第一章　随机向量

§1.1　均值向量和协方差矩阵

一、随机向量和随机矩阵

1 、设 $\mathbf{X}_p=(X_1,X_2,\ldots,X_p)'$, 若对于任意的 $i\in\{1,2,\ldots,p\}$, X_i 均为随机变量，则称 $\mathbf{X}_p$ 为 p 维随机向量.

2 、若矩阵 $X_{m\times n}=(X_{ij})_{m\times n}$ 中每一个元素 X_{ij} 均为随机变量，则称 X 为 $m\times n$ 随机矩阵.

二、均值向量和协方差矩阵

1 、设 $\mathbf{X}_p=(X_1,X_2,\ldots,X_p)'$, 若 $EX_i=\mu_i,i=1,2,\ldots,p$ 存在，则称 $E\mathbf{X}=(\mu_1,\mu_2,\ldots,\mu_p)'\overset{\triangle}{=}\mu'$ 为随机向量 $\mathbf{X}$ 的均值向量. 同理，称 $E\mathbf{X}_{m\times n}=(EX_{ij})_{m\times n}$ 为随机矩阵 X 的期望矩阵或均值矩阵.

其中

$$EX_{ij}=\begin{cases}\sum\limits_{i,j}x_{ij}p_{ij} & \begin{cases}\text{若}X_{ij}\text{为离散型随机变量,}\\ \text{概率分布律为}p_{ij}\end{cases}\\ \int_{-\infty}^{+\infty}xf_{ij}(x)\,dx & \begin{cases}\text{若}X_{ij}\text{为连续型随机变量,}\\ \text{概率密度函数为}f_{ij}(x)\end{cases}\end{cases}$$

2 、协方差矩阵

令 $\sigma_{ii}\overset{\triangle}{=}{\sigma_i}^2,i=1,2,\ldots,p$

$$\begin{aligned}\sigma_{ij} &= E(X_i-\mu_i)(X_j-\mu_j),\\ &= \begin{cases}\sum\limits_{x_i,x_j}(x_i-\mu_i)(x_j-\mu_j)p_{ij}, & i,j=1,\ldots,p\\ \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x_i-\mu_i)(x_j-\mu_j)f_{ij}(x_i,x_j)\,dx_i\,dx_j, & i,j=1,\ldots,p\end{cases}\end{aligned}$$

其中 $f_{ij}(x_i,x_j)$ 为随机向量 (X_i,X_j) 的联合概率密度.

若 $f(x_1,\ldots,x_p)$ 为随机向量 $\mathbf{X}$ 的联合概率密度，则

$$f_{ij}(x_i,x_j)=\iint\cdots\int\limits_{R^{p-2}}f(x_1,\ldots,x_p)\,dx_1\ldots dx_{i-1}\,dx_{i+1}\,\ldots\,dx_{j-1}\,dx_{j+1}\,\ldots\,dx_p$$

(不妨设 $i<j$). 则称

$$\Sigma=\begin{pmatrix}\sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p}\\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p}\\ \vdots & \vdots & \cdots & \vdots\\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp}\end{pmatrix}$$

为随机向量 $\mathbf{X}$ 的协方差矩阵. 由 Σ 的定义知,

$$\begin{aligned}
\Sigma &= \begin{pmatrix} E(X_1-\mu_1)^2 & E(X_1-\mu_1)(X_2-\mu_2) & \cdots & E(X_1-\mu_1)(X_p-\mu_p) \\ E(X_2-\mu_2)(X_1-\mu_1) & E(X_2-\mu_2)^2 & \cdots & E(X_2-\mu_2)(X_p-\mu_p) \\ \vdots & \vdots & \cdots & \vdots \\ E(X_p-\mu_p)(X_1-\mu_1) & E(X_p-\mu_p)(X_2-\mu_2) & \cdots & E(X_p-\mu_p)^2 \end{pmatrix} \\
&= E\begin{pmatrix} (X_1-\mu_1)^2 & (X_1-\mu_1)(X_2-\mu_2) & \cdots & (X_1-\mu_1)(X_p-\mu_p) \\ (X_2-\mu_2)(X_1-\mu_1) & (X_2-\mu_2)^2 & \cdots & (X_2-\mu_2)(X_p-\mu_p) \\ \vdots & \vdots & \cdots & \vdots \\ (X_p-\mu_p)(X_1-\mu_1) & (X_p-\mu_p)(X_2-\mu_2) & \cdots & (X_p-\mu_p)^2 \end{pmatrix} \\
&= E\left[\begin{pmatrix} X_1-\mu_1 \\ X_2-\mu_2 \\ \vdots \\ X_p-\mu_p \end{pmatrix}(X_1-\mu_1, X_2-\mu_2, \ldots, X_p-\mu_p)\right] \\
&= E(\mathbf{X}-\mu)(\mathbf{X}-\mu)' \triangleq cov\,\mathbf{X}
\end{aligned}$$

故随机向量 $\mathbf{X}$ 的协方差矩阵也可写成

$$cov\mathbf{X} = E(\mathbf{X}-\mu)(\mathbf{X}-\mu)'$$

下面分析协方差矩阵 Σ 的特点：由定义知

- 对于任意的 $i,j=1,\ldots,p$, 有 $\sigma_{ij}=\sigma_{ji}$, 即 Σ 为对称矩阵.
- 当 $i=j$ 时， σ_{ij} 为第 i 个分量 X_i 的方差.
- 对于任意的 $i,j=1,\ldots,p,\sigma_{ij}$ 表示 $\mathbf{X}$ 的第 i 个分量与第 j 个分量的协方差. 若 $\sigma_{ij}=0$, 即 $cov(X_i,X_j)=0$, 则称 X_i 与 X_j 是不相关的. 在概率论中，我们已经知道，若 X_i 与 X_j 相互独立，则 X_i 与 X_j 不相关，但反之未必成立.

3 、相关系数矩阵

令 $\rho_{ij}=\frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}}, i,j=1,\ldots,p$, 则 ρ_{ij} 为变量 X_i 与 X_j 的线性相关系数，它度量了随机变量 X_i 与 X_j 之间的线性相关程度. $|\rho_{ij}|$ 的值越大，说明 X_i 与 X_j 之间的线性相关程度越大，反之越小. 当 $\rho_{ij}>0$ 时， X_i 与 X_j 正相关，当 $\rho_{ij}<0$ 时， X_i 与 X_j 负相关. 称 $p\times p$ 阶矩阵

$$\rho=(\rho_{ij})_{p\times p}$$

为随机变量 $\mathbf{X}$ 的相关系数矩阵.

若记 $\mathbf{V}^{\frac{1}{2}} = \begin{pmatrix} \sqrt{\sigma_{11}} & & & \\ & \sqrt{\sigma_{22}} & & \\ & & \ddots & \\ & & & \sqrt{\sigma_{pp}} \end{pmatrix}_{p\times p}$

则 $\rho = \mathbf{V}^{-\frac{1}{2}}\Sigma\mathbf{V}^{-\frac{1}{2}}$

若 $\mathbf{V}^{\frac{1}{2}}$ 已知，则 ρ 与 Σ 可以相互确定．事实上，

$$\rho = \mathbf{V}^{-\frac{1}{2}}\Sigma\mathbf{V}^{-\frac{1}{2}} \Longleftrightarrow \Sigma = V^{\frac{1}{2}}\rho V^{\frac{1}{2}}$$

例 1.1.1 设 $\mathbf{X} = (X_1, X_2)$, 其联合概率分布律为

$X_1 \backslash X_2$	0	1	p_{x_1}
-1	0.24	0.06	0.3
0	0.16	0.14	0.3
1	0.40	0.00	0.4
p_{x_2}	0.8	0.2	1

求 $E\mathbf{X}$, $Cov\,\mathbf{X}$, ρ.

解:

$\mu_1 = (-1)\times 0.3 + 0\times 0.3 + 1\times 0.4 = 0.1$

$\mu_2 = 0\times 0.8 + 1\times 0.2 = 0.2$

$$E\mathbf{X} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} 0.1 \\ 0.2 \end{pmatrix}$$

$$\begin{aligned} \sigma_{11} &= E(X_1-\mu_1)^2 = EX_1{}^2 - \mu_1{}^2 \\ &= (-1)^2\times 0.3 + 0^2\times 0.3 + 1^2\times 0.4 - 0.1^2 \\ &= 0.69 \\ \sigma_{22} &= E(X_2-\mu_2)^2 \\ &= EX_2{}^2 - \mu_2{}^2 = 0^2\times 0.8 + 1^2\times 0.2 - 0.2^2 = 0.16 \end{aligned}$$

$$\begin{aligned} \sigma_{12} &= E(X_1-\mu_1)(X_2-\mu_2) \\ &= EX_1X_2 - \mu_1 EX_2 - \mu_2 EX_1 + \mu_1\mu_2 \\ &= EX_1X_2 - \mu_1\mu_2 \\ &= (-1)\times 0\times 0.24 + (-1)\times 1\times 0.06 + 0\times 0\times 0.16 \\ &\quad +0\times 1\times 0.14 + 1\times 0\times 0.40 + 1\times 1\times 0.00 - 0.1\times 0.2 \\ &= -0.06 - 0.02 = -0.08 \end{aligned}$$

由对称性知 $\sigma_{21} = -0.08$

所以

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} = \begin{pmatrix} 0.69 & -0.08 \\ -0.08 & 0.16 \end{pmatrix}$$

从而

$$\rho = \begin{pmatrix} 1 & \frac{-0.08}{\sqrt{0.69}\sqrt{0.16}} \\ \frac{-0.08}{\sqrt{0.69}\sqrt{0.16}} & 1 \end{pmatrix} = \begin{pmatrix} 1 & -0.96 \\ -0.96 & 1 \end{pmatrix}$$

三、随机向量的线性变换的均值向量和协方差矩阵

设 $\mathbf{X}$ 是 p 维随机向量， $\mathbf{A}$ 为 $m \times p$ 阶常数矩阵， $\mathbf{b}$ 为 m 维向量，令 $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$, 则 $\mathbf{Y}$ 为 m 维随机向量.

结论 1.1.1 $E\mathbf{Y} = \mathbf{A}\,E\,\mathbf{X} + \mathbf{b} = \mathbf{A}\mu + \mathbf{b}$.

证明：设 $\mathbf{A} = (a_{ij})_{m\times p}$, $\mathbf{b} = (b_1, b_2, \ldots, b_m)'$. 则

$Y_i = \sum_{k=1}^{p} a_{ik}X_k + b_i$, $i = 1, \ldots, m.E\,(Y_i) = \sum_{k=1}^{p} a_{ik}EX_k + b_i$, $i = 1, \ldots, m.$

$$\begin{aligned} E\mathbf{Y} &= \begin{pmatrix} EY_1 \\ EY_2 \\ \vdots \\ EY_m \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^{p} a_{1k}EX_k + b_1 \\ \sum_{k=1}^{p} a_{2k}EX_k + b_2 \\ \vdots \\ \sum_{k=1}^{p} a_{mk}EX_k + b_m \end{pmatrix} \\ &= \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mp} \end{pmatrix} \begin{pmatrix} EX_1 \\ EX_2 \\ \vdots \\ EX_p \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix} \\ &= \mathbf{A}\,E\,\mathbf{X} + \mathbf{b} \end{aligned}$$

结论 1.1.2 $Cov\mathbf{Y} = \mathbf{A}\,Cov\,\mathbf{X}\,\mathbf{A}' = \mathbf{A}\Sigma\mathbf{A}'$

证明： $$\begin{aligned} Cov\mathbf{Y} &= E\left[(\mathbf{Y}-E\mathbf{Y})(\mathbf{Y}-E\mathbf{Y})'\right] \\ &= E\left[(\mathbf{AX}+\mathbf{b}-\mathbf{A}\mu-\mathbf{b})(\mathbf{AX}+\mathbf{b}-\mathbf{A}\mu-\mathbf{b})'\right] \\ &= E\left[\mathbf{A}(\mathbf{X}-\mu)(\mathbf{X}-\mu)'\mathbf{A}'\right] \\ &= \mathbf{A}E\left[(\mathbf{X}-\mu)(\mathbf{X}-\mu)'\mathbf{A}'\right] \\ &= \mathbf{A}\left\{E\left[(\mathbf{X}-\mu)(\mathbf{X}-\mu)'\mathbf{A}'\right]'\right\}' \\ &= \mathbf{A}\left\{E\left[\mathbf{A}(\mathbf{X}-\mu)(\mathbf{X}-\mu)'\right]\right\}' \\ &= \mathbf{A}\left\{\mathbf{A}E\left[(\mathbf{X}-\mu)(\mathbf{X}-\mu)'\right]\right\}' \\ &= \mathbf{A}(\mathbf{A}\Sigma)' = \mathbf{A}\Sigma'\mathbf{A}' = \mathbf{A}\Sigma\mathbf{A}' \end{aligned}$$

例 1.1.2　设 $\mathbf{X}=(X_1,X_2)'$, $E\,\mathbf{X}=\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}$, $Cov\,\mathbf{X}=\begin{pmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{pmatrix}$.

求线性组合 $\begin{aligned} Y_1 &= X_1-X_2\\ Y_2 &= X_1+X_2\end{aligned}$ 的均值向量和协方差阵 .

解： 因为 $\mathbf{Y}=\begin{pmatrix}Y_1\\ Y_2\end{pmatrix}=\begin{pmatrix}1 & -1\\ 1 & 1\end{pmatrix}\begin{pmatrix}X_1\\ X_2\end{pmatrix}=\mathbf{AX}$

所以 $E\mathbf{Y}=\mathbf{A}\,E\,\mathbf{X}=\begin{pmatrix}1 & -1\\ 1 & 1\end{pmatrix}\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}=\begin{pmatrix}\mu_1-\mu_2\\ \mu_1+\mu_2\end{pmatrix}$

$$\begin{aligned} Cov\mathbf{Y} &= \mathbf{A}\,Cov\,\mathbf{X}\,\mathbf{A}' \\ &= \begin{pmatrix}1 & -1\\ 1 & 1\end{pmatrix}\begin{pmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{pmatrix}\begin{pmatrix}1 & 1\\ -1 & 1\end{pmatrix} \\ &= \begin{pmatrix}\sigma_{11}+\sigma_{22}-2\sigma_{12} & \sigma_{11}-\sigma_{22}\\ \sigma_{11}-\sigma_{22} & \sigma_{11}+\sigma_{22}+\sigma_{12}\end{pmatrix} \end{aligned}$$

特别当 $\sigma_{11}=\sigma_{22}$ 时， $\sigma_{11}-\sigma_{22}=0$, 说明随机变量 Y_1 与 Y_2 不相关，即方差相等的两个随机变量的和与差是不相关的.

例 1.1.3　设一元随机变量 $X\sim N\left(\mu,\sigma^2\right)$, $X_1,\ldots,X_n$ 是来自 X 的一个样本. $\overline{X}=\frac{1}{n}\sum\limits_{i=1}^{n}X_i$, 求 $E\overline{X}$, $D\overline{X}$.

解：构造向量 $\mathbf{X}=(X_1,\ldots,X_n)'$, 则

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i=\frac{1}{n}(1,1,\ldots,1)\begin{pmatrix}X_1\\X_2\\\vdots\\X_n\end{pmatrix}\triangleq\frac{1}{n}\mathbf{1}'\mathbf{X}$$

因为 $X_1,\ldots,X_n$ 是来自总体 X 的样本，所以 $X_1,\ldots,X_n$ 独立同分布，从而 $E\mathbf{X}=\mu\mathbf{1}$, $Cov\,\mathbf{X}=\sigma^2 I$, 由结论 1.1.1 与结论 1.1.2 知

$$E\overline{X}=\frac{1}{n}\mathbf{1}'E\mathbf{X}=\frac{1}{n}\mathbf{1}'\begin{pmatrix}EX_1\\EX_2\\\vdots\\EX_n\end{pmatrix}=\frac{1}{n}\mathbf{1}'\begin{pmatrix}\mu\\\mu\\\vdots\\\mu\end{pmatrix}=\frac{\mu}{n}\mathbf{1}'\mathbf{1}=\mu$$

即 $\overline{X}$ 是 μ 的无偏估计.

$$D\overline{X}=Cov\overline{X}=\frac{1}{n}\mathbf{1}'Cov\,\mathbf{X}\cdot\frac{1}{n}\mathbf{1}=\frac{\sigma^2}{n^2}\mathbf{1}'\mathbf{1}=\frac{\sigma^2}{n}$$

特别的，由向量 $\mathbf{X}_{p\times 1}$ 的各个分量的线性组合所得到的变量

$$\begin{aligned}Y&=a_1X_1+a_2X_2+\cdots+a_pX_p\\&\triangleq\mathbf{a}'\mathbf{X}\end{aligned}$$

是一元随机变量，从而得到结论：

$$\begin{aligned}EY&=\mathbf{a}'EX=\mathbf{a}'\mu=a_1\mu_1+a_2\mu_2+\cdots+a_p\mu_p\\DY&=\mathbf{a}'Cov\,\mathbf{X}\mathbf{a}=\mathbf{a}'\Sigma\mathbf{a}\end{aligned}$$

若 $\mathbf{X}$ 和 $\mathbf{Y}$ 分别为 p 维和 q 维随机变量，我们定义

$$Cov\,(\mathbf{X},\mathbf{Y})=E\left[(\mathbf{X}-E\,\mathbf{X})\,(\mathbf{Y}-E\mathbf{Y})'\right]$$

则有下面的结论

结论 1.1.3　若 $\mathbf{X}$ 和 $\mathbf{Y}$ 分别为 p 维和 q 维随机变量，$\mathbf{A}$ 和 $\mathbf{B}$ 分别为 $m\times p$ 和 $n\times q$ 维常数矩阵，则

$$Cov\,(\mathbf{AX},\mathbf{BY})=\mathbf{A}Cov\,(\mathbf{X},\mathbf{Y})\,\mathbf{B}'$$

证明：
$$\begin{aligned}Cov\,(\mathbf{AX},\mathbf{BY})&=E\left[(\mathbf{AX}-\mathbf{A}E\,\mathbf{X})\,(\mathbf{BY}-\mathbf{B}E\mathbf{Y})'\right]\\&=E\left[\mathbf{A}\,(\mathbf{X}-E\,\mathbf{X})\,(\mathbf{Y}-E\mathbf{Y})'\,\mathbf{B}'\right]\\&=\mathbf{A}E\left[(\mathbf{X}-E\,\mathbf{X})\,(\mathbf{Y}-E\mathbf{Y})'\right]\mathbf{B}'\\&=\mathbf{A}Cov\,(\mathbf{X},\mathbf{Y})\,\mathbf{B}'\end{aligned}$$

特别地，当 $\begin{array}{rcl} Y_1 &=& a_1X_1+\cdots+a_pX_p=\mathbf{a}'\mathbf{X} \\ Y_2 &=& b_1X_1+\cdots+b_pX_p=\mathbf{b}'\mathbf{X} \end{array}$ 时，

$$\begin{aligned} \text{有 } Cov\left(Y_1,Y_2\right) &= Cov\left(\mathbf{a}'\mathbf{X},\mathbf{b}'\mathbf{X}\right) \\ &= \mathbf{a}'Cov\left(\mathbf{X},\mathbf{X}\right)\mathbf{b} \\ &= \mathbf{a}'\Sigma\mathbf{b} \end{aligned}$$

§1.2 随机向量的二次型

设 $\mathbf{X}=(X_1,X_2,\ldots,X_p)'$ 为 p 维随机向量， $\mathbf{A}$ 为 $p\times p$ 阶对称矩阵，则称随机变量

$$\mathbf{X}'\mathbf{A}\mathbf{X}=\sum_{i=1}^{p}\sum_{j=1}^{p}a_{ij}X_iX_j$$

为 $\mathbf{X}$ 的二次型.

结论 1.2.1 设 $E\,\mathbf{X}=\mu$, $Cov\,\mathbf{X}=\Sigma$, 则

$$E\left(\mathbf{X}'\mathbf{A}\mathbf{X}\right)=\mu'\mathbf{A}\mu+tr\left(\mathbf{A}\Sigma\right)$$

其中 $tr\mathbf{A}$ 表示矩阵 $\mathbf{A}$ 的对角线上的元素之和，称为矩阵 $\mathbf{A}$ 的迹.

证明：因为

$$\begin{aligned} \mathbf{X}'\mathbf{A}\mathbf{X} &= \left(\mathbf{X}-\mu+\mu\right)'\mathbf{A}\left(\mathbf{X}-\mu+\mu\right) \\ &= \left(\mathbf{X}-\mu\right)'\mathbf{A}\left(\mathbf{X}-\mu\right)+\mu'\mathbf{A}\left(\mathbf{X}-\mu\right)+\left(\mathbf{X}-\mu\right)'\mathbf{A}\mu+\mu'\mathbf{A}\mu \end{aligned}$$

$E\left[\mu'\mathbf{A}\left(\mathbf{X}-\mu\right)\right]=\mu'\mathbf{A}E\left(\mathbf{X}-\mu\right)$

而

$=\mu'\mathbf{A}\left(E\,\mathbf{X}-\mu\right)=0$

又因为对任两个 n 维向量 $\mathbf{a}$ 、 $\mathbf{b}$, 有 $\mathbf{a}'\mathbf{b}=\mathbf{b}'\mathbf{a}$, 所以

$E\left[\left(\mathbf{X}-\mu\right)'\mathbf{A}\mu\right]=E\left[\mu'\mathbf{A}\left(\mathbf{X}-\mu\right)\right]=0.$

故

$$E\left(\mathbf{X}'\mathbf{A}\mathbf{X}\right)=E\left[\left(\mathbf{X}-\mu\right)'\mathbf{A}\left(\mathbf{X}-\mu\right)\right]+\mu'\mathbf{A}\mu$$

$$=E\left[tr\left(\mathbf{X}-\mu\right)'\mathbf{A}\left(\mathbf{X}-\mu\right)\right]+\mu'\mathbf{A}\mu$$

$$=E\left[tr\mathbf{A}\left(\mathbf{X}-\mu\right)\left(\mathbf{X}-\mu\right)'\right]+\mu'\mathbf{A}\mu$$

(因为$tr\mathbf{A}\mathbf{B}=tr\mathbf{B}\mathbf{A}$)

$$=trE\left[\mathbf{A}\left(\mathbf{X}-\mu\right)\left(\mathbf{X}-\mu\right)'\right]+\mu'\mathbf{A}\mu\ (\text{因为}trE\,\mathbf{X}=Etr\mathbf{X})$$

$$=tr\left(\mathbf{A}\Sigma\right)+\mu'\mathbf{A}\mu$$

特别地

(1) 若$\mu=0$, 则$E\left(\mathbf{X}'\mathbf{A}\mathbf{X}\right)=tr\left(\mathbf{A}\Sigma\right)$;

(2) 若$\Sigma=\sigma^2I$, 则$E\left(\mathbf{X}'\mathbf{A}\mathbf{X}\right)=\mu'\mathbf{A}\mu+\sigma^2tr\mathbf{A}$;

(3) 若$\mu=0$, $\Sigma=\sigma^2I$, 则$E\left(\mathbf{X}'\mathbf{A}\mathbf{X}\right)=\sigma^2tr\mathbf{A}$.

例 1.2.1　设 $X\sim N\left(\mu,\sigma^2\right)$, $X_1,\ldots,X_n$ 是来自 X 的一个样本.　$s^2=\frac{1}{n-1}\sum\limits_{i=1}^{n}\left(X_i-\overline{X}\right)^2$,　$\overline{X}=\frac{1}{n}\sum\limits_{i=1}^{n}X_i$, 求 Es^2.

解：令 $\mathbf{X}=\left(X_1,\ldots,X_n\right)'$, 因为 $X_1,\ldots,X_n$ 是来自 X 的样本，所以 $X_1,\ldots,X_n$ 独立同分布，从而

$$E\,\mathbf{X}=\mu\mathbf{1}\ ,\ Cov\,\mathbf{X}=\sigma^2\mathbf{I}$$

又因为　$\left(n-1\right)s^2=\sum\limits_{i=1}^{n}\left(X_i-\overline{X}\right)^2$

$$=\left(X_1-\overline{X},\ldots,X_n-\overline{X}\right)\begin{pmatrix}X_1-\overline{X}\\ \vdots\\ X_n-\overline{X}\end{pmatrix}$$

而　$\begin{pmatrix}X_1-\overline{X}\\ \vdots\\ X_n-\overline{X}\end{pmatrix}=\begin{pmatrix}X_1\\ \vdots\\ X_n\end{pmatrix}-\begin{pmatrix}\overline{X}\\ \vdots\\ \overline{X}\end{pmatrix}=\mathbf{X}-\mathbf{1}\overline{X}$

$$=\mathbf{X}-\tfrac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{X}=\left(\mathbf{I}-\tfrac{1}{n}\mathbf{1}\mathbf{1}'\right)\mathbf{X}$$

所以

$$\left(n-1\right)s^2=\mathbf{X}'\left(\mathbf{I}-\tfrac{1}{n}\mathbf{1}\mathbf{1}'\right)'\left(\mathbf{I}-\tfrac{1}{n}\mathbf{1}\mathbf{1}'\right)\mathbf{X}$$

$$=\mathbf{X}'\left(\mathbf{I}-\tfrac{1}{n}\mathbf{1}\mathbf{1}'\right)\mathbf{X}$$

$$
\begin{aligned}
E\left[(n-1)s^2\right] &= (E\mathbf{X})'\left(\mathbf{I}-\tfrac{1}{n}\mathbf{1}\mathbf{1}'\right)E\mathbf{X}+tr\left[\left(\mathbf{I}-\tfrac{1}{n}\mathbf{1}\mathbf{1}'\right)Cov\,\mathbf{X}\right] \\
&= \mu\mathbf{1}'\left(\mathbf{I}-\tfrac{1}{n}\mathbf{1}\mathbf{1}'\right)\mu\mathbf{1}+\sigma^2 tr\left(\mathbf{I}-\tfrac{1}{n}\mathbf{1}\mathbf{1}'\right) \\
&= 0+(n-1)\sigma^2 \;=\; (n-1)\sigma^2
\end{aligned}
$$

从而 $Es^2=\sigma^2$, 即样本方差 s^2 是总体方差 σ^2 的无偏估计.

§1.3 样本均值向量和样本协方差阵

每我们试图了解一个社会现象或自然现象时，常常选择若干个变量（不妨设有 p 个）或事物的特征进行观测或试验，从而出现多元数据.

我们将用记号 X_{ik} 表示第 k 个变量在第 i 次试验中的观测值，则 p 个变量的 n 个观测值可以表示如下：

$$
\begin{array}{c|ccccccc}
 & X_1 & X_2 & \cdots & X_k & \cdots & X_p \\
1 & X_{11} & X_{12} & \cdots & X_{1k} & \cdots & X_{1p} \\
2 & X_{21} & X_{22} & \cdots & X_{2k} & \cdots & X_{2p} \\
\vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
i & X_{i1} & X_{i2} & \cdots & X_{ik} & \cdots & X_{ip} \\
\vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
n & X_{n1} & X_{n2} & \cdots & X_{nk} & \cdots & X_{np}
\end{array}
$$

或者用一个 $n\times p$ 阶矩阵来表示这些数据，即

$$
\mathcal{X}=\begin{pmatrix}
X_{11} & X_{12} & \cdots & X_{1p} \\
X_{21} & X_{22} & \cdots & X_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
X_{n1} & X_{n2} & \cdots & X_{np}
\end{pmatrix}
$$ 称为样本数据集

在做实验之前，我们用随机变量 X_{ik} 表示第 k 个变量在第 i 次试验中的可能观测值，此时

$$
\mathbf{X}\overset{\text{按行分块}}{=\!=}\begin{pmatrix}
\mathbf{X}_1{}' \\
\mathbf{X}_2{}' \\
\vdots \\
\mathbf{X}_n{}'
\end{pmatrix}\overset{\text{按列分块}}{=\!=}(\mathbf{Y}_1,\mathbf{Y}_2,\ldots,\mathbf{Y}_p)
$$

$\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_n$ 看作 p 元总体 $\mathbf{X}$ 的一个样本, $\mathbf{Y}_1,\mathbf{Y}_2,\ldots,\mathbf{Y}_p$ 分别看作 $\mathbf{X}$ 每个分量的 n 个观测值.

1. 样本均值向量

令 $\overline{X}_k = \frac{1}{n}\sum_{i=1}^{n} X_{ik}\ ,\ k = 1, 2, \ldots, p$, 则称 $\overline{\mathbf{X}} = \left(\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_p\right)'$ 为样本均值向量.

2. 样本协方差阵

令 $s_{kk} = {s_k}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_{ik} - \overline{X}_k\right)^2 (k = 1, 2, \ldots, p)$ 表示第 k 个变量的样本方差.

$s_{ij} = \frac{1}{n-1}\sum_{l=1}^{n}\left(X_{li} - \overline{X}_i\right)\left(X_{lj} - \overline{X}_j\right) (i\,, j = 1, 2, \ldots, p)$ 表示第 i 个变量与第 j 个变量的样本协方差. 则称 $\mathbf{S} = (s_{ij})_{p\times p}$ 为样本协方差阵.

3. 样本相关系数矩阵

$$\mathbf{R} = (r_{ij})_{p\times p} \quad ,\text{其中} \quad r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}}\ ,\ i\,, j = 1, 2, \ldots, p$$

设 $\mathbf{D}^{\frac{1}{2}} = \begin{pmatrix} \sqrt{s_{11}} & & & \\ & \sqrt{s_{22}} & & \\ & & \ddots & \\ & & & \sqrt{s_{pp}} \end{pmatrix}$

则 $\mathbf{R} = \mathbf{D}^{-\frac{1}{2}}\mathbf{S}\mathbf{D}^{-\frac{1}{2}} \Longleftrightarrow \mathbf{S} = \mathbf{D}^{\frac{1}{2}}\mathbf{R}\mathbf{D}^{\frac{1}{2}}$

4. 样本均值向量和协方差矩阵的性质

结论 1.3.1　设 $\mathbf{X}_{p\times 1}$ 是 p 维随机向量, $E\,\mathbf{X} = \mu\,, Cov\,\mathbf{X} = \Sigma\,, \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ 是来自 $\mathbf{X}$ 的一个随机样本, 则

(1) $E\overline{\mathbf{X}} = \mu \quad Cov\overline{\mathbf{X}} = \frac{1}{n}\Sigma$

(2) $E\mathbf{S} = \Sigma$

证明: (1) 因为 $\overline{\mathbf{X}} = \frac{1}{n}\left(\mathbf{X}_1 + \mathbf{X}_2 + \cdots + \mathbf{X}_n\right)$

所以 $E\overline{\mathbf{X}} = \frac{1}{n}\left(E\mathbf{X}_1 + E\mathbf{X}_2 + \cdots + E\mathbf{X}_n\right)$

$= \frac{1}{n}\left(\mu + \mu + \cdots + \mu\right)$

$= \mu$

即样本均值 $\overline{\mathbf{X}}$ 是总体均值的无偏估计.

$$\begin{aligned} Cov\overline{\mathbf{X}} &= \frac{1}{n^2}Cov\left(\mathbf{X}_1 + \mathbf{X}_2 + \cdots + \mathbf{X}_n\right) \\ &= \frac{1}{n^2}\left(Cov\mathbf{X}_1 + Cov\mathbf{X}_2 + \cdots + Cov\mathbf{X}_n\right) \\ &= \frac{1}{n^2}\cdot n\Sigma = \frac{1}{n}\Sigma \end{aligned}$$

$$\left(\begin{array}{l} \text{因为 } \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \text{ 独立同分布} \\ \text{所以 } E\mathbf{X}_i = \mu\,,\ Cov\mathbf{X}_i = \Sigma\ ,\ i = 1, 2, \ldots, n \end{array}\right)$$

因为 $(n-1)\,\mathbf{S} =$

$$\begin{pmatrix} \sum_{l=1}^{n} (X_{l1}-\overline{X}_1)^2 & \sum_{l=1}^{n} (X_{l1}-\overline{X}_1)(X_{l2}-\overline{X}_2) & \cdots & \sum_{l=1}^{n} (X_{l1}-\overline{X}_1)(X_{lp}-\overline{X}_p) \\ \sum_{l=1}^{n} (X_{l2}-\overline{X}_2)(X_{l1}-\overline{X}_1) & \sum_{l=1}^{n} (X_{l2}-\overline{X}_2)^2 & \cdots & \sum_{l=1}^{n} (X_{l2}-\overline{X}_2)(X_{lp}-\overline{X}_p) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{l=1}^{n} (X_{lp}-\overline{X}_p)(X_{l1}-\overline{X}_1) & \sum_{l=1}^{n} (X_{lp}-\overline{X}_p)(X_{l2}-\overline{X}_2) & \cdots & \sum_{l=1}^{n} (X_{lp}-\overline{X}_p)^2 \end{pmatrix}$$

$$= \sum_{l=1}^{n} \begin{pmatrix} (X_{l1}-\overline{X}_1)^2 & (X_{l1}-\overline{X}_1)(X_{l2}-\overline{X}_2) & \cdots & (X_{l1}-\overline{X}_1)(X_{lp}-\overline{X}_p) \\ (X_{l2}-\overline{X}_2)(X_{l1}-\overline{X}_1) & (X_{l2}-\overline{X}_2)^2 & \cdots & (X_{l2}-\overline{X}_2)(X_{lp}-\overline{X}_p) \\ \vdots & \vdots & \ddots & \vdots \\ (X_{lp}-\overline{X}_p)(X_{l1}-\overline{X}_1) & (X_{lp}-\overline{X}_p)(X_{l2}-\overline{X}_2) & \cdots & (X_{lp}-\overline{X}_p)^2 \end{pmatrix}$$

$$= \sum_{l=1}^{n} \begin{pmatrix} X_{l1}-\overline{X}_1 \\ X_{l2}-\overline{X}_2 \\ \vdots \\ X_{lp}-\overline{X}_p \end{pmatrix} \left(X_{l1}-\overline{X}_1, X_{l2}-\overline{X}_2, \ldots, X_{lp}-\overline{X}_p\right)$$

$$= \sum_{l=1}^{n} \left(\mathbf{X}_l - \overline{\mathbf{X}}\right)\left(\mathbf{X}_l - \overline{\mathbf{X}}\right)'$$

$$= \sum_{l=1}^{n} \mathbf{X}_l\mathbf{X}_l' - \sum_{l=1}^{n} \overline{\mathbf{X}}\mathbf{X}_l' - \sum_{l=1}^{n} \mathbf{X}_l\overline{\mathbf{X}}' + \sum_{l=1}^{n} \overline{\mathbf{X}}\overline{\mathbf{X}}'$$

$$= \sum_{l=1}^{n} \mathbf{X}_l\mathbf{X}_l' - n\overline{\mathbf{X}}\cdot\overline{\mathbf{X}}'$$

$$E\left[(n-1)\,\mathbf{S}\right] = \sum_{l=1}^{n} E\left(\mathbf{X}_l\mathbf{X}_l'\right) - n\,E\left[\overline{\mathbf{X}}\cdot\overline{\mathbf{X}}'\right]$$

又因为对任意随机向量 $\mathbf{Y}$, 我们有

$$\begin{aligned} E\left(\mathbf{Y}\mathbf{Y}'\right) &= E\left[(\mathbf{Y}-E\mathbf{Y}+E\mathbf{Y})(\mathbf{Y}-E\mathbf{Y}+E\mathbf{Y})'\right] \\ &= E\left[(\mathbf{Y}-E\mathbf{Y})(\mathbf{Y}-E\mathbf{Y})'\right] + E\left[E\mathbf{Y}(\mathbf{Y}-E\mathbf{Y})'\right] \\ &\quad + E\left[(\mathbf{Y}-E\mathbf{Y})\cdot E\mathbf{Y}'\right] + E\mathbf{Y}\cdot E\mathbf{Y}' \\ &= Cov\mathbf{Y} + E\mathbf{Y}\cdot E\mathbf{Y}' \end{aligned}$$

而对于任意的 $l \in \{1,2,\ldots,n\}$, $E\mathbf{X}_l = \mu$, $Cov\mathbf{X}_l = \Sigma$

$$E\overline{\mathbf{X}} = \mu\ ,\ Cov\overline{\mathbf{X}} = \frac{1}{n}\Sigma$$

所以

$$\begin{aligned} E\left[(n-1)\,\mathbf{S}\right] &= \sum_{l=1}^{n}\left[\Sigma+\mu\cdot\mu'\right]-n\left[\mu\mu'+\tfrac{1}{n}\Sigma\right] \\ &= n\Sigma+n\mu\mu'-n\mu\mu'-\Sigma \\ &= (n-1)\,\Sigma \end{aligned}$$

故 $E\mathbf{S}=\Sigma$

即 样本协方差阵是总体协方差阵的无偏估计.

5. 随机变量的线性组合的样本值

在许多实际问题中，我们常常要考虑一个形如

$$Y=b_1X_1+b_2X_2+\cdots+b_pX_p$$

的线性组合生成的变量.

(1) 随机变量 Y 的样本均值和样本方差

为了用 $X_1,X_2,\ldots,X_p$ 的样本来表示 Y 的样本均值和样本方差，我们不妨设 $\mathbf{X}=(X_1,X_2,\ldots,X_p)'$, $\mathbf{b}=(b_1,b_2,\ldots,b_p)'$, 则

$$Y=\mathbf{b}'\mathbf{X}$$

若 $\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_n$ 是来自 $\mathbf{X}$ 的样本，则随机变量 Y 对应的样本为

$$Y_1=\mathbf{b}'\mathbf{X}_1\,,\quad Y_2=\mathbf{b}'\mathbf{X}_2\,,\ldots\,,\,Y_n=\mathbf{b}'\mathbf{X}_n$$

根据样本均值和样本方差的定义

$$\begin{aligned} \overline{Y} &= \tfrac{1}{n}\sum_{i=1}^{n}Y_i=\tfrac{1}{n}\sum_{i=1}^{n}\left(\mathbf{b}'\mathbf{X}_i\right)=\tfrac{1}{n}\mathbf{b}'\left(\sum_{i=1}^{n}\mathbf{X}_i\right)=\mathbf{b}'\left(\tfrac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\right) \\ &= \mathbf{b}'\overline{\mathbf{X}} \\ {s_Y}^2 &= \tfrac{1}{n-1}\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2=\tfrac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{b}'\mathbf{X}_i-\mathbf{b}'\overline{\mathbf{X}}\right)\left(\mathbf{b}'\mathbf{X}_i-\mathbf{b}'\overline{\mathbf{X}}\right)' \\ &= \tfrac{1}{n-1}\sum_{i=1}^{n}\mathbf{b}'\left(\mathbf{X}_i-\overline{\mathbf{X}}\right)\left(\mathbf{X}_i-\overline{\mathbf{X}}\right)'\mathbf{b} \\ &= \mathbf{b}'\left[\tfrac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{X}_i-\overline{\mathbf{X}}\right)\left(\mathbf{X}_i-\overline{\mathbf{X}}\right)'\right]\mathbf{b} \\ &= \mathbf{b}'\mathbf{S}\mathbf{b} \end{aligned}$$

故有以下结论：

结论 1.3.1 设 $Y=\mathbf{b}'\mathbf{X}$, $\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_n$ 是来自 $\mathbf{X}$ 的样本，则

$$\overline{Y}=\mathbf{b}'\overline{\mathbf{X}}\,,\quad {s_Y}^2=\mathbf{b}'\mathbf{S}\mathbf{b}$$

其中 $\overline{\mathbf{X}}$ 与 $\mathbf{S}$ 分别为随机向量 $\mathbf{X}$ 的样本均值向量和样本协方差阵.

(2) 两个线性组合的样本协方差

结论 1.3.2 $Y=\mathbf{b}'\mathbf{X}$, $Z=\mathbf{c}'\mathbf{X}$, $\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_n$ 是来自 $\mathbf{X}$ 的样本，则 Y 与 Z 的样本协方差 $s_{YZ}=\mathbf{b}'\mathbf{S}\mathbf{c}$. 其中 $\mathbf{S}$ 是 $\mathbf{X}$ 的样本协方差阵.

事实上，由样本协方差的定义知

$$\begin{aligned}
s_{YZ} &= \tfrac{1}{n-1}\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)\left(Z_i-\overline{Z}\right)\\
&= \tfrac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{b}'\mathbf{X}_i-\mathbf{b}'\overline{\mathbf{X}}\right)\left(\mathbf{c}'\mathbf{X}_i-\mathbf{c}'\overline{\mathbf{X}}\right)'\\
&= \tfrac{1}{n-1}\sum_{i=1}^{n}\mathbf{b}'\left(\mathbf{X}_i-\overline{\mathbf{X}}\right)\left(\mathbf{X}_i-\overline{\mathbf{X}}\right)'\mathbf{c}\\
&= \mathbf{b}'\left[\tfrac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{X}_i-\overline{\mathbf{X}}\right)\left(\mathbf{X}_i-\overline{\mathbf{X}}\right)'\right]\mathbf{c}\\
&= \mathbf{b}'\mathbf{S}\mathbf{c}
\end{aligned}$$

例 1.3.1 设三维随机向量 $\mathbf{X}=(X_1,X_2,X_3)'$ 的样本数据集为

$$\mathbf{X}=\begin{pmatrix}1&2&5\\4&1&6\\4&0&4\end{pmatrix}$$

令 $Y=2X_1+2X_2-X_3$, $Z=X_1-X_2+3X_3$, 试计算 Y,Z 的样本均值和样本方差以及 Y 与 Z 的样本协方差.

解：方法 1　直接由 Y,Z 的样本观测值计算.

由 $\mathbf{X}_1=(1,2,5)'$, $\mathbf{X}_2=(4,2,5)'$, $\mathbf{X}_3=(4,0,4)'$ 得

$$Y_1=2\times1+2\times2-5=1$$

$$Y_2=2\times4+2\times1-6=4$$

$$Y_3=2\times4+2\times0-4=4$$

所以 $\overline{Y}=\frac{1}{3}(1+4+4)=3$

$${s_Y}^2=\frac{1}{2}\left[(1-3)^2+(4-3)^2+(4-3)^2\right]=3$$

同理可得: $Z_1=14$, $Z_2=21$, $Z_3=16$

$$\overline{Z}=17\,,\quad {s_Z}^2=13$$

$$s_{YZ}=\frac{1}{2}\left[(1-3)(14-17)+(4-3)(21-17)+(4-3)(16-17)\right]$$

$$=\frac{9}{2}$$

方法 2　首先计算 $\mathbf{X}$ 的样本均值向量和样本协方差阵, 再由结论 1.3.1 和结论 1.3.2 来求 Y,Z 的样本均值、样本方差及样本协方差.

$$\overline{\mathbf{X}}=\frac{1}{3}(\mathbf{X}_1+\mathbf{X}_2+\mathbf{X}_3)=\begin{pmatrix}3\\1\\5\end{pmatrix}$$

$$S=\begin{pmatrix}3 & -\frac{3}{2} & 0\\ -\frac{3}{2} & 1 & \frac{1}{2}\\ 0 & \frac{1}{2} & 1\end{pmatrix}$$

由结论 1.3.1 和结论 1.3.2 知

$$\overline{Y} = (2,2,-1)\begin{pmatrix} 3 \\ 1 \\ 5 \end{pmatrix} = 3$$

$${s_Y}^2 = (2,2,-1)\begin{pmatrix} 3 & -\frac{3}{2} & 0 \\ -\frac{3}{2} & 1 & \frac{1}{2} \\ 0 & \frac{1}{2} & 1 \end{pmatrix}\begin{pmatrix} 2 \\ 2 \\ -1 \end{pmatrix} = 3$$

$$\overline{Z} = (1,-1,3)\begin{pmatrix} 3 \\ 1 \\ 5 \end{pmatrix} = 17$$

$${s_Z}^2 = (1,-1,3)\begin{pmatrix} 3 & -\frac{3}{2} & 0 \\ -\frac{3}{2} & 1 & \frac{1}{2} \\ 0 & \frac{1}{2} & 1 \end{pmatrix}\begin{pmatrix} 1 \\ -1 \\ 3 \end{pmatrix} = 13$$

$${s_{YZ}}^2 = (2,2,-1)\begin{pmatrix} 3 & -\frac{3}{2} & 0 \\ -\frac{3}{2} & 1 & \frac{1}{2} \\ 0 & \frac{1}{2} & 1 \end{pmatrix}\begin{pmatrix} 1 \\ -1 \\ 3 \end{pmatrix} = \frac{9}{2}$$

一般地，我们可以把结论 1.3.1 和结论 1.3.2 推广到 m 个线性组合的情形.

结论 1.3.3 设

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mp} \end{pmatrix}\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} \quad \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$$

$$\triangleq \mathbf{AX}$$

是来自 $\mathbf{X}$ 的样本,则 $\mathbf{Y}$ 的样本均值向量和样本协方差矩阵分别为：$\mathbf{A}\overline{\mathbf{X}}$ 和 $\mathbf{ASA}'$.

第二章　　多元正态分布

把我们熟悉的一元正态分布向多维推广，在多元分析中起着十分重要的作用. 本书中的大多数方法都是基于数据从一个多元正态分布生成的假设. 虽然实际的数据从来不会恰好是多元正态的，然而正态分布常常是“真实的”

总体分布的一种有效的近似.

正态分布的重要性在于它的双重作用,既可作为某些自然现象总体模型,又可作为许多统计量近似的抽样分布.

§2.1 多元正态的概率密度函数及其性质

多元正态分布是一元正态分布向 $p \geq 2$ 维的推广.

若随机变量 X 具有概率密度函数

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\},\ -\infty<x<+\infty$$

则称 X 为具有均值为 μ, 方差为 σ^2 的正态随机变量, 记为 $X\sim N(\mu,\sigma^2)$. 其图形如图 2-1-1.

一、多元正态的概率密度函数

1. 定义：若随机向量 $\mathbf{X}_{p\times 1}=(X_1,X_2,\ldots,X_p)'$ 具有概率密度函数

$$f(\mathbf{X})=\frac{1}{(2\pi)^{\frac{p}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{X}-\mu)'\mathbf{\Sigma}^{-1}(\mathbf{X}-\mu)\right\},\ \mathbf{X}\in R^p$$

其中 $\mu=(\mu_1,\mu_2,\ldots,\mu_p)'$, $\mathbf{\Sigma}$ 是对称正定矩阵, 则称 $\mathbf{X}$ 为 p 维正态随机向量或称 $\mathbf{X}$ 服从 p 元正态分布, 记作 $\mathbf{X}\sim N_p(\mu,\mathbf{\Sigma})$.

例 2.1.1 设 $\mathbf{X}=(X_1,X_2)'$, $E\mathbf{X}=(\mu_1,\mu_2)'\triangleq\mu'$, $Cov\,\mathbf{X}=\begin{pmatrix}\sigma_{11}&\sigma_{12}\\ \sigma_{21}&\sigma_{22}\end{pmatrix}\triangleq\mathbf{\Sigma}$,$(\sigma_{12}=\sigma_{21})$, $\mathbf{X}\sim N_2(\mu,\mathbf{\Sigma})$, 写出 $\mathbf{X}$ 的联合概率密度函数.

解：由定义知：

$f(X_1,X_2)=\frac{1}{2\pi|\mathbf{\Sigma}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{X}-\mu)'\mathbf{\Sigma}^{-1}(\mathbf{X}-\mu)\right\}$,

因为 $\begin{pmatrix}a&b\\ c&d\end{pmatrix}^{-1}=\frac{1}{ad-bc}\begin{pmatrix}d&-b\\ -c&a\end{pmatrix}$

所以 $\mathbf{\Sigma}^{-1}=\frac{1}{\sigma_{11}\sigma_{22}-{\sigma_{12}}^2}\begin{pmatrix}\sigma_{22}&-\sigma_{12}\\ -\sigma_{12}&\sigma_{11}\end{pmatrix}$

又因为

$$\begin{aligned}&(\mathbf{X}-\mu)'\mathbf{\Sigma}^{-1}(\mathbf{X}-\mu)\\ =\ &(X_1-\mu_1,X_2-\mu_2)\frac{1}{\sigma_{11}\sigma_{22}-{\sigma_{12}}^2}\begin{pmatrix}\sigma_{22}&-\sigma_{12}\\ -\sigma_{12}&\sigma_{11}\end{pmatrix}\begin{pmatrix}X_1-\mu_1\\ X_2-\mu_2\end{pmatrix}\\ =\ &\frac{1}{\sigma_{11}\sigma_{22}-{\sigma_{12}}^2}((X_1-\mu_1)\sigma_{22}-(X_2-\mu_2)\sigma_{12},\\ &-\sigma_{12}(X_1-\mu_1)+\sigma_{11}(X_2-\mu_2))\begin{pmatrix}X_1-\mu_1\\ X_2-\mu_2\end{pmatrix}\\ =\ &\frac{1}{{\sigma_1}^2{\sigma_2}^2(1-\rho^2)}\Big[(X_1-\mu_1)^2\sigma_{22}-2(X_1-\mu_1)(X_2-\mu_2)\sigma_{12}\\ &+(X_2-\mu_2)^2\sigma_{11}\Big]\\ =\ &\frac{1}{1-\rho^2}\left[\left(\frac{X_1-\mu_1}{\sigma_1}\right)^2-2\rho\frac{X_1-\mu_1}{\sigma_1}\cdot\frac{X_2-\mu_2}{\sigma_2}+\left(\frac{X_2-\mu_2}{\sigma_2}\right)^2\right]\end{aligned}$$

（其中 $\rho^2=\frac{\sigma_{12}^2}{\sigma_{11}\sigma_{22}}$ ， $\sigma_1^2=\sigma_{11}$ ， $\sigma_2^2=\sigma_{22}$ ）

所以

$$
\begin{aligned}
&f\left(X_1,X_2\right)\\
&=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2\left(1-\rho^2\right)}\cdot\right.\\
&\left.\left[\left(\frac{X_1-\mu_1}{\sigma_1}\right)^2-2\rho\frac{X_1-\mu_1}{\sigma_1}\frac{X_2-\mu_2}{\sigma_2}+\left(\frac{X_2-\mu_2}{\sigma_2}\right)^2\right]\right\}
\end{aligned}
$$

当 $\rho=0$ 时， X_1 与 X_2 不相关. 此时有

$$
\begin{aligned}
&f\left(X_1,X_2\right)=\frac{1}{\sqrt{2\pi}\sigma_1}\exp\left\{-\frac{\left(X_1-\mu_1\right)^2}{2{\sigma_1}^2}\right\}\\
&\cdot\frac{1}{\sqrt{2\pi}\sigma_2}\exp\left\{-\frac{\left(X_2-\mu_2\right)^2}{2{\sigma_2}^2}\right\}\\
&=f_1\left(X_1\right)f_2\left(X_2\right)
\end{aligned}
$$

所以 X_1 与 X_2 相互独立. 即对于二元正态变量来说， X_1 与 X_2 不相关当且仅当 X_1 与 X_2 相互独立. 但对于一般的变量而言，若 X_1 与 X_2 相互独立，则 X_1 与 X_2 不相关，但反之不成立.

例如：设 $X\sim N\left(0,1\right)$, $Y=X^2$. 考虑 X 与 Y 的相关性与独立性.

因为 $Cov\left(X,Y\right)=E\left(X-EX\right)\left(Y-EY\right)=EXY-EX\cdot EY=EX^3-EX\cdot EX^2=0$

所以 X 与 Y 不相关.

显然 X 与 Y 不独立 (从事件的角度考虑).

事实上，由于

$$
\begin{aligned}
&\Pr\left\{X<1,Y<5\right\}=\Pr\left\{X<1\right\}\\
\neq\ &\Pr\left\{X<1\right\}\Pr\left\{Y<5\right\}
\end{aligned}
\quad .
$$

所以， X 与 Y 不独立.

二元正态分布的概率密度曲面图如下：

2. 常数概率密度轮廓线

设 $\mathbf{X}\sim N_p(\mu,\mathbf{\Sigma})$, $f(\mathbf{X})$ 为 $\mathbf{X}$ 的概率密度函数，当 $f(\mathbf{X})=c_1$ 时

$$
\left(\mathbf{X}-\mu\right)'\mathbf{\Sigma}^{-1}\left(\mathbf{X}-\mu\right)=-2\ln\left\{c_1\left(2\pi\right)^{\frac{p}{2}}\left|\mathbf{\Sigma}\right|^{\frac{1}{2}}\right\}\triangleq c^2
$$

称满足条件 $\left(\mathbf{X}-\mu\right)'\mathbf{\Sigma}^{-1}\left(\mathbf{X}-\mu\right)=c^2$ 的所有 $\mathbf{X}$ 组成的集合为概率密度的轮廓线；它是一个以 μ 为中心，且轴为 $\pm c\sqrt{\lambda_i}\mathbf{e}_i$(其中 $\mathbf{\Sigma}\mathbf{e}_i=\lambda_i\mathbf{e}_i$, $i=1,2,\ldots,p$. $\mathbf{e}_i$ 为单位正交化的特征向量) 的椭球面.

例 2.1.2　在例 2.1.1 中，令 $\sigma_{11}=\sigma_{22}$, 求二元正态概率密度的轮廓线.

解：由定义知，轮廓线为： $\left\{\mathbf{X}:\ \left(\mathbf{X}-\mu\right)'\mathbf{\Sigma}^{-1}\left(\mathbf{X}-\mu\right)=c^2\right\}$

下面确定椭圆的长轴与短轴的长度与方向.

首先计算 $\mathbf{\Sigma}=\begin{pmatrix}\sigma_{11}&\sigma_{12}\\\sigma_{21}&\sigma_{22}\end{pmatrix}$ 的特征值.

$$
\text{因为}\left|\mathbf{\Sigma}-\lambda\mathbf{I}\right|=0\iff\begin{vmatrix}\sigma_{11}-\lambda&\sigma_{12}\\\sigma_{21}&\sigma_{11}-\lambda\end{vmatrix}=0
$$

$$\Longleftrightarrow (\sigma_{11}-\lambda)^2-{\sigma_{12}}^2=0 \Longleftrightarrow \lambda_{1,2}=\sigma_{11}\pm\sigma_{12}$$

其次，计算每个特征值 λ_i 对应的单位特征向量.

对 $\lambda_1=\sigma_{11}+\sigma_{12}$ 来说,

$$\text{因为}\quad \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{11} \end{pmatrix} \begin{pmatrix} e_{11} \\ e_{12} \end{pmatrix} = (\sigma_{11}+\sigma_{12}) \begin{pmatrix} e_{11} \\ e_{12} \end{pmatrix}$$

$$\Longleftrightarrow \quad \begin{cases} \sigma_{11}e_{11}+\sigma_{12}e_{12}= & (\sigma_{11}+\sigma_{12})\,e_{11} \\ \sigma_{12}e_{11}+\sigma_{11}e_{12}= & (\sigma_{11}+\sigma_{12})\,e_{12} \end{cases}$$

$$\Longrightarrow \quad e_{11}=e_{12}$$

故 λ_1 对应的单位特征向量为 $\left(\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}}\right)'$.

同理可得 λ_2 对应的单位特征向量为 $\left(\frac{1}{\sqrt{2}},-\frac{1}{\sqrt{2}}\right)'$.

当 $\sigma_{12}>0$ 时， $\lambda_1>\lambda_2$, 长轴的方向为 $\left(\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}}\right)$, 半长轴的长度为 $c\sqrt{\lambda_1}$.

当 $\sigma_{12}<0$ 时， $\lambda_1<\lambda_2$, 长轴的方向为 $\left(\frac{1}{\sqrt{2}},-\frac{1}{\sqrt{2}}\right)$, 半长轴的长度为 $c\sqrt{\lambda_2}$.

轮廓线的图形如下：

二、多元正态分布的性质

结论 2.1.1 若 $\mathbf{X}\sim N_p(\mu,\mathbf{\Sigma})$, 则对于任意 p 维向量 $\mathbf{a}$, 有 $\mathbf{a}'\mathbf{X}\sim N(\mathbf{a}'\mu,\mathbf{a}'\mathbf{\Sigma a})$, 反之，若对于任意的 p 维向量 $\mathbf{a}$, 有 $\mathbf{a}'\mathbf{X}\sim N(\mathbf{a}'\mu,\mathbf{a}'\mathbf{\Sigma a})$, 则 $\mathbf{X}\sim N_p(\mu,\mathbf{\Sigma})$.

因证明中涉及较多的数学知识，故略去.

推论：若 $\mathbf{X}\sim N_p(\mu,\mathbf{\Sigma})$, 则对于任意的 i, $X_i\sim N\left(\mu_i,\sigma_{ii}\right)$ $,i=1,2,\ldots,p$

$$X_i\pm X_j\sim N\left(\mu_i\pm\mu_j,\sigma_{ii}+\sigma_{jj}\pm 2\sigma_{ij}\right)$$

即正态随机向量任何一个分量都是正态随机变量，任何两个分量的和与差也为正态随机变量.

结论 2.1.2 若 $\mathbf{X}\sim N_p(\mu,\mathbf{\Sigma})$, 则 $\mathbf{Y}=\mathbf{A}_{m\times p}\mathbf{X}_{p\times 1}\sim N_m\left(\mathbf{A}\mu,\mathbf{A\Sigma A}'\right)$, 且 $\mathbf{Y}=\mathbf{AX}+\mathbf{d}\sim N_m\left(\mathbf{A}\mu+\mathbf{d},\mathbf{A\Sigma A}'\right)$, 其中 $\mathbf{d}$ 为 m 维常数向量.

证明：因为对于任意的 $\mathbf{a}\in\mathbf{R}^m$

$$\mathbf{a}'\mathbf{Y}=\mathbf{a}'\mathbf{AX}\overset{\triangle}{=}\mathbf{b}'\mathbf{X}$$

而 $\mathbf{X}\sim N_p\left(\mu,\mathbf{\Sigma}\right)$

由结论 2.1.1 的第一部分知 $\mathbf{b}'\mathbf{X}\sim N\left(\mathbf{b}'\mu,\mathbf{b}'\mathbf{\Sigma b}\right)$

即 $\mathbf{a}'\mathbf{Y} \sim N\left(\mathbf{a}'\mathbf{A}\mu, \mathbf{a}'\mathbf{A}\mathbf{\Sigma}\mathbf{A}'\mathbf{a}\right)$

又由 $\mathbf{a}$ 的任意性及结论 2.1.1 的第二部分知,

$$\mathbf{Y} \sim N_m\left(\mathbf{A}\mu, \mathbf{A}\mathbf{\Sigma}\mathbf{A}'\right).$$

推论 1　若 $\mathbf{X} \sim N_p(\mu, \mathbf{\Sigma})$, 则 $\mathbf{Y} = \mathbf{\Sigma}^{-\frac{1}{2}}\left(\mathbf{X} - \mu\right) \sim N_p\left(\mathbf{0}, \mathbf{I}\right)$

补充：若 $\mathbf{A}$ 为 $n \times n$ 阶半正定矩阵，则存在矩阵 $\mathbf{B}$, 使 $\mathbf{A} = \mathbf{B}\mathbf{B}'$; 若 $\mathbf{A}$ 为 $n \times n$ 阶正定矩阵，则存在对称矩阵 $\mathbf{B}$, 使 $\mathbf{A} = \mathbf{B} \cdot \mathbf{B}$.

证明：因为 $\mathbf{X} \sim N_p(\mu, \mathbf{\Sigma})$, 所以 $\mathbf{X} - \mu \sim N_p\left(\mathbf{0}, \mathbf{\Sigma}\right)$

$\mathbf{\Sigma}^{-\frac{1}{2}}(\mathbf{X} - \mu) \sim N_p(\mathbf{0}, \mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{\Sigma}\mathbf{\Sigma}^{-\frac{1}{2}})$ 即 $N_p(\mathbf{0}, \mathbf{I})$

所以 $\mathbf{Y} = \mathbf{\Sigma}^{-\frac{1}{2}}\left(\mathbf{X} - \mu\right) \sim N_p\left(\mathbf{0}, \mathbf{I}\right)$

推论 2　若 $\mathbf{X} \sim N_p(\mu, \mathbf{\Sigma})$, 则 $\left(\mathbf{X} - \mu\right)' \mathbf{\Sigma}^{-1}\left(\mathbf{X} - \mu\right) \sim \chi_p^2$

证明：因为 $\mathbf{Y} = \mathbf{\Sigma}^{-\frac{1}{2}}\left(\mathbf{X} - \mu\right) \sim N_p(\mathbf{0}, \mathbf{I})$

所以对于任意的 i, $Y_i \sim N(0,1)$, 且 $Y_1, Y_2, \ldots, Y_p$ 相互独立

根据 χ^2 分布的定义，有

$$\mathbf{Y}'\mathbf{Y} = \sum_{i=1}^{p} {Y_i}^2 \sim \chi_p^2$$

而 $\mathbf{Y}'\mathbf{Y} = (\mathbf{X} - \mu)'\mathbf{\Sigma}^{-1}(\mathbf{X} - \mu)$ (称之为 $\mathbf{X}$ 到 μ 的广义距离的平方), 所以 $\left(\mathbf{X} - \mu\right)' \mathbf{\Sigma}^{-1}\left(\mathbf{X} - \mu\right) \sim \chi_p^2$.

我们称 $\left\{\mathbf{X} : \left(\mathbf{X} - \mu\right)' \mathbf{\Sigma}^{-1}\left(\mathbf{X} - \mu\right) \leq \chi_p^2(\alpha)\right\}$ 为 $\mathbf{X}$ 的置信度为 $1 - \alpha$ 的置信椭球，其中 $\chi_p^2(\alpha)$ 满足 $\Pr\left\{\chi^2 \geq \chi_p^2(\alpha)\right\} = \alpha$

如图所示：

例 2.1.3　设 $\mathbf{X} \sim N_3(\mu, \mathbf{\Sigma})$, 求 $\begin{pmatrix} X_1 - X_2 \\ X_2 - X_3 \end{pmatrix}$ 的分布.

解：因为 $\begin{pmatrix} X_1 - X_2 \\ X_2 - X_3 \end{pmatrix} = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \triangleq \mathbf{A}\mathbf{X}$ 所以

$$\mathbf{A}\mu = \begin{pmatrix} \mu_1 - \mu_2 \\ \mu_2 - \mu_3 \end{pmatrix} \triangleq \mu^*$$

$$\mathbf{A}\mathbf{\Sigma}\mathbf{A}' = \begin{pmatrix} \sigma_{11} - 2\sigma_{12} + \sigma_{22} & \sigma_{12} + \sigma_{23} - \sigma_{22} - \sigma_{13} \\ \sigma_{12} + \sigma_{23} - \sigma_{22} - \sigma_{13} & \sigma_{22} - 2\sigma_{23} + \sigma_{33} \end{pmatrix} \triangleq \mathbf{\Sigma}^*$$

所以由结论 2.1.2 知, $\begin{pmatrix} X_1 - X_2 \\ X_2 - X_3 \end{pmatrix} \sim N_2\left(\mu^*, \mathbf{\Sigma}^*\right)$

结论 2.1.3　设 $\mathbf{X} \sim N_p(\mu, \mathbf{\Sigma})$, $\mathbf{X} = \begin{pmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{pmatrix} \begin{matrix} m \\ p-m \end{matrix}$,

$$\mu = \begin{pmatrix} \mu^{(1)} \\ \mu^{(2)} \end{pmatrix} \begin{matrix} m \times 1 \\ (p-m) \times 1 \end{matrix}, \quad \mathbf{\Sigma} = \begin{matrix} m & p-m \\ \end{matrix} \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix} \begin{matrix} m \\ p-m \end{matrix}$$

则 $\mathbf{X}^{(1)} \sim N_m\left(\mu^{(1)}, \mathbf{\Sigma}_{11}\right), \mathbf{X}^{(2)} \sim N_{p-m}\left(\mu^{(2)}, \mathbf{\Sigma}_{22}\right)$.

证明：令 $\mathbf{A} = (\mathbf{I} \quad \mathbf{0})$, 则

$$\mathbf{X}^{(1)} = \underset{m \times m \quad m \times (p-m)}{(\ \mathbf{I} \qquad \mathbf{0}\ )} \begin{pmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{pmatrix} \triangleq \mathbf{AX}$$

因为

$$\mathbf{A}\mu = \mu^{(1)}, \ \mathbf{A\Sigma A}' = \mathbf{\Sigma}_{11}$$

所以，由结论 2.1.2 知 $\mathbf{X}^{(1)} \sim N_m\left(\mu^{(1)}, \mathbf{\Sigma}_{11}\right)$.

同理可证： $\mathbf{X}^{(2)} \sim N_{p-m}\left(\mu^{(2)}, \mathbf{\Sigma}_{22}\right)$.

该结论说明多元正态变量的任意个分量所组成的向量仍是多元正态变量.

例 2.1.4 设 $\mathbf{X} \sim N_5(\mu, \mathbf{\Sigma})$, 求 (1) $(X_1, X_2)'$的分布.

(2) $(X_2, X_5)'$的分布.

解： (1) 方法 1 直接应用结论 2.1.3 得

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}\right)$$

方法 2

$$\text{因为} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix} \mathbf{X} \triangleq \mathbf{AX}$$

$$E\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \mathbf{A}\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

$$Cov\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \mathbf{A\Sigma A}' = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$$

所以根据结论 2.1.2 知, $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}\right)$.

方法 3 利用边缘分布与联合分布之间的关系来计算，但对于正态分布通常不使用此方法.

$$f(X_1, X_2) = \int\int\int_{R^3} f(\mathbf{X}) dX_3 dX_4 dX_5$$

(2) 因为 $\begin{pmatrix} X_2 \\ X_5 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \mathbf{X} \triangleq \mathbf{BX}$

$$E\begin{pmatrix} X_2 \\ X_5 \end{pmatrix} = \mathbf{B}\mu = \begin{pmatrix} \mu_2 \\ \mu_5 \end{pmatrix}$$

$$Cov\begin{pmatrix} X_2 \\ X_5 \end{pmatrix} = \mathbf{B\Sigma B}' = \begin{pmatrix} \sigma_{22} & \sigma_{25} \\ \sigma_{52} & \sigma_{55} \end{pmatrix}$$

所以根据结论 2.1.2 知，$\begin{pmatrix} X_2 \\ X_5 \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_2 \\ \mu_5 \end{pmatrix}, \begin{pmatrix} \sigma_{22} & \sigma_{25} \\ \sigma_{52} & \sigma_{55} \end{pmatrix}\right)$

另一种方法是将随机变量 $\mathbf{X}$ 的分量进行重新排列，不妨设重新排列后的向量为 $\mathbf{Y}$，则

$$\mathbf{Y} = \begin{pmatrix} X_2 \\ X_5 \\ X_1 \\ X_3 \\ X_4 \end{pmatrix}$$

根据 $\mathbf{X}$ 与 $\mathbf{Y}$ 之间的关系，可以得到随机向量 $\mathbf{Y}$ 的均值向量和协方差矩阵分别为：

$$\mu^* = \begin{pmatrix} \mu_2 \\ \mu_5 \\ \mu_1 \\ \mu_3 \\ \mu_4 \end{pmatrix}, \quad \mathbf{\Sigma}^* = \begin{pmatrix} \sigma_{22} & \sigma_{25} & \vdots & \sigma_{21} & \sigma_{23} & \sigma_{24} \\ \sigma_{52} & \sigma_{55} & \vdots & \sigma_{51} & \sigma_{53} & \sigma_{54} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \sigma_{12} & \sigma_{15} & \vdots & \sigma_{11} & \sigma_{13} & \sigma_{14} \\ \sigma_{32} & \sigma_{35} & \vdots & \sigma_{31} & \sigma_{33} & \sigma_{34} \\ \sigma_{42} & \sigma_{45} & \vdots & \sigma_{41} & \sigma_{43} & \sigma_{44} \end{pmatrix}$$

则由结论 2.1.3 知，$\begin{pmatrix} X_2 \\ X_5 \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_2 \\ \mu_5 \end{pmatrix}, \begin{pmatrix} \sigma_{22} & \sigma_{25} \\ \sigma_{52} & \sigma_{55} \end{pmatrix}\right)$.

结论 2.1.4　设 $\mathbf{X} = \begin{pmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{pmatrix} \begin{matrix} m \\ (p-m) \end{matrix} \sim N_p(\mu, \mathbf{\Sigma})$，则 $\mathbf{X}^{(1)}$ 与 $\mathbf{X}^{(2)}$ 相互独立 $\iff \mathbf{\Sigma}_{12} = 0$

证明：“$\Longrightarrow$”　因为 $\mathbf{X}^{(1)}$ 与 $\mathbf{X}^{(2)}$ 相互独立，所以 $\mathbf{X}^{(1)}$ 与 $\mathbf{X}^{(2)}$ 互不相关，从而 $\mathbf{\Sigma}_{12} = 0$.

“$\Longleftarrow$”　设 $\mathbf{\Sigma}_{12} = 0$,

$$
\begin{aligned}
f(\mathbf{X}) &= \frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{X}-\mu)'\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\mu)\right\} \\
&= \frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}_{11}|^{\frac{1}{2}}|\boldsymbol{\Sigma}_{22}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\left(\left(\mathbf{X}^{(1)}-\mu^{(1)}\right)',\right.\right. \\
&\quad \left.\left.\left(\mathbf{X}^{(2)}-\mu^{(2)}\right)'\right)\begin{pmatrix}\boldsymbol{\Sigma}_{11}^{-1} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Sigma}_{22}^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{X}^{(1)}-\mu^{(1)}\\ \mathbf{X}^{(2)}-\mu^{(2)}\end{pmatrix}\right\} \\
&= \frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{\Sigma}_{11}|^{\frac{1}{2}}|\boldsymbol{\Sigma}_{22}|^{\frac{1}{2}}}\cdot \\
&\quad \exp\left\{-\frac{1}{2}\left[\left(\mathbf{X}^{(1)}-\mu^{(1)}\right)'{\boldsymbol{\Sigma}_{11}}^{-1}\left(\mathbf{X}^{(1)}-\mu^{(1)}\right)\right.\right. \\
&\quad \left.\left.+\left(\mathbf{X}^{(2)}-\mu^{(2)}\right)'{\boldsymbol{\Sigma}_{22}}^{-1}\left(\mathbf{X}^{(2)}-\mu^{(2)}\right)\right]\right\} \\
&= \frac{1}{(2\pi)^{\frac{m}{2}}|\boldsymbol{\Sigma}_{11}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\left(\mathbf{X}^{(1)}-\mu^{(1)}\right)'{\boldsymbol{\Sigma}_{11}}^{-1}\left(\mathbf{X}^{(1)}-\mu^{(1)}\right)\right\} \\
&\quad \cdot\frac{1}{(2\pi)^{\frac{p-m}{2}}|\boldsymbol{\Sigma}_{22}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\left(\mathbf{X}^{(2)}-\mu^{(2)}\right)'{\boldsymbol{\Sigma}_{22}}^{-1}\left(\mathbf{X}^{(2)}-\mu^{(2)}\right)\right\} \\
&= f_1\left(\mathbf{X}^{(1)}\right)f_2\left(\mathbf{X}^{(2)}\right)
\end{aligned}
$$

根据相互独立的定义知， $\mathbf{X}^{(1)}$ 与 $\mathbf{X}^{(2)}$ 相互独立.

(注：$\begin{pmatrix}\boldsymbol{\Sigma}_{11} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Sigma}_{22}\end{pmatrix}^{-1} = \begin{pmatrix}{\boldsymbol{\Sigma}_{11}}^{-1} & \mathbf{0}\\ \mathbf{0} & {\boldsymbol{\Sigma}_{22}}^{-1}\end{pmatrix}$,
$\left|\begin{array}{cc}\boldsymbol{\Sigma}_{11} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Sigma}_{22}\end{array}\right| = |\boldsymbol{\Sigma}_{11}||\boldsymbol{\Sigma}_{22}|$.)

结论 2.1.5　设 $\mathbf{X}^{(1)} \sim N_m\left(\mu^{(1)}, \boldsymbol{\Sigma}_{11}\right)$, $\mathbf{X}^{(2)} \sim N_k\left(\mu^{(2)}, \boldsymbol{\Sigma}_{22}\right)$, 且 $\mathbf{X}^{(1)}$ 与 $\mathbf{X}^{(2)}$ 相互独立，则

$$
\begin{pmatrix}\mathbf{X}^{(1)}\\ \mathbf{X}^{(2)}\end{pmatrix} \sim N_{m+k}\left(\begin{pmatrix}\mu^{(1)}\\ \mu^{(2)}\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Sigma}_{11} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Sigma}_{22}\end{pmatrix}\right).
$$

证明：因为 $\mathbf{X}^{(1)}$ 与 $\mathbf{X}^{(2)}$ 相互独立，

$$
f_1\left(\mathbf{X}^{(1)}\right) = \frac{1}{(2\pi)^{\frac{m}{2}}|\boldsymbol{\Sigma}_{11}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\left(\mathbf{X}^{(1)}-\mu^{(1)}\right)'{\boldsymbol{\Sigma}_{11}}^{-1}\left(\mathbf{X}^{(1)}-\mu^{(1)}\right)\right\},
$$

$$
f_2\left(\mathbf{X}^{(2)}\right) = \frac{1}{(2\pi)^{\frac{k}{2}}|\boldsymbol{\Sigma}_{22}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\left(\mathbf{X}^{(2)}-\mu^{(2)}\right)'{\boldsymbol{\Sigma}_{22}}^{-1}\left(\mathbf{X}^{(2)}-\mu^{(2)}\right)\right\},
$$

所以，

$$f\left(\mathbf{X}^{(1)},\mathbf{X}^{(2)}\right)=f_1\left(\mathbf{X}^{(1)}\right)f_2\left(\mathbf{X}^{(2)}\right)$$

$$=\frac{1}{(2\pi)^{\frac{m+k}{2}}\left|\mathbf{\Sigma}_{11}\right|^{\frac{1}{2}}\left|\mathbf{\Sigma}_{22}\right|^{\frac{1}{2}}}\cdot$$
$$\exp\left\{-\frac{1}{2}\left[\left(\mathbf{X}^{(1)}-\mu^{(1)}\right)'\mathbf{\Sigma}_{11}{}^{-1}\left(\mathbf{X}^{(1)}-\mu^{(1)}\right)\right.\right.$$

$$\left.\left.+\left(\mathbf{X}^{(2)}-\mu^{(2)}\right)'\mathbf{\Sigma}_{22}{}^{-1}\left(\mathbf{X}^{(2)}-\mu^{(2)}\right)\right]\right\}$$

$$=\frac{1}{(2\pi)^{\frac{m+k}{2}}\left|\mathbf{\Sigma}_{11}\right|^{\frac{1}{2}}\left|\mathbf{\Sigma}_{22}\right|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\left(\left(\mathbf{X}^{(1)}-\mu^{(1)}\right)',\left(\mathbf{X}^{(2)}-\mu^{(2)}\right)'\right)\cdot\right.$$

$$\left.\begin{pmatrix}\mathbf{\Sigma}_{11}{}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{\Sigma}_{22}{}^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{X}^{(1)}-\mu^{(1)}\\ \mathbf{X}^{(2)}-\mu^{(2)}\end{pmatrix}\right\}$$

令 $\mathbf{X}=\begin{pmatrix}\mathbf{X}^{(1)}\\ \mathbf{X}^{(2)}\end{pmatrix}$, $\mu=\begin{pmatrix}\mu^{(1)}\\ \mu^{(2)}\end{pmatrix}$, $\mathbf{\Sigma}=\begin{pmatrix}\mathbf{\Sigma}_{11} & \mathbf{0}\\ \mathbf{0} & \mathbf{\Sigma}_{22}\end{pmatrix}$, 则

$$f\left(\mathbf{X}^{(1)},\mathbf{X}^{(2)}\right)=\frac{1}{(2\pi)^{\frac{m+k}{2}}\left|\mathbf{\Sigma}\right|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\left(\mathbf{X}-\mu\right)'\mathbf{\Sigma}^{-1}\left(\mathbf{X}-\mu\right)\right\}.$$

根据多元正态概率密度函数的定义知, $\mathbf{X}\sim N_{m+k}\left(\mu,\mathbf{\Sigma}\right)$.

推论：设 $\mathbf{X}\sim N_p\left(\mu,\mathbf{\Sigma}\right)$, $\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_n$ 是来自 $\mathbf{X}$ 的样本，则

$$\begin{pmatrix}\mathbf{X}_1\\ \mathbf{X}_2\\ \vdots\\ \mathbf{X}_n\end{pmatrix}\sim N_{np}\left(\begin{pmatrix}\mu\\ \mu\\ \vdots\\ \mu\end{pmatrix}_{np\times 1},\begin{pmatrix}\mathbf{\Sigma} & & & \\ & \mathbf{\Sigma} & & \\ & & \ddots & \\ & & & \mathbf{\Sigma}\end{pmatrix}_{np\times np}\right)$$

进一步，如果 $\mathbf{X}_i\sim N_p\left(\mu_i,\mathbf{\Sigma}_i\right)$, $i=1,2,\ldots,n$, 且相互独立，则

$$\begin{pmatrix}\mathbf{X}_1\\ \mathbf{X}_2\\ \vdots\\ \mathbf{X}_n\end{pmatrix}\sim N_{np}\left(\begin{pmatrix}\mu_1\\ \mu_2\\ \vdots\\ \mu_n\end{pmatrix}_{np\times 1},\begin{pmatrix}\mathbf{\Sigma}_1 & & & \\ & \mathbf{\Sigma}_2 & & \\ & & \ddots & \\ & & & \mathbf{\Sigma}_n\end{pmatrix}_{np\times np}\right)$$

三、正态随机向量的线性组合的性质

设 $\mathbf{X}_i\sim N_p\left(\mu_i,\mathbf{\Sigma}_i\right)$, $i=1,2,\ldots,n$, 且相互独立，下面考虑 $\mathbf{Y}=a_1\mathbf{X}_1+a_2\mathbf{X}_2+\ldots+a_n\mathbf{X}_n$ 的分布，其中 $a_1,a_2,\ldots,a_n$ 是常数.

因为 $\begin{pmatrix}\mathbf{X}_1\\ \mathbf{X}_2\\ \vdots\\ \mathbf{X}_n\end{pmatrix}\sim N_{np}\left(\begin{pmatrix}\mu_1\\ \mu_2\\ \vdots\\ \mu_n\end{pmatrix}_{np\times 1},\begin{pmatrix}\mathbf{\Sigma}_1 & & & \\ & \mathbf{\Sigma}_2 & & \\ & & \ddots & \\ & & & \mathbf{\Sigma}_n\end{pmatrix}_{np\times np}\right)$

而 $\mathbf{Y} = a_1\mathbf{X}_1 + a_2\mathbf{X}_2 + \ldots + a_n\mathbf{X}_n$

$$= (a_1\mathbf{I}, a_2\mathbf{I}, \ldots, a_n\mathbf{I}) \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{pmatrix}_{np\times 1}$$

$$\triangleq \mathbf{AZ} \quad \left(\text{其中}\mathbf{I}\text{为}p\times p\text{阶单位阵}\right)$$

所以

$$E\mathbf{Y} = \mathbf{A}E\mathbf{Z} = \mathbf{A}\begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} = a_1\mu_1 + a_2\mu_2 + \ldots + a_n\mu_n$$

$$Cov\mathbf{Y} = \mathbf{A}Cov\mathbf{Z}\mathbf{A}'$$

$$= (a_1\mathbf{I}, a_2\mathbf{I}, \ldots, a_n\mathbf{I}) \begin{pmatrix} \mathbf{\Sigma}_1 & & & \\ & \mathbf{\Sigma}_2 & & \\ & & \ddots & \\ & & & \mathbf{\Sigma}_n \end{pmatrix} \begin{pmatrix} a_1\mathbf{I} \\ a_2\mathbf{I} \\ \vdots \\ a_n\mathbf{I} \end{pmatrix}$$

$$= a_1^2\mathbf{\Sigma}_1 + a_2^2\mathbf{\Sigma}_2 + \ldots + a_n^2\mathbf{\Sigma}_n$$

故由结论 2.1.2 知,

$$\mathbf{Y} = a_1\mathbf{X}_1 + a_2\mathbf{X}_2 + \ldots + a_n\mathbf{X}_n \sim N_p\left(\sum_{i=1}^{n} a_i\mu_i\,,\, \sum_{i=1}^{n} a_i^2\mathbf{\Sigma}_i\right).$$

综上得:

结论 2.1.6　设 $\mathbf{X}_i \sim N_p\left(\mu_i, \mathbf{\Sigma}_i\right)$, $i = 1, 2, \ldots, n$, 且相互独立, 则

$$\mathbf{Y} = a_1\mathbf{X}_1 + a_2\mathbf{X}_2 + \ldots + a_n\mathbf{X}_n \sim N_p\left(\sum_{i=1}^{n} a_i\mu_i\,,\, \sum_{i=1}^{n} a_i^2\mathbf{\Sigma}_i\right).$$

结论 2.1.7　设 $\mathbf{X}_i \sim N_p\left(\mu_i, \mathbf{\Sigma}_i\right)$, $i = 1, 2, \ldots, n$, 且相互独立, $\mathbf{Y} = \sum\limits_{i=1}^{n} a_i\mathbf{X}_i$, $\mathbf{W} = \sum\limits_{i=1}^{n} b_i\mathbf{X}_i$, 则

$$\begin{pmatrix} \mathbf{Y} \\ \mathbf{W} \end{pmatrix} \sim N_{2p}\left(\begin{pmatrix} \sum\limits_{i=1}^{n} a_i\mu_i \\ \sum\limits_{i=1}^{n} b_i\mu_i \end{pmatrix}, \begin{pmatrix} \sum\limits_{i=1}^{n} a_i^2\mathbf{\Sigma}_i & \sum\limits_{i=1}^{n} a_ib_i\mathbf{\Sigma}_i \\ \sum\limits_{i=1}^{n} a_ib_i\mathbf{\Sigma}_i & \sum\limits_{i=1}^{n} b_i^2\mathbf{\Sigma}_i \end{pmatrix}\right)$$

事实上, 因为

$$\begin{pmatrix} \mathbf{Y} \\ \mathbf{W} \end{pmatrix} = \begin{pmatrix} a_1\mathbf{I} & a_2\mathbf{I} & \cdots & a_n\mathbf{I} \\ b_1\mathbf{I} & b_2\mathbf{I} & \cdots & b_n\mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{pmatrix} \triangleq \mathbf{BZ}$$

$$E\begin{pmatrix} \mathbf{Y} \\ \mathbf{W} \end{pmatrix} = \mathbf{B}E\mathbf{Z} = \mathbf{B}\begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} a_i\mu_i \\ \sum_{i=1}^{n} b_i\mu_i \end{pmatrix}$$

$$\begin{aligned} Cov\begin{pmatrix} \mathbf{Y} \\ \mathbf{W} \end{pmatrix} &= \mathbf{B}\begin{pmatrix} \mathbf{\Sigma}_1 & & & \\ & \mathbf{\Sigma}_2 & & \\ & & \ddots & \\ & & & \mathbf{\Sigma}_n \end{pmatrix}\mathbf{B}' \\ &= \begin{pmatrix} a_1\mathbf{\Sigma}_1 & a_2\mathbf{\Sigma}_2 & \cdots & a_n\mathbf{\Sigma}_n \\ b_1\mathbf{\Sigma}_1 & b_2\mathbf{\Sigma}_2 & \cdots & b_n\mathbf{\Sigma}_n \end{pmatrix} \begin{pmatrix} a_1\mathbf{I} & b_1\mathbf{I} \\ a_2\mathbf{I} & b_2\mathbf{I} \\ \vdots & \vdots \\ a_n\mathbf{I} & b_n\mathbf{I} \end{pmatrix} \\ &= \begin{pmatrix} \sum_{i=1}^{n} a_i^2\mathbf{\Sigma}_i & \sum_{i=1}^{n} a_ib_i\mathbf{\Sigma}_i \\ \sum_{i=1}^{n} a_ib_i\mathbf{\Sigma}_i & \sum_{i=1}^{n} b_i^2\mathbf{\Sigma}_i \end{pmatrix} \end{aligned}$$

例 2.1.5　设 $\mathbf{X} \sim N_3(\mu, \mathbf{\Sigma})$, 其中 $\mu = \begin{pmatrix} 3 \\ -1 \\ 1 \end{pmatrix}$, $\mathbf{\Sigma} = \begin{pmatrix} 3 & -1 & 1 \\ -1 & 1 & 0 \\ 1 & 0 & 2 \end{pmatrix}$, $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4$ 是来自 $\mathbf{X}$ 的样本.

求：　(1)$a'\mathbf{X} = a_1X_1 + a_2X_2 + a_3X_3$ 的分布;

(2) 求 $\frac{1}{2}\mathbf{X}_1 + \frac{1}{2}\mathbf{X}_2 + \frac{1}{2}\mathbf{X}_3 + \frac{1}{2}\mathbf{X}_4$ 和 $\mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3 - 3\mathbf{X}_4$ 的分布以及它们的联合分布.

解：　(1) 因 $\mathbf{X} \sim N_3(\mu, \mathbf{\Sigma})$, 可以得到 $\mathbf{a}'\mathbf{X} \sim N(\mathbf{a}'\mu, \mathbf{a}'\mathbf{\Sigma}\mathbf{a})$

而

$$\mathbf{a}'\mu = 3a_1 - a_2 + a_3$$

$$\begin{aligned} a'\mathbf{\Sigma}\mathbf{a} &= (a_1, a_2, a_3)\begin{pmatrix} 3 & -1 & 1 \\ -1 & 1 & 0 \\ 1 & 0 & 2 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}, \\ &= 3a_1^2 + a_2^2 + 2a_3^2 - 2a_1a_2 + 2a_1a_3 \end{aligned}$$

故 $\mathbf{a}'\mathbf{X} \sim N\left(3a_1 - a_2 + a_3, 3a_1^2 + a_2^2 + 2a_3^2 - 2a_1a_2 + 2a_1a_3\right)$.

(2) 令 $\mathbf{Y}=\frac{1}{2}\mathbf{X}_1+\frac{1}{2}\mathbf{X}_2+\frac{1}{2}\mathbf{X}_3+\frac{1}{2}\mathbf{X}_4=\left(\frac{1}{2}\mathbf{I},\frac{1}{2}\mathbf{I},\frac{1}{2}\mathbf{I},\frac{1}{2}\mathbf{I}\right)\begin{pmatrix}\mathbf{X}_1\\ \mathbf{X}_2\\ \mathbf{X}_3\\ \mathbf{X}_4\end{pmatrix}$

$$\mathbf{W}=\mathbf{X}_1+\mathbf{X}_2+\mathbf{X}_3-3\mathbf{X}_4=(\mathbf{I},\mathbf{I},\mathbf{I},-3\mathbf{I})\begin{pmatrix}\mathbf{X}_1\\ \mathbf{X}_2\\ \mathbf{X}_3\\ \mathbf{X}_4\end{pmatrix}$$

$$\begin{pmatrix}\mathbf{Y}\\ \mathbf{W}\end{pmatrix}=\begin{pmatrix}\frac{1}{2}\mathbf{I} & \frac{1}{2}\mathbf{I} & \frac{1}{2}\mathbf{I} & \frac{1}{2}\mathbf{I}\\ \mathbf{I} & \mathbf{I} & \mathbf{I} & -3\mathbf{I}\end{pmatrix}\begin{pmatrix}\mathbf{X}_1\\ \mathbf{X}_2\\ \mathbf{X}_3\\ \mathbf{X}_4\end{pmatrix}\triangleq\mathbf{B}\begin{pmatrix}\mathbf{X}_1\\ \mathbf{X}_2\\ \mathbf{X}_3\\ \mathbf{X}_4\end{pmatrix}$$

$\left(\text{其中}\mathbf{I}\text{为 3 阶单位矩阵}\right)$

根据结论 2.1.7 知，$\begin{pmatrix}\mathbf{Y}\\ \mathbf{W}\end{pmatrix}$ 服从 6 元正态分布，其均值向量为

$$\begin{aligned}E\begin{pmatrix}\mathbf{Y}\\ \mathbf{W}\end{pmatrix} &= \begin{pmatrix}\frac{1}{2}\mathbf{I} & \frac{1}{2}\mathbf{I} & \frac{1}{2}\mathbf{I} & \frac{1}{2}\mathbf{I}\\ \mathbf{I} & \mathbf{I} & \mathbf{I} & -3\mathbf{I}\end{pmatrix}\begin{pmatrix}\mu\\ \mu\\ \mu\\ \mu\end{pmatrix}=\begin{pmatrix}2\mu\\ \mathbf{0}\end{pmatrix}\\ &= (6,-2,2,0,0,0)'\end{aligned}$$

协方差矩阵为：

$$\begin{aligned}Cov\begin{pmatrix}\mathbf{Y}\\ \mathbf{W}\end{pmatrix} &= \begin{pmatrix}\frac{1}{2}\mathbf{I} & \frac{1}{2}\mathbf{I} & \frac{1}{2}\mathbf{I} & \frac{1}{2}\mathbf{I}\\ \mathbf{I} & \mathbf{I} & \mathbf{I} & -3\mathbf{I}\end{pmatrix}\begin{pmatrix}\mathbf{\Sigma} & & & \\ & \mathbf{\Sigma} & & \\ & & \mathbf{\Sigma} & \\ & & & \mathbf{\Sigma}\end{pmatrix}\begin{pmatrix}\frac{1}{2}\mathbf{I} & \mathbf{I}\\ \frac{1}{2}\mathbf{I} & \mathbf{I}\\ \frac{1}{2}\mathbf{I} & \mathbf{I}\\ \frac{1}{2}\mathbf{I} & -3\mathbf{I}\end{pmatrix}\\ &= \begin{pmatrix}\mathbf{\Sigma} & \mathbf{0}\\ \mathbf{0} & 12\mathbf{\Sigma}\end{pmatrix}\\ &= \begin{pmatrix}3 & -1 & 1 & 0 & 0 & 0\\ -1 & 1 & 0 & 0 & 0 & 0\\ 1 & 0 & 2 & 0 & 0 & 0\\ 0 & 0 & 0 & 36 & -12 & 12\\ 0 & 0 & 0 & -12 & 12 & 0\\ 0 & 0 & 0 & 12 & 0 & 24\end{pmatrix}\end{aligned}$$

根据结论 2.1.3 知，

$$\mathbf{Y}\sim N_3(\mu,\mathbf{\Sigma})\ ,\ \mathbf{W}\sim N_3(\mathbf{0},12\mathbf{\Sigma})$$

因为 $Cov(\mathbf{Y},\mathbf{W})=\mathbf{0}$，所以随机向量 $\mathbf{Y}$ 与 $\mathbf{W}$ 互不相关.

§2.2 参数的极大似然估计

1. 极大似然估计的定义

设 $\mathbf{X}\sim f(\mathbf{X},\theta)$, θ 为未知参数, $\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_n$ 是来自总体 $\mathbf{X}$ 的样本. 似然函数为 $L(\theta)=\prod\limits_{i=1}^{n}f(\mathbf{X}_i,\theta)$.

若存在 $\theta^*\in\Theta$, 使 $L(\theta^*)=\sup_{\theta\in\Theta}L(\theta)$, 其中 Θ 表示参数 θ 的取值范围, 则称 θ^* 为参数 θ 的极大似然估计 (Maximum Likelihood Estimator), 简记为 MLE.

2. 一元正态分布参数的极大似然估计

设 $X\sim N(\mu,\sigma^2)$, $X_1,X_2,\ldots,X_n$ 是来自 X 的样本, 则似然函数

$L(\mu,\sigma^2)=(\frac{1}{\sqrt{2\pi}\sigma})^n\exp\left\{-\sum\limits_{i=1}^{n}\frac{(X_i-\mu)^2}{2\sigma^2}\right\}$

$\ln L(\mu,\sigma^2)=-\frac{n}{2}\ln 2\pi--\frac{n}{2}\ln\sigma^2-\frac{1}{2\sigma^2}\sum\limits_{i=1}^{n}(X_i-\mu)^2$, 称之为对数似然函数.

$$\begin{cases}\frac{\partial\ln L(\mu,\sigma^2)}{\partial\mu}=\frac{1}{\sigma^2}\sum\limits_{i=1}^{n}(X_i-\mu)\overset{令}{=\!=}0\\ \frac{\partial\ln L(\mu,\sigma^2)}{\partial\sigma^2}=-\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum\limits_{i=1}^{n}(X_i-\mu)^2\overset{令}{=\!=}0\end{cases}$$, 称之为似然方程组.

解之得 $\hat{\mu}=\overline{X}$, $\hat{\sigma^2}=\frac{1}{n}\sum\limits_{i=1}^{n}\left(X_i-\overline{X}\right)^2=\frac{n-1}{n}s^2$

易验证 $\hat{\mu}$ 与 $\hat{\sigma^2}$ 使 $L(\mu,\sigma^2)$ 达到最大, 故参数 μ, σ^2 的最大似然估计为:

$$\hat{\mu}=\overline{X},\ \hat{\sigma^2}=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2=\frac{n-1}{n}s^2.$$

3. 多元正态分布参数的极大似然估计

为了更方便地求解参数的极大似然估计, 先引入如下引理.

结论 2.2.1 设 $\mathbf{B}$ 为 $p\times p$ 阶对称正定矩阵, 常数 $b>0$, 则对任意正定矩阵 $\mathbf{\Sigma}$, 有

$$\frac{1}{|\mathbf{\Sigma}|^b}e^{-tr(\mathbf{\Sigma}^{-1}\mathbf{B})/2}\leq\frac{1}{|\mathbf{B}|^b}(2b)^{bp}e^{-bp}$$

仅当 $\mathbf{\Sigma}=\frac{1}{2b}\mathbf{B}$ 时, 等号成立.

结论 2.2.1 设 $\mathbf{X}\sim N_p(\mu,\mathbf{\Sigma})$, $\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_n$ 是来自 $\mathbf{X}$ 的样本, 则

$$\hat{\mu}=\overline{\mathbf{X}},\ \hat{\mathbf{\Sigma}}=\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{X}_i-\overline{\mathbf{X}}\right)\left(\mathbf{X}_i-\overline{\mathbf{X}}\right)'=\frac{n-1}{n}\mathbf{S}$$

分别是 μ 和 $\mathbf{\Sigma}$ 的极大似然估计量, 其观测值称为 μ 和 $\mathbf{\Sigma}$ 的极大似然估计值.

证明: 因为似然函数

$$L(\mu,\mathbf{\Sigma})=\frac{1}{(2\pi)^{\frac{np}{2}}|\mathbf{\Sigma}|^{\frac{n}{2}}}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{X}_i-\mu)'\mathbf{\Sigma}^{-1}(\mathbf{X}_i-\mu)\right\}$$

下面求参数 $\hat{\mu}$, $\hat{\mathbf{\Sigma}}$ 使 $L(\mu, \mathbf{\Sigma})$ 达到最大值. 由于

$$\sum_{i=1}^{n} (\mathbf{X}_i - \mu)' \mathbf{\Sigma}^{-1} (\mathbf{X}_i - \mu)$$

$$= \sum_{i=1}^{n} \left(\mathbf{X}_i - \overline{\mathbf{X}} + \overline{\mathbf{X}} - \mu\right)' \mathbf{\Sigma}^{-1} \left(\mathbf{X}_i - \overline{\mathbf{X}} + \overline{\mathbf{X}} - \mu\right)$$

$$= \sum_{i=1}^{n} \left(\mathbf{X}_i - \overline{\mathbf{X}}\right)' \mathbf{\Sigma}^{-1} \left(\mathbf{X}_i - \overline{\mathbf{X}}\right) + \sum_{i=1}^{n} \left(\mathbf{X}_i - \overline{\mathbf{X}}\right)' \mathbf{\Sigma}^{-1} \left(\overline{\mathbf{X}} - \mu\right)$$

$$\sum_{i=1}^{n} \left(\overline{\mathbf{X}} - \mu\right)' \mathbf{\Sigma}^{-1} \left(\mathbf{X}_i - \overline{\mathbf{X}}\right) + \sum_{i=1}^{n} \left(\overline{\mathbf{X}} - \mu\right)' \mathbf{\Sigma}^{-1} \left(\overline{\mathbf{X}} - \mu\right)$$

$$= \sum_{i=1}^{n} \left(\mathbf{X}_i - \overline{\mathbf{X}}\right)' \mathbf{\Sigma}^{-1} \left(\mathbf{X}_i - \overline{\mathbf{X}}\right) + n \left(\overline{\mathbf{X}} - \mu\right)' \mathbf{\Sigma}^{-1} \left(\overline{\mathbf{X}} - \mu\right)$$

$$= \sum_{i=1}^{n} tr \left(\mathbf{X}_i - \overline{\mathbf{X}}\right)' \mathbf{\Sigma}^{-1} \left(\mathbf{X}_i - \overline{\mathbf{X}}\right) + n \left(\overline{\mathbf{X}} - \mu\right)' \mathbf{\Sigma}^{-1} \left(\overline{\mathbf{X}} - \mu\right)$$

$$= tr\mathbf{\Sigma}^{-1} \left[\sum_{i=1}^{n} \left(\mathbf{X}_i - \overline{\mathbf{X}}\right) \left(\mathbf{X}_i - \overline{\mathbf{X}}\right)'\right] + n \left(\overline{\mathbf{X}} - \mu\right)' \mathbf{\Sigma}^{-1} \left(\overline{\mathbf{X}} - \mu\right)$$

$$= (n-1) tr\mathbf{\Sigma}^{-1}\mathbf{S} + n \left(\overline{\mathbf{X}} - \mu\right)' \mathbf{\Sigma}^{-1} \left(\overline{\mathbf{X}} - \mu\right)$$

所以

$$L(\mu, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{\frac{np}{2}} |\mathbf{\Sigma}|^{\frac{n}{2}}} \exp\left\{-\frac{1}{2}\left[(n-1)\, tr\mathbf{\Sigma}^{-1}\mathbf{S} + n \left(\overline{\mathbf{X}} - \mu\right)' \mathbf{\Sigma}^{-1} \left(\overline{\mathbf{X}} - \mu\right)\right]\right\} \quad (*)$$

又因为 $\left(\overline{\mathbf{X}} - \mu\right)' \mathbf{\Sigma}^{-1} \left(\overline{\mathbf{X}} - \mu\right) \geq 0$ 所以, 要使 $L(\mu, \mathbf{\Sigma})$ 达到最大, 只要 $\left(\overline{\mathbf{X}} - \mu\right)' \mathbf{\Sigma}^{-1} \left(\overline{\mathbf{X}} - \mu\right)$ 达到最小. 即取 $\hat{\mu} = \overline{\mathbf{X}}$, 将 $\hat{\mu} = \overline{\mathbf{X}}$ 代入 $L(\mu, \mathbf{\Sigma})$ 得

$$L(\hat{\mu}, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{\frac{np}{2}} |\mathbf{\Sigma}|^{\frac{n}{2}}} \exp\left\{-\frac{1}{2} tr\left(\mathbf{\Sigma}^{-1} (n-1)\, \mathbf{S}\right)\right\}$$

根据结论 2.2.1 有

$$\begin{aligned} L(\hat{\mu}, \mathbf{\Sigma}) &\leq \frac{1}{(2\pi)^{\frac{np}{2}} |(n-1)\mathbf{S}|^{\frac{n}{2}}} n^{\frac{np}{2}} e^{-\frac{np}{2}} \\ &= \frac{1}{(2\pi)^{\frac{np}{2}} |\frac{n-1}{n}\mathbf{S}|^{\frac{n}{2}}} e^{-\frac{np}{2}} \end{aligned}$$

当 $\hat{\mathbf{\Sigma}} = \frac{n-1}{n}\mathbf{S}$ 时, $L(\hat{\mu}, \hat{\mathbf{\Sigma}})$ 达到最大.

故 $\mu, \mathbf{\Sigma}$ 的极大似然估计为

$$\hat{\mu} = \overline{\mathbf{X}}\,, \hat{\mathbf{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{X}_i - \overline{\mathbf{X}}\right) \left(\mathbf{X}_i - \overline{\mathbf{X}}\right)' = \frac{n-1}{n}\mathbf{S}$$

4. 充分统计量

由 $(*)$ 式知， $L(\mu,\mathbf{\Sigma})=f(\mu,\mathbf{\Sigma},\overline{\mathbf{X}},\mathbf{S})$, 即似然函数仅通过样本均值向量 $\overline{\mathbf{X}}$ 与协方差阵依赖于 $\mathbf{X}_1,\ldots,\mathbf{X}_n$. 我们称 $\overline{\mathbf{X}}$ 和 $\mathbf{S}$ 是参数 $\theta=(\mu,\mathbf{\Sigma})$ 的充分统计量. 它意味着，不管样本容量 n 有多大，数据矩阵 $\mathcal{X}$ 中关于 μ 和 $\mathbf{\Sigma}$ 的全部信息都包含在 $\overline{\mathbf{X}}$ 和 $\mathbf{S}$ 中. 但对于非正态总体，一般不成立. 故若不能把数据看作服从多元正态分布，单独依赖于 $\overline{\mathbf{X}}$ 和 $\mathbf{S}$ 的方法可能会忽视其他有用的信息. 因此，要慎重检查多元正态假设的合理性.

5. 极大似然估计的性质

若 $\hat{\mu}$ 与 $\hat{\mathbf{\Sigma}}$ 分别是 μ 与 $\mathbf{\Sigma}$ 的极大似然估计， $g(\mu,\mathbf{\Sigma})$ 是任一单调函数，则 $g(\hat{\mu},\hat{\mathbf{\Sigma}})$ 是 $g(\mu,\mathbf{\Sigma})$ 的极大似然估计.

§2.3 多元变量的抽样分布

假定 $\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_n$ 是来自均值为 μ, 协方差阵为 $\mathbf{\Sigma}$ 的正态总体的样本. 因为样本的信息全部集中在样本均值向量 $\overline{\mathbf{X}}$ 和样本协方差矩阵 $\mathbf{S}$ 中，所以通过研究 $\overline{\mathbf{X}}$ 和 $\mathbf{S}$ 的分布可以推断总体的一些性质，我们把 $\overline{\mathbf{X}}$ 和 $\mathbf{S}$ 所服从的分布称为抽样分布.

我们首先介绍与 χ^2 分布具有类似性质的威沙特分布.

定义. 设 $\mathbf{X}_j\sim N_p(\mathbf{0},\mathbf{\Sigma})\,,\ j=1,2,\ldots,m,$ 且相互独立，则称随机矩阵 $\mathbf{W}=\sum\limits_{j=1}^{m}\mathbf{X}_j\mathbf{X}_j'$ 所服从的分布为自由度是 m 的威沙特分布. 记作 $\mathbf{W}\sim\mathbf{W}_m(\mathbf{\Sigma})$, 当 $\mathbf{\Sigma}=\mathbf{I}$ 时，称为标准威沙特分布.

性质 1 若 $\mathbf{W}_1\sim\mathbf{W}_m(\mathbf{\Sigma})\,,\ \mathbf{W}_2\sim\mathbf{W}_n(\mathbf{\Sigma})$, 且 $\mathbf{W}_1$ 与 $\mathbf{W}_2$ 相互独立，则

$$\mathbf{W}_1+\mathbf{W}_2\sim\mathbf{W}_{m+n}(\mathbf{\Sigma})$$

性质 2 若 $\mathbf{W}\sim\mathbf{W}_m(\mathbf{\Sigma})$, 则 $\mathbf{CWC}'\sim\mathbf{W}_m(\mathbf{C\Sigma C}')$

自由度为 m 威沙特分布的概率密度函数为

$$f_m(\mathbf{W})=\frac{|\mathbf{W}|^{(m-p-1)/2}\exp\left\{-\frac{1}{2}tr\left(\mathbf{W\Sigma}^{-1}\right)\right\}}{2^{\frac{mp}{2}}\pi^{\frac{p(p-1)}{4}}|\mathbf{\Sigma}|^{\frac{m}{2}}\prod\limits_{i=1}^{p}\Gamma\left(\frac{1}{2}\left(m-i+1\right)\right)}$$

其中 $\Gamma(\cdot)$ 为伽马函数， $\mathbf{W}$ 为任一正定矩阵.

下面介绍 $\overline{\mathbf{X}}$ 和 $\mathbf{S}$ 的抽样分布.

一、 $\overline{\mathbf{X}}$ 和 $\mathbf{S}$ 的抽样分布

结论 2.3.1 设 $\mathbf{X}\sim N_p(\mu,\mathbf{\Sigma})\,,\ \mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_n$ 是来自总体 $\mathbf{X}$ 的样本,

$$\overline{\mathbf{X}}=\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\,,\ \mathbf{S}=\frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{X}_i-\overline{\mathbf{X}}\right)\left(\mathbf{X}_i-\overline{\mathbf{X}}\right)'$$

则 $(1)\overline{\mathbf{X}}\sim N_p(\mu,\frac{1}{n}\mathbf{\Sigma})$

$(2)(n-1)\mathbf{S}\sim\mathbf{W}_{n-1}\left(\mathbf{\Sigma}\right)$

$(3)\overline{\mathbf{X}}$ 和 $\mathbf{S}$ 是相互独立的.

证明： (1) 在结论 2.1.6 中，令 $\mu_1 = \mu_2 = \cdots = \mu_n$, $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \cdots = \mathbf{\Sigma}_n$

$$a_1 = a_2 = \cdots = a_n = \frac{1}{n}$$

则有 $\overline{\mathbf{X}} \sim N_p(\mu, \frac{1}{n}\mathbf{\Sigma})$

(2)因为 $(n-1)\mathbf{S} = \sum_{i=1}^{n} \left(\mathbf{X}_i - \overline{\mathbf{X}}\right)\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)'$

$$= \sum_{i=1}^{n} \left(\mathbf{X}_i - \mu\right)\left(\mathbf{X}_i - \mu\right)' - n\left(\overline{\mathbf{X}} - \mu\right)\left(\overline{\mathbf{X}} - \mu\right)'$$

而 $\mathbf{X}_i - \mu \sim N_p(\mathbf{0}, \mathbf{\Sigma})$, $i = 1, 2, \ldots, n$, 且相互独立,
所以 $\sum_{i=1}^{n} \left(\mathbf{X}_i - \mu\right)\left(\mathbf{X}_i - \mu\right)' \sim \mathbf{W}_n(\mathbf{\Sigma})$
又因为 $\overline{\mathbf{X}} - \mu \sim N_p(\mathbf{0}, \frac{1}{n}\mathbf{\Sigma}) \Longrightarrow \sqrt{n}\left(\overline{\mathbf{X}} - \mu\right) \sim N_p(\mathbf{0}, \mathbf{\Sigma})$
所以 $n\left(\overline{\mathbf{X}} - \mu\right)\left(\overline{\mathbf{X}} - \mu\right)' \sim \mathbf{W}_1(\mathbf{\Sigma})$
又因为 $\sum_{i=1}^{n} \left(\mathbf{X}_i - \mu\right)\left(\mathbf{X}_i - \mu\right)' = (n-1)\mathbf{S} + n\left(\overline{\mathbf{X}} - \mu\right)\left(\overline{\mathbf{X}} - \mu\right)'$, 且 $\overline{\mathbf{X}}$ 和 $\mathbf{S}$ 相互独立,
所以 $(n-1)\mathbf{S} \sim \mathbf{W}_{n-1}\left(\mathbf{\Sigma}\right)$.
(3) 略.

二、 $\overline{\mathbf{X}}$ 和 $\mathbf{S}$ 的大样本特性

在一元中，无论总体的分布类型如何，由中心极限定理知，只要样本容量 n 充分大，样本均值近似服从正态分布．这个结论对多元变量亦成立.

结论 2.3.2 (中心极限定理) 设 $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ 是来自任何有均值 μ 与协方差矩阵 $\mathbf{\Sigma}$ 的总体的独立观测结果，则对大样本容量有

$$\sqrt{n}\left(\overline{\mathbf{X}} - \mu\right) \sim N_p\left(\mathbf{0}, \mathbf{\Sigma}\right)$$

又因为当 n 充分大时， $\mathbf{S}$ 依概率收敛到 $\mathbf{\Sigma}$, 从而

$$n\left(\overline{\mathbf{X}} - \mu\right)' \mathbf{S}^{-1}\left(\overline{\mathbf{X}} - \mu\right) \sim \chi_p^2$$

§2.4 多元变量的正态性检验

正如我们所指出的，以后各章讨论的大多数统计方法都假定每个观测向量 $\mathbf{X}_j$ 来自多元正态分布．因此，当数据显示出与多元正态的中度至极端的背离时，必须要有能发现这种情况的方法存在.

根据正态分布的性质，多元正态分布的边缘分布是正态分布，且多元正态密度的轮廓线是椭球面，因此，我们提出下面几个问题：

1. $\mathbf{X}$ 的每个分量的边缘分布是否是正态？分量 X_i 的几个线性组合是否是正态？
2. 根据各种特征的观测结果所作出的散布图，是否呈现出正态总体期望的椭圆形状？

3. 是否存在应该进行检验以确保精确度的“杂乱”观测值？

一、一元正态性的检验

n 较小时的点图和 $n > 25$ 时的直方图有助于揭示一元分布的一个尾部比另一个长得多的情况. 如果关于一个变量 X_j 的直方图似乎较为对称，我们可以通过计算在某个区间的观测值的数目做进一步检验.

1. 因为若$X \sim N(\mu, \sigma^2)$, 则

$$\Pr\{\mu - \sigma < X < \mu + \sigma\} = 0.683$$
$$\Pr\{\mu - 2\sigma < X < \mu + 2\sigma\} = 0.954$$
$$\Pr\{\mu - 3\sigma < X < \mu + 3\sigma\} = 0.997$$

故当 n 充分大时，我们期望观测到位于区间

$$\left(\overline{X} - s, \overline{X} + s\right),\ \left(\overline{X} - 2s, \overline{X} + 2s\right),\ \left(\overline{X} - 3s, \overline{X} + 3s\right)$$

内观测结果的比例大约分别是 68.3%,95.4% 和 99.7%. 如果比例太少，建议采用比正态粗尾的总体分布，如 $\alpha < 2$ 时广义误差分布，或稳定分布族 (分形分布), 其中 α 越小，尾部越厚.

2. Q-Q 图

对每个分量的样本观测值作 Q-Q 图，若边缘分布不是正态分布，则联合分布不是正态的，但反之未必成立.

例：随机向量 (x, y) 的联合概率密度函数为

$$f(x,y) = \frac{1}{2\pi} e^{-\frac{1}{2}\left(x^2+y^2\right)} \left[1 - \frac{xy}{\left(1+x^2\right)\left(1+y^2\right)}\right]$$

显然不是二元正态分布，但

$$f_x(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},\ f_y(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}$$

均为标准正态分布.

设 $X_1, X_2, \ldots, X_n$ 是随机变量 X 的 n 个观测值，由小到大排序后得

$$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$$

根据经验分布函数

$$F_n(X) = \begin{cases} 0, & X < X_{(1)} \\ \frac{i}{n}, & X_{(i)} \le X < X_{(i+1)} \quad i = 1, 2, \ldots, n-1 \\ 1, & X_{(n)} \le X \end{cases}$$

若 $X \sim N(\mu, \sigma^2)$, 则应有

$$\Pr\left\{X \le X_{(i)}\right\} = \Pr\left\{\frac{X - \mu}{\sigma} \le \frac{X_{(i)} - \mu}{\sigma}\right\} = \frac{i}{n},\ i = 1, 2, \ldots, n$$

又因有 $\frac{X-\mu}{\sigma} \sim N(0, 1)$, 则

由 $\int_{-\infty}^{q(i)} \frac{1}{\sqrt{2\pi}} e^{-\frac{X^2}{2}} dx = \frac{i-\frac{1}{2}}{n}$ *or* $\frac{i}{n+1}$ 确定分位数 $q(i)$(为避免出现 1 的情况，将 $\frac{i}{n}$ 修正为 $\frac{i-\frac{1}{2}}{n}$ *or* $\frac{i}{n+1}$)

从而 $q(i) = \frac{X_{(i)}-\mu}{\sigma}$　即 $q(i)$ 与 $X_{(i)}$ 具有线性相关关系.

作 Q-Q 图的步骤：

(a) 将样本观测值排序　$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$

(b) 计算样本分位数对应的概率　$\frac{1-\frac{1}{2}}{n}, \frac{2-\frac{1}{2}}{n}, \ldots, \frac{n-\frac{1}{2}}{n}$

(c) 计算标准正态分布的分位数 $q(i)$,$q_{(i)}$ 满足

$$\Phi\left(q_{(i)}\right) = \frac{i-\frac{1}{2}}{n} \quad , i = 1, 2, \ldots, n$$

(d) 将数对 $\left(X_{(i)}, q_{(i)}\right)$,$i = 1, 2, \ldots, n$ 画在直角坐标系中，若呈直线状，则认为是正态的；否则，认为是非正态数据.

判断方法 1: 相关系数法

$$r_\theta = \frac{\sum_{i=1}^{n} \left(X_{(i)} - \overline{X}\right)\left(q_{(i)} - \overline{q}\right)}{\sqrt{\sum_{i=1}^{n} \left(X_{(i)} - \overline{X}\right)^2} \sqrt{\sum_{i=1}^{n} \left(q_{(i)} - \overline{q}\right)^2}}$$

对任意的 α, 若 $r_\theta < r_\alpha$, 则拒绝正态性假设，即认为数据不是来自正态总体.

判断方法 2: 一元线性回归方法

二、二元正态分布的检验

方法 1　轮廓线

如果观测值是从多元正态分布生成的，则每个二元分量是正态随机变量，其常数密度轮廓线应是椭圆；散布图显示一个近乎椭圆的形状，从而与这个结构一致.

当 $p = 2$ 时，　$(\mathbf{X} - \mu)' \mathbf{\Sigma}^{-1} (\mathbf{X} - \mu) \sim \chi_2^2$

对任意的 $\alpha > 0$; $\Pr\left\{(\mathbf{X} - \mu)' \mathbf{\Sigma}^{-1} (\mathbf{X} - \mu) \leq \chi_2^2(\alpha)\right\} = 1 - \alpha$

又由于 $\overline{\mathbf{X}}$ 和 $\mathbf{S}$ 分别是 μ 和 $\mathbf{\Sigma}$ 的无偏估计，故代替后得：

$$\Pr\left\{\left(\mathbf{X} - \overline{\mathbf{X}}\right)' \mathbf{S}^{-1} \left(\mathbf{X} - \overline{\mathbf{X}}\right) \leq \chi_2^2(\alpha)\right\} = 1 - \alpha$$

故有理由希望对位于该椭圆内的样本观测值有相同的百分比，否则，正态性的假设就是可疑的.

例 4.1.2

方法 2　卡方图 (对 $p \geq 2$ 均成立)

在判断一个数据集的联合正态性时，一种更正式的方法是基于广义平方距离.

$$d_j^2 = \left(\mathbf{X}_j - \overline{\mathbf{X}}\right)' \mathbf{S}^{-1} \left(\mathbf{X}_j - \overline{\mathbf{X}}\right), \ j = 1, 2, \ldots, n$$

其中 $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ 是样本观测值.

当总体是多元正态的，且 n 与 $n-p$ 都很大时，

$$d^2 = \left(\mathbf{X} - \overline{\mathbf{X}}\right)' \mathbf{S}^{-1} \left(\mathbf{X} - \overline{\mathbf{X}}\right) \sim \chi_p^2$$

构造卡方图的方法

(a) 计算 d_i^2, $i = 1, 2, \ldots, n$, 并排序得 $d_{(1)}^2 \leq d_{(2)}^2 \leq \cdots \leq d_{(n)}^2$

(b) 确定分位数点 $q_{(i)}$, 使

$$\int_0^{q_{(i)}} \chi_p^2(x)\, dx = \frac{i - \frac{1}{2}}{n},\ i = 1, 2, \ldots, n$$

(c) 将数对 $\left(q_{(i)},\ d_{(i)}^2\right)$, $i = 1, 2, \ldots, n$, 画在直角坐标系内.

(d) 若图形是一条通过原点且斜率为 1 的直线，则认为总体数据是正态数据. 一个系统的弯曲图形表明缺乏正态性，一个或两个远离直线的点表示大的距离，或是离群的观测值，这些都值得进一步探讨.

例 4.13

§2.5 多元变量的正态性检验

如果数据不是来自正态总体，则许多统计方法就不能直接使用，为此，我们考虑通过数据变换，使非正态数据变成更接近正态的数据. 在适当的数据变换后，就可以实现正态理论分析.

有一种稍作修改的幂变换称之为 Box – Cox 变换，可以改变数据的正态性，我们首先把注意力集中在一元的情况.

一、一元数据的正态性变换

令 $X^{(\lambda)} = \begin{cases} \frac{X^\lambda - 1}{\lambda}, & \lambda \neq 0; \\ \ln X, & \lambda = 0. \end{cases}$

给定观测值 $X_1, X_2, \ldots, X_n$, 选择 λ, 使得似然函数达到最大. 假设选择的 λ 使 $X_i^{(\lambda)} \sim N(\mu, \sigma^2)$, $i = 1, 2, \ldots, n$.

因为 $\frac{dX^{(\lambda)}}{dX} = X^{\lambda - 1}$

所以变换后的似然函数为

$$L(\lambda, \mu, \sigma^2) = \left(\tfrac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{-\tfrac{1}{2\sigma^2} \sum_{i=1}^n \left(X_i^{(\lambda)} - \mu\right)^2\right\} \cdot \prod_{i=1}^n X_i^{\lambda - 1}$$

$$\ln L(\lambda, \mu, \sigma^2) = -\tfrac{n}{2} \ln\left(2\pi\sigma^2\right) - \tfrac{1}{2\sigma^2} \sum_{i=1}^n \left(X_i^{(\lambda)} - \mu\right)^2 + \ln \prod_{i=1}^n X_i^{\lambda - 1}$$

$$\begin{cases} \frac{\partial \ln L(\lambda, \mu, \sigma^2)}{\partial \mu} = 0 \\ \frac{\partial \ln L(\lambda, \mu, \sigma^2)}{\partial \sigma^2} = 0 \end{cases} \quad \text{解之得:} \quad \begin{aligned} \hat{\mu} &= \overline{\mathbf{X}}^{(\lambda)} = \tfrac{1}{n} \sum_{i=1}^n X_i^{(\lambda)} \\ \hat{\sigma^2} &= \tfrac{1}{n} \sum_{i=1}^n \left(X_i^{(\lambda)} - \overline{X^{(\lambda)}}\right)^2 \end{aligned}$$

此时

$$l(\lambda)=\ln L(\lambda,\hat{\mu},\hat{\sigma^2})=-\tfrac{n}{2}\ln 2\pi-\tfrac{n}{2}\ln\tfrac{1}{n}\sum_{i=1}^{n}\left(X_i^{(\lambda)}-\overline{X^{(\lambda)}}\right)^2$$

$$-\tfrac{n}{2}+\ln\left(\prod_{i=1}^{n}X_i\right)^{\lambda-1}$$

$$=-\tfrac{n}{2}\ln 2\pi-\tfrac{n}{2}-\tfrac{n}{2}\left[\ln\tfrac{1}{n}\sum_{i=1}^{n}\left(X_i^{(\lambda)}-\overline{X^{(\lambda)}}\right)^2-\tfrac{2}{n}\ln\left(\prod_{i=1}^{n}X_i\right)^{\lambda-1}\right]$$

$$=-\tfrac{n}{2}\ln 2\pi-\tfrac{n}{2}-\tfrac{n}{2}\ln\tfrac{1}{n}\frac{\sum_{i=1}^{n}\left(X_i^{(\lambda)}-\overline{X^{(\lambda)}}\right)^2}{\left(\prod_{i=1}^{n}X_i\right)^{\frac{2(\lambda-1)}{n}}}$$

$$=-\tfrac{n}{2}\ln 2\pi-\tfrac{n}{2}-\tfrac{n}{2}\ln\tfrac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i^{(\lambda)}}{\left(\prod_{i=1}^{n}X_i\right)^{\frac{\lambda-1}{n}}}-\frac{\overline{X_{(\lambda)}}}{\left(\prod_{i=1}^{n}X_i\right)^{\frac{\lambda-1}{n}}}\right)^2$$

要使 $l(\lambda)=\ln L(\lambda,\mu,\sigma^2)$ 达到最大，由上式知，只要

$$\tfrac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i^{(\lambda)}}{\left(\prod_{i=1}^{n}X_i\right)^{\frac{\lambda-1}{n}}}-\frac{\overline{X_{(\lambda)}}}{\left(\prod_{i=1}^{n}X_i\right)^{\frac{\lambda-1}{n}}}\right)^2$$

$$\triangleq\tfrac{1}{n}\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2\ \text{达到最小}$$

其中 $Y_i=\dfrac{X_i^{\lambda}-1}{\lambda\left(\prod_{i=1}^{n}X_i\right)^{\frac{\lambda-1}{n}}}$，$i=1,2,\ldots,n$.

即求使 $S_\lambda^2=\tfrac{1}{n}\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2$ 达到最小的 λ 值，具体做法如下：

1. 给定一系列的 λ 值,由样本观测值 $X_1,X_2,\ldots,X_n$, 计算 Y_i，$i=1,2,\ldots,n$;
2. 计算 $S_\lambda^2=\tfrac{1}{n}\sum_{i=1}^{n}\left(Y_i-\overline{\mathbf{Y}}\right)^2$;
3. 以 $\left(\lambda,S_\lambda^2\right)$ 为数对作图;
4. 从图像上找到使 S_λ^2 达到最小的 $\hat{\lambda}$.

例 4.16

二、多元数据的正态性变换

对于多元观测值，必须对每个变量选择一个幂变换．设 $\lambda_1,\lambda_2,\ldots,\lambda_p$ 是对于 p 个测量特征的幂变换，每个 λ_k 通过使 $S_{\lambda_k}^2=\tfrac{1}{n}\sum_{i=1}^{n}\left(Y_i^{(\lambda_k)}-\overline{Y^{(\lambda_k)}}\right)^2$ 达到最小而得到的.

其中 $Y_i^{(\lambda_k)}=\dfrac{X_{ik}^{(\lambda_k)}-1}{\lambda_k\left(\prod\limits_{i=1}^{n}X_{ik}\right)^{\frac{\lambda_k-1}{n}}}$, $i=1,2,\ldots,n$

$\overline{Y^{(\lambda_k)}}=\frac{1}{n}\sum\limits_{i=1}^{n}Y_i^{(\lambda_k)}$, $k=1,2,\ldots,p$

$X_{1k},X_{2k},\ldots,X_{nk}$ 为第k个变量的n个观测值

则 $X^{(\lambda)}=\begin{pmatrix}X_1^{(\lambda_1)}\\ \vdots\\ X_p^{(\lambda_p)}\end{pmatrix}=\begin{pmatrix}\frac{X_1^{\lambda_1}-1}{\lambda_1}\\ \vdots\\ \frac{X_p^{\lambda_p}-1}{\lambda_p}\end{pmatrix}$

第三章　关于均值向量的推断

从本章开始, 我们转入多元统计学的方法论, 将集中讨论关于总体均值向量及其分量的统计推断问题. 虽然我们将从假设检验开始统计推断的讨论, 但我们的最终目的还是要基于联合置信域的形式给出均值向量诸分量的一个完整的统计分析.

多元分析的精髓之一就是必须对 p 个相关变量同时进行分析.

§3.1　均值向量的检验

一、一元正态分布的均值检验

设 $X\sim N(\mu,\sigma^2)$, $X_1,X_2,\ldots,X_n$ 是来自 X 的样本

$$H_0:\mu=\mu_0\ ,\ H_1:\mu\neq\mu_0$$

因为

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i\sim N\left(\mu,\frac{\sigma^2}{n}\right)$$

$$\frac{(n-1)s^2}{\sigma^2}=\frac{\sum\limits_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{\sigma^2}\sim\chi^2_{n-1}$$

所以 $\dfrac{(\overline{X}-\mu)/(\sigma/\sqrt{n})}{s/\sigma}=\dfrac{\overline{X}-\mu}{s/\sqrt{n}}\sim t_{n-1}$.

当 H_0 成立时,

$$\frac{\overline{X}-\mu_0}{s/\sqrt{n}}\sim t_{n-1}\text{ , 等价地 }\frac{\left(\overline{X}-\mu_0\right)^2}{s^2/n}\sim F_{1,n-1}$$

对于任意给定的 $\alpha>0$, 当 $\dfrac{|\overline{X}-\mu_0|}{s/\sqrt{n}}>t_{n-1}\left(\dfrac{\alpha}{2}\right)$ 时，拒绝 H_0.

等价地，当 $\dfrac{\left(\overline{X}-\mu_0\right)^2}{s^2/n}>F_{1,n-1}(\alpha)$ 时，拒绝 H_0.

p 值法　$p=\Pr\{F>F_1\}$, 其中 F_1 为统计量 $\dfrac{n\left(\overline{X}-\mu\right)^2}{s^2}$ 的观测值. 当 $p<\alpha$ 时，拒绝 H_0; 否则，不拒绝 H_0. α 通常取 0.01,0.05,0.1.

二、 p 元正态分布均值向量的检验

设 $\mathbf{X} \sim N_p(\mu, \mathbf{\Sigma})$, $\mu, \mathbf{\Sigma}$ 未知， $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ 是来自 $\mathbf{X}$ 的样本. 下面讨论假设

$$H_0 : \mu = \mu_0 \ , \ H_1 : \mu \neq \mu_0$$

的检验问题.

在一元中，检验统计量为

$$F = \frac{n\left(\overline{X} - \mu_0\right)^2}{s^2} = n\left(\overline{X} - \mu_0\right)'\left(s^2\right)^{-1}\left(\overline{X} - \mu_0\right).$$

将其推广到 p 元得到：

$$n\left(\overline{\mathbf{X}} - \mu_0\right)'\mathbf{S}^{-1}\left(\overline{\mathbf{X}} - \mu_{\mathbf{0}}\right)$$

$$= \left(\frac{\overline{\mathbf{X}} - \mu_{\mathbf{0}}}{1/\sqrt{n}}\right)'\left(\frac{(n-1)\,\mathbf{S}}{n-1}\right)^{-1}\left(\frac{\overline{\mathbf{X}} - \mu_{\mathbf{0}}}{1/\sqrt{n}}\right)$$

其中 $\dfrac{\overline{\mathbf{X}} - \mu_{\mathbf{0}}}{1/\sqrt{n}} \sim N_p(\mathbf{0}, \mathbf{\Sigma})$,

$$(n-1)\mathbf{S} = \sum_{i=1}^{n}\left(\mathbf{X_i} - \overline{\mathbf{X}}\right)\left(\mathbf{X_i} - \overline{\mathbf{X}}\right)' \sim W_{n-1}(\Sigma).$$

将上述统计量记为 T^2, 即 $T^2 = n\left(\overline{X} - \mu_0\right)' S^{-1}\left(\overline{X} - \mu_0\right)$, 称之为 Hotelling 统计量, 因为 Hotelling 首先求出了 T^2 的分布, 他证明了统计量 T^2 与 $\dfrac{(n-1)\,p}{n-p}F_{p,n-p}$ 同分布，或等价地，

$$\frac{(n-1)\,p}{n-p}T^2 \sim F_{p,n-p}$$

当 $p = 1$ 时，正是一元的情形.

因此，对给定的 $\alpha > 0$, 当 $T^2 > \dfrac{(n-1)\,p}{n-p}F_{p,n-p}(\alpha)$ 时，拒绝 H_0; 否则，不拒绝 H_0.

例 5.2

三、 Hotelling 统计量 T^2 与似然比检验

在构造检验方法时存在一个一般原理，即所谓的似然比方法，而 T^2 统计量则能从 $H_0 : \mu = \mu_0$ 的似然比检验导出.

设 $\mathbf{X} \sim N_p(\mu, \mathbf{\Sigma}), \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ 是来自 $\mathbf{X}$ 的样本. 对假设

$$H_0 : \mu = \mu_0 \ , \ H_1 : \mu \neq \mu_0$$

构造的似然比检验统计量为

$$\mathbf{\Lambda} = \frac{\sup_{\mathbf{\Sigma}} L(\mu_0, \mathbf{\Sigma})}{\sup_{\mu, \mathbf{\Sigma}} L(\mu, \mathbf{\Sigma})}$$

其中 $L(\mu, \mathbf{\Sigma})$ 为样本 $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ 的联合概率密度函数.

由参数$\mu, \mathbf{\Sigma}$的极大似然估计的推导知,当 $\hat{\mu}=\overline{\mathbf{X}}$, $\hat{\mathbf{\Sigma}}=\frac{1}{n}\sum\limits_{i=1}^{n}\left(\mathbf{X}_i-\overline{\mathbf{X}}\right)\left(\mathbf{X}_i-\overline{\mathbf{X}}\right)'$ 时, $L\left(\mu, \mathbf{\Sigma}\right)$ 达到最大, 且

$$\sup_{\mu, \mathbf{\Sigma}} L(\mu, \Sigma)=\frac{1}{(2\pi)^{\frac{np}{2}}\left|(n-1)\,\mathbf{S}\right|^{\frac{n}{2}}} n^{\frac{np}{2}} e^{-\frac{np}{2}}$$

同理

$$\sup_{\mu_0, \mathbf{\Sigma}} L(\mu, \mathbf{\Sigma})=\frac{1}{(2\pi)^{\frac{np}{2}}\left|\sum\limits_{i=1}^{n}\left(\mathbf{X}_i-\mu_0\right)\left(\mathbf{X}_i-\mu_0\right)'\right|^{\frac{n}{2}}} n^{\frac{np}{2}} e^{-\frac{np}{2}}$$

记 $\hat{\mathbf{\Sigma_1}}=\frac{1}{n}\sum\limits_{i=1}^{n}\left(\mathbf{X}_i-\mu_0\right)\left(\mathbf{X}_i-\mu_0\right)'$

所以

$$\Lambda=\frac{\left|\sum\limits_{i=1}^{n}\left(\mathbf{X}_i-\overline{\mathbf{X}}\right)\left(\mathbf{X}_i-\overline{\mathbf{X}}\right)'\right|^{\frac{n}{2}}}{\left|\sum\limits_{i=1}^{n}\left(\mathbf{X}_i-\mu_0\right)\left(\mathbf{X}_i-\mu_0\right)'\right|^{\frac{n}{2}}}$$

若 H_0 成立, 则 Λ 应比较大, 故当 $\Lambda<\Lambda_\alpha$ 时, 拒绝 H_0.

与 Λ 等价的统计量 $\Lambda^{\frac{2}{n}}=\dfrac{\left|\sum\limits_{i=1}^{n}\left(\mathbf{X}_i-\overline{\mathbf{X}}\right)\left(\mathbf{X}_i-\overline{\mathbf{X}}\right)'\right|}{\left|\sum\limits_{i=1}^{n}\left(\mathbf{X}_i-\mu_0\right)\left(\mathbf{X}_i-\mu_0\right)'\right|}$

威尔克斯统计量

临界值 Λ_α 不易确定, 幸好 T^2 与 Λ 之间存在一一对应的关系, 因此在检验时, 我们不需要知道 Λ 的分布.

结论 3.1.1　设 $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ 为取自正态总体的一个随机样本, 则

$$\begin{aligned}
&\Lambda^{\frac{2}{n}}=\left(1+\frac{T^2}{n-1}\right)^{-1}\\
\Longleftrightarrow\ &T^2=\frac{(n-1)\left|\sum\limits_{i=1}^{n}\left(\mathbf{X}_i-\mu_0\right)\left(\mathbf{X}_i-\mu_0\right)'\right|}{\left|\sum\limits_{i=1}^{n}\left(\mathbf{X}_i-\overline{\mathbf{X}}\right)\left(\mathbf{X}_i-\overline{\mathbf{X}}\right)'\right|}-(n-1)
\end{aligned}$$

由此可见， T^2 统计量可以通过两种方式获得:

$$T^2 = n\left(\overline{\mathbf{X}} - \mu_0\right)' \mathbf{S}^{-1}\left(\overline{\mathbf{X}} - \mu_0\right) \quad 或$$

$$T^2 = \frac{(n-1)\left|\sum_{i=1}^{n}\left(\mathbf{X}_i - \mu_0\right)\left(\mathbf{X}_i - \mu_0\right)'\right|}{\left|\sum_{i=1}^{n}\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)'\right|} - (n-1)$$

似然比检验在多元分析中是常用的一种方法，这种检验所具有的优良大样本性质在很大范围内都成立.

四、广义似然比方法

现在我们考虑广义似然比方法. 设 θ 是由总体所有未知参数所组成的向量， $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ 为来自总体 $\mathbf{X} \sim f(\mathbf{X}, \theta)$ 的一个随机样本， $L(\theta)$ 为似然函数. 参数向量 θ 在参数集 Θ 中取值. 考虑假设

$$H_0:\ \theta \in \Theta_0 \qquad H_1:\ \theta \overline{\in} \Theta_0$$

$$L(\theta) = \prod_{i=1}^{n} f\left(\mathbf{X_i}, \theta\right)$$

似然比

$$\Lambda = \frac{\sup\limits_{\theta \in \Theta_0} L\left(\theta\right)}{\sup\limits_{\theta \in \Theta} L\left(\theta\right)}$$

若 $\Lambda < C_\alpha$, 则拒绝 H_0; 否则不拒绝 H_0.C_α 为适当的常数.

凭直觉来看，如果 θ 在 Θ_0 上变换时所得的最大似然函数值比 θ 在 Θ 上变换时要小得多，就拒绝 H_0, 即认为 Θ_0 不包含 θ 的似真值.

在似然比方法的各种应用中，我们必须知道似然比检验统计量 Λ 的抽样分布, 于是在给定的显著性水平 $\alpha > 0$ 下, 选定检验法中的临界值 C_α. 但是, 当样本容量很大且满足一定的正则性条件时， $-2\ln\Lambda$ 的抽样分布与 χ^2 分布十分接近. 这一吸引人的特征部分地说明了为什么似然比方法如此流行.

结论 3.1.2　当总体的分布 $f(\mathbf{X}, \theta)$ 满足正则条件，且样本容量很大时

$$-2\ln\Lambda = -2\ln\left\{\frac{\sup_{\theta\in\Theta_0} L\left(\theta\right)}{\sup_{\theta\in\Theta} L\left(\theta\right)}\right\} \sim \chi^2_{\nu-\nu_0}$$

这里自由度 $\nu - \nu_0 = dim\Theta - dim\Theta_0$

例　设 $\mathbf{X} \sim N_p(\mu, \mathbf{\Sigma})$, $X_1, X_2, \ldots, X_n$ 为来自总体 $\mathbf{X} \sim f(\mathbf{X}, \theta)$ 的随机样本,

$$H_0:\ \mu = \mu_0\ ,\ H_1:\ \mu \neq \mu_0$$

$$\Theta = \left\{\left(\mu, \mathbf{\Sigma}\right), \mu \in R^p, \mathbf{\Sigma}\text{是非负定矩阵}\right\},\ dim\Theta = p + \frac{p\left(p+1\right)}{2}$$

$$\Theta_0 = \left\{\left(\mu_\mathbf{0}, \mathbf{\Sigma}\right), \mathbf{\Sigma}\text{是非负定矩阵}\right\},\ dim\Theta_0 = \frac{p\left(p+1\right)}{2}$$

所以 $\nu-\nu_0=p$, 即当 n 很大时

$$-2\ln\Lambda=-2\ln\left(\frac{\left|\sum\limits_{i=1}^{n}(\mathbf{X_i}-\overline{\mathbf{X}})(\mathbf{X_i}-\overline{\mathbf{X}})'\right|}{\left|\sum\limits_{i=1}^{n}(\mathbf{X_i}-\mu_\mathbf{0})(\mathbf{X_i}-\mu_\mathbf{0})'\right|}\right)^{\frac{n}{2}}$$

$$=-2\ln\left(1+\tfrac{T^2}{n-1}\right)^{-\frac{n}{2}}$$

$$=n\ln(1+\tfrac{T^2}{n-1})\sim\chi_p^2.$$

§3.2 置信域和均值分量的联合比较

为得到利用样本进行推断的基本方法，我们需要把一元置信区间的概念推广到多元置信域的情形. 设 θ 为总体未知参数向量， Θ 为 θ 所有可能取值的集合. 置信域就是 θ 可能的所有值集合. 置信域是通过样本数据来确定的，我们暂且用 $R(\mathcal{X})$ 来表示，这里 $\mathcal{X}=(\mathbf{X_1},\mathbf{X_2},\ldots,\mathbf{X_n})'$ 为数据矩阵.

若 $\Pr\{\theta_0\in R(\mathcal{X})\}=1-\alpha$

则称 $R(\mathcal{X})$ 为参数 θ 的置信水平为 $1-\alpha$ 的置信域，其中的概率是在参数 θ 的未知真值处计算的.

一、置信域

1. 一元的情形　$X\sim N(\mu,\sigma^2)$, $X_1,X_2,\ldots,X_n$ 是来自 X 的样本，求 μ 的置信区间.

   因为

   $$\frac{\overline{X}-\mu}{s/\sqrt{n}}\sim t_{n-1}$$

   所以，对给定的 α, $\Pr\left\{\left|\frac{\overline{X}-\mu}{s/\sqrt{n}}\right|\le t_{n-1}(\frac{\alpha}{2})\right\}=1-\alpha$

   等价地

   $$\Pr\left\{\overline{X}-t_{n-1}\left(\frac{\alpha}{2}\right)\cdot\frac{s}{\sqrt{n}}\le\mu\le\overline{X}+t_{n-1}\left(\frac{\alpha}{2}\right)\cdot\frac{s}{\sqrt{n}}\right\}=1-\alpha$$

   故参数 μ 的置信水平为 $1-\alpha$ 的置信区间为

   $$\left[\overline{X}-t_{n-1}\left(\frac{\alpha}{2}\right)\cdot\frac{s}{\sqrt{n}},\overline{X}+t_{n-1}\left(\frac{\alpha}{2}\right)\cdot\frac{s}{\sqrt{n}}\right]$$

2. p 元情形　设 $\mathbf{X}\sim N_p(\mu,\mathbf{\Sigma})$, $\mathbf{X_1},\mathbf{X_2},\ldots,\mathbf{X_n}$ 为来自总体 $\mathbf{X}$ 的随机样本，求参数 μ 的置信水平为 $1-\alpha$ 的置信域.

   因为

   $$T^2=n\left(\overline{\mathbf{X}}-\mu\right)'\mathbf{S^{-1}}\left(\overline{\mathbf{X}}-\mu\right)\sim\frac{(n-1)\,p}{n-p}F_{p,n-p}$$

所以，对给定的显著性水平 $\alpha>0$, 因

$$\Pr\left\{T^2 \leq \frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)\right\}=1-\alpha$$

即 $\Pr\left\{\mu:\, n\left(\overline{\mathbf{X}}-\mu\right)'\mathbf{S}^{-1}\left(\overline{\mathbf{X}}-\mu\right)\leq \frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)\right\}=1-\alpha$

故 μ 的置信水平为 $1-\alpha$ 的置信域为：

$$\left\{\mu:\, n\left(\overline{\mathbf{X}}-\mu\right)'\mathbf{S}^{-1}\left(\overline{\mathbf{X}}-\mu\right)\leq \frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)\right\}$$

有时也称为置信椭球, 它是以 $\overline{\mathbf{X}}$ 为中心, 以 $\frac{\sqrt{\lambda_i}c}{\sqrt{n}}\mathbf{e}_i$ 为轴方向的椭球, 其中 $c^2=\frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)$, $\mathbf{S}\mathbf{e}_i=\lambda_i\mathbf{e}_i, i=1,2,\ldots,p$, $\lambda_1\geq\lambda_2\geq\ldots\geq\lambda_p$.

(例 5.3 P181)

二、 T^2 联合置信区间

设 $\mathbf{X}\sim N_p(\mu,\mathbf{\Sigma})$, $\mathbf{a}'=(a_1,a_2,\ldots,a_p)'$.

考虑 $Z=\mathbf{a}'\mathbf{X}\sim N\left(\mathbf{a}'\mu,\mathbf{a}'\mathbf{\Sigma}\mathbf{a}\right)$

设 $\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_n$ 是来自 $\mathbf{X}$ 的样本, 样本均值向量为 $\overline{\mathbf{X}}$, 样本协方差矩阵为 $\mathbf{S}$, 则 Z 对应的样本为 $Z_1,Z_2,\ldots,Z_n$, 样本均值和样本方差分别为

$$\overline{Z}=\mathbf{a}'\overline{\mathbf{X}},\quad s_Z^2=\mathbf{a}'\mathbf{S}\mathbf{a}$$

问题是：对于任意一个向量 $\mathbf{a}$, 求 $\mathbf{a}'\mu$ 的置信区间.

因为 $\mathbf{a}'\mathbf{X}\sim N(\mathbf{a}'\mu,\mathbf{a}'\mathbf{\Sigma}\mathbf{a})$, 故同一元正态分布的情形, 对于给定的置信水平 $1-\alpha$, 有

$$\Pr\left\{\mathbf{a}'\overline{\mathbf{X}}-t_{n-1}\left(\frac{\alpha}{2}\right)\sqrt{\frac{\mathbf{a}'\mathbf{S}\mathbf{a}}{n}}\leq\mathbf{a}'\mu\leq\mathbf{a}'\overline{\mathbf{X}}+t_{n-1}\left(\frac{\alpha}{2}\right)\sqrt{\frac{\mathbf{a}'\mathbf{S}\mathbf{a}}{n}}\right\}=1-\alpha$$

故 $a'\mu$ 的置信水平为 $1-\alpha$ 的置信区间为：

$$\left[\mathbf{a}'\overline{\mathbf{X}}-t_{n-1}\left(\frac{\alpha}{2}\right)\sqrt{\frac{\mathbf{a}'\mathbf{S}\mathbf{a}}{n}},\ \mathbf{a}'\overline{\mathbf{X}}+t_{n-1}\left(\frac{\alpha}{2}\right)\sqrt{\frac{\mathbf{a}'\mathbf{S}\mathbf{a}}{n}}\right]$$

特别地, 当 $\mathbf{a}'=(0,\ldots,0,1,0,\ldots,0)$ 时, $\mathbf{a}'\mu=\mu_i$, $i=1,2,\ldots,p$.

故 μ_i 的置信水平为 $1-\alpha$ 的置信区间为：

$$\left[\overline{X}_i-t_{n-1}\left(\frac{\alpha}{2}\right)\sqrt{\frac{s_{ii}}{n}},\ \overline{X}_i+t_{n-1}\left(\frac{\alpha}{2}\right)\sqrt{\frac{s_{ii}}{n}}\right],\ i=1,2,\ldots,p \quad (3.2.1)$$

称之为单一 t 区间.

但这 p 个单一 t 区间都同时包含各自 μ_i 的概率不一定再是 $1-\alpha$.

例如：设 $\mathbf{X}\sim N_p(\mu,\mathbf{\Sigma})$, $\mathbf{\Sigma}=\begin{pmatrix}\sigma_{11}&0&0&0\\0&\sigma_{22}&0&0\\0&0&\ddots&0\\0&0&0&\sigma_{pp}\end{pmatrix}$.

则由 $\mathbf{X}$ 各分量的相互独立性知,

$$\Pr\left\{\mu_i \in \text{第}i\text{个单一 t 区间}, i=1,2,\ldots,p\right\}$$

$$=\prod_{i=1}^{p}\Pr\left\{\mu_i \in \text{第}i\text{个单一区间}\right\}$$

$$=(1-\alpha)^p<1-\alpha$$

故要使所有 $\mathbf{a}$ 对应的置信区间包含 $\mathbf{a}'\mu$ 的概率仍为 $1-\alpha$, 一个很自然的方法就是将随机变量放大，让其与变量 $\mathbf{a}$ 无关.

因为

$$\frac{\sqrt{n}\left(\mathbf{a}'\overline{\mathbf{X}}-\mathbf{a}'\mu\right)}{\sqrt{\mathbf{a}'\mathbf{S}\mathbf{a}}}\sim t_{n-1}$$

$$\frac{n\left(\mathbf{a}'\overline{\mathbf{X}}-\mathbf{a}'\mu\right)^2}{\mathbf{a}'\mathbf{S}\mathbf{a}}\sim F_{1,n-1}$$

根据定理:

$$\max_{\mathbf{X}\neq 0}\frac{\left(\mathbf{X}'\mathbf{d}\right)^2}{\mathbf{X}'\mathbf{B}\mathbf{X}}=\mathbf{d}'\mathbf{B}^{-1}\mathbf{d}$$

其中 $\mathbf{B}$ 是正定矩阵，得:

$$\max_{\mathbf{a}\neq 0}\frac{n\left[\mathbf{a}'\left(\overline{\mathbf{X}}-\mu\right)\right]^2}{\mathbf{a}'\mathbf{S}\mathbf{a}}=n\left(\overline{\mathbf{X}}-\mu\right)'\mathbf{S}^{-1}\left(\overline{\mathbf{X}}-\mu\right)=T^2\sim\frac{(n-1)\,p}{n-p}F_{p,n-p}$$

故对给定的 $\alpha>0$, 因为

$$\Pr\left\{T^2\leq\frac{(n-1)\,p}{n-p}F_{p,n-p}(\alpha)\right\}=1-\alpha$$

而

$$\frac{n\left[\mathbf{a}'\left(\overline{\mathbf{X}}-\mu\right)\right]^2}{\mathbf{a}'\mathbf{S}\mathbf{a}}\leq T^2$$

所以 $\Pr\left\{\dfrac{n\left[\mathbf{a}'\left(\overline{\mathbf{X}}-\mu\right)\right]^2}{\mathbf{a}'\mathbf{S}\mathbf{a}}\leq c^2\right\}\geq 1-\alpha$

即对于任意的 $\mathbf{a}$, 都有

$$\Pr\left\{\frac{\sqrt{n}\left[\mathbf{a}'\left(\overline{\mathbf{X}}-\mu\right)\right]}{\sqrt{\mathbf{a}'\mathbf{S}\mathbf{a}}}\leq c\right\}\geq 1-\alpha$$

由此可得结论:

结论 3.2.1　设 $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ 为来自总体 $N_p(\mu, \mathbf{\Sigma})$ 的样本，且 $\mathbf{\Sigma}$ 为正定矩阵，则对所有 $\mathbf{a}$, 区间

$$\left[\mathbf{a}'\overline{\mathbf{X}} - c \cdot \sqrt{\frac{\mathbf{a}'\mathbf{S}\mathbf{a}}{n}}, \mathbf{a}'\overline{\mathbf{X}} + c \cdot \sqrt{\frac{\mathbf{a}'\mathbf{S}\mathbf{a}}{n}}\right]$$

至少以概率 $1-\alpha$ 包含 $\mathbf{a}'\mu$, 我们称上述区间为 T^2- 联合置信区间.

特别地，当 $\mathbf{a}' = (0, \ldots, 0, 1, 0, \ldots, 0)$ 时，我们得到 μ_i 的 T^2- 联合置信区间为：

$$\left[\overline{X}_i - c \cdot \sqrt{\frac{s_{ii}}{n}}, \overline{X}_i + c \cdot \sqrt{\frac{s_{ii}}{n}}\right], i = 1, 2, \ldots, p \quad (3.2.2)$$

它与 (3.2.1) 的形式完全相同，只是临界值不同，对于给定的 $\alpha > 0$, 一般地，有

$$t_{n-1}\left(\frac{\alpha}{2}\right) < c = \sqrt{\frac{(n-1)\,p}{n-p} F_{p,n-p}(\alpha)}$$

另外，(3.2.2) 恰好为置信椭球在第 i 个坐标轴上的投影.

例 5.4　例 5.5

三、庞弗罗尼方法

如果 T^2 区间仅应用于 p 个均值向量，那么 T^2 区间显然就太宽了. 为了说明其原因，我们来考虑二元置信椭圆及联合置信区间. 若 μ_1 , μ_2 分别落在各自的 T^2 区间内，那么 (μ_1, μ_2) 就落在由此两区间围成的矩形内. 此矩形显然包含了置信椭圆. 虽然置信椭圆更小些，但它却以 0.95 的概率包含了均值向量 $\mu = (\mu_1, \mu_2)'$. 因此，对由 T^2 区间所围成的矩形而言，其包含两个均值分量 μ_1 与 μ_2 的概率要大于 0.95. 这一结果促使我们去考虑另一种进行多重比较的方法，即所谓的庞弗罗尼方法.

如果我们仅把注意力集中在有限个线性组合 $\mathbf{a}_1'\mu, \ldots, \mathbf{a}_m'\mu$ 的联合置信区间上，则可采用庞弗罗尼方法得到更精确的联合置信区间.

设 m 个线性组合 $\mathbf{a}_1'\mu, \ldots, \mathbf{a}_m'\mu$ 的置信区间分别用 $C_1, \ldots, C_m$ 来表示，且

$$\Pr\{\mathbf{a}_i'\mu \in C_i\} = 1 - \alpha_i, i = \overline{1, m}$$

因为

$$\begin{aligned}\Pr\{\mathbf{a}_i'\mu \in C_i,\ i = 1, \ldots, m\} &= 1 - \Pr\left\{\text{至少有一个} i,\ \mathbf{a}_i'\mu \overline{\in} C_i\right\} \\ &\geq 1 - \sum_{i=1}^{m} \Pr\{\mathbf{a}_i'\mu \overline{\in} C_i\} \\ &= 1 - \sum_{i=1}^{m} \alpha_i\end{aligned}$$

由此，

我们可以控制总体误差率 $\alpha_1 + \alpha_2 + \ldots + \alpha_m$, 而不管置信表示中隐藏着何种相关结构.

根据单 -t 区间有

$$P\left\{\mathbf{a}_i'\mu \in \left[\mathbf{a}_i'\overline{\mathbf{X}} - t_{n-1}\left(\frac{\alpha_i}{2}\right)\sqrt{\frac{\mathbf{a}_i'\mathbf{S}\mathbf{a}_i}{n}}, \mathbf{a}_i'\overline{\mathbf{X}} + t_{n-1}\left(\frac{\alpha_i}{2}\right)\sqrt{\frac{\mathbf{a}_i'\mathbf{S}\mathbf{a}_i}{n}}\right]\right\} = 1 - \alpha_i$$

$(i=1,2,\ldots,m)$

从而

$$\Pr\left\{\mathbf{a}_i'\mu\in\left[\mathbf{a}_i'\overline{\mathbf{X}}-t_{n-1}\left(\frac{\alpha_i}{2}\right)\sqrt{\frac{\mathbf{a}_i'\mathbf{S}\mathbf{a}_i}{n}},\mathbf{a}_i'\overline{\mathbf{X}}+t_{n-1}\left(\frac{\alpha_i}{2}\right)\sqrt{\frac{\mathbf{a}_i'\mathbf{S}\mathbf{a}_i}{n}}\right]^2,\ i=\overline{1,m}\right\}\geq 1-\sum_{i=1}^{m}\alpha_i$$

若取 $\alpha_i=\frac{\alpha}{m}$, 则

$$\Pr\left\{\mathbf{a}_i'\mu\in\left[\mathbf{a}_i'\overline{\mathbf{X}}-t_{n-1}\left(\frac{\alpha_i}{2}\right)\sqrt{\frac{\mathbf{a}_i'\mathbf{S}\mathbf{a}_i}{n}},\mathbf{a}_i'\overline{\mathbf{X}}+t_{n-1}\left(\frac{\alpha_i}{2}\right)\sqrt{\frac{\mathbf{a}_i'Sa_i}{n}}\right]^2,\ i=\overline{1,m}\right\}\geq 1-\alpha$$

即区间 $\left[\mathbf{a}_i'\overline{\mathbf{X}}-t_{n-1}\left(\frac{\alpha_i}{2}\right)\sqrt{\frac{\mathbf{a}_i'\mathbf{S}\mathbf{a}_i}{n}},\mathbf{a}_i'\overline{\mathbf{X}}+t_{n-1}\left(\frac{\alpha_i}{2}\right)\sqrt{\frac{\mathbf{a}_i'\mathbf{S}\mathbf{a}_i}{n}}\right]^2,\ i=\overline{1,m},$ 至少以概率 $1-\alpha$ 包含 $\mathbf{a}_i'\mu,\ i=\overline{1,m}$, 我们称上述区间为庞弗罗尼联合置信区间.

特别地，当 $\mathbf{a}'=(0,\ldots,0,1,0,\ldots,0)$ 时，$i=1,2,\ldots,p$, 我们可以得到 $\mu=(\mu_1,\ldots,\mu_p)'$ 的 p 个分量 $\mu_i,\ i=\overline{1,p}$ 的置信水平为 $1-\alpha$ 的联合置信区间为：

$$\begin{aligned}&\left[\overline{X}_1-t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{11}}{n}},\overline{X}_1+t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{11}}{n}}\right]\\&\left[\overline{X}_2-t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{22}}{n}},\overline{X}_2+t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{22}}{n}}\right]\\&\ldots\\&\left[\overline{X}_p-t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{pp}}{n}},\overline{X}_p+t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{pp}}{n}}\right]\end{aligned}\qquad(3.2.3)$$

将 (3.2.3) 与 (3.2.2) 进行比较，我们发现 (3.2.3) 仅以分位数 $t_{n-1}\left(\frac{\alpha}{2p}\right)$ 替代了 (3.2.2) 中的 c, 区间保持相同的结构.

例 5.6　(庞弗罗尼联合置信区间的构造以及 T^2 区间的比较)

$\forall i,j\quad \mu_i-\mu_j$ 的联合置信区间为：

$$\mu_i-\mu_j:\ \overline{X}_i-\overline{X}_j\pm t_{n-1}\left(\frac{\alpha}{2m}\right)\sqrt{\frac{s_{ii}+s_{jj}-2s_{ij}}{n}}\quad i\neq j$$

其中 $m=C_p^2$

§3.3　总体均值向量的大样本推断

当样本容量很大时，我们不需要总体的正态性假设就可以构造 μ 的假设检验及置信域，所有 μ 的大样本推断都是建立在 χ^2 分布的基础上的.

因为 $n\left(\overline{\mathbf{X}}-\mu_0\right)' S^{-1}\left(\overline{\mathbf{X}}-\mu_0\right) \sim \chi_p^2$

所以 $P\left\{n\left(\overline{\mathbf{X}}-\mu_0\right)' S^{-1}\left(\overline{\mathbf{X}}-\mu_0\right) \leq \chi_p^2(\alpha)\right\}=1-\alpha$

由此可推得大样本得假设检验与联合置信域，其过程概括在下面得两个结论中

结论 3.3.1　设 $\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n}$ 为来自总体均值为 μ, 协方差矩阵为正定阵 $\mathbf{\Sigma}$ 的一个随机样本，当 $n-p$ 很大时，若观测值

$$n\left(\overline{\mathbf{X}}-\mu_0\right)' S^{-1}\left(\overline{\mathbf{X}}-\mu_0\right)>\chi_p^2(\alpha)$$

则在显著性水平 α 下，拒绝 $H_0:\ \mu=\mu_0$

将上述检验与正态总体下均值检验作比较, 我们发现, 它们的检验统计量具有相同的结构,只是临界值不同,但当 n 相对于 p 较大时, $(n-1)pF_{p,n-p}(\alpha)/(n-p)$ 与 $\chi_p^2(\alpha)$ 近似相等，故两种检验基本上给出相同的结果.

结论 3.3.2　$\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n}$ 为来自总体均值为 μ, 协方差矩阵为正定阵 $\mathbf{\Sigma}$ 的一个随机样本，当 $n-p$ 很大时，对所有的 $\mathbf{a}$

$$a'\overline{\mathbf{X}} \pm \sqrt{\chi_p^2(\alpha)}\sqrt{\frac{\mathbf{a'Sa}}{n}}$$

包含 $\mathbf{a}'\mu$ 的概率近似为 $1-\alpha$. 特别地，我们有置信水平至少为 $1-\alpha$ 的联合置信表示:

$$\begin{aligned}
&\overline{X}_1 \pm \sqrt{\chi_p^2(\alpha)}\sqrt{s_{11}/n} \text{包含} \mu_1 \\
&\vdots \qquad\qquad\qquad\qquad\qquad\qquad (3.3.1) \\
&\overline{X}_p \pm \sqrt{\chi_p^2(\alpha)}\sqrt{s_{pp}/n} \text{包含} \mu_p
\end{aligned}$$

此外，对所有 $(\mu_i\,,\,\mu_k)\,,\, i\,,\, k=1,2,\ldots,p$, 以样本均值为中心的置信椭圆

$$n\left(\overline{X}_i-\mu_i\,,\,\overline{X}_k-\mu_k\right)\begin{pmatrix} s_{ii} & s_{ik} \\ s_{ki} & s_{kk} \end{pmatrix}^{-1}\begin{pmatrix} \overline{X}_i-\mu_i \\ \overline{X}_k-\mu_k \end{pmatrix} \leq \chi_p^2(\alpha)$$

包含 $(\mu_i\,,\,\mu_k)$ 的概率近似为 $1-\alpha$.

当样本容量很大时，均值分量的单 - t 置信区间为:

$$\overline{X}_i \pm U_{\frac{\alpha}{2}}\sqrt{\frac{s_{ii}}{n}}\ \ i=1,2,\ldots,p \quad (3.3.2)$$

其中 $U_{\frac{\alpha}{2}}$ 表示正态分布的 $\frac{\alpha}{2}$ 上侧分位数.

当 $m=p$ 时，单个均值的庞弗罗尼联合置信区间为:

$$\overline{X}_i \pm U(\frac{\alpha}{2p})\sqrt{\frac{s_{ii}}{n}}\ \ i=1,2,\ldots,p \quad (3.3.3)$$

有些统计学家更愿意使用基于 F 分布与 t 分布的分位数，而不愿意使用基于 χ^2 分布及标准正态分布的分位数，即将 (3.3.1) 、 (3.3.2) 和 (3.3.3) 中的 $\chi_p^2(\alpha)$,$U(\frac{\alpha}{2})$ 与 $U(\frac{\alpha}{2p})$ 分别换成 $F_{p,n-p}(\alpha)\,,\, t_{n-1}(\frac{\alpha}{2})$ 和 $t_{n-1}(\frac{\alpha}{2p})$.

关于大样本情况下，均值向量的比较亦有类似的结论：
对于任意的 $i \neq j$ $\mu_i - \mu_j$ 的庞弗罗尼联合置信区间为：

$$\overline{X}_i - \overline{X}_j \pm U\left(\frac{\alpha}{2m}\right)\sqrt{\frac{s_{ii}+s_{jj}-2s_{ij}}{n}}\ , i,j=1,2,\cdots,p.\ m=\mathrm{C}_p^2.$$

§3.4 多元质量控制图

为改进产品和服务的质量，需要考虑数据以查明引起质量变动的原因. 当我们监控某种服务活动时，就应收集数据，用以评估这个过程的能力与稳定性. 当一个过程稳定时，变动是由一些经常存在的原因引起的，没有一个原因是变动的主要来源.

任何控制图的目的，都是为了识别是否出现了引起变动的特殊原因，这些原因是来自常规过程的外部. 这些变动原因的出现表现需要某种适时的维修，不过也可能暗示需要改进此过程，控制图可使这些变动可视化，因而有利于我们区分变动的一般原因和特殊原因.

一个控制图一般是由按时间顺序标绘的数据点和两条水平线组成，这两条水平线成为控制限，用来表示由一般原因引起的变动量.

一、控制图

1. 一元 $\overline{X}$ 控制图

若 $X \sim N(\mu,\sigma^2)$, 则 $\Pr\{-3\sigma < X-\mu < 3\sigma\} = 99.97$ %设 $X_1, X_2 \ldots, X_n$ 是来自任意总体的样本， $EX=\mu$, $DX=\sigma^2$, 则当 n 充分大时，

$$\overline{X} \dot{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\frac{\overline{X}-\mu}{s/\sqrt{n}} \dot{\sim} N(0,1)$$

从而 $\Pr\left\{\overline{X} - \frac{3s}{\sqrt{n}} < \mu < \overline{X} + \frac{3s}{\sqrt{n}}\right\} = 99.97$ %

(a) 按时间顺序对各个观测值或样本均值作出标绘;
(b) 生成并画出代表所有观测值的样本均值中的中心线 $\overline{\overline{X}}$;
(c) 按如下公式计算并画出控制限

$$\text{控制上限}(UCL) = \overline{\overline{X}} + \frac{3s}{\sqrt{n}}$$

$$\text{控制下限}(LCL) = \overline{\overline{X}} - \frac{3s}{\sqrt{n}}$$

其目的是：为了在数据服从正态分布的假设下，或当样本容量 n 很大时，误将正常数据标成失控数据的可能性变得非常小. 所谓失控数据，是指暗示存在某种特殊变动原因的观测值.

例 3.4.1 (p196 例 5.8)

在考虑多于一个重要特征的情形下，需要用多元的方法来监控过程的稳定性. 这种方法能考虑诸特征之间的相关性，且能控制对波动的特

殊原因发出错误信号的总体概率. 变量间的高度相关使得评估一个总的误差率成为不可能，而这一点却可以通过大量一元控制图来实现.

两个最常见的多元控制图是椭圆控制图和 T^2 控制图.

2. 二元椭圆控制图

设 $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \sim N_p(\mu, \mathbf{\Sigma})$, $i.i.d$, 则

$$\mathbf{X}_j - \overline{\mathbf{X}} \sim N_p\left(0, \frac{n-1}{n}\mathbf{\Sigma}\right)$$

因为 $\mathbf{X}_j - \overline{\mathbf{X}}$ 与 $\mathbf{S}$ 不独立，故只能用近似的 χ^2 分布来确定其控制限.

当 $p = 2$ 时，椭圆控制图是一种较为直接的控制图. 95 %的置信椭圆由满足下面不等式的所有 $\mathbf{X}$ 构成:

$$\left(\mathbf{X} - \overline{\mathbf{X}}\right)' \mathbf{S}^{-1} \left(\mathbf{X} - \overline{\mathbf{X}}\right) \leq \chi_2^2(0.05)$$

例 3.4.2 (例 5.9)

当一个点落在控制椭圆之外时,需对两个分量分别构造一元控制图,进一步分析引起失控的原因.

3. T^2 控制图

一个 T^2 控制图可用来处理大量的特征. 与椭圆控制图不同的是,它并不仅仅局限于两个变量. 此外，与散点图不同的是，它的点是按时间顺序画出的，这样我们从图上可以看到各种变化模式和趋势.

对第 j 个点计算 T^2 统计量

$$T_j^2 = \left(\mathbf{X}_j - \overline{\mathbf{X}}\right)' \mathbf{S}^{-1} \left(\mathbf{X}_j - \overline{\mathbf{X}}\right) \quad j = 1, 2, \ldots, n$$

将 T_j^2 画在时间轴上，控制下限是 0, 控制上限定为 $UCL = \chi_p^2(0.05)$ 或 $UCL = \chi_p^2(0.01)$.

当多元 T^2 控制图发出第 j 个数据落在控制域外面的信号时，就应确定哪些变量要对此负责. 为此，我们经常选择基于庞弗罗尼区间的控制域来解决这一问题. 如果 x_{jk} 不落在区间

$$\left(\overline{X}_k - t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{kk}}{n}}, \overline{X}_k + t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{kk}}{n}}\right)$$

之中，则第 k 个变量失控，此处 p 为所测量变量的总数.

二、未来观测值的预测域

当一个过程稳定时，利用收集到的数据 $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ 来确定某一个未来观测值的预测域. 预期未来观测值将落入其内的区域称为预测域. 若过程稳定，则观测值可视为独立同分布的，且均服从 $N_p(\mu, \mathbf{\Sigma})$.

1. 椭球预测图

定理：设 $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ 独立同分布，且分布为 $N_p(\mu, \mathbf{\Sigma})$,$\mathbf{X}_0$ 为来自相同分布的一个未来观测值，则

$$T^2 = \frac{n}{n+1}\left(\mathbf{X}_0 - \overline{\mathbf{X}}\right)' \mathbf{S}^{-1} \left(\mathbf{X}_0 - \overline{\mathbf{X}}\right) \dot{\sim} \frac{(n-1)\,p}{n-p} F_{p,n-p}$$

且

$$\Pr\left\{\left(\mathbf{X}_0-\overline{\mathbf{X}}\right)'\mathbf{S}^{-1}\left(\mathbf{X}_0-\overline{\mathbf{X}}\right)\le\frac{\left(n^2-1\right)p}{n\left(n-p\right)}F_{p,n-p}\left(\alpha\right)\right\}=1-\alpha$$

证明：因为 $\mathbf{X}_0$ 是未来观测值，所以 $\mathbf{X}_0$ 与历史观测值 $\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_n$ 相互独立.

$$E\left(\mathbf{X}_0-\overline{\mathbf{X}}\right)=E\mathbf{X}_0-E\overline{\mathbf{X}}=0$$

$$Cov\left(\mathbf{X}_0-\overline{\mathbf{X}}\right)=Cov\mathbf{X}_0+Cov\overline{\mathbf{X}}=\mathbf{\Sigma}+\tfrac{1}{n}\mathbf{\Sigma}=\tfrac{n+1}{n}\mathbf{\Sigma}$$

故 $\sqrt{\frac{n}{n+1}}\left(\mathbf{X}_0-\overline{\mathbf{X}}\right)\sim N_p\left(0,\mathbf{\Sigma}\right)$

又因为 $(n-1)S\sim\mathbf{W}_{n-1}(\mathbf{\Sigma})$

所以 $\sqrt{\frac{n}{n+1}}\left(\mathbf{X}_0-\overline{\mathbf{X}}\right)'\left(\frac{(n-1)\mathbf{S}}{n-1}\right)^{-1}\sqrt{\frac{n}{n+1}}\left(\mathbf{X}_0-\overline{\mathbf{X}}\right)\sim\frac{(n-1)p}{n-p}F_{p,n-p}$

即 $T^2=\frac{n}{n+1}\left(\mathbf{X}_0-\overline{\mathbf{X}}\right)'\mathbf{S}^{-1}\left(\mathbf{X}_0-\overline{\mathbf{X}}\right)\sim\frac{(n-1)p}{n-p}F_{p,n-p}$

值得注意的是：在利用当前观测值来确定未来观测值的控制域之前，当前观测值必须是稳定的. 由于

$$\Pr\left\{\left(\mathbf{X}_0-\overline{\mathbf{X}}\right)'\mathbf{S}^{-1}\left(\mathbf{X}_0-\overline{\mathbf{X}}\right)\le\frac{\left(n^2-1\right)p}{n\left(n-p\right)}F_{p,n-p}(\alpha)\right\}=1-\alpha$$

故在获得任何新观测值 $\mathbf{X}_0$ 之前， $\mathbf{X}_0$ 落在预测椭球中的概率为 $1-\alpha$，落在椭球中的任何未来观测值被认为是稳定的或在控制中的. 椭球之外的观测值代表着可能的失控观测值或特殊原因变量.

2. T^2 预测值

对每一个新的观测值 $\mathbf{X}_i$，按时间顺序对

$$T_i^2=\frac{n}{n+1}\left(\mathbf{X}_i-\overline{\mathbf{X}}\right)'\mathbf{S}^{-1}\left(\mathbf{X}_i-\overline{\mathbf{X}}\right)$$

作图，令 $LCL=0\,,UCL=\frac{(n-1)p}{n-p}F_{p,n-p}(0.05)$

若 $T_k^2>UCL$，则认为第 k 个点 $\mathbf{X}_k$ 异常，应对其进行检查，并采取措施.

三、基于子样本均值的控制图

假定来自控制过程的每一个观测值随机向量独立同分布，且分布为 $N_p(\mu,\mathbf{\Sigma})$. 当抽样方法指定选取的单元数 $m>1$ 时，我们将采取与之前不同的处理方法：

第一个单元：	$\mathbf{X}_{11}$	$\mathbf{X}_{12}$	$\cdots$	$\mathbf{X}_{1m}$
第二个单元：	$\mathbf{X}_{21}$	$\mathbf{X}_{22}$	$\cdots$	$\mathbf{X}_{2m}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
第 n 个单元：	$\mathbf{X}_{n1}$	$\mathbf{X}_{n2}$	$\cdots$	$\mathbf{X}_{nm}$

第 i 个单元的样本均值为

$$\overline{\mathbf{X}}_i=\frac{1}{m}\sum_{j=1}^{m}\mathbf{X}_{ij}\ ,\ i=1,2,\ldots,n$$

第 i 个单元的样本协方差矩阵为

$$\mathbf{S}_i = \frac{1}{m-1}\sum_{j=1}^{m}\left(\mathbf{X}_{ij}-\overline{\mathbf{X}}_i\right)\left(\mathbf{X}_{ij}-\overline{\mathbf{X}}_i\right)' \ , i=1,2,\ldots,n$$

总的样本均值为

$$\overline{\overline{\mathbf{X}}} = \frac{1}{n}\sum_{i=1}^{n}\overline{\mathbf{X}}_i = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbf{X}_{ij}$$

总的样本协方差矩阵为

$$\begin{aligned}\mathbf{S} &= \tfrac{1}{nm-n}\left((m-1)\mathbf{S}_1+(m-1)\mathbf{S}_2+\cdots+(m-1)\mathbf{S}_n\right)\\ &= \tfrac{1}{n}\left(\mathbf{S}_1+\mathbf{S}_2+\cdots+\mathbf{S}_n\right)\end{aligned}$$

因为 $(m-1)\mathbf{S}_i \sim \mathbf{W}_{m-1}(\Sigma)$, $i=1,2,\ldots,n$
所以 $(nm-n)\mathbf{S} \sim \mathbf{W}_{nm-n}(\mathbf{\Sigma})$
又因为对于任意的 i , j , $\overline{\mathbf{X}}_i$ 与 $\mathbf{S}_j$ 相互独立
所以 $\overline{\mathbf{X}}_i - \overline{\overline{\mathbf{X}}}$ 与 $\mathbf{S}$ 相互独立
而

$$\begin{aligned}\overline{\mathbf{X}}_i-\overline{\overline{\mathbf{X}}} &= \overline{\mathbf{X}}_i - \tfrac{1}{n}\sum_{j=1}^{n}\overline{\mathbf{X}}_j\\ &= \left(1-\tfrac{1}{n}\right)\overline{\mathbf{X}}_i - \tfrac{1}{n}\overline{\mathbf{X}}_1-\cdots-\tfrac{1}{n}\overline{\mathbf{X}}_{i-1}-\tfrac{1}{n}\overline{\mathbf{X}}_{i+1}-\cdots-\tfrac{1}{n}\overline{\mathbf{X}}_n\end{aligned}$$

故 $E\left(\overline{\mathbf{X}}_i-\overline{\overline{\mathbf{X}}}\right) = E\left(\overline{\mathbf{X}}_i\right)-E\left(\overline{\mathbf{X}}\right)=0$

$$\begin{aligned}Cov\left(\overline{\mathbf{X}}_i-\overline{\overline{\mathbf{X}}}\right) &= \left(1-\tfrac{1}{n}\right)^2 Cov\left(\overline{\mathbf{X}}_i\right)+\tfrac{n-1}{n^2}Cov\left(\overline{\mathbf{X}}_1\right)\\ &= \tfrac{(n-1)^2}{n^2}\cdot\tfrac{1}{m}\mathbf{\Sigma}+\tfrac{n-1}{n^2}\cdot\tfrac{1}{m}\mathbf{\Sigma}\\ &= \tfrac{n-1}{nm}\mathbf{\Sigma}\end{aligned}$$

从而

$$\sqrt{\frac{nm}{n-1}}\left(\overline{\mathbf{X}}_i-\overline{\overline{\mathbf{X}}}\right) \sim N_p(0,\mathbf{\Sigma})$$

根据 T^2 分布的定义知：

$$\begin{aligned}&\sqrt{\tfrac{nm}{n-1}}\left(\overline{\mathbf{X}}_i-\overline{\overline{\mathbf{X}}}\right)'\left(\tfrac{(nm-n)\mathbf{S}}{nm-n}\right)^{-1}\sqrt{\tfrac{nm}{n-1}}\left(\overline{\mathbf{X}}_i-\overline{\overline{\mathbf{X}}}\right)'\\ &\qquad\sim \tfrac{(nm-n)p}{nm-n-p+1}\quad F_{p,nm-n-p+1}\end{aligned}$$

即 $T^2 = \frac{nm}{n-1}\left(\overline{\mathbf{X}}_i-\overline{\mathbf{X}}\right)'\mathbf{S}^{-1}\left(\overline{\mathbf{X}}_i-\overline{\mathbf{X}}\right)' \sim \frac{(nm-n)p}{nm-n-p+1}F_{p,nm-n-p+1}$

由此可得各子样本均值的椭球控制图为：

$$\left\{\overline{\mathbf{X}}:\ \left(\overline{\mathbf{X}}-\overline{\overline{\mathbf{X}}}\right)'\mathbf{S}^{-1}\left(\overline{\mathbf{X}}-\overline{\overline{\mathbf{X}}}\right)\leq\frac{(n-1)(m-1)p}{m(nm-n-p+1)}F_{p,nm-n-p+1}(0.05)\right\}$$

上式右端常用 $\chi_p^2(0.05)/m$ 来近似代替.

对那些落在控制椭球之外的点所对应的子样本应仔细检查，以确定被测量的质量特征的行为是否有变换.

若取 $LCL=0\,,UCL=\frac{(n-1)(m-1)p}{m(nm-n-p+1)}F_{p,nm-n-p+1}(0.05)$

将 $T_j^2=\left(\overline{\mathbf{X}}_j-\overline{\overline{\mathbf{X}}}\right)'\mathbf{S}^{-1}\left(\overline{\mathbf{X}}_j-\overline{\overline{\mathbf{X}}}\right)\quad j=1,2,\ldots,n$ 画在以时间为横坐标的坐标系之中，则得 T^2 控制图.

超过 UCL 的 T_j^2 对应的子样本对应于潜在的失控或存在特殊原因的波动.

四、未来子样本观测值的预测域

一旦数据取自某一稳定运行过程，便可用来构造未来观测子样本均值的预测限. 假设某个未来子样本为 $\mathbf{X}_{n+1,1}\,,\mathbf{X}_{n+1,2},\ldots,\mathbf{X}_{n+1,m},\overline{\mathbf{X}}=\frac{1}{m}\sum\limits_{i=1}^{m}\mathbf{X}_{n+1,i}$ 为未来子样本均值，则 $\overline{\mathbf{X}}-\overline{\overline{\mathbf{X}}}$ 服从多元正态分布，均值为 0, 且

$$\begin{aligned}Cov(\overline{\mathbf{X}}-\overline{\overline{\mathbf{X}}})\ &=Cov\left(\overline{\mathbf{X}}\right)+Cov\left(\overline{\overline{\mathbf{X}}}\right)=\tfrac{1}{m}\mathbf{\Sigma}+\tfrac{1}{nm}\mathbf{\Sigma}\\&=\tfrac{n+1}{nm}\mathbf{\Sigma}\end{aligned}$$

从而

$$\frac{nm}{n+1}\left(\overline{\mathbf{X}}-\overline{\overline{\mathbf{X}}}\right)'\mathbf{S}^{-1}\left(\overline{\mathbf{X}}-\overline{\overline{\mathbf{X}}}\right)\sim\frac{(nm-n)p}{nm-n-p+1}F_{p,nm-n-p+1}$$

对给定的显著性水平 α, 未来子样本均值的 $1-\alpha$ 预测椭球为：

$$\left\{\overline{\mathbf{X}}:\ \left(\overline{\mathbf{X}}-\overline{\overline{\mathbf{X}}}\right)'\mathbf{S}^{-1}\left(\overline{\mathbf{X}}-\overline{\overline{\mathbf{X}}}\right)\leq\frac{(n+1)(m-1)p}{m(nm-n-1)}F_{p,nm-n-p+1}(0.05)\right\}$$

若取 $LCL=0\,,UCL=\frac{(n+1)(m-1)p}{m(nm-n-1)}F_{p,nm-n-p+1}(0.05)\dot{\sim}\chi_p^2(0.05)$

将由未来子样本均值构成的量

$$T^2=\left(\overline{\mathbf{X}}-\overline{\overline{\mathbf{X}}}\right)'S^{-1}\left(\overline{\mathbf{X}}-\overline{\overline{\mathbf{X}}}\right)$$

按时间顺序作图，则得未来子样本均值的 T^2 预测图.

如果某个未来子样本均值落在预测椭球之外或控制上限上方，则意味着当前值在某种程度上不同于先前稳定过程的相应值，这也许是好事，也可能是坏事，但几乎可以肯定地说，我们必须认真地去查明引起这种变换的原因.

第四章　多个多元均值向量的比较

§4.1　成对比较与重复测量设计

为了考察不同实验条件下的响应是否明显不同，人们常常将这些条件下的实验测验值记录下来. 例如，对一种新药的效果的效果或广告宣传的饱和

度, 可通过比较 “处理” (新药与旧药) 前与处理后的测量值来确定. 在另一些情形下, 对相同或相似的实验单元可使用两种或更多种处理, 比较其响应, 便可评价各种处理引起的效应.

用来对两种处理进行比较或用来判断一种处理是否存在的一个合理方法, 就是将两种处理指配给相同的单元, 然后通过计算或对响应之差, 消除单元到单元的外来变差所引起的大部分影响, 从而完成对这些响应的分析.

一、一元成对比较

设 X_{j1} 表示在第 j 个试验下对处理 1 的响应 (或处理前的响应)

X_{j2} 表示在第 j 个试验下对处理 2 的响应 (或处理后的响应)

现选取 n 个实验单元, 令 $t_j = X_{j1} - X_{j2}\,,\ j = 1, 2, \ldots, n$

假设 $t_j \sim N(\mu, \sigma^2),\ j = 1, 2, \ldots, n,$ 且相互独立, 则

$$T = \frac{\overline{t} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

其中

$$\overline{t} = \frac{1}{n}\sum_{i=1}^{n} t_i \quad , \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(t_i - \overline{t}\right)^2$$

如果两种处理无显著性差异, 则应有 $\mu = 0$

下面对假设 $H_0:\ \mu = 0\ ,\ H_1:\ \mu \neq 0$ 进行检验.

对给定的显著性水平 $\alpha > 0,$ 因为:

$$\Pr\left\{|T| \leq t_{n-1}\left(\frac{\alpha}{2}\right)\right\} = 1 - \alpha$$

故 μ 对应的置信度为 $1 - \alpha$ 的置信区间为:

$$\left[\overline{t} - t_{n-1}\left(\frac{\alpha}{2}\right)\frac{s}{\sqrt{n}}\,,\ \overline{t} + t_{n-1}\left(\frac{\alpha}{2}\right)\frac{s}{\sqrt{n}}\right]$$

如果该置信区间包含原点, 则不拒绝 $H_0,$ 否则拒绝 $H_0.$

若置信区间的左端点大于 0, 则认为处理 1 比处理 2 对应的指标大; 若置信区间的右端点小于 0, 则认为处理 1 比处理 2 对应的指标小.

二、 p 元成对比较

下面分析对具有 p 个响应变量的两种不同的处理进行比较.

设 X_{ijk} 表示第 j 个实验单元中第 k 个响应变量的第 i 种处理下的测值.

$$i = 1, 2 \ \ , \ \ j = 1, 2, \ldots, n \ \ , \ \ k = 1, 2, \ldots, p.$$

令 $d_{jk} = X_{1jk} - X_{2jk}\,,\ j = 1, 2, \ldots, n\,,\ k = 1, 2, \ldots, p$

$$\mathbf{d}_j = \begin{pmatrix} d_{j1} \\ d_{j2} \\ \vdots \\ d_{jp} \end{pmatrix}\ ,\ j = 1, 2, \ldots, n$$

设 $E\mathbf{d}_j=\mu\,,\,Cov(\mathbf{d}_j)=\mathbf{\Sigma}$

如果 $\mathbf{d}_1,\ldots,\mathbf{d}_n$ 独立同分布，且服从 $N_p(\mu,\mathbf{\Sigma})$, 则

$$T^2=n\left(\overline{\mathbf{d}}-\mu\right)'\mathbf{S}^{-1}\left(\overline{\mathbf{d}}-\mu\right)\sim\frac{(n-1)\,p}{n-p}F_{p,n-p}$$

其中 $\overline{\mathbf{d}}=\frac{1}{n}\sum_{i=1}^{n}\mathbf{d}_i\,,\,\mathbf{S}=\frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{d}_i-\overline{\mathbf{d}}\right)\left(\mathbf{d}_i-\overline{\mathbf{d}}\right)'$. 故对于给定的显著性水平 $\alpha>0$, 若

$$T^2=n\overline{\mathbf{d}}'\mathbf{S}^{-1}\overline{\mathbf{d}}>\frac{(n-1)\,p}{n-p}F_{p,n-p}(\alpha)$$

则拒绝假设 $H_0:\ \mu=0$, 即认为两种不同的处理之间有显著性差异.

μ 的置信度为 $1-\alpha$ 的置信椭球为：

$$\left\{\mu:\ \left(\overline{\mathbf{d}}-\mu\right)'\mathbf{S}^{-1}\left(\overline{\mathbf{d}}-\mu\right)\leq\frac{(n-1)\,p}{n\,(n-p)}F_{p,n-p}(\alpha)\right\}$$

每个分量 μ_i 的置信度为 $1-\alpha$ 的联合置信区间为：

$$\overline{d}_i\pm\sqrt{\frac{(n-1)\,p}{n\,(n-p)}F_{p,n-p}\,(\alpha)}\sqrt{\frac{s_{ii}}{n}}$$

其中 $\overline{d}_i$ 为向量 $\overline{\mathbf{d}}$ 的第 i 个分量， s_{ii} 为协方差矩阵 $\mathbf{S}$ 的第 i 个对角线上的元素.

每个分量 μ_i 的置信度至少为 $1-\alpha$ 的庞弗罗尼联合置信区间为：

$$\overline{d}_i\pm t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{ii}}{n}}$$

其中 $t_{n-1}\left(\frac{\alpha}{2p}\right)$ 表示自由度为 $n-1$ 的 $\frac{\alpha}{2p}$ 上侧分位数.

三、一个响应变量多种不同处理的比较

一元成对比较的另一种推广，就是在一个响应变量的情况下对 q 种处理进行比较. 设 X_{ji} 表示 j 个实验单元施加第 i 种处理后的效应.

令 $\mathbf{X}_j=\begin{pmatrix}X_{j1}\\X_{j2}\\\vdots\\X_{jq}\end{pmatrix}\quad j=1,2,\ldots,n$

若 $\mathbf{X}_j\sim N_q(\mu,\mathbf{\Sigma})\,,\,j=1,2,\ldots,n$ 且相互独立，则要检验 q 种不同的处理之间是否有差异，只要检验均值向量 μ 的各个分量之间是否相等即可. 为此我们可以构造对比矩阵

$$\mathbf{C}=\begin{pmatrix}1&-1&&&&\\&1&-1&&&\\&&\ddots&\ddots&&\\&&&&1&-1\end{pmatrix}_{(q-1)\times q}$$

或

$$\mathbf{C}=\begin{pmatrix}1&-1&0&\cdots&0\\1&0&-1&\cdots&0\\\vdots&\vdots&\vdots&\ddots&\vdots\\1&0&0&\cdots&-1\end{pmatrix}_{(q-1)\times q}$$

那么假设 $H_0:\ \mu_1=\mu_2=\cdots=\mu_q$ 等价于 $H_0:\ \mathbf{C}\mu=0$
令 $\mathbf{Y}_j=\mathbf{C}\mathbf{X}_j\,,\ j=1,2,\ldots,n,$ 则 $\mathbf{Y}_j\sim N_{q-1}\left(\mathbf{C}\mu,\mathbf{C}\mathbf{\Sigma}\mathbf{C}'\right)$

$$\overline{\mathbf{Y}}=\frac{1}{n}\sum_{j=1}^{n}\mathbf{Y}_j=\frac{1}{n}\sum_{j=1}^{n}\mathbf{C}\mathbf{X}_j=\mathbf{C}\overline{\mathbf{X}}\sim N_{q-1}\left(\mathbf{C}\mu,\frac{1}{n}\mathbf{C}\mathbf{\Sigma}\mathbf{C}'\right)$$

$$\mathbf{S}_{\mathbf{Y}}=\mathbf{C}\mathbf{S}\mathbf{C}'$$

其中 $\mathbf{S}=\frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{X}_i-\overline{\mathbf{X}}\right)\left(\mathbf{X}_i-\overline{\mathbf{X}}\right)'$

根据 T^2 统计量的定义，有

$$T^2=n\left(\mathbf{C}\overline{\mathbf{X}}-\mathbf{C}\mu\right)'\left(\mathbf{C}\mathbf{S}\mathbf{C}'\right)^{-1}\left(\mathbf{C}\overline{\mathbf{X}}-\mathbf{C}\mu\right)$$

$$\sim\frac{(n-1)(q-1)}{n-q-1}F_{q-1,n-q+1}$$

当 H_0 成立时

$$T^2=n\left(\mathbf{C}\overline{\mathbf{X}}\right)'\left(\mathbf{C}\mathbf{S}\mathbf{C}'\right)^{-1}\left(\mathbf{C}\overline{\mathbf{X}}\right)\sim\frac{(n-1)(q-1)}{n-q-1}F_{q-1,n-q+1}$$

对给定的 $\alpha>0,$ 若

$$T^2=n\left(\mathbf{C}\overline{\mathbf{X}}\right)'\left(\mathbf{C}\mathbf{S}\mathbf{C}'\right)^{-1}\left(\mathbf{C}\overline{\mathbf{X}}\right)>\frac{(n-1)(q-1)}{n-q-1}F_{q-1,n-q+1}(\alpha)$$

则拒绝假设 $H_0,$ 即认为不同的处理之间有显著性差异.

对比向量 $\mathbf{C}\mu$ 的置信度为 $1-\alpha$ 的置信椭球为：

$$\left\{\mathbf{C}\mu:\ n\left(\mathbf{C}\overline{\mathbf{X}}-\mathbf{C}\mu\right)'\left(\mathbf{C}\mathbf{S}\mathbf{C}'\right)^{-1}\left(\mathbf{C}\overline{\mathbf{X}}-\mathbf{C}\mu\right)\leq\right.$$

$$\left.\frac{(n-1)(q-1)}{n-q-1}F_{q-1,n-q+1}(\alpha)\right\}$$

单个对比向量 $\mathbf{C}_i'\mu$ 的置信度为 $1-\alpha$ 的联合 T^2 置信区间为：

$$\mathbf{C}_i'\overline{\mathbf{X}}\pm\sqrt{\frac{(n-1)(q-1)}{n-q-1}F_{q-1,n-q+1}(\alpha)}\sqrt{\frac{\mathbf{C}_i'\mathbf{S}\mathbf{C}_i}{n}}\ ,\ i=1,2,\ldots,q-1$$

其中 $\mathbf{C}_i$ 为对比矩阵 $\mathbf{C}$ 中的第 i 行元素组成的向量.

§4.2 两个总体均值向量的比较

考虑两个相互独立的总体 $\mathbf{X}_1$ 和 $\mathbf{X}_2$, 设 $E\mathbf{X}_1=\mu_1\,, Cov\left(\mathbf{X}_1\right)=\mathbf{\Sigma}_1\,, E\mathbf{X}_2=\mu_2\,, Cov\left(\mathbf{X}_2\right)=\mathbf{\Sigma}_2\,,$

$\mathbf{X}_{11},\ldots,\mathbf{X}_{1n_1}$ 是来自总体 $\mathbf{X}_1$ 的样本，样本均值为 $\overline{\mathbf{X}}_1=\frac{1}{n_1}\sum\limits_{i=1}^{n_1}\mathbf{X}_{1i}$, 样本协方差矩阵

$\mathbf{S}_1=\frac{1}{n_1-1}\sum\limits_{i=1}^{n_1-1}\left(\mathbf{X}_{1i}-\overline{\mathbf{X}}_1\right)\left(\mathbf{X}_{1i}-\overline{\mathbf{X}}_1\right)'.$

$\mathbf{X}_{21},\mathbf{X}_{22},\ldots,\mathbf{X}_{2n_2}$ 是来自总体 $\mathbf{X}_2$ 的样本，样本均值为 $\overline{\mathbf{X}}_2=\frac{1}{n_2}\sum\limits_{i=1}^{n_2}\mathbf{X}_{2i}$, 样本协方差矩阵为

$\mathbf{S}_2=\frac{1}{n_2-1}\sum\limits_{i=1}^{n_2-1}\left(\mathbf{X}_{2i}-\overline{\mathbf{X}}_2\right)\left(\mathbf{X}_{2i}-\overline{\mathbf{X}}_2\right)'$. 我们的目的是要对两个总体的均值向量之间的关系作推断. 例如， $\mu_1=\mu_2$ 是否成立？若不成立，问哪些分量不相等？下面分几种情况来讨论.

一、首先假设两个总体 $\mathbf{X}_1$ 和 $\mathbf{X}_2$ 均服从正态分布，且 $\mathbf{\Sigma}_1=\mathbf{\Sigma}_2$, 求假设 $H_0:\ \mu_1=\mu_2$ 的检验统计量

因为所有的样本都来自正态总体，根据正态分布的性质知

$$\overline{\mathbf{X}}_1\sim N_p\left(\mu_1,\frac{1}{n_1}\mathbf{\Sigma}\right),\overline{\mathbf{X}}_2\sim N_p\left(\mu_2,\frac{1}{n_2}\mathbf{\Sigma}\right)$$

从而

$$\overline{\mathbf{X}}_1-\overline{\mathbf{X}}_2\sim N_p\left(\mu_1-\mu_2,\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\mathbf{\Sigma}\right)$$

$$\sqrt{\frac{n_1n_2}{n_1+n_2}}\left(\overline{\mathbf{X}}_1-\overline{\mathbf{X}}_2\right)\sim N_p\left(\sqrt{\frac{n_1n_2}{n_1+n_2}}(\mu_1-\mu_2),\mathbf{\Sigma}\right)$$

又因为 $\left(n_1-1\right)\mathbf{S}_1\sim\mathbf{W}_{n_1-1}\left(\mathbf{\Sigma}\right)\,,\left(n_2-1\right)\mathbf{S}_2\sim\mathbf{W}_{n_2-1}\left(\mathbf{\Sigma}\right)$

所以 $\left(n_1-1\right)\mathbf{S}_1+\left(n_2-1\right)\mathbf{S}_2\sim\mathbf{W}_{n_1+n_2-2}\left(\mathbf{\Sigma}\right)$ 且与 $\overline{\mathbf{X}}_1-\overline{\mathbf{X}}_2$ 相互独立.

根据 T^2 统计量的定义

$$T^2=\frac{n_1n_2}{n_1+n_2}\left(\overline{\mathbf{X}}_1-\overline{\mathbf{X}}_2-(\mu_1-\mu_2)\right)'\left(\frac{\left(n_1-1\right)\mathbf{S}_1+\left(n_2-1\right)\mathbf{S}_2}{n_1+n_2-2}\right)^{-1}\cdot$$

$$\left(\overline{\mathbf{X}}_1-\overline{\mathbf{X}}_2-(\mu_1-\mu_2)\right)$$

与

$$\frac{\left(n_1+n_2-2\right)p}{n_1+n_2-p-1}F_{p,n_1+n_2-p-1}$$

同分布.

1. 假设 $H_0:\ \mu_1=\mu_2$ 的检验

记 $\mathbf{S}_p=\frac{\left(n_1-1\right)\mathbf{S}_1+\left(n_2-1\right)\mathbf{S}_2}{n_1+n_2-2}$, 则当假设 $H_0:\ \mu_1=\mu_2$ 成立时,

$$T^2=\frac{n_1n_2}{n_1+n_2}\left(\overline{\mathbf{X}}_1-\overline{\mathbf{X}}_2\right)'\mathbf{S}_p^{-1}\left(\overline{\mathbf{X}}_1-\overline{\mathbf{X}}_2\right)$$

$$\sim\frac{\left(n_1+n_2-2\right)p}{n_1+n_2-p-1}F_{p,n_1+n_2-p-1}$$

故对给定的显著性水平 $\alpha>0$, 当

$$T^2=\frac{n_1n_2}{n_1+n_2}\left(\overline{\mathbf{X}}_1-\overline{\mathbf{X}}_2\right)'\mathbf{S}_p^{-1}\left(\overline{\mathbf{X}}_1-\overline{\mathbf{X}}_2\right)$$

$$>\frac{(n_1+n_2-2)\,p}{n_1+n_2-p-1}F_{p,n_1+n_2-p-1}(\alpha)$$

时，拒绝 H_0, 即认为两个正态总体的均值向量之间有显著性差异.

2. $\mu_1-\mu_2$ 的置信度为 $1-\alpha$ 的置信椭球为：

$$\left\{\mu_1-\mu_2:\frac{n_1n_2}{n_1+n_2}\left\{\overline{\mathbf{X}}_1-\overline{\mathbf{X}}_2-(\mu_1-\mu_2)\right\}'\mathbf{S}_p^{-1}\cdot\right.$$

$$\left.\left\{\overline{\mathbf{X}}_1-\overline{\mathbf{X}}_2-(\mu_1-\mu_2)\right\}\le c^2\right\}$$

其中 $c^2=\frac{(n_1+n_2-2)\,p}{n_1+n_2-p-1}F_{p,n_1+n_2-p-1}(\alpha)$, 置信椭球的中心在 $\overline{\mathbf{X}}_1-\overline{\mathbf{X}}_2$, 各主轴分别为 $\pm\sqrt{\lambda_i}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}ce_i$, 其中 $\mathbf{S}_p\mathbf{e}_i=\lambda_i\mathbf{e}_i$, $i=1,2,\ldots,p$.

3. T^2 联合置信区间

对任意 p 维向量 $\mathbf{a}$,$\mathbf{a}'\left(\overline{\mathbf{X}}_1-\overline{\mathbf{X}}_2\right)\pm c\sqrt{\mathbf{a}'\mathbf{S}_p\mathbf{a}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$

将以概率 $1-\alpha$ 包含 $\mathbf{a}'(\mu_1-\mu_2)$. 特别地,

$$A_i\triangleq\left\{\overline{X}_{1i}-\overline{X}_{2i}\pm c\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\sqrt{s_{ii,p}}\right\}\ \ i=1,2,\ldots,p$$

以概率 $1-\alpha$ 包含 $\mu_{1i}-\mu_{2i}$. 其中 $s_{ii,p}$ 表示矩阵 $\mathbf{S}_p$ 的第 i 个对角线上的元素.

如果区间 A_i 包含原点，则认为 μ_1 与 μ_2 的第 i 个分量之间无显著性差异. 如果区间 A_i 的左端点大于 0, 则认为 $\mu_{1i}>\mu_{2i}$; 如果区间 A_i 的右端点小于 0, 则认为 $\mu_{1i}<\mu_{2i}$.

4. 庞弗罗尼联合置信区间

由前面的讨论知，两个均值向量的 p 个分量差的 $1-\alpha$ 庞弗罗尼联合置信区间为：

$$\mu_{1i}-\mu_{2i}:\left(\overline{X}_{1i}-\overline{X}_{2i}\right)\pm t_{n_1+n_2-2}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\cdot\sqrt{s_{ii,p}}$$

$i=1,2,\ldots,p$

其中 $t_{n_1+n_2-2}\left(\frac{\alpha}{2p}\right)$ 是自由度为 n_1+n_2-2 的 t 分布的 $\frac{\alpha}{2p}$ 上侧分位数.

二、当 $\mathbf{\Sigma}_1\ne\mathbf{\Sigma}_2$ 时，求假设 $H_0:\ \mu_1=\mu_2$ 的检验统计量

1. $n_1=n_2=n$

令 $\mathbf{Y}=\mathbf{X}_1-\mathbf{X}_2$, 则 $\mathbf{Y}_i=\mathbf{X}_{1i}-\mathbf{X}_{2i}$, $i=1,2,\ldots,n$, $\overline{\mathbf{Y}}=\overline{\mathbf{X}}_1-\overline{\mathbf{X}}_2$

因为 $\mathbf{X_1}$ 与 $\mathbf{X_2}$ 相互独立，所以

$\mathbf{Y}_i\sim N_p\left(\mu_1-\mu_2,\mathbf{\Sigma}_1+\mathbf{\Sigma}_2\right)$, $i=1,2,\ldots,n$

$$\overline{\mathbf{Y}} \sim N_p\left(\mu_1-\mu_2, \frac{1}{n}\left(\mathbf{\Sigma}_1+\mathbf{\Sigma}_2\right)\right)$$

$$\sqrt{n}\overline{\mathbf{Y}} \sim N_p\left(\sqrt{n}(\mu_1-\mu_2), \mathbf{\Sigma}_1+\mathbf{\Sigma}_2\right)$$

$$\mathbf{S_Y}=\frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{Y}_i-\overline{\mathbf{Y}}\right)\left(\mathbf{Y}_i-\overline{\mathbf{Y}}\right)'$$

$(n-1)\,\mathbf{S_Y} \sim \mathbf{W}_{n-1}\left(\mathbf{\Sigma}_1+\mathbf{\Sigma}_2\right)$ 且与 $\overline{\mathbf{Y}}$ 相互独立

根据 T^2 统计量的定义知

$$\begin{aligned} T^2 &= n\left[\overline{\mathbf{Y}}-(\mu_1-\mu_2)\right]'\mathbf{S_Y}^{-1}\left[\overline{\mathbf{Y}}-(\mu_1-\mu_2)\right] \\ &= n\left[\left(\overline{\mathbf{X}}_1-\overline{\mathbf{X}}_2\right)-(\mu_1-\mu_2)\right]'\mathbf{S_Y}^{-1}\left[\left(\overline{\mathbf{X}}_1-\overline{\mathbf{X}}_2\right)-(\mu_1-\mu_2)\right] \end{aligned}$$

与 $\frac{(n-1)p}{n-p}F_{p,n-p}$ 同分布.

对任意给定的 $\alpha>0$, 当

$$n\left(\overline{\mathbf{X}}_1-\overline{\mathbf{X}}_2\right)'\mathbf{S_Y}^{-1}\left(\overline{\mathbf{X}}_1-\overline{\mathbf{X}}_2\right) > \frac{(n-1)\,p}{n-p}F_{p,n-p}\left(\alpha\right)$$

时, 拒绝 H_0, 即认为两个总体的均值向量之间有显著性差异.

2. $n_1 \neq n_2$, 不妨设 $n_1<n_2$

令 $\mathbf{Y}_j=\mathbf{X}_{1j}-\sqrt{\frac{n_1}{n_2}}\mathbf{X}_{2j}+\frac{1}{\sqrt{n_1n_2}}\sum_{k=1}^{n_1}\mathbf{X}_{2k}-\frac{1}{n_2}\sum_{k=1}^{n_2}\mathbf{X}_{2k}, j=1,2,\cdots,n_1,$

则 $\mathbf{Y}_j(j=1,2,\cdots,n_1)$ 是正态随机变量的线性组合.

因为

$$\begin{aligned} E\mathbf{Y}_j &= E\mathbf{X}_{1j}-\sqrt{\frac{n_1}{n_2}}E\mathbf{X}_{2j}+\frac{1}{\sqrt{n_1n_2}}\sum_{k=1}^{n_1}E\mathbf{X}_{2k}-\frac{1}{n_2}\sum_{k=1}^{n_2}E\mathbf{X}_{2k} \\ &= \mu_1-\sqrt{\frac{n_1}{n_2}}\mu_2+\frac{1}{\sqrt{n_1n_2}}\cdot n_1\mu_2-\mu_2 \\ &= \mu_1-\mu_2 \end{aligned}$$

$$
\begin{aligned}
CovY_j &= Cov\mathbf{X}_{1j} + Cov\left[\left(\frac{1}{\sqrt{n_1 n_2}} - \frac{1}{n_2} - \sqrt{\frac{n_1}{n_2}}\right)\mathbf{X}_{2j}\right] \\
&+ \left(\frac{1}{\sqrt{n_1 n_2}} - \frac{1}{n_2}\right)^2 Cov\left(\sum_{\substack{k=1 \\ k\neq j}}^{n_1} X_{2k}\right) \\
&+ \frac{1}{n_2^2}\sum_{k=n_1+1}^{n_2} Cov\left(X_{2k}\right) \\
&= \mathbf{\Sigma}_1 + \left(\frac{1}{\sqrt{n_1 n_2}} - \frac{1}{n_2} - \sqrt{\frac{n_1}{n_2}}\right)^2 \mathbf{\Sigma}_2 \\
&+ \left(\frac{1}{\sqrt{n_1 n_2}} - \frac{1}{n_2}\right)^2 \cdot (n_1 - 1)\,\mathbf{\Sigma}_2 \\
&+ \frac{1}{n_2^2}(n_2 - n_1)\,\mathbf{\Sigma}_2 \\
&= \mathbf{\Sigma}_1 + \frac{n_1}{n_2}\mathbf{\Sigma}_2
\end{aligned}
$$

所以

$$
\mathbf{Y}_j \sim N_p\left(\mu_1 - \mu_2, \mathbf{\Sigma}_1 + \frac{n_1}{n_2}\mathbf{\Sigma}_2\right) \quad , j = 1, 2, \ldots, n_1
$$

根据 T^2 统计量的定义知：

$$
\begin{aligned}
T^2 &= n_1\left\{\overline{\mathbf{Y}} - (\mu_1 - \mu_2)\right\}' \mathbf{S_Y}^{-1}\left\{\overline{\mathbf{Y}} - (\mu_1 - \mu_2)\right\} \\
&\sim \frac{(n_1 - 1)\,p}{n_1 - p} F_{p, n_1 - p}
\end{aligned}
$$

其中 $\mathbf{S_Y} = \frac{1}{n_1 - 1}\sum_{i=1}^{n}\left(\mathbf{Y}_i - \overline{\mathbf{Y}}\right)\left(\mathbf{Y}_i - \overline{\mathbf{Y}}\right)'$

对于给定的 $\alpha > 0$, 当

$$
n_1 \overline{\mathbf{Y}}' \mathbf{S_Y}^{-1} \overline{\mathbf{Y}} > \frac{(n_1 - 1)\,p}{n_1 - p} F_{p, n_1 - p}(\alpha)
$$

时，拒绝 H_0, 即认为两个总体的均值向量之间有显著性差异.

类似地，可讨论 $\mu_1 - \mu_2$ 地置信域.

如果总体的分布不是正态分布，但当样本容量 n_1 和 n_2 相对于 p 来说都很大时，我们可以用大样本方法对两个均值向量进行检验.

结论 4.1　设样本容量使 $n_1 - p$ 与 $n_2 - p$ 都很大，此时 $\mu_1 - \mu_2$ 地置信度为 $1 - \alpha$ 的近似置信椭球为：

$$
\begin{aligned}
&\left\{\mu_1 - \mu_2 : \left[\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)\right]' \left(\frac{\mathbf{S}_1}{n_1} + \frac{\mathbf{S}_2}{n_2}\right)^{-1} \cdot \right. \\
&\left. \left[\left(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2\right) - (\mu_1 - \mu_2)\right] \leq \chi_p^2(\alpha)\right\}
\end{aligned}
$$

且所有线性组合 $\mathbf{a}'(\mu_1-\mu_2)$ 的置信度为 $1-\alpha$ 的 T^2 联合置信区间为：

$$\mathbf{a}'\left(\overline{\mathbf{X}}_1-\overline{\mathbf{X}}_2\right)\pm\sqrt{\chi_p^2(\alpha)}\cdot\sqrt{\mathbf{a}'\left(\frac{1}{n_1}\mathbf{S}_1+\frac{1}{n_2}\mathbf{S}_2\right)\mathbf{a}}$$

证明：因为 $\overline{\mathbf{X}}_1$ 与 $\overline{\mathbf{X}}_2$ 相互独立，所以

$$E\left(\overline{\mathbf{X}}_1-\overline{\mathbf{X}}_2\right)=\mu_1-\mu_2$$

$$Cov\left(\overline{\mathbf{X}}_1-\overline{\mathbf{X}}_2\right)=Cov\left(\overline{\mathbf{X}}_1\right)+Cov\left(\overline{\mathbf{X}}_2\right)=\tfrac{1}{n_1}\mathbf{\Sigma}_1+\tfrac{1}{n_2}\mathbf{\Sigma}_2$$

根据中心极限定理：

$$\overline{\mathbf{X}}_1-\overline{\mathbf{X}}_2\dot{\sim}N_p\left(\mu_1-\mu_2,\frac{1}{n_1}\mathbf{\Sigma}_1+\frac{1}{n_2}\mathbf{\Sigma}_2\right)$$

从而

$$\left[\left(\overline{\mathbf{X}}_1-\overline{\mathbf{X}}_2\right)-(\mu_1-\mu_2)\right]'\left[\frac{1}{n_1}\mathbf{\Sigma}_1+\frac{1}{n_2}\mathbf{\Sigma}_2\right]^{-1}\cdot$$
$$\left[\left(\overline{\mathbf{X}}_1-\overline{\mathbf{X}}_2\right)-(\mu_1-\mu_2)\right]\dot{\sim}\chi_p^2$$

又因为当 n_1 , n_2 很大时， $\mathbf{S}_1$ 和 $\mathbf{S}_2$ 依概率收敛于 $\mathbf{\Sigma}_1$ 和 $\mathbf{\Sigma}_2$.

故用 $\mathbf{S}_1$ 和 $\mathbf{S}_2$ 分别替换 $\mathbf{\Sigma}_1$ 和 $\mathbf{\Sigma}_2$ 时，上式仍成立，即

$$\left[\left(\overline{\mathbf{X}}_1-\overline{\mathbf{X}}_2\right)-(\mu_1-\mu_2)\right]'\left[\frac{1}{n_1}\mathbf{S}_1+\frac{1}{n_2}\mathbf{S}_2\right]^{-1}\cdot$$
$$\left[\left(\overline{\mathbf{X}}_1-\overline{\mathbf{X}}_2\right)-(\mu_1-\mu_2)\right]\dot{\sim}\chi_p^2$$

§4.3 单因素多个总体均值向量的比较

在实际应用中，我们常常需要一对两个以上的总体进行比较，例如，某响应变量受因素 A 的影响. 当因素 A 取不同的水平时，得到响应变量的不同样本观测值，通常的问题是：根据样本观测值考察因素 A 的不同水平对响应变量的影响是否有显著性差异.

下面我们分只有一个响应变量和 p 个响应变量的情况分别讨论.

一、一元单因素方差分析

我们假设因素 A 有 a 个不同的水平 $A_1,A_2,\ldots,A_a$. 在第 i 个水平 A_i 下作独立重复试验得响应变量的观测值为： $X_{i1},X_{i2},\ldots,X_{in_i}$, $i=1,2,\ldots,a$, 且不同水平下响应变量的观测值相互独立. 如果我们把样本 $X_{i1},X_{i2},\ldots,X_{in_i}$ 看作是来自总体 X_i 的样本，且 $X_i\sim N\left(\mu_i,\sigma^2\right)$, $i=1,\ldots,a$, 则比较不同水平下的响应变量是否有显著性差异就归结为检验假设：

$$H_0:\ \mu_1=\mu_2=\cdots=\mu_a$$

是否成立. 若假设 H_0 成立，则认为 A 的不同水平之间无显著性差异.

现将上面的分析写成如下的形式

$$\begin{cases} X_{ij} = \mu_i + \varepsilon_{ij} \\ i = 1, 2, \ldots, a \ , \ j = 1, 2, \ldots, n_i \\ \varepsilon_{ij} \sim N(0, \sigma^2) \ \text{且相互独立} \end{cases} \tag{4.3.1}$$

我们称 (4.3.1) 为单因素方差分析模型.

如果令 $\mu = \frac{1}{n} \sum\limits_{i=1}^{a} \sum\limits_{j=1}^{n_i} E\left(X_{ij}\right) = \frac{1}{n} \sum\limits_{i=1}^{a} n_i \mu_i \ , \ n = n_1 + n_2 + \cdots + n_a$

$$\alpha_i = \mu_i - \mu \ , \ \text{易验证} \ \sum_{i=1}^{a} n_i \alpha_i = 0$$

则 (4.3.1) 可写成

$$\begin{cases} X_{ij} = \mu + \alpha_i + \varepsilon_{ij} \\ i = 1, 2, \ldots, a \ , \ j = 1, 2, \ldots, n_i \\ \varepsilon_{ij} \sim N(0, \sigma^2) \ \text{且相互独立} \\ \sum\limits_{i=1}^{a} n_i \alpha_i = 0 \end{cases} \tag{4.3.2}$$

由 α_i 的定义知，假设 $H_0 : \ \mu_1 = \mu_2 = \cdots = \mu_a$ 等价于假设

$$H_0 : \ \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$$

下面我们通过平方和分解的方式导出检验 H_0 的统计量.

首先我们引入记号 $\overline{X} = \frac{1}{n} \sum\limits_{i=1}^{a} \sum\limits_{j=1}^{n_i} X_{ij} \ , \ \overline{X}_i = \frac{1}{n} \sum\limits_{j=1}^{n_i} X_{ij}$

$$\begin{aligned} SS_T &\triangleq \sum_{i=1}^{a} \sum_{j=1}^{n_i} \left(X_{ij} - \overline{X}\right)^2 \\ &= \sum_{i=1}^{a} \sum_{j=1}^{n_i} \left(X_{ij} - \overline{X}_i + \overline{X}_i - \overline{X}\right)^2 \\ &= \sum_{i=1}^{a} \sum_{j=1}^{n_i} \left(X_{ij} - \overline{X}_i\right)^2 + \sum_{i=1}^{a} \sum_{j=1}^{n_i} \left(\overline{X}_i - \overline{X}\right)^2 \\ &= \sum_{i=1}^{a} \sum_{j=1}^{n_i} \left(X_{ij} - \overline{X}_i\right)^2 + \sum_{i=1}^{a} n_i \left(\overline{X}_i - \overline{X}\right)^2 \\ &\triangleq SS_E + SS_A \end{aligned}$$

其中 $SS_T = \sum\limits_{i=1}^{a} \sum\limits_{j=1}^{n_i} \left(X_{ij} - \overline{X}\right)^2$ 表示总的变差平方和，$SS_A = \sum\limits_{i=1}^{a} n_i \left(\overline{X}_i - \overline{X}\right)^2$ 表示因素 A 的不同水平引起的变差平方和，$SS_E = \sum\limits_{i=1}^{a} \sum\limits_{j=1}^{n_i} \left(X_{ij} - \overline{X}_i\right)^2$ 表示除了因素 A 之外的其它随机因素引起的变差平方和.

可以证明, 当 H_0 成立时, $\frac{SS_A}{\sigma^2} \sim \chi^2_{a-1}$, $\frac{SS_E}{\sigma^2} \sim \chi^2_{n-a}$, 且二者相互独立, 从而

$$F = \frac{SS_A/(a-1)}{SS_E/(n-a)} \sim F_{a-1,n-a}$$

根据 SS_A 与 SS_E 的定义知, 当 H_0 成立时, F 的值较小, 故 F 可作为检验 H_0 的统计量.

对给定的显著性水平 α, 当

$$F = \frac{SS_A/(a-1)}{SS_E/(n-a)} > F_{a-1,n-a}(\alpha)$$

时, 拒绝 H_0, 即认为因素 A 的不同水平之间有显著性差异.

等价地, 当 $p = \Pr\left\{F > \frac{SS_A/(a-1)}{SS_E/(n-a)}\right\} < \alpha$ 时, 拒绝 H_0, 其中 F 表示自由度分别为 $a-1$ 和 $n-a$ 的 F 随机变量.

一般地, 将上述计算过程列成表格 - 通常计算机软件的输出结果, 我们称之为一元方差分析表.

表 4.3.1　一元方差分析表

变差来源	平方和	自由度	均方和	F 统计量	p 值
因素A	SS_A	$a-1$	$MS_A = SS_A/(a-1)$	$F = \frac{MS_A}{MS_E}$	$\Pr\left\{F > \frac{MS_A}{MS_E}\right\}$
误差	SS_E	$n-a$	$MS_E = SS_E/(n-a)$		
总和	SS_T	$n-1$			

二、 p 元单因素方差分析

与一元单因素方差分析类似, 只是将单个响应变量变成 p 个响应变量. 设 $\mathbf{X}_i \sim N_p(\mu_i, \mathbf{\Sigma})$, $i = 1, 2, \ldots, a$, 且相互独立, 其中 $\mathbf{X}_i$ 和 μ_i 均为 p 维向量.

$\mathbf{X}_{i1}, \mathbf{X}_{i2}, \ldots, \mathbf{X}_{in_i}$ 是来自总体 $\mathbf{X}_i$ 的样本, 其模型如下:

$$\begin{cases} \mathbf{X}_{ij} = \mu + \alpha_i + \varepsilon_{ij} \\ i = 1, 2, \ldots, a\ ,\ j = 1, 2, \ldots, n_i \\ \varepsilon_{ij} \sim N_p(\mathbf{0}, \mathbf{\Sigma}) \text{ 且相互独立} \end{cases} \tag{4.3.3}$$

其中 $\mu = \frac{1}{n}\sum_{i=1}^{a} n_i\mu_i$, $n = \sum_{i=1}^{a} n_i$, $\alpha_i = \mu_i - \mu$, 我们称 (4.3.3) 为 p 元单因素方差分析模型.

检验假设 $H_0:\ \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$

因为 $\sum\limits_{i=1}^{a}\sum\limits_{j=1}^{n_i}\left(\mathbf{X}_{ij}-\overline{\mathbf{X}}\right)\left(\mathbf{X}_{ij}-\overline{\mathbf{X}}\right)'$

$$=\sum_{i=1}^{a}\sum_{j=1}^{n_i}\left(\mathbf{X}_{ij}-\overline{\mathbf{X}}_i+\overline{\mathbf{X}}_i-\overline{\mathbf{X}}\right)\left(\mathbf{X}_{ij}-\overline{\mathbf{X}}_i+\overline{\mathbf{X}}_i-\overline{\mathbf{X}}\right)'$$

$$=\sum_{i=1}^{a}\sum_{j=1}^{n_i}\left(\mathbf{X}_{ij}-\overline{\mathbf{X}}_i\right)\left(\mathbf{X}_{ij}-\overline{\mathbf{X}}_i\right)'+\sum_{i=1}^{a}n_i\left(\overline{\mathbf{X}}_i-\overline{\mathbf{X}}\right)\left(\overline{\mathbf{X}}_i-\overline{\mathbf{X}}\right)'$$

$$\triangleq \mathbf{W}+\mathbf{B}$$

其中 $\overline{\mathbf{X}}_i=\frac{1}{n_i}\sum\limits_{j=1}^{n_i}\mathbf{X}_{ij}$, $\overline{\mathbf{X}}=\frac{1}{n}\sum\limits_{i=1}^{a}\sum\limits_{j=1}^{n_i}\mathbf{X}_{ij}$, $n=\sum\limits_{i=1}^{a}n_i$

$$\mathbf{W}=\sum_{i=1}^{a}\sum_{j=1}^{n_i}\left(\mathbf{X}_{ij}-\overline{\mathbf{X}}_i\right)\left(\mathbf{X}_{ij}-\overline{\mathbf{X}}_i\right)',\ \mathbf{B}=\sum_{i=1}^{a}n_i\left(\overline{\mathbf{X}}_i-\overline{\mathbf{X}}\right)^2$$

可以证明，当 H_0 成立时，

$$\mathbf{W}\sim\mathbf{W}_{n-a}\left(\mathbf{\Sigma}\right)$$

$$\mathbf{B}\sim\mathbf{W}_{a-1}\left(\mathbf{\Sigma}\right)$$

且相互独立.

威尔克斯最早提出了检验假设 H_0 的统计量为：

$$\Lambda=\frac{|\mathbf{W}|}{|\mathbf{B}+\mathbf{W}|}=\frac{|\sum\limits_{i=1}^{a}\sum\limits_{j=1}^{n_i}\left(\mathbf{X}_{ij}-\overline{\mathbf{X}}_i\right)\left(\mathbf{X}_{ij}-\overline{\mathbf{X}}_i\right)'|}{|\sum\limits_{i=1}^{a}\sum\limits_{j=1}^{n_i}\left(\mathbf{X}_{ij}-\overline{\mathbf{X}}\right)\left(\mathbf{X}_{ij}-\overline{\mathbf{X}}\right)'|}$$

我们称之为威尔克斯统计量，其中 $|\mathbf{W}|$ 表示矩阵 $\mathbf{W}$ 的行列式. 当 Λ 较小时，拒绝 H_0, 即认为因素 A 的不同水平之间有显著性差异.

威尔克斯 Λ 统计量的优点是使用方便，而且与似然比准则密切相关，缺点是不易确定临界值. 下表针对一些特殊的 a 和 p, 给出了 Λ 的精确分布.

表 4.3.2　威尔克斯 $\Lambda=\frac{|\mathbf{W}|}{|\mathbf{B}+\mathbf{W}|}$ 的分布

p	a	抽样分布
1	≥ 2	$\frac{n-a}{a-1}\cdot\frac{1-\Lambda}{\Lambda}\sim F_{a-1,n-a}$
2	≥ 2	$\frac{n-a-1}{a-1}\cdot\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\sim F_{2(a-1),2(n-a-1)}$
≥ 1	2	$\frac{n-p-1}{p}\cdot\frac{1-\Lambda}{\Lambda}\sim F_{p,n-p-1}$
≥ 1	3	$\frac{n-p-2}{p}\cdot\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\sim F_{2p,2(n-p-2)}$

对于其他情形及大样本情形，经巴特利特修正的 Λ 可用于检验 H_0.

巴特利特证明了，若假设 H_0 成立，且 $\sum\limits_{i=1}^{a}n_i=n$ 充分大，则

$$-\left(n-1-\frac{a+p}{2}\right)\ln\Lambda=-\left(n-1-\frac{a+p}{2}\right)\ln\left(\frac{|\mathbf{W}|}{|\mathbf{B}+\mathbf{W}|}\right)$$

$$\dot{\sim}\chi^2_{p(a-1)}$$

故当 n 充分大时，若

$$-\left(n-1-\frac{a+p}{2}\right)\ln\left(\frac{|\mathbf{W}|}{|\mathbf{B}+\mathbf{W}|}\right)>\chi^2_{p(a-1)}(\alpha)$$

则以显著性水平 α 拒绝 H_0, 其中 $\chi^2_{p(a-1)}(\alpha)$ 表示自由度为 $p(a-1)$ 的 χ^2 分布的 α 上侧分位数.

三、不同水平效应的联合置信区间

当假设 H_0 被拒绝时，我们自然对导致拒绝假设的那些水平感兴趣. 对成对比较，用庞弗罗尼方法可构造 $\alpha_i-\alpha_j$(等价地 $\mu_i-\mu_j$) 的分量的联合置信区间，这些区间比用别的方法构造的区间要窄，而且只要求一元 t 分布的临界值.

结论 4.2　对模型 (4.3.3), 对所有的分量 $k=1,2,\ldots,p$ 及所有的差值 $\alpha_{ik}-\alpha_{jk}$, $i,j=1,2,\ldots,a$, 以至少为 $1-\alpha$ 的置信度有 $\alpha_{ik}-\alpha_{jk}=\mu_{ik}-\mu_{jk}$ 属于区间

$$\overline{X}_{ik}-\overline{X}_{jk}\pm t_{n-a}\left(\frac{\alpha}{pa(a-1)}\right)\sqrt{\frac{W_{ii}}{n-a}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}\qquad(4.3.4)$$

其中的 W_{ii} 表示矩阵 $\mathbf{W}$ 的第 i 个对角元素.

当区间 (4.3.4) 包含原点时, 认为 μ_{ik} 与 μ_{jk} 无显著性差异; 当 (4.3.4) 的左端点大于零时, 认为 $\mu_{ik}>\mu_{jk}$; 当 (4.3.4) 的右端点小于零时, 认为 $\mu_{ik}<\mu_{jk}$. 通过比较，可以找出相对较大的分量与较小的分量. 从而寻找最优水平.

§4.4　双因素多个总体均值向量的比较

假定响应变量值是在两个因素的不同的水平下得到的，且各种不同水平组合下的响应变量值相互独立. 设因素 A 有 a 个不同的水平, 因素 B 有 b 个不同的水平. 我们在各组不同的水平组合下做 c 次独立重复试验, 获得 c 各独立观测值. 用 X_{ijk} 表示在因素 A 的第 i 个水平下, 因素 B 的第 j 个水平下得到的第 k 个观测值，且假设这些值均服从正态分布，即 $X_{ijk}\sim N_p(\mu_{ij},\Sigma)$, $i=1,2,\ldots,a$, $j=1,2,\ldots,b$, $k=1,2,\ldots,c$. 下面我们分两种情况来讨论.

一、一元双因素方差分析

当只有一个响应变量值时，我们用 X_{ijk} 表示在水平组合 (A_i,B_j) 下指标观测值，　$i=1,2,\ldots,a$, $j=1,2,\ldots,b$, $k=1,2,\ldots,c$, 这时可将 X_{ijk} 写成如下形式：

$$\begin{cases}X_{ijk}=\mu_{ij}+\varepsilon_{ijk}\\ i=1,2,\ldots,a\ ,\ j=1,2,\ldots,b\ ,\ k=1,2,\ldots,c\\ \varepsilon_{ijk}\sim N(0,\sigma^2)\ \text{且相互独立}\end{cases}\qquad(4.4.1)$$

我们称 (4.4.1) 为双因素方差分析模型.

为了做统计分析，我们将 μ_{ij} 作适当的分解，为此，引入

$$\mu = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\mu_{ij}$$

$$\overline{\mu}_{i\cdot} = \frac{1}{b}\sum_{j=1}^{b}\mu_{ij}$$

$$\overline{\mu}_{\cdot j} = \frac{1}{a}\sum_{i=1}^{a}\mu_{ij}$$

$$\alpha_i = \overline{\mu}_{i\cdot} - \mu \ , \ i = 1, \ldots, a$$

$$\beta_j = \overline{\mu}_{\cdot j} - \mu \ , \ j = 1, \ldots, b$$

$$\gamma_{ij} = \mu_{ij} - \overline{\mu}_{i\cdot} - \overline{\mu}_{\cdot j} + \mu = (\mu_{ij} - \mu) - \alpha_i - \beta_j \ , \ i = 1, \ldots, a \, , \, j = 1, \ldots, b$$

其中 μ 为总平均，α_i 为因素 A 的水平 A_i 的效应，β_j 为因素 B 的水平 B_j 的效应．$\mu_{ij} - \mu$ 为水平组合 (A_i, B_j) 对指标值的效应．在许多情况下，$\mu_{ij} - \mu$ 并不等于水平 A_i 的效应 α_i 和 B_j 的效应 β_j 之和．γ_{ij} 为 A_i 和 B_j 的交互效应．通常把因素 A 和 B 对试验指标的交互效应设想为某一个因素的效应，称这个因素为 A 与 B 的交互作用，记为 $A \times B$. 易验证，

$$\sum_{i=1}^{a}\alpha_i = 0 \, , \, \sum_{j=1}^{b}\beta_j = 0 \, , \, \sum_{i=1}^{a}\sum_{j=1}^{b}\gamma_{ij} = 0$$

于是我们可以将 (4.4.1) 写成如下形式：

$$\begin{cases} X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \\ i = 1, 2, \ldots, a \ , \ j = 1, 2, \ldots, b \ , \ k = 1, 2, \ldots, c \\ \varepsilon_{ijk} \sim N(0, \sigma^2) \ \text{且相互独立} \\ \sum\limits_{i=1}^{a}\alpha_i = 0 \, , \, \sum\limits_{j=1}^{b}\beta_j = 0 \, , \, \sum\limits_{i=1}^{a}\sum\limits_{j=1}^{b}\gamma_{ij} = 0 \end{cases} \tag{4.4.2}$$

1. 无交互效应的情形

   假设 $\gamma_{ij} = 0 \, , \, i = 1, \ldots, a \, , \, j = 1, \ldots, b$, 即不存在交互效应．为简单计，我们只考虑 $c = 1$ 的情形，对 $c > 1$ 的情形，统计分析方法完全相同，此时 (4.4.2) 又可写成：

$$\begin{cases} X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \\ i = 1, 2, \ldots, a \ , \ j = 1, 2, \ldots, b \\ \varepsilon_{ij} \sim N(0, \sigma^2) \ \text{且相互独立} \\ \sum\limits_{i=1}^{a}\alpha_i = 0 \, , \, \sum\limits_{j=1}^{b}\beta_j = 0 \end{cases} \tag{4.4.3}$$

这就是无交互效应的两因素方差分析模型.

我们的目的是考察因素 A 或 B 的各水平对指标的影响是否有显著性差异，这归结为对假设为：

$$\text{或} \quad \begin{array}{l} H_1:\ \alpha_1 = \cdots = \alpha_a = 0 \\ H_2:\ \beta_1 = \cdots = \beta_b = 0 \end{array}$$

的检验. 下面采用与单因素方差分析模型类似的方法导出检验统计量.

$$\text{记} \quad \overline{X} = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b} X_{ij} \qquad \overline{X}_{i\cdot} = \frac{1}{b}\sum_{j=1}^{b} X_{ij}$$

$$\overline{X}_{\cdot j} = \frac{1}{a}\sum_{i=1}^{a} X_{ij} \qquad SS_T = \sum_{i=1}^{a}\sum_{j=1}^{b}\left(X_{ij} - \overline{X}\right)^2$$

其中 SS_T 为全部试验数据的总变差，称为总平方和，对其进行分解.

$$\begin{aligned} SS_T &= \sum_{i=1}^{a}\sum_{j=1}^{b}\left(X_{ij} - \overline{X}\right)^2 \\ &= \sum_{i=1}^{a}\sum_{j=1}^{b}\left[\left(X_{ij} - \overline{X}_{i\cdot} - \overline{X}_{\cdot j} + \overline{X}\right) + \left(\overline{X}_{i\cdot} - \overline{X}\right) + \left(\overline{X}_{\cdot j} - \overline{X}\right)\right]^2 \\ &= \sum_{i=1}^{a}\sum_{j=1}^{b}\left(X_{ij} - \overline{X}_{i\cdot} - \overline{X}_{\cdot j} + \overline{X}\right)^2 + \sum_{i=1}^{a} b\left(\overline{X}_{i\cdot} - \overline{X}\right)^2 \\ &\quad + a\sum_{j=1}^{b}\left(\overline{X}_{\cdot j} - \overline{X}\right)^2 \\ &\triangleq SS_E + SS_A + SS_B \end{aligned}$$

其中 SS_A 和 SS_B 分别表示因素 A 和因素 B 的不同水平引起的差异，分别称为因素 A 或因素 B 的平方和. SS_E 表示除了因素 A 和因素 B 之外，剩下的随机因素所引起的变差.

可以证明，当 H_1 成立时，$SS_A/\sigma^2 \sim \chi^2_{a-1}$, 且与 SS_E 相互独立，而 $SS_E/\sigma^2 \sim \chi^2_{(a-1)(b-1)}$, 于是当 H_1 成立时：

$$F_A = \frac{SS_A/(a-1)}{SS_E/[(a-1)(b-1)]} \sim F_{a-1,(a-1)(b-1)}$$

故对给定显著性水平 α, 当 $F_A > F_{a-1,(a-1)(b-1)}(\alpha)$ 时，我们拒绝假设 H_1, 即认为因素 A 的不同水平的效应有显著性差异.

同理，当 H_2 成立时

$$F_B = \frac{SS_B/(b-1)}{SS_E/[(a-1)(b-1)]} \sim F_{b-1,(a-1)(b-1)}$$

故对给定显著性水平 α, 当 $F_B > F_{b-1,(a-1)(b-1)}(\alpha)$ 时，我们拒绝假设 H_2, 即认为因素 B 的不同水平的效应有显著性差异.

与单因素情形一样，我们可以写出方差分析表，如表 4-4-1.

表 4-4-1　无交互效应的双因素方差分析表

变差来源	平方和	自由度	均方和	F 比	p 值
因素A	SS_A	$a-1$	$MS_A=\frac{SS_A}{a-1}$	$F_A=\frac{MS_A}{MS_E}$	$p_1=\Pr\{F_1>F_A\}$
因素B	SS_B	$b-1$	$MS_B=\frac{SS_B}{b-1}$	$F_B=\frac{MS_B}{MS_E}$	$p_2=\Pr\{F_2>F_B\}$
误差	SS_E	$(a-1)(b-1)$	$MS_E=\frac{SS_E}{(a-1)(b-1)}$		
总和	SS_T	$ab-1$			

其中 F_1 表示自由度分别为 $a-1$ 与 $(a-1)(b-1)$ 的 F 统计量，F_2 表示自由度分别为 $b-1$ 与 $(a-1)(b-1)$ 的 F 统计量. 当 p 值小于 α 时，拒绝假设.

如果经过 F_A 检验，H_1 被拒绝，则认为因素 A 的不同水平效应 $\alpha_1,\ldots,\alpha_a$ 不全相同鸵辉□讲罘治鲆谎我们希望比较 α_i 的大小，为此对 $\alpha_i-\alpha_j$ 作区间估计.

不难推出，$m=\dfrac{a(a-1)}{2}$ 个效应之差 $\alpha_i-\alpha_j$ 的置信度为 $1-\alpha$ 的庞弗罗尼置信区间为

$$\overline{X}_{i\cdot}-\overline{X}_{j\cdot}\pm\hat{\sigma}\sqrt{\frac{2}{b}}t_{(a-1)(b-1)}\left(\frac{\alpha}{2m}\right)\qquad(4.4.4)$$

其中 $\hat{\sigma}^2=\frac{SS_E}{(a-1)(b-1)}$.

同理，$m=\dfrac{b(b-1)}{2}$ 个效应之差 $\beta_i-\beta_j$ 的置信度为 $1-\alpha$ 的庞弗罗尼置信区间为

$$\overline{X}_{\cdot i}-\overline{X}_{\cdot j}\pm\hat{\sigma}\sqrt{\frac{2}{a}}t_{(a-1)(b-1)}\left(\frac{\alpha}{2m}\right)\qquad(4.4.5)$$

2. 关于交互效应的检验

在无交互效应时，对因素 A 、 B 各水平的每种组合只进行一次试验，即 $c=1$, 仍然可以进行统计检验. 而当要考虑因素 A 、 B 之间的交互作用 $A\times B$ 时，在各水平组合下需要做重复试验. 设每种组合下的试验次数均为 $c(c>1)$. 此时对应的统计模型即为 (4.3.2). 在该模型中，效应 α_i 并不能反映水平 A_i 的优劣，这是因为在交互效应存在的情况下，因子水平 A_i 的优劣还与因子 B 的水平有关系. 对不同的 B_j,A_i 的优劣也不相同. 因此，对这样的模型，检验 $\alpha_1=\cdots=\alpha_a=0$ 与检验 $\beta_1=\cdots=\beta_b=0$ 都是没有实际意义的，然而一个重要的检验问题是交互效应是否存在的检验，即检验

$$H_3:\ \gamma_{ij}=0\ ,\ i=1,\ldots,a\ ,\ j=1,\ldots,b$$

与前面方法类似，引进

$$\overline{X} = \frac{1}{abc}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c} X_{ijk}$$

$$\overline{X}_{ij\cdot} = \frac{1}{c}\sum_{k=1}^{c} X_{ijk}\ ,\ i = 1,\ldots,a\ ,\ j = 1,\ldots,b$$

$$\overline{X}_{i\cdot\cdot} = \frac{1}{bc}\sum_{j=1}^{b}\sum_{k=1}^{c} X_{ijk}\ ,\ i = 1,\ldots,a$$

$$\overline{X}_{\cdot j\cdot} = \frac{1}{ac}\sum_{i=1}^{a}\sum_{k=1}^{c} X_{ijk}\ ,\ j = 1,\ldots,b$$

作平方和分解，得

$$\begin{aligned} SS_T &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\left(X_{ijk} - \overline{X}\right)^2 \\ &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\left[\left(X_{ijk} - \overline{X}_{ij\cdot}\right) + \left(\overline{X}_{i\cdot\cdot} - \overline{X}\right)\right. \\ &\left. + \left(\overline{X}_{\cdot j\cdot} - \overline{X}\right) + \left(\overline{X}_{ij\cdot} - \overline{X}_{i\cdot\cdot} - \overline{X}_{\cdot j\cdot} + \overline{X}\right)\right]^2 \\ &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\left(X_{ijk} - \overline{X}_{ij\cdot}\right)^2 + bc\sum_{i=1}^{a}\left(\overline{X}_{i\cdot\cdot} - \overline{X}\right)^2 \\ &+ ac\sum_{j=1}^{b}\left(\overline{X}_{\cdot j\cdot} - \overline{X}\right)^2 \\ &+ c\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\overline{X}_{ij\cdot} - \overline{X}_{i\cdot\cdot} - \overline{X}_{\cdot j\cdot} + \overline{X}\right)^2 \\ &\stackrel{\triangle}{=} SS_E + SS_A + SS_B + SS_{A\times B} \end{aligned}$$

与前面的讨论类似，可以证明，当 H_3 成立时

$$F_{A\times B} = \frac{\frac{SS_{A\times B}}{(a-1)(b-1)}}{\frac{SS_E}{ab(c-1)}} \sim F_{(a-1)(b-1),ab(c-1)}$$

对给定的显著性水平 $\alpha > 0$, 当 $F_{A\times B} > F_{(a-1)(b-1),ab(c-1)}(\alpha)$ 时，我们拒绝假设 H_3, 即认为因素 A 与因素 B 之间的交互作用存在. 当 $F_{A\times B} <$

$F_{(a-1)(b-1),ab(c-1)}(\alpha)$ 时，我们接受假设 H_3, 即认为因素 A 与因素 B 之间的交互作用不存在. 这时我们可以进一步检验因素 A 和因素 B 的各水平之间的效应是否有差异，方法同无交互效应的情形.

本节讨论的几种检验可以归纳成表 4-4-2.

表 4-4-12　有交互效应的双因素方差分析表

变差来源	平方和	自由度	均方和	F 比	p 值
因素A	SS_A	$a-1$	$MS_A=\frac{SS_A}{a-1}$	$F_A=\frac{MS_A}{MS_E}$	$p_1=P\{F_1>F_A\}$
因素B	SS_B	$b-1$	$MS_B=\frac{SS_B}{a-1}$	$F_B=\frac{MS_B}{MS_E}$	$p_2=P\{F_2>F_B\}$
交互作用 $A\times B$	$SS_{A\times B}$	$(a-1)(b-1)$	$MS_{A\times B}=\frac{SS_{A\times B}}{(a-1)(b-1)}$	$F_{A\times B}=\frac{MS_{A\times B}}{MS_E}$	$p_3=P\{F_3>F_{A\times B}\}$
误差	SS_E	$ab(c-1)$	$MS_E=\frac{SS_E}{ab(c-1)}$		
总和	SS_T	$abc-1$			

其中 F_1 表示自由度分别为 $a-1$ 与 $ab(c-1)$ 的 F 随机变量;

F_2 表示自由度分别为 $b-1$ 与 $ab(c-1)$ 的 F 随机变量;

F_3 表示自由度分别为 $(a-1)(b-1)$ 与 $ab(c-1)$ 的 F 随机变量.

二、 p 元双因素方差分析

用类比方法, 我们可以对由 p 个响应变量组成的随机向量建立模型如下：

$$
\begin{cases}
\mathbf{X}_{ijk}=\mu+\alpha_i+\beta_j+\gamma_{ij}+\varepsilon_{ijk} \\
i=1,2,\ldots,a\ ,\ j=1,2,\ldots,b\ ,\ k=1,2,\ldots,c \qquad (4.4.6) \\
\varepsilon_{ijk}\sim N_p(0,\mathbf{\Sigma})\ \text{且相互独立} \\
\sum\limits_{i=1}^{a}\alpha_i=0\,,\ \sum\limits_{j=1}^{b}\beta_j=0\,,\ \sum\limits_{i=1}^{a}\sum\limits_{j=1}^{b}\gamma_{ij}=0
\end{cases}
$$

其中 $\mu = \frac{1}{abc}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c} E\left(\mathbf{X}_{ijk}\right) = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\mu_{ij}$

$\overline{\mu}_{i\cdot} = \frac{1}{b}\sum_{j=1}^{b}\mu_{ij}$

$\overline{\mu}_{\cdot j} = \frac{1}{a}\sum_{i=1}^{a}\mu_{ij}$

$\alpha_i = \overline{\mu}_{i\cdot} - \mu\ ,\ i = 1, \ldots, a$

$\beta_j = \overline{\mu}_{\cdot j} - \mu\ ,\ j = 1, \ldots, b$

$\gamma_{ij} = \mu_{ij} - \mu - \alpha_i - \beta_j\ ,\ i = 1, \ldots, a\ ,\ j = 1, \ldots, b$

且均为 p 维向量. 若记

$\overline{\mathbf{X}} = \frac{1}{abc}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\mathbf{X}_{ijk}$

$\overline{\mathbf{X}}_{ij\cdot} = \frac{1}{c}\sum_{k=1}^{c}\mathbf{X}_{ijk}\ ,\ i = 1, \ldots, a\ ,\ j = 1, \ldots, b$

$\overline{\mathbf{X}}_{i\cdot\cdot} = \frac{1}{bc}\sum_{j=1}^{b}\sum_{k=1}^{c}\mathbf{X}_{ijk}\ ,\ i = 1, \ldots, a$

$\overline{\mathbf{X}}_{\cdot j\cdot} = \frac{1}{ac}\sum_{i=1}^{a}\sum_{k=1}^{c}\mathbf{X}_{ijk}\ ,\ j = 1, \ldots, b$

则 $\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\left(\mathbf{X}_{ijk}-\overline{\mathbf{X}}\right)\left(\mathbf{X}_{ijk}-\overline{\mathbf{X}}\right)'$

$$=\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\left(\mathbf{X}_{ijk}-\overline{\mathbf{X}}_{ij\cdot}\right)\left(\mathbf{X}_{ijk}-\overline{\mathbf{X}}_{ij\cdot}\right)'$$

$$+bc\sum_{i=1}^{a}\left(\overline{\mathbf{X}}_{i\cdot\cdot}-\overline{\mathbf{X}}\right)\left(\overline{\mathbf{X}}_{i\cdot\cdot}-\overline{\mathbf{X}}\right)'+ac\sum_{j=1}^{b}\left(\overline{\mathbf{X}}_{\cdot j\cdot}-\overline{\mathbf{X}}\right)\left(\overline{\mathbf{X}}_{\cdot j\cdot}-\overline{\mathbf{X}}\right)'$$

$$+c\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\overline{\mathbf{X}}_{ij\cdot}-\overline{\mathbf{X}}_{i\cdot\cdot}-\overline{\mathbf{X}}_{\cdot j\cdot}+\overline{\mathbf{X}}\right)\left(\overline{\mathbf{X}}_{ij\cdot}-\overline{\mathbf{X}}_{i\cdot\cdot}-\overline{\mathbf{X}}_{\cdot j\cdot}+\overline{\mathbf{X}}\right)'$$

$$\triangleq SSP_E+SSP_A+SSP_B+SSP_{A\times B}$$

从上可以看出，从一元分析到多元分析的推广只不过是简单地将诸如 $\left(\overline{X}_{i\cdot\cdot}-\overline{X}\right)^2$ 这样的标量用相应的矩阵 $\left(\overline{\mathbf{X}}_{i\cdot\cdot}-\overline{\mathbf{X}}\right)\left(\overline{\mathbf{X}}_{i\cdot\cdot}-\overline{\mathbf{X}}\right)'$ 替代而已.

假设

$$H_1:\ \gamma_{ij}=0\ ,\ i=1,\ldots,a\ ,\ j=1,\ldots,b$$

的检验统计量为：

$$\Lambda_1^*=\frac{|SSP_E|}{|SSP_{A\times B}+SSP_E|}$$

若 Λ^* 较小，则拒绝 H_1. 对于大样本情形，上面的威尔克斯 Λ 统计量的临界值可以用 χ^2 分布的临界值来近似，即对于给定的显著性水平 $\alpha>0$, 若

$$-\left[ab\,(c-1)-\frac{p+1-(a-1)\,(b-1)}{2}\right]\ln\Lambda_1^*>\chi^2_{(a-1)(b-1)p}\,(\alpha)$$

则拒绝 H_1, 否则接受 H_1.

当假设 H_1 成立时，我们可以进一步检验因素 A 和因素 B 的效应. 首先考虑假设：

$$H_2:\ \alpha_1=\alpha_2=\cdots=\alpha_a=0$$

当 $\Lambda_2^*=\frac{|SSP_E|}{|SSP_A+SSP_E|}$ 具有较小值时，就拒绝 H_2.

对于大样本情形，对给定的显著性水平 $\alpha>0$, 若

$$-\left[ab\,(c-1)-\frac{p+1-(a-1)}{2}\right]\ln\Lambda_2^*>\chi^2_{(a-1)p}\,(\alpha)$$

则拒绝 H_2.

同理，对假设：

$$H_3:\ \beta_1=\beta_2=\cdots=\beta_b=0$$

当 $\Lambda_3^*=\frac{|SSP_E|}{|SSP_B+SSP_E|}$ 具有较小值时，就拒绝 H_3.

对于大样本情形，对给定的显著性水平 α, 若

$$-\left[ab\left(c-1\right)-\frac{p+1-\left(b-1\right)}{2}\right]\ln\Lambda_3^*>\chi_{(b-1)p}^2\left(\alpha\right)$$

则拒绝 H_3.

第五章　线性回归模型

回归分析是一种通过一组预测变量（自变量）来预测一个或多个响应变量（因变量）的统计方法. 在本章中，我们首先讨论预测一个响应变量的多重回归，然后将这个模型推广到预测若干个因变量的多元回归.

§5.1　一元多重线性回归

设因变量 Y 与 $p-1$ 个变量 $x_1,x_2,\ldots,x_{p-1}$ 有相关关系，则 Y 可表示成

$$Y=f\left(x_1,\ldots,x_{p-1}\right)+\varepsilon \tag{5.1.1}$$

其中 ε 为误差项，它是一个随机变量，且假定 $\varepsilon\sim N(0,\sigma^2)$

若 $f(x_1,\ldots,x_{p-1})=\beta_0+\beta_1x_1+\ldots+\beta_{p-1}x_{p-1}$, 则 (5.1.1) 可以改写成

$$\begin{cases} Y=\beta_0+\beta_1x_1+\ldots+\beta_{p-1}x_{p-1}+\varepsilon \\ \varepsilon\sim N\left(0,\sigma^2\right) \end{cases} \tag{5.1.2}$$

称 (5.1.2) 为一元多重(p 重)正态线性回归模型, ε 称为误差项, $\beta_0,\beta_1,\ldots,\beta_{p-1}$ 为未知参数，β_0 称为常数项，$\beta_1,\ldots,\beta_{p-1}$ 称为回归系数. 对模型 (5.1.2), 我们主要讨论以下三个问题：

1. 求未知参数 $\beta_0,\beta_1,\ldots,\beta_{p-1}$ 的估计值 $\hat{\beta}_0,\ldots,\hat{\beta}_{p-1}$, 以建立线性回归方程

$$\hat{Y}=\hat{\beta}_0+\hat{\beta}_1x_1+\ldots+\hat{\beta}_{p-1}x_{p-1}$$

2. 对回归方程进行合理性检验
3. 利用回归方程对未来进行预测

一、未知参数的最小二乘估计

为了求得模型 (5.1.2) 中未知参数的估计，现进行 $n(n>p)$ 次独立重复观测，得到 n 组观测值

$$(x_{i1},x_{i2},\ldots,x_{i,p-1},Y_i)\qquad i=1,2,\ldots,n$$

它们满足

$$\begin{cases} Y_i=\beta_0+x_{i1}\beta_1+\ldots+x_{i,p-1}\beta_{p-1}+\varepsilon_i \\ i=1,2,\ldots,n \\ \varepsilon_i\sim N(0,\sigma^2)\text{ 且相互独立} \end{cases} \tag{5.1.3}$$

$$\text{若记}\quad \mathbf{Y}=\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix},\qquad \mathbf{X}=\begin{pmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ 1 & x_{21} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{n,p-1} \end{pmatrix}$$

$$\beta=\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix},\quad \varepsilon=\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

将 (5.1.3) 写成矩阵的形式为：

$$\begin{cases} \mathbf{Y}=\mathbf{X}\beta+\varepsilon \\ \\ \varepsilon\sim N_n(0,\sigma^2\mathbf{I}) \end{cases} \tag{5.1.4}$$

获得参数向量 β 的估计的一个重要方法是最小二乘法. 这个方法是找 β 的估计，使得偏差向量 $\varepsilon=\mathbf{Y}-\mathbf{X}\beta$ 的长度的平方 $\parallel \mathbf{Y}-\mathbf{X}\beta \parallel^2$ 达到最小. 设 $Q(\beta)=\parallel \mathbf{Y}-\mathbf{X}\beta \parallel^2$

定义 5.1.1 若存在 $\hat{\beta}$, 使得

$$Q(\hat{\beta})=\min_{\beta} Q(\beta)$$

则称 $\hat{\beta}$ 为参数向量 β 的最小二乘估计.

$$\begin{aligned} \text{因为}\quad Q(\beta) &= \parallel \mathbf{Y}-\mathbf{X}\beta \parallel^2 \\ &= (\mathbf{Y}-\mathbf{X}\beta)'(\mathbf{Y}-\mathbf{X}\beta) \\ &= \mathbf{Y}'\mathbf{Y}-2\mathbf{Y}'\mathbf{X}\beta+\beta'\mathbf{X}'\mathbf{X}\beta \end{aligned}$$

对 β 求偏导数，并令其等于零，得到方程组

$$\mathbf{X}'\mathbf{X}\beta=\mathbf{X}'\mathbf{Y} \tag{5.1.5}$$

称它为正则方程. 这个线性方程组有唯一解的充要条件是 $\mathbf{X}'\mathbf{X}$ 的秩为 p, 等价地， $rank(\mathbf{X})=p$. 今后，在线性回归模型的讨论中，我们总假定这个条件满足，故 (5.1.5) 存在唯一解

$$\hat{\beta}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{5.1.6}$$

可以证明： $\hat{\beta}$ 确实使 $Q(\beta)$ 达到最小，故 $\hat{\beta}$ 为参数 β 的最小二乘估计. 由 $\hat{\beta}$ 可建立回归方程

$$\hat{Y}=\hat{\beta}_0+\hat{\beta}_1x_1+\ldots+\hat{\beta}_{p-1}x_{p-1}$$

为对该方程进行合理性检验，我们首先讨论 $\hat{\beta}$ 的性质.

二、最小二乘估计的性质

无论从理论上还是从应用上讲，最小二乘估计都是最重要的估计，其原因是这种估计具有许多优良性质.

定理 5.1.1　对于线性回归模型 (5.1.4), 最小二乘估计 $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ 具有性质：

1. $E\hat{\beta} = \beta$
2. $Cov(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

证明：　1. 因为 $E\mathbf{Y} = \mathbf{X}\beta$, 所以

$$E(\hat{\beta}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{Y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta = \beta$$

2. 因为 $Cov(\mathbf{Y}) = \sigma^2\mathbf{I}$, 所以

$$\begin{aligned} Cov(\hat{\beta}) &= Cov\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\right] \\ &= (\mathbf{X}'\mathbf{X})\mathbf{X}'Cov(\mathbf{Y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

定理证毕.

这个定理的第一条结论表明，最小二乘估计 $\hat{\beta}$ 是 β 的无偏估计，这就是说，我们的估计没有系统性偏差. 在实际问题中，当我们用 $\hat{\beta}$ 去估计 β 时，有时可能会偏高，有时则可能会偏低，但这些正负偏差在概率上平均起来等于零.

在线性回归模型 (5.1.4) 中，还有一个重要参数 σ^2, 它是模型误差项的方差，σ^2 反映了模型误差以及观测误差的大小，在回归分析中起着重要的作用，现在我们讨论 σ^2 的估计问题.

误差向量 $\varepsilon = \mathbf{Y} - \mathbf{X}\beta$, 它是一个不可观测的随机向量，用最小二乘估计 $\hat{\beta}$ 代替其中的 β 得到

$$\hat{\varepsilon} = \mathbf{Y} - \mathbf{X}\hat{\beta}$$

称为残差向量. 我们将 $\hat{\varepsilon}$ 看成误差向量 ε 的一个估计，很自然我们用

$$RSS = \hat{\varepsilon}'\hat{\varepsilon} = \sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

来衡量 σ^2 的大小，这里 RSS 是残差平方和（ Residual Sum of Squares, 简记为 RSS ），它的大小反映了实际数据与理论模型 (5.1.4) 的偏离程度或者说拟合程度. RSS 愈小，数据与模型拟合得愈好. 下面的定理给出了 RSS 的一个有用表达式以及利用 RSS 构造 σ^2 的无偏估计.

定理 5.1.2 $(1)RSS = \mathbf{Y}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y} = \mathbf{Y}'\mathbf{H}\mathbf{Y}$

$(2)\hat{\sigma}^2 = \frac{RSS}{n-p}$是$\sigma^2$的无偏估计

证明：(1) $RSS = \hat{\varepsilon}'\hat{\varepsilon} = (\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta})$

$= [(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y}]'[(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y}]$

$= \mathbf{Y}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y} = \mathbf{Y}'\mathbf{H}\mathbf{Y}$

H具有性质： $(a)\mathbf{H}^2 = \mathbf{H}$;

$(b)\mathbf{H}' = \mathbf{H}$;

$(c)\mathbf{X}'\mathbf{H} = 0, \mathbf{H}\mathbf{X} = 0.$

(2) 因为$E\mathbf{Y} = \mathbf{X}\beta, Cov(\mathbf{Y}) = \sigma^2\mathbf{I}$,所以

$E(RSS) = E[\mathbf{Y}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y}]$

$= \beta'\mathbf{X}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{X}\beta + \sigma^2 tr(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')$

$= \sigma^2[n - tr(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')]$

又因为 $tr\mathbf{A}\mathbf{B} = tr\mathbf{B}\mathbf{A}$, 所以

$$tr[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = tr[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}] = tr\mathbf{I}_p = p$$

于是

$$E(RSS) = \sigma^2(n-p)$$

即

$$E(\frac{RSS}{n-p}) = \sigma^2$$

故 $\hat{\sigma}^2 = \frac{RSS}{n-p}$ 是 σ^2 的无偏估计. 定理证毕.

在定理 5.1.1 和定理 5.1.2 中，我们只利用了 $E\mathbf{Y} = \mathbf{X}\beta$, $Cov\mathbf{Y} = \sigma^2\mathbf{I}$, 并未利用性质 $\mathbf{Y} \sim N_n(\mathbf{X}\beta, \sigma^2\mathbf{I})$, 但当 $\mathbf{Y} \sim N_n(\mathbf{X}\beta, \sigma^2\mathbf{I})$ 等价地 $\varepsilon \sim N_n(0, \sigma^2\mathbf{I})$ 时， $\hat{\beta}$ 和 $\hat{\sigma}^2$ 具有更好的性质，它们在后续讨论中具有更重要的作用.

定理 5.1.3 对于线性回归模型 (5.1.4), 有

(1) $\hat{\beta} \sim N_p(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$;

(2) $\frac{RSS}{\sigma^2} \sim \chi^2_{n-p}$;

(3) $\hat{\beta}$与RSS相互独立;

(4) $F = \frac{(\hat{\beta} - \beta)'(\mathbf{X}'\mathbf{X})(\hat{\beta} - \beta)/p}{RSS/(n-p)} \sim F_{p,n-p}$;

(5) β的置信水平为$1-\alpha$的置信椭球为：

$$\{\beta : (\beta - \hat{\beta})'(\mathbf{X}'\mathbf{X})(\beta - \hat{\beta}) \leq \frac{p}{n-p} \cdot RSS \cdot F_{p,n-p}(\alpha)\};$$

该椭球以 $\hat{\beta}$ 为中心，其方向与大小均由 $\mathbf{X}'\mathbf{X}$ 的特征向量和特征值所决定.

(6) β_i , $i = 1, 2, \ldots, p-1$ 的联合 T^2 置信区间为：

$$\hat{\beta}_i \pm \sqrt{\hat{Var}\beta_i} \cdot \sqrt{pF_{p,n-p}(\alpha)}.$$

这里 $\hat{Var}(\beta_i)$ 为矩阵 $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$ 对应于 $\hat{\beta}_i$ 的对角元

$$\begin{aligned} RSS &= \mathbf{Y}'\mathbf{H}\mathbf{Y} = (\mathbf{X}\beta+\varepsilon)'\mathbf{H}(\mathbf{X}\beta+\varepsilon) \\ &= \beta' X'\mathbf{H}(\mathbf{X}\beta+\varepsilon) + \varepsilon'\mathbf{H}\mathbf{X}\beta + \varepsilon'\mathbf{H}\varepsilon \\ &= \varepsilon'\mathbf{H}\varepsilon \end{aligned}$$

(TH. 若 $\mathbf{X} \sim N_p(0, \mathbf{I}), \mathbf{A}$ 为 $p \times p$ 阶对称矩阵，且 $rank(\mathbf{A}) = r$, 则当 $\mathbf{A}$ 满足 $\mathbf{A}^2 = \mathbf{A}$ 时，二次型 $\mathbf{X}'\mathbf{A}\mathbf{X} \sim \chi^2(r)$.)

因为 $rank(H) = n - p$, 所以 $RSS/\sigma^2 \sim \chi^2(n-p)$.

由于定理的证明用到较多的数学知识，故从略.

三、一般线性假设的显著性检验

由前面的讨论知，只要求得参数 β 的最小二乘估计 $\hat{\beta}$, 就可以建立回归方程. 但是，所建立的回归方程是否真正刻画了因变量和自变量之间实际的依赖关系呢？这一方面，也许是最重要的方面，是要把回归方程拿到实践中去考察；另一方面，我们可以做统计假设检验，这叫做回归方程的显著性检验. 另外，我们有时还希望研究因变量是否真正依赖一个或几个特定的自变量，这就导致了相应的回归系数的显著性检验.

我们首先讨论比较一般的线性假设

$$H: \ \mathbf{A}\beta = \mathbf{b} \tag{5.1.7}$$

的检验问题，这里 $\mathbf{A}$ 为 $m \times p$ 阶矩阵， $rank(\mathbf{A}) = m$; $\mathbf{b}$ 为 $m \times 1$ 维已知向量. 我们将会看到，实际应用中许多感兴趣的问题都可归结为形如的假设的检验问题. 我们先提出检验方法的基本思想，对模型 (5.1.4) 应用最小二乘法，对应的残差平方和为：

$$RSS = (\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta}) = \mathbf{Y}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y} \triangleq \mathbf{Y}'\mathbf{H}\mathbf{Y}$$

这里 $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ 为 β 的最小二乘估计. RSS 反映了实际数据与模型 (5.1.4) 拟合的程度. RSS 愈小表示数据与模型拟合得愈好. 现在，在模型 (5.1.4) 上加上线性假设 (5.1.7), 再应用最小二乘法，获得约束最小二乘估计

$$\hat{\beta}_H = \hat{\beta} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}')^{-1}(\mathbf{A}\hat{\beta} - \mathbf{b}) \tag{5.1.8}$$

相应的残差平方和

$$RSS_H = (\mathbf{Y} - \mathbf{X}\hat{\beta}_H)'\mathbf{Y} - \mathbf{X}\hat{\beta}_H)$$

很明显，加了约束条件 (5.1.7), 模型参数 β 的变化范围缩小了，因而残差平方和 RSS_H 就要变大，于是总有 $RSS_H \geq RSS$. 如果真正的参数确实满足约束条件 (5.1.7), 那么加上约束条件跟不加约束条件本质上是一样的. 这时对无约束模型和有约束模型，数据拟合的程度也应该一样. 因而刻画拟合程度的残差平方和之差 $RSS_H - RSS$ 应该比较小. 反过来，若真正的参数不满足

约束 (5.1.7), 则 $RSS_H - RSS$ 倾向于比较大，因此，当 $RSS_H - RSS$ 比较大时，我们就拒绝假设 (5.1.7), 否则就接受它. 在统计学上当我们谈到一个量大小时，往往有一个比较标准. 对现在的情况，我们把比较的标准取为 RSS, 于是用统计量 $\frac{RSS_H - RSS}{RSS}$ 的大小来决定是否拒绝假设 (5.1.7).

定理 5.1.4 对于模型 (5.1.4)

(1) $\frac{RSS}{\sigma^2} \sim \chi^2_{n-p}$;

(2) 若假设(5.1.7)成立, 则$\frac{RSS_H - RSS}{\sigma^2} \sim \chi^2_m$; 其中 $m = rank(\mathbf{A})$

(3) RSS与$RSS_H - RSS$相互独立;

(4) 当假设(5.1.7)成立时

$$F_H = \frac{(RSS_H - RSS)/m}{RSS/(n-p)} \sim F_{m,n-p} \tag{5.1.9}$$

这里 $F_{m,n-p}$ 表示自由度为 $m, n-p$ 的 F 分布.

我们也可以从另外一个角度导出检验 H 的统计量.

因为 $\hat{\beta} \sim N_p(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$

所以 $\mathbf{A}\hat{\beta} - \mathbf{b} \sim N_m(\mathbf{A}\beta - b, \sigma^2\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}')$

从而 $(\mathbf{A}\hat{\beta} - \mathbf{A}\beta)'[\sigma^2\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\hat{\beta} - \mathbf{A}\beta) \sim \chi^2_m$

又因为 $\frac{RSS}{\sigma^2} \sim \chi^2_{n-p}$ 且与 $\hat{\beta}$ 相互独立.

根据 F 的定义，当假设 H 成立时,

$$F_H = \frac{(\mathbf{A}\hat{\beta} - \mathbf{b})'(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}')^{-1}(\mathbf{A}\hat{\beta} - \mathbf{b})/m}{RSS/(n-p)} \tag{5.1.10}$$

$$\sim F_{m,n-p}$$

对给定的显著性水平 $\alpha > 0$,, 记 $F_{m,n-p}(\alpha)$ 为相应的 F 分布的上侧 α 分位数点，即

$$\Pr\{F_{m,n-p} > F_{m,n-p}(\alpha)\} = \alpha$$

则当 $F_H > F_{m,n-p}(\alpha)$ 时，我们就拒绝假设 H, 否则不拒绝假设 H. (5.1.9) 与 (5.1.10) 本质上相同.

例 5.1.1设 $\quad Y_1 = \beta_1 + \varepsilon_1$

$$Y_2 = 2\beta_1 - \beta_2 + \varepsilon_2$$

$$Y_3 = \beta_1 + 2\beta_2 + \varepsilon_3$$

我们欲检验 $H: \beta_1 = \beta_2$

将观测数据写成矩阵的形式

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 2 & -1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix}$$

假设 H 等价于

$$(1,-1)\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}=0$$

易求 $\beta=(\beta_1,\beta_2)'$ 的最小二乘估计

$$\hat{\beta}=\begin{pmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{pmatrix}=\begin{pmatrix}\frac{1}{6}(Y_1+2Y_2+Y_3)\\ \frac{1}{5}(-Y_2+2Y_3)\end{pmatrix}$$

$$\begin{aligned}RSS &= \mathbf{Y}'\mathbf{Y}-\hat{\beta}'\mathbf{X}'\mathbf{Y}=\mathbf{Y}'\mathbf{Y}-\hat{\beta}'(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\\ &= \mathbf{Y}'\mathbf{Y}-\hat{\beta}'\,(\mathbf{X}'\mathbf{X})\,\hat{\beta}\\ &= \textstyle\sum_{i=1}^{3}Y_i^2-6\hat{\beta}_1^2-5\hat{\beta}_2^2\end{aligned}$$

为了求在假设 $H:\beta_1=\beta_2$ 下的残差平方和 RSS_H, 我们将 $\beta_1=\beta_2$(记它们的公共值为 α) 代入原来模型，得到约简模型

$$\begin{pmatrix}Y_1\\ Y_2\\ Y_3\end{pmatrix}=\begin{pmatrix}1\\ 1\\ 3\end{pmatrix}\alpha+\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \varepsilon_3\end{pmatrix}$$

由此得 α 的最小二乘估计（也就是 β_1 和 β_2 的约束最小二乘估计）

$$\hat{\alpha}=\hat{\beta}_H=\frac{1}{11}(Y_1+Y_2+3Y_3)$$

于是

$$RSS_H=\mathbf{Y}'\mathbf{Y}-\hat{\alpha}\,'\tilde{\mathbf{X}}\,'\mathbf{Y}=\sum_{i=1}^{3}Y_i^2-\frac{1}{11}(Y_1+Y_2+3Y_3)^2$$

当 H 成立时， $F_H\sim F_{1,1}$

其中

$$F_H=\frac{RSS_H-RSS}{RSS}=\frac{6\hat{\beta}_1^2+5\hat{\beta}_2^2-\frac{1}{11}\left(Y_1+Y_2+3Y_3\right)^2}{\sum_{i=1}^{3}Y_i^2-6\hat{\beta}_1^2-5\hat{\beta}_2^2}$$

例 5.1.2　同一模型检验

假设我们对因变量 Y 和自变量 $X_1,X_2,\ldots,X_{p-1}$ 有两批观测数据．对第一批数据，有线性回归模型：

$$Y_i=\beta_0^{(1)}+\beta_1^{(1)}x_{i1}+\ldots+\beta_{p-1}^{(1)}x_{i,p-1}+\varepsilon_i\ ,\ i=1,2,\ldots,n_1$$

而对第二批数据，也有线性回归模型

$$Y_i=\beta_0^{(2)}+\beta_1^{(2)}x_{i1}+\ldots+\beta_{p-1}^{(2)}x_{i,p-1}+\varepsilon_i\ ,\ i=n_1+1,\ldots,n_1+n_2$$

其中所有误差 ε_i 都相互独立，且服从 $N(0,\sigma^2)$.

现在的问题是,考察这两批数据所反映的因变量 Y 与自变量 $X_1,\ldots,X_{p-1}$ 之间的依赖关系是不是完全一样,也就是要检验模型中的系数是否完全相等,即检验:

$$H:\beta_i^{(1)}=\beta_i^{(2)}\,,\ i=0,1,\ldots,p-1$$

这个问题具有广泛的应用背景. 例如, 这两批数据可以是同一公司在两个不同时间段上的数据, Y 是反映公司经济效应的某项指标,而自变量 $X_1,\ldots,X_{p-1}$ 是影响公司效益的内在和外在因素. 那么我们所要做的检验就是考察公司效应指标对诸因素的依赖关系在两个时间段上是否有了变化，也就是所谓经济结构的变化. 又譬如，在生物学研究中，有很多试验花费时间比较长，而为了保证结论的可靠性，又必须做一定数量的试验. 为此，很多试验要分配在几个试验室同时进行. 这时，前面讨论的数据就可以看作是来自两个不同试验室的观测数据，而我们检验的目的是考察两个试验室所得结论有没有差异. 类似的例子还可以举出很多.

为了要导出所需的检验统计量，我们首先把上面的两个模型写成矩阵形式

$$\begin{cases}\mathbf{Y}_1=\mathbf{X}_1\beta_1+\varepsilon_1\\ \varepsilon_1\sim N_{n_1}(\mathbf{0},\sigma^2\mathbf{I})\end{cases}$$

和

$$\begin{cases}\mathbf{Y}_2=\mathbf{X}_2\beta_2+\varepsilon_2\\ \varepsilon_2\sim N_{n_2}(\mathbf{0},\sigma^2\mathbf{I})\end{cases}$$

将其合并，便得到如下模型:

$$\begin{cases}\begin{pmatrix}\mathbf{Y}_1\\ \mathbf{Y}_2\end{pmatrix}=\begin{pmatrix}\mathbf{X}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{X}_2\end{pmatrix}\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}+\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\end{pmatrix}\\ \begin{pmatrix}\varepsilon_1\\ \varepsilon_2\end{pmatrix}\sim N_{n_1+n_2}\left(0,\sigma^2\mathbf{I}\right)\end{cases}\tag{5.1.11}$$

我们要检验的假设为

$$H:\ (\mathbf{I}_p\quad -\mathbf{I}_p)\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}=0\tag{5.1.12}$$

由模型 (5.1.11) 得参数 $\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$ 的最小二乘估计为

$$\begin{aligned}
\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} &= \left[\begin{pmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{pmatrix}' \begin{pmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{pmatrix} \right]^{-1} \begin{pmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{pmatrix}' \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2'\mathbf{X}_2 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}_1' & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2' \end{pmatrix} \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} \\
&= \begin{pmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-1} & \mathbf{0} \\ \mathbf{0} & (\mathbf{X}_2'\mathbf{X}_2)^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{X}_1'\mathbf{Y}_1 \\ \mathbf{X}_2'\mathbf{Y}_2 \end{pmatrix} \\
&= \begin{pmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Y}_1 \\ (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{Y}_2 \end{pmatrix}
\end{aligned}$$

于是 $\hat{\beta}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Y}_1$, $\hat{\beta}_2 = (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{Y}_2$

对应的残差平方和为

$$RSS = \mathbf{Y}_1'\mathbf{Y}_1 + \mathbf{Y}_2'\mathbf{Y}_2 - \hat{\beta}_1'\mathbf{X}_1'\mathbf{Y}_1 - \hat{\beta}_2'\mathbf{X}_2'\mathbf{Y}_2$$

因为 $\mathbf{A} = (\mathbf{I}_p \quad -\mathbf{I}_p)$, $rank(\mathbf{A}) = p$

$$\begin{aligned}
&\left(\mathbf{A}\hat{\beta}\right)' \left[\mathbf{A} \begin{pmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-1} & \mathbf{0} \\ \mathbf{0} & (\mathbf{X}_2'\mathbf{X}_2)^{-1} \end{pmatrix} \mathbf{A}' \right]^{-1} \left(\mathbf{A}\hat{\beta}\right) \\
&= \left(\hat{\beta}_1 - \hat{\beta}_2\right)' \left[(\mathbf{X}_1'\mathbf{X}_1)^{-1} + (\mathbf{X}_2'\mathbf{X}_2)^{-1} \right]^{-1} \left(\hat{\beta}_1 - \hat{\beta}_2\right)
\end{aligned}$$

所以检验假设 (5.1.12) 的统计量为

$$\begin{aligned}
F_H &= \frac{\left(\hat{\beta}_1 - \hat{\beta}_2\right)' \left[(\mathbf{X}_1'\mathbf{X}_1)^{-1} + (\mathbf{X}_2'\mathbf{X}_2)^{-1} \right]^{-1} \left(\hat{\beta}_1 - \hat{\beta}_2\right)/p}{RSS/(n_1 + n_2 - 2p)} \\
&\sim F_{p,n_1+n_2-2p}
\end{aligned}$$

据此，我们可以对假设 $H: \beta_1 = \beta_2$ 作出检验.

对给定的显著性水平 α, 若 $F > F_{p,n_1+n_2-2p}(\alpha)$, 则拒绝原假设，即认为两批数据不服从同一个线性回归模型.

四、回归方程的显著性检验

所谓回归方程的显著性检验就是检验假设：所有回归系数都等于零，即检验：

$$H: \ \beta_1 = \ldots = \beta_{p-1} = 0 \tag{5.1.13}$$

如果我们检验的结论是拒绝原假设 H, 则意味着至少有一个 $\beta_i \neq 0$, 当然也可能所有 β_i 都不等于零, 换句话说, 我们认为 Y 至少线性依赖于某一个自变量 X_i, 也可能线性依赖于所有自变量 $X_1, \ldots, X_{p-1}$. 如果检验的结论是接受原假设 H, 则意味着: 所有 $\beta_i = 0$, 即可以认为相对于误差而言, 所以自变量对因变量 Y 的影响是不重要的, 即认为回归方程没有意义. 若设 $\mathbf{A} = (0, \mathbf{I}_{p-1})$, $\mathbf{b} = \mathbf{0}$, 则 (5.1.13) 等价于

$$H: \ \mathbf{A}\beta = \mathbf{0} \quad rank(\mathbf{A}) = p - 1$$

将其代入模型 (5.1.13) 中, 得约简模型

$$\begin{cases} y_i = \beta_0 + \varepsilon_i; \\ i = 1, \cdots, n; \\ \varepsilon_i \sim N(0, \sigma^2)\text{且相互独立.} \end{cases} \tag{5.1.14}$$

由此得参数 β_0 的最小二乘估计为:

$$\hat{\beta}_{0H} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

于是对应的残差平方和为:

$$RSS_H = \mathbf{Y}'\mathbf{Y} - \hat{\beta}_{0H}\mathbf{1}'\mathbf{Y} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

这里 $\mathbf{1}$ 表示所有分量全为 1 的 n 维向量. 在回归分析中, 把这个特殊的残差平方和称为总平方和 (Total Sum of Squares, 简称为 TSS). 这是因为约简模型 (5.1.14) 中不包含任何回归自变量, 残差平方和 RSS_H 完全是 n 个观测数据 $y_1, \ldots, y_n$ 的变差平方和. 对于原模型 (5.1.3), 我们知道残差平方和:

$$RSS = \mathbf{Y}'\mathbf{Y} - \hat{\beta}'\mathbf{X}'\mathbf{Y}$$

于是

$$RSS_H - RSS = \hat{\beta}'\mathbf{X}'\mathbf{Y} - \hat{\beta}_{0H}\mathbf{1}'\mathbf{Y} \tag{5.1.15}$$

称之为回归平方和, 简记为 $SS_{\text{回}} = RSS_H - RSS$. 根据定理 5.1.4, 我们可以得到检验假设 (5.1.13) 的统计量.

$$F_{\text{回}} = \frac{SS_{\text{回}}/(p-1)}{RSS/(n-p)} \sim F_{p-1,n-p}$$

对给定的显著性水平 $\alpha > 0$, 当 $F_{\text{回}} > F_{p-1,n-p}(\alpha)$ 时, 我们拒绝原假设 H, 否则不拒绝 H.

现在我们从方差分析的角度来解释检验统计量 $F_{\text{回}}$. 根据前面的讨论知, (5.1.15) 中的 RSS_H 就是通常的总平方和 TSS, 于是有

$$TSS = RSS + SS_{\text{回}}$$

这就是说，我们把总平方和分解成两部分，一部分是回归平方和 $SS_{回} = \hat{\beta}'\mathbf{X}'\mathbf{Y} - \hat{\beta}_{0H}\mathbf{1}'\mathbf{Y}$. 它反映了回归自变量对因变量变动平方和的贡献，另一部分是残差平方和 RSS, 它是误差的影响，这里误差包括试验的随机误差和模型误差，后者是指重要回归自变量的遗漏，模型的非线性等，因此，检验统计量 $F_{回}$ 是把回归平方和与试验误差相比较，当回归平方和相对试验误差比较大时，我们就拒绝原假设. 通常我们把每个平方和除以相应的自由度，称为均方和，并列成下面的方差分析表

表 5.1.1　方差分析表

方差来源	平方和	自由度	均方和	F 比	p 值
回归	$SS_{回}$	$p-1$	$MS = \frac{SS_{回}}{(p-1)}$	$F_{回} = \frac{MS}{MR}$	$P\{F_{p-1,n-p} > F_{回}\}$
误差	RSS	$n-p$	$MR = \frac{RSS}{(n-p)}$		
总计	TSS	$n-1$			

需要强调的是，如果经过检验，不拒绝假设 $H: \beta_1 = \ldots = \beta_{p-1} = 0$, 则意味着，与模型的各种误差比较起来，诸自变量对 Y 的影响是不重要的，这里可能有两种情况，其一是，模型的各种误差太大，因而即使回归自变量对 Y 有一定影响，但相比这个较大的模型误差，也不算大，对这种情况，我们就要想办法缩小误差. 这包括从分析问题的专业背景入手，检查是否漏掉了重要自变量，或 Y 对某些回归自变量有非线性相依关系等. 其二是，回归自变量对 Y 的影响确实很小，对这种情况，我们就要放弃建立 Y 对诸自变量的线性回归.

五、回归系数的显著性检验

回归方程的显著性检验是对线性回归方程的一个整体性检验. 如果我们检验的结果是拒绝原假设, 则意味着因变量 Y 线性地依赖于自变量 $X_1, \ldots, X_{p-1}$ 这个回归自变量的整体. 但是，这并不排除 Y 并不依赖于其中某些自变量，即某些 β_i 可能等于零. 因此在回归方程显著性检验被拒绝之后，我们还要对每个自变量逐一做显著性检验，即对给定的 i,$1 \le i \le p-1$ 检验假设

$$H_i: \ \beta_i = 0 \tag{5.1.16}$$

$$\Longleftrightarrow H_i: \ \mathbf{A}\beta = 0 \text{ 其中 } \mathbf{A} = (\underbrace{0, \ldots, 0}_{i-1\text{个}}, 1, 0, \ldots, 0)$$

我们利用定理 5.1.4 可导出检验 H_i 的统计量，这里我们给出一种更直接的方法.

对于模型 (5.1.4),β 的最小二乘估计为 $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, 根据定理 5.1.3 知

$$\hat{\beta} \sim N_p\left(\beta, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}\right) \tag{5.1.17}$$

若记 $\mathbf{C}_{p\times p} = (c_{ij}) = (\mathbf{X}'\mathbf{X})^{-1}$, 则有

$$\hat{\beta}_i \sim N\left(\beta_i, \sigma^2 c_{i+1,i+1}\right)$$

于是当 H_i 成立时,

$$\frac{\hat{\beta}_i}{\sigma\sqrt{c_{i+1,i+1}}} \sim N(0,1)$$

又因为 $RSS/\sigma^2 \sim \chi^2_{n-p}$, 且与 $\hat{\beta}_i$ 相互独立，根据 t 分布的定义有

$$T_i = \frac{\hat{\beta}_i}{\sqrt{c_{i+1,i+1}}\hat{\sigma}} \sim t_{n-p} \tag{5.1.18}$$

其中 $\hat{\sigma}^2 = \frac{RSS}{n-p}$, 对给定的显著性水平 α, 当

$$|T_i| > t_{n-p}\left(\frac{\alpha}{2}\right)$$

时，我们就拒绝假设 H_i, 否则就不拒绝假设 H_i.

从 (5.1.17) 中可以看出, 回归系数 β_i 的最小二乘估计 $\hat{\beta}_i$ 的方差 $Var\left(\hat{\beta}_i\right) = \sigma^2 c_{i+1,i+1}$. 文献中常把 $Var^{1/2}\left(\hat{\beta}_i\right) = \sigma\sqrt{c_{i+1,i+1}}$ 称为 $\hat{\beta}_i$ 的标准误差. 它的一个估计为 $\hat{\sigma}\sqrt{c_{i+1,i+1}}$, 因此 (5.1.18) 所给出的 t 检验统计量就是回归系数最小二乘估计 $\hat{\beta}_i$ 与其标准误差估计的商. 在许多统计软件中，都给出如下回归系数估计表，有的称其为回归系数的方差分析表.

表 5.1.2　回归系数的方差分析表

系数	最小二乘估计	标准误差估计	统计量 t_i	p 值
β_0	$\hat{\beta}_0$	$\hat{\sigma}\sqrt{c_{11}}$	$T_0 = \frac{\hat{\beta}_0}{\hat{\sigma}\sqrt{c_{11}}}$	$P\{t_{n-p} > T_0\}$
β_1	$\hat{\beta}_1$	$\hat{\sigma}\sqrt{c_{22}}$	$T_1 = \frac{\hat{\beta}_1}{\hat{\sigma}\sqrt{c_{22}}}$	$P\{t_{n-p} > T_1\}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
β_{p-1}	$\hat{\beta}_{p-1}$	$\hat{\sigma}\sqrt{c_{pp}}$	$T_{p-1} = \frac{\hat{\beta}_{p-1}}{\hat{\sigma}\sqrt{c_{pp}}}$	$P\{t_{n-p} > T_{p-1}\}$

如果经过检验，不拒绝假设 H_i, 我们就认为回归自变量 X_i 对因变量 Y 无显著的影响，因而可以将其从回归方程中剔除. 将这个回归自变量从回归方程剔除之后，剩余变量的回归系数的估计也随之发生变化. 将 Y 对剩余的回归自变量重新作回归，然后再检验其回归系数是否为零，再剔除经检验认为对 Y 无显著影响的变量，这样的过程一直继续下去，直到对所有的自变量，经检验都认为对 Y 有显著的影响为止. 对回归系数做显著性检验的过程事实上也是回归自变量的选择过程. 关于回归自变量的选择，可参考茆诗松《线性回归分析》.

六、因变量的预测

所谓预测，就是对给定的回归自变量的值，预测对应的因变量所可能取的值. 这是回归分析最重要的应用之一，因为在线性回归模型中，回归自变量往往代表一组试验条件或生产条件或社会经济条件，由于试验或生产等方面的费用或花费时间长的原因，我们在有了经验回归模型之后，希望对一些感兴趣的试验、生产条件不真正去做试验，就对相应的因变量的取值能够作出估计和分析. 因此，预测就常常显得十分必要.

由模型 (5.1.4), 可求得参数 β 的最小二乘估计 $\hat{\beta}$, 从而建立回归方程

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_{p-1} x_{p-1} \tag{5.1.19}$$

现给定观测值 $\mathbf{X}_0 = (1, x_{01}, x_{02}, \ldots, x_{0,p-1})'$, 则根据回归方程就可求得对应的因变量的预测值为:

$$\hat{y}_0 = \mathbf{X}_0' \hat{\beta}$$

在应用上，有时区间预测更为人们所关心．所谓区间预测就是找一个区间，使得被预测量落在这个区间的概率达到预先给定的值.

因为　$y_0 = \beta_0 + \beta_1 x_{01} + \ldots + \beta_{p-1} x_{0,p-1} + \varepsilon_0$ 与$\hat{y}_0$相互独立

$$E(\hat{y}_0 - y_0) = E\hat{y}_0 - Ey = \mathbf{X}_0'\beta - \mathbf{X}_0'\beta = 0$$

$$\begin{aligned} Var(\hat{y}_0 - y_0) &= Var\hat{y}_0 + Vary_0 = Var(\mathbf{X}_0'\hat{\beta}) + Var(\varepsilon_0) \\ &= \mathbf{X}_0' Var(\hat{\beta})\mathbf{X}_0 + \sigma^2 \\ &= \sigma^2 \mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0 + \sigma^2 \end{aligned}$$

所以　$\hat{y}_0 - y_0 \sim N\left(0, \sigma^2\left(1 + \mathbf{X}_0'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}_0\right)\right)$

因为 $\hat{\sigma}^2$ 只依赖于 $y_1, \ldots, y_n$, 所以它也与 y_0 独立，而

$$\frac{(n-p)\,\hat{\sigma}^2}{\sigma^2} = \frac{RSS}{\sigma^2} \sim \chi^2_{n-p}$$

根据 t 分布的定义,

$$T = \frac{\hat{y}_0 - y_0}{\hat{\sigma}\sqrt{1 + \mathbf{X}_0'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}_0}} \sim t_{n-p}$$

故对给定的显著性水平 α, 有

$$\Pr\left\{\frac{|\hat{y}_0 - y_0|}{\hat{\sigma}\sqrt{1 + \mathbf{X}_0'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}_0}} \le t_{n-p}\left(\alpha/2\right)\right\} = 1 - \alpha$$

由此得 y_0 的置信水平为 $1-\alpha$ 的预测区间为

$$\left(\hat{y}_0 - t_{n-p}\left(\frac{\alpha}{2}\right)\hat{\sigma}\sqrt{1 + \mathbf{X}_0'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}_0}\,,\ \hat{y}_0 + t_{n-p}\left(\frac{\alpha}{2}\right)\hat{\sigma}\sqrt{1 + \mathbf{X}_0'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}_0}\right) \tag{5.1.20}$$

值得注意的是, 前面的讨论都是基于假设 $\varepsilon \sim N_n(0, \sigma^2\mathbf{I})$ 之上，但对于一个实际问题，变量 Y 与 $x_1, \ldots, x_{p-1}$ 之间到底是一个什么样的关系，我们并不十分清楚，样本数据是否存在异常值，是否存在周期性，我们往往从数据的表面并不能明显看出，用最小二乘估计模型的参数是在模型满足一些基本假设时才有效，如果模型的基本假设显著性出错，可能导致模型结论严重歪曲. 为了说明上述问题， 1973 年 Anscombe 构造了四组数据，见表 5.1.3.

表 5.1.3

第一组		第二组		第三组		第四组	
x	y	x	y	x	y	x	y
10	8.04	8	8.14	8	6.77	19	12.50
8	6.95	10	9.14	13	12.74	8	6.58
13	7.58	13	9.74	10	7.46	8	5.76
11	8.33	9	8.77	9	7.11	8	7.71
9	8.81	11	9.26	11	7.81	8	8.84
14	9.96	14	8.10	14	8.84	8	8.47
6	7.24	6	6.13	6	6.08	8	7.04
12	10.84	4	3.10	4	5.39	8	5.25
7	4.82	7	7.26	12	8.15	8	5.56
5	5.68	12	9.13	5	5.73	8	7.91
4	4.26	5	4.74	7	6.42	8	6.89

用第一组数据得到的检验回归方程为：

$$\hat{y} = 3.00009 + 0.50009x$$

相关系数 $r = 0.8164206$
决定系数 $r^2 = 0.6665$

$$F = \frac{MS}{MR} = \frac{27.51}{1.53} = 17.98$$

用第二组数据得到的检验回归方程为：

$$\hat{y} = 2.76455 + 0.53636x$$

相关系数 $r = 0.841742$
决定系数 $r^2 = 0.7085$

$$F = \frac{MS}{MR} = \frac{31.65}{1.45} = 21.8779$$

用第三组数据得到的检验回归方程为：

$$\hat{y} = 3.00246 + 0.49973x$$

相关系数 $r = 0.8162866$
决定系数 $r^2 = 0.6663$

$$F = \frac{MS}{MR} = 17.97225$$

用第四组数据得到的检验回归方程为：

$$\hat{y} = 3.00172 + 0.49991x$$

相关系数 $r = 0.816522$

决定系数 $r^2 = 0.6667081$

$$F = \frac{MS}{MR} = 18.0034$$

$$F_{1,9}(0.05) = 5.12$$

可以看出这四组数据所建的回归方程非常接近，决定系数 r^2 , F 统计量也都相近，且均通过显著性检验．这说明四组数据 y 与 x 之间都有十分显著的相关关系．然而，变量 y 与 x 之间是否就有线性相关关系呢？由上述四组数据的散点图 (见图 5.1.1) 可以看到，变量 y 与 x 之间的关系是很不相同的．

由图 5.1.1 的 (a) 可知，由直线 $\hat{y} = 3.00009 + 0.50009x$ 作为 y 与 x 间关系的拟合是合适的，回归方程刻画了变量 y 与 x 间的线性相关关系．由图 5.1.1 的 (b) 可知，变量 y 与 x 之间的相关关系应当是曲线关系，尽管回归方程也通过了显著性检验，但由直线方程 $\hat{y} = 2.76455 + 0.53636x$ 去揭示它们的相关关系很不合适．如果进一步作残差分析会发现残差点的分布不具有随机性原则．由图 5.1.1 的 (c) 可知，变量 y 与 x 之间存在着线性关系，但用直线 $\hat{y} = 3.00246 + 0.49973x$ 去拟合这种关系不太理想．因为第三组数据中第 2 对数据较其后数据大的多，进一步作残差分析，可以发现它是一个异常值．如果将它剔除，用其余 10 对数据重新计算得经验回归方程为：

$$\hat{y} = 4.00565 + 0.34539x$$

相关系数 $r = 0.9999967$
决定系数 $r^2 = 0.9999933$

$$F = 1198365$$

可以看出这个回归方程比起 $\hat{y} = 3.00246 + 0.49973x$ 显著得多. 用 $\hat{y} = 4.00565 + 0.34539x$ 去拟合 y 与 x 的关系会更精确些.

由图 5.1.1 的 (d) 可知，可以用 $\hat{y} = 3.00172 + 0.49991x$ 去拟合 y 与 x 的关系，但此直线的斜率完全取决于 (19,12.50) 这一个点，假如这个数据点是个异常值，我们无法精确地获得经验回归直线．这种情况所得到的经验回归方程是很不可信的．

这个例子告诉我们，当拒绝假设 $H_0 : \beta_1 = 0$ 时，并不能完全肯定地说 y 与 x 之间就存在线性相关关系．在实际应用中，不应局限于一种方法去分析判断．要得到确实可信的结果，我们还应对数据进行回归诊断．

七、模型的有效性——回归诊断

由前面的分析知道，当根据数据得到回归方程后，并不能说明回归方程描述了数据之间的真实关系，还需要对模型的正确性进行判断．回归诊断主要讨论两个问题： (1) 分析模型的假设 $\varepsilon \sim N_n(0, \sigma^2\mathbf{I})$ 是否正确．因为这些假设都是关于误差项的，所以很自然我们要从分析它们的“估计量”——残差的角度来解决，正是这个原因，这部分内容在文献中也称为残差分析. (2) 探测强影响数据．对参数估计或预测有异常大的影响的数据称为强影响数据或强影响点．在回归分析中，因变量 Y 的取值 y_i 具有随机性，而自变量 $X_1, \ldots, X_{p-1}$ 的取值 $\mathbf{x}_i' = (x_{i1}, \ldots, x_{i,p-1})$, $i = 1, \ldots, n$ 也只是许多可能取

到的值中的 n 组. 我们希望每组数据 $(\mathbf{x}_i', y_i)$ 对未知参数的估计有一定的影响, 但这种影响不能过大. 这样, 我们得到的经验回归方程就具有一定的稳定性. 不然的话, 如果个别一两组数据对估计有异常大的影响, 当我们剔除这些数据之后, 就能得到与原来差异很大的经验回归方程, 这样我们就有理由怀疑所建立的经验回归方程是否真正描述了因变量与诸自变量之间的客观存在的相依关系. 正是这个原因, 我们在做回归分析时, 有必要考察每组数据对参数估计的影响大小. 这部分内容在回归诊断中, 统称为影响分析. 影响分析只是研究探查强影响数据的统计方法, 至于对已确认的强影响数据如何处理, 这需要具体问题具体分析. 往往先要仔细核查数据获得的全过程, 如果强影响数据是由于试验条件失控或记录失误或其它一些过失所致, 那么这些数据就应该剔除. 不然的话, 应该考虑收集更多的数据 (从几何上讲, 这些数据应该跟强影响数据比较接近) 或采用一些稳健估计方法以缩小强影响数据对估计的影响, 从而获得较稳定的经验回归方程.

回归诊断是一个很复杂的问题, 实施起来很有点像医生给病人诊病, 有时一个症状往往是多种不同疾病的征兆, 必须从多方面做检查分析, 才能断言毛病出在什么地方. 在这方面, 理论虽然起一定指导作用, 但 "临床" 经验是十分重要的, 这一点从后面的讨论中可以看到. 我们先讨论第一个问题——残差分析.

考虑线性回归模型 (5.1.4), 如果用 $\mathbf{x}_1', \ldots, \mathbf{x}_n'$ 表示 X 的 n 个行向量, 则

$$\hat{\varepsilon}_i = y_i - \mathbf{x}_i'\hat{\beta}\,,\ i = 1, \ldots, n$$

为第 i 次试验或观测的残差. 我们把残差 $\hat{\varepsilon}_i$ 看作误差 ε_i 的一次观测值, 如果模型 (5.1.4) 正确的话, 它应该具有 ε_i 的一些特征, 因此我们应该首先研究残差的性质.

记 $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$, 称 $\hat{\mathbf{Y}}$ 为拟合值向量, 称其第 i 个分量 $\hat{y}_i = \mathbf{x}_i'\hat{\beta}$ 为第 i 个拟合值, 则

$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \stackrel{\triangle}{=} (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

残差向量 $\hat{\varepsilon}$ 可以表示为:

$$\hat{\varepsilon} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y} = \mathbf{H}\varepsilon$$

从这个表达式我们很容易证明残差向量的下列重要性质.

定理 5.1.3 (a)$E(\hat{\varepsilon}) = 0\,, Cov(\hat{\varepsilon}) = \sigma^2\mathbf{H}$

(b)$\hat{\varepsilon} \sim N_n\left(0, \sigma^2\mathbf{H}\right)$

我们看到 $Var(\hat{\varepsilon}_i) = \sigma^2 h_{ii}$, 这里 h_{ii} 为 $\mathbf{H}$ 的第 i 个对角元. 可见一般情况下残差 $\hat{\varepsilon}_i$ 的方差不相等, 这有碍于 $\hat{\varepsilon}_i$ 的实际应用, 将其标准化为 $\hat{\varepsilon}_i/(\sigma\sqrt{h_{ii}})$, 再用 $\hat{\sigma}$ 代替 σ 得到所谓学生化残差

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{h_{ii}}}\ ,\ \ i = 1, \ldots, n$$

这里 $\hat{\sigma}^2 = RSS/(n-p)$. 即使在 $\varepsilon \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$ 的条件下, r_i 的分布仍然比较复杂, 但可以近似地认为 r_i 相互独立服从 $N(0,1)$(详细讨论见陈希孺等 (1987)). 于是我们可以断言当 $\varepsilon \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$ 时, 学生化残差 $r_1, \ldots, r_n$ 近似

地看作来自总体 $N(0,1)$ 的一组随机样本. 根据正态分布的如下性质：若随机变量 $U \sim N(\mu, \sigma^2)$, 则

$$P\left(\mu - 2\sigma < U < \mu + 2\sigma\right) = 95.4\%$$

对于现在的情形， $\mu = 0\,, \sigma = 1$, 于是，大约应有 95.4拟合值向量 $\hat{y}$ 与残差 $\hat{\varepsilon}$ 相互独立，因而与学生化残差 $r_1, \ldots, r_n$ 也独立. 所以，如果我们以拟合值 $\hat{y}_i$ 为横轴， r_i 为纵轴，那么平面上的点 $(\hat{y}_i\,, r_i)\ , i = 1, \ldots, n$ 大致应落在宽度为 4 的水平带 $|r_i| \leq 2$ 区域内，且不呈现任何趋势，如图 5.1.2(a). 这样的以残差为纵轴而以拟合值 $\hat{y}_i$ 或其它量为横轴的图称为残差图，这是回归诊断的一个重要工具. 如果残差图具有图 5.1.2(a) 的性状，则我们可以认为，现在我们手头的数据与假设 $\varepsilon \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$ 没有明显不一致的征兆. 我们就可以认为，假设 $\varepsilon \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$ 基本上是合理的. 而图 5.1.2(b)-(d) 显示了误差等方差，即 $Var(\varepsilon_i) = \sigma^2\,, i = 1, \ldots, n$ 不满足. 其中图 (b) 表示了误差方差随 $\hat{y}_i$ 的增大而有增加的趋势. 而图 (c) 所表示的情形正好相反，即误差方差随 $\hat{y}_i$ 的增大而减少，但是图 (d) 表示对较大或较小的 $\hat{y}_i$, 误差方差偏小，而对中等大小的 $\hat{y}_i$, 误差方差偏大. 图 (e) 和 (f) 表明回归函数可能是非线性的，或误差 ε_i 之间有一定相关性或漏掉了一个或多个重要的回归自变量. 对于一批实际数据，这样的残差图究竟反映了哪一种情况，还需要做进一步的诊断. 这种由一种 “症状” 可能产生多种 “疾病” 的情况正是回归诊断的困难所在. 在具体处理时，和医生治病一样，临床经验是很重要的.

上面我们讲的是以拟合值 $\hat{y}_i$ 为横坐标的残差图. 为了从不同的角度分析残差，我们可以做其它一些残差图. 例如，如果因变量是按时间顺序观测的，那么 $y_1, \ldots, y_n$ 表示了分别在时刻 $t = t_1, \ldots, t_n$ 的因变量观测值，则我们可以取时间 t 或观测序号为 X 轴，构造 $(t_i\,, r_i)$ 或 $(i\,, r_i)$ 的残差图. 又譬如，我们也可将某个自变量 X_j 取做 X 轴等. 不同的残差图可能从不同角度提供一些有用信息.

从残差图诊断出可能的 “疾病”，也就是某些假设条件不成立，我们就需要对问题 “对症下药”，如果有症状使我们怀疑因变量 Y 对自变量的依赖不仅仅是线性关系，那么我们就可以考虑在回归自变量中增加某些自变量的二次项，如 X_i^2 或 X_2^2 或交叉项 X_1X_2 等. 至于增添哪些变量的二次项和哪些变量的交叉项，这就要通过对实际问题的分析或实际计算，看其实际效果. 若增加二次项 $X_1^2\,, X_2^2$ 和交叉项 X_1X_2, 可以通过引进新变量 $Z_1 = X_1^2\,, Z_2 = X_2^2\,, Z_3 = X_1X_2$, 把问题化成线性回归形式. 如果残差图显示了误差方差不相等，我们可以有两种 “治疗方案”. 其一是对因变量作变换，使变换过的新变量具有近似相等的方差. 重要的问题是如何选择所要做的变换. 虽然在理论上有一些原则可遵循 (参阅陈希孺等 (1987)p.122), 但在实际应用中还是要靠对具体情况的分析，提出一些可选择的变换，然后通过实际计算比较它们的客观效果. 另一种方法是应用加权最小二乘估计. 另外，还有一种因变量的变换，它是从综合角度考虑 (即要求对因变量变换过之后，新的因变量关于诸自变量具有线性相依关系，且误差服从正态、等方差、相互独立等) 提出的一种 “治疗方案”，在实际应用上效果比较好，这就是前面介绍过的著名的 Box-Cox 变换.

为了既简单又能说明问题，下面我们看一个一元线性回归的例子.

例 5.1.3　一公司为了研究产品的营销策略，对产品的销售情况进行了调查. 设 Y 表示某地区该产品的家庭人均购买量 (单位：元),X 表示家庭人均收入 (单位：元). 表 5.1.4 记录了 53 个家庭的数据.

表 5.1.4　家庭人均收入数据

i	X(元)	Y(元)	$\hat{y}_i$	$\hat{\varepsilon}_i$	$Z=\sqrt{Y}$	$\hat{z}_i'$	$\tilde{\varepsilon}_i$
1	679	0.790	1.669	-0.879	0.889	1.229	-0.340
2	292	0.440	0.244	0.196	0.663	0.860	-0.197
3	1012	0.560	2.896	-2.336	0.748	1.547	-0.798
4	493	0.790	0.984	-0.194	0.889	1.052	-0.163
5	582	2.700	1.312	1.388	1.643	1.137	0.506
6	1156	3.640	3.426	0.214	1.908	1.684	0.224
7	997	4.730	2.840	1.890	2.175	1.532	0.643
8	2189	9.500	7.230	2.270	3.082	2.668	0.414
9	1097	5.340	3.209	2.131	2.311	1.628	0.683
10	2078	6.850	6.822	0.028	2.617	2.562	0.055
11	1818	5.840	5.864	-0.024	2.417	2.315	0.102
12	1700	5.210	5.430	-0.220	2.283	2.202	0.080
13	747	3.250	1.920	1.330	1.803	1.294	0.509
14	2030	4.430	6.645	-2.215	2.105	2.517	-0.412
15	1643	3.160	5.220	-2.060	1.778	2.148	-0.370
16	414	0.500	0.693	-0.193	0.707	0.977	-0.270
17	354	0.170	0.472	-0.302	0.412	0.920	-0.507
18	1276	1.880	3.868	-1.988	1.371	1.798	-0.427
19	745	0.770	1.912	-1.412	0.877	1.292	-0.415
20	435	1.390	0.771	0.691	1.179	0.997	0.182
21	540	0.560	1.157	-0.597	0.748	1.097	-0.348
22	874	1.560	2.388	-0.828	1.249	1.415	-0.166
23	1543	5.280	4.851	0.429	2.298	2.052	0.245
24	1029	0.640	2.958	-2.318	0.800	1.563	-0.763
25	710	4.000	1.784	2.216	2.000	1.259	0.741
26	1434	0.310	4.450	-4.140	0.557	1.949	-1.392
27	837	4.200	2.251	1.949	2.049	1.380	0.670
28	1748	4.880	5.606	-0.726	2.209	2.248	-0.039
29	1381	3.480	4.255	-0.775	1.865	1.898	-0.033
30	1428	7.580	4.428	3.152	2.753	1.943	0.810
31	1255	2.630	3.791	-1.161	1.622	1.778	-0.156
32	1777	4.990	5.713	-0.723	2.234	2.275	-0.042
33	370	0.590	0.531	0.059	0.768	0.935	-0.167
34	2316	8.190	7.698	0.492	2.862	2.789	0.073
35	1130	4.790	3.330	1.460	2.189	1.659	0.530
36	463	0.510	0.874	-0.364	0.714	1.023	-0.309
37	770	1.740	2.004	-0.264	1.319	1.316	0.003
38	724	4.100	1.835	2.265	2.025	1.272	0.753

i	X(元)	Y(元)	$\hat{y}_i$	$\hat{\varepsilon}_i$	$Z=\sqrt{Y}$	$\hat{z}_i'$	$\tilde{\varepsilon}_i$
39	808	3.940	2.144	1.796	1.985	1.352	0.633
40	790	0.960	2.078	-1.118	0.980	1.335	-0.355
41	783	3.290	2.052	1.238	1.814	1.328	0.486
42	406	0.440	0.664	-0.224	0.663	0.969	-0.306
43	1242	3.240	3.743	-0.503	1.800	1.766	0.034
44	658	2.140	1.592	0.548	1.463	1.209	0.254
45	1746	5.710	5.599	0.111	2.390	2.246	0.144
46	468	0.640	0.892	-0.252	0.800	1.028	-0.228
47	1114	1.900	3.271	-1.371	1.378	1.644	-0.262
48	413	0.510	0.690	-0.180	0.714	0.976	-0.262
49	1787	8.330	5.750	2.580	2.886	2.285	0.601
50	3560	14.940	12.280	2.660	3.865	3.974	-0.109
51	1495	5.110	4.675	0.435	2.261	2.007	0.254
52	2221	3.850	7.348	-3.498	1.962	2.699	-0.736
53	1526	3.930	4.789	-0.859	1.982	2.036	-0.054

应用最小二乘法，求得 Y 对 X 的一元经验回归方程为

$$\hat{Y}=-0.8313+0.003683X$$

相应的残差 $\hat{\varepsilon}_i$ 和拟合值 $\hat{y}_i$ 也列在表 5.1.4 中. 图 5.1.3 是以 $\hat{y}_i$ 为横轴，残差 $\hat{\varepsilon}_i$ 为纵轴的残差图. 直观上容易看出，残差图从左向右逐渐散开呈漏斗状，这是误差方差不相等的一个征兆. 考虑对因变量 Y 作变换，先试变换 $Z=Y^{\frac{1}{2}}$，得到经验回归方程

$$\hat{Z}=0.5822+0.000953X$$

计算新的残差 $\bar{\varepsilon}_i$，残差图画在图 5.1.4 中，已无任何明显趋势.

这表明我们所用的变换是合适的，最后得到的经验回归方程为

$$\begin{aligned}\hat{Y} &= \hat{Z}^2=(0.5822+0.000953X)^2\\ &= 0.3390+0.0011X+0.00000091X^2\end{aligned}$$

现在我们讨论回归诊断的第二个问题：影响分析，即探查对估计或预测有异常大影响的数据. 为此，我们先引进一些记号，用 $\mathbf{Y}_{(i)}$，$\mathbf{X}_{(i)}$ 和 $\varepsilon_{(i)}$ 分别表示从 $\mathbf{Y}$、$\mathbf{X}$ 和 ε 剔除第 i 行所得到的向量或矩阵. 从线性回归模型 (5.1.4) 剔除第 i 组数据后，剩余的 $n-1$ 组数据的线性回归模型为：

$$\begin{cases}\mathbf{Y}_{(i)}=\mathbf{X}_{(i)}\beta+\varepsilon_{(i)}\\ \varepsilon_{(i)}\sim N_{n-1}(\mathbf{0},\sigma^2\mathbf{I}_{n-1})\end{cases}\tag{5.1.21}$$

将从这个模型求到的 β 的最小二乘估计记为 $\hat{\beta}_{(i)}$，则

$$\hat{\beta}_{(i)}=\left(\mathbf{X}_{(i)}{}'\mathbf{X}_{(i)}\right)^{-1}\mathbf{X}_{(i)}{}'\mathbf{Y}_{(i)}\tag{5.1.22}$$

很显然，向量 $\hat{\beta}-\hat{\beta}_{(i)}$ 反映了第 i 组数据对回归系数估计的影响大小，但它是一个向量，应用上不便于使用，于是需要考虑它的某种数量化函数． Cook 统计量就是其中应用最广泛的一种． Cook 统计量的定义为：

$$D_i=\frac{\left(\hat{\beta}-\hat{\beta}_{(i)}\right)'\mathbf{X}'\mathbf{X}\left(\hat{\beta}-\hat{\beta}_{(i)}\right)}{p\,\hat{\sigma}^2},\ i=1,\ldots,n \tag{5.1.23}$$

这里 $\hat{\sigma}^2=\parallel \mathbf{y}-\mathbf{X}\hat{\beta}\parallel^2/(n-p)$. 于是，对每一组观测数据，我们可以有一个数量 D_i 来刻画它对回归系数估计影响的大小．但要从 (5.1.23) 计算 $D_i,\ i=1,\ldots,n$ 很不方便，它需要计算 $\hat{\beta},\hat{\beta}_{(1)},\ldots,\hat{\beta}_{(n)}$, 因而需要计算 $n+1$ 个回归，计算量太大．下面的定理提供了计算 D_i 的简便公式，它只需要计算完全数据的线性回归模型 (5.1.4).

定理 5.1.5

$$D_i=\frac{1}{p}\left(\frac{1-h_{ii}}{h_{ii}}\right)r_i^2,\ i=1,\ldots,n \tag{5.1.24}$$

这里 h_{ii} 是矩阵 $\mathbf{H}=\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ 的第 i 个对角元， r_i 是学生化残差.

这个定理表明，在计算 Cook 统计量时，我们只需要从完全数据的线性回归模型算出学生化残差 r_i, 矩阵 $\mathbf{H}$ 的对角元 h_{ii} 就可以了，并不必对任何一个不完全数据的线性回归模型 (5.1.21) 进行计算.

在 (5.1.24) 中，若不考虑与 i 无关的因子 $\frac{1}{p}$, 则 Cook 统计量 D_i 被分解成两部分，其中一部分为

$$P_i=\frac{1-h_{ii}}{h_{ii}}$$

?(它是 h_{ii} 的单调增函数，因为 h_{ii} 度量了第 i 组数据 x_i 到试验中心 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ 的距离 (参见陈希孺等 (1987)). 因此，本质上 P_i 刻画了第 i 组数据距离其它数据的远近). 而另一部分为 r_i^2. 直观上，如果一组数据距离试验中心很远，并且对应的学生化残差又很大，那么它必定是强影响数据．但是，要给 Cook 统计量一个用以判定强影响数据的临界值是很困难的，在应用上要视具体问题的实际情况而定．关于这一点，还有另外一种做法．由定理 5.1.3 可知

$$\frac{\left(\hat{\beta}-\beta\right)'\mathbf{X}'\mathbf{X}\left(\hat{\beta}-\beta\right)}{p\,\hat{\sigma}^2}\sim F_{p,n-p}$$

对给定的 $\alpha(0<\alpha<1)$, 随机事件

$$\frac{\left(\hat{\beta}-\beta\right)'\mathbf{X}'\mathbf{X}\left(\hat{\beta}-\beta\right)}{p\,\hat{\sigma}^2}\le F_{p,n-p}(\alpha) \tag{5.1.25}$$

发生的概率为 $1-\alpha$. 这里 $F_{p,n-p}(\alpha)$ 表示自由度为 p,$n-p$ 的 F 分布的上侧 α 分位点．把 β 看作变量，(5.1.25) 表示 p 维空间中以 $\hat{\beta}$ 为中心的椭球，称为置信椭球，其置信系数为 $1-\alpha$. 在 (5.1.25) 左端用 $\hat{\beta}_{(i)}$ 代替 $\hat{\beta}$, 就得到 Cook 统计量 D_i. 因此，若 $D_i=F_{p,n-p}(\alpha)$, 就表明将第 i 组数据剔除后， $\hat{\beta}_{(i)}$ 从 $\hat{\beta}$ 处移到了 β 的置信系数为 $1-\alpha$ 的置信椭球上．这样我们可以借助于置信系数的大小来评价 D_i 的大小．例如，若 $D_i=F_{p,n-p}(0.20)$, 则表示第 i 组数

据剔除后， β 的最小二乘估计 $\hat{\beta}_{(i)}$ 落在了 β 的置信系数为 $1-0.80=20\%$ 的置信椭球上. D_i 对应的置信系数愈大，表明第 i 组数据的影响愈大. 但在大多数情况下，对给定的 D_i, 我们难于从 F 分布表求到对应的置信系数精确值.

§5.2 多元多重线性回归

在本节中，我们考虑 m 个响应变量 $Y_1, Y_2, \ldots, Y_m$ 与 $p-1$ 个预测变量 $X_1, X_2 \ldots, X_{p-1}$ 之间的关系. 假设每个响应都有自己的回归模型.

$$\begin{cases} Y_1 = \beta_{10} + \beta_{11}X_1 + \ldots + \beta_{1,p-1}X_{p-1} + \varepsilon_1 \\ Y_2 = \beta_{20} + \beta_{21}X_1 + \ldots + \beta_{2,p-1}X_{p-1} + \varepsilon_2 \\ \vdots \\ Y_m = \beta_{m0} + \beta_{m1}X_1 + \ldots + \beta_{m,p-1}X_{p-1} + \varepsilon_m \end{cases} \tag{5.2.1}$$

我们称 (5.2.1) 为多元多重线性回归模型. 误差项 $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_m)'$ 满足 $E(\varepsilon) = \mathbf{0}, Var(\varepsilon) = \mathbf{\Sigma}$. 由 $Var(\varepsilon) = \mathbf{\Sigma}$ 知, 与不同响应相联系的误差项彼此之间可能相关. 为建立与一元多重线性回归模型相一致的符号体系, 以 $(1, X_{j1}, \ldots, X_{j,p-1})'$ 代表第 j 次试验中的预测变量值，以 $Y_j = (Y_{j1}, Y_{j1}, \ldots, Y_{jm})'$ 代表第 j 个响应值， $\varepsilon_j = (\varepsilon_{j1}, \varepsilon_{j2}, \ldots, \varepsilon_{jm})'$ 代表第 j 个误差，用矩阵表示时，若记

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1,p-1} \\ 1 & X_{21} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{n,p-1} \end{pmatrix}_{n\times p}$$

$$\mathbf{Y} = \begin{pmatrix} Y_{11} & Y_{12} & \cdots & Y_{1m} \\ Y_{21} & Y_{22} & \cdots & Y_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \cdots & Y_{nm} \end{pmatrix} = \begin{pmatrix} Y_{(1)} & Y_{(2)} & \cdots & Y_{(m)} \end{pmatrix}$$

$$\beta = \begin{pmatrix} \beta_{10} & \beta_{20} & \cdots & \beta_{m0} \\ \beta_{11} & \beta_{21} & \cdots & \beta_{m1} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{1,p-1} & \beta_{2,p-1} & \cdots & \beta_{m,p-1} \end{pmatrix}_{p\times m} = \begin{pmatrix} \beta_{(1)} & \beta_{(2)} & \cdots & \beta_{(m)} \end{pmatrix}$$

$$\varepsilon = \begin{pmatrix} \varepsilon_{11} & \varepsilon_{12} & \cdots & \varepsilon_{1m} \\ \varepsilon_{21} & \varepsilon_{22} & \cdots & \varepsilon_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \cdots & \varepsilon_{nm} \end{pmatrix} = \begin{pmatrix} \varepsilon_1' \\ \varepsilon_2' \\ \vdots \\ \varepsilon_n' \end{pmatrix} = \begin{pmatrix} \varepsilon_{(1)} & \varepsilon_{(2)} & \cdots & \varepsilon_{(m)} \end{pmatrix}$$

则多元多重线性回归模型可写成以下形式

$$\underset{(n\times m)}{\mathbf{Y}} = \underset{(n\times p)}{\mathbf{X}}\ \underset{(p\times m)}{\beta} + \underset{(n\times m)}{\varepsilon} \tag{5.2.2}$$

其中 $E\left(\varepsilon_{(i)}\right)=0\,, Cov\left(\varepsilon_{(i)}\,,\varepsilon_{(j)}\right)=\sigma_{ij}\mathbf{I}\,, i,j=1,2,\ldots,m.$ 第 j 次试验的 m 个观测值有协方差阵 $\mathbf{\Sigma}=(\sigma_{ij})$, 但来自不同试验的观测值不相关, 即 $Cov\left(\varepsilon^{i}\,,\varepsilon^{j}\right)=0, i\neq j.\beta$ 和 σ_{ij} 是未知参数.

一、未知参数的估计

从 (5.2.2) 知, 第 i 个响应 $\mathbf{Y}_{(i)}\,, i=1,\ldots,m$ 服从线性回归模型

$$\begin{cases} \mathbf{Y}_{(i)}=\mathbf{X}\beta_{(i)}+\varepsilon_{(i)} \\ \\ E\varepsilon_{(i)}=\mathbf{0}\,, Cov\varepsilon_{(i)}=\sigma_{ii}\mathbf{I} \end{cases} \tag{5.2.3}$$

由 §6.1 的讨论可得 $\beta_{(i)}$ 的最小二乘估计为

$$\hat{\beta}_{(i)}=\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}_{(i)}\ ,\ i=1,2,\ldots,m$$

从而参数 β 的最小二乘估计为

$$\begin{aligned} \hat{\beta} &= \left(\hat{\beta}_{(1)}\ \hat{\beta}_{(2)}\ \ldots\ \hat{\beta}_{(m)}\right)=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left(\mathbf{Y}_{(1)}\,\mathbf{Y}_{(2)}\ \ldots\ \mathbf{Y}_{(m)}\right) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \end{aligned}$$

由 $\hat{\beta}$ 可以构造以下两个矩阵

预测值矩阵 $\quad\hat{\mathbf{Y}}=\mathbf{X}\hat{\beta}=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$

残差矩阵 $\qquad\hat{\varepsilon}=\mathbf{Y}-\hat{\mathbf{Y}}=\left[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\mathbf{Y}$

根据一元多重线性回归模型的性质, 很容易推得如下性质:

定理 5.2.1 对模型 (5.2.3), 有

(1) $E\hat{\beta}=\beta$, 即$\hat{\beta}$是参数β的无偏估计

(2) $Cov(\hat{\beta}_{(i)},\hat{\beta}_{(j)})=\sigma_{ij}(\mathbf{X}'\mathbf{X})^{-1}\,, i,j=1,2,\ldots,m$

(3) $Cov(\hat{\beta}\,,\hat{\varepsilon})=0$

(4) $E\left(\frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-p}\right)=\mathbf{\Sigma}$, 即$\frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-p}$是$\mathbf{\Sigma}$的无偏估计

证明: (1) 因为 $E(\hat{\beta}_{(i)})=\beta_{(i)}\,, i=1,2,\ldots,m$

所以 $\quad\begin{aligned} E\hat{\beta} &= \left(E\hat{\beta}_{(1)}\ E\hat{\beta}_{(2)}\ \ldots\ E\hat{\beta}_{(m)}\right) \\ &= (\beta_1\ \beta_2\ \ldots\ \beta_m) \\ &= \beta \end{aligned}$

故 $\hat{\beta}$ 是参数 β 的无偏估计.

(2) 因为

$$\begin{aligned}\hat{\beta}_{(i)}-\beta_{(i)} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_{(i)}-\beta_{(i)}\\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left(\hat{\beta}_{(i)}+\varepsilon_{(i)}\right)-\beta_{(i)}\\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon_{(i)}\end{aligned}$$

所以

$$Cov(\hat{\beta}_{(i)},\hat{\beta}_{(j)})=E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon_{(i)}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon_{(j)}\right)'\right]$$

$$=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E\left(\varepsilon_{(i)}\varepsilon'_{(j)}\right)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

$$=\sigma_{ij}(\mathbf{X}'\mathbf{X})^{-1}\ ,\ i,j=1,2,\ldots,m$$

(3) 因为

$\hat{\varepsilon}_{(i)}=\mathbf{Y}_{(i)}-\hat{\mathbf{Y}}_{(i)}=\left[I-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\varepsilon_{(i)}$

所以，对任意的 i,j

$$Cov(\hat{\beta}_{(i)}\,,\hat{\varepsilon}_{(j)})=E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon_{(i)}\cdot\left(\left(\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\varepsilon_{(j)}\right)'\right]$$

$$=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E\left(\varepsilon_{(i)}\cdot\varepsilon'_{(j)}\right)\left(\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)'$$

$$=\sigma_{ij}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left(\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)$$

$$=0$$

故 $\hat{\beta}$ 中每个元素与 $\hat{\varepsilon}$ 中的每个元素不相关，因此 $\hat{\beta}$ 与 $\hat{\varepsilon}$ 不相关.

(4) 因为对任意的 i,j 有

$$\begin{aligned}E\left(\hat{\varepsilon}'_{(i)}\hat{\varepsilon}_{(j)}\right) &= E\left(\varepsilon'_{(i)}\left(\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\varepsilon_{(j)}\right)\\ &= tr\left[\left(\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)E\varepsilon'_{(i)}\varepsilon_{(j)}\right]\\ &= tr\left[\left(\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\sigma_{ij}\mathbf{I}\right]\\ &= \sigma_{ij}tr\left(\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\\ &= \sigma_{ij}(n-p)\end{aligned}$$

所以 $E\left(\frac{\hat{\varepsilon}_{(i)}{}'\hat{\varepsilon}_{(j)}}{n-p}\right)=\sigma_{ij}$, 故 $\frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-p}$ 是协方差矩阵 $\mathbf{\Sigma}$ 的无偏估计. 由前面的讨论知，多元多重回归模型并没有提出新的计算问题. 最小二乘估计 $\hat{\beta}_{(i)}=$

$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}_{(i)}$ 是对每个响应变量逐个计算的，但要注意的是：模型要求对所有响应变量使用同一组预测变量. 一旦用数据拟合了一个多元多重回归模型，就应该用 §5.1 中所介绍的诊断方法对这个模型做检查，根据残差向量检验模型的正态性是否满足或离群值是否存在.

二、回归参数的检验

要对回归参数进行检验，还需要假设误差向量 $\varepsilon \sim N_m(\mathbf{0}, \mathbf{\Sigma})$. 此时，$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ 为 β 的极大似然估计量，且对于任意的 $i = 1, 2, \ldots, m$ 有 $\hat{\beta}_{(i)} \sim N_p(\beta_{(i)}\,, \sigma_{ii}(\mathbf{X}'\mathbf{X})^{-1})$ 且 $\mathbf{\Sigma}$ 的极大似然估计量为

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\hat{\varepsilon}\,'\hat{\varepsilon} = \frac{1}{n}\left(\mathbf{Y} - \mathbf{X}\hat{\beta}\right)'\left(\mathbf{Y} - \mathbf{X}\hat{\beta}\right)$$

它具有性质： (1) $n\hat{\mathbf{\Sigma}} \sim W_{n-p}(\mathbf{\Sigma})$

(2)$\hat{\mathbf{\Sigma}}$与$\hat{\beta}$相互独立

如果响应变量不依赖于预测变量 $X_q, \ldots, X_{p-1}$, 则在 m 个响应情况下，该假设变为：

$$H:\ \beta^{(2)} = 0 \quad \text{其中矩阵} \quad \beta = \begin{pmatrix} \beta^{(1)} \\ \beta^{(2)} \end{pmatrix} \begin{matrix} q \\ p-q \end{matrix}$$

令$\mathbf{X} = \underset{q}{(\mathbf{X}^{(1)}} \quad \underset{p-q}{\mathbf{X}^{(2)}}\ )$ ，我们可以将模型 (5.2.2) 写成

$$\mathbf{Y} = \mathbf{X}^{(1)}\beta^{(1)} + \mathbf{X}^{(2)}\beta^{(2)} + \varepsilon \tag{5.2.4}$$

在假设 $H:\ \beta^{(2)} = 0$ 下，因为

$$\mathbf{Y} = \mathbf{X}^{(1)}\beta^{(1)} + \varepsilon$$

所以假设 H 的似然比

$$\Lambda = \frac{\max_{\beta^{(1)},\mathbf{\Sigma}} L\left(\beta^{(1)}, \mathbf{\Sigma}\right)}{\max_{\beta,\mathbf{\Sigma}} L\left(\beta, \mathbf{\Sigma}\right)} = \frac{L(\hat{\beta}^{(1)}, \hat{\mathbf{\Sigma}}_1)}{L(\hat{\beta}, \hat{\mathbf{\Sigma}})} = \left(\frac{|\hat{\mathbf{\Sigma}}|}{|\hat{\mathbf{\Sigma}}_1|}\right)^{\frac{n}{2}}$$

其中 $\hat{\beta}^{(1)} = \left(\mathbf{X}^{(1)\prime}\mathbf{X}^{(1)}\right)^{-1}\mathbf{X}^{(1)\prime}\mathbf{Y}$, $\hat{\mathbf{\Sigma}}_1 = \frac{1}{n}\left(\mathbf{Y} - \mathbf{X}^{(1)}\hat{\beta}^{(1)}\right)'\left(\mathbf{Y} - \mathbf{X}^{(1)}\hat{\beta}^{(1)}\right)$

因此可等价地采用威尔克斯 Λ 统计量

$$\Lambda^{\frac{2}{n}} = \frac{|\hat{\mathbf{\Sigma}}|}{|\hat{\mathbf{\Sigma}}_1|}$$

定理 5.2.2 对于模型 (5.2.2), 若 $rank(\mathbf{X}) = p$, 且 $p + m < n$, 误差 ε 服从正态分布，在假设 $H:\ \beta^{(2)} = 0$ 下， $n\hat{\mathbf{\Sigma}}$ 服从 $W_{n-p}(\mathbf{\Sigma})$ 分布， $n\left(\hat{\mathbf{\Sigma}}_1 - \hat{\mathbf{\Sigma}}\right)$ 服从 $W_{p-q}(\mathbf{\Sigma})$ 分布，且二者相互独立， H 的似然比检验等价于当

$$-2\ln\Lambda = -n\ln\left(\frac{|\hat{\mathbf{\Sigma}}|}{|\hat{\mathbf{\Sigma}}_1|}\right) = -n\ln\left(\frac{|n\hat{\mathbf{\Sigma}}|}{|n\hat{\mathbf{\Sigma}} + n\left(\hat{\mathbf{\Sigma}}_1 - \hat{\mathbf{\Sigma}}\right)|}\right)$$

的值较大时拒绝 H. 当 n 充分大时 $(n \gg pm)$, 修正的统计量

$$-\left[n-p-\frac{1}{2}(m-p+q)\right]\ln\left(\frac{|\hat{\boldsymbol{\Sigma}}|}{|\hat{\boldsymbol{\Sigma}}_1|}\right)\sim\chi^2_{(p-q)m}$$

对给定的显著性水平 α, 当左边的式子大于 $\chi^2_{(p-q)m}(\alpha)$ 时，拒绝假设 H.

三、多元多重回归预测

如果具有正态误差 ε 的模型 $\mathbf{Y}=\mathbf{X}\beta+\varepsilon$ 经检验是正确的，则可利用该模型进行预测. 对于给定的预测变量值 $\mathbf{X}_0=(1,x_{01},\cdots,x_{0,p-1})'$, 因为

$$\hat{\beta}'\mathbf{X}_0\sim N_m\left(\beta'\mathbf{X}_0\,,\,\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0\Sigma\right)$$

$$n\hat{\boldsymbol{\Sigma}}\sim W_{n-p}(\boldsymbol{\Sigma})\ \text{且与}\hat{\beta}\text{相互独立}$$

根据 T^2 统计量的定义

$$T^2=\left[\frac{\hat{\beta}'\mathbf{X}_0-\beta'\mathbf{X}_0}{\sqrt{\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0}}\right]'\left(\frac{n\hat{\boldsymbol{\Sigma}}}{n-p}\right)^{-1}\left[\frac{\hat{\beta}'\mathbf{X}_0-\beta'\mathbf{X}_0}{\sqrt{\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0}}\right]$$

与 $\dfrac{(n-p)m}{n-p-m+1}F_{m,n-p-m+1}$ 同分布.

故回归函数在 $\mathbf{X}_0$ 的平均值 $E\mathbf{X}_0=\beta'\mathbf{X}_0$ 的置信水平为 $1-\alpha$ 的置信椭球为

$$\left(\hat{\beta}'\mathbf{X}_0-\beta'\mathbf{X}_0\right)'\hat{\boldsymbol{\Sigma}}^{-1}\left(\hat{\beta}'\mathbf{X}_0-\beta'\mathbf{X}_0\right)\le\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0\cdot\frac{nm}{n-p-m+1}F_{m,n-p-m+1}(\alpha)$$

其中 $F_{m,n-p-m+1}(\alpha)$ 为自由度为 m 和 $n-p-m+1$ 的 F 分布的上侧 α 分位数.

$E(\mathbf{Y}_i)=\mathbf{X}_0'\beta_{(i)}$ 的置信水平为 $1-\alpha$ 的联合 T^2 的置信区间为

$$\mathbf{X}_0'\hat{\beta}_{(i)}\pm\sqrt{\frac{nm}{n-p-m+1}F_{m,n-p-m+1}(\alpha)}\cdot\sqrt{\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0\hat{\sigma}_{ii}}\qquad(5.2.5)$$

这里 $\hat{\beta}_{(i)}$ 为 $\hat{\beta}$ 的第 i 列， $\hat{\sigma}_{ii}$ 是 $\hat{\boldsymbol{\Sigma}}$ 的第 i 个对角元.

下面我们对未来的某个预测变量 $\mathbf{X}_0$ 对应的响应 $\mathbf{Y}_0=\beta'\mathbf{X}_0+\varepsilon_0$ 进行预测. 因为 ε_0 独立于 ε, 所以

$$\mathbf{Y}_0-\hat{\beta}'\mathbf{X}_0=\left(\beta-\hat{\beta}\right)'\mathbf{X}_0+\varepsilon_0\sim N_m\left(\mathbf{0},\left(1+\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0\right)\boldsymbol{\Sigma}\right)$$

且与 $n\hat{\boldsymbol{\Sigma}}$ 独立. 故 $\mathbf{Y}_0$ 的置信水平为 $1-\alpha$ 的预测椭球为

$$\left(\mathbf{Y}_0-\hat{\beta}'\mathbf{X}_0\right)'\left(\frac{n\hat{\boldsymbol{\Sigma}}}{n-p}\right)^{-1}\left(\mathbf{Y}_0-\hat{\beta}'\mathbf{X}_0\right)$$

$$\le\left(1+\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0\right)\frac{(n-p)m}{n-p-m+1}F_{m,n-p-m+1}(\alpha)$$

单个响应 y_{0i} 的置信水平为 $1-\alpha$ 的联合 T^2 的置信区间为

$$\hat{\beta}_{(i)}{}'\mathbf{X}_0 \pm \sqrt{\frac{nm}{n-p-m+1}F_{m,n-p-m+1}(\alpha)} \cdot \sqrt{(1+\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0)\,\hat{\sigma}_{ii}} \quad (5.2.6)$$

第六章　主成分分析及典型相关分析

在实际问题的研究中，为了全面分析问题，往往涉及众多有关的变量. 但是，变量太多不但会增加计算的复杂性，而且也给合理地分析问题和解释问题带来困难. 一般说来，虽然每个变量都提供了一定的信息，但其重要性有所不同. 实际上，在很多情况下，众多变量间有一定的相关关系，人们希望利用这种相关性对这些变量加以 “改造”，用为数较少的新变量来反映原变量所提供的大部分信息，通过对新变量的分析达到解决问题的目的. 主成分分析及典型相关分析便是在这种降维的思维下产生的处理高维数据的统计方法.

本章主要介绍主成分分析及典型相关分析的概念和方法.

§6.1　主成分分析

主成分分析的基本方法是通过构造原变量的适当的线性组合，以产生一系列互不相关的新变量，从中选出少数几个新变量并使它们含有尽可能多的原变量带有的信息，从而使得用这几个新变量代替原变量分析问题和解决问题成为可能. 当研究的问题确定之后，变量中所含 “信息” 的大小通常用该变量的方差或样本方差来度量.

6.1.1　总体主成分

1. 总体主成分的定义

设 $X_1, X_2, \cdots, X_p$ 为某实际问题所涉及的 p 个随机变量. 记 $\mathbf{X} = (X_1, X_2, \cdots, X_p)'$，其协方差矩阵为

$$\mathbf{\Sigma} \triangleq (\sigma_{ij})_{p\times p} \triangleq E[\mathbf{X} - E(\mathbf{X})][\mathbf{X} - E(\mathbf{X})]', \quad (6.1.1)$$

它是一个 p 阶半正定矩阵. 设 $\mathbf{a}_i = (a_{i1}, a_{i2}, \cdots, a_{ip})'(i = 1, 2, \cdots, p)$ 为 p 个常数向量，考虑如下线性组合，

$$\begin{cases} Y_1 \triangleq \mathbf{a}_1'\mathbf{X} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p, \\ Y_2 \triangleq \mathbf{a}_2'\mathbf{X} = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p, \\ \qquad\qquad \vdots \\ Y_p \triangleq \mathbf{a}_p'\mathbf{X} = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p. \end{cases} \quad (6.1.2)$$

易知有

$$Var(Y_i) = Var(\mathbf{a}_i'\mathbf{X}) = \mathbf{a_i'\Sigma a_i}, i = 1, 2, \cdots, p, \quad (6.1.3)$$

$$Cov(Y_i, Y_j) = Cov(\mathbf{a}_i'\mathbf{X}, \mathbf{a}_j'\mathbf{X}) = \mathbf{a_i'\Sigma a_j}, i \neq j, i, j = 1, 2, \cdots, p. \quad (6.1.4)$$

如果我们希望用 Y_1 代替原来 p 个变量 $X_1, X_2, \cdots, X_p$, 这就要求 Y_1 尽可能地反映原 p 个变量的信息. 这里, "信息" 用 Y_1 的方差来度量, 即 $\mathrm{Var}(Y_1)$ 越大, 表示 Y_1 所含的 $X_1, X_2, \cdots, X_p$ 中的信息越多. 但由 (6.1.3) 式可知, 必须对 $\mathbf{a}_i$ 加以限制, 否则 $\mathrm{Var}(Y_1)$ 无界. 最方便的限制是要求所有 $\mathbf{a}_i$ 具有单位长度, 即

$$\mathbf{a}_i'\mathbf{a}_i = 1 \tag{6.1.5}$$

因此, 我们希望在约束条件 $\mathbf{a}_1'\mathbf{a}_1 = 1$ 之下, 求 $\mathbf{a}_1$ 使 $\mathrm{Var}(Y_1)$ 达到最大, 由此 $\mathbf{a}_1$ 所确定的随机变量 $Y_1 = \mathbf{a}_1'\mathbf{X}$ 称为 $X_1, X_2, \cdots, X_p$ 的第一主成分.

如果第一主成分 Y_1 还不足以反映原变量的信息, 考虑采用 Y_2. 为了有效地反映原变量的信息, Y_1 中已有的信息就不必要再包含在 Y_2 中, 用统计的语言来讲, 要求 Y_1 与 Y_2 不相关, 即

$$Cov(Y_1, Y_2) = \mathbf{a}_1'\mathbf{\Sigma}\mathbf{a}_2 = 0. \tag{6.1.6}$$

于是, 在约束条件 $\mathbf{a}_2'\mathbf{a}_2 = 1$ 及 $\mathbf{a}_1'\mathbf{\Sigma}\mathbf{a}_2 = 0$ 之下, 求 $\mathbf{a}_2$ 使 $\mathrm{Var}(Y_2)$ 达到最大, 由此 $\mathbf{a}_2$ 所确定的随机变量 $Y_2 = \mathbf{a}_2'\mathbf{X}$ 称为 $X_1, X_2, \cdots, X_p$ 的第二主成分.

一般地, 在约束条件 $\mathbf{a}_i'\mathbf{a}_i = 1$ 及 $Cov(Y_i, Y_k) = \mathbf{a}_i'\mathbf{\Sigma}\mathbf{a}_k = 0 (k = 1, 2, \cdots, i-1)$ 之下, 求 $\mathbf{a}_i$ 使 $\mathrm{Var}(Y_i)$ 达到最大, 由此 $\mathbf{a}_i$ 所确定的 $Y_i = \mathbf{a}_i'\mathbf{X}$ 称为 $X_1, X_2, \cdots, X_p$ 的第 i 主成分.

2 . 总体主成分的求法

利用线性代数知识, 可证明如下定理.

定理 6.1.1 设 $\mathbf{\Sigma}$ 是 $\mathbf{X}=(X_1, X_2, \cdots, X_p)'$ 的协方差矩阵, $\mathbf{\Sigma}$ 的特征值及相应的单位正交化特征向量分别为 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ 及 $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_p$, 则 $\mathbf{X}$ 的第 i 个主成分为

$$Y_i = \mathbf{e}_i'\mathbf{X} = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ip}X_p, i = 1, 2, \cdots, p, \tag{6.1.7}$$

并且有

$$\begin{cases} Var(Y_i) = \mathbf{e}_i'\mathbf{\Sigma}\mathbf{e}_i = \lambda_i, i = 1, 2, \cdots, p, \\ Cov(Y_i, Y_k) = \mathbf{e}_i'\mathbf{\Sigma}\mathbf{e}_i = 0, i \neq k. \end{cases} \tag{6.1.8}$$

证明 令 $\mathbf{P} = (\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_p)$, 则 $\mathbf{P}$ 为一正交矩阵, 且

$$\mathbf{P}'\mathbf{\Sigma}\mathbf{P} = \mathbf{\Lambda} \triangleq diag(\lambda_1, \lambda_2, \cdots, \lambda_p), \tag{6.1.9}$$

其中 $diag(\lambda_1, \lambda_2, \cdots, \lambda_p)$ 表示对角矩阵.

设 $Y_1 = \mathbf{a}_1\mathbf{X}$ 为 $\mathbf{X}$ 的第一主成分, 其中 $\mathbf{a}_1'\mathbf{a}_1 = 1$. 令

$$\mathbf{Z}_1 \triangleq (Z_{11}, Z_{12}, \cdots, Z_{1p})' = \mathbf{P}'\mathbf{a}_1,$$

则

$$\begin{aligned} &Var(Y_1) = \mathbf{a}_1'\mathbf{\Sigma}\mathbf{a}_1 = \mathbf{Z}_1'\mathbf{P}'\mathbf{\Sigma}\mathbf{P}\mathbf{Z}_1 = \lambda_1 Z_{11}^2 + \lambda_2 Z_{12}^2 + \cdots + \lambda_p Z_{1p}^2 \\ &\leq \lambda_1 \mathbf{Z}_1'\mathbf{Z}_1 = \lambda_1 \mathbf{a}_1'\mathbf{P}\mathbf{P}'\mathbf{a}_1 = \lambda_1, \end{aligned}$$

并且当 $\mathbf{Z}_1=(1,0,\cdots,0)'$ 时，等号成立. 这时

$$\mathbf{a}_1=\mathbf{P}\mathbf{Z}_1=\mathbf{e}_1.$$

由此可知，在约束 $\mathbf{a}_1'\mathbf{a}_1=1$ 之下，当 $\mathbf{a}_1=\mathbf{e}_1$ 时， $Var(Y_1)$ 达到最大，且

$$\max_{a_1'a_1=1}\{Var(Y_1)\}=Var(\mathbf{e}_1'\mathbf{X})=\mathbf{e}_1'\mathbf{\Sigma}\mathbf{e}_1=\lambda_1.$$

为证明一般情况，我们先证明如下结论：若 $\mathbf{a}'\mathbf{a}=1$, 则

$$\max_{\mathbf{a}\perp\{\mathbf{e}_1,\cdots,\mathbf{e}_{i-1}\}}\{\mathbf{a}'\mathbf{\Sigma}\mathbf{a}\}=\lambda_i, i=2,3,\cdots,p, \tag{6.1.10}$$

其中 $\mathbf{a}\perp\{\mathbf{e}_1,\cdots,\mathbf{e}_{i-1}\}$ 表示 $\mathbf{a}$ 与 $\mathbf{e}_1,\mathbf{e}_2,\cdots,\mathbf{e}_{i-1}$ 均正交，即 $\mathbf{a}'\mathbf{e}_k=0(k=1,2,\cdots,i-1)$.

事实上，令 $\mathbf{Z}\triangleq(Z_1,Z_2,\cdots,Z_p)'=\mathbf{P}'\mathbf{a}$, 则

$$\mathbf{a}=\mathbf{P}\mathbf{Z}=Z_1\mathbf{e}_1+Z_2\mathbf{e}_2+\cdots+Z_p\mathbf{e}_p, \tag{6.1.11}$$

且 $\mathbf{Z}'\mathbf{Z}=\mathbf{a}'\mathbf{a}=1$. 由此可得

$$0=\mathbf{a}'\mathbf{e}_k=Z_1\mathbf{e}_1'\mathbf{e}_k+Z_2\mathbf{e}_2'\mathbf{e}_k+\cdots+Z_p\mathbf{e}_p'\mathbf{e}_k=Z_k, k=1,2,\cdots,i-1,$$

因而

$$\begin{aligned}\mathbf{a}'\mathbf{\Sigma}\mathbf{a}&=\mathbf{Z}'\mathbf{P}'\mathbf{\Sigma}\mathbf{P}\mathbf{Z}=\lambda_1Z_1^2+\cdots+\lambda_{i-1}Z_{i-1}^2+\lambda_iZ_i^2+\cdots+\lambda_pZ_p^2\\&=\lambda_iZ_i^2+\cdots+\lambda_pZ_p^2\le\lambda_i\mathbf{Z}'\mathbf{Z}=\lambda_i.\end{aligned} \tag{6.1.12}$$

若取 $\mathbf{Z}=(0,\cdots,0,1,0,\cdots,0)'$, 其中第 i 个元素为 1, 其余均为零，这时由（ 6.1.11 ）和（ 6.1.12 ）知

$$\mathbf{a}=\mathbf{e}_i且\mathbf{a}'\mathbf{\Sigma}\mathbf{a}=\lambda_i. \tag{6.1.13}$$

因此，当 $\mathbf{a}=\mathbf{e}_i$ 时，它与 $\mathbf{e}_k(1\le k\le i-1)$ 均正交，且使 $\mathbf{a}'\mathbf{\Sigma}\mathbf{a}$ 达到其最大值 λ_i.

利用（ 6.1.10 ），取 $\mathbf{a}_i=\mathbf{e}_i$, 则

$$\mathbf{a}_i'\mathbf{a}_i=1且\mathbf{a}_i\perp\{\mathbf{e}_1,\mathbf{e}_2,\cdots\mathbf{e}_{i-1}\},$$

这时 $Y_i=\mathbf{e}_i'\mathbf{X}$ 且 $Var(Y_i)=\mathbf{a}_i'\mathbf{\Sigma}\mathbf{a}_i$ 达到最大值 $\lambda_i(i=2,3,\cdots,p)$, 并且有

$$Cov(Y_i,Y_k)=\mathbf{e}_i'\mathbf{\Sigma}\mathbf{e}_k=\lambda_k\mathbf{e}_i'\mathbf{e}_k=0(i\ne k).$$

定理得证.

由此定理可知，求 $\mathbf{X}$ 的各主成分，等价于求 $\mathbf{\Sigma}$ 的各特征值及相应的单位正交化特征向量. 按特征值由大到小所对应的特征向量为组合系数的 $X_1,X_2,\cdots,X_p$ 的线性组合分别为 $\mathbf{X}$ 的第一，第二，直至第 p 个主成分，而各主成分的方差等于相应的特征值.

下面进一步讨论 $X_1, X_2, \cdots, X_p$ 的方差与各主成分的方差之间的关系，以确定各主成分所包含的信息占 $X_1, X_2, \cdots, X_p$ 中总信息的份额. 易证下面结果：

定理 6.1.2 设 $Y_i = \mathbf{e_i'X}(i = 1, 2, \cdots, p)$ 为 $\mathbf{X}$ 的 p 个主成分，则

$$\sum_{i=1}^{p} Var(X_i) = \sum_{i=1}^{p} \sigma_{ii} = \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} Var(Y_i). \tag{6.1.14}$$

证明 由（ 6.1.9 ）知，

$$\sum_{i=1}^{p} \sigma_{ii} = tr(\mathbf{\Sigma}) = tr(\mathbf{P\Lambda P'}) = tr(\mathbf{\Lambda P'P}) = tr(\mathbf{\Lambda}) = \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} Var(Y_i).$$

此定理说明， $X_1, X_2, \cdots, X_p$ 各变量的方差之和等于各个主成分的方差之和，即 $\sum\limits_{i=1}^{p} \lambda_i$ 因此， $\dfrac{\lambda_k}{\sum\limits_{i=1}^{p} \lambda_i}$ 描述了第 k 个主成分提取的信息占总信息量的份额. 我们特给出如下定义：

定义 6.1.1 称 $\dfrac{\lambda_k}{\sum\limits_{i=1}^{p} \lambda_i}$ 为第 k 个主成分 Y_k 的贡献率，称 $\dfrac{\sum\limits_{i=1}^{m} \lambda_i}{\sum\limits_{i=1}^{p} \lambda_i}$ 为前 m 个主成分 $Y_1, Y_2, \cdots, Y_m$ 的累积贡献率.

累积贡献率表明了前 m 个主成分提取了 $X_1, X_2, \cdots, X_p$ 中的总信息量的份额. 在实际应用中，主成分个数的确定方法：通常选取 $m < p$, 使前 m 个主成分的累积贡献率达到一定的比例（如 80% 到 90% ）. 或用崖底碎石图（崖底碎石图就是 $\hat{\lambda}_i$ 对序号 i 的 $(i, \hat{\lambda}_i)$ 的图. 为确定主成分的个数，我们在该图上找拐弯处，选区一个拐弯点对应的序号，次序号后的特征值全部较小且彼此大小差不多. 这样选出的号码作为主成分的个数）来判断. 这样用前 m 个主成分代替原来的变量 $X_1, X_2, \cdots, X_p$ 而不至于损失太多的信息，从而达到减少变量个数的目的.

下面我们通过具体例子说明求总体主成分的方法.

例 6.1.1 设随机变量 $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3},)'$ 的协方差矩阵为

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix},$$

求 $\mathbf{X}$ 的各主成分.

易求得 $\mathbf{\Sigma}$ 的特征值及相应的单位正交化特征向量分别为

$$\begin{aligned} &\lambda_1 = 5.83, \mathbf{e}_1' = (0.383, -0.924, 0), \\ &\lambda_2 = 2.00, \mathbf{e}_2' = (0, 0, 1), \\ &\lambda_3 = 0.17, \mathbf{e}_3' = (0.924, 0.383, 0). \end{aligned}$$

因此 $\mathbf{x}$ 的主成分为

$$\begin{aligned} Y_1 &= \mathbf{e_1'X} = 0.383X_1 - 0.924X_2, \\ Y_2 &= \mathbf{e_2'X} = X_3, \\ Y_3 &= \mathbf{e_3'X} = 0.924X_1 + 0.383X_2. \end{aligned}$$

这里 X_3 是一个主成分是因为由 $\mathbf{\Sigma}$ 可知， X_3 和 X_1, X_2 均不相关.

如果我们只取第一主成分，则贡献率为

$$\frac{5.83}{5.83 + 2.00 + 0.17} = 73\%$$

若取前两个主成分，则累积贡献率为

$$\frac{5.83 + 2.00}{5.83 + 2.00 + 0.17} = 98\%$$

因此，用前两个主成分代替原来三个变量，其信息损失是很小的．用 R 编程如下：

```
 a=c(1,-2,0,-2,5,0,0,0,2)
 b=matrix(data=a,ncol=3,nrow=3,byrow=T)
 c=eigen(b)
 c
 $values
[1] 5.8284271 2.0000000 0.1715729

$vectors
           [,1] [,2]      [,3]
[1,] -0.3826834    0 0.9238795
[2,]  0.9238795    0 0.3826834
[3,]  0.0000000    1 0.0000000
```

3. 标准化变量的主成分

在实际问题中，不同的变量往往有不同的量纲，由于不同的量纲会引起个变量取值的分散程度差异较大，这时，总体方差则主要受方差较大的变量的控制．若用 Σ 求主成分，则优先照顾了方差大的变量，有时会造成很不合理的结果．为了消除由于量纲的不同可能带来的影响，常采用变量标准化的方法，即令

$$X_i^* = \frac{X_i - \mu_i}{\sqrt{\sigma_{ii}}}, i = 1, 2, \cdots, p, \tag{6.1.15}$$

其中 $\mu_i = E(X_i), \sigma_{ii} = Var(X_i)$. 这时 $\mathbf{X}^* = (X_1^*, X_2^*, \cdots, X_p^*)'$ 的协方差矩阵便是 $\mathbf{X} = (X_1, X_2, \cdots, X_p)'$ 的相关矩阵 $\rho = (\rho_{ij})_{p\times p}$, 其中

$$\rho_{ij} = E\left[\frac{X_i - \mu_i}{\sqrt{\sigma_{ii}}}\right]\left[\frac{X_j - \mu_i}{\sqrt{\sigma_{jj}}}\right] = \frac{Cov(X_i, X_j)}{\sqrt{\sigma_{ii}\sigma_{jj}}}. \tag{6.1.16}$$

利用 $\mathbf{X}$ 的相关矩阵 ρ 作主成分分析，平行于前面的结论，我们有下面的定理.

定理 6.1.3 设 $\mathbf{X}^* = (X_1^*, X_2^*, \cdots, X_P^*)'$ 为标准化的随机向量，其协方差矩阵（即 $\mathbf{X}$ 的相关矩阵）为 ρ , 则 X^* 的第 i 个主成分为

$$Y_i^* = (\mathbf{e}_i^*)'\mathbf{X}^* = e_{i1}^* \frac{X_1 - \mu_1}{\sqrt{\sigma_{11}}} + e_{i2}^* \frac{X_2 - \mu_2}{\sqrt{\sigma_{22}}} + \cdots + e_{ip}^* \frac{X_p - \mu_p}{\sqrt{\sigma_{pp}}}, \quad i = 1, 2, \cdots, p, \tag{6.1.17}$$

并且

$$\sum_{i=1}^{p} Var(Y_i^*) = \sum_{i=1}^{p} \lambda_i^* = \sum_{i=1}^{p} Var(X_i^*) = p, \tag{6.1.18}$$

其中 $\lambda_1^* \geq \lambda_2^* \cdots \geq \lambda_p^* \geq 0$ 为 ρ 的特征值， $\mathbf{e_1^*}, \mathbf{e_2^*}, \cdots, \mathbf{e_p^*}$ 为相应的单位正交化特征向量. 这时，第 i 个主成分的贡献率为

$$\frac{\lambda_i^*}{p}, i = 1, 2, \cdots, p, \tag{6.1.19}$$

前 m 个主成分的累积贡献率为

$$\frac{\sum_{i=1}^{m} \lambda_i^*}{p}. \tag{6.1.20}$$

例 6.1.2 设 $\mathbf{X} = (X_1, X_2)'$ 协方差矩阵为

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & 4 \\ 4 & 100 \end{bmatrix},$$

相应的相关矩阵为

$$\rho = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 1 \end{bmatrix}.$$

如果从 $\mathbf{\Sigma}$ 出发作主成分分析，易求得其特征值和相应的单位正交化特征向量为

$$\lambda_1 = 100.16, \quad \mathbf{e}_1 = (0.040, 0.999)',$$
$$\lambda_2 = 0.84, \quad \mathbf{e}_2 = (0.999, -0.040)'$$

$\mathbf{X}$ 的两个主成分分别为

$$Y_1 = 0.040X_1 + 0.999X_2, Y_2 = 0.999X_1 - 0.040X_2.$$

第一主成分的贡献率为

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{100.16}{101} = 99.2\%.$$

我们看到由于 X_2 的方差很大，它完全控制了提取信息量占 99.2% 的第一主成分（ X_2 在 Y_1 中的系数为 0.999 ），淹没了变量 X_1 的作用.

如果从 ρ 出发求主成分，可求得其特征值和相应的单位正交化特征向量为

$$\lambda_1^* = 1.4, \mathbf{e}_1^* = (0.707, 0.707)',$$
$$\lambda_2^* = 0.6, \mathbf{e}_2^* = (0.707, -0.707)'.$$

$\mathbf{X}^*$ 的两个主成分分别为

$$Y_1 = 0.707X_1^* + 0.707X_2^* = 0.707(X_1 - \mu_1) + 0.0707(X_2 - \mu_2),$$
$$Y_2 = 0.707X_1^* - 0.707X_2^* = 0.707(X_1 - \mu_1) - 0.0707(X_2 - \mu_2).$$

此时，第一个主成分的贡献率有所下降，为

$$\frac{\lambda_1^*}{p} = \frac{1.4}{2} = 70\%.$$

由此看到，原变量在第一主成分中的相对重要性由于标准化而有很大的变化．在由 $\mathbf{\Sigma}$ 所求得的第一主成分中的， X_1 和 X_2 的权重系数分别为 0.040 和 0.999, 主要由大方差的变量控制．而在由 ρ 所求得的第一主成分中， X_1 和 X_2 的权重系数反而成了 0.707 和 0.0707 , 即 X_1 的相对重要性得到提升．此例也表明，由 $\mathbf{\Sigma}$ 和 ρ 求得的主成分一般是不相同的，而且，其中一组主成分也不是第二组主成分的某简单函数．用 R 语言编程如下：

```
a=matrix(c(1,4,4,100),ncol=2,nrow=2)
b=diag(1/diag(a))
c=sqrt(b)%*%a%*%sqrt(b)
a1=eigen(a)
c1=eigen(c)
a
     [,1] [,2]
[1,]    1    4
[2,]    4  100

b
     [,1] [,2]
[1,]    1 0.00
[2,]    0 0.01
c
     [,1] [,2]
[1,]  1.0  0.4
[2,]  0.4  1.0


a1
$values
[1] 100.1613532   0.8386468
```

```
$vectors
           [,1]        [,2]
[1,] 0.04030552  0.99918740
[2,] 0.99918740 -0.04030552


 c1
$values
[1] 1.4 0.6

$vectors
          [,1]       [,2]
[1,] 0.7071068  0.7071068
[2,] 0.7071068 -0.7071068
```

在实际应用中，当涉及的各变量的变化范围差异较大时，从 ρ 出发求主成分比较合理.

6.1.2 样本主成分

前面讨论的是总体主成分，但在实际问题中，一般 $\mathbf{\Sigma}$（或 ρ ）是未知的，需要通过样本来估计. 设

$$\mathbf{X}_i = (X_{i1}, X_{i2}, \cdots, X_{ip})', i = 1, 2, \cdots, n \tag{6.1.21}$$

为取自 $\mathbf{X} = (X_1, X_2, \cdots, X_p)'$ 的一个容量为 n 的简单随机样本，则样本协方差矩阵及样本相关矩阵分别为

$$\mathbf{S} \triangleq (s_{ij})_{p\times p} = \frac{1}{n-1}\sum_{k=1}^{n}(\mathbf{X_k} - \overline{\mathbf{X}})(\mathbf{X_k} - \overline{\mathbf{X}})', \tag{6.1.22}$$

$$\mathbf{R} \triangleq (r_{ij})_{p\times p} = \left(\frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}\right), \tag{6.1.23}$$

其中

$$\overline{\mathbf{X}} = (\overline{X}_1, \overline{X}_2, \cdots, \overline{X}_p)', \overline{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij}, i = 1, 2, \cdots, p,$$
$$s_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(X_{ik} - \overline{X}_i)(X_{jk} - \overline{X}_j), i, j = 1, 2, \cdots, p.$$

分别以 $\mathbf{S}$ 和 $\mathbf{R}$ 作为 $\mathbf{\Sigma}$ 和 ρ 的估计，按前面所述方法求得的主成分称为样本主成分. 具体结论如下：

定理 6.1.4 设 $\mathbf{S} = (s_{ij})_{p\times p}$ 是样本协方差矩阵，其特征值为 $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$, 相应的单位正交化特征向量为 $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2 \cdots, \hat{\mathbf{e}}_p$, 则第 i 个样本主成分为

$$Y_i = \hat{\mathbf{e}}_i'\mathbf{X} = \hat{e}_{i1}X_1 + \hat{e}_{i2}X_2 + \cdots + \hat{e}_{ip}X_p, i = 1, 2, \cdots, p, \tag{6.1.24}$$

其中 $\mathbf{X}=(X_1,X_2,\cdots,X_p)'$ 为 $\mathbf{X}$ 的任一观测值. 当依次代入 $\mathbf{X}$ 的 n 个观测值 $\mathbf{X_k}=(X_{k1},X_{k2},\cdots,X_{kp})'(k=1,2,\cdots,n)$ 时，便得到第 i 个样本主成分 Y_i 的 n 个观测值 $Y_{ik}(k=1,2,\cdots,n)$. 这时

$$\begin{cases} Y_i\text{的样本方差}=\hat{\mathbf{e}}_i'\mathbf{S}\hat{\mathbf{e}}_i=\hat{\lambda}_i, i=1,2,\cdots,p, \\ Y_i\text{与}Y_j\text{的样本协方差}=\hat{\mathbf{e}}_i'\mathbf{S}\hat{\mathbf{e}}_j=0, i\neq j, \\ \text{样本总方差}=\sum\limits_{i=1}^{p}s_{ii}=\sum\limits_{i=1}^{p}\hat{\lambda}_i. \end{cases} \tag{6.1.25}$$

第 i 个样本主成分的贡献率定义为

$$\frac{\hat{\lambda}_i}{\sum\limits_{i=1}^{p}\hat{\lambda}_i}, i=1,2,\cdots,p, \tag{6.1.26}$$

前 m 个样本主成分的累积贡献率定义为

$$\frac{\sum\limits_{i=1}^{m}\hat{\lambda}_i}{\sum\limits_{i=1}^{p}\hat{\lambda}_i}. \tag{6.1.27}$$

同样，为了消除量纲的影响，我们可以对样本进行标准化，即令

$$\mathbf{X}_i^*=\left[\frac{X_{i1}-\overline{X}_1}{\sqrt{s_{11}}},\frac{X_{i2}-\overline{X}_2}{\sqrt{s_{22}}},\cdots,\frac{X_{ip}-\overline{X}_p}{\sqrt{s_{pp}}}\right]', i=1,2,\cdots,n, \tag{6.1.28}$$

则标准化数据的样本协方差矩阵即为原数据的样本相关矩阵 $\mathbf{R}$. 由 $\mathbf{R}$ 出发所得的样本主成分称为标准化样本主成分. 只要求出 $\mathbf{R}$ 的特征值及相应的单位正交化特征向量，类似上述结果可求得标准化样本主成分. 这时标准化样本的样本总方差为 p.

实际应用中，将样本 $\mathbf{X}_i(i=1,2,\cdots,n)$ 代入各主成分中，可得到各样本主成分的观测值 $Y_{ki}(k=1,2,\cdots,n;i=1,2,\cdots,p)$. 为便于理解和比较，我们用表格形式列出原始数据及它的各主成分观测值如下：

表 6.1.1　原始数据及它的主成分观测值

序号	原变量				主成分			
	X_1	X_2	$\cdots$	X_p	Y_1	Y_2	$\cdots$	Y_p
1	X_{11}	X_{12}	$\cdots$	X_{1p}	Y_{11}	Y_{12}	$\cdots$	Y_{1p}
2	X_{21}	X_{22}	$\cdots$	X_{2p}	Y_{21}	Y_{22}	$\cdots$	X_{2p}
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	X_{n1}	X_{n2}	$\cdots$	X_{np}	Y_{n1}	Y_{n2}	$\cdots$	Y_{np}

选取前 m(m<p) 个样本主成分，使累积贡献率达到一定的要求（如 80% 到 90% ），以前 m 个样本主成分的观测值代替原始数据作统计分析，这样便可达到降低原始数据维数的目的.

例 6.1.3 彩色胶卷的显影色彩很容易受显影液微小变化的影响. 为了对显影液进行质量控制，将胶卷在不同情况下曝光，然后通过红，绿，蓝滤色片

进行测量. 测量在高，中，低三种密度下进行，故每一胶卷共有 9 个指标，共做了 108 个试验（数据略）. 根据数据可求得样本协方差矩阵 **S** 为（为便于理解，我们写出了 **S** 各列所对应的指标，各行对应的指标相同，而 **S** 的对称部分未写出）

高			中			低		
红	绿	蓝	红	绿	蓝	红	绿	蓝
177	179	95	96	53	32	-7	-4	-3
	419	245	131	181	127	-2	1	4
		302	60	109	142	4	4	11
			153	102	42	4	3	2
				137	96	4	5	6
					128	2	2	8
						34	31	33
							39	39
								48

求出 **S** 的特征值和相应的单位正交化特征向量可得样本主成分的累积贡献率如下表：

i	$\hat{\lambda}_i$	累计贡献率 (%)
1	878.52	60.92
2	196.10	74.52
3	128.64	83.44
4	103.43	90.62
5	81.26	96.25
6	37.85	98.88
7	6.98	99.36
8	5.71	99.76
9	3.52	100.00

由此看到，前三个样本主成分提取的信息达 83.44%. 故我们可利用前三个样本主成分分析此问题.

将样本的前三个主成分的系数（即特征向量 $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \hat{\mathbf{e}}_3$ ）列于下表：

		$\hat{\mathbf{e}}_1$	$\hat{\mathbf{e}}_2$	$\hat{\mathbf{e}}_3$
	红	0.305	-0.485	-0.412
高	绿	0.654	-0.150	-0.182
	蓝	0.482	0.587	-0.235
	红	0.261	-0.491	0.457
中	绿	0.323	-0.038	0.495
	蓝	0.271	0.376	0.268
	红	0.002	0.057	0.256
低	绿	0.006	0.053	0.266
	蓝	0.014	0.088	0.282

由此系数可知，第一个样本主成分的信息主要来源于前六个变量，低密度时的三个变量的系数很小；第二个主成分的信息主要反映了高红，高蓝，中红和中蓝四个变量，注意到红和蓝的系数的符号相反，故第二主成分主要反映了红和蓝的对比；第三个主成分反映了高密度的各变量和中，低密度的对比.

另外，从样本协方差矩阵 $\mathbf{S}$ 可知，最后三个变量的样本方差相对于前六个变量的样本方差是比较小的，因此第一主成分主要由前六个变量起主导作用. 这时，采用样本相关矩阵求标准化样本主成分将会更加合理，我们将这一分析留作练习.

例 6.1.4 从 1975 年 1 月至 1976 年 12 月，对纽约证券交易所的五种股票（ Allied Chemical,du Pont,Union Carbide,Exxon 和 Texaco ）的周反弹率进行连续 100 周的观测，其中周反弹率 =（本周五收盘价 - 上周五收盘价）/ 上周五收盘价（具体数据可参看参考文件 [3]$p293, 294$ ）. 由于在一般经济条件下，股票有集聚的趋势，因此各股票的周反弹率存在相关关系. 设 $\mathbf{X} = (X_1, X_2, X_3, X_4, X_5)'$ 表示这五种股票的周反弹率，利用这 100 个观测值 $\mathbf{X_1}, \mathbf{X_2} \cdots, \mathbf{X_{100}}$ 可求得样本均值向量为

$$\overline{\mathbf{X}} = (0.0054, 0.0048, 0.0057, 0.0063, 0.0037)',$$

其样本相关矩阵为

$$\mathbf{R} = \begin{bmatrix} 1.000 & 0.577 & 0.509 & 0.387 & 0.462 \\ 0.577 & 1.000 & 0.599 & 0.389 & 0.322 \\ 0.509 & 0.599 & 1.000 & 0.436 & 0.426 \\ 0.387 & 0.389 & 0.436 & 1.000 & 0.523 \\ 0.462 & 0.322 & 0.426 & 0.523 & 1.000 \end{bmatrix},$$

由 $\mathbf{R}$ 出发作主成分分析.

令

$$X_i^* = \frac{X_i - \overline{X}_i}{\sqrt{s_{ii}}}, i = 1, 2, 3, 4, 5$$

为标准化样本，则 $\mathbf{R}$ 是 $\mathbf{X}^* = (X_1^*, X_2^*, \cdots, X_5^*)'$ 的样本协方差矩阵. 可求得

R 的特征值和相应的单位正交化特征向量为

$$\begin{aligned}
&\hat{\lambda}_1^* = 2.857, \hat{\mathbf{e}}_1^* = (0.464, 0.457, 0.470, 0.421, 0.421)',\\
&\hat{\lambda}_2^* = 0.809, \hat{\mathbf{e}}_2^* = (0.240, 0.509, 0.260, -0.526, -0.582)',\\
&\hat{\lambda}_3^* = 0.540, \hat{\mathbf{e}}_3^* = (-0.612, 0.178, 0.335, 0.541, -0.435)',\\
&\hat{\lambda}_4^* = 0.452, \hat{\mathbf{e}}_4^* = (0.387, 0.206, -0.662, 0.472, -0.382)',\\
&\hat{\lambda}_5^* = 0.343, \hat{\mathbf{e}}_5^* = (-0.451, 0.676, -0.400, -0.176, 0.385)'.
\end{aligned}$$

由此看到，前两个样本主成分的累计贡献率为

$$\frac{\hat{\lambda}_1^* + \hat{\lambda}_2^*}{p} = \frac{2.857 + 0.809}{5} = 73\%.$$

若用标准化样本变量 $X_i^*(i = 1, 2, \cdots, 5)$ 表示，前两个标准化样本主成分为

$$\begin{aligned}
y_1^* &= (\hat{\mathbf{e}}_1^*)'\mathbf{X}^* = 0.464X_1^* + 0.457X_2^* + 0.470X_3^* + 0.421X_4^* + 0.421X_5^*,\\
y_2^* &= (\hat{\mathbf{e}}_2^*)'\mathbf{X}^* = 0.240X_1^* + 0.509X_2^* + 0.260X_3^* - 0.526X_4^* - 0.582X_5^*.
\end{aligned}$$

这两个样本主成分有很明显的经济意义，第一主成分基本上是五个变量的等权重之和，它可以被简单的称之为市场成分．第二个主成分反映了化工股票（ Allied Chemical,du Pont,Union Carbide ）和石油股票（ Exxon,Texaco ）的对照，该量可以被称为工业成分．因此我们看到，这五种股票的周反弹率的大部分信息可由市场活动和与之不相关的工业活动反映．这种解释与经济学的结实是一致的．

其余主成分是不易解释的，它们可能反映了这五种股票各自的特点．但不管怎样，它们所包含的信息量是比较少的．

最后需要指出的是，虽然利用主成分本身可对所涉及的变量之间的关系在一定程度上作分析，但这往往并不意味着分析问题的结束．主成分分析只是作为对原问题进行统计分析的中间步骤，目的是利用主成分分析选取能反映原数据的大部分信息的几个主成分变量，以它们的观测值代替原始数据作分析，达到减少变量个数的目的．如利用主成分变量作回归分析（即主成分分析），聚类分析，等等．

§6.2 典型相关分析

典型相关分析着眼于识别和量化两组随机变量之间的相关关系，它是两个随机变量之间的相关关系在两组变量下的推广．

我们知道，两个随机变量 X,Y 的相关性可用它们的相关系数来度量，其定义为

$$\rho_{X,Y} = \frac{Cov(X, Y)}{\sqrt{Var(X) \cdot Var(Y)}}. \tag{6.2.1}$$

但在许多实际问题中，需要研究多个变量与多个变量间的相关关系．例如，工厂对原料的主要质量指标 $(X_1, X_2, \cdots, X_p)$ 进行测量，然后对产品的主要质量指标 $(Y_1, Y_2, \cdots, Y_q)$ 也进行测量，质量管理人员希望对 $(X_1, X_2, \cdots, X_p)$

与 $(Y_1, Y_2, \cdots, Y_q)$ 之间的相关性有所了解以提高产品质量；又如，在生物科学中，在研究某生物群状况 (用一组变量 $(X_1, X_2, \cdots, X_p)$ 描述) 与其生活环境 (用另一组变量 $(Y_1, Y_2, \cdots, Y_q)$ 刻画) 之间的关系，这对于保持生态平衡有指导意义. 虽然每个 $X_i (i = 1, 2, \cdots, p)$ 与每个 $Y_i (i = 1, 2, \cdots, q)$ 之间的相关关系也反映了两组变量中各对之间的联系，但不能反映这两组变量整体之间的相关性. 另外，当 p 和 q 较大时，只孤立地了解各对 (X_i, Y_j) 的相关性，也无助于实际问题的全面分析和解决.

受主成分分析思想的启发，我们可以把两组变量的相关性转化为两个变量的相关性来考虑，即考察一组变量的线性组合

$$Z = a_1X_1 + a_2X_2 + \cdots + a_pX_p = \mathbf{a}'\mathbf{X}$$

与另一组变量的线形组合

$$W = b_1Y_1 + b_2Y_2 + \cdots + b_qY_q = \mathbf{b}'\mathbf{Y}$$

的相关性. 为最大可能的提取 $(X_1, X_2, \cdots, X_p)$ 与 $(Y_1, Y_2, \cdots, Y_q)$ 之间的相关性，我们选择 $\mathbf{a}$ 和 $\mathbf{b}$, 使 Z 与 W 之间有最大的相关系数，这时称 Z 和 W 为第一对典型变量. 进一步，我们还可以确定第二对，第三对典型变量等等，并使各对典型变量之间互不相关 (即相关性不会被各对典型变量重复提这样，我们就将两组变量间的相关性凝结为少数几个典型变量对之间的相关性，通过对相关性较大的几个典型变量的研究来了解原来两组之间的相关关系，从而容易抓住问题的本质.

6.2.1 总体的典型变量与典型相关

设两组随机变量分别为

$$\mathbf{X} = (X_1, X_2, \cdots, X_p)', \mathbf{Y} = (Y_1, Y_2, \cdots, Y_q)',$$

令

$$\begin{cases} Cov(\mathbf{X}) = E[\mathbf{X} - E(\mathbf{X})][\mathbf{X} - E(\mathbf{X})]' \triangleq \mathbf{\Sigma_{11}}, \\ Cov(\mathbf{Y}) = E[\mathbf{Y} - E(\mathbf{Y})][\mathbf{Y} - E(\mathbf{Y})]' \triangleq \mathbf{\Sigma_{22}}, \\ Cov(\mathbf{X}, \mathbf{Y}) = E[\mathbf{X} - E(\mathbf{X})][\mathbf{Y} - E(\mathbf{Y})]' \triangleq \mathbf{\Sigma_{12}}, \\ Cov(\mathbf{Y}, \mathbf{X}) = E[\mathbf{Y} - E(\mathbf{Y})][\mathbf{X} - E(\mathbf{X})]' \triangleq \mathbf{\Sigma_{21}}. \end{cases} \tag{6.2.2}$$

则有 $\mathbf{\Sigma_{12}} = \mathbf{\Sigma_{21}'}$. 进一步假定 $\mathbf{\Sigma_{11}}$ 和 $\mathbf{\Sigma_{22}}$ 是满秩阵 (从而是正定矩阵)，令

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma_{11}} & \mathbf{\Sigma_{12}} \\ \mathbf{\Sigma_{21}} & \mathbf{\Sigma_{22}} \end{bmatrix}, \tag{6.2.3}$$

则 $\mathbf{\Sigma}$ 是 $(X_1, X_2, \cdots, X_p, Y_1, Y_2, \cdots, Y_q)'$ 的协方差矩阵，且不失一般性，可设 $p \le q$.

为研究 $\mathbf{X}$ 和 $\mathbf{Y}$ 的相关关系，考虑两组变量的线性组合

$$\begin{cases} U \triangleq \mathbf{a}'\mathbf{X} = a_1X_1 + a_2X_2 + \cdots + a_pX_p, \\ V \triangleq \mathbf{b}'\mathbf{Y} = b_1Y_1 + b_2Y_2 + \cdots + b_qY_q. \end{cases} \tag{6.2.4}$$

下面计算 U 和 V 的相关系数. 由于

$$\begin{cases} Var(U) = Var(\mathbf{a'X}) = \mathbf{a'\Sigma_{11}a} \\ Var(V) = Var(\mathbf{b'Y}) = \mathbf{b'\Sigma_{22}b} \\ Cov(U,V) = Cov(\mathbf{a'X}, \mathbf{b'Y}) = \mathbf{a'\Sigma_{12}b} \end{cases} \tag{6.2.5}$$

则 U 和 V 的相关系数为

$$\rho_{U,V} = \frac{\mathbf{a'\Sigma_{12}b}}{\sqrt{\mathbf{a'\Sigma_{11}a}}\sqrt{\mathbf{b'\Sigma_{22}b}}} \tag{6.2.6}$$

典型相关分析即确定 $\mathbf{a}$ 和 $\mathbf{b}$ 使得 $\rho_{U,V}$ 达到最大. 由相关系数的性质可知, 对任何非零常数 c, 有

$$\rho_{cU,cV} = \frac{Cov(cU,cV)}{\sqrt{Var(cU)}\sqrt{Var(cV)}} = \frac{c^2Cov(U,V)}{\sqrt{c^2Var(U)}\sqrt{c^2Var(V)}} = \rho_{U,V} \tag{6.2.7}$$

即给 $\mathbf{a}, \mathbf{b}$ 同时乘以非零常数 c,U 和 V 的相关系数不变, 故可对 $\mathbf{a}$ 和 $\mathbf{b}$ 加以适当约束, 以保证其惟一性. 由 (6.2.6) 可以看出, 使 $\rho_{U,V}$ 有最简单表示的约束为

$$\mathbf{a'\Sigma_{11}a} = 1, \mathbf{b'\Sigma_{22}b} = 1. \tag{6.2.8}$$

这等价于规定

$$Var(U) = Var(V) = 1.$$

于是典型相关分析即在约束条件 (6.2.8) 下, 确定 $\mathbf{a}$ 和 $\mathbf{b}$, 使 (6.2.6) 达到最大. 这时, 称 U,V 为典型变量. 如果只有一对 U,V 还不足以反映 $\mathbf{X}$ 和 $\mathbf{Y}$ 之间的相关性, 可进一步构造与 U,V 互不相关的另外一对典型变量, 如此等等. 具体地, 各对典型变量的定义如下:

第一对典型变量是 $U_1 = \mathbf{a_1'X}$和$V_1 = \mathbf{b_1'Y}$, 其中 U_1 和 V_1 具有单位方差且使 U_1 和 V_1 的相关系数达到最大.

第二对典型变量是 $U_2 = \mathbf{a_2'X}$和$V_2 = \mathbf{b_2'Y}$, 其中 U_2 和 V_2 具有单位方差且 U_2, V_2 和 U_1, V_1 均不相关, 即

$$Cov(U_2, U_1) = Cov(U_2, V_1) = Cov(V_2, U_1) = Cov(V_2, V_1) = 0. \tag{6.2.9}$$

在上述约束条件下并使 U_2 和 V_2 的相关系数达到最大.

一般地, 第 k 对 $(k \le p \le q)$ 典型变量是 $U_k = \mathbf{a_k'X}$ 和 $V_k = \mathbf{b_k'Y}$, 其中 U_k 和 V_k 具有单位方差, 且与前面 k-1 对典型变量中的每个 $U_i, V_i (i = 1, 2, \cdots, k-1)$ 均不相关, 在此条件下并使 U_k 和 V_k 的相关系数达到最大.

我们称第 k 对典型变量间的相关系数为第 k 个典型相关系数.

利用推导主成分类似的方法 (即定理 6.1.1 的证明), 可以给出各典型变量的具体表达式和相应的典型相关系数. 为便于理解后述定理的内容, 我们首先介绍正定矩阵的平方根矩阵的概念及其简单性质.

设 $\mathbf{A}$ 为 p 阶对称正定矩阵, 令 $\mathbf{P} = (\mathbf{e_1}, \mathbf{e_2}, \cdots, \mathbf{e_p})$, 其中 $\mathbf{e_i} (i = 1, 2, \cdots, p)$ 为 $\mathbf{A}$ 的 p 个单位正交化特征向量, $\lambda_i (i = 1, 2, \cdots, p)$ 为相应的特征值, 则 $\lambda_i > 0 (i = 1, 2, \cdots, n)$. 由线性代数知识知 $\mathbf{P}$ 为正交矩阵且

$$\mathbf{A} = \mathbf{P\Lambda P'}, \tag{6.2.10}$$

其中 $\mathbf{\Lambda} = diag(\lambda_1, \lambda_2, \cdots, \lambda_p)$. 令

$$\mathbf{\Lambda}^{\frac{1}{2}} = diag(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \cdots, \sqrt{\lambda_p}) = \begin{bmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sqrt{\lambda_p} \end{bmatrix}, \quad (6.2.11)$$

则 $\mathbf{A}$ 的平方根矩阵定义为

$$\mathbf{A}^{\frac{1}{2}} \triangleq \mathbf{P}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{P}'. \quad (6.2.12)$$

易证 $\mathbf{A}^{\frac{1}{2}}$ 有下列性质：

(i)$(\mathbf{A}^{\frac{1}{2}})' = \mathbf{A}^{\frac{1}{2}}$,即$\mathbf{A}^{\frac{1}{2}}$是对称矩阵;

(ii)$\mathbf{A}^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}} = \mathbf{P}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{P}'\mathbf{P}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{P}' = \mathbf{P}\mathbf{\Lambda}\mathbf{P}' = \mathbf{A}$;

(iii)$(\mathbf{A}^{\frac{1}{2}})^{-1} = \mathbf{P}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{P}'$,其中$\mathbf{\Lambda}^{-\frac{1}{2}} = (\mathbf{\Lambda}^{\frac{1}{2}})^{-1} = diag(\frac{1}{\sqrt{\lambda_1}}, \frac{1}{\sqrt{\lambda_2}}, \cdots, \frac{1}{\sqrt{\lambda_p}})$,

通常记$(\mathbf{A}^{\frac{1}{2}})^{-1}$为$\mathbf{A}^{-\frac{1}{2}}$;

(iv)$\mathbf{A}^{-\frac{1}{2}}\mathbf{A}^{-\frac{1}{2}} = \mathbf{A}^{-1}$.

有了矩阵平方根的概念，我们不加证明地给出总体典型相关分析的主要结果如下：

定理 6.2.1 $\mathbf{X} = (X_1, X_2, \cdots, X_p)', \mathbf{Y} = (Y_1, Y_2, \cdots, Y_q)', Cov(\mathbf{X}) = \mathbf{\Sigma_{11}}$, $Cov(\mathbf{Y}) = \mathbf{\Sigma_{22}}, Cov(\mathbf{X}, \mathbf{Y}) = \mathbf{\Sigma_{12}}$, 其中 $\mathbf{\Sigma_{11}}$ 和 $\mathbf{\Sigma_{22}}$ 均为满秩阵且 $p \le q$. 则 $\mathbf{X}, \mathbf{Y}$ 的第 k 对典型变量为

$$U_k = \mathbf{e_k'}\mathbf{\Sigma_{11}^{-\frac{1}{2}}}\mathbf{X}, V_k = \mathbf{f_k'}\mathbf{\Sigma_{22}^{-\frac{1}{2}}}\mathbf{Y}, k = 1, 2, \cdots, p. \quad (6.2.13)$$

其典型相关系数为

$$\rho_{U_k, V_k} = \rho_k, k = 1, 2, \cdots, p, \quad (6.2.14)$$

其中 $\rho_1^2 \ge \rho_2^2 \ge \cdots \ge \rho_p^2$ 为 p 阶矩阵

$$\mathbf{A} \triangleq \mathbf{\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}} \quad (6.2.15)$$

的特征值， $\mathbf{e_1}, \mathbf{e_2}, \cdots, \mathbf{e_p}$ 为相应的单位正交化特征向量， $\mathbf{f_1}, \mathbf{f_2}, \cdots, \mathbf{f_p}$ 为 q 阶矩阵

$$\mathbf{B} \triangleq \mathbf{\Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-\frac{1}{2}}} \quad (6.2.16)$$

的对应于前 p 个最大特征值（按由大到小的次序排列）的单位正交化特征向量，且 $\mathbf{f}_i = \frac{1}{\rho_i}\mathbf{\Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}e_i}$. 并且典型变量 U_k 和 $V_k (k = 1, 2, \cdots, p)$ 有如下性质：

$$\begin{cases} Var(U_k) = Var(V_k) = 1, k = 1, 2, \cdots, p, \\ Cov(U_k, U_l) = 0, k \ne l, \\ Cov(V_k, V_l) = 0, k \ne l, \\ Cov(U_k, V_l) = 0, k \ne l. \end{cases} \quad (6.2.17)$$

利用线性代数知识可知，$\mathbf{A}$ 和 $\mathbf{B}$ 有相同的非零特征值, 因此 $\mathbf{f_1}, \mathbf{f_2}, \cdots, \mathbf{f_p}$ 即是对应于 $\mathbf{B}$ 的前 p 个最大特征值 $\rho_1^2 \ge \rho_2^2 \ge \cdots \ge \rho_p^2$ 的单位正交化特征向量.

若 $\mathbf{Ae} = \lambda \mathbf{e}$, 即 $\mathbf{\Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}Ae} = \lambda \mathbf{\Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}e}$, 两边左乘以矩阵 $\mathbf{\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}}$ 得

$$\mathbf{\Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-\frac{1}{2}}\Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}e} = \lambda(\mathbf{\Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}e})$$

即 $\mathbf{B\Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}e} = \lambda(\mathbf{\Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}e})$

由此得, 若 $\rho_1^2 \geq \rho_2^2 \geq \cdots \geq \rho_p^2$ 为 p 阶矩阵 $\mathbf{A}$ 对应的特征值， $\mathbf{e_1}, \cdots, \mathbf{e_p}$ 为相应的单位正交化特征向量, 则 $\rho_1^2 \geq \rho_2^2 \geq \cdots \geq \rho_p^2$ 为 q 阶矩阵 $\mathbf{B}$ 的前 p 个特征值， $\mathbf{\Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}e_i} \triangleq \mathbf{f_i^*}$ 为第 i 个特征值对应的特征向量， $i = 1, \cdots, p$. 又因为

$$\mathbf{f_i^{*'}f_i^*} = \mathbf{e_i'\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-\frac{1}{2}}\Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}e_i}$$

$$= \mathbf{e_i'\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}e_i}$$

$$= \mathbf{e_i'Ae_i} = \rho_i^2\mathbf{e_i'e_i} = \rho_i^2$$

将 $\mathbf{f}_i^*$ 单位化后得 $\mathbf{f}_i = \frac{1}{\rho_i}\mathbf{\Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}e_i}$ 为矩阵 $\mathbf{B}$ 对应的第 i 个特征值对应的单位正交化特征向量， $i = 1, 2, \cdots, p$.

$$Cov(U_k, V_k) = Cov(\mathbf{e_k'\Sigma_{11}^{-\frac{1}{2}}X}, \mathbf{f_k'\Sigma_{22}^{-\frac{1}{2}}Y})$$

$$= \mathbf{e_k'\Sigma_{11}^{-\frac{1}{2}}}Cov(\mathbf{X}, \mathbf{Y})\mathbf{\Sigma_{22}^{-\frac{1}{2}}f_k}$$

$$= \mathbf{e_k'\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-\frac{1}{2}}}\frac{1}{\rho_k}\mathbf{\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}e_k}$$

$$= \frac{1}{\rho_k}\mathbf{e_k'\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}e_k}$$

$$= \frac{1}{\rho_k}\mathbf{e_k'Ae_k} = \frac{1}{\rho_k}\mathbf{e}_k'\rho_k^2\mathbf{e}_k$$

$$= \rho_k\mathbf{e}_k'\mathbf{e}_k = \rho_k$$

$\forall k = 1, \cdots, p,$

$$Var(U_k) = \mathbf{e_k'\Sigma_{11}^{-\frac{1}{2}}}Cov(X^{(1)})\mathbf{\Sigma_{11}^{-\frac{1}{2}}e_k} = \mathbf{e_k'\Sigma_{11}^{-\frac{1}{2}}\Sigma_{11}\Sigma_{11}^{-\frac{1}{2}}e_k} = \mathbf{e_k'e_k} = 1$$

$$Var(V_k) = \mathbf{f_k'\Sigma_{22}^{-\frac{1}{2}}}Cov(X^{(2)})\mathbf{\Sigma_{22}^{-\frac{1}{2}}f_k} = \mathbf{f_k'\Sigma_{22}^{-\frac{1}{2}}\Sigma_{22}\Sigma_{22}^{-\frac{1}{2}}f_k} = \mathbf{f_k'f_k} = 1$$

$k \neq l$,

$$Cov(U_k, V_l) = Cov(\mathbf{e_k'}\boldsymbol{\Sigma}_{\mathbf{11}}^{-\frac{1}{2}}X^{(1)}, \mathbf{f_k'}\boldsymbol{\Sigma}_{\mathbf{22}}^{-\frac{1}{2}}X^{(2)})$$

$$= \mathbf{e_k'}\boldsymbol{\Sigma}_{\mathbf{11}}^{-\frac{1}{2}}Cov(X^{(1)}, X^{(2)})\boldsymbol{\Sigma}_{\mathbf{22}}^{-\frac{1}{2}}\mathbf{f_l} = \mathbf{e_k'}\boldsymbol{\Sigma}_{\mathbf{11}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{12}}\boldsymbol{\Sigma}_{\mathbf{22}}^{-\frac{1}{2}}\mathbf{f_l}$$

$$= \mathbf{e_k'}\boldsymbol{\Sigma}_{\mathbf{11}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{12}}\boldsymbol{\Sigma}_{\mathbf{22}}^{-\frac{1}{2}}\frac{1}{\rho_l}\boldsymbol{\Sigma}_{\mathbf{22}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{21}}\boldsymbol{\Sigma}_{\mathbf{11}}^{-\frac{1}{2}}\mathbf{e_l}$$

$$= \frac{1}{\rho_l}\mathbf{e_k'}\mathbf{A}\mathbf{e_l} = \frac{1}{\rho_l}\mathbf{e}_k'\rho_l^2\mathbf{e_l} = \rho_l\mathbf{e}_k'e_l = 0$$

当 $k \neq l$ 时，同理 $Cov(U_k, U_l) = 0, Cov(V_k, V_l) = 0$.

如果对 $\mathbf{X}$ 和 $\mathbf{Y}$ 的各分量进行标准化，得

$$\mathbf{X}^* = (X_1^*, X_2^*, \cdots, X_p^*)', \mathbf{Y}^* = (Y_1^*, Y_2^*, \cdots, Y_q^*)',$$

其中

$$X_i^* = \frac{X_i - E(X_i)}{\sqrt{Var(X_i)}}, i = 1, 2, \cdots, p,$$
$$Y_j^* = \frac{Y_j - E(Y_j)}{\sqrt{Var(Y_j)}}, j = 1, 2, \cdots, q.$$

则有

$$Var(\mathbf{X}^*) = \boldsymbol{\rho}_{\mathbf{11}}, Var(\mathbf{Y}^*) = \boldsymbol{\rho}_{\mathbf{22}}, Cov(\mathbf{X}^*, \mathbf{Y}^*) = \rho_{12} = \rho_{21}',$$

其中 $\boldsymbol{\rho}_{\mathbf{11}}, \boldsymbol{\rho}_{\mathbf{22}}$ 分别为 $\mathbf{X}^*$ 和 $\mathbf{Y}^*$ 的相关矩阵，而

$$\rho = \begin{bmatrix} \rho_{11} & \rho_{12} \\ \rho_{21} & \rho_{22} \end{bmatrix}$$

为 $(X_1^*, \cdots, X_p^*, Y_1^*, \cdots, Y_q^*)'$ 的相关矩阵.

从 ρ 出发作典型相关分析，有类似于前述的结果. 即第 k 对典型相关变量为

$$\begin{cases} U_k^* = (\mathbf{a}_k^*)'\mathbf{X}^* = (\mathbf{e_k^*})'\boldsymbol{\rho}_{\mathbf{11}}^{-\frac{1}{2}}\mathbf{X}^*, \\ V_k^* = (\mathbf{b}_k^*)'\mathbf{Y}^* = (\mathbf{f_k^*})'\boldsymbol{\rho}_{\mathbf{22}}^{-\frac{1}{2}}\mathbf{Y}^*, \end{cases} \tag{6.2.18}$$

典型相关系数为

$$\rho_{U_k^*V_k^*} = \rho_k^*, k = 1, 2, \cdots, p,$$

其中 $\rho_1^{*2} \geq \rho_2^{*2} \geq \cdots \geq \rho_p^{*2}$ 是矩阵 $\boldsymbol{\rho}_{\mathbf{11}}^{-\frac{1}{2}}\boldsymbol{\rho}_{\mathbf{12}}\boldsymbol{\rho}_{\mathbf{22}}^{-\mathbf{1}}\boldsymbol{\rho}_{\mathbf{21}}\boldsymbol{\rho}_{\mathbf{11}}^{-\frac{1}{2}}$ 的 p 个特征值（从而也是矩阵 $\boldsymbol{\rho}_{\mathbf{22}}^{-\frac{1}{2}}\boldsymbol{\rho}_{\mathbf{21}}\boldsymbol{\rho}_{\mathbf{11}}^{-\mathbf{1}}\boldsymbol{\rho}_{\mathbf{12}}\boldsymbol{\rho}_{\mathbf{22}}^{-\frac{1}{2}}$ 的前 p 个最大特征值），$\mathbf{e_i^*}$ 和 $\mathbf{f_i^*}$ 分别为上述两矩阵对应于 ρ_i^{*2} 的单位正交化特征向量.

例 6.2.1 设 $\mathbf{X}^* = (X_1^*, X_2^*)'$ 和 $\mathbf{Y}^* = (Y_1^*, Y_2^*)'$ 是两组标准化随机变量. 已知 $(X_1^*, X_2^*, Y_1^*, Y_2^*)'$ 的相关矩阵为

$$\rho = \begin{bmatrix} \rho_{11} & \rho_{12} \\ \rho_{21} & \rho_{22} \end{bmatrix} = \begin{bmatrix} 1.0 & 0.4 & 0.5 & 0.6 \\ 0.4 & 1.0 & 0.3 & 0.4 \\ 0.5 & 0.3 & 1.0 & 0.2 \\ 0.6 & 0.4 & 0.2 & 1.0 \end{bmatrix}.$$

试求 $\mathbf{X}^*$ 和 $\mathbf{Y}^*$ 的典型变量和典型相关系数.

通过计算可得

$$\rho_{11}^{-\frac{1}{2}}=\begin{bmatrix}1.0681 & -0.2229\\ -0.2229 & 1.0681\end{bmatrix}, \rho_{22}^{-1}=\begin{bmatrix}1.0417 & -0.2083\\ -0.2083 & 1.0417\end{bmatrix},$$
$$\boldsymbol{\rho_{11}^{-\frac{1}{2}}\rho_{12}\rho_{22}^{-1}\rho_{21}\rho_{11}^{-\frac{1}{2}}}=\begin{bmatrix}0.4371 & 0.2178\\ 0.2178 & 0.1096\end{bmatrix}.$$

令

$$\begin{vmatrix}0.4371-\lambda & 0.2178\\ 0.2178 & 0.1096-\lambda\end{vmatrix}=\lambda^2-0.5467\lambda+0.0005=0.$$

可求得矩阵 $\boldsymbol{\rho_{11}^{-\frac{1}{2}}\rho_{12}\rho_{22}^{-1}\rho_{21}\rho_{11}^{-\frac{1}{2}}}$ 的特征值为

$$\lambda_1=\rho_1^{*2}=0.5458, \lambda_2=\rho_2^{*2}=0.0009.$$

解方程组

$$\begin{bmatrix}0.4371 & 0.2178\\ 0.2178 & 0.1096\end{bmatrix}\mathbf{e}=0.5458\mathbf{e},$$

得相应于 $\rho_1^{*2}=0.5458$ 的单位化特征向量为

$$\mathbf{e}_1^*=(0.8947, 0.4466)',$$

从而

$$\mathbf{a_1^*}=\boldsymbol{\rho_{11}^{-\frac{1}{2}}}\mathbf{e_1^*}=(0.8561, 0.2776)'.$$

同理可求得

$$\mathbf{b_1^*}=(0.5448, 0.7366)'.$$

因此第一对典型相关变量为

$$\begin{cases}U_1=(\mathbf{a_1^*})'\mathbf{X}=0.8561X_1^*+0.2776X_2^*,\\ V_1=(\mathbf{b_1^*})'\mathbf{Y}=0.5448Y_1^*+0.7366Y_2^*.\end{cases}$$

U_1 和 V_1 的第一个典型相关系数为

$$\rho_1^*=\sqrt{\rho_1^{*2}}=\sqrt{0.5458}=0.74.$$

类似地，可求出第二对典型相关变量 U_2 和 V_2, 其第二个典型相关系数为

$$\rho_2^*=\sqrt{0.0009}=0.03.$$

由于 ρ_2^* 非常小，因此，第二对典型变量只提供除第一对典型变量所提取的 $\mathbf{X}^*, \mathbf{Y}^*$ 相关信息以外的很小一部分相关信息，它对于分析 $\mathbf{X}^*, \mathbf{Y}^*$ 之间的相关性是不重要的.

6.2.2 样本的典型变量与典型相关

设 $\begin{pmatrix}\mathbf{X_i}\\ \mathbf{Y_i}\end{pmatrix}(i=1,2,\cdots,n)$ 为来自总体 $\begin{pmatrix}\mathbf{X}\\ \mathbf{Y}\end{pmatrix}$ 的一个容量为 n 的样本, 其中 $\mathbf{X_i}=(X_{i1},X_{i2},\cdots,X_{ip})',\mathbf{Y_i}=(Y_{i1},Y_{i2},\cdots,Y_{iq})'(i=1,2,\cdots,n)$, 则样本协方差矩阵为

$$\mathbf{S}_{(\mathbf{p+q})\times(\mathbf{p+q})}=\begin{pmatrix}\mathbf{S}_{11(p\times p)} & \mathbf{S}_{12(p\times q)}\\ \mathbf{S}_{21(q\times p)} & \mathbf{S}_{22(q\times q)}\end{pmatrix}, \tag{6.2.19}$$

其中

$$\begin{cases}\mathbf{S_{11}}=\frac{1}{n-1}\sum\limits_{i=1}^{n}(\mathbf{X_i}-\overline{\mathbf{X}})(\mathbf{X_i}-\overline{\mathbf{X}})',\overline{\mathbf{X}}=\frac{1}{n}\sum\limits_{i=1}^{n}\mathbf{X_i},\\ \mathbf{S_{22}}=\frac{1}{n-1}\sum\limits_{i=1}^{n}(\mathbf{Y_i}-\overline{\mathbf{Y}})(\mathbf{Y_i}-\overline{\mathbf{Y}})',\overline{\mathbf{Y}}=\frac{1}{n}\sum\limits_{i=1}^{n}\mathbf{Y_i},\\ \mathbf{S_{12}}=\frac{1}{n-1}\sum\limits_{i=1}^{n}(\mathbf{X_i}-\overline{\mathbf{X}})(\mathbf{Y_i}-\overline{\mathbf{Y}})'=\mathbf{S}_{21}'.\end{cases} \tag{6.2.20}$$

以 $\mathbf{S_{11}},\mathbf{S_{12}},\mathbf{S_{22}},\mathbf{S_{21}}$ 分别代替定理 6.2.1 中的 $\mathbf{\Sigma_{11}},\mathbf{\Sigma_{12}},\mathbf{\Sigma_{22}},\mathbf{\Sigma_{21}}$ 而得到的典型变量称为样本典型变量, 相应的典型相关系数称为样本典型相关系数. 具体地说, 样本典型变量为

$$\begin{cases}\hat{U}_k\triangleq\hat{\mathbf{a}}_\mathbf{k}'\mathbf{X}=\hat{\mathbf{e}}_\mathbf{k}'\mathbf{S}_{\mathbf{11}}^{-\frac{\mathbf{1}}{\mathbf{2}}}\mathbf{X},\\ \hat{V}_k\triangleq\hat{\mathbf{b}}_\mathbf{k}'\mathbf{Y}=\hat{\mathbf{f}}_\mathbf{k}'\mathbf{S}_{\mathbf{22}}^{-\frac{\mathbf{1}}{\mathbf{2}}}\mathbf{Y},\end{cases}\quad k=1,2,\cdots,p. \tag{6.2.21}$$

样本典型相关系数为

$$\rho_{\hat{U}_k,\hat{V}_k}=\hat{\rho}_k,k=1,2,\cdots,p. \tag{6.2.22}$$

其中 $\hat{\rho}_1^2\geq\hat{\rho}_2^2\geq\cdots\geq\hat{\rho}_p^2$ 是 $\mathbf{S}_{\mathbf{11}}^{-\frac{\mathbf{1}}{\mathbf{2}}}\mathbf{S_{12}}\mathbf{S}_{\mathbf{22}}^{-\mathbf{1}}\mathbf{S_{21}}\mathbf{S}_{\mathbf{11}}^{-\frac{\mathbf{1}}{\mathbf{2}}}$ 的特征值, $\hat{\mathbf{e}}_\mathbf{1},\hat{\mathbf{e}}_\mathbf{2},\cdots,\hat{\mathbf{e}}_\mathbf{p}$ 为相应的单位正交化特征向量, $\hat{\mathbf{f}}_\mathbf{1},\hat{\mathbf{f}}_\mathbf{2},\cdots,\hat{\mathbf{f}}_\mathbf{p}$ 为 $\mathbf{S}_{\mathbf{22}}^{-\frac{\mathbf{1}}{\mathbf{2}}}\mathbf{S_{21}}\mathbf{S}_{\mathbf{11}}^{-\mathbf{1}}\mathbf{S_{12}}\mathbf{S}_{\mathbf{22}}^{-\frac{\mathbf{1}}{\mathbf{2}}}$ 的相应于特征值 $\hat{\rho}_1^2\geq\hat{\rho}_2^2\geq\cdots\geq\hat{\rho}_p^2$ 的单位正交化特征向量.

另外, 为消除量纲的影响, 也可以对样本观测值进行标准化, 即令

$$\begin{cases}X_{ik}^*=\dfrac{X_{ik}-\overline{X}_k}{\sqrt{s_{kk}^{(1)}}},i=1,2,\cdots,n,k=1,2,\cdots,p,\\ Y_{ik}^*=\dfrac{Y_{ik}-\overline{Y}_k}{\sqrt{s_{kk}^{(2)}}},i=1,2,\cdots,n,k=1,2,\cdots,q,\end{cases} \tag{6.2.23}$$

其中 $s_{kk}^{(1)}$ 和 $s_{kk}^{(2)}$ 分别为 $\mathbf{S}_{11}$ 和 $\mathbf{S}_{22}$ 的主对角线上的第 k 个元素, $\overline{X}_k$ 和 $\overline{Y}_k$ 分别为 $\overline{\mathbf{X}}$和$\overline{\mathbf{Y}}$ 的第 k 个分量. 标准化样本 $\begin{bmatrix}\mathbf{X}_i^*\\ \mathbf{Y}_i^*\end{bmatrix}(i=1,2,\cdots,n)$(这里$\mathbf{X}_i^*=(X_{1i}^*,X_{2i}^*,\cdots,X_{pi}^*)',\mathbf{Y}_i^*=(Y_{1i}^*,Y_{2i}^*,\cdots,Y_{qi}^*)'$) 的样本协方差矩阵即为原样本的样本相关矩阵 $\mathbf{R}$. 令

$$\mathbf{R}_{(\mathbf{p+q})\times(\mathbf{p+q})}=\begin{bmatrix}\mathbf{R}_{11(p\times p)} & \mathbf{R}_{12(p\times q)}\\ \mathbf{R}_{21(q\times p)} & \mathbf{R}_{22(q\times q)}\end{bmatrix}, \tag{6.2.24}$$

以 $\mathbf{R_{11}}, \mathbf{R_{12}}, \mathbf{R_{21}}, \mathbf{R_{22}}$ 代替前面的 $\mathbf{S_{11}}, \mathbf{S_{12}}, \mathbf{S_{21}}, \mathbf{S_{22}}$, 则得到标准化样本的典型变量和典型相关系数. 在实际分析中, 为使典型变量易于理解, 通常从 $\mathbf{R}$ 出发, 求标准化样本的典型变量, 选择样本典型相关系数较大的少数几对样本典型变量, 以反映原来两组变量间的相关关系. 那么, 样本典型相关系数多大时, 才可认为相应的一对典型变量之间存在显著相关性呢？我们可用 Bartlett 检验来讨论此问题.

6.2.3 典型相关系数的显著性检验

本段中我们假定总体 $\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix}$ 服从 p+q 维正态分布 $N_{p+q}(\mu, \mathbf{\Sigma})$, 其中 $p \leq q$ 且

$$\mu = \begin{bmatrix} \mu_{\mathbf{1}} \\ \mu_{\mathbf{2}} \end{bmatrix} = \begin{bmatrix} E(\mathbf{X}) \\ E(\mathbf{Y}) \end{bmatrix}, \mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix}.$$

如果 $\mathbf{X}$ 和 $\mathbf{Y}$ 互不相关, 则有 $\mathbf{\Sigma}_{12} = 0$, 由定理 6.2.1 可知, 典型相关系数 $\rho_k = 0 (k = 1, 2, \cdots, p)$; 反之也有 $\mathbf{\Sigma}_{12} = 0$. 因此通过检验 $\rho_1 = \rho_2 = \cdots = \rho_p = 0$, 便可判断 $\mathbf{X}$ 和 $\mathbf{Y}$ 是否显著相关, 由于 $\rho_1^2 \geq \rho_2^2 \geq \cdots \geq \rho_p^2 \geq 0$, 这等价于检验 ρ_1 是否为零. 因此, 我们检验假设

$$H_0^{(1)} : \rho_1 = 0, H_1^{(1)} : \rho_1 \neq 0, \tag{6.2.25}$$

当接受 $H_0^{(1)}$ 时, 即认为 $\mathbf{X}$ 和 $\mathbf{Y}$ 不相关. 这时相关分析便无意义. 当拒绝 $H_0^{(1)}$ 时, 可进一步检验假设

$$H_0^{(2)} : \rho_1 \neq 0, \rho_2 = 0, H_1^{(2)} : \rho_2 \neq 0, \tag{6.2.26}$$

若接受 $H_0^{(2)}$, 则认为除第一对典型变量显著相关以外, 其余各对典型变量的相关性不显著. 因而在实际应用中, 可只考虑第一对典型相关变量, 以其反映 $\mathbf{X}$ 和 $\mathbf{Y}$ 的相关性. 若拒绝 $H_0^{(2)}$, 则需进一步检验 ρ_3 是否为零. 依次类推, 若假设 $H_0^{(k-1)} : \rho_1 \neq 0, \cdots, \rho_{k-2} \neq 0, \rho_{k-1} = 0$ 被拒绝, 则进一步检验

$$H_0^{(k)} : \rho_1 \neq 0, \cdots, \rho_{k-1} \neq 0, \rho_k = 0, H_1^{(k)} : \rho_k \neq 0, \tag{6.2.27}$$

若接受 $H_0^{(k)}$, 则只需要用前 k-1 对典型变量反映 $\mathbf{X}, \mathbf{Y}$ 的相关关系. 若拒绝 $H_0^{(k)}$, 则继续检验 ρ_{k+1} 是否为零, 等等.

上述假设的 Bartlett 检验方法如下, 在 $\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} \sim N_{p+q}(\mu, \mathbf{\Sigma})$ 条件下, 为检验 $H_0^{(1)}$, 令

$$W_1 = \prod_{i=1}^{p} (1 - \hat{\rho}_i^2), \tag{6.2.28}$$

当 H_0 为真且 n 充分大时, 统计量

$$A_1 = -[n - 1 - \frac{1}{2}(p + q + 1)] \ln W_1 \tag{6.2.29}$$

渐近服从自由度为 pq 的 χ^2 分布. 给定显著水平 j, 以 $\chi_j^2(pq)$ 记自由度为 pq 的 χ^2 分布的上侧 j 分位数, 则当 $A_1 \geq \chi_j^2(pq)$ 时, 拒绝 $H_0^{(1)}$. 否则接受 $H_0^{(1)}$, 检验结束, 即认为 $\mathbf{X}$ 和 $\mathbf{Y}$ 不相关.

当拒绝 $H_0^{(1)}$ 后, 为检验 $H_0^{(2)}$, 令

$$W_2 = \prod_{i=2}^{p}(1-\hat{\rho}_i^2), \tag{6.2.30}$$

$$A_2 = -[n-2-\frac{1}{2}(p+q+1)]\ln W_2. \tag{6.2.31}$$

当 $H_0^{(2)}$ 为真时, A_2 渐近服从自由度为 (p-1)(q-1) 的 χ^2 分布, 若 $A_2 \geq \chi_j^2[(p-1)(q-1)]$, 拒绝 $H_0^{(2)}$. 否则接受 $H_0^{(2)}$, 即认为只有第一对典型变量显著相关.

一般地, 若第 k-1 步检验拒绝 $H_0^{(k-1)}$, 则需检验 $H_0^{(k)}$, 令

$$W_k = \prod_{i=k}^{p}(1-\hat{\rho}_i^2), \tag{6.2.32}$$

$$A_k = -[n-k-\frac{1}{2}(p+q+1)]\ln W_k. \tag{6.2.33}$$

当 $H_0^{(k)}$ 为真时, A_k 渐近服从自由度为 (p-k+1)(q-k+1) 的 χ^2 分布, 当 $A_k \geq \chi_j^2[(p-k+1)(q-k+1)]$, 拒绝 $H_0^{(k)}$. 否则接受 $H_0^{(k)}$, 检验结束, 即认为只有前 k-1 对典型变量显著相关. 这时只需考虑前 k-1 对典型变量, 用以研究 $\mathbf{X}$ 和 $\mathbf{Y}$ 之间的相关关系. 若拒绝 $H_0^{(k)}$, 则继续检验假设 $H_0^{(k+1)}$.

对标准化样本, 检验方法完全相同, 只是将其中的 $\hat{\rho}_i^2$ 代之以矩阵 $\mathbf{R_{11}^{-\frac{1}{2}}R_{12}R_{22}^{-1}R_{21}R_{11}^{-\frac{1}{2}}}$ 的特征值即可.

典型相关系数的大样本检验为实际应用中典型变量对的取舍提供了一个参考准则. 下面举例说明典型相关分析方法的应用.

例 6.2.2 课程设置的典型相关分析.

数学专业通常开设下列课程: X_1 (数学分析), X_2 (高等代数), X_3 (解析几何), X_4 (普通物理); Y_1 (常微分方程), Y_2 (抽象代数), Y_3 (复变函数), Y_4 (概率统计), Y_5 (实变函数), Y_6 (偏微分方程), Y_7 (泛函分析), 等等. 前四门课程称为基础课, 后七门称为专业课. 这两组课程之间有一定的相关关系, 下面以学生的各门课的学业成绩为各变量的取值进行相关分析. 令

$$\mathbf{X} = (X_1, X_2, X_3, X_4)'$$

$$\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4, Y_5, Y_6, Y_7)'$$

抽取了某校 n=75 名学生的学业成绩, 计算其样本相关矩阵 (即标准化样本数据的样本协方差矩阵) 为

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{11(4\times4)} & \mathbf{R}_{12(4\times7)} \\ \mathbf{R}_{21(7\times4)} & \mathbf{R}_{22(7\times7)} \end{bmatrix} = \text{缺数据}$$

其中 $\mathbf{R_{12}} = \mathbf{R_{21}'}$, 未写出部分与相应左下角部分对称.

求得矩阵 $\mathbf{R_{11}^{-\frac{1}{2}}R_{12}R_{22}^{-1}R_{21}R_{11}^{-\frac{1}{2}}}$ 的四个特征值为

$$\hat{\rho}_1^2 = 0.953, \hat{\rho}_2^2 = 0.151, \hat{\rho}_3^2 = 0.114, \hat{\rho}_4^2 = 0.042,$$

求得相应于各 $\hat{\rho}_i^2 (i = 1, 2, 3, 4)$ 的典型变量的线性组合系数如下：

对

$$\begin{aligned}
&\hat{\rho}_1 = 0.976,\\
&\hat{\mathbf{a}}_1 = (0.743, 0.251, 0.336, 0.521)',\\
&\hat{\mathbf{b}}_1 = (0.131, 0.877, 0.378, 0.010, 0.216, -0.055, -0.185)'.
\end{aligned}$$

对

$$\begin{aligned}
&\hat{\rho}_2 = 0.389,\\
&\hat{\mathbf{a}}_2 = (-0.198, 0.612, -0.664, -0.143)',\\
&\hat{\mathbf{b}}_2 = (0.504, 0.593, -0.509, 0.534, 0.019, 0.306, -0.279)'.
\end{aligned}$$

对

$$\begin{aligned}
&\hat{\rho}_3 = 0.338,\\
&\hat{\mathbf{a}}_3 = (-0.666, 0.002, 0.987, -0.010)',\\
&\hat{\mathbf{b}}_3 = (-0.399, 0.843, 0.302, -0.361, 0.675, -0.385, -0.056)'.
\end{aligned}$$

对

$$\begin{aligned}
&\hat{\rho}_4 = 0.205,\\
&\hat{\mathbf{a}}_4 = (-0.605, 0.048, -0.481, 0.107)',\\
&\hat{\mathbf{b}}_4 = (0.234, 0.985, -0.036, -0.639, -0.263, -0.149, 0.190)'.
\end{aligned}$$

下面利用 Bartlett 检验判断各对典型变量相关的显著性. 取 $j = 0.10$, 将检验统计量的观测值及 χ^2 分布的上侧 j 分位数（临界值） $\chi_j^2(f)$ 及检验的结果列于下表：

i	$\hat{\rho}_i$	W_i	A_i	自由度f	临界值$\chi_j^2(f)$	结论
1	0.976	0.034	229.93	28	37.92	拒绝$H_0^{(1)}$
2	0.389	0.720	22.34	18	25.99	接受$H_0^{(2)}$
3	0.338	0.849	11.13	10	15.99	
4	0.205	0.958	2.92	4	7.78	

由于 $H_0^{(2)}$ 被接受，即除第一对典型变量显著相关外，其余各对的相关性均不显著. 因此只需考虑第一对样本典型变量：

$$\begin{aligned}
&\hat{U}_1 = 0.743X_1^* + 0.251X_2^* + 0.336X_3^* + 0.521X_4^*,\\
&\hat{V}_1 = 0.131Y_1^* + 0.877Y_2^* + 0.378Y_3^* + 0.010Y_4^* + 0.216Y_5^* - 0.055Y_6^* - 0.185Y_7^*,
\end{aligned}$$

其中 $X_i^* (i = 1, 2, 3, 4)$ 和 $Y_i^* (j = 1, 2, \cdots, 7)$ 为相应的样本标准化变量.

下面我们对 $(\hat{U}_1, \hat{V}_1)$ 的实际意义给予解释. 由于 $\rho_{\hat{U}_1,\hat{V}_1} = 0.976$, 即 $\hat{U}_1$ 与 $\hat{V}_1$ 具有高度正相关关系, 而 $\hat{U}_1$ 和 $\hat{V}_1$ 中各变量的权系数大部分为正 (尤其是绝对值较大的权系数), 因此一般说来, 学好基础课对学好专业课具有促进作用, 即基础课成绩较好者, 他们的专业课成绩一般也较好. 进一步, $\hat{U}_1$ 中 X_1^* (数学分析) 的权系数最大其次是 X_4^* (普通物理) 的权系数大 (这主要是因为普通物理是以数学分析为基础的物理课程, 这两门课程本身具有较高的相关性. 这也可由 $\mathbf{R}_{11}$ 中反映 X_1^* 与 X_4^* 的相关系数 (0.543) 较大看出), 因此, 基础课对专业课的促进作用主要体现在数学分析课程上. 而 $\hat{V}_1$ 中权系数最大的是 Y_2^* (抽象代数), 其次是 Y_3^* (复变函数) 和 Y_5^* (实变函数), 于是 $(\hat{U}_1, \hat{V}_1)$ 主要反映了 "数学分析——函数论" 这条主线, 它代表了数学专业所设置课程的主要特征.

第七章　判别分析

本章在阐述判别分析的基本思想和意义的基础上, 介绍两种实用的判别分析方法——距离判别和 Bayes 判别.

§7.1　判别分析的基本思想及意义

在科学研究中, 经嵽龅秸庋奈侍某研究对象以某种方式 (如先前的结果或经验) 已划分成若干类型, 而每一类型都是用一些指标 $\mathbf{X} = (X_1, X_2, \cdots, X_p)'$ 来表征的, 即不同类型的 $\mathbf{X}$ 的观测值在某种意义上有一定的差异. 当得到一个新样本观测值 (或个体) 的的关于指标 $\mathbf{X}$ 的观测值时, 要判断该样本观测值 (或个体) 属于这几个已知类型中的哪一个, 这类问题通常称为判别分析. 也就是说, 判别分析是根据所研究个体的某些指标的观测值来推断该个体所属类型的一种统计方法.

判别分析的应用十分广泛. 例如, 在工业生产中, 要根据某种产品的一些非破坏性测量指标判别产品的质量等级; 在经济分析中, 根据人均国民收入, 人均工农业产值, 人均消费水平等指标判断一个国家的经济发展程度; 在考古研究中, 根据挖掘的古人头盖骨的容量, 周长等判断此人的性别; 在地质勘探中, 根据某地的地质结构, 化探和物探等各项指标来判断该地的矿化类型; 在医学诊断中, 医生要根据某病人的化验结果和病情征兆判断病人患哪一种疾病, 等等. 值得注意的是, 作为一种统计方法, 判别分析所处理的问题一般都是机理不甚清楚或者基本不了解的复杂问题, 如果样本观测值的某些观测指标和其所属类型有必然的逻辑关系, 也就没有必要应用判别分析方法了.

用统计的语言来描述判别分析, 就是已知有 g 个总体 $G_1, G_2, \cdots, G_g$ (每个总体 G_i 可认为是属于 G_i 的指标 $\mathbf{X} = (X_1, X_2, \cdots, X_p)'$ 取值的全体), 它们的分布函数 $F_1(\mathbf{x}), F_2(\mathbf{x}), \cdots, F_g(\mathbf{x})$ 均为 p 维函数, 对于任一给定的新样本关于指标 $\mathbf{X}$ 的观测值 $\mathbf{X} = (x_1, x_2, \cdots, x_p)'$, 我们要判断该样本观测值应属于这 g 个总体中的哪一个.

在实际应用中, 通常由取自各总体的关于指标 $\mathbf{X}$ 的样本为该总体的代表, 该样本称为训练样本. 判别分析即提取训练样本中各总体的信息以构造

一定的准则来决定新样本观测值的归属问题. 训练样本往往是历史上对某现象长期观测或者是用昂贵的试验手段得到的, 因此对当前的新样本观测值, 我们自然希望将其指标值中的信息同各总体训练样本中的信息作比较, 使可在一定程度上判定新样本观测值的所属类型. 概括起来, 下述几个方面体现了判别分析的重要意义.

第一, 为未来的决策和行动提供参考. 例如, 以前对一些公司在破产前两年观测到某些重要的金融指标值. 现在, 要根据另一个同类型公司的这些指标的观测值, 预测该公司两年后是否将濒临破产的危险, 这便是一种判别, 其结论可以帮助该公司决策人员及早采取措施, 防止将来可能破产的结局.

第二, 避免产品的破坏. 例如, 一只灯泡的寿命只有将它用坏时才能得知; 一种材料的强度只有将它压坏时才能获得. 一般地, 我们希望根据一些非破坏性的测量指标, 便可将产品分出质量等级, 这也要用到判别分析.

第三, 减少获得直接分类信息的昂贵代价. 例如在医学诊断中, 一些疾病可用代价昂贵的化验和手术得到确诊, 但通常人们往往更希望通过便于观测 (从而也可能导致误诊) 的一些外部症状来诊断, 以避免过大的开支和对患者不必要的损伤.

第四, 在直接分类信息不能获得的情况下可用判别分析. 例如, 要判断某未署名的文学作品是否出自某已故作家之手, 很显然, 我们不能直接去问他. 这时可以用这位已故作家署名作品的写作特点 (用一些变量描述) 为训练样本, 用判别分析方法在一定程度上判定该未署名作品是否由该作家所作.

从以上例子也可以清楚地看到, 如果不是利用直接明确的分类信息来判断某样本观测值的归属问题, 难免会出现误判的情况. 判别分析的任务是依据训练样本所提供的信息, 建立在某种意义下最优 (如误判概率最小, 或误判损失最小等) 的准则来判定一个新样本属于哪一个总体. 根据判别准则的不同, 我们主要介绍距离判别和 Bayes 判别.

§7.2 距离判别

距离判别是通过定义样本指标 $\mathbf{X}$ 的观测值 $\mathbf{x}$ (p 维) 到各总体的距离, 以其大小判定样本观测值属于哪个总体. 常用的距离是 Mahalanobis 距离 (简称马氏距离), 其定义如下:

设 G 是 p 维总体, 数学期望为 μ, 协方差矩阵为 $\mathbf{\Sigma}$, 定义 p 维样本 $\mathbf{x}$ 到总体 G 的马氏距离为

$$d(\mathbf{x}, G) \hat{=} [(\mathbf{x}-\mu)' \mathbf{\Sigma^{-1}} (\mathbf{x}-\mu)]^{\frac{1}{2}}. \tag{7.2.1}$$

7.2.1 两总体的距离判别

设 G_1, G_2 为两个不同的 p 维总体, 数学期望分别为 μ_1和μ_2, 协方差矩阵分别为 $\mathbf{\Sigma_1}$和$\mathbf{\Sigma_2}$. 设 $\mathbf{x}$ 为一个待判样本观测值 (为方便计, 以后我们将不区分样本观测值和它的指标观测值 $\mathbf{x}$), 分别计算 $\mathbf{x}$ 到 G_1和G_2 的马氏距离

$d(\mathbf{x}, G_1)$和$d(\mathbf{x}, G_2)$, 哪个距离小，就判定 $\mathbf{x}$ 属于哪个总体. 即判别准则如下：

$$\begin{cases} \mathbf{x} \in G_1, 若 d(\mathbf{x}, G_1) < d(\mathbf{x}, G_2), \\ \mathbf{x} \in G_2, 若 d(\mathbf{x}, G_2) < d(\mathbf{x}, G_1). \end{cases} \tag{7.2.2}$$

下面分别就两总体的协方差矩阵相等和不等两种情况进一步讨论该判别准则.

1. 设 $\boldsymbol{\Sigma_1} = \boldsymbol{\Sigma_2} = \boldsymbol{\Sigma}$

此时，考察样本观测值 $\mathbf{x}$ 到两总体的马氏距离的平方差，由于

$$\begin{aligned} & d^2(\mathbf{x}, G_2) - d^2(\mathbf{x}, G_1) \\ & = (\mathbf{x} - \mu_{\mathbf{2}})' \boldsymbol{\Sigma^{-1}} (\mathbf{x} - \mu_{\mathbf{2}}) - (\mathbf{x} - \mu_{\mathbf{1}})' \boldsymbol{\Sigma^{-1}} (\mathbf{x} - \mu_{\mathbf{1}}) \\ & = \mathbf{x}' \boldsymbol{\Sigma^{-1}} \mathbf{x} - \mathbf{2x}' \boldsymbol{\Sigma^{-1}} \mu_{\mathbf{2}} + \mu_{\mathbf{2}}' \boldsymbol{\Sigma^{-1}} \mu_{\mathbf{2}} - \mathbf{x}' \boldsymbol{\Sigma^{-1}} \mathbf{x} + \mathbf{2x}' \boldsymbol{\Sigma^{-1}} \mu_{\mathbf{1}} - \mu_{\mathbf{1}}' \boldsymbol{\Sigma^{-1}} \mu_{\mathbf{1}} \\ & = 2\mathbf{x}' \boldsymbol{\Sigma^{-1}} (\mu_{\mathbf{1}} - \mu_{\mathbf{2}}) + \mu_{\mathbf{2}}' \boldsymbol{\Sigma^{-1}} \mu_{\mathbf{2}} - \mu_{\mathbf{1}}' \boldsymbol{\Sigma^{-1}} \mu_{\mathbf{1}} + \mu_{\mathbf{1}}' \boldsymbol{\Sigma^{-1}} \mu_{\mathbf{2}} - \mu_{\mathbf{2}}' \boldsymbol{\Sigma^{-1}} \mu_{\mathbf{1}} \\ & = 2\mathbf{x}' \boldsymbol{\Sigma^{-1}} (\mu_{\mathbf{1}} - \mu_{\mathbf{2}}) - (\mu_{\mathbf{1}} + \mu_{\mathbf{2}})' \boldsymbol{\Sigma^{-1}} (\mu_{\mathbf{1}} - \mu_{\mathbf{2}}) \\ & = 2[\mathbf{x} - \tfrac{\mathbf{1}}{\mathbf{2}} (\mu_{\mathbf{1}} + \mu_{\mathbf{2}})]' \boldsymbol{\Sigma^{-1}} (\mu_{\mathbf{1}} - \mu_{\mathbf{2}}) \\ & = 2(\mathbf{x} - \bar{\mu})' \boldsymbol{\Sigma^{-1}} (\mu_{\mathbf{1}} - \mu_{\mathbf{2}}). \end{aligned} \tag{7.2.3}$$

其中 $\bar{\mu} = \frac{1}{2}(\mu_{\mathbf{1}} + \mu_{\mathbf{2}})$. 令

$$W(\mathbf{x}) = (\mathbf{x} - \bar{\mu})' \boldsymbol{\Sigma^{-1}} (\mu_{\mathbf{1}} - \mu_{\mathbf{2}}), \tag{7.2.4}$$

则判别准则（ 7.2.2 ）此时可简化为

$$\begin{cases} \mathbf{x} \in G_1, 若 W(\mathbf{x}) \geq 0, \\ \mathbf{x} \in G_2, 若 W(\mathbf{x}) < 0. \end{cases} \tag{7.2.5}$$

进一步，令 $\mathbf{a}' = (\mu_{\mathbf{1}} - \mu_{\mathbf{2}})' \boldsymbol{\Sigma^{-1}}$, 则（ 7.2.4 ）中的 $W(\mathbf{x})$ 可表为

$$W(\mathbf{x}) = \mathbf{a}'(\mathbf{x} - \bar{\mu}). \tag{7.2.6}$$

上式表明，当 $\mu_{\mathbf{1}}, \mu_{\mathbf{2}}$ 及 $\boldsymbol{\Sigma}$ 均已知时，用以判别的函数 $W(\mathbf{x})$ 此时为 $\mathbf{x}$ 的线性函数，即判别函数是线性的. 线性判别函数因其使用方便而得到广泛的应用.

但在实际问题中， $\boldsymbol{\Sigma}$ 及 $\mu_{\mathbf{1}}, \mu_{\mathbf{2}}$ 通常是未知的，我们所具有的资料只是来自两个总体的训练样本. 这时，可通过训练样本对 $\boldsymbol{\Sigma}$ 及 $\mu_{\mathbf{1}}, \mu_{\mathbf{2}}$ 作估计. 设 $\mathbf{x}_{\mathbf{1}}^{(\mathbf{1})}, \mathbf{x}_{\mathbf{2}}^{(\mathbf{1})}, \cdots, \mathbf{x}_{\mathbf{n_1}}^{(\mathbf{1})}$ 为来自 G_1 的容量为 n_1 的训练样本， $\mathbf{x}_{\mathbf{1}}^{(\mathbf{2})}, \mathbf{x}_{\mathbf{2}}^{(\mathbf{2})}, \cdots, \mathbf{x}_{\mathbf{n_2}}^{(\mathbf{2})}$ 来

自 G_2 的容量为 n_2 的训练样本 (每个 $\mathbf{x}_i^{(k)}$(k=1,2) 均为 p 维列向量), 其各自的样本均值向量可作为 μ_1和μ_2 的估计，即

$$\hat{\mu}_k = \frac{1}{n_k}\sum_{i=1}^{n_k}\mathbf{x}_i^{(k)} = \bar{\mathbf{x}}^{(k)}, k = 1, 2, \tag{7.2.7}$$

联合各样本协方差矩阵

$$\mathbf{S}_k = \frac{1}{n_k - 1}\sum_{i=1}^{n_k}(\mathbf{x}_i^{(k)} - \bar{\mathbf{x}}^{(k)})(\mathbf{x}_i^{(k)} - \bar{\mathbf{x}}^{(k)})', k = 1, 2, \tag{7.2.8}$$

可得到 $\mathbf{\Sigma}$ 的一个无偏估计为

$$\hat{\mathbf{\Sigma}} \hat{=} \frac{(n_1 - 1)\mathbf{S_1} + (n_2 - 1)\mathbf{S_2}}{n_1 + n_2 - 2}. \tag{7.2.9}$$

这时，判别函数 $W(\mathbf{x})$ 的估计为

$$\hat{W}(\mathbf{x}) = (\mathbf{x} - \hat{\bar{\mu}})'\mathbf{\hat{\Sigma}^{-1}}(\hat{\mu_1} - \hat{\mu_2}), \tag{7.2.10}$$

其中 $\hat{\bar{\mu}} = \mathbf{\frac{1}{2}}(\hat{\mu_1} + \hat{\mu_2})$. 对于给定的一个新样本 $\mathbf{x}$, 视 $\hat{W}(\mathbf{x}) \geq 0$和$\hat{W}(\mathbf{x}) < \mathbf{0}$ 判定 $\mathbf{x}$ 属于 G_1 或 G_2.

2. 若 $\mathbf{\Sigma_1} \neq \mathbf{\Sigma_2}$

正如本节开始所述，可由 $d^2(\mathbf{x}, G_1)$和$d^2(\mathbf{x}, G_2)$ 的大小判定 $\mathbf{x}$ 属于哪个总体，或令

$$\begin{aligned} W(\mathbf{x}) &= d^2(\mathbf{x}, G_2) - d^2(\mathbf{x}, G_1) \\ &= (\mathbf{x} - \mu_2)'\mathbf{\Sigma_1^{-1}}(\mathbf{x} - \mu_2) - (\mathbf{x} - \mu_1)'\mathbf{\Sigma_2^{-1}}(\mathbf{x} - \mu_1), \end{aligned} \tag{7.2.11}$$

则

$$\begin{cases} \mathbf{x} \in G_1, 若W(\mathbf{x}) \geq 0, \\ \mathbf{x} \in G_2, 若W(\mathbf{x}) < 0. \end{cases} \tag{7.2.12}$$

这时，判别函数 $W(\mathbf{x})$ 为 $\mathbf{x}$ 的二次函数.

实际应用中，若 $\mu_1, \mu_2, \mathbf{\Sigma_1}, \mathbf{\Sigma_2}$ 未知，可用各总体的训练样本对它们作估计，从而得到判别函数 $W(\mathbf{x})$ 的估计为

$$\hat{W}(\mathbf{x}) = (\mathbf{x} - \hat{\mu}_2)'\mathbf{S_2^{-1}}(\mathbf{x} - \hat{\mu}_2) - (\mathbf{x} - \hat{\mu}_1)'\mathbf{S_1^{-1}}(\mathbf{x} - \hat{\mu}_1), \tag{7.2.13}$$

其中 $\hat{\mu}_1, \hat{\mu}_2$ 和 $\mathbf{S}_1, \mathbf{S}_2$ 如（ 7.2.7 ）和（ 7.2.8 ）所示.

7.2.2 多总体的距离判别

设有 g 个 p 维总体 $G_1, G_2, \cdots, G_g$, 均值向量分别为 $\mu_1, \mu_2, \cdots, \mu_g$, 协方差矩阵分别为 $\mathbf{\Sigma_1}, \mathbf{\Sigma_2}, \cdots, \mathbf{\Sigma_g}$. 类似两总体距离判别方法，计算新样本观测值 $\mathbf{x}$ 到各总体的距离，比较这 g 个距离，判定 $\mathbf{x}$ 属于其距离最短的总体 (若

最短距离不惟一，则可将 $\mathbf{x}$ 归于具有最短距离总体中的任一个，因此，不妨设最短距离惟一). 下面仍就各协方差矩阵相等和不等的情况予以详细讨论.

1. 若 $\boldsymbol{\Sigma_1} = \boldsymbol{\Sigma_2} = \cdots = \boldsymbol{\Sigma_g} = \boldsymbol{\Sigma}$

此时，由（ 7.2.3 ）式可知 x 到 G_j 和 G_i 的马氏距离的平方差为

$$d^2(\mathbf{x}, G_j) - d^2(\mathbf{x}, G_i) = 2[\mathbf{x} - \frac{1}{2}(\mu_\mathbf{i} + \mu_\mathbf{j})]'\boldsymbol{\Sigma}^{-1}(\mu_\mathbf{i} - \mu_\mathbf{j}). \tag{7.2.14}$$

令

$$W_{ij}(\mathbf{x}) = [\mathbf{x} - \frac{\mathbf{1}}{\mathbf{2}}(\mu_\mathbf{i} + \mu_\mathbf{j})]'\boldsymbol{\Sigma^{-1}}(\mu_\mathbf{i} - \mu_\mathbf{j}). \tag{7.2.15}$$

则 $\mathbf{x}$ 到 G_i 的距离最小等价于对所有的 $j \neq i$, 有 $W_{ij}(\mathbf{x}) > 0$, 从而判别准则为

$$\mathbf{x} \in G_i, 若对一切 j \neq i, W_{ij}(\mathbf{x}) > 0. \tag{7.2.16}$$

当 $\mu_\mathbf{1}, \mu_\mathbf{2}, \cdots, \mu_\mathbf{g}$ 和 $\boldsymbol{\Sigma}$ 未知时，可利用各总体的训练样本对其作估计. 设 $\mathbf{x_1^{(k)}}, \mathbf{x_2^{(k)}}, \cdots, \mathbf{x_{n_k}^{(k)}}$ 为来自总体 G_k 的训练样本 $(k = 1, 2, \cdots, g)$, 令

$$\hat{\mu}_k = \frac{1}{n_k}\sum_{i=1}^{n_k}\mathbf{x}_i^{(k)} = \bar{\mathbf{x}}^{(k)}, k = 1, 2, \cdots, g, \tag{7.2.17}$$

$$\mathbf{S}_k = \frac{1}{n_k - 1}\sum_{i=1}^{n}(\mathbf{x}_i^{(k)} - \bar{\mathbf{x}}^{(k)})(\mathbf{x}_i^{(k)} - \bar{\mathbf{x}}^{(k)})', k = 1, 2, \cdots, g, \tag{7.2.18}$$

利用 $\mathbf{S}_k (k = 1, 2, \cdots, g)$ 对 $\boldsymbol{\Sigma}$ 的联合估计为

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n - g}[(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2 + \cdots + (n_g - 1)\mathbf{S}_g], \tag{7.2.19}$$

其中 $n = \sum_{i=1}^{g} n_i$.

以 $\hat{\mu}_k (k = 1, 2, \cdots, g)$ 和 $\hat{\boldsymbol{\Sigma}}$ 代替（ 7.2.15 ）式中的 $\mu_k (k = 1, 2, \cdots, g)$ 及 $\boldsymbol{\Sigma}$, 便可得判别函数 $W_{ij}(\mathbf{x})$ 的估计 $\hat{W}_{ij}(\mathbf{x})$, 以 $\hat{W}_{ij}(\mathbf{x})$ 代替（ 7.2.16 ）中的 $W_{ij}(\mathbf{x})$ 进行判别.

2. 若 $\boldsymbol{\Sigma}_i (i = 1, 2, \cdots, g)$ 不全相同

这时只需直接计算

$$d^2(\mathbf{x}, G_i) = (\mathbf{x} - \mu_\mathbf{i})'\boldsymbol{\Sigma_i^{-1}}(\mathbf{x} - \mu_\mathbf{i}), i = 1, 2, \cdots, g, \tag{7.2.20}$$

若

$$\min_{1 \leq k \leq g}\{d^2(\mathbf{x}, G_k)\} = d^2(\mathbf{x}, G_i), 则判\mathbf{x} \in G_i. \tag{7.2.21}$$

同样，若 $\mu_\mathbf{i}, \boldsymbol{\Sigma_i} (i = 1, 2, \cdots, g)$ 未知，可用它们的估计量 $\hat{\mu}_i$ 和 $\mathbf{S_i}$（分别见（ 7.2.17 ）和（ 7.2.18 ））代入（ 7.2.20 ）计算 $\mathbf{x}$ 到各总体的距离.

7.2.3 判别准则的评价

当一个判别准则提出以后, 很自然的问题就是它们的优良性如何. 通常，一个判别准则的优劣，用它的误判概率来衡量. 以两总体为例，一个判别准

则的误判概率即 $\mathbf{x}$ 属于 G_1 而判归 G_2 或者相反的概率. 但只有当总体的分布完全已知时, 才有可能精确计算误判概率. 在实际应用中, 这种情况是很少见的, 因为在大多数情况下, 我们可利用的资料只是来自各总体的训练样本, 而总体的分布是未知的. 下面我们以两个总体为例, 介绍两种以训练样本为基础的评价判别准则优劣的方法. 它们也很容易推广到多个总体的情况.

1. 貌似误判率方法

当利用各总体的训练样本构造出判别准则后, 评价此准则优劣的一个可行的办法是通过对训练样本中的各样本逐个回判 (即将各样本观测值代入判别准则中进行再判别), 利用回判的误判率来衡量判别准则的效果, 具体方法如下:

设 G_1, G_2 为两个总体, $\mathbf{x_1^{(k)}}, \mathbf{x_2^{(k)}}, \cdots, \mathbf{x_{n_k}^{(k)}} (k = 1, 2)$ 为来自 G_1和G_2 的容量分别为 n_1 和 n_2 的训练样本, 以此按一定方法 (如距离判别法) 构造一个判别准则 (或判别函数), 以全体训练样本作为 $n_1 + n_2$ 个新样本, 逐个代入已建立的判别准则中判别其归属, 这个过程称为回判. 为明了起见, 将回判结果连同其实际分类列成如下的四格表 7.2.1.

表 7.2.1　两总体回判结果

回判情况 实际归类	G_1	G_2	合计
G_1	n_{11}	n_{12}	n_1
G_2	n_{21}	n_{22}	n_2

其中

n_{11} : 属于G_1的样本观测值被正确判归G_1的个数,
n_{12} : 属于G_1的样本观测值被错误判归G_2的个数,
n_{21} : 属于G_2的样本观测值被错误判归G_1的个数,
n_{22} : 属于G_2的样本观测值被正确判归G_2的个数.

很显然有

$$n_{11} + n_{12} = n_1, n_{21} + n_{22} = n_2.$$

定义貌似误判率为回归中判错样本观测值的比例, 记为 $\hat{\alpha}$, 即

$$\hat{\alpha} = \frac{n_{12} + n_{21}}{n_1 + n_2}. \tag{7.2.22}$$

$\hat{\alpha}$ 在一定程度上反映了某判别准则的误判率且对任何误判准则都易于计算. 但是, $\hat{\alpha}$ 是由建立判别函数的数据反过来又用作评估准则优劣的数据而得到的, 因此 $\hat{\alpha}$ 作为真实误判率的估计是有偏的, 往往要比真实的误判率来的小. 但作为误判概率的一种近似, 当训练样本容量较大时, 还是具有一定的参考价值.

2. 刀切法

刀切法也称为 Lachenbruch 删除法或交差确认法 (Cross-Validation). 其基本思想是每次剔除训练样本中的一个样本, 利用其余容量为 $n_1 + n_2 - 1$ 的训练样本建立判别准则 (或判别函数), 再用所建立的判别准则对删除的那个样本观测值作判断, 对训练样本中的每个样本观测值重复上述步骤, 以其误判的比例作为误判概率的估计. 具体地说:

（ 1 ）从总体 G_1 的容量 n_1 的训练样本开始，剔除其中的一个样本观测值，用剩余的容量为 n_1-1 的训练样本和总体 G_2 的容量为 n_2 的训练样本建立判别函数；

（ 2 ）用（ 1 ）中建立的判别函数对删除的那个样本观测值作判别；

（ 3 ）重复步骤（ 1 ）和（ 2 ），直到 G_1 的训练样本中的 n_1 个样本观测值依次被删除和判别，用 $n_{1M}^{(J)}$ 记误判的样本观测值个数；

（ 4 ）对总体 G_2 的训练样本重复步骤（ 1 ），（ 2 ）和（ 3 ），并用 $n_{2M}^{(J)}$ 记误判的样本观测值个数．则总的误判比例为

$$\hat{\alpha}_J \widehat{=} \frac{n_{1M}^{(J)}+n_{2M}^{(J)}}{n_1+n_2}. \tag{7.2.23}$$

可以证明它是实际误判概率的渐近无偏估计．

刀切法比貌似误判率法要更合理些，但缺点是计算量较大．在 SAS 等统计软件中有专门的计算程序，因此借助计算机的威力，刀切法还是值得推荐的一种评价判别准则优良性的方法．

最后需要指出的是，判别准则的误判率在一定的程度上还依赖于所考虑的各总体之间的分离程度．各总体之间相互离得越远，就越有可能建立有效的判别准则，否则，某些总体靠得很近，使用判别分析本身就意义不大，更不用说建立有效的判别准则了．另外，各总体的协方差矩阵是否相等，严格地说也需要进行统计检验．当各总体服从多元正态分布时，我们可以对各总体的均值向量是否相等进行统计检验以确定使用判别分析是否有意义．同时，也可对各总体的协方差阵是否相等进行检验以确定是采用线性判别函数还是二次判别函数（具体检验方法可参看文献 [8] 第五章）．但这些检验方法往往十分复杂，在实际应用中，我们可就协方差矩阵相等和不相等情况下，分别利用线性判别函数和二次判别函数作分析，通过貌似误判率方法或刀切法估计各情况下判别准则的优劣，以选择一个较优的判别准则． SAS 软件包含了检验协方差阵相等的程序，因此借助统计软件，可以进行更深入的统计分析．

§7.3 Bayes 判别

Bayes 统计的基本思想：假定对所研究的对象（总体）在抽样前已有一定的认识，常用先验概率分布来描述这种认识．然后基于抽取的样本再对先验认识作修正，得到所谓后验概率分布，而各种统计推断都基于后验概率分布来进行．将 Bayes 统计的思想用于判别分析，就得到 Bayes 判别方法．

7.3.1 **Bayes** 判别的基本思想

设 $G_1, G_2, \cdots, G_g$ 为 g 个 p 维总体，分别具有互不相同的 p 维概率密度函数 $f_1(\mathbf{x}), f_2(\mathbf{x}), \cdots, f_g(\mathbf{x})$．在进行判别分析之前，我们往往已对各总体有一定了解，实际中通常表现在某些总体较之其它总体出现的可能性会相对大一些．例如，对某厂生产的产品，正品总比次品多，即出现的样本观测值属于正品总体的可能性要比属于次品总体的可能性要相对大一些．又如，在全年

365 天中，发生大地震的可能性要比无大地震或无地震的可能性要小得多. 因此，一个合理的判别准则应该考虑到每个总体出现的可能性的大　聪妊楦怕史植迹一般来说，将一个随机样本观测值应该首先考虑判入有较大可能出现的总体中. 设这 g 个总体出现的先验概率分布为 $q_1, q_2, \cdots, q_g$, 显然应有

$$q_i \geq 0 (i = 1, 2, \cdots, g) 且 \sum_{i=1}^{g} q_i = 1. \tag{7.3.1}$$

除考虑各总体出现的先验概率外，还应考虑误判所造成的损失问题. 在大多数实际问题中，若将属于总体 G_1 的样本观测值判归为 G_2, 则会造成一定的损失，反之亦然，但造成损失的程度可能有所不同. 例如，将一个正品电子元件判为次品，所损失的只是生产厂家 (如果这种元件的成本不是很昂贵的话), 但若判为正品而使用在更大的系统中, 则有可能造成整个系统的损坏 (这种损失往往是很大的). 又如，将实际生病的人判为无病，有可能导致病情加重甚至死亡而造成损失. 反之将无病者诊断为有病，可给他们造成不必要的医疗费用支出和精神负担. 总之，在制定判别准则时，应考虑到误判的损失问题. 而这通常在判别分析前就是可以估计的，我们用表 7.3.1 的损失矩阵描述.

表 7.3.1 损失矩阵

判定为 实际为	G_1	G_2	$\cdots$	G_g
G_1	0	$c(2\|1)$	$\cdots$	$c(g\|1)$
G_2	$c(1\|2)$	0	$\cdots$	$c(g\|2)$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
G_g	$c(1\|g)$	$c(2\|g)$	$\cdots$	0

其中 $c(j|i)$ 表示将实际属于 G_i 的样本观测值判为 G_j 所造成的损失度量.

一个判别准则的实质就是对 $\mathbf{R}^p$ 空间作一个不相重叠的划分：$D_1, D_2, \cdots, D_g$, 若样本观测值 $\mathbf{x}$ 落入 D_i, 则判此样本观测值属于总体 G_i, 因此一个判别准则可简记为 $D = (D_1, D_2, \cdots, D_g)$.

以 $P(j|i, D)$ 表示在判别准则 D 之下将事实上来自 G_i 的样本观测值误判为来自 G_j 的概率，则

$$P(j|i, D) = \int_{D_j} f_i(\mathbf{x}) d\mathbf{x}, j = 1, 2, \cdots, g, j \neq i. \tag{7.3.2}$$

由此误判而造成的损失为 $c(j|i)(j = 1, 2, \cdots, g 且 j \neq i)$. 因此, 在一个给定的判别准则 D 之下对 G_i 而言所造成的损失,应该是误判为 $G_1, \cdots, G_{i-1}, G_{i+1}, \cdots, G_k$ 的所有损失，按照各误判概率加权求和，即在此判别规则下，将来自 G_i 的样本观测值错判为其它总体的期望损失为 (注意 $c(i|i) = 0$)

$$l_i \widehat{=} \sum_{j=1, j\neq i}^{g} P(j|i, D) c(j|i) = \sum_{j=1}^{g} P(j|i, D) c(j|i). \tag{7.3.3}$$

又由于各总体 G_i 出现的先验概率为 $q_i(i=1,2,\cdots,g)$, 故在判别准则 D 之下总期望损失为

$$L\widehat{=}\sum_{i=1}^{g} q_i l_i = \sum_{i=1}^{g}\sum_{j=1}^{g} q_i c(j|i)P(j|i,D). \tag{7.3.4}$$

我们看到，总期望损失 L 与判别准则 D 有关， Bayes 判别即选择 $D=(D_1,D_2,\cdots,D_g)$, 使 L 达到最小. 下面我们分两个及多个总体情形分别予以讨论.

7.3.2 两总体的 Bayes 判别

1. 一般总体

设 G_1,G_2 为两个 p 维总体，概率密度分别为 $f_1(\mathbf{x})$ 和 $f_2(\mathbf{x})$, 总体 G_1,G_2 的先验概率分布为 q_1 和 q_2, 误判损失分别为 $c(2|1)$ 和 $c(1|2)$. 对 $\mathbf{R}^p$ 的一个划分 $D=(D_1,D_2)$, 有

$$P(2|1,D)=\int_{D_2} f_1(\mathbf{x})d\mathbf{x},$$

$$P(1|2,D)=\int_{D_1} f_2(\mathbf{x})d\mathbf{x}.$$

根据（ 7.3.4 ），总期望损失为

$$\begin{aligned}
&L = q_1c(2|1)P(2|1,D)+q_2c(1|2)P(1|2,D)\\
&= q_1c(2|1)\textstyle\int_{D_2} f_1(\mathbf{x})d\mathbf{x} + q_2c(1|2)\int_{D_1} f_2(\mathbf{x})d\mathbf{x}\\
&= \textstyle\int_{D_1} q_2c(1|2)f_2(\mathbf{x})d\mathbf{x} - \int_{D_1} q_1c(2|1)f_1(\mathbf{x})d\mathbf{x}\\
&+ \textstyle\int_{D_1} q_1c(2|1)f_1(\mathbf{x})d\mathbf{x} + \int_{D_2} q_1c(2|1)f_1(\mathbf{x})d\mathbf{x}\\
&= \textstyle\int_{D_1}[q_2c(1|2)f_2(\mathbf{x}) - q_1c(2|1)f_1(\mathbf{x})]d\mathbf{x} + q_1c(2|1)
\end{aligned} \tag{7.3.5}$$

最后一个等式成立是因为 $\int_{D_1} f_1(\mathbf{x})d\mathbf{x}+\int_{D_2} f_1(\mathbf{x})d\mathbf{x}=1$. 由于第二项与 D 无关，要使 L 达到最小，只需第一项达到最小. 这只需选择 D_1 为（ 7.3.5 ）中的被积函数取非正值的范围即可，即取 D_1 为

$$\begin{aligned}
D_1 &= \{\mathbf{x}: q_2c(1|2)f_2(\mathbf{x}) - q_1c(2|1)f_1(\mathbf{x}) \le 0\}\\
&= \left\{\mathbf{x}: \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \ge \frac{q_2c(1|2)}{q_1c(2|1)}\right\},
\end{aligned} \tag{7.3.6}$$

此时,

$$D_2 = \{\mathbf{x}: \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{q_2c(1|2)}{q_1c(2|1)}\}. \tag{7.3.7}$$

因此，两一般总体的 Bayes 判别准则如下：对给定的样本观测值 $\mathbf{x}$, 计算两总体的概率密度函数在 $\mathbf{x}$ 处的值，判定

$$\begin{cases} \mathbf{x} \in G_1, 若 \dfrac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \dfrac{q_2 c(1|2)}{q_1 c(2|1)}, \\ \mathbf{x} \in G_2, 若 \dfrac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \dfrac{q_2 c(1|2)}{q_1 c(2|1)}. \end{cases} \tag{7.3.8}$$

下面给出此判别准则的几个特例：

（ 1 ） 等先验概率情形

实际应用中，若各总体的先验概率分布未知，一般有两种处理方法，如果训练样本是通过随机观测得到的，通常取先验概率为各个训练样本的容量占总观测数的比例. 如果对其先验概率分布基本不了解，可假定各总体的先验概率观测值相等. 在两总体情况下，即假定 $q_1 = q_2 = \frac{1}{2}$, 这时 Bayes 判别准则为

$$\begin{cases} \mathbf{x} \in G_1, 若 \dfrac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \dfrac{c(1|2)}{c(2|1)}, \\ \mathbf{x} \in G_2, 若 \dfrac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \dfrac{c(1|2)}{c(2|1)}. \end{cases} \tag{7.3.9}$$

（ 2 ） 等误判损失的情形

若误判损失难以确定，通常可假定 $c(1|2) = c(2|1)$. 这时， Bayes 判别准则为

$$\begin{cases} \mathbf{x} \in G_1, 若 \dfrac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \dfrac{q_2}{q_1}, \\ \mathbf{x} \in G_2, 若 \dfrac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \dfrac{q_2}{q_1}. \end{cases} \tag{7.3.10}$$

（ 3 ） 等先验概率及等误判损失情形

这时， $q_1 = q_2 = \frac{1}{2}, c(1|2) = c(2|1)$, 从而 Bayes 判别准则为

$$\begin{cases} \mathbf{x} \in G_1, 若 f_1(\mathbf{x}) \geq f_2(\mathbf{x}), \\ \mathbf{x} \in G_2, 若 f_1(\mathbf{x}) < f_2(\mathbf{x}). \end{cases} \tag{7.3.11}$$

应用中，总体的概率密度函数通常是未知的，我们可用的资料是来自各总体的训练样本. 通常的作法是利用训练样本对总体的概率密度作非参数估计 (如最邻近估计，核估计等). 由于这些估计涉及较多的统计和数学知识，在此不作进一步介绍. 下面只就正态总体情况作详细讨论.

2. 两正态总体的 Bayes 判别

设 G_1, G_2 为两个不同的 p 维正态总体，这时其概率密度为

$$f_i(\mathbf{x}) = (2\pi)^{-\frac{p}{2}} |\mathbf{\Sigma}_i|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mathbf{x} - \mu_\mathbf{i})' \mathbf{\Sigma_i^{-1}} (\mathbf{x} - \mu_\mathbf{i})\}, i = 1, 2, \tag{7.3.12}$$

其中 $\mu_\mathbf{i}$和$\mathbf{\Sigma_i}(i = 1, 2)$ 为两总体的均值向量和协方差矩阵， $|\mathbf{\Sigma_i}|$ 表示矩阵 $\mathbf{\Sigma_i}$ 的行列式 (i=1,2).

（ 1 ）若 $\mathbf{\Sigma_1 = \Sigma_2 = \Sigma}$

这时，由（ 7.2.3 ）和（ 7.2.4 ）可得

$$\begin{aligned}\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} &= \exp\{\frac{1}{2}[(\mathbf{x}-\mu_{\mathbf{2}})'\mathbf{\Sigma^{-1}}(\mathbf{x}-\mu_{\mathbf{2}}) - (\mathbf{x}-\mu_{\mathbf{1}})'\mathbf{\Sigma^{-1}}(\mathbf{x}-\mu_{\mathbf{1}})]\} \\ &= \exp\{\frac{1}{2}[d^2(\mathbf{x},G_2) - d^2(\mathbf{x},G_1)]\} \\ &= \exp\{W(\mathbf{x})\},\end{aligned} \tag{7.3.13}$$

其中 $W(\mathbf{x})$ 与（ 7.2.4 ）相同，即

$$W(\mathbf{x}) = [\mathbf{x} - \frac{\mathbf{1}}{\mathbf{2}}(\mu_{\mathbf{1}} + \mu_{\mathbf{2}})]'\mathbf{\Sigma^{-1}}(\mu_{\mathbf{1}} - \mu_{\mathbf{2}}). \tag{7.3.14}$$

从而 Bayes 判别准则（ 7.3.8 ）为

$$\begin{cases} \mathbf{x} \in G_1, 若 W(\mathbf{x}) \geq \ln\left[\dfrac{q_2 c(1|2)}{q_1 c(2|1)}\right], \\ \mathbf{x} \in G_2, 若 W(\mathbf{x}) < \ln\left[\dfrac{q_2 c(1|2)}{q_1 c(2|1)}\right]. \end{cases} \tag{7.3.15}$$

我们看到, 在总体服从正态分布的假定下，Bayes 判别函数与第二节的等协方差矩阵的距离判别函数是一样的，只是判别准则中的判别限有所差异，这是因为 Bayes 判别考虑了总体的先验概率分布和误判损失．若假定了等先验概率和等误判损失，则二者就完全一样了．但值得注意的是距离判别中并未假定 G_1 和 G_2 为正态总体.

实际应用中,若 $\mu_{\mathbf{1}}, \mu_{\mathbf{2}}$和$\mathbf{\Sigma}$ 未知,则可用训练样本估计,即用 $\hat{\mu}_{\mathbf{1}} = \bar{\mathbf{x}}^{(\mathbf{1})}, \hat{\mu}_{\mathbf{2}} = \bar{\mathbf{x}}^{(\mathbf{2})}$ 以及 $\hat{\mathbf{\Sigma}} = \dfrac{(n_1-1)\mathbf{S}_1 + (n_2-1)\mathbf{S}_2}{n_1+n_2-2}$ 代替 $W(\mathbf{x})$ 中的 $\mu_{\mathbf{1}}, \mu_{\mathbf{2}}$和$\mathbf{\Sigma}$, 其中 $\bar{\mathbf{x}}^{(\mathbf{1})}, \bar{\mathbf{x}}^{(\mathbf{2})}$ 及 $\mathbf{S_1}, \mathbf{S_2}$ 的定义见（ 7.2.7 ）和（ 7.2.8 ）.

例 7.3.1 下表数据是某气象站预报某地区有无春旱的观测资料， x_1和x_2 是与气象有关的综合预报因子．其中包括春旱发生的 6 个年份的 x_1, x_2 的观测值和无春旱的 8 个年份的相应观测值．其先验概率分别用训练样本的容量比例确定，即 $q_1 = \frac{6}{14}, q_2 = \frac{8}{14}$, 并假定误判损失 $c(1|2) = c(2|1)$. 试在正态总体及等协方差矩阵的假定下建立判别准则.

由表中所给数据可求得

$$\begin{aligned}&\hat{\mu}_1 = \bar{\mathbf{x}}^{(1)} = (25.32, -2.42)', \hat{\mu}_2 = \bar{\mathbf{x}}^{(2)} = (22.03, -1.19)', \\ &\hat{\mu}_1 - \hat{\mu}_2 = \bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)} = (3.29, -1.23)'.\end{aligned}$$

表 7.3.2 某气象站预报有无春旱的数据

	春旱				无春旱		
序号	x_1	x_2	$\tilde{W}(\mathbf{x})$	序号	x_1	x_2	$\tilde{W}(\mathbf{x})$
1	24.8	-2.0	6.886	1	22.1	-0.7	5.624
2	24.1	-2.4	6.907	2	21.6	-1.4	5.835
3	26.6	-3.0	7.790	3	22.0	-0.8	5.647
4^*	23.5	-1.9	6.527	4	22.8	-1.6	6.217
5	25.5	-2.1	7.100	5	22.7	-1.5	6.146
6	27.4	-3.1	8.029	6	21.5	-1.0	5.622
				7	22.1	-1.2	5.861
				8	21.4	-1.3	5.740

$\frac{1}{2}(\hat{\mu}_1+\hat{\mu}_2)=\frac{1}{2}(\bar{\mathbf{x}}^{(1)}+\bar{\mathbf{x}}^{(2)})=(23.68,-1.81)'$,

$\mathbf{S}_1=\begin{bmatrix}2.220 & -0.640\\ -0.640 & 0.270\end{bmatrix},\mathbf{S}_2=\begin{bmatrix}0.270 & 0.007\\ 0.007 & 0.107\end{bmatrix}$,

$\hat{\mathbf{\Sigma}}=\frac{1}{12}(5\mathbf{S}_1+7\mathbf{S}_2)=\begin{bmatrix}1.08 & -0.26\\ -0.26 & 0.17\end{bmatrix}$,

$\hat{\mathbf{\Sigma}}^{-1}=\frac{1}{0.116}\begin{bmatrix}0.17 & 0.26\\ 0.26 & 1.08\end{bmatrix}$,

$\ln\left[\frac{q_2c(1|2)}{q_1c(2|1)}\right]=\ln\left[\frac{4}{3}\right]=0.288.$

由（ 7.3.14 ）得判别函数

$$W(\mathbf{x})=[x_1-23.68,x_2+1.81]'\frac{1}{0.116}\begin{bmatrix}0.17 & 0.26\\ 0.26 & 1.08\end{bmatrix}\begin{bmatrix}3.29\\ -1.23\end{bmatrix}$$

$$=\frac{1}{0.116}[0.2395x_1-0.4730x_2-6.527].$$

为应用方便，令 $\tilde{W}(\mathbf{x})=0.2395x_1-0.4730x_2$, 由 $W(\mathbf{x})\geq 0.288$ 得 $\tilde{W}(\mathbf{x})\geq 6.560$, 从而判别准则为

$$\begin{cases}\mathbf{x}\in G_1(\text{春旱}),\text{若}\tilde{W}(\mathbf{x})\geq 6.560,\\ \mathbf{x}\in G_2(\text{无春旱}),\text{若}\tilde{W}(\mathbf{x})<6.560.\end{cases}$$

由此判别准则回判 14 个样本观测值，其 $\tilde{W}(\mathbf{x})$ 的值列入前表中各总体的最后一列．误判的只有一个，即春旱总体中的第 4 号样本观测值，貌似误判率只有 $\frac{1}{14}=0.07$.

（ 2 ）若 $\mathbf{\Sigma_1}\neq\mathbf{\Sigma_2}$

此时，经推导（留作习题）可得判别准则为

$$\begin{cases}\mathbf{x}\in G_1,\text{若}W^*(\mathbf{x})\geq K,\\ \mathbf{x}\in G_2,\text{若}W^*(\mathbf{x})<K.\end{cases}\tag{7.3.16}$$

其中

$$W^*(\mathbf{x}) = -\frac{1}{2}\mathbf{x}'(\boldsymbol{\Sigma_1^{-1}} - \boldsymbol{\Sigma_2^{-1}})\mathbf{x} + (\mu_1'\boldsymbol{\Sigma_1^{-1}} - \mu_2'\boldsymbol{\Sigma_2^{-1}})\mathbf{x}, \tag{7.3.17}$$

$$K = \ln\left[\frac{q_2 c(1|2)}{q_1 c(2|1)}\right] + \frac{1}{2}\ln\left[\frac{|\boldsymbol{\Sigma_1}|}{|\boldsymbol{\Sigma_2}|}\right] + \frac{1}{2}(\mu_1'\boldsymbol{\Sigma_1^{-1}}\mu_1 - \mu_2'\boldsymbol{\Sigma_2^{-1}}\mu_2). \tag{7.3.18}$$

若 μ_1, μ_2和$\boldsymbol{\Sigma_1}, \boldsymbol{\Sigma_2}$ 未知，可用训练样本估计，即 $\hat{\mu}_1 = \bar{\mathbf{x}}^{(1)}, \hat{\mu}_2 = \bar{\mathbf{x}}^{(2)}\hat{\boldsymbol{\Sigma}}_1 = \mathbf{S_1}, \hat{\boldsymbol{\Sigma}}_2 = \mathbf{S_2}$.

例 7.3.2（续例 7.3.1 ）对例 7.3.1 有关春旱预报的数据，假定两总体均服从二维正态分布，但 $\boldsymbol{\Sigma_1} \neq \boldsymbol{\Sigma_2}$, 试建立相应的 Bayes 判别准则.

由前例所求得的两总体的样本协方差阵 $\mathbf{S_1}$和$\mathbf{S_2}$ 可得

$$|\mathbf{S}_1| = 0.190, |\mathbf{S}_2| = 0.029,$$

$$\mathbf{S}_1^{-1} = \begin{bmatrix} 1.421 & 3.368 \\ 3.368 & 11.684 \end{bmatrix}, \mathbf{S}_2^{-1} = \begin{bmatrix} 3.690 & -0.241 \\ -0.241 & 9.310 \end{bmatrix},$$

$$\mathbf{S}_1^{-1} - \mathbf{S}_2^{-1} = \begin{bmatrix} -2.269 & 3.609 \\ 3.609 & 2.374 \end{bmatrix},$$

$$(\bar{\mathbf{x}}^{(1)})'\mathbf{S_1^{-1}} = (27.829, 57.002), (\bar{\mathbf{x}}^{(2)})'\mathbf{S_2^{-1}} = (81.577, -16.388),$$

$$(\bar{\mathbf{x}}^{(1)})'\mathbf{S_1^{-1}} - (\bar{\mathbf{x}}^{(2)})'\mathbf{S_2^{-1}} = (-53.748, 73.390).$$

代入（ 7.3.17 ）式可求得

$$W^*(\mathbf{x}) = 1.135x_1^2 - 1.187x_2^2 - 3.609x_1x_2 - 53.748x_1 + 73.390x_2.$$

又由于

$$(\bar{\mathbf{x}}^{(1)})'\mathbf{S_1^{-1}}\bar{\mathbf{x}}^{(1)} = 566.688, (\bar{\mathbf{x}}^{(2)})'\mathbf{S_2^{-1}}\bar{\mathbf{x}}^{(2)} = 1816.654,$$

由（ 7.3.18 ）可求得

$$K = 0.288 + \frac{1}{2}\ln\left(\frac{0.190}{0.029}\right) + \frac{1}{2}(566.688 - 1816.654) = -623.76.$$

故判别准则为

$$\begin{cases} \mathbf{x} \in G_1, 若W^*(\mathbf{x}) \geq -623.76, \\ \mathbf{x} \in G_2, 若W^*(\mathbf{x}) < -623.76. \end{cases}$$

利用此准则对愿 14 个样本进行回判得 $W^*(\mathbf{x})$ 值如表 7.3.3.

表 7.3.3 气象站预报有无春旱数据的回判结果

序号	1	2	3	4	5	6	7	8
春旱 G_1	-607.40	-610.34	-569.47	-618.86	-598.63	-552.95		
无春旱 G_2	-629.61	-627.35	-629.07	-624.24	-625.09	-627.91	-627.55	-627.43

由此表可知，所有样本回判结果均无误，即貌似误判率为零. 由于此题中两总体的训练样本容量均很小，因此还不能简单地认为此例中的判别准则较之例 7.3.1 中的准则为优，但从 $\mathbf{\Sigma_1}$和$\mathbf{\Sigma_2}$ 的估计量 $\mathbf{S_1}, \mathbf{S_2}$ 来看，二者确有较大差异，因此认为 $\mathbf{\Sigma_1} \neq \mathbf{\Sigma_2}$ 似乎更为合理，而此例的计算量要比例 7.3.1 大许多.

7.3.3 多总体的 Bayes 判别

1. 一般总体

设 $G_1, G_2, \cdots, G_g$ 为 g 个不同的 p 维总体,概率密度函数分别为 $f_1(\mathbf{x}), f_2(\mathbf{x}), \cdots, f_g(\mathbf{x})$, 各总体的先验概率为 $q_1, q_2, \cdots, q_g$, 误判损失为 $c(j|i)(i, j = 1, 2, \cdots, g, i \neq j)$. 为方便起见，记 $c(i|i) = 0(i = 1, 2, \cdots, g)$. 令

$$h_k(\mathbf{x}) \hat{=} \sum_{i=1}^{g} q_i f_i(\mathbf{x}) c(k|i). \tag{7.3.19}$$

在总期望损失（ 7.3.4 ）达到最小的条件下，可以证明其判别准则为

$$\mathbf{x} \in G_i, 若 \min_{1 \leq k \leq g} \{h_k(\mathbf{x})\} = h_i(\mathbf{x}). \tag{7.3.20}$$

即将样本观测值 $\mathbf{x}$ 判归为使 $h_k(\mathbf{x})(k = 1, 2, \cdots, g)$ 达到最小的那个总体.

特别地，若误判损失相同，即 $c(j|i) = c(i \neq j)$, 这时（ 7.3.19 ）为（注意 $c(i|i) = 0$ ）

$$h_i(\mathbf{x}) = c[\sum_{k=1}^{g} q_k f_k(\mathbf{x}) - q_i f_i(\mathbf{x})]. \tag{7.3.21}$$

由于 $\sum\limits_{k=1}^{g} q_k f_k(\mathbf{x})$ 与 i 无关，故 $h_i(\mathbf{x})$ 最小等价于 $q_i f_i(\mathbf{x})$ 最大. 从而在等误判损失下，多总体的 Bayes 判别准则为

$$\mathbf{x} \in G_i, 若 q_i f_i(\mathbf{x}) \geq q_j f_j(\mathbf{x}) 对一切 j \neq i 成立. \tag{7.3.22}$$

例 7.3.3 设有三个总体 G_1, G_2和G_3, 概率密度函数分别为 $f_1(\mathbf{x}), f_2(\mathbf{x}), f_3(\mathbf{x})$, 各总体的先验概率为 $q_1 = 0.05, q_2 = 0.60, q_3 = 0.35$, 误判损失如下表：

判定为 实际为	G_1	G_2	G_3
G_1	$c(1\|1) = 0$	$c(2\|1) = 10$	$c(3\|1) = 50$
G_2	$c(1\|2) = 500$	$c(2\|2) = 0$	$c(3\|2) = 200$
G_3	$c(1\|3) = 100$	$c(2\|3) = 50$	$c(3\|3) = 0$

现有一样本观测值 $\mathbf{x_0}$ 使 $f_1(\mathbf{x}_0) = 0.01, f_2(\mathbf{x}_0) = 0.85, f_3(\mathbf{x}_0) = 2$, 按 Bayes 判别准则，我们应判别 $\mathbf{x_0}$ 属于哪个总体？若假定误判损失均相同，情况又如何？

在给定的误判损失下，由（ 7.3.19 ）可得

$$h_1(\mathbf{x}_0) = q_2 f_2(\mathbf{x}_0)c(1|2) + q_3 f_3(\mathbf{x}_0)c(1|3)$$
$$= 0.60 \times 0.85 \times 500 + 0.35 \times 2 \times 100 = 325,$$

$$h_2(\mathbf{x}_0) = q_1 f_1(\mathbf{x}_0)c(2|1) + q_3 f_3(\mathbf{x}_0)c(2|3)$$
$$= 0.05 \times 0.01 \times 10 + 0.35 \times 2 \times 50 = 35.055,$$

$$h_3(\mathbf{x}_0) = q_1 f_1(\mathbf{x}_0)c(3|1) + q_2 f_2(\mathbf{x}_0)c(3|2)$$
$$= 0.05 \times 0.01 \times 50 + 0.60 \times 0.85 \times 200 = 102.025.$$

由于 $\min\limits_{1\le i\le 3}\{h_i(\mathbf{x}_0)\} = h_2(\mathbf{x}_0)$, 根据判别准则（ 7.3.20 ），我们判 $\mathbf{x}_0 \in G_2$.

如果假定误判损失相等，这时

$$q_1 f_1(\mathbf{x}_0) = 0.05 \times 0.01 = 0.0005,$$
$$q_2 f_2(\mathbf{x}_0) = 0.60 \times 0.85 = 0.510,$$
$$q_3 f_3(\mathbf{x}_0) = 0.35 \times 2 = 0.70.$$

由判别准则（ 7.3.22 ）知，我们应判 $\mathbf{x}_0 \in G_3$.

2. 正态总体

进一步假定 $G_1, G_2, \cdots, G_g$ 均为 p 维正态总体，其概率密度函数为

$$f_i(\mathbf{x}) = (2\pi)^{-\frac{p}{2}}|\mathbf{\Sigma}_i|^{-\frac{1}{2}}\exp\{-\frac{1}{2}(\mathbf{x}-\mu_\mathbf{i})'\mathbf{\Sigma_i^{-1}}(\mathbf{x}-\mu_\mathbf{i})\}, i = 1, 2, \cdots, g.$$

在误判损失不等的情况下，只能按准则（ 7.3.20 ）判别. 在等误判损失下，正态总体的 Bayes 判别有更明确的表达式.

易知，在等误判损失下， Bayes 判别准则（ 7.3.22 ）等价于

$$\mathbf{x} \in G_i, 若 \ln(q_i f_i(\mathbf{x})) \ge \ln(q_j f_j(\mathbf{x}))对一切 j \ne i 成立.$$

代入正态概率密度函数得

$$\ln(q_i f_i(\mathbf{x})) = \ln q_i - \frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{\Sigma_i}| - \frac{1}{2}(\mathbf{x}-\mu_\mathbf{i})'\mathbf{\Sigma_i^{-1}}(\mathbf{x}-\mu_\mathbf{i}).$$

令

$$d_i^Q(\mathbf{x}) = -\frac{1}{2}\ln|\mathbf{\Sigma_i}| - \frac{1}{2}(\mathbf{x}-\mu_\mathbf{i})'\mathbf{\Sigma_i^{-1}}(\mathbf{x}-\mu_\mathbf{i}) + \ln q_i, i = 1, 2, \cdots, g, \quad (7.3.23)$$

这时 Bayes 判别准则为

$$\mathbf{x} \in G_i, 若 \max_{1\le j\le g}\{d_k^Q(\mathbf{x})\} = d_i^Q(\mathbf{x}). \quad (7.3.24)$$

进一步，若各正态总体的协方差矩阵相等，即

$$\mathbf{\Sigma_1} = \mathbf{\Sigma_2} = \cdots = \mathbf{\Sigma_g} = \mathbf{\Sigma}$$

则（ 7.3.23 ）为

$$d_i^Q(\mathbf{x}) = -\frac{1}{2}\ln|\mathbf{\Sigma}| - \frac{\mathbf{1}}{\mathbf{2}}\mathbf{x}'\mathbf{\Sigma^{-1}}\mathbf{x} + \mu_\mathbf{i}'\mathbf{\Sigma^{-1}}\mathbf{x} - \frac{\mathbf{1}}{\mathbf{2}}\mu_\mathbf{i}'\mathbf{\Sigma^{-1}}\mu_\mathbf{i} + \ln q_i, i = 1, 2, \cdots, g.$$

由于前两项与 i 无关，若令

$$d_i(\mathbf{x}) = \mu_\mathbf{i}'\mathbf{\Sigma^{-1}}\mathbf{x} - \frac{\mathbf{1}}{\mathbf{2}}\mu_\mathbf{i}'\mathbf{\Sigma^{-1}}\mu_\mathbf{i} + \ln q_i, i = 1, 2, \cdots, g, \tag{7.3.25}$$

则 $d_i(\mathbf{x})$ 为 $\mathbf{x}$ 的线性函数．这时 Bayes 判别准则（ 7.3.24 ）简化为

$$\mathbf{x} \in G_i, 若 \max_{1\le k\le g}\{d_k(\mathbf{x})\} = d_i(\mathbf{x}). \tag{7.3.26}$$

实际应用中, 若已知 $G_1, G_2, \cdots, G_g$ 为正态总体, 但 $\mu_\mathbf{i}, \mathbf{\Sigma_i}(i = 1, 2, \cdots, g)$ 未知，则可用训练样本的样本均值 $\bar{\mathbf{x}}^{(i)}$ 和样本方差 $\mathbf{S}_i(i = 1, 2, \cdots, g)$ 作为 $\mu_\mathbf{i}$和$\mathbf{\Sigma_i}(i = 1, 2, \cdots, g)$ 的估计, 代入其概率密度函数中再作相应的判别分析. 若假定 $\mathbf{\Sigma}_i$ 均相等，则用 $\mathbf{S}_i(i = 1, 2, \cdots, g)$ 联合估计 $\mathbf{\Sigma}$, 即

$$\hat{\mathbf{\Sigma}} = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2 + \cdots + (n_g - 1)\mathbf{S}_g}{n_1 + n_2 + \cdots + n_g - g}$$

例 7.3.4 某商学院在招收研究生时,以学生在大学期间的平均学分 (GPA)x_1 和管理能力考试 (GMAT) 成绩 x_2 帮助录取研究生．对申请者划分为三类, G_1: 录取， G_2: 未录取， G_3: 待定．表 7.3.4 记录了近期报考者的 x_1, x_2 值和录取情况.

在误判损失相等，总体先验分布相同且三个总体服从协方差矩阵相等的正态分布的假定下建立 Bayes 判别准则．另外，假设一位申请者的 GPA 为 $x_1 = 3.21$,GMAT 成绩 $x_2 = 497$, 利用所建立的准则判别该申请者应归入哪一类.

缺表 7.3.4

由所给数据知 $n_1 = 31, n_2 = 28, n_3 = 26$, 并可求得

$$\bar{\mathbf{x}}^{(1)} = \begin{bmatrix} 3.40 \\ 561.23 \end{bmatrix}, \bar{\mathbf{x}}^{(2)} = \begin{bmatrix} 2.48 \\ 447.07 \end{bmatrix}, \bar{\mathbf{x}}^{(3)} = \begin{bmatrix} 2.99 \\ 446.23 \end{bmatrix},$$

$$S_1 = \begin{bmatrix} 0.0436 & 0.0581 \\ 0.0581 & 4618.2473 \end{bmatrix}, S_2 = \begin{bmatrix} 0.0336 & -1.1920 \\ -1.1920 & 3891.2540 \end{bmatrix},$$
$$S_3 = \begin{bmatrix} 0.0297 & -5.4038 \\ -5.4038 & 2246.9046 \end{bmatrix},$$

$$\hat{\mathbf{\Sigma}} = \frac{30\mathbf{S}_1 + 27\mathbf{S}_2 + 25\mathbf{S}_3}{31 + 28 + 26 - 3} = \begin{bmatrix} 0.0361 & -2.0188 \\ -2.0188 & 3655.9011 \end{bmatrix},$$
$$\hat{\mathbf{\Sigma}}^{-1} = \frac{1}{127.9025}\begin{bmatrix} 3655.9011 & 2.0188 \\ 2.0188 & 0.0361 \end{bmatrix}.$$

由于 $q_1 = q_2 = q_3$, 故（ 7.3.25 ）中的 $\ln q_i$ 可省去，从而可得

$$\hat{d}_1(\mathbf{x}) = (3.40, 561.23)\hat{\mathbf{\Sigma}}^{-1}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \frac{1}{2}(3.40, 561.23)\hat{\mathbf{\Sigma}}^{-1}\begin{bmatrix} 3.40 \\ 561.23 \end{bmatrix}$$
$$= 106.042x_1 + 0.212x_2 - 239.782,$$

同理可求得

$$\hat{d}_2(\mathbf{x}) = 77.943x_1 + 0.165x_2 - 133.607,$$
$$\hat{d}_3(\mathbf{x}) = 92.508x_1 + 0.173x_2 - 176.929.$$

因而 Bayes 判别准则为

$$\mathbf{x} = (x_1, x_2)' \in G_i, \text{若} \max\{\hat{d}_1(\mathbf{x}), \hat{d}_2(\mathbf{x}), \hat{d}_3(\mathbf{x})\} = \hat{d}_i(\mathbf{x}),$$

其中 i=1,2,3.

对新申请者， $\mathbf{x}_0 = (321, 497)'$, 可求得

$$\hat{d}_1(\mathbf{x}_0) = 205.98, \hat{d}_2(\mathbf{x}_0) = 198.60, \hat{d}_3(\mathbf{x}_0) = 206.00.$$

比较知 $\max\{\hat{d}_1(\mathbf{x}_0), \hat{d}_2(\mathbf{x}_0), \hat{d}_3(\mathbf{x}_0)\} = \hat{d}_3(\mathbf{x}_0)$, 故将该申请者归入待定类.

进一步，利用所建立的准则对训练样本进行回判，其结果表明（可见参考文献 [4]$P_{535,536}$ 的 SAS 程序输出结果），G_1 中的第 2,3,24 和 31 号申请者被误判入待定类 G_3; 属于 G_2 的第 58 和 59 号申请者被误判入待定类 G_3; 属于 G_3 的第 66 号申请者被误判入录取类 G_1. 从而其貌似误判率为

$$\hat{\alpha} = \frac{4+2+1}{n_1+n_2+n_3} = \frac{7}{85} = 0.082.$$

若利用刀切法评估此准则的优良性，其计算结果表明 G_1 中有 5 位申请者被误判入 G_3;G_2 中有 2 位被误判入 G_3;G_3 中有两位分别被误判入 G_1 和 G_2, 因此其误判比例为

$$\hat{\alpha}_J = \frac{9}{85} = 0.106$$

我们看到 $\hat{\alpha}_J$ 的确比 $\hat{\alpha}$ 大.

第八章　聚类分析

在实际问题中，经常要遇到分类的问题. 例如，在考古学中，要将某些古生物化石进行科学的分类；在生物学中，要根据各生物体的综合特征进行分类；在经济学中，为了研究不同地区城镇居民的收入及消费情况，往往需要划分为不同的类型去研究；在产品质量管理中，也要根据各产品的某些重要指标而将其分为一等品，二等品等等. 总之，科学的分类方法无论在自然科学，还是在社会科学中，都有着极其广泛的应用.

随着人类社会的发展与科学技术的进步，对分类学的要求也越来越高. 有时，只凭经验和专业知识还不能进行科学有效的分类，于是数学这一有力的工具被逐渐引入到分类学中，形成了一门新兴的学科——数值分类学. 随着多元分析方法的引进，从数值分析学中逐渐分离出了聚类分析这个分支.

聚类分析的基本思想是，从一批样品的多个指标变量中，定义能度量样品间或变量间相似程度（或亲疏关系）的统计量，在此基础上求出各样品（或变量）之间的相似程度度量值，按相似程度的大小，把样品（或变量）逐一分类，关系密切的类聚集到一个小的分类单位，关系疏远的类聚集到一个大的分类单位，直到所有的样品或变量都聚集完毕，把不同的类型一一划分出来，形成一个亲疏关系谱系图，用以更直观地显示分类对象（样品或变量）的差异和联系.

值得一提的是聚类分析和第七章的判别分析都是研究分类问题，但二者有本质的区别. 聚类分析一般上寻求客观分类的方法，事先对总体到底有几种类型无所知晓，而判别分析则是在总体类型划分已知，在各总体分布或来自各总体训练样本的基础上，对当前的新样品用统计的方法判定它们属于哪个总体.

聚类分析的历史还很短，由于在其发展过程中首先是着重于实用，因此相对而言理论上还不够完善. 无论聚类统计量还是聚类的方法，都还未最终定型. 目前，聚类统计量种类繁多，聚类方法也五花八门，但由于聚类分析方法能广泛地应用于解决实际问题，它和回归分析，判别分析一起被称为多元分析的三大实用方法.

本章将重点介绍一些常见的分类统计量和目前使用较为广泛的谱系聚类方法. 关于其它聚类方法，如动态聚类法，有序样品聚类法，分解法，加入法等等.

§8.1 分类统计量

聚类分析所研究的内容包括两个方面，一是对样品进行分类，设 n 个样品，每个样品均用 p 个指标的观测向量 $\mathbf{X}_i(i=1,2,\cdots,n)$ 来表征，要根据 $\mathbf{X}_i$ 间某种相似性度量，将这 n 个样品进行分类. 如某班有 n 个学生，根据每个学生的期末各科考试成绩将该班学生分类（如分为优，良，中，差四类等）. 另一方面是对变量进行分类，即对所考察的 p 个指标 $\mathbf{X}=(X_1,X_2,\cdots,X_p)'$，根据 n 个观测值 $\mathbf{X}_i=(X_{i1},X_{i2},\cdots,X_{ip})'(i=1,2,\cdots,n)$ 及某些相似性原则将这 p 个变量 $X_1,X_2,\cdots,X_p$ 进行分类. 如在服装设计中，往往要测量很多的指标（变量），如身高，上体长，臂长，肩宽，胸围，腰围等，有时需要对这些指标分类，以显示人体各部分的不同特点，以便于服装设计.

对样品进行分类的方法称为 Q 型聚类法，所用的统计量用“距离”这一术语描述；对变量进行分类的方法，称为 R 型聚类法，所用的统计量用“相似系数”描述. 下面分别介绍几种常用的距离和相似系数.

8.1.1 样品间的“相近性”度量 — 距离

和前一章相同，我们将不区分样品与它的指标观测值.

设每个样品 $\mathbf{X}_i$ 有 p 个指标，它们的观测值可表示为

$$\mathbf{X}_i=(X_{i1},X_{i2},\cdots,X_{ip})',i=1,2,\cdots,n. \tag{8.1.1}$$

这时，每个样品 $\mathbf{X}_i$ 可看成 p 维空间中的一个点，n 个样品就组成 p 组空间中的 n 个点，我们很自然地用各点之间的距离来衡量各样品之间的靠近程度.

设 $d(\mathbf{X_i}, \mathbf{X_j})$ 为样品 $\mathbf{X}_i$ 和 $\mathbf{X}_j$ 之间的距离，则一般要求它满足下列三个条件：

(i)$d(\mathbf{X_i}, \mathbf{X_j}) \geq 0$;且$d(\mathbf{X_i}, \mathbf{X_j}) = 0$当且仅当$\mathbf{X}_i = \mathbf{X}_j$;

(ii)$d(\mathbf{X_i}, \mathbf{X_j}) = d(\mathbf{X_j}, \mathbf{X_i})$;

(iii)$d(\mathbf{X_i}, \mathbf{X_j}) \leq d(\mathbf{X_i}, \mathbf{X_k}) + d(\mathbf{X_k}, \mathbf{X_j})$（三角不等式）

在聚类分析中，有时所用的距离并不满足（ iii ），我们在广义的角度上仍称它为距离．下面介绍几种聚类分析中的常用距离.

1.Minkowski 距离

$$d(\mathbf{X_i}, \mathbf{X_j}) \widehat{=} [\sum_{k=1}^{p} |X_{ik} - X_{jk}|^m]^{\frac{1}{m}}. \tag{8.1.2}$$

特别当 $m = 1, 2, \infty$ 时，分别可得到如下三种距离：

2. 绝对距离

$$d(\mathbf{X_i}, \mathbf{X_j}) \widehat{=} \sum_{k=1}^{p} |X_{ik} - X_{jk}|. \tag{8.1.3}$$

3. 欧氏距离

$$d(\mathbf{X_i}, \mathbf{X_j}) \widehat{=} [\sum_{k=1}^{p} (X_{ik} - X_{jk})^2]^{\frac{1}{2}}. \tag{8.1.4}$$

4.Chebishov 距离

$$d(\mathbf{X_i}, \mathbf{X_j}) \widehat{=} \max_{1 \leq k \leq p} \{|X_{ik} - X_{jk}|\}. \tag{8.1.5}$$

以上各距离在 Q 型聚类中是比较常用的，但当指标的测量值相差悬殊时，应先对数据进行标准化，然后再用标准化的数据计算距离，具体方法如下：

令

$$\bar{X}_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij}, j = 1, 2, \cdots, p, \tag{8.1.6}$$

$$s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_{ij} - \bar{X}_j)^2, j = 1, 2, \cdots, p, \tag{8.1.7}$$

$$r_j = \max_{1 \leq i \leq n} \{X_{ij}\} - \min_{1 \leq i \leq n} \{X_{ij}\}, j = 1, 2, \cdots, p. \tag{8.1.8}$$

对数据的标准化方法一般有如下两种：

$$X_{ij}^* = \frac{X_{ij} - \bar{X}_j}{s_j}, i = 1, 2, \cdots, n, j = 1, 2, \cdots, p. \tag{8.1.9}$$

或

$$X_{ij}' = \frac{X_{ij} - \min\limits_{1 \leq i \leq n} \{X_{ij}\}}{r_j}, i = 1, 2, \cdots, n, j = 1, 2, \cdots, p. \tag{8.1.10}$$

若考虑 p 个指标的相关性等问题，我们可以采用方差加权距离或 Mahalanobis 距离.

5. 方差加权距离

$$d(\mathbf{X_i},\mathbf{X_j})\hat{=}\left(\sum_{j=1}^{p}\frac{(X_{ij}-X_{jj})^2}{\sigma_j^2}\right)^{\frac{1}{2}}, \tag{8.1.11}$$

其中 σ_j^2 为第 j 个指标的方差.

6.Mahalanobis 距离

$$d(\mathbf{X_i},\mathbf{X_j})\hat{=}(\mathbf{X_i}-\mathbf{X_j})'\mathbf{\Sigma^{-1}}(\mathbf{X_i}-\mathbf{X_j}), \tag{8.1.12}$$

其中 $\mathbf{\Sigma}$ 为 p 个指标的协方差矩阵.

实用中，若 σ_j^2 和 $\mathbf{\Sigma}$ 未知，可用观测数据对其作估计，一般用（ 8.1.7 ）中 s_j^2 和样本协方差阵 $\mathbf{S}$（见第六章（ 6.1.22 ）式）作为 σ_j^2 和 $\mathbf{\Sigma}$ 的估计.

用聚类分析解决实际问题时，选用何种距离是十分重要的，这通常要结合有关专业的实际背景而定．距离的定义有很大的灵活性，有时可根据实际问题定义新的距离．下面我们举例说明此问题.

例 8.1.1 欧洲各国的语言有许多相似之处，有的甚至十分相近．当然，单词的词意会随着历史的进程而发生变化，但一般来说，表示数字 1,2,3,$\cdots$ 的单词的词意变化不大. 因此为了研究这些语言之间的历史关系, 也许通过比较各种语言对数字的表达比较恰当．表 8.1.1 列出了英语（ E ），挪威语（ N ），丹麦语（ Da ），荷兰语（ Du ），德语（ G ），法语（ Fr ），西班牙语（ S ），意大利语（ I ），波兰语（ P ），匈牙利语（ H ），和芬兰语（ Fi ） 11 种语言对数字 $1,2,\cdots,10$ 的拼写方法，我们希望定义出这 11 个数字拼写间的距离．缺表 8.1.1

显然，此问题无法直接用前述各公式计算距离．仔细观察表 8.1.1, 我们发现前三种文字（英语，挪威语，丹麦语）很相似，尤其是每个单词的第一个字母．于是产生一种定义距离的办法：用每种语言的 10 个数字表达中的第一个字母不同的个数来定义这两种语言之间的距离，例如英语和挪威语只有数字 1 和 8 的第一个字母不同，故这两种语言之间的距离定义为 2. 用此办法，我们便可定义出任意两种语言间的距离．写成矩阵形式列于表 8.1.2 中（矩阵中的对称部分未写出）：

表 8.1.2 欧洲 11 种语言之间的距离

$$
\begin{array}{c}
\\
E \\
N \\
Da \\
Du \\
G \\
Fr \\
S \\
I \\
P \\
H \\
Fi
\end{array}
\begin{array}{c}
\begin{array}{cccccccccccc} E & N & Da & Du & G & Fr & S & I & P & H & Fi \end{array} \\
\left[\begin{array}{cccccccccccc}
0 & & & & & & & & & & \\
2 & 0 & & & & & & & & & \\
2 & 1 & 0 & & & & & & & & \\
7 & 5 & 6 & 0 & & & & & & & \\
6 & 4 & 5 & 5 & 0 & & & & & & \\
6 & 6 & 6 & 9 & 7 & 0 & & & & & \\
6 & 6 & 5 & 9 & 7 & 2 & 0 & & & & \\
6 & 6 & 5 & 9 & 7 & 1 & 1 & 0 & & & \\
7 & 7 & 6 & 10 & 8 & 5 & 3 & 4 & 0 & & \\
9 & 8 & 8 & 8 & 9 & 10 & 10 & 10 & 10 & 0 & \\
9 & 9 & 9 & 9 & 9 & 9 & 9 & 9 & 9 & 8 & 0
\end{array}\right]
\end{array}
$$

8.1.2 变量间的“关联性”度量 — 相似系数

当对 p 个指标变量进行聚类时，用相似系数衡量变量间的关联程度．一般地，称 $c_{j\beta}$ 为变量 X_j 和 X_β 间的相似系数，如果对一切 $1 \le j, \beta \le p$ 满足：

(i)$|c_{j\beta}| \le 1$;

(ii)$c_{jj} = 1$;

(iii)$c_{j\beta} = c_{\beta j}$.

$c_{j\beta}$ 越接近于 1，说明变量 X_j 与 X_β 关系越密切．设 $(X_{1j}, X_{2j}, \cdots, X_{nj})'$ 表示对变量 X_j 的 n 个观测值 $(j = 1, 2, \cdots, p)$，常用的相似系数有

1. 夹角余弦

$$
c_{j\beta} \widehat{=} \frac{\sum\limits_{i=1}^{n} X_{ij} X_{i\beta}}{[\sum\limits_{i=1}^{n} X_{ij}^2 \cdot \sum\limits_{i=1}^{n} X_{i\beta}^2]^{\frac{1}{2}}}. \tag{8.1.13}
$$

若将变量 X_j 的 n 个观测值 $(X_{1j}, X_{2j}, \cdots, X_{nj})'$ 与变量 X_β 的相应 n 个观测值 $(X_{1\beta}, X_{2\beta}, \cdots, X_{n\beta})'$ 看成 n 维空间中的两个向量，$c_{j\beta}$ 正好是这两个向量夹角的余弦．这个统计量在图像识别中很有用.

2. 相关系数

从统计角度看，两个随机变量的相关系数是描述这两个变量关联性（线性关系）强弱的一个很有用的特征数字．因此，用任意两个变量的 n 个观测值对其相关系数的估计可作为两个变量关联性的一种度量．其定义为

$$
r_{j\beta} \widehat{=} \frac{\sum\limits_{i=1}^{n} (X_{ij} - \bar{X}_j)(X_{i\beta} - \bar{X}_\beta)}{[\sum\limits_{i=1}^{n} (X_{ij} - \bar{X}_j)^2 \cdot \sum\limits_{i=1}^{n} (X_{i\beta} - \bar{X}_\beta)^2]^{\frac{1}{2}}}, \tag{8.1.14}
$$

其中 $\bar{X}_j (j = 1, 2, \cdots, p)$ 的定义见（ 8.1.6 ）式． $r_{j\beta} (1 \le j, \beta \le p)$ 其实就是 $\mathbf{X} = (X_1, \cdots, X_p)'$ 的样本相关系数矩阵中的各元素.

当变量为定性变量（如性别，职业，等级）时，也可定义样品间的“距离”和变量间的“相似系数”，在此不作深入讨论.

§8.2 谱系聚类法

谱系聚类法是目前应用较为广泛的一种聚类方法. 有关该方法的研究内容相当丰富，而且许多统计软件（如 SAS ）中都有专门的程序. 谱系聚类法是根据古老的植物分类学的思想对研究对象进行分类的一种方法. 我们知道，在植物分类学中，分类的单位：门、纲、目、科、属、种，其中种是分类的基本单位. 分类单位越小它所包含的植物种类就越少，植物间的共同特征就越多. 利用这种分类思想，谱系聚类法首先视各样品（或变量）自成一类，然后把最相似的样品（或变量）聚为小类，再将已聚合的小类按其相似性再聚合，随着相似性的减弱，最后将一切子类都聚合到一个大类，从而得到一个按相似性大小聚结起来的一个谱系关系.

在谱系聚类法的合并过程中要涉及到两个类之间的距离（或相似系数）问题. 类与类之间的距离有许多定义方式，不同的定义方式就产生了不同的谱系聚类法. 本节中，我们首先引进三种类与类之间的距离，然后再详细介绍谱系聚类法.

8.2.1 类与类之间的距离

我们先就样品聚类的情形予以讨论，并为简单起见，以 i,j 等分别表示样品 $\mathbf{X}_i, \mathbf{X}_j$, 以 d_{ij} 简记样品 i 与 j 之间的距离 $d(\mathbf{X}_i, \mathbf{X}_j)$, 用 G_p 和 G_q 表示两个类，它们所包含的样品个数分别记为 n_p 和 n_q, 类 G_p 与 G_q 之间的距离用 $D(G_p, G_q)$ 表示. 下面给出三种最常用的类与类之间距离的定义.

1. 最短距离

$$D(G_p, G_q) \widehat{=} \min\{d_{ij} | i \in G_p, j \in G_q\}, \tag{8.2.1}$$

即定义 G_p 与 G_q 中最邻近的两个样品的距离为这两个类之间的距离.

类与类之间的最短距离有如下的递推公式，设 G_r 为由 G_p 与 G_q 合并所得，则 G_r 与其它类 $G_k (k \neq p, q)$ 的最短距离为

$$\begin{aligned} D(G_r, G_k) &= \min\{d_{ij} | i \in G_r, j \in G_k\} \\ &= \min\{\min\{d_{ij} | i \in G_p, j \in G_k\}, \min\{d_{ij} | i \in G_q, j \in G_k\}\} \\ &= \min\{D(G_p, G_k), D(G_q, G_k)\}. \end{aligned} \tag{8.2.2}$$

2. 最长距离

$$D(G_p, G_q) \widehat{=} \max\{d_{ij} | i \in G_p, j \in G_q\}, \tag{8.2.3}$$

即定义 G_p 与 G_q 中相距最远的两个样品之间的距离为这两个类之间的距离.

类与类之间的最长距离有如下的递推公式，设类 G_r 由 G_p 与 G_q 合并而成，则 G_r 到 $G_k(k \neq p,q)$ 的最长距离可由下式递推得到，

$$\begin{aligned} D(G_r,G_k) &= \max\{d_{ij}|i \in G_r, j \in G_k\} \\ &= \max\{\max\{d_{ij}|i \in G_p, j \in G_k\}, \max\{d_{ij}|i \in G_q, j \in G_k\}\} \\ &= \max\{D(G_p,G_k), D(G_q,G_k)\}. \end{aligned} \tag{8.2.4}$$

3. 类平均距离

$$D(G_p,G_q) \widehat{=} \frac{1}{n_p n_q} \sum_{i \in G_p} \sum_{j \in G_q} d_{ij}, \tag{8.2.5}$$

即用 G_p 与 G_q 中每两两样品间的距离的平均数作为这两个类之间的距离.

类平均距离的递推公式如下，设类 G_i 中的样品数（或变量数）为 n_i，类 G_r 由 G_p 与 G_q 合并而得，即 $G_r = \{G_p, G_q\}$，则 G_r 到其它类 $G_k(k \neq p,q)$ 的类平均距离有如下递推公式，

$$\begin{aligned} D(G_r,G_k) &= \frac{1}{n_k n_r} \sum_{i \in G_k} \sum_{j \in G_r} d_{ij} \\ &= \frac{1}{n_k n_r} [\sum_{i \in G_k} (\sum_{j \in G_p} d_{ij} + \sum_{i \in G_q} d_{ij})] \\ &= \frac{1}{n_k n_r} [n_k n_p D(G_k,G_p) + n_k n_q D(G_k,G_q)] \\ &= \frac{n_p}{n_r} D(G_k,G_p) + \frac{n_q}{n_r} D(G_k,G_q), \end{aligned} \tag{8.2.6}$$

其中 $n_r = n_p + n_q$.

以上类与类之间的距离，不但适用于对样品的聚类问题，而且也适合于对变量的聚类问题，这只要将 d_{ij} 用变量间的相似系数 $c_{j\beta}$ 代替，相应的“距离”可称之为类与类之间的相似系数. 对上述三种类与类的“距离”的定义也可用相似系数的术语作类似解释. 由于用 d_{ij} 和用 $c_{j\beta}$ 在具体的聚类过程中并无本质差异，为方便计，以后均用 d_{ij} 表示两个样品间的距离或两变量间的相似系数，两类间的相似系数也统称为两类间的距离.

8.2.2 谱系聚类法

有了样品之间的距离（或变量间的相似系数）以及类与类之间的距离的定义后，便可进行谱系聚类，其基本步骤归纳如下：

1.n 个样品（或变量）一开始就作为 n 个类，计算两两之间的距离（或相似系数）构成一个对称矩阵 $\mathbf{D} = (d_{ij})_{n \times n}$，其对角线上的元素全为零（对相似系数矩阵，其对角线上元素全为 1 ）. 显然，此时有 $D(G_p,G_q) = d_{pq}$. 记 $\mathbf{D}_{(0)} \widehat{=} \mathbf{D}$.

2. 选择 $\mathbf{D}_{(0)}$ 中对角线元素以外的下三角部分（或上三角部分）中的最小元素（相似系数矩阵则选择对角线元素以外的最大者），设其为 $D(G_p,G_q)$,

则将 G_p 与 G_q 合并成一个新类 $G_r=\{G_p,G_q\}$. 在 $\mathbf{D}_{(0)}$ 中划去 G_p 与 G_q 所对应的两行与两列，并加入由新类 $G_r=\{G_p,G_q\}$ 与剩下的未聚合的各类之间的距离所组成的一行和一列，得到一个新的距离矩阵 $\mathbf{D}_{(1)}$,$\mathbf{D}_{(1)}$ 是一个 n-1 阶对称阵（若在 $\mathbf{D}_{(0)}$ 中最小元素不惟一，对其它的最小元素也作如上相同处理，每合并两类，矩阵 $\mathbf{D}_{(0)}$ 则降低一阶）.

3. 由 $\mathbf{D}_{(1)}$ 出发，重复步骤 2 得到对称矩阵 $\mathbf{D}_{(2)}$, 从 $\mathbf{D}_{(2)}$ 出发得到 $\mathbf{D}_{(3)}$, 依次类推，直到 n 个样品（或变量）聚为一个大类为止.

4. 在合并过程中记下两类合并时样品（或变量）的编号以及合并两类时的距离或相似系数的大ㄎ颐浅浦□□□剑并绘成聚类的谱系图，然后可根据实际问题的背景和要求选定相应的临界水平以确定类的个数.

在给定 $\mathbf{D}_{(0)}$ 的基础上，采用类与类间不同的距离定义便得到不同的谱系聚类方法. 相应于前述三种类与类间距离的谱系聚类法我们分别称为最短距离法，最长距离法和类平均法. 下面我们通过具体例子说明这三种聚类方法的具体应用.

例 8.2.1 为研究辽宁、浙江、河南、甘肃、青海 5 省份 1991 年城镇居民生活消费的分布规律，需要用调查资料对这 5 个省分类，变量名称及原始数据如表 8.2.1 所示.

表 8.2.1　1991 **年辽宁等** 5 **省城镇居民月均消费数据**（单位：元 / 人）

变量 省份	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
辽宁	7.90	39.77	8.49	12.94	19.27	11.05	2.04	13.29
浙江	7.68	50.37	11.35	13.30	19.25	14.59	2.75	14.87
河南	9.42	27.93	8.20	8.14	16.17	9.42	1.55	9.76
甘肃	9.16	27.98	9.01	9.32	15.99	9.10	1.82	11.35
青海	10.06	28.64	10.52	10.05	16.18	8.39	1.96	10.81

其中，

X_1：人均粮食支出，　　X_5：人均衣着商品支出，
X_2：人均副食支出，　　X_6：人均日用品支出，
X_3：人均烟，酒，茶支出，　　X_7：人均燃料支出，
X_4：人均其它副食支出，　　X_8：人均非商品支出.

将每个省份看成一个样品，并以 1,2,3,4,5 分别表示辽宁、浙江、河南、甘肃、青海 5 省份，计算两组间的欧氏距离 $d_{ij}(i,j=1,2,\cdots,5)$. 如

$$d_{12}=d_{21}=[(7.90-7.68)^2+(39.77-50.37)^2+\cdots+(13.29-14.87)^2]^{\frac{1}{2}}=1.67,$$

$$d_{23}=d_{32}=[(7.68-9.42)^2+(50.37-27.93)^2+\cdots+(14.87-9.76)^2]^{\frac{1}{2}}=24.63,$$

等等. 从而得距离矩阵 $\mathbf{D}_{(0)}$ 如下（由于对称，只写出对角线及下三角部分并

在行和列位置上标出相应的类):

$$\mathbf{D}_{(0)}=\begin{array}{c} \\ \{1\} \\ \{2\} \\ \{3\} \\ \{4\} \\ \{5\} \end{array}\begin{array}{c} \begin{array}{ccccc} \{1\} & \{2\} & \{3\} & \{4\} & \{5\} \end{array} \\ \begin{bmatrix} 0 & & & & \\ 11.67 & 0 & & & \\ 13.80 & 24.63 & 0 & & \\ 13.12 & 24.06 & 2.20 & 0 & \\ 12.80 & 23.54 & 3.51 & 2.21 & 0 \end{bmatrix} \end{array}$$

$\mathbf{D}_{(0)}$ 中各元素数值的大小，反映了 5 个省之间消费水平的接近程度. 例如，甘肃省与河南省之间的欧氏距离最ㄎ. 20 ），反映了这两个省份城镇居民消费水平最接近.

1. 最短距离法

对此例，采用最短距离法的聚类过程如下：

首先,将五个省各看成一类,即令 $G_i=\{i\},(i=1,2,3,4,5)$, 这时 $D(G_i,G_j)=d_{ij}(i,j=1,2,3,4,5)$. 从 $\mathbf{D}_{(0)}$ 中看到，其中最小的元素为 $D(\{4\},\{3\})=d_{43}=2.20$, 故将 G_3 和 G_4 在水平 2.20 上合并成一个新类 $G_6=\{3,4\}$, 然后利用递推式（ 8.2.2 ）计算 G_6 与 G_1,G_2,G_5 之间的最短距离. 由于

$$D(G_6,G_i)=D(\{3,4\},\{i\})=\min\{D(\{3\},\{i\}),D(\{4\},\{i\})\},i=1,2,5,$$

故由 $\mathbf{D}_{(0)}$ 中数据可得

$$D(\{3,4\},\{1\})=\min\{d_{31},d_{41}\}=\min\{13.80,13.12\}=13.12,$$

$$D(\{3,4\},\{2\})=\min\{d_{32},d_{42}\}=\min\{24.63,24.06\}=24.06,$$

$$D(\{3,4\},\{5\})=\min\{d_{35},d_{45}\}=\min\{3.51,2.21\}=2.21.$$

在 $\mathbf{D}_{(0)}$ 中划去 $\{3\},\{4\}$ 所对应的行和列，并加上新类 $\{3,4\}$ 到其它类距离作为新的一行一列，得到

$$\mathbf{D}_{(1)}=\begin{array}{c} \\ \{3,4\} \\ \{1\} \\ \{2\} \\ \{5\} \end{array}\begin{array}{c} \begin{array}{cccc} \{3,4\} & \{1\} & \{2\} & \{5\} \end{array} \\ \begin{bmatrix} 0 & & & \\ 13.12 & 0 & & \\ 24.06 & 11.67 & 0 & \\ 2.21 & 12.80 & 23.54 & 0 \end{bmatrix} \end{array}$$

从 $\mathbf{D}_{(1)}$ 可知， $G_6=\{3,4\}$ 到 $G_5=\{5\}$ 的距离最小，为 2.21, 因此在水平 2.21 上将 G_6 和 G_5 合并得到一新类 $G_7=\{3,4,5\}$, 再由

$$D(G_7,G_i)=D(\{D_6,G_5\},G_i)=\min\{D(G_6,G_i),D(G_5,G_i)\}$$

$$=\min\{D(\{3,4\},\{i\}),D(\{5\},\{i\})\},i=1,2,$$

可得

$$D(\{3,4,5\},\{1\}) = \min\{D(\{3,4\},\{1\}), D(\{5\},\{1\})\} = \min\{13.12, 12.80\} = 12.80,$$

$$D(\{3,4,5\},\{2\}) = \min\{D(\{3,4\},\{2\}), D(\{5\},\{2\})\} = \min\{24.06, 23.54\} = 23.54.$$

在 $\mathbf{D}_{(1)}$ 中划去 $G_6 = \{3,4\}$ 和 $G_5 = \{5\}$ 所对应的行和列，加上 $G_7 = \{3,4,5\}$ 的相应行列得到

$$\mathbf{D}_{(2)} = \begin{array}{c} \\ \{3,4,5\} \\ \{1\} \\ \{2\} \end{array} \begin{array}{c} \begin{array}{ccc} \{3,4,5\} & \{1\} & \{2\} \end{array} \\ \begin{bmatrix} 0 & & \\ 12.80 & 0 & \\ 23.54 & 11.67 & 0 \end{bmatrix} \end{array}$$

$\mathbf{D}_{(2)}$ 中最短距离为 $D(\{2\},\{1\}) = 11.67$, 故在距离水平 11.67 上合并 G_1 和 G_2 得新类 $G_8 = \{1,2\}$. 至此我们仅有两类 $G_7 = \{3,4,5\}$ 和 $G_8 = \{1,2\}$, 其间最短距离为

$$D(G_7, G_8) = \min\{D(G_7, G_1), D(G_7, G_2)\}$$

$$= \min\{D(\{3,4,5\},\{1\}), D(\{3,4,5\},\{2\})\}$$

$$= \min\{12.80, 23.54\} = 12.80,$$

从而得

$$\mathbf{D}_{(3)} = \begin{array}{c} \\ \{3,4,5\} \\ \{1,2\} \end{array} \begin{array}{c} \begin{array}{cc} \{3,4,5\} & \{1,2\} \end{array} \\ \begin{bmatrix} 0 & \\ 12.80 & 0 \end{bmatrix} \end{array}$$

最后在距离水平 12.80 上将 G_7 和 G_8 合为一个包含所有 5 个省份的大类.

综上所述，我们在距离为 2.20 的水平上首先合并样品 3 和 4, 得新类 $G_6 = \{3,4\}$; 然后, 更新距离矩阵后又在距离为 2.21 的水平上合并 G_5 与 G_6 得新类 $G_7 = \{3,4,5\}$; 在距离为 11.67 的水平上又合并 G_1 与 G_2 得 $G_8 = \{1,2\}$, 最后在距离为 12.8 的水平上将 G_7 与 G_8 合并，形成一个大类. 将上述聚类过程连同各类合并时的水平用图表示出来，便可更清楚地看到整个聚类过程及相应的水平. 此图（见图 8.2.1 ）称为谱系图.

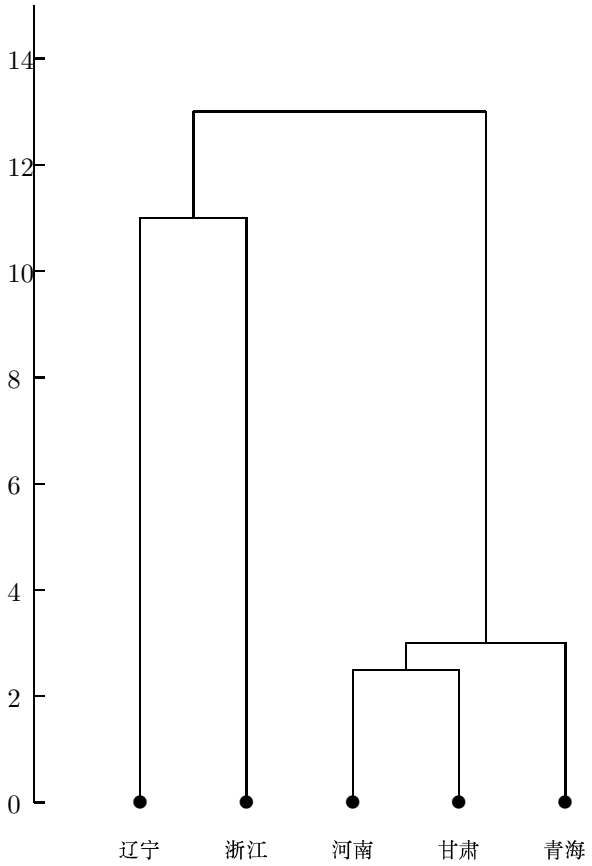

图 8.2.1　5 省份最小距离法谱系图

由此可见，将这 5 个省份分为两类比较合适，即河南、甘肃、青海为一类，辽宁和浙江为一类. 若想要类中的各个体更接近，可分为三类，即河南、甘肃、青海为一类，辽宁和浙江各自为一类.

2. 最长距离法

对本例采用最长距离法，其聚类过程如下：

(i) 先将 $\{3\},\{4\}$ 在距离水平 2.20 上合并得新类 $\{3,4\}$;

(ii) 求 $\{3,4\}$ 与其余各类 (即 $\{1\},\{2\},\{5\}$) 的最长距离

$$D(\{3,4\},\{1\}) = \max\{d_{31}, d_{41}\} = \max\{13.80, 13.12\} = 13.80,$$

$$D(\{3,4\},\{2\}) = \max\{d_{32}, d_{42}\} = \max\{24.63, 24.06\} = 24.63,$$

$$D(\{3,4\},\{5\}) = \max\{d_{35}, d_{45}\} = \max\{3.51, 2.21\} = 3.51.$$

更新后的距离矩阵为

$$\mathbf{D}_{(1)}=\begin{array}{c} \\ \{3,4\} \\ \{1\} \\ \{2\} \\ \{5\} \end{array}\begin{array}{c} \begin{array}{cccc} \{3,4\} & \{1\} & \{2\} & \{5\} \end{array} \\ \begin{bmatrix} 0 & & & \\ 13.80 & 0 & & \\ 24.63 & 11.67 & 0 & \\ 3.51 & 12.80 & 23.54 & 0 \end{bmatrix} \end{array}$$

(iii) 从 $\mathbf{D}_{(1)}$ 知，$D(\{3,4\},\{5\})=3.51$ 最小，在此距离水平上，将类 $\{3,4\}$ 与 $\{5\}$ 合并得新类 $\{3,4,5\}$. 根据递推公式（ 8.2.4 ），可求得 $\{3,4,5\}$ 到其它两类 $\{1\}$ 和 $\{2\}$ 的最长距离为

$$D(\{3,4,5\},\{1\})=\max\{D(\{3,4\},\{1\}),D(\{5\},\{1\})\}=\max\{13.80,12.80\}=13.80,$$

$$D(\{3,4,5\},\{2\})=\max\{D(\{3,4\},\{2\}),D(\{5\},\{2\})\}=\max\{24.63,23.54\}=23.63.$$

更新距离矩阵 $\mathbf{D}_{(1)}$ 得

$$\mathbf{D}_{(2)}=\begin{array}{c} \\ \{3,4,5\} \\ \{1\} \\ \{2\} \end{array}\begin{array}{c} \begin{array}{ccc} \{3,4,5\} & \{1\} & \{2\} \end{array} \\ \begin{bmatrix} 0 & & \\ 13.80 & 0 & \\ 24.63 & 11.67 & 0 \end{bmatrix} \end{array}$$

(iv) 由 $\mathbf{D}_{(2)}$ 知，在距离水平 11.67 上合并 {1},{2} 为一新类 $\{1,2\}$, 并由递推式（ 8.2.5 ）可求得 $\{1,2\}$ 到 $\{3,4,5\}$ 间的最长距离为

$$D(\{1,2\},\{3,4,5\})=\max\{D(\{1\},\{3,4,5\}),D(\{2\},\{3,4,5\})\}$$

$$=\max\{13.80,24.63\}=24.63.$$

更新的距离矩阵为

$$\mathbf{D}_{(3)}=\begin{array}{c} \\ \{3,4,5\} \\ \{1,2\} \end{array}\begin{array}{c} \begin{array}{cc} \{3,4,5\} & \{1,2\} \end{array} \\ \begin{bmatrix} 0 & \\ 24.63 & 0 \end{bmatrix} \end{array}$$

(v) 最后将 $\{1,2\}$ 与 $\{3,4,5\}$ 在距离水平 24.63 上合并为一个大类 $\{1,2,3,4,5\}$.

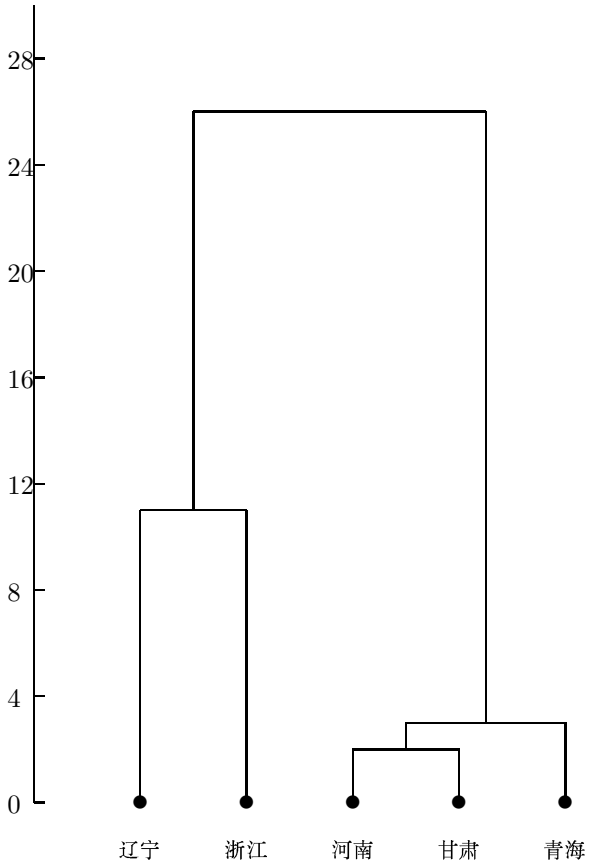

图 8.2.2　5 省份最长距离法谱系图

同样可画出最长距离法的谱系图如图 8.2.2. 我们看到，对此例所给数据而言，最长距离法的分类结果和最短距离法的分类结果相同，但各类合并时的距离水平随着类与类距离定义的不同而有所变化.

3. 类平均法

下面利用类平均法对本例作聚类分析

(i) 在距离水平 2.20 上合并 $\{3\},\{4\}$ 得新类 $\{3,4\}$;

(ii) 计算 $\{3,4\}$ 到其它类的类平均距离. 此时 $n_p = n_q = 1, n_r = 2$, 由 (8.2.6) 可得

$$D(\{3,4\},\{1\}) = \frac{1}{2}D(\{3\},\{1\}) + \frac{1}{2}D(\{4\},\{1\}) = \frac{1}{2}\times 13.80 + \frac{1}{2}\times 13.12 = 13.46,$$

$$D(\{3,4\},\{2\}) = \frac{1}{2}D(\{3\},\{2\}) + \frac{1}{2}D(\{4\},\{2\}) = \frac{1}{2}\times 24.63 + \frac{1}{2}\times 24.06 = 24.35,$$

$$D(\{3,4\},\{5\}) = \frac{1}{2}D(\{3\},\{5\}) + \frac{1}{2}D(\{4\},\{5\}) = \frac{1}{2}\times 3.51 + \frac{1}{2}\times 2.21 = 2.68.$$

从而更新的距离矩阵为

$$\mathbf{D}_{(1)}=\begin{array}{c} \\ \{3,4\} \\ \{1\} \\ \{2\} \\ \{5\} \end{array}\begin{array}{c} \begin{array}{cccc} \{3,4\} & \{1\} & \{2\} & \{5\} \end{array} \\ \begin{bmatrix} 0 & & & \\ 13.46 & 0 & & \\ 24.35 & 11.67 & 0 & \\ 2.86 & 12.80 & 23.54 & 0 \end{bmatrix} \end{array}$$

(iii) 由 $\mathbf{D}_{(1)}$ 知, 在距离水平 2.86 上应合并 $\{5\}$ 与 $\{3,4\}$ 为一新类 $\{3,4,5\}$, 利用递推式（ 8.2.6 ）可得新类 $\{3,4,5\}$ 到 $\{1\}$ 与 $\{2\}$ 的类平均距离分别为（这时, $n_p=2,n_q=1,n_r=3$ ）:

$$D(\{3,4,5\},\{1\})=\frac{2}{3}D(\{3,4\},\{1\})+\frac{1}{3}D(\{5\},\{1\})=\frac{2}{3}\times 13.46+\frac{1}{3}\times 12.80=13.24,$$

$$D(\{3,4,5\},\{2\})=\frac{2}{3}D(\{3,4\},\{2\})+\frac{1}{3}D(\{5\},\{2\})=\frac{2}{3}\times 24.35+\frac{1}{3}\times 23.54=24.08.$$

从而更新的距离矩阵为

$$\mathbf{D}_{(2)}=\begin{array}{c} \\ \{3,4,5\} \\ \{1\} \\ \{2\} \end{array}\begin{array}{c} \begin{array}{ccc} \{3,4,5\} & \{1\} & \{2\} \end{array} \\ \begin{bmatrix} 0 & & \\ 13.24 & 0 & \\ 24.08 & 11.67 & 0 \end{bmatrix} \end{array}$$

(iv) 从 $\mathbf{D}_{(2)}$ 可知, 在距离水平 11.67 上合并 $\{1\}$,$\{2\}$ 为新类 $\{1,2\}$, 由递推式(8.2.6)可得 $\{1,2\}$ 到 $\{3,4,5\}$ 的类平均距离为(此时 $n_p=n_q=1,n_r=2$)

$$\begin{aligned} &D(\{1,2\},\{3,4,5\})=\tfrac{1}{2}D(\{1\},\{3,4,5\})+\tfrac{1}{2}D(\{2\},\{3,4,5\}) \\ &=\tfrac{1}{2}\times 13.24+\tfrac{1}{2}\times 24.08=18.66. \end{aligned}$$

从而更新的距离矩阵为

$$\mathbf{D}_{(3)}=\begin{array}{c} \\ \{3,4,5\} \\ \{1,2\} \end{array}\begin{array}{c} \begin{array}{cc} \{3,4,5\} & \{1,2\} \end{array} \\ \begin{bmatrix} 0 & \\ 18.66 & 0 \end{bmatrix} \end{array}$$

(v) 最后在距离水平 18.66 上合并 $\{3,4,5\}$ 与 $\{1,2\}$ 成一个大类.

类平均法的谱系图如图 8.2.3 所示.

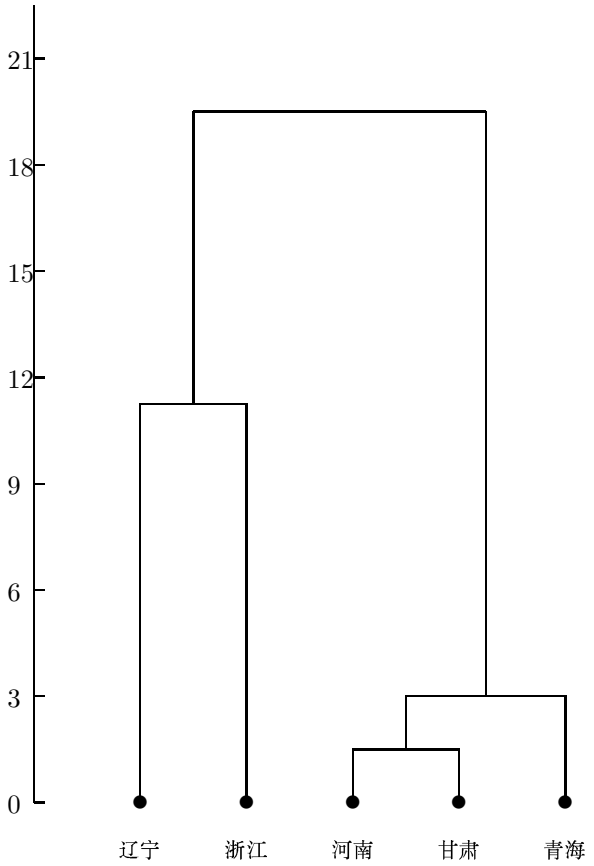

图 8.2.3　5 省份类平均法谱系图

综合以上分析, 我们看到对例 8.2.1 三种聚类方法的聚类结果是相同的, 只是不同的方法在各类合并室的距离水平有所不同. 但一般而言, 不同的聚类方法的聚类结果不尽相同, 下面我们以例 8.1.1 为例说明这一点.

例 8.2.2（续例 8.1.1 ）在例 8.1.1 关于欧洲 11 种语言的距离矩阵的基础上, 利用最短距离法、最长距离法和类平均法进行聚类.

由距离矩阵（见例 8.1.1 ）可知, 丹麦语和挪威语, 意大利语和法语, 意大利语和西班牙语之间的距离最　), 即在距离矩阵中,

$$d_{32} = d_{86} = d_{87} = 1,$$

对于各聚类方法来说, 我们均在距离水平为 1 的基础上首先合并 $\{3\}, \{2\}$（即丹麦语和挪威语）为一新类. 但由于 $d_{67} = 2$（即法语和西班牙语之间的距离为 2), 故在距离为 1 的水平上还不能将 $\{6\}, \{7\}, \{8\}$（即法语, 西班牙语和意大利语）合并为一类, 而只能合并 $\{8\}$ 和 $\{6\}$ 或者 $\{8\}$ 和 $\{7\}$ 为一类, 我们选择合并 $\{8\}$ 和 $\{6\}$. 接下来按不同的聚类方法作分析, 得到它们的聚类谱系图, 如图 8.2.4, 图 8.2.5, 图 8.2.6.

从最短距离法的谱系图上, 我们看到挪威语和丹麦语、法语和意大利语在最小距离水平 1 上合并为新类, 即这两对语言之间的相似性最大. 随着距

离水平的增加，英语并入到挪威语 – 丹麦语的类中，西班牙语并入到法语 – 意大利语的类中衫加和德语在距离为 5 的水平上合并成一类，在此水平上，波兰语又加入到德语 – 意大利语 – 西班牙语的类中．如此下来，直到所有 11 种语言归为一类.

将最长距离法的谱系图与最短距离法的谱系图作比较，我们看到，两者均在距离为 2 的水平上将英语、挪威语、丹麦语归入一类以及将法语、意大利语、西班牙语归入一类．波兰语仍然在距离为 5 的水平上并入到法语 – 意大利语 – 西班牙语的类中．两种聚类法对匈牙利语和芬兰语的分类方式相同，但这两种方法对德语和荷兰语的处理所不同，前者在距离为 5 的水平上将二者合并，且这两种语言保持在同一类中直到最后合并为一个大类．而最长距离法则在距离水平为 6 时，将德语并入英语 – 挪威语 – 丹麦语的类中，荷

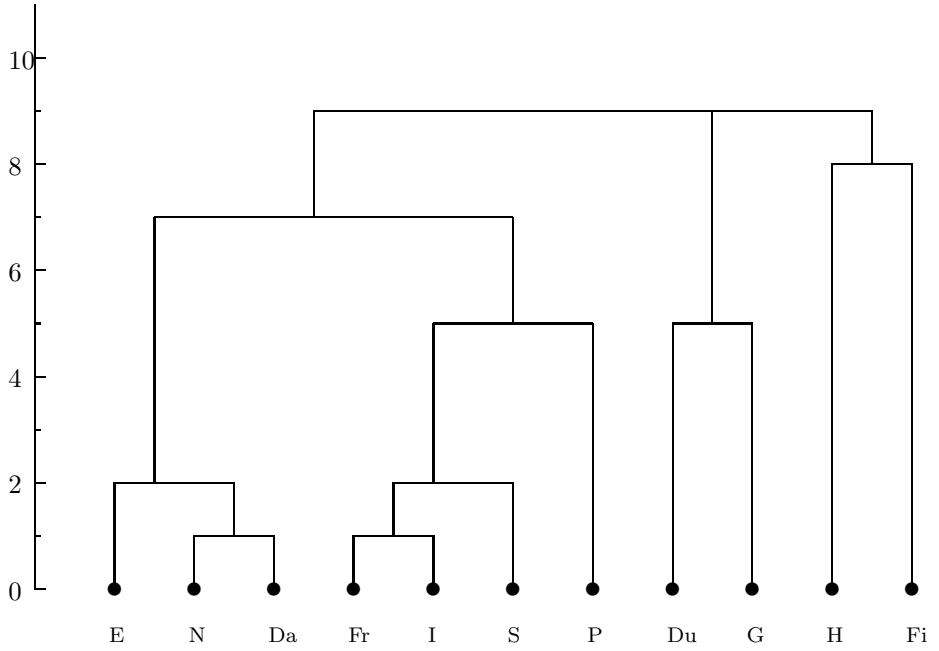

图 8.2.4　欧洲 11 种语言的最短距离法谱系图

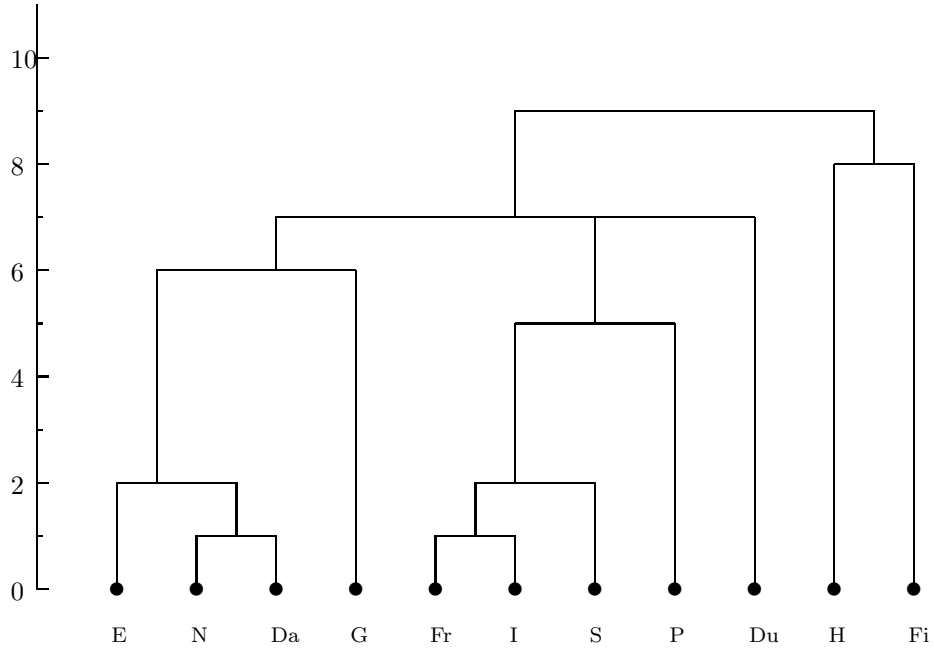

图 8.2.5　欧洲 11 种语言的最长距离法谱系图

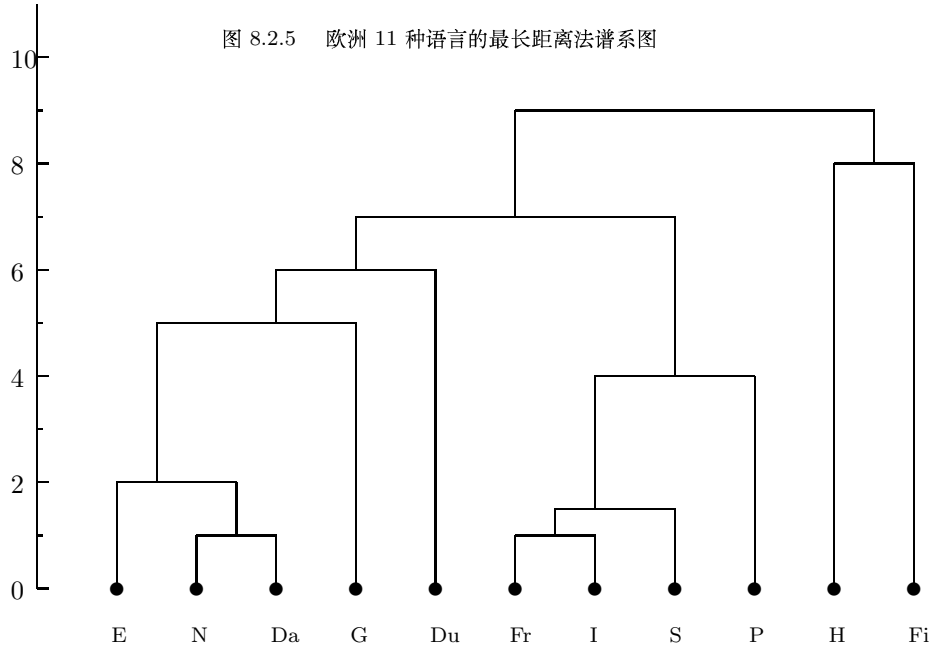

图 8.2.6　欧洲 11 种语言的类平均法谱系图

兰语则自成一类直到在距离水平为 7 时与英语 – 挪威语 – 丹麦语 – 德语类及法语 – 意大利语 – 西班牙语 – 波兰语的类合并. 在最终合并为一个大类之前，最短距离法有三类，而最长距离法只有两类.

相比之下，类平均法的谱系图和最长距离法的谱系图更相似.

由于各种聚类法所采用的类与类之间的距离的定义不同，产生不同的聚

类结果是很自然的. 对于一个具体问题, 比较好的作法是试探各种聚类方法, 同时, 对于一个给定的聚类法, 采用几种不同的样品间距离 (或变量间的相似系数) 进行聚类, 如果各种方法的聚类结果基本一致, 则认为其聚类结果是可信的. 另外, 一个经验的总结是最短距离法适用于样品散点图 (即将每个样品看成 p 维空间中的点所形成的图形) 是条形的, 甚至是 S 形的类, 而其它方法则更适合于椭球形的类.

前述的三种方法当然也适用于对变量的聚类, 下面给出一个对变量聚类的例子作为本书的结束.

例 8.2.3 (续例 6.1.4)利用从 1975 年 1 月至 1976 年 12 月对纽约证券交易所的 5 种股票 (Allied Chemical,du Pont,Union Carbide,EXXon 和 TeXaco) 的周反弹率的连续 100 周观测数据所求得的样本相关矩阵, 对这 5 种股票作聚类分析.

为方便计, 我们分别以 1,2,3,4,5 代表 Allied Chemical,du Pont,Union Carbide,EXXon 和 TeXaco 这 5 种股票 (即 5 个变量). 采用变量间的相关系数估计为相似性度量, 则由例 6.1.4 知, 这 5 个变量的样本相关矩阵 (即变量的相似系数矩阵) 为 (保留二位小数)

$$\mathbf{R}=(c_{ij})_{5\times5}=\begin{array}{c} \\ \{1\} \\ \{2\} \\ \{3\} \\ \{4\} \\ \{5\} \end{array}\begin{array}{c} \begin{array}{ccccc} \{1\} & \{2\} & \{3\} & \{4\} & \{5\} \end{array} \\ \begin{bmatrix} 1 & & & & \\ 0.58 & 1 & & & \\ 0.51 & 0.60 & 1 & & \\ 0.39 & 0.39 & 0.44 & 1 & \\ 0.46 & 0.32 & 0.43 & 0.52 & 1 \end{bmatrix} \end{array}$$

首先, 我们采用最短距离法进行聚类分析.

(i) 从 $\mathbf{R}$ 可知, $c_{32}=0.60$ 最大, 故在相似水平为 0.60 时, 将 $\{2\},\{3\}$ 两个变量合为一个新类 $\{2,3\}$, 根据递推公式 (8.2.2) 可求得 $\{2,3\}$ 到其它类的相似系数 (用 $C(\{2,3\},\{i\})$ 表示) 为

$$C(\{2,3\},\{1\})=\min\{C_{21},C_{31}\}=\min\{0.58,0.51\}=0.51,$$

$$C(\{2,3\},\{4\})=\min\{C_{42},C_{43}\}=\min\{0.39,0.44\}=0.39,$$

$$C(\{2,3\},\{5\})=\min\{C_{52},C_{53}\}=\min\{0.32,0.43\}=0.32.$$

更新相似矩阵 $\mathbf{R}$ 为

$$\mathbf{R}_{(1)}=\begin{array}{c} \\ \{2,3\} \\ \{1\} \\ \{4\} \\ \{5\} \end{array}\begin{array}{c} \begin{array}{cccc} \{2,3\} & \{1\} & \{4\} & \{5\} \end{array} \\ \begin{bmatrix} 1 & & & \\ 0.51 & 1 & & \\ 0.39 & 0.39 & 1 & \\ 0.32 & 0.46 & 0.52 & 1 \end{bmatrix} \end{array}$$

(ii) 由 $\mathbf{R}_{(1)}$ 可知， $\{4\}$ 和 $\{5\}$ 之间有最大的相似系数 0.52, 故在此相似水平上，将这两类合并为新类 $\{4,5\}$, 并计算它和其它类间的相似系数为

$$C(\{4,5\},\{2,3\}) = \min\{C(\{4\},\{2,3\}), C(\{5\},\{2,3\})\} = \min\{0.39, 0.32\} = 0.32,$$

$$C(\{4,5\},\{1\}) = \min\{C(\{4\},\{1\}), C(\{5\},\{1\})\} = \min\{0.39, 0.46\} = 0.39.$$

更新矩阵 $\mathbf{R}_{(1)}$ 得

$$\mathbf{R}_{(2)} = \begin{array}{c} \\ \{4,5\} \\ \{2,3\} \\ \{1\} \end{array} \begin{array}{c} \begin{array}{ccc} \{4,5\} & \{2,3\} & \{1\} \end{array} \\ \begin{bmatrix} 1 & & \\ 0.32 & 1 & \\ 0.39 & 0.51 & 1 \end{bmatrix} \end{array}$$

(iii) 由 $\mathbf{R}_{(2)}$ 可知，在相似水平 0.51 上合并 $\{1\}$ 与 $\{2,3\}$ 得新类 $\{1,2,3\}$, 并计算 $\{1,2,3\}$ 和 $\{4,5\}$ 的相似系数为

$$C(\{1,2,3\},\{4,5\}) = \min\{C(\{1\},\{4,5\}), C(\{2,3\},\{4,5\})\} = \min\{0.39, 0.32\} = 0.32.$$

更新 $\mathbf{R}_{(2)}$ 得

$$\mathbf{R}_{(3)} = \begin{array}{c} \\ \{1,2,3\} \\ \{4,5\} \end{array} \begin{array}{c} \begin{array}{cc} \{1,2,3\} & \{4,5\} \end{array} \\ \begin{bmatrix} 1 & \\ 0.32 & 1 \end{bmatrix} \end{array}$$

(iv) 最后，在相似水平 0.32 上将这 5 种股票合并为一个类.

最短距离法的谱系图如图 8.2.7 所示.

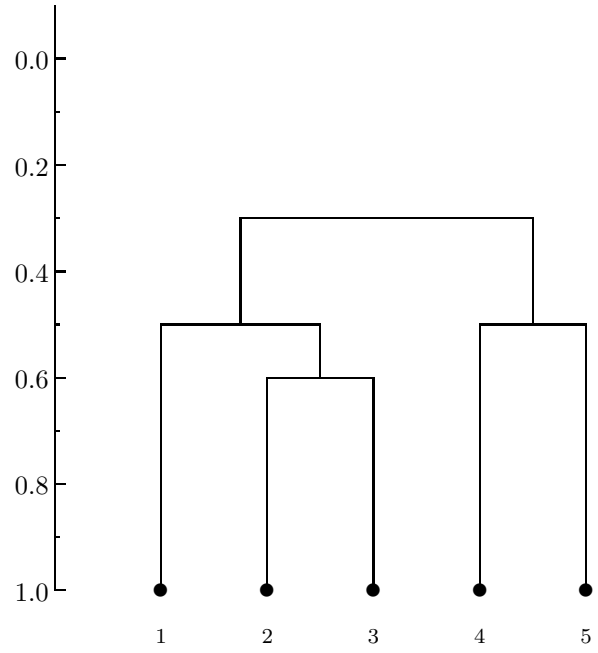

图 8.2.7　5 种股票最短距离法谱系图

若利用最长距离法，我们首先在相似系数为 0.60 在水平上合并 {2},{3} 为一新类 $\{2,3\}$; 在 0.58 水平上合并 $\{2,3\}$ 与 {1} 得 $\{1,2,3\}$; 在 0.52 水平上合并 {4},{5} 得 $\{4,5\}$; 最后在 0.46 的水平上合并为一个类.

其谱系图如图 8.2.8 所示.

利用类平均法，仍然在 0.60 的相似水平上得新类 $\{2,3\}$; 在 0.55 相似水平上合并 {1} 与 $\{2,3\}$; 在 0.52 的相似水平上合并 {4},{5} 得 $\{4,5\}$; 最后在 0.39 水平上合并为一个类.

其类平均法的谱系图如图 8.2.9 所示.

我们看到，各种方法对此例的聚类结果基本一致，在相似系数大约为 0.5 的水平上，可将这 5 种股票分为两类，即 Allied Chemical,du Pont 和 Union Carbide 为一类， EXXon 和 TeXaco 为一类. 事实上，前三者为化工股票，后二者为石油股票.

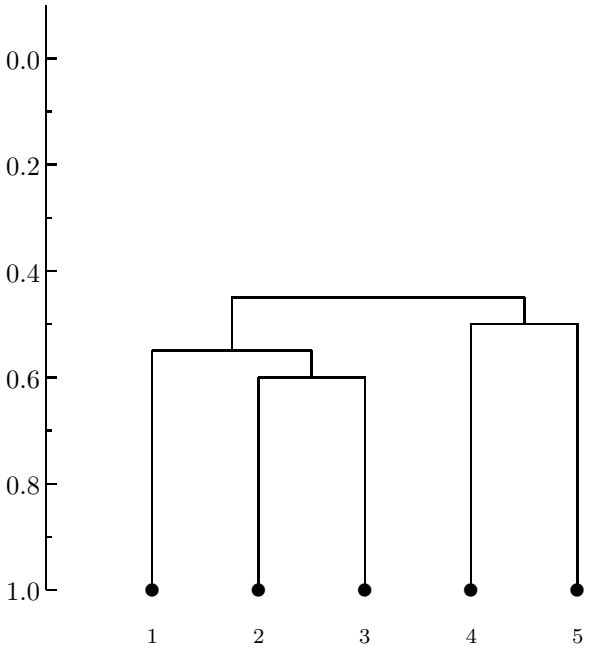

图 8.2.8　5 种股票最长距离法谱系图

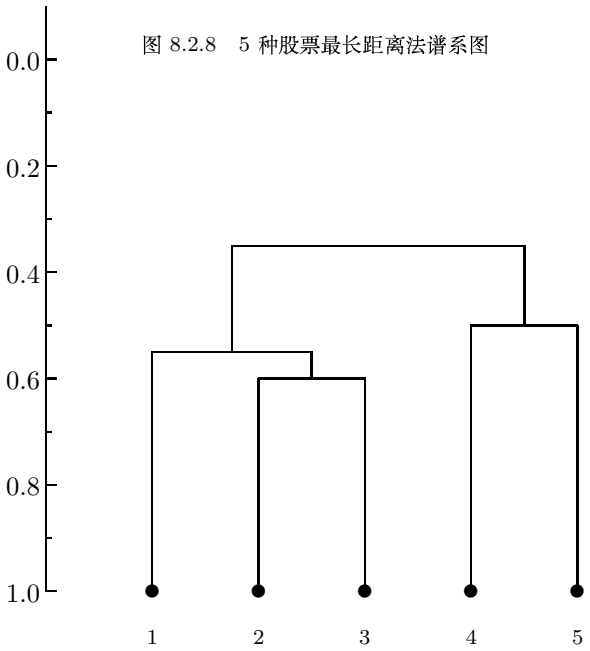

图 8.2.9　5 种股票的类平均法谱系图

附录　预备知识

多元统计分析是以数理统计的基本理论为基础，它的理论推导和具体方法的实现以矩阵代数为主要工具. 尽管读者已学过矩阵代数和数理统计的基本内容，但为了在以后的一些推导中用时更顺手些，本章主要对多元统计分析要用的矩阵代数和数理统计内容给予简单回顾.

§9.1　向量与矩阵

一、定义与基本运算

1. 有关定义. 由 $n \times p$ 个实数 $a_{ij}(i = 1, 2, \ldots, n; j = 1, 2, \ldots, p)$ 排成的一个矩阵数表，称为一个 n 行 p 列矩阵，并用 $\mathbf{A}$ 表示

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{np} \end{pmatrix}$$

有时也简记作 $\mathbf{A} = (a_{ij})_{n\times p}$, a_{ij} 称为矩阵 $\mathbf{A}$ 的元素. 当 $n = p$ 时，称 $\mathbf{A}$ 为方阵. 若 $p = 1$ 时，矩阵的形式为：

$$\begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{pmatrix}$$

我们称它为 n 维列向量. 当 $n = 1$ 时，矩阵的形式为 $(a_{11}, a_{12}, \cdots, a_{1p})$, 我们称它为 p 维行向量. 若 $\mathbf{A}$ 的元素全为零，则称 $\mathbf{A}$ 为零矩阵，记作 $\mathbf{A} = \mathbf{0}$. 若 $\mathbf{A}$ 为方阵, , 则称方阵中下标重复的元素 $a_{11}, a_{22}, \ldots, a_{nn}$ 为主对角元素，其余元素为非对角元素. 若方阵只有对角元素不为零，非对角元素全为零，则称 $\mathbf{A}$ 为对角阵，记作：

$$\mathbf{A} = diag\,(a_{11}, a_{22}, \cdots, a_{nn})$$

若对角矩阵其对角线上的元素全为 1, 则称 $\mathbf{A}$ 为 n 阶单位矩阵，记作 $\mathbf{A} = \mathbf{I}$.

如果将矩阵 $\mathbf{A}_{n\times p}$ 的行与列彼此交换，得到的新矩阵是一 p 行 n 列矩阵，记作

$$\mathbf{A}' = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \cdots & \cdots & \cdots & \cdots \\ a_{1p} & a_{2p} & \cdots & a_{np} \end{pmatrix}$$

称它为矩阵 $\mathbf{A}$ 的转置矩阵.

若 $\mathbf{A}$ 为方阵，且 $\mathbf{A}' = \mathbf{A}$, 则称 $\mathbf{A}$ 为对称阵.

若方阵 $\mathbf{A}$ 中，当 $i > j$ 时所有元素均为零，则称 $\mathbf{A}$ 为上三角阵；当 $i < j$ 时所有元素均为零，则称 $\mathbf{A}$ 为下三角阵.

2. 基本运算

(a) 两个 $n \times p$ 矩阵 $\mathbf{A} = (a_{ij})$, $\mathbf{B} = (b_{ij})$ 的和，定义为：

$$\mathbf{A} + \mathbf{B} = (a_{ij} + b_{ij})$$

(b) 实数 α 和 $n \times p$ 矩阵 $\mathbf{A}$ 的积记作 $\alpha\mathbf{A}$, 仍是 $n \times p$ 矩阵,

$$\alpha\mathbf{A} = (\alpha a_{ij})$$

(c) $n\times p$ 矩阵 $\mathbf{A}=(a_{ij})$ 和 $p\times r$ 矩阵 $\mathbf{B}=(b_{jk})$ 的积记作 $\mathbf{AB}$, 是 $n\times r$ 矩阵，它的第 (i,j) 元素为 $\sum\limits_{k=1}^{p}a_{ik}b_{kj}$, 即

$$\mathbf{AB}=\left(\sum_{k=1}^{p}a_{ik}b_{kj}\right)$$

对于矩阵的加法，数乘与乘的运算，容易验证：

对加法满足结合律和交换律

$$(\mathbf{A}+\mathbf{B})+\mathbf{C}=\mathbf{A}+(\mathbf{B}+\mathbf{C})$$

$$\mathbf{A}+\mathbf{B}=\mathbf{B}+\mathbf{A}$$

对乘法满足结合律

$(\alpha\beta)\,\mathbf{A}=\alpha\,(\beta\mathbf{A})$ ， $(\alpha\mathbf{A})\,\mathbf{B}=\alpha\,(\mathbf{AB})$

$(\mathbf{AB})\,\mathbf{C}=\mathbf{A}\,(\mathbf{BC})$

对乘法和加法满足分配律

$\alpha\,(\mathbf{A}+\mathbf{B})=\alpha\mathbf{A}+\alpha\mathbf{B}$, $(\alpha+\beta)\,\mathbf{A}=\alpha\mathbf{A}+\alpha\mathbf{B}$

$\mathbf{A}\,(\mathbf{B}+\mathbf{C})=\mathbf{AB}+\mathbf{BC}$, $(\mathbf{B}+\mathbf{C})\,\mathbf{A}=\mathbf{BA}+\mathbf{CA}$

此外，矩阵的转置运算还有如下关系式：

$(\mathbf{A}+\mathbf{B})'=\mathbf{A}'+\mathbf{B}'$, $(\alpha\mathbf{A})'=\alpha\mathbf{A}'$

$(\mathbf{AB})'=\mathbf{B}'\mathbf{A}'$

(d) 矩阵分块．在矩阵运算中，往往先将矩阵“分块”再进行运算，这样做特别是对高阶矩阵会起到简化运算的作用．例如，将两个 $n\times p$ 矩阵 $\mathbf{A}$ 、 $\mathbf{B}$ 分别分为四块：

$$\mathbf{A}=\begin{pmatrix}\mathbf{A}_{11}&\mathbf{A}_{12}\\ \mathbf{A}_{21}&\mathbf{A}_{22}\end{pmatrix},\ \mathbf{B}=\begin{pmatrix}\mathbf{B}_{11}&\mathbf{B}_{12}\\ \mathbf{B}_{21}&\mathbf{B}_{22}\end{pmatrix}$$

其中 $\mathbf{A}_{ij}$,$\mathbf{B}_{ij}$ 为 $n_i\times p_j$ 子矩阵, $i=1,2;j=1,2;n_1+n_2=n;p_1+p_2=p$, 则有

$$\mathbf{A}'=\begin{pmatrix}\mathbf{A}'_{11}&\mathbf{A}'_{21}\\ \mathbf{A}'_{12}&\mathbf{A}'_{22}\end{pmatrix},\ \mathbf{A}+\mathbf{B}=\begin{pmatrix}\mathbf{A}_{11}+\mathbf{B}_{11}&\mathbf{A}_{12}+\mathbf{B}_{12}\\ \mathbf{A}_{21}+\mathbf{B}_{21}&\mathbf{A}_{22}+\mathbf{B}_{22}\end{pmatrix}$$

$$\alpha\mathbf{A}=\begin{pmatrix}\alpha\mathbf{A}_{11}&\alpha\mathbf{A}_{12}\\ \alpha\mathbf{A}_{21}&\alpha\mathbf{A}_{22}\end{pmatrix}$$

又如，将 $n\times p$ 矩阵 $\mathbf{A}$ 分成如上四块，而将 $p\times r$ 矩阵 $\mathbf{B}$ 分成如下四块：

$$\mathbf{B}=\begin{pmatrix}\mathbf{B}_{11}&\mathbf{B}_{12}\\ \mathbf{B}_{21}&\mathbf{B}_{22}\end{pmatrix}$$

其中 $\mathbf{B}_{ij}$ 为 $p_i\times r_j$, $i=1,2;j=1,2;r_1+r_2=r$, 则容易验证有下列分块乘法规律：

$$\mathbf{AB}=\begin{pmatrix}\mathbf{A}_{11}\mathbf{B}_{11}+\mathbf{A}_{12}\mathbf{B}_{21}&\mathbf{A}_{11}\mathbf{B}_{12}+\mathbf{A}_{12}\mathbf{B}_{22}\\ \mathbf{A}_{21}\mathbf{B}_{11}+\mathbf{A}_{22}\mathbf{B}_{21}&\mathbf{A}_{21}\mathbf{B}_{12}+\mathbf{A}_{22}\mathbf{B}_{22}\end{pmatrix}$$

当然，还可以按以上原则将矩阵分成更多的块进行运算，行块和列块的数目也不必相同.

二、矩阵的逆和秩

1. 方阵的行列式. 由 n 阶方阵 $\mathbf{A}$ 中的元素组成的行列式，叫做方阵 $\mathbf{A}$ 的行列式，记为 $|\mathbf{A}|$ 或 $det\mathbf{A}$. 它有下面一些熟知的性质：
   (a) 若 $\mathbf{A}'$ 的某行 (或列) 元素全为零，则 $|\mathbf{A}| = \mathbf{0}$;
   (b) $|\mathbf{A}'| = |\mathbf{A}|$;
   (c) $|\alpha\mathbf{A}| = \alpha^n|\mathbf{A}|$;
   (d) 若 $\mathbf{A}$ 的两行 (或列) 成比例，则 $|\mathbf{A}| = \mathbf{0}$;
   (e) 若 $\mathbf{A}$ 的两行 (或列) 互换，所得矩阵之行列式等于 $-|\mathbf{A}|$;
   (f) 若将 $\mathbf{A}$ 的某一行 (列) 乘以一个常数加到另一行 (或列) 上，所得矩阵的行列式等于 $|\mathbf{A}|$.

   本书中还经常用到下列一些性质：
   (a) 若 $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_k$ 是 n 阶方阵，则

$$|\mathbf{A}_1\mathbf{A}_2\cdots\mathbf{A}_k| = |\mathbf{A}_1||\mathbf{A}_2|\cdots|\mathbf{A}_k|$$

   (b) 若分块矩阵 $\begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}$ 中， $\mathbf{A}_{12} = \mathbf{0}$ 或 $\mathbf{A}_{21} = \mathbf{0}$,$\mathbf{A}_{11}$ 为方阵，则

$$\begin{vmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{A}_{22} \end{vmatrix} = \begin{vmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{vmatrix} = |\mathbf{A}_{11}|\,|\mathbf{A}_{22}|$$

   (c) 若 $\mathbf{A}$ 和 $\mathbf{B}$ 分别为 $m \times n$ 和 $n \times m$ 阵，则

$$|\mathbf{I}_m + \mathbf{AB}| = |\mathbf{I}_n + \mathbf{BA}|$$

   (d) 若 $\mathbf{A}$ 为正交阵，则

$$|\mathbf{A}| = \pm 1$$

   (e) 若 $\mathbf{A}$ 为三角阵，则

$$|\mathbf{A}| = \prod_i a_{ii}$$

2. 逆矩阵. 设 $\mathbf{A}$ 为 n 阶方阵，如果有 n 阶方阵 $\mathbf{B}$, 使得

$$\mathbf{AB} = \mathbf{BA} = \mathbf{I}$$

   则称 $\mathbf{B}$ 是 $\mathbf{A}$ 的逆矩阵，记作 $\mathbf{A}^{-1}$. 逆矩阵有以下基本性质：
   (a) $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$
   (b) $(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$
   (c) 若方阵 $\mathbf{A}$ 和 $\mathbf{B}$ 均有逆矩阵存在，则

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$$

   (d) 设 $\mathbf{A}$ 为 n 阶可逆阵， $\mathbf{b}$ 和 $\mathbf{a}$ 为 n 维向量，则方程

$$\mathbf{Ab} = \mathbf{a}$$

的解为：

$$\mathbf{b}=\mathbf{A}^{-1}\mathbf{a}$$

(e) $\left|\mathbf{A}^{-1}\right|=|\mathbf{A}|^{-1}$

(f) 若 $\mathbf{A}$ 是正交矩阵，则

$$\mathbf{A}^{-1}=\mathbf{A}'$$

(g) 若 $\mathbf{A}$ 是对角阵，$\mathbf{A}=diag(a_{11},\ldots,a_{nn})$, 且 $a_{ii}\neq 0; i=1,2,\ldots,n$, 则 $\mathbf{A}^{-1}=diag(a_{11}^{-1},\ldots,a_{nn}^{-1})$;

(h) 设将可逆矩阵 $\mathbf{A}$ 分块为：

$\mathbf{A}=\begin{pmatrix}\mathbf{A}_{11} & \mathbf{A}_{12}\\ \mathbf{A}_{21} & \mathbf{A}_{22}\end{pmatrix}$, 其中 $\mathbf{A}_{11}$ 和 $\mathbf{A}_{22}$ 为方阵，若 $|\mathbf{A}_{11}|\neq 0$, 则

$$\mathbf{A}^{-1}=\begin{pmatrix}\mathbf{A}_{11}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}+\begin{pmatrix}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\\ -\mathbf{I}\end{pmatrix}\mathbf{B}^{-1}\left(\mathbf{A}_{21}\mathbf{A}_{11}^{-1},\,-\mathbf{I}\right)$$

其中 $\mathbf{B}=\mathbf{A}_{22}-\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$

若 $\mathbf{A}_{22}|\neq 0$, 则

$$\mathbf{A}^{-1}=\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{A}_{22}^{-1}\end{pmatrix}+\begin{pmatrix}-\mathbf{I}\\ \mathbf{A}_{22}^{-1}\mathbf{A}_{21}\end{pmatrix}\mathbf{D}^{-1}\left(-\mathbf{I},\,\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\right)$$

其中 $\mathbf{D}=\mathbf{A}_{11}-\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$

(i) 若 $|\mathbf{A}|\neq 0$,$|\mathbf{B}|\neq 0$, 则

$$\begin{pmatrix}\mathbf{A} & \mathbf{0}\\ \mathbf{C} & \mathbf{B}\end{pmatrix}^{-1}=\begin{pmatrix}\mathbf{A}^{-1} & \mathbf{0}\\ -\mathbf{B}^{-1}\mathbf{C}\mathbf{A}^{-1} & \mathbf{B}^{-1}\end{pmatrix}$$

$$\begin{pmatrix}\mathbf{A} & \mathbf{D}\\ \mathbf{0} & \mathbf{B}\end{pmatrix}^{-1}=\begin{pmatrix}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{D}\mathbf{B}^{-1}\\ \mathbf{0} & \mathbf{B}^{-1}\end{pmatrix}$$

(j) 设方阵 $\mathbf{A}$ 的行列式 $|\mathbf{A}|$ 分块为：

$$|\mathbf{A}|=\begin{vmatrix}\mathbf{A}_{11} & \mathbf{A}_{12}\\ \mathbf{A}_{21} & \mathbf{A}_{22}\end{vmatrix}$$

当 $|\mathbf{A}_{11}|\neq 0$ 时，则有 $|\mathbf{A}|=|\mathbf{A}_{11}|\cdot|\mathbf{A}_{22}-\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}|$

当 $|\mathbf{A}_{22}|\neq 0$ 时，则有 $|\mathbf{A}|=|\mathbf{A}_{22}|\cdot|\mathbf{A}_{11}-\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}|$

3. 矩阵的迹．设 $\mathbf{A}$ 为 n 阶方阵，$\mathbf{A}$ 的迹定义为它的对角线元素之和，记为 $tr\mathbf{A}$, 即

$$tr\mathbf{A}=\sum_{i=1}^{n}a_{ii}$$

对于方阵的迹显然有如下性质：

(a) $tr\mathbf{A}'=tr\mathbf{A}$

(b) $tr(\mathbf{A}+\mathbf{B})=tr\mathbf{A}+tr\mathbf{B}$
(c) $tr(\alpha\mathbf{A})=\alpha tr\mathbf{A}$
(d) $tr(\mathbf{AB})=tr(\mathbf{BA})$

4. 矩阵的秩. 设 $\mathbf{A}$ 为 $n\times p$ 矩阵, 若存在 $\mathbf{A}$ 的一个 r 阶子方阵的行列式不等于零, 而 $\mathbf{A}$ 的一切 $(r+1)$ 阶子方阵的行列式均为零, 则称 $\mathbf{A}$ 的秩为 r, 记作 $rank\mathbf{A}=r$. 矩阵的秩具有如下性质:
(a) $rank\mathbf{A}=0$, 当且仅当 $\mathbf{A}=\mathbf{0}$;
(b) 若 $\mathbf{A}$ 为 $n\times p$ 阵, 则 $0\le rank\mathbf{A}\le\min(n,p)$;
(c) $rank\mathbf{A}=rank\mathbf{A}'$;
(d) $rank\begin{pmatrix}\mathbf{A}&\mathbf{0}\\\mathbf{0}&\mathbf{B}\end{pmatrix}=rank\begin{pmatrix}\mathbf{0}&\mathbf{A}\\\mathbf{B}&\mathbf{0}\end{pmatrix}=rank\mathbf{A}+rank\mathbf{B}$;
(e) $rank(\mathbf{AB})\le\min(rank\mathbf{A},rank\mathbf{B})$;
(f) $rank(\mathbf{A}+\mathbf{B})\le rank\mathbf{A}+rank\mathbf{B}$;
(g) 若矩阵 $\mathbf{A}$ 和 $\mathbf{C}$ 为可逆阵, 则

$$rank(\mathbf{ABC})=rank\mathbf{B}$$

三、特征根与特征向量

1. 特征根与特征向量的概念及性质. 设 $\mathbf{A}$ 是 p 阶方阵, 方程

$$|\mathbf{A}-\lambda\mathbf{I}_p|=0$$

是 λ 的 p 次多项式, 它一定有 p 个根 (可以是复的, 也可能有重根), 则称这 p 个根为矩阵 $\mathbf{A}$ 的特征根. 若 λ_i 是 $\mathbf{A}$ 的特征根, 则 $|\mathbf{A}-\lambda_i\mathbf{I}_p|=0$, 这时一定存在一个非零的向量 $\mathbf{l}_i$, 使得

$$(\mathbf{A}-\lambda_i\mathbf{I}_p)\mathbf{l}_i=\mathbf{0}$$

或

$$\mathbf{A}\mathbf{l}_i=\lambda\mathbf{l}_i$$

则称 $\mathbf{l}_i$ 为 $\mathbf{A}$ 的对应 λ_i 的特征向量.

下面列举一些关于特征根和特征向量的性质:
(a) $\mathbf{A}$ 和 $\mathbf{A}'$ 有相同的特征根.
(b) 若 $\lambda_1,\lambda_2,\ldots,\lambda_p$ 为 $\mathbf{A}$ 的特征根, 则 $\mathbf{A}-k\mathbf{I}_p$ 的特征根为 $\lambda_1-k,\lambda_2-k,\ldots,\lambda_p-k$;$k\mathbf{A}$ 的特征根为 $k\lambda_1,k\lambda_2,\ldots,k\lambda_p$. 这里 k 为常数.
(c) 若 $\mathbf{A}=diag(a_{11},a_{22},\ldots,a_{pp})$, 则 $a_{11},a_{22},\ldots,a_{pp}$ 为 $\mathbf{A}$ 的 p 个特征根, 相应特征向量分别为 $\mathbf{e}_1=(1,0,\ldots,0)'$, $\mathbf{e}_2=(0,1,0,\ldots,0)',\ldots,\mathbf{e}_p=(0,\ldots,0,1)'$.
(d) 若乘积 $\mathbf{AB}$ 和 $\mathbf{BA}$ 有意义, 则 $\mathbf{AB}$ 和 $\mathbf{BA}$ 有相同的非零特征根.
(e) 若 $\lambda_1,\lambda_2,\ldots,\lambda_p$ 为 $\mathbf{A}$ 的特征根, $\mathbf{A}$ 可逆, 则 $\mathbf{A}^{-1}$ 的特征根为 $\lambda_1^{-1},\lambda_2^{-1},\ldots,\lambda_p^{-1}$.
(f) 若 $\mathbf{A}$ 是正交阵, 则 $|\lambda_i|=1$,$i=1,2,\ldots,p$.
(g) 若 $\mathbf{A}$ 为对称阵, 则 $\mathbf{A}$ 的特征根全为实数. 故可按大小次序排成 $\lambda_1\ge\lambda_2\ge\cdots\ge\lambda_p$.

(h) 若 $\mathbf{A}$ 为对称阵， λ_i,λ_j 是它的两个不相同的特征根，则相应的特征向量 $\mathbf{l}_i$ 和 $\mathbf{l}_j$ 正交. 若 λ_i 和 λ_j 相同，我们也可以选择使相应的 $\mathbf{l}_i$ 和 $\mathbf{l}_j$ 互相正交，这时 $\mathbf{A}$ 可表示为：

$$\mathbf{A}=\sum_{i=1}^{p}\lambda_i\mathbf{l}_i\mathbf{l}_i'$$

称它为 $\mathbf{A}$ 的谱分解.

(i) 若 $\mathbf{A}$ 是三角阵，则 $\mathbf{A}$ 的特征根正好是它的对角元素.

(j) $tr\mathbf{A}=\sum\limits_{i=1}^{p}\lambda_i$, $|\mathbf{A}|=\prod\limits_{i=1}^{p}\lambda_i$.

2. 特征根的极值性质. 特征根的极值性质在主成分分析的理论推导中很有用，下面我们讨论特征根的极值性质.

设 $\mathbf{A}$ 为 p 阶对称阵, 将 $\mathbf{A}$ 的特征根 $\lambda_1,\lambda_2,\ldots,\lambda_p$ 依大小顺序排列, 设 $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p$;$\mathbf{l}_1,\mathbf{l}_2,\ldots,\mathbf{l}_p$ 为相应的标准化特征向量，从谱分解式

$$\mathbf{A}=\sum_{i=1}^{p}\lambda_i\mathbf{l}_i\mathbf{l}_i'$$

$$\mathbf{I}=\sum_{i=1}^{p}\mathbf{l}_i\mathbf{l}_i'$$

可知，对任给 $\mathbf{x}$,$\mathbf{x}=\sum\limits_{i=1}^{p}a_i\mathbf{l}_i$, 有

$$\frac{\mathbf{x}'\mathbf{A}\mathbf{x}}{\mathbf{x}'\mathbf{x}}=\frac{\sum\limits_{i=1}^{p}\lambda_i a_i^2}{\sum\limits_{i=1}^{p}a_i^2}$$

式中， $\mathbf{x}'\mathbf{x}=\|\ \mathbf{x}\ \|^2$. 由于 $\mathbf{l}_1,\mathbf{l}_2,\ldots,\mathbf{l}_p$ 组成 p 维空间中的一组标准正交基,

$$\mathbf{x}'\mathbf{A}\mathbf{x}=\sum_{i=1}^{p}\lambda_i a_i^2\qquad \text{从而}$$

利用上面的等式可得到

$$\sup_{\mathbf{x}\neq\mathbf{0}}\mathbf{x}'\mathbf{A}\mathbf{x}/\mathbf{x}'\mathbf{x}=\sup_{\mathbf{a}\neq\mathbf{0}}\sum_{i=1}^{p}\lambda_i a_i^2/\mathbf{a}'\mathbf{a}=\lambda_1$$

$$\inf_{\mathbf{x}\neq\mathbf{0}}\frac{\mathbf{x}'\mathbf{A}\mathbf{x}}{\mathbf{x}'\mathbf{x}}=\inf_{\mathbf{a}\neq\mathbf{0}}\sum_{i=1}^{p}\lambda_i a_i^2/\mathbf{a}'\mathbf{a}=\lambda_p$$

仿照上述方法不难证明

$$\sup_{\mathbf{x}'\mathbf{l}_i=0} \mathbf{x}'\mathbf{A}\mathbf{x}/\mathbf{x}'\mathbf{x} = \lambda_{k+1}$$

$$i = 1, \ldots, k$$

$$\mathbf{x} \neq \mathbf{0}$$

若 $\mathbf{A}' = \mathbf{A}$,$\mathbf{B} > 0$,$\mu_1 \geq \mu_2 \geq \cdots \geq \mu_p$ 为 $\mathbf{A}$ 相对于 $\mathbf{B}$ 的特征根，类似可得

$$\sup_{\mathbf{x}\neq\mathbf{0}} \mathbf{x}'\mathbf{A}\mathbf{x}/\mathbf{x}'\mathbf{B}\mathbf{x} = \mu_1$$

$$\inf_{\mathbf{x}\neq\mathbf{0}} \mathbf{x}'\mathbf{A}\mathbf{x}/\mathbf{x}'\mathbf{B}\mathbf{x} = \mu_p$$

若 $\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_p$ 为对应于 $\mu_1, \mu_2, \ldots, \mu_p$ 的 $\mathbf{B}^{-\frac{1}{2}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}}$ 的特征向量，则

$$\sup_{\mathbf{x}'\mathbf{l}_i=0} \mathbf{x}'\mathbf{A}\mathbf{x}/\mathbf{x}'\mathbf{B}\mathbf{x} = \mu_{k+1}$$

$$i = 1, \ldots, k$$

$$\mathbf{x} \neq \mathbf{0}$$

四、正定阵、非负定阵和投影阵

1. 二次型的矩阵表示. p 个变量 $x_1, x_2, \ldots, x_p$ 的实二次型 $f(x_1, x_2, \ldots, x_p)$ 是指 $f(x_1, x_2, \ldots, x_p) = a_{11}x_1^2 + 2a_{12}x_1x_2 + \cdots + 2a_{1p}x_1x_p + a_{22}x_2^2 + 2a_{23}x_2x_3 + \cdots + 2a_{2p}x_2x_p + \cdots + a_{pp}x_p^2 = \sum_i^p \sum_j^p a_{ij}x_ix_j$, 其中 $a_{ij}(i, j = 1, 2, \ldots, p)$ 是给定的实常数，称为二次型的系数.

   利用矩阵乘法二次型可表示为矩阵形式

$$f(x_1, x_2, \ldots, x_p) = \sum_i^p \sum_i^p a_{ij}x_ix_j = \mathbf{x}'\mathbf{A}\mathbf{x}$$

其中 $\mathbf{x}' = (x_1, x_2, \ldots, x_p)$.

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{pmatrix} \text{ 为对称矩阵}$$

2. 正定阵、非负定阵. 设有实二次型 $f(x_1, x_2, \ldots, x_p) = \mathbf{x}'\mathbf{A}\mathbf{x}$, 如果对任何 $\mathbf{x} \neq \mathbf{0}$, 都有 $f(\mathbf{x}) > 0$, 则称 $f(\mathbf{x})$ 为正定二次型，并称对称矩阵 $\mathbf{A}$ 是正定的，记作 $\mathbf{A} > \mathbf{0}$; 如果对任何 $\mathbf{x} \neq \mathbf{0}$, 都有 $f(\mathbf{x}) < 0$, 则称 $f(\mathbf{x})$ 为负

定二次型，并称对称矩阵 $\mathbf{A}$ 是负定的，记作 $\mathbf{A}<\mathbf{0}$; 如果对任何 $\mathbf{x}$, 有 $f(\mathbf{x})\geq 0$, 则称 $\mathbf{A}$ 是非负定阵，记作 $\mathbf{A}\geq\mathbf{0}$. 下面列举正定阵和非负定阵的一些性质：

(a) 一个对称阵是正 (非负) 定的，当且仅当它的特征根为正 (非负).

(b) 若 $\mathbf{A}$ 为正定阵，则 $\mathbf{A}^{-1}$ 亦正定.

(c) 设 $\mathbf{A}$ 为 p 阶正定阵，$\mathbf{B}$ 是 $p\times g$ 阶矩阵 $(g\leq p)$, 且 $rank\mathbf{B}=g$, 则 $\mathbf{B}'\mathbf{A}\mathbf{B}$ 是正定的.

(d) 若 $\mathbf{A}$ 为正定阵，则 $C\mathbf{A}$ 亦为正定阵，其中 C 为正数.

(e) 若 $\mathbf{A}>\mathbf{0}$,$\mathbf{B}>\mathbf{0}$,$\mathbf{A}-\mathbf{B}>\mathbf{0}$, 则 $\mathbf{B}^{-1}-\mathbf{A}^{-1}>\mathbf{0}$, 且 $|\mathbf{A}|>|\mathbf{B}|$.

(f) 若 $\mathbf{A}>\mathbf{0}$, 将 $\mathbf{A}$ 分块为 $\mathbf{A}=\begin{pmatrix}\mathbf{A}_{11} & \mathbf{A}_{12}\\ \mathbf{A}_{21} & \mathbf{A}_{22}\end{pmatrix}$, 其中 $\mathbf{A}_{11}$ 为方阵，则 $\mathbf{A}_{11}>\mathbf{0}$,$\mathbf{A}_{22}>\mathbf{0}$.$\mathbf{A}_{11.2}=\mathbf{A}_{11}-\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}>\mathbf{0}$,$\mathbf{A}_{22.1}=\mathbf{A}_{22}-\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}>\mathbf{0}$.

(g) 若 $\mathbf{A}\geq\mathbf{0}$, 则必存在一个正交阵 $\mathbf{T}$, 使

$$\mathbf{T}'\mathbf{A}\mathbf{T}=diag(\lambda_1,\lambda_2,\ldots,\lambda_p)=\mathbf{\Lambda}$$

其中 $\lambda_1,\ldots,\lambda_p$ 为 $\mathbf{A}$ 的特征根，$\mathbf{T}$ 的列向量为相应的特征向量. 于是

$$\mathbf{A}=\mathbf{T}\mathbf{\Lambda}\mathbf{T}'$$

(h) 由性质 1,$\lambda_1,\lambda_2,\ldots,\lambda_p$ 均非负, 即 $\mathbf{\Lambda}\geq\mathbf{0}$. 记 $f(\mathbf{\Lambda})=diag\left(f(\lambda_1),\ldots,f(\lambda_p)\right)$,$f(\mathbf{A})=\mathbf{T}f(\mathbf{\Lambda})\mathbf{T}'$,

特别 $\mathbf{\Lambda}=diag\left(\lambda_1^{1/2},\ldots,\lambda_p^{1/2}\right)$,

$$\mathbf{A}^{\frac{1}{2}}=\mathbf{T}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{T}'$$

$\mathbf{A}^{\frac{1}{2}}$ 称为 $\mathbf{A}$ 的平方根，由于 $\mathbf{\Lambda}^{\frac{1}{2}}\geq\mathbf{0}$, 利用性质 3, 得 $\mathbf{A}^{\frac{1}{2}}\geq\mathbf{0}$. 综上所述，可得性质 9.

(i) 若 $\mathbf{A}\geq\mathbf{0}(>\mathbf{0})$, 则存在 $\mathbf{A}^{\frac{1}{2}}\geq\mathbf{0}(>\mathbf{0})$, 使得 $\mathbf{A}=\mathbf{A}^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}}$.

3. 投影阵. 设 $\mathbf{A}$ 为 p 阶方阵，若 $\mathbf{A}^2=\mathbf{A}$, 称 $\mathbf{A}$ 为幂等矩阵. 对称的幂等阵称为投影阵，投影阵具有如下性质：

(a) 若 $\mathbf{A}$ 是投影阵，则 $tr\mathbf{A}=rank\mathbf{A}$.

(b) 若 $\mathbf{A}$ 是投影阵，则 $\mathbf{I}-\mathbf{A}$ 也为投影阵.

(c) 若 $\mathbf{A}$ 是秩为 r 的投影阵，则 $\mathbf{A}$ 有 r 个特征根为 1, 其余为 0, 故满秩的投影阵必为单位阵.

(d) 若 $\mathbf{A}$ 和 $\mathbf{B}$ 均为投影阵，且 $\mathbf{A}+\mathbf{B}=\mathbf{I}$, 则 $\mathbf{A}\mathbf{B}=\mathbf{B}\mathbf{A}=\mathbf{0}$.

(e) 若 $\mathbf{X}$ 为 $n\times p$ 阵，$n\geq p$,$rank\mathbf{X}=p$, 则 $\mathbf{P}_x=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ 是投影阵，且 $rank(\mathbf{P}_x)=p$.

§9.2 矩阵的分解和微商

一、矩阵的分解

矩阵的分解在多元分析中十分有用. 下面列举出几个常用的结果.

1. 若 $\mathbf{A}$ 为 p 阶对称阵，则存在一个正交阵 $\mathbf{T}$ 使

$$\mathbf{T}'\mathbf{A}\mathbf{T}=diag(\lambda_1,\lambda_2,\ldots,\lambda_p)\equiv\mathbf{\Lambda}$$

从而

$$\mathbf{A}=\mathbf{T}\mathbf{\Lambda}\mathbf{T}'$$

其中 $\lambda_1,\lambda_2,\ldots,\lambda_p$ 为 $\mathbf{A}$ 的特征根，$\mathbf{T}$ 的列向量为 $\mathbf{A}$ 对应于 $\lambda_1,\lambda_2,\ldots,\lambda_p$ 的特征向量.

2. 若 $\mathbf{A}\geq\mathbf{0}$, 则存在 $\mathbf{A}^{\frac{1}{2}}\geq\mathbf{0}$, 使得

$$\mathbf{A}=\mathbf{A}^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}}$$

称 $\mathbf{A}^{\frac{1}{2}}$ 为 $\mathbf{A}$ 的平方根.

3. 若 $\mathbf{A}$ 为 n 阶正定阵，则存在唯一的其对角元素为正的上三角阵 $\mathbf{T}$ 使得

$$\mathbf{A}=\mathbf{T}'\mathbf{T}$$

称这一分解为乔列斯基 (cholesky) 分解. 由 $\mathbf{A}=\mathbf{T}'\mathbf{T}$ 可以求得 $\mathbf{T}$, 因为

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{1n} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$$

$$=\begin{pmatrix} t_{11} & 0 & \cdots & 0 \\ t_{21} & t_{22} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ t_{1n} & t_{2n} & \cdots & t_{nn} \end{pmatrix}\begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ 0 & t_{22} & \cdots & t_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & t_{nn} \end{pmatrix}$$

所以

$t_{11}=\sqrt{a_{11}}$

$t_{1j}=\frac{a_{1j}}{t_{11}}=\frac{a_{1j}}{\sqrt{a_{11}}}\,,\quad j=2,3,\ldots,n$

$t_{22}=\sqrt{a_{22}-t_{12}^2}=\sqrt{a_{22}-\frac{a_{12}^2}{a_{11}}}$

$t_{2j}=(a_{2j}-t_{12}t_{1j})/t_{22}\,,\ j=3,4,\ldots,n$

$t_{33}=\sqrt{a_{33}-t_{13}^2-t_{23}^2}$

$\cdots\cdots$

依次下去就可求得矩阵 $\mathbf{T}$ 的全部元素.

二、矩阵的微商

设 $\mathbf{X}=(x_1,x_2,\ldots,x_n)'$ 为实向量，$\mathbf{Y}=f(\mathbf{X})$ 为 $\mathbf{X}$ 的实函数，则 $f(\mathbf{x})$ 关于 $\mathbf{x}$ 的微商定义为：

$$\frac{\partial f(\mathbf{X})}{\partial\mathbf{X}}=\begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \\ \vdots \\ \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$$

例如由定义可得：

1. 若 $Y=\mathbf{x}'\mathbf{x}$, 则因 $\mathbf{Y}=x_1^2+x_2^2+\cdots+x_n^2$,

$$\frac{\partial \mathbf{Y}}{\partial x_j}=2x_j\ ,\ j=1,2,\ldots,n$$

从而

$$\frac{\partial(\mathbf{x}'\mathbf{x})}{\partial \mathbf{x}}=2\mathbf{x}$$

2. 若 $\mathbf{Y}=\mathbf{x}'\mathbf{A}\mathbf{x}$, 因 $Y=\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n}a_{ij}x_ix_j$, 则

$$\partial Y/\partial x_j=\sum_{i=1}^{n}\left(a_{ij}+a_{ji}\right)x_j\ ,\ j=1,2,\ldots,n$$

从而

$$\frac{\partial(\mathbf{x}'\mathbf{A}\mathbf{x})}{\partial \mathbf{x}}=\left(\mathbf{A}+\mathbf{A}'\right)\mathbf{x}$$

当 $\mathbf{A}$ 为对称阵时, $\mathbf{A}=\mathbf{A}'$, 则有

$$\frac{\partial(\mathbf{x}'\mathbf{A}\mathbf{x})}{\partial \mathbf{x}}=2\mathbf{A}\mathbf{x}$$

设 $\mathbf{X}=(x_{ij})_{n\times m}$,$\mathbf{Y}=f(\mathbf{X})$ 为矩阵 $\mathbf{X}$ 的实值函数, 则有

$$\frac{\partial \mathbf{Y}}{\partial \mathbf{X}}=\begin{pmatrix} \frac{\partial y}{\partial x_{11}} & \frac{\partial y}{\partial x_{12}} & \cdots & \frac{\partial y}{\partial x_{1m}} \\ \frac{\partial y}{\partial x_{21}} & \frac{\partial y}{\partial x_{22}} & \cdots & \frac{\partial y}{\partial x_{2m}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial y}{\partial x_{n1}} & \frac{\partial y}{\partial x_{n2}} & \cdots & \frac{\partial y}{\partial x_{nm}} \end{pmatrix}$$

如：若 $\mathbf{Y}=tr(\mathbf{X}'\mathbf{A}\mathbf{X})$, 式中 $\mathbf{X}'$ 为 $n\times m$ 阵, $\mathbf{A}$ 为 $m\times m$ 阵, 则

$$\frac{\partial tr\mathbf{X}'\mathbf{A}\mathbf{X}}{\partial \mathbf{X}}=\left(\mathbf{A}+\mathbf{A}'\right)\mathbf{X}$$

若 $\mathbf{A}$ 为对称阵, 即得

$$\frac{\partial tr\mathbf{X}'\mathbf{A}\mathbf{X}}{\partial \mathbf{X}}=2\mathbf{A}\mathbf{X}$$