第 13 讲　多元统计分析

司守奎
烟台市，海军航空大学
Email：sishoukui@163.com

13.1　聚类分析

聚类分析，亦称群分析或点群分析，它是研究多要素事物分类问题的数量方法。其基本原理是，根据样本自身的属性，用数学方法按照某些相似性或差异性指标，定量地确定样本之间的亲疏关系，并按这种亲疏关系程度对样本进行分类。

常见的聚类分析方法有系统聚类法、动态聚类法和模糊聚类法等。对样本进行分类称为 Q 型聚类分析，对指标进行分类称为 R 型聚类分析。

13.1.1　Q 型聚类分析

设有 n 个样品，每个样品测得 p 项指标（变量），原始数据阵为

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

其中 a_{ij}（$i=1,\cdots,n$；　$j=1,\cdots,p$）为第 i 个样品 ω_i 的第 j 个指标的观测数据。

1. 数据的变换处理

由于样本数据矩阵由多个指标组成，不同指标一般有不同的量纲，为消除量纲的影响，通常需要进行数据变换处理。

常用的数据变换方法有：

（1）中心化处理

中心化变换是一种坐标轴平移处理方法，它是先求出每个变量的样本平均值，再从原始数据中减去该变量的均值，就得到中心化变换后的数据。

设变换后的数据为 $\tilde{a}_{ij}$，则有

$$\tilde{a}_{ij}=a_{ij}-\mu_j,\ （i=1,\cdots,n；\ \ j=1,\cdots,p）$$

其中 $\mu_j=\dfrac{\sum\limits_{i=1}^{n}a_{ij}}{n}$。

（2）规格化变换

规格化变换是从数据矩阵的每一个变量中找出其最大值和最小值，这两者之差称为极差，然后从每个变量的原始数据中减去该变量中的最小值，再除以极差，就得到规格化数据，即有

$$\tilde{a}_{ij}=\frac{a_{ij}-\min\limits_{1\le i\le n}(a_{ij})}{\max\limits_{1\le i\le n}(a_{ij})-\min\limits_{1\le i\le n}(a_{ij})}\ （i=1,\cdots,n；\ \ j=1,\cdots,p）.$$

（3）标准化变换

首先对每个变量进行中心化变换，然后用该变量的标准差进行标准化，即有

$$\tilde{a}_{ij}=\frac{a_{ij}-\mu_j}{s_j}\ （i=1,\cdots,n；\ \ j=1,\cdots,p），$$

其中 $\mu_j=\dfrac{\sum_{i=1}^{n}a_{ij}}{n}$，$s_j=\sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}(a_{ij}-\mu_j)^2}$。

13.1.2　样品（或指标）间亲疏程度的测度计算

第 i 个样品 ω_i 由矩阵 A 的第 i 行所描述，所以任何两个样品 ω_k 与 ω_m 之间的相似性，都可以通过矩阵 A 中的第 k 行与第 m 行的相似程度来描述；任何两个变量 x_k 与 x_m 之间的相似性，可以通过第 k 列与第 m 列的相似程度来描述。

研究样品或变量的亲疏程度或相似程度的数量指标通常有两种：一种是相似系数，性质越接近的变量或样品，其取值越接近于 1 或-1，而彼此无关的变量或样品的相似系数则越接近于 0，相似的归为一类，不相似的归为不同类。另一种是距离，它将每个样品看成 p 维空间的一个点，n 个样品组成 p 维空间的 n 个点。用各点之间的距离来衡量各样品之间的相似程度（或靠近程度）。距离近的点归为一类，距离远的点属于不同的类。对于变量之间的聚类（R 型）常用相似系数来测度变量之间的亲疏程度，而对于样品之间的聚类分析，则常用距离来测度样品之间的亲疏程度。

1. 常用距离的计算

令 d_{ij} 表示样品 ω_i 与 ω_j 的距离。常用的距离有

（1）闵氏（Minkowski）距离

$$d_{ij}(q)=\left(\sum_{k=1}^{p}\left|a_{ik}-a_{jk}\right|^{q}\right)^{1/q}.$$

当 $q=1$ 时，

$$d_{ij}(1)=\sum_{k=1}^{p}\left|a_{ik}-a_{jk}\right|$$，即绝对值距离。

当 $q=2$ 时，

$$d_{ij}(2)=\left(\sum_{k=1}^{p}(a_{ik}-a_{jk})^2\right)^{1/2}$$，即欧氏距离。

当 $q=\infty$ 时，

$$d_{ij}(\infty)=\max_{1\le k\le p}\left|a_{ik}-a_{jk}\right|$$，即车比雪夫距离。

（2）马氏（Mahalanobis）距离

马氏距离是由印度统计学家马哈拉诺比斯于 1936 年定义的，故称为马氏距离。其计算公式为

$$d_{ij}=\sqrt{(A_i-A_j)\Sigma^{-1}(A_i-A_j)^T}$$，

这里 A_i 表示矩阵 A 的第 i 行，Σ 表示观测变量之间的协方差阵。

$$\Sigma=(\sigma_{ij})_{p\times p}$$，

其中，$\sigma_{ij}=\dfrac{1}{n-1}\sum_{k=1}^{n}(a_{ki}-\mu_i)(a_{kj}-\mu_j)$。

2. 相似系数的计算

研究样品（或变量）之间的关系，除了用距离表示外，还有相似系数。相似系数是描述样品（或变量）之间相似程度的一个统计量，常用的相似系数有：

（1）夹角余弦

将任何两个样品 ω_i 与 ω_j 看成 p 维空间的两个向量，这两个向量的夹角余弦用 $\cos\theta_{ij}$ 表示，则

$$\cos\theta_{ij}=\frac{\sum_{k=1}^{p}a_{ik}a_{jk}}{\sqrt{\sum_{k=1}^{p}a_{ik}^2}\cdot\sqrt{\sum_{k=1}^{p}a_{jk}^2}}.$$

当$\cos\theta_{ij}=1$时，说明两个样品ω_i与ω_j完全相似；$\cos\theta_{ij}$接近 1 时，说明ω_i与ω_j相似密切；$\cos\theta_{ij}=0$时，说明ω_i与ω_j完全不一样；$\cos\theta_{ij}$接近 0 时，说明ω_i与ω_j差别大。把所有两两样品的相似系数都计算出来，可排成相似系数矩阵

$$\Theta=\begin{bmatrix}\cos\theta_{11} & \cos\theta_{12} & \cdots & \cos\theta_{1n}\\ \cos\theta_{21} & \cos\theta_{22} & \cdots & \cos\theta_{2n}\\ \vdots & \vdots & & \vdots\\ \cos\theta_{n1} & \cos\theta_{n2} & \cdots & \cos\theta_{nn}\end{bmatrix},$$

其中$\cos\theta_{11}=\cdots=\cos\theta_{nn}=1$。根据$\Theta$可对$n$个样品进行分类，把比较相似的样品归为一类，不怎么相似的样品归为不同的类。

（2）皮尔逊相关系数

通常所说的相关系数，一般指变量间的相关系数，为了描述样品间的相似关系，也可类似给出定义，即第i个样品与第j个样品之间的相关系数定义为

$$r_{ij}=\frac{\sum_{k=1}^{p}(a_{ik}-\bar{\mu}_i)(a_{jk}-\bar{\mu}_j)}{\sqrt{\sum_{k=1}^{p}(a_{ik}-\bar{\mu}_i)^2}\cdot\sqrt{\sum_{k=1}^{p}(a_{jk}-\bar{\mu}_j)^2}},$$

其中，$\bar{\mu}_i=\dfrac{\sum_{k=1}^{p}a_{ip}}{p}$。

实际上，r_{ij}就是两个向量$A_i-\bar{A}_i$与$A_j-\bar{A}_j$的夹角余弦，其中$\bar{A}_i=\bar{\mu}_i[1,\ \cdots,\ 1]$。若将原始数据标准化，则$\bar{A}_i=\bar{A}_j=0$，这时$r_{ij}=\cos\theta_{ij}$。

$$R=(r_{ij})_{n\times n}=\begin{bmatrix}r_{11} & r_{12} & \cdots & r_{1n}\\ r_{21} & r_{22} & \cdots & r_{2n}\\ \vdots & \vdots & & \vdots\\ r_{n1} & r_{n2} & \cdots & r_{nn}\end{bmatrix},$$

其中$r_{11}=\cdots=r_{nn}$，可根据R对n个样品进行分类。

13.1.3　基于类间距离的系统聚类

系统聚类法是聚类分析方法中使用最多的方法。其基本思想是：距离相近的样品（或变量）先聚为一类，距离远的后聚成类，此过程一直进行下去，每个样品总能聚到合适的类中。它包括如下步骤：

（1）将每个样品（或变量）独自聚成一类，构造n个类；

（2）根据所确定的样品（或变量）距离公式，计算n个样品（或变量）两两间的距离，构造距离矩阵，记为$D_{(0)}$。

（3）把距离最近的两类归为一新类，其它样品（或变量）仍各自聚为一类，共聚成$n-1$类。

（4）计算新类与当前各类的距离，将距离最近的两个类进一步聚成一类，共聚成$n-2$类。以上步骤一直进行下去，最后将所有的样品（或变量）聚成一类。

（5）画聚类谱系图。

（6）决定类的个数及各类包含的样品数，并对类做出解释。

正如样品之间的距离可以有不同的定义方法一样，类与类之间的距离也有各种定义。例如可以定义类与类之间的距离为两类之间最近样品的距离，或者定义为两类之间最远样品的距离，也可以定义为两类重心之间的距离等。类与类之间用不同的方法定义距离，就产生了不同的系统聚类方法。常用的系统聚类方法有，最短距离法、最长聚类法、中间距离法、重心法、类平均法、可变类平均法、可变法和离差平方和法。

1. 最短距离法

最短距离法定义类G_i与G_j之间的距离为两类间最邻近的两样品之距离，即G_i与G_j两类间的距离D_{ij}定义为

$$D_{ij}=\min_{\omega_i\in G_i,\omega_j\in G_j} d_{ij}.$$

设类G_p与G_q合并成一个新类记为G_r，则任一类G_k与G_r的距离是

$$D_{kr}=\min_{\omega_i\in G_k,\omega_j\in G_r} d_{ij}=\min\{\min_{\omega_i\in G_k,\omega_j\in G_p} d_{ij},\min_{\omega_i\in G_k,\omega_j\in G_q} d_{ij}\}$$
$$=\min\{D_{kp},D_{kq}\}.$$

最短距离法聚类的步骤如下：

（1）定义样品之间距离，计算样品两两距离，得一距离阵记为$D_{(0)}=(d_{ij})_{n\times n}$，开始每个样品自成一类，显然这时$D_{ij}=d_{ij}$。

（2）找出$D_{(0)}$的非对角线最小元素，设为D_{pq}，则将G_p和G_q合并成一个新类，记为G_r，即$G_r=\{G_p,G_q\}$。

（3）给出计算新类与其它类的距离公式：

$$D_{kr}=\min\{D_{kp},D_{kq}\}.$$

将$D_{(0)}$中第p、q行及p、q列，用上面公式合并成一个新行新列，新行新列对应G_r，所得到的矩阵记为$D_{(1)}$。

（4）对$D_{(1)}$重复上述类似$D_{(0)}$的（2）、（3）两步得到$D_{(2)}$。如此下去，直到所有的元素并成一类为止。

如果某一步$D_{(k)}$中非对角线最小的元素不止一个，则对应这些最小元素的类可以同时合并。

为了便于理解最短距离法的计算步骤，下面举一个简单例子。

例 13.1　设抽出 5 个样品，每个样品只测 1 个指标，它们是 2,3,3.5,7,9，试用最短距离法对 5 个样品进行分类。

解　（1）定义样品间距离采用欧氏距离，计算样品两两距离，得距离矩阵$D_{(0)}$，如表 13.1 所示。

表 13.1　$D_{(0)}$ **表**

	$G_1=\{\omega_1\}$	$G_2=\{\omega_2\}$	$G_3=\{\omega_3\}$	$G_4=\{\omega_4\}$	$G_5=\{\omega_5\}$
$G_1=\{\omega_1\}$	0	1	1.5	5	7
$G_2=\{\omega_2\}$	1	0	0.5	4	6
$G_3=\{\omega_3\}$	1.5	0.5	0	3.5	5.5
$G_4=\{\omega_4\}$	5	4	3.5	0	2
$G_5=\{\omega_5\}$	7	6	5.5	2	0

（2）找出$D_{(0)}$中非对角线最小元素是 0.5，即$D_{23}=D_{32}=0.5$，则将G_2与G_3合并成

一个新类，记为$G_6=\{\omega_2,\omega_3\}$。

（3）计算新类G_6与其它类的距离，按公式

$D_{i6}=\min\{D_{i2},D_{i3}\}$，$i=1,4,5$.

即将表$D_{(0)}$的第2,3列取较小的一列，第2,3行取较小的一行得表$D_{(1)}$，如表13.2所示。

表13.2　$D_{(1)}$表

	$G_1=\{\omega_1\}$	$G_6=\{\omega_2,\ \omega_3\}$	$G_4=\{\omega_4\}$	$G_5=\{\omega_5\}$
$G_1=\{\omega_1\}$	0	1	5	7
$G_6=\{\omega_2,\ \omega_3\}$	1	0	3.5	5.5
$G_4=\{\omega_4\}$	5	3.5	0	2
$G_5=\{\omega_5\}$	7	5.5	2	0

（4）找出$D_{(1)}$中非对角线最小元素是1，则将相应的两类G_1和G_6合并为$G_7=\{\omega_1,\omega_2,\omega_3\}$，然后再按公式计算各类与$G_7$的距离，即将$G_1,G_6$相应的两行两列归并为一行一列，新的行（列）由原来的两行（列）中较小的一个组成，计算结果得表$D_{(2)}$，如表13.3所示。

表13.3　$D_{(2)}$表

	$G_7=\{\omega_1,\omega_2,\omega_3\}$	$G_4=\{\omega_4\}$	$G_5=\{\omega_5\}$
$G_7=\{\omega_1,\omega_2,\omega_3\}$	0	3.5	5.5
$G_4=\{\omega_4\}$	3.5	0	2
$G_5=\{\omega_5\}$	5.5	2	0

（5）找出$D_{(2)}$中非对角线最小元素是2，则将G_4与G_5合并成$G_8=\{\omega_4,\omega_5\}$，最后再按公式计算$G_7$与$G_8$的距离，即将$G_4,G_5$相应的两行两列归并成一行一列，新的行列由原来的两行（列）中较小的一个组成，得表$D_{(3)}$，如表13.4所示。

表13.4　$D_{(3)}$表

	$G_7=\{\omega_1,\omega_2,\omega_3\}$	$G_8=\{\omega_4,\omega_5\}$
$G_7=\{\omega_1,\omega_2,\omega_3\}$	0	3.5
$G_8=\{\omega_4,\omega_5\}$	3.5	0

最后，将G_7和G_8合并成G_9，上述合并过程可用图13.1表达。纵坐标的刻度是并类的距离。

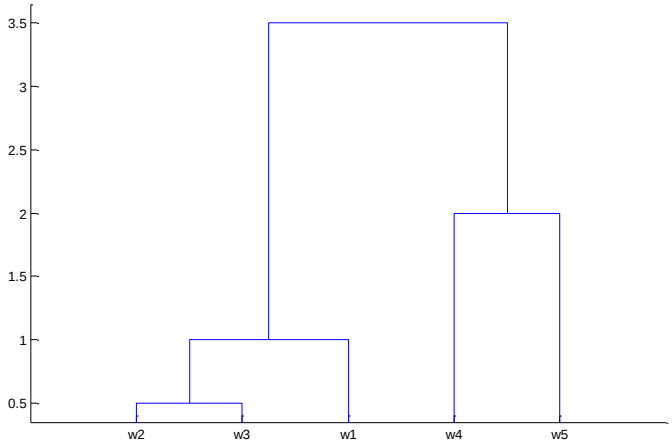

图 13.1　聚类图

由图 13.1 看到分成两类$\{\omega_1,\omega_2,\omega_3\}$及$\{\omega_4,\omega_5\}$比较合适。

计算及画聚类图的 MATLAB 程序如下

```
clc, clear
a=[2,3,3.5,7,9]';
x=pdist(a)　%求聚类对象两两之间的距离
xc=squareform(x)　%变换成距离方阵，方便观察对象之间的距离
y=linkage(x)　%产生聚类树
name1([1:5],1)='A'; name2=int2str([1:5]');
name=cellstr(strcat(name1,name2)); %构造标注名称的字符细胞数组
[h,t]=dendrogram(y,'label',name)　%画聚类图
n=input('请输入分类的类别数(输入后请回车)n=');
T=cluster(y,'maxclust',n)
```

最短距离法也可用于指标（变量）分类，分类时可以用距离，也可以用相似系数。但用相似系数时应找最大的元素并类，也就是把公式$D_{ik}=\min\{D_{ip},D_{iq}\}$中的 min 换成 max。

例 13.2　**表 13.5** 给出了某地区 9 个农业区的 7 项经济指标，试用最短距离法进行聚类分析。

表 13.5　某地区的经济指标数据

区代号	人均耕地 x_1	劳动耕地 x_2	水田比重 x_3	复种指数 x_4	粮食亩产 x_5	人均粮食 x_6	稻谷占粮食比 x_7
G_1	0.294	1.093	5.63	113.6	4510	1036	12.2
G_2	0.315	0.971	0.39	95.1	2773	683	0.85
G_3	0.123	0.316	5.28	148.5	6934	611	6.49
G_4	0.179	0.527	0.39	111	4458	632	0.92
G_5	0.081	0.212	72.04	217.8	12240	791	80.3
G_6	0.082	0.211	43.78	179.6	8973	636	48.1
G_7	0.075	0.181	65.15	194.7	1068	634	80.1
G_8	0.293	0.666	5.35	94.9	3679	771	7.8
G_9	0.167	0.414	2.9	94.8	4231	574	1.17

解　记区域G_i（$i=1,\cdots,9$）对应的指标变量x_j（$j=1,\cdots,7$）值为a_{ij}，构造数据矩阵$A=(a_{ij})_{9\times7}$。聚类分析的步骤如下：

（1）数据的规格化变换

为了消除不同指标变量的不同量纲之间的影响，对原始数据进行规格化变换。设变换后的数据为$\tilde{a}_{ij}$，变换公式为

$$\tilde{a}_{ij}=\frac{a_{ij}-\min\limits_{1\le i\le n}(a_{ij})}{\max\limits_{1\le i\le n}(a_{ij})-\min\limits_{1\le i\le n}(a_{ij})}\quad（i=1,\cdots,9；\ j=1,\cdots,7）.$$

（2）计算G_i间的两两之间的距离

由于数据已经进行了规格化处理，我们这里可以使用欧氏距离计算G_i与G_k之间的距离d_{ik}，计算公式为

$$d_{ik}=\left(\sum_{j=1}^{7}(a_{ij}-a_{kj})^2\right)^{1/2},\quad i,k=1,2,\cdots,9.$$

（3）最短距离法的聚类

最短距离法定义类G_i与G_k之间的距离为两类间最邻近的两样品之距离，即G_i与G_k两类间的距离D_{ik}定义为

$$D_{ik}=\min_{\omega_i\in G_i,\omega_k\in G_k}d_{ik}.$$

最短距离法的聚类过程如下：

i）开始每个样品自成一类，分成 9 类，显然这时$D_{ij}=d_{ij}$，构造距离矩阵$D_{(0)}=(D_{ij})_{9\times9}=(d_{ij})_{9\times9}$。

ii）找出$D_{(0)}$的非对角线最小元素，设为D_{pq}，则将G_p和G_q合并成一个新类，记为G_{10}，即$G_{10}=\{G_p,G_q\}$。

iii）给出计算新类$r=10$与其它类的距离公式

$$D_{kr}=\min\{D_{kp},D_{kq}\}.$$

将$D_{(0)}$中第p、q行及p、q列取最小值，合并成一个新行新列，新行新列对应G_{10}，所得到的矩阵记为$D_{(1)}$。

iv）对$D_{(1)}$重复上述类似$D_{(0)}$的 ii）和 iii）操作，得到$D_{(2)}$。如此下去，直到所有的样品并成一类为止。

如果某一步$D_{(k)}$中非对角线最小的元素不止一个，则对应这些最小元素的类可以同时合并。

利用 MATLAB 程序，画出的聚类图见图 13.2。

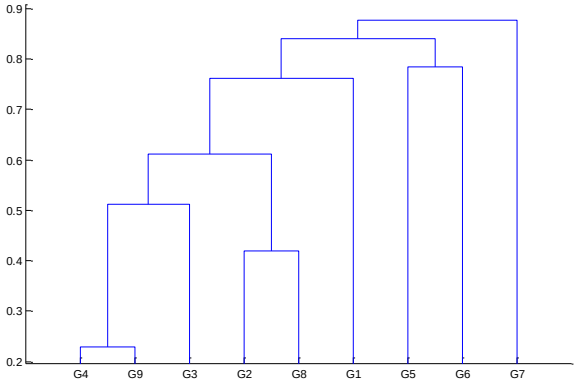

图 13.2　9 个农业区的聚类图

从图 13.2 的聚类图可以看出，如果把 9 个农业区分成 3，则$\{G_1,G_2,G_3,G_4,G_8,G_9\}$为

第一类，$\{G_5, G_6\}$为第二类，$\{G_7\}$为第三类。

计算及画聚类图的 MATLAB 程序如下

```
clc, clear
a=load('data132.txt'); %把表 13.5 中的数据保存在纯文本文件 data132.txt 中
[m,n]=size(a);
for j=1:n
    b(:,j)=(a(:,j)-min(a(:,j)))/(max(a(:,j))-min(a(:,j)));  %数据规格化
end
y=pdist(b)  %求聚类对象两两之间的距离
yc=squareform(y)  %变换成距离方阵，方便观察对象之间的距离
z=linkage(y)  %产生聚类树
name1([1:9],1)='G'; name2=int2str([1:9]');
name=cellstr(strcat(name1,name2)); %构造标注名称的字符细胞数组
[h,t]=dendrogram(z,'label',name)  %画聚类图
T=cluster(z,'maxclust',3)  %分成 3 类的聚类结果
```

2. 最长距离法

定义类G_i与类G_j之间距离为两类最远样品的距离，即

$$D_{qp} = \max_{\omega_i \in G_p, \omega_j \in A_q} d_{ij}.$$

最长距离法与最短距离法的合并步骤完全一样，也是将各样品先自成一类，然后将非对角线上最小元素对应的两类合并。设某一步将类G_p与G_q合并为G_r，则任一类G_k与G_r的距离用最长距离公式为

$$D_{kr} = \max_{\omega_i \in G_k, \omega_j \in G_r} d_{ij} = \max\{\max_{\omega_i \in G_k, \omega_j \in G_p} d_{ij}, \max_{\omega_i \in G_k, \omega_j \in G_q} d_{ij}\}$$
$$= \max\{D_{kp}, D_{kq}\}.$$

再找非对角线最小元素的两类并类，直至所有的样品全归为一类为止。

可见，最长距离法与最短距离法只有两点不同，一是类与类之间的距离定义不同；二是计算新类与其它类的距离所用的公式不同。

例 13.3（续例 13.1）设抽出 5 个样品，每个样品只测 1 个指标，它们是 2,3,3.5,7,9，试用最长距离法对 5 个样品进行分类。

解　这里我们使用马氏距离，利用 MATLAB 软件画出的聚类图见图 13.3，从图 13.3 可以看出聚类效果和例 13.1 是一样的。

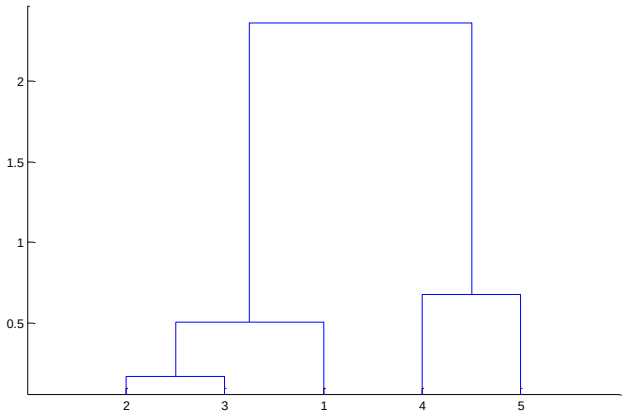

图 13.3　基于最长距离的聚类图

计算及画图的 MATLAB 程序如下

```
clc, clear
a=[2,3,3.5,7,9]';
x=pdist(a,'mahalanobis')  %求聚类对象两两之间的马氏距离
y=linkage(x,'complete')  %利用最长距离法产生聚类树
```

```
name=cellstr(int2str([1:5])); %构造标注名称的字符细胞数组
[h,t]=dendrogram(y,'label',name)   %画聚类图
n=input('请输入分类的类别数(输入后请回车)n=');
T=cluster(y,'maxclust',n)
```

13.1.4 MATLAB 聚类分析的相关命令

MATLAB中聚类分析相关命令的使用说明如下。

（1）pdist

B=pdist(A)计算 $m\times n$ 矩阵A（看作 m 个 n 维行向量，每行是一个对象的数据）中两两对象间的欧氏距离。对于有 m 个对象组成的数据集，共有 $(m-1)\cdot m/2$ 个两两对象组合。

输出B是包含距离信息的长度为 $(m-1)\cdot m/2$ 的向量。可用squareform函数将此向量转换为方阵，这样可使矩阵中的元素（i,j）对应原始数据集中对象i和j间的距离。

B=pdist(A,'metric')中用'metric'指定的方法计算矩阵A中对象间的距离。'metric'可取表13.6中特征字符串值。

表 13.6 'metric'取值及含义

字符串	含　义
'euclidean'	欧氏距离（缺省）
'seuclidean'	标准欧氏距离
'cityblock'	绝对值距离
'minkowski'	闵氏距离（Minkowski距离）
'chebychev'	车比雪夫距离（Chebychev距离）
'mahalanobis'	马氏距离（Mahalanobis距离）
'hamming'	海明距离（Hamming距离）
custom distance function	自定义函数距离
'cosine'	1-两个向量夹角的余弦
'correlation'	1-样本的相关系数
'spearman'	1-样本的Spearman秩相关系数
'jaccard'	1-Jaccard系数

B=pdist(A,'minkowski',p)用闵氏距离计算矩阵A中对象间的距离。p为闵氏距离计算用到的指数值，缺省为2。

（2）linkage

Z=linkage(B)使用最短距离算法生成具层次结构的聚类树。输入矩阵B为pdist函数输出的 $(m-1)\cdot m/2$ 维距离行向量。

Z=linkage(B, 'method')使用由'method'指定的算法计算生成聚类树。'method'可取表13.7中特征字符串值。

表 13.7 'method'取值及含义

字符串	含　义
'single'	最短距离（缺省）
'average'	无权平均距离
'centroid'	重心距离
'complete'	最大距离
'median'	赋权重心距离
'ward'	离差平方和方法（Ward方法）
'weighted'	赋权平均距离

输出Z为包含聚类树信息的 $(m-1)\times 3$ 矩阵。聚类树上的叶节点为原始数据集中的对象，由1到 m 。它们是单元素的类，级别更高的类都由它们生成。对应于Z中第 j 行每个新生成的类，其索引为 $m+j$ ，其中 m 为初始叶节点的数量。

第1列和第2列，即Z(:,[1:2])包含了被两两连接生成一个新类的所有对象的索引。生成的

新类索引为$m+j$。共有$m-1$个级别更高的类，它们对应于聚类树中的内部节点。

第三列Z(:,3)包含了相应的在类中的两两对象间的连接距离。

（3）cluster

T = cluster(Z,'cutoff',c)从连接输出（linkage）中创建聚类。cutoff为定义cluster函数如何生成聚类的阈值，其不同的值含义如表13.8所示。

表 13.8　cutoff 取值及含义

cutoff取值	含　义
0<cutoff<2	cutoff作为不一致系数的阈值。不一致系数对聚类树中对象间的差异进行了量化。如果一个连接的不一致系数大于阈值，则cluster函数将其作为聚类分组的边界。
2<=cutoff	cutoff作为包含在聚类树中的最大分类数

T = cluster(Z,'cutoff',c,'depth',d)从连接输出(linkage)中创建聚类。参数depth指定了聚类数中的层数，进行不一致系数计算时要用到。不一致系数将聚类树中两对象的连接与相邻的连接进行比较。详细说明见函数inconsistent。当参数depth被指定时，cutoff通常作为不一致系数阈值。

输出T为大小为m的向量，它用数字对每个对象所属的类进行标识。为了找到包含在类i中的来自原始数据集的对象，可用find(T==i)。

（4）zsore(A)

对数据矩阵进行标准化处理，处理方式为

$$\tilde{a}_{ij}=\frac{a_{ij}-\mu_j}{s_j},$$

其中μ_j, s_j是矩阵$A=(a_{ij})_{m\times n}$每一列的均值和标准差。

（5）H＝dendrogram(Z,P)

由linkage产生的数据矩阵Z画聚类树状图。P是结点数，默认值是30。

（6）T=clusterdata(A,cutoff)

将矩阵A的数据分类。A为$m\times n$矩阵，被看作m个n维行向量。它与以下几个命令等价

B=pdist(A)

Z=linkage(B,’single’)

T=cluster(Z,cutoff)

（7）squareform

将 pdist 的输出转换为方阵。

13.1.5　动态聚类法

用系统聚类法聚类时，随着聚类样本对象的增多，计算量会迅速增加，而且聚类结果—谱系图会十分复杂，不便于分析。特别是样品的个数很大（如$n\geq 100$）时，系统聚类法的计算量非常大，将占据大量的计算机内存空间和较多的计算时间，甚至会因计算机内存或计算时间的限制而无法进行。为了改进上述缺点，一个自然的想法是先粗略地分一下类，然后按某种最优原则进行修正，直到将类分得比较合理为止。基于这种思想就产生了动态聚类法，也称逐步聚类法。

动态聚类适用于大型数据。动态聚类法有许多种方法，这里介绍一种比较流行的动态聚类法—K均值法，它是一种快速聚类法，该方法得到的结果简单易懂，对计算机的性能要求不高，因而应用广泛。该方法由麦克奎因（Macqueen）于 1967 年提出。

算法的思想是假定样本集中的全体样本可分为C类，并选定C个初始聚类中心，然后，根据最小距离原则将每个样本分配到某一类中，之后不断迭代计算各类的聚类中心，并依据新的聚类中心调整聚类情况，直到迭代收敛或聚类中心不再改变。

K均值聚类算法最后将总样本集G划分为C个子集：$G_1, G_2, \cdots, G_C$，它们满足下面条件：

（1）$G_1 \bigcup G_2 \bigcup \cdots \bigcup G_C = G$；

（2）$G_i \bigcap G_j = \Phi$（$1 \le i < j \le C$）；

（3）$G_i \ne \Phi$，$G_i \ne G$（$1 \le i \le C$）。

记 m_i（$i = 1, \cdots, C$）为 C 个聚类中心，误差平法和的聚类准则

$$J_e = \sum_{i=1}^{C} \sum_{\omega \in G_i} \|\omega - m_i\|^2 ,$$

使 J_e 最小的聚类是误差平方和准则下的最优结果。

K 均值聚类算法描述如下：

（1）初始化。设总样本集 $G = \{\omega_j, j = 1,2,\cdots,n\}$ 是 n 个样本组成的集合，聚类数为 C（$2 \le C \le n$），将样本集 G 任意划分为 C 类，记为 $G_1, G_2, \cdots, G_C$，计算对应的 C 个初始聚类中心，记为 $m_1, m_2, \cdots, m_C$，并计算 J_e。

（2）$G_i = \Phi$（$i = 1,2,\cdots,C$），按最小距离原则将样本 ω_j（$j = 1,2,\cdots,n$）进行聚类，即

若 $d(\omega_j, G_i) = \min\limits_{1 \le k \le C} d(\omega_j, m_k)$，则 $\omega_j \in G_i$，$G_i = G_i \bigcup \{\omega_j\}$，$j = 1,2,\cdots,n$。

重新计算聚类中心

$$m_i = \frac{1}{n_i} \sum_{\omega_j \in G_i} \omega_j ,\quad i = 1,2,\cdots,C ;\quad j = 1,2,\cdots,n ,$$

式中，n_i 为当前 G_i 类中的样本数目。并重新计算 J_e。

（3）若连续两次迭代的 J_e 不变，则算法终止，否则算法转（2）。

注：实际计算时，可以不计算 J_e，只要聚类中心不发生变化，算法即可终止。

例 13.4　已知聚类的指标变量为 x_1, x_2，四个样本点的数据分别为

$\omega_1 = (1,3)$，$\omega_2 = (1.5,3.2)$，$\omega_3 = (1.3,2.8)$，$\omega_4 = (3,1)$.

试用 K 均值聚类分析把样本点分成 2 类。

解　现要分为两类 G_1 和 G_2 类，设初始聚类为 $G_1 = \{\omega_1\}$，$G_2 = \{\omega_2, \omega_3, \omega_4\}$，则初始聚类中心为

G_1 类：为 ω_1 值，即 $m_1 = (1,3)$。

G_2 类：$m_2 = \left(\frac{1.5 + 1.3 + 3}{3}, \frac{3.2 + 2.8 + 1}{3} \right) = (1.93, 2.33)$.

计算每个数据点到 G_1, G_2 聚类中心的距离

$d_{11} = \|\omega_1 - m_1\| = \sqrt{(1-1)^2 + (3-3)^2} = 0$，$d_{12} = \|\omega_1 - m_2\| = 1.14$；

$d_{21} = \|\omega_2 - m_1\| = 0.54$，$d_{22} = \|\omega_2 - m_2\| = 0.97$；

$d_{31} = \|\omega_3 - m_1\| = 0.36$，$d_{32} = \|\omega_3 - m_2\| = 0.78$；

$d_{41} = \|\omega_4 - m_1\| = 2.83$，$d_{42} = \|\omega_4 - m_2\| = 1.70$；

得到新的划分为：$G_1 = \{\omega_1, \omega_2, \omega_3\}$，$G_2 = \{\omega_4\}$，新的聚类中心为

G_1 类：$m_1 = \left(\frac{1 + 1.5 + 1.3}{3}, \frac{3 + 3.2 + 2.8}{3} \right) = (1.27, 3.0)$。

G_2 类：为 ω_4 值，即 $m_2 = (3,1)$.

重新计算每个样本点到 G_1, G_2 聚类中心的距离

$d_{11}=\|\omega_1-m_1\|=0.26$，$d_{12}=\|\omega_1-m_2\|=2.82$；

$d_{21}=\|\omega_2-m_1\|=0.31$，$d_{22}=\|\omega_2-m_2\|=2.66$；

$d_{31}=\|\omega_3-m_1\|=0.20$，$d_{32}=\|\omega_3-m_2\|=2.47$；

$d_{41}=\|\omega_4-m_1\|=2.65$，$d_{42}=\|\omega_4-m_2\|=0$；

所以，得新的划分为：$G_1=\{\omega_1,\omega_2,\omega_3\}$，$G_2=\{\omega_4\}$。

可见，新的划分与前面的相同，聚类中心没有改变，聚类结束。

计算的 MATLAB 程序如下

```
clc, clear
a=[1 3; 1.5 3.2; 1.3 2.8; 3 1]; %输入数据
[IDX,C]=kmeans(a,2) %IDX 返回的是聚类编号，C 的每一行是一个聚类中心。
```

例 13.5（续例 13.1）设抽出 5 个样品，每个样品只测 1 个指标，它们是 2,3,3.5,7,9，试用 K 均值聚类法把 5 个样品分成两类。

解　进行 K 均值聚类的 MATLAB 程序如下：

```
clc, clear
a=[2,3,3.5,7,9]'; %输入数据
[IDX,C]=kmeans(a,2) %IDX 返回的是聚类编号，C 的每一行是一个聚类中心。
```

聚类效果和例 13.1 的最短距离法的聚类效果是一样的。

13.2　判别分析

判别分析是多元统计分析中用于判别样本所属类型的一种统计分析方法，是一种在已知研究对象用某种方法分成若干类的情况下，确定新样本的观测数据，判定新样品所属类别的方法。它产生于 20 世纪 30 年代。

判别分析与聚类分析不同。判别分析要求具有一定的先验信息，是在已知研究对象分成若干类型（或组别）并已取得各种类型的一批已知样品的观测数据，然后在此基础上根据某些准则建立判别式，然后对未知类型的样品进行判别分类。对于聚类分析来说，对于一批给定样品要划分的类型事先并无先验信息，需要通过聚类分析以确定分类。因此，判别分析和聚类分析往往联合起来使用，例如判别分析要求先知道各类总体情况才能判断新样品的归类。当总体分类不清楚时，可先用聚类分析对原来的一批样品进行分类，然后再用判别分析建立判别式以对新样品进行判别。

常用的判别分析方法很多，有距离判别法、Fisher 判别法、Bayes 判别法和逐步判别法等。

13.2.1　距离判别法

距离判别法的基本思想为：根据已知分类的数据，分别计算各类的重心即分组（类）的均值，对任意给定的一个样品，若它与第 i 类的重心距离最近，就认为它来自第 i 类。因此，距离判别法又称为最近邻方法。

1. 两个总体的情形

设有两个总体（或称两类）G_1、G_2，现从第一个总体中抽取 n_1 个样品，从第二个总体中抽取 n_2 个样品，每个样品测量 p 个指标。两个总体的样本数据矩阵分别为 $A^{(1)}=(a_{ij}^{(1)})_{n_1\times p}$，$A^{(2)}=(a_{ij}^{(2)})_{n_2\times p}$。设 $\mu^{(1)}$、$\mu^{(2)}$，$\Sigma^{(1)}$、$\Sigma^{(2)}$ 分别为 G_1、G_2 的均值向量和协方差阵。

现取一个样品 $X=[x_1,\cdots,x_p]^T$，问 X 应判归为哪一类？

首先定义 X 到 G_1、G_2 总体的马氏距离，分别记为 $d(X,G_1)$ 和 $d(X,G_2)$，这里

$$d(X,G_i)=\sqrt{(X-\mu^{(i)})^T(\Sigma^{(i)})^{-1}(X-\mu^{(i)})}\,,\quad i=1,2\,,$$

按距离最近准则判别归类，则可写成

$$\begin{cases} X \in G_1, & \text{当} d(X,G_1) < d(X,G_2), \\ X \in G_2, & \text{当} d(X,G_1) > d(X,G_2), \\ \text{待判}, & \text{当} d(X,G_1) = d(X,G_2). \end{cases}$$

当 $\mu^{(1)}$、$\mu^{(2)}$，$\Sigma^{(1)}$、$\Sigma^{(2)}$ 未知时，可通过样本来估计。

2. 多个总体的情形

类似两个总体的讨论推广到多个总体。设有 k 个总体 $G_1,\cdots,G_k$，它们的均值和协方差阵分别为 $\mu^{(i)},\Sigma^{(i)}$，$i=1,\cdots,k$，从每个总体 G_i 中抽取 n_i 个样品，每个样品测 p 个指标。这 k 个总体的样本矩阵分别为 $A^{(1)}=(a_{ij}^{(1)})_{n_1\times p}$，…，$A^{(k)}=(a_{ij}^{(k)})_{n_k\times p}$。取一个样品 $X=[x_1,\cdots,x_p]^T$，问 X 应判归为哪一类？

类似地，定义 X 到各总体 G_i 的马氏距离

$$d(X,G_i)=\sqrt{(X-\mu^{(i)})^T(\Sigma^{(i)})^{-1}(X-\mu^{(i)})}\,,\quad i=1,\cdots,k\,.$$

X 到哪个总体近，就把 X 归于该类。

例 13.6　已经测得了 9 支 Af 和 6 支 Apf 的数据如下

Af：(1.24,1.27)，(1.36,1.74)，(1.38,1.64)，(1.38,1.82)，(1.38,1.90)，(1.40,1.70)，(1.48,1.82)，(1.54,1.82)，(1.56,2.08)

Apf：(1.14,1.78)，(1.18,1.96)，(1.20,1.86)，(1.26,2.00)，(1.28,2.00)，(1.30,1.96)

对触角和翼长分别为(1.24,1.80)，(1.28,1.84)与(1.40,2.04)的 3 个待判标本，用马氏距离判别法加以识别。

解　这里只给出计算的 MATLAB 程序：

```
clc, clear
x0=[1.24,1.27; 1.36,1.74; 1.38,1.64; 1.38,1.82; 1.38,1.90; 1.40,1.70
    1.48,1.82; 1.54,1.82; 1.56,2.08; 1.14,1.78; 1.18,1.96; 1.20,1.86
    1.26,2.00; 1.28,2.00; 1.30,1.96]; %输入已知样本数据
x=[1.24,1.80; 1.28,1.84; 1.40,2.04]; %输入待判样本点数据
g=[ones(9,1);2*ones(6,1)]; %已知样本数据的类别标号
[c,err]=classify(x,x0,g,'mahalanobis') %c 为分类结果，err 为错判率
```

分类结果是把前两个样本判为 Apf，第 3 个样本判为 Af。

13.2.2　Fisher 判别

Fisher 判别的基本思想是投影，即将表面上不易分类的数据通过投影到某个方向上，使得投影类与类之间得以分离的一种判别方法。

仅考虑两总体的情况，设两个 p 维总体为 G_1,G_2，且二阶矩都存在。Fisher 的判别思想是变换多元观测 X 到一元观测 y，使得由总体 G_1,G_2 产生的 y 尽可能的分离开来。

设在 p 维的情况下，X 的线性组合 $y=a^TX$，其中 a 为 p 维实向量。设 G_1,G_2 的均值向量分别为 μ_1,μ_2（均为 p 维），且有公共的协方差矩阵 Σ（$\Sigma>0$）。那么线性组合 $y=a^TX$ 的均值为

$$\mu_{y_1}=E(y\mid y=a^TX,X\in G_1)=a^T\mu_1,$$
$$\mu_{y_2}=E(y\mid y=a^TX,X\in G_2)=a^T\mu_2,$$

其方差为

$$\sigma_y^2=\mathrm{Var}(y)=a^T\Sigma a\,,$$

考虑比

$$\frac{(\mu_{y_1}-\mu_{y_2})^2}{\sigma_y^2}=\frac{[a^T(\mu_1-\mu_2)]^2}{a^T\Sigma a}=\frac{(a^T\delta)^2}{a^T\Sigma a}, \tag{1}$$

其中$\delta=\mu_1-\mu_2$为两总体均值向量差，根据 Fisher 的思想，我们要选择a使得（1）式达到最大。

定理 13.1　X 为 p 维随机变量，设 $y=a^TX$，当选取 $a=c\Sigma^{-1}\delta$，$c\neq 0$ 为常数时，（1）式达到最大。

特别当$c=1$时，线性函数

$$y=a^Tx=(\mu_1-\mu_2)^T\Sigma^{-1}X,$$

称为 Fisher 线性判别函数。令

$$K=\frac{1}{2}(\mu_{y_1}+\mu_{y_2})=\frac{1}{2}(a^T\mu_1+a^T\mu_2)=\frac{1}{2}(\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1+\mu_2).$$

定理 13.2　利用上面的记号，则有

$$\mu_{y_1}-K>0,\quad \mu_{y_2}-K<0.$$

由定理 13.2 得到如下的 Fisher 判别规则

$$\begin{cases} X\in G_1, \text{当}X\text{使得}(\mu_1-\mu_2)^T\Sigma^{-1}X\geq K, \\ X\in G_2, \text{当}X\text{使得}(\mu_1-\mu_2)^T\Sigma^{-1}X<K. \end{cases}$$

定义判别函数

$$W(X)=(\mu_1-\mu_2)^T\Sigma^{-1}X-K=(X-\frac{1}{2}(\mu_1+\mu_2))^T\Sigma^{-1}(\mu_1-\mu_2), \tag{2}$$

则判别规则可改写成

$$\begin{cases} X\in G_1, \text{当}X\text{使得}W(X)\geq 0, \\ X\in G_2, \text{当}X\text{使得}W(X)<0. \end{cases}$$

当总体的参数未知时，用样本对 μ_1,μ_2 及 Σ 进行估计，注意到这里的 Fisher 判别与距离判别一样不需要知道总体的分布类型，但两总体的均值向量必须有显著的差异才行，否则判别无意义。

例 13.7　云南某地盐矿的判别分析

已知云南某地盐矿分为钾盐及非钾盐（即钠盐）两类。我们已掌握的两类盐矿有关历史样本数据如表 13.9 所示。为对待判样本进行判别，需要进行判别分析。

表 13.9　云南某地盐矿的有关样本数据表

	样本	x_1	x_2	x_3	x_4
钾盐（A 类）	1	13.85	2.79	7.8	49.6
	2	22.31	4.67	12.31	47.8
	3	28.82	4.63	16.18	62.15
	4	15.29	3.54	7.58	43.2
	5	28.29	4.9	16.12	58.7
钠盐（B 类）	1	2.18	1.06	1.22	20.6
	2	3.85	0.8	4.06	47.1
	3	11.4	0	3.5	0
	4	3.66	2.42	2.14	15.1
	5	12.1	0	5.68	0
待判样本	1	8.85	3.38	5.17	26.1
	2	28.6	2.4	1.2	127
	3	20.7	6.7	7.6	30.8
	4	7.9	2.4	4.3	33.2
	5	3.19	3.2	1.43	9.9

	6	12.4	5.1	4.48	24.6

解　我们可以使用 Fisher 判别法进行判别，利用 MATLAB 软件求得的判别函数为

$$y=-38.2614+4.9088x_1+4.3617x_2-8.8504x_3+0.7449x_4.$$

判别准则为

$$\begin{cases} X\in \text{A类}, & \text{当}X\text{使得}y(X)>0, \\ X\in \text{B类}, & \text{当}X\text{使得}y(X)<0. \end{cases}$$

待判样品结果如表 13.10 所示。

表 13.10　待判样品结果表

待判样本	1	2	3	4	5	6
类别	B	A	A	B	B	A

计算的 MATLAB 程序如下

```
clc, clear
x0=[13.85    2.79    7.8 49.6
22.31    4.67    12.31    47.8
28.82    4.63    16.18    62.15
15.29    3.54    7.58    43.2
28.29    4.9 16.12    58.7
2.18    1.06    1.22    20.6
3.85    0.8 4.06    47.1
11.4    0    3.5 0
3.66    2.42    2.14    15.1
12.1    0    5.68    0]; %输入已知样本数据
x=[8.85 3.38    5.17    26.1
28.6    2.4 1.2 127
20.7    6.7 7.6 30.8
7.9 2.4 4.3 33.2
3.19    3.2 1.43    9.9
12.4    5.1 4.48    24.6]; %输入待判样本点数据
g=[ones(5,1);2*ones(5,1)]; %已知样本数据的类别标号
[c,err,P,logp,Coeffs]=classify(x,x0,g)
K=Coeffs(1,2).const    %线性判别的常数项
L=Coeffs(1,2).linear    %线性判别的一次项系数
f = @(x) K + x*L    %定义判别函数的匿名函数
ff=f(x)    %给出待判样本的判别值
```

或者是使用函数 fitcdiscr 进行判别，MATLAB 程序如下

```
clc, clear
x0=[13.85    2.79    7.8 49.6
22.31    4.67    12.31    47.8
28.82    4.63    16.18    62.15
15.29    3.54    7.58    43.2
28.29    4.9 16.12    58.7
2.18    1.06    1.22    20.6
3.85    0.8 4.06    47.1
```

```
11.4    0    3.5 0
3.66    2.42    2.14    15.1
12.1    0    5.68    0]; %输入已知样本数据
x=[8.85 3.38    5.17    26.1
28.6    2.4 1.2 127
20.7    6.7 7.6 30.8
7.9 2.4 4.3 33.2
3.19    3.2 1.43    9.9
12.4    5.1 4.48    24.6]; %输入待判样本点数据
g=[ones(5,1);2*ones(5,1)]; %已知样本数据的类别标号
obj=fitcdiscr(x0,g)
c=predict(obj,x)   %对待判样本进行分类
K=obj.Coeffs(1,2).Const   %线性判别的常数项
L=obj.Coeffs(1,2).Linear   %线性判别的一次项系数
f = @(x) K + x*L   %定义判别函数的匿名函数
ff=f(x)   %给出待判样本的判别值
```

如果进行线性分类，也可以利用线性最小二乘法。为了建立判别系统，引入判别函数

$$y = g(X) = c_1x_1 + c_2x_2 + c_3x_3 + c_4x_4 + c_5。$$

标志已知数据的准则为

$$g(X_0) = 1，当\ X_0\ 属于\ A\ 类时。$$

$$g(X_0) = -1，当\ X_0\ 属于\ B\ 类时。$$

记已知 A 类样本点的数据矩阵为$(a_{ij})_{5\times4}$，已知 B 类样本点的数据矩阵为$(b_{ij})_{5\times4}$。据所给的学习样本，可以得到关于判别函数中待定系数c_i（$i=1,2,3,4,5$）的线性方程组，这是包含 5 个未知数共 10 个方程的超定方程组。

$$\begin{cases} c_1a_{i1} + c_2a_{i2} + c_3a_{i3} + c_4a_{i4} + c_5 = 1, \quad i = 1,2,\cdots,5, \\ c_1b_{i1} + c_2b_{i2} + c_3b_{i3} + c_4b_{i4} + c_5 = -1, \quad i = 1,2,\cdots,5. \end{cases}$$

使用线性最小二乘法求解上述超定方程组，即求$c_1,\cdots,c_5$，使得

$$Q(c_1,c_2,\cdots,c_5) = \sum_{i=1}^{5}(c_1a_{i1} + c_2a_{i2} + c_3a_{i3} + c_4a_{i4} + c_5 - 1)^2 + \sum_{i=1}^{5}(c_1b_{i1} + c_2b_{i2} + c_3b_{i3} + c_4b_{i4} + c_5 + 1)^2$$

达到最小值，这就是超定方程组的最小二乘解。

利用 MATLAB 软件，求得的分类超平面为

$$y = g(X) = 0.2384x_1 + 0.2118x_2 - 0.4298x_3 + 0.0362x_4 - 1.8582.$$

对于未知样本 $X = [x_1,\cdots,x_4]^T$ 的判别准则为

$$\begin{cases} X \in \text{A类}, \quad 当X使得g(X) > 0, \\ X \in \text{B类}, \quad 当X使得g(X) < 0. \end{cases}$$

待判的 6 个样本点的判别结果和 Fisher 准则法的判别结果一致。

计算的 MATLAB 程序如下

```
clc, clear
x0=[13.85    2.79 7.8   49.6
22.31    4.67 12.31    47.8
28.82    4.63 16.18    62.15
```

```
15.29    3.54 7.58 43.2
28.29    4.9  16.12    58.7
2.18 1.06 1.22 20.6
3.85 0.8  4.06 47.1
11.4 0    3.5  0
3.66 2.42 2.14 15.1
12.1 0    5.68 0]; %输入已知样本数据
x=[8.85  3.38 5.17 26.1
28.6 2.4  1.2  127
20.7 6.7  7.6  30.8
7.9  2.4  4.3  33.2
3.19 3.2  1.43 9.9
12.4 5.1  4.48 24.6]; %输入待判样本点数据
xs=[x0,ones(10,1)]; %构造线性方程组的系数矩阵
b=[ones(5,1);-ones(5,1)]; %构造线性方程组的常数项列
cs=xs\b
f=[x,ones(6,1)]*cs   %计算待判样本的判别函数取值
ind=ones(6,1);   %分类编号的初值值
ind(find(f<0))=2   %最终结果的分类编号
```

13.3 主成分分析

例 13.8 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，原始数据见表 13.11。

表 13.11 30 个省市自治区的八项指标

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
内蒙古	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
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
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

解　用$i=1,2,\cdots,30$分别表示北京，天津，…，新疆 30 个省市自治区，第i个省市自治区的第 j 个指标变量x_j的取值记作a_{ij}，构造矩阵$A=(a_{ij})_{30\times 8}$。

（1）对原始数据进行标准化处理

将各指标值a_{ij}转换成标准化指标$\tilde{a}_{ij}$，

$$\tilde{a}_{ij}=\frac{a_{ij}-\mu_j}{s_j}，\quad（i=1,2,\cdots,30；\ \ j=1,2,\cdots,8）,$$

其中$\mu_j=\frac{1}{30}\sum_{i=1}^{30}a_{ij}$，$s_j=\sqrt{\frac{1}{29}\sum_{i=1}^{30}(a_{ij}-\mu_j)^2}$，（$j=1,2,\cdots,8$），即$\mu_j,s_j$为第 j 个指标的样本均值和样本标准差。对应地，称

$$\tilde{x}_j=\frac{x_j-\mu_j}{s_j}，\quad（\ j=1,2,\cdots,8\ ）$$

为标准化指标变量。

（2）计算相关系数矩阵R

相关系数矩阵$R=(r_{ij})_{8\times 8}$，

$$r_{ij}=\frac{\sum_{k=1}^{30}\tilde{a}_{ki}\cdot\tilde{a}_{kj}}{30-1}，\quad（i,j=1,2,\cdots,8）$$

式中$r_{ii}=1$，$r_{ij}=r_{ji}$，r_{ij}是第i个指标与第 j 个指标的相关系数。

（3）计算特征值和特征向量

计算相关系数矩阵R的特征值$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_8>0$，及对应的标准化特征向量$u_1,u_2,\cdots u_8$，其中$u_j=(u_{1j},u_{2j},\cdots,u_{8j})^T$，由特征向量组成8个新的指标变量

$$\begin{cases}y_1=u_{11}\tilde{x}_1+u_{21}\tilde{x}_2+\cdots+u_{81}\tilde{x}_8,\\ y_2=u_{12}\tilde{x}_1+u_{22}\tilde{x}_2+\cdots+u_{82}\tilde{x}_8,\\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\\ y_8=u_{18}\tilde{x}_1+u_{28}\tilde{x}_2+\cdots+u_{88}\tilde{x}_8.\end{cases}$$

式中y_1是第1主成分，y_2是第2主成分，…，y_8是第8主成分。

利用MATLAB软件求得相关系数矩阵的8个特征根及其贡献率如表13.12。

表 13.12　主成分分析结果

序号	特征根	贡献率	累计贡献率
1	3.7551	46.9391	46.9391
2	2.1967	27.4592	74.3983
3	1.2149	15.1861	89.5844
4	0.4024	5.0300	94.6144
5	0.2128	2.6600	97.2745
6	0.1380	1.7245	98.9990

7	0.0655	0.8182	99.8172
8	0.0146	0.1828	100

可以看出，前三个特征根的累计贡献率就达到 89.5844%，主成分分析效果很好。下面选取前三个主成分进行分析。前三个特征根对应的特征向量见表 13.13。

表 13.13　标准化变量的前 3 个主成分对应的特征向量

	$\tilde{x}_1$	$\tilde{x}_2$	$\tilde{x}_3$	$\tilde{x}_4$	$\tilde{x}_5$	$\tilde{x}_6$	$\tilde{x}_7$	$\tilde{x}_8$
第1特征向量	0.4566	0.3131	0.4705	0.2406	0.2507	-0.2624	-0.3197	0.4246
第2特征向量	0.2588	-0.4036	0.1087	-0.4874	0.4981	0.1700	0.4010	0.2879
第3特征向量	0.1097	0.2462	0.1923	0.3338	-0.2497	0.7228	0.3970	0.1914

由此可得三个主成分分别为

$$\begin{aligned} y_1 &= 0.4566\tilde{x}_1 + 0.3131\tilde{x}_2 + 0.4705\tilde{x}_3 + 0.2406\tilde{x}_4 \\ &\quad + 0.2507\tilde{x}_5 - 0.2624\tilde{x}_6 - 0.3197\tilde{x}_7 + 0.4246\tilde{x}_8, \\ y_2 &= 0.2588\tilde{x}_1 - 0.4036\tilde{x}_2 + 0.1087\tilde{x}_3 - 0.4874\tilde{x}_4 \\ &\quad + 0.4981\tilde{x}_5 + 0.1700\tilde{x}_6 + 0.4010\tilde{x}_7 + 0.2879\tilde{x}_8, \\ y_3 &= 0.1097\tilde{x}_1 + 0.2462\tilde{x}_2 + 0.1923\tilde{x}_3 + 0.3338\tilde{x}_4 \\ &\quad - 0.2497\tilde{x}_5 + 0.7228\tilde{x}_6 + 0.3970\tilde{x}_7 + 0.1914\tilde{x}_8. \end{aligned}$$

在第一主成分的表达式中第一、三、八项指标的系数较大，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，固定生产投资和工业总产值所刻划的反映经济发展状况的综合指标。

在第二主成分中，第二、四、五项指标的影响大，可将之看成是反映居民消费水平，职工平均工资和货物周转量的综合指标。

在第三主成分中，第六项指标影响最大，远远超过其它指标的影响，可单独看成是居民消费价格指数的影响。

计算的MATLAB程序如下

```
clc,clear
a=load('data138.txt');     %把原始数据保存在纯文本文件data138.txt中
b=zscore(a); %数据标准化
r=corrcoef(b)   %计算相关系数矩阵
%下面利用相关系数矩阵进行主成分分析，x的列为r的特征向量，即主成分的系数
[x,y,z]=pcacov(r) %y为r的特征值，z为各个主成分的贡献率
zz=cumsum(z) %求累积贡献率
```

13.4　因子分析

例 13.9　（续例 13.8）对全国 30 个省市自治区的经济发展八项指标作因子分析。

解　用 $i=1,2,\cdots,30$ 分别表示北京，天津，…，新疆 30 个省市自治区，第 i 个省市自治区的第 j 个指标变量 x_j 的取值记作 a_{ij}，构造矩阵 $A=(a_{ij})_{30\times 8}$。

（1）对原始数据进行标准化处理

将各指标值 a_{ij} 转换成标准化指标 $\tilde{a}_{ij}$，

$$\tilde{a}_{ij}=\frac{a_{ij}-\mu_j}{s_j},\quad (i=1,2,\cdots,30;\quad j=1,2,\cdots,8)$$

其中 $\mu_j=\frac{1}{30}\sum_{i=1}^{30}a_{ij}$， $s_j=\sqrt{\frac{1}{29}\sum_{i=1}^{30}(a_{ij}-\mu_j)^2}$，（$j=1,2,\cdots,8$），即 μ_j,s_j 为第 j 个指标的样本均值和样本标准差。对应地，称

$$\tilde{x}_j=\frac{x_j-\mu_j}{s_j}，\quad（j=1,2,\cdots,8）$$

为标准化指标变量。

（2）计算相关系数矩阵 R

相关系数矩阵 $R=(r_{ij})_{8\times8}$，

$$r_{ij}=\frac{\sum_{k=1}^{30}\tilde{a}_{ki}\cdot\tilde{a}_{kj}}{30-1}，\quad（i,j=1,2,\cdots,8）$$

式中 $r_{ii}=1$， $r_{ij}=r_{ji}$， r_{ij} 是第 i 个指标与第 j 个指标的相关系数。

（3）计算特征值，特征向量和初等因子载荷矩阵

计算相关系数矩阵 R 的特征值 $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_8>0$，及对应的标准化特征向量 $u_1,u_2,\cdots u_8$，其中 $u_j=(u_{1j},u_{2j},\cdots,u_{8j})^T$。

利用MATLAB软件求得相关系数矩阵的8个特征根及其贡献率如表13.14。

表 13.14　特征值及贡献率

序号	特征根	贡献率	累计贡献率
1	3.7551	46.9391	46.9391
2	2.1967	27.4592	74.3983
3	1.2149	15.1861	89.5844
4	0.4024	5.0300	94.6144
5	0.2128	2.6600	97.2745
6	0.1380	1.7245	98.9990
7	0.0655	0.8182	99.8172
8	0.0146	0.1828	100

可以看出，前三个特征根的累计贡献率就达到89. 5844%。

初等因子载荷矩阵 $\Lambda_1=[\sqrt{\lambda_1}u_1,\sqrt{\lambda_2}u_2,\cdots,\sqrt{\lambda_8}u_8]$。

（4）选择主因子，构造因子模型

根据（3）中的因子贡献率，选择 3 个主因子。对提取的 Λ_1 的前三列组成的因子载荷矩阵进行方差最大的正交旋转，得到矩阵

$$\Lambda_2=[\sqrt{\lambda_1}u_1,\sqrt{\lambda_2}u_2,\sqrt{\lambda_3}u_3]T=\begin{bmatrix}\alpha_{11}&\alpha_{12}&\alpha_{13}\\ \alpha_{21}&\alpha_{22}&\alpha_{23}\\ \vdots&\vdots&\vdots\\ \alpha_{81}&\alpha_{82}&\alpha_{83}\end{bmatrix},$$

这里 T 为正交矩阵。构造因子模型

$$\begin{cases}\tilde{x}_1=\alpha_{11}F_1+\alpha_{12}F_2+\alpha_{13}F_3,\\ \cdots\cdots\cdots\cdots\cdots\cdots\\ \tilde{x}_8=\alpha_{81}F_1+\alpha_{82}F_2+\alpha_{83}F_3.\end{cases}$$

利用 MATLAB 软件，求得旋转后的载荷矩阵见表 13.15。

表 13.15　旋转因子分析表

指标	主因子 1	主因子 2	主因子 3
$\tilde{x}_1$	0.9550	0.1253	−0.1309
$\tilde{x}_2$	0.2167	0.8410	−0.2129
$\tilde{x}_3$	0.8713	0.3522	−0.1375
$\tilde{x}_4$	0.0511	0.9268	−0.1143
$\tilde{x}_5$	0.7523	−0.5052	−0.1891
$\tilde{x}_6$	−0.1353	−0.0090	0.9687
$\tilde{x}_7$	−0.1026	−0.4944	0.8208
$\tilde{x}_8$	0.9439	0.1111	−0.0146

从表13.15可见，每个因子只有少数几个指标的因子载荷较大，因此可根据表13.15进行分类，将8个指标按高载荷分成三类，列于表13.16。

第一个因子在指标$\tilde{x}_1$、$\tilde{x}_3$、$\tilde{x}_8$有较大的载荷，这些是从GDP、固定资产投资、工业总产值等三个方面反映经济发展状况的，因此命名为总量因子。

第二个因子在指标$\tilde{x}_2$、$\tilde{x}_4$、$\tilde{x}_5$有较大的载荷，这些是从居民消费水平、职工平均工资、货物周转量这三个方面反映经济发展状况的，因此命名为消费因子。

第三个因子在指标$\tilde{x}_6$、$\tilde{x}_7$有较大的载荷，因此，命名为价格因子。

表 13.16　因子名称表

	高载荷指标	意义
主因子1	$\tilde{x}_1$：GDP $\tilde{x}_3$：固定资产投资 $\tilde{x}_8$：工业总产值	总量因子
主因子2	$\tilde{x}_2$：居民消费水平 $\tilde{x}_4$：职工平均工资 $\tilde{x}_5$：货物周转量	消费因子
主因子3	$\tilde{x}_6$：居民消费价格指数 $\tilde{x}_7$：商品零售价格指数	价格因子

计算的MATLAB程序如下

```
clc,clear
a=load('data138.txt');     %把原始数据保存在纯文本文件exam8.txt中
b=zscore(a); %数据标准化
r=corrcoef(b)   %计算相关系数矩阵
%下面利用相关系数矩阵进行主成分分析，x的列为r的特征向量，即主成分的系数
[x,y,z]=pcacov(r) %y为r的特征值，z为各个主成分的贡献率
zz=cumsum(z) %求累积贡献率
f1=repmat(sign(sum(x)),size(x,1),1); %构造与x同维数的元素为±1的矩阵
x2=x.*f1; %修改特征向量正负号,使得各特征向量的分量和为正
f2=repmat(sqrt(y)',size(x2,1),1);
lambda=x2.*f2       %构造全部因子的载荷矩阵
num=3;   %选择三个主因子
[lambda2,t]=rotatefactors(lambda(:,1:num),'method', 'varimax')   %对载荷矩阵进行旋转，
其中lambda2为旋转载荷矩阵，t为变换的正交矩阵
f3=repmat(sign(sum(lambda2)),size(lambda2,1),1); %构造元素为±1的矩阵
lambda3=lambda2.*f3   %修改载荷矩阵第2列的符号
```

习题13

13.1 家用电器故障实时检测问题

家用电器在日常生活中必不可少，如电饭煲能让煮饭更加便捷轻松、空调能帮助人们防暑降温、热水器能让我们在寒冷的冬天享受到舒适的沐浴条件等。但是随着时间的推移，电器的老化使其工作能力呈现衰减趋势，电器老化的成因有很多种，同时也是不可避免的。当用户提出电器需要维修时，电器的故障情况通常来说已经非常严重，为了保证用户的体验效果，现需要一种能够对电器运行状况进行实时监测并判别的方法来解决该问题。

现有某家电公司提供的经过脱敏处理后的某种电器运行数据，请你根据已有数据，并结合自己所掌握的知识，利用数学建模的方法来解决以下问题：

1) 电器在复杂的工作环境下工作时，有可能会导致传感器读取到异常数据，请针对这个问题给出你的解决方案。

2) 请根据附件一中的数据分析不同参数之间的相关性以及其对故障判别的重要程度。结合你之前所做的工作，建立一个该电器的故障判别模型并对附件一中的数据进行判别。

3) 请问你根据附件一的数据所建立的模型，是否依然适用于附件二中的数据？如果不适用，请给出你的修正方案。将你的模型修正后，请尝试着判断附件三中的数据，判断其状态为正常或故障（附件三中的数据标签已隐藏，你的判断结果将作为评奖时的参考）。

4) 除此之外，请你考虑一个问题，你的模型是否会出现误判？如果出现误判，那么将正常判断为故障和将故障判断为正常这两种错误，哪一种更应该避免？你是否能在模型中嵌入这一个影响因子？如果可以，请给出你的解决方案。

5) 在实际情况中，因家用电器的控制器计算能力有限（时钟周期仅仅有 40MHz 左右，桌面 CPU 的时钟周期大致为 4GHz），故其无法解决计算量特别大的模型。请问你能否在保证判断准确的前提下尽可能地降低自己计算复杂度？如果可以，请给出你的解决方案。