第八章、因子分析

December 24, 2018

8.1 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，并用少数几个潜在变量来表示其基本的数据结构。这几个潜在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，而潜在变量是不可观测的，称为潜在因子。

案例一： 基于因子分析的区域生活垃圾处理评价，《合作经济与科技》，2012年6月。

研究背景：对我国不同省份的生活垃圾处理能力进行因子分析，以反映我国不同地区的生活垃圾综合处理能力。研究造成不同地区垃圾处理能力差异的主要因素，为提高我国生活垃圾综合处理水平提供参考建议。

- 指标选取:
  1. 生活垃圾清运总量X_1;
  2. 无害化处理工厂数量X_2;
  3. 无害化处理能力X_3;
  4. 无害化处理量X_4;
  5. 粪便清运量X_5;
  6. 粪便无害化处理量X_6;
  7. 生活垃圾无害化处理率X_7;

- 主成分分析后，构成新的因子.
  1. F_1 共因子在无害化处理能力、无害化处理量等方面解释较强，故命名为生活垃圾处理能力优势因子;
  2. F_2 在生活垃圾处理结构等方面解释较强，命名为处理结构优势因子;
  3. F_3 在生活垃圾无害化处理率、无害化处理厂数等方面解释较强, 命名为处理设施优势因子.

- 处理结果分析
  1. 从F_1的得分来看，处于前列的几个省区来看，广东、山东、浙江、江苏等省经济发展水平、城市化水平较高，人口数量多，人口密度大，生活垃圾产生的数量大，处理量也大，处理能力较强。
  2. 从F2的得分来看，北京、上海、内蒙等地得分较高，上述地区在生活垃圾中的处理结构上有较为明显的优势。
  3. 从F3的得分来看，天津、重庆、北京等直辖市得分较高，居民的生活垃圾中，说明了上述地区在生活垃圾处理体系构建较为完整，从垃圾产生的源头到垃圾回收再利用等环节的设施较为完备。

案例二：企业形象或品牌形象研究

研究背景： 消费者可以通过一个有24个指标构成的评价体系，评价百货商场的24 个方面的优劣。但消费者主要关心的是三个方面，即商店的环境、商店的服务和商品的价格。

研究方法： 因子分析方法可以通过24个变量，找出反映商店环境、商店服务水平和商品价格的三个潜在的因子，对商店进行综合评价。而这三个公共因子可以表示为：

$$x_i = \mu_i + \alpha_{i1}F_1 + \alpha_{i2}F_2 + \alpha_{i3}F_3 + \epsilon_i, \quad i = 1, \ldots, 24.$$

称F_1, F_2, F_3 是不可观测的潜在因子。24个变量共享这三个因子，但是每个变量又有自己的个性，不被包含的部分ϵ_i, 称为特殊因子。

因子分析与其它统计方法的联系和区别：

1. 因子分析与回归分析差异： 因子分析中的因子是一个比较抽象的概念，而回归因子有非常明确的实际意义；
2. 主成分分析分析与因子分析的差异：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。
3. 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；
4. 因子分析：潜在变量和随机误差的线性组合表示原始变量。

一、数学模型

设$\{X_i, i = 1, \ldots, p\}$ 为p 个变量，因子分析模型表示为

$$X_i = \mu_i + \alpha_{i1}F_1 + \cdots + \alpha_{im}F_m + \epsilon_i, \quad i = 1, \ldots, p.$$

或

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} + \begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1m} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ \alpha_{p1} & \alpha_{p2} & \cdots & \alpha_{pm} \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_p \end{bmatrix}$$

或

$$\mathbf{X} - \boldsymbol{\mu} = \mathbf{AF} + \boldsymbol{\epsilon}.$$

1. 称$F_1, \ldots, F_m$ 为公共因子，是不可观测的变量，他们的系数称为因子载荷。
2. ϵ是特殊因子，是不能被前m个公共因子包含的部分。

满足：

(1) $\mathrm{Cov}(\mathbf{F}, \boldsymbol{\epsilon}) = 0$, 即公共因子$F$ 与特殊因子ϵ 不相关;

(2) $D(\mathbf{F}) = I$, I 为单位矩阵，即公共因子$F_1, \ldots, F_m$互不相关，方差为1;

(3) $D(\boldsymbol{\epsilon}) = diag(\sigma_1^2, \ldots, \sigma_p^2)$, 即特殊因子不相关(或进一步假设$\epsilon_i \sim N(0, \sigma_i^2)$.

二、因子分析模型的性质

1.原始变量$\mathbf{X}$的协方差矩阵的分解.

$$\mathbf{X}-\boldsymbol{\mu}=\mathbf{AF}+\boldsymbol{\epsilon}.$$

所以有

$$\mathsf{Var}[\mathbf{X}-\boldsymbol{\mu}]=\mathbf{A}\mathsf{Var}[\mathbf{F}]\mathbf{A}'+\mathsf{Var}[\boldsymbol{\epsilon}].$$

即

$$\Sigma_X=AA'+D.$$

其中：A是因子模型的系数，$D=\mathsf{Var}[\boldsymbol{\epsilon}]=diag(\sigma_1^2,\ldots,\sigma_p^2)$.
D 的主对角线上的元素值越小，则公共因子共享的成分越多。

2. 因子不受计量单位的影响.

将原始变量X做变换$X^* = CX$, 其中$C = \mathrm{diag}(c_1, c_2, \ldots, c_p)$, $c_i > 0$, $i = 1, \ldots, p$.

$$C(X - \mu) = C(AF + \epsilon),$$

令$\mu^* = C\mu$, $A^* = CA$, $\epsilon^* = C\epsilon$. 将模型改写:

$$X^* - \mu^* = A^*F + \epsilon^*,$$

满足:

(1) $\mathsf{E}[\mathbf{F}] = 0, \mathsf{E}[\boldsymbol{\epsilon}^*] = 0, \mathsf{Var}[\mathbf{F}] = 1.$

(2) $\mathsf{Var}(\boldsymbol{\epsilon}^*) = \mathrm{diag}((\sigma_1^*)^2, \ldots, (\sigma_p^*)^2)$, $(\sigma_i^*)^2 = c_i^2\sigma_i^2$, $i = 1, \ldots, p.$

(3) $\mathsf{Cov}(\mathbf{F}, \boldsymbol{\epsilon}^*) = \mathsf{E}[\mathbf{F}\epsilon^*] = 0.$

3. 因子载荷不是惟一的.

设T为一个$p \times p$的正交矩阵，令$\mathbf{A}^* = \mathbf{AT}$，$\mathbf{F}^* = \mathbf{T}'\mathbf{F}$，则模型可以表示为

$$\mathbf{X} - \mu = \mathbf{A}^*\mathbf{F}^* + \boldsymbol{\epsilon},$$

且满足条件因子模型的条件：

$$(1) \quad \mathsf{E}[\mathbf{F}^*] = \mathsf{E}[\mathbf{T}'\mathbf{F}] = \mathbf{0}, \mathsf{E}[\boldsymbol{\epsilon}] = \mathbf{0}.$$

$$(2) \quad \mathsf{Var}[\mathbf{F}^*] = \mathsf{Var}[\mathbf{T}'\mathbf{F}^*] = \mathbf{T}'\mathsf{Var}[\mathbf{T}'\mathbf{F}^*]\mathbf{T} = 1.$$

$$(3) \quad \mathsf{Var}(\boldsymbol{\epsilon}) = diag(\sigma_1^2, \ldots, \sigma_p^2),$$

$$(4) \quad \mathsf{Cov}(\mathbf{F}^*, \boldsymbol{\epsilon}) = \mathsf{E}[\mathbf{F}^*\epsilon'] = 0.$$

三、因子载荷矩阵中的几个统计特征

1. 因子载荷a_{ij}的统计意义.

因子载荷a_{ij}是第i个变量与第j个公共因子的相关系数, 模型为

$$X_i = a_{i1}F_1 + \cdots + a_{im}F_m + \epsilon_i,$$

在上式的左右两边乘以F_i, 再求数学期望

$$\mathsf{E}[X_iF_j] = a_{i1}\mathsf{E}[F_1F_j] + \cdots + a_{ij}\mathsf{E}[F_j^2] + \cdots + a_{im}\mathsf{E}[F_mF_j] + \mathsf{E}[\epsilon_iF_j],$$

根据公共因子的模型性质，有$\gamma_{X_i,F_j} = a_{ij}$ （载荷矩阵中第i行，第j列的元素）反映了第i个变量X_i与第j个公共因子F_j的相关程度。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h_i^2 是因子载荷矩阵的第i行的元素的平方和, 即$h_i^2 = \sum_{j=1}^{m} a_{ij}^2$.

统计意义： 对因子模型

$$X_i = a_{i1}F_1 + \cdots + a_{im}F_m + \epsilon_i,$$

两边求方差，

$$\mathsf{Var}[X_i] = a_{i1}^2\mathsf{Var}[F_1] + \cdots + a_{im}^2\mathsf{Var}[F_m] + \mathsf{Var}[\epsilon_i].$$

因此，有: $1 = \sum_{j=1}^{m} a_{i1}^2 + \sigma_i^2 = h_i^2 + \sigma_i^2$.

- 所有的公共因子和特殊因子对变量X_i的贡献为1,
  即$h_i^2 + \sigma_i^2 = 1$.
- 如果共同度$h_i^2 = \sum_{j=1}^{m} a_{i1}^2$ 非常靠近1, 则σ_i^2 很小, 则因子分析的效果好, 从原变量空间到公共因子空间的转化性质好.

3. 公共因子F_j 方差贡献的统计意义.

因子载荷矩阵中各列元素的平方和

$$q_j^2 = \sum_{i=1}^{p} a_{ij}^2, \quad j = 1, 2, \cdots, p.$$

称为第j个公共因子F_j 对所有分量$(X_1, \ldots, X_p)$ 的方差贡献和。

1. q_j^2表示第j个公共因子F_j对所有的分量$X = (X_1, \ldots, X_p)$ 的总影响.
2. q_j^2 越大, 说明F_j对X的贡献越大.
3. 把载荷矩阵A的各列平方和都计算出来, 使得相应的贡献有顺序：$q_1^2 \geq \ldots \geq q_m^2$, 此为依据找出最有影响的公共因子的相对重要性. 因此, 因子分析的关键是如何估计因子载荷矩阵.

4. 正交因子模型的几何解释.

把m 个公共因子和p个特殊因子看成是$m+p$个相互正交的单位向量, 由此构成$m+p$维空间的一个直角坐标系, 并称因子空间.

- 变量X_i 可以用因子空间中向量

$$P_i = (a_{i1}, \ldots, a_{im}, 0, \ldots, \sigma_i, \ldots, 0)',$$

表示，其中σ_i 是X_i 的对应于自己的特殊因子轴上的载荷.

- P_i 的长度等于1, 即

$$||P_i|| = \sqrt{a_{i1}^2 + \cdots + a_{im}^2 + \sigma_i^2} = 1.$$

- P_i与各因子轴F_j 的夹角余弦为

$$\cos\langle P_i, F_j\rangle = \|P_i\|\cos\langle P_i, F_j\rangle = a_{ij} = r_{P_iF_j},$$

这表明了P_i 与各公共因子的夹角余弦就等于其相应的坐标, 也就是等于变量X_i与各公共因子的相关系数.

- 因子空间中分别表示变量X_i 和X_j 的向量P_i 与P_j 的夹角余弦为他们的内积

$$\cos\langle P_i, P_j\rangle = \frac{P_i'P_j}{\|P_i\|\|P_j\|} = P_i'P_j = \sum_{t=1}^{m} a_{it}a_{jt} = r_{X_iX_j}$$

它恰好等于变量X_i 与X_j 的相关系数.

因子载荷矩阵的估计方法

一、主成分分析法

设随机向量$\mathbf{X}=(X_1,\ldots,X_p)'$ 的均值为$\boldsymbol{\mu}$, 协方差为$\boldsymbol{\Sigma}$，其特征根为$\lambda_1 \geq \ldots \geq \lambda_p \geq 0$，$\mathbf{U}=(\mathbf{u}_1,\ldots,\mathbf{u}_p)$ 为对应的标准化特征向量，则

$$\begin{aligned}\boldsymbol{\Sigma} &= \mathbf{U}\mathrm{diag}(\lambda_1,\ldots,\lambda_p)\mathbf{U}' \\ &= \lambda_1\mathbf{u}_1\mathbf{u}_1'+\cdots+\lambda_p\mathbf{u}_p\mathbf{u}_p' \\ &= \left(\sqrt{\lambda_1}\mathbf{u}_1,\ldots,\sqrt{\lambda_p}\mathbf{u}_p\right)\begin{pmatrix}\sqrt{\lambda_1}\mathbf{u}_1' \\ \vdots \\ \sqrt{\lambda_p}\mathbf{u}_p'\end{pmatrix}.\end{aligned}$$

上式给出的Σ 表达式是精确的，但在因子模型中没有价值.

因子分析目标：寻求用少数几个公共因子解释.

处理方法：略 去后面的$m+1 \sim p$ 项，有

$$\begin{aligned}
\boldsymbol{\Sigma} &\approx \hat{\mathbf{A}}\hat{\mathbf{A}}' + \hat{\mathbf{D}} \\
&= \lambda_1 \mathbf{u}_1 \mathbf{u}_1' + \cdots + \lambda_m \mathbf{u}_m \mathbf{u}_m' + \hat{\mathbf{D}} \\
&= [\sqrt{\lambda_1}\mathbf{u}_1, \ldots, \sqrt{\lambda_m}\mathbf{u}_m] \begin{pmatrix} \sqrt{\lambda_1}\mathbf{u}_1' \\ \vdots \\ \sqrt{\lambda_m}\mathbf{u}_m' \end{pmatrix} + \hat{\mathbf{D}}.
\end{aligned}$$

其中：$\hat{\mathbf{D}} = diag(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_p^2),\ \hat{\sigma}_i^2 = 1 - \sum_{j=1}^{m} a_{ij}^2.$

因子模型的主成分解

$$\begin{cases} \widehat{A} = \sqrt{\lambda_1}\mathbf{u}_1 + \cdots + \sqrt{\lambda_m}\mathbf{u}_m = (a_{ij})_{p\times m}, \\ \widehat{\sigma}_i^2 = 1 - \sum_{j=1}^m a_{ij}^2. \end{cases}$$

主成分分解有一个假定，即模型中的$m+1 \sim p$ 项公因子是不重要的, 因而, 从Σ的分解中忽略了$m+1 \sim p$ 项公因子，并将其归为残差. 即$E = \Sigma - (\hat{A}\hat{A}' + \hat{D}) = (\epsilon_{ij})_{p\times p}$为残差,

$$Q(m) = \sum_{i=1}^{p}\sum_{j=1}^{p}\epsilon_{ij}^2$$

为残差平方和, 有$Q(m) \leq \lambda_{m+1} + \ldots + \lambda_p$.

公因子数m的选取的方法：

- 根据实际问题的意义和专业理论知识来确定，
- 用确定主成分的个数的原则

$$\frac{\lambda_1 + \cdots + \lambda_m}{\lambda_1 + \cdots + \lambda_p} \geq P_0$$

一般$P_0 \geq 0.7$.

上面的讨论是针对总体的. 如果是样本,则用样本协方差阵S（或采用样本相关阵R）来代替Σ.

主成分估计的具体步骤：

1. 由样本计算出样本相关阵R(如果样本标准化, 则也可用样本协方差阵).
2. 求样本相关阵R的特征根和标准化特征向量，其特征根为$\lambda_1 \geq \ldots \geq \lambda_p \geq 0$，$\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_p)$ 为对应的标准化特征向量.
3. 根据特征根$\frac{\lambda_1+\cdots+\lambda_m}{\lambda_1+\cdots+\lambda_p} \geq P_0$, 确定公共因子个数$m$,
4. 求因子模型的因子载荷矩阵$\hat{A} = \sqrt{\lambda_1}\mathbf{u}_1 + \cdots + \sqrt{\lambda_m}\mathbf{u}_m$.
5. 根据因子载荷矩阵, 计算共同度$h_i^2 = \sum_{j=1}^{m} a_{ij}^2$.
6. 求特殊因子方差, $\sigma_i^2 = 1 - h_i^2$.
7. 结合专业知识，对m个公共因子作解释, 说明这些公共因子的意义.

二、主因子解

主因子解也可以看成是主成分方法的一种修正.

第一步：$R = AA' + D$, 有$R - D = AA' =: R^*$ 称为约相关阵.

第二步：假设已经得出特殊方差的初始估计$(\widehat{\sigma}_i^*)^2$, 可求出相应的共同度

$$(h_i^*)^2 = 1 - (\widehat{\sigma}_i^*)^2.$$

重新给出约相关矩阵R^*

$$R^* = \begin{pmatrix} (h_1^*)^2 & r_{12} & \cdots & r_{1p} \\ r_{21} & (h_2^*)^2 & \cdots & r_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ r_{p1} & r_{p2} & \cdots & (h_p^*)^2 \end{pmatrix}$$

第三步：计算R^*的特征根和单位正交特征向量, 取前m个特征根$\lambda_1^* \geq \ldots \geq \lambda_m^* \geq 0$,相应的特征向量为$(\mathbf{u}_1^*, \ldots, \mathbf{u}_m^*)$。因此，因子模型的主因子解为

$$\begin{cases} \widehat{A} = \sqrt{\lambda_1^*}\mathbf{u}_1^* + \cdots + \sqrt{\lambda_m^*}\mathbf{u}_m^*, \\ \widehat{\sigma}_i^2 = 1 - \sum_{j=1}^m a_{ij}^2. \end{cases}$$

第四步：重新计算约相关阵.....直到解平稳为止。

在实际中特殊因子的方差是未知的，公因子的方差，即共同度$(h_i^*)^2$也是未知的. 因此，常采用迭代主因子法.

迭代主因子法具体步骤如下:

1. 初始值利用主成分给出,
2. 利用上面的方法求出公共因子载荷矩阵, 并计算特殊因子方差，得出主因子解,
3. 重新根据特殊因子方差, 给出约相关阵R^*,
4. 重新利用上面的方法求出公共因子载荷矩阵, 并计算特殊因子方差，得出主因子解,
5. 重复上述过程, 直到解稳定为止.

公因子方差的初始估计的几种方法：

1. h_i^2 取为第i个变量与其他所有变量的多重相关系数的平方；
2. h_i^2 取为第i个变量与其他所有变量的相关系数中绝对值的最大值；
3. 取$h_i^2 = 1$，它等价于主成分解.

三、极大似然估计

如果假定公共因子F和特殊因子ϵ服从正态分布，那么可以得到因子载荷和特殊因子方差的极大似然估计. 设$X_1, \ldots, X_n$ 为来自正态总体$N_p(\mu, \Sigma)$的随机样本.

$$\Sigma_X = AA' + D$$

则似然函数为

$$L(\mu, A, D) = \prod_{i=1}^{d} \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x_i} - \mu)'\mathbf{\Sigma^{-1}}(\mathbf{x_i} - \mu)\right].$$

由极大似然估计能得出$\Sigma = AA' + D$的估计. 用样本协方差阵S(极大似然估计）代替总体的协方差阵Σ. 用求极值的方法, 给出如下方程,

$$\begin{cases} \widehat{\mu} = \overline{X}, \\ S\widehat{D}^{-1}\widehat{A} = \widehat{A}(I + \widehat{A}'\widehat{D}^{-1}\widehat{A}), \\ \widehat{D} = \mathrm{diag}(S - \widehat{A}\widehat{A}'). \end{cases}$$

根据方程给出因子载荷矩阵A 和特殊因子方差D的解.
但上面的方程组并不能给出唯一的A 和D 的估计. 因此, 在求解$\widehat{A}$ 和$\widehat{D}$ 时, 要加上一个唯一性条件：

$$\widehat{A}'\widehat{D}^{-1}\widehat{A} = \Lambda,$$

其中, Λ是一个对角矩阵。

在实际的计算中，由方程组迭代来求极大似然估计$\widehat{A}$ 和$\widehat{D}$ 的解.

$$\begin{cases} S\widehat{D}^{-1}\widehat{A} = \widehat{A}(I + \widehat{A}'\widehat{D}^{-1}\widehat{A}), \\ \widehat{D} = \mathrm{diag}(S - \widehat{A}\widehat{A}'). \end{cases}$$

即选给出初始的$\widehat{D}$的值，然后由第一个方程给出$\widehat{A}$ 的估计，然后再由第二个方程给出$\widehat{D}$ 的估计.....迭代进行下去.

R语言上机：三种估计方法的函数

- 主成分估计方法(factor.analy1)
- 主因子估计法(factor.analy2)
- 极大似然估计方法(factor.analy3)

例8.3.1 盐泉水化学分析资料的因子分析.

因子旋转

一、为什么要因子旋转？

建立了因子分析模型的目的不仅仅要找出公共因子以及对变量进行分组，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惟一的，所以应该对因子载荷阵进行旋转。使因子载荷阵的结构简化，使载荷矩阵中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
- 等量最大法

案例：奥运会十项全能运动项目得分数据的因子分析

(1) 选择的变量为：

- 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}

(2)收集资料后得相关阵R

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

(3) 因子分析后得载荷矩阵

变量	F_1	F_2	F_3	F_4	共同度
X_1	0.691	0.217	-0.58	-0.206	0.84
X_2	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。

(4) 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。

- 百米跑X_1，跳远X_2和400米跑X_5，需要腿部爆发力的项目在F_1 有较大的载荷，可以称为短跑速度因子；
- 铅球X_3，铁饼X_7和标枪X_9在F_2 上有较大的载荷，可以称为爆发性臂力因子；
- 百米跨栏X_6，撑杆跳远X_8，跳远X_2和为跳高X_4在F_3 上有较大的载荷，称为爆发腿力因子；
- 1500米跑主要体现在F_4上，称F_4为长跑耐力因子。

二、因子旋转理论

设因子模型：

$$\mathbf{X} - \boldsymbol{\mu} = \mathbf{AF} + \boldsymbol{\epsilon}.$$

满足：

- $\mathsf{Cov}(\mathbf{F}, \boldsymbol{\epsilon}) = 0$, 公共因子$F$和特殊因子$\epsilon$不相关;
- $D(\mathbf{F}) = I$, I 为单位矩阵，即公共因子$F_1, \ldots, F_m$互不相关，方差为1;
- $D(\boldsymbol{\epsilon}) = diag(\sigma_1^2, \ldots, \sigma_p^2)$, 即特殊因子不相关(或进一步可假设$\epsilon_i \sim N(0, \sigma_i^2)$).

如果$\boldsymbol{\mu} = 0$, 对F作正交变换$Z = \Gamma' F$

$$\mathbf{X} = A\Gamma\mathbf{Z} + \boldsymbol{\epsilon}.$$

满足:

- $D(\mathbf{Z}) = D(\Gamma' F) = \Gamma' D(F)\Gamma = I_m$, 即新公共因子$Z_1, \ldots, Z_m$互不相关，方差为1;
- $\mathsf{Cov}(\mathbf{Z}, \boldsymbol{\epsilon}) = 0$, 即特殊因子$\mathbf{Z}$与公共因子$\epsilon$不相关;
- $D(\mathbf{X}) = D(A\Gamma Z) + D(\boldsymbol{\epsilon}) = A\Gamma D(\mathbf{Z})\Gamma' A' + D = AA' + D$, 即特殊因子不相关, $\mathsf{E}[\epsilon_i^2] = \sigma_i^2$. （ 如果有正态的假设, 则$\epsilon_i \sim N(0, \sigma_i^2)$）).

说明$\mathbf{Z}$ 也是公因子向量.

三、变换后因子的共同度性质(因子载荷方差)

$A = (a_{ij})_{p\times m}$为公因子向量F的因子载荷矩阵，

$$h_i^2 = \sum_{t=1}^{m} a_{it}^2, \quad i = 1, \cdots, p$$

为变量X_i的共同度.

如果A的每一列(即因子载荷向量)数值越分散，相应的因子载荷向量的方差越大. 令

$$d_{ij}^2 = \frac{a_{ij}^2}{h_i^2}, \quad i = 1, \ldots, p, \; j = 1, \ldots, m.$$

定义第j列的方差为

$$V_j = \sum_{i=1}(d_{ij}^2 - \bar{d}_j)^2/p = p\sum_{i=1}^{p}\frac{a_{ij}^4}{h_i^4} - \left(\sum_{i=1}^{p}\frac{a_{ij}^2}{h_i^2}\right)^2,$$

其中$\bar{d}_j = \frac{1}{p}\sum_{i=1}^{p} d_{ij}^2$, $j = 1, \ldots, m$.

相应的因子载荷矩阵A的方差

$$V = \sum_{j=1}^{m} V_j = \frac{1}{p^2}\left\{\sum_{j=1}^{m}\left[p\sum_{i=1}^{p}\frac{a_{ij}^4}{h_i^4} - \left(\sum_{i=1}^{p}\frac{a_{ij}^2}{h_i^2}\right)^2\right]\right\}.$$

若V_j值越大，A的第j个因子载荷向量数值越分散.

四、旋转方法

(1)方差最大的正交旋转

方差最大的正交旋转就是要找一个正交矩阵Γ，使得$B = A\Gamma$的方差极大.

记$B = (b_{ij})_{p\times m}$, 则载荷矩阵B的方差为

$$V = \frac{1}{p^2}\left\{\sum_{j=1}^{m}\left[p\sum_{i=1}^{p}\frac{b_{ij}^4}{h_i^4} - \left(\sum_{i=1}^{p}\frac{b_{ij}^2}{h_i^2}\right)^2\right]\right\}.$$

采用求极值的方法给出b_{ij}的值。

设$m=2$, 则正交矩阵

$$\Gamma=\begin{bmatrix}\cos\phi & -\sin\phi\\ \sin\phi & \cos\phi\end{bmatrix}$$

$$A\Gamma=\begin{bmatrix}a_{11}\cos\phi+a_{12}\sin\phi & -a_{11}\sin\phi+a_{12}\cos\phi\\ \vdots & \vdots\\ a_{p1}\cos\phi+a_{p2}\sin\phi & -a_{p1}\sin\phi+a_{p2}\cos\phi\end{bmatrix}=\begin{bmatrix}b_{11} & b_{12}\\ \vdots & \vdots\\ b_{p1} & b_{p2}\end{bmatrix}$$

$$V=\frac{1}{p^2}\left\{\sum_{j=1}^{2}\left[p\sum_{i=1}^{p}\frac{b_{ij}^4}{h_i^4}-\left(\sum_{i=1}^{p}\frac{b_{ij}^2}{h_i^2}\right)^2\right]\right\}.$$

求V的偏导并令其为0

$$\frac{\partial V}{\partial \phi} = 0.$$

整理后得

$$\tan 4\phi = \frac{d - 2\alpha\beta/p}{c - (\alpha^2 - \beta^2)/p}.$$

各系数的意义见书page310.

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.

(2) 四次方最大旋转

- 四次方最大旋转是从简化载荷矩阵的行出发, 通过旋转初始因子, 使每个变量只在一个因子上又较高的载荷, 而在其它的因子上尽可能低的载荷。如果每个变量只在一个因子上有非零的载荷，这时的因子解释是最简单的。
- 四次方最大法通过使因子载荷矩阵中每一行的因子载荷平方的方差达到最大。
- 简化的准则:

$$Q = \sum_{i=1}^{p}\sum_{j=1}^{m}\left(b_{ij}^2 - \frac{1}{m}\right)^2.$$

(3)等量最大法

- 等量最大法把四次方最大法和方差最大法结合起来求Q和V的加权平均最大.
- 简化的准则:

$$E=\sum_{i=1}^{p}\sum_{j=1}^{m}b_{ij}^{4}-\gamma\sum_{i=1}^{p}\left(\sum_{j=1}^{m}b_{ij}^{2}\right)^{2}\Big/p.$$

关于因子旋转

- **正交因子：** 前面讨论的因子旋转后得到的公共因子之间是不相关的, 即采用的是正交旋转.
- **斜交因子：** 在实际中公共因子之间可能是有相关关系, 即这时候公共因子之间是相关的, 这时的因子模型称为斜交因子模型, 公共因子称为斜交公因子.
- **稀疏因子：** 如果变量很多, 希望得出的模型比较简洁, 则可采用稀疏因子模型的方法. 其主要思想为如果某个变量前的系数很小, 则有二种可能：(1)确实对这个因子有影响，但影响很小; (2)实际应该为零, 即对因子没有影响. 无论是哪种情形除去这个变量对接下去模型的讨论是有积极意义的.

因子得分

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。

公因子得分的计算方法很多种, 下面介绍几种常见的因子得分的计算方法.

(1)加权最小二乘法

考虑公因子F对随机变量X满足如下的方程

$$\mathbf{X} = \mathbf{AF} + \epsilon.$$

假设:

- (1) 因子载荷矩阵**A**已知;
- (2) 特殊因子方差已知, 即$\mathsf{Var}[\epsilon] = D = diag(\sigma_1^2, \cdots, \sigma_p^2)$, σ_i^2已知, 一般不相等.

采用加权最小二乘法去估计公共因子**F**的值.

- 用误差方差的倒数作为权重的误差平方和作为损失函数：

$$L(F_1,\ldots,F_m)=\sum_{i=1}\frac{\epsilon_i^2}{\sigma_i^2}=(X-AF)'D^{-1}(X-AF)=:\psi(F).$$

由$\frac{\partial\psi(F)}{\partial F}=0$,得出

$$\widehat{F}=(A'D^{-1}A)A'D^{-1}X.$$

(2) Bartlett得分

假设$X \sim N_p(AF, D)$, X的似然函数的对数为

$$L(F) = -\frac{1}{2}(X - AF)'D^{-1}(X - AF) - \frac{1}{2}\ln|2\pi D|,$$

由此得出的F的得分即为F的最大似然估计，即Bartlett得分.

注意：一般A 和D是未知的，所以用某一种估计代入.

- 采用主成分得分（注意此时一般不需要加权）, 经计算可得, 因子得分与样本主成分得分仅相差一个常数;
- 采用主因子得分;
- 采用极大似然估计;

(3) 回归法(Thompson得分)

考虑公因子F对随机变量X满足如下的回归方程

$$F_j = b_{j1}X_1 + \cdots + b_{jp}X_p + \epsilon_j,\ j = 1, \cdots, m.$$

由因子载荷矩阵A,

$$a_{ij} = \mathsf{E}[X_i F_j] = b_{j1}r_{i1} + \cdots + b_{jp}r_{ip},\ i = 1, \cdots, p.$$

即

$$\begin{cases} a_{1j} = b_{j1}r_{11} + \cdots + b_{jp}r_{1p} \\ \cdots\cdots \\ a_{pj} = b_{j1}r_{p1} + \cdots + b_{jp}r_{pp} \end{cases}$$

整理得上述方程组得

$$\widehat{F} = A'R^{-1}X.$$

其中R是$p \times p$样本相关阵, A 为因子载荷矩阵$(p \times m)$, 由回归法得出的因子得分也被称为Thompson得分.

Bartlett得分与Thompson得分的比较

记：

- Bartlett得分：$\widehat{F}(1) = (A'D^{-1}A)A'D^{-1}X.$
- Thompson得分：$\widehat{F}(2) = A'R^{-1}X = A'(AA' + D)^{-1}X.$

- 二种得分之间的关系：

$$\begin{aligned}
\widehat{F}(2) &= A'(AA'+D)^{-1}X \\
&= (I_m + A'D^{-1}A)^{-1}A'DX \\
&= (I_m + A'D^{-1}A)^{-1}(A'D^{-1}A')^{-1}(A'D^{-1}A')A'DX \\
&= (I_m + A'D^{-1}A)^{-1}(A'D^{-1}A')^{-1}\widehat{F}(1). \\
\widehat{F}(1) &= (A'D^{-1}A')(I_m + A'D^{-1}A)\widehat{F}(2) \\
&= (I_m + (A'D^{-1}A)^{-1})\widehat{F}(2).
\end{aligned}$$

当条件$A'D^{-1}A = \Lambda$为对解阵，并且对角元素近似等于0时，二种估计得出的得分几乎相等.

- Bartlett得分为无偏估计, Thompson得分就有偏的.

$$
\begin{aligned}
\mathsf{E}[\widehat{F}(1)|F] &= \mathsf{E}[(A'D^{-1}A)A'D^{-1}X|F] \\
&= \mathsf{E}[(A'D^{-1}A)A'D^{-1}(AF+\epsilon)|F] = F. \\
\mathsf{E}[\widehat{F}(2)|F] &= \mathsf{E}[(I_m + A'D^{-1}A)^{-1}A'DX|F] \\
&= \mathsf{E}[(I_m + A'D^{-1}A)^{-1}A'D(AF+\epsilon)|F] \\
&= (I_m + A'D^{-1}A)^{-1}A'DAF.
\end{aligned}
$$

- Thompson得分就有偏的有较小的均方误差.

$$\begin{aligned}
\mathsf{E}[(\widehat{F}(1)-F)(\widehat{F}(1)-F)'|F] &= (A'D^{-1}A)^{-1}. \\
\mathsf{E}[(\widehat{F}(2)-F)(\widehat{F}(2)-F)'|F] &= (I_m + A'D^{-1}A)^{-1}.
\end{aligned}$$

因子分析函数factanal()的用法

格式：Factanal(x, factors, scores=(“none”, “regression”, “Bartlett”), rotation=” varimax”).

- x 为数值矩阵或数据表单
- $factors$ 为因子个数
- $scores$ 为因子得分的计算方法，包括“regression”, “Barlett”.
- $rotation$ 为因子旋转方法，缺省为varimax，如果rotation=” none”则不做因子旋转.

例8.3.1 盐泉水化学分析资料的因子分析.

- $D831 <- read.table("ex831.txt", header = F)$,
- $fre <- factanal(D831, 3, scores = "Bartlett", rotation = "none")$,
- 输出结果:
  - Uniquenesses:（是特殊方差）
  - Loadings:（是因子载荷矩阵）
  - 所选择的三个因子的信息(见下表)

	Factor1	Factor2	Factor3	说明
SS loadings	4.12	1.11	0.53	公共因子对变量的总方差贡献
Proportion Var	0.59	0.16	0.08	方差贡献率
Cumulative Var	0.59	0.75	0.82	累积方差贡献率

采用R语句：$fre <- factanal(D831, 3, scores = "Bartlett", rotation = "varimax")$

Uniquenesses:

V1	V2	V3	V4	V5	V6	V7
0.005	0.896	0.005	0.005	0.168	0.005	0.163

Loadings:

	Factor1	Factor2	Factor3
V1	-0.126	0.978	-0.151
V2		-0.265	0.179
V3	0.346	-0.408	0.843
V4	0.488	-0.398	0.776
V5	-0.173	0.780	-0.439
V6	0.921	-0.231	0.304
V7	-0.897		-0.176

	Factor1	Factor2	Factor3
SS loadings	2.059	2.015	1.683
Proportion Var	0.294	0.288	0.240
Cumulative Var	0.294	0.582	0.822

从因子分析结果中，取得因子得分$fre\$scores$

	Factor1	Factor2	Factor3
[1,]	-0.862163502	-0.6388763	1.40456115
[2,]	-0.740963311	-0.3736757	1.37782368
[3,]	1.758591988	-0.2246278	2.27140180
[4,]	-0.829967344	-0.7186278	1.26724714
[5,]	-0.086671065	-0.2925362	1.22871197
[6,]	-0.750204596	-0.7176610	0.51667004
[7,]	3.604968573	-0.3849943	-0.42555246
[8,]	-0.033005572	-0.9603268	-0.96492010
[9,]	-0.312914778	-0.9301337	-0.63602943
[10,]	-0.358945743	-0.8997472	-0.91059281
[11,]	-0.070530205	-0.9459084	-1.17783404
[12,]	-0.386747182	-0.8526238	-0.95672738
[13,]	0.006548049	-0.9720722	-1.12597660
[14,]	-0.200859058	1.2628140	-0.14466116
[15,]	-0.094652433	1.5374205	-0.11500456
[16,]	-0.136494826	1.4613570	-0.10149526
[17,]	-0.213422430	1.5795045	-0.04639371
[18,]	-0.083498652	1.0385721	-0.47563940
[19,]	-0.093141625	1.4399269	-0.28473219
[20,]	-0.115926288	0.5922162	-0.70085667

因子分析可分为R 型和Q 型两种.

- 当研究对象是变量时, 属于R型因子分析, 以上的讨论都是以变量作为的研究对象, 即是R型因子分析.
- 当研究对象是样品时, 属于Q型因子分析, 其做法只要将R 型因子分析中的变量和样品的作用互相调换, 其余的处理方法不变. 样品间的相似度一般采用相似系数（如夹角余弦）.

案例：国民生活质量的因素分析

国家发展的最终目标，是为了全面提高全体国民的生活质量，满足广大国民日益增长的物质和文化的合理需求。在可持续发展消费的统一理念下，增加社会财富，创自更多的物质文明和精神文明，保持人类的健康延续和生生不息，在人类与自然协同进化的基础上，维系人类与自然的平衡，达到完整的代际公平和区际公平(即时间过程的最大合理性与空间分布的最大合理化)。

从1990年开始，联合国开发计划署(UYNP)首次采用“人文发展系数”指标对于国民生活质量进行测度。人文发展系数利用三类内涵丰富的指标组合，即

- 人的健康状况(使用出生时的人均预期寿命表达),
- 人的智力程度(使用组合的教育成就表达),
- 人的福利水平(使用人均国民收入或人均GDP 表达).

并且特别强调三类指标组合的整体表达内涵，去衡量一个国家或地区的社会发展总体状况以及国民生活质量的总水平。

在这个指标体系中有如下的指标：

- X_1: 预期寿命
- X_2: 成人识字率
- X_3: 综合入学率
- X_4: 人均GDP（美圆）
- X_5: 预期寿命指数
- X_6: 教育成就指数
- X_7: 人均GDP指数

旋转后的因子结构

Rotated Factor Pattern

	FACTOR1	FACTOR2	FACTOR3
X1	0.38129	0.41765	0.81714
X2	0.12166	0.84828	0.45981
X3	0.64803	0.61822	0.22398
X4	0.90410	0.20531	0.34100
X5	0.38854	0.43295	0.80848
X6	0.28207	0.85325	0.43289
X7	0.90091	0.20612	0.35052

FACTOR1为经济发展因子

FACTOR2为教育成就因子

FACTOR3为健康水平因子

被每个因子解释的方差和共同度

Variance explained by each factor

FACTOR1	FACTOR2	FACTOR3
2.439700	2.276317	2.009490

Final Communality Estimates: Total = 6.725507

X1	X2	X3	X4	X5	X6	X7
0.986	0.946	0.852	0.976	0.992	0.995	0.977

Standardized Scoring Coefficients标准化得分系数

	FACTOR1	FACTOR2	FACTOR3
X1	-0.18875	-0.34397	0.85077
X2	-0.24109	0.60335	-0.10234
X3	0.35462	0.50232	-0.59895
X4	0.53990	-0.17336	-0.10355
X5	-0.17918	-0.31604	0.81490
X6	-0.09230	0.62258	-0.24876

案例：生育率的影响因素分析

生育率受社会、经济、文化、计划生育政策等很多因素影响，但这些因素对生育率的影响并不是完全独立的，而是交织在一起，如果直接用选定的变量对生育率进行多元回归分析，最终结果往往只能保留两三个变量，其他变量的信息就损失了。因此，考虑用因子分析的方法，找出变量间的数据结构，在信息损失最少的情况下用新生成的因子对生育率进行分析。

选择的变量有：

- 多子率,
- 综合节育率,
- 初中以上文化程度比例,
- 城镇人口比例,
- 人均国民收入.

下表是1990年中国30个省、自治区、直辖市的数据

多子率（%）	综合节育率（%）	初中以上文化程度比例（%）	人均国民收入（元）	城镇人口比例（%）
0.94	89.89	64.51	3577	73.08
2.58	92.32	55.41	2981	68.65
13.46	90.71	38.2	1148	19.08
12.46	90.04	45.12	1124	27.68
8.94	90.46	41.83	1080	36.12
2.8	90.17	50.64	2011	50.86
8.91	91.43	46.32	1383	42.65
8.82	90.78	47.33	1628	47.17
0.8	91.47	62.36	4822	66.23
5.94	90.31	40.85	1696	21.24
2.6	92.42	35.14	1717	32.81
7.07	87.97	29.51	933	17.9
14.44	88.71	29.04	1313	21.36
15.24	89.43	31.05	943	20.4
3.16	90.21	37.85	1372	27.34
9.04	88.76	39.71	880	15.52
12.02	87.28	38.76	1248	28.91
11.15	89.13	36.33	976	18.23
22.46	87.72	38.38	1845	36.77
24.34	84.86	31.07	798	15.1
33.21	83.79	39.44	1193	24.05
4.78	90.57	31.26	903	20.25
21.56	86	22.38	654	19.93
14.09	80.86	21.49	956	14.72
32.31	87.6	7.7	865	12.59
11.18	89.71	41.01	930	21.49
13.8	86.33	29.69	938	22.04
25.34	81.56	31.3	1100	27.35
20.84	81.45	34.59	1024	25.82
39.6	64.9	38.47	1374	31.91

没有旋转的因子结构与旋转后的因子结构的比较

旋转后的因子结构

	Factor1	Factor2
x1	-0.35310	-0.87170
x2	0.07757	0.95154
x3	0.89114	0.25621
x4	0.92204	0.16655
x5	0.95149	0.15728

没有 旋转的因子结构

	Factor1	Factor2
x1	-0.76062	0.55316
x2	0.56898	-0.76662
x3	0.89184	0.25374
x4	0.87066	0.34618
x5	0.89076	0.36962

各旋转后的共同度				
0.88454023	0.91143998	0.85977061	0.87789453	0.93006369

Factor1可解释方差	Factor2可解释方差
2.9975429	2.1642615

在这个例子中我们得到了两个因子，

- 第一个因子是社会经济发展水平因子，
- 第二个是计划生育因子。

有了因子得分值后，则可以利用因子得分为变量，进行其他的统计分析。

因子分析的步骤

- **第一步：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。

- **第二步：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。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无关系，做因子分析是不恰当的。并且相关系数矩阵是估计因子结构的基础。

- **第三步：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60%才能符合要求。

- **第四步：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。

- **第五步：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变量的聚类分析，做回归分析中的回归因子。