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

**Student's Name: Amit Maindola** **Mobile No: +91 7470985613**

**Roll Number: B20079** **Branch: Computer Science & 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 | 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 can make our analysis noisy and can violate our assumptions as we can see the attributes pres, test and pedi have many outliers
2. Here we replaced outliers with median , this reduces the distortion of our analysis.
3. Initially, the values having bigger values use to overpower the ones with smaller values.So, the analysis will be more partial. Now after normalization, each value is not between 5 to 12, so they will have equal weightage in the analysis.

**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 |
| 2 | plas | 121.66 | 30.438 | 0 | 1 |
| 3 | pres (in mm Hg) | 72.197 | 11.147 | 0 | 1 |
| 4 | skin (in mm) | 20.437 | 15.699 | 0 | 1 |
| 5 | test (in mu U/mL) | 60.897 | 77.644 | 0 | 1 |

| | | | | | |
|---|---|---|---|---|---|
| 6 | BMI (in kg/m$^2$) | 32.198 | 6.411 | 0 | 1 |
| 7 | pedi | 0.428 | 0.245 | 0 | 1 |
| 8 | Age (in years) | 32.760 | 11.055 | 0 | 1 |

**Inferences:**

1. Initially, the values having bigger values use to overpower the ones with smaller values. So, the analysis will be more partial. Now after standardization, every value has a common mean of 0 with variance 1. So there isn't much variation.

**2 a.**


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

**Inferences:**

1. Attribute 2 is negatively correlated to the Attribute 1 as we can observe from the graph. The covariance will be negative.
2. Observing the density of the graph, the distribution of both the attributes seem to be symmetric. The mean of both the attribute is approximately 0.

**b.**

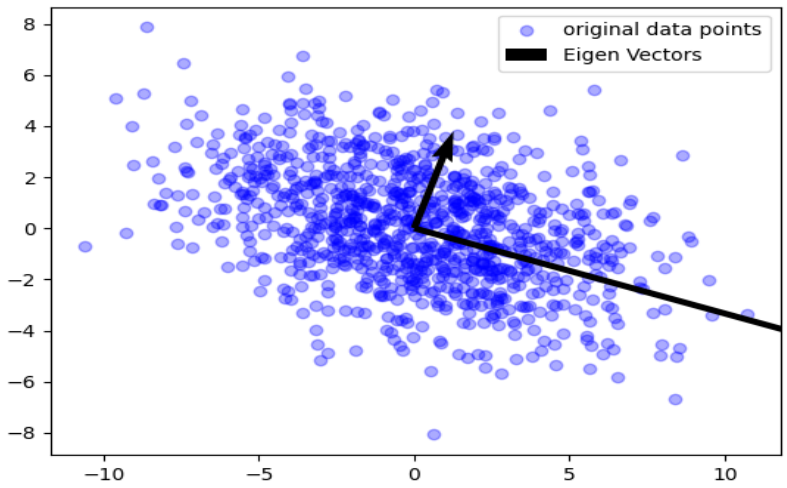


**Figure 2 Plot of 2D synthetic data and Eigen directions**

**Inferences:**

1. The spread along the first eigen vector is not much as compared to the spread along the second eigen vector. The data is more of a kind spread in a linear fashion with a smaller portion in the other vector's direction.
2. The density of points near the intersection of axis is very dense, and it gradually decreases as the spread increases. In other word, the number of points decreases as move far from the center.

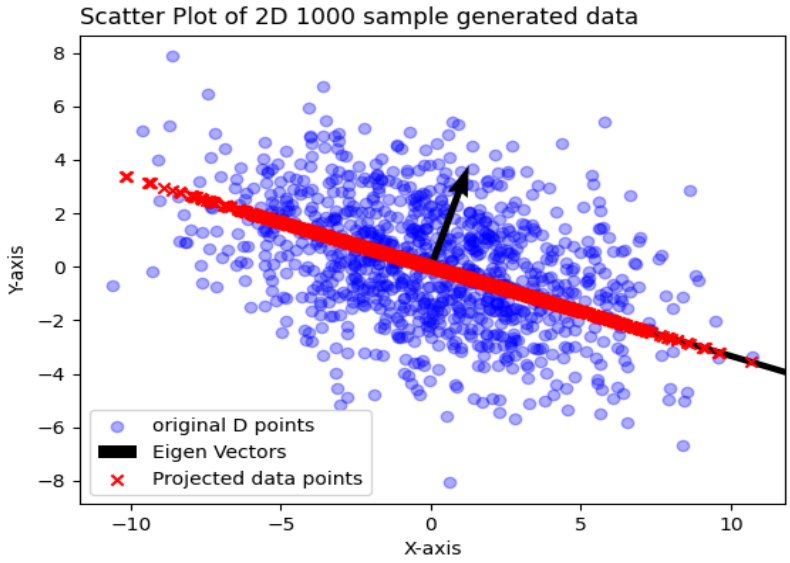


**Figure 3 Projected Eigen directions onto the scatter plot with 1st Eigen direction highlighted**

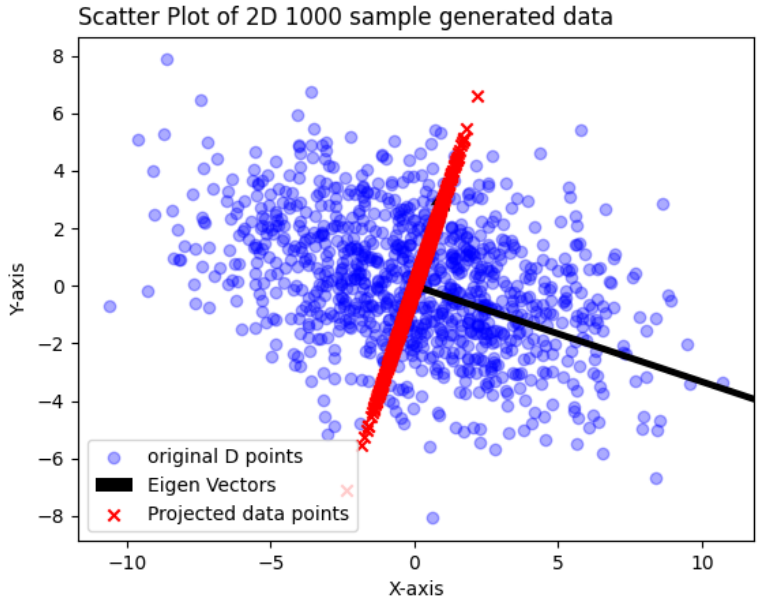


**Figure 4 Projected Eigen directions onto the scatter plot with 2nd Eigen direction highlighted**

**Inferences:**

1. The greater Eigen Value has more spread of data along its respective Eigen direction while the smaller Eigen Value has less spread along its respective Eigen direction.
2. Regarding the density, along the first Eigen vector (smaller line), the variance is not very large, so the spread is not so much varying. However, along the second Eigen Vector, the variance is high, so the spread is more, so the density actually is high near the intersection and spread is large.

**d.** Reconstruction error = 0.000

**Inferences:**

1. More the reconstruction error, more loss in the nature of data. So, the reconstruction error must not be very high. Here the reconstruction error was zero because the number of dimension remained same after reconstructing the data.

**3 a.**

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

| Direction | Variance | Eigenvalue |
|---|---|---|
| 1 | 1.984 | 1.842 |
| 2 | 1.984 | 1.842 |

**Inferences:**

1. Higher the values of Eigen Vector, more variance along that vector, so more strength along that direction. So, we can say that data will be more spread along the first Eigen vector.

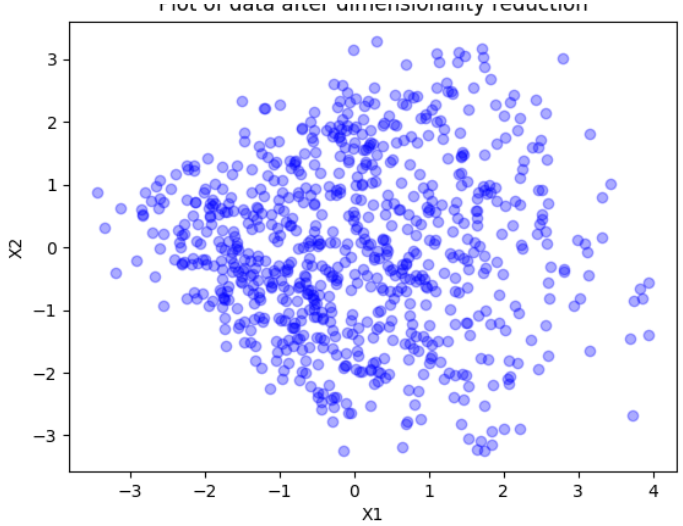


**Figure 5 Plot of data after dimensionality reduction**

**Inferences:**

1. As the density along positive slope is more so it seems by looking at the graph that the data is positively correlated

**b.**

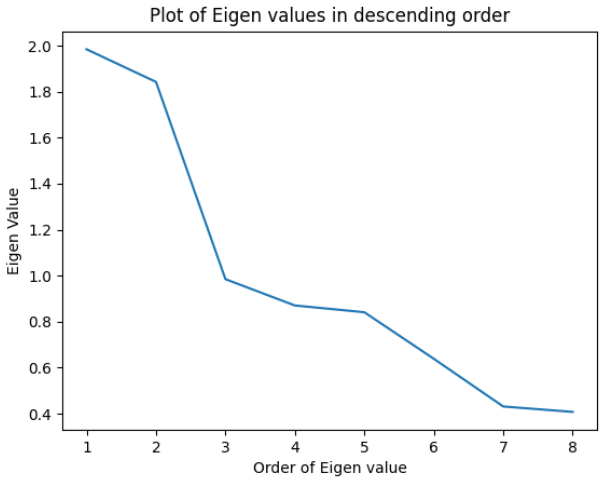


**Figure 6 Plot of Eigenvalues in descending order**

**Inferences:**

1. It drops rapidly from second to third Eigen value and then decreases gradually.
2. From the third Eigenvalue the rate of decrease changes substantially.

**c.**

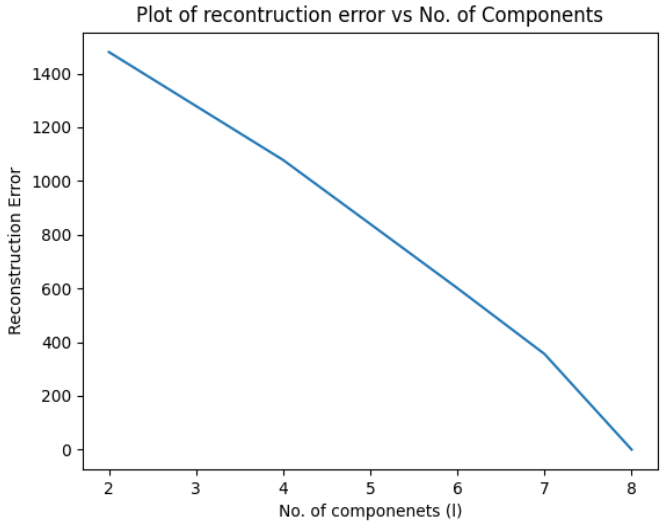


**Figure 7 Line plot to demonstrate reconstruction error vs. components**

**Inferences:**

1. More the magnitude of reconstruction error, lesser the quality of reconstructions. As we can see the RMSE increases, as we keep dropping the dimensions At l = d = 8, the reconstruction error is almost zero.

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

| | x1 | x2 |
|---|---|---|
| x1 | 1.984 | 0 |
| x2 | 0 | 1.984 |

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

| | x1 | x2 | x3 |
|---|---|---|---|
| x1 | 1.984 | 0 | 0 |
| x2 | 0 | 1.842 | 0 |

| | | | |
|---|---|---|---|
| x3 | 0 | 0 | 0.984 |

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

| | x1 | x2 | x3 | x4 |
|---|---|---|---|---|
| x1 | 1.984 | 0 | 0 | 0 |
| x2 | 0 | 1.842 | 0 | 0 |
| x3 | 0 | 0 | 0.984 | 0 |
| x4 | 0 | 0 | 0 | 0.870 |

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

| | x1 | x2 | x3 | x4 | x5 |
|---|---|---|---|---|---|
| x1 | 1.984 | 0 | 0 | 0 | 0 |
| x2 | 0 | 1.842 | 0 | 0 | 0 |
| x3 | 0 | 0 | 0.984 | 0 | 0 |
| x4 | 0 | 0 | 0 | 0.870 | 0 |
| x5 | 0 | 0 | 0 | 0 | 0.841 |

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

| | x1 | x2 | x3 | x4 | x5 | x6 |
|---|---|---|---|---|---|---|
| x1 | 1.984 | 0 | 0 | 0 | 0 | 0 |
| x2 | 0 | 1.842 | 0 | 0 | 0 | 0 |
| x3 | 0 | 0 | 0.984 | 0 | 0 | 0 |
| x4 | 0 | 0 | 0 | 0.870 | 0 | 0 |
| x5 | 0 | 0 | 0 | 0 | 0.841 | 0 |
| x6 | 0 | 0 | 0 | 0 | 0 | 0.639 |

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

| | x1 | x2 | x3 | x4 | x5 | x6 | x7 |
|---|---|---|---|---|---|---|---|
| x1 | 1.984 | 0 | 0 | 0 | 0 | 0 | 0 |
| x2 | 0 | 1.842 | 0 | 0 | 0 | 0 | 0 |
| x3 | 0 | 0 | 0.984 | 0 | 0 | 0 | 0 |
| x4 | 0 | 0 | 0 | 0.870 | 0 | 0 | 0 |
| x5 | 0 | 0 | 0 | 0 | 0.841 | 0 | 0 |
| x6 | 0 | 0 | 0 | 0 | 0 | 0.639 | 0 |
| x7 | 0 | 0 | 0 | 0 | 0 | 0 | 0.431 |

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

| | x1 | x2 | x3 | x4 | x5 | x6 | x7 | x8 |
|---|---|---|---|---|---|---|---|---|
| x1 | 1.984 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| x2 | 0 | 1.842 | 0 | 0 | 0 | 0 | 0 | 0 |
| x3 | 0 | 0 | 0.984 | 0 | 0 | 0 | 0 | 0 |
| x4 | 0 | 0 | 0 | 0.870 | 0 | 0 | 0 | 0 |
| x5 | 0 | 0 | 0 | 0 | 0.841 | 0 | 0 | 0 |
| x6 | 0 | 0 | 0 | 0 | 0 | 0.639 | 0 | 0 |
| x7 | 0 | 0 | 0 | 0 | 0 | 0 | 0.431 | 0 |
| x8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.408 |

**Inferences:**

1. As we want to reduce the dimensionality so while reconstructing the data we lost those dimensions that we chose to discard so the off diagonal elements are not correlated.
2. The diagonal elements are variances and off diagonal elements are zero we chose the Eigen vectors to project data on so the maximum variance is only along Eigen vectors(diagonal elements) and there is almost no covariance along the direction of other vectors.
3. The diagonal values keep decreasing as we move from left to right.
4. This is because we took the l strongest Eigen vectors(highest Eigen values) so the trend is decreasing.
5. From the magnitude of diagonal elements, the first component captures data variations the best.
6. From the value of diagonal elements, estimate the number of components that shall 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 as we took the l strongest vectors so for each value of l ranging from 2 to 8 the strongest Eigen vector i.e. highest Eigen value remains the same.
8. The magnitude of the 2nd diagonal element (topmost left corner) in each of the obtained covariance matrices is same as we took the l strongest vectors so for each value of l ranging from 2 to 8 the second strongest Eigen vector i.e. second highest Eigen value remains the same.

9. The 3rd, 4th, 5th, 6th, and 7th diagonal elements(if present) across covariance matrices are same because we chose l strongest vectors to project data on so the value and order always remains same and in the covariance matrixes where there isn't any 3rd, 4th, 5th, 6th or 7th diagonal present is because we took dimension less than 3, 4, 5, 6 or 7 respectively.

**d.**

Table 11 Covariance matrix for original data

| | pregs | plas | pres | skin | test | BMI | pedi | Age |
|---|---|---|---|---|---|---|---|---|
| pregs | 1 | 0.118 | 0.209 | -0.096 | -0.106 | 0.028 | -0.017 | 0.562 |
| plas | 0.118 | 1 | 0.204 | 0.060 | 0.174 | 0.228 | 0.062 | 0.264 |
| pres (in mm Hg) | 0.209 | 0.204 | 1 | 0.026 | -0.051 | 0.272 | 0.009 | 0.317 |
| skin (in mm) | -0.096 | 0.060 | 0.026 | 1 | 0.476 | 0.373 | 0.154 | -0.086 |
| test (in mu U/mL) | -0.106 | 0.174 | -0.051 | 0.476 | 1 | 0.167 | 0.190 | -0.059 |
| BMI (in kg/m$^2$) | 0.028 | 0.228 | 0.272 | 0.373 | 0.167 | 1 | 0.121 | 0.083 |
| pedi | -0.017 | 0.062 | 0.009 | 0.154 | 0.190 | 0.121 | 1 | 0.038 |
| Age (in years) | 0.562 | 0.264 | 0.317 | -0.086 | -0.059 | 0.083 | 0.038 | 1 |

**Inferences:**

1. The off-diagonal values in original matrix are close to zero but not that close to zero so as to treat them negligible but in the covariance matrix obtained after PCA l=8 reduction the off diagonal values are pretty close to zero that we considered them to be zero.
2. The magnitude of diagonal elements in original covariance matrix is one while in the covariance matrix obtained after PCA l=8 reduction the diagonal values are not one and in fact less in magnitude than that of original covariance matrix.
3. No, there isn't any trade of decrease in diagonal elements like covariance obtained after dimensionality reduction in fact in original matrix all diagonal values are equal.