

第 6 章 统计分析建模方法

一、复习题

1、主成分分析法（方法 1-基于协方差矩阵方法）

- 由 X 的协方差阵 Σ_X ，求出其特征根；
- 求出分别所对应的特征向量；
- 计算累积贡献率，给出恰当的主成分个数；
- 计算所选出的 k 个主成分的得分。将原始数据的中心化值。

2、主成分分析法（方法 2-基于相关系数矩阵方法）

- 计算相关系数矩阵；
- 计算特征值与特征向量；
- 计算主成分贡献率及累计贡献率；
- 计算主成分载荷；
- 各主成分的得分。

3、聚类分析法（方法 1-直接聚类法）

原理：先把各个分类对象单独视为一类，然后根据距离最小的原则，依次选出一对分类对象，并成新类。

- 性质 1：如果其中一个分类对象已归于一类，则把另一个也归入该类；
- 性质 2：如果一对分类对象正好属于已归的两类，则把这两类并为一类。

每一次归并，都划去该对象所在的列与列序相同的行。经过 $n-1$ 次就可以把全部分类对象归为一类，根据归并的先后顺序作出聚类谱系图。

4、聚类分析法（方法 2- k-means 聚类算法）

- 选择一个含有随机选择样本的 k 个簇的初始划分，计算这些簇的质心。
- 根据距离把剩余的每个样本分配到距离它最近的簇质心的一个划分。
- 计算被分配到每个簇的样本的均值向量，作为新的簇的质心。
- 重复 2,3 步，直到 k 个簇的质心点不再发生变化或准则函数收敛。
- **注：聚类分析中要考虑相似性度量（ L_k 范数）和数据处理方法（4 种方法）。**

二、主成分分析方法（PCA）

1.基础概念

- 多变量问题是经常会遇到的。变量太多，无疑会增加分析问题的难度与复杂性。
- 问题：能否用较少的新变量尽可能多地保留原来较多的变量所反映的信息？
- 事实上，这种想法是可以实现的。因为在许多实际问题中，多个变量之间是具有一定的相关关系的。
- 主成分分析原理：是把原来多个变量化为少数几个综合指标的一种统计分析方法，从数学角度来看，这是一种降维处理技术。
- 主成分分析方法就是综合处理这种问题的一种强有力的方法。
- 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) 如何作主成分分析？

- 当分析中所选择的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) 如何选择几个主成分。

- 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

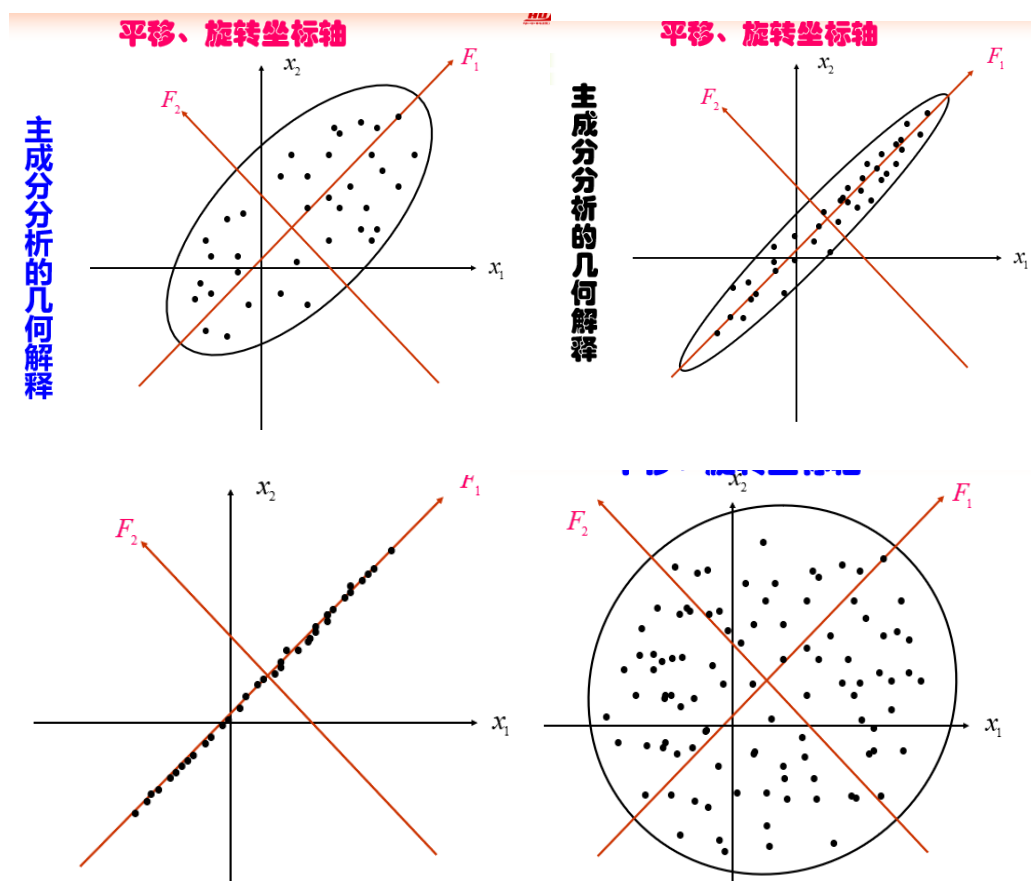
数。关于保留几个主成分，应该权衡主成分个数和保留的信息。

(3) 如何解释主成分所包含的几何意义或经济意义或其它。

主成分分析就是试图在力保数据信息丢失最少的原则下，对这种多变量的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。

2.几何解释

先假定数据只有二维，即只有两个变量，它们由横坐标和纵坐标所代表；因此每个观测值都有相应于这两个坐标轴的两个坐标值；如果这些数据形成一个椭圆形状的点阵（这在变量的二维正态的假定下是可能的）



PCA: 进一步解释

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，短轴如果退化成一点，那只有在长轴的方向才能够解释这些点的变化了；这样，由二维到一维的降维就自然完成了。

- 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。
- 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（球）的长短轴相差越大，降维也越有道理。
- 对于多维变量的情况和二维类似，也有高维的椭球，只不过无法直观地看见罢了。
- 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

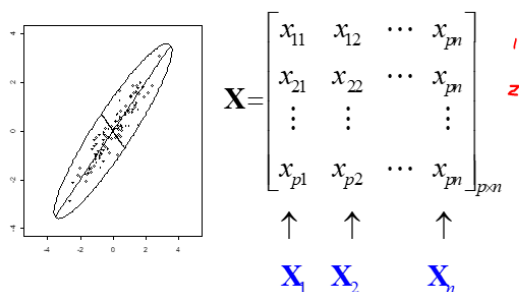
量；这样，主成分分析就基本完成了。

- 注意，和二维情况类似，高维椭圆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)。
- 正如二维椭圆有两个主轴，三维椭圆有三个主轴一样，有几个变量，就有几个主成分。
- 选择越少的主成分，降维就越好。
- 什么是标准呢？
- ——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有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。

3. 均值和协方差 特征值和特征向量

设有 n 个样本，每个样本观测 p 个指标（变量）：

X_1, X_2, \dots, X_n ，得到原始数据矩阵：



1. 样本均值

$$M = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

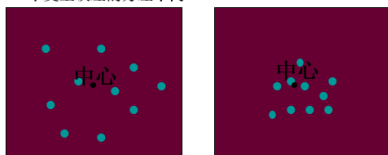
显然，样本均值是数据散列图的**中心**。

$$\bar{X}_k = X_k - M$$

于是 $p \times n$ 矩阵的列 B 具有零样本均值，称为平均偏差形式

$$B = [\bar{X}_1, \bar{X}_2, \dots, \bar{X}_n]$$

协方差表示的是两个变量的总体的误差，这与只表示一个变量误差的方差不同。



2. 样本协方差

$$S = \frac{1}{n-1}BB^T$$

注意：协方差是对称矩阵且半正定

协方差的大小在一定程度上反映了多变量之间的关系，但它还受变量自身度量单位的影响。



特征值与特征向量

定义 A 为 n 阶方阵， λ 为数， X 为 n 维非零向量，若

$$AX = \lambda X$$

则 λ 称为 A 的**特征值**， X 称为 A 的**特征向量**。

注 ① 特征向量 $X \neq 0$ ，特征值问题只针对于方阵；

② λ, X 并不一定唯一；

③ n 阶方阵 A 的特征值，就是使齐次线性方程组 $(\lambda I - A)x = 0$ 有非零解的 λ 值，即满足 $|\lambda I - A| = 0$ 的 λ 都是方阵 A 的特征值。

定义 称以 λ 为未知数的一元 n 次方程 $|\lambda I - A| = 0$ 为 A 的**特征方程**。

例1:

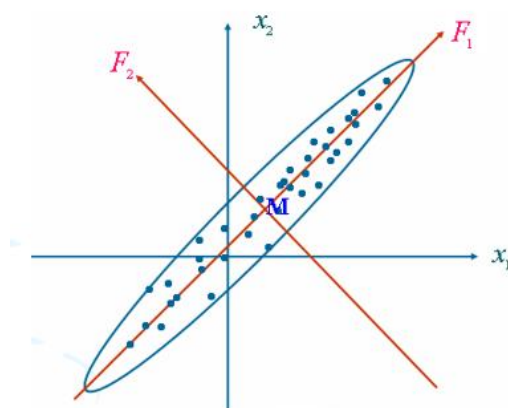
从一个总体中随机抽取4个样本作三次测量,每一个样本的观测向量为:

$$X_1 = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, X_2 = \begin{bmatrix} 4 \\ 2 \\ 13 \end{bmatrix}, X_3 = \begin{bmatrix} 7 \\ 8 \\ 1 \end{bmatrix}, X_4 = \begin{bmatrix} 8 \\ 4 \\ 5 \end{bmatrix}$$

计算样本均值 M 和协方差矩阵 S 以及 S 的特征值和特征向量。

$$M = \frac{1}{n} \sum_{i=1}^n X_i \quad S = \frac{1}{n-1} BB^T \quad SX = \lambda X$$

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 个样本，每个样本有两个观测变量 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信息将会有较大的损失。



如果我们将 x_1 轴和 x_2 轴先平移, 再同时按逆时针方向旋转 θ 角度, 得到新坐标轴 F_1 和 F_2 。 F_1 和 F_2 是两个新变量。

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。

4 主成分分析

基本原理:

假定有 n 个数据样本, 每个样本共有 p 个变量, 构成一个 $n \times p$ 阶的数据矩阵

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

当 p 较大时, 在 p 维空间中考察问题比较麻烦。为了克服这一困难, 就需要进行降维处理。——用较少的几个综合指标代替原来较多的变量指标, 而且使这些较少的综合指标既能尽量多地反映原来较多变量指标所反映的信息, 同时它们之间又是彼此独立的。

定义: 记 x_1, x_2, \dots, x_p 为原变量指标, z_1, z_2, \dots, z_m ($m \leq p$) 为新变量指标

$$\begin{cases} z_1 = l_{11}x_1 + l_{12}x_2 + \cdots + l_{1p}x_p \\ z_2 = l_{21}x_1 + l_{22}x_2 + \cdots + l_{2p}x_p \\ \dots\dots\dots \\ z_m = l_{m1}x_1 + l_{m2}x_2 + \cdots + l_{mp}x_p \end{cases}$$

系数 l_{ij} 的确定原则:

- ① z_i 与 z_j ($i \neq j$; $i, j=1, 2, \dots, m$) 相互无关;
- ② z_1 是 x_1, x_2, \dots, x_p 的一切线性组合中方差最大者, z_2 是与 z_1 不相关的 x_1, x_2, \dots, x_p 的所有线性组合中方差最大者; z_m 是与 z_1, z_2, \dots, z_{m-1} 都不相关的 x_1, x_2, \dots, x_p 的所有线性组合中方差最大者。|
则新变量指标 z_1, z_2, \dots, z_m 分别称为原变量指标 x_1, x_2, \dots, x_p 的第1, 第2, ..., 第 m 主成分。

主成分分析的实质就是确定原来变量 $x_j (j=1, 2, \dots, p)$ 在诸主成分 $z_i (i=1, 2, \dots, m)$ 上的荷载 $l_{ij} (i=1, 2, \dots, m; j=1, 2, \dots, p)$ 。

从数学上可以证明，它们分别是相关矩阵 m 个较大的特征值所对应的特征向量。

PCA 的性质：

1 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

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

2、若上述矩阵的特征根所对应的单位特征向量为 u_1, \dots, u_p

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

其中 $\lambda_i, i=1, 2, \dots, p$ 是 A 的特征根。

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

3、均值 $E(U^T x) = U^T M$

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。

4、方差为所有特征根之和

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

协方差矩阵 Σ 的对角线上的元素之和等于特征根之和。

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

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

来描述，称为累积贡献率。

PCA 常用统计量：

• 1. 特征根： λ_i

• 2. 各成分贡献率： $\frac{\lambda_i}{\sum \lambda_i}$

• 3. 前各成分累计贡献率： $\sum_{i=1}^k \lambda_i / \sum_{i=1}^p \lambda_i$

• 4. 特征向量：各成分表达式中标准化原始变量的系数向量——各成分的特征向量。

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 $F_1, F_2, \dots, F_k (k \leq p)$ 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 $\geq 80\%$ 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。

例 设 x_1, x_2, x_3 的协方差矩阵为

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

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

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

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，但应该取两个主成分。97.88%

主成分分析的步骤

一、基于协方差矩阵方法

$$\mathbf{X}_l = (x_{1l}, x_{2l}, \dots, x_{pl})' \quad (l = 1, 2, \dots, n)$$

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

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

第二步：求出分别所对应的特征向量 U_1, U_2, \dots, U_p ，

$$\mathbf{U}_i = (u_{1i}, u_{2i}, \dots, u_{pi})^T$$

第三步：计算累积贡献率，给出恰当的主成分个数。

$$F_i = \mathbf{U}_i^T \mathbf{X}, \quad i = 1, 2, \dots, k \quad (k \leq p)$$

第四步：计算所选出的 k 个主成分的得分。将原始数据中心化值：

$$\mathbf{X}_i^* = \mathbf{X}_i - \bar{\mathbf{X}} = (x_{1i} - \bar{x}_1, x_{2i} - \bar{x}_2, \dots, x_{pi} - \bar{x}_p)^T$$

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。

二、基于相关系数矩阵方法

(一) 计算相关系数矩阵

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix}$$

r_{ij} ($i, j=1, 2, \dots, p$) 为原变量 x_i 与 x_j 的相关系数, $r_{ij}=r_{ji}$, 其计算公式为

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

(二) 计算特征值与特征向量

① 解特征方程 $|\lambda I - R| = 0$, 常用雅可比法 (Jacobi) 求出特征值, 并使其按大小顺序排列 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$;

② 分别求出对应于特征值 λ_i 的特征向量 e_i ($i=1, 2, \dots, p$)

要求 $\|e_i\| = 1$, 即 $\sum_{j=1}^p e_{ij}^2 = 1$, 其中 e_{ij} 表示向量 \mathbf{i} 的第 j 个分量。

③ 计算主成分贡献率及累计贡献率

✓ 贡献率
$$\frac{\lambda_i}{\sum_{k=1}^p \lambda_k} \quad (i=1, 2, \dots, p)$$

✓ 累计贡献率
$$\frac{\sum_{k=1}^i \lambda_k}{\sum_{k=1}^p \lambda_k} \quad (i=1, 2, \dots, p)$$

一般取累计贡献率达85%~95%的特征值 $\lambda_1, \lambda_2, \dots, \lambda_m$ 所对应的第1、第2、...、第 m ($m \leq p$) 个主成分。

④ 计算主成分载荷

$$l_{ij} = p(z_i, x_j) = \sqrt{\lambda_i} e_{ij} (i, j = 1, 2, \dots, p)$$

⑤ 各主成分的得分

$$Z = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1m} \\ z_{21} & z_{22} & \cdots & z_{2m} \\ \vdots & \vdots & & \vdots \\ z_{m1} & z_{m2} & \cdots & z_{mm} \end{bmatrix}$$

6 主成分载荷阵中各列元素的平方和 $S_j = \sum_i l_{ij}^2$ 称为公共因子 F_j 对 X 诸变量的方差贡献之总和

例题：

例：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;

计算协方差矩阵:

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

结果:


17.4677	23.8520	-9.9489	-20.6757	-9.5172
23.8520	121.7307	-87.7656	-68.0747	-13.8078
-9.9489	-87.7656	110.3156	52.9433	26.8544
-20.6757	-68.0747	52.9433	55.7677	23.0128
-9.5172	-13.8078	26.8544	23.0128	41.5361

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
PRIN1	253.6856	210.2479	0.731467	0.731467
PRIN2	43.4377	6.9483	0.125247	0.856713
PRIN3	36.4894	29.1793	0.105212	0.961926
PRIN4	7.3101	1.4153	0.021078	0.983003
PRIN5	5.8948	0	0.016997	1

Eigenvectors

	PRIN1	PRIN2	PRIN3	PRIN4	PRIN5
X1	0.1343	0.0143	0.4766	-0.3259	0.8053
X2	0.6526	-0.4857	0.1822	0.5521	0.0155
X3	-0.5945	-0.2363	0.6916	0.2721	-0.1958
X4	-0.4172	-0.0789	-0.4457	0.5551	0.5594
X5	-0.169	-0.8378	-0.2506	-0.4547	0.0074

 第一主成份的贡献率为73.15%，第一主成份
 $Z_1 = 0.1343X_1 + 0.6526X_2 - 0.5945X_3 - 0.4172X_4 - 0.169X_5$

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	21.8039	13.0278	18.6173	10.8926	8.2699	-9.1318	-17.8851	-13.4903	-17.4549	-14.6463
排序	1	3	2	4	5	6	10	7	9	8

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	0.387	0.462
0.577	1	0.599	0.389	0.322
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. 0	0.068590	1.00000	
Eigenvectors					
	PRIN1	PRIN2	PRIN3	PRIN4	PRIN5
X1	0.463605	-.240339	-.611705	0.386635	0.451262
X2	0.457108	-.509305	0.178189	0.206474	-0.676223
X3	0.470176	-.260448	0.335056	-.662445	0.400007
X4	0.421459	0.525665	0.540763	0.472006	0.175599
X5	0.421224	0.581970	-.435176	-.382439	-0.385024

主成分分析结论

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。

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_i (即 $m = 1$)时，这个 Y_i 仍是使用全部 X 变量(p 个)得到的。例如要计算 Y_i 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. 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本在主分量中的地位。


3. 由主成分分析法构造回归模型。

把各主成分作为新自变量代替原来自变量 x 做回归分析。

4. 用主成分分析筛选回归变量。

回归变量的选择有着重的实际意义，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。

主成分分析方法应用实例：

 (1) 将表1中的数据作标准差标准化处理，然后计算相关系数矩阵（表2）。

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

表2 相关系数矩阵

	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉
X ₁	1	-0.327	-0.714	-0.336	0.309	0.408	0.79	0.156	0.744
X ₂	-0.33	1	-0.035	0.644	0.42	0.255	0.009	-0.078	0.094
X ₃	-0.71	-0.035	1	0.07	-0.74	-0.755	-0.93	-0.109	-0.924
X ₄	-0.34	0.644	0.07	1	0.383	0.069	-0.05	-0.031	0.073
X ₅	0.309	0.42	-0.74	0.383	1	0.734	0.672	0.098	0.747
X ₆	0.408	0.255	-0.755	0.069	0.734	1	0.658	0.222	0.707
X ₇	0.79	0.009	-0.93	-0.046	0.672	0.658	1	-0.03	0.89
X ₈	0.156	-0.078	-0.109	-0.031	0.098	0.222	-0.03	1	0.29
X ₉	0.744	0.094	-0.924	0.073	0.747	0.707	0.89	0.29	1

特征值：

0.0315	0	0	0	0	0	0	0	0	0
0	0.0453	0	0	0	0	0	0	0	0
0	0	0.1144	0	0	0	0	0	0	0
0	0	0	0.1926	0	0	0	0	0	0
0	0	0	0	0.3152	0	0	0	0	0
0	0	0	0	0	0.5074	0	0	0	0
0	0	0	0	0	0	1.0430	0	0	0
0	0	0	0	0	0	0	2.0895	0	0
0	0	0	0	0	0	0	0	4.6611	0

特征向量：（与特征值对应）

0.2334	-0.1128	-0.5593	0.3125	0.3548	-0.3747	-0.0599	0.3679	0.3421
0.0470	0.0190	-0.0322	-0.1099	0.7615	0.1551	-0.0276	-0.6135	0.0572
-0.6923	0.2028	-0.4671	0.2060	0.0450	-0.0678	0.0929	-0.0661	-0.4464
0.1395	-0.0040	0.0962	0.3946	-0.3098	-0.5977	0.0362	-0.6006	0.0193
0.0082	0.0622	-0.5798	-0.5077	-0.3957	0.0980	-0.0107	-0.3068	0.3765
-0.0788	-0.0397	-0.0445	0.6383	-0.1543	0.6204	0.1222	-0.1241	0.3793
-0.2354	0.7772	0.2411	0.0042	0.0687	-0.1476	-0.2461	0.0920	0.4322
0.0856	0.2313	0.0443	-0.0926	0.0784	-0.0855	0.9497	0.0695	0.0914
-0.6128	-0.5318	0.2458	-0.1358	0.0712	-0.2240	0.0898	0.0173	0.4464

(3) 对于特征值=4.661, 2.089, 1.043分别求出其特征向量 e_1, e_2, e_3 , 再计算各变量 x_1, x_2, \dots, x_9 在主成分 z_1, z_2, z_3 上的载荷。

$$l_{ij} = p(z_i, x_j) = \sqrt{\lambda_i} e_{ij} (i, j = 1, 2, \dots, p) \quad , \quad S_j = \sum_i l_{ij}^2$$

	z_1	z_2	z_3	占方差的百分数/%
x_1	0.739	0.532	-0.061	82.918
x_2	0.123	-0.887	-0.028	80.191
x_3	-0.964	-0.096	0.095	92.948
x_4	0.042	-0.868	0.037	75.346
x_5	0.813	-0.444	-0.011	85.811
x_6	0.819	-0.179	0.125	71.843
x_7	0.933	0.133	-0.251	95.118
x_8	0.197	0.1	0.97	98.971
x_9	0.964	0.025	0.092	92.939

表 4
主成分载荷

上述计算过程，可以借助于SPSS或Matlab软件系统实现。

(1)第 1 主成分 z_1 与 x_1, x_5, x_6, x_7, x_9 呈现出较强的正相关，与 x_3 呈现出较强的负相关，而这几个变量则综合反映了生态经济结构状况，因此可以认为第 1 主成分 z_1 是生态经济结构的代表。

(2)第 2 主成分 z_2 与 x_2, x_4, x_5 呈现出较强的正相关，与 x_1 呈现出较强的负相关，其中，除了 x_1 为人口总数外， x_2, x_4, x_5 都反映了人均占有资源量的情况，因此可以认为第 2 主成分 z_2 代表了人均资源量。

(3)第 3 主成分 z_3 与 x_8 呈现出的正相关程度最高，其次是 x_6 ，而与 x_7 呈负相关，因此可以认为第 3 主成分在一定程度上代表了农业经济结构。

(4)另外，表 4 中最后一列（占方差的百分数），在一定程度上反映了 3 个主成分 z_1, z_2, z_3 包含原变量（ x_1, x_2, \dots, x_9 ）的信息量多少。

显然，用 3 个主成分 z_1, z_2, z_3 代替原来 9 个变量（ x_1, x_2, \dots, x_9 ）描述农业生态经济系统，可以使问题更进一步简化、明了。

三、聚类分析法

- 对一个数据，既可以对变量(指标)进行分类(相当于对数据中的列分类)，也可以对观测值(事件，样品)来分类(相当于对数据中的行分类)。
- 当然，不一定事先假定有多少类，完全可以按照数据本身的规律来分类。
- 聚类分析 (cluster analysis)：对变量（列）的聚类称为 R 型聚类，而对观测值（行）聚类称为 Q 型聚类。它们在数学上是无区别的。

某地区九个农业区的七项经济指标数据

区代号	人均耕地 $X_1(\text{hm}^2/\text{人})$	劳均耕地 $x_2(\text{hm}^2/\text{个})$	水田比重 $X_3(\%)$	复种指数 $X_4(\%)$	粮食亩产 $x_5(\text{kg}/\text{hm}^2)$	人均粮食 $x_6(\text{kg}/\text{人})$	稻谷占粮食 比重 $x_7(\%)$
G_1	0.294	1.093	5.63	113.6	4510.5	1036.4	12.2
G_2	0.315	0.971	0.39	95.1	2773.5	683.7	0.85
G_3	0.123	0.316	5.28	148.5	6934.5	611.1	6.49
G_4	0.179	0.527	0.39	111	4458	632.6	0.92
G_5	0.081	0.212	72.04	217.8	12249	791.1	80.38
G_6	0.082	0.211	43.78	179.6	8973	636.5	48.17
G_7	0.075	0.181	65.15	194.7	10689	634.3	80.17
G_8	0.293	0.666	5.35	94.9	3679.5	771.7	7.8
G_9	0.167	0.414	2.9	94.8	4231.5	574.6	1.17

- 如果要对 100 个学生进行分类, 而仅知道他们的数学成绩, 则只好按照数学成绩分类; 这些成绩在直线上形成 100 个点。这样就可以把接近的点放到一类。
- 如果还知道他们的物理成绩, 这样数学和物理成绩就形成二维平面上的 100 个点, 也可以按照距离远近来分类。
- 三维或者更高维的情况也是类似; 只不过三维以上的图形无法直观地画出来而已。
- 上面的例子
- 数据样本 X , 由 d 个属性值组成: $X = (x_1, x_2, \dots, x_d)$, 其中 x_i 表示样本中的各属性, d 是样本或样本空间的维数(或属性个数)。
- 数据样本集 $X = \{X_1, X_2, \dots, X_n\}$, 第 i 个样本记为 $X_i = \{x_{i1}, \dots, x_{id}\}$, 许多情况下聚类的样本集看成是一个 $n \times d$ (n 个样本 $\times d$ 个属性)的**数据矩阵**:

$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1d} \\ . & . & . & . & . \\ x_{i1} & \dots & x_{if} & \dots & x_{id} \\ . & . & . & . & . \\ x_{n1} & \dots & x_{nf} & \dots & x_{nd} \end{bmatrix}$$

- 簇 C_i : 数据样本集 X 分成 k 个簇, 每个簇是相应数据样本的集合, 相似样本在同一簇中, 相异样本在不同簇中。
- 簇 C_i ($i=1, 2, \dots, k$) 中样本的数量 n_i 。簇记为 $C_i = \{X_{j1}^i, X_{j2}^i, \dots, X_{j n_i}^i\}$,
- C_i ($i=1, \dots, k$) 是 X 的子集:
 $C_1 \cup C_2 \cup \dots \cup C_k = X$ 且 $C_i \cap C_j = \Phi, i \neq j$

- 用下面的特征来描述簇：

①簇的质心 (centroid) :即样本的平均值, 是簇的“中间值” (middle), 但并不需要是簇中实际点。

令 n_i 表示簇 C_i 中样本的数量, m_i 表示对应样本的均值

②簇的半径, 是簇中两个点间的均方差的平方根。

- 聚类定义: 给定一数据样本集 $X = \{X_1, X_2, \dots, X_n\}$, 根据数据点间的相似程度将数据集合分成 k 簇: $\{C_1, C_2, \dots, C_k\}$ 的过程称为聚类,
- 相似样本在同一簇中, 相异样本在不同簇中。
- 关于同一簇中的样本比来自不同簇的样本更为相似的判断问题主要涉及以下两个独立的子问题:
 - a. 怎样度量样本之间的相似性;
 - b. 怎样衡量对样本集的一种划分的好坏。

1. 相似性度量

- 相异度矩阵 (dissimilarity matrix) 用来存储 n 个样本两两之间的相似性, 表现形式是一个 $n \times n$ 维的矩阵:

$$\begin{bmatrix} 0 & & & & \\ d(X_2, X_1) & 0 & & & \\ d(X_3, X_1) & d(X_3, X_2) & 0 & & \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d(X_n, X_1) & d(X_n, X_2) & \vdots & \vdots & 0 \end{bmatrix}$$

- $d(X_i, X_j)$ 是样本 X_i 和样本 X_j 间相异性的量化表示。
- 最明显的相似性度量是样本之间的距离。
- $X_i = \{x_{i1}, \dots, x_{id}\}$ 和 $X_j = \{x_{j1}, \dots, x_{jd}\}$ 是两个具有 d 个属性的两个样本。距离度量标准 $d(X_i, X_j)$ 表示第 i 个样本与第 j 个样本间的距离。
- 在聚类分析中, 最常用的距离定义如下:
- 最著名的距离度量标准是 d 维空间中的欧几里德距离:

$$d(X_i, X_j) = \left(\sum_{k=1}^d (x_{ik} - x_{jk})^2 \right)^{1/2}$$

- 更广义的 d 维空间中的度量为明考斯基距离 (Minkowski) 度量

$$d(X_i, X_j) = \left(\sum_{k=1}^d |x_{ik} - x_{jk}|^p \right)^{1/p}$$

- 通常也被称为 L_p 范数, 欧几里德距离即 L_2 范数。而 L_1 范数则常被称为曼哈坦 (Manhattan) 距离或城区距离。
- 加权的距离

例: 对于一个4维向量 $X_1 = \{1, 0, 1, 0\}$ 和 $X_2 = \{2, 1, -3, -1\}$, 这些距离的度量标准

$$L_1(X_1, X_2) = 1 + 1 + 4 + 1 = 7,$$

$$L_2(X_1, X_2) = (1 + 1 + 16 + 1)^{1/2} = 4.36$$

$$L_3(X_1, X_2) = (1 + 1 + 64 + 1)^{1/3} = 4.06。$$

$$L_p(X_i, X_j) = \left(\sum_{k=1}^d |x_{ik} - x_{jk}|^p \right)^{1/p}$$



聚类要素的数据处理

- 当分类要素的对象确定之后，在进行聚类分析之前，首先要对聚类要素进行数据处理。假设有n个聚类的对象，每一个聚类对象都有d个要素构成。

聚类对象	要素					
	x_1	x_2	\dots	x_j	\dots	x_d
1	x_{11}	x_{12}	\dots	x_{1j}	\dots	x_{1d}
2	x_{21}	x_{22}	\dots	x_{2j}	\dots	x_{2d}
\vdots	\dots	\dots	\dots	\dots	\dots	\dots
i	x_{i1}	x_{i2}	\dots	x_{ij}	\dots	x_{id}
\vdots	\dots	\dots	\dots	\dots	\dots	\dots
n	x_{n1}	x_{n2}	\dots	x_{nj}	\dots	x_{nd}

- 在聚类分析中，常用的聚类要素的数据处理方法有如下几种：
 - ① 总和标准化。分别求出各聚类要素所对应的数据的总和，以各要素的数据除以该要素的数据的总和，即

$$x'_{ij} = \frac{x_{ij}}{\sum_{i=1}^n x_{ij}} \quad (i=1, 2, \dots, n; j=1, 2, \dots, d)$$

这种标准化方法所得到的新数据满足

$$\sum_{i=1}^n x'_{ij} = 1 \quad (j=1, 2, \dots, d)$$

- ② 标准差标准化，即

$$x'_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \quad (i=1, 2, \dots, n; j=1, 2, \dots, d)$$

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

由这种标准化方法所得到的新数据，各要素的平均值为0，标准差为1，即有

$$\bar{x}'_j = \frac{1}{n} \sum_{i=1}^n x'_{ij} = 0 \quad s'_j = \sqrt{\frac{1}{n} \sum_{i=1}^n (x'_{ij} - \bar{x}'_j)^2} = 1$$

③ 极大值标准化，即

$$x'_{ij} = \frac{x_{ij}}{\max_i \{x_{ij}\}} \quad (i=1,2,\dots,n; j=1,2,\dots,d)$$

经过这种标准化所得的新数据，各要素的极大值为1，其余各数值小于1。

④ 极差的标准化，即

$$x'_{ij} = \frac{x_{ij} - \min_i \{x_{ij}\}}{\max_i \{x_{ij}\} - \min_i \{x_{ij}\}} \quad (i=1,2,\dots,n; j=1,2,\dots,d)$$

经过这种标准化所得的新数据，各要素的极大值为1，极小值为0，其余的数值均在0与1之间。

- 经过极差标准化处理后的某地区九个农业区的七项经济指标数据

区代号	人均耕地 $X_1(\text{hm}^2/\text{人})$	劳均耕地 $x_2(\text{hm}^2/\text{个})$	水田比重 $X_3(\%)$	复种指数 $X_4(\%)$	粮食亩产 $x_5(\text{kg}/\text{hm}^2)$	人均粮食 $x_6(\text{kg}/\text{人})$	稻谷占粮食 比重 $x_7(\%)$
G_1	0.91	1.00	0.07	0.15	0.18	1.00	0.14
G_2	1.00	0.87	0.00	0.00	0.00	0.24	0.00
G_3	0.20	0.15	0.07	0.44	0.44	0.08	0.07
G_4	0.44	0.38	0.00	0.13	0.18	0.13	0.00
G_5	0.03	0.03	1.00	1.00	1.00	0.45	1.00
G_6	0.03	0.03	0.61	0.69	0.65	0.13	0.59
G_7	0.00	0.00	0.90	0.81	0.84	0.13	1.00
G_8	0.91	0.53	0.07	0.00	0.10	0.43	0.09
G_9	0.38	0.26	0.04	0.00	0.15	0.00	0.00

- 九个农业区之间的曼哈坦 (Manhattan) 距离矩阵

$$D = (d_{ij})_{9 \times 9} = \begin{bmatrix} 0 & & & & & & & & \\ 1.52 & 0 & & & & & & & \\ 3.10 & 2.70 & 0 & & & & & & \\ 2.19 & 1.47 & 1.23 & 0 & & & & & \\ 5.86 & 6.02 & 3.64 & 4.77 & 0 & & & & \\ 4.72 & 4.46 & 1.86 & 2.99 & 1.78 & 0 & & & \\ 5.79 & 5.53 & 2.93 & 4.06 & 0.83 & 1.07 & 0 & & \\ 1.32 & 0.88 & 2.24 & 1.29 & 5.14 & 3.96 & 5.03 & 0 & \\ 2.62 & 1.66 & 1.20 & 0.51 & 4.84 & 3.06 & 3.32 & 1.40 & 0 \end{bmatrix}$$

直接聚类法:

- 原理：先把各个分类对象单独视为一类，然后根据距离最小的原则，依次选出一对分类对象，并成新类。
- 如果其中一个分类对象已归于一类，则把另一个也归入该类；

➤ 如果一对分类对象正好属于已归的两类，则把这两类并为一类。

每一次归并，都划去该对象所在的列与列序相同的行。经过 $n-1$ 次就可以把全部分类对象归为一类，这样就可以根据归并的先后顺序作出聚类谱系图。

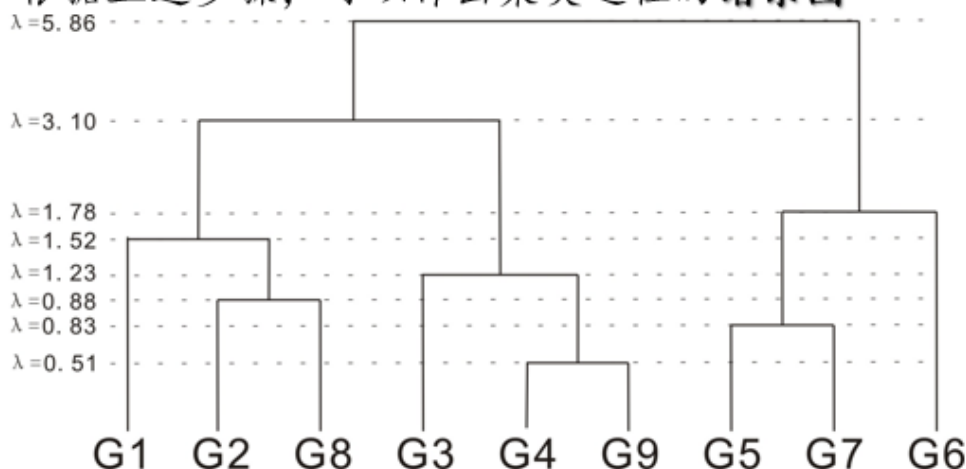
用直接聚类法对某地区的九个农业区进行聚类分析，步骤如下：

- ①在距离矩阵D中，除去对角线元素以外， $d_{49}=d_{94}=0.51$ 为最小者，
故将第4区与第9区并为一类，划去第9行和第9列；
- ②在余下的元素中，除对角线元素以外， $d_{75}=d_{57}=0.83$ 为最小者，
故将第5区与第7区并为一类，划掉第7行和第7列；
- ③在第二步之后余下的元素之中，除对角线元素以外， $d_{82}=d_{28}=0.88$ 为最小者，故将第2区与第8区并为一类，划去第8行和第8列；

$$D=(d_{ij})_{9 \times 9} = \begin{bmatrix} 0 & & & & & & & & \\ 1.52 & 0 & & & & & & & \\ 3.10 & 2.70 & 0 & & & & & & \\ 2.19 & 1.47 & 1.23 & 0 & & & & & \\ 5.86 & 6.02 & 3.64 & 4.77 & 0 & & & & \\ 4.72 & 4.46 & 1.86 & 2.99 & 1.78 & 0 & & & \\ 5.79 & 5.53 & 2.93 & 4.06 & 0.83 & 1.07 & 0 & & \\ 1.32 & 0.88 & 2.24 & 1.29 & 5.14 & 3.96 & 5.03 & 0 & \\ 2.62 & 1.66 & 1.20 & 0.51 & 4.84 & 3.06 & 3.32 & 1.40 & 0 \end{bmatrix}$$

- ④在第三步之后余下的元素中，除对角线元素以外， $d_{43}=d_{34}=1.23$ 为最小者，故将第3区与第4区并为一类，划去第4行和第4列，此时，第3、4、9区已归并为一类；
- ⑤在第四步之后余下的元素中，除对角线元素以外， $d_{21}=d_{12}=1.52$ 为最小者，故将第1区与第2区并为一类，划去第2行和第2列，此时，第1、2、8区已归并为一类；
- ⑥在第五步之后余下的元素中，除对角线元素以外， $d_{65}=d_{56}=1.78$ 为最小者，故将第5区与第6区并为一类，划去第6行和第6列，此时，第5、6、7区已归并为一类；
- ⑦在第六步之后余下的元素中，除对角线元素以外， $d_{31}=d_{13}=3.10$ 为最小者，故将第1区与第3区并为一类，划去第3行和第3列，此时，第1、2、3、4、8、9区已归并为一类；
- ⑧在第七步之后余下的元素中，除去对角线元素以外，只有 $d_{51}=d_{15}=5.86$ ，故将第1区与第5区并为一类，划去第5行和第5列，此时，第1、2、3、4、5、6、7、8、9、区均归并为一类；

• 根据上述步骤，可以作出聚类过程的谱系图



k-means 聚类算法

1. 选择一个含有随机选择样本的 k 个簇的初始划分，计算这些簇的质心。
2. 根据距离把剩余的每个样本分配到距离它最近的簇质心的一个划分。
3. 计算被分配到每个簇的样本的均值向量，作为新的簇的质心。
4. 重复 2,3 直到 k 个簇的质心点不再发生变化或准则函数收敛。

- 坐标表示5个点 $\{X_1, X_2, X_3, X_4, X_5\}$ 作为一个聚类分析的二维样本： $X_1 = (0, 2)$ ， $X_2 = (0, 0)$ ， $X_3 = (1.5, 0)$ ， $X_4 = (5, 0)$ ， $X_5 = (5, 2)$ 。假设要求的簇的数量 $k=2$ 。

- 第1步：由样本的随机分布形成两个簇：

$$C_1 = \{X_1, X_2, X_4\} \text{ 和 } C_2 = \{X_3, X_5\}。$$

这两个簇的质心 M_1 和 M_2 是：

$$M_1 = \{(0+0+5)/3, (2+0+0)/3\} = \{1.66, 0.66\}；$$

$$M_2 = \{(1.5+5)/2, (0+2)/2\} = \{3.25, 1.00\}；$$

- 样本初始随机分布之后，方差是：

$$e_1^2 = [(0-1.66)^2 + (2-0.66)^2] + [(0-1.66)^2 + (0-0.66)^2] + [(5-1.66)^2 + (0-0.66)^2] = 19.36；$$

$$e_2^2 = 8.12；$$

- 总体平方误差是： $E^2 = e_1^2 + e_2^2 = 19.36 + 8.12 = 27.48$

$$J_e = \sum_{i=1}^k \sum_{X \in C_i} |X - m_i|^2$$

第2步：取距离其中一个质心 (M_1 或 M_2) 最小的距离分配所有样本，簇内样本的重新分布如下：

$$d(M_1, X_1) = (1.66^2 + 1.34^2)^{1/2} = 2.14$$

$$d(M_2, X_1) = 3.40 \implies X_1 \in C_1；$$

$$d(M_1, X_2) = 1.79 \text{ 和 } d(M_2, X_2) = 3.40 \implies X_2 \in C_1$$

$$d(M_1, X_3) = 0.83 \text{ 和 } d(M_2, X_3) = 2.01 \implies X_3 \in C_1$$

$$d(M_1, X_4) = 3.41 \text{ 和 } d(M_2, X_4) = 2.01 \implies X_4 \in C_2$$

$$d(M_1, X_5) = 3.60 \text{ 和 } d(M_2, X_5) = 2.01 \implies X_5 \in C_2$$

$$\text{新簇 } C_1 = \{X_1, X_2, X_3\} \text{ 和 } C_2 = \{X_4, X_5\}$$

- 第3步：计算新的质心：

$$M_1 = \{0.5, 0.67\}； M_2 = \{5.0, 1.0\}。$$

- 相应的方差及总体平方误差分别是：

$$e_1^2 = 4.17； e_2^2 = 2.00； E = 6.17；$$

- 可以看出第一次迭代后，总体误差显著减小（从值 27.48 到 6.17）。

■ 在这个简单的例子中，第一次迭代同时也是最后一次迭代，因为如果继续分析新中心和样本间的距离，样本将会全部分给同样的簇，不将重新分配，算法停止。