多變量分析 HW2

劉昱維, 吳冠瑋, 廖永賦, 謝靖惟, 黃奎鈞

Q1: 8.1

- First Principle Component: $\lambda_1 = 6, Y_1 = -0.89X_1 - 0.45X_2$
- Second Principle Component: $\lambda_2 = 1, Y_2 = 0.45X_1 - 0.89X_2$
- Proprotion of variance explained by the first PC: $\frac{\lambda_1}{\Sigma\lambda_i} = 85.71$%

Q2: 8.2

$$\rho = diag(\Sigma)^{\frac{-1}{2}}\,\Sigma\,diag(\Sigma)^{\frac{-1}{2}} = \begin{pmatrix} 1.00 & 0.63 \\ 0.63 & 1.00 \end{pmatrix}$$

(a)

- First Principle Component: $\lambda_1 = 1.63, Y_1 = -0.71Z_1 - 0.71Z_2$
- Second Principle Component: $\lambda_2 = 0.37, Y_2 = 0.71Z_1 - 0.71Z_2$
- Proprotion of variance explained by the first PC: $\frac{\lambda_1}{\Sigma\lambda_i} = 81.62$%

(b)

Principle components from 8.1 & 8.2 are different since they came from two different matrix (covariance matrix vs. correlation matrix). Mathmetically, the difference resulted from two different matrices have different eigenvalues and eigenvectors, hence different principle component. This difference reflects the fact that scaling has an effect on the principle component.

(c)

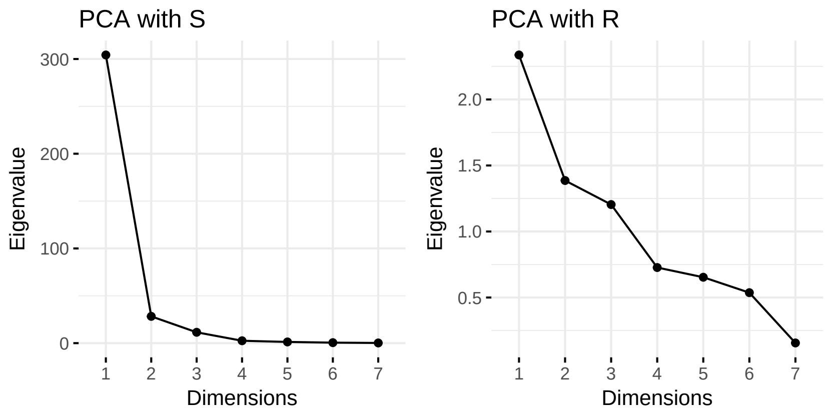
$\rho_{Y_i,\,Z_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}$

$\rho_{Y_1,\,Z_1} = \frac{(-0.71)(1.28)}{\sqrt{1}} = -0.9$

$\rho_{Y_1,\,Z_2} = -0.9$

$\rho_{Y_2,\,Z_1} = 0.43$

Q3: PCA with Covariance Matrix vs. Correlation Matrix



Principle Components: Covariance Matrix

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
wind	-0.01	0.08	-0.03	0.92	0.34	-0.01	-0.17
radiation	0.99	0.12	-0.01	0.00	0.00	0.00	0.00
CO	0.01	-0.10	0.18	-0.14	0.65	0.56	0.44
NO	0.00	0.01	0.13	-0.33	0.64	-0.50	-0.46
NO2	0.02	-0.15	0.96	0.10	-0.21	0.01	-0.11
O3	0.11	-0.97	-0.17	0.06	0.00	-0.05	-0.07
HC	0.00	-0.02	0.09	0.11	0.06	-0.66	0.74

Principle Components: Correlation Matrix

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
wind	-0.24	0.28	-0.64	0.17	-0.56	0.22	0.24
radiation	0.21	-0.53	-0.22	0.78	0.16	0.01	0.01
CO	0.55	-0.01	0.11	0.01	-0.57	0.11	-0.59
NO	0.38	0.43	0.41	0.29	0.06	0.45	0.46
NO2	0.50	0.20	-0.20	-0.04	-0.05	-0.74	0.34
O3	0.32	-0.57	-0.16	-0.51	-0.08	0.33	0.42
HC	0.32	0.31	-0.54	-0.14	0.57	0.27	-0.31

Interpretation

PCA with $\mathbf{S}$

The scree plot of PCA using $\mathbf{S}$ indicates that only one principle component is important, which explains 87.29% of the total variance.

The first principle component obtained by **covariance matrix** explains nearly all the variation in the data, this is probably due to the **large scale of the variable radiation** compared to orther variables. A better explanation of PCA should rely on the **correlation matrix** in this case.

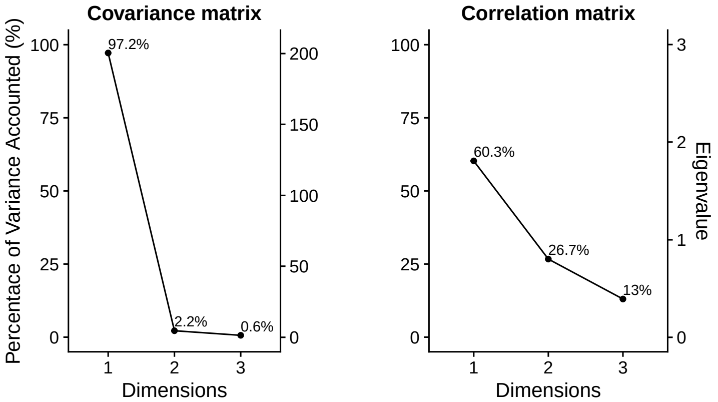
PCA with $\mathbf{R}$

According to the unity criterion and the scree plot (a not-so-obvious elbow at the forth PC), three principle components can well summarize the data, which explains 70.38% of the total variance.

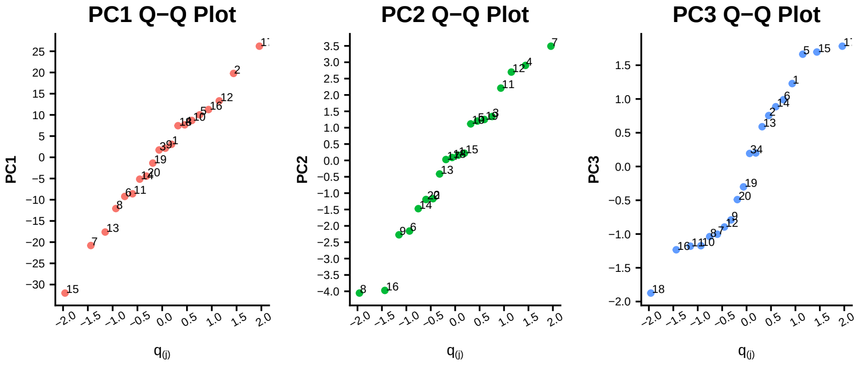
The interpretation of the first principle component obtained by the correlation matricies is straightforward: the variable **wind** constrasts with other variables consist of **pollutants**, probably because strong wind blows away pollutants, i.e. wind and polutants have opposite effect on air polution.

Interpretation of the remaining two principle components requires extensive knowledge on air pollution.

Q4



	PC1	PC2	PC3
sweet_rate	0.05	-0.57	-0.82
sodium	1.00	0.05	0.02
potassium	-0.03	0.82	-0.58

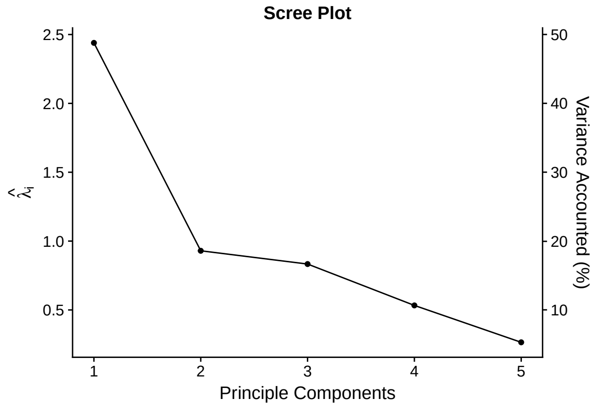


Suspect Observations?

From the three Q-Q plots above, there seems no obvious deviation from the straight lines, and there are no unique points that deviate from the main data in all three Q-Q plots. For detecting suspect observations, more techniques are needed.

Q5

(a) Scree Plot, PCs, Interpretation



	PC1	PC2	PC3	PC4	PC5
food1	-0.51	0.06	-0.40	-0.53	-0.54
food2	-0.52	-0.28	-0.41	0.07	0.69
food3	-0.40	0.10	0.77	-0.43	0.24
food4	-0.29	0.88	-0.07	0.37	0.05
food5	-0.48	-0.38	0.28	0.62	-0.41

Interpretation

By either the "elbow" of the scree plot or the unity criterion, the **first principle component** is sufficient, which explains **48.79%** of the total variance.

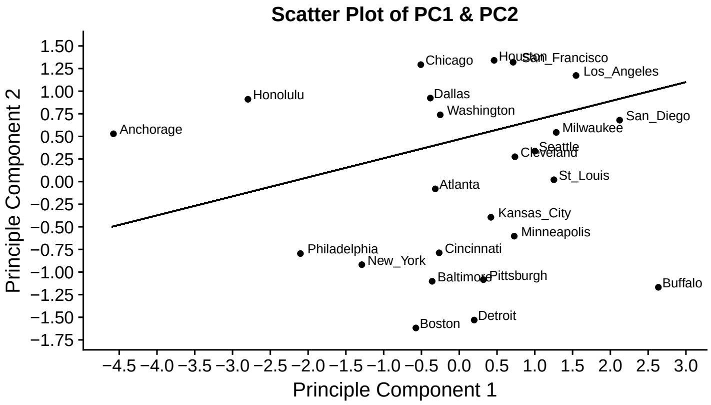
The data can be greatly reduced to 1 variable by the first principle component, and the "weight" of the first principle component on each variable have the same sign, indicating the price of each food has similar effect on the variation of the data.

(b) The First Two principal component scores

city	PC1	PC2
Anchorage	-4.58	0.53
Atlanta	-0.32	-0.08
Baltimore	-0.36	-1.10
Boston	-0.57	-1.62
Buffalo	2.64	-1.17
Chicago	-0.51	1.29
Cincinnati	-0.27	-0.79
Cleveland	0.74	0.27
Dallas	-0.38	0.92
Detroit	0.20	-1.53
Honolulu	-2.80	0.91
Houston	0.46	1.34
Kansas_City	0.42	-0.39
Los_Angeles	1.55	1.17
Milwaukee	1.28	0.54
Minneapolis	0.73	-0.60
New_York	-1.29	-0.92

city	PC1	PC2
Philadelphia	-2.10	-0.80
Pittsburgh	0.32	-1.08
St_Louis	1.25	0.02
San_Diego	2.12	0.68
San_Francisco	0.71	1.32
Seattle	1.00	0.34
Washington	-0.25	0.74

(c) Scatter Plot: PC1 vs. PC2



Two clusters seems to be identified on the scatter plot, ignoring **Buffalo**.

(d) Food Price in Different Cities

Using principle component 1 as indicator, since it has the same sign on all “weighted” food price, and it’s the most important principle component in explaining the variability of the data, the x-axis of the scatter above showed that **Buffalo** has the cheapest food price, and **Anchorage** has the most expensive food price.

Q6

(a)

Weight of the First 5 Pricinple Components on each Variable

	PC1	PC2	PC3	PC4	PC5
1	0.147	-0.092	0.024	-0.429	0.28
2	-0.26	0.126	-0.084	0.214	-0.311
3	-0.315	0.03	-0.059	0.12	-0.123
4	0.055	-0.116	-0.17	-0.375	0.15
5	0.38	-0.071	-0.133	-0.184	0.085
6	-0.122	-0.519	-0.235	0.002	-0.005
7	0.013	-0.218	0.3	-0.216	-0.447
8	-0.072	0.234	-0.393	-0.027	-0.195
9	0.24	0.055	-0.221	-0.324	-0.271
10	0.283	-0.13	-0.106	0.185	-0.128
11	-0.045	0.123	-0.498	0.012	0.073

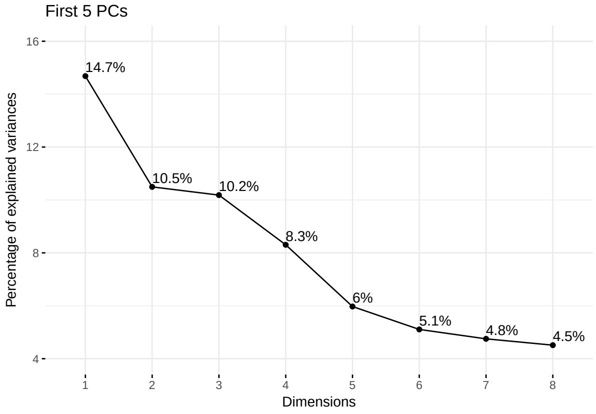
	PC1	PC2	PC3	PC4	PC5
12	-0.341	0.146	0.063	-0.345	-0.101
13	-0.363	0.134	-0.045	-0.298	0.079
14	0.027	0.225	-0.358	-0.117	-0.335
15	0.271	0.019	-0.007	-0.206	-0.231
16	0.177	-0.045	-0.122	0.127	-0.185
17	-0.209	-0.463	-0.188	0.024	0.016
18	-0.256	0.088	-0.017	-0.259	0.226
19	-0.181	-0.47	-0.222	-0.019	-0.09
20	-0.093	-0.117	0.318	-0.213	-0.413

Percentage of Variance Accounted

$\hat{\lambda}_1$ = 2.94, $\hat{\lambda}_2$ = 2.1, $\hat{\lambda}_3$ = 2.04, $\hat{\lambda}_4$ = 1.66, $\hat{\lambda}_5$ = 1.19.

The percentage of total variance accounted by the first 5 PCs is 49.64%.

(b)



First PC

The first PC group variables as:

- The variables 1, 4, 5, 7, 9, 10, 14, 15, 16 seem to reflect the trade-offs between cutting down energy usage and solving energy crysis.
- The variables 2, 3, 6, 8, 11, 12, 13, 17, 18, 19, 20 seem to reflect 2 attitudes:
    1. Little can be done to change the energy crysis.
    2. Favor solving energy crysis at the cost of pollution.

Second PC

The second PC group variables as:

- The variables 2, 3, 8, 9, 11, 12, 13, 14, 15, 18 mostly reflect the attitude that little can be done to solve the energy crysis.
- The variables 1, 4, 5, 6, 7, 10, 16, 17, 19, 20 seem to reflect that the energy crysis is changable despite some sacrifices such as pollution.

Third, Forth, and Fifth PC

The third, forth, and fifth PC are not obvious and easy to interpret, and in fact, they account for only 10%, 8%, and 6% of total variance, respectively.