# 应用多元统计分析

Mr.tan

2017 年 6 月 22 日

| 单词 | 中文 | 单词 | 中文 |
|---|---|---|---|
| Parameter Estimation | 参数估计 | Total Variance Explained | 总方差解释 |
| Hypothetical Test | 假设检验 | Cumulative | 累积的 |
| Discriminant Analysis | 判别分析 | Component matrix | 因子载荷矩阵 |
| Cluster Analysis | 聚类分析 | Factor analysis | 因子分析 |
| Principal Component Analysis | 主成分分析 | Rotated Component Matrix | 旋转因子载荷矩阵 |
| Correlation Matrix | 相关矩阵 | Component Score Coefficient Matrix | 因子得分系数矩阵 |
| Coefficient | 系数 | Communalities | 公因子方差 (共同度) |
| Significant Level | 显著水平 | Correspondence Analysis | 相应分析 |
| Extract | 提取 | Proportion of Inertia | 惯性比例 |
| Rotation | 旋转 | Score in Dimension | 维度得分 |
| 因子载荷矩阵除以对应的根号下特征值就是主成分系数 | | | |

## 1 多元分析概述

1. 多元统计方法用于解决什么问题?

答：多元统计分析方法在经济管理、农业、医学、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、文学等方面都有广泛的运用。可解决如下问题：

- 假设的提出与检验
- 数据或结构性化简
- 分类或组合
- 变量之间的关系
- 预测与决策

2. 应用多元统计分析主要包括哪些分析方法? 这些方法分别可用于解决哪方面的具体问题?

答：应用多元统计分析主要包括判别分析、聚类分析、主成分分析、因子分析、相应分析、典型相关分析等方法。

**判别分析**：在互联网上，淘宝根据客户购买商品的信息，通过判别分析预测客户下一次购买产品的类型。

**聚类分析**：银行希望根据客户过去的贷款数据，来预测新的贷款者核贷后逾期的机率，在此可以用聚类分析。

**主成分分析**：利用主成分分析对各地区城市的设施水平进行综合评价和排序。

**因子分析**：已知学生各科成绩，通过因子分析可以判断适合读文科还是理科。

**相应分析**：研究头发颜色与眼睛颜色的关系

**典型相关分析**：通过典型相关分析来反映我国财政收入与财政支出之间的关系。

**总结：已知每位学生六门课的成绩，可以用到聚类分析、判别分析、主成分分析、因子分析、相应分析、典型相关分析**

# 2 多元正态分布的参数估计

## 2.1 随机向量的数字特征

均值向量的性质：

$$
\begin{aligned}
&E(AX) = AE(X)\\
&E(AXB) = AE(X)B\\
&E(AX + BY) = AE(X) + BE(Y)
\end{aligned}
$$

协方差阵的性质：

$$
\begin{aligned}
&D(X + a) = D(X)\\
&D(AX) = AD(X)A' = A\Sigma A'\\
&D(AX, BY) = A\mathrm{cov}(X, Y)B'
\end{aligned}
$$

相关系数：

$$\rho_{ij} = \frac{\mathrm{cov}(X_i, X_j)}{\sqrt{D(X_i)}\sqrt{D(X_j)}}$$

## 2.2 威沙特分布 wishart

设 $X_{(a)} = (X_{a1}, X_{a2}, \cdots, X_{ap})' \sim N_p(\mu_a, \Sigma)$，$a = 1, 2, \cdots, n$，且相互独立，则由 $\boldsymbol{X}_{(a)}$ 组成的随机矩阵：

$$W_{p\times p} = \sum_{a=1}^{n} X_{(a)}X'_{(a)}$$

的分布称为非中心 Wishart 分布，记为 $W_p(n,\Sigma,Z)$

其中 $Z=(\mu_{a1},\cdots,\mu_{an})(\mu_{a1},\cdots,\mu_{an})'=\sum_{a=1}^{n}\mu_a\mu'_a$，$\mu_a$ 称为非中心参数；当 $\mu_a=0$ 时称为中心 Wishart 分布，记为 $W_p(n,\Sigma)$

Wishart 分布的基本性质：

1 .若 $X_{(a)}\sim N_p(\mu,\Sigma)$，且 $a=1,2,\cdots,n$ 相互独立，则样本离差阵 $\boldsymbol{S}=\sum_{a=1}^{n}(\boldsymbol{X}_{(a)}-\bar{\boldsymbol{X}})(\boldsymbol{X}_{(a)}-\bar{\boldsymbol{X}})'$ $W_p(n-1,\Sigma)$，其中 $\bar{\boldsymbol{X}}=\frac{1}{n}\sum_{a=1}^{n}\boldsymbol{X}_{(a)}$。

2．若 $S_i\sim W_p(n_i,\Sigma)$，$i=1,\cdots,k$，且相互独立，则 $\sum_{i=1}^{k}S_i\sim W_p(\sum_{i=1}^{k}n_i,\Sigma)$。

3．若 $X_{p\times p}\sim W_p(n,\Sigma)$，$C_{p\times p}$ 为非奇异阵，则 $CXC'\sim W_p(n,C\Sigma C')$。

### 2.3 荔枝

设三维均值向量 $X=(X1,X2,X3)'$ 的均值向量为 $E(x)=(5,0,1)'$，其协方差矩阵 $\Sigma=\begin{bmatrix}1&0&-2\\0&3&-1\\-2&-1&4\end{bmatrix}$，设 $A=\begin{bmatrix}0&2&5\\4&3&-1\end{bmatrix}$，求 $E(AX)$ 和 cov(AX)

解：

$$E(AX)=AE(X)=\begin{bmatrix}0&2&5\\4&3&-1\end{bmatrix}(5,0,1)'=\begin{bmatrix}-5\\19\end{bmatrix}$$

$$\operatorname{cov}(AX)=AD(X)A'=\begin{bmatrix}0&2&5\\4&3&-1\end{bmatrix}\begin{bmatrix}1&0&-2\\0&3&-1\\-2&-1&4\end{bmatrix}\begin{bmatrix}0&4\\2&3\\5&-1\end{bmatrix}=\begin{bmatrix}-88&95\\-55&69\end{bmatrix}$$

$$\rho=\begin{pmatrix}1&0&-2\Big/\sqrt{1}\sqrt{4}\\0&1&-1\Big/\sqrt{3}\sqrt{4}\\-2\Big/\sqrt{1}\sqrt{4}&-1\Big/\sqrt{3}\sqrt{4}&1\end{pmatrix}$$

## 3 多元正态分布均值向量和协差阵的检验

多元统计分析中的各种均值向量和协差阵的检验

其基本思想和步骤均可归纳为：

第一，提出待检验的假设 $H_0$ 和 $H_1$；

第二，给出检验的统计量及其服从的分布；

第三，给定检验水平，查统计量的分布表，确定相应的临界值，从而得到否定域；

第四，根据样本观测值计算出统计量的值，看是否落入否定域中，以便对待判假设做出决策（拒绝或接受）。

## 3.1 霍特林分布 Hotelling

设 $X \sim N_p(\mu, \Sigma)$，$S \sim W_p(n, \Sigma)$ 且 $X$ 与 $S$ 相互独立，$n \geq p$，则称统计量 $T^2 = nX'S^{-1}X$ 的分布为非中心霍特林 $T^2$ 分布，记为 $T^2 \sim T^{\mathbf{2}}(p, n, \mu)$。当 $\mu = 0$ 时，称 $T^2$ 服从（中心）霍特林 $T^2$ 分布。记为 $T^{\mathbf{2}}(p, n)$。

## 3.2 威尔克斯统计量 wilks

若 $\boldsymbol{A}_1 \sim W_p(n_1, \boldsymbol{\Sigma}), n_1 \geq p, \boldsymbol{A}_2 \sim W_p(n_2, \boldsymbol{\Sigma}), \boldsymbol{\Sigma} > 0$ 且 $A_1$ 和 $A_2$ 相互独立，则称

$$\Lambda = \frac{|\boldsymbol{A}_1|}{|\boldsymbol{A}_1 + \boldsymbol{A}_2|}$$

为威尔克斯统计量，$\Lambda$ 的分布称为威尔克斯分布，简记为 $\Lambda \sim \Lambda(p, n_1, n_2)$，其中 $n_1$，$n_2$ 为自由度。

# 4 判别分析

**判别分析：即根据历史上划分类别的有关资料和某种最优准则，确定一种判别方法，判定一个新的样本归属哪一类。**

某医院有部分患有肺炎、肝炎、冠心病、糖尿病等病人的资料，记录了每个患者若干项症状指标数据。现在想利用现有的这些资料找出一种方法，使得对于一个新的病人，当测得这些症状指标数据时，能够判定其患有哪种病

## 4.1 马氏距离

设 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 是来自均值向量为 $\mu$，协方差为 $\boldsymbol{\Sigma}(> 0)$ 的总体 $G$ 中的 $p$ 维样本，则总体 $G$ 内两点 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 之间的马氏距离定义为

$$D^2(\boldsymbol{X}, \boldsymbol{Y}) = (\boldsymbol{X} - \boldsymbol{Y})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{Y})$$

定义点 $X$ 到总体 $G$ 的马氏距离为

$$D^2(\boldsymbol{X}, G) = (\boldsymbol{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$$

## 4.2 两个总体的距离判别问题

问题：设有协方差矩阵 $\square$ 相等的两个总体 $G_1$ 和 $G_2$，其均值分别是 $\mu_1$ 和 $\mu_2$，对于一个新的样品 $X$，要判断它来自哪个总体。

$$\begin{aligned}
&D^2(\boldsymbol{X}, G_1) - D^2(\boldsymbol{X}, G_2) \\
&= (\boldsymbol{X} - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_1) - (\boldsymbol{X} - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_2) \\
&= -2\big(\boldsymbol{X} - \tfrac{\mu_1 + \mu_2}{2}\big)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\
&= -2(\boldsymbol{X} - \bar{\boldsymbol{\mu}})'\boldsymbol{\alpha} = -2\boldsymbol{\alpha}'(\boldsymbol{X} - \bar{\boldsymbol{\mu}})
\end{aligned}$$

其中 $\bar{\boldsymbol{\mu}} = \frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)$ 是两个总体均值的平均值，$\boldsymbol{\alpha} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$，记 $W(\boldsymbol{X}) = \boldsymbol{\alpha}'(\boldsymbol{X}-\bar{\boldsymbol{\mu}})$，则判别规则可表示为

$$\begin{cases} \boldsymbol{X} \in G_1, & if \quad W(\boldsymbol{X}) \geq 0 \\ \boldsymbol{X} \in G_2, & if \quad W(\boldsymbol{X}) < 0 \end{cases}$$

这里称 $W(\boldsymbol{X})$ 为两总体距离判别的判别函数，由于它是 $X$ 的线性函数，故又称为线性判别函数，$\alpha$ 称为判别系数。

## 4.3 贝叶斯判别法 Bayes

距离判别法虽然简单，便于使用。但是该方法也有它明显的不足之处。

第一，判别方法与总体各自出现的概率的大小无关；

第二，判别方法与错判之后所造成的损失无关。Bayes 判别法就是为了解决这些问题而提出的一种判别方法。

总平均损失：

$$g(R) = \sum_{i=1}^{k} q_i r(i,R) = \sum_{i=1}^{k} q_i \sum_{j=1}^{k} C(j|i) P(j|i,R)$$

所谓 Bayes 判别法则，就是要选择一种划分 $R_1, R_2, \cdots R_k$，使得总平均损失 $G(R)$ 达到极小。

## 4.4 费歇尔判别法 Fisher

### 4.4.1 费希尔判别的基本思想

**主要思想是通过将多维数据投影到某个方向上，投影的原则是将总体与总体之间尽可能的放开，然后再选择合适的判别规则，将新的样品进行分类判别。**

从 $k$ 个总体中抽取具有 $p$ 个指标的样品观测数据，借助方差分析的思想构造一个线性判别函数

$$U(\mathrm{X}) = u_1X_1 + u_2X_2 + \cdots + u_pX_p = \mathrm{u}'\mathrm{X}$$

其中系数 $\mathrm{u} = (u_1, u_2, \cdots, u_p)'$ 确定的原则是使得总体之间区别最大，而使每个总体内部的离差最小。

## 4.5 梨子

1. 简述欧几里得距离与马氏距离的区别和联系?

欧几里得距离的局限有，1. 在多元数据分析中，其度量不合理，2. 会受到实际问题中量纲的影响。

马氏距离定义为 $D^2(\boldsymbol{X},\boldsymbol{Y}) = (\boldsymbol{X}-\boldsymbol{Y})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{Y})$，当协方差矩阵为单位矩阵时即为欧几里得距离。

在一定程度上，欧几里得距离是马氏距离的特殊情况，马氏距离是欧几里得距离的推广。

2. 试述判别分析的实质？

判别分析就是希望利用已经测得的变量数据，找出一种判别函数，使得这一函数具有某种最优性质，能把属于不同类别的样本点尽可能地区别开来。实质上就是在某种意义上，以最优的性质对 $p$ 维空间 $R_p$ 构造一个“划分”，这个“划分”就构成了一个判别规则。

3. 简述距离判别法的基本思想？

距离判别问题分为:1. 两个总体的距离判别问题，2. 多个总体的判别问题。其基本思想都是分别计算样本与各个总体的距离（马氏距离），将距离近的判别为一类。

4. 简述贝叶斯判别法的基本思想和方法？

贝叶斯判别法的基本思想：选择一种划分 $R_1, R_2, \cdots R_k$，使得总平均损失 $G(R)$ 达到极小。

5. 简述费希尔判别法的基本思想？

主要思想是通过将多维数据投影到某个方向上，投影的原则是将总体与总体之间尽可能的放开，然后再选择合适的判别规则，将新的样品进行分类判别。

6. 试析距离判别法、贝叶斯判别法和费希尔判别法的异同

这个就将三者的思想答出来，再加以区别。

7. 设有两个二元总体 $G_1$ 和 $G_2$，从中分别抽取样本计算得到

$$\overline{X}^{(1)} = \begin{bmatrix} 5 \\ 1 \end{bmatrix}, \overline{X}^{(2)} = \begin{bmatrix} 3 \\ -2 \end{bmatrix}, \Sigma_p = \begin{bmatrix} 5.8 & 2.1 \\ 2.1 & 7.6 \end{bmatrix}$$

假设 $\Sigma_1 = \Sigma_2$，试用距离判别法建立判别函数和判别规则。样品 X=（6，0）’应属于哪个总体？

解：

$$\mu_1 = \overline{X}^{(1)} = \begin{bmatrix} 5 \\ 1 \end{bmatrix}, \mu_2 = \overline{X}^{(2)} = \begin{bmatrix} 3 \\ -2 \end{bmatrix}, \mu = \frac{\mu_1 + \mu_2}{2} = \begin{bmatrix} 4 \\ -0.5 \end{bmatrix}$$

判别函数：

$$W(X) = (X - \mu)' \Sigma^{-1} (\mu_1 - \mu_2)$$

判别规则：

$$\begin{cases} \boldsymbol{X} \in G_1, & if \quad W(\boldsymbol{X}) \geq 0 \\ \boldsymbol{X} \in G_2, & if \quad W(\boldsymbol{X}) < 0 \end{cases}$$

判断样本类别：

$$\begin{aligned} & W(X) = (X - \mu)' \Sigma^{-1} (\mu_1 - \mu_2) \\ & = \left( \begin{bmatrix} 6 \\ 0 \end{bmatrix} - \begin{bmatrix} 4 \\ -0.5 \end{bmatrix} \right)' \left( \frac{1}{39.67} \begin{bmatrix} 7.6 & -2.1 \\ -2.1 & 5.8 \end{bmatrix} \right) \left( \begin{bmatrix} 5 \\ 1 \end{bmatrix} - \begin{bmatrix} 3 \\ -2 \end{bmatrix} \right) \\ & = \frac{24.4}{39.67} > 0 \end{aligned}$$

# 5 聚类分析

Q 型聚类是对样品进行分类处理，R 型聚类是对变量进行分类处理。

设 $X_i$ 与 $X_j$ 是来自均值向量为 $\mu$，协方差为 $\Sigma$ 的总体 $G$ 中的 $p$ 维样品，则两个样品间的马氏距离为

$$d_{ij}^2(M) = (\boldsymbol{X}_i - \boldsymbol{X}_j)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_i - \boldsymbol{X}_j)$$

## 5.1 系统聚类法

**系统聚类法基本思想和步骤**

基本思想：距离相近的样品（或变量）先聚成类，距离相远的后聚成类，过程一直进行下去，每个样品（或变量）总能聚到合适的类中。

系统聚类过程是：假设总共有 n 个样品（或变量），

第一步：将每个样品（或变量）独自聚成一类，共有 n 类；

第二步：根据所确定的样品（或变量）“距离”公式，把距离较近的两个样品（或变量）聚合为一类，其它的样品（或变量）仍各自聚为一类，共聚成 n-1 类；

第三步：将“距离”最近的两个类进一步聚成一类，共聚成 n-2 类；...，以上步骤一直进行下去，最后将所有的样品（或变量）全聚成一类。

### 5.1.1 最短距离法

最短距离法进行聚类分析的步骤如下：

(1) 定义样品之间距离，计算样品的两两距离，得一距离阵记为 $D_{(0)}$，开始每个样品自成一类，显然这时 $D_{ij} = d_{ij}$。

(2) 找出距离最小元素，设为 $D_{pq}$，则将 $G_p$ 和 $G_q$ 合并成一个新类，记为 $G_r$，即 $G_r = \{G_p, G_q\}$。

(3) 按 (2) 计算新类与其它类的距离。

(4) 重复 (2)、(3) 两步，直到所有元素并成一类为止。如果某一步距离最小的元素不止一个，则对应这些最小元素的类可以同时合并。

## 5.2 粒子

1. 设有六个样品，每个只测量一个指标，分别是 1，2，5，7，9，10，试用最短距离法将它们分类。

解：

(1) 样品采用绝对值距离，计算样品间的距离阵 $D_{(0)}$。

表 1: $D_{(0)}$

| | G1 | G2 | G3 | G4 | G5 | G6 |
|---|---|---|---|---|---|---|
| G1 | 0 | | | | | |
| G2 | 1 | 0 | | | | |
| G3 | 4 | 3 | 0 | | | |
| G4 | 6 | 5 | 2 | 0 | | |
| G5 | 8 | 7 | 4 | 2 | 0 | |
| G6 | 9 | 8 | 5 | 3 | 1 | 0 |

(2)$D_{(0)}$ 中最小的元素是 $D_{12} = D_{56} = 1$，于是将 $G_1$ 和 $G_2$ 合并成 $G_7$，$G_5$ 和 $G_6$ 合并成 $G_8$（也可将 $G_1$、$G_2$，$G_5$、$G_6$ 合并为一类），并重新计算新类与其它类的距离 $D_{(1)}$。

表 2: $D_{(1)}$

| | G7 | G3 | G4 | G8 |
|---|---|---|---|---|
| G7 | 0 | | | |
| G3 | 3 | 0 | | |
| G4 | 5 | 2 | 0 | |
| G8 | 7 | 4 | 2 | 0 |

(3) 在 $D_{(1)}$ 中最小值是 $D_{34} = D_{48} = 2$，由于 $G_4$ 与 $G_3$ 合并，又与 $G_8$ 合并，因此 $G_3$、$G_4$、$G_8$ 合并成一个新类 $G_9$，其与其它类的距离 $D_{(2)}$。

表 3: $D_{(2)}$

| | G7 | G9 |
|---|---|---|
| G7 | 0 | |
| G9 | 3 | 0 |

(4) 最后将 $G_7$ 和 $G_9$ 合并成 $G_{10}$，这时所有的六个样品聚为一类，其过程终止。

最后再画出聚类谱系图，注意横轴刻度指的是聚类距离

### 5.2.1 最长距离法

最长距离法与最短距离法的聚类步骤一样的，只是需要注意的是**类与类之间的距离定义为最大的，计算新类与其他类的距离也是最大的，每次选择还是选最小的。**

### 5.2.2 重心法

重心法定义类间距离为两类重心（各类样品的均值）的距离

2. 设有六个样品，每个只测量一个指标，分别是1，2，5，7，9，10，试用重心法将它们聚类。

(1) 样品采用欧氏距离，计算样品间的平方距离阵 $D^2_{(0)}$

表 4: $D^2_{(0)}$

| | G1 | G2 | G3 | G4 | G5 | G6 |
|---|---|---|---|---|---|---|
| G1 | 0 | | | | | |
| G2 | 1 | 0 | | | | |
| G3 | 16 | 9 | 0 | | | |
| G4 | 36 | 25 | 4 | 0 | | |
| G5 | 64 | 49 | 16 | 4 | 0 | |
| G6 | 81 | 64 | 25 | 9 | 1 | 0 |

(2)$D^2_{(0)}$ 中最小的元素是 $D^2_{(12)} = D^2_{(56)} = 1$，于是将 $G_1$ 和 $G_2$ 合并成 $G_7$，$G_5$ 和 $G_6$ 合并成 $G_8$，并重新计算新类与其它类的距离得到距离阵 $D^2_{(1)}$

表 5: $D^2_{(1)}$

| | G1 | G2 | G3 | G4 |
|---|---|---|---|---|
| G1 | 0 | | | |
| G2 | 12.25 | 0 | | |
| G3 | 30.25 | 4 | 0 | |
| G4 | 64 | 20.25 | 6.25 | 0 |

(3) 在 $D^2_{(1)}$ 中最小值是 $D^2_{(34)} = 4$，那么 $G_3$ 与 $G_4$ 合并一个新类 $G_9$，其与其它类的距离 $D^2_{(2)}$

表 6: $D^2_{(2)}$

| | G7 | G9 | G8 |
|---|---|---|---|
| G7 | 0 | | |
| G9 | 20.25 | 0 | |
| G8 | 64 | 12.25 | 0 |

(4) 在 $D^2_{(2)}$ 中最小值是 $D^2_{(89)} = 12.5$，那么 $G_8$ 与 $G_9$ 合并一个新类，其与与其它类的距离 $D^2_{(3)}$

表 7: $D^2_{(3)}$

| | G7 | G10 |
|---|---|---|
| G7 | 0 | |
| G10 | 39.0625 | 0 |

(5) 最后将 $G_7$ 和 $G_{10}$ 合并成 $G_{11}$，这时所有的六个样品聚为一类，其过程终止。

## 5.3 K 均值聚类法

主要包括以下三个步骤：

1．将所有的样品分成 K 个初始类；

2．通过欧氏距离将某个样品划入离中心最近的类中，并对获得样品与失去样品的类，重新计算中心坐标；

3．重复步骤 2，直到所有的样品都不能再分配时为止。

### 5.3.1 栗子

3. 假定我们对 A、B、C、D 四个样品分别测量两个变量和得到结果见表

| 样品 | 变量 | |
|---|---|---|
| | X1 | X2 |
| A | 5 | 3 |
| B | -1 | 1 |
| C | 1 | -2 |
| D | -3 | -2 |

P87 页

4. 判别分析和聚类分析有何区别?

判别分析是根据一定的判别准则，判定一个样本归属于哪一类。具体而言，设有 n 个样本，对每个样本测得 p 项指标（变量）的数据，已知每个样本属于 k 个类别（或总体）中的某一类，通过找出一个最优的划分，使得不同类别的样本尽可能地区别开，并判别该样本属于哪个总体。聚类分析是分析如何对样品（或变量）进行量化分类的问题。在聚类之前，我们并不知道总体，而是通过一次次的聚类，使相近的样品（或变量）聚合形成总体。**通俗来讲，判别分析是在已知有多少类及是什么类的情况下进行分类，而聚类分析是在不知道类的情况下进行分类。**

5. 简述系统聚类的基本思想?

系统聚类的基本思想是：距离相近的样品（或变量）先聚成类，距离相远的后聚成类，过程一直进行下去，每个样品（或变量）总能聚到合适的类中。

6. 试述 K 均值法与系统聚类法的异同?

相同：K—均值法和系统聚类法一样，都是以距离的远近亲疏为标准进行聚类的。

不同：系统聚类对不同的类数产生一系列的聚类结果，而 K—均值法只能产生指定类数的聚类结果。

具体类数的确定，离不开实践经验的积累；有时也可以借助系统聚类法以一部分样品为对象进行聚类，其结果作为 K—均值法确定类数的参考

# 6 主成分分析

多变量、降维、线性组合、正交变换、坐标旋转

**基本思想：多个变量之间往往存在着一定程度的相关性。人们自然希望通过线性组合的方式，从这些指标中尽可能快地提取信息。当第一个线性组合不能提取更多的信息时，再考虑用第二个线性组合继续这个快速提取的过程，...，直到所提取的信息与原指标相差不多时为止。这就是主成分分析的思想**

$$\begin{cases} Y_1 = t_{11}X_1 + t_{12}X_2 + \cdots + t_{1p}X_p = T_1{}'\mathrm{X} \\ Y_2 = t_{21}X_1 + t_{22}X_2 + \cdots + t_{2p}X_p = T_2{}'\mathrm{X} \\ \cdots\cdots\cdots\cdots \\ Y_p = t_{p1}X_1 + t_{p2}X_2 + \cdots + t_{pp}X_p = T_p{}'\mathrm{X} \end{cases}$$

用矩阵表示为 $\mathrm{Y} = \mathrm{T}'\mathrm{X}$，其中 $\mathrm{Y} = (Y_1, Y_2, \cdots Y_p)'$，$\mathrm{T} = (T_1, T_2, \cdots, T_p)$。

结论：变量 X 的协方差矩阵为 $\boldsymbol{\Sigma}$。协方差矩阵对应的特征值就是 $\lambda$，对应的特征向量就是需要求解的系数向量 T

主成分的方差贡献率

$$\varphi_k = \frac{\lambda_k}{\sum_{k=1}^{p} \lambda_k}$$

**主成分分析的具体步骤可以归纳为：**

1. 将原始数据标准化；

2. 建立变量的相关系数阵 R(如果不标准化，这里就是协方差矩阵)；

3. 求 R 的特征根为 $\lambda_1^* \geq \cdots \geq \lambda_p^* \geq 0$，相应的特征向量为 $T_1^*, T_2^*, \cdots, T_p^*$；

4. 由累积方差贡献率确定主成分的个数 (m)，并写出主成分为 $Y_i = (T_i^*)'X, i = 1, 2, \cdots, m$

## 6.1 利用主成分进行综合评价

对主成分进行加权综合。我们利用主成分进行综合评价时，主要是将原有的信息进行综合，因此，要充分的利用原始变量提供的信息。**将主成分的权数根据它们的方差贡献率来确定，因为方差贡献率反映了各个主成分的信息含量多少**

## 6.2 李子

1. 试述主成分分析的基本思想?

我们处理的问题多是多指标变量问题，由于多个变量之间往往存在着一定程度的相关性，人们希望能通过线性组合的方式从这些指标中尽可能快的提取信息。当第一个组合不能提取更多信息时，再考虑第二个线性组合。继续这个过程，直到提取的信息与原指标差不多时为止。这就是主成分分析的基本思想。

2. 主成分分析的作用体现在何处?

一般说来，在主成分分析适用的场合，用较少的主成分就可以得到较多的信息量。以各个主成分为分量，就得到一个更低维的随机向量；**主成分分析的作用就是在降低数据“维数”的同时又保留了原数据的大部分信息。**

3. 已知 $X=(X_1,X_2,X_3)'$ 的协方差矩阵为

$$\Sigma=\begin{bmatrix} 11 & \sqrt{3}\Big/2 & 3/2 \\ \sqrt{3}\Big/2 & 21/4 & 5\sqrt{3}\Big/4 \\ 3/2 & 5\sqrt{3}\Big/4 & 31/4 \end{bmatrix}$$

试进行主成分分析?

方法：(1). 根据

$$|\Sigma-\lambda E|=0$$

求得特征值分别为 $\lambda_1=12$，$\lambda_2=8$，$\lambda_3=4$

(2). 计算特征值对应的特征向量，

$$|\Sigma-\lambda_1 E|\to\begin{bmatrix} 1 & 0 & -2 \\ 0 & \sqrt{3} & -1 \\ 0 & 0 & 0 \end{bmatrix}$$

得 $\alpha_1=\left(2\sqrt{3},1,\sqrt{3}\right)'$，同理计算 $\alpha_2$，$\alpha_3$

(3). 正交单位化

$$T_1=\frac{\alpha_1}{\|\alpha_1\|}=\left(\frac{\sqrt{3}}{2},\frac{1}{4},\frac{\sqrt{3}}{4}\right)'$$

同理计算 $T_2$，$T_3$

(4). 最后写出主成分的表达式

$$Y=T'X$$

4. 书上 P124 还有个证明题，那个需要说一下，记得哈。

# 7 因子分析

因子分析 (factor analysis) 也是一种降维、简化数据的技术。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“抽象”的变量来表示其

基本的数据结构。这几个抽象的变量被称作“因子”，能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，而因子一般是不可观测的潜在变量。

因子分析就是一种通过显在变量测评潜在变量，通过具体指标测评抽象因子的统计分析方法。

## 7.1 R 型因子分析模型

R 因子分析中的公共因子是不可直接观测但又客观存在的共同影响因素，每一个变量都可以表示成公共因子的线性函数与特殊因子之和，即

$$X_i = a_{i1}F_1 + a_{i2}F_2 + \cdots + a_{im}F_m + \varepsilon_i, i = 1, 2, \cdots, p$$

式中的 $F_1, F_2, \cdots, F_m$ 称为公共因子，$\varepsilon_i$ 称为 $X_i$ 的特殊因子。该模型可用矩阵表示为：

$$X = AF + \varepsilon$$

$a_{ij}$ 称为因子载荷，矩阵 $A$ 为因子载荷矩阵

因子分析与主成分分析有许多相似之处（因子分析的求解过程同主成分分析类似，也是从一个协方差阵出发的。），但这两种模型又存在明显的不同。主成分分析的数学模型本质上是一种线性变换，是将原始坐标变换到变异程度大的方向上去，归纳重要信息。而因子分析从本质上看是从显在变量去“提练”潜在因子的过程。

### 7.1.1 因子载荷阵的统计意义

### 7.1.2 因子载荷 $a_{ij}$ 的统计意义

对于因子模型

$$X_i = a_{i1}F_1 + a_{i2}F_2 + \cdots + a_{ij}F_j + \cdots + a_{im}F_m + \varepsilon_i$$

我们可以得到，$X_i$ 与 $F_j$ 的协方差为：

$$\mathrm{cov}(\mathrm{X_i}, \mathrm{F_j}) = \mathrm{a_{ij}}$$

$a_{ij}$ 是 $X_i$ 与 $F_j$ 的相关系数，它一方面表示 $X_i$ 对 $F_j$ 的依赖程度，绝对值越大，密切程度越高；另一方面也反映了变量 $X_i$ 对公共因子 $F_j$ 的相对重要性。

### 7.1.3 变量共同度 $h_i^2$ 的统计意义

设因子载荷矩阵为 $A$，**称第 $i$ 行元素的平方和**，即

$$h_i^2 = \sum_{j=1}^{m} a_{ij}^2 \quad i = 1, 2, \cdots, p$$

**为变量 $X_i$ 的共同度。**

### 7.1.4 公因子 $F_j$ 的方差贡献 $g_j^2$ 的统计意义

设因子载荷矩阵为 $A$，称第 $j$ 列元素的平方和，即

$$g_j^2 = \sum_{i=1}^{p} a_{ij}^2 \quad j = 1, 2, \cdots, m$$

为公共因子 $F_j$ 对 $X$ 的贡献，即 $g_j^2$ 表示同一公共因子 $F_j$ 对各变量所提供的方差贡献之总和，它是衡量每一个公共因子相对重要性的一个尺度。

## 7.2 因子载荷矩阵的求解

$$A = \left( \sqrt{\lambda_1^*} t_1^* \quad , \sqrt{\lambda_2^*} t_2^*, \quad \cdots, \quad \sqrt{\lambda_m^*} t_m^* \right)$$

## 7.3 公因子重要性的分析

### 7.3.1 因子旋转

因子分析的目标之一就是要对所提取的抽象因子的实际含义进行合理解释。有时直接根据特征根、特征向量求得的因子载荷阵难以看出公共因子的含义。这时需要通过因子旋转的方法，使每个变量仅在一个公共因子上有较大的载荷，而在其余的公共因子上的载荷比较小，至多达到中等大小。

### 7.3.2 因子得分

因子得分的估算公式

$$\widehat{F} = A' R^{-1} X$$

其中 $R$ 是 $X$ 的相关系数矩阵

例子：

$$A = \begin{bmatrix} 0.4 & 0.1 \\ 0.7 & 0.2 \\ -0.1 & 0.6 \\ 1 & 0.4 \end{bmatrix}, \quad X = \begin{bmatrix} 1 \\ 4 \\ 3 \\ 7 \end{bmatrix}, \quad R^{-1} = \begin{bmatrix} 1 & 0.2 & 0.1 & 0.4 \\ 0.2 & 1 & 0.2 & 0.6 \\ 0.1 & 0.2 & 1 & 0.7 \\ 0.4 & 0.6 & 0.7 & 1 \end{bmatrix}$$

计算 $F_1$(理科能力)，$F_2$(文科能力) 得分，并评价哪科能力强?

$$\widehat{F} = A^{'} R^{-1} X = \begin{bmatrix} 19.28 \\ 12.33 \end{bmatrix}$$

理科能力强

表 8: Correlation Matrix

| correlation | | 数学 | 物理 | 化学 | 语文 | 历史 | 英语 |
|---|---|---|---|---|---|---|---|
| | 数学 | 1 | 0.426 | 0.527 | -0.464 | -0.356 | -0.296 |
| | 物理 | 0.426 | 1 | 0.345 | -0.307 | -0.285 | -0.235 |
| | 化学 | 0.527 | 0.345 | 1 | -0.391 | -0.29 | -0.136 |
| | 语文 | -0.464 | -0.307 | -0.391 | 1 | 0.778 | 0.81 |
| | 历史 | -0.356 | -0.285 | -0.29 | 0.778 | 1 | 0.82 |
| | 英语 | -0.296 | -0.235 | -0.136 | 0.81 | 0.82 | 1 |

表 9: Communalities

| | Initial | Extraction |
|---|---|---|
| 数学 | 1 | 0.692 |
| 物理 | 1 | 0.51 |
| 化学 | 1 | 0.674 |
| 语文 | 1 | 0.865 |
| 历史 | 1 | 0.862 |
| 英语 | 1 | 0.914 |
| Extraction Method: Principal Component Analysis. | | |

表 10: Total Variance Explained

| Component | Initial Eigenvalues | | | Extraction Sums of Squared Loadings | | |
|---|---|---|---|---|---|---|
| | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % |
| 1 | 3.238 | 53.972 | 53.972 | 3.238 | 53.972 | 53.972 |
| 2 | 1.277 | 21.288 | 75.26 | 1.277 | 21.288 | 75.26 |
| 3 | 0.681 | 11.346 | 86.607 | | | |
| 4 | 0.458 | 7.634 | 94.24 | | | |
| 5 | 0.212 | 3.526 | 97.767 | | | |
| 6 | 0.134 | 2.233 | 100 | | | |

表 11: Component Matrix

| | Component | |
|---|---|---|
| | 1 | 2 |
| 数学 | -0.662 | 0.503 |
| 物理 | -0.53 | 0.478 |
| 化学 | -0.555 | 0.605 |
| 语文 | 0.9 | 0.233 |
| 历史 | 0.857 | 0.357 |
| 英语 | 0.816 | 0.498 |

表 12: Rotated Component Matrix

| | Component | |
|---|---|---|
| | 1 | 2 |
| 数学 | -0.245 | 0.795 |
| 物理 | -0.152 | 0.698 |
| 化学 | -0.099 | 0.815 |
| 语文 | 0.867 | -0.335 |
| 历史 | 0.904 | -0.209 |
| 英语 | 0.953 | -0.072 |

表 13: Component Score Coefficient Matrix

| | Component | |
|---|---|---|
| | 1 | 2 |
| 数学 | 0.064 | 0.439 |
| 物理 | 0.085 | 0.4 |
| 化学 | 0.137 | 0.484 |
| 语文 | 0.332 | -0.014 |
| 历史 | 0.378 | 0.073 |
| 英语 | 0.432 | 0.169 |

## 7.4 梨子

因子模型：

$$
\begin{aligned}
X1 &= -0.245F1 + 0.795F2 + \varepsilon_1 \\
X2 &= -0.152F1 + 0.698F2 + \varepsilon_2 \\
X3 &= -0.099F1 + 0.815F2 + \varepsilon_3 \\
X4 &= 0.867F1 - 0.335F2 + \varepsilon_4 \\
X5 &= 0.904F1 - 0.209F2 + \varepsilon_5 \\
X6 &= 0.953F1 - 0.072F2 + \varepsilon_6
\end{aligned}
$$

主成分分析模型：

$$
\begin{aligned}
Y1 &= \tfrac{-0.245}{\sqrt{3.238}}X1 + \tfrac{-0.152}{\sqrt{3.238}}X2 + \tfrac{-0.099}{\sqrt{3.238}}X3 + \tfrac{0.867}{\sqrt{3.238}}X4 + \tfrac{0.904}{\sqrt{3.238}}X5 + \tfrac{0.953}{\sqrt{3.238}}X6 \\
Y2 &= \tfrac{0.795}{\sqrt{1.277}}X1 + \tfrac{0.698}{\sqrt{1.277}}X2 + \tfrac{0.815}{\sqrt{1.277}}X3 + \tfrac{-0.335}{\sqrt{1.277}}X4 + \tfrac{-0.209}{\sqrt{1.277}}X5 + \tfrac{-0.072}{\sqrt{1.277}}X6
\end{aligned}
$$

因子得分表达式：

$$
\begin{aligned}
Y1 &= 0.065X1 + 0.084X2 + 0.136X3 + 0.333X4 + 0.378X5 + 0.432X6 \\
Y2 &= 0.441X1 + 0.398X2 + 0.485X3 - .014X4 + 0.072X5 + 0.169X6
\end{aligned}
$$

综合评价表达式：

$$
S = \frac{3.238}{3.238 + 1.277}Y1 + \frac{1.277}{3.238 + 1.277}Y2
$$

2. 试述因子分析与主成分分析的联系与区别？

答：因子分析与主成分分析的联系是：1. 两种分析方法都是一种降维、简化数据的技术。2. 两种分析的求解过程是类似的，都是从一个协方差阵出发，利用特征值、特征向量求解。如果说主成分分析是将原指标综合、归纳，那么因子分析可以说是将原指标给予分解、演绎。

因子分析与主成分分析的主要区别是：主成分分析本质上是一种线性变换，将原始坐标变换到变异程度大的方向上为止，突出数据变异的方向，归纳重要信息。而因子分析是从显在变量去提炼潜在因子的过程。此外，主成分分析不需要构造分析模型而因子分析要构造因子模型。

3. 因子分析主要可应用于哪些方面?

答：因子分析是一种通过显在变量测评潜在变量，通过具体指标测评抽象因子的统计分析方法。目前因子分析在心理学、社会学、经济学等学科中都有重要的应用。具体来说，

(1). 因子分析可以用于分类。如用考试分数将学生的学习状况予以分类。

(2). 因子分析可以用于探索潜在因素。即是探索未能观察的或不能观测的的潜在因素是什么，起的作用如何等。对我们进一步研究与探讨指示方向。在社会调查分析中十分常用。

(3). 因子分析可用于时空分解。如研究几个不同地点的不同日期的气象状况，就用因子分析将时间因素引起的变化和空间因素引起的变化分离开来从而判断各自的影响和变化规律。

(4). 因子分析可用于综合评价。

4. 设某客观现象可用 $X=(X1,X2,X3)'$ 来描述，在因子分析时，从约相关阵出发计算出特征值为 $\lambda_1=1.754,\lambda_2=1,\lambda_3=0.255$。由于 $(\lambda_1+\lambda_2)/(\lambda_1+\lambda_2+\lambda_3)\geq 85\%$，所以找前两个特征值所对应的公共因子即可，又知 $\lambda_1,\lambda_2$ 对应的**正则化特征向量**分别为 $(0.707,-0.316,0.632)^{'},(0,0.899,0.4470)^{'}$，求：

(1). 主成分模型

$$Y1=0.707X1-0.306X2+0.632X3$$
$$Y2=0X1+0.899X2+0.4470X3$$

这里要说明一下，样本的协方差矩阵求出的特征向量一定是个列向量。

(2). 计算因子载荷矩阵 $A$，并建立因子模型。

$$A=\left(\sqrt{\lambda_1}t_1,\sqrt{\lambda_2}t_2\right)=\begin{pmatrix}0.936 & 0\\ -0.418 & 0.899\\ 0.837 & 0.447\end{pmatrix}$$

因子模型：

$$\begin{pmatrix}X1\\X2\\X3\end{pmatrix}=\begin{pmatrix}0.936 & 0\\ -0.418 & 0.899\\ 0.837 & 0.447\end{pmatrix}\begin{pmatrix}F1\\F2\end{pmatrix}+\begin{pmatrix}\varepsilon_1\\\varepsilon_2\end{pmatrix}$$

(3). 计算共同度 $h_i^2(i=1,2,3)$

$$h_1^2=0.936^2+0^2=0.876$$
$$h_2^2=-0.418^2+0.899^2=0.983$$
$$h_3^2=0.837^2+0.447^2=0.900$$

(4). 计算第一公因子对 $X$ 的“贡献率”。

$$g_1^2=\lambda_1=1.754$$

# 8 相应分析

相应分析 (correspondence analysis) 也叫对应分析，其特点是它所研究的变量可以是定性的。通常意义下的相应分析，是指对两个定性变量（因素）的多种水平进行相应性研究。

在社会、经济以及其他领域中，进行数据分析时经常要处理**因素与因素之间的关系，及因素内部各个水平之间的相互关系。**

## 8.1 丽子

1. 什么是相应分析？它与因子分析有何关系？

相应分析是指两个定性变量的多种水平进行相应性研究。其特点是它所研究的变量可以是定性的。

相应分析与因子分析的关系是：在进行相应分析过程中，计算出过渡矩阵后，要分别对变量和样本进行因子分析。因此，因子分析是相应分析的基础。

2. 试述相应分析的基本思想和步骤?

相应分析基本思想是指对两个定性变量的多种水平进行分析。

相应分析基本步骤：(1). 建立列联表 (2). 通过列联表的转换，使得因素 A 和因素 B 具有对等性 (3). 对因素 B 进行因子分析 (4). 对因素 A 进行因子分析 (5). 把两个因素的各个水平的状况同时反映到具有相同坐标轴的因子平面上 (6). 根据因素 A 和因素 B 各个水平在平面图上的分布，描述两因素及各个水平之间的相关关系。

# 9 典型相关分析

$$M_1 = \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$
$$M_2 = \Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

因为 $\lambda = a'\Sigma_{12}b = Corr(U,V)$，求 $Corr(U,V)$ 最大值也就是求 $\lambda$ 的最大值，而求 $\lambda$ 的最大值又转化为求 $M_1$ 和 $M_2$ 的最大特征根。

综上所述，典型变量和典型相关系数的计算可归结为矩阵 $M_1$ 和 $M_2$ 特征根及相应特征向量的求解。如果矩阵 $M_1$ 和 $M_2$ 的秩为 $r$，则共有 $r$ 对典型变量，第 $k$ 对 $(1 \le k \le r)$ 典型变量的系数向量分别是矩阵 $M_1$ 和 $M_2$ 第 $k$ 特征根 $\lambda_k^2$ 相应的特征向量，典型相关系数为 $\lambda_k$ 。

## 9.1 狸子

1. 什么是典型相关分析? 简述其基本思想?

答：典型相关分析是研究两组变量之间相关关系的一种多元统计方法。用于揭示两组变量之间的内在联系。典型相关分析的目的是识别并量化两组变量之间的联系。将两组变量相关关系的分析转化为一组变量的线性组合与另一组变量线性组合之间的相关关系。

基本思想和主成分分析非常相似。首先在每组变量中找出变量的线性组合，使得两组的线性组合之间具有最大的相关系数。然后选取和最初挑选的这对线性组合不相关的线性组合，使其配对，并选取相关系数最大的一对，如此继续下去，直到两组变量之间的相关性被提取完毕为此。

2. 试分析一组变量的典型变量与其主成分的联系与区别?

答：一组变量的典型变量和其主成分都是经过线性变换计算矩阵特征值与特征向量得出的。主成分分析只涉及一组变量的相互依赖关系而典型相关则扩展到两组变量之间的相互依赖关系之中，度量了这两组变量之间联系的强度。