# IC 272: DATA SCIENCE - III
# LAB ASSIGNMENT – III
## Attribute normalization, standardization and dimension reduction of data

**Student's Name: Madhur Jajoo** **Mobile No: 7597389137**

**Roll Number: b20211** **Branch: DSE**

**1 a.**

**Table 1 Minimum and maximum attribute values before and after normalization**

| S. No. | Attribute | Before normalization | | After normalization | |
|---|---|---|---|---|---|
| | | Minimum | Maximum | Minimum | Maximum |
| 1 | pregs | 0 | 13 | 5 | 12 |
| 2 | plas | 44 | 199 | 5 | 12 |
| 3 | pres (in mm Hg) | 38 | 106 | 5 | 12 |
| 4 | skin (in mm) | 0 | 63 | 5 | 12 |
| 5 | test (in mu U/mL) | 0 | 318 | 5 | 12 |
| 6 | BMI (in kg/m$^2$) | 18.2 | 50 | 5 | 12 |
| 7 | pedi | 0.078 | 1.191 | 5 | 12 |
| 8 | Age (in years) | 21 | 66 | 5 | 12 |

**Inferences:**

1. outliers increase variability in data which leads to decrease in statistical power.
2. We have used median to replace the outliers, as median is the value which most frequent therefore we used median.
3. Before normalization the min. and max. value are dispersed over a big range but because of normalization they are now in a small range.

**b.** **Table 2 Mean and standard deviation before and after standardization**

| S. No. | Attribute | Before standardization | | After standardization | |
|---|---|---|---|---|---|
| | | Mean | Std. Deviation | Mean | Std. Deviation |
| 1 | pregs | 3.783 | 3.271 | 0 | 1.0 |
| 2 | plas | 121.656 | 30.438 | 0 | 1.0 |
| 3 | pres (in mm Hg) | 72.197 | 11.147 | 0 | 1.0 |
| 4 | skin (in mm) | 20.438 | 15.699 | 0 | 1.0 |
| 5 | test (in mu U/mL) | 60.919 | 77.636 | 0 | 1.0 |
| 6 | BMI (in kg/m$^2$) | 32.199 | 6.411 | 0 | 1.0 |
| 7 | pedi | 0.428 | 0.245 | 0 | 1.0 |
| 8 | Age (in years) | 32.76 | 11.055 | 0 | 1.0 |

**Inferences:**

1. After standardization the mean for all the attributes tends to zero and standard deviation tends towards 1

**a.**

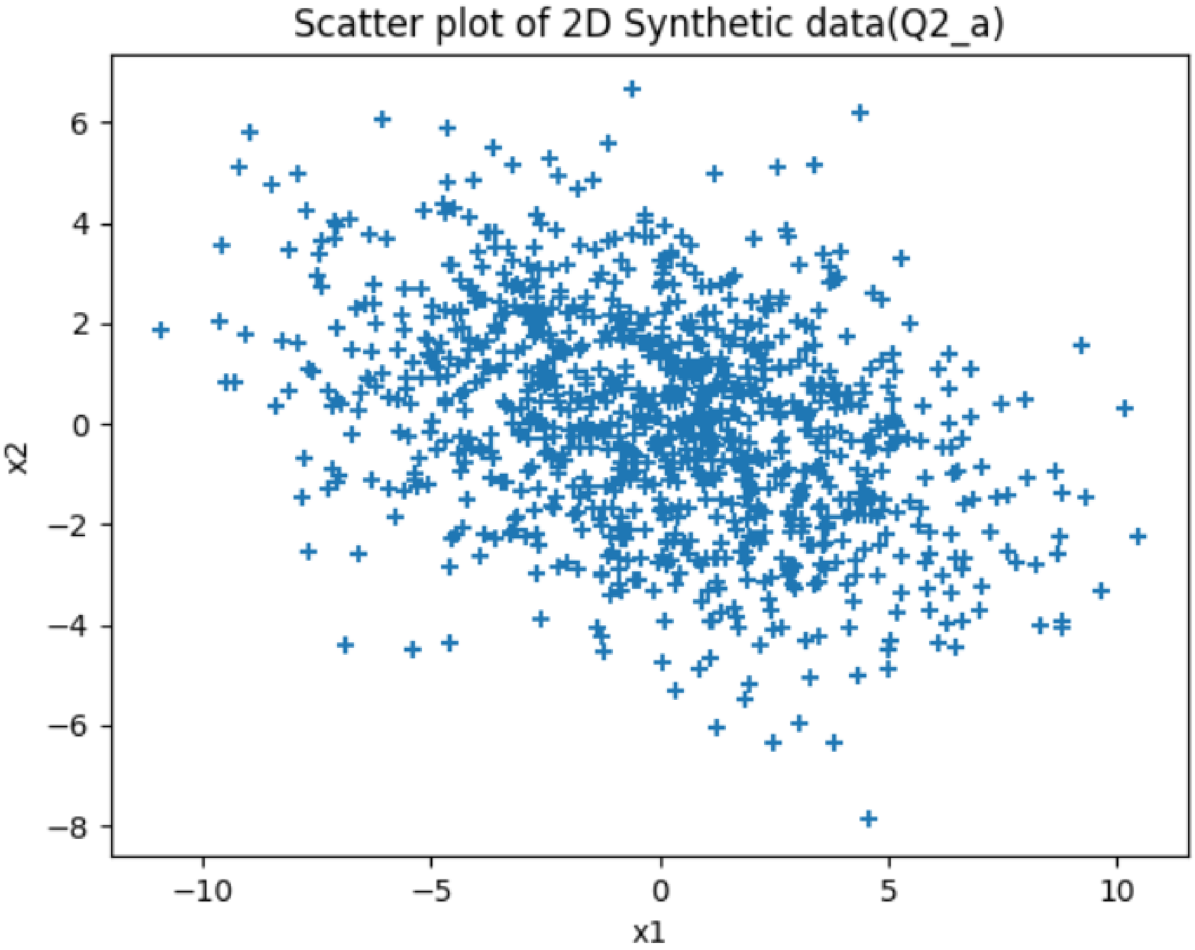

**Figure 1 Scatter plot of 2D synthetic data of 1000 samples**

**Inferences:**

1. Both the attributes are negatively correlated
2. Most of the values lies near the origin

**b.**

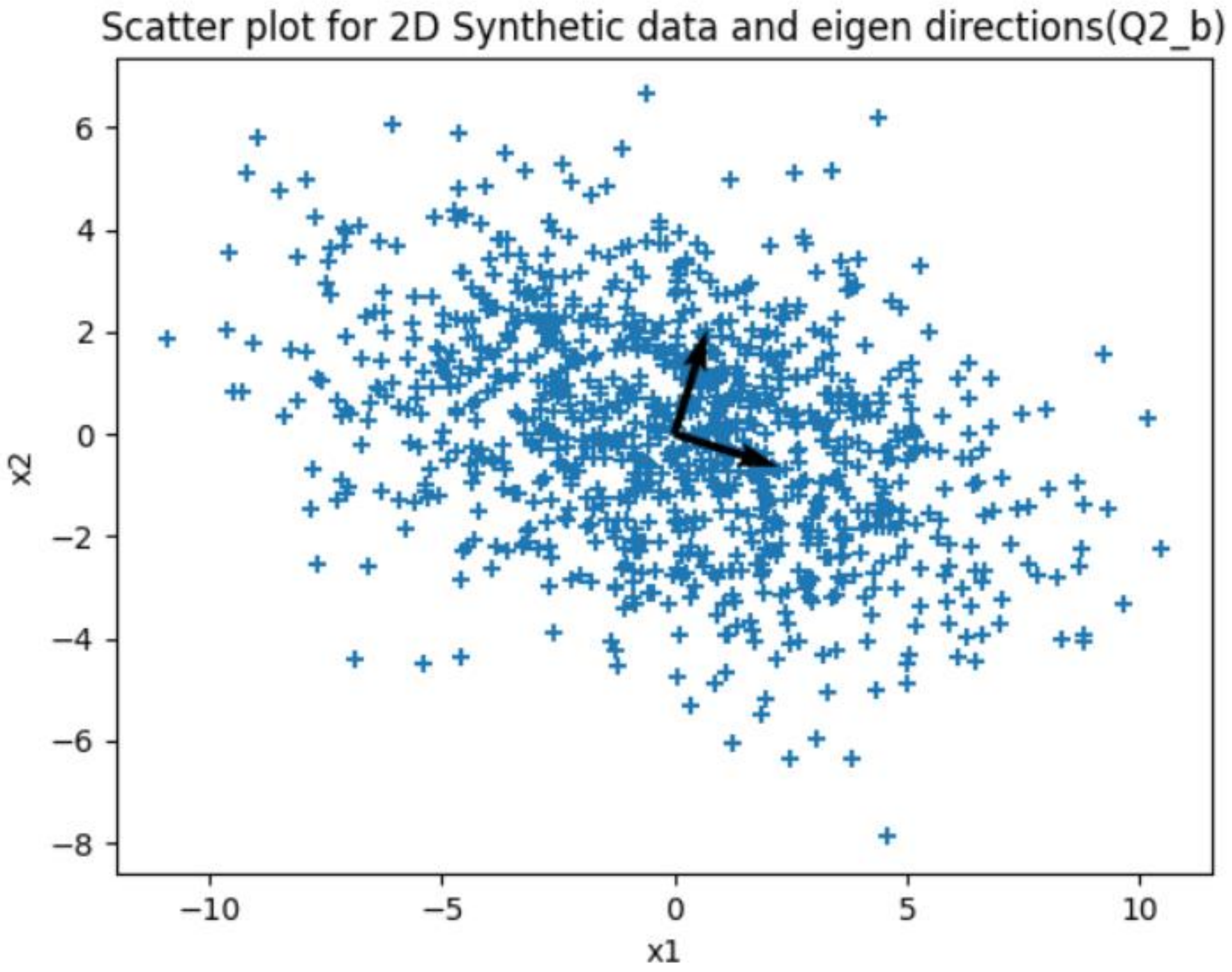

**Figure 2 Plot of 2D synthetic data and Eigen directions**

**Inferences:**

1. The spread starts from 4.111 and goes till 12.882
2. The density near the intersection of the vectors is high, and it decreases gradually as we go far from intersection.

**c.**

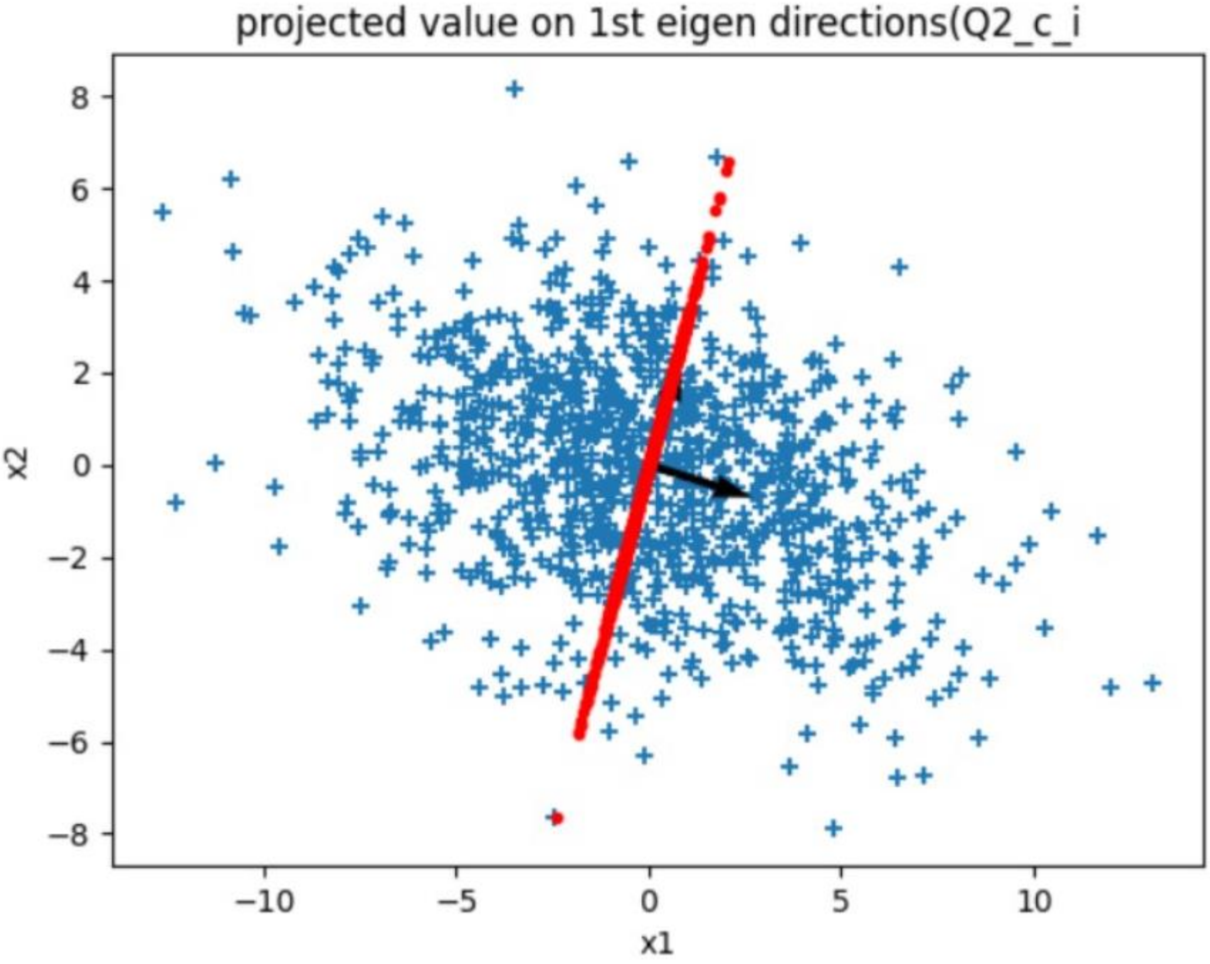

**Figure 3 Projected Eigen directions onto the scatter plot with 1st Eigen direction highlighted**

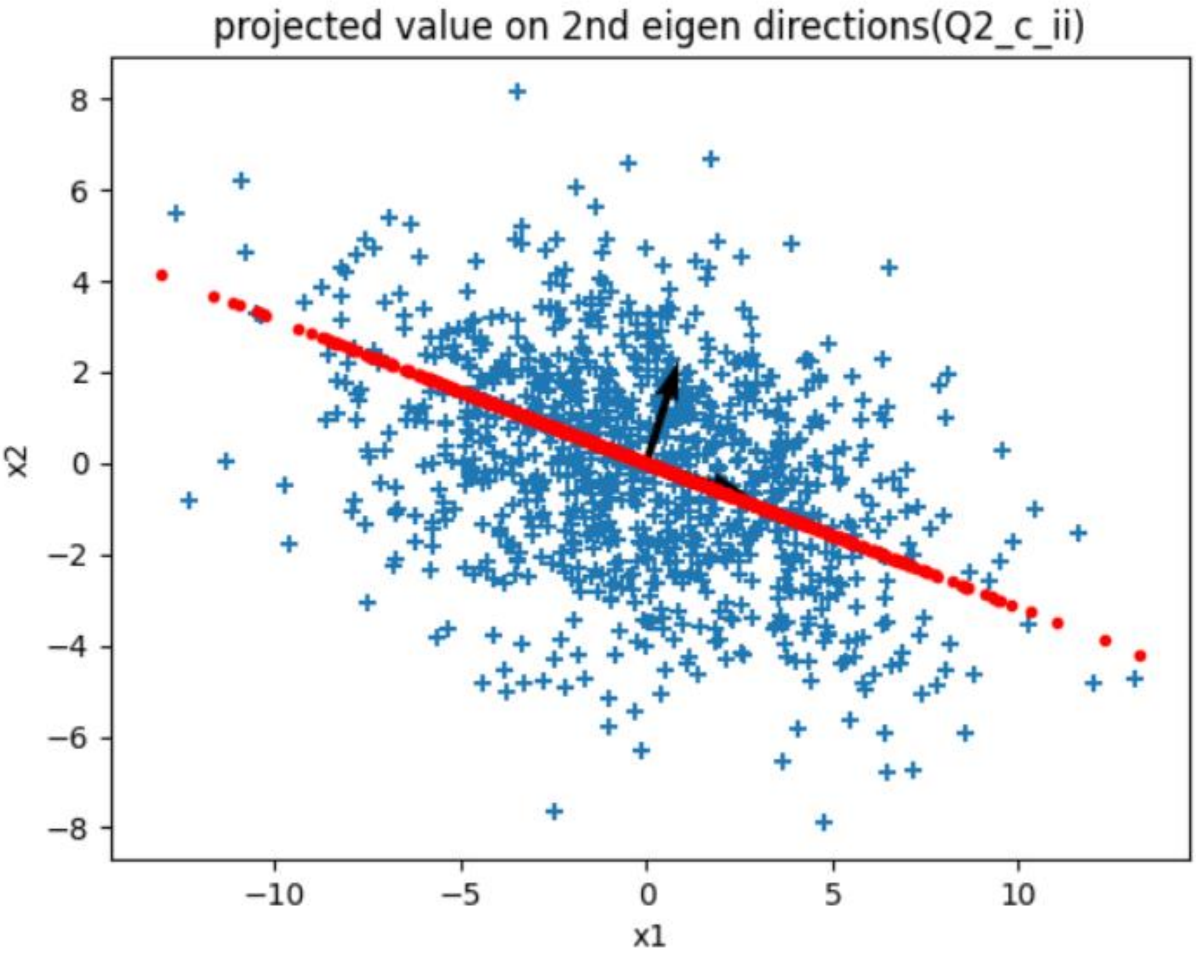

**Figure 4 Projected Eigen directions onto the scatter plot with 2nd Eigen direction highlighted**

**Inferences:**

1. From the projection on both of the vectors we can see that the spread is more for 1st eigen vector and less for 2nd eigen vector and the magnitude of the 1st eigen value is more than the 2nd, however both the eigen vectors have equal magnitude but different directions as both are orthogonal.
2. As we can observe from the projections of the data along the eigen vectors the spread is more for 1st eigen vector than 2nd eigen vector the magnitude of the 1st eigen value is more than 2nd means that data is more biased towards the first values or first eigen vector because the 1st eigen values is significantly higher than 2nd value so acc to dimensions first eigen is more useful to represent data and density of the data is more near the origin because mean of the data is 0.

**d.** Reconstruction error = 0

**Inferences:**

1. The lower the error the good the reconstruction.

**3**

**a.**

**Table 3 Variance and Eigenvalues of the projected data along the two directions**

| Direction | Variance | Eigenvalue |
|---|---|---|
| 1 | 1.992 | 1.992 |
| 2 | 1.853 | 1.853 |

**Inferences:**

1. Variance and eigenvalues are equal as they are calculated from covariance matrix they must be equal.

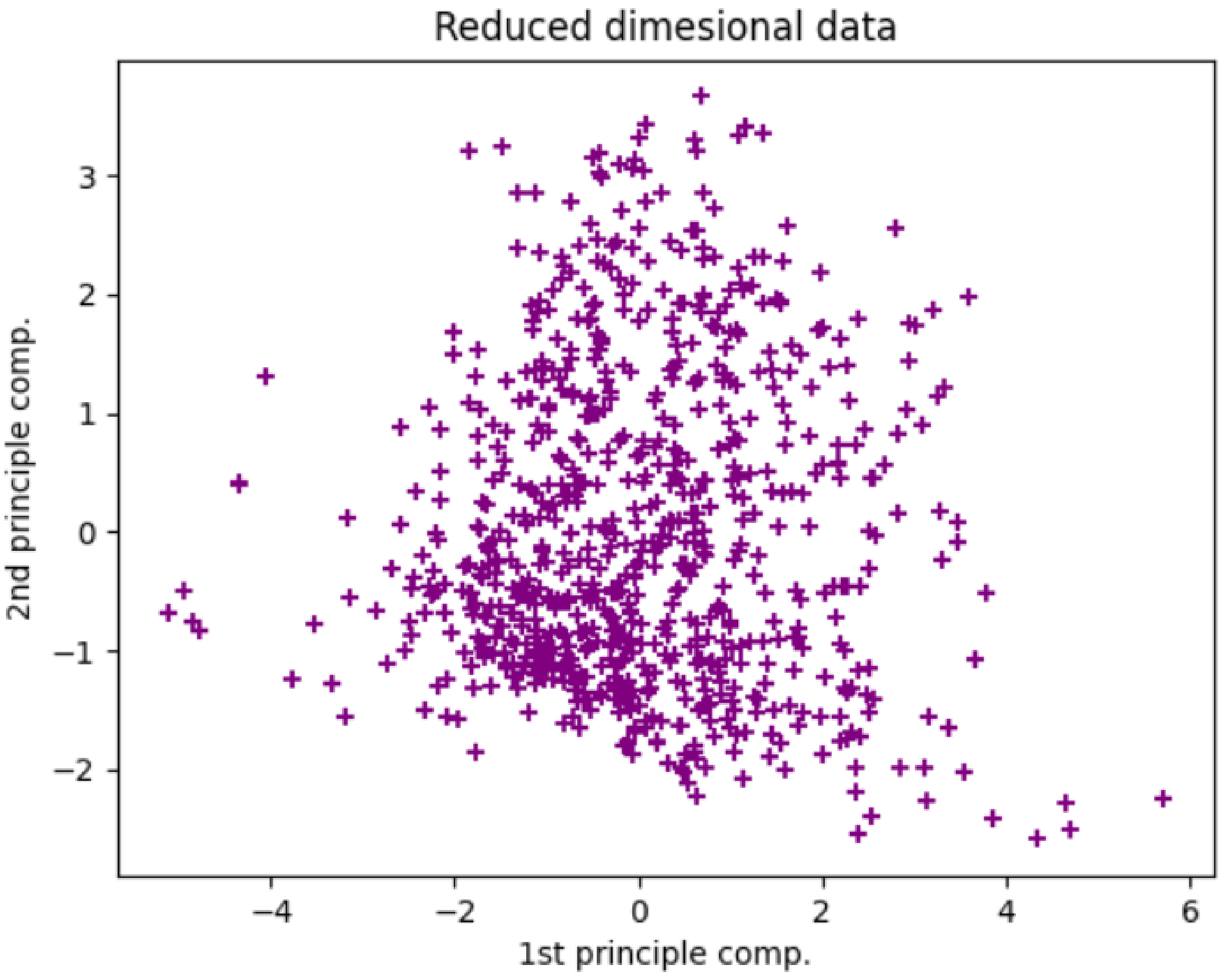

**Figure 5 Plot of data after dimensionality reduction**

**Inferences:**

1. The attributes are in slight positive correlation.

**b.**

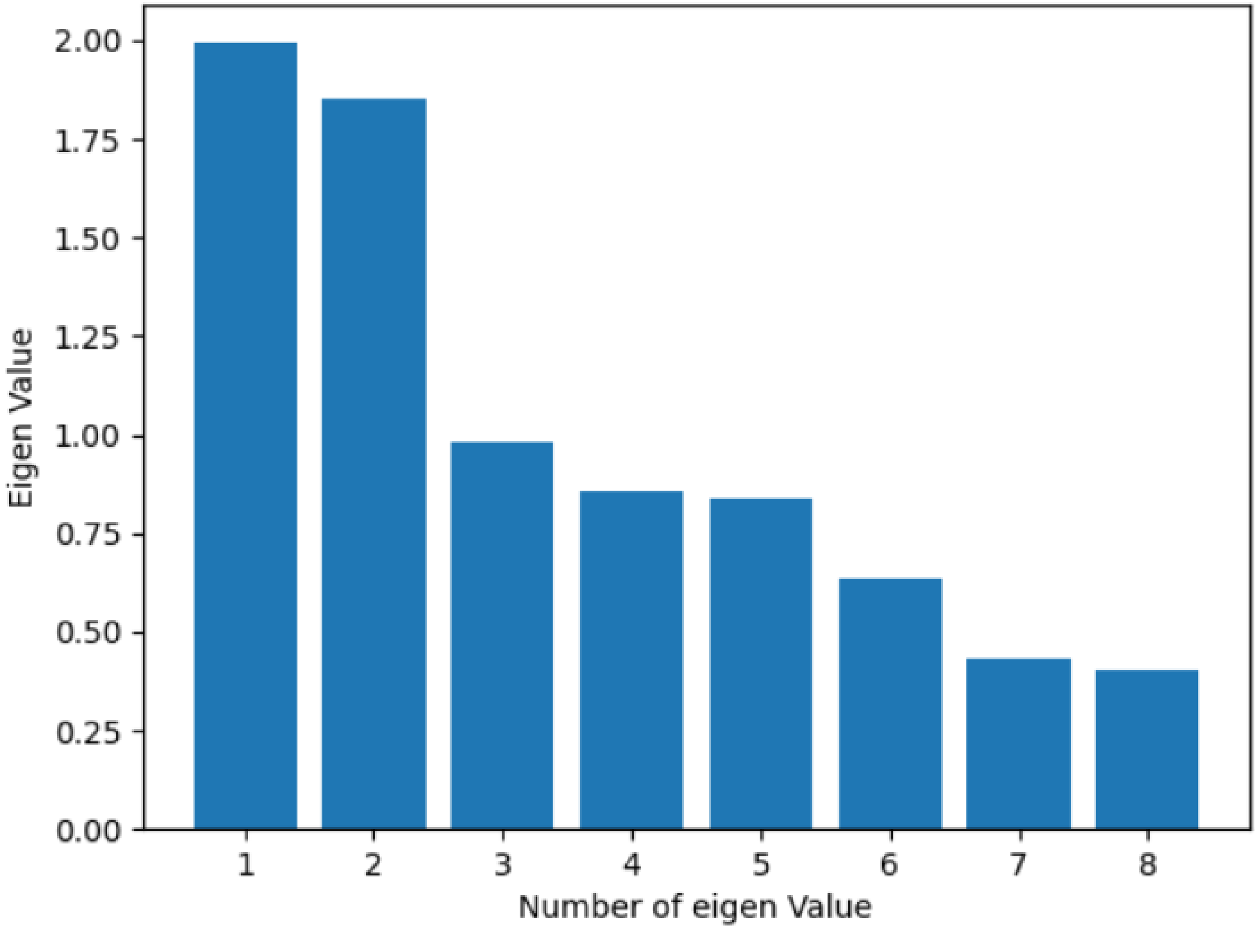

**Figure 6 Plot of Eigenvalues in descending order**

**Inferences:**

1. After the first two attributes they decrease rapidly and then gradually.
2. At eigen value equal to 1.875 it decreases substantially.

**c.**

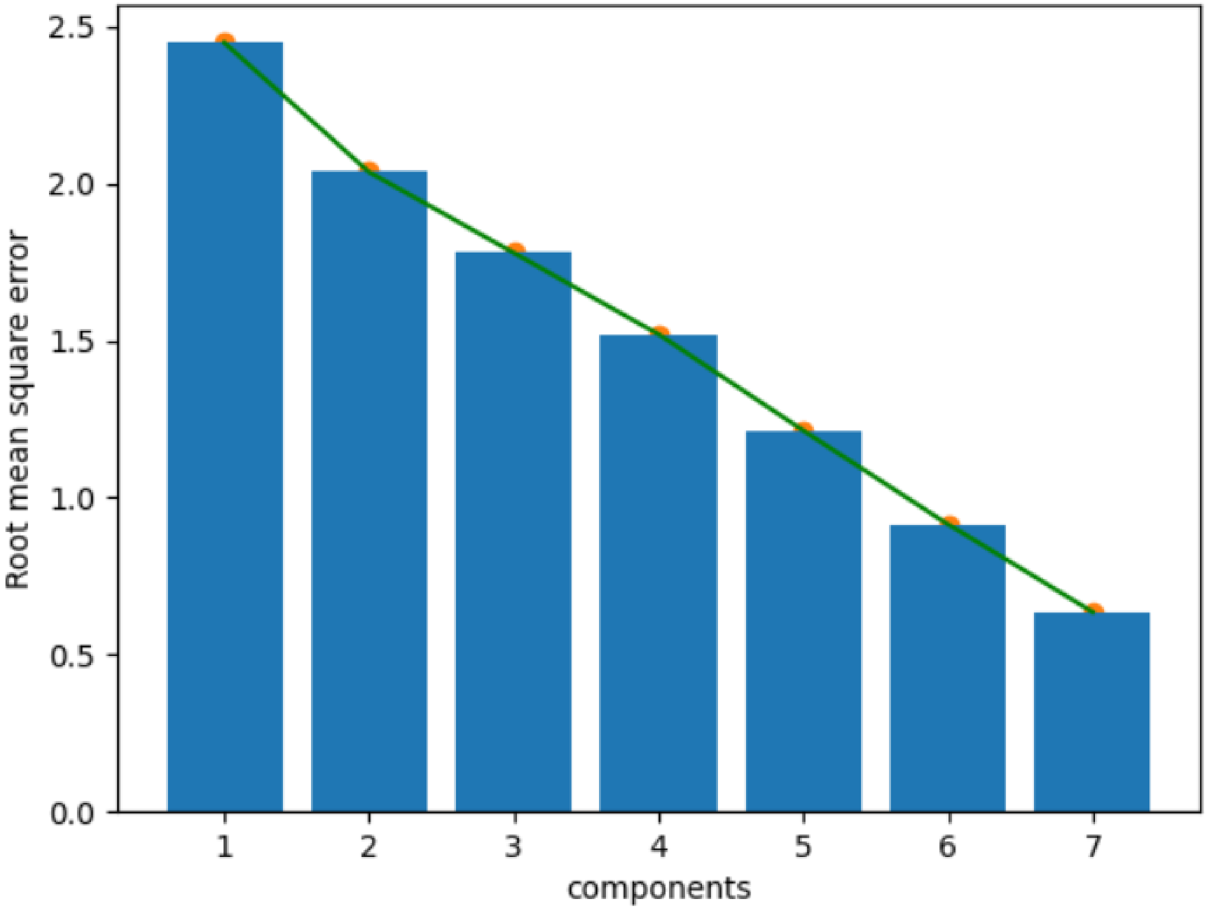

**Figure 7 Line plot to demonstrate reconstruction error vs. components**

**Inferences:**

1. The higher the magnitude of reconstruction the lower the Error is.

Table 4 Covariance matrix for dimensionally reduced data (l=2)

| | x1 | x2 |
|---|---|---|
| x1 | 1528.22 | 0 |
| x2 | 0 | 1421.57 |

Table 5 Covariance matrix for dimensionally reduced data (l=3)

| | x1 | x2 | x3 |
|---|---|---|---|
| x1 | 1528.22 | 0 | 0 |
| x2 | 0 | 1421.57 | 0 |
| x3 | 0 | 0 | 753.1 |

Table 6 Covariance matrix for dimensionally reduced data (l=4)

| | x1 | x2 | x3 | x4 |
|---|---|---|---|---|
| x1 | 1528.22 | 0 | 0 | 0 |
| x2 | 0 | 1421.57 | 0 | 0 |
| x3 | 0 | 0 | 753.1 | 0 |
| x4 | 0 | 0 | 0 | 658.32 |

Table 7 Covariance matrix for dimensionally reduced data (l=5)

| | x1 | x2 | x3 | x4 | x5 |
|---|---|---|---|---|---|
| x1 | 1528.22 | 0 | 0 | 0 | 0 |
| x2 | 0 | 1421.57 | 0 | 0 | 0 |
| x3 | 0 | 0 | 753.1 | 0 | 0 |
| x4 | 0 | 0 | 0 | 658.32 | 0 |
| x5 | 0 | 0 | 0 | 0 | 643.321 |

Table 8 Covariance matrix for dimensionally reduced data (l=6)

| | x1 | x2 | x3 | x4 | x5 | x6 |
|---|---|---|---|---|---|---|
| x1 | 1528.22 | 0 | 0 | 0 | 0 | 0 |
| x2 | 0 | 1421.57 | 0 | 0 | 0 | 0 |
| x3 | 0 | 0 | 753.1 | 0 | 0 | 0 |
| x4 | 0 | 0 | 0 | 658.32 | 0 | 0 |
| x5 | 0 | 0 | 0 | 0 | 643.321 | 0 |
| x6 | 0 | 0 | 0 | 0 | 0 | 488.125 |

Table 9 Covariance matrix for dimensionally reduced data (l=7)

| | x1 | x2 | x3 | x4 | x5 | x6 | x7 |
|---|---|---|---|---|---|---|---|
| x1 | 1528.22 | 0 | 0 | 0 | 0 | 0 | 0 |
| x2 | 0 | 1421.57 | 0 | 0 | 0 | 0 | 0 |
| x3 | 0 | 0 | 753.1 | 0 | 0 | 0 | 0 |
| x4 | 0 | 0 | 0 | 658.32 | 0 | 0 | 0 |
| x5 | 0 | 0 | 0 | 0 | 643.321 | 0 | 0 |
| x6 | 0 | 0 | 0 | 0 | 0 | 488.125 | 0 |
| x7 | 0 | 0 | 0 | 0 | 0 | 0 | 332.988 |

Table 10 Covariance matrix for dimensionally reduced data (l=8)

| | x1 | x2 | x3 | x4 | x5 | x6 | x7 | x8 |
|---|---|---|---|---|---|---|---|---|
| x1 | 1528.22 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| x2 | 0 | 1421.57 | 0 | 0 | 0 | 0 | 0 | 0 |
| x3 | 0 | 0 | 753.1 | 0 | 0 | 0 | 0 | 0 |
| x4 | 0 | 0 | 0 | 658.32 | 0 | 0 | 0 | 0 |
| x5 | 0 | 0 | 0 | 0 | 643.321 | 0 | 0 | 0 |
| x6 | 0 | 0 | 0 | 0 | 0 | 488.125 | 0 | 0 |
| x7 | 0 | 0 | 0 | 0 | 0 | 0 | 332.988 | 0 |
| x8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 310.349 |

**Inferences:**

1. Off diagonal values are all tending to zero which on rounding off becomes zero as the covariance matrix is symmetric therefore they are zero.
2. The diagonal values are significant while the off diagonal are tending to zero this is because covariance matrix is symmetric
3. As we go downwards in the diagonal the value decreases.
4. As we go lower the eigenvalues decreases resulting in decrease of corresponding projection value decreases and these results in decrease in value of variance.
5. The first eigen vector seems to be more reliable however 2nd is also good.
6. From the value of diagonal elements, the number of components that will give the optimum reconstruction with dimension reduction and eigen vectors.
7. The value of 1st diagonal element is equal as the eigen vectors are same.
8. The 2nd value is also equal as it is independent of other vectors.
9. All are equal

**d.**

Table 11 Covariance matrix for original data

| | pregs | plas | pres | skin | test | BMI | pedi | Age |
|---|---|---|---|---|---|---|---|---|
| pregs | 1 | 0.120 | 0.124 | -0.095 | -0.107 | -0.037 | -0.030 | 0.532 |
| plas | 0.120 | 1 | 0.125 | 0.040 | 0.169 | 0.186 | 0.112 | 0.225 |
| pres (in mm Hg) | 0.124 | 0.125 | 1 | 0.195 | 0.095 | 0.240 | 0.050 | 0.210 |
| skin (in mm) | -0.095 | 0.040 | 0.195 | 1 | 0.458 | 0.328 | 0.132 | -0.078 |
| test (in mu U/mL) | -0.107 | 0.169 | 0.095 | 0.458 | 1 | 0.164 | 0.106 | -0.078 |
| BMI (in kg/m$^2$) | -0.037 | 0.186 | 0.240 | 0.328 | 0.164 | 1 | 0.042 | 0.127 |
| pedi | -0.030 | 0.112 | 0.050 | 0.132 | 0.106 | 0.042 | 1 | 0.028 |
| Age (in years) | 0.532 | 0.225 | 0.210 | -0.078 | -0.055 | 0.127 | 0.028 | 1 |

**Inferences:**

1. The off-diagonal values have a similar trend as compared with the covariance matrix obtained above after PCA l=8 reduction but in case of PCA l=8 values were 0.
2. The magnitudes of diagonal values are 1 as standard deviation is 1 and correlation is also 1 so covariance is also 1 but the magnitudes above were decreasing as we moved down.
3. For the real data there is no trend of increasing or decreasing values while for the reduced after standardization a decreasing trend is there.