

# 填空题

1.  $T^2$  分布与  $F$  分布的关系为：若  $T^2 \sim T^2(p, n)$  则  $\frac{n-p+1}{np} T^2 \sim F(p, n-p+1)$

若  $T^2 \sim T^2(p, n)$ ，其中  $p$  为自由度参数， $n$  为样本容量，则与  $T^2$  分布相关的  $F$  分布的参数为  $(p, n-p+1)$ 。因此，给定  $T^2 \sim T^2(p, n)$ ，有  $\frac{n-p+1}{np} T^2 \sim F(p, n-p+1)$

这表示  $\frac{n-p+1}{np} T^2$  服从自由度为  $p$  和  $n-p+1$  的  $F$  分布。

$$n \rightarrow n$$

$$p \rightarrow p+1$$

$$(p, n) \rightarrow (p, n - [p+1])$$

2. 设  $X_{(1)}, \dots, X_{(n)}$  是来自正态总体  $X \sim N_p(\mu, \Sigma)$  的随机样本，但  $\Sigma = \Sigma_0$  已知时进行单总体均值向量的假设检验， $\bar{X} \sim N_p\left(\mu, \frac{\Sigma_0}{n}\right)$ ， $\sqrt{n}(\bar{X} - \mu) \sim N_p(0, \Sigma_0)$ ，则当  $H_0$  成立时，统计量  $T^2 = n \cdot (\bar{X} - \mu_0)^T \Sigma^{-1} (\bar{X} - \mu_0) \sim \chi^2(p)$

在这个情境下，我们要考虑 Hotelling's  $T^2$  统计量，其定义为：

$$T^2 = n \cdot (\bar{X} - \mu_0)^T \Sigma^{-1} (\bar{X} - \mu_0)$$

其中， $\bar{X}$  是样本均值， $\mu_0$  是假设的总体均值， $\Sigma_0$  是总体协方差矩阵。由题意可知

$$\bar{X} \sim N_p\left(\mu, \frac{\Sigma_0}{n}\right)，因此，$$

$$\sqrt{n}(\bar{X} - \mu_0) \sim N_p(0, \Sigma_0)$$

代入  $T^2$  的定义，有：

$$T^2 = n \cdot (\bar{X} - \mu_0)^T \Sigma^{-1} (\bar{X} - \mu_0) = (\sqrt{n}(\bar{X} - \mu_0))^T (\Sigma^{-1}) (\sqrt{n}(\bar{X} - \mu_0))$$

由于  $\sqrt{n}(\bar{X} - \mu_0) \sim N_p(0, \Sigma_0)$ ，我们可以令  $\sqrt{n}(\bar{X} - \mu_0) = Z$ ，其中

$Z \sim N_p(0, \Sigma_0)$ 。然后，我们可以重写  $T^2$  为：

$$T^2 = Z^T (\Sigma^{-1}) Z$$

由于  $Z \sim N_p(0, \Sigma_0)$ ，我们有  $Z^T \Sigma_0^{-1} Z \sim \chi_p^2$ ，因此：

$$T^2 \sim \chi_p^2$$

所以，当  $H_0$  成立时， $T^2$  统计量服从自由度为  $p$  的卡方分布，即  $T^2 \sim \chi_p^2$ 。

当  $H_0$  成立时，我们知道  $T^2$  统计量服从自由度为  $p$  的卡方分布，即  $T^2 \sim \chi_p^2$ 。如果我们要进行转换，使得  $T^2$  统计量服从  $T^2$  分布，我们可以考虑将  $T^2$  除以其自由度。

具体而言，如果  $T^2 \sim \chi_p^2$ ，那么我们可以定义一个新的统计量  $T_{\text{new}}^2$ ：

$$T_{\text{new}}^2 = \frac{T^2}{p}$$

这样的转换会使得  $T_{\text{new}}^2$  统计量服从自由度为  $p$  的 Hotelling's  $T^2$  分布。因此, 当  $H_0$  成立时,  $T_{\text{new}}^2$  统计量服从  $T^2(p, n)$  分布。

- 因子分析中, 因子载荷量的统计意义是利用标准化的数据计算的相关阵  $R$  的特征值所对应的特征向量。
- 常用的判别方法有距离判别、贝叶斯判别、费希尔判别。
- 设  $A$  是载荷矩阵, 则衡量 (公共) 因子重要性的一个量是  $A$  中列元素的平方和。
- 设  $x, y$  是从均值向量分别是  $\mu$ , 协方差阵为  $\Sigma$  的总体中抽取的两个样本, 则  $x$  与  $y$  之间的马氏距离  $d_{(x,y)} = \sqrt{(x - y)' \Sigma^{-1} (x - y)}$ 。
- 在 R 语言中散点图调用的函数为 `plot`。
- 设  $X \sim N_p(\mu, \Sigma)$ ,  $B$  为  $s \times p$  的常数矩阵,  $d$  为  $s$  维常数向量, 令  $Y = BX + d$ , 则  $Y \sim N_s(B\mu + d, B\Sigma B')$ 。
- 设  $X \sim N_2(\mu, \Sigma)$ , 其中  $X = (x_1, x_2)$   $\mu = (\mu_1, \mu_2)$   $\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ , 则  $x_1 + x_2$  与  $x_1 - x_2$  独立。

这两个随机变量的协方差矩阵是:

$$\text{cov}(x_1 + x_2, x_1 - x_2) = \text{cov}(x_1, x_1) - \text{cov}(x_1, x_2) + \text{cov}(x_2, x_1) - \text{cov}(x_2, x_2)$$

考虑到协方差矩阵  $\Sigma$  的形式:

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

我们有:

$$\text{cov}(x_1 + x_2, x_1 - x_2) = 0$$

协方差为 0, 相互独立

- $P$  个随机变量的总方差为协方差矩阵  $\Sigma$  的特征值之和。
- 样本主成分的总方差与原变量样本的总方差相等。

## 判断题

- 对于两组的判别, 最大后验概率法判别规则可使两个误判概率之和达到最小。×
- $R$  型聚类分析的分类对象是样品。×
- 在聚类分析中使用主成分的目的是为了更好地计算样品间的距离。×
- 在正交因子模型中, 因子载荷完全决定了原始变量之间的协方差或相关系数。✓
- 对应分析图中行点和列点之间的距离是没有意义的。×
- 费希尔判别既可用于分类也可用于分离, 且在实际应用中更多地用于分离。✓
- 马氏距离不受变量单位的影响。✓

8. 正交旋转将改变共性方差。×
9. 对应分析图中相近的行点和列点之间的关联性与它们离原点的远近无关。×
10. 主成分之间互不相关，它们的方向彼此垂直。✓

## 计算题

1. 设  $X \sim N_3(\mu, 2I_3)$ , 已知  $\mu = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}$   $A = \begin{bmatrix} 0.5 & -1 & 0.5 \\ -0.5 & 0 & -0.5 \end{bmatrix}$   $d = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$  试求

$Y = AX + d$  的分布。

$$\text{解: } Y = AX + d \quad \therefore Y \sim N_3(AN + d, A\Sigma A')$$

$$\text{其中 } AN + d = \begin{bmatrix} 0.5 & -1 & 0.5 \\ -0.5 & 0 & -0.5 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 2 \end{bmatrix} \\ = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

$$A\Sigma A' = \begin{bmatrix} 0.5 & -1 & 0.5 \\ -0.5 & 0 & -0.5 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 0.5 & -0.5 \\ -1 & 0 \\ 0.5 & -0.5 \end{bmatrix} \\ = \begin{bmatrix} 3 & -1 \\ -1 & 1 \end{bmatrix}$$

$$\therefore Y = AX + d \sim N_3\left(\begin{bmatrix} 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 3 & -1 \\ -1 & 1 \end{bmatrix}\right)$$

2. 设  $X = (x_1, x_2, x_3)' \sim N_3(\mu, \Sigma)$ , 其中  $\mu = \begin{bmatrix} 2 \\ -3 \\ 1 \end{bmatrix}$   $\Sigma = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 2 \end{bmatrix}$ 。

1. 试求  $3x_1 - x_2 + x_3$  的分布
2. 求条件分布  $(x_1, x_2) | x_3 = t_3$

解: (1). 令  $Y = 3x_1 - x_2 + x_3 = (3, -1, 1)x = Ax \sim N(A\mu, A\Sigma A')$

$$\text{其中 } A\mu = (3, -1, 1) \begin{bmatrix} 2 \\ -3 \\ 1 \end{bmatrix} = 10$$

$$A\Sigma A' = (3, -1, 1) \begin{bmatrix} 1 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ -1 \\ 1 \end{bmatrix} = 10$$

$$\therefore Y = 3x_1 - x_2 + x_3 \sim N(10, 10)$$

$$(2). \text{ 令 } \mu = \begin{bmatrix} \mu^{(1)} \\ \mu^{(2)} \end{bmatrix} \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

$$(x_1, x_2) | x_3 = t_3 \sim N(\mu_{1.2}, \Sigma_{11.2})$$

$$\begin{aligned} \text{其中 } \mu_{1.2} &= \mu^{(1)} + \Sigma_{12} \Sigma_{22}^{-1} (x^{(2)} - \mu^{(2)}) = \begin{bmatrix} 2 \\ -3 \end{bmatrix} + \begin{bmatrix} 1 \\ 2 \end{bmatrix} \frac{1}{2} (t_3 - 1) \\ &= \begin{bmatrix} \frac{3}{2} + \frac{1}{2} t_3 \\ -4 + t_3 \end{bmatrix} \end{aligned}$$

$$\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} = \begin{bmatrix} 1 & 1 \\ 1 & 3 \end{bmatrix} - \begin{bmatrix} 1 \\ 2 \end{bmatrix} \frac{1}{2} \begin{bmatrix} 1 & 2 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{bmatrix}$$

$$\therefore (x_1, x_2) | x_3 = t_3 \sim N\left(\begin{bmatrix} \frac{3}{2} + \frac{1}{2} t_3 \\ -4 + t_3 \end{bmatrix}, \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{bmatrix}\right)$$

3. 试检验假设  $H_0: \mu = \mu_0 \leftrightarrow H_1: \mu \neq \mu_0$  ( $\alpha = 0.05$ )

设误判损失、先验概率及目睹概率值如下图

|       | $G_1$            | $G_2$             | $G_3$            |
|-------|------------------|-------------------|------------------|
| $G_1$ | $L(1 1) = 0$     | $L(2 1) = 10$     | $L(3 1) = 200$   |
| $G_2$ | $L(1 2) = 20$    | $L(2 2) = 0$      | $L(3 2) = 100$   |
| $G_3$ | $L(1 3) = 60$    | $L(2 3) = 50$     | $L(3 3) = 0$     |
| 先验概率  | $p_1 = 0.05$     | $p_2 = 0.65$      | $p_3 = 0.3$      |
| 概率密度  | $f_1(x_0) = 0.1$ | $f_2(x_0) = 0.63$ | $f_3(x_0) = 2.4$ |

试根据最小ECM判别准则判断  $x_0$  应归属于哪一个总体。



解: 当  $i=1$  时,  $h_1(x_0) = P_2 L(1|2) f_2(x_0) + P_3 L(1|3) f_3(x_0)$   
 $= 0.65 \times 20 \times 0.63 + 0.3 \times 60 \times 2.4 = 51.39$

当  $i=2$  时,  $h_2(x_0) = P_1 L(2|1) f_1(x_0) + P_3 L(2|3) f_3(x_0)$   
 $= 0.05 \times 10 \times 0.1 + 0.3 \times 50 \times 2.4 = 36.05$

当  $i=3$  时,  $h_3(x_0) = P_1 L(3|1) f_1(x_0) + P_2 L(3|2) f_2(x_0)$   
 $= 0.05 \times 200 \times 0.1 + 0.65 \times 100 \times 2.4 = 41.95$

由ECM判别可知,  $h_2(x_0)$  最小,  $\therefore x_0 \in G_2$

4. 试检验假设  $H_0: \mu = \mu_0 \leftrightarrow H_1: \mu \neq \mu_0$  ( $\alpha = 0.05$ )

设误判损失、先验概率及目睹概率值如下图

|       | $G_1$             | $G_2$             | $G_3$             |
|-------|-------------------|-------------------|-------------------|
| $G_1$ | $L(1 1 = 0)$      | $L(2 1 = 30)$     | $L(3 1 = 60)$     |
| $G_2$ | $L(1 2 = 350)$    | $L(2 2 = 0)$      | $L(3 2 = 200)$    |
| $G_3$ | $L(1 3 = 120)$    | $L(2 3 = 70)$     | $L(3 3 = 0)$      |
| 先验概率  | $p_1 = 0.5$       | $p_2 = 0.2$       | $p_3 = 0.3$       |
| 概率密度  | $f_1(x_0) = 0.36$ | $f_2(x_0) = 0.12$ | $f_3(x_0) = 0.67$ |

试根据最小ECM判别准则判断  $x_0$  应归属于哪一个总体。

$$\begin{aligned}\text{解: 当 } i=1 \text{ 时, } h_1(x_0) &= P_2 L(1|2) f_2(x_0) + P_3 L(1|3) f_3(x_0) \\ &= 0.2 \times 350 \times 0.12 + 0.3 \times 120 \times 0.67 = 32.52\end{aligned}$$

$$\begin{aligned}\text{当 } i=2 \text{ 时, } h_1(x_0) &= P_1 L(2|1) f_1(x_0) + P_3 L(2|3) f_3(x_0) \\ &= 0.5 \times 30 \times 0.36 + 0.3 \times 70 \times 0.67 = 19.47\end{aligned}$$

$$\begin{aligned}\text{当 } i=3 \text{ 时, } h_1(x_0) &= P_1 L(3|1) f_1(x_0) + P_2 L(3|2) f_2(x_0) \\ &= 0.5 \times 60 \times 0.36 + 0.2 \times 200 \times 0.12 = 15.6\end{aligned}$$

由ECM判别可知,  $h_3(x_0)$  最小,  $\therefore x_0 \in G_3$

5. 设有5个样本数据1, 2, 5, 7, 10。使用最长, 最短距离法进行聚类分析, 写出聚类步骤中各步距离矩阵, 并画出谱系聚类图。

解: 最长距离

(1)  $D^{(0)}$ 

|       | $G_1$ | $G_2$ | $G_3$ | $G_4$ | $G_5$ |
|-------|-------|-------|-------|-------|-------|
| $G_1$ | 0     |       |       |       |       |
| $G_2$ | 1     | 0     |       |       |       |
| $G_3$ | 4     | 3     | 0     |       |       |
| $G_4$ | 6     | 5     | 2     | 0     |       |
| $G_5$ | 9     | 8     | 5     | 3     | 0     |

(2)  $D^{(1)}$  将  $G_1, G_2$  聚为一类  $G_6$ 

|       | $G_3$ | $G_4$ | $G_5$ | $G_6$ |                  |
|-------|-------|-------|-------|-------|------------------|
| $G_3$ | 0     |       |       |       | $G_6 = \{1, 2\}$ |
| $G_4$ | 2     | 0     |       |       |                  |
| $G_5$ | 5     | 3     | 0     |       |                  |
| $G_6$ | 4     | 6     | 9     | 0     |                  |

(3)  $D^{(2)}$  将  $G_3, G_4$  聚为一类  $G_7$ 

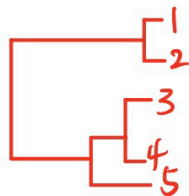
|       | $G_5$ | $G_6$ | $G_7$ |                  |
|-------|-------|-------|-------|------------------|
| $G_5$ | 0     |       |       | $G_7 = \{3, 4\}$ |
| $G_6$ | 9     | 0     |       |                  |
| $G_7$ | 5     | 6     | 0     |                  |

(4)  $D^{(3)}$  将  $G_5, G_7$  聚为一类  $G_8$ 

|       | $G_6$ | $G_8$ |                  |
|-------|-------|-------|------------------|
| $G_6$ | 0     |       | $G_8 = \{5, 7\}$ |
| $G_8$ | 9     | 0     |                  |

(5) 将  $G_6, G_8$  聚为一类  $G_9$ , 算法结束

谱系图



最短距离法.

 $D^{(0)}$ 

|       | $G_1$ | $G_2$ | $G_3$ | $G_4$ | $G_5$ |
|-------|-------|-------|-------|-------|-------|
| $G_1$ | 0     |       |       |       |       |
| $G_2$ | 1     | 0     |       |       |       |
| $G_3$ | 4     | 3     | 0     |       |       |
| $G_4$ | 6     | 5     | 2     | 0     |       |
| $G_5$ | 9     | 8     | 5     | 3     | 0     |

 $D^{(1)}$  $G_6 = \{1, 2\}$ 

|       | $G_3$ | $G_4$ | $G_5$ | $G_6$ |
|-------|-------|-------|-------|-------|
| $G_3$ | 0     |       |       |       |
| $G_4$ | 2     | 0     |       |       |
| $G_5$ | 5     | 3     | 0     |       |
| $G_6$ | 3     | 5     | 8     | 0     |

 $D^{(2)}$  $G_7 = \{5, 7\}$ 

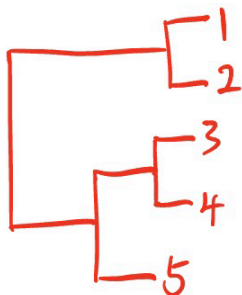
|       | $G_5$ | $G_6$ | $G_7$ |
|-------|-------|-------|-------|
| $G_5$ | 0     |       |       |
| $G_6$ | 8     | 0     |       |
| $G_7$ | 3     | 3     | 0     |

 $D^{(3)}$  $G_8 = \{5, 7, 10\}$ 

|       | $G_6$ | $G_8$ |
|-------|-------|-------|
| $G_6$ | 0     |       |
| $G_8$ | 3     | 0     |

 $G_9 = \{5, 7, 10, 1, 2\}$ 

谱系图.



6. 设三元总体  $X$  的协方差阵为  $\Sigma = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 6 \end{bmatrix}$ , 从  $\Sigma$  出发, 求总体主成分  $Z_1, Z_2, Z_3$ , 并求前

两个主成分的累积贡献率。

解:  $|\lambda I - \Sigma| = \begin{vmatrix} \lambda-1 & 0 & 0 \\ 0 & \lambda-3 & 0 \\ 0 & 0 & \lambda-6 \end{vmatrix} = (\lambda-1)(\lambda-3)(\lambda-6) = 0$

解得特征值  $\lambda_1=6, \lambda_2=3, \lambda_3=1$

当  $\lambda_1=6$  时,  $(\lambda I - \Sigma) = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$  解得特征向量  $u_1 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$

当  $\lambda_2=3$  时,  $(\lambda I - \Sigma) = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -3 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$  解得特征向量  $u_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$

当  $\lambda_3=1$  时,  $(\lambda I - \Sigma) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -5 \end{bmatrix} \rightarrow \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$  解得特征向量  $u_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$

所以第一主成分:  $z_1 = u_1'x = x_3$

第二主成分:  $z_2 = u_2'x = x_2$

第三主成分:  $z_3 = u_3'x = x_1$

前两个主成分的累积贡献率:  $\frac{6+3}{6+3+1} = 0.9$

这里有个错误的地方,  $u_3 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$

7. 设三元总体  $X$  的协方差阵为  $\Sigma = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$ , 从  $\Sigma$  出发, 求第一主成分  $Z_1$  的贡献率。

$$\text{解: } |\lambda I - \Sigma| = \begin{vmatrix} \lambda-1 & -2 \\ -2 & \lambda-1 \end{vmatrix} = (\lambda-1)^2 - 4 = (\lambda+1)(\lambda-3) = 0$$

解得特征值  $\lambda_1 = 3, \lambda_2 = -1$

$$\text{当 } \lambda_1 = 3 \text{ 时, } (3I - \Sigma) = \begin{bmatrix} 2 & -2 \\ -2 & 2 \end{bmatrix} \rightarrow \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -x_1 + x_2 \\ x_1 - x_2 \end{bmatrix} = 0$$

$$\text{解得 } u_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\text{当 } \lambda_2 = -1 \text{ 时, } (-I - \Sigma) = \begin{bmatrix} -2 & -2 \\ -2 & -2 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} x_1 + x_2 \\ x_1 + x_2 \end{bmatrix} = 0$$

$$\text{解得 } u_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

所以第一主成分:  $z_1 = u_1'x = x_1 + x_2$

第二主成分:  $z_2 = u_2'x = x_1 - x_2$

第一主成分的累积贡献率为  $\frac{3}{3+1-1} = 0.75$

#### 8. 简述费希尔判别的基本思想。

费希尔 (Fisher) 判别的基本思想是投影, 或降维。对于来自不同总体 (类) 的高维数据, 选择若干个好的投影方向将它们投影为低维数据, 使得这些来自不同类的低维数据之间有比较清晰的界限。对于新样品对应的高维数据点, 也将其以同样方向投影为一个低维数据点, 然后再利用一般的距离判别方法判断其属于哪一类。而衡量类与类之间是否分开的方法借助于一元方差分析的思想。

#### 9. 简述 $k$ 均值聚类法的基本思想及该算法的具体步骤。

$K$  均值聚类法又称为逐步聚类法, 其基本思想是, 开始时先选择一批凝聚点或给出一个初始分类, 让样品按照某种原则向凝聚点凝聚, 并对凝聚点进行不断的更新或迭代, 直至分类比较合理或迭代达到稳定为止, 这样就形成一个最终的分类结果。

$k$  均值聚类算法的具体步骤如下:

1. 第一步, 指定聚类数目  $k$ 。在  $k$  均值聚类过程中, 要求首先给出聚类数目。
2. 第二步, 确定  $k$  个初始中心凝聚点, 或者将所有样品分成  $k$  个初始类, 然后将这  $k$  个类的重心 (均值) 作为初始中心凝聚点。
3. 第三步, 根据最近原则进行聚类。依次计算每个观测点到  $k$  个中心凝聚点的距离, 并按照离中心凝聚点最近的原则, 将所有样品分配到最近的类中, 形成  $k$  个类。



4. 第四步，重新确定 $k$ 个类的中心点。依次计算各类的重心，即各类中所有样品点的均值，并以重心点作为 $k$ 个类的新的中心凝聚点。
5. 第五步，判断是否已经满足聚类算法的迭代终止条件，如果未满足则返回到第三步，并重复上述过程，直至满足迭代终止条件。

#### 10. 主成分分析的基本思想

主成分分析就是一种通过降维技术把多个指标约化为少数几个综合指标的统计分析方法。其基本思想是：设法将原来众多具有一定相关性的指标，重新组合成一组新的相互无关的综合指标来代替原来指标。数学上的处理就是将原来 $p$ 个指标的若干个线性组合作为新的指标。

#### 11. 因子分析的基本思想

基本思想：把每个研究变量分解为几个影响因素变量，将每个原始变量分解成两部分因素，一部分是由所有变量共同具有的少数几个公共因子组成的，另一部分是每个变量独自具有的因素，即特殊因子。

#### 12. 统聚类法的基本思想基本思想

首先将每个样品视为单点类，即每个样品自成一类。然后逐次进行类的合并，每次将具有最小距离的两个类合并在一起，合并后重新计算类与类之间的距离，这个过程一直继续到所有的样品归为一类为止，并把这个过程画成一张系统聚类图。

13. 为研究某地区的综合发展状况，研究人员收集了该地区10年的经济发展、社会状况等方面的统计数据，包括总人口数量( $x_1$ )，GDP( $x_2$ )，社会固定资产投资( $x_3$ )，城市化水平( $x_4$ )，人均居住面积( $x_5$ )，客运量( $x_6$ )，如图所示。

|    | $x_1$  | $x_2$   | $x_3$  | $x_4$ | $x_5$ | $x_6$   |
|----|--------|---------|--------|-------|-------|---------|
| 1  | 7296.0 | 292.13  | 69.97  | 0.095 | 5.7   | 12208.0 |
| 2  | 7395.0 | 346.57  | 79.6   | 0.099 | 6.1   | 12682.0 |
| 3  | 7494.0 | 395.38  | 85.0   | 0.103 | 6.11  | 13109.0 |
| 4  | 7564.0 | 459.83  | 96.46  | 0.107 | 6.65  | 14839.0 |
| 5  | 7637.0 | 581.56  | 140.15 | 0.123 | 6.9   | 17309.0 |
| 6  | 7711.0 | 680.46  | 194.33 | 0.132 | 7.77  | 19772.0 |
| 7  | 7818.0 | 742.05  | 223.08 | 0.125 | 9.15  | 26459.0 |
| 8  | 7958.0 | 892.29  | 297.77 | 0.132 | 9.61  | 5209.0  |
| 9  | 8061.0 | 1117.66 | 369.82 | 0.162 | 9.96  | 29035.0 |
| 10 | 8160.0 | 1293.94 | 305.54 | 0.182 | 10.25 | 30718.0 |

1. 试对该数据资料使用主成分法、主因子法、极大似然法求因子载荷阵的估计值。（累计贡献率  $\geq 90\%$ ）

```
1 > EcoData <- read.csv("EcoData.csv", head=TRUE)
2 > EcoData <- EcoData[, -1]
```

```

3 > print(EcoData)
4      X1      X2      X3      X4      X5      X6
5 1  7296  292.13  69.97  0.095  5.70 12208
6 2  7395  346.57  79.60  0.099  6.10 12682
7 3  7494  395.38  85.00  0.103  6.11 13109
8 4  7564  459.83  96.46  0.107  6.65 14839
9 5  7637  581.56 140.15  0.123  6.90 17309
10 6  7711  680.46 194.33  0.132  7.77 19772
11 7  7818  742.05 223.08  0.125  9.15 26459
12 8  7958  892.29 297.77  0.132  9.61 25209
13 9  8061 1117.66 369.82  0.162  9.96 29035
14 10 8160 1293.94 305.54  0.182 10.25 30718
15 > R <- cor(EcoData)
16 > library("psych")
17 > # 主成分法
18 > pc <- principal(r=R,nfactors=2,rotate="none");pc
19 Principal Components Analysis
20 Call: principal(r = R, nfactors = 2, rotate = "none")
21 Standardized loadings (pattern matrix) based upon correlation matrix
22      PC1   PC2   h2    u2 com
23 X1 0.99   0.00 0.98 0.0171 1.0
24 X2 0.99   0.12 1.00 0.0022 1.0
25 X3 0.97  -0.14 0.97 0.0322 1.0
26 X4 0.96   0.29 1.00 0.0024 1.2
27 X5 0.98  -0.16 0.99 0.0052 1.1
28 X6 0.99  -0.10 0.98 0.0183 1.0
29
30                      PC1  PC2
31 SS loadings              5.77 0.15
32 Proportion Var           0.96 0.03
33 Cumulative Var           0.96 0.99
34 Proportion Explained     0.97 0.03
35 Cumulative Proportion   0.97 1.00
36
37 Mean item complexity = 1.1
38 Test of the hypothesis that 2 components are sufficient.
39
40 The root mean square of the residuals (RMSR) is 0.01
41
42 Fit based upon off diagonal values = 1
43 > # 主因子法
44 > fa <- fa(r=R,nfactors=2,fm="pa",rotate="none");fa
45 Factor Analysis using method = pa
46 Call: fa(r = R, nfactors = 2, rotate = "none", fm = "pa")
47 Standardized loadings (pattern matrix) based upon correlation matrix
48      PA1   PA2   h2    u2 com
49 X1 0.99   0.00 0.98 0.02136 1.0
50 X2 0.99   0.11 1.00 0.00049 1.0
51 X3 0.97  -0.10 0.94 0.05797 1.0
52 X4 0.96   0.28 0.99 0.00761 1.2

```

```

53 X5 0.99 -0.18 1.00 -0.00439 1.1
54 X6 0.98 -0.10 0.98 0.02232 1.0
55
56                                PA1  PA2
57 SS loadings                    5.75 0.14
58 Proportion Var                 0.96 0.02
59 Cumulative Var                 0.96 0.98
60 Proportion Explained           0.98 0.02
61 Cumulative Proportion          0.98 1.00
62
63 Mean item complexity = 1.1
64 Test of the hypothesis that 2 factors are sufficient.
65
66 df null model = 15 with the objective function = 18.92
67 df of the model are 4 and the objective function was 0.67
68
69 The root mean square of the residuals (RMSR) is 0
70 The df corrected root mean square of the residuals is 0
71
72 Fit based upon off diagonal values = 1
73 > # 极大似然法
74 > f1 <- factanal(EcoData,2,rot="none");f1
75 Call:
76 factanal(x = EcoData, factors = 2, rotation = "none")
77
78 Uniquenesses:
79      X1      X2      X3      X4      X5      X6
80 0.019 0.005 0.055 0.005 0.005 0.016
81
82 Loadings:
83      Factor1 Factor2
84 X1 0.989
85 X2 0.996
86 X3 0.960 0.155
87 X4 0.970 -0.234
88 X5 0.974 0.216
89 X6 0.979 0.160
90
91                                Factor1 Factor2
92 SS loadings                    5.739 0.158
93 Proportion Var                 0.957 0.026
94 Cumulative Var                 0.957 0.983
95
96 Test of the hypothesis that 2 factors are sufficient.
97 The chi square statistic is 5.58 on 4 degrees of freedom.
98 The p-value is 0.233
99 >

```

2. 写出三种方法下的因子模型。

$$x_1 = 0.99f_1$$

$$x_2 = 0.99f_1 + 0.12f_2$$

$$x_3 = 0.97f_1 - 0.14f_2$$

■ 主成分法:

$$x_4 = 0.96f_1 + 0.29f_2$$

$$x_5 = 0.98f_1 - 0.16f_2$$

$$x_6 = 0.99f_1 - 0.1f_2$$

$$x_1 = 0.99f_1$$

$$x_2 = 0.99f_1 + 0.11f_2$$

$$x_3 = 0.97f_1 - 0.1f_2$$

■ 主因子法:

$$x_4 = 0.96f_1 + 0.28f_2$$

$$x_5 = 0.99f_1 - 0.18f_2$$

$$x_6 = 0.98f_1 - 0.1f_2$$

$$x_1 = 0.989f_1$$

$$x_2 = 0.996f_1$$

$$x_3 = 0.96f_1 + 0.155f_2$$

■ 极大似然法:

$$x_4 = 0.97f_1 - 0.234f_2$$

$$x_5 = 0.974f_1 - 0.216f_2$$

$$x_6 = 0.979f_1 - 0.16f_2$$

3. 写出三种方法中的因子方差，方差贡献率及方差累计贡献率。

| 公因子 | 主成分法方差 | 方差贡献率 | 方差累计贡献率 | 主因子法方差 | 方差贡献率 | 方差累计贡献率 | 极大似然法方差 | 方差贡献率 | 方差累计贡献率 |
|-----|--------|-------|---------|--------|-------|---------|---------|-------|---------|
| 1   | 5.77   | 0.96  | 0.96    | 5.75   | 0.96  | 0.96    | 5.739   | 0.957 | 0.957   |
| 2   | 0.15   | 0.03  | 0.99    | 0.14   | 0.02  | 0.98    | 0.158   | 0.026 | 0.983   |

14. 为了分析媒体网站的定位，对中国媒体网站进行评价。根据Alexa网站提供的评价指标，选取了 5 项指标作为媒体网站的评价标准，评价数据如图所示。

| 网站    | 流量 | 访问量 | 被连接数 | 速度 | 浏览页面数 |
|-------|----|-----|------|----|-------|
| 天极网   | 3  | 4   | 8    | 1  | 2     |
| 太平洋电脑 | 4  | 10  | 10   | 2  | 4     |
| 搜狐    | 5  | 95  | 20   | 3  | 4     |
| 新浪    | 5  | 114 | 20   | 1  | 5     |
| 网易    | 5  | 84  | 20   | 2  | 5     |
| TOM   | 4  | 44  | 4    | 1  | 5     |
| 中华网   | 4  | 14  | 10   | 2  | 5     |
| CCTV  | 4  | 12  | 15   | 3  | 3     |
| 新华网   | 1  | 1   | 10   | 2  | 1     |
| 人民网   | 3  | 7   | 8    | 3  | 2     |

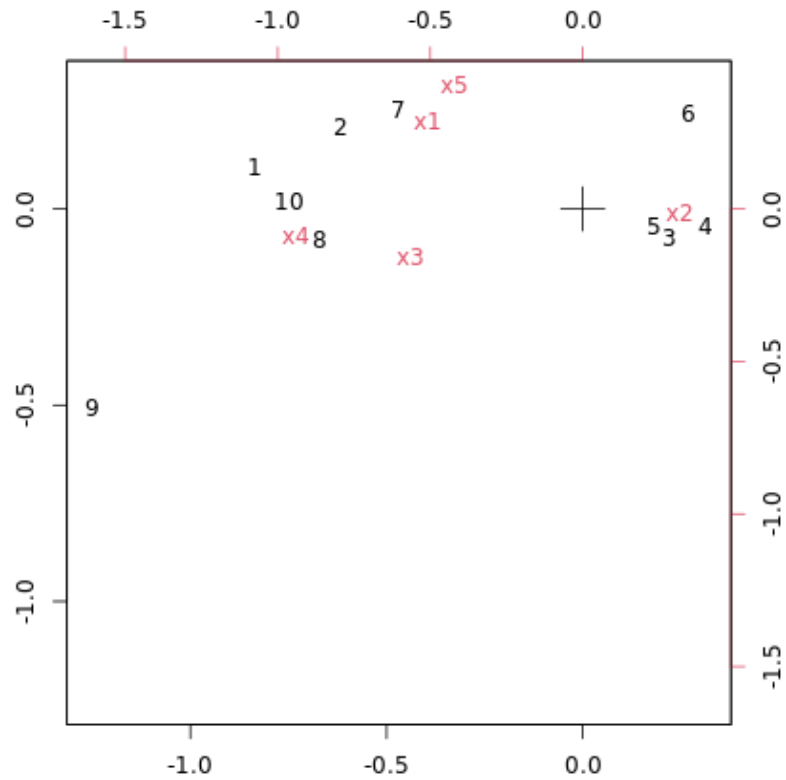
1. 取 $m = 2$ ，写出 $Q$ 型与 $R$ 型的因子载荷矩阵。

```

1 > data <- read.csv("Alexa.csv", head=TRUE)
2 > data <- data[, -1]
3 > library("MASS")
4 > ca1 <- corresp(data, 2)
5 > ca1
6 First canonical correlation(s): 0.4333964 0.1433618
7
8 Row scores:
9           [,1]      [,2]
10 [1,] -1.9322470  0.7437385
11 [2,] -1.4227035  1.4687414
12 [3,]  0.5117460 -0.5152798
13 [4,]  0.7261601 -0.3201458
14 [5,]  0.4219459 -0.3100543
15 [6,]  0.6266890  1.6841660
16 [7,] -1.0873964  1.7653308
17 [8,] -1.5429105 -0.5296207
18 [9,] -2.8878094 -3.5136091
19 [10,] -1.7240602  0.1395799
20
21 Column scores:
22           [,1]      [,2]
23 x1 -1.1696327  2.01246755
24 x2  0.7417411 -0.07074029
25 x3 -1.3044761 -1.10746752
26 x4 -2.1685686 -0.62338275
27 x5 -0.9637053  2.82395385

```

2. 试对该数据进行分析。



15. 对某地区的某类消费品的销量 $y$ 进行调查，它与下面 4 个变量相关:居民可收入支配 $x_1$ ，该类消费品评价价格指数 $x_2$ ，该产品的社会保有量 $x_3$ ，其他消费品评价价格指数 $x_4$ ，历史资料如图所示。试求



| $y$  | $x_1$ | $x_2$ | $x_3$ | $x_4$ |
|------|-------|-------|-------|-------|
| 8.4  | 82.9  | 92    | 17.1  | 94    |
| 9.6  | 88    | 93    | 21.3  | 96    |
| 10.4 | 99.9  | 96    | 25.1  | 97    |
| 11.4 | 105.3 | 94    | 29    | 97    |
| 12.2 | 117.7 | 100   | 34    | 100   |
| 14.2 | 131.0 | 101   | 40    | 101   |
| 15.8 | 148.2 | 105   | 44    | 104   |
| 17.9 | 161.8 | 112   | 49    | 109   |
| 19.6 | 174.2 | 112   | 51    | 111   |
| 20.8 | 184.7 | 112   | 53    | 111   |

1. 普通线性回归方程。

```

1 > data <- read.csv("biao001.csv", head=TRUE)
2 > lm.sol <- lm(y~x1+x2+x3+x4, data=data)
3 > summary(lm.sol)
4 Call:
5 lm(formula = y ~ x1 + x2 + x3 + x4, data = data)
6
7 Residuals:
8 1          2          3          4          5          6          7
9 0.024803  0.079476  0.012381 -0.007025 -0.288345  0.216090 -0.142085
10 0.158360
11 9          10
12 -0.135964  0.082310
13
14 Coefficients:
15             Estimate Std. Error t value Pr(>|t|)
16 (Intercept) -17.66768    5.94360   -2.973  0.03107 *
17 x1           0.09006    0.02095    4.298  0.00773 **
18 x2          -0.23132    0.07132   -3.243  0.02287 *
19 x3           0.01806    0.03907    0.462  0.66328
20 x4           0.42075    0.11847    3.552  0.01636 *
21 ---
22 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

22
23 Residual standard error: 0.2037 on 5 degrees of freedom
24 Multiple R-squared: 0.9988, Adjusted R-squared: 0.9978
25 F-statistic: 1021 on 4 and 5 DF, p-value: 1.827e-07

```

$$y = -17.66768 + 0.09006x_1 - 0.23132x_2 + 0.01806x_3 + 0.42075x_4$$

2. 利用主成分回归建立销售量 $y$ 与4个变量的回归方程。(保留4位小数)

```

1 > conomy<-read.csv("biao001.csv",head=TRUE)
2 > conomy.pr<-princomp(~x1+x2+x3+x4,data=conomy,cor=T)
3 > pre <- predict(conomy.pr)
4 > conomy $ z1 <- pre[,1]
5 > conomy $ z2 <- pre[,2]
6 > lm.sol <- lm(y~z1+z2,data=conomy)
7 > summary(lm.sol)
8 Call:
9 lm(formula = y ~ z1 + z2, data = conomy)
10
11 Residuals:
12      Min       1Q   Median       3Q      Max
13 -0.74323 -0.29223  0.01746  0.30807  0.80849
14
15 Coefficients:
16             Estimate Std. Error t value Pr(>|t|)
17 (Intercept) 14.03000     0.17125   81.927 1.06e-11 ***
18 z1           2.06119     0.08623   23.903 5.70e-08 ***
19 z2           0.62409     0.85665    0.729    0.49
20 ---
21 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
22
23 Residual standard error: 0.5415 on 7 degrees of freedom
24 Multiple R-squared: 0.9879, Adjusted R-squared: 0.9845
25 F-statistic: 285.9 on 2 and 7 DF, p-value: 1.945e-07
26 > beta <- coef(lm.sol)
27 > A <- loadings(conomy.pr)
28 > x.bar <- conomy.pr$center
29 > x.sd <- conomy.pr$scale
30 > coef <- (beta[2]*A[,1]+beta[3]*A[,2])/x.sd
31 > c(beta0,coef)
32 (Intercept)          x1          x2          x3          x4
33 -16.88460655  0.03420968  0.09376460  0.11954881  0.12360237

```

$$y = -16.8846 + 0.0342x_1 + 0.0938x_2 + 0.1195x_3 + 0.1236x_4$$

16. 将 $n = 1000$ 个人组成的样本按人的心理健康与其父母社会经济地位进行交叉分类, 心理健康程度分为:0(心理健康), 1(轻微症状), 2(中度症状), 3(心理受损)共4个等级, 父母经济地位按照高低分为A,B,C,D,E五个等级。分类结果见表。

|         | $A$ | $B$ | $C$ | $D$ | $E$ |
|---------|-----|-----|-----|-----|-----|
| 0(心理健康) | 80  | 60  | 65  | 40  | 21  |
| 1(轻微症状) | 91  | 73  | 62  | 47  | 31  |
| 2(中度症状) | 93  | 42  | 59  | 35  | 22  |
| 3(心理受损) | 46  | 51  | 32  | 24  | 26  |

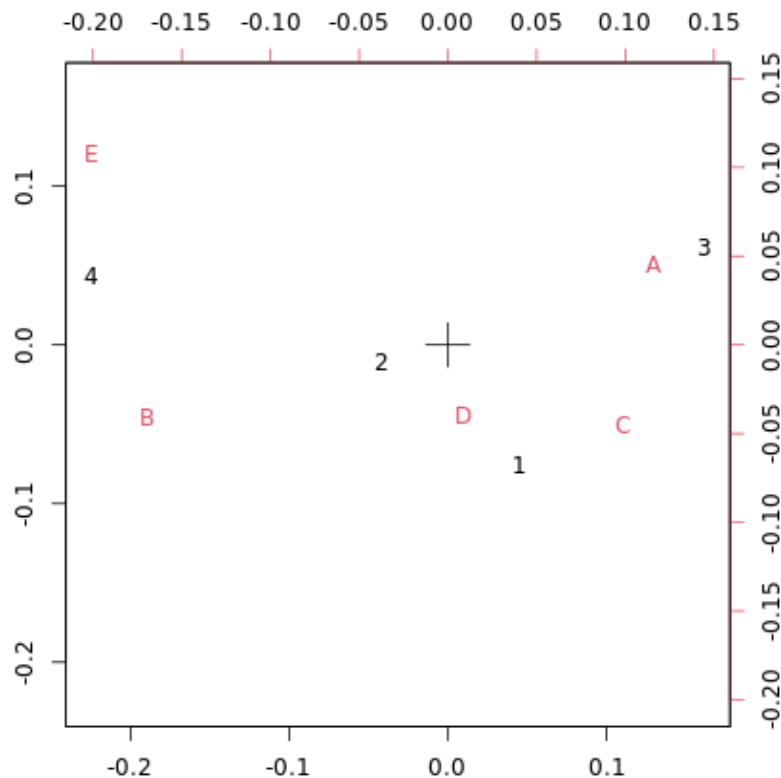
1. 取 $m = 2$ ，写出 $Q$ 型与 $R$ 型的因子载荷矩阵。

```

1 > data <- read.csv("biao002.csv", head=TRUE)
2 > data
3 index  A  B  C  D  E
4 1      0 80 60 65 40 21
5 2      1 91 73 62 47 31
6 3      2 93 42 59 35 22
7 4      3 46 51 32 24 26
8 > library("MASS")
9 > ca1 <- corresp(data, 2)
10 > ca1
11 First canonical correlation(s): 0.13611576 0.07820638
12
13 Row scores:
14      [,1]      [,2]
15 [1,]  0.5596349  1.1053713
16 [2,] -0.1772184  0.5860922
17 [3,]  0.9968915 -1.4080092
18 [4,] -1.9067299 -0.6404426
19
20 Column scores:
21      [,1]      [,2]
22 index -4.7797857 -8.8467932
23 A      0.7973588 -0.7688971
24 B     -1.1290698  0.9792588
25 C      0.7815115  0.2709922
26 D      0.1603097  0.6227149
27 E     -1.5710745 -0.7986631

```

2. 试对该数据进行分析。



17. 设标准化变量  $x = (x_1, x_2, x_3)'$  的相关矩阵  $R = \begin{bmatrix} 1 & 0.48 & 0.76 \\ 0.48 & 1 & 0.23 \\ 0.76 & 0.23 & 1 \end{bmatrix}$ ,  $R$  的特征值和相应

的单位正交特征向量为:

$$\lambda_1 = 2.0131760 \quad u_1 = (0.6624437, 0.4498215, 0.5990243)'$$

$$\lambda_2 = 0.7950835 \quad u_2 = (0.1218358, -0.853693, 0.5063243)'$$

$$\lambda_3 = 0.1917404 \quad u_3 = (0.7391376, -0.2624289, -0.6203279)'$$

1. 求  $m = 1$  时的因子模型的主成分分解。
2. 求  $m = 2$  时的因子载荷矩阵  $A$ 。
3. 计算共性方差  $h_i^2$  的值并解释其统计意义。
4. 求公因子  $f_1, f_2$  的方差贡献。
5. 计算相关系数  $cov(x_i, f_1) (i = 1, 2, 3)$ , 并说明哪个变量在公因子  $f_1$  上有最大载荷。(保留4位小数)

---


$$1. A = \sqrt{\lambda_1} u_1 = 1.4189 u_1 = (0.9399, 0.6383, 0.8499)$$

因子模型:

- $x_1 = 0.9399 f_1 + \varepsilon_1$
- $x_2 = 0.6383 f_1 + \varepsilon_2$
- $x_3 = 0.8499 f_1 + \varepsilon_3$

$$2. A = (\sqrt{\lambda_1}u_1, \sqrt{\lambda_2}u_2) = \begin{bmatrix} 0.9399 & 0.1086 \\ 0.6383 & -0.7612 \\ 0.8499 & 0.4515 \end{bmatrix}$$

3.
  - $h_1^2 = 0.9399^2 + 0.1086^2 = 0.8952$
  - $h_2^2 = 0.6383^2 + (-0.7612)^2 = 0.9869$
  - $h_3^2 = 0.8499^2 + 0.4515^2 = 0.9262$
  - $h_1^2$ 表示公因子 $f_1, f_2$ 对样本数据 $x_1$ 的方差贡献为0.8952
  - $h_2^2$ 表示公因子 $f_1, f_2$ 对样本数据 $x_2$ 的方差贡献为0.9869
  - $h_3^2$ 表示公因子 $f_1, f_2$ 对样本数据 $x_3$ 的方差贡献为0.4515
4.
  - $g_1^2 = 0.9399^2 + 0.6383^2 + 0.8499^2 = 2.0132$
  - $g_2^2 = 0.1086^2 + (-0.7612)^2 + 0.4515^2 = 0.7951$
  - $g_1^2$ 表示 $f_1$ 对变量 $x_1, x_2, x_3$ 的方差贡献为2.0132
  - $g_2^2$ 表示 $f_2$ 对变量 $x_1, x_2, x_3$ 的方差贡献为0.7951

$$5. cov(x_1, f_1) = 0.9399 \quad cov(x_2, f_1) = 0.6838 \quad cov(x_3, f_1) = 0.8499$$

比较可知 $cov(x_1, f_1) = 0.9399$ 最大, 故变量 $x_1$ 在公因子 $f_1$ 上有最大载荷。

$$18. \text{设标准化变量 } x = (x_1, x_2, x_3) \text{ 的相关矩阵 } R = \begin{bmatrix} 1 & 0.56 & 0.3 \\ 0.56 & 1 & 0.35 \\ 0.3 & 0.35 & 1 \end{bmatrix}, R \text{ 的特征值和相应}$$

的单位正交特征向量为:

$$\lambda_1 = 1.86 \quad u_1 = (0.53, 0.57, 0.61)'$$

$$\lambda_2 = 0.65 \quad u_2 = (-0.22, -0.49, 0.84)'$$

$$\lambda_3 = 0.39 \quad u_3 = (0.65, -0.64, -0.18)'$$

1. 求 $m = 1$ 时的因子模型的主成分分解。
2. 求 $m = 2$ 时的因子载荷矩阵 $A$ 。
3. 计算共性方差 $h_i^2$ 的值并解释其统计意义。
4. 求公因子 $f_1, f_2$ 的方差贡献。
5. 计算相关系数 $cov(x_i, f_1) (i = 1, 2, 3)$ , 并说明哪个变量在公因子 $f_1$ 上有最大载荷。(保留4位小数)

$$1. A = \sqrt{\lambda_1}u_1 = 1.3638u_1 = (0.7228, 0.7774, 0.8319)$$

因子模型:

- $x_1 = 0.7228f_1 + \varepsilon_1$
- $x_2 = 0.7774f_1 + \varepsilon_2$
- $x_3 = 0.8319f_1 + \varepsilon_3$

$$2. A = (\sqrt{\lambda_1}u_1, \sqrt{\lambda_2}u_2) = \begin{bmatrix} 0.7228 & -0.1774 \\ 0.7774 & -0.3950 \\ 0.8319 & 0.6772 \end{bmatrix}$$

3.   ▪  $h_1^2 = 0.7228^2 + (-0.1774)^2 = 0.5539$   
       ▪  $h_2^2 = 0.7774^2 + (-0.3950)^2 = 0.7604$   
       ▪  $h_3^2 = 0.8319^2 + 0.6772^2 = 1.1507$   
       ▪  $h_1^2$ 表示公因子 $f_1, f_2$ 对样本数据 $x_1$ 的方差贡献为0.5539  
       ▪  $h_2^2$ 表示公因子 $f_1, f_2$ 对样本数据 $x_2$ 的方差贡献为0.7604  
       ▪  $h_3^2$ 表示公因子 $f_1, f_2$ 对样本数据 $x_3$ 的方差贡献为1.1507
4.   ▪  $g_1^2 = 0.7228^2 + 0.7774^2 + 0.8319^2 = 1.8189$   
       ▪  $g_2^2 = (-0.1774)^2 + (-0.3950)^2 + 0.6712^2 = 0.6461$   
       ▪  $g_1^2$ 表示 $f_1$ 对变量 $x_1, x_2, x_3$ 的方差贡献为1.8189  
       ▪  $g_2^2$ 表示 $f_2$ 对变量 $x_1, x_2, x_3$ 的方差贡献为0.6461
5.  $cov(x_1, f_1) = 0.7228 \quad cov(x_2, f_1) = 0.7774 \quad cov(x_3, f_1) = 0.8319$   
     比较可知 $cov(x_3, f_1) = 0.8319$ 最大，故变量 $x_3$ 在公因子 $f_1$ 上有最大载荷。