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

Student's Name: Gaurav Kumar **Mobile No: 8529143452**

Roll Number: B20197 **Branch:**EE

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	5.0	12
2	plas	44	199	5.0	12
3	pres (in mm Hg)	38	106	5.0	12
4	skin (in mm)	0	63	5.0	12
5	test (in mu U/mL)	0.0	318.0	5.0	12
6	BMI (in kg/m^2)	18.2	50.0	5.0	12
7	pedi	0.078	1.191	5.0	12
8	Age (in years)	21	66	5.0	12

Inferences:

1. Since outliers increase the variability in the data and excluding outliers make our data statistically significant.
2. Pergs, test and pedi attributes have many outliers.
3. First calculated Q1(1st quartile),Q3(3rd quartile), and then IQR. From IQR,Q1,Q3 found the value of Lower Bound (Q1-(1.5*IQR)), Upper Bound (Q3+(1.5*IQR)) . And if the points in data were above above upper bound or if points where lower then the lower bound then the datum is a outlier . Replace it with median of the respective attribute.
4. Before normalization the data values of some attributes have variability which makes which will make some attributes overpower while training models.
5. Before normalization the data values of some attributes have variability which makes which will make some attributes overpower while training models.

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	-5.551	1.0
2	plas	121.656	30.438	-1.387	1.0
3	pres (in mm Hg)	72.196	11.146	4.001	1.0
4	skin (in mm)	20.437	15.698	-1.387	1.0
5	test (in mu U/mL)	60.919	77.635	-5.319	1.0
6	BMI (in kg/m^2)	32.198	6.410	-6.129	1.0
7	pedi	0.427	0.245	2.521	1.0
8	Age (in years)	32.760	11.055	2.127	1.0

Inferences:

1. Before standarlization the data values of some attributes have variability which makes which will make some attributes overpower while training models.
2. After the standardlization,the values gets centered around the mean with a unit standard deviation.

2 a.

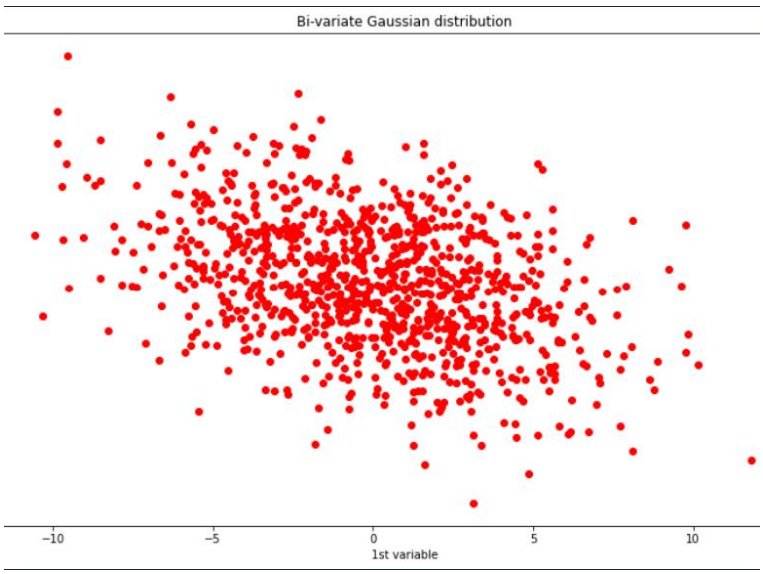

Figure 1 Scatter plot of 2D synthetic data of 1000 samples

Inferences:

1. Attribute 1 is negatively correlated with Attribute 2 with correlation coefficient = -0.3329 (it can change as data values are random).
2. According to density , points are near about best fit line or not variability data.

b.

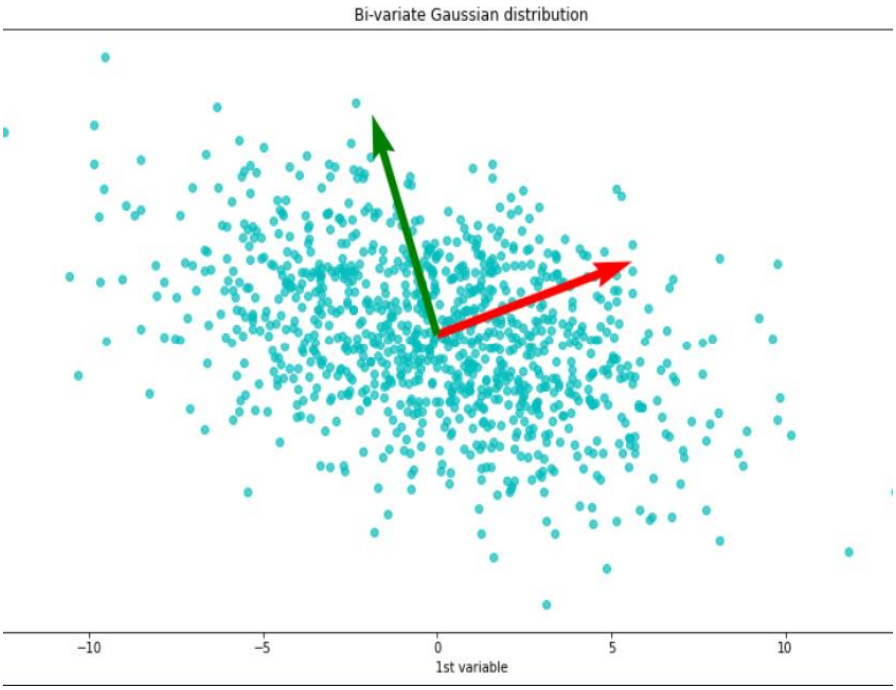

Figure 2 Plot of 2D synthetic data and Eigen directions

Inferences:

1. Eigenvalues gives the information of variance of data and according to eigenvalue in question data is not much variability or have high variance from mean along first eigen vector as compared to second eigen vector.
2. Near the intersection of Eigen axes , density of points are very high and gradually as we move away from it density decreases.

c.

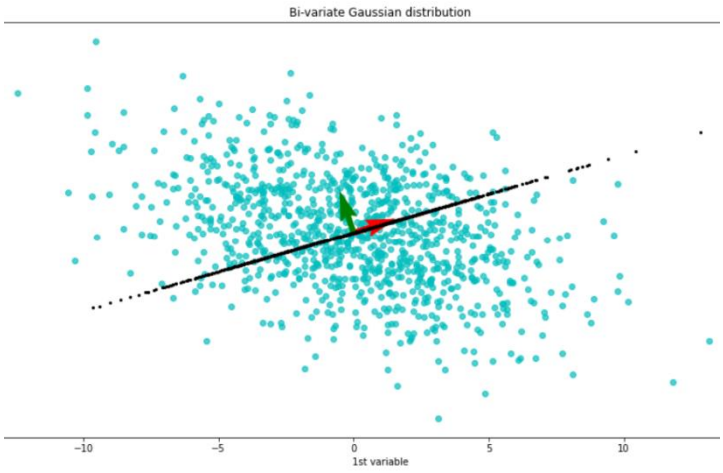

Figure 3 Projected Eigen directions onto the scatter plot with 1st Eigen direction highlighted

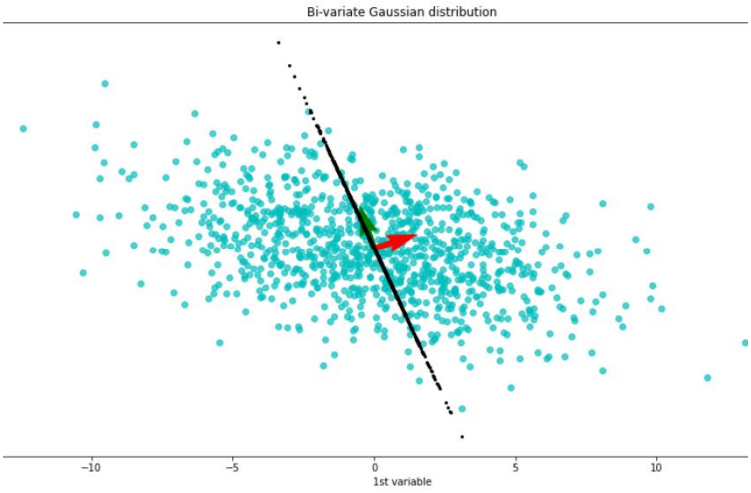

Figure 4 Projected Eigen directions onto the scatter plot with 2nd Eigen direction highlighted

Inferences:

1. Eigen value(=14) which is greater has more spread data along its respective Eigen direction while the Eigen value(=4) has less spread along its respective Eigen direction.
2. Regarding the density, along the first Eigen vector (smaller line), the variance is not very large, so the spread is not so much varying. However, along the second Eigen Vector, the variance is high, so the spread is more, so the density actually is high near the intersection and spread is large
3. Eigenvalues gives the information of variance of data and according to eigenvalue in question data is not much variability or have high variance from mean.

d. Reconstruction error = 3.338e-13 (nearly 0)

Inferences:

1. After reconstructing matrix , magnitude of reconstruction error indicates that how much data values are apart in original and reconstruction matrix. More the reconstruction error, more loss in the nature of data. Here the reconstruction error was zero because the number of dimension remained same after reconstructing the data.

3 a.

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

Direction	Variance	Eigenvalue
1	1.992	1.853
2	1.992	1.853

Inferences:

1. Higher the values of Eigen Vector, more variance along that vector, so more strength along that direction. So, we can say that data will be more spread along the first Eigen vector.

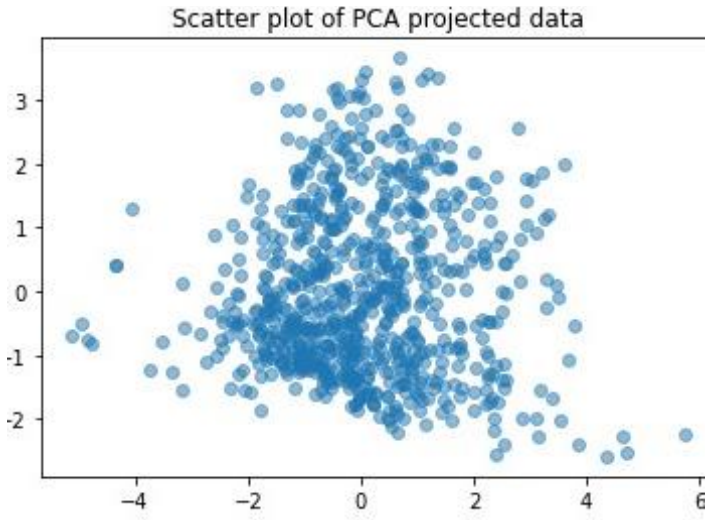

Figure 5 Plot of data after dimensionality reduction

Inferences:

1. As the density along positive slope is more so it seems by looking at the graph that the data is positively correlated.

b.

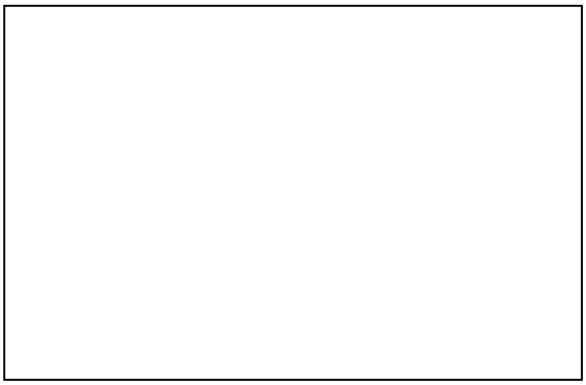

Figure 6 Plot of Eigenvalues in descending order

Inferences:

1. It drops rapidly from second to third Eigen value and then decreases gradually
2. From the third Eigenvalue the rate of decrease changes substantially.

c.

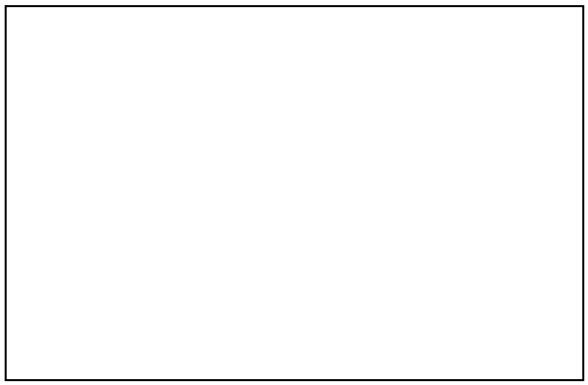

Figure 7 Line plot to demonstrate reconstruction error vs. components

Inferences:

1. After reducing dimension , when we reconstruct the new matrix to original matrix , reconstruction error measure the difference in the distance between predicted data from the original data.

2. After reducing dimension , when we reconstruct the new matrix to original matrix , reconstruction error measure the difference in the distance between predicted data from the original data.

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	0.853	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	0.6360	0	0
x7	0	0	0	0	0	0	0.434	0
x8	0	0	0	0	0	0	0	0.405

Inferences:

1. Observe off-diagonal elements and state the reason for the observed trend.
2. Observe the difference between diagonal and off-diagonal values and justify the reason for your observation.
3. Infer the trend in diagonal values.

4. Justify the reason for the increase/ decrease.
5. From the magnitude of diagonal elements, which component captures data variations the best?
6. From the value of diagonal elements, estimate the number of components that shall give the optimum reconstruction along with dimensionality reduction.
7. Observe the magnitude of the 1st diagonal element (topmost left corner) in each of the obtained covariance matrices. Is the magnitude the same or different? Ponder upon the underlying reason for your observation.
8. Observe the magnitude of the 2nd diagonal element in each of the obtained covariance matrices. Is the magnitude the same or different? Ponder upon the underlying reason for your observation.
9. Compare 3rd, 4th, 5th, 6th, and 7th diagonal elements across covariance matrices. Are they the same or different?
10. Inference 10(You may add or delete the number of inferences)

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	0.264	0.022	0.326
skin (in mm)	-0.097	0.06	0.026	1	0.473	0.153	0.153	-0.101
test (in mu U/mL)	0.108	0.180	-0.051	0.473	1	0.199	0.199	-0.074
BMI (in kg/m^2)	0.028	0.228	0.272	0.374	0.172	0.124	0.124	0.078
pedi	0.005	0.082	0.022	0.153	0.199	1	1	0.036
Age (in years)	0.561	0.274	0.326	-0.101	-0.074	0.036	0.036	1

Inferences:

1. Observe the off-diagonal values and compare with the covariance matrix obtained after PCA l=8 reduction.
2. Similarly, compare the magnitudes of diagonal values.
3. Is there any trade of increase or decrease in diagonal elements like/ unlike covariance obtained after dimensionality reduction?
4. Inference 4(You may add or delete the number of inferences)