南京理工大学
NANJING UNIVERSITY OF SCIENCE & TECHNOLOGY

主成分分析

一、主成分分析的基本思想

主成分分析：将原来较多的指标简化为少数几个新的综合指标的多元统计方法。

- 主成分分析得到的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。

❖ 主成分分析的运用：

1、对一组内部相关的变量作简化的描述

2、用来削减回归分析或群集分析(Cluster)中变量的数目

3、用来检查异常点

4、用来作多重共线性鉴定

二、数学模型

- 假设我们所讨论的实际问题中，有p个指标，我们把这p个指标看作p个随机变量，记为X_1，$X_2, \ldots, X_p$，主成分分析就是要把这p个指标的问题，转变为讨论p个指标的线性组合的问题，而这些新的指标$F_1, F_2, \ldots, F_k (k \leqslant p)$，按照保留主要信息量的原则充分反映原指标的信息，并且相互独立。

- 这种由讨论多个指标降为少数几个综合指标的过程在数学上就叫做降维。主成分分析通常的做法是，寻求原指标的线性组合F_i。

$$F_1 = u_{11}X_1 + u_{21}X_2 + \cdots + u_{p1}X_p$$

$$F_2 = u_{12}X_1 + u_{22}X_2 + \cdots + u_{p2}X_p$$

$$\cdots\cdots$$

$$F_p = u_{1p}X_1 + u_{2p}X_2 + \cdots + u_{pp}X_p$$

❖ 满足如下的条件：

1、每个主成分的系数平方和为1。即

$$u_{1i}^2 + u_{2i}^2 + \cdots + u_{pi}^2 = 1$$

2、主成分之间相互独立，即无重叠的信息。即

$$Cov（F_i，\ F_j）= 0，\ i \neq j，\ i，\ j = 1，\ 2，\ \cdots，\ p$$

3、主成分的方差依次递减，重要性依次递减，即

$$Var（F_1）\geq Var(F_2) \geq \cdots \geq Var(F_p)$$

F_1,F_2,…,F_p分别称为原变量的第一、第二、…、第p个主成分。

- 问题的关键：

1、如何进行主成分分析？（主成分分析的方法）

基于相关系数矩阵还是基于协方差矩阵做主成分分析。当分析中所选择的变量具有不同的量纲，变量水平差异很大，应该选择基于相关系数矩阵的主成分分析。

2、如何确定主成分个数？

主成分分析的目的是简化变量，一般情况下主成分的个数应该小于原始变量的个数。

❖ 主成分分析的目标：

1、从相关的$X_1, X_2, \ldots, X_k$，求出相互独立的新综合变量（主成分）$Y_1, Y_2, \ldots, Y_k$。

2、$Y=(Y_1, Y_2, \cdots, Y_k)'$所反映信息的含量无遗漏或损失的指标—方差，等于$X=(X_1, X_2, \cdots, X_k)'$的方差 。

X与Y之间的计算关系是：

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_k \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \cdots & a_{kk} \end{bmatrix} \begin{bmatrix} X_1 \\ \vdots \\ X_k \end{bmatrix} \quad \text{即} Y=AX$$

- 从协方差矩阵出发求解主成分的步骤：

1、求解各观测变量 $X_l = (x_{1l}, x_{2l}, \cdots, x_{pl})' (l = 1,2,\cdots, n)$ 的协方差矩阵。

2、由X的协方差阵Σ，求出其特征根，即解方程 $\Sigma u_i = \lambda_i u_i$，可得特征根 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$。

3、求解 $|\Sigma - \lambda I| = 0$ 可得各特征根对应的特征向量U_1，U_2，…，U_p 。

其中最大特征根的特征向量对应第一主成分的系数向量；第二大特征根对应的特征向量是第二大主成分的系数向量…… $F_i = U_i' X, i = 1, \cdots, k(k \leq p)$

4、计算累积贡献率，给出恰当的主成分个数。

5、计算所选出的k个主成分的得分。将原始数据的中心化值：

$$\mathbf{X}_i^* = \mathbf{X}_i - \overline{\mathbf{X}} = \left(x_{1i} - \overline{x}_1,\ x_{2i} - \overline{x}_2, \cdots,\ x_{pi} - \overline{x}_p\right)'$$

代入前k个主成分的表达式，分别计算出各样本k个主成分的得分。

6、对结果进行正确分析和合理解释.

- 从相关系数矩阵出发求解主成分的步骤：

1、标准化各观测变量数据。

2、求解标准化各观测变量的相关系数矩阵。

2、根据矩阵知识 $|\rho - \lambda I| = 0$求解相关系数矩阵的特征根。

3、求解各特征根对应的特征向量。$\rho u_i = \lambda_i u_i$

其中最大特征根的特征向量对应第一主成分的系数向量；第二大特征根对应的特征向量是第二大主成分的系数向量……

❖ 三、主成分性质

1、主成分的协方差阵为对角阵

2、p个随机变量的总方差为协方差矩阵Σ的所有特征根之和

$$\sum_{i=1}^{p} Var(F_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp}$$

说明主成分分析把p个随机变量的总方差分解成为p个不相关的随机变量的方差之和。

3、贡献率：

第i个主成分的方差在全部方差中所占比重 $\lambda_i \Big/ \sum_{i=1}^{p} \lambda_i$

称为贡献率，反映了原来p个指标多大的信息，有多大的综合能力。

4、累积贡献率：

前k个主成分共有多大的综合能力，用这k个主成分的方差和在全部方差中所占比重

来描述，称为累积贡献率。 $\sum_{i=1}^{k} \lambda_i \Big/ \sum_{i=1}^{p} \lambda_i$

□ 一、主成分个数的选取

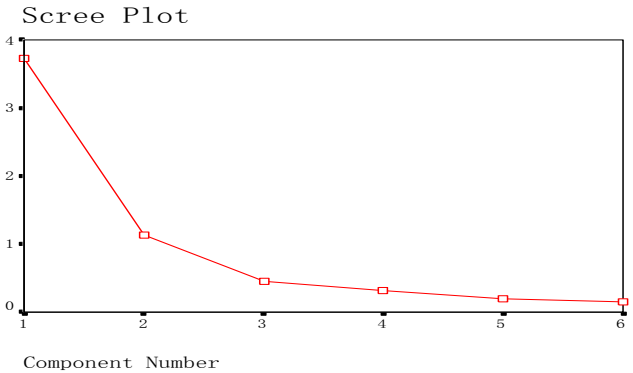
1. 累积贡献率达到85%以上

2. 根据特征根的变化来确定

数据标准化情况下：

3. 作碎石图

描述特征值的贡献



Matlab里的主成分分析函数

1.princomp

功能：主成分分析

格式：PC=princomp(X)

[PC,SCORE,latent,tsquare]=princomp(X)

说明：[PC,SCORE,latent,tsquare]=princomp(X)对数据矩阵X进行主成分分析，给出各主成分(PC)、所谓的Z-得分(SCORE)、X的方差矩阵的特征值(latent)和每个数据点的HotellingT^{2}统计量(tsquare)。

2.pcacov

功能：运用协方差矩阵进行主成分分析

格式：**PC=pcacov(X)**

[PC,latent,explained]=pcacov(X)

说明：**[PC,latent,explained]=pcacov(X)**通过协方差矩阵**X**进行主成分分析，返回主成分**(PC)**、协方差矩阵**X**的特征值**(latent)**和每个特征向量表征在观测量总方差中所占的百分数**(explained)**。

3.pcares

功能：主成分分析的残差

格式：**residuals=pcares(X,ndim)**

说明：**pcares(X,ndim)**返回保留**X**的**ndim**个主成分所获的残差。注意，**ndim**是一个标量，必须小于**X**的列数。而且，**X**是数据矩阵，而不是协方差矩阵。

4.barttest

功能：主成分的巴特力特检验

格式：**ndim=barttest(X,alpha)**

[ndim,prob,chisquare]=barttest(X,alpha)

说明：巴特力特检验是一种等方差性检验。

ndim=barttest(X,alpha)是在显著性水平**alpha**下，给出满足数据矩阵**X**的非随机变量的**n**维模型，**ndim**即模型维数，它由一系列假设检验所确定，**ndim=1**表明数据**X**对应于每个主成分的方差是相同的；**ndim=2**表明数据**X**对应于第二成分及其余成分的方差是相同的。