第10章　多元分析

多元分析（multivariate analysis）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，在很多工程领域有着广泛使用，并在使用中不断完善和创新。

10.1　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，比如有关世界时间进程的研究，就形成了历史学，有关世界空间地域的研究，就形成了地理学。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研究更清晰、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 analysis）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。

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。

10.1.1　Q 型聚类分析

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

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

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

当$q=1,2$或$q\to+\infty$时，则分别得到

（1）绝对值距离

$$d_1(x,y)=\sum_{k=1}^{p}\left|x_k-y_k\right|, \qquad (10.1)$$

（2）欧氏距离

$$d_2(x,y)=\left[\sum_{k=1}^{p}\left|x_k-y_k\right|^2\right]^{\frac{1}{2}}, \qquad (10.2)$$

（3）Chebyshev 距离

$$d_\infty(x,y)=\max_{1\le k\le p}\left|x_k-y_k\right|. \qquad (10.3)$$

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

$$d(x,y)=\sqrt{(x-y)^T\Sigma^{-1}(x-y)}\ ,\qquad (10.4)$$

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。

2　类与类间的相似性度量

如果有两个样本类G_1和G_2，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）

$$D(G_1,G_2)=\min_{\substack{x_i\in G_1\\ y_j\in G_2}}\{d(x_i,y_j)\},\qquad(10.5)$$

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）

$$D(G_1,G_2)=\max_{\substack{x_i\in G_1\\ y_j\in G_2}}\{d(x_i,y_j)\},\qquad (10.6)$$

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）

$$D(G_1,G_2)=d(\bar{x},\bar{y}), \qquad (10.7)$$

其中$\bar{x},\bar{y}$分别为G_1,G_2的重心。

（4）类平均法（group average method）

$$D(G_1,G_2)=\frac{1}{n_1 n_2}\sum_{x_i\in G_1}\sum_{x_j\in G_2} d(x_i,x_j), \qquad (10.8)$$

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

$$D_1 = \sum_{x_i \in G_1} (x_i - \bar{x}_1)^T (x_i - \bar{x}_1),$$

$$D_2 = \sum_{x_j \in G_2} (x_j - \bar{x}_2)^T (x_j - \bar{x}_2),$$

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

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

$$D(G_1, G_2) = D_{12} - D_1 - D_2. \qquad (10.9)$$

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。

3　聚类图

Q 型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 10.1(a)），可以用聚类图（如图 10.1（b））来表示聚类结果。

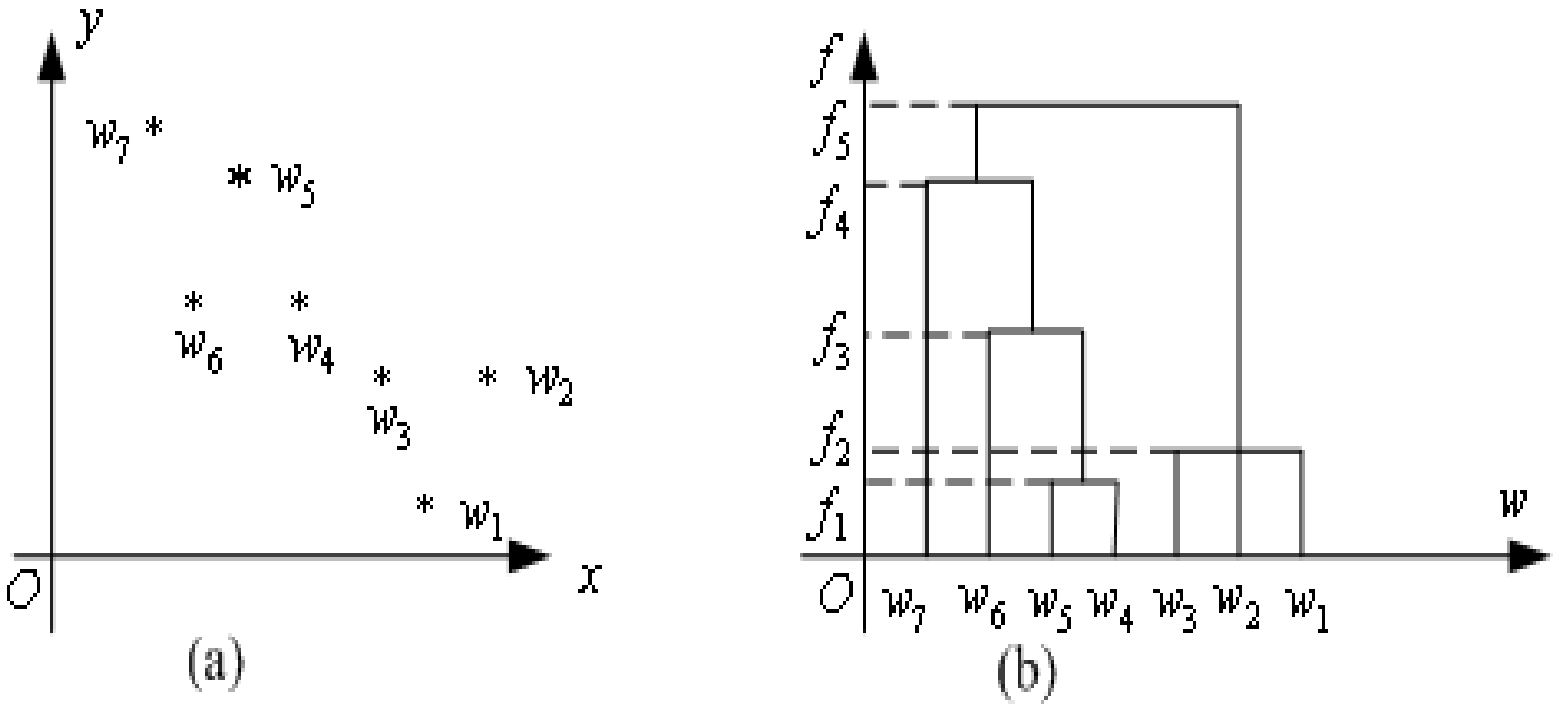

图 1　聚类方法示意图

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

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

当距离值为f_4分为两类

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

当距离值为f_3分为三类

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

当距离值为 f_2 分为四类

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

当距离值为 f_1 分为六类

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

$$G_5=\{w_6\},\ G_6=\{w_7\}$$

当距离小于 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。

4　最短距离法的聚类举例

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），由 Florek 等人于 1951 年和 Sneath 于 1957 年引入。下面举例说明最短距离法的计算步骤。

例 10.1　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 10.1。

表 10.1　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。

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

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

$$H_5=\{h_9\}$$

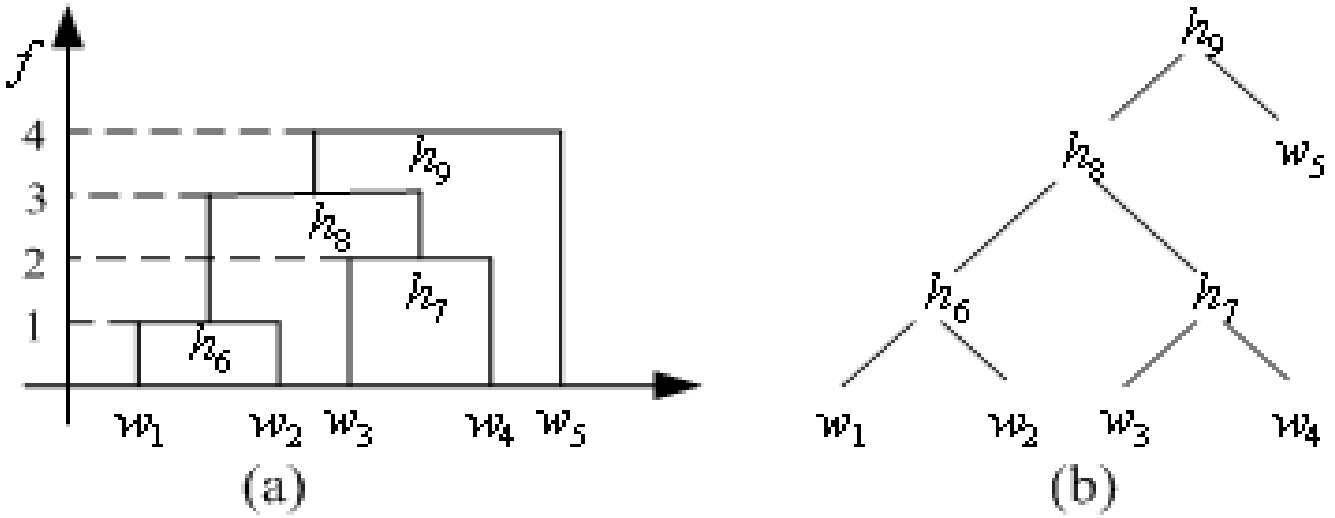

图 10.2　最短距离法

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 10.2（a））。这是一颗二叉树，如图 10.2（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的工作成绩较好，而w_1，w_2的工作成绩较差。

5　Matlab聚类分析的相关命令

Matlab中聚类分析相关命令的使用说明如下。

（1）pdist

Y=pdist(X)计算$m \times n$矩阵X（看作m个n维行向量）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10.2中特征字符串值。

表10.2 ’metric’取值及含义

字符串	含 义
'euclidean'	欧氏距离（缺省）
'seuclidean'	标准欧氏距离
'cityblock'	绝对值距离
'minkowski'	闵氏距离（Minkowski距离）
'chebychev'	车比雪夫距离（Chebychev距离）
'mahalanobis'	马氏距离（Mahalanobis距离）
'hamming'	海明距离（Hamming距离）

表10.2 ’metric’取值及含义

字符串	含 义
custom distance function	自定义函数距离
'cosine'	1-两个向量夹角的余弦
'correlation'	1-样本的相关系数
'spearman'	1-样本的Spearman秩相关系数
'jaccard'	1-Jaccard系数

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 10.3 中特征字符串值。

表10.3 ’method’取值及含义

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

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

第1列和第2列，即Z(:,[1:2])包含了被两两连接生成一个新类的所有对象的索引。生成的新类索引为$m+j$。共有$m-1$个级别更高的类，它们对应于聚类树中的内部节点。

第三列Z(:,3)包含了相应的在类中的两两对象间的连接距离。

(3）cluster

T = cluster(Z,'cutoff',c)从连接输出（linkage）中创建聚类。cutoff为定义cluster函数如何生成聚类的阈值，其不同的值含义如表10.4所示。

表10.4　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',c,'depth',d)从连接输出(linkage)中创建聚类。参数depth指定了聚类数中的层数，进行不一致系数计算时要用到。不一致系数将聚类树中两对象的连接与相邻的连接进行比较。详细说明见函数inconsistent。当参数depth被指定时，cutoff通常作为不一致系数阈值。

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

（4）zsore(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$\bar{x}_j, s_j$是矩阵$X=(x_{ij})_{m\times n}$每一列的均值和标准差。

（5）H＝dendrogram(Z,P)

由linkage产生的数据矩阵Z画聚类树状图。P是结点数，默认值是30。

（6）T=clusterdata(X,cutoff)

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

Y=pdist(X)

Z=linkage(Y,'single')

T=cluster(Z,cutoff)

(7)squareform

将pdist的输出转换为方阵。

10.1.2　R型聚类法

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　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

$$r_{jk} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\left[\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 \sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2\right]^{\frac{1}{2}}}, \qquad (10.10)$$

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

$$r_{jk}=\frac{\sum_{i=1}^{n}x_{ij}x_{ik}}{\left(\sum_{i=1}^{n}x_{ij}^2\sum_{i=1}^{n}x_{ik}^2\right)^{\frac{1}{2}}}. \tag{10.11}$$

各种定义的相似度量均应具有以下两个性质:

ⅰ）$|r_{jk}|\leq 1$，对于一切j,k；

ⅱ）$r_{jk}=r_{kj}$,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。

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，常用的有最长距离法、最短距离法等。

（1）最长距离法

在最长距离法中，定义两类变量的距离为

$$R(G_1,G_2)=\max_{\substack{x_j\in G_1\\ x_k\in G_2}}\{d_{jk}\},\qquad (10.12)$$

其中$d_{jk}=1-|r_{jk}|$或$d_{jk}^2=1-r_{jk}^2$，这时，$R(G_1,G_2)$与两类中相似性最小的两变量间的相似性度量值有关。

（2）最短距离法

在最短距离法中，定义两类变量的距离为

$$R(G_1,G_2)=\min_{\substack{x_j\in G_1\\ x_k\in G_2}}\{d_{jk}\},\qquad (10.13)$$

其中$d_{jk}=1-|r_{jk}|$或$d_{jk}^2=1-r_{jk}^2$，这时，$R(G_1,G_2)$与两类中相似性最大的两个变量间的相似性度量值有关。

例10.2　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 10.5）。

表10.5　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						

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 10.3。

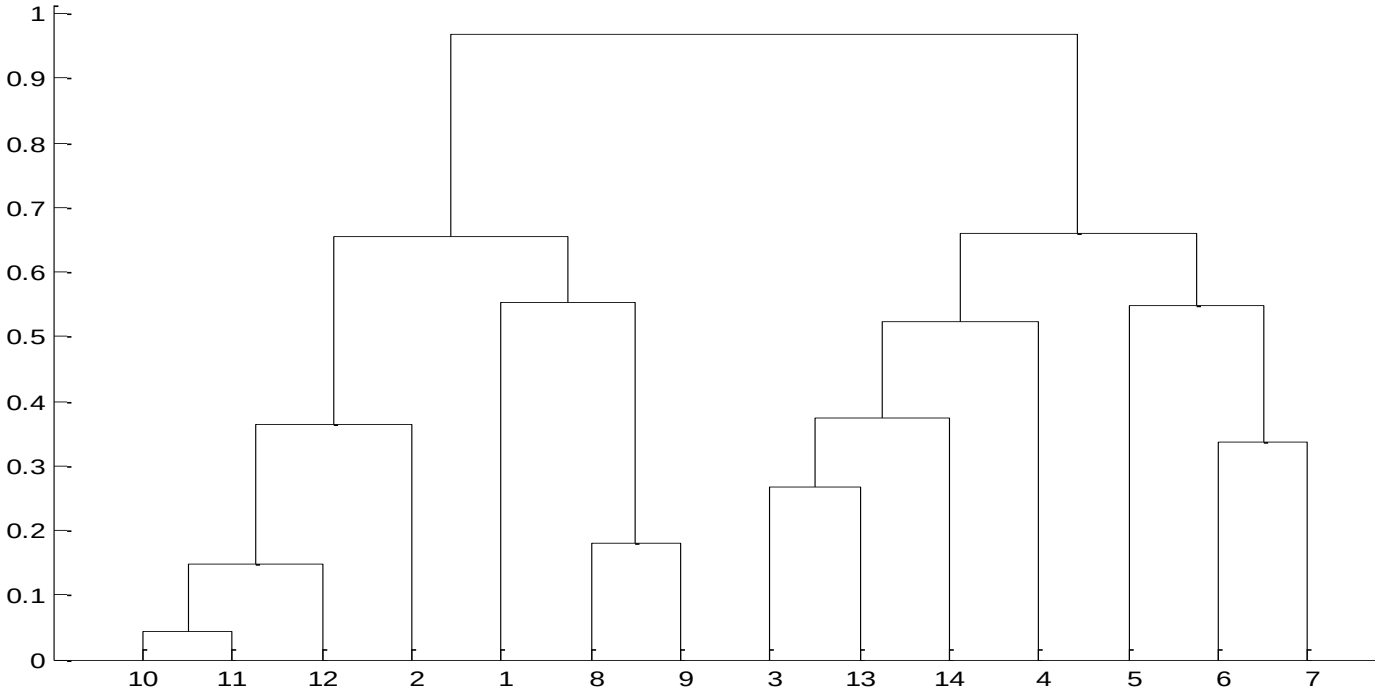

图10.3　成年女子14个部位指标的聚类图

通过聚类图，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。

10.1.3　聚类分析案例—我国各地区普通高等教育发展状况分析

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10.4。

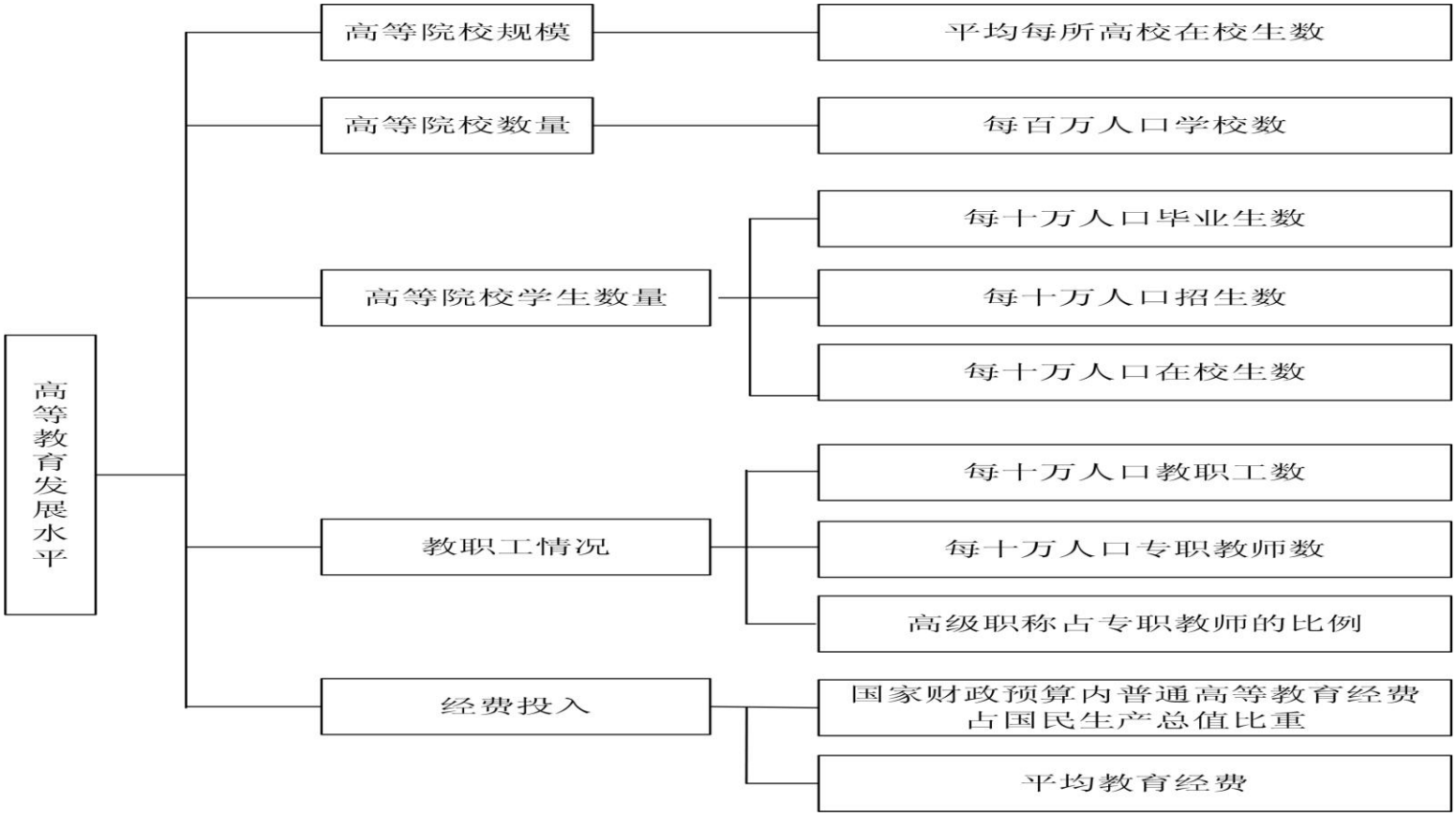

图10.4　高等教育的十项评价指标

（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》十项指标值见表 10.6。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。

表10.6　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

（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 10.7 所示。

表10.7　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10.5。

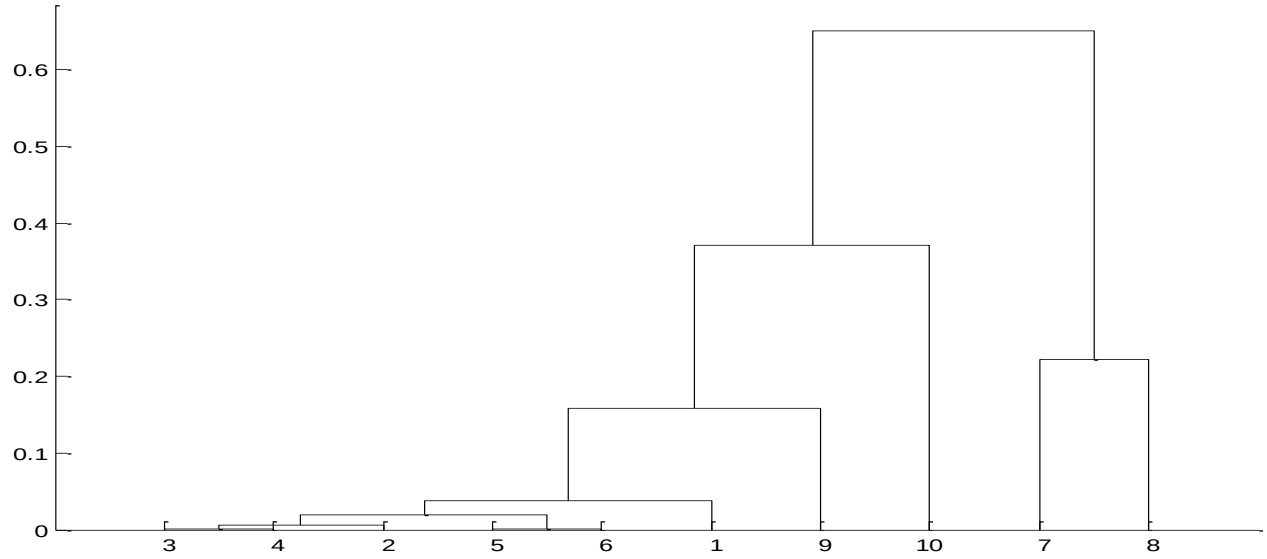

图 10.5　指标聚类树型图

从聚类图 10.5 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。

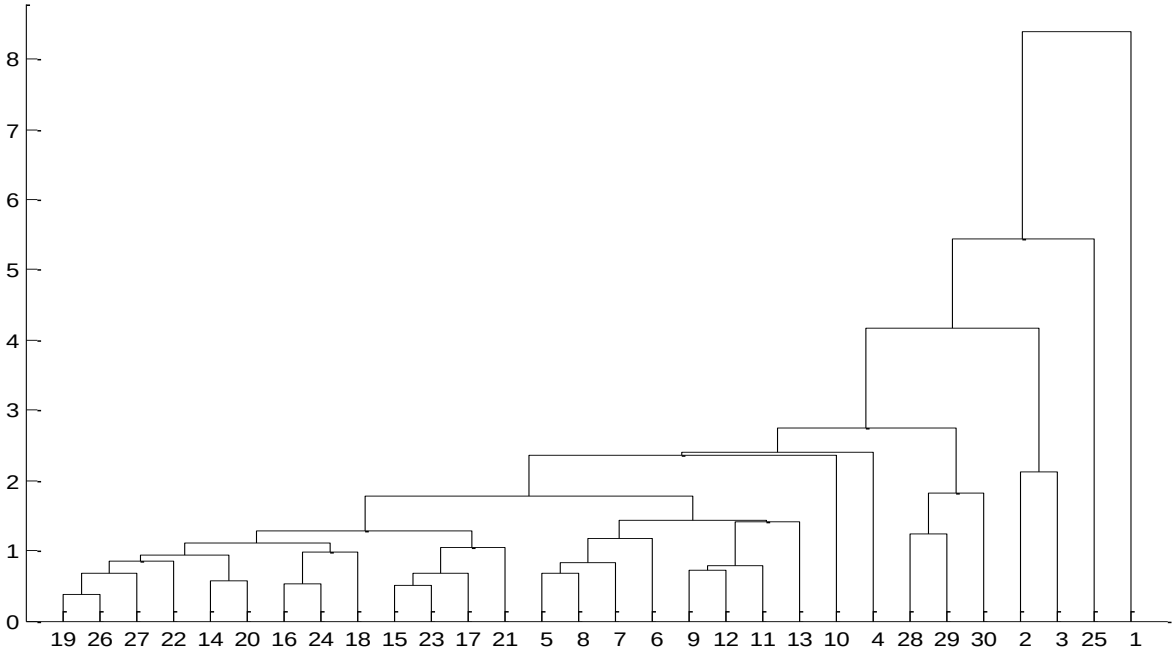

图 10.6　各地区聚类树型图

（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，样本间相似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 10.6。

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它地区不同的情形，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。

10.2　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。

10.2.1 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

$$s = c_1 x_1 + c_2 x_2 + \cdots + c_p x_p \quad , \qquad (10.14)$$

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

$$\mathbf{Var}(c_1X_1 + c_2X_2 + \cdots + c_pX_p) \qquad (10.15)$$

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，（10.15）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

$$c_1^2 + c_2^2 + \cdots + c_p^2 = 1 \qquad (10.16)$$

在此约束下，求（10.15）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

$$\begin{cases} Z_1 = c_{11}X_1 + c_{12}X_2 + \cdots + c_{1p}X_p, \\ Z_2 = c_{21}X_1 + c_{22}X_2 + \cdots + c_{2p}X_p, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ Z_p = c_{p1}X_1 + c_{p2}X_2 + \cdots + c_{pp}X_p, \end{cases} \tag{10.17}$$

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）使方差达到最大的主成分分析不用转轴（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。

10.2.2　特征值因子的筛选

设有 p 个指标变量 $x_1, x_2, \cdots, x_p$，它在第 i 次试验中的取值为

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

将它们写成矩阵形式

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{np} \end{bmatrix}, \tag{10.18}$$

矩阵 A 称为设计阵。

实际中确定（10.17）式中的系数就是采用矩阵A^TA的特征向量。因此，剩下的问题仅仅是将A^TA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，近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。

注：使用$\tilde{x}_i = (x_i - \mu_i)/\sigma_i$对数据进行标准化后，得到的标准化数据矩阵记为$\tilde{A}$，由于$x_1, x_2, \cdots, x_p$的相关系数矩阵$R = \tilde{A}^T \tilde{A}/(n-1)$，在主成分分析中我们只需计算相关系数矩阵$R$的特征值和特征向量即可。

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

$$\rho_i = \sum_{j=1}^{r} r^2(z_j, x_i), \qquad (10.19)$$

这里$r(z_j, x_i)$表示z_j与x_i的相关系数。

例10.3　设$x=[x_1,x_2,x_3]^T$，且

$$A^TA=\begin{bmatrix}1&-2&0\\-2&5&0\\0&0&0\end{bmatrix},$$

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。

10.2.4　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 10.1.3 节中的问题相同，我们这里就不重复叙述了。

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，分别为$x_1, x_2, \cdots, x_m$，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a_{ij}。将各指标值a_{ij}转换成标准化指标值$\tilde{a}_{ij}$，

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

其中$\mu_j = \frac{1}{n}\sum_{i=1}^{n} a_{ij}$，$s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(a_{ij} - \mu_j)^2}$，$(j = 1,2,\cdots,m)$，

即μ_j, s_j为第j个指标的样本均值和样本标准差。

对应地，称

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

（2）计算相关系数矩阵$\boldsymbol{R}$

相关系数矩阵$\boldsymbol{R}=(r_{ij})_{m\times m}$

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

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$\boldsymbol{R}$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$u_j = [u_{1j}, u_{2j}, \cdots, u_{mj}]^T$，由特征向量组成$m$个新的指标变量

$$y_1 = u_{11}\tilde{x}_1 + u_{21}\tilde{x}_2 + \cdots + u_{m1}\tilde{x}_m,$$
$$y_2 = u_{12}\tilde{x}_1 + u_{22}\tilde{x}_2 + \cdots + u_{m2}\tilde{x}_m,$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots,$$
$$y_m = u_{1m}\tilde{x}_1 + u_{2m}\tilde{x}_2 + \cdots + u_{mm}\tilde{x}_m,$$

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（$p \le m$）个主成分，计算综合评价值

ⅰ）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（一般取 $\alpha_p = 0.85, 0.90, 0.95$）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。

ii）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。

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10.10。

表10.10　主成分分析结果

序号	特征根	贡献率	累积贡献率
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，前两个特征根的累积贡献率就达到90%以上，主成分分析效果很好。下面选取前四个主成分（累积贡献率就达到98%）进行综合评价。前四个特征根对应的特征向量见表10.11。

表10.11　标准化变量的前4个主成分对应的特征向量

	$\tilde{x}_1$	$\tilde{x}_2$	$\tilde{x}_3$	$\tilde{x}_4$	$\tilde{x}_5$	$\tilde{x}_6$	$\tilde{x}_7$	$\tilde{x}_8$	$\tilde{x}_9$	$\tilde{x}_{10}$
1	0.34	0.35	0.36	0.36	0.36	0.36	0.22	0.12	0.31	0.24
2	-0.19	0.03	0.02	0.01	-0.05	-0.06	0.58	0.70	-0.19	-0.28
3	-0.16	-0.10	-0.09	-0.11	-0.15	-0.16	-0.03	0.35	0.12	0.86
4	-0.10	-0.22	-0.16	-0.16	-0.04	-0.00	0.08	0.07	0.89	0.24

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10.12。(表略)

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。

10.4　判别分析

判别分析（discriminant 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。

10.4.1　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 10.7）。

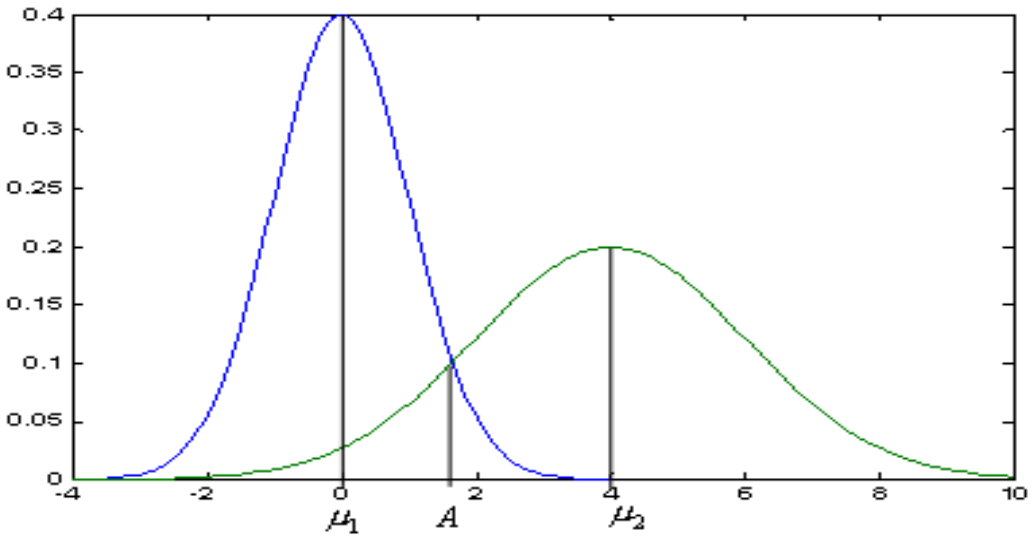

图 10.7　不同均值、方差的正态分布

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 10.7 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$1.17\sigma_2$，因此，应该认为A点距μ_2更近一点。

定义 10.1　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

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

要判断x来自哪一个总体，需要计算x到总体A和BMahalanobis 距离$d(x, A)$和$d(x, B)$，然后进行比较，若$d(x, A) \leq d(x, B)$，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

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

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

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

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

$$w(x)=(x-\bar{\mu})^T\Sigma^{-1}(\mu_1-\mu_2), \qquad (10.30)$$

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$n_1$个样本点，$x_1^{(2)}, x_2^{(2)}, \cdots, x_{n_2}^{(2)}$是来自总体$B$的$n_2$个样本点，则样本的均值与协方差为

$$\hat{\mu}_i = \bar{x}^{(i)} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_j^{(i)}, \quad i = 1,2, \tag{10.31}$$

$$\hat{\Sigma} = \frac{1}{n_1 + n_2 - 2}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(x_j^{(i)} - \bar{x}^{(i)})(x_j^{(i)} - \bar{x}^{(i)})^T = \frac{1}{n_1 + n_2 - 2}(S_1 + S_2), \tag{10.32}$$

其中

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

对于待测样本x，其判别函数定义为

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

其中

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

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

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

其中

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

$i=1,2.$

10.4.2 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

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

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

其方差为

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

考虑比

$$\frac{(\mu_{y_1}-\mu_{y_2})^2}{\sigma_y^2}=\frac{[a^T(\mu_1-\mu_2)]^2}{a^T\Sigma a}=\frac{(a^T\delta)^2}{a^T\Sigma a}, \qquad (10.33)$$

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（10.33）式达到最大。

定理 10.1　x为p维随机变量，设$y=a^T x$，当选取$a=c\Sigma^{-1}\delta$，$c\neq 0$为常数时，（10.33）式达到最大。

特别当$c=1$时，线性函数

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

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 10.2　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 10.2 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

$$\begin{aligned} W(x)&=(\mu_1-\mu_2)^T\Sigma^{-1}x-K \\ &=(x-\frac{1}{2}(\mu_1+\mu_2))^T\Sigma^{-1}(\mu_1-\mu_2) \end{aligned}, \tag{10.34}$$

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，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。

10.4.3 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，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

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

(10.35)

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，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（10.35）式达到极小。

定理 10.3　极小化平均误判损失(10.35)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$\frac{f_1(x)}{f_2(x)}=\frac{L(1|2)}{L(2|1)}\cdot\frac{p_2}{p_1}$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，得到两总体的 Bayes 判别准则

$$\begin{cases} x \in X_1, & \text{当}x\text{使得} \dfrac{f_1(x)}{f_2(x)} \geq \dfrac{L(1|2)}{L(2|1)} \cdot \dfrac{p_2}{p_1}, \\ x \in X_2, & \text{当}x\text{使得} \dfrac{f_1(x)}{f_2(x)} < \dfrac{L(1|2)}{L(2|1)} \cdot \dfrac{p_2}{p_1}. \end{cases} \tag{10.36}$$

应用此准则时仅仅需要计算

（1）新样本点 $x_0=[x_{01},x_{02},\cdots,x_{p0}]^T$ 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（10.36）的三种特殊情况：

（1）当 $p_2 / p_1 = 1$

$$\begin{cases} x \in X_1, \quad \text{当}x\text{使得} \dfrac{f_1(x)}{f_2(x)} \geq \dfrac{L(1|2)}{L(2|1)}, \\ x \in X_2, \quad \text{当}x\text{使得} \dfrac{f_1(x)}{f_2(x)} < \dfrac{L(1|2)}{L(2|1)}. \end{cases} \tag{10.37}$$

（2）当 $L(1|2)/L(2|1)=1$ 时

$$\begin{cases} x \in X_1, & \text{当}x\text{使得} \dfrac{f_1(x)}{f_2(x)} \geq \dfrac{p_2}{p_1}, \\ x \in X_2, & \text{当}x\text{使得} \dfrac{f_1(x)}{f_2(x)} < \dfrac{p_2}{p_1}. \end{cases} \tag{10.38}$$

（3） $p_1 / p_2 = L(1|2) / L(2|1) = 1$时

$$\begin{cases} x \in X_1, & \text{当}x\text{使得} \dfrac{f_1(x)}{f_2(x)} \geq 1, \\ x \in X_2, & \text{当}x\text{使得} \dfrac{f_1(x)}{f_2(x)} < 1. \end{cases} \tag{10.39}$$

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（10.37），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（10.38），如果上述两者都难以确定则利用规则（10.39），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 $x \in X_2$。

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

$$f_i(x) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\{-\frac{1}{2}(x-\mu_i)^T \Sigma^{-1}(x-\mu_i)\}. \quad (10.40)$$

定理 10.4　设总体 $X_i \sim N_p(\mu_i, \Sigma)(i=1,2)$，其中 $\Sigma > 0$，则使平均误判损失极小的划分为

$$\begin{cases} R_1 = \{x : W(x) \geq \beta\}, \\ R_2 = \{x : W(x) < \beta\}. \end{cases} \tag{10.41}$$

其中

$$W(x) = [x - \frac{1}{2}(\mu_1 + \mu_2)]^T \Sigma^{-1}(\mu_1 - \mu_2), \tag{10.42}$$

$$\beta = \ln \frac{L(1|2) \cdot p_2}{L(2|1) \cdot p_1}. \tag{10.43}$$

不难发现（10.42）式的$W(x)$与 Fisher 判别和马氏距离判别的线性判别函数（10.34），（10.30）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（10.31）和（10.32），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

$$W(x)=[x-\frac{1}{2}(\hat{\mu}_1+\hat{\mu}_2)]^T\hat{\Sigma}^{-1}(\hat{\mu}_1-\hat{\mu}_2) \qquad (10.44)$$

称为 Anderson 线性判别函数，判别的规则为

$$\begin{cases} x\in X_1, & \text{当}x\text{使得}W(x)\geq\beta, \\ x\in X_2, & \text{当}x\text{使得}W(x)<\beta, \end{cases} \qquad (10.45)$$

其中β由（10.43）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$\mu_1,\mu_2,\Sigma$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 10.3 可得判别区域

$$\begin{cases} R_1 = \{x : W(x) \geq K\}, \\ R_2 = \{x : W(x) < K\}, \end{cases} \tag{10.46}$$

其中

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

(10.47)

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

(10.48)

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 10.15　表 10.24 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 10.24　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 10.24 的数据计算得

$$\hat{\mu}_1 = [25.3167,\ -2.4167]^T,\ \hat{\mu}_2 = [22.0250,\ -1.1875]^T,$$

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

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 10.23 的数据代入$W(x)$，计算的结果填在表 10.23 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％。回代结果是春旱中有一个样本点误判。

10.4.4　应用举例

例 10.16　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 10.25。

表 10.25　生产厂家的数据

厂家	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
式样	9	7	8	8	9	8	7	4	3	6	2	1	6	8	2
包装	8	6	7	5	9	9	5	4	6	3	4	2	4	1	4
耐久性	7	6	8	5	3	7	6	4	6	3	5	2	5	3	5
类别	1	1	1	1	1	1	1	2	2	2	2	2	待判	待判	待判

今有 3 家新的厂家，得分分别为[6，4，5]，[8，1，3]，[2，4，5]，试对 3 个新厂家进行分类。