重慶工商大學

CHONGQING TECHNOLOGY AND BUSINESS UNIVERSITY

聚类

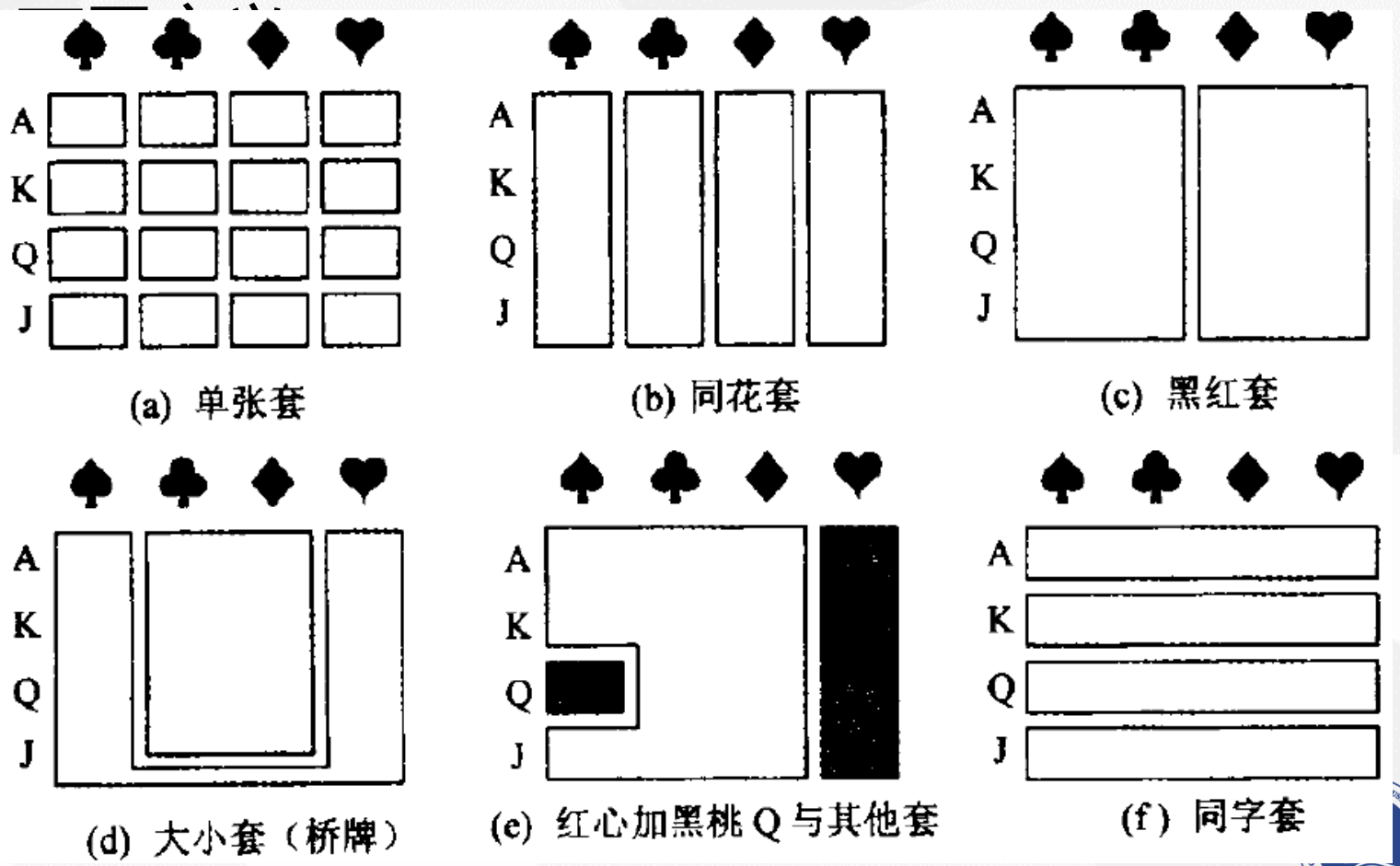
CLUSTERING

数学与统计学院 杨炜明

1 引言

- 判别分析和聚类分析是两种不同目的的分类方法，它们所起的作用是不同的。
- 判别分析方法假定组（或类）已事先分好，判别新样品应归属哪一组，对组的事先划分有时也可以通过聚类分析得到。
- 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



2 距离和相似系数

- 相似性度量：距离和相似系数。
- 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。
- 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

- 通常变量按测量尺度的不同可以分为间隔、有序和名义尺度变量三类。

➢间隔尺度变量：变量用连续的量来表示，如长度、重量、速度、温度等。

➢有序尺度变量：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。

➢名义尺度变量：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。

一、距离

- 设x_{ij}为第i个样品的第j个指标，数据矩阵列于表6.2.1。

表6. 2. 1　　数据矩阵

样品 \ 变量	x_1	x_2	…	x_p
1	x_{11}	x_{12}	…	x_{1p}
2	x_{21}	x_{22}	…	x_{2p}
⋮	⋮	⋮		⋮

距离d_{ij}一般应满足的四个条件

- (i) $d_{ij}\geq 0$，对一切i,j；
- (ii) $d_{ij}=0$，当且仅当第i个样品与第j个样品的各变量值相同；
- (iii) $d_{ij}=d_{ji}$，对一切i,j；
- (iv) $d_{ij}\leq d_{ik}+d_{kj}$，对一切i,j,k。

常用的距离

- 1.欧式(Euclidean)距离
- 2.兰氏(Lance和Williams)距离
- 3.切比雪夫(Chebychev)距离
- 4.明考夫斯基(Minkowski)距离
- 5.马氏(Mahalanobis)距离

Euclidean距离	$\sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$
Lance&Williams距离	$\sum_{k=1}^{p}\frac{\lvert x_{ik} - x_{jk}\rvert}{x_{ik} + x_{jk}}$
Chebychev距离	$\max\lvert x_i - y_i\rvert$
Minkovski距离	$\sqrt[q]{\sum_{i=1}^{p}\lvert x_i - y_i\rvert^q}$
Mahalanobis距离	$\sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)' \boldsymbol{S}^{-1}(\boldsymbol{x}_i - \boldsymbol{x}_j)}$

1.明考夫斯基距离

- 第i个样品与第j个样品间的明考夫斯基距离（简称明氏距离）定义为

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

这里q为某一自然数。明氏距离有以下三种特殊形式：

- （i）当$q=1$时，$d_{ij}(1)=\sum_{k=1}^{p}\left|x_{ik}-x_{jk}\right|$，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，$d_{ij}(2)=\left[\sum_{k=1}^{p}\left|x_{ik}-x_{jk}\right|^{2}\right]^{1/2}$，称为欧氏距离，这是聚类分析中最常用的一个距离；
- （iii）当$q=\infty$时，$d_{ij}(\infty)=\max_{1\le k\le p}\left|x_{ik}-x_{jk}\right|$，称为切比雪夫距离。

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

$$x_{ij}^{*}=\frac{x_{ij}-\bar{x}_j}{\sqrt{s_{jj}}},\quad i=1,2,\cdots,n,\quad j=1,2,\cdots,p$$

其中 $\bar{x}_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij}$ 和 $s_{jj}=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_j\right)^2$ 分别为第j个变量的样本均值和样本方差。

2.兰氏距离

- 当$x_{ji}>0$，$j=1,2,\cdots,n$，$i=1,2,\cdots,p$时，可以定义第i个样品与第j个样品间的兰氏距离为

$$d_{ij}(L)=\sum_{k=1}^{p}\frac{\left|x_{ik}-x_{jk}\right|}{x_{ik}+x_{jk}}$$

这个距离与各变量的单位无关。由于它对大的异常值不敏感，故适用于高度偏斜的数据。

明氏距离和兰氏距离都没有考虑变量间的相关性，因此这两种距离更适合各变量之间互不相关的情形。

3.马氏距离

- 第i个样品与第j个样品间的马氏距离为

$$d_{ij}(M)=\sqrt{(\boldsymbol{x}_i-\boldsymbol{x}_j)'\boldsymbol{S}^{-1}(\boldsymbol{x}_i-\boldsymbol{x}_j)}$$

其中$\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots,x_{ip})'$，$\boldsymbol{x}_j=(x_{j1},x_{j2},\cdots,x_{jp})'$，$\boldsymbol{S}$为样本协方差矩阵。

- 使用马氏距离的好处是考虑到了各变量之间的相关性，并且与各变量的单位无关。但马氏距离有一个很大的缺陷，就是马氏距离公式中的$\boldsymbol{S}$难以确定。没有关于不同类的先验知识，$\boldsymbol{S}$就无法计算。因此，在实际聚类分析中，马氏距离不是理想的距离。

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：性别(x_1)，取值为男和女；外语语种(x_2)，取值为英、日和俄；专业(x_3)，取值为统计、会计和金融；职业(x_4)，取值为教师和非教师；居住处(x_5)，取值为校内和校外；学历(x_6)，取值为本科和本科以下。

➢现有两名学员：

$$\boldsymbol{x}_1=（男，英，统计，非教师，校外，本科）'$$

$$\boldsymbol{x}_2=（女，英，金融，教师，校外，本科以下）'$$

➢这两名学员的第二个变量都取值“英”，称为配合的，第一个变量一个取值为“男”，另一个取值为“女”，称为不配合的。一般地，若记配合的变量数为m_1，不配合的变量数为m_2，则它们之间的距离可定义为

$$d_{12}=\frac{m_2}{m_1+m_2}$$

➢故按此定义本例中$\boldsymbol{x}_1$与$\boldsymbol{x}_2$之间的距离为2/3。

二、相似系数

- 聚类分析方法不仅用来对样品进行分类，而且可用来对变量进行分类，在对变量进行分类时，常常采用相似系数来度量变量之间的相似性。
- 变量之间的这种相似性度量，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。

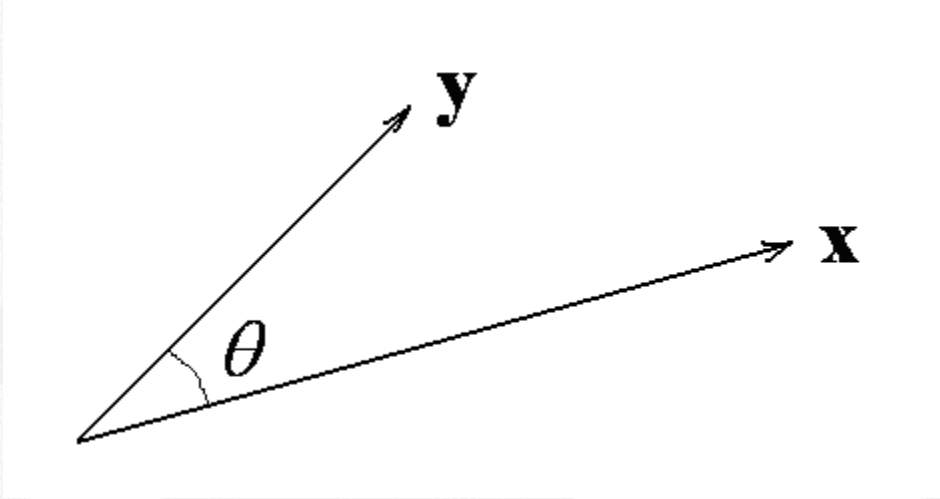
相似系数一般需满足的条件

(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



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

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

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

- 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，显然c_{ij}满足定义相似系数的三个条件，故可作为相似系数。

- 距离必须满足定义距离的四个条件，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

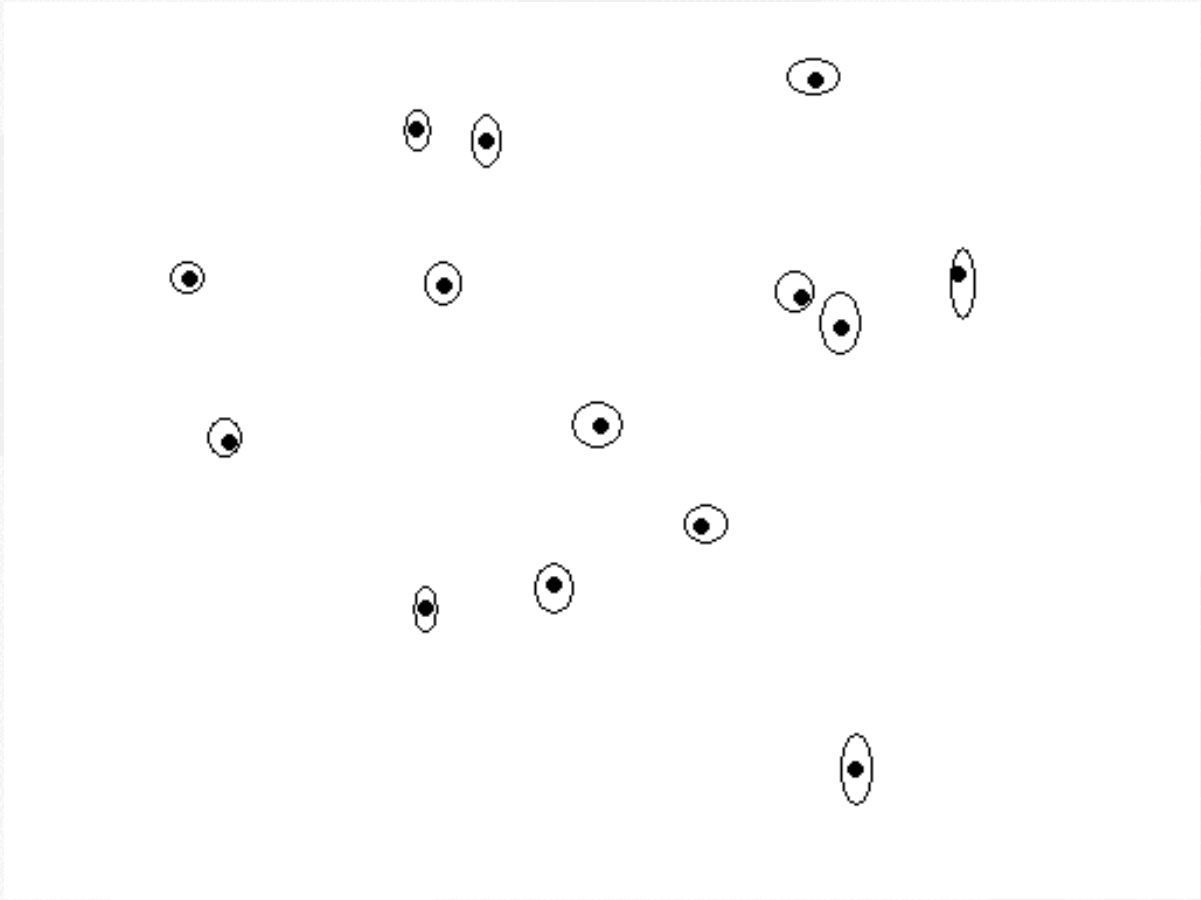
$$d_{ij}=\sqrt{2\left(1-c_{ij}\right)}$$

则d_{ij}满足距离定义的四个条件。

3 系统聚类法

- 系统聚类法(hierarchical clustering method)是聚类分析诸方法中用得最多的一种。
- 基本思想是：开始将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。
- 本节介绍的几种系统聚类方法，其区别在于类与类之间距离的计算方法不同。

一开始每个样品各自作为一类



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：$D_{KL}=d_{23}$

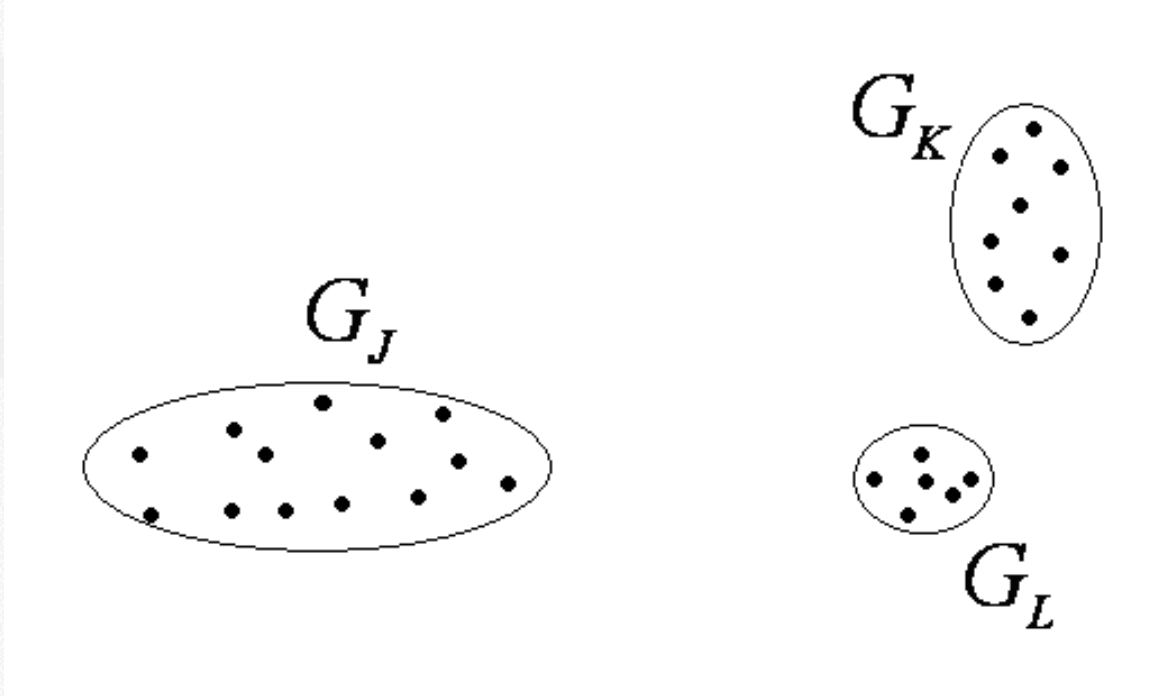
最短距离法的聚类步骤

- (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

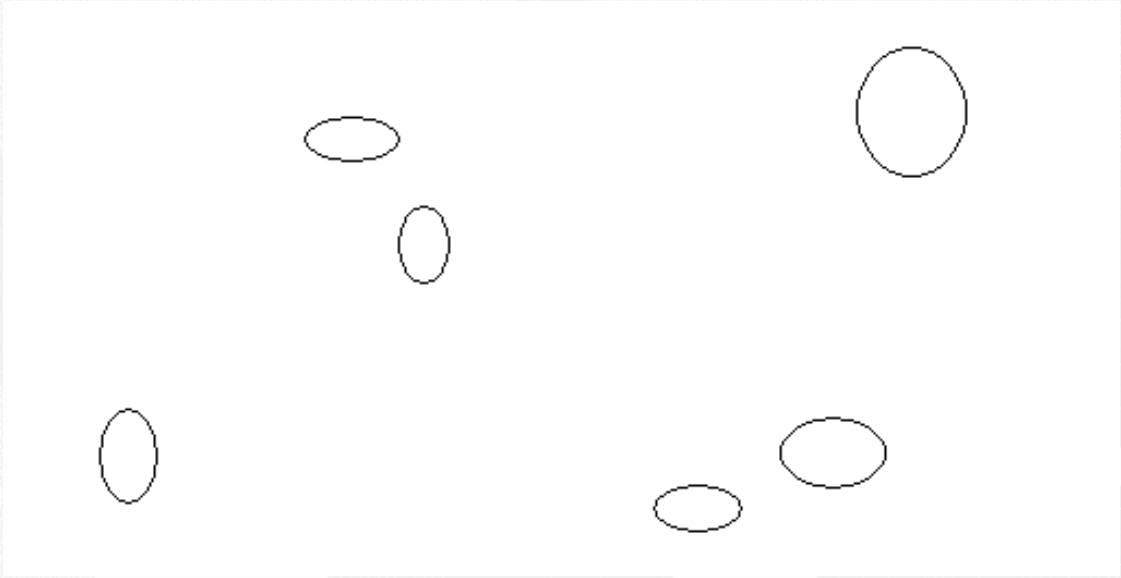


最短距离法的聚类步骤

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（6.3.2）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（tie），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。



- 由于最短距离法是用两类之间最近样本点的距离来聚的，因此该方法不适合对分离得很差的群体进行聚类

- 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$G_8 = G_5 \cup G_7$

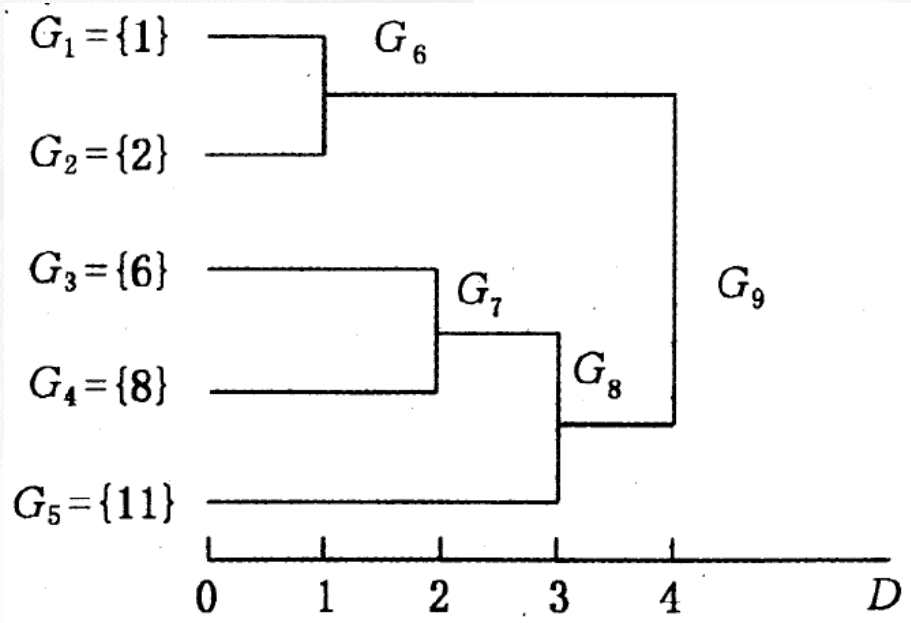


图6. 3. 2　最短距离法树形图

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

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

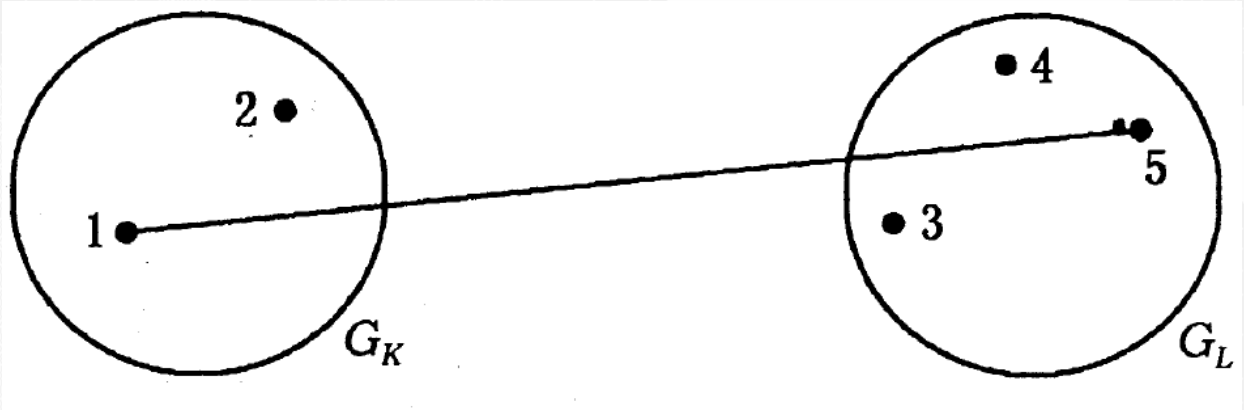
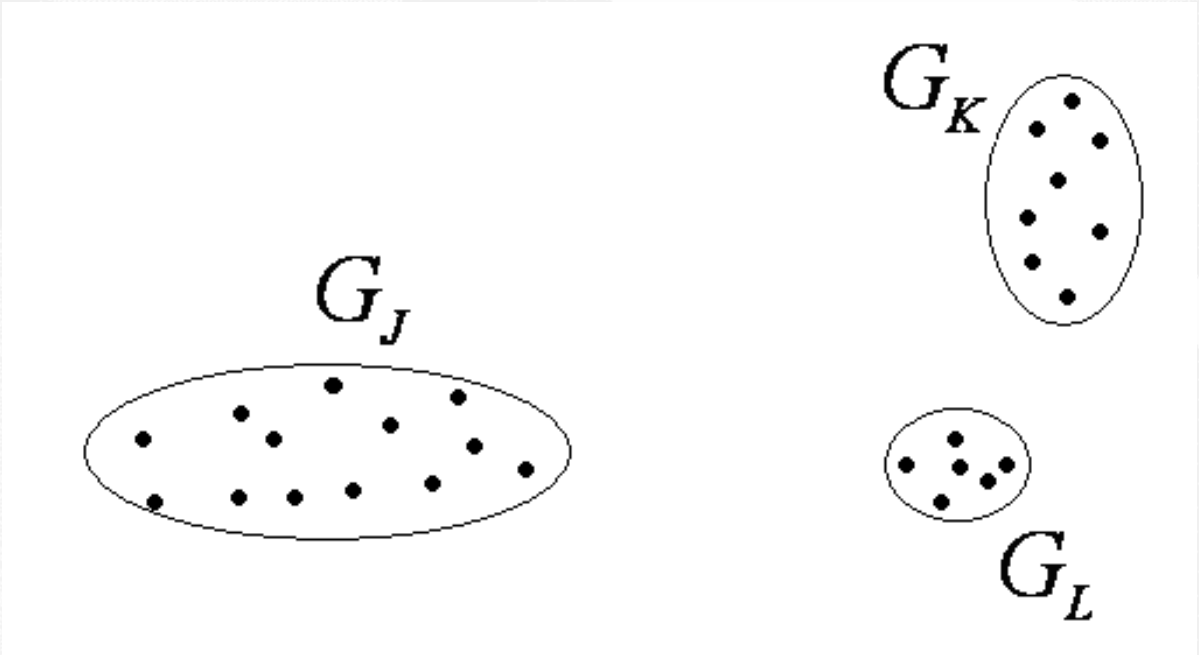


图6. 3. 3　最长距离法：　$D_{KL}=d_{15}$

- 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：

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



- 对例6.3.1采用最长距离法，其树形图如图6.3.4所示，它与图6.3.2有相似的形状，但并类的距离要比图6.3.2大一些，仍分成两类为宜。

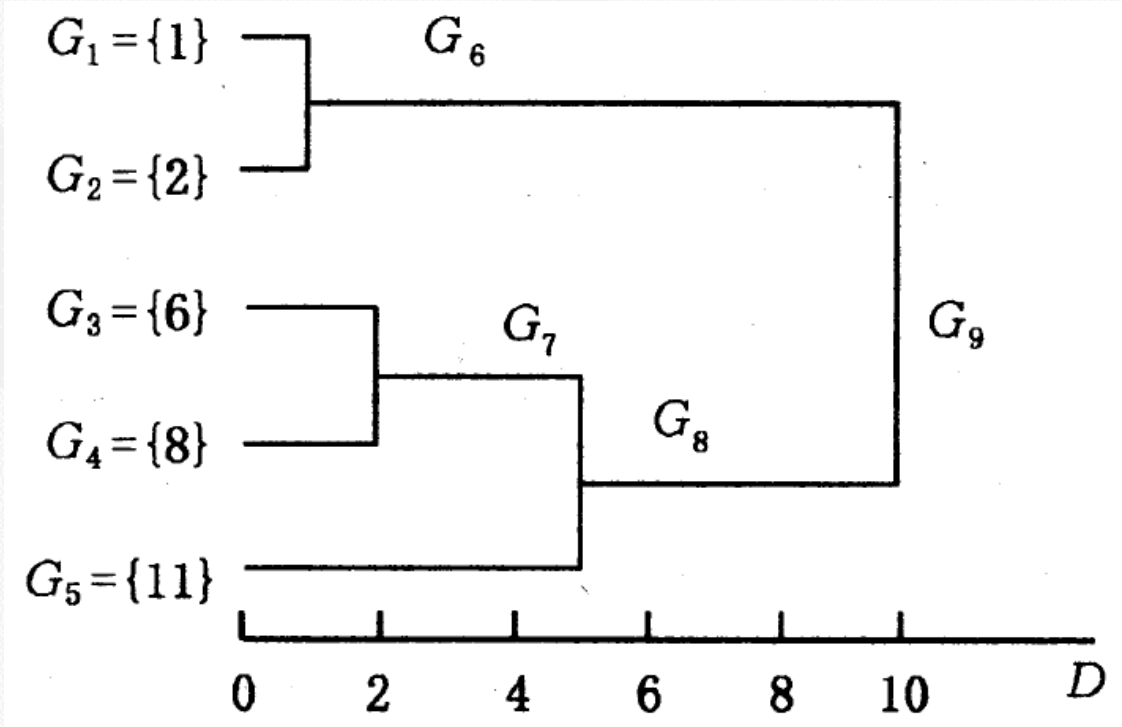
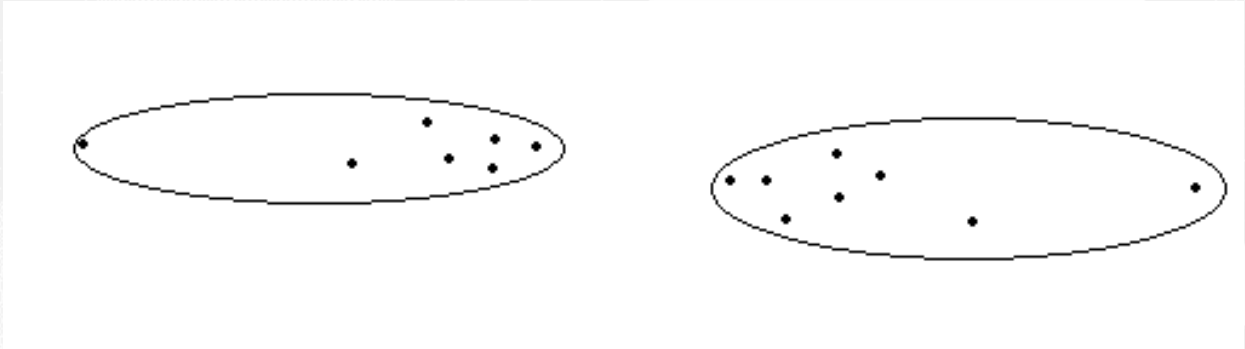


图6. 3. 4　最长距离法树形图

异常值的影响

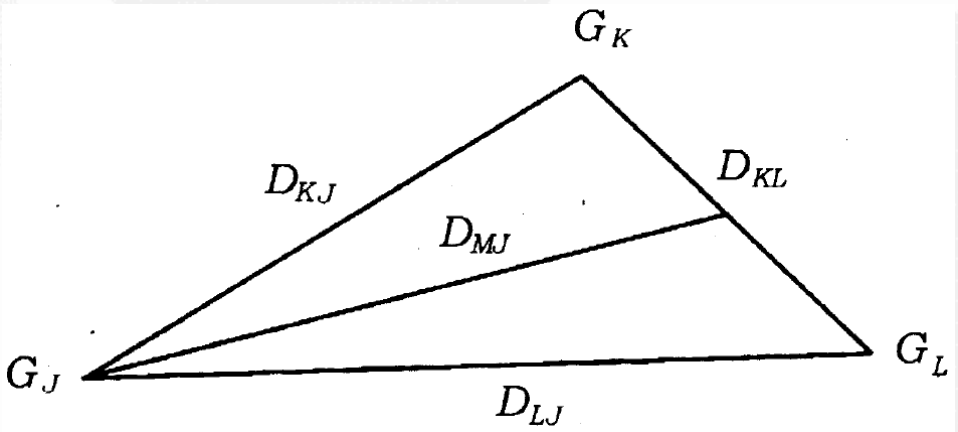
- 最长距离法容易被异常值严重地扭曲，一个有效的方法是将这些异常值单独拿出来后再进行聚类。



三、中间距离法

- 类与类之间的距离既不取两类最近样品间的距离，也不取两类最远样品间的距离，而是取介于两者中间的距离，称为中间距离法(median method)。
- 设某一步将G_K和G_L合并为G_M，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(如下图所示)，取D_{KL}边的中线作为D_{MJ}。D_{MJ}的计算公式为

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



四、类平均法

- 类平均法(average linkage method)有两种定义，一种定义方法是把类与类之间的距离定义为所有样品对之间的平均距离，即定义G_K和G_L之间的距离为

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

其中n_K和n_L分别为类G_K和G_L的样品个数，d_{ij}为G_K中的样品i与G_L中的样品j之间的距离，如图6.3.7所示。容易得到它的一个递推公式：

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

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

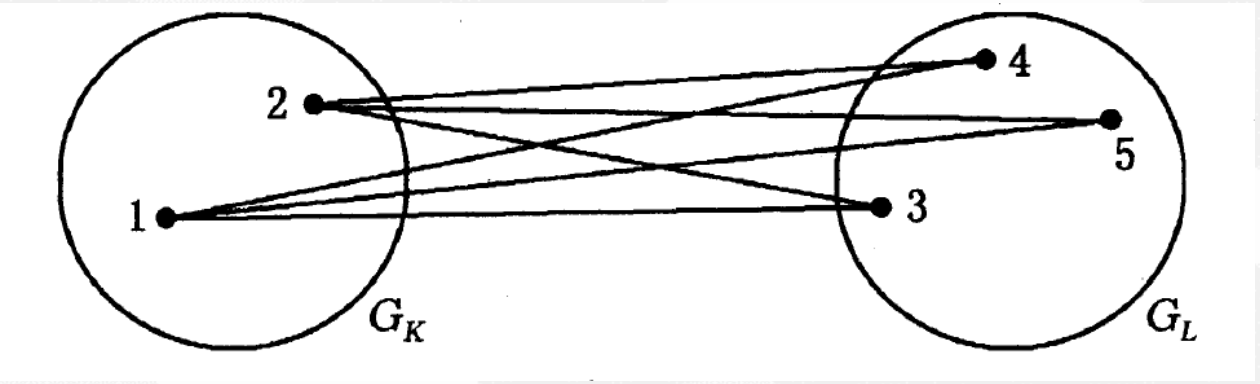


图6. 3. 7　类平均法：$D_{KL}=(d_{13}+d_{14}+d_{15}+d_{23}+d_{24}+d_{25})/6$

- 另一种定义方法是定义类与类之间的平方距离为样品对之间平方距离的平均值，即

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

- 它的递推公式为

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

- 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。

- 对例6.3.1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. 6 $\boldsymbol{D}_{(0)}^{2}$

	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. 7 $D^2_{(1)}$

	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. 8 $D^2_{(2)}$

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

表6. 3. 9　　$D^2_{(3)}$

	G_6	G_8
G_6	0	
G_8	51.17	0

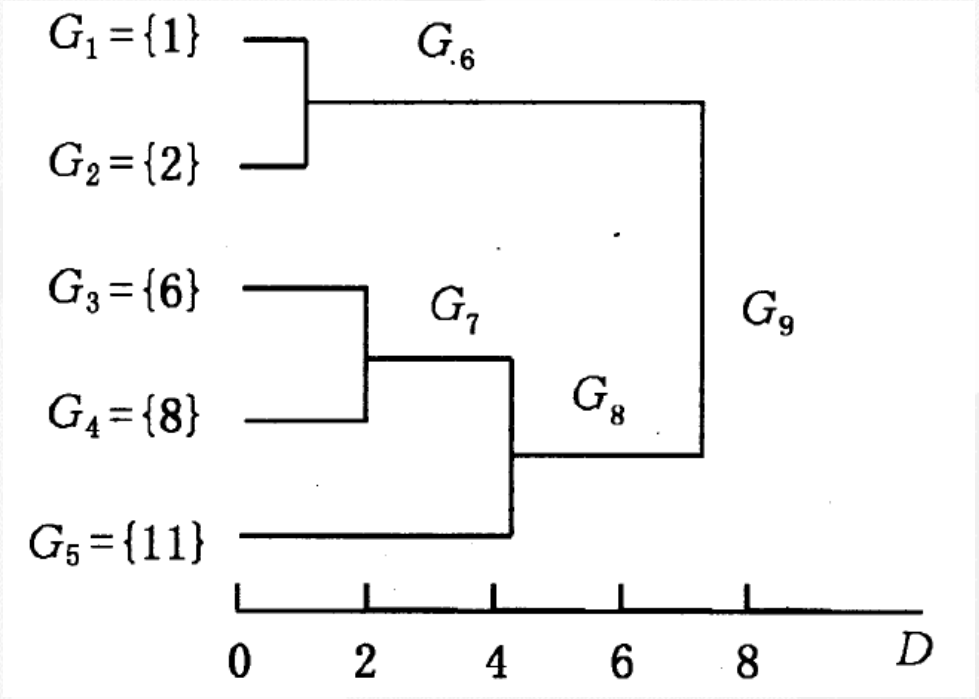


图6. 3. 8　类平均法树形图

五、重心法

- 类与类之间的距离定义为它们的重心(均值)之间的欧氏距离。设G_K和G_L的重心分别为 $\bar{\boldsymbol{x}}_K$和$\bar{\boldsymbol{x}}_L$，则G_K与G_L之间的平方距离为

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

- 这种系统聚类法称为重心法(centroid hierarchical method)，如图6.3.9所示。它的递推公式为

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

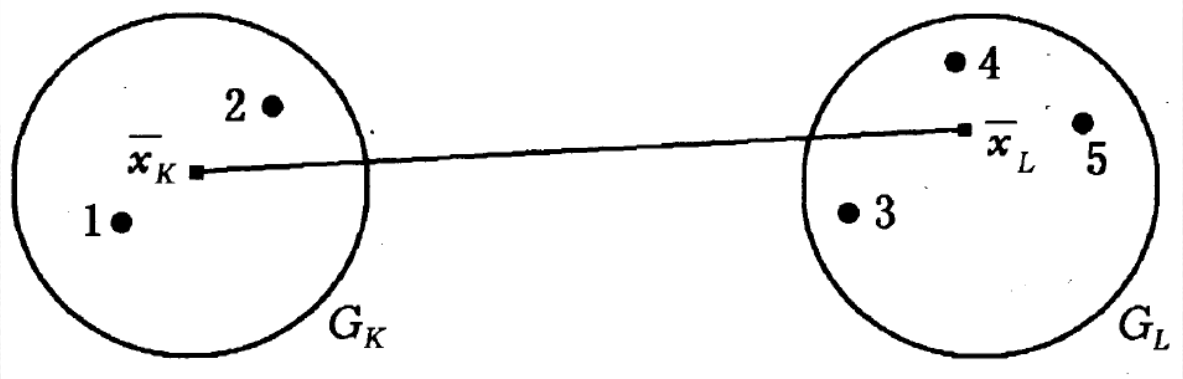


图6.3.9　重心法：$D_{KL} = d_{\bar{x}_K \bar{x}_L}$

- 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。

六、离差平方和法(Ward方法)

- 类中各样品到类重心（均值）的平方欧氏距离之和称为（类内）离差平方和。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

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

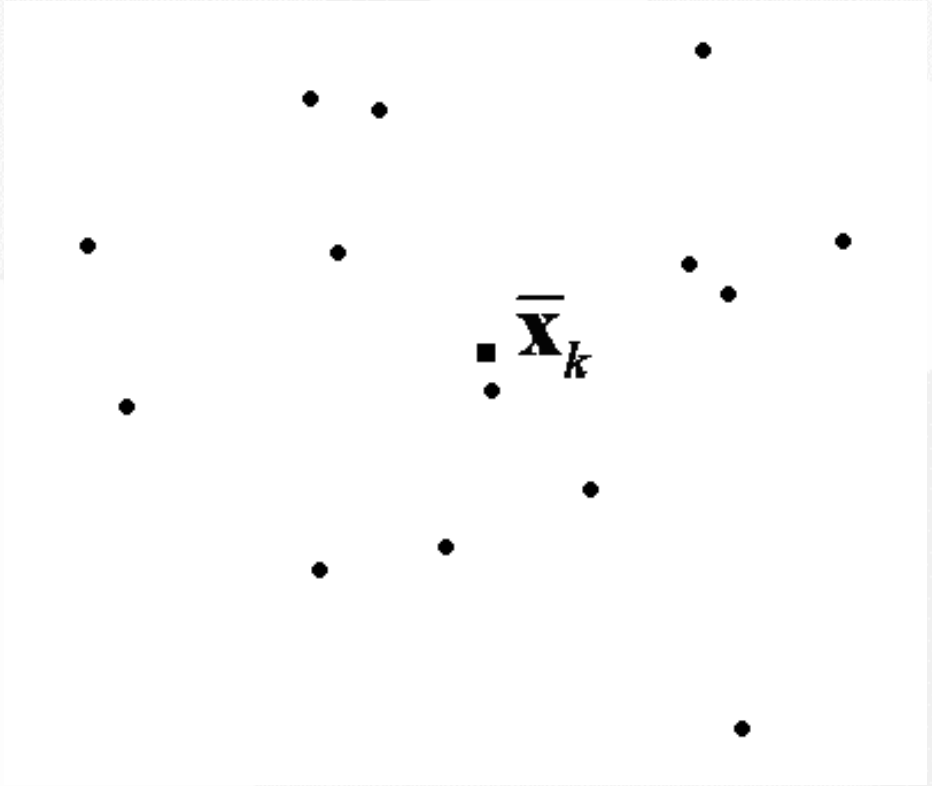
$$W_L = \sum_{i \in G_L} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_L)' (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_L)$$

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

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

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

这种系统聚类法称为离差平方和法或Ward方法(Ward's minimum variance method)。

- D_{KL}^2 也可表达为

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

- 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。

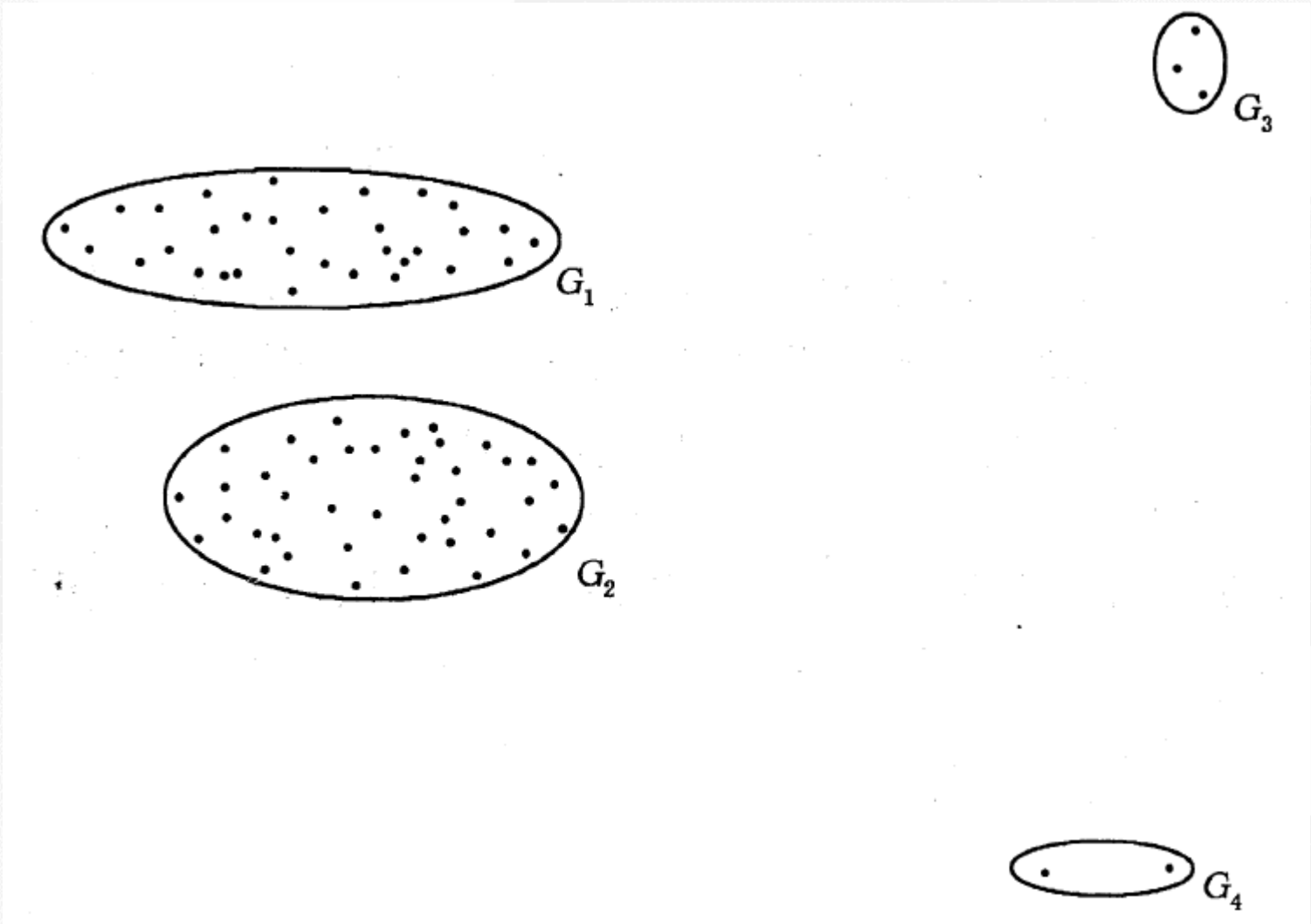


图6. 3. 10　离差平方和法与重心法的聚类比较

- 离差平方和法的平方距离递推公式为

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

- 对例6.3.1采用离差平方和法进行聚类。

表6. 3. 10 $D_{(0)}^2$

	G_1	G_2	G_3	G_4	G_5
G_1	0				
G_2	0.5	0			
G_3	12.5	8	0		
G_4	24.5	18	2	0	
G_5	50	40.5	12.5	4.5	0

表6. 3. 11 $D^2_{(1)}$

	G_6	G_3	G_4	G_5
G_6	0			
G_3	13.5	0		
G_4	28.17	2	0	
G_5	60.17	12.5	4.5	0

表6. 3. 12 $D^2_{(2)}$

	G_6	G_7	G_5
G_6	0		
G_7	30.25	0	
G_5	60.17	10.67	0

表6. 3. 13　　$D^2_{(3)}$

	G_6	G_8
G_6	0	
G_8	56.03	0

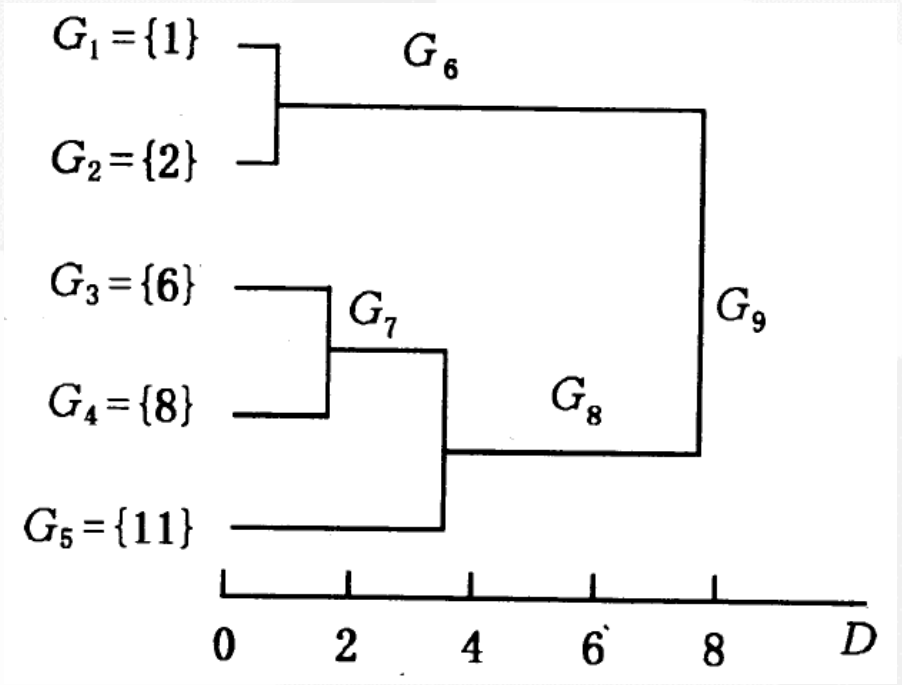


图6. 3. 11　离差平方和法树形图

- 例6.3.3　表6.3.14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. 14　　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

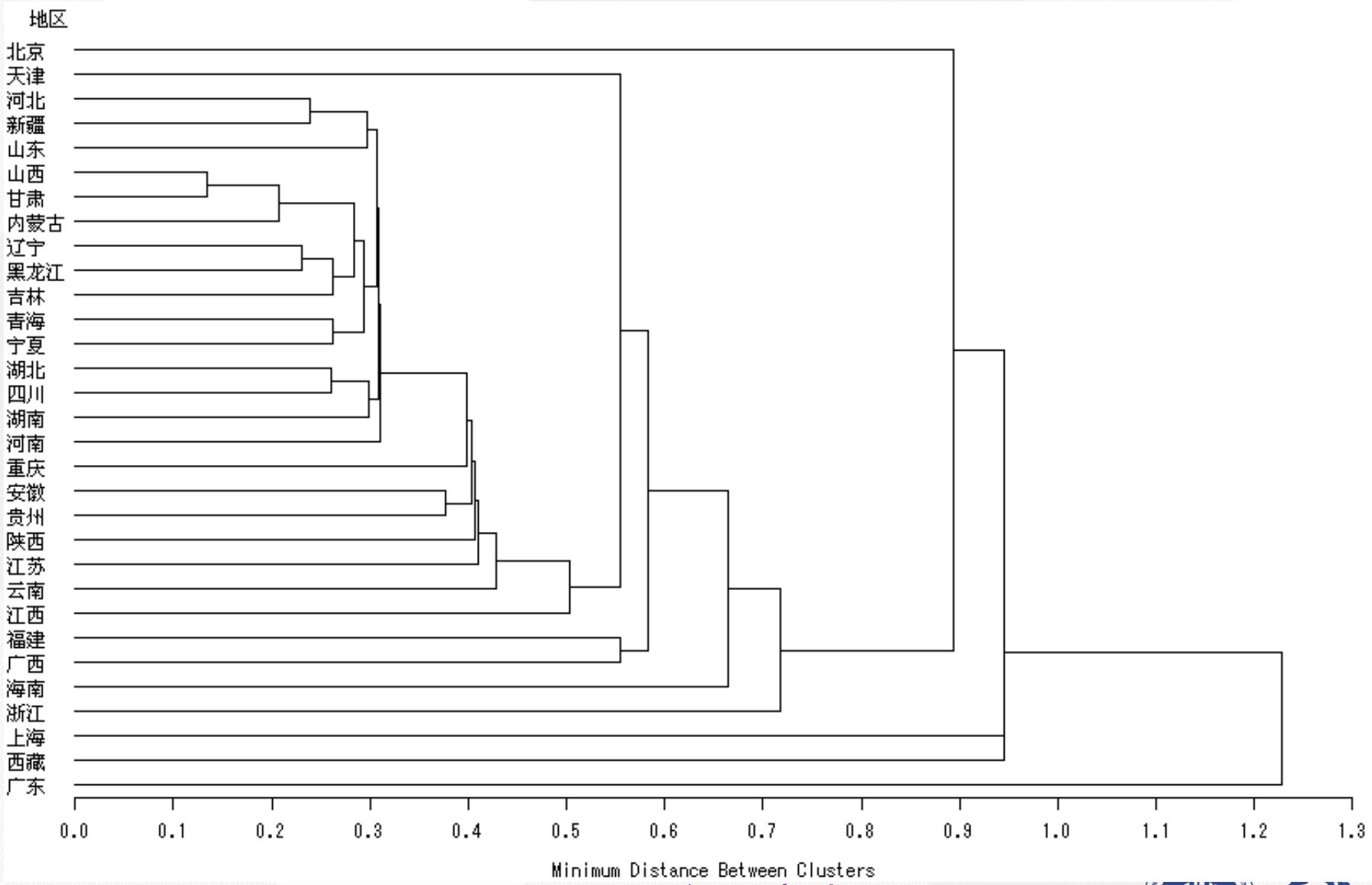


图6.3.12　最短距离法

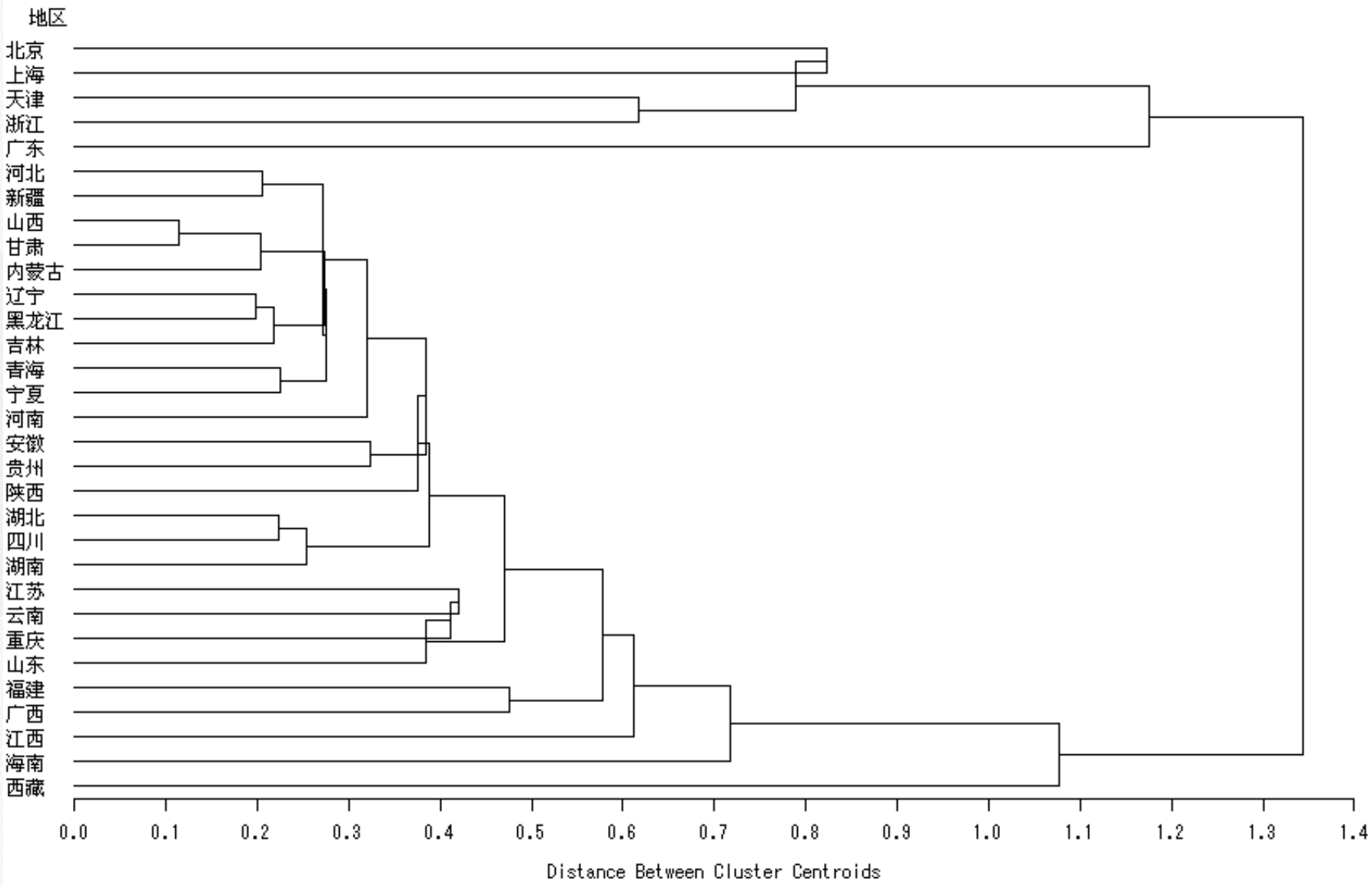


图6.3.13　重心法

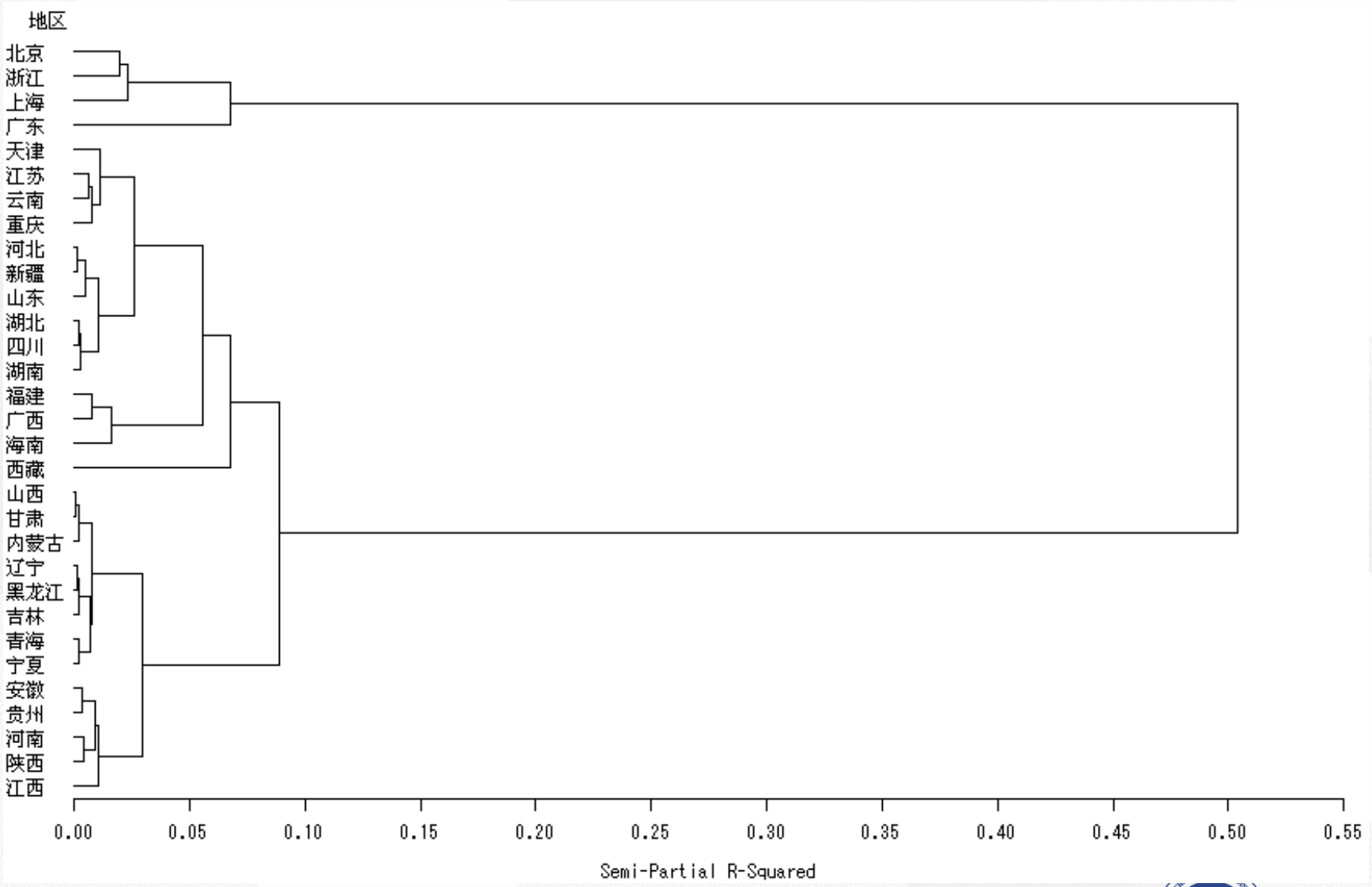


图6.3.14　离差平方和法

- 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，它将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沿海地区。

  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较落后地区，城镇居民的消费水平也是较低的。

- 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（书中介绍的）八种系统聚类法的递推公式统一为：

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

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6.3.15列出了上述八种方法四个参数的取值。

表6. 3. 15　　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	$\frac{1}{2}$
中间距离法	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0
可变法	$(1-\beta)/2$	$(1-\beta)/2$	$\beta(<1)$	0
类平均法	n_K/n_M	n_L/n_M	0	0
可变类平均法	$(1-\beta)n_K/n_M$	$(1-\beta)n_L/n_M$	$\beta(<1)$	0
重心法	n_K/n_M	n_L/n_M	$-\alpha_K\alpha_L$	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

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$D_1 \leq D_2 \leq D_3 \leq \cdots$，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、可变法、类平均法、可变类平均法和离差平方和法都具有单调性，但中间距离法和重心法不具有单调性。

八、类的个数

- 如果能够分成若干个很分开的类，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。

8.1 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。

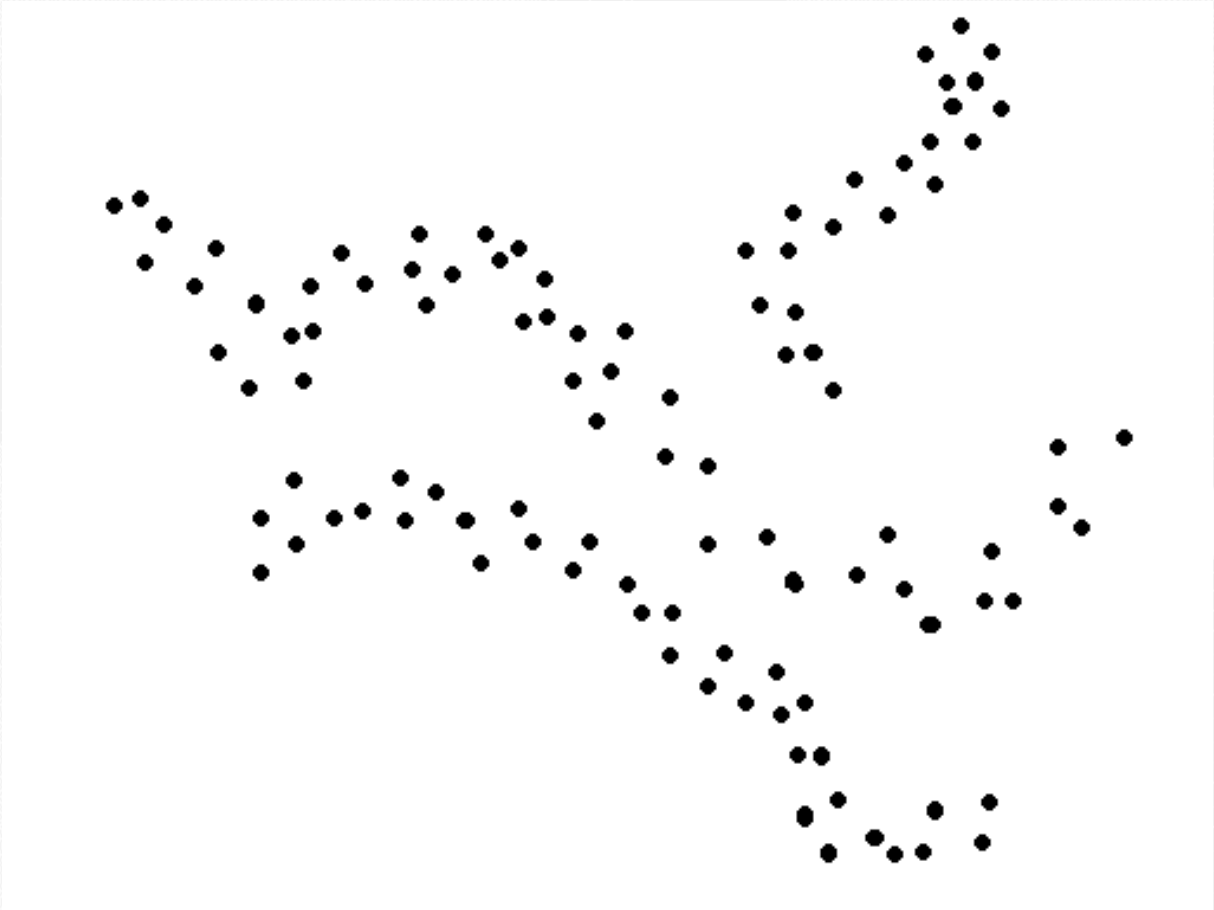
8.2 观测样品的散点图

- 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来确定类的个数。对于三个变量，可使用SAS软件的交互式数据分析菜单系统通过旋转三维坐标轴从各个角度来观测散点图。
- 如果变量个数超过三个，则可对每一可能考虑的聚类结果分别使用费希尔判别法进行降维，将所有样品的前两个或三个判别式得分制作成散点图，观测类之间是否分离得较好以决定分几类较为合适。

从散点图中进行主观聚类

- 观测散点图还有一个重要的用途，就是从直觉上来判断所采用的聚类方法是否合理，甚至有时直接从散点图中进行主观的分类，效果也许会好于正规的聚类方法，特别是在寻找“自然的”类方面。

寻找“自然的”类



8.3 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。

4 动态聚类法

- 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可以事先指定，也可以在聚类过程中确定。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。
- 动态聚类法有许多种方法，本节中，只讨论一种比较流行的动态聚类法——k均值法。k均值法是由麦奎因（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)对除凝聚点之外的所有样品逐个归类，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9.4.1　对例9.3.1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

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

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

由于1到$G_1^{(0)}$的距离小于到$G_2^{(0)}$的距离，因此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9.4.1。
- 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

类 \ 样品	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。

重庆工商大学

CHONGQING TECHNOLOGY AND BUSINESS UNIVERSITY