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

Student's Name: Priyanka Kumari **Mobile No:** 8328354314

Roll Number: B20307 **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	0.00	13.00	5.00	12.00
2	plas	44.00	199.00	5.00	12.00
3	pres (in mm Hg)	38.00	106.00	5.00	12.00
4	skin (in mm)	0.00	63.00	5.00	12.00
5	test (in mu U/mL)	0.00	318.00	5.00	12.00
6	BMI (in kg/m^2)	18.20	50.00	5.00	12.00
7	pedi	0.078	1.191	5.00	12.00
8	Age (in years)	21.00	66.00	5.00	12.00

Inferences:

1. presence of an outlier can make our data noisier; it can have disproportionate effect on statistical results which can result in misleading interpretations. Thus, outlier correction is needed to make data less noisy and analysis of data becomes much easier.
2. First step was to identify the outliers present in each attribute. So, all the values which are less than the lower quartile – (IQR*1.5) and greater than the upper quartile+(IQR*1.5) can be identified as outliers. Once we identified the outliers it can be replaced by median of that particular attribute. This method is more efficient as the data becomes less noisy and most of the outliers get replaced by median value and make aur data more reliable.
3. Initially data contained a large range of min and max values which lead to more variations however after normalization the min and max values fall under specified range of 5 to 12.
4. Min-Max normalization preserves the relationship among the original data values.

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	3.782552	3.270644	-1.324172e-16	1.0
2	plas	121.656250	30.438286	-5.637851e-18	1.0
3	pres (in mm Hg)	72.196615	11.146723	4.840240e-16	1.0
4	skin (in mm)	20.437500	15.698554	-1.445603e-19	1.0
5	test (in mu U/mL)	59.569010	78.415321	1.006140e-16	1.0
6	BMI (in kg/m^2)	32.198958	6.410558	3.896839e-15	1.0
7	pedi	0.427044	0.245323	1.051965e-15	1.0
8	Age (in years)	32.760417	11.055385	1.937108e-16	1.0

Inferences:

1. Before standardization we had wide variations in the mean and standard deviation values. After standardization the data became less variant with standard deviation value as 1.0.
2. Data standardization is the process of converting data to a common format, Standardization promotes productivity by eliminating inefficiency. This is the result of eliminating ambiguity and providing quality control.

2 a.

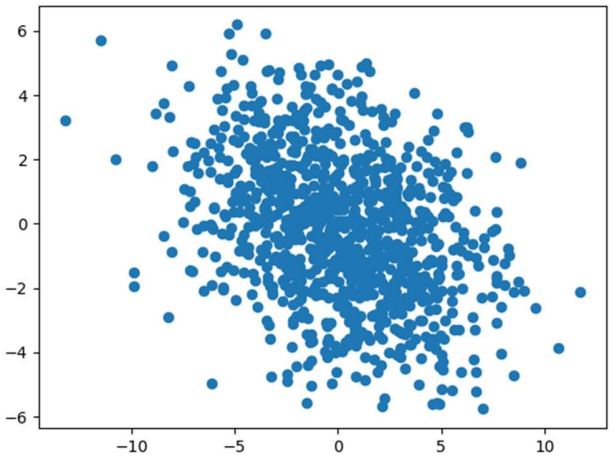


Figure 1 Scatter plot of 2D synthetic data of 1000 samples

Inferences:

1. from the above data attribute 1 is negatively correlated with attribute 2. As the value of attribute 1 increases value of attribute 2 decreases.
2. Density becomes higher in the range from -5 to 5.

b.

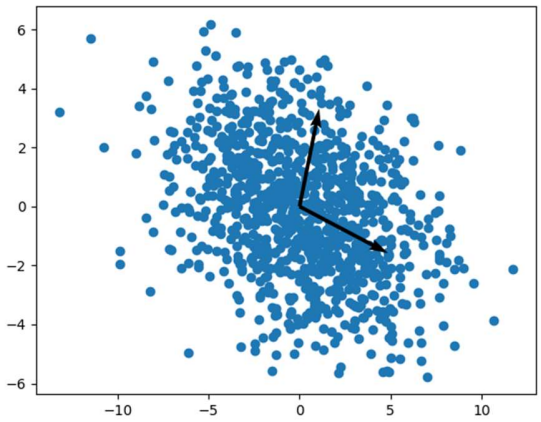


Figure 2 Plot of 2D synthetic data and Eigen directions

Inferences:

1. Data spread is more up to the magnitude of vectors.
2. From the above graph, the density of data is more near the intersection of the 2 vectors while it reduces as we move along the vector direction.

c.

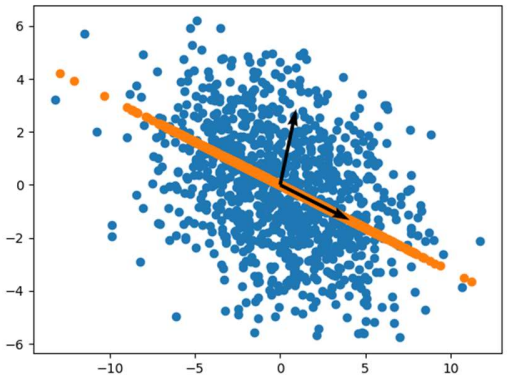


Figure 3 Projected Eigen directions onto the scatter plot with 1st Eigen direction highlighted

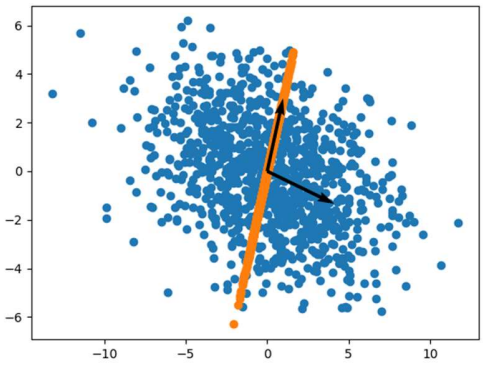


Figure 4 Projected Eigen directions onto the scatter plot with 2nd Eigen direction highlighted

Inferences:

1. Magnitude of first eigen vector is more than that of second.
2. Variance of projections along the eigen vectors is equal to the magnitude of their corresponding eigen values.

d. Reconstruction error = 0.00

Inferences:

1. The accuracy of the reconstructed data depends on the reconstruction error. Smaller the reconstruction error larger the accuracy of the reconstruction data.

3 a.

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

Direction	Variance	Eigenvalue
1	1.992	-0.000
2	0.2490	0.232

Inferences:

1. Variance of the data in one of the eigendirections is slightly higher than the other. It shows the data is spread more out of the line of first eigendirection
2. Inference 2(You may add or delete the number of inferences)

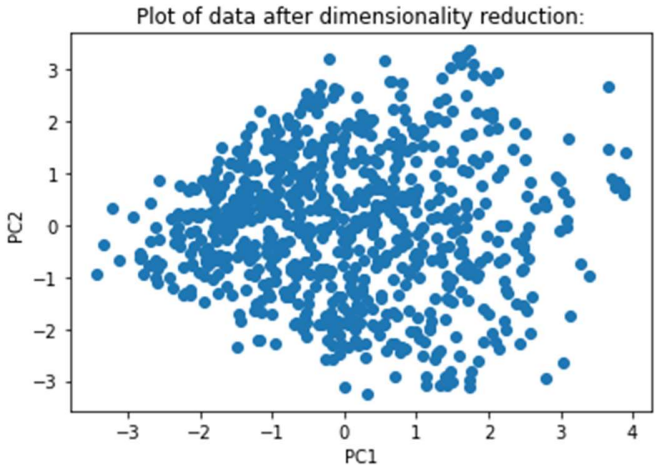
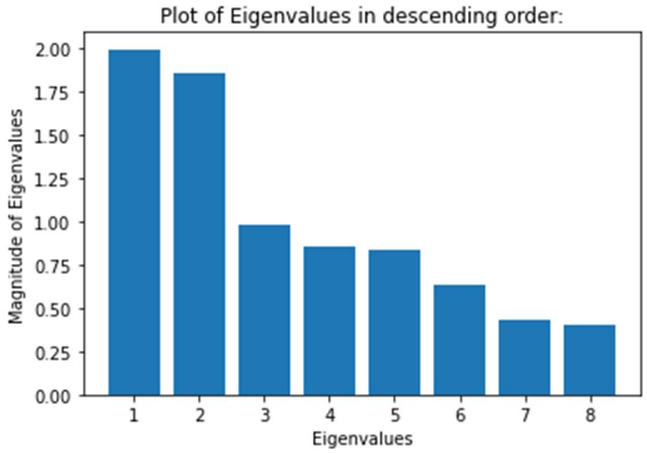


Figure 5 Plot of data after dimensionality reduction

Inferences:

1. The scatter plot appears to be symmetrically distributed among the axes. All the values range between -4 and +4.
2. They seem to have a week positive correlation.

b.



Inferences:

1. The eigen values tend to decrease rapidly.
2. The Eigenvalue from where the rate of decrease changes substantially is 1.853.

c.

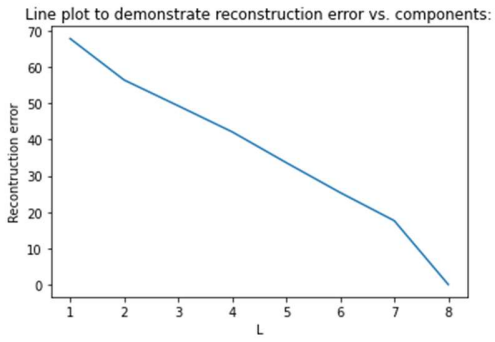


Figure 6 Line plot to demonstrate reconstruction error vs. components

Inferences:

1. The magnitude of reconstruction error affects the quality of reconstruction inversely.

2. As the reconstruction error decreases, the components value increases.

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

	x1	x2
x1	1.99	0
x2	0	1.853

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

	x1	x2	x3
x1	1.99	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	0.585	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 all zero since the new attributes are independent of one another, they are uncorrelated.

2.The diagonal elements are non-zero values since these values depicts the level of projection on the eigenvectors of these attributes. Rest non diagonal elements are all zero since they are become uncorrelated after performing the pca analysis.

3. We see a decreasing trend of the diagonal values from top to bottom.

4. Since these attributes are ordered according to the magnitude of their projection on eigenvectors or their eigenvalues. Hence lower the magnitudes of eigenvalues, lower is the variance as shown in the covariance matrices.

5.The component x1 captures the data variation in the best way since it has the highest magnitude of projection.

6.From the value of diagonal elements, the first two components will give the optimum reconstruction along with dimensionality reduction.

7.We observe the magnitude of the 1st diagonal element (topmost left corner) in each of the obtained covariance matrices is same since the variance of this component is unaffected by other components as all of them are uncorrelated.

8.We observe the magnitude of the 2nd diagonal element in each of the obtained covariance matrices is same since the other components doesn't affect the projection of this attribute on the eigenvectors.

9.Also the 3rd, 4th, 5th, 6th, and 7th diagonal elements across covariance matrices are all same. The reason is same as former that the projection of all these components is independent of each other. This is what PCA analysis does to our data i.e., it makes all the components uncorrelated.

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	0.06	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)	-0.0097	0.06	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 values of the dimensionally reduced matrix is zero whereas in the original matrix we find them to be non-zero. This is because of the pca analysis performed to obtain our dimensionally reduced data which makes our components or attributes uncorrelated thus making variance values between different attributes equal to zero.

2.The diagonal elements of original matrix is 1 since the variance between attribute and itself will always be unity whereas in our pca analysis performed data we obtain less variance values for all the diagonal elements since this represents the magnitude of projection of datapoints of attributes on eigendirections.