Assignment 1 — Airpollution

Haoran Duan[1]

School of Computing, Newcastle University, UK

Abstract. Use some statistical method and PCA method to analyze the Airpollution data for the first assignment.

1 Introduction

Import necessary packages, and check the first few data to understand it.

Table 1. 6 Heads of Data

	SMI	SMN	SMX	PMI	PMN	PMX	PM2	PERWH	NONP	GE65	LPOP
PROVIDEN	30	163	349	56	119	223	116.10	97.90	83.90	109	58.56
JACKSON1	29	70	161	27	74	124	21.30	60.00	69.10	64	52.72
JOHNSTOW	88	123	245	70	166	452	15.80	98.70	73.30	103	54.48
JERSEYC	155	229	340	63	147	253	1357.20	93.10	87.30	103	57.86
HUNTINGT	60	70	137	56	122	219	18.10	97.00	73.20	93	54.06
DESMOIN	31	88	188	61	183	329	44.80	95.90	87.10	97	54.25

2 First Part

2.1 Numerical and Graphical Summaries

Get the data and make it as a matrix.

Simpley check the structure of the data R :

```
library(nclSLR)
data(airpollution)
air_Data = airpollution
air_Matrix = as.matrix(air_Data)
dim(air_Data)
```

we can know that in this data:

$$\begin{aligned} Individuals &= 80 \\ Variables &= 11 \end{aligned} \tag{1}$$

Table 2. Mean and Variance Vector

	SMIN	SMEAN	SMAX	PMIN	PMEAN	PMAX	PM2	PERWH	NONP	GE65	LPOP
Mean	47.1	99.7	219.9	44.5	116.7	275.5	72.9	87.3	81.8	85.9	56.6

	SMIN	SMEAN	SMAX	PMIN	PMEAN	PMAX	PM2	PERWH	NON	GE65	LPOP
Var	913.2	2542.9	14409.4	337.9	1508.4	25312.5	23920.2	107.8	45.5	465.5	14.9

Table 2. From mean, suspend particular has bigger range than sulphat and the PERWH has the bigger value than others. From variance, we can infer the suspend particular is more unstable than the suiphatt, and there are more uncertainty in GE65 in other 5 variables not include sulphat and suspend particular. But the sulphat and suspend particular are the problem should be considered firstly.All the values of each variables are continuous,and we can infer that we can use the Linear Regression or classifier after making label if we need.

But for now,then, Get the scatter plot. R : Fig 1

```
panel.hist <- function(x, ...){
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(usr[1:2], 0, 1.5) )
    h <- hist(x, plot = FALSE)
    breaks <- h$breaks; nB <- length(breaks)
    y <- h$counts; y <- y/max(y)
    rect(breaks[-nB], 0, breaks[-1], y, ...)
}
pairs(air_Matrix, col = c,
       diag.panel = panel.hist)
```

If the two variables have correlations, it can be easily found, for example, we can infer PERWH has relationship with NONPOOR, but PM2 tend to have the least correlation with others.

Then calculate the correlation matrix: R : Fig 2

```
panel.cor <- function(x, y){
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- round(cor(x, y), digits=2)
    txt <- paste0(r)
    cex.cor <- strwidth(txt)
    text(0.5, 0.5, txt)
}

# Create the plots
pairs(air_Data,
    lower.panel = panel.cor,
    upper.panel = panel.cor)
```

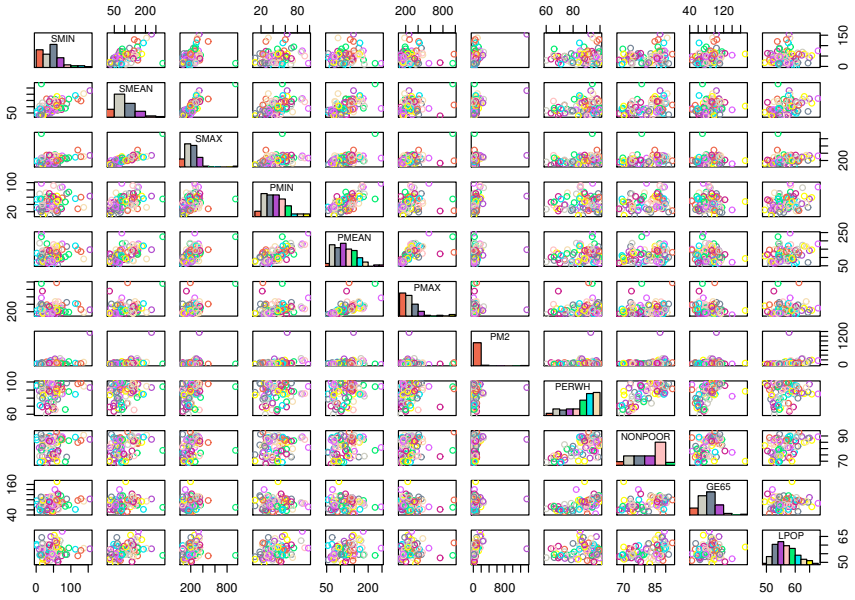

Fig. 1. ScatterPlot

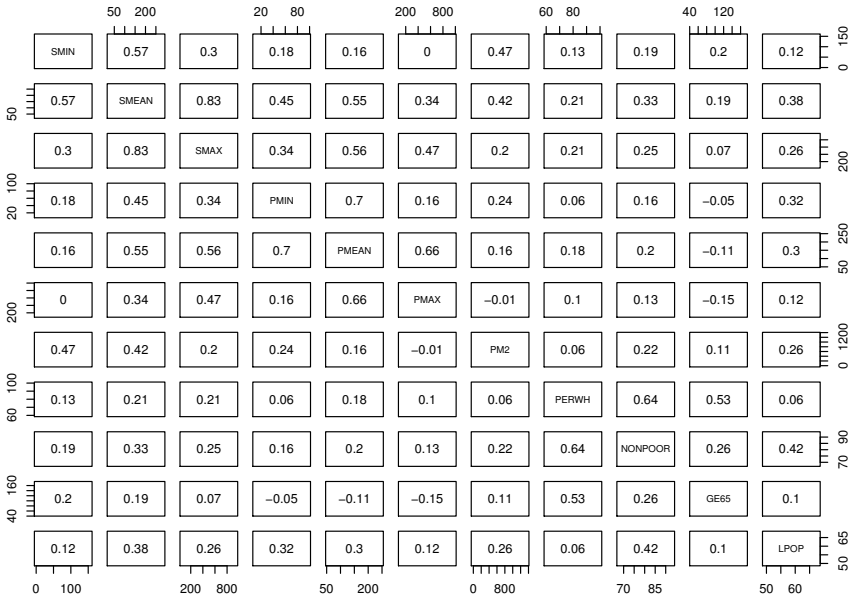

Fig. 2. ScatterPlot of Correalation

And we can know that if the value is positive, so the two variables will be correlation, and the bigger the value is, the stronger the linear relationship is. We can see that is no surprising that the Max,mean and Min value of one pollution have strong positive linear correlation. And the PERWH has strong positive linear relationship with NONPOOR, but PERWH and PM2 just have a little correlation.

Also, we can check these correlation from a heatmap, but it is not as obvious as two methods before if there are much data.

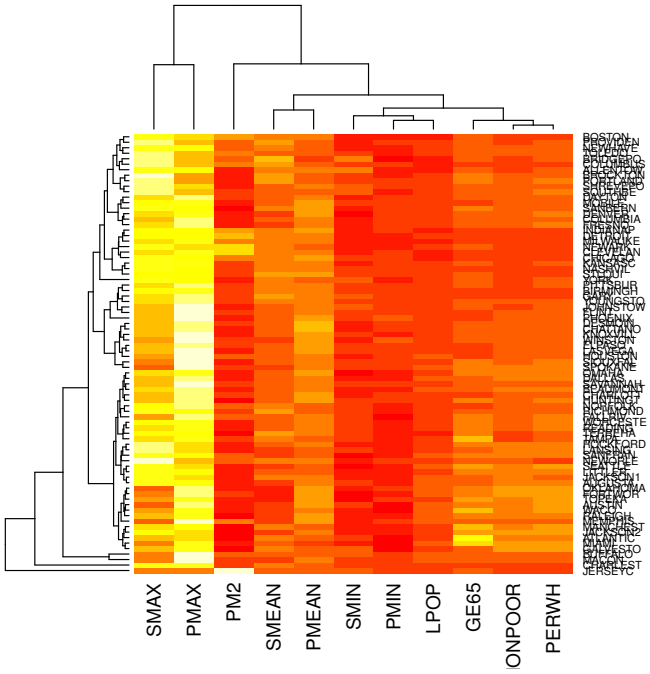

Fig. 3. ScatterPlot of Correalation

2.2 Total Variation and Generalised Variance

R :

```
#Before Standardization
gv = det(var(air_Matrix))
tv = sum(diag(var(air_Matrix)))
```

$$\begin{aligned} \textit{Generalised Variance} &= 8.72131e+29 \\ \textit{Total Variation} &= 69577.97 \end{aligned} \tag{2}$$

2.3 Standardise and Verify

Standardization or standardisation is the process of implementing and developing technical standards based on the consensus of different parties that include firms, users, interest groups, standards organizations and governments Standardization can help to maximize compatibility, interoperability, safety, repeatability, or quality. It can also facilitate commoditization of formerly custom processes. This view includes the case of "spontaneous standardization processes", to produce de facto standards. [1]

Standardise the data R :

```
sdr_Matrix = scale(air_Matrix, center=T,scale=T)
```

Check Mean R :

```
colMeans(sdr_Matrix)
as.integer(colMeans(sdr_Matrix))
>[1] 0 0 0 0 0 0 0 0 0 0 0
```

Check Covariance Matrix R :

```
var(sdr_Matrix)
cor(air_Matrix)
```

Table 3. Covariance Matrix

	SMIN	SMEAN	SMAX	PMIN	PMEAN	PMAX	PM2	PERWH	NONP	GE65	LPOP
SMIN	1.00	0.57	0.30	0.18	0.16	-0.00	0.47	0.13	0.19	0.20	0.12
SMEAN	0.57	1.00	0.83	0.45	0.55	0.34	0.42	0.21	0.33	0.19	0.38
SMAX	0.30	0.83	1.00	0.34	0.56	0.47	0.20	0.21	0.25	0.07	0.26
PMIN	0.18	0.45	0.34	1.00	0.70	0.16	0.24	0.06	0.16	-0.05	0.32
PMEAN	0.16	0.55	0.56	0.70	1.00	0.66	0.16	0.18	0.20	-0.11	0.30
PMAX	-0.00	0.34	0.47	0.16	0.66	1.00	-0.01	0.10	0.13	-0.15	0.12
PM2	0.47	0.42	0.20	0.24	0.16	-0.01	1.00	0.06	0.22	0.11	0.26
PERWH	0.13	0.21	0.21	0.06	0.18	0.10	0.06	1.00	0.64	0.53	0.06
NONP	0.19	0.33	0.25	0.16	0.20	0.13	0.22	0.64	1.00	0.26	0.42
GE65	0.20	0.19	0.07	-0.05	-0.11	-0.15	0.11	0.53	0.26	1.00	0.10
LPOP	0.12	0.38	0.26	0.32	0.30	0.12	0.26	0.06	0.42	0.10	1.00

Table 4. Correlation Matrix

	SMIN	SMEAN	SMAX	PMIN	PMEAN	PMAX	PM2	PERWH	NONP	GE65	LPOP
SMIN	1.00	0.57	0.30	0.18	0.16	-0.00	0.47	0.13	0.19	0.20	0.12
SMEAN	0.57	1.00	0.83	0.45	0.55	0.34	0.42	0.21	0.33	0.19	0.38
SMAX	0.30	0.83	1.00	0.34	0.56	0.47	0.20	0.21	0.25	0.07	0.26
PMIN	0.18	0.45	0.34	1.00	0.70	0.16	0.24	0.06	0.16	-0.05	0.32
PMEAN	0.16	0.55	0.56	0.70	1.00	0.66	0.16	0.18	0.20	-0.11	0.30
PMAX	-0.00	0.34	0.47	0.16	0.66	1.00	-0.01	0.10	0.13	-0.15	0.12
PM2	0.47	0.42	0.20	0.24	0.16	-0.01	1.00	0.06	0.22	0.11	0.26
PERWH	0.13	0.21	0.21	0.06	0.18	0.10	0.06	1.00	0.64	0.53	0.06
NONP	0.19	0.33	0.25	0.16	0.20	0.13	0.22	0.64	1.00	0.26	0.42
GE65	0.20	0.19	0.07	-0.05	-0.11	-0.15	0.11	0.53	0.26	1.00	0.10
LPOP	0.12	0.38	0.26	0.32	0.30	0.12	0.26	0.06	0.42	0.10	1.00

3 Second Part

Standardised data is better

I think using the PCA based on standardised data is always the better way. First, I think data in real life must be multi-dimensions.These dimensions have different identities and the values of them are various–too big or too small. The values of different things which have different identities cant be compare directly. So we can use STANDARDIZATION to make them has the ZERO-MEAN and ONE-SD, the process will make these things can be compared easily.Second for PCA, we want to find some correlation from different and use less informations to represent the more information, it is necessary to make these variables can be used together. Third, we can if the data is too small or too big, and the gap of each of them should be smaller, so standardization have the effects like Data-Normalization in Machine Learning I think.

Interpretation of Principle Components

First PC

I think all coefficients of first PC have same sign, so the first PC can be called the 'size', and the first PC can be interpreted as the average measures for whole data, the GE65 is the least and sulphate is the most.

	SMIN	SMEAN	SMAX	PMIN	PMEAN	PMAX	PM2	PERWH	NONPOOR	GE65	LPOP
PC1	0.26	0.45	0.40	0.31	0.39	0.25	0.24	0.21	0.28	0.11	0.27
PC2	0.19	-0.01	-0.13	-0.23	-0.34	-0.34	0.15	0.46	0.37	0.54	0.04

Second PC

Firstly, most coefficients of sulphate and suspend particular are negative and the coefficients of others are positive, so the second PC can be interpreted as

the contrast of these twoo group. Secondly, from the values of coefficients, it can be interpreted that second PC are focus **less** on using LPOP and SMEAN to compare.

How many Principle Components will be recommended
From Table 5, summary of PCs, we can choose the PC4 - PC7.
From , Fig 4, first nine PCs can be needed.
So, in conclusion, I think we can choose the PC2 - PC7.

Table 5. Summary of Principle Components

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
Standard deviation	1.9589	1.3779	1.1793	1.0215	0.8790	0.8215	0.7378	0.6607	0.4573	0.3294	0.2896
Proportion of Variance	0.3488	0.1726	0.1264	0.0949	0.0703	0.0613	0.0495	0.0397	0.0190	0.0099	0.0076
Cumulative Proportion	0.3488	0.5214	0.6479	0.7427	0.8130	0.8743	0.9238	0.9635	0.9825	0.9924	1.0000

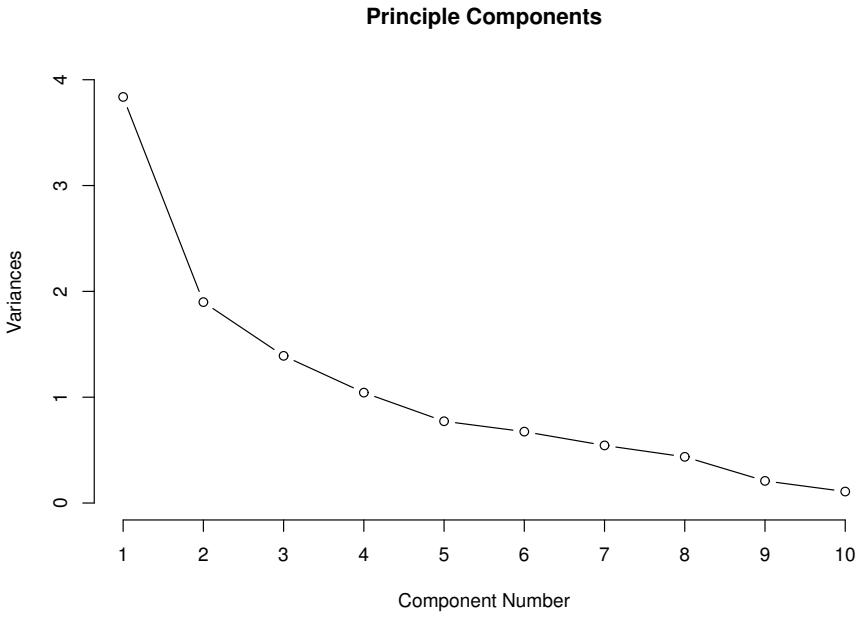

Fig. 4. Scree Daigram

Scatter plot of Principle Components
From graph, I think if the Principle Componets is very good, it will show some

cluster of group, but it is mess in the scatter plot, and there are two cities which have the longest distance, I think these two principle components are focus least on them.

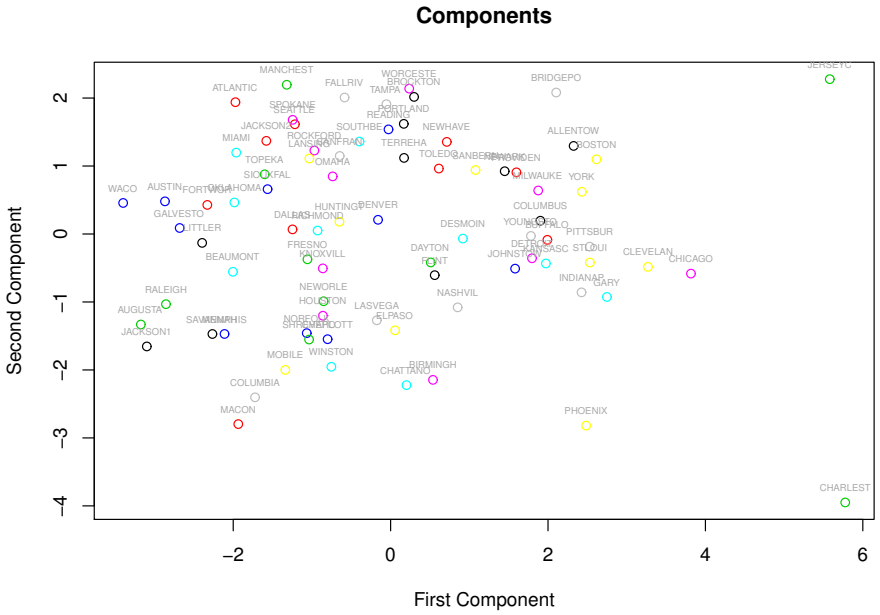

Fig. 5. Components

References

1. Xie, Z., Hall, J., McCarthy, I.P., Skitmore, M. and Shen, L., 2016. Standardization efforts: The relationship between knowledge dimensions, search processes and innovation outcomes. Technovation, 48, pp.69-78.