

# 因子分析 FACTOR ANALYSIS

数学与统计学院 杨炜明



#### 1 引言

因子分析(factor analysis)是一种数据简化 的技术。它通过研究众多变量之间的内部依赖关系, 探求观测数据中的基本结构,并用少数几个假想变 量来表示其基本的数据结构。这几个假想变量能够 反映原来众多变量的主要信息。原始的变量是可观 测的显在变量,而假想变量是不可观测的潜在变量, 称为因子。



## 注:

主成分分析分析与因子分析也有不同,主成分分析仅仅是变量变换,而因子分析需要构造因子模型。

主成分分析:原始变量的线性组合表示新的综合变量,即主成分;

因子分析:潜在的假想变量和随机影响变量的线性组合表示原始变量。



#### 2 因子分析模型

#### 一、数学模型

设  $X_i(i=1,2,\cdots,p)$  p 个变量,如果表示为

$$X_i = \mu_i + l_{i1}F_1 + \dots + l_{im}F_m + \varepsilon_i \qquad (m \le p)$$

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

或
$$\mathbf{X} = \mathbf{\mu} + L\mathbf{F} + \varepsilon$$



称 $F_1, F_2, \dots, F_m$ 为公共因子,是不可观测的变量,他们的系数称为因子载荷。 $\varepsilon_i$ 是特殊因子,是不能被前m个公共因子包含的部分,并且满足:

 $cov(F,\varepsilon) = 0$ ,  $F,\varepsilon$  不相关;

$$D(F) = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & \ddots & & \\ & & 1 \end{bmatrix} = I$$

即  $F_1, F_2, \dots, F_m$  互不相关,方差为1。



$$D(arepsilon) = egin{bmatrix} oldsymbol{\psi}_1 & & & & & \ & oldsymbol{\psi}_2 & & & & \ & & \ddots & & \ & & oldsymbol{\psi}_p \end{bmatrix}$$

即互不相关,方差不一定相等。



# 用矩阵的表达方式

$$X - \mu = LF + \varepsilon$$

$$E(F) = 0$$

$$E(\varepsilon) = 0$$

$$Var(F) = I$$

$$Var(\mathbf{\varepsilon}) = diag(\psi_1, \psi_2, \dots, \psi_p)$$

$$cov(F, \varepsilon) = E(F\varepsilon') = 0$$



#### 二、因子载荷矩阵中的几个统计特征

#### 1、因子载荷lii的统计意义

因子载荷 lii 是第i个变量与第j个公共因子的相关系数

假定变量已经标准化,模型为  $X_i = l_{i1}F_1 + \cdots + l_{im}F_m + \varepsilon_i$ 

在上式的左右两边乘以 $F_j$ ,再求数学期望

$$E(X_iF_j) = l_{i1}E(F_1F_j) + \dots + l_{ij}E(F_jF_j) + \dots + l_{im}E(F_mF_j) + E(\varepsilon_iF_j)$$

根据公共因子的模型性质,有

 $\gamma_{x_iF_j} = l_{ij}$ (载荷矩阵中第i行,第j列的元素)反映了第i个变量与第j个公共因子的相关重要性。绝对值越大,相关的密切程度越高。



#### 2、变量共同度的统计意义

**定义:** 变量  $X_i$  的共同度是因子载荷矩阵的第i行的元素的平方和。记为  $h_i^2 = \sum_{i=1}^m l_{ij}^2$ 。

#### 统计意义:

$$X_i = l_{i1}F_1 + \dots + l_{im}F_m + \varepsilon_i$$
 两边求方差
$$Var(X_i) = l_{i1}^2 Var(F_1) + \dots + l_{im}^2 Var(F_m) + Var(\varepsilon_i)$$

$$1 = \sum_{j=1}^{m} l_{ij}^2 + \sigma_i^2$$

所有的公共因子和特殊因子对变量  $X_i$  的贡献为1。如果  $\sum_{i,j}^m l_{i,j}^2$  非常 靠近1, $\psi_i$  非常小,则因子分析的效果好,从原变量空间,接度之高大学子空间的转化性质好。

## 3、公共因子<math>F方差贡献的统计意义

因子载荷矩阵中各列元素的平方和

$$S_j = \sum_{i=1}^p l_{ij}^2$$

称为相应公共因子 $F_j$ 对总方差的贡献。衡量 $F_j$ 的相对重要性。



## 3 因子载荷矩阵的估计方法(主成分分析法)

设随机向量  $x = (x_1, x_2, \dots, x_p)$  的均值为 $\mu$ , 协方差为 $\Sigma$ , 为 $\Sigma$ 的特征根, 为对应的

标准化特征户量。则

$$u_1, u_2, \cdots, u_p$$

$$\Sigma = LL' + \Psi = U egin{bmatrix} \lambda_1 & & & \ & \lambda_2 & & \ & \ddots & \ & & \lambda_p \end{bmatrix} U'$$



$$\begin{bmatrix} \mathbf{u}_{1} & \mathbf{u}_{2} & \cdots & \mathbf{u}_{p} \end{bmatrix} \begin{pmatrix} \lambda_{1} & 0 \\ \ddots & \lambda_{p} \end{pmatrix} \begin{bmatrix} \mathbf{u}_{1}' \\ \mathbf{u}_{2}' \\ \vdots \\ \mathbf{u}_{p}' \end{bmatrix}$$

$$= \lambda_{1} \mathbf{u}_{1} \mathbf{u}_{1}' + \lambda_{2} \mathbf{u}_{2} \mathbf{u}_{2}' + \cdots + \lambda_{p} \mathbf{u}_{p} \mathbf{u}_{p}'$$

$$= \begin{bmatrix} \sqrt{\lambda_{1}} \mathbf{u}_{1} & \sqrt{\lambda_{2}} \mathbf{u}_{2}' \\ \vdots \\ \sqrt{\lambda_{p}} \mathbf{u}_{p}' \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_{1}} \mathbf{u}_{1}' \\ \sqrt{\lambda_{2}} \mathbf{u}_{2}' \\ \vdots \\ \sqrt{\lambda_{p}} \mathbf{u}_{p}' \end{bmatrix}$$

上式给出的Σ表达式是精确的,然而,它实际上是毫无价值的,因为我们的。 目的是寻求用少数几个公共因子解释,故略去后面的p-m + 1项的员员,有多人多

$$\Sigma \approx \hat{\mathbf{A}}\hat{\mathbf{A}}' + \hat{\mathbf{D}} = \lambda_1 \mathbf{u}_1 \mathbf{u}_1' + \lambda_2 \mathbf{u}_2 \mathbf{u}_2' + \dots + \lambda_m \mathbf{u}_m \mathbf{u}_m' + \hat{\mathbf{D}}$$

$$= \left[ \sqrt{\lambda_1} \mathbf{u}_1 \quad \sqrt{\lambda_2} \mathbf{u}_2 \quad \cdots \quad \sqrt{\lambda_m} \mathbf{u}_m \right] \begin{bmatrix} \sqrt{\lambda_1} \mathbf{u}_1' \\ \sqrt{\lambda_2} \mathbf{u}_2' \\ \vdots \\ \sqrt{\lambda_p} \mathbf{u}_m' \end{bmatrix} + \hat{\Psi} \approx \hat{L} \hat{L}' + \hat{\Psi}$$

其中
$$\hat{\Psi} = diag(\hat{\psi}_1, \hat{\psi}_2, \dots, \hat{\psi}_m)$$

上式有一个假定,一模型中的特殊因子是不重要的,因而从Σ的分解中忽略了特殊因子的方差。



注: 残差矩阵 S-LL'-Ψ



例 假定某地固定资产投资率  $x_1$  通货膨胀率  $x_2$ 失业率  $x_3$ 相关系数矩阵为

$$\begin{bmatrix} 1 & 1/5 & -1/5 \\ 1/5 & 1 & 2/5 \\ -1/5 & -2/5 & 1 \end{bmatrix}$$

试用主成分分析法求因子分析模型。



特征根为
$$\lambda_1 = 1.55$$
  $\lambda_2 = 0.85$   $\lambda_3 = 0.6$ 

$$\mathbf{U} = \begin{bmatrix} 0.475 & 0.883 & 0 \\ 0.629 & -0.331 & 0.707 \\ -0.629 & 0.331 & 0.707 \end{bmatrix}$$

$$L = \begin{cases} 0.475\sqrt{1.55} & 0.883\sqrt{0.85} & 0\\ 0.629\sqrt{1.55} & -0.331\sqrt{0.85} & 0.707\sqrt{0.6}\\ -0.629\sqrt{1.55} & 0.331\sqrt{0.85} & 0.707\sqrt{0.6} \end{cases}$$

$$= \begin{bmatrix} 0.569 & 0.814 & 0 \\ 0.783 & -0.305 & 0.548 \\ -0.783 & 0.305 & 0.548 \end{bmatrix}$$



$$x_1 = 0.569F_1 + 0.814F_2$$

$$x_2 = 0.783F_1 - 0.305F_2 + 0.548F_3$$

$$x_3 = -0.783F_1 + 0.305F_2 + 0.548F_3$$

可取前两个因子F1和 $F_2$ 为公共因子,第一公因子 $F_1$ 物价就业因子,对X的贡献为1.55。第二公因子 $F_2$ 为投资因子,对X的贡献为0.85。共同度分别为1,0.706,0.706。



#### 4 因子旋转(正交变换)

#### (一) 为什么要旋转因子

建立了因子分析数学目的不仅仅要找出公共因子以 及对变量讲行分组, 更重要的要知道每个公共因子的意 义,以便进行进一步的分析,如果每个公共因子的含义 不清,则不便于进行实际背景的解释。由于因子载荷阵 是不唯一的, 所以应该对因子载荷阵进行旋转。目的是 使因子载荷阵的结构简化, 使载荷矩阵每列或行的元素 平方值向0和1两极分化。有三种主要的正交旋转法。四 次方最大法、方差最大法和等量最大法。 重磨工商大學

# 变换后因子的共同度

旋转方法

#### 设 $\Gamma$ 正交矩阵,做正交变换 $\mathbf{B} = L\Gamma$

$$\mathbf{B} = (b_{ij})_{p \times p} = (\sum_{k=1}^{m} l_{ik} \gamma_{kj})$$

$$h_i^2(\mathbf{B}) = \sum_{j=1}^{m} b_{ij}^2 = \sum_{j=1}^{m} (\sum_{k=1}^{m} l_{ik} \gamma_{kj})^2$$

$$= \sum_{j=1}^{m} \sum_{k=1}^{m} l_{ik}^2 \gamma_{kj}^2 + \sum_{j=1}^{m} \sum_{k=1}^{m} \sum_{t=1}^{m} l_{ik} l_{it} \gamma_{kj} \gamma_{tj}$$

$$= \sum_{k=1}^{m} l_{ik}^2 \sum_{j=1}^{m} \gamma_{kj}^2 = \sum_{k=1}^{m} l_{ik}^2 = h_i^2(L)$$

变换后因子的共同度没有发生变化!



#### 变换后因子贡献

#### 设Г正交矩阵, 做正交变换

$$\mathbf{B} = L\Gamma$$

$$\mathbf{B} = (b_{ij})_{p \times p} = (\sum_{k=1}^{m} l_{ik} \gamma_{kj})$$

$$S_{j}^{2}(\mathbf{B}) = \sum_{i=1}^{p} b_{ij}^{2} = \sum_{i=1}^{p} (\sum_{k=1}^{m} l_{ik} \gamma_{kj})^{2}$$

$$= \sum_{i=1}^{p} \sum_{k=1}^{m} l_{ik}^{2} \gamma_{kj}^{2} + \sum_{i=1}^{p} \sum_{k=1}^{m} \sum_{t=1}^{m} l_{ik} l_{it} \gamma_{kj} \gamma_{tj}$$

$$= \sum_{i=1}^{p} l_{ik}^{2} \sum_{k=1}^{m} \gamma_{kj}^{2} = S_{j}^{2}(L) \sum_{k=1}^{m} \gamma_{kj}^{2}$$

变换后因子的贡献发生了变化!



#### 1、方差最大法

方差最大法从简化因子载荷矩阵的每一列出发,使和每个因子 有关的载荷的平方的方差最大。当只有少数几个变量在某个因子 上有较高的载荷时,对因子的解释最简单。方差最大的直观意义 是希望通过因子旋转后,使每个因子上的载荷尽量拉开距离,一 部分的载荷趋于±1,另一部分趋于0。当*m*=2时,

$$A = \begin{bmatrix} l_{11} & l_{12} \\ l_{21} & l_{22} \\ \vdots & \vdots \\ l_{p1} & l_{p2} \end{bmatrix} \qquad \begin{aligned} X_1 &= l_{11}F_1 + l_{12}F_2 \\ X_2 &= l_{21}F_1 + l_{22}F_2 \\ \vdots & \vdots \\ X_p &= l_{p1}F_1 + l_{p2}F_2 \end{aligned}$$



设旋转矩阵为: 
$$T = \begin{pmatrix} \cos \varphi & -\sin \varphi \\ \sin \varphi & \cos \varphi \end{pmatrix}$$

$$D$$

$$D$$

$$B = LT = L \begin{pmatrix} \cos \varphi & -\sin \varphi \\ \sin \varphi & \cos \varphi \end{pmatrix}$$

$$= \begin{pmatrix} l_{11} \cos \varphi + l_{12} \sin \varphi & -l_{11} \sin \varphi + l_{12} \cos \varphi \\ \vdots & \vdots \\ l_{p1} \cos \varphi + l_{p2} \sin \varphi & -l_{p1} \sin \varphi + l_{p1} \cos \varphi \end{pmatrix}$$

$$= \begin{pmatrix} b_{11} & b_{12} \\ \vdots & \vdots \\ b_{p1} & b_{p2} \end{pmatrix}$$



# 奥运会十项全能运动项目得分数据的因子分析

百米跑成绩  $X_1$ 

跳远成绩  $X_2$ 

铅球成绩  $X_3$ 

跳高成绩  $X_4$ 

400米跑成绩  $X_5$ 

百米跨栏  $X_6$ 

铁饼成绩  $X_7$ 

撑杆跳远成绩  $X_8$ 

标枪成绩  $X_9$ 

1500米跑成绩  $X_{10}$ 



								]
) 1								
5 0.42	1							
0.51	0.38	1						
0.49	0.19	0.29	1					
0.52	0.36	0.46	0.34	1				
3 0.31	0.73	0.27	0.17	0.32	1			
0.36	0.24	0.39	0.23	0.33	0.24	1		
0.21	0.44	0.17	0.13	0.18	0.34	0.24	1	
7 0.09	-0.08	0.18	0.39	0.01	-0.02	0.17	-0.02	1
	40.5130.4900.5230.3100.3610.21	5     0.42     1       4     0.51     0.38       3     0.49     0.19       0     0.52     0.36       8     0.31     0.73       0     0.36     0.24       1     0.21     0.44	5       0.42       1         4       0.51       0.38       1         3       0.49       0.19       0.29         0       0.52       0.36       0.46         8       0.31       0.73       0.27         0       0.36       0.24       0.39         1       0.21       0.44       0.17	5     0.42     1       4     0.51     0.38     1       3     0.49     0.19     0.29     1       0     0.52     0.36     0.46     0.34       8     0.31     0.73     0.27     0.17       0     0.36     0.24     0.39     0.23       1     0.21     0.44     0.17     0.13	5     0.42     1       4     0.51     0.38     1       3     0.49     0.19     0.29     1       0     0.52     0.36     0.46     0.34     1       8     0.31     0.73     0.27     0.17     0.32       0     0.36     0.24     0.39     0.23     0.33       1     0.21     0.44     0.17     0.13     0.18	5       0.42       1         4       0.51       0.38       1         3       0.49       0.19       0.29       1         0       0.52       0.36       0.46       0.34       1         3       0.31       0.73       0.27       0.17       0.32       1         0       0.36       0.24       0.39       0.23       0.33       0.24         1       0.21       0.44       0.17       0.13       0.18       0.34	5       0.42       1         4       0.51       0.38       1         3       0.49       0.19       0.29       1         0       0.52       0.36       0.46       0.34       1         8       0.31       0.73       0.27       0.17       0.32       1         0       0.36       0.24       0.39       0.23       0.33       0.24       1         1       0.21       0.44       0.17       0.13       0.18       0.34       0.24	5 0.42 1 4 0.51 0.38 1 3 0.49 0.19 0.29 1 0 0.52 0.36 0.46 0.34 1 8 0.31 0.73 0.27 0.17 0.32 1



变量	$ F_1 $	$\overline{F_2}$	$\overline{F_3}$	$F_4$	共同度
$X_1$	0.691	0.217	-0.58	-0.206	0.84
$X_{\gamma}$	0.789	0.184	-0.193	0.092	0.7
$X_3$	0.702	0.535	0.047	-0.175	0.8
$X_4$	0.674	0.134	0.139	0.396	0.65
$X_5$	0.62	0.551	-0.084	-0.419	0.87
$X_6$	0.687	0.042	-0.161	0.345	0.62
$X_{7}$	0.621	-0.521	0.109	-0.234	0.72
$X_{8}$	0.538	0.087	0.411	0.44	0.66
$X_{9}$	0.434	-0.439	0.372	-0.235	0.57
$X_{10}$	0.147	0.596	0.658	-0.279	0.89

因子载荷矩阵可以看出,除第一因子在所有的变量在公共因子 上有较大的正载荷,可以称为一般运动因子。其他的3个因子不太 容易解释。似乎是跑和投掷的能力对比,似乎是长跑耐力和强跑速 度的对比。于是考虑旋转因子,得下表

变量	$F_1$	$F_2$	$F_3$	$F_4$	共同度
$X_1$	0.844*	0.136	0.156	-0.113	0.84
$X_2$	0.631*	0.194	0.515*	-0.006	0.7
$X_3$	0.243	0.825*	0.223	-0.148	0.81
$X_4$	0.239	0.15	0.750*	0.076	0.65
$X_5$	0.797*	0.075	0.102	0.468	0.87
$X_6$	0.404	0.153	0.635*	-0.17	0.62
$X_7$	0.186	0.814*	0.147	-0.079	0.72
$X_8$	-0.036	0.176	0.762*	0.217	0.66
$X_9$	-0.048	0.735*	0.11	0.141	0.57
$X_{10}$	0.045	-0.041	0.112	0.934*	0.89



通过旋转,因子有了较为明确的含义。 $X_1$ 百米跑, $X_2$ 跳远和 $X_5$ 400米跑,需要爆发力的项目在 $F_1$ 有较大的载荷, $F_1$ 可以称为短跑速度因子;

 $X_3$ 铅球, $X_7$ 铁饼和 $X_9$ 标枪在 $F_2$ 上有较大的载荷,可以称为爆发性臂力因子;

 $X_6$ 百米跨栏, $X_8$ 撑杆跳远, $X_2$ 跳远和为 $X_4$ 跳高在  $F_3$ 上有较大的载荷, $F_3$ 爆发腿力因子; $F_4$ 长跑耐力因子。



## 5 因子得分

#### (一) 因子得分的概念

前面我们主要解决了用公共因子的线性组合来表示一组观测变量的有关问题。如果我们要使用这些因子做其他的研究,比如把得到的因子作为自变量来做回归分析,对样本进行分类或评价,这就需要我们对公共因子进行测度,即给出公共因子的值。



#### 因子分析的数学模型为:

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1m} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2m} \\ \vdots & \vdots & & \vdots \\ \alpha_{p1} & \alpha_{p2} & \cdots & \alpha_{pm} \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix}$$

原变量被表示为公共因子的线性组合,当载荷矩阵旋转之后,公共因子可以做出解释,通常的情况下,我们还想反过来把公共因子表示为原标量的线性组合。

因子得分函数: 
$$F_j = \beta_{j1}X_1 + \cdots + \beta_{jp}X_p$$
  $j = 1, \cdots, m$ 

可见,要求得每个因子的得分,必须求得分函数的系数,而由于p>m,所以不能得到精确的得分,只能通过**,**使度工商大学

#### 2、Thompson方法

$$\hat{F}_{j} = b_{j1}X_{1} + \dots + b_{jp}X_{p}$$
  $j = 1, \dots, m$ 

$$\begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1p} \\ b_{21} & b_{22} & \cdots & b_{2p} \\ \vdots & \vdots & & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mp} \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_m \end{bmatrix}$$



$$\alpha_{ij} = \gamma_{x_i F_j} = E(X_i F_j)$$

$$= E[X_i(b_{j1} X_1 + \dots + b_{jp} X_p)]$$

$$= b_{j1} \gamma_{i1} + \dots + b_{jp} \gamma_{ip}$$

$$= \begin{bmatrix} r_{i1} & r_{i2} & \dots & r_{ip} \end{bmatrix} \begin{bmatrix} b_{j1} \\ b_{j2} \\ \vdots \\ b_{jp} \end{bmatrix}$$

则,我们有如下的方程组:



$$\begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1p} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2p} \\ \vdots & \vdots & & \vdots \\ \gamma_{p1} & \gamma_{p2} & \cdots & \gamma_{pp} \end{bmatrix} \begin{bmatrix} b_{j1} \\ b_{j2} \\ \vdots \\ b_{jp} \end{bmatrix} = \begin{bmatrix} l_{1j} \\ l_{2j} \\ \vdots \\ l_{pj} \end{bmatrix}$$
  $j=1, 2, \cdots, m$ 

$$\begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1p} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2p} \\ \vdots & \vdots & & \vdots \\ \gamma_{p1} & \gamma_{p2} & \cdots & \gamma_{pp} \end{bmatrix}$$
为原始变量的相关系数矩阵



```
为第j个因子得分函数的系数
为载荷矩阵的第j列
```

注: 共需要解m次才能解出 所有的得分函数的系数。



# 6 因子分析的步骤

一、因子分析通常包括以下五个步骤

#### 选择分析的变量

用定性分析和定量分析的方法选择变量,因子分析的前提条件是观测变量间有较强的相关性,因为如果变量之间 无相关性或相关性较小的话,他们不会有共享因子,所以 原始变量间应该有较强的相关性。

#### 计算所选原始变量的相关系数矩阵

#### 提取公共因子

这一步要确定因子求解的方法和因子的个数。需要根据研究者的设计方案或有关的经验或知识事先确定。因子个数的确定可以根据因子方差的大小。只取方差大于1(或特征值大于1)的那些因子,因为方差小于1的因子其贡献可能很小;按照因子的累计方差贡献率来确定,一般认为要达到85%才能符合要求;

#### 因子旋转

通过坐标变换使每个原始变量在尽可能少的因子之间有密切的关系,这样因子解的实际意义更容易解释,并为每个潜在因子赋予有实际意义的名字。



#### 计算因子得分

求出各样本的因子得分,有了因子得分值,则可以在许多分析中使用这些因子,例如以因子的得分做聚类分析的变量,做回归分析中的回归因子。



