因子分析

王宁宁

广州医科大学统计学系

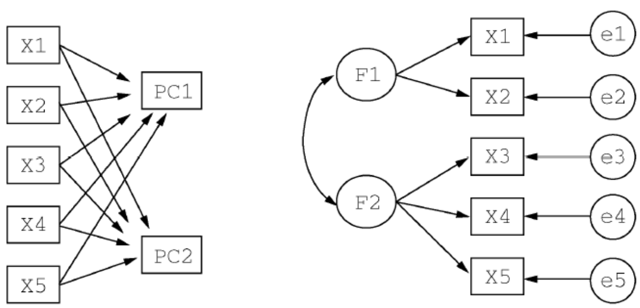
概述

- 因子分析（Factor Analysis，FA）是指研究从变量群中提取共性因子的统计技术。
- 最早由英国心理学家 C.E.斯皮尔曼提出。他发现学生的各科成绩之间存在着一定的相关性，一科成绩好的学生，往往其他各科成绩也比较好，从而推想是否存在某些潜在的共性因子，或称某些一般智力条件影响着学生的学习成绩。
- 因子分析可在许多变量中找出隐藏的具有代表性的因子。将相同本质的变量归入一个因子，可减少变量的数目，还可检验变量间关系的假设。

因子分析一般可以分为：

- 探索性因子分析（Explorey Factor Analysis）
- 验证性因子分析 (Confirmatory Factor Analysis)

比较因子分析和主成份分析



(a) 主成分分析模型　　(b) 因子分析模型

模型如下：

$$
\begin{aligned}
X_1 - \mu_1 &= \ell_{11}F_1 + \ell_{12}F_2 + \cdots + \ell_{1m}F_m + \varepsilon_1 \\
X_2 - \mu_2 &= \ell_{21}F_1 + \ell_{22}F_2 + \cdots + \ell_{2m}F_m + \varepsilon_2 \\
&\vdots \qquad\qquad\qquad \vdots \\
X_p - \mu_p &= \ell_{p1}F_1 + \ell_{p2}F_2 + \cdots + \ell_{pm}F_m + \varepsilon_p
\end{aligned}
$$

F_i是第i个公共因子 ϵ_i是第i个特殊因子 l_{ij}是第第j个公共因子的第i个荷载因子

写成矩阵形式：

$$\underset{(p\times1)}{\mathbf{X}} = \underset{(p\times1)}{\boldsymbol{\mu}} + \underset{(p\times m)}{\mathbf{L}}\underset{(m\times1)}{\mathbf{F}} + \underset{(p\times1)}{\boldsymbol{\varepsilon}}$$

$$\begin{aligned}
\mu_i &= \textit{mean} \text{ of variable } i \\
\varepsilon_i &= i\text{th } \textit{specific factor} \\
F_j &= j\text{th } \textit{common factor} \\
\ell_{ij} &= \textit{loading} \text{ of the } i\text{th variable on the } j\text{th factor}
\end{aligned}$$

下面考虑：

$$\begin{aligned}
(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})' &= (\mathbf{LF}+\boldsymbol{\varepsilon})(\mathbf{LF}+\boldsymbol{\varepsilon})' \\
&= (\mathbf{LF}+\boldsymbol{\varepsilon})((\mathbf{LF})'+\boldsymbol{\varepsilon}') \\
&= \mathbf{LF}(\mathbf{LF})' + \boldsymbol{\varepsilon}(\mathbf{LF})' + \mathbf{LF}\boldsymbol{\varepsilon}' + \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'
\end{aligned}$$

对上式左右同时取期望：

$$\begin{aligned}
\boldsymbol{\Sigma} &= \text{Cov}(\mathbf{X}) = E(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})' \\
&= \mathbf{L}E(\mathbf{FF}')\mathbf{L}' + E(\boldsymbol{\varepsilon}\mathbf{F}')\mathbf{L}' + \mathbf{L}E(\mathbf{F}\boldsymbol{\varepsilon}') + E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') \\
&= \mathbf{LL}' + \boldsymbol{\Psi}
\end{aligned}$$

模型假定：

$\mathbf{F}$ and $\boldsymbol{\varepsilon}$ are independent

$E(\mathbf{F}) = \mathbf{0}$, $\text{Cov}(\mathbf{F}) = \mathbf{I}$

$E(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\text{Cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Psi}$, where $\boldsymbol{\Psi}$ is a diagonal matrix

方差结构：

1. $\text{Cov}(\mathbf{X}) = \mathbf{LL}' + \boldsymbol{\Psi}$

or

$$\text{Var}(X_i) = \ell_{i1}^2 + \cdots + \ell_{im}^2 + \psi_i$$

$$\text{Cov}(X_i, X_k) = \ell_{i1}\ell_{k1} + \cdots + \ell_{im}\ell_{km}$$

2. $\text{Cov}(\mathbf{X}, \mathbf{F}) = \mathbf{L}$

or

$$\text{Cov}(X_i, F_j) = \ell_{ij}$$

因子分析是方差协方差分析的一种方法。

目的：

- 求公共因子
- 求荷载因子

下面说明上述的荷载因子并不是唯一的：

$$\mathbf{X} - \boldsymbol{\mu} = \mathbf{LF} + \boldsymbol{\varepsilon} = \mathbf{LTT}'\mathbf{F} + \boldsymbol{\varepsilon} = \mathbf{L}^*\mathbf{F}^* + \boldsymbol{\varepsilon}$$

$$\mathbf{L}^* = \mathbf{LT} \quad \text{and} \quad \mathbf{F}^* = \mathbf{T}'\mathbf{F}$$

$$E(\mathbf{F}^*) = \mathbf{T}'E(\mathbf{F}) = \mathbf{0}$$

$$\text{Cov}(\mathbf{F}^*) = \mathbf{T}'\text{Cov}(\mathbf{F})\mathbf{T} = \mathbf{T}'\mathbf{T} = \underset{(m\times m)}{\mathbf{I}}$$

荷载矩阵在任意一个正交矩阵的作用下都会不会改变方差结构：

$$\boldsymbol{\Sigma} = \mathbf{LL}' + \boldsymbol{\Psi} = \mathbf{LTT}'\mathbf{L}' + \boldsymbol{\Psi} = (\mathbf{L}^*)(\mathbf{L}^*)' + \boldsymbol{\Psi}$$

选择公因子数目：

- 碎石图平行分析
- 主观

求公因子的方法：

- 主成份法
- 极大似然法
- 最小残差法
- 加权最小二乘法
- R 提供了 6 种不同的方法(stats 提供了一种，包 psych 提供另外五种)

求荷载因子：

- 最大方差法 varimax
- 斜交旋转 promax
- 最大分位数 quartimax
- bentlerT
- R（包 psych） 提供了 15 种不同的方法

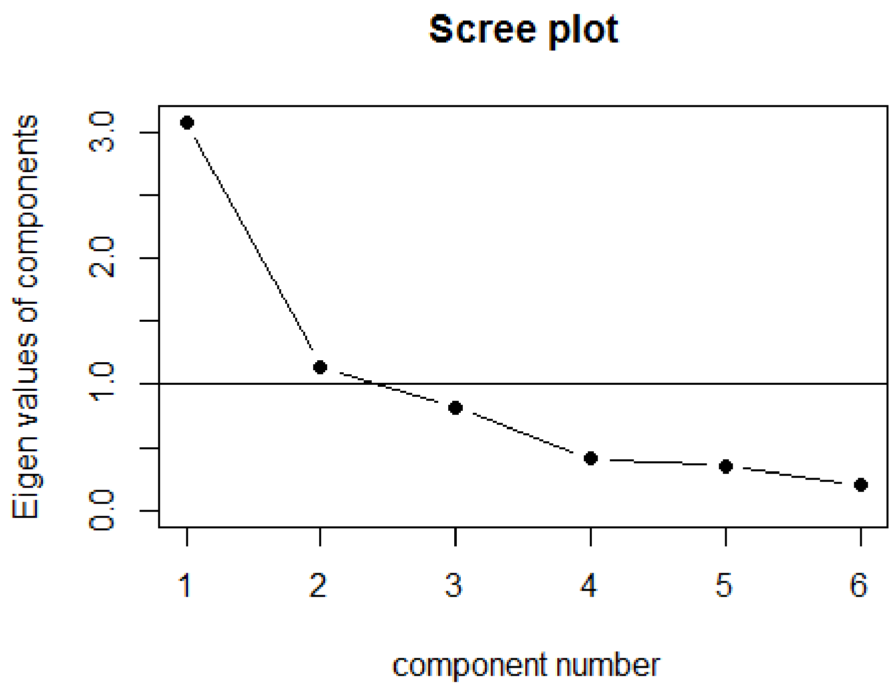
例子

112 个人参与了六个测验，包括非语言的普通智力测验（general）、画图测验（picture）、积木图案测验（blocks）、迷津测验（maze）、阅读测验（reading）和词汇测验（vocab）。我们如何用一组较少的、潜在的心理学因素来解释参与者的测验得分呢？

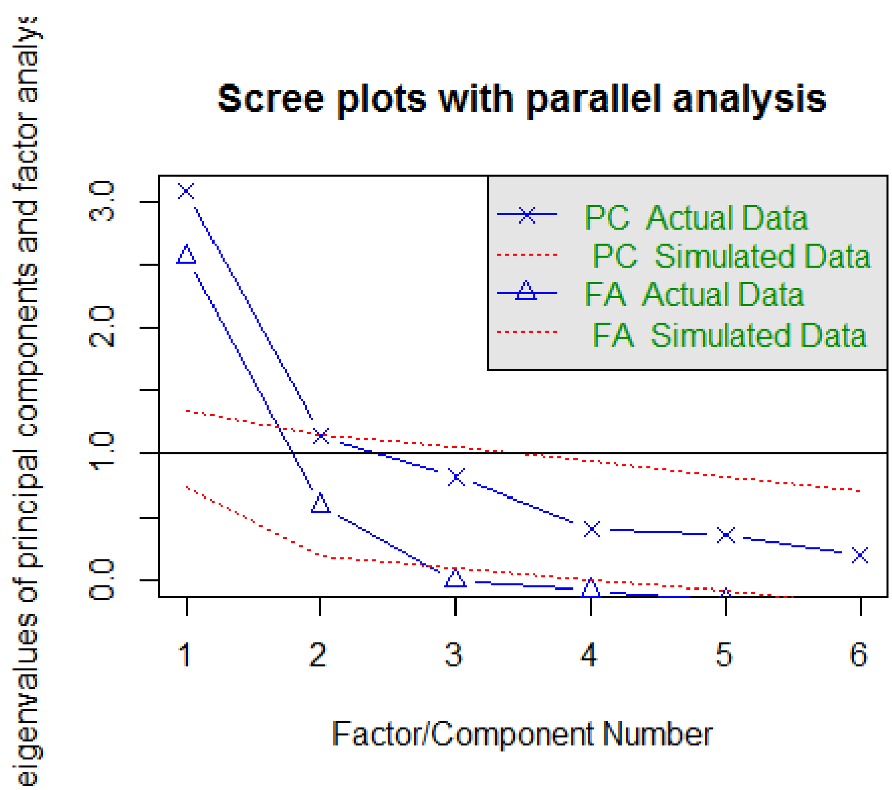
```
help(ability.cov)

#install.packages("psych")
library(psych)
covariances <- ability.cov$cov
# convert covariances to correlations
correlations <- cov2cor(covariances)
correlations

##           general   picture    blocks      maze   reading     vocab
## general 1.0000000 0.4662649 0.5516632 0.3403250 0.5764799 0.5144058
## picture 0.4662649 1.0000000 0.5724364 0.1930992 0.2629229 0.2392766
## blocks  0.5516632 0.5724364 1.0000000 0.4450901 0.3540252 0.3564715
## maze    0.3403250 0.1930992 0.4450901 1.0000000 0.1839645 0.2188370
## reading 0.5764799 0.2629229 0.3540252 0.1839645 1.0000000 0.7913779
## vocab   0.5144058 0.2392766 0.3564715 0.2188370 0.7913779 1.0000000
```

- 碎石图
- 平行分析

```
scree(correlations,factors = F)
```




```r
#win.graph(width = 12,height = 9,pointsize = 8)
fa.parallel(correlations, n.obs = 112, fa = "both", main = "Scree plots
 with parallel analysis")
```



```
## Parallel analysis suggests that the number of factors =  2  and the
number of components =  1
```

建议选择两个公共因子

没有旋转的因子分析，选择主成份法：

```
fa <- fa(correlations, nfactors = 2, rotate = "none", fm = "pa")
fa
```

```
## Factor Analysis using method =  pa
## Call: fa(r = correlations, nfactors = 2, rotate = "none", fm = "pa")
## Standardized loadings (pattern matrix) based upon correlation matrix
##          PA1   PA2   h2    u2 com
## general 0.75  0.07 0.57 0.432 1.0
## picture 0.52  0.32 0.38 0.623 1.7
## blocks  0.75  0.52 0.83 0.166 1.8
## maze    0.39  0.22 0.20 0.798 1.6
## reading 0.81 -0.51 0.91 0.089 1.7
## vocab   0.73 -0.39 0.69 0.313 1.5
##
##                        PA1  PA2
## SS loadings           2.75 0.83
## Proportion Var        0.46 0.14
## Cumulative Var        0.46 0.60
## Proportion Explained  0.77 0.23
## Cumulative Proportion 0.77 1.00
##
## Mean item complexity =  1.5
## Test of the hypothesis that 2 factors are sufficient.
##
## The degrees of freedom for the null model are  15  and the objective
 function was  2.48
## The degrees of freedom for the model are 4  and the objective functi
on was  0.07
##
## The root mean square of the residuals (RMSR) is  0.03
## The df corrected root mean square of the residuals is  0.06
##
## Fit based upon off diagonal values = 0.99
## Measures of factor score adequacy
##                                                  PA1  PA2
## Correlation of scores with factors              0.96 0.92
## Multiple R square of scores with factors        0.93 0.84
## Minimum correlation of possible factor scores   0.86 0.68
```

最大方差法的旋转：

```
fa.varimax <- fa(correlations, nfactors = 2, rotate = "varimax", fm = "
pa")
fa.varimax
```

```
## Factor Analysis using method =  pa
## Call: fa(r = correlations, nfactors = 2, rotate = "varimax", fm = "p
a")
## Standardized loadings (pattern matrix) based upon correlation matrix
##          PA1  PA2   h2    u2 com
## general 0.49 0.57 0.57 0.432 2.0
## picture 0.16 0.59 0.38 0.623 1.1
## blocks  0.18 0.89 0.83 0.166 1.1
## maze    0.13 0.43 0.20 0.798 1.2
## reading 0.93 0.20 0.91 0.089 1.1
## vocab   0.80 0.23 0.69 0.313 1.2
##
##                        PA1  PA2
## SS loadings           1.83 1.75
## Proportion Var        0.30 0.29
## Cumulative Var        0.30 0.60
## Proportion Explained  0.51 0.49
## Cumulative Proportion 0.51 1.00
##
## Mean item complexity =  1.3
## Test of the hypothesis that 2 factors are sufficient.
##
## The degrees of freedom for the null model are  15  and the objective
 function was  2.48
## The degrees of freedom for the model are 4  and the objective functi
on was  0.07
##
## The root mean square of the residuals (RMSR) is  0.03
## The df corrected root mean square of the residuals is  0.06
##
## Fit based upon off diagonal values = 0.99
## Measures of factor score adequacy
##                                                   PA1  PA2
## Correlation of scores with factors               0.96 0.92
## Multiple R square of scores with factors         0.91 0.85
## Minimum correlation of possible factor scores    0.82 0.71
```

利用斜交旋转提取因子：

```
fa.promax <- fa(correlations, nfactors = 2, rotate = "promax", fm = "pa
")
```

```
## Warning in fac(r = r, nfactors = nfactors, n.obs = n.obs, rotate =
## rotate, : A Heywood case was detected. Examine the loadings carefull
y.
```

```
fa.promax
```

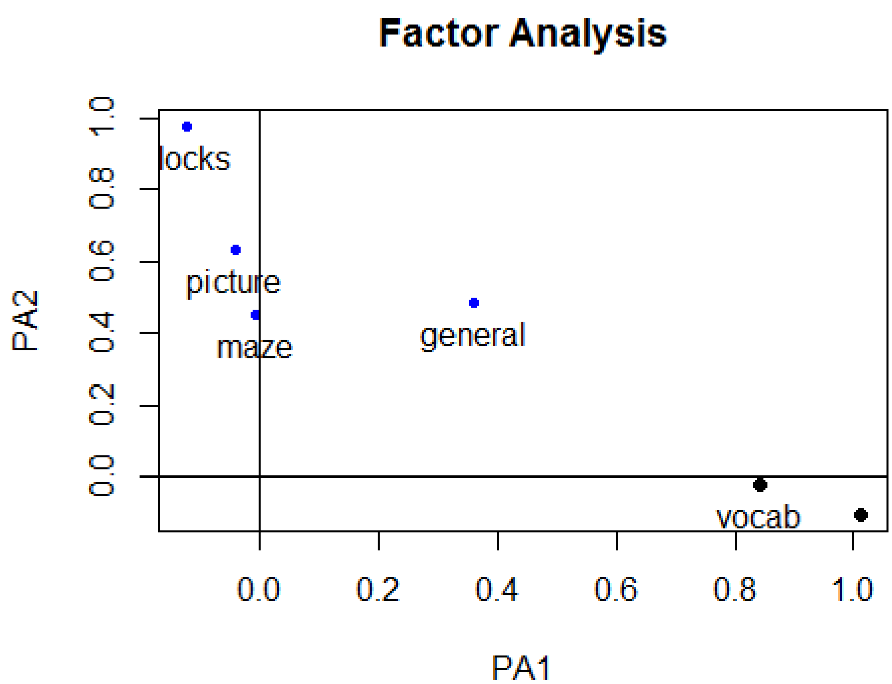
```
## Factor Analysis using method =  pa
## Call: fa(r = correlations, nfactors = 2, rotate = "promax", fm = "pa
")
```

```
##
##  Warning: A Heywood case was detected.
## Standardized loadings (pattern matrix) based upon correlation matrix
##           PA1   PA2   h2    u2 com
## general  0.36  0.49 0.57 0.432 1.8
## picture -0.04  0.64 0.38 0.623 1.0
## blocks  -0.12  0.98 0.83 0.166 1.0
## maze    -0.01  0.45 0.20 0.798 1.0
## reading  1.01 -0.11 0.91 0.089 1.0
## vocab    0.84 -0.02 0.69 0.313 1.0
##
##                        PA1  PA2
## SS loadings           1.82 1.76
## Proportion Var        0.30 0.29
## Cumulative Var        0.30 0.60
## Proportion Explained  0.51 0.49
## Cumulative Proportion 0.51 1.00
##
##  With factor correlations of
##      PA1  PA2
## PA1 1.00 0.57
## PA2 0.57 1.00
##
## Mean item complexity =  1.2
## Test of the hypothesis that 2 factors are sufficient.
##
## The degrees of freedom for the null model are  15  and the objective
 function was  2.48
## The degrees of freedom for the model are 4  and the objective functi
on was  0.07
##
## The root mean square of the residuals (RMSR) is  0.03
## The df corrected root mean square of the residuals is  0.06
##
## Fit based upon off diagonal values = 0.99
## Measures of factor score adequacy
##                                                   PA1  PA2
## Correlation of scores with factors               0.97 0.94
## Multiple R square of scores with factors         0.93 0.89
## Minimum correlation of possible factor scores    0.86 0.77

# Calculate factor loading matrix

fsm <- function(oblique) {
  if (class(oblique)[2]=="fa" & is.null(oblique$Phi)) {
    warning("Object doesn't look like oblique EFA")
  } else {
    P <- unclass(oblique$loading)
    F <- P %*% oblique$Phi
    colnames(F) <- c("PA1", "PA2")
```

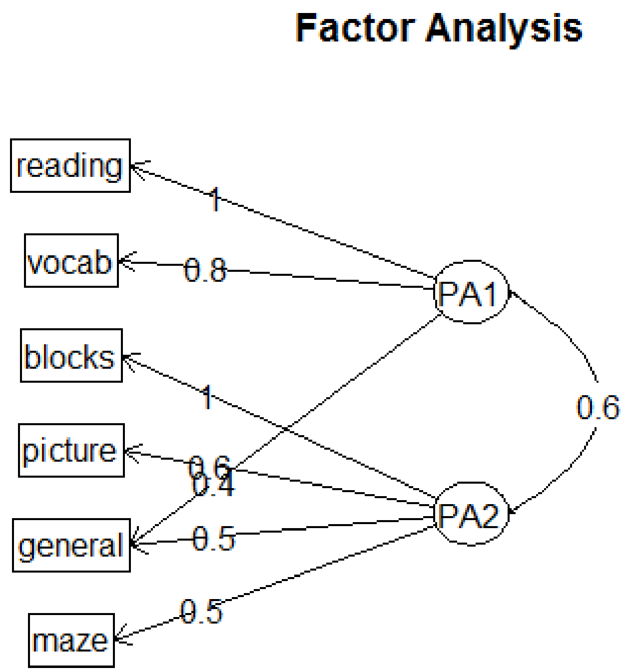
```r
    return(F)
  }
}

fsm(fa.promax)
```

```
##               PA1       PA2
## general 0.6398556 0.6927493
## picture 0.3250348 0.6133638
## blocks  0.4365629 0.9075015
## maze    0.2525385 0.4496097
## reading 0.9503302 0.4720320
## vocab   0.8285707 0.4586943
```

```r
factor.plot(fa.promax, labels =rownames(fa.promax$loadings))
```




```r
fa.diagram(fa.promax, simple = FALSE)
```



主成份分析和因子分析的异同

- 因子分析中是把变量表示成各因子的线性组合，而主成分分析中则是把主成分表示成个变量的线性组合。
- 主成分分析的重点在于解释个变量的总方差，而因子分析则把重点放在解释各变量之间的协方差。
- 主成分分析中不需要有假设(assumptions)，因子分析则需要一些假设。因子分析的假设包括：各个共同因子之间不相关，特殊因子（specific factor）之间也不相关，共同因子和特殊因子之间也不相关。
- 主成分分析中，当给定的协方差矩阵或者相关矩阵的特征值是唯一的时候，的主成分一般是独特的；而因子分析中因子不是独特的，可以旋转得到不同的因子。
- 在因子分析中，因子个数需要分析者指定，而指定的因子数量不同而结果不同。在主成分分析中，成分的数量是一定的，一般有几个变量就有几个主成分。和主成分分析相比，由于因子分析可以使用旋转技术帮助解释因子，在解释方面更加有优势。大致说来，当需要寻找潜在的因子，并对这些因子进行解释的时候，更加倾向于使用因子分析，并且借助旋转技术帮助更好解释。而如果想把现有的变量变成少数几个新的变量（新的变量几乎带有原来所有变量的信息）来进入后续的分析，则可以使用主成分分析。

练习

```r
d <- read.csv("http://statstudy.github.io/data/simCog.csv")
head(d)
```

```
##    Knowledge OralExpression  Deduction MentalRotation Visualization
## 1 -0.1937224     -0.3194100 -0.8310500     -0.2354839   -0.00874325
## 2 -0.2468147      0.8653481  0.6461765      1.5121300    0.35422918
## 3 -0.7808090      0.1591759 -1.4743426     -0.8502143   -1.27522722
## 4 -1.4049924     -0.7473637 -0.5223564      0.2194446   -1.04352583
## 5 -0.4138702     -0.7118456  0.4500109     -0.2184520    0.03457733
## 6 -0.4649210      0.7466816 -0.5005259      0.5053426    0.32853127
##    Vocabulary   Analogies Quantitative PatternRecognition
## 1  0.03621763 -0.82876798  -0.03681655        -0.87668865
## 2  0.39317066  0.43736748  -0.02804369         0.99727512
## 3 -0.90394664  0.58931766   0.25639006        -0.86685970
## 4 -1.11491005 -0.52233132  -1.51637252        -0.95695461
## 5  0.52532648  0.74736073  -0.67334222        -0.07362392
## 6 -1.70224923  0.02511076   1.81908129         0.01534438
```