第七章 因子分析

汇报什么？

- 假定你是一个公司的财务经理，掌握了公司的所有财务数据，比如固定资产、流动资金、每一笔借贷的数额和期限、各种税费、工资支出、原料消耗、产值、利润、折旧等等。
- 如果让你向老总介绍公司近期财务状况，你能够把这些指标和数字都原封不动地罗列出去吗？
- 当然不能。否则……
- 你必须要把各个方面作出高度概括，从几个方面（综合指标）简单明了地把情况说清楚。

- 常见有很多变量的数据。这一类数据所涉及的问题可以推广到对企业、对学校、对区域进行分析、排序、分类等问题。
- 比如全国或各个地区的带有许多经济和社会变量的数据；各个学校的研究、教学等各种指标的数据等等。
- 变量很多，有一些是相关的。希望找出它们的少数“代表”。
- 因子分析（factor analysis）就是把变量维数降低以便于描述、理解和分析的方法。

引　言

因子分析(factor analysis)是一种数据化简的技术。它通过研究众多原始变量之间的相互关系，探索变量集的基本结构，并力图用少数几个假想的潜在变量来表示之；

由于这几个潜在变量能够反映众多原始变量的主要信息，这样就达到数据化简和降维评价的目的。原始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。

因子分析技术的发端

Spearman's Example

有一组古典文学、法语、英语、数学和音乐的测验成绩，从它们的相关性表明存在一个潜在的“智力”因子（$\mathbf{F_1}$）。而另一组变量，表示身体健康的得分，它们对应另一个潜在的因子（$\mathbf{F_2}$）。记这些变量为$(\mathbf{X_1},\ldots,\mathbf{X_p})$. 我要寻求下面这样的结构：

$$X_1 - \mu_1 = a_{11}F_1 + a_{12}F_2 + \mathrm{L} \; + a_{1m}F_m + \varepsilon_1$$

$$X_2 - \mu_2 = a_{21}F_1 + a_{22}F_2 + \mathrm{L} \; + a_{2m}F_m + \varepsilon_2$$

L L

$$X_p - \mu_p = a_{p1}F_1 + a_{p2}F_2 + \mathrm{L} \; + a_{pm}F_m + \varepsilon_p$$

or with matrix notation,

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

再例如，糕点行业品种繁多，多至成百上千种，要研究糕点行业的价格变动，似乎不那么容易。

但仔细分析起来，无论哪种样式的糕点，用料不外乎面粉、食油和糖等几大类原料，各种糕点的价格变动与面粉、食油和糖的价格变动密切相关。那么，面粉、食油和糖的价格就可以看作是众多糕点价格的公共因子。这三个公共因子与原有指标的关系可以表示为：

$$x_i = \alpha_{i1}F_1 + \alpha_{i2}F_2 + \alpha_{i3}F_3 + \varepsilon_i \qquad i = 1, \cdots, n$$

称 F_1、F_2、F_3 是不可观测的潜在因子。n个变量（指标）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。

要了解和掌握糕行业的物价变动，只要抓住上述三种潜在因子的价格变动即可。

可见，如果把握了主要矛盾、或者矛盾的主要方面，分析和解决问题的思路就打开了。因子分析的作用正在于此。

§ 1 因子分析模型

一、数学模型

设有 $X_i\ (i=1,2,\cdots,n)$ n 个变量，如果均可表示为

$$X_i=\alpha_{i1}F_1+\alpha_{i2}F_2+\cdots+\alpha_{im}F_m+\varepsilon_i \qquad (m\le n)$$

$$\begin{bmatrix}X_1\\X_2\\\vdots\\X_n\end{bmatrix}=\begin{bmatrix}\alpha_{11}&\alpha_{12}&\cdots&\alpha_{1m}\\\alpha_{21}&\alpha_{22}&\cdots&\alpha_{2m}\\\vdots&\vdots&&\vdots\\\alpha_{n1}&\alpha_{n2}&\cdots&\alpha_{nm}\end{bmatrix}\begin{bmatrix}F_1\\F_2\\\vdots\\F_m\end{bmatrix}+\begin{bmatrix}\varepsilon_1\\\varepsilon_2\\\vdots\\\varepsilon_n\end{bmatrix}$$

或 $\mathbf{X}=\mathbf{AF}+\boldsymbol{\varepsilon}$

称为 $F_1, F_2, \cdots, F_m$ 公共因子，是不可观测的变量，它们的系数称为因子载荷（变量与因子的相关）。ε_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：

$$\operatorname{cov}(F, \varepsilon) = 0,$$ 即 F, ε 不相关；

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

即 $F_1, F_2, \cdots, F_m$ 互不相关，方差为1。

$$\operatorname{cov}(\varepsilon)=\begin{bmatrix}\sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_n^2\end{bmatrix}$$

即特殊因子互不相关，方差不一定相等，$\varepsilon_i \sim N(0,\sigma_i^2)$ 。

用矩阵的表达方式

$$\mathbf{X}-\boldsymbol{\mu}=\mathbf{AF}+\boldsymbol{\varepsilon} \qquad E(\mathbf{F})=\mathbf{0}$$

$$E(\boldsymbol{\varepsilon})=\mathbf{0} \qquad Var(\mathbf{F})=\mathbf{I}$$

$$\mathrm{cov}(\mathbf{F},\boldsymbol{\varepsilon})=E(\mathbf{F}\boldsymbol{\varepsilon}')=\begin{pmatrix} E(F_1\varepsilon_1) & E(F_1\varepsilon_2) & \mathrm{L} & E(F_1\varepsilon_p) \\ E(F_2\varepsilon_1) & E(F_2\varepsilon_2) & \mathrm{L} & E(F_2\varepsilon_p) \\ \mathrm{M} & \mathrm{M} & & \mathrm{M} \\ E(F_p\varepsilon_1) & E(F_p\varepsilon_2) & \mathrm{L} & E(F_p\varepsilon_p) \end{pmatrix}=\mathbf{0}$$

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

二、因子分析模型的性质

1、原始变量X的协方差矩阵的分解

Q $\mathbf{X}-\boldsymbol{\mu}=\mathbf{AF}+\boldsymbol{\varepsilon}$

$\therefore$ $Var(\mathbf{X}-\boldsymbol{\mu})=\mathbf{A}Var(\mathbf{F})\mathbf{A}'+Var(\boldsymbol{\varepsilon})$

$\boldsymbol{\Sigma}_x=\mathbf{AA}'+\mathbf{D}$

A是因子模型的系数

$$Var(\boldsymbol{\varepsilon})=\mathbf{D}=diag(\sigma_1^2,\sigma_2^2,\mathrm{L}\ ,\sigma_p^2)$$

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、模型不受计量单位的影响

将原始变量X做变换$\mathbf{X^*=CX}$,这里

$\mathbf{C}=\mathrm{diag}(c_1,c_2,\ldots,c_n),c_i>0$。

$$\mathbf{C(X-\mu)=C(AF+\varepsilon)}$$

$$\mathbf{CX=C\mu+CAF+C\varepsilon}$$

$$\mathbf{X^*=C\mu+CAF+C\varepsilon}$$

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

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

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

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

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

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

3、因子载荷不是惟一的

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

$\mathbf{X}^* = \boldsymbol{\mu} + \mathbf{A}^*\mathbf{F}^* + \boldsymbol{\varepsilon}$ 且满足条件因子模型的条件

$$E(\mathbf{T}'\mathbf{F}) = \mathbf{0} \qquad E(\boldsymbol{\varepsilon}) = \mathbf{0}$$

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

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

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

§2 因子负荷矩阵A的估计方法

（一）主成分法（Principal Component）

（二）极大似然法（Maximum Likelihood ）

（三）主因子法

……

（一）主成分分析法

设随机向量 $\mathbf{x} = (x_1, x_2, \cdots, x_p)'$ 的均值为$\boldsymbol{\mu}$，协方差为$\boldsymbol{\Sigma}$，

$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ 为$\boldsymbol{\Sigma}$的特征根，

$\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_p$ 为对应的标准化特征向量，则

$$\boldsymbol{\Sigma} = \mathbf{U} \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \end{bmatrix} \mathbf{U}' = \mathbf{A}\mathbf{A}' + \mathbf{D}$$

$$\begin{bmatrix}\mathbf{u}_1 & \mathbf{u}_2 & \mathrm{L} & \mathbf{u}_p\end{bmatrix}\begin{pmatrix}\lambda_1 & & 0\\ & \mathrm{O} & \\ 0 & & \lambda_p\end{pmatrix}\begin{bmatrix}\mathbf{u}_1'\\ \mathbf{u}_2'\\ \mathrm{M}\\ \mathbf{u}_p'\end{bmatrix}$$

$$=\lambda_1\mathbf{u}_1\mathbf{u}_1'+\lambda_2\mathbf{u}_2\mathbf{u}_2'+\mathrm{L}\ +\lambda_m\mathbf{u}_m\mathbf{u}_m'+\lambda_{m+1}\mathbf{u}_{m+1}\mathbf{u}_{m+1}'+\mathrm{L}\ +\lambda_p\mathbf{u}_p\mathbf{u}_p'$$

$$=\begin{bmatrix}\sqrt{\lambda_1}\mathbf{u}_1 & \sqrt{\lambda_2}\mathbf{u}_2 & \cdots & \sqrt{\lambda_p}\mathbf{u}_p\end{bmatrix}\begin{bmatrix}\sqrt{\lambda_1}\mathbf{u}_1'\\ \sqrt{\lambda_2}\mathbf{u}_2'\\ \vdots\\ \sqrt{\lambda_p}\mathbf{u}_p'\end{bmatrix}$$

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项的贡献，有

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

$$= \left[\sqrt{\lambda_1}\mathbf{u}_1 \quad \sqrt{\lambda_2}\mathbf{u}_2 \quad \mathrm{L} \quad \sqrt{\lambda_m}\mathbf{u}_m \right]_{p\times m} \begin{bmatrix} \sqrt{\lambda_1}\mathbf{u}_1' \\ \sqrt{\lambda_2}\mathbf{u}_2' \\ \mathrm{M} \\ \sqrt{\lambda_p}\mathbf{u}_m' \end{bmatrix}_{m\times p} + \hat{\mathbf{D}} \approx \hat{\mathbf{A}}\hat{\mathbf{A}}' + \hat{\mathbf{D}}$$

其中$\hat{\mathbf{D}} = diag(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \mathrm{L} \ , \hat{\sigma}_p^2)$

$$\hat{\sigma}_i^2 = s_{ii} - \sum_{j=1}^{m} a_{ij}^2$$

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

注：残差矩阵

$$\mathbf{S-\hat{A}\hat{A}'-\hat{D}}$$

其中S为样本的协方差矩阵。

（二）主因子法

主因子方法是对主成分方法的修正，假定我们首先对变量进行标准化变换。则

$$R=AA'+D$$

$$R^*=AA'=R-D$$

称R^*为约(或再生)相关矩阵，　R^*对角线上的元素是 h_i^2　　，

而不是1。

$$R^* = \mathbf{R} - \hat{\mathbf{D}} = \begin{bmatrix} \hat{h}_1^2 & r_{12} & \mathrm{L} & r_{1p} \\ r_{21} & \hat{h}_2^2 & \mathrm{L} & r_{2p} \\ \mathrm{M} & \mathrm{M} & & \mathrm{M} \\ r_{p1} & r_{p2} & \mathrm{L} & \hat{h}_p^2 \end{bmatrix}$$

直接求$\mathbf{R}^*$的前p个特征根和对应的正交特征向量。得如下的矩阵：

$$\mathbf{A} = \left[\sqrt{\lambda_1^*}\mathbf{u}_1^* \quad \sqrt{\lambda_2^*}\mathbf{u}_2^* \quad \mathrm{L} \quad \sqrt{\lambda_p^*}\mathbf{u}_p^* \right]$$

$\mathbf{R}^*$特征根：$\lambda_1^* \geq \mathrm{L} \geq \lambda_p^* \geq 0$

正交特征向量：$\mathbf{u}_1^*, \mathbf{u}_2^*, \mathrm{L}, \mathbf{u}_p^*$

当特殊因子 ε_i 的方差不为0且已知的，问题非常好解决。

$$\mathbf{R}^{*}=\mathbf{R}-\begin{bmatrix}\sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & 0 & \\ & & & \sigma_p^2\end{bmatrix}$$

$$=\begin{bmatrix}\sqrt{\lambda_1^{*}}\mathbf{u}_1^{*} & \sqrt{\lambda_2^{*}}\mathbf{u}_2^{*} & \mathrm{L} & \sqrt{\lambda_p^{*}}\mathbf{u}_p^{*}\end{bmatrix}\begin{bmatrix}\sqrt{\lambda_1^{*}}\mathbf{u}_1'^{*} \\ \sqrt{\lambda_2^{*}}\mathbf{u}_2'^{*} \\ \mathrm{M} \\ \sqrt{\lambda_p^{*}}\mathbf{u}_p'^{*}\end{bmatrix}$$

$$\mathbf{A} = \left[\sqrt{\lambda_1^*}\,\mathbf{u}_1^* \quad \sqrt{\lambda_2^*}\,\mathbf{u}_2^* \quad \mathrm{L} \quad \sqrt{\lambda_m^*}\,\mathbf{u}_m^* \right]$$

$$\mathbf{D} = \begin{pmatrix} 1-\hat{h}_1^2 & & 0 \\ & \mathrm{O} & \\ 0 & & 1-\hat{h}_p^2 \end{pmatrix}$$

在实际的应用中，个性方差矩阵一般都是未知的，可以通过一组样本来估计。**估计的方法有如下几种：**

首先，求h_i^2的初始估计值，构造出 $\mathbf{R}^*$

1）取$h_i^2=1$，在这个情况下主因子解与主成分解等价；

2）取 $h_i^2=R_i^2$ ，R_i^2为x_i与其他所有的原始变量x_j的复相关系数的平方，即x_i对其余的p-1个x_j的回归方程的判定系数，这是因为x_i 与公共因子的关系是通过其余的p-1个x_j 的线性组合联系起来的；

2）取 $\hat{h}_i^2 = \max|r_{ij}|(j \neq i)$，这意味着取$x_i$与其余的$x_j$的简单相关系数的绝对值最大者；

4）取 $h_i^2 = \dfrac{1}{p-1}\sum_{j=1,i\neq j}^{p} r_{ij}$，其中要求该值为正数。

5）取 $h_i^2 = 1/r^{ii}$，其中 r^{ii} 是 $\mathbf{R}^{-1}$ 的对角元素。

（三）极大似然估计法（略）

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 $\mathbf{x_1, x_2, \cdots, x_n}$ 为来自正态总体 $\mathbf{N_p(\mu,\Sigma)}$的随机样本。

$$\mathbf{\Sigma = AA' + \Sigma_\varepsilon}$$

$$L(\hat{\boldsymbol{\mu}}, \hat{\mathbf{A}}, \hat{\mathbf{D}}) = f(\mathbf{X}) = f(X_1) \cdot f(X_2) \mathrm{L}\ \ f(X_n)$$

$$= \prod_{i=1}^{n} (2\pi)^{-p/2} |\Sigma|^{1/2} \exp[-\frac{1}{2}(x_i - \mu)'\Sigma^{-1}(x_i - \mu)]$$

$$= \left[(2\pi)^{p} |\Sigma|\right]^{-n/2} \exp[-\frac{1}{2} \sum_{i=1}^{n} (\mathbf{X}_i - \mu)'\Sigma^{-1}(\mathbf{X}_i - \mu)]$$

它通过$\boldsymbol{\Sigma}$依赖$\mathbf{A}$和$\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$。上式并不能唯一确定$\mathbf{A}$,为此可添加一个唯一性条件：

$$\mathbf{A}'\boldsymbol{\Sigma}_{\varepsilon}^{-1}\mathbf{A} = \boldsymbol{\Lambda}$$

这里$\boldsymbol{\Lambda}$式一个对角矩阵，用数值极大化的方法可以得到极大似然估计 $\hat{\mathbf{A}}$和$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}}$。极大似然估计$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}}$和$\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$将使 $\hat{\mathbf{A}}'\hat{\boldsymbol{\Sigma}}_{\varepsilon}^{-1}\hat{\mathbf{A}} = \hat{\boldsymbol{\Lambda}}$ 为对角阵，且似然函数达到最大。

相应的共同度的似然估计为：

$$h_i^2 = \hat{a}_{i1}^2 + \hat{a}_{i2}^2 + \cdots + \hat{a}_{im}^2$$

第J个因子对总方差的贡献：

$$S_j^2 = \hat{a}_{1j}^2 + \hat{a}_{2j}^2 + \cdots + \hat{a}_{pj}^2$$

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：$\lambda_1 = 1.55 \quad \lambda_2 = 0.85 \quad \lambda_3 = 0.6$

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

$$\mathbf{A} = \begin{bmatrix} 0.475\sqrt{1.55} & 0.883\sqrt{0.85} & 0 \\ 0.629\sqrt{1.55} & -0.331\sqrt{0.85} & 0.707\sqrt{0.6} \\ -0.629\sqrt{1.55} & 0.331\sqrt{0.85} & 0.707\sqrt{0.6} \end{bmatrix}$$

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

$$x_1 = 0.569F_1 + 0.814F_2$$

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

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

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。第一公因子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。

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

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

试用主因子分析法求因子分析模型。假定用$\hat{h}_i^2 = \max |r_{ij}| (j \neq i)$代替初始的 h_i^2　$h_1^2 = \frac{1}{5}, h_2^2 = 1, h_3^2 = \frac{2}{5}$。

$$R^* = \begin{bmatrix} 1/5 & 1/5 & -1/5 \\ 1/5 & 1 & 2/5 \\ -1/5 & -2/5 & 2/5 \end{bmatrix} = \frac{1}{5}\begin{bmatrix} 1 & 1 & -1 \\ 1 & 5 & -2 \\ -1 & -2 & 2 \end{bmatrix}$$

特征根为：$\lambda_1 = 0.9123$

$\lambda_2 = 0.0877$

$\lambda_3 = 0$

对应的非零特征向量为：$\begin{bmatrix} 0.369 & 0.929 \\ 0.657 & -0.261 \\ -0.657 & 0.261 \end{bmatrix}$

$$\begin{bmatrix} 0.369\sqrt{0.9123} & 0.929\sqrt{0.0877} \\ 0.657\sqrt{0.9123} & -0.261\sqrt{0.0877} \\ -0.657\sqrt{0.9123} & 0.261\sqrt{0.0877} \end{bmatrix} = \begin{bmatrix} 0.352 & 0.275 \\ 0.628 & -0.077 \\ -0.628 & 0.077 \end{bmatrix}$$

$$x_1 = 0.352F_1 + 0.275F_2 + \varepsilon_1$$

$$x_2 = 0.625F_1 - 0.077F_2 + \varepsilon_2$$

$$x_1 = -0.682F_1 + 0.077F_2 + \varepsilon_3$$

□ □ □ □ □ □ ：

$$h_1^2 = 0.352^2 + o.275^2 = 0.18129$$

$$h_2^2 = 0.625^2 + 0.077^2 = 0.3966$$

$$h_3^2 = 0.682^2 + 0.077^2 = 0.4710$$

例：

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

撑杆跳高成绩 X_8

标枪成绩 X_9

1500米跑成绩 X_{10}

选取了二战以后139名参加了奥林匹克运动会十项全能项目的运动员，获取了他们在每一个项目上的成绩数据，并对数据进行了标准化的处理。用主成分法抽取公共因子，进行因子分析。

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

变量	F_1	F_2	F_3	F_4	共同度
X_1 100m	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	0.8
X_4 跳高	0.674	0.134	0.139	0.396	0.65
X_5 400m	0.62	0.551	-0.084	-0.419	0.87
X_6 110m	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	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} 1500m	0.147	0.596	0.658	-0.279	0.89

因子载荷矩阵可以看出，第一因子在所有的变量上均有较大的正载荷，可以解释为一般运动因子。其他的3个因子不太容易解释。有的似乎是跑和投掷的能力对比，有的似乎是长跑耐力和短跑速度的对比。

变量	F_1	F_2	F_3	F_4	共同度
X_1 100m	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	0.8
X_4 跳高	0.674	0.134	0.139	0.396	0.65
X_5 400m	0.62	0.551	-0.084	-0.419	0.87
X_6 110m	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	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} 1500m	0.147	0.596	0.658	-0.279	0.89

因子载荷矩阵可以看出，第一因子在所有的变量上均有较大的正载荷，可以解释为一般运动因子。其他的3个因子不太容易解释。有的似乎是跑和投掷的能力对比，有的似乎是长跑耐力和短跑速度的对比。

§3 因子载荷矩阵中的几个统计特征

一、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 $X_i = a_{i1}F_1 + \cdots + a_{im}F_m + \varepsilon_i$

根据公共因子的模型性质，有

$\gamma_{x_i F_j} = \alpha_{ij}$ （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个公共因子的相关密切程度。绝对值越大，相关的密切程度越高。

二、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 $h_i^2 = \sum_{j=1}^{m} a_{ij}^2$。

统计意义：

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

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

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

所有的公共因子和特殊因子对变量 X_i 的方差贡献为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。

三、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

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

称为 F_j 对所有的 $X_i\ (i=1,\cdots,p)$ 的方差贡献.

S_j 是对公因子 F_j 相对重要性的一种衡量。

§4 因子旋转

为什么要旋转因子？

因子分析的目的不仅仅要找出几个公共因子，更重要的要弄清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，所以可以对因子载荷阵进行旋转，使因子载荷阵的结构简化。主要的正交旋转法：方差最大法(Varimax)等。

变量	F_1	F_2	F_3	F_4	共同度
X_1 100m	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 400m	0.797*	0.075	0.102	0.468	0.87
X_6 110m	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	-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} 1500m	0.045	-0.041	0.112	0.934*	0.89

注：此处跳远为三级跳远。

变量	F_1	F_2	F_3	F_4	共同度
X_1 100m	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 400m	0.797*	0.075	0.102	0.468	0.87
X_6 110m	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	-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} 1500m	0.045	-0.041	0.112	0.934*	0.89

注：此处跳远为三级跳远。

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

X_3 铅球，X_7 铁饼和 F_2 标枪在 上有较大的载荷，可以称为投掷（臂力）因子；

X_6 百米跨栏，X_8 撑杆跳远，F_4 跳远和 F_3 跳高在 上有较大的载荷，弹跳（腿力）因子；长跑（耐力）因子。

§ 5 因子得分

从前面的例子中我们可以看到，公共因子可以帮助我们了解原始变量之间的依赖关系，探求变量的基本结构，并据此对变量分组。但我们有时还要用这些因子做其他研究，比如对观测对象进行分类或评价，实现数据化简和降维评价的目的。这样就还需要估计出每个观测对象在几个公共因子上的得分。估计的方法主要有：巴特莱特法、回归法等。

因子分析的数学模型为：

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1m} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2m} \\ \vdots & \vdots & & \vdots \\ \alpha_{n1} & \alpha_{n2} & \cdots & \alpha_{nm} \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix}$$

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变量的线性组合。

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

可见，要求得每个因子的得分，必须求得得分函数的系数，而由于$n>m$（未知参数大于观测个数），所以不能得到精确的得分，只能通过估计。

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

旋转后的负荷矩阵

	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

$$x_i = \alpha_{i1}F_1 + \alpha_{i2}F_2 + \alpha_{i3}F_3 + \varepsilon_i$$

X1=-0.21522F1-0.27397F2+0.89092F3

X2=0.63973F1-0.28739F2-0.28755F3 ……

X7=-0.11041F1+0.97851F2-0.07246F3

	高载荷指标	因子命名
因子1	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2	X6：万人拥有的大学生数（人） X7：万人拥有的科学家、工程师数（人）	人力资源因子
因子3	X1;人口（万人） X3:GDP(亿元)	社会经济总量因子

Standardized Scoring Coefficients
标准化得分系数

	FACTOR1	FACTOR2	FACTOR3
X1	0.05764	-0.06098	0.50391
X2	0.22724	-0.09901	-0.07713
X3	0.14635	0.12957	0.59715
X4	0.47920	0.11228	0.17062
X5	0.45583	0.07419	0.10129
X6	0.05416	0.48629	0.04099
X7	0.05790	0.48562	0.04822

F1=0.05764X1+0.22724X2+0.14635X3+.....+0.05790X7

F2=-0.06098X1-0.09901X2+0.12957X3+.....+0.48562X7

F3=0.50391X1-0.07713X2+0.59715X3+.....+0.04822X7

前三个因子得分（降维评价）

REGION	自然资源	人力资源	经济总量
beijing	-0.08169	4.23473	-0.37983
tianjin	-0.47422	1.31789	-0.87891
hebei	-0.22192	-0.35802	0.86263
shanxi1	-0.48214	-0.32643	-0.54219
neimeng	0.54446	-0.66668	-0.92621
liaoning	-0.20511	0.46377	0.34087
jilin	-0.21499	0.10608	-0.57431
heilongj	0.10839	-0.11717	-0.02219
shanghai	-0.20069	2.38962	-0.04259

可在三维空间（图形）中对省区市进行分类及评价

§ 6 因子分析的步骤和建议

因子分析通常包括以下五个步骤

选择要分析的变量并标准化

计算所选原始变量的相关系数矩阵

提取公共因子（主成分+极大似然）

因子旋转

计算因子得分

对因子分析结果的考虑：

公因子的累积方差贡献率、变量共同度

公因子个数、 因子的内涵

一些注意事项

- 因子分析依赖于原始变量，也只能反映原始变量的信息。所以原始变量的选择很重要。
- 另外，如果原始变量本质上都独立，那么降维就可能失败。数据越相关，降维效果就越好。
- 在得到分析的结果时，并不一定会都得到如我们例子那样清楚的结果。这与问题的性质，选取的原始变量以及数据的质量等都有关系
- 在用因子得分进行排序时要特别小心，特别是对于敏感问题。由于原始变量不同，因子的选取不同，排序可以很不一样。

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，但实际上，很多因子分析并没有产生如此明确的结果。虽然我们可以从累计方差贡献率、共同度等指标中获取部分的信息，但不幸的是，评价因子分析效果的法则在很大程度上仍要依赖一个

“哇！”准则

也许这正是统计学艺术性一面的体现。