第9章 因子分析及R使用

武慧
wuh@hit.edu.cn

经济管理学院
哈尔滨工业大学（威海）

1

因子分析的直观理解

1904年，英国心理学家Charles Spearman研究了33名学生在古典语、法语和英语三门成绩。

三门成绩高度相关

$$\mathbf{R} = \begin{array}{l} \text{Classics} \\ \text{French} \\ \text{English} \end{array} \begin{pmatrix} 1.00 & & \\ 0.83 & 1.00 & \\ 0.78 & 0.67 & 1.00 \end{pmatrix}$$

可否被一个“公共因子”刻画？
语言能力？

本章内容

1. 因子分析
2. 因子旋转
3. 因子分析的R实现

3

1. 因子分析简介

4

因子分析的直观理解

- 对于$\boldsymbol{p}$个原始变量$Y_1, Y_2, \cdots, Y_p$来说，那些高度相关的变量很可能会遵循一个共同的潜在结构——或可称之为公共因子。
- 因子分析旨在提出因子模型来研究如何用几个公共因子，记作$F_1, F_2, \cdots, F_m$，通常$m < p$，来刻画原始变量之间的相关性。
- 然而，这些“公共因子”通常是无法观测的，故称为潜变量。这在心理学、社会学及行为科学等学科中非常常见，比如“智力”和“社会阶层”。

5

单因子模型

- Charles Spearman基于学生3门语言成绩的数据提出了单因子模型：

$$
\begin{aligned}
Y_1 &= l_1 F + \varepsilon_1 \\
Y_2 &= l_2 F + \varepsilon_2 \\
Y_3 &= l_3 F + \varepsilon_3
\end{aligned}
$$

F ：公共因子（Common factor）
ε_j：特殊因子（Specific factor）
l_i：系数/载荷（Loading）

- 大多数时候一个公共因子是不够的，错综复杂的变量可能需要多个公共因子刻画，这就是我们将要学习的正交因子模型。

6

正交因子模型

- 假设可观测随机向量$\boldsymbol{y}=\left(Y_1,\cdots,Y_p\right)'$的均值为$\boldsymbol{\mu}$，协方差矩阵为$\boldsymbol{\Sigma}$。正交因子模型假定$\boldsymbol{y}$线性依赖于$m$个不可观测公共因子$\mathbf{f}=(F_1,\cdots,F_m)'$和$p$个不可观测的特殊因子$\boldsymbol{\varepsilon}=\left(\varepsilon_1,\cdots,\varepsilon_p\right)'$，通常$m<p$：

$$
\begin{aligned}
Y_1-\mu_1&=l_{11}F_1+l_{12}F_2+\ldots+l_{1m}F_m+\varepsilon_1\\
Y_2-\mu_2&=l_{21}F_1+l_{22}F_2+\ldots+l_{2m}F_m+\varepsilon_2\\
&\cdots\\
Y_p-\mu_p&=l_{p1}F_1+l_{p2}F_2+\ldots+l_{pm}F_m+\varepsilon_p
\end{aligned}
$$

其中，系数l_{jk}称为第j个变量在第k个因子上的载荷，体现了该公共因子对此变量的解释力。

矩阵形式和假设

- 使用矩阵记号，上述模型可写为

$$\mathbf{y}_{p\times 1}-\boldsymbol{\mu}_{p\times 1}=\mathbf{L}_{p\times m}\mathbf{f}_{m\times 1}+\boldsymbol{\varepsilon}_{p\times 1}$$

其中 $\mathbf{f}$ 和 $\boldsymbol{\varepsilon}$ 假设满足

正交性 (Orthogonality)

$$E(\mathbf{f})=\mathbf{0}_{m\times 1},\qquad COV(\mathbf{f})=E(\mathbf{f}\mathbf{f}')=\mathbf{I}_{m\times m}$$

$$E(\boldsymbol{\varepsilon})=\mathbf{0}_{p\times 1},\qquad COV(\boldsymbol{\varepsilon})=E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')=\boldsymbol{\Psi}_{p\times p}=\begin{bmatrix}\psi_1 & 0 & \cdots & 0\\ 0 & \psi_2 & \cdots & 0\\ \vdots & \vdots & & \vdots\\ 0 & 0 & \cdots & \psi_p\end{bmatrix}$$

而且 $\mathbf{f}$ 与 $\boldsymbol{\varepsilon}$ 是无关的：$COV(\boldsymbol{\varepsilon},\mathbf{f})=\mathbf{O}_{p\times m}$

8

协方差矩阵分解

矩阵形式模型：$\mathbf{y}-\boldsymbol{\mu}=\mathbf{Lf}+\boldsymbol{\varepsilon}$

元素形式模型：$Y_j-\mu_j=l_{j1}F_1+l_{j2}F_2+\ldots+l_{jm}F_m+\varepsilon_j, j=1,\ldots,p$

$\boldsymbol{\Sigma}=COV(\mathbf{y})=\mathbf{LL}'+\boldsymbol{\Psi}$

- $\sigma_{jj}=var(Y_j)=h_j^2+\psi_j$ ，其中 $h_j^2=l_{j1}^2+\ldots+l_{jm}^2$
  - h_j^2 体现了**共同性(communality)**，指由 m 个公共因子贡献的方差
  - ψ_j 称为**唯一性(uniqueness)**或**个体方差(specific variance)**，指无法由公共因子贡献的方差部分
- $\sigma_{jj'}=cov(Y_j,Y_j')=l_{j1}l_{j'1}+\ldots+l_{jm}l_{j'm},\ j\neq j'$

$COV(\mathbf{y},\mathbf{f})=\mathbf{L}$，因此载荷 l_{jk} 测度了第 j 个变量与第 k 个公共因子之间的关联 $cov(Y_j,F_k)=l_{jk}$

9

单选题 1分

因子分析的主要目的是：

A 发现数据中的模式和趋势

B 通过公共因子刻画原始变量之间的相关性

C 预测因变量的变化

D 计算各个变量的均值

10

单选题 1分

在因子分析中，哪些因素被认为是公共因子?

A 只与某个特定变量相关的因子

B 只能用于定性数据的因子

C 可以解释多个变量之间的相关性的因子

D 与原始变量没有任何关系的因子

11

2. 因子旋转

12

因子及载荷的不唯一性

$$\mathbf{y}-\boldsymbol{\mu}=\mathbf{Lf}+\boldsymbol{\varepsilon}$$

- 以上因子模型等价于$\boldsymbol{y}-\boldsymbol{\mu}=\boldsymbol{L}^{*}\boldsymbol{f}^{*}+\boldsymbol{\varepsilon}$，其中$\boldsymbol{L}^{*}=\boldsymbol{LT}$，$\boldsymbol{f}^{*}=\boldsymbol{T}'\boldsymbol{f}$，$\boldsymbol{T}$是满足$\boldsymbol{TT}'=\boldsymbol{T}'\boldsymbol{T}=\boldsymbol{I}$的任意正交矩阵。

➢ 新的载荷矩阵$\boldsymbol{L}^{*}$任然满足：

$$\boldsymbol{\Sigma}=\mathbf{LL}'+\boldsymbol{\Psi}=\mathbf{L}^{*}\mathbf{L}^{*\prime}+\boldsymbol{\Psi}$$

➢ 新的因子$\boldsymbol{f}^{*}$仍然满足：

$$E(\mathbf{f}^{*})=\mathbf{0},\ COV(\mathbf{f}^{*})=\mathbf{I}\text{和}\ COV(\mathbf{f}^{*},\boldsymbol{\varepsilon})=\mathbf{O}$$

因子及其载荷矩阵并不唯一，可按照任意的正交矩阵$\boldsymbol{T}$提供的方向旋转。寻找使因子及载荷结构更简单、解释更清晰的旋转方向$\boldsymbol{T}$。

因子旋转直觉理解

$$
\begin{aligned}
Y_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + \ldots + l_{1m}F_m + \varepsilon_1 \\
Y_2 - \mu_2 &= l_{21}F_1 + l_{22}F_2 + \ldots + l_{2m}F_m + \varepsilon_2 \\
&\ldots \\
Y_p - \mu_p &= l_{p1}F_1 + l_{p2}F_2 + \ldots + l_{pm}F_m + \varepsilon_p
\end{aligned}
$$

- 载荷矩阵$\boldsymbol{L}$代表原始变量和公共因子之间的关系（协方差/相关系数）。
- 从几何的角度，$\hat{\boldsymbol{L}}$的第j行的载荷构成了原始变量Y_j在因子/载荷空间的坐标。

14

因子旋转直觉理解例子

一个12岁的女孩对她身边的7个人进行9分制评分。评分基于五个维度进行的，分别是“友好(kind)”、“聪明(intelligent)”、“快乐(happy)”、“受人喜爱(likeable)”和“公正(just)”：

Perception Data: Ratings on Five Adjectives for Seven People

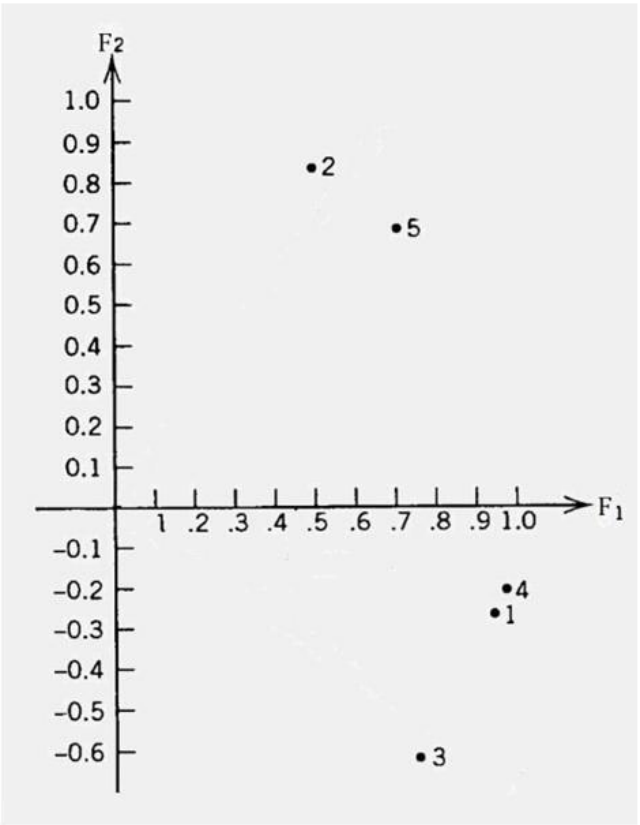
People	Kind	Intelligent	Happy	Likeable	Just
FSM1[a]	1	5	5	1	1
SISTER	8	9	7	9	8
FSM2	9	8	9	9	8
FATHER	9	9	9	9	9
TEACHER	1	9	1	1	9
MSM[b]	9	7	7	9	9
FSM3	9	7	9	9	7

[a]Female schoolmate 1.
[b]Male schoolmate.

15

因子旋转直觉理解例子

Variables	Loadings	
	$\hat{l}_{j1}$	$\hat{l}_{j2}$
1 Kind	.969	−.231
2 Intelligent	.519	.807
3 Happy	.785	−.587
4 Likeable	.971	−.210
5 Just	.704	.667



16

因子旋转直觉理解

$$
\begin{aligned}
Y_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + \ldots + l_{1m}F_m + \varepsilon_1 \\
Y_2 - \mu_2 &= l_{21}F_1 + l_{22}F_2 + \ldots + l_{2m}F_m + \varepsilon_2 \\
&\ldots \\
Y_p - \mu_p &= l_{p1}F_1 + l_{p2}F_2 + \ldots + l_{pm}F_m + \varepsilon_p
\end{aligned}
$$

- 希望得到每个原始变量都由某个因子主要决定（对应载荷数值很大），而与其他因子关系不大（对应载荷接近0）。
- 从几何的角度，因子旋转的目标是让坐标轴靠近尽可能多的点。

17

因子旋转的两种类型

- 正交旋转：原来垂直的坐标轴经过旋转后仍保持垂直，角度和距离都保持不变，共同度也不变，点的相对位置也维持原状；只有参考系改变了。
- 斜交旋转：不要求轴保持垂直，因此旋转更加自由，也更容易让轴靠近更多的点。

正交因子旋转（最大方差法）

$$
\begin{aligned}
Y_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + \ldots + l_{1m}F_m + \varepsilon_1 \\
Y_2 - \mu_2 &= l_{21}F_1 + l_{22}F_2 + \ldots + l_{2m}F_m + \varepsilon_2 \\
&\cdots \\
Y_p - \mu_p &= l_{p1}F_1 + l_{p2}F_2 + \ldots + l_{pm}F_m + \varepsilon_p
\end{aligned}
$$

- 最常用的为最大方差法，其目标是通过旋转因子载荷矩阵来使得每个因子在某些变量上具有较大的载荷（即该因子对这些变量的影响很大），而在其他变量上具有较小的载荷。
- 即最大化因子载荷矩阵列（即因子对每个变量的贡献）中方差的平方和，使得每个因子的载荷尽可能地集中在少数几个变量上，从而使得因子的解释更加清晰。

19

正交因子旋转（最大方差法）

Variables	Principal Component Loadings		Graphically Rotated Loadings		Varimax Rotated Loadings		Communalities
	F1	F2	F1	F2	F1	F2	$\hat{h}_i^2$
Kind	.969	−.231	**.927**	.367	**.951**	.298	.993
Intelligent	.519	.807	−.037	**.959**	.033	**.959**	.921
Happy	.785	−.587	**.980**	−.031	**.975**	−.103	.960
Likeable	.971	−.210	**.916**	.385	**.941**	.317	.987
Just	.704	.667	.194	**.950**	.263	**.933**	.940
Variance accounted for	3.263	1.538	2.696	2.106	2.811	1.991	4.802
Proportion of total variance	.653	.308	.539	.421	.562	.398	.960
Cumulative proportion	.653	.960	.539	.960	.562	.960	.960

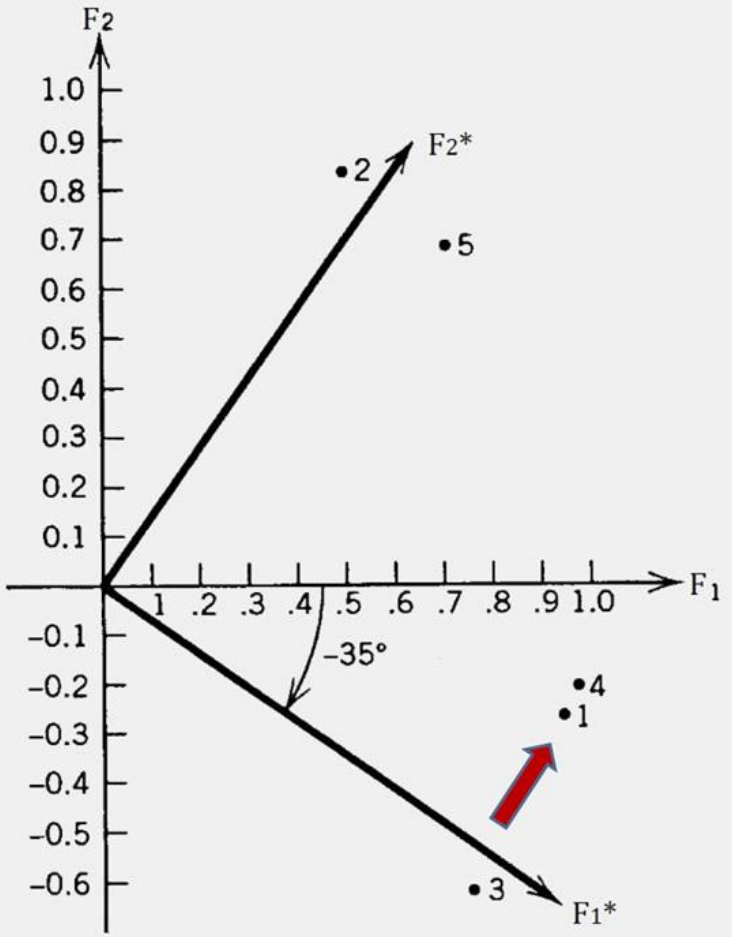
旋转后的载荷很容易解释：

- 第一个因子与变量kind, happy, likeable高度相关，可描述为亲和力；
- 第二个因子主要由intelligence和just构成，可解释为人的逻辑思维能力。

20

斜交因子旋转

- 回想12岁女孩的例子，如果旋转后的坐标轴允许不再垂直（即斜交旋转），代表F_1^*的轴可以更加靠近第1个和第4个变量对应的点。



斜交因子旋转

- 不像正交旋转中使用正交矩阵T，斜交旋转使用一个更一般的非奇异变换矩阵Q来得到$f^* = Q'f$，那么

$$COV(\mathbf{f}^*) = \mathbf{Q}'\mathbf{I}\mathbf{Q} = \mathbf{Q}'\mathbf{Q} \neq \mathbf{I}$$

因此新的因子之间是相关的，不是正交的。

- 由于距离和角度不再保持不变，f^*的共同度与f的也不同。

- 当不要求坐标轴相互垂直时，旋转后的坐标轴更容易“穿过”多数坐标点。

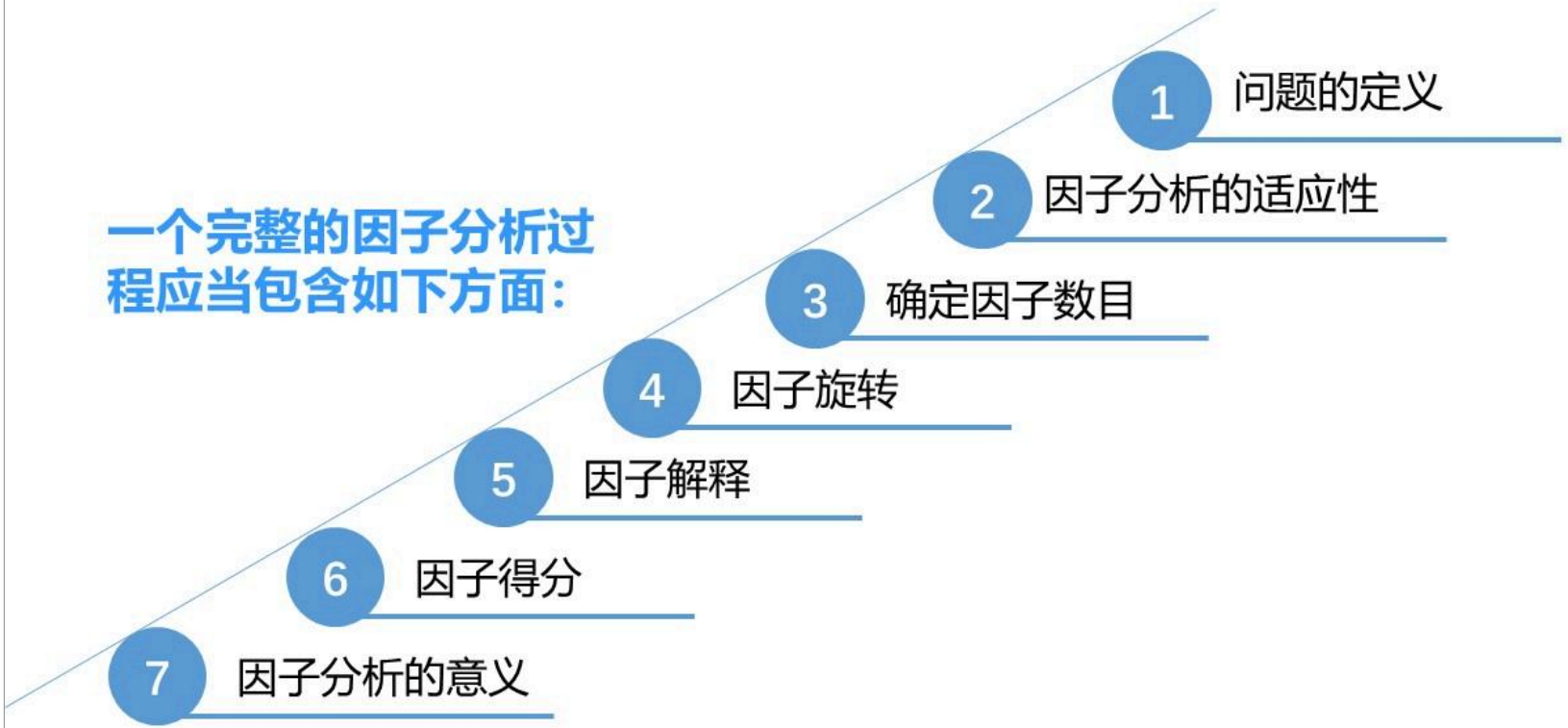
22

3. 因子分析的R实现

23

实际中如何进行因子分析

一个完整的因子分析过程应当包含如下方面：



24

案例

【例】下表记录了52位学生6门功课的考试分数的部分数据，依次表示数学、物理、化学、语文、历史、英语成绩：

```
library(readxl)
test <-
read_excel("test_score.csv")
View(test_score)
```

	Y1	Y2	Y3	Y4	Y5	Y6
1	65	61	72	84	81	79
2	77	77	76	64	70	55
3	67	63	49	65	67	57
4	78	84	75	62	71	64
5	66	71	67	52	65	57
6	83	100	79	41	67	50
7	86	94	97	51	63	55
8	67	84	53	58	66	56
9	69	56	67	75	94	80
10	77	90	80	68	66	60
11	84	67	75	60	70	63
12	62	67	83	71	85	77
13	91	74	97	62	71	66
14	82	70	83	68	77	85
15	66	61	77	62	73	64

相关性分析

```
# 样本相关系数矩阵
R<-round(cor(test), 3); R
```

```
> # 样本相关系数矩阵
> R<-round(cor(test), 3); R
       Y1     Y2     Y3     Y4     Y5     Y6
Y1  1.000  0.647  0.696 -0.561 -0.456 -0.439
Y2  0.647  1.000  0.573 -0.503 -0.351 -0.458
Y3  0.696  0.573  1.000 -0.380 -0.274 -0.244
Y4 -0.561 -0.503 -0.380  1.000  0.813  0.835
Y5 -0.456 -0.351 -0.274  0.813  1.000  0.819
Y6 -0.439 -0.458 -0.244  0.835  0.819  1.000
```

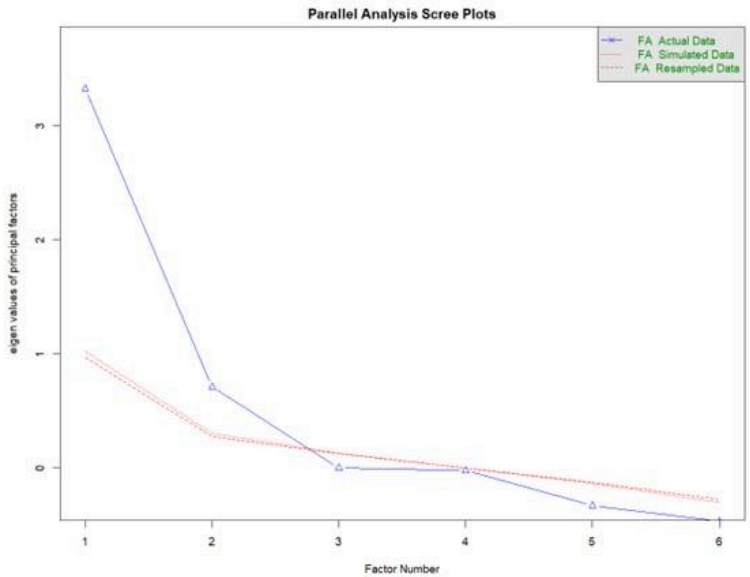
- 待估计的因子模型初步定为：

$$Y_j - \mu_j = l_{j1}F_1 + l_{j2}F_2 + \varepsilon_j,\ j = 1, \ldots, 6$$

决定因子个数

```
library(psych)
# 绘制碎石图确定因子数量
fa.parallel(test, fm = "ml", fa = "fa", n.iter = 100)
```

- fm：因子提取方法，可选择极大似然法（“ml”），主因子法（“pa”）。



执行因子分析

```
library(psych)
# 基于主因子法估计载荷
fa_pa <- fa(test, nfactors=2, fm = "pa",
     rotate="varimax")

# 基于极大似然法估计载荷
fa_ml <- fa(test, nfactors=2, fm = "ml",
     rotate="varimax")
```

- rotate：因子旋转方法，“none”：不旋转，“varimax”：正交旋转（最大方差旋转），“oblimin”：斜交旋转方法。

28

查看因子分析结果

```
> print(fa_pa)
Factor Analysis using method =  pa
Call: fa(r = test, nfactors = 2, rotate = "varimax", fm = "pa")
Standardized loadings (pattern matrix) based upon correlation matrix
     PA1   PA2   h2   u2 com
Y1 -0.32  0.82 0.77 0.23 1.3
Y2 -0.31  0.67 0.54 0.46 1.4
Y3 -0.11  0.81 0.66 0.34 1.0
Y4  0.85 -0.36 0.85 0.15 1.3
Y5  0.86 -0.20 0.78 0.22 1.1
Y6  0.90 -0.20 0.86 0.14 1.1

                       PA1  PA2
SS loadings           2.49 1.98
Proportion Var        0.41 0.33
Cumulative Var        0.41 0.74
Proportion Explained  0.56 0.44
Cumulative Proportion 0.56 1.00

Mean item complexity =  1.2
Test of the hypothesis that 2 factors are sufficient.
```

- 因子载荷：PA1，PA2；
- 公共因子贡献的方差：h2，u2；
- 个体方差：com；
- 每个因子所解释的方差（特征值）：SS loadings；
- 每个因子所解释的方差比例：Proportion Var；
- 累积方差比例：Cumulative Var。

29

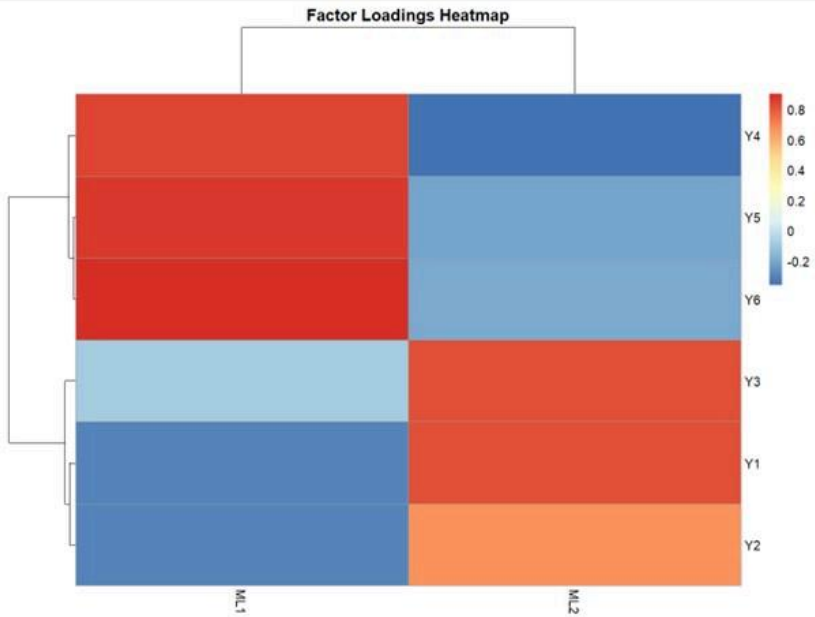
查看因子分析结果

```
> print(fa_ml)
Factor Analysis using method =  ml
Call: fa(r = test, nfactors = 2, rotate = "varimax", fm = "ml")
Standardized loadings (pattern matrix) based upon correlation matrix
     ML1   ML2   h2   u2 com
Y1 -0.31  0.82 0.77 0.23 1.3
Y2 -0.31  0.67 0.54 0.46 1.4
Y3 -0.10  0.81 0.67 0.33 1.0
Y4  0.85 -0.36 0.85 0.15 1.4
Y5  0.86 -0.22 0.79 0.21 1.1
Y6  0.90 -0.21 0.85 0.15 1.1

                       ML1  ML2
SS loadings           2.47 2.00
Proportion Var        0.41 0.33
Cumulative Var        0.41 0.75
Proportion Explained  0.55 0.45
Cumulative Proportion 0.55 1.00

Mean item complexity =  1.2
Test of the hypothesis that 2 factors are sufficient.
```

30

因子载荷的热图

```
library(pheatmap)
# 绘制热图
pheatmap(fa_ml$loadings, main = "Factor Loadings Heatmap")
```

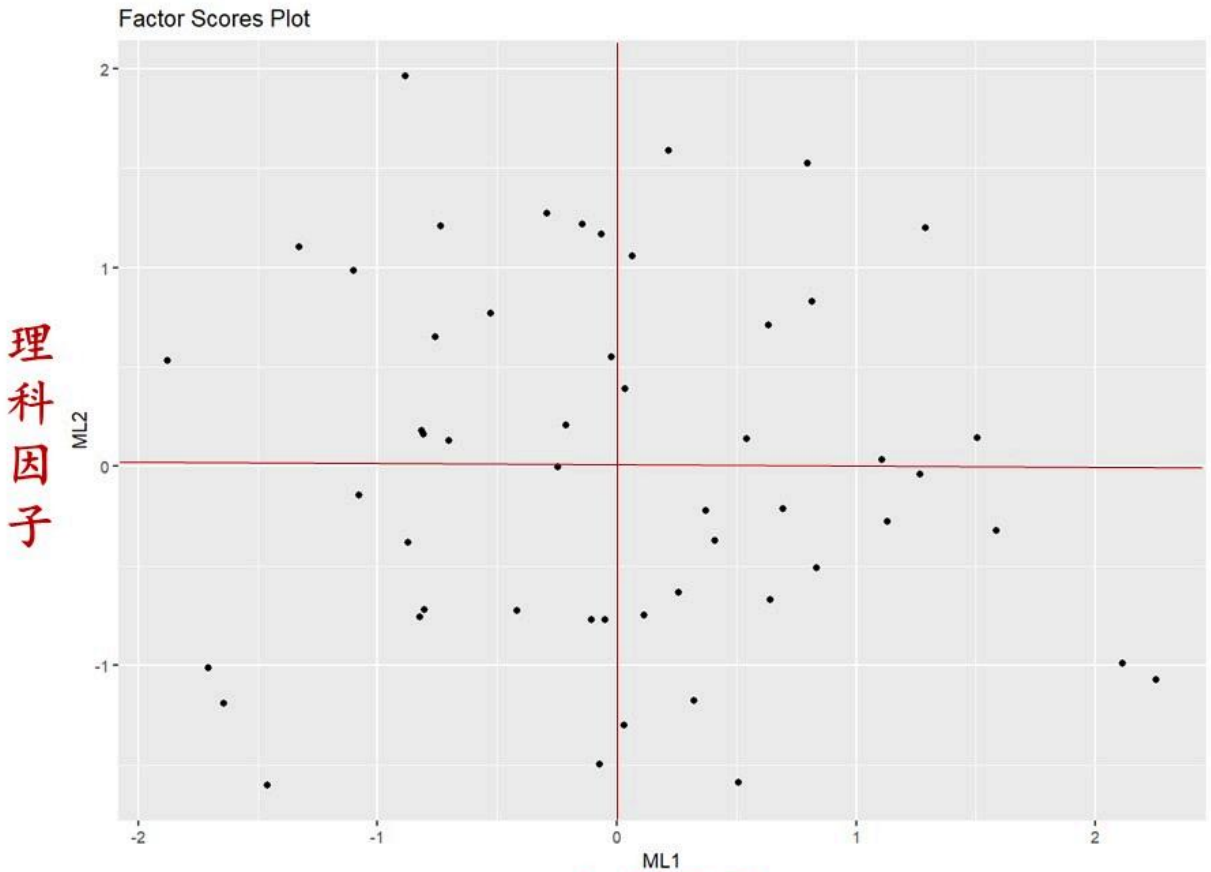


ML1：文科因子；
ML2：理科因子。

因子得分

```
# 获取因子得分
fa_scores <- fa_ml$scores

# 可视化因子得分，使用 ggplot2 绘制散点图
library(ggplot2)
ggplot(as.data.frame(fa_scores), aes(x = ML1, y = ML2)) +
     geom_point() +
     labs(title = "Factor Scores Plot")
```

32



33

单选题 1分

在psych::fa函数中，若要指定一个特定的因子提取方法（如极大似然法），应该使用哪个参数?

A fm = "ml"

B extract = "ml"

C method = "pa"

D fm = “pa"

34

单选题 1分

在psych::fa函数中，参数rotate = "none"表示:

A 进行正交旋转

B 进行斜交旋转

C 不进行旋转

D 提取主成分

35

本章小结

- 因子分析简介：因子分析旨在提出因子模型来研究如何用几个公共因子来刻画原始变量之间的相关性。
- 因子旋转：正交旋转、斜交旋转。
- 因子分析的R实现：psych::fa。