
SPSS实现主成分分析与因子分析（一）

作者：王江源 来源：爱科学

本文原地址：<https://www.iikx.com/news/statistics/6060.html>

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

SPSS实现主成分分析与因子分析(一)

一、主成分分析

(1)问题提出

在问题研究中，为了不遗漏和准确起见，往往会面面俱到，取得大量的指标来进行分析。比如为了研究某种疾病的影响因素，我们可能会收集患者的人口学资料、病史、体征、化验检查等等数十项指标。如果将这些指标直接纳入多元统计分析，不仅会使模型变得复杂不稳定，而且还有可能因为变量之间的多重共线性引起较大的误差。有没有一种办法能对信息进行浓缩，减少变量的个数，同时消除多重共线性？

这时，主成分分析隆重登场。

(2)主成分分析的原理

主成分分析的本质是坐标的旋转变换，将原始的 n 个变量进行重新的线性组合，生成 n 个新的变量，他们之间互不相关，称为 n 个“成分”。同时按照方差最大化的原则，保证第一个成分的方差最大，然后依次递减。这 n 个成分是按照方差从大到小排列的，其中前 m 个成分可能就包含了原始变量的大部分方差(及变异信息)。那么这 m 个成分就成为原始变量的“主成分”，他们包含了原始变量的大部分信息。

注意得到的主成分不是原始变量筛选后的剩余变量，而是原始变量经过重新组合后的“综合变量”。

我们以最简单的二维数据来直观的解释主成分分析的原理。假设现在有两个变量 X_1 、 X_2 ，在坐标上画出散点图如下：

可见，他们之间存在相关关系，如果我们将坐标轴整体逆时针旋转45°，变成新的坐标系Y1、Y2，如下图：

根据坐标变化的原理，我们可以算出：

$$Y1 = \sqrt{2}/2 * X1 + \sqrt{2}/2 * X2$$

$$Y2 = \sqrt{2}/2 * X1 - \sqrt{2}/2 * X2$$

其中 \sqrt{x} 为x的平方根。

通过对X1、X2的重新进行线性组合，得到了两个新的变量Y1、Y2。

此时，Y1、Y2变得不再相关，而且Y1方向变异(方差)较大，Y2方向的变异(方差)较小，这时我们可以提取Y1作为X1、X2的主成分，参与后续的统计分析，因为它携带了原始变量的大部分信息。

至此我们解决了两个问题：降维和消除共线性。

对于二维以上的数据，就不能用上面的几何图形直观的表示了，只能通过矩阵变换求解，但是本质思想是一样的。

二、因子分析

(一)原理和方法：

因子分析是主成分分析的扩展。

在主成分分析过程中，新变量是原始变量的线性组合，即将多个原始变量经过线性(坐标)变换得到新的变量。

因子分析中，是对原始变量间的内在相关结构进行分组，相关性强的分在一组，组间相关性较弱，这样各组变量代表一个基本要素(公共因子)。通过原始变量之间的复杂关系对原始变量进行分解，得到公共因子和特殊因子。将原始变量表示成公共因子的线性组合。其中公共因子是所有原始变量中所共同具有的特征，而特殊因子则是原始变量所特有的部分。因子分析强调对新变量(因子)的实际意义的解释。

举个例子：

比如在市场调查中我们收集了食品的五项指标(x1-x5):味道、价格、风味、是否快餐、能量，经过因子分析，我们发现了：

$$x1 = 0.02 * z1 + 0.99 * z2 + e1$$

$$x2 = 0.94 * z1 - 0.01 * z2 + e2$$

$$x_3 = 0.13 * z_1 + 0.98 * z_2 + e_3$$

$$x_4 = 0.84 * z_1 + 0.42 * z_2 + e_4$$

$$x_5 = 0.97 * z_1 - 0.02 * z_2 + e_1$$

(以上的数字代表实际为变量间的相关系数，值越大，相关性越大)

第一个公因子 z_1 主要与价格、是否快餐、能量有关，代表“价格与营养”

第二个公因子 z_2 主要与味道、风味有关，代表“口味”

e_1-5 是特殊因子，是公因子中无法解释的，在分析中一般略去。

同时，我们也可以将公因子 z_1 、 z_2 表示成原始变量的线性组合，用于后续分析。

(二)使用条件：

(1)样本量足够大。通常要求样本量是变量数目的5倍以上，且大于100例。

(2)原始变量之间具有相关性。如果变量之间彼此独立，无法使用因子分析。在SPSS中可用KMO检验和Bartlett球形检验来判断。

(3)生成的公因子要有实际的意义，必要时可通过因子旋转(坐标变化)来达到。

三、主成分分析和因子分析的联系与区别

联系：两者都是降维和信息浓缩的方法。生成的新变量均代表了原始变量的大部分信息且互相独立，都可以用于后续的回归分析、判别分析、聚类分析等等。

区别：

(1)主成分分析是按照方差最大化的方法生成的新变量，强调新变量贡献了多大比例的方差，不关心新变量是否有明确的实际意义。

(2)因子分析着重要求新变量具有实际的意义，能解释原始变量间的内在结构。

下一篇文章，将介绍主成分分析和因子分析的在SPSS中的实现。

更多统计方法 请访问 <https://www.iikx.com/news/statistics/>

本文版权归原作者所有，请勿用于商业用途，[爱科学iikx.com](http://iikx.com)转发