第二十九章　多元分析

多元分析（multivariate analyses）是多变量的统计分析方法，是数理统计中应用广泛的一个重要分支，其内容庞杂，视角独特，方法多样，深受工程技术人员的青睐和广泛使用，并在使用中不断完善和创新。由于变量的相关性，不能简单地把每个变量的结果进行汇总，这是多变量统计分析的基本出发点。

§1　聚类分析

将认识对象进行分类是人类认识世界的一种重要方法，比如有关世界的时间进程的研究，就形成了历史学，也有关世界空间地域的研究，则形成了地理学。又如在生物学中，为了研究生物的演变，需要对生物进行分类，生物学家根据各种生物的特征，将它们归属于不同的界、门、纲、目、科、属、种之中。事实上，分门别类地对事物进行研究，要远比在一个混杂多变的集合中更清晰、明了和细致，这是因为同一类事物会具有更多的近似特性。在企业的经营管理中，为了确定其目标市场，首先要进行市场细分。因为无论一个企业多么庞大和成功，它也无法满足整个市场的各种需求。而市场细分，可以帮助企业找到适合自己特色，并使企业具有竞争力的分市场，将其作为自己的重点开发目标。

通常，人们可以凭经验和专业知识来实现分类。而聚类分析（cluster analyses）作为一种定量方法，将从数据分析的角度，给出一个更准确、细致的分类工具。

1.1　相似性度量

1.1.1　样本的相似性度量

要用数量化的方法对事物进行分类，就必须用数量化的方法描述事物之间的相似程度。一个事物常常需要用多个变量来刻画。如果对于一群有待分类的样本点需用 p 个变量描述，则每个样本点可以看成是 R^p 空间中的一个点。因此，很自然地想到可以用距离来度量样本点间的相似程度。

记 Ω 是样本点集，距离 $d(\cdot,\cdot)$ 是 $\Omega\times\Omega\to R^+$ 的一个函数，满足条件：

1）$d(x,y)\geq 0$，$x,y\in\Omega$；

2）$d(x,y)=0$ 当且仅当 $x=y$；

3）$d(x,y)=d(y,x)$，$x,y\in\Omega$；

4）$d(x,y)\leq d(x,z)+d(x,y)$，$x,y,z\in\Omega$。

这一距离的定义是我们所熟知的，它满足正定性，对称性和三角不等式。在聚类分析中，对于定量变量，最常用的是 Minkowski 距离

$$d_q(x,y)=\left[\sum_{k=1}^{p}|x_k-y_k|^q\right]^{\frac{1}{q}},\quad q>0$$

当$q=1,2$或$q\to+\infty$时，则分别得到

1）绝对值距离

$$d_1(x,y)=\sum_{k=1}^{q}|x_k-y_k|,\tag{1}$$

2）欧氏距离

$$d_2(x,y)=\left[\sum_{k=1}^{p}|x_k-y_k|^2\right]^{\frac{1}{2}},\tag{2}$$

3）Chebyshev 距离

$$d_\infty(x,y)=\max_{1\le k\le p}|x_k-y_k|。\tag{3}$$

在 Minkowski 距离中，最常用的是欧氏距离，它的主要优点是当坐标轴进行正交旋转时，欧氏距离是保持不变的。因此，如果对原坐标系进行平移和旋转变换，则变换后样本点间的距离和变换前完全相同。

值得注意的是在采用 Minkowski 距离时，一定要采用相同量纲的变量。如果变量的量纲不同，测量值变异范围相差悬殊时，建议首先进行数据的标准化处理，然后再计算距离。在采用 Minkowski 距离时，还应尽可能地避免变量的多重相关性（multicollinearity）。多重相关性所造成的信息重叠，会片面强调某些变量的重要性。由于 Minkowski 距离的这些缺点，一种改进的距离就是马氏距离，定义如下

4）马氏（Mahalanobis）距离

$$d(x,y)=\sqrt{(x-y)^T\Sigma^{-1}(x-y)}\tag{4}$$

其中x,y为来自p维总体Z的样本观测值，Σ为Z的协方差矩阵，实际中Σ往往是不知道的，常常需要用样本协方差来估计。马氏距离对一切线性变换是不变的，故不受量纲的影响。

此外，还可采用样本相关系数、夹角余弦和其它关联性度量作为相似性度量。近年来随着数据挖掘研究的深入，这方面的新方法层出不穷。

1.1.2　类与类间的相似性度量

如果有两个样本类G_1和G_2，我们可以用下面的一系列方法度量它们间的距离：

1）最短距离法（nearest neighbor or single linkage method）

$$D(G_1,G_2)=\min_{\substack{x_i\in G_1\\ y_j\in G_2}}\{d(x_i,y_j)\}, \tag{5}$$

它的直观意义为两个类中最近两点间的距离。

2）最长距离法（farthest neighbor or complete linkage method）

$$D(G_1,G_2)=\max_{\substack{x_i\in G_1\\ y_j\in G_2}}\{d(x_i,y_j)\}, \tag{6}$$

它的直观意义为两个类中最远两点间的距离。

3）重心法（centroid method）

$$D(G_1,G_2)=d(\bar{x},\bar{y}), \tag{7}$$

其中 $\bar{x},\bar{y}$ 分别为 G_1,G_2 的重心。

4）类平均法（group average method）

$$D(G_1,G_2)=\frac{1}{n_1n_2}\sum_{x_i\in G_1}\sum_{x_j\in G_2}d(x_i,x_j), \tag{8}$$

它等于 G_1,G_2 中两两样本点距离的平均，式中 n_1,n_2 分别为 G_1,G_2 中的样本点个数。

5）离差平方和法（sum of squares method）

若记

$$D_1=\sum_{x_i\in G_1}(x_i-\bar{x}_1)^T(x_i-\bar{x}_1),\quad D_2=\sum_{x_j\in G_2}(x_j-\bar{x}_2)^T(x_j-\bar{x}_2),$$

$$D_{12}=\sum_{x_k\in G_1\cup G_2}(x_k-\bar{x})^T(x_k-\bar{x}),$$

其中

$$\bar{x}_1=\frac{1}{n_1}\sum_{x_i\in G_1}x_i,\quad \bar{x}_2=\frac{1}{n_2}\sum_{x_j\in G_2}x_j,\quad \bar{x}=\frac{1}{n_1+n_2}\sum_{x_k\in G_1\cup G_2}x_k$$

则定义

$$D(G_1,G_2)=D_{12}-D_1-D_2 \tag{9}$$

事实上，若 G_1,G_2 内部点与点距离很小，则它们能很好地各自聚为一类，并且这两类又能够充分分离（即 D_{12} 很大），这时必然有 $D=D_{12}-D_1-D_2$ 很大。因此，按定义可以认为，两类 G_1,G_2 之间的距离很大。离差平方和法最初是由 Ward 在 1936 年提出，

后经 Orloci 等人 1976 年发展起来的，故又称为 Ward 方法。

1.2　系统聚类法

1.2.1　系统聚类法的功能与特点

系统聚类法是聚类分析方法中最常用的一种方法。它的优点在于可以指出由粗到细的多种分类情况，典型的系统聚类结果可由一个聚类图展示出来。

例如，在平面上有 7 个点 $w_1, w_2, \cdots, w_7$（如图 1（a）），可以用聚类图（如图 1（b））来表示聚类结果。

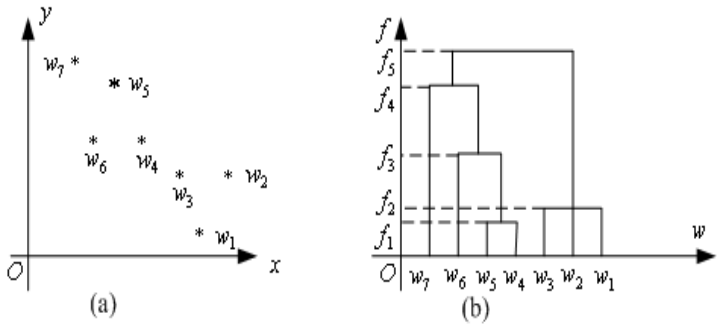

图 1　聚类方法示意图

记 $\Omega = \{w_1, w_2, \cdots, w_7\}$，聚类结果如下：当距离值为 f_5 时，分为一类

$$G_1 = \{w_1, w_2, w_3, w_4, w_5, w_6, w_7\};$$

距离值为 f_4 分为两类：

$$G_1 = \{w_1, w_2, w_3\},\quad G_2 = \{w_4, w_5, w_6, w_7\};$$

距离值为 f_3 分为三类：

$$G_1 = \{w_1, w_2, w_3\},\quad G_2 = \{w_4, w_5, w_6\},\quad G_3 = \{w_7\};$$

距离值为 f_2 分为四类：

$$G_1 = \{w_1, w_2, w_3\},\quad G_2 = \{w_4, w_5\},\quad G_3 = \{w_6\},\quad G_4 = \{w_7\}$$

距离值为 f_1 分为六类：

$$G_1 = \{w_4, w_5\},\ G_2 = \{w_1\},\ G_3 = \{w_2\},\ G_4 = \{w_3\},\ G_5 = \{w_6\},\ G_6 = \{w_7\}$$

距离小于 f_1 分为七类，每一个点自成一类。

怎样才能生成这样的聚类图呢？步骤如下：设$\Omega=\{w_1,w_2,\cdots,w_7\}$，

1）计算n个样本点两两之间的距离$\{d_{ij}\}$，记为矩阵$D=(d_{ij})_{n\times n}$；

2）首先构造n个类，每一个类中只包含一个样本点，每一类的平台高度均为零；

3）合并距离最近的两类为新类，并且以这两类间的距离值作为聚类图中的平台高度；

4）计算新类与当前各类的距离，若类的个数已经等于1，转入步骤5），否则，回到步骤3）；

5）画聚类图；

6）决定类的个数和类。

显而易见，这种系统归类过程与计算类和类之间的距离有关，采用不同的距离定义，有可能得出不同的聚类结果。

1.2.2　最短距离法与最长距离法

如果使用最短距离法来测量类与类之间的距离，即称其为系统聚类法中的最短距离法（又称最近邻法），最先由Florek等人1951年和Sneath1957年引入。下面举例说明最短距离法的计算步骤。

例1　设有5个销售员w_1,w_2,w_3,w_4,w_5，他们的销售业绩由二维变量(v_1,v_2)描述，见表1。

表1　销售员业绩表

销售员	v_1（销售量）百件	v_2（回收款项）万元
w_1	1	0
w_2	1	1
w_3	3	2
w_4	4	3
w_5	2	5

记销售员$w_i(i=1,2,3,4,5)$的销售业绩为(v_{i1},v_{i2})。如果使用绝对值距离来测量点与点之间的距离，使用最短距离法来测量类与类之间的距离，即

$$d(w_i,w_j)=\sum_{k=1}^{2}\left|v_{ik}-v_{jk}\right|,\quad D(G_p,G_q)=\min_{\substack{w_i\in G_p\\ w_j\in G_q}}\{d(w_i,w_j)\}$$

由距离公式 $d(\cdot,\cdot)$，可以算出距离矩阵。

$$\begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \end{array} \begin{array}{c} \begin{array}{ccccc} w_1 & w_2 & w_3 & w_4 & w_5 \end{array} \\ \begin{bmatrix} 0 & 1 & 4 & 6 & 6 \\ & 0 & 3 & 5 & 5 \\ & & 0 & 2 & 4 \\ & & & 0 & 4 \\ & & & & 4 \end{bmatrix} \end{array}$$

第一步，所有的元素自成一类 $H_1=\{w_1,w_2,w_3,w_4,w_5\}$。每一个类的平台高度为零，即 $f(w_i)=0(i=1,2,3,4,5)$。显然，这时 $D(G_p,G_q)=d(w_p,w_q)$。

第二步，取新类的平台高度为 1，把 w_1,w_2 合成一个新类 h_6，此时的分类情况是

$$H_2=\{h_6,w_3,w_4,w_5\}$$

第三步，取新类的平台高度为 2，把 w_3,w_4 合成一个新类 h_7，此时的分类情况是

$$H_3=\{h_6,h_7,w_5\}$$

第四步，取新类的平台高度为 3，把 h_6,h_7 合成一个新类 h_8，此时的分类情况是

$$H_4=\{h_8,w_5\}$$

第五步，取新类的平台高度为 4，把 h_8 和 w_5 合成一个新类 h_9，此时的分类情况是

$$H_5=\{h_9\}$$

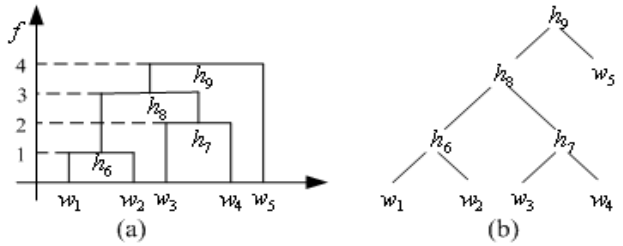

图 2　最短距离法

这样，h_9 已把所有的样本点聚为一类，因此，可以转到画聚类图步骤。画出聚类

图（如图 2（a））。这是一颗二叉树，如图 2（b）。

有了聚类图，就可以按要求进行分类。可以看出，在这五个推销员中 w_5 的工作成绩最佳， w_3, w_4 的工作成绩最好，而 w_1, w_2 的工作成绩较差。

完全类似于以上步骤，但以最长距离法来计算类间距离，就称为系统聚类法中的最长距离法。

计算的 MATLAB 程序如下：

```
clc,clear
a=[1,0;1,1;3,2;4,3;2,5];
[m,n]=size(a);
d=zeros(m,m);
for i=1:m
    for j=i+1:m
        d(i,j)=mandist(a(i,:),a(j,:)');
    end
end
d
nd=nonzeros(d);
nd=union(nd,nd)
for i=1:m-1
    nd_min=min(nd);
    [row,col]=find(d==nd_min);tm=union(row,col);
    tm=reshape(tm,1,length(tm));
    s(i)={char(['第',int2str(i),'次合成，平台高度为',num2str(nd_min),'
时的分类结果为：',int2str(tm)])};
    %上面大括号{}代表建立数组
    nd(find(nd==nd_min))=[];
    if length(nd)==0
        break
    end
end
s(:)
```

或者使用MATLAB统计工具箱的相关命令，编写如下程序：

```
clc,clear
a=[1,0;1,1;3,2;4,3;2,5];
y=pdist(a,'cityblock');yc=squareform(y)
z=linkage(y)
[h,t]=dendrogram(z)
```

MATLAB中相关命令的使用说明如下：

1）pdist

Y=pdist(X)计算$m\times n$矩阵X（被看作m个大小为n的向量）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Y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。

Y=pdist(X,’metric’)中用’metric’指定的方法计算矩阵X中对象间的距离。’metric’可取表2中特征字符串值。

表2 ’metric’取值及含义

字符串	含 义
’Euclid’	欧氏距离（缺省）
’SEuclid’	标准欧氏距离
’Mahal’	马氏距离（Mahalanobis距离）
’CityBlock’	绝对值距离
’Minkowski’	闵氏距离（Minkowski距离）

Y=pdist(X,’minkowski’,p)用闵氏距离计算矩阵X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(Y)使用最短距离算法生成具层次结构的聚类树。输入矩阵Y为pdist函数输出的$(m-1)\cdot m/2$维距离行向量。

Z=linkage(Y,’method’)使用由’method’指定的算法计算生成聚类树。’method’可取表3中特征字符串值。

表3 ’method’取值及含义

字符串	含 义
’single’	最短距离（缺省）
’complete’	最大距离
’average’	平均距离
’centroid’	重心距离
’ward’	离差平方和方法（Ward方法）

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中行j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(i,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(i,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)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4所示。

表4　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,depth,flag)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。

参数flag重载参数cutoff的缺省含义。如flag为’inconsistent’，则cutoff作为不一致系数的阈值。如flag为’cluster’，则cutoff作为分类的最大数目。

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(X)

对数据矩阵进行标准化处理，处理方式为

$$\tilde{x}_{ij}=\frac{x_{ij}-\bar{x}_j}{s_j}$$

其中矩阵$X=(x_{ij})_{m\times n}$看作是m个大小为n的向量，$\bar{x}_j,s_j$是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(X,cutoff)

将矩阵X的数据分类。X为$m\times n$矩阵，被看作m个大小为n的向量。它与以下几个命令等价：

Y=pdist(X,’euclid’)

Z=linkage(Y,’single’)

T=cluster(Z,cutoff)

7)squareform

将pdist的输出转换为方阵。

8）cophenet

c=cophenet(Z,Y) 计算相干系数，它是将Z中的距离信息（由linkage()函数产生）和Y中的距离信息（由pdist()）函数产生进行比较。Z为$(m-1)\times 3$矩阵，距离信息包含在第三列。Y是$(m-1)\cdot m/2$维的行向量。

例如，给定距离为Y的一组对象$\{1,2,\cdots,m\}$，函数linkage()生成聚类树。cophenet()函数用来度量这种分类的失真程度，即由分类所确定的结构与数据间的拟合程度。

输出值c为相干系数。对于要求很高的解，该值的幅度应非常接近1。它也可用来比较两种由不同算法所生成的分类解。

Z(:,3)和Y之间的相干系数定义为

$$c=\frac{\sqrt{\sum_{i<j}(y_{ij}-y)(z_{ij}-z)}}{\sqrt{\sum_{i<j}(y_{ij}-y)^2\sum_{i<j}(z_{ij}-z)^2}}$$

其中y_{ij}为Y中对象i和j间的距离；z_{ij}为Z(:,3)中对象i和j间的距离；y和z分别为Y和Z(:,3)的平均距离。

1.3　变量聚类法

在实际工作中，变量聚类法的应用也是十分重要的。在系统分析或评估过程中，为避免遗漏某些重要因素，往往在一开始选取指标时，尽可能多地考虑所有的相关因素。而这样做的结果，则是变量过多，变量间的相关度高，给系统分析与建模带来很大的不便。因此，人们常常希望能研究变量间的相似关系，按照变量的相似关系把它们聚合成若干类，进而找出影响系统的主要因素。

1.3.1　变量相似性度量

在对变量进行聚类分析时，首先要确定变量的相似性度量，常用的变量相似性度量有两种。

1）相关系数

记变量x_j的取值$(x_{1j},x_{2j},\cdots,x_{nj})^T\in R^n(j=1,2,\cdots,m)$。则可以用两变量$x_j$与$x_k$的样本相关系数作为它们的相似性度量

$$r_{jk}=\frac{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k)}{\left[\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2\right]^{\frac{1}{2}}},\qquad(10)$$

在对变量进行聚类分析时，利用相关系数矩阵是最多的。

2）夹角余弦

也可以直接利用两变量 x_j 与 x_k 的夹角余弦 r_{jk} 来定义它们的相似性度量，有

$$r_{jk}=\frac{\sum_{i=1}^{n}x_{ij}x_{ik}}{\left(\sum_{i=1}^{n}x_{ij}^2\sum_{i=1}^{n}x_{ik}^2\right)^{\frac{1}{2}}} \tag{11}$$

各种定义的相似度量均应具有以下两个性质：

a）$|r_{jk}|\le 1$，对于一切 j,k；

b）$r_{jk}=r_{kj}$，对于一切 j,k。

$|r_{jk}|$ 越接近1，x_j 与 x_k 越相关或越相似。$|r_{jk}|$ 越接近零，x_j 与 x_k 的相似性越弱。

1.3.2　变量聚类法

类似于样本集合聚类分析中最常用的最短距离法、最长距离法等，变量聚类法采用了与系统聚类法相同的思路和过程。在变量聚类问题中，常用的有最大系数法、最小系数法等。

1）最大系数法

在最大系数法中，定义两类变量的距离为

$$R(G_1,G_2)=\max_{\substack{x_j\in G_1\\x_k\in G_2}}\{r_{jk}\}, \tag{12}$$

这时，$R(G_1,G_2)$ 等于两类中最相似的两变量间的相似性度量值。

2）最小系数法

在最小系数法中，定义两类变量的距离为

$$R(G_1,G_2)=\min_{\substack{x_j\in G_1\\x_k\in G_2}}\{r_{jk}\}, \tag{13}$$

这时，$R(G_1,G_2)$ 等于两类中相似性最小的两个变量间的相似性度量值。

例2　服装标准制定中的变量聚类法。

在服装标准制定中，对某地成年女子的各部位尺寸进行了统计，通过14个部位的测量资料，获得各因素之间的相关系数表（见表2）。

表5　成年女子各部位相关系数

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}

x_1	1												
x_2	0.366	1											
x_3	0.242	0.233	1										
x_4	0.28	0.194	0.59	1									
x_5	0.36	0.324	0.476	0.435	1								
x_6	0.282	0.262	0.483	0.47	0.452	1							
x_7	0.245	0.265	0.54	0.478	0.535	0.663	1						
x_8	0.448	0.345	0.452	0.404	0.431	0.322	0.266	1					
x_9	0.486	0.367	0.365	0.357	0.429	0.283	0.287	0.82	1				
x_{10}	0.648	0.662	0.216	0.032	0.429	0.283	0.263	0.527	0.547	1			
x_{11}	0.689	0.671	0.243	0.313	0.43	0.302	0.294	0.52	0.558	0.957	1		
x_{12}	0.486	0.636	0.174	0.243	0.375	0.296	0.255	0.403	0.417	0.857	0.852	1	
x_{13}	0.133	0.153	0.732	0.477	0.339	0.392	0.446	0.266	0.241	0.054	0.099	0.055	1
x_{14}	0.376	0.252	0.676	0.581	0.441	0.447	0.44	0.424	0.372	0.363	0.376	0.321	0.627	1

其中 x_1 – 上体长，x_2 – 手臂长，x_3 – 胸围，x_4 – 颈围，x_5 – 总肩围，x_6 – 总胸宽，x_7 – 后背宽，x_8 – 前腰节高，x_9 – 后腰节高，x_{10} – 总体长，x_{11} – 身高，x_{12} – 下体长，x_{13} – 腰围，x_{14} – 臀围。用最大系数法对这14个变量进行系统聚类，分类结果如图3。

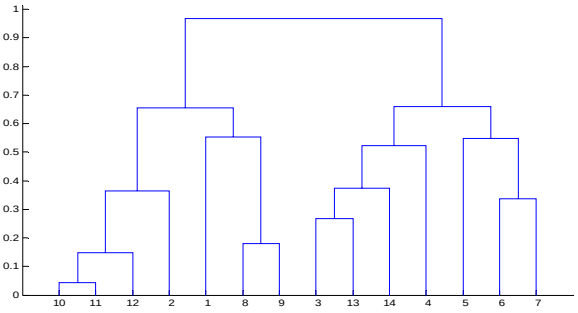

图3　成年女子14个部位指标的聚类图

计算的MATLAB程序如下：

```
%把下三角相关系数矩阵粘贴到纯文本文件ch.txt中
a=textread('ch.txt');
for i=1:14
    a(i,i)=0;
end
b=a(:);b=nonzeros(b);b=b';b=1-b;
z=linkage(b,'complete');
y=cluster(z,2)
dendrogram(z)
ind1=find(y==2);ind1=ind1'
ind2=find(y==1);ind2=ind2'
```

可以看出，人体的变量大体可以分为两类：一类反映人高、矮的变量，如上体长，手臂长，前腰节高，后腰节高，总体长，身高，下体长；另一类是反映人体胖瘦的变量，如胸围，颈围，总肩围，总胸宽，后背宽，腰围，臀围。

§2 聚类分析案例—我国各地区普通高等教育发展状况分析

聚类分析又称群分析，是对多个样本（或指标）进行定量分类的一种多元统计分析方法。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。本案例运用Q型和R型聚类分析方法对我国各地区普通高等教育的发展状况进行分析。

1．案例研究背景

近年来，我国普通高等教育得到了迅速发展，为国家培养了大批人才。但由于我国各地区经济发展水平不均衡，加之高等院校原有布局使各地区高等教育发展的起点不一致，因而各地区普通高等教育的发展水平存在一定的差异， 不同的地区具有不同的特点。对我国各地区普通高等教育的发展状况进行聚类分析，明确各类地区普通高等教育发展状况的差异与特点，有利于管理和决策部门从宏观上把握我国普通高等教育的整体发展现状，分类制定相关政策，更好的指导和规划我国高教事业的整体健康发展。

2．案例研究过程

（1）建立综合评价指标体系

高等教育是依赖高等院校进行的，高等教育的发展状况主要体现在高等院校的相关方面。遵循可比性原则，从高等教育的五个方面选取十项评价指标，具体如图4。

（2）数据资料

指标的原始数据取自《中国统计年鉴，1995》和《中国教育统计年鉴，1995》除以各地区相应的人口数得到十项指标值见表 6。其中：x_1 为每百万人口高等院校数；x_2 为每十万人口高等院校毕业生数；x_3 为每十万人口高等院校招生数；x_4 为每十万人口高等院校在校生数；x_5 为每十万人口高等院校教职工数；x_6 为每十万人口高等院校专职

教师数；x_7为高级职称占专职教师的比例；x_8为平均每所高等院校的在校生数；x_9为国家财政预算内普通高教经费占国内生产总值的比重；x_{10}为生均教育经费。

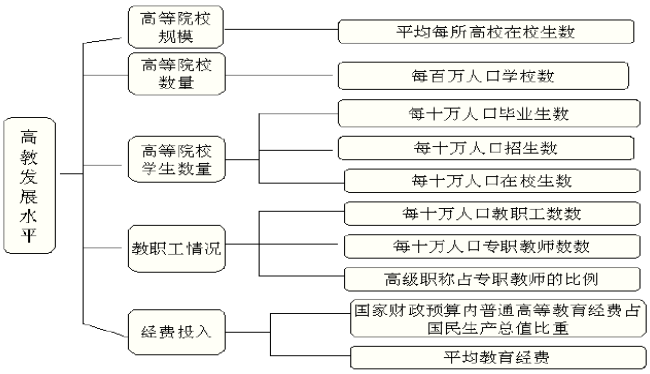

图4　高等教育的十项评价指标

表6　我国各地区普通高等教育发展状况数据

地区	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
北京	5.96	310	461	1557	931	319	44.36	2615	2.20	13631
上海	3.39	234	308	1035	498	161	35.02	3052	.90	12665
天津	2.35	157	229	713	295	109	38.40	3031	.86	9385
陕西	1.35	81	111	364	150	58	30.45	2699	1.22	7881
辽宁	1.50	88	128	421	144	58	34.30	2808	.54	7733
吉林	1.67	86	120	370	153	58	33.53	2215	.76	7480
黑龙江	1.17	63	93	296	117	44	35.22	2528	.58	8570
湖北	1.05	67	92	297	115	43	32.89	2835	.66	7262
江苏	.95	64	94	287	102	39	31.54	3008	.39	7786
广东	.69	39	71	205	61	24	34.50	2988	.37	11355
四川	.56	40	57	177	61	23	32.62	3149	.55	7693
山东	.57	58	64	181	57	22	32.95	3202	.28	6805
甘肃	.71	42	62	190	66	26	28.13	2657	.73	7282
湖南	.74	42	61	194	61	24	33.06	2618	.47	6477
浙江	.86	42	71	204	66	26	29.94	2363	.25	7704
新疆	1.29	47	73	265	114	46	25.93	2060	.37	5719
福建	1.04	53	71	218	63	26	29.01	2099	.29	7106
山西	.85	53	65	218	76	30	25.63	2555	.43	5580
河北	.81	43	66	188	61	23	29.82	2313	.31	5704

安徽	.59	35	47	146	46	20	32.83	2488	.33	5628
云南	.66	36	40	130	44	19	28.55	1974	.48	9106
江西	.77	43	63	194	67	23	28.81	2515	.34	4085
海南	.70	33	51	165	47	18	27.34	2344	.28	7928
内蒙古	.84	43	48	171	65	29	27.65	2032	.32	5581
西藏	1.69	26	45	137	75	33	12.10	810	1.00	14199
河南	.55	32	46	130	44	17	28.41	2341	.30	5714
广西	.60	28	43	129	39	17	31.93	2146	.24	5139
宁夏	1.39	48	62	208	77	34	22.70	1500	.42	5377
贵州	.64	23	32	93	37	16	28.12	1469	.34	5415
青海	1.48	38	46	151	63	30	17.87	1024	.38	7368

（3）R型聚类分析

定性考察反映高等教育发展状况的五个方面十项评价指标，可以看出，某些指标之间可能存在较强的相关性。比如每十万人口高等院校毕业生数、每十万人口高等院校招生数与每十万人口高等院校在校生数之间可能存在较强的相关性， 每十万人口高等院校教职工数和每十万人口高等院校专职教师数之间可能存在较强的相关性。为了验证这种想法，运用MATLAB软件计算十个指标之间的相关系数，相关系数矩阵如表6所示。

表6　相关系数矩阵

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
x_1	1.0000	0.9434	0.9528	0.9591	0.9746	0.9798	0.4065	0.0663	0.8680	0.6609
x_2	0.9434	1.0000	0.9946	0.9946	0.9743	0.9702	0.6136	0.3500	0.8039	0.5998
x_3	0.9528	0.9946	1.0000	0.9987	0.9831	0.9807	0.6261	0.3445	0.8231	0.6171
x_4	0.9591	0.9946	0.9987	1.0000	0.9878	0.9856	0.6096	0.3256	0.8276	0.6124
x_5	0.9746	0.9743	0.9831	0.9878	1.0000	0.9986	0.5599	0.2411	0.8590	0.6174
x_6	0.9798	0.9702	0.9807	0.9856	0.9986	1.0000	0.5500	0.2222	0.8691	0.6164
x_7	0.4065	0.6136	0.6261	0.6096	0.5599	0.5500	1.0000	0.7789	0.3655	0.1510
x_8	0.0663	0.3500	0.3445	0.3256	0.2411	0.2222	0.7789	1.0000	0.1122	0.0482
x_9	0.8680	0.8039	0.8231	0.8276	0.8590	0.8691	0.3655	0.1122	1.0000	0.6833
x_{10}	0.6609	0.5998	0.6171	0.6124	0.6174	0.6164	0.1510	0.0482	0.6833	1.0000

可以看出某些指标之间确实存在很强的相关性，因此可以考虑从这些指标中选取

几个有代表性的指标进行聚类分析。为此，把十个指标根据其相关性进行R型聚类，再从每个类中选取代表性的指标。首先对每个变量（指标）的数据分别进行标准化处理。变量间相近性度量采用相关系数，类间相近性度量的计算选用类平均法。聚类树型图见图5。

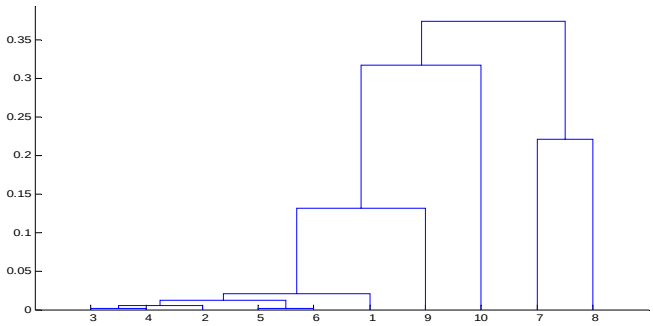

图 5　指标聚类树型图

计算的 MATLAB 程序如下：

```
load gj.txt     %把原始数据保存在纯文本文件 gj.txt 中
r=corrcoef(gj);   %计算相关系数矩阵
d=tril(r);          %取出相关系数矩阵的下三角元素
for i=1:10           %对角线元素化成零
    d(i,i)=0;
end
d=d(:);
d=nonzeros(d);   %取出非零元素
d=d';d=1-d;
z=linkage(d)
dendrogram(z)
```

从聚类图中可以看出，每十万人口高等院校招生数、每十万人口高等院校在校生数、每十万人口高等院校教职工数、每十万人口高等院校专职教师数、每十万人口高等院校毕业生数 5 个指标之间有较大的相关性，最先被聚到一起。如果将 10 个指标分为 6 类，其它 5 个指标各自为一类。这样就从十个指标中选定了六个分析指标：

x_1：每百万人口高等院校数；

x_2：每十万人口高等院校毕业生数；

x_7：高级职称占专职教师的比例；

x_8：平均每所高等院校的在校生数；

x_9：国家财政预算内普通高教经费占国内生产总值的比重；

x_{10}：生均教育经费。

可以根据这六个指标对30 个地区进行聚类分析。

（4）Q 型聚类分析

根据这六个指标对30个地区进行聚类分析。首先对每个变量的数据分别进行标准化处理，样本间相近性采用欧氏距离度量，类间距离的计算选用类平均法。聚类树型图见图6。

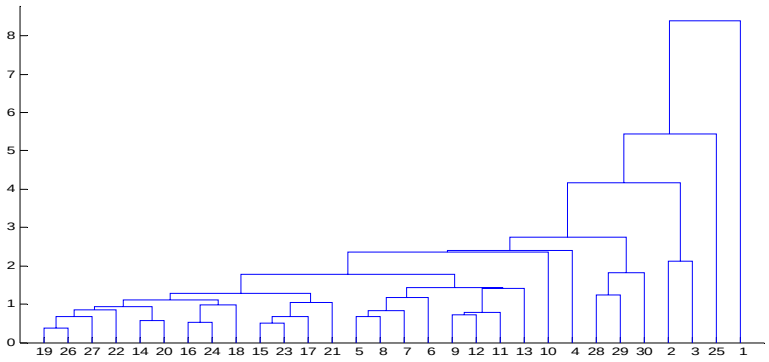

图6　各地区聚类树型图

计算的MATLAB程序如下：

```
load gj.txt   %把原始数据保存在纯文本文件gj.txt中
gj(:,3:6)=[];
gj=zscore(gj);
y=pdist(gj);
z=linkage(y)
dendrogram(z,'average')
```

4．案例研究结果

各地区高等教育发展状况存在较大的差异，高教资源的地区分布很不均衡。如果根据各地区高等教育发展状况把30 个地区分为三类，结果为：

第一类：北京；第二类：西藏；第三类：其他地区。

如果根据各地区高等教育发展状况把30个地区分为四类，结果为：

第一类：北京；第二类：西藏；第三类：上海天津；第四类：其他地区。

如果根据各地区高等教育发展状况把30个地区分为五类，结果为：

第一类：北京；第二类：西藏；第三类：上海天津；第四类：宁夏、贵州、青海；第五类：其他地区。

从以上结果结合聚类图中的合并距离可以看出，北京的高等教育状况与其它地区相比有非常大的不同，主要表现在每百万人口的学校数量和每十万人口的学生数量以及国

家财政预算内普通高教经费占国内生产总值的比重等方面远远高于其他地区，这与北京作为全国的政治、经济与文化中心的地位是吻合的。上海和天津作为另外两个较早的直辖市，高等教育状况和北京是类似的状况。宁夏、贵州和青海的高等教育状况极为类似，高等教育资源相对匮乏。西藏作为一个非常特殊的民族地区，其高等教育状况具有和其他地区不同的情形，被单独聚为一类，主要表现在每百万人口高等院校数比较高，国家财政预算内普通高教经费占国内生产总值的比重和生均教育经费也相对较高，而高级职称占专职教师的比例与平均每所高等院校的在校生数又都是全国最低的。这正是西藏高等教育状况的特殊之处：人口相对较少，经费比较充足，高等院校规模较小，师资力量薄弱。其他地区的高等教育状况较为类似，共同被聚为一类。针对这种情况，有关部门可以采取相应措施对宁夏、贵州、青海和西藏地区进行扶持，促进当地高等教育事业的发展。

§3 主成分分析

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，主成分分析和聚类分析有很大的不同，它有严格的数学理论作基础。

主成分分析的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变异，将我们手中许多相关性很高的变量转化成彼此相互独立或不相关的变量。通常是选出比原始变量个数少，能解释大部分资料中的变异的几个新变量，即所谓主成分，并用以解释资料的综合性指标。由此可见，主成分分析实际上是一种降维方法。

3.1 基本思想及方法

如果用 $x_1,x_2,\cdots,x_p$ 表示 p 门课程， $c_1,c_2,\cdots,c_p$ 表示各门课程的权重，那么加权之和就是

$$s=c_1x_1+c_2x_2+\cdots+c_px_p \tag{14}$$

我们希望选择适当的权重能更好地区分学生的成绩。每个学生都对应一个这样的综合成绩，记为 $s_1,s_2,\cdots,s_n$， n 为学生人数。如果这些值很分散，表明区分得好，即是说，需要寻找这样的加权，能使 $s_1,s_2,\cdots,s_n$ 尽可能的分散，下面来看它的统计定义。

设 $X_1,X_2,\cdots,X_p$ 表示以 $x_1,x_2,\cdots,x_p$ 为样本观测值的随机变量，如果能找到 $c_1,c_2,\cdots,c_p$， 使得

$$\mathrm{Var}(c_1X_1+c_2X_2+\cdots+c_pX_p) \tag{15}$$

的值达到最大，则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 p 个变量的最大变异。当然，（15）式必须加上某种限制，否则权值可选择无穷大而没有意

义，通常规定

$$c_1^2 + c_2^2 + \cdots + c_p^2 = 1 \tag{16}$$

在此约束下，求（15）式的最优解。由于这个解是 $p-$ 维空间的一个单位向量，它代表一个“方向”，它就是常说的主成分方向。

一个主成分不足以代表原来的 p 个变量，因此需要寻找第二个乃至第三、第四主成分，第二个主成分不应该再包含第一个主成分的信息，统计上的描述就是让这两个主成分的协方差为零，几何上就是这两个主成分的方向正交。具体确定各个主成分的方法如下。

设 Z_i 表示第 i 个主成分，$i = 1,2,\cdots,p$，可设

$$\begin{cases} Z_1 = c_{11}X_1 + c_{12}X_2 + \cdots + c_{1p}X_p \\ Z_2 = c_{21}X_1 + c_{22}X_2 + \cdots + c_{2p}X_p \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ Z_p = c_{p1}X_1 + c_{p2}X_2 + \cdots + c_{pp}X_p \end{cases} \tag{17}$$

其中对每一个 i，均有 $c_{i1}^2 + c_{i2}^2 + \cdots + c_{ip}^2 = 1$，且 $(c_{11}, c_{12}, \cdots, c_{1p})$ 使得 $\mathrm{Var}(Z_1)$ 的值达到最大；$(c_{21}, c_{22}, \cdots, c_{2p})$ 不仅垂直于 $(c_{11}, c_{12}, \cdots, c_{1p})$，而且使 $\mathrm{Var}(Z_2)$ 的值达到最大；$(c_{31}, c_{32}, \cdots, c_{3p})$ 同时垂直于 $(c_{11}, c_{12}, \cdots, c_{1p})$ 和 $(c_{21}, c_{22}, \cdots, c_{2p})$，并使 $\mathrm{Var}(Z_3)$ 的值达到最大；以此类推可得全部 p 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。

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，这是主成分分析的最大问题，回归分析是不存在这种情况的，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。

2）为使方差达到最大的主成分分析，所以不用转轴（由于统计软件常把主成分分析和因子分析放在一起，后者往往需要转轴，使用时应注意）。

3）主成分的保留。用相关系数矩阵求主成分时，Kaiser主张将特征值小于1的主成分予以放弃（这也是SPSS软件的默认值）。

4）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。

下面我们直接通过主成分估计（principle estimate）进一步阐述主成分分析的基本思想和相关概念。

3.2　主成分估计

主成分估计（principal component estimate）是Massy在1965年提出的，它是回归系数参数的一种线性有偏估计(biased estimate)，同其它有偏估计，如岭估计(ridge estimate)等一样，是为了克服最小二乘(LS)估计在设计阵病态（即存在多重共线性）时表现出的不稳定性而提出的。

主成分估计采用的方法是将原来的回归自变量变换到另另一组变量，即主成分，选择其中一部分重要的主成分作为新的自变量（此时丢弃了一部分，影响不大的自变量，这实际达到了降维的目的），然后用最小二乘法对选取主成分后的模型参数进行估计，最后再变换回原来的模型求出参数的估计。

设有 p 个回归（自）变量 $x_1, x_2, \cdots, x_p$，它在第 i 次试验中的取值为

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

将它们写成矩阵形式

$$X = (x_1, x_2, \cdots, x_p) = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \tag{18}$$

（注意这里 $x_1, x_2, \cdots, x_p$ 既表示回归自变量，又表示这些变量的观测值列向量，从上下文中我们容易区分开。）（18）即为设计阵，考虑线性模型

$$Y = \beta_0 1 + X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I), \tag{19}$$

其中 Y 为 $n \times 1$ 向量，β_0 为未知参数，1为所有元素均为1的 n 维列向量，β 为 $p \times 1$ 未知参数向量，ε 为 $n \times 1$ 误差向量。假定 X 已经标准化（即 X 的每个分量 x_j 均已标准化，如果未标准化，需要作变量的标准化变换 $(x_{ij} - \bar{x}_j)/s_j$，其中 $\bar{x}_j, s_j$ 为 x_j 各分量的均值和标准差。），此时

$$\hat{\beta}_0 = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{20}$$

对于自变量的任意一个线性组合

$$z = c_1 x_1 + c_2 x_2 + \cdots + c_p x_p, \quad \sum_{j=1}^{p} c_j^2 = 1, \tag{21}$$

将 z 视为一个新的变量。于是 z 在第 i 次试验中的取值为

$$z_i = c_1 x_{i1} + c_2 x_{i2} + \cdots + c_p x_{ip} \quad (i = 1,2,\cdots,n) \tag{22}$$

由于 X 已经标准化，因此

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} c_j x_{ij} = \frac{1}{n}\sum_{j=1}^{p} c_j \sum_{i=1}^{n} x_{ij} = 0 \tag{23}$$

记 $w = (c_1, c_2, \cdots, c_p)^T$，则

$$M_2^* = \frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^2 = \frac{1}{n}\sum_{i=1}^{n} z_i^2 = \frac{1}{n}(Xw)^T(Xw) \tag{24}$$

对于新变量 z 来说，如果在 n 次试验之下它的取值变化不大，即是说 M_2^* 较小，则这个新变量可以去掉。反之，M_2^* 较大，那么这个新变量有较大的变化，它的作用比较明显。注意到 z_i 的取值与 c_i 的选取有关。因此，我们总是希望所选择的 $c_i (i = 1,2,\cdots,p)$，使 M_2^* 达到最大，这才说明新变量在新建的回归模型中有较大的影响。

如果 $X^T X$ 的特征值 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$，它们所对应的标准化正交特征向量为 $\eta_1, \eta_2, \cdots, \eta_p$，则 $M_2^* = (Xw)^T(Xw)/n$ 的最大值在 $w = \eta_1$ 时达到，且最大值为 λ_1 / n。此时新变量 z 即为

$$z = X\eta_1$$

常记 $z_1 = X\eta_1$，并称之为自变量的第一主成分。一般地，如果已经确定了 k 个主成分

$$z_i = X\eta_i \quad (i = 1,2,\cdots,k), \tag{25}$$

则第 $k+1$ 个主成分 $z_{k+1} = Xw$ 可由下面两个条件决定：

1）$w^T \eta_i = 0$，$i = 1,2,\cdots,k$，$w^T w = 1$；

2）在条件1）之下，使 M_2^* 达到最大。

由二次型的条件极值可知，第 $k+1$ 个主成分就是 $z_{k+1} = X\eta_{k+1}$，这样，总共可以找到 p

个主成分 $z_i = X\eta_i$（$i = 1,2,\cdots,p$）。

现在回到线性模型(19)，将 $x_1, x_2, \cdots, x_p$ 变换为主成分 $z_1, z_2, \cdots, z_p$ 之后再求 β 的估计，令

$$Z = (z_1, z_2, \cdots, z_p) = \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \vdots & \vdots & & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{pmatrix} \tag{26}$$

记 $Q = (\eta_1, \eta_2, \cdots, \eta_p)_{p\times p}$，$Q$ 为标准化正交阵，且 $Z = XQ$，引入新参数 $\alpha = Q^T\beta$，或者 $\beta = Q\alpha$，则

$$Y = \beta_0 1 + ZQ^T\beta + \varepsilon = \beta_0 1 + Z\alpha + \varepsilon, \tag{27}$$

其中

$$Z^TZ = Q^TX^TXQ = Q^T(X^TX)Q = \Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix} \tag{28}$$

式（27）称之为模型（19）的典则形式。由式（28）可知，X^TX 的特征值 λ_i 度量了第 i 个主成分 z_i 在 n 次试验中取值变化的大小。如果 $\lambda_i \approx 0$，则该主成分在 n 次试验中取值的变化很小，它的作用可以并入模型（27）中的常数项 β_0。这相当于在典则形式中剔除变量 z_i。

如果 $\lambda_{r+1} = \cdots = \lambda_p \approx 0$，则剔除 $z_{r+1}, z_{r+2}, \cdots, z_p$，只剩下 α 的前 r 个分量 $\alpha_1, \alpha_2, \cdots, \alpha_r$，设它的最小二乘估计为 $\hat{\alpha}_1, \hat{\alpha}_2, \cdots, \hat{\alpha}_r$，而 α 后面的 $p-r$ 个分量则以0作为它们的估计，然后由关系式 $\beta = Q\alpha$ 即可确定 β 的估计，我们称之为 β 的主成分估计，实际步骤如下：

先将 Q, α 分块，即

$$Q=(Q_1,Q_2)，\ \alpha=\begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix} \tag{29}$$

其中Q_1为$p\times r$矩阵，α_1为r维向量，从而α的主成分估计为

$$\hat{\alpha}=\begin{pmatrix}\hat{\alpha}_1 & 0\end{pmatrix}^T \tag{30}$$

从而得到β的主成分估计

$$\hat{\beta}=(Q_1,Q_2)\begin{pmatrix}\hat{\alpha}_1\\ 0\end{pmatrix}=Q_1\hat{\alpha}_1 \tag{31}$$

理论上表明：主成分估计在设计阵病态时优于LS估计，但（31）在特征值为1的附近存在跳跃，会影响计算的稳定性，杨虎在1989年给出的单参数主成分估计解决了这个问题。

定义1　若存在$1\le r<p$，使$\lambda_r\ge 1>\lambda_{r+1}$，记

$$A=\mathrm{diag}(\frac{\lambda_1-1+\theta}{\lambda_1},\cdots,\frac{\lambda_r-1+\theta}{\lambda_r},\theta\lambda_{r+1},\cdots,\theta\lambda_p) \tag{32}$$

这里$\theta\in(\lambda_p,1)$为平稳参数，我们称$\hat{\beta}=QAQ^TQ_1\hat{\alpha}_1$为$\beta$的单参数主成分估计。

例3　Hald水泥问题，考察含如下四种化学成分

$x_1=3\mathrm{Ca0}\cdot\mathrm{Al_2O_3}$的含量（%），　　$x_2=3\mathrm{Ca0}\cdot\mathrm{SiO_2}$的含量（%），

$x_3=4\mathrm{Ca0}\cdot\mathrm{Al_2O_3}\cdot\mathrm{Fe_2O_3}$的含量（%），$x_4=2\mathrm{Ca0}\cdot\mathrm{SiO_2}$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7，对数据实施标准化，则$X^TX/12$就是样本相关系数阵（见表8）。

表7　Hald水泥

序号	x_1	x_2	x_3	x_4	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

表8 Hald水泥数据的样本相关系数阵

	x_1	x_2	x_3	x_4
x_1	1	0.2286	-0.8241	-0.2454
x_2	0.2286	1	-0.1392	-0.9730
x_3	-0.8241	-0.1392	1	0.0295
x_4	-0.2454	-0.9730	0.0295	1

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

$$\eta_1^T = (0.476, 0.5639, -0.3941, -0.5479)$$

$$\eta_2^T = (-0.509, 0.4139, 0.605, -0.4512)$$

$$\eta_3^T = (0.6755, -0.3144, 0.6377, -0.1954)$$

对Hald数据直接作线性回归得经验回归方程

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

再由（31）式计算出主成分估计，即可获得如下主成分回归方程

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

两个方程的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。

计算的MATLAB程序如下：

```
clc,clear
load sn.txt  %把原始的x1,x2,x3,x4,y的数据保存在纯文本文件sn.txt中
[m,n]=size(sn);num=3; %num为选取的主成分的个数
```

```
mu=mean(sn);sigma=std(sn);
snb=zscore(sn); %数据标准化
b=snb(:,1:end-1); %x1,x2,x3,x4的数据赋给b
r=cov(b);  %标准化数据的协方差阵就是相关系数阵
[x,y,z]=pcacov(r);
f=repmat(sign(sum(x)),size(x,1),1);
x=x.*f;
%以下是普通的最小二乘法回归
r=[ones(m,1),b]\snb(:,end);  %标准化数据的回归方程系数
bzh=mu./sigma;
ch10=mu(end)-bzh(1:end-1)*r(2:end)*sigma(end)  %原始数据的常数项
fr=r(2:end);fr=fr';
ch1=fr./sigma(1:end-1)*sigma(end) %原始数据的x1,x2等等系数
%以下是主成分回归
pval=b*x(:,1:num);
rp=[ones(m,1),pval]\snb(:,end);  %主成分数据的回归方程系数
beta=x(:,1:num)*rp(2:num+1);              %标准化数据的回归方程系数
ch20=mu(end)-bzh(1:end-1)*beta*sigma(end)  %原始数据的常数项
fr=beta';
ch2=fr./sigma(1:end-1)*sigma(end) %原始数据的x1,x2等等系数
check1=sqrt(sum((sn(:,1:end-1)*ch1'+ch10-sn(:,end)).^2)/(m-n))
check2=sqrt(sum((sn(:,1:end-1)*ch2'+ch20-sn(:,end)).^2)/(m-num-1))
```

3.3 特征值因子的筛选

回到主成分分析，实际中确定（17）式中的系数就是采用（28）式中矩阵的特征向量。因此，剩下的问题仅仅是将 X^TX 的特征值按由大到小的次序排列之后，如何筛选这些特征值？一个实用的方法是删去 $\lambda_{r+1},\lambda_{r+2},\cdots,\lambda_p$ 后，这些删去的特征值之和占整个特征值之和 $\sum\lambda_i$ 的15%以下，换句话说，余下的特征值所占的比重（定义为累积贡献率）将超过85%，当然这不是一种严格的规定，今年来文献中关于这方面的讨论很多，有很多比较成熟的方法，这里不一一介绍。

单纯考虑累积贡献率有时是不够的，还需要考虑选择的主成分对原始变量的贡献值，我们用相关系数的平方和来表示，如果选取的主成分为 $z_1,z_2,\cdots,z_r$，则它们对原变量 x_i 的贡献值为

$$\rho_i = \sum_{j=1}^{r} r^2(z_j, x_i) \tag{33}$$

例4　设 $X = (x_1, x_2, x_3)$，且

$$X^T X = \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

则可算得 $\lambda_1 = 5.8284$，$\lambda_2 = 0.1716$，如果我们仅取第一个主成分，由于其累积贡献率已经达到 97.14%，似乎很理想了，但如果进一步计算主成分对原变量的贡献值，容易发现

$$\rho_3 = r^2(z_1, x_3) = 0$$

可见，第一个主成分对第三个变量的贡献值为0，这是因为 x_3 和 x_1, x_2 都不相关。由于在第一个主成分中一点也不包含 x_3 的信息，这时只选择一个主成分就不够了，需要再取第二个主成分。

例5　研究纽约股票市场上五种股票的周回升率。这里，周回升率＝（本星期五市场收盘价－上星期五市场收盘价）/上星期五市场收盘价。从1975年1月到1976年12月，对这五种股票作了100组独立观测。因为随着一般经济状况的变化，股票有集聚的趋势，因此，不同股票周末回升率是彼此相关的。

设 $x_1, x_2, \cdots, x_5$ 分别为五只股票的周回升率，则从数据算得

$$\bar{x}^T = (0.0054, 0.0048, 0.0057, 0.0063, 0.0037)$$

$$R = \begin{pmatrix} 1.000 & 0.577 & 0.509 & 0.387 & 0.462 \\ 0.577 & 1.000 & 0.599 & 0.389 & 0.322 \\ 0.509 & 0.599 & 1.000 & 0.436 & 0.426 \\ 0.387 & 0.389 & 0.436 & 1.000 & 0.523 \\ 0.462 & 0.322 & 0.426 & 0.523 & 1.000 \end{pmatrix}$$

这里 R 是标准化数据的协方差矩阵，R 的特征值和标准正交特征向量为

$$\lambda_1 = 2.857,\ \lambda_2 = 0.809,\ \lambda_3 = 0.540,\ \lambda_4 = 0.452,\ \lambda_5 = 0.343,$$

$$\eta_1^T = (0.464, 0.457, 0.470, 0.421, 0.421)$$

$$\eta_2^T = (0.240, 0.509, 0.260, -0.526, -0.582)$$

标准化变量的前两个主成分为

$$z_1 = 0.464\tilde{x}_1 + 0.457\tilde{x}_2 + 0.470\tilde{x}_3 + 0.421\tilde{x}_4 + 0.421\tilde{x}_5$$

$$z_2 = 0.240\tilde{x}_1 + 0.509\tilde{x}_2 + 0.260\tilde{x}_3 - 0.526\tilde{x}_4 - 0.582\tilde{x}_5$$

它们的累积贡献率为

$$\frac{\lambda_1 + \lambda_2}{\sum_{i=1}^{5} \lambda_i} \times 100\% = 73\%$$

这两个主成分具有重要的实际解释，第一主成分大约等于这五种股票周回升率和的一个常数倍，通常称为股票市场主成分，简称市场主成分；第二主成分代表化学股票（在 z_2 中系数为正的三只股票都是化学工业上市企业）和石油股票（在 z_2 中系数为负的两只股票恰好都为石油板块的上市企业）的一个对照，称之为工业主成分。这说明，这些股票周回升率的大部分变差来自市场活动和与它不相关的工业活动。关于股票价格的这个结论与经典的证券理论吻合。至于其它主成分解释较为困难，很可能表示每种股票自身的变差，好在它们的贡献率很少，可以忽略不计。

§4　主成分分析案例－我国各地区普通高等教育发展水平综合评价

主成分分析试图在力保数据信息丢失最少的原则下，对多变量的截面数据表进行最佳综合简化，也就是说，对高维变量空间进行降维处理。本案例运用主成分分析方法综合评价我国各地区普通高等教育的发展水平。

问题与第2节中的问题相同，我们这里就不重复叙述了。

4.1　主成分分析法的步骤

主成分分析法进行评价的步骤如下：

1）对原始数据进行标准化处理

假设进行主成分分析的指标变量有 m 个：$x_1, x_2, \cdots, x_m$，共有 n 个评价对象，第 i 个评价对象的第 j 个指标的取值为 x_{ij}。将各指标值 x_{ij} 转换成标准化指标 $\tilde{x}_{ij}$，

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \quad (i = 1,2,\cdots,n; \quad j = 1,2,\cdots,m)$$

其中 $\bar{x}_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij}$，$s_j=\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2$，（$j=1,2,\cdots,m$），即 $\bar{x}_j,s_j$ 为第 j 个指标的样本均值和样本标准差。对应地，称

$$\tilde{x}_i=\frac{x_i-\bar{x}_i}{s_i}，\quad（i=1,2,\cdots,m）$$

为标准化指标变量。

2）计算相关系数矩阵 R

相关系数矩阵 $R=(r_{ij})_{m\times m}$

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

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 $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_m\geq 0$，及对应的特征向量 $u_1,u_2,\cdots,u_m$，其中 $u_j=(u_{1j},u_{2j},\cdots,u_{nj})^T$，由特征向量组成 m 个新的指标变量

$$\begin{cases}y_1=u_{11}\tilde{x}_1+u_{21}\tilde{x}_2+\cdots+u_{n1}\tilde{x}_n\\ y_2=u_{12}\tilde{x}_1+u_{22}\tilde{x}_2+\cdots+u_{n2}\tilde{x}_n\\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\\ y_m=u_{1m}\tilde{x}_1+u_{2m}\tilde{x}_2+\cdots+u_{nm}\tilde{x}_n\end{cases}$$

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。

4）选择 p（$p\leq m$）个主成分，计算综合评价值

① 计算特征值 $\lambda_j(j=1,2,\cdots,m)$ 的信息贡献率和累积贡献率。称

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

为主成分 y_j 的信息贡献率；

$$\alpha_p = \frac{\sum_{k=1}^{p} \lambda_k}{\sum_{k=1}^{m} \lambda_k}$$

为主成分 $y_1, y_2, \cdots, y_p$ 的累积贡献率，当 α_p 接近于1（ $\alpha_p = 0.85, 0.90, 0.95$ ）时，则选择前 p 个指标变量 $y_1, y_2, \cdots, y_p$ 作为 p 个主成分，代替原来 m 个指标变量，从而可对 p 个主成分进行综合分析。

② 计算综合得分

$$Z = \sum_{j=1}^{p} b_j y_j$$

其中 b_j 为第 j 个主成分的信息贡献率，根据综合得分值就可进行评价。

4.2 基于主成分分析法的综合评价

定性考察反映高等教育发展状况的五个方面十项评价指标，可以看出，某些指标之间可能存在较强的相关性。比如每十万人口高等院校毕业生数、每十万人口高等院校招生数与每十万人口高等院校在校生数之间可能存在较强的相关性， 每十万人口高等院校教职工数和每十万人口高等院校专职教师数之间可能存在较强的相关性。为了验证这种想法，计算十个指标之间的相关系数。

可以看出某些指标之间确实存在很强的相关性，如果直接用这些指标进行综合评价，必然造成信息的重叠，影响评价结果的客观性。主成分分析方法可以把多个指标转化为少数几个不相关的综合指标，因此，可以考虑利用主成分进行综合评价。

利用MATLAB软件对十个评价指标进行主成分分析，相关系数矩阵的前几个特征根及其贡献率如表7。

表7 主成分分析结果

序号	特征根	贡献率	累计贡献率
1	7.5022	75.0216	75.0216
2	1.577	15.7699	90.7915
3	0.5362	5.3621	96.1536
4	0.2064	2.0638	98.2174
5	0.145	1.4500	99.6674
6	0.0222	0.2219	99.8893

可以看出，前两个特征根的累计贡献率就达到90%以上，主成分分析效果很好。下面选取前四个主成分（累计贡献率就达到98%）进行综合评价。前四个特征根对应的特

征向量见表8。

表8　标准化变量的前4个主成分对应的特征向量

	$\tilde{x}_1$	$\tilde{x}_2$	$\tilde{x}_3$	$\tilde{x}_4$	$\tilde{x}_5$	$\tilde{x}_6$	$\tilde{x}_7$	$\tilde{x}_8$	$\tilde{x}_9$	$\tilde{x}_{10}$
第1特征向量	0.3497	0.3590	0.3623	0.3623	0.3605	0.3602	0.2241	0.1201	0.3192	0.2452
第2特征向量	-0.1972	0.0343	0.0291	0.0138	-0.0507	-0.0646	0.5826	0.7021	-0.1941	-0.2865
第3特征向量	−0. 1639	−0. 1084	−0. 0900	−0. 1128	−0. 1534	−0. 1645	−0. 0397	0. 3577	0. 1204	0. 8637
第4特征向量	-0.1022	-0.2266	-0.1692	-0.1607	-0.0442	-0.0032	0. 0812	0. 0702	0. 8999	0. 2457

由此可得四个主成分分别为

$$y_1 = 0.3497\tilde{x}_1 + 0.359\tilde{x}_2 + \cdots + 0.2452\tilde{x}_{10}$$

$$y_2 = -0.1972\tilde{x}_1 + 0.0343\tilde{x}_2 + \cdots - 0.286\tilde{x}_{10}$$

$$y_3 = -0.1639\tilde{x}_1 - 0.1084\tilde{x}_2 + \cdots + 0.8637\tilde{x}_{10}$$

$$y_4 = -0.1022\tilde{x}_1 - 0.2266\tilde{x}_2 + \cdots - 0.2457\tilde{x}_{10}$$

从主成分的系数可以看出，第一主成分主要反映了前六个指标（学校数、学生数和教师数方面）的信息，第二主成分主要反映了高校规模和教师中高级职称的比例，第三主成分主要反映了生均教育经费，第四主成分主要反映了国家财政预算内普通高教经费占国内生产总值的比重。把各地区原始十个指标的标准化数据代入四个主成分的表达式，就可以得到各地区的四个主成分值。

分别以四个主成分的贡献率为权重，构建主成分综合评价模型：

$$Z = 0.7502y_1 + 0.1577y_2 + 0.0536y_3 + 0.0206y_4$$

把各地区的四个主成分值代入上式，可以得到各地区高教发展水平的综合评价值以及排序结果如表9。

表9　排名和综合评价结果

地区	北京	上海	天津	陕西	辽宁	吉林	黑龙江	湖北	江苏	广东
名次	1	2	3	4	5	6	7	8	9	10
综合评价值	8.6043	4.4738	2.7881	0.8119	0.7621	0.5884	0.2971	0.2455	0.0581	0.0058
地区	四川	山东	甘肃	湖南	浙江	新疆	福建	山西	河北	安徽

名次	11	12	13	14	15	16	17	18	19	20
综合评价值	-0.268	-0.3645	-0.4879	-0.5065	-0.7016	-0.7428	-0.7697	-0.7965	-0.8895	-0.8917
地区	云南	江西	海南	内蒙古	西藏	河南	广西	宁夏	贵州	青海
名次	21	22	23	24	25	26	27	28	28	30
综合评价值	-0.9557	-0.9610	-1.0147	-1.1246	-1.1470	-1.2059	-1.2250	-1.2513	-1.6514	-1.68

```
clc,clear
load gj.txt   %把原始数据保存在纯文本文件gj.txt中
gj=zscore(gj); %数据标准化
r=corrcoef(gj);  %计算相关系数矩阵
[x,y,z]=pcacov(r);
f=repmat(sign(sum(x)),size(x,1),1);
x=x.*f;
df=gj*x(:,1:4)
tf=df*z(1:4)/100;
[stf,ind]=sort(tf,'descend')
```

4.3 结论

各地区高等教育发展水平存在较大的差异，高教资源的地区分布很不均衡。北京、上海、天津等地区高等教育发展水平遥遥领先，主要表现在每百万人口的学校数量和每十万人口的教师数量、学生数量以及国家财政预算内普通高教经费占国内生产总值的比重等方面。陕西和东北三省地区高等教育发展水平也比较高。贵州、广西、河南、安徽等地区高等教育发展水平比较落后，这些地区的高等教育发展需要政策和资金的扶持。值得一提的是西藏、新疆、甘肃等经济不发达地区的高等教育发展水平居于中上游水平，可能是由于人口等原因。

§5 因子分析

因子分析（factor analysis）是由英国心理学家Spearman在1904年提出来的，他成功地解决了智力测验得分的统计分析，长期以来，教育心理学家不断丰富、发展了因子分析理论和方法，并应用这一方法在行为科学领域进行了广泛的研究。

因子分析可以看成主成分分析的推广，它也是多元统计分析中常用的一种降维方式，因子分析所涉及的计算与主成分分析也很类似，但差别也是很明显的：1）主成分分析把方差划分为不同的正交成分，而因子分析则把方差划归为不同的起因因子；2）因子分析中特征值的计算只能从相关系数矩阵出发，且必须将主成分转换成因子。

因子分析有确定的模型，观察数据在模型中被分解为公共因子、特殊因子和误差三部分。初学因子分析的最大困难在于理解它的模型，我们先看如下几个例子。

例6 为了解学生的知识和能力，对学生进行了抽样命题考试，考题包括的面很广，但总的来讲可归结为学生的语文水平、数学推导、艺术修养、历史知识、生活知识等五

个方面，我们把每一个方面称为一个（公共）因子，显然每个学生的成绩均可由这五个因子来确定，即可设想第i个学生考试的分数X_i能用这五个公共因子$F_1,F_2,\cdots,F_5$的线性组合表示出来

$$X_i=\mu_i+a_{i1}F_1+a_{i2}F_2+\cdots+a_{i5}F_5+U_i,\ (i=1,2,\cdots,N) \tag{34}$$

线性组合系数$a_{i1},a_{i2},\cdots,a_{i5}$称为因子载荷（loadings），它分别表示第$i$个学生在这五个因子方面的能力；$\mu_i$是总平均，$U_i$是第$i$个学生的能力和知识不能被这五个因子包含的部分，称为特殊因子，常假定$U_i\sim N(0,\sigma_i^2)$，不难发现，这个模型与回归模型在形式上是很相似的，但这里$F_1,F_2,\cdots,F_5$的值却是未知的，有关参数的意义也有很大的差异。

因子分析的首要任务就是估计因子载荷a_{ij}的方差σ_i^2，然后给因子F_i一个合理的解释，若难以进行合理的解释，则需要进一步作因子旋转，希望旋转后能发现比较合理的解释。

例7　诊断时，医生检测了病人的五个生理指标：收缩压、舒张压、心跳间隔、呼吸间隔和舌下温度，但依据生理学知识，这五个指标是受植物神经支配的，植物神经又分为交感神经和副交感神经，因此这五个指标可用交感神经和副交感神经两个公共因子来确定，从而也构成了因子模型。

例8　Holjinger和Swineford在芝加哥郊区对145名七、八年级学生进行了24个心理测验，通过因子分析，这24个心理指标被归结为4个公共因子，即词语因子、速度因子、推理因子和记忆因子。

特别需要说明的是这里的因子和试验设计里的因子（或因素）是不同的，它比较抽象和概括，往往是不可以单独测量的。

5.1　因子分析模型

设有p个原始变量$x_i(i=1,2,\cdots,p)$，它们可能相关，也可能独立，将x_i标准化得到新变量z_i，则可以建立因子分析模型如下：

$$z_i=a_{i1}F_1+a_{i2}F_2+\cdots+a_{im}F_m+c_iU_i\ (i=1,2,\cdots,p), \tag{35}$$

其中$F_j(j=1,2,\cdots,m)$出现在每个变量的表达式中，称为公共因子，它们的含义要根据具体问题来解释，$U_i(i=1,2,\cdots,p)$仅与变量z_i有关，称为特殊因子，系数a_{ij},c_i

（$i=1,2,\cdots,p$， $j=1,2,\cdots,m$）称为因子载荷，$A=(a_{ij})$称为载荷矩阵。

可以将（35）式表示为如下的矩阵形式

$$z=AF+CU \tag{36}$$

其中 $z=(z_1,z_2,\cdots,z_p)^T$， $F=(F_1,F_2,\cdots,F_m)^T$， $U=(U_1,U_2,\cdots,U_p)^T$，

$$A=(a_{ij})_{p\times m}，\quad C=\mathrm{diag}(c_1,c_2,\cdots,c_p)$$

对此模型通常需要假设

1）各特殊因子之间以及特殊因子与所有公共因子之间均相互独立，即

$$\begin{cases}\mathrm{Cov}(U)=\mathrm{diag}(\sigma_1^2,\sigma_2^2,\cdots,\sigma_p^2)\\ \mathrm{Cov}(F,U)=0\end{cases} \tag{37}$$

2）各公共因子都是均值为0，方差为1的独立正态随机变量，其协方差矩阵为单位阵 I_m，即 $F\sim N(0,I_m)$。当因子 F 的各个分量相关时，$\mathrm{Cov}(F)$ 不再是对角阵，这样的模型称为斜交因子模型，我们不考虑这种模型。

m 个公共因子对第 i 个变量方差的贡献称为第 i 共同度，记为 h_i^2，

$$h_i^2=a_{i1}^2+a_{i2}^2+\cdots+a_{im}^2 \tag{38}$$

而特殊因子的方差称为特殊方差或者特殊值（即（37）式中的 σ_i^2， $i=1,2,\cdots,p$），从而第 i 个变量的方差有如下分解

$$\mathrm{Varz}_i=h_i^2+\sigma_i^2，\quad i=1,2,\cdots,p \tag{39}$$

因子分析的一个基本问题是如何估计因子载荷，亦即如何求解因子模型（35），我们下面仅仅介绍最常用的基于样本相关系数矩阵 R 的主成分解。

设 $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p$ 为样本相关系数矩阵 R 的特征值，$\eta_1,\eta_2,\cdots,\eta_p$ 为相应的标准正交化特征向量。设 $m<p$，则样本相关系数矩阵 R 的主成分因子分析的载荷矩阵 A 为

$$A=(\sqrt{\lambda_1}\eta_1,\sqrt{\lambda_2}\eta_2,\cdots,\sqrt{\lambda_m}\eta_m)， \tag{40}$$

特殊因子的方差用 $R-AA^T$ 的对角元来估计，即

$$\sigma_i^2 = 1 - \sum_{j=1}^{m} a_{ij}^2 \qquad (41)$$

例9（续例5） 我们考虑样本相关系数矩阵 R 的前两个样本主成分，对 $m=1$ 和 $m=2$，因子分析主成分解见表10，对 $m=2$，残差矩阵 $R - AA^T - \mathrm{Cov}(U)$ 为

$$\begin{bmatrix} 0 & -0.1274 & -0.1643 & -0.0689 & 0.0173 \\ -0.1274 & 0 & -0.1223 & 0.0553 & 0.0118 \\ -0.1643 & -0.1234 & 0 & -0.0193 & -0.0171 \\ -0.0689 & 0.0553 & -0.0193 & 0 & -0.2317 \\ 0.0173 & 0.0118 & -0.0171 & -0.2317 & 0 \end{bmatrix}$$

表10　因子分析主成分解

变量	一个因子		两个因子		
	因子载荷估计 F_1	特殊方差	因子载荷估计		特殊方差
			F_1	F_2	
1	0.7836	0.3860	0.7836	-0.2162	0.3393
2	0.7726	0.4031	0.7726	-0.4581	0.1932
3	0.7947	0.3685	0.7947	-0.2343	0.3136
4	0.7123	0.4926	0.7123	0.4729	0.2690
5	0.7119	0.4931	0.7119	0.5235	0.2191
累积贡献	0.571342		0.571342	0.733175	

由这两个因子解释的总方差比一个因子大很多。然而，对 $m=2$，残差矩阵负元素较多，这表明 AA^T 产生的数比 R 中对应元素（相关系数）要大。

第一个因子 F_1 代表了一般经济条件，称为市场因子，所有股票在这个因子上的载荷都比较大，且大致相等，第二个因子是化学股和石油股的一个对照，两者分别有比较大的负、正载荷。可见 F_2 使不同的工业部门的股票产生差异，通常称之为工业因子。归纳起来，我们有如下结论：股票回升率由一般经济条件、工业部门活动和各公司本身特殊活动三部分决定，这与例5的结论基本一致。

计算的MATLAB程序如下：

```
clc,clear
r=[1.000 0.577 0.509 0.387 0.462
   0.577 1.000 0.599 0.389 0.322
```

```
    0.509  0.599  1.000  0.436  0.426
    0.387  0.389  0.436  1.000  0.523
    0.462  0.322  0.426  0.523  1.000];
[vec,val,con]=pcacov(r);
f1=repmat(sign(sum(vec)),size(vec,1),1);
vec=vec.*f1;       %特征向量正负号转换
f2=repmat(sqrt(val)',size(vec,1),1);
a=vec.*f2
a1=a(:,1);    %一个因子的载荷矩阵
tm=r-a1*a1';
tcha1=diag(tm)    %一个因子的特殊方差
a2=a(:,[1,2]);  %两个因子的载荷矩阵
tm=r-a2*a2';
tcha2=diag(tm)  %两个因子的特殊方差
ccha2=r-a2*a2'-diag(tcha2)  %求两个因子时的残差矩阵
gong=cumsum(con)       %求累积贡献率
```

5.2 因子旋转

上面主成分解是不唯一的，因为对 A 作任何正交变换都不会改变原来的 AA^T，即设 Q 为 m 阶正交矩阵，$B=AQ$ 则有 $BB^T=AA^T$，载荷矩阵的这种不唯一性表明看是不利的，但我们却可以利用这种不变性，通过适当的因子变换，使变换后新的因子具有更鲜明的实际意义或可解释性，比如，我们可以通过正交变换使 B 中有尽可能多的元素等于或接近于0，从而使因子载荷矩阵结构简单化，便于做出更有实际意义的解释。

由于正交变换是一种旋转变换，如果我们选取方差最大的正交旋转，即将各个因子旋转到某个位置，使每个变量在旋转后的因子轴上的投影向最大、最小两级分化，从而使每个因子中的高载荷只出现在少数的变量上，在最后得到的旋转因子载荷矩阵中，每列元素除几个值外，其余的均接近于0。

5.2.1 考虑两个因子的平面正交旋转

设因子载荷矩阵为

$$A=(a_{ij}),\ i=1,2,\cdots,p,\ \ j=1,2 \tag{42}$$

取正交矩阵

$$Q=\begin{pmatrix}\cos\phi & -\sin\phi\\ \sin\phi & \cos\phi\end{pmatrix} \tag{43}$$

这是逆时针旋转，如作顺时针旋转，只需将（43）式次对角线上的两个元素对换即可。并记

$$B = AQ = (b_{ij})\,,\ i = 1,2,\cdots,p\,,\ \ j = 1,2 \tag{44}$$

称 B 为旋转因子载荷矩阵，此时模型（36）变为

$$z = B(Q^T F) + CU \tag{45}$$

同时，公共因子 F 也随之变为 $Q^T F$，现在希望通过旋转，将变量分为主要由不同因子说明的两个部分，因此，要求 $(b_{11}^2, b_{21}^2, \cdots, b_{p1}^2)$ 和 $(b_{12}^2, b_{22}^2, \cdots, b_{p2}^2)^T$ 这两列数据分别求得的方差尽可能的大。

下面考虑相对方差

$$V_j = \frac{1}{p}\sum_{i=1}^{p}\left(\frac{b_{ij}^2}{h_i^2}\right)^2 - \left(\frac{1}{p}\sum_{i=1}^{p}\frac{b_{ij}^2}{h_i^2}\right)^2\,,\ \ j = 1,2 \tag{46}$$

取 b_{ij} 是为了消除 b_{ij} 符号的影响，除以 h_i^2 是为了消除各个变量对公共因子依赖程度不同的影响，正交旋转的目的是为了使总方差 $V = V_1 + V_2$ 达到最大。令 $\frac{dV}{d\phi} = 0$，经计算，ϕ 应满足

$$\tan 4\phi = \frac{D_0 - 2A_0B_0 / p}{C_0 - (A_0^2 - B_0^2)/p} \tag{47}$$

其中

$$\begin{cases} A_0 = \sum_{i=1}^{p} u_i, & B_0 = \sum_{i=1}^{p} v_i \\ C_0 = \sum_{i=1}^{p} (u_i^2 - v_i^2), & D_0 = 2\sum_{i=1}^{p} u_i v_i \\ u_i = \left(\frac{a_{i1}}{h_i}\right)^2 - \left(\frac{a_{i2}}{h_i}\right)^2, & v_i = \frac{2a_{i1}a_{i2}}{h_i^2} \end{cases} \tag{48}$$

当 $m = 2$ 时，还可以通过图解法，凭直觉将坐标轴旋转一个角度 ϕ，一般的做法是先对变量聚类，利用这些类很容易确定新的公共因子。

5.2.2　公共因子数 $m > 2$ 的情形

可以每次考虑不同的两个因子的旋转，从 m 个因子中每次选两个旋转，共有

$m(m-1)/2$种选择，这样共有$m(m-1)/2$旋转，做完这$m(m-1)/2$次旋转就算完成了一个循环，然后重新开始第二个循环，每经一个循环，A阵的各列的相对方差和V只会变大，当第k次循环后的$V^{(k)}$与上一次循环的$V^{(k-1)}$比较变化不大时，就停止旋转。

例10　设某三个变量的样本相关系数矩阵为

$$R=\begin{pmatrix} 1 & -1/3 & 2/3 \\ -1/3 & 1 & 0 \\ 2/3 & 0 & 1 \end{pmatrix}$$

试从R出发，作因子分析。

解　1）求R的特征值及其相应的特征向量。

由特征方程$\det(R-\lambda I)=0$可得三个特征值，依大小次序记为$\lambda_1=1.7454$，

$\lambda_2=1$，$\lambda_3=0.2546$，由于前面两个特征值的累积方差贡献率已达91.51%，因而只要取两个主因子就行了，下面给出了前两个特征值对应的特征向量：

$$\eta_1^T=(0.7071, 0.3162, -0.6325)$$

$$\eta_2^T=(0, 0.8944, 0.4472)$$

2）求因子载荷矩阵A

由（40）式即可算出

$$A=\begin{pmatrix} 0.9342 & 0 \\ -0.4178 & 0.8944 \\ 0.8355 & 0.4472 \end{pmatrix}$$

3）对载荷矩阵A作正交旋转

对载荷矩阵A作正交旋转，使得到的矩阵$A_1=AQ$的方差和最大。计算结果为

$$Q=\begin{pmatrix} 0.9320 & -0.3625 \\ 0.3625 & 0.9320 \end{pmatrix},\quad A_1=\begin{pmatrix} 0.8706 & -0.3386 \\ -0.0651 & 0.9850 \\ 0.9408 & 0.1139 \end{pmatrix}$$

求解的MATLAB程序如下：

```
clc,clear
r=[1 -1/3 2/3;-1/3 1 0;2/3 0 1];
[vec,val,con]=pcacov(r);num=2;
fl=repmat(sign(sum(vec)),size(vec,1),1);
```

```
vec=vec.*f1;      %特征向量正负号转换
f2=repmat(sqrt(val)',size(vec,1),1);
a=vec.*f2    %载荷矩阵
[b,t]=rotatefactors(a(:,1:num),'method', 'varimax')
```

例11　在一项关于消费者爱好的研究中，随机的邀请一些顾客对某种新食品进行评价，共有5项指标（变量，1－味道，2－价格，3－风味，4－适于快餐，5－能量补充），均采用7级打分法，它们的相关系数矩阵

$$R=\begin{pmatrix} 1 & 0.02 & 0.96 & 0.42 & 0.01 \\ 0.02 & 1 & 0.13 & 0.71 & 0.85 \\ 0.96 & 0.13 & 1 & 0.5 & 0.11 \\ 0.42 & 0.71 & 0.5 & 1 & 0.79 \\ 0.01 & 0.85 & 0.11 & 0.79 & 1 \end{pmatrix}$$

从相关系数矩阵R可以看出，变量1和3、2和5各成一组，而变量4似乎更接近（2，5）组，于是，我们可以期望，因子模型可以取两个、至多三个公共因子。

R的前两个特征值为2.8531和1.8063，其余三个均小于1，这两个公共因子对样本方差的累计贡献率为0.9319，于是，我们选$m=2$，因子载荷、贡献率和特殊方差的估计列入表11中。

表11　因子分析表

变量	因子载荷估计		旋转因子载荷估计		共同度	特殊方差（未旋转）
	F_1	F_2	Q^TF_1	Q^TF_2		
1	0.5599	0.8161	0.027	0.9854	0.9795	0.0205
2	0.7773	-0.5242	0.8734	0.0034	0.8789	0.1211
3	0.6453	0.7479	0.1329	0.9705	0.9759	0.0241
4	0.9391	-0.1049	0.8178	0.4035	0.8929	0.1071
5	0.7982	-0.5432	0.9734	-0.0179	0.9322	0.0678
特征值	2.8531	1.8063				
累积贡献	57.0618	93.1885				

因为$AA^T+\mathrm{Cov}(U)$与R比较接近，所以从直观上，我们可以认为两个因子的模型给出了数据较好的拟合。另一方面，五个贡献值都比较大，表明了这两个公共因子确实解释了每个变量方差的绝大部分。

很明显，变量2，4，5在Q^TF_1上有大载荷，而在Q^TF_2上的载荷较小或可忽略。相反，变量1，3在Q^TF_2上有大载荷，而在Q^TF_1上的载荷却是可以忽略。因此，我们有

理由称Q^TF_1为营养因子，Q^TF_2为滋味因子。旋转的效果一目了然。

计算的MATLAB程序如下：

```
clc,clear
load li11.txt  %把原始的相关系数矩阵保存在纯文本文件li11.txt中
r=li11;num=2; %num为因子的个数
[vec,val,con]=pcacov(r);
f1=repmat(sign(sum(vec)),size(vec,1),1);
vec=vec.*f1;      %特征向量正负号转换
f2=repmat(sqrt(val)',size(vec,1),1);
a=vec.*f2
a1=a(:,[1:num])    %因子的载荷矩阵
tm=r-a1*a1';
tcha=diag(tm)    %因子的特殊方差
ccha=r-a1*a1'-diag(tcha)  %求残差矩阵
gong=cumsum(con(1:num))      %求累积贡献率
[b1,b2]=factoran(r,2,'xtype','cov','rotate','varimax')    %求旋转因子载荷矩阵和特殊方差
```

在因子分析中，一般人们的重点是估计因子模型的参数，即载荷矩阵，有时公共因子的估计，即所谓因子得分，也是需要的，因子得分可以用于模型诊断，也可以作下一步分析的原始数据，需要指出的是，因子得分的计算并不是通常意义下的参数估计，它是对不可观测的随机向量F_i取值的估计。通常可以用加权最小二乘法和回归法来估计因子得分。

§6　因子分析案例

因子分析(factor analysis)是一种数据简化的技术。它通过研究众多变量之间的内部依赖关系，探求观测数据中的基本结构，并用少数几个假想变量来表示其基本的数据结构。这几个假想变量能够反映原来众多变量的主要信息。原始的变量是可观测的显在变量，而假想变量是不可观测的潜在变量，称为因子。

因子分析与回归分析不同，因子分析中的因子是一个比较抽象的概念，而回归因子有非常明确的实际意义。

主成分分析分析与因子分析也有不同，主成分分析仅仅是变量变换，而因子分析需要构造因子模型。

主成分分析:原始变量的线性组合表示新的综合变量，即主成分。

因子分析：潜在的假想变量和随机影响变量的线性组合表示原始变量。

下面我们首先总结一下因子分析的原理。

6.1　因子分析的原理

6.1.1　因子分析模型

1. 数学模型

设 $X_i(i=1,2,\cdots,p)$ 个变量，如果表示为

$$X_i=\mu_i+a_{i1}F_1+\cdots+a_{im}F_m+\varepsilon_i\text{，（}m\le p\text{）}\qquad(49)$$

或

$$\begin{bmatrix}X_1\\X_2\\\vdots\\X_p\end{bmatrix}=\begin{bmatrix}\mu_1\\\mu_2\\\vdots\\\mu_p\end{bmatrix}+\begin{bmatrix}a_{11}&a_{12}&\cdots&a_{1m}\\a_{21}&a_{22}&\cdots&a_{2m}\\\vdots&\vdots&&\vdots\\a_{p1}&a_{p2}&\cdots&a_{pm}\end{bmatrix}\begin{bmatrix}F_1\\F_2\\\vdots\\F_m\end{bmatrix}+\begin{bmatrix}\varepsilon_1\\\varepsilon_2\\\vdots\\\varepsilon_p\end{bmatrix}$$

或

$$X-\mu=AF+\varepsilon$$

其中

$$X=\begin{bmatrix}X_1\\X_2\\\vdots\\X_p\end{bmatrix},\ \mu=\begin{bmatrix}\mu_1\\\mu_2\\\vdots\\\mu_p\end{bmatrix},\ A=\begin{bmatrix}a_{11}&a_{12}&\cdots&a_{1m}\\a_{21}&a_{22}&\cdots&a_{2m}\\\vdots&\vdots&&\vdots\\a_{p1}&a_{p2}&\cdots&a_{pm}\end{bmatrix},\ \varepsilon=\begin{bmatrix}\varepsilon_1\\\varepsilon_2\\\vdots\\\varepsilon_p\end{bmatrix}$$

称 $F_1,F_2,\cdots,F_p$ 为公共因子，是不可观测的变量，它们的系数称为载荷因子。ε_i 是特殊因子，是不能被前 m 个公共因子包含的部分。并且满足

$E(F)=0$，$E(\varepsilon)=0$，$\mathrm{Cov}(F)=I_m$，

$D(\varepsilon)=Cov(\varepsilon)=\mathrm{diag}(\sigma_1^2,\sigma_2^2,\cdots,\sigma_m^2)$，$\mathrm{cov}(F,\varepsilon)=0$。

2．因子分析模型的性质

（1）原始变量 X 的协方差矩阵的分解

由 $X-\mu=AF+\varepsilon$，得 $\mathrm{Cov}(X-\mu)=A\mathrm{Cov}(F)A^T+\mathrm{Cov}(\varepsilon)$，即

$$Cov(X)=AA^T+\mathrm{diag}(\sigma_1^2,\sigma_2^2,\cdots,\sigma_m^2)$$

$\sigma_1^2,\sigma_2^2,\cdots,\sigma_m^2$ 的值越小，则公共因子共享的成分越多。

（2）载荷矩阵不是唯一的

设 T 为一个 $p\times p$ 的正交矩阵，令 $\tilde{A}=AT$，$\tilde{F}=T^TF$，则模型可以表示为

$$X = \mu + \tilde{A}\tilde{F} + \varepsilon$$

3．因子载荷矩阵中的几个统计性质

（1）因子载荷 a_{ij} 的统计意义

因子载荷 a_{ij} 是第 i 个变量与第 j 个公共因子的相关系数，反映了第 i 个变量与第 j 个公共因子的相关重要性。绝对值越大，相关的密切程度越高。

（2）变量共同度的统计意义

变量 X_i 的共同度是因子载荷矩阵的第 i 行的元素的平方和。记为 $h_i^2 = \sum_{j=1}^{m} a_{ij}^2$ 。对（49）式两边求方差，得

$$\mathrm{Var}(X_i) = a_{i1}^2\mathrm{Var}(F_1) + \cdots + a_{im}^2\mathrm{Var}(F_m) + \mathrm{Var}(\varepsilon_i)$$

即

$$1 = \sum_{j=1}^{m} a_{ij}^2 + \sigma_i^2$$

可以看出所有的公共因子和特殊因子对变量 X_i 的贡献为1。如果 $\sum_{j=1}^{m} a_{ij}^2$ 非常靠近1，σ_i^2 非常小，则因子分析的效果好，从原变量空间到公共因子空间的转化效果好。

（3）公共因子 F_j 方差贡献的统计意义

因子载荷矩阵中各列元素的平方和

$$S_j = \sum_{i=1}^{p} a_{ij}^2$$

称为 $F_j(j = 1,2,\cdots,m)$ 对所有的 X_i 的方差贡献和。衡量 F_j 的相对重要性。

6.1.2　因子载荷矩阵的估计方法

1.主成分分析法

见第五节。

2．主因子法

主因子方法是对主成分方法的修正，假定我们首先对变量进行标准化变换。则

$$R = AA^T + D$$

$$R^* = AA^T = R - D$$

称 R^* 为约相关系数矩阵，R^* 对角线上的元素是 h_i^2，而不是1。

$$R^* = R - \hat{D} = \begin{bmatrix} \hat{h}_1^2 & r_{12} & \cdots & r_{1p} \\ r_{21} & \hat{h}_2^2 & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & \hat{h}_p^2 \end{bmatrix}$$

直接求 R^* 的前 p 个特征值和对应的正交特征向量。得到如下的矩阵

$$A = [\sqrt{\lambda_1^*}u_1^* \quad \sqrt{\lambda_2^*}u_2^* \quad \cdots \quad \sqrt{\lambda_p^*}u_p^*]$$

其中 R^* 的特征值：$\lambda_1^* \geq \lambda_2^* \geq \cdots \geq \lambda_p^*$，对应的正交特征向量为 $u_1^*, u_2^*, \cdots, u_p^*$。

在实际应用中，特殊因子的方差一般都是未知的，可以通过一组样本来估计。估计的方法有如下几种：

1）取 $\hat{h}_i^2 = 1$，在这个情况下主因子解与主成分解等价。

2）取 $\hat{h}_i^2 = R_i^2$，R_i^2 为 x_i 与其它所有的原始变量 x_j 的复相关系数的平方，即 x_i 对其余的 $p-1$ 个 x_j 的回归方程的判定系数，这是因为 x_i 与公共因子的关系是通过其余的 $p-1$ 个 x_j 的线性组合联系起来的。

3）取 $\hat{h}_i^2 = \max|r_{ij}|(j \neq i)$，这意味着取 x_i 与其余的 x_j 的简单相关系数的绝对值最大者。

4）取 $\hat{h}_i^2 = \dfrac{1}{p-1}\sum\limits_{\substack{j=1 \\ j\neq i}}^{p} r_{ij}$，其中要求该值为正数。

5）取 $\hat{h}_i^2 = 1/r^{ii}$，其中 r^{ii} 是 R^{-1} 的对角元素。

3．极大似然估计法（略）

例12　假定某地固定资产投资率 x_1，通货膨胀率 x_2，失业率 x_3，相关系数矩阵为

$$\begin{bmatrix} 1 & 1/5 & -1/5 \\ 1/5 & 1 & -2/5 \\ -1/5 & -2/5 & 1 \end{bmatrix}$$

试用主成分分析法求因子分析模型。

解　特征值为 $\lambda_1=1.5464$，$\lambda_2=0.8536$，$\lambda_3=0.6$，特征向量

$$u_1=\begin{bmatrix}0.4597\\0.628\\-0.628\end{bmatrix},\quad u_2=\begin{bmatrix}0.8881\\-0.3251\\0.3251\end{bmatrix},\quad u_3=\begin{bmatrix}0\\0.7071\\0.7071\end{bmatrix}$$

载荷矩阵

$$A=\left[\sqrt{\lambda_1}u_1\quad\sqrt{\lambda_2}u_2\quad\sqrt{\lambda_3}u_3\right]=\begin{bmatrix}0.5717&0.8205&0\\0.7809&-0.3003&0.5477\\-0.7809&0.3003&0.5477\end{bmatrix}$$

$$x_1=0.5717F_1+0.8205F_2$$

$$x_2=0.7809F_1-0.3003F_2+0.5477F_3$$

$$x_3=-0.7809F_1+0.3003F_2+0.5477F_3$$

可取前两个因子 F_1 和 F_2 为公共因子，第一公因子 F_1 为物价因子，对 X 的贡献为1.5464，第二公因子 F_2 为投资因子，对 X 的贡献为0.8536。共同度分别为1，0.7，0.7。

计算的MATLAB程序为：

```
clc,clear
r=[1 1/5 -1/5;1/5 1 -2/5;-1/5 -2/5 1];
[vec,val,con]=pcacov(r);num=2;
f1=repmat(sign(sum(vec)),size(vec,1),1);
vec=vec.*f1;      %特征向量正负号转换
f2=repmat(sqrt(val)',size(vec,1),1);
a=vec.*f2    %载荷矩阵
s1=sum(a.^2,1)
tt=a.^2;tt=tt(:,1:num);
s2=sum(tt,2)
```

例13　假定某地固定资产投资率 x_1，通货膨胀率 x_2，失业率 x_3，相关系数矩阵为

$$\begin{bmatrix}1&1/5&-1/5\\1/5&1&-2/5\\-1/5&-2/5&1\end{bmatrix}$$

试用主因子分析法求因子分析模型。

解　假定用 $\hat{h}_i^2 = \max|r_{ij}|(j \neq i)$ 代替初始的 h_i^2。则有 $h_1^2 = \frac{1}{5}$，$h_2^2 = \frac{2}{5}$，$h_3^2 = \frac{2}{5}$ 。

$$R^* = \begin{bmatrix} 1/5 & 1/5 & -1/5 \\ 1/5 & 2/5 & -2/5 \\ -1/5 & -2/5 & 2/5 \end{bmatrix}$$

特征值为 $\lambda_1 = 0.9123$，$\lambda_2 = 0.0877$，$\lambda_3 = 0$。非零特征值对应的特征向量为

$$u_1 = \begin{bmatrix} 0.369 \\ 0.6572 \\ -0.6572 \end{bmatrix}，\quad u_2 = \begin{bmatrix} 0.9294 \\ -0.261 \\ 0.261 \end{bmatrix}$$

取两个主因子，求得载荷矩阵

$$A = \begin{bmatrix} 0.3525 & 0.2752 \\ 0.6277 & -0.0773 \\ -0.6277 & 0.0773 \end{bmatrix}$$

6.1.3　因子旋转（正交变换）

建立因子分析数学模型目的不仅仅要找出公共因子以及对变量进行分组，更重要的要知道每个公共因子的意义，以便进行进一步的分析，如果每个公共因子的含义不清，则不便于进行实际背景的解释。由于因子载荷阵是不唯一的，所以应该对因子载荷阵进行旋转。目的是使因子载荷阵的结构简化，使载荷矩阵每列或行的元素平方值向0和1两级分化。有三种主要的正交旋转法，四次方最大法、方差最大法和等量最大法。

1．方差最大法

方差最大法从简化因子载荷矩阵的每一列出发，使和每个因子有关的载荷的平方的方差最大。当只有少数几个变量在某个因子上有较高的载荷时，对因子的解释最简单。方差最大的直观意义是希望通过因子旋转后，使每个因子上的载荷尽量拉开距离，一部分的载荷趋于±1，另一部分趋于0。

2．四次方最大旋转

四次方最大旋转是从简化载荷矩阵的行出发，通过旋转初始因子，使每个变量只在一个因子上有较高的载荷，而在其它的因子上尽可能低的载荷。如果每个变量只在一个因子上有非零的载荷，这时的因子解释是最简单的。

四次方最大法通过使因子载荷矩阵中每一行的因子载荷平方的方差达到最大。

3．等量最大法

等量最大法把四次方最大法和方差最大法结合起来，求它们的加权平均最大。

6.2　因子得分

1．因子得分的概念

前面我们主要解决了用公共因子的线性组合来表示一组观测变量的有关问题。如果

我们要使用这些因子做其他的研究，比如把得到的因子作为自变量来做回归分析，对样本进行分类或评价，这就需要我们对公共因子进行测度，即给出公共因子的值。

因子分析的数学模型为：

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pm} \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix}$$

原变量被表示为公共因子的线性组合，当载荷矩阵旋转之后，公共因子可以做出解释，通常的情况下，我们还想反过来把公共因子表示为原变量的线性组合。

因子得分函数

$$F_j = \beta_{j1}X_1 + \cdots + \beta_{jp}X_p, \quad j = 1,2,\cdots,m$$

可见，要求得每个因子的得分，必须求得分函数的系数，而由于 $p > m$，所以不能得到精确的得分，只能通过估计。

（1）巴特莱特因子得分(加权最小二乘法）

把 $x_i - \mu_i$ 看作因变量，把因子载荷矩阵

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pm} \end{bmatrix}$$

看成自变量的观测。

$$\begin{cases} x_{i1} - \mu_1 = a_{11}f_1 + a_{12}f_2 + \cdots + a_{1m}f_m + \varepsilon_1 \\ x_{i2} - \mu_2 = a_{21}f_1 + a_{22}f_2 + \cdots + a_{2m}f_m + \varepsilon_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ x_{ip} - \mu_p = a_{p1}f_1 + a_{p2}f_2 + \cdots + a_{pm}f_m + \varepsilon_p \end{cases}$$

由于特殊因子的方差相异，所以用加权最小二乘法求得分。使

$$\sum_{j=1}^{p}[(x_{ij} - \mu_i) - (a_{i1}\hat{f}_1 + a_{i2}\hat{f}_2 + \cdots a_{im}\hat{f}_m)]^2 / \sigma_i^2$$

最小的 $\hat{f}_1,\cdots,\hat{f}_m$ 是相应个案的因子得分。

用矩阵表达有

$$x - \mu = AF + \varepsilon$$

则要使

$$(x-\mu-AF)^T D^{-1}(x-\mu-AF) \qquad (50)$$

达到最小，其中

$$D=\begin{bmatrix}\sigma_1^{-2} & & \\ & \ddots & \\ & & \sigma_p^{-2}\end{bmatrix}$$

使（50）式取得最小值的 F 是相应个案的因子得分。

计算得 F 满足

$$A^T D^{-1}F = A^T D^{-1}A(x-\mu)$$

解之得

$$\hat{F} = (A^T D^{-1}A)^{-1}A^T D^{-1}(x-\mu)$$

（2）回归方法

下面我们简单介绍一下回归方法的思想。

不妨设

$$\begin{bmatrix}X_1\\X_2\\\vdots\\X_p\end{bmatrix}=\begin{bmatrix}a_{11}&a_{12}&\cdots&a_{1m}\\a_{21}&a_{22}&\cdots&a_{2m}\\\vdots&\vdots&&\vdots\\a_{p1}&a_{p2}&\cdots&a_{pm}\end{bmatrix}\begin{bmatrix}F_1\\F_2\\\vdots\\F_m\end{bmatrix}$$

则有

$$\hat{F}_j = b_{j1}X_1+\cdots+b_{jp}X_p,\quad j=1,2,\cdots,m$$

记

$$\begin{bmatrix}b_{11}&b_{12}&\cdots&b_{1p}\\b_{21}&b_{22}&\cdots&b_{2p}\\\vdots&\vdots&&\vdots\\b_{m1}&b_{m2}&\cdots&b_{mp}\end{bmatrix}=\begin{bmatrix}b_1\\b_2\\\vdots\\b_m\end{bmatrix}$$

$$a_{ij}=\gamma_{X_iF_j}=E(X_iF_j)=E[X_i(b_{j1}X_1+\cdots+b_{jp}X_p)]$$

$$=b_{j1}\gamma_{i1}+\cdots+b_{jp}\gamma_{ip}=[\gamma_{i1}\quad\gamma_{i2}\quad\cdots\quad\gamma_{ip}]\begin{bmatrix}b_{j1}\\b_{j2}\\\vdots\\b_{jp}\end{bmatrix}$$

则我们有如下的方程组

$$\begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1p} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2p} \\ \vdots & \vdots & & \vdots \\ \gamma_{p1} & b_{p2} & \cdots & b_{pp} \end{bmatrix} \begin{bmatrix} b_{j1} \\ b_{j2} \\ \vdots \\ b_{jp} \end{bmatrix} = \begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{pj} \end{bmatrix}$$

其中

$$\begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1p} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2p} \\ \vdots & \vdots & & \vdots \\ \gamma_{p1} & b_{p2} & \cdots & b_{pp} \end{bmatrix}, \begin{bmatrix} b_{j1} \\ b_{j2} \\ \vdots \\ b_{jp} \end{bmatrix}, \begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{pj} \end{bmatrix}$$

分别为原始变量的相关系数矩阵，第 j 个因子得分函数的系数，载荷矩阵的第 j 列。

用矩阵表示有

$$[b_1^T \quad b_2^T \quad \cdots \quad b_m^T] = R^{-1}A$$

6.3　因子分析的步骤

1．选择分析的变量

用定性分析和定量分析的方法选择变量，因子分析的前提条件是观测变量间有较强的相关性，因为如果变量之间无相关性或相关性较小的话，他们不会有共享因子，所以原始变量间应该有较强的相关性。

2．计算所选原始变量的相关系数矩阵

相关系数矩阵描述了原始变量之间的相关关系。可以帮助判断原始变量之间是否存在相关关系，这对因子分析是非常重要的，因为如果所选变量之间无关系，做因子分析是不恰当的。并且相关系数矩阵是估计因子结构的基础。

3．提出公共因子

这一步要确定因子求解的方法和因子的个数。需要根据研究者的设计方案或有关的经验或知识事先确定。因子个数的确定可以根据因子方差的大小。只取方差大于1(或特征值大于1)的那些因子，因为方差小于1的因子其贡献可能很小；按照因子的累计方差贡献率来确定，一般认为要达到60％才能符合要求。

4．因子旋转

通过坐标变换使每个原始变量在尽可能少的因子之间有密切的关系，这样因子解的实际意义更容易解释，并为每个潜在因子赋予有实际意义的名字。

5．计算因子得分

求出各样本的因子得分。有了因子得分值，则可以在许多分析中使用这些因子，例如以因子的得分做聚类分析的变量，做回归分析中的回归因子。

6.4　我国上市公司赢利能力与资本结构的实证分析

已知上市公司的数据见表12。

表12　上市公司数据

公司	销售净利率 x_1	资产净利率 x_2	净资产收益率 x_3	销售毛利率 x_4	资产负利率 x
歌华有线	43.31	7.39	8.73	54.89	15.35
五粮液	17.11	12.13	17.29	44.25	29.69
用友软件	21.11	6.03	7	89.37	13.82
太太药业	29.55	8.62	10.13	73	14.88
浙江阳光	11	8.41	11.83	25.22	25.49
烟台万华	17.63	13.86	15.41	36.44	10.03
方正科技	2.73	4.22	17.16	9.96	74.12
红河光明	29.11	5.44	6.09	56.26	9.85
贵州茅台	20.29	9.48	12.97	82.23	26.73
中铁二局	3.99	4.64	9.35	13.04	50.19
红星发展	22.65	11.13	14.3	50.51	21.59
伊利股份	4.43	7.3	14.36	29.04	44.74
青岛海尔	5.4	8.9	12.53	65.5	23.27
湖北宜化	7.06	2.79	5.24	19.79	40.68
雅戈尔	19.82	10.53	18.55	42.04	37.19
福建南纸	7.26	2.99	6.99	22.72	56.58

试用因子分析法对上述企业进行综合评价。

1. 对原始数据进行标准化处理

假设进行因子分析的指标变量有 p 个：$x_1, x_2, \cdots, x_p$，共有 n 个评价对象，第 i 个评价对象的第 j 个指标的取值为 x_{ij}。将各指标值 x_{ij} 转换成标准化指标 $\tilde{x}_{ij}$，

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \quad (i = 1,2,\cdots,n;\ \ j = 1,2,\cdots,p)$$

其中 $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$，$s_j = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$，（$j = 1,2,\cdots,p$），即 $\bar{x}_j, s_j$ 为第 j 个指标的样本均值和样本标准差。对应地，称

$$\tilde{x}_i = \frac{x_i - \bar{x}_i}{s_i}, \quad (i = 1,2,\cdots,p)$$

为标准化指标变量。

2. 计算相关系数矩阵 R

相关系数矩阵 $R = (r_{ij})_{p\times p}$

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

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．计算初等载荷矩阵

计算相关系数矩阵R的特征值$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p\geq 0$，及对应的特征向量$u_1,u_2,\cdots,u_p$，其中$u_j=(u_{1j},u_{2j},\cdots,u_{nj})^T$，初等载荷矩阵

$$A=[\sqrt{\lambda_1}u_1\quad\sqrt{\lambda_2}u_2\quad\cdots\quad\sqrt{\lambda_p}u_p]$$

4．选择m（$m\leq p$）个主因子，进行因子旋转

根据初等载荷矩阵，计算各个公共因子的贡献率，并选择m个主因子。对提取的因子载荷矩阵进行旋转，得到矩阵$B=\hat{A}T$（其中$\hat{A}$为A的前m列，T为正交矩阵），构造因子模型

$$\begin{cases}\tilde{x}_1=b_{11}F_1+\cdots+b_{1m}F_m\\ \cdots\cdots\cdots\cdots\cdots\cdots\\ \tilde{x}_p=b_{p1}F_1+\cdots+b_{pm}F_m\end{cases}$$

利用MATLAB程序计算得旋转后的因子贡献及贡献率见表13、因子载荷阵见表14。

表13 贡献率数据

因子	贡献	贡献率	累计贡献率
1	1.7794	44.49	44.49
2	1.6673	41.68	86.17

表14 旋转因子分析表

指标	主因子1	主因子2
销售净利率	0.893	0.0082
资产净利率	0.372	0.8854
净资产收益率	-0.2302	0.9386
销售毛利率	0.8892	0.0494

5.计算因子得分，并进行综合评价

我们用回归方法求单个因子得分函数

$$\hat{F}_j=b_{j1}\tilde{x}_1+\cdots+b_{jp}\tilde{x}_p，\quad j=1,2,\cdots,m$$

记

$$\begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1p} \\ b_{21} & b_{22} & \cdots & b_{2p} \\ \vdots & \vdots & & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mp} \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

则有

$$[b_1^T \quad b_2^T \quad \cdots \quad b_m^T] = R^{-1}A$$

计算得各个因子得分函数

$$F_1 = 0.531\tilde{x}_1 + 0.1615\tilde{x}_2 - 0.1831\tilde{x}_3 + 0.5015\tilde{x}_4$$

$$F_2 = -0.045\tilde{x}_1 + 0.5151\tilde{x}_2 + 0.581\tilde{x}_3 - 0.0199\tilde{x}_4$$

利用综合因子得分公式

$$F = \frac{44.49F_1 + 41.68F_2}{86.17}$$

计算出16家上市公司赢利能力的综合得分见表15。

表15　上市公司综合排名表

排名	1	2	3	4	5	6	7	8
F_1	0.0315	0.0025	0.9789	0.4558	-0.0563	1.2791	1.5159	1.2477
F_2	1.4691	1.4477	0.3959	0.8548	1.3577	-0.1564	-0.5814	-0.9729
F	0.7269	0．7016	0．6969	0．6488	0．6277	0．5847	0．5014	0.1735
公司	烟台万华	五粮液	贵州茅台	红星发展	雅戈尔	太太药业	歌华有线	用友软件
排名	9	10	11	12	13	14	15	16
F_1	-0.0351	0.9313	-0.6094	-0.9859	-1.7266	-1.2509	-0.8872	-0.891
F_2	0.3166	-1.1949	0.1544	0.3468	0.2639	-0.7424	-1.1091	-1.2403
F	0.135	-0.0972	-0.2399	-0.3412	-0.7637	-1.0049	-1.1091	-1.2403
公司	青岛海尔	红河光明	浙江阳光	伊利股份	方正科技	中铁二局	福建南纸	湖北宜化

我们通过相关分析，在显著水平为0.05的情况下，得出赢利能力 F 与资产负债率 x 之间的相关系数为-0.6987，这表明两者存在中度相关关系。因子分析法的回归方程为：

$$F = 0.829 - 0.0268x$$

回归方程在显著性水平0.05的情况下，通过了假设检验。

计算的MATLAB程序如下：

```
clc,clear
```

```
load data.txt    %把原始数据保存在纯文本文件data.txt中
data=reshape(data,[16,5]);
m=size(data,1);
x=data(:,5);data=data(:,1:4),num=2;
data=zscore(data); %数据标准化
r=cov(data);
[vec,val,con]=pcacov(r);  %进行主成分分析的相关计算
val,con
f1=repmat(sign(sum(vec)),size(vec,1),1);
vec=vec.*f1;      %特征向量正负号转换
f2=repmat(sqrt(val)',size(vec,1),1);
a=vec.*f2    %载荷矩阵
%如果指标变量多，选取的主因子个数少，可以直接使用factoran进行因子分析
%本题中4个指标变量，选取2个主因子，factoran无法实现
[b,t]=rotatefactors(a(:,1:num),'method', 'varimax')  %旋转变换
bz=[b,a(:,num+1:end)]   %旋转后的载荷矩阵
gx=sum(bz.^2)              %计算因子贡献
gxv=gx/sum(gx)             %计算因子贡献率
dfxsh=inv(r)*b             %计算得分函数的系数
df=data*dfxsh             %计算各个因子的得分
zdf=df*gxv(1:num)'/sum(gxv(1:num))          %对各因子的得分进行加权求和
[szdf,ind]=sort(zdf,'descend')       %对企业进行排名
xianshi=[df(ind,:)';zdf(ind)';ind'] %显示计算结果
[x_zdf_coef,p]=corrcoef([zdf,x])     %计算相关系数
[d1,d1int,d2,d2int,stats]=regress(zdf,[ones(m,1),x]) %回归分析计算
```

6.4　主成分分析法与因子分析法数学模型的异同比较

1．相同点

在以下几方面是相同的：指标的标准化，相关系数矩阵及其特征值和特征向量，用累计贡献率确定主成分、因子个数 m，单个主成分与综合主成分的分析评价、单因子与综合因子的分析评价步骤。

2．不同点

不同之处见表 16。

表 16　主成分分析与因子分析法的不同点

主成分分析数学模型	因子分析的一种数学模型

$F_i = a_{1i}x_1 + a_{2i}x_2 + \cdots a_{pi}x_p$ $= a_i^T x, \quad i = 1,2,\cdots,m$	$x_j = b_{j1}F_1 + b_{j2}F_2 + \cdots + b_{jm}F_m + \varepsilon_j$ $j = 1,2,\cdots,p$
$A = (a_{ij})_{p\times m} = (a_1, a_2, \cdots, a_m)$，$Ra_1 = \lambda_i a_i$ R 为相关系数矩阵，λ_i, a_i 是相应的特征值和单位特征向量，$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$	因子载荷矩阵 $B = (b_{ij})_{p\times m} = \hat{B}C$，$\hat{B} = (\sqrt{\lambda_1}a_1, \cdots, \sqrt{\lambda_m}a_m)$ 为初等因子载荷矩阵（λ_i, a_i 同左），C 为正交旋转矩阵
$A^T A = I$（A 为正交矩阵）	$B^T B \neq I$（B 为非正交阵）
用 A 的第 i 列绝对值大的对应变量对 F_i 命名	将 B 的第 j 列绝对值大的对应变量归为 F_j 一类并由此对 z_j 命名
$\lambda_1, \lambda_2, \cdots \lambda_m$ 互不相同时，a_{ij} 唯一	相关系数 $r_{X_i F_j} = b_{ij}$ 不是唯一的
协方差 $\mathrm{cov}(F_i, F_j) = \lambda_i \delta_{ij}$，$\delta_{ij} = \begin{cases} 0, i \neq j \\ 1, i = j \end{cases}$	协方差 $\mathrm{cov}(F_i, F_j) = \delta_{ij}$，$\delta_{ij} = \begin{cases} 0, i \neq j \\ 1, i = j \end{cases}$
λ_i（特征值）为主成分 F_i 的方差	$v_i = \sum_{k=1}^{p} b_{ki}^2 (\neq \lambda_i)$ 为因子 F_i 对 x 的贡献
主成分 F_j 是由 x 确定的	因子 F_i 是不可观测的
主成分函数 $(F_1, F_2, \cdots, F_m)^T = A^T x$	因子得分函数 $(F_1, F_2, \cdots, F_m)^T = R^{-1}Bx$
主成分 F_i 中 x 的系数平方和 $\sum_{k=1}^{p} a_{ki}^2 = 1$，无特殊因子	$\sum_{i=1}^{m} b_{ji}^2 + \sigma_j^2 = h_j^2 + \sigma_j^2 = 1$，$h_j^2$ 称为共同度，σ_j^2 称为特殊方差
综合主成分函数：$F = \sum_{i=1}^{m} (\lambda_i / p) F_i$，	综合因子得分函数：$F = \sum_{i=1}^{m} (v_i / p) F_i$，

其中 $p=\sum_{i=1}^{m}\lambda_i$	其中 $p=\sum_{i=1}^{m}v_i$

§7　判别分析

判别分析（distinguish analysis）是根据所研究的个体的观测指标来推断该个体所属类型的一种统计方法，在自然科学和社会科学的研究中经常会碰到这种统计问题。例如在地质找矿中我们要根据某异常点的地质结构、化探和物探的各项指标来判断该异常点属于哪一种矿化类型；医生要根据某人的各项化验指标的结果来判断该人属于什么病症；调查了某地区的土地生产率、劳动生产率、人均收入、费用水平、农村工业比重等指标，来确定该地区属于哪一种经济类型地区等等。该方法起源于 1921 年 Pearson 的种族相似系数法，1936 年 Fisher 提出线性判别函数，并形成把一个样本归类到两个总体之一的判别法。

判别问题用统计的语言来表达，就是已有 q 个总体 $X_1,X_2,\cdots,X_q$，它们的分布函数分别为 $F_1(x),F_2(x),\cdots F_q(x)$，每个 $F_i(x)$ 都是 p 维函数。对于给定的样本 X，要判断它来自哪一个总体？当然，应该要求判别准则在某种意义下是最优的，例如错判的概率最小或错判的损失最小等。我们仅介绍最基本的几种判别方法，即距离判别，Bayes 判别和 Fisher 判别。

7.1　距离判别

距离判别是简单、直观的一种判别方法，该方法适用于连续性随机变量的判别类，对变量的概率分布没有什么限制。

1．Mahalanobis 距离的概念

通常我们定义的距离是 Euclid 距离（简称欧氏距离）。但在统计分析与计算中，Euclid 距离就不适用了，看一下下面的例子（见图 6）。

为简单起见，考虑一维 $p=1$的情况。设 $X\sim N(0,1)$，$Y\sim N(4,2^2)$。从图 6 上来看，A 点距 X 的均值 $\mu_1=0$ 较近，距 Y 的均值 $\mu_2=4$ 较远。但从概率角度来分析问题，情况并非如此。经计算，A 点的 x 值为 1.66，也就是说，A 点距 $\mu_1=0$ 是$1.66\sigma_1$，而 A 点距 $\mu_2=4$ 却只有$1.77\sigma_2$，因此，应该认为 A 点距 μ_2 更近一点。

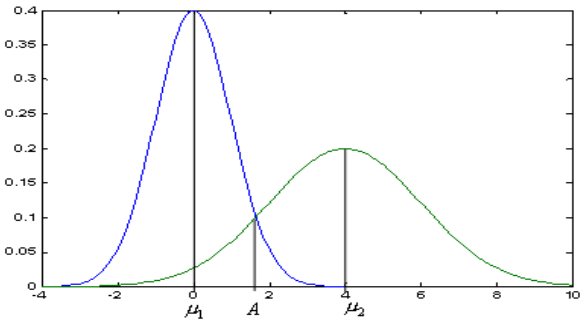

图 6　不同均值、方差的正态分布

定义 2　设 x, y 是从均值为 μ，协方差为 Σ 的总体 A 中抽取的样本，则总体 A 内两点 x 与 y 的 Mahalanobis 距离（简称马氏距离）定义为

$$d(x,y)=\sqrt{(x-y)^T\Sigma^{-1}(x-y)}$$

定义样本 x 与总体 A 的 Mahalanobis 距离为

$$d(x,A)=\sqrt{(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

2．距离判别的判别准则和判别函数

在这里讨论两个总体的距离判别，分协方差相同和协方差不同两种进行讨论。

设总体 A 和 B 的均值向量分别为 μ_1 和 μ_2，协方差阵分别为 Σ_1 和 Σ_2，今给一个样本 x，要判断 x 来自哪一个总体。

首先考虑协方差相同，即

$$\mu_1\neq\mu_2,\quad \Sigma_1=\Sigma_2=\Sigma$$

要判断 x 来自哪一个总体，需要计算 x 到总体 A 和 B Mahalanobis 距离 $d(x,A)$ 和 $d(x,B)$，然后进行比较，若 $d(x,A)\le d(x,B)$，则判定 x 属于 A；否则判定 x 来自 B。由此得到如下判别准则：

$$x\in\begin{cases}A, & d(x,A)\le d(x,B)\\ B, & d(x,A)>d(x,B)\end{cases}$$

现在引进判别函数的表达式，考察 $d^2(x,A)$ 与 $d^2(x,B)$ 之间的关系，有

$$d^2(x,B)-d^2(x,A)=(x-\mu_2)^T\Sigma^{-1}(x-\mu_2)-(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)$$

$$= 2(x-\bar{\mu})^T \Sigma^{-1}(\mu_1-\mu_2)$$

其中 $\bar{\mu}=\dfrac{\mu_1+\mu_2}{2}$ 是两个总体的均值。

令

$$w(x)=(x-\bar{\mu})^T \Sigma^{-1}(\mu_1-\mu_2) \tag{51}$$

称 $w(x)$ 为两总体距离的判别函数，因此判别准则变为

$$x\in\begin{cases}A, & w(x)\geq 0\\ B, & w(x)<0\end{cases}$$

在实际计算中，总体的均值与协方差阵是未知的，因此总体的均值与协方差需要用样本的均值与协方差来代替，设 $x_1^{(1)},x_2^{(1)},\cdots,x_{n_1}^{(1)}$ 是来自总体 A 的 n_1 个样本，$x_1^{(2)},x_2^{(2)},\cdots,x_{n_2}^{(2)}$ 是来自总体 B 的 n_2 个样本，则样本的均值与协方差为

$$\hat{\mu}_i=\bar{x}^{(i)}=\frac{1}{n_i}\sum_{j=1}^{n_i}x_j^{(i)}\ ,\quad j=1,2 \tag{52}$$

$$\hat{\Sigma}=\frac{1}{n_1+n_2-2}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(x_j^{(i)}-\bar{x}^{(i)})(x_j^{(i)}-\bar{x}^{(i)})^T=\frac{1}{n_1+n_2-2}(S_1+S_2) \tag{53}$$

其中

$$S_i=\sum_{j=1}^{n_i}(x_j^{(i)}-\bar{x}^{(i)})(x_j^{(i)}-\bar{x}^{(i)})^T\ ,\quad i=1,2$$

对于待测样本 x，其判别函数定义为

$$\hat{w}(x)=(x-\bar{x})^T\hat{\Sigma}^{-1}(\bar{x}^{(1)}-\bar{x}^{(2)})\ ,$$

其中

$$\bar{x}=\frac{\bar{x}^{(1)}+\bar{x}^{(2)}}{2}$$

其判别准则为

$$x\in\begin{cases}A, & \hat{w}(x)\geq 0\\ B, & \hat{w}(x)<0\end{cases}$$

再考虑协方差不同的情况，即

$\mu_1 \neq \mu_2$， $\Sigma_1 \neq \Sigma_2$

对于样本 x，在方差不同的情况下，判别函数为

$$w(x) = (x-\mu_2)^T \Sigma_2^{-1}(x-\mu_2) - (x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1)$$

与前面讨论的情况相同，在实际计算中总体的均值与协方差是未知的，同样需要用样本的均值与协方差来代替。因此，对于待测样本 x，判别函数定义为

$$\hat{w}(x) = (x-\overline{x}^{(2)})^T \hat{\Sigma}_2^{-1}(x-\overline{x}^{(2)}) - (x-\overline{x}^{(1)})^T \hat{\Sigma}_1^{-1}(x-\overline{x}^{(1)})$$

其中

$$\hat{\Sigma}_i = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(x_j^{(i)} - \overline{x}^{(i)})(x_j^{(i)} - \overline{x}^{(i)})^T = \frac{1}{n_i - 1}S_i \text{， } i = 1,2 \text{。}$$

7.2 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 X_1, X_2，且都有二阶矩存在。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 X_1, X_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 $y = a^T x$，其中 a 为 p 维实向量。设 X_1, X_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 $y = a^T x$ 的均值为

$$\mu_{y_1} = E(y \mid x \in X_1) = a^T \mu_1$$

$$\mu_{y_2} = E(y \mid x \in X_2) = a^T \mu_2$$

其方差为

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

考虑比

$$\frac{(\mu_{y_1} - \mu_{y_2})^2}{\sigma_y^2} = \frac{[a^T(\mu_1 - \mu_2)]^2}{a^T \Sigma a} = \frac{(a^T \delta)^2}{a^T \Sigma a} \qquad (54)$$

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（54）式达到最大。

定理 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，（54）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 2　利用上面的记号，取$a^T=(\mu_1-\mu_2)^T\Sigma^{-1}$，则有

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

由定理 2 我们得到如下的 Fisher 判别规则：

$$\begin{cases}x\in X_1, & \text{当}x\text{使得}(\mu_1-\mu_2)^T\Sigma^{-1}x\geq K\\ x\in X_2, & \text{当}x\text{使得}(\mu_1-\mu_2)^T\Sigma^{-1}x<K\end{cases}$$

定义判别函数

$$W(x)=(\mu_1-\mu_2)^T\Sigma^{-1}x-K=(x-\frac{1}{2}(\mu_1+\mu_2))^T\Sigma^{-1}(\mu_1-\mu_2)\qquad(55)$$

则判别规则可改写成

$$\begin{cases}x\in X_1, & \text{当}x\text{使得}W(x)\geq 0\\ x\in X_2, & \text{当}x\text{使得}W(x)<0\end{cases}$$

当总体的参数未知时，我们用样本对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。

7.3　Bayes 判别

Bayes 判别和 Bayes 估计的思想方法是一样的，即假定对研究的对象已经有一定的认识，这种认识常用先验概率来描述，当我们取得一个样本后，就可以用样本来修正已有的先验概率分布，得出后验概率分布，再通过后验概率分布进行各种统计推断。

1．误判概率与误判损失

设有两个总体X_1和X_2，根据某一个判别规则，将实际上为X_1的个体判为X_2或

者将实际上为 X_2 的个体判为 X_1 的概率就是误判概率，一个好的判别规则应该使误判概率最小。除此之外还有一个误判损失问题或者说误判产生的花费（cost）问题，如把 X_1 的个体误判到 X_2 的损失比 X_2 的个体误判到 X_1 严重得多，则人们在作前一种判断时就要特别谨慎。譬如在药品检验中把有毒的样品判为无毒后果比无毒样品判为有毒严重得多，因此一个好的判别规则还必须使误判损失最小。

为了说明问题，我们仍以两个总体的情况来讨论。设所考虑的两个总体：X_1 与 X_2 分别具有密度函数 $f_1(x)$ 与 $f_2(x)$，其中 x 为 p 维向量。记 Ω 为 x 的所有可能观测值的全体，称它为样本空间，R_1 为根据我们的规则要判为 X_1 的那些 x 的全体，而 $R_2 = \Omega - R_1$ 是要判为 X_2 的那些 x 的全体。显然 R_1 与 R_2 互斥完备。某样本实际是来自 X_1，但被判为 X_2 的概率为

$$P(2|1) = P(x \in R_2 \mid X_1) = \int_{R_2} \cdots \int f_1(x)dx$$

来自 X_2，但被判为 X_1 的概率为

$$P(1|2) = P(x \in R_1 \mid X_2) = \int_{R_1} \cdots \int f_2(x)dx$$

类似地，来自 X_1 被判为 X_1 的概率，来自 X_2 被判为 X_2 的概率分别为

$$P(1|1) = P(x \in R_1 \mid X_1) = \int_{R_1} \cdots \int f_1(x)dx$$

$$P(2|2) = P(x \in R_2 \mid X_2) = \int_{R_2} \cdots \int f_2(x)dx$$

又设 p_1, p_2 分别表示总体 X_1 和 X_2 的先验概率，且 $p_1 + p_2 = 1$，于是

$$P(\text{正确地判为}X_1) = P(\text{来自}X_1, \text{被判为}X_1) = P(x \in R_1 \mid X_1) \cdot P(X_1) = P(1|1) \cdot p_1$$

$$P(\text{误判到}X_1) = P(\text{来自}X_2, \text{被判为}X_1) = P(x \in R_1 \mid X_2) \cdot P(X_2) = P(1|2) \cdot p_2$$

类似地有

$$P(\text{正确地判为}X_2) = P(2|2) \cdot p_2$$

$$P(\text{误判到}X_2) = P(2|1)\cdot p_1$$

设 $L(1|2)$ 表示来自 X_2 误判为 X_1 引起的损失，$L(2|1)$ 表示来自 X_1 误判为 X_2 引起的损失，并规定 $L(1|1) = L(2|2) = 0$。

将上述的误判概率与误判损失结合起来，定义平均误判损失（expected cost of misclassification，简记为 ECM）如下：

$$\text{ECM}(R_1, R_2) = L(2|1)P(2|1)p_1 + L(1|2)P(1|2)p_2\,, \tag{56}$$

一个合理的判别规则应使 ECM 达到极小。

2．两总体的 Bayes 判别

由上面叙述知道，我们要选择样本空间 Ω 的一个划分：R_1 和 $R_2 = \Omega - R_1$ 使得平均损失（56）式达到极小。

定理 3　极小化平均损失（56）的区域 R_1 和 R_2 为

$$R_1 = \left\{x : \frac{f_1(x)}{f_2(x)} \geq \frac{L(1|2)}{L(2|1)}\cdot\frac{p_2}{p_1}\right\}$$

$$R_2 = \left\{x : \frac{f_1(x)}{f_2(x)} < \frac{L(1|2)}{L(2|1)}\cdot\frac{p_2}{p_1}\right\}$$

（当 $\dfrac{f_1(x)}{f_2(x)} < \dfrac{L(1|2)}{L(2|1)}\cdot\dfrac{p_2}{p_1}$ 时，即 x 为边界点，它可归入 R_1，R_2 的任何一个，为了方便就将它归入 R_1）。

由上述定理，我们得到两总体的 Bayes 判别准则：

$$\begin{cases} x \in X_1, & \text{当}x\text{使得}\dfrac{f_1(x)}{f_2(x)} \geq \dfrac{L(1|2)}{L(2|1)}\cdot\dfrac{p_2}{p_1} \\ x \in X_2, & \text{当}x\text{使得}\dfrac{f_1(x)}{f_2(x)} < \dfrac{L(1|2)}{L(2|1)}\cdot\dfrac{p_2}{p_1} \end{cases} \tag{57}$$

应用此准则时仅仅需要计算：

1）新样本点 $x_0 = (x_{01}, x_{02}, \cdots, x_{0p})^T$ 的密度函数比 $f_1(x_0) / f_2(x_0)$；

2）损失比 $L(1|2) / L(2|1)$；

3）先验概率比 p_2 / p_1。

损失和先验概率以比值的形式出现是很重要的，因为确定两种损失的比值（或两总体的先验概率的比值）往往比确定损失本身（或先验概率本身）来得容易。下面列举（57）的三种特殊情况：

1）当 $p_2 / p_1 = 1$

$$\begin{cases} x \in X_1, & \text{当}x\text{使得} \dfrac{f_1(x)}{f_2(x)} \geq \dfrac{L(1|2)}{L(2|1)} \\ x \in X_2, & \text{当}x\text{使得} \dfrac{f_1(x)}{f_2(x)} < \dfrac{L(1|2)}{L(2|1)} \end{cases} \tag{58}$$

2）当 $L(1|2)/L(2|1)=1$ 时

$$\begin{cases} x \in X_1, & \text{当}x\text{使得} \dfrac{f_1(x)}{f_2(x)} \geq \dfrac{p_2}{p_1} \\ x \in X_2, & \text{当}x\text{使得} \dfrac{f_1(x)}{f_2(x)} < \dfrac{p_2}{p_1} \end{cases} \tag{59}$$

3） $p_1 / p_2 = L(1|2)/L(2|1)=1$ 时

$$\begin{cases} x \in X_1, & \text{当}x\text{使得} \dfrac{f_1(x)}{f_2(x)} \geq 1 \\ x \in X_2, & \text{当}x\text{使得} \dfrac{f_1(x)}{f_2(x)} < 1 \end{cases} \tag{60}$$

对于具体问题，如果先验概率或者其比值都难以确定，此时就利用规则（58），同样如误判损失或者其比值都是难以确定，此时就利用规则（59），如果上述两者都难以确定则利用规则（60），最后这种情况是一种无可奈何的办法，当然判别也变得很简单：若 $f_1(x) \geq f_2(x)$，则判 $x \in X_1$，否则判 $x \in X_2$。

我们将上述的两总体 Bayes 判别应用于正态总体 $X_i \sim N_p(\mu_i, \Sigma_i)$（$i=1,2$），分两种情况讨论。

1）$\Sigma_1 = \Sigma_2 = \Sigma$，（$\Sigma > 0$），此时 X_i 的密度为

$$f_i(x) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\{-\frac{1}{2}(x-\mu_i)^T \Sigma^{-1}(x-\mu_i)\} \tag{61}$$

定理 4　设总体 $X_i \sim N_p(\mu_i, \Sigma_i)$（$i=1,2$），其中 $\Sigma > 0$，则使平均误判损失极小的划分为

$$\begin{cases} R_1 = \{x : W(x) \geq \beta\} \\ R_2 = \{x : W(x) < \beta\} \end{cases} \tag{62}$$

其中

$$W(x) = [x - \frac{1}{2}(\mu_1 + \mu_2)]^T \Sigma^{-1}(\mu_1 - \mu_2) \tag{63}$$

$$\beta = \ln \frac{L(1|2) \cdot p_2}{L(2|1) \cdot p_1} \tag{64}$$

不难发现（63）式的 $W(x)$ 与 Fisher 判别和马氏距离判别的线性判别函数（55），（51）是一致的。判别规则也只是判别限不一样。

如果总体的 μ_1, μ_2 和 Σ 未知，用式（52）和（53），算出总体样本的 $\hat{\mu}_1, \hat{\mu}_2$ 和 $\hat{\Sigma}$，来代替 μ_1, μ_2 和 Σ，得到的判别函数

$$W(x) = [x - \frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2)]^T \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_2) \tag{65}$$

称为 Anderson 线性判别函数，判别的规则为

$$\begin{cases} x \in X_1, & \text{当} x \text{使得} W(x) \geq \beta \\ x \in X_2, & \text{当} x \text{使得} W(x) < \beta \end{cases} \tag{66}$$

其中 β 由（64）所决定。

这里应该指出，总体参数用其估计来代替，所得到的规则，仅仅只是最优（在平均误判损失达到极小的意义下）规则的一个估计，这时对于一个具体问题来讲，我们并没有把握说所得到的规则能够使平均误判损失达到最小，但当样本的容量充分大时，估计 $\hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma}$ 分别和 μ_1, μ_2, Σ 很接近，因此我们有理由认为“样本”判别规则的性质会很好。

2）$\Sigma_1 \neq \Sigma_2$（$\Sigma_1 > 0$，$\Sigma_2 > 0$）

由于误判损失极小化的划分依赖于密度函数之比 $f_1(x) / f_2(x)$ 或等价于它的对数 $\ln(f_1(x) / f_2(x))$，把协方差矩阵不等的两个多元正态密度代入这个比后，包含 $|\Sigma_i|^{1/2}$

（$i=1,2$）的因子不能消去，而且 $f_i(x)$ 的指数部分也不能组合成简单表达式，因此，

对于 $\Sigma_1 \neq \Sigma_2$ 时，由定理 3 可得判别区域：

$$\begin{cases} R_1 = \{x : W(x) \geq K\} \\ R_2 = \{x : W(x) < K\} \end{cases} \tag{67}$$

其中

$$W(x) = -\frac{1}{2} x^T (\Sigma_1^{-1} - \Sigma_2^{-1}) x + (\mu_1^T \Sigma_1^{-1} - \mu_2^T \Sigma_2^{-1}) x \tag{68}$$

$$K = \ln\left(\frac{L(1|2)p_2}{L(2|1)p_1}\right) + \frac{1}{2}\ln\frac{|\Sigma_1|}{|\Sigma_2|} + \frac{1}{2}(\mu_1^T \Sigma_1^{-1} \mu_1 - \mu_2^T \Sigma_2^{-1} \mu_2) \tag{69}$$

显然，判别函数 $W(x)$ 是关于 x 的二次函数，它比 $\Sigma_1 = \Sigma_2$ 时的情况复杂得多。如果 $\mu_i, \Sigma_i (i=1,2)$ 未知，仍可采用其估计来代替。

例 14　表 17 是某气象站预报有无春旱的实际资料，x_1 与 x_2 都是综合预报因子（气象含义从略），有春旱的是 6 个年份的资料，无春旱的是 8 个年份的资料，它们的先验概率分别用 $6/14$ 和 $8/14$ 来估计，并设误判损失相等，试建立 Anderson 线性判别函数。

表 17　某气象站有无春旱的资料

序　号		1	2	3	4	5	6	7	8
春旱	x_1	24.8	24.1	26.6	23.5	25.5	27.4		
	x_2	-2.0	-2.4	-3.0	-1.9	-2.1	-3.1		
	$W(x_1, x_2)$	3.0156	2.8796	10.0929	-0.0322	4.8098	12.0960		
无春旱	x_1	22.1	21.6	22.0	22.8	22.7	21.5	22.1	21.4
	x_2	-0.7	-1.4	-0.8	-1.6	-1.5	-1.0	-1.2	-1.3
	$W(x_1, x_2)$	-6.9371	-5.6602	-6.8144	-2.4897	-3.0303	-7.1958	-5.2789	-6.4097

由表 17 的数据计算得

$$\hat{\mu}_1 = (25.3167,\ -2.4167)^T,\quad \hat{\mu}_2 = (22.0250,\ -1.1875)^T$$

$$S_1=\begin{pmatrix}11.0683 & -3.2883\\ -3.2883 & 1.3483\end{pmatrix},\quad S_2=\begin{pmatrix}1.9150 & -0.4425\\ -0.4425 & 0.7488\end{pmatrix}$$

$$\hat{\Sigma}=\begin{pmatrix}1.0819 & -0.3109\\ -0.3109 & 0.1748\end{pmatrix},\quad \beta=\ln\frac{p_2}{p_1}=0.288$$

将上述计算结果代入 Anderson 线性判别函数得

$$W(x)=W(x_1,x_2)=2.0893x_1-3.3165x_2-55.4331$$

判别限为 0.288，将表 17 的数据代入$W(x)$，计算的结果填在表 17 中$W(x_1,x_2)$相应的栏目中，错判的只有一个，即春旱中的第 4 号，与历史资料的拟合率达 93%。

计算的 MATLAB 程序如下：

```
clc,clear
a=[24.8      24.1      26.6      23.5      25.5      27.4
-2.0      -2.4      -3.0      -1.9      -2.1      -3.1]';
b=[22.1      21.6      22.0      22.8      22.7      21.5      22.1      21.4
-0.7      -1.4      -0.8      -1.6      -1.5      -1.0      -1.2      -1.3]';
n1=6;n2=8;
mu1=mean(a);mu2=mean(b);
mu1=mu1',mu2=mu2'
s1=(n1-1)*cov(a),s2=(n2-1)*cov(b)
sigma2=(s1+s2)/(n1+n2-2)
beta=log(8/6)
syms x1 x2
x=[x1;x2];
wx=(x-0.5*(mu1+mu2))'*inv(sigma2)*(mu1-mu2);
digits(6),wx=vpa(wx)
ahat=subs(wx,{x1,x2},{a(:,1),a(:,2)})
bhat=subs(wx,{x1,x2},{b(:,1),b(:,2)})
```

下面我们编写$\Sigma_1\neq\Sigma_2$情形下的 MATLAB 程序：

```
clc,clear
p1=6/14;p2=8/14;
a=[24.8      24.1      26.6      23.5      25.5      27.4
-2.0      -2.4      -3.0      -1.9      -2.1      -3.1]';
b=[22.1      21.6      22.0      22.8      22.7      21.5      22.1      21.4
-0.7      -1.4      -0.8      -1.6      -1.5      -1.0      -1.2      -1.3]';
```

```
n1=6;n2=8;
mu1=mean(a);mu2=mean(b);
mu1=mu1',mu2=mu2'
cov1=cov(a),cov2=cov(b)
k=log(p2/p1)+0.5*log(det(cov1)/det(cov2))+0.5*(mu1'*inv(cov1)*mu1-mu2'*inv(
cov2)*mu2)
syms x1 x2
x=[x1;x2];
wx=-0.5*x.'*(inv(cov1)-inv(cov2))*x+(mu1'*inv(cov1)-mu2'*inv(cov2))*x;
digits(6),wx=vpa(wx);
wx=simple(wx)
ahat=subs(wx,{x1,x2},{a(:,1),a(:,2)})
bhat=subs(wx,{x1,x2},{b(:,1),b(:,2)})
ahat>=k,bhat<k
```

分类正确率为 100%。

7.4　应用举例

例 15　某种产品的生产厂家有 12 家，其中 7 家的产品受消费者欢迎，属于畅销品，定义为 1 类；5 家的产品不大受消费者欢迎，属于滞销品，定义为 2 类。将 12 家的产品的式样，包装和耐久性进行了评估后，得分资料见表 18。

表 18　生产厂家的数据

厂家	1	2	3	4	5	6	7	8	9	10	11	12	13
式样	9	7	8	8	9	8	7	4	3	6	2	1	6
包装	8	6	7	5	9	9	5	4	6	3	4	2	4
耐久性	7	6	8	5	3	7	6	4	6	3	5	2	5
类别	1	1	1	1	1	1	1	2	2	2	2	2	待判

今有一新得厂家，得分为（6，4，5），该厂的产品是否受欢迎。

利用如下的 MATLAB 程序：

```
train=[9    7    8    8    9    8    7    4    3    6    2    1
8    6    7    5    9    9    5    4    6    3    4    2
7    6    8    5    3    7    6    4    6    3    5    2]';
sample=[6 4 5];
group=[ones(7,1);2*ones(5,1)];
[x1,y1]=classify(sample,train,group,'mahalanobis')
[x2,y2]=classify(sample,train,group,'linear')
```

求得利用马氏距离和线性分类方法都把新厂家分在第一类。

§8 典型相关分析（Canonical correlation analysis）

8.1 典型相关分析的基本思想

通常情况下，为了研究两组变量

$$(x_1, x_2, \cdots, x_p)，(y_1, y_2, \cdots, y_q)$$

的相关关系，可以用最原始的方法，分别计算两组变量之间的全部相关系数，一共有 pq 个简单相关系数，这样又繁琐又不能抓住问题的本质。如果能够采用类似于主成分的思想，分别找出两组变量的各自的某个线性组合，讨论线性组合之间的相关关系，则更简捷。

首先分别在每组变量中找出第一对线性组合，使其具有最大相关性，

$$\begin{cases} u_1 = a_{11}x_1 + a_{21}x_2 + \cdots + a_{p1}x_p \\ v_1 = b_{11}y_1 + b_{21}y_2 + \cdots + b_{q1}y_q \end{cases}$$

然后再在每组变量中找出第二对线性组合，使其分别与本组内的第一线性组合不相关，第二对本身具有次大的相关性。

$$\begin{cases} u_2 = a_{12}x_1 + a_{22}x_2 + \cdots + a_{p2}x_p \\ v_2 = b_{12}y_1 + b_{22}y_2 + \cdots + b_{q2}y_q \end{cases}$$

u_2 与 u_1、v_2 与 v_1 不相关，但 u_2 和 v_2 相关。如此继续下去，直至进行到 r 步，两组变量的相关性被提取完为止，可以得到 r 组变量，这里 $r \le \min(p, q)$。

8.2 典型相关的数学描述

研究两组随机变量之间的相关关系，可用复相关系数（也称全相关系数）。1936 年 Hotelling 将简单相关系数推广到多个随机变量与多个随机变量之间的相关关系的讨论中，提出了典型相关分析。

实际问题中，需要考虑两组变量之间的相关关系的问题很多，例如，考虑几种主要产品的价格（作为第一组变量）和相应这些产品的销售量（作为第二组变量）之间的相关关系；考虑投资性变量（如劳动者人数、货物周转量、生产建设投资等）与国民收入变量（如工农业国民收入、运输业国民收入、建筑业国民收入等）之间的相关关系等等。

复相关系数描述两组随机变量 $X = (x_1, x_2, \cdots, x_p)$ 与 $Y = (y_1, y_2, \cdots, y_p)$ 之间的相关程度。其思想是先将每一组随机变量作线性组合，成为两个随机变量：

$$u = a^T X = \sum_{i=1}^{p} a_i x_i \,,\ v = b^T Y = \sum_{i=1}^{q} b_i y_i \qquad (70)$$

再研究 u 与 v 的相关系数。由于 u, v 与投影向量 a, b 有关，所以 r_{uv} 与 a, b 有关，

$r_{uv}=r_{uv}(a,b)$。我们取在$a^T\Sigma_{XX}a=1$和$b^T\Sigma_{YY}b=1$的条件下使r_{uv}达到最大的a,b作为投影向量，这样得到的相关系数为复相关系数：

$$r_{uv}=\max_{\substack{a^T\Sigma_{XX}a=1\\ b^T\Sigma_{YY}b=1}} r_{uv}(a,b) \tag{71}$$

将两组变量的协方差矩阵分块得：

$$\mathrm{Cov}\begin{pmatrix}X\\Y\end{pmatrix}=\begin{pmatrix}\mathrm{Var}(X) & \mathrm{Cov}(X,Y)\\ \mathrm{Cov}(Y,X) & \mathrm{Var}(Y)\end{pmatrix}=\begin{pmatrix}\Sigma_{XX} & \Sigma_{XY}\\ \Sigma_{YX} & \Sigma_{YY}\end{pmatrix} \tag{72}$$

此时

$$r_{uv}=\frac{\mathrm{Cov}(a^TX,b^TY)}{\sqrt{D(a^TX)}\sqrt{D(b^TY)}}=\frac{a^T\Sigma_{XY}b}{\sqrt{a^T\Sigma_{XX}a}\sqrt{b^T\Sigma_{YY}b}}=a^T\Sigma_{XY}b \tag{73}$$

因此问题转化为在$a^T\Sigma_{XX}a=1$和$b^T\Sigma_{YY}b=1$的条件下求$a^T\Sigma_{XY}b$的极大值。

根据条件极值的求法引入lagrange乘数，可将问题转化为求

$$S(a,b)=a^T\Sigma_{XY}b-\frac{\lambda}{2}(a^T\Sigma_{XX}a-1)-\frac{\gamma}{2}(b^T\Sigma_{YY}b-1) \tag{74}$$

的极大值，其中λ,γ是Lagrange乘数。

由极值的必要条件得方程组：

$$\begin{cases}\dfrac{\partial S}{\partial a}=\Sigma_{XY}b-\lambda\Sigma_{XX}a=0\\[2ex] \dfrac{\partial S}{\partial b}=\Sigma_{YX}a-\gamma\Sigma_{YY}b=0\end{cases} \tag{75}$$

将上二式分别左乘a^T与b^T，则得

$$\begin{cases}a^T\Sigma_{XY}b=\lambda a^T\Sigma_{XX}a=\lambda\\ b^T\Sigma_{YX}a=\gamma b^T\Sigma_{YY}b=\gamma\end{cases} \tag{76}$$

注意$\Sigma_{XY}=\Sigma_{YX}^T$，所以

$$\lambda=\gamma=a^T\Sigma_{XY}b \tag{77}$$

代入方程组（75）得：

$$\begin{cases}\Sigma_{XY}b-\lambda\Sigma_{XX}a=0\\ \Sigma_{YX}a-\lambda\Sigma_{YY}b=0\end{cases} \tag{78}$$

以 Σ_{YY}^{-1} 左乘（78）第二式得 $\lambda b=\Sigma_{YY}^{-1}\Sigma_{YX}a$，所以

$$b=\frac{1}{\lambda}\Sigma_{YY}^{-1}\Sigma_{YX}a$$

代入（78）第一式得：

$$(\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}-\lambda^2\Sigma_{XX})a=0 \tag{79}$$

同理可得

$$(\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}-\lambda^2\Sigma_{YY})b=0 \tag{80}$$

记

$$M_1=\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\text{，}\ M_2=\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY} \tag{81}$$

则得

$$M_1a=\lambda^2a\text{，}\ M_2b=\lambda^2b \tag{82}$$

说明 λ^2 既是 M_1 又是 M_2 的特征根，a,b 就是其相应于 M_1 和 M_2 的特征向量。M_1 和 M_2 的特征根非负，均在0和1之间，相等的非零特征根数目等于 $\min(p,q)$，不妨设为 q。

设 $M_1a=\lambda^2a$ 的特征根排序为 $\lambda_1^2\geq\lambda_2^2\geq\cdots\geq\lambda_q^2$，其余 $p-q$ 个特征根为0，我们称 $\lambda_1,\lambda_2,\cdots,\lambda_q$ 为典型相关系数。相应从 $M_1a=\lambda^2a$ 解出的特征向量为 $a_1,a_2,\cdots,a_p$，从 $M_2b=\lambda^2b$ 解出的特征向量为 $b_1,b_2,\cdots,b_q$，从而可得 q 对线性组合：

$$u_i=a_i^TX\text{，}\ v_i=b_i^TY\text{，}\ i=1,2,\cdots,q \tag{83}$$

称每一对变量为典型变量。求典型相关系数和典型变量归结为求 M_1 和 M_2 的特征根和特征向量。

还可以证明，当 $i\neq j$ 时，

$$\mathrm{Cov}(u_i,u_j)=\mathrm{Cov}(a_i^TX,a_j^TX)=a_i^T\Sigma_{XX}a_j=0 \tag{84}$$

$$\mathrm{Cov}(v_i,v_j)=\mathrm{Cov}(b_i^TY,b_j^TY)=b_i^T\Sigma_{YY}b_j=0 \tag{85}$$

表示一切典型变量都是不相关的，并且其方差为1，

$$\mathrm{Cov}(u_i,u_j)=E(u_iu_j)=\delta_{ij} \tag{86}$$

$$\mathrm{Cov}(v_i,v_j)=E(v_iv_j)=\delta_{ij} \tag{87}$$

其中

$$\delta_{ij}=\begin{cases}1, & i=j\\ 0, & i\neq j\end{cases} \tag{88}$$

X 与 Y 的同一对典型变量 u_i 和 v_i 之间的相关系数为 λ_i，不同对的典型变量 u_i 和 v_j（$i\neq j$）之间不相关，也就是说协方差为0，即

$$Cov(u_i,v_j)=E(u_iv_j)=\begin{cases}\lambda_i, & i=j\\ 0, & i\neq j\end{cases} \tag{89}$$

当总体的均值向量 μ 和协差阵 Σ 未知时，无法求总体的典型相关系数和典型变量，因而需要给出样本的典型相关系数和典型变量。

设 $X_{(1)},\cdots,X_{(n)}$ 和 $Y_{(1)},\cdots,Y_{(n)}$ 为来自总体容量为 n 的样本，这时有协方差阵的无偏估计：

$$\hat{\Sigma}_{XX}=\frac{1}{n-1}\sum_{i=1}^{n}(X_{(i)}-\overline{X})(X_{(i)}-\overline{X})^T \tag{90}$$

$$\hat{\Sigma}_{YY}=\frac{1}{n-1}\sum_{i=1}^{n}(Y_{(i)}-\overline{Y})(Y_{(i)}-\overline{Y})^T \tag{91}$$

$$\hat{\Sigma}_{XY}=\hat{\Sigma}_{YX}^T=\frac{1}{n-1}\sum_{i=1}^{n}(X_{(i)}-\overline{X})(Y_{(j)}-\overline{Y})^T \tag{92}$$

其中 $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_{(i)}$，$\overline{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_{(i)}$，用 $\hat{\Sigma}$ 代替 Σ 并按（81）和（82）求出 $\hat{\lambda}_i$ 和 $\hat{a},\hat{b}$，

称 $\hat{\lambda}_i$ 为样本典型相关系数，称 $\hat{u}_i=\hat{a}_i^TX$，$\hat{v}_i=\hat{b}_i^TY$，（$i=1,\cdots,q$）为样本的典型变量。

计算时也可从样本的相关系数矩阵出发求样本的典型相关系数和典型变量，将相关系数矩阵 R 取代协方差阵，计算过程是一样的。

如果复相关系数中的一个变量是一维的，那么也可以称为偏相关系数。偏相关系数

是描述一个随机变量 y 与多个随机变量（一组随机变量） $X=(x_1,x_2,\cdots,x_p)^T$ 之间的关系。其思想是先将那一组随机变量作线性组合，成为一个随机变量：

$$u=c^TX=\sum_{i=1}^{p}c_ix_i \tag{93}$$

再研究 y 与 u 的相关系数。由于 u 与投影向量 c 有关，所以 r_{yu} 与 c 有关，$r_{yu}=r_{yu}(c)$。我们取在 $c^T\Sigma_{XX}c=1$ 的条件下使 r_{yu} 达到最大的 c 作为投影向量得到的相关系数为偏相关系数：

$$r_{yu}=\max_{c^T\Sigma_{XX}c=1}r_{yu}(c) \tag{94}$$

其余推导与计算过程与复相关系数的类似。

8.3　原始变量与典型变量之间的相关性

（1）原始变量与典型变量之间的相关系数

设原始变量相关系数矩阵

$$R=\begin{bmatrix}R_{11} & R_{12}\\ R_{21} & R_{22}\end{bmatrix}$$

X 典型变量系数矩阵

$$A=[a_1\quad a_2\quad\cdots\quad a_r]_{p\times r}=\begin{bmatrix}a_{11} & a_{12} & \cdots & a_{1r}\\ a_{12} & a_{22} & \cdots & a_{2r}\\ \vdots & \vdots & & \vdots\\ a_{p1} & a_{p2} & \cdots & a_{pr}\end{bmatrix}$$

Y 典型变量系数矩阵

$$B=[b_1\quad b_2\quad\cdots\quad b_r]_{p\times r}=\begin{bmatrix}b_{11} & b_{12} & \cdots & b_{1r}\\ b_{12} & b_{22} & \cdots & b_{2r}\\ \vdots & \vdots & & \vdots\\ b_{q1} & b_{q2} & \cdots & b_{qr}\end{bmatrix}$$

则有

$$\operatorname{cov}(x_i,u_j)=\operatorname{cov}(x_i,\sum_{k=1}^{p}a_{kj}x_k)=\sum_{k=1}^{p}a_{kj}\operatorname{cov}(x_i,x_k)$$

x_i 与 u_j 的相关系数

$$\rho(x_i,u_j)=\sum_{k=1}^{p}a_{kj}\operatorname{cov}(x_i,x_k)/\sqrt{D(x_i)}$$

同理可计算得

$$\rho(x_i,v_j)=\sum_{k=1}^{q}b_{kj}\operatorname{cov}(x_i,y_k)/\sqrt{D(x_i)}$$

$$\rho(y_i,u_j)=\sum_{k=1}^{p}a_{kj}\operatorname{cov}(y_i,x_k)/\sqrt{D(y_i)}$$

$$\rho(y_i,v_j)=\sum_{k=1}^{q}b_{kj}\operatorname{cov}(y_i,y_k)/\sqrt{D(y_i)}$$

（2）各组原始变量被典型变量所解释的方差

X 组原始变量被 u_i 解释的方差比例

$$m_{u_i}=\sum_{k=1}^{p}\rho^2(u_i,x_k)/p,$$

X 组原始变量被 v_i 解释的方差比例

$$m_{v_i}=\sum_{k=1}^{p}\rho^2(v_i,x_k)/p$$

Y 组原始变量被 u_i 解释的方差比例

$$n_{u_i}=\sum_{k=1}^{q}\rho^2(u_i,y_k)/q$$

Y 组原始变量被 v_i 解释的方差比例

$$n_{v_i}=\sum_{k=1}^{q}\rho^2(v_i,y_k)/q$$

8.4　典型相关系数的检验

在实际应用中，总体的协方差矩阵常常是未知的，类似于其他的统计分析方法，需要从总体中抽出一个样本，根据样本对总体的协方差或相关系数矩阵进行估计，然后利用估计得到的协方差或相关系数矩阵进行分析。由于估计中抽样误差的存在，所以估计以后还需要进行有关的假设检验。

1．计算样本的协方差阵

假设有 X 组和 Y 组变量，样本容量为 n，观测值矩阵为

$$\begin{bmatrix} x_{11} & \cdots & x_{1p} & y_{11} & \cdots & y_{1q} \\ x_{21} & \cdots & x_{2p} & y_{21} & \cdots & y_{2p} \\ \vdots & & \vdots & \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} & y_{n1} & \cdots & y_{nq} \end{bmatrix}$$

对应的标准化数据矩阵为

$$Z=\begin{bmatrix} \dfrac{x_{11}-\bar{x}_1}{\sigma_x^1} & \cdots & \dfrac{x_{1p}-\bar{x}_p}{\sigma_x^p} & \dfrac{y_{11}-\bar{y}_1}{\sigma_y^1} & \cdots & \dfrac{y_{1q}-\bar{y}_q}{\sigma_y^q} \\ \dfrac{x_{21}-\bar{x}_1}{\sigma_x^1} & \cdots & \dfrac{x_{2p}-\bar{x}_p}{\sigma_x^p} & \dfrac{y_{21}-\bar{y}_1}{\sigma_y^1} & \cdots & \dfrac{y_{2q}-\bar{y}_q}{\sigma_y^q} \\ \vdots & & \vdots & \vdots & & \vdots \\ \dfrac{x_{n1}-\bar{x}_1}{\sigma_x^1} & \cdots & \dfrac{x_{np}-\bar{x}_p}{\sigma_x^p} & \dfrac{y_{n1}-\bar{y}_1}{\sigma_y^1} & \cdots & \dfrac{y_{2q}-\bar{y}_q}{\sigma_y^q} \end{bmatrix}$$

样本的协方差

$$\hat{\Sigma}=\frac{1}{n-1}Z^TZ=\frac{1}{n-1}\begin{bmatrix} S_{XX} & S_{XY} \\ S_{YX} & S_{YY} \end{bmatrix}=\frac{1}{n-1}S$$

2．整体检验（$H_0:\Sigma_{XY}=0$；$H_1:\Sigma_{XY}\neq 0$）

$$H_0:\rho_1=\rho_2=\cdots\rho_r=0$$

$H_1:\rho_i(i=1,2,\cdots,r)$ 中至少 ρ_1 不为零

检验的统计量为

$$\Lambda_1=\frac{|S|}{|S_{XX}||S_{YY}|}$$

经计算得

$$\Lambda_1=\left|I-S_{XX}^{-1}S_{XY}S_{YY}^{-1}S_{YX}\right|=\prod_{i=1}^{p}(1-\lambda_i)$$

若 Λ_1 小，则支持 H_1。

在原假设为真的情况下，检验的统计量

$$Q_1 = -\left[n - \frac{1}{2}(p+q+3)\right]\ln \Lambda_1$$

近似服从自由度为 pq 的 χ^2 分布。在给定的显著水平 α 下，如果 $Q_1 \geq \chi_\alpha^2(pq)$，则拒绝原假设，认为至少第一对典型变量之间的相关性显著。

3．部分总体典型相关系数为零的检验

$H_0 : \rho_2 = \rho_3 = \cdots \rho_r = 0$

$H_1 : \rho_2, \rho_3, \cdots, \rho_r$ 至少有一非零

若原假设 H_0 被接受，则认为只有第一对典型变量是有用的；若原假设 H_0 被拒绝，则认为第二对典型变量也是有用的，并进一步检验假设。

$H_0 : \rho_3 = \rho_4 = \cdots \rho_r = 0$

$H_1 : \rho_3, \rho_4, \cdots, \rho_r$ 至少有一非零

如此进行下去，直至对某个 k

$H_0 : \rho_{k+1} = \rho_{k+2} = \cdots \rho_r = 0$

$H_1 : \rho_{k+1}, \rho_{k+2}, \cdots, \rho_r$ 至少有一非零

检验的统计量

$$\Lambda_{k+1} = \prod_{i=k+1}^{r}(1-\lambda_i)，\quad Q = -[n-k-\frac{1}{2}(p+q+3) + \sum_{i=2}^{k+1}\lambda_i^{-1}]\ln \Lambda_{k+1}$$

近似服从自由度为 $(p-k)(q-k)$ 的 χ^2 分布。在给定的显著水平 α 下，如果 $Q \geq \chi_\alpha^2((p-k)(q-k))$，则拒绝原假设，认为至少第 $k+1$ 对典型变量之间的相关性显著。

8.5　典型相关分析案例

8.5.1　职业满意度典型相关分析

某调查公司从一个大型零售公司随机调查了 784 人，测量了 5 个职业特性指标和 7 个职业满意变量，有关的变量见表 19。讨论两组指标之间是否相联系。

表 19　指标变量表

X 组	x_1－用户反馈， x_2－任务重要性， x_3－任务多样性， x_4－任务特殊性 x_5－自主性
Y 组	y_1－主管满意度， y_2－事业前景满意度， y_3－财政满意度， y_4－工作强度满意度 y_5－公司地位满意度， y_6－工作满意度， y_7－总体满意度

相关系数矩阵数据见表 20。

表 20　相关系数矩阵数据

	x_1	x_2	x_3	x_4	x_5	y_1	y_2	y_3	y_4	y_5	y_6	y_7
x_1	1.00	0.49	0.53	0.49	0.51	0.33	0.32	0.20	0.19	0.30	0.37	0.21
x_2	0.49	1.00	0.57	0.46	0.53	0.30	0.21	0.16	0.08	0.27	0.35	0.20
x_3	0.53	0.57	1.00	0.48	0.57	0.31	0.23	0.14	0.07	0.24	0.37	0.18
x_4	0.49	0.46	0.48	1.00	0.57	0.24	0.22	0.12	0.19	0.21	0.29	0.16
x_5	0.51	0.53	0.57	0.57	1.00	0.38	0.32	0.17	0.23	0.32	0.36	0.27
y_1	0.33	0.30	0.31	0.24	0.38	1.00	0.43	0.27	0.24	0.34	0.37	0.40
y_2	0.32	0.21	0.23	0.22	0.32	0.43	1.00	0.33	0.26	0.54	0.32	0.58
y_3	0.20	0.16	0.14	0.12	0.17	0.27	0.33	1.00	0.25	0.46	0.29	0.45
y_4	0.19	0.08	0.07	0.19	0.23	0.24	0.26	0.25	1.00	0.28	0.30	0.27
y_5	0.30	0.27	0.24	0.21	0.32	0.34	0.54	0.46	0.28	1.00	0.35	0.59
y_6	0.37	0.35	0.37	0.29	0.36	0.37	0.32	0.29	0.30	0.35	1.00	0.31
y_7	0.21	0.20	0.18	0.16	0.27	0.40	0.58	0.45	0.27	0.59	0.31	1.00

一些计算结果的数据见下面的表格。

表 21　X 组的典型变量

	u_1	u_2	u_3	u_4	u_5
x_1	0.421704	-0.34285	0.857665	-0.78841	0.030843
x_2	0.195106	0.668299	-0.44343	-0.26913	0.983229

x_3	0.167613	0.853156	0.259213	0.468757	-0.91414
x_4	-0.02289	-0.35607	0.423106	1.042324	0.524367
x_5	0.459656	-0.72872	-0.97991	-0.16817	-0.43924

表 22　原始变量与本组典型变量之间的相关系数

	u_1	u_2	u_3	u_4	u_5
x_1	0.829349	-0.10934	0.48534	-0.24687	0.061056
x_2	0.730368	0.436584	-0.20014	0.002084	0.485692
x_3	0.753343	0.466088	0.105568	0.301958	-0.33603
x_4	0.615952	-0.22251	0.205263	0.661353	0.302609
x_5	0.860623	-0.26604	-0.38859	0.148424	-0.12457
	v_1	v_2	v_3	v_4	v_5
y_1	0.756411	0.044607	0.339474	0.129367	-0.33702
y_2	0.643884	0.358163	-0.17172	0.352983	-0.33353
y_3	0.387242	0.037277	-0.17673	0.53477	0.414847
y_4	0.377162	0.791935	-0.00536	-0.28865	0.334077
y_5	0.653234	0.108391	0.209182	0.437648	0.434613
y_6	0.803986	-0.2416	-0.23477	-0.40522	0.196419
y_7	0.502422	0.162848	0.4933	0.188958	0.067761

表 23　原始变量与对应组典型变量之间的相关系数

	v_1	v_2	v_3	v_4	v_5
x_1	0.459216	0.025848	-0.05785	0.017831	0.003497
x_2	0.404409	-0.10321	0.023854	-0.00015	0.027816
x_3	0.417131	-0.11019	-0.01258	-0.02181	-0.01924
x_4	0.341056	0.052602	-0.02446	-0.04777	0.01733
x_5	0.476532	0.062893	0.046315	-0.01072	-0.00713

	u_1	u_2	u_3	u_4	u_5
y_1	0.41883	-0.01055	-0.04046	-0.00934	-0.0193
y_2	0.356523	-0.08467	0.020466	-0.0255	-0.0191
y_3	0.214418	-0.00881	0.021064	-0.03863	0.023758
y_4	0.208837	-0.18722	0.000639	0.020849	0.019133
y_5	0.3617	-0.02562	-0.02493	-0.03161	0.02489
y_6	0.445172	0.057116	0.027981	0.029268	0.011249
y_7	0.278194	-0.0385	-0.05879	-0.01365	0.003881

表24　典型相关系数

1	2	3	4	5
0.553706	0.236404	0.119186	0.072228	0.05727

可以看出，所有五个表示职业特性的变量与u_1有大致相同的相关系数，u_1视为形容职业特性的指标。第一对典型变量的第二个成员v_1与y_1, y_2, y_5, y_6有较大的相关系数，说明v_1主要代表了主管满意度，事业前景满意度，公司地位满意度和工种满意度。而u_1和v_1之间的相关系数0.5537。

u_1和v_1解释的本组原始变量的比率：

$$m_{u_1} = 0.5818，\ n_{v_1} = 0.3721$$

X组的原始变量被u_1到u_5解释了100%，Y组的原始变量被v_1到v_5解释了80.3%。

计算的MATLAB程序如下：

```
clc,clear
load da.txt    %原始的相关系数矩阵保存在纯文本文件da.txt中
%r为相关系数矩阵
r=da;
n1=5;n2=7;num=min(n1,n2);
s1=r(1:n1,1:n1);
s12=r(1:n1,n1+1:end);
```

```
s21=s12';
s2=r(n1+1:end,n1+1:end);
m1=inv(s1)*s12*inv(s2)*s21;
m2=inv(s2)*s21*inv(s1)*s12;
[x1,y1]=eig(m1);
%以下是特征向量归一化，满足a's1a=1
gu1=x1'*s1*x1;
gu1=sqrt(diag(gu1));  %求典型相关系数
gu1=gu1'.*sign(sum(x1));    %每个特征向量的最大分量为正
gu1=repmat(gu1,length(gu1),1);
a=x1./gu1;
y1=diag(y1);      %取出特征值
[y1,ind1]=sort(y1,'descend');  %特征值按照从大到小排列
a=a(:,ind1(1:num))      %取出X组的系数阵
y1=sqrt(y1(1:num))      %计算典型相关系数
flag=1;
xlswrite('bk1.xls',a,'Sheet1','A1')    %把计算结果写到Excel文件中去
flag=n1+2;
str=char(['A',int2str(flag)]);
xlswrite('bk1.xls',y1','Sheet1',str)
[x2,y2]=eig(m2);
%以下是特征向量归一化，满足b's2b=1
gu2=x2'*s2*x2;
gu2=sqrt(diag(gu2));
gu2=gu2'.*sign(sum(x2));
gu2=repmat(gu2,length(gu2),1);
b=x2./gu2;
y2=diag(y2);
[y2,ind2]=sort(y2,'descend');
b=b(:,ind2(1:num))
y2=sqrt(y2(1:num))  %计算典型相关系数
flag=flag+2;
str=char(['A',int2str(flag)]);
xlswrite('bk1.xls',b,'Sheet1',str)
flag=flag+n2+1;
str=char(['A',int2str(flag)]);
xlswrite('bk1.xls',y2','Sheet1',str)
x_u_r=s1*a;       %x,u的相关系数
x_u_r=x_u_r(:,1:num)
```

```
flag=flag+2;
str=char(['A',int2str(flag)]);
xlswrite('bk1.xls',x_u_r,'Sheet1',str)
y_v_r=s2*b;       %y,v的相关系数
y_v_r=y_v_r(:,1:num)
flag=flag+n1+1;
str=char(['A',int2str(flag)]);
xlswrite('bk1.xls',y_v_r,'Sheet1',str)
x_v_r=s12*b;      %x,v的相关系数
x_v_r=x_v_r(:,1:num)
flag=flag+n2+1;
str=char(['A',int2str(flag)]);
xlswrite('bk1.xls',x_v_r,'Sheet1',str)
y_u_r=s21*a;      %y,u的相关系数
y_u_r=y_u_r(:,1:num)
flag=flag+n1+1;
str=char(['A',int2str(flag)]);
xlswrite('bk1.xls',y_u_r,'Sheet1',str)
mu=sum(x_u_r.^2)/n1     %x组原始变量被u_i解释的方差比例
mv=sum(x_v_r.^2)/n1     %x组原始变量被v_i解释的方差比例
nu=sum(y_u_r.^2)/n2     %y组原始变量被u_i解释的方差比例
nv=sum(y_v_r.^2)/n2     %y组原始变量被v_i解释的方差比例
```

8.5.2 中国城市竞争力与基础设施的典型相关分析

1．导言

随着经济全球化和我国加入 WTO，作为区域中心的城市在区域经济发展中的作用越来越重要，城市间的竞争也愈演愈烈，许多有识之士甚至断言，21 世纪，国家之间、区域之间、国际企业之间的竞争将突出地表现为城市层面上的竞争。因此，为了应对新的经济社会环境，积极探索影响城市竞争力的因素，研究提高城市综合实力的方法，充分发挥其集聚与扩散作用，以进一步带动整个区域经济建设，已成为一项重要的战略课题，城市竞争力研究已受到学术界的高度重视。钟卫东和张伟（2002）分析了城市竞争力评价中存在的问题，应用综合指数修正法构建城市竞争力的三级评价指标体系，并提出了纵横因子评价法；徐康宁（2002）提出建立测度城市竞争力指标体系的四个原则和三级指标共确定了 69 个具体指标；沈正平、马晓冬、戴先杰和翟仁祥（2002）构建了测度城市竞争力的指标体系，并用因子分析、聚类分析等方法对新亚欧大陆桥经济带 25 个样本城市的竞争力进行了评价；倪鹏飞(2002)提出城市竞争力与基础设施竞争力假说，并运用主成分分析和模糊曲线分析法进行了分析检验；此外，郝寿义、成起宏(1999)、上海社会科学院(2001)、唐礼智(2001)和和宁越敏(2002)等都对城市竞争力问题作了可贵的探索。但通过查阅上述文献发现，现有成果在城市竞争力评价方法上尚存在一些缺陷

和不足，有许多问题需要进一步探讨。下面将典型相关分析方法引入到城市竞争力评价问题中，对城市竞争力与城市基础设施的相关性进行实证分析，并据此提出了相应的政策建议。

2．典型相关分析法的基本思想

统计分析中，我们用简单相关系数反映两个变量之间的线性相关关系。1936 年 Hotelling 将线性相关性推广到两组变量的讨论中，提出了典型相关分析方法。它的基本思想是仿照主成分分析法中把多变量与多变量之间的相关化为两个变量之间相关的做法，首先在每组变量内部找出具有最大相关性的一对线性组合，然后再在每组变量内找出第二对线性组合，使其本身具有最大的相关性，并分别与第一对线性组合不相关。如此下去，直到两组变量内各变量之间的相关性被提取完毕为止。有了这些最大相关的线性组合，则讨论两组变量之间的相关，就转化为研究这些线性组合的最大相关，从而减少了研究变量的个数。典型相关分析的过程如下：

假设有两组随机变量 $X=(x_1,\cdots,x_p)^T$，$Y=(y_1,\cdots,y_q)^T$，Z 为 $p+q$ 维总体的 n 次中心化观测数据阵：

$$Z=\begin{pmatrix} x_{11} & \cdots & x_{1p} & y_{11} & \cdots & y_{1q} \\ x_{21} & \cdots & x_{2p} & y_{21} & \cdots & y_{2q} \\ \vdots & & \vdots & \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} & y_{n1} & \cdots & y_{nq} \end{pmatrix}$$

第一步，计算相关系数阵 R，并将 R 剖分为 $R=\begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix}$，其中 R_{11}，R_{22} 分别为第一组变量和第二组变量的相关系数阵，$R_{12}=R_{21}^T$ 为第一组与第二组变量的相关系数阵。

第二步，求典型相关系数及典型变量。首先求 $M_1=R_{11}^{-1}R_{12}R_{22}^{-1}R_{21}$ 的特征根 λ_i^2，特征向量 a_i；$M_2=R_{22}^{-1}R_{21}R_{11}^{-1}R_{12}$ 的特征根 λ_i^2，特征向量 b_i，则有

则典型变量为

$$u_1=a_1^TX,v_1=b_1^TY;u_2=a_2^TX,v_2=b_2^TY;\cdots;u_t=a_t^TX,v_t=b_t^TY(t\le\min(p,q))$$

第三步，典型相关系数 λ_i 的显著性检验。

第四步，典型结构与典型冗余分析。

典型结构指原始变量与典型变量之间的相关系数阵 $R(X,V)$，据此可以计算任一个典型变量 u_k 或 v_k 解释本组变量 X(或Y) 总变差的百分比 $R_d(X;u_k)$（或 $R_d(Y;v_k)$）。同时可求得前 t 个典型变量 $u_1,\cdots,u_t$(或 $v_1,\cdots,v_t$) 解释本组变量 X(或Y) 总变差的累计百分比 $R_d(X;u_1,\cdots,u_t)$ 或 $R_d(Y;v_1,\cdots,v_t)$。

典型冗余分析用来研究典型变量解释另一组变量总变差百分比的问题。第二组典

型变量v_k解释第一组变量X总变差的百分比$R_d(X;v_k)$（或第一组中典型变量解释的变差被第二组中典型变量重复解释的百分比）简称为第一组典型变量的冗余测度；第一组典型变量u_k解释第二组变量Y总变差的百分比$R_d(Y;u_k)$（或第二组中典型变量解释的变差被第一组中典型变量重复解释的百分比）简称为第二组典型变量的冗余测度。冗余测度的大小表示这对典型变量能够对另一组变差相互解释的程度大小。

3．城市竞争力与基础设施关系的典型相关分析

（1）城市竞争力指标与基础设施指标

城市竞争力主要取决于产业经济效益、对外开放程度、基础设施、市民素质、政府管理及环境质量等因素。城市基础设施是以物质形态为特征的城市基础结构系统，是指城市可利用的各种设施及质量，包括交通、通讯、能源动力系统，住房储备，文、卫、科教机构和设施等。基础设施是城市经济、社会活动的基本载体，它的规模、类型、水平直接影响着城市产业的发展和价值体系的形成，因此，基础设施竞争力是城市竞争力的重要组成部分，对提高城市竞争力非常重要。

我们选取了从不同的角度表现城市竞争力的四个关键性指标，构建了城市竞争力指标体系：市场占有率、GDP 增长率、劳动生产率和居民人均收入。城市基础设施指标体系主要包含六个指标：对外设施指数（由城市货运量和客运量指标综合构成），对内基本设施指数（由城市能源、交通、道路、住房等具体指标综合而成），每百人拥有电话机数，技术性设施指数（是城市现代交通、通讯、信息设施的综合指数，由港口个数、机场等级、高速公路、高速铁路、地铁个数、光缆线路数等加权综合构成），文化设施指数（由公共藏书量、文化馆数量、影剧院数量等指标加权综合构成），卫生设施指数（由医院个数、万人医院床位数综合构成）。

我们选取了 20 个最具有代表性的城市，城市名称和竞争力、基础设施各项指标数据如表 25、表 26。

表 25　城市竞争力表现要素得分

城市	劳动生产率 y_1	市场占有率 y_2	居民人均收入 y_3	长期经济增长率 y_4	城市	劳动生产率 y_1	市场占有率 y_2	居民人均收入 y_3	长期经济增长率 y_4
上海	45623.05	2.5	8439	16.27	青岛	33334.62	0.63	6222	11.63
深圳	52256.67	1.3	18579	21.5	武汉	24633.27	0.59	5573	16.39
广州	46551.87	1.13	10445	11.92	温州	39258.78	-0.69	9034	22.43
北京	28146.76	1.38	7813	15	福州	38201.47	-0.34	7083	18.53
厦门	38670.43	0.12	8980	26.71	重庆	16524.32	0.44	5323	12.22
天津	26316.96	1.37	6609	11.07	成都	31855.63	-0.02	6019	11.88
大连	45330.53	0.56	6070	12.4	宁波	22528.8	-0.16	9069	15.7
杭州	45853.89	0.28	7896	13.93	石家庄	21831.94	-0.15	5497	13.56
南京	35964.64	0.74	6497	8.97	西安	19966.36	-0.15	5344	12.43
珠海	55832.61	-0.12	13149	9.22	哈尔滨	19225.71	-0.16	4233	10.16

数据来源：倪鹏飞等:《城市竞争力蓝皮书：中国城市竞争力报告 NO.1》，北京，社会科学出版社 2003 年版。

表 26　城市基础设施构成要素得分

城市	对外设施指数 x_1	对内设施指数 x_2	每百人电话数 x_3	技术设施指数 x_4	文化设施指数 x_5	卫生设施指数 x_6	城市	对外设施指数 x_1	对内设施指数 x_2	每百人电话数 x_3	技术设施指数 x_4	文化设施指数 x_5	卫生设施指数 x_6
上海	1.03	0.42	50	2.15	1.23	1.64	青岛	0.01	-0.14	24	0.37	-0.4	-0.49
深圳	1.34	0.13	131	0.33	-0.27	-0.64	武汉	0.02	-0.47	28	0.03	0.15	0.26
广州	1.07	0.4	48	1.31	0.49	0.09	温州	-0.47	0.03	45	-0.76	-0.46	-0.75
北京	-0.43	0.19	20	0.87	3.57	1.8	福州	-0.45	-0.2	34	-0.45	-0.34	-0.52
厦门	-0.53	0.25	32	-0.09	-0.33	-0.84	重庆	0.72	-0.83	13	0.05	-0.09	0.56
天津	-0.11	0.07	27	0.68	-0.12	0.87	成都	0.37	-0.54	21	-0.11	-0.24	-0.02
大连	0.35	0.06	31	0.28	-0.3	-0.16	宁波	0.01	0.38	40	-0.17	-0.4	-0.71
杭州	-0.5	0.27	38	-0.78	-0.12	1.61	石家庄	-0.81	-0.49	22	-0.38	-0.21	-0.59
南京	0.31	0.25	43	0.49	-0.09	-0.06	西安	-0.24	-0.91	18	-0.05	-0.27	0.61
珠海	-0.28	0.84	37	-0.79	-0.49	-0.98	哈尔滨	-0.53	-0.77	27	-0.45	-0.18	1.08

数据来源：倪鹏飞等:《城市竞争力蓝皮书：中国城市竞争力报告 NO.1》，北京，社会科学出版社 2003 年版。

（2）城市竞争力与基础设施的典型相关分析

将上述经过整理的指标数据利用 MATLAB 软件的 CANONCORR 函数进行处理，得出如下结果。

① 典型相关系数及其检验

典型相关系数及其检验如表 27 所示

表 27　典型相关系数

序号	1	2	3	4
典型相关系数	0.9601	0.9499	0.6470	0.3571

由上表可知，前两个典型相关系数均较高，表明相应典型变量之间密切相关。但要确定典型变量相关性的显著程度，尚需进行相关系数的 χ^2 统计量检验，具体做法是：比较统计量 χ^2 计算值与临界值的大小，据比较结果判定典型变量相关性的显著程度。其结果如表 28 所示。

表 28　相关系数检验表

序号	自由度	x^2 计算值	x^2 临界值(显著水平 0.05)
1	24	74.9775	3.7608e-007
2	15	40.8284	3.3963e-004

3	8	9.2942	0.3181
4	3	2.0579	0.5605

注：表中的 e-007 表示10^{-7}。

从上表看这 4 对典型变量均通过了 χ^2 统计量检验，表明相应典型变量之间相关关系显著，能够用城市基础设施变量组来解释城市竞争力变量组。

② 典型相关模型

鉴于原始变量的计量单位不同，不宜直接比较，本文采用标准化的典型系数，给出典型相关模型，如下表 29 所示：

表 29　典型相关模型

1	$u_1 = 0.1535x_1 + 0.3423x_2 + 0.4913x_3 + 0.3372x_4 + 0.1149x_5 + 0.1419x_6$ $v_1 = 0.1395y_1 + 0.7185y_2 + 0.427y_3 + 0.0285y_4$
2	$u_2 = -0.2134x_1 - 0.2637x_2 - 0.3953x_3 + 0.869x_4 - 0.2429x_5 + 0.3856x_6$ $v_2 = 0.1322y_1 - 0.7361y_2 + 0.772y_3 + 0.0059y_4$

由表 29 第一组典型相关方程可知，基础设施方面的主要因素是 x_2, x_3, x_4（典型载荷分别为 0.3423，0.4913，0.3372），说明基础设施中影响城市竞争力的主要因素是对内设施指数（x_2）、每百人电话数（x_3）和技术设施指数（x_4）；城市竞争力的第一典型变量 v_1 与 y_2 呈高度相关，说明在城市竞争力中，市场占有率（y_2）占有主要地位。根据第二组典型相关方程，x_4（技术设施指数）是基础设施方面的主要因素，而居民人均收入（y_3）（典型载荷为 0.869），是反映城市竞争力的一个重要指标。由于第一组典型变量占有信息量比重较大，所以总体上基础设施方面的主要因素按重要程度依次是 x_3, x_2, x_4，反映城市竞争力的主要指标是 y_2, y_3。

③ 典型结构

结构分析是依据原始变量与典型变量之间的相关系数给出的，如表 30 所示。

表 30　结构分析(相关系数)

	u_1	u_2	v_1	v_2

x_1	-0.71449	0.094452	-0.68599	0.089723
x_2	-0.63728	-0.34418	-0.61185	-0.32695
x_3	-0.71902	-0.54257	-0.69034	-0.5154
x_4	-0.72322	0.632013	-0.69437	0.600373
x_5	-0.41018	0.468804	-0.39381	0.445334
x_6	-0.1968	0.725205	-0.18895	0.688899
	v_1	v_2	u_1	u_2
y_1	-0.62924	-0.49738	-0.60414	-0.47248
y_2	-0.8475	0.529457	-0.81369	0.502951
y_3	-0.69906	-0.70239	-0.67117	-0.66722
y_4	-0.16928	-0.38871	-0.16253	-0.36925

由表30知，x_1, x_2, x_3, x_4与“基础设施组”的第一典型变量u_1均呈高度相关，说明对外设施、对内设施、每白人电话数和技术设施在反映城市基础设施方面占有主导地位，其中又以技术设施居于首位。x_2与基础设施组的第二变量和竞争力组的第二变量都呈高度相关，但由于u_2, v_2所含信息量比较低，故总体上看x_2对城市竞争力影响较小。“竞争力组”的第一典型变量v_1与y_2的相关系数均比较高，体现了y_2在反映城市竞争力中占有主导地位。y_3与v_1呈较高相关，与v_2呈高相关，但v_2凝聚的信息量有限，因而y_3在“竞争力”中的贡献低于y_2。由于第一对典型变量之间的高度相关，导致“基础设施组”中四个主要变量与“竞争力组”中的第一典型变量呈高度相关；而“竞争力组”中的y_2则与“影响组”的第一典型变量也呈高度相关。这种一致性从数量上体现了“基础设施组”对“竞争力组”的本质影响作用，与指标的实际经济联系非常吻合，说明典型相关分析结果具有较高的可信度。

值得一提的是，与线性回归模型不同，相关系数与典型系数可以有不同的符号。如基础设施方面的u_2与x_5相关系数为正值（0.468804），而典型系数却为负值（-0.2429）；竞争力方面的v_2与y_3，相关系数为负值（-0.70239），而典型系数却为正值（0.772）。由于出现这种反号的情况，称x_5, y_3为抑制变量(Suppressor)。由表30的相关系数还可以看出，“影响组”的第一典型变量u_1对y_2（市场占有率）有相当高的预测能力，系数值

为-0.81369，而对 y_4（长期经济增长率）预测能力较差，系数值仅为-0.16253。

④ 典型冗余分析与解释能力

典型相关系数的平方的实际意义是一对典型变量之间的共享方差在两个典型变量各自方差中的比例。

典型冗余分析用来表示各典型变量对原始变量组整体的变差解释程度，分为组内变差解释和组间变差解释，典型冗余分析的结果见表31和表32。

表31　被典型变量解释的 X 组原始变量的方差

被本组的典型变量解释			典型相关系数平方	被对方 Y 组典型变量解释		
	比例	累计比例			比例	累积比例
u_1	0.3606	0.3606	0.9218	v_1	0.3324	0.3324
u_2	0.2612	0.6218	0.9024	v_2	0.2357	0.5681
u_3	0.0631	0.6849	0.4186	v_3	0.0264	0.5945
u_4	0.0795	0.7644	0.1275	v_4	0.0101	0.6046

表32　被典型变量解释的 Y 组原始变量的方差

被本组的典型变量解释			典型相关系数平方	被对方 X 组典型变量解释		
	比例	累计比例			比例	累积比例
v_1	0.4079	0.4079	0.9218	u_1	0.4079	0.4079
v_2	0.2644	0.6723	0.9024	u_2	0.2930	0.7009
v_3	0.0648	0.7371	0.4186	u_3	0.1549	0.8558
v_4	0.0184	0.7555	0.1275	u_4	0.1442	1

从上表 31 和表 32 可以看出，两对典型变量 u_1、u_2 和 v_1、v_2 均较好地预测了对应的那组变量，而且交互解释能力也比较强。来自城市“竞争力组”的方差被“基础设施组”典型变量 u_1、u_2 解释的比例和为 70.09%；来自“基础设施组”的方差被“竞争力组”典型变量 v_1、v_2 解释的方差比例和为 56.81%。城市竞争力变量组被其自身及其对立典型变量解释的百分比、基础设施变量组被其自身及其对立典型变量解释的百分比均较高，尤其是第一对典型变量具有较高的解释百分比，反映两者之间较高的相关性。

4．城市竞争力与基础设施关系的经济分析

根据城市竞争力与基础设施关系的典型相关分析结果，城市竞争力与基础设施之间的关系可从下列三个方面进行阐述：

（1）市场占有率是决定城市竞争力水平的首要指标，每白人电话数、设施指数和技术设施指数是影响城市竞争力的主要基础设施变量。

市场占有率是企业竞争力大小的最直接表现，它反映一个城市域外产品需求的大小和其产品在全部城市产品市场中的份额，反映了一个城市创造价值的相对规模。根据典型载荷的大小可知，影响市场占有率的最主要因素是技术设施指数。技术设施指数是城市现代交通、通讯、信息设施的综合指数，由先进交通设施指标港口个数、机场等级、高速公路、高速铁路、地铁个数、光缆线路数加权而成，是一个主客观结合指标，它代表了一个城市的物流和信息流传播水平和扩散速度。第一典型变量显示，城市竞争力中的市场占有率与基础设施关系最密切，影响一个城市市场占有率的基础设施因素主要是交通和信息设施，这也是与信息时代的发展相一致的。因此，第一典型变量真实的反映了城市竞争力与基础设施力之间的本质联系，它将市场占有率从竞争力中提取出来，强调了信息基础设施建设对提升城市竞争力的重要性。

（2）城市居民人均收入是反映城市竞争力的另外一个重要变量。

城市居民人均收入和长期经济增长率综合反映了城市在域内和域外创造价值的状况。城市居民人均收入是城市创造价值在其域内成员收益上的直接反映，而城市吸引、占领、争夺、控制资源和市场创造价值的能力、潜力及持续性决定于 GDP 的长期增长，即 GDP 增长率反映了城市价值扩展的速度和潜力。因此，居民人均收入可以综合反映出一个城市吸引、控制资源和创造市场价值的能力和潜力。基础设施建设中的对内设施指数通过城市能源、交通、道路、住房和卫生设施条件等影响并制约着城市吸引、利用资源并创造价值的能力和水平。由于现在城市的竞争不再是自然资源的单一竞争，人才竞争已成为竞争的主要对象和核心，占有人才便控制了城市竞争的制高点，也就决定了城市创造价值的能力和潜力。而城市能源是价值创造的基础，交通、道路、住房及卫生设施等决定着城市利用资源和对人才的吸引力。因此，城市基础设施中的对内设施建设对提升城市竞争力具有重要作用。第二对典型变量还说明，每百人电话数和技术设施指数与居民人均收入和长期经济增长率反方向增长，电话和技术设施方面的投资在一定程度上影响了城市利用资源、创造价值的水平。因为电话的数量和技术设施投资必然要占用城市有限的人力、物力资源，短时期内会影响城市居民人均收入水平和 GDP 的增长。

（3）劳动生产率在我国城市竞争力中的作用尚不明显。

从以上典型分析结果可以得出，目前我国劳动生产率在城市竞争力中的重要作用尚不明显，这可能源于两个原因：一是我国各城市的劳动生产率低，对城市竞争力的贡献率不高；二是城市基础设施建设与劳动生产率之间的相关度不高。但相关研究成果显示，中国目前的劳动生产率并不低，不能否认劳动生产率在城市竞争力中的作用（张金昌，2002），如果这一结论成立，则对这一问题唯一的解释就是城市基础设施建设与劳动生产率的关联度不高。

计算的 MATLAB 程序如下：

```
clc,clear
load x.txt      %原始的x组的数据保存在纯文本文件x.txt中
load y.txt      %原始的y组的数据保存在纯文本文件y.txt中
n1=size(x,2);n2=size(y,2);
x=zscore(x);y=zscore(y);      %标准化数据
```

```
n=size(x,1);
%a,b返回的是典型变量的系数，r返回的是典型相关系数
%u,v返回的是典型变量的值，stats返回的是假设检验的一些统计量的值
[a,b,r,u,v,stats]=canoncorr(x,y)
x_u_r=x'*u/(n-1)     %计算x,u的相关系数
y_v_r=y'*v/(n-1)     %计算y,v的相关系数
x_v_r=x'*v/(n-1)     %计算x,v的相关系数
y_u_r=y'*u/(n-1)     %计算y,u的相关系数
mu=sum(x_u_r.^2)/n1     %x组原始变量被u_i解释的方差比例
mv=sum(x_v_r.^2)/n1     %x组原始变量被v_i解释的方差比例
nu=sum(y_u_r.^2)/n2     %y组原始变量被u_i解释的方差比例
nv=sum(y_v_r.^2)/n2     %y组原始变量被v_i解释的方差比例
val=r.^2                %典型系数的平方
```

习题二十九

1．表 33 是 1999 年中国省、自治区的城市规模结构特征的一些数据，试通过聚类分析将这些省、自治区进行分类。

表 33　城市规模结构特征数据

省、自治区	城市规模（万人）	城市首位度	城市指数	基尼系数	城市规模中位值（万人）
京津冀	699.70	1.4371	0.9364	0.7804	10.880
山西	179.46	1.8982	1.0006	0.5870	11.780
内蒙古	111.13	1.4180	0.6772	0.5158	17.775
辽宁	389.60	1.9182	0.8541	0.5762	26.320
吉林	211.34	1.7880	1.0798	0.4569	19.705
黑龙江	259.00	2.3059	0.3417	0.5076	23.480
苏沪	923.19	3.7350	2.0572	0.6208	22.160
浙江	139.29	1.8712	0.8858	0.4536	12.670
安徽	102.78	1.2333	0.5326	0.3798	27.375
福建	108.50	1.7291	0.9325	0.4687	11.120
江西	129.20	3.2454	1.1935	0.4519	17.080
山东	173.35	1.0018	0.4296	0.4503	21.215
河南	151.54	1.4927	0.6775	0.4738	13.940
湖北	434.46	7.1328	2.4413	0.5282	19.190
湖南	139.29	2.3501	0.8360	0.4890	14.250
广东	336.54	3.5407	1.3863	0.4020	22.195
广西	96.12	1.2288	0.6382	0.5000	14.340
海南	45.43	2.1915	0.8648	0.4136	8.730

川渝	365.01	1.6801	1.1486	0.5720	18.615
云南	146.00	6.6333	2.3785	0.5359	12.250
贵州	136.22	2.8279	1.2918	0.5984	10.470
西藏	11.79	4.1514	1.1798	0.6118	7.315
陕西	244.04	5.1194	1.9682	0.6287	17.800
甘肃	145.49	4.7515	1.9366	0.5806	11.650
青海	61.36	8.2695	0.8598	0.8098	7.420
宁夏	47.60	1.5078	0.9587	0.4843	9.730
新疆	128.67	3.8535	1.6216	0.4901	14.470

2. 表 34 是我国 1984—2000 年宏观投资的一些数据，试利用主成分分析对投资效益进行分析和排序。

表34　1984—2000年宏观投资效益主要指标

年份	投资效果系数（无时滞）	投资效果系数（时滞一年）	全社会固定资产交付使用率	建设项目投产率	基建房屋竣工率
1984	0.71	0.49	0.41	0.51	0.46
1985	0.40	0.49	0.44	0.57	0.50
1986	0.55	0.56	0.48	0.53	0.49
1987	0.62	0.93	0.38	0.53	0.47
1988	0.45	0.42	0.41	0.54	0.47
1989	0.36	0.37	0.46	0.54	0.48
1990	0.55	0.68	0.42	0.54	0.46
1991	0.62	0.90	0.38	0.56	0.46
1992	0.61	0.99	0.33	0.57	0.43
1993	0.71	0.93	0.35	0.66	0.44
1994	0.59	0.69	0.36	0.57	0.48
1995	0.41	0.47	0.40	0.54	0.48
1996	0.26	0.29	0.43	0.57	0.48
1997	0.14	0.16	0.43	0.55	0.47
1998	0.12	0.13	0.45	0.59	0.54
1999	0.22	0.25	0.44	0.58	0.52
2000	0.71	0.49	0.41	0.51	0.46

3. 表35资料为25名健康人的7项生化检验结果，7项生化检验指标依次命名为 $x_1, x_2, \cdots, x_7$，请对该资料进行因子分析。

表35　检验数据

x_1	x_2	x_3	x_4	x_5	x_6	x_7

3.76	3.66	0.54	5.28	9.77	13.74	4.78
8.59	4.99	1.34	10.02	7.5	10.16	2.13
6.22	6.14	4.52	9.84	2.17	2.73	1.09
7.57	7.28	7.07	12.66	1.79	2.1	0.82
9.03	7.08	2.59	11.76	4.54	6.22	1.28
5.51	3.98	1.3	6.92	5.33	7.3	2.4
3.27	0.62	0.44	3.36	7.63	8.84	8.39
8.74	7	3.31	11.68	3.53	4.76	1.12
9.64	9.49	1.03	13.57	13.13	18.52	2.35
9.73	1.33	1	9.87	9.87	11.06	3.7
8.59	2.98	1.17	9.17	7.85	9.91	2.62
7.12	5.49	3.68	9.72	2.64	3.43	1.19
4.69	3.01	2.17	5.98	2.76	3.55	2.01
5.51	1.34	1.27	5.81	4.57	5.38	3.43
1.66	1.61	1.57	2.8	1.78	2.09	3.72
5.9	5.76	1.55	8.84	5.4	7.5	1.97
9.84	9.27	1.51	13.6	9.02	12.67	1.75
8.39	4.92	2.54	10.05	3.96	5.24	1.43
4.94	4.38	1.03	6.68	6.49	9.06	2.81
7.23	2.3	1.77	7.79	4.39	5.37	2.27
9.46	7.31	1.04	12	11.58	16.18	2.42
9.55	5.35	4.25	11.74	2.77	3.51	1.05
4.94	4.52	4.5	8.07	1.79	2.1	1.29
8.21	3.08	2.42	9.1	3.75	4.66	1.72
9.41	6.44	5.11	12.5	2.45	3.1	0.91

4．为了了解家庭的特征与其消费模式之间的关系。调查了70个家庭的下面两组变量：

$$\begin{cases} x_1: \text{每年去餐馆就餐的频率} \\ x_2: \text{每年外出看电影频率} \end{cases}, \quad \begin{cases} y_1: \text{户主的年龄} \\ y_2: \text{家庭的年收入} \\ y_3: \text{户主受教育程度} \end{cases}$$

已知相关系数矩阵见表36，试对两组变量之间的相关性进行典型相关分析。

表36　相关系数矩阵

	x_1	x_2	y_1	y_2	y_3
x_1	1	0.8	0.26	0.67	0.34
x_2	0.8	1	0.33	0.59	0.34
y_1	0.26	0.33	1	0.37	0.21

y_2	0.67	0.59	0.37	1	0.35
y_3	0.34	0.34	0.21	0.35	1

5．近年来我国淡水湖水质富营养化的污染日趋严重，如何对湖泊水质的富营养化进行综合评价与治理是摆在我们面前的一项重要任务。表 37 和表 38 分别为我国 5 个湖泊的实测数据和湖泊水质评价标准。

表 37　全国 5 个主要湖泊评价参数的实测数据

	总磷（mg/L）	耗氧量（mg/L）	透明度（L）	总氮（mg/L）
杭州西湖	130	10.3	0.35	2.76
武汉东胡	105	10.7	0.4	2.0
青海湖	20	1.4	4.5	0.22
巢湖	30	6.26	0.25	1.67
滇池	20	10.13	0.5	0.23

表 38　湖泊水质评价标准

评价参数	极贫营养	贫营养	中营养	富营养	极富营养
总磷	<1	4	23	110	>660
耗氧量	<0.09	0.36	1.8	7.1	>27.1
透明度	>37	12	2.4	0.55	<0.17
总氮	<0.02	0.06	0.31	1.2	>4.6

（1）试利用以上数据，分析总磷、耗氧量、透明度和总氮这 4 种指标对湖泊水质富营养化所起作用。

（2）对上述 5 个湖泊的水质进行综合评估，确定水质等级。