Dimensionality Reduction of Air pollution dataset

2024-06-05

sample variance covariance matirx

```
           V1          V2         V3         V4         V5         V6        V7
V1  2.5000000  -2.7804878 -0.3780488 -0.4634146 -0.5853659 -2.2317073 0.1707317
V2 -2.7804878 300.5156794  3.9094077 -1.3867596  6.7630662 30.7909408 0.6236934
V3 -0.3780488   3.9094077  1.5220674  0.6736353  2.3147503  2.8217189 0.1416957
V4 -0.4634146  -1.3867596  0.6736353  1.1823461  1.0882695 -0.8106852 0.1765389
V5 -0.5853659   6.7630662  2.3147503  1.0882695 11.3635308  3.1265970 1.0441347
V6 -2.2317073  30.7909408  2.8217189 -0.8106852  3.1265970 30.9785134 0.5946574
V7  0.1707317   0.6236934  0.1416957  0.1765389  1.0441347  0.5946574 0.4785134
```

Prinicipal components

```
             [,1]        [,2]        [,3]          [,4]          [,5]
[1,] -0.010039244  0.07622439  0.03087761  0.9203045748  0.3423859285
[2,]  0.993199405  0.11615518  0.00659069 -0.0002118679  0.0022391022
[3,]  0.014062314 -0.09956775 -0.18282641 -0.1382922410  0.6500776063
[4,] -0.004710175  0.01320423 -0.13021553 -0.3277842624  0.6431560485
[5,]  0.024255644 -0.15038113 -0.95526318  0.1023719020 -0.2065840405
[6,]  0.112429558 -0.97335904  0.16981025  0.0632480276 -0.0002935726
             [,6]         [,7]
[1,]  0.011779079 -0.169729925
[2,]  0.003353218 -0.001781987
[3,] -0.563893916  0.443577538
[4,]  0.497513370 -0.462855916
[5,] -0.009009299 -0.105029951
[6,]  0.051067254 -0.066992404
```

Proportion of variation explained by components

```
[1] 0.8729480 0.9540751 0.9869680 0.9942105 0.9978816 0.9993986 1.0000000
```

Rotation matrix and Rotated Matrix

We print the first 6 rotation matrix

```
            PC1         PC2         PC3           PC4           PC5
V1 -0.010039244  0.07622439 -0.03087761  0.9203045748  0.3423859285
V2  0.993199405  0.11615518 -0.00659069 -0.0002118679  0.0022391022
V3  0.014062314 -0.09956775  0.18282641 -0.1382922410  0.6500776063
V4 -0.004710175  0.01320423  0.13021553 -0.3277842624  0.6431560485
V5  0.024255644 -0.15038113  0.95526318  0.1023719020 -0.2065840405
V6  0.112429558 -0.97335904 -0.16981025  0.0632480276 -0.0002935726
V7  0.002340785 -0.02382046  0.08519558  0.1095073458  0.0619613872
            PC6         PC7
V1  0.011779079  0.169729925
V2  0.003353218  0.001781987
V3 -0.563893916 -0.443577538
V4  0.497513370  0.462855916
V5 -0.009009299  0.105029951
V6  0.051067254  0.066992404
V7  0.657012233 -0.738019426
```

Rotated Matrix

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
[1,]	98.53743	1.68335056	-10.922375	7.671521	6.437940	-0.91493174
[2,]	107.03165	6.31183037	-8.204509	6.449050	5.491224	1.82350163
[3,]	103.07426	5.47537509	-4.240009	6.103658	6.308310	1.89719321
[4,]	89.25190	-5.38624670	-5.721540	10.043097	6.748370	1.91041451
[5,]	91.69310	-0.35249529	-6.406308	6.073788	4.674973	1.52490326
[6,]	91.01811	-2.98783250	-10.100598	8.421807	5.242621	1.70432383
[7,]	85.40721	-6.72516384	-10.311113	8.710481	8.219120	2.37343650
[8,]	73.61844	-8.68060650	-19.012868	5.918690	4.251706	2.06081230
[9,]	82.94824	-2.85894196	-9.183337	7.555855	4.384082	0.46913544
[10,]	64.88189	-3.23817127	-11.736649	8.339944	4.978701	1.45492909
[11,]	71.09526	3.76399279	-10.080577	5.046172	6.155467	1.51348234
[12,]	91.45284	1.96605734	-10.736791	6.293531	3.849517	1.33566430
[13,]	73.08760	-4.25932545	-16.862196	6.951391	6.145520	0.68622319
[14,]	70.53124	-0.01792283	-9.796422	9.876827	5.378624	1.32137233
[15,]	72.78687	-2.26775044	-6.277805	10.086816	5.358817	1.01079506
[16,]	77.78716	-1.91358004	-7.230992	9.267824	4.821043	1.00677277
[17,]	76.41820	1.11487976	-5.867365	7.953243	4.890467	0.85645731
[18,]	71.34020	2.00377109	-15.535571	8.001552	5.019246	1.69355082
[19,]	67.04804	4.12342020	-12.606650	8.845661	4.617820	1.02617873
[20,]	69.25593	2.14994992	-8.210374	8.436002	5.440832	2.28353141
[21,]	62.33281	1.41158547	-13.622606	9.639324	6.097034	1.70494861
[22,]	88.20941	3.57152964	-6.057425	8.479973	5.903171	1.35492113
[23,]	80.98388	-3.22683866	-11.108752	8.601342	4.363862	2.17460901
[24,]	30.12700	0.83660181	-5.279299	4.163815	4.811180	1.98847502
[25,]	85.27872	-14.37275903	-6.299833	7.401450	4.308958	1.73953816
[26,]	84.23259	3.13025226	-5.931839	7.698808	4.901751	1.89362309
[27,]	78.96671	-3.28701510	-9.187966	6.446906	4.025818	1.50535078
[28,]	79.70696	-1.25759631	-4.972509	8.418936	3.596148	2.14750656
[29,]	62.68501	-1.91145443	-8.022537	5.728023	5.047198	1.81402949
[30,]	37.08713	1.70363935	-6.728872	9.624167	4.839304	1.05779760
[31,]	71.52497	0.08296047	-8.766108	8.261418	4.259519	0.81266332
[32,]	52.82750	-3.49815418	-10.748120	7.722638	3.523090	1.42723398
[33,]	48.33498	0.25723886	-8.495746	3.523168	7.467690	1.42728455
[34,]	77.42915	-16.15197133	-5.914726	7.495178	3.578713	1.67066135
[35,]	35.87521	-5.26755049	-4.695749	9.717155	5.627471	0.19666509
[36,]	85.74045	-1.03674251	-7.123949	8.236317	4.434608	0.36480720
[37,]	86.90180	-2.54526746	-3.821755	5.432864	3.378777	1.02587951
[38,]	87.79771	-9.82315639	-10.351257	5.648002	5.174394	-0.48884436
[39,]	81.50814	-16.69319469	-5.671538	6.977282	8.016046	1.55678822
[40,]	79.32667	2.14755578	-7.231390	6.495843	5.579780	0.07126855
[41,]	69.40013	-7.56777958	-9.110095	6.362184	5.302702	0.49723252
[42,]	40.40214	-0.91937872	-5.664223	6.966926	6.241381	0.98063075

	[,7]
[1,]	0.32659703
[2,]	-0.05921296
[3,]	0.30104238
[4,]	0.54501380
[5,]	0.37195260
[6,]	0.66176709
[7,]	1.00121454
[8,]	-0.35835603

```
 [9,]  0.74261175
[10,]  0.80404601
[11,]  0.18434497
[12,]  0.15281001
[13,] -0.40919830
[14,] -0.38365802
[15,]  0.18974656
[16,]  0.24553660
[17,]  0.82808563
[18,]  0.34858835
[19,] -0.08368001
[20,] -0.77453485
[21,]  0.23522629
[22,]  0.24110835
[23,]  0.19797187
[24,]  0.59497935
[25,]  0.94970021
[26,] -0.02561131
[27,]  0.01303617
[28,] -0.26539262
[29,] -0.01027084
[30,]  0.44950775
[31,]  0.52190571
[32,]  1.18646051
[33,]  0.51884971
[34,] -0.28463326
[35,] -0.20528704
[36,] -0.33700879
[37,] -0.09207350
[38,]  0.08221666
[39,] -0.48128872
[40,]  0.19713424
[41,]  0.71703390
[42,] -0.53247934
```

Find the rotation matrix and the rotated version of the air-pollution data.

correlation matrix

```
           V1          V2         V3          V4         V5         V6
V1  1.0000000 -0.10144191 -0.1938032 -0.26954261 -0.1098249 -0.2535928
V2 -0.1014419  1.00000000  0.1827934 -0.07356907  0.1157320  0.3191237
V3 -0.1938032  0.18279338  1.0000000  0.50215246  0.5565838  0.4109288
V4 -0.2695426 -0.07356907  0.5021525  1.00000000  0.2968981 -0.1339521
V5 -0.1098249  0.11573199  0.5565838  0.29689814  1.0000000  0.1666422
V6 -0.2535928  0.31912373  0.4109288 -0.13395214  0.1666422  1.0000000
V7  0.1560979  0.05201044  0.1660323  0.23470432  0.4477678  0.1544506
           V7
V1 0.15609793
V2 0.05201044
V3 0.16603235
V4 0.23470432
V5 0.44776780
```

```
V6 0.15445056
V7 1.00000000
```

Principal Components based on Correlation

```
           [,1]         [,2]       [,3]         [,4]        [,5]         [,6]
[1,]  0.2368211  0.278445138  0.6434744  0.172719491  0.56053441 -0.223579220
[2,] -0.2055665 -0.526613869  0.2244690  0.778136601 -0.15613432 -0.005700851
[3,] -0.5510839 -0.006819502 -0.1136089  0.005301798  0.57342221 -0.109538907
[4,] -0.3776151  0.434674253 -0.4070978  0.290503052 -0.05669070 -0.450234781
[5,] -0.4980161  0.199767367  0.1965567 -0.042428178  0.05021430  0.744968707
[6,] -0.3245506 -0.566973655  0.1598465 -0.507915905  0.08024349 -0.330583071
            [,7]
[1,] -0.24146701
[2,] -0.01126548
[3,]  0.58524622
[4,] -0.46088973
[5,] -0.33784371
[6,] -0.41707805
```

Proportion

```
[1] 0.3338261 0.5318262 0.7038356 0.8077051 0.9010589 0.9777287 1.0000000
```

Rotation Matrix

We print the first six rotation matrix

```
          PC1          PC2        PC3          PC4         PC5          PC6
V1 -0.2368211  0.278445138  0.6434744  0.172719491 -0.56053441  0.223579220
V2  0.2055665 -0.526613869  0.2244690  0.778136601  0.15613432  0.005700851
V3  0.5510839 -0.006819502 -0.1136089  0.005301798 -0.57342221  0.109538907
V4  0.3776151  0.434674253 -0.4070978  0.290503052  0.05669070  0.450234781
V5  0.4980161  0.199767367  0.1965567 -0.042428178 -0.05021430 -0.744968707
V6  0.3245506 -0.566973655  0.1598465 -0.507915905 -0.08024349  0.330583071
           PC7
V1 -0.24146701
V2 -0.01126548
V3  0.58524622
V4 -0.46088973
V5 -0.33784371
V6 -0.41707805
```

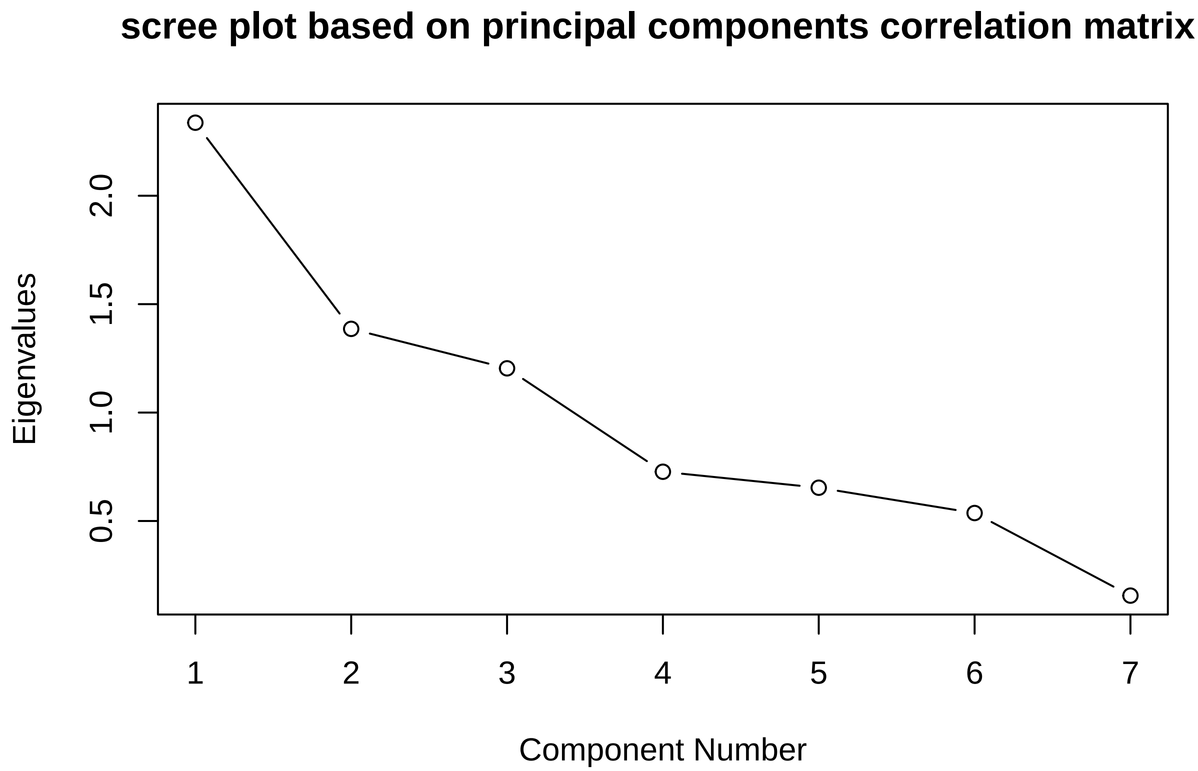
Rotated Matrix

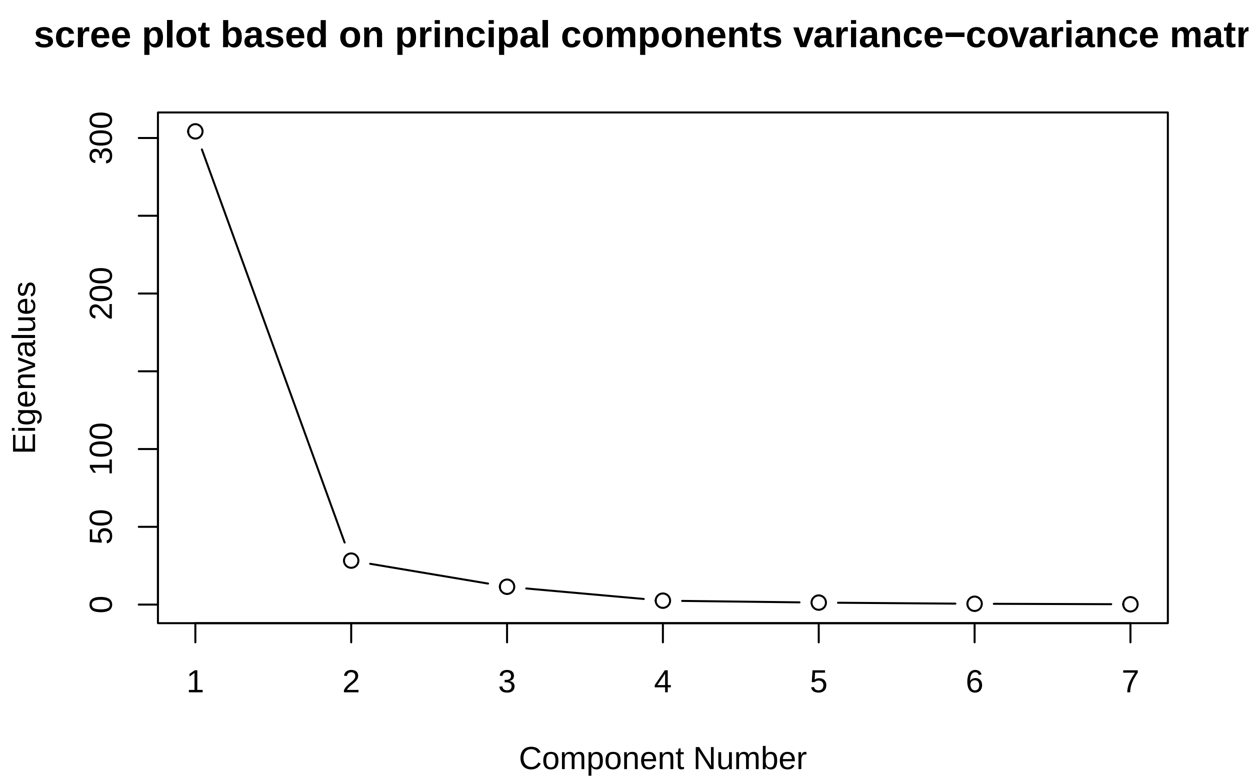
```
           PC1       PC2      PC3      PC4         PC5         PC6       PC7
 [1,] 32.07517 -50.08180 30.25585 73.39863  6.80393555 -1.74746119 -6.623730
 [2,] 30.73816 -53.23514 31.03816 82.01169 11.50408010 -0.28848797 -6.121602
 [3,] 28.24838 -52.49472 29.51390 78.56094 11.00015655  2.99916653 -5.142243
 [4,] 29.36230 -48.39725 30.94012 62.28024  6.03966191  4.25050687 -9.104846
 [5,] 28.43304 -48.55707 27.81296 66.60113  9.15877158  1.34436847 -6.986544
```

```
 [6,] 31.26548 -47.50738 30.40880 64.84511  7.51287261 -0.15687391  -8.744584
 [7,] 32.94572 -44.60658 29.68463 59.27379  5.30740937  2.41026661  -9.607065
 [8,] 34.71331 -37.33718 25.59883 49.50911  5.31160156 -5.96377064 -12.028687
 [9,] 28.33819 -43.94827 27.47922 58.85022  6.33202899 -0.72837889  -7.511095
[10,] 25.44512 -31.91473 24.28962 46.09487  3.64389646 -2.04181396  -7.538291
[11,] 24.35220 -32.79391 21.66996 55.09526  6.03732017 -1.56365900  -4.853909
[12,] 29.45145 -46.05708 28.11964 67.95516  9.19864482 -2.62725557  -7.086685
[13,] 31.67910 -35.42640 25.00207 52.16188  3.52265682 -4.76096927  -9.558445
[14,] 23.68925 -34.08418 25.78314 52.34760  3.72790077 -1.10768786  -7.478134
[15,] 23.20238 -37.87230 26.52904 52.21691  3.89339116  1.68013440  -7.277478
[16,] 24.96505 -40.58405 27.20447 55.89244  5.18438287  0.74009073  -7.430182
[17,] 23.02662 -39.03449 25.46387 56.55019  5.92994203  1.00899885  -5.250528
[18,] 27.13299 -31.73218 25.24424 54.24463  5.04410190 -5.43949470  -7.006180
[19,] 22.68265 -29.54838 24.06014 52.29520  4.12082104 -4.49122240  -5.793168
[20,] 21.90191 -32.66010 23.90890 52.78264  5.02332935 -0.16750079  -6.761694
[21,] 23.81322 -26.83530 24.11786 47.67169  2.61825282 -3.55370651  -6.712037
[22,] 25.30966 -44.07377 28.23403 66.85897  7.37995363  1.42064000  -5.670993
[23,] 28.83220 -41.46768 28.31443 57.52393  6.55498080 -1.39897310  -9.137941
[24,] 11.86638 -12.33409 11.31494 23.43789  1.61781621  0.70397475  -2.753863
[25,] 32.49657 -51.44889 29.32286 53.25999  6.70206076  4.03217811 -11.634073
[26,] 24.17313 -42.23894 26.80629 63.56840  7.88937297  1.06471847  -5.969710
[27,] 27.57927 -41.67876 25.64438 55.85015  6.89813902 -0.63406565  -8.270702
[28,] 23.51480 -42.30162 26.84404 57.35025  7.30445894  1.79877281  -7.706051
[29,] 22.69814 -31.51688 20.77312 45.29907  4.79783968  0.22314372  -6.624423
[30,] 12.36204 -15.09797 17.31091 29.08258 -0.30572561 -0.52863017  -3.793965
[31,] 23.49283 -35.80212 24.93120 52.53222  4.99862752 -1.25441152  -6.207732
[32,] 21.46387 -25.93446 21.11682 36.83906  2.97800834 -2.47919145  -6.531073
[33,] 20.11821 -21.49889 15.10973 36.89814  2.51021885  0.46193638  -3.970370
[34,] 30.30610 -48.10402 27.25952 46.66476  5.38009435  3.94114565 -12.860187
[35,] 13.95643 -18.52803 17.12968 24.16171 -2.26897713  2.36208743  -6.081804
[36,] 26.52700 -45.38329 27.81570 62.08790  6.42792126  0.29564850  -7.592753
[37,] 26.04699 -48.47167 25.95338 61.45403  8.82923697  2.41714505  -7.285488
[38,] 34.06236 -50.06774 27.42682 58.42125  5.75927778  0.07425294 -10.272768
[39,] 33.50418 -49.41521 27.20204 50.37195  3.86387351  6.94240112 -12.852880
[40,] 24.66290 -40.00607 24.46876 59.61626  5.90610650  0.02027429  -5.153181
[41,] 27.59943 -38.12718 23.65201 46.55564  3.94922096  0.51975287  -8.238789
[42,] 14.91493 -18.58075 15.51149 30.31962  0.06711863  1.52157052  -4.908664
```

We can see that for PCA based on the covariance matrix , variables with larger variances will dominate the principal components, which may not always be good however PCA based on the correlation matrix standardizes the variables making it less sensitive to differences in scale among variables. Also, PCA based on the covariance matrix is sensitive to outliers but PCA based on the correlation matrix is less sensitive to outliers because it standardizes the variables, making them less affected by extreme values.

Scatterplot for marginal distribution of each pairs after outlier removed





Yes, the data can be summarized in three or fewer dimensions. As we can see from the two scree plots above.

For the scree plot of the correlation matrix, An elbow occurs in the plot at about $i = 3$. So three or fewer sample principal components can effectively summarize the total sample variance. The first three components accounts for about 70% of the total sample variance.

Using the scree plot and the proportion of variance explained, it appears as if 2 or 3 components can be retained. Three components will about explain 98% of the variability and two components explains about 95% of variablity in the data.
Clearly we can also see that it makes a difference whether the components obtained from the sample Correlation matrix or the sample covariance matrix should be used.

PART E

Interpret the principal components in parts (a) and (b).

For the principal components using the covariance matrix we have that 87% of the variation in the data is explained by the first component, 95% explains of the variation is explained by the first two principal component and 98% of the variation in the data can be explained by the first three components. Sample variation is summarized very well by the first three principal components.By summarizing the dataset with the first three principal components, we effectively reduce the dimensionality to 3 principal components.

Using the correlation matrix , we have that about 33% of the variation is the data can be explained by the first principal component and about 70% of the variation in the data can be explained by the first three principal component,

We can see that the principal components from the covariance matrix explains more variability in the data than from the correlation matrix. These results however can be affected by outliers and higher loadings.

APPENDIX

RCODE

```r
###  Construct a principal component analysis of these data using the sample variance-covariance matrix

data1= as.matrix(data)


### sample variance covariance matirx

sigma = var(data1);sigma
total.variance = sum(diag(sigma))


### Prinicipal components

# variance
lamba = eigen(sigma)$values
e = eigen(sigma)$vectors
head(e)

### Proportion of variation explained by components

##  finding t-squared
prop = cumsum(lamba)/sum(lamba);prop

### Rotation matrix and Rotated Matrix

##  finding t-squared
## Rotation matrix (Rotated)

pc = prcomp(data1)

## rotation matrix
rotat = pc$rotation


###  Rotated Matrix

##  finding t-squared
## Rotation matrix (Rotated)
head(data1%*%e)


###  Construct a principal component analysis of these data using the sample correlation matrix R: Find

## correlation matrix

cat=cor(data1);cat

###  Principal Components based on Correlation
```

```r
n= length(cat[,1])
p = length(cat[1,]) ## number of variables
xbar2<-cbind(apply(cat,2,mean))
S1= var(cat)
lambda2 = eigen(cat)$values
e2=eigen(cat)$vectors
head(e2)


###  Proportion

prop2 = cumsum(lambda2)/sum(lambda2);prop2


## Rotation Matrix

pca_cor <- prcomp(data, scale = TRUE)

# Rotation matrix
rotation_matrix_cor <- pca_cor$rotation;
head(rotation_matrix_cor)


### Rotated Matrix

##  finding t-squared
# Rotated data
rotated_data_cor <- data1%*% rotation_matrix_cor
head(rotated_data_cor)


## PART D
## Scatterplot for marginal distribution of each pairs after outlier removed

pca_cor <- prcomp(data, scale = TRUE)

# Rotation matrix
rotation_matrix_cor <- pca_cor$rotation

# Rotated data
rotated_data_cor <- data1%*% rotation_matrix_cor

plot(pca_cor$sdev^2, type = "b",main = "scree plot based on principal components correlation matrix ", 

plot(pca_cor$sdev^2, type = "b",main = "scree plot based on principal components correlation matrix ", 

pc = prcomp(data1)
plot(pc$sdev^2, type = "b",main = "scree plot based on principal components variance-covariance matrix 
```