

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 $\Omega \times \Omega \rightarrow R^+$ 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

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

当 $q = 1, 2$ 或 $q \rightarrow +\infty$ 时，则分别得到

(1) 绝对值距离

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

(2) 欧氏距离

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

(3) Chebyshev 距离

$$d_\infty(x, y) = \max_{1 \leq k \leq p} |x_k - y_k|. \quad (10.3)$$

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

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

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)

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

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)

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

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)

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

其中 \bar{x}, \bar{y} 分别为 G_1, G_2 的重心。

(4) 类平均法 (group average method)

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

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

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

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, \dots, 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。

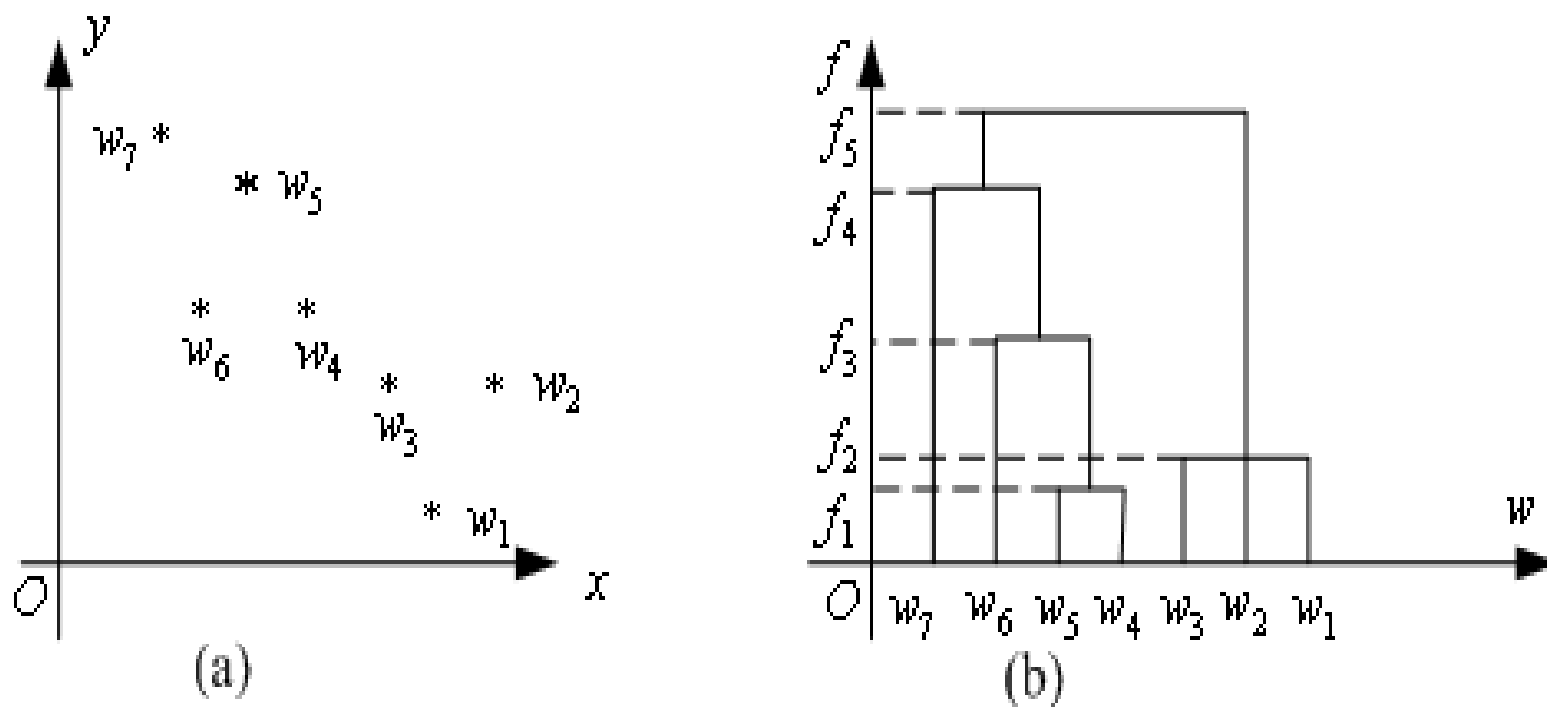


图 1 聚类方法示意图

记 $\Omega = \{w_1, w_2, \dots, w_7\}$, 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

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

当距离值为 f_3 分为三类

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

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 $\Omega = \{w_1, w_2, \dots, w_7\}$,

(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

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

由距离公式 $d(\cdot, \cdot)$ ，可以算出距离矩阵。

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

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

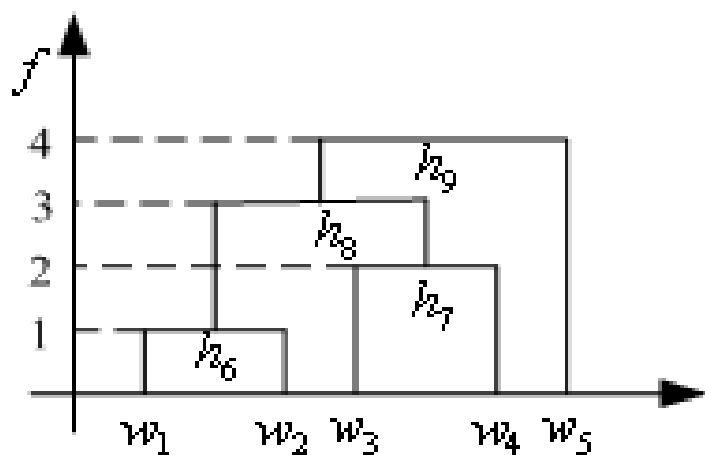
$$H_3 = \{h_6, h_7, w_5\}$$

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

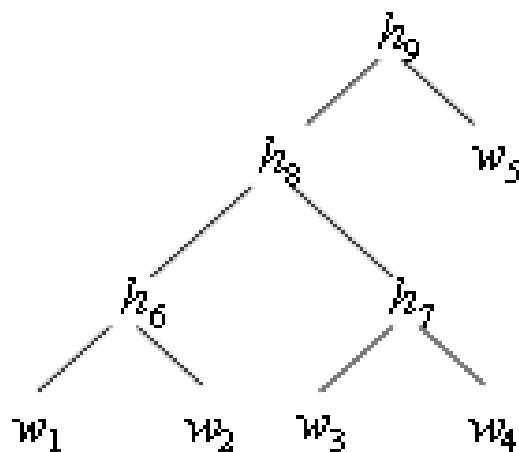
$$H_4 = \{h_8, w_5\}$$

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

$$H_5 = \{h_9\}$$



(a)



(b)

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

$Y = \text{pdist}(X)$ 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。

$Y = \text{pdist}(X, 'metric')$ 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'	切比雪夫距离（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

$Y = \text{pdist}(X, 'minkowski', p)$ 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

$T = \text{cluster}(Z, 'cutoff', c)$ 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
$0 < \text{cutoff} < 2$	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。
$2 \leq \text{cutoff}$	cutoff作为包含在聚类树中的最大分类数

$T = \text{cluster}(Z, 'cutoff', c, 'depth', d)$ 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 $\text{find}(T==i)$ 。

(4) zscore(X)

对数据矩阵进行标准化处理，处理方式

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

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

$(x_{1j}, x_{2j}, \dots, x_{nj})^T \in R^n (j = 1, 2, \dots, m)$ 。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

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

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

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

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 and 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

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

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

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

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	0.3 66	1												
x_3	0.2 42	0.2 33	1											
x_4	0.2 8	0.1 94	0.5 9	1										
x_5	0.3 6	0.3 24	0.4 76	0.4 35	1									
x_6	0.2 82	0.2 62	0.4 83	0.4 7	0.4 52	1								
x_7	0.2 45	0.2 65	0.5 4	0.4 78	0.5 35	0.6 63	1							
x_8	0.4 48	0.3 45	0.4 52	0.4 04	0.4 31	0.3 22	0.2 66	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。

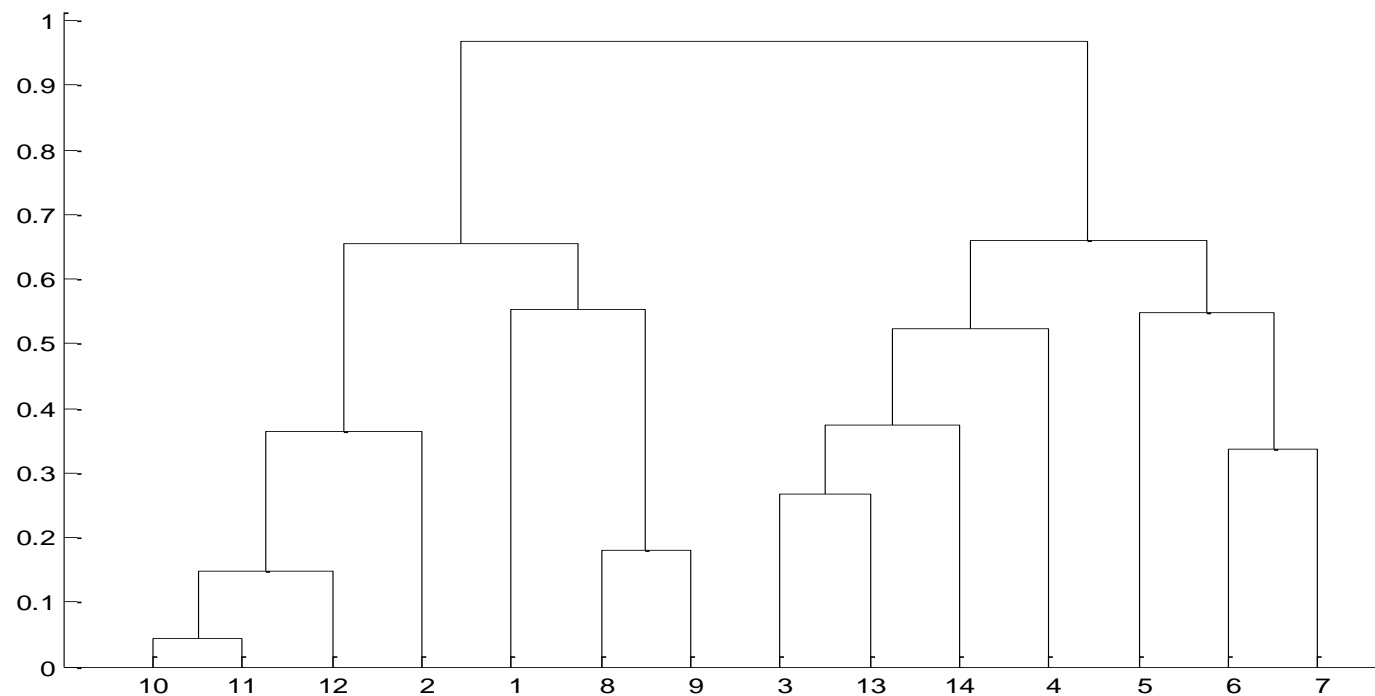


图10.3 成年女子14个部位指标的聚类图

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。

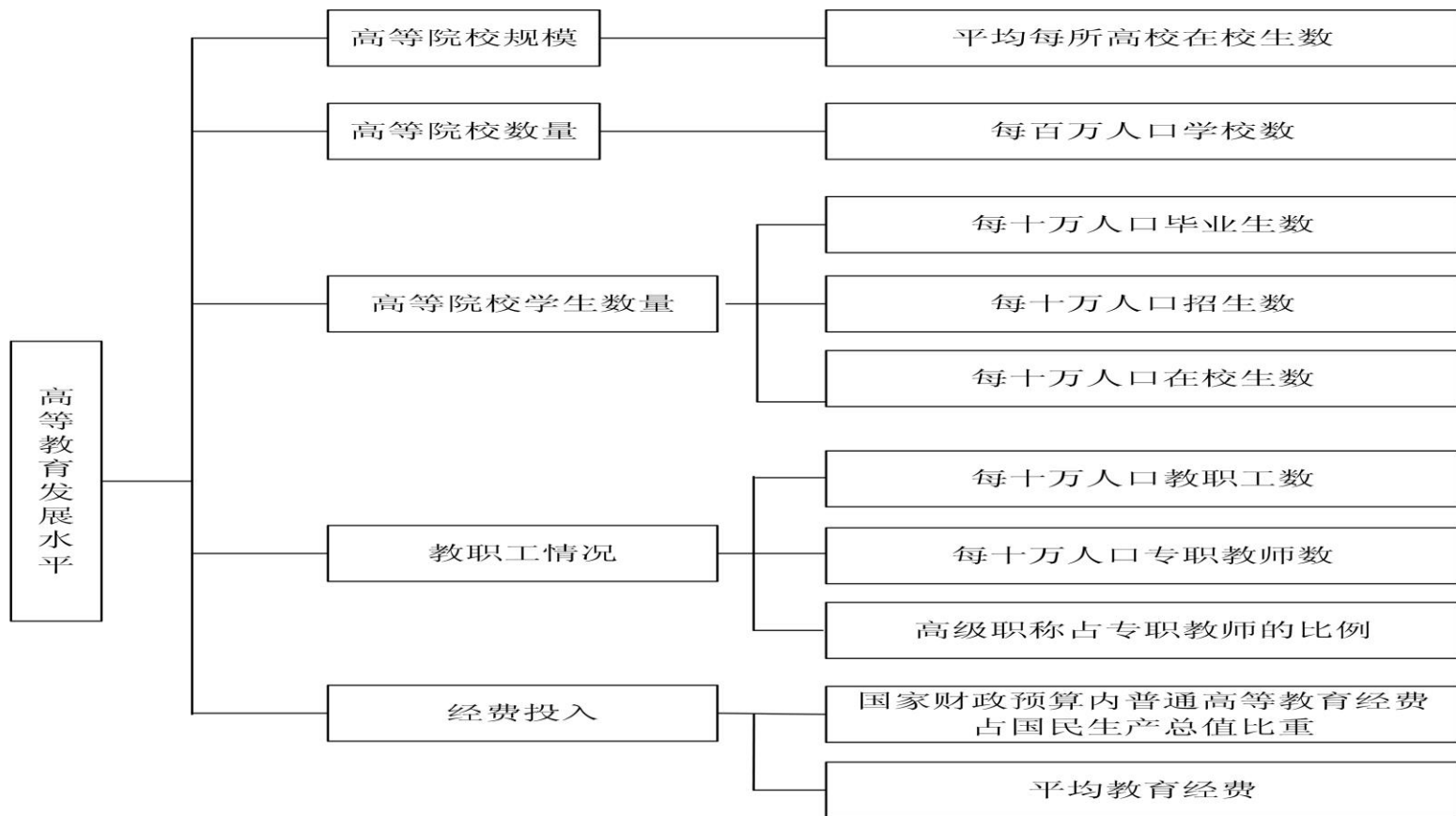


图10.4 高等教育的十项评价指标

(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 为每十万人人口高等院校毕业生数； x_3 为每十万人人口高等院校招生数； x_4 为每十万人人口高等院校在校生数； x_5 为每十万人人口高等院校教职工数； x_6 为每十万人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	1363 1
上海	3.39	234	308	1035	498	161	35.02	3052	.90	1266 5
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。比如每十万人人口高等院校毕业生数、每十万人人口高等院校招生数与每十万人人口高等院校在校生数之间可能存在较强的相关性，每十万人人口高等院校教职工数和每十万人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。

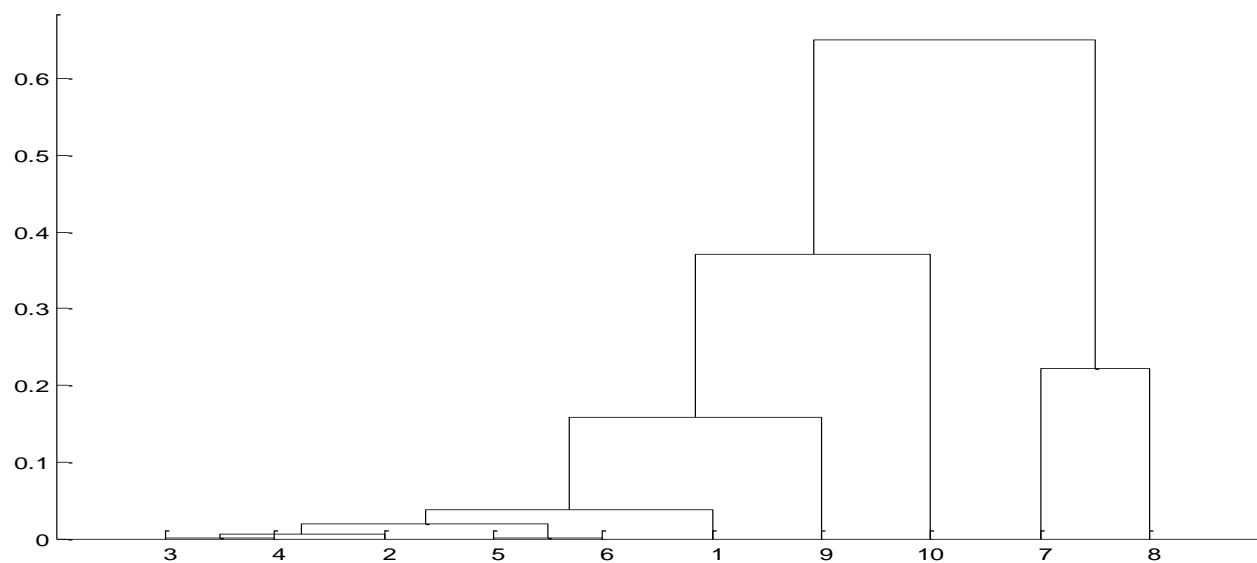


图 10.5 指标聚类树型图

从聚类图 10.5 中可以看出，每十万人人口高等院校招生数、每十万人人口高等院校在校生数、每十万人人口高等院校教职工数、每十万人人口高等院校专职教师数、每十万人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。

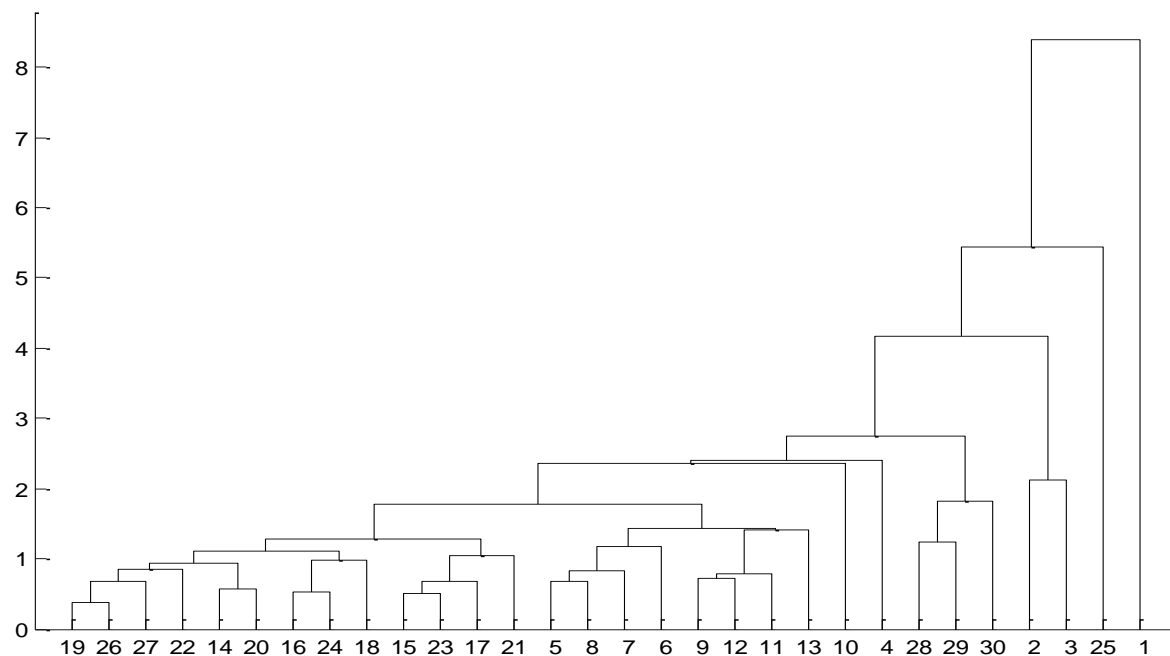


图 10.6 各地区聚类树型图

(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, \dots, x_p 表示 p 门课程, c_1, c_2, \dots, c_p 表示各门课程的权重, 那么加权之和就是

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

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, \dots, 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, \dots, s_n 尽可能的分散, 下面来看它的统计定义。

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

$$\text{Var}(c_1 X_1 + c_2 X_2 + \dots + c_p X_p) \quad (10.15)$$

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

$$c_1^2 + c_2^2 + \dots + c_p^2 = 1 \quad (10.16)$$

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, $i = 1, 2, \dots, p$, 可设

[illegible]

其中对每一个 i , 均有 $c_{i1}^2+c_{i2}^2+\cdots+c_{ip}^2=1$, 且

$[c_{1-1}, c_{1-2}, \dots, c_p]$ 使得 $V a \mathbf{Z}_1(\cdot)$ 的值达到最大；

$[c_{21}, c_{22}, \cdots, c_{2p}]$ 不仅垂直于 $[c_{11}, c_{12}, \cdots, c_{p1}]$, 而且使

VarZ₂ 的值达到最大;

$[c_{31}, c_{32}, \dots, c_{3p}]$ 同时垂直于 $[c_{11}, c_{12}, \dots, c_{1p}]$ 和 $[c_{21}, c_{22}, \dots, c_{2p}]$ ，并使 $\text{Var}(Z_3)$ 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，所以实际中可以先将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, \dots, x_p ，它在第 i 次试验中的取值为

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

将它们写成矩阵形式

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

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 $A^T A$ 的特征向量。因此，剩下的问题仅仅是将 $A^T A$ 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 $\lambda_{r+1}, \lambda_{r+2}, \dots, \lambda_p$ 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, \dots, 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, \dots, z_r ，则它们对原变量 x_i 的贡献值为

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

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

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

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, \dots, 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} ，

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

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

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

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

(2) 计算相关系数矩阵 R

相关系数矩阵 $R = (r_{ij})_{m \times m}$

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

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

$$y_1 = u_{11}\tilde{x}_1 + u_{21}\tilde{x}_2 + \cdots + u_{m1}\tilde{x}_m,$$

$$y_2 = u_{12}\tilde{x}_1 + u_{22}\tilde{x}_2 + \cdots + u_{m2}\tilde{x}_m,$$

$$\cdots \cdots \cdots,$$

$$y_m = u_{1m}\tilde{x}_1 + u_{2m}\tilde{x}_2 + \cdots + u_{mm}\tilde{x}_m,$$

式中 y_1 是第 1 主成分， y_2 是第 2 主成分， \cdots ， 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

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

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

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, \dots, 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, \dots, 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。比如每十万人人口高等院校毕业生数、每十万人人口高等院校招生数与每十万人人口高等院校在校生数之间可能存在较强的相关性，每十万人人口高等院校教职工数和每十万人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。