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

Student's Name: SARTHAK SAPTAMI KUMAR JHA **Mobile No: 8825319259**

Roll Number: B20317 **Branch: CSE**

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	0.00	13.00	5.00	12.00
2	plas	44.00	199.00	5.00	12.00
3	pres (in mm Hg)	38.00	106.00	5.00	12.00
4	skin (in mm)	0.00	63.00	5.00	12.00
5	test (in mu U/mL)	0.00	318.00	5.00	12.00
6	BMI (in kg/m^2)	18.20	50.00	5.00	12.00
7	pedi	0.078	1.191	5.00	12.00
8	Age (in years)	21.00	66.00	5.00	12.00

Inferences:

1. Outlier correction is necessary, else the outliers will dominate the min-max normalization process.
2. We compute the median of the data without including any outliers, and then replace the outliers with the median. The use of median for replacement is better because, median depends on the order of the data. Changing the lowest or highest data point does not affect the order of the data, thus the median is not affected by the change in value.
3. We observe that after min-max normalization, each attribute is restricted between 5 and 12, thus giving each attribute a common 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	3.78	3.27	0	1
2	plas	121.66	30.44	0	1
3	pres (in mm Hg)	72.2	11.15	0	1
4	skin (in mm)	20.44	15.7	0	1
5	test (in mu U/mL)	59.57	78.42	0	1

6	BMI (in kg/m^2)	32.2	6.41	0	1
7	pedi	0.43	0.25	0	1
8	Age (in years)	32.76	11.06	0	1

Inferences:

1. The mean and standard deviation are standardized for each attribute, with each having mean 0 and standard deviation 1.

2 a.

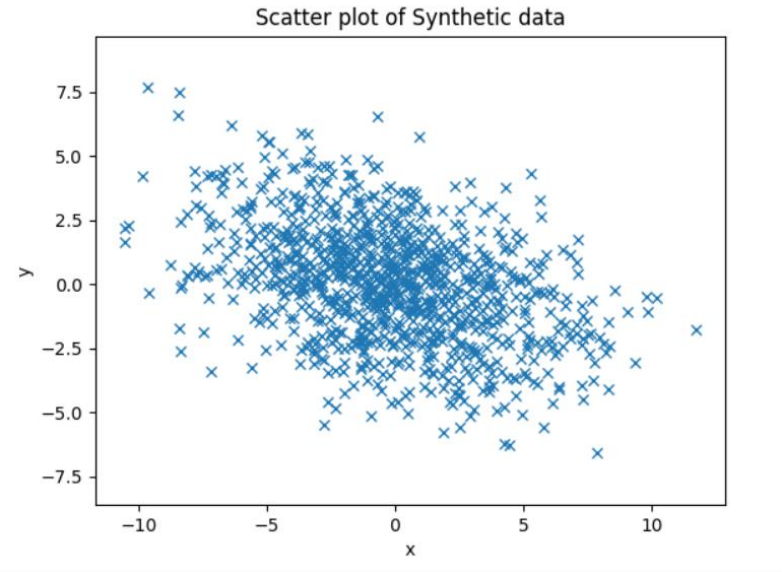

Figure 1 Scatter plot of 2D synthetic data of 1000 samples

Inferences:

1. Attribute 1 and Attribute 2 are negatively and strongly correlated.
2. The points are densely packet close to the origin.

b.

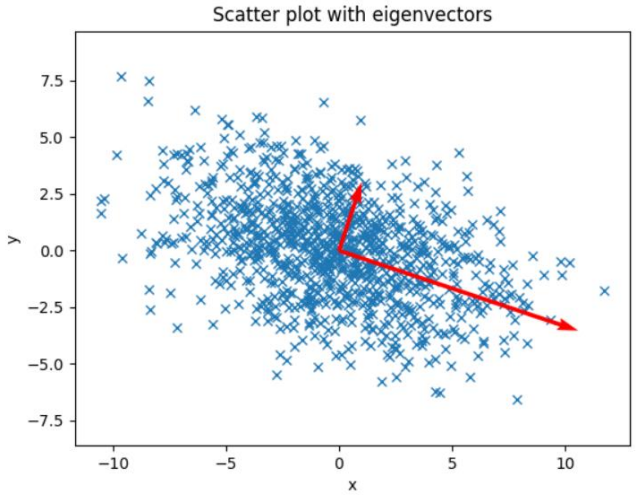

Figure 2 Plot of 2D synthetic data and Eigen directions

Inferences:

1. The eigenvalue with higher magnitude has higher spread of the data in its corresponding eigenvector's direction.
2. There is a higher density of data points close to the intersection of the eigenvectors, and sparse distribution at points away from the intersection of the eigenvectors.

c.

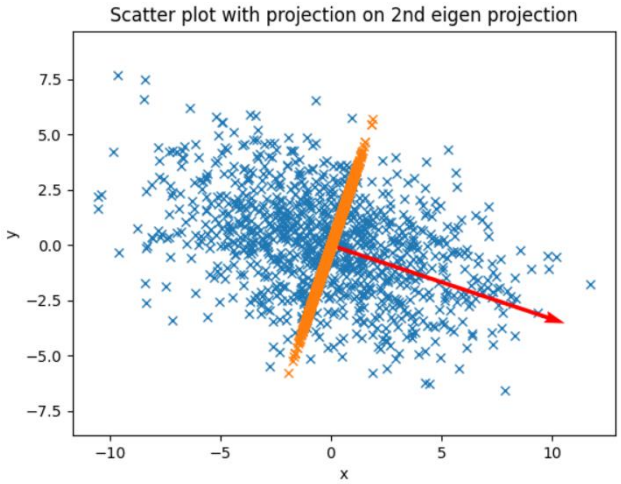

Figure 3 Projected Eigen directions onto the scatter plot with 1st Eigen direction highlighted

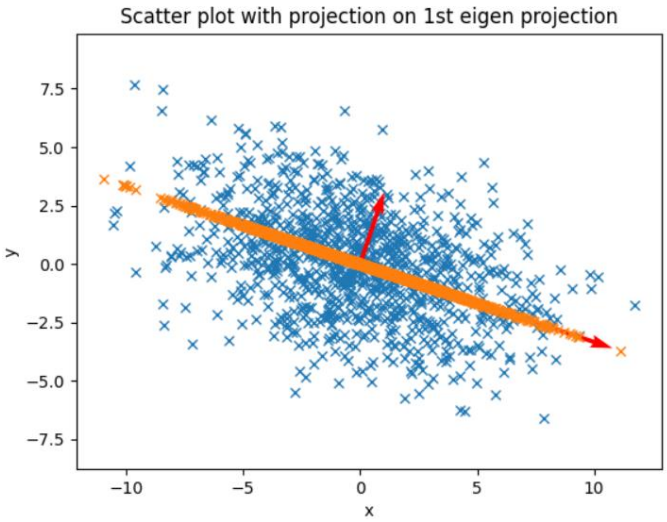

Figure 4 Projected Eigen directions onto the scatter plot with 2nd Eigen direction highlighted

Inferences:

1. The eigenvalue with higher magnitude has higher magnitude has higher spread on its eigenvector as compared to the other eigenvalue.
2. The spread of points is more for the eigenvalue with higher magnitude. The variance also follows the same trend. The density of points is highest at the point of intersection of the eigenvectors.

d. Reconstruction error = 1.660

Inferences:

1. Higher reconstruction error leads to lower quality of reconstruction of data.

3 a.

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

Direction	Variance	Eigenvalue
1	1.98	1.98
2	1.84	1.84

Inferences:

1. The eigenvalue having higher magnitude has higher amount of variance in the direction of its corresponding eigenvector.

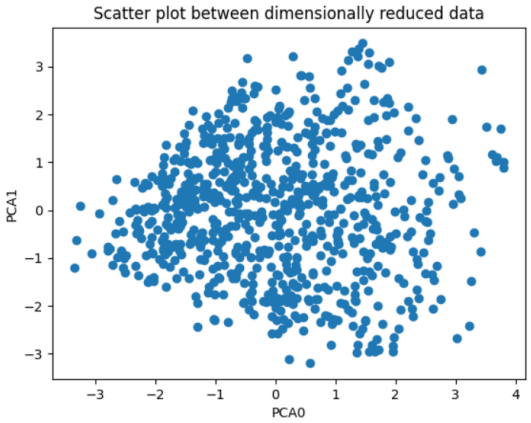

Figure 5 Plot of data after dimensionality reduction

Inferences:

1. There is no correlation between the dimensionally reduced data.
2. The scatter plot correctly shows the relation between dimensionally reduced data, as the dimensionally reduced data must be uncorrelated, which can be seen in the scatter plot.

b.

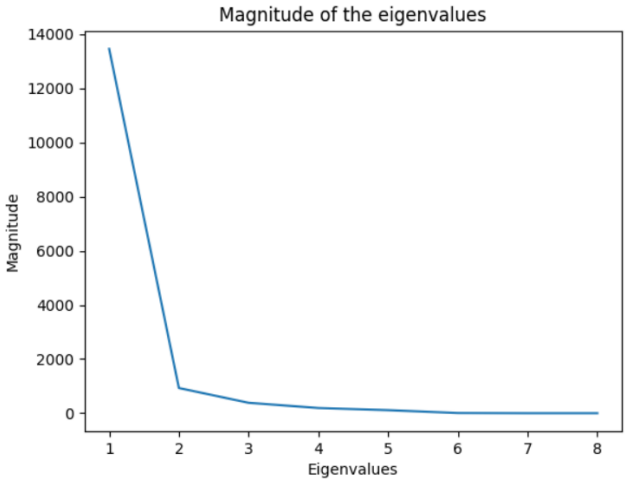

Figure 6 Plot of Eigenvalues in descending order

Inferences:

1. The value of eigenvalues drops significantly in the first couple of eigenvalues; thus, they contribute the highest to the dimensionality reduction, the rest of eigenvalues are low in value and decline slowly.
2. The eigenvalue 2 has the maximum change, as eigenvalue 1 is very high close to 13450, while eigenvalue 3 is close to 390.
3. Using the original, non-normalized and non-standardized data is better for eigenvalue visualization.
4. **c.**

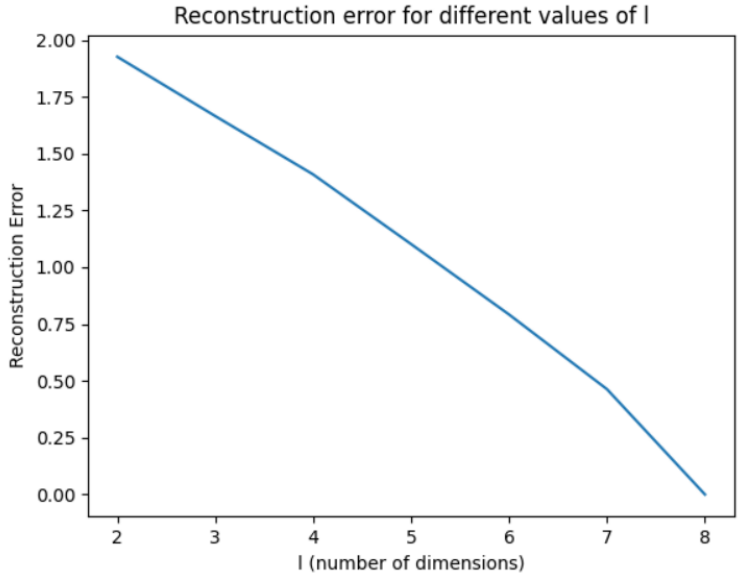

Figure 7 Line plot to demonstrate reconstruction error vs. components

Inferences:

1. If we have very high reconstruction error, it will lead to poor quality of reconstruction with larger margin of errors.
2. The greater number of components we choose to include in the dimensionality reduction, the lesser amount of reconstruction error is seen in the data.

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

	x1	x2
x1	1.98	0
x2	0	1.84

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

	x1	x2	x3
x1	1.98	0	0
x2	0	1.84	0
x3	0	0	0.98

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

	x1	x2	x3	x4
x1	1.98	0	0	0
x2	0	1.84	0	0
x3	0	0	0.98	0
x4	0	0	0	0.85

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

	x1	x2	x3	x4	x5
x1	1.98	0	0	0	0
x2	0	1.84	0	0	0
x3	0	0	0.98	0	0
x4	0	0	0	0.85	0
x5	0	0	0	0	0.84

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

	x1	x2	x3	x4	x5	x6
x1	1.98	0	0	0	0	0
x2	0	1.84	0	0	0	0
x3	0	0	0.98	0	0	0
x4	0	0	0	0.85	0	0
x5	0	0	0	0	0.84	0
x6	0	0	0	0	0	0.64

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

	x1	x2	x3	x4	x5	x6	x7
x1	1.98	0	0	0	0	0	0
x2	0	1.84	0	0	0	0	0
x3	0	0	0.98	0	0	0	0
x4	0	0	0	0.85	0	0	0
x5	0	0	0	0	0.84	0	0
x6	0	0	0	0	0	0.64	0
x7	0	0	0	0	0	0	0.45

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

	x1	x2	x3	x4	x5	x6	x7	x8
x1	1.98	0	0	0	0	0	0	0
x2	0	1.84	0	0	0	0	0	0
x3	0	0	0.98	0	0	0	0	0
x4	0	0	0	0.85	0	0	0	0
x5	0	0	0	0	0.84	0	0	0
x6	0	0	0	0	0	0.64	0	0
x7	0	0	0	0	0	0	0.45	0
x8	0	0	0	0	0	0	0	0.40

Inferences:

1. All the off-diagonal elements are zero because all the attributes are uncorrelated with each other in the reduced representation.
2. The diagonal elements represent the variance of each attribute of the data.
3. The values of the diagonal elements continue decreasing as they are arranged in a manner such that the eigenvalue with highest value is placed first, followed by less significant eigenvalues.
4. There is a decrease in the values of diagonal elements as lesser significant eigenvalues are used in each of the next values of l which are visible in next values of x.
5. Since the value of x1 is highest, we can say that x1, captures the most data, but for optimal amount of data capture, we must include as many values of x as possible.
6. From the values of diagonal elements, we can say that the first three will capture close to 60% of the data, and can be optimally used for reconstruction along with dimensionality reduction.
7. The magnitude of the first diagonal element is same in all covariance matrices, because each time, we use the most significant eigenvalue to calculate the first column.
8. For the same reason as above, after the most significant eigenvalue is used, the next largest is used for dimensionality reduction in the next column.

9. They are all same for the reasons as stated above.
10. Using covariance matrix 3 or l=3 can be sufficient for reconstruction of the data as it captures around 60% of the data.

d.

Table 11 Covariance matrix for original data

	pregs	plas	pres	skin	test	BMI	pedi	Age
pregs	1	0.11	0.20	-0.09	-0.10	0.02	0.004	0.56
plas	0.11	1	0.30	0.06	0.15	0.22	0.08	0.27
pres (in mm Hg)	0.20	0.30	1	0.02	-0.04	0.27	0.02	0.32
skin (in mm)	-0.09	0.06	0.02	1	0.45	0.37	0.15	-0.10
test (in mu U/mL)	-0.10	0.15	-0.04	0.45	1	0.16	0.19	-0.07
BMI (in kg/m^2)	0.02	0.22	0.27	0.37	0.16	1	0.12	0.07
pedi	0.004	0.08	0.02	0.15	0.19	0.12	1	0.03
Age (in years)	0.56	0.27	0.32	-0.10	-0.07	0.07	0.03	1

Inferences:

1. The off-diagonal values in original data have some value, because columns in original data have some correlation with each other, while in dimensionally reduced data, the columns are all uncorrelated, and thus the off-diagonal elements have values 0.
2. The diagonal elements all have values, 1 as the original data has been standardized so, its diagonal elements, which represent variance have values 1. In the dimensionally reduced data, the columns have some values of variance dependent on the eigenvalues used for forming them.
3. No, all the diagonal elements have the same value 1 since, it was standardized data.