Compulsory HWA3, Multivariate Analysis, 2023HT

Shaobo Jin

1. (1pt) Determine the population principal components Z_1 and Z_2 for the covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix}.$$

Calculate also the proportion of the total population variance explained by the first principal component. Note: `princomp` and `prcomp` cannot be used for this task.
Solution: We only need to find the eigenvalues and eigenvectors of $\boldsymbol{\Sigma}$. Note that

$$\det(\lambda \boldsymbol{I} - \boldsymbol{\Sigma}) = \det \begin{bmatrix} \lambda - 5 & -2 \\ -2 & \lambda - 2 \end{bmatrix} = \lambda^2 - 7\lambda + 6$$

Hence, the eigenvalues are 6 and 1. When $\lambda = 6$, the eigenvector is $\left(2/\sqrt{5}, 1/\sqrt{5}\right)$. When $\lambda = 1$, the eigenvector is $\left(1/\sqrt{5}, -2/\sqrt{5}\right)$. Hence the population principal components are

$$\begin{aligned} Z_1 &= \frac{2}{\sqrt{5}} X_1 + \frac{1}{\sqrt{5}} X_2, \\ Z_2 &= \frac{1}{\sqrt{5}} X_1 - \frac{2}{\sqrt{5}} X_2. \end{aligned}$$

The proportion of the total population variation explained by the first component is $6/7$.

2. (1pt) In this task, we will focus on principal component regression. Suppose that our linear regression model is

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i, \quad i = 1, 2, ..., n,$$

where $\boldsymbol{x}_i$ is a $r \times 1$ vector. The OLS estimator $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y}$ requires $\left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}$, which can be problematic if the correlations among variables in $\boldsymbol{x}$ are large (known as multicollinearity). Multicollinearity makes $\left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}$ hard or impossible to compute numerically even though $\boldsymbol{X}^T \boldsymbol{X}$ is invertible. One alternative is to conduct PCA to $\boldsymbol{X}$ and regress y on the principal components instead. Let the matrix of principal component scores be $\boldsymbol{Z}$ ($n \times r$).

   (a) Suppose that we want to regress Y on the first principal component and the second principle components

$$y_i = \beta_1 z_1 + \beta_2 z_2 + e_i.$$

Let $\boldsymbol{Z}_1$ be the $n \times 2$ matrix of first two component scores. Show that the corresponding OLS estimator is $\boldsymbol{Z}_1^T \boldsymbol{y}$. Do you think multicollinearity in $\boldsymbol{x}$ will cause issues when you regress Y on the principal components?
Solution: The OLS estimator is

$$\left(\boldsymbol{Z}_1^T \boldsymbol{Z}_1\right)^{-1} \boldsymbol{Z}_1^T \boldsymbol{y} \quad = \quad \boldsymbol{Z}_1^T \boldsymbol{y},$$

since the principal components have variances 1 and are orthogonal to each other ($\boldsymbol{Z}_1^T \boldsymbol{Z}_1 = \boldsymbol{I}$). Since we only need to invert an identity matrix to get the OLS estimator, the multicollinearity issue no longer matters.

(b) Suppose that you decided to add more principal components into the model (using the first four principal components) as

$$y_i \quad = \quad \gamma_1 z_1 + \gamma_2 z_2 + \gamma_3 z_3 + \gamma_4 z_4 + e_i.$$

Let $\boldsymbol{Z}_2$ be the $n \times 4$ matrix of first four component scores. Show that the OLS estimator of γ_1 and γ_2 remain the same as the OLS estimator of β_1 and β_2, even though two more variables are added into the model.
Solution: The OLS estimator of γ's is

$$\left(\boldsymbol{Z}_2^T \boldsymbol{Z}_2\right)^{-1} \boldsymbol{Z}_2^T \boldsymbol{y} \quad = \quad \boldsymbol{Z}_2^T \boldsymbol{y} = \begin{bmatrix} \boldsymbol{Z}_1^T \boldsymbol{y} \\ \boldsymbol{Z}_3^T \boldsymbol{y} \end{bmatrix},$$

where

$$\boldsymbol{Z}_2 \quad = \quad \begin{bmatrix} \boldsymbol{Z}_1 & \boldsymbol{Z}_3 \end{bmatrix}.$$

Hence, the OLS estimator of γ_1 and γ_2 is the same as the OLS estimator of β_1 and β_2. They are simply $\boldsymbol{Z}_1^T \boldsymbol{y}$.

3. (1pt) Consider the orthogonal factor analysis model

$$\boldsymbol{X} \quad = \quad \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{F} + \boldsymbol{e}.$$

If you perform orthogonal factor rotation, are the communalities after rotation the same as the communalities before rotation?
Solution: Communalities are the diagonal elements of $\boldsymbol{L}\boldsymbol{L}^T$. If we perform orthogonal factor rotation with an orthogonal matrix $\boldsymbol{T}$, the factor loading matrix is $\boldsymbol{L}\boldsymbol{T}^T$. Then,

$$\left(\boldsymbol{L}\boldsymbol{T}^T\right)\left(\boldsymbol{L}\boldsymbol{T}^T\right)^T \quad = \quad \boldsymbol{L}\boldsymbol{L}^T.$$

Hence the communalities remain unchanged.

4. (1pt) Consider the Bartlett factor scores

$$\boldsymbol{f}_B \quad = \quad \left(\boldsymbol{L}^T \boldsymbol{\Psi}^{-1} \boldsymbol{L}\right)^{-1} \boldsymbol{L}^T \boldsymbol{\Psi}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}),$$

and the regression factor scores

$$\boldsymbol{f}_R = \left(\boldsymbol{L}^T\boldsymbol{\Psi}^{-1}\boldsymbol{L}+\boldsymbol{\Phi}^{-1}\right)^{-1}\boldsymbol{L}^T\boldsymbol{\Psi}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}),$$

where $\boldsymbol{\Phi}=\operatorname{cov}(\boldsymbol{F})$. Show that the Bartlett score is conditionally unbiased but the regression score is conditionally biased, that is, $\mathbb{E}(\boldsymbol{f}_B \mid \boldsymbol{F})=\boldsymbol{F}$ and $\mathbb{E}(\boldsymbol{f}_R \mid \boldsymbol{F})\neq\boldsymbol{F}$.
Solution: Note that $\mathbb{E}(\boldsymbol{x} \mid \boldsymbol{F})=\boldsymbol{\mu}+\boldsymbol{L}\boldsymbol{F}$. Hence,

$$\begin{aligned}\mathbb{E}(\boldsymbol{f}_B \mid \boldsymbol{F}) &= \left(\boldsymbol{L}^T\boldsymbol{\Psi}^{-1}\boldsymbol{L}\right)^{-1}\boldsymbol{L}^T\boldsymbol{\Psi}^{-1}\mathbb{E}(\boldsymbol{x}-\boldsymbol{\mu} \mid \boldsymbol{F})\\ &= \left(\boldsymbol{L}^T\boldsymbol{\Psi}^{-1}\boldsymbol{L}\right)^{-1}\boldsymbol{L}^T\boldsymbol{\Psi}^{-1}\boldsymbol{L}\boldsymbol{F}\\ &= \boldsymbol{F},\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}(\boldsymbol{f}_R \mid \boldsymbol{F}) &= \left(\boldsymbol{L}^T\boldsymbol{\Psi}^{-1}\boldsymbol{L}+\boldsymbol{\Phi}^{-1}\right)^{-1}\boldsymbol{L}^T\boldsymbol{\Psi}^{-1}\mathbb{E}(\boldsymbol{x}-\boldsymbol{\mu} \mid \boldsymbol{F})\\ &= \left(\boldsymbol{L}^T\boldsymbol{\Psi}^{-1}\boldsymbol{L}+\boldsymbol{\Phi}^{-1}\right)^{-1}\boldsymbol{L}^T\boldsymbol{\Psi}^{-1}\boldsymbol{L}\boldsymbol{F}\\ &\neq \boldsymbol{F}.\end{aligned}$$

5. (1pt) (Changing the signs in CCA when obtaining the eigenvectors) Let the first canonical variate pair is

$$\begin{aligned}U_1 &= \left(\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{e}_1\right)^T\boldsymbol{X}^{(1)},\\ V_1 &= \left(\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{f}_1\right)^T\boldsymbol{X}^{(2)}.\end{aligned}$$

Consider

$$\begin{aligned}U_1^* &= \left(-\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{e}_1\right)^T\boldsymbol{X}^{(1)},\\ V_1^* &= \left(-\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{f}_1\right)^T\boldsymbol{X}^{(2)}.\end{aligned}$$

Which of the following pair(s) still form(s) a canonical variate pair, (U_1,V_1^*), (U_1^*,V_1), and (U_1^*,V_1^*)?
Solution: U_1 and V_1 form the first canonical variate pair, then we must have

$$\operatorname{cor}(U_1,V_1) = \rho_1^* > 0.$$

It must be positive since $\rho_1^{*2}\geq\rho_2^{*2}\geq\cdots\geq\rho_p^{*2}\geq 0$. Hence, only (U_1^*,V_1^*) is still the first canonical variate pair. The other two will change the correlation from positive to negative. This means that we must be careful when choosing the sign of the eigenvector, if you want to keep the canonical correlation to be positive. However, changing the signs do not change the squared correlation.

6. (5p) Consider the data set HWA1.RData that we used in HWA1. You can download the data set from the homepage of HWA1. We will ignore the variable Station in this task.

   (a) (1pt) Perform PCA to the data set. Report the principal component loading matrix (you can take screenshot for this). Be clear on how you choose the number of principal components. State also how you will group the variables if you only take the first two components.
   Solution: We start with principal component in R.

```
pca <- prcomp(Data, scale = TRUE, center = TRUE)
```

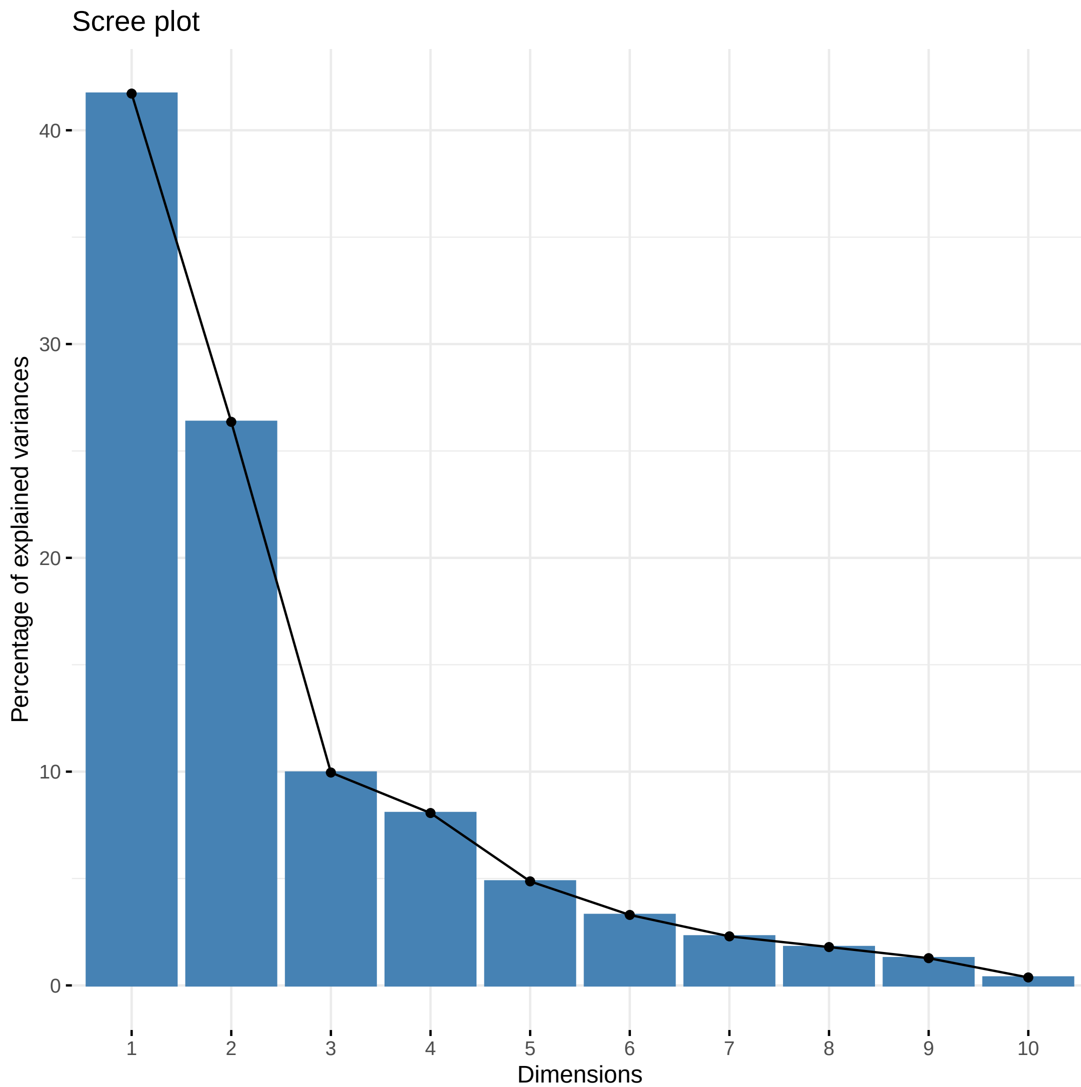
The The rotation matrix is

```
pca$rotation

##              PC1         PC2         PC3          PC4
## PM25  0.29919948 -0.39264148  0.07241718 -0.251325326
## SO2   0.35155897 -0.11050576  0.14095255 -0.422350771
## NO2   0.36413596 -0.31950300  0.04772319  0.035923104
## CO    0.37848464 -0.30453271 -0.00672722 -0.166700944
## O3   -0.36062696 -0.11415785  0.28170802 -0.470863874
## TEMP -0.39530390 -0.32778360  0.09673517 -0.002198435
## PRES  0.34083129  0.34589347 -0.08302296  0.154647836
## DEWP -0.30092609 -0.44905081 -0.03743775  0.128886416
## RAIN -0.08858854 -0.08991666 -0.92876341 -0.316912341
## WSPM -0.10076187  0.43468468  0.11380856 -0.605555132
##              PC5         PC6         PC7         PC8
## PM25  0.43025040 -0.18228935  0.17856772  0.14785699
## SO2  -0.73191099  0.15736012  0.12844583  0.29752581
## NO2   0.01891799  0.24536761 -0.78342434 -0.26918694
## CO    0.29943374 -0.07366741  0.29033698 -0.09144453
## O3   -0.10670449 -0.64217676 -0.22831954 -0.23980190
## TEMP -0.01998794  0.17193244 -0.20274248  0.27244713
## PRES  0.05850866 -0.50645175 -0.35312158  0.56081505
## DEWP  0.07887694  0.05621666 -0.03479938  0.56107598
## RAIN -0.07218211 -0.06036792 -0.08645890 -0.04745947
## WSPM  0.40315255  0.41367438 -0.16416655  0.21328015
##               PC9         PC10
## PM25 -0.633394681  0.120921093
## SO2  -0.020091000 -0.008214029
## NO2  -0.003550316 -0.122326578
## CO    0.735133176  0.089492123
## O3    0.088359127 -0.120934007
## TEMP  0.115688851  0.753826283
## PRES  0.119560200  0.141291632
## DEWP  0.126685696 -0.590199255
```

```
## RAIN -0.025566626  0.042384155
## WSPM  0.076061377 -0.096767687
```

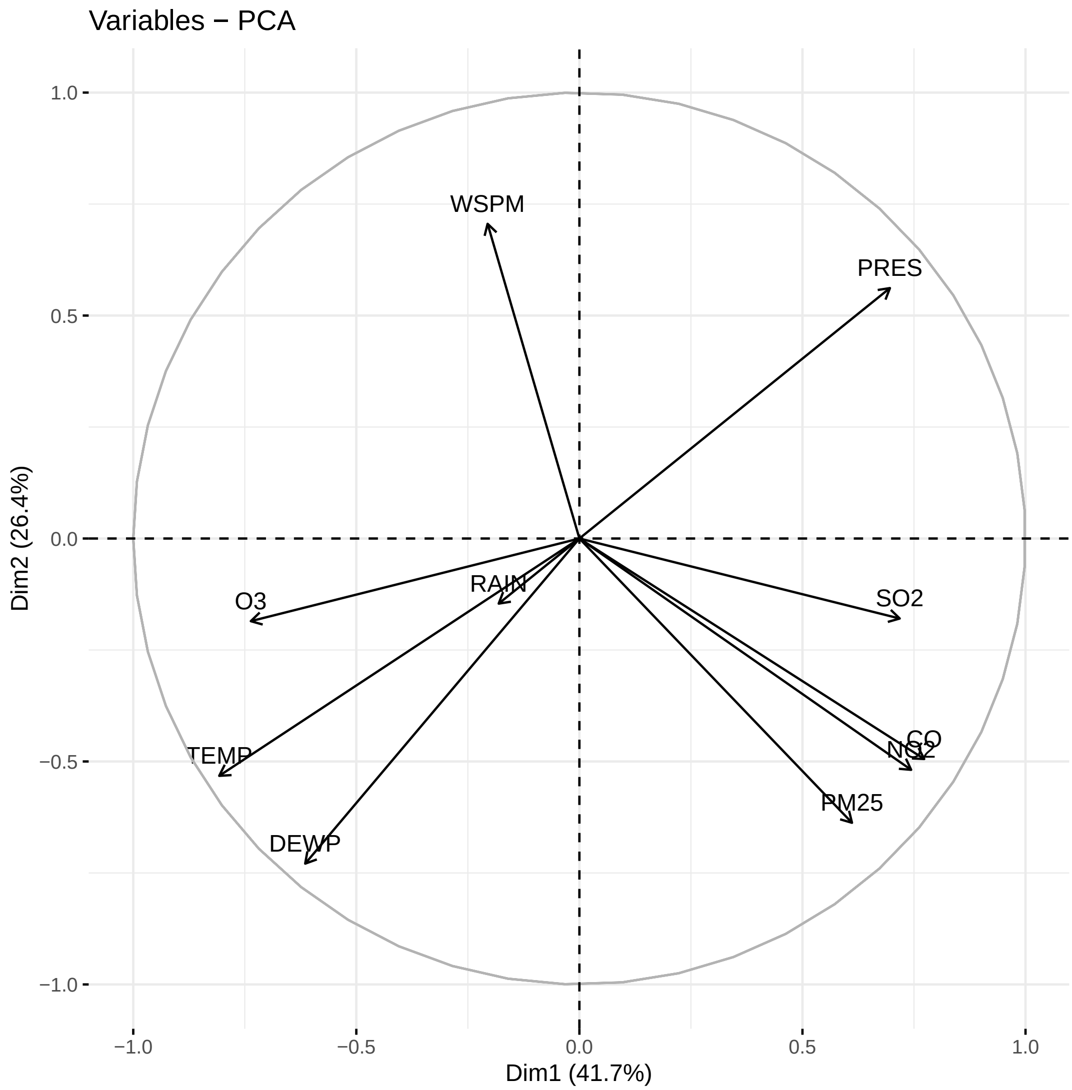
To decide the number of components, we use the scree plot

```
fviz_eig(pca)
```



The elbow occurs at 3 components. Hence, we set the number of components to be 3. However, if we want the proportion of variation be explained to be 95%, we need to choose 7 components. If we only have the first two components, the graph of variables is

```r
fviz_pca_var(pca)
```

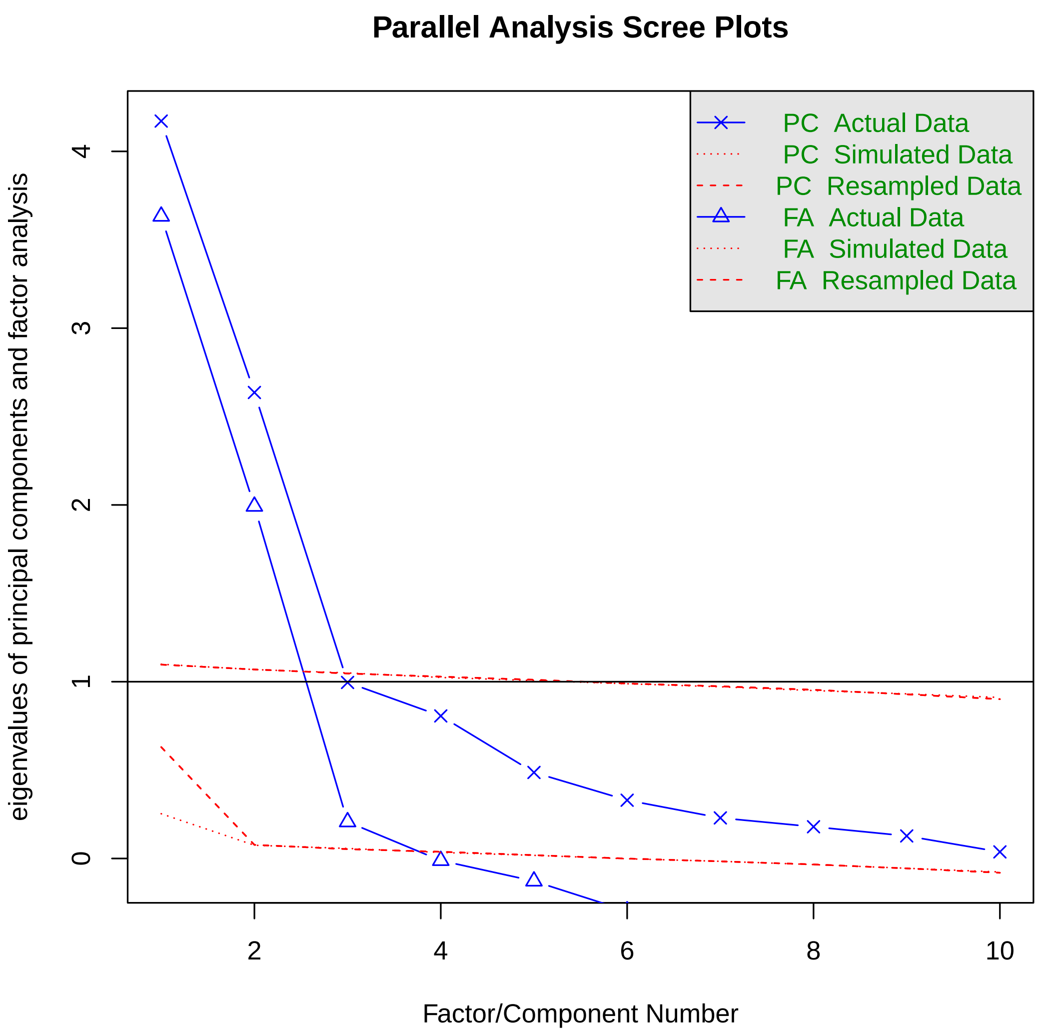


We perhaps can group them into four groups.

(b) (1pt) Perform factor analysis with maximum likelihood to the data set with both orthogonal rotation and oblique rotation. Report the factor loading matrices and the correlation matrix of factors (you can take screenshot for this task). Be clear on how you choose the number of factors and which rotation methods have been used. Please also interpret the factors.

Solution: We start the psych package in R. To determine the number of factors, we start with parallel analysis

```r
fa.parallel(Data)
```




```
## Parallel analysis suggests that the number of factors =  3  and the number o
```

The suggested number of factors is 3. Hence, we start with 3 factors and fit the model by ML. The rotation methods are varimax for orthogonal rotation and geomin for oblique rotation.

```
Varimax <- fa(Data, nfactors = 3, rotate = "varimax", fm = "ml")
Varimax

## Factor Analysis using method =  ml
## Call: fa(r = Data, nfactors = 3, rotate = "varimax", fm = "ml")
## Standardized loadings (pattern matrix) based upon correlation matrix
##         ML2   ML1   ML3   h2    u2 com
## PM25   0.91  0.01  0.06 0.83 0.171 1.0
## SO2    0.60 -0.26 -0.27 0.50 0.501 1.8
## NO2    0.81 -0.19  0.09 0.69 0.305 1.1
```

```
## CO    0.92 -0.21  0.02 0.89 0.114 1.1
## O3   -0.28  0.75 -0.14 0.66 0.337 1.3
## TEMP -0.21  0.90  0.35 0.97 0.031 1.4
## PRES  0.09 -0.84 -0.22 0.75 0.247 1.2
## DEWP  0.02  0.77  0.61 0.96 0.037 1.9
## RAIN -0.05  0.07  0.23 0.06 0.940 1.3
## WSPM -0.43 -0.05 -0.58 0.53 0.471 1.9
##
##                        ML2  ML1  ML3
## SS loadings           3.00 2.82 1.03
## Proportion Var        0.30 0.28 0.10
## Cumulative Var        0.30 0.58 0.68
## Proportion Explained  0.44 0.41 0.15
## Cumulative Proportion 0.44 0.85 1.00
##
## Mean item complexity =  1.4
## Test of the hypothesis that 3 factors are sufficient.
##
## df null model =  45  with the objective function =  8.19 with Chi Square =
## df of  the model are 18  and the objective function was  0.42
##
## The root mean square of the residuals (RMSR) is  0.03
## The df corrected root mean square of the residuals is  0.05
##
## The harmonic n.obs is  2418 with the empirical chi square  213.94  with prob
## The total n.obs was  2418  with Likelihood Chi Square =  1001.41  with prob
##
## Tucker Lewis Index of factoring reliability =  0.875
## RMSEA index =  0.15  and the 90 % confidence intervals are  0.143 0.158
## BIC =  861.18
## Fit based upon off diagonal values = 0.99
## Measures of factor score adequacy
##                                                    ML2  ML1
## Correlation of (regression) scores with factors   0.96 0.95
## Multiple R square of scores with factors          0.93 0.91
## Minimum correlation of possible factor scores     0.86 0.82
##                                                    ML3
## Correlation of (regression) scores with factors   0.87
## Multiple R square of scores with factors          0.76
## Minimum correlation of possible factor scores     0.52
```

```
Geomin <- fa(Data, nfactors = 3, rotate = "geominQ", fm = "ml")
```

```
## Loading required namespace:  GPArotation
```

```
Geomin
```

```
## Factor Analysis using method =  ml
## Call: fa(r = Data, nfactors = 3, rotate = "geominQ", fm = "ml")
## Standardized loadings (pattern matrix) based upon correlation matrix
##        ML1   ML2   ML3   h2    u2 com
## PM25  0.23  0.97  0.00 0.83 0.171 1.1
## SO2  -0.07  0.69  0.26 0.50 0.501 1.3
## NO2  -0.04  0.78 -0.11 0.69 0.305 1.0
## CO   -0.01  0.93 -0.03 0.89 0.114 1.0
## O3    0.85 -0.01  0.39 0.66 0.337 1.4
## TEMP  0.90 -0.10 -0.11 0.97 0.031 1.1
## PRES -0.90 -0.07 -0.03 0.75 0.247 1.0
## DEWP  0.74  0.01 -0.43 0.96 0.037 1.6
## RAIN  0.00 -0.13 -0.25 0.06 0.940 1.5
## WSPM -0.01 -0.24  0.62 0.53 0.471 1.3
##
##                        ML1  ML2  ML3
## SS loadings           2.96 2.96 0.92
## Proportion Var        0.30 0.30 0.09
## Cumulative Var        0.30 0.59 0.68
## Proportion Explained  0.43 0.43 0.13
## Cumulative Proportion 0.43 0.87 1.00
##
##  With factor correlations of
##       ML1   ML2   ML3
## ML1  1.00 -0.37 -0.35
## ML2 -0.37  1.00 -0.28
## ML3 -0.35 -0.28  1.00
##
## Mean item complexity =  1.2
## Test of the hypothesis that 3 factors are sufficient.
##
## df null model =  45  with the objective function =  8.19 with Chi Square =
## df of  the model are 18  and the objective function was  0.42
##
## The root mean square of the residuals (RMSR) is  0.03
## The df corrected root mean square of the residuals is  0.05
##
## The harmonic n.obs is  2418 with the empirical chi square  213.94  with prob
## The total n.obs was  2418  with Likelihood Chi Square =  1001.41  with prob
##
## Tucker Lewis Index of factoring reliability =  0.875
## RMSEA index =  0.15  and the 90 % confidence intervals are  0.143 0.158
## BIC =  861.18
## Fit based upon off diagonal values = 0.99
```

```
## Measures of factor score adequacy
##                                                    ML1  ML2
## Correlation of (regression) scores with factors   0.99 0.97
## Multiple R square of scores with factors          0.97 0.94
## Minimum correlation of possible factor scores     0.94 0.89
##                                                    ML3
## Correlation of (regression) scores with factors   0.92
## Multiple R square of scores with factors          0.84
## Minimum correlation of possible factor scores     0.68
```

It is seen that the likelihood ratio tests suggest that 3 factors are enough for our data set. In orthogonal factor analysis, the covariance matrix of the factors is the identity matrix. For oblique factor analysis with geomin rotation, the estimated factor correlation matrix is

```
Geomin$Phi

##             ML1        ML2        ML3
## ML1  1.0000000 -0.3687723 -0.3450983
## ML2 -0.3687723  1.0000000 -0.2788809
## ML3 -0.3450983 -0.2788809  1.0000000
```

In order to interpret the factors, we need to look at the factor loading matrices. We start with the orthogonal rotation. The loading matrix is

```
Varimax$loadings

##
## Loadings:
##      ML2    ML1    ML3
## PM25  0.908
## SO2   0.601 -0.260 -0.265
## NO2   0.807 -0.190
## CO    0.918 -0.209
## O3   -0.280  0.752 -0.136
## TEMP -0.208  0.896  0.350
## PRES        -0.836 -0.217
## DEWP         0.770  0.608
## RAIN                0.230
## WSPM -0.435        -0.581
##
##                  ML2   ML1   ML3
## SS loadings    3.001 2.815 1.029
## Proportion Var 0.300 0.282 0.103
## Cumulative Var 0.300 0.582 0.685
```

The first factor is about the chemical compounds in the air. The second factor is weather related values and the chemical compound O3. The third factor is less clear. The loading matrix of the geomin rotation is

```
Geomin$loadings

##
## Loadings:
##       ML1    ML2    ML3
## PM25  0.234  0.970
## SO2          0.690  0.260
## NO2          0.785 -0.109
## CO           0.928
## O3    0.853         0.394
## TEMP  0.904        -0.111
## PRES -0.903
## DEWP  0.743        -0.435
## RAIN        -0.131 -0.245
## WSPM        -0.242  0.620
##
##                  ML1   ML2   ML3
## SS loadings    2.976 2.986 0.882
## Proportion Var 0.298 0.299 0.088
## Cumulative Var 0.298 0.596 0.684
```

The second factor is about the chemical compounds in the air. The frist factor is weather related values and the chemical compound O3. The third factor is less clear still, maybe can be viewed as a general factor.

(c) (1pt) If we use factor analysis estimated by principal component, the diagonal elements in $\hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T + \hat{\boldsymbol{\Psi}} - \boldsymbol{S}$ are all zero. Use R or other software to show that the diagonal elements are still zero if you apply maximum likelihood to the covariance matrix. If you use the R function `fa` and store the result in the object `fit`, then `fit$uniquenesses` extracts the uniqueness.

Solution: We apply ML to the covariance matrix as

```
fit <- fa(Data, nfactors = 3, rotate = "varimax", fm = "ml", covar = TRUE)
```

To extract the unique variance from the orthogonal factor analysis model, we use

```
fit$uniquenesses

##         PM25          SO2          NO2           CO           O3          TEM
## 1.044115e+03 1.500818e+02 2.183535e+02 1.611077e+05 5.745298e+02 4.011541e+0
##         PRES         DEWP         RAIN         WSPM
## 2.458521e+01 4.999994e-03 9.666088e-02 3.151171e-01
```

The diagonal elements of the estimated covariance matrix is

```
diag(fit$loadings %*% t(fit$loadings) + fit$uniquenesses)

##         PM25          SO2          NO2            CO           O3          TEM
## 5.108373e+03 3.179802e+02 8.004398e+02 1.006660e+06 1.483611e+03 1.159991e+0
##         PRES         DEWP         RAIN         WSPM
## 1.025880e+02 1.811638e+02 1.045007e-01 5.994389e-01
```

The diagonal elements of the sample covariance matrix is

```
diag(cov(Data))

##         PM25          SO2          NO2            CO           O3          TEM
## 5.108373e+03 3.179802e+02 8.004398e+02 1.006660e+06 1.483611e+03 1.159991e+0
##         PRES         DEWP         RAIN         WSPM
## 1.025880e+02 1.811638e+02 1.045007e-01 5.994389e-01
```

They are the same as the diagonal elements of the estimated covariance matrix.

(d) (1pt) Let PM25, SO2, NO2, CO, and O3 be one set of variables, and TEMP, PRES, DEWP, RAIN, and WSPM be another set of variables. Perform CCA. Report the coefficients of linear combinations for the first canonical variate pair. Report also the corresponding canonical correlation.
Solution: We apply CCA in R as

```
ccpack <- cc(X = Data[, 1 : 5], Y = Data[, 6 : 10]) # Package uses cov()
```

The first canonical correlation coefficient is

```
ccpack$cor[1]

## [1] 0.8185748
```

The coefficients for the linear combinations are

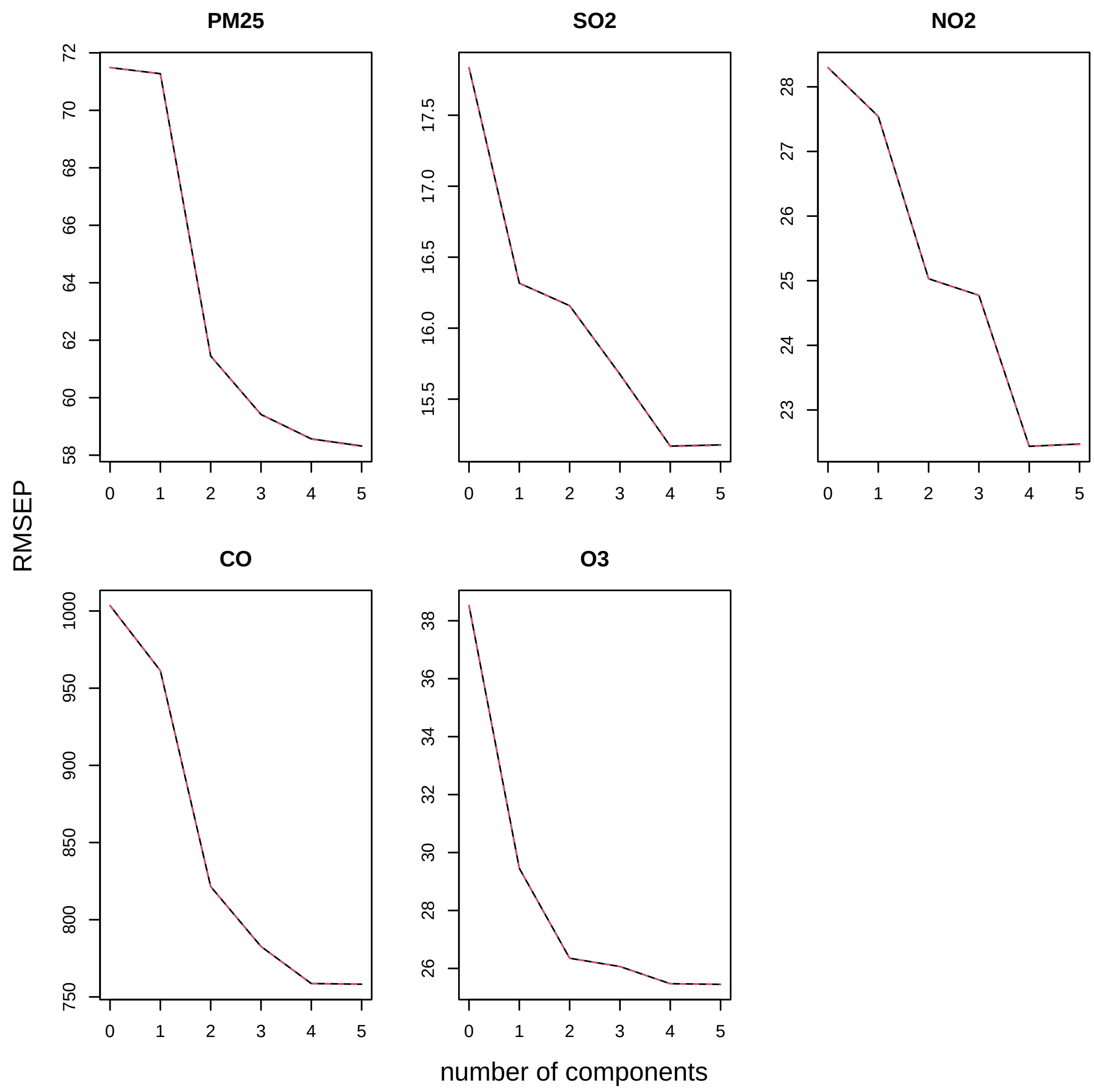
```
ccpack$xcoef

##               [,1]         [,2]          [,3]          [,4]
## PM25 -0.0004157534  0.012040370  0.0110873240  0.0001718376
## SO2  -0.0132318600 -0.040634021 -0.0029806976  0.0516578251
## NO2   0.0063022801  0.010164411 -0.0591650002 -0.0152666426
## CO   -0.0003740553  0.000139664  0.0004185406  0.0004138483
## O3    0.0195229936  0.007761541 -0.0102963866  0.0169331148
##               [,5]
## PM25 -0.0227359322
## SO2  -0.0263138184
## NO2   0.0003316964
## CO    0.0020698342
## O3    0.0114290161
```

```
ccpack$ycoef
```

```
##              [,1]        [,2]        [,3]        [,4]
## TEMP  0.14252314 -0.13393450 -0.18777256 -0.09277520
## PRES  0.01072161 -0.02438936 -0.04022331 -0.17910813
## DEWP -0.04363418  0.14105466  0.14077905 -0.04343464
## RAIN  0.03450408 -0.60369021  0.43385005 -0.77496655
## WSPM  0.37578242 -0.11649950  1.50094422 -0.19096981
##              [,5]
## TEMP  0.03438711
## PRES -0.02757160
## DEWP -0.05818116
## RAIN  3.04643223
## WSPM -0.52721649
```

(e) (1pt) Let PM25, SO2, NO2, CO, and O3 be one set of variables, and TEMP, PRES, DEWP, RAIN, and WSPM be another set of variables. Perform PLS regression that regress the first set on the second set. Report the estimated regression coefficients (you can take screenshots for this task).
Solution: We apply PLS in R as

```
PLS <- plsr(cbind(PM25, SO2, NO2, CO, O3) ~ TEMP + PRES + DEWP + RAIN + WSPM,
            ncomp = 5, data = Data, validation = "CV")
plot(RMSEP(PLS))
```



We consider 4 components. The estimated coefficients are

```
coef(PLS, ncomp = 1 : 4, intercept = TRUE)

## , , 1 comps
##
##                      PM25          SO2          NO2
## (Intercept) -8.892021e+01 -1.868974e+02 -1.330855e+02
## TEMP        -2.835203e-01 -3.386565e-01 -3.079481e-01
## PRES         1.718703e-01  2.052939e-01  1.866784e-01
## DEWP        -1.177396e-01 -1.406364e-01 -1.278839e-01
## RAIN        -8.026516e-04 -9.587431e-04 -8.718070e-04
## WSPM        -2.102899e-02 -2.511850e-02 -2.284082e-02
##                        CO           O3
## (Intercept) -6.900207e+03 752.282616584
## TEMP        -1.360655e+01   1.162831572
```

```
## PRES         8.248304e+00  -0.704909691
## DEWP        -5.650494e+00   0.482897809
## RAIN        -3.852041e-02   0.003291999
## WSPM        -1.009212e+00   0.086248414
##
## , , 2 comps
##
##                      PM25           SO2           NO2
## (Intercept) -8.627059e+02 -2.370468e+02 -379.36902603
## TEMP        -6.594072e+00 -7.476459e-01   -2.31649511
## PRES         1.009406e+00  2.595749e-01    0.45325255
## DEWP         5.795140e+00  2.425798e-01    1.75409066
## RAIN         6.814219e-03 -4.650905e-04    0.00155252
## WSPM        -1.491749e+00 -1.204365e-01   -0.49094731
##                        CO            O3
## (Intercept) -1.758708e+04  1.033596e+03
## TEMP        -1.007626e+02  3.457060e+00
## PRES         1.981564e+01 -1.009399e+00
## DEWP         7.601321e+01 -1.666756e+00
## RAIN         6.667738e-02  5.228522e-04
## WSPM        -2.132155e+01  6.209351e-01
##
## , , 3 comps
##
##                      PM25          SO2          NO2
## (Intercept) 2085.13349463 501.15254214 297.41226307
## TEMP          -9.32306522  -1.43104184  -2.94303260
## PRES          -1.86958090  -0.46138234  -0.20772118
## DEWP           5.95847050   0.28348111   1.79158900
## RAIN          -0.06709666  -0.01897389  -0.01541635
## WSPM          -2.20999359  -0.30029967  -0.65584593
##                       CO           O3
## (Intercept) 29237.452146 300.52071709
## TEMP         -144.110872   4.13571246
## PRES          -25.915217  -0.29344676
## DEWP           78.607613  -1.70737310
## RAIN           -1.107349   0.01890317
## WSPM          -32.730410   0.79954978
##
## , , 4 comps
##
##                    PM25         SO2         NO2          CO
## (Intercept) 2066.874309 494.0888690 278.7818455 28892.76067
## TEMP          -8.047968  -0.9377630  -1.6420111  -120.03997
```

```
## PRES          -1.837371  -0.4489219  -0.1748568   -25.30718
## DEWP           4.697739  -0.2042403   0.5052252    54.80790
## RAIN          -4.314030  -1.6619251  -4.3486954   -81.27970
## WSPM         -18.589102  -6.6366538 -17.3679618  -341.93034
##                      O3
## (Intercept) 310.3508722
## TEMP          3.4492414
## PRES         -0.3107873
## DEWP         -1.0286360
## RAIN          2.3053149
## WSPM          9.6175320
```