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

Student's Name: Abhay Vijayvargiya **Mobile No: 6377967485**

Roll Number: B20176 **Branch:DSE**

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	318.0	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 are unusual values in the dataset. If not removed it can produce poor transformation of data and can decrease the statistical analysis of the data.
2. Parameter used for outlier deletion is Inter Quartile Range i.e., remove any data point that is above (Q3 + IQR) or below (Q1 – IQR).
3. Before normalization each attribute has its own scale of data and this can cause poor analysis of data. After normalization all attributes come under same scale so that we can analyze data better

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	3.801	3.276	0.0	1
2	plas	120.95	31.77	0.0	1
3	pres (in mm Hg)	68.48	18.66	0.0	1
4	skin (in mm)	20.42	15.705	0.0	1
5	test (in mu U/mL)	59.82	78.23	0.0	1
6	BMI (in kg/m^2)	31.61	7.43	0.0	1

7	pedi	0.672	1.25	0.0	1
8	Age (in years)	32.49	11.41	0.0	1

Inferences:

1. Before standardization different attributes have different standardization. After standardization all the attributes have 0 mean and 1 standard deviation.
2. Standardization puts all the attributes in same scale and allows us to compare data between different types of variables.

2 a.

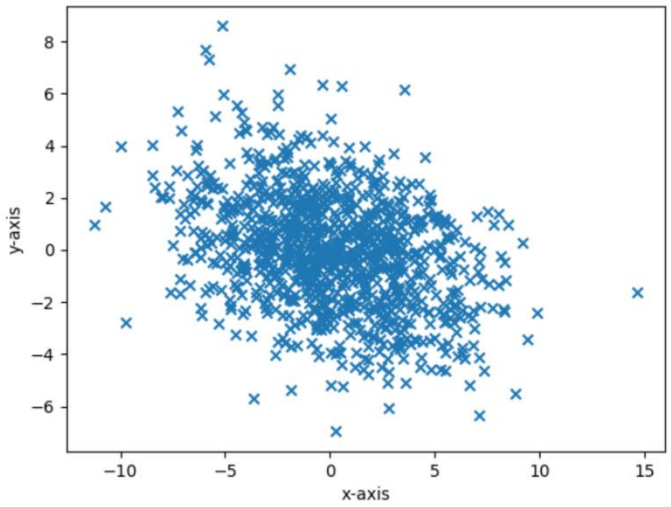

Figure 1 Scatter plot of 2D synthetic data of 1000 samples

Inferences:

1. Attribute 1 is negatively corelated with attribute 2
2. Density is more about (x = 0, y = 0), since we have generated the data samples with zero mean.

b.

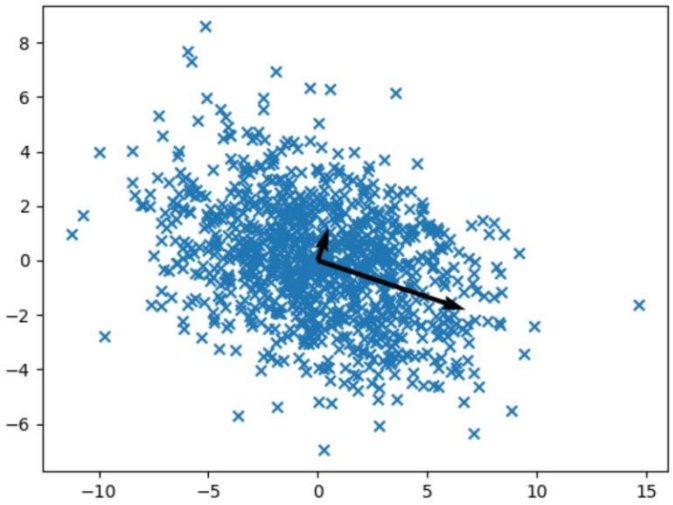

Figure 2 Plot of 2D synthetic data and Eigen directions

Inferences:

1. Eigen values indicates the spread of the data in the direction corresponding to its eigen vector. Greater the magnitude of eigenvalue greater the spread in its eigen vector direction.
2. Density is greater in the intersection of the eigenvalues and gradually moves in the direction of eigen vector which have greater magnitude of eigen value.

c.

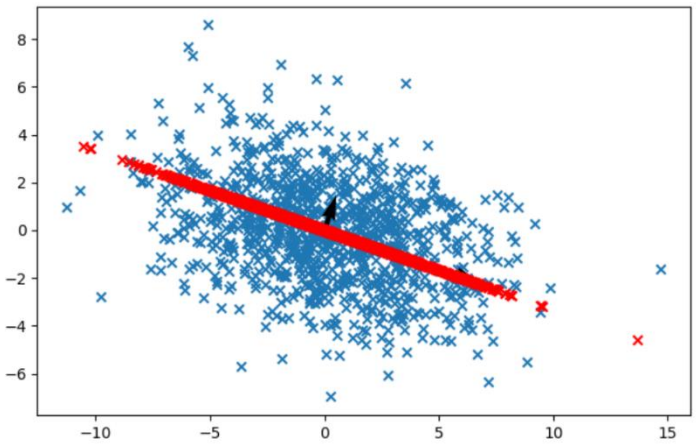

Figure 3 Projected Eigen directions onto the scatter plot with 1st Eigen direction highlighted

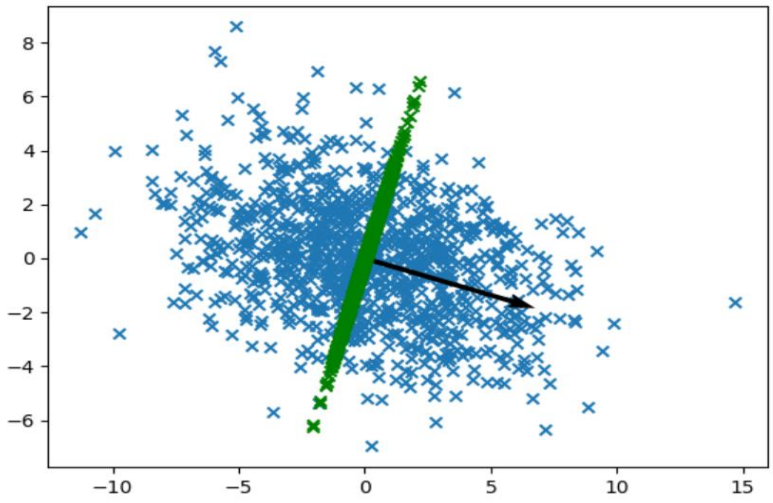

Figure 4 Projected Eigen directions onto the scatter plot with 2nd Eigen direction highlighted

Inferences:

1. Magnitude of 1st eigenvalue = 14.0, Magnitude of 2nd eigenvalue = 4.0.
2. Variance of the data along the eigen direction having greater eigen value is greater than along the direction having less eigen value.
3. Both eigen value and the variance of the projected data tells the spread the data along that eigen direction

d. Reconstruction error = 0.000 (in the range e^{-16})

Inferences:

1. If the value of reconstruction error is high then the data obtained after reconstruction does not represent original data appreciatory.
2. Reconstruction error can be minimized by using more principal components.

3 a.

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

Direction	Variance	Eigenvalue
1	1.992	1.993
2	1.853	1.854

Inferences:

1. Both the variance of projected data and eigenvalues are approximately equal since both the parameters tell the spread of the data in particular direction.

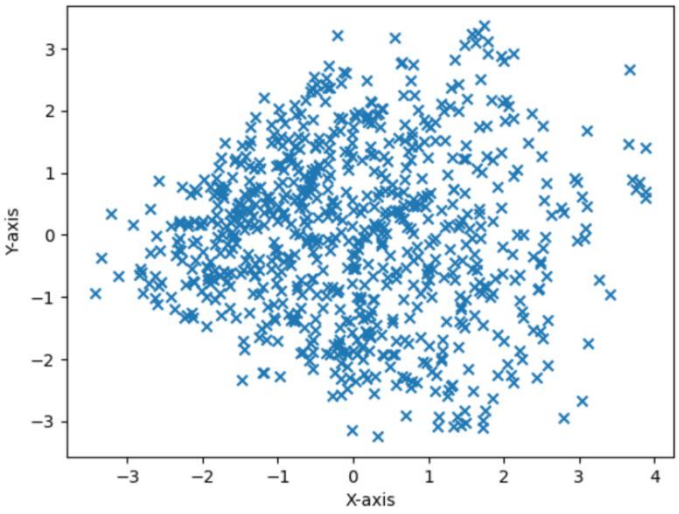

Figure 5 Plot of data after dimensionality reduction

Inferences:

1. There is approximately zero correlation among the two attributes. After performing PCA we get the attributes having zero correlation.

b.

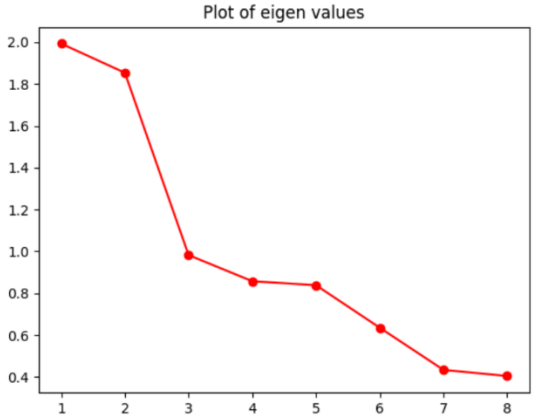

Figure 6 Plot of Eigenvalues in descending order

Inferences:

1. The subsequent eigen values decrease rapidly.
2. From eigen value = 1.853 the rate of decrease increases suddenly.
3. The first two eigen values represent almost all the data in their respective direction. Other eigen values contribute very less in their eigen directions.

c.

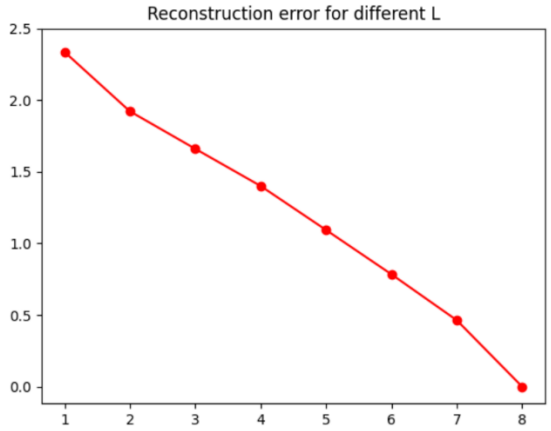

Figure 7 Line plot to demonstrate reconstruction error vs. components

Inferences:

1. If the value of reconstruction error is high then the data obtained after reconstruction does not represent original data appreciatory.
2. Reconstruction error can be minimized by using more principal components as can be seen clearly in above plot.

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

	x1	x2
x1	1.992	0.0
x2	0.0	1.853

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

	x1	x2	x3
x1	1.992	0.0	0.0
x2	0.0	1.853	0.0
x3	0.0	0.0	0.982

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

	x1	x2	x3	x4
x1	1.992	0.0	0.0	0.0
x2	0.0	1.853	0.0	0.0
x3	0.0	0.0	0.982	0.0
x4	0.0	0.0	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	0.0	0.0
x2	0.0	1.853	0.0	0.0	0.0
x3	0.0	0.0	0.982	0.0	0.0
x4	0.0	0.0	0.0	0.858	0.0
x5	0.0	0.0	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	0.0	0.0	0.0
x2	0.0	1.853	0.0	0.0	0.0	0.0
x3	0.0	0.0	0.982	0.0	0.0	0.0
x4	0.0	0.0	0.0	0.858	0.0	0.0
x5	0.0	0.0	0.0	0.0	0.839	0.0
x6	0.0	0.0	0.0	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	0.0	0.0	0.0
x2	0.0	1.853	0.0	0.0	0.0	0.0	0.0
x3	0.0	0.0	0.982	0.0	0.0	0.0	0.0
x4	0.0	0.0	0.0	0.858	0.0	0.0	0.0
x5	0.0	0.0	0.0	0.0	0.839	0.0	0.0
x6	0.0	0.0	0.0	0.0	0.0	0.636	0.0
x7	0.0	0.0	0.0	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	0.0	0.0	0.0	0.0
x2	0.0	1.853	0.0	0.0	0.0	0.0	0.0	0.0
x3	0.0	0.0	0.982	0.0	0.0	0.0	0.0	0.0
x4	0.0	0.0	0.0	0.858	0.0	0.0	0.0	0.0
x5	0.0	0.0	0.0	0.0	0.839	0.0	0.0	0.0
x6	0.0	0.0	0.0	0.0	0.0	0.636	0.0	0.0
x7	0.0	0.0	0.0	0.0	0.0	0.0	0.434	0.0
x8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.405

Inferences:

1. The off-diagonal values are 0. After performing PCA we get the uncorrelated features or attributes.
2. The off-diagonal values are 0 and the diagonal values contain the variance of that column attribute. Since each attribute is uncorrelated the diagonal values represent the spread in that eigen direction.
3. It is decreased rapidly after 2nd diagonal element.
4. The reason for decrease is the less spread along that eigen direction.
5. The first diagonal element gives best variation of the data in its direction
6. 5 components can give optimum reconstruction with dimensional reduction.
7. The first diagonal entry is same in all the matrices since it's the highest eigenvalue and thus contained in any no. of components.
8. The 3rd, 4th,5th,6th and 7th diagonal entries are close yet decreasing. However the rate of decrease is decreasing.

d.

Table 11 Covariance matrix for original data

	pregs	plas	pres	skin	test	BMI	pedi	Age
pregs	1.00	0.118	0.209	-0.097	-0.108	0.028	0.005	0.561
plas	0.118	1.00	0.205	0.06	0.18	0.228	0.082	0.274
pres (in mm Hg)	0.209	0.205	1.00	0.026	-0.051	0.272	0.022	0.326
skin (in mm)	-0.097	0.06	0.026	1.00	0.473	0.374	0.153	-0.101
test (in mu U/mL)	-0.108	0.18	-0.051	0.473	1.00	0.172	0.199	-0.074
BMI (in kg/m^2)	0.028	0.228	0.272	0.374	0.172	1.00	0.124	0.078
pedi	0.005	0.082	0.022	0.153	0.199	0.124	1.00	0.036
Age (in years)	0.561	0.274	0.326	-0.101	-0.074	0.078	0.036	1.00

Inferences:

1. In the above matrix the off-diagonal entries represent the covariance of the attributes and diagonal entries represents the variance of the attribute which is 1 since our data is standardized.
2. All the diagonal values are 1 since our data is standardized.
3. There is no decrease in the diagonal entries unlike the covariance matrix which is obtained after dimensionality reduction.