探索性因子分析

- 每一个变量的变异性可以归结为**公共因子**+**特殊因子**
- 目的：寻找少量公共因子以解释一组输入变量
- 公共因子可用于进一步分析

正交因子分析

- 假设 $\mathrm{X} = \left(\mathrm{X}_1, \cdots, X_p\right)^{\mathrm{T}} \in R^p$

$$\mathrm{X}_1 \ - \mu_1 = a_{11}F_1 + a_{12}F_2 + \cdots + a_{1q}F_q + \varepsilon_1$$

$$\mathrm{X}_2 \ - \mu_2 = a_{21}F_1 + a_{22}F_2 + \cdots + a_{2q}F_q + \varepsilon_2$$

$$\cdots \quad \cdots$$

$$\mathrm{X}_p \ - \mu_p = a_{p1}F_1 + a_{p2}F_2 + \cdots + a_{pq}F_q + \varepsilon_p$$

- 写成矩阵形式：$X \ - \mu = AF + \varepsilon$

载荷矩阵　公因子　特殊因子

正交因子分析

- 正交因子不可观测，为识别它们，做如下假定
- $E(F_i) = 0, var(F_i) = 1, cov\left(F_i, F_j\right) = 0 \ (i \neq j)$
- $E(\varepsilon_k) = 0, var(\varepsilon_k) = \sigma_k^2,\ cov(\varepsilon_k, \varepsilon_m) = 0 \ (k \neq m)$
- $cov(F_i, \varepsilon_k) = 0$

正交因子分析

- 通过以上假定，可以得到如下结论
- $var(X_k) = a_{k1}^2 + a_{k2}^2 + \cdots + a_{kq}^2 + \sigma_k^2$
  - $a_{k1}^2 + a_{k2}^2 + \cdots + a_{kq}^2$ 称为X_k的共性方差
  - σ_k^2 称为 X_k 的特殊方差
- $cov(X_k, X_m) = a_{k1}a_{m1} + \cdots + a_{kq}a_{mq}$

公共因子的解释

- 解释公共因子F_i时，可以通过对载荷系数的绝对值较大的输入来解释
  - ✓载荷系数的正负本身没有意义
  - ✓正负对比有意义

模型估计

- $\Sigma = AA^{\top} + \Psi$
- 因子载荷矩阵A有无穷多个解：
- 对于任意的正交矩阵$\mathrm{Q} \in R^{q\times q}$，令$A^* = AQ$，有 $A^*\mathrm{A}^{*\top} = (AQ)(AQ)^{\top} = AA^{\top}$
- 因此，可以先得到载荷矩阵估计的初始值，再经过旋转得到更好的解释

模型估计：主成分法

- 令$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$表示Σ的特征值，则Σ有如下分解

$$\Sigma = \lambda_1 v_1 v_1^\top + \lambda_2 v_2 v_2^\top + \cdots + \lambda_p v_p v_p^\top$$

- 令载荷矩阵 $\hat{A}$ 的第i列为 $\sqrt{\lambda_i}\, v_i$ ，则有 $\hat{A}\,\hat{A}^\top = \lambda_1 v_1 v_1^\top + \lambda_2 v_2 v_2^\top + \cdots + \lambda_q v_q v_q^\top$

- 令 $\widehat{\sigma_k^2} = \Sigma_{kk} - \sum_{i=1}^{q} \widehat{a_{ki}^2}$

模型估计：最大似然估计

- 假定 $F_1, \cdots, F_q$ 都服从多元正态分布，由于载荷矩阵的不唯一性，需要附加一个方便计算的唯一性条件：
  - $A^{\top}\Psi^{-1}\mathrm{A}$ 为对角矩阵
- 然后可以得到A和Ψ 最大似然估计。

因子旋转

得到因子载荷矩阵的初步估计后，可进行因子旋转，旋转后的载荷矩阵需要满足下列几个条件:

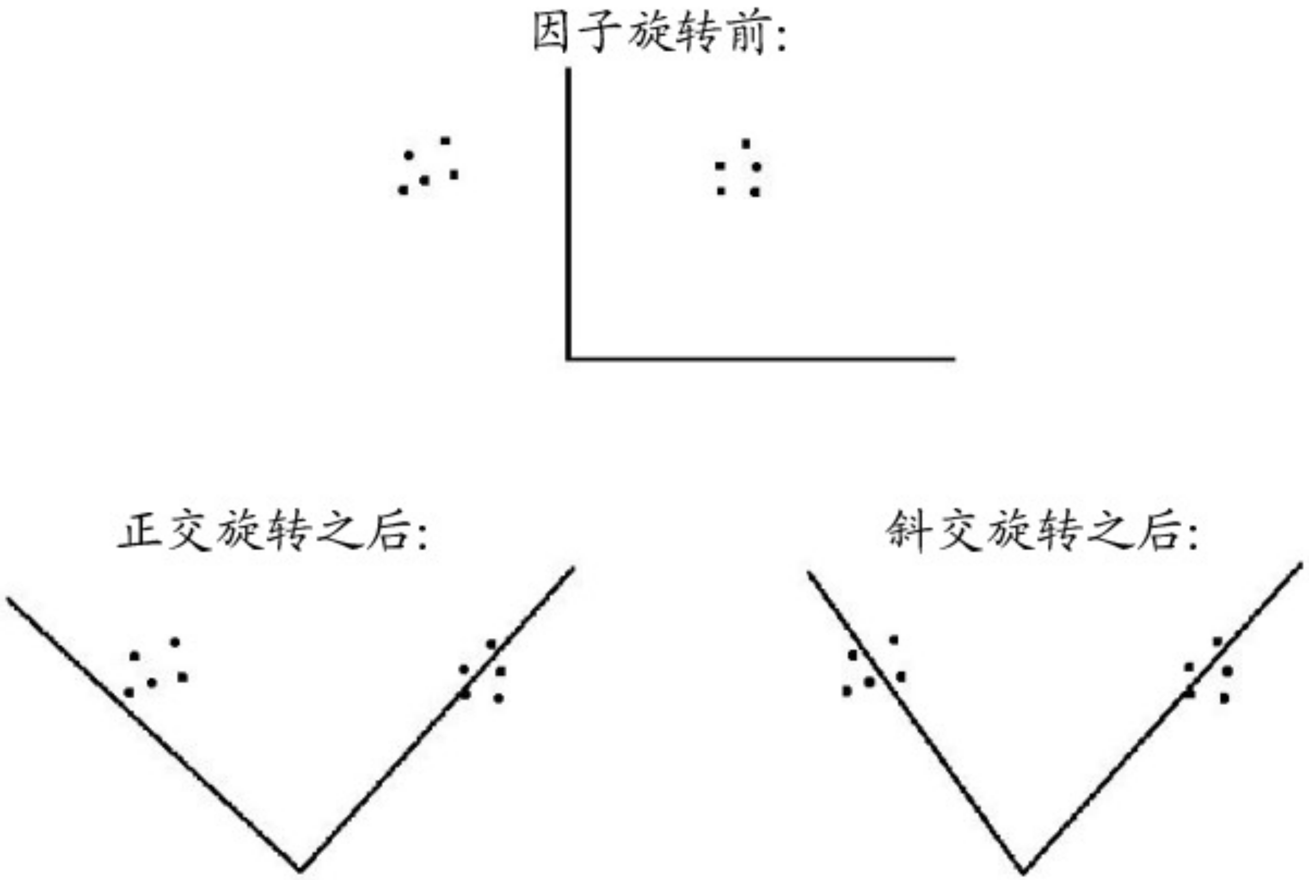
- 对于任意因子而言，只有少数输入变量在该因子上的载荷的绝对值较大，其余变量在该因子上的载荷接近于0;
- 对于任意输入变量而言，它只在少数因子上的载荷的绝对值较大，在其它因子上的载荷接近于0;
- 任何两个因子对应的载荷呈现不同的模式，因而在解释时这两个因子具有不同的含义。

因子旋转

- 正交旋转：采用正交矩阵对因子进行旋转，保持了因子之间的正交性；
- 斜交旋转：采用非正交矩阵对因子进行旋转，可以更好地简化载荷矩阵，提高因子的可解释性，但旋转后的因子之间存在相关性。

选择哪一类旋转依赖于对因子之间相关性的假定。

因子旋转



最大方差旋转（varimax rotation）

- 应用最广泛的因子旋转方法：
  - 它是一种正交旋转
  - 目的是使得载荷平方方差最大化：

$$\sum_{k=1}^{p}\sum_{i=1}^{q}\left(a_{ki}^{2}-\frac{1}{pq}\sum_{k'=1}^{p}\sum_{i'=1}^{q}a_{k'i'}^{2}\right)^{2}$$

探索性因子分析

- Kaiser准则：共性方差占总方差比例大于平均解释比 $\sum_{k=1}^{p} a_{ki}^2 / \sum_{k=1}^{p} \Sigma_{kk} > 1/p$
- 使用崖底碎石图（scree plot），选择拐点之前的一点
- 如果载荷矩阵由最大似然估计而得，可以使用假设检验

因子得分

- 对公共因子的 $F = (F_1, \cdots, F_q)$ 的估计值被称为“因子得分”
  - 可以通过最小二乘等方法来估计
  - 其形成的降维数据可以用于进一步分析

用最小二乘估计求因子得分

- 假设收集到的样本矩阵为：$\mathbb{X} \in R^{N\times p}$ ，那么因子模型可以写为：

$$\mathbb{X}^\top = A\mathbb{F}^\top + \mathbb{E}^\top$$

其中，$\mathbb{F} = \left(F^{(1)}, \cdots, F^{(n)}\right)^\top \in R^{n\times q}$为因子矩阵。

- 假设已知A，最小二乘的目标函数为：

$$\|\mathbb{X}^\top - A\mathbb{F}^\top\|_F^2$$

求解以上目标函数，可以得到因子得分：

$$\widehat{\mathbb{F}}^\top = (A^\top A)^{-1}A^\top\mathbb{X}^\top$$

探索性因子分析

- 因子分析和主成分分析的区别：

✓主成分分析解释输入变量的总变异，因子分析解释公共变异性

✓因子分析有“模型”，因子不可观测，存在识别性问题

其他降维方法

- Kernalized PCA
- Manifold learning