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

Student's Name: Nikhil **Branch:**

Roll Number: B20219 **CSE**

Mobile No: 8949463760

1 a.

Table 1 Minimum and maximum attribute values before and after normalization

```
The minimum and maximum values before performing the Min-Max normalization of the attributes is
 {'pregs': [0.0, 13.0], 'plas': [44.0, 199.0], 'pres': [38.0, 106.0], 'skin': [0.0, 63.0], 'test': [0.0, 318.0], 'BMI': [18.2, 50.0], 'pedi': [0.078, 1.191], 'Age': [21.0, 66.0]}
The minimum and maximum values after performing the Min-Max normalization of the attributes is
 {'pregs': [5.0, 12.0], 'plas': [5.0, 12.0], 'pres': [5.0, 12.0], 'skin': [5.0, 12.0], 'test': [5.0, 12.0], 'BMI': [5.0, 12.0], 'pedi': [5.0, 12.0], 'Age': [5.0, 12.0]}
```

Inferences:

1. Most of the outliers are noise. Removing the outliers is a part of reduction of noise in the data.
2. For a normal distribution, about 99.8% data lies between μ-3σ and μ+3σ. Taking the upper bound as Q3+1.5IQR simplifies the calculation and it approximates to about μ+2.7σ. Since out data doesn't have much of a skewness, this is indeed a good approximation to declare the domain of the attribute.
3. Before normalization, the attributes take values in their actual domain, but after normalization, the difference between all the attribute values is scaled down to a constant factor and the minimum of the data is shifted to some fixed value, so now the data in all attributes lies in the same range, and the relationship between them does not change. This type of scaling technique needs good domain understanding, and the main advantage of it is, we don't have to make any assumptions to scale the data down.

b.

Table 2 Mean and standard deviation before and after standardization

```
The mean before standardization is
 {'pregs': 3.783, 'plas': 121.656, 'pres': 72.197, 'skin': 20.438, 'test': 60.919, 'BMI': 32.199, 'pedi': 0.428, 'Age': 32.76}
The standard deviation before standardization is
 {'pregs': 3.271, 'plas': 30.438, 'pres': 11.147, 'skin': 15.699, 'test': 77.636, 'BMI': 6.411, 'pedi': 0.245, 'Age': 11.055}
The mean after standardization is
 {'pregs': -0.0, 'plas': -0.0, 'pres': 0.0, 'skin': -0.0, 'test': -0.0, 'BMI': -0.0, 'pedi': 0.0, 'Age': 0.0}
The standard deviation after standardization is
 {'pregs': 1.0, 'plas': 1.0, 'pres': 1.0, 'skin': 1.0, 'test': 1.0, 'BMI': 1.0, 'pedi': 1.0, 'Age': 1.0}
```

Inferences:

1. Before standardization, all the attributes have their respective actual means and standard deviations. After standardization, means and standard deviations of all the attributes become 0 and 1 respectively. This type of scaling assumes the data to be a normal distribution, and scales the attribute values in such a way that the new mean and standard deviations are 0 and 1 respectively. This method does not require domain understanding.

2 a.

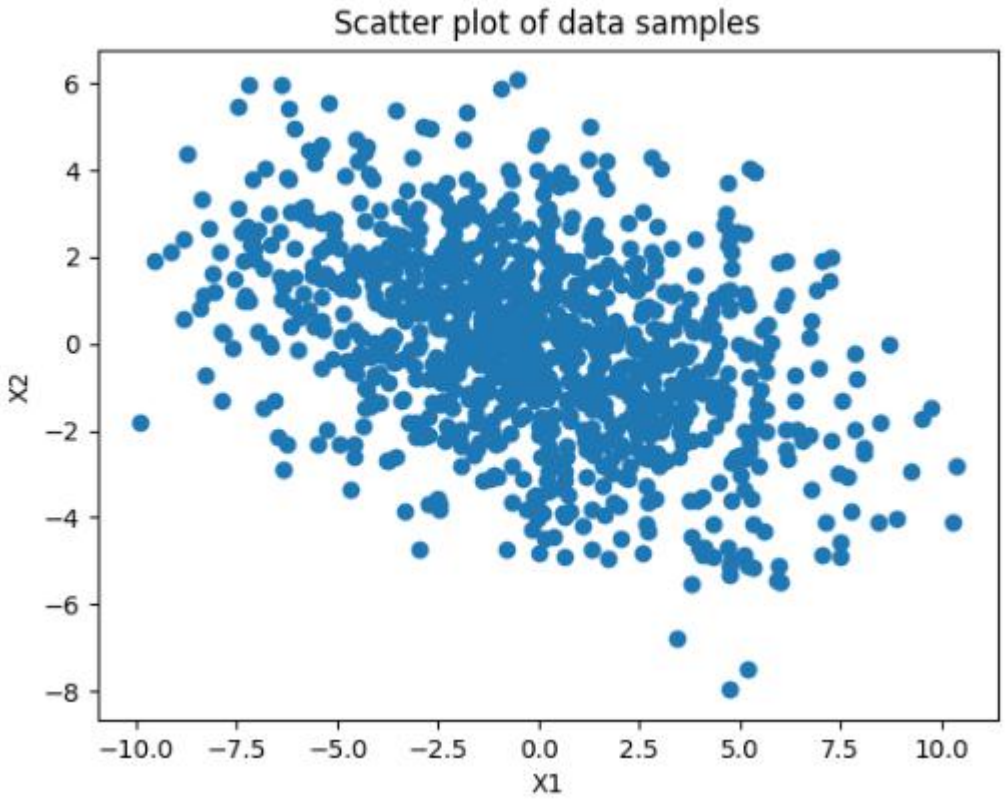

Figure 1 Scatter plot of 2D synthetic data of 1000 samples

Inferences:

1. The plot shows a decreasing trend of value of X2 with respect to increase in value of X1. So it can be concluded that the attributes have a moderately negative correlation correlation coefficient.

2. The density of points is maximum around (0,0), and it keeps on decreasing as we go beyond the origin, which means that more number of samples in the data lie close to 0, and the probability of finding a sample farther from origin decreases as the distance from the origin decreases.

b.

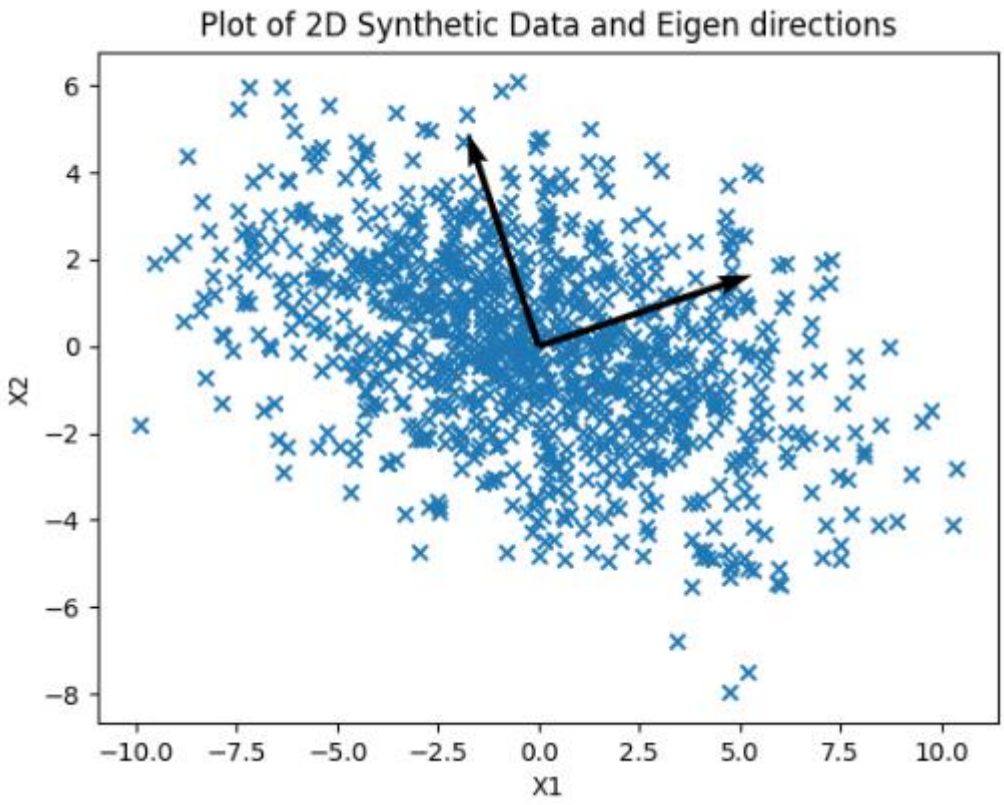

Figure 2 Plot of 2D synthetic data and Eigen directions

Inferences:

1. The data is more spread along the first eigen direction, and this can be seen by the first eigen value being greater than the second eigenvalue.
2. Near the intersection there is very high density or points and it gradually decreases as we go further from the intersection. Since the intersection is at the origin, it means that more number of samples in the data lie close to 0, and the probability of finding a sample farther from origin decreases as the distance from the origin decreases.

c.

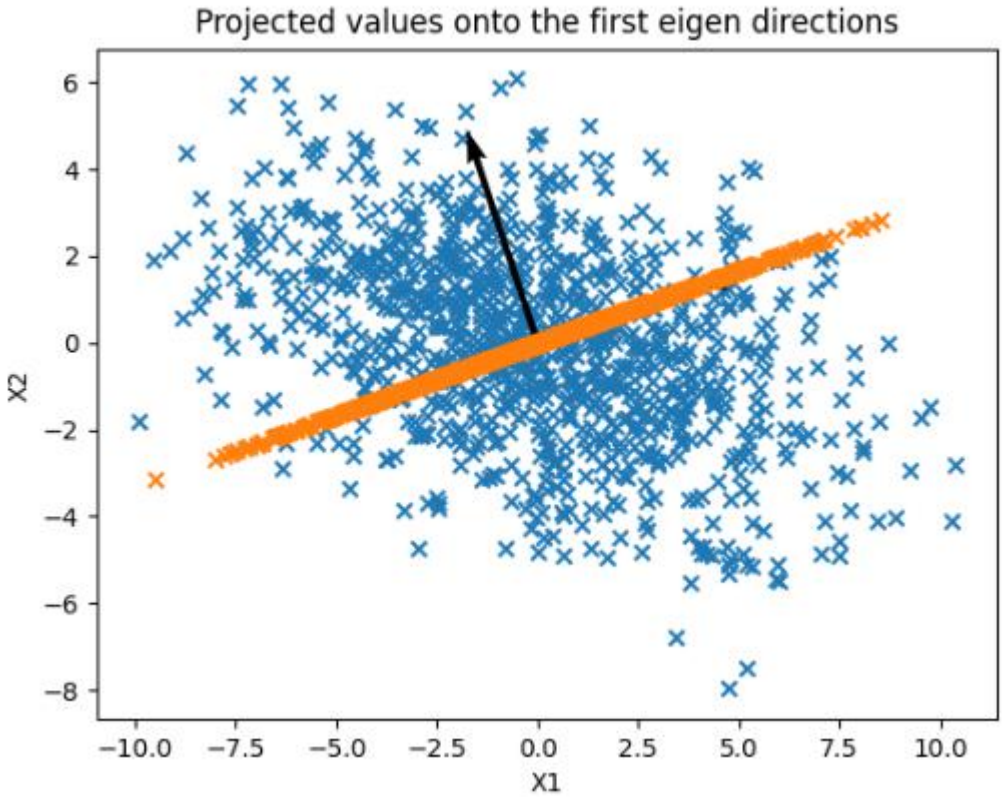

Figure 3 Projected Eigen directions onto the scatter plot with 1st Eigen direction highlighted

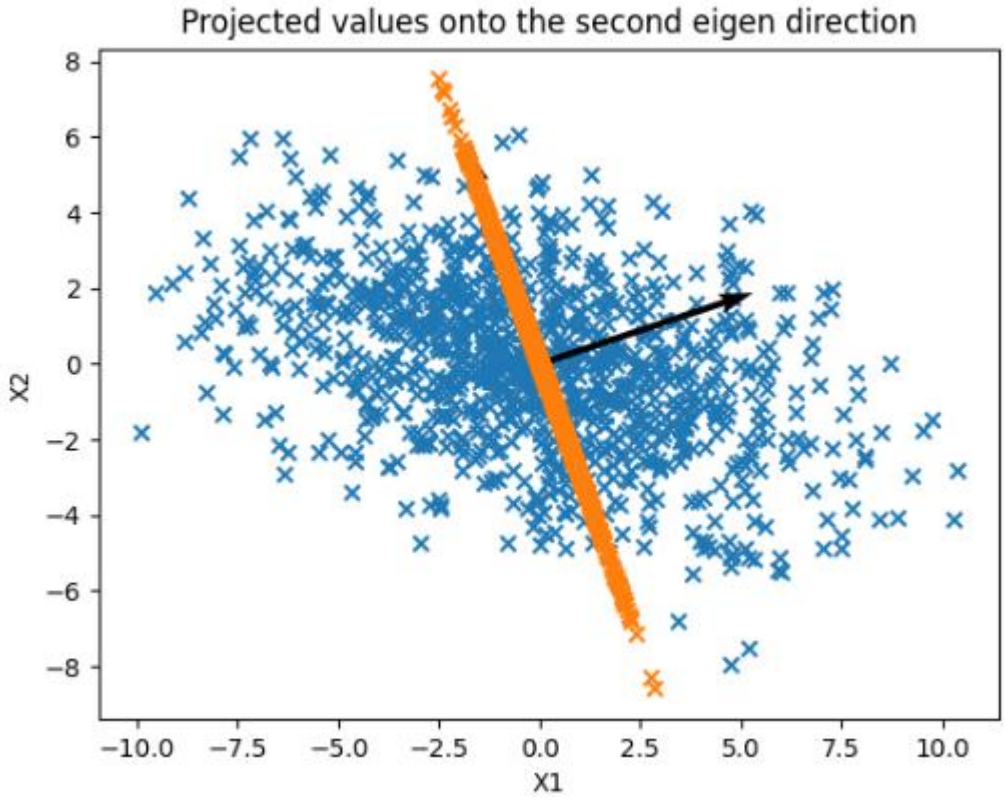

Figure 4 Projected Eigen directions onto the scatter plot with 2nd Eigen direction highlighted

Inferences:

1. The data is more spread along the first eigen direction, so this can be said that the first eigenvalue is greater than the second eigenvalue.
2. Near the intersection there is very high density or points and it gradually decreases as we go further from the intersection. Since the intersection is at the origin, it means that more number of samples in the data lie close to 0, and the probability of finding a sample farther from origin decreases as the distance from the origin decreases.

d. Reconstruction error = 0.000

Inferences:

1. We see that we have no reconstruction error which means that this reconstruction is 100% accurate. This error is zero because we used all the eigenvectors for reconstruction, not only the significant ones.

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. Interestingly, the variance of data projected along the 2 directions is exactly equal to the eigenvalues of the respective directions.

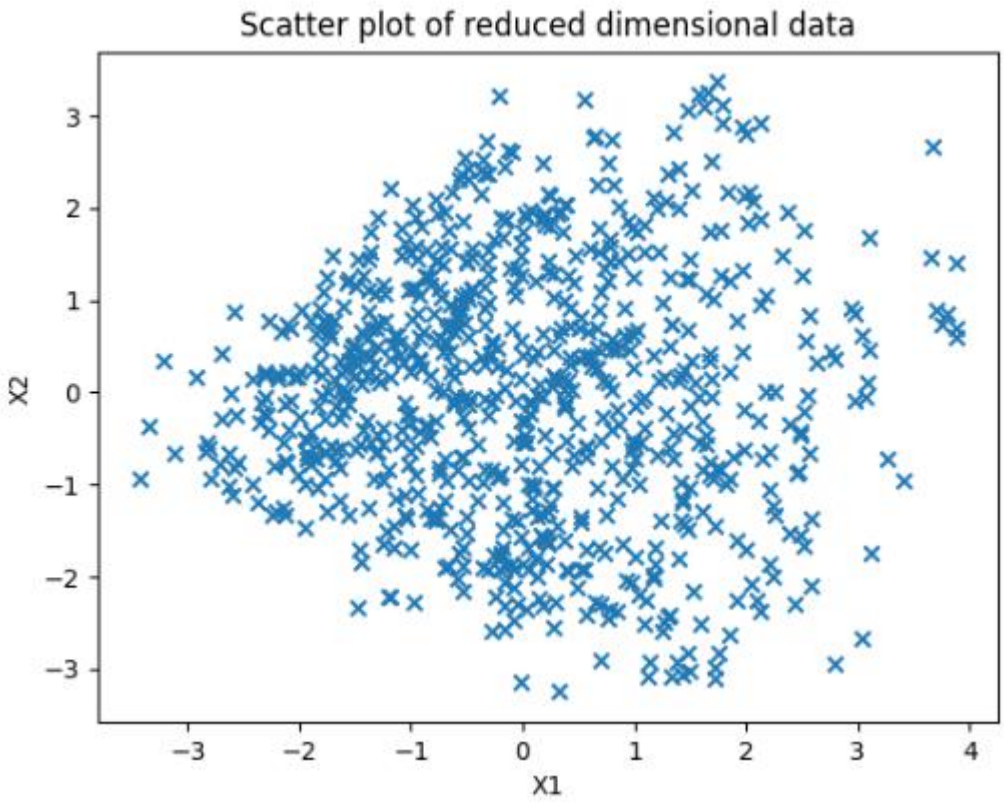

Figure 5 Plot of data after dimensionality reduction

Inferences:

1. Since we see no definite trend of increase or decrease of X2 on increase of X1, the given attributes are uncorrelated.

b.

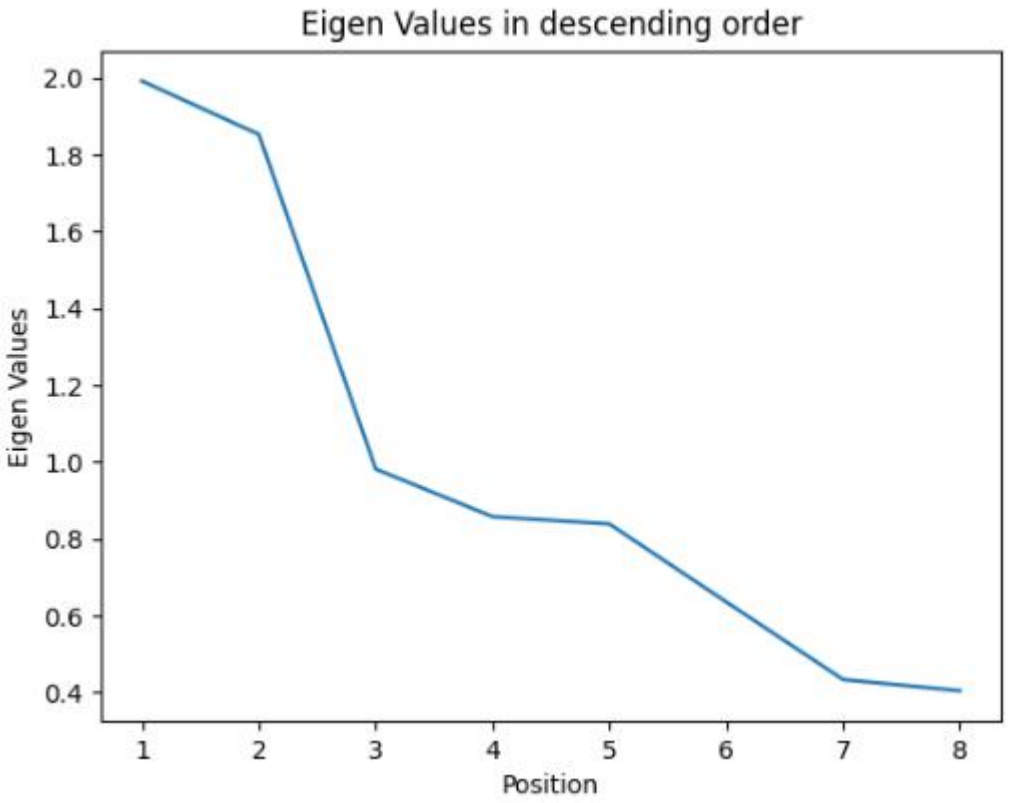

Figure 6 Plot of Eigenvalues in descending order

Inferences:

1. The values of the eigenvalues decrease rapidly after the second eigenvalue and then they decrease only gradually.
2. From second eigenvalue, the values of other eigen values decrease substantially.

c.

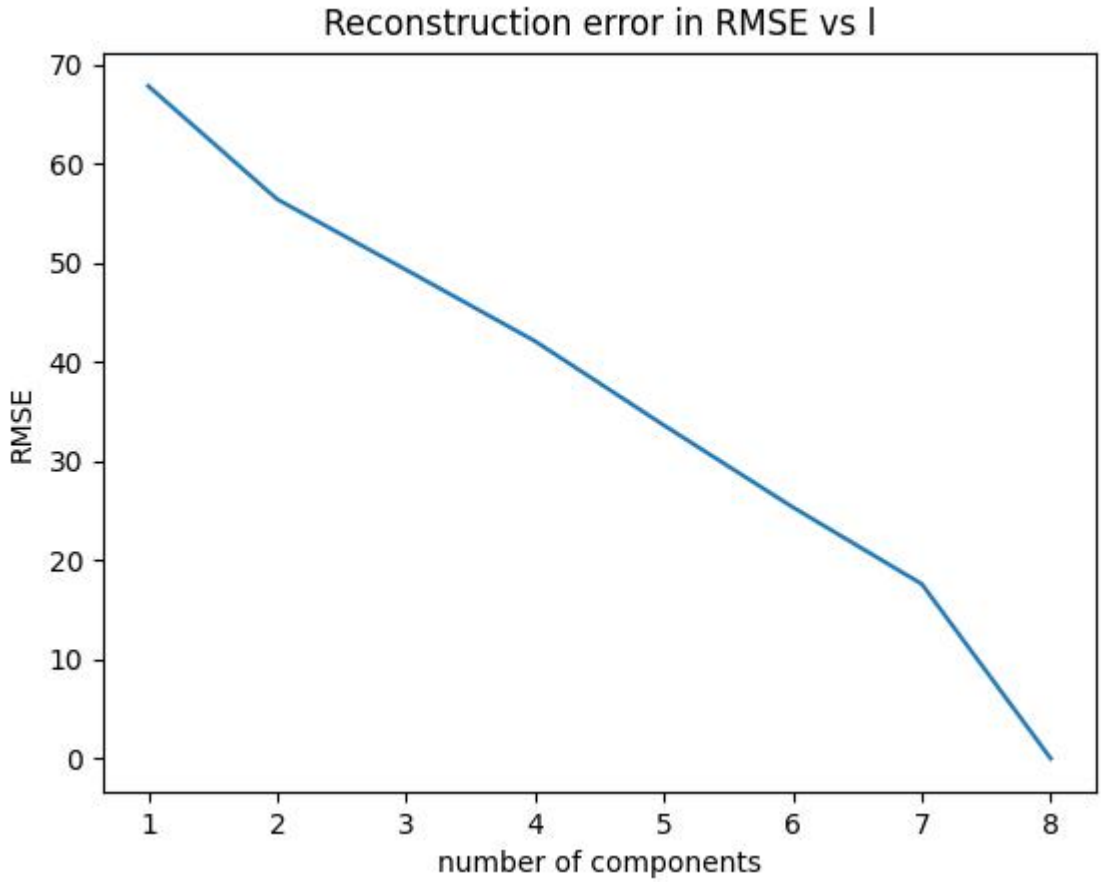

Figure 7 Line plot to demonstrate reconstruction error vs. components

Inferences:

1. We observe that greater the number of components, lower the reconstruction error. The quality of reconstruction depends upon the choice and number of significant eigenvalues. Larger eigenvalues and greater number of eigenvalues taken as significant reduces the reconstruction error.

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

	x1	x2
x1	1.992	0.000
x2	0.000	1.853

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

	x1	x2	x3
x1	1.992	0.000	0.000
x2	0.000	1.853	-0.000
x3	0.000	-0.000	0.982

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

	x1	x2	x3	x4
x1	1.992	0.000	0.000	-0.000
x2	0.000	1.853	-0.000	-0.000
x3	0.000	-0.000	0.982	-0.000
x4	-0.000	-0.000	-0.000	0.858

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

	x1	x2	x3	x4	x5
x1	1.992	0.000	0.000	0.000	0.000
x2	0.000	1.853	-0.000	0.000	0.000
x3	0.000	-0.000	0.982	0.000	-0.000
x4	0.000	0.000	0.000	0.858	-0.000
x5	0.000	0.000	-0.000	-0.000	0.839

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

	x1	x2	x3	x4	x5	x6
x1	1.992	0.000	0.000	0.000	0.000	-0.000
x2	0.000	1.853	-0.000	0.000	0.000	-0.000
x3	0.000	-0.000	0.982	0.000	0.000	-0.000
x4	0.000	0.000	0.000	0.858	0.000	0.000
x5	0.000	0.000	0.000	0.000	0.839	0.000
x6	-0.000	-0.000	-0.000	0.000	0.000	0.636

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

	x1	x2	x3	x4	x5	x6	x7
x1	1.992	-0.000	-0.000	-0.000	0.000	-0.000	0.000
x2	-0.000	1.853	-0.000	0.000	-0.000	0.000	-0.000
x3	-0.000	-0.000	0.982	0.000	-0.000	-0.000	0.000
x4	-0.000	0.000	0.000	0.858	0.000	0.000	-0.000
x5	0.000	-0.000	-0.000	0.000	0.839	-0.000	0.000
x6	-0.000	0.000	-0.000	0.000	-0.000	0.636	-0.000
x7	0.000	-0.000	0.000	-0.000	0.000	-0.000	0.434

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

	x1	x2	x3	x4	x5	x6	x7	x8
x1	1.992	-0.000	-0.000	-0.000	0.000	-0.000	0.000	-0.000
x2	-0.000	1.853	-0.000	0.000	-0.000	0.000	-0.000	0.000
x3	-0.000	-0.000	0.982	0.000	-0.000	-0.000	0.000	0.000
x4	-0.000	0.000	0.000	0.858	0.000	0.000	-0.000	-0.000
x5	0.000	-0.000	-0.000	0.000	0.839	-0.000	0.000	-0.000
x6	-0.000	0.000	-0.000	0.000	-0.000	0.636	-0.000	-0.000
x7	0.000	-0.000	0.000	-0.000	0.000	-0.000	0.434	0.000
x8	-0.000	0.000	0.000	-0.000	-0.000	-0.000	0.000	0.405

Inferences:

1. The off-diagonal values are 0, and this is because the data is projected onto orthonormal vectors, hence the new attributes are independent of each other.
2. The diagonal elements represent the variance among the values of each attribute. This is highest for the attribute representing projection of whole data on most significant eigenvector and lowest on the attribute representing least significant attribute, and since all the attributes represent projection on orthonormal vectors, the non diagonal entries, i.e. the covariances of the attributes among each other are 0.
3. We see a decreasing trend in eigenvalues as we go down on the diagonal. This is because the attributes x1,x2,.. are arranged in decending order of the significance of the eigenvector on which they are projected on. That's why, lower the significance of the eigenvector which they represent, lower the variance, which is exactly what is observed in the covariance matrices.
4. From the magnitude of diagonal entries in the matrices, we see that the attribute x1 shows most variance, hence it represents variance in the data in the most better way.

5. From part(b), we saw a drastic decrease in eigenvalues after second eigenvalue. This means that the significant eigenvalues are the first and second eigenvalues only. So the first 2 eigenvectors can be used to have optimum dimensionality reduction and less reconstruction error.
6. We observe that magnitude of first diagonal entry is same in all the matrices, and it has to be same, because it represents variance in the data projected on first eigenvector in all the matrices, which should not change based on our choice of dimensionality: the projected data will be same, no matter how many other vectors we project the data on.
7. We observe that magnitude of second diagonal entry is same in all the matrices, and it has to be same, because it represents variance in the data projected on second eigenvector in all the matrices, which should not change based on our choice of dimensionality: the projected data will be same, no matter how many other vectors we project the data on.
8. We observe that magnitude of 3rd,4th,5th,6th,7th diagonal entries is same in all the matrices, and they have to be same, because they represent variances in the data projected on the respective eigenvectors in all the matrices, which should not change based on our choice of dimensionality: the projected data will be same, no matter how many other vectors we project the data on.

d.

Table 11 Covariance matrix for original data

The covariance matrix of original data is

	pregs	plas	pres	skin	test	BMI	pedi	Age
pregs	1.000	0.118	0.209	-0.097	-0.108	0.028	0.005	0.561
plas	0.118	1.000	0.205	0.060	0.180	0.228	0.082	0.274
pres	0.209	0.205	1.000	0.026	-0.051	0.272	0.022	0.326
skin	-0.097	0.060	0.026	1.000	0.473	0.374	0.153	-0.101
test	-0.108	0.180	-0.051	0.473	1.000	0.172	0.199	-0.074
BMI	0.028	0.228	0.272	0.374	0.172	1.000	0.124	0.078
pedi	0.005	0.082	0.022	0.153	0.199	0.124	1.000	0.036
Age	0.561	0.274	0.326	-0.101	-0.074	0.078	0.036	1.000

Inferences:

1. The off diagonal values in the original data are not 0, unlike the data obtained after dimensionality reduction. This is because the original data attributes are not necessarily orthonormal, so covariance between the attributes is not necessarily 0.

2. We observe that the diagonal values in the original data are all equal to one, which is expected because this is the standardised form of the original data.
3. Interestingly, sum of all the diagonal values in original data and the dimensionally reduced data for l = 8 is same- 8. So yes, there is a tradeoff between the diagonal entries of both the covariance matrices.