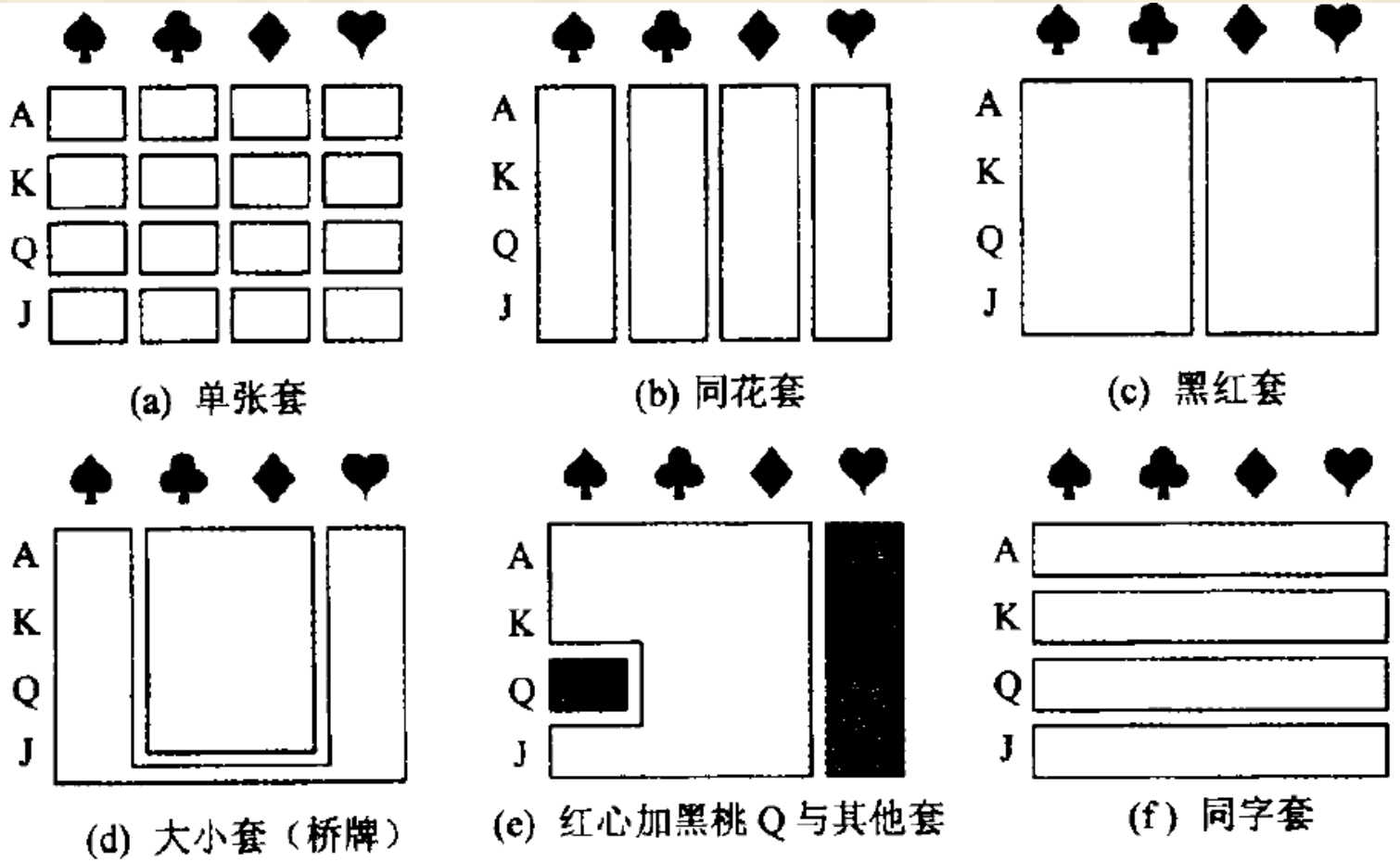
第六章　聚类分析

- §6.1　引言
- §6.2　距离和相似系数
- §6.3　系统聚类法
- §6.4　动态聚类法

§6.1 引言

- 聚类分析：将分类对象分成若干类，相似的归为同一类，不相似的归为不同的类。
- 聚类分析和判别归类有着不同的分类目的，彼此之间既有区别又有联系。
- 聚类分析分为Q型（分类对象为样品）和R型（分类对象为变量）两种。

相似性的不同定义



§6.2 距离和相似系数

- 相似性度量：距离和相似系数。
- 样品之间的距离和相似系数有着各种不同的定义，而这些定义与变量的类型有着非常密切的关系。
- 变量的测量尺度：间隔、有序和名义尺度。
  - 间隔变量：变量用连续的量来表示，如长度、重量、速度、温度等。
  - 有序变量：变量度量时不用明确的数量表示，而是用等级来表示，如某产品分为一等品、二等品、三等品等有次序关系。
  - 名义变量：变量用一些类表示，这些类之间既无等级关系也无数量关系，如性别、职业、产品的型号等。

- 间隔变量也称为定量变量，有序变量和名义变量统称为定性变量或属性变量或分类变量。
- 对于间隔变量，距离常用来度量样品之间的相似性，相似系数常用来度量变量之间的相似性。
- 本章主要讨论具有间隔尺度变量的样品聚类分析方法。

- 一、距离
- 二、相似系数

一、距离

- 设$\boldsymbol{x}=(x_1,x_2,\cdots,x_p)'$ 和$\boldsymbol{y}=(y_1,y_2,\cdots,y_p)'$为两个样品，则所定义的距离一般应满足如下三个条件：
- (i)非负性：$d(\boldsymbol{x},\boldsymbol{y})\geq 0$，$d(\boldsymbol{x},\boldsymbol{y})=0$当且仅当$\boldsymbol{x}=\boldsymbol{y}$；
- (ii)对称性：$d(\boldsymbol{x},\boldsymbol{y})=d(\boldsymbol{y},\boldsymbol{x})$；
- (iii)三角不等式：$d(\boldsymbol{x},\boldsymbol{y})\leq d(\boldsymbol{x},\boldsymbol{z})+d(\boldsymbol{z},\boldsymbol{y})$。

常用的距离

- 1.明考夫斯基（Minkowski）距离
- 2.兰氏（Lance和Williams）距离
- 3.马氏距离

1.明考夫斯基距离

- 明考夫斯基距离（简称明氏距离）：

$$d(\boldsymbol{x},\boldsymbol{y})=\left[\sum_{i=1}^{p}|x_i-y_i|^q\right]^{1/q}$$

这里$q\geq 1$。

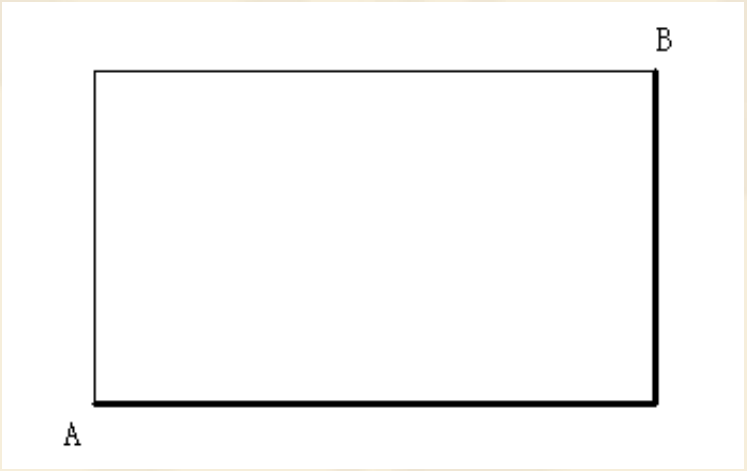
- 明氏距离的三种特殊形式：

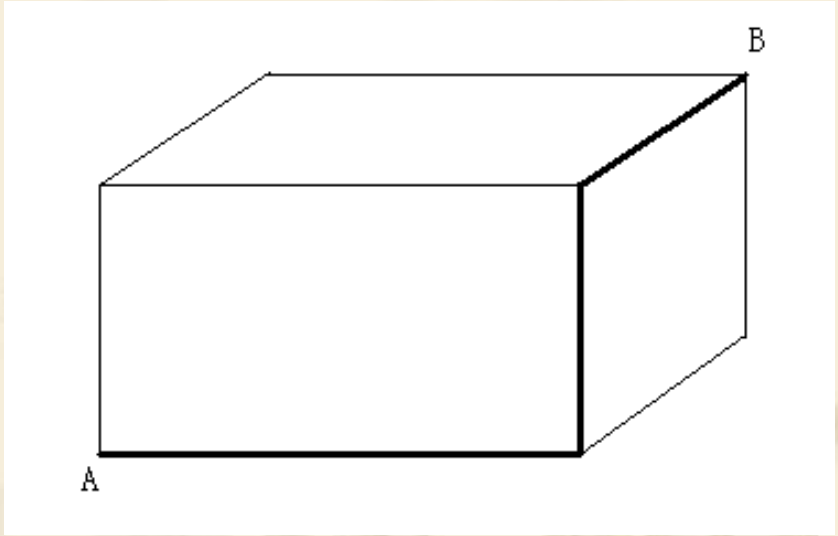
- (i)当$q=1$时，$d(\boldsymbol{x},\boldsymbol{y})=\sum_{i=1}^{p}|x_i-y_i|$，称为绝对值距离，常被形象地称作“城市街区”距离；

- (ii)当$q=2$时，$d(\boldsymbol{x},\boldsymbol{y})=\left[\sum_{i=1}^{p}|x_i-y_i|^2\right]^{1/2}=\sqrt{(\boldsymbol{x}-\boldsymbol{y})'(\boldsymbol{x}-\boldsymbol{y})}$，这是欧氏距离，它是聚类分析中最常用的一个距离；

- (iii)当$q=\infty$时，$d(\boldsymbol{x},\boldsymbol{y})=\max_{1\leq i\leq p}|x_i-y_i|$，称为切比雪夫距离。

绝对值距离图示





对各变量的数据作标准化处理

- 当各变量的单位不同或测量值范围相差很大时，应先对各变量的数据作标准化处理。最常用的标准化处理是，令

$$x_i^* = \frac{x_i - \bar{x}_i}{\sqrt{s_{ii}}}, \quad i = 1, 2, \cdots, p$$

其中$\bar{x}_i$ 和s_{ii}分别为x_i的样本均值和样本方差。

2.兰氏距离

- 当所有的数据皆为正时，可以定义$\boldsymbol{x}$与$\boldsymbol{y}$之间的兰氏距离为

$$d(\boldsymbol{x},\boldsymbol{y})=\sum_{i=1}^{p}\frac{|x_i-y_i|}{x_i+y_i}$$

- 该距离与各变量的单位无关，且适用于高度偏斜或含异常值的数据。

3.马氏距离

- $\boldsymbol{x}$和$\boldsymbol{y}$之间的马氏距离为

$$d(\boldsymbol{x},\boldsymbol{y})=\sqrt{(\boldsymbol{x}-\boldsymbol{y})'\boldsymbol{S}^{-1}(\boldsymbol{x}-\boldsymbol{y})}$$

其中$\boldsymbol{S}$为样本协差阵。

- 聚类过程中的类一直变化着，$\boldsymbol{S}$一般难以确定，除非有关于不同类的先验知识。因此，在实际聚类分析中，马氏距离一般不是理想的距离。

名义尺度变量的一种距离定义

- 例6.2.1　某高校举办一个培训班，从学员的资料中得到这样六个变量：

x_1：性别（男，女）
x_2：外语语种（英语，非英语）
x_3：专业（统计，非统计）
x_4：职业（教师，非教师）
x_5：居住处（校内，校外）
x_6：学位（硕士，学士）

- 现有两名学员：

$\boldsymbol{x}$=(男，英语，统计，非教师，校外，学士)′
$\boldsymbol{y}$=(女，英语，非统计，教师，校外，硕士)′

- 一般地，若记

m_1：配合的变量数

m_2：不配合的变量数

则它们之间的距离可定义为

$$d(\boldsymbol{x},\boldsymbol{y})=\frac{m_2}{m_1+m_2}$$

- 故按此定义，本例中$\boldsymbol{x}$ 与$\boldsymbol{y}$ 之间的距离为2/3。

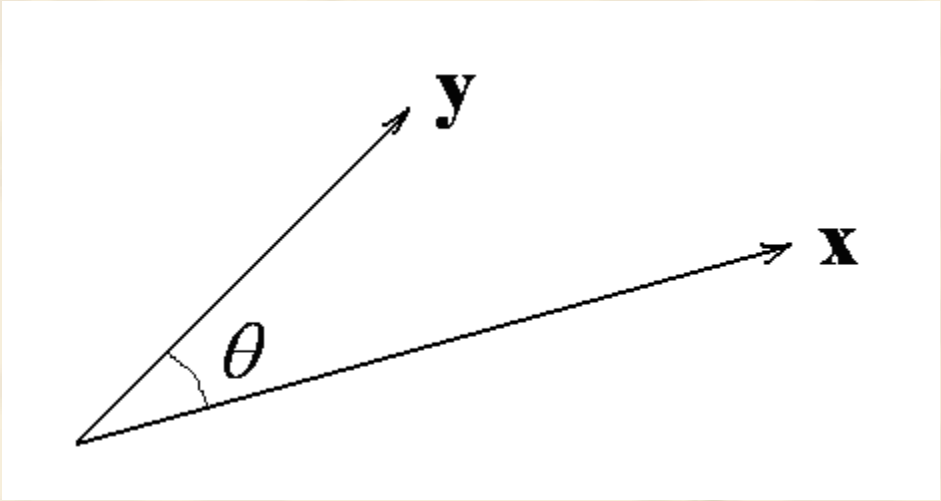
二、相似系数

- 变量之间的相似性度量，在一些应用中要看相似系数的大小，而在另一些应用中要看相似系数绝对值的大小。
- 相似系数（或其绝对值）越大，认为变量之间的相似性程度就越高；反之，则越低。
- 聚类时，比较相似的变量倾向于归为一类，不太相似的变量归属不同的类。

变量间相似系数一般应满足的条件

- (1)$c_{ij}=\pm 1$，当且仅当$x_i=ax_j+b$，$a(\neq 0)$ 和b是常数；
  (2)$|c_{ij}|\leq 1$，对一切i,j；
  (3)$c_{ij}=c_{ji}$，对一切i,j。

两个向量的夹角余弦



$$\cos(\theta) = \frac{\boldsymbol{x}'\boldsymbol{y}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|}$$

1.夹角余弦

- 变量x_i与x_j的夹角余弦定义为

$$c_{ij}(1)=\frac{\sum_{k=1}^{n}x_{ki}x_{kj}}{\left[\left(\sum_{k=1}^{n}x_{ki}^2\right)\left(\sum_{k=1}^{n}x_{kj}^2\right)\right]^{1/2}}$$

它是R^n中变量x_i的观测向量$(x_{1i},x_{2i},\cdots,x_{ni})'$与变量$x_j$的观测向量$(x_{1j},x_{2j},\cdots,x_{nj})'$之间夹角$\theta_{ij}$的余弦函数，即$c_{ij}(1)=\cos\theta_{ij}$。

2.相关系数

- 变量x_i与x_j的相关系数为

$$c_{ij}(2) = r_{ij} = \frac{\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)}{\left\{\left[\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)^2\right]\left[\sum_{k=1}^{n}(x_{kj}-\bar{x}_j)^2\right]\right\}^{1/2}}$$

- 如果变量x_i与x_j是已标准化了的，则它们间的夹角余弦就是相关系数。

- 相似系数除常用来度量变量之间的相似性外有时也用来度量样品之间的相似性，同样，距离有时也用来度量变量之间的相似性。
- 由距离来构造相似系数总是可能的，如令

$$c_{ij}=\frac{1}{1+d_{ij}}$$

这里d_{ij}为第i个样品与第j个样品的距离，c_{ij}可作为相似系数，用来度量样品之间的相关性。

- 距离必须满足定义距离的三个条件，所以不是总能由相似系数构造。高尔（Gower）证明，当相似系数矩阵(c_{ij})为非负定时，如令

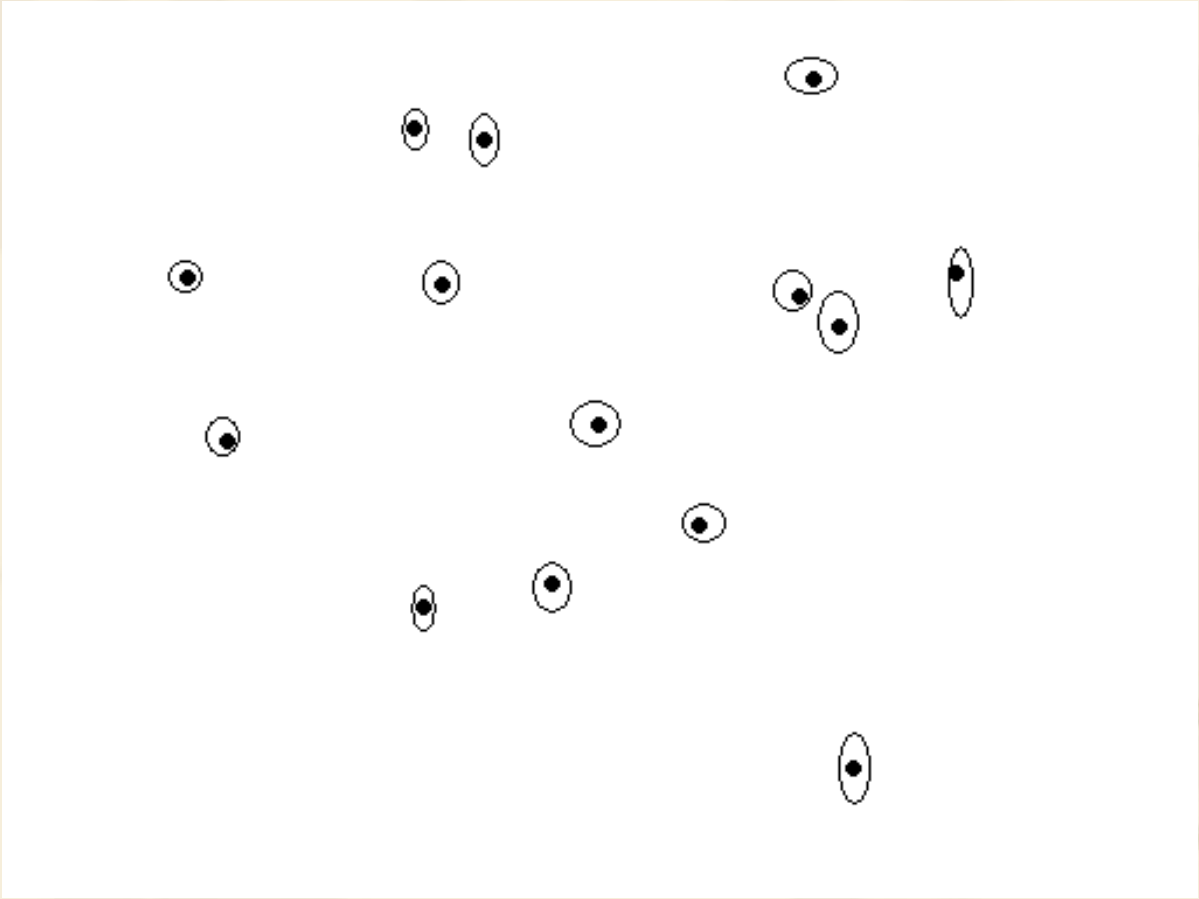
$$d_{ij}=\sqrt{2\left(1-c_{ij}\right)}$$

则d_{ij}满足距离定义的三个条件。

§6.3 系统聚类法

- 系统聚类法（或层次聚类法）是通过一系列相继的合并或相继的分割来进行的，分为聚集的和分割的两种，适用于样品数目n不是非常大的情形。
- 聚集系统法的基本思想是：开始时将n个样品各自作为一类，并规定样品之间的距离和类与类之间的距离，然后将距离最近的两类合并成一个新类，计算新类与其他类的距离；重复进行两个最近类的合并，每次减少一类，直至所有的样品合并为一类。

一开始每个样品各自作为一类



- 分割系统法的聚类步骤与聚集系统法正相反。由n个样品组成一类开始，按某种最优准则将它分割成两个尽可能远离的子类，再用同样准则将每一子类进一步地分割成两类，从中选一个分割最优的子类，这样类数将由两类增加到三类。如此下去，直至所有n个样品各自为一类或采用某种停止规则。
- 聚集系统法最为常用，本节着重介绍其中常用的六种方法并略提另两种方法，所有这些聚类方法的区别在于类与类之间距离的定义不同。

§6.3 系统聚类法

- 一、最短距离法
- 二、最长距离法
- 三、类平均法
- 四、重心法
- *五、中间距离法
- 六、离差平方和法（Ward方法）
- *七、系统聚类法的统一
- 八、系统聚类法的性质
- 九、使用图形作聚类及对效果的评估
- 十、对变量的聚类
- 十一、类的个数

一、最短距离法

- 定义类与类之间的距离为两类最近样品间的距离，即

$$D_{KL} = \min_{i \in G_K, j \in G_L} d_{ij}$$

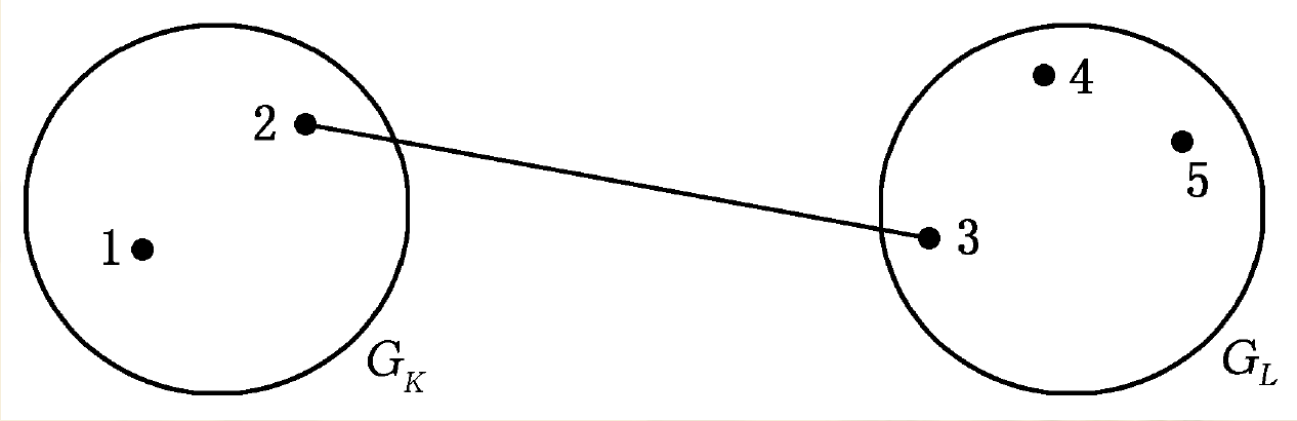


图6. 3. 1　最短距离法：$D_{KL}=d_{23}$

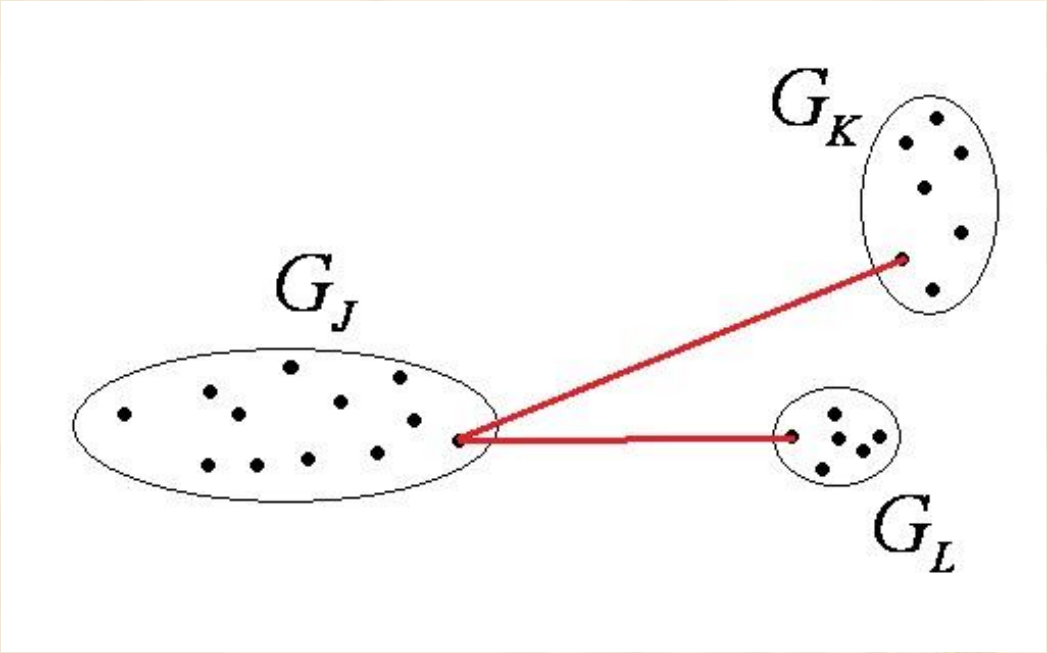
最短距离法的聚类步骤

- (1)规定样品之间的距离，计算n个样品的距离矩阵$\boldsymbol{D}_{(0)}$，它是一个对称矩阵。
- (2)选择$\boldsymbol{D}_{(0)}$中的最小元素，设为D_{KL}，则将G_K和G_L合并成一个新类，记为G_M，即$G_M = G_K \cup G_L$。
- (3)计算新类G_M与任一类G_J之间距离的递推公式为

$$D_{MJ} = \min_{i \in G_M, j \in G_J} d_{ij} = \min\left\{ \min_{i \in G_K, j \in G_J} d_{ij}, \min_{i \in G_L, j \in G_J} d_{ij} \right\}$$

$$= \min\left\{ D_{KJ}, D_{LJ} \right\}$$

递推公式的图示理解



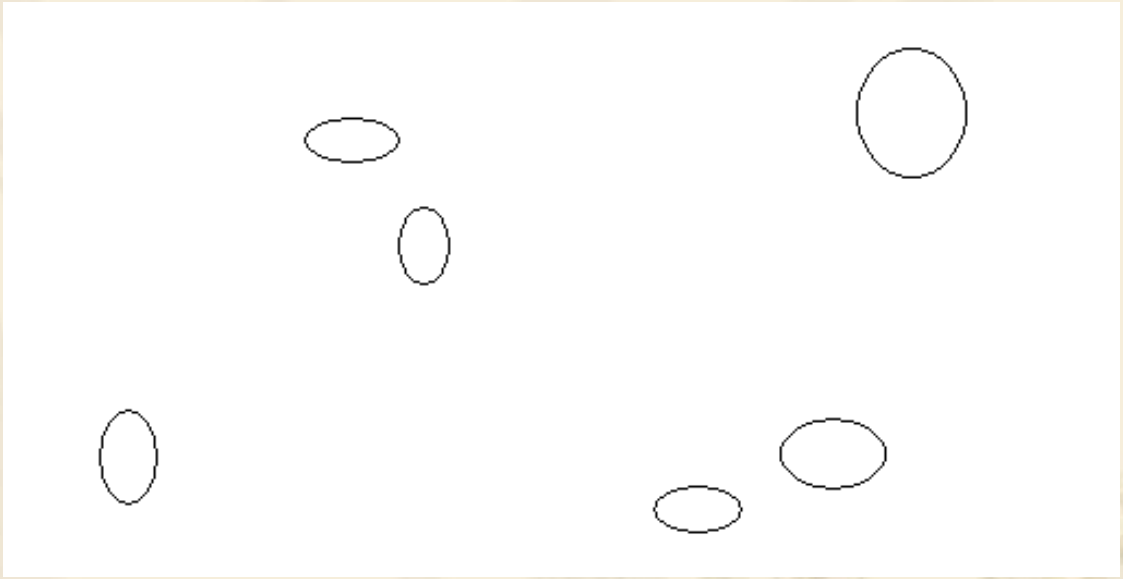
最短距离法的聚类步骤（续）

在$\boldsymbol{D}_{(0)}$中，G_K和G_L所在的行和列合并成一个新行新列，对应G_M，该行列上的新距离值由上述递推公式求得，其余行列上的距离值不变，这样就得到新的距离矩阵，记作$\boldsymbol{D}_{(1)}$。

- (4)对$\boldsymbol{D}_{(1)}$重复上述对$\boldsymbol{D}_{(0)}$的两步得$\boldsymbol{D}_{(2)}$，如此下去直至所有元素合并成一类为止。

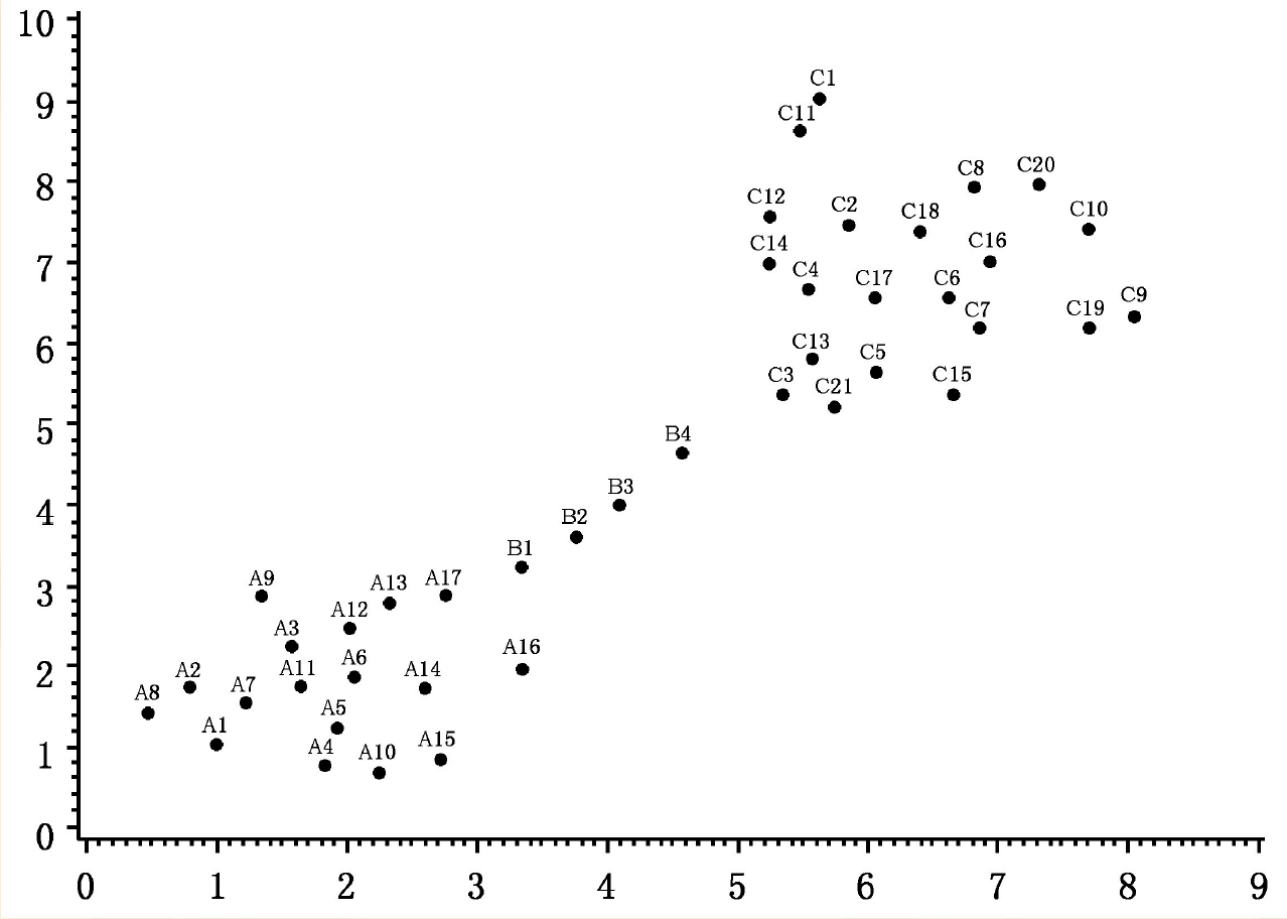
- 如果某一步$\boldsymbol{D}_{(m)}$中最小的元素不止一个，则称此现象为结，对应这些最小元素的类可以任选一对合并或同时合并。最短距离法最容易产生结，且有一种挑选长链状聚类的倾向，称为链接倾向。

结的图示：



- 最短距离法不适合对分离得很差的群体进行聚类。

一个最短距离法产生链接的例子
（例6.3.4）



- 例6.3.1　设有五个样品，每个只测量了一个指标，分别是1，2，6，8，11，试用最短距离法将它们分类。

- 记$G_1=\{1\}$，$G_2=\{2\}$，$G_3=\{6\}$，$G_4=\{8\}$，$G_5=\{11\}$，样品间采用绝对值距离。

表6. 3. 1　　$D_{(0)}$

	G_1	G_2	G_3	G_4	G_5
G_1	0				
G_2	1	0			
G_3	5	4	0		
G_4	7	6	2	0	
G_5	10	9	5	3	0

表6. 3. 2 $D_{(1)}$

	G_6	G_3	G_4	G_5
G_6	0			
G_3	4	0		
G_4	6	2	0	
G_5	9	5	3	0

其中$G_6 = G_1 \cup G_2$

表6. 3. 3 $D_{(2)}$

	G_6	G_7	G_5
G_6	0		
G_7	4	0	
G_5	9	3	0

其中$G_7 = G_3 \cup G_4$

表6. 3. 4　　$D_{(3)}$

	G_6	G_8
G_6	0	
G_8	4	0

其中$G_6 = G_1 \cup G_2$

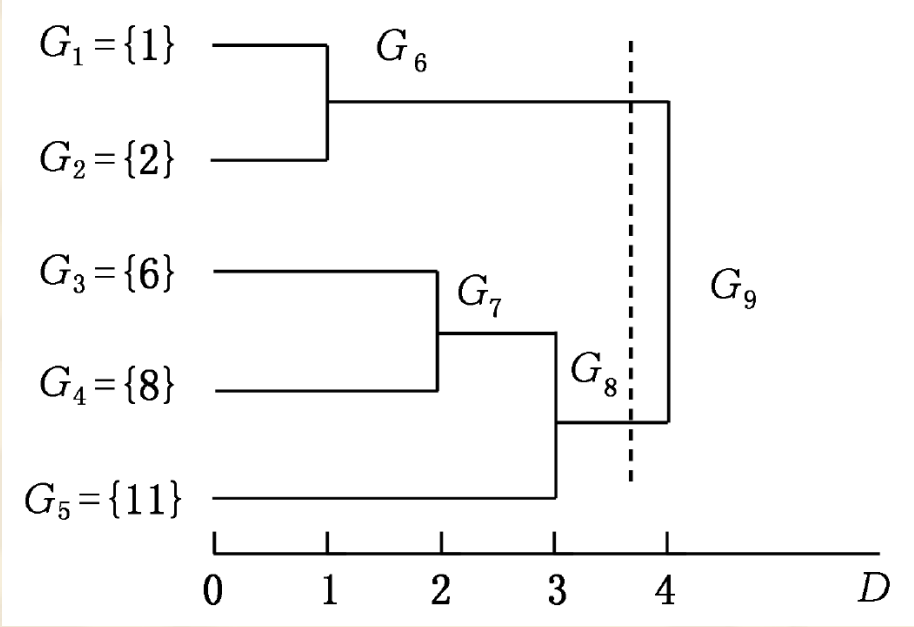


图6. 3. 2　最短距离法树形图

二、最长距离法

- 类与类之间的距离定义为两类最远样品间的距离，即

$$D_{KL} = \max_{i \in G_K, j \in G_L} d_{ij}$$

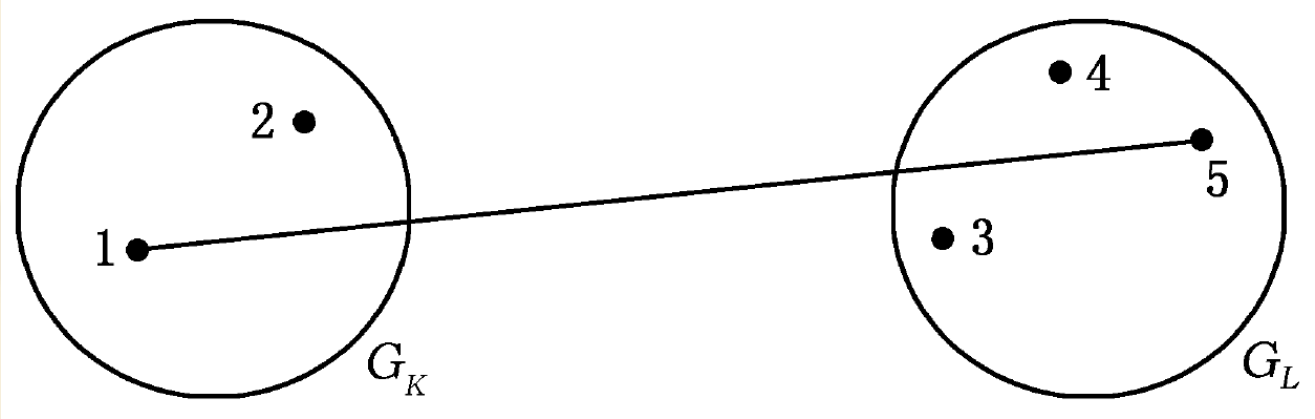
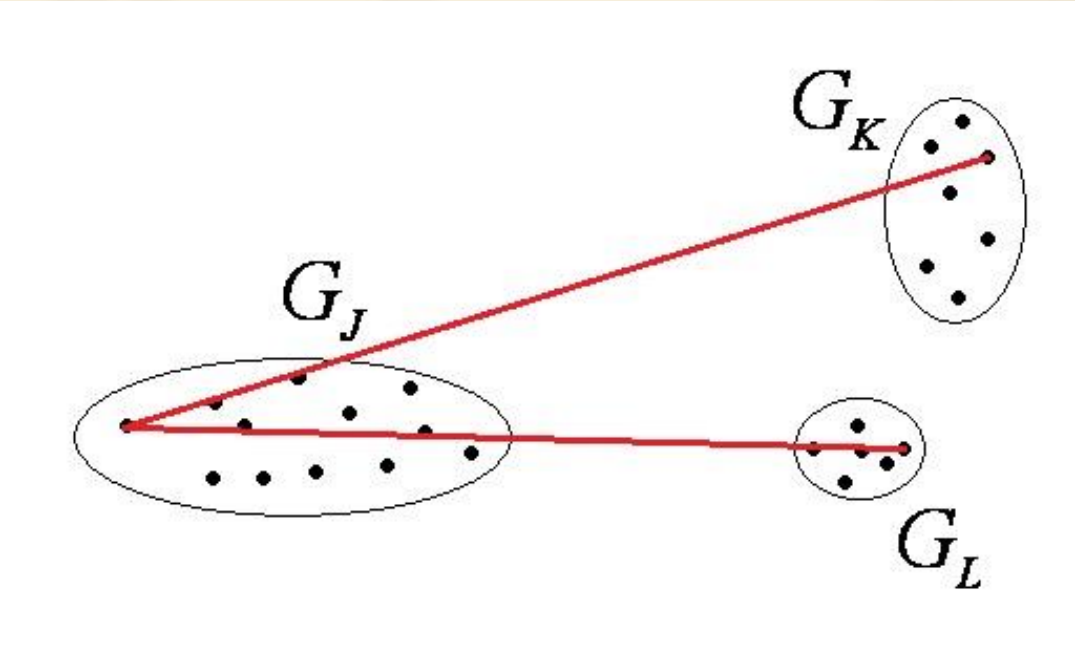


图6. 3. 3　最长距离法：$D_{KL}=d_{15}$

- 最长距离法与最短距离法的并类步骤完全相同，只是类间距离的递推公式有所不同。
- 递推公式：

$$D_{MJ} = \max\{D_{KJ}, D_{LJ}\}$$



- 对例6.3.1采用最长距离法。

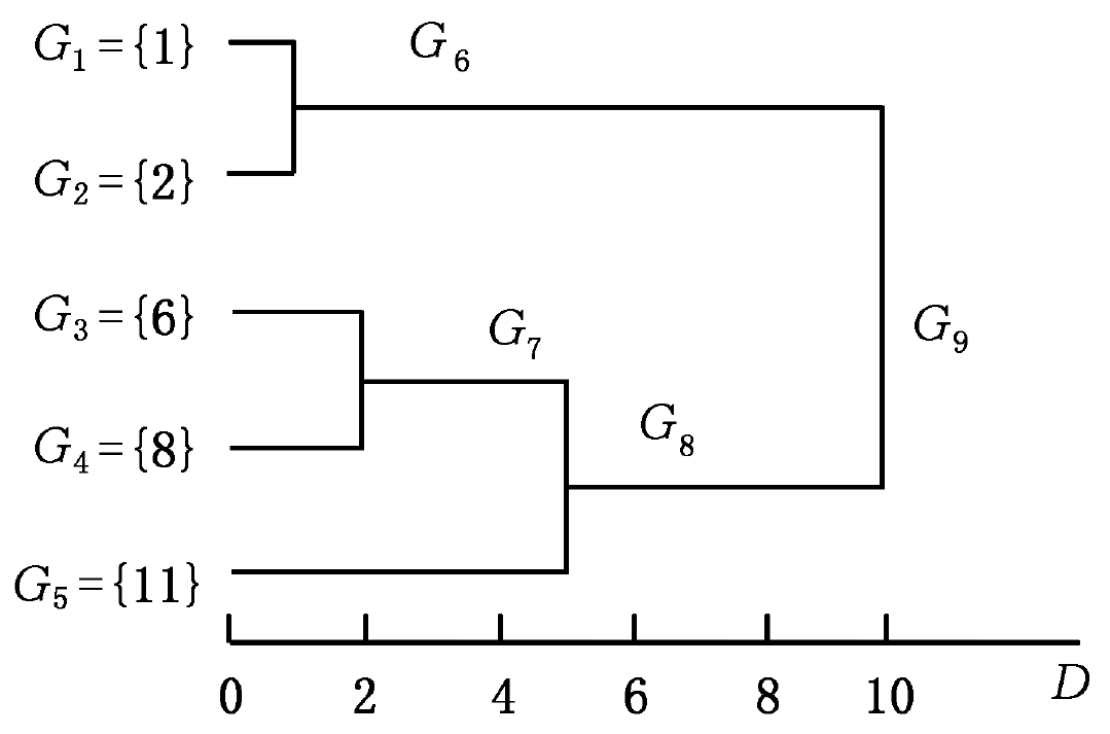
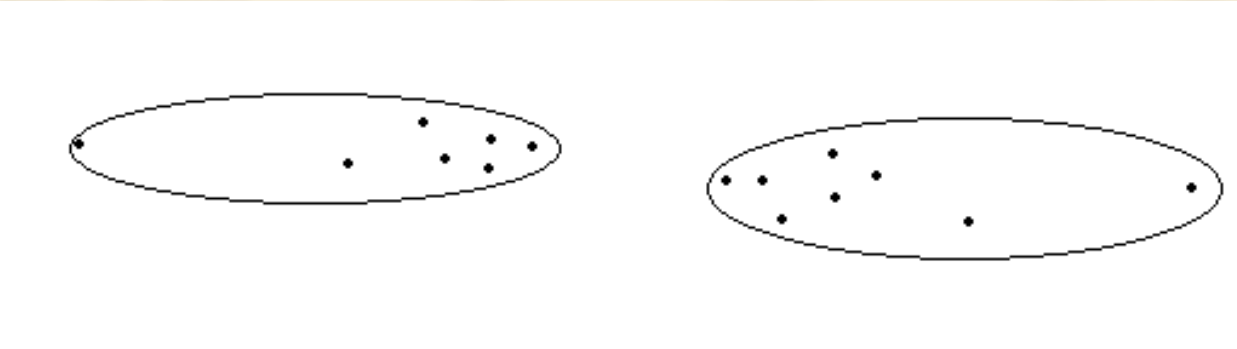


图6. 3. 4　最长距离法树形图

异常值的影响

- 最长距离法容易被异常值严重地扭曲。



三、类平均法

- 有两种定义。
- 定义1：类G_K和G_L之间的距离定义为

$$D_{KL}=\frac{1}{n_K n_L}\sum_{i\in G_K,\,j\in G_L} d_{ij}$$

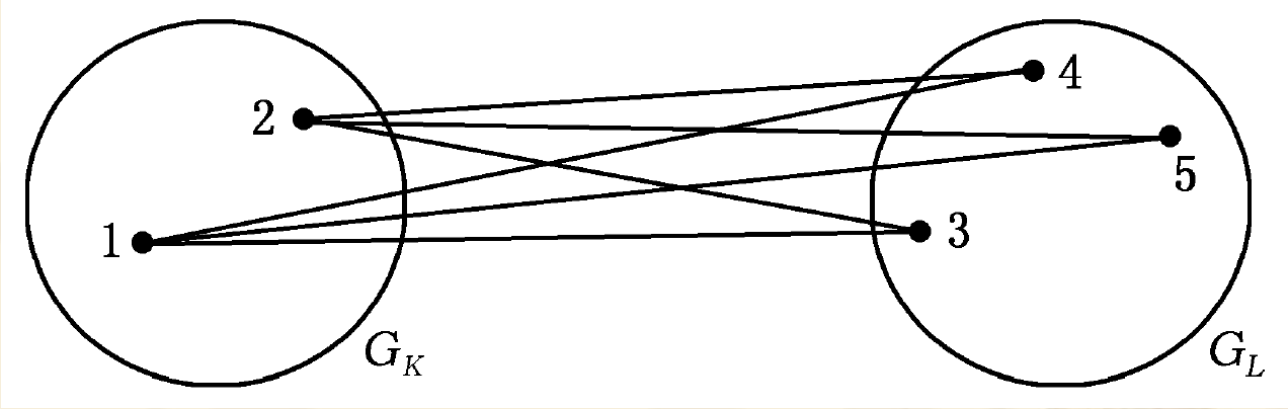


图6.3.5　类平均法

- 递推公式：

$$D_{MJ} = \frac{1}{n_M n_J} \sum_{i \in G_M, j \in G_J} d_{ij} = \frac{1}{n_M n_J} \left(\sum_{i \in G_K, j \in G_J} d_{ij} + \sum_{i \in G_L, j \in G_J} d_{ij} \right)$$

$$= \frac{n_K}{n_M} D_{KJ} + \frac{n_L}{n_M} D_{LJ}$$

- 定义2：类G_K和G_L之间的平方距离定义为

$$D_{KL}^2 = \frac{1}{n_K n_L} \sum_{i \in G_K, j \in G_L} d_{ij}^2$$

- 递推公式：

$$D_{MJ}^2 = \frac{n_K}{n_M} D_{KJ}^2 + \frac{n_L}{n_M} D_{LJ}^2$$

- 类平均法较好地利用了所有样品之间的信息，在很多情况下它被认为是一种比较好的系统聚类法。

- 例6.3.2　在例6.3.1中采用（使用平方距离的）类平均法进行聚类。一开始将$\boldsymbol{D}_{(0)}$的每个元素都平方，并记作$\boldsymbol{D}_{(0)}^2$。

表6. 3. 5　　$\boldsymbol{D}_{(0)}^2$

	G_1	G_2	G_3	G_4	G_5
G_1	0				
G_2	1	0			
G_3	25	16	0		
G_4	49	36	4	0	
G_5	100	81	25	9	0

表6. 3. 6 $D^2_{(1)}$

	G_6	G_3	G_4	G_5
G_6	0			
G_3	20.5	0		
G_4	42.5	4	0	
G_5	90.5	25	9	0

表6. 3. 7 $D^2_{(2)}$

	G_6	G_7	G_5
G_6	0		
G_7	31.5	0	
G_5	90.5	17	0

表6. 3. 8　　$D^2_{(3)}$

	G_6	G_8
G_6	0	
G_8	51.17	0

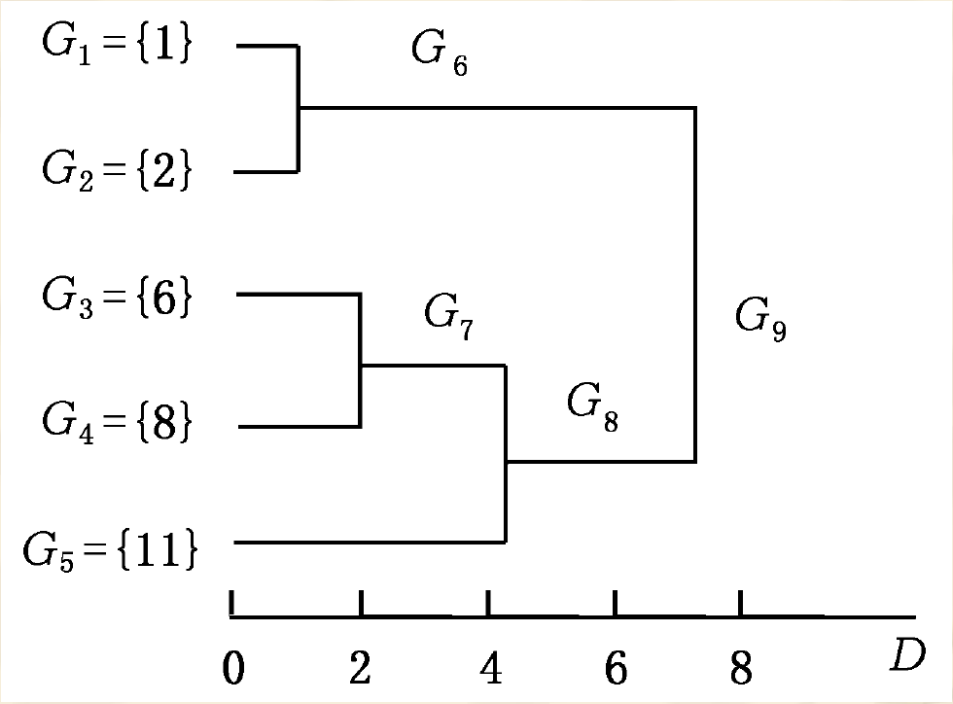


图6. 3. 6　类平均法树形图

四、重心法

- 设类G_K和G_L的重心（均值）分别为$\bar{\boldsymbol{x}}_K$和$\bar{\boldsymbol{x}}_L$，则G_K与G_L之间的平方距离定义为

$$D_{KL}^2 = d_{\bar{x}_K \bar{x}_L}^2 = (\bar{\boldsymbol{x}}_K - \bar{\boldsymbol{x}}_L)'(\bar{\boldsymbol{x}}_K - \bar{\boldsymbol{x}}_L)$$

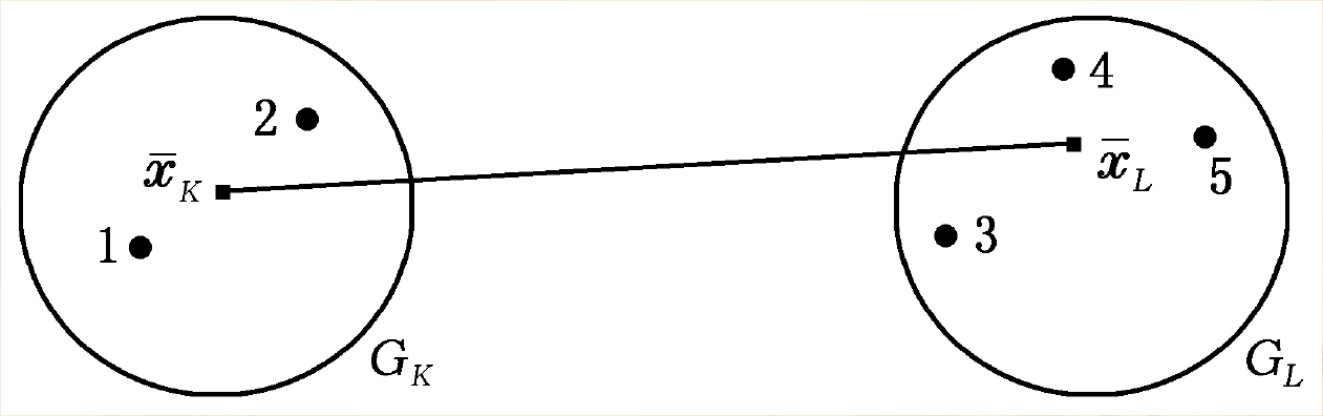


图6.3.7　重心法

- $G_M = G_K \cup G_L$的重心是

$$\bar{x}_M = \frac{n_K \bar{x}_K + n_L \bar{x}_L}{n_M}$$

其中$n_M = n_K + n_L$为G_M的样品个数。

- 递推公式：

$$D_{MJ}^2 = \frac{n_K}{n_M} D_{KJ}^2 + \frac{n_L}{n_M} D_{LJ}^2 - \frac{n_K n_L}{n_M^2} D_{KL}^2$$

- 与其他系统聚类法相比，重心法在处理异常值方面更稳健，但是在别的方面一般不如类平均法或离差平方和法的效果好。

*五、中间距离法

- 设$G_M = G_K \cup G_L$，对于任一类G_J，考虑由D_{KJ}，D_{LJ}和D_{KL}为边长组成的三角形，取D_{KL}边的中线作为D_{MJ}。D_{MJ}的计算公式为

$$D_{MJ}^2 = \frac{1}{2}D_{KJ}^2 + \frac{1}{2}D_{LJ}^2 - \frac{1}{4}D_{KL}^2$$

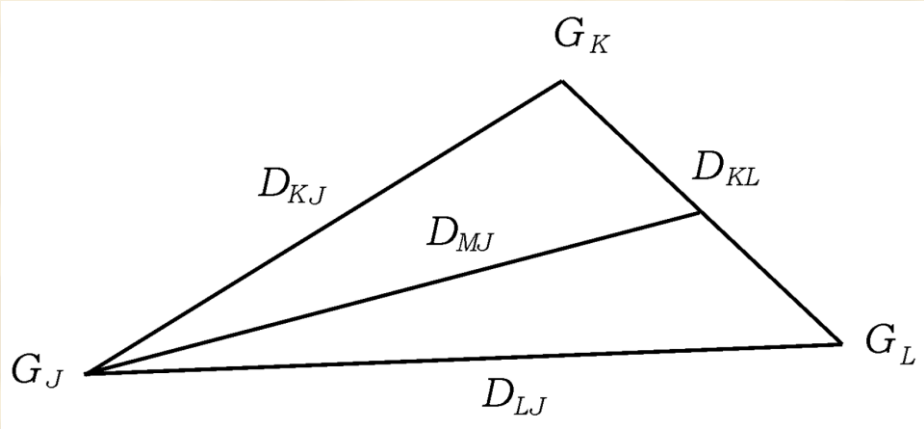


图6. 3. 8　中间距离法的几何表示

六、离差平方和法(Ward方法)

- （类内）离差平方和：类中各样品到类重心（均值）的平方欧氏距离之和。
- 设类G_K和G_L合并成新类G_M，则G_K, G_L和G_M的离差平方和分别是

$$W_K = \sum_{i \in G_K} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_K)' (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_K)$$

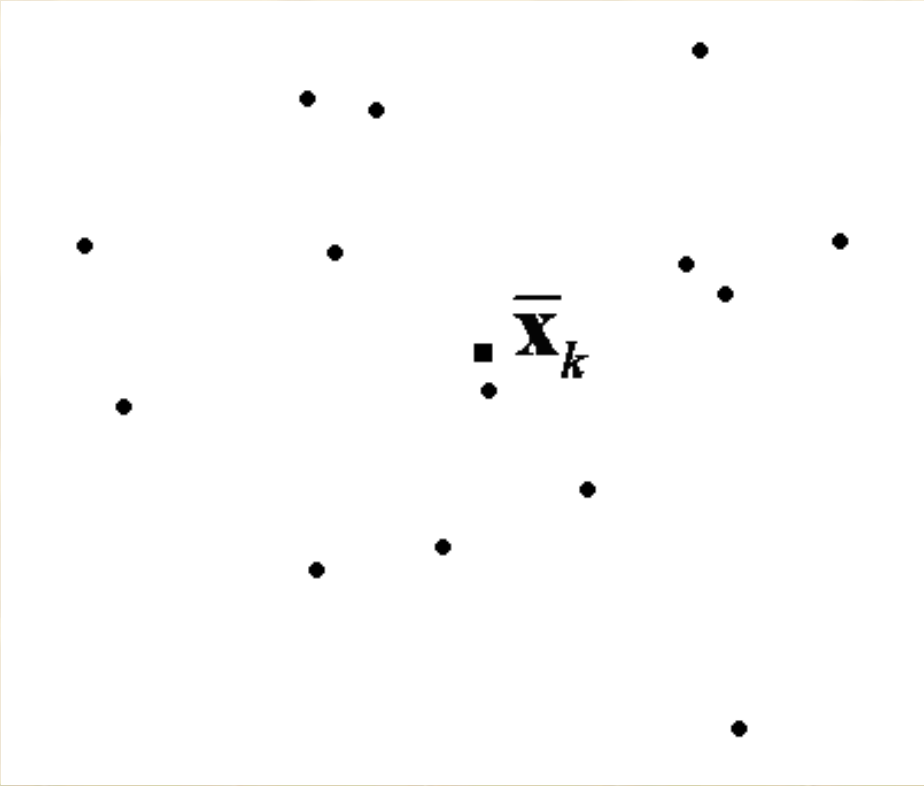
$$W_L = \sum_{i \in G_L} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_L)' (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_L)$$

$$W_M = \sum_{i \in G_M} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_M)' (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_M)$$

对固定的类内样品数，它们反映了各自类内样品的分散程度。

类内离差平方和的几何解释

- 类内离差平方和W_K是类G_K内各点到类重心点$\bar{x}_k$的直线距离之平方和。



- 定义G_K和G_L之间的平方距离为

$$D_{KL}^2 = W_M - W_K - W_L$$

- D_{KL}^2也可表达为

$$D_{KL}^2 = \frac{n_K n_L}{n_M}(\bar{x}_K - \bar{x}_L)'(\bar{x}_K - \bar{x}_L)$$

  - $\frac{n_K n_L}{n_M} = \frac{n_K n_L}{n_K + n_L} = \frac{1}{1/n_L + 1/n_K}$，当$n_K = n_L$时，$\frac{n_K n_L}{n_M} = \frac{n_K}{2}$

- 离差平方和法使得两个大的类倾向于有较大的距离，因而不易合并；相反，两个小的类却因倾向于有较小的距离而易于合并。这往往符合我们对聚类的实际要求。

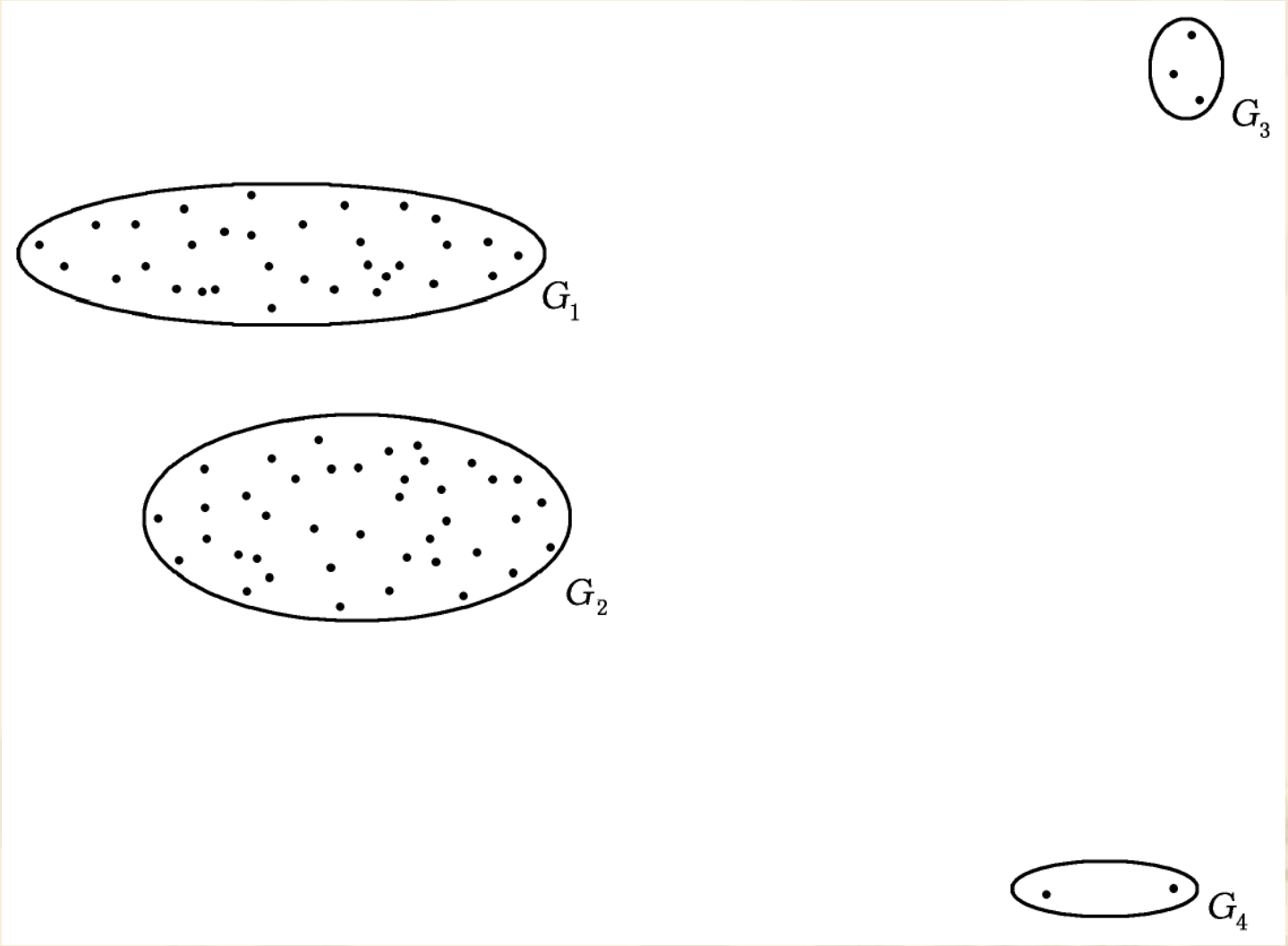


图6.3.9　离差平方和法与重心法的聚类比较

- 递推公式：

$$D_{MJ}^2 = \frac{n_J + n_K}{n_J + n_M} D_{KJ}^2 + \frac{n_J + n_L}{n_J + n_M} D_{LJ}^2 - \frac{n_J}{n_J + n_M} D_{KL}^2$$

- 对例6.3.1采用离差平方和法进行聚类。

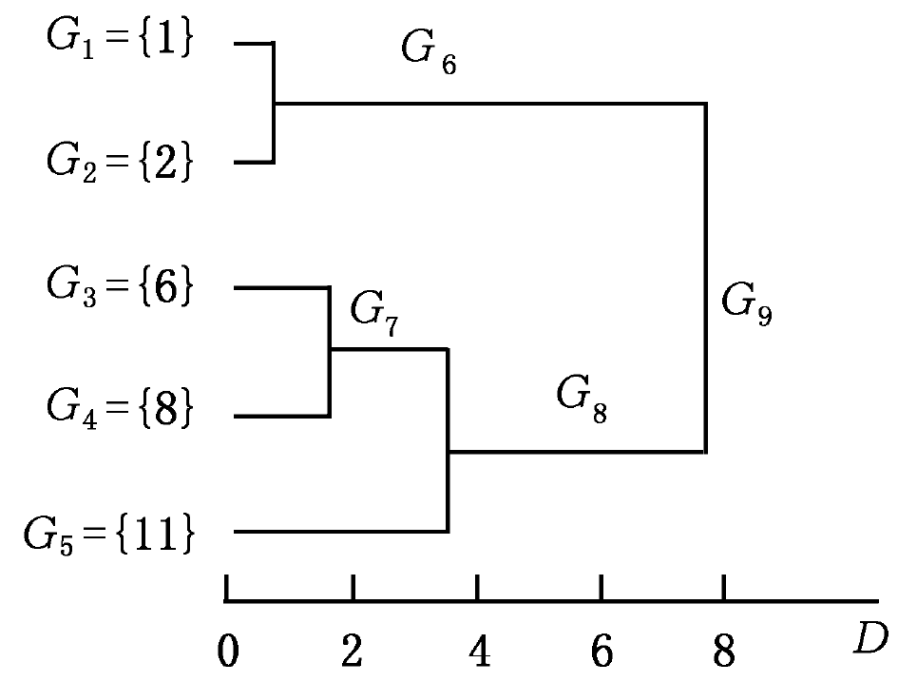


图6. 3. 10　离差平方和法树形图

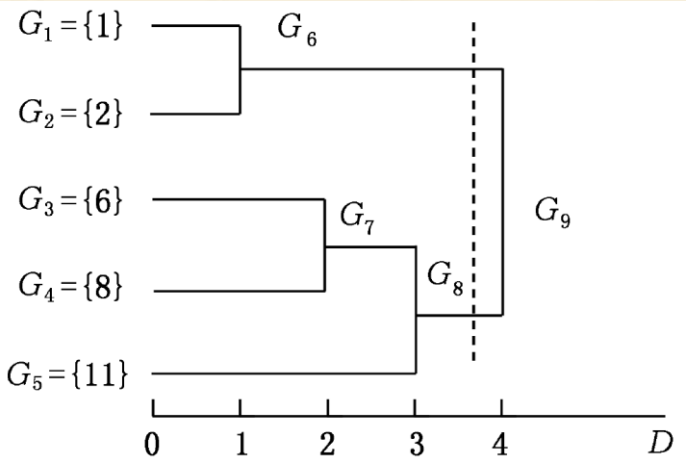


图 6.3.2　最短距离法树形图

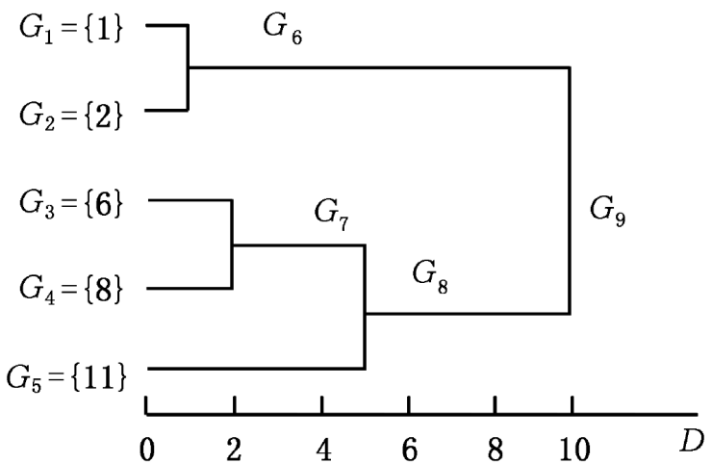


图 6.3.4　最长距离法树形图

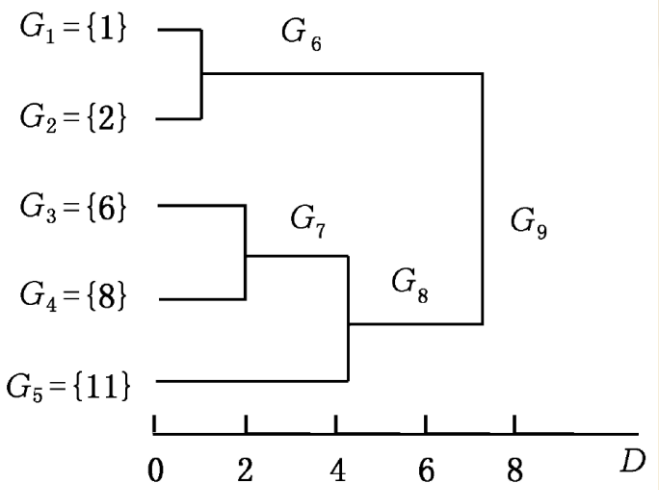


图 6.3.6　类平均法树形图

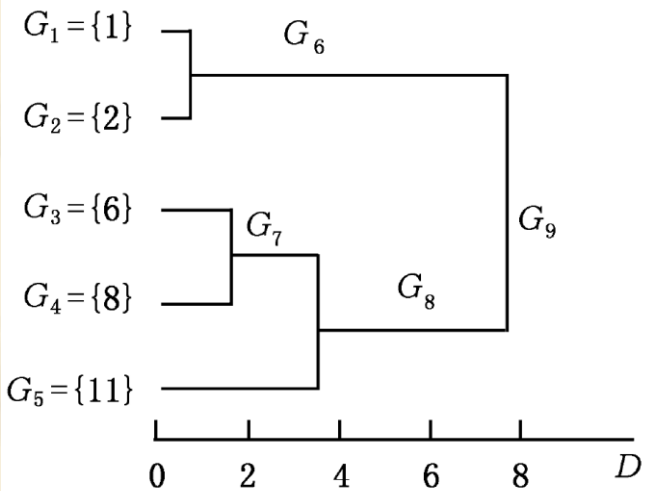


图 6.3.10　离差平方和法树形图

- 例6.3.3　表6.3.9列出了1999年全国31个省、直辖市和自治区的城镇居民家庭平均每人全年消费性支出的八个主要变量数据。这八个变量是

  x_1：食品　　x_5：交通和通讯

  x_2：衣着　　x_6：娱乐教育文化服务

  x_3：家庭设备用品及服务　　x_7：居住

  x_4：医疗保健　　x_8：杂项商品和服务

- 分别用最短距离法、重心法和Ward方法对各地区作聚类分析。为同等地对待每一变量，在作聚类前，先对各变量作标准化变换。

表6. 3. 9　　消费性支出数据　　单位：元

地区	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
北京	2959.19	730.79	749.41	513.34	467.87	1141.82	478.42	457.64
天津	2459.77	495.47	697.33	302.87	284.19	735.97	570.84	305.08
河北	1495.63	515.9	362.37	285.32	272.95	540.58	364.91	188.63
山西	1406.33	477.77	290.15	208.57	201.5	414.72	281.84	212.1
内蒙古	1303.97	524.29	254.83	192.17	249.81	463.09	287.87	192.96
辽宁	1730.84	553.9	246.91	279.81	239.18	445.2	330.24	163.86
吉林	1561.86	492.42	200.49	218.36	220.69	459.62	360.48	147.76
黑龙江	1410.11	510.71	211.88	277.11	224.65	376.82	317.61	152.85
上海	3712.31	550.74	893.37	346.93	527	1034.98	720.33	462.03
江苏	2207.58	449.37	572.4	211.92	302.09	585.23	429.77	252.54
浙江	2629.16	557.32	689.73	435.69	514.66	795.87	575.76	323.36
安徽	1844.78	430.29	271.28	126.33	250.56	513.18	314	151.39
福建	2709.46	428.11	334.12	160.77	405.14	461.67	535.13	232.29
江西	1563.78	303.65	233.81	107.9	209.7	393.99	509.39	160.12
山东	1675.75	613.32	550.71	219.79	272.59	599.43	371.62	211.84

河南	1427.65	431.79	288.55	208.14	217	337.76	421.31	165.32
湖北	1783.43	511.88	282.84	201.01	237.6	617.74	523.52	182.52
湖南	1942.23	512.27	401.39	206.06	321.29	697.22	492.6	226.45
广东	3055.17	353.23	564.56	356.27	811.88	873.06	1082.82	420.81
广西	2033.87	300.82	338.65	157.78	329.06	621.74	587.02	218.27
海南	2057.86	186.44	202.72	171.79	329.65	477.17	312.93	279.19
重庆	2303.29	589.99	516.21	236.55	403.92	730.05	438.41	225.8
四川	1974.28	507.76	344.79	203.21	240.24	575.1	430.36	223.46
贵州	1673.82	437.75	461.61	153.32	254.66	445.59	346.11	191.48
云南	2194.25	537.01	369.07	249.54	290.84	561.91	407.7	330.95
西藏	2646.61	839.7	204.44	209.11	379.3	371.04	269.59	389.33
陕西	1472.95	390.89	447.95	259.51	230.61	490.9	469.1	191.34
甘肃	1525.57	472.98	328.9	219.86	206.65	449.69	249.66	228.19
青海	1654.69	437.77	258.78	303	244.93	479.53	288.56	236.51
宁夏	1375.46	480.89	273.84	317.32	251.08	424.75	228.73	195.93
新疆	1608.82	536.05	432.46	235.82	250.28	541.3	344.85	214.4

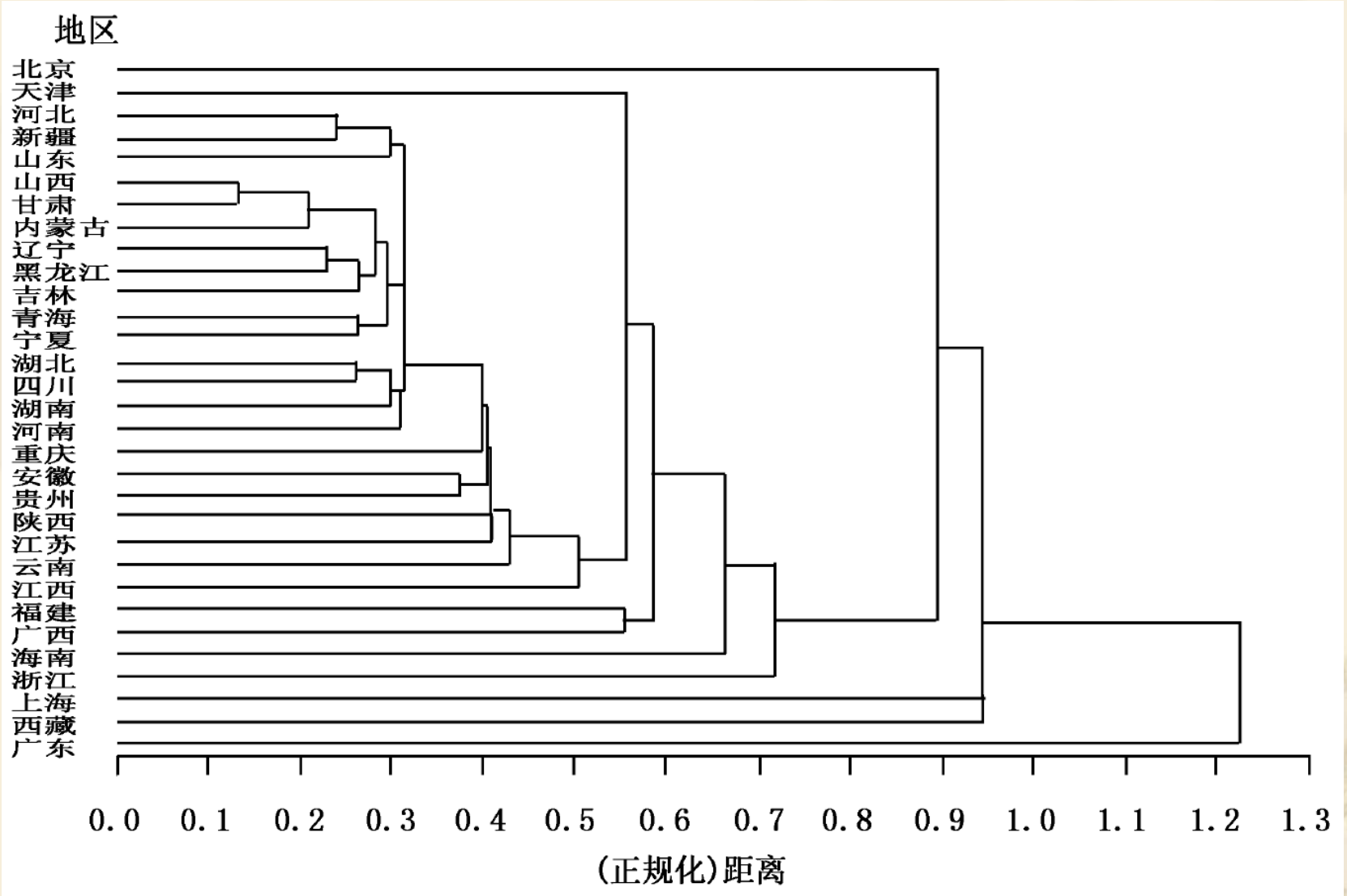

地区
北京
天津
河北
新疆
山东
山西
甘肃
内蒙古
辽宁
黑龙江
吉林
青海
宁夏
湖北
四川
湖南
河南
重庆
安徽
贵州
陕西
江苏
云南
江西
福建
广西
海南
浙江
上海
西藏
广东
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3
(正规化)距离

图6.3.11　最短距离法

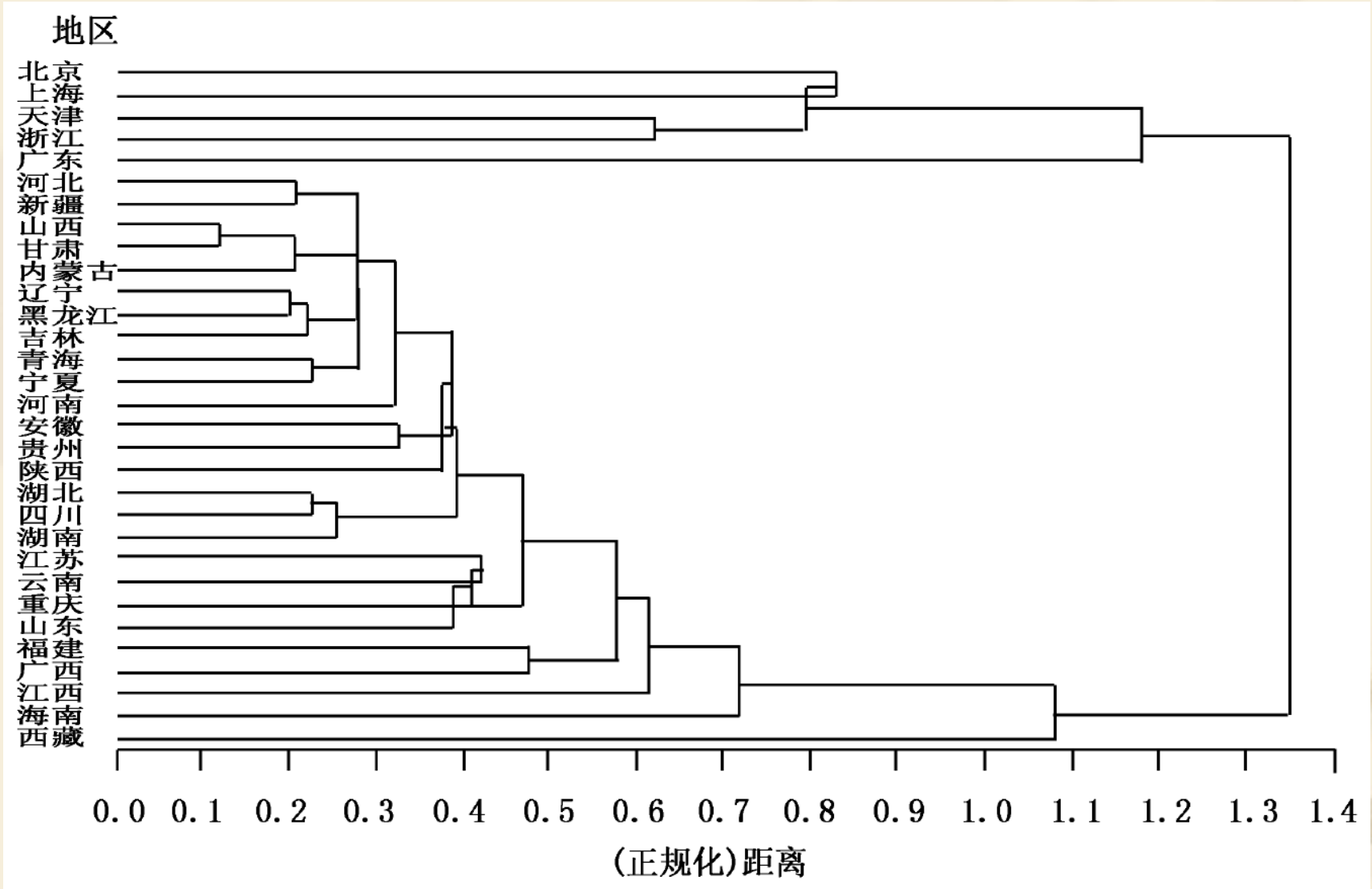

地区
北京
上海
天津
浙江
广东
河北
新疆
山西
甘肃
内蒙古
辽宁
黑龙江
吉林
青海
宁夏
河南
安徽
贵州
陕西
湖北
四川
湖南
江苏
云南
重庆
山东
福建
广西
江西
海南
西藏
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4
(正规化)距离

图6.3.12　重心法

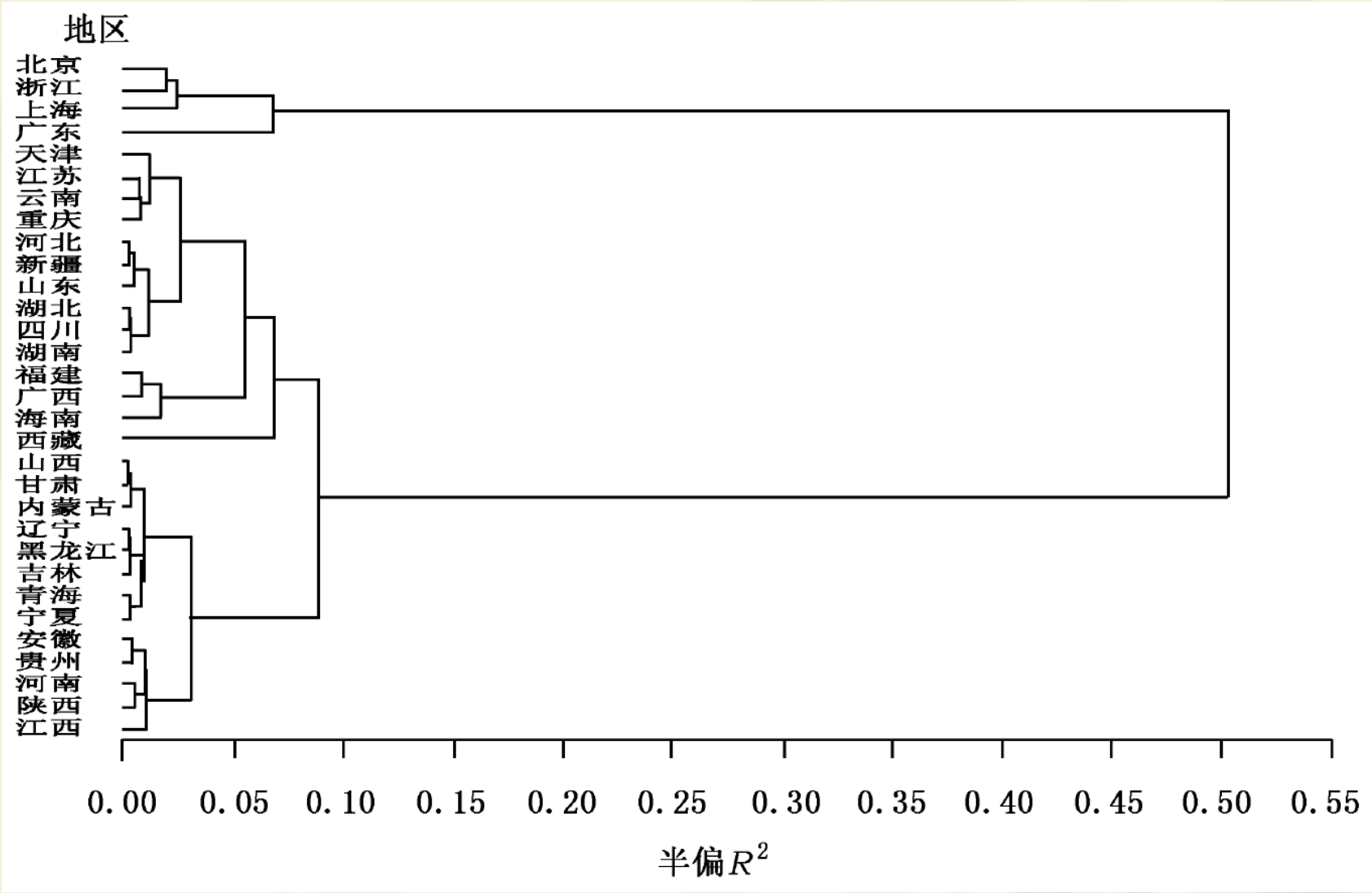


图6.3.13　离差平方和法

- 从这三个树形图来看，只有Ward方法较好地符合了我们的实际聚类要求。可将31个地区分为以下三类：
- 第I类：北京、浙江、上海和广东。这些都是我国经济最发达、城镇居民消费水平最高的地区。

  第II类：天津、江苏、云南、重庆、河北、新疆、山东、湖北、四川、湖南、福建、广西、海南和西藏。这些地区在我国基本上属于经济发展水平和城镇居民消费水平中等的地区。

  第III类：山西、甘肃、内蒙古、辽宁、黑龙江、吉林、青海、宁夏、安徽、贵州、河南、陕西和江西。这些地区在我国基本上属于经济欠发达地区，城镇居民的消费水平也较低。
- 如果分为五类，则广东和西藏将各自为一类。

*七、系统聚类法的统一

- Lance和Williams于1967年将递推公式统一为：

$$D_{MJ}^2 = \alpha_K D_{KJ}^2 + \alpha_L D_{LJ}^2 + \beta D_{KL}^2 + \gamma \left| D_{KJ}^2 - D_{LJ}^2 \right|$$

其中$\alpha_K, \alpha_L, \beta, \gamma$是参数，不同的系统聚类法，它们有不同的取值。表6.3.10列出了上述八种方法四个参数的取值。

表6. 3. 10 系统聚类法参数表

方　法	α_K	α_L	β	γ
最短距离法	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
最长距离法	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
类平均法	n_K/n_M	n_L/n_M	0	0
重心法	n_K/n_M	n_L/n_M	$-\alpha_K\alpha_L$	0
* 中间距离法	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0
离差平方和法	$\frac{n_J+n_K}{n_J+n_M}$	$\frac{n_J+n_L}{n_J+n_M}$	$-\frac{n_J}{n_J+n_M}$	0
* 可变法	$(1-\beta)/2$	$(1-\beta)/2$	$\beta(<1)$	0
* 可变类平均法	$(1-\beta)n_K/n_M$	$(1-\beta)n_L/n_M$	$\beta(<1)$	0

八、系统聚类法的性质

- 1.单调性
- *2.空间的浓缩与扩张
- 3.一个说明性的例子

1.单调性

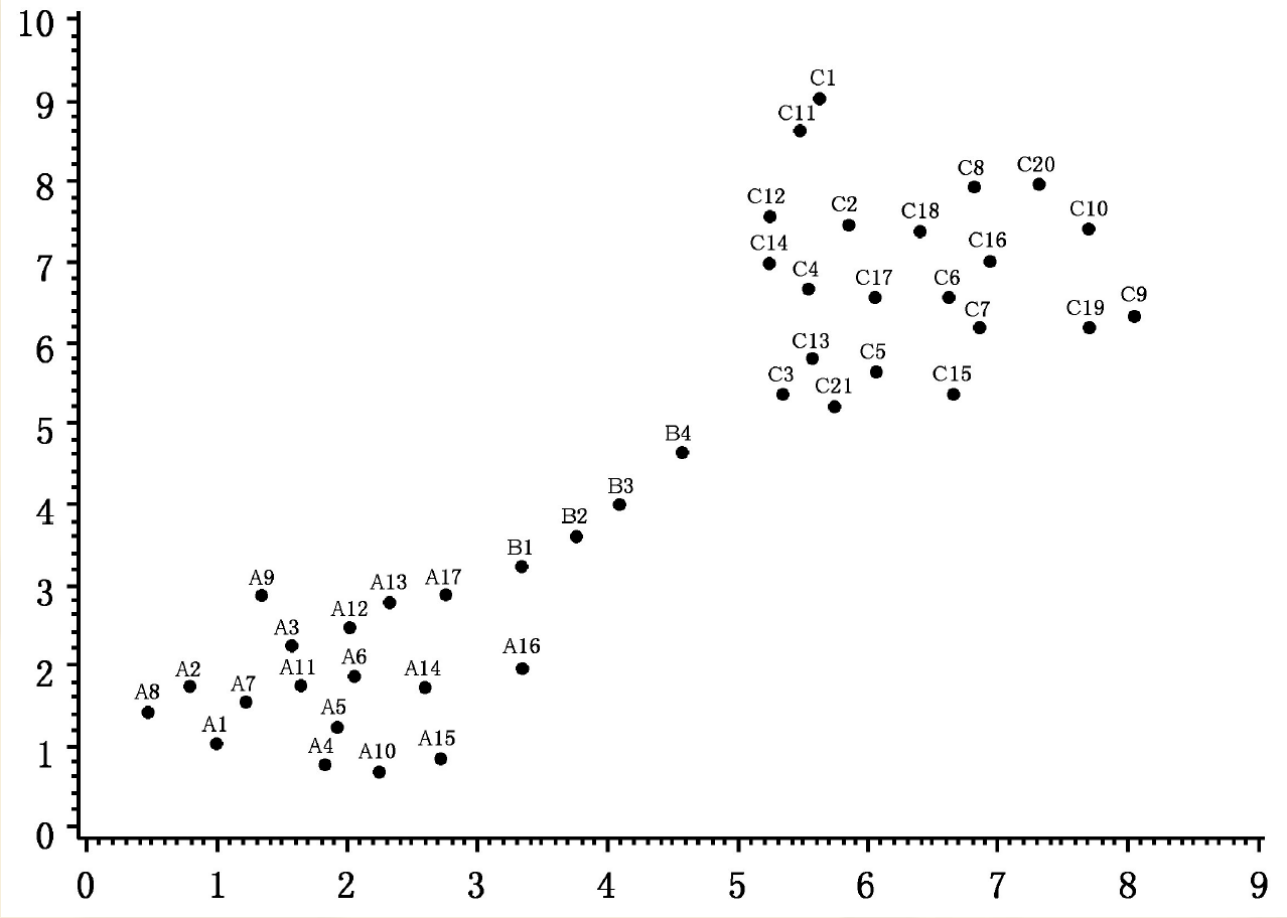
- 令D_i是系统聚类法中第i次并类时的距离，如果一种系统聚类法能满足$D_1 \le D_2 \le D_3 \le \cdots$，则称它具有单调性。
- 这种单调性符合系统聚类法的思想，先合并较相似的类，后合并较疏远的类。
- 最短距离法、最长距离法、类平均法、离差平方和法、可变法和可变类平均法都具有单调性，但重心法和中间距离法不具有单调性。

*2.空间的浓缩与扩张

- 设$\boldsymbol{A}=(a_{ij})$和$\boldsymbol{B}=(b_{ij})$是两个元素非负的同阶矩阵，若$a_{ij}\geq b_{ij}$(对一切i, j)，则记作$\boldsymbol{A}\geq\boldsymbol{B}$。该记号仅在本节中使用。
- 设有两种系统聚类法，它们在第i步的距离矩阵分别为$\boldsymbol{A}_i$和$\boldsymbol{B}_i$，$i=0,1,\cdots,n-1$，若$\boldsymbol{A}_i\geq\boldsymbol{B}_i$，$i=1,\cdots,n-1$，则称第一种方法比第二种方法使空间扩张，或第二种方法比第一种方法使空间浓缩。
- 设聚类中的某步将类G_K和G_L合并成新类G_M，由于接下来的一步在计算类之间的距离时，老类之间的距离仍保持不变，故比较不同聚类法的聚类距离我们只需比较任一老类G_J到新类G_M的距离即可。用$\boldsymbol{D}(*)$表示用“*”方法聚类时的距离矩阵。

- 以类平均法为基准，有如下一些结论：
  - (1) D(短)≤ D(平)， D(重)≤ D(平)。
  - (2) D(长)≥ D(平)。
  - (3) 当0＜β＜1时， D(变平)≤ D(平)；当β＜0时， D(变平)≥ D(平)。
- 太浓缩的方法不够灵敏，太扩张的方法可能因灵敏度过高而容易失真。
- 类平均法比较适中，它既不太浓缩也不太扩张，因此它在这方面是比较理想的。最短距离法是一种非常浓缩的方法，容易出现链接倾向。

3.一个说明性的例子（例6.3.4）



- (1)采用最短距离法。可以算得：
  - 当聚成两类时，C_1和C_{11}组成一类，其余所有的点组成另一类，这里出现了链接现象；
  - 当聚成三类时，C_1和C_{11}组成第I类，其余的C点组成第II类，所有的A点和B点组成第III类。
- (2)采用类平均法。经算得：
  - 当聚成两类时，一类由所有C点构成，另一类由所有A点和所有B点构成；
  - 当聚成三类时，A点群、B点群和C点群各自作为一类。

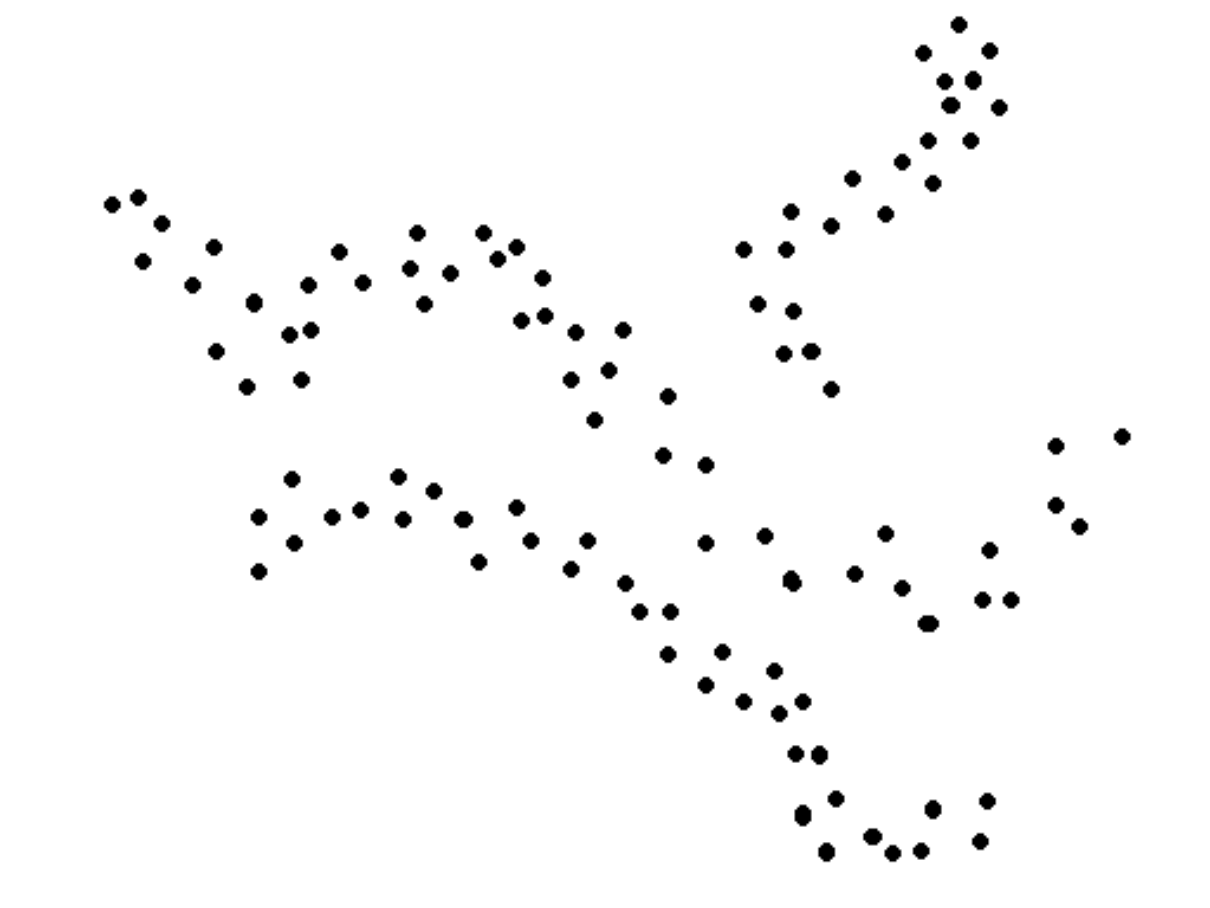
九、使用图形作聚类及对聚类效果的评估

- 1.使用图形作直观的聚类
- 2.使用图形对聚类效果的评估

1.使用图形作直观的聚类

- 当p=2时，可以直接在散点图上进行主观的聚类，其效果未必逊于、甚至好于正规的聚类方法，特别是在寻找“自然的”类和符合我们实际需要的类方面。
- 当p=3时，我们可使用统计软件产生三维旋转图，通过三维旋转从各个角度来观测散点图，作直观的聚类。但由于其视觉效果及易操作性远不如平面散点图，故实践中很少采用。
- 当$p\geq3$时，有时我们可采用主成分分析或因子分析的技术将维数降至2（或3）维，然后再生成散点图（或旋转图），从直觉上进行主观的聚类。

寻找“自然的”类



2.使用图形对聚类效果的评估

- 经聚类分析已将类分好之后，常常希望从统计的角度看一下聚类的效果：不同类之间是否分离得较好，同一类内的样品（或变量）是否彼此相似。
- 通常可通过构造图形作直观的观测，所使用的图形有如下两种：
  - (1)将p维数据画于平面图上，方法有平行（坐标）图、星形图、切尔诺夫脸谱图、星座图和安德鲁曲线图等；
  - (2)使用费希尔判别的降维方法，将p维数据降至2（或3）维再构造散点图（或旋转图）。
  - 如果方法(2)能够成功，则往往更值得推荐，尤其在样品数很大的场合下。

- 例6.3.5　在例6.3.3中，为了从原始数据的直观图形上来看一下按Ward方法聚成三类的效果，使用JMP软件的聚类结果中带有的并排平行图（或称轮廓图）。
- 平行图中的八个变量轴相互平行等间隔，各变量轴上的坐标是已标准化了的值。
- 前两类中的高亮轮廓线分别属于广东和西藏，它们在类内显得较为异类，需要时皆可自成一类。

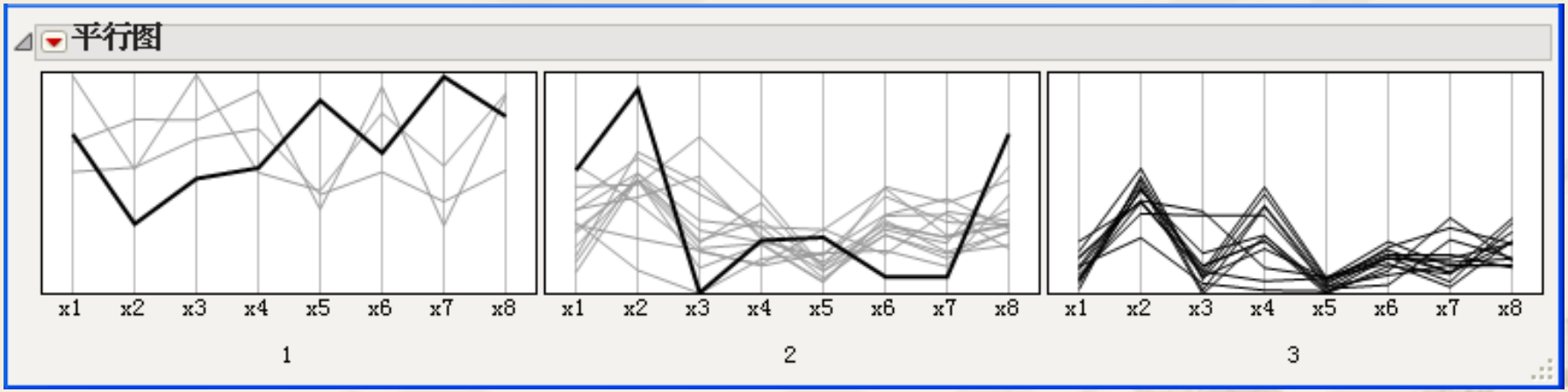


图6.3.15 Ward方法所分三类的平行图

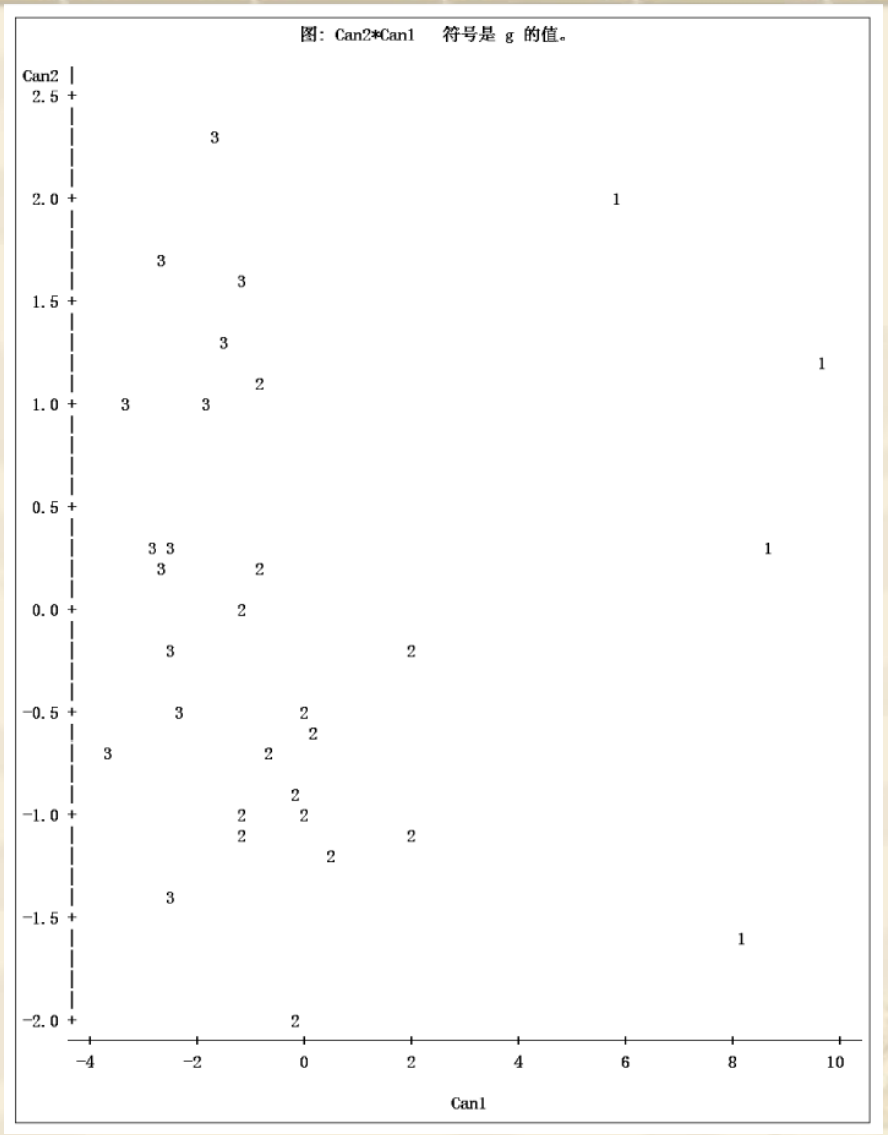


图6.3.16 按图6.3.13分三类的两个判别函数得分的散点图

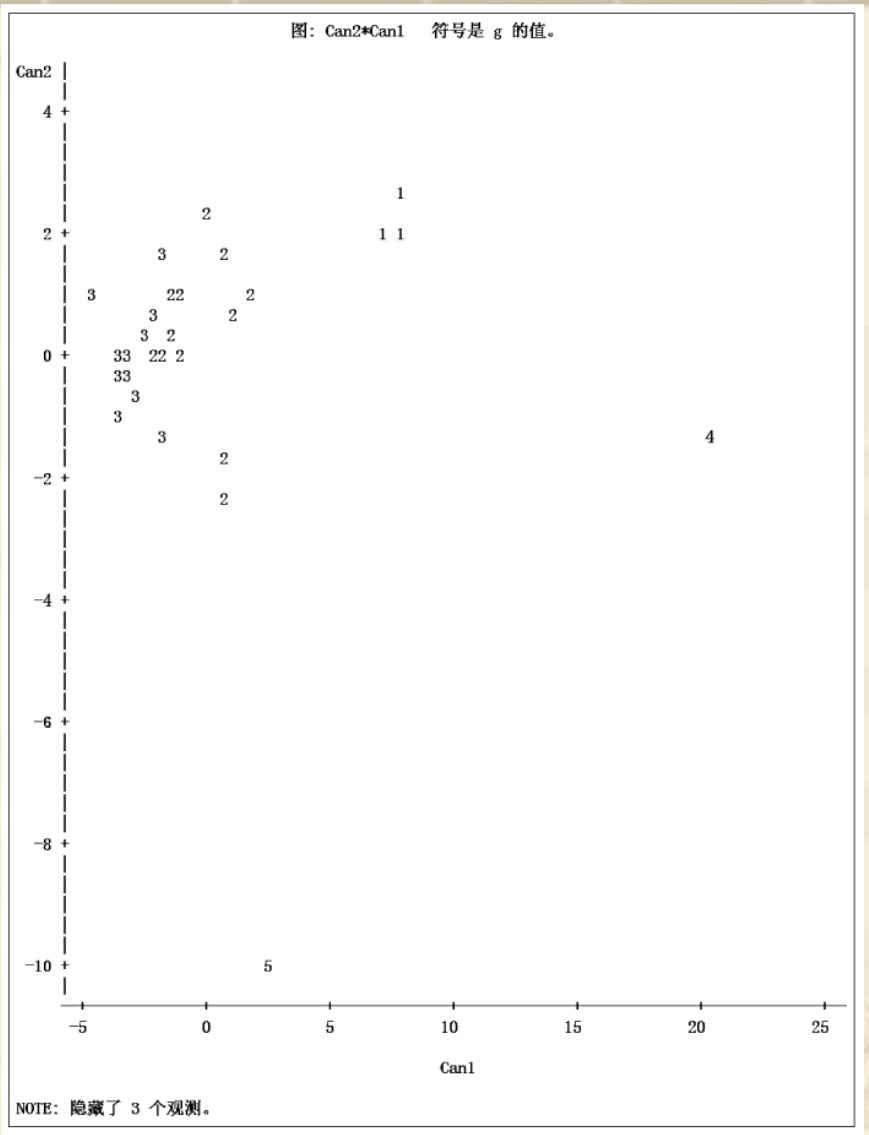


图6.3.17 按图6.3.13分五类的两个判别函数得分的散点图

十、对变量的聚类

- 最短距离法、最长距离法和类平均法都属于连接方法，它们既可以用于样品的聚类，也能够用于变量的聚类。不过并非所有的系统聚类方法都适用于对变量的聚类。

❖ 例6.3.7 对305名女中学生测量八个体型指标：

x_1：身高　　x_5：体重

x_2：手臂长　　x_6：颈围

x_3：上肢长　　x_7：胸围

x_4：下肢长　　x_8：胸宽

表6. 3. 11　　各对变量之间的相关系数

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
x_1	1.000							
x_2	0.846	1.000						
x_3	0.805	0.881	1.000					
x_4	0.859	0.826	0.801	1.000				
x_5	0.473	0.376	0.380	0.436	1.000			
x_6	0.398	0.326	0.319	0.329	0.762	1.000		
x_7	0.301	0.277	0.237	0.327	0.730	0.583	1.000	
x_8	0.382	0.415	0.345	0.365	0.629	0.577	0.539	1.000

- 单从该相关矩阵就可直观地判断出聚成两类：$\{x_1,x_2,x_3,x_4\}$和$\{x_5,x_6,x_7,x_8\}$，这两类的特征明显，其类内变量分别都是身材方面的“纵向”指标和“横向”指标。
- 分别用最短距离法、最长距离法和(6.3.5)式的类平均法对变量进行聚类，这三种方法的类与类之间的相似系数分别定义为两类变量间的最大、最小和平均相关系数，每次聚类时合并两个相似系数最大的类。
- 从图6.3.18可见，聚成两类：$\{x_1,x_2,x_3,x_4\}$和$\{x_5,x_6,x_7,x_8\}$。
- 最短距离法和类平均法也都有与此相同的聚成两类的结果。

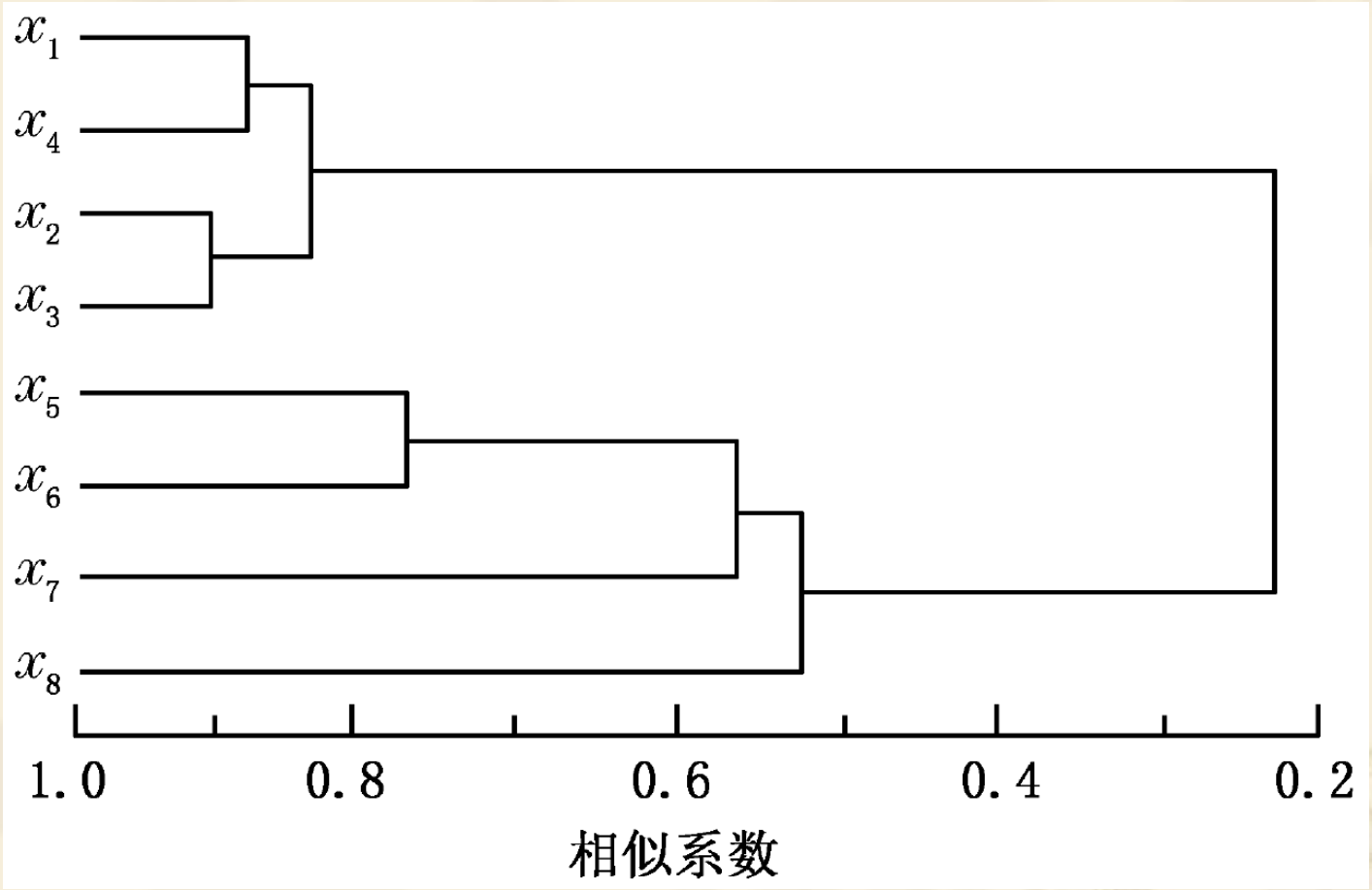


图6.3.18　八个体型变量的最长距离法树形图

十一、类的个数

- 如果能够分成若干很分开的类，则类的个数就比较容易确定；反之，如果无论怎样分都很难分成明显分开的若干类，则类个数的确定可能就比较困难了。
- 确定类个数的常用方法有：

  1.给定一个阈值T。

  2.观测样品的散点图。

  *3.使用统计量。

1.给定一个阈值T

- 通过观测树形图，给出一个你认为合适的阈值T，要求类与类之间的距离要大于T，有些样品可能会因此而归不了类或只能自成一类。这种方法有较强的主观性，这是它的不足之处。

2.观测样品的散点图

- 如果样品只有两个（或三个）变量，则可通过观测数据的散点图（或旋转图）来主观确定类的个数。
- 如果变量个数超过三个，则可对每一可能考虑的聚类结果，将所有样品的前两个（或三个）费希尔判别函数得分制作成散点图（或旋转图），目测类之间是否分离得较好。
- 该图既能帮助我们评估聚类效果的好坏，也能帮助我们判断所定的类数目是否恰当。
- 例6.3.8　在例6.3.6中，图6.3.16显示分为三类是合适的，图6.3.17显示分为五类也是合适的。最终到底分为几类还需综合考虑，该例分成三类似乎更符合实际的需要。

*3.使用统计量

- （1）R^2统计量。
- （2）半偏R^2统计量。
- （3）伪F统计量。
- （4）伪t统计量。

§6.4 动态聚类法

- 在系统聚类法中，对于那些先前已被“错误”分类的样品不再提供重新分类的机会，而动态聚类法（或称逐步聚类法）却允许样品从一个类移动到另一个类中。
- 动态聚类法的计算量要比建立在距离矩阵基础上的系统聚类法小得多。因此，使用动态聚类法计算机所能承受的样品数目n要远远超过使用系统聚类法所能承受的n。

- 动态聚类法的基本思想是，选择一批凝聚点或给出一个初始的分类，让样品按某种原则向凝聚点凝聚，对凝聚点进行不断的修改或迭代，直至分类比较合理或迭代稳定为止。类的个数k需先指定一个。
- 选择初始凝聚点（或给出初始分类）的一种简单方法是采用随机抽选（或随机分割）样品的方法，可以要求凝聚点之间至少应间隔某个距离值。
- 动态聚类法只能用于对样品的聚类，而不能用于对变量的聚类。
- 动态聚类法有许多种方法，在这一节中，我们将讨论一种比较流行的动态聚类法——k均值法。它是由麦奎因（MacQueen，1967）提出并命名的一种算法。

k均值法的基本步骤

- (1)选择k个样品作为初始凝聚点，或者将所有样品分成k个初始类，然后将这k个类的重心（均值）作为初始凝聚点。
- (2)对所有的样品逐个归类，将每个样品归入凝聚点离它最近的那个类（通常采用欧氏距离），该类的凝聚点更新为这一类目前的均值，直至所有样品都归了类。
- (3)重复步骤(2)，直至所有的样品都不能再分配为止。

- 最终的聚类结果在一定程度上依赖于初始凝聚点或初始分类的选择。经验表明，聚类过程中的绝大多数重要变化均发生在第一次再分配中。
- 例6.4.1　对例6.3.1采用k均值法聚类，指定k=2，具体步骤如下：
- (1) 随意将这些样品分成$G_1^{(0)}=\{1,6,8\}$和$G_2^{(0)}=\{2,11\}$两类，则这两个初始类的均值分别是5和$6\frac{1}{2}$。
- (2)计算1到两个类(均值)的欧氏距离

$$d\left(1,G_1^{(0)}\right)=|1-5|=4$$

$$d\left(1,G_2^{(0)}\right)=\left|1-6\frac{1}{2}\right|=5\frac{1}{2}$$

1不用重新分配，计算6到两个类的距离

$$d\left(6,G_1^{(0)}\right)=\left|6-5\right|=1$$

$$d\left(6,G_2^{(0)}\right)=\left|6-6\frac{1}{2}\right|=\frac{1}{2}$$

故6应重新分配到$G_2^{(0)}$中，修正后的两个类为$G_1^{(1)}=\{1,8\}$和$G_2^{(1)}=\{2,6,11\}$，新的类均值分别为$4\frac{1}{2}$和$6\frac{1}{3}$。计算

$$d\left(8,G_1^{(1)}\right)=\left|8-4\frac{1}{2}\right|=3\frac{1}{2}$$

$$d\left(8,G_2^{(1)}\right)=\left|8-6\frac{1}{3}\right|=1\frac{2}{3}$$

结果8重新分配到$G_2^{(1)}$中，两个新类为 $G_1^{(2)}=\{1\}$，$G_2^{(2)}=\{2,6,8,11\}$，其类均值分别为1和 $6\frac{3}{4}$。再计算

$$d\left(2,G_1^{(2)}\right)=|2-1|=1$$

$$d\left(2,G_2^{(2)}\right)=\left|2-6\frac{3}{4}\right|=4\frac{3}{4}$$

重新分配2到$G_1^{(2)}$中，两个新类为$G_1^{(3)}=\{1,2\}$，$G_2^{(3)}=\{6,8,11\}$，其类均值分别为$1\frac{1}{2}$和$8\frac{1}{3}$。

- (3)再次计算每个样品到类均值的距离，结果列于表6.4.1。
- 最终得到的两个类为{1,2}和{6,8,11}。

表6. 4. 1　　各样品到类均值的距离

样品 / 类	1	2	6	8	11
$G_1^{(3)}=\{1,2\}$	$\frac{1}{2}$	$\frac{1}{2}$	$4\frac{1}{2}$	$6\frac{1}{2}$	$9\frac{1}{2}$
$G_2^{(3)}=\{6,8,11\}$	$7\frac{1}{3}$	$6\frac{1}{3}$	$2\frac{1}{3}$	$\frac{1}{3}$	$2\frac{2}{3}$

- 例6.4.2　对例6.3.3使用k均值法进行聚类，聚类前对各变量作标准化变换，聚类结果如下：

第I类：北京、上海和浙江。

第II类：广东。

第III类：天津、江苏、福建、山东、湖南、广西、重庆、四川和云南。

第IV类：河北、山西、内蒙古、辽宁、吉林、黑龙江、安徽、江西、河南、湖北、海南、贵州、陕西、甘肃、青海、宁夏和新疆。

第V类：西藏。

- 由于k均值法对凝聚点的初始选择有一定敏感性，故再试一下其他初始的凝聚点也许是个不错的想法。如果不同初始凝聚点的选择产生明显不同的最终聚类结果，或者迭代的收敛是极缓慢的，那么可能表明没有自然的类可以形成。
- k均值法有时也可用来改进系统聚类的结果，例如，先用类平均法聚类，然后将其各类的重心作为k均值法的初始凝聚点重新聚类，这可使得系统聚类时错分的样品能有机会获得重新的分类。不过，k均值法能否有效地改善系统聚类，我们不能一概而论，还应视聚类的最终结果而定。