主成分分析的计算步骤

样本观测数据矩阵为：

$$X=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

第一步：对原始数据进行标准化处理

$$x_{ij}^{*}=\frac{x_{ij}-\bar{x}_j}{\sqrt{\mathrm{var}(x_j)}} \quad (i=1,2,\cdots,n;\ j=1,2,\cdots,p)$$

其中 $\bar{x}_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij}$

$$\mathrm{var}(x_j)=\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2$$

$$(j=1,2,\cdots,p)$$

第二步：计算样本相关系数矩阵

$$R=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix}$$

为方便，假定原始数据标准化后仍用 X 表示，则经标准化处理后的数据的相关系数为:

$$r_{ij}=\frac{1}{n-1}\sum_{t=1}^{n}x_{ti}x_{tj}$$

$$(i,j=1,2,\cdots,p)$$

第三步：用雅克比方法求相关系数矩阵 R 的特征值（$\lambda_1,\lambda_2\cdots\lambda_p$）和相应的特征向量 $a_i=(a_{i1},a_{i2},\cdots a_{ip}),i=1,2\cdots p$。

第四步：选择重要的主成分，并写出主成分表达式

主成分分析可以得到 p 个主成分，但是，由于各个主成分的方差是递减的，包含的信息量也是递减的，所以实际分析时，一般不是选取 p 个主成分，而是根据各个主成分累计贡献率的大小选取前 k 个主成分，这里贡献率就是指某个主成分的方差占全部方差的比重，

实际也就是某个特征值占全部特征值合计的比重。即

$$\text{贡献率}=\frac{\lambda_i}{\sum_{i=1}^{p}\lambda_i}$$

贡献率越大，说明该主成分所包含的原始变量的信息越强。主成分个数k的选取，主要根据主成分的累积贡献率来决定，即一般要求累计贡献率达到85%以上，这样才能保证综合变量能包括原始变量的绝大多数信息。

另外，在实际应用中，选择了重要的主成分后，还要注意主成分实际含义解释。主成分分析中一个很关键的问题是如何给主成分赋予新的意义，给出合理的解释。一般而言，这个解释是根据主成分表达式的系数结合定性分析来进行的。主成分是原来变量的线性组合，在这个线性组合中个变量的系数有大有小，有正有负，有的大小相当，因而不能简单地认为这个主成分是某个原变量的属性的作用，线性组合中各变量系数的绝对值大者表明该主成分主要综合了绝对值大的变量，有几个变量系数大小相当时，应认为这一主成分是这几个变量的总和，这几个变量综合在一起应赋予怎样的实际意义，这要结合具体实际问题和专业，给出恰当的解释，进而才能达到深刻分析的目的。

第五步：计算主成分得分

根据标准化的原始数据，按照各个样品，分别代入主成分表达式，就可以得到各主成分下的各个样品的新数据，即为主成分得分。具体形式可如下。

$$\begin{pmatrix} F_{11} & F_{12} & \cdots & F_{1k} \\ F_{21} & F_{22} & \cdots & F_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ F_{n1} & F_{n2} & \cdots & F_{nk} \end{pmatrix}$$

第六步：依据主成分得分的数据，则可以进行进一步的统计分析

其中，常见的应用有主成份回归，变量子集合的选择，综合评价等。