CS6105 Linear Algebra and Random Processes Assignment (PCA & Test of fit)

Anju Gopinath (CS15M057), Arulkumar S (CS15S023)

November 26, 2015

1 Principle Component Analysis (PCA)

PCA is the process of finding the principle directions in which the given dataset has more variance, such that when we project the data on the principle directions, we will be able to represent the data with lesser dimension. Thus reducing the dimension to represent the given data. i.e., we will be able to transform a (possibly) larger number of variables into a (smaller) number of uncorrelated variables.

1.1 Procedure followed

- First, the given data **X** (as a whole, including all classes) is mean-subtracted.

  $Y = X - \mu_X$

- The covariance matrix of the mean-subtracted data (Y) is found using $(YY^T)/N$.

- Next, to find the Eigen vector decomposition (EVD) of covariance matrix, the below trick is used to reduce the computation.
  $(YY^T)\phi^{'} = \phi^{'}\wedge$, where $\phi^{'}$ is a matrix, in which every column is an Eigen vector of YY^T & $\wedge$ is a diagonal matrix of Eigen values of YY^T

  Now multiply Y^T on both sides, we get

  $(Y^TY)Y^T\phi^{'} = (Y^T\phi^{'})\wedge$

  Thus, the Eigen vectors of Y^TY is $Y^T\phi^{'}$

  For the current face recognition problem, every sample is of 10304 dimensions & there are 400 samples available in total. By using the above linear algebra trick, we need to find EVD for a 400 x 400 matrix , rather than 10304 x 10304 matrix. Thus, reducing the computation cost.

- The Eigen vectors of covariance matrix are sorted in the descending order of Eigen values, since the Eigen values represent the magnitude variance along the direction of particular Eigen vector.

- The given data is projected onto the space, in which the top-k eigen vectors are the basis.

  Projected images $= Y * Q_{pca}$, where Q_{pca} is a matrix whose columns are Eigen vectors of covariance matrix & Y is a matrix of mean-subracted data.

- To reconstruct the data, the Projected images are multiplied by inverse basis matrix & the frobenius norm is calculated for the error between original data & reconstructed data.

- Only the Eigen vectors which has more than 1% variance are considered as Basis for space of Principle components.

- After the transformed data is collected, the data is splitted into 70% training data & 30% test data.
- The Bayes classifier with full, diagonal covariance matrix (a parametric method) & K-NN (3 neighborhood) (a non-parametric method) are used to classify the data & the perfromance is measured.

1.2 Observations

The dimensions of the dataset is reduced from 10304 to ~20 dimensions when considering the Eigen vectors which have less than 1% variation.

The parametric method (Bayes classifier) gives classification efficiency of 88.33% with Same covariance matrix for all classes.

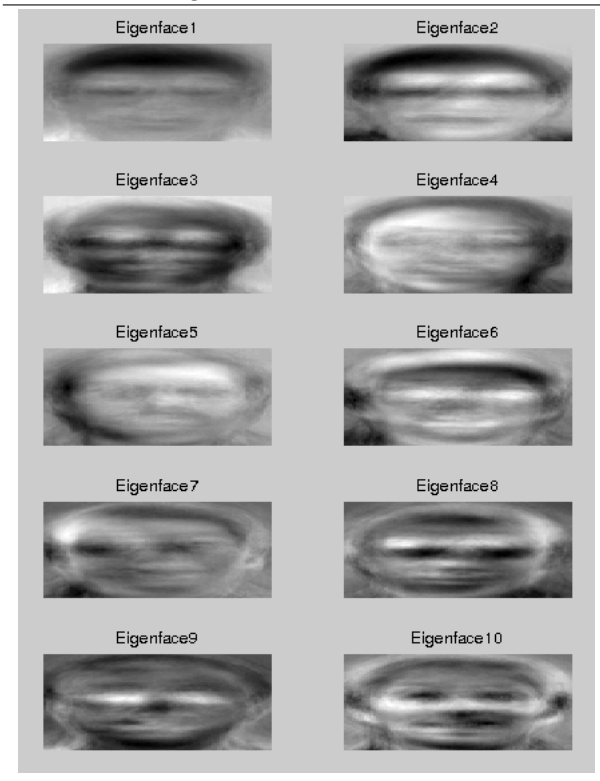
The non-parametric method (K-NN (3 neighborhood)) gives classification efficiency of ~86.67%.

1.3 Plots and Images

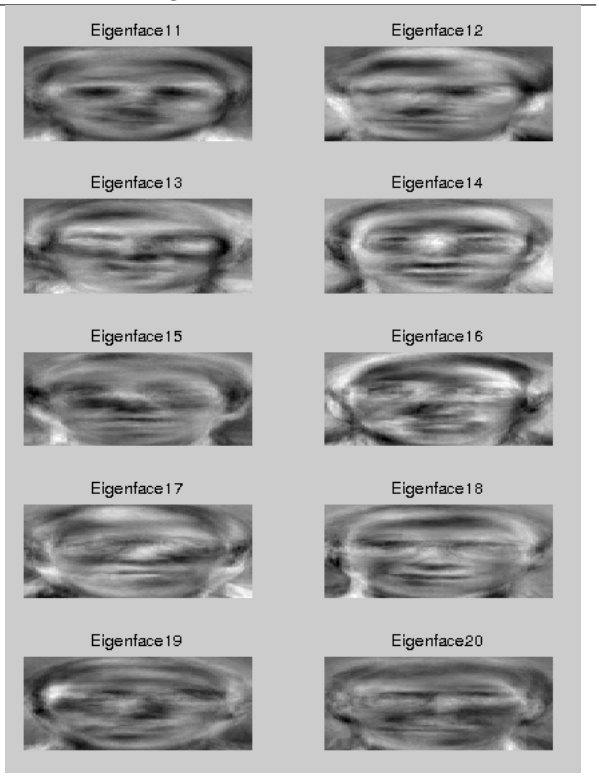
1.3.1 Eigen faces

Below are the top varying eigen faces found from the covariance matrix from all the training data. These eigen faces will act as standard orthogonal basis & all the images in training data will be projected on these eigen faces to reduce the dimension.

eigen faces from 1 to 10



eigen faces from 11 to 20



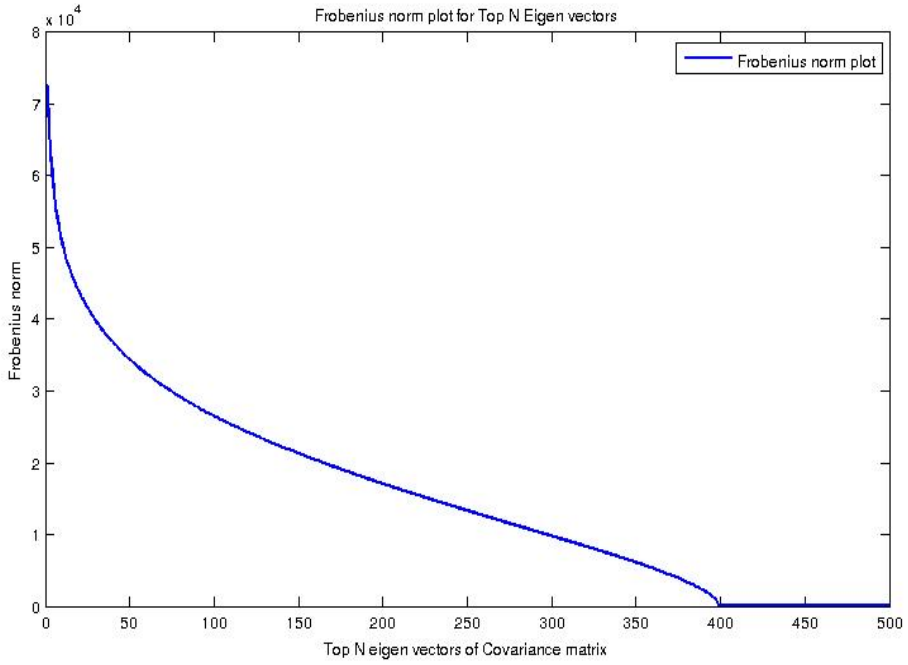
1.3.2 consider the original image

1.3.3 Relative error plot for reconstructed image

top most N eigen vectors	reconstructed image	relative error image
Top most 25		
Top most 50		
Top most 75		
Top most 100		
Top most 125		
Top most 150		

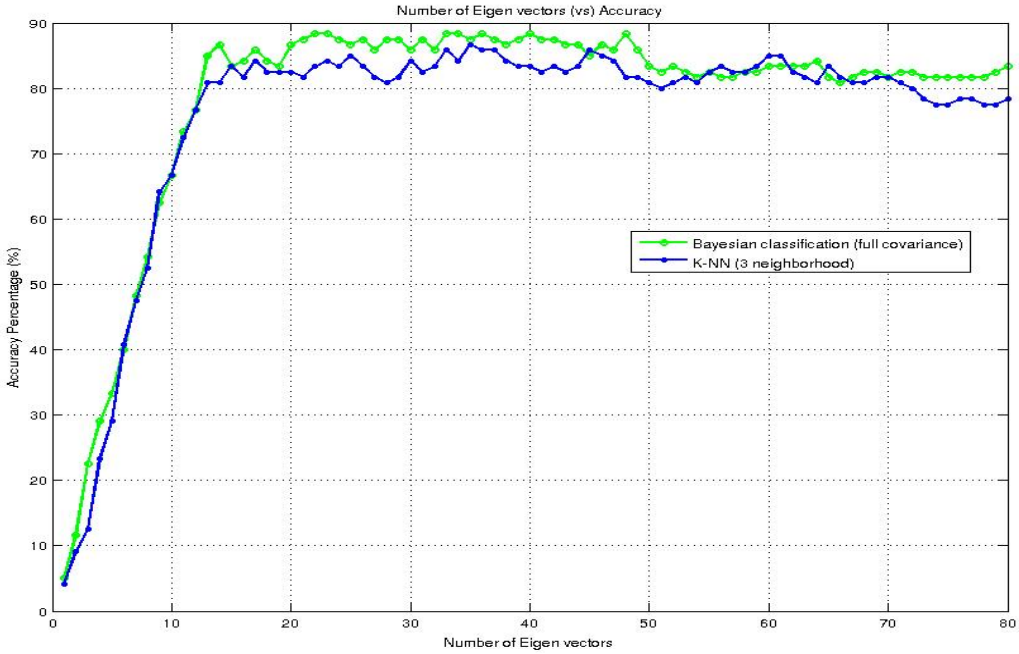
1.3.4 Frobenius error plot

In the below graph, we can see that the Frobenius norm is getting reduced, as we add more and more Eigen vectors. Since there are only 400 examples in our dataset, there can be atmost 400 non-zero eigen vectors. We can verify that after adding 400 eigen vectors, the full image would have been completely reconstructed.



1.3.5 Variation of accuracy w.r.t., Top N eigen vectors

The parametric (Bayesian decision framework) and non-parametric (K-NN (3 neighborhood)) model reach their best accuracies (88.33% and 86.67% respectively) when we include ~20 top Eigen vectors of covariance matrix. Even when we add more Eigen vectors after achieving this high performance, the performance is not getting increased.



2 TEST OF FIT

The test of fit of a statistical model describes how well it fits a set of observations.It summarizes the discrepancy between observed and expected values for the model in question.
Here, we have performed the Chi-Squared test and Student's t test

2.1 Chi-Squared Test

$$\tilde{\chi}^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k}$$

where O is the observed value and E is the expected value.

2.2 Student's t test

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{{s_1}^2}{N_1} + \frac{{s_2}^2}{N_2}}}$$

where x is the mean, s2 is the variance and n the number of observations.

2.3 Procedure for Chi-Squared Test

- Given the data, divide it into bins and calculate the no of points falling in each bin. This is the observed value.
- Calculate the parameters of the data for the assumed distribution using Maximum Likelihood Estimation and Method of Moments
- Plug in the parameters into the formula for the assumed distribution.
- Divide the area calculated into bins and calculate the points falling in each bin. This is the expected value
- Perform chi-squared test. If the value obtained is less than the critical value, hypothesis H0 is accepted, else, H1 is accepted, where H0 is the hypothesis that the assumed distribution is correct and H1 is the hypothesis that it is incorrect.

2.4 Procedure for Student's t Test

- Given the data, divide it into training and testing data.
- Calculate the various parameters for both the datasets using Maximum Likelihood Estimation and Method of Moments for the assumed distribution.
- Perform student's t test.If the value obtained is less than the critical value, hypothesis H0 is accepted, else, H1 is accepted, where H0 is the hypothesis that the assumed distribution is correct and H1 is the hypothesis that it is incorrect.

2.5 Observations

Table 1: Chi Squared -Continuous Data - Gaussian Distribution -MLE

bins	X^2	Critical Value
14	144.46	16.919
18	94.3532	27.5
22	87.0058	32.67

Table 2: Chi Squared- Continuous Data-Gaussian Distribution -MOM

bins	X^2	Critical Value
14	144.2861	16.919
18	94.1887	27.5
22	86.8380	32.67

Table 3: Students t - Continuous Data - MLE

train:test	t value
600:400	.7686
700:300	.0362
500:500	1.1743

Table 4: Students t - Continuous Data - MOM

train:test	t value
600:400	.7678
700:300	.0361
500:500	1.1732

Table 5: Chi Squared - Discrete Data

X^2 value	Critical Value
4.1889	5.009

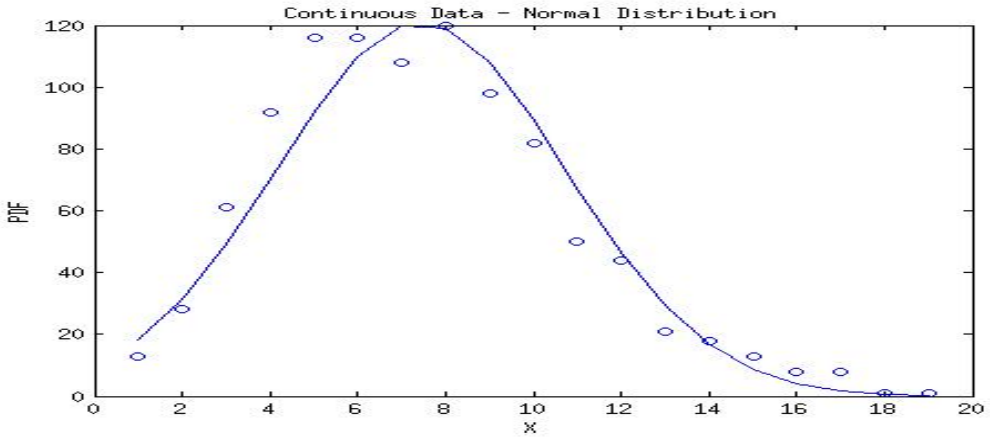
Table 6: Students t - Discrete Data - MLE

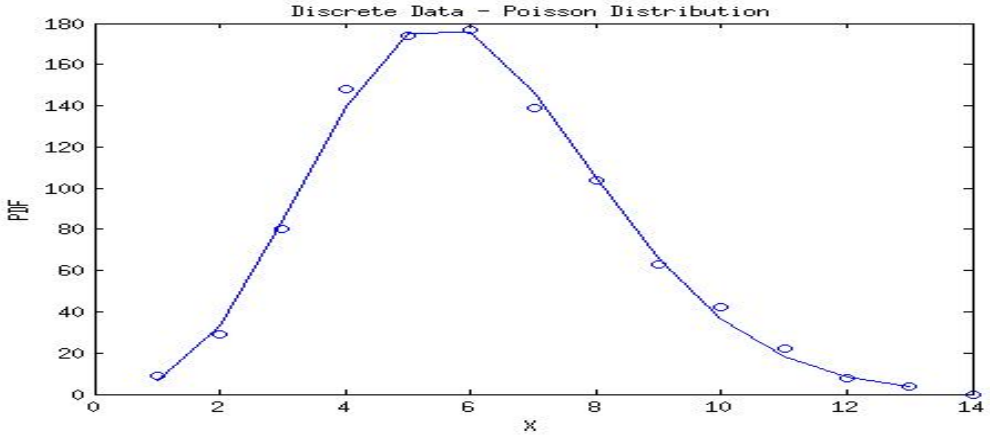
train:test	t value
600:400	.3182
700:300	.5005
500:500	.9136

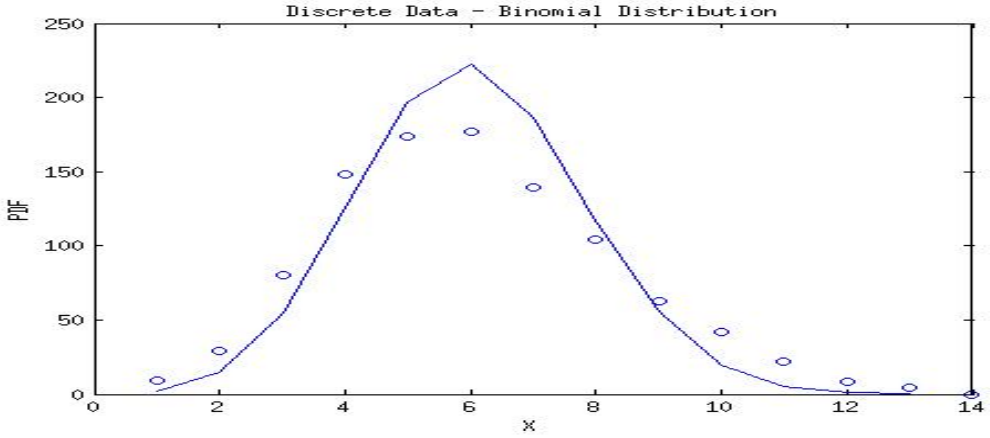
Table 7: Students t - Discrete Data - MOM

train:test	t value
600:400	.3179
700:300	.4998
500:500	.9127

2.6 Plots


Continuous Data - Normal Distribution
PDF
X
0
20
40
60
80
100
120
0
2
4
6
8
10
12
14
16
18
20


Discrete Data - Poisson Distribution
PDF
X
0
20
40
60
80
100
120
140
160
180
0
2
4
6
8
10
12
14


Discrete Data - Binomial Distribution
PDF
X
0
50
100
150
200
250
0
2
4
6
8
10
12
14