HUDM6122 Homework_05

Chenguang Pan

2023-03-20

0.1 Github Address

All my latest homework can be found on Github: https://github.com/cgpan/hudm6122_homeworks . Thanks for checking if interested.

0.2 Ex 5.1

Show how the result rises from the assumptions of uncorrelated factors, independence of the specific variates, and independence of common factors and specific variances. What form does take if the factors are allowed to be correlated?

MY SOLUTION:
Based on the assumption of Exploratory Factor Analysis(EFA), a set of observed variables **x** assumed to be linked to a set of latent variables **f**. Therefore, we can have a regression model in matrix form

$$\boldsymbol{x} = \boldsymbol{\Lambda f} + \boldsymbol{u}$$

, where $\boldsymbol{\Lambda}$ is a $q \times k$ matrix of factor loadings (a.k.a., the coefficients of the regression model), and the u is the vector of unexplained error of each observed variables.

Let's take the variance of the formula above

$$V(\boldsymbol{x}) = V(\boldsymbol{\Lambda f} + \boldsymbol{u})$$

. Based on the operation rule of variance, like

$$V(a+b) = V(a) + V(b) + 2Cov(ab)$$

, we combined the two formulas above, then

$$V(\boldsymbol{x}) = V(\boldsymbol{\Lambda f} + \boldsymbol{u}) = V(\boldsymbol{\Lambda f}) + V(\boldsymbol{u}) + 2Cov(\boldsymbol{\Lambda f u})$$

. Since the we assumed that the error terms are uncorrelated with the factors, therefore the $Cov(\boldsymbol{\Lambda f u}) = 0$. Then, we can continue to drive the variance formula as

$$V(\boldsymbol{x}) = V(\boldsymbol{\Lambda f}) + V(\boldsymbol{u}) = \boldsymbol{\Lambda} V(\boldsymbol{f}) \boldsymbol{\Lambda}^T + \Psi$$

. In addition, we assumed that the factors are uncorrelated with each other. The $V(\boldsymbol{f})$ is actually an identity matrix. Therefore, the formula can be written as

$$V(\boldsymbol{x}) = \boldsymbol{\Lambda} V(\boldsymbol{f}) \boldsymbol{\Lambda}^T + \Psi = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \Psi$$

. Finally, the formula can be written as

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \Psi$$

.

If we allow the factors to be correlated with each other, then the $V(\boldsymbol{f})$ is not an identity matrix. Let's use the greek letter Φ to represent the variance matrix of loadings $\mathbf{f}$. Thus, the formula should be

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \Psi$$

.

0.3 Ex 5.2

Show that the communalities in a factor analysis model are unaffected by the transformation ...

MY SOLUTION:
This question mentioned that we need to use the transformed factor loadings $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}\boldsymbol{M}$. Let's assume that $\boldsymbol{M}$ is an $k \times k$ orthogonal matrix. We can re-write the the basic regression equation linking the observed and the factors as:

$$\boldsymbol{x} = (\boldsymbol{\Lambda}\boldsymbol{M})(\boldsymbol{M}^T\boldsymbol{f}) + \boldsymbol{u}$$

.
Using the rule of variance, we can have

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda}\boldsymbol{M})(\boldsymbol{\Lambda}\boldsymbol{M})^T + \Psi$$

. Since the $\boldsymbol{M}$ is a orthogonal matrix and $\boldsymbol{M}\boldsymbol{M}^T = \boldsymbol{I}$. Therefore, the variance equation can be written as

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \Psi$$

. That is, the transformed factor loadings $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}\boldsymbol{M}$ will not influence the communalities (i.e., $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T$) in the a factor analysis model.

0.4 Ex 5.3

Give a formula for the proportion of variance explained by the jth factor estimated by the principal factor approach.

MY SOLUTION:
The proportion of variance explained by the jth factor represents the proportion of the total variance in the observed variables that is accounted for by that factor alone. Therefore, the formula could be

$$Proportion_j = \frac{\sum_{i=1}^{q} \lambda_{ij}^2}{\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T}$$

.

0.5 Ex 5.4

Apply the factor analysis model separately to the life expectancies of men and women and compare the results.

MY SOLUTION:
The textbook does not provide the original dataset. Based on the code in the `MVA`, I create the dataset via a separated r file named "HW05 Test". This file created the `life.rdata` and `life.csv` dataset in the same file folder.

```
> load("life.rdata")
> head(life)
           m0 m25 m50 m75 w0 w25 w50 w75
Algeria    63  51  30  13 67  54  34  15
Cameroon   34  29  13   5 38  32  17   6
Madagascar 38  30  17   7 38  34  20   7
Mauritius  59  42  20   6 64  46  25   8
Reunion    56  38  18   7 62  46  25  10
Seychelles 62  44  24   7 69  50  28  14
>
> # subset the male and female dataset
> life_male <- life[,1:4]
> life_female <- life[,5:8]
>
> # test the number of factors needed for the male and female dataset separately
> sapply(1, function(f)
+   factanal(life_male, factors=f, method="mle")$PVAL)
   objective
0.0007284301
> sapply(1, function(f)
+   factanal(life_female, factors=f, method="mle")$PVAL)
   objective
4.738464e-12
```

When test the number of the factors from 1 to larger number, there is always a warning that `N factors are too many for N variables`. More details can be found on Page 143 of the textbook or here https://stats.stackexchange.com/questions/593452/efa-n-factors-are-too-many-for-n-variables

The results suggest that an one-factor solution might be adequate to account for the observed covariances in the data.

Next, I run the one-factor solution for both male and female datasets.

```
> factanal(life_male, factors = 1, method="mle")

Call:
factanal(x = life_male, factors = 1, method = "mle")

Uniquenesses:
   m0   m25   m50   m75
0.594 0.552 0.005 0.434

Loadings:
    Factor1
m0  0.638
m25 0.669
m50 0.998
m75 0.752

               Factor1
SS loadings      2.415
Proportion Var   0.604

Test of the hypothesis that 1 factor is sufficient.
```

```
The chi square statistic is 14.45 on 2 degrees of freedom.
The p-value is 0.000728
> factanal(life_female, factors = 1, method="mle")

Call:
factanal(x = life_female, factors = 1, method = "mle")

Uniquenesses:
   w0   w25   w50   w75
0.220 0.005 0.115 0.526

Loadings:
    Factor1
w0  0.883
w25 0.998
w50 0.941
w75 0.689

               Factor1
SS loadings      3.134
Proportion Var   0.784

Test of the hypothesis that 1 factor is sufficient.
The chi square statistic is 52.15 on 2 degrees of freedom.
The p-value is 4.74e-12
```

0.6 Ex 5.6

The matrix below shows the correlations between ratings on nine statements about pain made by 123 people suffering from extreme pain. Each statement was scored on a scale from 1 to 6, ranging from agreement to disagreement. The nine pain statements were as follows:

MY SOLUTION:
First, to change the lower triangular matrix into the complete correlation matrix.

```
> library(Matrix)
> # import the correlation matrix
> corr_lower <- matrix(c(1, 0, 0, 0, 0, 0, 0, 0, 0,
+                        -0.04, 1, 0,0,0,0,0,0,0,
+                        0.61, -0.07, 1,0,0,0,0,0,0,
+                        0.45, -0.12, 0.59, 1,0,0,0,0,0,
+                        0.03, 0.49, 0.03, -0.08, 1,0,0,0,0,
+                        -0.29, 0.43, -0.13, -0.21, 0.47, 1,0,0,0,
+                        -0.30, 0.30, -0.24, -0.19, 0.41, 0.63,1,0,0,
+                        0.45, -0.31,0.59,0.63,-0.14,-0.13,-0.26,1,0,
+                        0.30,-0.17,-0.32,0.37,-0.24,-0.15,-0.29,0.40,1),9,9, byrow = T)
> # generate a complete correlation matrix
> corr_symmetric <- forceSymmetric(corr_lower, uplo="L")
```

The correlation matrix looks good. Next, I run the PCA first.

```
> # run the PCA first
> # use prcomp to calculate the principal components
> pca <- prcomp(corr_symmetric, scale. = FALSE)
> # get the PCA results
> summary(pca)
Importance of components:
                          PC1    PC2     PC3     PC4    PC5     PC6     PC7
Standard deviation     1.1594 0.5014 0.31822 0.18698 0.1734 0.15725 0.09446
Proportion of Variance 0.7466 0.1396 0.05624 0.01942 0.0167 0.01373 0.00496
Cumulative Proportion  0.7466 0.8863 0.94250 0.96192 0.9786 0.99235 0.99730
                           PC8       PC9
Standard deviation     0.06968 3.631e-17
Proportion of Variance 0.00270 0.000e+00
Cumulative Proportion  1.00000 1.000e+00
> # draw the scree plot
> plot(pca, type = "l",
+       main = "Scree Plot")
```