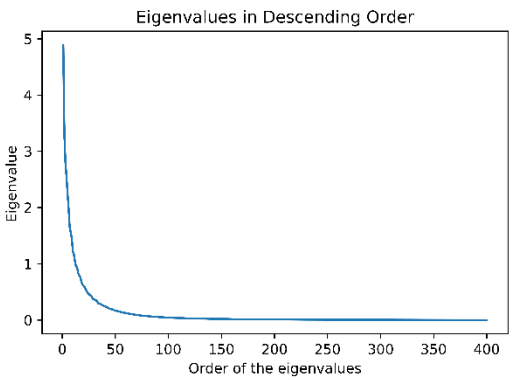
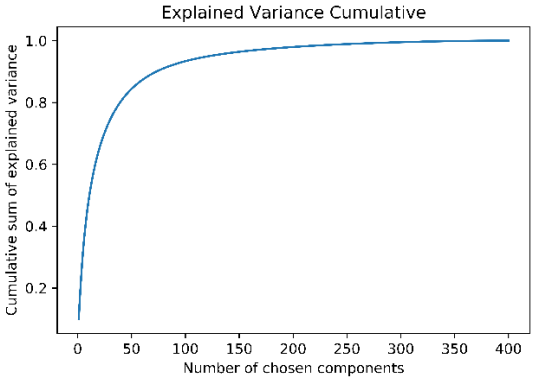
IE452/IE552: Algebraic and Geometric Methods in Data Analysis

Project-II Report

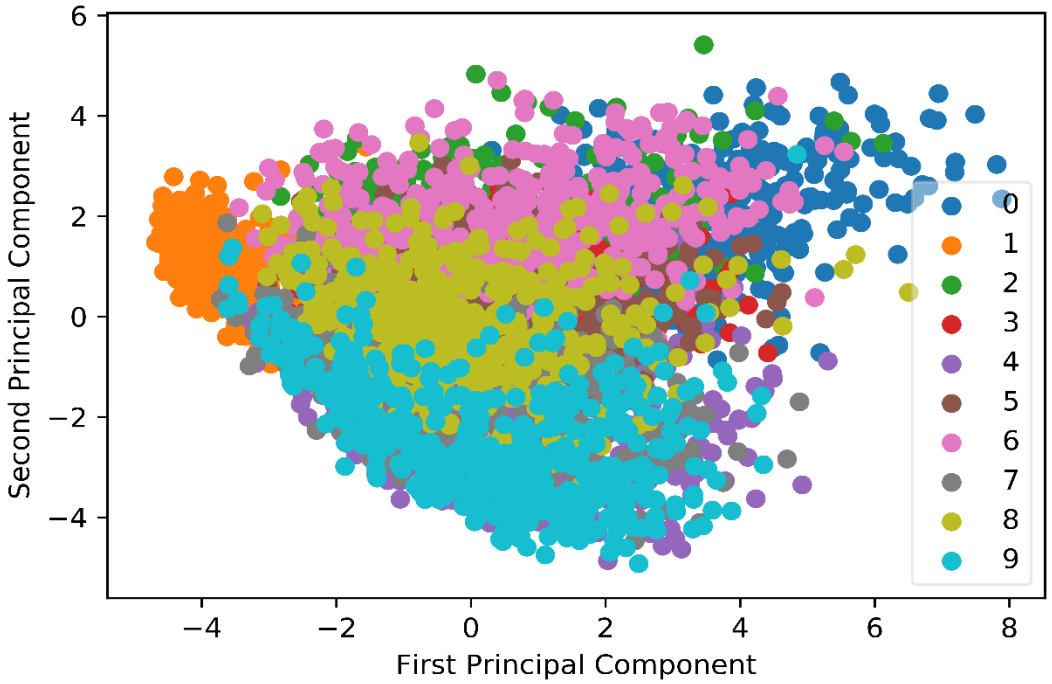
Batıhan Akça – 21502824
Emre Can Durmuş – 21601443

Question 1:



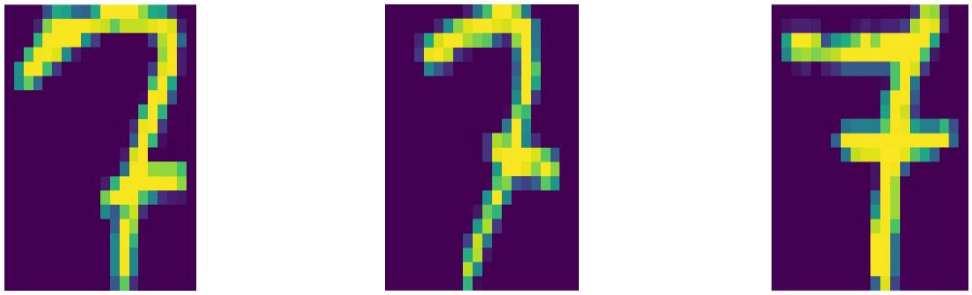


Eigenvalue change significantly reduces after 50. Eigenvalue and the eigenvalues begin converging to zero. Therefore, just by examining the plot on the left choosing 50 components seems reasonable.



Along the first principal component digit 0s are located on the most positive side and the digit 1s are located on the most negative side dominantly. This is the difference between having a density as a line in the middle like 1s or not like 0s. On the plot also there are 7s located near 1s which is expected

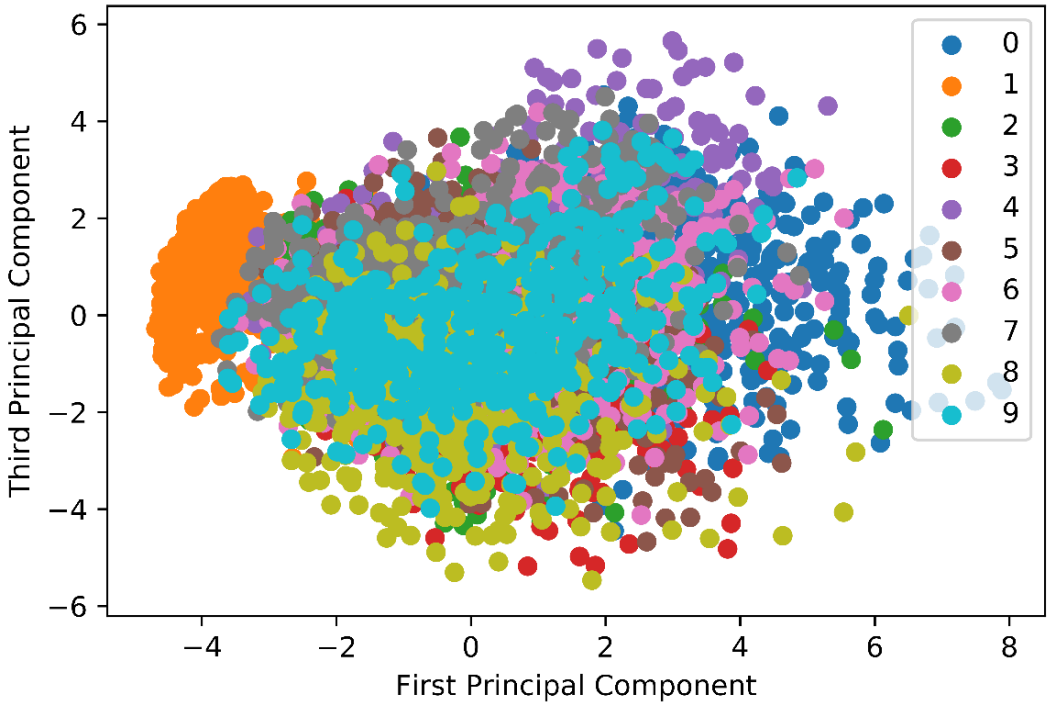
because shape-wise 1s and 7s are similar to each other. An ideal 1 and 7 may be disguisinable according to the first principal component because 7 does not contain density in the middle as a line as much as 1; however, since the dataset is about the handwritten digits, this may not be the case.



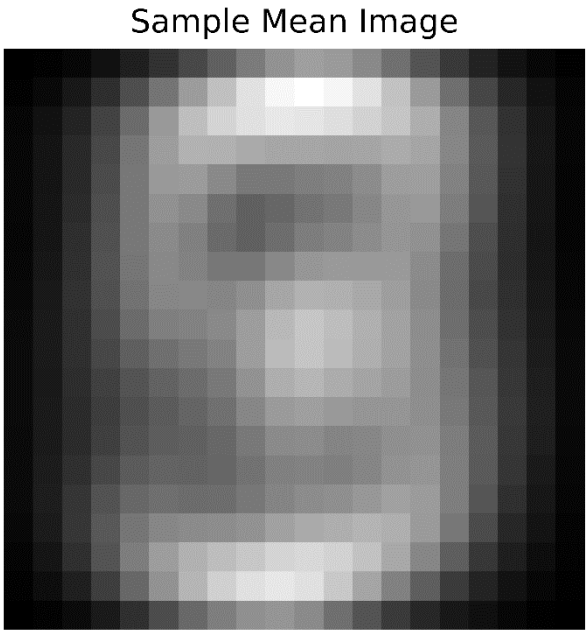
Some silly examples of 7s that have their vertical lines straight like 1 unlike an ideal 7 where its vertical and horizontal lines are not perpendicular.

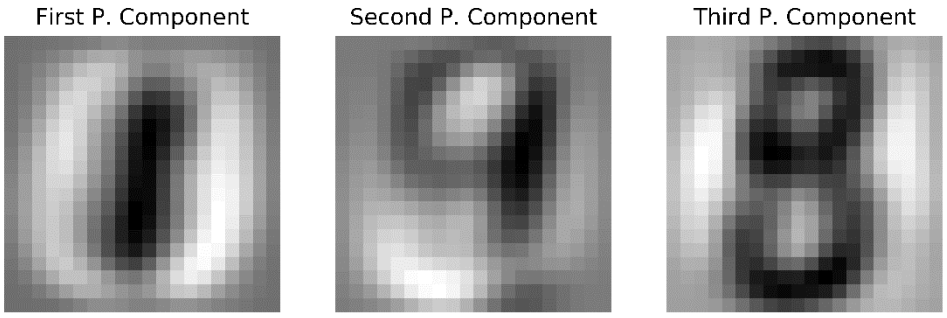
The first principal component also has the general circular shape information which means whether the digit has a circular shape or not. Hence, 1s and 7s are separated from others completely and located on the most negative side along the first principal component, but other digits are distributed on a large range since they have circular shapes and its level may change because of the handwritten diversity.

Along the second principal component we see a transition from the most positive values to most negative values as 3 layers digit 6, digit 8, digit 9 which shows that the second component has the information of the location of the curvy shapes it means along the second component digit 6 has the most positive values because its curvy shape is in the bottom and digit 9 has the most negative values because its curvy shape is in the top of the digit. Like a hybrid of digit 6 and digit 9, digit 8 is distributed between digit 6 and digit 9 on the scatter plot.



Along the third principal component, we see a transition from the most positive values to the most negative values as digit 4, digit 7, digit 1, digit 8, and digit 3. It shows that the third component has the information of overflowed pixels from the center to side edges which means unlike digit 3 or digit 8, digit 4 and digit 7 have some high-density pixels near the edges. It may also be inferred that there is a mixture of 7s and 1s along the third principle that may be happened because those two are similar according to the pixel densities which are near to the side edges.





The sample mean image shows that overall outer edges are empty except some pixels on the top and the bottom edge. This means that for those pixels, we expect very low variances near zero in covariance matrix and we expect high variances for the pixels that we can distinguish digits by densities of those.

The first principal component looks like a blurry digit 0 which is expected because on the scatter plot it is examined that the first principal component has the information of circular shapes. According to the first component, it may be inferred that we expect high variance for those pixels which forms a blurry 0, so we expect high pixel densities and low pixel densities in the data for those pixels and as a result, we saw dark greys in the sample mean image because the average is expected to be low.

As a more general comment on covariance matrix, by examining the sample mean image we may expect high variances for greyish pixels because we infer that we expect high pixel densities and low pixel densities for those pixels at the same so the average become grey and we may expect low variances

for the pixels which are clearly white or black because for white ones we expect very high pixel densities most of the time and for black one we expect very low pixel densities most of the time.

On the code 20 subspaces selected between 1 and 200 and projected.

Question 2:

First Eigenvalue	Second Eigenvalue	Eigengap
9.94426597026179	9.769810047931962	0.17445592232982854

Perturbed matrix's unique maximal vector is found by ordering eigenvalues in descending order and detecting the index of highest eigenvalue, but as a result, a complex-valued vector is found.

Our motivation to find the best/optimal r-rank projection is to minimize the mean-squared error.

$$\text{Mean squared error: } \mathrm{E}[\|X - \Pi_V(X)\|_2^2]$$

$$Where\ \Pi_V(X)\ is\ the\ projected\ vector\ along\ the\ V\ subspace\ of\ \Sigma's\ eigenvectors.$$

The mean-squared error can be written as for the optimal choice of V as

$$\mathrm{E}[\|X - \Pi_V(X)\|_2^2] = \sum_{j=r+1}^{d} \gamma_j^2\ (\Sigma)$$

Where the eigenvalues are ordered as $\gamma_1(\Sigma) \geq \gamma_2(\Sigma) \geq \gamma_3(\Sigma) \geq \cdots \gamma_d(\Sigma) \geq 0$

Therefore, for a fixed r it can be said that the optimal subspace V is spanned by the top r eigenvectors of Σ.

$$Projection\ Z = \sum_{j=1}^{r} \gamma_j\ \ (\Sigma)(v_j v_j^T)$$

Fixed Dimension r	Reconstruction Error
10	436.76227930048503
50	148.4758418856656
100	0

I hereby pledge on my honor that the work done on this assignment/examination is solely my own and I have not given and/or received any aid on this assignment/examination. I have also read and accept the 'Academic Integrity Code' of Bilkent Universtity.

Batıhan Akça Emre Can Durmuş