治理技术专题

定量政治分析方法

Quantitative Analysis II

苏 毓 淞

清华大学社会科学院政治学系

第四讲　因子分析

因子分析 (factor analysis)

- 因子分析是主成分分析 (principle component) 方法的延伸 (extension)。
- 一种降维的方法。
- 从不同变量之间“相关性矩阵”的“依赖关系”出发，归纳降维成几个综合因子。
- 起源于 1904 年 Charles Spearman 的学生考试成绩研究。

基本思路

- 根据变量之间的相关性进行分组，让同组内的变量具有较高的相关性，但不同组间变量的相关性则较低。
- 每组变量代表一个基本结构，称之为“公共因子”。
- 原始变量可以拆解为：
  1. 以公共因子为要素的线性函数；
  2. 与公共因子无关特殊因子。
- 使用公共因子可以降维处理变量，也可以对不同样本进行分类。

基本理论和模型

- Spearman 研究 33 名学生在古典法 (C)、法语 (F)、英语 (E)、数学 (M)、判别 (D)、和音乐 (S) 六门考试的相关性得到以下相关矩阵：

	C	F	E	M	D	S
C	1.00	0.83	0.78	0.70	0.66	0.63
F	0.83	1.00	0.67	0.67	0.65	0.57
E	0.78	0.67	1.00	0.64	0.54	0.51
M	0.70	0.67	0.64	1.00	0.45	0.51
D	0.66	0.65	0.54	0.45	1.00	0.40
S	0.63	0.57	0.51	0.51	0.40	1.00

- 他发现一个规律，在不考虑对角线要素的情况下，C 列和 E 列有：

$$\frac{0.83}{0.67} \approx \frac{0.70}{0.64} \approx \frac{0.66}{0.54} \approx \frac{0.63}{0.51} \approx 1.2$$

- 所以每个考试成绩都遵守：

$$X_i = \alpha_i F + e_i$$

基本理论和模型

- 如果有 m 个公共因子，则：

$$X_i = \alpha_{i1}F_1 + a_{i2}F_2 + \cdots + a_{im}F_m + e_i$$

- e_i 为特殊因子，var(e_i) 就是 X_i 的特殊度 (Uniqueness) 或残余方差，表示 X_i 与公共因子无关的部分。
- a_{ij} 称为因子载荷 (factor loading)，a_{ij} 的绝对值越大 ($|a_{ij}| \leq 1$)，表明 X_i 与 F_j 的载荷量越大，相依程度越高。

基本理论和模型

- 假定 X_i 和 F 的方差均为 1。

$$\text{var}(X_i) = \alpha_{i1}^2 \text{var}(F_1) + a_{i2}^2 \text{var}(F_2) + \cdots + a_{im}^2 \text{var}(F_m) + \text{var}(e_i)$$
$$\text{var}(X_i) = \alpha_{i1}^2 + a_{i2}^2 + \cdots + a_{im}^2 + \text{var}(e_i)$$
$$\text{var}(X_i) = h_i^2 + \sigma_i^2$$

- $\alpha_{i1}^2 + a_{i2}^2 + \cdots + a_{im}^2 = h^2$ 表示公共因子解释 X_i 方差的比例，称之为 X_i 的共同度，h_i^2 越大，表示公共因子能解释 X_i 方差的比例越大，因子分析的效果也就越好

基本理论和模型

- p 个 X 与 m 个 F 的关系为：

$$
\begin{aligned}
X_1 &= \alpha_{11}F_1 + \alpha_{12}F_2 + \cdots + \alpha_{1m} + e_1 \\
X_2 &= \alpha_{21}F_1 + \alpha_{22}F_2 + \cdots + \alpha_{2m} + e_2 \\
&\vdots \\
X_p &= \alpha_{p1}F_1 + \alpha_{p2}F_2 + \cdots + \alpha_{pm} + e_p
\end{aligned}
$$

- 各个 F 之间是独立的 $\text{cov}(\boldsymbol{F}) = 0$，各个 e 之间也是独立的 $\text{cov}(\boldsymbol{e}) = 0$。
- 公共因子对于 p 个 X 贡献度 (Contribution)：

$$
g^2 = \alpha_{11}^2 + a_{21}^2 + \cdots + a_{p1}^2
$$

- g^2 越大，则公共因子解释这 p 个 X 的贡献就越大。

主要概念

- 因子载荷：某个因子与某个原变量的相关系数，主要反映该公共因子对相应原变量的贡献力大小。
- 变量共同度：对某一个原变量来说，其在所有因子上的载荷的平方和就叫做该变量的共同度。它反映了所有公共因子对该原变量的方差（变异）的解释程度。如果因子分析结果中大部分变量的共同度都高于 0.8，说明提取的公共因子已经基本反映了原变量 80%以上的信息，因子分析效果较好。变量共同度是衡量因子分析效果的常用指标。
- 公共因子的方差贡献：是某公共因子对所有原变量载荷的平方和, 它反映该公共因子对所有原始总变异的解释能力，等于因子载荷矩阵中某一列载荷的平方和。一个因子的方差贡献越大，说明该因子就越重要。

检验

- 确定原有若干变量是否适合于做因子分析的基本依据是原有变量的相关性矩阵。
- 如果相关矩阵中的相关系数大都小于 0.3，而且未达到显著性水平，则说明变量间的相关性普遍较低，它们存在潜在共同因子的可能性较小，就不再适合于做因子分析；
- 如果相关系数都比较大，则可以进行因子分析。
- 三个用于判断因子分析适合度的指标：
  - 巴特利特球形检验（Bartlett Test of Sphericity）；
  - 反像相关矩阵检验（Anti-image correlation matrix）；
  - KMO(Kaiser-Meyer-Olkin) 检验。

巴特利特球形检验

- 该检验首先假设变量相关矩阵为单位阵 (identity matrix)，即对角线为 1、非对角线为 0，然后检验实际相关矩阵与此差异性。如果差异性显著，则拒绝单位阵假设，即认为原变量间的相关性显著，适合于作因子分析，否则不能作因子分析。
- H_0: 相关系数矩阵是一个单位阵
- 目的是拒绝 H_0

反像相关矩阵检验

- 反像相关矩阵检验以原变量的偏相关矩阵为基础。将偏相关矩阵中的每个元素（偏相关系数）取反 (即取负) 得到反像相关矩阵。如果原变量间相互作用较大，则控制了这些相互作用后的偏相关系数较小，此时反像相关矩阵中的元素的绝对值比较小，则适合于做因子分析，反之则不适合于作因子分析。
- `cov2cor(solve(corMat))`

KMO

- KMO 检验是依据变量间的简单相关与偏相关的比较。其计算公式为所有原变量简单相关系数的平方和除以简单相关系数平方和加偏相关系数平方和。即：

$$KMO = \frac{\sum\sum_{i=j} r_{ij}^2}{\sum\sum_{i=j} r_{ij}^2 + \sum\sum_{i=j} p_{ij}^2}$$

- 与反像相关检验的本质一样，如原变量间相互作用较大，变量间的偏相关系数就会相对较小，简单相关系数则相对较大。从上面的公式看出，KMO 值就大，适合于因子分析，反之则 KMO 值较小而不适合于做因子分析。Kaiser 提供的判断标准是：至少 > 0.6

0.00 to 0.49	无法接受
0.50 to 0.59	糟糕的
0.60 to 0.69	普通
0.70 to 0.79	中等
0.80 to 0.89	很好
0.90 to 1.00	非常好

Cronbach's Alpha

- 克隆巴赫系数（Cronbach's alpha）是检视内在信度 (internal consistency, reliability) 的一种方法
- 要做信度分析需先检查每个问项是否都是同方向的（即都是正面问法，也就是题间的相关系数都是正的），如有一题与其它题相关系数都是负的，应考虑将此题先“变号”或“删除”后再进行计算 *alpha* 系数。如有受测者乱答，可将它的数据删除后再算 α 值。
- 对问卷调查当有题目与其它题目是负相关时须注意是否反向问法。如是，则应先将得分反向，再计算 α 信度或是删除该题。若为测验，则不能做反向处理，只能做删除题目。
- α 值在 0.7 以上即通过检验，α 值在 0.5 以下则不通过检验，但这些不是固定的标准，

Square Multiple Correlation, SMC

- 其实就是回归的 R^2
- 即每一个变量回归其他变量的 R^2。
- 如果有 SMC 过小的变量，可以考虑删除。

求解因子载荷

- 主成分法：操作因子分析必须先使用主成分法，使用变量间的相关矩阵求得，但是得出的 e_i 之间彼此不独立，所以答案不正确，但是会返回几个主成分，可以供我们使用以下方法参考用。
- 主轴因子法：类似主成分法，差别在利用共同度代替相关矩阵对角线上的 1。
- 最大似然法 (Maximum Likelihood)：假定公共因子和特殊因子遵从联合正态分布不，进行最大似然法求解。
- 因子旋转 (rotation)：以上方法的到的因子载荷矩阵都不是唯一解，所以要进行因子旋转，旋转的目的是要是因子载荷系数尽可能接近 0 或 1。

因子分析步骤

- 根据研究问题选取变量。
- 标准化原始变量后，求相关性矩阵。
- 求解公共因子和因子载荷矩阵。
- 因子旋转
- 因子得分（predictive factor)
- 进一步分析。

因子旋转的目的

- 使每个变量在尽可能少的因子上有比较高的载荷；让某个变量在某个因子上的载荷趋于 1，而在其他因子上的载荷趋于 0。
- 要求每一列上的载荷大部分为很小的值，每一行中只有少量的最好只有一个较大的载荷值；每两列中大载荷与小载荷的排列模式应该不同。

因子旋转的方法：

- varimax: 方差最大旋转。简化对因子的解释（常用）。
- direct oblimin: 直接斜交旋转。允许因子之间具有相关性。
- quartmax: 四次最大正交旋转。简化对变量的解释
- equamax: 平均正交旋转。
- promax: 斜交旋转方法。（常用）。