因子分析

1. 因子分析的基本概念

因子分析与PCA

1. PCA是对原始变量的一种线性组合，对组合的解释；而因子分析更像一种回归分析，构建一些因子，再通过因子进行回归分析
2. PCA中，主成分是原始变量的线性组合；因子分析中，变量是主成分的线性组合
3. PCA分析不会因为提取个数的改变而变化；因子会随着因子的个数而变化

2. 正交因子模型

$$x = \mu + Af + \epsilon$$

模型解释

$x = (x_1, \ldots, x_p)$ 为 p 维随机向量，表示对于一个观测（个体）而言的各种指标、变量；
$\mu = (\mu_1, \ldots, \mu_p)$ 为 p 维均值向量 *【注意：该变量为公式中的非随机变量】（？对于1个obs为什么会有均值）*
A 为 $p \times m$ 的因子载荷矩阵，表示各个因子如何线性组合成为原始变量；
$f = (f_1, \ldots, f_m)$ 为 m 维因子向量，也就是因子模型中的因子；
$\epsilon = (\epsilon_1, \ldots, \epsilon_p)$ 为 p 维随机向量，相当于回归模型中的误差项

模型假设

简而言之，主要就是对f限定0均值、同单位方差；对ϵ限定0均值、不相关；对f, ϵ之间不相关。具体而言：

1. 各个因子之间互相独立（互不相关，说明彼此的信息之间没有重叠）：

   $$Var(f) = I$$

   （注：独立性主要体现在协方差部分为0；对角线为1即方差为1更多的是为了模型设定的简洁性，在后面的因子旋转等可见，对角线为1总可以通过线性变换得到）

2. 误差/特殊因子与各个因子之间互相独立：

   $$Cov(f, \epsilon) = 0$$

3. 误差/特殊因子之间互相独立：（但此处不要求同方差）

   $$ar(\epsilon) = diag(\sigma_1^2, \ldots, \sigma_p^2) := D$$

4. 为模型简洁性，要求因子和误差期望为0：

   $$E(f) = 0, E(\epsilon) = 0$$

模型性质

1. **变量x的方差分解**

因子分析模型的核心模型为：$x = \mu + Af + \epsilon$，在模型满足上述假定的基础上，可以通过纯粹的代数计算进行如下推导：

$$\Sigma_x = Var(\mu + Af + \epsilon) = AA' + V(\epsilon) = AA' + D$$

上式含义丰富，它表明，变量向量x彼此之间的协方差内容完全由因子载荷A确定（即Σ_X的非对角线部分）

进一步，对于标准化后的随机向量，方差等于相关系数，故

$$R = \Sigma_X = AA' + D$$

□

上述讨论的内容是对于一个完全符合前述假设的理论正交因子模型的性质。在实践中，我们希望通过确定一些尽可能少的因子（尽可能少以起到化简的作用）$f_1, \ldots, f_m$来尽可能地对原始变量向量的变异性进行解释，即使得

$$AA' + D \approx \Sigma_X$$

公式左侧为后面需要拟合的内容，公式右侧为实证数据的真实情况。我们希望能类似回归分析中寻找最优的β一样，找到一个尽可能近似的拟合组合$AA' + D$

2. **尺度变换不变性（因子模型不受单位影响）**

简而言之，就是对一个符合正交模型假设的变量向量x进行变换：$x^* = Cx$，其中C为正的对角矩阵，这样得到的新的变量仍然是符合前述正交因子模型假设的新变量。

上述变换的意义就在于，C矩阵可以理解成对应不同的变量分量各自有不同的单位，但不论单位怎么变换，正交因子模型分解始终成立。

3. **因子载荷不唯一性（正交旋转不变性）**

不难承认下述恒等变形：

$$x = \mu + Af + \epsilon = \mu + (AT)(T'f) + \epsilon := \mu + A^*f^* + \epsilon$$

其中，$T \in \mathbb{R}^{m \times m}$为任意正交矩阵

这说明只要对分解的因子进行正交旋转（正交变换），该分解等式总是成立的，即一个变量向量x对应着无数种分解方式（故在实证中也可以通过该性质不断旋转，以最终确定解释性最强的一组因子分解）

因子载荷矩阵A的统计性质

1. $cov(x, f) = A$
   该式可以通过协方差性质直接推得。进一步对于标准化向量x方差等于相关系数。

□

在推导后续的内容之前，首先讨论一下变量x的方差.

在前面x的方差分解中，已经提到$\Sigma_X = AA' + D$，若将这个表达式具体地对应到每个分变量x_i，将矩阵具体写开，即有：

$$var(x_i) = a_{i1}^2 + \ldots + a_{im}^2 + \sigma_i^2 := h_i^2 + \sigma_i^2 \quad (1)$$

若再求和求得各个x_i的总方差，则有：

$$\sum_{i=1}^{p} var(x_i) = \sum_{i=1}^{p} a_{i1}^2 + \ldots + \sum_{i=1}^{p} a_{im}^2 + \sum_{i=1}^{p} \sigma_i^2$$
$$:= g_1^2 + \ldots + g_m^2 + \sum \sigma_i^2 \quad (2)$$

后续会反复提到上述两个方差内容。

另外，后续三个性质都是针对A的元素展开的，上面的讨论已知A矩阵的含义即为$cov(x, f)$，因此具体地以表格将A每个元素展示如下。

	f_1	f_2	$\ldots$	f_m
x_1	$cov(x_1, f_1)$	$cov(x_1, f_2)$	$\ldots$	$cov(x_1, f_m)$
x_2	$cov(x_2, f_1)$	$cov(x_2, f_2)$	$\ldots$	$cov(x_2, f_m)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
x_p	$cov(x_p, f_1)$	$cov(x_p, f_2)$	$\ldots$	$cov(x_p, f_m)$

2. **A的行平方和【全部因子-单个变量】**
   参上，$\Sigma_X = AA' + D$，仔细观察$AA' := \mathrm{A}$，不难发现A的第i个对角线元素就是原先A的对应第i行的平方和。而由上面（1）式，这恰恰就是h_i^2。

   总结一下，对于单个变量x_i，其变异性信息（方差）包括h_i, σ_i^2两个部分，而前面的部分就是因子载荷矩阵A的第i行平方和。这一结论也是自然的，因为再上表中，每一行就表示着x_i与$f_1, \ldots, f_m$的相关系数关系，只需要记忆这种关系是通过二范数（平方和）进行描述的即可。

   为叙述方便，这里称h_i^2为公共因子$f_1, \ldots, f_m$对单个变量x_i的方差贡献，成为**共性方差**；后面的$\sigma_i^2 = var(\epsilon)$为**特殊方差** *（这里也有一个存粹记忆上的技巧：h_i^2正好对应着拼音里行的声母h，因此是行元素平方和）*

3. **A的列平方和【单个因子-全部变量】**

   通过上述（2）式可以发现，各个原始变量x_i汇总的方差也可以认为由$\sum g_i^2$和$\sum \sigma_i^2$构成。而通过观察不难发现，g_i^2就是A的第i列平方和。

   也就是说，总的原始变量x_i的汇总方差信息，由各个因子$f_1, \ldots, f_m$的方差贡献（g_i^2）之和，与特殊方差（σ_i^2）之和构成。而每个因子对全部变量的方差贡献，就是这里说的A的列平方和。

   这一结论同样是自然的，仿照2中所属，观察A的含义形式，每一列就表示一个因子f与各原始变量$x_1, \ldots, x_p$之间的方差对应关系。只不过描述的形式是通过平方和二范数表示的而已。

   统计学意义上，g_j^2反映了单个公共因子f_j对全部原始变量$x_1, \ldots, x_p$的影响，是某个因子对全部方差的贡献

程度的度量指标，表明某个因子解释了全部变量多大程度的变异性；量化地，也将某因子的方差贡献g_j^2占比全部x_i的总方差的比例，成为贡献率

注：写到这里的突发奇想。这里其实更多的展现出一种"信息"的思想。矩阵A中的元素表示了两个变量之间的交互信息（协方差矩阵），而这种交互包含了两方面内容，即信息-交互。更数学地讲，对应数学形式$\varphi(x)$，x就是这里的信息，$\varphi(\cdot)$就是这里的交互方法（为了和因子模型中的因子f(factor)相区分，这里的general形式的函数用希腊字母φ表示）。当我们对列元素求解平方和的时候，我们相当于收集到了关于$x_1,\ldots,x_p,f_j$的有关信息，后续无论如何操作，都是围绕着这些信息进行的处理。在没有额外的假设，或者补充条件的 情况下，是无法逃出这些变量的控制范围的。（即使有，可能也是一种数值上的巧合，或者是其中蕴含着更深层次的联系）。而取二范数（平方和）的操作，就相当是一种对信息的处理方式，或者交互方法，是如何整合这些信息得到我们想要的内容。

4. **A全部元素平方和【全部因子-全部变量】**
   通过矩阵运算，可以知道A全部元素平方和有两种表示方法：

$$tr(AA')=\sum_{i=1}^{p}h_i^2$$

$$tr(A'A)=\sum_{j=1}^{m}g_j^2$$

   经过上述的讨论，不难发现，全部元素平方和即是全部原始变量总方差中，去除无法解释的$\sum\sigma^2$部分后的剩余内容。相当于全部因子对总方差的一个贡献程度。

3. 因子模型的参数估计

在前面的记号中，我们认为$x=(x_1,\ldots,x_p)$是指的对于一个观测具有p个变量，把这p个变量记作一个向量。为了后面叙述方便，这里再次明确：$\mathbf{x_1,\ldots,x_n}$表示有n个样本观测$\mathbf{x}$，这里的每个$\mathbf{x}$都是一个p维的向量，即$\mathbf{x_i}=(x_{i1},\ldots,x_{ip})'$

这一小节的主要任务在于：在前述内容中，我们主要都是在讨论一个理论的因子模型设置是怎么样的，一个理论的模型具有哪些性质。就好比给实际的模型假设了一个$y=X\beta+\epsilon$的线性方程组假设一样。但是在实际的模型中，都是通过一系列参数估计对该内容进行近似逼近。本节的内容就是相当于引入实际数据，探究如何近似拟合出这样一个因子模型。

再具体一点，我们知道当有了一组数据，那么这些变量就一定能计算出一个数据之间的协方差矩阵Σ_X。我们可以进一步施加规定，人为地认为$\mathbf{x}$是由以下几个部分组成的：$\mathbf{x}=\mu+Af+\epsilon$（即我们的因子模型主体内容）。因此使用这个模型，就是要看这三部分到底都长什么样子。不过虽说是三部分，但是这里面μ是非随机部分，表示某个变量x_i对于全部样本的均值水平；ϵ是我们无法观测到的误差波动项，根本无法求；f是我们设定的内容，就好比$y=X\beta+\epsilon$中的X；最后其实真正要关注的就是这个载荷矩阵A了。因此后续的模型估计，说到底都是对于A的估计。这时，又想起对于一个完全的理论模型$\Sigma_X\equiv AA'+D$。而只要由一组数据，我们都是有Σ_X的，因此对于真实模型，我们希望尽可能地满足$\Sigma_X\approx AA'+D$，这样就可以一路回推，完成模型。后续的几个方法也是以此为基础（灵感）进行讨论的。

主成分法

对于样本的协方差矩阵S，由于协方差矩阵一定是实对称的，因此一定可以进行谱分解如下：

$$S\equiv\hat{\lambda}_1\hat{t}_1\hat{t}_1'+\ldots+\hat{\lambda}_m\hat{t}_m\hat{t}_m'+\ldots.+\hat{\lambda}_p\hat{t}_p\hat{t}_p'$$

这里的p表示变量个数，m表示因子个数。

这里通过人为选择，指定一个m使得$\sum_{i=1}^{m}\hat{\lambda}_i/\sum_{i=1}^{p}\hat{\lambda}_i$相对占比已经比较高，此时剩余的几项占比已经很低，则可以做出近似：

$$S\approx\hat{\lambda}_1\hat{t}_1\hat{t}_1'+\ldots+\hat{\lambda}_m\hat{t}_m\hat{t}_m'+\hat{D}:=\hat{A}\hat{A}'+\hat{D}$$

相当于是把略去几项的对角线项保留，存入$\hat{D}$中，非对角项则彻底抛弃，因此是约等于。又由代数运算可以知道，$\hat{A}=(\sqrt{\hat{\lambda}_1}\hat{t}_1,\ldots,\sqrt{\hat{\lambda}_m}\hat{t}_m)$

另外定义**残差矩阵**：

$$S-(\hat{A}\hat{A}'+\hat{D})$$

再另外，在许多场景下，也适合用标准化的协方差（相关系数矩阵）$\hat{R}$出发进行因子分析。

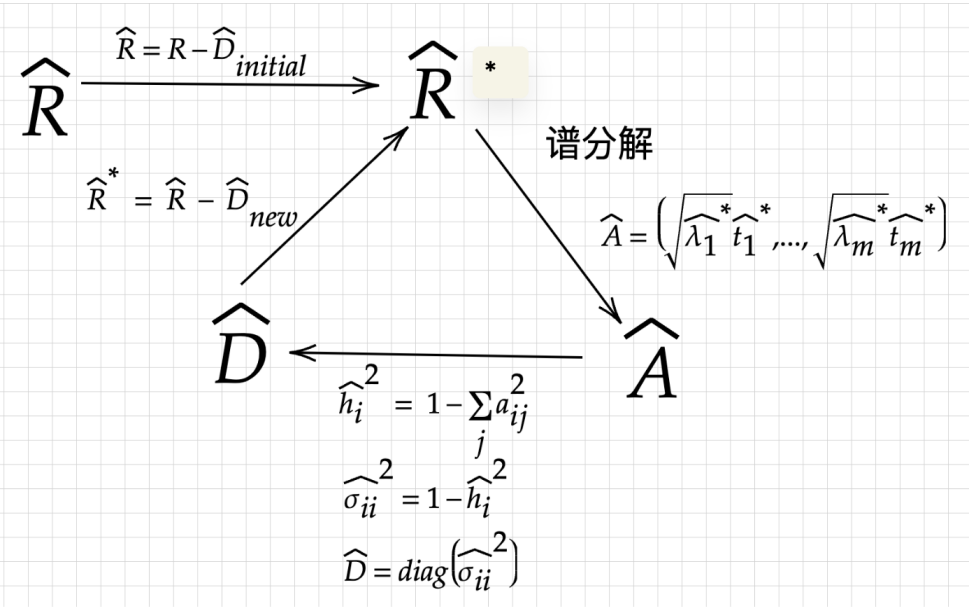
主因子法

主因子法利用的公式与上述主成分法基本相同，但是是一种通过**递归算法**求解更精确解的计算方法，其大体思想如下：

在原先的计算过程中，我们认为（对于一个符合因子模型假设的理论数据）有：$R=AA'+D$（为计算方便，这里采用的是各变量分别标准化后的结果，相当于是相关系数矩阵），则有：$R^*=R-D=AA'$。因此，若能给出一个关于随机误差项ϵ的方差较为精准的估计（即$\hat{D}$），剩余部分就可以直接谱分解了。

具体步骤如下：

1. 首先给出$\hat{D}$的初始估计（这个在后面会具体提及实证中的常用做法）
2. 根据已知实际数据计算其相关系数矩阵R_X
3. 由1、2计算：$R^*=R_X-\hat{D}$
4. 由3对R^*类似地进行谱分解，得到$R^*=T\Lambda T'=\hat{\lambda_1}\hat{t_1}\hat{t_1}'+\ldots+\hat{\lambda_m}\hat{t_m}\hat{t_m}'+\ldots+\hat{\lambda_p}\hat{t_p}\hat{t_p}'$ 并取前m个较大的部分作为近似（认为后面的余项趋近于0）；因此有$R^*\approx\hat{\lambda_1}\hat{t_1}\hat{t_1}'+\ldots+\hat{\lambda_m}\hat{t_m}\hat{t_m}'$。
5. 再将其写作$R^*=AA'$的形式，最终得到$\hat{A}$的最新估计：$\hat{A}=(\sqrt{\hat{\lambda}_1^*}\hat{t}_1^*,\ldots,\sqrt{\hat{\lambda}_m^*}\hat{t}_m^*)$
6. 再由$\hat{A}$的最新估计，更新$\hat{D}$的最新估计：由A的性质可知，其行元素平方和为$\hat{h_i^2}$，（对于一个标准化后的数据）$\hat{\sigma_i^2}=1-\hat{h_i^2}$，而$\hat{D}=\mathrm{diag}(\sigma_i^2)$，因此由这几个公式即可完成对$\hat{D}$的最新估计



D的初始估计

下面给出三种比较常用的初始估计

1. $\hat{\sigma}_i^2=\frac{1}{r^{ii}}$，其中$r_{ii}$是$\hat{R}^{-1}$的对角线元素
2. $\hat{h_i}^2=\max_{j\neq i}\left\vert r_{ij} \right\vert $
3. $\hat{h_i}^2=1$

极大似然法

极大似然和前两个方法略有不同；其主要思想还是通过对于误差项ϵ进行了分布假设，求得似然函数，再通过寻找使得似然函数最大化参数得到的。

具体来说，对于一个正交因子模型，已经对f,ϵ的均值、方差进行了假设。因此这里进一步假设二者都服从相应的正态分布，且彼此独立。则作为线性组合的结果，$\mathbf{x}\sim N_p(\mu,\Sigma_X)=N_p(\mu,AA'+D)$。这时我们便顺利得到了由参数构成的分布情况，按照MLE的一般求解规则，我们有

$$\hat{\Theta}=(\hat{\mu},\hat{A},\hat{D})=\max L(\mu,A,D)$$

再接着通过代入多元正态函数的表达式，求解极值，比较容易求得:

$$\mu=\bar{x}$$

另两项则满足如下方程：

$$\hat{\Sigma}\hat{D}^{-1}\hat{A}=\hat{A}(I_m+\hat{A}'\hat{D}^{-1}\hat{A})$$
$$\hat{D}=\mathrm{diag}(\hat{\Sigma}-\hat{A}\hat{A}')$$

其中$\hat{\Sigma}$是样本协方差矩阵。

但是需要指出，上述两个方程无法得到唯一的$\hat{A}$，额外添加约束，如令$A'D^{-1}A$为对角阵，则可以得到唯一解。该求解过程亦为数值迭代求解过程。

另外需要指出，虽然上述内容是针对正态分布假设的，但即使数据不服从正态分布，仍然可以尝试通过该MLE公式得到参数估计。最终还是要通过残差矩阵是否接近于0判断是否是一个较为合理的估计。

4. 因子旋转

因子旋转是因子分析中的一个重要步骤，其目的是为了更好地解释因子模型，使得因子与变量之间的关系更加清晰。因子旋转的基本思想是：在因子载荷矩阵A的基础上，通过线性变换，使得新的因子载荷矩阵A^*满足某种特定的性质，从而达到更好的解释效果。

其理论基础即为上面曾提到过的：因子正交旋转不变性（这里只对正交旋转进行讨论）

$x=\mu+Af+\epsilon$，其核心部分就是这个$x=Af$，若把新的因子$f=(f_1,\dots,f_p)$看作是坐标轴（基向量），则$A'=(a_1,\dots,a_p)$ 中的各个a_i相当于是在这一个坐标系下的坐标。

一般意义上，我们希望通过因子旋转使得因子载荷矩阵A尽可能每一行只有一个元素比较大，其余都相对很小（若是在标准化数据上，则为每行只有一个的绝对值接近1，其余接近0）。这个时候得到的因子往往由于彼此差距性显著，故因子的解释性往往较强。

因子旋转的性质

事实上，$A^*A^{*\prime}=AA'$，故

1. 因子旋转不改变共性方差h_i^2
2. 因子旋转不改变因子的累积贡献率
3. 因子旋转不改变残差矩阵

最大方差旋转法

5. 因子得分

在前面的建模过程中，我们已经得到的是：每个变量x_i都表示为了一些因子$f_1,\dots,f_m$的线性组合。或者说，因子模型中最重要的部分就是对于一个样本而言：$\mathbf{x}=Af$，即对于一个观测样本，其每个具体指标（变量x_i）都可以认为是因子的一些线性组合。而线性组合的系数就表示了某个因子与某个具体变量之间的协方差。通过竖着解读因子载荷矩阵，看某个因子f_i和哪些变量之间的关联程度很强，和其他哪些的关联程度很弱，就可以大致理解这个因子表示的是哪个方面的指标。这样就相当于认为构造了几个新的自变量（因子），每个自变量（因子）可以把具有类似作用/含义的原始变量x_i归纳成一个更概括性的指标。

但是由于因子载荷矩阵描述的是f,x之间的关联程度，还不涉及到x的具体取值。因此上一段讲述的因子分析的作用，还是只局限于把许多自变量x_i进行一个类似聚类的处理。但是比如在金融交易市场中，量化投资者就希望把一系列复杂的观测变量（如股价、波动率、成交量、成交频率）等等抽象成几个大的因子，并且后续就利用这几个因子继续进行操作。换言之，进一步，我们不满足于仅仅得到f,x的关系，我们还想通过x的取值，再加上f,x的相互作用关系，得到一系列f的取值，后续就可以利用这几个比较抽象的，少数的几个f的“得分”进行其他定量分析。

加权最小二乘

因子模型$x=\mu+Af+\epsilon$，事实上很像OLS的一般模型假定：$y=X\beta+\epsilon$，因此完全可以借用回归分析的做法进行估计。

稍有不同的是，OLS的模型假定ϵ是同方差的，但因子模型中无此假设。故应当使用加权最小二乘进行估计，称该WLS结果为Bartlett's 因子得分，具体地（下表达式完全可以通过WLS的$\hat{\beta}$结果类比得到：

$$\hat{\mathbf{f}}=(\hat{f_1},\dots,\hat{f_m})=(A'D^{-1}A)^{-1}\hat{A}'\hat{D}^{-1}(x-\mu)$$

在实际数据中，用样本估计值$\hat{A},\hat{D},\bar{x}$分别替代上表达式即可得到因子得分。

回归法

回归法大致的思路如下：在MSE准则下，f最优的预测值就是条件期望，即$\hat{f}=E(f|x)$，故回归法的重点就是求解这个条件期望。

假设$(f;\epsilon)'$服从多元联合正态分布，则$(f;x)'$亦服从多元正态分布；而多元正态分布中有关于给定多元正态中的一部分的条件下其余部分分布的一个现成结论，应用此结论可以直接求得条件期望。

$$\hat{f}=A'(AA'+D)^{-1}(x-\mu)\equiv(I+A'D^{-1}A)^{-1}A'D^{-1}$$