IC 272: DATA SCIENCE - III

LAB ASSIGNMENT – III

Attribute normalization, standardization and dimension reduction of data

Student's Name: Arpit singh **Mobile No: 6265104315**

Roll Number: B20084 **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. Outliers increase the range of the data too much which is not good for proper data interpretation.
2. We use median to replace outliers as it is the most common value in the data.
3. Before normalization minimum and maximum values of the attributes are varying a lot but after normalisation every attribute hai same minimum and maximum value.

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	7.036	1.761	0	1
2	plas	8.507	1.374	0	1
3	pres (in mm Hg)	8.52	1.147	0	1
4	skin (in mm)	7.27	1.744	0	1
5	test (in mu U/mL)	6.34	1.708	0	1
6	BMI (in kg/m^2)	8.081	1.411	0	1
7	pedi	7.199	1.541	0	1
8	Age (in years)	6.829	1.719	0	1

Inferences:

1. After standardization, standard deviation of all the attributes becomes 1 and mean becomes 0 so now the deviation in the values of every attribute is equal.

2 a.

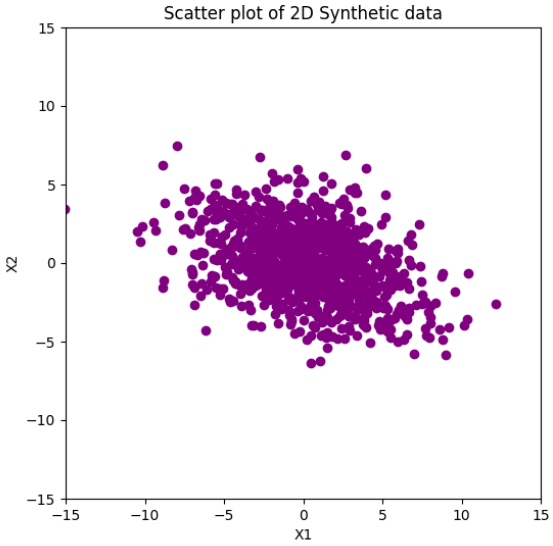

Figure 1 Scatter plot of 2D synthetic data of 1000 samples

Inferences:

1. Attribute 1 and attribute 2 both are negatively correlated as we can observe that decrease in the value of Y-axis increases the value in X-axis.
2. Density of the points is concentrated at the origin in the range of -5 to 5.

b.

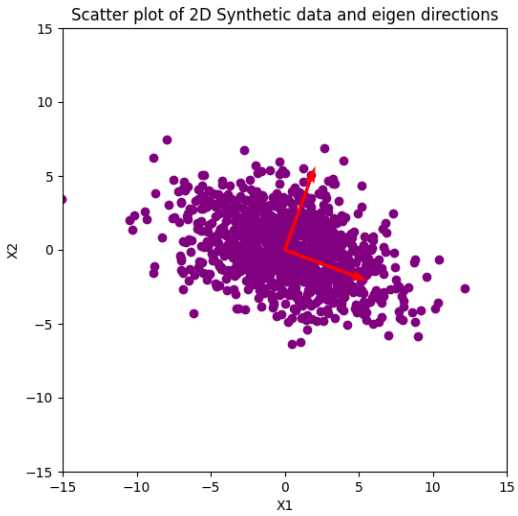

Figure 2 Plot of 2D synthetic data and Eigen directions

Inferences:

1. Spread of data is influenced by the magnitude of Eigenvalues, very less number of points are located beyond the magnitude of Eigenvalues.
2. The density of points near the intersection of Eigen axes is high as compared to the points gradually away from it.

c.

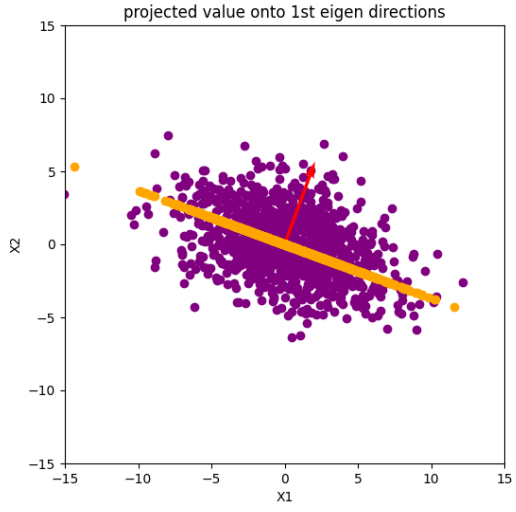

Figure 3 Projected Eigen directions onto the scatter plot with 1st Eigen direction highlighted

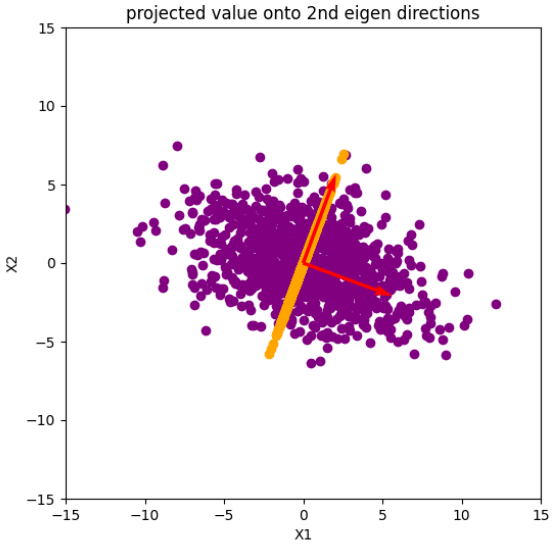

Figure 4 Projected Eigen directions onto the scatter plot with 2nd Eigen direction highlighted

Inferences:

1. We can observe that the spread of the 1st projection of the eigenvector is more as compared to the 2nd projection of the eigenvector and the magnitude of the 2nd projection of the eigenvector is more as compared to the 1st eigenvector.
2. As we can observe from that the spread is more for 1st eigenvector than 2nd eigenvector the magnitude of the 2nd eigen value is more than 1st means that data is more biased towards the 2nd values or 2nd eigenvector because the 2nd eigenvalues is significantly higher than 1st value so first eigen is more useful to represent data.

d. Reconstruction error = 0.0

Inferences:

1. Here, the magnitude of reconstruction error was almost 0. If the reconstruction error is near zero that means the quality of reconstruction is good. a high value of reconstruction error will result in deviation of data from the original data which may give false results while interpretation of the data.

3 a.

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

Direction	Variance	Eigenvalue
1	1.995	1.995
2	1.855	1.855

Inferences:

1. The variance of the projected data along the two directions is equal to the Eigenvalues of the two directions of projection.

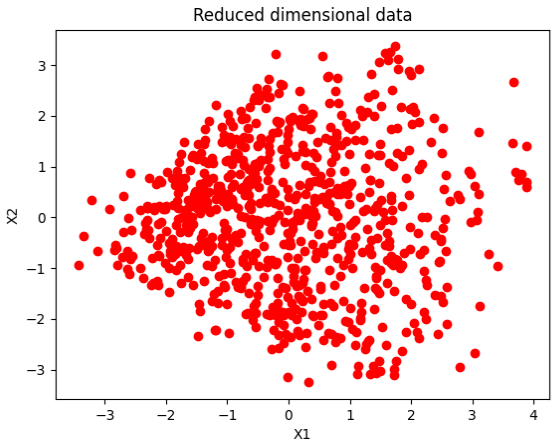

Figure 5 Plot of data after dimensionality reduction

Inferences:

1. There is no correlation between the two attributes obtained after dimensionality reduction from the spread of data points as we can not observe any increasing or decreasing trend.

b.

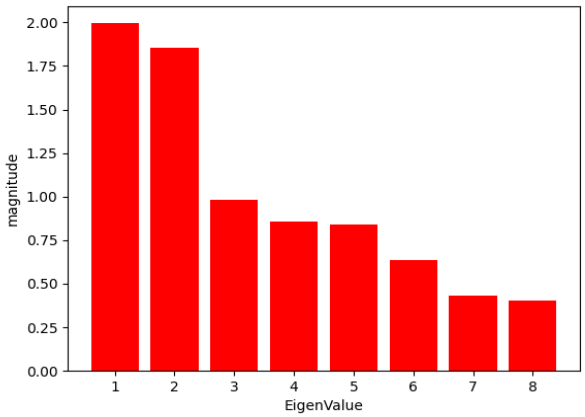

Figure 6 Plot of eigenvalues in decreasing order

Inferences:

1. After the first two eigenvalues they decrease rapidly then again the decrease becomes stable.
2. Eigenvalue equal to 2 is the value after which the rate of decrease changes substantially

c.

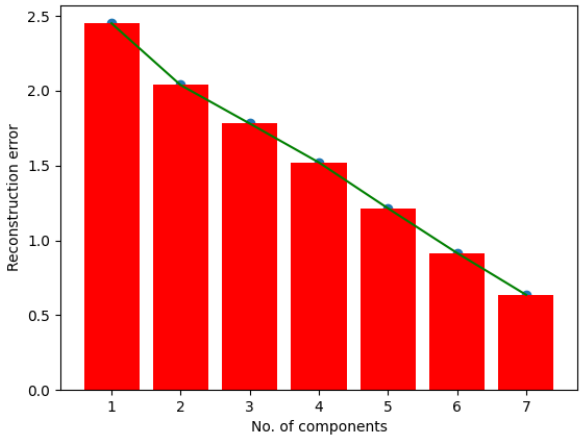

Figure 7 line plot to demonstrate reconstruction error vs. components

Inferences:

1. The magnitude of reconstruction error is directly proportional to the quality of reconstruction. As higher the error lower the quality of reconstruction.

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

	x1	x2
x1	1530.212	0
x2	0	1423.428

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

	x1	x2	x3
x1	1530.212	0	0
x2	0	1423.428	0
x3	0	0	754.083

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

	x1	x2	x3	x4
x1	1530.212	0	0	0
x2	0	1423.428	0	0
x3	0	0	754.083	0
x4	0	0	0	659.18

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

	x1	x2	x3	x4	x5
x1	1530.212	0	0	0	0
x2	0	1423.428	0	0	0
x3	0	0	754.083	0	0
x4	0	0	0	659.18	0
x5	0	0	0	0	644.16

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

	x1	x2	x3	x4	x5	x6
x1	1530.212	0	0	0	0	0
x2	0	1423.428	0	0	0	0
x3	0	0	754.083	0	0	0
x4	0	0	0	659.18	0	0
x5	0	0	0	0	644.16	0

x6	0	0	0	0	0	488.762

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

	x1	x2	x3	x4	x5	x6	x7
x1	1530.212	0	0	0	0	0	0
x2	0	1423.428	0	0	0	0	0
x3	0	0	754.083	0	0	0	0
x4	0	0	0	659.18	0	0	0
x5	0	0	0	0	644.16	0	0
x6	0	0	0	0	0	488.762	0
x7	0	0	0	0	0	0	333.422

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

	x1	x2	x3	x4	x5	x6	x7	x8
x1	1530.212	0	0	0	0	0	0	0
x2	0	1423.428	0	0	0	0	0	0
x3	0	0	754.083	0	0	0	0	0
x4	0	0	0	659.18	0	0	0	0
x5	0	0	0	0	644.16	0	0	0
x6	0	0	0	0	0	488.762	0	0
x7	0	0	0	0	0	0	333.422	0
x8	0	0	0	0	0	0	0	310.754

Inferences:

1. Value of off-diagonal elements is 0 as data is symmetric.
2. The diagonal values are significant while the off diagonal values are approximately 0 this is because the covariance matrix is symmetric.
3. diagonal values decrease from up to down .
4. As we go lower the eigenvalues decrease so corresponding projection value also decreases and these results in decrease in value of variance.
5. The 1st eigenvector is more reliable.

6. From the value of diagonal elements, the number of components that will give the optimum reconstruction with dimension reduction and eigenvectors should be 8.
7. The value of the 1st diagonal element is equal as the eigenvectors are the same.
8. The 2nd value is also equal as it is independent of other vectors.
9. 3rd, 4th, 5th, 6th, and 7th diagonal elements across covariance matrices. All are equal.

d.

Table 11 Covariance matrix for original data

	pregs	plas	pres	skin	test	BMI	pedi	Age
pregs	1	0.120	0.124	-0.095	-0.107	-0.037	-0.030	0.532
plas	0.120	1	0.125	0.040	0.169	0.186	0.112	0.225
pres (in mm Hg)	0.124	0.125	1	0.195	0.095	0.240	0.050	0.210
skin (in mm)	-0.095	0.040	0.195	1	0.458	0.328	0.132	-0.078
test (in mu U/mL)	-0.107	0.169	0.095	0.458	1	0.164	0.106	-0.078
BMI (in kg/m^2)	-0.037	0.186	0.240	0.328	0.164	1	0.042	0.127
pedi	-0.030	0.112	0.050	0.132	0.106	0.042	1	0.028
Age (in years)	0.532	0.225	0.210	-0.078	-0.055	0.127	0.028	1

Inferences:

1. The off-diagonal values have a similar trend as compared with the covariance matrix obtained after PCA l=8 reduction but in case of PCA l=8 values were 0.
2. The magnitudes of diagonal values are 1 as standard deviation is 1 and correlation is also 1 so covariance is also 1.
3. For the real data there is no such trend of increasing or decreasing values but for reduced that after standardization there is a decreasing trend in the data.