《多元统计分析》试题

系别：　　姓名：　　学号：

1. 设 $X=(X_1,\ X_2,\ X_3)^{'}\sim N_3(\mu,\ \Sigma)$，其中

$$\mu^{'}=(2,-3,1),\quad \Sigma=\begin{pmatrix}1&1&1\\1&3&2\\1&2&2\end{pmatrix}$$

(1) 试求 $3X_1-2X_2+X_3$ 的分布；

(2) 构造向量 $a_{2\times1}$ 使 X_2 和 $X_2-a^{'}\begin{pmatrix}X_1\\X_3\end{pmatrix}$ 为独立的；

(3) $Y_1=X_1-X_2,\quad Y_2=X_1+X_2+X_3$, 求 Y_1 和 Y_2 的相关系数；

(4) 求条件分布 $(X_2\mid X_1=2.5,\ X_3=1.5)$.

解答：

(1) $3X_1-2X_2+X_3=(3,\ -2,\ 1)X\sim N(\mu,\sigma^2)$, 其中 $\mu=13,\ \sigma^2=9$.

(2) 设 $a^{'}=(a_1,\ a_2)$, 这样 $X_2=(0,1,0)X,\quad X_2-a^{'}\begin{pmatrix}X_1\\X_3\end{pmatrix}=(-a_1,1,-a_2)X$.. 要 X_2 和 $X_2-a^{'}\begin{pmatrix}X_1\\X_3\end{pmatrix}$ 相互独立独立，只需 $(0,1,0)\Sigma(-a_1,1,-a_2)^{'}=0$ 即可。即 a_1,a_2 满足 $a_1+2a_2=3$. 故可取 $a_1=1,a_2=1$, 即 $a^{'}=(1,1)$.

(3) $Z=\begin{pmatrix}Y_1\\Y_2\end{pmatrix}=\begin{pmatrix}X_1-X_2\\X_1+X_2+X_3\end{pmatrix}=\begin{pmatrix}1&-1&0\\1&1&1\end{pmatrix}\begin{pmatrix}X_1\\X_2\\X_3\end{pmatrix}=AX$, 所以 Z 为二元正态分布 $N_2(\mu_Z,\Sigma_Z)$，其中均值为

$$\mu_Z=A\mu=\begin{pmatrix}1&-1&0\\1&1&1\end{pmatrix}\begin{pmatrix}2\\-3\\1\end{pmatrix}=\begin{pmatrix}5\\0\end{pmatrix},$$

协差阵为

$$\Sigma_Z=A\Sigma A^{'}=\begin{pmatrix}1&-1&0\\1&1&1\end{pmatrix}\begin{pmatrix}1&1&1\\1&3&2\\1&2&2\end{pmatrix}\begin{pmatrix}1&1\\-1&1\\0&1\end{pmatrix}=\begin{pmatrix}2&-3\\-3&14\end{pmatrix}.$$

这样可算出 Y_1 与 Y_2 之间的相关系数为

$$\rho_{Y_1Y_2} = \frac{cov(Y_1,\ Y_2)}{\sqrt{var(Y_1)var(Y_2)}} = \frac{-3}{\sqrt{2\cdot 14}} = -0.567.$$

(4) $E(X_2|X_1=2.5, X_3=1.5) = -3 + (0,0)\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}^{-1}\begin{pmatrix} 2.5-2 \\ 1.5-1 \end{pmatrix} = -2.5,$

$$Var(X_2|X_1=2.5, X_3=1.5) = 3 - (1,2)\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}^{-1}\begin{pmatrix} 1 \\ 2 \end{pmatrix} = 1.$$

所以我们有 $(X_2|X_1=2.5, X_3=1.5) \sim N(-2.5, 1)$. #

2. 考虑三个总体的 Bayes 判别问题，设先验概率、误判损失以及密度函数在新的样本点 x_0 处的值如下表：

			真实总体	
		G_1	G_2	G_3
分	G_1	$c(1\|1)=0$	$c(1\|2)=500$	$c(1\|3)=100$
	G_2	$c(2\|1)=10$	$c(2\|2)=0$	$c(2\|3)=50$
类	G_3	$c(3\|1)=50$	$c(3\|2)=200$	$c(3\|3)=0$
先验概率		$q_1=0.05$	$q_2=0.6$	$q_3=0.35$
$p_i(x_0)$		$p_1(x_0)=0.01$	$p_2(x_0)=0.85$	$p_3(x_0)=2$

(1) 利用 ECM 最小的规则，应该把 x_0 判给哪一个总体？

(2) 若所有的误判损失相同，此时 x_0 应判给哪一个总体？并计算其后验概率，根据后验概率最大原则， x_0 应判给哪一个总体？

解答： *(1)* $h_j(x) = \sum_{i=1}^{3} q_i\ p_i(x)\ c(j|i), j=1,2,3$, 分别算得

$h_1(x_0) = q_2p_2(x_0)c(1|2) + q_3p_3(x_0)c(1|3) = 0.6*0.85*500 + 0.35*2*100 = 195,$

$h_2(x_0) = q_1p_1(x_0)c(2|1) + q_3p_3(x_0)c(2|3) = 0.05*0.01*10 + 0.35*2*50 = 35,$

$h_3(x_0) = q_1p_1(x_0)c(3|1) + q_2p_2(x_0)c(3|2) = 0.05*0.01*100 + 0.6*0.85*200 = 102.$

故根据 *ECM* 最小的规则，应把 x_0 判 G_2.

(2) 若所有的误判损失都相等，记作 c, 则我们可以计算得

$h_1(x_0) = 0.6*0.85*c + 0.35*2*c = 1.21c,$

$h_2(x_0) = 0.05*0.01*c + 0.35*2*c = 0.7005c,$

$h_3(x_0) = 0.05*0.01*c + 0.6*0.85*c = 0.5105c.$

故把 x_0 判 G_3 。

现在来计算后验概率： $P(G_i|x) = q_ip_i(x)/\sum_{l=1}^{k} q_lp_l(x)$

$P(G_1|x) = (0.05*0.01)/(0.05*0.01 + 0.6*0.85 + 0.35*2) = 0.0005/1.2105$

$P(G_2|x) = (0.6*0.85)/(0.05*0.01 + 0.6*0.85 + 0.35*2) = 0.51/1.2105$

$P(G_3|x) = (0.35 * 2)/(0.05 * 0.01 + 0.6 * 0.85 + 0.35 * 2) = 0.7/1.2105$

故把 x_0 判 G_3。 #

3. 对标准化变量 Z_1, Z_2, 和 Z_3, 相关阵为

$$R = \begin{pmatrix} 1.0 & 0.63 & 0.45 \\ 0.63 & 1.0 & 0.35 \\ 0.45 & 0.35 & 1.0 \end{pmatrix}$$

算得 R 的特征根和特征向量为

$$\lambda_1 = 1.96, \quad e_1' = (0.625,\ 0.593,\ 0.507);$$
$$\lambda_2 = 0.68, \quad e_2' = (-0.219,\ -0.491,\ 0.843);$$
$$\lambda_3 = 0.36, \quad e_1' = (0.749,\ -0.638,\ -0.177).$$

(1). 考虑 $m = 1$ 正交因子模型，用主成分法估计因子载荷阵 A 和特殊方差阵 Ψ. 并把估计值代入写出写出正交因子模型。

(2). 若考虑 $m = 2$ 正交因子模型，做 (1) 要求的问题。并求两个公因子所能解释的总方差的比例是多少？

解答： *(1)* $\widehat{A} = \sqrt{\lambda_1}e_1 = \sqrt{1.96}\begin{pmatrix} 0.625 \\ 0.593 \\ 0.507 \end{pmatrix} = \begin{pmatrix} 0.875 \\ 0.830 \\ 0.710 \end{pmatrix}$,

$$\widehat{A}\widehat{A}' = \begin{pmatrix} 0.766 & 0.726 & 0.621 \\ 0.726 & 0.689 & 0.589 \\ 0.621 & 0.589 & 0.504 \end{pmatrix}, \quad \widehat{\Psi} = diag(R - \widehat{A}\widehat{A}') = \begin{pmatrix} 0.234 & 0 & 0 \\ 0 & 0.311 & 0 \\ 0 & 0 & 0.496 \end{pmatrix}.$$

把估计值代入因子模型为

$$\begin{cases} Z_1 = 0.875F_1 + \epsilon_1 \\ Z_2 = 0.830F_1 + \epsilon_2 \\ Z_3 = 0.710F_1 + \epsilon_3 \end{cases}$$

且满足

$$\begin{cases} E(F) = 0, \quad cov(F_1) = 1 \\ E(\varepsilon) = (0,0,0)', \quad cov(\varepsilon) = \Psi = diag(0.234, 0.311, 0.496) \\ cov(\varepsilon, F_1) = 0 \end{cases},$$

(2) $\widehat{A} = (\sqrt{\lambda_1}e_1,\ \sqrt{\lambda_2}e_2) = \begin{pmatrix} 0.875 & -0.181 \\ 0.830 & -0.405 \\ 0.710 & 0.695 \end{pmatrix}$,

$$\widehat{A}\widehat{A}' = \begin{pmatrix} 0.798 & 0.804 & 0.489 \\ 0.804 & 0.853 & 0.308 \\ 0.489 & 0.308 & 0.987 \end{pmatrix}, \quad \widehat{\Psi} = diag(R - \widehat{A}\widehat{A}') = \begin{pmatrix} 0.202 & 0 & 0 \\ 0 & 0.147 & 0 \\ 0 & 0 & 0.013 \end{pmatrix}.$$

把估计值代入因子模型为

$$\begin{cases} Z_1 = 0.875F_1 - 0.181F_2\epsilon_1 \\ Z_2 = 0.830F_1 - 0.405F_2\epsilon_2 \\ Z_3 = 0.710F_1 + 0.695F_2\epsilon_3 \end{cases}$$

且满足

$$\begin{cases} E(F) = 0, \quad cov(F_1) = 1 \\ E(\varepsilon) = (0,0,0)', \quad cov(\varepsilon) = \Psi = diag(0.108, 0.147, 0.013) \\ cov(\varepsilon, F_1) = 0 \end{cases},$$

4. 五个样本两两之间的距离矩阵是

$$D_0 = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & & & & \\ 4 & 0 & & & \\ 6 & 9 & 0 & & \\ 1 & 7 & 10 & 0 & \\ 6 & 3 & 5 & 8 & 0 \end{pmatrix} \end{array}$$

试用最短距离法将这个五个样本进行分类。

解答：

	G_1	G_2	G_3	G_4	G_5	$G_6=\{G_1,G_4\}$	$G_7=\{G_2,G_5\}$	$G_8=\{G_6,G_7\}$	$G_9\{G_3,G_8\}$
G_1	*0*								
G_2	*4*	*0*							
G_3	*6*	*9*	*0*						
G_4	**1**	*7*	*10*	*0*					
G_5	*6*	**3**	*5*	*8*	*0*				
G_6	×	*4*	*6*	×	*6*	*0*			
G_7	×	×	*5*	×	×	**4**	*0*		
G_8	×	×	**5**	×	×	×	×	*0*	
G_8	×	×	×	×	×	×	×	×	*0*

将上述并类过程画成系统聚类图如下：

这样若取阈值 $T=3.5$，则分成三类 $\{1,4\},\{2,5\},\{3\}$. 但若取阈值 $T=4.2$，则分成两类 $\{1,4,2,5\},\{3\}$.

5. $X=(X_1,\ X_2,\ X_3)'$ 的协差阵为

$$\Sigma=\begin{pmatrix}\sigma^2 & \sigma^2\rho & 0\\ \sigma^2\rho & \sigma^2 & \sigma^2\rho\\ 0 & \sigma^2\rho & \sigma^2\end{pmatrix},\quad -\frac{1}{\sqrt{2}}<\rho<\frac{1}{\sqrt{2}}.$$

试求三个主成分及每个主成分的方差贡献率.

6. 求 μ_1 和 $\mu_1-\mu_2$ 的同时区间 (*Bonferroni* 方法) 和联合置信区间 ($\alpha=0.05$), 其中样本观察矩阵为

$$X=\begin{pmatrix}X_1'\\ X_2'\\ X_3'\\ X_4'\end{pmatrix}=\begin{pmatrix}3 & 6\\ 4 & 4\\ 5 & 7\\ 4 & 7\end{pmatrix}.$$

($F_{2,\ 2}(0.05)=19,\ \ t_3(0.0125)=4.1765$)

解答：$\overline{X}=\frac{1}{4}\sum_{k=1}^{4}X_i=\begin{pmatrix}4\\6\end{pmatrix},\quad A=\sum_{k=1}^{4}(X_i-\overline{X})(X_i-\overline{X})'=\begin{pmatrix}2 & 1\\ 1 & 6\end{pmatrix}.$

$$\mu_1=(1,0)\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}:=a_1'\mu,\quad \mu_1-\mu_2=(1,-1)\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}:=a_2'\mu.$$

联合置信区间为 ($n=4,p=2$)：

$$\begin{aligned}\mu_1 &\in a_1'\overline{X}\pm\sqrt{\frac{p}{n(n-p)}F_{p,\ n-p}(\alpha)a_1'Aa_1}\\ &= (1,0)\begin{pmatrix}4\\6\end{pmatrix}\pm\sqrt{\frac{2}{2\cdot4}\ F_{2,\ 2}(0.05)\cdot2}=4\pm3.08=(0.92,7.08)\\ \mu_1-\mu_2 &\in a_2'\overline{X}\pm\sqrt{\frac{p}{n(n-p)}F_{p,\ n-p}(\alpha)a_2'Aa_2}\\ &= (1,-1)\begin{pmatrix}4\\6\end{pmatrix}\pm\sqrt{\frac{2}{2\cdot4}\ F_{2,\ 2}(0.05)\cdot6}=-2\pm5.34=(-7.34,3.34)\end{aligned}$$

同时置信区间为：

$$\begin{aligned}
\mu_1 &\in a_1'\overline{X} \pm t_{n-1}(\alpha/2k)\sqrt{a_1'Aa_1}/\sqrt{n(n-1)} \\
&= 4 \pm t_3(0.0125)\sqrt{a_{11}}/\sqrt{4\cdot 3} = 4 \pm 1.705 = (2.295, 5.705), \\
\mu_1-\mu_2 &\in a_2'\overline{X} \pm t_{n-1}(\alpha/2k)\sqrt{a_2'Aa_2}/\sqrt{n(n-1)} \\
&= -2 \pm t_3(0.0125)\sqrt{6}/\sqrt{4\cdot 3} = -2 \pm 2.953 = (-4.953, 0.953),
\end{aligned}$$

7. 简要回答：

(1)Wishart 分布和 Wilks 分布的定义；

(2)T^2 分布。解答：*(1) Wishart* 分布：设 $X_i = (X_{i1}, \cdots, X_{ip})' \sim N_p(\mu_i, \Sigma), \quad i = 1, \cdots, n.$ 且相互独立，则由 $X_i, \quad i = 1, \cdots, n$ 组成的随机矩阵

$$W_{p\times p} = \sum_{i=1}^{n} X_i X_i'.$$

W 的分布为非中心的 *Wishart* 分布，记为 $W_p(n, \Sigma, \Delta)$, 其中 $\Delta = \sum_{i=1}^{n} \mu_i \mu_i'$.

Wilks 分布：若 $A \sim W_p(n,\ \Sigma)$ 与 $B \sim W_p(m,\ \Sigma)$ 独立，$n > p,\ m > p,\ \Sigma > 0$，则称

$$\Lambda = \frac{|A|}{|A+B|}$$

服从 Wilks 分布，记作 $\Lambda \sim \Lambda_{p,\ n,\ m}$.

(2) Hotelling T^2 分布：设 $A \sim W_p(n, I_p)$ 与 $\mu \sim N_p(\mu, I_p)$ 独立，$n > p$, 称

$$T^2 = n\mu' A^{-1}\mu$$

为服从非中心的 Hotelling T^2 分布，记作 $T^2 \sim T^2_{p,\ n,\ \mu}$. 若 $\mu = 0$, 称 T^2 服从（中心）的 T^2 分布，记作 $T^2 \sim T_{p,\ n}$.