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

Student's Name: Akshar Singh **Mobile No: 7428357700**

Roll Number: B20147 **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. Outlier correction was needed as outliers indicate erroneous data which if not corrected makes data incorrect for prediction.
2. The method used was to replace the values beyond upper and lower whisker with the median since the median of the data is not greatly influenced by outliers.
3. Normalization scales each and every data value in an attribute between a particular range which in this case is [5,12].

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	3.783	3.271	0	1
2	plas	121.656	30.438	0	1
3	pres (in mm Hg)	72.197	11.147	0	1
4	skin (in mm)	20.438	15.699	0	1
5	test (in mu U/mL)	60.919	77.636	0	1
6	BMI (in kg/m^2)	32.199	6.411	0	1

7	pedi	0.428	0.245	0	1
8	Age (in years)	32.760	11.055	0	1

Inferences:

1. The attributes of the transformed data is rescaled such that it has 0 mean and unit variance i.e. standard deviation of 1.

2 a.

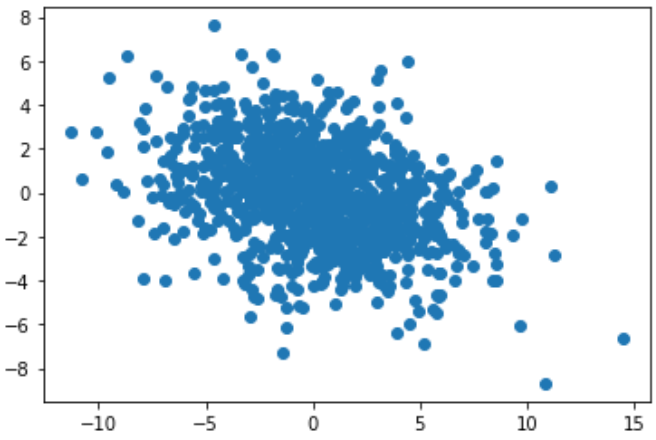

Figure 1 Scatter plot of 2D synthetic data of 1000 samples

Inferences:

1. The attribute 1 and attribute 2 seems to be negatively uncorrelated as with increase in values in one attribute other decreases.
2. The data seems to be denser when the x-value is between -5 and 5 and y-value is between -4 and 4.

b.

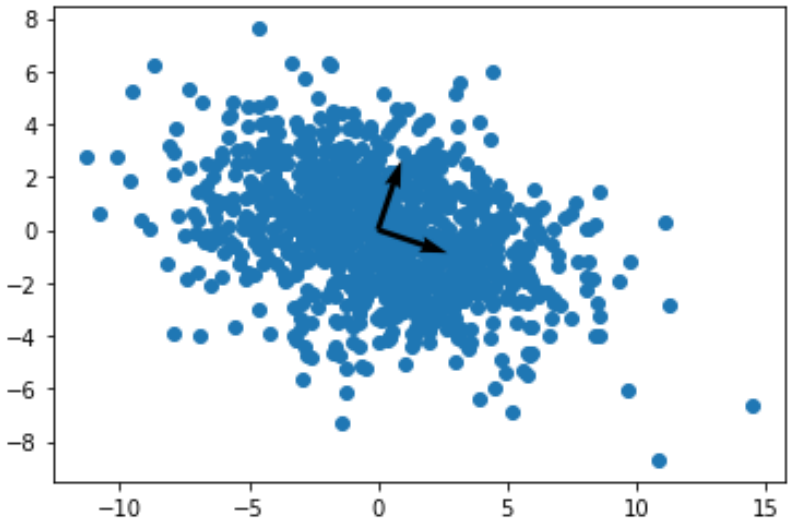

Figure 2 Plot of 2D synthetic data and Eigen directions

Inferences:

1. The data seems to be more spread in that direction where Eigenvalue is more.
2. The points are much more denser near the intersection of Eigen axis than away from it.

c.

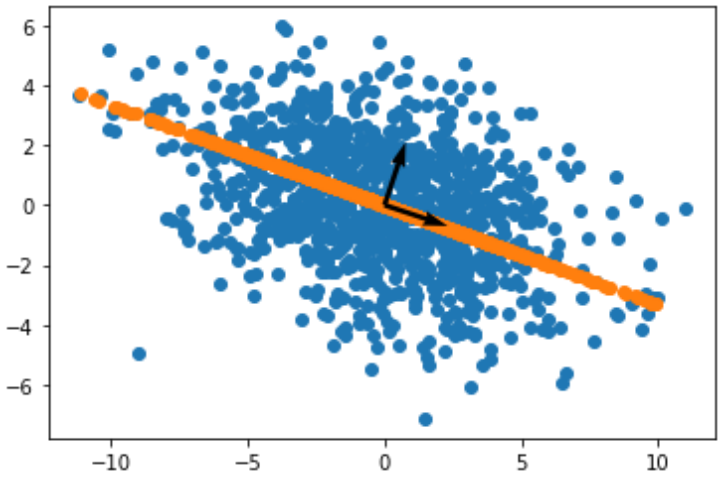

Figure 3 Projected Eigen directions onto the scatter plot with 1st Eigen direction highlighted

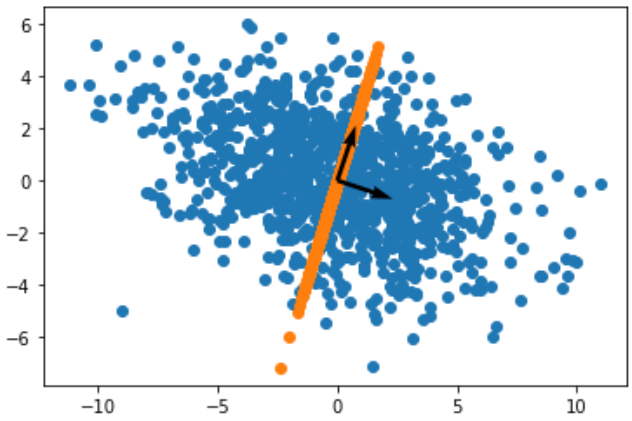

Figure 4 Projected Eigen directions onto the scatter plot with 2nd Eigen direction highlighted

Inferences:

1. The eigen value for the first eigen direction is 14 whereas for the other eigen direction is 4.
2. The data points seem to be denser at the intersection points of the eigen axis. The data is more spread along the Eigen direction where the eigen value is more.

d. Reconstruction error = 0

Inferences:

1. Reconstructed data is said to be loseless when the reconstruction error is close to 0.

3 a.

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

Direction	Variance	Eigenvalue
1	1.992	1.992
2	1.853	1.853

Inferences:

1. The variance is same as the eigen value for both the directions.

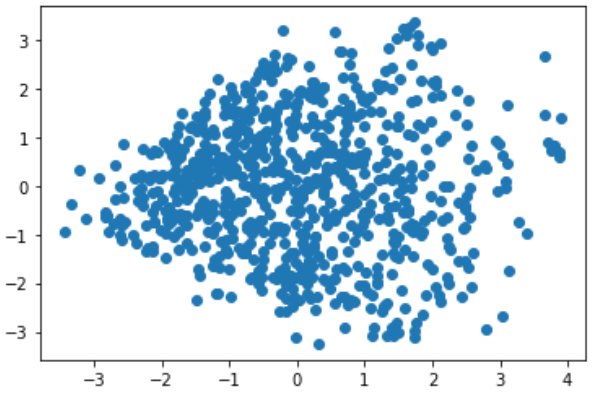

Figure 5 Plot of data after dimensionality reduction

Inferences:

1. The two attributes obtained after dimensionality reduction seems to be uncorrelated.

b.

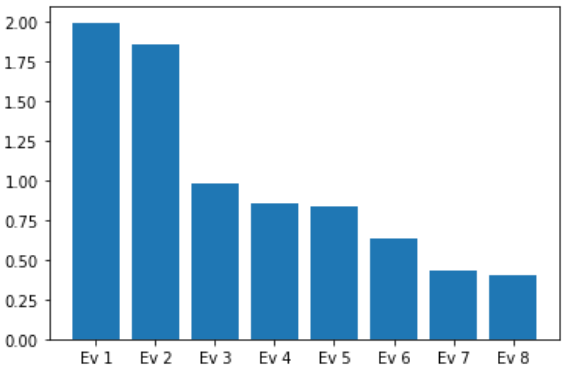

Figure 6 Plot of Eigenvalues in descending order

Inferences:

1. Infer whether the subsequent Eigenvalues decrease gradually or rapidly
2. There is a gradual decrease between the Eigen value1 and Eigen value 2 and from Eigen value 3 onwards but there is a rapid decrease between Eigen value 2 and Eigen value 3.
3. There is a substantial decrease between Eigen value 2 and Eigen value3.

c.

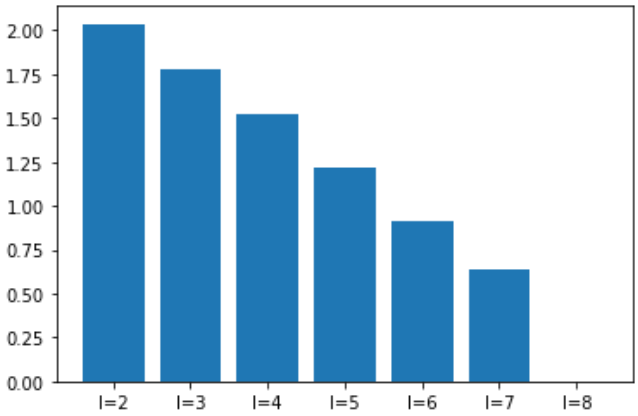

Figure 7 Line plot to demonstrate reconstruction error vs. components

Inferences:

1. Higher the magnitude of reconstruction error lower the quality of reconstruction data.

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

	x1	x2
x1	1.992	0
x2	0	1.853

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

	x1	x2	x3
x1	1.992	0	0
x2	0	1.853	0
x3	0	0	0.982

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

	x1	x2	x3	x4
x1	1.992	0	0	0
x2	0	1.853	0	0
x3	0	0	0.982	0
x4	0	0	0	0.858

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

	x1	x2	x3	x4	x5
x1	1.992	0	0	0	0
x2	0	1.853	0	0	0
x3	0	0	0.982	0	0
x4	0	0	0	0.858	0
x5	0	0	0	0	0.839

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

	x1	x2	x3	x4	x5	x6
x1	1.992	0	0	0	0	0
x2	0	1.853	0	0	0	0
x3	0	0	0.982	0	0	0
x4	0	0	0	0.858	0	0
x5	0	0	0	0	0.839	0
x6	0	0	0	0	0	0.636

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

	x1	x2	x3	x4	x5	x6	x7
x1	1.992	0	0	0	0	0	0
x2	0	1.853	0	0	0	0	0
x3	0	0	0.982	0	0	0	0
x4	0	0	0	0.858	0	0	0
x5	0	0	0	0	0.839	0	0
x6	0	0	0	0	0	0.636	0
x7	0	0	0	0	0	0	0.434

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

	x1	x2	x3	x4	x5	x6	x7	x8
x1	1.992	0	0	0	0	0	0	0
x2	0	1.853	0	0	0	0	0	0
x3	0	0	0.982	0	0	0	0	0
x4	0	0	0	0.858	0	0	0	0
x5	0	0	0	0	0.839	0	0	0
x6	0	0	0	0	0	0.636	0	0
x7	0	0	0	0	0	0	0.434	0
x8	0	0	0	0	0	0	0	0.405

Inferences:

1. The off-diagonal elements are zero since attributes are uncorrelated with each other.
2. The diagonal elements are not zero as it tells the variance of the particular attribute whereas the off-diagonal elements are zero since attributes are uncorrelated with each other.
3. The diagonal elements are not zero as it tells the variance of the particular attribute.
4. The diagonal values are decreasing sequentially.
5. There is a decrease because of the descending eigen values.
6. The first component captures the data variations the best since it's the highest.
7. From the value of diagonal elements, the first two components shall give the optimum reconstruction along with dimensionality reduction.
8. The magnitude of the 1st diagonal element in each of the covariance matrices is same because the same eigen vector having highest value of eigen vector has been used in each of these.
9. The magnitude of the 2nd diagonal element in each of the covariance matrices is same because the same eigen vector having 2nd highest value of eigen vector has been used in each of these.
10. The 3rd, 4th, 5th, 6th, and 7th diagonal elements across covariance matrices are also same .

d.

Table 11 Covariance matrix for original data

	pregs	plas	pres	skin	test	BMI	pedi	Age
pregs	1	0.118	0.208	-0.097	-0.108	0.0283	0.004	0.561
plas	0.118	1	0.205	0.060	0.179	0.228	0.0817	0.274
pres (in mm Hg)	0.209	0.205	1	0.026	-0.051	0.272	0.022	0.326
skin (in mm)	-0.097	0.060	0.025	1	0.472	0.374	0.153	-0.101
test (in mu U/mL)	-0.108	0.179	-0.051	0.473	1	0.172	0.199	-0.074
BMI (in kg/m^2)	0.0283	0.228	0.272	0.374	0.172	1	0.124	0.078
pedi	0.004	0.082	0.022	0.158	0.199	0.124	1	0.036
Age (in years)	0.561	0.274	0.326	-0.101	-0.074	0.078	0.036	1

Inferences:

1. The off-diagonal values are not zero as the attributes are correlated with each other which was not the case with PCA l=8.
2. The diagonal elements don't have the same values as the PCA l=8.
3. There is no such trend as obtained in the dimensional reduced data.