Beijing Multi-Site Air-Quality Data Set
Final Project

May 3, 2022

Professor Jonathan Reuning-Scherer

Multivariate Statistical Methods Final Project

Introduction and Background:

Outdoor air pollution is a major public health issue that affects people in low-income, middle-income, and high-income countries. In 2016, it was anticipated that fine particles 2.5 microns in diameter or smaller, commonly known as PM2, were responsible for 4.2 million premature deaths worldwide attributable to exposure to ambient (outdoor) air pollution in urban and rural areas. Cancer and respiratory disorders. (WHO)

Pollution can be measured by monitoring air quality indicators. The data for this study was collected from the Beijing Environmental Monitoring Center's air quality monitoring station between March 1, 2013, and February 28, 2017. Data on air pollution collected every hour. Carbon monoxide (CO), lead (Pb), nitrogen dioxide (NO2), suspended particle matter containing dust, soot, smoke, and soot (SPM), sulfur dioxide (SO2), and tropospheric ozone (O3) were the six "typical" air pollutants that were monitored.

Design and primary questions:

Some of the main questions we may want to answer, based on those air pollutants and relevant meteorological variables, are that how can we explain the variability of this dataset in a lower dimensional world while preserving most of the variability between those variables, what might be the potential factors that cause this variability and make the variables correlated with each other, and how can we categorize different measurements with different time point into different clusters and how can we interpret such clustering in words. The tests that correspond to all of three questions above are PCA (Principal Component Analysis), Factor Analysis, and Cluster Analysis. To perform those tests, some of data cleaning and transforming are needed, which will be illustrated below.

Data Description:

Beijing Multi-Site Air-Quality Data Set from 12 nationally controlled air-quality monitoring sites. There are 35064 observations of 18 feature variables, the variables may be grouped as follows:

- 5 Categorical variables: year, month, day, hour, wd, station
- 13 Continuous variables(unit):

  No., NO2(ug/m^3), CO(ug/m^3), O3(ug/m^3), PM2.5 concentration (ug/m^3), PM10 concentration (ug/m^3), SO2(ug/m^3), TEMP(degree Celsius), PRES pressure (hPa), DEWP(degree Celsius), RAIN precipitation (mm), WSPM wind speed (m/s)

-

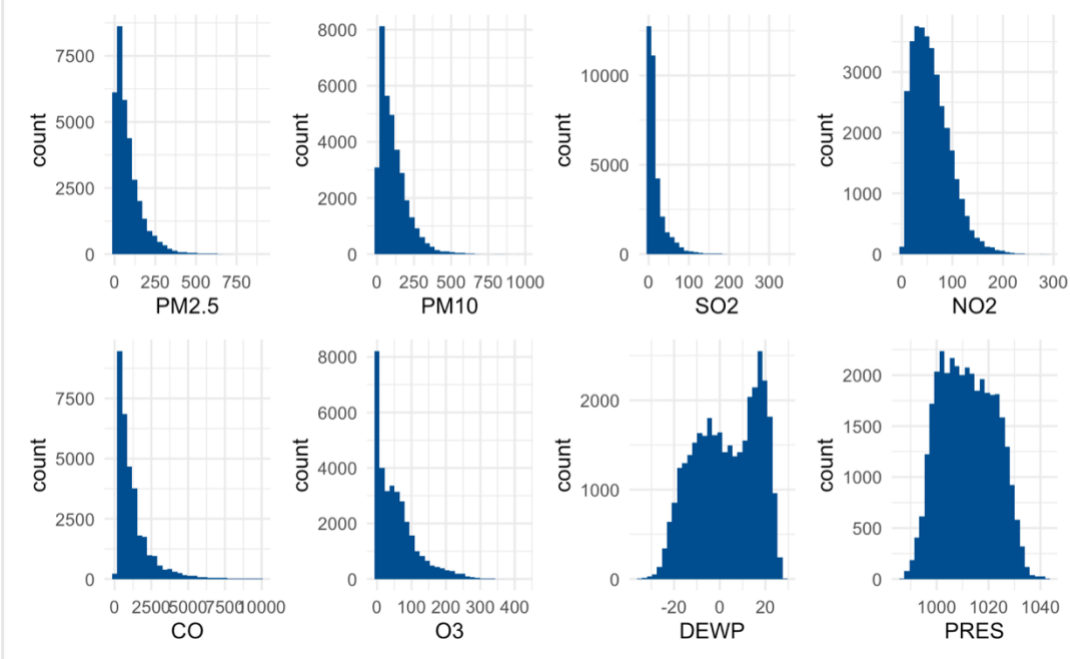
```
       No             year          month            day             hour          PM2.5            PM10
 Min.   :    1   2013:7344   1      : 2976   1      : 1152   0      : 1461   Min.   :  3.00   Min.   :  2.0
 1st Qu.: 8767   2014:8760   3      : 2976   2      : 1152   1      : 1461   1st Qu.: 22.00   1st Qu.: 38.0
 Median :17532   2015:8760   5      : 2976   3      : 1152   2      : 1461   Median : 58.00   Median : 87.0
 Mean   :17532   2016:8784   7      : 2976   4      : 1152   3      : 1461   Mean   : 82.77   Mean   :110.1
 3rd Qu.:26298   2017:1416   8      : 2976   5      : 1152   4      : 1461   3rd Qu.:114.00   3rd Qu.:155.0
 Max.   :35064               10     : 2976   6      : 1152   5      : 1461   Max.   :898.00   Max.   :984.0
                             (Other):17208   (Other):28152   (Other):26298   NA's   :925      NA's   :718
      SO2              NO2                CO              O3              TEMP             PRES             DEWP
 Min.   :  0.2856   Min.   :  2.00   Min.   :  100   Min.   :  0.2142   Min.   :-16.80   Min.   : 985.9   Min.   :-35.300
 1st Qu.:  3.0000   1st Qu.: 30.00   1st Qu.:  500   1st Qu.:  8.0000   1st Qu.:  3.10   1st Qu.:1003.3   1st Qu.: -8.100
 Median :  9.0000   Median : 53.00   Median :  900   Median : 42.0000   Median : 14.50   Median :1011.4   Median :  3.800
 Mean   : 17.3759   Mean   : 59.31   Mean   : 1263   Mean   : 56.3534   Mean   : 13.58   Mean   :1011.8   Mean   :  3.123
 3rd Qu.: 21.0000   3rd Qu.: 82.00   3rd Qu.: 1500   3rd Qu.: 82.0000   3rd Qu.: 23.30   3rd Qu.:1020.1   3rd Qu.: 15.600
 Max.   :341.0000   Max.   :290.00   Max.   :10000   Max.   :423.0000   Max.   : 40.50   Max.   :1042.0   Max.   : 28.500
 NA's   :935        NA's   :1023     NA's   :1776    NA's   :1719       NA's   :20       NA's   :20       NA's   :20
      RAIN               wd                WSPM            station
 Min.   : 0.00000   Length:35064       Min.   : 0.000   Length:35064
 1st Qu.: 0.00000   Class :character   1st Qu.: 0.900   Class :character
 Median : 0.00000   Mode  :character   Median : 1.400   Mode  :character
 Mean   : 0.06742                      Mean   : 1.708
 3rd Qu.: 0.00000                      3rd Qu.: 2.200
 Max.   :72.50000                      Max.   :11.200
 NA's   :20                            NA's   :14
```

- How data collected: The Beijing Municipal Environmental Monitoring Center obtained hourly air pollution data from 12 nationally controlled air-quality monitoring sites, from March 1st, 2013, to February 28th, 2017.
- Sources of error: Noticed that PM2.5 has 925 NAs, Pm10 has 718 NAs, SO2 has 935 NAs, NO2 has 1023 NAs, CO has 1776 NAs, O3 has 1719 NAs, TEMP has 20 NAs, PRES has 20 NAs, DEMO has 20 NAs, RAIN has 20 NAs, WSPM has 14 NAs. Delete these NAs before computing the dataset. Ignore the errors caused by the equipment.

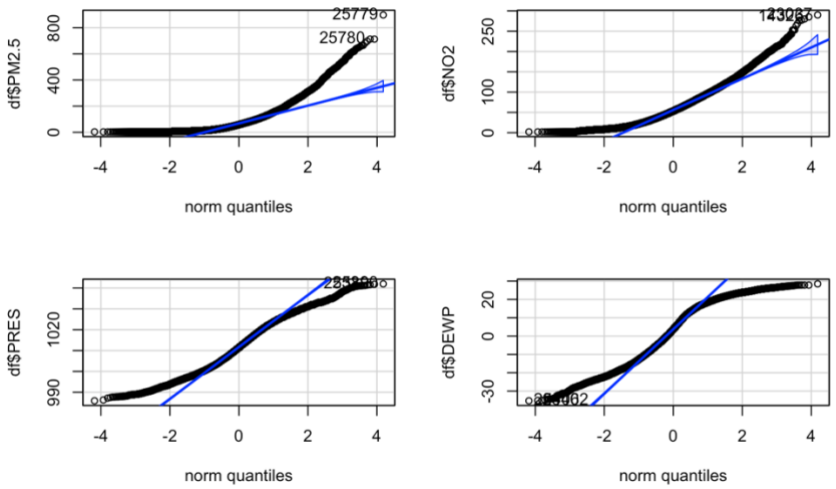
Data Plots and Summary Statistics:

First of all, make the histogram of each numeric variables to detect the outliers, for the main 6 main air pollutants, most of them are the right-skewed distribution. As a result, delete each NA appeared in the Beijing Multi-Site Air-Quality Data Set. As for the outliers, using the Grubb's Test to detect and eliminate them, then compute the normal quantile plots of each numeric variables again.

Attached the histogram plots:



I have drawn normal quantile plots for 13 continuous data and selected four representative plot examples. For example, the graph of pressure in the data shows better normal characteristics, which can also be obtained from the above histogram. However, most data do not conform to the normal distribution due to the obvious right-skew characteristics, such as PM2.5.



Methods & Results

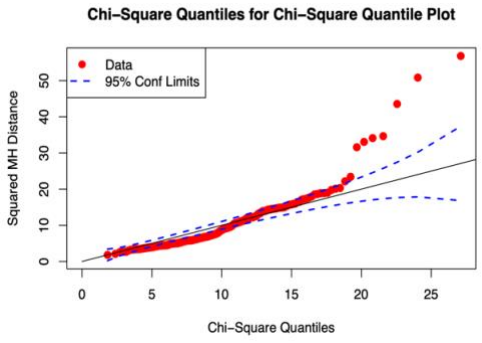
- *PCA:*

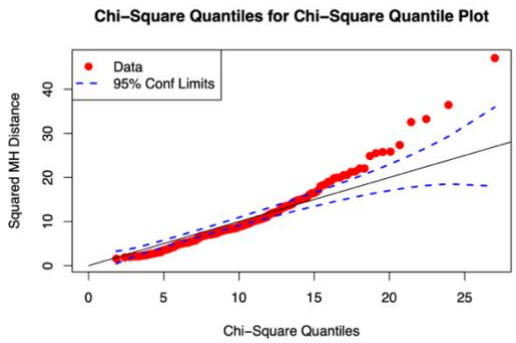
The very first method we are conducting here is PCA (Principal Component Analysis), which examines the question that how we can explain the most of variability in our dataset in a lower dimensional world. Everything below is completed with R studio and relevant packages in R.

At the beginning of our study, we decided to choose the first 200 rows instead of the whole dataset to better visualizing the plots and to better explaining our plots later. Besides, only the numeric columns have been selected, since the PCA is more fitted to work with numeric variables. Moreover, of course, any row with NA or null value will be dropped. After doing these manipulations of dataset, we ended up with a 200*10 dataset.

To conduct the PCA, we need to firstly check some of assumptions.

1. The relationship between variables should be linear.
2. The variables in dataset should follow a multivariate normal distribution. (Not need for PCA, but needed for parallel analysis)
3. The correlations between variables should be large enough to conduct PCA.

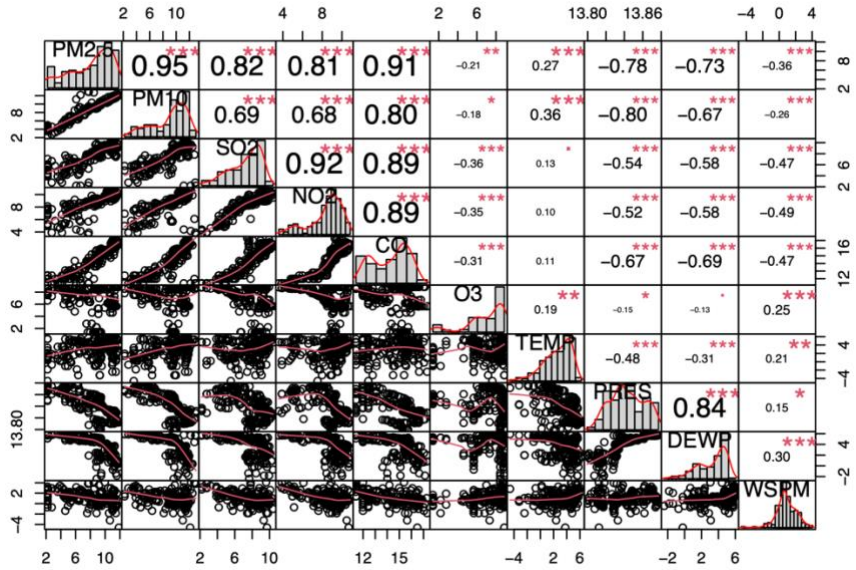




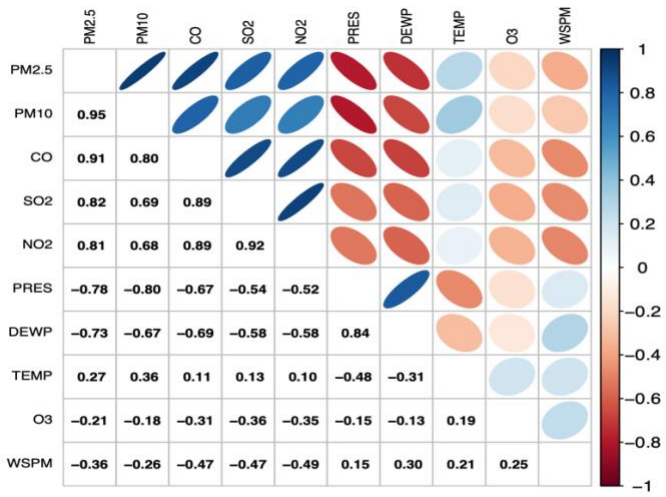
The plot on the left-hand side is the one before the transformation, and the one on the right-hand side is the one after the transformation. The transformation applied here is log(x^2) where x is every variable in the dataset. The reason why we must square it before taking the log is that some of variables take the negative value and negative value is not in the domain of log function. So, we must square it firstly.

Although it is still not a perfect normal distribution based on the quantile plot, it is good enough to continue our further analysis.

(skip to the next page)



This plot shows the scatter plots and its potential relationship between each pair of variables. It is obvious to observe that most of pairs have clear linear relationship. Therefore, it is appropriate to conduct PCA in this case. Meanwhile, correlations are good indicators to see if the relationship is strongly linear, Based on the plot above, we could see that correlations are decently large enough to indicate linear relationship.



This is another plot, probably to be considered as a more artistic one, to demonstrate the facts that linearity assumption and correlation assumption are satisfied. We could easily see that most pairs of variables have strong correlation (Over half of correlations are over 0.5), which guarantee that we could reduce our dimensions while explaining most of variability in our dataset. (The main target of PCA)

Before formally conducting PCA, we may want to determine how many principal components would be more efficient to explain the variability while keeping the number of PC as small as possible. There are various tools we can perform to determine this, including total variance, eigenvalue>1 criteria, scree plot and parallel analysis plot as well.

Total variance:

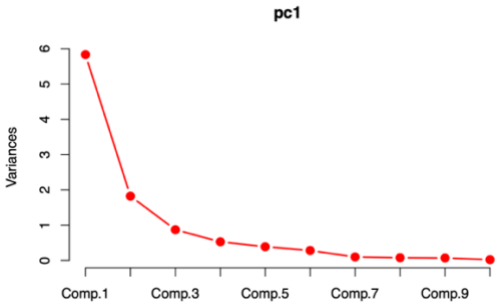
	Comp1	Comp2	Comp3	Comp4	Comp5
Cumulative Proportion	0.58	0.77	0.85	0.91	0.94

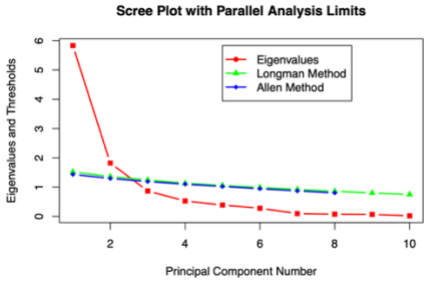
The table above shows the total variance that has been explained by including first couples of components in our PCA model.
The rule of thumb here is to choose number of components that explain over 80% of variations; Therefore, first three principal components should be included, since they explain 85% of variability.

Eigenvalue criteria:

	Comp1	Comp2	Comp3
Eigenvalue	5.83	1.82	0.87

The rule we applied here is to choose number of components where the eigenvalue is smaller than one. Therefore, choosing first three PC would be the best solution according to this method.





One the LHS, we have scree plot that determines the number of PC at its elbow. One the RHS, we have our Parallel Analysis plot that determines number of PC at where eigenvalues are smaller than the other two curves. Parallel Analysis is fitted to do here because we have a multivariate normal distribution between our variables, which has been illustrated before. Both of plots indicate that choosing first 3 components would be decent.

Consequently, all four methods support the idea that we should select three PCs .

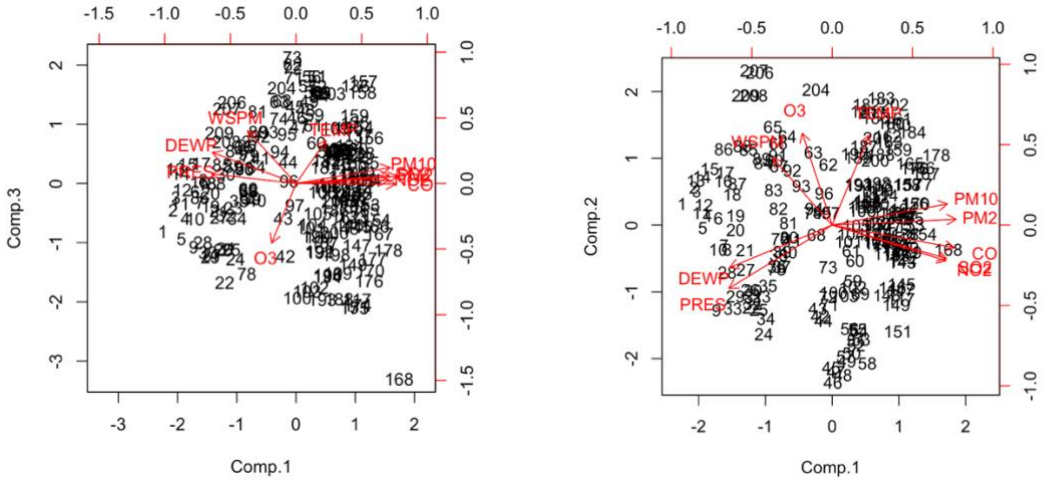
```
Loadings:
      Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10
PM2.5  0.40   0.03   0.07   0.15   0.13   0.31   0.21   0.10   0.06   0.80
PM10   0.37   0.12   0.16   0.13   0.37   0.48   0.27  -0.03  -0.29  -0.53
SO2    0.37  -0.19   0.08  -0.02  -0.48  -0.04  -0.17   0.67  -0.31  -0.10
NO2    0.37  -0.21   0.03   0.00  -0.49  -0.03   0.01  -0.72  -0.25   0.03
CO     0.39  -0.13  -0.01   0.14  -0.12   0.03  -0.04   0.02   0.85  -0.26
O3    -0.10   0.53  -0.61   0.07  -0.42   0.39   0.06   0.03   0.03  -0.04
TEMP   0.12   0.52   0.44  -0.69  -0.14  -0.04   0.09  -0.01   0.13   0.01
PRES  -0.33  -0.36   0.10  -0.10  -0.29   0.13   0.79   0.11   0.09  -0.04
DEWP  -0.33  -0.24   0.32  -0.10  -0.13   0.69  -0.47  -0.05   0.08   0.04
WSPM  -0.19   0.39   0.54   0.66  -0.26  -0.13   0.04  -0.01   0.01  -0.01
```

After doing PCA, the plot above shows all loadings for each component. We will only look at the first three components since we have decided to only include the first three PC.

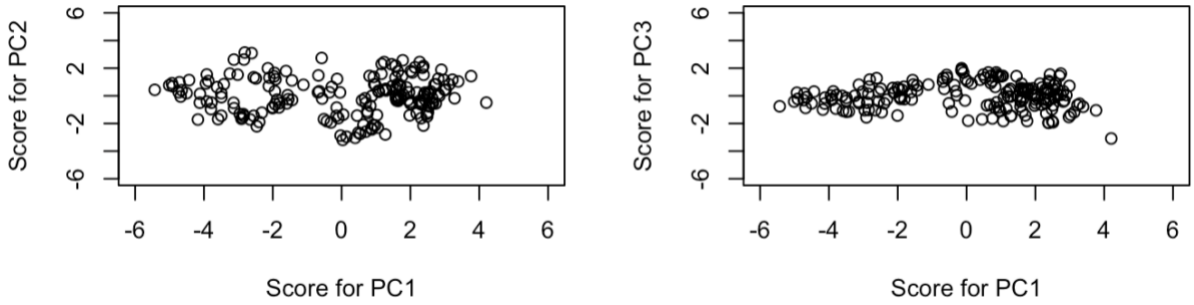
Interpretations for components:

The first PC hugely correlates with PM2.5, PM10, SO2, NO2 and CO, and weakly correlates with Temp and WSPM. We could interpret this PC as the one that separate the air pollutants with environmental factors. The second PC counts hugely on

O3 and Temp. Since O3 and Temp have strong positive relationship empirically, we could simply consider this component as the one that distinguish measurements by its temperature. The third PC behaves very similar to components 2, the difference is that PC3 correlates more heavily on few variables while PC2 is more spread out. Therefore, we could say that PC3 does a more concentrated job at distinguishing measurements by its temperature and PC2 does a similar job but considering some of meteorological variables, like PRES in the meantime.



These two biplots help us to better visualize our loading coefficients listed above. For example, we could see that PM2.5 weights a lot on Comp1 since it is parallel to X-axis and the norm of the vector is long etc.



The scores plot between PC1, PC2 and PC1, PC3 above indicates that our analysis is very effective. If we look the two plots, we could easily see that most of

variations could be explained by PC1 followed by PC2 and then PC3 (The flatter the pattern of scatter plot is, the better PCA we have done). But the gaps of explaining the variations between three PCs are huge, and PC1 does an incredibly good job at capturing most of the variations, since dots are distributed evenly along the parallel line of the x-axis.

- *Factor Analysis:*

So far, we have found that many of variables in our datasets are strongly correlated, and we have done PCA to explain the total variation while reducing dimension to three.

Naturally, we may want to ask if there are some factors causing this strong correlation, which raise to our study of factor analysis. Evaluation of correlation has been done before at the part of PCA, which ensures that we have strong correlation to conduct factor analysis.

For the dataset, we will just keep using the cleaning version of data from PCA model, because every requirement about the dataset in factor analysis is the same as the one for PCA.

The extraction methods, in this study, will include iterative PCA, PAF and Maximum likelihood in this study. Iterative PCA assumes all communalities equal to 0 at the first iteration, and then compute the max change in total communalities to revise the estimates of total communalities at each iteration, which, however, does not guarantee the communalities to be converged. Iterative PAF does the same procedure, the only difference is that it uses Principal Axis Factoring instead of PCA. But every iteration follows the same rule. Maximum likelihood method uses MLE to estimate the communalities based on the assumption that all factors are independently normally distributed.

We are only using one rotation method, which is varimax. Varimax rotation basically seeks to rotate the plot such that the variance of loadings within factors can be maximized. And it works in our case, since varimax works on the columns of dataset, which is our case here.

Another main question here is how to determine the number of factors. PCA is a great method to do this. And based on the selection at former part of PCA, we have

decided to include three PCs in our model. So, we will keep using 3 factors in factor analysis part.

Three factor analysis models have the following loadings.

Maximum likelihood with Varimax rotation

```
Loadings:
      Factor1 Factor2 Factor3
PM2.5  0.650   0.702   0.231
PM10   0.441   0.829   0.337
SO2    0.886   0.296   0.160
NO2    0.890   0.284   0.161
CO     0.838   0.463   0.145
O3    -0.371   0.194  -0.516
TEMP           0.524  -0.120
PRES  -0.354  -0.870   0.238
DEWP  -0.502  -0.671   0.321
WSPM  -0.541

                Factor1 Factor2 Factor3
SS loadings       3.711   3.080   0.687
Proportion Var    0.371   0.308   0.069
Cumulative Var    0.371   0.679   0.748
```

PAF with Varimax rotation

```
         PA1   PA3   PA2   h2    u2 com
PM2.5  0.71  0.65  0.18 0.95 0.047 2.1
PM10   0.55  0.73  0.17 0.87 0.129 2.0
SO2    0.81  0.32  0.30 0.86 0.143 1.6
NO2    0.84  0.29  0.28 0.86 0.141 1.5
CO     0.87  0.40  0.20 0.95 0.053 1.5
O3    -0.28  0.13 -0.60 0.45 0.548 1.5
TEMP  -0.10  0.61 -0.12 0.40 0.604 1.1
PRES  -0.47 -0.79  0.31 0.94 0.060 2.0
DEWP  -0.65 -0.52  0.40 0.85 0.146 2.6
WSPM  -0.63  0.17 -0.07 0.43 0.572 1.2

                       PA1  PA3  PA2
SS loadings           4.05 2.61 0.90
Proportion Var        0.40 0.26 0.09
Cumulative Var        0.40 0.67 0.76
Proportion Explained  0.54 0.35 0.12
Cumulative Proportion 0.54 0.88 1.00
```

Iterative PCA with varimax rotation

```
         PA1   PA3   PA2   h2    u2 com
PM2.5  0.70  0.65  0.20 0.95 0.047 2.2
PM10   0.54  0.74  0.19 0.87 0.131 2.0
SO2    0.80  0.33  0.32 0.86 0.144 1.7
NO2    0.83  0.30  0.30 0.86 0.142 1.5
CO     0.85  0.41  0.23 0.95 0.053 1.6
O3    -0.27  0.13 -0.61 0.47 0.534 1.5
TEMP  -0.10  0.61 -0.12 0.40 0.604 1.1
PRES  -0.47 -0.80  0.29 0.94 0.059 1.9
DEWP  -0.65 -0.53  0.37 0.84 0.160 2.6
WSPM  -0.63  0.16 -0.08 0.43 0.568 1.2

                       PA1  PA3  PA2
SS loadings           3.95 2.68 0.93
Proportion Var        0.39 0.27 0.09
Cumulative Var        0.39 0.66 0.76
Proportion Explained  0.52 0.35 0.12
Cumulative Proportion 0.52 0.88 1.00
```

```
Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 71.42 on 18 degrees of freedom.
The p-value is 2.59e-08
```

Based on this loading information, we could see that all of three have done a good job of capturing the variation within variables, since their cumulative var are all close to 0.8. It is also good to see that the hypothesis testing under the ML with varimax model suggests that three factors are sufficient, which follows our assumption of number of factors = 3.

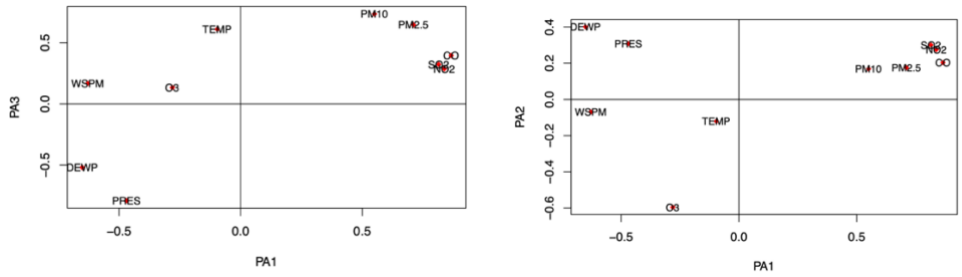
To better understand and to choose the best model among these three, we could use residual correlation matrix and root mean squared residuals to make comparison between these three models.

	ML model	PAF	PCA
root mean squared residuals	0.030	0.02179	0.0218
*	6.67%	6.67%	6.67%

*: proportion of residual greater than 0.05 in absolute value in residual matrix

Based on this table, since all three models have the same proportion, we may want to determine the best one by comparing their root mean squared residuals. Based on that rule, we can tell that iterative PAF with varimax rotation model is the best one, since it has lowest root mean squared residuals.

The last thing about factor analysis is to give an empirical interpretation for underlying factors. In order to see this clearly, we will use the loading plots of iterative PAF with varimax rotation model to see how variables are distributed by factors. And we will try to explain them in words.



PA1, our first factor, distinguishes PRES, DEWP, WSPM from PM10, PM2.5. It basically separates air pollutants variables from meteorological variables. Therefore, this factor can be directly explained as whether the variable is an air pollutant variable or a meteorological variable.

PA2 distinguishes O3 and Temp from the rest of all variables. And scientifically, O3 should have positive relationship with Temp, and high O3 in air causes temperature to raise. So, this factor could be understood as whether this variable is strongly correlated with temperature or not. Intuitively, those variables that are correlated to temperature in air will behave similarly, so that this factor makes sense here.

PA3, like PA2, distinguishes DEWP (dew point temperature) and PRES(pressure) from the rest of variables. We notice that DEWP and PRES, like O3 and Temp, are highly correlated. The relationship is causal. High pressure in air causes dew point temperature to raise as well. Consequently, this factor is saying that those variables that will be accompanied by change of pressure should be highly correlated, so that this factor will explain their high correlation.

- *Cluster Analysis:*

Another question we may also be curious about is that how we can categorize all these measurements from different time point into different kinds, and what might be the potential reason for such clustering, which raise to our study of cluster analysis.

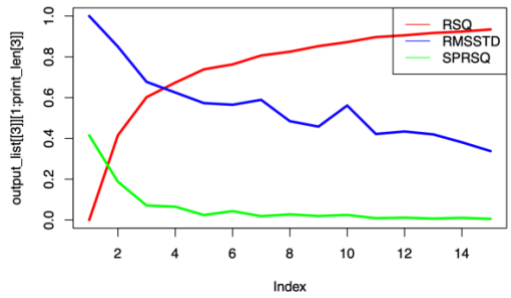
We will still use the same cleaning version of dataset from PCA, since we are still only interested in numerical variables with all NA value deleted from the raw dataset. The differences are that we will only use the first 40 rows of dataset to better visualize our Dendrograms and to explore their potential relationships. Meanwhile, scaling is needed in this case, since we do not want to have any variable that dominates the decision of clustering.

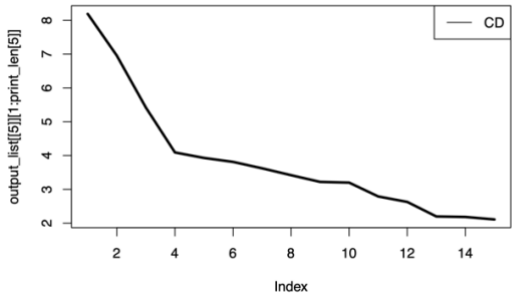
The distance metrics we are going to use are Maximum and Euclidean, which are the two most common distance metrics. Euclidean distance measures the differences between two rows by giving equal weights to all variables, which is the most nature setting.

Maximum distance, by words, measures the differences by the biggest gap between all variables of two rows of dataset.

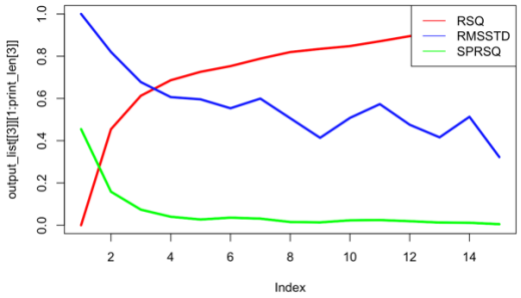
The agglomeration methods we are using here are Ward and Complete. Ward methods minimizes internal sums of squares, which is good for space conserving. Complete method defines the maximum distance as the distance between clusters, which might be a problem of detecting the outlier, but still a common way to do agglomeration. We may want to firstly determine the number of clusters for each of combinations of methods above before making the Dendrograms.

Euclidean and complete:

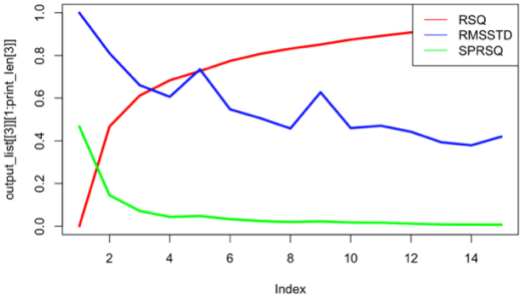




Maximum and Wards:



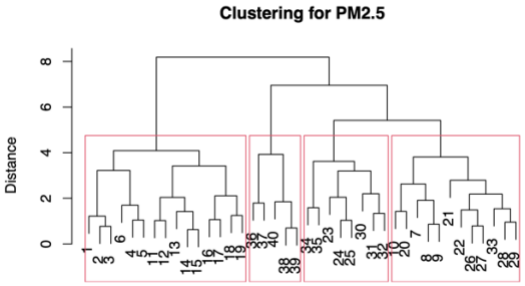
Euclidean and Wards:



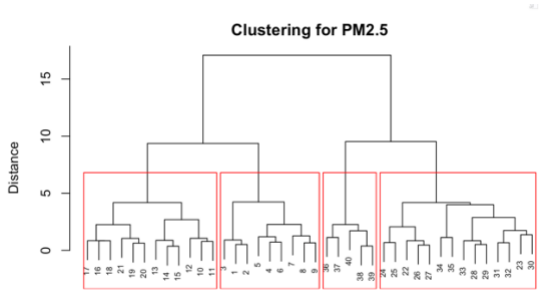
All of three models we have picked (Euclidean and complete, Maximum and Wards, Euclidean and Wards) have indicated that four clusters would be the best choice since RMSSTD reach the local minimum, SPRSQ and RSQ reach their elbows at clusters=4 point; Meanwhile, CD plot for Euclidean and Complete reaches its elbow at cluster=4 as well. All the plots point out the result that choosing cluster=4 would be the best option.

Then it is the time to perform clustering analysis and to get the Dendrograms.

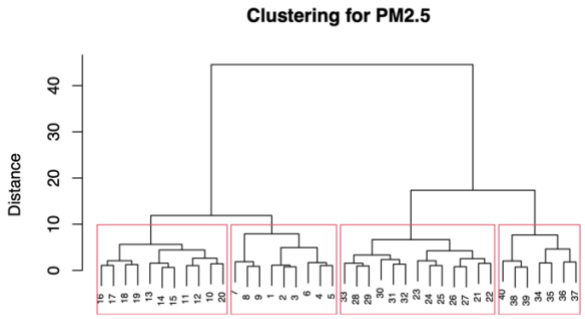
Euclidean and Complete:



Maximum and Wards:



Euclidean and Wards:



From the first Dendrogram (Euclidean and Complete), we could see that roughly first 20 observations are categorized into one kind and the rest 20 observations are categorized into another. The clustering sizes, if four clusters are assumed, are roughly 15, 10, 10, 5 in four clusters.

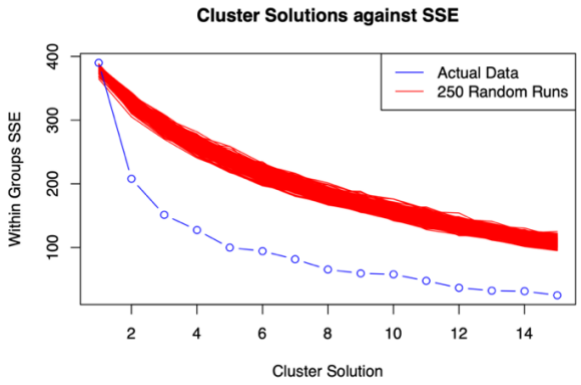
The rest of two Dendrograms are pretty similar to each other, they all cluster roughly 10 observations in a group followed by the number of rows. For instance, row 1-10 are in a category, row 10-20 are in another group etc. One of the interpretations for such clustering might be that row 0-10 are daytime measurements while 10-20 are nighttime measurements, so there is a significant differences of air condition between daytime and nighttime. Moreover, if we only cluster it into 2 clusters, the result supports that there are significant differences of air condition between different date, since it separates row 0-20 from row 20-40, which are basically measurements from two different dates.

It is not hard to see that Wards agglomeration makes more sense here when it comes to interpretation of clustering. Complete agglomeration makes the clustering hard

to explain, probably because there are outliers in our first 40 rows that make this method not working ideally.

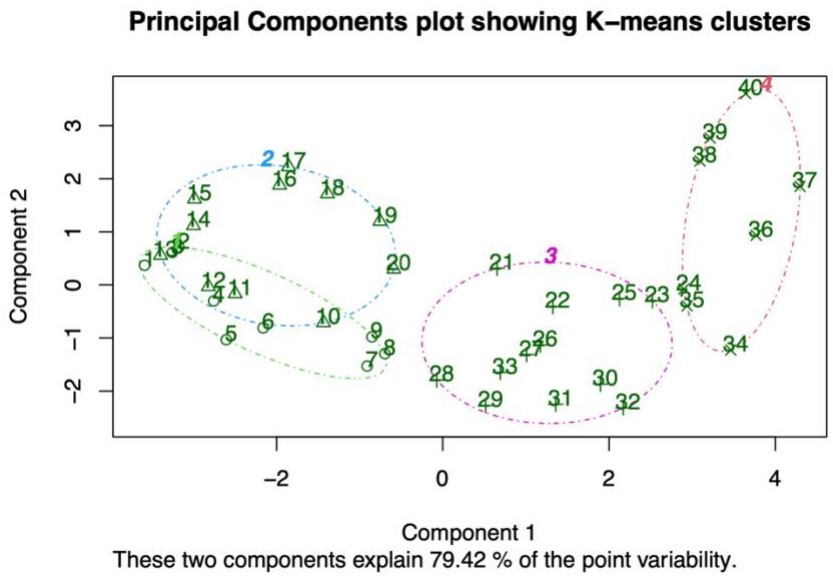
Non-Hierarchical Clustering with K-Means

A more efficient and famous method of clustering is K-means method. Since our data has been scaled, we could directly conduct the K-means method in Rstudio. The first thing we want to do is to determine the number of clusters.



To determine the optimal number of clusters, we are looking for the cluster solution point where the gap between random runs and actual data is maximized. It is, based on the plot, roughly at the cluster solution = 4, so that we decide to set number of clusters equal to 4.

If we run K-means clustering analysis based on clusters = 4, we ended up with clustering result below. The plot shows the clustering result in a 2 PCs plot.



The clustering result is very similar to Maximum and Wards, Euclidean and Wards methods that have been discussed before. So, the interpretation should be similar as well. Such clustering indicates that there are significant differences of air conditions between daytime and nighttime, and between different dates.

Conclusion and Discussion:

Each of three multivariate analysis gives us a good insight about what the data is uncovering about the air condition in Beijing.

PCA method reduces the dimensions of dataset while keeping the most of variations between the variables that we selected. We have checked all assumptions and have done all necessary transformation on data to meet the assumptions. Besides, making uses of variety of methods to determine the most appropriate number of PCs; Fortunately, all methods point out that three PCs would be the best. Once we formally conducted the PCA, we came up with specific explanations for the empirical meaning of each PCs. Finally, the score plot tells us that our PCA successfully captured most of variations.

Factor analysis, beyond the PCA, tells us what might be the potential factors that make our variables strongly correlated. We fitted factor analysis models with varimax rotation and three different extraction methods with three predetermined factors (Based

on the result of PCA). Comparing the root mean squared residuals, we concluded that iterative PAF with varimax rotation model is the best one. Lastly, by looking at the loading plots, we have interpretated what might be the meanings of three factors in the real world.

Clustering analysis helps us to see how the observations might be categorized based on their corresponding variables. The models we have fitted were Euclidean and complete, Maximum and Wards, Euclidean and Wards with respect to different distance metrics and agglomeration methods. We also concluded that all of three should chose four clusters. By looking at the Dendrograms, we not only summarized the pattern of clustering but also discovered some of potential reasons about why such clustering make sense here. At last, K-means method was also performed, the clustering result was like that of Maximum and Wards; Therefore, explanation would be similar as well.