**主成分分析法**

PCA（Principal Component Analysis）是一种常用的数据分析方法。PCA通过线性变换将原始数据变换为一组各维度线性无关的表示，可用于提取数据的主要特征分量，常用于高维数据的降维。

作者：匿名用户

链接：https://www.zhihu.com/question/35889196/answer/168433870

来源：知乎

著作权归作者所有。商业转载请联系作者获得授权，非商业转载请注明出处。

通过主成分分析所得来的新变量是原始变量的线性组合，每个主成分都是由原有P个变量线组合得到，在诸多主成分z中，Z1在总方差中占的比重最大，说明它综合原有变量的能力最强，其余主成分在总方差中占的比重依次递减，说明越往后的主成分综合原信息的能力越弱。以后的分析可以用前面几个方差最大的主成分来进行，一般情况下，要求前几个z所包含的信息不少于原始信息的85％ ，这样既减少了变量的数目，又能够用较少的主成分反映原有变量的绝大部分信息。如利用主成分来消除多元回归方程的多重共线性，利用主成分来筛选多元线性回归方程中的变量等。通过因子分析得来的新变量是对每一个原始变量进行内部剖析。打比喻来说，原始变量就如成千上万的糕点，每一种糕点的原料都有面粉、油、糖及相应的不同原料，这其中，面粉、油、糖是所有糕点的共同材料，这正好象是因子分析中的新变量即因子变量。正确选择因子变量后，如果想考虑成千上万糕点的物价变动， 只需重点考虑面粉、油、糖等公共因子的物价变动即可。所以因子分析不是对原始变量的重新组合，而是对原始变量进行分解，分解为公共因子与特殊因子两部分。即因子分析就是要利用少数几个公共因子去解释较多个要观测变量中存在的复杂关系，它把原始变量分解为两部分因素，一部分是由所有变量共同具有的少数几个公共因子构成的，另一部分是每个原始变量独自具有的因素，即特殊因子。1，因子分析中是把变量表示成各因子的线性组合，而主成分分析中则是把主成分表示成各个变量的线性组合。 在主成分分析中,最终确定的新变量是原始变量的线性组合,如原始变量为x1 ,x2 ,. . . ,x3 ,经过坐标变换,将原有的p个相关变量xi 作线性变换,每个主成分都是由原有p 个变量线性组合得到。在诸多主成分Zi 中,Z1 在方差中占的比重最大,说明它综合原有变量的能力最强,越往后主成分在方差中的比重也小,综合原信息的能力越弱。2，主成分分析的重点在于解释个变量的总方差，而因子分析则把重点放在解释各变量之 间的协方差。3，主成分分析中不需要有假设(assumptions),因子分析则需要一些假设。因子分析的假设包括：各个共同因子之间不相关，特殊因子（specific factor）之间也不相关，共同因子和特殊因子之间也不相关。4，主成分分析中，当给定的协方差矩阵或者相关矩阵的特征值是唯一的时候，的主成分 一般是独特的；而因子分析中因子不是独特的，可以旋转得到不到的因子。5，在因子分析中，因子个数需要分析者指定（spss根据一定的条件自动设定，只要是特 征值大于1的因子进入分析），而指 定的因子数量不同而结果不同。在主成分分析中，成分的数量是一定的，一般有几个变量就有几个主成分。和主成分分析相比，由于因子分析可以使用旋转技术帮助解释因子，在解释方面更加有优势。大致说来，当需要寻找潜在的因子，并对这些因子进行解释的时候， 更加倾向于使用因子分析，并且借助旋转技术帮助更好解释。而如果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。当然，这中情况也可以使用因子得分做到。所以这种区分不是绝对的。总得来说，主成分分析主要是作为一种探索性的技术，在分析者进行多元数据分析之前 ，用主成分分析来分析数据，让自己对数据有一个大致的了解是非常重要的。主成分分析一般很少单独使用：a，了解数据。(screening the data),b,和cluster analysis一 起使用，c，和判别分析一起使用，比如当变量很多，个案数不多，直接使用判别分析可能无解，这时候可以使用主成份发对变量简化。（reduce dimensionality）d,在多元回归中，主成分分析可以帮助判断是否存在共线性（条件指数），还可以用来处理共线性。在算法上，主成分分析和因子分析很类似，不过，在因子分析中所采用的协方差矩阵的对角元素不再是变量的方差，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）。应用中的优缺点比较

(一) 主成分分析

1、优点。首先它利用降维技术用少数几个综合变量来代替原始多个变量,这些综合变量集中了原始变量的大部分信息。其次它通过计算综合主成分函数得分,对客观经济现象进行科学评价。再次它在应用上侧重于信息贡献影响力综合评价。

2、缺点。当主成分的因子负荷的符号有正有负时,综合评价函数意义就不明确。命名清晰性低。

(二) 因子分析

1、优点:第一它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;第二,它通过旋转使得因子变量更具有可解释性,命名清晰性高。2、缺点:在计算因子得分时,采用的是最小二乘法,此法有时可能会失效。