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

**Student's Name:** Vishal Sharma

**Mobile No:** 9540140310

**Roll Number:** B20239

**Branch:** Data Science and Engineering

**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 | 0.07 | 2.42 | 5 | 12 |
| 8 | Age (in years) | 21 | 81 | 5 | 12 |

**Inferences:**

1. Outliers correction is necessary because they can make many of our assumption go wrong, thus the data is not much reliable.
2. Replacing the outliers with median is preferred because the change in data might be huge if it is replaced with mean
3. Earlier, some attributes have a dominating effect on other attributes (like test), which will give wrong analysis. But after normalization, the data has a definite range and thus, there is no such domination of any attribute. Thus, the analysis will be more accurate.

**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 | 3.845 | 3.369 | 0 | 1 |
| 2 | plas | 120.894 | 31.972 | 0 | 1 |
| 3 | pres (in mm Hg) | 69.105 | 19.355 | 0 | 1 |
| 4 | skin (in mm) | 20.536 | 15.952 | 0 | 1 |
| 5 | test (in mu U/mL) | 79.8 | 115.244 | 0 | 1 |
| 6 | BMI (in $kg/m^2$) | 32 | 7.884 | 0 | 1 |
| 7 | pedi | 0.471 | 0.331 | 0 | 1 |
| 8 | Age (in years) | 32.24 | 11.76 | 0 | 1 |

**Inferences:**

1. Earlier, some attributes have a dominating effect on other attributes (like test), which will give wrong analysis. But after normalization, the data has a definite range and thus, there is no such domination of any attribute. Thus, the analysis will be more accurate.

**2 a.**

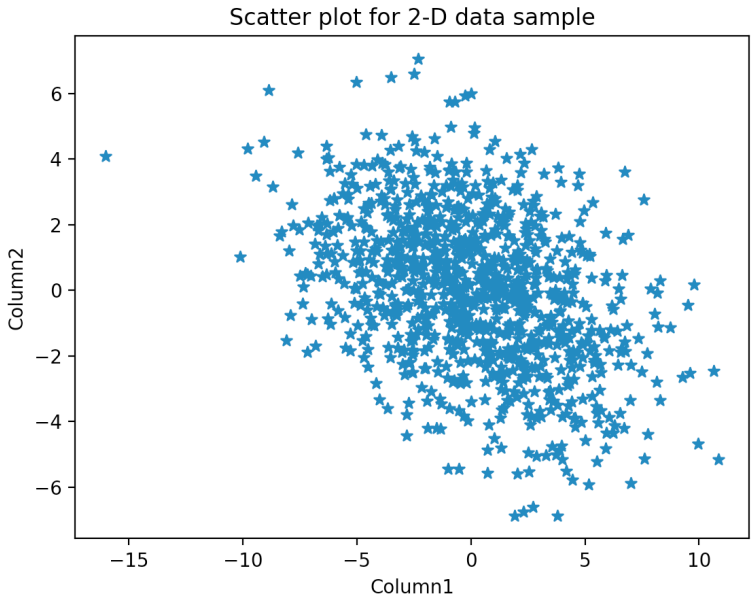


**Figure 1 Scatter plot of 2D synthetic data of 1000 samples**

**Inferences:**

1. From the above plot, we can see that there is a negative slope between the two attributes. Thus 'Column2' and 'Column2' are negatively correlated.
2. From the density of scatter plot, the distribution seems to be symmetric. Thus the mean can be inferred as 0.

**b.**

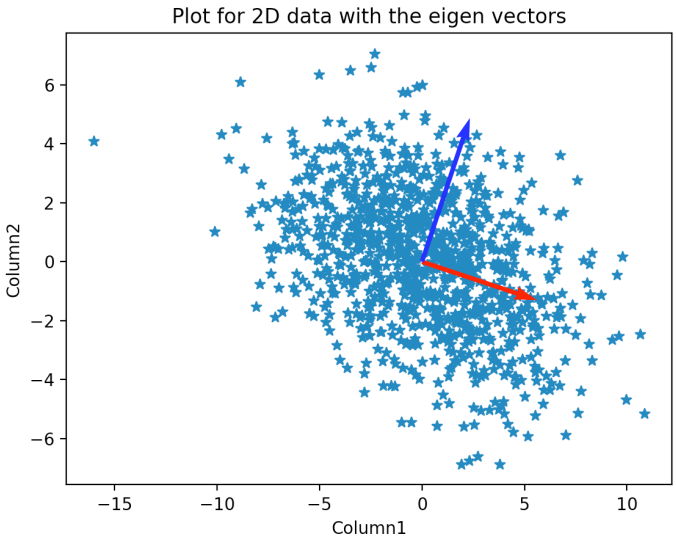


**Figure 2 Plot of 2D synthetic data and Eigen directions**

**Inferences:**

1. The spread for eigenvalue 4 is not much compared with 14.
2. The density is quite high on the intersection of these eigen vectors but decreases as moved further from the intersection point.

**c.**

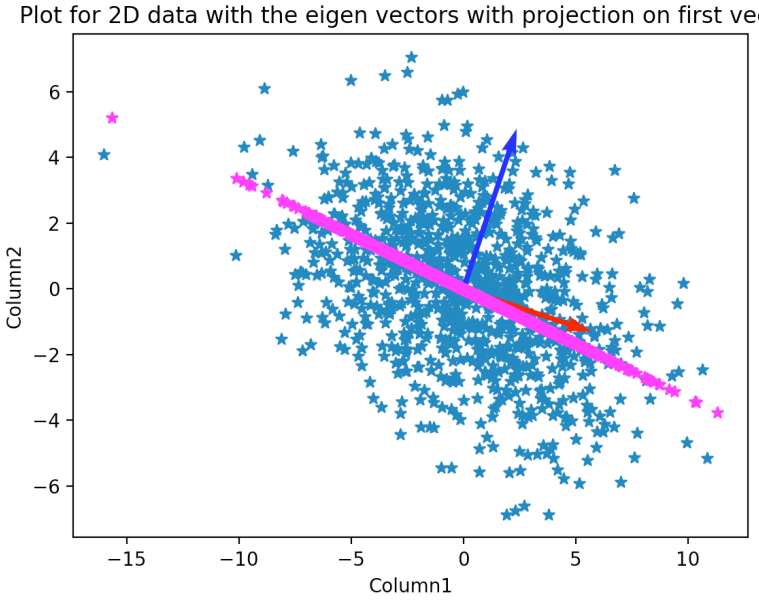


**Figure 3 Projected Eigen directions onto the scatter plot with 1st Eigen direction highlighted**

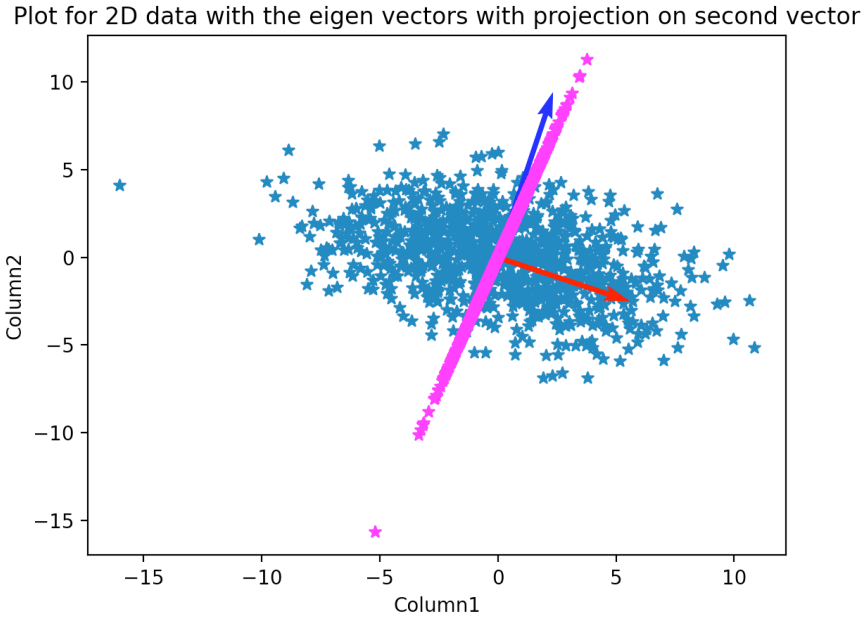


**Figure 4 Projected Eigen directions onto the scatter plot with 2nd Eigen direction highlighted**

**Inferences:**

1. The eigenvalue 14 has more spread along its direction than the eigenvalue 4.
2. As for the density, for the first eigen vector, the eigenvalue is small, and thus the variance or the spread is not much in this direction. But for the second eigenvector, the spread or the variance is more.
3. Inference 3(You may add or delete the number of inferences)
   Note: The scatter plots above are for illustration purposes. Replace it with the scatter plot obtained by you. Rename x-axis legend with x1 and y-axis legend with x2.

**d.** Reconstruction error = 0

**Inferences:**

1. Greater the reconstruction error, greater is the loss in nature of data. Here the reconstruction is very low, approximately zero. Thus, it can inferred that there is approximately no loss in nature, reason being that the dimensions even after reduction remained the same.

**3 a.**

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

| Direction | Variance | Eigenvalue |
|---|---|---|
| 1 | 2.094 | 2.094 |
| 2 | 1.731 | 1.731 |

**Inferences:**

1. Higher the value of eigen vector, more is the variance. Thus, more will be the spread. Thus, the data is spread more towards first eigen vector.

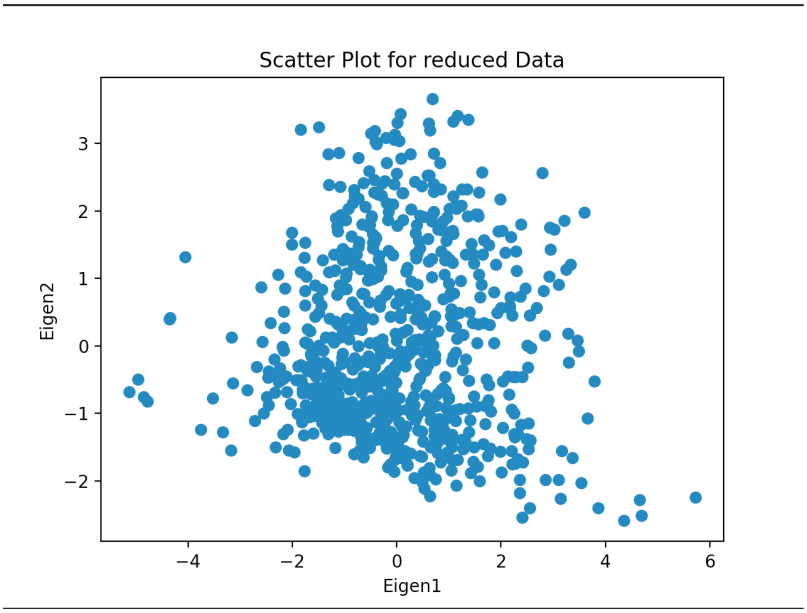


**Figure 5 Plot of data after dimensionality reduction**

**Inferences:**

1. From the above scatter plot, it is observable that the data does not have a positive or a negative slop. All data points are randomly spread. Thus, both the data are independent.

**b.**

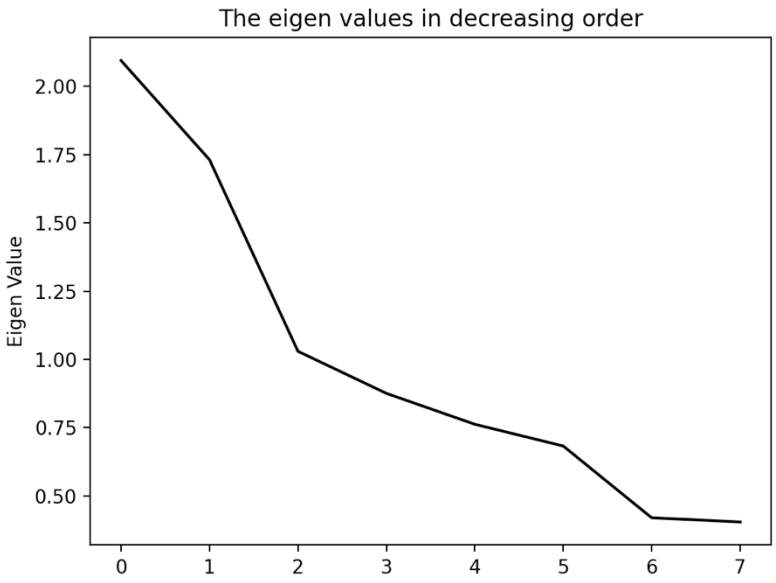


**Fig 6 Plot of Eigenvalues in decreasing order**

**Inferences:**

1. There is a gradual drop in the value of eigenvalues from second to third.
2. From third eigenvalue, the rate of decrease changes substantially.

**c.**

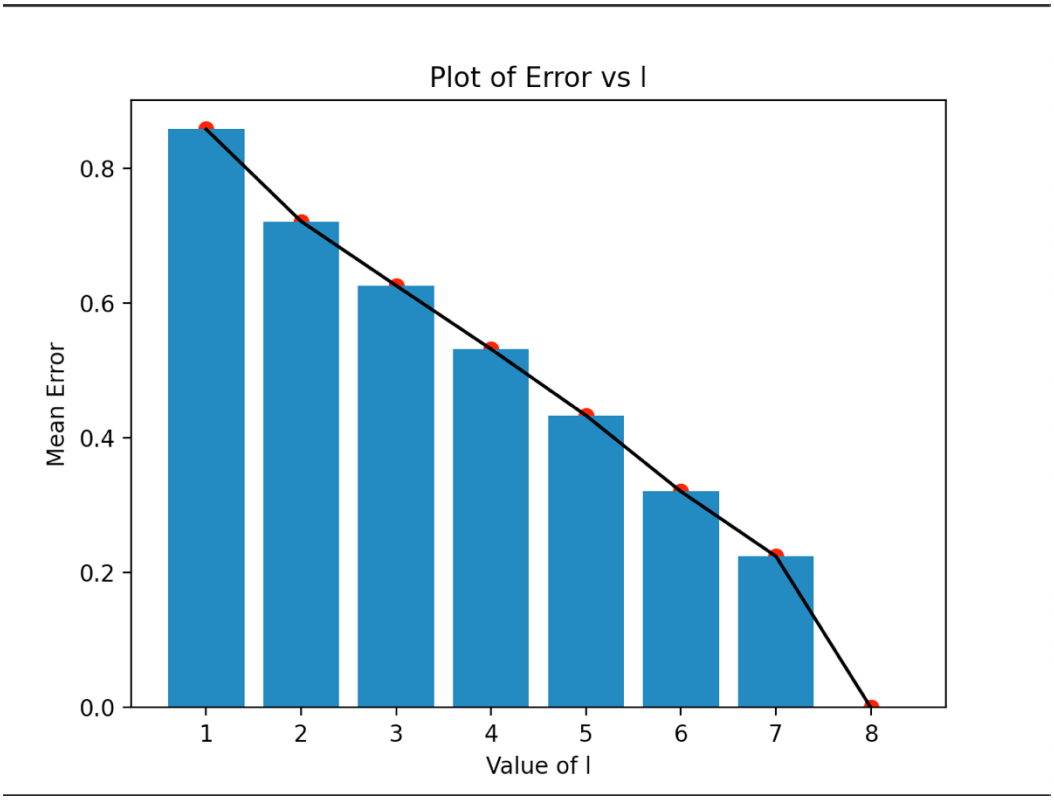


**Figure 6 Line plot to demonstrate reconstruction error vs. components**

**Inferences:**

1. More the value of reconstruction error, lesser is the quality of reconstruction, thus more chances of giving wrong analysis. For L = 8, the error is almost zero. But as l keeps increasing, the error increases.

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

| | x1 | x2 |
|---|---|---|
| x1 | 2.094 | 0 |
| x2 | 0 | 1.731 |

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

| | x1 | x2 | x3 |
|---|---|---|---|
| x1 | 2.094 | 0 | 0 |
| x2 | 0 | 1.731 | 0 |
| 0 | 0 | 0 | 1.03 |

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

| | x1 | x2 | x3 | x4 |
|---|---|---|---|---|
| x1 | 2.094 | 0 | 0 | 0 |
| x2 | 0 | 1.731 | 0 | 0 |
| x3 | 0 | 0 | 1.03 | 0 |
| x4 | 0 | 0 | 0 | 0.876 |

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

| | x1 | x2 | x3 | x4 | x5 |
|---|---|---|---|---|---|
| x1 | 2.094 | 0 | 0 | 0 | 0 |
| x2 | 0 | 1.731 | 0 | 0 | 0 |
| x3 | 0 | 0 | 1.03 | 0 | 0 |
| x4 | 0 | 0 | 0 | 0.876 | 0 |
| x5 | 0 | 0 | 0 | 0 | 0.762 |

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

| | x1 | x2 | x3 | x4 | x5 | x6 |
|---|---|---|---|---|---|---|
| x1 | 2.094 | 0 | 0 | 0 | 0 | 0 |
| x2 | 0 | 1.731 | 0 | 0 | 0 | 0 |
| x3 | 0 | 0 | 1.03 | 0 | 0 | 0 |
| x4 | 0 | 0 | 0 | 0.876 | 0 | 0 |
| x5 | 0 | 0 | 0 | 0 | 0.762 | 0 |
| x6 | 0 | 0 | 0 | 0 | 0 | 0.683 |

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

| | x1 | x2 | x3 | x4 | x5 | x6 | x7 |
|---|---|---|---|---|---|---|---|
| x1 | 2.094 | 0 | 0 | 0 | 0 | 0 | 0 |
| x2 | 0 | 1.731 | 0 | 0 | 0 | 0 | 0 |
| x3 | 0 | 0 | 1.03 | 0 | 0 | 0 | 0 |
| x4 | 0 | 0 | 0 | 0.876 | 0 | 0 | 0 |
| x5 | 0 | 0 | 0 | 0 | 0.762 | 0 | 0 |
| x6 | 0 | 0 | 0 | 0 | 0 | 0.683 | 0 |
| x7 | 0 | 0 | 0 | 0 | 0 | 0 | 0.42 |

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

| | x1 | x2 | x3 | x4 | x5 | x6 | x7 | x8 |
|---|---|---|---|---|---|---|---|---|
| x1 | 2.094 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| x2 | 0 | 1.731 | 0 | 0 | 0 | 0 | 0 | 0 |
| x3 | 0 | 0 | 1.03 | 0 | 0 | 0 | 0 | 0 |
| x4 | 0 | 0 | 0 | 0.876 | 0 | 0 | 0 | 0 |
| x5 | 0 | 0 | 0 | 0 | 0.762 | 0 | 0 | 0 |
| x6 | 0 | 0 | 0 | 0 | 0 | 0.683 | 0 | 0 |
| x7 | 0 | 0 | 0 | 0 | 0 | 0 | 0.42 | 0 |
| x8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.404 |

**Inferences:**

1. All the off-diagonal elements are zero, indicating that the respective attributes are uncorrelated. As the dimension of the data is reduced, those which were correlated were lost.
2. The diagonal elements indicate the eigenvalues or moreover the variance of the respective attributes. While the off-diagonal elements, being uncorrelated are zero. As we chose to project the data on the eigen vectors, thus giving no correlation for other columns.
3. The diagonal values keep decreasing as we move from left to right.
4. As me took the eigen values in the sorted order, l = 1 has maximum variance.
5. From the magnitude, it can be inferred that first component captures data variations the best.
6. From the values, the seventh and eighth components shall be optimum for data reduction and reconstruction.
7. The highest value, i.e., the first diagonal element, remains the same for every l.
8. The second largest diagonal element remains the same irrespective of l.
9. The value of $3^{rd}$, $4^{th}$ … $7^{th}$ largest elements are all same irrespective of l. This is due to the fact that eigenvalues don't change with l.

**d.**

Table 11 Covariance matrix for original data

| | pregs | plas | pres | skin | test | BMI | pedi | Age |
|---|---|---|---|---|---|---|---|---|
| pregs | 1 | 0.129 | 0.141 | -0.082 | -0.074 | 0.018 | -0.034 | 0.544 |
| plas | 0.129 | 1 | 0.153 | 0.057 | 0.331 | 0.221 | 0.137 | 0.264 |
| pres (in mm Hg) | 0.141 | 0.153 | 1 | 0.207 | 0.089 | 0.282 | 0.041 | 0.24 |
| skin (in mm) | -0.082 | 0.057 | 0.207 | 1 | 0.437 | 0.393 | 0.184 | -0.114 |
| test (in mu U/mL) | -0.074 | 0.331 | 0.089 | 0.437 | 1 | 0.198 | 0.185 | -0.114 |
| BMI (in $kg/m^2$) | 0.018 | 0.221 | 0.282 | 0.393 | 0.198 | 1 | 0.141 | 0.036 |
| pedi | -0.034 | 0.137 | 0.041 | 0.184 | 0.185 | 0.141 | 1 | 0.034 |
| Age (in years) | 0.544 | 0.264 | 0.24 | -0.114 | -0.114 | 0.036 | 0.034 | 1 |

**Inferences:**

1. The off-diagonal elements are close to 0 but not so small to be called as zero. But after PCA, all of them gets reduced to almost zero.
2. The magnitude of covariance for diagonal elements indicates the variance, which is 1. But after PCA for l = 8, all of them gets reduced.
3. There is no such trade of decrease in diagonal elements.