第8章　主成分分析及R应用

第8章 主成分分析及R应用

- §8.1 主成分分析的背景
- §8.2 主成分分析的基本理论
- §8.3 主成分分析的实例
- §8.4 R语言中主成分分析的常用函数

§8.1 主成分分析的背景

主成分分析（principal components analysis）也称主分量分析，最早可追溯到 K. Pearson 于1901年开创的非随机变量的多元转化分析；是由霍特林（Hotelling）于1933年推广到随机变量。

主成分分析是利用降维的思想，在损失很少信息的前提下把多个指标转化为几个综合指标的多元统计方法。

通常，把转化生成的综合指标称之为主成分，其中的每个主成分都是原始变量的线性组合，且各个主成分之间互不相关。这就使得主成分比原始变量具有某些更优越的性能。

用主成分分析方法的原因：

- 变量个数太多，反映的信息有重叠；
- 变量个数太多，在高维空间研究样本的分布等太复杂；
- 变量个数太多，不易画散点图等；
- 在回归中，自变量个数太多可能存在共线性关系，使参数估计不稳定。

适合做主成分分析的案例展示

下表给出的是美国50个州每100000个人中七种犯罪数量比率的数据。这七种犯罪分别是：

x_1：杀人罪　　x_5：夜盗罪

x_2：勒索罪　　x_6：盗窃罪

x_3：抢劫罪　　x_7：汽车犯罪

x_4：伤害罪

state	x_1	x_2	x_3	x_4	x_5	x_6	x_7
Alabama	14.2	25.2	96.8	278.3	1135.5	1881.9	280.7
Alaska	10.8	51.6	96.8	284	1331.7	3369.8	753.3
Arizona	9.5	34.2	138.2	312.3	2346.1	4467.4	439.5
Arkansas	8.8	27.6	83.2	203.4	972.6	1862.1	183.4
California	11.5	49.4	287	358	2139.4	3499.8	663.5
Colorado	6.3	42	170.7	292.9	1935.2	3903.2	477.1
Connecticut	4.2	16.8	129.5	131.8	1346	2620.7	593.2
Delaware	6	24.9	157	194.2	1682.6	3678.4	467
Florida	10.2	39.6	187.9	449.1	1859.9	3840.5	351.4
Georgia	11.7	31.1	140.5	256.5	1351.1	2170.2	297.9
Hawaii	7.2	25.5	128	64.1	1911.5	3920.4	489.4
Idaho	5.5	19.4	39.6	172.5	1050.8	2599.6	237.6
Illinois	9.9	21.8	211.3	209	1085	2828.5	528.6
Indiana	7.4	26.5	123.2	153.5	1086.2	2498.7	377.4
Iowa	2.3	10.6	41.2	89.8	812.5	2685.1	219.9
Kansas	6.6	22	100.7	180.5	1270.4	2739.3	244.3
Kentucky	10.1	19.1	81.1	123.3	872.2	1662.1	245.4
Louisiana	15.5	30.9	142.9	335.5	1165.5	2469.9	337.7
Maine	2.4	13.5	38.7	170	1253.1	2350.7	246.9
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

相关系数矩阵							
	x1	x2	x3	x4	x5	x6	x7
x1	1.0000	0.6012	0.4837	0.6486	0.3858	0.1019	0.0688
x2	0.6012	1.0000	0.5919	0.7403	0.7121	0.6140	0.3489
x3	0.4837	0.5919	1.0000	0.5571	0.6372	0.4467	0.5907
x4	0.6486	0.7403	0.5571	1.0000	0.6229	0.4044	0.2758
x5	0.3858	0.7121	0.6372	0.6229	1.0000	0.7921	0.5580
x6	0.1019	0.6140	0.4467	0.4044	0.7921	1.0000	0.4442
x7	0.0688	0.3489	0.5907	0.2758	0.5580	0.4442	1.0000

- 该相关矩阵表明，变量之间存在一定的相关性，即彼此之间信息有不少是重复的，从而有一定的降维空间。
- 该案例可用主成分分析进行降维，降维之后再进行分析。

§8.2 主成分分析的基本理论

设对某一目标的研究涉及 p 个指标，分别用 $X_1, X_2, \cdots, X_P$ 表示，这 p 个指标构成的随机向量为 $\mathbf{X} = (X_1, X_2, \cdots, X_p)'$。

设随机向量 $\mathbf{X}$ 的均值为 $\boldsymbol{\mu}$，协方差矩阵为 $\boldsymbol{\Sigma}$。

对 X 进行线性变换，可以形成新的综合变量，用 Y 表示：

$$\begin{cases} Y_1 = u_{11}X_1 + u_{12}X_2 + \cdots + u_{1p}X_p = u_1'X & Y_1: \text{第1主成分} \\ Y_2 = u_{21}X_1 + u_{22}X_2 + \cdots + u_{2p}X_p = u_2'X & Y_2: \text{第2主成分} \\ \cdots\cdots\cdots\cdots & \cdots\cdots \\ Y_p = u_{p1}X_1 + u_{p2}X_2 + \cdots + u_{pp}X_p = u_p'X & Y_p: \text{第}p\text{主成分} \end{cases}$$

一、主成分分析的确定思想

主成分分析的基本思想就是在保留原始变量尽可能多信息的前提下达到降维的目的，从而简化问题的复杂性并抓住问题的主要矛盾。(数理统计学中，一般认为随机变量的方差越大，所含信息越多.)

保留原始变量尽可能多的信息，其含义是：少数几个新变量包含了原始所有变量大部分信息，也就是少数几个新变量的方差之和尽可能接近原始所有变量的方差之和。

对新变量还有另一个要求：新变量包含的信息不重叠.

$$\begin{cases} Y_1 = u_{11}X_1 + u_{12}X_2 + \cdots + u_{1p}X_p = u_1^{'}X \\ Y_2 = u_{21}X_1 + u_{22}X_2 + \cdots + u_{2p}X_p = u_2^{'}X \\ \cdots\cdots\cdots\cdots \\ Y_p = u_{p1}X_1 + u_{p2}X_2 + \cdots + u_{pp}X_p = u_p^{'}X \end{cases}$$

1. 主成分的定义

主成分需满足的三个约束条件：

(1) 主成分的方差依次递减，重要性依次递减，即

$$Var(Y_1) \geq Var(Y_2) \geq \cdots \geq Var(Y_p)$$

(2) 主成分之间互不相关，即无重叠的信息。即

$$Cov(Y_i,\ Y_j) = 0,\ i \neq j,\ i,\ j = 1,\ 2,\ \cdots,\ p$$

(3) 每个主成分的系数平方和为1。即

$$u_{i1}^2 + u_{i2}^2 + \cdots + u_{ip}^2 = 1,\ \ i = 1, 2, \cdots, p$$

$$\begin{cases} Y_1 = u_{11}X_1 + u_{12}X_2 + \cdots + u_{1p}X_p = u_1^{'}X \\ Y_2 = u_{21}X_1 + u_{22}X_2 + \cdots + u_{2p}X_p = u_2^{'}X \\ \cdots\cdots\cdots\cdots \\ Y_p = u_{p1}X_1 + u_{p2}X_2 + \cdots + u_{pp}X_p = u_p^{'}X \end{cases}$$

希望在约束条件 $\|u_1\|=1$ 下寻求向量u_1，使得 $Var(Y_1)=u_1^{'}\Sigma u_1$ 达到最大，Y_1就称为第一主成分。

如果第一主成分所含信息不够多，则需考虑第二主成分：在约束条件$\|u_2\|=1$，$\mathrm{Cov}(Y_1, Y_2)=0$下使得$Var(Y_2)=u_2^{'}\Sigma u_2$ 达到最大，所求的Y_2称为第二主成分.

2. 主成分的求解与结论

(1) X是p维随机向量，$D(X)=\Sigma.$

$(2)Y_1=\mu_1^T X,\mu_1$是一个p维列向量，$\|\mu_1\|=1.$

$(3)D(Y_1)=\mu_1^T\Sigma\mu_1.$

目标：在所有p维单位列向量中，找使得$D(Y_1)=\mu_1^T\Sigma\mu_1$取到最大值.

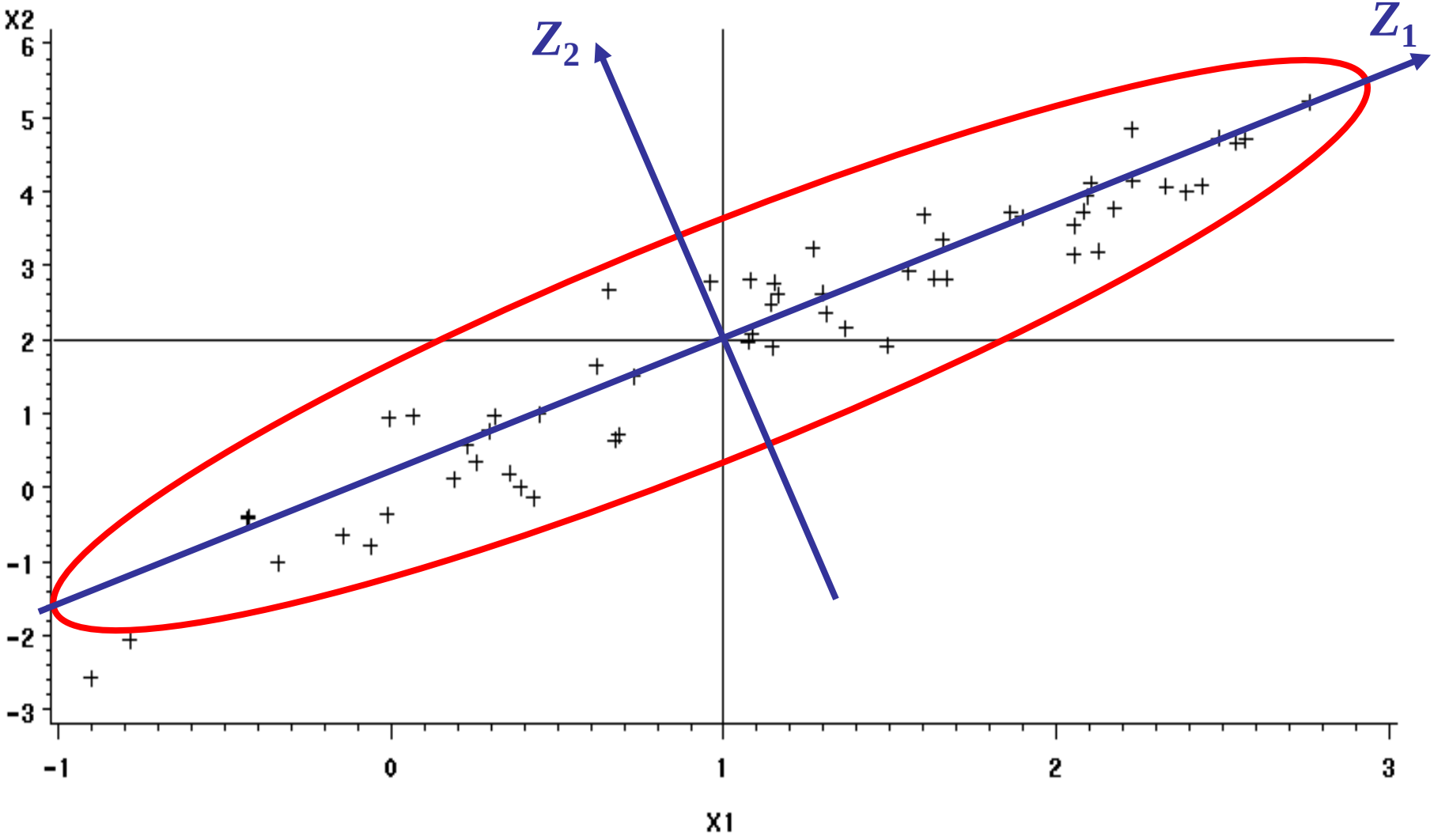
结论：Σ的特征值为$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p,$

相应的正交单位特征向量为$u_1,u_2,\cdots,u_p,$则

第一主成分为：$Y_1=\mu_1^T X.$　第k主成分为：$Y_k=\mu_k^T X.$

随机向量X的主成分Y(随机向量)，是以Σ的单位正交特征向量作为系数的X的线性组合，Y的分量之间互不相关，且Y的各分量的方差为Σ的从大到小排列的特征根。

3. 主成分的几何理解



- 例1：　设 $X=(X_1, X_2, X_3)'$的协方差矩阵为

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

问：由该协差阵出发，求解三个主成分。

- 例1：　设 $X=(X_1, X_2, X_3)'$的协方差矩阵为

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

问：由该协差阵出发，求解三个主成分。

[提示]：

（1）若手动计算，思路：

求 $|\Sigma - \lambda \cdot I| = 0$ 的p个根，即得到Σ的p个特征根.

代入每个特征根，求矩阵（ $\Sigma - \lambda \cdot I$ ）所对应的齐次线性方程组的一组（单位化的）基础解系，即得到单位正交特征向量.

（2）R软件实现，可处理高维矩阵，且更方便简单。

- 例1：　设 $X=(X_1, X_2, X_3)'$的协方差矩阵为

$$\boldsymbol{\Sigma}=\begin{pmatrix}1 & -2 & 0\\ -2 & 5 & 0\\ 0 & 0 & 2\end{pmatrix}$$

解：该协差阵的特征值为

$$\lambda_1=5.83，\lambda_2=2.00，\lambda_3=0.17$$

相应的特征向量为

$$u_1=\begin{pmatrix}-0.383\\ 0.924\\ 0.000\end{pmatrix}, u_2=\begin{pmatrix}0\\ 0\\ 1\end{pmatrix}, u_3=\begin{pmatrix}0.924\\ 0.383\\ 0.000\end{pmatrix}.$$

R语言代码：求矩阵的特征根和单位正交特征向量

```
x=matrix(c(1,-2,0,-2,5,0,0,0,2),3,3)        #给出矩阵
eigen(x)      #求得x的特征根和单位正交特征向量
```

- 例1：　设 $X=(X_1, X_2, X_3)'$的协方差矩阵为

$$\boldsymbol{\Sigma}=\begin{pmatrix}1 & -2 & 0\\ -2 & 5 & 0\\ 0 & 0 & 2\end{pmatrix}$$

解：该协差阵的特征值为

$$\lambda_1=5.83，\lambda_2=2.00，\lambda_3=0.17$$

相应的特征向量为

$$u_1=\begin{pmatrix}-0.383\\ 0.924\\ 0.000\end{pmatrix}, u_2=\begin{pmatrix}0\\ 0\\ 1\end{pmatrix}, u_3=\begin{pmatrix}0.924\\ 0.383\\ 0.000\end{pmatrix}.$$

- 例1：　设 $X=(X_1, X_2, X_3)'$的协方差矩阵为

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

解：该协差阵的特征值为

$$\lambda_1=5.83，\ \lambda_2=2.00，\ \lambda_3=0.17$$

相应的特征向量为

$$u_1 = \begin{pmatrix} -0.383 \\ 0.924 \\ 0.000 \end{pmatrix}, u_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, u_3 = \begin{pmatrix} 0.924 \\ 0.383 \\ 0.000 \end{pmatrix}.$$

则：　第一主成分为 $Y_1 = \mathrm{u}_1'X$

第二主成分为 $Y_2 = \mathrm{u}_2'X$

第三主成分为 $Y_3 = \mathrm{u}_3'X$

4. 主成分的性质

性质1. Y 的协方差阵为对角阵： $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_p)$ ，

即 $\mathrm{Var}(Y_i) = \lambda_i$ ， $i=1, 2, \cdots, p$, 且 $Y_1, Y_2, \cdots, Y_p$ 互不相关。

性质2. 记 $\boldsymbol{\Sigma} = (\sigma_{ij})_{p\times p}$ ， 有 $\sum_{i=1}^{p}\lambda_i = \sum_{i=1}^{p}\sigma_{ii}$ ， 即 $\sum_{i=1}^{p}\mathrm{Var}(Y_i) = \sum_{i=1}^{p}\mathrm{Var}(X_i)$ 。

性质3. Y 的p个分量按方差大小、由大到小排列。

称 $\alpha_k = \dfrac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$ 为第k个主成分的方差贡献率。

称 $\dfrac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$ 为主成分前k个主成分的累积贡献率。

性质4. 主成分负荷： $\rho(Y_i, X_j) = \sqrt{\lambda_i}u_{ij} \Big/ \sqrt{\sigma_{jj}}, (i, j = 1, 2, \cdots, p).$

表示 Y_i 与 X_j 之间的相关系数，记为 $\rho_{Yi, Xj}$

（详见教材P222）

【注】主成分负荷量是主成分解释的依据之一，

负荷量的绝对值大小、及其与原始变量的对应关系，

可用于探究该主成分的可能成因。

*** 性质5** $\sum_{j=1}^{p} \rho^2_{Y_i, X_j} \cdot \sigma_{jj} = \lambda_i$

*** 性质6** $\sum_{i=1}^{p} \rho^2_{Y_i, X_j} = 1$

- 例1：设$X=(X_1, X_2, X_3)'$的协方差矩阵为

$$\boldsymbol{\Sigma}=\begin{pmatrix}1 & -2 & 0\\ -2 & 5 & 0\\ 0 & 0 & 2\end{pmatrix}$$

解：其特征值为

$$\lambda_1=5.83，\ \lambda_2=2.00，\ \lambda_3=0.17$$

相应的特征向量为

$$\mathrm{u}_1=\begin{pmatrix}-0.383\\ 0.924\\ 0.000\end{pmatrix},\quad \mathrm{u}_2=\begin{pmatrix}0\\ 0\\ 1\end{pmatrix},\quad \mathrm{u}_3=\begin{pmatrix}0.924\\ 0.383\\ 0.000\end{pmatrix}$$

第一主成分为 $Y_1=\mathrm{u}_1'X$ ，……

易知：$\sigma_{11}+\sigma_{22}+\sigma_{33}=1+5+2=\lambda_1+\lambda_2+\lambda_3=5.83+2.00+0.17=8$

若只取一个主成分，则其方差贡献率为

$$5.83/(5.83+2.00+0.17)=0.72875=72.875\%$$

- 例1：设$X=(X_1, X_2, X_3)'$的协方差矩阵为

$$\boldsymbol{\Sigma}=\begin{pmatrix}1 & -2 & 0\\ -2 & 5 & 0\\ 0 & 0 & 2\end{pmatrix}$$

（续）进一步，前两个主成分占总方差的（5.83+2）/8=0.98 。此时，两个成分Y_1和Y_2可以用于代替原先的三个变量，且信息损失较少。

$$\rho_{Y_1},\ X_1=\frac{u_{11}\sqrt{\lambda_1}}{\sqrt{\sigma_{11}}}=\frac{-0.383\times\sqrt{5.83}}{\sqrt{1}}=\text{-}0.925,$$

$$\rho_{Y_1},\ X_2=\frac{u_{12}\sqrt{\lambda_1}}{\sqrt{\sigma_{22}}}=\frac{0.924\times\sqrt{5.83}}{\sqrt{5}}=0.998,$$

$$\rho_{Y_1},\ X_3=\frac{u_{13}\sqrt{\lambda_1}}{\sqrt{\sigma_{33}}}=\frac{0\times\sqrt{5.83}}{\sqrt{2}}=0,$$

$$\rho_{Y_2},\ X_1=\rho_{Y_2},\ X_2=0\ ,\ \ \rho_{Y_2},\ X_3=\frac{1\times\sqrt{\lambda_2}}{\sqrt{\sigma_{33}}}=\frac{\sqrt{2}}{\sqrt{2}}=1。$$

5. 主成分个数的确定方法

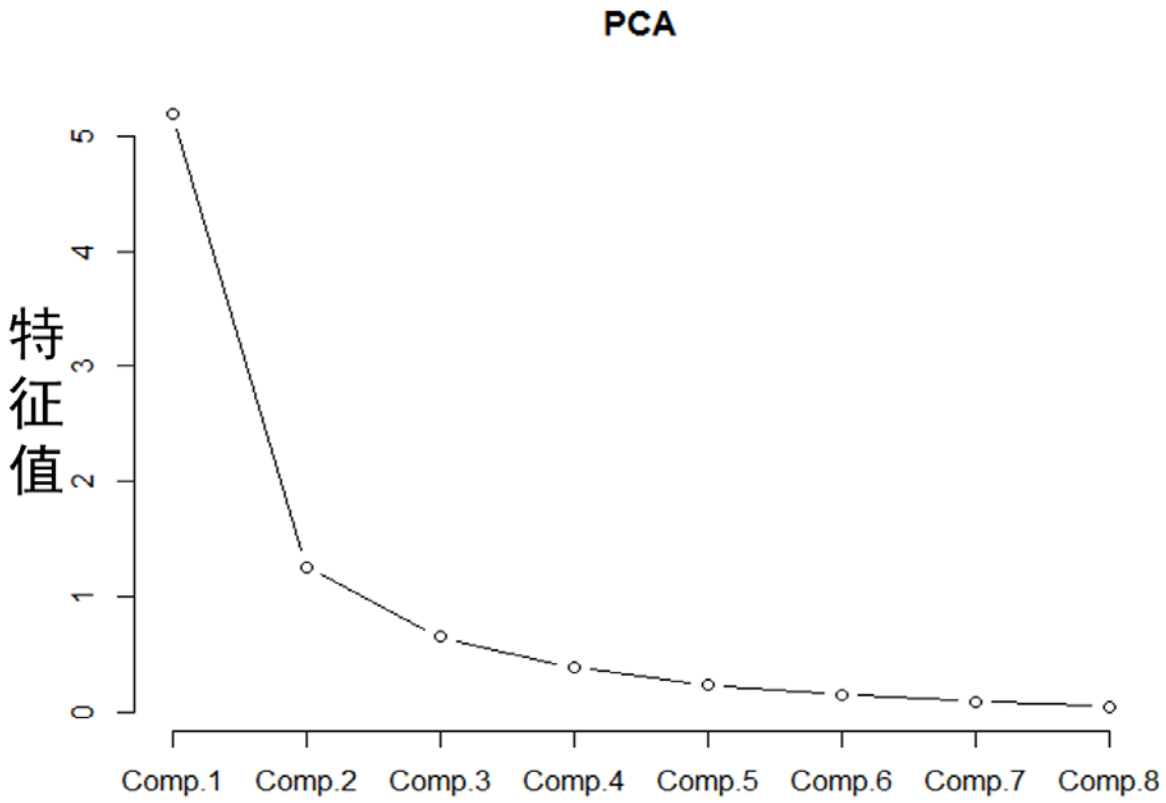
进行主成分分析的目的之一：为了减少变量的个数。所以，一般不会取 p 个主成分，而是取 $m < p$ 个主成分；

m 取多少合适？这是一个实际问题，通常以所取 m 使得累积贡献率达到80％以上为宜：

$$\frac{\sum_{i=1}^{m}\lambda_i}{\sum_{i=1}^{p}\lambda_i} \geq 80\%$$

既能使损失信息不太多，又达到减少变量、简化问题的目的。实际分析中，最常见的情况是主成分为2到3个。当然，这不绝对。

碎石图



选取主成分还可根据特征值的变化来确定：由碎石图可知，第二个及第三个特征值变化的趋势已经开始趋于平稳，所以，取前两个或是前三个主成分是比较合适的。这种方法确定的主成分个数，与按累积贡献率确定的主成分个数往往一致。在实际应用中，有些研究者习惯于保留特征值大于1的那些主成分。

6. 数据是否需要标准化？

	x1	x2
1	147	32
2	171	57
3	175	64
4	159	41
5	155	38
6	152	35
7	158	44
8	154	41
9	164	54
10	168	57
11	166	49
12	159	47
13	164	46
14	177	63

$$Y = 0.656X_1 + 0.755X_2$$

	x3	x4
1	1.47	32000
2	1.71	57000
3	1.75	64000
4	1.59	41000
5	1.55	38000
6	1.52	35000
7	1.58	44000
8	1.54	41000
9	1.64	54000
10	1.68	57000
11	1.66	49000
12	1.59	47000
13	1.64	46000
14	1.77	63000

$$Y = 0 \cdot X_1 + 1 \cdot X_2$$

	x5	x6
1	1470	32
2	1710	57
3	1750	64
4	1590	41
5	1550	38
6	1520	35
7	1580	44
8	1540	41
9	1640	54
10	1680	57
11	1660	49
12	1590	47
13	1640	46
14	1770	63

$$Y = 0.994 \cdot X_1 + 0.110 \cdot X_2$$

变量标准化：$X_i^* = \dfrac{X_i - \mu_i}{\sqrt{\sigma_{ii}}} \quad i = 1, 2, \cdots, p$

其中，μ_i与 σ_{ii} 分别表示变量 X_i 的期望与方差。

$$E(X_i^*) = 0, \mathrm{var}(X_i^*) = 1, Cov\left(X_i^*, X_j^*\right) = \rho_{X_i, X_j}.$$

$X^* = \begin{pmatrix} X_1^* \\ \vdots \\ X_p^* \end{pmatrix}$ 的协方差矩阵为 $X = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}$ 的相关系数矩阵.

从相关阵R出发求解主成分

R的特征值为$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$,

相应的正交单位特征向量为$u_1, u_2, \cdots, u_p$,则

第k主成分为：$Y_k = \mu_k{}^T X^*$.

由相关阵出发所求得主成分，仍有前述的各种性质.

1. Y的协方差矩阵为对角阵 $\boldsymbol{\Lambda}$;

2. $\sum_{i=1}^{p} \mathrm{var}(Y_i) = \sum_{i=1}^{p} \mathrm{var}(X_i^*) = p.$

3. 第k个主成分的方差占总方差的比例,

   即:第k个主成分的方差贡献率为 $\alpha_k = \lambda_k / p$ 。

   前m个主成分的累积方差贡献率为 $\sum_{i=1}^{m} \lambda_i / p$ 。

4. $\rho(Y_i, X_j^*) = u_{ij}\sqrt{\lambda_i}$。

确定主成分个数的准则

1. 由累积贡献率确定；
2. 由特征值确定（碎石图）.

例2　X的协方差矩阵$\Sigma=\begin{bmatrix}1 & 4\\ 4 & 100\end{bmatrix}$，相关矩阵 $R=\begin{bmatrix}1 & 0.4\\ 0.4 & 1\end{bmatrix}$

（1）Σ的特征值与特征向量是

$\lambda_1=100.16$，　　$u_1'=[0.040，0.999]$

$\lambda_2=0.84$，　　$u_2'=[0.999，-0.040]$

主成分为$Y_1=0.040X_1+0.999X_2$，$Y_2=0.999X_1-0.040X_2$

（2）R的特征值与特征向量是

$\lambda_1=1.4$，　　$u_1'=[0.707，0.707]$

$\lambda_2=0.6$，　　$u_2'=[0.707，-0.707]$

主成分为$Y_1=0.707X_1^*+0.707X_2^*=0.707\dfrac{X_1-\mu_1}{\sqrt{\sigma_{11}}}+0.707\dfrac{X_2-\mu_2}{\sqrt{\sigma_{22}}}$

$=0.707(X_1-\mu_1)+0.0707(X_2-\mu_2)$

$Y_2=0.707X_1^*-0.707X_2^*=0.707\dfrac{X_1-\mu_1}{\sqrt{\sigma_{11}}}-0.707\dfrac{X_2-\mu_2}{\sqrt{\sigma_{22}}}$

$=0.707(X_1-\mu_1)-0.0707(X_2-\mu_2)$

7. 主成分分析的注意事项

（1）主成分分析，最好以相关系数矩阵为主：

对于度量单位不同的指标，或取值范围彼此差异大的指标，不直接使用协差阵出发的分析；而应考虑数据的标准化，即由相关阵出发求主成分。

（2）协方差矩阵与相关系数矩阵需要估计

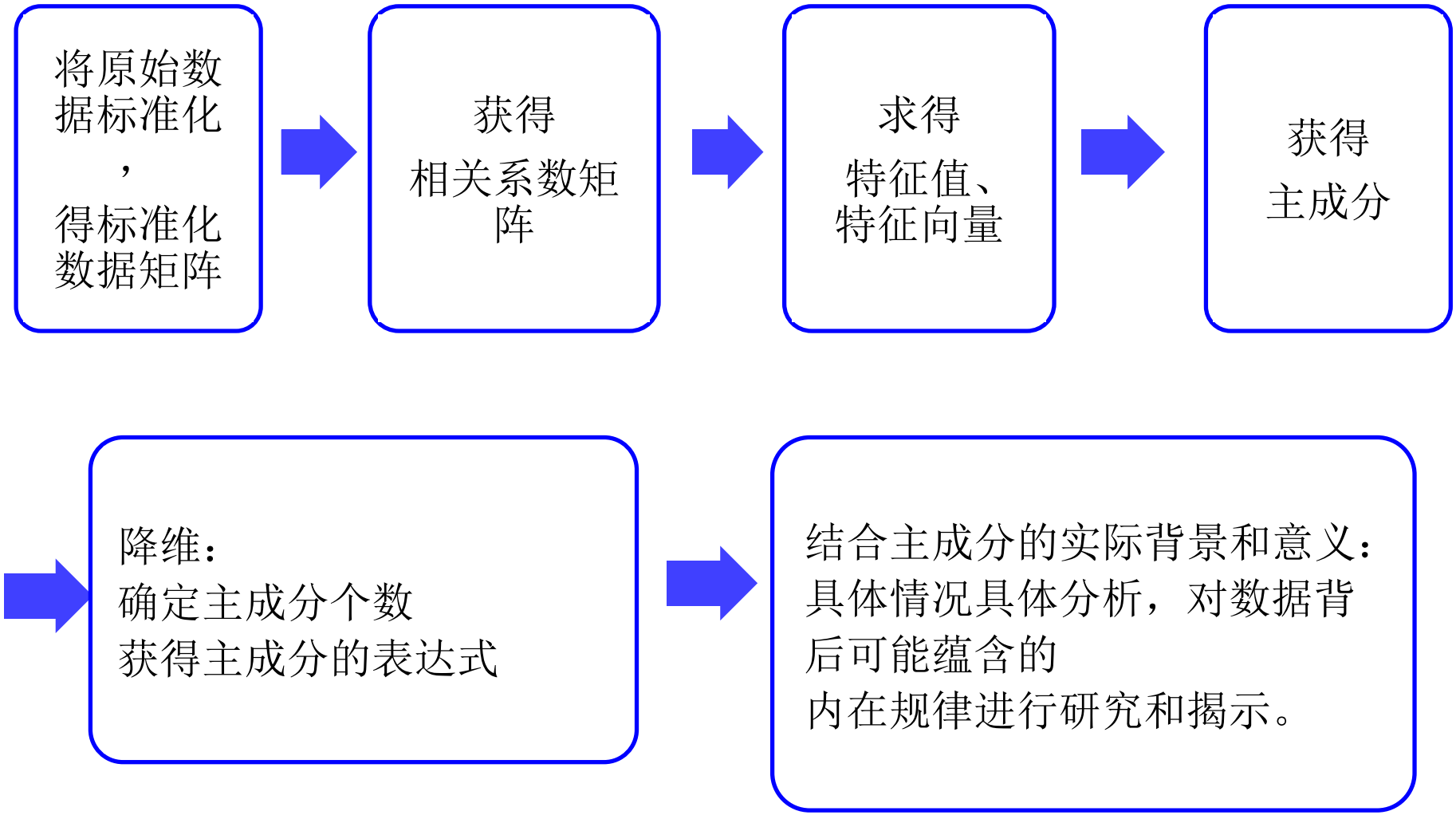
随机向量X的协方差矩阵 $\sum$ 和相关系数矩阵 R 往往是未知的，需要根据数据进行估计。

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

$$\overline{x_i}=\frac{1}{n}\sum_{k=1}^{n}x_{ki}, i=1,\ldots,p; \quad S_{ij}=\frac{1}{n-1}\sum_{k=1}^{n}(x_{ki}-\overline{x_i})(x_{kj}-\overline{x_j}); \quad r_{ij}=\frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}}.$$

$$\hat{\Sigma}=\begin{pmatrix} S_{11} & \cdots & S_{1p} \\ \vdots & \ddots & \vdots \\ S_{p1} & \cdots & S_{pp} \end{pmatrix}, \hat{R}=\begin{pmatrix} 1 & \cdots & r_{1p} \\ \vdots & \ddots & \vdots \\ r_{p1} & \cdots & 1 \end{pmatrix}.$$

总结：主成分分析的步骤


将原始数据标准化，得标准化数据矩阵
获得相关系数矩阵
求得特征值、特征向量
获得主成分
降维：
确定主成分个数
获得主成分的表达式
结合主成分的实际背景和意义：
具体情况具体分析，对数据背后可能蕴含的
内在规律进行研究和揭示。

§8.3 主成分分析的实例

- 实例. 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：

x_1：身高　　x_4：手臂长

x_2：坐高　　x_5：肋围

x_3：胸围　　x_6：腰围

所得样本相关系数矩阵R：

男子身材六项指标的样本相关系数矩阵R

	x_1	x_2	x_3	x_4	x_5	x_6
x_1	1.00					
x_2	0.79	1.00				
x_3	0.36	0.31	1.00			
x_4	0.76	0.55	0.35	1.00		
x_5	0.25	0.17	0.64	0.16	1.00	
x_6	0.51	0.35	0.58	0.38	0.63	1.00

---摘自《王学民教授-应用多元统计分析》

表：前三个特征值、特征向量以及贡献率

特征向量	u_1	u_2	u_3
x_1^* ： 身高	0.469	−0.365	0.092
x_2^* ： 坐高	0.404	−0.397	0.613
x_3^* ： 胸围	0.394	0.397	−0.279
x_4^* ： 手臂长	0.408	−0.365	−0.705
x_5^* ： 肋围	0.337	0.569	0.164
x_6^* ： 腰围	0.427	0.308	0.119
特征值	3.287	1.406	0.459
贡献率	0.548	0.234	0.077
累计贡献率	0.548	0.782	0.859

- 前三个主成分分别为

$$y_1 = 0.469x_1^* + 0.404x_2^* + 0.394x_3^* + 0.408x_4^* + 0.337x_5^* + 0.427x_6^*$$

$$y_2 = -0.365x_1^* - 0.397x_2^* + 0.397x_3^* - 0.365x_4^* + 0.569x_5^* + 0.308x_6^*$$

$$y_3 = 0.092x_1^* + 0.613x_2^* - 0.279x_3^* - 0.705x_4^* + 0.164x_5^* + 0.119x_6^*$$

- 根据累计贡献率可考虑取前面两个或三个主成分。
- 称第一主成分为（身材）大小成分，称第二主成分为形状成分（或胖瘦成分），称第三主成分为臂长成分。
- 可考虑取前两个主成分。

§8.4　R语言：主成分分析

R语言中的主成分分析函数

pc= princomp(data, cor=FALSE, scores=FLASE)

1. cor=FALSE: 根据协方差矩阵进行主成分分析；

cor=TRUE: 根据相关系数矩阵进行主成分分析；

2. scores=FALSE: 不输出主成分得分；

scores=TRUE: 输出主成分得分。

pc$loadings：	输出特征向量；主成分的系数；
pc$sdev：	特征值的平方根；主成分的标准差；
pc$sdev^2：	特征值；主成分的方差；
pc$scores：	主成分得分。

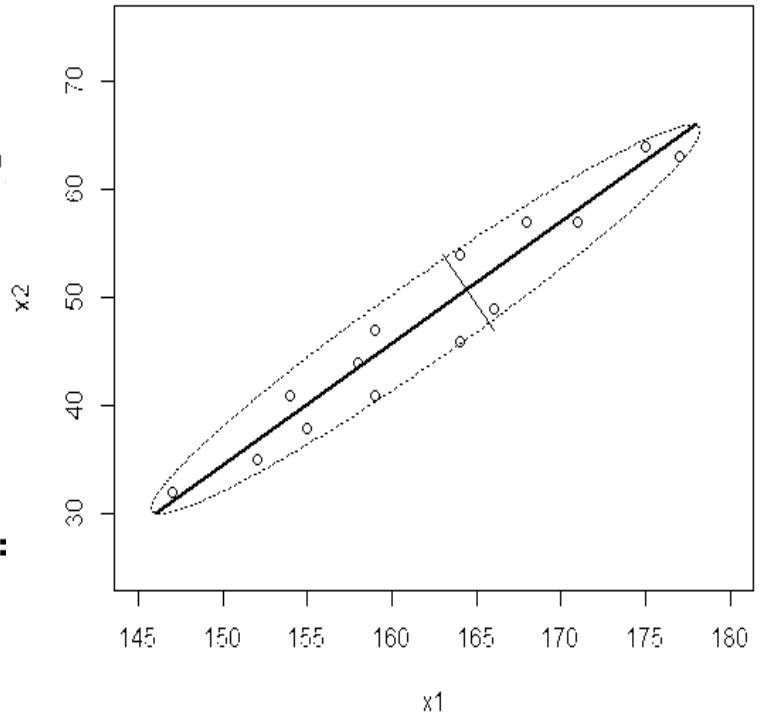
案例一：身高与体重数据的主成分分析

```
x1=c(147,171,175,159,155,152,158,154,164,168,166,159,164,177)
x2=c(32,57,64,41,38,35,44,41,54,57,49,47,46,63)
```

```
cor(x1,x2)
```

0.9672

```
plot(x1,x2,xlim=c(145,180),ylim=c(25
lines(c(146,178),c(30,66),lwd=2)
lines(c(163,166),c(54,47))
library(shape)
lines(getellipse(24,3,mid=c(162,48),angle=
```



1. 主成分分析函数

```
x1=c(147,171,175,159,155,152,158,154,164,168,166,159,164,177)
x2=c(32,57,64,41,38,35,44,41,54,57,49,47,46,63)
X=data.frame(x1,x2)
```

```
pc= princomp(X, cor=FALSE, scores=TRUE)
```

```
>pc$loadings        #输出系数： u1, u2
      Loadings:
     Comp.1 Comp.2
  x1  0.656  0.755
  x2  0.755 -0.656
```

```
>pc$sdev^2          #输出： 主成分方差=主成分标准差的平方
   Comp.1  Comp.2
  176.331   2.884
```

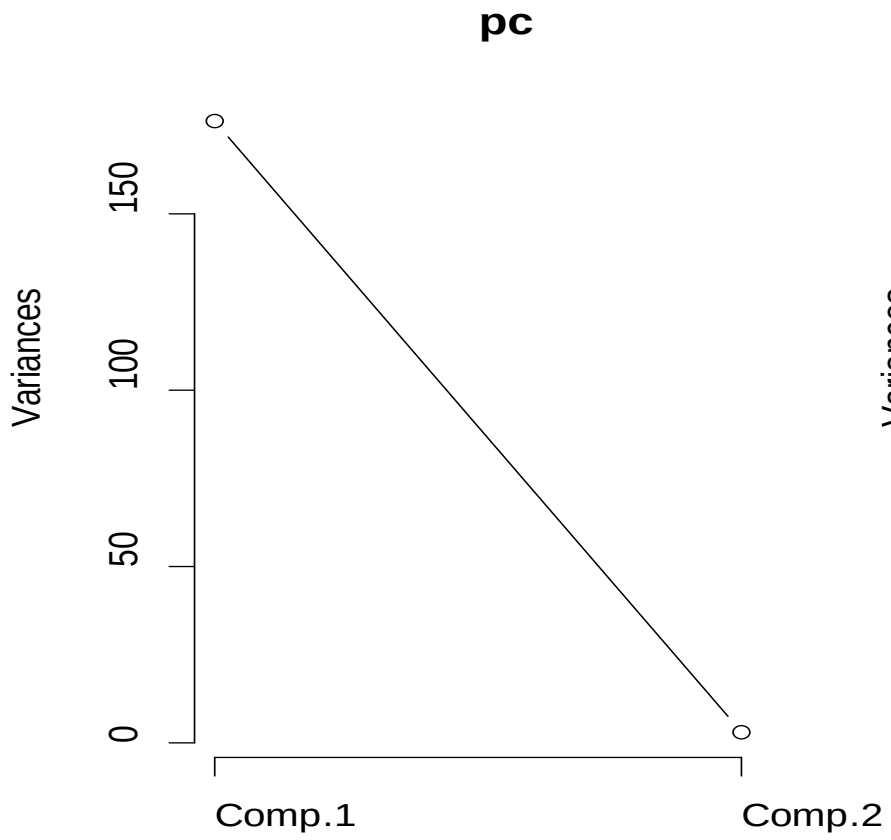
2. 确定主成分的个数：贡献率与碎石图

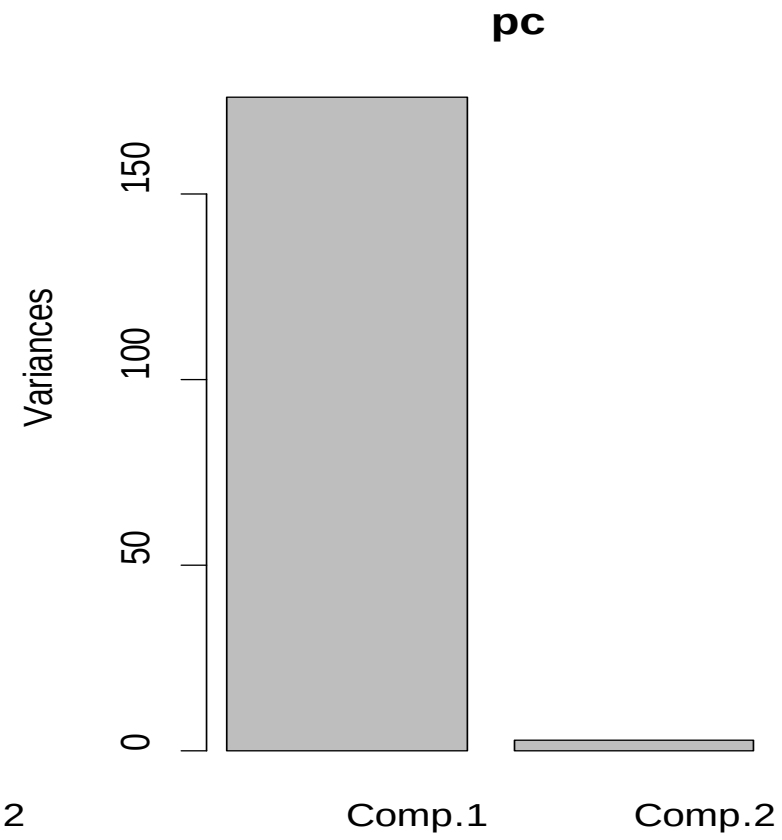
summary(pc)　#贡献率

```
> summary(pc)
Importance of components:
                                Comp.1      Comp.2
Standard deviation     12.7959253 1.63632584
Proportion of Variance  0.9839102 0.01608984
Cumulative Proportion   0.9839102 1.00000000
```

碎石图的绘制：便于直观观察特征根的变化情况

screeplot(pc, type=“lines”)　　#碎石图
screeplot(pc, type=“barplot”)　#条形图形式





3. 主成分表达式

>pc$loadings　　#输出系数：$u_1$，$u_2$

Loadings:

	Comp.1	Comp.2
x1	0.656	0.755
x2	0.755	-0.656

$$Z_1 = 0.656X_1 + 0.755X_2 \qquad Z_2 = 0.755X_1 - 0.656X_2$$

第一主成分：系数都为正，且值相差不大。身高大、体重大的学生，第一主成分的值就较大。

第二主成分：身高系数为正、体重系数为负，且绝对值相差不大：身高大、体重小的学生，第二主成分的值就较大。身高小、体重大的学生，第二主成分的值就小。

4. 主成分得分

$Z_1 = 0.656X_1 + 0.755X_2$　　$Z_2 = 0.755X_1 - 0.656X_2$

pc$scores

```
> pc$scores
             Comp.1      Comp.2
 [1,] -21.7469609 -1.0753729
 [2,]  12.8653792  0.6526072
 [3,]  20.7733282 -0.9172151
 [4,]  -7.0833638  2.0835700
 [5,] -11.9712295  1.0305890
 [6,] -16.2033944  0.7326289
 [7,]  -5.4740021 -0.6385533
 [8,] -10.3618678 -1.6915342
 [9,]   6.0104111 -2.6654363
[10,]  10.8982768 -1.6124553
[11,]   3.5467085  2.1231094
[12,]  -2.5532388 -1.8506348
[13,]  -0.0297556  2.5801701
[14,]  21.3297090  1.2485274
```

```
> X
    x1 x2
1  147 32
2  171 57
3  175 64
4  159 41
5  155 38
6  152 35
7  158 44
8  154 41
9  164 54
10 168 57
11 166 49
12 159 47
13 164 46
14 177 63
```

$$Z_1 = 0.656X_1 + 0.755X_2 \qquad Z_2 = 0.755X_1 - 0.656X_2$$

```
> X
    x1 x2
1  147 32
2  171 57
3  175 64
4  159 41
5  155 38
6  152 35
7  158 44
8  154 41
9  164 54
10 168 57
11 166 49
12 159 47
13 164 46
14 177 63
```

```
> scale(X,scale=FALSE)
                x1          x2
 [1,] -15.071429 -15.7142857
 [2,]   8.928571   9.2857143
 [3,]  12.928571  16.2857143
 [4,]  -3.071429  -6.7142857
 [5,]  -7.071429  -9.7142857
 [6,] -10.071429 -12.7142857
 [7,]  -4.071429  -3.7142857
 [8,]  -8.071429  -6.7142857
 [9,]   1.928571   6.2857143
[10,]   5.928571   9.2857143
[11,]   3.928571   1.2857143
[12,]  -3.071429  -0.7142857
[13,]   1.928571  -1.7142857
[14,]  14.928571  15.2857143
attr(,"scaled:center")
       x1        x2
162.07143  47.71429
```

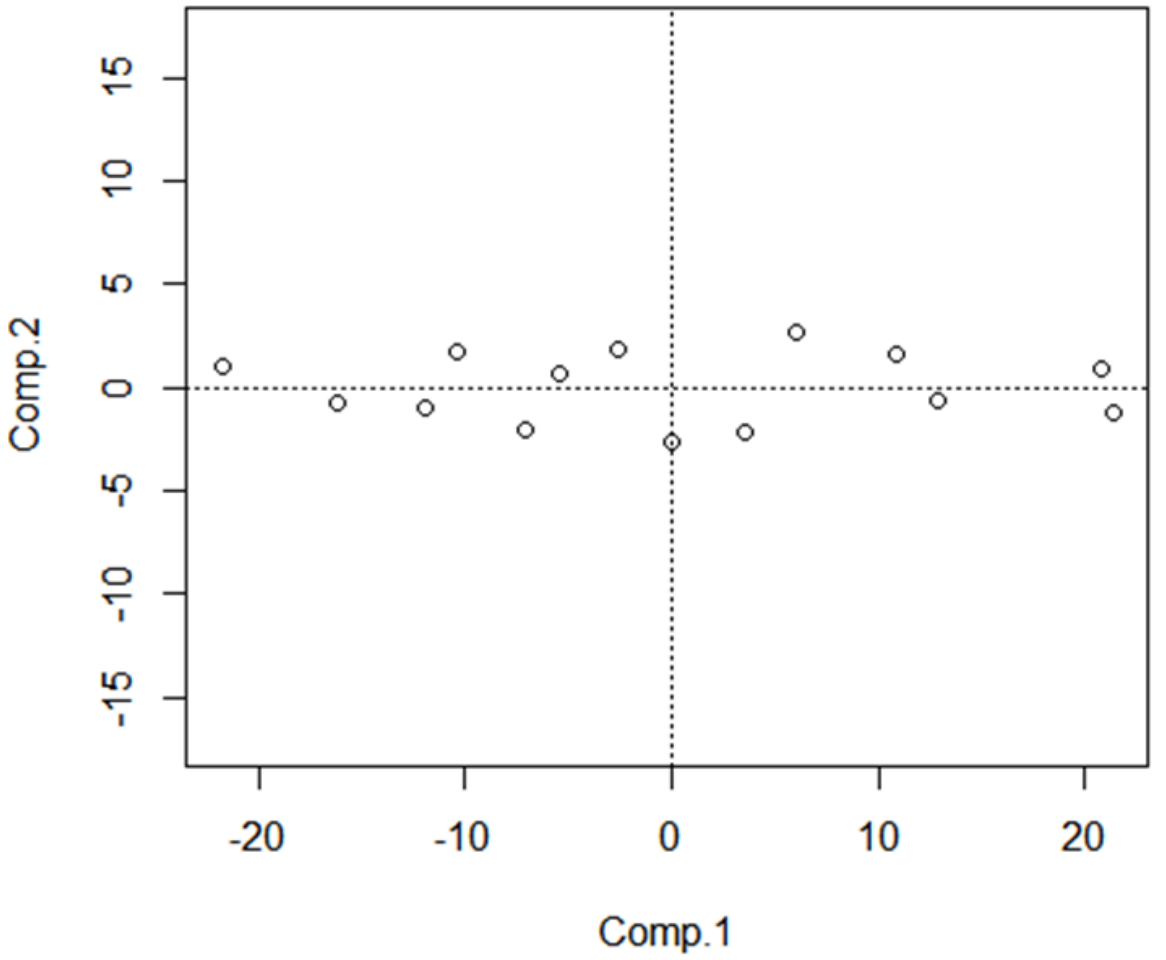
```
> scale(X,scale=FALSE)%*%pc$loadings
            Comp.1     Comp.2
 [1,] -21.7469609 -1.0753729
 [2,]  12.8653792  0.6526072
 [3,]  20.7733282 -0.9172151
 [4,]  -7.0833638  2.0835700
 [5,] -11.9712295  1.0305890
 [6,] -16.2033944  0.7326289
 [7,]  -5.4740021 -0.6385533
 [8,] -10.3618678 -1.6915342
 [9,]   6.0104111 -2.6654363
[10,]  10.8982768 -1.6124553
[11,]   3.5467085  2.1231094
[12,]  -2.5532388 -1.8506348
[13,]  -0.0297556  2.5801701
[14,]  21.3297090  1.2485274
```

5. 主成分得分散点图

pc$scores

	Comp. 1	Comp. 2
[1,]	−21.747	1.075
[2,]	12.865	−0.653
[3,]	20.773	0.917
[4,]	−7.083	−2.084
[5,]	−11.971	−1.031
[6,]	−16.203	−0.733
[7,]	−5.474	0.639
[8,]	−10.362	1.692
[9,]	6.010	2.665
[10,]	10.898	1.612
[11,]	3.547	−2.123
[12,]	−2.553	1.851
[13,]	−0.030	−2.580
[14,]	21.330	−1.249

plot(pc$scores,asp=1);abline(h=0,v=0,lty=3)



6. 双标图

```
biplot(pc$scores,pc$loadings)
abline(h=0,v=0,lty=3)
```

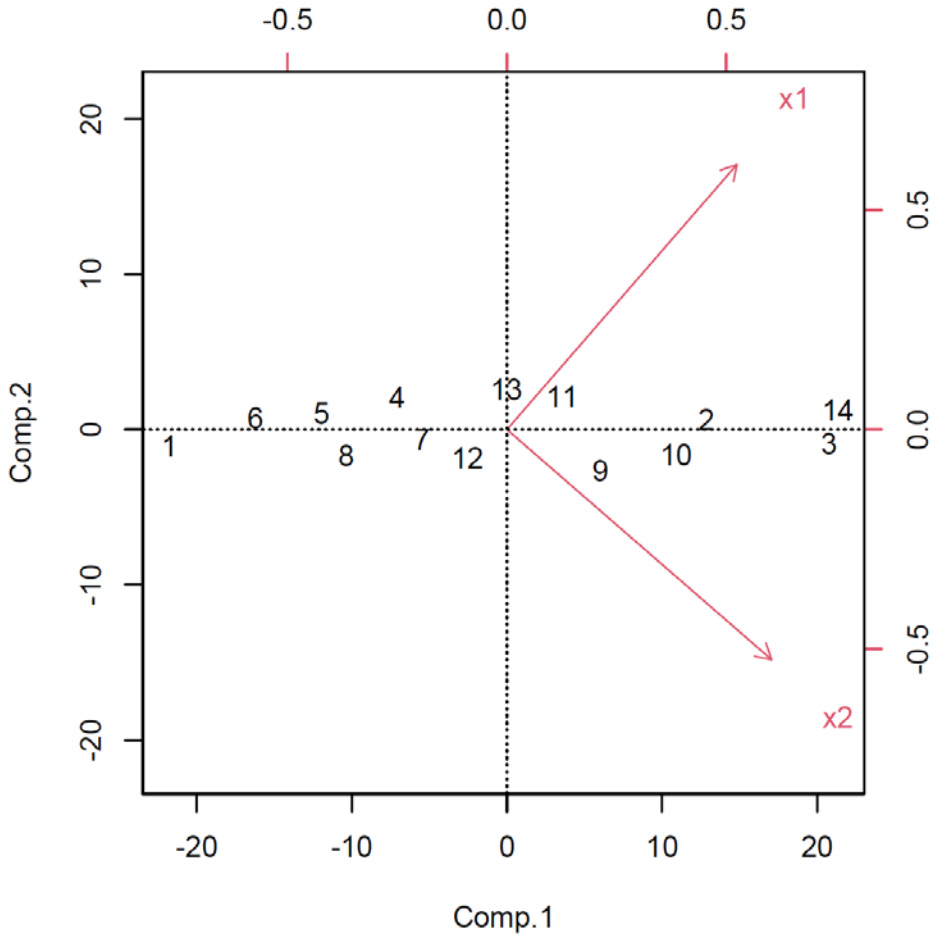
```
>pc$loadings
Loadings:
   Comp.1 Comp.2
x1  0.656  0.755
x2  0.755 -0.656
```

$$Z_1 = 0.656X_1 + 0.755X_2$$

$$Z_2 = 0.755X_1 - 0.656X_2$$

$$X_1 = 0.656Z_1 + 0.755Z_2$$

$$X_2 = 0.755Z_1 - 0.656Z_2$$



案例二：身体四项指标的主成分分析

在某中学随机抽取某年级30名学生，测量其身高X_1、体重X_2、胸围X_3和坐高X_4。对这30名中学生的四项指标数据做主成分分析。

序号	X_1	X_2	X_3	X_4	序号	X_1	X_2	X_3	X_4
1	148	41	72	78	16	152	35	73	79
2	139	34	71	76	17	149	47	82	79
3	160	49	77	86	18	145	35	70	77
4	149	36	67	79	19	160	47	74	87
5	159	45	80	86	20	156	44	78	85
6	142	31	66	76	21	151	42	73	82
7	153	43	76	83	22	147	38	73	78
8	150	43	77	79	23	157	39	68	80
9	151	42	77	80	24	147	30	65	75
10	139	31	68	74	25	157	48	80	88
11	140	29	64	74	26	151	36	74	80
12	161	47	78	84	27	144	36	68	76
13	158	49	78	83	28	141	30	67	76
14	140	33	67	77	29	139	32	68	73
15	137	31	66	73	30	148	38	70	78

1. 输入数据

```
student<-data.frame(
   X1=c(148, 139, 160, 149, 159, 142, 153, 150, 151, 139,
        140, 161, 158, 140, 137, 152, 149, 145, 160, 156,
        151, 147, 157, 147, 157, 151, 144, 141, 139, 148),
   X2=c(41, 34, 49, 36, 45, 31, 43, 43, 42, 31,
        29, 47, 49, 33, 31, 35, 47, 35, 47, 44,
        42, 38, 39, 30, 48, 36, 36, 30, 32, 38),
   X3=c(72, 71, 77, 67, 80, 66, 76, 77, 77, 68,
        64, 78, 78, 67, 66, 73, 82, 70, 74, 78,
        73, 73, 68, 65, 80, 74, 68, 67, 68, 70),
   X4=c(78, 76, 86, 79, 86, 76, 83, 79, 80, 74,
        74, 84, 83, 77, 73, 79, 79, 77, 87, 85,
        82, 78, 80, 75, 88, 80, 76, 76, 73, 78)
)
```

2. 做主成分分析以及基本结果

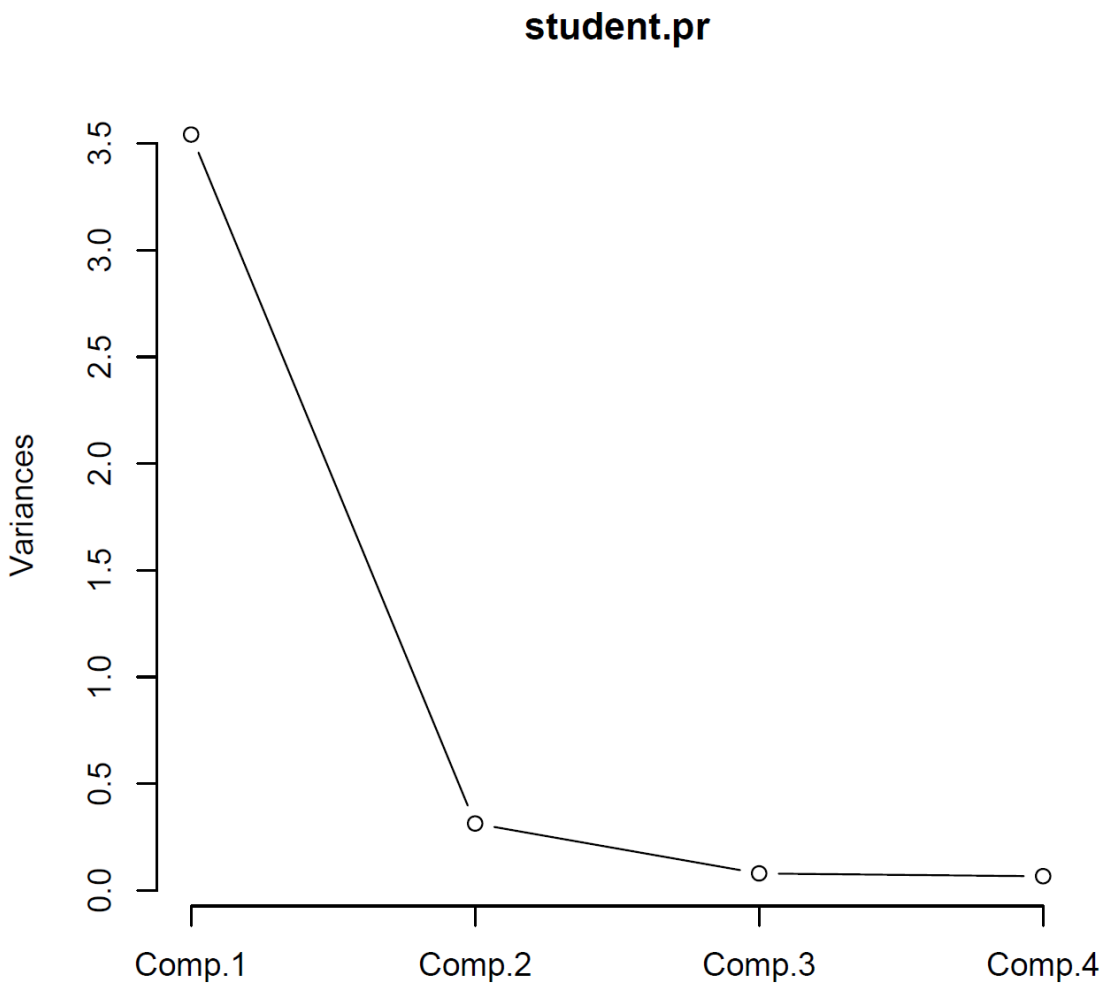
```
student.pr<-princomp(student, cor=TRUE)
# 作主成分分析，选择使用相关系数矩阵
summary(student.pr, loadings=TRUE)
# 显示结果，方差累积率以及载荷矩阵
```

```
Importance of components:
                          Comp.1     Comp.2     Comp.3     Comp.4
Standard deviation     1.8817805 0.55980636 0.28179594 0.25711844
Proportion of Variance 0.8852745 0.07834579 0.01985224 0.01652747
Cumulative Proportion  0.8852745 0.96362029 0.98347253 1.00000000

Loadings:
   Comp.1 Comp.2 Comp.3 Comp.4
X1  0.497  0.543  0.450  0.506
X2  0.515 -0.210  0.462 -0.691
X3  0.481 -0.725 -0.175  0.461
X4  0.507  0.368 -0.744 -0.232
```

3. 碎石图

screeplot(student.pr,type="lines") ###碎石图



由累积方差贡献率和碎石图可以看到，前两个主成分的累计方差贡献率达到96%，另外两个属成分可以舍弃，达到降维的目的。

4. 主成分

$$Z_1 = 0.497X_1^* + 0.515X_2^* + 0.481X_3^* + 0.507X_4^*$$

$$Z_2 = 0.543X_1^* - 0.210X_2^* - 0.725X_3^* + 0.368X_4^*$$

```
Loadings:
    Comp.1 Comp.2 Comp.3 Comp.4
X1   0.497  0.543  0.450  0.506
X2   0.515 -0.210  0.462 -0.691
X3   0.481 -0.725 -0.175  0.461
X4   0.507  0.368 -0.744 -0.232
```

第一主成分：系数都为正，且值都在0.5左右，它反映了中学生身材魁梧程度：身体高大的学生，他的4个部分的尺寸都比较大，第一主成分的值就较大；而身材矮小的学生，他的4部分的尺寸都比较小，第一主成分的值较小。因此，称第一主成分为魁梧因子。

第二主成分：身高和坐高都系数为正，体重和腰围的系数为负，“细高”的学生第二主成分的值大，“矮胖”的学生第二主成分值小。因此，称第二主成分为体型因子。

5. 主成分得分

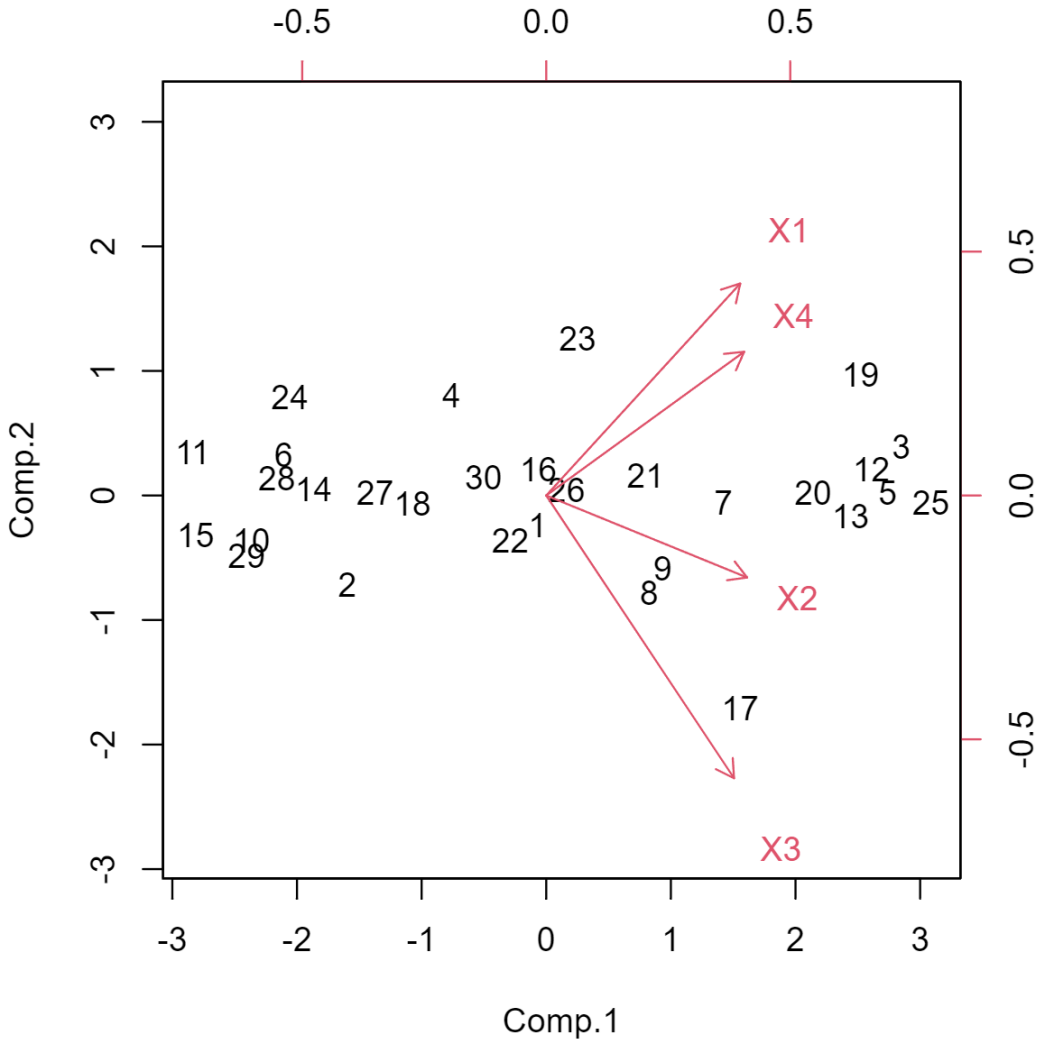
predict(student.pr)

student.pr$scores

	Comp.1	Comp.2	Comp.3	Comp.4
[1,]	-0.06990950	-0.23813701	0.35509248	-0.266120139
[2,]	-1.59526340	-0.71847399	-0.32813232	-0.118056646
[3,]	2.84793151	0.38956679	0.09731731	-0.279482487
[4,]	-0.75996988	0.80604335	0.04945722	-0.162949298
[5,]	2.73966777	0.01718087	-0.36012615	0.358653044
[6,]	-2.10583168	0.32284393	-0.18600422	-0.036456084
[7,]	1.42105591	-0.06053165	-0.21093321	-0.044223092
[8,]	0.82583977	-0.78102576	0.27557798	0.057288572
[9,]	0.93464402	-0.58469242	0.08814136	0.181037746
[10,]	-2.36463820	-0.36532199	-0.08840476	0.045520127

6. 双标图

biplot(student.pr$scores, student.pr$loadings)



综合得分：P227-综合得分及排名。

根据主成分，计算每个样本的一维综合分并排名。

实现方法：可安装mvstats包，调用以下函数

princomp.rank(PC, m=2, plot=T)

（注意：关于此综合得分的计算公式和解释的方法，学界有争议，因此不建议使用。

（虽然有些特殊情况下综合得分勉强可用。）

详见王学民、何晓群(P127-128)等编著的多元统计书籍。）

注意：关于主成分分析的应用

1. 在一些应用中，用前面少数几个主成分，替代众原始变量以作分析，这些主成分本身就成了分析的目标。一般需要给出这前几个主成分一个符合实际背景和意义的解释。

当然，这需要根据实际问题进行分析判断。有时，这些主成分的含义较为明确，容易找到。但有时，主成分的含义未必容易找到，这时需要具体问题具体分析，考虑是否使用其它分析方法，揭示数据蕴含的内在规律性。

注意：关于主成分分析的应用

2. 在更多的另一些应用中，主成分只是要达到目的的一个中间结果或步骤，而非目的本身。

例如：主成分聚类、主成分回归、评估正态性，等等。此时的主成分，也可不必给出解释。

总之，若使用主成分分析方法，应注意实际数据背景，具体问题具体分析，扬长避短。

小结

- 理解主成分分析的思想和方法的原理。
- 掌握主成分分析的基本步骤、注意事项及应用。
- 掌握 R 语言软件对实际数据进行主成分分析的编程实现：学会确定主成分的个数，并通过上机练习对分析结果进行恰当的解释。
- 更多实例，也可参考王学民教授等在中国慕课网发布的MOOC课程。

（致谢：本课件及参考资料的部分内容，综合选自以下课程或教材：
中国人民大学出版社-多元统计分析;
中国人民大学六西格玛质量管理研究中心；
清华大学出版社-实用多元统计分析（第6版译著）；
高等教育出版社-多元统计分析及R语言建模；
北京大学出版社-应用多元统计分析；
上海财经大学出版社-应用多元统计分析，
等书籍或相关的MOOC课材料，我校材料及网络开放资源，
等。）

说明：本课件的涉及资料，仅供学生学习之用，
不得外传无关人员，不得上传网络！
更不得用于牟利！

The end!

【附1】：主成分有时可作为自变量建立线性模型

在考虑因变量Y与p个自变量$X_1, \cdots, X_p$的回归模型中，当自变量间有较强的线性相关（多重共线性）时，利用经典的回归方法求回归系数的最小二乘估计，一般效果较差。利用主成分的性质，可由前m个主成分来建立主成分回归模型：

$$Y = b_0 + b_1 Z_1 + \cdots + b_m Z_m (m \leq p)$$

这样既简化了回归方程的结构，且消除了变量间相关性带来的影响；但另一方面，主成分回归也给回归模型的解释带来一定的复杂性，因为主成分是原始变量的线性组合，不是直接观测的变量，其含义有时不明确。在求得主成分回归方程后，经常又使用逆变换将其变为原始变量的回归方程。

【附2】：利用主成分检查多元正态性

设$D(X)=\Sigma$,如果Σ是对角矩阵，即p维向量的分量间不相关，这时把p元正态性检验问题可转化为p个一元正态性检验问题。但一般Σ不是对角矩阵，即分量间是相关的。利用主成分分析方法，求得X的p个主成分$Z_1, Z_2, \cdots, Z_p$(不相关)，并由原样本值计算p个主成分得分值，作为p个不相关的综合变量的样本值。这时就把p元正态性检验问题化为p个一元综合变量（主成分）的正态性检验。这就是多元正态性检验的主成分检验法。实际检验时，利用主成分的性质，只需对前$m(m<p)$个主成分得分数据逐个做正态性检验。

【附3】：有特殊结构的协方差矩阵的主成分

存在某些模式的协方差和相关矩阵，它们的主成分能够表示成简单的形式。设Σ是对角矩阵

$$\Sigma=\begin{bmatrix}\sigma_{11} & 0 & \cdots & 0\\ 0 & \sigma_{22} & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \sigma_{pp}\end{bmatrix}$$

令$e_i{}'=[0,\ldots,0,1,0,\ldots,0]$，第i个位置上为1，我们看到

$$\begin{bmatrix}\sigma_{11} & 0 & \cdots & 0\\ 0 & \sigma_{22} & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \sigma_{pp}\end{bmatrix}\begin{bmatrix}0\\ \vdots\\ 0\\ 1\\ 0\\ \vdots\\ 0\end{bmatrix}=\begin{bmatrix}0\\ \vdots\\ 0\\ 1\sigma_{ii}\\ 0\\ \vdots\\ 0\end{bmatrix}\quad 或\quad \Sigma e_i=\sigma_{ii}e_i$$

并得到结论：（σ_{ii}，e_i）是第i个特征值-特征向量对。由于线性组合$e_i{}'X=X_i$，此组主成分正是原来那组不相关随机变量。

【附4】：求解齐次线性方程组

定理：n 元齐次线性方程组 $Ax = 0$ 有非零解 $\Leftrightarrow$ $R(A) < n$

例：求解齐次线性方程组

$$\begin{cases} x_1 + x_2 - x_3 - x_4 = 0 \\ 2x_1 - 5x_2 + 3x_3 + 2x_4 = 0 \\ 7x_1 - 7x_2 + 3x_3 + x_4 = 0 \end{cases}$$

提问：为什么只对系数矩阵 A 进行初等行变换变为行最简形矩阵？

答：因为齐次线性方程组 $Ax = 0$ 的常数项都等于零，于是必有 $R(A, 0) = R(A)$，所以可从 $R(A)$ 判断齐次线性方程组的解的情况．齐次线性方程组永远有解，至少有零解。

定义：下列三种变换称为矩阵的初等行变换：

✓对调两行，简称对换变换，记作 $r_i \leftrightarrow r_j$

✓以非零常数 k 乘某一行的所有元素，简称倍乘变换
记作 $r_i \times k$；

✓某一行加上另一行的 k 倍，简称倍加变换，记作 $r_i + kr_j$

.

例：求解齐次线性方程组 $\begin{cases} x_1+x_2-x_3-x_4=0 \\ 2x_1-5x_2+3x_3+2x_4=0 \\ 7x_1-7x_2+3x_3+x_4=0 \end{cases}$

解：对方程组的系数矩阵作初等行变换，得阶梯阵

$$A=\begin{pmatrix} 1 & 1 & -1 & -1 \\ 2 & -5 & 3 & 2 \\ 7 & -7 & 3 & 1 \end{pmatrix} \sim \begin{pmatrix} 1 & 1 & -1 & -1 \\ 0 & -7 & 5 & 4 \\ 0 & -14 & 10 & 8 \end{pmatrix}$$

$$\sim \begin{pmatrix} 1 & 1 & -1 & -1 \\ 0 & -7 & 5 & 4 \\ 0 & 0 & 0 & 0 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & -\frac{2}{7} & -\frac{3}{7} \\ 0 & 1 & -\frac{5}{7} & -\frac{4}{7} \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

行阶梯形矩阵　　行最简形矩阵

$$\begin{pmatrix} 1 & 0 & -\frac{2}{7} & -\frac{3}{7} \\ 0 & 1 & -\frac{5}{7} & -\frac{4}{7} \\ 0 & 0 & 0 & 0 \end{pmatrix} \Rightarrow \begin{cases} x_1 = \frac{2}{7}x_3 + \frac{3}{7}x_4 \\ x_2 = \frac{5}{7}x_3 + \frac{4}{7}x_4 \\ x_3 = x_3 \\ x_4 = x_4 \end{cases}$$

这是原方程组的同解方程组，其中 x_3, x_4 为自由变量。

故原方程的基础解系为：$\mathbf{u}_1 = \begin{pmatrix} \frac{2}{7} \\ \frac{5}{7} \\ 1 \\ 0 \end{pmatrix}$，$\mathbf{u}_2 = \begin{pmatrix} \frac{3}{7} \\ \frac{4}{7} \\ 0 \\ 1 \end{pmatrix}$