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

**Student's Name: Kuldeep Jain Dugar** **Branch:**

**Roll Number: b20112** **CSE**

**Mobile No: 8986388665**

**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 | 17 | 5 | 12 |
| 2 | plas | 0 | 199 | 5 | 12 |
| 3 | pres (in mm Hg) | 0 | 122 | 5 | 12 |
| 4 | skin (in mm) | 0 | 99 | 5 | 12 |
| 5 | test (in mu U/mL) | 0 | 846 | 5 | 12 |
| 6 | BMI (in kg/$m^2$) | 0 | 67,1 | 5 | 12 |
| 7 | pedi | 0.078 | 2.42 | 5 | 12 |
| 8 | Age (in years) | 21 | 81 | 5 | 12 |

**Inferences:**

1. The need for outlier correction is to bring uniformity in data for more appropriate results .
2. Normalization method makes the data uniform by equalizing the max n min .
3. Max and min becomes uniform for all attributes.

**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 | 3.78 | 3.27 | 0 | 1 |
| 2 | plas | 121.66 | 30.43 | 0 | 1 |
| 3 | pres (in mm Hg) | 72.2 | 11.14 | 0 | 1 |

| | | | | | |
|---|---|---|---|---|---|
| 4 | skin (in mm) | 20.43 | 15.69 | 0 | 1 |
| 5 | test (in mu U/mL) | 60.92 | 77 | 0 | 1 |
| 6 | BMI (in kg/m$^2$) | 32 | 6.4 | 0 | 1 |
| 7 | pedi | 0.427 | 0.24 | 0 | 1 |
| 8 | Age (in years) | 32.76 | 11.05 | 0 | 1 |

**Inferences:**

1. All attributes have same mean and std. dev.

**2 a.**

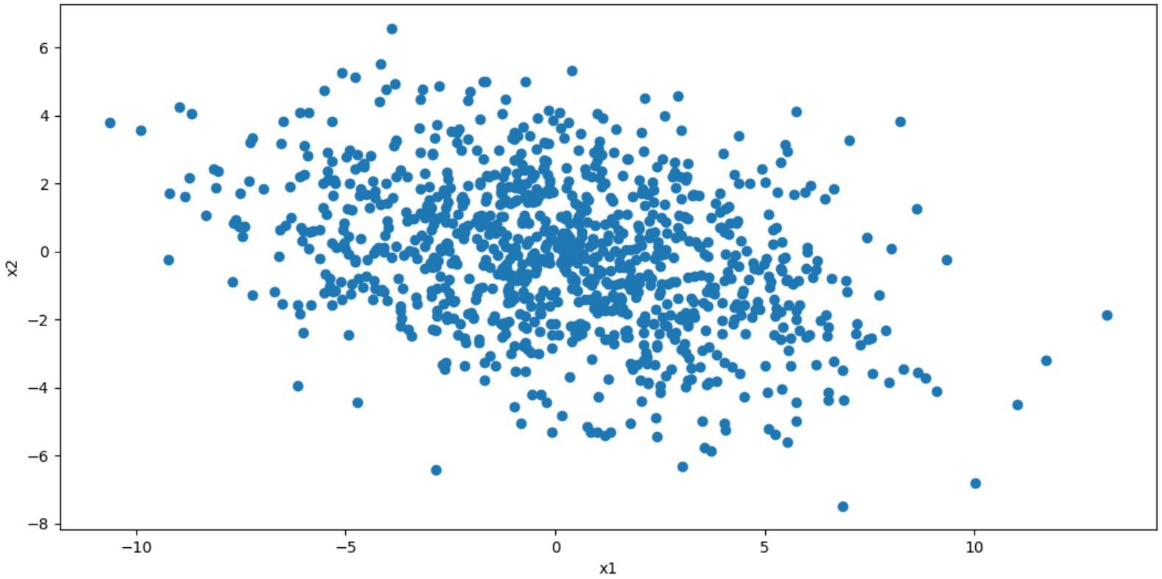

**Figure 1 Scatter plot of 2D synthetic data of 1000 samples**

**Inferences**

Correlation - negative

Density of points higher in the centre.

**b.**

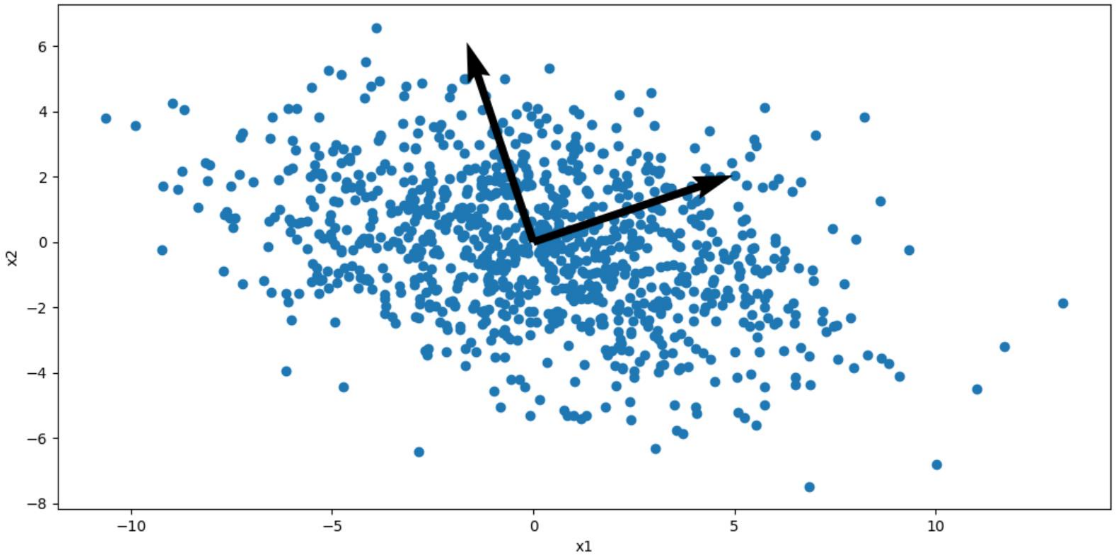

**Figure 2 Plot of 2D synthetic data and Eigen directions**

**Inferences:**

1. Spread of data based upon the magnitude of Eigenvalues- More eigen value more spread
2. Density of points near the intersection of Eigen axes is higher and gradually fades as we go away from it.

**c.**

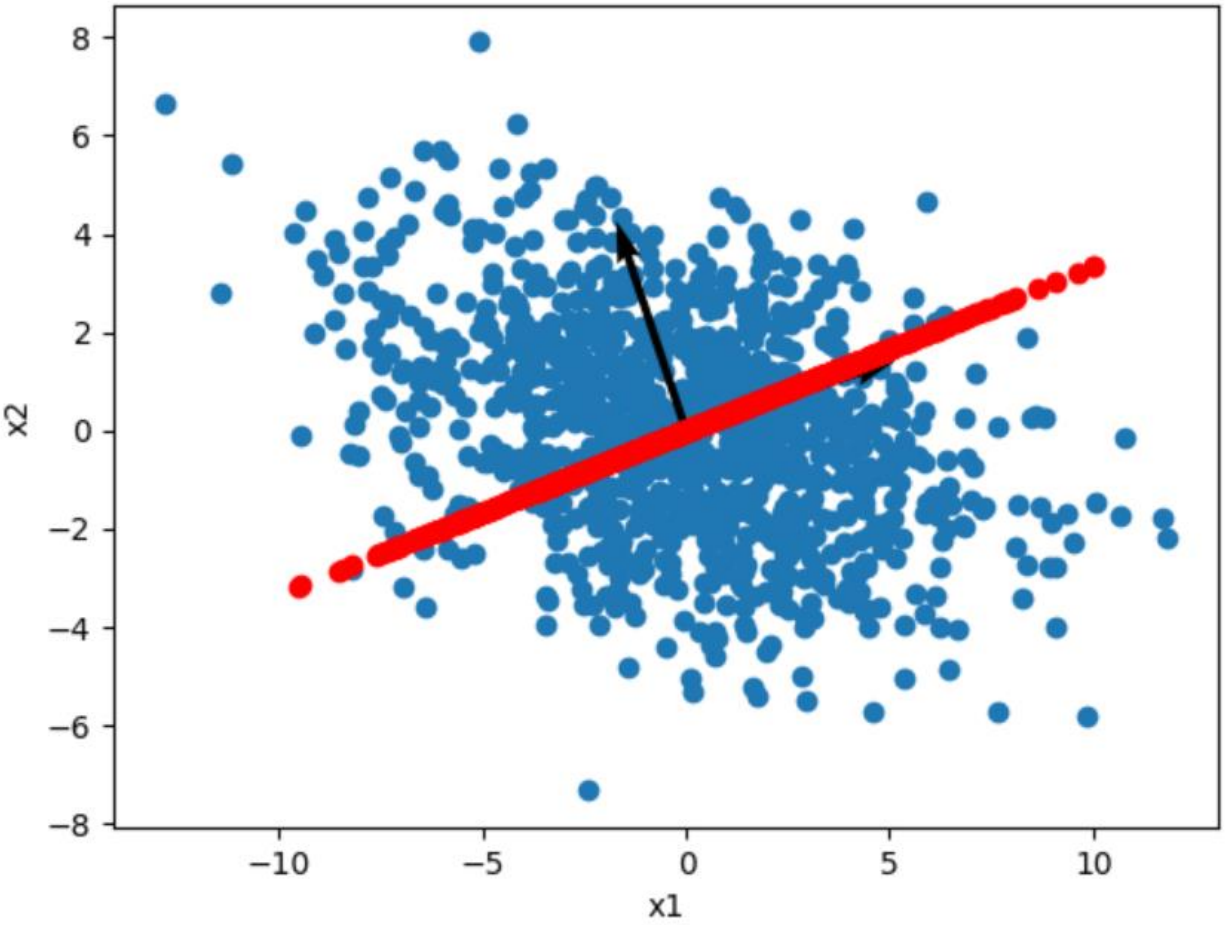

**Figure 3 Projected Eigen directions onto the scatter plot with 1st Eigen direction highlighted**

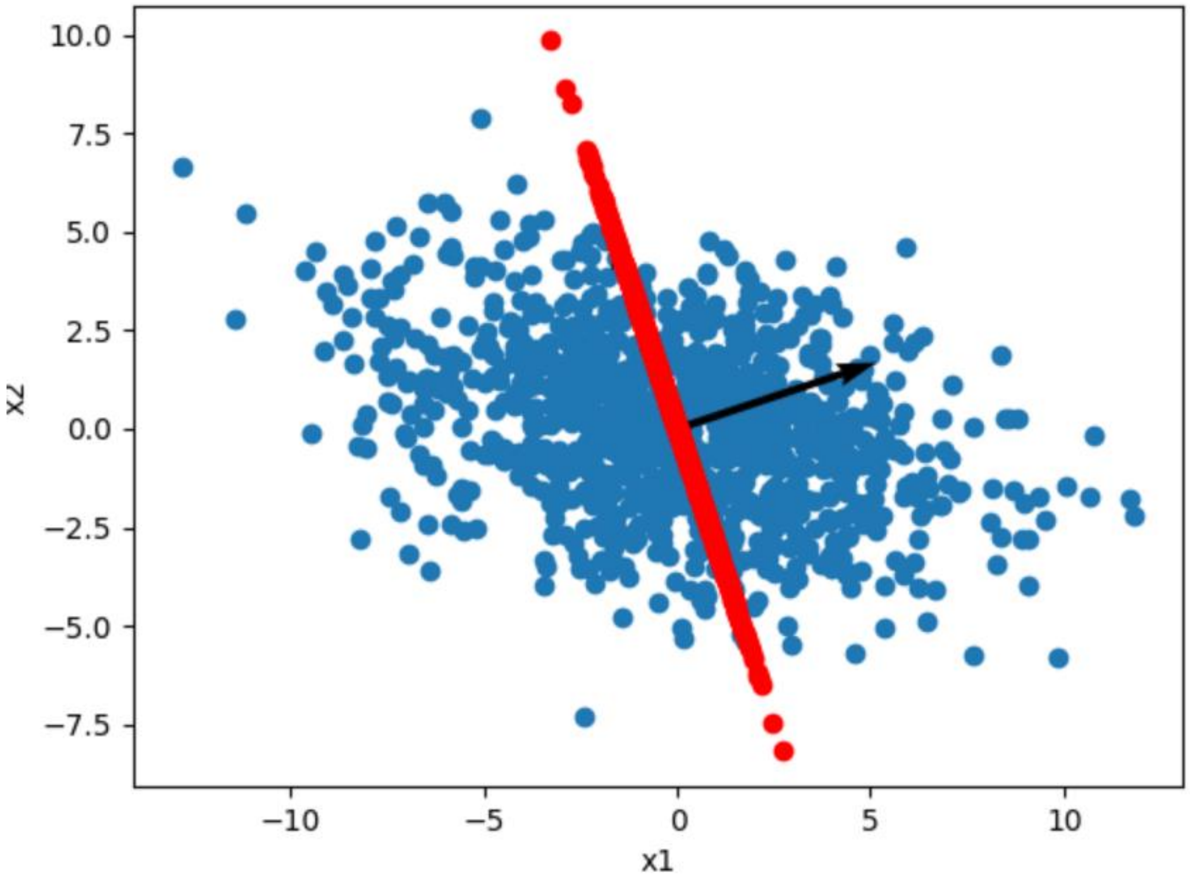

**Figure 4 Projected Eigen directions onto the scatter plot with 2nd Eigen direction highlighted**

**Inferences:**

1. The first eigenvalue is 14 ,and the other one is 4.0 ,therefore 1st eigen direction is according to the first eigenvalue and the 2nd eigen direction is according to the other eigenvalue(4.0).

2. the first eigen direction, there is more density of points ,as it is in the direction of the first eigenvalue which is greater then the 2nd eigenvalue ,therefore the other eigen direction has less density of points.

.

**d.** Reconstruction error -,

**Inferences:**

1. Infer how the magnitude of reconstruction error affects the quality of reconstruction.
2. Inference 2(You may add or delete the number of inferences)

**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. Compare variance of the projected data along the two directions with the Eigenvalues of the two directions of projection
2. Inference 2(You may add or delete the number of inferences)

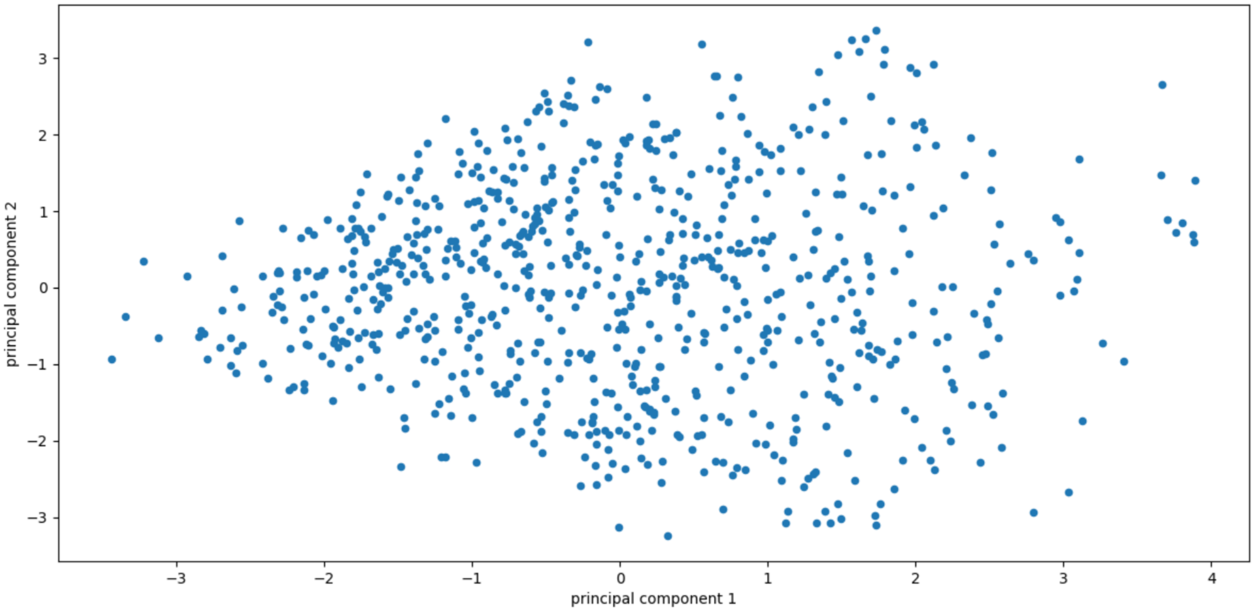

**Figure 5 Plot of data after dimensionality reduction**

**Inferences:**

The two attributes obtained after dimensionality reduction from the spread of data points are negatively correlated with very low magnitude.

**b.**

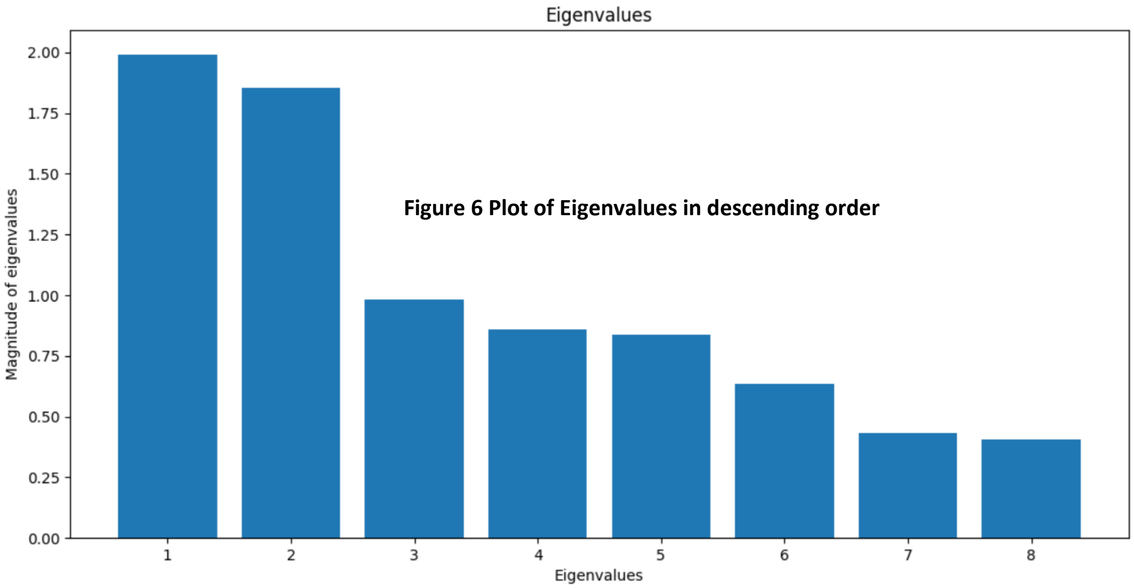

**Figure 6 Plot of Eigenvalues in descending order**

**Inferences:**

1. The subsequent Eigenvalues firstly decrease rapidly,then decreases gradually.
2. The eigenvalue with magnitude 0.981 ,then the eigenvalue decreases substantially.

**c.**

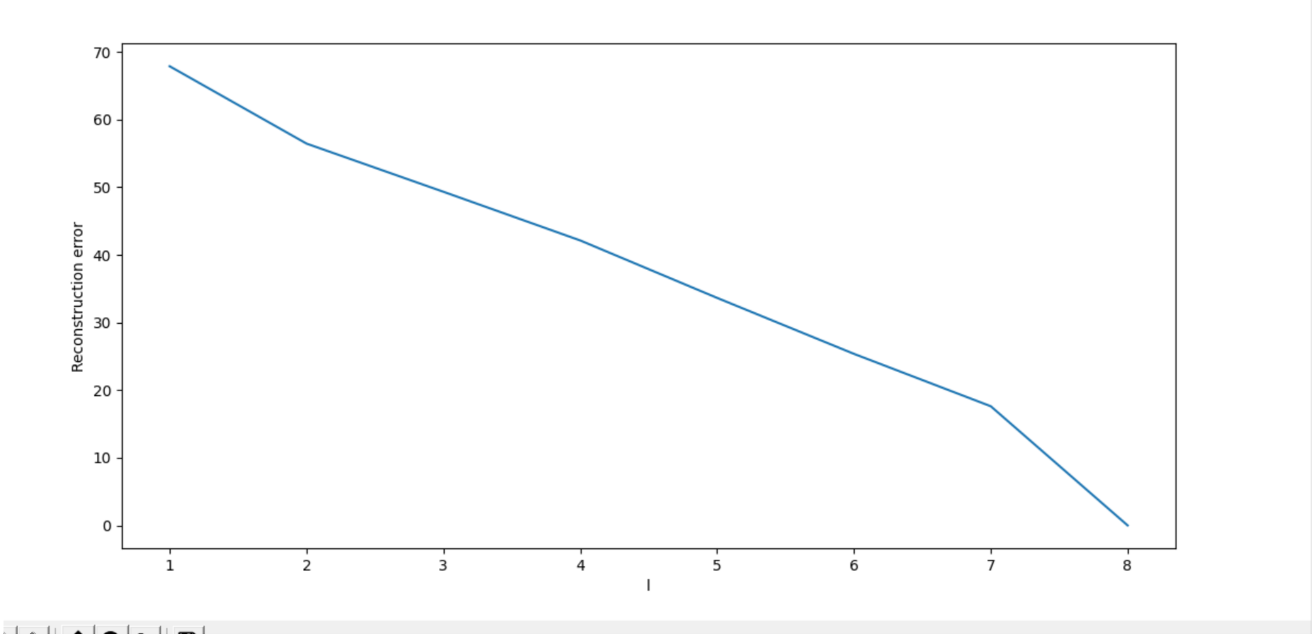

**Figure 7 Line plot to demonstrate reconstruction error vs. components**

**Inferences:**

If the magnitude of reconstruction error is large then, it will affects the quality of reconstruction of original data back and it will also affect the prediction of any attribute , therefore it will be better if the reconstruction error is small.

.

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

| | x1 | x2 |
|---|---|---|
| x1 | 1.992 | 0 |
| x2 | 0 | 1.853 |

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

| | x1 | x2 | x3 |
|---|---|---|---|
| x1 | 1.992 | 0 | 0 |
| x2 | 0 | 1.853 | 0 |
| x3 | 0 | 0 | 0.982 |

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

| | x1 | x2 | x3 | x4 |
|---|---|---|---|---|
| x1 | 1.992 | 0 | 0 | 0 |
| x2 | 0 | 1.853 | 0 | 0 |
| x3 | 0 | 0 | 0.982 | 0 |
| x4 | 0 | 0 | 0 | 0.858 |

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

| | x1 | x2 | x3 | x4 | x5 |
|---|---|---|---|---|---|
| x1 | 1.992 | 0 | 0 | 0 | 0 |
| x2 | 0 | 1.853 | 0 | 0 | 0 |
| x3 | 0 | 0 | 0.982 | 0 | 0 |
| x4 | 0 | 0 | 0 | 0.858 | 0 |
| x5 | 0 | 0 | 0 | 0 | 0.839 |

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

| | x1 | x2 | x3 | x4 | x5 | x6 |
|---|---|---|---|---|---|---|
| x1 | 1.992 | 0 | 0 | 0 | 0 | 0 |
| x2 | 0 | 1.853 | 0 | 0 | 0 | 0 |
| x3 | 0 | 0 | 0.982 | 0 | 0 | 0 |
| x4 | 0 | 0 | 0 | 0.858 | 0 | 0 |
| x5 | 0 | 0 | 0 | 0 | 0.839 | 0 |
| x6 | 0 | 0 | 0 | 0 | 0 | 0.636 |

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

| | x1 | x2 | x3 | x4 | x5 | x6 | x7 |
|---|---|---|---|---|---|---|---|
| x1 | 1.992 | 0 | 0 | 0 | 0 | 0 | 0 |
| x2 | 0 | 1.853 | 0 | 0 | 0 | 0 | 0 |
| x3 | 0 | 0 | 0.982 | 0 | 0 | 0 | 0 |
| x4 | 0 | 0 | 0 | 0.858 | 0 | 0 | 0 |
| x5 | 0 | 0 | 0 | 0 | 0.839 | 0 | 0 |
| x6 | 0 | 0 | 0 | 0 | 0 | 0.636 | 0 |
| x7 | 0 | 0 | 0 | 0 | 0 | 0 | 0.434 |

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

| | x1 | x2 | x3 | x4 | x5 | x6 | x7 | x8 |
|---|---|---|---|---|---|---|---|---|
| x1 | 1.992 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| x2 | 0 | 1.853 | 0 | 0 | 0 | 0 | 0 | |
| x3 | 0 | 0 | 0.982 | 0 | 0 | 0 | 0 | 0 |
| x4 | 0 | 0 | 0 | 0.858 | 0 | 0 | 0 | 0 |
| x5 | 0 | 0 | 0 | 0 | 0.839 | 0 | 0 | 00 |
| x6 | 0 | 0 | 0 | 0 | 0 | 0.636 | 0 | 0 |
| x7 | 0 | 0 | 0 | 0 | 0 | 0 | 0.434 | 0 |
| x8 | 0 | 0 | 00 | 0 | 0 | 0 | 0 | 0.405 |

**Inferences:**

1. . The off-daigonal elements values are zero because data is transformed such that these attributes are no more correlated with each other ,therefore their variance value is zero.
2. The diagonal values are non-zero because diagonal elements are the values of the eigenvalues which represents the variance of the attributes, therefore they are non-zero.
3. The trend of diagonal values are according to variance of the attributes with each other.
4. The decrease in values of diagonal values is gradually because variance(eigenvalues) decrease.
5. From the magnitude of diagonal elements, 1.992 component captures data variations the best.
6. From the value of diagonal elements, the number of components =**2** should give the optimum reconstruction along with dimensionality reduction.
7. the magnitude of the 1st diagonal element (topmost left corner) in each of the obtained covariance matrices is same because it represents the variations of the variables,therefore it will not get change.
8. The magnitude of the 2nd diagonal element in each of the obtained covariance matrices is same in magnitude because it represents the variations(2nd eigen value),therefore it will not get change.
9. Compare 3rd, 4th, 5th, 6th, and 7th diagonal elements across covariance matrices are same in magnitude because they represent the variation of the attributes,therefore they will not get change

**d.**

Table 11 Covariance matrix for original data

| | pregs | plas | pres | skin | test | BMI | pedi | Age |
|---|---|---|---|---|---|---|---|---|
| **pregs** | **1.000** | **0.118** | **0.209** | **-0.097** | **-0.108** | **0.028** | **0.005** | **0.561** |
| **plas** | **0.118** | **1.000** | **0.205** | **0.060** | **0.180** | **0.228** | **0.082** | **0.274** |
| **pres** | **0.209** | **0.205** | **1.000** | **0.026** | **-0.051** | **0.272** | **0.022** | **0.326** |
| **skin** | **-0.097** | **0.060** | **0.026** | **1.000** | **0.473** | **0.374** | **0.153** | **-0.101** |
| **test** | **-0.108** | **0.180** | **-0.051** | **0.473** | **1.000** | **0.172** | **0.199** | **-0.074** |
| **BMI** | **0.028** | **0.228** | **0.272** | **0.374** | **0.172** | **1.000** | **0.124** | **0.078** |
| **pedi** | **0.005** | **0.082** | **0.022** | **0.153** | **0.199** | **0.124** | **1.000** | **0.036** |
| **Age** | **0.561** | **0.274** | **0.326** | **-0.101** | **-0.074** | **0.078** | **0.036** | **1.000** |

**Inferences:**

1. The off-diagonal values represent the variances of one attribute with the other attribute therefore it is zero while in the covariance matrix obtained after PCA l=8 reduction the off-diagonal values are zero because they are no more correlated with each other.

2. In this covariance matrix the diagonal elements represent the variance of the attribute with itself, while in the covariance matrix obtained after PCA l=8 reduction the diagonal elements represent the variance of the whole data into 8 different components.
3. Is there any trade of increase or decrease in diagonal elements like/ unlike covariance obtained after dimensionality reduction? – no because it will not get change in this data

**Guidelines for Report (Delete this while you submit the report):**

- **The plot/graph/figure/table should be centre justified with sequence number and caption.**
- **Inferences should be written as a numbered list.**
- **Use specific and technical terms to write inferences.**
- **Values observed/calculated should be rounded off to three decimal places.**
- **The quantities which have units should be written with units.**