

Attribute normalization, standardization and dimension reduction of data

Student's Name: ISHAAN GUPTA Mobile No: 9179242114

Roll Number: B20292 Branch: MECHANICAL ENGINEERING

1 a.

Table 1 Minimum and maximum attribute values before and after normalization

S. No.	Attribute	Before no	rmalization	After normalization		
		Minimum	Maximum	Minimum	Maximum	
1	pregs	0	17	5	12	
2	plas	0	199	5	12	
3	pres (in mm Hg)	0	122	5	12	
4	skin (in mm)	0	99	5	12	
5	test (in mu U/mL)	0	846	5	12	
6	BMI (in kg/m²)	0	67.1	5	12	
7	pedi	0.078	2.420	5	12	
8	Age (in years)	21	81	5	12	

Inferences:

- 1. The need for outlier correction is that statistical patterns and conclusions might differ between analyses.
- 2. In min-max normalization minimum value is scaled to 5 and maximum value is scaled to 12 in this case.
- 3. All the minimum value changes to 5 and all the maximum value changes to 12.

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	3.845	3.370	0	1	
2	plas	120.895	31.973	0	1	
3	pres (in mm Hg)	69.105	19.365	0	1	
4	skin (in mm)	20.536	15.952	0	1	
5	test (in mu U/mL)	79.799	115.244	0	1	
6	BMI (in kg/m²)	31.993	7.884	0	1	
7	pedi	0.472	0.331	0	1	
8	Age (in years)	33.241	11.760	0	1	

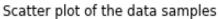


Attribute normalization, standardization and dimension reduction of data

Inferences:

1. In standardization mean is scaled to 0 and standard deviation is scaled to 12 in this case.

2 a.



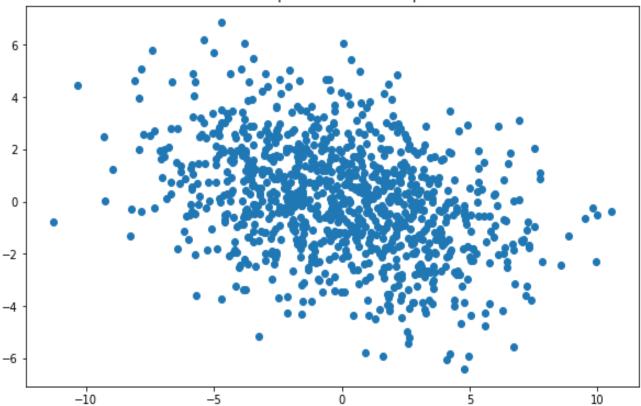


Figure 1 Scatter plot of 2D synthetic data of 1000 samples

Inferences:

- 1. Attribute 1 is negatively correlated to attribute 2.
- 2. The density of points is high on origin.

b.



Attribute normalization, standardization and dimension reduction of data

Eigen directions on scatter plot of the data samples

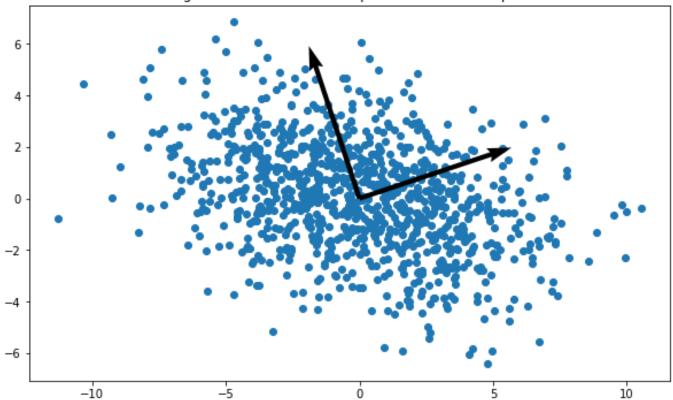


Figure 2 Plot of 2D synthetic data and Eigen directions

Inferences:

- 1. There is not much spreaded data based upon the magnitude of Eigenvalues.
- 2. The density of points is high on origin.

c.



Attribute normalization, standardization and dimension reduction of data

Projected values on eigenvector1

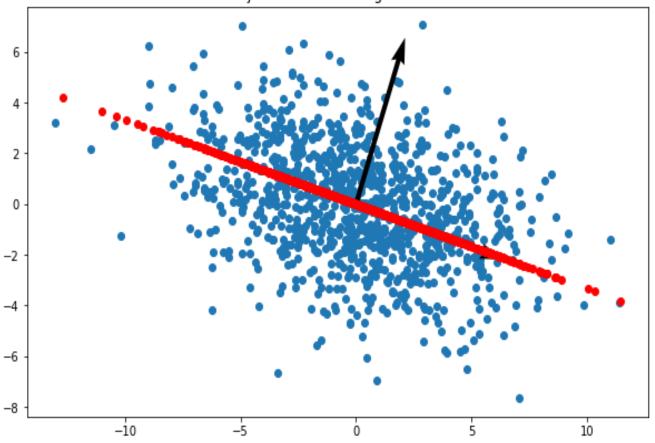


Figure 3 Projected Eigen directions onto the scatter plot with 1st Eigen direction highlighted



Attribute normalization, standardization and dimension reduction of data

Projected values on eigenvector2

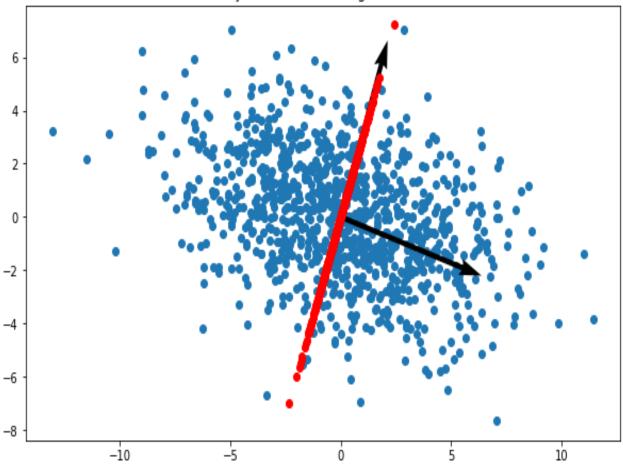


Figure 4 Projected Eigen directions onto the scatter plot with 2nd Eigen direction highlighted

Inferences:

- 1. Compare and contrast the magnitude of Eigenvalues
- 2. Infer variance of data along the Eigen axes from spread & density of points and relate it to the magnitude of Eigenvalues.

d. Reconstruction error = 3.1463713616948717e-16

- 1. Infer how the magnitude of reconstruction error affects the quality of reconstruction.
- 2. Inference 2(You may add or delete the number of inferences)



Attribute normalization, standardization and dimension reduction of data

3 a.

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

Direction	Variance	Eigenvalue
1	2.091	2.094
2	1.728	1.731

Inferences:

1. Variance 1 is more than Variance 2.

Scatter plot of reduced dimensional data

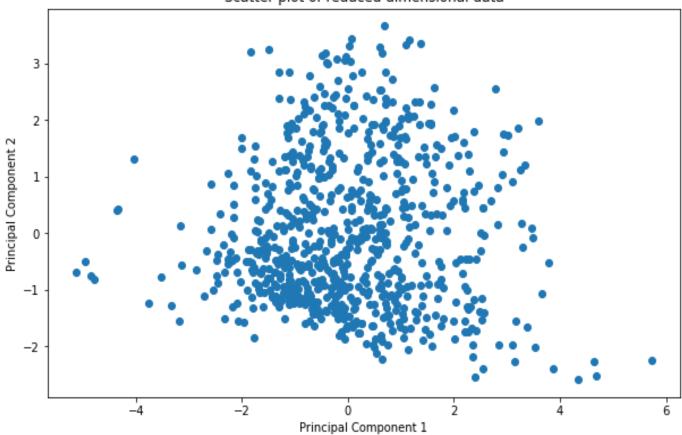


Figure 5 Plot of data after dimensionality reduction



Attribute normalization, standardization and dimension reduction of data

- 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.

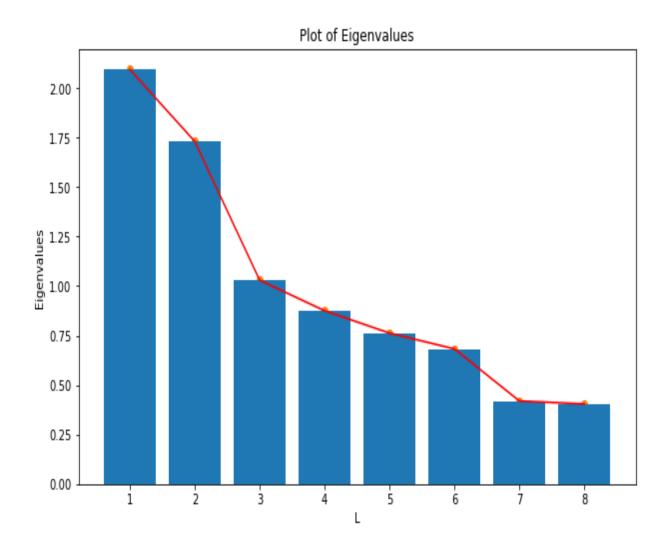


Figure 6 Plot of Eigenvalues in descending order



Attribute normalization, standardization and dimension reduction of data

Inferences:

- 1. Infer whether the subsequent Eigenvalues decrease gradually or rapidly
- 2. Identify the Eigenvalue from where the rate of decrease changes substantially
- 3. Inference 3(You may add or delete the number of inferences)

 Note: The plot above is for illustration purposes. Replace it with the plot obtained by you. Rename x-axis legend with Eigenvalues and y-axis legend with magnitude.

c.

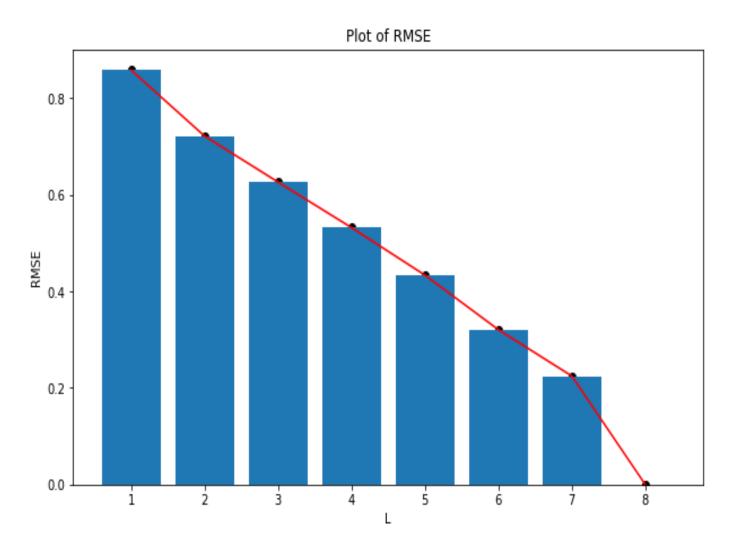


Figure 7 Line plot to demonstrate reconstruction error vs. components



Attribute normalization, standardization and dimension reduction of data

Inferences:

- 1. Infer how the magnitude of reconstruction error affects the quality of reconstruction.
- 2. Inference 2(You may add or delete the number of inferences)

 Note: The plot above is for illustration purposes. Replace it with the plot obtained by you. Rename x-axis legend with No. of components and y-axis legend with Reconstruction error.

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

	0	1
0	2.094379945288804	1.4683314682917324e-15
1	1.4683314682917324e-15	1.7312101406197233

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

	0	1	2
0	2.0943799452888046	-	-2.408619443254577e-
		2.0380626058307957e-	16
		16	
1	-	1.731210140619726	-
	2.0380626058307957e-		3.5318698566954133e-
	16		17
2	-2.408619443254577e-	-	1.029629869184153
	16	3.5318698566954133e-	
		17	

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

	0	1	2	3
0	2.094379945288807	-3.242372327458084e-	4.6319604677972634e-	1.5748665590510694e-
		17	17	16
1	-3.242372327458084e-	1.7312101406197253	1.030611204084891e-	-
	17		16	1.1579901169493159e-
				16
2	4.6319604677972634e-	1.030611204084891e-	1.029629869184154	5.141476119254962e-
	17	16		16
3	1.5748665590510694e-	-	5.141476119254962e-	0.8755290438080346
	16	1.1579901169493159e-	16	
		16		

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

	0	1	2	3	4
--	---	---	---	---	---



Attribute normalization, standardization and dimension reduction of data

0	2.0943799452888	1.2043097216272	-	6.4847446549161	9.2639209355945
	05	884e-16	2.7791762806783	68e-17	27e-17
			576e-17		
1	1.2043097216272	1.7312101406197	7.1795387250857	-	-
	884e-16	218	58e-17	7.4111367484756	1.5285469543730
				21e-17	968e-16
2	-	7.1795387250857	1.0296298691841	-	1.4011680415086
	2.7791762806783	58e-17	54	1.0190313029153	722e-16
	576e-17			979e-16	
3	6.4847446549161	-	-	0.8755290438080	-
	68e-17	7.4111367484756	1.0190313029153	354	1.8991037917968
		21e-17	979e-16		78e-16
4	9.2639209355945	-	1.4011680415086	-	0.7623443855511
	27e-17	1.5285469543730	722e-16	1.8991037917968	708
		968e-16		78e-16	

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

	0	1	2	3	4	5
0	2.0943799452	-	5.5583525613	4.6319604677	4.3077232350	-
	888046	4.0298056069	56715e-17	97263e-18	51455e-16	2.2233410245
		83619e-16				42686e-16
1	-	1.7312101406	8.6849258771	-	3.7055683742	-
	4.0298056069	197244	19868e-17	4.6319604677	378105e-17	3.7055683742
	83619e-16			97263e-18		378105e-17
2	5.5583525613	8.6849258771	1.0296298691	-	8.8586243946	-
	56715e-17	19868e-17	841534	3.9371663976	62266e-17	4.6840700230
				276737e-16		59982e-16
3	4.6319604677	-	-	0.8755290438	-	1.7717248789
	97263e-18	4.6319604677	3.9371663976	080335	5.0719967122	32453e-16
		97263e-18	276737e-16		380035e-16	
4	4.3077232350	3.7055683742	8.8586243946	-	0.7623443855	4.3424629385
	51455e-16	378105e-17	62266e-17	5.0719967122	511717	59934e-19
				380035e-16		
5	-	-	-	1.7717248789	4.3424629385	0.6826283879
	2.2233410245	3.7055683742	4.6840700230	32453e-16	59934e-19	464935
	42686e-16	378105e-17	59982e-16			

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

	0	1	2	3	4	5	6
--	---	---	---	---	---	---	---



Attribute normalization, standardization and dimension reduction of data

			1		1	1	
0	2.09437994	5.46571335	-	4.16876442	-	-	1.55170675
	52888037	2000771e-	2.22334102	1017537e-	9.26392093	3.93716639	6712083e-
		16	4542686e-	17	5594526e-	76276737e-	16
			16		18	16	
1	5.46571335	1.73121014	2.14228171	1.85278418	3.75188797	-	7.87433279
	2000771e-	06197242	63562342e-	71189053e-	89157833e-	7.87433279	5255347e-
	16		17	17	16	5255347e-	17
						17	
2	-	2.14228171	1.02962986	-	-	-	3.12657331
	2.22334102	63562342e-	91841554	3.70556837	3.02235420	1.81225453	57631526e-
	4542686e-	17		4237811e-	52377144e-	30256792e-	17
	16			16	16	16	
3	4.16876442	1.85278418	-	0.87552904	5.55835256	2.87471046	-
	1017537e-	71189053e-	3.70556837	38080354	1356715e-	5326676e-	3.01077430
	17	17	4237811e-		17	16	4068221e-
			16				17
4	-	3.75188797	-	5.55835256	0.76234438	-	-
	9.26392093	89157833e-	3.02235420	1356715e-	55511708	1.88173394	9.95871500
	5594526e-	16	52377144e-	17		00426382e-	5764116e-
	18		16			18	17
5	-	-	-	2.87471046	-	0.68262838	-
	3.93716639	7.87433279	1.81225453	5326676e-	1.88173394	79464933	2.08438221
	76276737e-	5255347e-	30256792e-	16	00426382e-		05087685e-
	16	17	16		18		17
6	1.55170675	7.87433279	3.12657331	-	-	-	0.41981617
	6712083e-	5255347e-	57631526e-	3.01077430	9.95871500	2.08438221	97057532
	16	17	17	4068221e-	5764116e-	05087685e-	
				17	17	17	

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

	0	1	2	3	4	5	6	7
0	2.0943799	5.4657133	-	4.1687644	-	-	1.5517067	8.3375288
	45288803	52000771	2.2233410	21017537	9.2639209	3.9371663	56712083	42035074
	7	e-16	24542686	e-17	35594526	97627673	e-16	e-17
			e-16		e-18	7e-16		



Attribute normalization, standardization and dimension reduction of data

1	5.4657133	1.7312101	2.1422817	1.8527841	3.7518879	-	7.8743327	-
	52000771	40619724	16356234	87118905	78915783	7.8743327	95255347	2.4317792
	e-16	2	2e-17	3e-17	3e-16	95255347	e-17	45593563
						e-17		2e-17
2	-	2.1422817	1.0296298	-	-	-	3.1265733	2.1770214
	2.2233410	16356234	69184155	3.7055683	3.0223542	1.8122545	15763152	19864713
	24542686	2e-17	4	74237811	05237714	33025679	6e-17	7e-16
	e-16			e-16	4e-16	2e-16		
3	4.1687644	1.8527841	-	0.8755290	5.5583525	2.8747104	-	-
	21017537	87118905	3.7055683	43808035	61356715	65326676	3.0107743	3.2423723
	e-17	3e-17	74237811	4	e-17	e-16	04068221	27458084
			e-16				e-17	e-17
4	-	3.7518879	-	5.5583525	0.7623443	-	-	-
	9.2639209	78915783	3.0223542	61356715	85551170	1.8817339	9.9587150	1.1579901
	35594526	3e-16	05237714	e-17	8	40042638	05764116	16949315
	e-18		4e-16			2e-18	e-17	9e-17
5	-	-	-	2.8747104	-	0.6826283	-	-
	3.9371663	7.8743327	1.8122545	65326676	1.8817339	87946493	2.0843822	1.6211861
	97627673	95255347	33025679	e-16	40042638	3	10508768	63729042
	7e-16	e-17	2e-16		2e-18		5e-17	e-17
6	1.5517067	7.8743327	3.1265733	-	-	-	0.4198161	-
	56712083	95255347	15763152	3.0107743	9.9587150	2.0843822	79705753	8.1059308
	e-16	e-17	6e-17	04068221	05764116	10508768	2	1864521e-
				e-17	e-17	5e-17		18
7	8.3375288	-	2.1770214	-	-	-	-	0.4044620
	42035074	2.4317792	19864713	3.2423723	1.1579901	1.6211861	8.1059308	47895868
	e-17	45593563	7e-16	27458084	16949315	63729042	1864521e-	3
		2e-17		e-17	9e-17	e-17	18	

- 1. Off-diagonal elements tend to 0 as eigen vectors are orthonormal.
- 2. Attributes are uncorrelated.
- 3. Diagonal values are in decreasing order.



Attribute normalization, standardization and dimension reduction of data

d.

Table 11 Covariance matrix for original data

	pregs	plas	pres	skin	test	BMI	pedi	Age
pr	1.0000000	0.129458	0.1412819	-	-	0.0176830	-	0.5443412
eg	00000002	67149927	77407140	0.081671	0.073534	90727830	0.0335226	28402340
S	9	276	27	77444900	61435162	645	72962613	1
				726	811		2	
pl	0.1294586	0.999999	0.1525895	0.057327	0.331357	0.2210710	0.1373372	0.2635143
as	71499272	99999999	86568664	89073817	10992020	69458983	99828370	19824333
	76	93	42	688	867	08	67	6
pr	0.1412819	0.152589	1.0000000	0.207370	0.088933	0.2818052	0.0412649	0.2395279
es	77407140	58656866	00000001	53840307	37837319	88849910	47930098	46421363
	27	442	3	038	289	9	536	66
sk	-	0.057327	0.2073705	0.999999	0.436782	0.3925732	0.1839275	-
in	0.0816717	89073817	38403070	99999999	57012001	04159037	72954162	0.1139702
	74449007	688	38	62	25	9	76	62367741
	26							38
te	-	0.331357	0.0889333	0.436782	0.999999	0.1978590	0.1850709	-
st	0.0735346	10992020	78373192	57012001	9999999	56493100	29168098	0.0421629
	14351628	867	89	25	53	82	75	54735376
	11							79
В	0.0176830	0.221071	0.2818052	0.392573	0.197859	1.0000000	0.1406469	0.0362418
М	90727830	06945898	88849910	20415903	05649310	00000001	52545105	70092294
1	645	308	9	79	082	8	34	085
р	-	0.137337	0.0412649	0.183927	0.185070	0.1406469	1.0000000	0.0335613
е	0.0335226	29982837	47930098	57295416	92916809	52545105	00000000	12434805
di	72962613	067	536	276	875	34	9	576
	2							
Α	0.5443412	0.263514	0.2395279	-	-	0.0362418	0.0335613	1.0000000
ge	28402340	31982433	46421363	0.113970	0.042162	70092294	12434805	0000001
	1	36	66	26236774	95473537	085	576	6
				138	679			

- 1. Off-diagonal values are weakly correlated, therefore, they are non 0.
- 2. The magnitudes of diagonal values are 1 as they are standardized.