主成分分析法原理简介

1.什么是主成分分析法

主成分分析也称主分量分析，是揭示大样本、多变量数据或样本之间内在关系的一种方法，旨在利用降维的思想，把多指标转化为少数几个综合指标，降低观测空间的维数，以获取最主要的信息。

在统计学中，主成分分析（principal components analysis, PCA）是一种简化数据集的技术。它是一个线性变换。这个变换把数据变换到一个新的坐标系统中，使得任何数据投影的第一大方差在第一个坐标(称为第一主成分)上，第二大方差在第二个坐标(第二主成分)上，依次类推。主成分分析经常用减少数据集的维数，同时保持数据集的对方差贡献最大的特征。这是通过保留低阶主成分，忽略高阶主成分做到的。这样低阶成分往往能够保留住数据的最重要方面。但是，这也不是一定的，要视具体应用而定。

2.主成分分析的基本思想

在实证问题研究中，为了全面、系统地分析问题，我们必须考虑众多影响因素。这些涉及的因素一般称为指标，在多元统计分析中也称为变量。因为每个变量都在不同程度上反映了所研究问题的某些信息，并且指标之间彼此有一定的相关性，因而所得的统计数据反映的信息在一定程度上有重叠。在用统计方法研究多变量问题时，变量太多会增加计算量和增加分析问题的复杂性，人们希望在进行定量分析的过程中，涉及的变量较少，得到的信息量较多。主成分分析正是适应这一要求产生的，是解决这类题的理想工具。

对同一个体进行多项观察时必定涉及多个随机变量X_1，X_2，…，X_p，它们之间都存在着相关性，一时难以综合。这时就需要借助主成分分析来概括诸多信息的主要方面。我们希望有一个或几个较好的综合指标来概括信息，而且希望综合指标互相独立地各代表某一方面的性质。

任何一个度量指标的好坏除了可靠、真实之外，还必须能充分反映个体间的变异。如果有一项指标，不同个体的取值都大同小异，那么该指标不能用来区分不同的个体。由这一点来看，一项指标在个体间的变异越大越好。因此我们把“变异大”作为“好”的标准来寻求综合指标。

3.主成分分析法的基本原理

主成分分析法是一种降维的统计方法，它借助于一个正交变换，将其分量相关的原随机向量转化成其分量不相关的新随机向量，这在代数上表现为将原随机向量的协方差阵变换成对角形阵，在几何上表现为将原坐标系变换成新的正交坐标系，使之指向样本点散布最开的 p 个正交方向，然后对多维变量系统进行降维处理，使之能以一个较高的精度转换成低维变量系统，再通过构造适当的价值函数，进一步把低维系统转化成一维系统。

4.主成分的一般定义

设有随机变量X_1，X_2，…，X_p，样本标准差记为S_1，S_2，…，S_p。首先作标准化变换：

$$C_j=a_{j1}x_1+a_{j2}x_2+\ \dots\ +a_{jp}x_p\ ,\ j=1,2,\dots,p$$

我们有如下的定义：

(1) 若$C_1=a_{11}x_1+a_{12}x_2+\dots+a_{1p}x_p$，，且使 $Var(C_1)$最大，则称C_1为第一主成分；

(2) 若$C_2=a_{21}x_1+a_{22}x_2+\dots+a_{2p}x_p$，，$(a_{21}，a_{22}，\dots，a_{2p})$垂直于$(a_{11}，a_{12}，\dots，a_{1p})$，且使$Var(C_2)$最大，则称$C_2$为第二主成分；

(3) 类似地，可有第三、四、五…主成分，至多有 p 个。

5.主成分的性质

主成分C_1，C_2，…，C_p具有如下几个性质：

(1) 主成分间互不相关，即对任意i和j，C_i和C_j的相关系数

$$Corr(C_i,C_j)=0 \quad i \quad j$$

(2) 组合系数(a_{i1}，a_{i2}，…，a_{ip})构成的向量为单位向量，

(3) 各主成分的方差是依次递减的， 即 $Var(C_1)\geq Var(C_2)\geq\ldots\geq Var(C_p)$

(4) 总方差不增不减， 即

$$Var(C_1)+Var(C_2)+\ldots+Var(C_p)=Var(x_1)+Var(x_2)+\ldots+Var(x_p)=p$$

这一性质说明，主成分是原变量的线性组合，是对原变量信息的一种改组，主成分不增加总信息量，也不减少总信息量。

(5) 主成分和原变量的相关系数 $Corr(C_i，x_j)=a_{ij}=a_{ij}$

(6) 令X_1，X_2，…，X_p的相关矩阵为R，(a_{i1}，a_{i2}，…，a_{ip})则是相关矩阵R的第i个特征向量(eigenvector)。而且，特征值l_i就是第i主成分的方差， 即 $Var(C_i)=l_i$

其中l_i为相关矩阵R的第i个特征值(eigenvalue)

$$l_1\geq l_2\geq\ldots\geq l_p\geq 0$$

6.主成分数目的选取

前已指出，设有p个随机变量，便有p个主成分。由于总方差不增不减，C_1，C_2等前几个综合变量的方差较大,而C_p，C_{p-1}等后几个综合变量的方差较小, 严格说来，只有前几个综合变量才称得上主(要)成份，后几个综合变量实为“次 ”(要)成份。实践中总是保留前几个，忽略后几个。

保留多少个主成分取决于保留部分的累积方差在方差总和中所占百分比(即累计贡献率)，它标志着前几个主成分概括信息之多寡。实践中，粗略规定一个

百分比便可决定保留几个主成分；如果多留一个主成分，累积方差增加无几，便不再多留。

7.主成分分析的主要作用

概括起来说，主成分分析主要由以下几个方面的作用。

1．主成分分析能降低所研究的数据空间的维数。即用研究m维的Y空间代替p维的X空间(m＜p)，而低维的Y空间代替 高维的x空间所损失的信息很少。即：使只有一个主成分Y_l(即 m＝1)时，这个Y_l仍是使用全部X变量(p个)得到的。例如要计算Yl的均值也得使用全部x的均值。在所选的前m个主成分中，如果某个X_i的系数全部近似于零的话，就可以把这个X_i删除，这也是一种删除多余变量的方法。

2．有时可通过因子负荷a_{ij}的结论，弄清X变量间的某些关系。

3．多维数据的一种图形表示方法。我们知道当维数大于 3 时便不能画出几何图形，多元统计研究的问题大都多于 3 个变量。要把研究的问题用图形表示出来是不可能的。然而，经过主成分分析后，我们可以选取前两个主成分或其中某两个主成分，根据主成分的得分，画出 n 个样品在二维平面上的分布况，由图形可直观地看出各样品在主分量中的地位，进而还可以对样本进行分类处理，可以由图形发现远离大多数样本点的离群点。

4．由主成分分析法构造回归模型。即把各主成分作为新自变量代替原来自变量x做回归分析。

5．用主成分分析筛选回归变量。回归变量的选择有着重的实际意义，为了使模型本身易于做结构分析、控制和预报，好从原始变量所构成的子集合中选择最佳变量，构成最佳变量集合。用主成分分析筛选变量，可以用较少的计算量来选择量，获得选择最佳变量子集合的效果。

8.主成分分析法的计算步骤

1、原始指标数据的标准化采集p 维随机向量$x=(x_1,X_2,...,X_p)^T$)，n 个样品$x_i=(x_{i1},x_{i2},...,x_{ip})^T$ ，i=1,2,…,n，

n＞p，构造样本阵，对样本阵元进行如下标准化变换：

$$Z_{ij}=\frac{x_{ij}-\bar{x}_j}{s_j},i=1,2,...,n;j=1,2,...,p$$

其中$\bar{x}_j=\frac{\sum_{i=1}^{n}x_{ij}}{n},s_j^2=\frac{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}{n-1}$，得标准化阵 Z。

2、对标准化阵 Z 求相关系数矩阵

$$R=[r_{ij}]_p xp=\frac{Z^TZ}{n-1}$$

其中，$r_{ij}=\frac{\sum z_{kj}\cdot z_{kj}}{n-1},i,j=1,2,...,p$ 。

3、解样本相关矩阵 R 的特征方程$|R-\lambda I_p|=0$，得 p 个特征根,确定主成分

按$\frac{\sum_{j=1}^{m}\lambda_j}{\sum_{j=1}^{p}\lambda_j}\geq 0.85$确定m 值，使信息的利用率达 85%以上，对每个$\lambda_j$, j=1,2,...,m, 解方程组$Rb=\lambda_j b$得单位特征向量$b_j^o$ 。

4、将标准化后的指标变量转换为主成分

$$U_{ij}=z_i^Tb_j^o,j=1,2,...,m$$

U_1称为第一主成分,U_2 称为第二主成分,…,U_p 称为第p 主成分。

5 、对 m 个主成分进行综合评价

对 m 个主成分进行加权求和，即得最终评价值，权数为每个主成分的方差贡献率。

参考文献

1 李成，孙旭，程福臻，用主成分分析法研究星团谱线的等值高度，天文学报，第 43 卷第 2 期，2002 年 5 月

2 Principal components analysis, Wikipedia

3 主成分分析法，MBAlib

4 Principal Components and Factor Analysis, StatSoft