Assignment2 MAST90138

Muhan Guan 1407870

2023-10-04

Tutor's name:Shangyu Chen

Time of Tutorial class: Thursday 10AM

Problem 1

(a)

Because the property of matrix: $(-Q)^T = -(Q^T)$,

Then we have $\Sigma = (-Q_1)(-Q_1)^T + \Psi = (-1)(-1)Q_1Q_1^T + \Psi = Q_1Q_1^T + \Psi$

Therefore, if the equation is satisfied for $Q = Q_1$, it is also satisfied for $Q = -Q_1$ with the same Ψ.

(b)

(i)

G should also be orthogonal,hence $GG^T = 1$. Additionally , if q=1,then the dimension of G is 1*1,which means $G^2 = 1$.

So the only two 1*1 orthogonal matrices $G_1 = [1], G_2 = [-1]$

(ii)

when $q = 1, then\ G_1 = 1, G_2 = -1$, let $Q_G = QG1,\ -Q_G = G_2Q$

$$\Sigma = (Q_G)(Q_G)^T + \Psi = (QG_1)(QG_1)^T + \Psi = Q \times (1) \times (1) \times Q^T + \Psi = QQ^T + \Psi$$

$$\Sigma = (-Q_G)(-Q_G)^T + \Psi = (G_2Q)(G_2Q)^T + \Psi = (-1) \times Q \times Q^T \times (-1) + \Psi = QQ^T + \Psi$$

So the equation in (a) still can be satisfied, when $q = 1\ and\ Q_G = QG$

(c)

If the below condition can be satisfied,

$$p(p+1)/2 > pq + p - q(q-1)/2$$

we can find the unique factor loadings and specific variance.

When $p = 3$ and $q = 1$: $p(p+1)/2 = 6, pq + p - q(q-1)/2 = 6$, then $6 \geq 6$ the condition is satisfied ,hence there is an unique solution.

By equation(1) we have:

$$\begin{bmatrix} 4 & -2 & 3 \\ -2 & 5 & -1 \\ 3 & -1 & 6 \end{bmatrix} = \begin{bmatrix} q_1 \\ q_2 \\ q_3 \end{bmatrix} \begin{bmatrix} q_1 & q_2 & q_3 \end{bmatrix} + \begin{bmatrix} \psi_{11} & 0 & 0 \\ 0 & \psi_{22} & 0 \\ 0 & 0 & \psi_{33} \end{bmatrix}$$

Then we can get the following six equations:

$$q_1 q_2 = -2$$

$$q_1 q_3 = 3$$

$$q_2 q_3 = -1$$

$$q_1^2 + \psi_{11} = 4$$

$$q_2^2 + \psi_{22} = 5$$

$$q_3^2 + \psi_{33} = 6$$

Solve them we can get :

$$Q_1 = \begin{bmatrix} -\sqrt{6} \\ \sqrt{\frac{2}{3}} \\ -\sqrt{\frac{3}{2}} \end{bmatrix}, \ Q_2 = \begin{bmatrix} \sqrt{6} \\ -\sqrt{\frac{2}{3}} \\ \sqrt{\frac{3}{2}} \end{bmatrix}, \ \Psi = \begin{bmatrix} -2 & 0 & 0 \\ 0 & \frac{13}{3} & 0 \\ 0 & 0 & \frac{9}{2} \end{bmatrix}$$

This unique Ψ is not interpretable. The $\Psi = Var(U)$,where U is specific factors.The variance makes statistical sense if and only if all of its elements are positive,but here the ψ_1 is negative, hence the Ψ is not interpretable .

Problem 2

(a)

```r
library(mlbench)
data(BostonHousing)
summary(BostonHousing)
```

```
##       crim                zn             indus       chas        nox
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   0:471   Min.   :0.3850
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1: 35   1st Qu.:0.4490
##  Median : 0.25651   Median :  0.00   Median : 9.69           Median :0.5380
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14           Mean   :0.5547
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10           3rd Qu.:0.6240
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74           Max.   :0.8710
##        rm             age              dis              rad
##  Min.   :3.561   Min.   :  2.90   Min.   : 1.130   Min.   : 1.000
##  1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100   1st Qu.: 4.000
##  Median :6.208   Median : 77.50   Median : 3.207   Median : 5.000
##  Mean   :6.285   Mean   : 68.57   Mean   : 3.795   Mean   : 9.549
##  3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188   3rd Qu.:24.000
##  Max.   :8.780   Max.   :100.00   Max.   :12.127   Max.   :24.000
##       tax           ptratio            b              lstat
##  Min.   :187.0   Min.   :12.60   Min.   :  0.32   Min.   : 1.73
##  1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38   1st Qu.: 6.95
##  Median :330.0   Median :19.05   Median :391.44   Median :11.36
##  Mean   :408.2   Mean   :18.46   Mean   :356.67   Mean   :12.65
##  3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23   3rd Qu.:16.95
##  Max.   :711.0   Max.   :22.00   Max.   :396.90   Max.   :37.97
##       medv
##  Min.   : 5.00
##  1st Qu.:17.02
##  Median :21.20
##  Mean   :22.53
##  3rd Qu.:25.00
##  Max.   :50.00
```

```r
#remove the binary variable
BostonHousing <- subset(BostonHousing, select = -chas)
```

```r
# transform each variable according to the requirement.
XBoston <- data.frame(
  X1 = log(BostonHousing$crim),
  X2 = BostonHousing$zn / 10,
  X3 = log(BostonHousing$indus),
  X5 = log(BostonHousing$nox),
  X6 = log(BostonHousing$rm),
  X7 = (BostonHousing$age^2.5) / 10000,
  X8 = log(BostonHousing$dis),
  X9 = log(BostonHousing$rad),
  X10 = log(BostonHousing$tax),
  X11 = exp(0.4 * BostonHousing$ptratio) / 1000,
  X12 = BostonHousing$b / 100,
  X13 = sqrt(BostonHousing$lstat),
  X14 = log(BostonHousing$medv)
)
```

```r
#  check the transformed data
head(XBoston)
```

```
##          X1  X2        X3         X5       X6       X7       X8        X9
## 1 -5.064036 1.8 0.8372475 -0.6198967 1.883275 3.432567 1.408545 0.0000000
## 2 -3.600502 0.0 1.9558605 -0.7571525 1.859574 5.529585 1.602836 0.6931472
## 3 -3.601235 0.0 1.9558605 -0.7571525 1.971996 2.918119 1.602836 0.6931472
## 4 -3.430523 0.0 0.7793249 -0.7808861 1.945624 1.419592 1.802073 1.0986123
## 5 -2.672924 0.0 0.7793249 -0.7808861 1.966693 2.162710 1.802073 1.0986123
## 6 -3.511570 0.0 0.7793249 -0.7808861 1.860975 2.639947 1.802073 1.0986123
##        X10       X11    X12      X13      X14
## 1 5.690359 0.4548647 3.9690 2.231591 3.178054
## 2 5.488938 1.2364504 3.9690 3.023243 3.072693
## 3 5.488938 1.2364504 3.9283 2.007486 3.546740
## 4 5.402677 1.7722408 3.9463 1.714643 3.508556
## 5 5.402677 1.7722408 3.9690 2.308679 3.589059
## 6 5.402677 1.7722408 3.9412 2.282542 3.356897
```

```r
dim(XBoston)
```

```
## [1] 506  13
```

(b)

```r
# scale and cneter the data
scaled_XBoston = scale(XBoston)

# apply factor analysis
fit_XBoston <- factanal(scaled_XBoston, factors = 3,rotation = "varimax")
print(fit_XBoston)
```

```
## 
## Call:
## factanal(x = scaled_XBoston, factors = 3, rotation = "varimax")
## 
## Uniquenesses:
##    X1    X2    X3    X5    X6    X7    X8    X9   X10   X11   X12   X13   X14 
## 0.096 0.575 0.309 0.144 0.519 0.259 0.118 0.109 0.213 0.686 0.752 0.166 0.137 
## 
## Loadings:
##     Factor1 Factor2 Factor3
## X1   0.552   0.725   0.270 
## X2  -0.586  -0.159  -0.238 
## X3   0.629   0.411   0.357 
## X5   0.790   0.414   0.247 
## X6  -0.164          -0.669 
## X7   0.769   0.252   0.293 
## X8  -0.871  -0.316  -0.152 
## X9   0.274   0.893   0.135 
## X10  0.348   0.767   0.277 
```

```
## X11  0.180   0.340   0.406
## X12 -0.181  -0.392  -0.248
## X13  0.407   0.259   0.775
## X14 -0.211  -0.304  -0.852
##
##                 Factor1 Factor2 Factor3
## SS loadings       3.515   2.876   2.523
## Proportion Var    0.270   0.221   0.194
## Cumulative Var    0.270   0.492   0.686
##
## Test of the hypothesis that 3 factors are sufficient.
## The chi square statistic is 306.8 on 42 degrees of freedom.
## The p-value is 5.87e-42
```

```r
#check if the method used is MLE or not
print(fit_XBoston$method)
```

```
## [1] "mle"
```

(c)

```r
#loading matrix
Q=loadings(fit_XBoston)
print(Q, digit=4)# loadings
```

```
##
## Loadings:
##     Factor1 Factor2 Factor3
## X1   0.5525  0.7247  0.2705
## X2  -0.5858 -0.1587 -0.2377
## X3   0.6287  0.4105  0.3566
## X5   0.7898  0.4141  0.2468
## X6  -0.1644         -0.6691
## X7   0.7688  0.2518  0.2934
## X8  -0.8709 -0.3164 -0.1515
## X9   0.2736  0.8932  0.1347
## X10  0.3480  0.7673  0.2772
## X11  0.1800  0.3405  0.4065
## X12 -0.1813 -0.3917 -0.2483
## X13  0.4072  0.2587  0.7752
## X14 -0.2111 -0.3043 -0.8520
##
##                 Factor1 Factor2 Factor3
## SS loadings      3.5155  2.8757  2.5232
## Proportion Var   0.2704  0.2212  0.1941
## Cumulative Var   0.2704  0.4916  0.6857
```

```r
Psi=diag(fit_XBoston$uniquenesses)
round(Psi, 4)#specific variance
```

```
##         [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]   [,8]   [,9]  [,10]
##  [1,] 0.0964 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
##  [2,] 0.0000 0.5752 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
##  [3,] 0.0000 0.0000 0.3091 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
##  [4,] 0.0000 0.0000 0.0000 0.1439 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
##  [5,] 0.0000 0.0000 0.0000 0.0000 0.5188 0.0000 0.0000 0.0000 0.0000 0.0000
##  [6,] 0.0000 0.0000 0.0000 0.0000 0.0000 0.2594 0.0000 0.0000 0.0000 0.0000
##  [7,] 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.1185 0.0000 0.0000 0.0000
##  [8,] 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.1092 0.0000 0.0000
##  [9,] 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.2133 0.0000
## [10,] 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.6865
## [11,] 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
## [12,] 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
## [13,] 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
##       [,11]  [,12] [,13]
##  [1,] 0.000 0.0000 0.000
##  [2,] 0.000 0.0000 0.000
##  [3,] 0.000 0.0000 0.000
##  [4,] 0.000 0.0000 0.000
##  [5,] 0.000 0.0000 0.000
##  [6,] 0.000 0.0000 0.000
##  [7,] 0.000 0.0000 0.000
##  [8,] 0.000 0.0000 0.000
##  [9,] 0.000 0.0000 0.000
## [10,] 0.000 0.0000 0.000
## [11,] 0.752 0.0000 0.000
## [12,] 0.000 0.1663 0.000
## [13,] 0.000 0.0000 0.137
```

(d)

Then Communality of j th variable in the fitted model can be expressed as :

Option1:

$$\text{Communality} = \sum_{\ell=1}^{q} q_{Y,j\ell}^2$$

Option2:

$$\text{Communality} = var(Y_j) - \psi_{Y,j} = 1 - Uniqueness_i$$

```r
library(MASS)
# option1:
communality_1 = round(rowSums(Q^2),4)
# option2:
communality_2 = round((1 - fit_XBoston$uniquenesses), 4)
# print
cat("Communalities by option1:\n", communality_1, "\n")
```

```
## Communalities by option1:
##  0.9036 0.4248 0.6909 0.8561 0.4812 0.7406 0.8815 0.8908 0.7867 0.3135 0.248 0.8337 0.863
```

```r
cat("Communalities by option2:\n", communality_2, "\n")
```

```
## Communalities by option2:
##  0.9036 0.4248 0.6909 0.8561 0.4812 0.7406 0.8815 0.8908 0.7867 0.3135 0.248 0.8337 0.863
```

```r
difference=fractions(communality_1)-fractions(communality_2)#verify the consistency of results
cat("Difference between two options:", difference, "\n")
```

```
## Difference between two options: 0 0 0 0 0 0 0 0 0 0 0 0 0
```

(e)

Comparatively, there exists a distinction in the loadings of the two fitted models. Specifically, the signs of the loadings are inverted, and the order of the columns varies between the two models.

However, no discrepancies are observed regarding the communalities and specific variances of these two fitted models.

This underscores the notion that alterations in the sign of the loadings or permutations of the column order do not effect the interpretability of the model. This is because, for the loadings $Q_G = QG$, as long as G satisfies the constraints $||G_j|| = 1$ and $G_j^T G_k = 0$ for $j \neq k$, the solution of the model still remains valid.

Problem 3

(a)

```r
data=read.table(file='/Users/guanmuhan/Downloads/WheatAssignment2 .txt')
X=as.matrix(data[,1:7])
# scale and cneter the data
X=scale(X)
# apply factor analysis
fit_X <- factanal(X, factors = 3,rotation = "varimax")
fit_X
```

```
##
## Call:
## factanal(x = X, factors = 3, rotation = "varimax")
##
## Uniquenesses:
##    V1    V2    V3    V4    V5    V6    V7
## 0.005 0.005 0.052 0.016 0.005 0.005 0.089
##
## Loadings:
##    Factor1 Factor2 Factor3
## V1  0.892   0.435  -0.109
## V2  0.929   0.349  -0.109
## V3  0.201   0.937  -0.173
## V4  0.974   0.163
## V5  0.768   0.626  -0.115
```

```
## V6         -0.159   0.983
## V7  0.952
##
##                 Factor1 Factor2 Factor3
## SS loadings       4.149   1.635   1.044
## Proportion Var    0.593   0.234   0.149
## Cumulative Var    0.593   0.826   0.975
##
## Test of the hypothesis that 3 factors are sufficient.
## The chi square statistic is 226.16 on 3 degrees of freedom.
## The p-value is 9.34e-49
```

(b)

The parameter `uniquenesses` in this output represents how many variance of each of seven scaled variables has not explained by three factors.

```
cat('The percentage of variance of each variable explained by 3 factors:' ,"\n")
```

```
## The percentage of variance of each variable explained by 3 factors:
```

```
round((1 - fit_X$uniquenesses) * 100, 4)
```

```
##      V1      V2      V3      V4      V5      V6      V7
## 99.5000 99.5000 94.8124 98.4168 99.5000 99.5000 91.0659
```

(c)

$$corr(X,F) = D^{-\frac{1}{2}}Q = corr(Y,F) = Q_Y$$

Therefore,the correlation between each original variables X_i and each factors F_i can be calculated by the correlation between the scaled data and their factors.

```
cat('The correlation between Factor1 and each of seven variables:' ,"\n")
```

```
## The correlation between Factor1 and each of seven variables:
```

```
fit_X$loadings[,1]
```

```
##         V1         V2         V3         V4         V5         V6         V7
##  0.8920499  0.9292023  0.2014453  0.9744354  0.7681174 -0.0592640  0.9519551
```

```
cat("\n")
```

```
cat('The correlation between Factor2 and each of seven variables:' ,"\n")
```

```
## The correlation between Factor2 and each of seven variables:
```

```r
fit_X$loadings[,2]
```

```
##          V1          V2          V3          V4          V5          V6
##  0.43542338  0.34870896  0.93677119  0.16328811  0.62626585 -0.15928709
##          V7
##  0.04079483
```

```r
cat("\n")
```

```r
cat('The correlation between Factor3 and each of seven variables:' ,"\n")
```

```
## The correlation between Factor3 and each of seven variables:
```

```r
fit_X$loadings[,3]
```

```
##          V1          V2          V3          V4          V5          V6
## -0.10921593 -0.10856629 -0.17321364 -0.08933156 -0.11473627  0.98291212
##          V7
##  0.05270147
```

`Factor1` is most correlated with **area** ,**perimeter** ,**length of kernel** and **length of kernel groove**,hence we can probably rename the Factor1 as **Size and Shape Factor.**

`Factor2` is most correlated with **compactness. (Compactness Factor)**

`Factor3` is most correlated with **asymmetry coefficient. (Asymmetry Factor)**