第八章 降维分析

8.1 主成分分析

8.2 因子分析

第八章 降维分析

随着大数据时代的发展，人们获取、存储数据的方式越来越方便。人们在描述一个对象时，总是希望更加全面、利用更多的指标，从更多的维度，来对对象的特征进行描述。

$$x = [x_1, x_2, \cdots, x_p]^T$$

相应地，这就导致我们会经常面临大量高维数据的处理问题。然而，很多数据分析处理方法（聚类分析、判别分析等），在处理高维数据时，总会带来一些困难，或者是导致效果不好。

第八章 降维分析

降维是寻找适当的低维空间来表达原数据的过程。我们希望数据表达方式的改变带来以下效果：

（1）以发现数据结构模式为目标探索高维数据，这种数据模式应该可以满足形成统计假设；

（2）当数据降维到二维或三维时，使用散点图可视化数据；

（3）使用统计方法分析数据，比如聚类、回归、判别分析等。

第八章 降维分析

本章描述的降维方法是构造、创建包含初始变量的新变量（比如线性组合），寻找从高维空间到低维空间的映射，并保持高维变量的信息。通常，这种映射可以是线性的或者非线性的。

主成分分析　principal component analysis，PCA

因子分析　factor analysis，FA

第八章 降维分析

8.1 主成分分析PCA

1. 主成分分析PCA基本思想

主成分分析（principal component analysis）是 1901 年 Pearson 对非随机变量引入的，1933 年 Hotelling 将此方法推广到随机向量的情形。

PCA将多指标转化为几个综合指标的多元统计分析的方法，主要目的是希望用较少的变量去解释原来资料中的大部分信息。通常选出的变量要比原始指标的变量少，能解释大部分资料中变异的几个新指标变量，即所谓的主成分。

第八章 降维分析

8.1 主成分分析PCA

1. 主成分分析PCA基本思想

具体而言，主成分分析通过将原来指标重新组合成一组新的相互无关的几个综合指标，来消除原有指标间的相关性，由几个相互无关的综合指标尽可能多地反映原来指标的信息，从而实现降维的一种方法。所构造的几个综合指标就称为 **主成分。**

第八章 降维分析

8.1 主成分分析PCA

2. 主成分分析相关定义与定理

对于m维随机向量 $(x_1, x_2, \cdots, x_m)$ 通过线性组合方式，构造一个新的指标（变量）

$$y_1 = c_1 x_1 + c_2 x_2 + \cdots + c_m x_m$$

如何选取线性组合系数 $l_1=[c_1, c_2, \ldots c_m]$?

我们希望所构造的新指标 y_1, 其方差能够最大，$\mathrm{Var}(y_1)$—> max

由于方差反映了数据差异的程度，因此也就表明我们抓住了这m个变量的最大的变异特征。此时， y_1 称为第一主成分

第八章 降维分析

8.1 主成分分析PCA

2. 主成分分析相关定义与定理

定义 8-1 设 $\boldsymbol{X}=(x_1,x_2,\cdots,x_m)'$ 为 m 维随机向量，则 $\boldsymbol{X}$ 的第 1，2，…，m 主成分定义为

$$y_i=\boldsymbol{l}_i'\boldsymbol{X}\ ,\quad \boldsymbol{l}_i'\boldsymbol{l}_i=1\quad (i=1,2,\cdots,m)$$

它们满足

（1）第一主成分 y_1 是一切形如 $y=\boldsymbol{l}'\boldsymbol{X}$，$\boldsymbol{l}'\boldsymbol{l}=1$ 使 y 的方差达极大者；

（2）第二主成分 y_2 是一切形如 $y=\boldsymbol{l}'\boldsymbol{X}$，$\boldsymbol{l}'\boldsymbol{l}=1$ 且与 y_1 不相关使 y 的方差达极大者；

（3）第 i 主成分 $y_i(i\le m)$ 是一切形如 $y=\boldsymbol{l}'\boldsymbol{X}$，$\boldsymbol{l}'\boldsymbol{l}=1$ 且与 $y_1,y_2,\cdots,y_{i-1}$ 不相关使 y 的方差达极大者；

第八章 降维分析

8.1 主成分分析PCA

2. 主成分分析相关定义与定理

定理 8-1 设 $\boldsymbol{X}=(x_1,x_2,\cdots,x_m)'$ 为 m 维随机向量，协方差阵为 Σ，Σ 的 m 个特征值为 $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_m\geq 0$，相应的标准正交化特征向量为 $\boldsymbol{l}_1,\boldsymbol{l}_2,\cdots,\boldsymbol{l}_m$，则 $\boldsymbol{X}$ 的第 i 主成分 $y_i=\boldsymbol{l}_i'\boldsymbol{X}$ 。

定义 8-2 称 $\lambda_i/\sum_{j=1}^{m}\lambda_j$ 为主成分 y_i 的贡献率，

$\sum_{i=1}^{k}\lambda_i/\sum_{j=1}^{m}\lambda_j$ 为主成分 $y_1,y_2,\cdots,y_k$ 的累计贡献率。

第八章 降维分析

8.1 主成分分析PCA

3. 主成分分析的具体实施

$(x_1, x_2, \cdots, x_m)$

设y_i表示第i个主成分，$i=1,2,\cdots,m$，可设

$$\begin{cases} y_1 = c_{11}x_1 + c_{12}x_2 + \cdots + c_{1m}x_m, \\ y_2 = c_{21}x_1 + c_{22}x_2 + \cdots + c_{2m}x_m, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ y_m = c_{m1}x_2 + c_{m2}x_2 + \cdots + c_{mm}x_m, \end{cases}$$

其中对每一个i，均有$c_{i1}^2 + c_{i2}^2 + \cdots + c_{im}^2 = 1$，且$[c_{11}, c_{12}, \cdots, c_{1m}]$使得$\mathrm{Var}(y_1)$的值达到最大；$[c_{21}, c_{22}, \cdots, c_{2m}]$不仅垂直于$[c_{11}, c_{12}, \cdots, c_{1m}]$，而且使$\mathrm{Var}(y_2)$的值达到最大；

8.1 主成分分析PCA

3. 主成分分析的具体实施

$(x_1, x_2, \cdots, x_m)$

$[c_{31}, c_{32}, \cdots, c_{3m}]$ 同时垂直于 $[c_{11}, c_{12}, \cdots, c_{1m}]$ 和 $[c_{21}, c_{22}, \cdots, c_{2m}]$，并使$\mathrm{Var}(y_3)$的值达到最大；以此类推可得全部 m个主成分，这项工作用手做是很繁琐的，但借助于计算机很容易完成。剩下的是如何确定主成分的个数，我们总结在下面几个注意事项中。

第八章 降维分析

8.1 主成分分析PCA

3. 主成分分析的具体实施

$(x_1, x_2, \cdots, x_m)$

（1）主成分分析的结果受量纲的影响，由于各变量的单位可能不一样，如果各自改变量纲，结果会不一样，所以实际中可以先把各变量的数据标准化，然后使用协方差矩阵或相关系数矩阵进行分析。

第八章 降维分析

8.1 主成分分析PCA

3. 主成分分析的具体实施

$(x_1, x_2, \cdots, x_m)$

（2）在实际研究中，由于主成分的目的是为了降维，减少变量的个数，故一般选取少量的主成分（不超过5或6个），只要它们能解释变异的70%～80%（称累积贡献率）就行了。

8.1 主成分分析PCA

4. 主成分分析的步骤

$(x_1, x_2, \cdots, x_m)$

设有m个指标变量$x_1, x_2, \cdots, x_m$，它在第i次观测中的取值为

$$a_{i1}, a_{i2}, \cdots, a_{im} \quad (i = 1, 2, \cdots, n),$$

将它们写成矩阵形式

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix},$$

矩阵A称为观测阵。

8.1 主成分分析PCA

4. 主成分分析的步骤

对于观测数据矩阵$A=(a_{ij})_{n\times m}$。按如下步骤进行 PCA 分析

（1）对原来的m个指标进行标准化，得到标准化的指标变量

$$y_j=\frac{x_j-\mu_j}{s_j},\quad j=1,2,\cdots,m,$$

其中，$\mu_j=\frac{1}{n}\sum_{i=1}^{n}a_{ij}$，$s_j=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(a_{ij}-\mu_j)^2}$。对应地，得到标准化的数据矩阵$B=(b_{ij})_{n\times m}$，其中$b_{ij}=\frac{a_{ij}-\mu_j}{s_j}$，$i=1,2,\cdots,n$，$j=1,2,\cdots,m$。

$$A=\begin{bmatrix}a_{11} & a_{12} & \cdots & a_{1m}\\ a_{21} & a_{22} & \cdots & a_{2m}\\ \vdots & \vdots & & \vdots\\ a_{n1} & a_{n2} & \cdots & a_{nm}\end{bmatrix}$$

$$B=(b_{ij})_{n\times m}=\begin{bmatrix}b_{11} & b_{12} & \cdots & b_{1m}\\ b_{21} & b_{22} & \cdots & b_{2m}\\ \vdots & \vdots & & \vdots\\ b_{n1} & b_{n2} & \cdots & b_{nm}\end{bmatrix}$$

8.1 主成分分析PCA

4. 主成分分析的步骤

$$\hat{\rho}_{ij}=\frac{\sigma_{ij}}{\sqrt{\hat{\sigma}_{ii}\times\hat{\sigma}_{jj}}},i,j=1,2,\cdots,m$$

$$B=(b_{ij})_{n\times m}=\begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \vdots & \vdots & & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nm} \end{bmatrix}$$

（2）根据标准化的数据矩阵B求出协方差阵 $\sum$ 或 相关系数矩阵 $R=(r_{ij})_{m\times m}$，

.

（3）计算协方差阵 $\sum$ 或 相关系数矩阵R的特征值$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_m$，及对应的标准正交化特征向量$u_1,u_2,\cdots,u_m$，其中$u_j=[u_{1j},u_{2j},\cdots,u_{mj}]^T$，由特征向量组成$m$个新的指标变量

$$\begin{cases} y_1=u_{11}x_1+u_{21}x_2+\cdots+u_{m1}x_m, \\ y_2=u_{12}x_1+u_{22}x_2+\cdots+u_{m2}x_m, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ y_m=u_{1m}x_1+u_{2m}x_2+\cdots+u_{mm}x_m. \end{cases}$$

式中y_1是第 1 主成分，y_2是第 2 主成分，…，y_m是第m主成分。

8.1 主成分分析PCA

4. 主成分分析的步骤

（4）计算主成分贡献率及累计贡献率，主成分y_j的贡献率为

$$w_j=\frac{\lambda_j}{\sum_{k=1}^{m}\lambda_k},\quad j=1,2,\cdots,m,$$

前i个主成分的累计贡献率为

$$\sum_{k=1}^{i}\lambda_k \Big/ \sum_{k=1}^{m}\lambda_k .$$

一般取累计贡献率达 85%以上的特征值$\lambda_1,\lambda_2,\cdots,\lambda_k$所对应的第 1、第 2、…、第$k(k\le p)$主成分。

（5）最后利用得到的主成分$y_1,y_2,\cdots,y_k$分析问题，或者继续进行评价、回归、聚类等其他建模。

第八章 降维分析

8.1 主成分分析PCA

例 8-1 设 $X=(x_1,x_2,x_3)'$ 的协方差阵为 Σ ，试讨论主成分 y_1,y_2,y_3 累计贡献率，并计算 $X=(1,2,3)'$ 的第一、二主成分。其中

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

解：（1）求特征值，由

$$\begin{vmatrix}1-\lambda & -2 & 0\\ -2 & 5-\lambda & 0\\ 0 & 0 & 2-\lambda\end{vmatrix}=(2-\lambda)[(1-\lambda)(5-\lambda)-4]=0$$

即 $(2-\lambda)[\lambda^2-6\lambda+1]=0$，解得三个特征值分别为

8.1 主成分分析PCA

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

（2）求特征向量，由

$$\begin{bmatrix} 1-\lambda_i & -2 & 0 \\ -2 & 5-\lambda_i & 0 \\ 0 & 0 & 2-\lambda_i \end{bmatrix} \begin{pmatrix} \xi \\ \eta \\ \zeta \end{pmatrix} = 0$$

将 $\lambda_1 = 5.83$ 代入上式得

$$\begin{bmatrix} -4.83 & -2 & 0 \\ -2 & -0.83 & 0 \\ 0 & 0 & -3.83 \end{bmatrix} \begin{pmatrix} \xi \\ \eta \\ \zeta \end{pmatrix} = 0 \Longrightarrow \begin{aligned} \zeta &= 0 \\ -4.83\xi - 2\eta &= 0 \\ -2(2.415\xi + \eta) &= 0 \end{aligned}$$

令 $\xi = 1$，得 $\eta = -2.415$；

或令 $\eta = 1$，得 $\xi = -1/2.415 = -0.414$；

解得相应的特征向量（标准化）为

8.1 主成分分析PCA

$$l_1 = \frac{1}{\sqrt{1+(-2.415)^2}}\begin{pmatrix}1\\-2.415\\0\end{pmatrix} = \begin{pmatrix}0.383\\-0.924\\0\end{pmatrix}$$

同理解得 $\lambda_2 = 2.00$，$\lambda_3 = 0.17$ 相应的特征向量（标准化）为

$$\boldsymbol{l}_2 = \begin{bmatrix}0\\0\\1\end{bmatrix}，\quad \mathbf{l}_3 = \begin{bmatrix}0.924\\0.383\\0.00\end{bmatrix}$$

如果我们只取第一主成分，贡献率可达

$$5.83/(5.83+2.00+0.17) = 72.88\%$$

这似乎很理想，如果进一步计算对每个向量的贡献率，

8.1 主成分分析PCA

故考虑第二主成分，这时贡献率可达

$$(5.83+2.00)/(5.83+2.00+0.17)=97.88\%$$

$$y_1=\boldsymbol{l}_1'\boldsymbol{X}=\begin{pmatrix}0.383, & -0.924, & 0.00\end{pmatrix}\begin{pmatrix}1\\2\\3\end{pmatrix}=-1.465$$

$$y_2=\boldsymbol{l}_2'\boldsymbol{X}=\begin{pmatrix}0, & 0, & 1\end{pmatrix}\begin{pmatrix}1\\2\\3\end{pmatrix}=3$$

$$l_1=\frac{1}{\sqrt{1+(-2.415)^2}}\begin{pmatrix}1\\-2.415\\0\end{pmatrix}=\begin{pmatrix}0.383\\-0.924\\0\end{pmatrix}\qquad \boldsymbol{l}_2=\begin{bmatrix}0\\0\\1\end{bmatrix},\ \mathbf{l}_3=\begin{bmatrix}0.924\\0.383\\0.00\end{bmatrix}$$

8.1 主成分分析PCA

5. 主成分分析的Python实现

3. sklearn.decomposition 模块的 PCA 函数

sklearn.decomposition 模块的 PCA 函数实现主成分分析，其调用格式为：

```
sklearn.decomposition.PCA(n_components=None, copy=True)
```

其中，n_components：类型为 int 或字符串，缺省时默认为 None，所有成分被保留；赋值为 int，比如 n_components=2，将提取两个主成分；赋值为（0，1）上的浮点数，将自动选择主成分的个数，使得满足信息贡献率的要求。

8.1 主成分分析PCA

5. 主成分分析的Python实现

copy：类型：bool，True 或者 False，缺省时默认为 True；表示是否在运行算法时，将原始训练数据复制一份。若为 True，则运行 PCA 算法后，原始训练数据的值不会有任何改变，因为是在原始数据的副本上进行运算；若为 False，则运行 PCA 算法后，原始训练数据的值会改，因为是在原始数据上进行降维计算。

8.1 主成分分析PCA

5. 主成分分析的Python实现

例 8-2　对 10 名男中学生的身高x_1、胸围x_2和体重x_3进行测量，得数据见表，对其做主成分分析。

表　男中学生的身高、胸围及体重数据

序号	身高x_1(cm)	胸围x_2(cm)	体重x_3(kg)	序号	身高x_1(cm)	胸围x_2(cm)	体重x_3(kg)
1	149.5	69.5	38.5	6	156.1	74.5	45.5
2	162.5	77	55.5	7	172.0	76.5	51.0
3	162.7	78.5	50.8	8	173.2	81.5	59.5
4	162.2	87.5	65.5	9	159.5	74.5	43.5
5	156.5	74.5	49.0	10	157.7	79	53.5

8.1 主成分分析PCA

5. 主成分分析的Python实现

把表中的 5 行 8 列数据保存到文本文件 data81.txt 中。

编写的 Python 程序如下：

```
#程序文件 Pex81.py
import numpy as np
from sklearn.decomposition import PCA
a=np.loadtxt("data81.txt")
b=np.r_[a[:,1:4],a[:,-3:]] # 构造数据矩阵
md=PCA().fit(b) # 构造并训练模型
print("特征值为： ",md.explained_variance_)
```

8.1 主成分分析PCA

5. 主成分分析的Python实现

```
print("各主成分的贡献率： ",md.explained_variance_ratio_)
print("奇异值为： ",md.singular_values_)
print("各主成分的系数：\n",md.components_) # 每行是一个主成分
"""下面直接计算特征值和特征向量，和库函数进行对比"""
cf=np.cov(b.T) # 计算协方差阵
c,d=np.linalg.eig(cf) #求特征值和特征向量
print("特征值为： ",c)
print("特征向量为：\n",d)
print("各主成分的贡献率为： ",c/np.sum(c))
```

8.1 主成分分析PCA

5. 主成分分析的Python实现

```
"""由相关系数阵进行分析"""
print("由相关系数阵进行分析")
cf=np.corrcoef(b.T)  #计算相关系数阵
print("相关系数阵：\n",cf)
c,d=np.linalg.eig(cf) #求特征值和特征向量
print("特征值为： ",c)
print("特征向量为：\n",d)
print("各主成分的贡献率为： ",c/np.sum(c))
```

8.1 主成分分析PCA

5. 主成分分析的Python实现

PCA函数分析结果

程序运行结果如下:

特征值为: [110.00413886 25.32447973 1.56804807]

各主成分的贡献率: [0.80355601 0.18498975 0.01145425]

奇异值为: [31.46485738 15.09703009 3.75665179]

各主成分的系数:

[[-0.55915657 -0.42128705 -0.71404562]
[0.82767368 -0.33348264 -0.45138188]
[-0.04796048 -0.84338992 0.53515721]]

8.1 主成分分析PCA

5. 主成分分析的Python实现

由协方差阵分析结果

```
协方差阵：
 [[51.74544444 18.98666667 34.41922222]
 [18.98666667 23.45555556 36.19555556]
 [34.41922222 36.19555556 61.69566667]]
特征值为： [110.00413886  25.32447973   1.56804807]
特征向量为：
 [[ 0.55915657  0.82767368 -0.04796048]
 [ 0.42128705 -0.33348264 -0.84338992]
 [ 0.71404562 -0.45138188  0.53515721]]
各主成分的贡献率为： [0.80355601 0.18498975 0.01145425]
```

8.1 主成分分析PCA

5. 主成分分析的Python实现

由相关系数阵分析结果

```
相关系数阵：
 [[1.         0.54499159 0.60916881]
 [0.54499159 1.         0.95149181]
 [0.60916881 0.95149181 1.        ]]
特征值为：  [2.42061868 0.53427412 0.0451072 ]
特征向量为：
 [[ 0.49833746  0.86375934 -0.07469656]
 [ 0.60627136 -0.4087742  -0.68215738]
 [ 0.61975383 -0.29465819  0.72738005]]
各主成分的贡献率为：  [0.80687289 0.17809137 0.01503573]
```

8.1 主成分分析PCA

5. 主成分分析的Python实现

注意：

（1）从上面程序运行结果可以看出，PCA 函数使用协方差阵做的主成分分析。主成分分析也可以使用相关系数阵，两者计算结果略有差异，使用相关系数阵做主成分分析，相当于对数据进行了标准化处理。

（2）从程序的运行结果看，主成分的系数可以相差一个负号，因为特征向量乘以–1后仍然为特征向量。

8.1 主成分分析PCA

5. 主成分分析的Python实现

各主成分的贡献率分别为 80.36%，18.50%，1.15%。因此，前两个主成分的累计贡献率已达 98.86%，应用中可取前两个主成分

$$y_1 = 0.5592x_1 + 0.4213x_2 + 0.7140x_3,$$

$$y_2 = 0.8277x_1 - 0.3335x_2 - 0.4514x_3.$$

8.1 主成分分析PCA

6. 主成分分析进行评价

例 8-3　根据 2008 年安徽统计年鉴资料，选择x_1（工业总产值的现价）、x_2（工业销售按当年价的产值）、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）6 项指标进行主成分分析，下表列出了安徽省各市大中型工业企业主要经济指标的统计数据。(1) 选取指标是否合适？ (2) 给出各市大中型工业企业排名。

8.1 主成分分析PCA

6. 主成分分析进行评价

表　安徽省各市大中型工业企业主要经济指标　　（单位：亿元）

地区	X_1	X_2	X_3	X_4	X_5	X_6
合肥市	1932.27	1900.53	653.83	570.95	1810.70	119.53
淮北市	367.05	366.08	186.16	252.07	395.43	32.82
亳州市	86.89	85.38	40.85	51.71	83.26	8.95
宿州市	154.27	147.07	30.68	57.96	146.30	-1.27
蚌埠市	197.21	193.28	104.56	90.15	182.60	7.85
阜阳市	244.17	231.55	56.37	121.96	224.04	26.49
淮南市	497.74	483.69	206.80	501.37	496.59	27.76
滁州市	308.91	296.99	118.65	76.90	277.42	19.32
六安市	191.77	189.05	70.19	62.31	191.98	23.08
马鞍山市	905.32	894.61	351.52	502.99	1048.02	53.88
巢湖市	254.99	242.38	106.66	75.48	234.76	19.65
芜湖市	867.07	852.34	418.82	217.76	806.94	37.01
宣城市	219.36	207.07	82.58	54.74	192.74	11.02
铜陵市	570.33	563.33	224.23	190.77	697.91	20.61
池州市	59.11	57.32	16.97	40.33	56.56	6.03
安庆市	430.58	426.25	103.08	147.05	442.04	0.79
黄山市	65.03	64.36	28.38	8.58	60.48	2.88

8.1 主成分分析PCA

6. 主成分分析进行评价

设有m个指标变量 $x_1, x_2, \cdots, x_m$

$$A=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}$$

矩阵 A ⟶ 标准化的数据矩阵 B

求出矩阵 B 的协方差阵 $\sum_{m\times m}$ 或 相关系数矩阵 $R=(r_{ij})_{m\times m}$

$\sum$ 或 R ⟶ 特征值 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$，标准正交化特征向量 $u_1, u_2, \cdots, u_m$

选定主成分的个数 k，则由 $[u_1, u_2, \cdots, u_k]_{m\times k}$，可构造 $[y_1, y_2, \cdots, y_k]_{n\times k}$ k 个主成分

$$[y_1, y_2, \cdots, y_k]_{n\times k} = B_{n\times m} \cdot [u_1, u_2, \cdots, u_k]_{m\times k}$$

$$\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1k} \\ y_{21} & y_{22} & \cdots & y_{2k} \\ \vdots & \vdots & & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nk} \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \vdots & \vdots & & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nm} \end{bmatrix} \cdot \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1k} \\ u_{21} & u_{22} & \cdots & u_{2k} \\ \vdots & \vdots & & \vdots \\ u_{m1} & u_{m2} & \cdots & u_{mk} \end{bmatrix}$$

8.1 主成分分析PCA

6. 主成分分析进行评价

$$w_j = \frac{\lambda_j}{\sum_{i=1}^{m} \lambda_i}, \quad j = 1, 2, \cdots, k$$

$$\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1k} \\ y_{21} & y_{22} & \cdots & y_{2k} \\ \vdots & \vdots & & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nk} \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_k \end{bmatrix} = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{bmatrix}$$

8.1 主成分分析PCA

6. 主成分分析进行评价

解　（1）利用 Python 软件，求得相关系数矩阵

$$R=\begin{bmatrix}1.0000 & 1.0000 & 0.9754 & 0.8231 & 0.9914 & 0.9375\\ 1.0000 & 1.0000 & 0.9758 & 0.8236 & 0.9920 & 0.9369\\ 0.9754 & 0.9758 & 1.0000 & 0.8245 & 0.9712 & 0.9127\\ 0.8231 & 0.8236 & 0.8245 & 1.0000 & 0.8502 & 0.8020\\ 0.9914 & 0.9920 & 0.9712 & 0.8502 & 1.0000 & 0.9212\\ 0.9375 & 0.9369 & 0.9127 & 0.8020 & 0.9212 & 1.0000\end{bmatrix},$$

由于$r_{12}=r_{21}=1$，表明指标x_1,x_2完全线性相关，所以选取的指标不合适，只需保留x_1,x_2中的一个指标，这里我们删除指标x_1。

8.1 主成分分析PCA

6. 主成分分析进行评价

（2）各市大中型工业企业排名的数学原理我们这里就不赘述了。只给出计算结果和 Python 程序。

第一主成分信息贡献率达到 92.2%，选取一个主成分进行评价即可，主成分及信息贡献率计算结果见表。根据第一主成分的评价结果见表。

特征值	特征向量	贡献率
4.6100	$[0.4595, 0.4552, 0.4158, 0.4600, 0.4441]^T$	92.2009%
0.2475	$[0.2517, 0.2103, -0.9054, 0.1315, 0.2354]^T$	4.9501%
0.1050	$[0.1926, 0.3702, -0.0390, 0.3029, -0.8559]^T$	2.1007%
0.0322	$[-0.3510, 0.7779, 0.0275, -0.5153, 0.0738]^T$	0.6431%
0.0053	$[-0.7518, 0.0803, -0.0719, 0.6434, 0.0965]^T$	0.1053%

8.1 主成分分析PCA

6. 主成分分析进行评价

表 各市第一主成分得分排名

排名	地区	得分	排名	地区	得分	排名	地区	得分
1	合肥市	6.2827	7	安庆市	-0.5654	13	宣城市	-1.1219
2	马鞍山市	2.6810	8	滁州市	-0.6892	14	亳州市	-1.4888
3	芜湖市	1.6979	9	阜阳市	-0.7568	15	宿州市	-1.5324
4	淮南市	0.9914	10	巢湖市	-0.8118	16	池州市	-1.6711
5	铜陵市	0.5144	11	六安市	-0.9758	17	黄山市	-1.7484
6	淮北市	0.2516	12	蚌埠市	-1.0575			

8.1 主成分分析PCA

6. 主成分分析进行评价

```
#程序文件 Pex82.py
import numpy as np
from scipy.stats import zscore
a=np.loadtxt("data82.txt")
print("相关系数阵为： \n",np.corrcoef(a.T))
b=np.delete(a,0,axis=1) #删除第 1 列数据
c=zscore(b); r=np.corrcoef(c.T) #数据标准化并计算相关系数阵
d,e=np.linalg.eig(r) #求特征值和特征向量
rate=d/d.sum()  #  计算各主成分的贡献率
print("特征值为： ",d)
```

8.1 主成分分析PCA

6. 主成分分析进行评价

```
print("特征向量为： \n",e)
print("各主成分的贡献率为： ",rate)
k=1; #提出主成分的个数
F=e[:,:k]; score_mat=c.dot(F) #计算主成分得分矩阵
score1=score_mat.dot(rate[0:k]) # 计算各评价对象的得分
score2=-score1 # 通过观测，调整得分的正负号
print("各评价对象的得分为： ",score2)
index=score1.argsort()+1 # 排序后的每个元素在原数组中的位置
print("从高到低各个城市的编号排序为： ",index)
```

第八章 降维分析

8.1 主成分分析PCA

7. 主成分-回归分析

主成分回归分析采用的方法是将原来的回归自变量变换到另一组变量，即主成分，以主成分作为新的自变量，达到了降维的目的，然后对选取主成分作为回归变量进行回归分析，最后再变换回原来的模型求出参数的估计。

第八章 降维分析

8.1 主成分分析PCA

7. 主成分-回归分析

例　Hald水泥问题，考察含如下四种化学成分 x_1 =3CaO·Al_2O_3的含量（%），x_2 =3CaO·SiO_2的含量（%），x_3 =4CaO·Al_2O_3·Fe_2O_3的含量（%），x_4 =2CaO·SiO_2的含量（%），的某种水泥，每一克所释放出的热量（卡）y与这四种成分含量之间的关系数据共13组，见表，对数据实施标准化得到数据矩阵$\tilde{A}$，样本相关系数阵。

第八章 降维分析

8.1 主成分分析PCA

7. 主成分-回归分析

表　Hald水泥

序号	x1	x2	x3	x4	y
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

第八章 降维分析

8.1 主成分分析PCA

7. 主成分-回归分析

表　相关系数矩阵

r =

1.0000	0.2286	-0.8241	-0.2454
0.2286	1.0000	-0.1392	-0.9730
-0.8241	-0.1392	1.0000	0.0295
-0.2454	-0.9730	0.0295	1.0000

第八章 降维分析

8.1 主成分分析PCA

7. 主成分-回归分析

相关系数阵的四个特征值依次为2.2357，1.5761，0.1866，0.0016。最后一个特征值接近于零，前三个特征值之和所占比例（累积贡献率）达到0.999594。于是我们略去第4个主成分。其它三个保留的特征值对应的三个特征向量分别为

$$\eta_1^T = [0.476, 0.5639, -0.3941, -0.5479],$$

$$\eta_2^T = [-0.509, 0.4139, 0.605, -0.4512],$$

$$\eta_3^T = [0.6755, -0.3144, 0.6377, -0.1954],$$

8.1 主成分分析PCA

7. 主成分-回归分析

即取前三个主成分，分别为

$$z_1 = 0.476\tilde{x}_1 + 0.5639\tilde{x}_2 - 0.3941\tilde{x}_3 - 0.5479\tilde{x}_4,$$
$$z_2 = -0.509\tilde{x}_1 + 0.4139\tilde{x}_2 + 0.605\tilde{x}_3 - 0.4512\tilde{x}_4,$$
$$z_3 = 0.6755\tilde{x}_1 - 0.3144\tilde{x}_2 + 0.6377\tilde{x}_3 - 0.1954\tilde{x}_4.$$

8.1 主成分分析PCA

7. 主成分-回归分析

对Hald数据直接作线性回归得经验回归方程

$$\hat{y} = 62.4054 + 1.5511x_1 + 0.5102x_2 + 0.102x_3 - 0.144x_4.$$

作主成分回归分析，得到如下回归方程

$$\hat{y} = 0.657z_1 + 0.0083z_2 + 0.3028z_3,$$

化成标准化变量的回归方程为

$$\hat{y} = 0.513\tilde{x}_1 + 0.2787\tilde{x}_2 - 0.0608\tilde{x}_3 - 0.4229\tilde{x}_4,$$

恢复到原始的自变量，得到如下主成分回归方程

$$\hat{y} = 85.7433 + 1.3119x_1 + 0.2694x_2 - 0.1428x_3 - 0.3801x_4.$$

第八章 降维分析

8.1 主成分分析PCA

7. 主成分-回归分析

区别在于后者具有更小的均方误差，因而更稳定。此外前者所有系数都无法通过显著性检验。

第八章 降维分析

8.1 主成分分析PCA

8. 主成分-聚类分析

例 7.4　Iris 数据集由 Fisher 于 1936 收集整理。Iris 也称鸢尾花卉数据集，是一类多重变量分析的数据集。数据集包含 150 个数据集，分为 3 类，每类 50 个数据，每个数据包含 4 个属性，数据格式如表所示。可通过花萼长度，花萼宽度，花瓣长度，花瓣宽度 4 个属性预测鸢尾花卉属于（Setosa，Versicolour，Virginica）三个种类中的哪一类。

表　Iris 数据集数据（全部数据见数据文件 iris.csv）

	Sepal_Length	Sepal_Width	Petal_Length	Petal_Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3	1.4	0.2	setosa
⋮	⋮	⋮	⋮	⋮	⋮
149	6.2	3.4	5.4	2.3	virginica
150	5.9	3	5.1	1.8	virginica

可以先采用PCA对4个属性进行降维，再进行聚类分析