1 因子分析

因子分析的 R 函数为 factanal (极大似然法), 以及 princomp/prcomp (主成分方法或主因子方法)。我们仅考虑前者。主要函数包括因子分析函数 factanal，因子旋转函数 varimax，promax。调用方式如下

```
##总体因子分析（基于协方差矩阵）
 > myFA = factanal(covmat=my.cov, factors=2,n.obs=..)
  #covmat: covariance/correlation matrix
  #factors: number of factors
  #n.obs: 若给出样本量，则基于协方差/相关系数矩阵可以计算拟合优度检验。
  #若不给n.obs，则不能计算拟合优度检验（如例1）

##样本因子分析（基于原始数据）
 > myFA = factanal(x=my.data, factors=2, scores=...) # use data.frame
 #scores: 因子的预测
 > varimax(L)
  #L: loadings obtained from factor analysis (princomp or factanal)
 > promax(L) #non-orthogonal rotation
```

下面两个例子分别演示总体因子分析（即只有协方差或相关系数矩阵的情形）和样本因子分析（即有原始数据的情形）.

例 **1.** (总体因子分析，课本例 9.3) 一项问卷调查中，消费者对一种新的食品产品的 5 个属性（"Taste", "Good buy", "Flavor", "Suitabl for snack", "Lots of energy"）进行了评价打分（1-7，1 分最低，7 分最高），5 个属性的样本相关系数如下：

Attribute (Variable)		1	2	3	4	5
Taste	1	1.00	.02	.96	.42	.01
Good buy for money	2	.02	1.00	.13	.71	.85
Flavor	3	.96	.13	1.00	.50	.11
Suitable for snack	4	.42	.71	.50	1.00	.79
Provides lots of energy	5	.01	.85	.11	.79	1.00

下面对该相关系数矩阵进行因子分析（因子个数取为 2）。两因子模型将 5 个变量的打分 $\mathbf{x} = (x_1, x_2, ..., x_5)^\top$ 表示为

$$\mathbf{x} = \boldsymbol{\mu} + \mathbf{LF} + \boldsymbol{\epsilon} = \boldsymbol{\mu} + \mathbf{l}_1 F_1 + \mathbf{l}_2 F_2 + \boldsymbol{\epsilon},$$

其中 $\boldsymbol{\mu} = E(\mathbf{x})$, $\mathbf{F} = (F_1, F_2)^\top \sim N_2(\mathbf{0}, I_2)$ 代表两个独立的因子，$\mathbf{l}_1, \mathbf{l}_2$ 分别是 F_1, F_2 的载荷,$L = (\mathbf{l}_1, \mathbf{l}_2)$ 为 5×2 载荷矩阵, $\boldsymbol{\epsilon} \sim N_5(\mathbf{0}, \Psi)$ 为特殊因子 (误差), $\Psi = diag(\psi_1, ..., \psi_5)$ 对角元为 5 个特殊方差。需要估计的参数为 $\boldsymbol{\mu}$, L, Ψ, 但因这里我们仅有相关系数矩阵，所以 $\boldsymbol{\mu}$ 不能估计，这并不影响 L,Ψ 的估计以及因子模型的解释。

数据分析结果如下：

```
> R=matrix(c(1,0.02,0.96,0.42,0.01,0.02,1,0.13,0.71,0.85,0.96,0.13,
       1,0.5,0.11,0.42,0.71,0.5,1,0.79,0.01,0.85,0.11,0.79,1),5,5)
> name=c("Taste","Good.buy","Flavor","Suitable.for.snack","Lots.of.energy")
> dimnames(R)=list(name,name)

> (myFA = factanal(covmat=R, factors=2)) # 两因子的因子分析模型
 # 本例我们不知道样本量，故不指定 n.obs，因而分析结果不给出拟合优度检验。
       Call:
       factanal(factors = 2, covmat = R)

       Uniquenesses:
       Taste    Good.buy    Flavor Suitable.for.snack   Lots.of.energy
       0.028       0.237     0.040              0.168            0.052

       Loadings:
                          Factor1  Factor2
       Taste                        0.985
       Good.buy            0.873
       Flavor              0.131    0.971
       Suitable.for.snack  0.817    0.405
       Lots.of.energy      0.973

                      Factor1 Factor2
       SS loadings      2.396   2.078
       Proportion Var   0.479   0.416
       Cumulative Var   0.479   0.895

       The degrees of freedom for the model is 1 and the fit was 0.0233
```

输出结果主要是 Ψ, L 的估计，以及由载荷计算得到的方差贡献率，细节如下：

- Uniqueness（特殊方差）Ψ

  5 个变量的个特殊方差（uniqueness, specific variance）分别为 0.028, 0.237, 0.040, 0.168, 0.052, 其中 “good buy” 和”Suitable for snack” 的 uniquenss 值远大于另外三个变量，这说明两个公共因子不能很好地解释这两个变量。从另外一个角度讲，“Good buy” 和 “ Suitable for snack” 两个变量作为问卷问题让消费者打分可能不是十分恰当，这两个问题以及 “Lots of energy” 的目的可能是希望得到消费者对新产品的 “实用性” 方面的感受，但显然它们在测量 “ 实用性” 的方面不如 “Lots of energy” 精准。从这个角度看，该问卷设计中的问题有改进余地。为了更好地拟合得到更小的 uniqueness, 我们可以增加因子个数（注意载荷旋转的目的是得到更好解释的因子，并不改变 uniqueness），也可尝试用主因子方法。

- 载荷 L 和因子的解释

  拟合得到的两因子模型为：

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \end{pmatrix} + \begin{pmatrix} 0 \\ 0.873 \\ 0.131 \\ 0.817 \\ 0.973 \end{pmatrix} F_1 + \begin{pmatrix} 0.985 \\ 0 \\ 0.971 \\ 0.405 \\ 0 \end{pmatrix} F_2 + error$$

  其中变量 2,4,5 即 Goog buy, Suitable for snack, Lots of energy 在 F_1 上的载荷值分别是 0.873, 0.817, 0.973, 而 Taste 和 Flavor 的载荷接近于 0 (分别为 0, $0.131 \approx 0$)，即 F_1 与 Taste，flavor 不

相关，与其他变量正相关。第一因子载荷说明 Goog buy, Suitable for snack, Lots of energy, 甚至包括 Flavor 可能是同一类变量，F_1 的含义应该是这三或四个变量的某种共同含义，我们可称之为“实用性”，因为 Lots of energy 的载荷最大 (0.973), 这说明该变量是消费者在实用性考量上最为看重的一个方面，其次是 Good buy(花钱合算) 和“是不是适合当作小吃”, 而 Flavor 在实用性考虑中最不重要 (0.131).

第二个因子 F_2 在 Taste, Flavor, Suitable for snack 上的载荷分别是 0.985, 0.971, 0.405 这三个变量都与口味有关，这说明 F_2 可能是口味因子。载荷的取值 Taste 最大，Flavor 次之，这说明对于口味的考量方面，Taste 最重要，Flavor 也很重要，而 Suitable for snack 显得不那么重要。

- 共性方差、方差贡献

  Communality(共性方差)：输出结果中没有列出各变量的共性方差 $h_i^2 = 1 - \text{uniquess}_i, i = 1,...,5$。由所给的特殊方差可计算得到：$(h_1^2, h_2^2, h_3^2, h_4^2, h_5^2) = (0.972, 0.763, 0.960, 0.832, 0.948)$。两个因子能解释变量 1，3，5 的方差的 90% 以上。对变量 2，4 的解释能力稍差，分别是 76.3%,83.2%.

  另外，总方差为 $tr(R) = 5$。输出结果的倒数第 4 行给出了因子载荷的第 1，2 列的平方和（SS）分别是 2.396，2.078，所以它们解释总方差的比率分别为 2.396/5=0.479, 2.078/5=0.416（输出结果的倒数第 3 行），两个因子的累积方差解释比例为 0.895（倒数第 2 行）。

- 因子旋转/载荷旋转 （注意载荷旋转是为了得到更好的解释，并不能改进拟合效果）

  函数 factanal 的缺省旋转方法为 rotation = ”varimax”(旋转后使得 varimax 最大)，所以上述结果中的载荷是载荷的原始极大似然估计通过正交旋转得到的。下面尝试 promax 旋转，得到如下载荷：

```
> promax(loadings(myFA))
                Loadings:
                                Factor1 Factor2
                Taste                     1.004
                Good_buy           0.892 -0.103
                Flavor                    0.975
                Suitable_for_snack 0.786  0.313
                Lots_of_energy     0.997 -0.137
```

  与 varimax 旋转的结果并没有多大差异。

例 **2.** （样本因子分析）数据集 geopol 给出了 41 个国家和地区的 10 项指标/变量数据，描述如下

变量	描述
pop	population
GDP	Gross Internal Product per habitant
pop.inc	rate of increase of the population
urban	rate of urban population
illiter	rate of illiteracy in the population
student	rate of students in the population
life	expected lifetime of people
nutri	rate of nutritional needs realized
magz	number of newspapers and magazines per 1000 habitants
tv	number of television per 1000 habitants

41 个国家和地区的名称缩写如下：

AFS	South Africa	DAN	Denmark	MAR	Marocco
ALG	Algeria	EGY	Egypt	MEX	Mexico
BRD	Germany	ESP	Spain	NOR	Norway
GBR	Great Britain	FRA	France	PER	Peru
ARS	Saudi Arabia	GAB	Gabun	POL	Poland
ARG	Argentine	GRE	Greece	POR	Portugal
AUS	Australia	HOK	Hong Kong	SUE	Sweden
AUT	Austria	HON	Hungary	SUI	Switzerland
BEL	Belgium	IND	India	THA	Tailand
CAM	Cameroon	IDO	Indonesia	URS	USSR
CAN	Canada	ISR	Israel	USA	USA
CHL	Chile	ITA	Italia	VEN	Venezuela
CHN	China	JAP	Japan	YOU	Yugoslavia
CUB	Cuba	KEN	Kenia		

你可以从原始数据 geopol 计算得到 10 个指标的协方差矩阵或相关系数矩阵，然后应用总体因子分析（并指定样本量 n.obs），也可基于原始数据直接进行因子分析，两者结果相同。

```
geopol= read.table("http://staff.ustc.edu.cn/~ynyang/vector/data/geopol.txt",head=T)
R=cor(geopol)
#R=cov(geopol)
myfa =factanal(covmat=R, factors=2, n.obs=41) # 总体因子分析
myfa2=factanal(x = geopol, factors = 2) # 样本因子分析

> myfa2
Call:
factanal(x = geopol, factors = 2)

Uniquenesses:
pop     GDP pop.inc   urban illiter student    life   nutri    magz      tv
0.956   0.005   0.243   0.589   0.169   0.530   0.111   0.525   0.325   0.215

Loadings:
        Factor1 Factor2
pop     -0.122  -0.170
GDP      0.276   0.959
pop.inc -0.770  -0.406
urban    0.493   0.409
illiter -0.829  -0.379
student  0.466   0.503
life     0.807   0.488
nutri    0.649   0.231
magz     0.391   0.722
tv       0.483   0.743

                Factor1 Factor2
SS loadings      3.289   3.041
Proportion Var   0.329   0.304
Cumulative Var   0.329   0.633

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 45.9 on 26 degrees of freedom.
The p-value is 0.00937
```

输出结果最后三行计算拟合优度统计量（chi-square statistic）

$$W = n_B\left\{\log(|LL^\top + \Psi|/|R|)\right\},$$

其中 n_B 是 Bartlett 校正的样本量 $n_B = (n - 1 - (2p + 4m + 5)/6) \approx n$，及其 p 值

$$pvalue = P(\chi^2_{[(p-m)^2-p-m]/2} \geq W)$$

p 值越大，拟合效果越好。本例的拟合优度 p 值较小，说明拟合不太好。拟合 3 因子模型：

```
myfa3=factanal(x = geopol, factors = 3) # 样本因子分析
Call:
factanal(x = geopol, factors = 3)

Uniquenesses:
pop     GDP pop.inc   urban illiter student    life   nutri    magz      tv
0.819   0.005   0.005   0.482   0.174   0.351   0.141   0.489   0.311   0.190

Loadings:
       Factor1 Factor2 Factor3
pop                    -0.420
GDP      0.271   0.908   0.311
pop.inc -0.910  -0.406
urban    0.383   0.282   0.541
illiter -0.755  -0.282  -0.419
student  0.344   0.349   0.639
life     0.753   0.418   0.343
nutri    0.664   0.174   0.201
magz     0.413   0.698   0.175
tv       0.444   0.646   0.442

               Factor1 Factor2 Factor3
SS loadings      3.112   2.385   1.537
Proportion Var   0.311   0.239   0.154
Cumulative Var   0.311   0.550   0.703

Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 24.94 on 18 degrees of freedom.
The p-value is 0.127
```

拟合有较大改进，p 值 0.127。因子 1 和 2 的载荷类似, 人口增长率 pop.inc 和文盲率 illiter 的载荷都是负数，因此因子 1，2 都和这两个指标负相关，大概都代表了国家发达程度（一个代表物质发达程度，一个代表文明发达程度）。我们可画出载荷散点图进行考察, 并结合每个国家的因子得分以及国家背景知识了解两个因子的含义。以两因子模型为例，因子载荷和因子得分散点图：

```
loadings (myfa2)->L
plot(L,type="n")
text(L,  rownames(L))

# 因子得分 (score):
myfa2 = factanal(x=geopol, factors=2,scores ="Bartlett" )  #
s = myfa2$score
```

```
plot (s,type="n")
text(s, rownames(geopol))
```

练习 1. 下面给出了 112 个孩子的 6 项测试（general: 综合, picture: 绘画,blocks: 积木,maze: 迷宫,reading: 阅读,vocab: 词汇量）成绩的协方差矩阵

	general	picture	blocks	maze	reading	vocab
general	24.641	5.991	33.520	6.023	20.755	29.701
picture	5.991	6.700	18.137	1.782	4.936	7.204
blocks	33.520	18.137	149.831	19.424	31.430	50.753
maze	6.023	1.782	19.424	12.711	4.757	9.075
reading	20.755	4.936	31.430	4.757	52.604	66.762
vocab	29.701	7.204	50.753	9.075	66.762	135.292

试对该数据做因子分析，两因子模型是否能很好地拟合数据? （提示：首先标准化，把该矩阵转换为相关系数矩阵）。

练习 2. （课本 Table 9.12）一家公司对其销售员工的销售业绩进行了评估，同时还希望能找到一种或几种能揭示员工销售潜力的测试或考试。公司随机抽取了 50 名销售人员，并对他们的三种销售表现做了评估：销售额增长速度，销售获利成绩，对新客户的销售能力。这三种表现的度量都以 100 作为平均业绩，分数越高表现越好。另外，这 50 人接受了 4 项测试，分别用于测试他们的创造力，物理推理能力 (mechanical reasoning)，抽象推理能力，和数学能力。

```
>sales= read.table("http://staff.ustc.edu.cn/~ynyang/vector/data/T9-12.DAT")
>dimnames(sales)[[2]]=c("Sale.growth", "Sales. profitabiity", "NewAccount.sales",
 "Creativity", "Manchanicl.reasoning.test", "Abstract.reasoning.test", "Math.test")
>sales
```

1. 首先将数据标准化（标准化函数 scale），求 $m = 2$ 和 $m = 3$ 个公共因子的因子分析的极大似然解。
2. 由 (a) 的解，求 $m = 2$ 和 $m = 3$ 的旋转载荷，比较两组旋转载荷，解释因子的含义。
3. 对 $m = 2, 3$, 给出公共方差、个性方差的估计值, 以及残差矩阵 $R - (LL^\top + \Psi)$，你倾向于选择 $m = 2$ 还是 $m = 3$?
4. 计算每个人的因子得分 (factor score), 并画出因子得分的散点图（因子得分计算：函数 factanal 中指定 scores="regression" 或 "Bartlett"）。

提示: 公司的主要目的是找出与销售业绩有关的测试。

2 结构方程模型（sem）

R 程序包 lavvan 和 sem 的提供了结构方程模型的分析工具。lavaan 语法简洁，但不提供路径图的画法；sem 语法较为复杂，但提供了路径图的绘制函数。

例 **3.** 我们以第十五讲例 1 的确认性因子分析 (CFA) 为例，学习使用 R 程序包 lavvan 和 sem 的用法。该数据是 220 个学生 6 门课程成绩的的相关系数矩阵，如下：

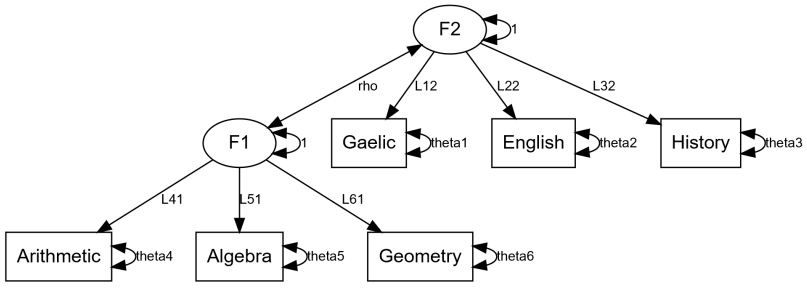
$$\mathbf{R} = \begin{array}{c} \begin{array}{cccccc} \text{Gaelic} & \text{English} & \text{History} & \text{Arithmetic} & \text{Algebra} & \text{Geometry} \end{array} \\ \begin{bmatrix} 1.0 & .439 & .410 & .288 & .329 & .248 \\ & 1.0 & .351 & .354 & .320 & .329 \\ & & 1.0 & .164 & .190 & .181 \\ & & & 1.0 & .595 & .470 \\ & & & & 1.0 & .464 \\ & & & & & 1.0 \end{bmatrix} \end{array}$$

```
#相关系数矩阵 r:
r=matrix(0,6,6)
r[lower.tri(r)]=c(0.439,0.410,0.288,0.329,0.248,
0.351, 0.354,0.320,0.329, 0.164,0.190,0.181,0.595,0.470,0.464)
r= r+t(r)
diag(r)=1
rownames(r)=colnames(r)=c("Gaelic","English","History",
 "Arithmetic","Algebra","Geometry")
```

假设如下结构方程模型（确认性因子分析模型）：

$$\begin{aligned} Gaelic &= L_{12}F_2 + \epsilon_1 \\ English &= L_{22}F_2 + \epsilon_2 \\ History &= L_{32}F_2 + \epsilon_3 \qquad \text{其中} \epsilon_i, i = 1,...,6 \text{独立}, var(\epsilon_i) = \theta_i \\ Arithmetic &= L_{41}F_1 + \epsilon_4 \qquad \epsilon's \text{与} (F_1, F_2) \text{独立} \\ Algebra &= L_{51}F_1 + \epsilon_5 \qquad var(F_1) = var(F_2) = 1, cov(F_1, F_2) = \rho \\ Geometry &= L_{61}F_1 + \epsilon_6 \end{aligned}$$

该模型以如下路径图表示 (绘图工具参见最后一页)：



2.1 lavaan 程序包

在 lavaan 中, 路径图模型结构以两个单引号 之间的一串命令指定，其中因子与显变量的关系以 =~ 指定，相关关系以 ~~ 指定，回归关系以 ~ 指定 (参见下面代码中的 mymodel='...')。结构方差模型的主要函数是 sem()

```
### MODEL SPECIFICATION:
> mymodel= '
   F=~variable1+variable2+... ##definition of factors
   F1~~F2 ## F1 and F2 are correlated
   F1~F2 # F1=a+b*F2+error ##regression
  '
 ### FIT THE SEM MODEL SPECIFIED BY mymodel
> sem(model = mymodel, data = , sample.cov = , sample.nobs =, ...)
 # model:model structure, data:original data,
 # sample.cov:covariance matrix or correlation matrix,
 # sample.nobs: sample size n
```

使用 lavann 程序包分析例 3 数据的代码如下：

```
## 例 3
> library(lavaan)
# 指定模型:
> model = '
 # latent variable definitions
  F1 =~ Arithmetic+Algebra+Geometry # “=~”read as “is manifested by”
  F2 =~ Gaelic+English+History
 # residual correlations （方差，协方差）
  F1~~F2  # F1 and F2 are correlated
        '
 > fit <- sem(model=model, sample.cov=r, sample.nobs=220, std.lv=TRUE )
#sample.cov 指定协方差矩阵或相关系数矩阵，同时需要指定 sample.nobs 样本量
# 也可以用 data = .. 指定原始数据
#std.lv=T 因子的方差假设为 1
> summary(fit)
```

练习 3. 例 3 中假设 F_1, F_2 是相关的, 试修改上述代码中 model 指定的最后一行的相关性假设修改为回归关系：$F_1 = \rho F_2 + error$ ，并运行 sem 分析.

附录：**sem** 程序包（**sem** 的句法过于复杂，可忽略）

sem 程序包同样需要先指定模型结构，函数为 specifyModel()，结构方程分析的主要函数和 lavaan 一样也是 sem()（注意：在同一个 R 环境中如果同时加载 lavaan 和 sem 程序包，主函数 sem() 会相互冲突）。路径图的绘制函数为 pathDiagram()。细节如下：

```
### MODEL SPECIFICATION:
> mod=c("f1 -> variable1, parameter, initial.value",...)
> mymodel= specifyModel(text=mod)
> mysem = sem(model=mymodel, S=r, N= 220)
       #S: covariance or correlation matrix, N: sample size
> pathDiagram(mysem) # draw the path diagram
```

在 sem 程序包中，上页的路径图所代表的模型如下指定：

```
library(sem)   # 载入 sem
mod <- c(  "F2   -> Gaelic, L12, NA",
        # 表示 Gaelic = L12*F2+error, 最后一个是 L12 的初值，这里 NA 表示不设初值
        "F2   -> English, L22, NA",
        "F2   -> History, L32, NA",
        "F1 -> Arithmetic, L41, NA",
        "F1 -> Algebra, L51, NA",
        "F1 -> Geometry, L61, NA",
        "F2   <-> F1, rho, NA", #CFA
        "F1   <-> F1, NA, 1",  # 表示 F1 的方差已知，为 1
        "F2   <-> F2, NA, 1",
        "Gaelic     <-> Gaelic, theta1, NA", #Gaelic 的方差为 theta1, 不赋初值 (NA)
        "English    <-> English, theta2, NA",
        "History    <-> History, theta3, NA",
        "Arithmetic <-> Arithmetic, theta4, NA",
        "Algebra    <-> Algebra, theta5, NA",
        "Geometry   <-> Geometry, theta6, NA")
```

其中每一项描述路径图中的一个单向箭头或双向箭头及其相关的参数, 包含逗号分隔的 3 各部分，分别表示箭头、参数和参数初值：

- 箭头：单箭头 $A \to B$ 表示回归方程 $B = \beta A + error$。双箭头 $A \leftrightarrow B$ 表示 A,B 相关。箭头后面第一个位置为变量 A，第二个位置为参数 β，第三个位置为参数 β 的初值（NA 表示不赋初值）
- 参数：当箭头是单向时，参数为载荷/回归系数 (β), 系数/载荷 β 会标在箭头附近；当箭头是双向时，参数为协方差（$A \leftrightarrow B$）或方差 ($A \leftrightarrow A$), 协方差或方差会标记在双箭头附近。

使用 sem 程序包画出例 3 的路径图，代码如下：

```
# 描述/指定模型:
mymodel =specifyModel(text=mod)
# 运行 sem
mysem <- sem(model=mymodel, S=r, N= 220)
# 画路径图:
pathDiagram(mysem,ignore.double=FALSE,edge.labels="both",rank.direction="TB")
#ignore.double: 是否不画双向箭头，箭头标号 edge.labels 可选"name","value","both";
#rank.direction 可选"TB" (从上到下）或"LR"（从左到右）
```