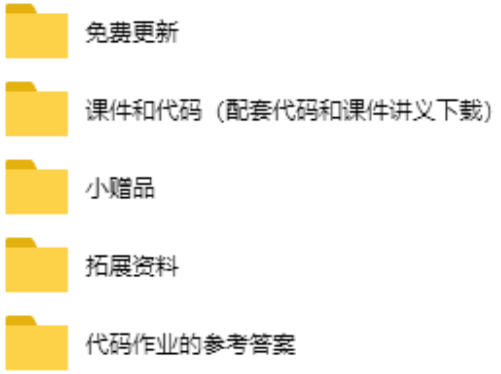
温馨提示

（1）视频中提到的附件可在**售后群的群文件**中下载。

包括**讲义、代码、优秀的作业、我视频中推荐的资料**等。



（2）关注我的**微信公众号《数学建模学习交流》**，后台发送**“软件”**两个字，可获得常见的建模软件下载方法；发送**“数据”**两个字，可获得建模数据的获取方法；发送**“画图”**两个字，可获得数学建模中常见的画图方法。另外，也可以看看公众号的历史文章，里面发布的都是对大家有帮助的技巧。

（3）**购买更多优质精选的数学建模资料**，可关注我的微信公众号《数学建模学习交流》，在后台发送**“买”**这个字即可进入店铺进行购买。

（4）视频价格不贵，但价值很高。单人购买观看只需要**58元**，和另外两名队友一起购买人均仅需**46元**，视频本身也是下载到本地观看的，所以请大家**不要侵犯知识产权**，对视频或者资料进行二次销售。

数学建模学习交流

第十四讲:主成分分析

本讲将介绍主成分分析(Principal Component Analysis,PCA)，主成分分析是一种降维算法，它能将多个指标转换为少数几个主成分，这些主成分是原始变量的线性组合，且彼此之间互不相关，其能反映出原始数据的大部分信息。一般来说，当研究的问题涉及到多变量且变量之间存在很强的相关性时，我们可考虑使用主成分分析的方法来对数据进行简化。

问题的提出

在实际问题研究中，多变量问题是经常会遇到的。变量太多，无疑会增加分析问题的难度与复杂性，而且在许多实际问题中，多个变量之间是具有一定的相关关系的。

因此，人们会很自然地想到，能否在相关分析的基础上，**用较少的新变量代替原来较多的旧变量，而且使这些较少的新变量尽可能多地保留原来变量所反映的信息？**

事实上，这种想法是可以实现的，主成分分析方法就是综合处理这种问题的一种强有力的工具。

主成分分析是把原来多个变量划为少数几个综合指标的一种统计分析方法。

从数学角度来看，这是一种**降维**处理技术。

数据降维的作用

降维是将高维度的数据（指标太多）保留下最重要的一些特征，去除噪声和不重要的特征，从而实现提升数据处理速度的目的。

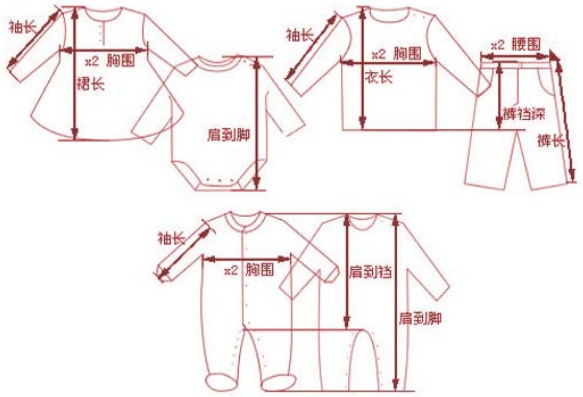
在实际的生产和应用中，降维在一定的信息损失范围内，可以为我们节省大量的时间和成本。降维也成为应用非常广泛的数据预处理方法。

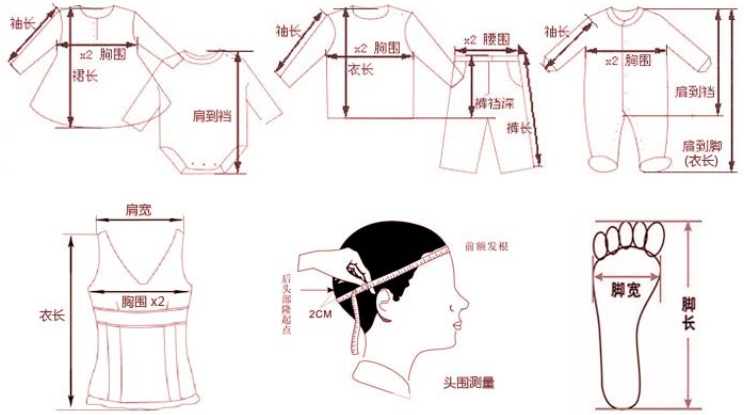
降维具有如下一些优点：

- 使得数据集更易使用；
- 降低算法的计算开销；
- 去除噪声；
- 使得结果容易理解。

一个简单的例子

例如，某人要做一件上衣要测量很多尺寸，如身长、袖长、胸围、腰围、肩宽、肩厚等十几项指标，但某服装厂要生产一批新型服装绝不可能把尺寸的型号分得过多？而是从多种指标中综合成几个少数的综合指标，做为分类的型号，利用主成分分析将十几项指标综合成3项指标，一项是反映长度的指标，一项是反映胖瘦的指标，一项是反映特殊体型的指标。





主成分分析的思想

假设有n个样本，p个指标，则可构成大小为$n \times p$的样本矩阵x:

$$x = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = (x_1, x_2, \cdots, x_p)$$

假设我们想找到新的一组变量$z_1, z_2, \cdots, z_m (m \leq p)$，且它们满足:

$$\begin{cases} z_1 = l_{11}x_1 + l_{12}x_2 + \cdots + l_{1p}x_p \\ z_2 = l_{21}x_1 + l_{22}x_2 + \cdots + l_{2p}x_p \\ \quad \vdots \\ z_m = l_{m1}x_1 + l_{m2}x_2 + \cdots + l_{mp}x_p \end{cases}$$

主成分分析的思想

$$\begin{cases} z_1 = l_{11}x_1 + l_{12}x_2 + \cdots + l_{1p}x_p \\ z_2 = l_{21}x_1 + l_{22}x_2 + \cdots + l_{2p}x_p \\ \quad \vdots \\ z_m = l_{m1}x_1 + l_{m2}x_2 + \cdots + l_{mp}x_p \end{cases}$$

系数l_{ij}的确定原则：

(1)z_i与$z_j(i \neq j;\ i,j = 1,2,\cdots,m)$相互无关；

(2)z_1是$x_1,x_2,\cdots,x_p$的一切线性组合中方差最大者；

(2)z_2是与z_1不相关的$x_1,x_2,\cdots,x_p$的所有线性组合中方差最大者；

(3)依次类推，z_m是与z_1，$z_2,\cdots,z_{m-1}$不相关的$x_1,x_2,\cdots,x_p$的所有线性组合中方差最大者。

(4)新变量指标$z_1,z_2,\cdots,z_m$分别称为原变量指标$x_1,x_2,\cdots,x_p$的第一，第二，…，第m主成分。

(是不是看到了典型相关分析的影子)

严谨的数学符号

设对某一事物的研究涉及 p 个指标，分别用 X_1，X_2，…，X_p 表示，这 p 个指标构成的 p 维随机向量为 $\boldsymbol{X}=(X_1, X_2, \cdots, X_p)'$。设随机向量 $\boldsymbol{X}$ 的均值为 $\boldsymbol{\mu}$，协方差矩阵为 $\boldsymbol{\Sigma}$。

对 $\boldsymbol{X}$ 进行线性变换，可以形成新的综合变量，用 $\boldsymbol{Y}$ 表示，也就是说，新的综合变量可以由原来的变量线性表示，即满足下式：

$$\begin{cases} Y_1 = u_{11}X_1 + u_{12}X_2 + \cdots + u_{1p}X_p \\ Y_2 = u_{21}X_1 + u_{22}X_2 + \cdots + u_{2p}X_p \\ \cdots\cdots\cdots\cdots \\ Y_p = u_{p1}X_1 + u_{p2}X_2 + \cdots + u_{pp}X_p \end{cases} \tag{5.1}$$

由于可以任意地对原始变量进行上述线性变换，由不同的线性变换得到的综合变量 $\boldsymbol{Y}$ 的统计特性也不尽相同。因此为了取得较好的效果，我们总是希望 $Y_i = \boldsymbol{u}_i'\boldsymbol{X}$ 的方差尽可能大且各 Y_i 之间互相独立，由于

$$\mathrm{var}(Y_i) = \mathrm{var}(\boldsymbol{u}_i'\boldsymbol{X}) = \boldsymbol{u}_i'\boldsymbol{\Sigma}\boldsymbol{u}_i$$

而对任给的常数 c，有

$$\mathrm{var}(c\boldsymbol{u}_i'\boldsymbol{X}) = c\boldsymbol{u}_i'\boldsymbol{\Sigma}\boldsymbol{u}_i c = c^2\boldsymbol{u}_i'\boldsymbol{\Sigma}\boldsymbol{u}_i$$

严谨的数学符号

因此对 u_i 不加限制时，可使 var(Y_i) 任意增大，问题将变得没有意义。我们将线性变换约束在下面的原则之下：

1. $u_i'u_i=1$，即 $u_{i1}^2+u_{i2}^2+\cdots+u_{ip}^2=1$ $(i=1, 2, \cdots, p)$；
2. Y_i 与 Y_j 相互无关 $(i\neq j;\ i, j=1, 2, \cdots, p)$；
3. Y_1 是 $X_1, X_2, \cdots, X_p$ 的一切满足原则 1 的线性组合中方差最大者；Y_2 是与 Y_1 不相关的 $X_1, X_2, \cdots, X_p$ 所有线性组合中方差最大者；$\cdots$，Y_p 是与 $Y_1, Y_2, \cdots, Y_{p-1}$ 都不相关的 $X_1, X_2, \cdots, X_p$ 的所有线性组合中方差最大者。

基于以上三条原则决定的综合变量 $Y_1, Y_2, \cdots, Y_p$ 分别称为原始变量的第一、第二、…、第 p 个主成分。其中，各综合变量在总方差中占的比重依次递减，在实际研究工作中，通常只挑选前几个方差最大的主成分，从而达到简化系统结构、抓住问题实质的目的。

参考教材：《应用多元统计分析》王学民

PCA详细的证明过程可看视频：https://www.bilibili.com/video/av32709936

（证明过程需要一定的多元统计基础和较强的线性代数基础）

PCA的计算步骤

假设有n个样本，p个指标，则可构成大小为$n \times p$的样本矩阵x:

$$x = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = (x_1, x_2, \cdots, x_p)$$

1.我们首先对其进行标准化处理：

按列计算均值$\overline{x_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$和标准差$S_j = \sqrt{\frac{\sum_{i=1}^{n}\left(x_{ij} - \overline{x_j}\right)^2}{n-1}}$，计算得

标准化数据$X_{ij} = \frac{x_{ij} - \overline{x_j}}{S_j}$，原始样本矩阵经过标准化变为：

$$X = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} = (X_1, X_2, \cdots, X_p)$$

PCA的计算步骤

原始样本矩阵经过标准化变为：

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

2.计算标准化样本的协方差矩阵

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

$$\text{其中} r_{ij}=\frac{1}{n-1}\sum_{k=1}^{n}(X_{ki}-\overline{X_i})(X_{kj}-\overline{X_j})=\frac{1}{n-1}\sum_{k=1}^{n}X_{ki}X_{kj}$$

(注意：上面12两步可直接合并为一步：直接计算x矩阵的样本相关系数矩阵)

$$R=\frac{\sum_{k=1}^{n}(x_{ki}-\overline{x_i})(x_{kj}-\overline{x_j})}{\sqrt{\sum_{k=1}^{n}(x_{ki}-\overline{x_i})^2\sum_{k=1}^{n}(x_{kj}-\overline{x_j})^2}}$$

PCA的计算步骤

3.计算R的特征值和特征向量

特征值：$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ （R是半正定矩阵，且$tr(R)=\sum_{k=1}^{p}\lambda_k = p$）

特征向量：$a_1 = \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{p1} \end{bmatrix}$, $a_2 = \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{p2} \end{bmatrix}$, $\cdots$, $a_p = \begin{bmatrix} a_{1p} \\ a_{2p} \\ \vdots \\ a_{pp} \end{bmatrix}$

（软件都会帮我们算好，不用自己算，$Matlab$中计算特征值和特征向量的函数：$eig(R)$）

4.计算主成分贡献率以及累计贡献率

贡献率 $= \dfrac{\lambda_i}{\sum_{k=1}^{p}\lambda_k} \ (i=1,2,\cdots,p)$　　累计贡献率 $= \dfrac{\sum_{k=1}^{i}\lambda_k}{\sum_{k=1}^{p}\lambda_k} \ (i=1,2,\cdots,p)$

PCA的计算步骤

$$\text{贡献率} = \frac{\lambda_i}{\sum_{k=1}^{p}\lambda_k}\ (i=1,2,\cdots,p) \qquad \text{累计贡献率} = \frac{\sum_{k=1}^{i}\lambda_k}{\sum_{k=1}^{p}\lambda_k}\ (i=1,2,\cdots,p)$$

5.写出主成分

一般取累计贡献率超过80%的特征值所对应的第一、第二、...、第$m(m \leqslant p)$个主成分。

第i个主成分: $F_i = a_{1i}X_1 + a_{2i}X_2 + \cdots + a_{pi}X_p \quad (i=1,2,\cdots,m)$

6.根据系数分析主成分代表的意义

对于某个主成分而言，指标前面的系数越大，代表该指标对于该主成分的影响越大。

7.利用主成分的结果进行后续的分析

（1）主成分得分：可用于评价类模型吗？（千万别用!!!）

（2）主成分可用于聚类分析（方便画图）

（3）主成分可用于回归分析

教材例题1讲解

在制定服装标准的过程中，对128名成年男子的身材进行了测量，每人测得的指标中含有这样六项：身高（x1）、坐高（x2）、胸围（x3）、手臂长（x4）、肋围（x5）和腰围（x6）。所得样本相关系数矩阵（对称矩阵哦）列于下表。

	x_1	x_2	x_3	x_4	x_5	x_6
x_1	1.000	0.79	0.36	0.76	0.25	0.51
x_2	0.79	1.000	0.31	0.55	0.17	0.35
x_3	0.36	0.31	1.000	0.35	0.64	0.58
x_4	0.76	0.55	0.35	1.000	0.16	0.38
x_5	0.25	0.17	0.64	0.16	1.000	0.63
x_6	0.51	0.35	0.58	0.38	0.63	1.000

注意：本题给我们的数据直接就是样本相关系数矩阵，一般来说，大家自己建模的时候，得到的数据是最原始的数据（每一列是指标，每一行是样本）。

参考教材：《应用多元统计分析》王学民

计算关键变量

经过计算，相关系数矩阵的特征值、相应的特征向量以及贡献率列于下表：

特征向量	a1	a2	a3	a4	a5	a6
x1：身高	0.469	−0.365	0.092	-0.122	-0.080	-0.786
x2：坐高	0.404	−0.397	0.613	0.326	0.027	0.443
x3：胸围	0.394	0.397	−0.279	0.656	0.405	-0.125
x4：手臂长	0.408	−0.365	−0.705	-0.108	-0.235	0.371
x5：肋围	0.337	0.569	0.164	-0.019	-0.731	0.034
x6：腰围	0.427	0.308	0.119	-0.661	0.490	0.179
特征值	3.287	1.406	0.459	0.426	0.295	0.126
贡献率	0.548	0.234	0.077	0.071	0.049	0.021
累计贡献率	0.548	0.782	0.859	0.930	0.979	1.000

从表中可以看到前三个主成分的累计贡献率达85.9%，因此可以考虑只取前面三个主成分，它们能够很好地概括原始变量。

写出主成分并简要分析

$$F_1 = 0.469X_1 + 0.404X_2 + 0.394X_3 + 0.408X_4 + 0.337X_5 + 0.427X_6$$
$$F_2 = -0.365X_1 - 0.397X_2 + 0.397X_3 - 0.365X_4 + 0.569X_5 + 0.308X_6$$
$$F_3 = 0.092X_1 + 0.613X_2 - 0.279X_3 - 0.705X_4 + 0.164X_5 + 0.119X_6$$

X_i均是标准化后的指标，x_i：身高、坐高、胸围、手臂长、肋围和腰围

第一主成分F_1对所有（标准化）原始变量都有近似相等的正载荷，故称第一主成分为（身材）大小成分。

第二主成分F_2在X_3, X_5, X_6上有中等程度的正载荷，而在X_1, X_2, X_4上有中等程度的负载荷，称第二主成分为形状成分（或胖瘦成分）。

第三主成分F_3在X_2上有大的正载荷，在X_4上有大的负载荷，而在其余变量上的载荷都较小，可称第三主成分为臂长成分。

注：由于第三主成分的贡献率不高（7.65％）且实际意义也不太重要，因此我们也可以考虑只取前两个主成分进行分析。

主成分分析的说明

在主成分分析中，我们首先应保证所提取的前几个主成分的累计贡献率达到一个较高的水平，其次对这些被提取的主成分必须都能够给出符合实际背景和意义的解释。

主成分的解释其含义一般多少带有点模糊性，不像原始变量的含义那么清楚、确切，这是变量降维过程中不得不付出的代价。因此，提取的主成分个数m通常应明显小于原始变量个数p（除非p本身较小），否则维数降低的“利”可能抵不过主成分含义不如原始变量清楚的“弊”。

如果原始变量之间具有较高的相关性，则前面少数几个主成分的累计贡献率通常就能达到一个较高水平，也就是说，此时的累计贡献率通常较易得到满足。

主成分分析的困难之处主要在于要能够给出主成分的较好解释，所提取的主成分中如有一个主成分解释不了，整个主成分分析也就失败了。

主成分分析是变量降维的一种重要、常用的方法，简单的说，该方法要应用得成功，一是靠原始变量的合理选取，二是靠“运气”。

——参考教材：《应用多元统计分析》王学民

教材例题2讲解

在1984年洛杉矶奥运会田径统计手册中，有55个国家和地区的如下八项男子径赛运动记录：
x1: 100米（单位：秒）　　x5: 1500米（单位：分）
x2: 200米（单位：秒）　　x6: 5000米（单位：分）
x3: 400米（单位：秒）　　x7: 10000米（单位：分）
x4: 800米（单位：秒）　　x8: 马拉松（单位：分）
其样本相关系数矩阵如下表所示：

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
x_1	1.000							
x_2	0.923	1.000						
x_3	0.841	0.851	1.000					
x_4	0.756	0.807	0.870	1.000				
x_5	0.700	0.775	0.835	0.918	1.000			
x_6	0.619	0.695	0.779	0.864	0.928	1.000		
x_7	0.633	0.697	0.787	0.869	0.935	0.975	1.000	
x_8	0.520	0.596	0.705	0.806	0.866	0.932	0.943	1.000

计算关键变量

特征向量	a1	a2	…
x1：100米	0.318	0.567	…
x2：200米	0.337	0.462	…
x3：400米	0.356	0.248	…
x4：800米	0.369	0.012	…
x5：1500米	0.373	−0.140	…
x6：5000米	0.364	−0.312	…
x7：10000米	0.367	−0.307	…
x8：马拉松	0.342	−0.439	…
特征值	6.622	0.878	…
贡献率	0.828	0.110	…
累计贡献率	0.828	0.937	…

对主成分的简要分析

特征向量	a1	a2
x1：100米	0.318	0.567
x2：200米	0.337	0.462
x3：400米	0.356	0.248
x4：800米	0.369	0.012
x5：1500米	0.373	−0.140
x6：5000米	0.364	−0.312
x7：10000米	0.367	−0.307
x8：马拉松	0.342	−0.439

由上表可知，前两个主成分的累计贡献率已高达93.7%，第一主成分F1在所有变量上有几乎相等的正载荷，可称为在径赛项目上的强弱成分。第二主成分F2在各个指标上的载荷基本上逐个递减，反映了速度与耐力成绩的对比 。

Matlab代码

PCA.m

```
%% 第一步：对数据x标准化为X
X=zscore(x);
%% 第二步：计算样本协方差矩阵
R = cov(X);
%% 注意：以上两步可合并为下面一步：直接计算样本相关系数矩阵
R = corrcoef(x);
%% 第三步：计算R的特征值和特征向量
[V,D] = eig(R);  % V 特征向量矩阵  D 特征值构成的对角矩阵
%% 第四步：计算主成分贡献率和累计贡献率
lambda = diag(D);  % diag函数用于得到一个矩阵的主对角线元素值(返回的是列向量)
lambda = lambda(end:-1:1);  % 因为lambda向量是从小大到排序的，我们将其调个头
% 计算贡献率
contribution_rate = lambda / sum(lambda);
% 计算累计贡献率  cumsum是求累加值的函数
cum_contribution_rate = cumsum(lambda)/ sum(lambda);
% 注意：这里的特征向量要和特征值一一对应，之前特征值相当于颠倒过来了，因此特征向量的各列需要颠倒过来
%  rot90函数可以使一个矩阵逆时针旋转90度，然后再转置，就可以实现将矩阵的列颠倒的效果
V=rot90(V)';
```

注：更加详细的代码注释可见源代码

利用Matlab进行主成分分析

特征向量	a1	a2	a3	…
x1：食品	0.401	−0.077	0.415	…
x2：衣着	0.132	0.749	0.332	…
x3：家庭设备	0.375	0.065	−0.442	…
x4：医疗	0.320	0.345	−0.478	…
x5：交通	0.388	−0.232	0.279	…
x6：娱乐	0.406	0.027	−0.310	…
x7：居住	0.326	−0.496	−0.034	…
x8：杂项	0.396	0.096	0.345	…
特征值	5.098	1.352	0.574	…
贡献率	0.637	0.169	0.072	…
累计贡献率	0.637	0.806	0.878	…

对结果的解释

特征向量	a1	a2	a3
x1：食品	0.401	−0.077	0.415
x2：衣着	0.132	0.749	0.332
x3：家庭设备	0.375	0.065	−0.442
x4：医疗	0.320	0.345	−0.478
x5：交通	0.388	−0.232	0.279
x6：娱乐	0.406	0.027	−0.310
x7：居住	0.326	−0.496	−0.034
x8：杂项	0.396	0.096	0.345
特征值	5.098	1.352	0.574
贡献率	0.637	0.169	0.072
累计贡献率	0.637	0.806	0.878

从上表可以看出，前两个和前三个主成分的累计贡献率分别达到80.6%和87.8%，第一主成分F1在所有变量(除在x2上的载荷稍偏小外)上都有近似相等的正载荷，反映了综合消费性支出的水平，因此第一主成分可称为综合消费性支出成分。第二主成分F2在变量x2上有很高的正载荷，在变量x4上有中等的正载荷，而在其余变量上有负载荷或很小的正载荷。可以认为这个主成分度量了受地区气候影响的消费性支出(主要是衣着，其次是医疗保健)在所有消费性支出中占的比重(也可理解为一种消费倾向)，第二主成分可称为消费倾向成分。第三主成分很难给出明显的解释，因此我们只取前面两个主成分。

注：解释什么的太难了！上面这段话我是写不出来的。

主成分分析的滥用：主成分得分

对主成分分析中综合得分方法的质疑

王学民

（发表于《统计与决策》，2007 年 4 月下）

摘要：在作主成分分析时，国内近年来流行一种通过建立综合评价函数来对各样品进行综合排名的方法。本文对这一方法的不科学性作了阐述，并指出在综合评价函数中对各主成分使用贡献率加权是错中加错。

关键词：主成分；信息量；综合评价函数；综合得分

除了王老师说的几点原因之外，我的补充：

（1）主成分是降维算法，你既然已经有数据了，为什么不把这些数据的信息全部用上呢？主成分分析是会损失原始数据的信息的。

（2）指标可能有各种类型（极大、极小、区间等），主成分只有标准化的过程，并没有正向化的过程呀。

主成分分析的滥用：主成分得分

求指标对应的系数

$$F_1 = 0.353ZX_1 + 0.042ZX_2 - 0.041ZX_3 + 0.364ZX_4 + 0.367ZX_5 + 0.366ZX_6 + 0.352ZX_7 + 0.364ZX_8 + 0.298ZX_9 + 0.355ZX_{10}$$

$$F_2 = 0.175ZX_1 - 0.741ZX_2 + 0.609ZX_3 - 0.004ZX_4 + 0.063ZX_5 - 0.061ZX_6 - 0.022ZX_7 + 0.158ZX_8 - 0.046ZX_9 - 0.115ZX_{10}$$

$$F = (72.2/84.5)\,F_1 + (12.3/84.5)\,F_2$$

城市	F1	F2	F	排名
广东	5.224739	0.114592	4.478657	1
江苏	2.254315	0.234636	1.959442	2
山东	1.962522	0.500242	1.749029	3
浙江	1.160716	-0.19308	0.963062	4
上海	0.296827	-2.35794	-0.09077	5
辽宁	-1.24298	1.960091	-0.77534	6
河北	-1.35286	0.408853	-1.09565	7
福建	-1.97451	-0.06651	-1.69594	8
天津	-3.04194	-1.00948	-2.7452	9
广西	-3.28683	0.408604	-2.7473	10

小石老师的视频中就对主成分分析进行了滥用。

主成分分析用于聚类

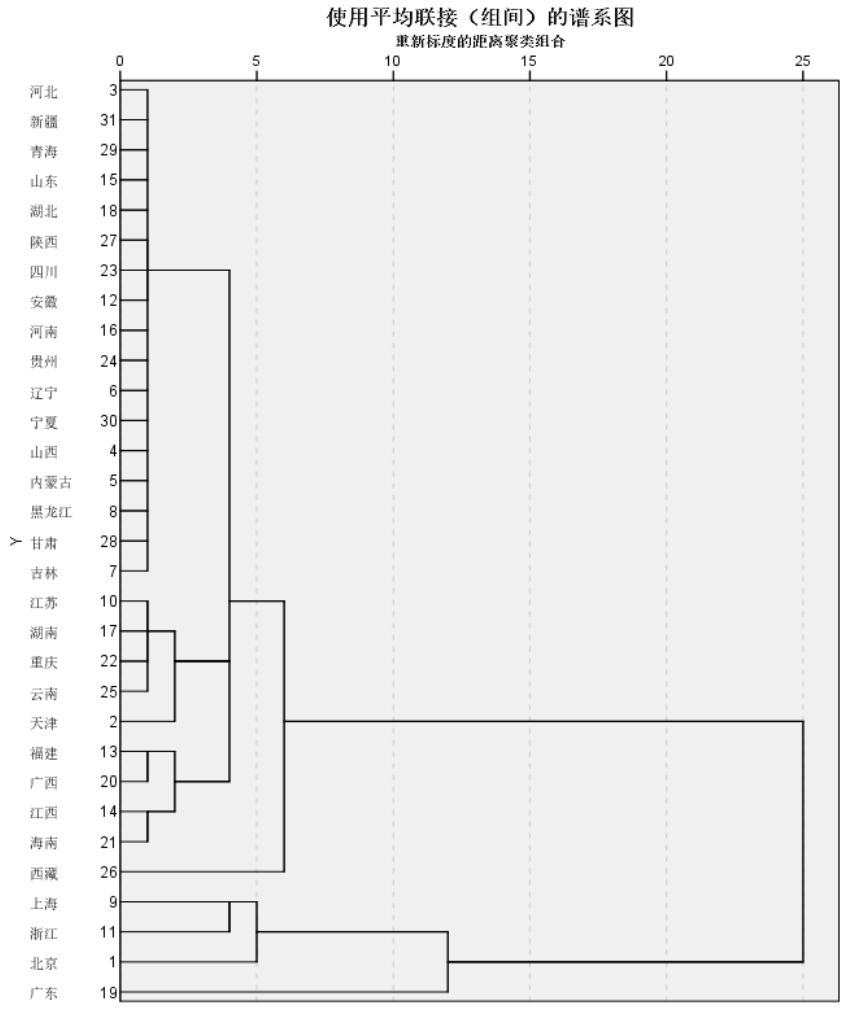
计算出第一主成分和第二主成分的值，将其视为两个新的指标（可以在图上直观的展示各样本的分布情况）

PCA.m

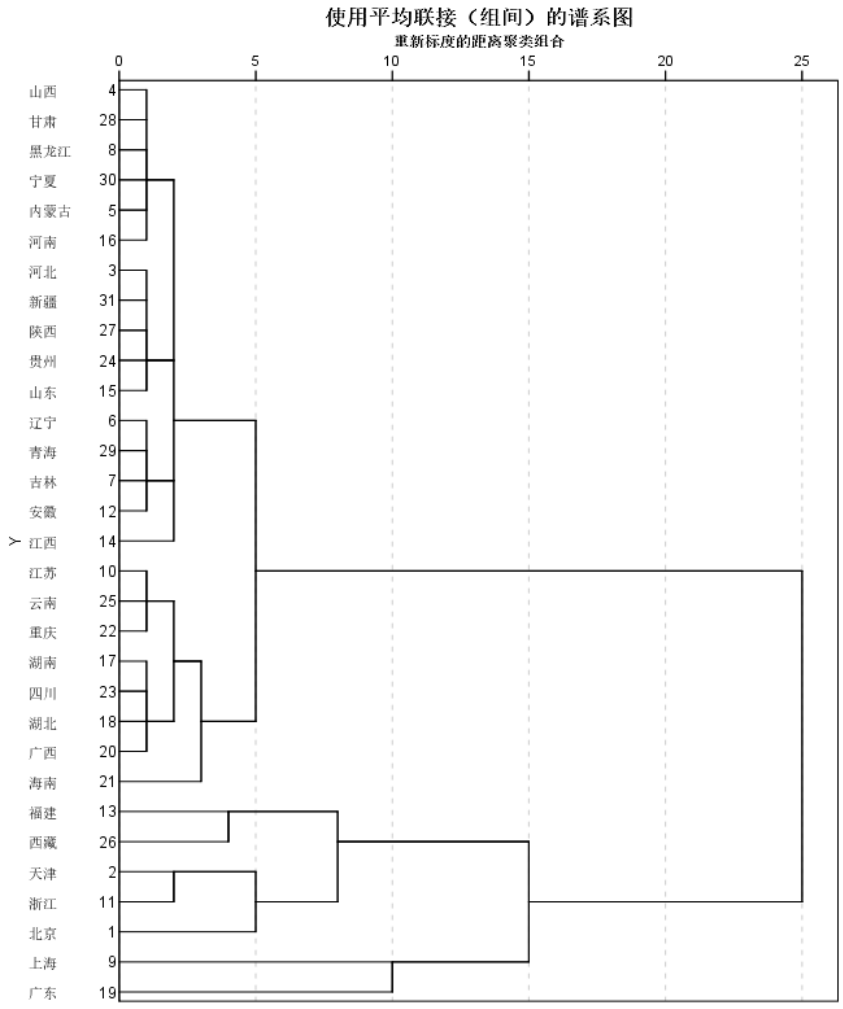
```
%% 计算我们所需要的主成分的值
m =input('请输入需要保存的主成分的个数: ');
F = zeros(n,m); %初始化保存主成分的矩阵（每一列是一个主成分）
for i = 1:m
    ai = V(:,i)';  % 将第i个特征向量取出，并转置为行向量
    Ai = repmat(ai,n,1);  % 将这个行向量重复n次，构成一个n*p的矩阵
    F(:, i) = sum(Ai .* X, 2); % 注意，对标准化的数据求了权重后要计算每一行的和
end

%%主成分聚类： 将主成分指标所在的F矩阵复制到Excel表格，然后再用Spss聚类
% 在Excel第一行输入指标名称（F1,F2, ..., Fm）
% 双击Matlab工作区的F,进入变量编辑中，然后复制里面的数据到Excel表格
% 导出数据之后，我们后续的分析就可以在Spss中进行。
```

Spss聚类

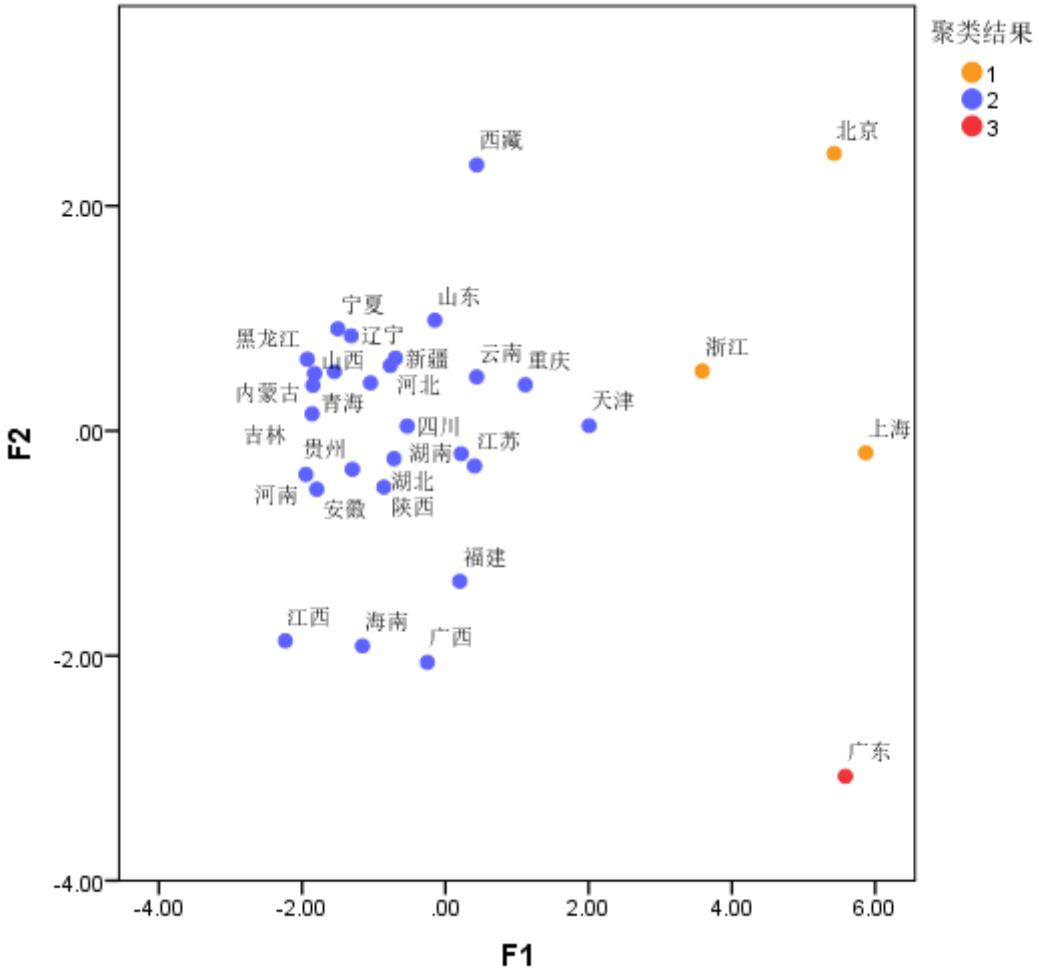


使用主成分的聚类结果



第十讲的聚类结果

聚类效果图



个人觉得，主成分聚类最大的意义就是能帮我们可视化最后的聚类效果，毕竟，使用主成分是会降低部分信息的。言外之意，只有在指标个数特别多，且指标之间存在很强的相关性时才用主成分聚类。

主成分回归

如果数据矩阵 $\boldsymbol{X}$ 不满列秩(列秩小于 K)，即某一解释变量可以由其他解释变量线性表出，则存在“严格多重共线性”。此时，$(\boldsymbol{X}'\boldsymbol{X})^{-1}$不存在，总体参数 $\boldsymbol{\beta}$ 不可识别，无法定义最小二乘估计量。严格多重共线性在现实数据中很少出现，即使出现，Stata 也会自动识别并删去多余的解释变量。

较常见的是近似(非严格)的多重共线性。其表现为，如果将第 k 个解释变量 x_k 对其余的解释变量 $\{x_1,\cdots,x_{k-1},x_{k+1},\cdots,x_K\}$ 进行回归，所得到的可决系数(记为 R_k^2)较高①。在存在近似多重共线性的情况下，OLS 仍然是最佳线性无偏估计，即在所有线性无偏估计中仍具有最小的方差。但这并不意味着 OLS 估计量方差在绝对意义上小。由于存在多重共线性，矩阵$(\boldsymbol{X}'\boldsymbol{X})$变得几乎不可逆，故从某种意义上来说，$(\boldsymbol{X}'\boldsymbol{X})^{-1}$变得很“大”，致使方差 $\mathrm{Var}(\boldsymbol{b}\mid\boldsymbol{X})=\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$增大，使得对系数的估计变得不准确。在这种情况下，只要数据矩阵 $\boldsymbol{X}$ 中的元素轻微变化，就可能引起$(\boldsymbol{X}'\boldsymbol{X})^{-1}$很大的变化，进而导致 OLS 估计值 $\boldsymbol{b}$ 发生很大变化。通常的“症状”是，虽然整个回归方程的 R^2 较大、F 检验也很显著，但单个系数的 t 检验却不显著，或者系数估计值不合理，甚至符号与理论预期相反。另一可能“症状”是，增减解释变量使得系数估计值发生较大变化(比如，最后加入的解释变量与已有解释变量构成多重共线性)。直观来看，如果两个(或多个)解释变量之间高度相关，则不容易区分它们各自对被解释变量的单独影响力。在极端情况下，一个变量刚好是另一变量的倍数，则完全无法区分。

主成分回归可用来解决多重共线性的问题。

例子：第七讲的课后作业

探究棉花单产和五个指标之间的关系。

年份	单产	种子费	化肥费	农药费	机械费	灌溉费
1990	1017	106.05	495.15	305.1	45.9	56.1
1991	1036.5	113.55	561.45	343.8	68.55	93.3
1992	792	104.55	584.85	414	73.2	104.55
1993	861	132.75	658.35	453.75	82.95	107.55
1994	901.5	174.3	904.05	625.05	114	152.1
1995	922.5	230.4	1248.75	834.45	143.85	176.4
1996	916.5	238.2	1361.55	720.75	165.15	194.25
1997	976.5	260.1	1337.4	727.65	201.9	291.75
1998	1024.5	270.6	1195.8	775.5	220.5	271.35
1999	1003.5	286.2	1171.8	610.95	195	284.55
2000	1069.5	282.9	1151.55	599.85	190.65	277.35
2001	1168.5	317.85	1105.8	553.8	211.05	290.1
2002	1228.5	319.65	1213.05	513.75	231.6	324.15
2003	1023	368.4	1274.1	567.45	239.85	331.8
2004	1144.5	466.2	1527.9	487.35	408	336.15
2005	1122	449.85	1703.25	555.15	402.3	358.8
2006	1276.5	537	1888.5	637.2	480.75	428.4
2007	1233	565.5	2009.85	715.65	562.05	456.9

主成分的解释

特征值为：

4.0945	0.7927	0.0817	0.0207	0.0104

贡献率为：

0.8189	0.1585	0.0163	0.0041	0.0021

累计贡献率为：

0.8189	0.9774	0.9938	0.9979	1.0000

与特征值对应的特征向量矩阵为：

种子费	0.4810	0.2384	-0.0178	0.0039	0.8435
化肥费	0.4875	-0.0792	-0.3359	0.7565	-0.2662
农药费	0.2814	-0.9224	-0.0045	-0.2443	0.1012
机械费	0.4732	0.2683	-0.4613	-0.6058	-0.3527
灌溉费	0.4773	0.1185	0.8210	-0.0321	-0.2882

从表中可以看出，前两个主成分的累计贡献率为97.74%，第一主成分F1在所有变量(除在x3上的载荷稍偏小外)上都有近似相等的正载荷，反映了在种植投入上较为综合的水平，因此第一主成分可称为综合投入成分。第二主成分F2在变量x3（农药）上有很高的负载荷，在变量x2上有较低的负载荷，而在其余变量上均为正载荷。可以认为这个主成分度量了受土壤环境影响的投入(主要是农药，其次是机械费用)在所有投入中占的比重。

注：我这里可能解释的不太好，大家自己做题的时候如果用到了主成分分析，一定得好好琢磨下主成分代表的含义~~~

在Stata中回归结果

注意:Y一定要是标准化后的哦~

```
. reg Y F1 F2
```

Source	SS	df	MS
Model	11.9878887	2	5.99394433
Residual	5.01211134	15	.334140756
Total	17	17	1

Number of obs	= 18
F(2, 15)	= 17.94
Prob > F	= 0.0001
R-squared	= 0.7052
Adj R-squared	= 0.6659
Root MSE	= .57805

Y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
F1	.3543677	.0692848	5.11	0.000	.2066906	.5020447
F2	.4908666	.1574691	3.12	0.007	.1552292	.8265041
_cons	7.64e-16	.1362475	0.00	1.000	-.2904046	.2904046

Source	chi2	df	p
Heteroskedasticity	5.93	5	0.3127
Skewness	2.23	2	0.3280
Kurtosis	0.85	1	0.3577
Total	9.01	8	0.3415

怀特异方差检验

Stata回归代码：

reg Y F1 F2

Stata异方差检验代码:

estat imtest,white

关于主成分回归的看法

问题1：之前学过逐步回归，逐步回归也可以用来解决多重共线性问题，我该用逐步回归还是今天学习的主成分分析呢?

如果你能够很好的解释清楚主成分代表的含义，那么我建议你在正文中既用主成分分析，又用逐步回归（多分析点没啥坏处，只要你能保证你不分析错就行），如果你解释不清楚，那么还是用逐步回归吧。

问题2：主成分回归后，需要将原来的变量带回到回归方程吗?

我觉得没必要，要是你带回去了，那和普通的回归有什么区别呢。主成分的核心作用就是降维，带回去了维度也没降下来呀。

课后作业

（1）想弄懂主成分分析原理的同学，可看该视频：
https://www.bilibili.com/video/av32709936

（2）王学民老师的Mooc视频（多元统计分析）
https://www.icourse163.org/course/tufc-1003381022

（3）完成论文小作业（作业数据.xlsx）。

要求如下：

（1）先直接回归，并用异方差和多重共线性检验。

（2）再使用逐步回归，分析结果。

（3）使用主成分回归，分析结果。

地区	工业总产值y	资产合计x1	工业增加值x2	实收资本x3	长期负债x4	业务收入x5	业务成本x6
1	491.7	380.31	158.39	121.54	22.74	439.65	344.44
2	21.12	30.55	6.4	12.4	3.31	21.17	17.71

部分数据见上表