经济统计分析方法

保险系　孙佳美副教授

本讲义用于《经济统计分析方法》实验课程

统计科学即统计学，是一门研究收集数据、表现数据、分析数据、解释数据，从而认识数量规律的方法论科学。

一位资深的海外统计学家说：统计就和柴、米、油、盐、酱、醋、茶一样，存在的时候并不是很突出，一旦不见了，人生就是黑白的了。

“没有统计，其它科学可以存在，但是很渺小”

一个医生说：“医学若无统计就不是科学”

台湾辅仁大学一教授说“统计即生活，统计即人生”

一留美学者说“统计是20世纪人类最伟大的发现之一”

（三）统计研究的特点

1. 数量性

统计总是用数字作为语言来表述事实。

2. 总体性

研究大量个别事物构成的现象整体的数据

课程简介

- 多元统计分析简称多元分析，是统计学的一个重要分支，也是近三、四十年迅速发展的一个分支。随着电子计算机的普及和软件的发展，信息储存手段以及数据信息的成倍增长，多元分析的方法已广泛应用于自然科学和社会科学的各个领域。国内国外实际应用中卓有成效的成果，已证明了多元分析方法是处理多维数据不可缺少的重要工具，并日益显示出无比的魅力。
- 多元分析在工业、农业、医学、经济学、教育学、体育科学、生态学、地质学、社会学、考古学、环境保护、军事科学、甚至文学中都有广泛应用，足见其应用的深度和广度。

课程简介

- **第一章　绪论**
- §1.1什么是多元统计分析
- §1.2 多元分析解决那些类型的实际问题
- **案例1：**起名为“波澜”恰当吗？
- **案例2：**企业信用评估？
- **案例3：**后40回出自谁的手笔？

课程简介

- 第二章 多元数据图的表示法
- 主要内容：轮廓图、雷达图、调和曲线图、星座图、盒形图、直方图、正态P-P图、Q-Q图
- 第三章　聚类分析
- 聚类分析、距离和相似系数、八种系统聚类分析方法、系统聚类的基本性质
- 第四章 判别分析
- 主要内容：距离判别法、费歇（Fisher）判别法、逐步辨别、贝叶斯（Bayes）判别法

课程简介

- **第五章 主成分分析**
- 主要内容：主成分分析的基本思想、主成分分析的数学模型及几何解释、主成分分析的推导及数学解释、计算实例
- **第六章 因子分析**
- 主要内容：因子分析的基本思想、因子分析的数学模型、因子载荷阵的估计方法和因子旋转、因子得分、计算实例
- **第七章 对应分析**
- 主要内容：对应分析的基本思想、对应分析方法的原理、计算实例

课程简介

- **第八章 相关分析和回归分析**
- 主要内容：相关分析概念及其基本思想、实例；线性回归、曲线回归的模型、统计检验、操作步骤、例子。
- **第九章 Logistc回归分析**
- 主要内容：Logistc回归分析的基本思想、Logistc回归分析的原理、计算实例

教材及主要参考书

- 授课教材：
- 于秀林，任雪松编著．**多元统计分析**，中国统计出版社，1999年8月
- 主要参考书：
- 1.卢文岱主编．SPSS for Windows 统计分析（第3版），电子工业出版社，2007年4月
- 2. 吴喜之编著．统计学：从数据到结论，中国统计出版社，2006年10月
- 3.薛薇编著．SPSS统计分析方法及应用，电子工业出版社，2004年9月
- 4. 王淑芬．应用统计学，北京大学出版社，中国林业出版社，2007年
- 5. 何晓群编著．多元统计分析，中国人民大学出版社，2007年4月

多元统计分析案例

案例1：起名为“ 波澜”恰当吗

案例2：客户信用程度评估

案例3：后40回出自谁的手笔

中美纯水有限公司欲为其新推出的一种纯水产品起一个合适的名字，为此专门委托了当地的策划咨询公司，取了一个名字“波澜”。

一个好的名字至少应该满足两个条件：

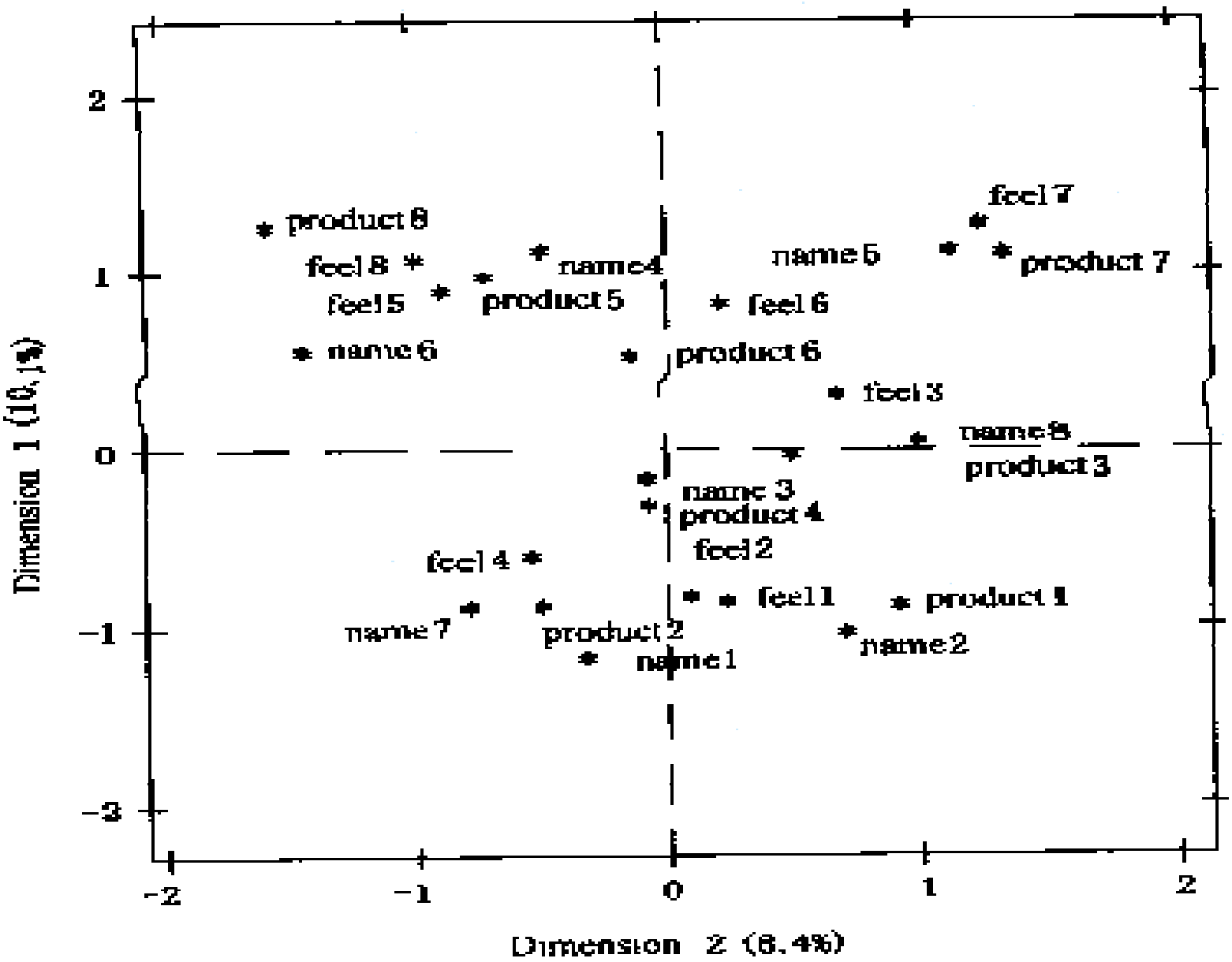
（1）会使消费者联想到正确的产品“纯水”；

（2）会使消费者产生与正确产品密切相关的联想，如“纯净”、“清爽”等。

后来中美纯水有限公司委托调查统计研究所，进行了一次全面的市场研究，在调查中还包括简单的名称测试。

调查的代码和含义如下：

代码	含义	代码	含义	代码	含义
Name1	玉泉	Product1	雪糕	Feel1	清爽
Name2	雪源	Product2	纯水	Feel2	甘甜
Name3	春溪	Product3	碳酸饮料	Feel3	欢快
Name4	期望	Product4	果汁饮料	Feel4	纯净
Name5	波澜	Product5	保健食品	Feel5	安闲
Name6	天山绿	Product6	空调	Feel6	个性
Name7	中美纯	Product7	洗衣机	Feel7	兴奋
Name8	雪浪花	Product8	毛毯	Feel8	高档


product 8
feel 8
name 4
feel 5
product 5
feel 7
name 5
product 7
feel 6
name 6
product 6
feel 3
name 8
product 3
name 3
product 4
feel 2
feel 4
feel 1
product 1
name 7
product 2
name 2
name 1
2
1
0
-1
-2
-2
-1
0
1
2
Dimension 1 (10.1%)
Dimension 2 (8.4%)

由直观图可以看出，“波澜”（Name5）与“洗衣机”（Product7）产品相联系，引起的感觉是“兴奋”，因此“波澜”不是合适的纯净水品牌名称。中美纯水公司的产品是“纯水”（Product2），他们如果想要使该名称给人们一种“纯净”（Feel4）的感觉，那么“中美纯”（Name7）将是最好的商品名称。如果想要使该名称给人们一种“清爽”（Feel1）的感觉，那么“玉泉”（Name1）将是最好的商品名称。中美纯水公司接受了调查统计研究所的建议，没有用“波澜”这个名称，而用了“中美纯”作为品牌的名称。实践证明，它的确是一个成功的品牌名称。

NEXT

HISTORICAL	DATA:								
			Loan Repayment Probability						
Loan				Yrs at	Yrs at	Yrs at	Yrs at		
Record	Monthly	Monthly	Home	Present	Previous	Present	Previous	No. of	
Number	Income	Expenses	Owner?	Job	Job	Address	Address	Depend.	Output
1	3000	1500	0	2	8	6	2	5	3
2	850	425	1	3	3	25	25	1	3
3	1000	3000	0	0.1	0.3	0.1	0.3	4	1
4	9000	2250	1	8	4	5	3	2	5
5	4000	1000	1	3	5	3	2	1	4
6	3500	2500	0	0.5	0.5	0.5	2	1	1
7	2200	1200	1	6	3	1	4	1	3
8	4500	3500	0	8	2	10	1	5	2
9	1200	1000	0	0.5	0.5	1	0.5	3	1
10	800	800	0	0.1	1	5	1	3	1
11	7500	3000	1	10	3	10	3	4	5
12	3000	1000	1	20	5	15	10	1	5
13	2500	700	1	10	5	15	5	3	5
14	3000	2600	1	6	1	3	4	2	2
15	7000	3700	1	10	4	10	1	4	4
16	3000	2800	0	1	2	3	4	3	1
17	4500	1500	1	6	4	4	9	3	4

TEST:									Braincel	
									Advice	Output
1	3000	1500	0	2	8	6	2	5		3
2	850	425	1	3	3	25	25	1		3
3	1000	3000	0	0.1	0.3	0.1	0.3	4		1
4	9000	2250	1	8	4	5	3	2		5
5	4000	1000	1	3	5	3	2	1		4
6	3500	2500	0	0.5	0.5	0.5	2	1		1
7	2200	1200	1	6	3	1	4	1		3
8	4500	3500	0	8	2	10	1	5		2
9	1200	1000	0	0.5	0.5	1	0.5	3		1
10	800	800	0	0.1	1	5	1	3		1
11	7500	3000	1	10	3	10	3	4		5
12	3000	1000	1	20	5	15	10	1		5
13	2500	700	1	10	5	15	5	3		5
14	3000	2600	1	6	1	3	4	2		2
15	7000	3700	1	10	4	10	1	4		4
16	3000	2800	0	1	2	3	4	3		1
17	4500	1500	1	6	4	4	9	3		4

PREDICTION:									
				Yrs at	Yrs at	Yrs at	Yrs at		
	Monthly	Monthly	Home	Present	Previous	Present	Previous	No. of	Braincel
	Income	Expenses	Owner?	Job	Job	Address	Address	Depend.	Advice
New:	2500	1500	0	3	2	3	4	1	1.0460221

NEXT

第二章 多元数据图表示法

本章主要内容

- 主要内容：掌握多元数据的图表示法

  轮廓图、雷达图、调和曲线图、星座图、盒形图、直方图、正态P-P图、Q-Q图
- 做图工具：EXCEL、SPSS

轮廓图 1

作图步骤为：

(1) 作平面坐标系，横坐标取 p 个点表示 p 个变量。

(2) 对给定的一次观测值，在 p 个点上的纵坐标(即高度)和它对应的变量取值成正比。

(3) 连接 p 个高度的顶点得一折线，则一次观测值的轮廓为一条多角折线形。n 次观测值可画出 n 条折线，构成轮廓图。

轮廓图2—例题

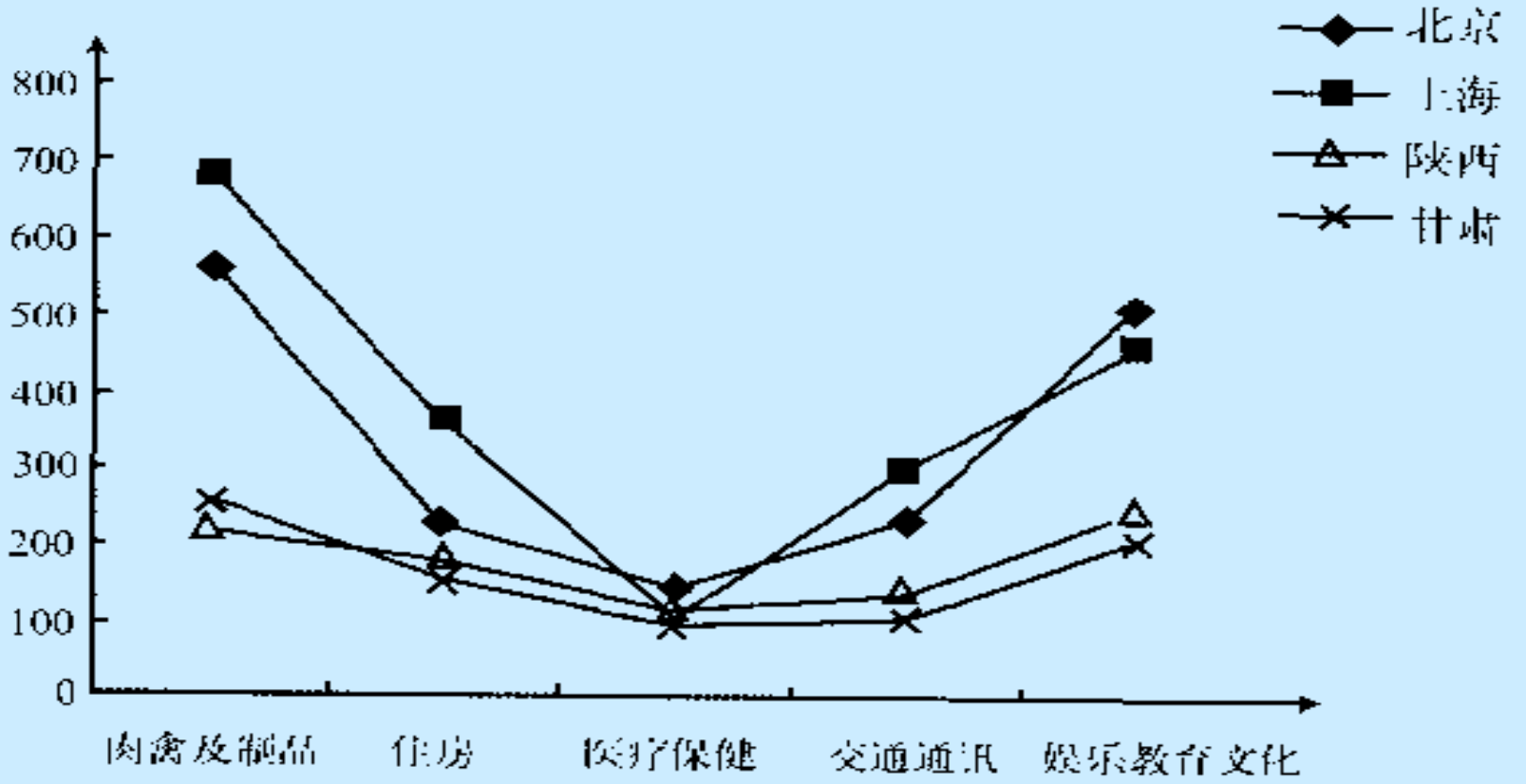
例 考察北京、上海、陕西、甘肃四个省市人均生活消费支出情况，选取以下五项指标，具体数据如下表（摘自1996年中国统计年鉴）：

（单位：元）

	肉禽及制品	住　房	医疗保健	交通和通讯	文娱用品及服务
北京	563.51	227.78	147.76	235.99	510.78
天津	678.92	365.07	112.82	301.46	465.88
陕西	237.38	174.48	119.78	141.07	245.57
甘肃	253.41	156.13	102.96	108.13	212.20

轮廓图3

下图画出四条折线为北京、上海、陕西、甘肃五项指标的数据即四个省市五项指标的轮廓。



雷达图1

作图步骤是：

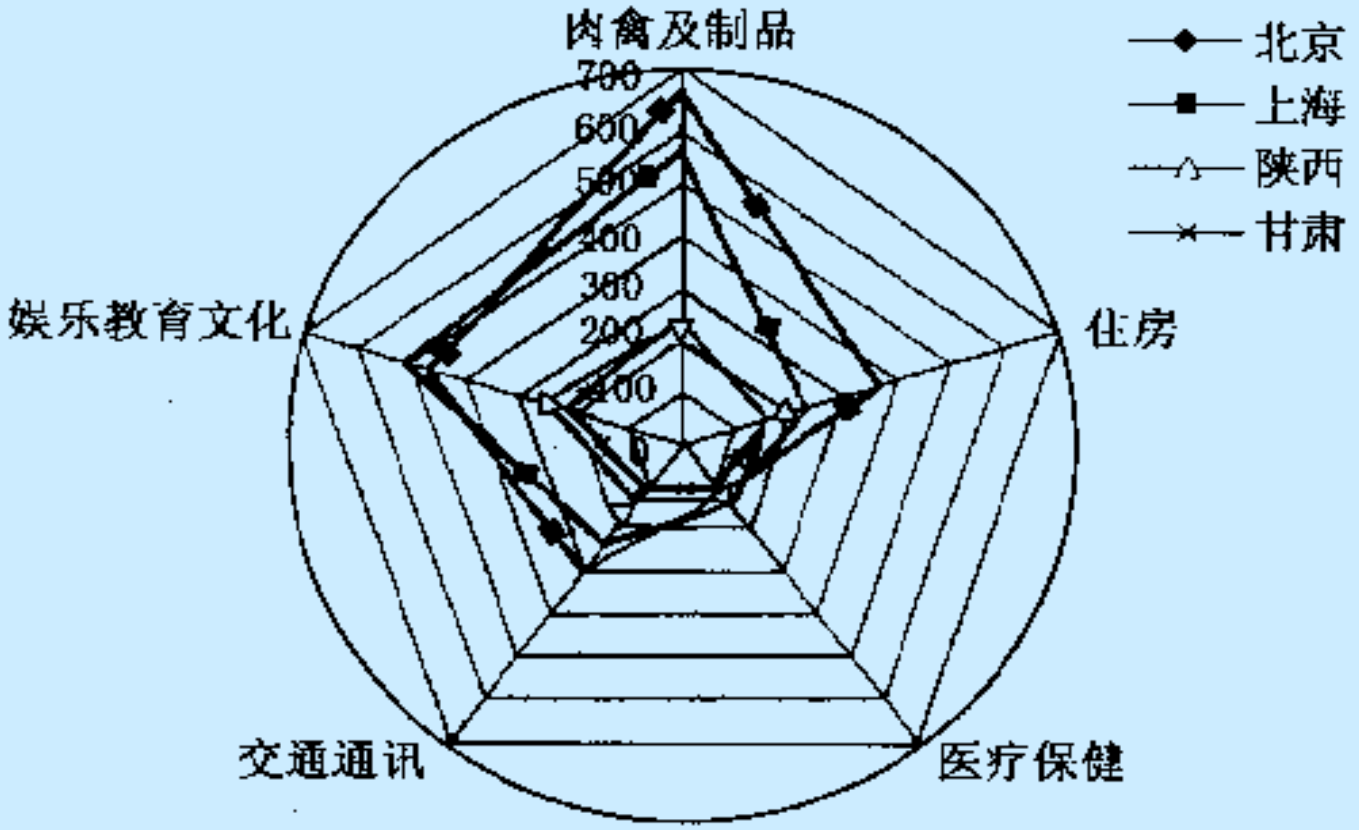
（1）作一圆，并把圆周分为 p 等分。

（2）连接圆心和各分点，把这 p 条半径依次定义为各变量的坐标轴，并标以适当的刻度。

（3）对给定的一次观测值，把它的 p 个分量值分别点在相应的坐标轴上，然后连接成一个 p 边形，这个 p 边形就是 p 元观测值的图示，n 次观测值可画出 n 个 p 边形。

雷达图2

将上例数据用雷达图表示如下：



雷达图3

当观测次数 n 较大时，为使图形清晰，每张图可以只画少数几次观测数据，甚至每张图只画一次观测值。为了获得较好的效果，在雷达图中适当分配变量的坐标轴，并选取合适的尺度是十分重要的，比如把要进行对比的指标其坐标轴分别放在左和右或正上方和正下方，以便根据图形偏左、偏右或偏上、偏下进行对比和分析。

值得注意的是，这里坐标轴只有正半轴，因而只能表示非负数据，若有负数据，只能通过合理变换使之非负才行。

调和曲线图1

调和曲线图是D. F. Andrews1972 年提出的三角多项式作图法，所以又称为三角多项式图，其思想是把高维空间中的一个样品点对应于二维平面上的一条曲线。

设 p 维数据 $X=(x_1,x_2,\cdots,x_p)'$ 对应的曲线是

$$f_x(t)=\frac{x_1}{\sqrt{2}}+x_2\sin t+x_3\cos t+x_4\sin 2t+x_5\cos 2t+\cdots$$

$$-\pi\leqslant t\leqslant\pi$$

上式当 t 在区间（$-\pi$，π）上变化时，其轨迹是一条曲线。

调和曲线图2

上例数据北京、上海、陕西、甘肃分别对应的曲线为：

$$f_1(t)=\frac{563.51}{\sqrt{2}}+227.78\sin t+147.76\cos t+235.99\sin 2t+510.78\cos 2t$$

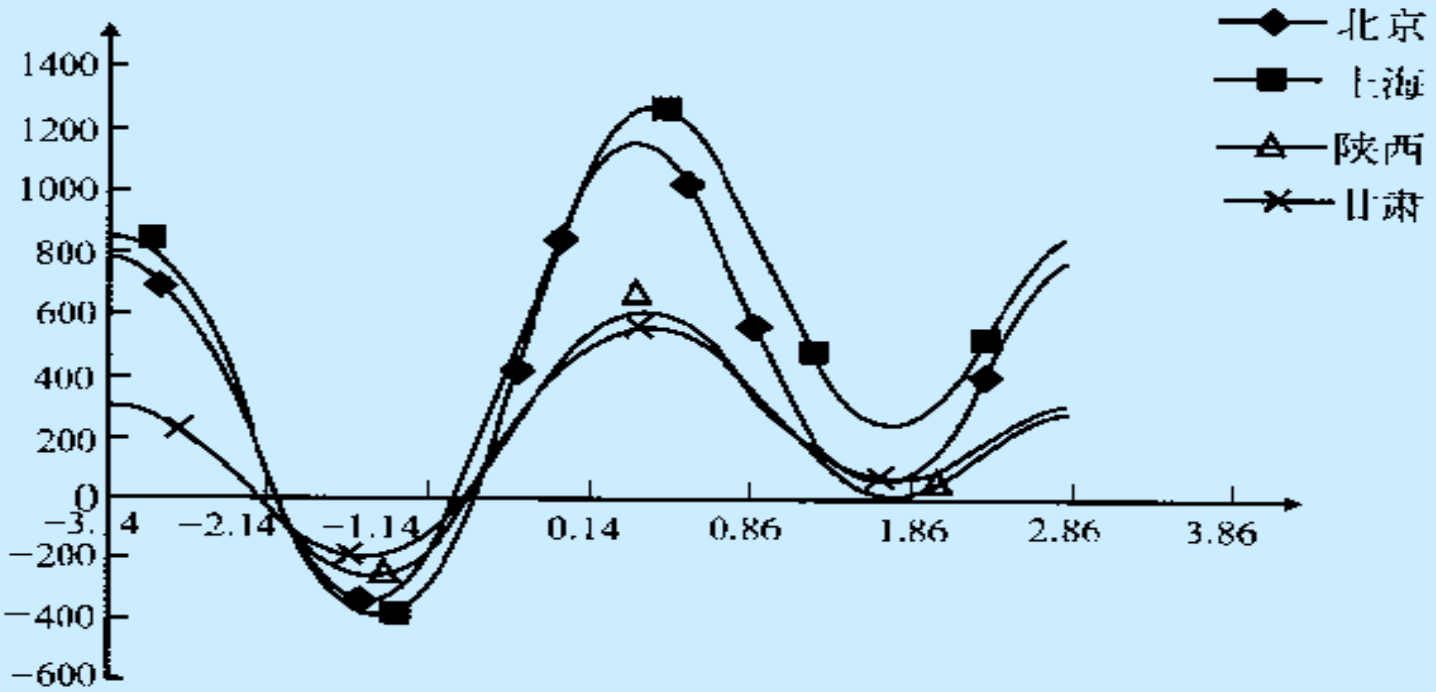
$$f_2(t)=\frac{678.92}{\sqrt{2}}+365.07\sin t+112.82\cos t+301.46\sin 2t+465.88\cos 2t$$

$$f_3(t)=\frac{237.38}{\sqrt{2}}+174.48\sin t+119.78\cos t+141.07\sin 2t+245.57\cos 2t$$

调和曲线图3

$$f_4(t)=\frac{253.41}{\sqrt{2}}+156.13\sin t+102.96\cos t+108.13\sin 2t$$
$$+212.20\cos 2t$$

它们的图形如下：



调和曲线图4

n 次观测对应 n 条曲线画在同一平面上就是一张调和曲线图。

在多项式的图表示中，当各变量的数值太悬殊时，最好先标准化后再作图。

作调和曲线时一般要借助计算机作图，这种图对聚类分析帮助很大，如果选择聚类统计量为距离的话，同类的曲线非常靠近拧在一起，不同类的曲线拧成不同的束，非常直观。

星座图 1

星座图是将高维空间中的样品点投影到平面上的一个半圆内，用投影点表示样品点。

具体的作图步骤是：

（1）将数据 $\{X_{ki}\}$ 变换为角度 $\{\theta_{ki}\}$，使 $0 \leqslant \theta_{ki} \leqslant \pi$，常取变换方法如下（极差标准化）：

$$\theta_{ki} = \frac{X_{ki} - \min\limits_{L=1\cdots n} X_{Li}}{\max\limits_{L=1\cdots n} X_{Li} - \min\limits_{L=1\cdots n} X_{Li}} \times 180^\circ \qquad \begin{matrix} k=1, \cdots, n \\ i=1, \cdots, p \end{matrix}$$

星座图2

(2) 适当地选一组权系数 $w_1, w_2, \cdots, w_p$，其中 $w_i \geqslant 0$ 且 $\sum_{i=1}^{p} w_i = 1$。重要的变量相应的权数可取大一点。最简单的取法为 $w_i = \frac{1}{p}$，$i = 1, \cdots, p$。

(3) 画出一个半径为 1 的上半圆及半圆底边的直径。

星座图3

（4）对给定的第 k 次观测 $X_k=(x_{k1},\cdots,x_{kp})'$ 对应着上半圆内的一个星号 * 和一条由折线表示的路径。路径的折点坐标是

$$\begin{cases} U_k^{(L)}=\sum_{i=1}^{L} W_i\cos\theta_{ki} & L=1,\cdots,p \\ V_k^{(L)}=\sum_{i=1}^{L} W_i\sin\theta_{ki} & k=1,\cdots,n \end{cases}$$

星星位于路径的终点，其坐标为 $(U_K^{(P)},V_K^{(P)})$。

星座图4

比如取 $k=1$，则第一个样品点的观测值 $X_1=(x_{11},\cdots,x_{1p})'$，其路径折点的坐标依次为：

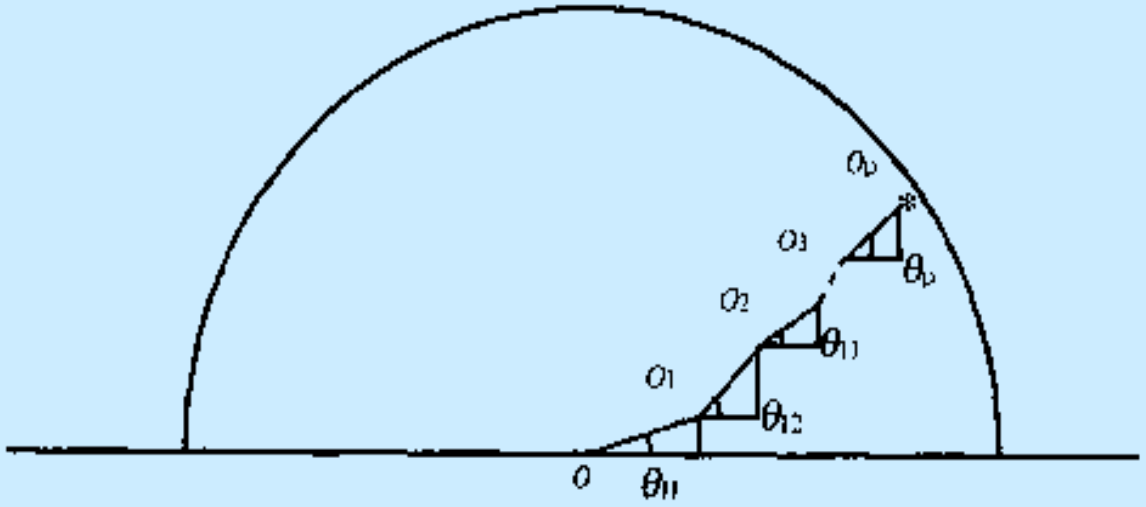
$$\begin{cases}U_1^{(1)}=W_1\cos\theta_{11}\\ V_1^{(1)}=W_1\sin\theta_{11}\end{cases},\qquad \begin{cases}U_1^{(2)}=W_1\cos\theta_{11}+W_2\cos\theta_{12}\\ V_1^{(2)}=W_1\sin\theta_{11}+W_2\sin\theta_{12}\end{cases},\qquad \cdots$$

$$\cdots\begin{cases}U_1^{(p)}=W_1\cos\theta_{11}+\cdots+W_p\cos\theta_{1p}\\ V_1^{(p)}=W_1\sin\theta_{11}+\cdots+W_p\sin\theta_{1p}\end{cases}$$

将这些坐标 $(U_1^{(1)},V_1^{(1)})$，$(U_1^{(2)},V_1^{(2)})$，$\cdots$，$(U_1^{(p)},V_1^{(p)})$ 所对应的点分别记为 o_1，o_2，$\cdots$，o_p，连接 o_1，$\cdots$，o_p 即为第一个样品点的路径。

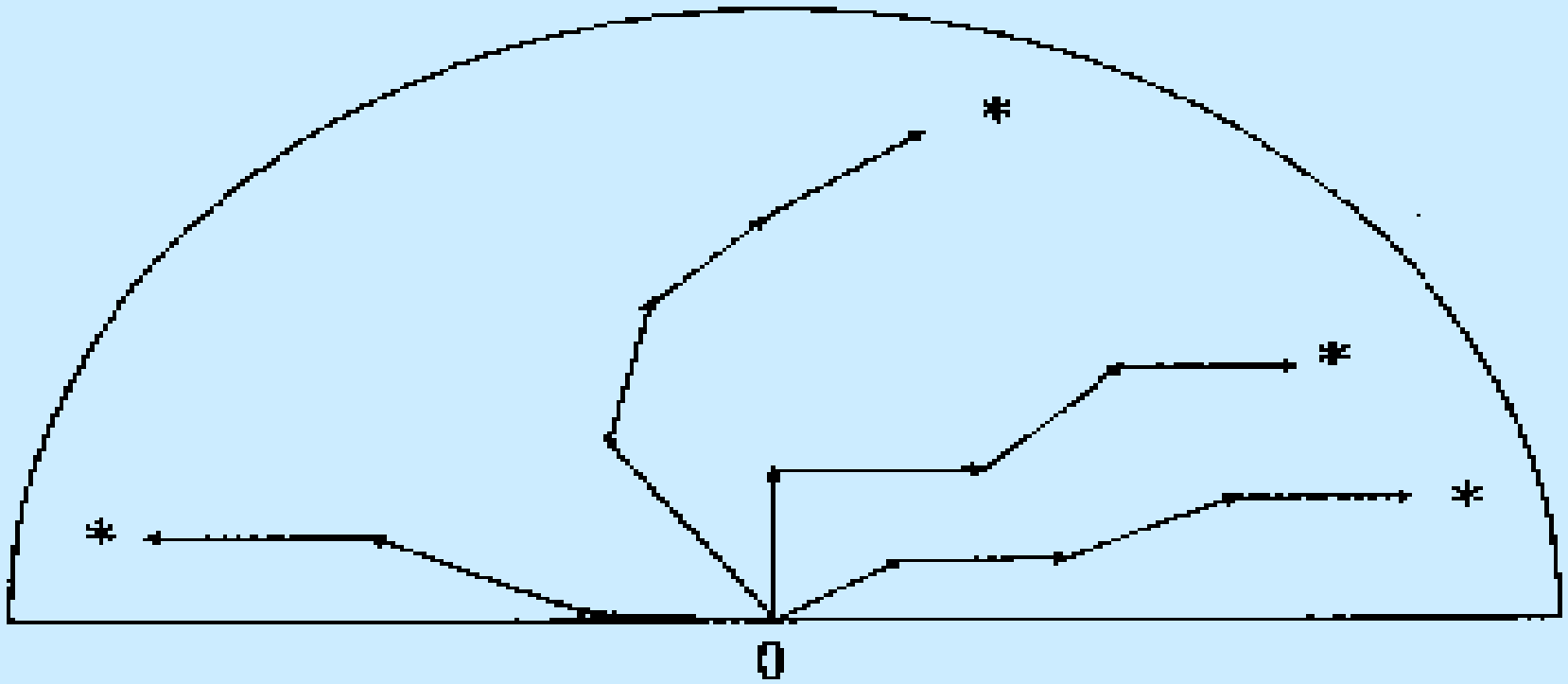
星座图5

从上面表达式不难看出路径终点的横坐标就是点 o_1 到点 o_p 的横坐标之和，终点的纵坐标是点 o_1 到点 o_p 的纵坐标之和。下面画出第一个样品点的路径折线和星星位置。



如果将 n 个样品点的路径折线和星星位置都画出来，就很像天文学中星座的图象，如下图所示。

星座图6



星座图7

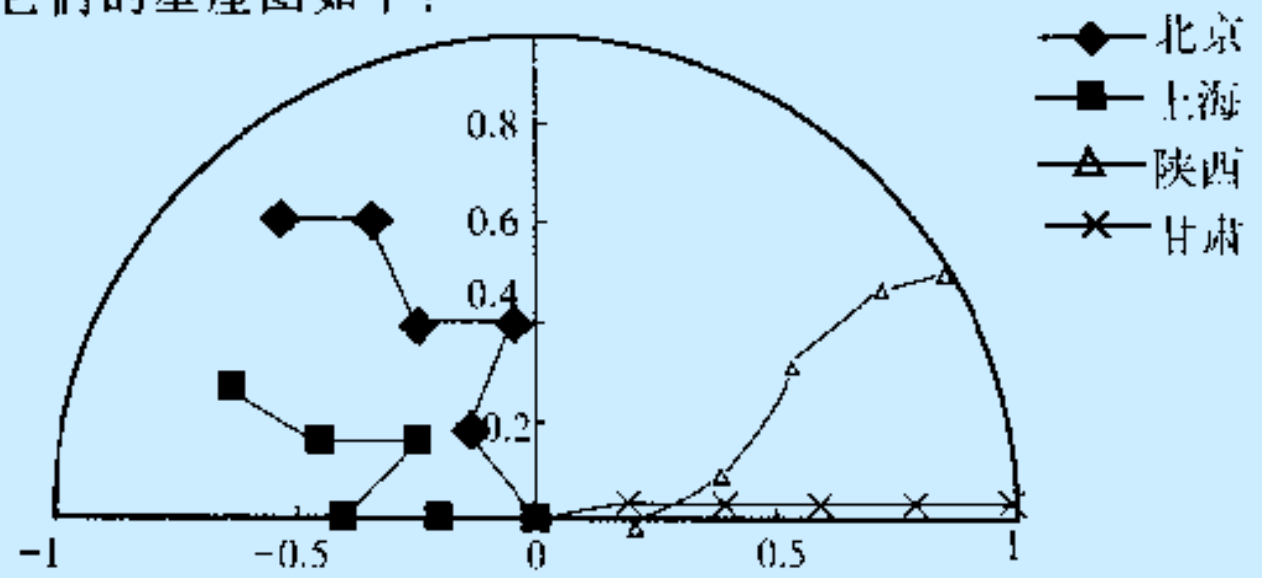
北京的 θ_{1i}（$i=1, \cdots, 5$）是 133°，35°，180°，119°，180°

上海的 θ_{2i}（$i=1, \cdots, 5$）是 180°，180°，40°，180°，153°

陕西的 θ_{3i}（$i=1, \cdots, 5$）是 0°，16°，68°，30°，20°

甘肃的 θ_{4i}（$i=1, \cdots, 5$）是 7°，0°，0°，0°，0°。

它们的星座图如下：



星座图8

从上述不难看到，星星的位置和路径与权数的选取有关，取不同的权数，画出的星座图也不同，究竟哪个最好，目前还无公认的最优法。一般权数选取的原则依实际问题的需要而定。通常情况对较重要指标取权数大些，次要指标取权数小些，如果指标的重要程度相差不大或难以区分，则选取等权。

多元数据的图表示法还有很多，如脸谱图、树形图、塑像图等等。此处不再作更多的介绍。

盒形图

盒形图也叫箱图，箱图是按分组变量值并列显示，箱图的结构如下：

- 矩形框为箱图主体，箱的上边线与下边线之差称为箱长，也称为”内四分位限“(国内——些统计书中称为“百分位差”)，它包含了变量约50％的数值，系统以默认的红色显示，箱体矩形框上、中、下3条平行线依次表示变量的75％、50％、25％分位数。
- 触须线，即中间的竖线。它向上触及和向下触及的两条横线分别表示变量本体的最大值和最小值。本体由除去奇异值和极端值以外的变量值组成，也称为本体值。
- Outlier(奇异值)，位于箱本体上下用圆圈标记的点，指从箱的上下边沿算起，对应的变量值超过箱长的1.5倍的那些值。由于选定的标识变量为Name，所以奇异值旁边标注姓名。
- Extreme(极端值)，系统默认用“*”标记。它们指从箱的上下边沿算起，其对应的变量值超过路长的3倍以上的那些点。

生成箱图的SPSS操作：

- Graphs—Boxptots命令即可

直方图

- 注意：纵坐标为频率/组距
  横坐标长度为组距

正态概率分布图

- Normal p-p plots
- Normal q-q plots
- 注：有关图的作法可参考：
- 郝黎仁等编著．SPSS实用统计分析（第十三章的相关内容），北京：中国水利水电出版仕，2002

第三章　聚类分析
Clustering Analysis

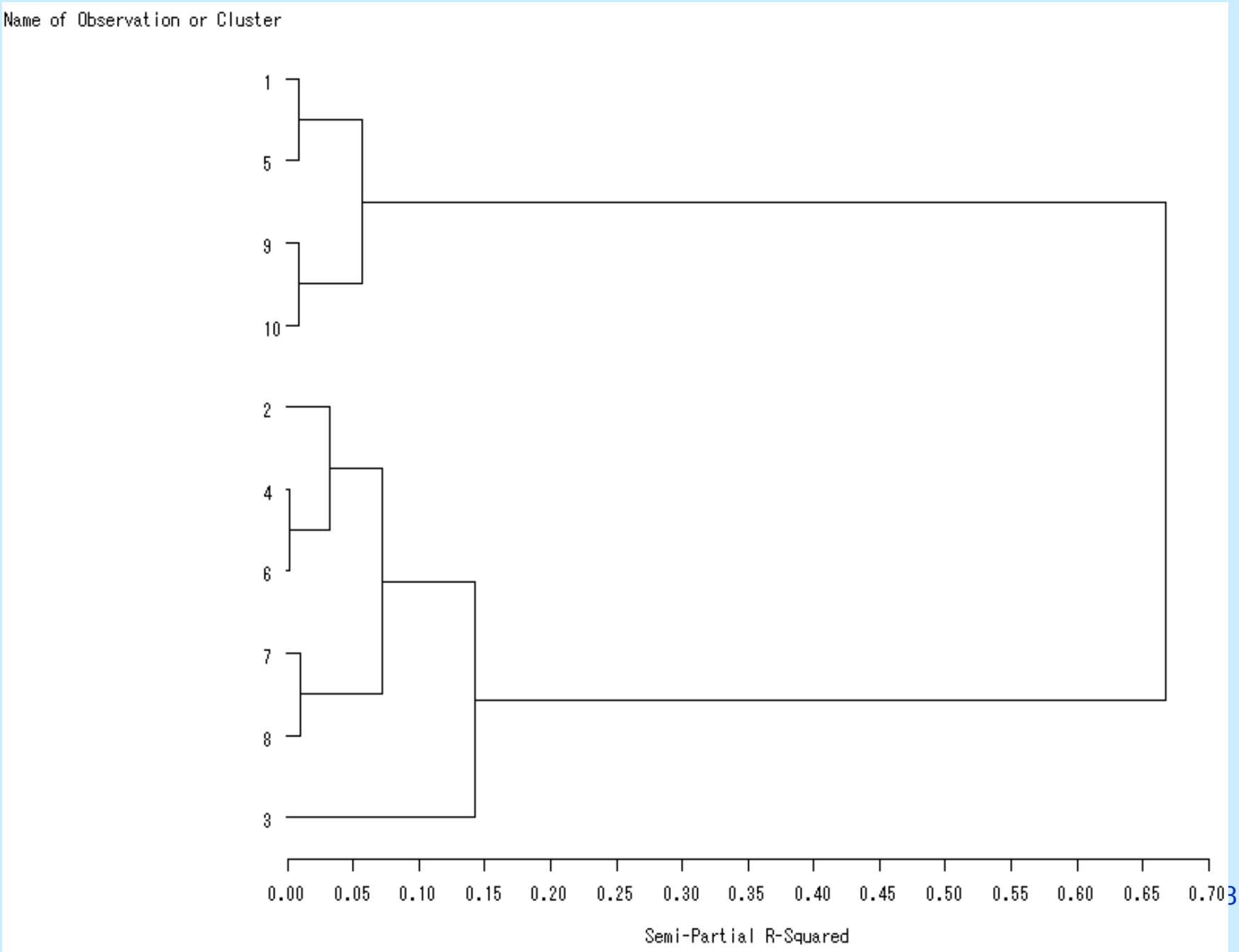
本章主要内容

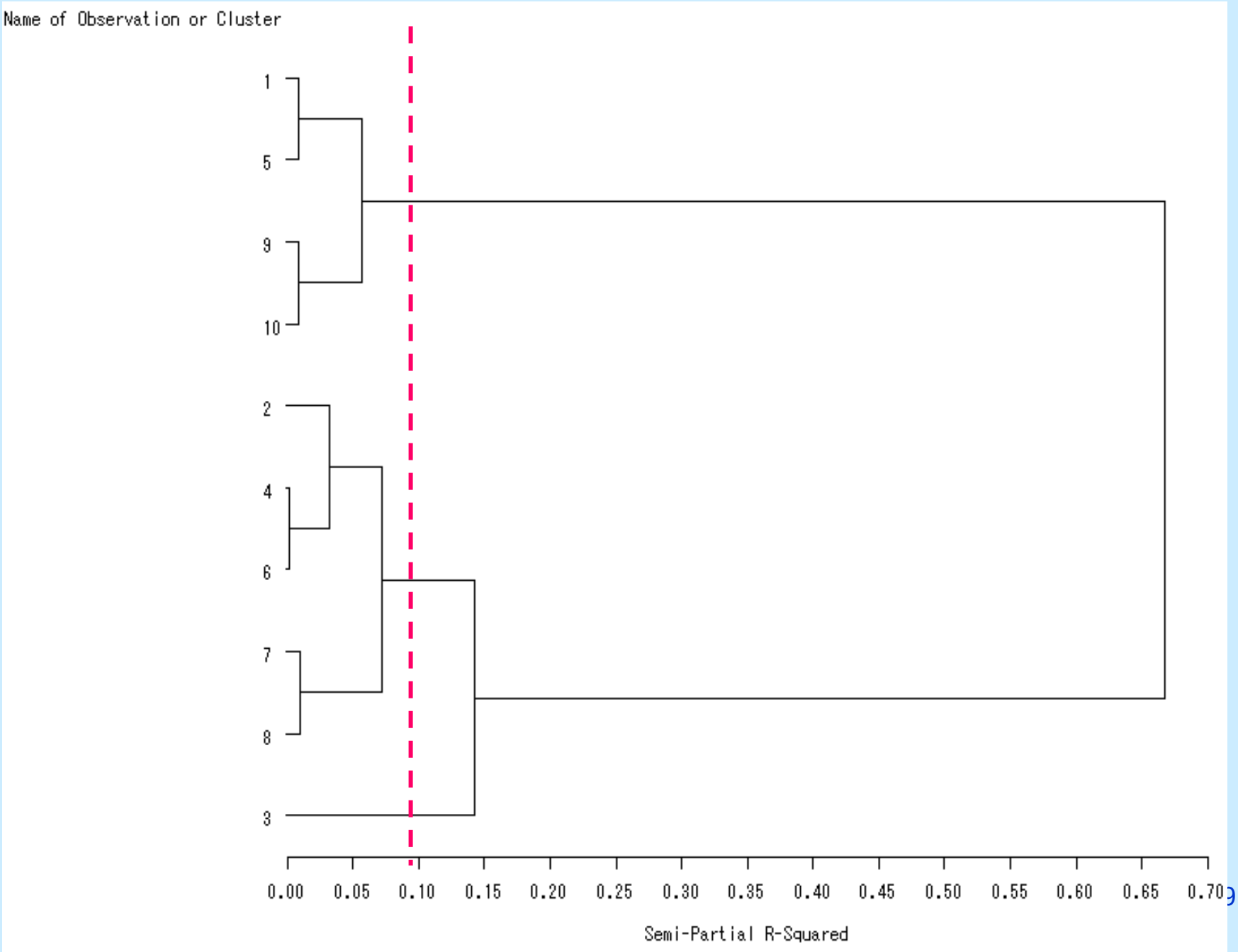
- 聚类分析的概念
- 距离和相似系数
- 八种系统聚类分析方法
- 系统聚类的基本性质
- 快速聚类
- 聚类分析应注意的问题

§1 什么是聚类分析

例 对10位应聘者做智能检验。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，选择合适的统计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


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
0.703
Semi-Partial R-Squared


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

我们直观地来看，这个分类是否合理？

计算4号和6号得分的离差平方和：

$(21-20)^2+(23-23)^2+(22-22)^2=1$

计算1号和2号得分的离差平方和：

$(28-18)^2+(29-23)^2+(28-18)^2=236$

计算1号和3号得分的离差平方和为482，由此可见一般，分类可能是合理的，欧氏距离很大的应聘者没有被聚在一起。

由此，我们的问题是如何来选择样品间相似的测度指标，如何将有相似性的类连接起来？

聚类分析根据一批样品的许多观测指标，按照一定的数学公式具体地计算一些样品或一些参数(指标)的相似程度，把相似的样品或指标归为一类，把不相似的归为一类。

例如对上市公司的经营业绩进行分类；据经济信息和市场行情，客观地对不同商品、不同用户及时地进行分类。又例如当我们对企业的经济效益进行评价时，建立了一个由多个指标组成的指标体系，由于信息的重叠，一些指标之间存在很强的相关性，所以需要将相似的指标聚为一类，从而达到简化指标体系的目的。

思考：样本点之间按什么刻画相似程度

思考：样本点和小类之间按什么刻画相似程度

思考：小类与小类之间按什么来刻画相似程度

§2 相似系数和距离

一、变量测量尺度的类型

为了将样本进行分类，就需要研究样品之间的关系；而为了将变量进行分类，就需要研究变量之间的关系。但无论是样品之间的关系，还是变量之间的关系，都是用变量来描述的，变量的类型不同，描述方法也就不同。通常，变量按照测量它们的尺度不同，可以分为三类。

(1) 间隔尺度。指标度量时用数量来表示，其数值由测量或计数、统计得到，如长度、重量、收入、支出等。一般来说，计数得到的数量是离散数量，测量得到的数量是连续数量。在间隔尺度中如果存在绝对零点，又称比例尺度。

(2)顺序尺度。指标度量时没有明确的数量表示，只有次序关系，或虽用数量表示，但相邻两数值之间的差距并不相等，它只表示一个有序状态序列。如评价酒的味道，分成好、中、次三等，三等有次序关系，但没有数量表示。

(3)名义尺度。指标度量时既没有数量表示也没有次序关系，只有一些特性状态，如眼睛的颜色，化学中催化剂的种类等。在名义尺度中只取两种特性状态的变量是很重要的，如电路的开和关，天气的有雨和无雨，人口性别的男和女，医疗诊断中的“十”和“一”，市场交易中的买和卖等都是此类变量。

二、数据的变换处理

所谓数据变换，就是将原始数据矩阵中的每个元素，按照某种特定的运算把它变成为一个新值，而且数值的变化不依赖于原始数据集合中其它数据的新值。

1、中心化变换

中心化变换是一种坐标轴平移处理方法，它是先求出每个变量的样本平均值，再从原始数据中减去该变量的均值，就得到中心化变换后的数据。

设原始观测数据矩阵为：

$$\mathbf{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

$$x_{ij}^{*} = x_{ij} - \bar{x}_j \qquad (i = 1,2,3,\cdots,n;\ j = 1,2,3,\cdots,p)$$

中心化变换的结果是使每列数据之和均为0，即每个变量的均值为0，而且每列数据的平方和是该列变量样本方差的(n–1)倍，任何不同两列数据之交叉乘积是这两列变量样本协方差的(n–1)倍，所以这是一种很方便地计算方差与协方差的变换。

2、极差规格化变换

规格化变换是从数据矩阵的每一个变量中找出其最大值和最小值，这两者之差称为极差，然后从每个变量的每个原始数据中减去该变量中的最小值，再除以极差，就得到规格化数据。即有：

$$x_{ij}^{*}=\frac{x_{ij}-\min\limits_{i=1,2,\cdots,n}(x_{ij})}{R_j} \qquad (i=1,2,3,\cdots,n;\ j=1,2,3,\cdots,p)$$

$$R_j=\max_{i=1,2,\cdots,n}(x_{ij})-\min_{i=1,2,\cdots,n}(x_{ij})$$

$$0\le x_{ij}^{*}\le 1$$

经过规格化变换后，数据矩阵中每列即每个变量的最大数值为1，最小数值为0，其余数据取值均在0－1之间；并且变换后的数据都不再具有量纲，便于不同的变量之间的比较。

3、标准化变换

标准化变换也是对变量的数值和量纲进行类似于规格化变换的一种数据处理方法。首先对每个变量进行中心化变换，然后用该变量的标准差进行标准化。即有：

$$x_{ij}^{*}=\frac{x_{ij}-\bar{x}_j}{S_j}\qquad (i=1,2,3,\cdots,n;\ j=1,2,3,\cdots,p)$$

$$S_j=\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2$$

经过标准化变换处理后，每个变量即数据矩阵中每列数据的平均值为0，方差为1，且也不再具有量纲，同样也便于不同变量之间的比较。变换后，数据短阵中任何两列数据乘积之和是两个变量相关系数的（n－1）倍，所以这是一种很方便地计算相关矩阵的变换。

4．对数变换

对数变换是将各个原始数据取对数，将原始数据的对数值作为变换后的新值。即：

$$x_{ij}^{*} = \log(x_{ij})$$

三、样品间亲疏程度的测度

研究样品或变量的亲疏程度的数量指标有两种，一种叫**相似系数**，性质越接近的变量或样品，它们的相似系数越接近于1或—l，而彼此无关的变量或样品它们的相似系数则越接近于0，相似的为一类，不相似的为不同类；另一种叫**距离**，它是将每一个样品看作p维空间的一个点，并用某种度量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。

变量之间的聚类即R型聚类分析，常用相似系数来测度变量之间的亲疏程度。而样品之间的聚类即Q型聚类分析，则常用距离来测度样品之间的亲疏程度。

注:变量聚类放到因子分析后面

1、定义距离的准则

定义距离要求满足第i个和第j个样品之间的距离如下四个条件（距离可以自己定义，只要满足距离的条件）

$d_{ij} \geq 0$对一切的i和j成立;

$d_{ij} = 0$当且仅当$i = j$成立;

$d_{ij} = d_{ji}$ 0对一切的i和j成立;

$d_{ij} \leq d_{ik} + d_{kj}$对于一切的$i$和$j$成立.

2、常用距离的算法

(1) 明氏距离测度

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

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

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

明考夫斯基距离主要有以下两个缺点：

①明氏距离的值与各指标的量纲有关，而各指标计量单位的选择有一定的人为性和随意性，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从而使该距离的数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。

(2)杰氏距离

这是杰斐瑞和马突斯塔(Jffreys & Matusita)所定义的一种距离，其计算公式为：

$$d_{ij}(J)=\left[\sum_{k=1}^{p}(\sqrt{x_{ik}}-\sqrt{x_{jk}})^2\right]^{1/2}$$

(3)兰氏距离

这是兰思和维廉姆斯(Lance & Williams)所给定的一种距离，其计算公式为：

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

这是一个自身标准化的量，由于它对大的奇异值不敏感，这样使得它特别适合于高度偏倚的数据。虽然这个距离有助于克服明氏距离的第一个缺点，但它也没有考虑指标之间的相关性。

(4)马氏距离

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

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

分别表示第i个样品和第j样品的p指标观测值所组成的列向量，即样本数据矩阵中第i个和第j个行向量的转置，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欧氏距离。显然，马氏距离与上述各种距离的主要不同就是马氏距离考虑了观测变量之间的相关性。如果假定各变量之间相互独立，即观测变量的协方差矩阵是对角矩阵，则马氏距离就退化为用各个观测指标的标准差的倒数作为权数进行加权的欧氏距离。因此，马氏距离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，为了对马氏距离和欧氏距离进行一下比较，以便更清楚地看清二者的区别和联系，现考虑一个例子。

例如，假设有一个二维正态总体，它的分布为：

$$N_2\left[\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & 0.9\\0.9 & 1\end{pmatrix}\right] \qquad \Sigma^{-1}=\frac{1}{0.19}\begin{pmatrix}1 & -0.9\\-0.9 & 1\end{pmatrix}$$

设$A(1,1)$和$B=(1,-1)$两点。

$$d_{A\mu}(M)=1.05 \qquad d_{A\mu}(U)=2$$

$$d_{B\mu}(M)=20 \qquad d_{B\mu}(U)=2$$

(5) 斜交空间距离

由于各变量之间往往存在着不同的相关关系，用正交空间的距离来计算样本间的距离易变形，所以可以采用斜交空间距离。

$$d_{ij}=\left[\frac{1}{p^2}\sum_{h=1}^{p}\sum_{k=1}^{p}(x_{ih}-x_{jh})(x_{ik}-x_{jk})\gamma_{hk}\right]^{1/2}$$

当各变量之间不相关时，斜交空间退化为欧氏距离。

2、相似系数的算法

（1）相似系数

设 $\mathbf{x}_i=\left(x_{i1},\ x_{i2},\ \cdots,\ x_{ip}\right)'$ 和 $\mathbf{x}_j=(x_{j1},x_{j2},\cdots,x_{jp})'$ 是第 i 和 j 个样品的观测值，则二者之间的相似测度为：

其中 $$\gamma_{ij}=\frac{\sum_{k=1}^{p}(x_{ik}-\overline{x}_i)(x_{jk}-\overline{x}_j)}{\sqrt{[\sum_{k=1}^{p}(x_{ik}-\overline{x}_i)^2][\sum_{k=1}^{p}(x_{jk}-\overline{x}_j)^2]}}$$

（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

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

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

五、距离和相似系数选择的原则

一般说来，同一批数据采用不同的亲疏测度指标，会得到不同的分类结果。产生不同结果的原因，主要是由于不同的亲疏测度指标所衡量的亲疏程度的实际意义不同，也就是说，不同的亲疏测度指标代表了不同意义上的亲疏程度。因此我们在进行聚类分析时，应注意亲疏测度指标的选择。通常，选择亲疏测度指标时，应注意遵循的基本原则主要有：

(1)所选择的亲疏测度指标在实际应用中应有明确的意义。如在经济变量分析中，常用相关系数表示经济变量之间的亲疏程度。

(2) 亲疏测度指标的选择要综合考虑已对样本观测数据实施了的变换方法和将要采用的聚类分析方法。如在标准化变换之下，夹角余弦实际上就是相关系数；又如若在进行聚类分析之前已经对变量的相关性作了处理，则通常就可采用欧氏距离，而不必选用斜交空间距离。此外，所选择的亲疏测度指标，还须和所选用的聚类分析方法一致。如聚类方法若选用离差平方和法，则距离只能选 用欧氏距离。

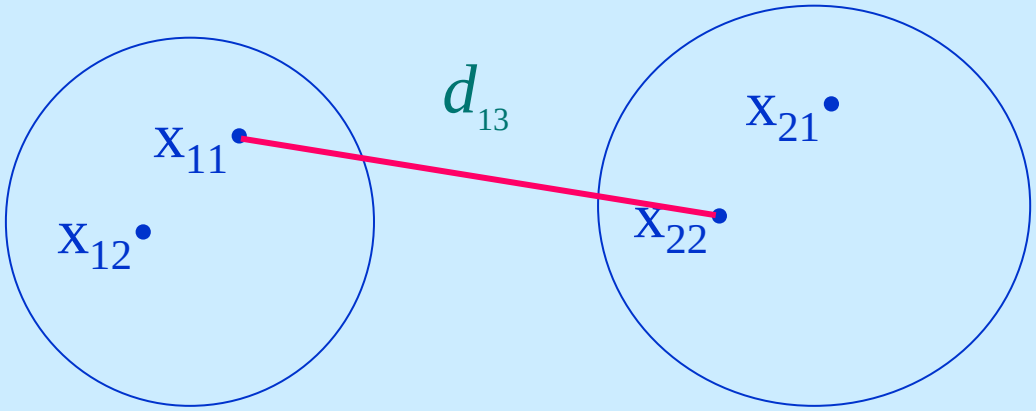
(3)适当地考虑计算工作量的大小。如对大样本的聚类问题，不适宜选择斜交空间距离，因采用该距离处理时，计算工作量太大。样品间或变量间亲疏测度指标的选择是一个比较复杂且带主规性的问题，我们应根据研究对象的特点作具体分折，以选择出合适的亲疏测度指标。实践中，在开始进行聚类分析时，不妨试探性地多选择几个亲疏测度指标，分别进行聚类，然后对聚类分析的结果进行对比分析，以确定出合适的亲疏测度指标。

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。

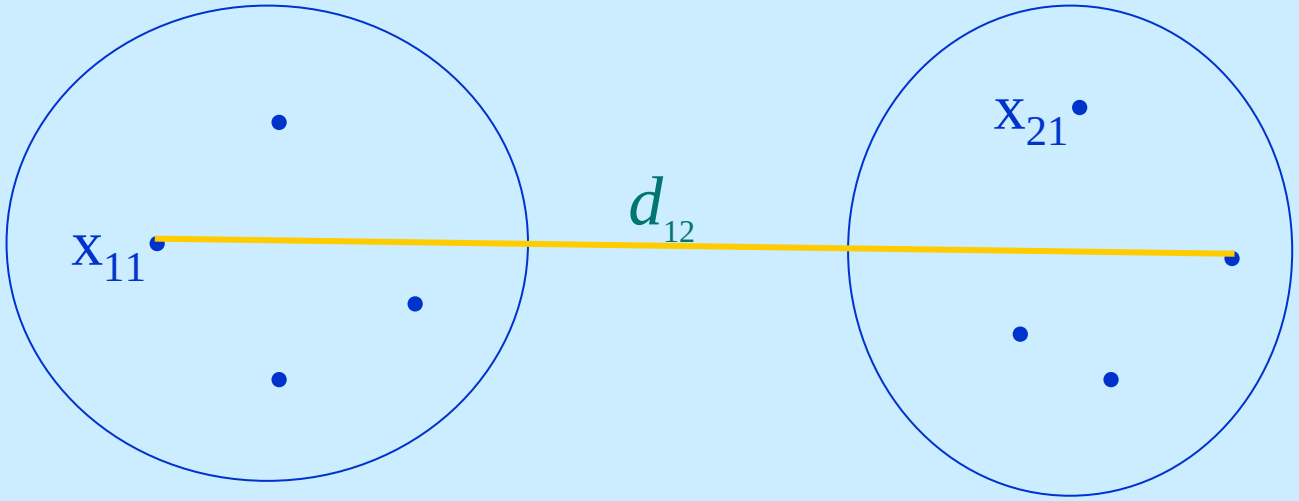
G_q \ G_p	G_1	G_2	$\cdots$	G_n
G_1	0	d_{12}	$\cdots$	d_{1n}
G_2	d_{21}	0		d_{2n}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
G_n	d_{n1}	d_{n2}	$\cdots$	0

四、样本数据与小类、小类与小类之间的度量

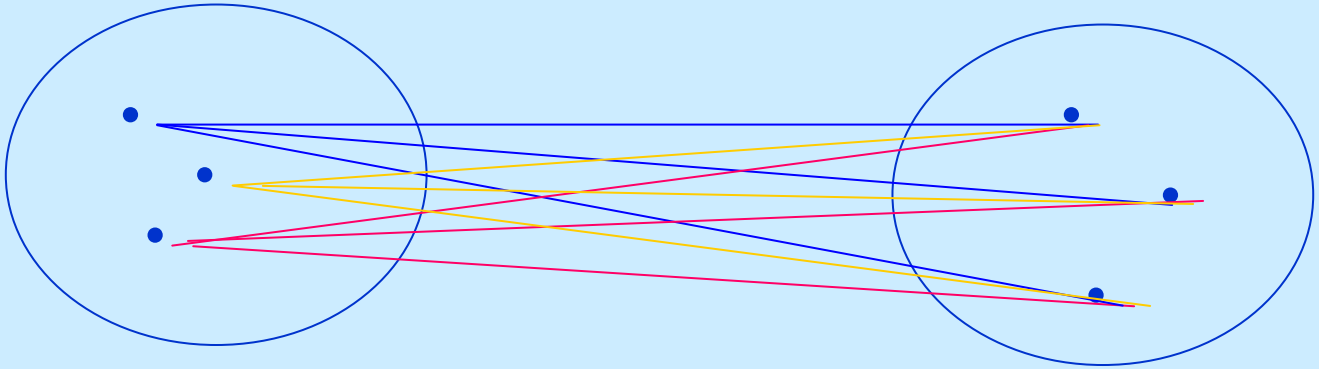
1、最短距离（Nearest Neighbor)



最长距离（Furthest Neighbor）



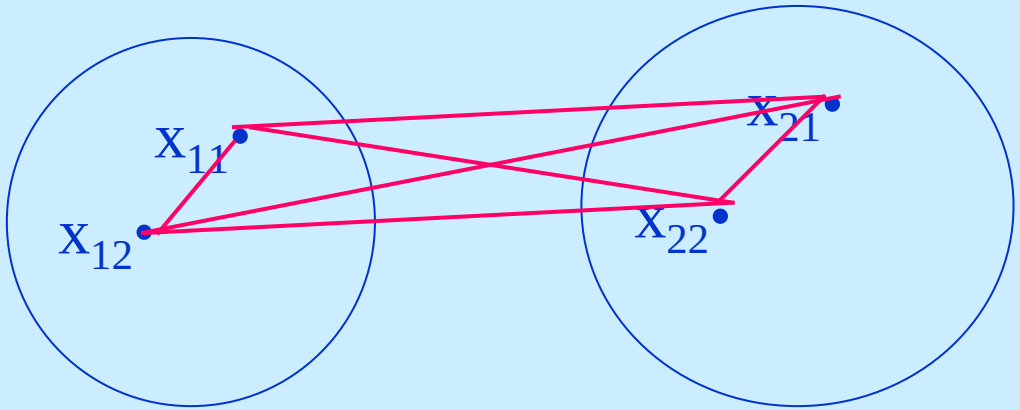
组间平均连接（Between-group Linkage)



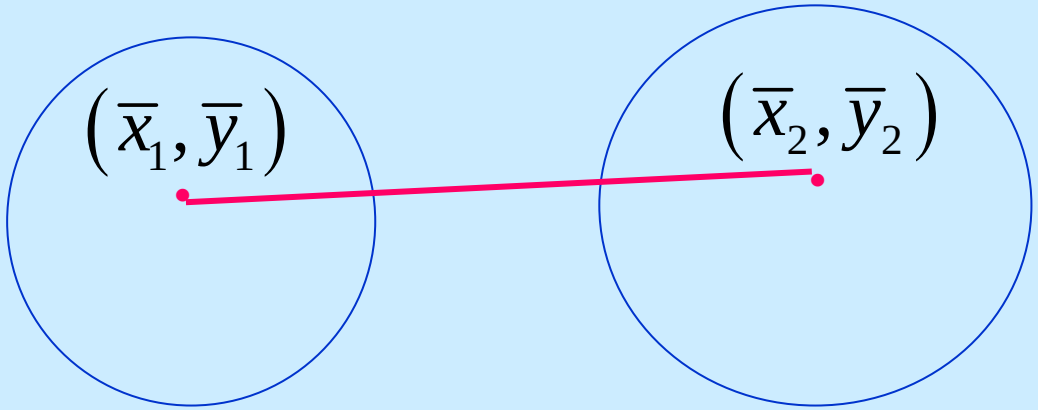
$$\frac{d_1 + \cdots + d_9}{9}$$

1、组内平均连接法（Within-group Linkage)

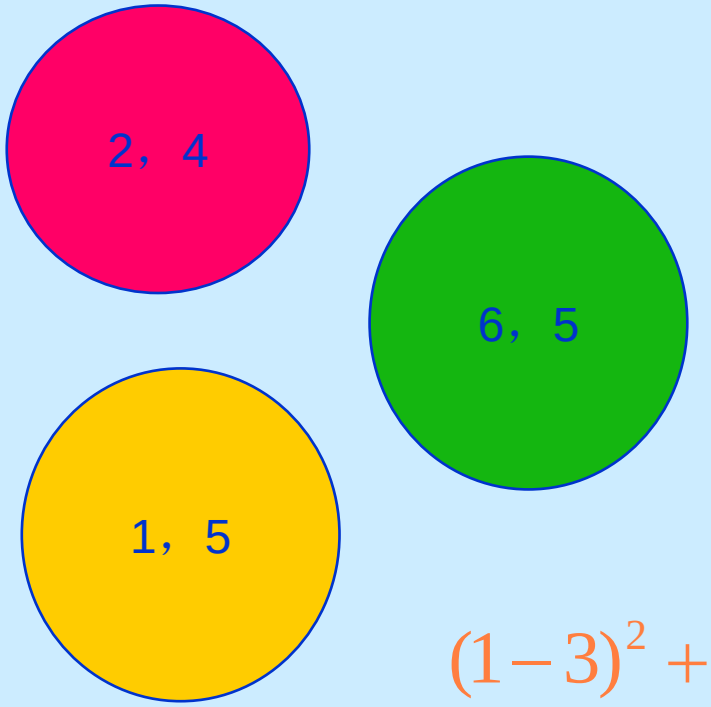
$$\frac{d_1 + d_2 + d_3 + d_4 + d_5 + d_6}{6}$$



重心法（Centroid clustering):均值点的距离



离差平方和法连接



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

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

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

红绿（2，4，6，5）8.75

离差平方和增加8.75－2.5＝6.25

黄绿（6，5，1，5）14.75

离差平方和增加14.75－8.5＝6.25

黄红（2，4，1，5）10－10＝0

故按该方法的连接和黄红首先连接。

§3　系统聚类方法

（一）方法

开始各样本自成一类。

1、根据样品的特征，规定样品之间的距离 d_{ij}，共有 C_n^2 个。将所有列表，记为D（0）表，该表是一张对称表。所有的样本点各自为一类。

2、选择D（0）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 $G_r=\{G_p,\ G_q\}$。

3、利用递推公式计算新类与其它类之间的距离。分别删除**D**（0）表的第p，q行和第p，q列，并新增一行和一列添上的结果，产生**D**（1）表。

4、在D（1）表再选择最小的非零数，其对应的两类有构成新类，再利用递推公式计算新类与其它类之间的距离。分别删除D（1）表的相应的行和列，并新增一行和一列添上的新类和旧类之间的距离。结果，产生D（2）表。类推直至所有的样本点归为一类为止。

（二）常用的种类

1、 最短距离法

设抽取五个样品，每个样品只有一个变量，它们是1，2，3.5，7，9。用最短距离法对5个样品进行分类。首先采用绝对距离计算距离矩阵：

$D(0)$

	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

然后 G_1 和 G_2 被聚为新类 G_6，得 $D(1)$：

	G_6	G_3	G_4	G_5
G_6	0			
G_3	1.5	0		
G_4	5	3.5	0	
G_5	7	5.5	2	0

定义距离：$D_{pq} = Min\{d_{ij}:\ \mathbf{x}_i \in G_p,\ \mathbf{x}_j \in G_q\}$

递推公式：$D_{rl} = Min\{D_{pl},\ D_{ql}\}$　　$l \neq p,\ q$

最短距离法的递推公式

定义距离：$D_{pq} = Min\{d_{ij}:\ \mathbf{x_i} \in G_p,\ \mathbf{x_j} \in G_q\}$

递推公式：$D_{rl} = Min\{D_{pl},\ D_{ql}\}$　　$l \neq p,\ q$

假设第p类和第q类合并成第类，第r类与其它各旧类的距离按最短距离法为：

$$D_{rl} = Min\{d_{ij}:\ \mathbf{x_i} \in G_r,\ \mathbf{x_j} \in G_l\}$$

$$= Min\{d_{ij}:\ \mathbf{x_i} \in (G_p \cup G_q),\ \mathbf{x_j} \in G_l\}$$

$$= Min\{d_{ij}:\ \mathbf{x_i} \in (G_p \cup G_q),\ \mathbf{x_j} \in G_l\}$$

$$= Min\{D_{ql}, D_{pl}\}$$

	G_7	G_4	G_5
G_7	0		
G_4	3.5	0	
G_5	5.5	2	0

	G_8	G_7
G_8	0	
G_7	3.5	0

各步聚类的结果：

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

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

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

(1,2,3,4,5)

2、最长距离法

用最长距离法对5个样品进行分类。首先采用绝对距离计算距离矩阵：

	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

然后和被聚为新类，得：

	G_6	G_3	G_4	G_5
G_6	0			
G_3	2.5	0		
G_4	6	3.5	0	
G_5	8	5.5	2	0

最长距离法的递推公式

定义距离：$D_{pq} = Max\left\{d_{ij}:\ \mathbf{x_i} \in G_p,\ \mathbf{x_j} \in G_q\right\}$

递推公式：$D_{rl} = Max\left\{D_{pl},\ D_{ql}\right\}$　　$l \neq p,\ q$

假设第p类和第q类合并成第类，第r类与其它各旧类的距离按最长距离法为：

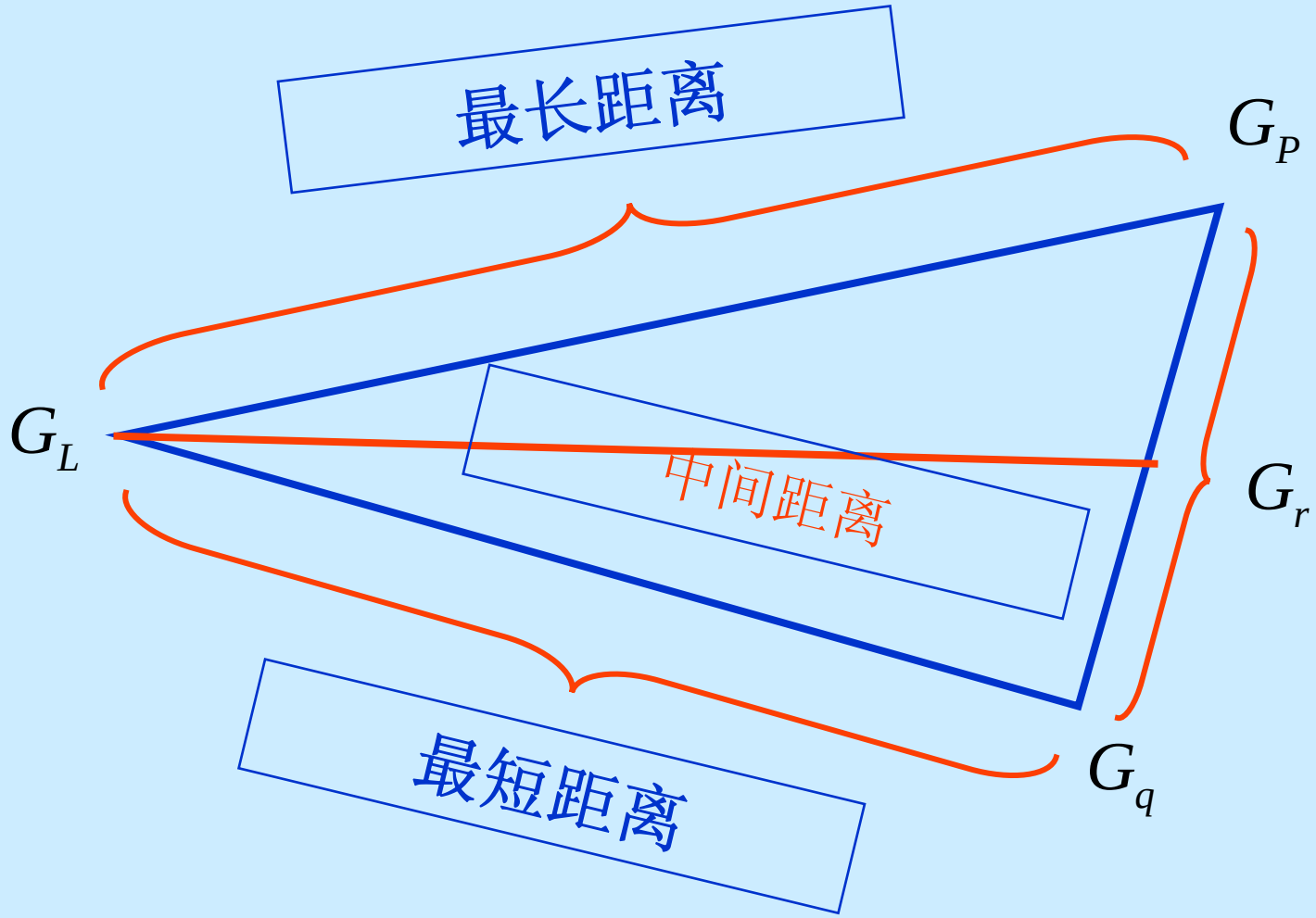
$$D_{rl} = Max\left\{d_{ij}:\ \mathbf{x_i} \in G_r,\ \mathbf{x_j} \in G_l\right\}$$

$$= Max\left\{d_{ij}:\ \mathbf{x_i} \in \left(G_p \cup G_q\right),\ \mathbf{x_j} \in G_l\right\}$$

$$= Max\left\{d_{ij}:\ \mathbf{x_i} \in \left(G_p \cup G_q\right),\ \mathbf{x_j} \in G_l\right\}$$

$$= Max\left\{D_{ql}, D_{pl}\right\}$$

3、中间距离法



用中间距离法对5个样品进行分类。首先采用绝对距离计算距离平方矩阵：

$D(0)$

	G_1	G_2	G_3	G_4	G_5
G_1	0				
G_2	1	0			
G_3	6.25	2.25	0		
G_4	36	25	12.25	0	
G_5	64	49	30.25	4	0

中间距离法的递推公式

递推公式：$D_{lr}^2 = \frac{1}{2}D_{lp}^2 + \frac{1}{2}D_{lq}^2 - \frac{1}{4}D_{pq}^2$

递推公式：$D_{kr}^2 = \frac{1}{2}D_{kp}^2 + \frac{1}{2}D_{kq}^2 + \beta D_{pq}^2, \quad -\frac{1}{4} < \beta < 0$

$$D_{63} = \frac{D_{13} + D_{23}}{2} - \frac{1}{4}D_{12}$$

$$D_{63} = \frac{6.25 + 2.25}{2} - \frac{1}{4} \times 1 = 4$$

	G_6	G_3	G_4	G_5
G_6	0			
G_3	4	0		
G_4	30.25	12.25	0	
G_5	56.25	30.25	4	0

4、类平均法

类平均法定义类间的距离是两类间样品的距离的平均数。对应我们前面讨论的组间

	G_1	G_2	G_3	G_4	G_5
G_1	0				
G_2	1	0			
G_3	6.25	2.25	0		
G_4	36	25	12.25	0	
G_5	64	49	30.25	4	0

然后和被聚为新类，得 $D(1)$ ：

	G_6	G_3	G_4	G_5
G_6	0			
G_3	4.25	0		
G_4	30.25	12.25	0	
G_5	56.25	30.25	4	0

递推公式：$D_{rk}^2 = \dfrac{n_p D_{kp}^2 + n_q D_{kq}^2}{n_p + n_q}$

类平均法的递推公式

$$D_{pq}^2 = \frac{1}{n_p n_q} \sum_{x_i \in G_p} \sum_{x_j \in G_q} d_{ij}^2$$

假设第p类和第q类合并成第类，第r类与其它各旧类的距离按最短距离法为：

$$D_{rl}^2 = \frac{1}{(n_p + n_q) n_l} \sum_{x_i \in (G_p \cup G_q)} \sum_{x_j \in G_l} d_{ij}^2$$

$$= \frac{1}{(n_p + n_q) n_l} \left(\sum_{x_i \in G_p} \sum_{x_j \in G_l} d_{ij}^2 + \sum_{x_i \in G_q} \sum_{x_j \in G_l} d_{ij}^2 \right)$$

$$= \frac{1}{(n_p + n_q) n_l} \left(\frac{n_p n_l}{n_p n_l} \sum_{x_i \in G_p} \sum_{x_j \in G_l} d_{ij}^2 + \frac{n_q n_l}{n_q n_l} \sum_{x_i \in G_q} \sum_{x_j \in G_l} d_{ij}^2 \right)$$

$$= \frac{1}{(n_p + n_q) n_l} \left(n_p n_l D_{pl} + n_q n_l D_{lq} \right)$$

$$= \frac{n_p D_{pl} + n_q D_{lq}}{n_p + n_q}$$

p类和q类与L类的距离的加权平均数

5、可变类平均法

类平均法的递推公式中，没有反映G_p类和G_q类的距离有多大，进一步将其改进，加入D^2_{Pq}，并给定系数$\beta<1$，则类平均法的递推公式改为：

$$D_{rl}^2=(1-\beta)\frac{n_pD_{pl}^2+n_qD_{ql}^2}{n_p+n_q}+\beta D_{pq}^2$$

用此递推公式进行聚类就是可变类平均法。递推公式由：

p类和q类与L类的距离的加权平均数

p类和q类的距离

两项的加权和构成，β的大小根据哪项更重要而定。

6、离差平方和法

G_1 和 G_2 被聚为新类，重心为 $\overline{X}_6 = (1+2)/2 = 1.5$

如G_1和G_2为一类，则离差平方和

$$S_{12} = (1-1.5)^2 + (2-1.5)^2 = 0.5$$

如 G_1和G_3为一类，则离差平方和

$$S_{13} = (1-2.25)^2 + (2-2.25)^2 = 3.125$$

类似于方差分析的想法，如果类分得恰当，同类内的样品之间的离差平方和应较小，而类间的离差平方和应当较大。

离差平方和法的思路是，当k固定时，选择使S达到最小的分类。先让n个样品各自成一类，然后缩小一类，每缩小一类离差平方和就要增大，选择使S^2增加最小的两类合并，直到所有的样品归为一类为止。离差平方和法定义类间的平方距离为

	G_1	G_2	G_3	G_4	G_5
G_1	0				
G_2	0.5	0			
G_3	3.125	1.125	0		
G_4	18	12.50	6.125	0	
G_5	32	24.50	15.125	2	0

定义距离为离差平方和的增量：$D_{pq}^2 = S_r^2 - S_p^2 - S_q^2$

其中 S_r^2 是由G_p和G_q合并成的G_r类的类内离差平方和。可以证明离差平方和的聚类公式为

递推公式：$D_{rk}^2 = \frac{n_k + n_p}{n_r + n_k} D_{pk}^2 + \frac{n_k + n_q}{n_r + n_k} D_{qk}^2 - \frac{n_k}{n_k + n_r} D_{pq}^2$

7、可变方法

如果让中间距离法的递推公式前两项的系数也依赖于β，则递推公式为：

$$D_{kr}^2 = \frac{1-\beta}{2}(D_{kp}^2 + D_{kq}^2) + \beta D_{pq}^2, \quad \beta < 1$$

用上式作为递推公式的系统聚类法称为可变法。

8、重心法

用重心法对5个样品进行分类。首先采用绝对距离计算距离平方矩阵：

	G_1	G_2	G_3	G_4	G_5
G_1	0				
G_2	1	0			
G_3	6.25	2.25	0		
G_4	36	25	12.25	0	
G_5	64	49	30.25	4	0

重心法，也称为样品的均值法。设Gp和Gq 为两个类

$$\overline{X}_p = \frac{1}{n_p}\sum_{x_i \in G_p}^{n_p} x_i \qquad \overline{X}_q = \frac{1}{n_q}\sum_{x_i \in G_q}^{n_q} x_i$$

分别为G_p和G_q的重心，类与类之间的距离定义为两个类重心（类内样品平均值）间的平方距离。

设某一步G_p和G_q的重心分别为为和，类内的样品数分别为和，如果要把G_p和G_q合并为G_r类，则G_r类的样品数$n_r=n_p+n_q$，G_r类的重心为$\overline{X}_p$和$\overline{X}_q$的加权算术平均数：

$$\overline{X}_r = \frac{n_p\overline{X}_p + n_q\overline{X}_q}{n_p + n_q}$$

重心法递推公式

假设第p类和第q类合并成第类，第r类与其它各旧类的距离按重心法为：

$$D_{rl} = \left\| \bar{\mathbf{x}}_r - \bar{\mathbf{x}}_l \right\| = \left\| \frac{1}{n_p + n_q} \sum_{x_i \in (G_p + G_q)} x_i - \frac{1}{n_l} \sum_{x_i \in G_l} x_i \right\|$$

$$= \left\| \frac{1}{n_p + n_q} \sum_{x_i \in G_p} + \frac{1}{n_p + n_q} \sum_{x_i \in G_q} x_i - \frac{1}{n_l} \sum_{x_i \in G_l} x_i \right\|$$

$$= \left\| \frac{n_p}{n_r n_p} \sum_{x_i \in G_p} x_i + \frac{n_q}{n_r n_q} \sum_{x_i \in G_q} x_i - \frac{1}{n_l} \sum_{x_i \in G_l} x_i \right\|$$

$$= \left\| \frac{n_p}{n_r} \overline{x}_p + \frac{n_q}{n_r} \overline{x}_q - \overline{x}_l \right\|$$

$$= \left\| \frac{n_p}{n_r} \overline{x}_p + \frac{n_q}{n_r} \overline{x}_q - \frac{n_p}{n_r} \overline{x}_l - \frac{n_q}{n_r} \overline{x}_l \right\|$$

$$= \left\| \frac{n_p}{n_r} \overline{x}_p - \frac{n_p}{n_r} \overline{x}_l + \frac{n_q}{n_r} \overline{x}_q - \frac{n_q}{n_r} \overline{x}_l \right\|$$

$$= \frac{n_p}{n_r} D^2_{pl} + \frac{n_q}{n_r} D^2_{ql} - \frac{n_p n_q}{n_r^2} D^2_{pq}$$

G_4和G_6的距离为

$$D_{46}^2 = \frac{1}{2}D_{41}^2 + \frac{1}{2}D_{42}^2 - \frac{1}{4}D_{12}^2$$

$$= \frac{1}{2} \times 36 + \frac{1}{2} \times 25 - \frac{1}{4} \times 1 = 30.25$$

（三）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。但是这个问题又是不可回避的。下面我们介绍几种方法。

1、给定阈值——通过观测聚类图，给出一个合适的阈值T。要求类与类之间的距离不要超过T值。例如我们给定T=0.35，当聚类时，类间的距离已经超过了0.35，则聚类结束。

总离差平方和的分解（准备知识）

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

$$\bar{x}_1 \quad \bar{x}_2 \quad \bar{x}_p$$

$$\text{总离差平方和} = (x_{11} - \bar{x}_1)^2 + \cdots + (x_{n1} - \bar{x}_1)^2$$

$$\cdots + (x_{1p} - \bar{x}_p)^2 + \cdots (x_{np} - \bar{x}_p)^2$$

如果这些样品被分成两类

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n_1 1} & x_{n_1 2} & \cdots & x_{n_1 p} \end{bmatrix} \qquad \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n_2 1} & x_{n_2 2} & \cdots & x_{n_2 p} \end{bmatrix}$$

$$\bar{x}_1^{(1)} \quad \bar{x}_2^{(1)} \quad \bar{x}_p^{(1)} \qquad \bar{x}_1^{(2)} \quad \bar{x}_2^{(2)} \quad \bar{x}_p^{(2)}$$

一组的离差平方和 $=(x_{11}-\bar{x}_1^{(1)})^2+\cdots+(x_{n_1 1}-\bar{x}_1^{(1)})^2$

$\cdots+(x_{1p}-\bar{x}_p^{(1)})^2+\cdots(x_{n_1 p}-\bar{x}_p^{(1)})^2$

二组的离差平方和 $=(x_{11}-\bar{x}_1^{(2)})^2+\cdots+(x_{n_2 1}-\bar{x}_1^{(2)})^2$

$\cdots+(x_{1p}-\bar{x}_p^{(2)})^2+\cdots(x_{n_2 p}-\bar{x}_p^{(2)})^2$

可以证明：

总离差平方和

＝组内离差平方和＋组间离差平方和

令T为总离差平方和

令P_G为分为G类的组内离差平方和。

2、统计量 $R^2=1-\frac{P_G}{T}$

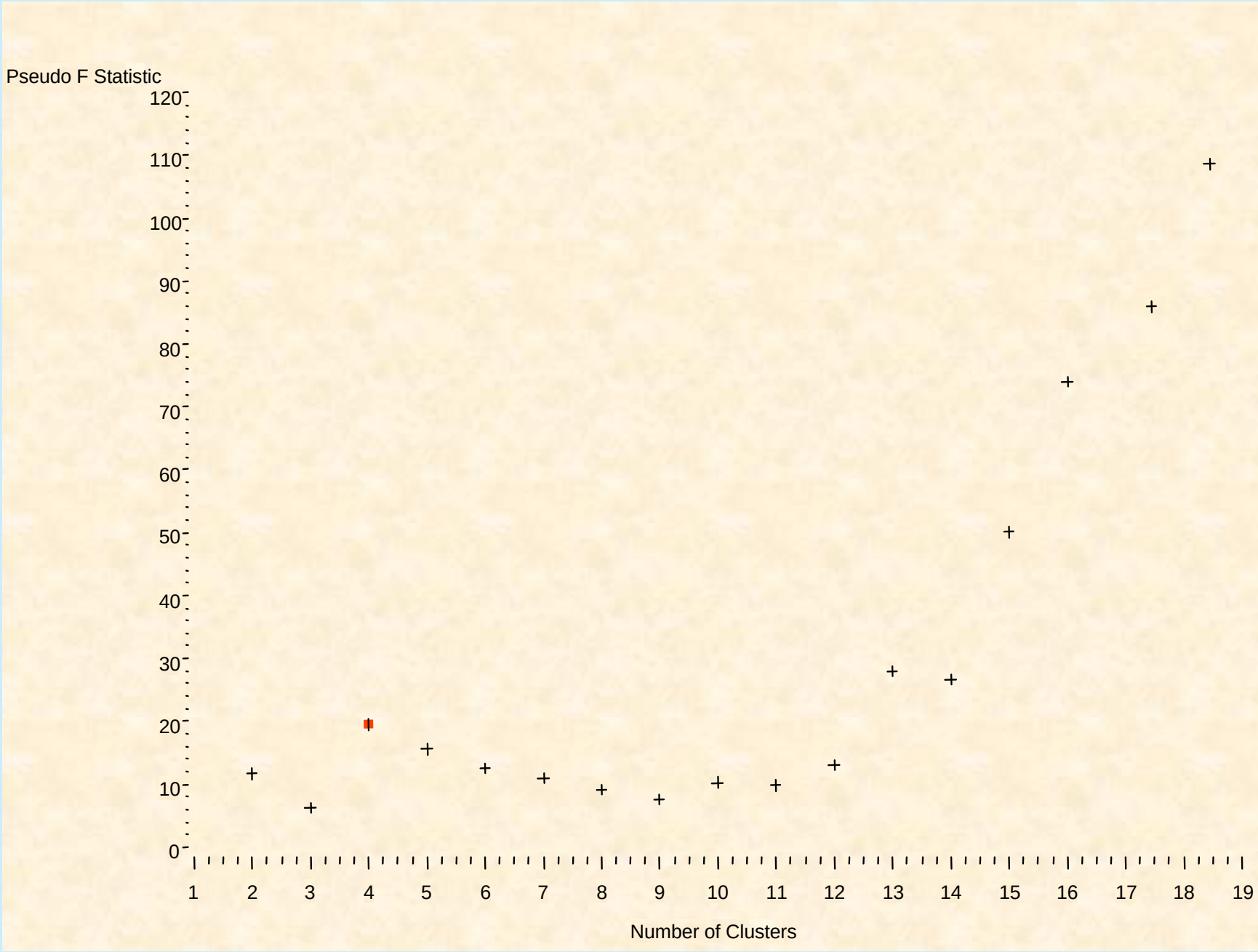
其中T是数据的总离差平方和，P_G是组内离差平方和。

R^2比较大，说明分G个类时类内的离差平方和比较小，也就是说分G类是合适的。但是，分类越多，每个类的类内的离差平方和就越小，R^2也就越大；所以我们只能取合适的G，使得 R^2足够大，而G本生很 R^2小，随着G的增加，R^2 的增幅不大。比如，假定分4类时， R^2 =0.8；下一次合并分三类时，下降了许多，=0.32，则分4 类是合适的。

3、伪F统计量的定义为

$$F=\frac{(T-P_G)/(G-1)}{P_G/(n-G)}$$

伪F统计量用于评价聚为G类的效果。如果聚类的效果好，类间的离差平方和相对于类内的离差平方和大，所以应该取伪F统计量较大而类数较小的聚类水平。


Pseudo F Statistic
120
110
100
90
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
Number of Clusters

4、伪 t^2 统计量的定义为

$$t^2 = \frac{B_{KL}}{(W_K + W_L)/(N_K + N_L - 2)}$$

其中W_L和 W_K分别是的类内离差平方和，是将K和L合并为第M类的离差平方和

$$B_{KL} = W_M - W_L - W_K$$

为合并导致的类内离差平方和的增量。用它评价合并第K和L类的效果，伪t^2 统计量大说明不应该合并这两类，应该取合并前的水平。

五、 系统聚类法的基本性质

（一） 单调性

在聚类分析过程中，并类距离分别为l_k（k=1，2，3，……）若满足 $l_1 < l_2 < \cdots < l_k < l_{k+1} < \cdots$ ，则称该聚类方法具有单调性。可以证明除了重心法和中间距离法之外，其他的系统聚类法均满足单调性的条件。

（二）空间的浓缩和扩张

1、 定义矩阵的大小

设同阶矩阵D（A）和D（B），如果D（A）的每一个元素不小于D（B）的每一个元素，则记为 $D(A) \geq D(B)$。

2、空间的浓缩和扩张

设有两种系统聚类法A和B，他们在第i步的距离矩阵分别为A_i和B_i（I=1，2，3...），若$A_i>B_i$，则称第一种方法A比第二种方法B使空间扩张，或第二种方法比第一种方法浓缩。

3、方法的比较

D（短）≤ D（平），D（重）≤ D（平）；
D（长）≥ D（平）；
当$0<\beta<1$，D（变平）≤ D（平）；
当 $\beta<0$ ，D（变平）≥ D（平）。

六、主要的步骤

1、选择变量

（1）和聚类分析的目的密切相关

（2）反映要分类变量的特征

（3）在不同研究对象上的值有明显的差异

（4）变量之间不能高度相关

2、计算相似性

相似性是聚类分析中的基本概念，他反映了研究对象之间的亲疏程度，聚类分析就是根据对象之间的相似性来分类的。有很多刻画相似性的测度

3、聚类

选定了聚类的变量，计算出样品或指标之间的相似程度后，构成了一个相似程度的矩阵。这时主要涉及两个问题：

（1）选择聚类的方法

（2）确定形成的类数

4、聚类结果的解释和证实

- 对聚类结果进行解释是希望对各个类的特征进行准确的描述，给每类起一个合适的名称。这一步可以借助各种描述性统计量进行分析，通常的做法是计算各类在各聚类变量上的均值，对均值进行比较，还可以解释各类产别的原因。

- 如果是变量聚类分析，聚类分析做完之后，各类中仍有较多的指标。也就是说聚类分析并没有达到降维的目的。这就需要在每类中选出一个代表指标，具体做法是：假设某类中有 k 个指标，首先分别计算类内指标之间的相关指数 $\gamma_{ij}^2 \quad (i \neq j)\,(i = 1,2,\cdots,k)$，然后计算某个指标与类内其他指标之间相关指数的平均数，即

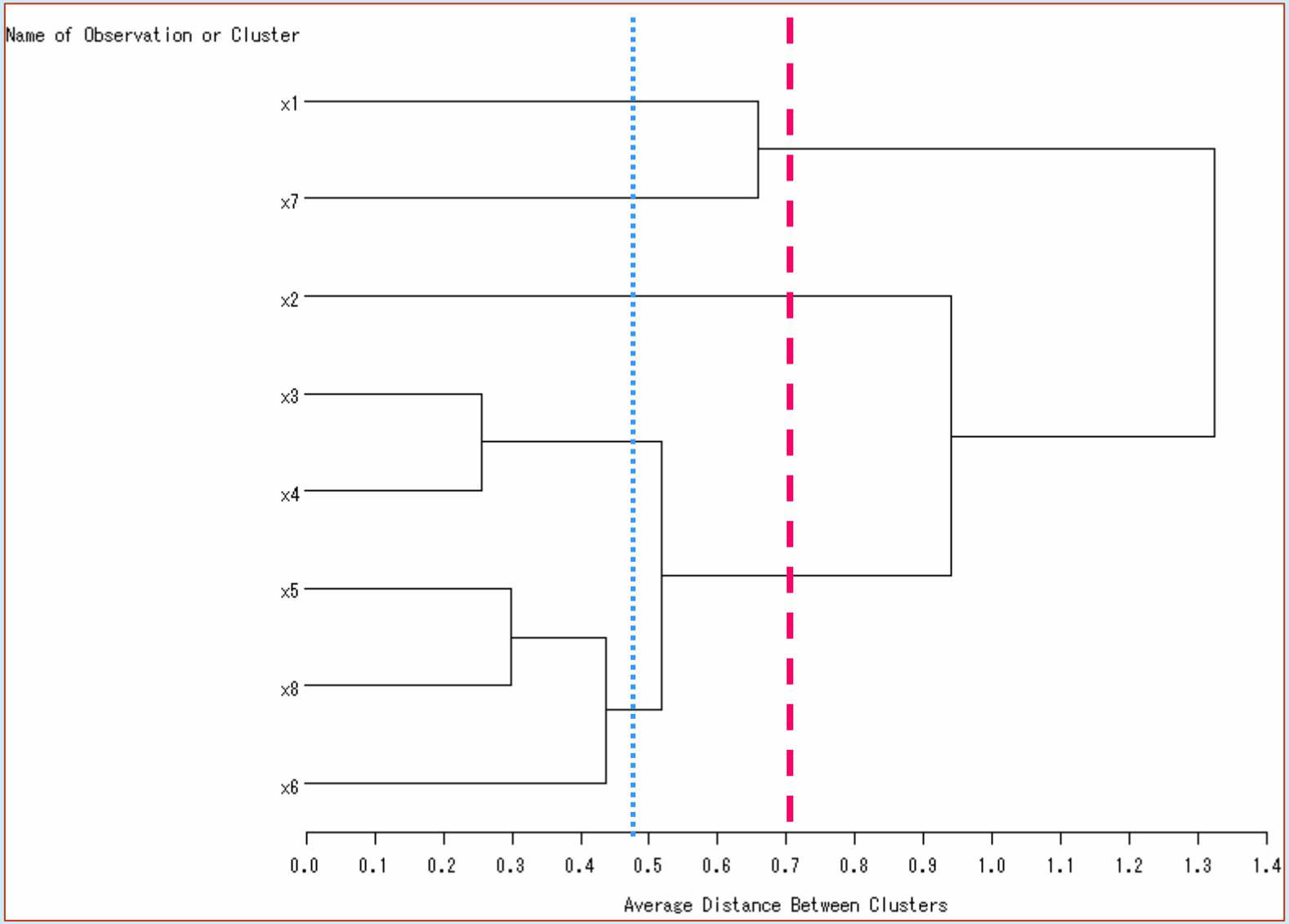
$$\overline{R}_i^2 = \frac{\sum\limits_{i \neq j} \gamma_{ij}}{k-1}$$

取 x_i 最大的 $\overline{R}_i^2$ ，做为该类的代表。

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

例（于秀林教材91页例）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 l—每千人户居民拥有的蜂窝移动电话数；

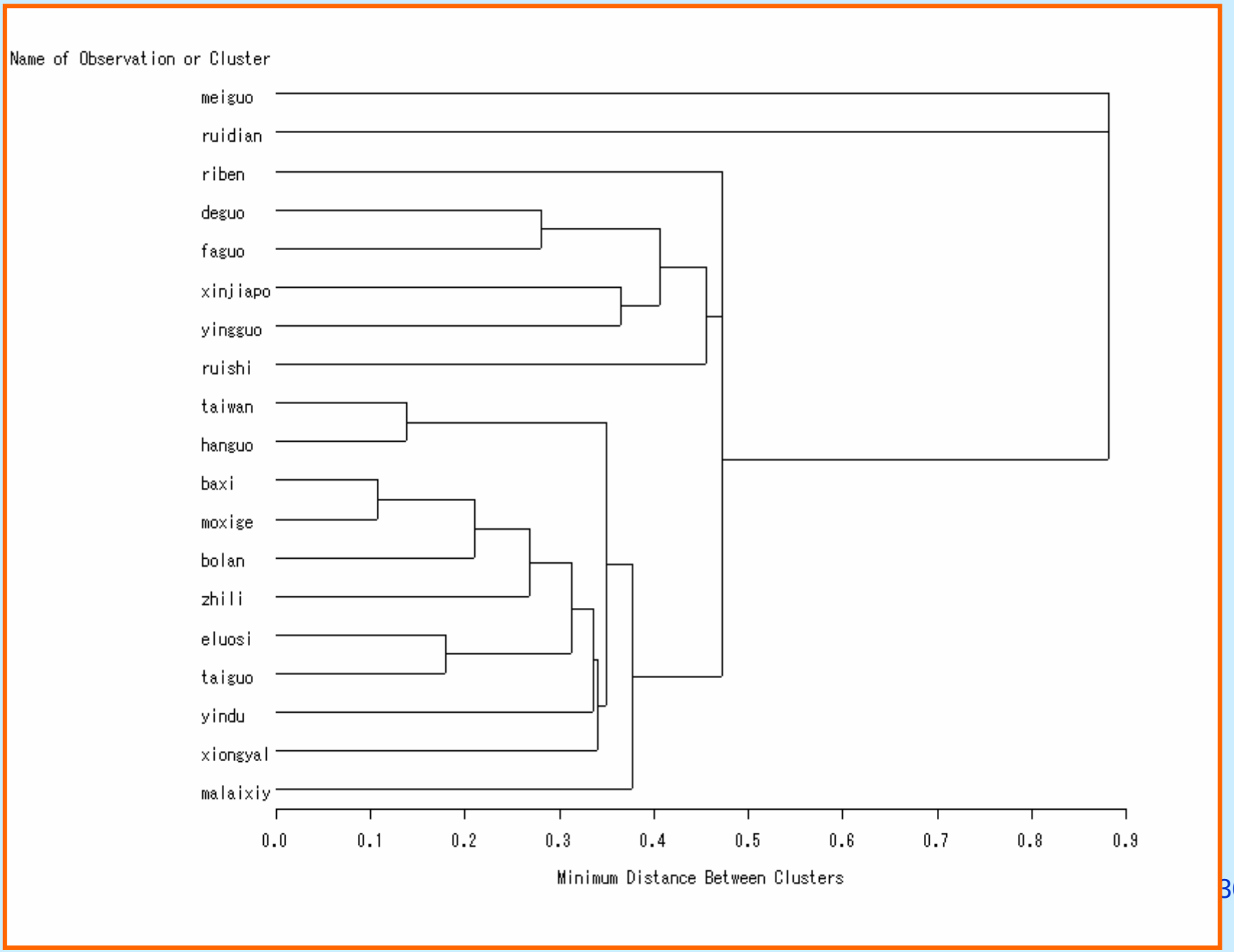
fee—高峰时期每三分钟国际电话的成本；

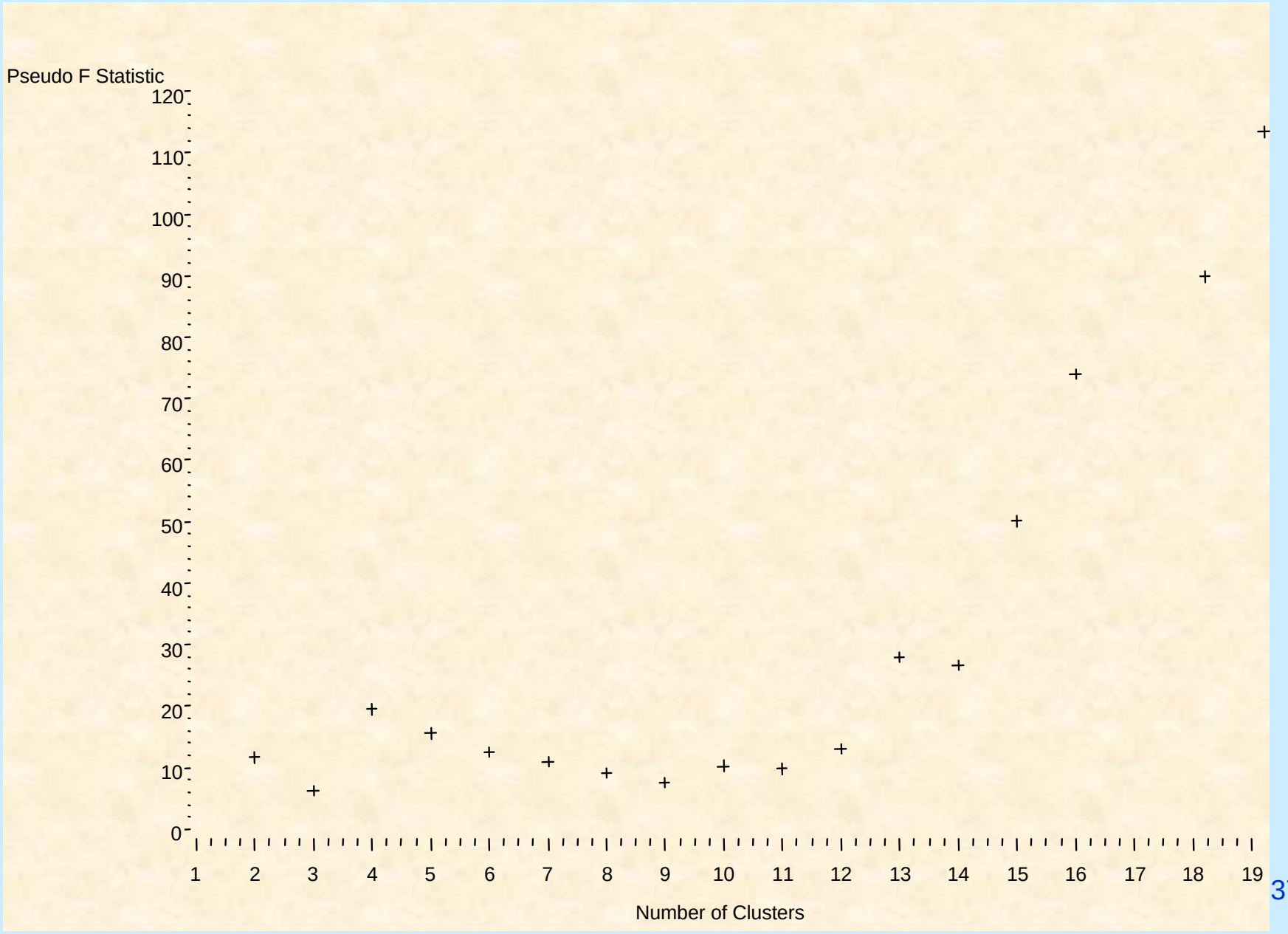
comp—每千人拥有的计算机数；

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

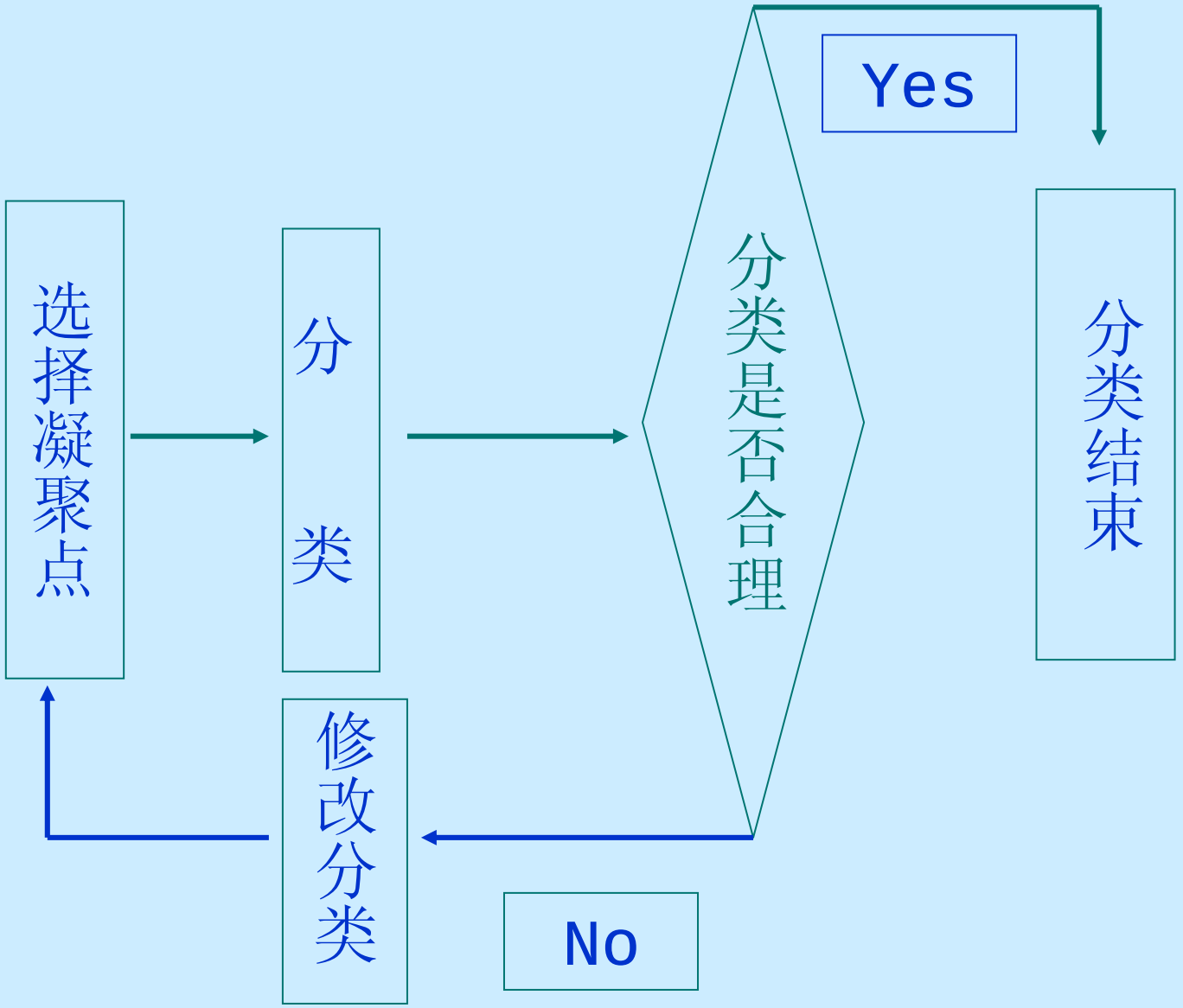

Pseudo F Statistic
120
110
100
90
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
Number of Clusters

§2 动态（快速）聚类

一、思想

系统聚类法是一种比较成功的聚类方法。然而当样本点数量十分庞大时，则是一件非常繁重的工作，且聚类的计算速度也比较慢。比如在市场抽样调查中，有4万人就其对衣着的偏好作了回答，希望能迅速将他们分为几类。这时，采用系统聚类法就很困难，而动态聚类法就会显得方便，适用。

动态聚类解决的问题是：假如有个样本点，要把它们分为类，使得每一类内的元素都是聚合的，并且类与类之间还能很好地区别开。动态聚类使用于大型数据。


选择凝聚点
分 类
分类是否合理
Yes
分类结束
修改分类
No

用一个简单的例子来说明动态聚类法的工作过程。例如我们要把图中的点分成两类。快速聚类的步骤：

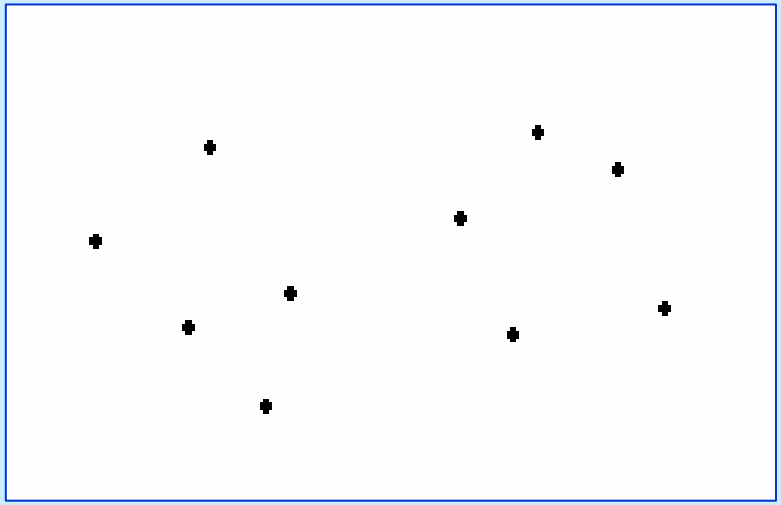
1、随机选取两个点 $x_1^{(1)}$ 和 $x_2^{(1)}$作为聚核。

2、对于任何点x_k，分别计算$d(x_k, x_1^{(1)})$和$d(x_k, x_2^{(1)})$

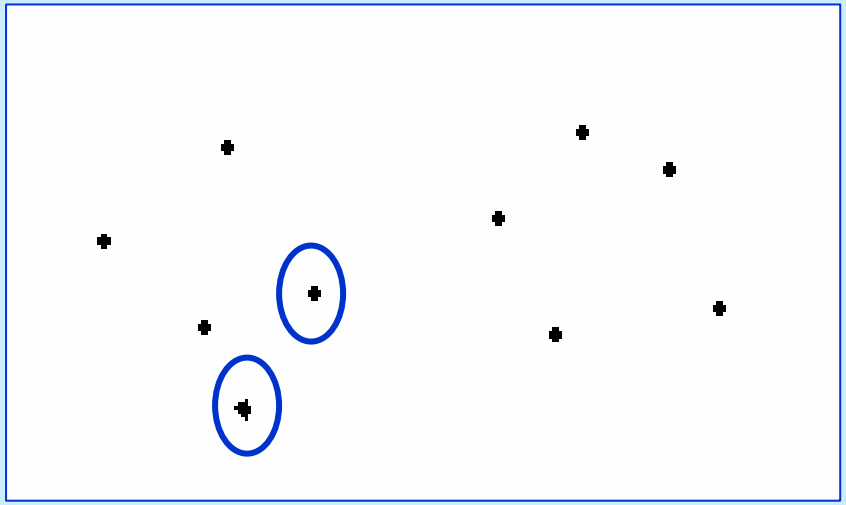
3、若$d(x_k, x_1^{(1)}) \prec d(x_k, x_2^{(1)})$，则将 x_k划为第一类，否则划给第二类。于是得图（b）的两个类。

4、分别计算两个类的重心，则得 $x_1^{(2)}$ 和 $x_2^{(2)}$，以其为新的聚核，对空间中的点进行重新分类，得到新分类。

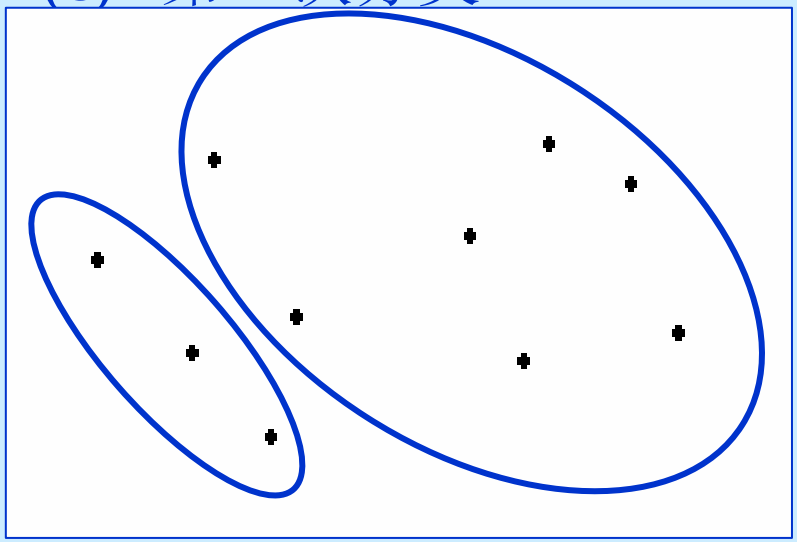
（a）空间的群点



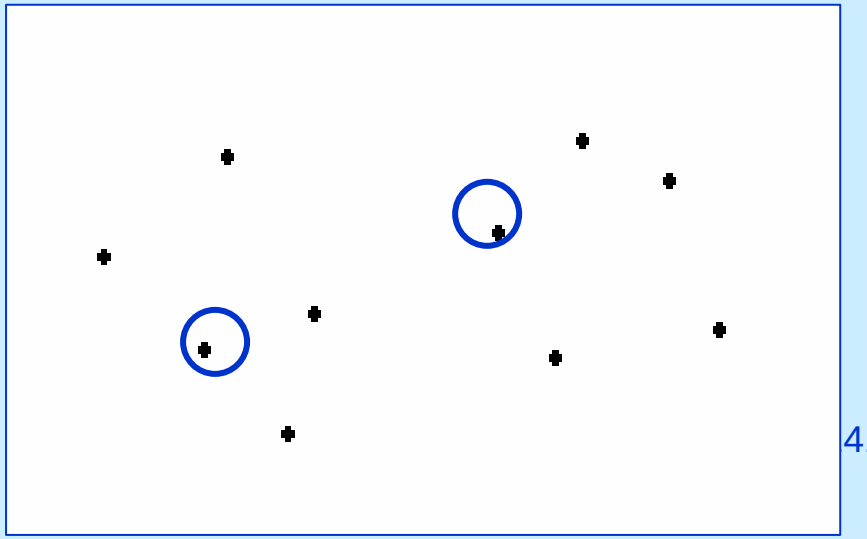
(b) 任取两个聚核



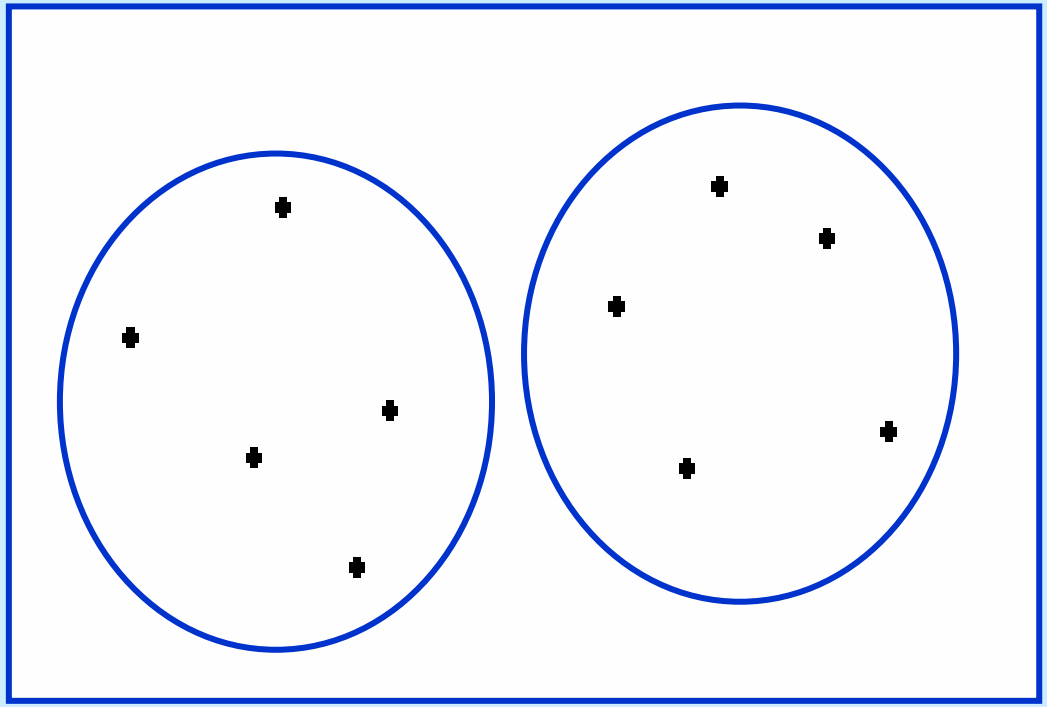
(c) 第一次分类



(d) 求各类中心



(e) 第二次分类



二、选择凝聚点和确定初始分类

凝聚点就是一批有代表性的点，是欲形成类的中心。凝聚点的 选择直接决定初始分类，对分类结果也有很大的影响，由于凝聚点 的不同选择，其最终分类结果也将出现不同。故选择时要慎重．通 常选择凝聚点的方法有：

(1) 人为选择，当人们对所欲分类的问题有一定了解时，根据经验，预先确定分类个数和初始分类，并从每一类中选择一个有代表性的样品作为凝聚点。

(2) 将数据人为地分为A类，计算每一类的重心，就将这些重心作为凝聚点。

(3) 用密度法选择凝聚点。以某个正数d为半径，以每个样品为球心，落在这个球内的样品数(不包括作为球心的样品)就叫做这个样品的密度。计算所有样品点的密度后，首先选择密度最大的样品作为第一凝聚点，并且人为地确定一个正数D(一般D＞ d，常取D＝2d)。然后选出次大密度的样品点，若它与第一个凝 聚点的距离大于D，则将其作为第二个凝聚点；否则舍去这点，再选密度次于它的样品。这样，按密度大小依次考查，直至全部样品考查完毕为止．此方法中，d要给的合适，太大了使凝聚点个数太 少，太小了使凝聚点个数太多。

(4) 人为地选择一正数d，首先以所有样品的均值作为第一凝聚点。然后依次考察每个样品，若某样品与已选定的凝聚点的距 离均大于d，该样品作为新的凝聚点，否则考察下一个样品。

(5) 随机地选择，如果对样品的性质毫无所知，可采用随机数表来选择，打算分几类就选几个凝聚点。或者就用前A个样品作为凝聚点(假设分A类)。这方法一般不提倡使用。

三、衡量聚类结果的合理性指标和算法终止的标准

定义　设 P_i^n 表示在第n次聚类后得到的第i类集合，$i=1,2,3,\cdots,k$ ，$A_i^{(n)}$ 为第n次聚类所得到的聚核。

定义　$u_n \triangleq \sum_{i=1}^{k} \sum_{x \in P_i^n} d^2(x, A_i^{(n)})$

$$A_i^j = \frac{1}{n_i} \sum_{x_l \in P_i^j} x_l \qquad j=1,2,\cdots$$

第j+1步的新聚核。

若分类不合理时，$u_n \triangleq \sum_{i=1}^{k} \sum_{x \in P_i^n} d^2(x, A_i^{(n)})$ 会很大，随着分类的过程，逐渐下降，并趋于稳定。

定义　第i类中所有元素与其重心的距离的平方和:

$$D(A_i^n, P_i^n) \triangleq \sum_{x_l \in P_i^n} d^2(x_l, A_i^n)$$

$$u_n \triangleq \sum_{i=1}^{k} \sum_{x_l \in P_i^n} d^2(x_l, A_i^{(n)}) = \sum_{i=1}^{k} D(A_i^n, P_i^n)$$

为所有K个类中所有元素与其重心的距离的平方和。

定义算法终止的标准是 $\dfrac{|u_{n+1} - u_n|}{u_{n+1}} \le \varepsilon$

ε是事前给定的一个充分小量　。

五、动态聚类步骤为：

第一，选择若干个观测值点为“凝聚点”；

第二，可选择地，通过分配每个“凝聚点”最近的类里来形成临时分类。每一次对一个观测值点进行归类，“凝聚点”更新为这一类目前的均值；

- 第三，可选择地，通过分配每个“凝聚点”最近的类里来形成临时分类。所有的观测值点分配完后，这些类的“凝聚点”用临时类的均值代替。该步骤可以一直进行直到“凝聚点”的改变很小或为零时止；
- 第四，最终的分类有分配每一个观测到最近的“凝聚点”而形成。

例 我国经济发展的总目标是到2000年人民生活达到小康标准，因此，了解各地区目前对小康生活质量的实现程度。对各地区实现小康生活质量的状况进行综合评价，对各级政府部门具有重要意义。数据是1990年全国30个省在经济（jj）、教育（jy）、健康（jk）和居住环境（jz）四个方面对小康标准已经实现的程度，1表示已经达到或超过小康水平，0表示低于或多或少刚达到温饱水平。希望利用该数据对15个地区进行分类研究。

	jj	jy	jk	jz	类别	距离
beijngsh	0.7258	0.9413	1.0000	0.5000	1	0.29550
anghai	0.5346	0.9848	1.0000	0.5000	1	0.14909
ianjin	0.3246	0.9733	1.0000	0.5000	1	0.16173
henna	0.2301	0.4621	1.0000	1.0000	2	0.22252
ejiang	0.5025	0.2374	1.0000	0.8882	2	0.34448
jilin	0.3446	0.7755	0.8280	0.5000	1	0.18212
elongji	0.2891	0.7835	0.8080	0.5000	1	0.22322
fujian	0.1406	0.3524	1.0000	0.7102	2	0.27468
uangxi	0.0939	0.6498	0.4435	1.0000	2	0.51560
anhui	0.1104	0.0802	1.0000	0.9545	2	0.34050
ingxia	0.2708	0.3127	0.5425	0.9053	2	0.29445
hunan	0.0618	0.5687	0.4385	0.5000	3	0.41704
jiangxi	0.0549	0.3042	0.3520	0.6155	3	0.15540
inghai	0.0751	0.0118	0.0000	0.8258	3	0.37720
uizhou	0.0286	0.0600	0.0590	0.5000	3	0.25968

四、有序样本聚类法

（一）功能范畴与数据类型

有序样本聚类法又称为最优分段法。该方法是由费歇在1958年提出的。它主要适用于样本由一个变量描述的情况。或者将多变量综合成为一个变量来分析。

设 $\Omega=\{\varpi_1,\varpi_2,\cdots,\varpi_n\}$ 是样本点构成的集合，样本点ϖ_i在函数$V(\varpi)$上的取值为 v_i 。若 $V(\varpi_i)=V(\varpi_j)$，则将视为一类。不妨假设 $v_1<v_2<\cdots<v_m$要将 $v_1,v_2,\cdots,v_m$分为 K 类；即 $P=(P_1,P_2,\cdots,P_k)$分类时不能打乱样本点的顺序，即每一类必须呈的 $\{\varpi_1,\varpi_2,\cdots$形式，即有序样本聚类。

系统聚类开始n个样品各自自成一类，然后逐步并类，直至所有的样品被聚为一类为止。而有序聚类则相反，开始所有的样品为一类，然后分为二类、三类等，直到分成n类。每次分类都要求产生的离差平方和的增量最小。

例 $\Omega=\{\varpi_1,\varpi_2,\cdots,\varpi_4\}$

$$\nu_1=V(\varpi_1)=1.1$$

$$\nu_2=V(\varpi_2)=V(\varpi_3)=1.3$$

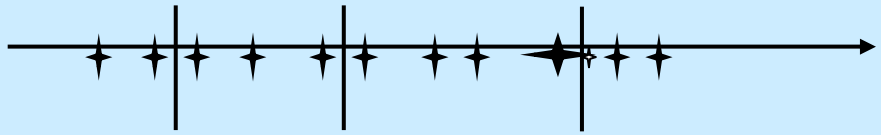
$$\nu_3=V(\varpi_4)=3.1$$

这里n=4，m=3。若将其分为两类，其结果应该是

$$P=\{(\nu_1,\nu_2),\nu_3\}$$

对应中的点是$\{(w_1,w_2,w_2),w_4\}$。

有序样本聚类法常常被用于系统的评估问题，被用来对样本点进行分类划级。例如，十二个地区的经济发展指数，排列出来以后，需要划分他们的等级。一种方法是按照行政命令。规定三个经济发达地区，四个中等发达的地区，三个一般地区，两个发展较差地区。



这种行政上的规定往往是不客观、不合理的。合理的分类应该把发展情况最近似的地区划入同一类。这就是有序样本聚类的工作思路。

（二）有序聚类的步骤

设有序样品x(1),x(2),…,x(n)。他们可以是从小到达排列，也可以是按时间的先后排列。

1、定义类的直径

设某类G中包含的样品有

$$\{\mathbf{X}_{(i)},\mathbf{X}_{(i+1)},\cdots,\mathbf{X}_{(j)}\} \qquad (j>i)$$

该类的均值向量为

$$\overline{\mathbf{X}}_{\mathbf{G}}=\frac{1}{j-i+1}\sum_{t=i}^{j}\mathbf{X}_{(t)}$$

用$D(i,j)$表示这一类的直径，常用的直径有，欧氏距离：

$$D(i,j)=\sum_{t=i}^{j}(\mathbf{x}_{(t)}-\overline{\mathbf{X}}_{\mathbf{G}})'(\mathbf{x}_{(t)}-\overline{\mathbf{X}}_{\mathbf{G}})$$

当是单变量的时，也可以定义直径为：

$$D(i,j)=\sum_{t=i}^{j}|x_{(t)}-\tilde{X}_G| \qquad \tilde{X}_G\text{是中位数}$$

2、定义分类的损失函数

用b(n,k)表示将n个有序的样品分为k类的某种分法：

$$G_1 = \{j_1, j_1+1, \cdots, j_2-1\}$$

$$G_2 = \{j_2, j_2+1, \cdots, j_3-1\}$$

$$\cdots \qquad \cdots$$

$$G_k = \{j_k, j_k+1, \cdots, n\}$$

定义这种分类法的损失函数为：各类的直径之和。

$$L[b(n,k)] = \sum_{t=1}^{k} D(i_t, i_{t+1}-1)$$

由损失函数的构造可以看出，损失函数是各类的直径之和。如果分类不好，则各类的直径之和大，否则比较小。

当n和k固定时， L[b(n,k)]越小表示各类的离差平方和越小，分类是合理的。因此要寻找一种分法b(n,k)，使分类损失函数L[b(n,k)]达到最小。记该分法为P[n,k]。

3、最优解的求法

若分类数k是已知的，求分类法b(n,k)，使它在损失函数意义下达到最小，其求法如下：

首先，找出分点j_k，使

$$P(n,2):\{1,2,\cdots,j_k-1\},\{j_k,j_k+1,\cdots,n\}$$

且$$L[P(n,2)]=\min_{2\le j\le n}[D(1,j_k-1)+D(j_k,n)]$$

于是得第k类 $G_k=\{j_k,j_k+1,\cdots,n\}$

然后，找出j_{k-1}，使它满足

$$P(j_k-1,2):\{1,2,\cdots,j_{k-1}-1\},\{j_{k-1},j_{k-1}+1,\cdots,j_k-1\}$$

且$$L[P(j_k-1,2)]=\min_{2\le j\le j_k-1}[D(1,j_{k-1}-1)+D(j_{k-1},j_k-1)]$$

于是得第k-1类 $$G_{k-1}=\{j_{k-1},j_{k-1}+1,\cdots,j_k-1\}$$

再然后，找出j_{k-2}，使它满足

$$P(j_{k-1}-1,2):\{1,2,\cdots,j_{k-2}-1\},\{j_{k-2},j_{k-2}+1,\cdots,j_{k-1}-1\}$$

且$$L[P(j_{k-1}-1,2)]=\min_{2\le j\le j_{k-1}-1}[D(1,j_{k-2}-1)+D(j_{k-2},j_{k-1}-1)]$$

于是得第k-2类 $G_{k-2}=\{j_{k-2},j_{k-2}+1,\cdots,j_{k-1}-1\}$

类推。一直可以得到所有类G1，G2，...Gk，这就是所求得最优解。

4、L[b(n,k)]的递推公式

$$\begin{cases} L[p(n,2)] = \min\limits_{2\le j\le n}\{D(1,j-1)+D(j,n)\} \\ L[p(n,k)] = \min\limits_{k\le j\le n}\{D(j-1,k-1)+D(j,n)\} \end{cases}$$

以上的两个公式的含义是，如果要找到n个样品分为k个类的最优分割，应建立在将j-1（j＝2,3,…,n)个样品分为k-1类的最优分割的基础上。

分析儿童的生长期。有如下的资料是1-11岁的男孩平均每年的增重：

年龄	1	2	3	4	5	6	7	8	9	10	11
增加重量（公斤）	9.3	1.8	1.9	1.7	1.5	1.3	1.4	2.0	1.9	2.3	2.1

问男孩的发育可分为几个阶段。

$D(i, j)$

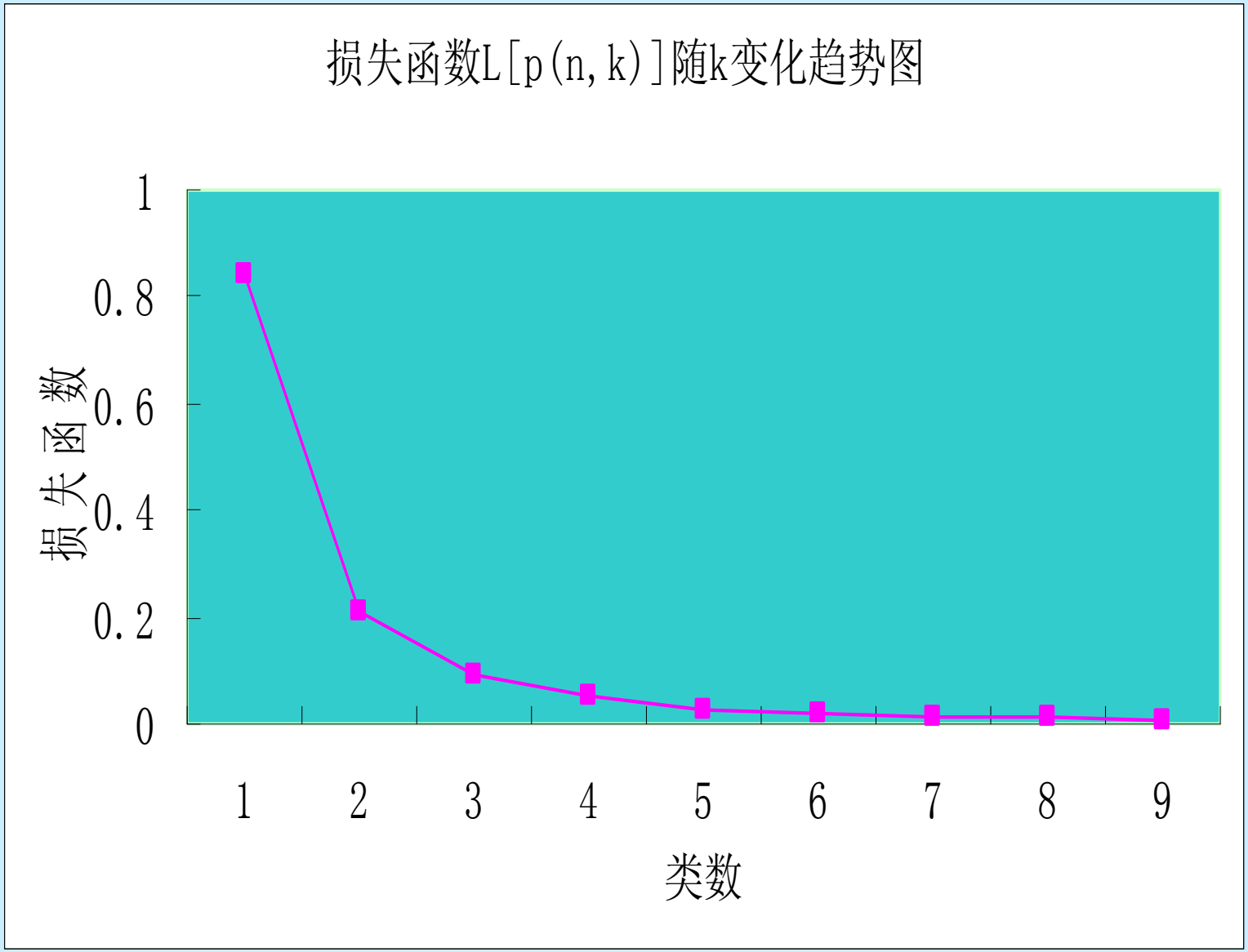
	i									
j	1	2	3	4	5	6	7	8	9	10
2	28.125									
3	37.007	0.005								
4	42.208	0.02	0.02							
5	45.992	0.088	0.08	0.02						
6	49.128	0.232	0.2	0.08	0.02					
7	51.1	0.28	0.232	0.088	0.02	0.005				
8	51.529	0.417	0.393	0.308	0.29	0.287	0.18			
9	51.98	0.469	0.454	0.393	0.388	0.37	0.207	0.00508		
10	52.029	0.802	0.8	0.774	0.773	0.708	0.42	0.087	0.08	
11	52.182	0.909	0.909	0.895	0.889	0.793	0.452	0.088	0.08	0.02

$$D(1,2) = (9.3 - 5.55)^2 + (1.8 - 5.55)^2 = 28.125$$

$$D(1,3) = (9.3 - 4.333)^2 + (1.8 - 4.333)^2 + (1.9 - 4.333)^2$$
$$= 37.007$$

最小损失函数L[p(n,k)

n	k								
	2	3	4	5	6	7	8	9	10
3	0.005/2								
4	0.02/2	0.005/4							
5	0.088/2	0.020/5	0.005/5						
6	0.232/2	0.040/5	0.02/6	0.005/6					
7	0.280/2	0.040/5	0.025/6	0.010/6	0.005/6				
8	0.417/2	0.280/8	0.040/8	0.025/8	0.010/8	0.005/8			
9	0.469/2	0.285/8	0.045/8	0.030/8	0.015/8	0.010/8	0.005/8		
10	0.802/2	0.367/8	0.127/8	0.045/10	0.030/10	0.015/10	0.010/10	0.005/10	
11	0.909/2	0.368/8	0.128/8	0.065/10	0.045/11	0.030/11	0.015/11	0.010/11	0.005/11


损失函数L[p(n,k)]随k变化趋势图
1
0.8
0.6
0.4
0.2
0
损失函数
1
2
3
4
5
6
7
8
9
类数

分类数	误差函数	最优分割结果
2	0.8392	1,2-11
3	0.21069	1,2-6,7-11
4	0.09199	1,2-4,5-7,8-11
5	0.05102	1,2-4,5-7,8-9,10-11
6	0.02586	1,2-3,4-5,6-7,8-9,10-11
7	0.02055	1,2-3,4-5,6-7,8-9,10,11
8	0.01523	1,2-3,4,5,6-7,8-9,10,11
9	0.01016	1,2-3,4,5,6,7,8-9,10,11
10	0.00508	1,2-3,4,5,6,7,8,9,10,11

第四章 判别分析

判别

- 有一些昆虫的性别很难看出，只有通过解剖才能够判别；
- 但是雄性和雌性昆虫在若干体表度量上有些综合的差异。于是统计学家就根据已知雌雄的昆虫体表度量（这些用作度量的变量亦称为预测变量）得到一个标准，并利用这个标准来判别其他未知性别的昆虫。
- 这样的判别虽然不能保证百分之百准确，但至少大部分判别都是对的，而且用不着杀死昆虫来进行判别了。

判别分析(discriminant analysis)

- 这就是本章要讲的是判别分析。
- 判别分析和前面的聚类分析有什么不同呢？
- 主要不同点就是，在聚类分析中一般人们事先并不知道或一定要明确应该分成几类，完全根据数据来确定。
- 而在判别分析中，至少有一个已经明确知道类别的“训练样本”，利用这个数据，就可以建立判别准则，并通过预测变量来为未知类别的观测值进行判别了。

判别分析例子

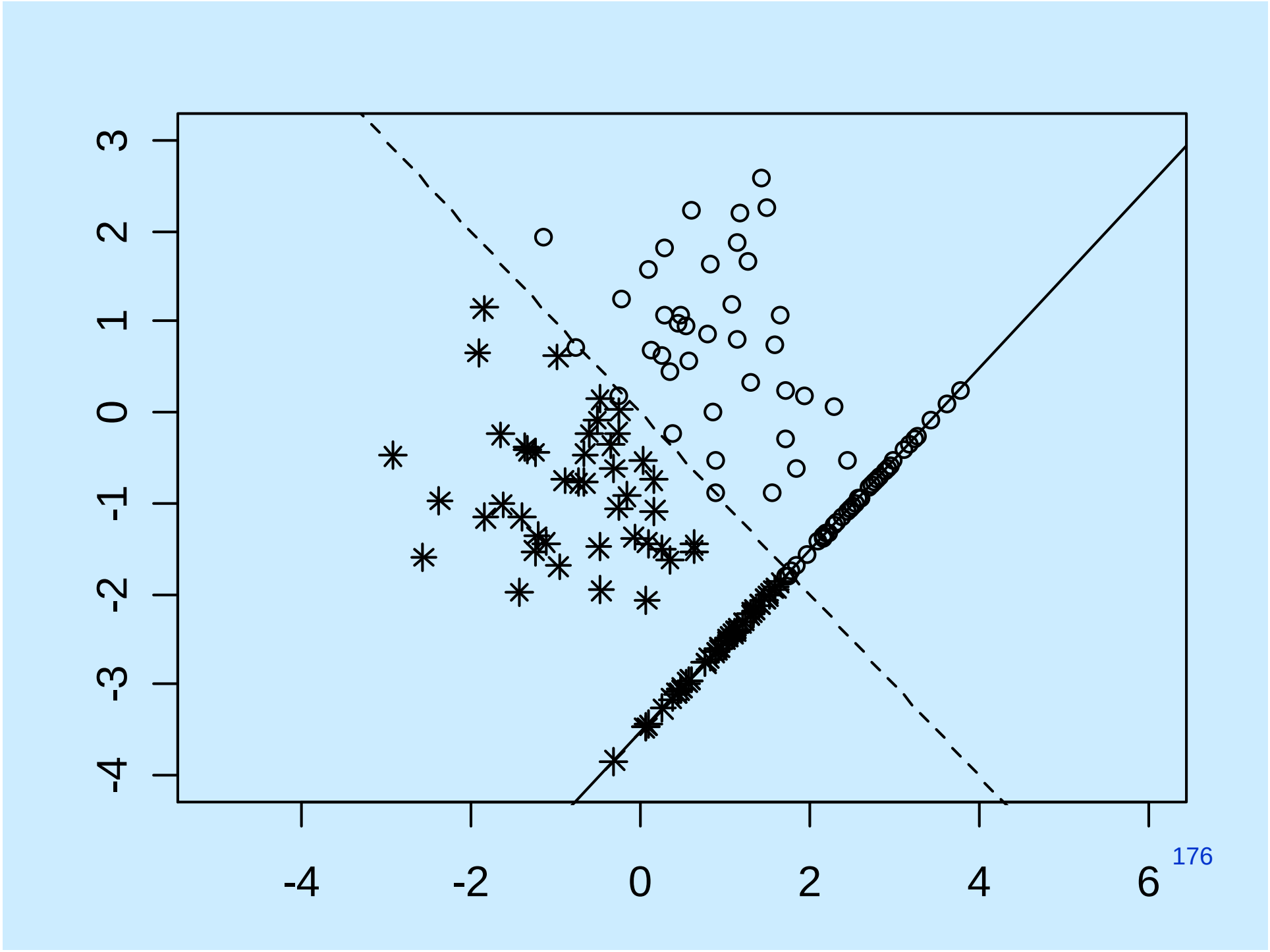
- 数据**Disc.sav** 企业打分:企图用一套打分体系来描绘企业的状况。该体系对每个企业的一些指标（变量）进行评分。
- 这些指标包括：企业规模（**is**）、服务**(se)**、雇员工资比例**(sa)**、利润增长**(prr)**、市场份额**(ms)**、市场份额增长**(msr)**、流动资金比例**(cp)**、资金周转速度**(cs)**等等。
- 另外，有一些企业已经被某杂志划分为上升企业、稳定企业和下降企业。
- 我们希望根据这些企业的上述变量的打分和它们已知的类别（三个类别之一：**group-1**代表上升，**group-2**代表稳定，**group-3**代表下降）找出一个分类标准，以对没有被该刊物分类的企业进行分类。
- 该数据有**90**个企业（**90**个观测值），其中**30**个属于上升型，**30**个属于稳定型，**30**个属于下降型。这个数据就是一个**“训练样本”**。

根据距离判别的思想

- **Disc.sav**数据有**8**个用来建立判别标准**(**或判别函数**)**的（预测）变量，另一个（**group**）是类别。
- 因此每一个企业的打分在这**8**个变量所构成的**8**维空间中是一个点。这个数据有**90**个点，
- 由于已经知道所有点的类别了，所以可以求得每个类型的中心。这样只要定义了如何计算距离，就可以得到任何给定的点（企业）到这三个中心的三个距离。
- 显然，最简单的办法就是离哪个中心距离最近，就属于哪一类。通常使用的距离是所谓的**Mahalanobis**距离。用来比较到各个中心距离的数学函数称为判别函数**(discriminant function).**这种根据远近判别的方法，原理简单，直观易懂。

Fisher判别法(先进行投影)

- 所谓**Fisher**判别法，就是一种先投影的方法。
- 考虑只有两个（预测）变量的判别分析问题。
- 假定这里只有两类。数据中的每个观测值是二维空间的一个点。见图。
- 这里只有两种已知类型的训练样本。其中一类有**38**个点（用**“o”**表示），另一类有**44**个点（用**“*”**表示）。按照原来的变量（横坐标和纵坐标），很难将这两种点分开。
- 于是就寻找一个方向，也就是图上的虚线方向，沿着这个方向朝和这个虚线垂直的一条直线进行投影会使得这两类分得最清楚。可以看出，如果向其他方向投影，判别效果不会比这个好。
- 有了投影之后，再用前面讲到的距离远近的方法来得到判别准则。这种首先进行投影的判别方法就是**Fisher**判别法。


3
2
1
0
-1
-2
-3
-4
-4
-2
0
2
4
6

Fisher判别法的数学

逐步判别法(仅仅是在前面的方法中加入变量选择的功能)

- 有时，一些变量对于判别并没有什么作用，为了得到对判别最合适的变量，可以使用逐步判别。也就是，一边判别，一边引进判别能力最强的变量，
- 这个过程可以有进有出。一个变量的判别能力的判断方法有很多种，主要利用各种检验，例如**Wilks' Lambda**、**Rao's V**、**The Squared Mahalanobis Distance**、**Smallest F ratio**或**The Sum of Unexplained Variations**等检验。其细节这里就不赘述了；这些不同方法可由统计软件的各种选项来实现。逐步判别的其他方面和前面的无异。

Disc.sav例子

- 利用**SPSS**软件的逐步判别法淘汰了不显著的流动资金比例**(cp)**，还剩下七个变量。用$x_1, x_2, x_3, x_4, x_5, x_6, x_7$分别表示标准化后的变量**is**，**se**，**sa**，**prr**，**ms**，**msr**，**cs**，得到两个典则判别函数（**Canonical Discriminant Function Coefficients**）：

$$F1 = -3.166 + 0.035x_1 + 3.283x_2 + 0.037x_3 - 0.007x_4 + 0.068x_5 - 0.023x_6 - 0.385x_7$$

$$F2 = -4.384 + 0.005x_1 + 0.567x_2 + 0.041x_3 + 0.012x_4 + 0.048x_5 + 0.044x_6 - 0.159x_7$$

这两个函数实际上是由**Fisher**判别法得到的向两个方向的投影。这两个典则判别函数的系数是下面的**SPSS**输出得到的：

Disc.sav例子

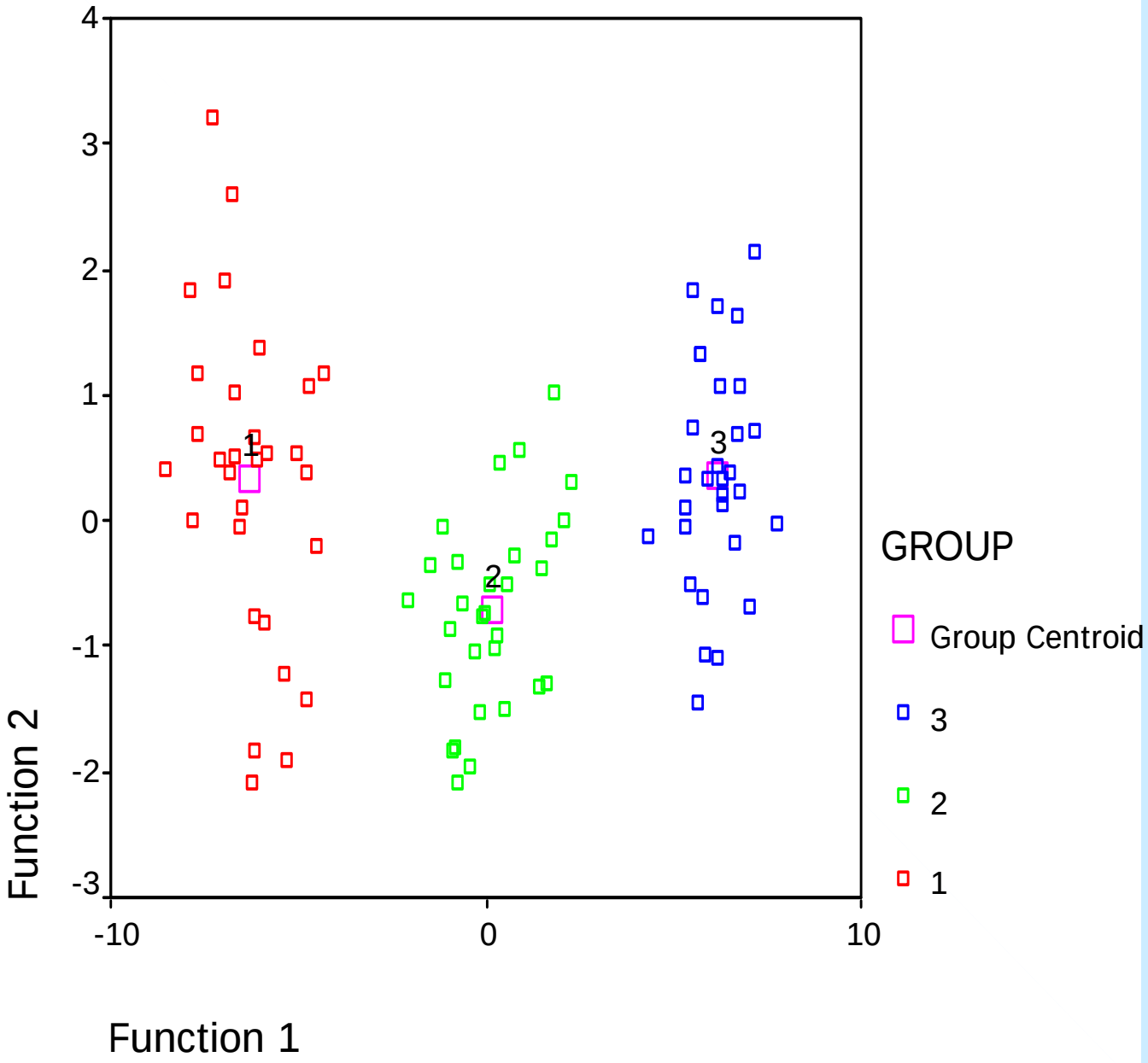
- 根据这两个函数，从任何一个观测值（每个观测值都有**7**个变量值）都可以算出两个数。把这两个数目当成该观测值的坐标，这样数据中的**150**个观测值就是二维平面上的**150**个点。它们的点图在下面图中。

Canonical Discriminant Function Coefficients

	Function	
	1	2
IS	.035	.005
SE	3.283	.567
SA	.037	.041
PRR	-.007	.012
MS	.068	.048
MSR	-.023	.044
CS	-.385	-.159
(Constant)	-3.166	-4.384

Unstandardized coefficients

Canonical Discriminant Functions



Disc.sav例子

- 从上图可以看出，第一个投影（相应于来自于第一个典则判别函数横坐标值）已经能够很好地分辨出三个企业类型了。这两个典则判别函数并不是平等的。其实一个函数就已经能够把这三类分清楚了。SPSS的一个输出就给出了这些判别函数（投影）的重要程度：

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	26.673[a]	99.0	99.0	.982
2	.262[a]	1.0	100.0	.456

a. First 2 canonical discriminant functions were used in the analysis.

前面说过，投影的重要性是和特征值的贡献率有关。该表说明第一个函数的贡献率已经是99%了，而第二个只有1%。当然，二维图要容易看一些。投影之后，再根据各点的位置远近算出具体的判别公式（SPSS输出）：

Disc.sav例子

- 具体的判别公式（SPSS输出），由一张分类函数表给出：

Classification Function Coefficients

	GROUP		
	1.00	2.00	3.00
IS	.118	.338	.554
SE	.770	21.329	41.616
SA	.345	.542	.811
PRR	.086	.029	-.001
MS	.355	.743	1.203
MSR	.368	.173	.081
CS	7.531	5.220	2.742
(Constant)	-57.521	-53.704	-96.084

Fisher's linear discriminant functions

该表给出了三个线性分类函数的系数。把每个观测点带入三个函数，就可以得到分别代表三类的三个值，哪个值最大，该点就属于相应的那一类。当然，用不着自己去算，计算机软件的选项可以把这些训练数据的每一个点按照这里的分类法分到某一类。当然，我们一开始就知道这些训练数据的各个观测值的归属，但即使是这些训练样本的观测值（企业）按照这里推导出的分类函数来分类，也不一定全都能够正确划分。

Disc.sav例子

- 下面就是对我们的训练样本的分类结果（SPSS）：

Classification Results [b,c]

			Predicted Group Membership			
		GROUP	1.00	2.00	3.00	Total
Original	Count	1.00	30	0	0	30
		2.00	0	30	0	30
		3.00	0	0	30	30
	%	1.00	100.0	.0	.0	100.0
		2.00	.0	100.0	.0	100.0
		3.00	.0	.0	100.0	100.0
Cross-validated[a]	Count	1.00	30	0	0	30
		2.00	0	30	0	30
		3.00	0	0	30	30
	%	1.00	100.0	.0	.0	100.0
		2.00	.0	100.0	.0	100.0
		3.00	.0	.0	100.0	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. 100.0% of original grouped cases correctly classified.

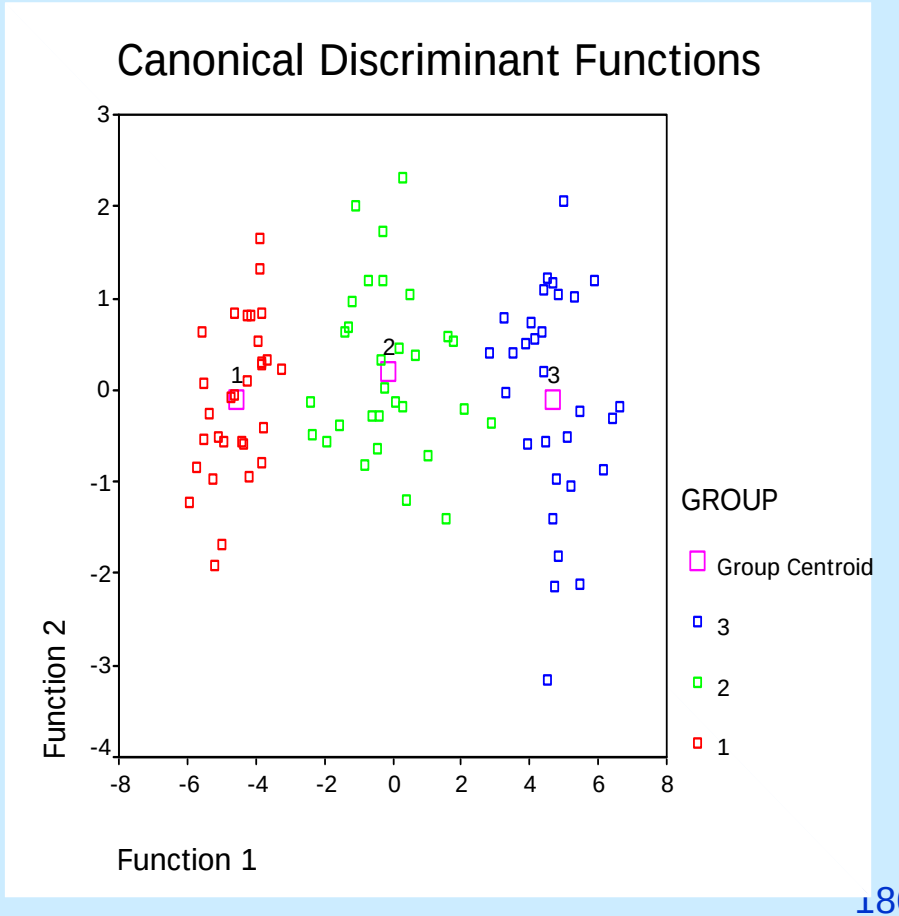
c. 100.0% of cross-validated grouped cases correctly classified.

误判和正确判别率

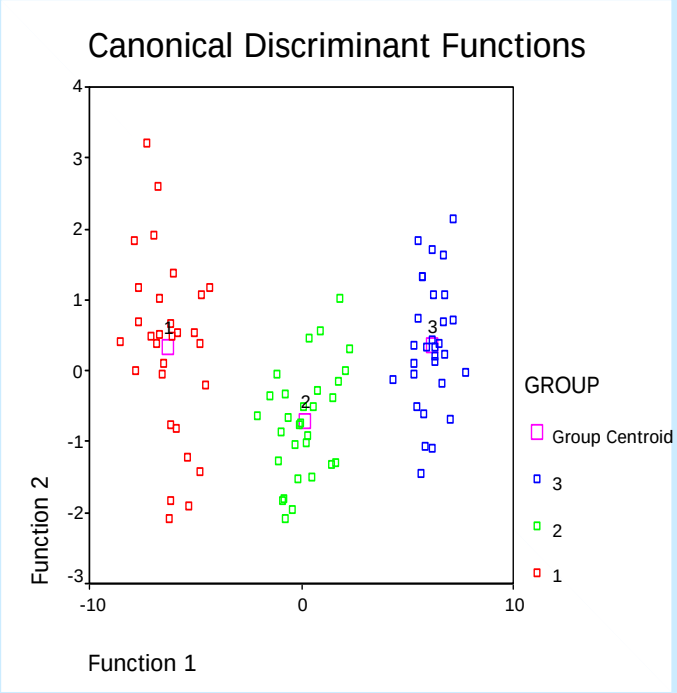
- 从这个表来看，我们的分类能够**100%**地把训练数据的每一个观测值分到其本来的类。
- 该表分成两部分；上面一半（**Original**）是用从全部数据得到的判别函数来判断每一个点的结果（前面三行为判断结果的数目，而后三行为相应的百分比）。
- 下面一半（**Cross validated**）是对每一个观测值，都用缺少该观测的全部数据得到的判别函数来判断的结果。
- 这里的判别结果是**100%**判别正确，但一般并不一定。

Disc.sav例子

- 如果就用这个数据，但不用所有的变量，而只用4个变量进行判别：企业规模（is）、服务(se)、雇员工资比例(sa)、资金周转速度(cs)。结果的图形和判别的正确与否就不一样了。下图为两个典则判别函数导出的150个企业的二维点图。它不如前面的图那么容易分清楚了



原先的图



Disc.sav例子

- 下面是基于4个变量时分类结果表：

Classification Results [b,c]

		GROUP	Predicted Group Membership 1.00	2.00	3.00	Total
Original	Count	1.00	30	0	0	30
		2.00	2	27	1	30
		3.00	0	0	30	30
	%	1.00	100.0	.0	.0	100.0
		2.00	6.7	90.0	3.3	100.0
		3.00	.0	.0	100.0	100.0
Cross-validated[a]	Count	1.00	30	0	0	30
		2.00	2	27	1	30
		3.00	0	0	30	30
	%	1.00	100.0	.0	.0	100.0
		2.00	6.7	90.0	3.3	100.0
		3.00	.0	.0	100.0	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. 96.7% of original grouped cases correctly classified.

c. 96.7% of cross-validated grouped cases correctly classified.

- 这个表的结果是有87个点（96.7%）得到正确划分，有3个点被错误判别；其中第二类有两个被误判为第一类，有一个被误判为第三类。

判别分析要注意什么？

- 训练样本中必须有所有要判别的类型，分类必须清楚，不能有混杂。
- 要选择好可能由于判别的预测变量。这是最重要的一步。当然，在应用中，选择的余地不见得有多大。
- 要注意数据是否有不寻常的点或者模式存在。还要看预测变量中是否有些不适宜的；这可以用单变量方差分析（**ANOVA**）和相关分析来验证。
- 判别分析是为了正确地分类，但同时也要注意使用尽可能少的预测变量来达到这个目的。使用较少的变量意味着节省资源和易于对结果进行解释。

判别分析要注意什么？

- 在计算中需要看关于各个类的有关变量的均值是否显著不同的检验结果（在**SPSS**选项中选择**Wilks' Lambda**、**Rao's V**、**The Squared Mahalanobis Distance**或**The Sum of Unexplained Variations**等检验的计算机输出），以确定是否分类结果是仅仅由于随机因素。
- 此外成员的权数（**SPSS**用**prior probability**，即"先验概率"，和贝叶斯统计的先验概率有区别）需要考虑；一般来说，加权要按照各类观测值的多少，观测值少的就要按照比例多加权。
- 对于多个判别函数，要弄清各自的重要性。
- 注意训练样本的正确和错误分类率。研究被误分类的观测值，看是否可以找出原因。

SPSS选项

- 打开disc.sav数据。然后点击Analyze－Classify－Discriminant，
- 把group放入Grouping Variable，再定义范围，即在Define Range输入1－3的范围。然后在Independents输入所有想用的变量；但如果要用逐步判别，则不选Enter independents together，而选择Use stepwise method，
- 在方法（Method）中选挑选变量的准则（检验方法；默认值为Wilks' Lambda）。
- 为了输出Fisher分类函数的结果可以在Statistics中的Function Coefficient选 Fisher和Unstandardized ，在Matrices中选择输出所需要的相关阵；
- 还可以在Classify中的Display选summary table, Leave-one-out classification；注意在Classify选项中默认的Prior Probability为All groups equal表示所有的类都平等对待，而另一个选项为Compute from group sizes，即按照类的大小加权。
- 在Plots可选 Combined-groups, Territorial map等。

判别分析

(Discriminant Analysis)

和聚类分析的关系

- 判别分析和聚类分析都是分类.
- 但判别分析是在已知对象有若干类型和一批已知样品的观测数据后的基础上根据某些准则建立判别式. 而做聚类分析时类型并不知道.
- 可以先聚类以得知类型,再进行判别.

距离判别法

- 假设有两个总体G_1和G_2, 如果能够定义点x到它们的距离$D(x,G_1)$和$D(x,G_2)$, 则
- 如果$D(x,G_1) < D(x,G_2)$则$x \in G_1$
- 如果$D(x,G_2) < D(x,G_1)$则$x \in G_2$
- 如果$D(x,G_1) = D(x,G_2)$则待判

Mahalanobis距离

- 假设$\mu^{(1)}$, $\mu^{(2)}$, $\Sigma^{(1)}$, $\Sigma^{(2)}$分别为G_1和G_2的均值向量和协差阵,则点x到G_i的马氏距离定义为

  $D^2(x,G_i)=(x-\mu^{(i)})'(\Sigma^{(i)})^{-1}(x-\mu^{(i)})$

- 其他一些距离为马氏距离的特殊情况,因此我们着重讨论马氏距离.马氏距离的好处是可以克服变量之间的相关性干扰,并且消除各变量量纲的影响.

线性判别函数：当$\Sigma^{(1)}=\Sigma^{(2)}=\Sigma$时

$$D^2(x,G_2)-D^2(x,G_1)=2[x-\frac{1}{2}(\mu^{(1)}+\mu^{(2)})]'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)})$$

记

$$\bar{\mu}=\frac{1}{2}(\mu^{(1)}+\mu^{(2)});W(x)=(x-\bar{\mu})'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)})$$

如果W(x)>0即D(x,G_1)<D(x,G_2)则 x∈G_1
如果W(x)<0即D(x,G_1)>D(x,G_2)则 x∈G_2
如果W(x)=0即D(x,G_1)=D(x,G_2)则待判

当$\mu^{(1)}, \mu^{(2)}, \Sigma$ 已知时,

令$a= \Sigma^{-1}(\mu^{(1)}- \mu^{(2)}) \equiv (a_1,\ldots, a_p)'$,则

$$W(x)=(x-\bar{\mu})'a=a'(x-\bar{\mu})=(a_1,...,a_p)\begin{pmatrix} x_1-\bar{\mu}_1 \\ \vdots \\ x_p-\bar{\mu}_p \end{pmatrix}$$

$$=a_1(x_1-\bar{\mu}_1),...,a_p(x_p-\bar{\mu}_p)$$

显然$W(x)$为$x_1,\ldots, x_p$的线性函数, 称为线性判别函数; a称为判别系数.

当$\mu^{(1)}, \mu^{(2)}, \Sigma$ 未知时，
可通过样本来估计：

$$\hat{\mu}^{(i)} = \frac{1}{n_i}\sum_{k=1}^{n_2} x_k^{(i)} = \bar{x}^{(i)}, \hat{\Sigma} = \frac{1}{n_1 + n_2 - 2}(S_1 + S_2),$$

$$S_i = \sum_{t=1}^{n_i}(x_t^{(i)} - \bar{x}^{(i)})(x_t^{(i)} - \bar{x}^{(i)})', \bar{x} = \frac{1}{2}(\bar{x}^{(1)} + \bar{x}^{(2)})$$

$x_1^{(i)}, \ldots, x_{n_i}^{(i)}$ 为来自$\mathbf{G_i}$的样本为$\boldsymbol{(i=1,2)}$

判别函数为

$$W(x) = (x - \bar{x})'\hat{\Sigma}^{-1}(\bar{x}^{(1)} - \bar{x}^{(2)})$$

非线性判别函数:当$\Sigma^{(1)} \neq \Sigma^{(2)}$时

$$D^2(x,G_2)-D^2(x,G_1)$$

$$=(x-\mu^{(2)})'(\Sigma^{(2)})^{-1}(x-\mu^{(2)})-(x-\mu^{(1)})'(\Sigma^{(1)})^{-1}(x-\mu^{(1)})$$

这是x的一个二次函数, 按照距离最近原则, 判别准则仍然为

如果$W(x)>0$即$D(x,G_1)<D(x,G_2)$则 $x\in G_1$

如果$W(x)<0$即$D(x,G_1)>D(x,G_2)$则 $x\in G_2$

如果$W(x)=0$即$D(x,G_1)=D(x,G_2)$则待判

多总体时的线性判别函数:当$\Sigma^{(1)}=\ldots=\Sigma^{(k)}=\Sigma$时

$$D^2(x,G_i)=(x-\mu^{(i)})'(\Sigma^{(i)})^{-1}(x-\mu^{(i)}),\ i=1,\ldots,k$$

记

$$W_{ij}(x)=\frac{1}{2}[D^2(x,G_i)-D^2(x,G_j)]$$

$$=[x-\frac{1}{2}(\mu^{(i)}+\mu^{(j)})]'\Sigma^{-1}(\mu^{(i)}-\mu^{(j)}),\ i,j=1,\ldots,k$$

相应的准则为:

如果对一切$j\neq i$, $W_{ij}(x)>0$, 则 $x\in G_i$

如果有某一个$W_{ij}(x)=0$, 则待判

非线性判别函数:当$\Sigma^{(1)},\ldots,\Sigma^{(k)}$不等时

$$W_{ij}(x)=(x-\mu^{(j)})'(\Sigma^{(j)})^{-1}(x-\mu^{(j)})$$

$$-(x-\mu^{(i)})'(\Sigma^{(i)})^{-1}(x-\mu^{(i)})$$

相应的准则为:

如果对一切$j\neq i$, $W_{ij}(x)>0$, 则 $x\in G_i$

如果有某一个$W_{ij}(x)=0$, 则待判.

当$\mu^{(i)}$, $\Sigma^{(i)}$ 未知时, 可通过样本来估计

$$S_i=\sum_{t=1}^{n_i}(x_t^{(i)}-\bar{x}^{(i)})(x_t^{(i)}-\bar{x}^{(i)})'$$

$$\hat{\mu}^{(i)}=\frac{1}{n_i}\sum_{t=1}^{n_i}x_t^{(i)}=\bar{x}^{(i)},\ \hat{\Sigma}^{(i)}=\frac{1}{n_i-1}S_i,\ i=1,\ldots,k$$

费歇(Fisher)判别法

- 并未要求总体分布类型
- 工作原理就是对原数据系统进行坐标变换,寻求能够将总体尽可能分开的方向.
- 点**x**在以**a**为法方向的投影为 **a′x**
- 各组数据的投影为

$$G_i: a'x_1^{(i)}, \cdots, a'x_{n_i}^{(i)}, i = 1, \ldots, k$$

将G_m组中数据投影的均值记为 $a'\bar{x}^{(m)}$ 有

$$a'\bar{x}^{(m)}=\frac{1}{n_m}\sum_{i=1}^{n_m}a'\bar{x}_i^{(m)},m=1,...,k$$

记k组数据投影的总均值为 $a'\bar{x}$ 有

$$a'\bar{x}=\frac{1}{n}\sum_{m=1}^{k}\sum_{i=1}^{n_m}a'\bar{x}_i^{(m)}$$

组间离差平方和为:

$$SSG = \sum_{m=1}^{k} n_m (a'\bar{x}^{(m)} - a'\bar{x})^2$$

$$= a'[\sum_{m=1}^{k} n_m (\bar{x}^{(m)} - \bar{x})(\bar{x}^{(m)} - \bar{x})']a = a'Ba;$$

这里 $B \equiv \sum_{m=1}^{k} n_m (\bar{x}^{(m)} - \bar{x})(\bar{x}^{(m)} - \bar{x})']$ 组内离差平方和为:

$$SSE = \sum_{m=1}^{k} \sum_{i=1}^{n_m} (a'x_i^{(m)} - a'\bar{x}^{(m)})^2$$

$$= a'[\sum_{m=1}^{k} \sum_{i=1}^{n_m} (x_i^{(m)} - \bar{x}^{(m)})(x_i^{(m)} - \bar{x}^{(m)})']a = a'Ea;$$

这里 $$E = \sum_{m=1}^{k} \sum_{i=1}^{n_m} (x_i^{(m)} - \bar{x}^{(m)})(x_i^{(m)} - \bar{x}^{(m)})'$$

注:$L=|E|/|B+E|$为有Wilks分布的检验零假设$H_0:\mu^{(1)}=\ldots=\mu^{(k)}$的似然比统计量. Wilks分布常用$\chi^2$分布近似(Bartlett)

希望寻找a使得SSG尽可能大而SSE尽可能小，即

$$\Delta(a)=\frac{a'Ba}{a'Ea}\rightarrow \max$$

使 $\frac{a'Ba}{a'Ea}$ 最大的值为方程$|B-\lambda E|=0$的最大特征根λ_1.

记方程$|B-\lambda E|=0$的全部特征根为$\lambda_1 \geqslant \ldots \geqslant \lambda_r>0$, 相应的特征向量为$v_1,\ldots,v_r$. $\Delta(a)$的大小可以估计判别函数$y_i(x)=v_i' x\ (=a' x)$的效果. 记p_i为判别能力(效率), 有

$$p_i=\frac{\lambda_i}{\sum_{h=1}^{r}\lambda_h}$$

m个判别函数的判别能力定义为

$$\sum_{i=1}^{m} p_i = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{h=1}^{r} \lambda_h}$$

据此来确定选择多少判别函数。再看逐步判别法。（即回到前面）

m个判别函数的判别能力定义为

$$\sum_{i=1}^{m} p_i = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{h=1}^{r} \lambda_h}$$

下面以两总体**(k=2)**为例来发现阈值**.** 它们的均值 $\bar{x}^{(1)}, \bar{x}^{(2)}$ 的投影分别为 $v_1'\bar{x}^{(1)}, v_1'\bar{x}^{(2)}$

当总体方差相等时阈值为

$$\bar{\mu} = (v_1'\bar{x}^{(1)} + v_1'\bar{x}^{(2)})/2 = v_1'(\bar{x}^{(1)} + \bar{x}^{(2)})/2$$

总体方差不等时,注意到$v_1'\bar{x}_1^{(1)},\ldots,v_1'\bar{x}_{n_1}^{(1)}$的样本方差为

$$s_1^2=\frac{1}{n_1-1}v_1'\left[\sum_{i=1}^{n_1}(\bar{x}_i^{(1)}-\bar{x}^{(1)})(\bar{x}_i^{(1)}-\bar{x}^{(1)})'\right]v_1=\frac{1}{n_1-1}v_1'A_1v_1$$

类似地,第二组数据投影的样本方差为

$$s_2^2=\frac{1}{n_2-1}v_1'A_2v_1$$

于是阈值

$$\mu^*=\frac{s_2v_1'\bar{x}^{(1)}+s_1v_1'\bar{x}^{(2)}}{s_1+s_2}$$

如 $v_1'\bar{x}^{(2)}<v_1'\bar{x}^{(1)}$

判别规则为

$$y(x)>\bar{\mu}(or\ \mu^*)\Rightarrow x\in G_1$$

$$y(x)<\bar{\mu}(or\ \mu^*)\Rightarrow x\in G_2$$

$$y(x)=\bar{\mu}(or\ \mu^*)\Rightarrow x\ undecided$$

用m个线性判别函数$y_i(x) = v_i'x, i=1,\ldots,m$,时, 先将样本点在$L(v_i,\ldots,v_m)$空间投影再按照p>1情况的距离判别法来制定判别规则. 判别能力为

$$\sum_{i=1}^{m} p_i = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{h=1}^{r} \lambda_h}$$

于秀林书上介绍了对用一个和m个判别函数的加权和不加权方法. 记$y(x)= v'x$, 其在G_i上的样本均值和方差, 以及总均值为

$$\bar{y}^{(i)} = v'\bar{x}^{(i)}, \sigma_i^2 = v's^{(i)}v, \bar{y} = v'\bar{x}$$

m=1时, 不加权法:

$$|y(x)-\bar{y}^{(i)}|=\min_j|y(x)-\bar{y}^{(j)}|\Rightarrow x\in G_i$$

m=1时, 加权法: 按大小排列 $\bar{y}^{(1)},...,\bar{y}^{(k)}\Rightarrow\bar{y}(1)\le\cdots\le\bar{y}(k)$

相应的标准差为 $\sigma(1),...,\sigma(k)$ 令

$$d_{i,i+1}=\frac{\sigma(i+1)\bar{y}(i)+\sigma(i)\bar{y}(i+1)}{\sigma(i+1)+\sigma(i)},i=1,...,k-1$$

$D_{i,i+1}$可为相应两类的分界点

$$d_{i-1,i}\le y(x)\le d_{i,i+1}\Rightarrow x\in G_i$$

m>1时, 不加权法: 记 $\bar{y}_l^{(i)} = c^{(l)\prime}\bar{x}^{(i)}, l = 1,\ldots,m; i = 1,\ldots,k$

对x=(x$_1$,…,)', y$_l$(x)=v$^{(l)}$'x

$$D_i^2 = \sum_{l=1}^{m}[y_l(x) - \bar{y}_l^{(i)}]^2, i = 1,\ldots,k$$

则 $D_\gamma^2 = \min_i D_i^2 \Rightarrow x \in G_\gamma$

m>1时, 加权法: 记

$$D_i^2 = \sum_{l=1}^{m}[y_l(x) - \bar{y}_l^{(i)}]^2\lambda_l, i = 1,\ldots,k$$

则 $D_\gamma^2 = \min_i D_i^2 \Rightarrow x \in G_\gamma$

Bayes判别法

- 不用判别式,而用比较新给样品属于各个总体的条件概率$P(l|x)$, $l=1,\ldots,k$, 的大小(将新样品判归为来自概率最大的总体).
- 先给出对于k个总体的先验概率$q_1,\ldots,q_k$. 如各总体密度为$\{f_k(x)\}$, 则后验概率为(g=1,…k):

  $P(g|x)=q_g f_g(x)/\Sigma_i q_i f_i(x)$
- 当且仅当$P(h|x)=\max_g P(g|x)$, 判x来自第h总体.
- 也可以用使错判的损失最小来判别. 如果c(i|j)为来自j总体的个体被错判到第i总体的损失. 定义平均错判损失(ECM)为

  $ECM=\Sigma_{i=1} q_i[\Sigma_{l\neq i} P(l|i)c(l|i)]$

逐步判别法

- 前面判别用了所有变量.
- 但是各变量所起作用并不一样.
- 要有进有出,引进“最重要的”并剔除不显著的. 根据是假设检验(比如似然比检验).
  - 检验的零假设是各组变量均值相等. **Lambda (Wilks’ Lambda统计量)** 接近**0**表示组均值不同,接近**1**表示组均值没有不同. **Chi-square**是**lambda**的卡方转换**(Bartelett**近似**)**, 用于确定其显著性.

鸢尾花数据(花瓣,花萼的长宽) 5个变量:花瓣长(slen),花瓣宽(swid), 花萼长(plen), 花萼宽(pwid), 分类号(1:Setosa, 2:Versicolor, 3:Virginica)(data14-04)

no	slen	swid	plen	pwid	spno
1	50	33	14	2	1
4	46	36	10	2	1
11	48	31	16	2	1
13	49	36	14	1	1
14	44	32	13	2	1
18	51	38	16	2	1
19	50	30	16	2	1
21	51	38	19	4	1
22	49	30	14	2	1
25	50	36	14	2	1
35	55	35	13	2	1
41	44	30	13	2	1
42	47	32	16	2	1
45	50	32	12	2	1
46	43	30	11	1	1

Statistics→Classify →Discriminant:

Variables: independent (slen,swid,plen,pwid) Grouping(spno) Define range(min-1,max-3)

Classify: prior probability(All group equal) use covariance matrix (Within-groups) Plots (Combined-groups, Separate-groups, Territorial map) Display (Summary table)

Statistics: Descriptive (Means) Function Coefficients (Fisher's, Unstandardized) Matrix (Within-groups correlation, Within-groups covariance, Separate-groups covariance, Total covariance)

Save: (Predicted group membership, Discriminant Scores, Probability of group membership)

鸢尾花数据(数据分析过程简明表)

Analysis Case Processing Summary

Unweighted Cases		N	Percent
Valid		150	100.0
Excluded	Missing or out-of-range group codes	0	.0
	At least one missing discriminating variable	0	.0
	Both missing or out-of-range group codes and at least one missing discriminating variable	0	.0
	Total	0	.0
Total		150	100.0

鸢尾花数据(原始数据的描述)

Group Statistics

分类		Mean	Std. Deviation	Valid N (listwise)	
				Unweighted	Weighted
刚毛鸢尾花	花萼长	50.06	3.525	50	50.000
	花萼宽	34.28	3.791	50	50.000
	花瓣长	14.62	1.737	50	50.000
	花瓣宽	2.46	1.054	50	50.000
变色鸢尾花	花萼长	59.36	5.162	50	50.000
	花萼宽	27.66	3.147	50	50.000
	花瓣长	42.60	4.699	50	50.000
	花瓣宽	13.26	1.978	50	50.000
佛吉尼亚鸢尾花	花萼长	66.38	7.128	50	50.000
	花萼宽	29.82	3.218	50	50.000
	花瓣长	55.60	5.540	50	50.000
	花瓣宽	20.26	2.747	50	50.000
Total	花萼长	58.60	8.633	150	150.000
	花萼宽	30.59	4.363	150	150.000
	花瓣长	37.61	17.682	150	150.000
	花瓣宽	11.99	7.622	150	150.000

鸢尾花数据(合并类内相关阵和协方差阵)

Pooled Within-Groups Matrices [a]

		花萼长	花萼宽	花瓣长	花瓣宽
Covariance	花萼长	29.960	8.767	16.129	4.340
	花萼宽	8.767	11.542	5.033	3.145
	花瓣长	16.129	5.033	18.597	4.287
	花瓣宽	4.340	3.145	4.287	4.188
Correlation	花萼长	1.000	.471	.683	.387
	花萼宽	.471	1.000	.344	.452
	花瓣长	.683	.344	1.000	.486
	花瓣宽	.387	.452	.486	1.000

a. The covariance matrix has 147 degrees of freedom.

鸢尾花数据(总协方差阵)

Covariance Matrices [a]

分类		花萼长	花萼宽	花瓣长	花瓣宽
刚毛鸢尾花	花萼长	12.425	9.922	1.636	1.033
	花萼宽	9.922	14.369	1.170	.930
	花瓣长	1.636	1.170	3.016	.607
	花瓣宽	1.033	.930	.607	1.111
变色鸢尾花	花萼长	26.643	8.288	18.290	5.578
	花萼宽	8.288	9.902	8.127	4.049
	花瓣长	18.290	8.127	22.082	7.310
	花瓣宽	5.578	4.049	7.310	3.911
佛吉尼亚鸢尾花	花萼长	50.812	8.090	28.461	6.409
	花萼宽	8.090	10.355	5.804	4.456
	花瓣长	28.461	5.804	30.694	4.943
	花瓣宽	6.409	4.456	4.943	7.543
Total	花萼长	74.537	-4.683	130.036	53.507
	花萼宽	-4.683	19.036	-33.056	-12.083
	花瓣长	130.036	-33.056	312.670	129.803
	花瓣宽	53.507	-12.083	129.803	58.101

a. The total covariance matrix has 149 degrees of freedom.

鸢尾花数据(特征值表)

Eigenvalue:用于分析的前两个典则判别函数的特征值, 是组间平方和与组内平方和之比值. 最大特征值与组均值最大的向量对应, 第二大特征值对应着次大的组均值向量

典则相关系数**(canonical correlation):**是组间平方和与总平方和之比的平方根.被平方的是由组间差异解释的变异总和的比.

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	30.419[a]	99.0	99.0	.984
2	.293[a]	1.0	100.0	.476

a. First 2 canonical discriminant functions were used in the analysis.

鸢尾花数据(Wilks' Lambda统计量)

检验的零假设是各组变量均值相等. Lambda接近0表示组均值不同,接近1表示组均值没有不同. Chi-square是lambda的卡方转换, 用于确定其显著性.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	.025	538.950	8	.000
2	.774	37.351	3	.000

鸢尾花数据(有关判别函数的输出)

Standardized Canonical Discriminant Function Coefficients

	Function	
	1	2
花萼长	-.346	.039
花萼宽	-.525	.742
花瓣长	.846	-.386
花瓣宽	.613	.555

标准化的典则判别函数系数(使用时必须用标准化的自变量)

$$y_1 = -0.346x_1 - 0.525x_2 + 0.846x_3 + 0.613x_4$$

$$y_2 = 0.039x_1 + 0.742x_2 - 0.386x_3 + 0.555x_4$$

鸢尾花数据(有关判别函数的输出)

Canonical Discriminant Function Coefficients

	Function	
	1	2
花萼长	-.063	.007
花萼宽	-.155	.218
花瓣长	.196	-.089
花瓣宽	.299	.271
(Constant)	-2.526	-6.987

Unstandardized coefficients

典则判别函数系数

$$y_1 = -0.063x_1 - 0.155x_2 + 0.196x_3 + 0.299x_4 - 2.526$$

$$y_2 = 0.007x_1 + 0.218x_2 - 0.089x_3 + 0.271x_4 - 6.948$$

鸢尾花数据(有关判别函数的输出)

这是类均值(重心)处的典则判别函数值

Functions at Group Centroids

分类	Function	
	1	2
刚毛鸢尾花	-7.392	.219
变色鸢尾花	1.763	-.737
佛吉尼亚鸢尾花	5.629	.518

Unstandardized canonical discriminant functions evaluated at group means

这是典则判别函数(前面两个函数)在类均值(重心)处的值

鸢尾花数据(用判别函数对观测量分类结果)

Classification Processing Summary

Processed		150
Excluded	Missing or out-of-range group codes	0
	At least one missing discriminating variable	0
Used in Output		150

先验概率(没有给)

Prior Probabilities for Groups

分类	Prior	Cases Used in Analysis	
		Unweighted	Weighted
刚毛鸢尾花	.333	50	50.000
变色鸢尾花	.333	50	50.000
佛吉尼亚鸢尾花	.333	50	50.000
Total	1.000	150	150.000

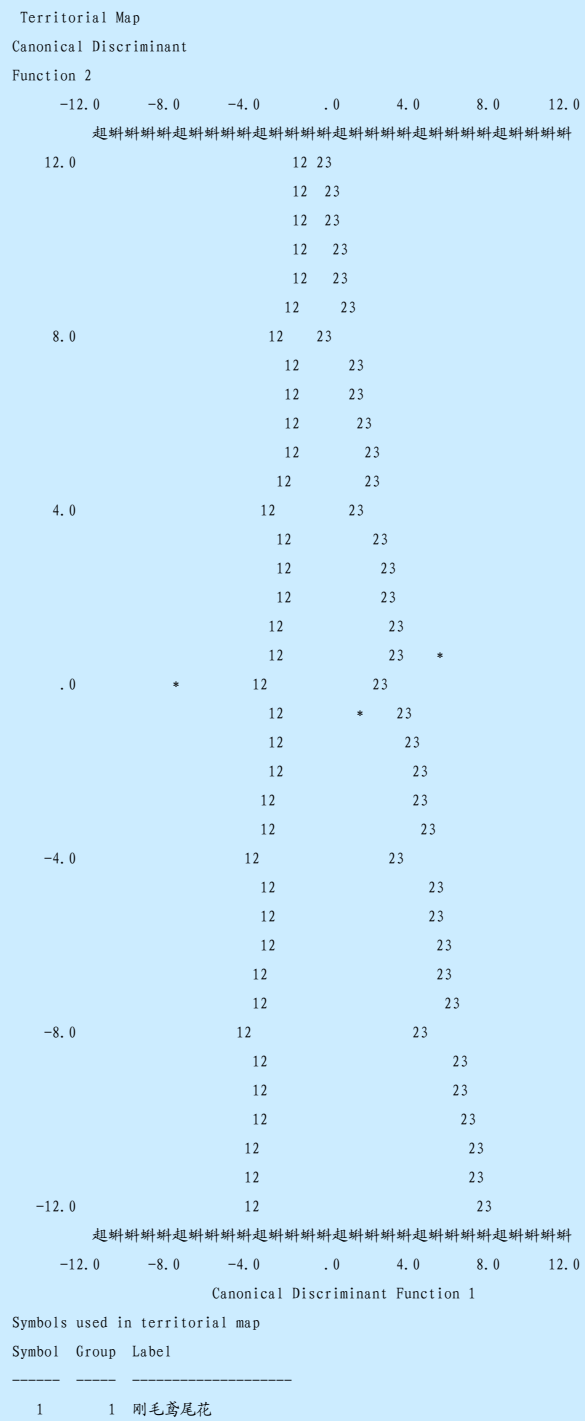
费歇判别函数系数

把自变量代入三个式子,哪个大归谁.

Classification Function Coefficients

	分类		
	刚毛鸢尾花	变色鸢尾花	佛吉尼亚鸢尾花
花萼长	1.687	1.101	.865
花萼宽	2.695	1.070	.747
花瓣长	-.880	1.001	1.647
花瓣宽	-2.284	.197	1.695
(Constant)	-80.268	-71.196	-103.890

Fisher's linear discriminant functions

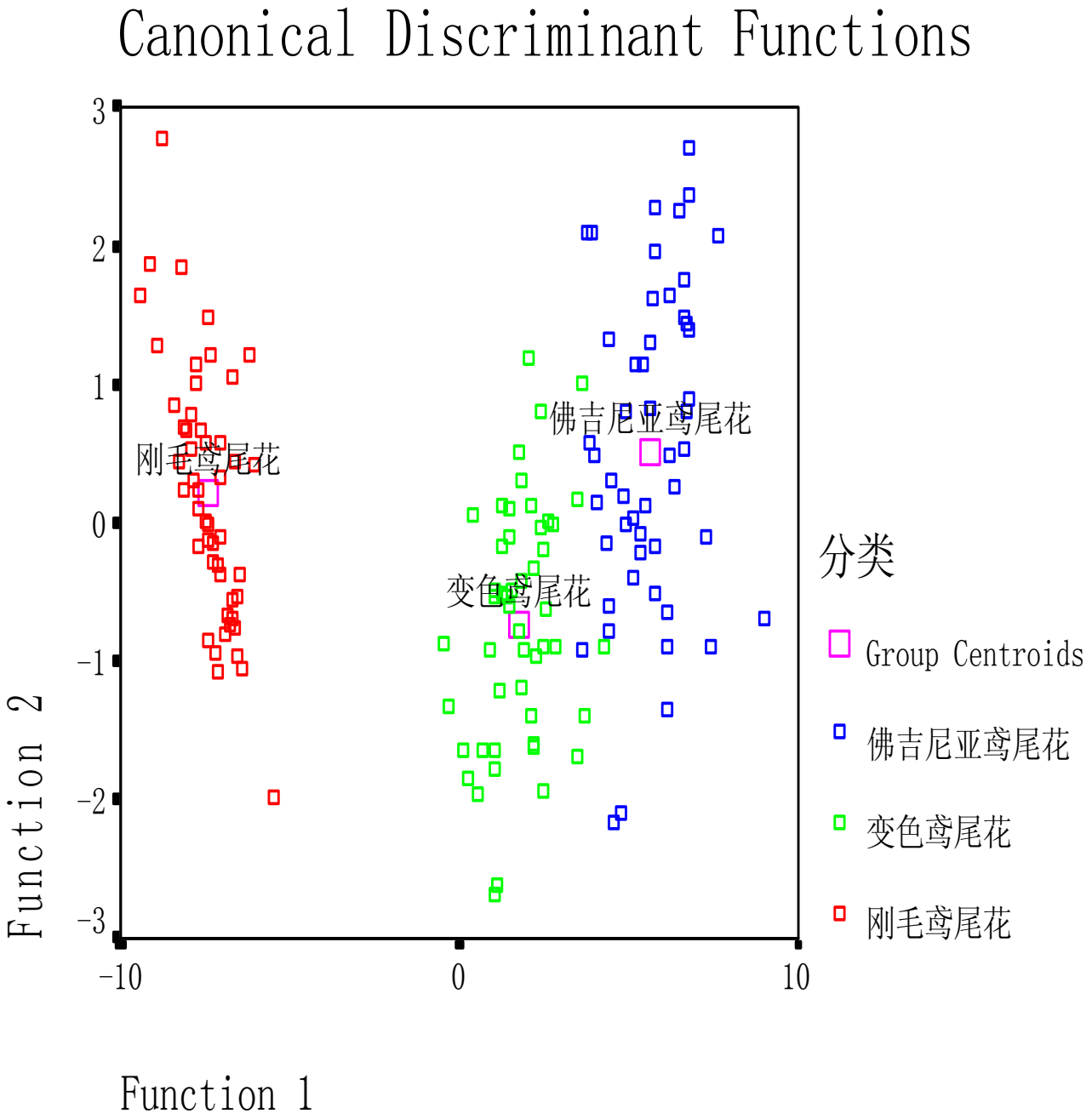


鸢尾花数据

Territory Map(区域图)

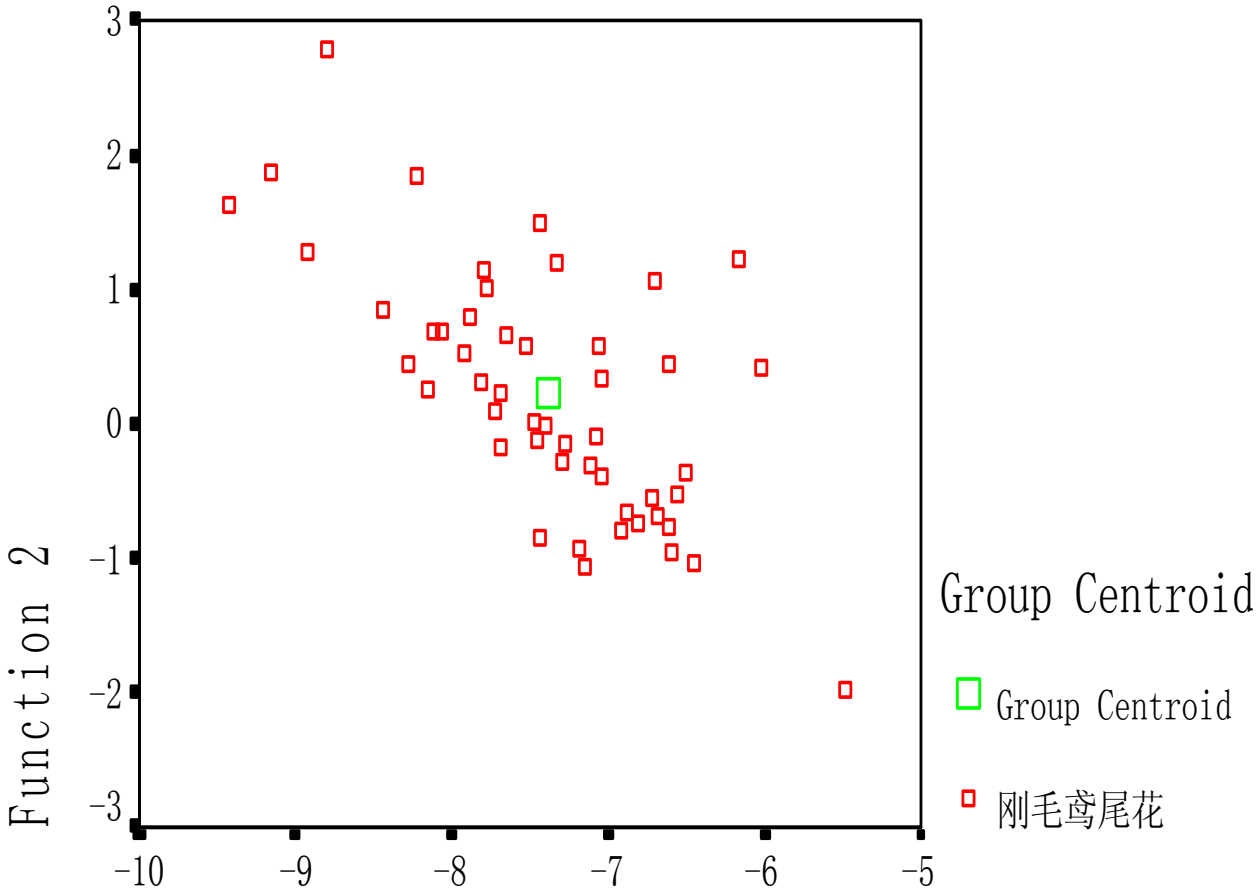
Canonical Discriminate Function 1 Versus Canonical Discriminate Function 2

三种鸢尾花的典则变量值把一个典则变量组成的坐标平面分成三个区域. *为中心坐标.


Canonical Discriminant Functions
3
2
1
0
-1
-2
-3
-10
0
10
Function 2
Function 1
刚毛鸢尾花
佛吉尼亚鸢尾花
变色鸢尾花
分类
Group Centroids
佛吉尼亚鸢尾花
变色鸢尾花
刚毛鸢尾花

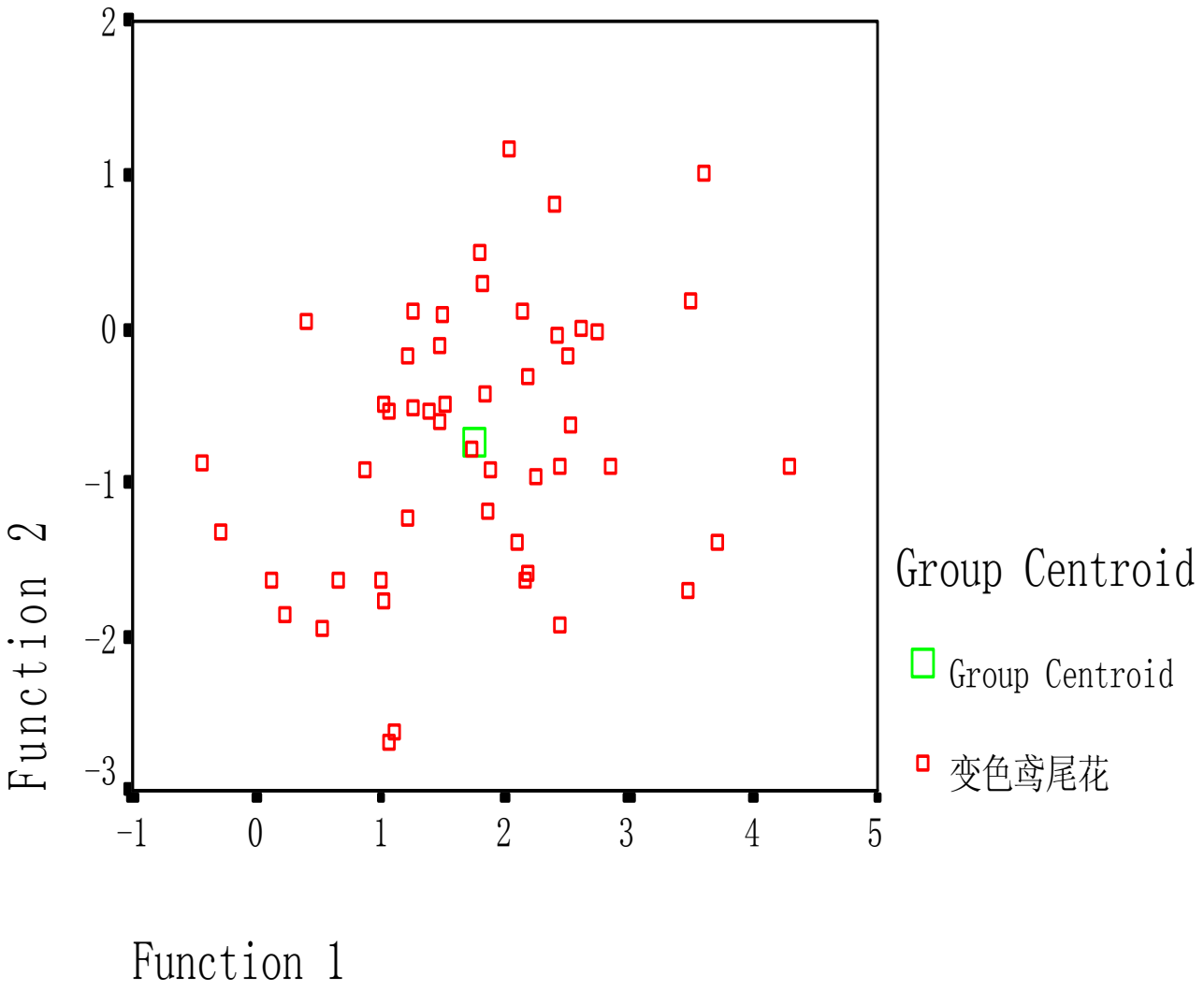
Canonical Discriminant Functions

分类 = 刚毛鸢尾花



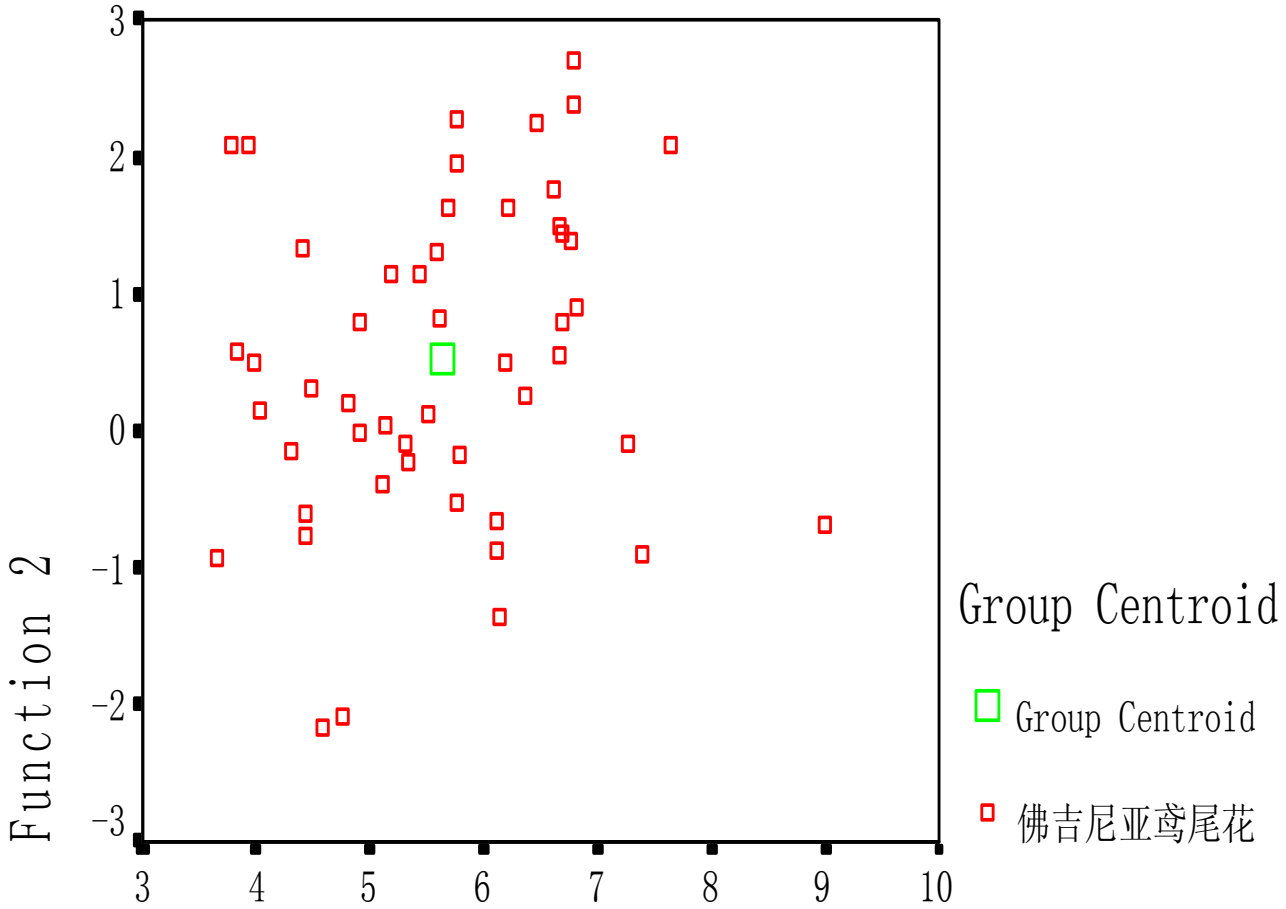
Canonical Discriminant Functions

分类 = 变色鸢尾花



Canonical Discriminant Functions

分类 = 佛吉尼亚鸢尾花



鸢尾花数据(预测分类结果小结)

Classification Results [a]

		分类	Predicted Group Membership			Total
			刚毛鸢尾花	变色鸢尾花	佛吉尼亚鸢尾花	
Original	Count	刚毛鸢尾花	50	0	0	50
		变色鸢尾花	0	48	2	50
		佛吉尼亚鸢尾花	0	1	49	50
	%	刚毛鸢尾花	100.0	.0	.0	100.0
		变色鸢尾花	.0	96.0	4.0	100.0
		佛吉尼亚鸢尾花	.0	2.0	98.0	100.0

a. 98.0% of original grouped cases correctly classified.

可以看出分错率

判别分析结束

第五章 主成分分析

本章主要内容

- 主成分分析的基本思想、主成分分析的数学模型及几何解释、主成分分析的推导及数学解释、计算实例

§1　基本思想

一项十分著名的工作是美国的统计学家斯通(stone)在1947年关于国民经济的研究。他曾利用美国1929—1938年各年的数据，得到了17个反映国民收入与支出的变量要素，例如雇主补贴、消费资料和生产资料、纯公共支出、净增库存、股息、利息外贸平衡等等。

在进行主成分分析后，竟以97.4%的精度，用三新变量就取代了原17个变量。根据经济学知识，斯通给这三个新变量分别命名为总收入F1、总收入变化率F2和经济发展或衰退的趋势F3。更有意思的是，这三个变量其实都是可以直接测量的。斯通将他得到的主成分与实际测量的总收入I、总收入变化率ΔI以及时间t因素做相关分析，得到下表：

	F1	F2	F3	i	i	t
F1	1					
F2	0	1				
F3	0	0	1			
i	0.995	-0.041	0.057	1		
Δi	-0.056	0.948	-0.124	-0.102	1	
t	-0.369	-0.282	-0.836	-0.414	-0.112	1

主成分分析是把各变量之间互相关联的复杂关系进行简化分析的方法。

在社会经济的研究中，为了全面系统的分析和研究问题，必须考虑许多经济指标，这些指标能从不同的侧面反映我们所研究的对象的特征，但在某种程度上存在信息的重叠，具有一定的相关性。

主成分分析试图在力保数据信息丢失最少的原则下，对这种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。

很显然，识辨系统在一个低维空间要比在一个高维空间容易得多。

在力求数据信息丢失最少的原则下，对高维的变量空间降维，即研究指标体系的少数几个线性组合，并且这几个线性组合所构成的综合指标将尽可能多地保留原来指标变异方面的信息。这些综合指标就称为主成分。要讨论的问题是：

（1） 基于相关系数矩阵还是基于协方差矩阵做主成分分析。当分析中所选择的经济变量具有不同的量纲，变量水平差异很大，应该选择基于相关系数矩阵的主成分分析。

（2） 选择几个主成分。主成分分析的目的是简化变量，一般情况下主成分的个数应该小于原始变量的个数。关于保留几个主成分，应该权衡主成分个数和保留的信息。

（3）如何解释主成分所包含的经济意义。

§2　数学模型与几何解释

假设我们所讨论的实际问题中，有p个指标，我们把这p个指标看作p个随机变量，记为X_1，X_2，…，X_p，主成分分析就是要把这p个指标的问题，转变为讨论p个指标的线性组合的问题，而这些新的指标F_1，F_2，…，F_k（$k \leqslant p$），按照保留主要信息量的原则充分反映原指标的信息，并且相互独立。

这种由讨论多个指标降为少数几个综合指标的过程在数学上就叫做降维。主成分分析通常的做法是，寻求原指标的线性组合F_i。

$$
\begin{aligned}
F_1 &= u_{11}X_1 + u_{21}X_2 + \cdots + u_{p1}X_p \\
F_2 &= u_{12}X_1 + u_{22}X_2 + \cdots + u_{p2}X_p \\
&\cdots\cdots \\
F_p &= u_{1p}X_1 + u_{2p}X_2 + \cdots + u_{pp}X_p
\end{aligned}
$$

满足如下的条件：

每个主成分的系数平方和为**1**。即

$$u_{1i}^2 + u_{2i}^2 + \cdots + u_{pi}^2 = 1$$

主成分之间不相关，即无重叠的信息。即

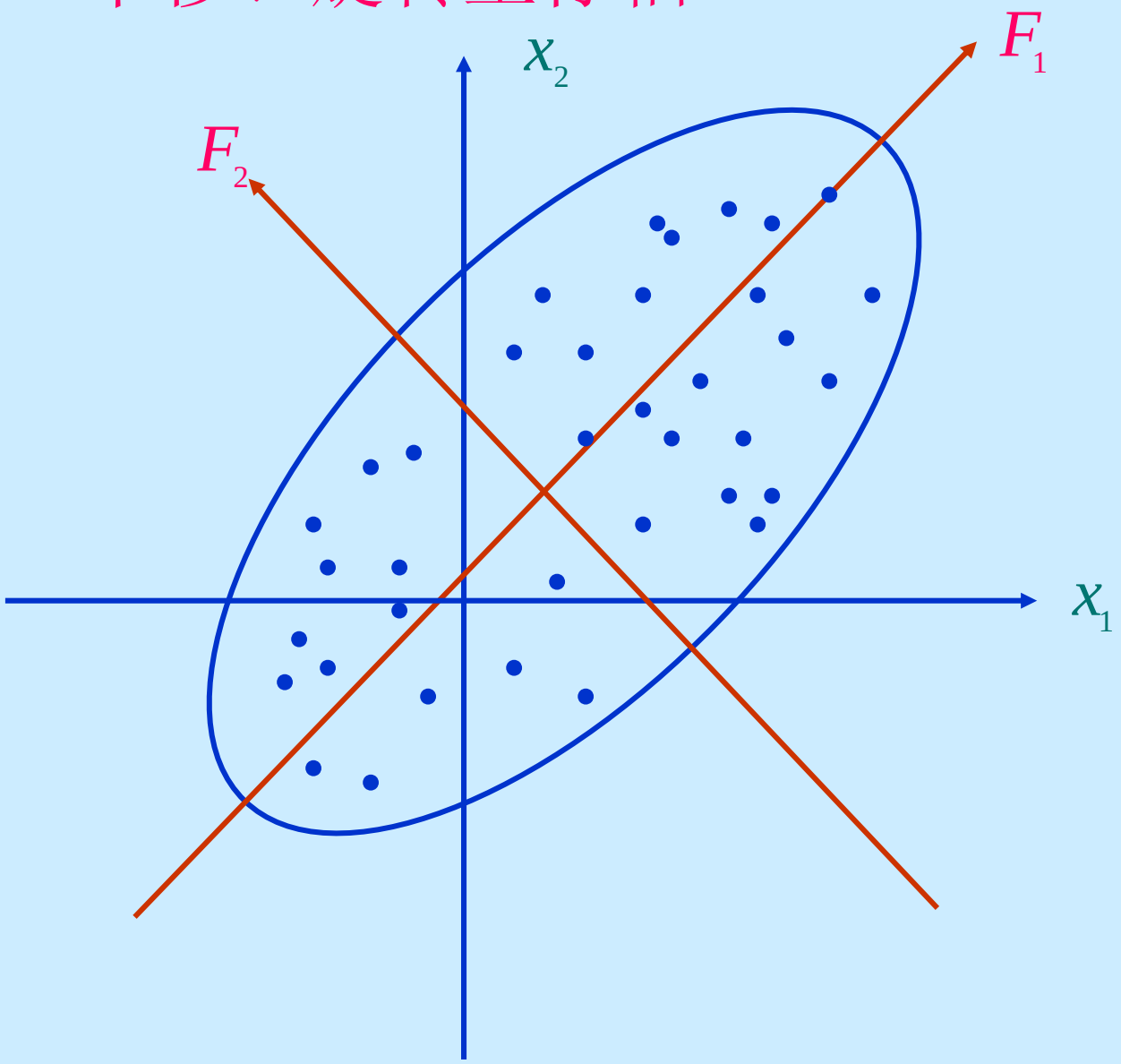
$$Cov（F_i，\ F_j）= 0，\ i \neq j，\ i，\ j = 1，\ 2，\ \cdots，\ p$$

主成分的方差依次递减，重要性依次递减，即

$$Var（F_1）\geq Var(F_2) \geq \cdots \geq Var(F_p)$$

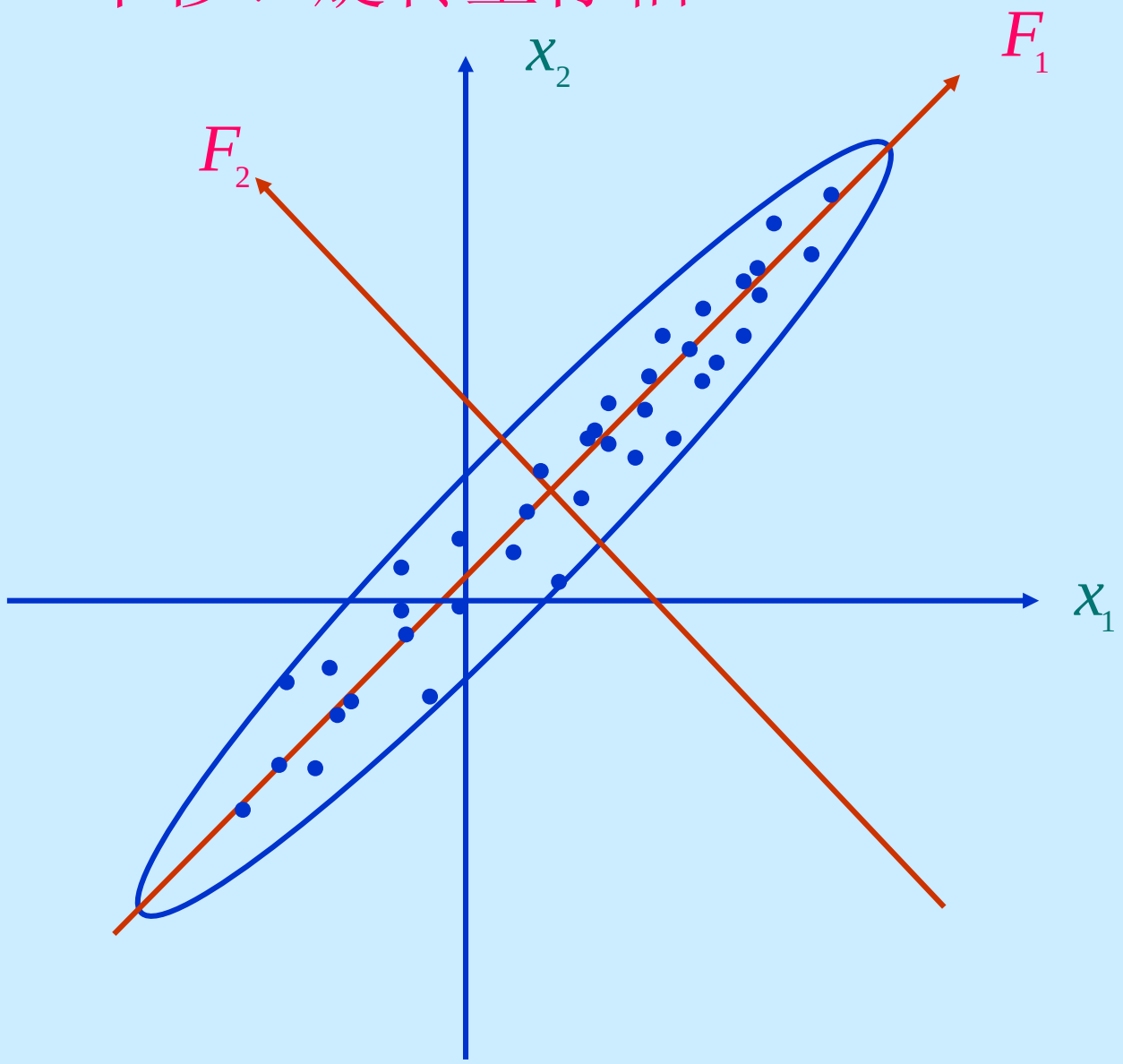
平移、旋转坐标轴

主成分分析的几何解释



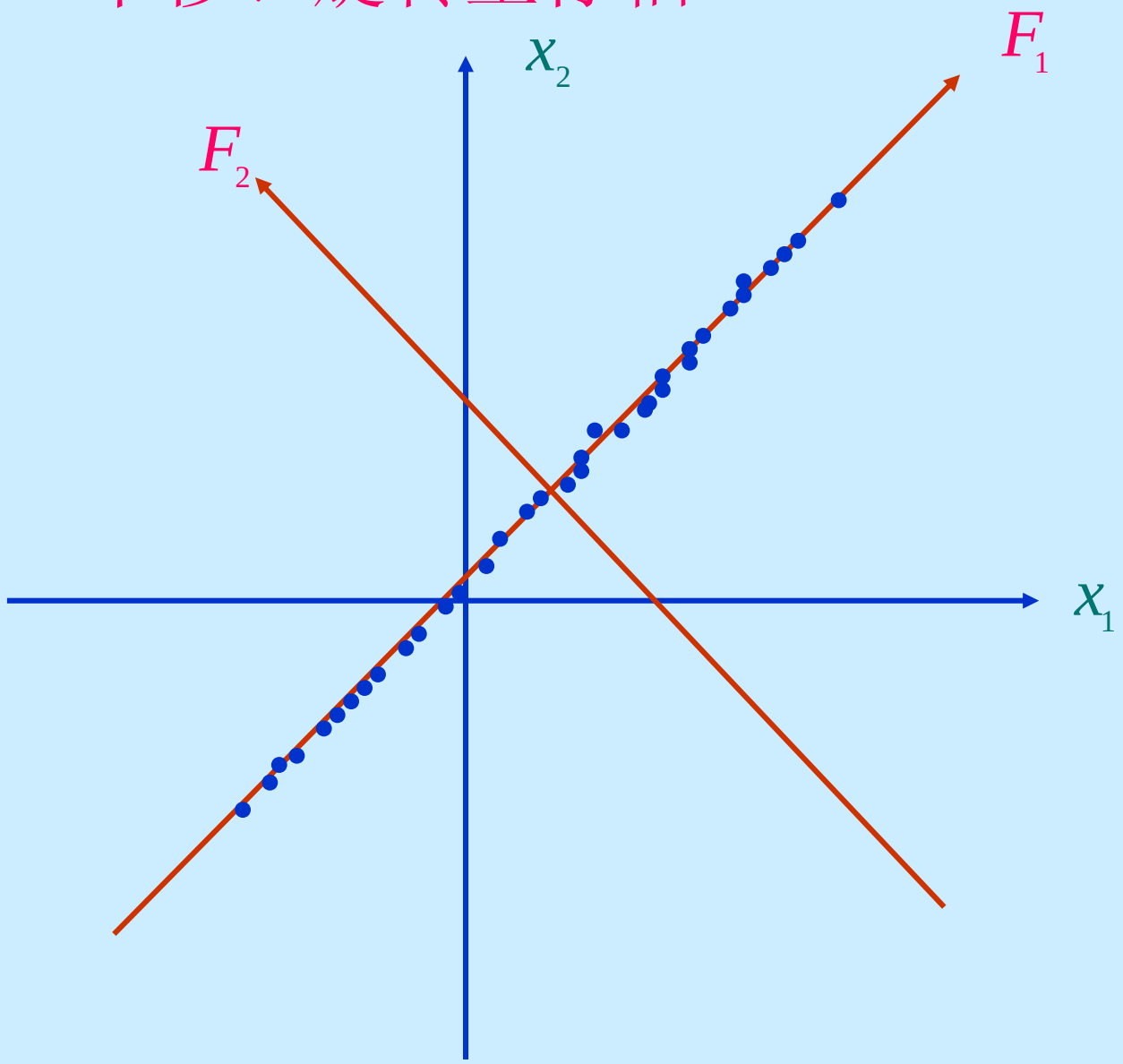
平移、旋转坐标轴

主成分分析的几何解释



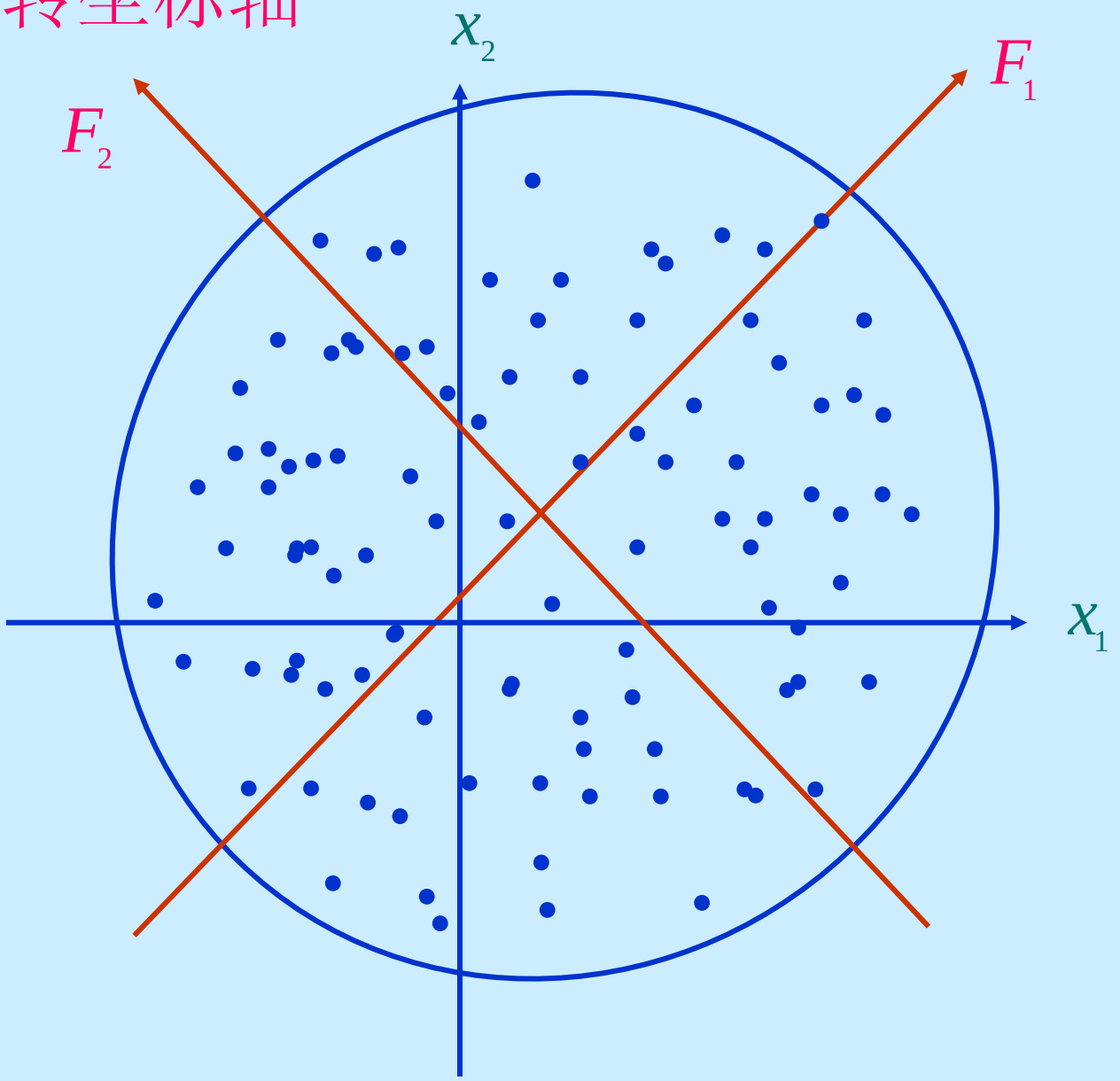
平移、旋转坐标轴

主成分分析的几何解释



平移、旋转坐标轴

主成分分析的几何解释



为了方便，我们在二维空间中讨论主成分的几何意义。设有n个样品，每个样品有两个观测变量x_1和x_2，在由变量x_1和x_2 所确定的二维平面中，n个样本点所散布的情况如椭圆状。由图可以看出这n个样本点无论是沿着x_1 轴方向或x_2轴方向都具有较大的离散性，其离散的程度可以分别用观测变量x_1 的方差和x_2 的方差定量地表示。显然，如果只考虑x_1和x_2 中的任何一个，那么包含在原始数据中的经济信息将会有较大的损失。

- 如果我们将xl 轴和x2轴先平移，再同时按逆时针方向旋转θ角度，得到新坐标轴Fl和F2。Fl和F2是两个新变量。

根据旋转变换的公式：

$$\begin{cases} y_1 = x_1 \cos\theta + x_2 \sin\theta \\ y_2 = -x_1 \sin\theta + x_2 \cos\theta \end{cases}$$

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \mathbf{U}'\mathbf{x}$$

$\mathbf{U}'$为旋转变换矩阵，它是正交矩阵，即有

$$\mathbf{U}' = \mathbf{U}^{-1}, \mathbf{U}'\mathbf{U} = \mathbf{I}$$

旋转变换的目的是为了使得n个样品点在F_1轴方向上的离 散程度最大，即F_1的方差最大。变量F_1代表了原始数据的绝大 部分信息，在研究某经济问题时，即使不考虑变量F_2也无损大局。经过上述旋转变换原始数据的大部分信息集中到F_1轴上，对数据中包含的信息起到了浓缩作用。

F_1，F_2除了可以对包含在X_1，X_2中的信息起着浓缩作用之外，还具有不相关的性质，这就使得在研究复杂的问题时避免了信息重叠所带来的虚假性。二维平面上的个点的方差大部分都归结在F_1轴上，而F_2轴上的方差很小。F_1和F_2称为原始变量x_1和x_2的综合变量。F简化了系统结构，抓住了主要矛盾。

如果图中的椭圆是相当扁平的，则可以只考虑F_1方向上（长轴）的波动，而忽略F2方向的波动。这样，二维降为一维。

一般的，找主成分的问题就是找p维空间中椭球体的主轴问题。

§3　主成分的推导及性质

一、两个线性代数的结论

1、若A是p阶实对称阵，则一定可以找到正交阵U，使

$$\mathbf{U}^{-1}\mathbf{A}\mathbf{U}=\begin{bmatrix}\lambda_1 & 0 & \cdots & 0\\ 0 & \lambda_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \lambda_p\end{bmatrix}_{p\times p}$$

其中 $\lambda_i, i=1.2.\cdots p$ 是A的特征根。

2、若上述矩阵的特征根所对应的单位特征向量为 $\mathbf{u}_1,\cdots,\mathbf{u}_p$

令 $$\mathbf{U}=(\mathbf{u}_1,\cdots,\mathbf{u}_p)=\begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \vdots & \vdots & & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pp} \end{bmatrix}$$

则实对称阵 $\mathbf{A}$ 属于不同特征根所对应的特征向量是正交的，即有$\mathbf{U}'\mathbf{U}=\mathbf{U}\mathbf{U}'=\mathbf{I}$

二、主成分的推导

（一） 第一主成分

设X的协方差阵为 $\boldsymbol{\Sigma}_{\mathbf{x}}=\begin{bmatrix}\sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p}\\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p}\\ \vdots & \vdots & & \vdots\\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2\end{bmatrix}$

由于 Σ_x 为非负定的对称阵，则有利用线性代数的知识可得，必存在正交阵U，使得

$$\mathbf{U}'\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{U}=\begin{bmatrix}\lambda_1 & & 0\\ & \ddots & \\ 0 & & \lambda_p\end{bmatrix}$$

其中λ_1，λ_2，…，λ_p为Σ_x的特征根，不妨假设$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$。而U恰好是由特征根相对应的特征向量所组成的正交阵。

$$\mathbf{U}=(\mathbf{u_1},\cdots,\mathbf{u_p})=\begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \vdots & \vdots & & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pp} \end{bmatrix}$$

$$\mathbf{U_i}=\left(u_{1i},\ u_{2i},\cdots,\ u_{pi}\right)' \qquad i=1,2,\cdots,P$$

下面我们来看，是否由U的第一列元素所构成为原始变量的线性组合是否有最大的方差。

设有P维正交向量 $\mathbf{a}_1 = (a_{11}, a_{21}, \cdots, a_{p1})'$

$$F_1 = a_{11}X_1 + \cdots + a_{p1}X_p = \mathbf{a}_1'\mathbf{X}$$

$$V(F_1) = \mathbf{a}_1'\Sigma\mathbf{a}_1 = \mathbf{a}_1'\mathbf{U}\begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \end{bmatrix}\mathbf{U}'\mathbf{a}_1$$

$$= \mathbf{a}_1'\left[\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_p\right]\begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \end{bmatrix}\begin{bmatrix} \mathbf{u}_1' \\ \mathbf{u}_2' \\ \vdots \\ \mathbf{u}_p' \end{bmatrix}\mathbf{a}_1$$

$$= \sum_{i=1}^{p} \lambda_i \mathbf{a}'\mathbf{u}_i\mathbf{u}_i'\mathbf{a}$$

$$= \sum_{i=1}^{p} \lambda_i (\mathbf{a}'\mathbf{u}_i)^2$$

$$\leq \lambda_1 \sum_{i=1}^{p} (\mathbf{a}'\mathbf{u}_i)^2$$

$$= \lambda_1 \sum_{i=1}^{p} \mathbf{a}'\mathbf{u}_i\mathbf{u}_i'\mathbf{a}$$

$$= \lambda_1 \mathbf{a}'\mathbf{U}\mathbf{U}'\mathbf{a} = \lambda_1 \mathbf{a}'\mathbf{a} = \lambda_1$$

当且仅当$a_1 = u_1$时，即 $F_1 = u_{11}X_1 + \cdots + u_{p1}X_p$ 时，

有最大的方差λ_1。因为$Var(F_1) = U'_1 \Sigma_x U_1 = \lambda_1$。

如果第一主成分的信息不够，则需要寻找第二主成分。

（二） 第二主成分

在约束条件 $\mathrm{cov}(F_1, F_2) = 0$ 下，寻找第二主成分

$$F_2 = u_{12}X_1 + \cdots + u_{p2}X_p$$

因为 $\mathrm{cov}(F_1, F_2) = \mathrm{cov}(u_1'x, u_2'x) = u_2'\Sigma u_1 = \lambda_1 u_2'u_1 = 0$

所以 $u_2'u_1 = 0$

则，对p维向量 u_2，有

$$V(F_2) = u_2'\Sigma u_2 = \sum_{i=1}^{p}\lambda_i \mathbf{u}_2'\mathbf{u}_i\mathbf{u}_i'\mathbf{u}_2 = \sum_{i=1}^{p}\lambda_i(\mathbf{u}_2'\mathbf{u}_i)^2 \le \lambda_2\sum_{i=2}^{p}(\mathbf{u}_2'\mathbf{u}_i)^2$$

$$(\because u_2'u_1 = 0)$$

$$= \lambda_2 \sum_{i=1}^{p} \mathbf{u}_2' \mathbf{u}_i \mathbf{u}_i' \mathbf{u}_2$$

$$= \lambda_2 \mathbf{u}_2' \mathbf{U}\mathbf{U}' \mathbf{u}_2 = \lambda_2 \mathbf{u}_2' \mathbf{u}_2 = \lambda_2$$

所以如果取线性变换：$F_2 = u_{12}X_1 + u_{22}X_2 + \cdots + u_{p2}X_p$

则F_2的方差次大。

类推

$$F_1 = u_{11}X_1 + u_{21}X_2 + \cdots + u_{p1}X_p$$

$$F_2 = u_{12}X_1 + u_{22}X_2 + \cdots + u_{p2}X_p$$

$$\cdots\cdots$$

$$F_p = u_{1p}X_1 + u_{2p}X_2 + \cdots + u_{pp}X_p$$

写为矩阵形式：

$$\mathbf{F} = \mathbf{U}'\mathbf{X}$$

$$\mathbf{U} = (\mathbf{u_1}, \cdots, \mathbf{u_p}) = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \vdots & \vdots & & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pp} \end{bmatrix}$$

$$\mathbf{X} = (X_1, X_2, \cdots, X_p)'$$

§4　主成分的性质

一、均值　$E(\mathbf{U}'x)=\mathbf{U}'\mu$

二、方差为所有特征根之和

$$\sum_{i=1}^{p}Var(F_i)=\lambda_1+\lambda_2+\cdots+\lambda_p=\sigma_1^2+\sigma_2^2+\cdots+\sigma_p^2$$

说明主成分分析把P个随机变量的总方差分解成为P个不相关的随机变量的方差之和。

协方差矩阵Σ的对角线上的元素之和等于特征根之和。

三、精度分析

1）贡献率：第i个主成分的方差在全部方差中所占比重 $\lambda_i \Big/ \sum_{i=1}^{p} \lambda_i$ ，称为贡献率，反映了原来P个指标多大的信息，有多大的综合能力 。

2）累积贡献率：前k个主成分共有多大的综合能力，用这k个主成分的方差和在全部方差中所占比重

$$\sum_{i=1}^{k} \lambda_i \Big/ \sum_{i=1}^{p} \lambda_i$$

来描述，称为累积贡献率。

我们进行主成分分析的目的之一是希望用尽可能少的主成分F_1，F_2，…，F_k（$k \leqslant p$）代替原来的P个指标。到底应该选择多少个主成分，在实际工作中，主成分个数的多少取决于能够反映原来变量80%以上的信息量为依据，即当累积贡献率≥80%时的主成分的个数就足够了。最常见的情况是主成分为2到3个。

四、原始变量与主成分之间的相关系数

$$F_j = u_{1j}x_1 + u_{2j}x_2 + \cdots + u_{pj}x_p \qquad j = 1,2,\cdots,m, m \le p$$

$$\mathbf{F} = \mathbf{U}'\mathbf{X} \qquad \mathbf{UF} = \mathbf{X}$$

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \vdots & \vdots & & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pp} \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_p \end{bmatrix}$$

$$Cov(x_i, F_j) = Cov(u_{i1}F_1 + u_{i2}F_2 + \cdots + u_{ip}F_p, F_j) = u_{ij}\lambda_j$$

$$\rho(x_i, F_j) = \frac{u_{ij}\lambda_j}{\sigma_i\sqrt{\lambda_j}} = \frac{u_{ij}\sqrt{\lambda_j}}{\sigma_i}$$

可见，x_i 和 F_j 的相关的密切程度取决于对应线性组合系数的大小。

主成分 / 原始变量	F_1	F_2	…	F_p
x_1	$\frac{\sqrt{\lambda_1}}{\sqrt{\sigma_1^2}}u_{11}$	$\frac{\sqrt{\lambda_2}}{\sqrt{\sigma_1^2}}u_{12}$	…	$\frac{\sqrt{\lambda_j}}{\sqrt{\sigma_i^2}}u_{ij}$
x_2	$\frac{\sqrt{\lambda_j}}{\sqrt{\sigma_i^2}}u_{ij}$	$\frac{\sqrt{\lambda_j}}{\sqrt{\sigma_i^2}}u_{ij}$	…	$\frac{\sqrt{\lambda_j}}{\sqrt{\sigma_i^2}}u_{ij}$
⋮ ⋮	⋮ ⋮			⋮ ⋮
x_p	$\frac{\sqrt{\lambda_j}}{\sqrt{\sigma_i^2}}u_{ij}$	$\frac{\sqrt{\lambda_j}}{\sqrt{\sigma_i^2}}u_{ij}$	…	$\frac{\sqrt{\lambda_j}}{\sqrt{\sigma_i^2}}u_{ij}$

五、原始变量被主成分的提取率

前面我们讨论了主成分的贡献率和累计贡献率，他度量了F_1，F_2，……，F_m分别从原始变量X_1，X_2，……X_P中提取了多少信息。那么X_1，X_2，……X_P各有多少信息分别F_1，F_2，……，F_m被提取了。应该用什么指标来度量？我们考虑到当讨论F_1分别与X_1，X_2，……X_P的关系时，可以讨论F_1分别与X_1，X_2，……X_P的相关系数，但是由于相关系数有正有负，所以只有考虑相关系数的平方。

$$Var(x_i) = Var(u_{i1}F_1 + u_{i2}F_2 + \cdots + u_{ip}F_p)$$

则 $$u_{i1}^2\lambda_1 + u_{i2}^2\lambda_2 + \cdots + u_{im}^2\lambda_m + \cdots + u_{ip}^2\lambda_p = \sigma_i^2$$

$u_{ij}^2\lambda_j$ 是Fj 能说明的第i 原始变量的方差

$u_{ij}^2\lambda_j / \sigma_i^2$ 是Fj 提取的第i 原始变量信息的比重

如果我们仅仅提出了m个主成分，则第i 原始变量信息的被提取率为：

$$\Omega_i = \sum_{j=1}^{m} \lambda_j u_{ij}^2 / \sigma_i^2 = \sum_{j=1}^{m} \rho_{ij}^2$$

例 设 x_1, x_2, x_3 的协方差矩阵为

$$\Sigma = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

解得特征根为 $\lambda_1 = 5.83,\ \lambda_2 = 2.00,\ \lambda_3 = 0.17$

$$U_1 = \begin{bmatrix} 0.383 \\ -0.924 \\ 0.000 \end{bmatrix} \quad U_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \quad U_3 = \begin{bmatrix} 0.924 \\ 0.383 \\ 0.000 \end{bmatrix}$$

第一个主成分的贡献率为5.83/（5.83+2.00+0.17）=72.875%，尽管第一个主成分的贡献率并不小，但在本题中第一主成分不含第三个原始变量的信息，所以应该取两个主成分。

	X_i与F_1的相关系数	平方	X_i与F_2的相关系数	平方	信息提取率
x_i	$\rho(x_i,F_1)=\rho_{i1}$	ρ_{i1}^2	$\rho(x_i,F_2)=\rho_{i2}$	ρ_{i2}^2	Ω_i
1	0.925	0.855	0	0	0.855
2	-0.998	0.996	0	0	0.996
3	0	0	1	1	1

$$\rho_{11}=\sqrt{\lambda_1}u_{11}\Big/\sqrt{\sigma_1^2}=\sqrt{5.83}*0.383/\sqrt{1}=0.925$$

$$\rho_{12}=\sqrt{\lambda_1}u_{21}\Big/\sqrt{\sigma_2^2}=\sqrt{2}*(-0.924)/\sqrt{5}=-0.998$$

$$\rho_{13}=0$$

定义：如果一个主成分仅仅对某一个原始变量有作用，则称为特殊成分。如果一个主成分所有的原始变量都起作用称为公共成分。

(该题无公共因子）

六、载荷矩阵

$$\begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1m} \\ u_{21} & u_{22} & \cdots & u_{2m} \\ \vdots & \vdots & & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pm} \end{bmatrix}$$

§5　主成分分析的步骤

一、基于协方差矩阵

在实际问题中，X的协方差通常是未知的，于是用X的样本相关阵 $\hat{\Sigma}_x$ 来近似。

$$\mathbf{X}_l = \left(x_{1l},\ x_{2l},\cdots,\ x_{pl}\right)'(l=1,2,\cdots,\ n)$$

$$\hat{\Sigma}_x = \left(\frac{1}{n-1}\sum_{l=1}^{n}(x_{il}-\bar{x}_i)(x_{jl}-\bar{x}_j)\right)_{p\times p}$$

第一步：由X的协方差阵 Σ_x，求出其特征根，即解方程 $|\mathbf{\Sigma}-\lambda\mathbf{I}|=0$ ，可得特征根 $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p\geq 0$ 。

第二步：求出分别所对应的特征向量U_1，U_2，…，U_p，

$$\mathbf{U_i} = \left(u_{1i},\ u_{2i},\cdots,\ u_{pi}\right)'$$

第三步：计算累积贡献率，给出恰当的主成分个数。

$$F_i = \mathbf{U_i'X},\ \ i = 1,2,\cdots,\ \ k(k \le p)$$

第四步：计算所选出的k个主成分的得分。将原始数据的中心化值：

$$\mathbf{X_i^*} = \mathbf{X_i} - \overline{\mathbf{X}} = \left(x_{1i} - \overline{x}_1,\ \ x_{2i} - \overline{x}_2,\cdots,\ \ x_{pi} - \overline{x}_p\right)'$$

代入前k个主成分的表达式，分别计算出各单位k个主成分的得分，并按得分值的大小排队。注意：SPSS中给出的得分系数矩阵的列不是单位化的，需要乘以sqrt（特征值）才对。

二、基于相关系数矩阵

如果变量有不同的量纲，则必须基于相关系数矩阵进行主成分分析。不同的是计算得分时应采用标准化后的数据。

例一　应收账款是指企业因对外销售产品、材料、提供劳务及其它原因，应向购货单位或接受劳务的单位收取的款项，包括应收销货款、其它应收款和应收票据等。出于扩大销售的竞争需要，企业不得不以赊销或其它优惠的方式招揽顾客，由于销售和收款的时间差，于是产生了应收款项。应收款赊销的效果的好坏，不仅依赖于企业的信用政策，还依赖于顾客的信用程度。由此，评价顾客的信用等级，了解顾客的综合信用程度，做到“知己知彼，百战不殆”，对加强企业的应收账款管理大有帮助。某企业为了了解其客户的信用程度，采用西方银行信用评估常用的5C方法，5C的目的是说明顾客违约的可能性。

1、品格（用X_1表示），指顾客的信誉，履行偿还义务的可能性。企业可以通过过去的付款记录得到此项。

2、能力（用X_2表示），指顾客的偿还能力。即其流动资产的数量和质量以及流动负载的比率。顾客的流动资产越多，其转化为现金支付款项的能力越强。同时，还应注意顾客流动资产的质量，看其是否会出现存货过多过时质量下降，影响其变现能力和支付能力。

3、资本（用X_3表示），指顾客的财务势力和财务状况，表明顾客可能偿还债务的背景。

4、附带的担保品（用X_4表示），指借款人以容易出售的资产做抵押。

5、环境条件（用X_5表示），指企业的外部因素，即指非企业本身能控制或操纵的因素。

首先并抽取了10家具有可比性的同类企业作为样本，又请8位专家分别给10个企业的5个指标打分，然后分别计算企业5个指标的平均值，如表。

76.5	81.5	76	75.8	71.7	85	79.2	80.3	84.4	76.5
70.6	73	67.6	68.1	78.5	94	94	87.5	89.5	92
90.7	87.3	91	81.5	80	84.6	66.9	68.8	64.8	66.4
77.5	73.6	70.9	69.8	74.8	57.7	60.4	57.4	60.8	65
85.6	68.5	70	62.2	76.5	70	69.2	71.7	64.9	68.9 ;

Total Variance = 485.31477778

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
PRIN1	410.506	367.242	0.845854	0.84585
PRIN2	43.264	22.594	0.089146	0.93500
PRIN3	20.670	12.599	0.042591	0.97759
PRIN4	8.071	5.266	0.016630	0.99422
PRIN5	2.805	.	0.005779	1.00000

Eigenvectors

	PRIN1	PRIN2	PRIN3	PRIN4	PRIN5
X1	0.468814	-.830612	0.021406	0.254654	-.158081
X2	0.484876	0.329916	0.014801	-.287720	-.757000
X3	0.472744	-.021174	-.412719	-.588582	0.509213
X4	0.461747	0.430904	-.240845	0.706283	0.210403
X5	0.329259	0.122930	0.878054	-.084286	0.313677

第一主成份的贡献率为84.6%，第一主成份

$$Z_1=0.469X_1+0.485X_2+0.473X_3+0.462X_4+0.329X_5$$

的各项系数大致相等，且均为正数，说明第一主成份对所有的信用评价指标都有近似的载荷，是对所有指标的一个综合测度，可以作为综合的信用等级指标。可以用来排序。将原始数据的值中心化后，代入第一主成份Z_1的表示式，计算各企业的得分，并按分值大小排序:

序号	1	2	3	4	5	6	7	8	9	10
得分	3.16	13.6	-9.01	35.9	25.1	-10.3	-4.36	-33.8	-6.41	-13.8
排序	4	3	7	1	2	8	5	10	6	9

在正确评估了顾客的信用等级后，就能正确制定出对其的信用期、收帐政策等，这对于加强应收帐款的管理大有帮助。

例二　基于相关系数矩阵的主成分分析。对美国纽约上市的有关化学产业的三个证券和石油产业的2个证券做了100周的收益率调查。下表是其相关系数矩阵。

1）利用相关系数矩阵做主成分分析。

2）决定要保留的主成分个数，并解释意义。

1	0.577	0.509	0.0063	0.0037
0.577	1	0.599	0.389	0.52
0.509	0.599	1	0.436	0.426
0.387	0.389	0.436	1	0.523
0.462	0.322	0.426	0.523	1

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
PRIN1	2.85671	2.04755	0.571342	0.57134
PRIN2	0.80916	0.26949	0.161833	0.73317
PRIN3	0.53968	0.08818	0.107935	0.84111
PRIN4	0.45150	0.10855	0.090300	0.93141
PRIN5	0.34295	.	0.068590	1.00000

Eigenvectors

	PRIN1	PRIN2	PRIN3	PRIN4	PRIN5
X1	0.463605	-.240339	-.611705	0.386635	-.451262
X2	0.457108	-.509305	0.178189	0.206474	0.676223
X3	0.470176	-.260448	0.335056	-.662445	-.400007
X4	0.421459	0.525665	0.540763	0.472006	-.175599
X5	0.421224	0.581970	-.435176	-.382439	0.385024

§6 主成分分析主要有以下几方面的应用

根据主成分分析的定义及性质，我们已大体上能看出主成分分析的一些应用。概括起来说，主成分分析主要有以下几方面的应用。

1. 主成分分析能降低所研究的数据空间的维数。即用研究m维的Y空间代替p维的X空间(m＜p)，而低维的Y空间代替 高维的x空间所损失的信息很少。即使只有一个主成分Y_1(即 m＝1)时，这个Y_1仍是使用全部X变量(p个)得到的。例如要计算Y_1的均值也得使用全部x的均值。在所选的前m个主成分中，如果某个X_i的系数全部近似于零的话，就可以把这个X_i删除，这也是一种删除多余变量的方法。

2．有时可通过因子负荷a_{ij}的结构，弄清X变量间的某些关系。

3．多维数据的一种图形表示方法。我们知道当维数大于3时便不能画出几何图形，多元统计研究的问题大都多于3个变量。要把研究的问题用图形表示出来是不可能的。然而，经过主成分分析后，我们可以选取前两个主成分或其中某两个主成分，根据主成分的得分，画出n个样品在二维平面上的分布况，由图形可直观地看出各样品在主分量中的地位。

4．由主成分分析法构造回归模型。即把各主成分作为新自变量代替原来自变量x做回归分析。

5．用主成分分析筛选回归变量。回归变量的选择有着重要的实际意义，为了使模型本身易于做结构分析、控制和预报，好从原始变量所构成的子集合中选择最佳变量，构成最佳变量集合。用主成分分析筛选变量，可以用较少的计算量来选择量，获得选择最佳变量子集合的效果。

- 以下是吴喜之教授关于主成分分析和因子分析的课件，大家可以参考其中的SPSS操作。

统计学

—从数据到结论

第十章主成分分析和因子分析

汇报什么？

- 假定你是一个公司的财务经理，掌握了公司的所有数据，这包括众多的变量，如：固定资产、流动资金、借贷的数额和期限、各种税费、工资支出、原料消耗、产值、利润、折旧、职工人数、分工和教育程度等等。
- 如果让你向上级或有关方面介绍公司状况，你能够把这些指标和数字都原封不动地摆出去吗？

需要高度概括

- 在如此多的变量之中，有很多是相关的。人们希望能够找出它们的少数“代表”来对它们进行描述。
- 需要把这种有很多变量的数据进行高度概括。

10.1 主成分分析

- 本章介绍两种把变量维数降低以便于描述、理解和分析的方法：主成分分析（principal component analysis）和因子分析（factor analysis）。
- 实际上主成分分析可以说是因子分析的一个特例。在引进主成分分析之前，先看下面的例子。

成绩数据（student.txt）

- 100个学生的数学、物理、化学、语文、历史、英语的成绩如下表（部分）。

学生代码	数学	物理	化学	语文	历史	英语
1	65	61	72	84	81	79
2	77	77	76	64	70	55
3	67	63	49	65	67	57
4	80	69	75	74	74	63
5	74	70	80	84	81	74
6	78	84	75	62	71	64
7	66	71	67	52	65	57
8	77	71	57	72	86	71
9	83	100	79	41	67	50
…	…	…	…	…	…	…

SPSS数据形式

	math	phys	chem	literat	history	english
1	65	61	72	84	81	79
2	77	77	76	64	70	55
3	67	63	49	65	67	57
4	80	69	75	74	74	63
5	74	70	80	84	81	74
6	78	84	75	62	71	64
7	66	71	67	52	65	57
8	77	71	57	72	86	71
9	83	100	79	41	67	50
10	86	94	97	51	63	55
11	74	80	88	64	73	66
12	67	84	53	58	66	56
13	81	62	69	56	66	52
14	71	64	94	52	61	52
15	78	96	81	80	89	76
16	69	56	67	75	94	80

从本例可能提出的问题

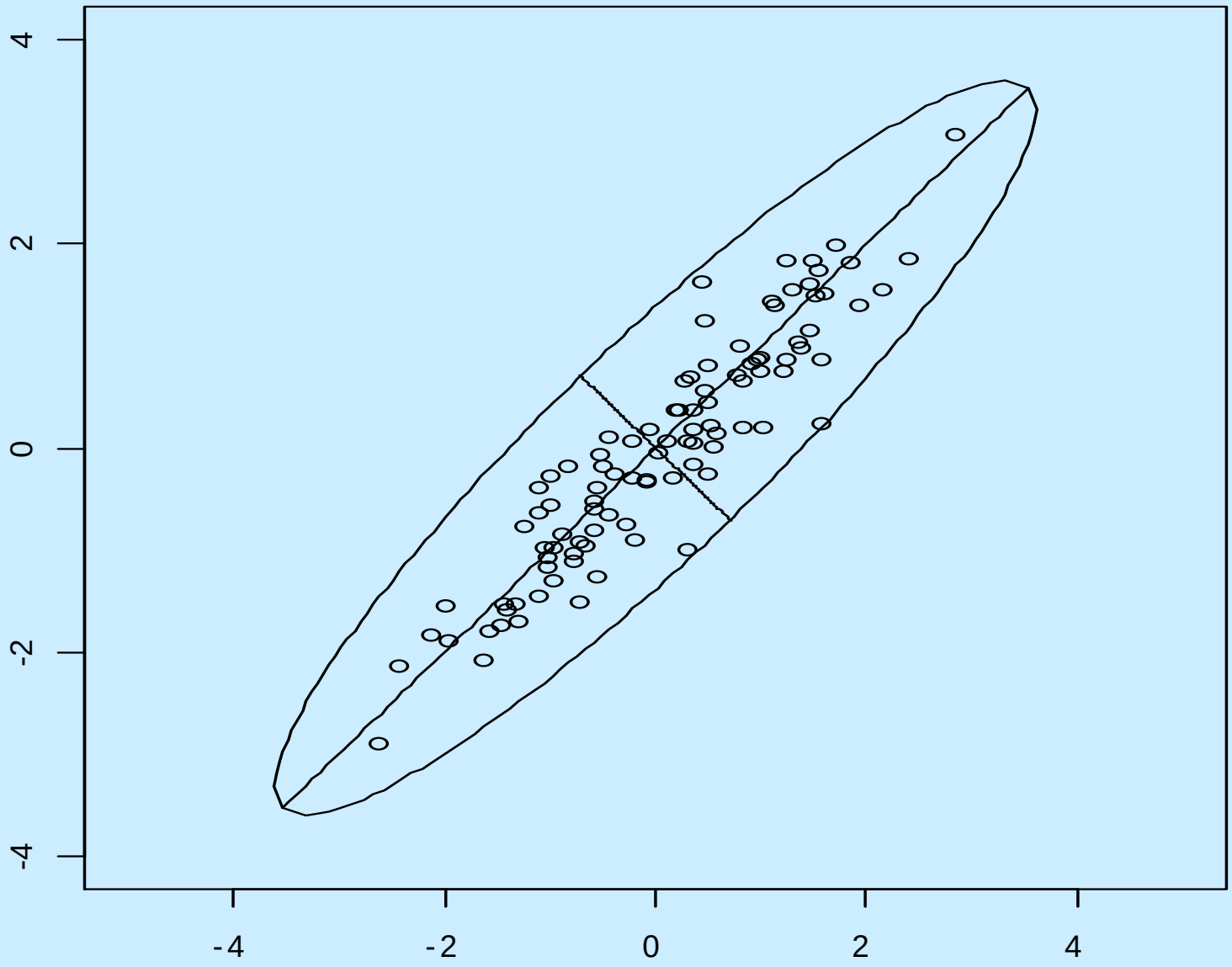
- 目前的问题是，能否把这个数据的6个变量用一两个综合变量来表示呢？
- 这一两个综合变量包含有多少原来的信息呢？
- 能否利用找到的综合变量来对学生排序或据此进行其他分析

空间的点

- 例中数据点是六维的；即每个观测值是6维空间中的一个点。希望把6维空间用低维空间表示。
- 先假定只有二维，即只有两个变量，由横坐标和纵坐标所代表；
- 每个观测值都有相应于这两个坐标轴的两个坐标值；

空间的点

- 如果这些数据形成一个椭圆形状的点阵（这在二维正态的假定下是可能的）该椭圆有一个长轴和一个短轴。在短轴方向上数据变化很少；
- 在极端的情况，短轴如退化成一点，长轴的方向可以完全解释这些点的变化，由二维到一维的降维就自然完成了。



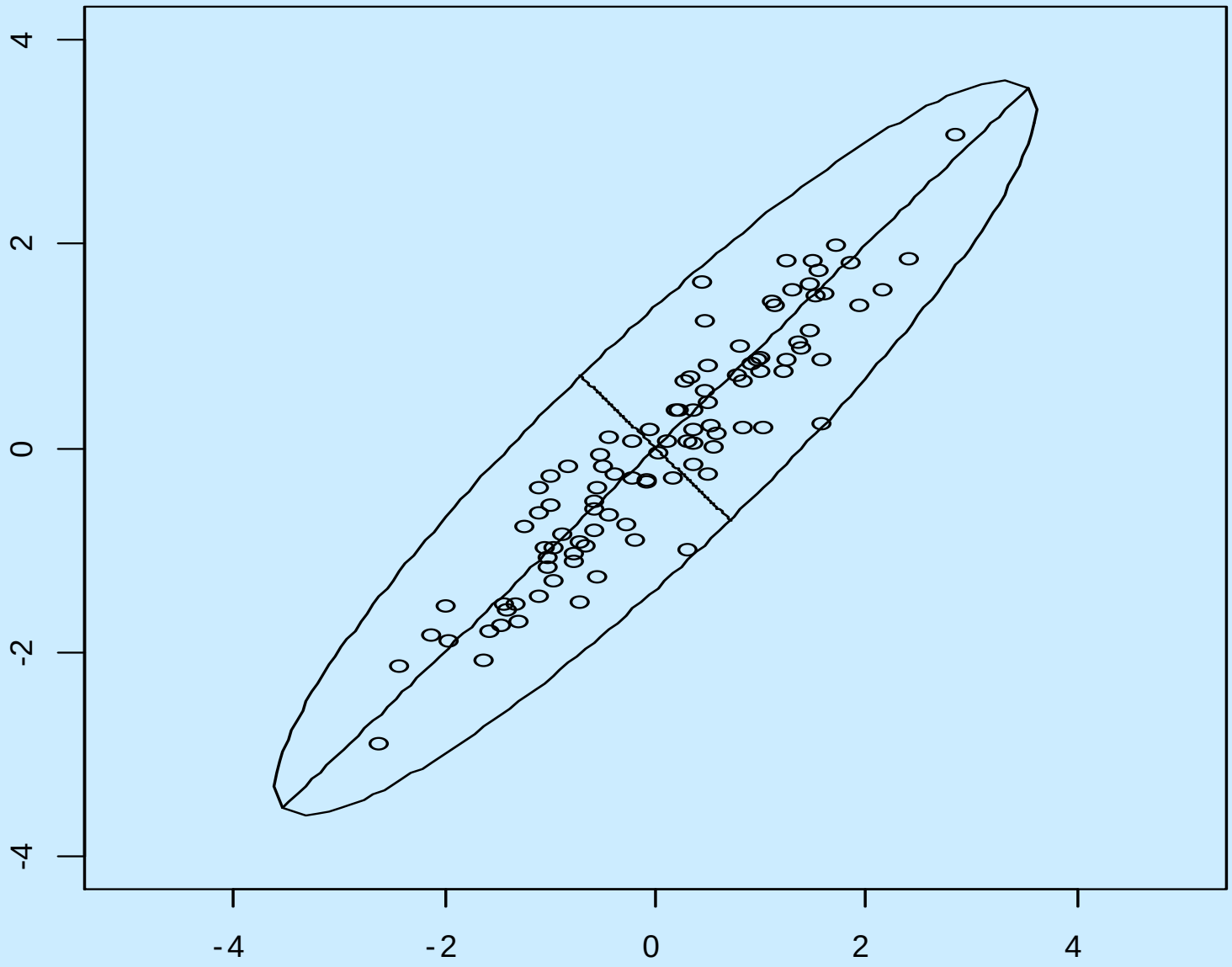
300

椭圆的长短轴

- 当坐标轴和椭圆的长短轴平行，那么代表长轴的变量就描述了数据的主要变化，而代表短轴的变量就描述了数据的次要变化。
- 但是，坐标轴通常并不和椭圆的长短轴平行。因此，需要寻找椭圆的长短轴，并进行变换，使得新变量和椭圆的长短轴平行。

椭圆的长短轴

- 如果长轴变量代表了数据包含的大部分信息，就用该变量代替原先的两个变量（舍去次要的一维），降维就完成了。
- 椭圆的长短轴相差得越大，降维也越有道理。


4
2
0
-2
-4
-4
-2
0
2
4

主轴和主成分

- 多维变量的情况和二维类似，也有高维的椭球，只不过不那么直观罢了。
- 首先把高维椭球的主轴找出来，再用代表大多数数据信息的最长的几个轴作为新变量；这样，主成分分析就基本完成了。

主轴和主成分

- 正如二维椭圆有两个主轴，三维椭球有三个主轴一样，有几个变量，就有几个主轴。
- 和二维情况类似，高维椭球的主轴也是互相垂直的。
- 这些互相正交的新变量是原先变量的线性组合，叫做主成分**(principal component)**。

主成分之选取

- 选择越少的主成分，降维就越好。什么是标准呢？
- 那就是这些被选的主成分所代表的主轴的长度之和占了主轴长度总和的大部分。
- 有些文献建议，所选的主轴总长度占所有主轴长度之和的大约**85%**即可，其实，这只是一个大体的说法；具体选几个，要看实际情况而定。

主成分分析的数学

- 要寻找方差最大的方向。即，使*向量**X*的线性组合***a'X***的方差最大的方向**a.**
- 而**Var(*a*'X)=*a*'Cov(X)*a*;**由于**Cov(X)**未知；于是用**X**的样本相关阵**R**来近似. 要寻找向量**a**使得***a'Ra***最大**(**注意相关阵和协方差阵差一个常数）
- 这涉及相关阵和特征值。回顾一下吧!
- 选择几个主成分呢?要看“贡献率.”

- 对于我们的数据，SPSS输出为

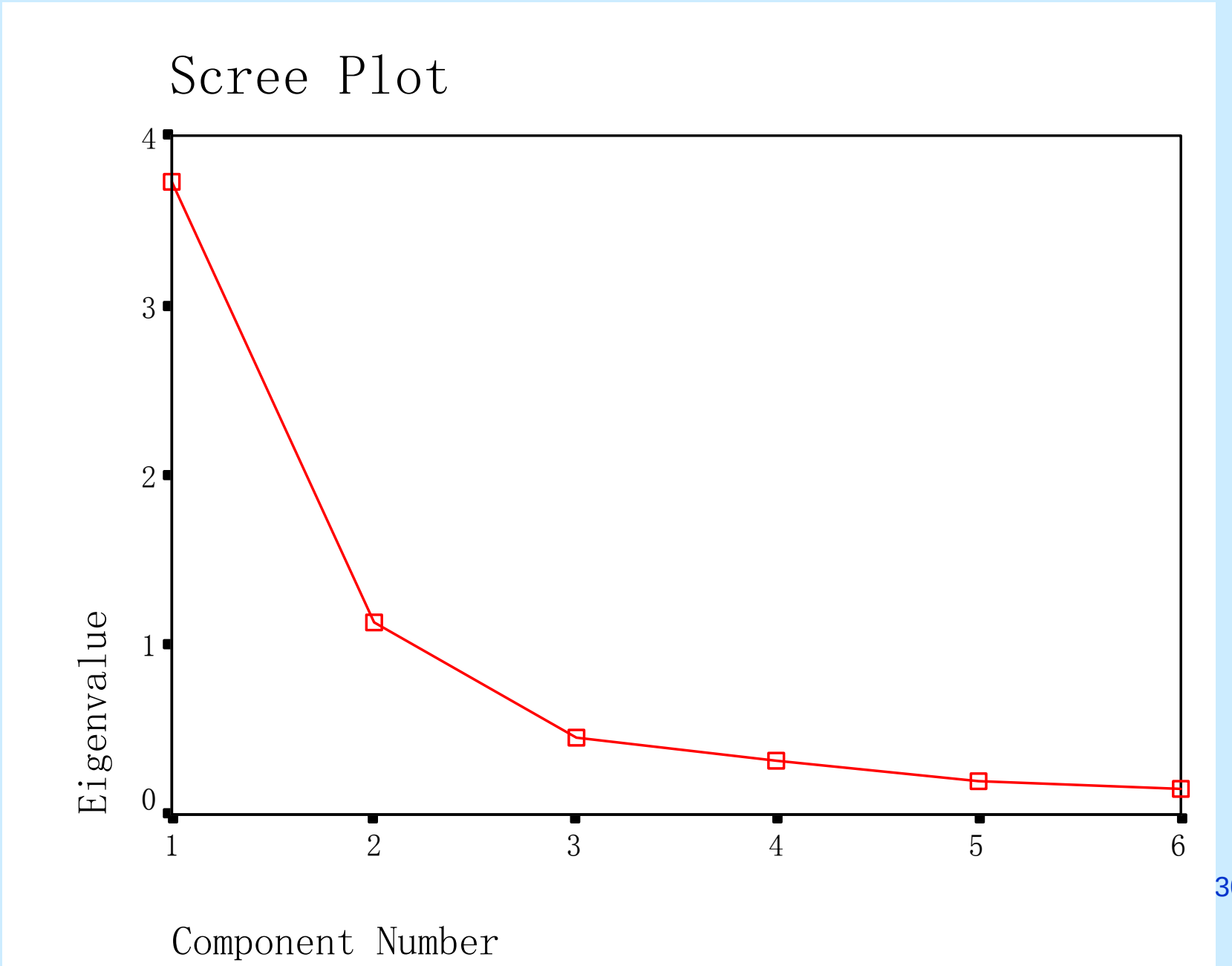
Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.735	62.254	62.254	3.735	62.254	62.254
2	1.133	18.887	81.142	1.133	18.887	81.142
3	.457	7.619	88.761			
4	.323	5.376	94.137			
5	.199	3.320	97.457			
6	.153	2.543	100.000			

Extraction Method: Principal Component Analysis.

- 这里的**Initial Eigenvalues**就是这里的六个主轴长度，又称特征值（数据相关阵的特征值）。头两个成分特征值累积占了总方差的**81.142%**。后面的特征值的贡献越来越少。

- 特征值的贡献还可以从**SPSS**的所谓碎石图看出



- 怎么解释这两个主成分。主成分是原始六个变量的线性组合。这由下表给出。

Component Matrix [a]

	Component					
	1	2	3	4	5	6
MATH	-.806	.353	-.040	.468	.021	.068
PHYS	-.674	.531	-.454	-.240	-.001	-.006
CHEM	-.675	.513	.499	-.181	.002	.003
LITERAT	.893	.306	-.004	-.037	.077	.320
HISTORY	.825	.435	.002	.079	-.342	-.083
ENGLISH	.836	.425	.000	.074	.276	-.197

Extraction Method: Principal Component Analysis.

a. 6 components extracted.

- 这里每一列代表一个主成分作为原来变量线性组合的系数（比例）。比如第一主成分为数学、物理、化学、语文、历史、英语这六个变量的线性组合，系数（比例）为**-0.806, -0.674, -0.675, 0.893, 0.825, 0.836**。

- 如用$x_1, x_2, x_3, x_4, x_5, x_6$分别表示原先的六个变量，而用$y_1, y_2, y_3, y_4, y_5, y_6$表示新的主成分，那么，第一和第二主成分为

$$y_1 = -0.806x_1 - 0.674x_2 - 0.675x_3 + 0.893x_4 + 0.825x_5 + 0.836x_6$$

$$y_2 = 0.353x_1 + 0.531x_2 + 0.513x_3 + 0.306x_4 + 0.435x_5 + 0.425x_6$$

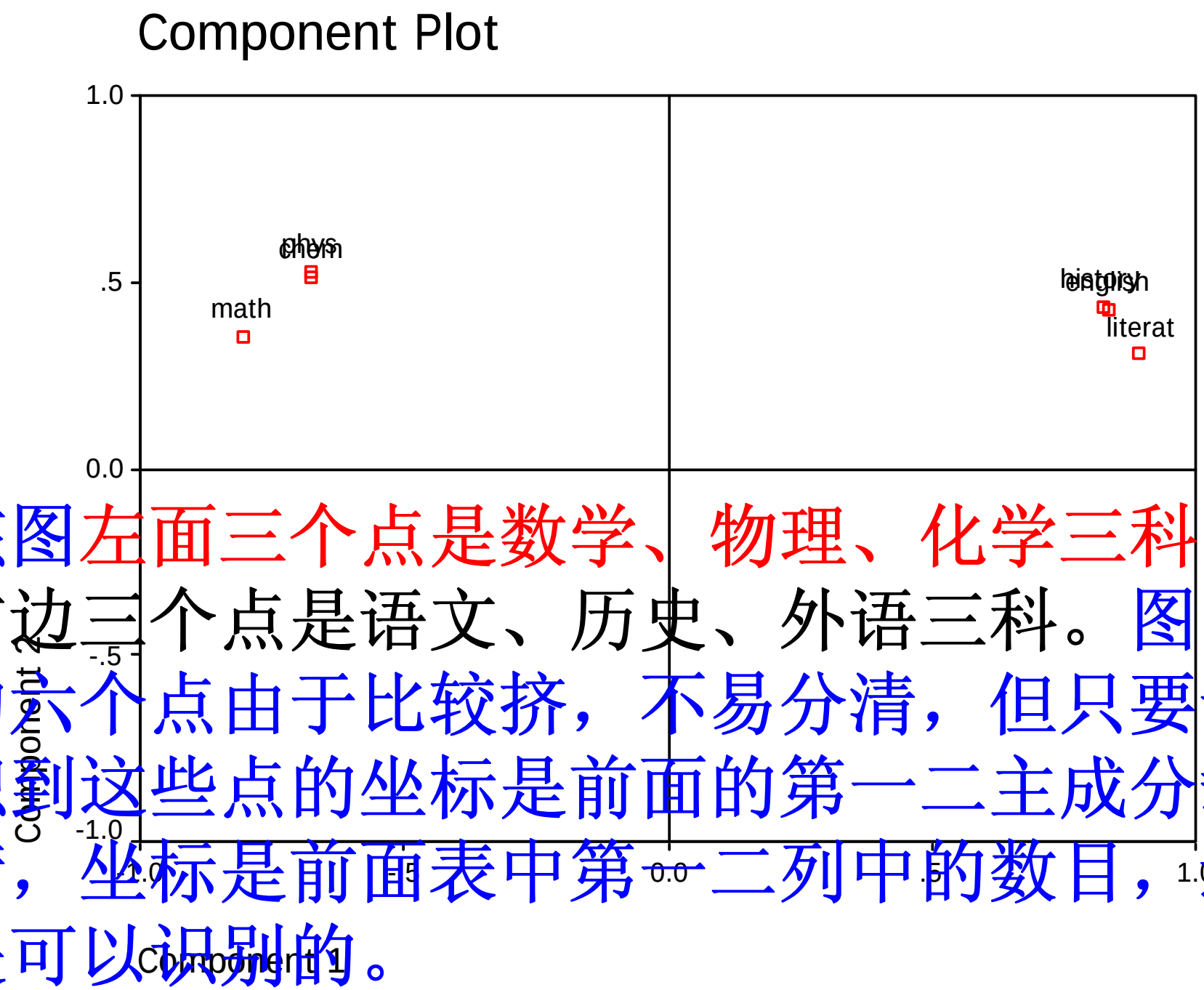
- 这些系数称为主成分载荷（**loading**），它表示主成分和相应的原先变量的相关系数。

$$y_1 = -0.806x_1 - 0.674x_2 - 0.675x_3 + 0.893x_4 + 0.825x_5 + 0.836x_6$$

$$y_2 = 0.353x_1 + 0.531x_2 + 0.513x_3 + 0.306x_4 + 0.435x_5 + 0.425x_6$$

- 比如y_1表示式中x_1的系数为**-0.806**，这就是说第一主成分和数学变量的相关系数为**-0.806**。
- 相关系数(绝对值）越大，主成分对该变量的代表性也越大。可以看得出，第一主成分对各个变量解释得都很充分。而最后的几个主成分和原先的变量就不那么相关了。

- 可以把第一和第二主成分的载荷点出一个二维图以直观地显示它们如何解释原来的变量的。这个图叫做载荷图。



该图左面三个点是数学、物理、化学三科，右边三个点是语文、历史、外语三科。图中的六个点由于比较挤，不易分清，但只要认识到这些点的坐标是前面的第一二主成分载荷，坐标是前面表中第一二列中的数目，还是可以识别的。

10.2 因子分析

- 主成分分析从原理上是寻找椭球的所有主轴。原先有几个变量，就有几个主成分。
- 而因子分析是事先确定要找几个成分，这里叫因子（**factor**）（比如两个），那就找两个。
- 这使得在数学模型上，因子分析和主成分分析有不少区别。而且因子分析的计算也复杂得多。根据因子分析模型的特点，它还多一道工序：因子旋转（**factor rotation**）；这个步骤可以使结果更好。

10.2 因子分析

- 对于计算机，因子分析并不费事。
- 从输出的结果来看，因子分析也有因子载荷（**factor loading**）的概念，代表了因子和原先变量的相关系数。但是在因子分析公式中的因子载荷位置和主成分分析不同。
- 因子分析也给出了二维图；其解释和主成分分析的载荷图类似。

- 主成分分析与因子分析的公式上的区别

$$y_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p$$

$$y_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p$$

……

$$y_p = a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pp}x_p$$

主成分分析

因子分析(m<p)

$$x_1 - \mu = a_{11}f_1 + a_{12}f_2 + \cdots + a_{1m}f_m + \varepsilon_1$$

$$x_2 - \mu = a_{21}f_1 + a_{22}f_2 + \cdots + a_{2m}f_m + \varepsilon_2$$

……

$$x_p - \mu = a_{p1}f_1 + a_{p2}f_2 + \cdots + a_{pm}f_m + \varepsilon_p$$

$$f_1 = \beta_{11}x_1 + \beta_{12}x_2 + \cdots + \beta_{1p}x_p$$

$$f_2 = \beta_{21}x_1 + \beta_{22}x_2 + \cdots + \beta_{2p}x_p$$

……

$$f_m = \beta_{m1}x_1 + \beta_{m2}x_2 + \cdots + \beta_{mp}x_p$$

因子得分

- 对于我们的数据，SPSS因子分析输出为

Rotated Component Matrix [a]

	Component	
	1	2
MATH	-.387	.790
PHYS	-.172	.841
CHEM	-.184	.827
LITERAT	.879	-.343
HISTORY	.911	-.201
ENGLISH	.913	-.216

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

- 这个表说明六个变量和因子的关系。为简单记，我们用$x_1, x_2, x_3, x_4, x_5, x_6$来表示**math**（数学），**phys**（物理），**chem**（化学），**literat**（语文），**history**（历史），**english**（英语）等变量。这样因子f_1和f_2与这些原变量之间的关系是（注意，和主成分分析不同，这里把成分（因子）写在方程的右边，把原变量写在左边；但相应的系数还是主成分和各个变量的线性相关系数，也称为因子载荷）：

$$x_1 = -0.387\, f_1 + 0.790\, f_2;$$

$$x_2 = -0.172\, f_1 + 0.841\, f_2;$$

$$x_3 = -0.184\, f_1 + 0.827\, f_2$$

$$x_4 = 0.879\, f_1 - 0.343\, f_2;$$

$$x_5 = 0.911\, f_1 - 0.201\, f_2;$$

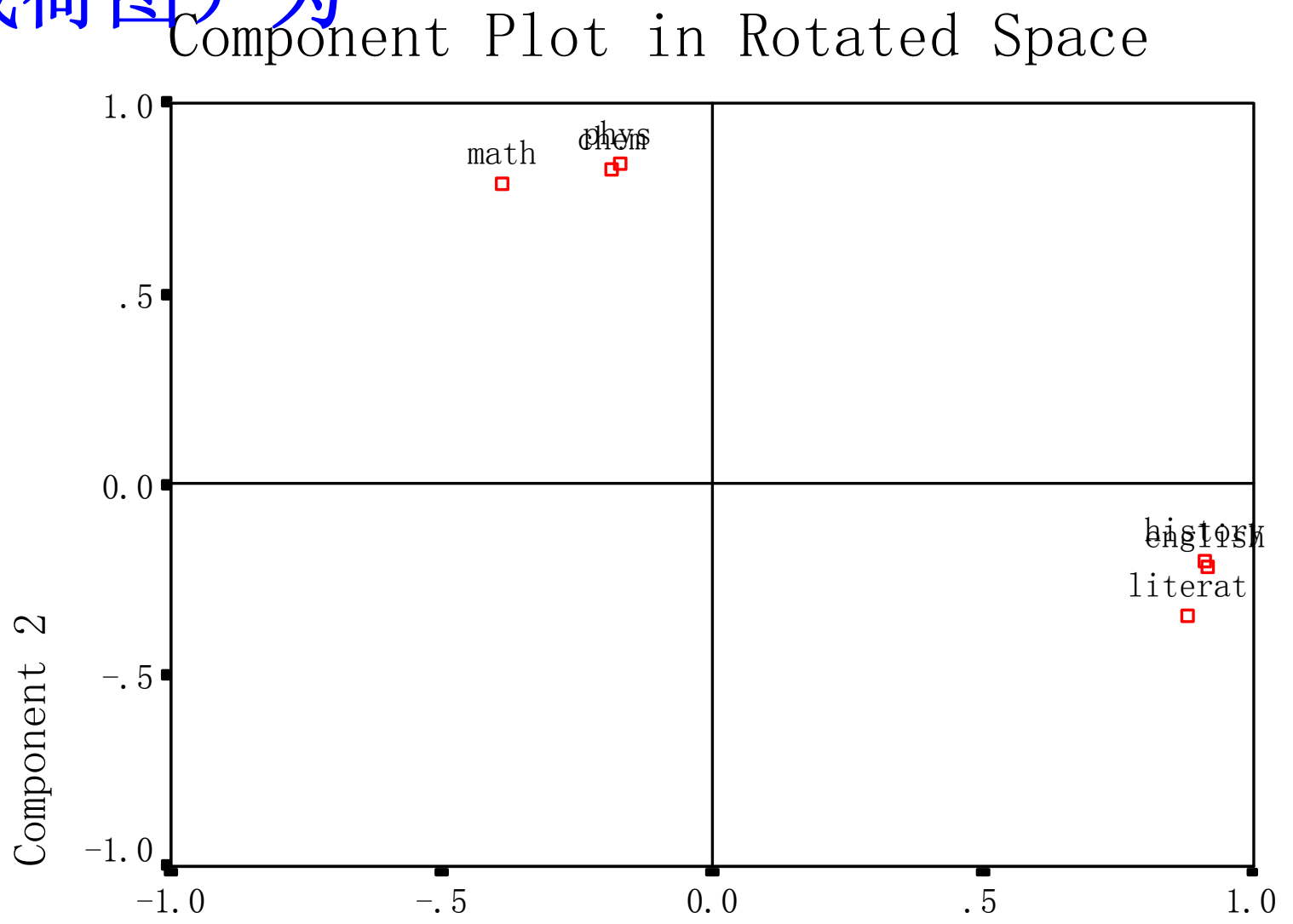
$$x_6 = 0.913\, f_1 - 0.216\, f_2$$

这里，第一个因子主要和语文、历史、英语三科有很强的正相关；而第二个因子主要和数学、物理、化学三科有很强的正相关。

因此可以给第一个因子起名为“文科因子”，而给第二个因子起名为“理科因子”。

从这个例子可以看出，因子分析的结果比主成分分析解释性更强。

- 这些系数所形成的散点图（在SPSS中也称载荷图）为



可以直观看出每个因子代表了一类学科

计算因子得分

- 可以根据输出

Component Score Coefficient Matrix

	Component	
	1	2
MATH	.036	.377
PHYS	.165	.474
CHEM	.155	.462
LITERAT	.357	.052
HISTORY	.417	.151
ENGLISH	.413	.142

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

- 算出每个学生的第一个因子和第二个因子的大小，即算出每个学生的因子得分f_1和f_2。

- 该输出说明第一和第二主因子为（习惯上用字母*f*来表示因子）可以按照如下公式计算，该函数称为因子得分（**factor score**）。

$$f_1 = 0.036x_1 + 0.165x_2 + 0.155x_3 + 0.357x_4 + 0.417x_5 + 0.413x_6$$

$$f_2 = 0.377x_1 + 0.474x_2 + 0.462x_3 + 0.052x_4 + 0.151x_5 + 0.142x_6$$

人们可以根据这两套因子得分对学生分别按照文科和理科排序。当然得到因子得分只是**SPSS**软件的一个选项。

10.3因子分析和主成分分析的一些注意事项

- 可以看出，因子分析和主成分分析都依赖于原始变量，也只能反映原始变量的信息。所以原始变量的选择很重要。
- 另外，如果原始变量都本质上独立，那么降维就可能失败，这是因为很难把很多独立变量用少数综合的变量概括。数据越相关，降维效果就越好。

10.3因子分析和主成分分析的一些注意事项

- 在得到分析的结果时，并不一定会都得到如我们例子那样清楚的结果。这与问题的性质，选取的原始变量以及数据的质量等都有关系
- 在用因子得分进行排序时要特别小心，特别是对于敏感问题。由于原始变量不同，因子的选取不同，排序可以很不一样。

SPSS实现(因子分析与主成分分析)

- 拿**student.sav**为例，选**Analyze－Data Reduction－Factor**进入主对话框；
- 把**math、phys、chem、literat、history、english**选入**Variables**，然后点击**Extraction**，
- 在**Method**选择一个方法（如果是主成分分析，则选**Principal Components**），
- 下面的选项可以随意，比如要画碎石图就选**Scree plot**，另外在**Extract**选项可以按照特征值的大小选主成分（或因子），也可以选定因子的数目；
- 之后回到主对话框（用**Continue**）。然后点击**Rotation**，再在该对话框中的**Method**选择一个旋转方法（如果是主成分分析就选**None**），
- 在**Display**选**Rotated solution**（以输出和旋转有关的结果）和**Loading plot**（以输出载荷图）；之后回到主对话框（用**Continue**）。
- 如果要计算因子得分就要点击**Scores**，再选择**Save as variables**（因子得分就会作为变量存在数据中的附加列上）和计算因子得分的方法（比如**Regression**）；要想输出**Component Score Coefficient Matrix**表，就要选择**Display factor score coefficient matrix**；之后回到主对话框（用**Continue**）。这时点**OK**即可。

第六章　因子分析
Factor Analysis

§1 引言

因子分析(factor analysis)是一种数据简化的技术。它通过研究众多变量之间的内部依赖关系，探求观测数据中的基本结构，并用少数几个假想变量来表示其基本的数据结构。这几个假想变量能够反映原来众多变量的主要信息。原始的变量是可观测的显在变量，而假想变量是不可观测的潜在变量，称为因子。

例如，在企业形象或品牌形象的研究中，消费者可以通过一个有24个指标构成的评价体系，评价百货商场的24个方面的优劣。

但消费者主要关心的是三个方面，即商店的环境、商店的服务和商品的价格。因子分析方法可以通过24个变量，找出反映商店环境、商店服务水平和商品价格的三个潜在的因子，对商店进行综合评价。而这三个公共因子可以表示为：

$$x_i = \mu_i + \alpha_{i1}F_1 + \alpha_{i2}F_2 + \alpha_{i3}F_3 + \varepsilon_i \quad i = 1, \cdots, 24$$

称 F_1、F_2、F_3 是不可观测的潜在因子。24个变量共享这三个因子，但是每个变量又有自己的个性，不被包含的部分 ε_i ，称为特殊因子。

注：

因子分析与回归分析不同，因子分析中的因子是一个比较抽象的概念，而回归因子有非常明确的实际意义；

主成分分析分析与因子分析也有不同，主成分分析仅仅是变量变换，而因子分析需要构造因子模型。

主成分分析:原始变量的线性组合表示新的综合变量，即主成分；

因子分析：潜在的假想变量和随机影响变量的线性组合表示原始变量。

§ 2 因子分析模型

一、数学模型

设 $X_i(i=1,2,\cdots,p)$ p 个变量，如果表示为

$$X_i=\mu_i+a_{i1}F_1+\cdots+a_{im}F_m+\varepsilon_i \qquad (m\le p)$$

$$\text{或}\begin{bmatrix}X_1\\X_2\\\vdots\\X_p\end{bmatrix}=\begin{bmatrix}\mu_1\\\mu_2\\\vdots\\\mu_p\end{bmatrix}+\begin{bmatrix}\alpha_{11}&\alpha_{12}&\cdots&\alpha_{1m}\\\alpha_{21}&\alpha_{22}&\cdots&\alpha_{2m}\\\vdots&\vdots&&\vdots\\\alpha_{p1}&\alpha_{p2}&\cdots&\alpha_{pm}\end{bmatrix}\begin{bmatrix}F_1\\F_2\\\vdots\\F_m\end{bmatrix}+\begin{bmatrix}\varepsilon_1\\\varepsilon_2\\\vdots\\\varepsilon_p\end{bmatrix}$$

$$\text{或}\mathbf{X}-\boldsymbol{\mu}=\mathbf{AF}+\varepsilon$$

称为 $F_1, F_2, \cdots, F_m$ 公共因子，是不可观测的变量，他们的系数称为因子载荷。ε_i 是特殊因子，是不能被前m个公共因子包含的部分。并且满足：

$\mathrm{cov}(F, \varepsilon) = 0$，　F, ε 即不相关；

$$D(F) = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix} = I$$

即 $F_1, F_2, \cdots, F_m$ 互不相关，方差为1。正交因子模型，否则为斜交因子模型

$$D(\varepsilon)=\begin{bmatrix}\sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_p^2\end{bmatrix}$$

即互不相关，方差不一定相等， $\varepsilon_i \sim N(0,\sigma_i^2)$。

用矩阵的表达方式

$$\mathbf{X}-\boldsymbol{\mu}=\mathbf{A}\mathbf{F}+\boldsymbol{\varepsilon} \qquad E(\mathbf{F})=\mathbf{0}$$

$$E(\boldsymbol{\varepsilon})=\mathbf{0} \qquad Var(\mathbf{F})=\mathbf{I}$$

$$\operatorname{cov}(\mathbf{F},\boldsymbol{\varepsilon})=E(\mathbf{F}\boldsymbol{\varepsilon}')=\begin{pmatrix} E(F_1\varepsilon_1) & E(F_1\varepsilon_2) & \cdots & E(F_1\varepsilon_p) \\ E(F_2\varepsilon_1) & E(F_2\varepsilon_2) & \cdots & E(F_2\varepsilon_p) \\ \vdots & \vdots & & \vdots \\ E(F_p\varepsilon_1) & E(F_p\varepsilon_2) & \cdots & E(F_p\varepsilon_p) \end{pmatrix}=\mathbf{0}$$

$$Var(\boldsymbol{\varepsilon})=diag(\sigma_1^2,\sigma_2^2,\cdots,\sigma_p^2)$$

二、因子分析模型的性质

1、原始变量X的协方差矩阵的分解

$$\because \quad \mathbf{X}-\boldsymbol{\mu}=\mathbf{AF}+\boldsymbol{\varepsilon}$$

$$\therefore \quad Var(\mathbf{X}-\boldsymbol{\mu})=\mathbf{A}Var(\mathbf{F})\mathbf{A}'+Var(\boldsymbol{\varepsilon})$$

$$\boldsymbol{\Sigma}_x=\mathbf{AA}'+\mathbf{D}$$

A是因子模型的系数

$$Var(\boldsymbol{\varepsilon})=\mathbf{D}=diag(\sigma_1^2,\sigma_2^2,\cdots,\sigma_p^2)$$

D的主对角线上的元素值越小，则公共因子共享的成分越多。见书176页说明

2、模型不受计量单位的影响

将原始变量X做变换**X*=CX**,这里

C＝diag(c_1,c_2,…,c_n),c_i>0。

$$\mathbf{C(X-\mu)=C(AF+\varepsilon)}$$

$$\mathbf{CX=C\mu+CAF+C\varepsilon}$$

$$\mathbf{X^*=C\mu+CAF+C\varepsilon}$$

$$\mathbf{X^*=\mu^*+A^*F^*+\varepsilon^*} \qquad \mathbf{F^*=F}$$

$$E(\mathbf{F}^*) = \mathbf{0}$$

$$E(\boldsymbol{\varepsilon}^*) = \mathbf{0}$$

$$Var(\mathbf{F}^*) = \mathbf{I}$$

$$Var(\boldsymbol{\varepsilon}^*) = diag(\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2)$$

$$\text{cov}(\mathbf{F}^*, \boldsymbol{\varepsilon}^*) = E(\mathbf{F}^* \boldsymbol{\varepsilon}^{*\prime}) = \mathbf{0}$$

3、因子载荷不是惟一的

设T为一个p×p的正交矩阵，令**$A^*=AT$**，**$F^*=T'F$**，则模型可以表示为

$$\mathbf{X}^* = \boldsymbol{\mu} + \mathbf{A}^*\mathbf{F}^* + \boldsymbol{\varepsilon}$$ 且满足条件因子模型的条件

$$E(\mathbf{T}'\mathbf{F}) = \mathbf{0} \qquad E(\boldsymbol{\varepsilon}) = \mathbf{0}$$

$$Var(\mathbf{F}^*) = Var(\mathbf{T}'\mathbf{F}) = \mathbf{T}'Var(\mathbf{F})\mathbf{T} = \mathbf{I}$$

$$Var(\boldsymbol{\varepsilon}) = diag(\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2)$$

$$\mathrm{cov}(\mathbf{F}^*, \boldsymbol{\varepsilon}) = E(\mathbf{F}^*\boldsymbol{\varepsilon}') = \mathbf{0}$$

三、 因子载荷矩阵中的几个统计特征

1、因子载荷a_{ij}的统计意义

因子载荷 a_{ij} 是第i个变量与第j个公共因子的相关系数

模型为 $X_i = a_{i1}F_1 + \cdots + a_{im}F_m + \varepsilon_i$

在上式的左右两边乘以F_j ，再求数学期望

$$E(X_iF_j) = a_{i1}E(F_1F_j) + \cdots + \alpha_{ij}E(F_jF_j) + \cdots + a_{im}E(F_mF_j) + E(\varepsilon_iF_j)$$

根据公共因子的模型性质，有

$\gamma_{x_iF_j} = \alpha_{ij}$ （载荷矩阵中第i行，第j列的元素）反映了第i个变量与第j个公共因子的相关重要性。绝对值越大，相关的密切程度越高。（见书175页）

2、变量共同度的统计意义

定义：变量 X_i 的共同度是因子载荷矩阵的第i行的元素的平方和。记为 $h_i^2 = \sum_{j=1}^{m} a_{ij}^2$。

统计意义：

$$X_i = a_{i1}F_1 + \cdots + a_{im}F_m + \varepsilon_i \quad \text{两边求方差}$$

$$Var(X_i) = a^2_{i1}Var(F_1) + \cdots + a^2_{im}Var(F_m) + Var(\varepsilon_i)$$

$$1 = \sum_{j=1}^{m} a_{ij}^2 + \sigma_i^2$$

所有的公共因子和特殊因子对变量 X_i 的贡献为1。如果 $\sum_{j=1}^{m} a_{ij}^2$ 非常靠近1，σ_i^2 非常小，则因子分析的效果好，从原变量空间到公共因子空间的转化性质好。

3、公共因子F_j方差贡献的统计意义

因子载荷矩阵中各列元素的平方和

$$S_j = \sum_{i=1}^{p} a_{ij}^2$$

称为所有的 $F_j\,(j=1,\cdots,m)$ 对 X_i 的方差贡献和。衡量 F_j 的相对重要性。见书176页

§ 3 因子载荷矩阵的估计方法

（一）主成分分析法

设随机向量 $\mathbf{x}=(x_1,x_2,\cdots,x_p)'$ 的均值为 $\boldsymbol{\mu}$，协方差为 $\boldsymbol{\Sigma}$，$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p\geq 0$ 为 Σ 的特征根，$\mathbf{u}_1,\mathbf{u}_2,\cdots,\mathbf{u}_p$ 为对应的标准化特征向量，则

$$\boldsymbol{\Sigma}=\mathbf{U}\begin{bmatrix}\lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p\end{bmatrix}\mathbf{U}'=\mathbf{A}\mathbf{A}'+\mathbf{D}$$

$$\begin{bmatrix}\mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_p\end{bmatrix}\begin{pmatrix}\lambda_1 & & 0\\ & \ddots & \\ 0 & & \lambda_p\end{pmatrix}\begin{bmatrix}\mathbf{u}_1'\\ \mathbf{u}_2'\\ \vdots\\ \mathbf{u}_p'\end{bmatrix}$$

$$= \lambda_1\mathbf{u}_1\mathbf{u}_1' + \lambda_2\mathbf{u}_2\mathbf{u}_2' + \cdots + \lambda_m\mathbf{u}_m\mathbf{u}_m' + \lambda_{m+1}\mathbf{u}_{m+1}\mathbf{u}_{m+1}' + \cdots + \lambda_p\mathbf{u}_p\mathbf{u}_p'$$

$$= \begin{bmatrix}\sqrt{\lambda_1}\mathbf{u}_1 & \sqrt{\lambda_2}\mathbf{u}_2 & \cdots & \sqrt{\lambda_p}\mathbf{u}_p\end{bmatrix}\begin{bmatrix}\sqrt{\lambda_1}\mathbf{u}_1'\\ \sqrt{\lambda_2}\mathbf{u}_2'\\ \vdots\\ \sqrt{\lambda_p}\mathbf{u}_p'\end{bmatrix}$$

上式给出的**Σ**表达式是精确的，然而，它实际上是毫无价值的，因为我们的目的是寻求用少数几个公共因子解释，故略去后面的p-m项的贡献，有

$$\boldsymbol{\Sigma} \approx \hat{\mathbf{A}}\hat{\mathbf{A}}' + \hat{\mathbf{D}} = \lambda_1 \mathbf{u}_1 \mathbf{u}_1' + \lambda_2 \mathbf{u}_2 \mathbf{u}_2' + \cdots + \lambda_m \mathbf{u}_m \mathbf{u}_m' + \hat{\mathbf{D}}$$

$$= \left[\sqrt{\lambda_1}\mathbf{u}_1 \quad \sqrt{\lambda_2}\mathbf{u}_2 \quad \cdots \quad \sqrt{\lambda_m}\mathbf{u}_m \right]_{p\times m} \begin{bmatrix} \sqrt{\lambda_1}\mathbf{u}_1' \\ \sqrt{\lambda_2}\mathbf{u}_2' \\ \vdots \\ \sqrt{\lambda_p}\mathbf{u}_m' \end{bmatrix}_{m\times p} + \hat{\mathbf{D}} \approx \hat{\mathbf{A}}\hat{\mathbf{A}}' + \hat{\mathbf{D}}$$

其中$\hat{\mathbf{D}} = diag(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \cdots, \hat{\sigma}_p^2)$

$$\hat{\sigma}_i^2 = s_{ii} - \sum_{j=1}^{m} a_{ij}^2$$

上式有一个假定，模型中的特殊因子是不重要的，因而从$\boldsymbol{\Sigma}$的分解中忽略了特殊因子的方差。

注：残差矩阵

$$\mathbf{S-\hat{A}\hat{A}'-\hat{D}}$$

其中S为样本的协方差矩阵。

（二）主因子法

主因子方法是对主成分方法的修正，假定我们首先对变量进行标准化变换。则

$$R=AA'+D$$

$$R^*=AA'=R-D$$

称R^*为约相关矩阵，R^*对角线上的元素是h_i^2，而不是1。

$$R^* = \mathbf{R} - \hat{\mathbf{D}} = \begin{bmatrix} \hat{h}_1^2 & r_{12} & \cdots & r_{1p} \\ r_{21} & \hat{h}_2^2 & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & \hat{h}_p^2 \end{bmatrix}$$

直接求**R***的前p个特征根和对应的正交特征向量。得如下的矩阵：

$$\mathbf{A} = \left[\sqrt{\lambda_1^*}\mathbf{u}_1^* \quad \sqrt{\lambda_2^*}\mathbf{u}_2^* \quad \cdots \quad \sqrt{\lambda_p^*}\mathbf{u}_p^* \right]$$

$\mathbf{R}^*$特征根：$\lambda_1^* \geq \cdots \geq \lambda_p^* \geq 0$

正交特征向量：$\mathbf{u}_1^*, \mathbf{u}_2^*, \cdots, \mathbf{u}_p^*$

当特殊因子 ε_i 的方差不为且已知的，问题非常好解决。

$$\mathbf{R}^{*}=\mathbf{R}-\begin{bmatrix}\sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_p^2\end{bmatrix}$$

$$=\begin{bmatrix}\sqrt{\lambda_1^{*}}\mathbf{u}_1^{*} & \sqrt{\lambda_2^{*}}\mathbf{u}_2^{*} & \cdots & \sqrt{\lambda_p^{*}}\mathbf{u}_p^{*}\end{bmatrix}\begin{bmatrix}\sqrt{\lambda_1^{*}}\mathbf{u}_1'^{*} \\ \sqrt{\lambda_2^{*}}\mathbf{u}_2'^{*} \\ \vdots \\ \sqrt{\lambda_p^{*}}\mathbf{u}_p'^{*}\end{bmatrix}$$

$$\mathbf{A} = \left[\sqrt{\lambda_1^*}\,\mathbf{u}_1^* \quad \sqrt{\lambda_2^*}\,\mathbf{u}_2^* \quad \cdots \quad \sqrt{\lambda_m^*}\,\mathbf{u}_m^* \right]$$

$$\mathbf{D} = \begin{pmatrix} 1-\hat{h}_1^2 & & 0 \\ & \ddots & \\ 0 & & 1-\hat{h}_p^2 \end{pmatrix}$$

在实际的应用中，个性方差矩阵一般都是未知的，可以通过一组样本来估计。**估计的方法有如下几种：**

首先，求h_i^2的初始估计值，构造出 $\mathbf{R}^*$

1）取$h_i^2=1$，在这个情况下主因子解与主成分解等价；

2）取 $h_i^2=R_i^2$ ，R_i^2为x_i与其他所有的原始变量x_j的复相关系数的平方，即x_i对其余的$p-1$个x_j的回归方程的判定系数，这是因为x_i 与公共因子的关系是通过其余的$p-1$个x_j 的线性组合联系起来的；

2）取 $\hat{h}_i^2 = \max|r_{ij}|(j \neq i)$ ，这意味着取x_i与其余的x_j的简单相关系数的绝对值最大者；

4）取 $h_i^2 = \frac{1}{p-1}\sum_{j=1,i\neq j}^{p} r_{ij}$ ，其中要求该值为正数。

5）取 $h_i^2 = 1/r^{ii}$，其中 r^{ii} 是 $\mathbf{R}^{-1}$ 的对角元素。

（三）极大似然估计法

如果假定公共因子**F**和特殊因子**ε**服从正态分布，那么可以得到因子载荷和特殊因子方差的极大似然估计。设 $\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n$ 为来自正态总体 $\mathbf{N_p(\mu,\Sigma)}$ 的随机样本。

$$\boldsymbol{\Sigma}=\mathbf{A}\mathbf{A}'+\boldsymbol{\Sigma}_{\varepsilon}$$

$$L(\hat{\boldsymbol{\mu}},\hat{\mathbf{A}},\hat{\mathbf{D}})=f(\mathbf{X})=f(X_1)\cdot f(X_2)\cdots f(X_n)$$

$$=\prod_{i=1}^{n}(2\pi)^{-p/2}|\Sigma|^{1/2}\exp[-\frac{1}{2}(x_i-\mu)'\Sigma^{-1}(x_i-\mu)]$$

$$=\left[(2\pi)^{p}|\Sigma|\right]^{-n/2}\exp[-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{X}_i-\mu)'\Sigma^{-1}(\mathbf{X}_i-\mu)]$$

它通过$\boldsymbol{\Sigma}$依赖$\mathbf{A}$和$\boldsymbol{\Sigma}_{\varepsilon}$。上式并不能唯一确定$\mathbf{A}$,为此可添加一个唯一性条件：

$$\mathbf{A}'\boldsymbol{\Sigma}_{\varepsilon}^{-1}\mathbf{A}=\boldsymbol{\Lambda}$$

这里$\boldsymbol{\Lambda}$式一个对角矩阵，用数值极大化的方法可以得到极大似然估计 $\hat{\mathbf{A}}$和$\hat{\boldsymbol{\Sigma}}_{\varepsilon}$。极大似然估计 $\hat{\mathbf{A}}$、$\hat{\boldsymbol{\Sigma}}_{\varepsilon}$和$\hat{\boldsymbol{\mu}}=\bar{\mathbf{x}}$将使 $\hat{\mathbf{A}}'\hat{\boldsymbol{\Sigma}}_{\varepsilon}^{-1}\hat{\mathbf{A}}=\hat{\boldsymbol{\Lambda}}$为对角阵，且似然函数达到最大。

相应的共同度的似然估计为：

$$\hat{h}_i^2=\hat{a}_{i1}^2+\hat{a}_{i2}^2+\cdots+\hat{a}_{im}^2$$

第J个因子对总方差的贡献：

$$S_j^2=\hat{a}_{1j}^2+\hat{a}_{2j}^2+\cdots+\hat{a}_{pj}^2$$

例 假定某地固定资产投资率 x_1，通货膨胀率 x_2，失业率 x_3，相关系数矩阵为

$$\begin{bmatrix} 1 & 1/5 & -1/5 \\ 1/5 & 1 & 2/5 \\ -1/5 & -2/5 & 1 \end{bmatrix}$$

试用主成分分析法求因子分析模型。

特征根为：$\lambda_1 = 1.55 \quad \lambda_2 = 0.85 \quad \lambda_3 = 0.6$

$$\mathbf{U}' = \begin{bmatrix} 0.475 & 0.883 & 0 \\ 0.629 & -0.331 & 0.707 \\ -0.629 & 0.331 & 0.707 \end{bmatrix}$$

$$\mathbf{A} = \begin{bmatrix} 0.475\sqrt{1.55} & 0.883\sqrt{0.85} & 0 \\ 0.629\sqrt{1.55} & -0.331\sqrt{0.85} & 0.707\sqrt{0.6} \\ -0.629\sqrt{1.55} & 0.331\sqrt{0.85} & 0.707\sqrt{0.6} \end{bmatrix}$$

$$= \begin{bmatrix} 0.569 & 0.814 & 0 \\ 0.783 & -0.305 & 0.548 \\ -0.783 & 0.305 & 0.548 \end{bmatrix}$$

$$x_1 = 0.569F_1 + 0.814F_2$$

$$x_2 = 0.783F_1 - 0.305F_2 + 0.548F_3$$

$$x_3 = -0.783F_1 + 0.305F_2 + 0.548F_3$$

可取前两个因子F1和F_2为公共因子，第一公因子F_1物价就业因子，对X的贡献为1.55。第一公因子F_2为投资因子，对X的贡献为0.85。共同度分别为1，0.706，0.706。

假定某地固定资产投资率x_1，通货膨胀率x_2，失业率x_3，相关系数矩阵为

$$\begin{bmatrix} 1 & 1/5 & -1/5 \\ 1/5 & 1 & 2/5 \\ -1/5 & -2/5 & 1 \end{bmatrix}$$

试用主因子分析法求因子分析模型。假定用$\hat{h}_i^2 = \max|r_{ij}|(j \neq i)$代替初始的$h_i^2$。$h_1^2 = \frac{1}{5}, h_2^2 = 1, h_3^2 = \frac{2}{5}$。

$$R^* = \begin{bmatrix} 1/5 & 1/5 & -1/5 \\ 1/5 & 1 & 2/5 \\ -1/5 & -2/5 & 2/5 \end{bmatrix} = \frac{1}{5}\begin{bmatrix} 1 & 1 & -1 \\ 1 & 5 & -2 \\ -1 & -2 & 2 \end{bmatrix}$$

特征根为：$\lambda_1 = 0.9123$

$\lambda_2 = 0.0877$

$\lambda_3 = 0$

对应的非零特征向量为：$\begin{bmatrix} 0.369 & 0.929 \\ 0.657 & -0.261 \\ -0.657 & 0.261 \end{bmatrix}$

$$\begin{bmatrix} 0.369\sqrt{0.9123} & 0.929\sqrt{0.0877} \\ 0.657\sqrt{0.9123} & -0.261\sqrt{0.0877} \\ -0.657\sqrt{0.9123} & 0.261\sqrt{0.0877} \end{bmatrix} = \begin{bmatrix} 0.352 & 0.275 \\ 0.628 & -0.077 \\ -0.628 & 0.077 \end{bmatrix}$$

$$x_1 = 0.352F_1 + 0.275F_2 + \varepsilon_1$$

$$x_2 = 0.625F_1 - 0.077F_2 + \varepsilon_2$$

$$x_1 = -0.682F_1 + 0.077F_2 + \varepsilon_3$$

新的共同度为：

$$h_1^2 = 0.352^2 + o.275^2 = 0.18129$$

$$h_2^2 = 0.625^2 + 0.077^2 = 0.3966$$

$$h_3^2 = 0.682^2 + 0.077^2 = 0.4710$$

§4 因子旋转（正交变换）

（一）为什么要旋转因子

建立了因子分析数学目的不仅仅要找出公共因子以及对变量进行分组，更重要的要知道每个公共因子的意义，以便进行进一步的分析，如果每个公共因子的含义不清，则不便于进行实际背景的解释。由于因子载荷阵是不惟一的，所以应该对因子载荷阵进行旋转。目的是使因子载荷阵的结构简化，使载荷矩阵每列或行的元素平方值向0和1两极分化。有三种主要的正交旋转法。四次方最大法、**方差最大法**和等量最大法。

（二）旋转方法

变换后因子的共同度

设Γ正交矩阵，做正交变换 $\mathbf{B}=\mathbf{A}\Gamma$

$$\mathbf{B}=(b_{ij})_{p\times p}=(\sum_{l=1}^{m}a_{il}\gamma_{lj})$$

$$h_i^2(\mathbf{B})=\sum_{j=1}^{m}b_{ij}^2=\sum_{j=1}^{m}(\sum_{l=1}^{m}a_{il}\gamma_{lj})^2$$

$$=\sum_{j=1}^{m}\sum_{l=1}^{m}a_{il}^2\gamma_{lj}^2+\sum_{j=1}^{m}\sum_{l=1}^{m}\sum_{\substack{t=1\\ j\neq l}}^{m}a_{il}a_{it}\gamma_{lj}\gamma_{tj}$$

$$=\sum_{l=1}^{m}a_{il}^2\sum_{j=1}^{m}\gamma_{lj}^2=\sum_{l=1}^{m}a_{il}^2=h_i^2(\mathbf{A})$$

变换后因子的共同度没有发生变化！

变换后因子贡献

设Γ正交矩阵，做正交变换 $\mathbf{B}=\mathbf{A}\Gamma$

$$\mathbf{B}=(b_{ij})_{p\times p}=(\sum_{l=1}^{q}a_{il}\gamma_{lj})$$

$$S_j^2(\mathbf{B})=\sum_{i=1}^{p}b_{ij}^2=\sum_{i=1}^{p}(\sum_{l=1}^{q}a_{il}\gamma_{lj})^2$$

$$=\sum_{i=1}^{p}\sum_{l=1}^{q}a_{il}^2\gamma_{lj}^2+\sum_{i=1}^{p}\sum_{l=1}^{q}\sum_{\substack{t=1\\t\neq l}}^{q}a_{il}a_{it}\gamma_{lj}\gamma_{tj}$$

$$=\sum_{i=1}^{p}a_{il}^2\sum_{l=1}^{q}\gamma_{lj}^2=S_j^2(\mathbf{A})\sum_{l=1}^{q}\gamma_{lj}^2$$

变换后因子的贡献发生了变化！

1、方差最大法

方差最大法从简化因子载荷矩阵的每一列出发，使和每个因子有关的载荷的平方的方差最大。当只有少数几个变量在某个因子上又较高的载荷时，对因子的解释最简单。方差最大的直观意义是希望通过因子旋转后，使每个因子上的载荷尽量拉开距离，一部分的载荷趋于±1，另一部分趋于0。

$$A=\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ \vdots & \vdots \\ a_{p1} & a_{p2} \end{bmatrix}$$

$$X_1 = a_{11}F_1 + a_{12}F_2$$
$$X_2 = a_{21}F_1 + a_{22}F_2$$
$$\cdots$$
$$X_p = a_{p1}F_1 + a_{p2}F_2$$

设旋转矩阵为：$T=\begin{pmatrix}\cos\varphi & -\sin\varphi\\ \sin\varphi & \cos\varphi\end{pmatrix}$

则 $B=AT=A\begin{pmatrix}\cos\varphi & -\sin\varphi\\ \sin\varphi & \cos\varphi\end{pmatrix}$

$$=\begin{pmatrix}a_{11}\cos\varphi+a_{12}\sin\varphi & -a_{11}\sin\varphi+a_{12}\cos\varphi\\ \vdots & \vdots\\ a_{p1}\cos\varphi+a_{p2}\sin\varphi & -a_{p1}\sin\varphi+a_{p1}\cos\varphi\end{pmatrix}$$

$$=\begin{pmatrix}a_{11}^{*} & a_{12}^{*}\\ \vdots & \vdots\\ a_{p1}^{*} & a_{p2}^{*}\end{pmatrix}$$

令 $d_{ij}=\dfrac{a_{ij}^{*}}{h_i}\qquad i=1,2,\cdots,p;\ j=1,2$

$\bar{d}_j=\dfrac{1}{p}\sum_{i=1}^{p}d_{ij}^{2}$(这是列和）

简化准则为：$V(\theta)=\sum_{j=1}^{m}\sum_{i=1}^{p}(d_{ij}^{2}-\bar{d}_j)^2=\max$

即：$V_1+V_2+V_3\cdots+V_m=\max\qquad(8.4.2)$

令 $\dfrac{\partial V}{\partial\theta}=0$，则可以解出$\theta_0$

旋转矩阵为：$T=\begin{pmatrix}\cos\theta_0 & -\sin\theta_0\\ \sin\theta_0 & \cos\theta_0\end{pmatrix}$

$$\mathbf{T} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{pmatrix} \qquad \mathbf{T}' = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{pmatrix}$$

$$\mathbf{T}\mathbf{T}' = \begin{pmatrix} 1 & & \\ & 1 & \\ & & 1 \end{pmatrix}$$

1、四次方最大旋转

四次方最大旋转是从简化载荷矩阵的行出发，通过旋转初始因子，使每个变量只在一个因子上又较高的载荷，而在其它的因子上尽可能低的载荷。如果每个变量只在一个因子上又非零的载荷，这是的因子解释是最简单的。

四次方最大法通过使因子载荷矩阵中每一行的因子载荷平方的方差达到最大。

简化准则为：$Q=\sum_{i=1}^{p}\sum_{j=1}^{m}(b_{ij}^2-\frac{1}{m})^2=\max$

$$Q=\sum_{i=1}^{p}\sum_{j=1}^{m}(b_{ij}^2-\frac{1}{m})^2=\sum_{i=1}^{p}\sum_{j=1}^{m}(b_{ij}^4-2\frac{1}{m}b_{ij}^2+\frac{1}{m^2})$$

$$=\sum_{i=1}^{p}\sum_{j=1}^{m}(b_{ij}^4-2\sum_{i=1}^{p}\sum_{j=1}^{m}\frac{1}{m}b_{ij}^2+\sum_{i=1}^{p}\sum_{j=1}^{m}\frac{1}{m^2})$$

$$=\sum_{i=1}^{p}\sum_{j=1}^{m}(b_{ij}^4-2\sum_{i=1}^{p}\sum_{j=1}^{m}\frac{1}{m}b_{ij}^2+\sum_{i=1}^{p}\sum_{j=1}^{m}\frac{1}{m^2})$$

$$=\sum_{i=1}^{p}\sum_{j=1}^{m}(b_{ij}^4-2+\frac{p}{m})$$

最终的简化准则为：$Q=\sum_{i=1}^{p}\sum_{j=1}^{m}b_{ij}^4=MAX$

3、等量最大法

等量最大法把四次方最大法和方差最大法结合起来求Q和V的加权平均最大。

最终的简化准则为：

$$E = \sum_{i=1}^{p}\sum_{j=1}^{m} b_{ij}^4 - \gamma \sum_{j=1}^{m}(\sum_{i=1}^{p} b_{ij}^2)^2 / p = MAX$$

权数γ等于m/2，因子数有关。

§5 因子得分

（一）因子得分的概念

前面我们主要解决了用公共因子的线性组合来表示一组观测变量的有关问题。如果我们要使用这些因子做其他的研究，比如把得到的因子作为自变量来做回归分析，对样本进行分类或评价，这就需要我们对公共因子进行测度，即给出公共因子的值。

因子分析的数学模型为：

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1m} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2m} \\ \vdots & \vdots & & \vdots \\ \alpha_{p1} & \alpha_{p2} & \cdots & \alpha_{pm} \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix}$$

原变量被表示为公共因子的线性组合，当载荷矩阵旋转之后，公共因子可以做出解释，通常的情况下，我们还想反过来把公共因子表示为原变量的线性组合。

因子得分函数：$F_j = \beta_{j1}X_1 + \cdots + \beta_{jp}X_p \qquad j = 1, \cdots, m$

可见，要求得每个因子的得分，必须求得分函数的系数，而由于p>m，所以不能得到精确的得分，只能通过估计

1、巴特莱特因子得分(加权最小二乘法)

1） 巴特莱特因子得分计算方法的思想

把 $x_i - \mu_i$ 看作因变量；把因子载荷矩阵

$$\begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1m} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2m} \\ \vdots & \vdots & & \vdots \\ \alpha_{p1} & \alpha_{p2} & \cdots & \alpha_{pm} \end{bmatrix}$$

看成自变量的观测；把某个个案的得分 F_{ij} 看着最小二乘法需要求的系数 。

$$\begin{cases} x_{i1}-\mu_1 = a_{11}f_1 + a_{12}f_2 + \cdots + a_{1m}f_m + \varepsilon_1 \\ x_{i2}-\mu_2 = a_{21}f_1 + a_{22}f_2 + \cdots + a_{2m}f_m + \varepsilon_2 \\ \cdots \\ x_{ip}-\mu_p = a_{p1}f_1 + a_{p2}f_2 + \cdots + a_{pm}f_m + \varepsilon_m \end{cases}$$

由于特殊因子的方差相异，所以用加权最小二乘法求得分，每个各案作一次，要求出所有样品的得分，需要作n次。

$$\sum_{j=1}^{p}\left[(x_{ij}-\mu_i)-(a_{i1}\hat{f}_1 + a_{i2}\hat{f}_2 + \cdots + a_{im}\hat{f}_m)\right]^2 / \sigma_i^2$$

使上式最小的$\hat{f}_1,\cdots,\hat{f}_m$是相应个案的因子得分。

用矩阵表达：$\mathbf{x-\mu=AF+\varepsilon}$

$$(\mathbf{x-\mu-AF})'\mathbf{D}^{-1}(\mathbf{x-\mu-AF})'=\min$$

$$\text{其中}\mathbf{D}^{-1}=\begin{pmatrix}\sigma_1^{-2} & & 0\\ & \ddots & \\ 0 & & \sigma_2^{-2}\end{pmatrix}$$

满足上式的F是相应个案的因子得分。

$$\frac{\partial}{\partial \mathbf{F}}(\mathbf{x}-\boldsymbol{\mu}-\mathbf{AF})'\mathbf{D}^{-1}(\mathbf{x}-\boldsymbol{\mu}-\mathbf{AF})'=0$$

$$-2\mathbf{A}'\mathbf{D}^{-1}(\mathbf{x}-\boldsymbol{\mu}-\mathbf{AF})'=0$$

$$\therefore \qquad \mathbf{A}'\mathbf{D}^{-1}(\boldsymbol{\varepsilon})'=0$$

$$\mathbf{D}^{-1}(\mathbf{x}-\boldsymbol{\mu})=\mathbf{D}^{-1}\mathbf{AF}+\mathbf{D}^{-1}\boldsymbol{\varepsilon}$$

$$\mathbf{A}'\mathbf{D}^{-1}(\mathbf{x}-\boldsymbol{\mu})=\mathbf{A}'\mathbf{D}^{-1}\mathbf{AF}+\mathbf{A}'\mathbf{D}^{-1}\boldsymbol{\varepsilon}$$

$$\mathbf{A}'\mathbf{D}^{-1}(\mathbf{x}-\boldsymbol{\mu})=\mathbf{A}'\mathbf{D}^{-1}\mathbf{AF}$$

$$\left[\mathbf{A}'\mathbf{D}^{-1}\mathbf{A}\right]^{-1}\mathbf{A}'\mathbf{D}^{-1}(\mathbf{x}-\boldsymbol{\mu})=\hat{\mathbf{F}}$$

2）得分估计的无偏性

- 如果将**f**和**ε**不相关的假定加强为相互独立，则

$$E(\hat{\mathbf{F}}/\mathbf{F}) = \left[\mathbf{A}'\mathbf{D}^{-1}\mathbf{A}\right]^{-1}\mathbf{A}'\mathbf{D}^{-1}E[(\mathbf{x}-\boldsymbol{\mu})/\mathbf{F})$$

$$= \left[\mathbf{A}'\mathbf{D}^{-1}\mathbf{A}\right]^{-1}\mathbf{A}'\mathbf{D}^{-1}E(\mathbf{AF}+\boldsymbol{\varepsilon}/\mathbf{F})$$

$$= \left[\mathbf{A}'\mathbf{D}^{-1}\mathbf{A}\right]^{-1}\mathbf{A}'\mathbf{D}^{-1}\mathbf{AF}$$

$$= \mathbf{A}^{-1}\mathbf{D}\mathbf{A}'^{-1}\mathbf{A}'\mathbf{D}^{-1}\mathbf{AF} = \mathbf{F}$$

$\hat{F}$的估计精度

3）

$$\hat{F}-F=\left[\mathbf{A}'\mathbf{D}^{-1}\mathbf{A}\right]^{-1}\mathbf{A}'\mathbf{D}^{-1}(\mathbf{AF}+\boldsymbol{\varepsilon})-\mathbf{F}$$

$$=\left[\mathbf{A}'\mathbf{D}^{-1}\mathbf{A}\right]^{-1}\mathbf{A}'\mathbf{D}^{-1}\boldsymbol{\varepsilon}$$

$$E(\hat{\mathbf{F}}-\mathbf{F})(\hat{\mathbf{F}}-\mathbf{F})'$$

$$=\left[\mathbf{A}'\mathbf{D}^{-1}\mathbf{A}\right]^{-1}\mathbf{A}'\mathbf{D}^{-1}E\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right)\mathbf{D}^{-1}\mathbf{A}\left[\mathbf{A}'\mathbf{D}^{-1}\mathbf{A}\right]^{-1}$$

$$=\left[\mathbf{A}'\mathbf{D}^{-1}\mathbf{A}\right]^{-1}\mathbf{A}'\mathbf{D}^{-1}\mathbf{D}\mathbf{D}^{-1}\mathbf{A}\left[\mathbf{A}'\mathbf{D}^{-1}\mathbf{A}\right]^{-1}$$

$$=\left[\mathbf{A}'\mathbf{D}^{-1}\mathbf{A}\right]^{-1}$$

2、回归方法

1) 思想

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1m} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2m} \\ \vdots & \vdots & & \vdots \\ \alpha_{n1} & \alpha_{n2} & \cdots & \alpha_{nm} \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\hat{F}_j = b_{j1}X_1 + \cdots + b_{jp}X_p \qquad j = 1, \cdots, m$$

$$\begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1p} \\ b_{21} & b_{22} & \cdots & b_{2p} \\ \vdots & \vdots & & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mp} \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_m \end{bmatrix}$$

$$\alpha_{ij} = \gamma_{x_i F_j} = E(X_i F_j)$$

$$= E[X_i(b_{j1}X_1 + \cdots + b_{jp}X_p)]$$

$$= b_{j1}\gamma_{i1} + \cdots + b_{jp}\gamma_{ip}$$

$$= \begin{bmatrix} r_{i1} & r_{i2} & \cdots & r_{ip} \end{bmatrix} \begin{bmatrix} b_{j1} \\ b_{j2} \\ \vdots \\ b_{jp} \end{bmatrix}$$

则，我们有如下的方程组：

$$\begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1p} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2p} \\ \vdots & \vdots & & \vdots \\ \gamma_{p1} & \gamma_{p2} & \cdots & \gamma_{pp} \end{bmatrix} \begin{bmatrix} b_{j1} \\ b_{j2} \\ \vdots \\ b_{jp} \end{bmatrix} = \begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{pj} \end{bmatrix} \quad j=1, 2, \cdots, m$$

$$\begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1p} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2p} \\ \vdots & \vdots & & \vdots \\ \gamma_{p1} & \gamma_{p2} & \cdots & \gamma_{pp} \end{bmatrix}$$ 为原始变量的相关系数 矩阵

$$\begin{bmatrix} b_{j1} \\ b_{j2} \\ \vdots \\ b_{jp} \end{bmatrix}$$ 为第 j个因子得分函数的系数

$$\begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{pj} \end{bmatrix}$$ 为载荷矩阵的第 j列

注：共需要解**m**次才能解出 所有的得分函数的系数。

§ 6 因子分析的步骤、展望和建议

一、 因子分析通常包括以下五个步骤

选择分析的变量

用定性分析和定量分析的方法选择变量，因子分析的前提条件是观测变量间有较强的相关性，因为如果变量之间无相关性或相关性较小的话，他们不会有共享因子,所以原始变量间应该有较强的相关性。

计算所选原始变量的相关系数矩阵

相关系数矩阵描述了原始变量之间的相关关系。可以帮助判断原始变量之间是否存在相关关系，这对因子分析是非常重要的，因为如果所选变量之间无关系，做因子分析是不恰当的。并且相关系数矩阵是估计因子结构的基础。

- 补充：KMO和Bartlett检验。 KMO值越接近1，则越适合作因子分析，KMO越小，则越不适合作因子分析。一个KMO的度量标准是：0.9以上非常适合；0.8适合；0.7一般；0.6不太适合；0.5以下不适合。
- 一般认为巴特利特球度检验值大于100适合做因子分析。

提取公共因子

这一步要确定因子求解的方法和因子的个数。需要根据研究者的设计方案或有关的经验或知识事先确定。因子个数的确定可以根据因子方差的大小。只取方差大于1(或特征值大于1)的那些因子，因为方差小于1的因子其贡献可能很小；按照因子的累计方差贡献率来确定，一般认为要达到60%才能符合要求；

因子旋转

通过坐标变换使每个原始变量在尽可能少的因子之间有密切的关系，这样因子解的实际意义更容易解释,并为每个潜在因子赋予有实际意义的名字。

计算因子得分

求出各样本的因子得分，有了因子得分值，则可以在许多分析中使用这些因子，例如以因子的得分做聚类分析的变量，做回归分析中的回归因子。

因子分析是十分主观的，在许多出版的资料中，因子分析模型都用少数可阐述因子提供了合理解释。实际上，绝大多数因子分析并没有产生如此明确的结果。不幸的是，评价因子分析质量的法则尚未很好量化，质量问题只好依赖一个

“哇！”准则

如果在仔细检查因子分析的时候，研究人员能够喊出“哇，我明白这些因子”的时候，就可看着是成功运用了因子分析方法。

§7 因子分析和主成分分析的一些注意事项

- 可以看出，因子分析和主成分分析都依赖于原始变量，也只能反映原始变量的信息。所以原始变量的选择很重要。
- 另外，如果原始变量都本质上独立，那么降维就可能失败，这是因为很难把很多独立变量用少数综合的变量概括。数据越相关，降维效果就越好。

§8 因子分析和主成分分析的一些注意事项

- 在得到分析的结果时，并不一定会都得到如我们例子那样清楚的结果。这与问题的性质，选取的原始变量以及数据的质量等都有关系
- 在用因子得分进行排序时要特别小心，特别是对于敏感问题。由于原始变量不同，因子的选取不同，排序可以很不一样。

SPSS实现(因子分析与主成分分析)

- 拿**student.sav**为例，选**Analyze－Data Reduction－Factor**进入主对话框；
- 把**math、phys、chem、literat、history、english**选入**Variables**，然后点击**Extraction**，
- 在**Method**选择一个方法（如果是主成分分析，则选**Principal Components**），
- 下面的选项可以随意，比如要画碎石图就选**Scree plot**，另外在**Extract**选项可以按照特征值的大小选主成分（或因子），也可以选定因子的数目；
- 之后回到主对话框（用**Continue**）。然后点击**Rotation**，再在该对话框中的**Method**选择一个旋转方法（如果是主成分分析就选**None**），
- 在**Display**选**Rotated solution**（以输出和旋转有关的结果）和**Loading plot**（以输出载荷图）；之后回到主对话框（用**Continue**）。
- 如果要计算因子得分就要点击**Scores**，再选择**Save as variables**（因子得分就会作为变量存在数据中的附加列上）和计算因子得分的方法（比如**Regression**）；要想输出**Component Score Coefficient Matrix**表，就要选择**Display factor score coefficient matrix**；之后回到主对话框（用**Continue**）。这时点**OK**即可。

- 上机练习：
- 1.

北京市各区县主要指标因子分析.sav

- 2. 自己搜集有关数据，进行保险公司绩效的因子分析。
- 主要参考文章：
- 1.陈飞跃. 基于因子分析的寿险公司经营绩效研究，现代商贸工业，2008年1月
- 2.张邯,马广军,田高良.我国保险公司绩效影响因素的实证研究,当代经济科学,2007年5月

阅读文章

- 王芳.主成分分析与因子分析的异同比较及应用，统计教育，2003 年第5期
- 林海明.如何用SPSS 软件一步算出主成分得分值，统计与信息论坛，2007年9月

第七章 对应分析

行和列变量的相关问题

- 在因子分析中，或者只对变量（列中的变量）进行分析，或者只对样品（观测值或行中的变量）进行分析；而且利用载荷图来描述各个变量之间的接近程度。
- 典型相关分析也只研究列中两组变量之间的关系。

行和列变量的相关问题

- 然而，在很多情况下，所关心的不仅仅是行或列本身变量之间的关系，而是行变量和列变量的相互关系；
- 这就是因子分析等方法所没有说明的了。先看一个例子。

例子（数据ChMath.txt)

- 为了考察汉字具有的抽象图形符号的特性能否会促进儿童空间和抽象思维能力。该数据以列联表形式展示在表中：
- 在研究读写汉字能力与数学的关系的研究时，人们取得了232个美国亚裔学生的数学成绩和汉字读写能力的数据。

例子（数据ChMath.txt)

- 该数据关于汉字读写能力的变量有三个水平：
- “纯汉字”意味着可以完全自由使用纯汉字读写，
- “半汉字”意味着读写中只有部分汉字（比如日文），
- 而“纯英文”意味着只能够读写英文而不会汉字。而数学成绩有4个水平（A、B、C、D）。

		数学成绩				总和
		数学A	数学B	数学C	数学F	
汉字使用	纯汉字	47	31	2	1	81
	半汉字	22	32	21	10	85
	纯英文	10	11	25	20	66
Total		79	74	48	31	232

人们可以对这个列联表进行前面所说的χ^2检验来考察行变量和列变量是否独立。结果在下面表中（通过Analyze－Descriptive Statistics－Crosstabs）

所有的检验都很显著，看来两个变量的确不独立。

Chi-Square Tests

	Value	Df	Asymp. Sig. (2-sided)
Pearson Chi-Square	75.312(a)	6	.000
Likelihood Ratio	84.090	6	.000
Linear-by-Linear Association	64.086	1	.000
N of Valid Cases	232		

对应分析

- 但是如何用象因子分析的载荷图那样的直观方法来展示这两个变量各个水平之间的关系呢？这就是对应分析（**correspondence analysis**）方法。
- 对应分析方法被普遍认为是探索性数据分析的内容，因此，读者只要能够会用数据画出描述性的点图，并能够理解图中包含的信息即可。

对应分析

- 处理列联表的问题仅仅是对应分析的一个特例。一般地，
- 对应分析常规地处理连续变量的数据矩阵；这些数据具有如在主成分分析、因子分析、聚类分析等时所处理的数据形式。

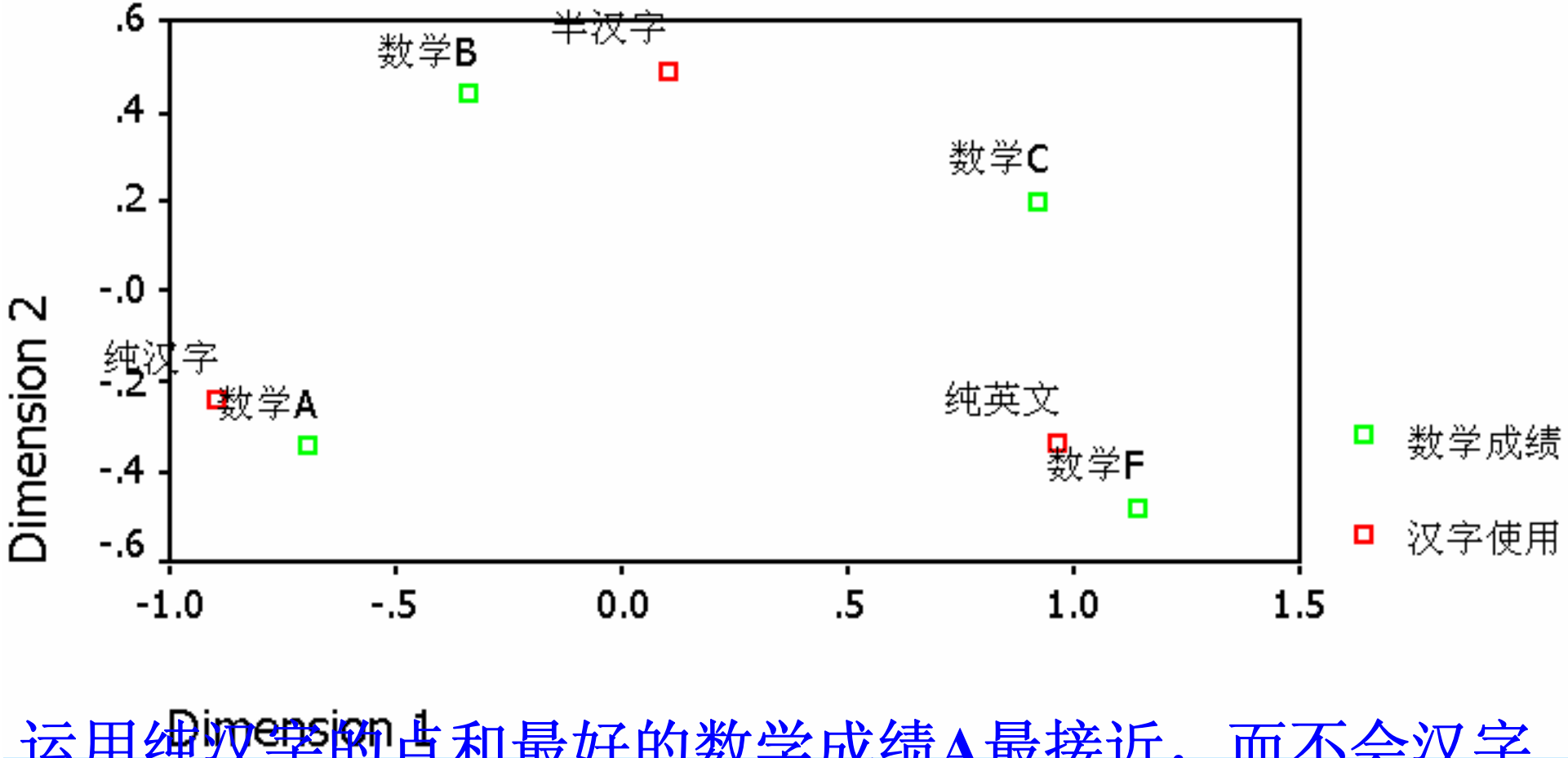
对应分析

- 在对应分析中，根据各行变量的因子载荷和各列变量的因子载荷之间的关系，行因子载荷和列因子载荷之间可以两两配对。
- 如果对每组变量选择前两列因子载荷，则两组变量就可画出两因子载荷的散点图。
- 由于这两个图所表示的载荷可以配对，于是就可以把这两个因子载荷的两个散点图画到同一张图中，并以此来直观地显示各行变量和各列变量之间的关系。

对应分析

- 由于列联表数据形式和一般的连续变量的数据形式类似，所以也可以用对应分析的数学方法来研究行变量各个水平和列变量各个水平之间的关系；
- 虽然对不同数据类型所产生结果的解释有所不同，数学的原理是一样的。下面通过对**ChMath.txt**数据的计算和结果分析来介绍对应分析。

首先看对应分析结果的一个主要SPSS展示，然后再解释该图的来源和解释。



运用纯汉字的点和最好的数学成绩A最接近，而不会汉字只会英文的点与最差的数学成绩F（或者D，虽然在纵坐标稍有差距）最接近，而用部分汉字的和数学成绩B接近。

对应分析的数学原理是什么？

结果解释

- 根据**SPSS**对数据**ChMath.sav**的计算，得到一些表格。
- 其中第一个就是下面的各维的汇总表。这里所涉及的是行与列因子载荷之间的关系；选择行和列变量的显著的因子载荷的标准是一样的。选择多少就涉及几维。为了画出散点图，就至少要选择两维了。

Summary

Dimension	Singular Value	Inertia	Chi Square	Sig.	Proportion of Inertia		Confidence Singular Value	
					Accounted for	Cumulative	Standard Deviation	Correlation 2
1	.552	.305			.939	.939	.047	.174
2	.141	.020			.061	1.000	.065	
Total		.325	75.312	.000(a)	1.000	1.000		

a 6 degrees of freedom

表中的术语

- **Inertia**－惯量，为每一维到其重心的加权距离的平方。它度量行列关系的强度。
- **Singular Value**－奇异值（是惯量的平方根），反映了是行与列各水平在二维图中分量的相关程度，是对行与列进行因子分析产生的新的综合变量的典型相关系数。
- **Chi Square**－就是关于列联表行列独立性χ^2检验的χ^2统计量的值，和前面表中的相同。其后面的**Sig**为在行列独立的零假设下的**p-**值，注释表明自由度为**(4-1)×(3-1)=6**，**Sig.**值很小说明列联表的行与列之间有较强的相关性。
- **Proportion of Inertia**－惯量比例，是各维度（公因子）分别解释总惯量的比例及累计百分比，类似于因子分析中公因子解释能力的说明。

解释

- 从该表可以看出，由于第一维的惯量比例占了总比例的93.9%，因此，其他维的重要性可以忽略（虽然画图时需要两维，但主要看第一维—横坐标）。
- 在SPSS的输出中还有另外两个表分别给出了画图中两套散点图所需要的两套坐标。

Overview Row Points(a)

汉字使用	Mass	Score in Dimension		Inertia	Contribution				
					Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
		1	2		1	2	1	2	Total
纯汉字	.349	-.897	-.240	.158	.509	.142	.982	.018	1.000
半汉字	.366	.102	.491	.015	.007	.627	.144	.856	1.000
纯英文	.284	.970	-.338	.152	.485	.231	.970	.030	1.000
Active Total	1.000			.325	1.000	1.000			

Overview Column Points(a)

数学成绩	Mass	Score in Dimension		Inertia	Contribution				
					Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
		1	2		1	2	1	2	Total
数学A	.341	-.693	-.345	.096	.296	.288	.940	.060	1.000
数学B	.319	-.340	.438	.029	.067	.433	.703	.297	1.000
数学C	.207	.928	.203	.100	.323	.061	.988	.012	1.000
数学F	.134	1.140	-.479	.100	.315	.218	.957	.043	1.000
Active Total	1.000			.325	1.000	1.000			

a Symmetrical normalization

解释

- 该表给出了图中三个汉字使用点的坐标：纯汉字(-.897,-.240)，半汉字(.102,.491)，纯英文(.970,-.338)，以及四个数学成绩点的坐标：数学**A(-.693,-.345)**，数学**B(-.340,.438)**，数学**C(.928,.203)**，数学**C(1.140,-.479)**。
- 两表中的概念不必记；其中**Mass**为行与列的边缘概率；**Score in Dimension**是各维度的分值 (二维图中的坐标)；**Inertia:**就是前面所提到的惯量，为每一行/列到其重心的加权距离

SPSS的实现

- 打开**ChMath.sav**数据，其形式和本章开始的列联表有些不同。其中**ch**列代表汉字使用的三个水平；而**math**列代表数学成绩的四个水平；第一列**count**实际上是**ch**和**math**两个变量各个水平组合的出现数目，也就是列联表中间的数目。
- 由于**count**把很大的本应有**232**行的原始数据简化成只有**12**行的汇总数据，在进行计算之前必须进行加权。也就是点击图标中的小天平，再按照**count**加权即可。

SPSS的实现

- 原始数据的预处理，即将待分析的两组原始数据组织成两个**SPSS**变量的形式；如果没有原始数据而只有交叉分组下的频数数据，则在对应分析前要对数据进行加权处理，指定加权变量。
- 加权之后，选择**Analyze－Data Reduction－Correspondence Analysis**，
- 然后把“汉字使用”选入**Row**（行），再点击**Define Range**来定义其范围为**1(Minimum value)**到**3(Maximum value)**，之后点击**Update**。
- 类似地，点击**Continue**之后，把“数学成绩”选入**Column (**列**)**，并以同样方式定义其范围为**1**到**4**。
- 由于其他选项可以用默认值，就可以直接点击**OK**来运行了。这样就得到上述表格和点图。

附录
对应分析的数学

因子分析对变量和对样品要分别对待. 对应分析把变量和样本同时反映到相同坐标轴(因子轴)的一张图形上.

数学上, 令$\mathbf{A}=[a_{ij}]$为$\mathbf{n}\times\mathbf{p}$矩阵, $\mathbf{x}=[x_i]$ 为n-(列)向量, $\mathbf{y}=[y_j]$ 为p-(列)向量. 那么(r,x,y)称为对应分析问题$C_0(A)$的解, 如果

$$rx_i = \sum_{j=1}^{n} \frac{a_{ij} y_j}{a_{i.}} \quad (i = 1, \ldots, n)$$

$$ry_j = \sum_{i=1}^{m} \frac{a_{ij} x_i}{a_{.j}} \quad (j = 1, \ldots, p).$$

$$rx_i = \sum_{j=1}^{n} \frac{a_{ij} y_j}{a_{i.}} \quad (i = 1, \ldots, n)$$

$$ry_j = \sum_{i=1}^{m} \frac{a_{ij} x_i}{a_{.j}} \quad (j = 1, \ldots, p).$$

行记分(**row score**) x_i和列记分y_j的加权均值成比例, 而列记分y_j和行记分x_i的加权均值成比例. 数值**r**为行列记分的相关(在典型相关的意义上).

记$R=diag(a_{i.})$, $C=diag(a_{.i})$, $R^{1/2}= diag(a_{.i}^{1/2})$,
则上面式子为
$rx=R^{-1}Ay$; $ry=C^{-1}A'x$
或
$rR^{1/2}x=(R^{-1/2}AC^{-1/2})C^{1/2}y$;
$rC^{1/2}y=(C^{-1/2}A' R^{-1/2})R^{1/2}x= (R^{-1/2} A C^{-1/2})'R^{1/2}x$
X为一个解的条件是下面特征值问题有解(最大特征值为1是平凡解, 两组非零特征值相同!)

$$r^2(R^{\frac{1}{2}}x)=(R^{-\frac{1}{2}}AC^{-\frac{1}{2}})(R^{-\frac{1}{2}}AC^{-\frac{1}{2}})'(R^{\frac{1}{2}}x)$$

$$r^2(C^{\frac{1}{2}}y)=(R^{-\frac{1}{2}}AC^{-\frac{1}{2}})'(R^{-\frac{1}{2}}AC^{-\frac{1}{2}})(C^{\frac{1}{2}}y)$$

令

$$Z \equiv (R^{-\frac{1}{2}} A C^{-\frac{1}{2}}), v \equiv R^{\frac{1}{2}} x, u \equiv C^{\frac{1}{2}} y$$

前面的特征值问题可以写成

$$r^2 u = Z'Zu$$

$$r^2 v = ZZ'v$$

两个特征值问题有同样的非零特征值.
如**U**是**Z'Z**的特征向量, 则**ZU**是**ZZ'**的特征向量.

Z'Z的特征根为$\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p$; Z'Z相应的特征向量为$u_1, u_2 \ldots, u_p$. ZZ'相应的特征向量为$v_1, v_2 \ldots, v_n$.对最大的m个特征值得因子载荷阵

$$F = \begin{bmatrix} u_{11}\sqrt{\lambda_1} & u_{12}\sqrt{\lambda_2} & \cdots & u_{1m}\sqrt{\lambda_m} \\ u_{21}\sqrt{\lambda_1} & u_{22}\sqrt{\lambda_2} & \cdots & u_{2m}\sqrt{\lambda_m} \\ \vdots & \vdots & \ddots & \vdots \\ u_{p1}\sqrt{\lambda_1} & u_{p2}\sqrt{\lambda_2} & \cdots & v_{pm}\sqrt{\lambda_m} \end{bmatrix} \quad G = \begin{bmatrix} v_{11}\sqrt{\lambda_1} & v_{12}\sqrt{\lambda_2} & \cdots & v_{1m}\sqrt{\lambda_m} \\ v_{21}\sqrt{\lambda_1} & v_{22}\sqrt{\lambda_2} & \cdots & v_{2m}\sqrt{\lambda_m} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1}\sqrt{\lambda_1} & v_{n2}\sqrt{\lambda_2} & \cdots & v_{nm}\sqrt{\lambda_m} \end{bmatrix}$$

可以对变量和样品作两两因子载荷图.

返回

第八章 相关分析

1 问题的提出

- 对于现实世界，不仅要知其然，而且要知其所以然。顾客对商品和服务的反映对于企业是至关重要的，
- 但是仅仅有满意顾客的比例是不够的；商家希望了解什么是影响顾客观点的因素，及这些因素如何起作用。
- 类似地，医疗卫生部门不能仅仅知道某流行病的发病率，而且想知道什么变量影响发病率，以及如何影响。

1 问题的提出

- 发现变量之间的统计关系，并且用此规律来帮助我们进行决策才是统计实践的最终目的。
- 一般来说，统计可以根据目前所拥有的信息（数据）来建立人们所关心的变量和其他有关变量的关系。这种关系一般称为**模型（model）**。

1 问题的提出

- 假如用Y表示感兴趣的变量，用X表示其他可能与Y有关的变量（X也可能是若干变量组成的向量）。则所需要的是建立一个函数关系$Y=f(X)$。
- 这里Y称为因变量或响应变量(dependent variable, response variable)，而X称为自变量，也称为解释变量或协变量(independent variable, explanatory variable, covariate)。建立这种关系的过程就叫做回归(regression)。

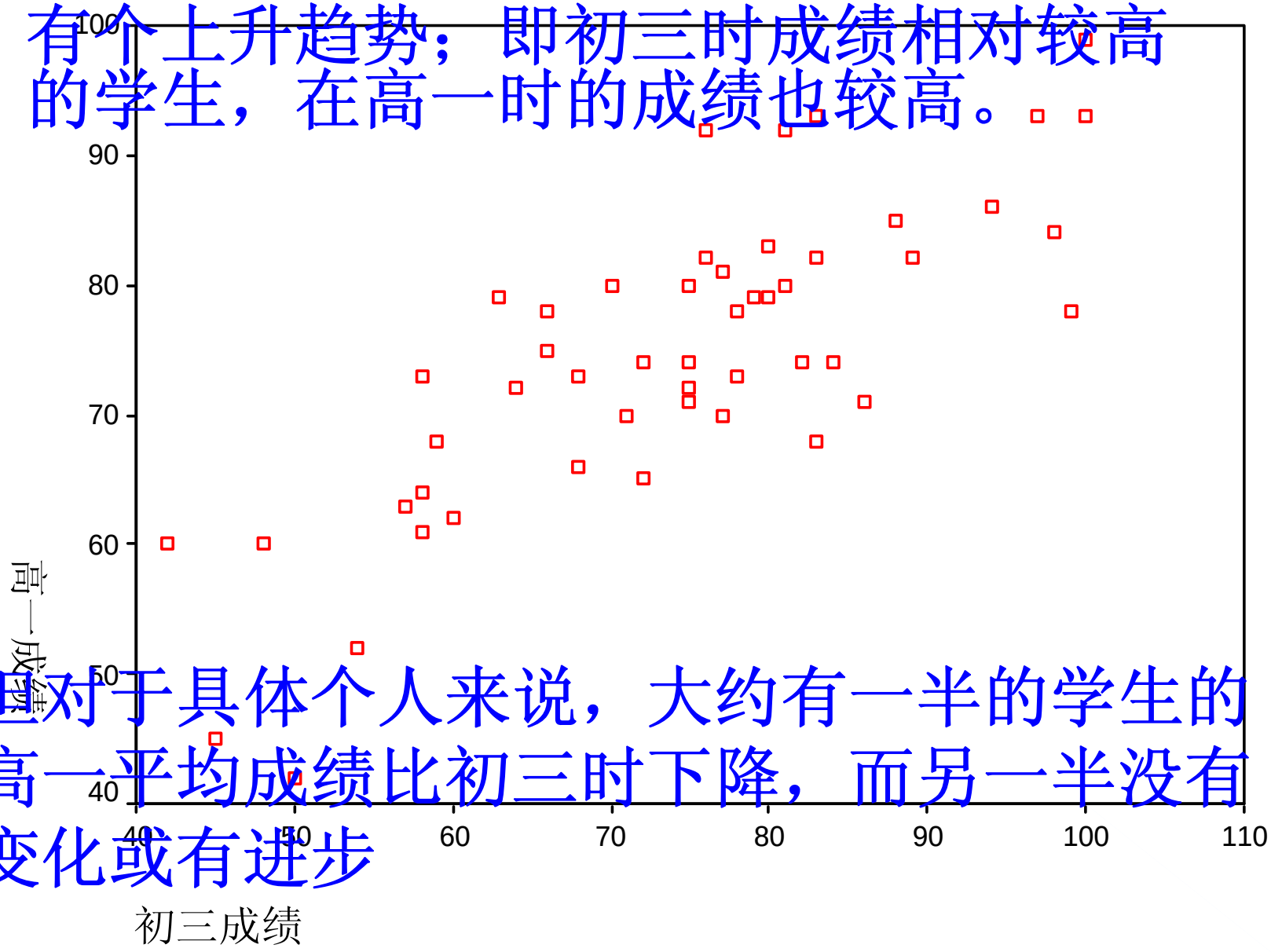
1 问题的提出

- 一旦建立了回归模型，除了对变量的关系有了进一步的定量理解之外，还可以利用该模型（函数）通过自变量对因变量做预测（**prediction**）。
- 这里所说的预测，是用已知的自变量的值通过模型对未知的因变量值进行估计；它并不一定涉及时间先后。
- 先看几个后面还要讨论的数值例子。

1 问题的提出

- 例1 有50个从初中升到高中的学生，为了比较初三的成绩是否和高中的成绩相关，得到了他们在初三和高一的各科平均成绩(数据在highschool.txt)。这两个成绩的散点图展示在图7.1中。

50名同学初三和高一成绩的散点图



有个上升趋势；即初三时成绩相对较高的学生，在高一时的成绩也较高。

但对于具体个人来说，大约有一半的学生的高一平均成绩比初三时下降，而另一半没有变化或有进步

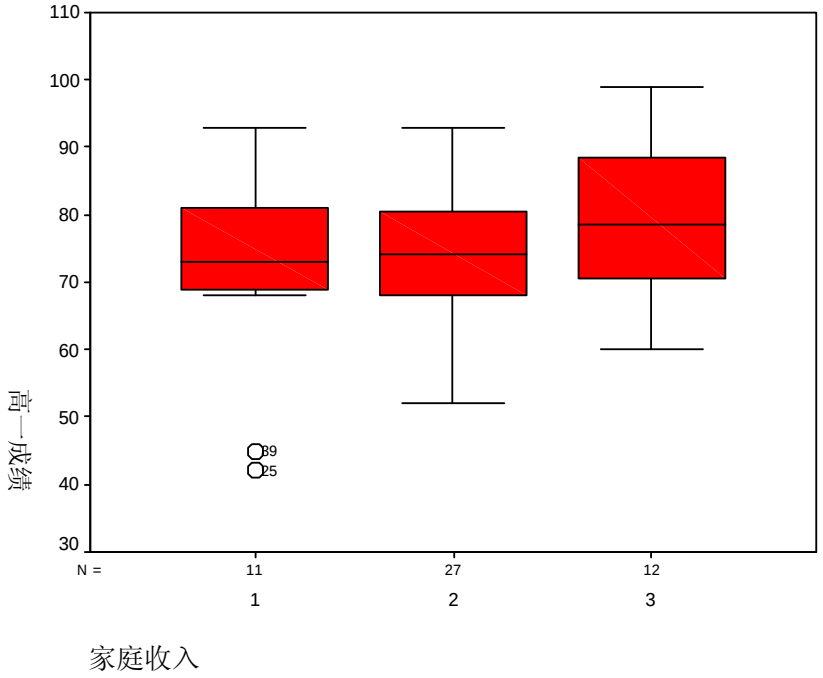
1 问题的提出

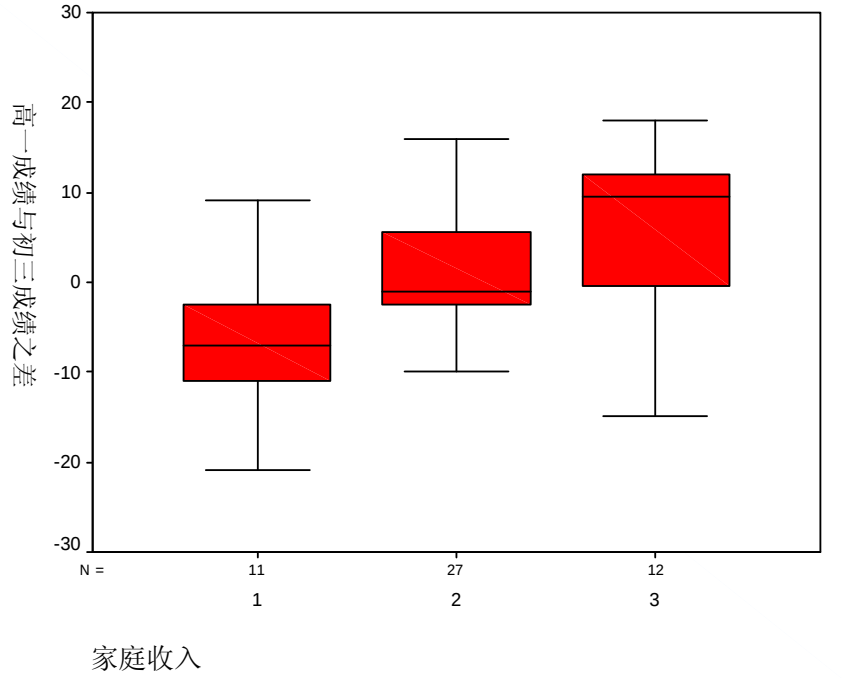
- 目前的问题是怎么判断这两个变量是否相关、如何相关及如何度量相关？
- 能否以初三成绩为自变量，高一成绩为因变量来建立一个回归模型以描述这样的关系，或用于预测。

1 问题的提出

- 该数据中，除了初三和高一的成绩之外，还有一个定性变量（没有出现在上面的散点图中）。它是学生在高一时的家庭收入状况；它有三个水平：低、中、高，分别在数据中用1、2、3表示。

为研究家庭收入情况对学生成绩变化的影响，下面点出两个盒形图，左边一个是不同收入群体的高一成绩的盒形图，右边一个是不同收入群体的高一和初三成绩之差的盒形图。





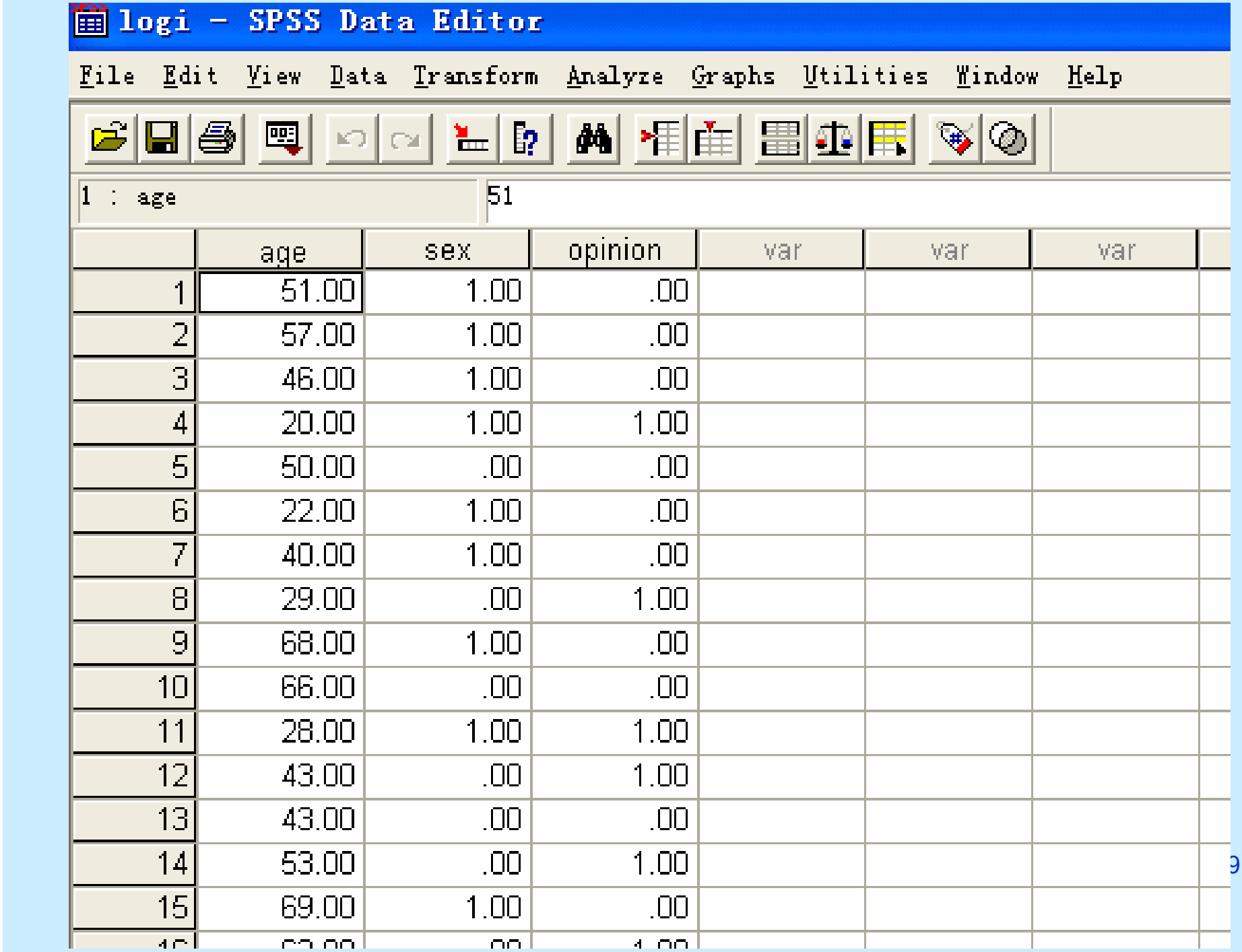
•可以看出收入高低对高一成绩稍有影响，但不如收入对成绩的变化（高一和初三成绩之差）的影响那么明显。

1 问题的提出

- 到底学生在高一的家庭收入对成绩有影响吗？是什么样的影响？
- 是否可以取初三成绩（这是定量变量）或（和）家庭收入（定性变量）为自变量，而取高一成绩为因变量，来建立一个描述这些变量之间关系的回归模型呢？

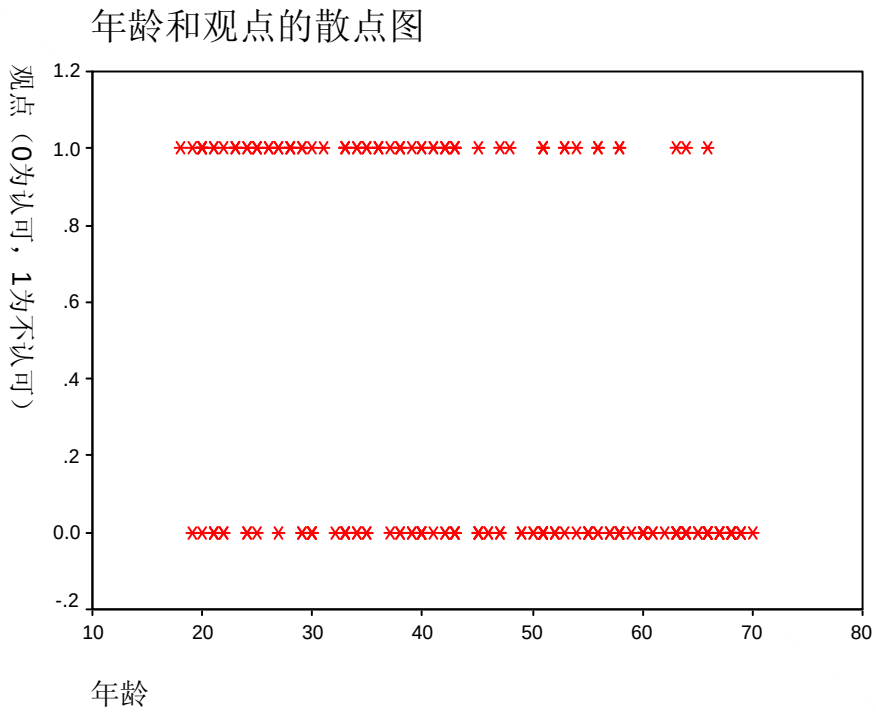
1 问题的提出

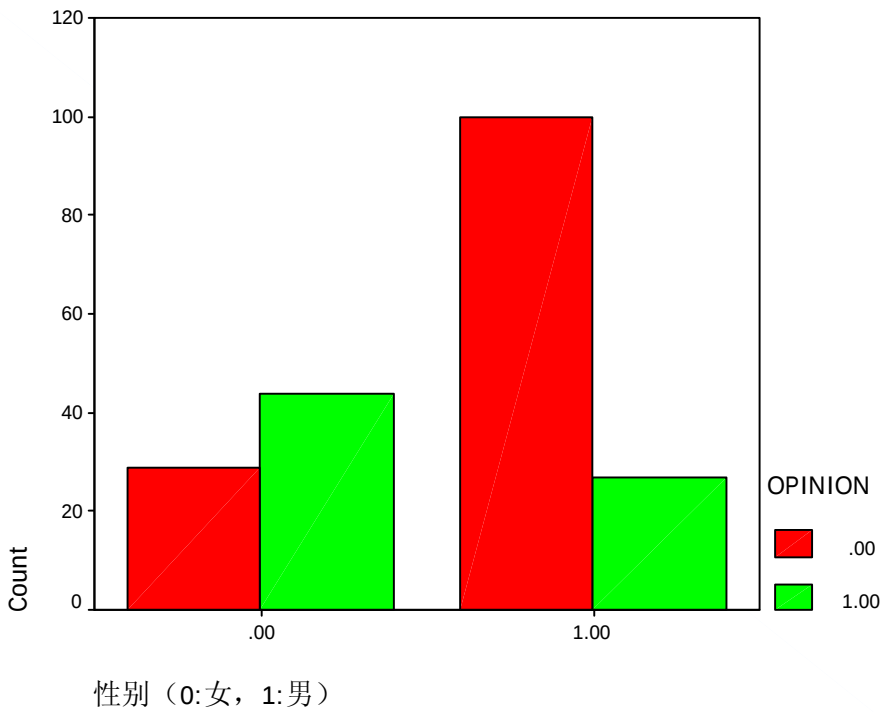
- 例2 这是200个不同年龄和性别的人对某项服务产品的认可的数据（logi.txt）。这里年龄是连续变量，性别是有男和女（分别用1和0表示）两个水平的定性变量，而变量观点则为包含认可（用1表示）和不认可（用0表示）两个水平的定性变量（见下页数据）。
- 想要知道的是年龄和性别对观点有没有影响，有什么样的影响，以及能否用统计模型表示出这个关系。



	age	sex	opinion	var	var	var
1	51.00	1.00	.00			
2	57.00	1.00	.00			
3	46.00	1.00	.00			
4	20.00	1.00	1.00			
5	50.00	.00	.00			
6	22.00	1.00	.00			
7	40.00	1.00	.00			
8	29.00	.00	1.00			
9	68.00	1.00	.00			
10	66.00	.00	.00			
11	28.00	1.00	1.00			
12	43.00	.00	1.00			
13	43.00	.00	.00			
14	53.00	.00	1.00			
15	69.00	1.00	.00			

年龄和观点的散点图(左)和性别与观点的条形图；

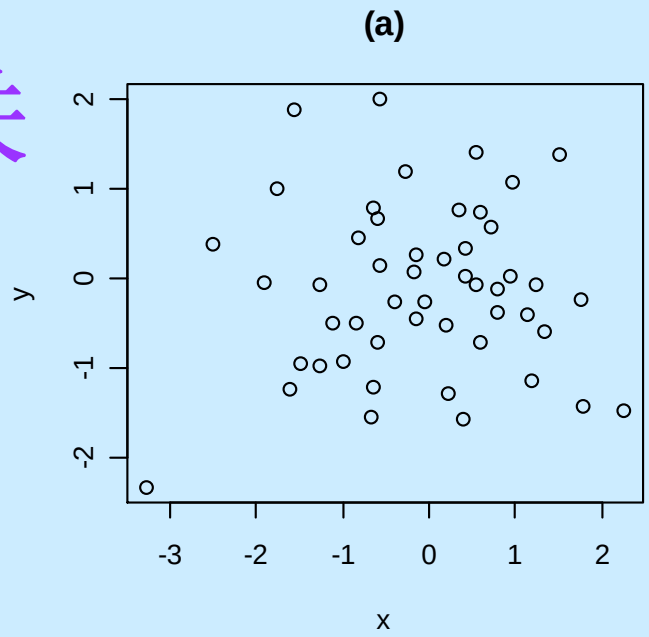




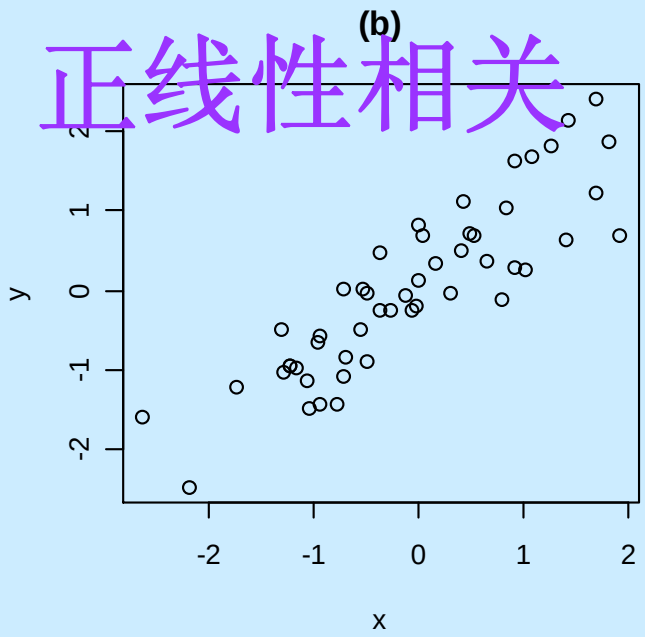
§2 定量变量的相关

- 如果两个定量变量没有关系，就谈不上建立模型或进行回归，但怎样才能发现两个变量有没有关系呢？
- 最简单的直观办法就是画出它们的散点图。下面是四组数据的散点图；每一组数据表示了两个变量x和y的样本。

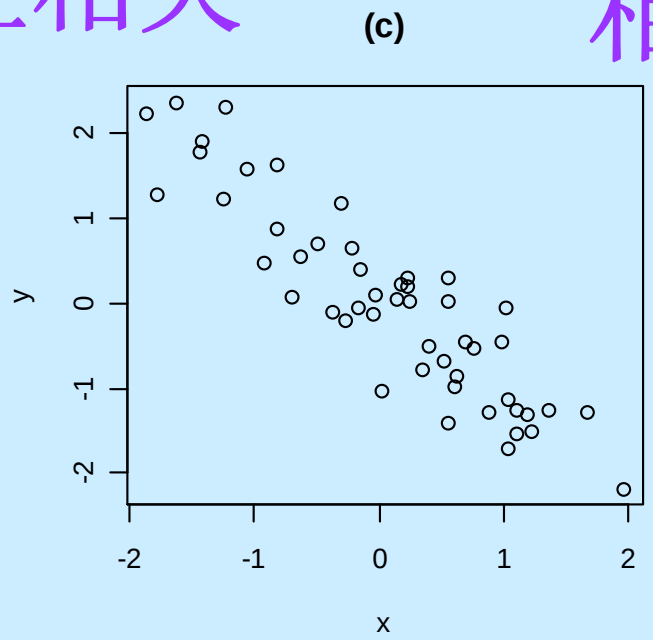
不相关



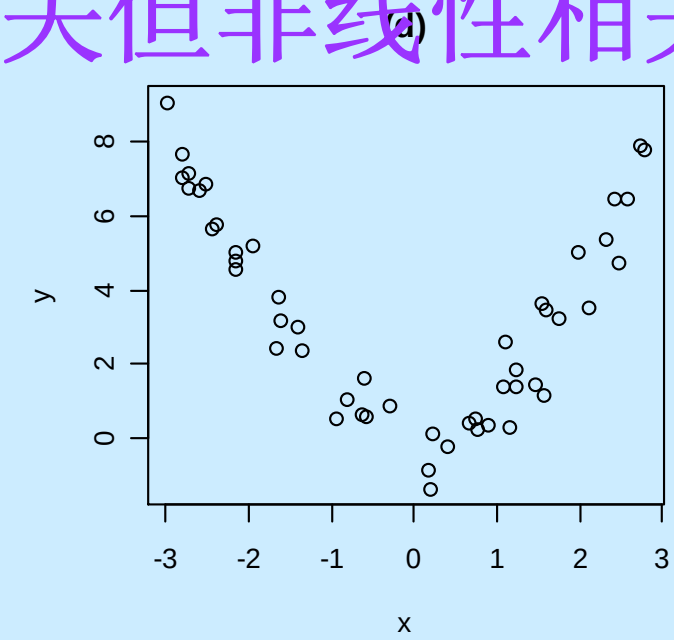
正线性相关



负线性相关



相关但非线性相关



2 定量变量的相关

- 但如何在数量上描述相关呢？下面引进几种对相关程度的度量。
- **Pearson相关系数（Pearson's correlation coefficient）**又叫相关系数或线性相关系数。它一般用字母*r*表示。它是由两个变量的样本取值得到，这是一个描述线性相关强度的量，取值于-1和1之间。当两个变量有很强的线性相关时，相关系数接近于1（正相关）或-1（负相关），而当两个变量不那么线性相关时，相关系数就接近0。

§2 定量变量的相关

- **Kendall τ 相关系数（Kendall's τ）**这里的度量原理是把所有的样本点配对（如果每一个点由x和y组成的坐标(x,y)代表，一对点就是诸如(x_1,y_1)和(x_2,y_2)的点对），然后看每一对中的x和y的观测值是否同时增加（或减少）。比如由点对(x_1,y_1)和(x_2,y_2)，可以算出乘积$(x_2-x_1)(y_2-y_1)$是否大于0；如果大于0，则说明x和y同时增长或同时下降，称这两点**协同（concordant）**；否则就是不协同。如果样本中协同的点数目多，两个变量就更加相关一些；如果样本中不协同（discordant）的点数目多，两个变量就不很相关。

§2 定量变量的相关

- **Spearman 秩相关系数（Spearman rank correlation coefficient 或Spearman's ρ）**它和Pearson相关系数定义有些类似，只不过在定义中把点的坐标换成各自样本的秩（即样本点大小的“座次”）。Spearman相关系数也是取值在-1和1之间，也有类似的解释。通过它也可以进行不依赖于总体分布的非参数检验。

§2 定量变量的相关

- 人们可能会问，上面的三种对相关的度量都是在其值接近1或-1时相关，而接近于0时不相关。到底如何才能够称为“接近”呢？
- 这很难一概而论。但在计算机输出中都有和这些相关度量相应的检验和*p*-值；因此可以根据这些结果来判断是否相关(见下面例7.1的继续)。

§2 定量变量的相关

- 例1（继续）得到初三和高一成绩的**Pearson**相关系数，**Kendall τ** 相关系数和**Spearman** 秩相关系数分别为**0.795, 0.595**和**0.758**。
- 这三个统计量相关的检验(零假设均为不相关)全部显著，***p*-值**都是**0.000**。注意这种**0.000**的表示并不表示这些***p*-值**恰好等于零，只是小数点前三位是**0**而已。

SPSS的相关分析

- 相关分析(hischool.sav）
- 利用SPSS选项： Analize－Correlate－Bivariate
- 再把两个有关的变量(这里为j3和s1)选入，选择Pearson，Spearman和Kendall就可以得出这三个相关系数和有关的检验结果了(零假设均为不相关)。
- 偏相关分析：参考SPSS实用统计分析书

第九章 回归分析和Logistic回归分析

介绍：

1、回归分析的概念和模型

2、回归分析的过程

回归分析的概念

- 寻求有关联（相关）的变量之间的关系
- 主要内容：
  - 从一组样本数据出发，确定这些变量间的定量关系式
  - 对这些关系式的可信度进行各种统计检验
  - 从影响某一变量的诸多变量中，判断哪些变量的影响显著，哪些不显著
  - 利用求得的关系式进行预测和控制

回归分析的模型

- 按是否线性分：线性回归模型和非线性回归模型
- 按自变量个数分：简单的一元回归，多元回归
- 基本的步骤：利用SPSS得到模型关系式，是否是我们所要的，要看回归方程的显著性检验（F检验）和回归系数b的显著性检验(T检验)，还要看拟合程度R^2(相关系数的平方,一元回归用R Square，多元回归用Adjusted R Square)

回归分析的过程

- 在回归过程中包括：
  - Liner：线性回归
  - Curve Estimation：曲线估计
  - Binary Logistic： 二分变量逻辑回归
  - Multinomial Logistic：多分变量逻辑回归
  - Ordinal 序回归
  - Probit：概率单位回归
  - Nonlinear：非线性回归
  - Weight Estimation：加权估计
  - 2-Stage Least squares：二段最小平方法
  - Optimal Scaling 最优编码回归
- 我们只讲前面3个简单的。

线性回归(Liner)

- 一元线性回归方程: $y=a+bx$
  - a称为截距
  - b为回归直线的斜率
  - 用R^2判定系数判定一个线性回归直线的拟合程度：用来说明用自变量解释因变量变异的程度（所占比例）
- 多元线性回归方程: $y=b_0+b_1x_1+b_2x_2+\ldots+b_nx_n$
  - b_0为常数项
  - b_1、b_2、…、b_n称为y对应于x_1、x_2、…、x_n的偏回归系数
  - 用Adjusted R^2调整判定系数判定一个多元线性回归方程的拟合程度：用来说明用自变量解释因变量变异的程度（所占比例）
- 一元线性回归模型的确定:一般先做散点图(Graphs ->Scatter->Simple),以便进行简单地观测（如：Salary与Salbegin的关系)
- 若散点图的趋势大概呈线性关系，可以建立线性方程，若不呈线性分布，可建立其它方程模型，并比较R^2 (-->1)来确定一种最佳方程式（曲线估计）
- 多元线性回归一般采用逐步回归方法-Stepwise

逐步回归方法的基本思想

- 对全部的自变量$x_1,x_2,\ldots,x_p$,按它们对Y贡献的大小进行比较，并通过F检验法，选择偏回归平方和显著的变量进入回归方程，每一步只引入一个变量，同时建立一个偏回归方程。当一个变量被引入后，对原已引入回归方程的变量，逐个检验他们的偏回归平方和。如果由于引入新的变量而使得已进入方程的变量变为不显著时，则及时从偏回归方程中剔除。在引入了两个自变量以后，便开始考虑是否有需要剔除的变量。只有当回归方程中的所有自变量对Y都有显著影响而不需要剔除时，在考虑从未选入方程的自变量中，挑选对Y有显著影响的新的变量进入方程。不论引入还是剔除一个变量都称为一步。不断重复这一过程，直至无法剔除已引入的变量，也无法再引入新的自变量时，逐步回归过程结束。

线性回归分析实例

- 实例：Data02-01建立一个以初始工资Salbegin 、工作经验prevexp 、工作时间jobtime 、工作种类jobcat 、受教育年限edcu等为自变量，当前工资Salary为因变量的回归模型。

1. 先做数据散点图,观测因变量Salary与自变量Salbegin之间关系是否有线性特点
   - Graphs ->Scatter->Simple
   - X Axis： Salbegin
   - Y Axis： Salary
2. 若散点图的趋势大概呈线性关系，可以建立线性回归模型
   - Analyze->Regression->Linear
   - Dependent: Salary
   - Independents: Salbegin,prevexp,jobtime,jobcat,edcu等变量
   - Method: Stepwise
   - 比较有用的结果：
     - 拟合程度Adjusted R^2：越接近1拟合程度越好
     - 回归方程的显著性检验Sig
     - 回归系数表Coefficients的Model最后一个中的回归系数B和显著性检验Sig
     - 得模型： Salary=-15038.6+1.37Salbegin+5859.59jobcat-19.55prevexp+154.698jobtime+539.64edcu

曲线估计(Curve Estimation)

- 对于一元回归，若散点图的趋势不呈线性分布，可以利用曲线估计方便地进行线性拟合(liner)、二次拟合(Quadratic)、三次拟合(Cubic)等。采用哪种拟合方式主要取决于各种拟合模型对数据的充分描述(看修正 Adjusted R^2 -->1)

不同模型的表示		
模型名称	回归方程	相应的线性回归方程
Linear(线性)	$Y=b_0+b_1t$	
Quadratic(二次)	$Y=b_0+b_1t+b_2t^2$	
Compound(复合)	$Y=b_0(b_1^t)$	$Ln(Y)=ln(b_0)+ln(b_1)t$
Growth(生长)	$Y=e^{b0+b1t}$	$Ln(Y)=b_0+b_1t$
Logarithmic(对数)	$Y=b_0+b_1ln(t)$	
Cubic(三次)	$Y=b_0+b_1t+b_2t^2+b_3t^3$	
S	$Y=e^{b0+b1/t}$	$Ln(Y)=b_0+b_1/t$
Exponential(指数)	$Y=b_0*e^{b1*t}$	$Ln(Y)=ln(b_0)+b_1t$
Inverse(逆)	$Y=b_0+b_1/t$	
Power(幂)	$Y=b_0(t^{b1})$	$Ln(Y)=ln(b_0)+b_1ln(t)$
Logistic(逻辑)	$Y=1/(1/u+b_0b_1^t)$	$Ln(1/Y-1/u)=ln(b_0+ln(b_1)t)$

曲线估计(Curve Estimation)分析实例

- 实例：

王淑芬P177例6.9

- 建立若干曲线模型（可试着选用所有模型Models)
  - Analyze->Regression-> Curve Estimation
  - Dependent:
  - Independent:
  - Models: 全选(除了最后一个逻辑回归)
  - 选Plot models：输出模型图形
  - 比较有用的结果：各种模型的Adjusted R^2，并比较哪个大，哪个模型的拟合效果就较好。

二项逻辑回归(Binary Logistic)

- 在现实中，经常需要判断一些事情是否将要发生，候选人是否会当选？为什么一些人易患冠心病？为什么一些人的生意会获得成功？此问题的特点是因变量只有两个值，不发生(0)和发生(1)。这就要求建立的模型必须因变量的取值范围在0～1之间。
- Logistic回归模型
  - Logistic模型：在逻辑回归中，可以直接预测观测量相对于某一事件的发生概率。包含一个自变量的回归模型和多个自变量的回归模型公式：

$$prob(event) = \frac{1}{1+e^{-z}}$$

其中： $z=B_0+B_1X_1+\ldots B_pX_p$(P为自变量个数）。某一事件不发生的概率为Prob(no event)＝1-Prob(event) 。因此最主要的是求$B_0,B_1,\ldots B_p$(常数和系数)

  - 数据要求：因变量应具有二分特点。自变量可以是分类变量和定距变量。如果自变量是分类变量应为二分变量或被重新编码为指示变量。
  - 回归系数：几率和概率的区别。几率=发生的概率/不发生的概率。如从52张桥牌中抽出一张A的几率为(4/52)/(48/52)=1/12，而其概率值为4/52=1/13

根据回归系数表，可以写出回归模型公式中的z。然后根据回归模型公式Prob(event)进行预测。

Logistic 回归基本概念

- 线性回归模型的一个局限性是要求因变量是定量变量（定距变量、定比变量）而不能是定性变量（定序变量、定类变量）。但是在许多实际问题中，经常出现因变量是定性变量（分类变量）的情况。可用于处理分类因变量的统计分析方法有：判别分别(Discriminant analysis)、Probit 分析、Logistic 回归分析和对数线性模型等。在社会科学中，应用最多的是Logistic回归分析。Logistic 回归分析根据因变量取值类别不同，又可以分为Binary Logistic 回归分析和Multinomial Logistic 回归分析，Binary Logistic 回归模型中因变量只能取两个值1 和0（虚拟因变量），而Multinomial Logistic 回归模型中因变量可以取多个值。

- 我们可以采用多种方法对取值为0、1 的因变量进行分析。通常以p表示事件发生的概率（事件未发生的概率为1-p），并把p 看作自变量Xi 的线性函数，即
- $$p = P(y = 1) = F(\beta_i X_i) \qquad i = 1, 2, \cdots, k$$
- 不同形式的F(·)，就有不同形式的模型，最简单的莫过于使F(·)为一线性函数，即
- $$p = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon \qquad (1)$$
- 我们可能会认为可用普通最小二乘法对上式进行估计，但因p 的值一定在区间[0,1] 内，而且当p 接近于0 或1 时，自变量即使有很大变化，p 的值也不可能变化很大，所以对上式直接用普通最小二乘法进行估计是行不通的。

- 我们引入*p* 的Logistic 变换（或称为*p* 的Logit 变换），
- 即

$$\theta(p) = \log it(p) = \ln(\frac{p}{1-p})$$

- 其中，*p*/(1-*p*)；logit(*p*)是因变量Y=1 的差异比(odds ratio)或似然比(likelihood ratio)的自然对
- 数，称为对数差异比(log odds ratio)、对数似然比(log likelihood ratio)或分对数。

用 θ (*p*) 代替式(1)中的*p*，得

$$\theta(p)=\ln\left(\frac{p}{1-p}\right)=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k+\varepsilon \qquad (2)$$

将 p 由 θ 来表示，得

$$p=\frac{e^{\theta}}{1+e^{\theta}}=\frac{e^{\beta_0+\beta_1X_1+\cdots+\beta_kX_k+\varepsilon}}{1+e^{\beta_0+\beta_1X_1+\cdots+\beta_kX_k+\varepsilon}}$$

可见，当*p* 从0 变化1 时， θ (*p*) 从-∞变到+∞

二项Logistic回归方程中回归系数的含义

- 1. odds ratio定义：见前
- 2.回归系数的含义：

$$Logit(p) = Ln(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k$$

- 当其他解释变量保持不变时， x_i 每增加一个单位将引起发生比扩大 e^{β_i} 倍，当回归系数为负时发生比缩小。

Logistic 回归模型的估计与检验

- 对于模型(2)，采用最大似然估计法(Maximum likelihood estimation，MLE)进行估计，它与用于估计一般线性回归模型参数的普通最小二乘法（OLS）形成对比。OLS 通过使得样本观测数据的残差平方和最小来选择参数，而最大似然估计法通过最大化对数似然值(log likelihood)估计参数。最大似然估计法是一种迭代算法，它以一个预测估计值作为参数的初始值，根据算法确定能增大对数似然值的参数的方向和变动。估计了该初始函数后，对残差进行检验并用改进的函数进行重新估计，直到收敛为止（即对数似然不再显著变化）。

一些常用的检验统计量

- **1）-2 对数似然值(-2 log likelihood，-2LL)**

  似然(likelihood)即概率，特别是由自变量观测值预测因变量观测值的概率。与任何概率一样，似然的取值范围在0、1 之间。对数似然值(log likelihood，LL)是它的自然对数形式，由于取值范围在[0,1]之间的数的对数值负数，所以对数似然值的取值范围在0 至-∞之间。对数似然值通过最大似然估计的迭代算法计算而得。因为-2LL 近似服从卡方分布且在数学上更为方便，所以-2LL 可用于检验Logistic 回归的显著性。-2LL 反映了在模型中包括了所有自变量后的误差，用于处理因变量无法解释的变动部分的显著性问题，又称为拟合劣度卡方统计量(Badness-of-fit Chi-square)。当-2LL 的实际显著性水平大于给定的显著性水平α时，因变量的变动中无法解释的部分是不显著的，意味着回归方程的拟合程度越好。

- -2LL 的计算公式为：

$$-2LL = -2\sum_{i=1}^{n}(y_i \ln(\hat{p}_i) + (1-y_i)\ln(1-\hat{p}_i))$$

其中，y_i 为各样本观测值，$\hat{p}_i$ 为第 i 个样本的预测概率，即

$$\hat{p}_i = \exp(\hat{\theta}_i)/[1+\exp(\hat{\theta}_i)]$$

$$= \exp(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik})/[1+\exp(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik})]。$$

- 2)拟合优度(Goodness of Fit)统计量

Logistic 回归的拟合优度统计量计算公式为：

$$\sum_{i=1}^{n} \frac{(y_i - \hat{p}_i)^2}{\hat{p}_i(1-\hat{p}_i)}$$

- 实际应用中，经常采用分类表(Classification Table)来反映拟合效果。另外，还经常采用分类图(Classification Plot)来反映拟合效果。它也可以用于发现奇异值，也可以用于判断是否应按其它规则来划分两个预测组。

- 3)Cox 和 Snell 的 （Cox & Snell’s R-Square）：

Cox 和 Snell 的 R^2试图在似然值基础上模仿线性回归模型的 R^2解释Logistic 回归模型，但它的最大值一般小于1，解释时有困难。其计算公式为：

$$R_{CS}^2 = 1 - \left(\frac{l(0)}{l(\hat{\beta})} \right)^2$$

其中，$l(\hat{\beta})$ 表示当前模型的似然值(likelihood)，$l(0)$ 表示初始模型的似然值。

4）Nagelkerke 的R^2（Nagelkerke's R-Square）

- 为了对Cox 和 Snell 的R^2 进一步调整，使得取值范围在0 和1 之间，Nagelkerke 把Cox和 Snell 的R^2 除以它的最大值，即：

$$R_N^2 = R_{CS}^2 / \max(R_{CS}^2)$$

这里的 $\max(R_{CS}^2) = 1 - [L(0)]^2$

- 5）Hosmer 和 Lemeshow 的拟合优度检验统计量（Hosmer and Lemeshow's Goodness of Fit Test Statistic）
- 与一般拟合优度检验不同，Hosmer 和 Lemeshow 的拟合优度检验通常把样本数据根据预测概率分为十组，然后根据观测频数和期望频数构造卡方统计量（即Hosmer 和 Lemeshow 的拟合优度检验统计量，简称H-L 拟合优度检验统计量），最后根据自由度为8的卡方分布计算其p 值并对 Logistic 模型进行检验。如果该p 值小于给定的显著性水平α（如α=0.05），则拒绝因变量的观测值与模型预测值不存在差异的零假设，表明模型的预测值与观测值存在显著差异。如果p 值大于α，我们没有充分的理由拒绝零假设，表明在可接受的水平上模型的估计拟合了数据。

- 6)Wald 统计量

同线性回归方程的参数显著性检验类似，Wald 统计量用于判断一个变量是否应该包含在模型中，检验步骤为：

（a）提出假设：

$$H_0 : \beta_i = 0 \quad (i = 1,2,\cdots,k)$$

$$H_1 : \beta_i \neq 0$$

（b）构造检验统计量——Wald 统计量

如果自变量Xi 不是分类变量，那么Wald 统计量为：

$$Wald_i = \frac{b_i^2}{Var(b_i)}$$

如果自变量Xi 是分类变量，Wald 统计量的公式较繁，这里从略。

Wald 统计量近似服从于自由度等于参数个数的卡方分布。

（c）作出统计判断。

Logistic回归的SPSS实现

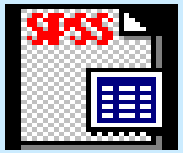
- 二分类Logistic回归的SPSS实现：
- （1）在主菜单中选择[Analyze]=>[Regression]=>[Binary Logistic]
- （2）在[Logistic Regression]对话框中，选择Y 进入[Dependent]框作为因变量，选择X 进入[Covariates]作为自变量。单击[Method]的下拉菜单，SPSS 提供了7 种方法：
- [Enter]：所有自变量强制进入回归方程；
- [Forward: Conditional]：以假定参数为基础作似然比检验，向前逐步选择自变量；
- [Forward: LR]：以最大局部似然为基础作似然比检验，向前逐步选择自变量；
- [Forward: Wald]：作Wald 概率统计法，向前逐步选择自变量；
- [Backward: Conditional]：以假定参数为基础作似然比检验，向后逐步选择自变量；
- [Backward: LR]：以最大局部似然为基础作似然比检验，向后逐步选择自变量；
- [Backward: Wald]：作Wald 概率统计法，向后逐步选择自变量。

- （3）单击[Logistic Regression]对话框中的[Options]按钮，在显示的子对话框中选择[Classification plots]和[Hosmer-Lemeshow goodness-of-fit]等选项，并单击[Continue]返回主对话框。
- （4）单击主对话框中[OK]按钮，输出结果。

二项逻辑回归(Binary Logistic)实例

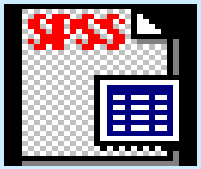
- 实例

薛薇相关回归分析（消费行为logistic回归）.sav



割草机logistic回归例子.sav

多分变量Logistic回归分析实例



data09-03.sav

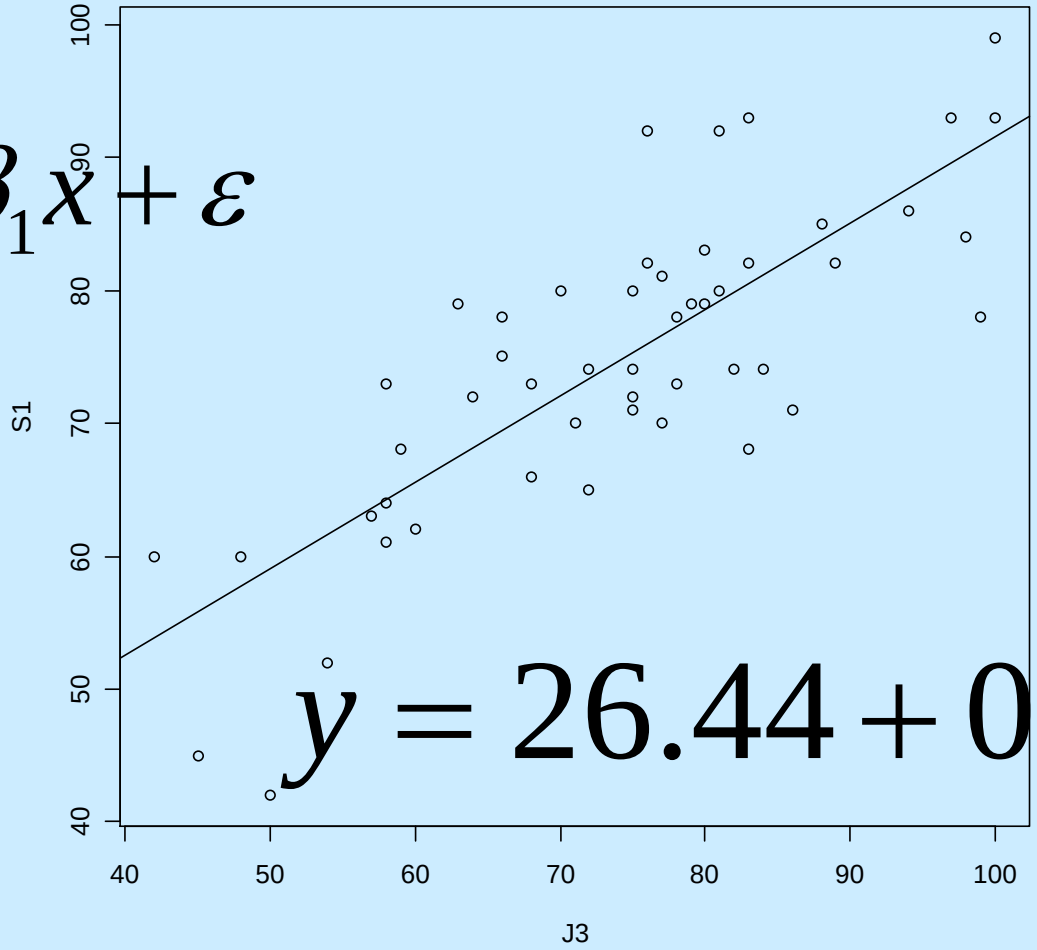
补充：回归分析

- 以下的讲义是吴喜之教授有关回归分析的讲义，很简单，但很实用

定量变量的线性回归分析

- 对例**1(highschoo.sav)**的两个变量的数据进行线性回归，就是要找到一条直线来最好地代表散点图中的那些点。

$$y = \beta_0 + \beta_1 x + \varepsilon$$



检验问题等

- 对于系数$\beta_1=0$的检验
- 对于拟合的F检验
- R^2(决定系数)及修正的R^2.

多个自变量的回归

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

如何解释拟合直线?

什么是逐步回归方法?

自变量中有定性变量的回归

- 例**1(highschoo.sav)**的数据中,还有一个自变量是定性变量“收入”,以虚拟变量或哑元**(dummy variable)**的方式出现;这里收入的“低”,“中”,“高”，用**1,2,3**来代表.所以,如果要用这种哑元进行前面回归就没有道理了.
- 以例**1**数据为例,可以用下面的模型来描述:

$$y=\beta_0+\beta_1 x+\alpha_1+\varepsilon,\quad \text{代表家庭收入的哑元=1时,}$$
$$=\beta_0+\beta_1 x+\alpha_2+\varepsilon,\quad \text{代表家庭收入的哑元=2时,}$$
$$=\beta_0+\beta_1 x+\alpha_3+\varepsilon,\quad \text{代表家庭收入的哑元=3时。}$$

自变量中有定性变量的回归

- 现在只要估计β_0, β_1,和α_1, α_2, α_3即可。
- 哑元的各个参数α_1, α_2, α_3本身只有相对意义，无法三个都估计，只能够在有约束条件下才能够得到估计。
- 约束条件可以有很多选择，一种默认的条件是把一个参数设为**0**，比如α_3**=0**，这样和它有相对意义的α_1和α_2就可以估计出来了。
- 对于例**1**，对β_0, β_1, α_1, α_2, α_3的估计分别为**28.708, 0.688, -11.066, -4.679, 0**。这时的拟合直线有三条，对三种家庭收入各有一条：

$$y = 28.708 + 0.688x - 11.066, \quad \text{（低收入家庭），}$$

$$y = 28.708 + 0.688x - 4.679, \quad \text{（中等收入家庭），}$$

$$y = 28.708 + 0.688x, \quad \text{（高收入家庭）。}$$

SPSS实现(hischool.sav)

- **Analize － General linear model － Univariate，**
- **在Options中选择Parameter Estimates，**
- **再在主对话框中把因变量（s1）选入Dependent Variable，把定量自变量(j3)选入Covariate，把定量因变量（income）选入Factor中。**
- **然后再点击Model，在Specify Model中选Custom，**
- **再把两个有关的自变量选入右边，再在下面Building Term中选Main effect。**
- **Continue-OK，就得到结果了。输出的结果有回归系数和一些检验结果。**

注意

- 这里进行的线性回归，仅仅是回归的一种，也是历史最悠久的一种。
- 但是，任何模型都是某种近似；
- 线性回归当然也不另外。
- 它被长期广泛深入地研究主要是因为数学上相对简单。
- 它已经成为其他回归的一个基础。
- 总应该用批判的眼光看这些模型。

SPSS的回归分析

- 自变量和因变量都是定量变量时的线性回归分析：
  - 菜 单 ： **Analize － Regression － Linear**
  - 把有关的自变量选入**Independent**，把因变量选入**Dependent**，然后**OK**即可。如果自变量有多个（多元回归模型，选Method: Stepwise ），只要都选入就行。

SPSS的回归分析

- 自变量中有定性变量（哑元）和定量变量而因变量为定量变量时的线性回归分析（**hischool.sav**)
  - 菜单：**Analize－General linear model－Univariate**，
  - 在**Options**中选择**Parameter Estimates**，
  - 再在主对话框中把因变量（**s1**）选入**Dependent Variable**，把定量自变量**(j3)**选入**Covariate**，把定性因变量（**income**）选入**Factor**中。
  - 点击**Model**，在**Specify Model**中选**Custom**，再把两个有关的自变量选入右边，再在下面**Building Term**中选**Main effect**。然后就**Continue-OK**。