ST 437/537: Applied Multivariate and Longitudinal Data Analysis

Factor Analysis: Overview

Arnab Maity
NCSU Department of Statistics
SAS Hall 5240 919-515-1937 amaity[at]ncsu.edu

Introduction

The primary purpose of *factor analysis* is to describe the covariance structure of multiple variables in terms of a few underlying, *unobservable*, random variables called *factors*.

According to Johnson and Wichern (2007), Karl Pearson and Charles Spearman and others were the proponents of modern factor analysis models in the early 20th century. Charles Spearman proposed his “single factor” theory of intelligence in 1904. Specifically, Spearman considered several measures of the mental ability of children (examination scores in several subjects such as classics, French, English, Mathematics, and music) and proposed that a *single unobserved variable* can explain the relationship among these variables, called the “general intelligence,” g.

In general, suppose the observable variables can be grouped by their correlations (all variables in a group are highly correlated among themselves but small correlations with other groups). Then it might be that a single underlying factor controls the variables in each group. In other words, variables or factors like “intelligence” can not be measured directly; we can only observe their impact through some observable variables or *manifest variables*, and infer about the underlying factors by inspecting covariance among the manifest variables. Factor analysis formally attempts to confirm such structures.

There two types of factor analysis, as we discuss below.

Exploratory Factor Analysis (EFA)

Exploratory factor analysis is used to investigate whether any factors are underlying the covariance structure among the manifest variables *without making assumptions* about which factors are related to which of the manifest variables. Main reasons for performing such an analysis are to

- investigate the structure of the covariance relationship among the manifest variables
- data/dimension reduction
- scoring of different attributes via so-called *factor scores*

The dimension reduction aspect of factor analysis often plays an essential role in multivariate statistics. In the situation, where one observes data on a large number of variables, but only has a few observations (small sample size), factor analysis can help reduce the number of variables by grouping highly correlated variables.

The model used in EFA is called the *common factor model*, that is, a set of factors contribute to the covariance among the manifest variables. In the case of Spearman's single factor model, suppose the manifest variable are $X = (X_1, \ldots, X_p)$. The the single factor model is

$$\begin{aligned} X_1 - \mu_1 &= \lambda_1 f + u_1 \\ \vdots & \qquad \vdots \\ X_p - \mu_p &= \lambda_p f + u_p. \end{aligned}$$

Here f is an unobserved random variable called the *common factor*, λ_i (called *loadings*; unknown) quantifies the strength of the association between the common factor and the i-th manifest variable X_i, and u_i is an unobserved random variable describes the part of X_i not explained by the common factor f. This model can be generalized to accommodate multiple factors as well. For example, a three-factor model can be written as below:

$$\begin{aligned} X_1 - \mu_1 &= \lambda_{11} f_1 + \lambda_{12} f_2 + \lambda_{13} f_3 + u_1 \\ \vdots & \qquad \vdots \\ X_p - \mu_p &= \lambda_{p1} f_1 + \lambda_{p2} f_2 + \lambda_{p3} f_3 + u_p, \end{aligned}$$

where we now have three common factors, f_1, f_2 and f_3. The loadings λ_{ij} quantifies the relationship between the j-th common factor, f_j, with the i-th manifest variable, X_i.

Our goal is to estimate/predict all the unknown components on the right-hand side of the equations.

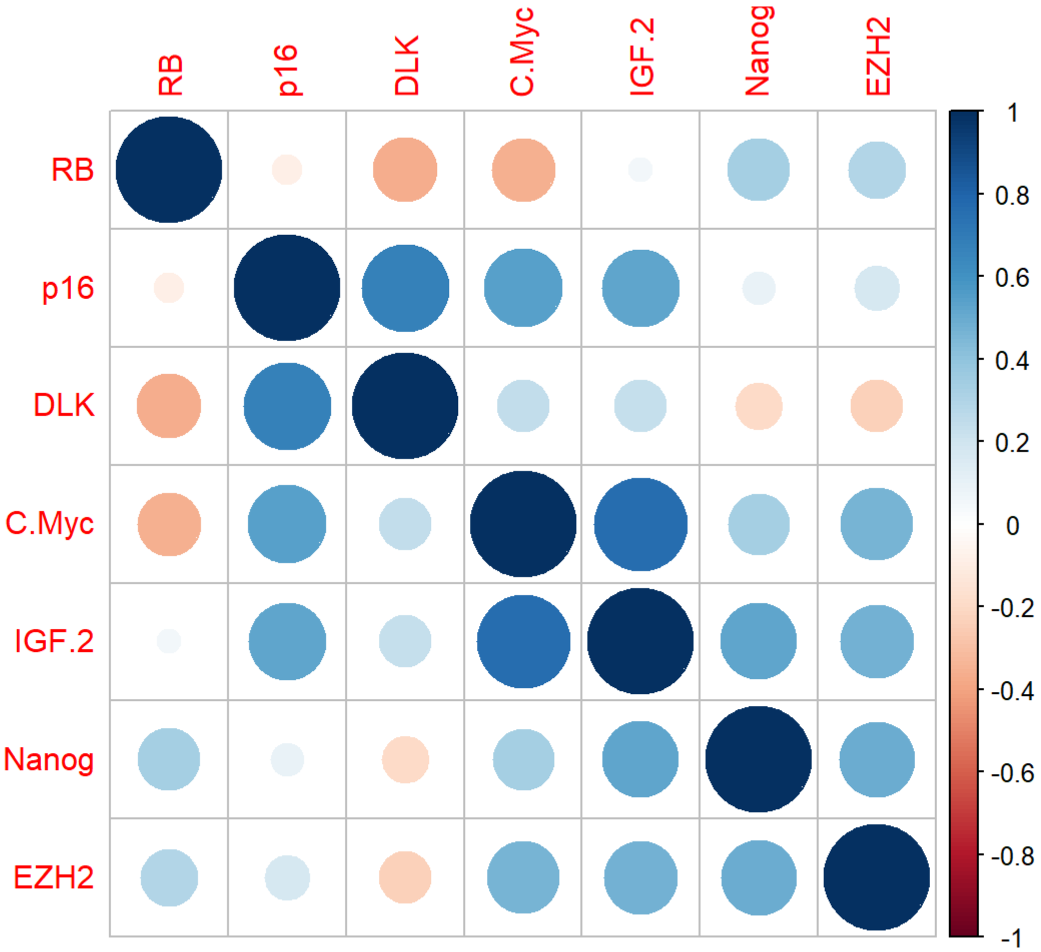
As a quick example, consider the Hemangioma data [Table 8.2 of Applied Multivariate Statistics with R by Daniel Zelterman. New York: Springer]. The dataset contains age (in days) at the time of surgery and expression of seven genetic markers for infants who were surgically treated for hemangioma. The hypothesis of interest is that exposure to the tumor might influence the expression measurements of some of these genes over time.

```
# read the data
dat <- read.table("data/hemangioma.txt", header = T)

# snapshot
head(dat)
```

```
##   Age       RB      p16       DLK     Nanog     C.Myc     EZH2     IGF.2
## 1  81 2.046149 3.067127  308974.7  94.17336  6.489601 2.764101 11175.689
## 2  95 6.540000 1.900000   70988.3 381.83000  1.000000 7.090000  5340.170
## 3  95 3.610000 3.820000  153060.6 237.28000  0.000000 5.570000  6310.240
## 4 165 1.912267 3.735868  596991.6  88.23737  0.000000 2.469633  7008.523
## 5 286 2.625436 5.168293  369600.6 282.29727 12.225828 1.628923  7104.238
## 6 299 2.870760 5.755246 1119257.5 176.75143  8.764235 3.511469  9342.126
```

```
# visual of the correlation matrix
# arrange the variables according their correlation
library(corrplot)
corrplot( cor(dat[, -1]), order = "hclust" )
```



Based on the grouping seen in the correlation matrix, it is conceivable that one factor might be controlling `C.Myc`, `IGF.2`, `Nanog` and `EZH2`. Perhaps, there is another factor controlling the other variables. At this point, we do not have any concrete hypothesis about how many factors there might be, what they are or which of them are related to which of the manifest variables. This is a situation where EFA can be applied.

Confirmatory Factor Analysis (CFA)

The exploratory factor analysis is typically used in preliminary/pilot studies to find whether a factor analysis is useful for a given multivariate dataset. The EFA is used to determine how many factors there might be and how are they related to the manifest variables. The second type of factor analysis, a confirmatory factor analysis, seeks to *formally test* whether a *pre-specified factor model* fits the covariance among the manifest variable well enough.

For example, let us consider the ability data (actually a correlation matrix) in the `MVA` package (Figure 7.1 in Everitt and Hothorn). Six variables were recorded [Calsyn and Kenny (1977)] for 556 eighth-grade students:

SCA: self-concept of ability;

PPE: perceived parental evaluation;

PTE: perceived teacher evaluation;

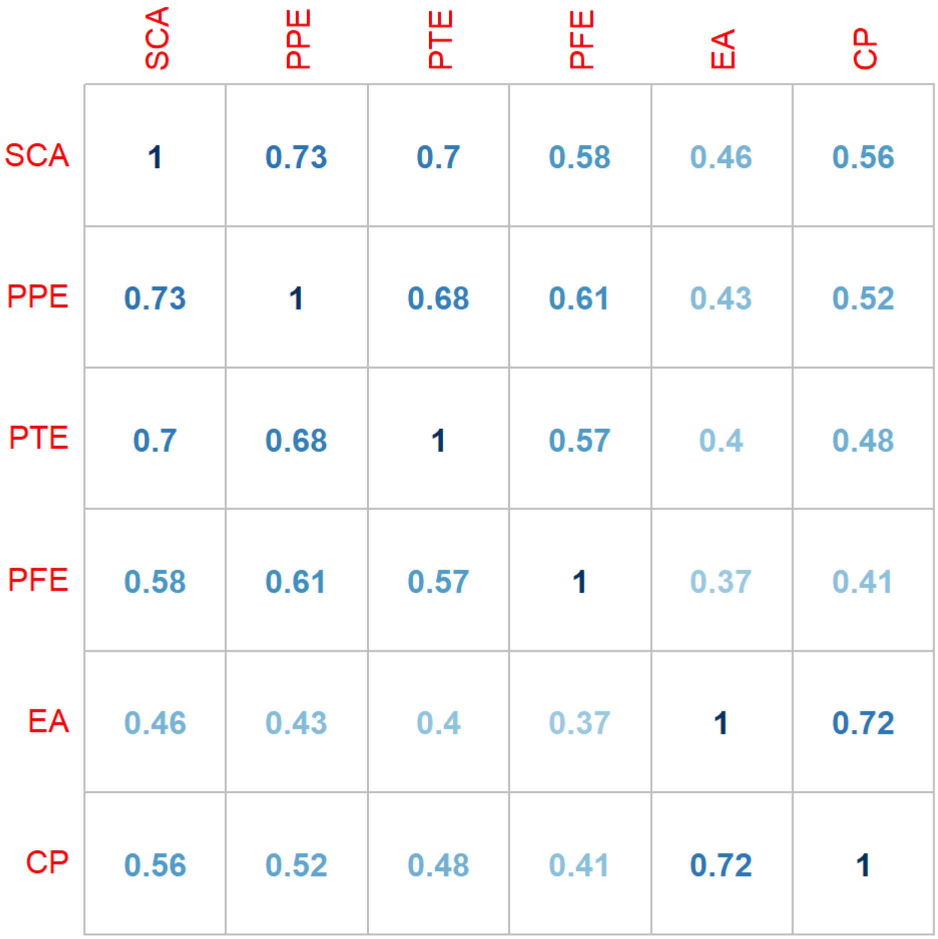
PFE: perceived friend’s evaluation;

EA: educational aspiration;

CP: college plans.

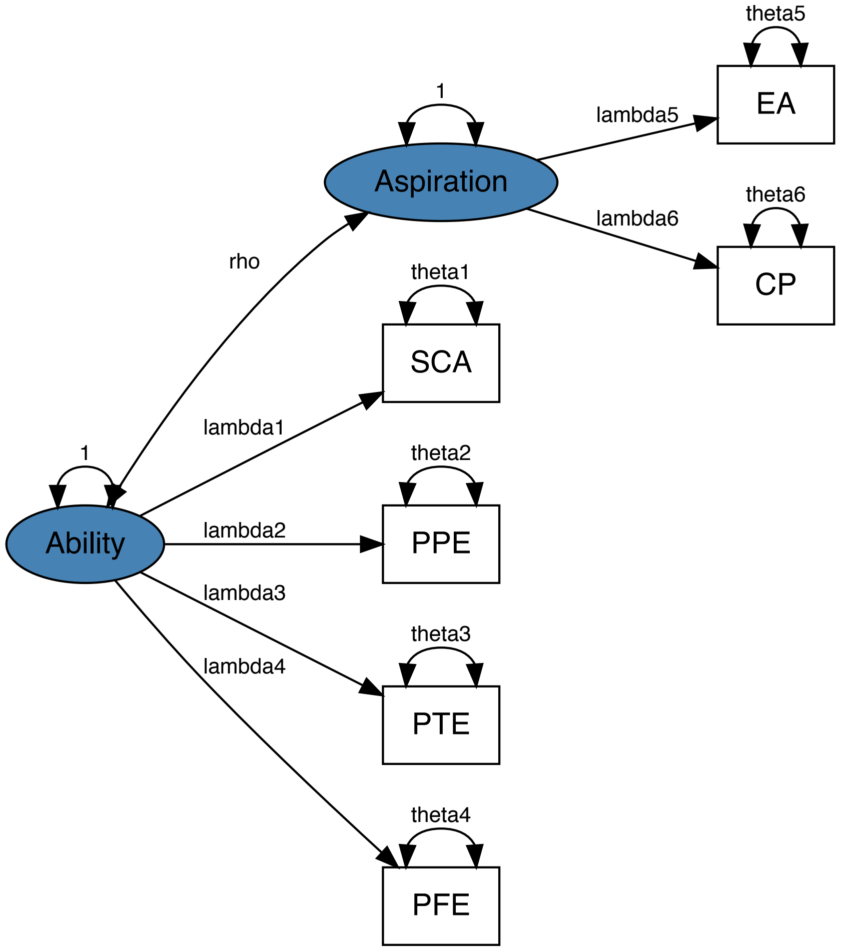
The ability dataset shows the correlation among these variables.

```
##      SCA  PPE  PTE  PFE   EA   CP
## SCA 1.00 0.73 0.70 0.58 0.46 0.56
## PPE 0.73 1.00 0.68 0.61 0.43 0.52
## PTE 0.70 0.68 1.00 0.57 0.40 0.48
## PFE 0.58 0.61 0.57 1.00 0.37 0.41
## EA  0.46 0.43 0.40 0.37 1.00 0.72
## CP  0.56 0.52 0.48 0.41 0.72 1.00
```



	SCA	PPE	PTE	PFE	EA	CP
SCA	1	0.73	0.7	0.58	0.46	0.56
PPE	0.73	1	0.68	0.61	0.43	0.52
PTE	0.7	0.68	1	0.57	0.4	0.48
PFE	0.58	0.61	0.57	1	0.37	0.41
EA	0.46	0.43	0.4	0.37	1	0.72
CP	0.56	0.52	0.48	0.41	0.72	1

Calsyn and Kenny (1977) postulated that there are two factors, and they relate to the manifest variables as follows:



The variables in the ellipses are factors, and the variables in the squares are manifest variables. Confirmatory factor analysis can be use here to formally test whether this specfic factor model fit the data well.

Main page: **ST 437/537: Applied Multivariate and Longitudinal Data Analysis (https://maityst537.wordpress.ncsu.edu/)**