因子分析

# 目录

前言

一、什么是因子分析？

1.1 因子分析应用背景

1.2 因子分析算法的基本步骤

1.3 因子分析算法的数学解释

1.3.1 因子模型

1.3.2 因子载荷矩阵的求解

1.3.3 因子载荷矩阵的旋转

1.3.4 因子得分

二、因子分析的应用实例

三、主成分分析（PCA）与因子分析（FA）的联系与区别

总结

# 前言

在学习数据降维时，了解到因子分析（FA）算法是其中的一种方式，因此，在这里对因子分析算法做一个简要的归纳、梳理。

# 一、什么是因子分析？

因子分析法是指： 研究从变量群中提取共性因子的统计技术，这里的共性因子指的是不同变量之间内在的隐藏因子。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。

其基本思想是： 根据相关性大小把变量分组，使得同组内的变量之间相关性较高，但不同组的变量不相关或相关性较低，每组变量代表一个基本结构一即公共因子。

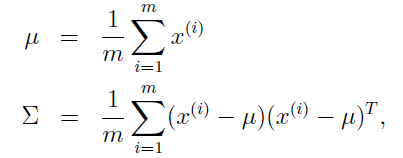
因子分析有两个核心问题： 一是如何构造因子变量，二是如何对因子变量进行命名解释。

因子分析类型： R型因子分析与Q型因子分析，就像聚类分析分为R型和Q型一样，R型的因子分析是对变量作因子分析，Q型因子分析是对样品作因子分析，本文是以R型因子分析展开。

## 1.1 因子分析应用背景

因子分析用于处理高斯数据，主要应用于以下两种情形：

假如有 m 个样本，每个样本的维度是 n, 如果 n » m；这时哪怕拟合出一个高斯模型都很困难，更不用说高斯混合, 为什么呢?其实，这和解多元线性方程组是一样的道理，就是自变量的个数多于非线性相关的方程的个数，这必然导致解的不唯一，虽然在解方程的时候可以随便选一个解满足方程组，但是对于某一实际数据集，往往样本对应的概率分布在客观上都是唯一的，只是我们无法简单地用概率论中的几个典型的分布准确表示出来罢了！

m 个样本的维度都较低。用高斯分布对数据建模，用最大似然估计去估计均值（期望）和方差：

我们会发现，协方差矩阵 Σ 是奇异的，即 Σ 不可逆，Σ-1 不存在，且有：

但是这两项在计算多元高斯分布时，又都是必不可少的。所以，除非 m 比 n 大一定较合适的数值，否则对方差和均值的最大似然估计将会很难找到正确的值。

## 1.2 因子分析算法的基本步骤

应用因子分析算法时，常常有如下几个基本步骤：

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；也可以用KMO 和 Bartlett 的检验来判断是否适合做因子分析，一般来说KMO的值越接近于1越好，大于zhi0.5的话适合做因dao子分析，你的KMO值是0.674大于0.5。Bartlett 的检验主要看Sig.越小越好，你的接近于0.由此可以得出，你的数据适合做因子分析。

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。

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。

计算因子变量的得分 。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。

具体而言：

(1) 相关性检验，一般采用KMO检验法和Bartlett球形检验法两种方法来对原始变量进行相关性检验；

(2) 输入原始数据Xn\*p，计算样本均值和方差，对数据样本进行标准化处理；

(3) 计算样本的相关矩阵R；

(4) 求相关矩阵R的特征根和特征向量；

(5) 根据系统要求的累积贡献率确定公共因子的个数；

(6) 计算因子载荷矩阵A；

(7) 对载荷矩阵进行旋转，以求能更好地解释公共因子；

(8) 确定因子模型；

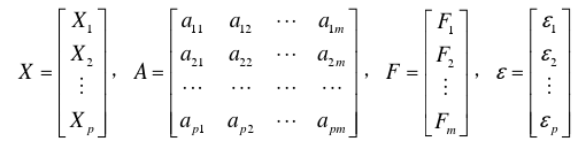
(9) 根据上述计算结果，求因子得分，对系统进行分析

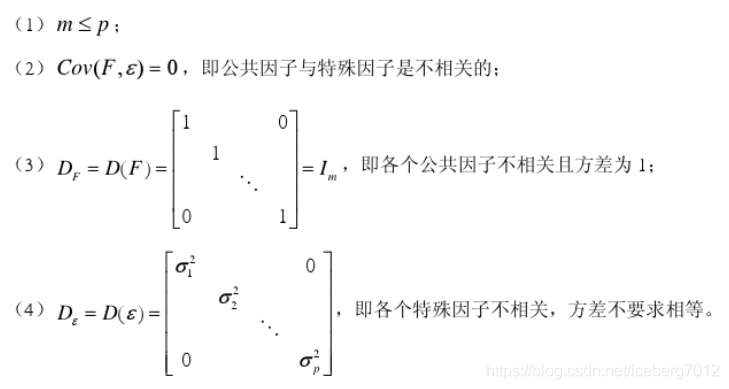
## 1.3 因子分析算法的数学解释

### 1.3.1 因子模型

因子分析中的公共因子是不可直接观测但又客观存在的共同影响因素，每个变量都可以表示成公共因子的线性函数与特殊因子之和，即：

式中的F1，F2，…，Fm称为公共因子，εi称为Xi的特殊因子。该模型可用矩阵表示为：

这里：

且满足：

模型中的矩阵A称为因子载荷矩阵，aij称为因子“载荷”，是第i个变量在第j个因子上的负荷，如果把变量Xi看成m维空间中的一个点，则aij表示它在坐标轴Fj上的投影。

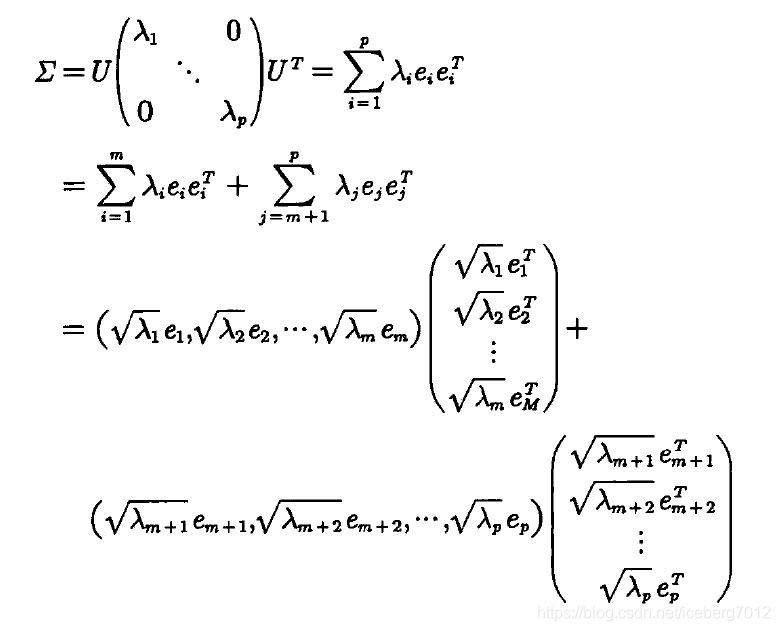
### 1.3.2 因子载荷矩阵的求解

因子载荷矩阵的求解方法有很多，主要有以下三种：主成分分析法；主因子法；极大似然估计法。（其中以主成分分析法最为常用）

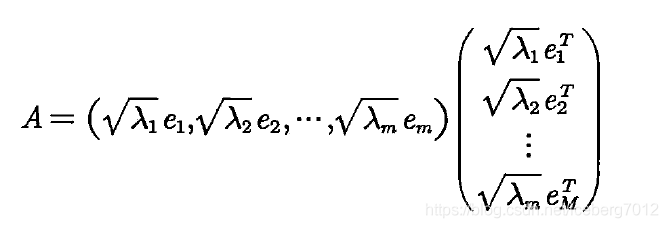
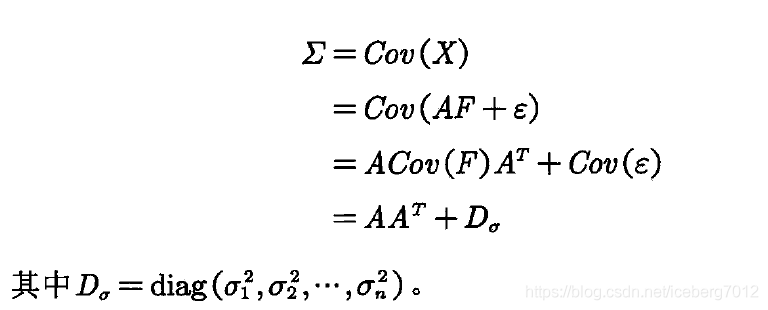
#### 1.主成分分析法

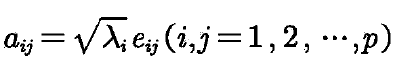
原理及主要计算步骤：

（1）计算原始数据Ｘ的协方差阵Σ；

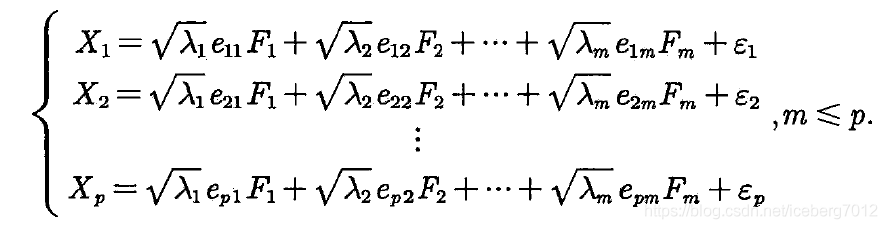
（2）计算协方差阵Σ的特征根，按数值大小表示为λ\_1≥λ\_2≥⋯≥λ\_p,相应的单位特征向量表示为e\_1,e\_2,…,e\_p，特征向量矩阵表示为U。此时协方差阵Σ有如下表示方式1：

基于公式1和模型假设，我们还可以得到协方差阵Σ有如下式2的表示方法：

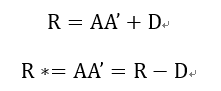
结合公式1和2，我们可以得到因子载荷矩阵的估计：

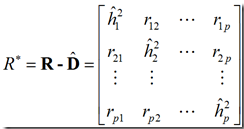
其中：

其中λ\_i表示第i个特征值，e\_ij表示λ\_i相对应的第i个特征向量的第j个分量。

得到载荷矩阵后，我们可以将因子模型表示为：

#### 2.主因子法

主因子方法是对主成分方法的修正，假定我们首先对变量进行标准化变换。则

称R为约相关矩阵，R对角线上的元素是h\_i\_2，而不是1。设h^\_i\_2是h\_i\_2的初始估计，则：

#### 3.极大似然估计法

详见极大似然估计法

### 1.3.3 因子载荷矩阵的旋转

设Q为m阶正交矩阵，令B=AQ，则：

由于，上一小节得到的因子载荷矩阵Ａ并不是唯一的，事实上对矩阵A做正交变换后得到的新矩阵都可以看作因子载荷矩阵。

我们在得到一个因子载荷矩阵的估计时，有可能会出现多个变量均在同一个因子上出现较大因子载荷，或者一个变量在多个因子上具有较大的载荷，此时很难对因子进行解释或命名，此时我们希望通过对因子载荷矩阵进行旋转得到新的简化后的因子载荷矩阵，新的因子载荷之间区分度更高，便于因子分析和命名。

载荷矩阵的旋转分正交旋转和斜交旋转两类。正交旋转的常用方法有方差最大法、四次方最大法和等量最大法。斜交旋转常用方法有最小斜交旋转法、四次方最小法、斜交旋转等。

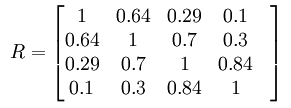
### 1.3.4 因子得分

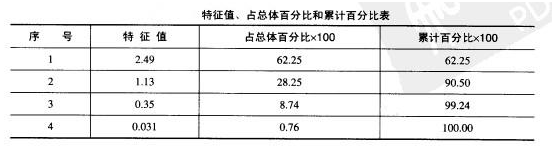
得到因子载荷矩阵之后，有时我们希望利用公共因子进行其他研宄，比如进行聚类分析或回归分析，此时我们希望能通过原始变量对公共因子进行估计，即得到因子得分。

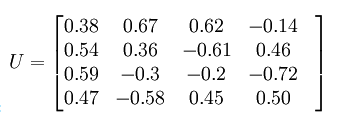
对于模型X=AF+ε，如果不考虑特殊因子ε的影响，可以得到X=AF，但矩阵A是pXm阶，模型中我们要求m<=p，通常因子个数远小于变量个数，即m<p，因此载荷矩阵不可逆，无法直接得到F的估计。

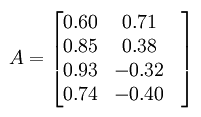
公共因子估计的常用方法是回归法和Bartlett法（加权最小二乘法）。

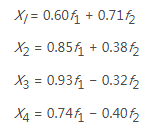
# 二、因子分析的应用实例

假设某一社会经济系统问题，其主要特性可用4个指标表示，它们分别是生产、技术、交通和环境。其相关矩阵为：

相应的特征值、占总体百分比和累计百分比如下表：

对应特征值的特征向量矩阵为：

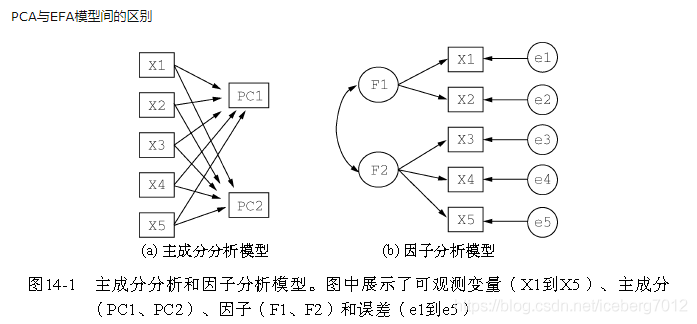
假如要求所取特征值反映的信息量占总体信息量的90％以上，则从累计特征值所占百分比看，只需取前两项即可。也就是说，只需取两个主要因子。对应于前两列特征值的特征向量，可求的其因子载荷矩阵A为：

于是，该问题的因子模型为：

因子分析：由以上可以看出，两个因子中，f1是全面反映生产、技术、交通和环境的因子，而f2却不同，它反映了对生产和技术这两项增长有利，而对交通和环境增长不利的因子。也就是说，按照原有统计资料得出的相关矩阵分析的结果是如果生产和技术都随f2增长了，将有可能出现交通紧张和环境恶化的问题，f2反映了这两方面的相互制约状况。

Python编程应用示例见：因子分析（KMO检验和Bartlett’s球形检验）

# 三、主成分分析（PCA）与因子分析（FA）的联系与区别

主成分分析（PCA）是一种数据降维技巧，它能将大量相关变量转化为一组很少的不相关变量，这些无关变量称为主成分。探索性因子分析（EFA）是一系列用来发现一组变量的潜在结构的方法。它通过寻找一组更小的、潜在的或隐藏的结构来解释已观测到的、显式的变量间的关系。

主成分（PC1和PC2）是观测变量（X1到X5）的线性组合。形成线性组合的权重都是通过最大化各主成分所解释的方差来获得，同时还要保证个主成分间不相关。相反，因子（F1和F2）被当做是观测变量的结构基础或“原因”，而不是它们的线性组合。代表观测变量方差的误差（e1到e5）无法用因子来解释。图中的圆圈表示因子和误差无法直接观测，但是可通过变量间的相互关系推导得到。

两者之间的区别与联系，具体而言有如下几种：

联系：

1. PCA和因子分析都是数据降维的重要方法，都对原始数据进行标准化处理，都消除了原始指标的相关性对综合评价所造成的信息重复的影响，都属于因素分析法，都基于统计分析方法；

2. 二者均应用于高斯分布的数据，非高斯分布的数据采用ICA算法；

3. 二者构造综合评价时所涉及的权数具有客观性，在原始信息损失不大的前提下，减少了后期数据挖掘和分析的工作量。

区别：

1. 原理不同； PCA的基本原理是利用降维（线性变换)的思想，在损失很少信息的前提下把多个指标转化为几个不相关的主成分，每个主成分都是原始变量的线性组合；

而FA基本原理是从原始变量相关矩阵内部的依赖关系出发，把因子表达成能表示成少数公共因子和仅对某一个变量有作用的特殊因子的线性组合（因子分析是主成分的推广，相对于主成分分析，更倾向于描述原始变量之间的相关关系）；

2.假设条件不同； 主成分分析不需要有假设，而因子分析需要假设各个共同因子之间不相关，特殊因子（specificfactor）之间也不相关，共同因子和特殊因子之间也不相关；

3. 求解方法不同； 主成分分析的求解方法从协方差阵出发，而因子分析的求解方法包括主成分法、主轴因子法、极大似然法、最小二乘法、a因子提取法等；

4. 降维后的“维度”数量不同，即因子数量和主成分的数量； 主成分分析的数量最多等于维度数；而因子分析中的因子个数需要分析者指定（SPSS和SAS根据一定的条件自动设定，只要是特征值大于1的因子主可进入分析），指定的因子数量不同而结果也不同。

5. 线性表示方法不同； 因子分析是把变量表示成各公因子的线性组合；主成分分析中则是把主成分表示成各变量的线性组合。

6. 主成分和因子的变化不同； 主成分分析：当给定的协方差矩阵或者相关矩阵的特征值唯一时，主成分一般是固定的独特的；因子分析：因子不是固定的，可以旋转得到不同的因子。

7.解释重点不同； 主成分分析：重点在于解释个变量的总方差；因子分析：则把重点放在解释各变量之间的协方差。

8.算法上的不同； 主成分分析：协方差矩阵的对角元素是变量的方差；因子分析：所采用的协方差矩阵的对角元素不在是变量的方差，而是和变量对应的共同度（变量方差中被各因子所解释的部分）。

9.优点不同； 对于因子分析，可以使用旋转技术，使得因子更好的得到解释，因此在解释主成分方面因子分析更占优势；其次因子分析不是对原有变量的取舍，而是根据原始变量的信息进行重新组合，找出影响变量的共同因子，化简数据；如果仅仅想把现有的变量变成少数几个新的变量（新的变量几乎带有原来所有变量的信息）来进入后续的分析，则可以使用主成分分析，不过一般情况下也可以使用因子分析。

综合来看，因子分析在实现中可以使用旋转技术，因此可以得到更好的因子解释，这一点比主成分占优势；另外，因子分析不需要舍弃原有变量，而是站到原有变量间的共性因子作为下一步应用的前提，其实就是由表及里去发现内在规律。但是，主成分分析由于不需要假设条件，并且可以最大限度的保持原有变量的大多数特征，因此适用范围更广泛，尤其是宏观的未知数据的稳定度更高。

总结

因子分析跟主成分分析一样，由于侧重点都是进行数据降维，因此很少单独使用，大多数情况下都会有一些模型组合使用。例如：

(1) 因子分析（主成分分析）+多元回归分析：判断并解决共线性问题之后进行回归预测；

(2) 因子分析（主成分分析）+聚类分析：通过降维后的数据进行聚类并分析数据特点，但因子分析会更适合，原因是基于因子的聚类结果更容易解释，而基于主成分的聚类结果很难解释；

(3) 因子分析（主成分分析）+分类：数据降维（或数据压缩）后进行分类预测，这也是常用的组合方法。

因子分析通过寻找公共因子的方式达到数据降维的目的（因子分析还可以用于分析不同变量之间的内在联系），主成分分析则是求特征矩阵，实现数据的降维。

因子分析的主要作用：

(1) 寻求基本数据结构；

(2) 用少数因子，描述具有相关性的多个指标；

(3) 数据简化，即降维。

1) 强相关问题会对分析带来困难

2) 通过因子分析可以找出少数的几个因子替代原来的变量做回归分析、聚类分析和判别分析