James May

Garrett Bauer

Peter Dorich

Assignment 4 Report

K-Mean

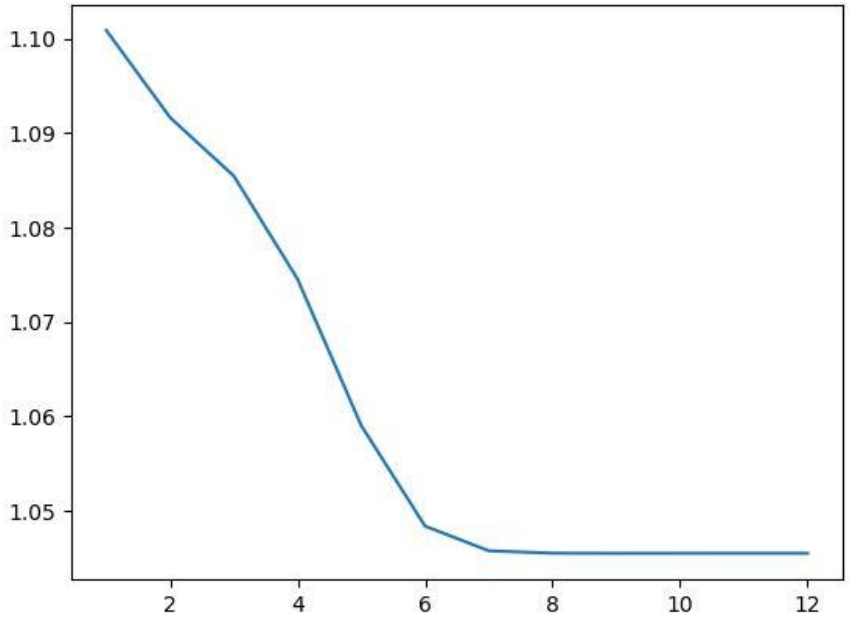


Figure 1: SSE is plotted as a function of iterations, where k = 2, and the number of points in each cluster stops changing at the 12th iteration.

The sum of squares error (SSE) is shown to converge in figure 1, roughly around iteration 7, but precisely at iteration 12. The algorithm was run with a large range starting seed positions and this was found to have the highest iteration count and therefore chosen to show convergence of the SSE.

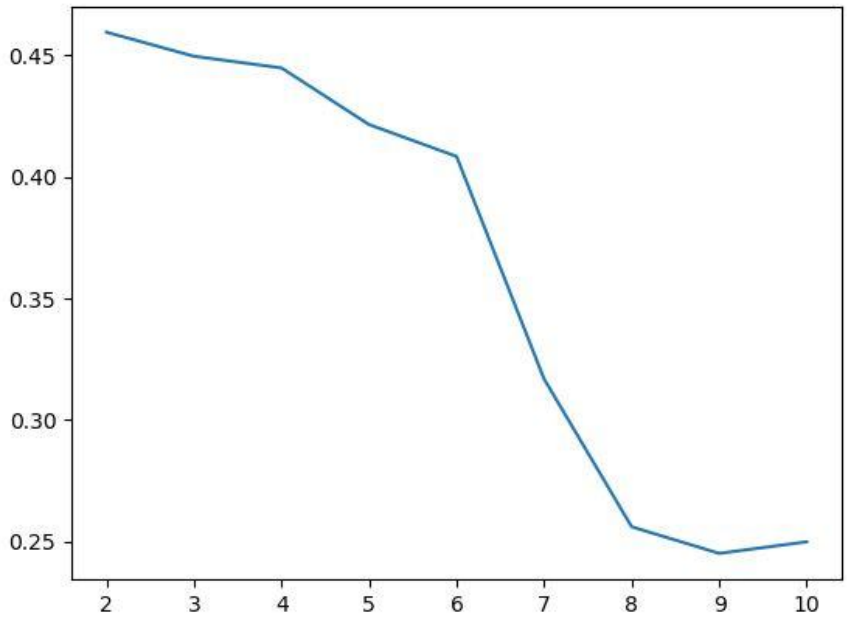


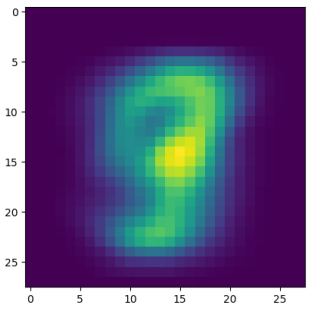
Figure 2: SSE is plotted as a function of k. Each k is run 10 times and the lowest final SSE value is plotted for each k.

As seen in figure 2, the SSE drops dramatically after iteration 6 and appears to start converging at iteration 8. Since we know the SSE will continue to decrease with an increasing k value, we can use the dramatic decrease in slope, after k=8, as an indicator that there are likely only 8 groupings in this data.

Principal Component Analysis

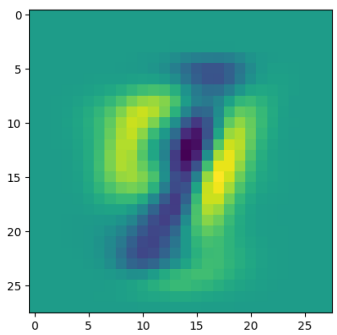
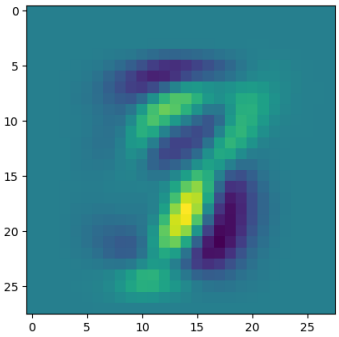
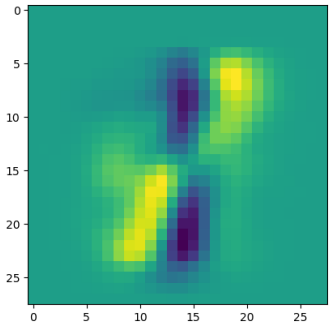
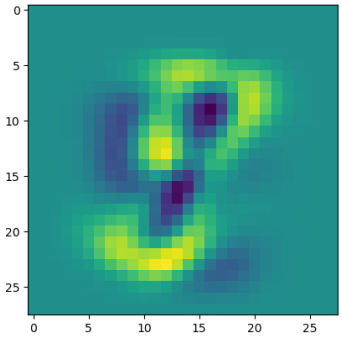
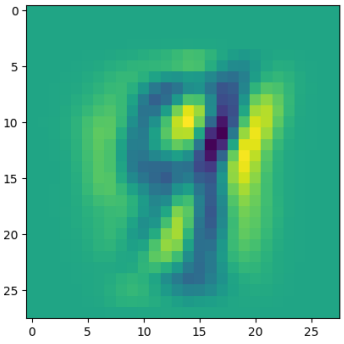
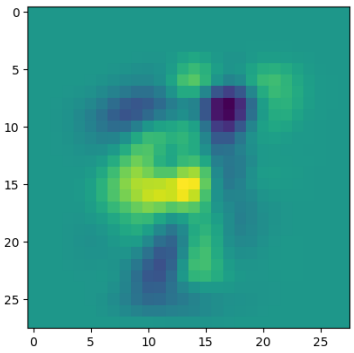
The top ten eigenvalues that we found in decreasing order are: [0.07296579, 0.04655825, 0.04006685, 0.02886374, 0.02557144, 0.02411997, 0.01766808, 0.01599164, 0.01394514, 0.01242548]

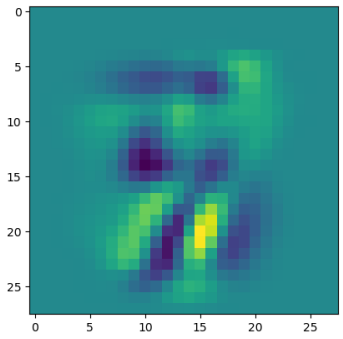
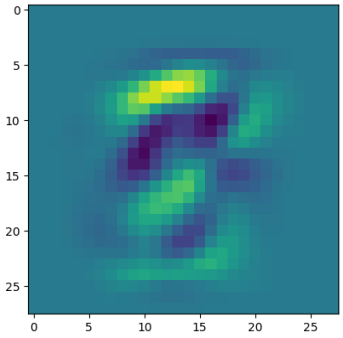
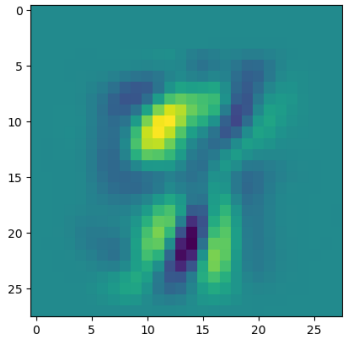
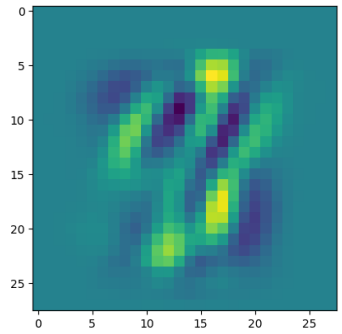
Below is the mean image generated by taking the mean value of each pixel in the dataset. This essentially gives us an "average" picture of all the images we are analyzing.



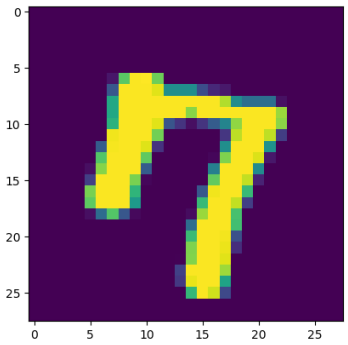
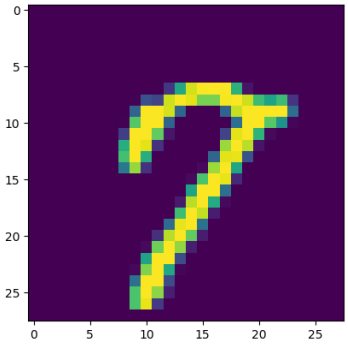
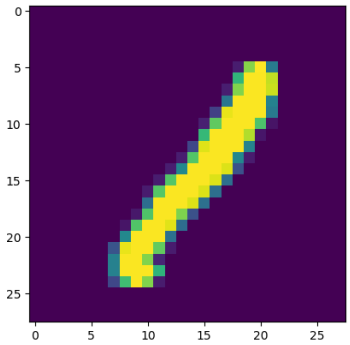
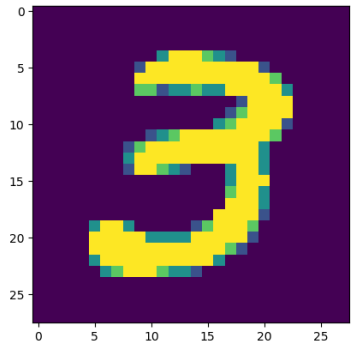
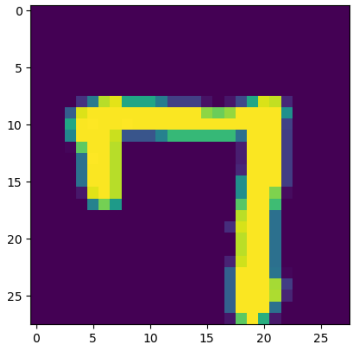
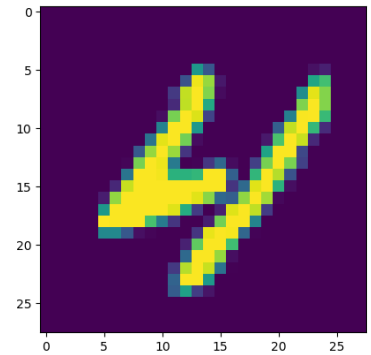
The following ten images are plots of the top ten eigenvectors (from left to right, top to bottom). They represent viewing the data from a certain plane to try and preserve as much variation as possible. The

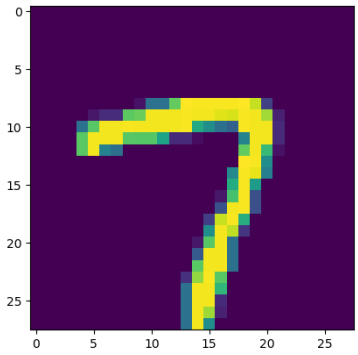
plot on the top left represents the first eigenvector, which should show the most variation in the data, while each plot after that shows progressively less variation.

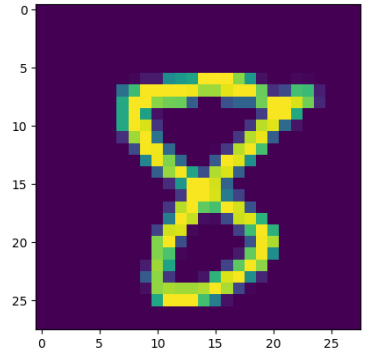







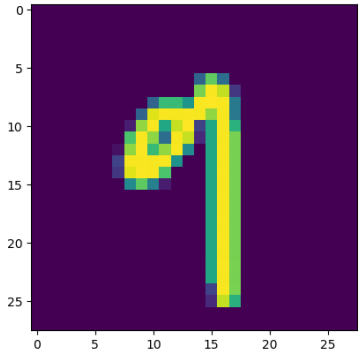





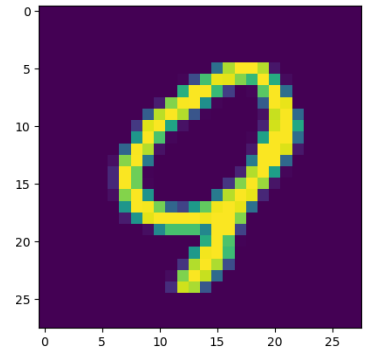
The ten images following this section are the plots of data points that have the highest values for each dimension after reduction has been applied. If you compare them to the corresponding eigenvector plots you can see that they tend to resemble each other, especially the first few images.


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