MATH 444 **Due Date: 01/30/2022**

Assignment 1

Ellis Wright (enw30)

1)The goal of this part is to visualize three dimensional data from the data set 'DataAssignment1S2022.mat'. Even though the data exists in three dimensions it is easy to see that we only need two dimensions to describe it as shown below in figures 1 and 2.

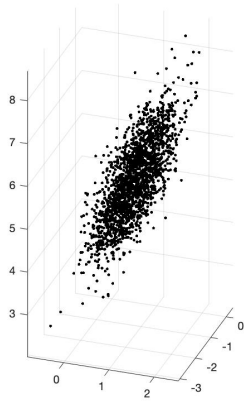

Figure 1: 3D View

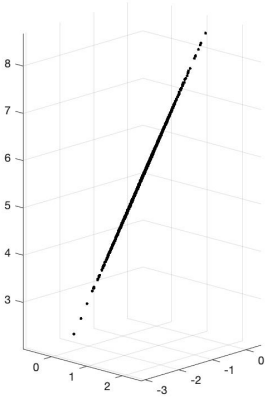

Figure 2: 2D View

If the data were in a higher dimension the only way we could get a visual is to project it down into two of the dimensions for an easy graph as shown in figure 3.

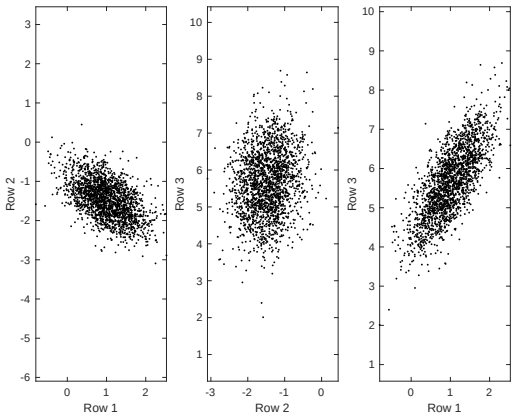

Figure 3: Projections onto two dimensions

Next we can focus on the differences of when we center the data or not. After centering the data and computing the singular value decomposition (SVD), we can see the the singular values, $\sigma_{\mathbf{c}} \approx \{43.3, 25.7, 0.1\}$. This confirms that the data can be treated as a linear combination of just two feature vectors. When we do not center the data the singular values, $\sigma \approx \{268.3, 25.8, 16.0\}$. This is because the first feature vector will correspond to the center of the data cloud. We can visualize the effect centering has on the data by plotting the principle components versus each other.

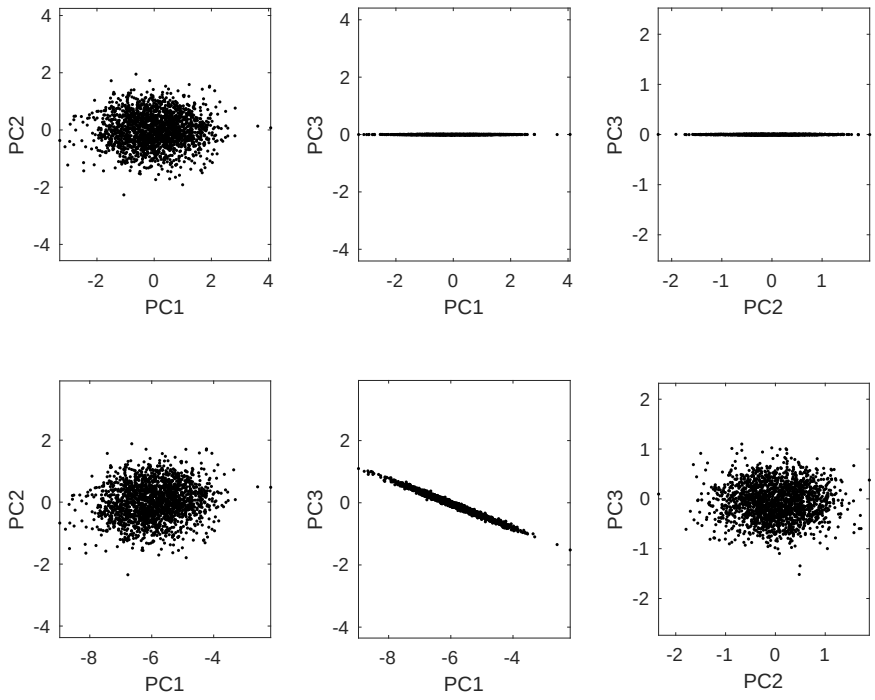

Figure 4: The top three graphs correspond to the principle components of the centered data while the bottom three graphs correspond to the principle components of the non-centered data.

As we can see from the centered data, all the information is stored in the first two principle components. This is not the case when we do not center the data.

2) For this section we want to investigate if PCA can be applied to the 'IrisDataAnnotated.mat' data. After applying the SVD we get singular values of $\sigma \approx \{25.1, 6.0, 3.4, 1.9\}$. These values suggest the first principle component, sepal length, does the best at classifying the type of Iris. We can plot PC2 vs PC1 to see if there is a presence of clusters.

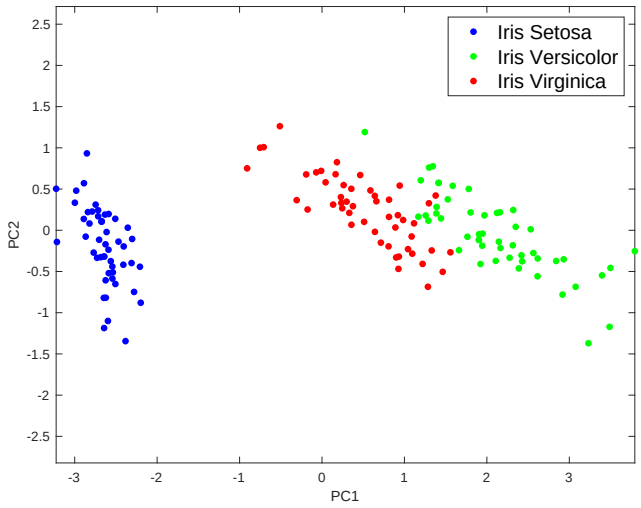

Figure 5: The iris data set shows presence of clusters when projected onto the PC2 vs PC1 plane.

As we can see there are three clusters. Iris setosa is the easiest to differentiate, while versicolor and virginica are more nuanced, yet still separated. The singular values show that the first PC is the most necessary so let's view a histogram of the data in one dimension to make it easy to see if we can separate the classes.

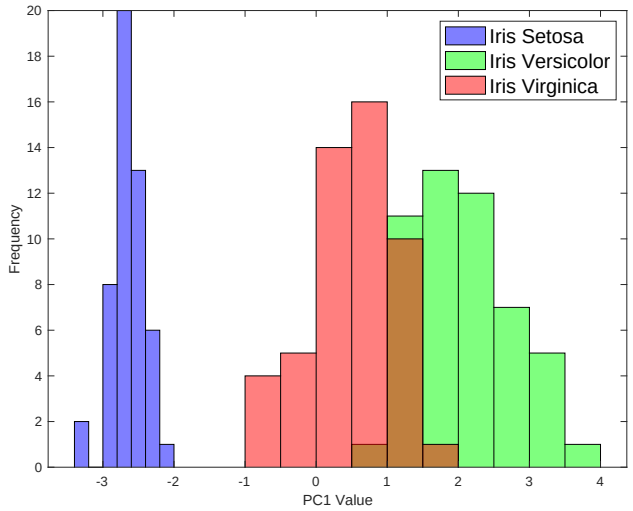

Figure 6: This shows a histogram of the data on the first principle component.

3) In this section we are going to investigate the 'HandwrittedDigits.mat' data with PCA. Without centering, we will specifically look at the data corresponding to the digits 0, 5 and 8. To do this we will create three data matrices each corresponding the data that is 0, 5, and 8. Then we will take the SVD of each to view the feature vectors (Figure 7) as well as several approximations (Figure 8) and their residuals (Figure 10).

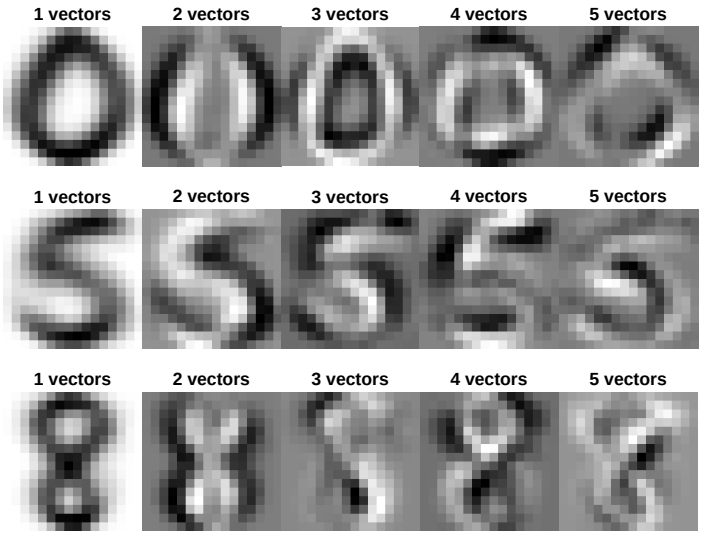

Figure 7: An image of the first five feature vectors for each digit.

Now, we want to see how PCA does when projecting the data down into smaller dimensions.

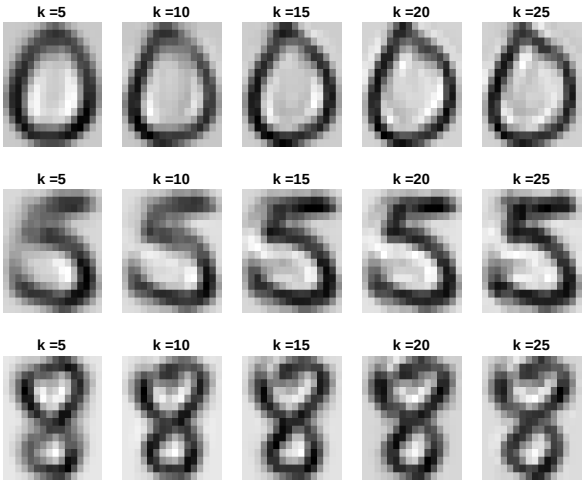

Figure 8: An image of different projected approximations for each digit.

For reference, here is an image of the handwritten digits we are approximating.

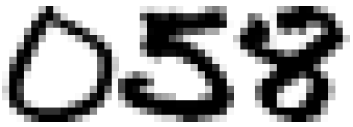

Figure 9: The reference images we are trying to approximate.

Let's see what the residuals look like when we project the data to 25 dimensions.

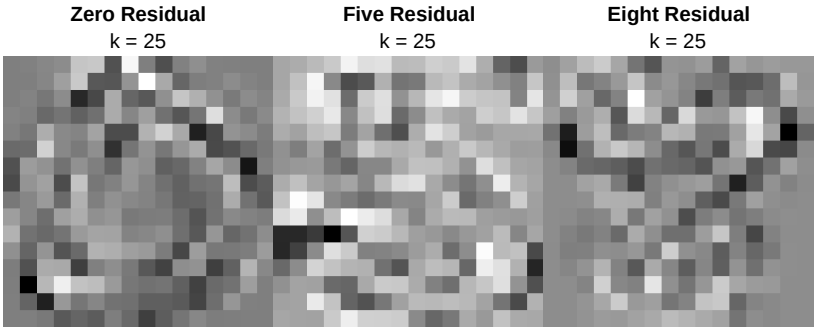

Figure 10: An image of the residual from an approximation.

Finally, let's see how the error improved as we increased the dimensions of the projection.

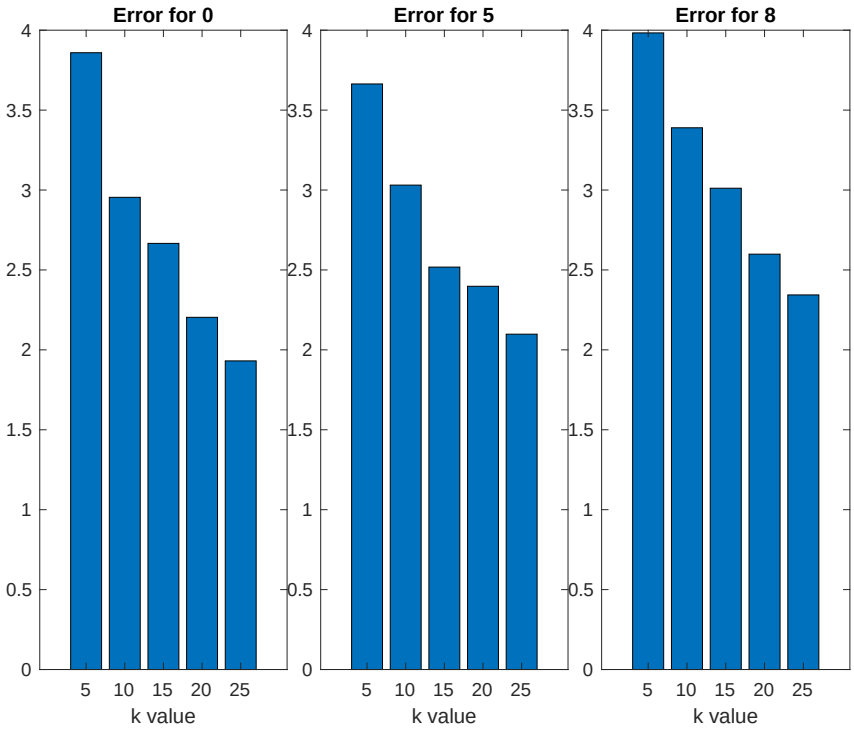

Figure 11: The euclidean norm of the residual for each digit.

As we can see, the error kept getting better as we projected into higher dimensions. We would expect the error to diminish as we get closer to the true dimension of the set. We would like to find a middle ground that has sufficiently small error with lower computational cost than the higher dimensional data set.