PCA & Factor Analysis on Air Pollution Data R K Puram, New Delhi

Vilas Wakale

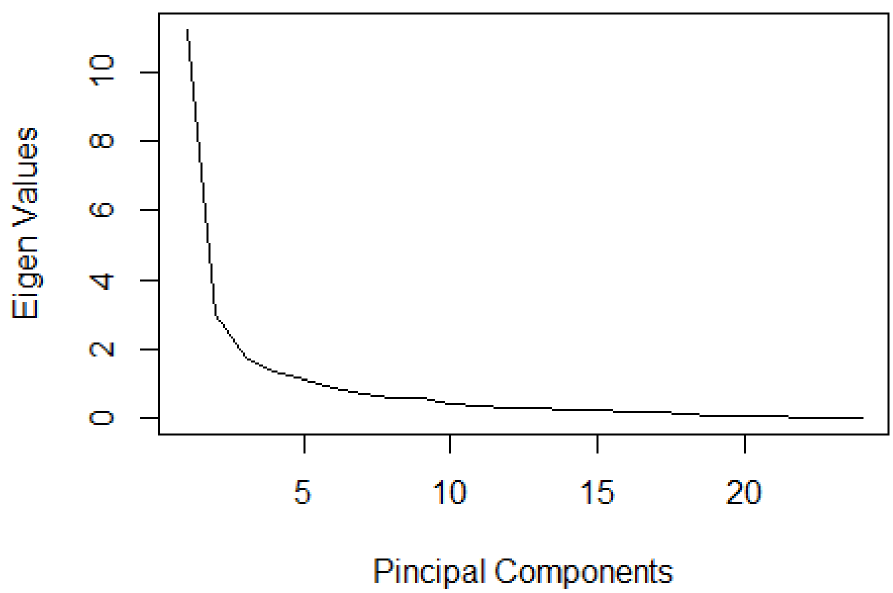
7 January 2018

```
#==============================================================================
#
#
# PRINCIPAL COMPONENT ANALYSIS
#
#==============================================================================
#
#install.packages("car")
#install.packages("nortest")
#
library(psych)
library(car)

library(foreign)
library(MASS)
library(lattice)
library(nortest) # Anderson Darling
#
getwd()

## [1] "D:/Vilas/00 Great Lakes Engagement/07 BACP Course Mentoring/02 BACP.O
CT/03 Module 3 Advanced Statistics/03 Week 3 Data Reduction Techniques"

setwd("D:\\Vilas\\00 Great Lakes Engagement\\07 BACP Course Mentoring\\02 BAC
P.OCT\\03 Module 3 Advanced Statistics\\03 Week 3 Data Reduction Techniques")

RKP=read.csv("RKP_Delhi_Edited.csv", header = TRUE)
names(RKP)

str(RKP)

head(RKP)

summary(RKP)

RKP <- na.omit(RKP)
summary(RKP)

str(RKP)

## 'data.frame':    347 obs. of  27 variables:
#
```

```
# Understanding Correlation
#
RKPCorr <- cor(RKP[-c(1,2,22)])
# Ignoring Non Numeric and unwanted variables such as Sr. No., Date, Weather
RKPCorr

# Barlett Sphericity Test for checking the possibility
# of data dimension reduction
print(cortest.bartlett(RKPCorr,nrow(RKP)))

## $chisq
## [1] 9585.526
##
## $p.value
## [1] 0
##
## $df
## [1] 276

#
# Finding out the Eigen Values and Eigen Vectors.
A<-eigen(RKPCorr)
eigenvalues<-A$values
eigenvectors<-A$vectors
eigenvalues

##  [1] 11.21500818  2.98569744  1.78929681  1.34697073  1.11005651
##  [6]  0.89010563  0.73051047  0.60189395  0.57041627  0.41180518
## [11]  0.38091795  0.32580388  0.31021327  0.27113681  0.24302887
## [16]  0.21258838  0.16697205  0.14337784  0.09715854  0.08089151
## [21]  0.04637548  0.03109583  0.02014172  0.01853671

#
# We will consider Components which are having eigenvalues > 1 unit
# i.e. PC1 - PC5.
#
eigenvectors

# Getting the Loadings and Communality
pc<-principal(RKP[-c(1,2,22)],nfactors = length(RKP[-c(1,2,22)]),rotate="none
")
pc

## Principal Components Analysis
## Call: principal(r = RKP[-c(1, 2, 22)], nfactors = length(RKP[-c(1,
##     2, 22)]), rotate = "none")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                       PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8   PC9
## NO                   0.86  0.06  0.24  0.23  0.08  0.18  0.02 -0.08  0.00
## CO                   0.69 -0.15  0.23  0.06 -0.02  0.37 -0.01 -0.20 -0.29
## NO2                  0.68  0.37  0.23 -0.02 -0.29 -0.30  0.08 -0.04 -0.20
```

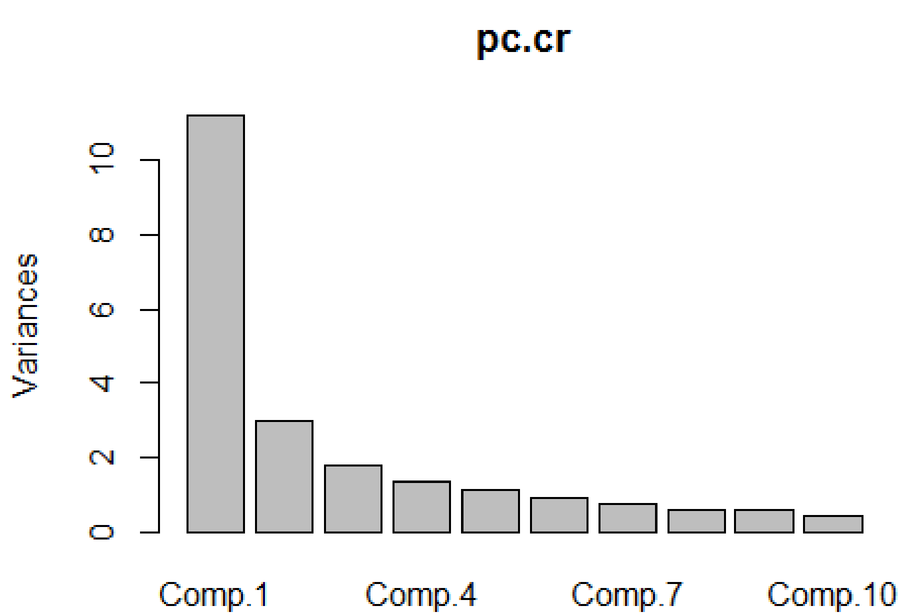
```
## O3                    0.07  0.79 -0.03  0.14 -0.27 -0.02  0.24 -0.10 -0.03
## SO2                   0.44  0.44 -0.30  0.36 -0.12  0.11 -0.51  0.01 -0.02
## PM2.5                 0.88  0.09 -0.10 -0.17  0.22 -0.03  0.03 -0.18  0.05
## Benzene               0.91 -0.23  0.13  0.06  0.01 -0.03  0.00  0.01  0.17
## Toulene               0.85 -0.26  0.27  0.18 -0.05 -0.01  0.03  0.06  0.19
## P_Xylene              0.86 -0.18  0.34  0.15  0.03  0.12 -0.01  0.00  0.17
## NOx                   0.83 -0.06  0.32  0.27  0.00  0.15  0.04 -0.07 -0.07
.

.

.

## P_Xylene             -0.01  0.01  0.04  0.08  0.06 -0.07  1  7.8e-16 1.7
## NOx                   0.05 -0.05 -0.02 -0.09 -0.03 -0.05  1  1.1e-15 2.1
## PD_PM2.5             -0.01 -0.14  0.11 -0.01 -0.01  0.02  1  6.7e-16 2.0
## PD_PM10               0.01  0.14 -0.09  0.00  0.00 -0.01  1  7.8e-16 2.8
## PD_NO2                0.02  0.00 -0.01  0.01  0.01  0.00  1  1.2e-15 4.7
## PD_SO2                0.05 -0.02  0.00 -0.01  0.00  0.01  1  1.2e-15 4.7
## PD_CO                -0.01  0.01  0.00  0.00  0.00  0.00  1  3.3e-16 3.7
##
##                         PC1  PC2  PC3  PC4  PC5  PC6  PC7  PC8  PC9 PC10
## SS loadings           11.22 2.99 1.79 1.35 1.11 0.89 0.73 0.60 0.57 0.41
## Proportion Var         0.47 0.12 0.07 0.06 0.05 0.04 0.03 0.03 0.02 0.02
## Cumulative Var         0.47 0.59 0.67 0.72 0.77 0.81 0.84 0.86 0.88 0.90
## Proportion Explained   0.47 0.12 0.07 0.06 0.05 0.04 0.03 0.03 0.02 0.02
## Cumulative Proportion  0.47 0.59 0.67 0.72 0.77 0.81 0.84 0.86 0.88 0.90
##                       PC11 PC12 PC13 PC14 PC15 PC16 PC17 PC18 PC19 PC20
## SS loadings           0.38 0.33 0.31 0.27 0.24 0.21 0.17 0.14 0.10 0.08
## Proportion Var        0.02 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.00 0.00
## Cumulative Var        0.92 0.93 0.94 0.96 0.97 0.97 0.98 0.99 0.99 1.00
## Proportion Explained  0.02 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.00 0.00
## Cumulative Proportion 0.92 0.93 0.94 0.96 0.97 0.97 0.98 0.99 0.99 1.00
##                       PC21 PC22 PC23 PC24
## SS loadings           0.05 0.03 0.02 0.02
## Proportion Var        0.00 0.00 0.00 0.00
## Cumulative Var        1.00 1.00 1.00 1.00
## Proportion Explained  0.00 0.00 0.00 0.00
## Cumulative Proportion 1.00 1.00 1.00 1.00
##
## Mean item complexity =  3.2
## Test of the hypothesis that 24 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0
##  with the empirical chi square  0  with prob <  NA
##
## Fit based upon off diagonal values = 1
```

```r
# Interpreting the variance
#
part.pca<-eigenvalues/sum(eigenvalues)*100
part.pca

## [1] 46.72920074 12.44040599 7.45540336 5.61237805 4.62523545
## [6] 3.70877346 3.04379364 2.50789148 2.37673444 1.71585490
## [11] 1.58715810 1.35751615 1.29255530 1.12973672 1.01262029
## [16] 0.88578492 0.69571688 0.59740768 0.40482725 0.33704796
## [21] 0.19323116 0.12956594 0.08392384 0.07723628

# The 5 PC's are able to explain 75% of Variance.

#Plotting SCREE Graphs
plot(eigenvalues,type="lines",
     xlab="Pincipal Components",ylab="Eigen Values")
```



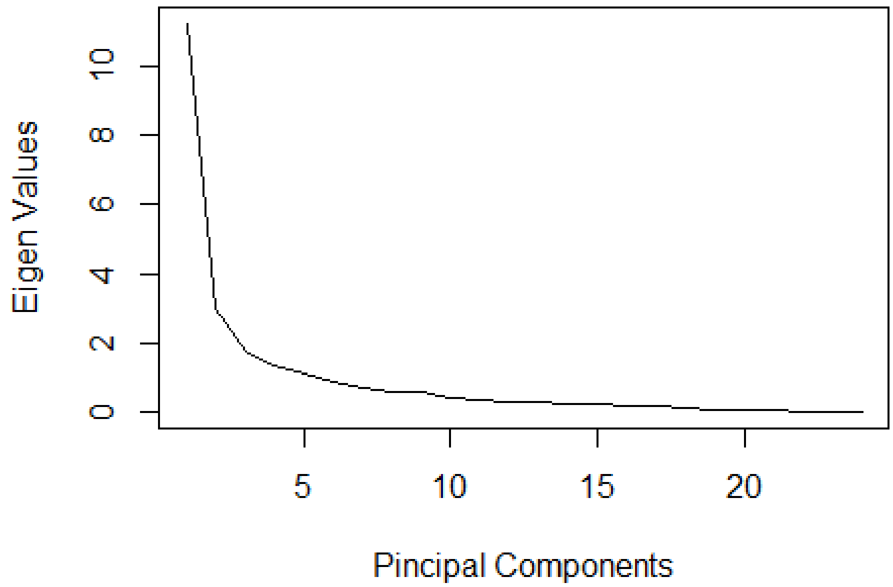

```r
# Principal Components Scoring and Perceptual Map
RKPsc<-scale(RKP[-c(1,2,22)])
z<-as.matrix(RKPsc%*%eigenvectors)
z

pc.cr<-princomp(RKPsc,cor=TRUE)
summary(pc.cr)

## Importance of components:
##                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5
## Standard deviation     3.348882 1.7279171 1.33764599 1.16059068 1.05359219
```

```
## Proportion of Variance 0.467292 0.1244041 0.07455403 0.05612378 0.04625235
## Cumulative Proportion  0.467292 0.5916961 0.66625010 0.72237388 0.76862624
##                            Comp.6     Comp.7     Comp.8     Comp.9
## Standard deviation     0.94345410 0.85469905 0.77581825 0.75525907
## Proportion of Variance 0.03708773 0.03043794 0.02507891 0.02376734
## Cumulative Proportion  0.80571397 0.83615191 0.86123082 0.88499817
##                           Comp.10    Comp.11    Comp.12    Comp.13
## Standard deviation     0.64172048 0.61718550 0.57079232 0.55696793
## Proportion of Variance 0.01715855 0.01587158 0.01357516 0.01292555
## Cumulative Proportion  0.90215672 0.91802830 0.93160346 0.94452901
##                           Comp.14   Comp.15     Comp.16     Comp.17
## Standard deviation     0.52070799 0.4929796 0.461073074 0.408622138
## Proportion of Variance 0.01129737 0.0101262 0.008857849 0.006957169
## Cumulative Proportion  0.95582638 0.9659526 0.974810430 0.981767599
##                            Comp.18     Comp.19    Comp.20     Comp.21
## Standard deviation     0.378652669 0.311702646 0.28441433 0.215349664
## Proportion of Variance 0.005974077 0.004048272 0.00337048 0.001932312
## Cumulative Proportion  0.987741676 0.991789948 0.99516043 0.997092739
##                            Comp.22      Comp.23      Comp.24
## Standard deviation     0.176340086 0.1419215358 0.1361495780
## Proportion of Variance 0.001295659 0.0008392384 0.0007723628
## Cumulative Proportion  0.998388399 0.9992276372 1.0000000000

# Check for Cumulative Proportion

plot(pc.cr)
```



```
#==============================================================================
#
#
# FACTOR ANALYSIS
#
#==============================================================================
#
RKPCorr<-cor(RKP[-c(1,2,22)])
round(RKPCorr, 2)

#
# Kaiser-Meyer-Olkin (KMO) Test : For finding Measure of Sampling Adequacy
KMO(r=RKPCorr)

## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = RKPCorr)
## Overall MSA =  0.88
## MSA for each item =
##                NO                CO               NO2                O3
##              0.84              0.97              0.88              0.73
##               SO2             PM2.5           Benzene           Toulene
##              0.87              0.90              0.92              0.90
##          P_Xylene               NOx              PM10     WindDirection
##              0.90              0.85              0.90              0.80
##               NH3                RH              Temp         WindSpeed
##              0.91              0.66              0.81              0.95
## VerticalWindSpeed             Solar       BarPressure          PD_PM2.5
##              0.64              0.91              0.90              0.89
##           PD_PM10            PD_NO2            PD_SO2             PD_CO
##              0.89              0.90              0.88              0.98

#
# Bartlett's Test of Sphericity:
print(cortest.bartlett(RKPCorr,nrow(RKP)))

## $chisq
## [1] 9585.526
##
## $p.value
## [1] 0
##
## $df
## [1] 276

#
# Finding out the Eigen Values and Eigen Vectors.
A<-eigen(RKPCorr)
eigenvalues<-A$values
eigenvectors<-A$vectors
eigenvalues
```

```
##  [1] 11.21500818  2.98569744  1.78929681  1.34697073  1.11005651
##  [6]  0.89010563  0.73051047  0.60189395  0.57041627  0.41180518
## [11]  0.38091795  0.32580388  0.31021327  0.27113681  0.24302887
## [16]  0.21258838  0.16697205  0.14337784  0.09715854  0.08089151
## [21]  0.04637548  0.03109583  0.02014172  0.01853671

#Plotting SCREE Graphs
plot(eigenvalues,type="lines",
     xlab="Pincipal Components",ylab="Eigen Values")
```




```
# Factor Analysis using Principal Axis Factoring using 5 factors
#
solution<-fa(r=RKPCorr,nfactors=5,rotate = "none",fm="pa")
solution

## Factor Analysis using method =  pa
## Call: fa(r = RKPCorr, nfactors = 5, rotate = "none", fm = "pa")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                        PA1   PA2   PA3   PA4   PA5   h2    u2 com
## NO                    0.86  0.06  0.29  0.15  0.20 0.89 0.107 1.4
## CO                    0.67 -0.12  0.19 -0.01  0.09 0.50 0.495 1.3
## NO2                   0.68  0.37  0.21  0.00 -0.41 0.81 0.188 2.5
## O3                    0.06  0.71  0.01  0.19 -0.15 0.57 0.433 1.3
## SO2                   0.42  0.38 -0.17  0.38  0.04 0.50 0.501 3.3
## PM2.5                 0.88  0.10 -0.14 -0.18  0.15 0.86 0.142 1.2
## Benzene               0.92 -0.23  0.13  0.01  0.01 0.91 0.089 1.2
## Toulene               0.85 -0.26  0.30  0.10 -0.01 0.89 0.107 1.5
```

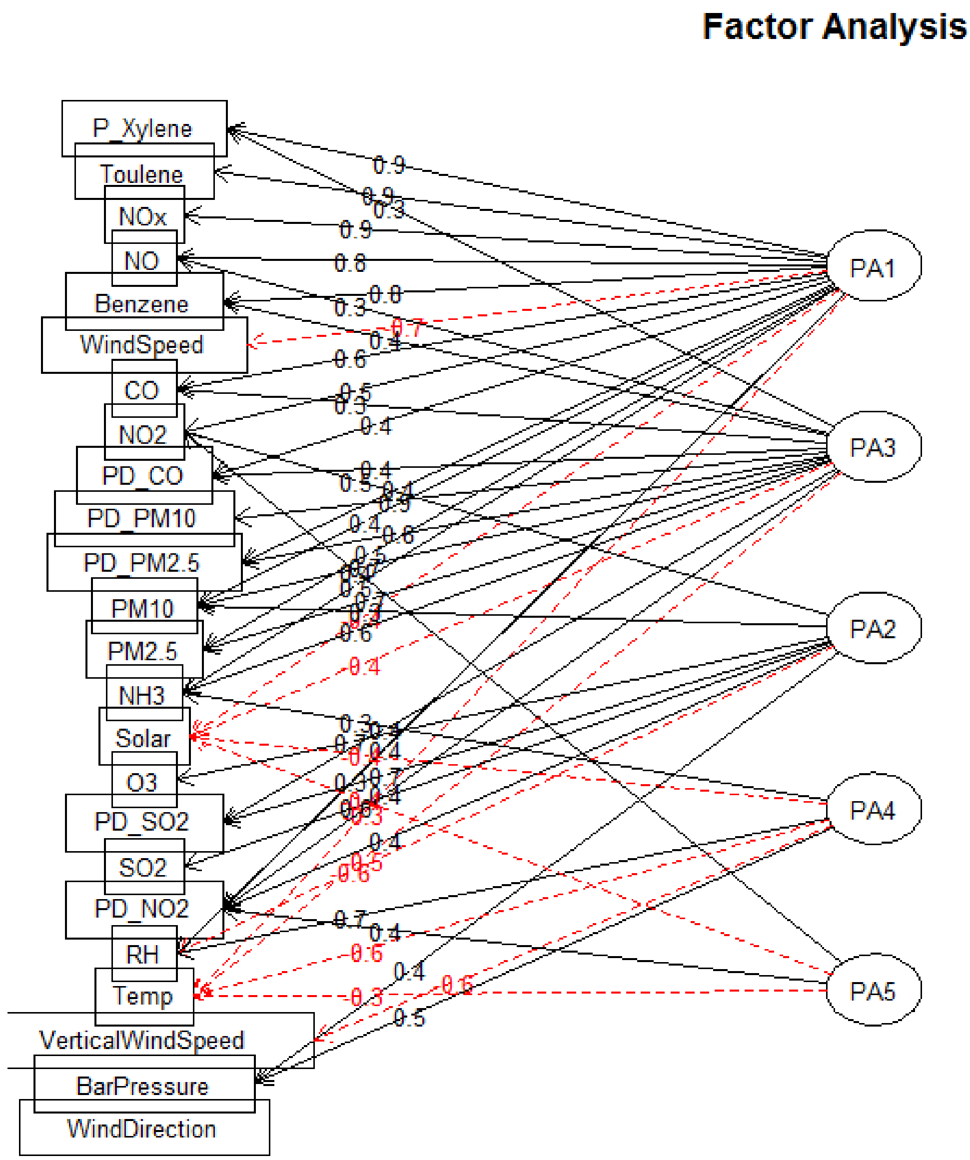
```
## P_Xylene           0.87 -0.18  0.38  0.04  0.13 0.95 0.052 1.5
## NOx                0.83 -0.06  0.37  0.17  0.12 0.86 0.135 1.5
## PM10               0.79  0.32 -0.12 -0.22  0.23 0.84 0.163 1.8
## WindDirection      0.26  0.18  0.05  0.08  0.23 0.16 0.842 3.1
## NH3                0.82 -0.09 -0.19 -0.03 -0.11 0.73 0.272 1.2
## RH                 0.30 -0.91 -0.18  0.09 -0.12 0.96 0.038 1.4
## Temp              -0.70  0.29  0.43 -0.10  0.15 0.79 0.211 2.2
## WindSpeed         -0.66  0.11 -0.31 -0.07  0.21 0.59 0.408 1.8
## VerticalWindSpeed -0.10  0.28  0.32 -0.36 -0.16 0.34 0.656 3.5
## Solar             -0.63  0.41  0.25  0.11  0.23 0.69 0.306 2.5
## BarPressure        0.52  0.08 -0.37  0.39  0.00 0.56 0.439 2.8
## PD_PM2.5           0.84  0.09 -0.28 -0.31  0.08 0.89 0.108 1.6
## PD_PM10            0.75  0.28 -0.29 -0.37  0.13 0.88 0.124 2.2
## PD_NO2             0.62  0.37  0.05 -0.13 -0.36 0.67 0.328 2.4
## PD_SO2             0.59  0.45 -0.25  0.28 -0.01 0.69 0.307 2.8
## PD_CO              0.60 -0.12  0.01 -0.14  0.02 0.40 0.600 1.2
##
##                         PA1  PA2  PA3  PA4  PA5
## SS loadings           11.00 2.72 1.50 0.99 0.73
## Proportion Var         0.46 0.11 0.06 0.04 0.03
## Cumulative Var         0.46 0.57 0.63 0.68 0.71
## Proportion Explained   0.65 0.16 0.09 0.06 0.04
## Cumulative Proportion  0.65 0.81 0.90 0.96 1.00
##
## Mean item complexity =  2
## Test of the hypothesis that 5 factors are sufficient.
##
## The degrees of freedom for the null model are  276  and the objective func
tion was  28.43
## The degrees of freedom for the model are 166  and the objective function w
as  5.24
##
## The root mean square of the residuals (RMSR) is  0.03
## The df corrected root mean square of the residuals is  0.04
##
## Fit based upon off diagonal values = 1
## Measures of factor score adequacy
##                                                   PA1  PA2  PA3  PA4  PA5
## Correlation of (regression) scores with factors   0.99 0.98 0.95 0.89 0.88
## Multiple R square of scores with factors          0.99 0.97 0.91 0.78 0.78
## Minimum correlation of possible factor scores     0.97 0.94 0.81 0.57 0.56
```

```
# Explore Loading if Factors can be balanced.

solution1 <-fa(r=RKPCorr,nfactors=5,rotate = "varimax",fm="pa")
solution1
```

```
## Factor Analysis using method =  pa
## Call: fa(r = RKPCorr, nfactors = 5, rotate = "varimax", fm = "pa")
## Standardized loadings (pattern matrix) based upon correlation matrix
```

```
##                      PA1   PA3   PA2   PA4   PA5   h2    u2 com
## NO                  0.82  0.32  0.29  0.09 -0.14 0.89 0.107 1.7
## CO                  0.64  0.30  0.03  0.10  0.00 0.50 0.495 1.5
## NO2                 0.52  0.28  0.49 -0.21  0.42 0.81 0.188 3.9
## O3                 -0.09  0.00  0.70 -0.27  0.03 0.57 0.433 1.3
## SO2                 0.18  0.16  0.63  0.22 -0.06 0.50 0.501 1.6
## PM2.5               0.50  0.74  0.21  0.14  0.00 0.86 0.142 2.0
## Benzene             0.80  0.43  0.04  0.25  0.13 0.91 0.089 1.8
## Toulene             0.88  0.25  0.03  0.19  0.12 0.89 0.107 1.3
## P_Xylene            0.92  0.30  0.04  0.09 -0.03 0.95 0.052 1.2
## NOx                 0.88  0.22  0.19  0.10 -0.06 0.86 0.135 1.3
## PM10                0.39  0.75  0.33 -0.02 -0.12 0.84 0.163 2.0
## WindDirection       0.20  0.15  0.21  0.00 -0.23 0.16 0.842 3.8
## NH3                 0.48  0.55  0.16  0.31  0.25 0.73 0.272 3.2
## RH                  0.33  0.03 -0.56  0.68  0.29 0.96 0.038 2.8
## Temp               -0.32 -0.47 -0.08 -0.61 -0.31 0.79 0.211 3.1
## WindSpeed          -0.70 -0.14 -0.10 -0.03 -0.27 0.59 0.408 1.4
## VerticalWindSpeed  0.00  0.03 -0.01 -0.57  0.12 0.34 0.656 1.1
## Solar              -0.37 -0.44  0.16 -0.42 -0.41 0.69 0.306 4.2
## BarPressure         0.18  0.25  0.44  0.51  0.04 0.56 0.439 2.8
## PD_PM2.5            0.36  0.84  0.15  0.14  0.09 0.89 0.108 1.5
## PD_PM10             0.24  0.87  0.24  0.00  0.02 0.88 0.124 1.3
## PD_NO2              0.36  0.42  0.43 -0.19  0.39 0.67 0.328 4.3
## PD_SO2              0.21  0.36  0.69  0.20  0.01 0.69 0.307 1.9
## PD_CO               0.45  0.42 -0.02  0.11  0.09 0.40 0.600 2.2
##
##                         PA1  PA3  PA2  PA4  PA5
## SS loadings            6.53 4.58 2.76 2.12 0.96
## Proportion Var         0.27 0.19 0.12 0.09 0.04
## Cumulative Var         0.27 0.46 0.58 0.67 0.71
## Proportion Explained  0.39 0.27 0.16 0.12 0.06
## Cumulative Proportion 0.39 0.66 0.82 0.94 1.00
##
## Mean item complexity =  2.2
## Test of the hypothesis that 5 factors are sufficient.
##
## The degrees of freedom for the null model are  276  and the objective func
tion was  28.43
## The degrees of freedom for the model are 166  and the objective function w
as  5.24
##
## The root mean square of the residuals (RMSR) is  0.03
## The df corrected root mean square of the residuals is  0.04
##
## Fit based upon off diagonal values = 1
## Measures of factor score adequacy
##                                                    PA1  PA3  PA2  PA4  PA5
## Correlation of (regression) scores with factors   0.98 0.96 0.94 0.93 0.89
## Multiple R square of scores with factors          0.97 0.92 0.88 0.87 0.79
## Minimum correlation of possible factor scores     0.94 0.84 0.76 0.73 0.58
```

```
# Solution 1 is giving better balanced Loadings.
#

# Draw the Factor Diagram
fa.diagram(solution1,simple=FALSE)
```




```
#==============================================================================
#
#
#                                 T H E - E N D
#
#==============================================================================
```