Homework 3

6

> ✎ Note
>
> 假设随机向量$(X,Y)'$服从二元正态分布,即
>
> $$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$$
>
> .
> 记$(x_1,y_1)',\cdots,(x_m,y_m)'$为来自该二元正态分布的容量为$m$的一组的独立同分布随机样本.此外,额外观测了来自总体$X\sim N(\mu_1,\sigma_1^2)$容量为$n-m$的一组独立同分布随机样本$x_{m+1},\cdots,x_n$,其中$n>m$.试给出所有未知参数的极大似然估计.

$$Sol.\ 已知, g(x)=\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left\{-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right\},$$

$$f(x,y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_1)^2}{\sigma_1^2}+\frac{(y-\mu_2)^2}{\sigma_2^2}-\frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}\right)\right\}.$$

$$则似然函数为 \ln L_1=\ln(\sum_{i=m+1}^{n}g(x_i))=-\frac{n-m}{2}\ln(2\pi)-\frac{n-m}{2}\ln(\sigma_1^2)-\frac{1}{2\sigma_1^2}\sum_{i=m+1}^{n}(x_i-\mu_1)^2,$$

$$\ln L_2=-m\ln(2\pi)-\frac{m}{2}\ln(\sigma_1^2\sigma_2^2(1-\rho^2))$$

$$-\frac{1}{2(1-\rho^2)}\sum_{i=1}^{m}\left(\frac{(x_i-\mu_1)^2}{\sigma_1^2}+\frac{(y_i-\mu_2)^2}{\sigma_2^2}-\frac{2\rho(x_i-\mu_1)(y_i-\mu_2)}{\sigma_1\sigma_2}\right).$$

$$有 \ln L=\ln L_1+\ln L_2,$$

$$联立\begin{cases} \frac{\partial\ln L}{\partial\mu_1}=\frac{1}{1-\rho^2}\sum_{i=1}^{m}\left(\frac{x_i-\mu_1}{\sigma_1^2}-\frac{\rho(y_i-\mu_2)}{\sigma_1\sigma_2}\right)+\sum_{i=m+1}^{n}\frac{(x_i-\mu_1)}{\sigma_1^2}=0 & (1) \\ \frac{\partial\ln L}{\partial\mu_2}=\frac{1}{1-\rho^2}\sum_{i=1}^{m}\left(\frac{y_i-\mu_2}{\sigma_2^2}-\frac{\rho(x_i-\mu_1)}{\sigma_1\sigma_2}\right)=0 & (2) \\ \frac{\partial\ln L}{\partial\sigma_1^2}=-\frac{n}{2\sigma_1^2}+\frac{1}{2(1-\rho^2)}\sum_{i=1}^{m}\left(\frac{(x_i-\mu_1)^2}{\sigma_1^4}-\frac{\rho(x_i-\mu_1)(y_i-\mu_2)}{\sigma_1^3\sigma_2}\right)+\frac{1}{2\sigma_1^4}\sum_{i=m+1}^{n}(x_i-\mu_1)^2=0 & (3) \\ \frac{\partial\ln L}{\partial\sigma_2^2}=-\frac{m}{2\sigma_2^2}+\frac{1}{2(1-\rho^2)}\sum_{i=1}^{m}\left(\frac{(y_i-\mu_2)^2}{\sigma_2^4}-\frac{\rho(x_i-\mu_1)(y_i-\mu_2)}{\sigma_1\sigma_2^3}\right)=0 & (4) \\ \frac{\partial\ln L}{\partial\rho}=\frac{m\rho}{1-\rho^2}-\frac{\rho}{(1-\rho^2)^2}\left(\sum_{i=1}^{m}\left(\frac{(x_i-\mu_1)^2}{\sigma_1^2}+\frac{(y_i-\mu_2)^2}{\sigma_2^2}\right)\right)-\frac{1+\rho^2}{(1-\rho^2)^2}\sum_{i=1}^{m}\left(\frac{(x_i-\mu_1)(y_i-\mu_2)}{\sigma_1\sigma_2}\right)=0 & (5) \end{cases}$$

$$(2)\Rightarrow\sum_{i=1}^{m}\frac{\rho(x_i-\mu_1)}{\sigma_1\sigma_2}=\sum_{i=1}^{m}\frac{y_i-\mu_2}{\sigma_2^2}代入(1)可得\hat{\mu}_1=\frac{1}{n}\sum_{i=1}^{n}x_1,$$

7

> ✎ Note
>
> 假设从该二元正态总体$N_2(\mu,\mathbf{\Sigma})$中随机产生n个模拟样本，其中
>
> $$\mu=\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \Sigma=\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

,
且$\sigma_1=1,\sigma_2=2,\rho=0.6$.针对不同的样本量$n=50,100,200$,重复模拟1000次.
(1)试计算参数$\mu_1,\mu_2,\sigma_1,\sigma_2$和$\rho$估计的平均值、偏差和标准差,并通过QQ图和直方图展示估计的好坏.进一步,随着样本量的变化,说明结果有什么变化;
(2)基于式子(5.41)和(5.42)编写程序,分别计算相关系数ρ的95%的平均置信区间和区间长度,并进行比较哪个置信区间最优.进一步,随着样本量的变化,平均区间长度有什么变化.

$$Sol.(1)\text{利用}MLE,\text{可知}\hat{\mu}=\overline{x},\hat{\mathbf{\Sigma}}=\mathbf{V}/n,\hat{\rho}=\frac{\sum\limits_{i=1}^{n}(x_{i1}-\overline{x}_1)(x_{i2}-\overline{x}_2)}{\sqrt{\sum\limits_{i=1}^{n}(x_{i1}-\overline{x}_1)^2\sum\limits_{i=1}^{n}(x_{i2}-\overline{x}_2)^2}},$$

$$\text{其中}\overline{x}=\begin{pmatrix}\overline{x}_1\\ \overline{x}_2\end{pmatrix}=\begin{pmatrix}\frac{1}{n}\sum\limits_{i=1}^{n}x_{i1}\\ \frac{1}{n}\sum\limits_{i=1}^{n}x_{i2}\end{pmatrix},\mathbf{V}=\begin{pmatrix}\sum\limits_{i=1}^{n}(x_{i1}-\overline{x}_1)^2 & \sum\limits_{i=1}^{n}(x_{i1}-\overline{x}_1)(x_{i2}-\overline{x}_2)\\ \sum\limits_{i=1}^{n}(x_{i2}-\overline{x}_2)(x_{i1}-\overline{x}_1) & \sum\limits_{i=1}^{n}(x_{i2}-\overline{x}_2)^2\end{pmatrix}.$$

$$(2)(5.41)\text{与}(5.42)\text{置信区间分别为}\left[r(n)-\frac{1-r^2(n)}{\sqrt{n}}z_{1-\alpha/2},r(n)+\frac{1-r^2(n)}{\sqrt{n}}z_{1-\alpha/2}\right],$$
$$\left[\frac{1}{2}\ln\frac{1+r(n)}{1-r(n)}-\frac{1}{\sqrt{n}}z_{1-\alpha/2},\frac{1}{2}\ln\frac{1+r(n)}{1-r(n)}+\frac{1}{\sqrt{n}}z_{1-\alpha/2}\right].$$
其中$z_{1-\alpha/2}$为标准正态分布的上$\alpha/2$分位点.

9

Note

假设$x_1,\cdots,x_n$为来自$0-1$分布的独立同分布的简单随机样本，其分布律为
$\Pr(x_1=1)=p,\Pr(x_1=0)=1-p$,其中$0<p<1$.根据中心极限定理,有$\sqrt{n}(\overline{x}-p)\xrightarrow{d}N(0,p(1-p))$,
其中$\overline{x}=\frac{1}{n}\sum\limits_{i=1}^{n}x_i$.
(1)试用Fisher Z变换方法构造p的置信水平为$1-\alpha$的置信区间;
(2)取$p=0.6$,从$0-1$分布中随机产生样本量$n=50,100,200$的随机数,重复1000次试验,编写程序,计算p的95%的平均置信区间和区间长度,并观察随着样本量的变化,平均区间长度有什么变化.

$$Sol.(1)\text{由}\mathbf{Fisher}Z\text{变换方法},\sqrt{n}(\overline{x}-p)\xrightarrow{d}N(0,p(1-p))\text{只需让}(f'(p))^2p(1-p)=1\text{即可}.$$
$$\mathrm{d}f=\frac{1}{\sqrt{p(1-p)}}\mathrm{d}p,f(p)=\int\frac{1}{\sqrt{p(1-p)}}\mathrm{d}p\overset{\text{令}p=\sin^2y}{=\!=\!=\!=}\int\frac{2\cos y\sin y}{\sin y\cos y}dy=2y=\arcsin(\sqrt{p})$$
$$\text{故}\sqrt{n}\left[\arcsin(\sqrt{\overline{x}})-\arcsin(\sqrt{p})\right]\xrightarrow{d}N(0,1),\text{从而}p\text{的置信水平为}1-\alpha\text{的置信区间为}:$$
$$\left(\left[\sin\left(\arcsin(\overline{x})-\frac{z_{1-\alpha}}{\sqrt{n}}\right)\right]^2,\left[\sin\left(\arcsin(\overline{x})+\frac{z_{1-\alpha}}{\sqrt{n}}\right)\right]^2\right).$$

11

Note

设$\mathbf{X}$和$\mathbf{Y}$是相互独立的随机向量,且$\mathbf{X}\sim N_p(\mu_1,\mathbf{\Sigma}),\mathbf{Y}\sim N_p(\mu_2,\mathbf{\Sigma})$,其中$\mathbf{\Sigma}>0$.进一步假设$x_1,\cdots,x_n$为来自总体$\mathbf{X}$的独立同分布的随机样本,$y_1,\cdots,y_m$为来自总体$\mathbf{Y}$的独立同分布的随机样本,$n,m>p$.

(1)试证明参数$(\mu_1,\mu_2,\mathbf{\Sigma})$的充分完备统计量为$(\overline{x},\overline{y},\mathbf{V}_1+\mathbf{V}_2)$,其中

$$\mathbf{V}_1=\sum_{i=1}^{n}(x_i-\overline{x})(x_i-\overline{x})',\mathbf{V}_2=\sum_{i=1}^{m}(y_i-\overline{y})(y_i-\overline{y})'$$

;
(2)试求参数$(\mu_1,\mu_2,\mathbf{\Sigma})$的极大似然估计,他们是无偏估计吗?
(3)试求参数$(\mu_1,\mu_2,\mathbf{\Sigma})$的已知最小协方差矩阵无偏估计，它们是不是唯一存在的?
(4)$\Delta^2=(\mu_1-\mu_2)'\mathbf{\Sigma}^{-1}(\mu_1-\mu_2)$通常用来表示两个正态分布$N_p(\mu_1,\mathbf{\Sigma})$和$N_p(\mu_2,\mathbf{\Sigma})$之间的距离,试求$\Delta^2$的极大似然估计,并且是无偏估计吗?若不是,请给出$\Delta^2$的无偏估计.

$$(1) Proof.\ f(x_1,\cdots,x_n,y_1,\cdots,y_m)=\left(\frac{1}{(2\pi)^{\frac{p}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}}\right)^n\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}[(x_i-\mu_1)'\mathbf{\Sigma}^{-1}(x_i-\mu_1)]\right\}$$

$$\cdot\left(\frac{1}{(2\pi)^{\frac{p}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}}\right)^m\exp\left\{-\frac{1}{2}\sum_{i=1}^{m}[(y_i-\mu_2)'\mathbf{\Sigma}^{-1}(y_i-\mu_2)]\right\}$$

$$=\frac{1}{(2\pi)^{\frac{p(m+n)}{2}}|\mathbf{\Sigma}|^{\frac{m+n}{2}}}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}[(x_i-\mu_1)'\mathbf{\Sigma}^{-1}(x_i-\mu_1)]-\frac{1}{2}\sum_{i=1}^{m}[(y_i-\mu_2)'\mathbf{\Sigma}^{-1}(y_i-\mu_2)]\right\}$$

$$=\frac{1}{(2\pi)^{\frac{p(m+n)}{2}}|\mathbf{\Sigma}|^{\frac{m+n}{2}}}\mathrm{etr}\left\{-\frac{1}{2}\sum_{i=1}^{n}[(x_i-\mu_1)'\mathbf{\Sigma}^{-1}(x_i-\mu_1)]-\frac{1}{2}\sum_{i=1}^{m}[(y_i-\mu_2)'\mathbf{\Sigma}^{-1}(y_i-\mu_2)]\right\}$$

$$=\frac{1}{(2\pi)^{\frac{p(m+n)}{2}}|\mathbf{\Sigma}|^{\frac{m+n}{2}}}\mathrm{etr}\left\{-\frac{\mathbf{\Sigma}^{-1}}{2}\sum_{i=1}^{n}[(x_i-\mu_1)(x_i-\mu_1)']-\frac{\mathbf{\Sigma}^{-1}}{2}\sum_{i=1}^{m}[(y_i-\mu_2)(y_i-\mu_2)']\right\}$$

$$=\frac{1}{(2\pi)^{\frac{p(m+n)}{2}}|\mathbf{\Sigma}|^{\frac{m+n}{2}}}\mathrm{etr}\left\{-\frac{\mathbf{\Sigma}^{-1}}{2}\sum_{i=1}^{n}[(x_i-\overline{x})(x_i-\overline{x})'+(\overline{x}-\mu_1)(\overline{x}-\mu_1)']\right\}$$

$$\cdot\,\mathrm{etr}\left\{-\frac{\mathbf{\Sigma}^{-1}}{2}\sum_{i=1}^{n}[(y_i-\overline{y})(y_i-\overline{y})'+(\overline{y}-\mu_2)(\overline{y}-\mu_2)']\right\}$$

$$=\frac{1}{(2\pi)^{\frac{p(m+n)}{2}}|\mathbf{\Sigma}|^{\frac{m+n}{2}}}\mathrm{etr}\left\{-\frac{\mathbf{\Sigma}^{-1}}{2}(\mathbf{V}_1+\mathbf{V}_2+n(\overline{x}-\mu_1)(\overline{x}-\mu_1)'+m(\overline{y}-\mu_2)(\overline{y}-\mu_2)')\right\}$$

由$\mathrm{Neyman-Fisher}$因子判别法则可知$(\mu_1,\mu_2,\mathbf{\Sigma})$的充分完备统计量为$(\overline{x},\overline{y},\mathbf{V}_1+\mathbf{V}_2)$.

$$(2) Sol.\ L=\frac{1}{(2\pi)^{\frac{p(m+n)}{2}}|\mathbf{\Sigma}|^{\frac{m+n}{2}}}\mathrm{etr}\left\{-\frac{\mathbf{\Sigma}^{-1}}{2}(\mathbf{V}_1+\mathbf{V}_2)\right\}$$

$$\cdot\exp\left\{-\frac{n(\overline{x}-\mu_1)'\mathbf{\Sigma}^{-1}(\overline{x}-\mu_1)+m(\overline{y}-\mu_2)'\mathbf{\Sigma}^{-1}(\overline{y}-\mu_2))}{2}\right\}$$

$$\ln L=-\frac{p(m+n)}{2}\ln(2\pi)-\frac{m+n}{2}\ln|\mathbf{\Sigma}|-\frac{1}{2}\mathrm{tr}(\mathbf{\Sigma}^{-1}(\mathbf{V}_1+\mathbf{V}_2))$$

$$-\frac{n(\overline{x}-\mu_1)'\mathbf{\Sigma}^{-1}(\overline{x}-\mu_1)+m(\overline{y}-\mu_2)'\mathbf{\Sigma}^{-1}(\overline{y}-\mu_2))}{2}$$

注意$\mathbf{\Sigma}>0$, 故当$\hat{\mu}_1=\overline{x},\hat{\mu}_2=\overline{y}$, 似然函数取得最大值，代入原式并去掉无关项后,

$$\ln L'=-(m+n)\ln|\mathbf{\Sigma}|-\mathrm{tr}(\mathbf{\Sigma}^{-1}(\mathbf{V}_1+\mathbf{V}_2))=(m+n)\ln|\mathbf{\Sigma}^{-1}|-\mathrm{tr}(\mathbf{\Sigma}^{-1}(\mathbf{V}_1+\mathbf{V}_2))$$

令$\dfrac{\partial\ln L'}{\partial(\mathbf{\Sigma}^{-1})}=(m+n)\mathbf{\Sigma}-(\mathbf{V}_1+\mathbf{V}_2)=0$, 可得$\hat{\mathbf{\Sigma}}=\dfrac{\mathbf{V}_1+\mathbf{V}_2}{m+n}$,

$E(\hat{\mu}_1)=E(\overline{x})=\mu_1, E(\hat{\mu}_2)=E(\overline{y})=\mu_2, E(\hat{\mathbf{\Sigma}})=\dfrac{m+n-2}{m+n}\mathbf{\Sigma}$, 故$\hat{\mu}_1,\hat{\mu}_2$为$UE$, $\hat{\mathbf{\Sigma}}$不是UE.

(3)由(1)(2)$\hat{\mu}_1,\hat{\mu}_2,\dfrac{m+n}{m+n-2}\hat{\mathbf{\Sigma}}$为$UE$, 且为充分完备统计量的函数，故为$UMVUE$, 是唯一的.

(4)$\hat{\Delta}^2=(\hat{\mu}_1-\hat{\mu}_2)'\hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_1-\hat{\mu}_2)=(\overline{x}-\overline{y})'\left(\dfrac{\mathbf{V}_1+\mathbf{V}_2}{m+n}\right)^{-1}(\overline{x}-\overline{y})$,

又$\overline{x}\sim N_p(\mu_1,\mathbf{\Sigma}/n),\overline{y}\sim N_p(\mu_2,\mathbf{\Sigma}/n),\hat{\mathbf{\Sigma}}\sim W_p(m+n-2,\mathbf{\Sigma})$,

又$\overline{x}$与$\overline{y}$独立, 有$\overline{x}-\overline{y}\sim N_p(\mu_1-\mu_2,2\mathbf{\Sigma}/n)$, 又$\overline{x}-\overline{y}$与$\hat{\mathbf{\Sigma}}$独立,

$$E(\hat{\Delta}^2)=E[(\hat{\mu}_1-\hat{\mu}_2)'\hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_1-\hat{\mu}_2)]\cdot(m+n)=(m+n)(\mu_1,\mu_2)'\cdot\frac{\Sigma^{-1}}{m+n-2-p-1}(\mu_1,\mu_2)$$

故$\mathbf{\Delta}^{2\text{的}}$一个$UE$为$(m+n-3-p)(\overline{x}-\overline{y})'(\mathbf{V}_1+\mathbf{V}_2)^{-1}(\overline{x}-\overline{y})$.