多 元 统 计 分 析 方 法

——— 因 子 分 析

引言

事物的表现是多方面的，事物之间的相互作用也是交叉重叠和具有层次性的，所以我们期望对事物进行准确描述的时候总会陷入一种两难：一方面，对事物的各种表现的观测越全面，对事物的认识就越准确和越完整；另一方面，对事物的观测越全面，得到的描述变量就越多，对事物的特性的表述却变得更加困难了！

显然，在高维度空间中描述事物比在低维度的空间中描述事物更客观，却更困难。这一矛盾如何解决呢？统计学提供了最有效的方法和手段，即下面要学习因子分析。

目录

一、基本理论

1.什么是因子分析？

因子分析是将具有错综复杂关系的变量（或样本）综合为少数几个因子，以再现原始变量和因子之间的相互关系，探讨多个能够直接测量，并且具有一定相关性的实测指标是如何受少数几个内在的独立因子所支配，并且在条件许可时借此尝试对变量进行分类。

2. 因子分析的基本思想

根据变量间相关性的大小把变量分组，使得同组内的变量之间的相关性（共性）较高，并用一个公共因子来代表这个组的变量，而不同组的变量相关性较低（个性）。

因子分析将每个原始变量分解成两部分因素，一部分是由所有变量共同具有的少数几个公共因子组成的，另一部分是每个变量独自具有的因素，即特殊因子。

3. 因子分析的目的

因子分析的目的，通俗来讲就是简化变量维数。即要使因素结构简单化，希望以最少的共同因素（公共因子），能对总变异量作最大的解释，因而抽取得因子越少越好，但抽取的因子的累积解释的变异量越大越好。

例：

在企业形象或品牌形象的研究中，消费者可以通过一个由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。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：

$$x_i=a_{i1}F_1+a_{i2}F_2+a_{i3}F_3+\varepsilon_i \quad (i=1,2,3\cdots24)$$

称 F_1 、F_2 、F_3 是不可观测的潜在因子,也称为公共因子。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。

二、因子分析模型

因子分析是通过研究多个变量间相关系数矩阵（或协方差矩阵）的内部依赖关系，找出能综合所有变量的少数几个综合指标，这几个综合指标是不可测量的，但它更能反映事物的本质，通常称为因子。各个因子间是独立的、互不相关，所有变量都可以表示成公因子的线性组合。

1. 数学模型

设有N个样本，P个指标，$X=(x_1,x_2,.....,x_p)^T$为随机向量，要寻找公共因子为$F=(F_1,F_2,.....,F_M)^T$，则模型：

$$X_1=a_{11}F_1+a_{12}F_2. \ . \ . \ +a_{1m}F_m+\varepsilon_1$$

$$X_2=a_{21}F_1+a_{22}F_2. \ . \ . \ +a_{2m}F_m+\varepsilon_2$$

$$\vdots$$

$$X_p=a_{p1}F1+a_{p2}F_2. \ . \ . \ +a_{pm}F_m+\varepsilon_p$$

被称为因子模型。

矩阵$A=(a_{ij})$称为因子载荷矩阵，a_{ij}为因子载荷，其实质就是公因子F_i和变量X_j的相关系数。ε为特殊因子，代表公因子以外的影响因素所导致的（不能被公共因子所解释的）变量变异，实际分析时忽略不计。上述模型表示成矩阵形式为：$X=AF+\varepsilon$。

对求得的公因子，需要观察它们在哪些变量上有较大的载荷，再据此说明该公因子的实际含义。但对于分析得到的初始因子模型，其因子载荷矩阵往往比较复杂，难以对公因子F_i给出一个合理的解释，此时可以考虑进一步做因子旋转，以求旋转后能得到更加合理的解释。

因子分析得到的模型有两个特点：其一，模型不受量纲的影响；其 二，因子载荷不是唯一的，通过因子轴的旋转，可以得到新的因子载荷阵，使意义更加明显。

2. 各统计量的意义

（1）特征值（Eigenwalue）：它可以被看成是公因子响力度的指标，代表引入该因子后可以解释平均多少个原始变量的信息。 如果特征值小于1，说明该因子的解释力度还不如直接引入一个原变量的平均解释力度大，因此一般可以用特征值大于1作为纳入标准。

（2）累计贡献率：前 k个主成分的累计贡献率指按照方差贡献率从大到小排列，前 k 个主成分累计提取了多少的原始信息，即前面 k 个主成分累计提取了x_1，x_2，…，x_p多少的信息。 一般来说，如果前 k 个主成分的累计贡献率达到85%，表明前 k 个主成分包含了全部测量指标所具有的主要信息，这样既减少了变量的个数，又便于对实际问题的分析和研究。

（3）因子载荷a_{ij}：因子载荷a_{ij}为第 i 个变量在第 j 个因子上的载荷，实际上就是 x_i 与 F_j 的相关系数，表示变量 x_i 依赖因子 F_j 的程度，或者说反映了第 i个变量 x_i 对于第 j 公因子 F_j 的相对重要性。其绝对值越大，则表示公因子 F_j 与 x_i 的关系越密切。

（4）变量共同度：变量共同度也称为公共方差，记为h_j^2，表示全部公因子对变量x_i的总方差所作出的贡献，或者变量x_i的信息能够被k个公因子所描述的程度，数值在0~1之间。取值越大，说明该变量能被公共因子解释的信息比例越高。变量x_i的共同度为因子载荷矩阵A中第i 行元素的平方和，即:$h_i^2=$ ，$(j=1,2\sum_{j=1}^{m}a_{ij}^2 \ldots\ldots k)$

如果大部分变量的共同度都在0.8上，则说明提取出的公因子已经基本反映了各原始变量80%以上的信息，因子分析效果理想。

（5）公因子的方差贡献：

公因子F_j的方差贡献定义为因子载荷矩阵中第 j 列元素的平方和，即：

$$S_j=\sum_{i=1}^{n}a_{ij}^2 \quad (i=1,2,3,\ldots\ldots k)$$

它所反映的是该因子对所有原始变量总方差的解释能力，其值越大，说明该因子的重要性越高。

三、因子分析的基本步骤

因子分析中需要解决两个问题：一是如何来构造少量的并且能够尽可能的反映原有信息的因子；二是如何对析取出的因子进行命名解释。

其基本步骤如下：

1. 确定待分析的原始变量是否适合进行因子分析，即进行因子分析的前提假设是否满足。

2. 因子提取

3. 因子旋转

4. 计算因子得分

1. 确定待分析的原始变量是否适合进行因子分析

由于因子分析是从众多原始变量中构造出少数几个有代表意义的因子，这就要求原变量之间具有较强的相关性。如果原变量间不存在相关关系，或者说没有共同成分的话，就无法、也没有必要再去析取因子，因为原变量本身就已经是最小的不能再缩减的变量集。

因此，因子分析时，需要对原变量进行相关分析。如果在计算出的相关矩阵，大部分相关系数都小于0.3，并且未通过统计检验，则变量不适合于进行因子分析。

此外，SPSS的因子分析过程也提供了用于检验变量是否合适于做因子分析的方法：

方法一：KMO检验

KMO（Kaiser-Meyer-Olkin)检验统计量是用于比较变量间简单相关系数和偏相关系数的指标。主要应用于多元统计的因子分析。

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：

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

（0≤KMO≤1）

其中 r_{ij}^2 是变量i和j的简单相关系数，p_{ij}^2 是变量i和变量j的偏相关系数。

如果KMO值越接近1，则越适合于做因子分析，如果KMO越小，则越不适合于做因子分析，其判断标准如下：

0.9<KMO：非常适合

0.8<KMO<0.9 :适合

0.7<KMO<0.8：一般

0.6<KMO<0.7：不太适合

KMO < 0.5：不合适

方法二：巴特利特（Bartlett ）球形检验

该检验首先假设变量相关矩阵为单位阵（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。

方法三：反映象相关矩阵检验

将偏相关矩阵中的每个元素取反，得到反映像相关矩阵。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，此时反映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。

2. 因子提取

- 因子提取方法

因子分析中，析取因子的方法有许多种，在“抽取”对话框中的“方法”下拉列表框中，可以选择不同的分析方法。

（1）主成分法：为默认选项，也是最常用的使用方法之一。

此方法将原有的P个相关变量X_i作线性变换后转成另一组不相关的变量Y_i，即：

$$y_1=u_{11}x_1+u_{21}x_2+\ldots+u_{p1}x_p$$

$$y_2=u_{12}x_1+u_{22}x_2+\cdots+u_{p2}x_p$$

.....

$$y_P=u_{1P}x_1+u_{2P}x_2+\cdots+u_{pp}x_p$$

该方程组要求：

$$u_{1k}^2+u_{2k}^2+u_{3k}^2+\cdots+u_{pk}^2=1 \quad (k=1, 2, 3, \cdots p)$$

系数u_{ij}依照两个原则来确定：

★ y_i与y_j $(i \neq j, i, j=1, 2, 3, \cdots p)$互不相关；

★ y_1是$x_1, x_2, x_3, \cdots, x_p$的一切线性组合（系数满足上述方程组）中方差最

y_2是与y_1不相关的 $x_1, x_2, x_3, \cdots, x_p$的一切线性组合中方差次大的；$y_P$是与$y_1$， y_2， y_3, $\cdots y_p$都不相关

的$x_1, x_2, x_3, \cdots, x_p$的一切线性组合中方差最小的；即 y_1，y_2，y_3，y_4，……，y_p为原有变量的第

1、第2、第3和第p个主成分。

通过选取前面几个方差最大的主成分，一方面能够用较少变量反映原有变量的绝大部分信息（一般方差的累计贡献率应大于85%），另一方面减少了数据分析和处理的复杂程度。

（2）未加权的最小平方法：该方法使实际的相关阵和再生的相关阵之差的平方和达到最小。