《多元统计与矩阵分析》样题 参考答案

一、单项选择 (25 题，每题 2 分，共 50 分)

1. Q 型聚类是指对______________进行聚类 A

(A) 样品 （B）变量 （C） 总体 (D) 元素

2. R 型聚类是指对______________进行聚类 B

(A) 样品 （B）变量 （C） 总体 (D)元素

3. 一元正态总体中，用于参数检验的 F 分布推广到多元正态总体，对应______C

(A) Wishart 分布 (B) Hotelling T^2 分布 (C) Wilks 分布 (D) Gauss 分布

4. 有因子分析模型，

$$\boldsymbol{X} = \boldsymbol{AF} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{X} = (x_{ij})_{p\times n} \qquad \boldsymbol{A} = (a_{ij})_{p\times m} \qquad \boldsymbol{F} = (f_{ij})_{m\times n} \qquad \boldsymbol{\varepsilon} = (\varepsilon_{ij})_{p\times n}$$

则下面肯定正确的是__________ C

(A) $\sum_{j=1}^{m} a_{ij} \le 1$ （B）$\sum_{j=1}^{m} a_{ij} \ge 1$ （C）$\sum_{j=1}^{m} a_{ij}^2 \le 1$ (D) $\sum_{j=1}^{m} a_{ij}^2 \ge 1$

5. 对二元(X, Y)数据进行主成分分析，计算得到协方差矩阵如下，

$$\boldsymbol{\Sigma} = \begin{array}{c c} & \begin{array}{cc} \boldsymbol{X} & \boldsymbol{Y} \end{array} \\ \begin{array}{c} \boldsymbol{X} \\ \boldsymbol{Y} \end{array} & \begin{bmatrix} 9 & -3 \\ -3 & 7 \end{bmatrix} \end{array}$$

则该数据的分布与图 1_______最类似。 C

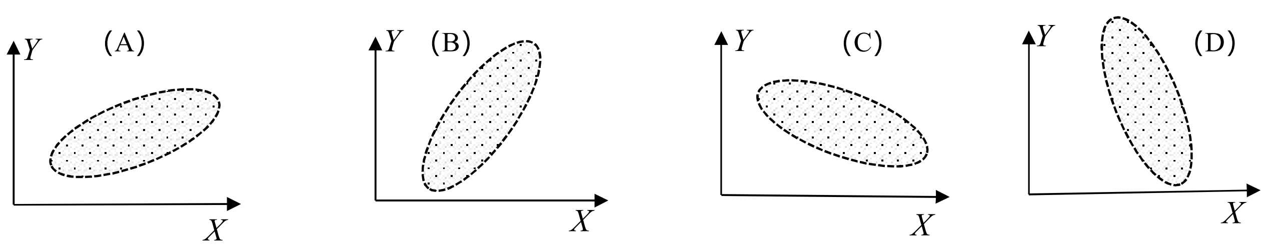


图 1 数据可视化

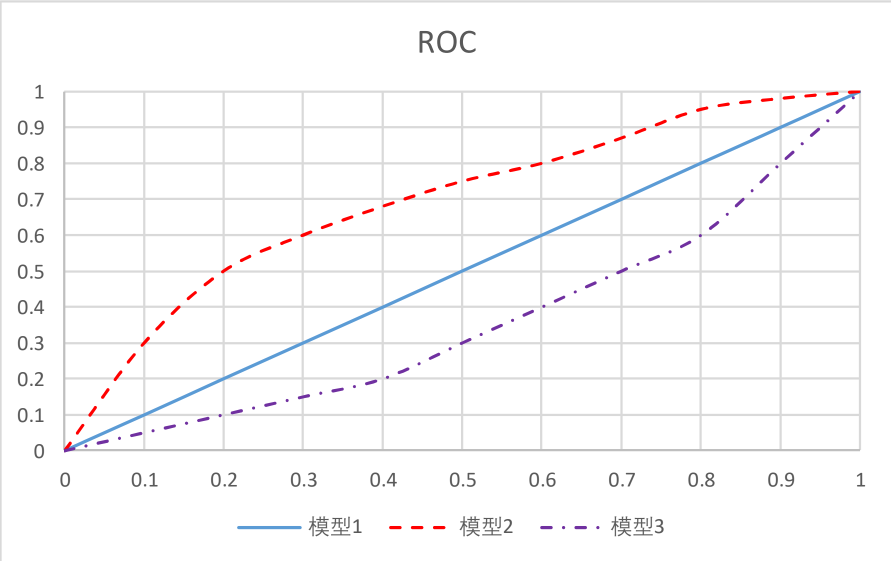
6. 下图展示了三个逻辑回归模型的 ROC 曲线。根据图中的结果，下面说法正确的是？B

(A) 模型 1 的分类性能是最差的

(B) 模型 2 的分类性能是最好的

(C) 模型 3 分类性能比模型 1 好

(D) 模型 3 分类性能比模型 2 好



，

二、简答（4 题，共 20 分）

1. 在因子分析中，共同因子的方差贡献怎么定义？有什么统计学意义？

参照 P136

2. 因子分析与对应分析异同？

两者都是描述数据间关系的一种实用的多元统计分析技术，用于揭示数据间的联系

两者都是利用降维的思想简化数据表达。

对应分析又称为 R-Q 因子分析，实质是对变量和样品同时分析，揭示两者的内

在关系；而因子分析一般指 R 因子分析，只对变量进行分析。

三、证明题（2 题，共 15 分）

1. 设有因子分析模型：

$$\boldsymbol{X}_i = w_{i1}\boldsymbol{F}_1 + w_{i2}\boldsymbol{F}_2 + \cdots + w_{im}\boldsymbol{F}_m + \boldsymbol{r}_i$$

$$(m \le p, i = 1,2,\dots,p)$$

方差$\boldsymbol{var}(\boldsymbol{r}_i) = s_i^2$

试证明：$\sum_{j=1}^{m} w_{ij}^2 + s_i^2 = 1$

参考 P136 页

2. 有原始数据

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x_{11}} & \boldsymbol{x_{12}} & \cdots & \boldsymbol{x_{1p}} \\ \boldsymbol{x_{21}} & \boldsymbol{x_{22}} & \cdots & \boldsymbol{x_{2p}} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{x_{n1}} & \boldsymbol{x_{n2}} & \cdots & \boldsymbol{x_{np}} \end{bmatrix}' = (\boldsymbol{X_1}\ \boldsymbol{X}_2\ \dots \boldsymbol{X}_p)'$$

试写出其第 k 主成分的数学表达式，并证明。

参考 P101、P102

3. [因子分析]

P141 关于因子分析的因子旋转，有类似下面的关于因子共同度性质的一个证明。这一段证明中，$\sum_{j=1}^{m}\sum_{l=1}^{m}\sum_{\substack{t=1\\t\neq l}}^{m} a_{il}a_{it}\gamma_{lj}\gamma_{tj}$这段是等于 0 的，试给出详细证明（注意说明原理）。

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

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

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

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

证明:

$$\sum_{j=1}^{m}\sum_{l=1}^{m}\sum_{\substack{t=1\\t\neq l}}^{m} a_{il}a_{it}\gamma_{lj}\gamma_{tj} = \sum_{\substack{t=1\\t\neq l}}^{m}\sum_{l=1}^{m} a_{il}a_{it}\sum_{j=1}^{m}\gamma_{lj}\gamma_{tj} = 0$$

注意$\sum_{j=1}^{m}\gamma_{lj}\gamma_{tj} = 0$

是因为正交阵的任意两个向量点积为 0

四、计算题（2 题，共 15 分）

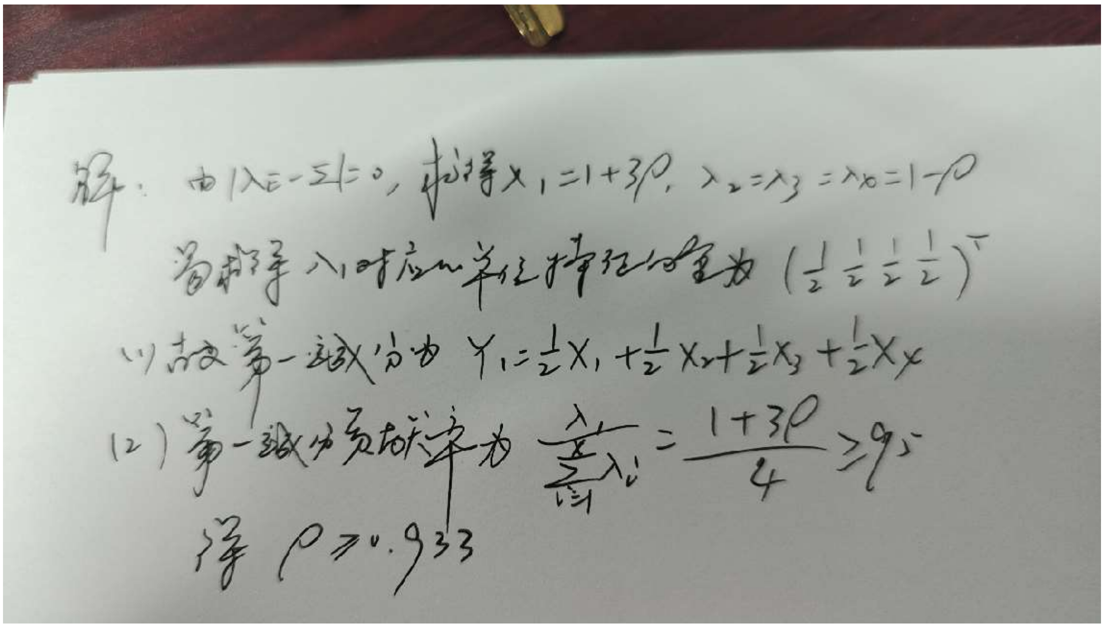
1.

设 $X=(X_1,X_2,X_3,X_4)^T \sim N_4(0,\Sigma)$，协方差阵 $\Sigma=\begin{pmatrix}1 & \rho & \rho & \rho\\ \rho & 1 & \rho & \rho\\ \rho & \rho & 1 & \rho\\ \rho & \rho & \rho & 1\end{pmatrix}$，

$$0<\rho\le 1$$

(1) 试从 Σ 出发求 X 的第一总体主成分；

(2) 试问当 ρ 取多大时才能使第一主成分的贡献率达 95%以上。



解：由 $|\lambda E-\Sigma|=0$，求得 $\lambda_1=1+3\rho$，$\lambda_2=\lambda_3=\lambda_4=1-\rho$

当特征根 λ_1 时对应的单位特征向量为 $(\frac{1}{2}\ \frac{1}{2}\ \frac{1}{2}\ \frac{1}{2})^T$

(1) 故第一主成分为 $Y_1=\frac{1}{2}X_1+\frac{1}{2}X_2+\frac{1}{2}X_3+\frac{1}{2}X_4$

(2) 第一主成分贡献率为 $\frac{\lambda_1}{\sum_{i=1}^{4}\lambda_i}=\frac{1+3\rho}{4}\ge 95\%$

得 $\rho\ge 0.933$

参考求解过程：
r2-r1，r3-r1，r4-1
c1-c2，c1-c3，c1-c4

2. 设三元总体 X 的协方差阵为 $\Sigma=\begin{pmatrix}1 & 0 & 0\\ 0 & 3 & 0\\ 0 & 0 & 6\end{pmatrix}$，从 Σ 出发，求总体主成分 F_1,F_2,F_3，并求前两个主成分的累积贡献率。

解：特征方程 $|\lambda E-\Sigma|=0$

求得特征根：$\lambda_1=6$，$\lambda_2=3$，$\lambda_3=1$

相应的特征向量为 $(0,0,1)^T$，$(0,1,0)^T$，$(1,0,0)$

所以 $F_1=X_3$，$F_2=X_2$，$F_3=X_1$

前两个主成分的累积贡献率 $=\frac{6+3}{6+3+1}=90\%$.

3. 设三维随机向量 $X\sim N_3(\mu,\Sigma)$，其中 $\Sigma=\begin{pmatrix}4&1&0\\1&3&0\\0&0&2\end{pmatrix}$，问 X_1 与 X_2 是否独立？

$(X_1,X_2)'$ 和 X_3 是否独立？为什么？

答：因为 $cov(X_1,X_2)=1$，所以 X_1 与 X_2 不独立。

协方差矩阵 $\Sigma=\begin{pmatrix}\Sigma_{11}&\Sigma_{12}\\\Sigma_{21}&\Sigma_{22}\end{pmatrix}$，$(X_1,X_2)'$

的协方差矩阵为 Σ_{11}，

$cov((X_1,X_2)',X_3)=\Sigma_{12}=0$

所以 $(X_1,X_2)'$ 与 X_3 是不相关的，因为 X 符合正态分布，而正态分布中，不相关与相互独立是等价的，所以 $(X_1,X_2)'$ 与 X_3 是独立的.

4. 使用Fisher判别进行三维数据进行二分类，类别标记分别为G_1，G_2。假设两类数据同分布，判别系数 $\alpha = (0.45\ \ 0.23\ \ 0.86)'$，两类数据的中心点分别为

$$\bar{x}^{(1)} = (2.1\ \ 4.5\ \ 9.7)',\quad \bar{x}^{(2)} = (10.4\ \ 5.5\ \ 1.1)'。$$

试写出其判别准则，并判断下面两个数据属于哪一类？

$$x_1 = (9.1\ 5.5\ 7.8)' \quad x_2 = (6.9\ 5.6\ 4.9)'$$

y(x) = a'x=0.45x1+0.23x2+0.86x3

y(1)=0.45*2.1+0.23*4.5+0.86*9.7=10.322

y(2)=0.45*10.4+0.23*5.5+0.86*1.1=6.891

Yc=(y(1)+y(2))/2=8.6065

对数据 $x_1 = (9.1\ 5.5\ 7.8)'$，有

y(x1) =0.45*9.1+0.23*5.5+0.86*7.8=12.068>8.6065,属于 G1

对数据 $x_2 = (6.9\ 5.6\ 4.9)'$

y(x2) =0.45*6.9+0.23*5.6+0.864.9=8.607≈8.6065

待判,（由于 8.607>8.6065，所以答属于 G1 也对）