Applied Multivariate-HW5

ID : 111024517 Name : 鄭家豪

due on 03/30

The code for the results is attached to the Rmd flie

1.

Data Resource:https://www.kaggle.com/datasets/hellbuoy/car-price-prediction

(a)

After standardizing each variable,each PC and the summary of PCA are the following:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
wheelbase	0.31	-0.28	0.11	-0.25		-0.08	0.11	-0.41	0.4	-0.37
carlength	0.35	-0.15	0.1	-0.16	-0.02	-0.01	0.12	-0.13	0.25	0.8
carwidth	0.35	-0.09	-0.08	-0.06	-0.23	-0.14	0.14	-0.45	-0.71	-0.12
carheight	0.12	-0.41	0.45	-0.42	0.11	0.11	-0.47	0.35	-0.24	-0.08
curbweight	0.37	-0.05	-0.06	0.03	-0.1	-0.1	-0.01	0.11	0.11	0.01
enginesize	0.33	0.05	-0.22	0.23	0.01	-0.3	-0.5		0.32	-0.22
boreratio	0.28	0.02	0.15	0.41	0.18	0.8	-0.07	-0.2	0.01	-0.08
stroke	0.06	-0.1	-0.77	-0.41	0.38	0.27	-0.04	0.05	-0.05	0.02
compressionratio	0.02	-0.5	-0.27	0.19	-0.63	0.21	0.17	0.37	0.07	-0.06
horsepower	0.3	0.3	-0.14	0.15	-0.19	-0.05	-0.41	0.15	-0.23	0.21
peakrpm	-0.09	0.46	0.03	-0.5	-0.55	0.32	-0.17	-0.15	0.18	-0.07
citympg	-0.32	-0.3	-0.08	0.1	-0.09	-0.02	-0.33	-0.36	0.06	
highwaympg	-0.33	-0.24	-0.09	0.11	-0.08	0.02	-0.37	-0.36	-0.05	0.3

	PC11	PC12	PC13
wheelbase	0.48	0.16	-0.1
carlength	-0.18	0.14	0.15
carwidth	-0.23	0.03	0.01
carheight	-0.07	0.02	
curbweight	0.03	-0.89	-0.11
enginesize	-0.48	0.23	-0.07

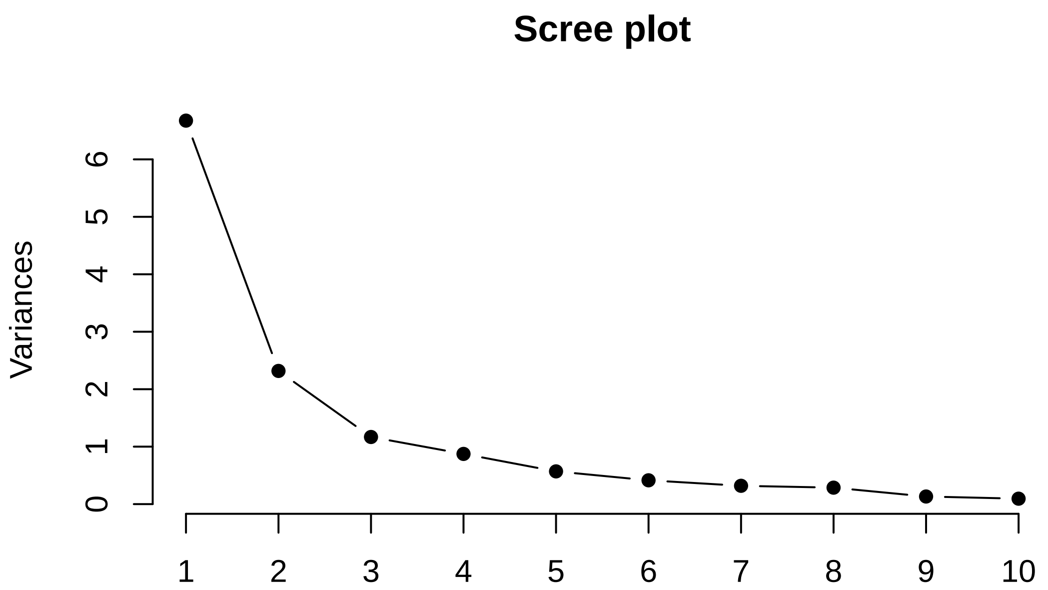
```
boreratio        -0.06 -0.02  0.01
stroke            0.02 -0.01  0.02
compressionratio        0.16 -0.02
horsepower        0.64  0.16  0.13
peakrpm          -0.16 -0.06 -0.02
citympg           0.03 -0.22   0.7
highwaympg        0.09 -0.07 -0.66

Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     2.5834 1.5225 1.08097 0.93479 0.75476 0.64388 0.56472
Proportion of Variance 0.5134 0.1783 0.08988 0.06722 0.04382 0.03189 0.02453
Cumulative Proportion  0.5134 0.6917 0.78157 0.84878 0.89260 0.92450 0.94903
                           PC8     PC9    PC10    PC11    PC12    PC13
Standard deviation     0.53549 0.36382 0.30765 0.27973 0.22585 0.14012
Proportion of Variance 0.02206 0.01018 0.00728 0.00602 0.00392 0.00151
Cumulative Proportion  0.97108 0.98127 0.98855 0.99457 0.99849 1.00000
```

The first row presents the standard deviation of each PC.

The second row presents the proportion of its explained variance.

And the last row shows the cumulative explained proportion.

(b)



Observe the scree plot,the slopes after PC 3 does not decrease sharply .And see cumulative explained proportion from (a),PC1 PC2 PC3 have accumulated nearly 80% of the variation explained. Thus,the proper number is 3.

To interpret the PCs, we can examine the variables with high positive or negative loadings (i.e.relatively larger |coefficient|) for each PC. These variables represent the features of the data that contribute the most to that PC.

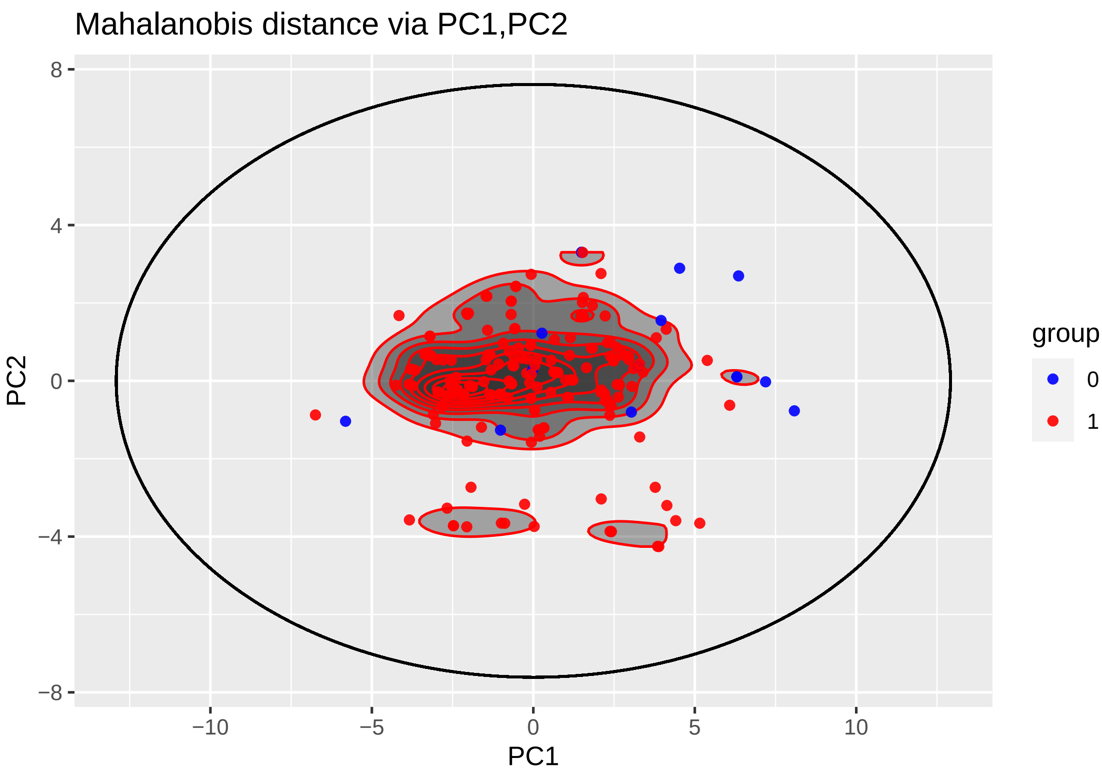
(c)

I calculate the mahalanobis distances of all observations,and calculate the total number of distances less than 25.

```
The proportion : 0.9268293
```

(d)

Here I draw the contour and the ellipse of $(v_j - \bar{v})'\Lambda^{-1}(v_j - \bar{v})$, as well as highlight the points selected in (c) with red and highlight the points not selected in (c) with blue. By the way,the demonstrating of contour plot is using the values of $(v_j - \bar{v})'\Lambda^{-1}(v_j - \bar{v})$.

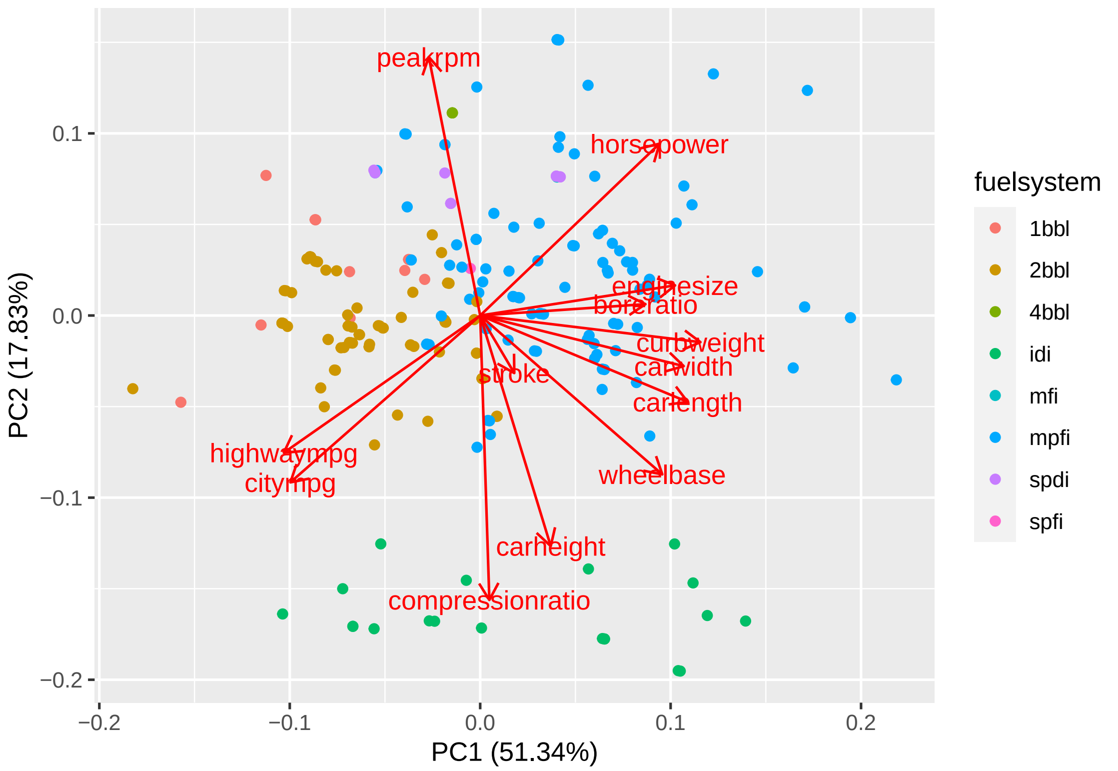


The contour across PC1 PC2 dimensions is consistent,it suggests that the structure of the data is extremely well-captured by first two PC,although its Mahalanobis distance via PC1 PC2 is not well captured.

(e)

```
Cor(PC1,Price)= 0.8380415

Cor(PC2,Price)= 0.1125982
```

(f)



From the above results, different fuel systems show distinct patterns in the biplot. For example, the fuel system "idi" is mainly located at the lower end of PC2. Furthermore, when observing the vectors of the continuous variables, compressionratio and carheight have a larger proportion of explaining the variation in PC2. This implies that the fuel system "idi" is more affected by compressionratio and carheight. Similar patterns can be observed for other fuel systems, and the biplot can be used to interpret these results.

2.

2\. (a)

By SVD, $\underset{n\times p}{X_C} = U_{n\times n}\Lambda^*_{n\times p}V^T_{p\times p}$.

where $Col(U)$ is the eigenvector set of $X_C X_C^T$

$Col(V)$ is the eigenvector set of $X_C X_C$

$\Lambda^* = \left[\begin{array}{c} diag(\lambda_i^*) \\ \hline 0_{(n-p)\times p}\end{array}\right]$, where $\frac{(\lambda_i^*)^2}{n-1} = \lambda_i$: eigenvalue of S,

denoted Λ as $diag(\lambda_i)$.

$(\Lambda^{*T}\Lambda^* = (n-1)\Lambda)$

The PC $Y = U\Lambda^*\Lambda^{-\frac{1}{2}} = \sqrt{n-1}\,U$, and

$$X_C = \sqrt{n-1}\,U\Lambda^* V^T\frac{1}{\sqrt{n-1}} \Rightarrow \sqrt{n-1}\,U = \sqrt{n-1}\,X_C V(\Lambda^*)^{-1}$$

So,
$$\begin{aligned}
\|Y_i - Y_j\|_2^2 &= \left(\sqrt{n-1}\,(X_i - X_j)^T V(\Lambda^*)^{-1}\right)\left(\sqrt{n-1}\,(X_i - X_j)^T V(\Lambda^*)^{-1}\right)^T \\
&= (n-1)(X_i - X_j)^T V(\Lambda^{*T}\Lambda^*)^{-1}V^T(X_i - X_j) \\
&= (X_i - X_j)^T\left[(n-1)V(\Lambda^{*T}\Lambda^*)^{-1}V^T\right](X_i - X_j) \\
&= (X_i - X_j)^T\left[(n-1)V((n-1)\Lambda)^{-1}\Lambda^T\right](X_i - X_j) \\
&= (X_i - X_j)^T S^{-1}(X_i - X_j).
\end{aligned}$$

(b) From (a), we know the squared Euclidean distance via PC space is the same as the Mahalanobis distance via original space.

That is noted that $\bar{Y} = 0_{p\times 1}$

$$\|Y_i - \bar{Y}\|_2^2 = \|Y_i - 0_{p\times 1}\|_2^2 = (X_i - \bar{X})^T S^{-1}(X_i - \bar{X})$$

This does reflect the variance of that variable.

(c) For any $\vec{x}, \vec{y} \in \mathbb{R}^p$, the angle between $\vec{x}, \vec{y}$ is defined by

$$\cos\theta = \frac{\langle\vec{x}, \vec{y}\rangle}{\|\vec{x}\|_2\cdot\|\vec{y}\|_2}, \quad \langle a, b\rangle = E(ab),\ \|a\|_2 = \sqrt{\langle a, a\rangle}$$

$$\rho_{X,Y} = \frac{E((X-u_X)(Y-u_Y))}{\sqrt{E((X-u_X)^2)E((Y-u_Y)^2)}} = \frac{\langle X-u_X, Y-u_Y\rangle}{\|X-u_X\|_2\cdot\|Y-u_Y\|_2} = \cos\theta$$

So, we have the concept of correlation as the cosine of an angle.

Thus in biplot.

the angle between the variables and the inner products between observations and variables reflects how about both are related. #

3.

`The mean expression profiles :(round to 3 decimals)`

	AURKA	AURKB	BUB1	CENPA	CENPF	KIF2C	PLK1	TTK	ERBB2	ESR1	PGR
n1p	9.650	8.689	7.508	9.426	9.125	9.671	9.161	8.934	8.486	7.357	4.248
T1p	8.427	8.670	6.939	9.068	8.257	9.081	8.697	8.934	8.486	6.597	4.574
n2p	8.994	8.557	7.185	9.093	8.225	9.547	8.362	8.747	9.611	8.249	5.455
T2p	8.810	8.551	7.244	9.041	8.604	9.465	8.508	8.615	8.922	6.727	3.980
n3p	8.976	8.119	7.102	8.948	8.693	9.353	8.609	8.295	9.382	8.605	4.579
T3p	8.969	8.394	7.148	9.277	8.940	9.691	8.558	9.215	8.916	6.717	3.701
n4p	9.238	8.192	7.150	9.222	8.523	9.185	8.437	8.293	10.570	8.369	5.121
T4p	8.538	8.450	7.270	9.175	8.482	9.380	8.535	8.679	9.092	6.411	3.470
n1R	7.834	7.257	6.389	8.368	7.479	8.603	6.840	6.651	9.540	9.347	6.528
T1R	8.651	8.610	6.762	9.042	8.205	9.060	7.800	8.374	9.012	6.569	4.265
n2R	8.077	7.599	6.521	8.371	7.892	8.757	7.802	7.092	9.715	9.297	6.244
T2R	8.589	8.045	7.024	8.854	8.498	9.331	8.323	8.342	9.156	7.398	4.385
n3R	8.284	7.602	6.540	8.446	8.066	8.910	7.713	7.437	9.550	9.383	6.017
T3R	8.855	8.402	7.208	9.132	8.851	9.507	8.470	8.641	8.914	7.246	4.054
n4R	8.418	7.851	6.761	8.573	8.038	8.883	7.844	7.798	9.680	9.464	5.467
T4R	8.920	8.310	7.079	9.041	8.334	9.477	8.231	8.574	8.764	6.875	4.037

The above is the mean express profiles,for simplicity,the values of table with dimesion 16x11 are rounded to the third decimal place.

Columns are 11 genes.

Rows are the combinations of TNBC Status, STAGE, and pCR:

- nip: (nonTNBC,Ti,pCR) , i=1,2,3,4.
- TiR: (TNBC,Ti,RD), i=1,2,3,4.

Next , the following is the distance matrix:

```
The distance matrix via 1-|correlation|:(round to 3 decimals)
```

	n1p	T1p	n2p	T2p	n3p	T3p	n4p	T4p	n1R	T1R	n2R	T2R	n3R	T3R	n4R	T4R
n1p	0.000	0.040	0.111	0.023	0.092	0.022	0.181	0.036	0.718	0.058	0.528	0.047	0.444	0.020	0.349	0.022
T1p	0.040	0.000	0.103	0.020	0.141	0.020	0.192	0.020	0.756	0.033	0.564	0.057	0.493	0.035	0.391	0.028
n2p	0.111	0.103	0.000	0.071	0.036	0.080	0.039	0.079	0.377	0.053	0.242	0.030	0.187	0.057	0.122	0.052
T2p	0.023	0.020	0.071	0.000	0.086	0.007	0.130	0.003	0.659	0.017	0.480	0.020	0.411	0.006	0.316	0.005
n3p	0.092	0.141	0.036	0.086	0.000	0.093	0.046	0.097	0.372	0.101	0.215	0.031	0.157	0.054	0.095	0.073
T3p	0.022	0.020	0.080	0.007	0.093	0.000	0.147	0.010	0.685	0.027	0.507	0.025	0.426	0.008	0.331	0.008
n4p	0.181	0.192	0.039	0.130	0.046	0.147	0.000	0.132	0.293	0.107	0.167	0.063	0.136	0.110	0.088	0.117
T4p	0.036	0.020	0.079	0.003	0.097	0.010	0.132	0.000	0.673	0.021	0.493	0.024	0.428	0.013	0.328	0.012
n1R	0.718	0.756	0.377	0.659	0.372	0.685	0.293	0.673	0.000	0.591	0.042	0.500	0.063	0.598	0.123	0.615
T1R	0.058	0.033	0.053	0.017	0.101	0.027	0.107	0.021	0.591	0.000	0.440	0.032	0.377	0.025	0.292	0.016
n2R	0.528	0.564	0.242	0.480	0.215	0.507	0.167	0.493	0.042	0.440	0.000	0.333	0.012	0.424	0.044	0.450
T2R	0.047	0.057	0.030	0.020	0.031	0.025	0.063	0.024	0.500	0.032	0.333	0.000	0.271	0.009	0.196	0.016
n3R	0.444	0.493	0.187	0.411	0.157	0.426	0.136	0.428	0.063	0.377	0.012	0.271	0.000	0.350	0.018	0.376
T3R	0.020	0.035	0.057	0.006	0.054	0.008	0.110	0.013	0.598	0.025	0.424	0.009	0.350	0.000	0.262	0.006
n4R	0.349	0.391	0.122	0.316	0.095	0.331	0.088	0.328	0.123	0.292	0.044	0.196	0.018	0.262	0.000	0.286
T4R	0.022	0.028	0.052	0.005	0.073	0.008	0.117	0.012	0.615	0.016	0.450	0.016	0.376	0.006	0.286	0.000

Finally,use this matrix to conduct MDS: