信陽師範学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

第12章 MATLAB多元统计分析

讲授人：牛言涛

日期：2020年4月24日

目录

CONTENTS

主成分分析产生的背景

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

在许多领域的研究与应用中，通常需要对含有多个指标（变量）的数据进行观测，收集大量数据后进行分析寻找规律。多指标（变量）大数据集无疑会为研究和应用提供丰富的信息，但是也在一定程度上增加了数据采集的工作量。更重要的是在很多情形下，许多变量之间可能存在相关性，从而增加了问题分析的复杂性。

如果分别对每个指标进行分析，分析往往是孤立的，不能完全利用数据中的信息，因此盲目减少指标会损失很多有用的信息，从而产生错误的结论。

因此需要找到一种合理的方法，在减少需要分析的指标同时，尽量减少原指标包含信息的损失，以达到对所收集数据进行全面分析的目的。由于各变量之间存在一定的相关关系，因此可以考虑将关系紧密的变量变成尽可能少的新变量，使这些新变量是两两不相关的，那么就可以用较少的综合指标分别代表存在于各个变量中的各类信息。主成分分析与因子分析就属于这类降维算法。

主成分分析产生的背景

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

- 数据降维是一种对高维度特征数据预处理方法。
- 降维目的：将高维度的数据保留最重要的一些特征，去除噪声和不重要的特征，从而提升数据处理速度。在实际的生产和应用中，降维在一定的信息损失范围内，可以节省大量的时间和成本。
- 特征降维一般分为两类：特征选择和特征抽取。

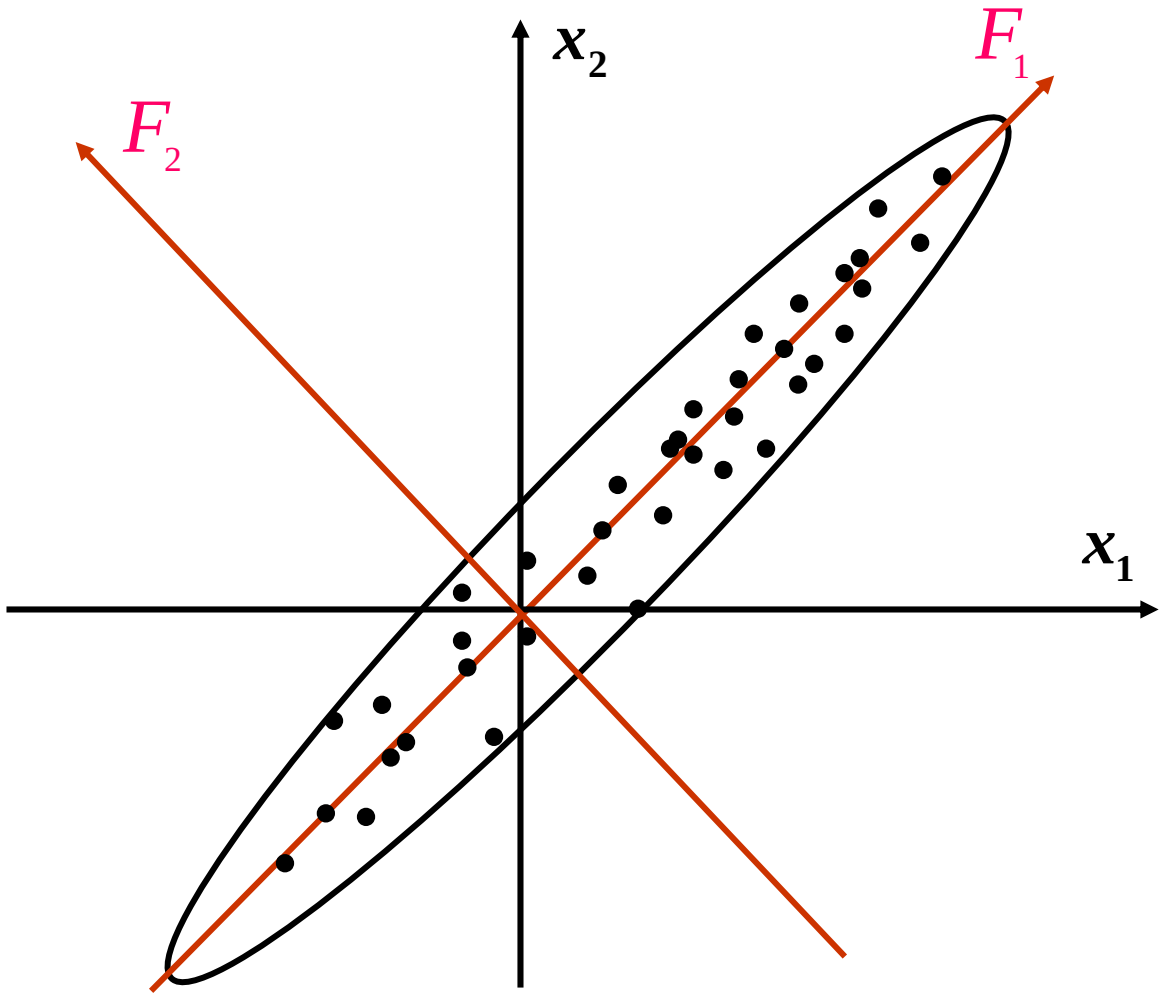
  1、特征选择：就是简单的从高纬度的特征中选择其中一个子集来作为新的特征；

  2、特征抽取：就是将高纬度的特征经过一些函数映射到低纬度，将其作为新的特征。
- 降维具有如下一些优点：

  1）使得数据集更易使用；2) 降低算法的计算开销；3) 去除噪声；4) 使得结果容易理解。
- 降维的算法有很多，比如奇异值分解(SVD)、主成分分析(PCA)、因子分析(FA)、独立成分分析(ICA)。

主成分分析产生的背景

旋转坐标轴



$$F_1 = x_1\cos\theta + x_2\sin\theta$$

$$F_2 = -x_1\sin\theta + x_2\cos\theta$$

$$\begin{bmatrix} F_1 \\ F_2 \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

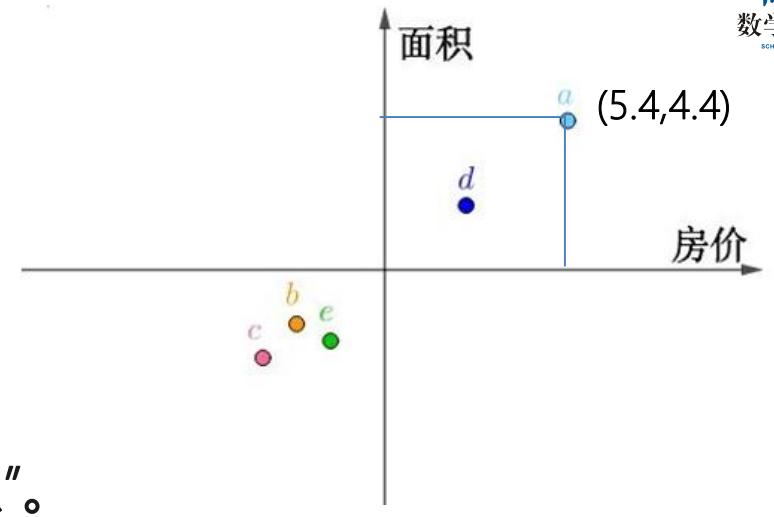
- 旋转变换的目的：使得n个样本点在$F1$轴方向上的离散程度最大，即$F1$的方差最大。变量$F1$代表了原始数据的绝大部分信息，在研究某经济问题时，即使不考虑变量$F2$也损失不多的信息。
- $F1$ 与 $F2$ 除了浓缩作用外，还具有不相关性。
- $F1$称为第一主成分，$F2$称为第二主成分。

1. 主成分分析原理解析

	房价(百万元)	面积(百平米)
a	10	9
b	2	3
c	1	2
d	7	6.5
e	3	2.5

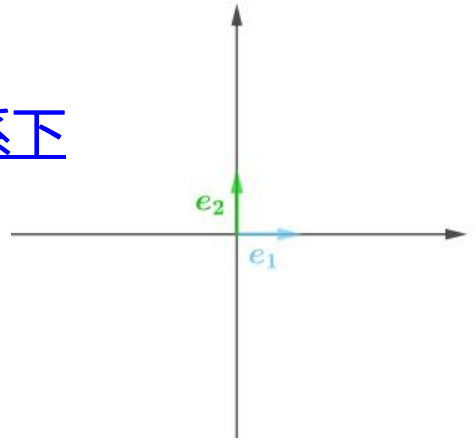
中心化 →

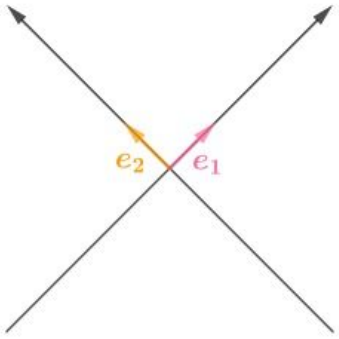
	房价(百万元)	面积(百平米)
a	5.4	4.4
b	−2.6	−1.6
c	−3.6	−2.6
d	2.4	1.9
e	−1.6	−2.1



- 将二维数据绘制在坐标轴上，横纵坐标分别为“房价”、“面积”。
- 如何降维呢？从线性代数的角度来看，二维坐标系总有各自的标准正交基（两两正交、模长为1）$\boldsymbol{e_1}, \boldsymbol{e_2}$.
- 考虑一个点：在某坐标系有点$\boldsymbol{a} = \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix}$，它表示在该坐标系下

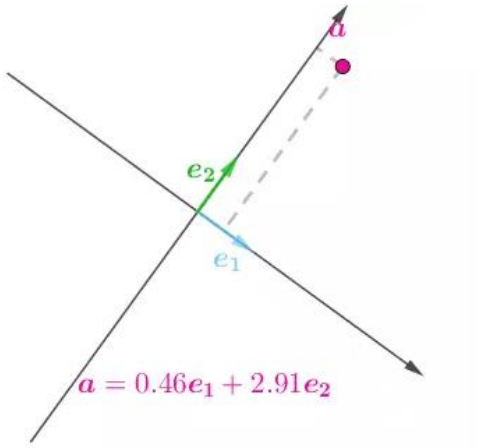
  标准正交基的线性组合$\boldsymbol{a} = \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix} = \boldsymbol{x}\boldsymbol{e_1} + \boldsymbol{y}\boldsymbol{e_2}$.

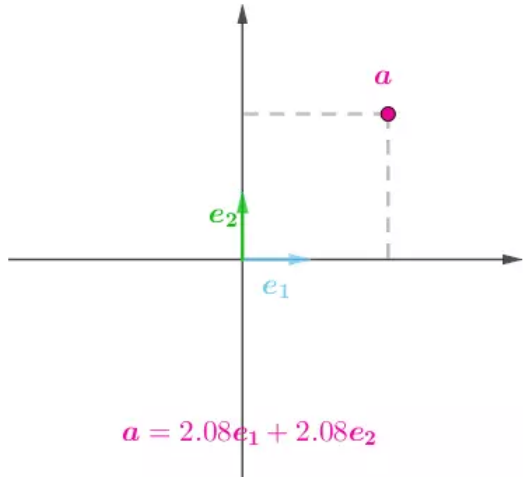


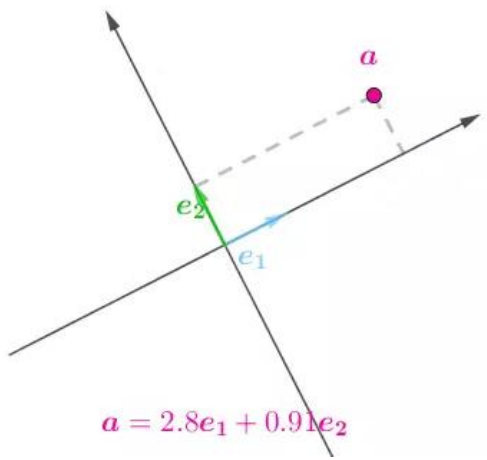


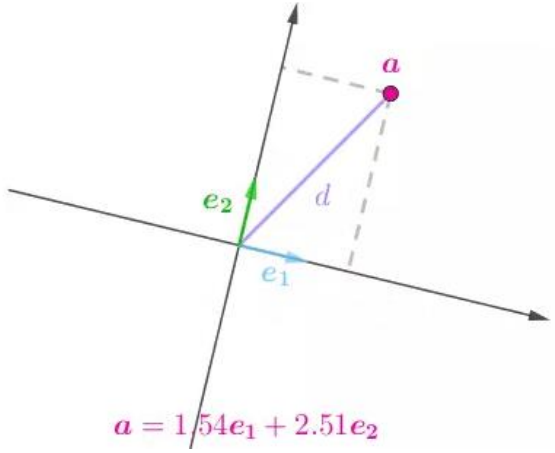
1. 主成分分析原理解析

只是在不同坐标系中，(x, y)的值会有所不同（旋转的坐标系表示不同的坐标系，下图左3个）



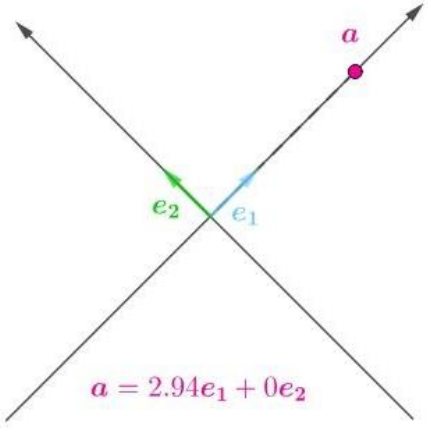






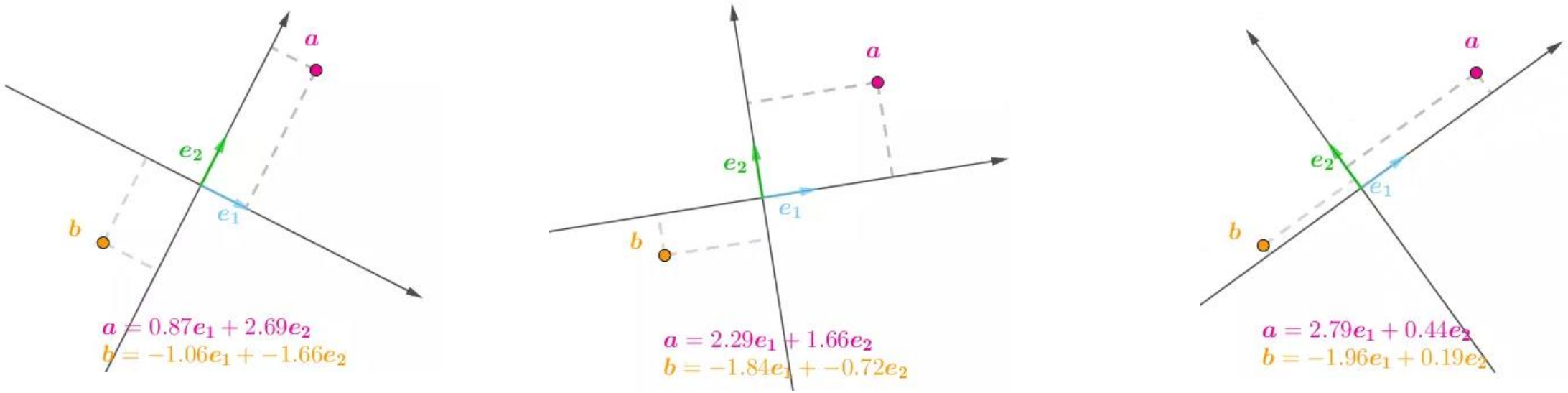
因为a到原点的距离d不会因为坐标系改变而改变，即$d^2 = x^2 + y^2$。

如上图右四所示。所以，在某坐标系下分配给x较多，那么分配给y的就必然比较少，反之亦然。最极端（理想）的情况是，在某个坐标系下，全部分配给了x，使得$y = 0$（右图所示）．那么在这个坐标系中，就可以降维了，去掉e_2并不会损失信息。



1. 主成分分析原理解析

如果有两个点$a=\begin{pmatrix}x_1\\y_1\end{pmatrix}, b=\begin{pmatrix}x_2\\y_2\end{pmatrix}$，则变量$X=(x_1,x_2)$，$Y=(y_1,y_2)$，做类似坐标系统旋转：



- 为了降维，应该选择尽量多分配给x_1, x_2，即变量X，少分配给y_1, y_2的坐标系。
- 如果$X=(x_1,x_2)$，$Y=(y_1,y_2)$表示两个数据指标，这恰是主成分分析的思想，即降维，从二维降到一维，使得X表示更多有贡献的信息，而Y损失的信息最少。

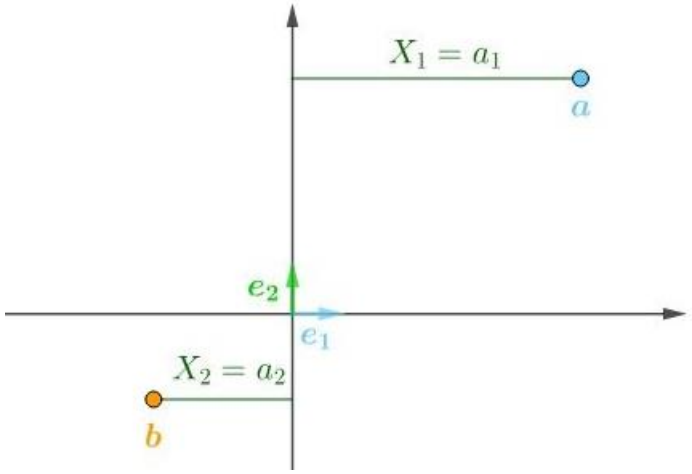
1. 主成分分析原理解析

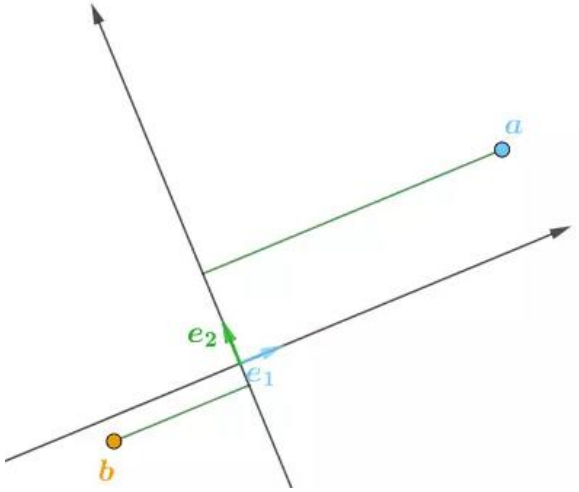
主成分分析：假设有数据如下表

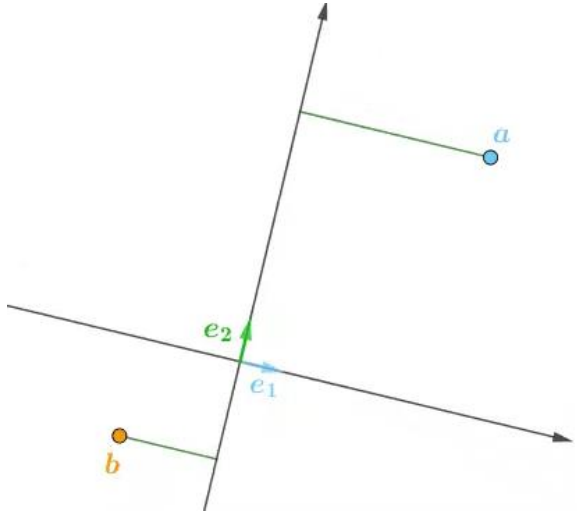
	X	Y
a	a_1	b_1
b	a_2	b_2

，表示有两个点$\boldsymbol{a}=\begin{pmatrix}\boldsymbol{x_1}\\ \boldsymbol{y_1}\end{pmatrix},\boldsymbol{b}=\begin{pmatrix}\boldsymbol{x_2}\\ \boldsymbol{y_2}\end{pmatrix}$，这两个点在初始坐标系下（即自然基$\boldsymbol{e_1}=\begin{pmatrix}\boldsymbol{1}\\ \boldsymbol{0}\end{pmatrix},\boldsymbol{e_2}=\begin{pmatrix}\boldsymbol{0}\\ \boldsymbol{1}\end{pmatrix}$）下坐标值为：$\boldsymbol{a}=\begin{pmatrix}\boldsymbol{x_1}\\ \boldsymbol{y_1}\end{pmatrix}=\begin{pmatrix}\boldsymbol{a_1}\\ \boldsymbol{b_1}\end{pmatrix},\boldsymbol{b}=\begin{pmatrix}\boldsymbol{x_2}\\ \boldsymbol{y_2}\end{pmatrix}=\begin{pmatrix}\boldsymbol{a_2}\\ \boldsymbol{b_2}\end{pmatrix}$，

如下左图；随着坐标系的不同，$\boldsymbol{X_1},\boldsymbol{Y_1}$的值会不断变化，如下右两图。







要想尽量多分配给$\boldsymbol{X_1},\boldsymbol{X_2}$，借鉴最小二乘法的思想，就是让：$\boldsymbol{X_1^2+X_2^2=\sum_{i=0}^{2}X_i^2}$最大。

1. 主成分分析原理解析

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

假设：$e_1=\begin{pmatrix}e_{11}\\e_{12}\end{pmatrix}, e_2=\begin{pmatrix}e_{21}\\e_{22}\end{pmatrix}$，根据点积的几何意义（一个向量在另一个向量方向上的投影）有：

$$X_1=a\cdot e_1=\begin{pmatrix}a_1\\b_1\end{pmatrix}\cdot\begin{pmatrix}e_{11}\\e_{12}\end{pmatrix}=a_1e_{11}+b_1e_{12},\quad X_2=b\cdot e_1=\begin{pmatrix}a_2\\b_2\end{pmatrix}\cdot\begin{pmatrix}e_{11}\\e_{12}\end{pmatrix}=a_2e_{11}+b_2e_{12}$$

则 $X_1^2+X_2^2=\left(a_1e_{11}+b_1e_{12}\right)^2+\left(a_2e_{11}+b_2e_{12}\right)^2=\left(a_1^2+a_2^2\right)e_{11}^2+2\left(a_1b_1+a_2b_2\right)e_{11}e_{12}+\left(b_1^2+b_2^2\right)e_{12}^2$

上式其实是一个二次型 $X_1^2+X_2^2=e_1^T\underbrace{\begin{pmatrix}a_1^2+a_2^2 & a_1b_1+a_2b_2\\ a_1b_1+a_2b_2 & b_1^2+b_2^2\end{pmatrix}}_{P}e_1=e_1^TPe_1$

这里矩阵P就是二次型，是一个对称矩阵，可以进行如下的奇异值分解$P=U\Sigma U^T$，其中U为正交矩阵，即$UU^T=I$，而Σ是对角矩阵：$\Sigma=\begin{pmatrix}\sigma_1 & 0\\ 0 & \sigma_2\end{pmatrix}$，其中$\sigma_1,\sigma_2$是奇异值，$\sigma_1>\sigma_2$.

1. 主成分分析原理解析

将P代回去，$X_1^2+X_1^2=e_1^TPe_1=e_1^TU\Sigma U^Te_1=\left(U^Te_1\right)^T\Sigma\left(U^Te_1\right)=n^T\Sigma n$，所得的$n$也是单位向量，即

$n=\begin{pmatrix}n_1\\n_2\end{pmatrix}\Rightarrow n_1^2+n_2^2=1$，继续回代 $X_1^2+X_1^2=n^T\Sigma n=\begin{pmatrix}n_1 & n_2\end{pmatrix}\begin{pmatrix}\sigma_1 & 0\\0 & \sigma_2\end{pmatrix}\begin{pmatrix}n_1\\n_2\end{pmatrix}=\sigma_1n_1^2+\sigma_2n_2^2$

最初求最大值的问题就转化为了：

$$X_1^2+X_2^2=\max\left(\sum_{i=0}^{2}X_i^2\right)\Leftrightarrow\begin{cases}\max\left(\sigma_1n_1^2+\sigma_2n_2^2\right)\\n_1^2+n_2^2=1\\\sigma_1>\sigma_2\end{cases}$$

结果是当$n_1=1,n_2=0$时取到极值。

因此可以推出要寻找的主元1，即：$\mathrm{n}=\begin{pmatrix}1\\0\end{pmatrix}=U^Te_1\Rightarrow e_1=U\begin{pmatrix}1\\0\end{pmatrix}$。总结如下：

$e_1=\begin{cases}P=U\Sigma U^T\\\text{最大奇异值}\sigma_1\text{ 对应的奇异向量}\end{cases}$，$e_2=\begin{cases}P=U\Sigma U^T\\\text{最小奇异值}\sigma_2\text{ 对应的奇异向量}\end{cases}$

1. 主成分分析原理解析

已知 $X=\begin{pmatrix}a_1\\a_2\end{pmatrix}$，$Y=\begin{pmatrix}b_1\\b_2\end{pmatrix}$，$P=\begin{pmatrix}a_1^2+a_2^2 & a_1b_1+a_2b_2\\ a_1b_1+a_2b_2 & b_1^2+b_2^2\end{pmatrix}=\begin{pmatrix}X\cdot X & X\cdot Y\\ X\cdot Y & Y\cdot Y\end{pmatrix}$

	X	Y
a	a_1	b_1
b	a_2	b_2

“中心化”后的样本方差和样本协方差：

$$Var(X)=\frac{1}{n}\sum_{i=1}^{n}X_i^2=\frac{1}{n}X\cdot X,\quad Cov(X,Y)=\frac{1}{n}\sum_{i=1}^{n}X_iY_i=\frac{1}{n}X\cdot Y$$

比较可得新矩阵，即协方差矩阵，且P、Q都可以进行奇异值分解：

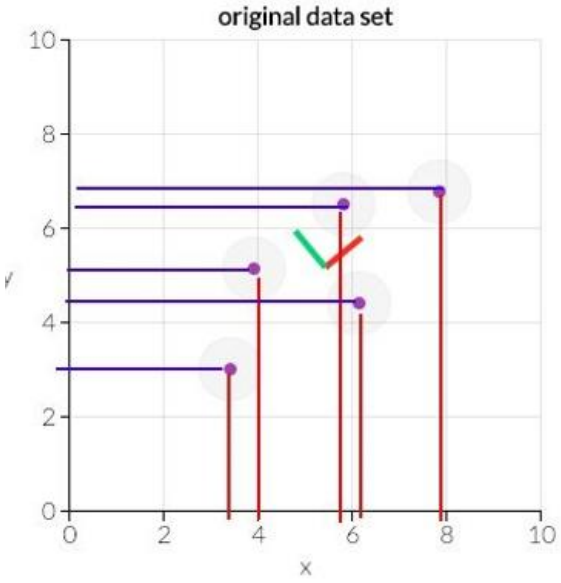
$$Q=\frac{1}{n}P=\begin{pmatrix}Var(X) & Cov(X,Y)\\ Cov(X,Y) & Var(Y)\end{pmatrix},\quad P=U\begin{pmatrix}\sigma_1 & 0\\ 0 & \sigma_2\end{pmatrix}U^T,\quad Q=\frac{1}{n}P=U\begin{pmatrix}\frac{\sigma_1}{n} & 0\\ 0 & \frac{\sigma_2}{n}\end{pmatrix}U^T$$

可见，协方差矩阵Q的奇异值分解和P相差无几，只是奇异值缩小了n倍，但不妨碍奇异值之间的大小关系，所以实际问题中，往往都是直接分解协方差矩阵Q。

1. 主成分分析原理解析

$$Q=\frac{1}{n}P=\begin{pmatrix} Var(X) & Cov(X,Y) \\ Cov(X,Y) & Var(Y) \end{pmatrix} \xrightarrow{\text{标准化后}} \begin{pmatrix} \sigma_X^2 & \sigma(X,Y) \\ \sigma(X,Y) & \sigma_Y^2 \end{pmatrix} \xrightarrow{\text{新的维度pc1和pc2}} \begin{pmatrix} \sigma_{pc1}^2 & 0 \\ 0 & \sigma_{pc2}^2 \end{pmatrix}$$

- 数据的全部信息包含在协方差矩阵所描述的全部变异中。
- σ_X^2代表数据点在维度x（图x轴）上投影点的分散程度，它是通过所有点到数据中心点（即x的平均值）的平均平方和计算得到的；σ_Y^2代表数据点在维度y（图y轴）上投影点的分散程度；方差越大，数据越分散，也就意味着信息量越多，该特征越有区分度。
- 协方差代表维度x和维度y之间的相关程度，协方差越大，也就意味着噪声越大，信息的冗余程度越高。
- 主成分分析的目的就是要最小化噪声，最大化提取出数据中包含的信息。它可以找到新的维度，从而剔除不同维度之间的冗余信息，即让不同维度之间的相关为0。因此，主成分分析在各个维度之间相关性很高时尤为有用。
- 对角线上的方差是新维度上的方差，非对角线代表维度之间的相关，它等于0意味着主成分分析得到的新维度完全剔除了特征之间的冗余信息。



$$\begin{pmatrix} \sigma_{pc1}^2 & 0 \\ 0 & \sigma_{pc2}^2 \end{pmatrix}$$

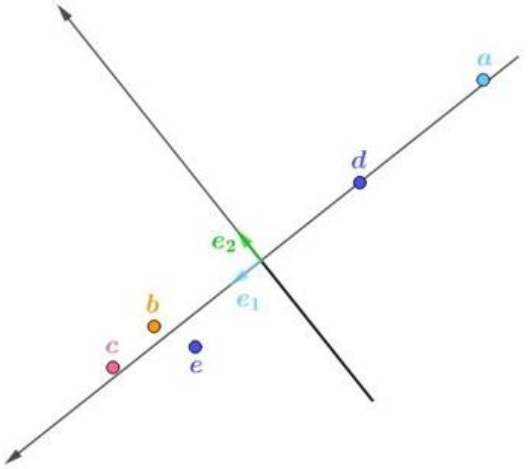
1. 主成分分析原理解析

由前例中心化后的两个向量 $X=(5.4,-2.6,-3.6,2.4,-1.6)^T, Y=(4.4,-1.6,-2.6,1.9,-2.1)^T$ ，

组成协方差矩阵： $Q=\begin{pmatrix} Var(X) & Cov(X,Y) \\ Cov(X,Y) & Var(Y) \end{pmatrix}=\frac{1}{5}\begin{pmatrix} X\cdot X & X\cdot Y \\ X\cdot Y & Y\cdot Y \end{pmatrix}=\frac{1}{5}\begin{pmatrix} 57.2 & 45.2 \\ 45.2 & 36.7 \end{pmatrix}$

进行奇异值分解： $Q\approx\begin{pmatrix} -0.78 & -0.62 \\ -0.62 & 0.78 \end{pmatrix}\begin{pmatrix} 18.66 & 0 \\ 0 & 0.12 \end{pmatrix}\begin{pmatrix} -0.78 & -0.62 \\ -0.62 & 0.78 \end{pmatrix}$

根据之前的分析，主元1应该匹配最大奇异值对应的奇异向量，主元2匹配最小奇异值对应的奇异向量，即： $e_1=\begin{pmatrix} -0.78 \\ -0.62 \end{pmatrix}, e_2=\begin{pmatrix} -0.62 \\ 0.78 \end{pmatrix}.$



如下算出新坐标，比如对于a： $X_1=a\cdot e_1=-6.94$， $X_2=a\cdot e_2=0.084$.

以此类推，得到心得数据表：

	主元1	主元2
a	−6.94	0.084
b	3.02	0.364
c	4.42	0.204
d	−3.05	−0.006
e	2.55	−0.646

主元2整体来看，数值很小，丢掉的损失信息也非常小，这样就实现了非理想情况下的降维。

2. 主成分分析数学模型

设$X_1, X_2, \ldots X_p$，为实际问题的p个n维随机变量(p项指标)记$X = (X_1, X_2, \ldots, X_p)^T$，其协方差矩阵为

$$\Sigma = (\sigma_{ij})_p = E[(X - E(X))(X - E(X))^T]$$

它是一个p阶的非负定矩阵。设变量$x_1, x_2, \ldots x_p$经过线性变换后得到新的综合变量$Y_1, Y_2, \ldots, Y_p$，即

$$\begin{cases} Y_1 = l_{11}x_1 + l_{12}x_2 + \cdots + l_{1p}x_p \\ Y_2 = l_{21}x_1 + l_{22}x_2 + \cdots + l_{2p}x_p \\ \cdots \\ Y_p = l_{p1}x_1 + l_{p2}x_2 + \cdots + l_{pp}x_p \end{cases}$$

，或 $Y_i = l_{i1}X_1 + l_{i2}X_2 + \cdots + l_{ip}X_p$, $i = 1,2,\cdots, p$

其中系数$l_i = (l_{i1}, l_{i2}, \cdots, l_{ip}), i = 1, 2, \cdots, p$ 为常数向量。要求满足以下条件：

(1) 系数向量是单位向量，即 $l_{i1}^2 + l_{i2}^2 + \cdots + l_{ip}^2 = 1$, $i = 1,2,\cdots, p$

(2) 不同的主成分不相关，即 $\mathrm{cov}(Y_i, Y_j) = 0$, $(i \neq j, i, j = 1,2,\cdots, p)$

(3) 各主成分的方差递减，即 $\mathrm{var}(Y_1) \geq \mathrm{var}(Y_2) \geq \cdots \geq \mathrm{var}(Y_p) \geq 0$

于是，称Y_1为第一主成分，Y_2为第二主成分，依此类推，Y_p称为第p个主成分。主成分又叫主分量。这里l_{ij}称为主成分的系数。

2. 主成分分析数学模型

当总体$X=(X_1,X_2,\cdots,X_p)^T$的协方差矩阵$\sum=(\sigma_{ij})_p$已知时，可根据下面的定理求出主成分。

定理1 设p维随机向量X的协方差矩阵Σ的特征值满足$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p\geq 0$，相应的单位正交特征向量为$e_1,e_2,\ldots,e_p$，则$X$的第$i$个主成分为

$$Y_i=e_i^TX=e_{i1}X_1+e_{i2}X_2+\cdots+e_{ip}X_p\quad(i=1,2,\cdots,p)$$

其中 $e_i=(e_{i1},e_{i2},\cdots,e_{ip})^T$，且
$$\begin{cases}Var(Y_k)=e_k^T\sum e_k=\lambda_k,\ (k=1,2,\cdots,p)\\ \mathrm{cov}(Y_k,Y_j)=e_k^T\sum e_j=0,\ (k\neq j,k,j=1,2,\cdots,p)\end{cases}$$

定理表明：求X的主成分等价于求它的协方差矩阵的所有特征值及相应的正交单位化特征向量。

推论：若记$Y=(Y_1,Y_2,\cdots,Y_p)^T$为主成分向量，矩阵$p=(e_1,e_2,\ldots,e_p)$，则$Y=p^TX$，且Y的协方差

$\Sigma_Y=P^T\Sigma P=\Lambda=Diag(\lambda_1,\lambda_2,\cdots,\lambda_p)$，主成分的总方差$\sum_{i=1}^{p}Var(Y_i)=\sum_{i=1}^{p}Var(X_i)$。

2. 主成分分析数学模型

此性质表明主成分分析是将p个原始变量的总方差分解为p个不相关变量$Y_1, Y_2, \cdots, Y_p$的方差之和。

由于$Var(Y_k) = \lambda_k$，因此 $\lambda_k \Big/ \sum_{k=1}^{p} \lambda_k$ 描述了第k个主成分提取的信息占总信息的份额。

称 $\lambda_k \Big/ \sum_{k=1}^{p} \lambda_k$ 为第k个主成分的贡献率，表示第k个主成分提取的信息占总信息的百分比。

称前m个主成分的贡献率之和 $\sum_{k=1}^{m} \lambda_k \Big/ \sum_{k=1}^{p} \lambda_k$ 为累计贡献率，它表示前m个主成分综合提供总信息的程度。通常$m < p$，且累计贡献率达到85%以上.

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。

2. 主成分分析数学模型

主成分是原来变量的线性组合，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面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。

定理2：设$\boldsymbol{Y}=(\boldsymbol{Y_1},\boldsymbol{Y_2},\cdots,\boldsymbol{Y_p})^{\boldsymbol{T}}$为总体$\boldsymbol{X}=(\boldsymbol{X_1},\boldsymbol{X_2},\cdots,\boldsymbol{X_p})^{\boldsymbol{T}}$的主成分向量，则主成分$\boldsymbol{Y_i}$与变量$\boldsymbol{X_j}$的相关系数

$$\rho_{Y_iX_j}=\frac{\sqrt{\lambda_i}}{\sqrt{\sigma_{jj}}}e_{ij}(i,j=1,2,\cdots,p).\quad \Sigma_{YX}=\left(\rho_{Y_iX_j}\right)_p=\left(\sqrt{diag(\Sigma)}\right)^{-1}P\sqrt{\Lambda}$$

称$\boldsymbol{Y_i}$与$\boldsymbol{X_j}$的相关系数为因子载荷量，由于因子载荷量与主成分的系数向量成正比，与标准差成反比，因此因子载荷量的绝对值大小刻画了该主成分的成因，可以解释第$\boldsymbol{j}$个变量对第$\boldsymbol{i}$个主成分的重要程度。 $diag(\Sigma)$ 表示协方差矩阵的主对角线元素组成的对角矩阵。Λ是特征值对角矩阵。

3. 标准化变量的主成分

实际问题经常遇到不同的指标具有不同的量纲，有时会导致各指标取值的分散程度较大，这样在计算协方差矩阵时，可能出现总体的方差主要受方差较大的数据的控制，可能造成不合理的结果。为消除量纲的影响，对原始数据进行标准化。即令

$$X_i^* = \frac{X_i - \mu_i}{\sqrt{\sigma_{ii}}} \quad (i = 1, 2, \cdots, p)$$

其中$\mu_i = EX_i$，X^*的协方差矩阵就是原始数据的相关系数矩阵。计算标准化变量的主成分公式为：

$$Y_i^* = (e_i^*)^T X^* = e_{i1}^* X_1^* + e_{i2}^* X_2^* + \cdots + e_{ip}^* X_p^* \quad (i = 1, 2, \cdots, p)$$

3. 标准化变量的主成分

性质1：总体方差和等于向量的维数 $\sum_{i=1}^{p} Var(Y_i^*) = \sum_{i=1}^{p} Var(X_i^*) = \sum_{i=1}^{p} \lambda_i^* = p$

其中λ_i^*是相关系数矩阵的特征值。

性质2：标准化变量的第i个主成分的贡献率 $\lambda_i^* / p,\ i = 1, 2, \cdots, p$

标准化变量的前m个主成分的累积贡献率 $\sum_{i=1}^{m} \lambda_i^* / p$

性质3：主成分Y_i^*与标准化数据X_j^*的相关系数 $\rho(Y_i^*, X_j^*) = \sqrt{\lambda_i^*} e_{ij}^*$

值得注意的是同一个总体，分别从协方差矩阵和相关系数矩阵出发进行主成份分析，所得的主成份的贡献率可以不同。

4. 样本主成分分析

实际问题中，总体的协方差矩阵Σ一般是未知的，具有的资料只是来自于X的一个容量为n的样本观测数据。设$x_i=(x_{i1},x_{i2},\cdots,x_{ip})^T, i=1,2,\cdots,n$ 为取自总体的一个容量为n的简单随机样本，可知样本协方差矩阵及样本相关矩阵分别为：

$$S=(s_{ij})_{p\times p}=\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x})(x_k-\bar{x})^T,\quad R=(r_{ij})_{p\times p}=\frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}$$

其中 $\bar{x}=(\bar{x}_1,\bar{x}_2,\cdots,\bar{x}_p)^T$ $\bar{x}_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij}$; $s_{jk}=\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k)$ $j,k=1,2,\cdots,p$

分别以S和R作为总体Σ和ρ的估计，然后按总体主成分分析的方法作样本主成分分析。关于样本主成份，有如下结论：

设$S_{p\times p}$为样本协方差矩阵，其特征值$\hat{\lambda}_1\geq\hat{\lambda}_2\geq\cdots\geq\hat{\lambda}_p\geq 0$，相应的单位正交化特征向量$\hat{e}_1,\hat{e}_2,\cdots,\hat{e}_p$，第$k$个样本主成份表为 $y_k=\hat{e}_k^T x=\hat{e}_{k1}x_1+\hat{e}_{k2}x_2+\cdots+\hat{e}_{kp}x_p$。当依次代入观测值$x_i$时，便得到第$k$个样本主成分的$n$个观测值 $y_{1k},y_{2k},\cdots,y_{nk}$ 称其为第k个样本主成份的得分。

4. 样本主成分分析

同总体主成份分析一样，为了消除量纲的影响，可对样本进行标准化，即令

$$x_i^* = \left(\frac{x_{i1}-\bar{x}_1}{\sqrt{s_{11}}}, \frac{x_{i2}-\bar{x}_2}{\sqrt{s_{22}}}, \cdots, \frac{x_{ip}-\bar{x}_p}{\sqrt{s_{pp}}}\right)^T (i = 1, 2, \cdots, n)$$

由于标准化的样本数据的协方差矩阵也即原始数据的样本相关系数矩阵，由样本相关系数矩阵出发作主成份分析即可。

- 由样本观测数据矩阵进行主成分分析的步骤为：
  1. 对原始数据进行标准化处理；X = zscore(data)
  2. 计算样本相关系数矩阵； R = corrcoef(X)
  3. 求相关系数矩阵的特征值和相应的特征向量；[coeff,D] = eig(R)
  4. 选择重要的主成分，并写出主成分表达式； F = coeff*(x1,..,xp)′;
  5. 计算主成分得分；score = X*coeff
  6. 依据主成分得分的数据，进一步从事统计分析.

5. 主成分分析MATLAB代码实现

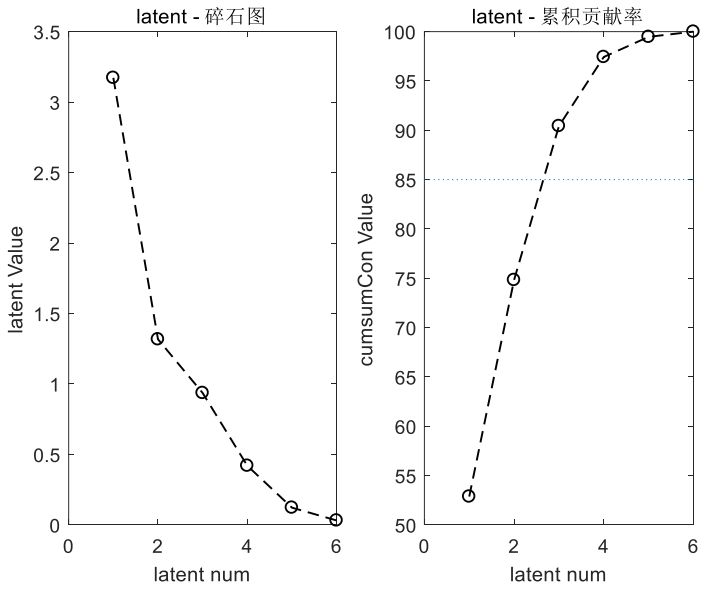
信阳师范学院 数学与统计学院 SCHOOL OF MATHEMATICS AND STATISTICS

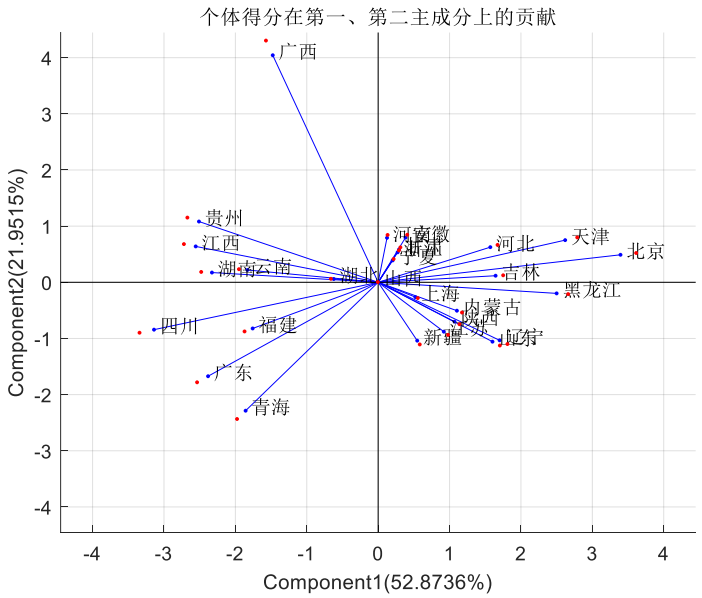
PCA_demo.m

```
function pca = PCA_demo(X,varnames)
% PCA_demo用于实现主成分分析，输入参数X为样本数据，一行代表一个样本
% varnames表示样本的行名称，可以是字符元胞数组或汉字组成的元胞数组
% 输出参数pca为结构体，包含载荷系数、个体得分、特征值、贡献率和累积贡献率

    %% 要求输入两个参数，如果不足，提示警告并返回
    if nargin < 2
        warning('输入参数数目不足！调用格式：pca = PCA_demo(X,varnames)')
        pca = [];
        return
    end

    %% 1、标准化矩阵
    X = zscore(X); %标准化矩阵X

    %% 2、计算相关系数矩阵
    R = corrcoef(X);

    %% 3、第三步：计算特征向量和特征值
    %计算矩阵R的特征向量矩阵coeff和特征值矩阵D
    [coeff,D] = eig(R); % 特征值由小到大
    % 特征值矩阵由大到小排序diag(sort(diag(D),'descend'))
    D = rot90(rot90(D));
    latent = diag(D);  % 特征值矩阵转换为特征值向量，取对角线元素
    coeff = fliplr(coeff);  % 特征向量矩阵coeff对应latent从大到小排序

    %% 4、计算贡献率和累计贡献率
    explained = latent/sum(latent)*100;
    cumsumCon = cumsum(explained);

    %% 5、计算得分
    score = X * coeff;

    %% 6、可视化
    subplot(1,2,1)
    plot(latent,'ko--','LineWidth',1) %从大到小特征值
    xlabel('latent num'); ylabel('latent Value')
    title('latent - 碎石图')
    subplot(1,2,2)
    plot(cumsumCon,'ko--','LineWidth',1) %从大到小特征值
    hl = refline(0,85); hl.LineStyle = ':';
    xlabel('latent num'); ylabel('cumsumCon Value')
    title('latent - 累积贡献率')
    % 个体得分在第一、第二主成分上的贡献
    figure
    biplot(score(:,[1,2]),'Score',score(:,[1,2]),'VarLabels',varnames)
    title('个体得分在第一、第二主成分上的贡献')
    ytext = strcat('Component2(',num2str(explained(2)),'%)');
    xtext = strcat('Component1(',num2str(explained(1)),'%)');
    xlabel(xtext); ylabel(ytext);

    %% 7、输出参数组合
    pca.coeff = coeff;
    pca.score = score;
    pca.latent = latent;
    pca.explained = explained;
    pca.cumsumCon = cumsumCon;
end
```

5. 主成分分析MATLAB代码实现

```
>> [data,text] = xlsread('bodyzb.xlsx');
>> varnames = text(2:end,1);
>> pca = PCA_demo(data,varnames)
pca =
  包含以下字段的 struct:
        coeff: [6×6 double]
        score: [28×6 double]
       latent: [6×1 double]
    explained: [6×1 double]
    cumsumCon: [6×1 double]
```






```
>> pca.coeff
ans =
    0.5224   -0.1951    0.1906    0.2547   -0.2159    0.7357
    0.5255   -0.0811    0.1665    0.3890   -0.3120   -0.6640
    0.5111   -0.1810    0.1046   -0.3363    0.7563   -0.0996
    0.3465   -0.0463   -0.7410   -0.4563   -0.3469   -0.0102
    0.1884    0.6567    0.4714   -0.4963   -0.2540    0.0156
    0.1850    0.6994   -0.3921    0.4650    0.3148    0.0871
```

调用程序，输出coeff为载荷系数矩阵，score为个体得分矩阵，latent对应各主成分特征值，explained为各主成分贡献率，cumsumCon为累积贡献率。同时绘制两个图。

6. 主成分分析的MATLAB函数

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

pcacov函数用来根据协方差矩阵或相关系数矩阵进行主成分分析，调用格式如下：

- [COEFF,latent,explained]=pcacov(V)
  - ✓ 输入参数V是总体或样本的协方差矩阵或相关系数矩阵，对于p维总体，V是p*p的矩阵。
  - ✓ COEFF是p个主成分的系数矩阵，它是p*p的矩阵，它的第i列是第i个主成分的系数向量。
  - ✓ latent是p个主成分的方差构成的向量，即V的p个特征值的大小(从大到小)构成的向量。
  - ✓ explained是p个主成分的贡献率向量，已经转化为百分比。

6. 主成分分析的MATLAB函数

例1：波特兰水泥数据，其ingredients为成分百分比：3CaO.Al2O3（铝酸三钙），3CaO.SiO2（硅酸三钙），4CaO.Al2O3.Fe2O3（铁酸四铝），2CaO.SiO2（β-硅酸二钙）

```
>> load hald
>> covx = cov(ingredients);  %计算协方差矩阵
>> [COEFF,latent,explained] = pcacov(covx)
COEFF =  %主成份变换矩阵
   -0.0678   -0.6460    0.5673    0.5062
   -0.6785   -0.0200   -0.5440    0.4933
    0.0290    0.7553    0.4036    0.5156
    0.7309   -0.1085   -0.4684    0.4844
latent =  %主成份方差向量
  517.7969  67.4964  12.4054  0.2372
explained =  %各主成份贡献率向量
   86.5974  11.2882  2.0747    0.0397
```

前两个主成份的累计贡献率为：86.5974%+11.2882%=97.8856%，因此，若用前两个主成分代替原来四个变量，其信息损失为2.1144%，很小。若要求累积贡献率大于等于85%，则可用一个主成分代替原来四个变量。

$$Y_1 = -0.0678X_1 - 0.6785X_2 + 0.0290X_3 + 0.7309X_4$$

$$Y_2 = -0.6460X_1 - 0.0200X_2 + 0.7553X_3 - 0.1085X_4$$

6. 主成分分析的MATLAB函数

```
>> S = diag(diag(covx)) %协方差矩阵的主对角线元组成的对角矩阵
>> SYX = inv(sqrt(S))*COEFF*sqrt(diag(latent))
SYX =
   -0.2623   -0.9023    0.3397    0.0419
   -0.9922   -0.0106   -0.1231    0.0154
    0.1031    0.9688    0.2219    0.0392
    0.9936   -0.0532   -0.0986    0.0141
```

$$\rho_{Y_i X_j} = \frac{\sqrt{\lambda_i}}{\sqrt{\sigma_{jj}}} e_{ij}(i, j = 1, 2, \cdots, p).$$

$$\Sigma_{YX} = \left(\rho_{Y_i X_j}\right)_p = \left(\sqrt{diag(\Sigma)}\right)^{-1} P\sqrt{\Lambda}$$

所以，SYX的第一列元素依次为Y_1与X_1,X_2,X_3,X_4的相关系数，即其余各列的元素类推。结果表明Y_1与X_2,X_4高度相关，Y_2与X_1,X_3高度相关等。

7. 样本主成分分析

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

原始数据的主成分分析

`coeff = pca(X)` 返回 n×p 数据矩阵 X 的主成分系数，也称为载荷。X 的行对应于观测值，列对应于变量。系数矩阵是 p×p 矩阵。coeff 的每列包含一个主成分的系数，并且这些列按成分方差的降序排列。默认情况下，pca 将数据中心化，并使用奇异值分解 (SVD) 算法。

`coeff = pca(X,Name,Value)` 使用由一个或多个 Name,Value 对组参数指定的用于计算和处理特殊数据类型的附加选项，返回上述语法中的任何输出参数。

例如，您可以指定 pca 返回的主成分数或使用 SVD 以外的其他算法。

`[coeff,score,latent] = pca( ___ )` 还在 score 中返回主成分分数，在 latent 中返回主成分方差。您可以使用上述语法中的任何输入参数。

主成分分数是 X 在主成分空间中的表示。score 的行对应于观测值，列对应于成分。

主成分方差是 X 的协方差矩阵的特征值。

`[coeff,score,latent,tsquared] = pca( ___ )` 还返回 X 中每个观测值的 Hotelling T 方统计量。

`[coeff,score,latent,tsquared,explained,mu] = pca( ___ )` 还返回 explained（即每个主成分解释的总方差的百分比）和 mu（即 X 中每个变量的估计均值）。

指定可选的、以逗号分隔的 Name,Value 对组参数。Name 为参数名称，Value 为对应的值。Name 必须放在引号内。示例： 'Algorithm','eig','Centered',false,'Rows','all','NumComponents',3 指定 pca 使用特征值分解算法，不将数据中心化，使用所有观测值，并仅返回前三个主成分。

7. 样本主成分分析

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

值	说明
'svd'	默认值。X 的奇异值分解 (SVD)。
'eig'	协方差矩阵的特征值分解 (EIG)。当观测值数目 n 超过变量的数目 p 时，EIG 算法比 SVD 更快，但不太准确，因为协方差的条件数是 X 的条件数的平方。
'als'	交替最小二乘 (ALS) 算法。此算法通过将 X 分解为 n×k 左因子矩阵 L 和 p×k 右因子矩阵 R 来计算最佳秩 k 逼近，其中 k 是主成分的数量。分解使用从随机初始值开始的迭代方法。 ALS 能够更好地处理缺失值。它倾向于采用成对删除 ('Rows','pairwise')，而不是采用整行删除 ('Rows','complete') 处理缺失值。它可以很好地处理随机缺失少量数据的数据集，但对于稀疏数据集可能表现不佳。

'Options' - 迭代的选项

结构体

迭代的选项，指定为以逗号分隔的对组，其中包含 'Options' 和由 statset 函数创建的结构体。pca 在 options 结构体中使用以下字段。

字段名称	说明
'Display'	显示输出的级别。选项包括 'off'、'final' 和 'iter'。
'MaxIter'	允许的最大步数。默认值为 1000。与优化设置不同，达到 MaxIter 值即视为收敛。
'TolFun'	用来指定代价函数的终止容差的正数。默认值为 1e-6。
'TolX'	正数，用来指定 ALS 算法中左因子矩阵和右因子矩阵的元素中相对变化的收敛阈值。默认值为 1e-6。

样本主成分分析——案例分析1

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

例1：各省份男子身材指标数据，进行主成分分析。

地区	身高x1	坐高x2	体重x3	胸围x4	肩宽x5	骨盆宽x6	地区	身高x1	坐高x2	体重x3	胸围x4	肩宽x5	骨盆宽x6
北京	173.28	93.62	60.1	86.72	38.97	27.51	江苏	171.36	92.53	58.39	87.09	38.23	27.04
天津	172.09	92.83	60.38	87.39	38.62	27.82	浙江	171.24	92.61	57.69	83.98	39.04	27.07
河北	171.46	92.73	59.74	85.59	38.83	27.46	安徽	170.49	92.03	57.56	87.18	38.54	27.57
山西	170.08	92.25	58.04	85.92	38.33	27.29	河南	170.43	92.38	57.87	84.87	38.78	27.37
内蒙古	170.61	92.36	59.67	87.46	38.38	27.14	青海	170.27	91.94	56	84.52	37.16	26.81
辽宁	171.69	92.85	59.44	87.45	38.19	27.1	福建	169.43	91.67	57.22	83.87	38.41	26.6
吉林	171.46	92.93	58.7	87.06	38.58	27.36	江西	168.57	91.4	55.96	83.02	38.74	26.97
黑龙江	171.6	93.28	59.75	88.03	38.68	27.22	湖北	169.88	91.89	56.87	86.34	38.37	27.19
山东	171.6	92.26	60.5	87.63	38.79	26.63	湖南	167.94	90.91	55.97	86.77	38.17	27.16
陕西	171.16	92.62	58.72	87.11	38.19	27.18	广东	168.82	91.3	56.07	85.87	37.61	26.67
甘肃	170.04	92.17	56.95	88.08	38.24	27.65	广西	168.02	91.26	55.28	85.63	39.66	28.07
宁夏	170.61	92.5	57.34	85.61	38.52	27.36	四川	167.87	90.96	55.79	84.92	38.2	26.53
新疆	171.39	92.44	58.92	85.37	38.83	26.47	贵州	168.15	91.5	54.56	84.81	38.44	27.38
上海	171.83	92.79	56.85	85.35	38.58	27.03	云南	168.99	91.52	55.11	86.23	38.3	27.14

样本主成分分析——案例分析1

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

```
[data,text] = xlsread('bodyzb.xlsx');
varnames = text(1,2:end);
rownames = text(2:end,1);
sddata = zscore(data); %中心化
[coeff,score,latent,~,explained] = pca(sddata);
% 计算贡献率
n = size(data,2);
resultl = cell(n+1,4);
resultl(1,:) = {'特征值','差值','贡献率','累积贡献率'};
resultl(2:end,1) = num2cell(latent);
resultl(2:end-1,2) = num2cell(-diff(latent));
resultl(2:end,3:4) = num2cell([explained,cumsum(explained)]);
% 取前2个主成分
result2 = cell(n+1,4);
result2(1,:) = {'变量','第一主成分','第二主成分','第三主成分'};
result2(2:end,1) = varnames;
result2(2:end,2:end) = num2cell(coeff(:,1:3));
```

```
resultl =
 7×4 cell 数组
  '特征值'      '差值'       '贡献率'       '累积贡献率'
    [3.1724]    [1.8553]    [52.8736]    [  52.8736]
    [1.3171]    [0.3809]    [21.9515]    [  74.8251]
    [0.9362]    [0.5162]    [15.6040]    [  90.4291]
    [0.4201]    [0.2976]    [ 7.0011]    [  97.4302]
    [0.1225]    [0.0907]    [ 2.0411]    [  99.4713]
    [0.0317]          []    [ 0.5287]    [ 100.0000]
```

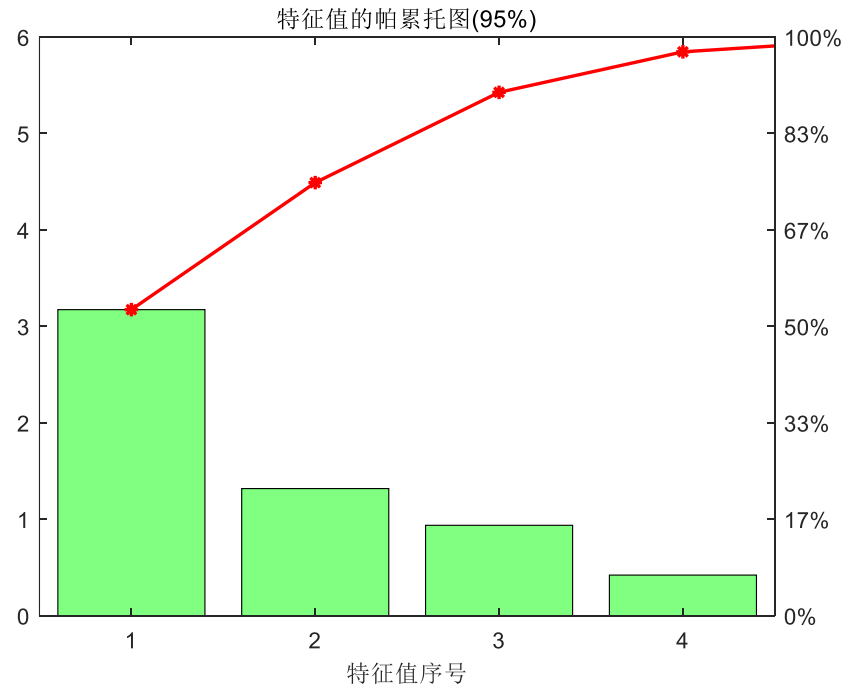
样本主成分分析——案例分析1

result2 =

7×4 cell 数组

'变量'	'第一主成分'	'第二主成分'	'第三主成分'
'身高x1'	[0.5224]	[-0.1951]	[-0.1906]
'坐高x2'	[0.5255]	[-0.0811]	[-0.1665]
'体重x3'	[0.5111]	[-0.1810]	[-0.1046]
'胸围x4'	[0.3465]	[-0.0463]	[0.7410]
'肩宽x5'	[0.1884]	[0.6567]	[-0.4714]
'骨盆宽x6'	[0.1850]	[0.6994]	[0.3921]

```
>> h = pareto(latent)
>> h(1).FaceColor = 'g';
>> h(2).Color = 'r';
>> h(2).LineWidth = 1.5;
>> h(2).Marker = '*';
>> alpha(0.5)
```



主成分与变量的关系，第一主成分对各个变量解释得都很充分，而第二、三主成分对各变量的解释并非全部充分。

$$Y_1 = 0.5224x_1 + 0.5255x_2 + 0.5111x_3 + 0.3465x_4 + 0.1884x_5 + 0.1850x_5$$

$$Y_2 = -0.1951x_1 - 0.0811x_2 - 0.1810x_3 - 0.0463x_4 + 0.6567x_5 + 0.6994x_5$$

$$Y_2 = -0.1906x_1 - 0.1665x_2 - 0.1046x_3 + 0.7410x_4 - 0.4714x_5 + 0.3921x_5$$

样本主成分分析——案例分析1

$$Y_1 = 0.5224x_1 + 0.5255x_2 + 0.5111x_3 + 0.3465x_4 + 0.1884x_5 + 0.1850x_5$$

$$Y_2 = -0.1951x_1 - 0.0811x_2 - 0.1810x_3 - 0.0463x_4 + 0.6567x_5 + 0.6994x_5$$

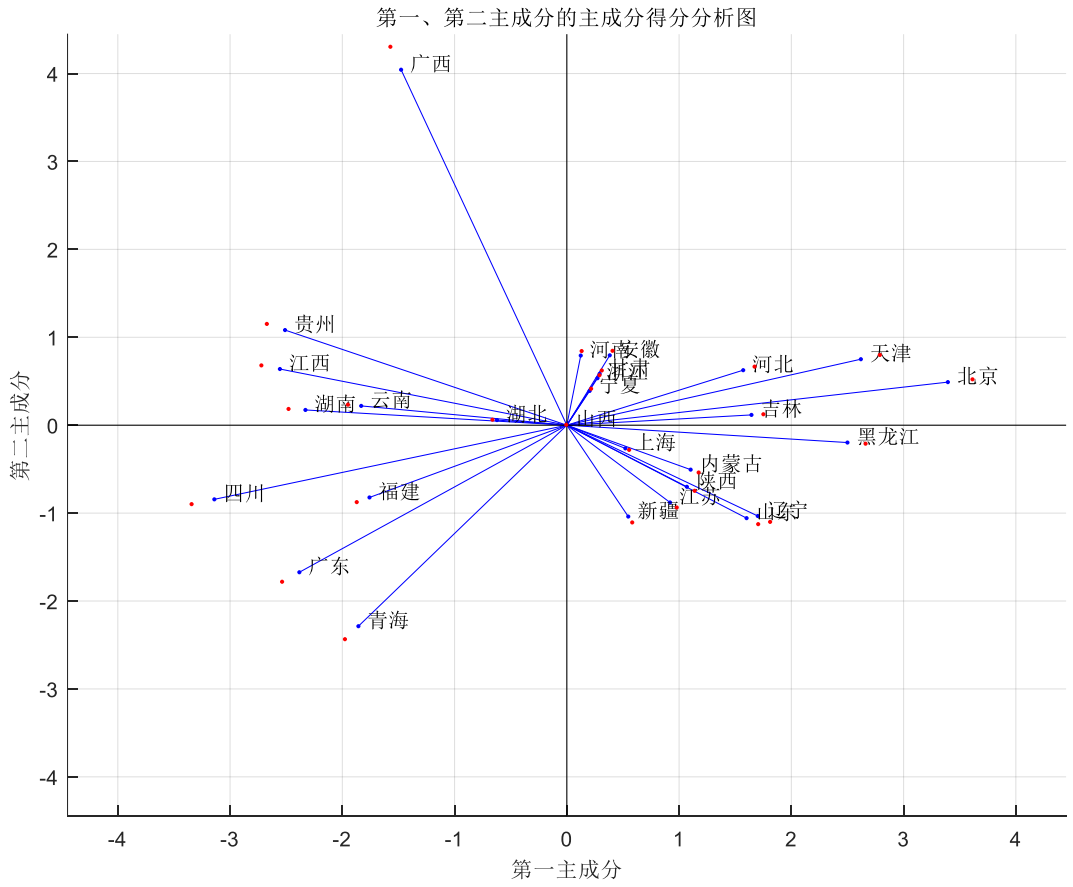
$$Y_2 = -0.1906x_1 - 0.1665x_2 - 0.1046x_3 + 0.7410x_4 - 0.4714x_5 + 0.3921x_5$$

- 每一个选中的主成分所代表的特征可以用来给这些成分起名字，但并不一定都合理。
- 第一主成分Y1与各变量均呈现出正相关，说明一个人身材的综合指标问题，试想一个人又高又胖时，Y1值较大，或者一个人又小又瘦，Y1值较小。此外，Y1与身高X1、做高X2和体重X3有较强的正相关，与肩宽X5、骨盆宽X6呈现较弱的正相关，此类身材更像是H型，比较匀称。
- 第二主成分Y2与身高X1、做高X2、体重X3和胸围X4均呈现较弱的负相关，与肩宽X5和骨盆宽X6呈现较强的正相关。试想一个人又矮又胖，则前四个变量值相对较小，而肩宽和骨盆宽较大，则值Y2较大，试想一个人又高又瘦的I型，则前三个变量值相对较大，而肩宽和骨盆宽较小，则Y2值较小，故Y2可有肩宽和骨盆宽来解释，如大骨架型身材，或者高矮与胖瘦的协调成分。
- 第三主成分Y3与身高X1、做高X2、体重X3和肩宽X5呈现负相关，与胸围X4和骨盆宽X6呈现正相关，且与胸围X4呈现较强的正相关，与肩宽X5呈现稍强的负相关，故Y3可以解释为与胸围相关，身材在厚度上的因素，如身材比较单薄。

样本主成分分析——案例分析1

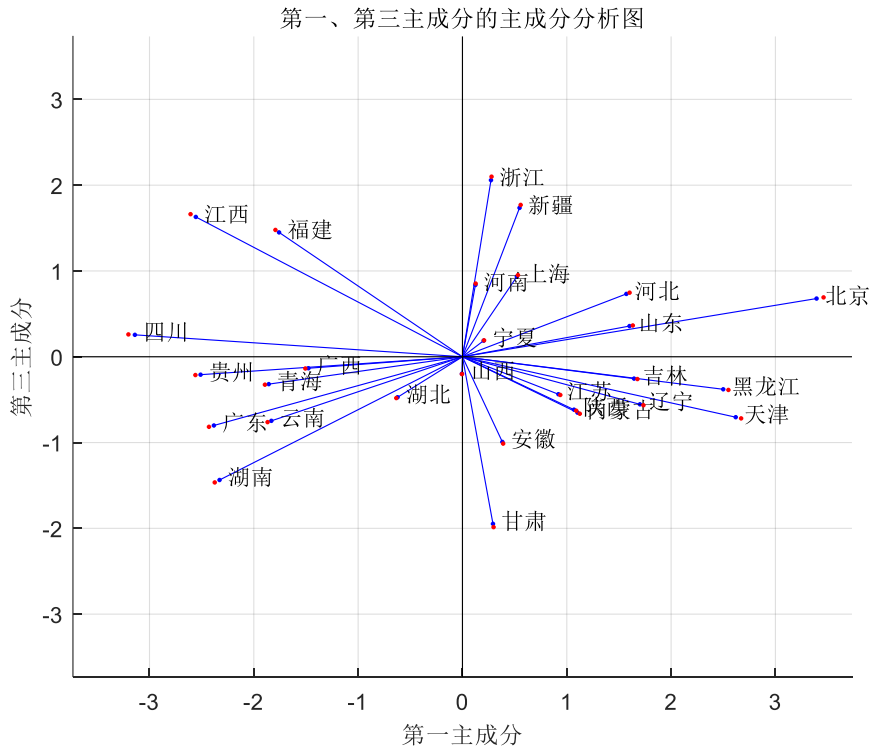
- 在身体综合指标和身体协调度这两个方面，各个省份还是有区别的。总体上说广西是比较特殊，有着较大的肩宽和骨盆宽，且在身高上不占优势。
- 圆心位置附近：上海、山西、宁夏、浙江、甘肃、河南、安徽、湖北等地两个维度绝对值不大，处于两个维度的均衡状态。
- 北京、天津、四川、江西、湖南等地在身体综合指标方面比较突出，身材相对来说较好。具体来说，北京和天津在身高、做高、体重等方面比较突出，且身材比例较为协调，而四川、江西和湖南等地身高、坐高和体重方面不占优势，但是身材比例也较为协调。
- 广西、青海等地在高矮与胖瘦的协调成分中比较突出。观看实际数据，发现广西人肩宽和骨盆宽较其他省份较大，且身高不占优势，而青海肩宽和骨盆宽较其他省份较小。
- 黑龙江、吉林等地，在身高上并不特别明显，相反，湖南、云南等地身高上不占优势，但都没有较宽的肩宽和骨盆宽。

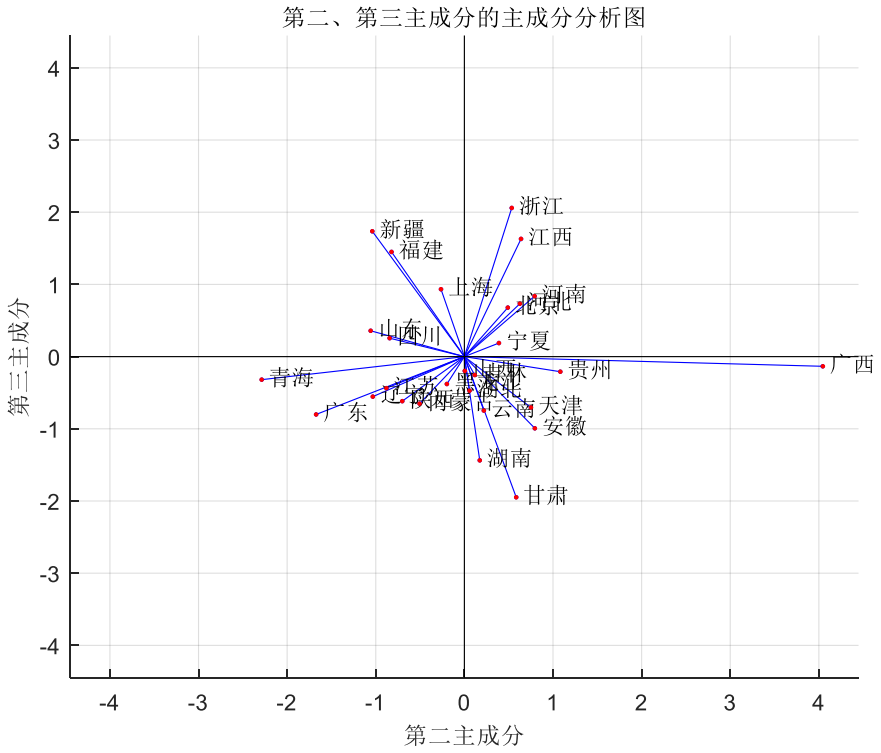
```
biplot(score(:,1:2),'Score',score(:,1:2),'VarLabels',rownames)
```



样本主成分分析——案例分析1

```
biplot(score(:,[1,3]),'Score',score(:,[1,3]),'VarLabels',rownames)
biplot(score(:,2:3),'Score',score(:,2:3),'VarLabels',rownames)
```

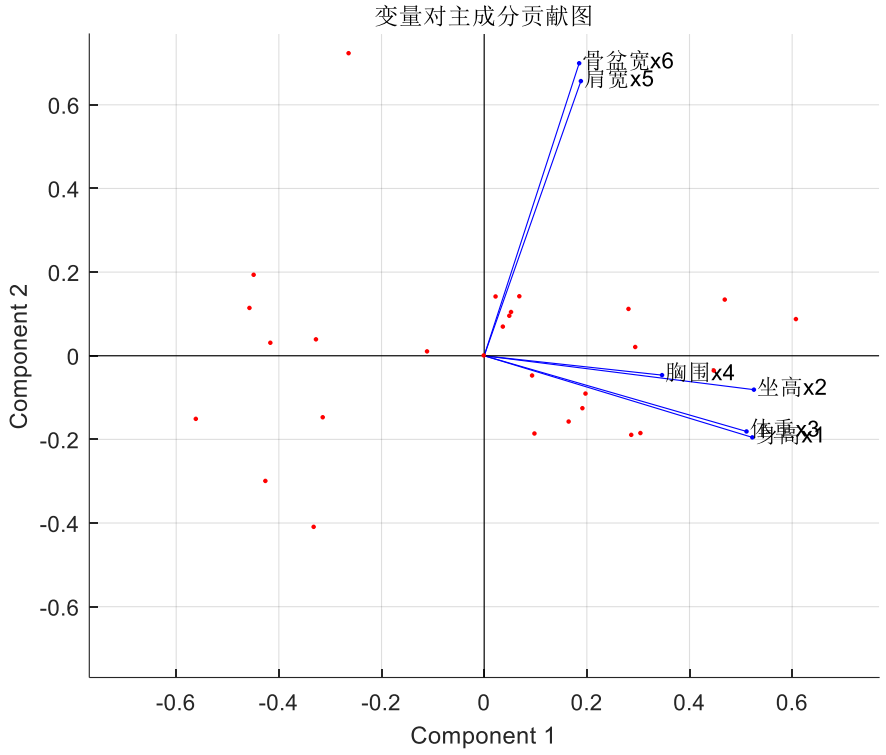


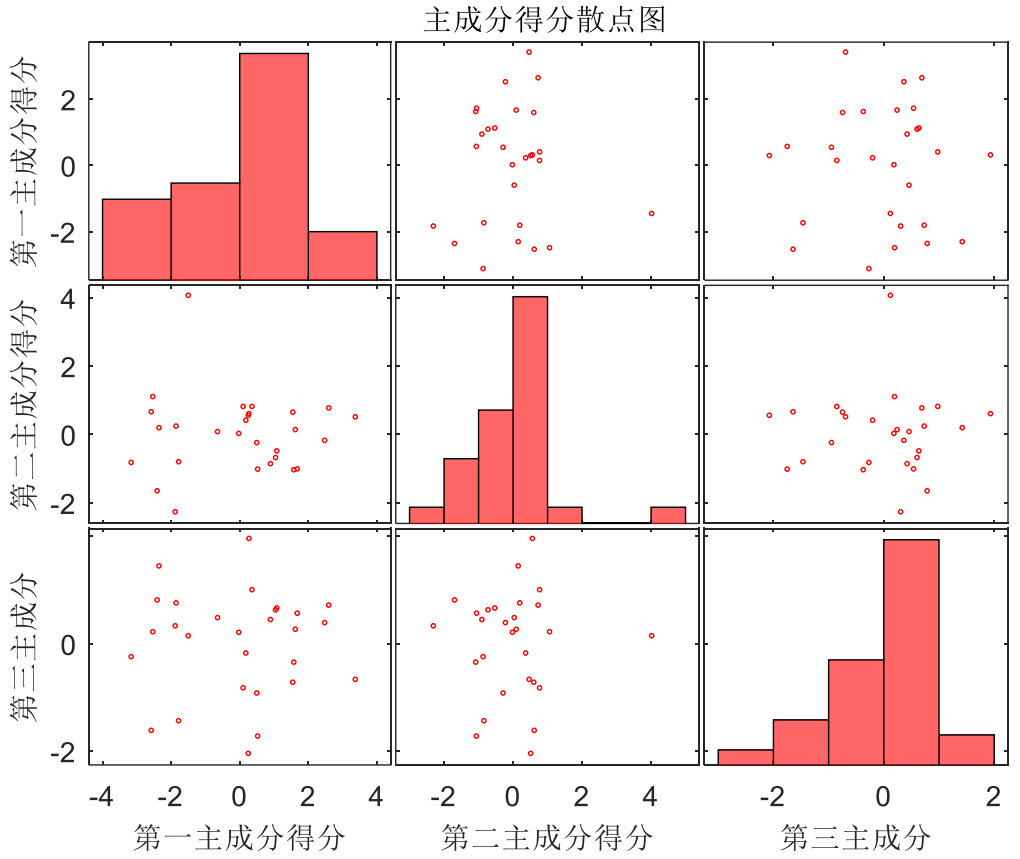


在第三主成分上，浙江和新疆表现的比较突出，胸围较小，而甘肃正好相反。

样本主成分分析——案例分析1

```
biplot(coeff(:,1:2),'Score',score(:,1:2),'VarLabels',varnames)
name = {'第一主成分得分','第二主成分得分','第三主成分'};
gplotmatrix(score(:,1:3),[],[],'r','o',1.5,'off',[],name,name)
title('主成分得分散点图')
```





从图中看出，身高、坐高、体重和胸围对第一主成分贡献较大，而肩宽与骨盆宽对第二主成分贡献较多。

样本主成分分析——案例分析1

以各个主成分的贡献率为权重，由主成分得分和对应权重线性加权求和可得到综合评价模型：

$$f = \left(52.874Y_1 + 21.952Y_2 + 15.604Y_3 + 7.0019Y_4 + 2.041Y_5 + 0.529Y_6\right)/100$$

```
>> F = data*coeff;
>> OA_evaluation = F*explained/100;
>> [OAE,ind] = sort(OA_evaluation,'descend');  %按照综合评价得分降序排列;
>> res = cell(length(OAE),2);
>> res(:,1) = num2cell(OAE);
>> res(:,2) = rownames(ind);
res =
  28×2 cell 数组
    [108.5903]    '北京'
    [108.5140]    '天津'
    [108.3931]    '黑龙江'
    [108.1091]    '山东'
    [107.9326]    '辽宁'
    [107.7491]    '吉林'
    [107.7467]    '内蒙古'
    [107.5903]    '河北'
    [107.5236]    '陕西'
    [107.4212]    '江苏'
    [107.2292]    '甘肃'
    [107.1940]    '安徽'
    [106.9601]    '新疆'
    [106.7654]    '山西'
    [106.6968]    '宁夏'
    [106.6920]    '上海'
    [106.5966]    '河南'
    [106.4728]    '湖北'
    [106.4334]    '浙江'
    [105.8901]    '广西'
    [105.7792]    '湖南'
    [105.7276]    '云南'
    [105.5097]    '广东'
    [105.4990]    '福建'
    [105.4138]    '青海'
    [105.0808]    '贵州'
    [104.9760]    '四川'
    [104.8738]    '江西'
```

样本主成分分析——案例分析2

例2：主要城市废气中主要污染物排放情况 (2013年)，进行主成分分析并进行综合评价

```
[data,text] = xlsread('Z0816C-2013Orig.xls');
d = zscore(data);
[coeff,score,latent,~,explained] = pca(d);
varnames = text(4,2:end); %变量名称
rownames = text(7:end,1); %31个主要城市
% 计算贡献率
n = size(data,2);
resultl = cell(n+1,4);
resultl(1,:) = {'特征值','差值','贡献率','累积贡献率'};
resultl(2:end,1) = num2cell(latent);
resultl(2:end-1,2) = num2cell(-diff(latent));
resultl(2:end,3:4) = num2cell([explained,cumsum(explained)]);
```

```
>> resultl
resultl =
  7×4 cell 数组
    '特征值'     '差值'      '贡献率'      '累积贡献率'
    [3.4040]    [1.6854]    [56.7327]    [  56.7327]
    [1.7185]    [1.3367]    [28.6421]    [  85.3748]
    [0.3818]    [0.0509]    [ 6.3631]    [  91.7379]
    [0.3309]    [0.2248]    [ 5.5153]    [  97.2532]
    [0.1061]    [0.0475]    [ 1.7691]    [  99.0223]
    [0.0587]          []    [ 0.9777]    [      100]
```

前三个主成分累积贡献率达到91.74%，故取前三个主成分。

样本主成分分析——案例分析2

```
result2 = cell(n+1,4);
result2(1,:) = {'变量','第一主成分','第二主成分','第三主成分'};
result2(2:end,1) = varnames;
result2(2:end,2:end) = num2cell(coeff(:,1:3));
>> result2
result2 =
  7×4 cell 数组
    '变量'          '第一主成分'    '第二主成分'    '第三主成分'
    '工业二氧化硫'    [  0.4260]    [ -0.4309]    [ -0.1712]
    '工业氮氧化物'    [  0.4292]    [ -0.3180]    [  0.6343]
    '工业烟(粉)尘'    [  0.4475]    [ -0.3222]    [ -0.4219]
    '生活二氧化硫'    [  0.4532]    [  0.2691]    [ -0.2280]
    '生活氮氧化物'    [  0.3788]    [  0.4719]    [  0.4884]
    '生活烟尘'        [  0.2917]    [  0.5606]    [ -0.3160]
```

$$Y_1 = 0.4260x_1 + 0.4292x_2 + 0.4475x_3 + 0.4532x_4 + 0.3788x_5 + 0.2917x_6$$

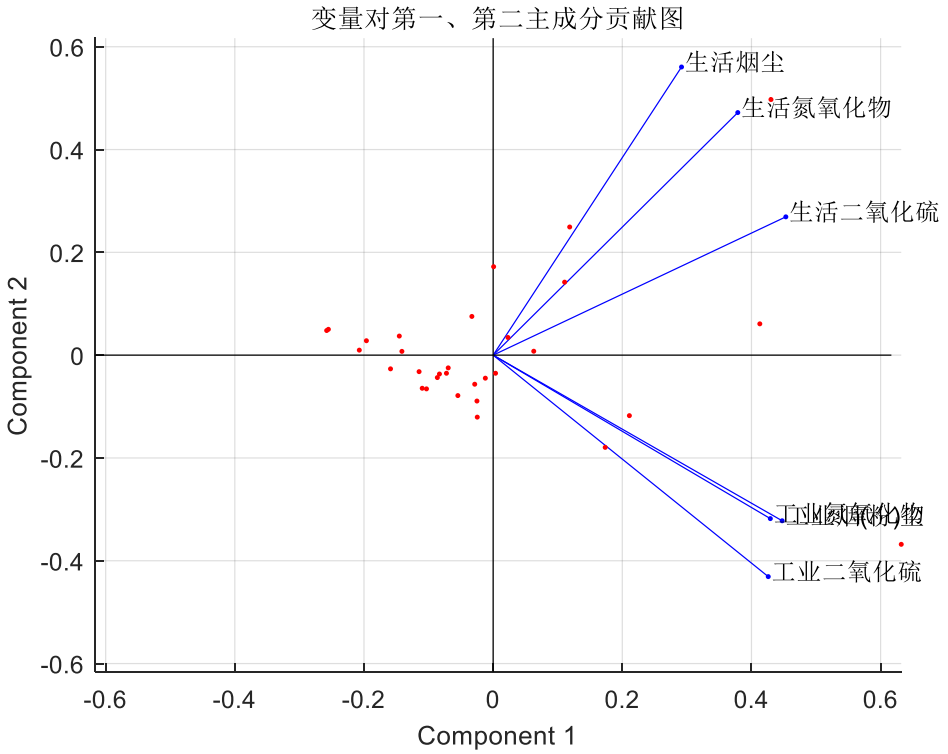
$$Y_2 = -0.4309x_1 - 0.318x_2 - 0.3222x_3 + 0.2691x_4 + 0.4719x_5 + 0.5606x_6$$

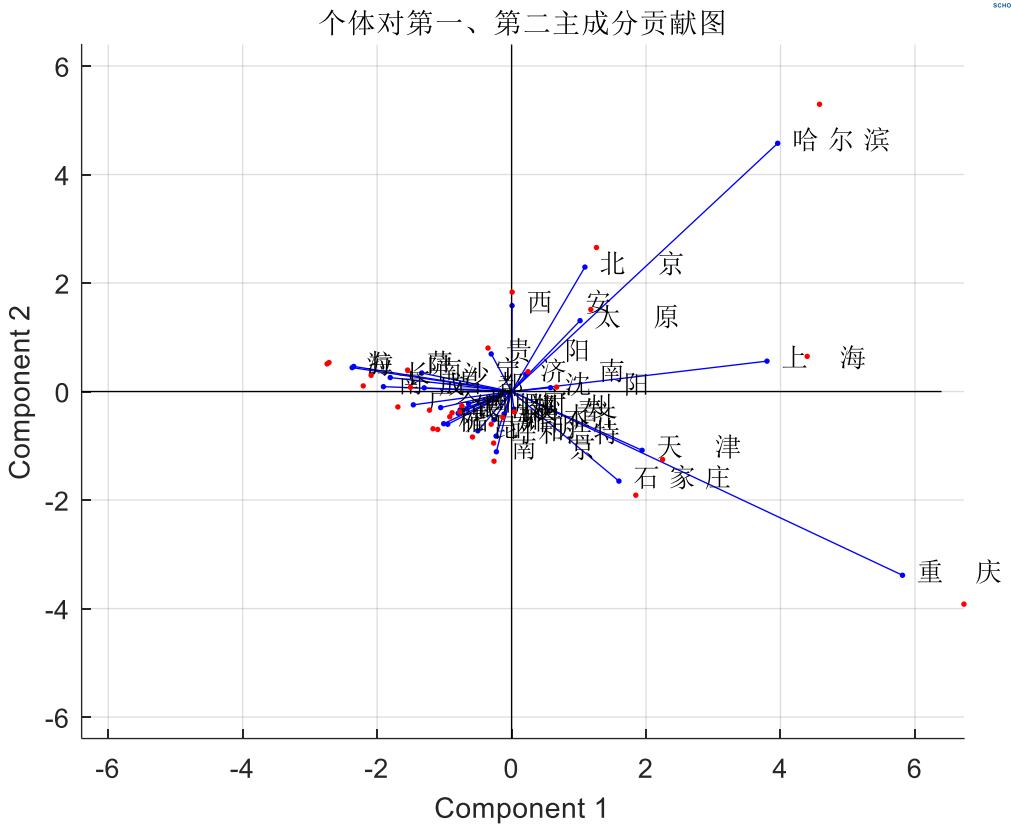
$$Y_2 = -0.1712x_1 + 0.6343x_2 - 0.4219x_3 - 0.228x_4 + 0.4884x_5 - 0.316x_6$$

第一主成分各成分载荷系数比较均衡，可视为综合指标评价，即污染与工业、生活污染均有关系。第二主成分可见工业污染、生活污染的载荷系数负正相反，有所区分，可从相关系数分析出发，工业污染之间相关性较大，生活污染之间相关性也较大，但它们之间的相关系数较小。第三主成分可见工业氮氧化物和生活氮氧化物载荷系数较大，故第三主成分代表氮氧化合物污染排放指标。

样本主成分分析——案例分析2

```
>> biplot(coeff(:,1:2),'Score',score(:,1:2),'VarLabels',varnames)
>> title('变量对第一、第二主成分贡献图')
>> biplot(score(:,1:2),'Score',score(:,1:2),'VarLabels',rownames)
>> title('个体对第一、第二主成分贡献图')
```





污染比较严重的前三个城市依次为重庆、哈尔滨和上海。

样本主成分分析——案例分析3

例3：在对某湖泊水质进行环境监测时，设15个监测点，每个监测点监测指标为5项，用主成分分析法确定最佳的检测布设点。其中DO溶解氧、COD化学需氧量、BOD生化需氧量、T-N为总氮，T-P为总磷。

点位	DO	COD	BOD	T-N	T-P	点位	DO	COD	BOD	T-N	T-P
1	4.3	4.74	4.23	3.66	0.105	9	6.2	4.24	2.33	0.71	0.068
2	5.9	4.61	2.59	2.92	0.081	10	7.4	3.99	2.84	0.74	0.063
3	7	3.94	2.92	1.71	0.072	11	8.1	4.43	3.44	0.86	0.07
4	6.9	3.92	3.11	1.32	0.075	12	7.7	4.31	3.5	0.93	0.074
5	7.4	4.02	3.1	1.26	0.076	13	5.7	4.88	5.02	1.84	0.134
6	6.9	3.75	3.15	1.05	0.096	14	6.8	4.73	4.34	1.39	0.109
7	6.7	4.44	3.14	1.02	0.072	15	5.5	5.93	5.06	2.81	0.24
8	6.8	4.35	4.08	1.27	0.11						

样本主成分分析——案例分析3

```
>> [wq,text] = xlsread('water_quality.xlsx');
>> wqzs = zscore(wq);
>> [coeff,score,latent,~,explained] = pca(wqzs);
>> coeff   %载荷系数矩阵
coeff =
   -0.4180    0.5645   -0.1879    0.6664    0.1649
    0.4836    0.2255   -0.5549    0.1115   -0.6285
    0.4336    0.4508    0.7538    0.1387   -0.1455
    0.4230   -0.5528    0.0089    0.6768    0.2395
    0.4736    0.3489   -0.2974   -0.2572    0.7066
>> latent = latent'  %特征值
latent =
    3.5195    0.9347    0.2503    0.1686    0.1268
```

```
>> contribution = cumsum(explained)  %累积贡献率
contribution =
   70.3898
   89.0844
   94.0905
   97.4630
  100.0000
```

从结果来看，第一个主成分的贡献率就达到了70.3898%，前二个主成分的累积贡献率达到了89.0844%，所以只用前两个主成分就可以了。

$$y1 = -0.4180x_1 + 0.4836x_2 + 0.4336x_3 + 0.4230x_4 + 0.4736x_5$$

$$y2 = 0.5645x_1 + 0.2255x_2 + 0.4508x_3 - 0.5528x_4 + 0.3489x_5$$

样本主成分分析——案例分析3

$$y1 = -0.4180x_1 + 0.4836x_2 + 0.4336x_3 + 0.4230x_4 + 0.4736x_5$$

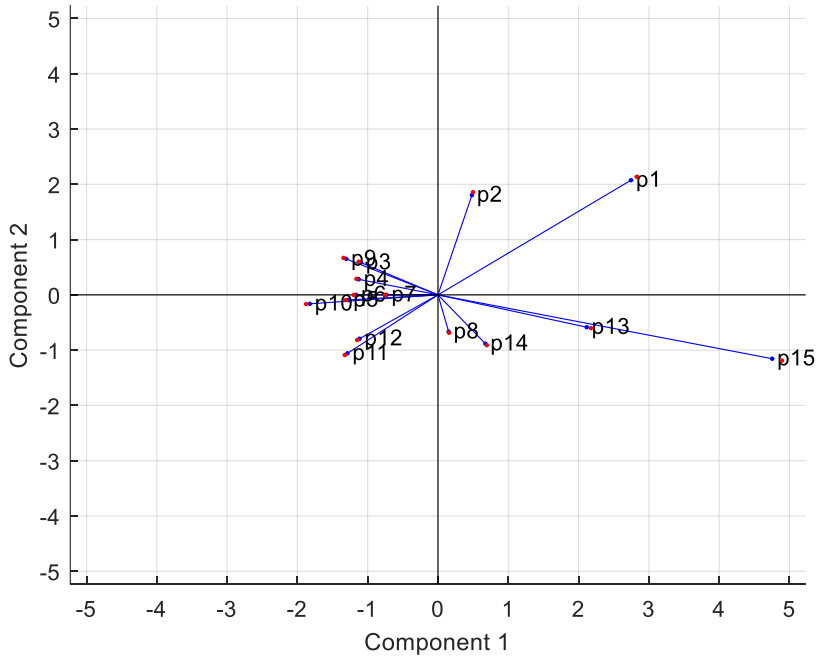
$$y2 = 0.5645x_1 + 0.2255x_2 + 0.4508x_3 - 0.5528x_4 + 0.3489x_5$$

从第一主成分y1的表达式来看，主要反映了有机污染物和水质自净作用的对比程度，该值越大，说明水质越好，自净能力强；

第二主成分y2主要反映在第1主成分值大体固定的条件下水体中的氨的形成富营养化程度的度量，随着T-N项权值的增加，说明富营养化引起水质的下降。

```
>> biplot(score(:,1:2),'Score',score(:,1:2),'VarLabels',text(2:end,1))
```

在第一、第二主成分得分最高的前三依次为P15、P1、P13，故最佳的检测布设点可设在P15.



主成分分析用于综合评价

例4：根据2008年安徽统计年鉴资料，选择X_1：工业总产值(现价)，X_2：工业销售产值(当年价)，X_3：流动资产年平均余额，X_4：固定资产净值年平均余额，X_5：业务收入，X_6：利润总额等六项指标进行主成分分析. (1)选取指标是否合适？ (2) 给出各市大中型工业企业排名。

地 区	x1	x2	x3	x4	x5	x6	地 区	x1	x2	x3	x4	x5	x6
合肥市	1932.3	1900.5	653.83	570.95	1810.7	119.53	马鞍山市	905.32	894.61	351.52	502.99	1048	53.88
淮北市	367.05	366.08	186.16	252.07	395.43	32.82	巢湖市	254.99	242.38	106.66	75.48	234.76	19.65
亳州市	86.89	85.38	40.85	51.71	83.26	8.95	芜湖市	867.07	852.34	418.82	217.76	806.94	37.01
宿州市	154.27	147.07	30.68	57.96	146.3	-1.27	宣城市	219.36	207.07	82.58	54.74	192.74	11.02
蚌埠市	197.21	193.28	104.56	90.15	182.6	7.85	铜陵市	570.33	563.33	224.23	190.77	697.91	20.61
阜阳市	244.17	231.55	56.37	121.96	224.04	26.49	池州市	59.11	57.32	16.97	40.33	56.56	6.03
淮南市	497.74	483.69	206.8	501.37	496.59	27.76	安庆市	430.58	426.25	103.08	147.05	442.04	0.79
滁州市	308.91	296.99	118.65	76.9	277.42	19.32	黄山市	65.03	64.36	28.38	8.58	60.48	2.88
六安市	191.77	189.05	70.19	62.31	191.98	23.08							

主成分分析用于综合评价

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

```
[ahdata,text] = xlsread('anhui.xlsx');
%删除第一列，因为第一列与第二列完全相关，可计算相关系数
ahdata(:,1) = [];
adzs = zscore(ahdata);
[coeff,score,latent,~,explained] = pca(adzs);
cscon = cumsum(explained)  %计算累积贡献率
indiscore = adzs*coeff*(explained/100);  %计算综合评价得分，可选择前两个主成分计算
[is,ind] = sort(indiscore,'descend');
city = text(2:end,1);
city = city(ind);  %排序
res = cell(length(city),2);
res(:,1) = city;
res(:,2) = num2cell(is);
```

特征值	特征向量	贡献率	累积贡献率
4.6100	(0.4595, 0.4552, 0.4158, 0.4600, 0.4441)	0.9220	0.9220
0.2475	(-0.2517, -0.2103, 0.9054,-0.1315,-0.2354)	0.0495	0.9715
0.1050	(0.1926, 0.3702, -0.0390, 0.3029, -0.8559)	0.0210	0.9925
0.0322	(-0.3510, 0.7779, 0.0275, -0.5153, 0.0738)	0.0064	0.9989
0.0053	(0.7518, -0.0803, 0.0719, -0.6434, -0.0965)	0.0011	1.0000

```
res =
  17×2 cell 数组
    '合 肥 市'      [ 6.0702]
    '马鞍山市'      [ 2.6351]
    '芜 湖 市'      [ 1.6096]
    '淮 南 市'      [ 1.0376]
    '铜 陵 市'      [ 0.4814]
    '淮 北 市'      [ 0.2671]
    '安 庆 市'      [-0.5576]
    '滁 州 市'      [-0.6813]
    '阜 阳 市'      [-0.7242]
    '巢 湖 市'      [-0.7959]
    '六 安 市'      [-0.9511]
    '蚌 埠 市'      [-1.0292]
    '宣 城 市'      [-1.0992]
    '亳 州 市'      [-1.4443]
    '宿 州 市'      [-1.4914]
    '池 州 市'      [-1.6208]
    '黄 山 市'      [-1.7060]
```

主成分分析用于分类

例5：瑞士银行纸币数据为200×6的矩阵，其中100行是真纸币数据，100行是假币数据。六项指标为：纸币长度，左、右侧纸币高度，上、下图廓内骨架距离以及对角线长度。选择两个主成分的得分做出平面图形，能否从图形上分别真假纸币？

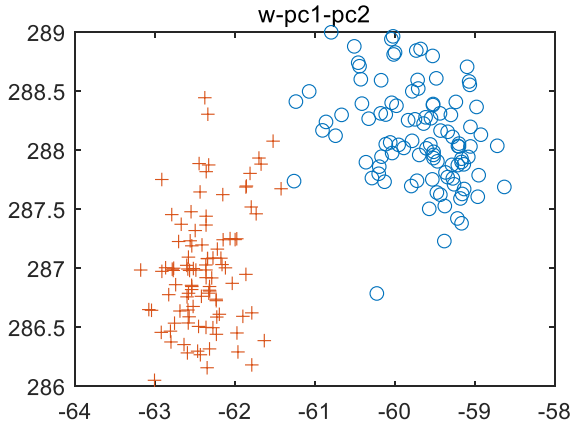
序号	长度	左侧高度	右侧高度	图廓下边距	图廓上边距	对角线长度
1	214.8	131	131.1	9	9.7	141
2	214.6	129.7	129.7	8.1	9.5	141.7
3	214.8	129.7	129.7	8.7	9.6	142.2
4	214.8	129.7	129.6	7.5	10.4	142
5	215	129.6	129.7	10.4	7.7	141.8
6	215.7	130.8	130.5	9	10.1	141.4
…	…	…	…	…	…	…

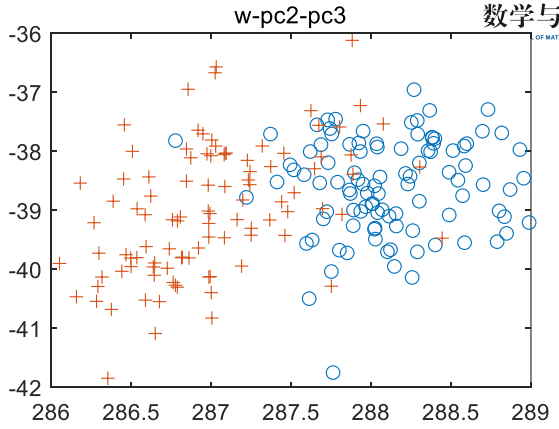
```
swiss = xlsread('swissbanknotes.xlsx');
swiss(:,1) = [];
sw = zscore(swiss);
[coeff,score,latent,~,explained] = pca(sw);
```

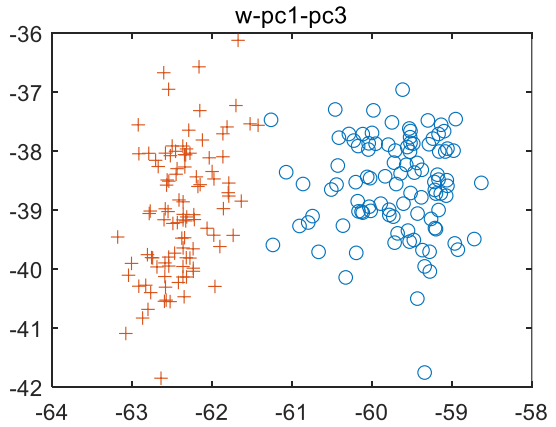
利用主成分分析可以计算出主成分的得分，如果主成分的贡献率较大，则其提取了原始数据的主要信息，因此可以利用主成分分析进行分类。

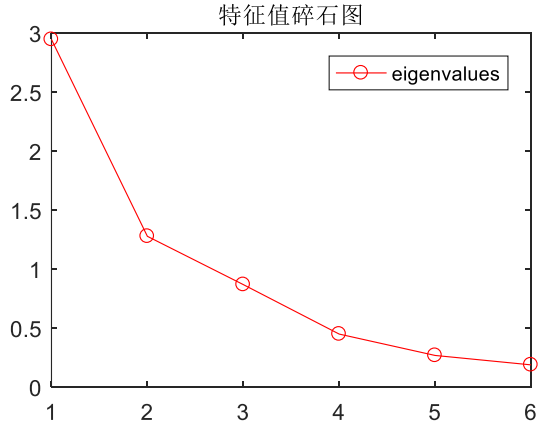
主成分分析用于分类

```
F = swiss*coeff;%计算主成分得分
subplot(2,2,1)
plot(F(1:100,1),F(1:100,2),'o',F(101:200,1),F(101:200,2),'+')
title('w-pc1-pc2')
subplot(2,2,2)
plot(F(1:100,2),F(1:100,3),'o',F(101:200,2),F(101:200,3),'+')
title('w-pc2-pc3')
subplot(2,2,3)
plot(F(1:100,1),F(1:100,3),'o',F(101:200,1),F(101:200,3),'+')
title('w-pc1-pc3')
subplot(2,2,4)
plot(latent,'-or') %从大到小特征值
legend('eigenvalues ')
title('特征值碎石图')
```









从图可以看出，第一，第二两个主成分区分度最好，第一，第三两个主成分也能区分真假币，但第二、第三两个主成分有部分样本交叉，不能很好地区分真假纸币。这从另一个侧面反映前两个主成分代表了较大的方差信息。


01
IDEA
02
LEARNING
03
CREATIVE
INSPIRATION
EDUCATION
BUSINESS
PHOTO BOOK

感谢聆听