A Report on Principal Component Analysis (PCA)

Date: September 29, 2018

Prepared by: Ratnakar Rao Mallayagari (201550807)

Contents:

1. Problem Statement:

Given a dataset of images, perform ***Principal Component Analysis (PCA)*** on those images and come up with a representation in lower dimensional space. Also construct the images using optimal number of principal components and analyze the loss incurred and variance captured using various number of principal components.

2. A Brief Note on PCA:

PCA is a technique used to represent the data in a lower dimensional space. The following are some of the notes on PCA

- It is aimed at retaining as much of information as possible by minimizing the square error loss and consequently maximizing the variance captured in the new space.
- This representation of data with minimum square error loss or maximum variance is achieved by projecting the data from higher dimensional space onto the sub-space with principal components as its basis.
- PCA is an unsupervised technique where it does not use the class information of the dataset
- It is used in feature extraction procedures.

3. PCA Procedure:

In this section, the procedure to perform PCA analysis is described with variables, conventions and mathematical formulations involving vectors, matrices and operations on them.

N – Number of Samples
D – Number of dimensions
X – Given dataset with NxD dimensions

Note: *The typesetting for the below mathematical formulations is prepared by me using LATEX in overleaf. The link to this overleaf latex is mentioned in Appendix section.*

$$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{1}$$

$$C = X - \mu \tag{2}$$

$$S = C^T C \tag{3}$$

Once we have the *Covariance Matrix(S)* of the centered data, the eigen vector set U of S are calculated and sorted in descending order of their respective eigen values λ. The first *k* eigen vectors in this descending order of their eigen values will form the principal basis in *k* dimensions. When the Centered data, *C* is projected onto this *k* dimensional basis B we get the coefficients of *C* in the new principal basis. dimension of B is *Dxk* where *k* is the number of principal components considered.

$$Z = C.B \tag{4}$$

Thus, we see that dimension of *Z* is *Nxk*. We can visualize this as, *all the N samples are represented in k dimensional space* as opposed to D dimensional space earlier. This results in the representation of original data in lower dimensional space.

Note: There are conventions to standardize the data after centering so that the covariance matrix is taken on the centered and standardized dataset.

4. PCA on High Dimensional Data (D>>N):

We see that the dimensions of *Covariance Matrix(S)* is *DxD*. For high-dimensional dataset with large number of dimensions as compared to the number of samples, the Covariance Matrix, S turns out to be huge and the calculation of eigen vectors is complex. Due to this, we resort to a trick to ease the calculations in high-dimensional space.

$$Sv = \lambda v \tag{5}$$

$$X^T X v = \lambda v \tag{6}$$

Now, pre-multiplying both sides with *X*

$$X X^T X v = \lambda X v \tag{7}$$

Substituting, C=*Xv* in the above equation

$$X X^T C = \lambda C \tag{8}$$

The above is again an eigen value problem. In high-dimensional dataset, we find the eigen vectors of XX^T. We see that XX^T dimension is *NxN* which is much less that *DxD*. Premultiplying the above equation with X^T gives us,

$$X^T X X^T C = \lambda X^T C \tag{9}$$

It is clear from the above equation that X^TC is the eigen vector set of X^TX. Thus, we conclude that, in order to find the eigen vectors of X^T, we first find the eigen vectors of XX^T and pre-multiply these eigen vectors with X^T to get the eigen vectors of X^TX

Once we have the eigen vectors of X^TX, the rest of the procedure for PCA is same as discussed previously. We will have to sort them in descending order of their respective eigen values λ. The first *k* eigen vectors in this descending order of their eigen values will form the principal basis in *k* dimensions. If *B* is principal basis with dimensions *Dxk*, then projecting *C*, having dimensions *NxD*, onto *B* will result in new matrix *Z*, with dimensions *Nxk*.

5. About Dataset:

- The dataset used to demonstrate *Principal Component Analysis (PCA)* has 520 number of samples
- Each original image has dimension 256x256x3. For this demonstration of PCA, the gray scale image is used after removing color gradient. Thus, the image used for this demonstration has 256x256 dimension i.e., after flattening it has 65536 dimensions.
- Please note that all the 520 images are used in Principal Component Analysis. But, only 8 of the 520 original samples are shown below for brevity (one image per person).

  For the sake of referencing these images, let us use the following names for these images

alice **andy** **bob** **bran** **mary** **abc** **pqr** **john**

6. Analysis Overview:

Given the above dataset with 520 images, PCA has been performed using various number of principal components. The following are the observations after running PCA with various number of principal components ranging from 1 to 520.

- *Mean squared error (MSE)* reduces considerably with each increment in number of principal components(k).

- When considering the *error percentage* with respect to the centered and standardized image we see that the loss is less than 20% with just k=32 components

- *Variance* contribution of initial components is much higher than variance contribution of later components.

- Overall *variance* captured is very high and reached 99% with just first 13 principal components

- Improvement in image *clarity* as the number of principal components increase

- A short summary of the characteristic data for Principal Component Analysis

- Detailed image clarity improvements for all the persons with increasing number of principal components

- Scatter plots of images belonging to two classes having high *between-class* variance

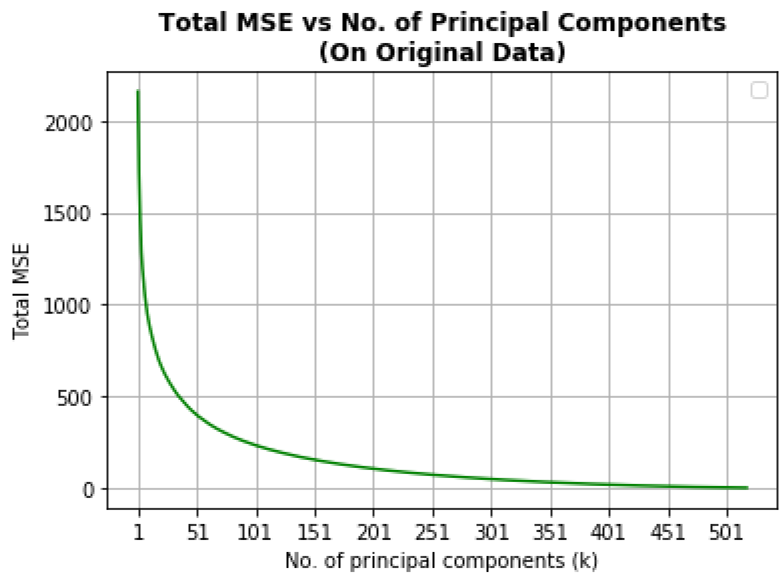
The above observations are substantiated with graphs, images, tables and results in the following section.

7. Substantiation of the Observations:

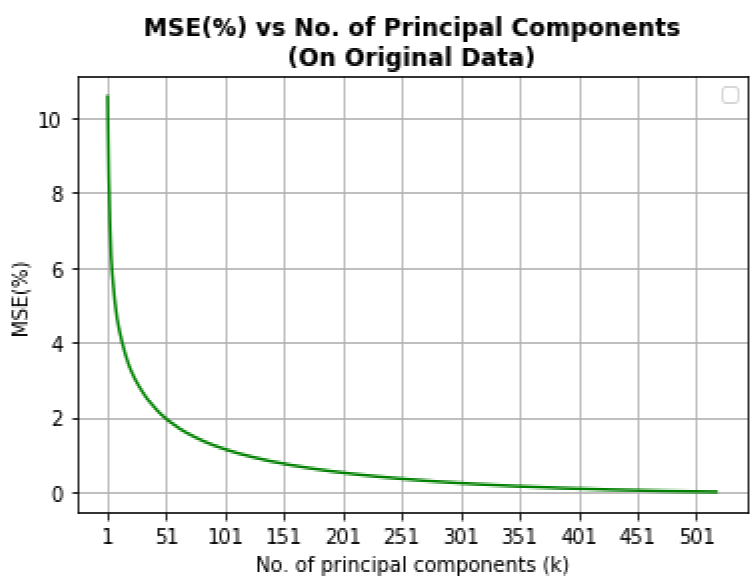
Observation 1: *Mean squared error(MSE)* reduces considerably with each increment in the number of principal components (k).

Case 1 a): Total MSE with respect to *original input data*

Total MSE = Mean of sum of squared error across all input samples

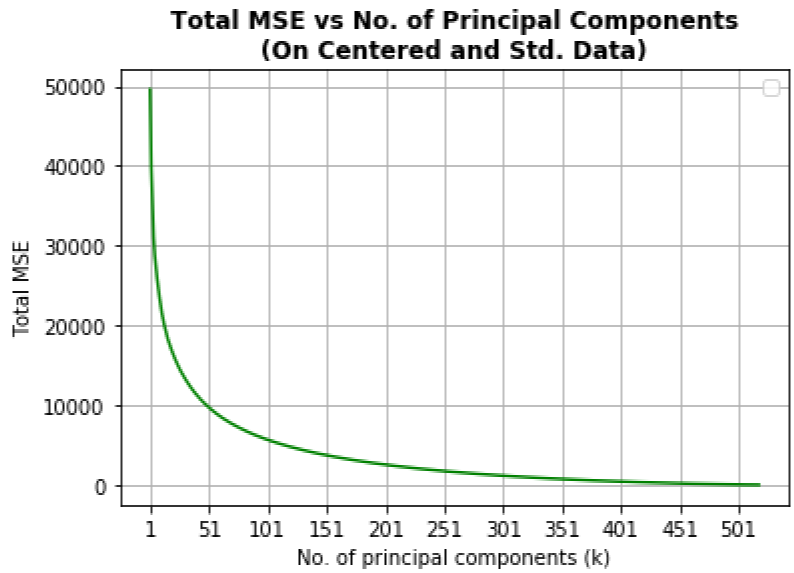


Case 1 b): MSE % with respect to *original input data*



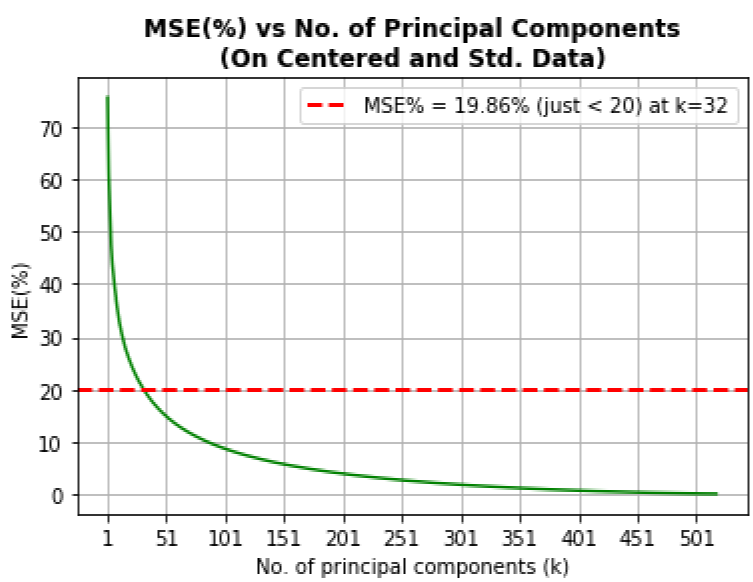
Case 2 a): Total MSE with respect to *centered and standardized input data*

Total MSE = Mean of sum of squared error across all centered and standardized input samples

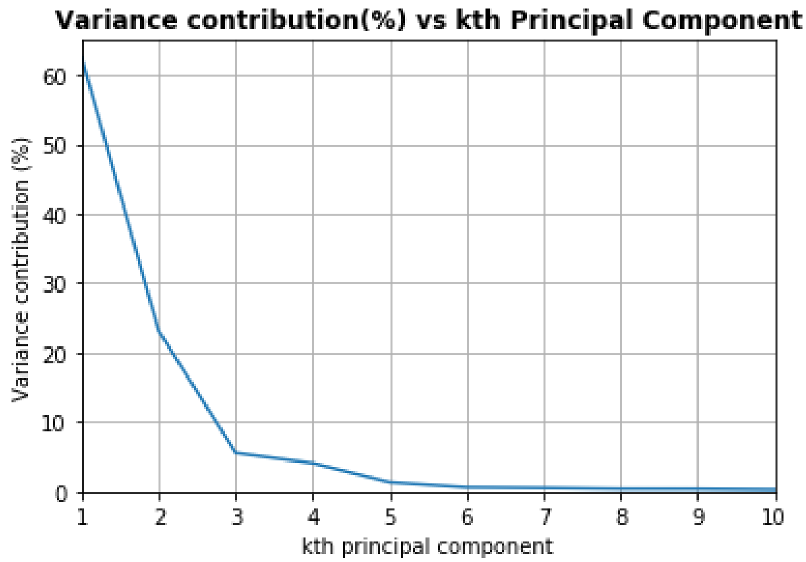


Case 2 b): MSE % with respect to *centered and standardized input samples data*

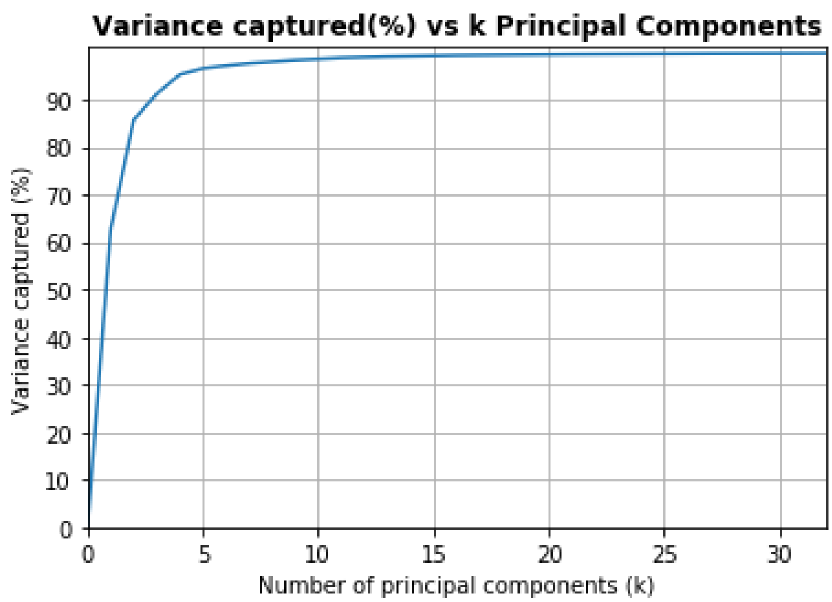
Observation 2: When considering the *error percentage* with respect to the original images, we see that the error percentage is less than 20% with just k=32 components



Observation 3: *Variance* contribution of initial components is much higher than variance contribution of later components.



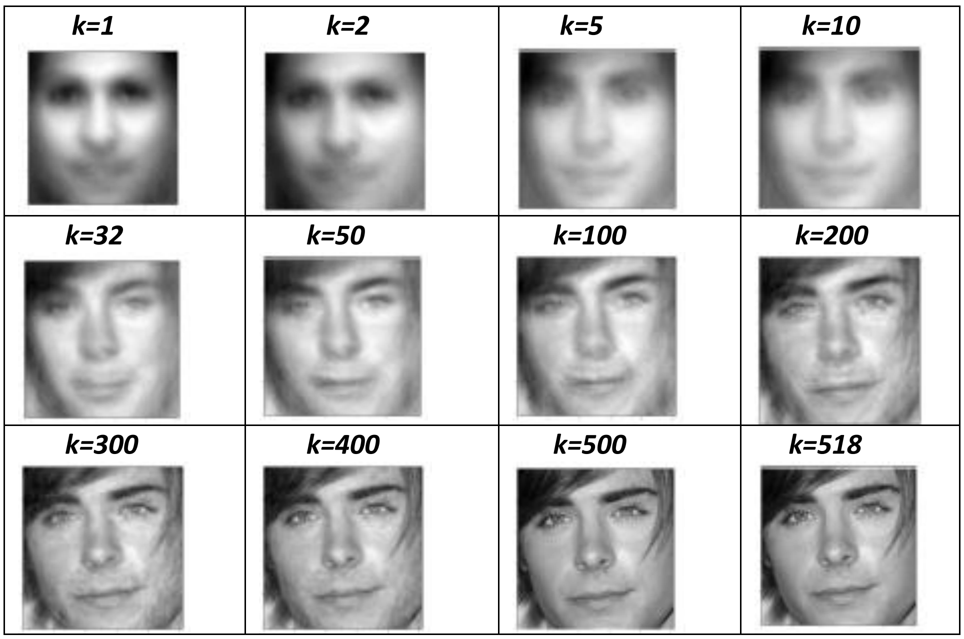
Observation 4: Overall *variance* captured is very high and reached 99% with just first 13 principal components



Continued…

Observation 5*: Improvement in image clarity with increase in number of principal components.*

As the number of principal components are increase, the error reduces, and variance captured increases. As a result, the amount of loss of information in the image is less. Thus, we see steady improvement in the clarity of the image. The following figure demonstrates the improvement in the image clarity of 'John' when we increase the number of principal components in PCA. For the sake of brevity an image of one person is considered to show the improvement in clarity.



Improvement in image clarity with increase in number of principal components (k).

Continued…

Observation 6: *Summary of Empirical Results*
Loss and variance measurements with increasing number of principal components. The table below shows the observations for few number of principal components for brevity.

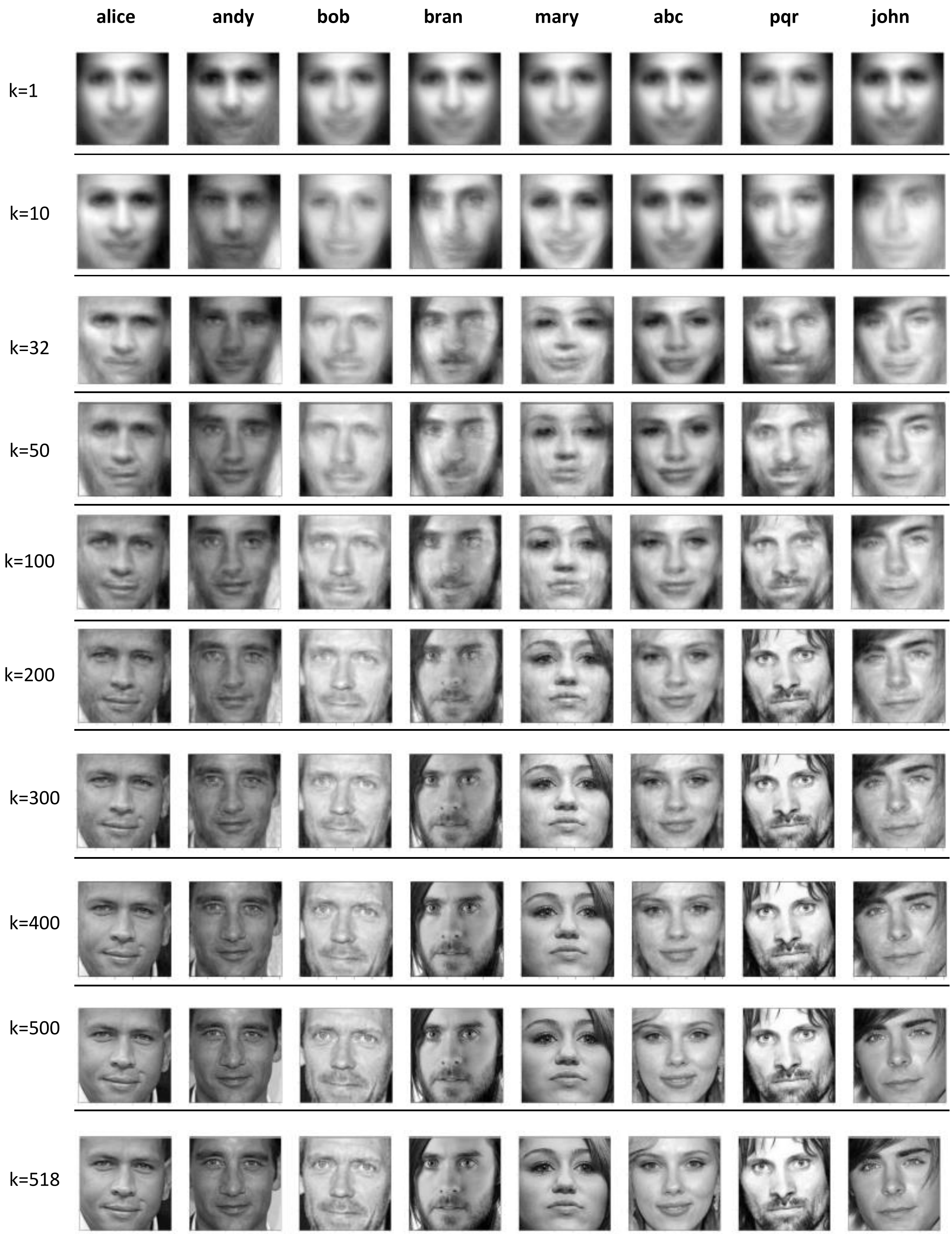
No. of Principal Components (k)	Original Loss	Original Loss Percent	Variance Captured Percentage	Centered & Std. Loss	Centered & Std. Loss Percent
1	2159.02	10.58	62.54	49586.35	75.66
5	1187.43	5.82	96.56	28782.35	43.92
10	913.22	4.47	98.54	22511.78	34.35
20	675.65	3.31	99.46	16598.82	25.33
32	532.52	2.61	99.73	13017.60	19.86
50	402.80	1.97	99.88	9836.70	15.01
100	234.12	1.15	99.97	5696.04	8.69
150	153.83	0.75	99.99	3734.70	5.70
200	104.85	0.51	99.99	2542.17	3.88
250	71.59	0.35	100.00	1732.30	2.64
300	47.69	0.23	100.00	1151.58	1.76
350	30.18	0.15	100.00	728.10	1.11
400	17.23	0.08	100.00	415.28	0.63
450	7.82	0.04	100.00	187.70	0.29
500	1.33	0.01	100.00	31.57	0.05
518	0.00	0.00	100.00	0.00	0.00

Note:

a) From the above results, it can be clearly seen that, as we increase the number of principal components *Total MSE* reduces and *Total Variance* captured increases and thus loss of information is minimum when we include all the principal components. It can also be seen that first few principal components capture the significant amount of information.
b) We see that we reach minimum with 518 components instead of 520 because 2 of the images in the given dataset are noisy without or cut image of a person.

Continued…

Observation 7: Resulting image clarity of each person with increasing number of principal components (k)



Observation 8: *Scatter plots of images of two persons with high between-class variance*
Note: *For brevity, considering one image of* ***andy*** *and one image of* ***mary***

<u>Case 1: Image constructed with 1 principal component</u>

Scatter plot of images belonging to **andy** and **mary**

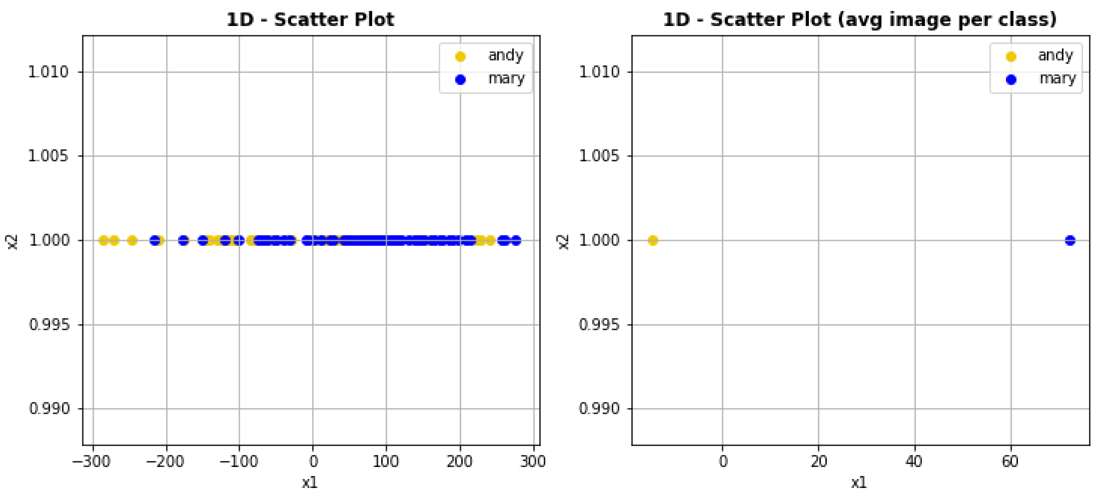


Figure: Left: all images of andy and mary in 1-D. Right: avg. image of andy and mary in 1-D

<u>Case 2: Images with 2 principal components and their scatter plots in 2-D:</u>

Continued…

Scatter plot of all the images belonging to **andy** and **mary**

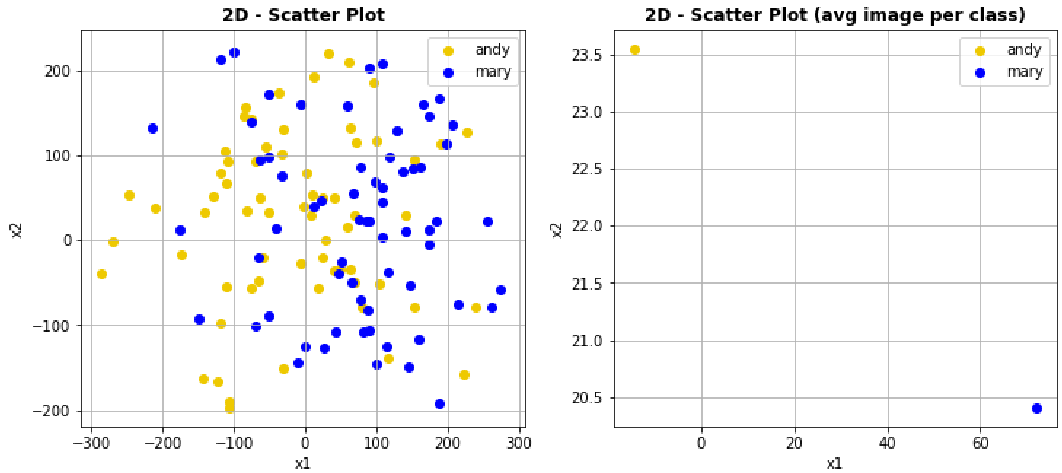


Figure: Left: all images of andy and mary in 2-D. Right: avg. image of andy and mary in 2-D

__Case 3: Images with 3 principal components and their scatter plots in 3-D:__

Scatter plot of all the images belonging to **andy** and **mary**

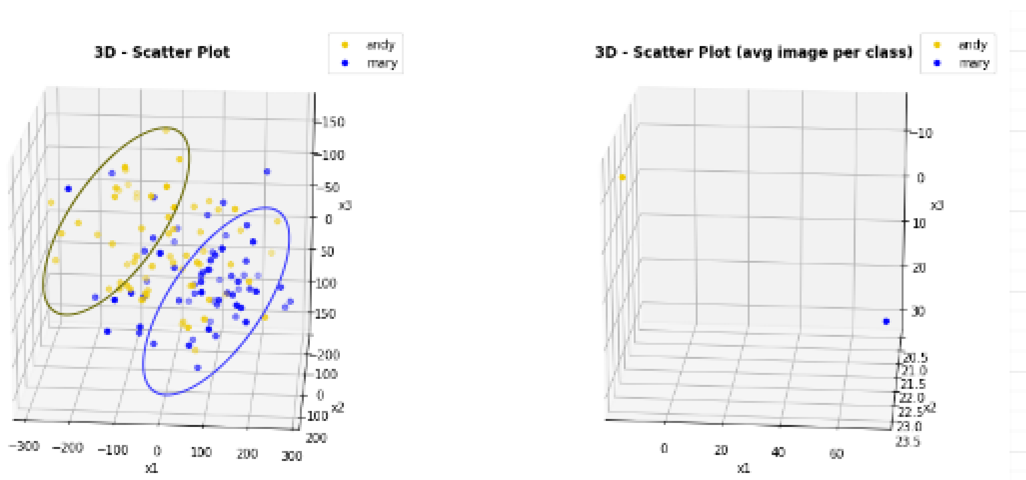


Figure: Left: all images of andy and mary in 3-D. Right: avg. image of andy and mary in 3-D

Observation Note: As the number of principal components or dimensions increases, we see that the similar images start forming clusters and move farther apart so that they are clearly distinguishable between classes

8. APPENDIX

As the following has comprehensive data, they are put in this Section for complete reference.

- Code Snippet of PCA implementation
- Scatter plots including images of all the classes irrespective of their intra-class variance and inter-class variance.
- Complete Results of PCA components by incrementing the number of principal components gradually

a) Code snippet:

```
def pca(X,num_components):
    Xbar, mu, std = normalize(X)
    O=(1/N)*(Xbar@Xbar.T)
    eig_vals, eig_vecs = eig(O)
    U = Xbar.T @ eig_vecs
    U = normalize_by_col(U)
    U = U[:,0:num_components]
    Z=Xbar@U
    reconstructed_data=Z@U.T
    centered_loss = np.square(Xbar - reconstructed_data).sum(axis=1).mean()
    centered_loss_percent = ((np.square(reconstructed_data -
Xbar).sum(axis=1).mean())/(np.square(Xbar).sum(axis=1).mean()))*100
    var_captured_percent =
((np.square(eig_vals[0:num_components]).sum())/(np.square(eig_vals).sum()))*100
    reconstructed_data=reconstructed_data*std+mu.T
    original_loss = np.square(X - reconstructed_data).sum(axis=1).mean()
    original_loss_percent = ((original_loss)/(np.square(X).sum(axis=1).mean()))*100
    return Z, U, reconstructed_data, centered_loss, centered_loss_percent, var_captured_percent,
original_loss, original_loss_percent
```

b) Scatter plots of all the images in the dataset

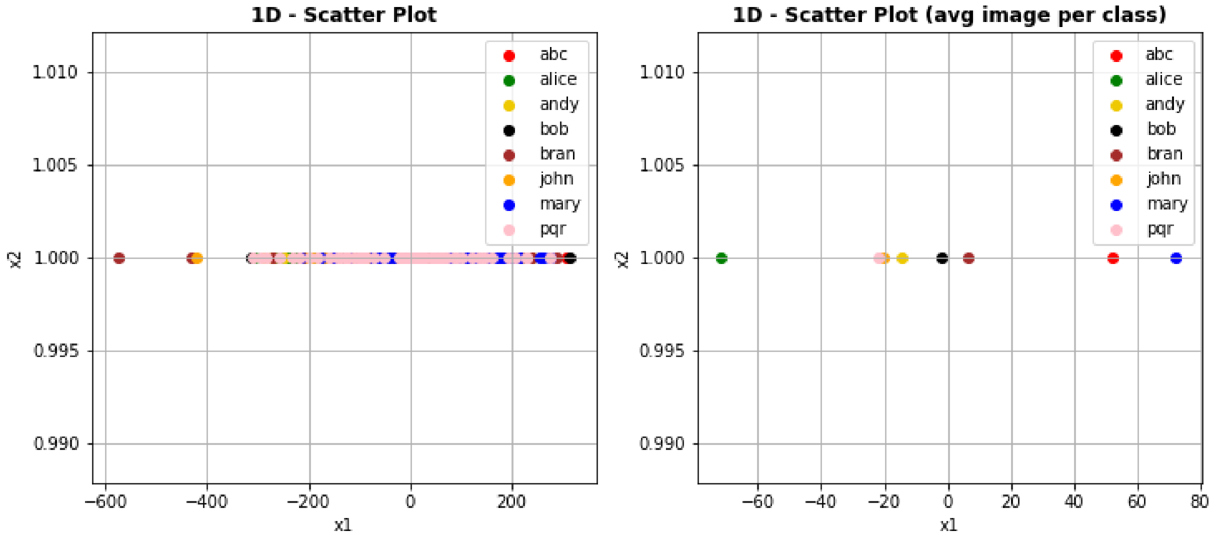


Figure: Left: all images in dataset in 1-D. Right: avg. image of all persons in 1-D

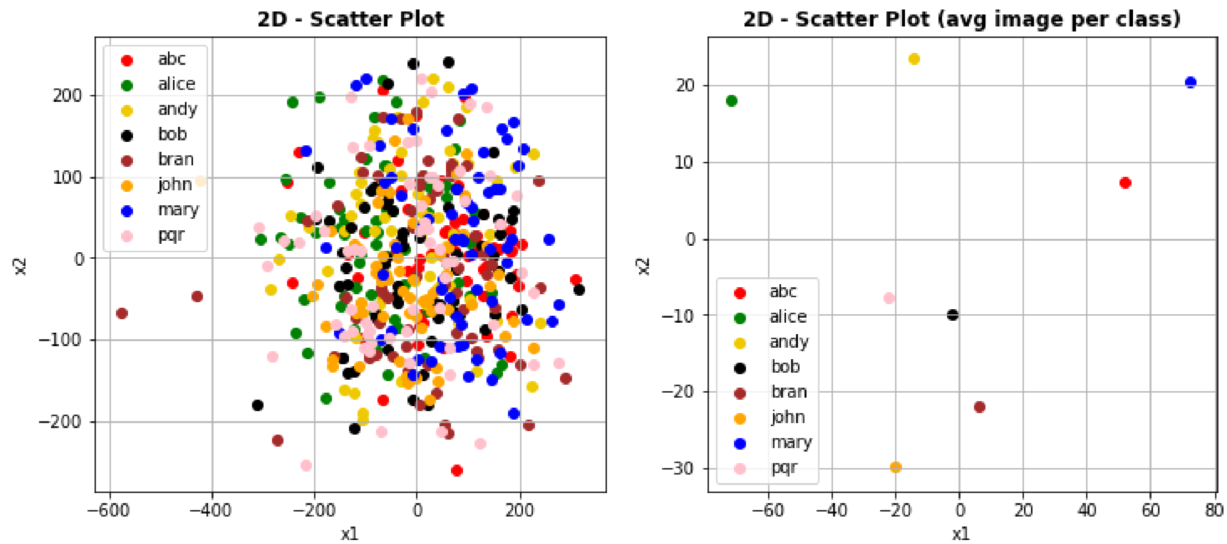


Figure: Left: all images in dataset in 2-D. Right: avg. image of all persons in 2-D

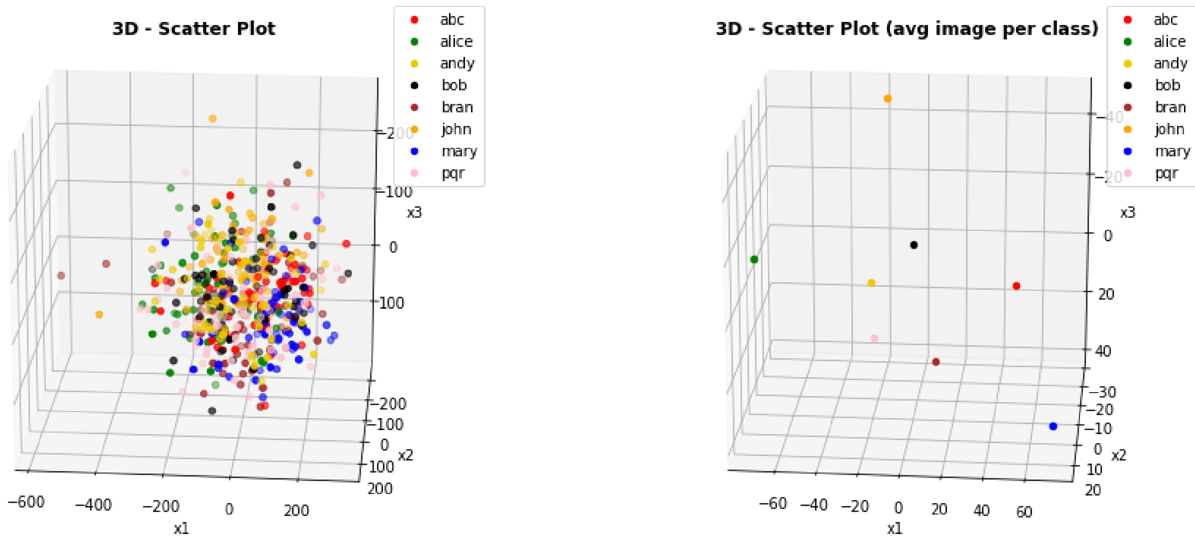


Figure: Left: all images in dataset in 3-D. Right: avg. image of all persons in 3-D

c) Complete PCA Results

All the PCA results data from my program is captured in a file and shared at my google drive link for reference if required.

https://drive.google.com/file/d/1igs_KkqzBdi32h028AWGzV9jFf8YIVw0/view?usp=sharing

d) Mathematical formulation of PCA

The link to the typesetting prepared by me for mathematical formulation of PCA is as below:

https://www.overleaf.com/read/kxqsyqvvrfhz