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

Student's Name: Khushi Baghel **Mobile No:** 7417130808

Roll Number: B20249 **Branch:** Engineering Physics

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	0.0	13.0	5.0	12.0
2	plas	44.0	199.0	5.0	12.0
3	pres (in mm Hg)	38.0	106.0	5.0	12.0
4	skin (in mm)	0.0	63.0	5.0	12.0
5	test (in mu U/mL)	0.0	318.0	5.0	12.0
6	BMI (in kg/m^2)	18.20	50.0	5.0	12.0
7	pedi	0.078	1.191	5.0	12.0
8	Age (in years)	21.0	66.0	5.0	12.0

Inferences:

1. Outliers increase the variability of the data so in order to increase the viability of our analysis we remove or replace the outliers.
2. As data is not normally distributed we are replacing outliers with median rather than mean.
3. Before normalization, the attributes having bigger values use to overpower the attribute with smaller values. So, the analysis will be more partial. Now after normalization each value is now between 5 and 12, implying that they will be given equal weightage in the analysis.

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	3.782	3.270	0	1.0
2	plas	121.65	30.438	0	1.0
3	pres (in mm Hg)	72.19	11.146	0	1.0
4	skin (in mm)	20.43	15.698	0	1.0
5	test (in mu U/mL)	60.91	77.635	0	1.0
6	BMI (in kg/m^2)	32.19	6.410	0	1.0

7	pedi	0.427	0.245	0	1.0
8	Age (in years)	32.760	11.055	0	1.0

Inferences:

1. Before standardization, the attributes having larger values use to overpower the attribute with smaller values. So, the analysis will be more partial. Now after standardization, every value has a common mean of 0 and variance 1.
2. Standardization is better than normalization since there is no input range, therefore there is no out of bound error.

2 a.

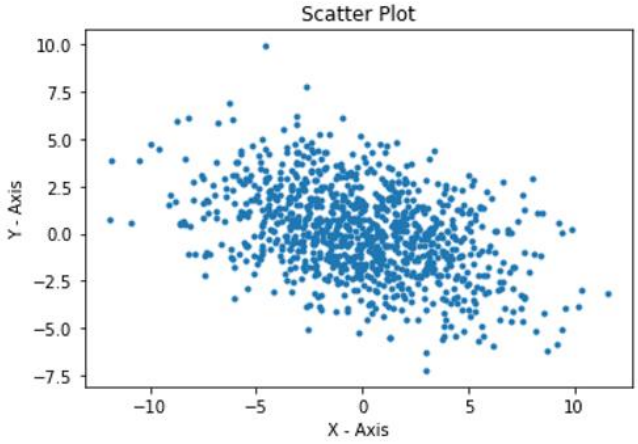

Figure 1 Scatter plot of 2D synthetic data of 1000 samples

Inferences:

1. Attribute 2 is negatively correlated with the Attribute 1 according to the graph.
2. The distribution of both attributes appears to be symmetric based on the density of the graph. The mean of both attributes is close to zero.

b.

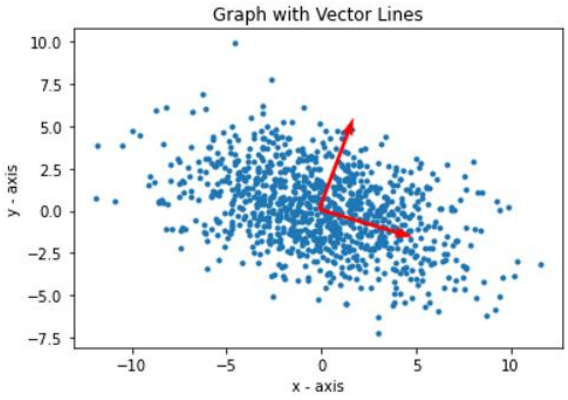

Figure 2 Plot of 2D synthetic data and Eigen directions

Inferences:

1. We may deduce from the above plot that the dots are more spread out along the first eigenvector, therefore the eigenvalue for the first vector is greater.
2. The density of points near the axis intersection is quite high, and it gradually decreases as the dispersion rises. In other words, as one moves out from the center, the number of points decreases.

c.

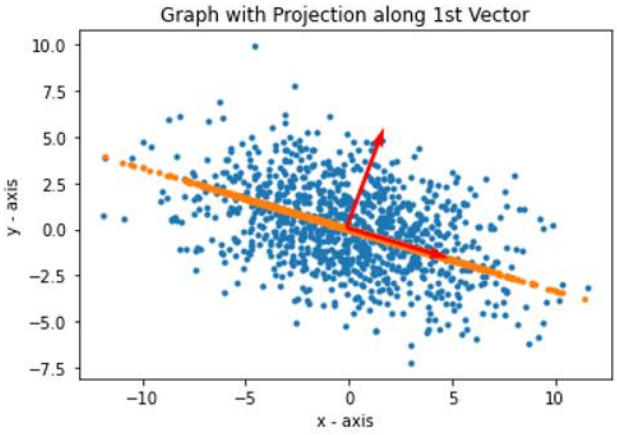

Figure 3 Projected Eigen directions onto the scatter plot with 1st Eigen direction highlighted

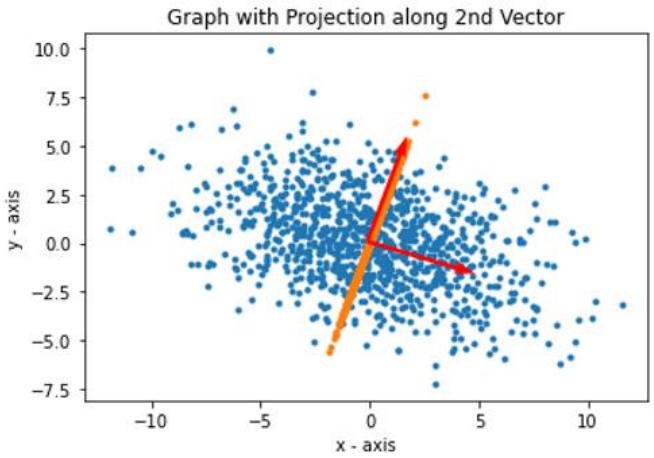

Figure 4 Projected Eigen directions onto the scatter plot with 2nd Eigen direction highlighted

Inferences:

1. The Eigen Values are: 1st Vector : 14, 2nd Vector : 4.
2. The data is much more spread along first eigen vector as compared to data along the axes of second eigen vector. This can be related to their corresponding eigen values. Higher the eigen values, more will be the spread.

d. Reconstruction error = 0 (round to three decimal places).

Inferences:

1. More the reconstruction error, more loss in the nature of data. So, the reconstruction error must not be very high.

3 a.

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

Direction	Variance	Eigenvalue
1	1.9924	1.9924
2	1.8534	1.8534

Inferences:

1. The eigenvalues in the covariance matrix of modified data are the same as the variances in the table above.
2. Higher the Eigen Vector values, greater the variance along that vector, and therefore greater the strength along that vector. As a result, we may claim that data will be more evenly distributed over the first Eigen vector.

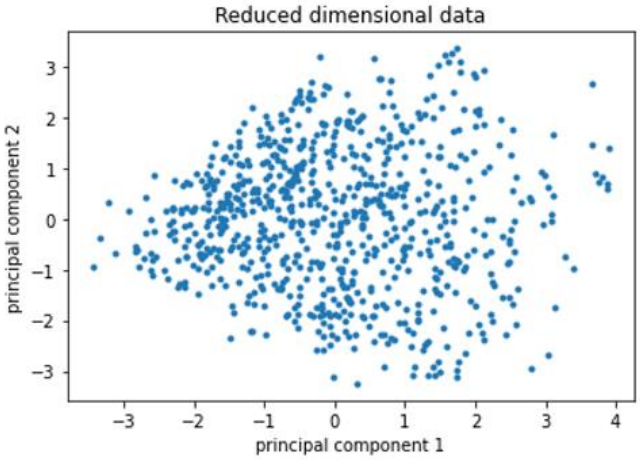

Figure 5 Plot of data after dimensionality reduction

Inferences:

1. The correlation between two attributes will be zero as we are projecting the data on orthonormal eigen vector during pca reduction. It can also be verified from the plot.

b.

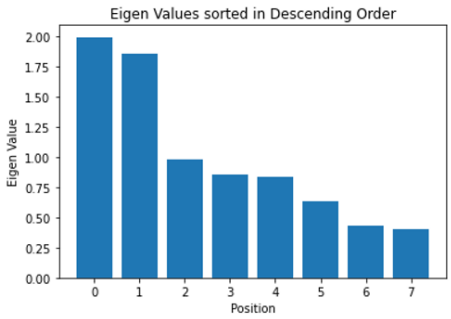

Figure 6 Plot of Eigenvalues in descending order

Inferences:

1. It drops significantly from second Eigen Value to third, and then it gradually decreases.

2. At the second Eigen Value, it gets the highest drop.

c.

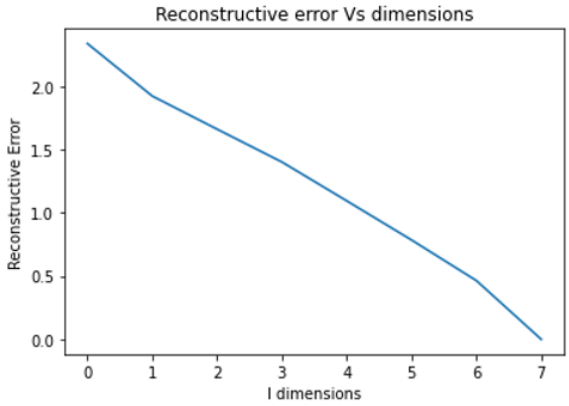

Figure 7 Line plot to demonstrate reconstruction error vs. components

Inferences:

1. More the magnitude of reconstruction error, lesser the quality of reconstructions. As we can see the Euclidean distance increases, as we keep reducing the dimensions.
2. At l = 8, the reconstruction error is almost negligible.

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. Off-diagonal elements are zero, and all of the reduced attributes are independent of one another, therefore the covariance is zero.
2. Off-diagonal elements are 0, diagonal elements are the corresponding variance of the attributes.
3. The diagonal values decrease on increasing the dimension, the trend for eigenvalue decrease when we move from x1 to x8.
4. Justify the reason for the increase/ decrease.
5. As the diagonal element of X1 is maximum therefore it capture variance better than other attributes.
6. We can observe a significant dip in variance after X2 therefore, 2 components gives the optimum reconstruction along with dimensionality reduction.
7. Yes the value of top element is same in all the cases, on increasing the number of attributes var(X1) remains the same.
8. Yes the value of 2nd diagonal element is same in all the cases, on increasing the number of attributes var(X2) remains the same.
9. Yes the values of diagonal elements is same in all the cases, on increasing the number of attributes var(Xi) remains the same.

d.

Table 11 Covariance matrix for original data

	pregs	plas	pres	skin	test	BMI	pedi	Age
pregs	1	0.118	0.209	-0.097	-0.108	0.028	0.005	0.561
plas	0.118	1	0.205	0.060	0.180	0.228	0.082	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	0.026	1	0.473	0.374	0.153	-0.101
test (in mu U/mL)	-0.108	0.180	-0.051	0.473	1	0.172	0.199	-0.074
BMI (in kg/m^2)	0.028	0.228	0.272	0.374	0.172	1	0.124	0.078

pedi	0.005	0.082	0.022	0.153	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 elements for original and reconstructed(l=8) are symmetrical whereas the off-diagonal elements for reduced(l=8) are 0.
2. The diagonal elements for original and reconstructed(l=8) are 1 whereas the diagonal elements for reduced(l=8) are var(Xi).
3. After reconstructing the reduced data(l = 8) the dataset obtained was similar to the original dataset therefore the covariance matrix was also similar but if we compute the cov matrix for reduced data(l = 8) without reconstruction we will get the variance of attributes on diagonal elements and off-diagonal elements are 0.