应用多元统计分析

第二章部分习题解答

2-1 设3维随机向量$X\sim N_3(\mu,2I_3)$，已知

$$\mu=\begin{pmatrix}2\\0\\0\end{pmatrix}, A=\begin{pmatrix}0.5 & -1 & 0.5\\ -0.5 & 0 & -0.5\end{pmatrix}, d=\begin{pmatrix}1\\2\end{pmatrix}.$$

试求$Y=AX+d$的分布.

解:利用性质2,即得二维随机向量$Y\sim N_2(\mu_y,\Sigma_y)$,
其中:

$$\mu_y=A\mu+d=\begin{pmatrix}1\\-1\end{pmatrix}+\begin{pmatrix}1\\2\end{pmatrix}=\begin{pmatrix}2\\1\end{pmatrix},$$

$$\Sigma_y=A(2I_3)A'=2AA'=\begin{pmatrix}3 & -1\\ -1 & 1\end{pmatrix}.$$

2-2 设$X=(X_1,X_2)'\sim N_2(\mu,\Sigma)$，其中

$$\mu=\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix},\Sigma=\sigma^2\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}.$$

（1）试证明X_1+X_2和X_1-X_2相互独立.

（2）试求X_1+X_2和X_1-X_2的分布.

解: (1) 记$Y_1=X_1+X_2=(1,1)'X$,

$Y_2=X_1-X_2=(1,-1)'X$,

利用性质2可知Y_1, Y_2为正态随机变量。又

$$\mathrm{Cov}(Y_1,Y_2)=(1\ 1)\Sigma\begin{pmatrix}1\\ -1\end{pmatrix}=\sigma^2(1+\rho\ \ 1+\rho)\begin{pmatrix}1\\ -1\end{pmatrix}=0$$

故X_1+X_2和X_1-X_2相互独立.

或者记

$$Y=\begin{pmatrix}Y_1\\Y_2\end{pmatrix}=\begin{pmatrix}X_1+X_2\\X_1-X_2\end{pmatrix}=\begin{pmatrix}1&1\\1&-1\end{pmatrix}\begin{pmatrix}X_1\\X_2\end{pmatrix}=CX$$

则 $Y\sim N_2(C\mu,C\Sigma C')$

因 $\Sigma_Y=C\Sigma C'=\begin{pmatrix}1&1\\1&-1\end{pmatrix}\sigma^2\begin{pmatrix}1&\rho\\\rho&1\end{pmatrix}\begin{pmatrix}1&1\\1&-1\end{pmatrix}$

$$=\sigma^2\begin{pmatrix}1+\rho&1+\rho\\1-\rho&\rho-1\end{pmatrix}\begin{pmatrix}1&1\\1&-1\end{pmatrix}=\sigma^2\begin{pmatrix}2(1+\rho)&0\\0&2(1-\rho)\end{pmatrix}$$

由定理**2.3.1**可知$\boldsymbol{X_1+X_2}$ 和$\boldsymbol{X_1-X_2}$相互独立.

(2) 因

$$Y=\begin{pmatrix}X_1+X_2\\X_1-X_2\end{pmatrix}\sim N_2\left(\begin{pmatrix}\mu_1+\mu_2\\\mu_1-\mu_2\end{pmatrix},\sigma^2\begin{pmatrix}2(1+\rho)&0\\0&2(1-\rho)\end{pmatrix}\right)$$

$$\therefore X_1+X_2\sim N(\mu_1+\mu_2,2\sigma^2(1+\rho));$$

$$X_1-X_2\sim N(\mu_1-\mu_2,2\sigma^2(1-\rho)).$$

2-3 设$X^{(1)}$和$X^{(2)}$ 均为p维随机向量,已知

$$X=\begin{pmatrix}X^{(1)}\\X^{(2)}\end{pmatrix}\sim N_{2p}\left(\begin{pmatrix}\mu^{(1)}\\\mu^{(2)}\end{pmatrix},\begin{bmatrix}\Sigma_1&\Sigma_2\\\Sigma_2&\Sigma_1\end{bmatrix}\right),$$

其中$\mu^{(i)}$ (i＝1，2)为p维向量,Σ_i (i＝1，2)为p阶矩阵,

(1) 试证明$X^{(1)}+X^{(2)}$和$X^{(1)}-X^{(2)}$ 相互独立.

(2) 试求$X^{(1)}+X^{(2)}$ 和$X^{(1)}-X^{(2)}$ 的分布.

解 :(1) 令

$$Y=\begin{pmatrix}X^{(1)}+X^{(2)}\\X^{(1)}-X^{(2)}\end{pmatrix}=\begin{pmatrix}I_p&I_p\\I_p&-I_p\end{pmatrix}\begin{pmatrix}X^{(1)}\\X^{(2)}\end{pmatrix}=CX$$

则 $Y \sim N_{2p}(C\mu, C\Sigma C')$

因 $\mathrm{D}(Y) = C\mathrm{D}(X)C' = \begin{pmatrix} I_p & I_p \\ I_p & -I_p \end{pmatrix} \begin{pmatrix} \Sigma_1 & \Sigma_2 \\ \Sigma_2 & \Sigma_1 \end{pmatrix} \begin{pmatrix} I_p & I_p \\ I_p & -I_p \end{pmatrix}$

$$= \begin{pmatrix} \Sigma_1 + \Sigma_2 & \Sigma_1 + \Sigma_2 \\ \Sigma_1 - \Sigma_2 & \Sigma_2 - \Sigma_1 \end{pmatrix} \begin{pmatrix} I_p & I_p \\ I_p & -I_p \end{pmatrix}$$

$$= \begin{pmatrix} 2(\Sigma_1 + \Sigma_2) & O \\ O & 2(\Sigma_1 - \Sigma_2) \end{pmatrix}$$

由定理2.3.1可知$X^{(1)}+X^{(2)}$和$X^{(1)}-X^{(2)}$相互独立.

(2) 因

$$Y=\begin{pmatrix} X^{(1)}+X^{(2)} \\ X^{(1)}-X^{(2)} \end{pmatrix} \sim N_{2p}\left(\begin{pmatrix} \mu^{(1)}+\mu^{(2)} \\ \mu^{(1)}-\mu^{(2)} \end{pmatrix},\begin{pmatrix} 2(\Sigma_1+\Sigma_2) & O \\ O & 2(\Sigma_1-\Sigma_2) \end{pmatrix}\right)$$

所以 $X^{(1)}+X^{(2)} \sim N_p(\mu^{(1)}+\mu^{(2)}, 2(\Sigma_1+\Sigma_2));$

$X^{(1)}-X^{(2)} \sim N_p(\mu^{(1)}-\mu^{(2)}, 2(\Sigma_1-\Sigma_2)).$

注意:由D(X)≥0,可知 $(\Sigma_1 - \Sigma_2) \geq 0$.

2-11 已知$X=(X_1,X_2)'$的密度函数为

$$f(x_1,x_2)=\frac{1}{2\pi}\exp\left\{-\frac{1}{2}(2x_1^2+x_2^2+2x_1x_2-22x_1-14x_2+65)\right\}$$

试求X的均值和协方差阵.

解一:求边缘分布及$\mathrm{Cov}(X_1,X_2)=\sigma_{12}$

$$f_1(x_1)=\int_{-\infty}^{\infty}f(x_1,x_2)dx_2=\frac{1}{2\pi}e^{-\frac{1}{2}(2x_1^2-22x_1+65)}\int_{-\infty}^{\infty}e^{-\frac{1}{2}(x_2^2+2x_1x_2-14x_2)}dx_2$$

$$=\frac{1}{2\pi}e^{-\frac{1}{2}(2x_1^2-22x_1+65)}\int_{-\infty}^{\infty}e^{-\frac{1}{2}(x_2^2+2x_2(x_1-7)+(x_1-7)^2)}dx_2\times e^{\frac{1}{2}(x_1-7)^2}$$

$$=\frac{1}{2\pi}e^{-\frac{1}{2}(2x_1^2-22x_1+65-x_1^2+14x_1-49)}\int_{-\infty}^{\infty}e^{-\frac{1}{2}(x_2-x_1+7)^2}dx_2$$

$$=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x_1^2-8x_1+16)}\times\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-\frac{1}{2}(x_2-x_1+7)^2}dx_2$$

$$=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x_1-4)^2} \qquad \therefore X_1\sim N(4,1).$$

类似地有

$$f_2(x_2)=\int_{-\infty}^{\infty}f(x_1,x_2)dx_1=\cdots=\frac{1}{\sqrt{2\pi}\sqrt{2}}e^{-\frac{1}{4}(x_2-3)^2}$$

$$\therefore X_2\sim N(3,2).$$

$$\sigma_{12} = \mathrm{Cov}(X_1, X_2) = \mathrm{E}[(X_1 - \mathrm{E}(X_1))(X_2 - \mathrm{E}(X_2)]$$

$$= \mathrm{E}[(X_1 - 4)(X_2 - 3)] \qquad \left(令 \begin{cases} u_1 = x_1 - 4 \\ u_2 = x_2 - 3 \end{cases} \right)$$

$$= \iint (x_1 - 4)(x_2 - 3) f(x_1, x_2) dx_1 dx_2$$

$$= \iint u_1 u_2 \frac{1}{2\pi} \exp[-\frac{1}{2}(2u_1^2 + u_2^2 + 2u_1 u_2)] du_1 du_2$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u_1 e^{-\frac{u_1^2}{2}} \left[\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} u_2 e^{-\frac{1}{2}(u_2 + u_1)^2} du_2 \right] du_1$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u_1 e^{-\frac{u_1^2}{2}} \frac{1}{\sqrt{2\pi}} \left[\underbrace{\boxed{\int_{-\infty}^{\infty} (u_2 + u_1) e^{-\frac{1}{2}(u_2 + u_1)^2} du_2}}_{\| \atop 0} - u_1 \underbrace{\boxed{\int_{-\infty}^{\infty} e^{-\frac{1}{2}(u_2 + u_1)^2} du_2}}_{\| \atop \sqrt{2\pi}} \right] du_1$$

$$= -\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u_1^2 e^{-\frac{u_1^2}{2}} du_1 = -1$$

所以

$$\mathrm{E}(X)=\begin{pmatrix}4\\3\end{pmatrix}=\mu,\quad \mathrm{D}(X)=\begin{pmatrix}1&-1\\-1&2\end{pmatrix}=\Sigma$$

且 $f(x_1,x_2)=\dfrac{1}{2\pi}\exp[-\dfrac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)]$

故 $\boldsymbol{X=(X_1,X_2)'}$ 为二元正态分布.

解二:比较系数法

设 $f(x_1,x_2)=\dfrac{1}{2\pi}\exp\left\{-\dfrac{1}{2}(2x_1^2+x_2^2+2x_1x_2-22x_1-14x_2+65)\right\}$

$$=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2\sigma_1^2\sigma_2^2(1-\rho^2)}[\sigma_2^2(x_1-\mu_1)^2-2\sigma_1\sigma_2\rho(x_1-\mu_1)(x_2-\mu_2)+\sigma_1^2(x_2-\mu_2)^2]\right\}$$

比较上下式相应的系数,可得:

$$\begin{cases}\sigma_1\sigma_2\sqrt{1-\rho^2}=1\\ \sigma_2^2=2\\ \sigma_1^2=1\\ -\rho\sigma_1\sigma_2=1\\ -2\mu_1\sigma_2^2+2\rho\sigma_1\sigma_2\mu_2=-22\\ -2\mu_2\sigma_1^2+2\rho\sigma_1\sigma_2\mu_1=-14\\ \mu_1^2\sigma_2^2+\mu_2^2\sigma_1^2-2\rho\sigma_1\sigma_2\mu_1\mu_2=65\end{cases}$$

$$\Rightarrow\begin{cases}\sigma_2=\sqrt{2}\\ \sigma_1=1\\ \rho=-1/\sqrt{2}\end{cases}$$

$$\Rightarrow\begin{cases}4\mu_1+2\mu_2=22\\ 2\mu_1+2\mu_2=14\end{cases}\qquad\begin{cases}\mu_1=4\\ \mu_2=3\end{cases}$$

故$X=(X_1,X_2)'$为二元正态随机向量.且

$$\mathrm{E}(X)=\begin{pmatrix}4\\3\end{pmatrix}=\mu,\quad \mathrm{D}(X)=\begin{pmatrix}1&-1\\-1&2\end{pmatrix}=\Sigma$$

解三:两次配方法

(1)第一次配方 $2x_1^2+2x_1x_2+x_2^2=(x_1+x_2)^2+x_1^2$

因$2x_1^2+2x_1x_2+x_2^2=(x_1,x_2)\begin{pmatrix}2&1\\1&1\end{pmatrix}\begin{pmatrix}x_1\\x_2\end{pmatrix}$,而$\begin{pmatrix}2&1\\1&1\end{pmatrix}=\begin{pmatrix}1&1\\1&0\end{pmatrix}\begin{pmatrix}1&1\\1&0\end{pmatrix}=BB'$,

令$y=\begin{pmatrix}y_1\\y_2\end{pmatrix}=\begin{pmatrix}1&1\\1&0\end{pmatrix}\begin{pmatrix}x_1\\x_2\end{pmatrix}=\begin{pmatrix}x_1+x_2\\x_1\end{pmatrix}$,则$2x_1^2+2x_1x_2+x_2^2=y_1^2+y_2^2$

(2)第二次配方.由于$\begin{cases}x_1=y_2\\x_2=y_1-y_2\end{cases}$

$$2x_1^2 + x_2^2 + 2x_1x_2 - 22x_1 - 14x_2 + 65$$
$$= y_1^2 + y_2^2 - 22y_2 - 14(y_1 - y_2) + 65$$
$$= y_1^2 - 14y_1 + 49 + y_2^2 - 8y_2 + 16$$
$$= (y_1 - 7)^2 + (y_2 - 4)^2$$

即

$$\frac{1}{2\pi} e^{-\frac{1}{2}(2x_1^2 + x_2^2 + 2x_1x_2 - 22x_1 - 14x_2 + 65)} \overset{\begin{cases} x_1 = y_2 \\ x_2 = y_1 - y_2 \end{cases}}{=} \frac{1}{2\pi} e^{-\frac{1}{2}[(y_1 - 7)^2 + (y_2 - 4)^2]}$$
$$= g(y_1, y_2)$$

设函数 $g(y_1, y_2)$ 是随机向量 $\boldsymbol{Y}$ 的密度函数.

(3) 随机向量 $Y=\begin{pmatrix}Y_1\\Y_2\end{pmatrix}\sim N_2\left(\begin{pmatrix}7\\4\end{pmatrix}, I_2\right)$

(4) 由于 $X=\begin{pmatrix}X_1\\X_2\end{pmatrix}=\begin{pmatrix}0&1\\1&-1\end{pmatrix}\begin{pmatrix}Y_1\\Y_2\end{pmatrix}=CY$

$$\begin{pmatrix}0&1\\1&-1\end{pmatrix}\begin{pmatrix}7\\4\end{pmatrix}=\begin{pmatrix}4\\3\end{pmatrix},\quad \begin{pmatrix}0&1\\1&-1\end{pmatrix}I_2\begin{pmatrix}0&1\\1&-1\end{pmatrix}=\begin{pmatrix}1&-1\\-1&2\end{pmatrix}$$

故 $X=CY\sim N_2\left(\begin{pmatrix}4\\3\end{pmatrix}, \begin{pmatrix}1&-1\\-1&2\end{pmatrix}\right)$

$$\mathrm{E}(X)=\begin{pmatrix}4\\3\end{pmatrix},\quad \mathrm{D}(X)=\begin{pmatrix}1&-1\\-1&2\end{pmatrix}$$

2-12 设X_1~N(0,1),令

$$X_2=\begin{cases}-X_1, & 当 -1\le X_1\le 1,\\ X_1, & 其它.\end{cases}$$

(1)证明X_2~N(0,1);

(2)证明(X_1, X_2)不是二元正态分布.

证明(1):任给x,当$x\le -1$时

$$P\{X_2\le x\}=P\{X_1\le x\}=\Phi(x)$$

当$x\ge 1$时, $P\{X_2\le x\}$

$$=P\{X_2\le -1\}+P\{-1<X_2\le 1\}+P\{1<X_2\le x\}$$

$$=P\{X_1\le -1\}+P\{-1<-X_1\le 1\}+P\{1<X_1\le x\}$$

$$=P\{X_1\le x\}=\Phi(x)$$

当-1≤*x*≤1时,

$$P\{X_2 \le x\} = P\{X_2 \le -1\} + P\{-1 < X_2 \le x\}$$
$$= P\{X_1 \le -1\} + P\{-x \le X_1 < 1\}$$
$$= P\{X_1 \le -1\} + P\{-1 < X_1 < x\}$$
$$= P\{X_1 \le x\} = \Phi(x)$$

$$\therefore \quad X_2 \sim N(0,1).$$

(2) 考虑随机变量*Y*= *X*₁-*X*₂ ,显然有

$$Y = X_1 - X_2 = \begin{cases} X_1 + X_1, & \text{当-}1 \le X_1 \le 1 \\ 0 & \text{其它} \end{cases}$$

$$P\{Y=0\}=P\{X_1>1\text{或}X_1<-1\}$$
$$=P\{X_1>1\}+P\{X_1<-1\} \quad (X_1\sim N(0,1))$$
$$=2\Phi(-1)=0.3174\neq 0$$

若(X_1, X_2)是二元正态分布,则由性质4可知,它的任意线性组合必为一元正态. 但$Y=X_1-X_2$不是正态分布,故(X_1, X_2)不是二元正态分布.

2-17 设$X \sim N_p(\mu, \Sigma)$, $\Sigma > 0$, X的密度函数记为$f(x;\mu,\Sigma)$. (1) 任给$a>0$, 试证明概率密度等高面

$$f(x;\mu,\Sigma) = a$$

是一个椭球面.

(2) 当$p=2$且 $\Sigma = \sigma^2\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ $(\rho>0)$时,

概率密度等高面就是平面上的一个椭圆，试求该椭圆的方程式，长轴和短轴.

证明(1):任给$a>0$, 记$a_0 = (2\pi)^{p/2}|\Sigma|^{1/2}$,当$0<a<\dfrac{1}{a_0}$时,

$$f(x;\mu,\Sigma) = a \Leftrightarrow (x-\mu)'\Sigma^{-1}(x-\mu) = b^2$$

其中 $b^2 = -2\ln[a(2\pi)^{p/2}|\Sigma|^{1/2}] = -2\ln[aa_0] > 0,$

因$\Sigma > 0$, Σ的特征值记为$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$, λ_i对应的特征向量记特$l_i (i = 1,2,\cdots,p)$,则有Σ^{-1}的谱谱分解

$$\Sigma^{-1} = \sum_{i=1}^{p} \frac{1}{\lambda_i} l_i l_i'$$

(见附录§5 P390)

令$y_i = (x-\mu)' l_i \ (i = 1,2,\cdots,p)$,则概率密度等高面为

$$(x-\mu)'\Sigma^{-1}(x-\mu) = (x-\mu)' \sum_{i=1}^{p} \frac{1}{\lambda_i} l_i l_i' (x-\mu) = b^2$$

$$\Leftrightarrow \quad \sum_{i=1}^{p} \frac{1}{\lambda_i} y_i^2 = b^2$$

$$\Leftrightarrow \quad \frac{y_1^2}{\lambda_1 b^2}+\frac{y_2^2}{\lambda_2 b^2}+\cdots+\frac{y_p^2}{\lambda_p b^2}=1$$

故概率密度等高面 $f(x;\mu,\Sigma)=a$ 是一个椭球面.

(2)当 $p=2$ 且 $\Sigma=\sigma^2\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}(\rho>0)$ 时， $|\Sigma|=\sigma^4(1-\rho^2)$.

由 $|\Sigma-\lambda I_p|=\begin{vmatrix}\sigma^2-\lambda & \sigma^2\rho\\ \sigma^2\rho & \sigma^2-\lambda\end{vmatrix}=(\sigma^2-\lambda)^2-\sigma^4\rho^2$

$$=(\sigma^2-\lambda-\sigma^2\rho)(\sigma^2-\lambda+\sigma^2\rho)=0$$

可得Σ的特征值 $\lambda_1=\sigma^2(1+\rho),\lambda_2=\sigma^2(1-\rho)$.

λ_i (i=1,2)对应的特征向量为

$$l_1=\begin{pmatrix}\frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}}\end{pmatrix}\quad l_1=\begin{pmatrix}\frac{1}{\sqrt{2}}\\ -\frac{1}{\sqrt{2}}\end{pmatrix}$$

由(1)可得椭圆方程为

$$\frac{y_1^2}{\sigma^2(1+\rho)b^2}+\frac{y_2^2}{\sigma^2(1-\rho)b^2}=1$$

其中 $b^2=-2\ln[a(2\pi)\ |\Sigma|^{1/2}]=-2\ln[2\pi\sigma^2\sqrt{1-\rho^2}a]$,

长轴半径为 $d_1=b\sigma\sqrt{1+\rho}$, 方向沿着l_1方向(b>0);

短轴半径为$d_2=b\sigma\sqrt{1-\rho}$, 方向沿着l_2方向.

2-19 为了了解某种橡胶的性能，今抽了十个样品，每个测量了三项指标：硬度、变形和弹性，其数据见表。试计算样本均值，样本离差阵，样本协差阵和样本相关阵.

解：求 A, S, R.

$$A=\sum_{\alpha=1}^{10}(X_{(\alpha)}-\bar{X})(X_{(\alpha)}-\bar{X})'$$

$$=X'\left[I_n-\frac{1}{n}J\right]X \qquad J_{n\times n}=\begin{pmatrix}1&1&\cdots&1\\1&1&\cdots&1\\\vdots&\vdots&&\vdots\\1&1&\cdots&1\end{pmatrix}$$

$$S=\frac{1}{n-1}A=(s_{ij})_{p\times p}.$$

$$R = D_s S D_s \overset{\text{或}}{=} D_a A D_a$$

其中 $D_s = \begin{pmatrix} \frac{1}{\sqrt{s_{11}}} & & 0 \\ & \frac{1}{\sqrt{s_{22}}} \ddots & \\ 0 & & \frac{1}{\sqrt{s_{pp}}} \end{pmatrix}$

$$D_a = \begin{pmatrix} \frac{1}{\sqrt{a_{11}}} & \ddots & 0 \\ 0 & & \frac{1}{\sqrt{a_{pp}}} \end{pmatrix}$$

应用多元统计分析

第三章习题解答

3-1 设$X \sim N_n(\mu, \sigma^2 I_n)$，$A$为对称幂等阵, 且$\mathrm{rk}(A)=r(r\le n)$, 证明

$$\frac{1}{\sigma^2}X'AX \sim \chi^2(r,\delta), \quad \text{其中 } \delta=\frac{1}{\sigma^2}\mu'A\mu.$$

证明　因A为对称幂等阵，而对称幂等阵的特征值非0即1, 且只有r个非0特征值，即存在正交阵Γ(其列向量r_i为相应特征向量)，使

$$\Gamma'A\Gamma=\begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}, \text{记 } \Gamma=(r_1,\cdots,r_n)$$

令 $Y=\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}=\Gamma' X$（即 $X=\Gamma Y$），则

$$Y \sim N_n(\Gamma'\mu,\sigma^2\Gamma' I_n\Gamma)=N_n(\Gamma'\mu,\sigma^2 I_n)$$

$$\frac{1}{\sigma^2}X'AX=\frac{1}{\sigma^2}Y'\Gamma' A\Gamma Y=\frac{1}{\sigma^2}Y'\begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}Y=\frac{1}{\sigma^2}\sum_{i=1}^{r} Y_i^2$$

因为 $Y_i \sim N(r_i'\mu,\sigma^2)$ $(i=1,2,\cdots,r)$，且相互独立

所以 $\xi=\dfrac{1}{\sigma^2}X'AX=\dfrac{1}{\sigma^2}\sum\limits_{i=1}^{r} Y_i^2 \sim \chi^2(r,\delta)$，

其中非中心参数为

$$\delta = \frac{1}{\sigma^2}\sum_{i=1}^{r}(r_i'\mu)^2 = \frac{1}{\sigma^2}[\mu'(r_1r_1' + \cdots + r_rr_r')\mu]$$

$$= \frac{1}{\sigma^2}\mu'(r_1,\cdots,r_r)\begin{bmatrix} r_1' \\ \vdots \\ r_r' \end{bmatrix}\mu = \frac{1}{\sigma^2}\mu'\Gamma\begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}\Gamma'\mu$$

$$= \frac{1}{\sigma^2}\mu' A\mu.$$

第三章 多元正态总体参数的检验

3-2 设$X \sim N_n(\mu,\sigma^2 I_n)$， A，B为n阶对称阵.
若$AB=0$，证明$X'AX$与$X'BX$相互独立.

证明的思路：记$\mathrm{rk}(A)=r$.

因A为n阶对称阵，存在正交阵Γ，使得

$$\Gamma' A\Gamma=\mathrm{diag}(\lambda_1,\ldots,\lambda_r,0,\ldots,0)$$

令$Y=\Gamma'X$，则$Y \sim N_n(\Gamma'\mu,\sigma^2 I_n)$，

且 $$\xi = X'AX = (\Gamma Y)' A\Gamma\Gamma = Y'\Gamma' A\Gamma\Gamma = \sum_{i=1}^{r} \lambda_i Y_i^2$$

又因为

$$\mathbf{X}'BX = Y'\Gamma'B\Gamma Y = Y'HY$$

其中$H=\Gamma'B\Gamma$。如果能够证明$\mathbf{X}'BX$可表示为Y_{r+1}，…, Y_n的函数，即H只是右下子块为非0的矩阵。

则$\mathbf{X}'AX$与$\mathbf{X}'BX$相互独立。

证明 记$\mathrm{rk}(A)=r$.

若$r=n$, 由$AB=O$, 知$B=O_{n\times n}$, 于是$X'AX$与$X'BX$独立;

若$r=0$时, 则$A=0$, 则两个二次型也是独立的.

以下设$0<r<n$.因A为n阶对称阵, 存在正交阵Γ,使得

$$\Gamma'A\Gamma=\begin{bmatrix} D_r & 0 \\ 0 & 0 \end{bmatrix}, D_r=\begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_r \end{bmatrix}$$

其中$\lambda_i \neq 0$为A的特征值(i=1, …, r). 于是

$$A = \Gamma \left[\begin{array}{c:c} D_r & 0 \\ \hdashline 0 & 0 \end{array}\right] \Gamma', AB = \Gamma \left[\begin{array}{c:c} D_r & 0 \\ \hdashline 0 & 0 \end{array}\right] \Gamma' \cdot B\Gamma\Gamma',$$

令 $$\underset{n\times n}{H} = \Gamma' B \Gamma \triangleq \left[\begin{array}{c:c} H_{11} & H_{12} \\ \hdashline H_{21} & H_{22} \end{array}\right],$$ 其中 H_{11} 为r阶方阵.

$$AB = \Gamma \left[\begin{array}{c:c} D_r & 0 \\ \hdashline 0 & 0 \end{array}\right] \left[\begin{array}{c:c} H_{11} & H_{12} \\ \hdashline H_{21} & H_{22} \end{array}\right] \Gamma' = \Gamma \left[\begin{array}{c:c} D_r H_{11} & D_r H_{12} \\ \hdashline 0 & 0 \end{array}\right] \Gamma'$$

由$AB=O$可得$D_r H_{11}=O$ ， $D_r H_{12}=O$.

因D_r为满秩阵, 故有$H_{11}=O_{r\times r}$, $H_{12}=O_{r\times(n-r)}$.

由于H为对称阵，所以$H_{21}=O_{(n-r)\times r}$. 于是

$$H = \Gamma' B \Gamma = \begin{array}{c} \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & H_{22} \end{array}\right] \end{array} \begin{array}{l} r \\ n-r \end{array}$$

令$\boldsymbol{Y}=\boldsymbol{\Gamma}'\ \boldsymbol{X}$，则$\boldsymbol{Y}\sim N_{\boldsymbol{n}}(\boldsymbol{\Gamma}'\ \boldsymbol{\mu},\boldsymbol{\sigma}^2\boldsymbol{I}_{\boldsymbol{n}})$，且

$$\xi = X'AX = (\Gamma Y)' A\Gamma\Gamma = Y'\Gamma' A\Gamma\Gamma = \sum_{i=1}^{r} \lambda_i Y_i^2$$

$$\eta = X'BX = Y'\Gamma' B\Gamma\Gamma = Y'HY = (Y_{r+1}, \cdots, Y_n) H_{22} \begin{bmatrix} Y_{r+1} \\ \vdots \\ Y_n \end{bmatrix}$$

由于$\boldsymbol{Y}_1$，$\ldots$, $\boldsymbol{Y}_{\boldsymbol{r}}$, $\boldsymbol{Y}_{\boldsymbol{r+1}}$, $\ldots$, $\boldsymbol{Y}_{\boldsymbol{n}}$相互独立，故$\boldsymbol{X'AX}$与$\boldsymbol{X'BX}$相互独立.

3-3 设$X \sim N_p(\mu,\Sigma),\Sigma>0$, A和B为p阶对称阵,
试证明
$(X-\mu)'A(X-\mu)$与$(X-\mu)'B(X-\mu)$相互独立

$$\Leftrightarrow \Sigma A\Sigma B\Sigma=0_{p\times p}.$$

证明 由于 $\Sigma=\Sigma^{\frac{1}{2}}\cdot\Sigma^{\frac{1}{2}}>0$, 令

$$Y=\Sigma^{-\frac{1}{2}}(X-\mu)\sim N_p(0,I_p)$$

(记 $\Sigma^{-\frac{1}{2}}=\left(\Sigma^{\frac{1}{2}}\right)^{-1}$)

$$\xi = (X-\mu)'A(X-\mu) = Y'\Sigma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}}Y \triangleq Y'CY$$

$$\eta = (X-\mu)'B(X-\mu) = Y'\Sigma^{\frac{1}{2}}B\Sigma^{\frac{1}{2}}Y \triangleq Y'DY.$$

由“1. 结论6”知ξ与η相互独立⇔

$$CD = O \Longleftrightarrow \Sigma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}} \cdot \Sigma^{\frac{1}{2}}B\Sigma^{\frac{1}{2}} = O$$

$$\Longleftrightarrow \Sigma A \Sigma B \Sigma = O$$

3-4 试证明Wishart分布的性质(4)和T^2分布的性质(5).

性质4 分块Wishart矩阵的分布:设$X_{(\alpha)} \sim N_p(0,\Sigma)$ ($\alpha=1,\ldots,n$)相互独立，其中

$$\Sigma = \begin{array}{cc} \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} & \begin{array}{c} r \\ p-r \end{array} \end{array}$$

又已知随机矩阵

$$W = \sum_{\alpha=1}^{n} X_{(\alpha)} X'_{(\alpha)} = \begin{array}{cc} \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} & \begin{array}{c} r \\ p-r \end{array} \end{array} \sim W_p(n,\Sigma)$$

则 ① $W_{11} \sim W_r(n,\Sigma_{11})$，$W_{22} \sim W_{p-r}(n,\Sigma_{22})$

② 当$\Sigma_{12}=0$时，W_{11}与W_{22}相互独立.

证明：设

$$X_{(\alpha)} = \begin{pmatrix} X_{(\alpha)}^{(1)} \\ X_{(\alpha)}^{(2)} \end{pmatrix} \begin{matrix} r \\ p-r \end{matrix}, \text{则} \begin{matrix} X_{(\alpha)}^{(1)} \sim N_r(0, \Sigma_{11}), \\ X_{(\alpha)}^{(2)} \sim N_{p-r}(0, \Sigma_{22}), \end{matrix}$$

记 $$\underset{n\times p}{X} = (x_{ij}) = \left(\underset{n\times r}{X(1)} \mid \underset{n\times(p-r)}{X(2)} \right)$$，则

$$W = X'X = \begin{pmatrix} X(1)'X(1) & X(1)'X(2) \\ X(2)'X(1) & X(2)'X(2) \end{pmatrix} = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix},$$

即

$$W_{11} = X(1)'X(1), \quad W_{22} = X(2)'X(2)$$

由定义3.1.4可知

$$W_{11}=X(1)'X(1)=\sum_{\alpha=1}^{n}(X_{(\alpha)}^{(1)})'X_{(\alpha)}^{(1)}\sim W_r(n,\Sigma_{11});$$

$$W_{22}=X(2)'X(2)=\sum_{\alpha=1}^{n}(X_{(\alpha)}^{(2)})'X_{(\alpha)}^{(2)}\sim W_{p-r}(n,\Sigma_{22}).$$

当$\Sigma_{12}=O$时，对$\alpha=1,2,\ldots,n$，$X_{(\alpha)}^{(1)}$与$X_{(\alpha)}^{(2)}$相互独立。故有W_{11}与W_{22}相互独立。

性质5 在非退化的线性变换下,T^2统计量保持不变.

证明: 设$X_{(\alpha)}$ ($\alpha=1,\ldots,n$) 是来自p元总体$N_p(\mu,\Sigma)$的随机样本, $\overline{X}$和A_x分别表示正态总体X的样本均值向量和离差阵,则由性质1有

$$T_x^2 = n(n-1)(\overline{X}-\mu)' A_x^{-1} (\overline{X}-\mu) \sim T^2(p, n-1).$$

令 $Y_{(i)} = CX_{(i)} + d(i=1,\ldots,n)$

其中C是$p\times p$非退化常数矩阵，d是$p\times 1$常向量。

则 $Y_{(i)} \sim N_p(C\mu + d, C\Sigma C') \quad (i=1,2,\ldots,n)$

$$\bar{Y} = C\bar{X} + d, \qquad \text{记}\mu_{\mathrm{y}} = C\mu + d$$

$$A_y = \sum_{i=1}^{n}(Y_{(i)} - \bar{Y})(Y_{(i)} - \bar{Y})'$$

$$= C[\sum_{i=1}^{n}(X_{(i)} - \bar{X})(X_{(i)} - \bar{X})']C' = CA_xC'$$

$$T_y^2 = n(n-1)(\bar{Y} - \mu_y)' A_y^{-1} (\bar{Y} - \mu_y)$$

$$= n(n-1)(\bar{X} - \mu)' C' [CA_xC']^{-1} C(\bar{X} - \mu)$$

$$= n(n-1)(\bar{X} - \mu) A_x^{-1} (\bar{X} - \mu) = T_x^2$$

所以 $$T_x^2 = T_y^2$$

3-5 对单个p维正态总体$N_p(\mu,\Sigma)$均值向量的检验问题，试用似然比原理导出检验$H_0:\mu=\mu_0(\Sigma=\Sigma_0$已知)的似然比统计量及分布.

P66当$\Sigma=\Sigma_0$已知μ的检验

解: 总体$X\sim N_p(\mu,\Sigma_0)(\Sigma_0>0)$，设$X_{(\alpha)}(\alpha=1,\dots,n)$ $(n>p)$为来自p维正态总体X的样本. 似然比统计量为

$$\lambda=\max_{\mu=\mu_0}L(\mu_0,\Sigma_0)\Big/\max_{\mu}L(\mu,\Sigma_0)$$

$$分子=\frac{1}{|2\pi\Sigma_0|^{n/2}}\exp\left[-\frac{1}{2}\sum_{\alpha=1}^{n}(X_{(\alpha)}-\mu_0)'\Sigma_0^{-1}(X_{(\alpha)}-\mu_0)\right]$$

$$=\frac{1}{|2\pi\Sigma_0|^{n/2}}\exp\left[-\frac{1}{2}\mathrm{tr}[\Sigma_0^{-1}\sum_{\alpha=1}^{n}(X_{(\alpha)}-\mu_0)(X_{(\alpha)}-\mu_0)']\right]$$

$$分子=\frac{1}{|2\pi\Sigma_0|^{n/2}}\exp\left[-\frac{1}{2}\mathrm{tr}[\Sigma_0^{-1}A_0]\right]$$

$$分母=L(\bar{X},\Sigma_0)=\max_{\mu}L(\mu,\Sigma_0)$$

$$=\frac{1}{|2\pi\Sigma_0|^{n/2}}\exp\left[-\frac{1}{2}\sum_{\alpha=1}^{n}(X_{(\alpha)}-\bar{X})'\Sigma_0^{-1}(X_{(\alpha)}-\bar{X})\right]$$

$$=\frac{1}{|2\pi\Sigma_0|^{n/2}}\exp\left[-\frac{1}{2}\mathrm{tr}[\Sigma_0^{-1}\sum_{\alpha=1}^{n}(X_{(\alpha)}-\bar{X})(X_{(\alpha)}-\bar{X})']\right]$$

$$=\frac{1}{|2\pi\Sigma_0|^{n/2}}\exp\left[-\frac{1}{2}\mathrm{tr}[\Sigma_0^{-1}A\ ]\right]$$

$$\lambda = \max_{\mu=\mu_0} L(\mu_0,\Sigma_0) \Big/ \max_{\mu} L(\mu,\Sigma_0)$$

$$= \exp\left[\mathrm{tr}[\frac{1}{2}\Sigma_0^{-1}A] - \mathrm{tr}[\frac{1}{2}\Sigma_0^{-1}A_0] \right]$$

$$= \exp\left[\mathrm{tr}[\frac{1}{2}\Sigma_0^{-1}A] - \mathrm{tr}[\frac{1}{2}\Sigma_0^{-1}(A + n(\bar{X}-\mu_0)(\bar{X}-\mu_0)'] \right]$$

$$= \exp\left[-\frac{n}{2}\mathrm{tr}[(\bar{X}-\mu_0)'\Sigma_0^{-1}(\bar{X}-\mu_0)] \right]$$

$$= \exp\left[-\frac{n}{2}(\bar{X}-\mu_0)'\Sigma_0^{-1}(\bar{X}-\mu_0) \right]$$

$$\ln\lambda = -\frac{n}{2}(\overline{X}-\mu_0)'\Sigma_0^{-1}(\overline{X}-\mu_0)$$

$$-2\ln\lambda = n(\overline{X}-\mu_0)'\Sigma_0^{-1}(\overline{X}-\mu_0) \overset{\text{def}}{=} \xi$$

因 $\overline{X} \overset{H_0\text{下}}{\sim} N_p(\mu_0, \frac{1}{n}\Sigma_0)$, $\sqrt{n}(\overline{X}-\mu_0) \overset{H_0\text{下}}{\sim} N_p(0,\Sigma_0)$

所以由§3“一、2.的结论1”可知

$$\xi = -2\ln\lambda \sim \chi^2(p).$$

3-6（均值向量各分量间结构关系的检验）设总体 $X \sim N_p(\mu, \Sigma)(\Sigma>0)$，$X_{(\alpha)}$ （$\alpha=1,\ldots,n$）($n>p$)为来自p维正态总体X的样本，记 $\mu=(\mu_1,\ldots,\mu_p)'$. C为$k\times p$常数($k<p$)，rank(C)=k, r为已知k维向量. 试给出检验$H_0: C\mu=r$的检验统计量及分布.

解：令 $Y_{(\alpha)}=CX_{(\alpha)}(\alpha=1,2,\cdots,n)$

则$Y_{(\alpha)}$ ($\alpha=1,\ldots,n$) 为来自k维正态总体Y的样本，且

$$Y_{(\alpha)} \sim N_k(C\mu, C\Sigma C');\text{记}\mu_y = C\mu, \Sigma_y = C\Sigma C'.$$

检验 $H_0: C\mu = r \Longleftrightarrow H_0: \mu_y = r$

这是单个**k**维正态总体均值向量的检验问题.利用 §**3.2**当$\boldsymbol{\Sigma_y = C\Sigma C'}$ 未知时均值向量的检验给出的结论, 取检验统计量:

$$F = \frac{n-k}{(n-1)k} T^2 \overset{H_0\text{下}}{\sim} F(k, n-k)$$

其中 $T^2 = (n-1)n(\bar{Y} - r)' \left[A_y\right]^{-1} (\bar{Y} - r).$

$$= (n-1)n(C\bar{X} - r)' \left[CAC'\right]^{-1} (C\bar{X} - r).$$

$$A = \sum_{i=1}^{n} (X_{(i)} - \bar{X})(X_{(i)} - \bar{X})'.$$

3-7 设总体X～Np(μ，Σ) (Σ>0), $X_{(\alpha)}$ (α=1,…,n)(n>p)为来自p维正态总体X的样本，样本均值为X，样本离差阵为A.记μ=($\mu_1,\ldots,\mu_p$)′.为检验$H_0:\mu_1=\mu_2=\ldots=\mu_p$,$H_1:\mu_1,\mu_2,\ldots,\mu_p$至少有一对不相等.令

$$C=\begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & -1 \end{pmatrix}_{(p-1)\times p},$$

则上面的假设等价于H_0：$C\mu=0_{p-1}$,H_1：$C\mu\neq 0_{p-1}$

试求检验H_0 的似然比统计量和分布.

解：$H_0:\mu_1=\mu_2=\cdots=\mu_p,$

$H_1:\mu_1,\mu_2,\cdots,\mu_p$ 至少有一对不相等.

$$\Longleftrightarrow \quad H_0: C\mu = 0, H_1: C\mu \neq 0,$$

利用3-6的结果知，检验$\mathbf{H_0}$的似然比统计量及分布为：

$$F = \frac{n-(p-1)}{(n-1)(p-1)} T^2 \overset{H_0\text{下}}{\sim} F(p-1, n-p+1),$$

其中 $$T^2 = (n-1)n(C\overline{X})'[CAC']^{-1} C\overline{X}$$

$$\sim T^2(n-1, p-1).(H_0\text{下})$$

(注意:3-6中的*k*在这里为*p*-1)

3-8 假定人体尺寸有这样的一般规律:身高(X_1),胸围(X_2)和上半臂围(X_3)的平均尺寸比例是6：4：1.假设$X_{(\alpha)}$(α＝1,…,n)为来自总体X=(X_1,X_2,X_3)′的随机样本.并设X～$N_3(\mu,\Sigma)$，试利用表3.5中男婴这一组数据检验三个尺寸(变量)是否符合这一规律(写出假设H_0,并导出检验统计量).

解：检验三个尺寸(变量)是否符合这一规律的问题可提成假设检验问题.因为

$$\mu_1:\mu_2:\mu_3=6:4:1 \Longleftrightarrow C\mu=0$$

其中 $C=\begin{pmatrix} 1 & 0 & -6 \\ 0 & 1 & -4 \end{pmatrix}_{2\times 3}$,

注意: $\dfrac{\mu_1}{\mu_3}=\dfrac{6}{1}$, 且 $\dfrac{\mu_2}{\mu_3}=\dfrac{4}{1}$

$$\Leftrightarrow \begin{cases} \mu_1-6\mu_3=0 \\ \mu_2-4\mu_3=0 \end{cases}.$$

$$\text{或}\quad C=\begin{pmatrix}2 & -3 & 0\\ 1 & 0 & -6\end{pmatrix},\text{或}\quad C=\begin{pmatrix}2 & -3 & 0\\ 0 & 1 & -4\end{pmatrix}$$

检验的假设**H**$_0$为 $H_0: C\mu=0, H_1: C\mu\neq 0,$

利用3-6的结论，取检验统计量为：

$$F=\frac{n-2}{2(n-1)}T^2 \overset{H_0\text{下}}{\sim} F(2,n-2)$$

$$T^2=(n-1)n(C\overline{X})'[XAC']^{-1}C\overline{X}.$$

由男婴测量数据(***p*=3,*n*=6**)计算可得

T^2=47.1434, *F*=18.8574, *p*值=0.009195<α=0.05,

故否定**H**$_0$,即认为这组数据与人类的一般规律不一致.

3-9 对单个p维正态总体$N_p(\mu,\Sigma)$协差阵的检验问题，试用似然比原理导出检验$H_0:\Sigma=\Sigma_0$的似然比统计量及分布.

解:总体$X \sim N_p(\mu, \Sigma)$，设$X_{(\alpha)}$（$\alpha=1, \dots, n$）为来自p维正态总体X的样本. 似然比统计量为

$$\lambda = \max_{\mu} L(\mu, \Sigma_0) \Big/ \max_{\mu,\Sigma} L(\mu, \Sigma)$$

分子当$\hat{\mu}=\overline{X}$达最大,且最大值

$$L(\overline{X}, \Sigma_0) = \frac{1}{|2\pi\Sigma_0|^{n/2}} \exp\left[-\frac{1}{2} \sum_{\alpha=1}^{n} (X_{(\alpha)} - \overline{X})' \Sigma_0^{-1} (X_{(\alpha)} - \overline{X}) \right]$$

$$=\frac{1}{|2\pi\Sigma_0|^{n/2}}\exp\left[-\frac{1}{2}\mathrm{tr}[\Sigma_0^{-1}\sum_{\alpha=1}^{n}(X_{(\alpha)}-\overline{X})(X_{(\alpha)}-\overline{X})']\right]$$

$$=(2\pi)^{-\frac{np}{2}}|\Sigma_0|^{-\frac{n}{2}}\mathrm{etr}\left[-\frac{1}{2}\Sigma_0^{-1}A\right]$$

$$\text{分母}=L(\overline{X},\frac{1}{n}A)=\max_{\mu,\Sigma}L(\mu,\Sigma)$$

$$=\left(\frac{n}{2\pi e}\right)^{\frac{np}{2}}|A|^{-\frac{n}{2}}=(2\pi)^{-\frac{np}{2}}\left(\frac{n}{e}\right)^{\frac{np}{2}}|A|^{-\frac{n}{2}}$$

$$\lambda = \max_{\mu} L(\mu\ ,\Sigma_0) \Big/ \max_{\mu,\Sigma} L(\mu,\Sigma)$$

$$= \frac{|\Sigma_0|^{-\frac{n}{2}} \operatorname{etr}\left(-\frac{1}{2}\Sigma_0^{-1}A\right)}{\left(\frac{n}{e}\right)^{\frac{np}{2}} |A|^{-\frac{n}{2}}}$$

$$= \left(\frac{e}{n}\right)^{\frac{np}{2}} \operatorname{etr}\left(-\frac{1}{2}\Sigma_0^{-1}A\right) |\Sigma_0^{-1}A|^{\frac{n}{2}}$$

由定理**3.2.1**,当n充分大时,有 $-2\ln\lambda \sim \chi^2\left(\frac{p(p+1)}{2}\right).$

3-10 对两个p维正态总体$N_p(\mu^{(1)},\Sigma)$和$N_p(\mu^{(2)},\Sigma)$均值向量的检验问题，试用似然比原理导出检验H_0: $\mu^{(1)}=\mu^{(2)}$的似然比统计量及分布.

解:设 $X_{(\alpha)}^{(i)}$ $(\alpha=1,\dots,n_i)$为来自总体$X\sim N_p(\mu^{(i)},\Sigma)$的随机样本($i$=1,2),且相互独立,$\Sigma>0$未知.检验$H_0$似然比统计量为

$$\lambda=\max_{\mu,\Sigma>0}L(\mu\ ,\Sigma)\Big/\max_{\mu^{(1)},\mu^{(2)},\Sigma>0}L(\mu^{(1)},\mu^{(2)},\Sigma)$$

记

$$A_i=\sum_{\alpha=1}^{n_i}(X_{(\alpha)}^{(i)}-\bar{X}^{(i)})(X_{(\alpha)}^{(i)}-\bar{X}^{(i)})' \quad (i=1,2) \quad n=n_1+n_2$$

其中 $\bar{X}^{(i)}=\dfrac{1}{n_i}\sum\limits_{\alpha=1}^{n_i}X_{(\alpha)}^{(i)}$ $(i=1,2)$, 记$\bar{X}=\dfrac{1}{n}\sum\limits_{i=1}^{2}\sum\limits_{\alpha=1}^{n_i}X_{(\alpha)}^{(i)}$,

$$T=\sum_{i=1}^{2}\sum_{j=1}^{n_k}(X_{(j)}^{(i)}-\overline{X})(X_{(j)}^{(i)}-\overline{X})'$$

$$=\sum_{i=1}^{2}A_i+\sum_{i=1}^{2}n_i(\overline{X}^{(i)}-\overline{X})(\overline{X}^{(i)}-\overline{X})'=A+B$$

其中 $\boldsymbol{A=A_1+A_2}$称为组内离差阵.$\boldsymbol{B}$称为组间离差阵.

分子当$\hat{\mu}=\overline{X},\hat{\Sigma}=\dfrac{T}{n}=\dfrac{A+B}{n}$达最大,且最大值为

$$L\ (\overline{X},\frac{T}{n})=(2\pi)^{-\frac{np}{2}}\left(\frac{n}{e}\right)^{\frac{np}{2}}|T|^{-\frac{n}{2}}$$

分母当$\hat{\mu}^{(1)} = \overline{X}^{(1)}, \hat{\mu}^{(2)} = \overline{X}^{(2)}, \hat{\Sigma} = \dfrac{A}{n}$达最大,

且最大值为 $L\ (\overline{X}^{(1)}, \overline{X}^{(2)}, \dfrac{A}{n}) = (2\pi)^{-\frac{np}{2}} \left(\dfrac{n}{e}\right)^{\frac{np}{2}} |A|^{-\frac{n}{2}}$

似然比统计量 $\lambda = \dfrac{|A|^{n/2}}{|T|^{n/2}} = \left(\dfrac{|A|}{|A+B|}\right)^{n/2} = \Lambda^{n/2}$

因为 $T = A + B = A + \sum_{i=1}^{2} n_i (\overline{X}^{(i)} - \overline{X})(\overline{X}^{(i)} - \overline{X})'$

$$= A + \frac{n_1 n_2}{n} (\overline{X}^{(1)} - \overline{X}^{(2)})(\overline{X}^{(1)} - \overline{X}^{(2)})'$$

$$|T|=\left|A+\frac{n_1n_2}{n}(\overline{X}^{(1)}-\overline{X}^{(2)})(\overline{X}^{(1)}-\overline{X}^{(2)})'\right|$$

$$=\begin{vmatrix} A & -\sqrt{\frac{n_1n_2}{n}}(\overline{X}^{(1)}-\overline{X}^{(2)}) \\ \sqrt{\frac{n_1n_2}{n}}(\overline{X}^{(1)}-\overline{X}^{(2)})' & 1 \end{vmatrix}$$

$$=|A|\left[1+\frac{n_1n_2}{n}(\overline{X}^{(1)}-\overline{X}^{(2)})'A^{-1}(\overline{X}^{(1)}-\overline{X}^{(2)})\right]$$

所以 $\dfrac{|A|}{|T|}=\dfrac{1}{1+\frac{n_1n_2}{n}(\overline{X}^{(1)}-\overline{X}^{(2)})'A^{-1}(\overline{X}^{(1)}-\overline{X}^{(2)})}$

由于 $\sqrt{\frac{n_1 n_2}{n}}(\overline{X}^{(1)}-\overline{X}^{(2)}) \overset{H_0\text{下}}{\sim} N_p(0,\Sigma)$

$$A = A_1 + A_2 \sim W_p(n-2,\Sigma),(n = n_1 + n_2)$$

由定义3.1.5可知

$$T^2 = (n-2)\frac{n_1 n_2}{n}(\overline{X}^{(1)}-\overline{X}^{(2)})' A^{-1}(\overline{X}^{(1)}-\overline{X}^{(2)})$$

$$\sim T^2(p,n-2)$$

由 $\Lambda = \frac{|A|}{|T|} = \frac{1}{1+\frac{1}{n-2}T^2}$， 或 $\frac{1}{n-2}T^2 = \frac{1-\Lambda}{\Lambda}$

可取检验统计量为

$$F=\frac{(n-2)-p+1}{p}\frac{T^2}{n-2}=\frac{n-p-1}{p}\frac{1-\Lambda}{\Lambda}$$

$$\overset{H_0\text{下}}{\sim} F(p,n-p-1)$$

检验假设$\mathbf{H_0}$的否定域为

$$\{\lambda<\lambda_\alpha\}\Longleftrightarrow\{\Lambda<\Lambda_\alpha\}\Longleftrightarrow\{T^2>T_\alpha^2\}$$

$$\Longleftrightarrow\{F>F_\alpha\}$$

3-11 表3.5给出15名2周岁婴儿的身高(X_1)，胸围(X_2)和上半臂围(X_3)的测量数据.假设男婴的测量数据$X_{(\alpha)}$ ($\alpha=1,\ldots,6$)为来自总体$N_3(\mu^{(1)}, \Sigma)$的随机样本.女婴的测量数据$Y_{(\alpha)}$ ($\alpha=1,\ldots,9$)为来自总体$N_3(\mu^{(2)}, \Sigma)$的随机样本.试利用表3.5中的数据检验$H_0:\mu^{(1)}=\mu^{(2)}$ ($\alpha=0.05$).

解:这是两总体均值向量的检验问题. 检验统计量取为($p=3,n=6,m=9$):

$$F=\frac{n+m-p-1}{(n+m-2)p}T^2 \overset{H_0下}{\sim} F(p,n+m-p-1)$$

其中 $T^2=(n+m-2)\dfrac{nm}{n+m}(\overline{X}-\overline{Y})'(A_1+A_2)^{-1}(\overline{X}-\overline{Y})$

故检验统计量为

$$F=\frac{n+m-p-1}{p}\times\frac{nm}{n+m}(\overline{X}-\overline{Y})'(A_1+A_2)^{-1}(\overline{X}-\overline{Y})$$

用观测数据代入计算可得:

$$T^2=5.3117, F=1.4982,$$

显著性概率值 $p=0.2693>0.05=\alpha$

故$\mathbf{H_0}$相容.

3-12 在地质勘探中，在A、B、C三个地区采集了一些岩石，测其部分化学成分见表3.6.假定这三个地区岩石的成分遵从$N_3(\mu^{(i)}, \Sigma_i)$ (i=1，2，3)(α=0.05).

(1) 检验H0：$\Sigma_1=\Sigma_2=\Sigma_3$；H1：$\Sigma_1, \Sigma_2, \Sigma_3$不全等;

(2) 检验H0：$\mu^{(1)}=\mu^{(2)}$, H1：$\mu^{(1)}\neq\mu^{(2)}$;

(3) 检验H0: $\mu^{(1)}=\mu^{(2)}=\mu^{(3)}$, H1:存在$i\neq j$, 使$\mu^{(i)}\neq\mu^{(j)}$;

(4) 检验三种化学成分相互独立.

解:(4) 设来自三个总体的样本为(p=3, k=3)

$$X_{(i)}^{(t)} \sim N_p(\mu^{(t)}, \Sigma), (t=1,\cdots,k; i=1,\cdots,n_t)$$

检验$H_0:\sigma_{12}=\sigma_{13}=\sigma_{23}=0, H_1:\sigma_{12},\sigma_{13},\sigma_{23}$不全相等.

检验$\mathbf{H_0}$的似然比统计量为 $\lambda = \max\limits_{\mu^{(i)},\sigma_{ii}} L(\mu^{(i)},\sigma_{ii}) \Big/ \max\limits_{\mu^{(i)},\Sigma} L(\mu^{(i)},\Sigma)$

$$D=\begin{pmatrix}\sigma_{11} & 0 & 0\\ 0 & \sigma_{22} & 0\\ 0 & 0 & \sigma_{33}\end{pmatrix}=\mathrm{diag}(\Sigma),$$

似然比统计量的分子为

$$L(\overline{X}^{(t)},\hat{D})=\max L(\mu^{(t)};D)$$

$$=(2\pi)^{-\frac{np}{2}}\left|\hat{D}\right|^{-\frac{n}{2}}\exp[-\frac{1}{2}\mathrm{tr}(\hat{D}^{-1}A)]$$

$$\hat{D}=\begin{pmatrix} a_{11}/n & 0 & 0 \\ 0 & a_{22}/n & 0 \\ 0 & 0 & a_{33}/n \end{pmatrix}=\frac{1}{n}\operatorname{diag}(A),$$

$$A=\sum_{t=1}^{k}A_t=\sum_{t=1}^{k}\sum_{i=1}^{n_t}(X_{(i)}^{(t)}-\overline{X}^{(t)})(X_{(i)}^{(t)}-\overline{X}^{(t)})'$$

称为合并组内离差阵.

$$\left|\hat{D}\right|=\left(\frac{1}{n}\right)^{p}\prod_{i=1}^{p}a_{ii},\quad \hat{D}^{-1}=n\begin{pmatrix} a_{11}^{-1} & 0 & 0 \\ 0 & a_{22}^{-1} & 0 \\ 0 & 0 & a_{33}^{-1} \end{pmatrix},$$

$$\mathrm{tr}(\hat{D}^{-1}A) = n\sum_{i=1}^{p} a_{ii}^{-1} \times a_{ii} = np$$

$$L(\overline{X}^{(t)}, \hat{D}) = (2\pi)^{-\frac{np}{2}} \left|\hat{D}\right|^{-\frac{n}{2}} \exp[-\frac{1}{2}\mathrm{tr}(\hat{D}^{-1}A)]$$

$$= (2\pi)^{-\frac{np}{2}} \left(\frac{1}{n}\right)^{-\frac{np}{2}} \left(\prod_{i=1}^{p} a_{ii}\right)^{-\frac{n}{2}} \exp\left(-\frac{np}{2}\right)$$

$$= \left(\frac{n}{2\pi \mathrm{e}}\right)^{\frac{np}{2}} \left(\prod_{i=1}^{p} a_{ii}\right)^{-\frac{n}{2}}$$

似然比统计量的分母为

$$L(\overline{X}^{(t)}, \frac{1}{n}A) = \max L(\mu^{(t)}; \Sigma)$$

$$= (2\pi)^{-\frac{np}{2}} \left| \frac{1}{n}A \right|^{-\frac{n}{2}} \exp\{-\frac{1}{2}\mathrm{tr}[(\frac{1}{n}A)^{-1}A]\}$$

$$= (2\pi)^{-\frac{np}{2}} \left(\frac{1}{n}\right)^{-\frac{np}{2}} |A|^{-\frac{n}{2}} \exp\left(-\frac{np}{2}\right) = \left(\frac{n}{2\pi\mathrm{e}}\right)^{\frac{np}{2}} |A|^{-\frac{n}{2}}$$

检验H_0的似然比统计量可化为:

$$\lambda=\frac{\max\limits_{\mu^{(i)},\sigma_{ii}} L(\mu^{(i)},\sigma_{ii})}{\max\limits_{\mu^{(i)},\Sigma} L(\mu^{(i)},\Sigma)}=\frac{\left(\dfrac{n}{2\pi e}\right)^{np/2}\left(\prod\limits_{i=1}^{p} a_{ii}\right)^{-n/2}}{\left(\dfrac{n}{2\pi e}\right)^{np/2}|A|^{-n/2}}$$

$$=\frac{\left(\prod\limits_{i=1}^{p} a_{ii}\right)^{-n/2}}{|A|^{-n/2}}=\left(\frac{|A|}{\prod\limits_{i=1}^{p} a_{ii}}\right)^{\frac{n}{2}}=V^{\frac{n}{2}}$$

Box证明了，在H_0成立下当$n\to\infty$时，

$$\xi=-b\ln V\sim\chi^2(f),$$

其中

$$b=13-\frac{3}{2}-\frac{3^3-3}{3(3^2-3)}=\frac{61}{6}=10.1667$$

$$f=\frac{1}{2}[3\times4-\sum_{i=1}^{3}1\times2]=3$$

$$V=0.7253,\ \xi=-b\ln V=3.2650,$$

因 $p=0.3525>0.05$.

故H_0相容，即随机向量的三个分量(三种化学成分)相互独立.

或者利用定理3.2.1,当n充分大时，

$$\xi=-2\ln\lambda \sim \chi^2(f),$$

其中 $f=p+p(p+1)/2-(p+p)=3,$

$$V=0.7253,\ \lambda=0.1240\ ,$$

$$\xi=-2\ln\lambda=-n\times\ln V=4.1750,$$

因 $p=0.2432>0.05.$

故H_0相容，即随机向量的三个分量(三种化学成分)相互独立.

3-13 对表3.3给出的三组观测数据分别检验是否来自4维正态分布.

(1) 对每个分量检验是否一维正态?

(2) 利用χ^2图检验法对三组观测数据分别检验是否来自4维正态分布.

应用多元统计分析

第四章部分习题解答

第四章 回归分析

4-1 设 $\begin{cases} y_1 = a + \varepsilon_1, \\ y_2 = 2a - b + \varepsilon_2, \\ y_3 = a + 2b + \varepsilon_3, \end{cases} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{bmatrix} \sim N_3(0, \sigma^2 I_3),$

(1) 试求参数***a,b***的最小二乘估计;

解**:**用矩阵表示以上模型**:**

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 2 & -1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix} \overset{\text{def}}{=} X\beta + \varepsilon$$

则

$$\hat{\beta} = \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = (X'X)^{-1} X'Y = \left[\begin{pmatrix} 1 & 2 & 1 \\ 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 2 & -1 \\ 1 & 2 \end{pmatrix} \right]^{-1} \begin{pmatrix} 1 & 2 & 1 \\ 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$

$$=\begin{pmatrix}6&0\\0&5\end{pmatrix}^{-1}\begin{pmatrix}y_1+2y_2+y_3\\-y_2+2y_3\end{pmatrix}=\begin{pmatrix}\dfrac{1}{6}(y_1+2y_2+y_3)\\\dfrac{1}{5}(-y_2+2y_3)\end{pmatrix}$$

(2) 试导出检验H_0：*a*=*b*的似然比统计量，并指出当假设成立时，这个统计量的分布是什么?

解:样本的似然函数为

$$L(a,b,\sigma^2)=\frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^3}\exp\left[-\frac{1}{2\sigma^2}[(y_1-a)^2+(y_2-2a+b)^2+(y_3-a-2b)^2]\right]$$

$$\le L(\hat{a},\hat{b},\sigma^2)=\frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^3}\exp\left[-\frac{1}{2\sigma^2}[(y_1-\hat{a})^2+(y_2-2\hat{a}+\hat{b})^2+(y_3-\hat{a}-2\hat{b})^2]\right]$$

令 $$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{3}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}[(y_1 - \hat{a})^2 + \cdots] = 0$$

可得 $$\hat{\sigma}^2 = \frac{1}{3}\left[(y_1 - \hat{a})^2 + (y_2 - 2\hat{a} + \hat{b})^2 + (y_3 - \hat{a} - 2\hat{b})^2\right]$$

似然比统计量的分母为

$$L(\hat{a}, \hat{b}, \hat{\sigma}^2) = (2\pi)^{-\frac{3}{2}} (\hat{\sigma}^2)^{-\frac{3}{2}} \exp[-\frac{3}{2}].$$

当**H**$_0$：$\boldsymbol{a=b=a_0}$成立时,样本的似然函数为

$$L(a_0, \sigma^2) = \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^3} \exp\left[-\frac{1}{2\sigma^2}[(y_1 - a_0)^2 + (y_2 - a_0)^2 + (y_3 - 3a_0)^2]\right]$$

令 $$\frac{\partial L(a_0,\sigma^2)}{\partial a_0}=L(a_0,\sigma^2)\left(-\frac{2}{2\sigma^2}[-(y_1-a_0)-(y_2-a_0)-3(y_3-3a_0)\right)=0$$

可得 $$\hat{a}_0=\frac{1}{11}(y_1+y_2+3y_3)$$

令 $$\frac{\partial \ln L(\hat{a}_0,\sigma^2)}{\partial \sigma^2}=-\frac{3}{2\sigma^2}+\frac{1}{2(\sigma^2)^2}[(y_1-\hat{a}_0)^2+\cdots]=0$$

可得 $$\hat{\sigma}^2=\frac{1}{3}\left[(y_1-\hat{a}_0)^2+(y_2-\hat{a}_0)^2+(y_3-3\hat{a}_0)^2\right]\overset{\text{drf}}{=}\hat{\sigma}_0^2$$

似然比统计量的分子为

$$L(\hat{a}_0,\hat{\sigma}_0{}^2)=(2\pi)^{-\frac{3}{2}}(\hat{\sigma}_0{}^2)^{-\frac{3}{2}}\exp[-\frac{3}{2}].$$

似然比统计量为

$$\lambda=\frac{L(\hat{a}_0,\hat{\sigma}_0^{\,2})}{L(\hat{a},\hat{b},\hat{\sigma}^2)}=(\hat{\sigma}_0^{\,2})^{-\frac{3}{2}}\Big/(\hat{\sigma}^2)^{-\frac{3}{2}}=\left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^{\,2}}\right)^{\frac{3}{2}}=V^{\frac{3}{2}}$$

以下来讨论与V等价的统计量分布:

$$\hat{\sigma}^2=\frac{1}{3}\left[(y_1-\hat{a})^2+(y_2-2\hat{a}+\hat{b})^2+(y_3-\hat{a}-2\hat{b})^2\right]$$

$$=\frac{1}{3}\left[(y_1-\hat{y}_1)^2+(y_2-\hat{y}_2)^2+(y_3-\hat{y}_3)^2\right]$$

$$=\frac{1}{3}(Y-X\hat{\beta})'(Y-X\hat{\beta})=\frac{1}{3}Y'(I_3-X(X'X)^{-1}X')Y$$

$$=\frac{1}{3}Y'AY\text{，且}\operatorname{rank}(A)=\operatorname{tr}(A)=3-2=1$$

因 $Y \sim N_3(X\beta, \sigma^2 I_3)$, A为对称幂等阵,

$$\frac{Y'AY}{\sigma^2} \sim \chi^2(1,\delta), \text{因}\ \delta = \frac{1}{\sigma^2}(X\beta)' AX\beta = 0$$

$$\therefore \quad \frac{Y'AY}{\sigma^2} \sim \chi^2(1)$$

当**H$_0$：*$a=b=a_0$*成立时,回归模型为**

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 3 \end{pmatrix} a_0 + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix} \overset{\text{def}}{=} Za_0 + \varepsilon, \text{且}\ Y \sim N_3(Za_0, \sigma^2 I_3)$$

$$\hat{\sigma}_0^2 = \frac{1}{3}\left[(y_1 - \hat{a}_0)^2 + (y_2 - \hat{a}_0)^2 + (y_3 - 3\hat{a}_0)^2\right]$$

$$=\frac{1}{3}(Y-Z\hat{a}_0)'(Y-Z\hat{a}_0)=\frac{1}{3}Y'(I_3-Z(Z'Z)^{-1}Z')Y$$

$$=\frac{1}{3}Y'BY$$

考虑 $\hat{\sigma}_0^2-\hat{\sigma}^2=\frac{1}{3}Y'(B-A)Y$

$$B-A=X(X'X)^{-1}X'-Z(Z'Z)^{-1}Z'$$

$$=\frac{1}{330}\begin{pmatrix} 25 & 80 & -35 \\ & 256 & -112 \\ & & 49 \end{pmatrix}$$

经验证:① ***B-A*是对称幂等阵;**

② **rank(*B-A*)=tr(*B-A*)=2-1=1;**

③ $A(B\text{-}A)=O_{3\times3}$.由第三章§3.1的结论6知

$Y'AY$与$Y'(B-A)Y$相互独立;也就是

$\hat{\sigma}_0^2-\hat{\sigma}^2$与$\hat{\sigma}^2$相互独立.

由第三章§3.1的结论4知(**H**$_0$：$a=b$成立时)

$$\frac{Y'(B-A)Y}{\sigma^2}\sim\chi^2(1,\delta),\text{因}\,\delta=\frac{1}{\sigma^2}(Za_0)'(B-A)Za_0=0$$

$$\therefore\quad\frac{3(\hat{\sigma}_0^2-\hat{\sigma}^2)}{\sigma^2}=\frac{Y'(B-A)Y}{\sigma^2}\sim\chi^2(1)$$

所以

$$\xi=\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2-\hat{\sigma}^2}=\frac{Y'AY}{Y'(B-A)Y}\sim F(1,1)$$

因$\lambda=V^{\frac{3}{2}}$，$V=\dfrac{\hat{\sigma}^2}{\hat{\sigma}_0^2}$，故$\xi=\dfrac{V}{1-V}$或$V=\dfrac{\xi}{1+\xi}$，

否定域为

$$\{\lambda\le\lambda_\alpha\}\Longleftrightarrow\{V\le V_\alpha\}\Longleftrightarrow\{\xi\ge f_\alpha\}$$

4-2 在多元线性回归模型**(4.1.3)**中**(*p*=1)**，试求出参数向量$\boldsymbol{\beta}$和$\boldsymbol{\sigma}^2$的最大似然估计.

解:模型**(4.1.3)**为
$$\begin{cases} Y = C\beta + \varepsilon \\ \varepsilon \sim N_n(0, \sigma^2 I_n) \end{cases}$$

样本的似然函数为

$$L(\beta,\sigma^2) = (2\pi)^{-\frac{n}{2}}(\sigma^2)^{-\frac{n}{2}}\exp\left(-\frac{1}{2\sigma^2}(Y-C\beta)'(Y-C\beta)\right)$$

$$\ln L(\beta,\sigma^2) = \ln(2\pi)^{-\frac{n}{2}} + \ln(\sigma^2)^{-\frac{n}{2}} - \frac{1}{2\sigma^2}(Y-C\beta)'(Y-C\beta)$$

$$= \ln(2\pi)^{-\frac{n}{2}} + \ln(\sigma^2)^{-\frac{n}{2}} - \frac{1}{2\sigma^2}(Y'Y - 2Y'C\beta - \beta'C'C\beta)$$

令
$$\begin{cases}\dfrac{\partial \ln L}{\partial \beta}=-\dfrac{1}{2\sigma^2}\left[-2(Y'C)'+2C'C\beta\right]=0\\ \dfrac{\partial \ln L}{\partial \sigma^2}=-\dfrac{n}{2\sigma^2}+\dfrac{1}{2(\sigma^2)^2}\left[(Y-C\beta)'(Y-C\beta)\right]=0\end{cases}$$

可得参数向量β和σ^2的最大似然估计为:

$$\begin{cases}\hat{\beta}=(C'C)^{-1}C'Y\\ \hat{\sigma}^2=\dfrac{1}{n}(Y-C\hat{\beta})'(Y-C\hat{\beta})\end{cases}.$$

4-6 称观测向量Y和估计向量$\hat{Y}$的相关系数R为全相关系数.即

$$R=\frac{\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2\times\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2}}\quad(\text{其中}\bar{\hat{y}}=\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i),$$

试证明:(1) $\bar{\hat{y}}=\bar{y}$;

(2) $R^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2\Big/\sum_{i=1}^{n}(y_i-\bar{y})^2$;

⑶ 残差平方和$Q(\hat{\beta})=(1-R^2)\sum_{i=1}^{n}(y_i-\bar{y})^2$.

证明:(1)估计向量为 $\hat{Y}=C\hat{\beta}=C(C'C)^{-1}C'Y=HY$

$$\bar{\hat{y}}=\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i=\frac{1}{n}1_n'\hat{Y}=\frac{1}{n}1_n'HY=\frac{1}{n}(H1_n)'Y$$

$$=\frac{1}{n}1_n'Y=\bar{y}.$$

(因$1_n\in C$张成的空间,这里有$H1_n=1_n$)

(2) 因 $\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})=\sum_{i=1}^{n}(y_i-\hat{y}_i+\hat{y}_i-\bar{y})(\hat{y}_i-\bar{y})$

$$=\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

上式第一项为:

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})=(Y-\hat{Y})'(\hat{Y}-\bar{y}1_n)$$

$$=(Y-C\hat{\beta})'(C\hat{\beta}-\bar{y}1_n)=Y'C\hat{\beta}-\hat{\beta}'C'C\hat{\beta}-\bar{y}(Y-\hat{Y})'1_n$$

$$=Y'C\hat{\beta}-('C'Y)C\hat{\beta}-0=0$$

$$R^2=\frac{\left[\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})\right]^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2\times\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2}=\frac{\left[\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2\right]^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2\times\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2},$$

所以

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{U}{l_{yy}}.$$

(3) 残差平方和Q为

$$Q(\hat{\beta}) = l_{yy} - U = l_{yy} - l_{yy}R^2$$

$$= (1 - R^2)l_{yy} = (1 - R^2)\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

4-7 在多对多的多元线性回归模型中，给定 $\boldsymbol{Y}_{n\times p}$, $\boldsymbol{X}_{n\times m}$, 且rank($\boldsymbol{X}$)=$\boldsymbol{m}$, C=($1_n|\boldsymbol{X}$). 则

$$\begin{aligned}Q(\beta) &= (Y - C\beta)'(Y - C\beta) \\ &= (Y - C\hat{\beta})'(Y - C\hat{\beta}) + (\hat{\beta} - \beta)'C'C(\hat{\beta} - \beta)\end{aligned}$$

其中$\hat{\boldsymbol{\beta}} = (\boldsymbol{C}'\boldsymbol{C})^{-1}\boldsymbol{C}'\boldsymbol{Y}$.

证明:
$$\begin{aligned}Q(\beta) &= (Y - C\beta)'(Y - C\beta) \\ &= (Y - C\hat{\beta} + C\hat{\beta} - C\beta)'(Y - C\hat{\beta} + C\hat{\beta} - C\beta) \\ &= (Y - C\hat{\beta})'(Y - C\hat{\beta}) + (\hat{\beta} - \beta)'C'C(\hat{\beta} - \beta)\end{aligned}$$

$C'(Y - C\hat{\beta}) = O$, 故交叉项=$\boldsymbol{O}$.

4-8 在多对多的回归模型中，令

$$Q(\beta)=(Y-C\beta)'(Y-C\beta).$$

试证明$\hat{\beta}\triangleq(C'C)^{-1}C'Y$是在下列四种意义下达最小：

(1) $\mathrm{tr}Q(\hat{\beta})\leqslant\mathrm{tr}Q(\beta)$；

(2) $Q(\hat{\beta})\leqslant Q(\beta)$；

(3) $|Q(\hat{\beta})|\leqslant|Q(\beta)|$；

(4) $\mathrm{ch}_1(Q(\hat{\beta}))\leqslant\mathrm{ch}_1(Q(\beta))$，其中$\mathrm{ch}_1(A)$表示$A$的最大特征值.

以上β是$(m+1)\times p$的任意矩阵.

第四章 回归分析

$$(1)\ 因\ \mathrm{tr}[Q(\beta)]=\mathrm{tr}[Q(\hat{\beta})]+\mathrm{tr}[(C\hat{\beta}-C\beta)'(C\hat{\beta}-C\beta)]$$

$$\geqslant \mathrm{tr}[Q(\hat{\beta})],$$

故 $\hat{\beta}$ 使 $Q(\beta)$ 在迹的意义下达最小；

或者：因 $\mathrm{tr}[Q(\beta)]=\mathrm{tr}[(Y-C\beta)'(Y-C\beta)]$

$$=\mathrm{tr}[E'E]=\sum_{i=1}^{n}\sum_{j=1}^{p}\varepsilon_{ij}^{2}$$

因 $\hat{\beta}=(C'C)^{-1}C'Y$ 是由拉直模型下得出的最小二乘估计量。即

$$\mathrm{tr}\,Q(\hat{\beta})=\sum_{i=1}^{n}\sum_{j=1}^{p}\hat{\varepsilon}_{ij}^{2}\leqslant\sum_{i=1}^{n}\sum_{j=1}^{p}\varepsilon_{ij}^{2}=\mathrm{tr}\,Q(\beta);$$

故 $\hat{\beta}$ 使 $Q(\beta)$ 在迹的意义下达最小；

(2) 因 $Q(\beta)=Q(\hat\beta)+(\hat\beta-\beta)'C'C(\hat\beta-\beta)$

$$\geqslant Q(\hat\beta),$$

故 $\hat\beta$ 使 $Q(\beta)$ 在非负定的意义下达最小；

以上不等式的等号仅当 $\beta=\hat\beta$ 时成立。

$$\text{等号成立} \Leftrightarrow C(\hat\beta-\beta)=0$$

$$\Leftrightarrow (C'C)^{-1}C'\bullet C(\hat\beta-\beta)=0$$

$$\Leftrightarrow \beta=\hat\beta.$$

(3) 设 $|Q(\hat{\beta})|\neq 0$，则 $Q^{-1}(\hat{\beta})$ 存在。因

$$
\begin{aligned}
|Q(\beta)| &= |Q(\hat{\beta})+(C\hat{\beta}-C\beta)'(C\hat{\beta}-C\beta)| \\
&= \begin{vmatrix} Q(\hat{\beta}) & -(C\hat{\beta}-C\beta)' \\ (C\hat{\beta}-C\beta) & I_n \end{vmatrix} \\
&= |Q(\hat{\beta})|\,|I_n+A|
\end{aligned}
$$

其中 $A=(C\hat{\beta}-C\beta)Q^{-1}(\hat{\beta})(C\hat{\beta}-C\beta)'$，显然 A 是 n 阶对称且非负定阵。

设 A 的特征值为 λ_i（$i=1,2,\cdots,n$；且 $\lambda_i \geqslant 0$），则 $I_n + A$ 的特征值为 $\lambda_i + 1 \geqslant 1$（$i=1,2,\cdots,n$），

故 $|I_n + A| = \prod_{i=1}^{n}(1+\lambda_i) \geqslant 1$，所以

$$|Q(\beta)| = |Q(\hat{\beta})||I_n + A| \geqslant |Q(\hat{\beta})|.$$

当 $|Q(\hat{\beta})| = 0$ 时，必有 $|Q(\hat{\beta})| \leqslant |Q(\beta)|$。

故 $\hat{\beta}$ 使 $Q(\beta)$ 在行列式的意义下达最小；

(4) 设 $Q(\beta)$ 的特征值为 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p$，对任给 $x \neq 0, (x \in R^p)$，

见附录P394定理7.2(7.5)式

$$ch_1(Q(\beta)) = \lambda_1 = \sup_{x \neq 0} \frac{x'Q(\beta)x}{x'x} = \sup_{\|x\|=1} x'Q(\beta)x$$

$$= \sup_{\|x\|=1} [x'Q(\hat{\beta})x + x'(C\hat{\beta} - C\beta)'(C\hat{\beta} - C\beta)x]$$

$$\geqslant \sup_{\|x\|=1} [x'Q(\hat{\beta})x] = ch_1[Q(\hat{\beta})]。$$

故 $\hat{\beta}$ 使 $Q(\beta)$ 在最大特征根的意义下达最小；

应用多元统计分析

第五章部分习题解答

第五章 判别分析

5-1 已知总体G_i (*m*=1)的分布为: $N(\mu^{(i)},\sigma_i^2)$ (*i*=1,2) ,按距离判别准则为(不妨设$\mu^{(1)}>\mu^{(2)}$,$\sigma_1<\sigma_2$)

$$\begin{cases} x\in G_1, 若 \quad \mu^* < x < \mu_*, \\ x\in G_2, 若 x\le \mu^* \quad x\ge \mu_*, \end{cases}$$

其中 $\mu^*=\dfrac{\sigma_1\mu^{(2)}+\sigma_2\mu^{(1)}}{\sigma_1+\sigma_2}$ 试求错判概率**P(2|1)**和**P(1|2).**

解:

$$P(2\,|\,1)=P\{X\le \mu^*\,|\,X\sim N(\mu^{(1)},\sigma_1^2)\}$$

$$+P\{X\ge \mu_*\,|\,X\sim N(\mu^{(1)},\sigma_1^2)\}$$

$$=P\left\{\frac{X-\mu^{(1)}}{\sigma_1}\le\frac{\mu^*-\mu^{(1)}}{\sigma_1}\right\}+P\left\{\frac{X-\mu^{(1)}}{\sigma_1}\ge\frac{\mu_*-\mu^{(1)}}{\sigma_1}\right\}$$

记

$$b=\frac{\mu^{*}-\mu^{(1)}}{\sigma_1}=\left[\frac{\sigma_1\mu^{(2)}+\sigma_2\mu^{(1)}}{\sigma_1+\sigma_2}-\mu^{(1)}\right]\Big/\sigma_1=\frac{\mu^{(2)}-\mu^{(1)}}{\sigma_1+\sigma_2},$$

$$a=-\frac{\mu_{*}-\mu^{(1)}}{\sigma_1}=-\left[\frac{\sigma_2\mu^{(1)}-\sigma_1\mu^{(2)}}{\sigma_2-\sigma_1}-\mu^{(1)}\right]\Big/\sigma_1=\frac{\mu^{(2)}-\mu^{(1)}}{\sigma_2-\sigma_1},$$

$$\therefore P(2\,|\,1)=P\{U\le b\}+P\{U\ge -a\}\qquad (U\sim N(0,1))$$

$$=\Phi(b)+\Phi(a)$$

$$P(1|2) = P\{\mu^* < X < \mu_* \mid X \sim N(\mu^{(2)}, \sigma_2^2)\}$$

$$= P\left\{\frac{X - \mu^{(2)}}{\sigma_2} < \frac{\mu_* - \mu^{(2)}}{\sigma_2}\right\} - P\left\{\frac{X - \mu^{(2)}}{\sigma_2} \le \frac{\mu^* - \mu^{(2)}}{\sigma_2}\right\}$$

$$= P\{U < -a\} - P\{U \le -b\}$$

$$= \Phi\left(\frac{\mu^{(1)} - \mu^{(2)}}{\sigma_2 - \sigma_1}\right) - \Phi\left(\frac{\mu^{(1)} - \mu^{(2)}}{\sigma_1 + \sigma_2}\right) \quad . \quad .$$

$$= \Phi(b) - \Phi(a)$$

5-2 设三个总体的分布分别为: G1为$N(2,0.5^2)$, G2为$N(0,2^2)$,G3为$N(3,1^2)$.试问样品x=2.5应判归哪一类?

(1) 按距离准则;

(2) 按Bayes准则 $\left(q_1 = q_2 = q_3 = \frac{1}{3}, L(j\,|\,i) = \begin{cases} 1, i \neq j \\ 0, i = j \end{cases}\right)$

解:(1)按距离准则,当样品x=2.5时,

$$d_1^2(x) = \frac{(2.5-2)^2}{0.5^2} = 1, d_2^2(x) = \frac{(2.5-0)^2}{2^2} = 1.5625,$$

$$d_3^2(x) = \frac{(2.5-3)^2}{1^2} = 0.25,$$

因0.25<1<1.5625,所以样品x=2.5判归G_3.

(2)按Bayes准则

解一:广义平方距离判别法

样品X到G_t的广义平方距离的计算公式为

$$D_t^2(X)=d_t^2(X)+g_1(t)+g_2(t),(t=1,2,3).$$

其中$g_1(t)=\ln|\sigma_t^2|, g_2(t)=0.$ 当样品x=2.5时,

$$D_1^2(x)=1+\ln(0.5)^2=-0.3863,$$

$$D_2^2(x)=1.5625+\ln 2^2=2.9488,$$

$$D_3^2(x)=0.25+\ln 1=0.25,$$

因样品到G_1的广义平方距离最小,所以将样品x=2.5 判归G_1.

解二:利用定理5.2.1的推论,计算 $q_t f_t(x),(t=1,2,3)$

当样品x=2.5时,

$$f_1(x)=\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left[-\frac{1}{2\sigma_1^2}(x-\mu^{(1)})^2\right]=\frac{1}{1.2533}\exp\left[-\frac{1}{2}\times 1\right]$$

$$=\frac{1}{1.2533}\times 0.6065=0.4839$$

所以 $q_1 f_1(x)=0.1613$,类似可得

$$q_2 f_2(x)=0.0304, q_3 f_3(x)=0.1174,$$

因**0.1613>0.1174>0.0304,**所以样品x=**2.5**判归**G$_1$.**

解三:后验概率判别法,

计算样品x已知,属G_t的后验概率:

$$P(t\mid x)=\frac{q_t f_t(x)}{\sum_{i=1}^{3} q_i f_i(x)} \quad (t=1,2,3)$$

当样品x=2.5时,经计算可得

$$P(1\mid x=2.5)=\frac{0.1613}{0.1613+0.0304+0.1174}=\frac{0.1613}{0.3091}=0.5218,$$

$$P(2\mid x=2.5)=\frac{0.0304}{0.3091}=0.0984,\ P(3\mid x=2.5)=\frac{0.1174}{0.3091}=0.3798,$$

因**0.5218>0.3798>0.0984**,所以样品x=**2.5**判归G_1.

5−3 设总体G_i的均值为$\mu^{(i)}(i=1,2)$,同协差阵Σ.

记$\bar{\mu}=\frac{1}{2}(a'\mu^{(1)}+a'\mu^{(2)})$,(其中$a=\Sigma^{-1}(\mu^{(1)}-\mu^{(2)})$),

试证明(1)$\mathrm{E}(a'X\mid G_1)>\bar{\mu}$;(2)$\mathrm{E}(a'X\mid G_2)<\bar{\mu}$.

解:$\mathrm{E}(a'X\mid G_1)-\bar{\mu}=a'\mu^{(1)}-\frac{1}{2}(a'\mu^{(1)}+a'\mu^{(2)})=\frac{1}{2}(a'\mu^{(1)}-a'\mu^{(2)})$

$$=\frac{1}{2}(\mu^{(1)}-\mu^{(2)})'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)})>0,(因\Sigma>0)$$

类似可证: $\mathrm{E}(a'X\mid G_2)-\bar{\mu}=-\frac{1}{2}(\mu^{(1)}-\mu^{(2)})'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)})<0$,.

即 $\mathrm{E}(a'X\mid G_1)>\bar{\mu}$, $\mathrm{E}(a'X\mid G_2)<\bar{\mu}$.

由此题的结论可得出判别法:

$$\begin{cases} a'X > \bar{\mu} & 判 X \in G_1, \\ a'X < \bar{\mu} & 判 X \in G_2. \end{cases}$$

$$\Leftrightarrow \begin{cases} W(X) > 0 & 判 X \in G_1, \\ W(X) < 0 & 判 X \in G_2, \end{cases}$$

其中$W(X) = a'(X - \mu^*)$

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

$$\mu^* = \frac{1}{2}(\mu^{(1)} + \mu^{(2)}).$$

5-4 设有两个正态总体G_1和G_2,已知$(m=2)$

$$\mu^{(1)}=\begin{pmatrix}10\\15\end{pmatrix},\mu^{(2)}=\begin{pmatrix}20\\25\end{pmatrix},\Sigma_1=\begin{pmatrix}18&12\\12&32\end{pmatrix},\Sigma_2=\begin{pmatrix}20&-7\\-7&5\end{pmatrix}.$$

先验概率$q_1=q_2$,而$L(2|1)=10, L(1|2)=75$.试问样品

$X_{(1)}=\begin{pmatrix}20\\20\end{pmatrix}$及$X_{(2)}=\begin{pmatrix}15\\20\end{pmatrix}$各应判归哪一类?

(1) 按Fisher准则

解:取$A=\Sigma_1+\Sigma_2=\begin{pmatrix}18&12\\12&32\end{pmatrix}+\begin{pmatrix}20&-7\\-7&5\end{pmatrix}=\begin{pmatrix}38&5\\5&37\end{pmatrix}$(组内)

$$B=\sum_{i=1}^{2}(\mu^{(i)}-\bar{\mu})(\mu^{(i)}-\bar{\mu})'=\frac{1}{2}(\mu^{(1)}-\mu^{(2)})(\mu^{(1)}-\mu^{(2)})'$$

或取$B=(\mu^{(1)}-\mu^{(2)})(\mu^{(1)}-\mu^{(2)})'$

$$=\begin{pmatrix}10-20\\15-25\end{pmatrix}(-10,\ -10)=\begin{pmatrix}100 & 100\\100 & 100\end{pmatrix}\text{(组间)}$$

类似于例**5.3.1**的解法, $\boldsymbol{A^{-1}B}$的特征根就等于

$$d^2=(\mu^{(1)}-\mu^{(2)})'A^{-1}(\mu^{(1)}-\mu^{(2)})$$

$$=(-10,-10)\begin{pmatrix}37 & -5\\-5 & 38\end{pmatrix}\begin{pmatrix}-10\\-10\end{pmatrix}\frac{1}{1381}=\frac{6500}{1381}=4.7067$$

取$a=\frac{1}{d}A^{-1}(\mu^{(1)}-\mu^{(2)})=\frac{-1}{\sqrt{65\times1381}}\begin{pmatrix}32\\33\end{pmatrix}$, 则$a'Aa=1$,

且a满足: $Ba=\lambda Aa \quad (\lambda=d^2)$.

判别效率$\Delta(a)=\dfrac{a'Ba}{a'Aa}=\lambda=4.7067.$

Fisher线性判别函数为$u(X)=a'X=\dfrac{-1}{\sqrt{89765}}(32X_1+33X_2)$

判别准则为 $\begin{cases}判X\in G_1 & 当u(X)>u^* \\ 判X\in G_2 & 当u(X)\le u^*\end{cases}$,

阈值为$u^*=\dfrac{\sigma_2\overline{u}^{(1)}+\sigma_1\overline{u}^{(2)}}{\sigma_1+\sigma_2}=-4.2964.$ 其中

$$\sigma_1^2=a'\Sigma_1 a=\frac{1}{89765}(32,33)\begin{pmatrix}18 & 12\\ 12 & 32\end{pmatrix}\begin{pmatrix}32\\ 33\end{pmatrix}=\frac{78624}{89765}=0.8759$$

$$\sigma_2^2=a'\Sigma_2 a=\frac{1}{89765}(32,33)\begin{pmatrix}20 & -7\\ -7 & 5\end{pmatrix}\begin{pmatrix}32\\ 33\end{pmatrix}=\frac{11141}{89765}=0.1241$$

$$\bar{u}^{(1)}=a'\mu^{(1)}=\frac{-1}{\sqrt{89765}}(32,33)\begin{pmatrix}10\\15\end{pmatrix}=\frac{-815}{\sqrt{89765}}=-2.7202$$

$$\bar{u}^{(2)}=a'\mu^{(2)}=\frac{-1}{\sqrt{89765}}(32,33)\begin{pmatrix}20\\25\end{pmatrix}=\frac{-1465}{\sqrt{89765}}=-4.8897$$

$$\bar{u}^{(1)}>\bar{u}^{(2)}$$

$$X_{(1)}=\begin{pmatrix}20\\20\end{pmatrix}\quad ,u(X_{(1)})=\frac{-1}{\sqrt{89765}}(32,33)\begin{pmatrix}20\\20\end{pmatrix}=-4.3390$$

$$u(X_{(1)})=-4.3390<u^{*},\quad \therefore \quad X_{(1)}\in G_2.$$

$$X_{(1)}=\begin{pmatrix}15\\20\end{pmatrix}\quad ,u(X_{(2)})=\frac{-1}{\sqrt{89765}}(32,33)\begin{pmatrix}15\\20\end{pmatrix}=-3.8050$$

$$u(X_{(2)})=-3.8050>u^{*}\quad \therefore \quad X_{(2)}\in G_1.$$

(2)Bayes准则(假设$\Sigma_1=\Sigma_2=\begin{pmatrix}18 & 12\\ 12 & 32\end{pmatrix}=\Sigma$)

解:由定理5.2.1,只须计算

$h_1(X)=q_2L(1|2)f_2(X),h_2(X)=q_1L(2|1)f_1(X),$

并比较大小,判X属损失最小者.考虑

$$\frac{h_1(X)}{h_2(X)}=\frac{L(1|2)f_2(X)}{L(2|1)f_1(X)}=\frac{75}{10}\bullet\frac{f_2(X)}{f_1(X)}$$

$$=7.5\exp\{-\frac{1}{2}(X-\mu^{(2)})'\Sigma^{-1}(X-\mu^{(2)})+$$

$$\frac{1}{2}(X-\mu^{(1)})'\Sigma^{-1}(X-\mu^{(1)})\}$$

$$= 7.5\exp\{-(X-\bar{\mu})'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)})(\bar{\mu}=\begin{pmatrix}15\\20\end{pmatrix})$$

$$= 7.5\exp\{\frac{10}{216}(X-\bar{\mu})'\begin{pmatrix}10\\3\end{pmatrix}\}.$$

当$X_{(1)}=\begin{pmatrix}20\\20\end{pmatrix}$时, $\frac{h_1(X_{(1)})}{h_2(X_{(1)})}=7.5\exp\{\frac{125}{54}\}=75.9229>1$

因$h_1(X)>h_2(X)$, 故判$X_{(1)}\in G_2$.

当$X_{(2)}=\begin{pmatrix}15\\20\end{pmatrix}$时, $\frac{h_1(X_{(2)})}{h_2(X_{(2)})}=7.5\exp\{0\}=7.5>1$

因$h_1(X)>h_2(X)$, 故判$X_{(2)}\in G_2$.

5–5 已知$X_{(i)}^{(t)}(t=1,2;i=1,2\ldots,n_i)$为来自$G_t$的样本.

记$d=\overline{X}^{(1)}-\overline{X}^{(2)}$,(其中$\overline{X}^{(t)}=\frac{1}{n_t}\sum_{i=1}^{n_t}X_{(i)}^{(t)}(t=1,2)$)

$$S=\frac{1}{n_1+n_2-2}(A_1+A_2).$$

试证明:$a=S^{-1}(\overline{X}^{(1)}-\overline{X}^{(2)})$使比值$\frac{(a'd)^2}{a'Sa}$达最大值,

且最大值为马氏距离D^2

(其中$D^2=(\overline{X}^{(1)}-\overline{X}^{(2)})'S^{-1}(\overline{X}^{(1)}-\overline{X}^{(2)})$).

$$解:\Delta(a)=\frac{(a'd)^2}{a'Sa}=\frac{(a'd)(a'd)'}{a'Sa}$$

$$=\frac{a'(\overline{X}^{(1)}-\overline{X}^{(2)})(\overline{X}^{(1)}-\overline{X}^{(2)})'a}{a'Sa}\overset{\text{def}}{=}\frac{a'Ba}{a'Sa}\le\lambda_1$$

其中λ_1为$S^{-1}B$的最大特征值,且仅当$a=\lambda_1$对应的特征向量时等号成立.

$$又S^{-1}B=(\overline{X}^{(1)}-\overline{X}^{(2)})(\overline{X}^{(1)}-\overline{X}^{(2)})'S^{-1}与$$

$$D^2=(\overline{X}^{(1)}-\overline{X}^{(2)})'S^{-1}(\overline{X}^{(1)}-\overline{X}^{(2)})$$

有相同的特征值.故$\lambda_1=D^2$;

以下来验证a就是D^2对应的一个特征向量：

$$S^{-1}Ba = S^{-1}(\overline{X}^{(1)} - \overline{X}^{(2)})(\overline{X}^{(1)} - \overline{X}^{(2)})'S^{-1}(\overline{X}^{(1)} - \overline{X}^{(2)})$$

$$= S^{-1}(\overline{X}^{(1)} - \overline{X}^{(2)}) \bullet D^2$$

$$= D^2 a.$$

故当取$a = S^{-1}(\overline{X}^{(1)} - \overline{X}^{(2)})$时，比值$\Delta(a) = D^2$达最大值.

5－6 设两个p维正态总体$N_p(\mu^{(i)},\Sigma)(i=1,2)$.设$\mu^{(1)},\mu^{(2)},\Sigma$已知.线性判别函数

$$W(X)=(X-\overline{\mu})\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}),\overline{\mu}=\frac{1}{2}(\mu^{(1)}+\mu^{(2)}),$$

判别准则为
$$\begin{cases}\text{判}X\in G_1, & \text{当}W(X)>0,\\ \text{判}X\in G_2, & \text{当}W(X)\le 0,\end{cases}$$

试求错判概率$P(2|1)$和$P(1|2)$.

解:记$a=\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}),W(X)=(X-\overline{\mu})'a$是$X$的线性函数,当$X\in G_1$时,$W(X)\sim N_1(\nu_1,\sigma_1^2)$,且

$$\nu_1 = E(W(X)) = (\mu^{(1)} - \overline{\mu})'a = \frac{1}{2}(\mu^{(1)} - \mu^{(2)})'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)})$$

$$= \frac{1}{2}d^2 \quad [\text{其中} d^2 = (\mu^{(1)} - \mu^{(2)})'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)})]$$

$$\sigma_1^2 = D(W(X)) = D[a'(X - \overline{\mu})] = a'D(X - \overline{\mu})a = a'\Sigma a$$

$$= (\mu^{(1)} - \mu^{(2)})'\Sigma^{-1} \bullet \Sigma \bullet \Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) = d^2$$

$$\therefore P(2|1) = P\{W(X) \le 0 \,|\, X \in G_1\} = P\{\frac{W(X) - \nu_1}{\sigma_1} \le \frac{0 - \nu_1}{\sigma_1}\}$$

$$= P\{U \le -\frac{1}{2}d^2 / d\} = \Phi(-\frac{1}{2}d) = 1 - \Phi(\frac{1}{2}d).$$

其中 $\quad U = \dfrac{W(X) - \nu_1}{\sigma_1} \sim N(0,1).$

$$当X \in G_2时, W(X) \sim N_1(\nu_2, \sigma_2^2), 且$$

$$\nu_2 = (\mu^{(2)} - \overline{\mu})'a = -\frac{1}{2}d^2, \sigma_2^2 = d^2$$

$$\therefore P(1|2) = P\{W(X) > 0 \mid X \in G_2\} = P\{\frac{W(X) - \nu_2}{\sigma_2} > \frac{0 - \nu_2}{\sigma_2}\}$$

$$= P\{U > \frac{1}{2}d^2 / d\} = 1 - \Phi(\frac{1}{2}d).$$

其中 $U = \dfrac{W(X) - \nu_2}{\sigma_2} \sim N(0,1).$

应用多元统计分析

第六章部分习题解答

6-1 证明下列结论：

(1) 两个距离的和所组成的函数仍是距离；

(2) 一个正常数乘上一个距离所组成的函数仍是距离；

(3) 设d为一个距离，$c>0$为常数，则 $d^* = \dfrac{d}{d+c}$ 仍是一个距离；

(4) 两个距离的乘积所组成的函数不一定是距离；

证明：(1)设$d^{(1)}$和$d^{(2)}$为距离，令$d = d^{(1)} + d^{(2)}$.

以下来验证d满足作为距离所要求的3个条件.

① $d_{ij}=d_{ij}^{(1)}+d_{ij}^{(2)}\geq 0$,且仅当$X_{(i)}=X_{(j)}$时$d_{ij}=0$;

② $d_{ij}=d_{ij}^{(1)}+d_{ij}^{(2)}=d_{ji}^{(1)}+d_{ji}^{(2)}=d_{ji}$,对一切$i,j$;

③ $d_{ij}=d_{ij}^{(1)}+d_{ij}^{(2)}\leq d_{ik}^{(1)}+d_{kj}^{(1)}+d_{ik}^{(2)}+d_{kj}^{(2)}$

$=d_{ik}+d_{kj}$,对一切i,k,j.

(2) 设d是距离, $a>0$为正常数. 令$d^*=ad$,显然有

① $d_{ij}^{*}=cd_{ij}\geq 0$,且仅当$X_{(i)}=X_{(j)}$时$d_{ij}^{*}=0$;

② $d_{ij}^{*}=cd_{ij}=cd_{ji}=d_{ji}^{*}$,对一切$i,j$;

③ $$d_{ij}^* = cd_{ij} \le c(d_{ik} + d_{kj}) = cd_{ik} + cd_{kj}$$
$$= d_{ik}^* + d_{kj}^*, \text{对一切} i, k, j.$$

故*d*=ad*是一个距离.

(3) 设*d*为一个距离, *c*>0为常数, 显然有

① $$d_{ij}^* = \frac{d_{ij}}{d_{ij} + c} \ge 0, \text{且仅当} X_{(i)} = X_{(j)} \text{时} d_{ij}^* = 0;$$

② $$d_{ij}^* = \frac{d_{ij}}{d_{ij} + c} = \frac{d_{ji}}{d_{ji} + c} = d_{ji}^*, \text{对一切} i, j;$$

③ $$d_{ij}^{*}=\frac{d_{ij}}{d_{ij}+c}=\frac{1}{1+c/d_{ij}}\le\frac{1}{1+c/(d_{ik}+d_{kj})}$$

$$=\frac{d_{ik}+d_{kj}}{d_{ik}+d_{kj}+c}=\frac{d_{ik}}{d_{ik}+d_{kj}+c}+\frac{d_{kj}}{d_{ik}+d_{kj}+c}$$

$$\le\frac{d_{ik}}{d_{ik}+c}+\frac{d_{kj}}{d_{kj}+c}\qquad(\text{因}d_{ik}\ge 0,d_{kj}\ge 0)$$

$$=d_{ik}^{*}+d_{kj}^{*}\qquad\text{对一切}i,k,j.$$

故***d****是一个距离.

(4) 设$d^{(1)}$和$d^{(2)}$是距离,令$d^{*}=d^{(1)}\bullet d^{(2)}$.

d^{*}虽满足前2个条件,但不一定满足三角不等式.

下面用反例来说明d^{*}不一定是距离.

设$d_{ij}^{(1)}=d_{ij}^{(2)}=\left\|X_{(i)}-X_{(j)}\right\|(m=1)$,则$d_{ij}^{*}=\left\|X_{(i)}-X_{(j)}\right\|^{2}$.

当$X_{(i)}=0, X_{(j)}=1, X_{(k)}=0.5$时,$d_{ij}^{*}=1, d_{ik}^{*}=\dfrac{1}{4}, d_{kj}^{*}=\dfrac{1}{4}$.

显然不满足$d_{ij}^{*}\leq d_{ik}^{*}+d_{kj}^{*}$.

第六章 聚类分析

6-2 试证明二值变量的相关系数为(6.2.2)式，夹角余弦为(6.2.3)式.

设变量X_i和X_j是二值变量，它们的n次观测值记为x_{ti}, x_{tj} (t=1,…,n). x_{ti}, x_{tj} 的值或为0，或为1.由二值变量的列联表（表6.5）可知：变量X_i取值1的观测次数为$a+b$,取值0的观测次数为$c+d$;变量X_i和X_j取值均为1的观测次数为a,取值均为0的观测次数为d 等等。利用两定量变量相关系数的公式：

$$r_{ij}=\frac{\sum_{t=1}^{n}(x_{ti}-\bar{x}_i)(x_{tj}-\bar{x}_j)}{\sqrt{\sum_{t=1}^{n}(x_{ti}-\bar{x}_i)^2}\sqrt{\sum_{t=1}^{n}(x_{tj}-\bar{x}_j)^2}}$$

$$\sum_{t=1}^{n}(x_{ti}-\bar{x}_i)(x_{tj}-\bar{x}_j)=\sum_{t=1}^{n}x_{ti}x_{tj}-n\bar{x}_i\bar{x}_j=a-n\frac{a+b}{n}\frac{a+c}{n}$$

$$=\frac{1}{n}[an-(a+b)(a+c)]=\frac{1}{n}[a(a+b+c+d)-(a+b)(a+c)]$$

$$=\frac{ad-bc}{n}$$

$$\sum_{t=1}^{n}(x_{ti}-\bar{x}_i)^2=\sum_{t=1}^{n}x_{ti}^2-n\bar{x}_i^2=a+b-n\left(\frac{a+b}{n}\right)^2$$

$$=\frac{(a+b)}{n}[n-(a+b)]=\frac{1}{n}(a+b)(c+d)$$

$$\sum_{t=1}^{n}(x_{tj}-\bar{x}_j)^2=\sum_{t=1}^{n}x_{tj}^2-n\bar{x}_j^2=a+c-n\left(\frac{a+c}{n}\right)^2$$

$$=\frac{(a+c)}{n}[n-(a+c)]=\frac{1}{n}(a+c)(b+d)$$

故二值变量的相关系数为：

$$C_{ij}(7)=\frac{\sum_{t=1}^{n}(x_{ti}-\bar{x}_i)(x_{tj}-\bar{x}_j)}{\sqrt{\sum_{t=1}^{n}(x_{ti}-\bar{x}_i)^2}\sqrt{\sum_{t=1}^{n}(x_{tj}-\bar{x}_j)^2}}=\frac{ad-bc}{\sqrt{(a+b)(c+d)}\sqrt{(a+c)(b+d)}}$$

(6.2.2)

利用两定量变量夹角余弦的公式：

$$\cos\alpha_{ij} = \frac{\sum_{t=1}^{n} x_{ti}x_{tj}}{\sqrt{\sum_{t=1}^{n} x_{ti}^2}\sqrt{\sum_{t=1}^{n} x_{tj}^2}}$$

其中

$$\sum_{t=1}^{n} x_{ti}x_{tj} = a,\ \sum_{t=1}^{n} x_{ti}^2 = a+b,\sum_{t=1}^{n} x_{tj}^2 = a+c$$

故有 $c_{ij}(9) = \cos\alpha_{ij} = \dfrac{a}{\sqrt{(a+b)(a+c)}}$ (6.2.3)

6-3 下面是5个样品两两间的距离阵

$$D^{(0)} = D^{(1)} = \begin{pmatrix} 0 & & & & \\ 4 & 0 & & & \\ 6 & 9 & 0 & & \\ 1 & 7 & 10 & 0 & \\ 6 & 3 & 5 & 8 & 0 \end{pmatrix}$$

试用最长距离法、类平均法作系统聚类，并画出谱系聚类图.

解:用最长距离法:

① 合并$\{X_{(1)},X_{(4)}\}=\mathbf{CL4}$,

并类距离 $D_1=1$.

$$D^{(2)} = \begin{array}{cc} \begin{pmatrix} 0 & & & \\ 9 & 0 & & \\ 3 & 5 & 0 & \\ 7 & 10 & 8 & 0 \end{pmatrix} & \begin{matrix} X_{(2)} \\ X_{(3)} \\ X_{(5)} \\ CL4 \end{matrix} \end{array}$$

② 合并$\{X_{(2)},X_{(5)}\}$=CL3,并类距离 $D_2=3$.

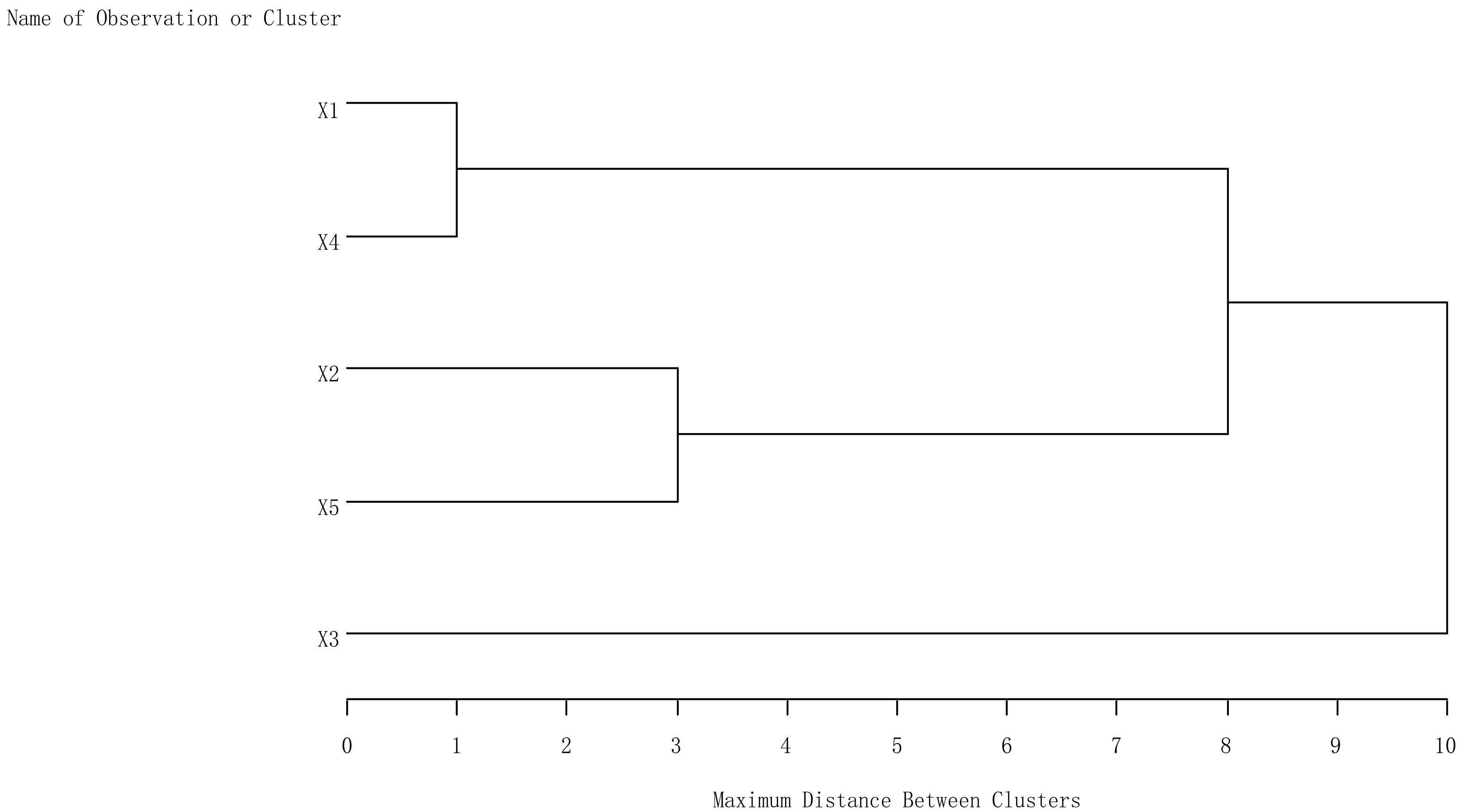
$$D^{(3)}=\begin{array}{c}\begin{pmatrix}0 & & \\ 10 & 0 & \\ 9 & 8 & 0\end{pmatrix}\end{array}\begin{array}{l}X_{(3)}\\ CL4\\ CL3\end{array}$$

③ 合并{CL3,CL4}=CL2,并类距离 $D_3=8$.

$$D^{(4)}=\begin{pmatrix}0 & \\ 10 & 0\end{pmatrix}\begin{array}{l}X_{(3)}\\ CL2\end{array}$$

④ 所有样品合并为一类CL1,并类距离 $D_4=10$.

最长距离法的谱系聚类图如下:



用类平均法:

$$D^{(0)} = D^{(1)} = \begin{pmatrix} 0 & & & & \\ 4 & 0 & & & \\ 6 & 9 & 0 & & \\ 1 & 7 & 10 & 0 & \\ 6 & 3 & 5 & 8 & 0 \end{pmatrix}$$

① 合并$\{X_{(1)},X_{(4)}\}$=CL4,并类距离 $D_1=1$.

$$D^{(2)} = \begin{array}{c} \begin{pmatrix} 0 & & & \\ 9^2 & 0 & & \\ 3^2 & 5^2 & 0 & \\ 65/2 & 136/2 & 100/2 & 0 \end{pmatrix} \end{array} \begin{array}{l} X_{(2)} \\ X_{(3)} \\ X_{(5)} \\ CL4 \end{array}$$

② 合并$\{X_{(2)},X_{(5)}\}$=CL3,并类距离 D_2=3.

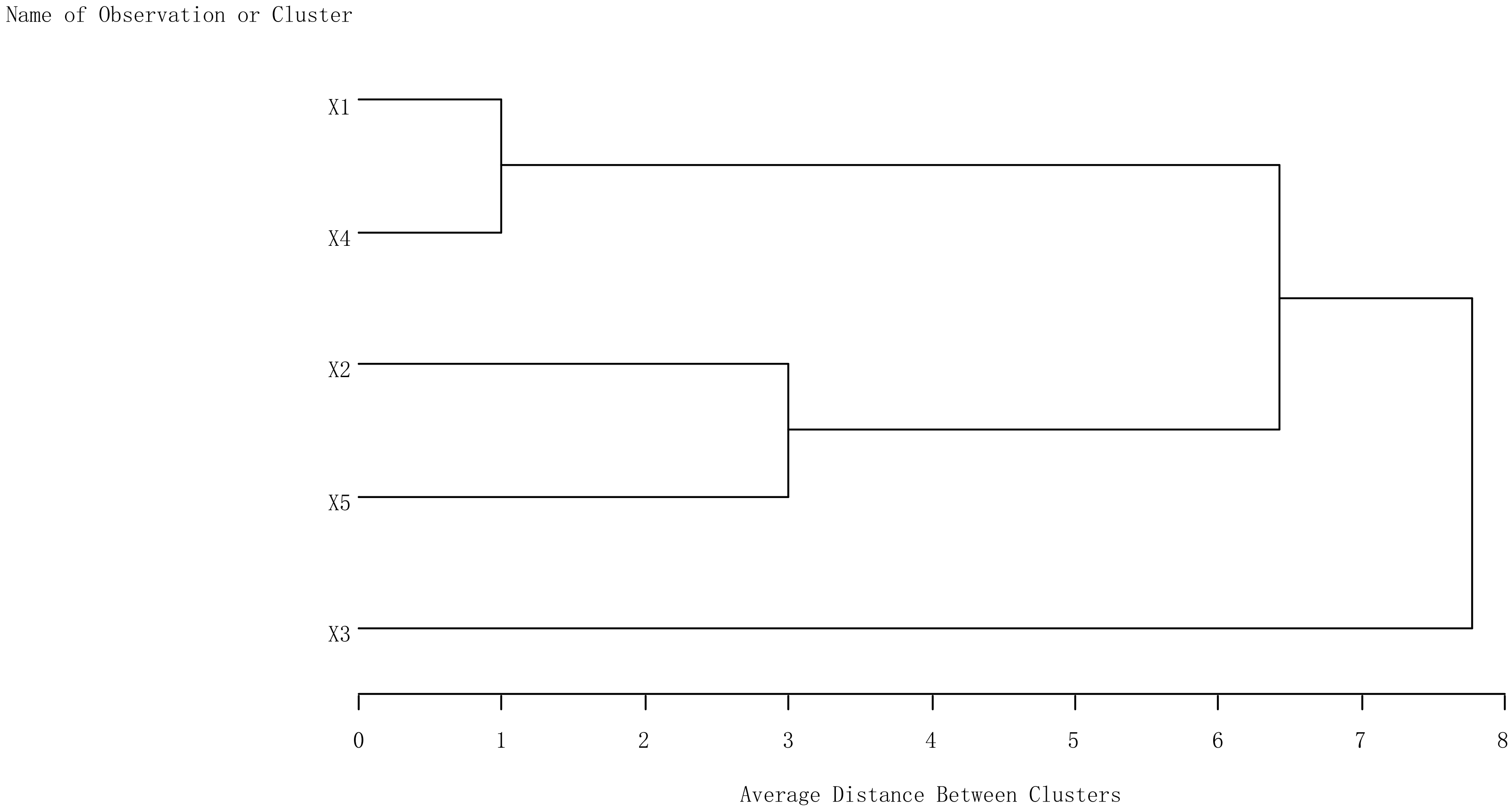
$$D^{(3)}=\begin{array}{c}\begin{pmatrix}0 & & \\ 136/2 & 0 & \\ 106/2 & 165/4 & 0\end{pmatrix}\end{array}\begin{array}{l}X_{(3)}\\ CL4\\ CL3\end{array}$$

③ 合并{CL3,CL4}=CL2,并类距离 $D_3=(165/4)^{1/2}$.

$$D^{(4)}=\begin{pmatrix}0 & \\ 121/2 & 0\end{pmatrix}\begin{array}{l}X_{(3)}\\ CL2\end{array}$$

④ 所有样品合并为一类CL1,并类距离 $D_4=(121/2)^{1/2}$.

类平均法的谱系聚类图如下:



6-4 利用距离平方的递推公式

$$D_{kr}^2 = \alpha_p D_{pk}^2 + \alpha_q D_{qk}^2 + \beta D_{pq}^2 + \gamma \,| D_{pk}^2 - D_{qk}^2 |$$

来证明当$\gamma=0,\alpha_p\geq 0,\alpha_q\geq 0,\alpha_p+\alpha_q+\beta\geq 1$时,系统聚类中的类平均法、可变类平均法、可变法、**Ward**法的单调性.

证明：设第L次合并G_p和G_q为新类G_r后,并类距离$D_L=D_{pq}$,且必有$D_{pq}^2\leq D_{ij}^2$.新类G_r与其它类G_k的距离平方的递推公式,当$\gamma=0,\alpha_p\geq 0,\alpha_q\geq 0$, $\alpha_p+\alpha_q+\beta\geq 1$时

$$D_{kr}^2 = \alpha_p D_{pk}^2 + \alpha_q D_{qk}^2 + \beta D_{pq}^2 \geq (\alpha_p + \alpha_q + \beta) D_{pq}^2 \geq D_{pq}^2$$

这表明新的距离矩阵中类间的距离均$\geq D_{pq}=D_L$,故有$D_{L+1}\geq D_L$，即相应的聚类法有单调性.

对于类平均法，因

$$\gamma=0,\alpha_p=\frac{n_p}{n_r}\geq 0,\alpha_q=\frac{n_q}{n_r}\geq 0,$$

$$\alpha_p+\alpha_q+\beta=\frac{n_p}{n_r}+\frac{n_q}{n_r}+0=1\geq 1$$

故类平均法具有单调性。

对于可变类平均法，因

$$\gamma=0,\alpha_p=(1-\beta)\frac{n_p}{n_r}\geq 0,\alpha_q=(1-\beta)\frac{n_q}{n_r}\geq 0,(\beta<1)$$

$$\alpha_p+\alpha_q+\beta=(1-\beta)\frac{n_p}{n_r}+(1-\beta)\frac{n_q}{n_r}+\beta=1\geq 1$$

故可变类平均法具有单调性。

对于可变法，因

$$\gamma = 0, \alpha_p = \frac{1-\beta}{2} \geq 0, \alpha_q = \frac{1-\beta}{2} \geq 0, (\beta < 1)$$

$$\alpha_p + \alpha_q + \beta = \frac{1-\beta}{2} + \frac{1-\beta}{2} + \beta = 1 \geq 1$$

故可变法具有单调性。

对于离差平方和法，因

$$\gamma = 0, \alpha_p = \frac{n_k + n_p}{n_r + n_k} \geq 0, \alpha_q = \frac{n_k + n_q}{n_r + n_k} \geq 0,$$

$$\alpha_p + \alpha_q + \beta = \frac{n_k + n_p}{n_r + n_k} + \frac{n_k + n_q}{n_r + n_k} - \frac{n_k}{n_r + n_k} = 1 \geq 1$$

故离差平方和法具有单调性。

6-5 试从定义直接证明最长和最短距离法的单调性.

证明：先考虑最短距离法：

设第L步从类间距离矩阵 $D^{(L-1)}=\left(D_{ij}^{(L-1)}\right)$出发，假设

$$D_{pq}^{(L-1)}=\min D_{ij}^{(L-1)}$$

故合并G_p和G_q为一新类G_r，这时第L步的并类距离：

$$D_L=D_{pq}^{(L-1)}$$

且新类G_r与其它类G_k的距离由递推公式可知

$$D_{rk}^{(L)}=\min(D_{pk}^{(L-1)},D_{qk}^{(L-1)})\ge D_{pq}^{(L-1)}=D_{(L)}\quad(k\neq p,q)$$

设第$L+1$步从类间距离矩阵 $D^{(L)}=\left(D_{ij}^{(L)}\right)$ 出发，

因 $D_{rk}^{(L)} \geq D_{pq}^{(L-1)} = D_L \quad (k \neq p,q)$

$$D_{ij}^{(L)} = D_{ij}^{(L-1)} \geq D_L \quad (i,j \neq r,p,q)$$

故第L＋1步的并类距离:

$$D_{L+1} = \min(D_{ij}^{(L)}) \geq D_L,$$

即最短距离法具有单调性.

类似地,可以证明最长距离法也具有单调性.

6-6 设A,B,C为平面上三个点,它们之间的距离为

$$d_{AB}^2 = d_{AC}^2 = 1.1,\quad d_{BC}^2 = 1.0$$

将三个点看成三个二维样品,试用此例说明中间距离法和重心法不具有单调性.

解:按中间距离法, 取 β =−1/4, 将B和C合并为一类后, 并类距离D_1=1, 而A与新类G_r={B, C} 的类间平方距离为

$$D_{Ar}^2 = \frac{1}{2}(D_{AB}^2 + D_{AC}^2) - \frac{1}{4}D_{BC}^2$$
$$= 0.5\times(1.1+1.1) - 0.25\times 1$$
$$= 1.1 - 0.25 = 0.85$$

当把**A**与{**B**，**C**}并为一类时，并类距离

$$D_2 = \sqrt{0.85} = 0.922 < 1 = D_1$$

故中间距离法不具有单调性。

按重心法, 将B和C合并为一类后, 并类距离 $\boldsymbol{D}_1$=1, 而**A**与新类$\boldsymbol{G}_r$={B, C}的类间平方距离为

$$D_{Ar}^2 = \frac{n_B}{n_r} D_{AB}^2 + \frac{n_C}{n_r} D_{AC}^2 - \frac{n_B}{n_r}\frac{n_C}{n_r} D_{BC}^2$$

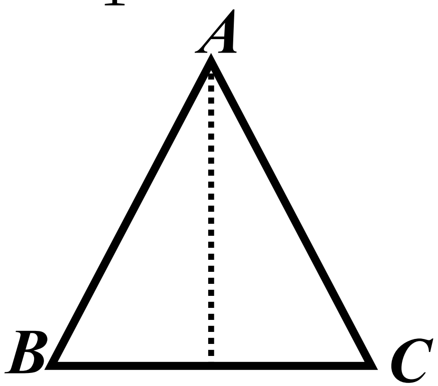
$$= 0.5 \times 1.1 + 0.5 \times 1.1 - 0.25 \times 1$$

$$= 1.1 - 0.25 = 0.85$$

当把A与{B，C}并为一类时，并类距离

$$D_2=\sqrt{0.85}=0.922<1=D_1$$

故重心法法不具有单调性。



并类过程如下：

$$D^{(1)}=\begin{array}{cc}\begin{pmatrix}0 & 1.1 & 1.1\\ & 0 & 1.0\\ & & 0\end{pmatrix} & \begin{array}{c}A\\B\\C\end{array}\end{array}\rightarrow D^{(2)}=\begin{array}{cc}\begin{pmatrix}0 & 0.85\\ & 0\end{pmatrix} & \begin{array}{c}A\\G_r\end{array}\end{array}$$

$$\rightarrow D^{(3)}=(0)$$

6-7 试推导重心法的距离递推公式(6.3.2);

$$D_{rk}^2=\frac{n_p}{n_r}D_{pk}^2+\frac{n_q}{n_r}D_{qk}^2-\frac{n_p n_q}{n_r^2}D_{pq}^2$$

解一：利用 $\overline{X}^{(r)}=\frac{1}{n_r}\left(n_p\overline{X}^{(p)}+n_q\overline{X}^{(q)}\right)$

如果样品间的距离定义为欧氏距离，则有

$$D_{rk}^2=(\overline{X}^{(k)}-\overline{X}^{(r)})'(\overline{X}^{(k)}-\overline{X}^{(r)})$$

$$=\left(\frac{n_p+n_q}{n_r}\overline{X}^{(k)}-\frac{n_p}{n_r}\overline{X}^{(p)}-\frac{n_q}{n_r}\overline{X}^{(q)}\right)'(\ldots)$$

$$
\begin{aligned}
D_{rk}^2 &= \left(\frac{n_p}{n_r}\right)^2 (\bar{X}^{(k)} - \bar{X}^{(p)})'(\ldots) + \left(\frac{n_q}{n_r}\right)^2 (\bar{X}^{(k)} - \bar{X}^{(q)})'(\ldots) \\
&\quad + \frac{n_p n_q}{n_r^2} (\bar{X}^{(k)} - \bar{X}^{(p)})' \underline{(\bar{X}^{(k)} - \bar{X}^{(q)})} \\
&\quad + \frac{n_p n_q}{n_r^2} (\bar{X}^{(k)} - \bar{X}^{(q)})' \underline{(\bar{X}^{(k)} - \bar{X}^{(p)})} \\
&= \frac{n_p^2}{n_r^2} D_{pk}^2 + \frac{n_q^2}{n_r^2} D_{qk}^2 + \frac{n_p n_q}{n_r^2} (\bar{X}^{(k)} - \bar{X}^{(p)})' \underline{(\bar{X}^{(k)} - \bar{X}^{(p)} + \bar{X}^{(p)} - \bar{X}^{(q)})} \\
&\quad + \frac{n_p n_q}{n_r^2} (\bar{X}^{(k)} - \bar{X}^{(q)})' \underline{(\bar{X}^{(k)} - \bar{X}^{(q)} + \bar{X}^{(q)} - \bar{X}^{(p)})}
\end{aligned}
$$

$$D_{rk}^2 = \frac{n_p^2}{n_r^2} D_{pk}^2 + \frac{n_q^2}{n_r^2} D_{qk}^2 + \frac{n_p n_q}{n_r^2} D_{pk}^2 + \frac{n_p n_q}{n_r^2} D_{qk}^2$$

$$+ \frac{n_p n_q}{n_r^2} (\overline{X}^{(k)} - \overline{X}^{(p)})' (\overline{X}^{(p)} - \overline{X}^{(q)})$$

$$- \frac{n_p n_q}{n_r^2} (\overline{X}^{(k)} - \overline{X}^{(q)})' (\overline{X}^{(p)} - \overline{X}^{(q)})$$

$$= \frac{n_p}{n_r} D_{pk}^2 + \frac{n_q}{n_r} D_{qk}^2 - \frac{n_p n_q}{n_r^2} D_{pq}^2$$

解二:因样品间的距离定义为欧氏距离,利用

$$\overline{X}^{(r)}=\frac{1}{n_r}\left(n_p\overline{X}^{(p)}+n_q\overline{X}^{(q)}\right)$$

$$D_{rk}^2=(\overline{X}^{(k)}-\overline{X}^{(r)})'(\overline{X}^{(k)}-\overline{X}^{(r)})$$

$$=\left(\overline{X}^{(k)}-\frac{1}{n_r}(n_p\overline{X}^{(p)}+n_q\overline{X}^{(q)})\right)'(\ldots)$$

$$=\overline{X}^{(k)\prime}\overline{X}^{(k)}-2\frac{n_p}{n_r}\overline{X}^{(k)\prime}\overline{X}^{(p)}-2\frac{n_q}{n_r}\overline{X}^{(k)\prime}\overline{X}^{(q)}$$

$$+\frac{1}{n_r^2}\left[n_p^2\overline{X}^{(p)\prime}\overline{X}^{(p)}+2n_pn_q\overline{X}^{(p)\prime}\overline{X}^{(q)}+n_q^2\overline{X}^{(q)\prime}\overline{X}^{(q)}\right]$$

利用
$$\overline{X}^{(k)'}\overline{X}^{(k)}=\frac{1}{n_r}\left(n_p\overline{X}^{(k)'}\overline{X}^{(k)}+n_q\overline{X}^{(k)'}\overline{X}^{(k)}\right)$$

$$\frac{n_q^2}{n_r^2}=\frac{1}{n_r^2}(n_qn_r-n_qn_p);\frac{n_p^2}{n_r^2}=\frac{1}{n_r^2}(n_pn_r-n_pn_q);$$

$$D_{rk}^2=\frac{n_p}{n_r}(\overline{X}^{(k)'}\overline{X}^{(k)}-2\overline{X}^{(k)'}\overline{X}^{(p)}+\overline{X}^{(p)'}\overline{X}^{(p)})$$

$$+\frac{n_q}{n_r}(\overline{X}^{(k)'}\overline{X}^{(k)}-2\overline{X}^{(k)'}\overline{X}^{(q)}+\overline{X}^{(q)'}\overline{X}^{(q)})$$

$$-\frac{n_pn_q}{n_r^2}(\overline{X}^{(p)'}\overline{X}^{(p)}-2\overline{X}^{(p)'}\overline{X}^{(q)}+\overline{X}^{(q)'}\overline{X}^{(q)})$$

故有
$$D_{rk}^2=\frac{n_p}{n_r}(\overline{X}^{(k)}-\overline{X}^{(p)})'(\overline{X}^{(k)}-\overline{X}^{(p)})$$
$$+\frac{n_q}{n_r}(\overline{X}^{(k)}-\overline{X}^{(q)})'(\overline{X}^{(k)}-\overline{X}^{(q)})$$
$$-\frac{n_p n_q}{n_r^2}(\overline{X}^{(p)}-\overline{X}^{(q)})'(\overline{X}^{(p)}-\overline{X}^{(q)})$$
$$=\frac{n_p}{n_r}D_{pk}^2+\frac{n_q}{n_r}D_{qk}^2-\frac{n_p n_q}{n_r^2}D_{pq}^2$$

6-8 试推导**Ward**法的距离递推公式(6.3.3);

解：Ward法把两类合并后增加的离差平方和看成类间的平方距离，即把类G_p和G_q的平方距离定义为 $D_{pq}^2=W_r-(W_p+W_q)$. 利用W_r的定义：

$$W_r=\sum_{t=1}^{n_r}(X_{(t)}^{(r)}-\overline{X}^{(r)})'(X_{(t)}^{(r)}-\overline{X}^{(r)})$$

$$=\sum_{t=1}^{n_p}(X_{(t)}^{(p)}-\overline{X}^{(r)})'(X_{(t)}^{(p)}-\overline{X}^{(r)})$$

$$+\sum_{t=1}^{n_q}(X_{(t)}^{(q)}-\overline{X}^{(r)})'(X_{(t)}^{(q)}-\overline{X}^{(r)})$$

$$W_r = \sum_{t=1}^{n_p} (X_{(t)}^{(p)} - \overline{X}^{(p)} + \overline{X}^{(p)} - \overline{X}^{(r)})'(\cdots)$$

$$+ \sum_{t=1}^{n_q} (X_{(t)}^{(q)} - \overline{X}^{(q)} + \overline{X}^{(q)} - \overline{X}^{(r)})'(\cdots)$$

$$= \sum_{t=1}^{n_p} (X_{(t)}^{(p)} - \overline{X}^{(p)})'(\cdots) + \sum_{t=1}^{n_p} (\overline{X}^{(p)} - \overline{X}^{(r)})'(\cdots) + 0 + 0$$

$$+ \sum_{t=1}^{n_q} (X_{(t)}^{(q)} - \overline{X}^{(q)})'(\cdots) + \sum_{t=1}^{n_q} (\overline{X}^{(q)} - \overline{X}^{(r)})'(\cdots) + 0 + 0$$

把$\overline{X}^{(r)} = \frac{1}{n_r}\left(n_p \overline{X}^{(p)} + n_q \overline{X}^{(q)}\right)$代入：$\overline{X}^{(p)} - \overline{X}^{(r)} = \frac{n_q}{n_r}(\overline{X}^{(p)} - \overline{X}^{(q)})$

$$\overline{X}^{(q)} - \overline{X}^{(r)} = \frac{n_p}{n_r}(\overline{X}^{(q)} - \overline{X}^{(p)})$$

$$W_r = W_p + W_q + \left(\frac{n_q}{n_r}\right)^2 \sum_{t=1}^{n_p} (\overline{X}^{(p)} - \overline{X}^{(q)})'(\cdots)$$

$$+ \left(\frac{n_p}{n_r}\right)^2 \sum_{t=1}^{n_q} (\overline{X}^{(q)} - \overline{X}^{(p)})'(\cdots)$$

$$= W_p + W_q + \left(\frac{n_q}{n_r}\right)^2 n_p (\overline{X}^{(p)} - \overline{X}^{(q)})'(\overline{X}^{(p)} - \overline{X}^{(q)})$$

$$+ \left(\frac{n_p}{n_r}\right)^2 n_q (\overline{X}^{(p)} - \overline{X}^{(q)})'(\overline{X}^{(p)} - \overline{X}^{(q)})$$

$$= W_p + W_q + \frac{n_p n_q}{n_r} (\overline{X}^{(p)} - \overline{X}^{(q)})'(\overline{X}^{(p)} - \overline{X}^{(q)})$$

$$D_{pq}^2 = W_r - (W_p + W_q) = \frac{n_p n_q}{n_p + n_q}(\bar{X}^{(p)} - \bar{X}^{(q)})'(\bar{X}^{(p)} - \bar{X}^{(q)})$$

$$= \frac{n_p n_q}{n_r} D_{pq}^2(\text{重})$$ （当样品间的距离定义为欧氏距离时）

记$\boldsymbol{G_r}$＝{G_p,G_q},则新类$\boldsymbol{G_r}$与其它类G_k的平方距离为

$$D_{rk}^2 = \frac{n_r n_k}{n_r + n_k}(\bar{X}^{(r)} - \bar{X}^{(k)})'(\bar{X}^{(r)} - \bar{X}^{(k)})$$

$$= \frac{n_r n_k}{n_r + n_k} D_{rk}^2(\text{重})$$

利用重心法的递推公式(6-7题已证明)可得:

$$D_{rk}^2=\frac{n_r n_k}{n_r+n_k}\left[\frac{n_p}{n_r}D_{pk}^2(\text{重})+\frac{n_q}{n_r}D_{qk}^2(\text{重})-\frac{n_p n_q}{n_r^2}D_{pq}^2(\text{重})\right]$$

$$=\frac{n_r n_k}{n_r+n_k}\left[\frac{n_p}{n_r}(\overline{X}^{(p)}-\overline{X}^{(k)})'(\cdots)+\frac{n_q}{n_r}(\overline{X}^{(q)}-\overline{X}^{(k)})'(\cdots)-\frac{n_p n_q}{n_r^2}(\overline{X}^{(p)}-\overline{X}^{(q)})'(\cdots)\right]$$

$$=\frac{n_k n_p}{n_r+n_k}(\overline{X}^{(p)}-\overline{X}^{(k)})'(\cdots)+\frac{n_k n_q}{n_r+n_k}(\overline{X}^{(q)}-\overline{X}^{(k)})'(\cdots)$$

$$-\frac{n_k}{n_r+n_k}\frac{n_p n_q}{n_r}(\overline{X}^{(p)}-\overline{X}^{(k)})'(\cdots)$$

$$=\frac{n_p+n_k}{n_r+n_k}D_{pk}^2+\frac{n_q+n_k}{n_r+n_k}D_{qk}^2-\frac{n_k}{n_r+n_k}D_{pq}^2$$

6-9 设有5个样品,对每个样品考察一个指标得数据为1，2,5,7,10.试用离差平方和法求5个样品分为k类(k＝5,4,3，2,1)的分类法b_k及相应的总离差平方和$W(k)$.

解：①计算样品间的欧氏平方距离阵

$$D^{(1)}=D^{(1)}=\frac{1}{2}\begin{pmatrix}0&&&&\\1&0&&&\\16&9&0&&\\36&25&4&0&\\81&64&25&9&0\end{pmatrix}=\begin{pmatrix}0&&&&\\0.5&0&&&\\8&4.5&0&&\\18&12.5&2&0&\\40.5&32&12.5&4.5&0\end{pmatrix}$$

② 合并 {1,2} ＝ CL4,并类距离D_1=(0.5)$^{1/2}$=0.707，并利用递推公式计算新类与其它类的平方距离得

$$D^{(2)}=\begin{array}{cc}\begin{pmatrix}0&&&\\49/6&0&&\\121/6&2&0&\\289/2&12.5&4.5&0\end{pmatrix}&\begin{array}{c}CL4\\5\\7\\10\end{array}\end{array}$$

③合并 {5,7} = **CL3**,并类距离$D_2=(2)^{1/2}=1.414$，并利用递推公式计算新类与其它类的平方距离得

$$D^{(3)}=\begin{array}{l}\begin{pmatrix}0 & & \\ 81/4 & 0 & \\ 32/3 & 289/2 & 0\end{pmatrix}\end{array}\begin{array}{l}CL3\\ CL4\\ 10\end{array}$$

④ 合并 {**CL3,10**}={5,7,10} = **CL2**,并类距离$D_3=(32/3)^{1/2}=3.266$，并利用递推公式计算新类与其它类的平方距离得

$$D^{(4)}=\begin{pmatrix}0 & \\ 245/6 & 0\end{pmatrix}\begin{array}{l}CL2\\ CL4\end{array}$$

⑤ 合并 **{CL4,CL2}={1,2,5,7,10}** ＝ **CL1**,并类距离D_4 =$(245/6)^{1/2}=6.39$，并利用递推公式计算新类与其它类的平方距离得

$$D^{(5)}=(0)CL1$$

⑥分类法b_k及相应的总离差平方和$W(k)$:

k=5	{1},{2},{5},{7},{10}	W(5)=0
k =4	{1,2}, {5},{7},{10}	W(4)=0.5
k =3	{1,2}, {5,7},{10}	W(3)=2.5
k =2	{1,2}, {5,7,10}	W(2)=13.666
k =1	{1,2,5,7,10}	W(1)=54

应用多元统计分析

第七章 主成分分析

7-1 设$\mathbf{X}=(X_1, X_2)'$的协方差阵 $\Sigma=\begin{pmatrix}1 & 4\\ 4 & 100\end{pmatrix}$, 试从Σ和相关阵R出发求出总体主成分，并加以比较.

解：从协差阵 $\Sigma=\begin{pmatrix}1 & 4\\ 4 & 100\end{pmatrix}$ 出发得总体主成分为

$$Z_1=0.040\ X_1+0.999\ X_2(\mathrm{Var}(Z_1)=\lambda_1=100.1614);$$

$$Z_2=0.999\ X_1-0.040\ X_2(\mathrm{Var}(Z_2)=\lambda_2=0.8386);$$

从相关阵 $R=\begin{pmatrix}1 & 0.4\\ 0.4 & 1\end{pmatrix}$ 出发得总体主成分为(带 * 者为标准化变量)

$$\begin{cases}Y_1=0.707\ X_1^*+0.707\ X_2^*(\mathrm{Var}(Z_1^*)=1.4);\\ Y_2=0.707\ X_1^*-0.707\ X_2^*(\mathrm{Var}(Z_2^*)=0.6);\end{cases}$$

或者(因 $\sigma_1^2=1,\sigma_2^2=100$)

$$\begin{cases}Y_1=0.707\ (X_1-\mu_1)+0.0707\ (X_2-\mu_2),\\ Y_2=0.707\ (X_1-\mu_1)-0.0707\ (X_2-\mu_2).\end{cases}$$

比较：

① 由 Σ 或 R 出发所得主成分不同；

② 由 Σ 出发时，第一主成分 Z_1 解释的总方差比例为

$$\frac{100.1614}{101}=0.9917 \text{（即 } 99.17\%\text{）；}$$

由 R 出发时，第一主成分 Y_1 解释的总方差比例为

$$\frac{1.4}{2}=0.7 \text{（}70\%\text{）；}$$

③ 由于 X_2 的方差大（$Var(X_2)=100$），故 Z_1 完全由 X_2 控制（系数为 0.999）。而原变量标准化后（$\rho=0.4$），结论相反，即 Z_1 主要由 X_1 控制（系数分别为 0.707 和 0.0707）。

④ 变量标准化后 $\rho(X_1^*, Y_1)=\sqrt{1.4}\times 0.707=0.8365$,

$$\rho(X_2^*, Y_1)=\sqrt{1.4}\times 0.707=0.8365,$$

即标准化后得第一主成分 Y_1 与 X_1^* 和 X_2^* 的相关系数相等.原始变量与第一主成分 Z_1 的相关系数不相等:

$$\rho(X_1, Z_1)=\sqrt{100.1614}\times 0.040/1=0.4003,$$

$$\rho(X_2, Z_1)=\sqrt{100.1614}\times 0.999/10=0.9998.$$

7-2 设$X=(X_1, X_2)'\sim N_2(0,\Sigma)$,协方差$\Sigma=\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}$

其中ρ为X_1和X_2的相关系数(ρ>0).

(1) 试从Σ出发求X的两个总体主成分;

(2) 求X的等概密度椭园的主轴方向;

(3) 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: (1)由 Σ(即 R)出发可得:

$$Y_1=\frac{\sqrt{2}}{2}X_1+\frac{\sqrt{2}}{2}X_2\quad (Var(Y_1)=\lambda_1=1+\rho);$$

$$Y_2=\frac{\sqrt{2}}{2}X_1-\frac{\sqrt{2}}{2}X_2\quad (Var(Y_2)=\lambda_2=1-\rho);$$

(2) 等概密度椭园为

$$(X-\mu)'\Sigma^{-1}(X-\mu)=C^2\quad (\mu=0,\Sigma=\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix})$$

由 Σ的特征值为$\lambda_1=1+\rho$, $\lambda_2=1-\rho$. 相应的特征向量为

$$a_1=(\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2})',\ a_2=(\frac{\sqrt{2}}{2},-\frac{\sqrt{2}}{2})',$$

则可得出 Σ^{-1}的特征值为$\frac{1}{\lambda_i}$ $(i=1,2)$,相应的特征向量仍为

a_1 和 a_2. 故 Σ^{-1}的谱分解式为

$$\Sigma^{-1} = \sum_{i=1}^{2} \frac{1}{\lambda_i} a_i a_i',$$

故

$$X'\Sigma^{-1}X = X' \cdot \sum_{i=1}^{2} \frac{1}{\lambda_i} a_i a_i' \cdot X = \sum_{i=1}^{2} \frac{1}{\lambda_i} [X' a_i]^2$$

$$= \frac{Y_1^2}{\lambda_1} + \frac{Y_2^2}{\lambda_2} = C^2 \quad (\text{这里 } Y_i = X' a_i)$$

$$\Longleftrightarrow \frac{Y_1^2}{\lambda_1 C^2} + \frac{Y_2^2}{\lambda_2 C^2} = 1 \quad (\text{等概密度椭园})$$

椭园长轴的方向为 $e_1 = (\sqrt{2}/2, \sqrt{2}/2)'$（即第一主成分的方向上），椭园短轴的方向为 $e_2 = (\sqrt{2}/2, -\sqrt{2}/2)'$（即第二主成分的方向上）.

(3) 当 $\frac{1+\rho}{2} \geqslant 0.95$，即 $\rho \geqslant 0.9$ 时第一主成分的贡献率达 95%以上.

7-3 设p维总体X的协差阵为

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix} \quad (0 < \rho \le 1).$$

(1) 试证明总体的第一主成分

$$Z_1 = \frac{1}{\sqrt{p}}(X_1 + X_2 + \cdots + X_p);$$

(2) 试求第一主成分的贡献率.

解：(1) 因 Σ 的最大特征值 $\lambda_1=\sigma^2[1+(p-1)\rho]$，而 $\lambda_2=\lambda_3=\cdots=\lambda_p=(1-\rho)\sigma^2$. 且最大特征值 λ_1 对应的单们特征向量为 $a_1=\frac{1}{\sqrt{p}}(1,1,\cdots,1)'$，故第一主成分为

$$Z_1=\frac{1}{\sqrt{p}}(X_1+X_2+\cdots+X_p).$$

(2) 因 $\mathrm{Var}(Z_1)=\lambda_1=\sigma^2[1+(p-1)\rho]$，故第一主成分的贡献率为

$$\frac{\lambda_1}{p\sigma^2}=\rho+\frac{1-\rho}{p}.$$

7-4　设总体$X=(X_1,\ldots,X_p)'\sim N_p(\mu,\Sigma)$ $(\Sigma>0)$,等概率密度椭球为　$(X-\mu)'\Sigma^{-1}(X-\mu)=C^2$($C$为常数).
试问椭球的主轴方向是什么?

解:　等概密度椭球为

$$(X-\mu)'\Sigma^{-1}(X-\mu)=C^2$$

设 Σ 的特征值为 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p>0$, 相应的单位正交特征向量为 $a_1, a_2, \cdots, a_p$. 则 Σ^{-1}的谱分解式为

$$\Sigma^{-1}=\sum_{i=1}^{p}\frac{1}{\lambda_i}a_ia_i',$$

$$\text{故}(X-\mu)'\cdot\sum_{i=1}^{p}\frac{1}{\lambda_i}a_ia_i'\cdot(X-\mu)=\sum_{i=1}^{p}\frac{1}{\lambda_i}[(X-\mu)'a_i]^2$$

$$=\frac{Z_1^2}{\lambda_1}+\frac{Z_2^2}{\lambda_2}+\cdots+\frac{Z_p^2}{\lambda_p}=C^2\quad(\text{这里 } Z_i=(X-\mu)'a_i)$$

$$\Longleftrightarrow\frac{Z_1^2}{\lambda_1C^2}+\frac{Z_2^2}{\lambda_2C^2}+\cdots+\frac{Z_p^2}{\lambda_pC^2}=1\quad(\text{等概密度椭球})$$

椭球的第 i 个主轴的方向在 X 的第 i 主成分的方向上, 其半长轴与$\sqrt{\lambda_i}$成比例, 且比例常数为 C.

7-5 设3维总体X的协差阵为

$$\Sigma = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

试求总体主成分.

解:总体主成分为

$$Z_i = X_i (i = 1,2,3)$$

主成分向量为

$$Z = (X_1, X_2, X_3)' \text{或} Z = (X_2, X_1, X_3)'$$

三个主成分的方差分别为**4,4,2.**

7-6 设3维总体X的协差阵为 $\Sigma=\begin{pmatrix}\sigma^2 & \rho\sigma^2 & 0\\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2\\ 0 & \rho\sigma^2 & \sigma^2\end{pmatrix}$

试求总体主成分，并计算每个主成分解释的方差比例

解：当 $0<\rho<\frac{1}{\sqrt{2}}$ 时，总体主成分为

$Z_1=\frac{X_1+\sqrt{2}X_2+X_3}{2}$，$\mathrm{Var}(Z_1)=\lambda_1=\sigma^2(1+\sqrt{2}\rho)$，解释的方差比例为 $\frac{1+\sqrt{2}\rho}{3}$；

$Z_2=\frac{X_1-X_3}{\sqrt{2}}$，$\mathrm{Var}(Z_2)=\lambda_2=\sigma^2$，解释的方差比例为 $\frac{1}{3}$；

$Z_3=\frac{X_1-\sqrt{2}X_2+X_3}{2}$，$\mathrm{Var}(Z_3)=\lambda_3=\sigma^2(1-\sqrt{2}\rho)$，解释的方差比例为 $\frac{1-\sqrt{2}\rho}{3}$.

7-7 设4维随机向量X的协差阵是

$$\Sigma=\begin{pmatrix}\sigma^2 & \sigma_{12} & \sigma_{13} & \sigma_{14}\\ \sigma_{12} & \sigma^2 & \sigma_{14} & \sigma_{13}\\ \sigma_{13} & \sigma_{14} & \sigma^2 & \sigma_{12}\\ \sigma_{14} & \sigma_{13} & \sigma_{12} & \sigma^2\end{pmatrix},$$

其中 $\sigma_{12}\ge\sigma_{13}\ge\sigma_{14},\sigma^2+\sigma_{14}\ge\sigma^2+\sigma_{13}.$

试求X的主成分.

解：　X的总体主成分为

$$Z_1 = \frac{1}{2}(X_1 + X_2 + X_3 + X_4),$$

$$Var(Z_1) = \lambda_1 = \sigma^2 + \sigma_{12} + \sigma_{13} + \sigma_{14},$$

$$Z_2 = \frac{1}{2}(X_1 + X_2 - X_3 - X_4),$$

$$Var(Z_2) = \lambda_2 = \sigma^2 + \sigma_{12} - \sigma_{13} - \sigma_{14},$$

$$Z_3 = \frac{1}{2}(X_1 - X_2 + X_3 - X_4),$$

$$Var(Z_3) = \lambda_3 = \sigma^2 - \sigma_{12} + \sigma_{13} - \sigma_{14},$$

$$Z_4 = \frac{1}{2}(X_1 - X_2 - X_3 + X_4),$$

$$Var(Z_4) = \lambda_4 = \sigma^2 - \sigma_{12} - \sigma_{13} + \sigma_{14}.$$

7-8 (1) 设数据阵 X 的第 j 列记为 x_j，则 $b_j=(b_{j1},b_{j2},\cdots,b_{jm})'$ 的最小二乘估计为 （记 $\Lambda_m=diag(\lambda_1,\lambda_2,\cdots,\lambda_m)$）

$$\hat{b}_j=(Z^{*\prime}Z^*)^{-1}Z^{*\prime}x_j \quad (\text{注意}: Z^*=XA^*)$$

$$=(A^{*\prime}X'XA^*)^{-1}A^{*\prime}X'\cdot Xe_j$$

$$=[(n-1)\Lambda_m]^{-1}[(n-1)A^*\Lambda_m]'e_j$$

$$=A^{*\prime}e_j=(a_{1j},a_{2j},\cdots,a_{mj})';$$

(2) X_j 对 $Z_1,\cdots,Z_m$ 的回归平方和为

$$U_j=(\hat{x}_j-\bar{x}_j)'(\hat{x}_j-\bar{x}_j) \quad (\bar{x}_j=0)$$

$$=\hat{x}_j'\hat{x}_j \quad (\text{注意}: \hat{x}_j=Z^*\hat{b}_j=Z^*A^{*\prime}e_j)$$

$$=e_j'A^*Z^{*\prime}\cdot Z^*A^{*\prime}e_j=e_j'A^*[(n-1)\Lambda_m]A^{*\prime}e_j$$

$$=(n-1)e_j'\left[\sum_{t=1}^{m}\lambda_t a_t a_t'\right]e_j=(n-1)\sum_{t=1}^{m}\lambda_t[e_j'a_t a_t'e_j]$$

$$= (n-1)\sum_{i=1}^{m}\lambda_i a_{ji}^2 = (n-1)\sum_{i=1}^{m}\rho^2(X_j, Z_i)$$

$$= (n-1)v_j. \left[记\ v_j = \sum_{i=1}^{m}\rho^2(X_j, Z_i)\right].$$

由平方和分解公式及 $\bar{x}_j=0$ 可得 X_j 的残差平方和为

$$Q_j = x_j' x_j - \hat{x}_j' \hat{x}_j$$

$$= (n-1)(1-v_j).$$

X_j 的决定系数

$$R_j^2 = \frac{U_j}{U_j + Q_j} = \frac{(n-1)v_j}{n-1} = v_j.$$

7-9 (1) 证明一　直接由样本的似然函数求 λ_1 的最大似然估计量.

证明二　因 $\Sigma = \lambda_1 I_p$ 的特征值 $\lambda_1 = \lambda_2 = \cdots = \lambda_p \triangleq \lambda_1$，由主成分的性质(2)知

$$p\lambda_1 = \sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} \quad (*)$$

其中 $\sigma_{ii} = \mathrm{Var}(X_i)$ $(i = 1, 2, \cdots, p)$. 由 X_i 的次观测数据 $(x_{1i}, x_{2i}, \cdots, x_{ni})$ 可得 σ_{ii}的最大似然估计量为

$$\hat{\sigma}_{ii} = \frac{1}{n}\sum_{t=1}^{n}(x_{ti} - \bar{x}i)^2,$$

$(*)$右边的最大似然估计量为

$$\frac{1}{n}\sum_{i=1}^{p}\sum_{t=1}^{n}(x_{ti} - \bar{x}i)^2$$

故

$$\hat{\lambda}_1 = \frac{1}{np}\sum_{i=1}^{p}\sum_{t=1}^{n}(x_{ti} - \bar{x}i)^2.$$

(2) 对任给正交阵 $B=(b_1, b_2, \cdots, b_p)$，令

$$Z = B'X = \begin{pmatrix} b_1'X \\ \vdots \\ b_p'X \end{pmatrix} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_p \end{pmatrix},$$

以下由主成分的定义来说明 $Z=(Z_1, \cdots, Z_p)'$ 是 X 的主成分.

这里 $Z_i = b_i'X$ $(i=1,2,\cdots,p)$ 满足：

① $b_i'b_i=1$ $(i=1,2,\cdots,p)$；

② 因 $COV(Z) = COV(B'X) = B'\Sigma B = \lambda_1 I_p$，即 Z 的 p 个分量 $Z_1, \cdots, Z_p$ 互不相关。且

$$Var(Z_i) = \lambda_i (i = 1,2,\cdots,p)$$

$$Cov(Z_i, Z_j) = b_i'\Sigma b_j = \lambda_1 b_i'b_j = 0 (当\ j = 1,2,, i-1).$$

③ 设 $a'a=1$，$a'b_j=0$ $(j=1,2,, i-1)$，则

$$Var(a'X) = a'\Sigma a = \lambda_1 \leqslant Var(Z_i) = \lambda_1,$$

由主成分的定义可知 $Z_i = b_i'X$ $(i=1,2,\cdots,p)$ 为 X 的第 i 个主成分.

7-10 证明：若 $L'X$ 是 X 的主成分。设 Σ 的特征值为 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p \geqslant 0$，则有

$$L'\Sigma L = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_p) \triangleq \Lambda_p，从而$$

$$L'(\Sigma + \sigma^2 I_p) L = L'\Sigma L + L'\sigma^2 I_p L = \Lambda_p + \sigma^2 I_p$$

$$= \mathrm{diag}(\lambda_1 + \sigma^2, \lambda_2 + \sigma^2, \cdots, \lambda_p + \sigma^2).$$

即 $\lambda_1 + \sigma^2 \geqslant \lambda_2 + \sigma^2 \geqslant \cdots \geqslant \lambda_p + \sigma^2$ 为 $\Sigma + \sigma^2 I_p$ 的特征值，L 的列向量为相应的特征向量. 从而 $L'Y$ 是 Y 的主成分.

反之，若 $L'Y$ 是 Y 的主成分。Y 的协差阵 $\Sigma + \sigma^2 I_p$ 的特征值记为 $\upsilon_1 \geqslant \upsilon_2 \geqslant \cdots \geqslant \upsilon_p \geqslant 0$，则有

$$L'(\Sigma + \sigma^2 I_p) L = \mathit{diag}(\upsilon_1, \upsilon_2, \cdots, \upsilon_p)，从而$$

$$L'\Sigma L = \mathit{diag}(\upsilon_1 - \sigma^2, \upsilon_2 - \sigma^2, \cdots, \upsilon_p - \sigma^2).$$

因 $\Sigma > 0$，$\upsilon_1 - \sigma^2 \geqslant \upsilon_2 - \sigma^2 \geqslant \cdots \geqslant \upsilon_p - \sigma^2$ 为 Σ 的特征值，L 的列向量为相应的特征向量. 从而 $L'X$ 是 X 的主成分.

第七章 主成分分析

7-11 (1) 八个指标若综合为三个主成分，可解释原变量信息的86.66%．若综合为四个主成分，可解释原变量信息的94.68%．

(2) 按第一主分量得分由小到大对13个行业排的次序为：8,10,12,7,9,11,13,6,4,3,2,1,5.

7-12 六个指标若综合为二个主成分，可解释原变量信息的81.25%．若综合为三个主成分，可解释原变量信息的91.38%．

按第一主分量得分由小到大对16个地区农民的生活水平排的次序为：山西，河北，河南，江西，内蒙，黑龙江，福建，安徽，山东，吉林，江苏，辽宁，天津，浙江，北京，上海．

应用多元统计分析

第八章习题解答

8–1 设标准化随机变量X_1, X_2, X_3的协差阵(即相关阵)为

$$R=\begin{pmatrix} 1 & 0.63 & 0.45 \\ 0.63 & 1 & 0.35 \\ 0.45 & 0.35 & 1 \end{pmatrix},$$

试求$m=1$的正交因子模型.

解:设随机向量符合正交因子模型,则相关阵满足:

$$R=\begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \end{pmatrix}\begin{pmatrix} a_{11} & a_{21} & a_{31} \end{pmatrix}+\begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{pmatrix}.$$

比较等号两边可得出:

$$\begin{cases} a_{11}^2+\sigma_1^2=1 \\ a_{21}^2+\sigma_2^2=1 \\ a_{31}^2+\sigma_3^2=1 \\ a_{11}a_{21}=0.63 \\ a_{11}a_{31}=0.45 \\ a_{31}a_{21}=0.35 \end{cases}$$

$$\frac{a_{21}}{a_{31}}=\frac{0.63}{0.45}=\frac{7}{5},\ a_{21}=\frac{7}{5}a_{31}$$

$$a_{31}\frac{7}{5}a_{31}=0.35,$$

$$a_{31}^2=\frac{0.35\times5}{7}=0.25$$

$$\Rightarrow a_{31}=0.5,\ a_{21}=0.7,\ a_{11}=0.9,$$

$$\sigma_1^2=1-a_{11}^2=1-0.81=0.19,$$

$$\sigma_2^2=1-a_{21}^2=0.51,\ \sigma_3^2=1-a_{31}^2=0.75$$

故 $m=1$的正交因子模型为

$$\begin{cases} X_1 = 0.9F_1 + \varepsilon_1 \\ X_2 = 0.7F_1 + \varepsilon_2 \\ X_3 = 0.5F_1 + \varepsilon_3 \end{cases}$$

特殊因子$\boldsymbol{\varepsilon}$=$(\varepsilon_1, \varepsilon_2,\ldots,\varepsilon_p)'$的协差阵$\boldsymbol{D}$为：

$$D = \begin{pmatrix} 0.19 & 0 & 0 \\ 0 & 0.51 & 0 \\ 0 & 0 & 0.75 \end{pmatrix}$$

$8-2$ $8-1$ R

$\lambda_1=1.9633$ $l_1=(0.6250,0.5932,0.5075)'$,

$\lambda_2=0.6795$ $l_2=(-0.2186,-0.4911,0.8432)'$,

$\lambda_3=0.3672$ $l_3=(0.7494,-0.6379,-0.1772)'$.

(1) $m=1$, ,

$Q(1)$.

解: $m=1$的因子模型的主成分解为:

$$A=(\sqrt{\lambda_1}l_1)=\begin{pmatrix}0.8757\\0.8312\\0.7111\end{pmatrix},D=\begin{pmatrix}0.2331&0&0\\0&0.3091&0\\0&0&0.4943\end{pmatrix}$$

$$E_1 = R - (AA' + D)$$

$$= \begin{pmatrix} 1 & 0.63 & 0.45 \\ & 1 & 0.35 \\ & & 1 \end{pmatrix} - \begin{pmatrix} 1 & 0.7279 & 0.6227 \\ & 1 & 0.5911 \\ & & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 0 & -0.0979 & -0.1727 \\ & 0 & -0.2411 \\ & & 0 \end{pmatrix}$$

故 $Q(1)=\sum_{i=1}^{3}\sum_{j=1}^{3}\varepsilon_{ij}^2=2\times(0.0979^2+0.1727^2+0.2411^2)$

$=0.1951$

(2)取公因子个数 $m=2$时,求因子模型的主成分解，并计算误差平方和 $Q(2)$.

$: m=2$:

$$A=(\sqrt{\lambda_1}l_1,\sqrt{\lambda_2}l_2)=\begin{pmatrix} 0.8757 & -0.1802 \\ 0.8312 & -0.4048 \\ 0.7111 & 0.6950 \end{pmatrix},$$

$$D = \begin{pmatrix} 0.2007 & 0 & 0 \\ 0 & 0.1452 & 0 \\ 0 & 0 & 0.01131 \end{pmatrix}$$

$$m = 2$$

$$\begin{cases} X_1 = 0.8757F_1 - 0.1802F_2 + \varepsilon_1 \\ X_2 = 0.8312F_1 - 0.4048F_2 + \varepsilon_2 \\ X_3 = 0.7111F_1 + 0.6950F_2 + \varepsilon_3 \end{cases}$$

$$E_2 = R - (AA' + D) = \begin{pmatrix} 1 & 0.63 & 0.45 \\ & 1 & 0.35 \\ & & 1 \end{pmatrix} - (AA' + D)$$

$$AA'+D=\begin{pmatrix} 1 & 0.8008 & 0.4975 \\ & 1 & 0.3097 \\ & & 1 \end{pmatrix}$$

$$E_2=\begin{pmatrix} 0 & -0.1708 & -0.0475 \\ & 0 & 0.0403 \\ & & 0 \end{pmatrix}$$

$$Q(2)=\sum_{i=1}^{3}\sum_{j=1}^{3}\varepsilon_{ij}^2=2\times(0.1708^2+0.0475^2+0.0403^2)$$

$$=\ 0.06611$$

或者利用习题8-4的结果:

$$Q(m)=\sum_{i=1}^{p}\sum_{j=1}^{p}\varepsilon_{ij}^2=\sum_{j=m+1}^{p}\lambda_j^2-\sum_{i=1}^{p}(\sigma_i^2)^2\le\sum_{j=m+1}^{p}\lambda_j^2,$$

$$Q(1)=(\lambda_2^2+\lambda_3^2)-[(\sigma_1^2)^2+(\sigma_2^2)^2+(\sigma_3^2)^2]$$

$$=0.6795^2+0.3672^2-[0.2331^2+0.3091^2+0.4943^2]$$

$$=0.5966-0.3943=0.2023$$

$$Q(2)=\lambda_3^2-[(\sigma_1^2)^2+(\sigma_2^2)^2+(\sigma_3^2)^2]$$

$$=0.3672^2-[0.2007^2+0.1452^2+0.01131^2]$$

$$=0.1348-0.06149=0.07331$$

(3) 试求误差平方和$Q(m)$＜0.1的主成分解.

因$Q(2)$=0.07331＜0.1，故m=2的主成分解满足要求.

第八章 因子分析

8-3 验证下列矩阵关系式(**A**为**$p \times m$**阵)

$$(1)\quad (I+A'D^{-1}A)^{-1}A'D^{-1}A = I-(I+A'D^{-1}A)^{-1};$$

$$(2)\quad (AA'+D)^{-1} = D^{-1}-D^{-1}A(I+A'D^{-1}A)^{-1}A^{-1}D^{-1};$$

$$(3)\quad A'(AA'+D)^{-1} = (I_m+A'D^{-1}A)^{-1}A'D^{-1}.$$

解：利用分块矩阵求逆公式求以下分块矩阵的逆：

$$B=\begin{array}{cc} \begin{pmatrix} D & -A \\ A' & I_m \end{pmatrix} & \begin{array}{c} p \\ m \end{array} \end{array}$$

记$B_{22\bullet 1} = I_m + A'D^{-1}A, \qquad B_{11\bullet 2} = D + AA',$

利用附录中分块求逆的二个公式(**4.1**)和(**4.2**)有：

$$B^{-1}=\begin{pmatrix} D & -A \\ A' & I_m \end{pmatrix}^{-1}=\begin{pmatrix} B^{11} & B^{12} \\ B^{21} & B^{22} \end{pmatrix}$$

$$=\begin{pmatrix} D^{-1}-D^{-1}AB_{22\bullet 1}^{-1}A'D^{-1} & D^{-1}AB_{22\bullet 1}^{-1} \\ -B_{22\bullet 1}^{-1}A'D^{-1} & B_{22\bullet 1}^{-1} \end{pmatrix}$$

$$=\begin{pmatrix} B_{11\bullet 2}^{-1} & B_{11\bullet 2}^{-1}A \\ -A'B_{11\bullet 2}^{-1} & I_m-A'B_{11\bullet 2}^{-1}A \end{pmatrix}$$

由逆矩阵的对应块相等，即得：

$$\begin{cases} B_{11\bullet 2}^{-1} = D^{-1} - D^{-1}AB_{22\bullet 1}^{-1}A'D^{-1} = B^{11} \\ A'B_{11\bullet 2}^{-1} = B_{22\bullet 1}^{-1}A'D^{-1} = B^{21} \\ I_m - A'B_{11\bullet 2}^{-1}A = B_{22\bullet 1}^{-1} = B^{22} \end{cases}$$

把$\boldsymbol{B_{22\cdot 1}}$和$\boldsymbol{B_{11\cdot 2}}$式代入以上各式，可得：

$$\begin{cases} (D + AA')^{-1} = D^{-1} - D^{-1}A(I_m + A'D^{-1}A)^{-1}A'D^{-1} \quad (2) \\ A'(D + AA')^{-1} = (I_m + A'D^{-1}A)^{-1}A'D^{-1} \quad (3) \\ I_m - A'(D + AA')^{-1}A = (I_m + A'D^{-1}A)^{-1} \end{cases}$$

由第三式和第二式即得 $I_m - (I_m + A'D^{-1}A)^{-1} = A'(D + AA')^{-1}A$

$$= (I_m + A'D^{-1}A)^{-1}A'D^{-1}A \quad (1)$$

第八章 因子分析

8-4 证明公因子个数为m的主成分解，其误差平方和$Q(m)$满足以下不等式

$$Q(m)=\sum_{i=1}^{p}\sum_{j=1}^{p}\varepsilon_{ij}^{2}\le\sum_{j=m+1}^{p}\lambda_{j}^{2},$$

其中$E=S-(AA'+D)=(\varepsilon_{ij})$,$A$,$D$是因子模型的主成分估计.

解： 设样本协差阵S有以下谱分解式：

$$S=\sum_{i=1}^{p}\lambda_i l_i l_i'=\sum_{i=1}^{m}\lambda_i l_i l_i'+\sum_{i=m+1}^{p}\lambda_i l_i l_i'$$

其中$\lambda_1\ge\lambda_2\ge\cdots\ge\lambda_p\ge 0$为$S$的特征值，$l_i$为相应的标准特征向量。

设A,D是因子模型的主成分估计,即

$$A=\left(\sqrt{\lambda_1}l_1 \;\cdots\; \sqrt{\lambda_m}l_m\right),$$

若记 $B=\left(\sqrt{\lambda_{m+1}}l_{m+1} \;\cdots\; \sqrt{\lambda_p}l_p\right)$, 有

$$S=(A\,|\,B)\begin{pmatrix} A' \\ B' \end{pmatrix}=AA'+BB'$$

则 $D=\mathrm{diag}(BB')$

$$E=S-(AA'+D)=BB'-D, \quad BB'=E+D.$$

因

$$B'B=\begin{pmatrix}\sqrt{\lambda_{m+1}}l'_{m+1}\\ \vdots\\ \sqrt{\lambda_p}l'_p\end{pmatrix}(\sqrt{\lambda_{m+1}}l_{m+1},\cdots,\sqrt{\lambda_p}l_p)=\begin{pmatrix}\lambda_{m+1} & & 0\\ & \ddots & \\ 0 & & \lambda_p\end{pmatrix}$$

故 $\sum_{j=m+1}^{p}\lambda_j^2=\mathrm{tr}(B'B\cdot B'B)=\mathrm{tr}(BB'\cdot BB')$

$$=\mathrm{tr}[(E+D)'(E+D)]=\mathrm{tr}[E'E+E'D+D'E+D'D]$$

$$=Q(m)+0+0+\sum_{i=1}^{p}(\sigma_i^2)^2$$

所以 $Q(m)=\sum_{i=1}^{p}\sum_{j=1}^{p}\varepsilon_{ij}^2=\sum_{j=m+1}^{p}\lambda_j^2-\sum_{i=1}^{p}(\sigma_i^2)^2\le\sum_{j=m+1}^{p}\lambda_j^2,$

第八章 因子分析

8-5 试比较主成分分析和因子分析的相同之处与不同点.

因子分析与主成分分析的不同点有:

(1) 主成分分析不能作为一个模型来描述,它只是通常的变量变换,而因子分析需要构造因子模型;

(2) 主成分分析中主成分的个数和变量个数p相同,它是将一组具有相关关系的变量变换为一组互不相关的变量(注意应用主成分分析解决实际问题时,一般只选取前$m(m<p)$个主成分),而因子分析的目的是要用尽可能少的公共因子,以便构造一个结构简单的因子模型;

(3) 主成分分析是将主成分表示为原变量的线性组合,而因子分析是将原始变量表示为公因子和特殊因子的线性组合,用假设的公因子来“解释”相关阵的内部依赖关系.

这两种分析方法又有一定的联系.当估计方法采用主成分法,因子载荷阵**A**与主成分的系数相差一个倍数;因子得分与主成分得分也仅相差一个常数.这种情况下可把因子分析看成主成分分析的推广和发展.

这两种方法都是降维的统计方法,它们都可用来对样品或变量进行分类.