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

Student's Name: Abhay Gupta **Mobile No:** 9511334630

Roll Number: B20075 **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	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. Mostly, outliers indicate the noise in the data. So, they are replaced by median to get the outlier corrected data.
2. As the mean value is highly influenced by the outliers, it is advised to replace the outliers with the median value.
3. Data after normalization is the scaled version of the original data, so that the data fall within a small specified range.
4. Before the normalization, data was having varying range among attributes. But, after the normalization process, whole of the data is lying within 5 and 12.

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	3.782	3.270	0	1
2	plas	121.656	30.438	0	1
3	pres (in mm Hg)	72.196	11.146	0	1
4	skin (in mm)	20.437	15.698	0	1

5	test (in mu U/mL)	60.897	77.644	0	1
6	BMI (in kg/m^2)	32.198	6.410	0	1
7	pedi	0.427	0.245	0	1
8	Age (in years)	32.760	11.055	0	1

Inferences:

1. Before the standardization process, every attribute was having different mean and variance. But after standardization process data is rescaled in such a way that mean and variance of each attribute becomes 0 and 1 respectively.
2. The mean of the data was not exactly 0, but it was of the order 10^{-17}, which can be assumed as 0.

2 a.

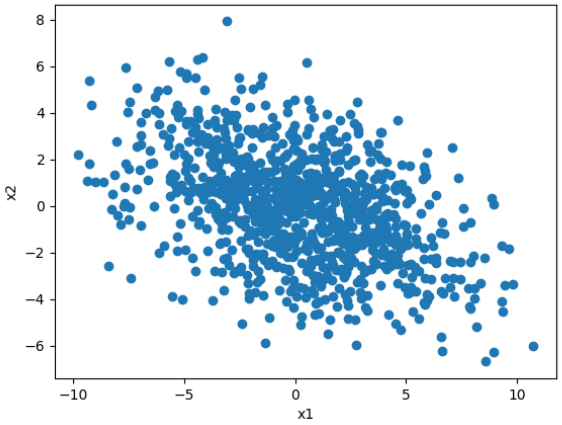

Figure 1 Scatter plot of 2D synthetic data of 1000 samples

Inferences:

1. According to the spread of the data, attribute 1 is negatively correlated to attribute 2.
2. Density of the data is maximum in the range of -5 to 5.

b.

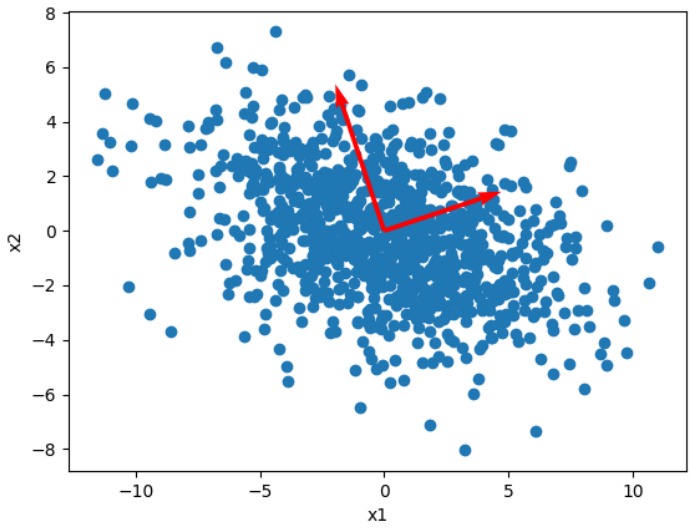

Figure 2 Plot of 2D synthetic data and Eigen directions

Inferences:

1. Spread/Variance of data is high if the magnitude of eigen values is high.
2. Density of points near the intersection of eigen vectors is maximum. It reduces gradually as we move away from the point of intersection

c.

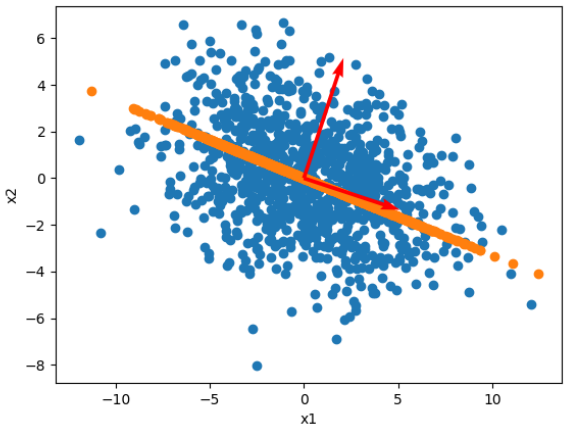

Figure 3 Projected Eigen directions onto the scatter plot with 1st Eigen direction highlighted

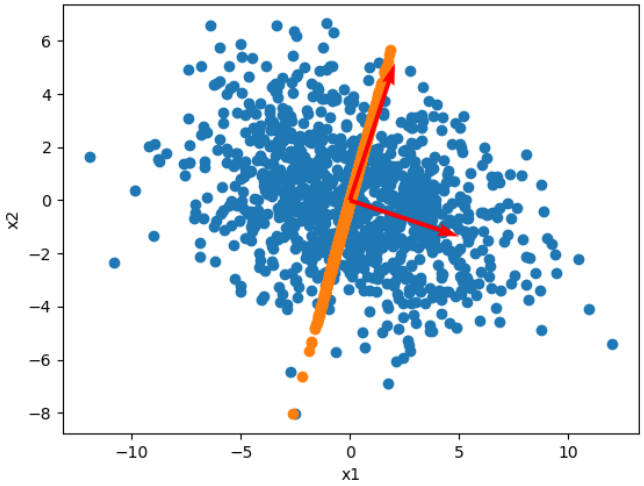

Figure 4 Projected Eigen directions onto the scatter plot with 2nd Eigen direction highlighted

Inferences:

1. Magnitudes of eigenvalues are 14.07 and 4.1
2. If the magnitude of eigen value is high, then the spread is high.

d. Reconstruction error = 0

Inferences:

1. If the reconstruction error is low, then reconstructed data will be close to the original data.
2. Here reconstruction error is of the order of 10^{-16}. So, it is given as 0.

3 a.

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

Direction	Variance	Eigenvalue
1	1.992	1.989
2	1.853	1.850

Inferences:

1. From the above table, we can see that eigenvalues are almost equal to the variance of the projected data.

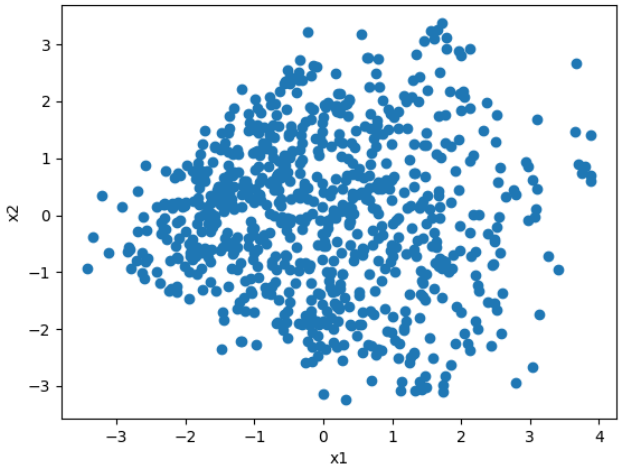

Figure 5 Plot of data after dimensionality reduction

Inferences:

1. From the above scatter plot we can say that both the attributes are not correlated.

2. So, it is seen that after applying PCA we get uncorrelated data.

b.

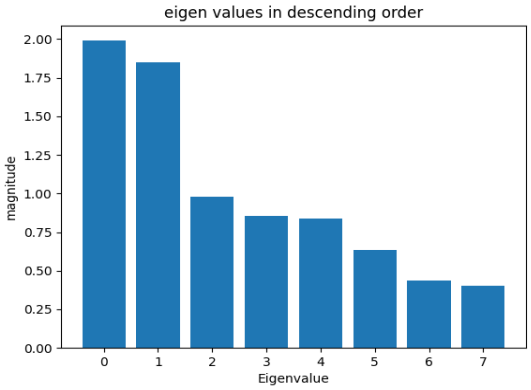

Figure 6 Plot of Eigenvalues in descending order

Inferences:

1. It drops significantly from second to third eigen value. But decreases gradually from third eigen value.
2. From second value rate change significantly.

.

c.

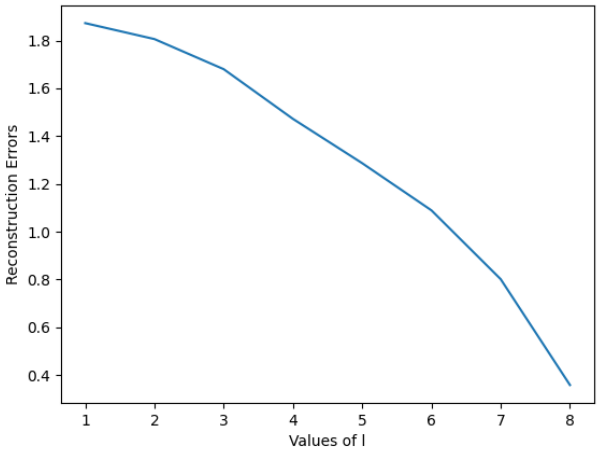

Figure 7 Line plot to demonstrate reconstruction error vs. components

Inferences:

1. As the magnitude of reconstruction increases, the quality of data becomes low.
2. As l increases, reconstruction error decreases.

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

	x1	x2
x1	1.989	0
x2	0	1.85

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

	x1	x2	x3
x1	1.989	0	0
x2	0	1.85	0
x3	0	0	0.980

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

	x1	x2	x3	x4
x1	1.989	0	0	0
x2	0	1.85	0	0
x3	0	0	0.98	0
x4	0	0	0	.857

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

	x1	x2	x3	x4	x5
x1	1.989	0	0	0	0
x2	0	1.85	0	0	0
x3	0	0	0.98	0	0
x4	0	0	0	.857	0
x5	0	0	0	0	0.837

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

	x1	x2	x3	x4	x5	x6
x1	1.989	0	0	0	0	0
x2	0	1.85	0	0	0	0
x3	0	0	0.98	0	0	0
x4	0	0	0	.857	0	0
x5	0	0	0	0	0.837	0
x6	0	0	0	0	0	0.635

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

	x1	x2	x3	x4	x5	x6	x7
x1	1.989	0	0	0	0	0	0
x2	0	1.85	0	0	0	0	0
x3	0	0	0.98	0	0	0	0
x4	0	0	0	.857	0	0	0
x5	0	0	0	0	0.837	0	0
x6	0	0	0	0	0	0.635	0
x7	0	0	0	0	0	0	0.433

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

	x1	x2	x3	x4	x5	x6	x7	x8
x1	1.989	0	0	0	0	0	0	0
x2	0	1.85	0	0	0	0	0	0
x3	0	0	0.98	0	0	0	0	0
x4	0	0	0	.857	0	0	0	0
x5	0	0	0	0	0.837	0	0	0
x6	0	0	0	0	0	0.635	0	0
x7	0	0	0	0	0	0	0.433	0
x8	0	0	0	0	0	0	0	0.404

Inferences:

1. Off diagonal elements are almost 0. Because after PCA we get uncorrelated data
2. Diagonal elements represents variance(spread). While off diagonal elements represents covariance.
3. Magnitude of diagonal elements decreases from x1 to x8.
4. From diagonal elements we can say that 2 eigen vectors can give good projected data.
5. From diagonal elements we can say that 2 eigen vectors can give good reconstruction data.

6. Magnitude of 1st diagonal element is same in each of the above table. This is because it represents the variance of attribute 1, which should be same .
7. Magnitude of 2nd diagonal element is same in each of the above table. This is because it represents the variance of attribute 1, which should be same.
8. Magnitude of 3rd, 4th, 5th, 6th, and 7th diagonal elements are same.

d.

Table 11 Covariance matrix for original data

	pregs	plas	pres	skin	test	BMI	pedi	Age
pregs	0.998	0.117	0.208	-0.096	-0.108	0.028	0.004	0.56
plas	0.117	0.998	0.204	0.05	0.178	0.227	0.081	0.273
pres (in mm Hg)	0.208	0.204	0.998	0.025	-0.050	0.271	0.022	0.325
skin (in mm)	-0.096	0.059	0.025	0.998	0.471	0.373	0.152	-0.101
test (in mu U/mL)	-0.108	0.178	-0.050	0.471	0.998	0.171	0.198	-0.073
BMI (in kg/m^2)	0.028	0.227	0.271	0.373	0.171	0.998	0.123	0.077
pedi	0.004	0.081	0.022	0.152	0.198	0.123	0.998	0.036
Age (in years)	0.560	0.273	0.325	-0.101	-0.073	0.077	0.036	0.998

Inferences:

1. The off-diagonal elements of covariance matrix of original data are significant values. But the off-diagonal elements of the covariance matrix obtained after PCA l=8 reduction are of the order 10^{-16}.
2. The diagonal elements represents the variance. Some of the diagonal element of the covariance matrix obtained after PCA l=8 reduction are greater than that of original data and some are lower.
3. Diagonal elements did not change much unlike the non-diagonal elements which reduced significantly after PCA.
4. As we know that after applying PCA, data becomes uncorrelated due to which the non-diagonal elements tends to 0 in the covariance matrix obtained after PCA l=8 reduction.