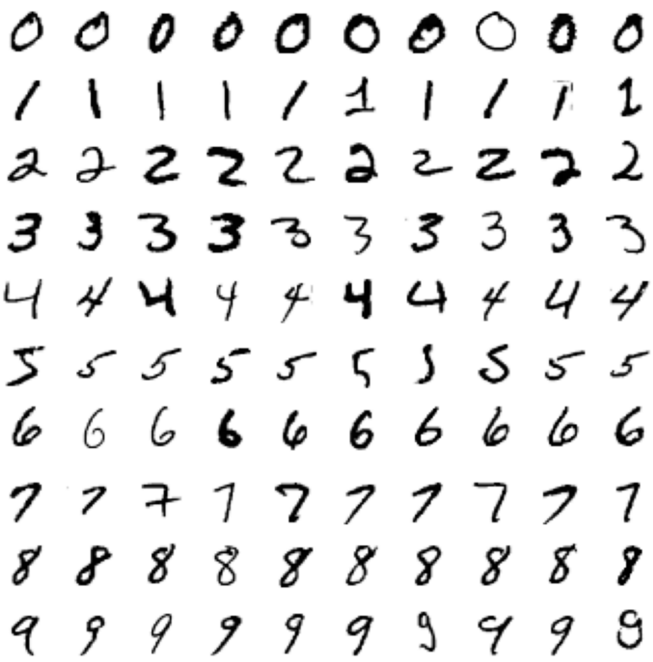
CMPE 481 Data Analysis and Visualization
Assignment 3 – Visualization Using PCA
Zuhal Didem Aytaç – 2018400045

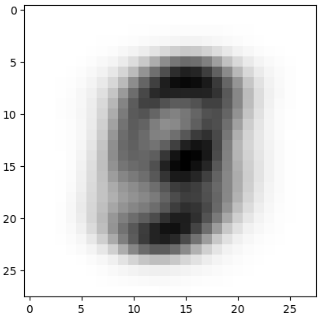
1) 10 Sample Digit Images Per Class

10 sample digit images per digit class are given below. The image is constructed using *plot_ten_sample_digits* function.



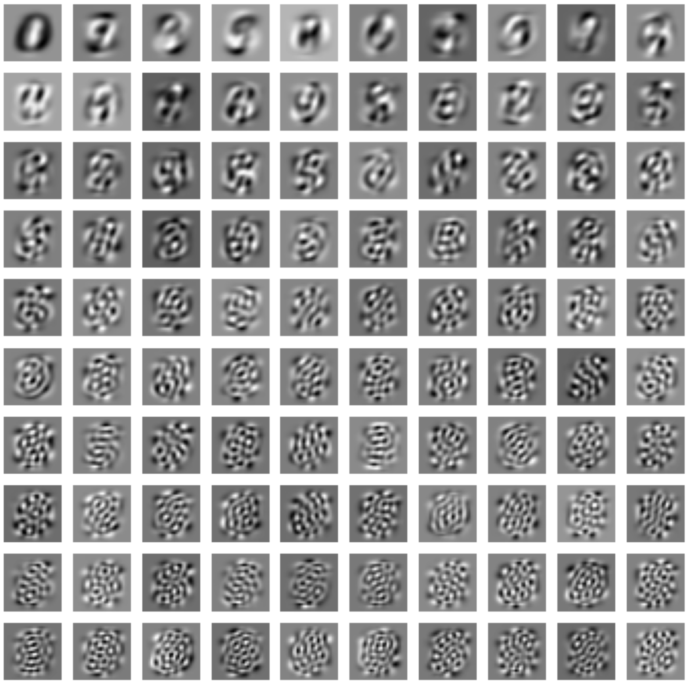
2.1) Mean Image using Training Set

Mean image of the train dataset is given below. The image is constructed using *plot_mean_image* function.



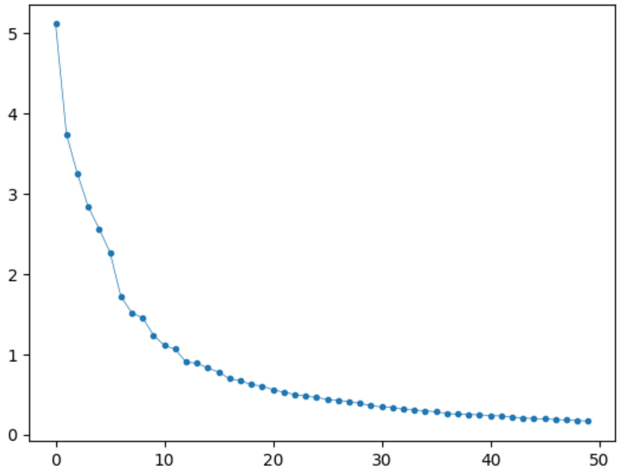
2.2) Eigenvectors using Training Set

Plot of largest 100 eigenvectors is given below. The image is constructed using *plot_100_eigenvectors* function.



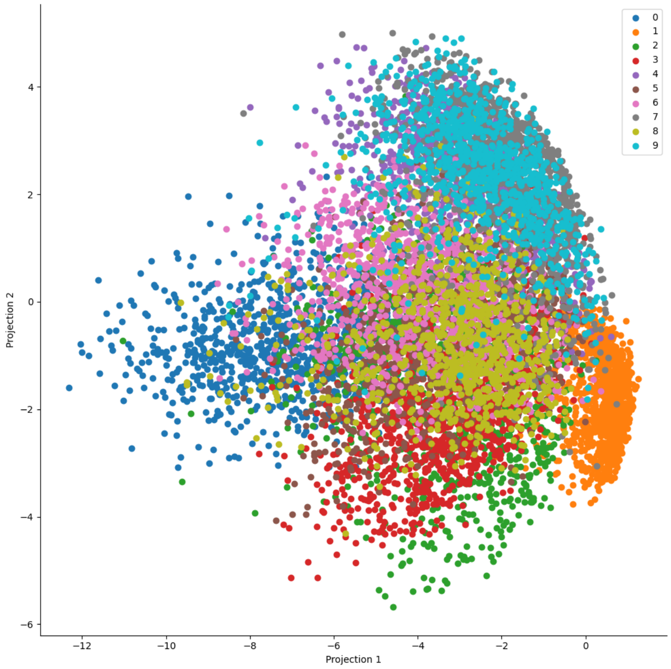
2.3) Eigenvalues using Training Set

Plot of largest 50 eigenvalues is given below. The image is constructed using *plot_50_eigenvalues* function.



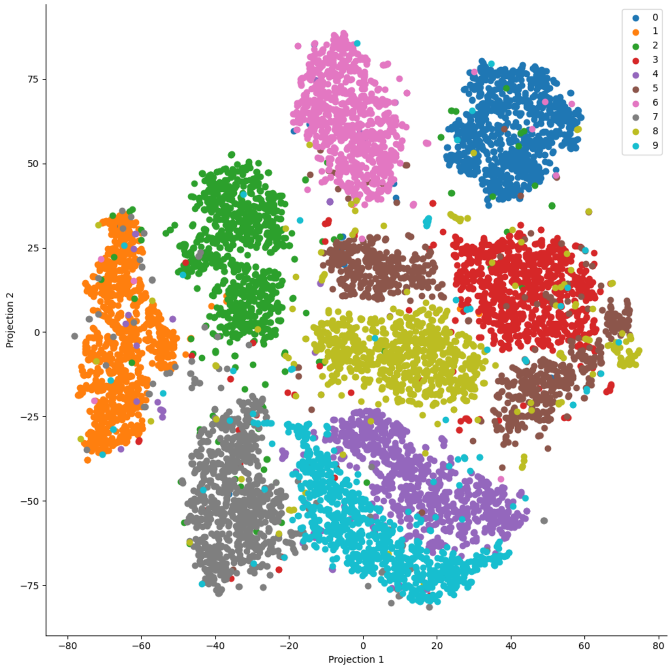
3) 2D Digits using PCA

Plot of 2D visualization of the test dataset using PCA is given below. The image is constructed using *plot_pca* function.



4) 2D Digits using t-SNE

Plot of 2D visualization of the test dataset using t-SNE is given below. The image is constructed using *plot_tsne* function.



5) t-SNE & PCA Comparison

We use them both for dimensionality reduction and we both reduce our dataset to two dimensions and plot them on 2D surfaces as above. As seen in the plots in previous page, the 2D plot of digits using t-SNE proposes a much cleaner clustering of different digits, however we cannot see that distinction in PCA. By looking at the resulting plots, we can say that, the t-SNE method gives a clear visualization of the similarities between same digits. However, PCA does not give a clustering of the data points with the same label.

The difference between the resulting plots lies in the fundamentals of the PCA and t-SNE methods. As PCA aims to preserve maximum variance. The disadvantage is that during PCA, small variations between digits that look alike is out of focus and the whole dataset looks on top of each other. On the other hand, t-SNE is good at preserving small variances and visualizing a clean clustering. T-SNE has the advantage of identifying tightly related data points. The disadvantage of t-SNE is that the algorithm is more complex in terms of time.

Another insight from the plots is that, we can see that similar looking digits' clusters are near or on top of each other in PCA, like 7 and 9 are on top of each other, 6-9-8 have really close clusters. However, we can see that, the nearest two clusters to 6 are 2 and 0, the nearest two clusters to 9 are 4 and 7. That observation suggests that, the clusters positions in t-SNE does not give any clues about the closeness of the relation between the clusters.

Another disadvantage of PCA is that we can still view some noise in the reduced plot, and the amount is not ignorable. Also, we can see that on the plot, there are 2 different and separate clusters for digit 5.

Overall, we can say that t-SNE successes to preserve the global structures and proposes a good clustering for datapoints of the same label in a 2D plot. However, it is complex in terms of time, it may produce noisy outputs and it loses the closeness relationship between different clusters during dimensionality reduction and fails to reflect that on a 2D plot.