多元统计分析

第9章　因子分析及R应用

第9章 因子分析及R应用

- §1 因子分析的背景和基本思想
- §2 正交因子模型及其统计性质
- §3 因子载荷矩阵的估计方法
- §4 因子旋转
- §5 因子得分
- §6 因子分析的步骤
- §7 R语言中因子分析的常用函数及实例

§1 因子分析的背景和基本思想

因子分析（Factor Analysis）思想的提出，始于1904年 Charles Spearman，Karl Pearson等学者对测定智力、学生考试成绩等的研究。近年来，随着计算机技术的发展，人们将因子分析的理论成功地应用于多个研究领域，如：心理学、医学、气象、地质、经济学等领域；这些应用使得因子分析的理论和方法更加丰富。

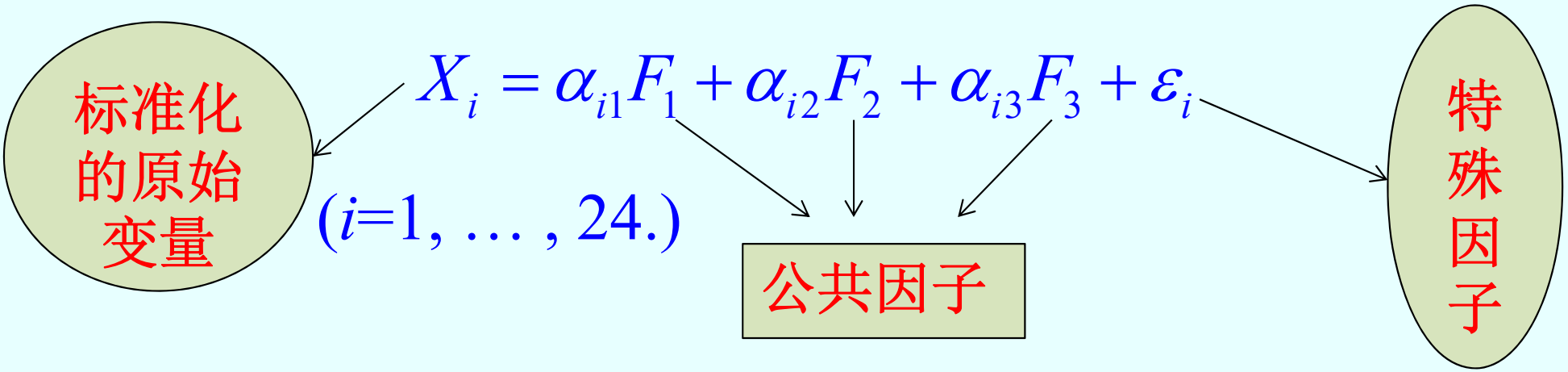
因子分析可以看作主成分分析的推广，但二者有所区别。它也是利用降维的思想：由研究原始变量内部的依赖关系出发，把具有复杂关系的原始变量，归结为少数几个综合变量的多元统计方法。

因子分析往往更倾向描述原始变量间的相关关系；因此：人们做因子分析时，出发点经常直接选择原始变量的相关系数矩阵。

§1 因子分析的背景和基本思想

- 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，用少数几个假想变量/潜变量，探求观测数据中的基本结构。
- 因子分析获得的潜变量，能够反映原始变量的主要信息。
- 注意：原始变量是可观测的显变量；
  潜变量是不可观测的，因而常称因子(factor)。

引例：在企业形象或品牌形象的研究中，消费者可以通过一个有24个变量指标构成的评价体系，评价百货商场的优劣。

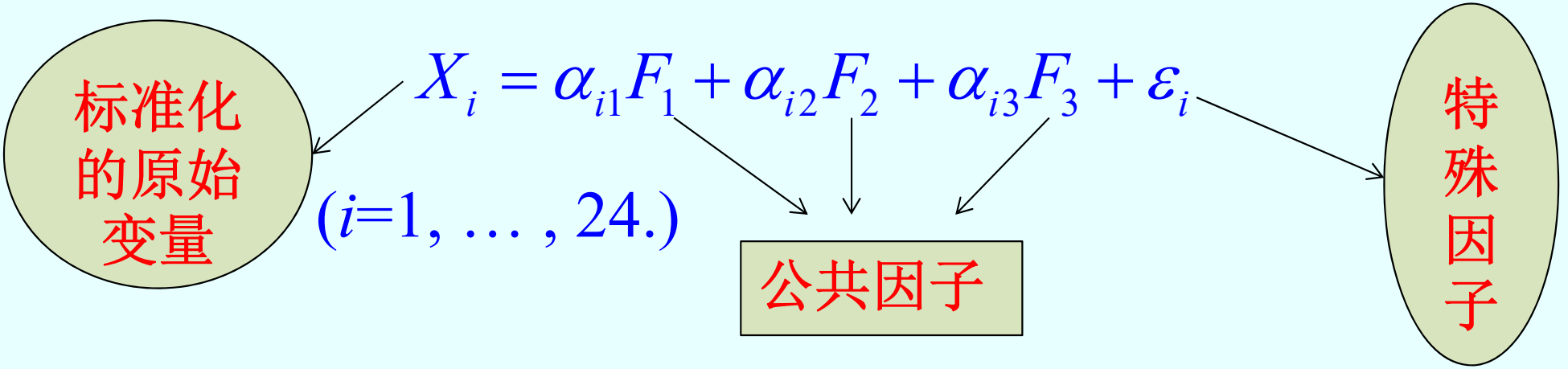


引例：在企业形象或品牌形象的研究中，消费者可以通过一个有24个变量指标构成的评价体系，评价百货商场的优劣。

注：消费者主要关心的是3个方面，即：

商店环境、商店服务、商品价格。

假如使用因子分析方法后，可以从24个变量：找出反映商店的上述方面的3个潜变量（因子），则可方便地对商店进行综合评价。



§ 2 正交因子模型及其统计性质

正交因子模型：

设 $X_i\,(i=1,2,\cdots,p)$为对第i个原始变量进行标准化后的变量，并将其表示为

$$X_i = a_{i1}F_1 + \cdots + a_{im}F_m + \varepsilon_i \qquad (m \le p)$$

当上式中的各个量满足一定的约束条件（见下页），则称：$\{F_i\}$为公共因子，它们是不可观测的假想变量（即：潜变量）；其系数$\{a_{ij}\}$称为因子载荷。

并且，称$\{\varepsilon_i\}$是特殊因子，它们表示随机误差。

§ 2 正交因子模型及其统计性质

正交因子模型需满足的约束条件：

1. $E(F)=0$ ， $Var(F)=\begin{bmatrix}1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1\end{bmatrix}=I$

即：公共因子间互不相关；各公共因子的均值为0，方差为1。

2. $Var(\varepsilon)=\begin{bmatrix}\sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_p^2\end{bmatrix}$

即：特殊因子间互不相关，

$$\varepsilon_i \sim N(0,\sigma_i^2)$$

3. $\mathrm{Cov}(F,\varepsilon)=0$ 即：F与ε不相关.

§ 2 正交因子模型及其统计性质

正交因子模型：

设 $X_i\,(i=1,2,\cdots,p)$为对第i个原始变量进行标准化后的变量，并将其表示为

$$\boldsymbol{X_i = a_{i1}F_1 + \cdots + a_{im}F_m + \varepsilon_i} \qquad (m \le p)$$

记：$X=(X_1,\cdots,X_p)',F=(F_1,\cdots,F_m)',\varepsilon=(\varepsilon_1,\cdots,\varepsilon_p)'$

§ 2 正交因子模型及其统计性质

正交因子模型：

设 $X_i\,(i=1,2,\cdots,p)$为对第i个原始变量进行标准化后的变量，并将其表示为

$$X_i = a_{i1}F_1 + \cdots + a_{im}F_m + \varepsilon_i \qquad (m \le p)$$

记：$X=(X_1,\cdots,X_p)',F=(F_1,\cdots,F_m)',\varepsilon=(\varepsilon_1,\cdots,\varepsilon_p)'$

则上述表达式进一步为：$X = AF + \varepsilon$

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

§ 2 正交因子模型及其统计性质

正交因子模型：

设 $X_i\,(i=1,2,\cdots,p)$为对第i个原始变量进行标准化后的变量，并将其表示为

$$\boldsymbol{X_i = a_{i1}F_1 + \cdots + a_{im}F_m + \varepsilon_i} \qquad (m \le p)$$

记：$X=(X_1,\cdots,X_p)',F=(F_1,\cdots,F_m)',\varepsilon=(\varepsilon_1,\cdots,\varepsilon_p)'$

则上述表达式进一步为：$X = AF + \varepsilon$

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

因子载荷矩阵

特殊因子：随机误差

§ 2 正交因子模型及其统计性质

正交因子模型需满足的约束条件：

1. $E(F)=0$ ， $Var(F)=\begin{bmatrix}1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1\end{bmatrix}=I$

即：公共因子间互不相关；各公共因子的均值为0，方差为1。

2. $Var(\varepsilon)=\begin{bmatrix}\sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_p^2\end{bmatrix}$

即：特殊因子间互不相关，

$$\varepsilon_i \sim N(0,\sigma_i^2)$$

3. $\mathrm{Cov}(F,\varepsilon)=0$ 即：F与ε不相关.

§ 2 正交因子模型及其统计性质

1、X的协方差矩阵的分解

$Var(F) = I$

由 $X = AF + \varepsilon$

$\Rightarrow \quad Var(X) = AVar(F)A' + Var(\varepsilon)$

记 $D = Var(\varepsilon) = diag(\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2)$

对角阵

$\Rightarrow R = AA' + D$

注：D的主对角线元素的值越小，则公共因子共享的信息越多。

§ 2 正交因子模型及其统计性质

2、因子载荷 a_{ij} 的统计意义

因子载荷 a_{ij} 是第 i 变量与第 j 公因子的相关系数。

证明：由 $X_i = a_{i1}F_1 + \cdots + a_{im}F_m + \varepsilon_i$

根据正交因子模型性质，有：

$$\rho_{X_i,\,F_j} = \mathrm{Cov}(X_i, F_j) = Cov(\sum_{k=1}^{m} a_{ik}F_k + \varepsilon_i, F_j)$$

$$= Cov(\sum_{k=1}^{m} a_{ik}F_k, F_j) + Cov(\varepsilon_i, F_j)$$

$$= \sum_{k=1}^{m} a_{ik}Cov(F_k, F_j) + 0 = a_{ij}Cov(F_j, F_j) = a_{ij}$$

§ 2 正交因子模型及其统计性质

2、因子载荷 a_{ij} 的统计意义

因子载荷 a_{ij} 是第 i 变量与第 j 公因子的相关系数。

证明：由 $X_i = a_{i1}F_1 + \cdots + a_{im}F_m + \varepsilon_i$

根据正交因子模型性质，有：

$$\rho_{X_i,F_j} = \mathrm{Cov}(X_i, F_j) = Cov(\sum_{k=1}^{m} a_{ik}F_k + \varepsilon_i, F_j) = Cov(\sum_{k=1}^{m} a_{ik}F_k, F_j) + Cov(\varepsilon_i, F_j)$$

$$= \sum_{k=1}^{m} a_{ik}Cov(F_k, F_j) + 0 = a_{ij}Cov(F_j, F_j) = a_{ij}$$

- **因子载荷 a_{ij}**（载荷矩阵A的第i行，第j列的元素）**反映：**

  第 i 个变量与第 j 个公共因子之间的相关系数的大小。

  它的绝对值越大，则二者的线性相关程度越高。

§ 2 正交因子模型及其统计性质

定义1：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 $h_i^2 = \sum_{j=1}^{m} a_{ij}^2$.

变量共同度的统计意义：

$X_i = a_{i1}F_1 + \cdots + a_{im}F_m + \varepsilon_i$ 两边求方差

$$\Rightarrow Var(X_i) = a^2{}_{i1}Var(F_1) + \cdots + a^2{}_{im}Var(F_m) + Var(\varepsilon_i)$$

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

§ 2 正交因子模型及其统计性质

定义1：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 $h_i^2 = \sum_{j=1}^{m} a_{ij}^2$.

变量共同度的统计意义：

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

因此，所有的公共因子和特殊因子对变量 X_i 的贡献为1。如果 $\sum_{j=1}^{m} a_{ij}^2$ 非常靠近1，σ_i^2 非常小，则因子分析的效果好，从原始变量空间到公共因子空间的转化性质一般而言就好。

§ 2 正交因子模型及其统计性质

定义1：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 $h_i^2 = \sum_{j=1}^{m} a_{ij}^2$.

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

如：

§ 2 正交因子模型及其统计性质

定义2：因子载荷矩阵中第 j 列元素的平方和

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

称为 F_j 对所有变量X 的方差贡献, 用于衡量它的相对重要性。

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

如:

§ 2 正交因子模型及其统计性质

定义2（续）

- **F_j所解释的总方差的比例（方差贡献率）**可以进一步表示为 S_j / p。

  注：原始变量已作标准化。A所有元素平方和<=p。

- **$F_1, \cdots, F_m$所解释的总方差的累积比例（累积方差贡献率）为** $\left(\sum_{j=1}^{m} S_j\right) / p$ 。

  注：一般选取m的个数，使得该累积贡献率>80%。

例1. 某次运动会的男子径赛运动涉及如下八个变量：

x_1：100米（秒）	x_5：1500米（分）
x_2：200米（秒）	x_6：5000米（分）
x_3：400米（秒）	x_7：10000米（分）
x_4：800米（秒）	x_8：马拉松（分）

表：八项男子径赛运动成绩记录的样本相关矩阵R

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
x_1	1.000							
x_2	0.923	1.000						
x_3	0.841	0.851	1.000					
x_4	0.756	0.807	0.870	1.000				
x_5	0.700	0.775	0.835	0.918	1.000			
x_6	0.619	0.695	0.779	0.864	0.928	1.000		
x_7	0.633	0.697	0.787	0.869	0.935	0.975	1.000	
x_8	0.520	0.596	0.705	0.806	0.866	0.932	0.943	1.000

- 例1. 由相关阵R出发，假设已分别得到 $m=1$ 和 $m=2$ 时正交因子模型的因子载荷矩阵，如下：

变　量	$m=1$		$m=2$		
	因子载荷矩阵 F_1	变量的共同度	因子载荷矩阵 F_1	F_2	变量的共同度
x_1^* ：100米	**0.817**		**0.817**	**0.531**	
x_2^* ：200米	**0.867**		**0.867**	**0.432**	
x_3^* ：400米	**0.915**		**0.915**	**0.233**	
x_4^* ：800米	**0.949**		**0.949**	**0.012**	
x_5^* ：1500米	**0.959**		**0.959**	**-0.131**	
x_6^* ：5000米	**0.938**		**0.938**	**-0.292**	
x_7^* ：10000米	**0.944**		**0.944**	**-0.287**	
x_8^* ：马拉松	**0.880**		**0.880**	**-0.411**	
因子的累积方差贡献率					

- 例1. 由相关阵R出发，假设已分别得到 m=1 和 m=2 时正交因子模型的因子载荷矩阵，如下：

变　量	m=1		m=2		
	因子载荷矩阵 F_1	变量的共同度	因子载荷矩阵 F_1	F_2	变量的共同度
x_1^* ：100米	0.817	0.668	0.817	0.531	0.950
x_2^* ：200米	0.867	0.752	0.867	0.432	0.939
x_3^* ：400米	0.915	0.838	0.915	0.233	0.892
x_4^* ：800米	0.949	0.900	0.949	0.012	0.900
x_5^* ：1500米	0.959	0.920	0.959	-0.131	0.938
x_6^* ：5000米	0.938	0.879	0.938	-0.292	0.965
x_7^* ：10000米	0.944	0.891	0.944	-0.287	0.973
x_8^* ：马拉松	0.880	0.774	0.880	-0.411	0.943
因子的累积方差贡献率	0.828		0.828	0.938	

§3　因子载荷矩阵的估计方法

（1）主成分法（线性代数-对称矩阵的谱分解）

（2）主轴因子法

（3）极大似然法

§3　因子载荷矩阵的估计方法

主成分法

由主成分分析：

$$\begin{cases} Y_1 = u_{11}X_1 + u_{12}X_2 + \cdots + u_{1p}X_p \\ Y_2 = u_{21}X_1 + u_{22}X_2 + \cdots + u_{2p}X_p \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ Y_p = u_{p1}X_1 + u_{p2}X_2 + \cdots + u_{pp}X_p \end{cases}$$

，即 $Y = U'X$

§3　因子载荷矩阵的估计方法

主成分法

由主成分分析：

$$\begin{cases} Y_1 = u_{11}X_1 + u_{12}X_2 + \cdots + u_{1p}X_p \\ Y_2 = u_{21}X_1 + u_{22}X_2 + \cdots + u_{2p}X_p \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ Y_p = u_{p1}X_1 + u_{p2}X_2 + \cdots + u_{pp}X_p \end{cases}$$

，即 $Y = U'X$

因为U为正交矩阵，有：$X = UY$

$$\begin{cases} X_1 = u_{11}Y_1 + u_{21}Y_2 + \cdots + u_{p1}Y_p \\ X_2 = u_{12}Y_1 + u_{22}Y_2 + \cdots + u_{p2}Y_p \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ X_p = u_{1p}Y_1 + u_{2p}Y_2 + \cdots + u_{pp}Y_p \end{cases}$$

§3　因子载荷矩阵的估计方法

主成分法

由主成分分析：

$$\begin{cases} Y_1 = u_{11}X_1 + u_{12}X_2 + \cdots + u_{1p}X_p \\ Y_2 = u_{21}X_1 + u_{22}X_2 + \cdots + u_{2p}X_p \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ Y_p = u_{p1}X_1 + u_{p2}X_2 + \cdots + u_{pp}X_p \end{cases}$$

，即 $Y = U'X$

因为U为正交矩阵，有：$X = UY$

$$\begin{cases} X_1 = u_{11}Y_1 + u_{21}Y_2 + \cdots + u_{p1}Y_p + \varepsilon_1 \\ X_2 = u_{12}Y_1 + u_{22}Y_2 + \cdots + u_{p2}Y_p + \varepsilon_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ X_p = u_{1p}Y_1 + u_{2p}Y_2 + \cdots + u_{pp}Y_p + \varepsilon_p \end{cases}$$

若取前m个Y, 则:

$$\begin{cases} X_1 = u_{11}\sqrt{\lambda_1}\dfrac{1}{\sqrt{\lambda_1}}Y_1 + u_{21}\sqrt{\lambda_2}\dfrac{1}{\sqrt{\lambda_2}}Y_2 + \cdots + u_{m1}\sqrt{\lambda_m}\dfrac{1}{\sqrt{\lambda_m}}Y_m + \varepsilon_1 \\ X_2 = u_{12}\sqrt{\lambda_1}\dfrac{1}{\sqrt{\lambda_1}}Y_1 + u_{22}\sqrt{\lambda_2}\dfrac{1}{\sqrt{\lambda_2}}Y_2 + \cdots + u_{m2}\sqrt{\lambda_m}\dfrac{1}{\sqrt{\lambda_m}}Y_m + \varepsilon_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ X_p = u_{1p}\sqrt{\lambda_1}\dfrac{1}{\sqrt{\lambda_1}}Y_1 + u_{2p}\sqrt{\lambda_2}\dfrac{1}{\sqrt{\lambda_2}}Y_2 + \cdots + u_{mp}\sqrt{\lambda_m}\dfrac{1}{\sqrt{\lambda_m}}Y_m + \varepsilon_p \end{cases}$$

令 $F_j = Y_j / \sqrt{\lambda_j}$ ，

记 $a_{ij} = u_{ji} \times \sqrt{\lambda_j}$

主成分法

则：

$$\begin{cases} X_1 = a_{11}F_1 + a_{12}F_2 + \cdots + a_{1m}F_m + \varepsilon_1 \\ X_2 = a_{21}F_1 + a_{22}F_2 + \cdots + a_{2m}F_m + \varepsilon_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ X_p = a_{p1}F_1 + a_{p2}F_2 + \cdots a_{pm}F_m + \varepsilon_p \end{cases}$$

§3　因子载荷矩阵的估计方法

- 思考：　由主成分法求得因子载荷矩阵A，那么，
  (1) 每个公共因子的方差贡献是多少？
  (2) 每个公共因子的方差贡献率是多少？

§3　因子载荷矩阵的估计方法

（1）主成分法（线性代数-对称矩阵的谱分解）

（2）主轴因子法

假定m个公共因子只能解释原始变量的大部分方差，并假设特殊因子的方差可以估算得到，则可以利用公共因子方差（或共同度）来代替相关矩阵主对角线上的元素1，并以新得到的矩阵R*=R-D（称为约相关矩阵）出发，根据R*=AA’对其求特征根和单位正交特征向量，得到因子解 A。

（3）极大似然法

假定公共因子和特殊因子服从正态分布，得到因子载荷和特殊因子方差的极大似然估计。

§4　因子旋转

建立因子分析的数学目的，不仅仅是找出公共因子、对变量分组，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，则不便于对实际意义进行解释。**由于因子载荷阵是不惟一的，所以对因子载荷阵进行旋转，在理论上是合理的。**

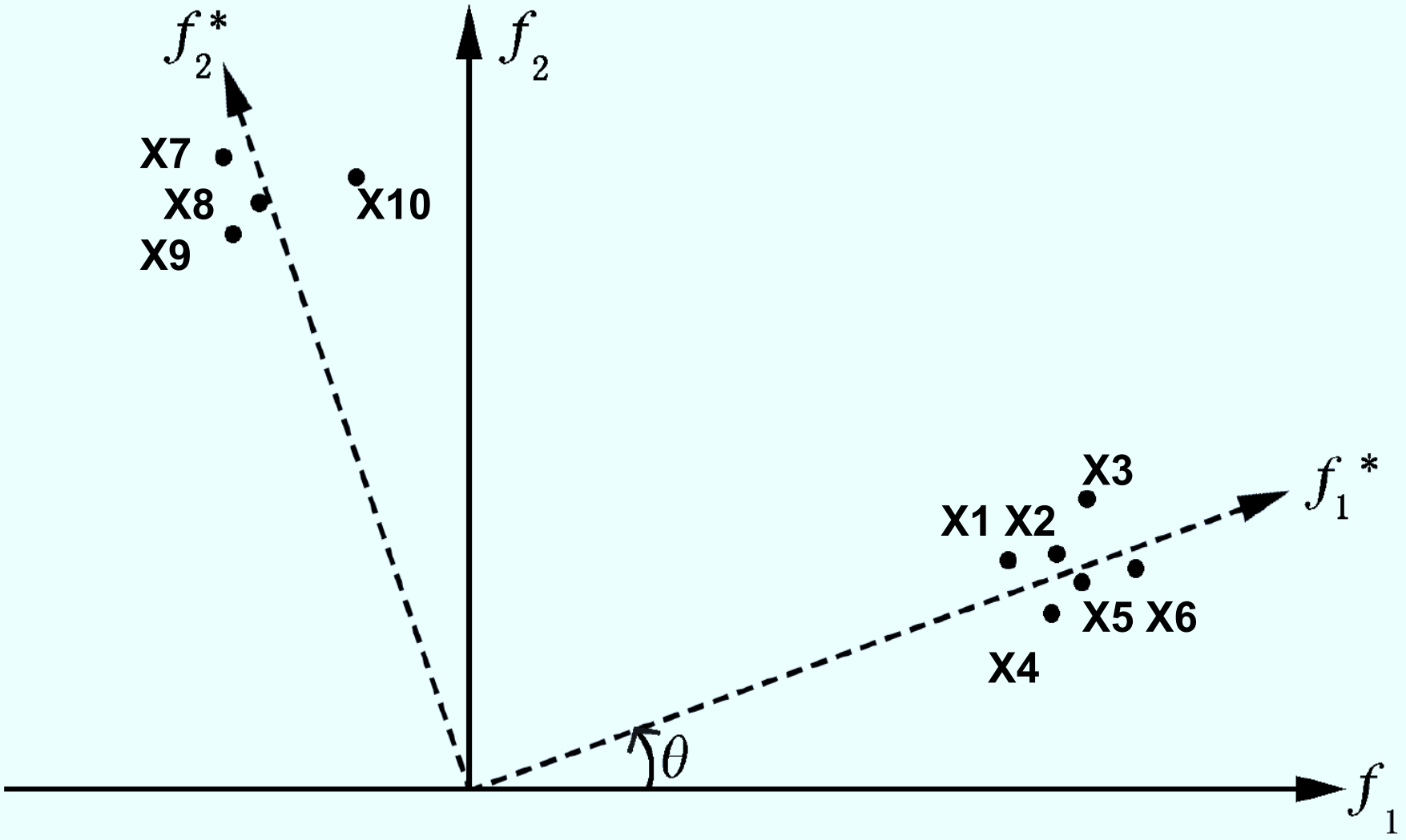
- 因子旋转的目的：使因子载荷阵的结构简化，便于解释。
- 因子旋转分为两大类：正交旋转、斜交旋转。具体地：

（1）正交旋转后的模型仍为正交因子模型；

（2）斜交旋转后的模型为斜交因子模型，

即公因子间不满足正交性。

正交因子旋转的示意图



正交旋转后，可能产生更简单的因子载荷结构

正交因子旋转

【原理】：因子载荷矩阵不是惟一的。

设T为一个m×m的正交矩阵，令$A^*=AT$，$F^*=T'F$，

则：正交旋转后，正交因子模型可以进一步表示为

$$\mathbf{X} = \mathbf{A}^*\mathbf{F}^* + \boldsymbol{\varepsilon}$$

正交因子旋转

【原理】：因子载荷矩阵不是惟一的。

设T为一个m×m的正交矩阵，令$A^*=AT$，$F^*=T'F$，

则：正交旋转后，正交因子模型可以进一步表示为

$$\mathbf{X} = \mathbf{A}^*\mathbf{F}^* + \boldsymbol{\varepsilon} \quad \text{(仍满足约束条件)}$$

$$E(\mathbf{F}^*) = \mathbf{0} \qquad E(\boldsymbol{\varepsilon}) = \mathbf{0}$$

$$Var(\mathbf{F}^*) = Var(\mathbf{T}'\mathbf{F}) = \mathbf{T}'Var(\mathbf{F})\mathbf{T} = \mathbf{I}$$

$$Var(\boldsymbol{\varepsilon}) = diag(\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2)$$

$$\mathrm{Cov}(\mathbf{F}^*, \boldsymbol{\varepsilon}) = E(\mathbf{F}^*\boldsymbol{\varepsilon}') = \mathbf{0}$$

注：正交旋转后，变量的共同度保持不变。

例2：奥运会十项全能运动项目得分数据的因子分析

X_1 百米跑

X_2 跳远

X_3 铅球

X_4 跳高

X_5 400米跑

X_6 百米跨栏

X_7 铁饼

X_8 撑杆跳远

X_9 标枪

X_{10} 1500米跑

例2：奥运会十项全能运动项目得分数据的因子分析

相关系数矩阵

$$
R = \begin{bmatrix}
1 & & & & & & & & & \\
0.59 & 1 & & & & & & & & \\
0.35 & 0.42 & 1 & & & & & & & \\
0.34 & 0.51 & 0.38 & 1 & & & & & & \\
0.63 & 0.49 & 0.19 & 0.29 & 1 & & & & & \\
0.40 & 0.52 & 0.36 & 0.46 & 0.34 & 1 & & & & \\
0.28 & 0.31 & 0.73 & 0.27 & 0.17 & 0.32 & 1 & & & \\
0.20 & 0.36 & 0.24 & 0.39 & 0.23 & 0.33 & 0.24 & 1 & & \\
0.11 & 0.21 & 0.44 & 0.17 & 0.13 & 0.18 & 0.34 & 0.24 & 1 & \\
-0.07 & 0.09 & -0.08 & 0.18 & 0.39 & 0.01 & -0.02 & 0.17 & -0.02 & 1
\end{bmatrix}
$$

未旋转的因子载荷阵

变量	F_1	F_2	F_3	F_4	共同度
X_1 百米	0.691	0.217	-0.58	-0.206	0.84
X_2 跳远	0.789	0.184	-0.193	0.092	0.7
X_3 铅球	0.702	0.535	0.047	-0.175	0.81
X_4 跳高	0.674	0.134	0.139	0.396	0.65
X_5 400米	0.62	0.551	-0.084	-0.419	0.87
X_6 百米跨	0.687	0.042	-0.161	0.345	0.62
X_7 铁饼	0.621	-0.521	0.109	-0.234	0.72
X_8 撑杆跳	0.538	0.087	0.411	0.44	0.66
X_9 标枪	0.434	-0.439	0.372	-0.235	0.57
X_{10} 1500	0.147	0.596	0.658	-0.279	0.89

可以看出：第一因子在所有的变量在公共因子上有较大的正载荷，可称为一般运动因子。

但是其他的3个因子，不太容易解释，似乎是跑和投掷的能力对比，长跑耐力和短跑速度的对比。

正交旋转后的因子载荷阵

变量	F_1	F_2	F_3	F_4	共同度
X_1 百米	0.844*	0.136	0.156	-0.113	0.84
X_2 跳远	0.631*	0.194	0.515*	-0.006	0.7
X_3 铅球	0.243	0.825*	0.223	-0.148	0.81
X_4 跳高	0.239	0.15	0.750*	0.076	0.65
X_5 400米	0.797*	0.075	0.102	0.468	0.87
X_6 百米跨	0.404	0.153	0.635*	-0.17	0.62
X_7 铁饼	0.186	0.814*	0.147	-0.079	0.72
X_8 撑杆跳	-0.036	0.176	0.762*	0.217	0.66
X_9 标枪	-0.048	0.735*	0.11	0.141	0.57
X_{10} 1500	0.045	-0.041	0.112	0.934*	0.89

例2：奥运会十项全能运动项目得分数据的因子分析

通过正交因子旋转，希望因子有较明确的含义，如：

（1）X1百米跑，X2跳远和 X5四百米跑，

在F1有较大载荷，因此 F1为（短跑）速度因子；

（2）X3铅球，X7铁饼和 X9标枪，

在F2上有较大载荷，因此 F2为爆发性臂力因子；

（3）X6百米跨栏，X8撑杆跳远，X2跳远和 X4跳高，

在F3上有较大载荷，因此 F3为爆发性腿部力量因子；

（4）F4可概括为长跑耐力因子。

【注】：正交旋转后，变量的共同度没有发生变化！

正交因子旋转的常用方法

- **方差最大法：Varimax方法**

  直观意义：从简化因子载荷阵的每列出发，使各个因子的载荷两极分化。当只有少数几个变量在某个因子上有较高的载荷时，对因子的解释会更简单。

  即：通过因子旋转，使得每个因子的载荷尽量拉开距离，希望一部分载荷趋于±1，另一部分载荷趋于0。

  （注：旋转角度的求解较为复杂，软件可直接实现，教材P250）

- **其它正交旋转法：**四次方最大法、等量最大法，等。

* 斜交因子旋转

- **设Q为m×m的非奇异矩阵（及满足若干其它的约束条件），令$A^*=AQ$，$F^*=Q'F$, 其中Q'Q 不为单位阵。**

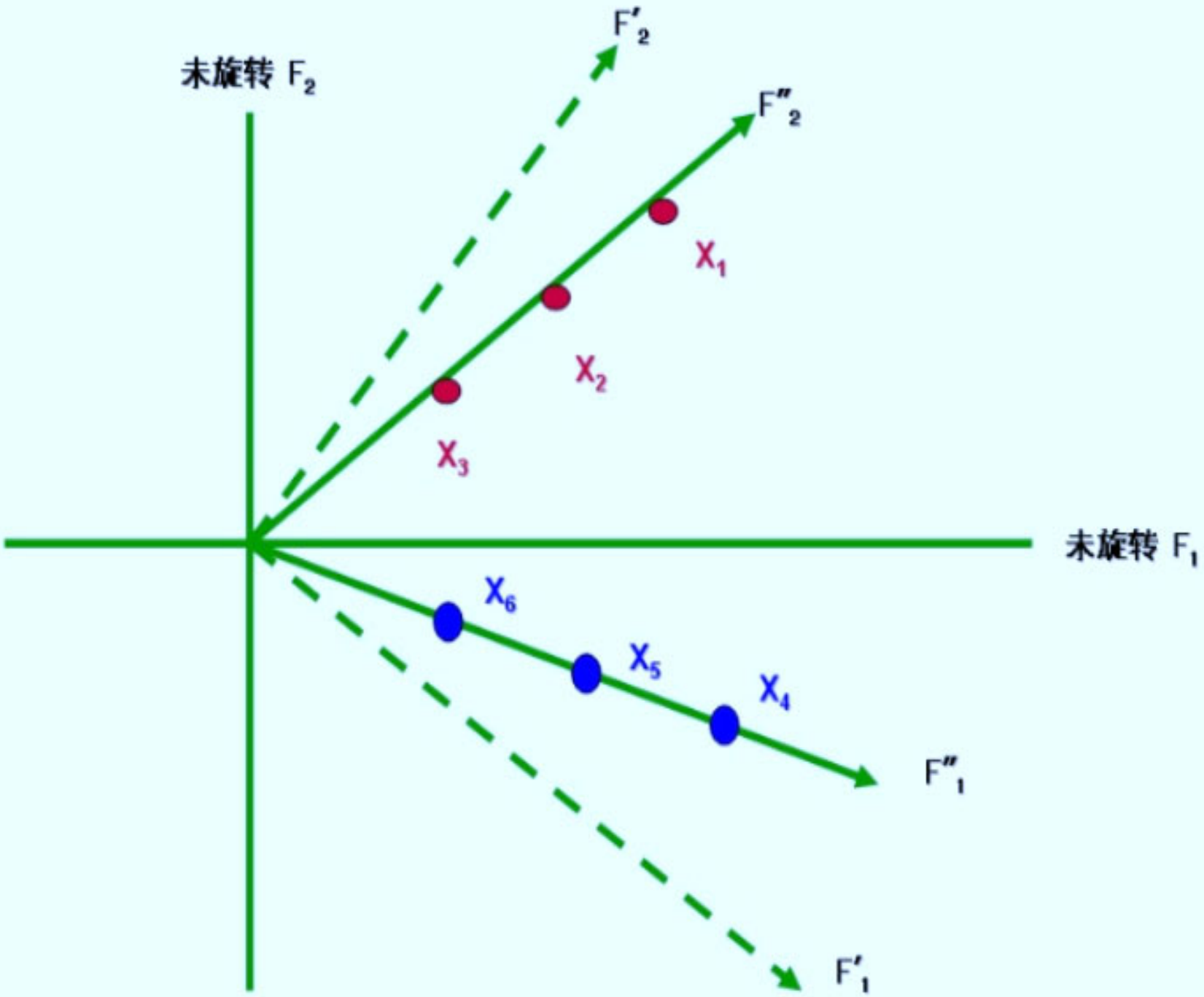
  则：斜交旋转后的因子模型，因子间不再正交：

$$\mathbf{X} = \mathbf{A}^*\mathbf{F}^* + \boldsymbol{\varepsilon}^*$$

$$Var(\mathbf{F}^*) = Var(Q'\mathbf{F}) = Q'Var(\mathbf{F})Q \neq \mathbf{I}$$

- 注：斜交旋转后，变量的共同度不再保持不变。
  斜交旋转往往易使坐标轴“穿过”更多的坐标点。
  常用的斜交旋转方法，如：Direct Oblimin方法（直接斜交旋转），Promax方法（最优斜交旋转），等。

正交旋转与斜交旋转的因子载荷结构示意图



§5　因子得分

前面主要解决了用公因子的线性组合，表示原始变量的问题。确定选用的因子模型，可能是正交的，也可能是斜交的。

如果要使用这些因子做进一步的其他研究，如：把得到的因子作为自变量做回归分析；或根据因子分析的结果，对样本进行分类或评价，就需要给出每个样本对应的公因子的值，即因子得分。

§5　因子得分

Thompson因子得分的计算方法（也称回归法）：

因子得分的计算公式：

$$\hat{F} = A'R^{-1}x$$

A为因子载荷阵, R为原始变量X的相关阵, x为X的观测值。

（注：经加权最小二乘的推导思想，可得Bartlett因子得分，该因子得分通过合理近似，也可导出上述公式， 详见《实用多元统计分析》译著P401、高惠璇著《应用多元统计分析》。）

这样，在得到一组样本值后，就可以求得因子得分。当因子数m=2时，还可绘制样本点的二维因子得分图，直观描述样本分布情况，从而便于把研究工作引向深入。

§5　因子得分

（*选看）Thompson因子得分的一种简要推导思路：

假设 $\begin{cases} \boldsymbol{x} = \boldsymbol{AF} + \varepsilon \\ \boldsymbol{F} \sim N_m(0, \mathrm{I_m})\ \varepsilon \sim N_p(0, \boldsymbol{D}) \\ \boldsymbol{F} \text{和} \varepsilon \text{独立} \end{cases}$

构造随机向量(F, x)’，并根据多元正态的假设，有：

$$\begin{pmatrix} \boldsymbol{F} \\ \boldsymbol{x} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I}_m & \boldsymbol{0} \\ \boldsymbol{A} & \boldsymbol{I}_p \end{pmatrix} \begin{pmatrix} \boldsymbol{F} \\ \varepsilon \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} \boldsymbol{F} \\ \boldsymbol{x} \end{pmatrix} \sim N_{p+m}\left(\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{I}_m & \boldsymbol{A}' \\ \boldsymbol{A} & \boldsymbol{AA}' + \boldsymbol{D} \end{pmatrix} \right)$$

$$\Rightarrow \boldsymbol{F} \mid \boldsymbol{x} \sim N_m(\mathrm{A'R^{-1}x},\ \mathrm{I_m\text{-}A'R^{-1}A}) \Rightarrow \text{不妨取} \hat{\mathrm{F}} = \mathrm{A'R^{-1}}x$$

例3:人均要素变量因子分析。对我国32个省市自治区的要素状况作因子分析。指标体系中有如下指标:

X1 ： 人口（万人）　　X2 ： 面积（万平方公里）

X3 ： GDP（亿元）　　X4 ： 人均水资源（立方米/人）

X5：人均生物量（吨/人）　X6：万人拥有的大学生数（人）

X7：万人拥有科学家、工程师数（人）

Rotated Factor Pattern

	FACTOR1	FACTOR2	FACTOR3
X1	-0.21522	-0.27397	0.89092
X2	0.63973	-0.28739	-0.28755
X3	-0.15791	0.06334	0.94855
X4	0.95898	-0.01501	-0.07556
X5	0.97224	-0.06778	-0.17535
X6	-0.11416	0.98328	-0.08300
X7	-0.11041	0.97851	-0.07246

旋转后的因子结构

例3：

X1= -0.215F1 -0.273F2 +0.890F3 +e1

X2= 0.639F1 -0.287F2 -0.287F3 +e2

X3= -0.157F1 +0.063F2 +0.948F3 +e3

X4= 0.958F1 -0.015F2 -0.075F3 +e4

X5= 0.972F1 -0.067F2 -0.175F3 +e5

X6= -0.114F1 +0.983F2 -0.083F3 +e6

X7= -0.110F1 +0.978F2 -0.072F3 +e7

	高载荷指标	因子命名
因子F1	X2：面积（万平方公里） X4：人均水资源（立方米/人） X5：人均生物量（吨/人）	自然资源因子
因子F2	X6：万人拥有的大学生数（人） X7：万人拥有的科学家、工程师人数	人力资源因子
因子F3	X1：人口（万人） X3：GDP(亿元)	经济发展总量因子

例3：

因子得分系数矩阵

	F1	F2	F3
X1	0.057	-0.060	0.503
X2	0.227	-0.099	-0.077
X3	0.146	0.129	0.597
X4	0.479	0.112	0.170
X5	0.455	0.074	0.101
X6	0.054	0.486	0.040
X7	0.057	0.485	0.048

于是，因子得分的计算公式：

$$F1=0.057X_1+0.227X_2+0.146X_3+0.479X_4+0.455X_5+0.054X_6+0.057X_7$$

$$F2=-0.060X_1-0.099X_2+0.129X_3+0.112X_4+0.074X_5+0.486X_6+0.485X_7$$

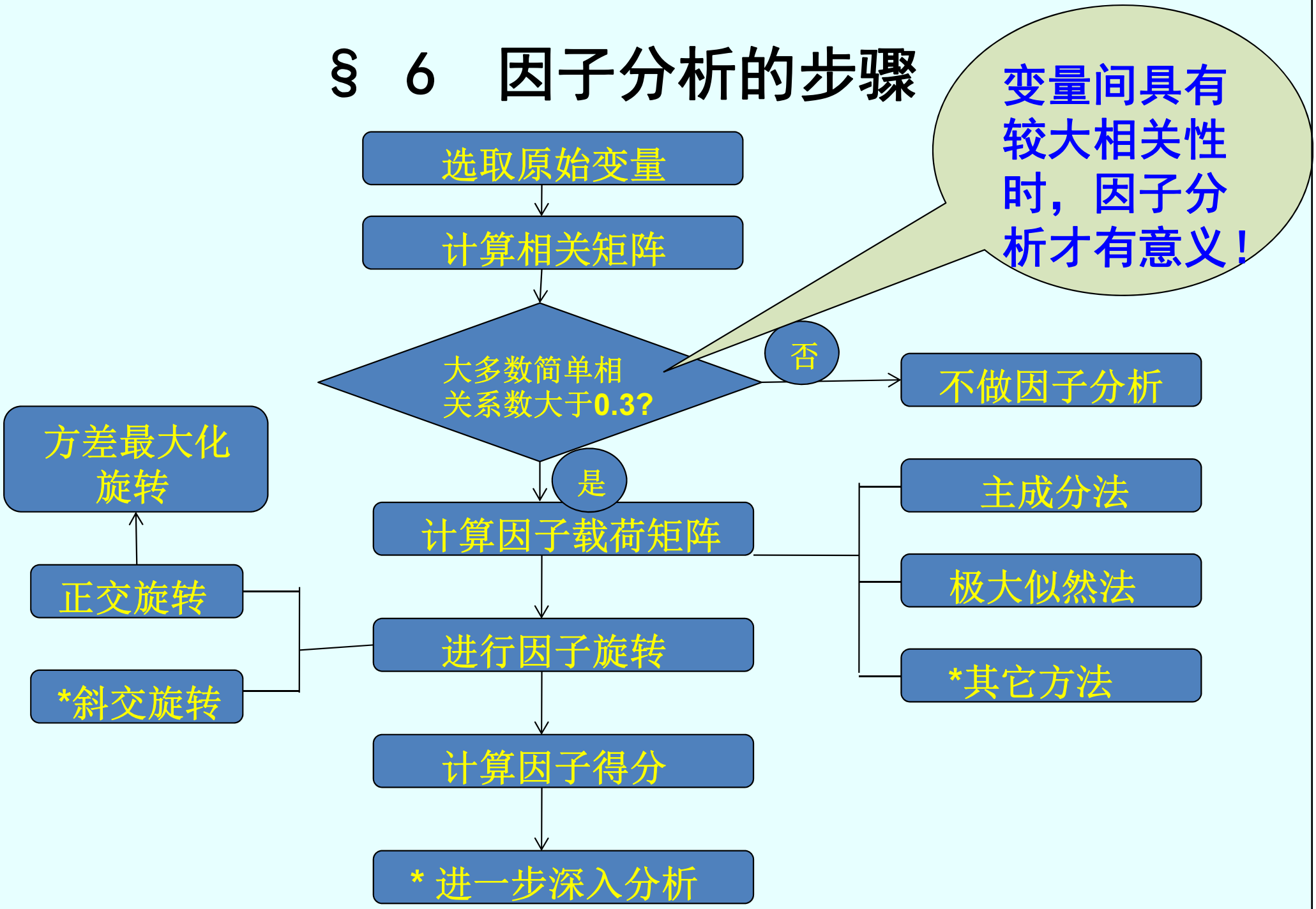
$$F3=0.503X_1-0.077X_2+0.597X_3+0.170X_4+0.101X_5+0.040X_6+0.048X_7$$

例3：

三个因子得分(仅列出部分省市)

样本 (REGION)	F1	F2	F3
北京	-0.081	4.234	-0.379
天津	-0.474	1.317	-0.878
河北	-0.221	-0.358	0.862
山西	-0.482	-0.326	-0.542
内蒙	0.544	-0.666	-0.926
辽宁	-0.205	0.463	0.340
吉林	-0.214	0.106	-0.574
黑龙江	0.108	-0.117	-0.022
…			

§ 6 因子分析的步骤



因子分析的逻辑框图

§ 6 因子分析的步骤

因子分析通常包括以下主要步骤：

1. 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。

2. 计算所选原始变量的相关系数矩阵

相关系数矩阵描述了原始变量之间的相关关系。相关系数矩阵是估计因子结构的基础，可以帮助判断原始变量之间是否存在相关关系，这对因子分析是非常重要的。因为如果所选变量之间均无显著的线性关系，则因子分析的意义不大。

§ 6 因子分析的步骤

因子分析通常包括以下主要步骤：

3. 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综合确定。因子个数可以根据因子的累积方差贡献率>80%而初步确定，但不绝对，有时需根据实际情况，具体问题具体分析。

4. 因子旋转

通过因子旋转使每个原始变量与尽可能少的因子有密切的关系，这样因子解的实际意义更容易解释，并尝试为每个因子（潜变量）赋予有实际意义的名字。

§ 6 因子分析的步骤

因子分析通常包括以下主要步骤：

5. 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，便于聚类更为可视化；再如，做回归分析中的回归自变量。这些延伸应用，同样需要对实际问题具体分析，酌情进行。

§6 因子分析的步骤

- **注意：** 因子分析是十分主观的，在许多出版的资料中，因子分析模型都用少数可阐述因子提供了合理解释。实际上，绝大多数因子分析并没有产生如此明确的结果。不幸的是，评价因子分析质量的法则尚未很好量化。
- 【“哇”原则】：实际分析中，如果在仔细检查因子分析的时候，研究人员能够喊出“哇，我明白这些因子”的时候，就可看着是成功运用了因子分析方法。

§7　R语言中因子分析的常用函数及实例

【教材例9.1】水泥行业上市公司经营业绩因子模型实证分析。

x1:主营业务利润率；**x2:**销售毛利率；**x3:**速动比率

x4:资产负债率；**x5:**主营业务收入增长率；**x6:**营业利润增长率

数据如下表所示：【详见教材P243】

	x1	x2	x3	x4	x5	x6
冀东水泥	33.8	34.75	0.67	59.77	15.49	16.35
大同水泥	27.54	28.04	2.36	35.29	-20.96	-46.45
四川双马	22.86	23.47	0.61	42.83	5.48	-49.22
牡丹江	19.05	19.95	1	48.51	-12.32	-65.99
西水股份	20.84	21.17	1.08	48.45	65.09	54.81
狮头股份	28.14	28.84	2.51	24.52	-6.43	-15.94
太行股份	30.45	31.13	1.02	46.14	6.57	-16.59
海螺水泥	36.29	36.96	0.27	58.31	70.85	117.59
尖峰集团	16.94	17.26	0.61	52.04	9.03	-94.05
四川金顶	28.74	29.4	0.6	65.46	-33.97	-55.02
祁连山	33.31	34.3	1.17	45.8	12.18	39.46
华新水泥	25.08	26.12	0.64	69.35	22.38	-10.2
福建水泥	34.51	35.44	0.38	61.61	23.91	-163.99
天鹅股份	25.52	26.73	1.1	47.02	-4.51	-68.79

【例9.1】水泥行业上市公司经营业绩因子模型实证分析。

```
X=read.table("clipboard",header=T)   #读取数据存入X
cor(X) #计算数据X的相关系数矩阵
```

	x1	x2	x3	x4	x5	x6
x1	1.00000000	0.9991983	-0.09974689	0.18850763	0.2010041	0.29778271
x2	0.99919830	1.0000000	-0.10420434	0.19672979	0.1903570	0.28747808
x3	-0.09974689	-0.1042043	1.00000000	-0.83715637	-0.4087603	0.01518741
x4	0.18850763	0.1967298	-0.83715637	1.00000000	0.2585103	-0.02928244
x5	0.20100410	0.1903570	-0.40876032	0.25851029	1.0000000	0.58029333
x6	0.29778271	0.2874781	0.01518741	-0.02928244	0.5802933	1.00000000

【例9.1】水泥行业上市公司经营业绩因子模型实证分析。

factanal(X, 3, rotation="none")　#极大似然法，未旋转

Call:
factanal(x = X, factors = 3, rotation = "none")
Uniquenesses:

注：共同度=1−Uniquenesses

x1	x2	x3	x4	x5	x6
0.005	0.005	0.005	0.271	0.005	0.548

Loadings:

	Factor1	Factor2	Factor3
x1	0.950	-0.307	
x2	0.948	-0.310	
x3	-0.340	-0.782	0.517
x4	0.363	0.561	-0.531
x5	0.454	0.693	0.556
x6	0.383	0.163	0.527

	Factor1	Factor2	Factor3
SS loadings	2.402	1.623	1.140
Proportion Var	0.400	0.271	0.190
Cumulative Var	0.400	0.671	0.861

The degrees of freedom for the model is 0 and the fit was 1.1422

由结果可以看出，前三个因子所解释的方差占整个方差的**86%**以上，基本能全面地反映：六项财务指标信息。

例9.1：用极大似然法，并进行因子旋转。

factanal(X,3,rotation="varimax")　# varimax旋转的因子分析

Call:
factanal(x = X, factors = 3, rotation = "varimax")

注：共同度=1−Uniquenesses

Uniquenesses:

x1	x2	x3	x4	x5	x6
0.005	0.005	0.005	0.271	0.005	0.548

Loadings:

	Factor1	Factor2	Factor3
x1	0.983		0.155
x2	0.985		0.142
x3		-0.990	-0.124
x4	0.127	0.844	
x5		0.293	0.953
x6	0.210		0.631

	Factor1	Factor2	Factor3
SS loadings	1.998	1.800	1.367
Proportion Var	0.333	0.300	0.228
Cumulative Var	0.333	0.633	0.861

The degrees of freedom for the model is 0 and the fit was 1.1422

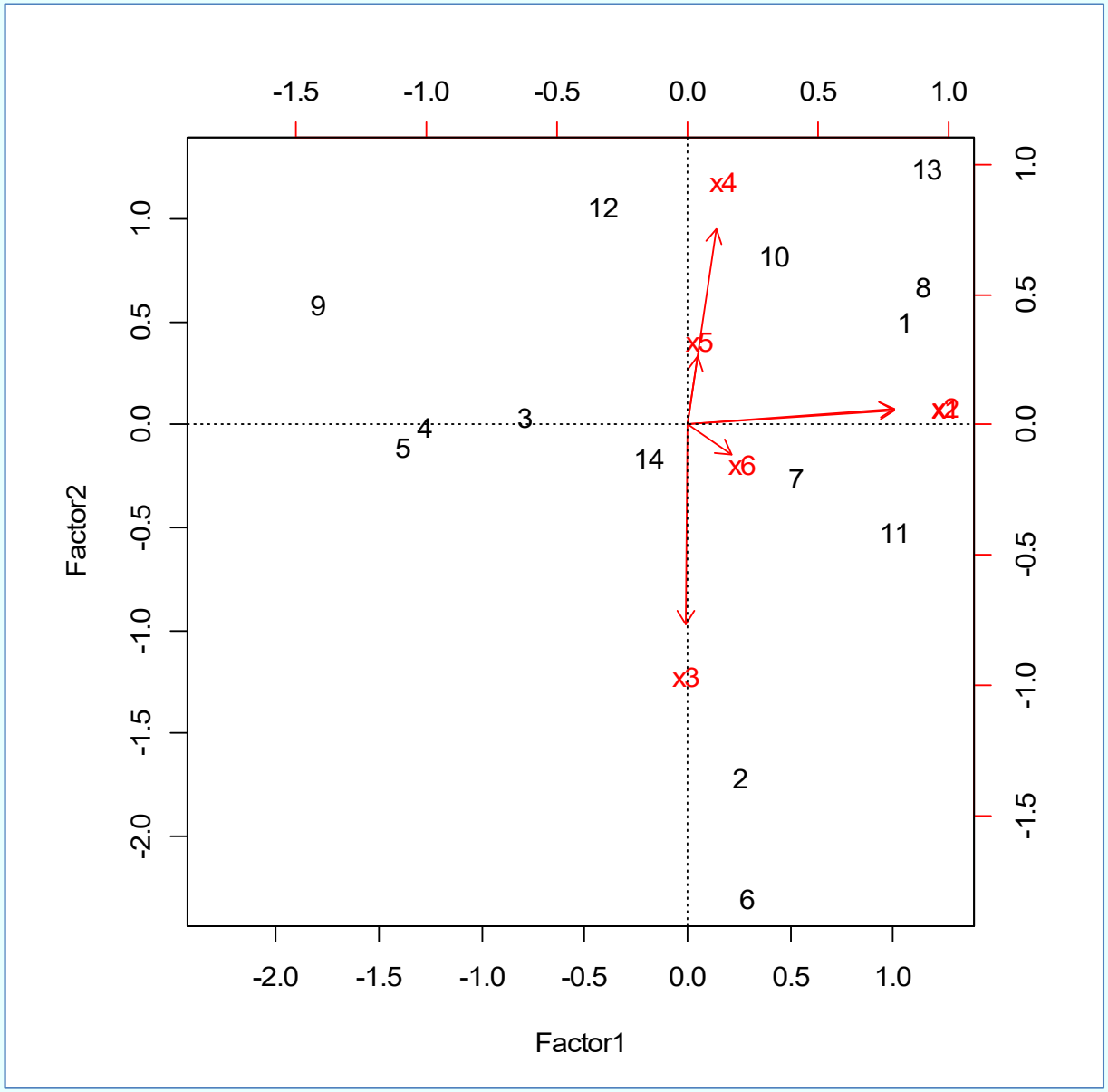
例9.1：用极大似然法并旋转后，计算因子得分。

```
F1 = factanal(X,3, rotation="varimax",
                          scores="regression")
F1$scores
```

	Factor1	Factor2	Factor3
冀东水泥	1.0571	0.49858	-0.01932
大同水泥	0.2508	-1.97182	-0.55062
四川双马	-0.7619	0.61936	-0.35643
牡丹江	-1.2622	0.10831	-0.82490
西水股份	-1.4124	-0.36520	2.09840
狮头股份	0.2993	-2.28407	0.06540
太行股份	0.5368	-0.01725	-0.16548
海螺水泥	1.1383	0.86089	1.85549
尖峰集团	-1.7990	0.62143	-0.20236
四川金顶	0.4397	0.83905	-1.87521
祁连山	1.0220	-0.27756	0.10237
华新水泥	-0.4381	0.53317	0.26013
福建水泥	1.1144	0.91988	0.13561
天鹅股份	-0.1847	-0.08479	-0.52308

例9.1：极大似然法下，绘制的前两个因子的双信息图。

```
biplot(F1$scores, F1$loadings)
abline(h=0,v=0,lty=3)
```



【例9.1】水泥行业上市公司经营业绩因子模型实证分析。

```
library(mvstats) #载入程序包mvstats
factpc(X,3) #主成分法的因子分析，注：此函数默认不旋转。
```

$Vars

	Vars	Vars.Prop	Vars.Cum
Factor1	2.570	0.4283	42.83
Factor2	1.713	0.2855	71.38
Factor3	1.249	0.2082	92.19

$loadings

	Factor1	Factor2	Factor3
x1	0.7829	0.5029	-0.3624
x2	0.7811	0.4964	-0.3756
x3	-0.5786	0.7685	0.0802
x4	0.5951	-0.6990	-0.2415
x5	0.6317	-0.1457	0.6557
x6	0.5084	0.3367	0.6943

由两种分析方法的结果看出：对这个数据集，主成分法要比极大似然法的提取效果好些。因为极大似然法要求数据来自多元正态分布，这个假设有时不满足。

例9.1：用主成分法，并进行因子旋转、计算因子得分。

```
library(mvstats) #载入程序包mvstats
F1=factpc(X,3,rotation="varimax",scores=regression)
```

$`Vars`

	Vars	Vars.Prop	Vars.Cum
Factor1	2.014	33.56	33.56
Factor2	1.938	32.30	65.87
Factor3	1.580	26.33	92.19

$loadings

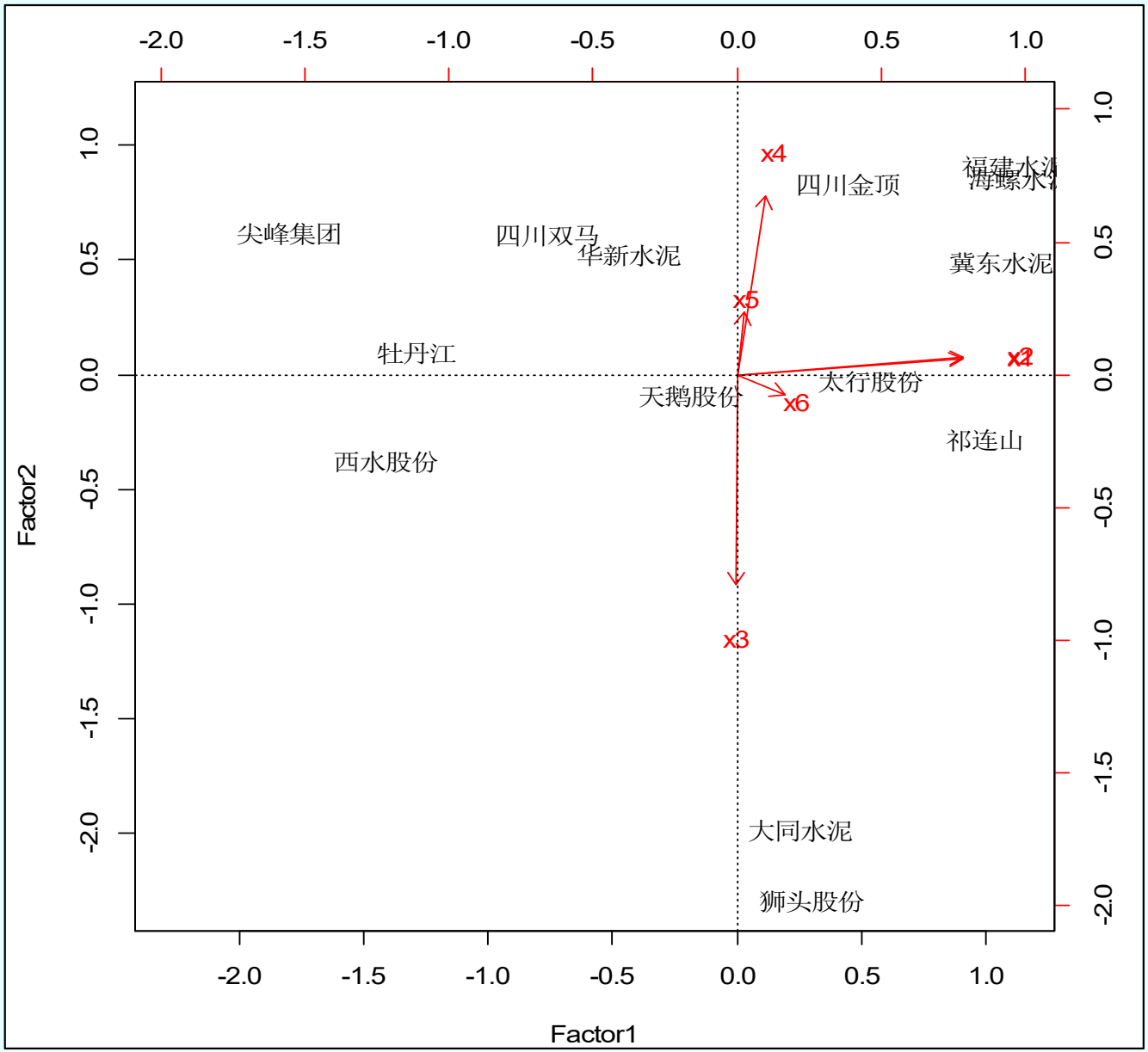
	Factor1	Factor2	Factor3
x1	0.986709	0.07216	0.135305
x2	0.988140	0.07913	0.122314
x3	-0.009491	-0.95685	-0.127000
x4	0.135286	0.93954	0.004538
x5	0.044103	0.32942	0.860082
x6	0.208451	-0.14120	0.889083

$scores

	Factor1	Factor2	Factor3
[1,]	1.0571	0.508465	0.22544
[2,]	0.2509	-1.704706	-0.68039
[3,]	-0.7922	0.052388	-0.14079
[4,]	-1.2794	-0.001121	-0.59625
[5,]	-1.3825	-0.096118	1.91289
[6,]	0.2910	-2.290232	-0.06280
[7,]	0.5235	-0.246292	-0.04099
[8,]	1.1476	0.681631	2.13317
[9,]	-1.7982	0.594084	-0.39758
[10,]	0.4175	0.832941	-1.27718
[11,]	1.0061	-0.507764	0.48519
[12,]	-0.4092	1.074736	0.24757
[13,]	1.1592	1.253210	-1.19980
[14,]	-0.1915	-0.151222	-0.60849

例9.1：主成分法下，绘制的前两个因子的双信息图。

```
biplot(F1$scores,F1$loadings)
abline(h=0,v=0,lty=3)
```



§7　R语言中因子分析的常用函数及实例

更多实例，可参见：
上机课件；
王学民教授-应用多元统计分析**MOOC**课程，等等.

因子分析总结

- 因子分析目的：建立因子模型，对数据降维和简化。
- 基本概念：公共因子、特殊因子、因子载荷、
  变量共同度、方差贡献、因子旋转、因子得分。
- 与主成分分析的区别：

主成分分析：用原始变量的线性组合表示新的综合变量，即主成分；
涉及的只是一般的变量变换，不作为模型来描述；
用主成分法进行系数矩阵的估计。

因子分析：潜变量和随机误差变量的线性组合表示原始变量。
构造因子模型，并伴有几个关键性的假定。
正交因子模型的估计方法不仅限于主成分法；
而且，分析允许做进一步的因子旋转。

- 注：本节主要实现的是R型因子分析（针对变量的相关阵）。

小 结

- 理解并掌握正交因子分析模型的表达式及其含义；
- 理解主成分法求解正交因子分析模型的基本原理；掌握正交因子模型的统计性质。
- 理解因子旋转的作用及意义。
- 掌握R软件进行因子分析的步骤，以及对结果的正确解释方法。

（致谢：本课件及参考资料的部分内容，综合选自以下课程或教材：
中国人民大学出版社-多元统计分析；
中国人民大学六西格玛质量管理研究中心；
清华大学出版社-实用多元统计分析（第6版译著）；
高等教育出版社-多元统计分析及R语言建模；
北京大学出版社-应用多元统计分析；
上海财经大学出版社-应用多元统计分析，
等书籍或相关的MOOC课材料，我校材料及网络开放资源，
等。）

The end!

[附1]：正交旋转后变量的共同度

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

$$\mathbf{B}=(b_{ij})_{p\times p}=(\sum_{l=1}^{m}a_{il}\gamma_{lj})$$

$$h_i^2(\mathbf{B})=\sum_{j=1}^{m}b_{ij}^2=\sum_{j=1}^{m}(\sum_{l=1}^{m}a_{il}\gamma_{lj})^2$$

$$=\sum_{j=1}^{m}\sum_{l=1}^{m}a_{il}^2\gamma_{lj}^2+\sum_{j=1}^{m}\sum_{l=1}^{m}\sum_{\substack{t=1\\ j\neq l}}^{m}a_{il}a_{it}\gamma_{lj}\gamma_{tj}=\sum_{l=1}^{m}a_{il}^2\sum_{j=1}^{m}\gamma_{lj}^2=\sum_{l=1}^{m}a_{il}^2=h_i^2(\mathbf{A})$$

正交旋转后变量的共同度没有发生变化！

[附2]：正交旋转后因子的方差贡献

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

$$\mathbf{B}=(b_{ij})_{p\times p}=(\sum_{l=1}^{q}a_{il}\gamma_{lj})$$

$$S_j^2(\mathbf{B})=\sum_{i=1}^{p}b_{ij}^2=\sum_{i=1}^{p}(\sum_{l=1}^{q}a_{il}\gamma_{lj})^2$$

$$=\sum_{i=1}^{p}\sum_{l=1}^{q}a_{il}^2\gamma_{lj}^2+\sum_{i=1}^{p}\sum_{l=1}^{q}\sum_{\substack{t=1\\t\neq l}}^{q}a_{il}a_{it}\gamma_{lj}\gamma_{tj}\quad=\sum_{i=1}^{p}a_{il}^2\sum_{l=1}^{q}\gamma_{lj}^2=S_j^2(\mathbf{A})\sum_{l=1}^{q}\gamma_{lj}^2$$

正交旋转后因子的方差贡献发生了变化！

[附3]：

Charles Spearman 提出因子分析的数据场景举例。

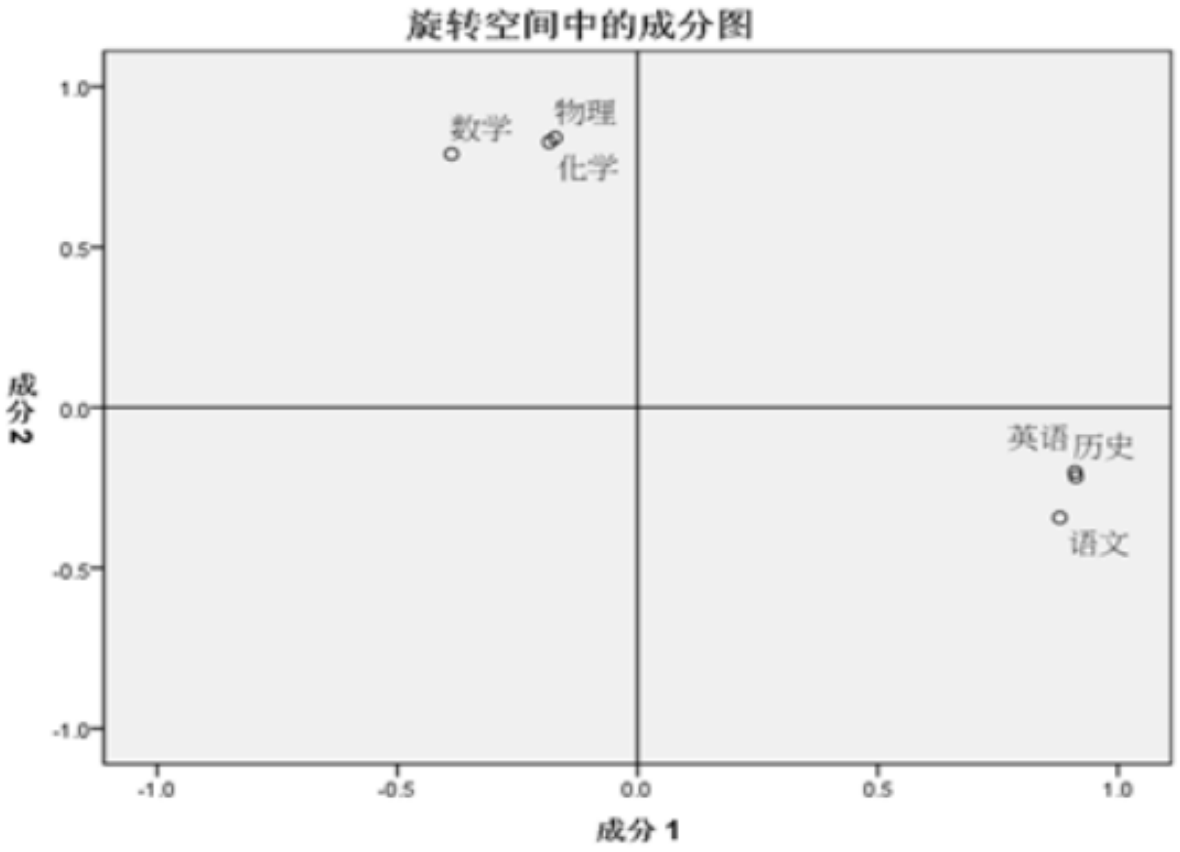
数学	物理	化学	语文	历史	英语
100	100	100	59	73	67
99	100	99	53	63	60
87	84	100	74	81	76
91	85	100	70	65	76
87	98	87	68	78	64
85	91	95	63	76	66
79	95	83	89	89	79

解释的总方差

成份	初始特征值			提取平方和载入			旋转平方和载入		
	合计	方差的 %	累积 %	合计	方差的 %	累积 %	合计	方差的 %	累积 %
1	3.735	62.254	62.254	3.735	62.254	62.254	2.649	44.147	44.147
2	1.133	18.887	81.142	1.133	18.887	81.142	2.220	36.995	81.142
3	.457	7.619	88.761						
4	.323	5.376	94.137						
5	.199	3.320	97.457						
6	.153	2.543	100.000						

旋转成份矩阵[a]

	成份	
	1	2
数学	-.387	.790
物理	-.172	.841
化学	-.184	.827
语文	.879	-.343
历史	.911	-.201
英语	.913	-.216



$$X_i = \alpha_{i1}F_1 + \alpha_{i2}F_2 + \varepsilon_i \qquad i = 1, \cdots, 6$$