IC 272: DATA SCIENCE - III
LAB ASSIGNMENT – III
Attribute Normalization, Standardization and Dimension Reduction of Data

Student's Name: Rashmi **Mobile No:** 7015331137

Roll Number: B19218 **Branch:** Engineering Physics

1 a.

Table 1 Minimum and Maximum Attribute Values Before and After Min-Max Normalization

S. No.	Attribute	Before Min-Max Normalization		After Min-Max Normalization	
		Minimum	Maximum	Minimum	Maximum
1	Temperature (in °C)	10.085	31.375	3.000	9.000
2	Humidity (in $g.m^{-3}$)	34.206	99.720	3.000	9.000
3	Pressure (in mb)	992.655	1037.604	3.000	9.000
4	Rain (in ml)	0.000	2470.500	3.000	9.000
5	Lightavgw/o0 (in lux)	0.000	10565.352	3.000	9.000
6	Lightmax (in lux)	2259.000	54612.000	3.000	9.000
7	Moisture (in %)	0.000	100.000	3.000	9.000

Inferences:

1. 'Rain' and 'Pressure' attributes have maximum number of outliers whereas 'Lightmax' and 'Moisture' have zero outliers.
2. After min-max normalization, data points are linearly transformed having range of 3-9.

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	Temperature (in °C)	21.370	4.125	-0.000	1.000
2	Humidity (in $g.m^{-3}$)	83.992	17.566	-0.000	1.000
3	Pressure (in mb)	1014.761	6.121	-0.000	1.000
4	Rain (in ml)	168.400	399.689	0.000	1.000
5	Lightavgw/o0 (in lux)	2197.392	2220.820	0.000	1.000
6	Lightmax (in lux)	21788.623	22064.993	0.000	1.000
7	Moisture (in %)	32.386	33.653	0.000	1.000

Inferences:

1. In standardization, the distribution of data points as gaussian distribution is linearly transformed it into standard gaussian distribution with mean μ=0 and σ=1.

2 a.

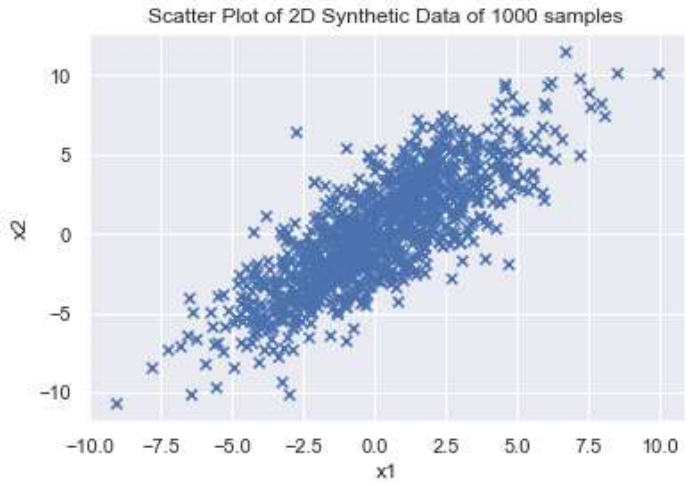

Figure 1 Scatter Plot of 2D Synthetic Data of 1000 samples

Inferences:

1. Attributes are positively correlated.
2. Data points are highly dense around the mean (0,0). We can see that as distance from mean is increasing, their density is decreasing.
3. Shape of the distribution is elliptical.

b.

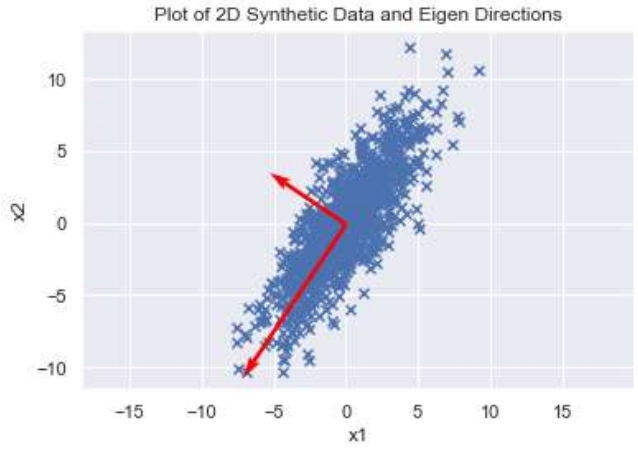

Figure 2 Plot of 2D Synthetic Data and Eigen Directions

Inferences:

1. The data is highly spread in that eigen direction which corresponds to high eigen value.
2. Eigen axis intersects at origin. Here data points are highly dense. As we go away from it, density decreases.

c.

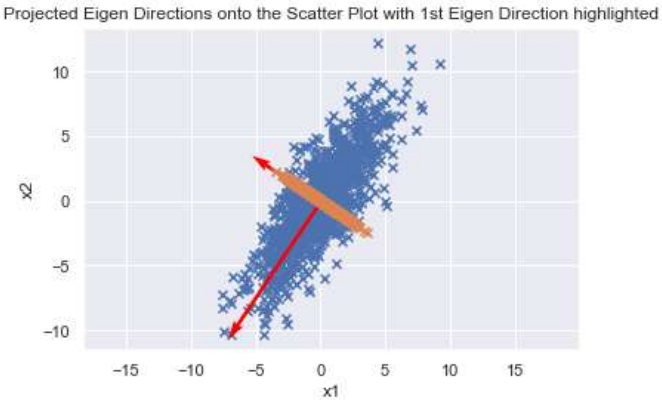

Figure 3 Projected Eigen Directions onto the Scatter Plot with 1st Eigen Direction highlighted

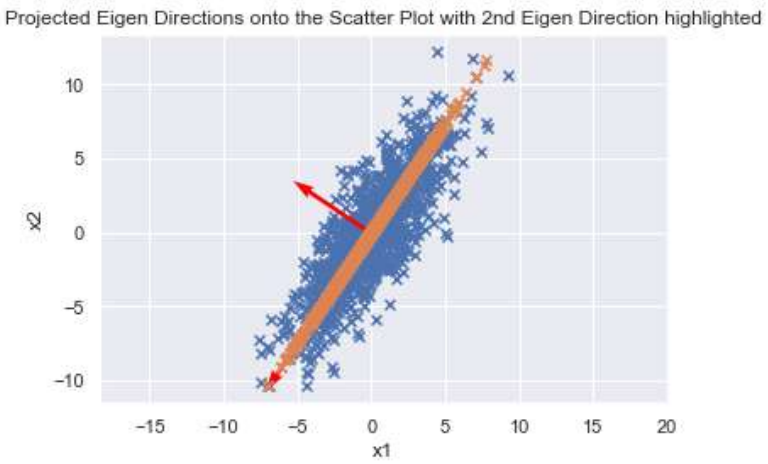

Figure 4 Projected Eigen Directions onto the Scatter Plot with 2nd Eigen Direction highlighted

Inferences:

1. Fig-3 corresponds to small eigen value and its projections lie in small range. On other hand, projection for Fig-4 lie on larger range, as it corresponds to larger eigen value.
2. Variance of projections is large on that eigen direction which has large eigen value.

d. Reconstruction Error = 1.009886618774658e-16 (approx. 0)

Inferences:

1. Here, l=d=2 i.e. lower dimension and actual dimension are same. Therefore, error is almost zero in this case.
2. As dimension of data decreases, the quality of reconstruction decreases.

3 a.

Table 3 Variance and Eigen Values of the projected data along the two directions

Direction	Variance	Eigen Value
1	2.202	2.202
2	1.421	1.421

Inferences:

1. Variances of the projected data are almost same as their respective eigen values.

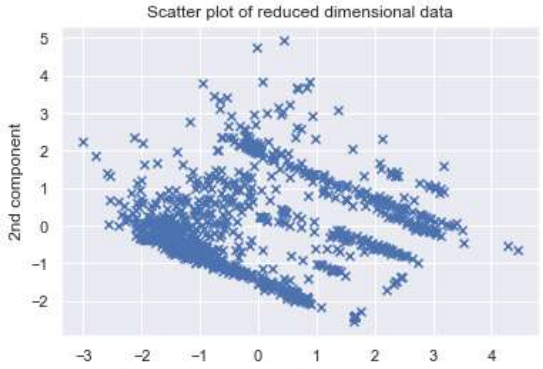

Figure 5 Plot of Landslide Data after dimensionality reduction

Inferences:

1. Data points are highly dispersed after dimensionality reduction.
2. It is comparably high dense in the origin region.

b.

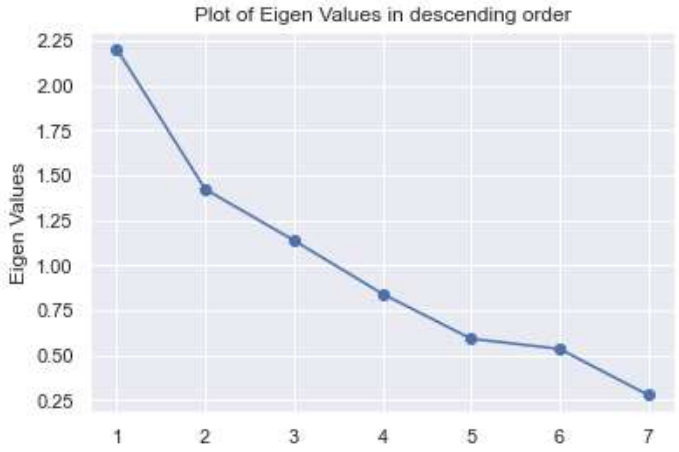

Figure 6 Plot of Eigen Values in descending order

Inferences:

1. The magnitude of eigen values is decreasing gradually after second eigen value. Initial change between 1st and 2nd eigen value is Sharp.
2. Rate of decrease has changed after second eigen values.
3. Therefore, use l=2 for dimension reduction, it will conserve most of the data.

c.

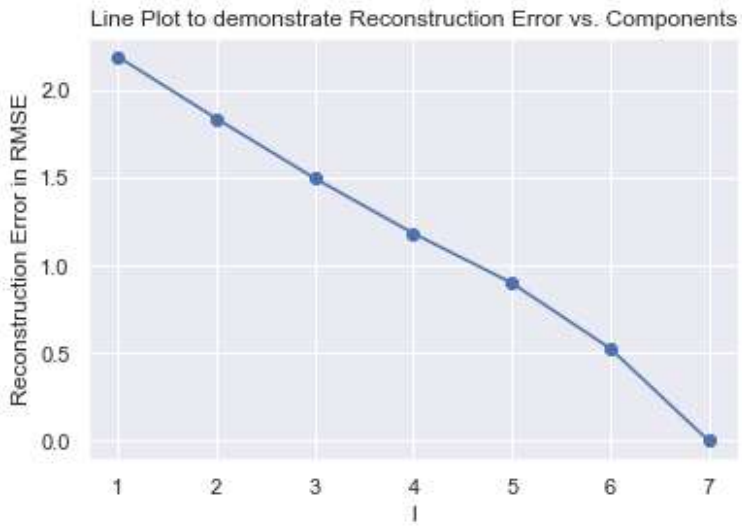

Figure 7 Line Plot to demonstrate Reconstruction Error vs. Components

Inferences:

1. Reconstruction error increases as we decrease the lower dimension.
2. l=3 or 4 is the best compensation between data loss and dimensionality reduction.