Explanatory Factor Analysis & Confirmatory Factor Analysis of Air Quality Dataset (S/18/847)

01.Introduction

In this study, I am going to analyze a data set by using a 'Factor Analysis (FA)' technique. Factor Analysis is a method of dimension reduction by modeling observed variables and their covariance structure in terms of a smaller number of underlying unobservable (latent) 'factors'. Because, with a large number of variables, the dispersion matrix may be too large to study, difficult to interpret properly, and difficult to manage. To interpret the data in a more meaningful form, it is necessary to reduce the number of variables to a few. So, I will find out how many factors are needed, how they explain the variability of data and their covariance structure.

02.Methodology

I use the 'Air Quality' data set that contains 9358 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality chemical Multisensory Device. The device was located on the field in a significantly polluted area, at road level within an Italian city. Data were recorded from March 2004 to February 2005(one year) representing the longest freely available recordings of on-field deployed air quality chemical sensor device response. Because there are some missing values, I cleaned the air quality data set and then it contains 9357 instances of 13 sensor measurements over one hour (True hourly average concentration) to study gas emissions namely CO, Non-Metanic Hydrocarbons, Benzene, Total Nitrogen Oxides (NO_x) and Nitrogen Dioxide (NO2). The data set contains only the numerical information:

01. CO(GT) – True hourly averaged concentration CO in mg/m^3(reference analyzer)
02. PT08.S1(CO) – Hourly averaged sensor response (CO targeted)
03. NMHC(GT) – True hourly averaged overall Non-Metanic Hydro Carbons concentration in microg/m^3(reference analyzer)
04. C6H6(GT) – True hourly averaged Benzene concentration in microg/m^3 (reference analyzer)
05. PT08.S2(NMHC) – Hourly averaged sensor response (NMHC targeted)
06. NO_x(GT) – True hourly averaged NO_x concentration in ppb (reference analyzer)
07. PT08.S3(NO_x) – hourly averaged sensor response (NO_x targeted)
08. NO2(GT) – True hourly averaged NO2 concentration in microg/m^3 (reference analyzer)
09. PT08.S4(NO2) – Hourly averaged sensor response (NO2 targeted)
10. PT08.S5(O3) – Hourly averaged sensor responses (O3 targeted)
11. T – Temperature (C)
12. RH – Relative Humidity (%)
13. AH – Absolute Humidity

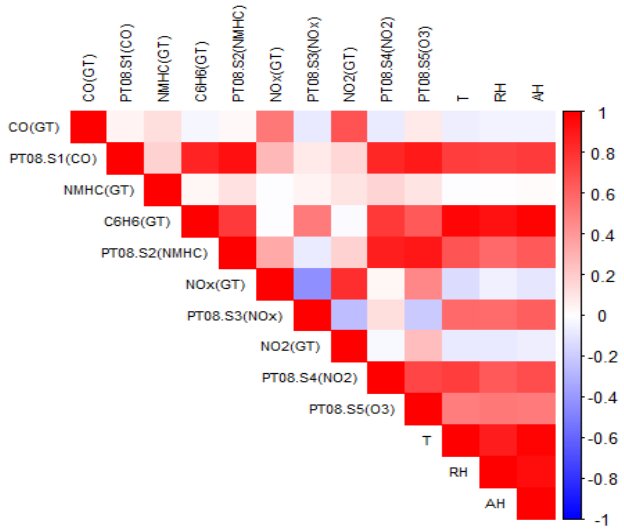
Since different variables have different measurement units, I standardized the data set. Since this data set contains 13 variables, the analysis and interpretation process are not very easy. Therefore. Data reduction is required.
For this study, I am going to perform Exploratory factor analysis and confirmatory factor analysis. In the explanatory factory analysis, I will use both the principal component method and maximum likelihood method with varimax rotation and without rotation.

03.Result and discussion

3.1 Exploratory Factor Analysis

3.1.1 Correlation matrix

Figure 1:



Since the variables have different units of measurement and I wish each variable to receive equal weight in the analysis, then the variable should be standardized before conducting an FA. Therefore, the variance-covariance matrix of the standardized data is equivalent to FA using a correlation matrix (Figure 1). In that figure you can see some variables are strong (positively/dark red), negatively/dark blue) correlated, and some are moderately and low correlated variables.

3.1.2. KMO test

KMO test output gives an overall MSA = 0.73. A KMO value of 0.73 falls within the range of acceptability for factor analysis. However, it is essential to check the KMO value because it indicates how well variables correlate with each other, with values closer to 1 suggesting better suitability.

3.1.3. Eigen values & variances explained by each variable

Table 1:

Component	Eigen Value	Proportion (%)	Cumulative Proportion (%)
CO(GT)	6.5848	50.6524	**50.6524**
PT08.S1(CO)	2.9448	22.6526	**73.3049**
NMHC(GT)	1.4088	10.8372	**84.1421**
C6H6(GT)	1.0323	7.9407	**92.0828**
PT08.S2(NMHC)	0.3938	3.0289	95.1116
NO_x (GT)	0.2507	1.9288	97.0404
PT08.S3(NOx)	0.1316	1.0121	98.0525
NO2(GT)	0.1057	0.8129	98.8654
PT08.S4(NO2)	0.0765	0.5585	99.4539
PT08.S5(O3)	0.0398	0.3062	99.7601
T	0.0282	0.2171	99.9771
RH	0.0026	0.0201	99.9972
AH	0.0004	0.0028	100.00

In Table 1 by considering eigenvalues, there are only four eigenvalues that are greater than one. Therefore, using eigenvalues we can say that the four-factor model is sufficient for this analysis.

In Table 1 by considering cumulative proportion variance explained by the first four factors = 0.9208. Therefore, we can conclude that the factor model explains 92.08% of the total variance which is sufficient interpretation for the dataset with a slight loss of information. Thus, future analysis will be done based on four factors.

3.1.4. Scree Plot

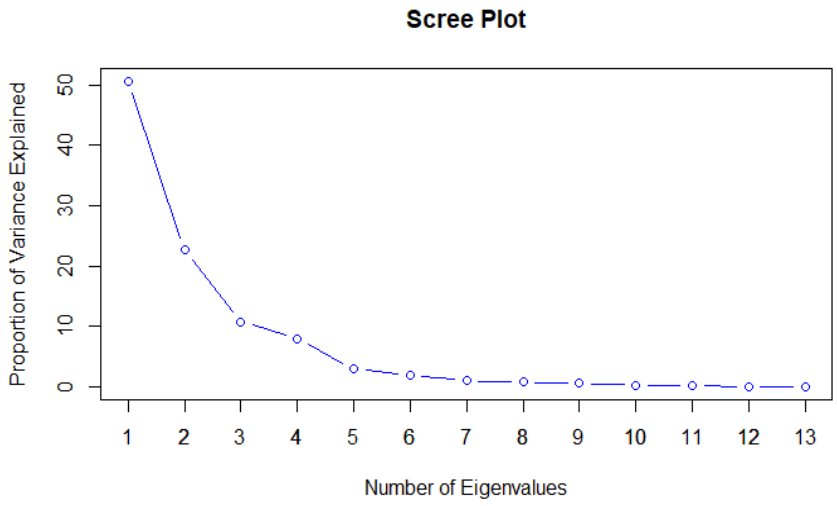


Figure 2:

The non-steep ("Elbow") of the graph can be seen after the fifth factor. Therefore, only the first four factors can represent data adequately.

3.1.4. Factor Loadings

3.1.4.1 Unrotated Factor Loadings & Rotated Factor Loadings

Unrotated	Loadings	From	PC	Method		Rotated	Loadings	From	PC	Method
	FA1	FA2	FA3	FA4			FA1	FA3	FA2	FA4
CO	-0.0100	0.5201	0.4817	0.1279		CO	-0.0650	0.0292	0.7006	0.1518
PT08.S1	0.9324	0.2525	-0.1663	0.0773		PT08.S1	0.9126	0.3278	0.1217	0.1075
NMHC	0.0882	0.1002	0.0111	0.5772		NMHC	0.0885	-0.0041	0.0684	0.5818
C6H6	0.9817	-0.1428	0.1257	-0.0601		C6H6	0.6737	0.7408	-0.0065	-0.0295
PT08.S2	0.8664	0.3521	-0.3073	0.0216		PT08.S2	0.9648	0.1518	0.1145	0.0483
NOx	0.0944	0.8912	0.2062	-0.1844		NOx	0.2889	-0.2584	0.8397	-0.1557
PT08.S3	0.3610	-0.6530	0.5307	0.1044		PT08.S3	-0.2017	0.8723	-0.1860	0.1147
NO2	0.0426	0.8040	0.4876	0.0333		NO2	0.0750	-0.0759	0.9335	0.0655
PT08.S4	0.8469	0.0663	-0.2698	0.1632		PT08.S4	0.8284	0.3036	-0.0941	0.1837
PT08.S5	0.7576	0.4798	-0.3044	-0.0470		PT08.S5	0.9225	0.0258	0.2166	-0.0207
T	0.9249	-0.2726	0.1562	-0.0496		T	0.5679	0.7905	-0.0925	-0.0229
RH	0.8766	-0.2334	0.1933	-0.0970		RH	0.5257	0.7661	-0.0386	-0.0699
AH	0.9365	-0.2662	0.2297	-0.0800		AH	0.5439	0.8408	-0.0412	-0.0508

Table 2: Table 3:

You can see in Table 2 that without factor rotation original factor loading cannot be readily interpretable. In Factor 3 there is no correlated variable. Therefore, we apply factor rotation called "Varimax" rotation in Table 3 that changes the pattern matrix but leaves the model unchanged: it re-parametrizes the same model. Table 3 is obtained when some of the loadings are high, and others close to 0. Therefore, after rotation, we can interpret data properly you can see it in Figure 3.

- **Factor Diagram**

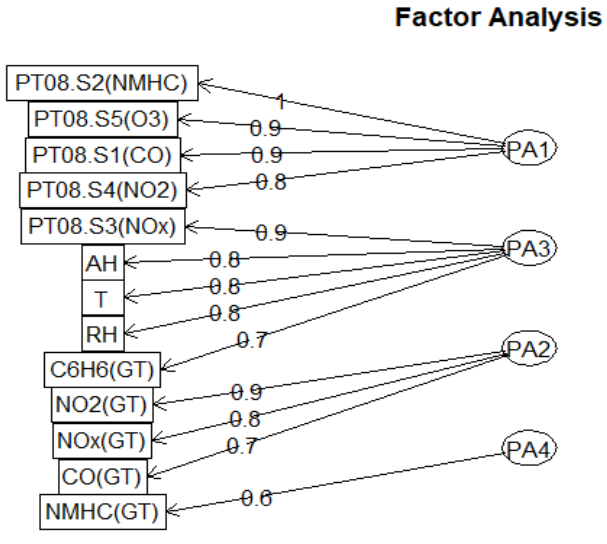


Figure 3:

3.1.4.2. Communalities

Variable	*Communality (Unrotated)*
CO(GT)	0.5190
PT08.S1(CO)	0.9667
NMHC(GT)	0.3511
C6H6(GT)	1.0036
PT08.S2(NMHC)	0.9694
NO_x(GT)	0.8797
PT08.S3(NOx)	0.8493
NO2(GT)	0.8871
PT08.S4(NO2)	0.8211
PT08.S5(O3)	0.8990
T	0.9565
RH	0.8696
AH	1.0070

Table 4:

One assessment of how well this model performs can be obtained from the Communalities. We can see in Table 4: except for two variables other variables that are close to one. That indicates that the model explains most of the variation in those variables, **C6H6(GT), AH, PT08.S1(CO), PT08S2(NMHC),** and **T**. The model explained **AH** best. And not bad for the **NO_x (GT), PT08.S3(NO_x), NO2(GT), PT08.S4(NO2), PT08.S5(O3), RH**. Other variables **NMHC(GT)** and **CO(GT**) do not do a good job, explaining only about half of the variation or less.

3.1.5 Factor Analysis using the Maximum Likelihood Method and Principal Component

Maximum Likelihood Method

```
Proportion Var        0.50 0.18 0.13 0.02
Cumulative Var        0.50 0.68 0.81 0.83
Proportion Explained  0.60 0.22 0.16 0.02
Cumulative Proportion 0.60 0.82 0.98 1.00
```

Principal Component Method

```
                       PA1  PA2  PA3  PA4
SS loadings           6.52 2.79 1.21 0.45
Proportion Var        0.50 0.21 0.09 0.03
Cumulative Var        0.50 0.72 0.81 0.84
Proportion Explained  0.59 0.25 0.11 0.04
Cumulative Proportion 0.59 0.85 0.96 1.00
```

- The ML method explains 83% of the total proportion of variation of the dataset and the PC method explains 84% of the total population variation of the dataset. Therefore, we can conclude that both methods do factor analysis same way as the data set.

3.1.6 Bartlett – correlated Likelihood ratio Test

Test	Df	Chi-squared	P-value
H_0: Four factors are sufficient Vs H_1: More factors are needed	32	291.8	0.000001< 0.05(Significance level)

The model Probability value (1.1e^-43) is less than 0.05. Therefore, we can conclude that the 4-factor model is not sufficient at a 5% significance level. We can not properly fit a factor model to describe this particular data and conclude that the factor model does not work with this particular dataset. This is perhaps due to some non-linearity. Whatever the case, it does not look like this yields a good-fitting factor model.

3.2. Confirmatory Factor Analysis

```
  Estimator                                      ML
  Optimization method                        NLMINB
  Number of model parameters                     31

  Number of observations                       9357

Model Test User Model:

  Test statistic                           97749.905
  Degrees of freedom                          60
  P-value (Chi-square)                         0.000

Model Test Baseline Model:

  Test statistic                          247795.890
  Degrees of freedom                          78
  P-value                                      0.000

User Model versus Baseline Model:

  Comparative Fit Index (CFI)                  0.606
  Tucker-Lewis Index (TLI)                     0.487

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)           -97571.609
  Loglikelihood unrestricted model (H1)   -48696.657

  Akaike (AIC)                            195205.219
  Bayesian (BIC)                          195426.679
  Sample-size adjusted Bayesian (SABIC)   195328.166
```

In the confirmatory factor analysis also, we can see CFI = 0.606 & TLI = 0.487. These values are also not good. Because if it is a good model those two values are nearly equal to 1.

04. Conclusion and recommendations

- According to the analysis, Kaiser method here I can find four eigenvalues that are greater than 1. Therefore, we conclude using the Kaiser method 4 factors are retained to represent the data. As well as those four factors explained 92.08% of the total population variance also greater than 80%. As well as in the scree plot also gives a clear "Elbow" after the fifth factor. Then 4 factors are retained. Finally, I hope the 4-factor model can represent the data adequately.
- Without rotating the data, it is difficult to interpret the relationship between factors and variables. Therefore, we rotate the loading matrix using the "varimax" method. Then following the result, we can obtain.

- Factor 1: Strongly correlated with **PT08.S2(NMHC), PT08.S5(O3), PT08.S1(CO), PT08.S4(NO2)** and they are positively correlated with factor 1. And AH, T, RH, and C6H6 variables are moderately correlated with factor 1.
- Factor 2: Strongly correlated with **NO2(GT), NOX(GT), CO(GT)** and they also positively associate with factor 2.
- Factor 3: Strongly correlated with **PT08.S3(NOX), AH, T. RH, C6H6(GT)** and they also positively associate with factor 3.
- Factor 4: It is strongly correlated with NMHC(GT) and its also positively correlated with factor 4.

- The total communality is 10.9791. The proportion of the total variance explained by the four factors is (10.9791/13) *100=**84.45%.**
- Though it explained almost more than 90% of the total variance, in the Bartlett – likelihood ratio test we cannot adequately fit the model at a 5% significance level. We cannot properly fit a factor model to describe this particular data and conclude that the factor model does not work with this particular dataset. Therefore, the next step could be to drop variables from the data set to obtain a better-fitting model. After dropping several variables also, I cannot properly fit a factor model. Therefore, factor models do not work with this dataset.
- The recommendations consider collecting additional data to enhance the stability and reliability of the factor analysis result.

05. References

Data set - https://archive.ics.uci.edu/dataset/360/air+quality

Kim, J., & Mueller, C. W. (1978). Factor Analysis: Statistical Methods and Practical Issues. SAGE Publications.

RPubs - Exploratory Factor Analysis in R

Factor Analysis on "Women Track Records" Data with R and Python | by Rukshan Pramoditha | Towards Data Science

Intro Guide to Factor Analysis (python) | by Wayne Wooyoung Hong | Medium

06. Appendices

6.1. Part of the data set.

	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT08.S5(O3)	T	RH	AH
1	2.6	1360	150	11.9	1046	166	1056	113	1692	1268	13.6	48.9	0.7578
2	2.0	1292	112	9.4	955	103	1174	92	1559	972	13.3	47.7	0.7255
3	2.2	1402	88	9.0	939	131	1140	114	1555	1074	11.9	54.0	0.7502
4	2.2	1376	80	9.2	948	172	1092	122	1584	1203	11.0	60.0	0.7867
5	1.6	1272	51	6.5	836	131	1205	116	1490	1110	11.2	59.6	0.7888
6	1.2	1197	38	4.7	750	89	1337	96	1393	949	11.2	59.2	0.7848
7	1.2	1185	31	3.6	690	62	1462	77	1333	733	11.3	56.8	0.7603
8	1.0	1136	31	3.3	672	62	1453	76	1333	730	10.7	60.0	0.7702
9	0.9	1094	24	2.3	609	45	1579	60	1276	620	10.7	59.7	0.7648
10	0.6	1010	19	1.7	561	-200	1705	-200	1235	501	10.3	60.2	0.7517

6.2. R codes

```
#Import Required Libraries

library(tidyverse)
library(factoextra)
library(psych)
library(corrplot)
library(ggplot2)
library(skimr)
library(performance)
library(GPArotation)
library(lavaan)

# Read the data set
Air_quality <- read_csv("/d/4th year/ST405/Mini Project Practice/AirQualityUCI.csv")
view(Air_quality)
```

```
# Clean the data set
Air_quality <- Air_quality[, -c(1,2,16,17)]
view(Air_quality)
missing_count <- colSums(is.na(Air_quality))
missing_count
Air_quality <- na.omit(Air_quality)
#Summary of the data set
summary(Air_quality)
#Structure of the data set
str(Air_quality)
#Dimension of the data set
dim(Air_quality)
#Covariance Matrix
Cov_air_quality <- cov(Air_quality)
cov(Air_quality)
# Standardized Data
Cor_air_quality <- cor(Air_quality)
Cor_air_quality
#Correlation plot
corrplot(Cor_air_quality, method = "color", type = "upper", order = "original",
      tl.cex = 0.7, tl.col = "black", col = colorRampPalette(c("blue", "white", "red"))(200))
#Apply KMO Test
KMO(Air_quality)
#calculate Eigenvalues & Eigen vectors
eigen_air_quality <- eigen(Cor_air_quality)
#extract eigenvalues
eigenvalues <- eigen_air_quality$values
round(eigenvalues,4)
#extract eigen vectors
eigenvectors <- eigen_air_quality$vectors
eigenvectors
#Variance explained
varianced_explained <- eigenvalues/sum(eigenvalues)
round(varianced_explained,4)
varianced_explained_prop <- varianced_explained*100
round(varianced_explained_prop,4)
#Explained Cumulative Proportion
cummulative_varianced_explained <-cumsum(varianced_explained)
round(cummulative_varianced_explained,4)
cummulative_varianced_explained_prop <- cummulative_varianced_explained*100
round(cummulative_varianced_explained_prop,4)
#scree plot
num_eigenvalues <- 1:length(varianced_explained_prop)
```

```
plot(num_eigenvalues, varianced_explained_prop,type = "b",
    xlab = "Number of Eigenvalues", ylab = "Proportion of Variance Explained",
    main = "Scree Plot",col = "blue") + axis(1, at = num_eigenvalues)
#Factor Analysis from "Principal Component method"
air_quality_PC<- fa(Cor_air_quality ,nfactors =4 ,rotate = "none",n.obs
= nrow(Air_quality) ,covar = TRUE,fm = "pa")
air_quality_PC
#Get unrotated loading from PC method
unrotated_pc_loadings <-
as.data.frame(unclass(air_quality_PC$loadings))
round(unrotated_pc_loadings,4)
#Factor Analysis from "Maximum Likelihood Method"
air_quality_ML <- fa(Cor_air_quality,nfactors = 4,rotate = "none",n.obs
= nrow(Air_quality) , covar = TRUE, fm = 'ml')
air_quality_ML
# Rotate the Pc method factor loadings using "Varimax" method
air_quality_PC_rotate <- fa(Cor_air_quality ,nfactors = 4,rotate =
"varimax",n.obs = nrow(Air_quality) ,covar = TRUE,fm = 'pa')
air_quality_PC_rotate
rotated_pc_loadings <-
as.data.frame(unclass(air_quality_PC_rotate$loadings))
round(rotated_pc_loadings,4)
#Rotated Pc communalities
rotated_pc_com <-
as.data.frame(unclass(air_quality_PC_rotate$communality))
round(rotated_pc_com,4)
#Graph Factor Loading Matrices
fa.diagram(air_quality_PC_rotate)
#Comfirmatory FA
variables<-Air_quality
new_names <- c(
  "PT08.S5(O3)" = "PT08_S5_O3",
  "PT08.S1(CO)" = "PT08_S1_CO",
  "PT08.S4(NO2)" = "PT08_S4_NO2",
  "PT08.S3(NOx)" = "PT08_S3_NOX",
  "AH" = "AH",
  "T" = "T",
  "RH" = "RH",
  "C6H6(GT)" = "C6H6_GT",
  "NO2(GT)" = "NO2_GT",
  "NOx(GT)" = "NOX_GT",
  "CO(GT)" = "CO_GT",
  "NMHC(GT)" = "NMHC_GT",
  "PT08.S2(NMHC)"="PT08_S2_NMHC")
```

```
colnames(variables)<-new_names
variables<-scale(variables)
library(semPlot)model <- model <-model <- '
Factor1 =~ PT08_S2_NMHC + PT08_S5_O3 + PT08_S1_CO + PT08_S4_NO2
Factor2 =~ NO2_GT + NOX_GT + CO_GT
Factor3 =~ PT08_S3_NOX + AH + T +RH + C6H6_GT
Factor4 =~ NMHC_GT
'
fit <- cfa(model, data = variables)
summary(fit,fit.measures = TRUE)
```