# 主成分分析法

## 简介：

是一种使用最广泛的数据降维算法，（非监督的极其学习方法），通过正交变换将一组可能存在相关性的变量数据，转换为一组线性不相关的变量，转换后的变量被称为主成分。

多变量之间可能存在相关性，从而增加了问题分析的复杂性。

## 方式：

将多个变量通过线性变换以选出较少个数重要变量的一种多元统计分析方法。由原始指标综合形成的几个新指标。依据主成分所含信息量的大小成为第一主成分，第二主成分等等。 PCA 实质上是一个基变换，使得变换后的数据有最大的方差，也就是通过对坐标轴的旋转和坐标原点的平移，使得其中一个轴（主轴）与数据点之间的方差最小，坐标转换后去掉高方差的[正交](https://so.csdn.net/so/search?q=%E6%AD%A3%E4%BA%A4&spm=1001.2101.3001.7020" \t "https://blog.csdn.net/wujianing_110117/article/details/_blank)轴，得到降维数据集。

## 目的：

在减少需要分析的指标同时，尽量减少原指标包含信息的损失，以达到对所收集数据进行全面分析的目的。由于各变量之间存在一定的相关关系，因此可以考虑将关系紧密的变量变成尽可能少的新变量，使这些新变量是两两不相关的，那么就可以用较少的综合指标分别代表存在于各个变量中的各类信息。

主成分与原始变量之间的关系：  
（1）主成分保留了原始变量绝大多数信息。  
（2）主成分的个数大大少于原始变量的数目。  
（3）各个主成分之间互不相关。  
（4）每个主成分都是原始变量的线性组合。

## 实现方法：

### 方法一：（使用SVD进行降维）

假定有  维数据样本 ，共有  个样本，每行是  维， 实矩阵可以分解为：

SVD分解矩阵A的步骤：

* 求的特征值、特征向量，用单位化的特征向量构成
* 求的特征值、特征向量，用单位化的特征向量构成
* 求或的特征值、特征向量，构成

中的大于0的数字，前个大于0的个数，就是最大主成分个数，对应的中的特征向量，就是最大特征向量组。

对原数据进行降维

### 方法二：（使用协方差矩阵和散度矩阵）

样本均值：



样本方差：



样本X和Y的协方差：







得出相关系数矩阵，求得特征值特征向量，特征向量转置后构成新的矩阵P，用P的第一行数据和X相乘，达到降维的效果。

### 方法三：（特征是分解矩阵）

对A矩阵求解特征值特征向量





将特征值代入求得特征向量



### 方法四：

1. **对原始数据计算均值和方差，进行标准化计算处理**

假定进行因子分析的指标变量有p个：，一共有n个评价对象，

评价对象为行，指标变量为列，

第i个评价对象的第j个指标的取值为



标准化公式：



样本均值：



样本标准差：



标准化公式：



1. **求样本相关系数矩阵R**

****

计算公式：



1. **求相关系数矩阵的特征根 和相应的标准正交的特征向量**

根据****求出特征值****，以及对应的特征向量****

****

由特征向量组成p个新的指标变量：

****

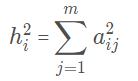
其中y1--yp都是主成分

1. **确定公共因子数**

<1>根据方差贡献率（方差解释率）和累计方差贡献率之和获取公共因子，

<2>还可以根据特征向量获取公共因子（特征向量不全，有0，则排除），

<3>还可以用求变量共同度寻找共同度高的因子



1. **计算综合评价值**
2. 计算特征值的信息贡献率和累计贡献率

信息贡献率：



累计贡献率：



当α接近（0.85、0.90、0.95），则选择前j个指标变量为j个主成分，

1. 计算综合得分



（3）