一、因子分析

1 因子分析的基本思想

1.1 因子分析的基本出发点

将原始指标综合成较少的指标，这些指标能够反映原始指标的绝大部分信息（方差），这些综合指标之间没有相关性。

1.2 因子变量的特点

（1）这些综合指标称为因子变量，是原变量的重造；

（2）个数远远少于原变量个数，但可反映原变量的绝大部分方差；

（3）不相关性；

（4）可命名解释性。

2 因子分析的基本步骤

（1）确认待分析的原始变量是否适合作因子分析；

（2）构造因子变量；

（3）利用旋转方法使因子变量具有可解释性；

（4）计算每个样本的因子变量得分。

3 因子分析的数学模型

数学模型（xi为标准化的原始变量；Fi为因子变量；k<p）



也可以矩阵的形式表示为：

X=AF+*ε*

F：因子变量；

A：因子载荷阵；

*a*ij：因子载荷；

*ε*：特殊因子。

4 因子分析的相关概念

（1）因子载荷

在因子变量不相关的条件下，aij就是第*i*个原始变量与第*j*个因子变量的相关系数。aij绝对值越大，则Xi与Fi的关系越强。

（2）变量的共同度(Communality)

也称公共方差。Xi的变量共同度为因子载荷矩阵A中第*i*行元素的平方和。

可见：Xi的共同度反应了全部因子变量对Xi总方差的解释能力。

（3）因子变量Fj的方差贡献

因子变量Fj的方差贡献为因子载荷矩阵A中第j列各元素的平方和

可见：因子变量Fj的方差贡献体现了同一因子Fj对原始所有变量总方差的解释能力，Sj/p表示了第j个因子解释原所有变量总方差的比例。

5 原有变量是否适合作因子分析

计算原有变量的相关系数矩阵，一般小于0.3就不适合作因子分析。

6 确定因子变量--主成份分析

6.1主成份分析法的数学模型



将原有的P个相关变量Xi作线性变换后转成另一组不相关的变量Yi

该方程组要求：

系数uij依照两个原则来确定：

yi与yj(i≠j,i,j=1,2,3,…p)互不相关；

y1是x1,x2,x3,…,xp的一切线性组合（系数满足上述方程组）中方差最大的；y2是与y1不相关的x1,x2,x3,…,xp的一切线性组合中方差次大的；yP是与y1, y2, y3,…yp都不相关的x1,x2,x3,…,xp的一切线性组合中方差最小的；

y1在总方差中所占比例最大，它综合原有变量的能力最强，其余变量在总方差中所占比例依次递减，即：其余变量综合原有变量的能力依次减弱。

6.2主成份分析的基本步骤

（1）将原始数据标准化；

（2）计算变量间简单相关系数矩阵R；

（3）求R的特征值λ1≥λ2≥λ3≥…λp≥0及对应的单位特征向量*μ*1,*μ*2*,μ*3*,…μ*p；

（4）得到：yi=*u*1i*x*1+*u*2i*x*2+…+*u*pi*x*p

6.3确定因子变量—计算因子载荷





7确定因子变量个数

确定k个因子变量

（1）根据特征值λi确定：**取特征值大于1的特征根**；

（2）根据累计贡献率：**一般累计贡献率应在70%以上**；







（3）通过观察碎石图的方式确定因子变量的个数。



8因子变量的命名解释

（1）发现：

aij的绝对值可能在某一行的许多列上都有较大的取值，或aij的绝对值可能在某一列的许多行上都有较大的取值。

（2）表明：

某个原有变量xi可能同时与几个因子都有比较大的相关关系，也就是说，某个原有变量xi的信息需要由若干个因子变量来共同解释；同时，虽然一个因子变量可能能够解释许多变量的信息，但它却只能解释某个变量的一少部分信息，不是任何一个变量的典型代表。

（3）结论：因子变量的实际含义不清楚

通过某种手段使：每个变量在尽可能少的因子上又比较高的载荷，即：在理想状态下，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。这样：一个因子变量就能够成为某个变量的典型代表，它的实际含义也就清楚了。

9计算因子得分

因子得分是因子变量构造的最终体现。

基本思想：是将因子变量表示为原有变量的线性组合，即：通过因子得分函数计算因子得分。

因子得分可看作各变量值的权数总和，权数的大小表示了变量对因子的重要程度。



[prince](https://github.com/MaxHalford/Prince)：PCA、CA、MCA、FAMD分析库

**factor analysis**

例如，一个学生的英语、数据、语文成绩都很好，那么潜在的共性因子可能是智力水平高。因此，因子分析的过程其实是寻找共性因子和个性因子并得到最优解释的过程。

因子分析有两个核心问题：一是如何构造因子变量，二是如何对因子变量进行命名解释。因子分析有下面4个基本步骤：

1. 确定原有若干变量是否适合于因子分析，因子分析的基本逻辑是从原始变量中构造出少数几个具有代表意义的因子变量，这就要求原有变量之间要具有比较强的相关性，否则，因子分析将无法提取变量间的“共性特征”（变量间没有共性还如何提取共性？）。实际应用时，可以使用相关性矩阵进行验证，如果相关系数小于0.3，那么变量间的共性较小，不适合使用因子分析。
2. 构造因子变量，因子分析中有多种确定因子变量的方法，如基于主成分模型的主成分分析法和基于因子分析模型的主轴因子法、极大似然法、最小二乘法等。其中基于主成分模型的主成分分析法是使用最多的因子分析方法之一。选取原则（1）根据特征值λi确定：**取特征值大于1的特征根**；（2）根据累计贡献率：**一般累计贡献率应在70%以上**；
3. 利用旋转使得因子变量更具有可解释性。在实际分析工作中，主要是因子分析得到因子和原变量的关系，从而对新的因子能够进行命名和解释，否则其不具有可解释性的前提下对比PCA就没有明显的可解释价值。
4. 计算因子变量的得分，因子变量确定以后，对每一样本数据，希望得到它们在不同因子上的具体数据值，这些数值就是因子得分，它和原变量的得分相对应。

方法。使您可以选择因子旋转的方法。可用的方法有最大方差、直接 Oblimin、最大四次方值、最大平衡值或最优斜交。

最大方差法 (Varimax Method). 一种正交旋转方法，它使得对每个因子有高负载的变量的数目达到最小。该方法简化了因子的解释。

直接 Oblimin 方法。一种斜交（非正交）旋转方法。当 delta 等于 0（缺省值）时，解是最斜交的。delta 负得越厉害，因子的斜交度越低。要覆盖缺省的 delta 值 0，请输入小于等于 0.8 的数。

最大四次方值法 (Quartimax Method). 一种旋转方法，它可使得解释每个变量所需的因子最少。该方法简化了观察到的变量的解释。

最大平衡值法 (Equamax Method). 一种旋转方法，它是简化因子的最大方差法与简化变量的最大四次方值法的组合。它可以使得高度依赖因子的变量的个数以及解释变量所需的因子的个数最少。

最优斜交旋转 (Promax Rotation). 斜交旋转，可使因子相关联。该旋转可比直接最小斜交旋转更快地计算出来，因此适用于大型数据集。

输出。使您可以在旋转解上包含输出以及前两个或前三个因子的载荷图。

旋转解 (Rotated Solution). 必须选择旋转方法才能获得旋转解。对于正交旋转，会显示已旋转的模式矩阵和因子转换矩阵。对于斜交旋转，会显示模式、结构和因子相关性矩阵。

载荷图 (Factor Loading Plot). 前三个因子的三维因子载荷图。对于双因子解，那么显示二维图。如果只抽取了一个因子，那么不显示图。如果要求旋转，那么图会显示旋转解。

因子旋转有称为正交变换，建立因子分析的目的不仅是找出公共因子以及对变量分组，更重要的是知道每个公共因子的含义。

由于因子载荷矩阵是不唯一的，所以应该对因子载荷矩阵进行旋转。目的是使因子载荷矩阵的结构简化，使载荷矩阵每列或者每行的元素平方值向 0 或者 1 两级分化。其方法有 3 种：

方差最大化

四次方最大旋转

等量最大法