第八章 聚类和判别分析

聚类分析
（Cluster Analysis）

§1 概述

一、引言：俗话说“物以类聚、人以群分”，对客观事物分门别类地进行考察，大大地简轻了人的认知负荷，并促进知识的专门化和系统化。可以说，分类是人类经验积累和知识发展的必然结果。

在生物、医学、社会、经济、管理、地址、人口、考古、教育等众多领域都存在着大量的分类问题。过去的很长时间里，人们主要靠专业知识和经验做定性分析和处理。随着科技进步和社会发展，对研究对象的分类变得愈来愈复杂和细致，需要考虑的变量和因素越来越多，很多时候仅凭直观知识经验已很难做出科学、准确的分类。

例如， 对10位应聘者做智能测验。3项指标X，Y和Z分别表示数学推理能力，空间想象能力和语言理解能力。其得分如下，请选择合适的方法对应聘者进行分类。

应聘者	1	2	3	4	5	6	7	8	9	10
X	28	18	11	21	26	20	16	14	24	22
Y	29	23	22	23	29	23	22	23	29	27
Z	28	18	16	22	26	22	22	24	24	24

数学方法被逐渐引入到分类学中，形成了数值分类学。近年来，随着数理统计的发展，多元分析技术被引入数值分类学中，就产生了聚类分析这一分支，并在实践中得到广泛的应用。聚类分析同时也是数据发掘（**data mining**）和知识发现的重要途径之一。

二、聚类分析的作用：

- Q型聚类：对样品聚类，使我们的对样品间亲疏关系的认识深化。
- R型聚类：对变量（指标）聚类，简化指标体系。

《红楼梦》的作者

众所周知，《红楼梦》一书共120回，一般认为前80回为曹雪芹所写，后40回为高鹗所续，长期以来对这个问题一直有争议。

1985、1986复旦大学李贤平教授带领他的学生作了这项有意义的工作，他们创造性想法是将120回看成是120个样本，然后确定与情节无关的虚词作为变量，数出每一回单位文字里变量出现的次数，作为数据，用多元分析中的聚类分析法进行分类，果然将120回分成两类即前80回为一类，后40回为一类，很形象地证实了不是出自同一人的手笔。

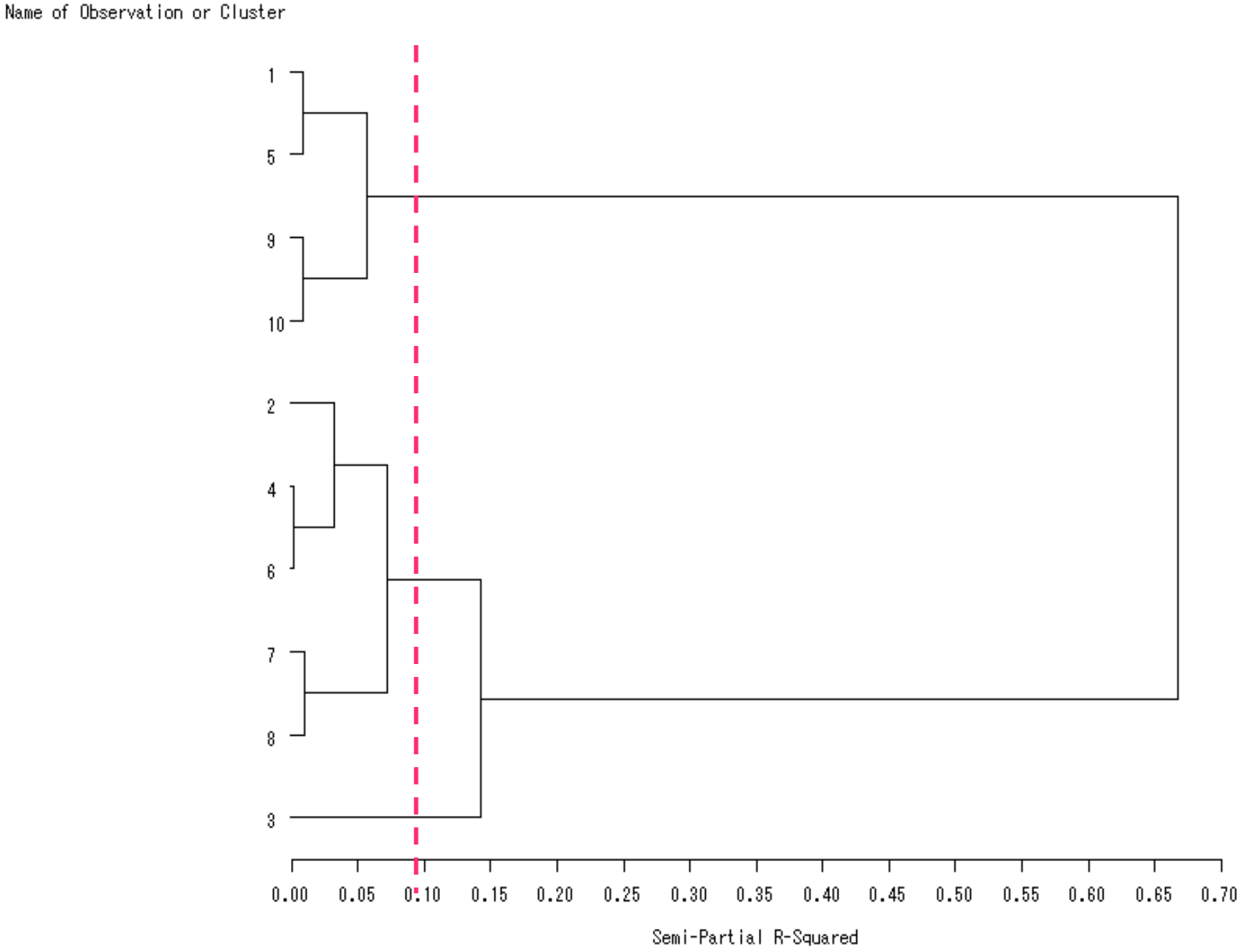
- 此后又进一步分析前80回是否为曹雪芹所写?这时又找了一本曹雪芹的其它著作，做了类似计算，结果证实了用词手法完全相同，断定为曹雪芹一人手笔。这个论证在红学界轰动很大，多元统计分析方法的结果支持了红学界观点，使红学界大为赞叹。

❖ 类似地,关于新大陆的发现者哥伦布的籍贯是意大利的佛罗伦萨还是西班牙的加泰罗尼亚，争议不断。统计学和语言学的结合，为这一论争做出了令人信服的裁断。

三、主要的聚类分析方法

- 系统聚类(Hierarchical clustering)：直观。
- 快速聚类(K-mean clustering):快速，动态。
- 模糊聚类：模糊数学方法、处理定性数据

- 下面是用系统聚类法对前述例子（X、Y、Z） 进行聚类所得的谱系图（分类树）。


Name of Observation or Cluster
1
5
9
10
2
4
6
7
8
3
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
Semi-Partial R-Squared

思考：

聚类分析的基本思想是先要找到一组指标，然后根据观测对象在这一系列指标上的取值，计算出它们之间的亲疏关系（相似性程度），并根据这种亲疏关系逐次地将对象纳入到特定的类别中。

回忆前例谱系图

然则样品或变量间的亲疏关系如何度量？

§2 相似性的度量

研究样品或变量间的亲疏程度/相似性的数量指标有两种：

一种叫**距离（多用于Q型聚类）**，它是将每一个样品看作p维空间的一个点，并测量点与点之间的距离，距离较近的归为一类，距离较远的点应属于不同的类；

另一种叫**相似系数（多用于R型聚类）**，性质越接近的变量，其相似系数越接近于1或—l，而彼此关系越弱的变量，其相似系数越接近于0，相似系数高的为一类，相似系数低的归为不同类。

一、常用距离

1、明氏距离

设 $\mathbf{x_i} = \left(x_{i1}, \quad x_{i2}, \quad \cdots, \quad x_{ip}\right)'$ 和 $\mathbf{x_j} = (x_{j1}, x_{j2}, \cdots, x_{jp})'$ 是第i和 j 个样品的观测值，则二者之间的明考斯基距离为：

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

特别 地：

g = 1时，街区距离 $d_{ij} = \sum_{k=1}^{p} | x_{ik} - x_{jk} |$

g = 2时，欧氏距离 $d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$

g = ∞时，切比雪夫距离 $d_{ij} = \max_{1 \le k \le p} | x_{ik} - x_{jk} |$

明考斯基距离的不足：

①明氏距离的值与各指标的量纲有关，各变量计量单位的不同不仅使此距离的实际意义难以说清，而且，任何一个变量计量单位的改变都会使此距离的数值改变，这使得距离值依赖于各变量计量单位的选择。

②明氏距离的定义没有考虑各个变量之间的相关性和重要性。实际上，明考夫斯基距离是把各个变量都同等看待，将两个样品在各个变量上的离差简单地进行了综合。

改进方法（1）数据变换（2）选用其他距离

2、马氏距离

这是印度著名统计学家马哈拉诺比斯(Mahalanobis)所定义的一种距离，其计算公式为：

$$d_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)'\Sigma^{-1}(\mathbf{x}_i - \mathbf{x}_j)$$

其中，$\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})'$ $\mathbf{x}_j = (x_{j1}, x_{j2}, \cdots, x_{jp})'$

Σ 表示观测变量之间的协方差短阵。在实践应用中，若总体协方差矩阵Σ未知，则可用样本协方差矩阵作为估计代替计算。

马氏距离又称为广义平方距离或统计距离，它不仅考虑了观测变量之间的相关性，而且也考虑到了各个观测指标取值的差异程度。

二、相似系数

1、相关系数

设 $\mathbf{x}_i=(x_{1i},\ x_{2i},\ \cdots,\ x_{ni})'$ 和 $\mathbf{x_j}=(x_{1j},x_{2j},\cdots,x_{nj})'$ 是第 i 和 j 个变量的观测值，则二者间皮尔逊相关系数的计算公式为:

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

2、夹角余弦

夹角余弦时从向量集合的角度所定义的一种测度变量之间亲疏程度的相似系数。设在n维空间的向量X_i、X_j，其夹角余弦 C_{ij}的计算公式为：

$$\mathbf{x}_i = \left(x_{1i}, x_{2i}, \cdots, x_{ni}\right)' \quad \mathbf{x}_j = \left(x_{1j}, x_{2j}, \cdots, x_{nj}\right)'$$

$$c_{ij} = \cos\alpha_{ij} = \frac{\sum_{k=1}^{n} x_{ki}x_{kj}}{\sqrt{\sum_{k=1}^{n} x_{ki}^2 \sum_{k=1}^{n} x_{kj}^2}}$$

SPSS 常用的距离和相似性测度

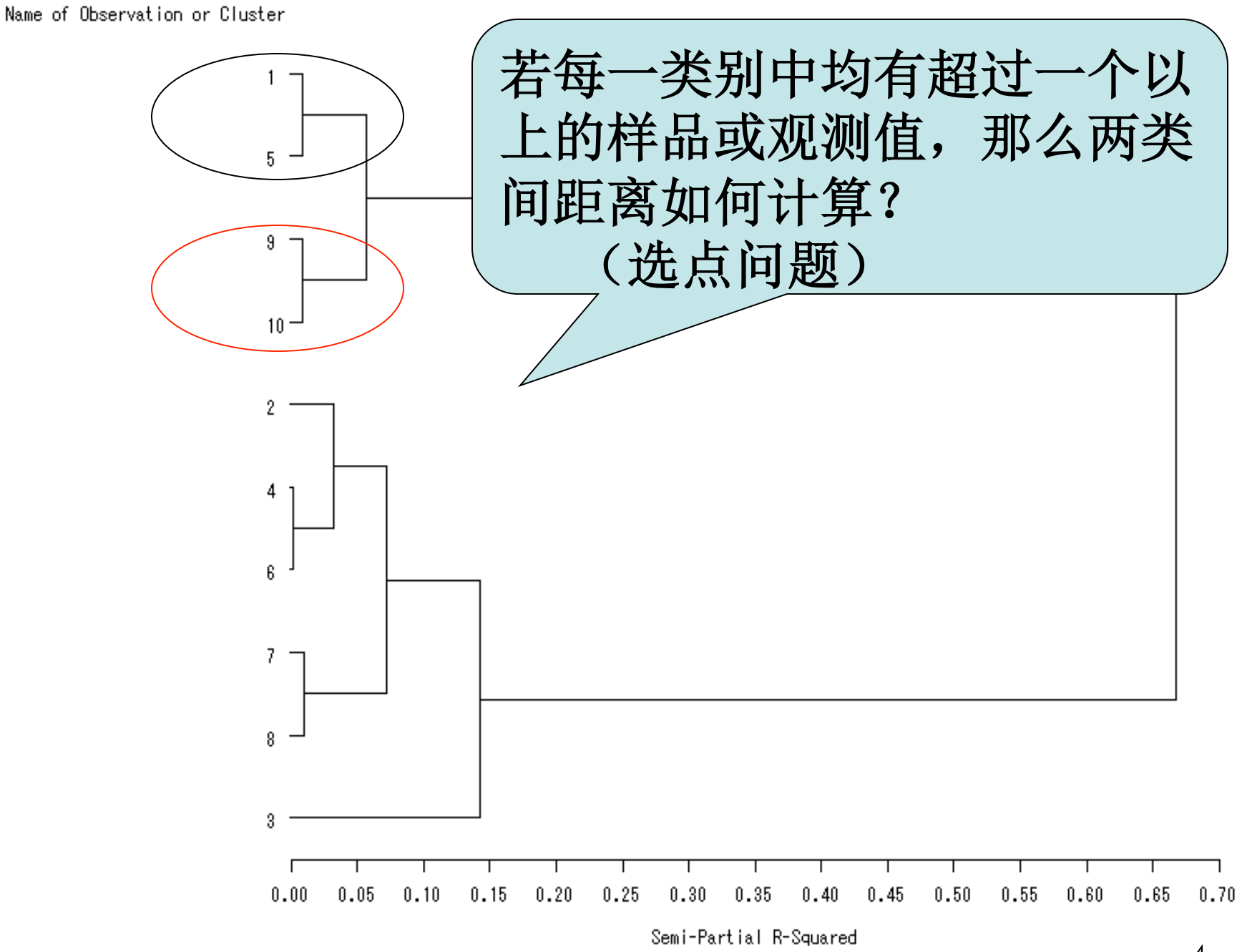
- **Squared Euclidean distance** （欧氏距离平方）
- **Cosine** （夹角余弦）
- **Pearson correlation**（皮尔逊相关）
- **Chebychev**（切比雪夫距离）
- **Block**（街区距离）
- **Minkowski**（明考斯基距离）
- **Customized**（自定义）

非负性 / 对称性 / 三角不等式

至此，我们已经可以根据所选择的距离构成样本点间的距离表,样本点之间被连接起来。

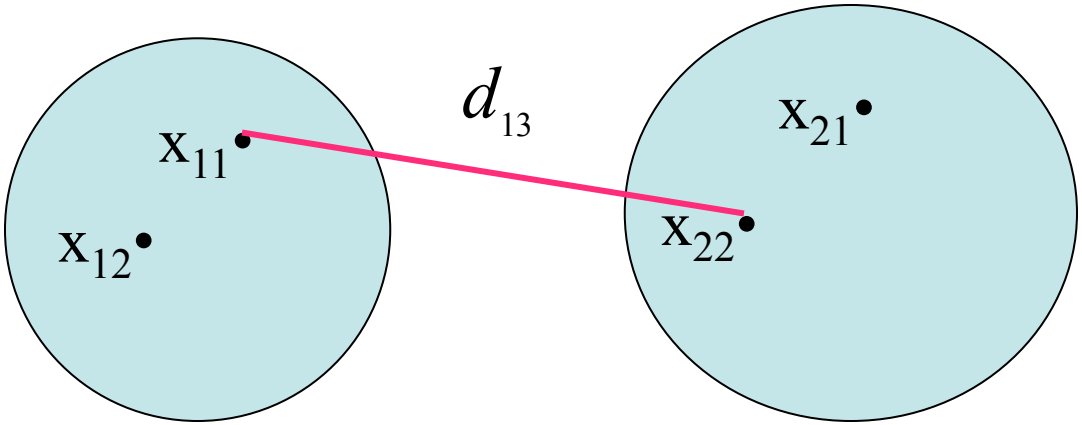
	G_1	G_2	…	G_n
G_1	0	d_{12}	…	d_{1n}
G_2	d_{21}	0		d_{2n}
⋮	⋮	⋮		⋮
G_n	d_{n1}	d_{n2}	…	0

回忆我们的第一个例子，选用一种距离测度，我们很容易地能算出**10**个应聘者之间的距离。

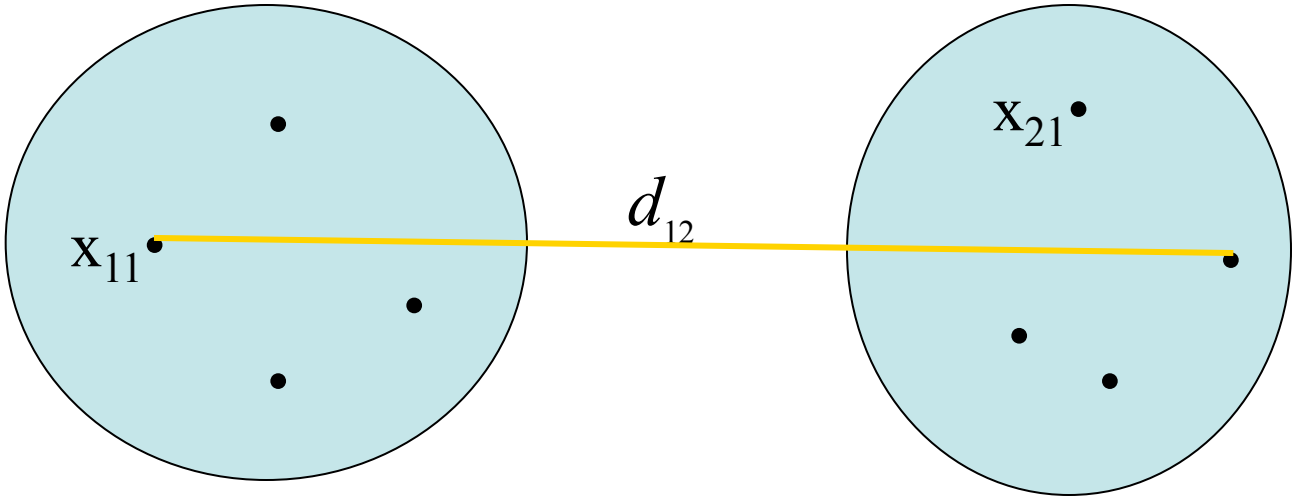

Name of Observation or Cluster
1
5
9
10
2
4
6
7
8
3
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
Semi-Partial R-Squared
若每一类别中均有超过一个以上的样品或观测值，那么两类间距离如何计算？
（选点问题）

三、类间的距离度量

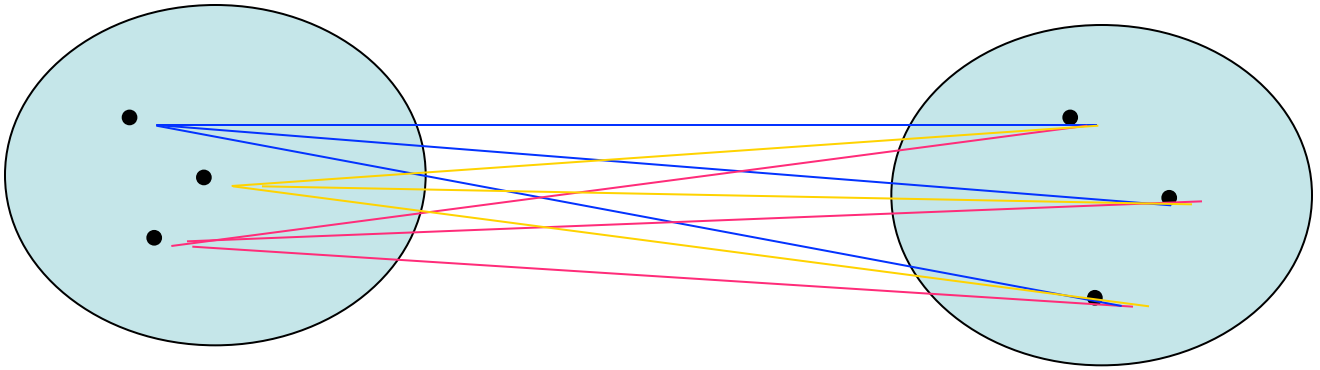
1 、最近邻法（Nearest Neighbor)（Single Linkage）



2、最远邻法（Farthest Neighbor ）（Complete Linkage）



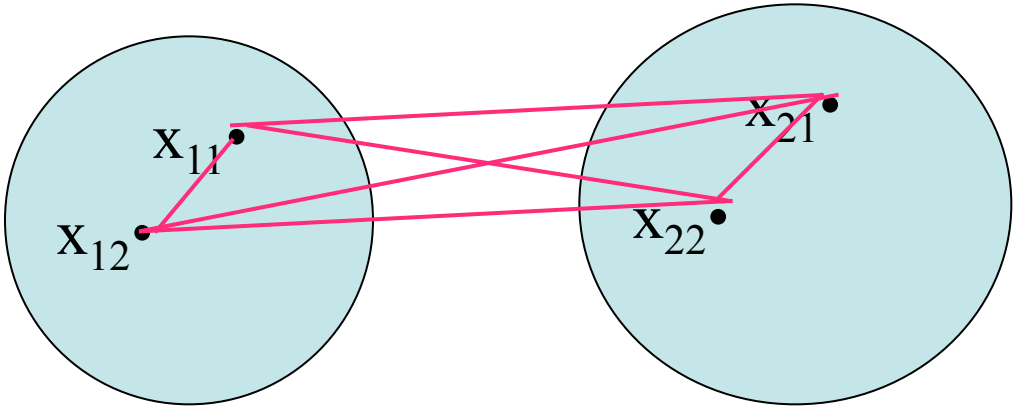
3、组间平均连接（Between-group Linkage)



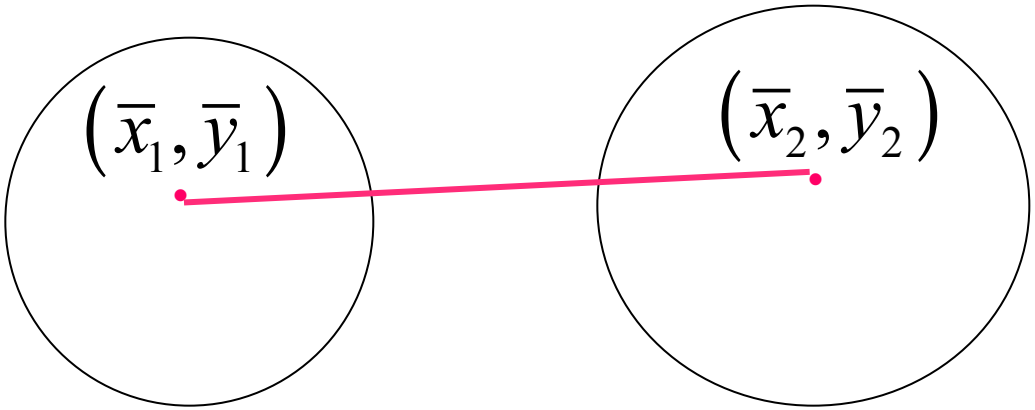
$$\frac{d_1 + \cdots + d_9}{9}$$

4、组内平均连接法（Within-group Linkage)

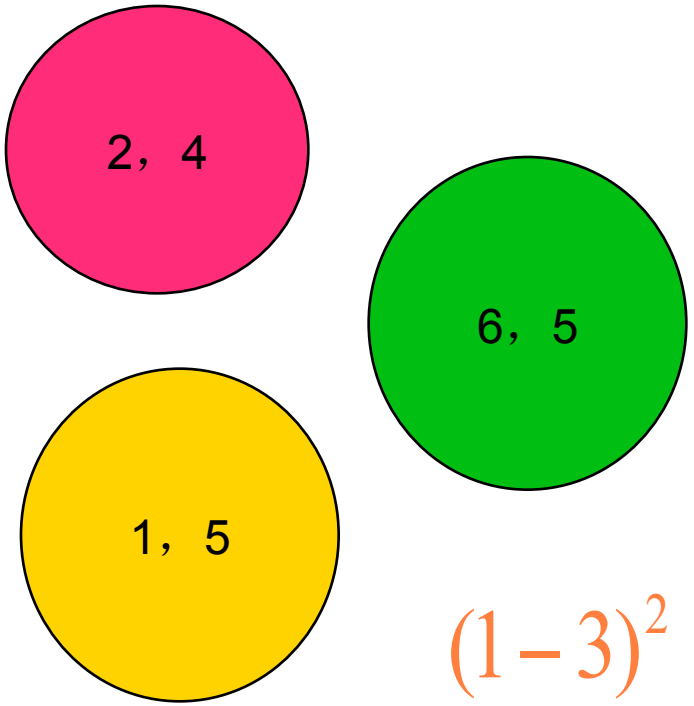
$$\frac{d_1 + d_2 + d_3 + d_4 + d_5 + d_6}{6}$$



5、重心法（Centroid clustering)



6 离差平方和法连接（Ward Method）



$$(2-3)^2+(4-3)^2=2$$

$$(6-5.5)^2+(5-5.5)^2=0.5$$

$$(1-3)^2+(5-3)^2=8$$

（1）红绿（2，4，6，5）8.75

离差平方和增加8.75－2.5＝6.25

（2）黄绿（6，5，1，5）14.75

离差平方和增加14.75－8.5＝6.25

（3）黄红（2，4，1，5）10

离差平方和增加10－10＝0

故按该方法的连接和黄红首先连接。

Clustering Method of SPSS

- **Between groups linkage**（组间连接）
- **Within groups linkage**（组内连接）
- **Nearest neighbor**（最近邻法）
- **Furthest neighbor**（最远邻法）
- **Centroid clustering**（重心法）
- **Median clustering**（中数法）
- **Ward’s Method**（瓦尔德法）

§3 系统聚类

一、系统聚类的基本过程

- 当分类对象和分类变量很多的时候，即使世界上最大、运行速度最快的计算机，也难以在较短的时间内考察所有可能的分类情况。由于这个原因，人们提出了各种各样的聚类算法，以期能在较短的时间内找到合理的分类。

- 系统聚类又称分层聚类**(Hierarchical Clustering)**,是应用最广泛的聚类算法，它通过逐次的合并（或分割）来完成聚类过程。

我们现在通过一个很简单的例子，来演示系统聚类的过程。设抽取五个样品，每个样品只观测一个变量，它们的观测值1，2，3.5，7，9。样品间关系的测度采用街区距离。

D（1）

	G_1	G_2	G_3	G_4	G_5
G_1	0				
G_2	1	0			
G_3	2.5	1.5	0		
G_4	6	5	3.5	0	
G_5	8	7	5.5	2	0

1、首先计算各样品之间的距离d_{ij}，共有C_5^2个。将所有列表，记为D（1）表，该表是一张对称表。所有的样本点各自为一类。

2、选择D（1）表中最小的非零数，不妨假设 d_{pq}，于是将 G_p 和 G_q 合并为一类，记为 $G_{pq}=\{G_p, G_q\}$。这里 $G_{pq}=G_{12}$。

D（1）

	G_1	G_2	G_3	G_4	G_5
G_1	0				
G_2	1	0			
G_3	2.5	1.5	0		
G_4	6	5	3.5	0	
G_5	8	7	5.5	2	0

最近邻法（Nearest Neighbor）

3、G_1 和 G_2 被聚为新类 G_{12}，重新计算各类之间的距离（最近邻法）得 D（2）：

D（2）

	G_{12}	G_3	G_4	G_5
G_{12}	0			
G_3	1.5	0		
G_4	5	3.5	0	
G_5	7	5.5	2	0

4、D（2）表中最小的非零数是1.5，表明剩余的4类中，G_{12}和G_3的距离最近，应该再合并这两类，所得新类不妨记做G_{123}

最近邻法（Nearest Neighbor）

5、计算G_{123}、G_4、G_5等剩余各类之间的距离（最近邻法）得　D（3）：

D（3）

	G_{123}	G_4	G_5
G_{123}	0		
G_4	3.5	0	
G_5	5.5	2	0

6、D（3）表中最小的非零数是2，表明剩余的3类中，G_4和G_5的距离最近，应该再合并这两类，所得新类记做G_{45}。

最近邻法（Nearest Neighbor）

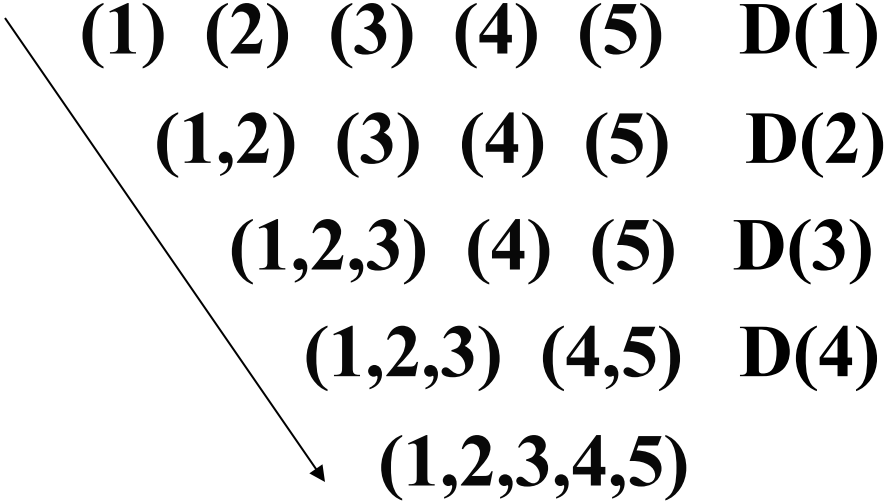
7、计算G_{123}、G_{45}两类之间的距离（最近邻法）得D（4）：

D（4）

	G_{123}	G_{45}
G_{123}	0	
G_{45}	3.5	0

8、D（4）表中最小的非零数是3.5，表明剩余两类G_{123}和G_{45}的距离为3.5。再合并这两类，最后所有样品归为一类，记做G_{12345}。遂告成功。

由以上步骤可知，各步聚类的结果依次为：



(1) (2) (3) (4) (5) D(1)

(1,2) (3) (4) (5) D(2)

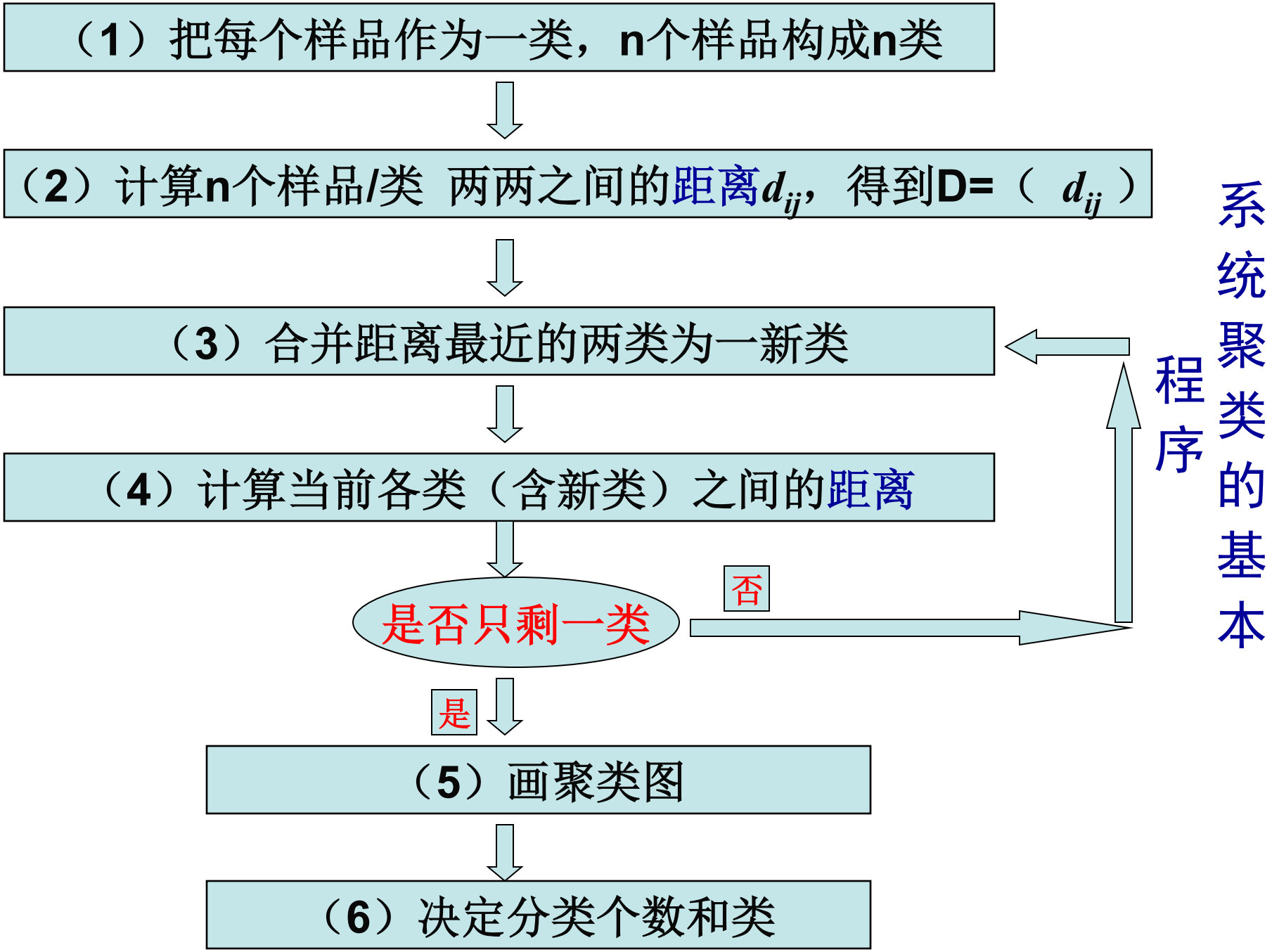
(1,2,3) (4) (5) D(3)

(1,2,3) (4,5) D(4)

(1,2,3,4,5)

注意：我们在这一次的系统聚类中，采用最近邻法来度量类间距离，当然也可以采用最远邻法、重心法、瓦尔德法等。同样，距离本身的计算也可以采用欧氏距离、马氏距离等。 (先选点，再选公式)

系统聚类的基本程序


（1）把每个样品作为一类，n个样品构成n类
（2）计算n个样品/类 两两之间的距离d_{ij}，得到D=（ d_{ij} ）
（3）合并距离最近的两类为一新类
（4）计算当前各类（含新类）之间的距离
是否只剩一类
否
是
（5）画聚类图
（6）决定分类个数和类

二、确定类的个数

在聚类分析过程中类的个数如何来确定才合适呢？这是一个十分困难的问题，人们至今仍未找到令人满意的方法，主要的障碍是对类的结构和内容很难给出一个统一的定义，更多的时候人们需要具体问题具体分析。但是这个问题又是不可回避的，于是一些统计学家提出了开发了若干统计量作为判据，但其原理较复杂。

不过Demirmen提出了根据分类树来进行直观判断的4条准则，值得推荐。

1、各类重心之间的距离必须足够大，即类必须是明显的；

2、各类所包含的数目不应过多；

3、分类的数目应符合使用需要；

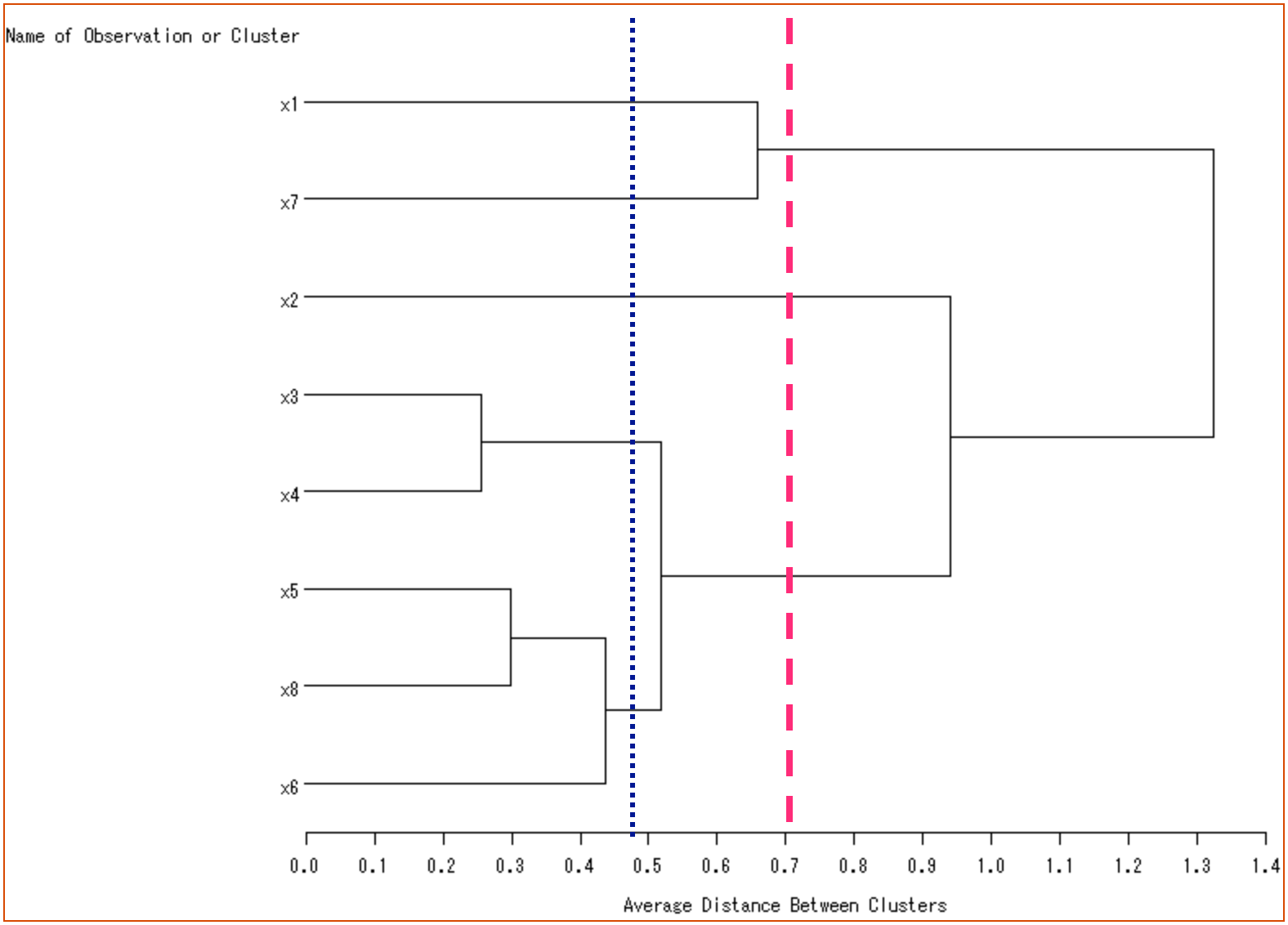
4、若采用不同聚类方法处理，应得到大致类似的结果。

§4 系统聚类应用举例

例：　某公司下属30个企业，公司为了考核下属企业的经济效益，设计了8个指标。为了避免重复，需要对这8个指标进行筛选，建立一个恰当的经济效益指标体系。通过计算30个企业8个指标的相关系数距离，数据是$1-r^2$。得如下表：

	x1	x2	x3	x4	x5	x6	x7	x8
x1	0							
x2	0.60	0						
x3	0.43	0.46	0					
x4	0.47	0.45	0.12	0				
x5	0.57	0.45	0.23	0.22	0			
x6	0.38	0.40	0.21	0.29	0.22	0		
x7	0.31	0.79	0.65	0.70	0.80	0.66	0	
x8	0.45	0.45	0.27	0.23	0.14	0.19	0.77	0

试用系统聚类法将它们聚类。


Name of Observation or Cluster
x1
x7
x2
x3
x4
x5
x8
x6
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
Average Distance Between Clusters

根据美国等20个国家和地区的信息基础设施的发展状况进行分类。

Call—每千人拥有的电话线数；

move—每千人户居民拥有的蜂窝移动电话数；

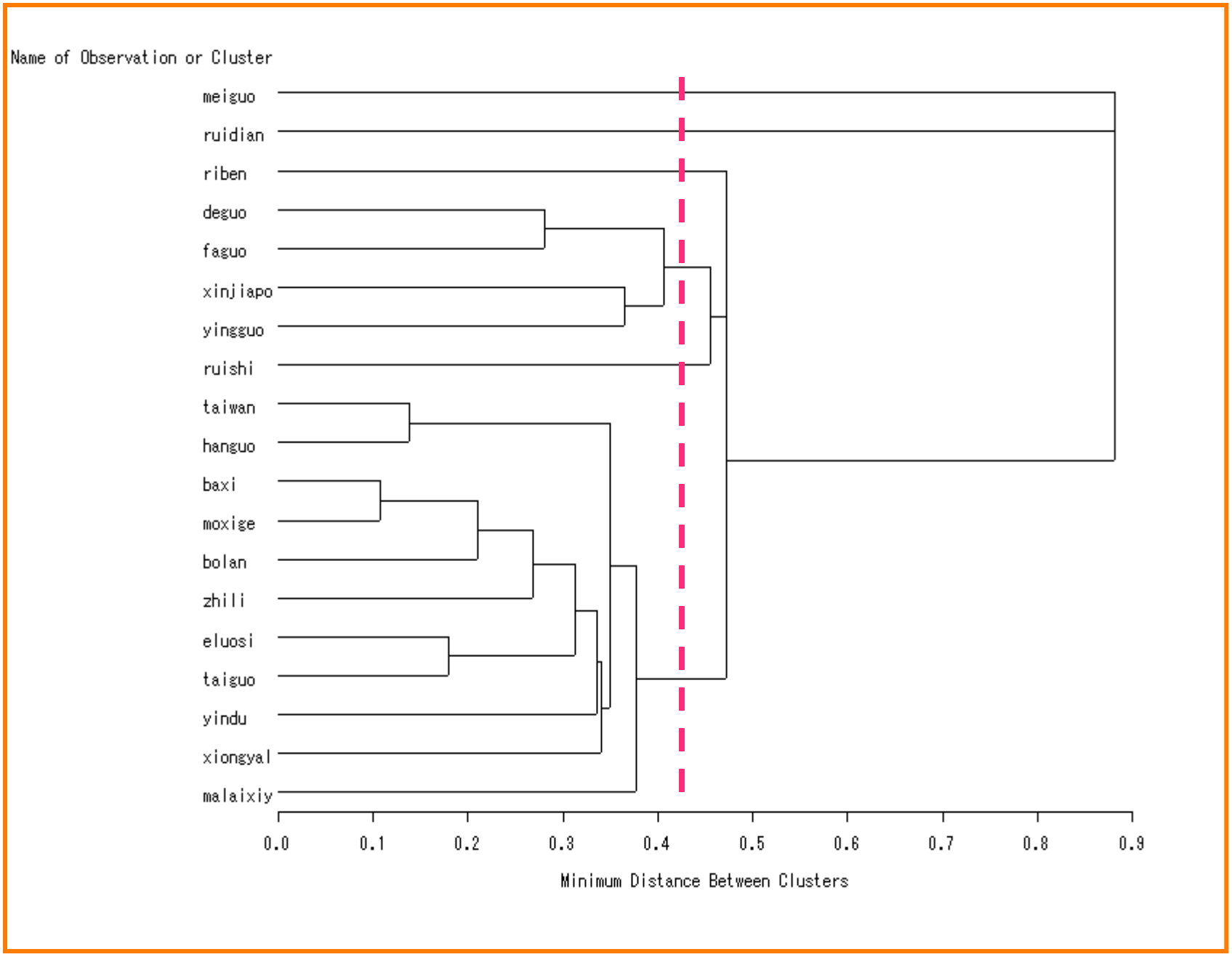
fee—高峰时期每三分钟国际电话的成本；

comp—每千人拥有的计算机数；

mips—每千人计算机功率（每秒百万指令）；

net—每千人互联网络户主数。

国家	call	movel	fee	comp	mips	net
meiguo	631.6	161.9	0.36	403	26073	35.34
riben	498.4	143.2	3.57	176	10223	6.26
deguo	557.6	70.60	2.18	199	11571	9.84
ruidian	684.1	281.8	1.4	246	16660	29.39
ruishi	644	93.5	1.98	234	13621	22.68
xinjiapo	498.4	147.5	2.5	284	13578	13.49
taiwan	469.4	56.1	3.68	119	6911	1.72
hanguo	434.5	73	3.36	99	5795	1.66
baxi	81.9	16.3	3.02	19	876	0.52
zhili	138.6	8.20	1.4	31	1411	1.28
moxige	92.2	9.8	2.61	31	1751	0.35
eluosi	174.9	5	5.12	24	1101	0.48
bolan	169	6.5	3.68	40	1796	1.45
xiongyali	262.2	49.4	2.66	68	3067	3.09
malaixiya	195.5	88.4	4.19	53	2734	1.25
taiguo	78.6	27.8	4.95	22	1662	0.11
yindu	13.6	0.30	6.28	2	101	0.01
faguo	559.1	42.9	1.27	201	11702	4.76
yingguo	521.10	122.5	0.98	248	14461	11.91


Name of Observation or Cluster
meiguo
ruidian
riben
deguo
faguo
xinjiapo
yingguo
ruishi
taiwan
hanguo
baxi
moxige
bolan
zhili
eluosi
taiguo
yindu
xiongyal
malaixiy
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Minimum Distance Between Clusters

判 别 分 析
（Discriminant Analysis）

§1 判别分析概述

一、概念

聚类分析：根据一些特征，计算观测样品或变量间的亲疏关系，并据此对它们分类或分组。

判别分析：根据已知的分类或分组情况，建立分类的规则，并据此预测未知类属的新样品所属的类或组。

同聚类分析一样，判别分析也是一种重要的多元分析技术和数据发掘方法，在科学研究和社会经济领域有着广泛的应用。

二、应用举例

中小企业的破产模型

为了研究中小企业的破产模型，选定4个经济指标：

X1总负债率（现金收益/总负债）

X2收益性指标（纯收入/总财产）

X3短期支付能力（流动资产/流动负债）

X4生产效率性指标（流动资产/纯销售额）

对17个破产企业（1类）和21个正常运行企业（2类）进行了调查，得如下资料：

总负债率	收益性指标	短期支付能力	生产效率指标	类别
-.45	-.41	1.09	.45	1
-.56	-.31	1.51	.16	1
.06	.02	1.01	.40	1
-.07	-.09	1.45	.26	1
-.10	-.09	1.56	.67	1
-.14	-.07	.71	.28	1
-.23	-.30	.22	.18	1
.07	.02	1.31	.25	1
.01	.00	2.15	.70	1
-.28	-.23	1.19	.66	1
.15	.05	1.88	.27	1
.37	.11	1.99	.38	1
-.08	-.08	1.51	.42	1
.05	.03	1.68	.95	1
.01	.00	1.26	.60	1
.12	.11	1.14	.17	1
-.28	-.27	1.27	.51	1
.51	.10	2.49	.54	2
.08	.02	2.01	.53	2

.38	.11	3.27	.55	2
.19	.05	2.25	.33	2
.32	.07	4.24	.63	2
.31	.05	4.45	.69	2
.12	.05	2.52	.69	2
-.02	.02	2.05	.35	2
.22	.08	2.35	.40	2
.17	.07	1.80	.52	2
.15	.05	2.17	.55	2
-.10	-1.01	2.50	.58	2
.14	-.03	.46	.26	2
.14	.07	2.61	.52	2
-.33	-.09	3.01	.47	2
.48	.09	1.24	.18	2
.56	.11	4.29	.45	2
.20	.08	1.99	.30	2
.47	.14	2.92	.45	2
.17	.04	2.45	.14	2
.58	.04	5.06	.13	2
.04	.01	1.50	.71	待判
-.06	-.06	1.37	.40	待判

.07	-.01	1.37	.34	待判
-.13	-.14	1.42	.44	待判
.15	.06	2.23	.56	待判
.16	.05	2.31	.20	待判
.29	.06	1.84	.38	待判
.54	.11	2.33	.48	待判

企业序号	判别类型	判别函数得分	判别为1的概率	判别为2的概率
1	1	-.56509	.69479	.30521
2	1	-.89817	.80234	.19766
3	1	-.59642	.70620	.29380
4	1	-1.02182	.83420	.16580
5	2	.25719	.35312	.64688
6	2	.34253	.32005	.67995
7	2	.27925	.34442	.65558
8	2	1.24010	.09012	.90988

为什么我们知道某些观测值属于某个特定总体而对其余的观测值却不能肯定？

1、有时我们要根据经验，对未来的情形进行预测，而真正的结果要将来才能确定；

2、“完备的”信息可能要毁坏对象；

3、直接的“确定性”信息不可获得，或者即使能获得，但成本过于昂贵。

其他应用举例

1、根据对矿石样品的分析，判断某矿场蕴藏矿石的级别；

2、根据一组生理指标，确定某儿童所患的脑膜炎是由细菌还是病毒所致；

3、根据企业的财务信息，确定其经营状况、信用等级甚至是否即将破产；

4、根据对研究生的学业成绩记录、入学考试成绩及面试情况，预测其能否获得学位；

5、根据笔迹分析的情况，判定某地发现的《兰亭序》摹本是米芾还是董其昌的手笔。

三、 判别分析的基本思想

判别分析的主要目的是为了识别个体所属的类别。其基本思想是根据已掌握的、历史上每个类别的若干样本的数据信息，总结出客观事物分类的规律性，建立判别公式和判别准则。然后，当遇到新的样本点时，就根据总结出来的判别公式和判别准则，判断该样本点所属的类别。

？ 怎样建立判别准则

§2 距离判别

距离判别最直观的想法是计算样品到各类总体重心（平均数）的距离，哪个距离最小就将它判归哪个总体，所以，我们首先考虑的是是否能够构造一个恰当的距离，通过比较样本与各类别之间距离的大小，判别其所属类别。

我们学过的距离有哪些？

一、距离的计算

设 $\mathbf{x}=(x_1,x_2,\cdots,x_m)'$ 和 $\mathbf{y}=(y_1,y_2,\cdots,y_m)'$ 是从期望 $\mu=(\mu_1,\mu_2,\cdots,\mu_m)'$ 和方差阵 $\Sigma=(\sigma_{ij})_{m\times m}>0$ 的总体G抽得的两个观测值，则

X与Y之间的马氏距离

$$d^2(\mathbf{x},\mathbf{y})=(\mathbf{x}-\mathbf{y})'\Sigma^{-1}(\mathbf{x}-\mathbf{y})$$

样本X和G_i类之间的马氏距离定义为X与G_i类重心间的距离：

$$d^2(\mathbf{x},G_i)=(\mathbf{x}-\mu_i)'\Sigma^{-1}(\mathbf{x}-\mu_i)\qquad i=1,2,\cdots,k$$

马氏距离和欧氏距离之间的差别

马氏距离 $d^2(\mathbf{x},G)=(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$

欧氏平方距离 $d^2(\mathbf{x},G)=(\mathbf{x}-\boldsymbol{\mu})'(\mathbf{x}-\boldsymbol{\mu})$

马氏距离的特点：

1、马氏距离是变量标准化后的欧式距离，故马氏距离不受计量单位的影响;

2、马氏距离考虑了变量间的相关性;

二、两总体的距离判别

设有两个p维总体G_1和G_2，x是一个具有同样维数的样品。一个最直观的想法是，若能计算样品x到总体 G_1和 G_2 的距离 $d(x,G_1)$ 和 $d(x,G_2)$， 则可以将x判归与其距离最近的类别中。该判别准则的数学模型可描述如下：

$$\begin{cases} x\in G_1, & \text{如} d^2(x, G_1) < d^2(x, G_2) \\ x\in G_2, & \text{如} d^2(x, G_2) < d^2(x, G_1) \\ \text{待判}, & \text{如} d^2(x,G_1) = d^2(x,G_2) \end{cases}$$

1、 G_1和G_2协方差相等（$\sum_1=\sum_2$）

当总体为正态总体，且它们的协方差矩阵相同的时候，样品x到两总体重心的马氏距离为

$$d^2(\mathbf{x},G_1)=(\mathbf{x}-\mu_1)'\Sigma_1^{-1}(\mathbf{x}-\mu_1)$$

$$d^2(\mathbf{x},G_2)=(\mathbf{x}-\mu_2)'\Sigma_2^{-1}(\mathbf{x}-\mu_2)$$

当总体不是正态分布时，有时也用马氏距离描述样品x到两总体的远近。若 μ_1、μ_2、$\sum$未知，可用样本数据来估计。

当$\sum_1=\sum_2=\sum$时，可做如下推导：

$$W(\mathrm{x}) = d^2(\mathrm{x}, G_1) - d^2(\mathrm{x}, G_2)$$

$$= (\mathrm{x} - \mu_1)'\Sigma^{-1}(\mathrm{x} - \mu_1) - (\mathrm{x} - \mu_2)'\Sigma^{-1}(\mathrm{x} - \mu_2)$$

令 $\bar{\mu} = \dfrac{\mu_1 + \mu_2}{2}$ $\qquad \alpha = \Sigma^{-1}(\mu_1 - \mu_2) = (a_1, a_2, \cdots, a_p)'$

则有 $W(\mathrm{x}) = -2(\mathrm{x} - \bar{\mu})'\alpha$

当 μ_1、μ_2 和Σ已知时，$\alpha = \Sigma^{-1}(\mu_1 - \mu_2)$ 是一个已知的p维向量，W（x）是X的线性函数，称为线性判别函数，α称为判别系数。用线性判别函数进行判别分析非常直观，使用起来最方便，在实际中的应用也最广泛。

进而，判别法可表示为

$$\begin{cases} x\in G_1, & \text{如} d^2(x, G_1) < d^2(x, G_2) \\ x\in G_2, & \text{如} d^2(x, G_2) < d^2(x, G_1) \\ \text{待判}, & \text{如} d^2(x, G_1) = d^2(x, G_2) \end{cases}$$

$$\begin{cases} \mathrm{x}\in G_1, & \text{如} W(\mathrm{x}) < 0 \\ \mathrm{x}\in G_2, & \text{如} W(\mathrm{x}) > 0 \\ \text{待判}, & \text{如} W(\mathrm{x}) = 0 \end{cases}$$

2、 G_1和G_2协方差不等（$\sum_1 \neq \sum_2$）

无法进一步简化判别函数W(x)，只能直接用

$$W(\mathrm{x}) = d^2(\mathrm{x}, G_2) - d^2(\mathrm{x}, G_1)$$
$$= (\mathrm{x} - \mu_2)' \Sigma_2^{-1} (\mathrm{x} - \mu_2) - (\mathrm{x} - \mu_1)' \Sigma_1^{-1} (\mathrm{x} - \mu_1)$$

作判别函数，它是x的二次函数。

例　在对企业的考核中，可以根据企业的生产经营情况把企业分为优秀企业和一般企业。考核企业经营状况的指标有：

资金利润率

劳动生产率

产品净值率

三个指标的均值向量和协方差矩阵已知。现有二个企业，观测值分别为

（7.8，39.1，9.6）和（8.1，34.2，6.9），问这两个企业应该属于哪一类？

变量	均值向量		协方差矩阵		
	优秀	一般			
资金利润率	**13.5**	**5.4**	**68.39**	**40.24**	**21.41**
劳动生产率	**40.7**	**29.8**	**40.24**	**54.58**	**11.67**
产品净值率	**10.7**	**6.2**	**21.41**	**11.67**	**7.90**

线性判别函数：

$$y = -0.60581x_1 + 0.25362x_2 + 1.83679x_3 - 18.7359$$

$$y_1 = -0.60581 \times 7.8 + 0.25362 \times 39.1 + 1.83679 \times 9.6 - 18.73596$$

$$= 4.0892 > 0$$

$$y_2 = -0.60581 \times 8.1 + 0.25362 \times 34.2 + 1.83679 \times 6.9 - 18.73596$$

$$= -2.2956 < 0$$

二、 多总体的距离判别法

设有K个总体，分别有均值向量μ_i和协方差阵Σ_i (i=1,2,…,k) 。又设x是一个待判样品。则 x 与 G_i 的距离为

$$d^2(\mathbf{y}, G_i) = (\mathbf{y} - \mu_i)'\Sigma_i^{-1}(\mathbf{y} - \mu_i)$$

最直接的考虑是：计算 x 与 所有总体的距离，找出其中的最小者，将 x 归入该类。也可采取与两总体距离判别类似的方法，即构造判别函数和判别准则。

$$V_{ij}(\mathrm{x}) = d^2(\mathrm{x}, G_i) - d^2(\mathrm{x}, G_j)$$

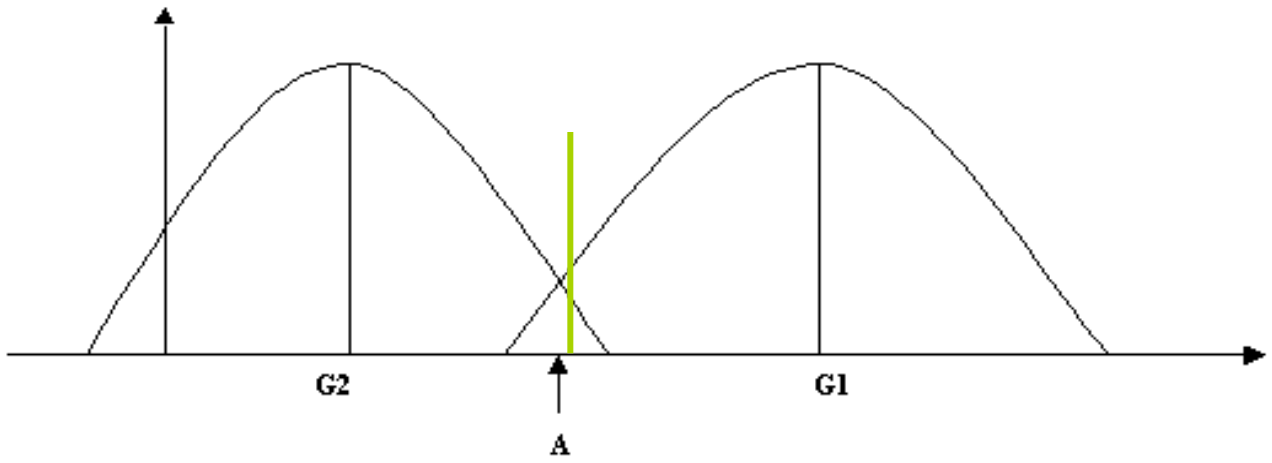
$$= (\mathrm{x} - \mu_i)'\Sigma_i^{-1}(\mathrm{x} - \mu_i) - (\mathrm{x} - \mu_j)'\Sigma_j^{-1}(\mathrm{x} - \mu_j)$$

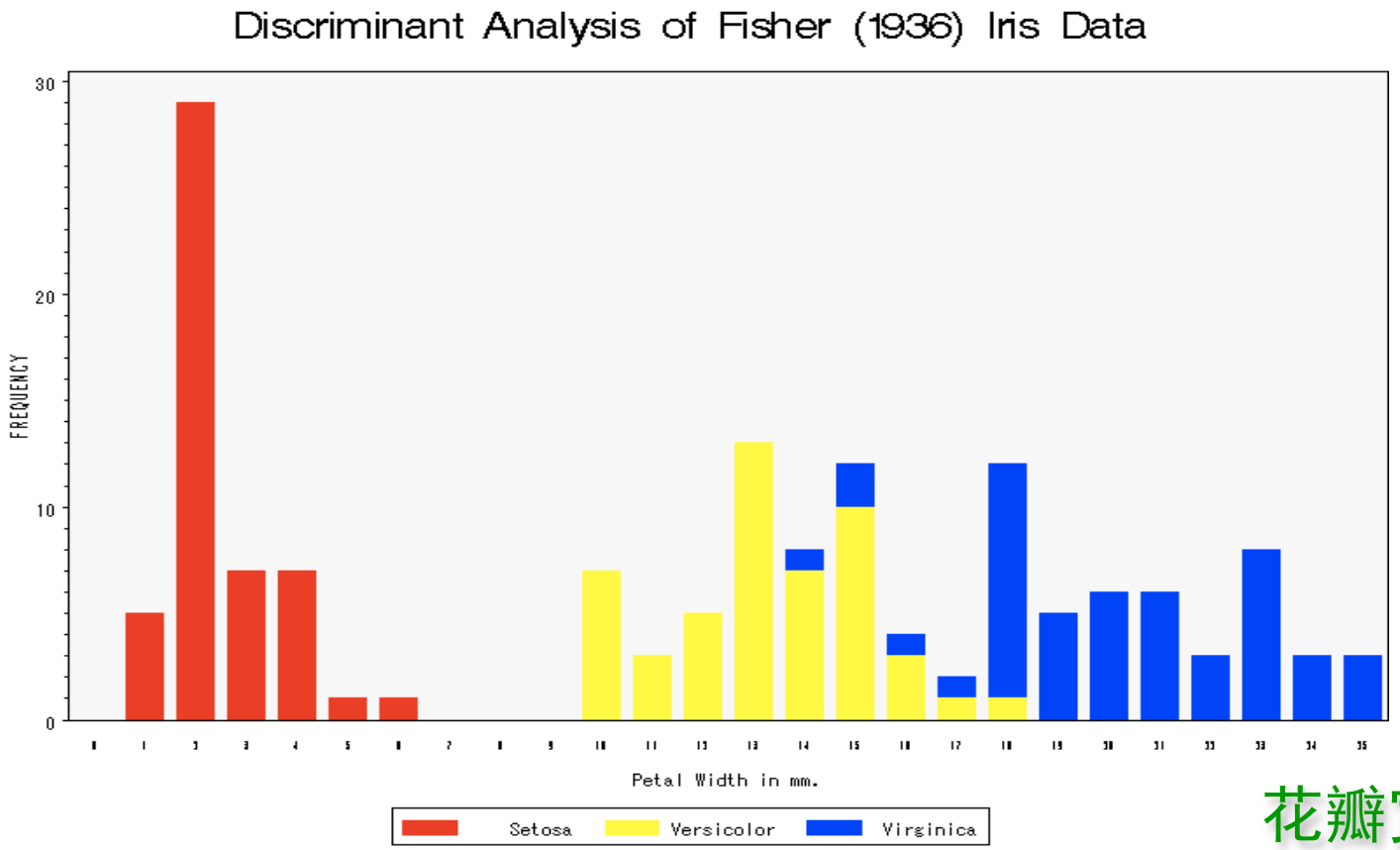
$$\begin{cases} \mathrm{x} \in G_i, \quad \text{若} V_{ij}(\mathrm{x}) < 0, \ \forall j \neq i \\ \text{待判，最短距离不止一个} \end{cases}$$

三、判别函数预测精度评估

由上面的分析可以看出，利用距离进行判别是合理的，但是这并不意谓着不会发生误判。误判率或错分概率就是对判别函数精确性的一种直观度量。下面以一元正态分布为例加以说明。

两总体分别服从 $N(\mu_1,\sigma^2)$ $N(\mu_2,\sigma^2)$

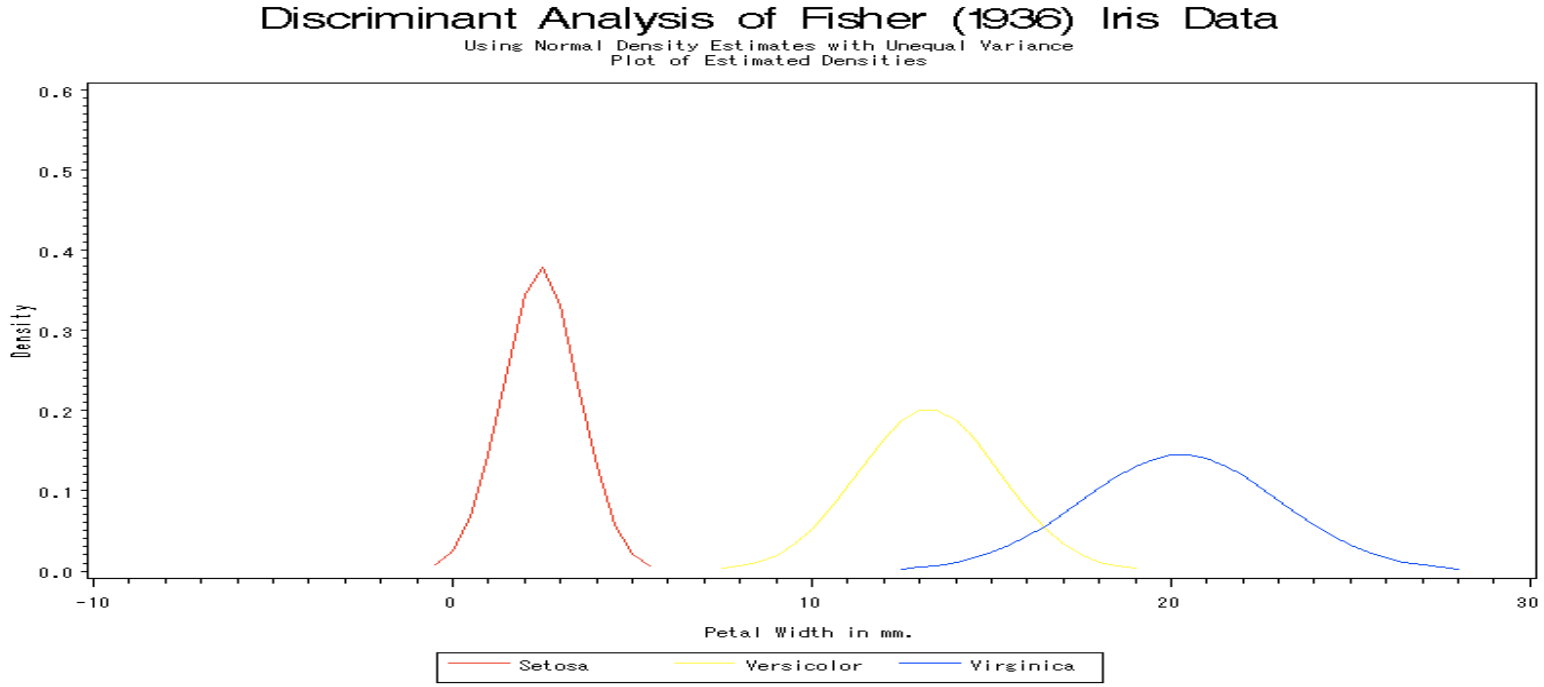


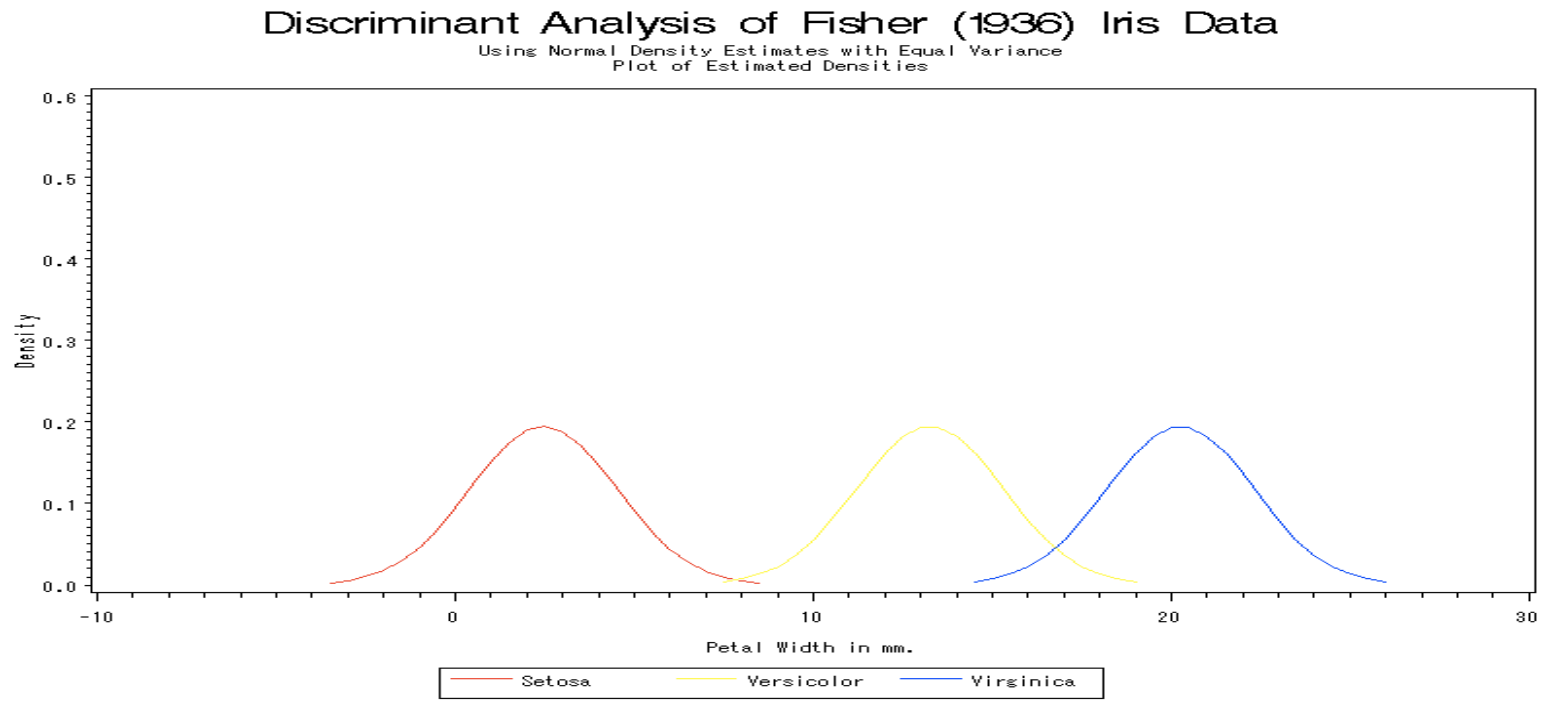


率频

花瓣宽度

刚毛鸢尾　杂色鸢尾　弗吉尼亚鸢尾


Discriminant Analysis of Fisher (1936) Iris Data
Using Normal Density Estimates with Unequal Variance
Plot of Estimated Densities
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
-10
0
10
20
30
Petal Width in mm.
Setosa
Versicolor
Virginica


Discriminant Analysis of Fisher (1936) Iris Data
Using Normal Density Estimates with Equal Variance
Plot of Estimated Densities
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
-10
0
10
20
30
Petal Width in mm.
Setosa
Versicolor
Virginica

若总体分布函数已知，则错分概率或误判率可以直接计算。但通常的情况是，总体分布函数及有关参数均未知，此时只能对误判率进行估计，相对简单的方法主要有以下三种：

简单误判率法

分半法

交叉验证法

简单误判率法

思路：用求得的判别函数对样本数据进行分类，计算其误判率。

特点：计算简便、易于理解，适用于任意判别方法。

判类 / 原类	G_1	G_2	$\cdots$	G_k	合计
G_1	m_{11}	m_{12}	$\cdots$	m_{1k}	n_1
G_2	m_{21}	m_{22}	$\cdots$	m_{2k}	n_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
G_k	m_{k1}	m_{k2}	$\cdots$	m_{kk}	n_k

简单误判率：

$$p = \frac{1}{n}\sum_{i=i}^{k}\sum_{\substack{j=1 \\ j\neq i}}^{k} m_{ij}$$

分半法

简单误判率的计算简便、理解直观。但不幸的是，它倾向于低估实际的误判率。除非每一类别中的样本容量n_i都充分大，否则这一缺陷难以克服。

一种修正的方法是将整个样本数据分为两部分：计算部分和验证部分。计算部分的数据用于构造判别函数；验证部分的数据用于评估判别函数。误判率用验证部分样品被错分的频率来估计。

分半法克服了简单误判率的缺陷，但它需要较大的样本容量，而且被评估的函数不是用全部数据构造的，损失了部分信息。

交叉验证法

交叉验证法又称刀切法、“留一个在外”法。

其基本思想是：

（1）用删除第1个观测的样本数据集计算出判别函数，然后用此判别函数来判别第1个观测。

（2）对每一类别中的每个观测样品（2,3,...）都如此这般。

（3）用总错判次数除以总观测数，作为误判率的估计。

交叉核实检查比较严格，较好地说明判别函数的有效性，即使在样本较小时也适用。通常情况下，交叉验证得到的误判率估计要大于简单误判率，也说明了后者的估计过于乐观。

总负债率	收益性指标	短期支付能力	生产效率指标	类别
-.45	-.41	1.09	.45	1
-.56	-.31	1.51	.16	1
.06	.02	1.01	.40	1
-.07	-.09	1.45	.26	1
-.10	-.09	1.56	.67	1
-.14	-.07	.71	.28	1
-.23	-.30	.22	.18	1
.07	.02	1.31	.25	1
.01	.00	2.15	.70	1
-.28	-.23	1.19	.66	1
.15	.05	1.88	.27	1
.37	.11	1.99	.38	1
-.08	-.08	1.51	.42	1
.05	.03	1.68	.95	1
.01	.00	1.26	.60	1
.12	.11	1.14	.17	1
-.28	-.27	1.27	.51	1
.51	.10	2.49	.54	2
.08	.02	2.01	.53	2

.38	.11	3.27	.55	2
.19	.05	2.25	.33	2
.32	.07	4.24	.63	2
.31	.05	4.45	.69	2
.12	.05	2.52	.69	2
-.02	.02	2.05	.35	2
.22	.08	2.35	.40	2
.17	.07	1.80	.52	2
.15	.05	2.17	.55	2
-.10	-1.01	2.50	.58	2
.14	-.03	.46	.26	2
.14	.07	2.61	.52	2
-.33	-.09	3.01	.47	2
.48	.09	1.24	.18	2
.56	.11	4.29	.45	2
.20	.08	1.99	.30	2
.47	.14	2.92	.45	2
.17	.04	2.45	.14	2
.58	.04	5.06	.13	2
.04	.01	1.50	.71	待判
-.06	-.06	1.37	.40	待判

中小企业破产模型判别函数的效果评估

Classification Results[b,c]

		□ □ □ □ □ 1, 正 □ □ □ □ □ □ 2	Predicted Group Membership		Total
			1	2	
Original	Count	1	15	2	17
		2	5	16	21
		Ungrouped cases	4	4	8
	%	1	88.2	11.8	100.0
		2	23.8	76.2	100.0
		Ungrouped cases	50.0	50.0	100.0
Cross-validated[a]	Count	1	15	2	17
		2	6	15	21
	%	1	88.2	11.8	100.0
		2	28.6	71.4	100.0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 81.6% of original grouped cases correctly classified.

c. 78.9% of cross-validated grouped cases correctly classified.

Ungrouped case 为例中的待判样品

§3 判别分析中的若干问题

一、判别分析的假定

1、当被解释变量为定性变量，而解释变量为定量变量，判别分析是最适合的；有时解释变量是定性变量，也可以经过转化，视为定量变量。

2、已知的两个或多个总体的均值必须存在显著的差异，即总体是分离良好的；否则，错分概率会很高，判别过程是徒劳的；

3、判别变量服从多元正态分布，且各总体的协方差相等，此时判别准则是较为稳健的。如果判别变量不服从正态分布，可对其进行变换。

二、判别变量的选择

判别分析是根据已有样本数据中包含的信息，对新样品进行归类的数学方法，其分类结果与样品在现实中的真实类属可能不尽相同。影响因素：

- 样本容量和样本信息的准确性
- 评价指标体系的充分性（全面无遗漏）
- 评价指标体系的必要性（指标间相关小、区分能力强）

解决方案：逐步判别

变量选择和逐步判别

变量选择是否恰当，是影响判别分析效果的关键因素。如果在某个判别问题中，忽略了有关键影响的变量，则判别函数难有好的效果。而另一方面，如果判别变量个数太多，计算量必然大，会影响估计的精度。特别是当引入了一些判别能力不强的变量时，判别的效果也会大受影响。

逐步判别的思想是：先选择判别力最高的变量进入判别函数中，然后是判别力次高的变量，依此类推。同时，检验已选入变量的判别力有无变化，将判别力低的剔除。直至既无新变量选入，也无已有变量被剔除为止。

逐步判别需较大的样本容量，否则结果不稳。

中小企业的破产模型

在中小企业破产模型的研究中，选定了

X1总负债率（现金收益/总负债）

X2收益性指标（纯收入/总财产）

X3短期支付能力（流动资产/流动负债）

X4生产效率性指标（流动资产/纯销售额）

等4个经济指标，对17个破产企业“1”和

21个运行良好企业“2”进行了调查。

如果这些指标是用来做判别分析的变量，它们之间没有显著性差异是不恰当的，所以要检验（MANOVA）所选择的指标在不同类型企业之间是否有显著的差异。

Dependent Variable: x1 （对X1进行的检验）

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.87466791	0.87466791	16.90	0.0002
Error	36	1.86300840	0.05175023		
Corrected Total	37	2.73767632			

X1在类间有显著性差异

Dependent Variable: x2 （对X2进行的检验）

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.08312077	0.08312077	1.95	0.1710
Error	36	1.53370028	0.04260279		
Corrected Total	37	1.61682105			

X2在类间没有显著性差异

类似地，还可以对X3、X4进行检验

x1,x2,x3,x4均为判别变量

Classification Results[b,c]

		□□□□□1，正□□□□□2	Predicted Group Membership		Total
			1	2	
Original	Count	1	15	2	17
		2	5	16	21
		Ungrouped cases	4	4	8
	%	1	88.2	11.8	100.0
		2	23.8	76.2	100.0
		Ungrouped cases	50.0	50.0	100.0
Cross-validated[a]	Count	1	15	2	17
		2	6	15	21
	%	1	88.2	11.8	100.0
		2	28.6	71.4	100.0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 81.6% of original grouped cases correctly classified.

c. 78.9% of cross-validated grouped cases correctly classified.

x1, x3为判别变量

Classification Results[b,c]

		□ □ □ □ □ 1, 正 □ □ □ □ □ □ 2	Predicted Group Membership		Total
			1	2	
Original	Count	1	15	2	17
		2	4	17	21
		Ungrouped cases	4	4	8
	%	1	88.2	11.8	100.0
		2	19.0	81.0	100.0
		Ungrouped cases	50.0	50.0	100.0
Cross-validated[a]	Count	1	15	2	17
		2	5	16	21
	%	1	88.2	11.8	100.0
		2	23.8	76.2	100.0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 84.2% of original grouped cases correctly classified.

c. 81.6% of cross-validated grouped cases correctly classified.

三、样本容量的确定

- 判别分析对样本量与预测变量个数的比率很敏感。
- 许多研究建议比率为每个预测变量20个观测值。
- 如果比率较低，则判别效果会受到影响。
- 建议最小的样本量是平均每个预测变量不少于5个观测。

四、其他判别方法

- 贝叶斯判别：考虑先验概率和错判损失，
  使期望错分代价最小。
- 费希尔判别：利用类似方差分析的原理，
  最大限度的分离总体。