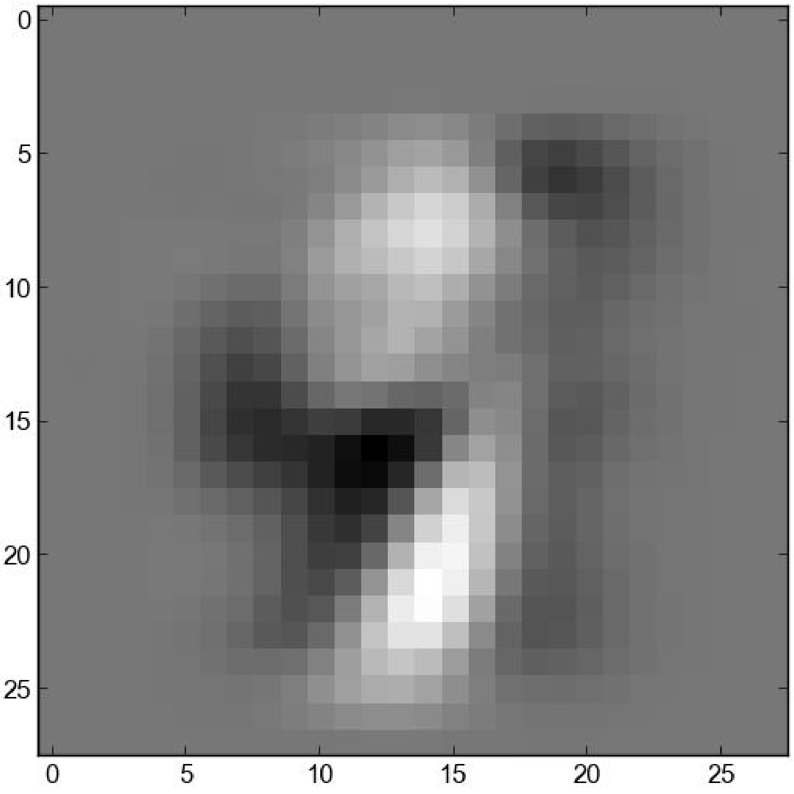
Below are the images plotted. The first image is plotted using mean vector, and the rest are plotted from with top 10 eigen vectors.

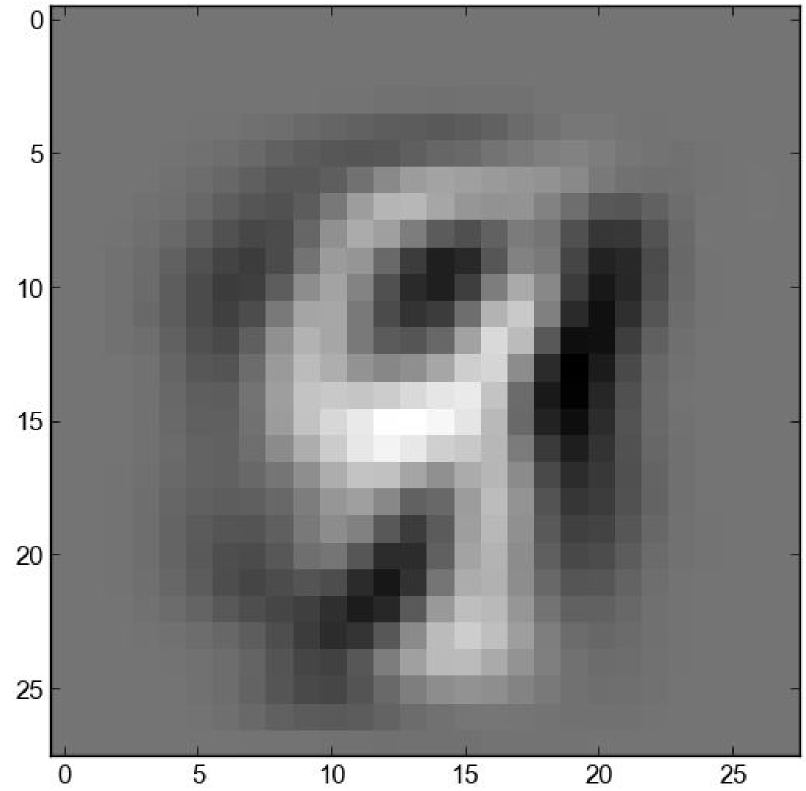


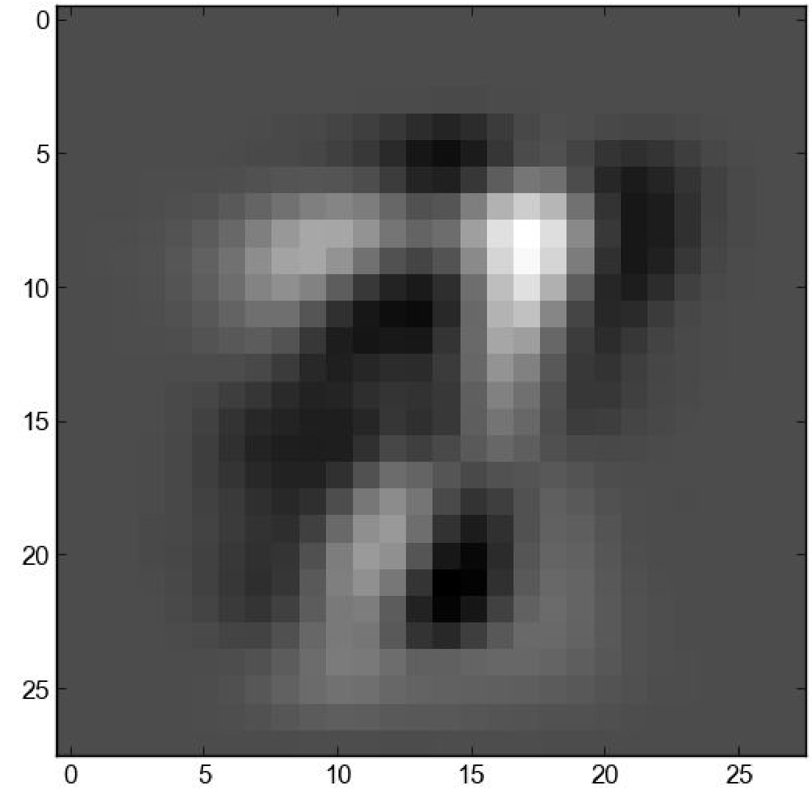

0
5
10
15
20
25
0
5
10
15
20
25

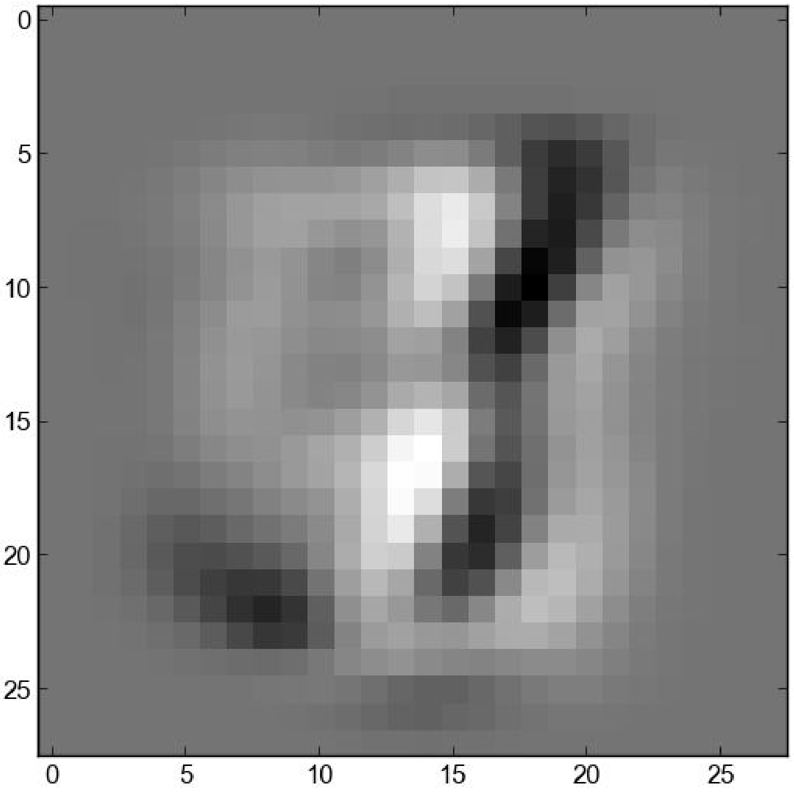

0
5
10
15
20
25
0
5
10
15
20
25

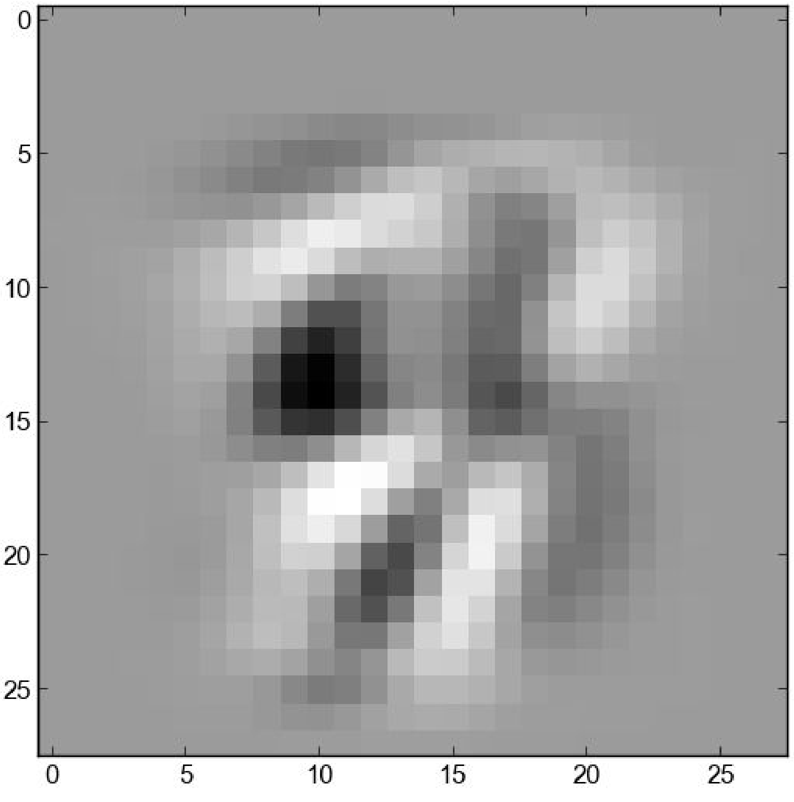

0
5
10
15
20
25
0
5
10
15
20
25

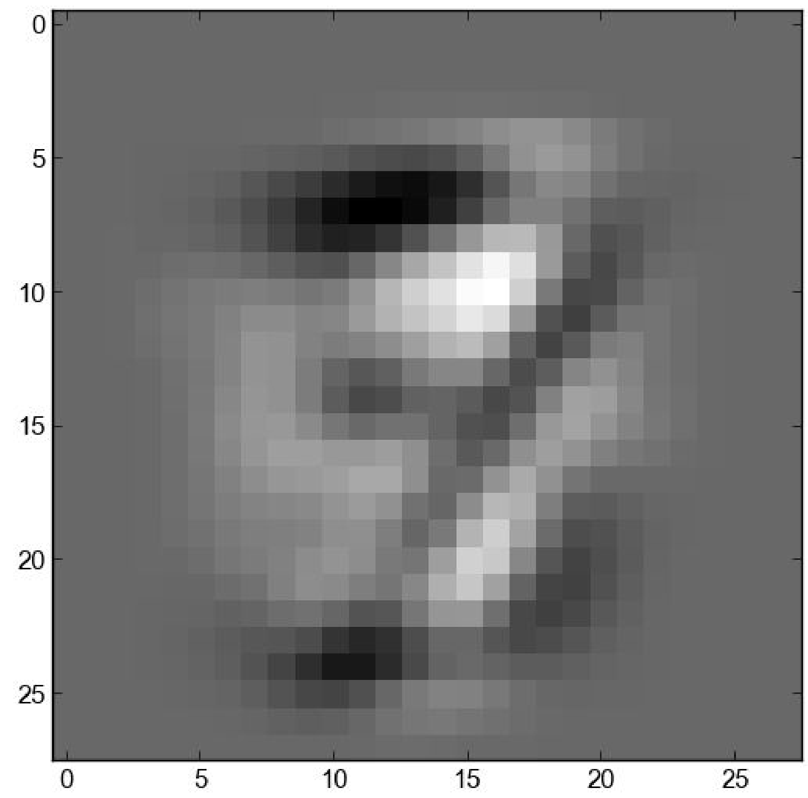

0
5
10
15
20
25
0
5
10
15
20
25

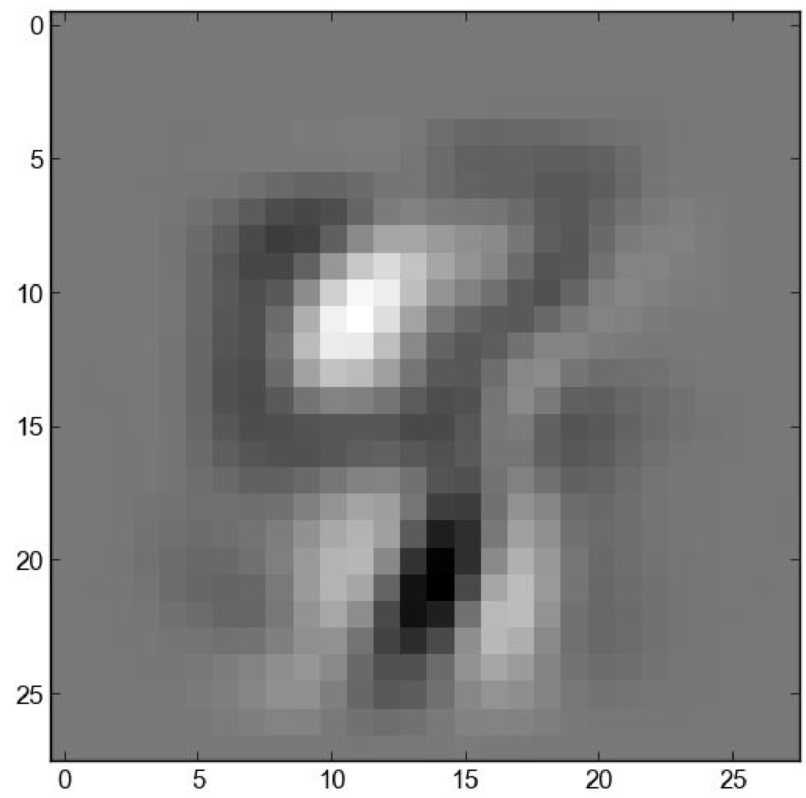

0
5
10
15
20
25
0
5
10
15
20
25


0
5
10
15
20
25
0
5
10
15
20
25


0
5
10
15
20
25
0
5
10
15
20
25


0
5
10
15
20
25
0
5
10
15
20
25


0
5
10
15
20
25
0
5
10
15
20
25

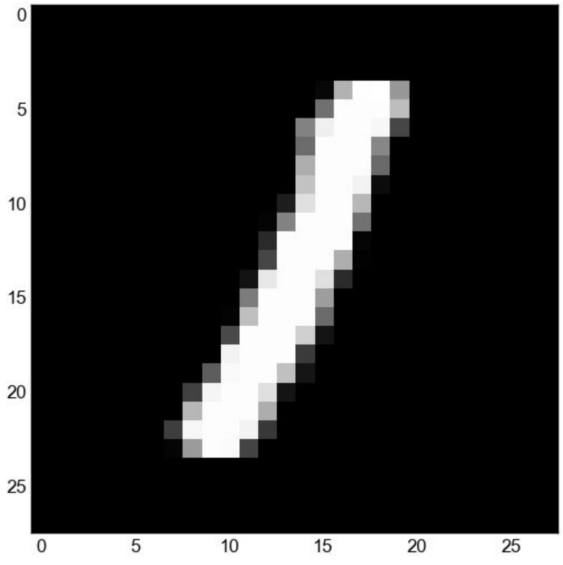

0
5
10
15
20
25
0
5
10
15
20
25

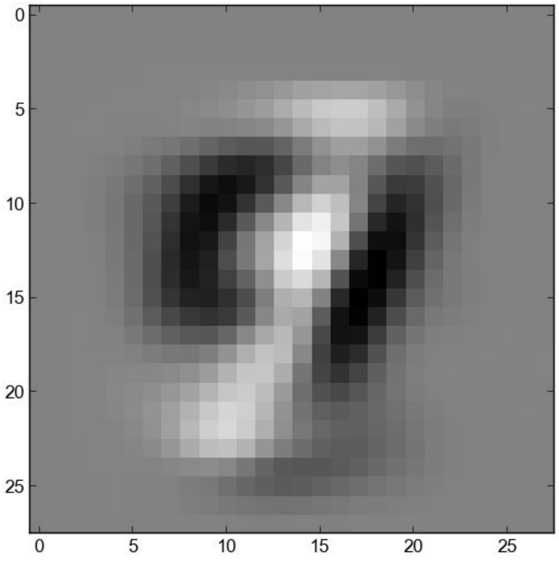
For the image plotted using mean vector, we can see the general trend of most handwritten digit. The image shows a digit very similar to '9', or 'g'. While the rest of the space around the digit is black, meaning value of 0.

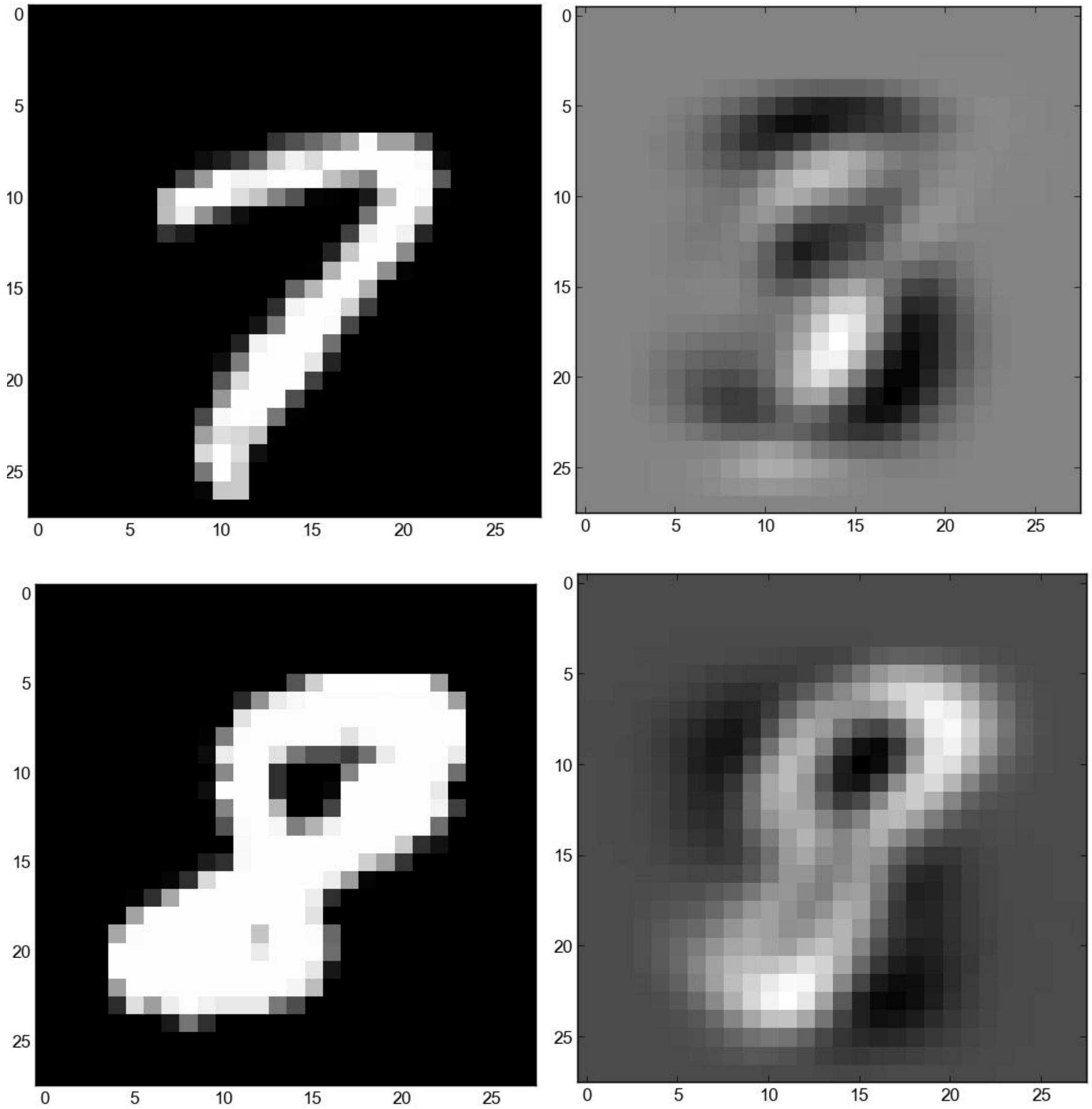
For the rest of the images for the top 10 eigen vectors derived from the data, most of them do not show any legible digits. Since the eigenvectors can contain negative values, they are normalized into gray scale of 0 to 255, causing the negative values to create dark spots, while positive values create white spots in the image. One important thing to note is that the position of the handwritten digit is still preserved, they are mostly gray in the image meaning those values are very close to 0. This means that each eigen vector preserve information about all the handwritten digit data, and the higher the eigen value, the more information it preserves.

Problem 3.

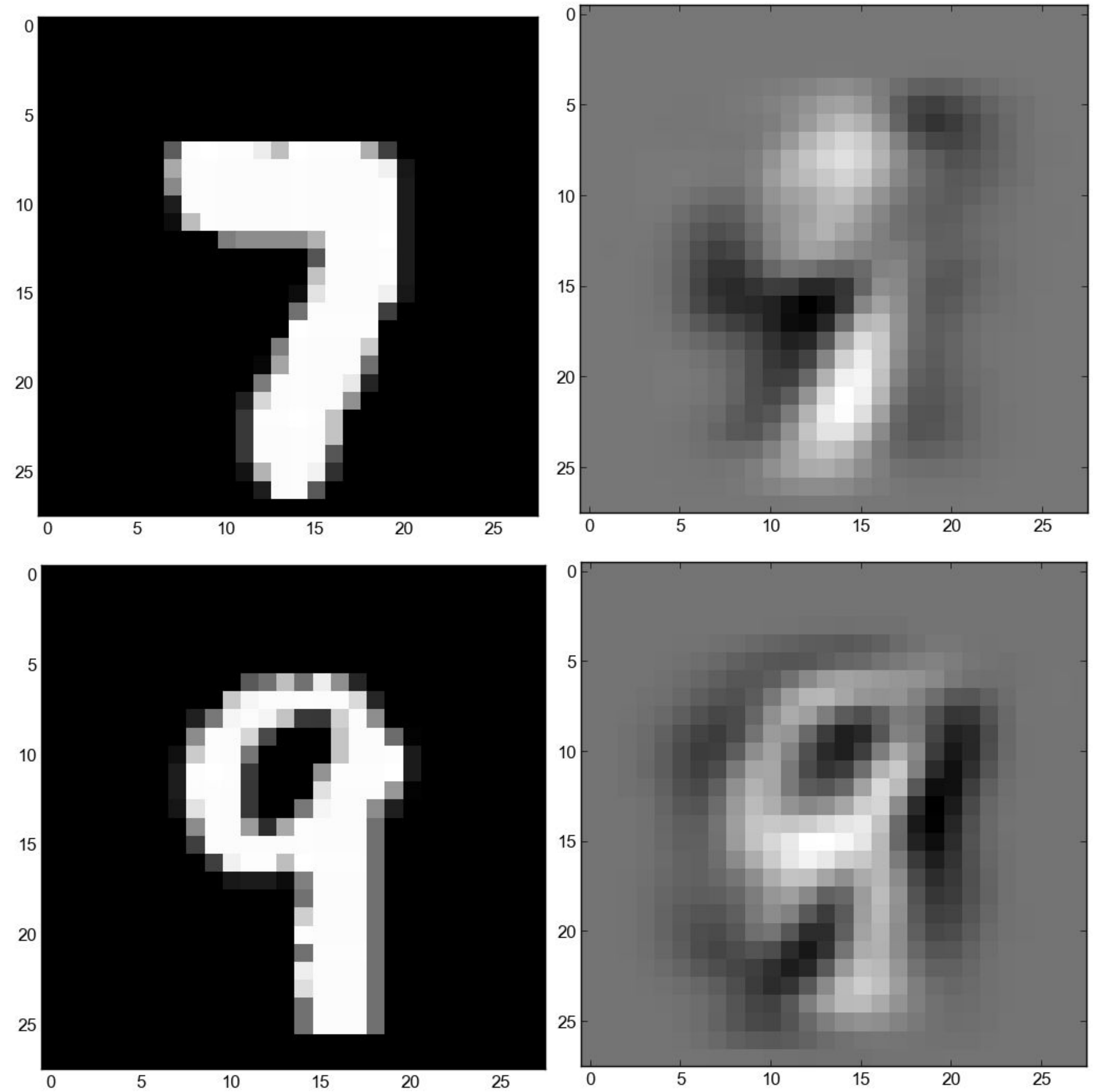
Below are the images plotted that has the largest value in the dimensions in terms of top 10 eigen vectors, derived from the data. When comparing the images of the value and the corresponding eigen vectors, we can see the the resemblance between the 2 in terms of white spots. This is because white spots in hand-written digit data has higher values (closer to 255), while black spots has lower values (closer to 0), and the eigen vector images, white spots has positive values while black spots has negative values. This explains the resemblance between each set of images as the higher value in the dimension of eigen vectors means resemblance to those eigen vectors.

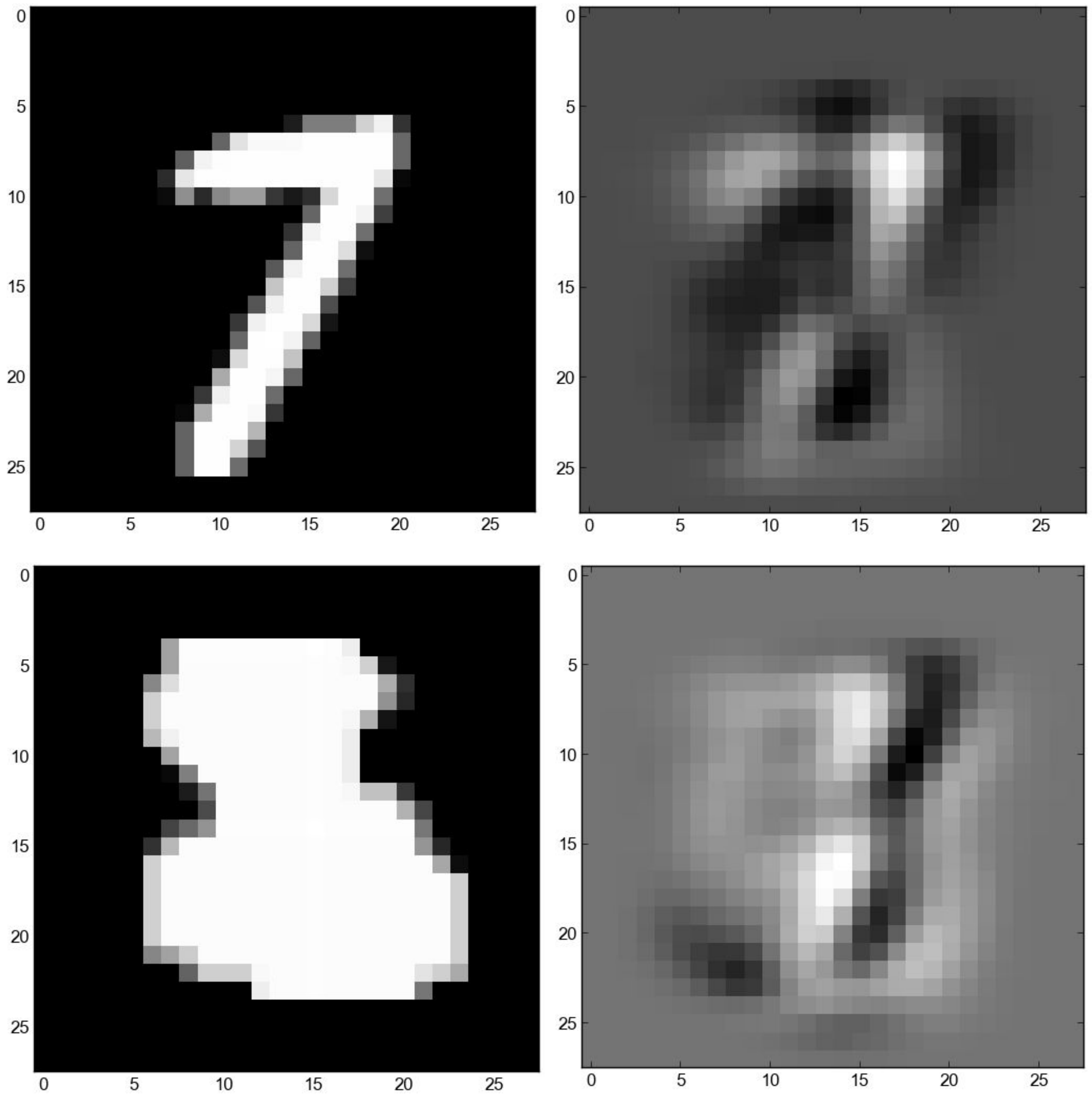


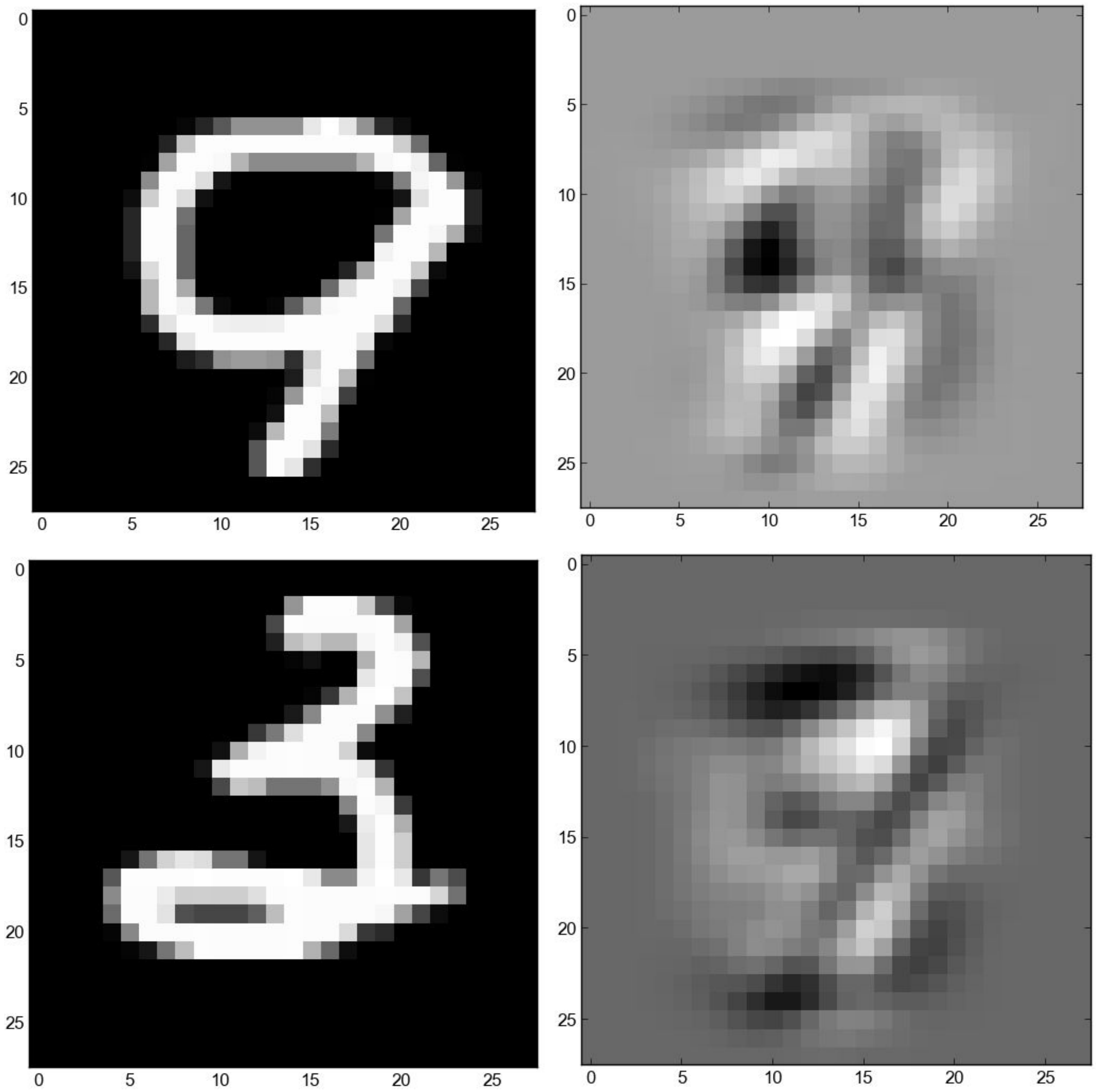


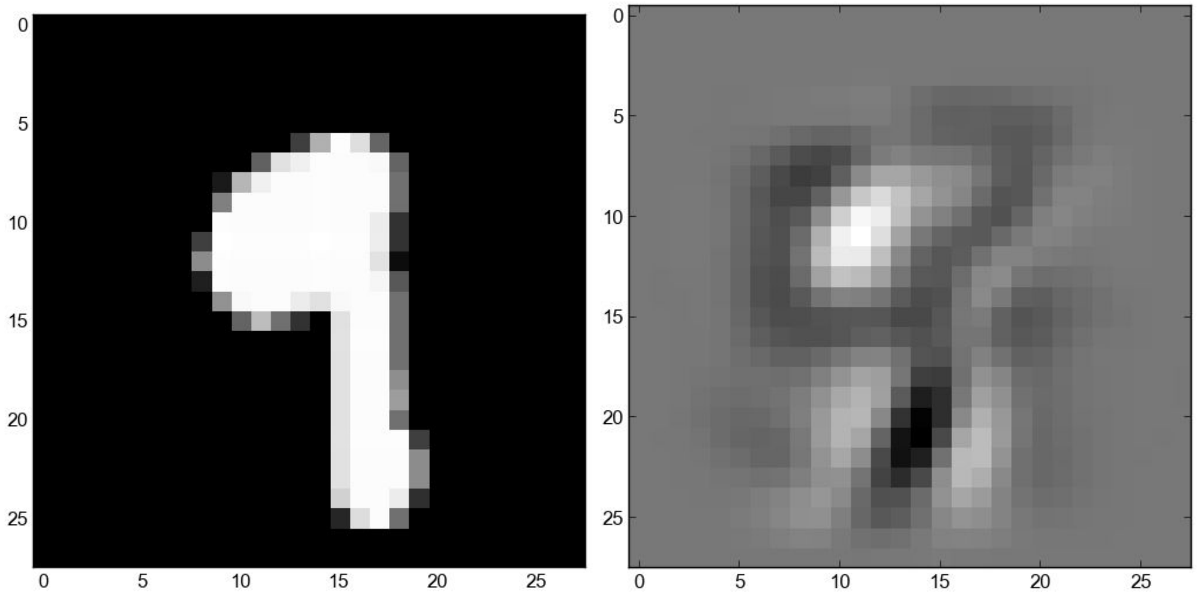


0
5
10
15
20
25
0
5
10
15
20
25
0
5
10
15
20
25
0
5
10
15
20
25
0
5
10
15
20
25
0
5
10
15
20
25
0
5
10
15
20
25
0
5
10
15
20
25


0
5
10
15
20
25
0 5 10 15 20 25
0
5
10
15
20
25
0 5 10 15 20 25
0
5
10
15
20
25
0 5 10 15 20 25
0
5
10
15
20
25
0 5 10 15 20 25


0 5 10 15 20 25
0 5 10 15 20 25
0 5 10 15 20 25
0 5 10 15 20 25
0 5 10 15 20 25
0 5 10 15 20 25
0 5 10 15 20 25
0 5 10 15 20 25


0
5
10
15
20
25
0
5
10
15
20
25
0
5
10
15
20
25
0
5
10
15
20
25
0
5
10
15
20
25
0
5
10
15
20
25
0
5
10
15
20
25
0
5
10
15
20
25


0
5
10
15
20
25
0
5
10
15
20
25
0
5
10
15
20
25
0
5
10
15
20
25