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

**Student's Name: Tarun Singla** **Mobile No: 8872526396**

**Roll Number: b19198** **Branch:EE**

**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 | 100.090 | 3.000 | 9.000 |

**Inferences:**

1. After Replacing outlier with median (calculated from non-outliers) outliers are removed
2. After min-max normalization Data points are linearly transformed in the range from 3-9
3. Behavior of the Data points are not changed

**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) | 31.376 | 4.125 | 0.00 | 1.0 |
| 2 | Humidity (in g.m$^{-3}$ ) | 83.991 | 17.565 | 0.00 | 1.0 |
| 3 | Pressure (in mb) | 1014.793 | 6.121 | 0.00 | 1.0 |
| 4 | Rain (in ml) | 171.467 | 399.550 | 0.00 | 1.0 |
| 5 | Lightavgw/o0 (in lux) | 2237.892 | 2206.423 | 0.00 | 1.0 |
| 6 | Lightmax (in lux) | 21788.620 | 22064.993 | 0.00 | 1.0 |
| 7 | Moisture (in %) | 32.386 | 33.653 | 0.00 | 1.0 |

**Inferences:**

1. In Standardization we assume points to be a Gaussian Distribution and than linearly transformed into standard Gaussian Distribution
2. May be the behavior of data may have changed

**2 a.**

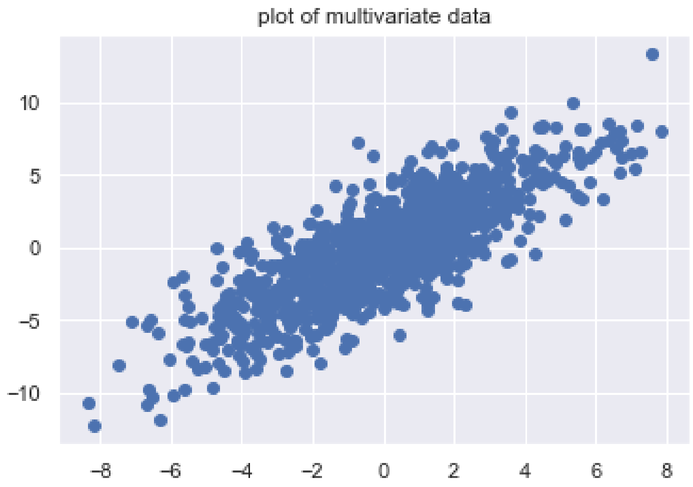

**Figure 1 Scatter Plot of 2D Synthetic Data of 1000 samples**

**Inferences:**

1. With the help of plot, we can say that attributes are positively correlated.
2. Data points are highly dense around the mean (0,0). As distance from mean is increasing, their density is decreasing
3. Shape of the distribution is elliptical.

**b.**

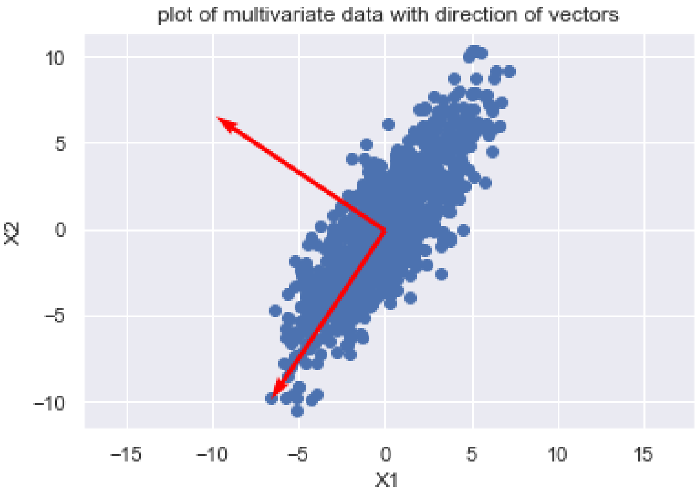

**Figure 2 Plot of 2D Synthetic Data and Eigen Directions**

**Inferences:**

1. The data is highly spread in the eigen direction of eigen value 18.1691. That means data point is highly spread in that eigen direction which corresponds to high eigen value.
2. Eigen axis intersects at origin. Where data points are highly dense. As we go away from it, density decreases.

**c.**

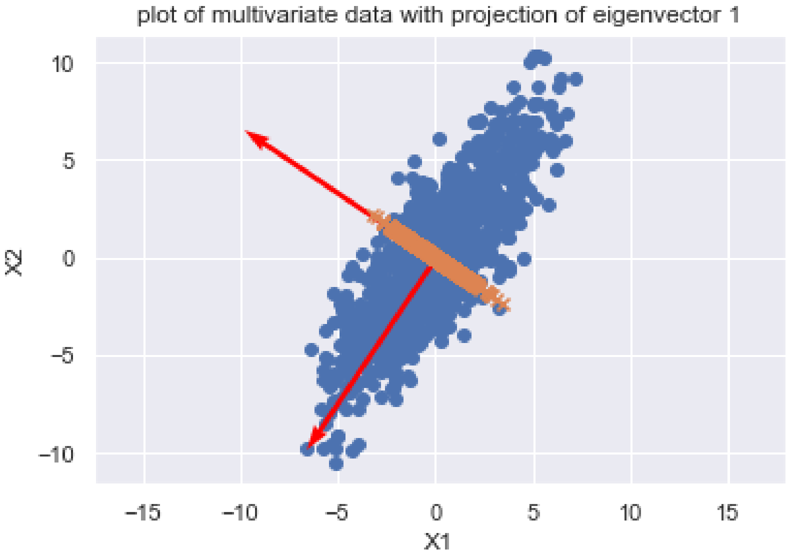

**Figure 3 Projected Eigen Directions onto the Scatter Plot with 1st Eigen Direction highlighted**

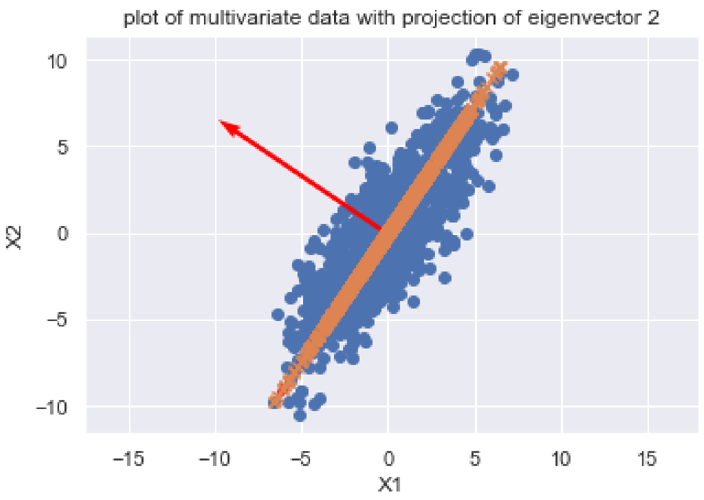

**Figure 4 Projected Eigen Directions onto the Scatter Plot with 2nd Eigen Direction highlighted**

**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, because it corresponds to larger eigen value..
2. . Variance of projections is also large on that eigen direction which has large eigen value i.e. eigen value=18.1691.

**d. Reconstruction** Error = (report only up to three decimal places) = 0.000 (almost zero)

**Inferences:**

1. 1. In this case, l=d=2 (lower dimension and actual dimension are same). So, error is almost zero. As dimension 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 | 2.223 | 2.224 |
| 2 | 1.429 | 1.430 |

**Inferences:**

1. Variances of the projected data are almost same as their respective eigen values.
2. High Eigen value means, more information holds in that eigen direction

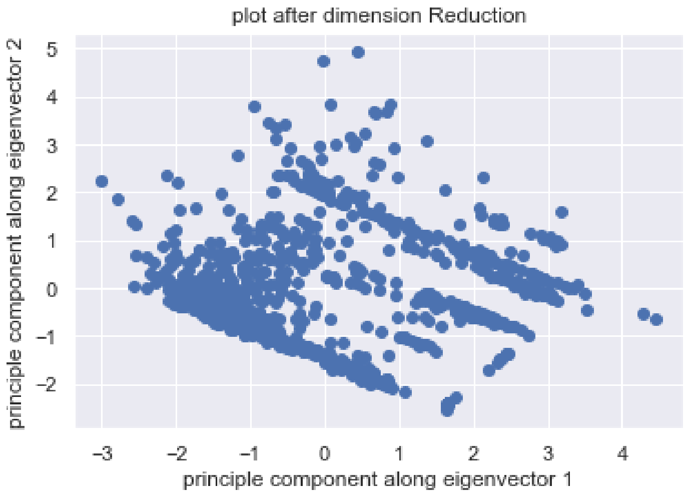

**Figure 5 Plot of Landslide Data after dimensionality reduction**

**Inferences:**

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

**b.**

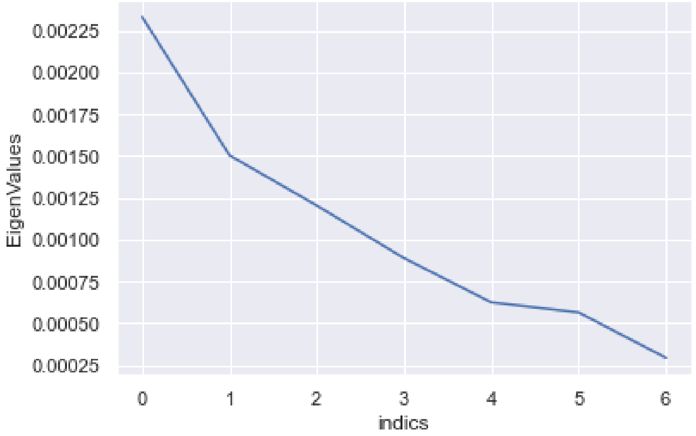

**Figure 6 Plot of Eigen Values in descending order**

**Inferences:**

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

**c.**

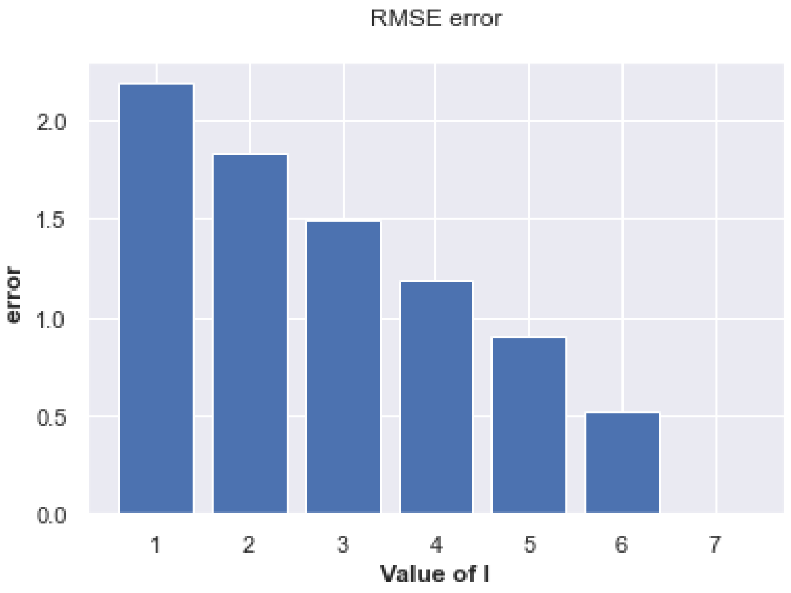

**Figure 7 Line Plot to demonstrate Reconstruction Error vs. Components**

**Inferences:**

1. Reconstruction error increases as we decrease the lower dimension.
2. When l = d RMSE error approaches to zero.

**Guidelines for Report (Delete this while you submit the report):**

- **The plot/graph/figure/table should be centre justified with sequence number and caption.**
- **Inferences should be written as a numbered list.**
- **Use specific and technical terms to write inferences.**

- **Values observed/calculated should be rounded off to three decimal places.**
- **The quantities which have units should be written with units.**