第六章

特征值与特征向量的计算

特征值与特征向量

已知n阶方阵A，求数λ和非零向量x使得

$$Ax = \lambda x$$

《线性代数》中介绍的方法

1）计算特征方程$|\lambda E - A| = 0$, 求得矩阵A的所有特征值。

2）对每个特征值λ_i, 求解齐次方程组$(\lambda E - A)x = 0$, 该方程组的所有非零解都是特征值λ_i所对应的特征向量。

注：该方法不适用于计算高阶方阵（n较大时），因为高次多项式的求根问题通常是数值不稳定。

Matlab函数

- [V,D]=eig(A)

例1　　A=[1,-1;2,4];

[V,D]=eig(A)

ans

V=

-0.7071	0.4472
0.7071	-0.8944

方阵A的特征向量矩阵

D=

2	0
0	3

方阵A的特征值矩阵

幂法和反幂法

幂法 用于计算矩阵按模<u>最大</u>的特征值及其相应的特征向量, 特别适用于大型稀疏矩阵.

反幂法 用于计算矩阵按模<u>最小</u>的特征值及其特征向量, 也可用来计算对应于一个给定近似特征值的特征向量.

➢ 幂法

设A为n阶实矩阵, 其特征值为$\lambda_1, \lambda_2, \ldots, \lambda_n$, 相应的特征向量为$u_1, u_2, \ldots, u_n$. 且满足条件

$$\boxed{|\lambda_1| >} |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n|$$

$u_1, u_2, \ldots, u_n$线性无关.

此时λ_1一定是实数!

- 幂法：求λ_1及其相应的特征向量.

- λ_1通常称为主特征值.

➢ 幂法基本思想

- 给定初始非零向量$x^{(0)}$，由矩阵A构造一向量序列

$$\begin{cases} x^{(1)} = Ax^{(0)} \\ x^{(2)} = Ax^{(1)} = A^2 x^{(0)} \\ \cdots \\ x^{(k+1)} = Ax^{(k)} = A^{k+1} x^{(0)} \\ \cdots \end{cases}$$

- 在一定条件下, 当k充分大时: $\lambda_1 \approx \dfrac{x_i^{(k+1)}}{x_i^{(k)}}$
- 相应的特征向量为: $x^{(k+1)}$

➢ 幂法的理论依据

对任意向量$x^{(0)}$，有 $x^{(0)}=\sum_{i=1}^{n}\alpha_i u_i$，设$\alpha_1$不为零.

$$x^{(k+1)}=Ax^{(k)}=A^{k+1}x^{(0)}$$

$$=\sum_{i=1}^{n}A^{k+1}\alpha_i u_i=\sum_{i=1}^{n}\alpha_i\lambda_i^{k+1}u_i$$

$$=\lambda_1^{k+1}\left[\alpha_1 u_1+\left(\frac{\lambda_2}{\lambda_1}\right)^{k+1}\alpha_2 u_2+\cdots+\left(\frac{\lambda_n}{\lambda_1}\right)^{k+1}\alpha_n u_n\right]$$

$$\approx\lambda_1^{k+1}\alpha_1 u_1$$

- 故 $\lambda_1\approx x_i^{(k+1)}\big/x_i^{(k)}$
- $x^{(k+1)}$为λ_1的特征向量的近似向量(除一个因子外).

- 如果$x^{(0)}$的选取恰恰使得$\alpha_1=0$，幂法仍能进行．因为计算过程中会有舍入误差, 迭代若干次后, 必然会产生一个向量$x^{(k)}$, 它在u_1方向上的分量不为零, 这样以后的计算就满足所设条件.

- 因为 $x^{(k)} \approx \lambda_1^k \alpha_1 u_1$，计算过程中可能会出现上溢($|\lambda_1|>1$)或下溢成为0 ($|\lambda_1|<1$). 为避免出现这一情形, 实际计算时每次迭代所求的向量都要归一化.

➢ 归一化过程

设有一向量$x \neq 0$, 将其归一化得到向量

$$y = \frac{x}{\max(x)}$$

其中$\max(x)$表示向量x的绝对值最大的分量，即如果有

$$|x_{i_0}| = \max_{1 \le i \le n} |x_i|$$

则$\max(x) = x_{i_0}$，且i_0为所有绝对值最大的分 量中的最小下标.

例　$x=(1, -8, 7)^T$, 则 $\max(x)=-8$，归一化向量为

$$y = \frac{x}{\max(x)} = (-0.125, 1, -0.875)^T$$

定理 设n阶实矩阵A有n个线性无关的特征向量u_1, $u_2, \ldots, u_n$, 主特征值λ_1满足$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \ldots \geq |\lambda_n|$, 则对任取非零初始向量$x^{(0)}=y^{(0)}\neq 0$ ($\alpha_1\neq 0$), 按下述方法构造向量序列 $\{x^{(k)}\}$, $\{y^{(k)}\}$

$$\begin{cases} x^{(k)} = Ay^{(k-1)} \\ \alpha_k = \max\left(x^{(k)}\right) \quad (k = 1, 2, \cdots) \\ y^{(k)} = \dfrac{x^{(k)}}{\alpha_k} \end{cases}$$

则有 (1) $\lim\limits_{k\to\infty} y_k = \dfrac{u_1}{\max(u_1)}$,

(2) $\lim\limits_{k\to\infty} \alpha_k = \lambda_1$.

例　用幂法求矩阵 $A=\begin{bmatrix}2 & -1 & 0\\ 0 & 2 & -1\\ 0 & -1 & 2\end{bmatrix}$

的按模最大的特征值和相应的特征向量.

取 $x^{(0)}=(0, 0, 1)^T$, 要求误差不超过10^{-3}.

解　$y^{(0)}=x^{(0)}=(0,0,1)^T,$

$$x^{(1)}=Ay^{(0)}=(0,-1,2)^T,\quad \alpha_1=\max(x^{(1)})=2,$$

$$y^{(1)}=\frac{x^{(1)}}{\alpha_1}=(0,-0.5,1)^T$$

$$x^{(2)}=Ay^{(1)}=(0.5,-2,2.5)^T,\quad \alpha_2=\max(x^{(2)})=2.5,$$

$$y^{(2)} = \frac{x^{(2)}}{\alpha_2} = (0.2, -0.8, 1)^T$$

$$x^{(3)} = Ay^{(2)} = \left(1.2, -2.6, 2.8\right)^T, \ \alpha_3 = \max(x^{(3)}) = 2.8,$$

$$\alpha_4 = 2.9285714,$$

$$\alpha_5 = 2.9756097,$$

$$\alpha_6 = 2.9918619,$$

$$\alpha_7 = 2.9972799,$$

$$\alpha_8 = 2.9990924,$$

$$\alpha_9 = 2.9996973.$$

$$\because \left|\alpha_9 - \alpha_8\right| = \left|2.9996973 - 2.99909241\right| = 0.0006049 < 10^{-3}$$

$$\therefore \lambda_1 \approx \alpha_9 = 2.9996973.$$

```
function y = maxa(x)
k=1;n=length(x);
for i=2:n
   if abs(x(i))>abs(x(k))
       k=i
   end
end
y=x(k);

A=[2,-1,0;0,2,-1;0,-1,2];
x0=[0;0;1];
y=x0/maxa(x0);
x1=A*y;
while (abs(maxa(x1)-maxa(x0)))>0.001
   x0=x1;
   y=x0/maxa(x0);
   x1=A*y;
end;
y
maxa(x1)
```

命令行窗口

```
>> eigex

y =

    0.9220
   -0.9997
    1.0000


ans =

    2.9997
```

例 用幂法求矩阵 $A=\begin{bmatrix} 2 & 4 & 6 \\ 3 & 9 & 15 \\ 4 & 16 & 36 \end{bmatrix}$

的最大特征值及其对应的特征向量

解:

```
A=[2,4,6;3,9,15;4,16,36];
[lambda,x,~,iter]=pow(A)
```

```
lambda =
  43.8800
x =
   0.1859
   0.4460
   1.0000
iter =
   10
```

➢ 反幂法

反幂法 用于计算矩阵按模最小的特征值及其特征向量，也可用来计算对应于一个给定近似特征值的特征向量, 是目前求特征向量最有效的方法.

反幂法

设A为n阶实可逆矩阵，其特征值满足

$$|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_{n-1}| > |\lambda_n|$$

对应的特征向量分别 $u_1, u_2, \ldots, u_n$, 则A^{-1}的特征值满足

$$\frac{1}{|\lambda_n|} > \frac{1}{|\lambda_{n-1}|} \geq \cdots \geq \frac{1}{|\lambda_2|} \geq \frac{1}{|\lambda_1|},$$

对应的特征向量分别 $u_n, u_{n-1}, \ldots, u_2, u_1$.

- 反幂法：计算λ_n以及相应的特征向量.

反幂法基本思想

- 对于A^{-1}应用幂法迭代，可求得矩阵A^{-1}的主特征值$1/\lambda_n$，从而求得A的按模最小的特征值λ_n.

反幂法迭代公式为

- 任取初始向量$x^{(0)}=y^{(0)}\neq 0$，构造向量序列

$$\begin{cases} x^{(k+1)} = A^{-1}y^{(k)} \\ y^{(k+1)} = \dfrac{x^{(k+1)}}{\max(x^{(k+1)})} \end{cases} \quad (k=0,1,2,\cdots).$$

- 迭代向量$x^{(k+1)}$可以通过解方程组求得

$$Ax^{(k+1)} = y^{(k)}$$

- 当k充分大时 $\begin{cases} \max(x_k) \approx \dfrac{1}{\lambda_n} \\ y^{(k)} \approx \dfrac{u_n}{\max(u_n)} \end{cases}$

定理 设A为非奇异矩阵且有n个线性无关的特征向量，其对应的特征值满足

$$|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_{n-1}| > |\lambda_n| > 0,$$

则对任何初始非零向量$x^{(0)}$ ($\alpha_n \neq 0$)，由反幂法构造的向量序列$\{x^{(k)}\}$, $\{y^{(k)}\}$满足

$$(1)\quad \lim_{k\to\infty} y^{(k)} = \frac{u_n}{\max(u_n)}$$

$$(2)\quad \lim_{k\to\infty} \max(x^{(k)}) = \frac{1}{\lambda_n}.$$

- 收敛速度比值为 $\left|\dfrac{\lambda_n}{\lambda_{n-1}}\right|$.

例：利用反幂法求解矩阵 $\begin{pmatrix} 2 & 4 & 6 \\ 3 & 9 & 15 \\ 4 & 16 & 36 \end{pmatrix}$

按模最小特征值及其对应的特征向量

解:

```
A=[2,4,6;3,9,15;4,16,36];
[lambda,x,~,iter]=antipow(A)
```

```
lambda =
   0.4025
x =
   1.0000
  -0.7085
   0.2061
iter =
   13
```

应用-----主成分分析

什么是主成分分析?

主成分分析（Principal Components Analysis ， PCA)也称为主分量分析，是一种通过降维来简化数据结构的方法，即如何把多个变量转化为少数几个综合变量，而这几个综合变量可以反映原来多个变量的大部分信息。

主成分分析的步骤

设有 n 个样品，每个样品观测 p 个指标，将原始数据写成矩阵

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

1．将原始数据标准化。

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，需要对原始数据进行标准化．即令

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

2．建立变量的相关系数阵：

$$R=\left(r_{ij}\right)_{p\times p} \qquad r_{ij}=\frac{\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)^2\sum_{k=1}^{n}(x_{kj}-\bar{x}_j)^2}}$$

3．求R的特征根 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p>0$ 及相应的单位特征向量：

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

4. 写出主成分

$$F_i = a_{1i}X_1 + a_{2i}X_2 + \cdots + a_{pi}X_p \quad i=1, \cdots, p$$

计算主成分贡献率及累计贡献率

- ✓ 贡献率

$$\frac{\lambda_i}{\sum_{k=1}^{p}\lambda_k} \quad (i = 1,2,\cdots,p)$$

- ✓ 累计贡献率

$$\frac{\sum_{k=1}^{i}\lambda_k}{\sum_{k=1}^{p}\lambda_k} \quad (i = 1,2,\cdots,p)$$

选择主成分的方法(1)

☯贡献率：第i 个主成分的贡献率为

$$r_i = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}$$

☯累积贡献率：前m个主成分的累积贡献率为

(Cumulative)

$$\eta_m = r_1 + r_2 + \cdots + r_m$$

☯选择法则：$\eta_m \geq 85\%$

☯保留m 个主成分

选择主成分的方法(2)

- 特征值大于1原则

若

$$\begin{cases} \lambda_m \geq 1 \\ \lambda_{m+1} < 1 \end{cases}$$

则保留m个主成分

实例演示

例1 对全国30个省市自治区经济发展基本情况的八项指标作主成分分析，原始数据如下：

省份	GDP X_1	居民消费水平 X_2	固定资产投资 X_3	职工平均工资 X_4	货物周转量 X_5	居民消费价格指数 X_6	商品零售价格指数 X_7	工业总产值 X_8
北京	1394.89	2505	519.01	8144	373.9	117.3	112.6	843.43
天津	920.11	2720	345.46	6501	342.8	115.2	110.6	582.51
河北	2849.52	1258	704.87	4839	2033.3	115.2	115.8	1234.85
山西	1092.48	1250	290.9	4721	717.3	116.9	115.6	697.25
内蒙	832.88	1387	250.23	4134	781.7	117.5	116.8	419.39
辽宁	2793.37	2397	387.99	4911	1371.1	116.1	114	1840.55
吉林	1129.2	1872	320.45	4430	497.4	115.2	114.2	762.47
黑龙江	2014.53	2334	435.73	4145	824.8	116.1	114.3	1240.37
上海	2462.57	5343	996.48	9279	207.4	118.7	113	1642.95
江苏	5155.25	1926	1434.95	5943	1025.5	115.8	114.3	2026.64

续表

省份	GDP X_1	居民消费水平 X_2	固定资产投资 X_3	职工平均工资 X_4	货物周转量 X_5	居民消费价格指数 X_6	商品零售价格指数 X_7	工业总产值 X_8
浙江	3524.79	2249	1006.39	6619	754.4	116.6	113.5	916.59
安徽	2003.58	1254	474	4609	908.3	114.8	112.7	824.14
福建	2160.52	2320	553.97	5857	609.3	115.2	114.4	433.67
江西	1205.11	1182	282.84	4211	411.7	116.9	115.9	571.84
山东	5002.34	1527	1229.55	5145	1196.6	117.6	114.2	2207.69
河南	3002.74	1034	670.35	4344	1574.4	116.5	114.9	1367.92
湖北	2391.42	1527	571.68	4685	849	120	116.6	1220.72
湖南	2195.7	1408	422.61	4797	1011.8	119	115.5	843.83
广东	5381.72	2699	1639.83	8250	656.5	114	111.6	1396.35
广西	1606.15	1314	382.59	5105	556	118.4	116.4	554.97

续表

省份	GDP X_1	居民消费水平 X_2	固定资产投资 X_3	职工平均工资 X_4	货物周转量 X_5	居民消费价格指数 X_6	商品零售价格指数 X_7	工业总产值 X_8
海南	364.17	1814	198.35	5340	232.1	113.5	111.3	64.33
四川	3534	1261	822.54	4645	902.3	118.5	117	1431.81
贵州	630.07	942	150.84	4475	301.1	121.4	117.2	324.72
云南	1206.68	1261	334	5149	310.4	121.3	118.1	716.65
西藏	55.98	1110	17.87	7382	4.2	117.3	114.9	5.57
陕西	1000.03	1208	300.27	4396	500.9	119	117	600.98
甘肃	553.35	1007	114.81	5493	507	119.8	116.5	468.79
青海	165.31	1445	47.76	5753	61.6	118	116.3	105.8
宁夏	169.75	1355	61.98	5079	121.8	117.1	115.3	114.4
新疆	834.57	1469	376.95	5348	339	119.7	116.7	428.76

数据来源：1996 年《中国统计年鉴》。

第一步 将原始数据标准化。

第二步 建立指标之间的相关系数阵R如下

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	1.000	.267	.951	.191	.617	−.274	−.264	.874
X_2	.267	1.000	.426	.718	−.151	−.234	−.593	.363
X_3	.951	.426	1.000	.400	431	−.282	−.359	.792
X_4	.191	.718	.400	1.000	−.356	−.134	−.539	.104
X_5	.617	−.151	.431	−.356	1.000	−.255	.022	.659
X_6	−.274	−.234	−.282	−.134	−.255	1.000	.760	−.126
X_7	−.264	−.593	−.359	−.539	.022	.760	1.000	−.192
X_8	.874	.363	.792	−.104	.659	−.126	−.192	1.000

第三步 求R的特征值和特征向量。

主成分	特征值	方差贡献率	累计贡献率
1	3.755	46.943	46.943
2	2.195	27.443	74.386
3	1.214	15.178	89.564
4	.403	5.033	94.596
5	.213	2.660	97.256
6	.139	1.737	98.993
7	6.594E-02	.824	99.817
8	1.462E-02	.183	100.000

从上表看，前3个特征值累计贡献率已达89．564%，说明前3个主成分基本包含了全部指标具有的信息，我们取前3个特征值，并计算出相应的特征向量：

第一特征向量 a_1	第二特征向量 a_2	第三特征向量 a_3
0.470641	0.107995	0.19241
0.456708	0.258512	0.109819
0.424712	0.287536	0.19241
−0.31944	0.400931	0.397525
0.312729	−0.40431	0.24505
0.250802	0.498801	−0.24777
0.240481	−0.48868	0.332179
−0.26267	0.167392	0.723351

因而前三个主成分为：

第一主成分：

$$F_1 = 0.470641X_1 + 0.456708X_2 + 0.424712X_3 - 0.31944X_4 + 0.312729X_5 + 0.250802X_6 + 0.240481X_7 - 0.26267X_8$$

第二主成分：

$$F_2 = 0.107995X_1 + 0.258512X_2 + 0.287536X_3 + 0.400931X_4 - 0.40431X5 + 0.498801X_6 - 0.48868X_7 + 0.167392X_8$$

第三主成分：

$$F_3 = 0.19241X_1 + 0.109819X_2 + 0.19241X_3 + 0.397525X_4 + 0.24505X_5 - 0.24777X_6 + 0.332179X_7 + 0.723351X_8$$

在第一主成分的表达式中第一、二、三项指标的系数较大，这三个指标起主要作用，我们可以把第一主成分看成是由国内生产总值、固定资产投资和居民消费水平所该划的反映经济发展状况的综合指标；

在第二主成分中，第四、五、六、七项指标的影响大，且第六、七项指标的影响尤其大，可将之看成是反映物价指数、职工工资和货物周转量的综合指标；

在第三主成分中，第八项指数影响最大，远超过其它指标的影响，可单独看成是工业总产值的影响。

在**MATLAB**中，运用样本数据矩阵进行主成分分析的命令为**princomp**，调用格式：

①**PC=pca(X)**

②**[PC,SCORE,latent,tsquare]=pca(X)**

其中输入**X**是样本数据矩阵，输出**PC**为主成分矩阵，**SCORE**是样本主成分的得分，**latent**是**X**的方差矩阵的特征值向量，**tsquare**是每个数据点的**HotellingT2**统计量。

例2　影响我国粮食安全生产的主要因素有以下几个方面: 有效灌溉面积，粮食播种面积，成灾面积，财政投入，农业劳动力 ，农村用电量，农业机械总动力 及农业化肥施用量，具体数据如表2所示。对粮食安全生产因素作主成分分析。

表2 影响我国粮食安全生产的主要因素表

年份	有效灌溉面积/万公顷	粮食播种面积/万公顷	成灾面积/万公顷	财政投入/万元	农业劳动力/万人	农村用电量/万千瓦	农机总动力/万千瓦	化肥施用量/万千千克
1990	4740.31	11346.60	178.20	221.76	33336.00	844.50	28707.70	647.58
1991	4782.21	11231.40	278.10	243.55	34186.30	963.20	29388.60	701.28
1992	4859.01	11056.00	259.00	269.04	34037.00	1106.90	30308.40	732.55
1993	4872.79	11050.90	231.30	323.42	33258.20	1244.80	31816.60	787.98
1994	4875.91	10854.40	313.80	399.70	32690.30	1473.90	33802.50	829.53
1995	4928.12	11006.00	222.70	430.22	32335.00	1655.70	36118.10	898.40
1996	5038.14	11254.80	212.30	510.07	32260.40	1812.70	38546.90	957.00
1997	5123.85	11291.20	303.10	560.77	32434.90	1980.10	42015.60	995.23
1998	5229.56	11378.70	251.80	626.02	32626.40	2042.10	45207.70	1020.88
1999	5315.80	11316.10	267.30	677.46	32911.80	2173.40	48996.10	1031.08
2000	5382.00	10846.30	343.70	766.89	32798.00	2421.30	52573.60	1036.63
2001	5424.90	10608.00	317.90	917.96	32451.00	2610.80	55172.10	1063.28
2002	5435.50	10389.10	271.60	1102.70	31991.00	2993.40	57929.90	1083.08
2003	5401.42	9941.00	325.20	1134.86	31259.60	3432.90	60386.50	1102.90
2004	5447.80	10160.60	163.00	1693.79	30596.00	3933.00	64027.90	1157.30
2005	5502.93	10427.80	199.70	1792.40	29975.50	4375.70	68397.80	1191.45
2006	5575.05	10495.80	246.30	2161.35	28886.35	4895.80	72522.10	1231.90
2007	5651.83	10563.80	250.60	3404.70	22543.4	5509.90	76589.60	1276.70
2008	5847.17	10679.30	222.80	4544.01	20078.6	5713.20	82190.41	1309.19

解：由于各个指标的单位不同，且各指标的方差相差很大，所以首先对样本数据进行无量纲的变换，然后对标准化的样本数据进行主成分分析。程序如下：

```
%将表2中各指标值作为矩阵X输入，省略号表示数据省略了
X=[4740.31, 11346.60, 178.20, 221.76, 33336.00, 844.50, 28707.70, 647.58;……5847.17, 10679.30, 222.80, 4544.01, 20078.6, 5713.20, 82190.41,1309.19];

X1=zscore(X); %对样本数据标准化
[pc,la,tent]=princomp(X1);   %主成分分析，pc是特征向量矩阵，la得分矩阵，tent特征值
tents=sum(tent);            %特征值总和
gxl= tent/ tents;            %各个主成分贡献率
```

输出结果

pc = %正交单位化特征向量

0.3933	-0.1518	-0.0544	0.3944	0.5494	0.3701	-0.1396	0.4533
-0.2821	0.3173	-0.7669	0.4403	0.0325	-0.1907	-0.0141	-0.0121
-0.0479	-0.8701	-0.4379	-0.1998	-0.0337	-0.0838	0.0257	-0.0115
0.3856	0.1952	-0.2572	-0.3364	0.2998	0.0210	0.7181	-0.1668
-0.3605	-0.2487	0.3850	0.4799	0.2640	-0.3102	0.5109	-0.0512
0.4083	0.0358	0.0265	-0.0014	-0.2123	-0.7233	0.0088	0.5128
0.4070	-0.0643	0.0422	0.1814	0.2707	-0.3433	-0.3636	-0.6872
0.3905	-0.1175	-0.0151	0.4841	-0.6466	0.2861	0.2659	-0.1692

gxl = 0.7388
0.1416
0.0765
0.0355
0.0043
0.0027
0.0005
0.0001 %样本各主成分的贡献率

由于第一、二主成分的累计贡献率为**88.04%**。可取前两个主成分，即

$$y_1 = e_1^T (x_1^*, x_2^*, x_3^*, x_4^*, x_5^*, x_6^*, x_7^*, x_8^*)^T$$

$$y_2 = e_2^T (x_1^*, x_2^*, x_3^*, x_4^*, x_5^*, x_6^*, x_7^*, x_8^*)^T$$

其中

e_1^T=(0.3933,-0.2821,-0.0479,0.3856,-0.3605,0.4083,0.4070,0.3905),

e_2^T=(-0.1518,0.3173,-0.8701,0.1952,-0.2487,0.0358,-0.0643, -0.1175).

第一主成分反映了我国在农业生产基础设施投入情况，第二主成分反映了我国粮食生产抗灾能力与劳动力情况。

- 主成分的实际意义，要结合具体问题和有关专业知识才能给出合理的解释。虽然利用主成分本身可对所研究的问题在一定程度上做分析，但主成分分析往往并不是最终目的，通常是将主成分综合原始数据信息，达到降低原始数据维数的目的，进而将少数几个主成分的得分做为新数据，对其再做进一步分析，如：基于主成分的回归分析、聚类分析等。

主成分分析用于综合评价的一般步骤

① 若各指标的属性不同(成本型、利润型、适度型)，则将原始数据矩阵A统一标准化，得到属性一致的指标矩阵B.

② 计算指标矩阵的相关系数矩阵R.

③ 计算R的特征值与相应的特征向量.

④ 根据特征值计算累计贡献率，确定主成分的个数，而特征向量就是主成分的系数向量。

⑤ 计算主成分的数值（即主成分得分）

主成分分析用于综合评价的一般步骤

利用相关系数矩阵**R**计算特征值与特征向量，则主成分得分为：

$$F = B^{*}V$$

其中，V是特征向量矩阵，B^*是将矩阵标准化以后的矩阵（即**zscore(B)**）

⑥ 计算综合评价值，进行排序．若为效益型矩阵，则评价值越大排名越靠前；若为成本型矩阵，则评价值越小排名越靠前**.**

通常计算综合评价值的公式为**:Z=FW**

其中**F**是主成分得分矩阵，**W**是将特征值归一化后得到的权向量**.**

例3 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。

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

解：首先输入数据

```
A=[data];                    % data即表中数据
R=corrcoef(A);
```

得到的相关系数矩阵为：

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

由于$r_{12}=r_{21}=1$，表明指标x_1,x_2完全线性相关，故只需保留一个指标x_2.

```
A1=A(:,2:6)./[ones(17,1)*std(A(:,2:6))];    % 消除量纲
[v,d]=eig(corrcoef(A1));                    % 计算特征值与特征向量
w=sum(d)/sum(sum(d));                       % 计算贡献率
F=[A1-ones(17,1)*mean(A1)]*d(:,5);% 计算主成分得分
[F1,I1]=sort(F,'descend');                  % I1各名次的序号
[F2,I2]=sort(I1);                           % I2给出各市排名
```

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)	0.9220
0.2475	(-0.2517, -0.2103, 0.9054,-0.1315,-0.2354)	0.0495
0.1050	(0.1926, 0.3702, -0.0390, 0.3029, -0.8559)	0.0210
0.0322	(-0.3510, 0.7779, 0.0275, -0.5153, 0.0738)	0.0064
0.0053	(0.7518, -0.0803, 0.0719, -0.6434, -0.0965)	0.0011

各市第一主成分得分排名

地　　区	得分	排　名	地　　区	排　名	得分	地　　区	得分	排　名
合　肥	15.4075	1	淮　南	5	0.5295	宣　城	-2.1845	11
淮　北	1.3498	4	滁　州	10	-0.8388	铜　陵	-0.6297	8
亳　州	-2.5201	12	六　安	7	-0.2293	池　州	-2.9935	14
宿　州	-4.1769	17	马鞍山	2	4.7641	安　庆	-3.8430	16
蚌　埠	-2.6984	13	巢　湖	9	-0.7853	黄　山	-3.5041	15
阜　阳	0.3236	6	芜　湖	3	2.0291			

总 结

主成分分析法通过研究指标体系的内在结构关系，从而将多个指标转化为少数几个相互独立且包含原来指标大部分信息（80%或85%以上）的综合指标。其优点在于它确定的权数是基于数据分析而得出的指标之间的内在结构关系，不受主观因素的影响，有较好的客观性，而且得出的综合指标（主成分）之间相互独立，减少信息的交叉，这对分析评价极为有利。

总 结

主成分分析的应用范围非常广泛，诸如投资组合风险管理、企业效益的综合评价、图像特征识别、机械加工或传感器故障检测、灾害损失分析等. 将主成分分析与聚类分析、判别分析以及回归分析方法相结合，还可以解决更多实际问题.

实验题

1. 根据 1998 年部分地区洪灾损失数据（见表 1）进行主成分分析，看看哪些省受灾较轻？受灾最重的是哪几个省？其中 x1—x12 分别为：受灾面积、成灾面积、绝收面积、受灾/万人次、成灾/万人次、死亡（人）、伤病（人）、紧急转移（人）、倒塌房屋/万间、损坏房屋/万间、死亡大牲畜/万头、直接经济损失/亿元。

表 1　1998 年部分地区洪灾损失指标数据

地区	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12
蒙	130.4	107.7	64.8	448.0	375.0	147	105476	97.7	37.0	59.0	36.0	164.0
吉	109.7	64.7	26.7	306.1	214.5	7	311000	98.3	56.0	62.3	14.8	140.0
黑	242.9	160.6	93.7	581.0	521.0	2	316844	156.4	82.0	75.0	16.1	218.0
皖	199.6	130.8	57.1	1562.0	1012.8	93	262461	100.4	26.9	49.3	1.0	130.5
闽	69.7	23.7	4.4	597.5	455.8	146	8402	195.3	68.9	167.8	100.0	87.9
赣	241.6	193.6	87.8	2381.0	1702.6	237	126505	304.6	117.9	168.2	50.7	434.2
鄂	254.0	169.0	44.5	1939.0	1534.7	353	891200	247.4	86.9	241.2	27.0	357.0
湘	213.0	141.3	39.9	2178.0	1652.4	854	224400	350.8	93.9	224.5	83.9	422.8
贵	79.3	54.8	28.2	1378.5	992.8	315	38000	81.1	10.7	76.3	6.6	114.9
川	128.2	75.4	16.9	1757.9	1044.3	581	14806	37.0	21.2	43.1	13.1	74.7
渝	65.3	49.4	6.7	904.0	668.2	304	21715	39.1	17.1	27.1	4.5	55.5
滇	39.9	16.1	3.5	462.7	52.8	166	235	7.6	10.4	14.8	1.5	23.1
陕	40.8	26.2	4.8	650.0	475.0	215	3069	12.7	9.9	24.6	1.8	43.0

实验题

2. 国内 35 个大城市某年的 10 项社会经济统计指标数据见下表 2，对其进行主成分分析，给出各地区主成分得分及综合得分与排名。

城市名称	年底总人口(万人)	非农业人口比(%)	农业总产值(万元)	工业总产值(万元)	客运总量(万人)	货运总量(万吨)	地方财政预算内收入(万元)	城乡居民年底储蓄余额(万元)	在岗职工人数(万人)	在岗职工工资总额(万元)
北京	1 249.90	0.597 8	1 843 427	19 999 706	20 323	45 562	2 790 863	26 806 646	410.80	5 773 301
天津	910.17	0.580 9	1 501 136	22 645 502	3 259	26 317	1 128 073	11 301 931	202.68	2 254 343
石家庄	875.40	0.233 2	2 918 680	6 885 768	2 929	1 911	352 348	7 095 875	95.60	758 877
太原	299.92	0.656 3	236 038	2 737 750	1 937	11 895	203 277	3 943 100	88.65	654 023
呼和浩特	207.78	0.441 2	365 343	816 452	2 351	2 623	105 783	1 396 588	42.11	309 337
沈阳	677.08	0.629 9	1 295 418	5 826 733	7 782	15 412	567 919	9 016 998	135.45	1 152 811
大连	545.31	0.494 6	1 879 739	8 426 385	10 780	19 187	709 227	7 556 796	94.15	965 922
长春	691.23	0.406 8	1 853 210	5 966 343	4 810	9 532	357 096	4 803 744	102.63	884 447
哈尔滨	927.09	0.462 7	2 663 855	4 186 123	6 720	7 520	481 443	6 450 020	172.79	1 309 151
上海	1 313.12	0.738 4	2 069 019	54 529 098	6 406	44 485	4 318 500	25 971 200	336.84	5 605 445
南京	537.44	0.534 1	989 199	13 072 737	14 269	11 193	664 299	5 680 472	113.81	1 357 861
杭州	616.05	0.355 6	1 414 737	12 000 796	17 883	11 684	449 593	7 425 967	96.90	1 180 947
宁波	538.41	0.254 7	1 428 235	10 622 866	22 215	10 298	501 723	5 246 350	62.15	824 034
合肥	429.95	0.318 4	628 764	2 514 125	4 893	1 517	233 628	1 622 931	47.27	369 577
福州	583.13	0.273 3	2 152 288	6 555 351	8 851	7 190	467 524	5 030 220	69.59	680 607

实验题

厦　门	128.99	0.486 5	333 374	5 751 124	3 728	2 570	418 758	2 108 331	46.93	657 484
南　昌	424.20	0.398 8	688 289	2 305 881	3 674	3 189	167 714	2 640 460	62.08	479 ,555
济　南	557.63	0.408 5	1 486 302	6 285 882	5 915	11 775	460 690	4 126 970	83.31	756 696
青　岛	702.97	0.369 3	2 382 320	11 492 036	13 408	17 038	658 435	4 978 045	103.52	961 704
郑　州	615.36	0.342 4	677 425	5 287 601	10 433	6 768	387 252	5 135 338	84.66	696 848
武　汉	740.20	0.586 9	1 211 291	7 506 085	9 793	15 442	604 658	5 748 055	149.20	1 314 766
长　沙	582.47	0.310 7	1 146 367	3 098 179	8 706	5 718	323 660	3 461 244	69.57	596 986
广　州	685.00	0.621 4	1 600 738	23 348 139	22 007	23 854	1 761 499	20 401 811	182.81	3 047 594
深　圳	119.85	0.793 1	299 662	20 368 295	8 754	4 274	1 847 908	9 519 900	91.26	1 890 338
南　宁	285.87	0.406 4	720 486	1 149 691	5 130	3 293	149 700	2 190 918	45.09	371 809
海　口	54.38	0.835 4	44 815	717 461	5 345	2 356	115 174	1 626 800	19.01	198 138
重　庆	3 072.34	0.206 7	4 168 780	8 585 525	52 441	25 124	898,912	9 090 969	223.73	1 606 804
成　都	1 003.56	0.335	1 935 590	5 894 289	40 140	19 632	561 189	7 479 684	132.89	1 200 671
贵　阳	321.50	0.455 7	362 061	2 247 934	15 703	4 143	197 908	1 787 748	55.28	419 681
昆　明	473.39	0.386 5	793 356	3 605 729	5 604	12 042	524 216	4 127 900	88.11	842 321
西　安	674.50	0.409 4	739 905	3 665 942	10 311	9 766	408 896	5 863 980	114.01	885 169
兰　州	287.59	0.544 5	259 444	2 940 884	1 832	4 749	169 540	2 641 568	65.83	550 890
西　宁	133.95	0.522 7	65 848	711 310	1 746	1 469	49 134	855 051	27.21	219 251
银　川	95.38	0.570 9	171 603	661 226	2 106	1 193	74 758	814 103	23.72	178 621
乌鲁木齐	158.92	0.824 4	78 513	1 847 241	2 668	9 041	254 870	2 365 508	55.27	517 622

Q & A

- 有什么问题吗？