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

Student's Name: Aryan Tiwari **Mobile No: 8982562898**

Roll Number: B20187 **Branch:** Electrical Engineering

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	50.0	5	12
7	pedi	0.078	1.191	5	12
8	Age (in years)	21	66	5	12

Inferences:

1. Most of the time outliers are indicators of noisy data and are hence removed
2. Median is used replace the outliers as it is the central value of skewed data.
3. The varying range of different attributes becomes common.

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	3.78	3.27	0	1
2	plas	121.65	30.43	0	1
3	pres (in mm Hg)	72.19	11.14	0	1
4	skin (in mm)	20.43	15.69	0	1
5	test (in mu U/mL)	60.91	77.63	0	1
6	BMI (in kg/m^2)	32.19	6.41	0	1
7	pedi	0.427	0.245	0	1
8	Age (in years)	32.76	11.055	0	1

Inferences:

1. Standardisation converts every attribute to normal distribution with mean =0 and variance = 1.

2 a.

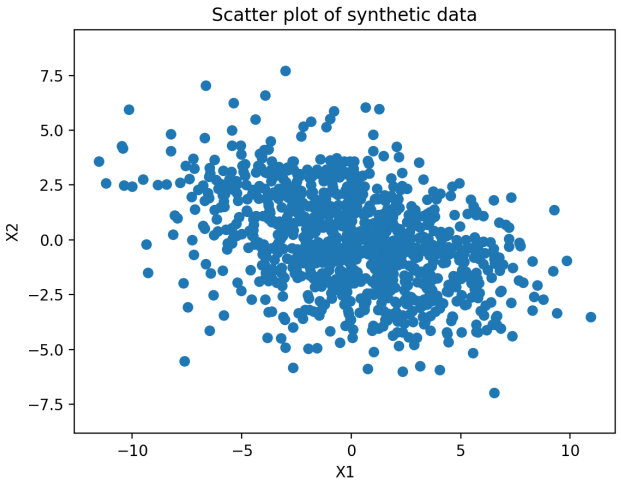

Figure 1 Scatter plot of 2D synthetic data of 1000 samples

Inferences:

1. They are negatively correlated.

b.

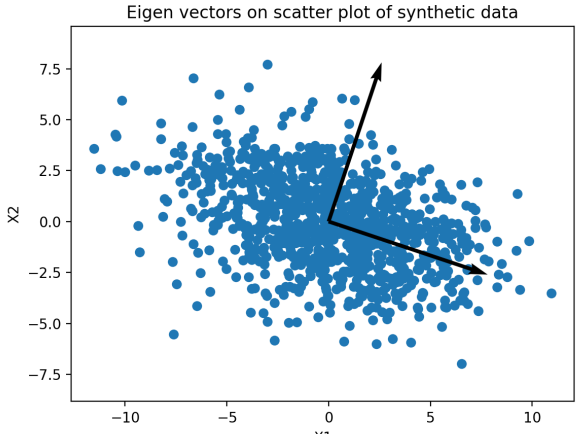

Figure 2 Plot of 2D synthetic data and Eigen directions

Inferences:

1. The data is spread more across the 1st eigen vector than the 2nd which shows the greater variance of 1st.
2. The plot is very densely covered at intersection.

c.

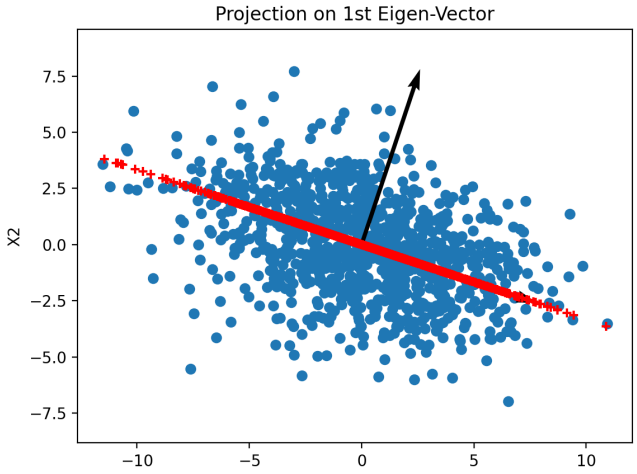

Figure 3 Projected Eigen directions onto the scatter plot with 1st Eigen direction highlighted

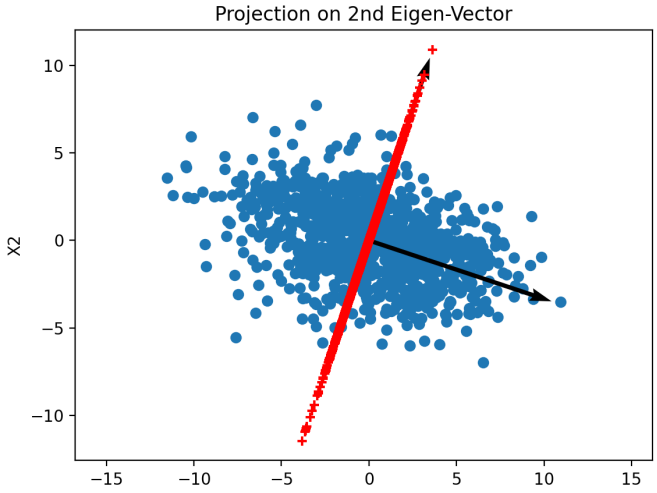

Figure 4 Projected Eigen directions onto the scatter plot with 2nd Eigen direction highlighted

Inferences:

1. The data is denser along the 1st eigen vector.
2. The variance along 1st eigen column is more.

d. Reconstruction error = 0.0

Inferences:

1. Since number of eigen vectors selected for reconstruction is same as dimension of original data, there is no loss.

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 of PCA data frame is same as 2 largest eigen values.

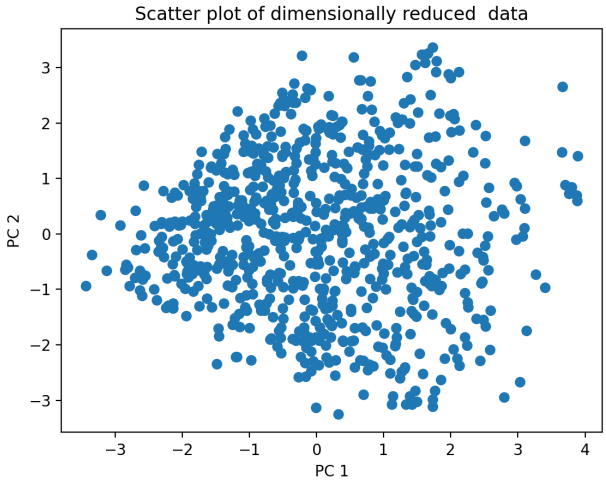

Figure 5 Plot of data after dimensionality reduction

Inferences:

1. Infer the correlation between the two attributes obtained after dimensionality reduction from the spread of data points
2. Inference 2(You may add or delete the number of inferences)
   Note: The scatter plots above are for illustration purposes. Replace it with the scatter plot obtained by you. Rename x-axis legend with x1 and y-axis legend with x2.

b.

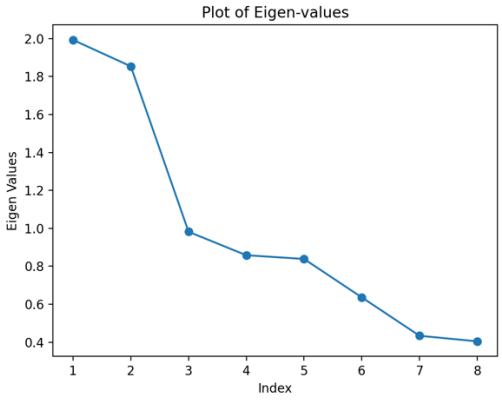

Figure 6 Plot of Eigenvalues in descending order

Inferences:

1. The eigen-values decrease rapidly
2. The eigenvalue decrease significantly after 1.85.

c.

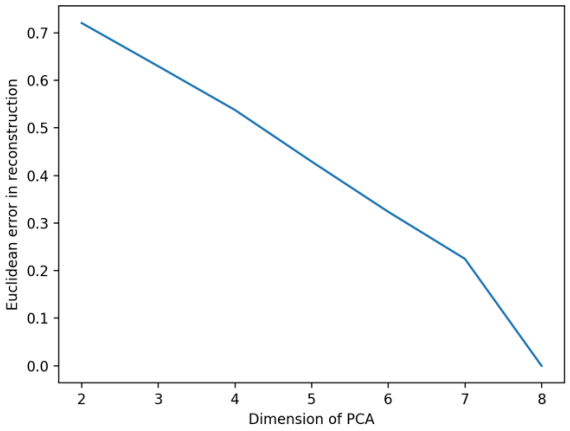

Figure 7 Line plot to demonstrate reconstruction error vs. components

Inferences:

1. Higher the magnitude of variance of better is the quality of reconstruction.

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

	0	1
0	1.992463047089108	-2.8949752923732895e-16
1	-2.8949752923732895e-16	1.8534221929648187

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

	0	1	2
0	1.9924630470891052	-2.2001812222037e-16	9.611317970679321e-17
1	-2.2001812222037e-16	1.8534221929648182	-4.238243828034496e-16
2	9.611317970679321e-17	-4.238243828034496e-16	0.9818791372481482

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

	0	1	2	3
0	1.9924630470891058	7.157826410392958e-16	-2.0641173834621555e-16	-6.80319193707723e-17
1	7.157826410392958e-16	1.8534221929648191	1.5676291208201362e-16	-2.2316640785082596e-16

2	-2.0641173834621555e-16	1.5676291208201362e-16	0.9818791372481482	2.967349674682622e-18
3	-6.80319193707723e-17	-2.2316640785082596e-16	2.967349674682622e-18	0.8583073274875039

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

	0	1	2	3	4
0	1.9924630470891045	3.8445271882717285e-16	3.688198522483571e-16	-9.727116982374253e-17	-4.713019775983716e-16
1	3.8445271882717285e-16	1.8534221929648207	3.670828670729331e-16	6.947940701695894e-18	-1.1464102157798227e-16
2	3.688198522483571e-16	3.670828670729331e-16	0.9818791372481508	-1.5053871520341105e-16	3.1960527227801113e-16
3	-9.727116982374253e-17	6.947940701695894e-18	-1.5053871520341105e-16	0.8583073274875052	4.4524719996701019e-16
4	-4.713019775983716e-16	-1.1464102157798227e-16	3.1960527227801113e-16	4.4524719996701019e-16	0.8387495985802278

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

	0	1	2	3	4	5
0	1.992463047089105	8.569126865424937e-17	3.896636743534475e-16	-3.057093908746193 6e-16	6.600543666611e-17	-2.547578257288494 6e-17
1	8.569126865424937e-17	1.8534221929648163	-1.065350907593370 6e-16	-2.014902803491809 4e-16	-2.258080728051165 7e-16	9.148121923899594e-17
2	3.896636743534475e-16	-1.065350907593370 6e-16	0.9818791372481492	1.273789128644247 3e-16	-2.617057664305453 5e-16	8.684925877119867e-19
3	-3.057093908746193 6e-16	-2.014902803491809 4e-16	1.273789128644247 3e-16	0.8583073274875033	1.818044483610425 7e-16	2.541788306703748e-16

4	6.6005436666111e-17	-2.2580807280511657e-16	-2.6170576643054535e-16	1.8180444836104257e-16	0.8387495985802265	-1.8006746318561686e-16
5	-2.5475782572884946e-17	9.1481219238995594e-17	8.6849258771198867e-19	2.5417883067037748e-16	-1.8006746318561686e-16	0.6364083688676926

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

	0	1	2	3	4	5	6
0	1.9924630470891072	1.1394622750781268e-15	-1.8933138412121313e-16	-2.3159802338986317e-17	1.0653509075933706e-16	1.4127479426781653e-16	-9.842915994069184e-17
1	1.1394622750781268e-15	1.853422192964821	2.1307018151867411e-16	-6.021548608136442e-17	4.4756318020091060e-16	7.6427347718654840e-17	2.3159802338986317e-17
2	-1.8933138412121313e-16	2.1307018151867411e-16	0.9818791372481492	-2.3449299868223645e-16	-2.4086194432545770e-16	3.7055683742378105e-17	1.3895881403391788e-17
3	-2.3159802338986317e-17	-6.021548608136442e-17	-2.3449299868223645e-16	0.8583073274875037	4.5740609619497970e-17	2.2609757033435390e-16	-8.684925877119869e-18
4	1.0653509075933706e-16	4.4756318020091060e-16	-2.4086194432545770e-16	4.5740609619497970e-17	0.8387495985802265	-5.732051078899114e-17	9.5534184648318560e-17
5	1.4127479426781653e-16	7.6427347718654840e-17	3.7055683742378105e-17	2.2609757033435390e-16	-5.732051078899114e-17	0.6364083688676925	-3.297376858013177e-16
6	-9.842915994069184e-17	2.3159802338986317e-17	1.3895881403391788e-17	-8.684925877119869e-18	9.5534184648318560e-17	-3.297376858013177e-16	0.4341428198613043

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

	0	1	2	3	4	5	6	7
0	1.99246305	1.14E-15	-1.89E-16	-2.08E-17	1.07E-16	1.41E-16	-9.84E-17	-2.36E-16
1	1.14E-15	1.85342219	2.13E-16	-4.40E-17	4.48E-16	7.64E-17	2.32E-17	-1.32E-16
2	-1.89E-16	2.13E-16	0.98187914	-2.38E-16	-2.41E-16	3.71E-17	1.39E-17	1.71E-16
3	-2.08E-17	-4.40E-17	-2.38E-16	0.85830733	4.57E-17	2.26E-16	-8.68E-18	-4.40E-17
4	1.07E-16	4.48E-16	-2.41E-16	4.57E-17	0.8387496	-5.73E-17	9.55E-17	2.93E-16
5	1.41E-16	7.64E-17	3.71E-17	2.26E-16	-5.73E-17	0.63640837	-3.30E-16	-1.86E-16
6	-9.84E-17	2.32E-17	1.39E-17	-8.68E-18	9.55E-17	-3.30E-16	0.43414282	-1.16E-18
7	-2.36E-16	-1.32E-16	1.71E-16	-4.40E-17	2.93E-16	-1.86E-16	-1.16E-18	0.40462751

Inferences:

1. The off-diagonal elements are tending towards 0 because the eigen-vectors are orthonormal and hence the projection don't overlap which makes the attributes uncorrelated.
2. The diagonal elements are variance of individual attributes and hence non-zero but the off diagonal elements represent correlation coefficient of PCA df which is 0.
3. The diagonal values are in decreasing order.
4. PCA of sklearn sorts the eigen-values(variance) in decreasing order.
5. The 1st diagonal element is biggest and hence the 1st column (0th index) captures the data best.
6. The 1st diagonal element is same for every matrix as it is the highest eigen-value of the dataframe.
7. The 2nd diagonal element is also same for every matrix as it is fixed(2nd highest eigen value).
8. The diagonal elements of covariance matrix won't change as they are eigen values or variance and are pre-calculated.

d.

Table 11 Covariance matrix for original data

	pregs	plas	pres	skin	Test	BMI	pedi	Age
pregs	1	1.18E-01	2.09E-01	-9.67E-02	-1.08E-01	2.83E-02	4.52E-03	5.61E-01
plas	1.18E-01	1	2.05E-01	6.00E-02	1.80E-01	2.28E-01	8.16E-02	2.74E-01
pres	2.09E-01	2.05E-01	1	2.56E-02	-5.10E-02	2.72E-01	2.25E-02	3.26E-01
skin	-9.67E-02	6.00E-02	2.56E-02	1	4.73E-01	3.74E-01	1.53E-01	-1.01E-01
test	-1.08E-01	1.80E-01	-5.10E-02	4.73E-01	1	1.72E-01	1.99E-01	-7.37E-02
BMI	2.83E-02	2.28E-01	2.72E-01	3.74E-01	1.72E-01	1	1.24E-01	7.77E-02
pedi	4.52E-03	8.16E-02	2.25E-02	1.53E-01	1.99E-01	1.24E-01	1	3.61E-02
Age	5.61E-01	2.74E-01	3.26E-01	-1.01E-01	-7.37E-02	7.77E-02	3.61E-02	1

Inferences:

1. The off-diagonal elements are non-zero as the attributes are correlated.
2. The diagonal values of original data covariance matrix are 1 as the data is standardized, while the diagonal elements decrease in l=8 PCA covariance matrix .
3. The columns of original matrix are equally important with equal variance while the importance of each column decreases along the PCA matrix.