Explanatory Factor Analysis and Confirmatory Factor Analysis on Gas Turbine CO and NOx (NO+NO2) Emission

S/18/843

1 Introduction

Explanatory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) are statistical techniques used to uncover underlying patterns in datasets. EFA explores the relationships among observed variables to identify latent factors, while CFA validates a pre-specified factor structure. In this report, we employ both EFA and CFA to analyze our dataset, aiming to understand its underlying structure and validate our theoretical model. Through these analyses, we aim to gain deeper insights into the dataset's complexity and relationships, facilitating more informed decision-making.

2 Methodology

I perform factor analysis using the Gas Turbine CO and NOx (NO+NO2) Emission dataset. In order to analyses flue gas emissions, specifically the dataset comprises 7385 occurrences of 11 sensor measures aggregated over one hour (by means of average or total) from a gas turbine located in Turkey's northwest region. The only variables in the dataset are numerical ones.

Variable Information

"AT" - Ambient temperature

"AP" - Ambient pressure

"AH" - Ambient humidity

"AFDP" - Air filter difference pressure

"GTEP" - Gas turbine exhaust pressure

"TIT" - Turbine inlet temperature

"TAT" - Turbine after temperature

"CDP" - Compressor discharge pressure

"TEY" - Turbine energy yield

"CO" - Carbon monoxide

"NOx" - Nitrogen oxides

Since different variables have different measurement units, I standardized the whole data set. Following that, I'll concentrate on explanatory factor analysis using techniques such as Eigen values and Eigen vectors, factor loadings, and communalities. I'll also concentrate on a confirmatory factor model that involves a few latent variables.

3 Results and Discussion

3.1 Exploratory Factor analysis

Before proceeding with the analysis, it's important to check whether the dataset is suitable for factor analysis. To do this, we perform the Kaiser-Meyer-Olkin (KMO) test.

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = mydata)
Overall MSA =  0.71
MSA for each item =
  AT   AP   AH AFDP GTEP  TIT  TAT  TEY  CDP   CO  NOX
0.40 0.30 0.41 0.94 0.96 0.66 0.46 0.70 0.84 0.88 0.71
```

Removing the variable (AP), which had a low contribution to the overall MSA value, resulted in an increase of the overall MSA value to 0.76. This indicates an improvement over the previous model.

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = mydata)
Overall MSA =   0.76
MSA for each item =
  AT   AH AFDP GTEP  TIT  TAT  TEY  CDP   CO  NOX
0.46 0.58 0.92 0.98 0.68 0.46 0.76 0.82 0.90 0.77
```

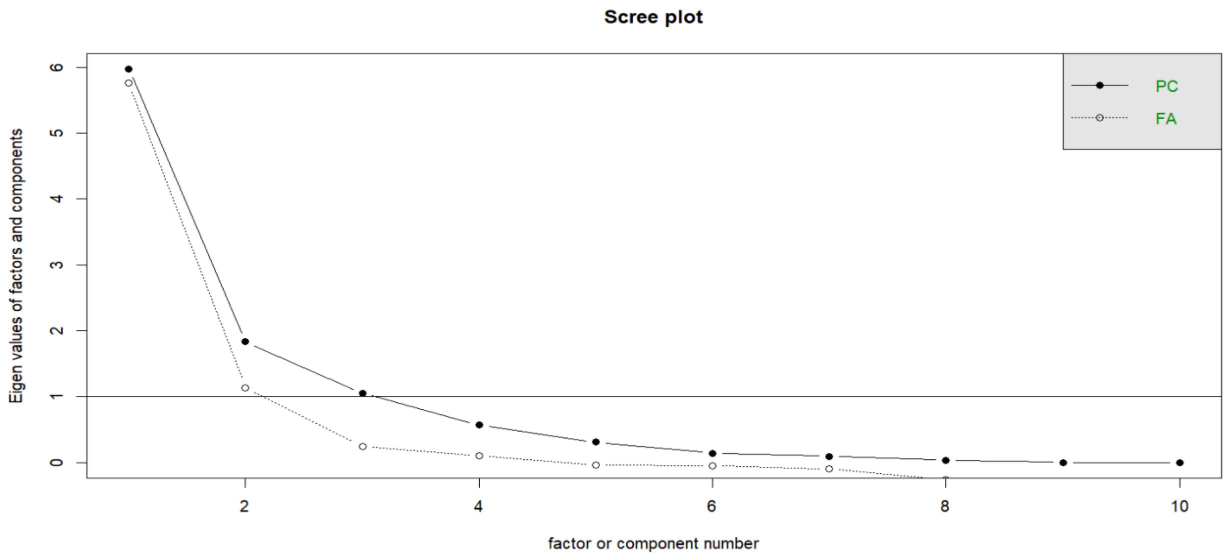
3.1.1 Eigen values and variances of each Component.

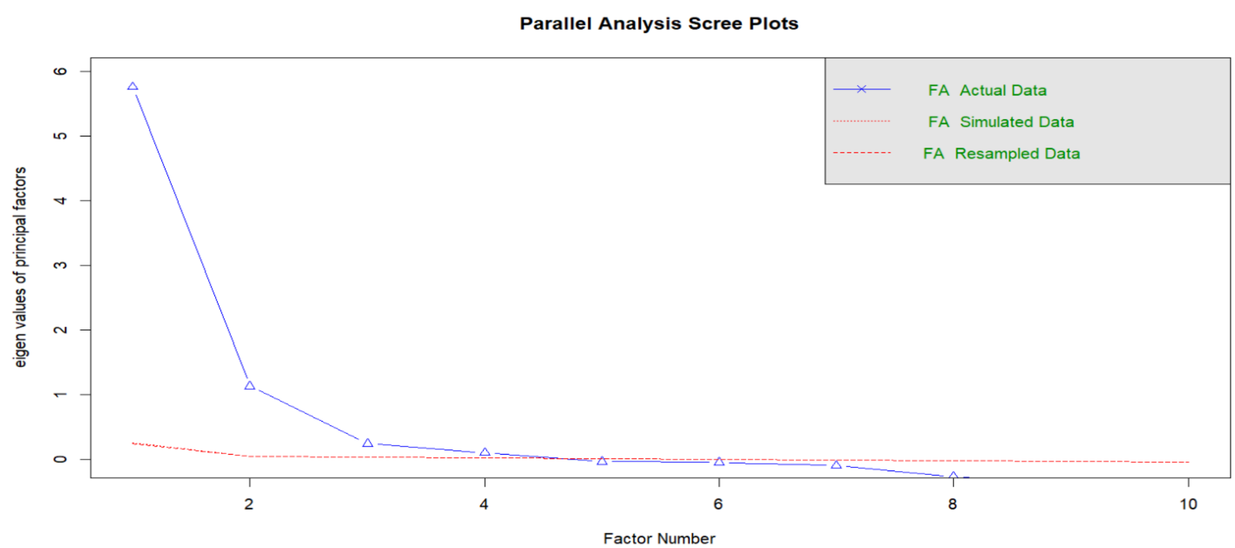
Component	Eigen value	Proportion	Cumulative Proportion
1	5.9704118081	0.8642863	0.8642863
2	1.8308611850	0.09405783	0.95834414
3	1.0526185444	0.03153228	0.98987642
4	0.5662281614	0.009092296	0.99896871
5	0.3075746872	0.0006682286	0.9996369431
6	0.1413411380	0.0002790371	0.9999159802
7	0.0958687573	5.872675e-05	9.999747e-01
8	0.0327451634	2.521936e-05	9.999999e-01
9	0.0016093547	7.368529e-08	1.000000e+00
10	0.0007412006	7.732737e-18	1.000000e+00

Considering the eigenvalues, we observe that the first three eigenvalues are greater than 1. This suggests that three factors are sufficient for analyzing this dataset.

The cumulative proportion column indicates that the first three components explain 98.99% of the total variance, suggesting that they capture most of the variability in the dataset. This suggests that three factors for analyzing this dataset would be sufficient.

3.1.2 Scree Plot





Parallel Scree plot also suggest that three factors are sufficient to analyze this dataset.

3.1.3 Factor Analyze using PC method and ML method

The factor loadings using PC method and ML method with varimax rotation are given below

```
Using PC
                        PA1  PA2  PA3
SS loadings            4.22 2.53 2.45
Proportion Var         0.42 0.25 0.24
Cumulative Var         0.42 0.67 0.92
Proportion Explained   0.46 0.28 0.27
Cumulative Proportion  0.46 0.73 1.00

Tucker Lewis Index of factoring reliability =  -13219141
RMSEA index =  2462.25  and the 90 % confidence intervals are  NA
2462.403
```

```
Using ML Method
                        ML1  ML2  ML3
SS loadings            4.55 2.00 1.59
Proportion Var         0.45 0.20 0.16
Cumulative Var         0.45 0.66 0.81
Proportion Explained   0.56 0.25 0.20
Cumulative Proportion  0.56 0.80 1.00

Tucker Lewis Index of factoring reliability =  0.815
RMSEA index =  0.291  and the 90 % confidence intervals are  0.287
0.296
```

In the Principal Components (PC) method, 92% of the total variance is explained, but high values of the Tucker Lewis Index and RMSEA suggest poor model fitting. On the other hand, the Maximum Likelihood (ML) method explains 81% of the total variance, with good model fit indicated by a Tucker Lewis Index of 0.815 and RMSEA index of 0.291. Hence, the ML method is more suitable for this dataset.

3.1.4 Factor Loadings

variable	Factor 1	Factor 2	Factor 3
AT	0.07591591	0.98657682	-0.126290676
AH	-0.13126564	-0.46399169	-0.041725906
AFDP	0.77213094	0.46170752	0.356543972
GTEP	0.82280374	0.18508061	0.414068306

TIT	0.94692698	0.28056918	0.148346519
TAT	-0.30526152	0.11392010	-0.942935739
TEY	0.91652243	0.09010936	0.386144046
CDP	0.87874242	0.19153182	0.434368026
CO	-0.70984902	-0.32110048	0.164182505
NOX	-0.38114384	-0.57234464	-0.001886582

Interpretation of factors:

Factor 1: The factor loadings indicate strong positive correlations with variables AFDP, GTEP, TIT, TEY, and CDP. Conversely, negative correlations are observed with variable CO. Among these correlations, the strongest associations are found with variables TIT, TEY, and CDP.

Factor 2: The factor loading show strong positive correlation with variable AT. While moderate negative correlation with AH and NOX variables

Factor 3: positively correlated with variables AFDP, GTEP, TIT, TEY, and CDP, and negatively correlated with variable TAT. Among these, the strongest negative correlations are observed with variable TAT.

3.1.5 Communalities

variables	Communalities
AT	0.9950464
AH	0.2342600
AFDP	0.9364836
GTEP	0.8827134
TIT	0.9973965
TAT	0.9952902
TEY	0.9972403
CDP	0.9975483
CO	0.6339470
NOX	0.4728526

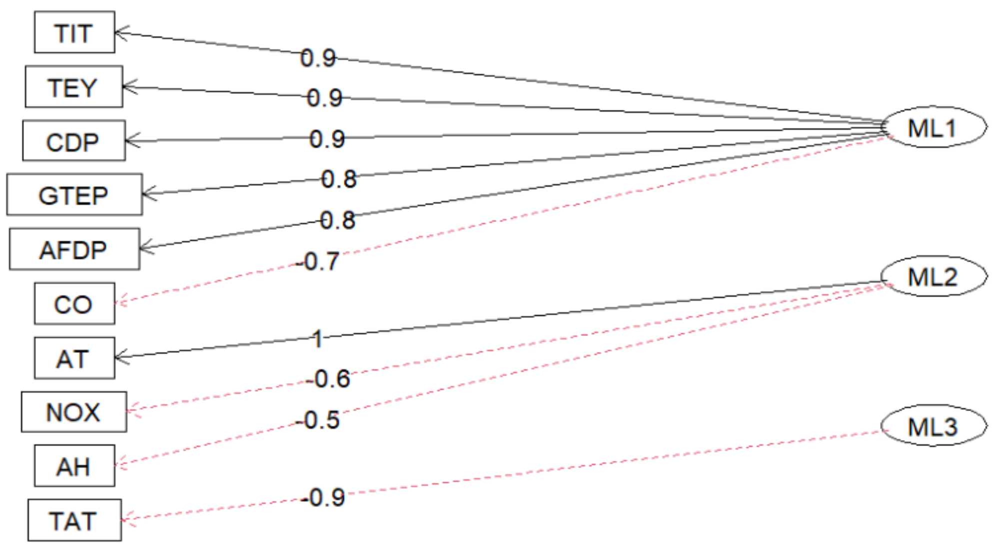
With a high communality of 0.995, variable AT mainly explains its variance by the components that were obtained during the analysis.

The high communalities of the variables AFDP, GTEP, TIT, TAT, and TEY, which range from 0.882 to 0.997, indicate that the factors contribute to a sizable amount of the variance in these variables.

With a moderate communality of 0.634, variable CO shows that the factors account for a significant portion of its variance.

Compared to other variables, variable NOX has a lower communality of 0.473, indicating that a smaller percentage of its variance is explained by the factors.

3.1.6 FA Diagram



3.1.7 Hypothesis Testing

H0 -: Three factors are sufficient

VS

H1 -: More factors are needed

```
The harmonic n.obs is  7385 with the empirical chi square  2039.31  with prob
<  0
```

p value(0.05) > 0. We cannot reject H0. Therefore we can conclude that 3 factors are sufficient to at 5% significance level

3.2 Confirmatory Factor Analysis.

```
model <-'
Factor1 = ~CDP+TEY+GTEP+AFDP+TIT+CO
Factor2 = ~AT+NOX+AH
Factor3 = ~ TAT
```

```
Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                        22

  Number of observations                          7384

Model Test User Model:

  Test statistic                             60132.857
  Degrees of freedom                                33
  P-value (Chi-square)                           0.000

Model Test Baseline Model:

  Test statistic                            152586.133
  Degrees of freedom                                45
  P-value                                        0.000

User Model versus Baseline Model:

  Comparative Fit Index (CFI)                    0.606
  Tucker-Lewis Index (TLI)                       0.463
```

4 Conclusions and Recommendations

Three variables are identified by the analysis and the empirical chi-squared test supports their presence, suggesting that these factors sufficiently describe the dataset. 81% of the total variance in the dataset is explained by the two-factor model.

After applying the varimax rotation to the factor loadings obtained from the ML method, we observed that Factor 1 exhibits strong positive correlations with AFDP, GTEP, TIT, TEY, CO and CDP. Additionally, there are both positive and negative correlations between Factor 1 and other variables. Notably, only the variable TAT shows a strong negative correlation with Factor 3. And the variable AT strongly correlated with factor 2 and NOX and AH negative moderately correlated with factor 3. However, without employing any factor rotation technique, the factor loadings do not provide a clear conclusion about the model.

In CFA, the p-value being less than 0.001 suggests that the user model fits better than the baseline model. However, the CFI and TLI values are slightly low, indicating a poor fit. The RMSEA value of 0.497 also suggests a poor fit, while the SRMR value of 0.172 indicates a moderate fit.

Overall, although the model offers some understanding of the connections between variables and factors, it seems to be not well-suited to the data according to the fit indices. It may require further adjustments or alternative modeling strategies to enhance its fit.

5 References

https://archive.ics.uci.edu/dataset/551/gas+turbine+co+and+nox+emission+data+set

https://online.stat.psu.edu/stat505/book/export/html/691

https://bookdown.org/sz_psyc490/r4psychometics/factor-analysis.html

6 Appendix

6.1 Part of dataset

	A	B	C	D	E	F	G	H	I	J	K	L
1	AT	AP	AH	AFDP	GTEP	TIT	TAT	TEY	CDP	CO	NOX	
2	1.9532	1020.1	84.985	2.5304	20.116	1048.7	544.92	116.27	10.799	7.4491	113.25	
3	1.2191	1020.1	87.523	2.3937	18.584	1045.5	548.5	109.18	10.347	6.4684	112.02	
4	0.94915	1022.2	78.335	2.7789	22.264	1068.8	549.95	125.88	11.256	3.6335	88.147	
5	1.0075	1021.7	76.942	2.817	23.358	1075.2	549.63	132.21	11.702	3.1972	87.078	
6	1.2858	1021.6	76.732	2.8377	23.483	1076.2	549.68	133.58	11.737	2.3833	82.515	
7	1.8319	1021.7	76.411	2.841	23.495	1076.4	549.92	133.58	11.829	2.0812	81.193	
8	2.074	1022	75.974	2.7981	22.945	1073.7	549.98	131.53	11.687	2.2529	83.171	
9	1.7824	1022.6	73.535	2.8327	23.337	1075.7	550.01	133.18	11.745	3.735	85.749	
10	1.593	1023.2	72.873	2.8729	23.654	1078.5	550.06	135.38	11.772	3.6398	86.491	
11	1.6819	1023.8	72.441	2.9058	23.463	1077.9	550.12	134.86	11.742	3.5866	86.328	
12	1.9002	1024.5	71.376	2.9126	23.562	1078.2	550.12	134.98	11.77	3.5605	84.117	
13	1.7797	1025.1	68.528	2.8725	23.276	1077	550.03	134.21	11.782	3.6902	85.317	

6.2 R codes

```
library(tidyverse)
library(psych)
library(ggplot2)
library(corrplot)
library(ggcorrplot)
library(nFactors)
library(skimr)
library(performance)
```

```
library(lavaan)
mydata<-read.csv("../Data/mydata.csv")
mydata
str(mydata)
any(is.na(mydata))
mydata <- apply(mydata,2,scale)
head(mydata)
KMO(mydata)
mydata <- mydata[,-c(2)]
#head(mydata)
KMO(mydata)
cortest.bartlett(mydata)
mydata_cov <- cov(mydata)
mydata_cov
mydata_cov_eigen <- eigen(mydata_cov)
# eigen values
mydata_cov_eigen$values
# eigen vectors
mydata_cov_eigen$vectors
pca<-princomp(mydata_cov)
pca
summary(pca)
scree(mydata)
mydata_PC<- fa(mydata_cov ,nfactors = 3,rotate = "varimax",n.obs
= 7385 ,covar = TRUE,fm = "pa",max.iter = 1000)
mydata_PC
rotated_pc_loadings <-as.data.frame(unclass(mydata_PC$loadings))
rotated_pc_loadings
rotated_pc_com <-as.data.frame(unclass(mydata_PC$communality))
rotated_pc_com
mydata_ML <- fa(mydata_cov,nfactors = 3,rotate = "varimax",n.obs
= 7385 , covar = TRUE, fm = 'ml')
```

```
mydata_ML
rotated_ml_loadings <-as.data.frame(unclass(mydata_ML$loadings))
rotated_ml_loadings
rotated_ml_com <- as.data.frame(unclass(mydata_ML$communality))
rotated_ml_com
fa.diagram(mydata_ML)
features <-mydata[,c("AFDP","GTEP","TIT","TEY","CDP","CO","NOX","AT","AH","TAT")]
#define the CFA model
model <- '
 Factor1 =~CDP+TEY+GTEP+AFDP+TIT+CO
 Factor2 =~AT+NOX+AH
 Factor3=~ TAT
'# Fit the CFA model
fit <- cfa(model, data = features)
# Assess model fit
#summary(fit, fit.measures = TRUE)
# Standardized estimates (factor loadings)
#parameterEstimates(fit, standardized = TRUE, ci = TRUE)
summary(fit,fit.measures=T,standardized=T)
```