第1章 矩阵运算与高等代数

● 常用结论

■ 若A,B是正交矩阵，则A',A^{-1},AB也是正交矩阵，$|A|=\pm1$

■ 若α是A属于λ_0的特征向量,则$A\alpha=\lambda_0\alpha$,且$k\alpha$也为特征向量

■ 相似矩阵特征多项式相同；AB与BA非零特征值及重数相同

■ 关于实对称矩阵A：特征值为实数；不同特征值的特征向量必正交；必正交相似于对角矩阵；必可对角化；对角线元素介于特征值之间；正定等价于特征值均大于0；正定等价于存在实可逆矩阵C的分解$A=C'C$；若正定则存在唯一正定矩阵C使得$A=C'C$；若正定则$|A|\leqslant a_{11}\cdots a_{nn}$

● 施密特三角化分解（QR分解）

设A为$n\times m$矩阵$(n\geqslant m)$，则存在分解$A=QR$,其中Q为$n\times m$列满秩矩阵，R为对角元素非负的上三角矩阵

● 施密特正交化

设$\alpha_1,\cdots,\alpha_s$线性无关，令$\beta_i=\alpha_i-\sum_{j=1}^{i-1}\frac{(\alpha_i,\beta_j)}{(\beta_j,\beta_j)}\beta_j$,则$\beta_1,\cdots,\beta_s$正交

● 矩阵拉直与Kronecker积

■ 拉直：设$\mathbf{A}=(\mathbf{a}_1,\cdots,\mathbf{a}_n)$是一个$m\times n$矩阵,则$\mathrm{Vec}(\mathbf{A})=\begin{pmatrix}\mathbf{a}_1\\ \vdots\\ \mathbf{a}_n\end{pmatrix}$

■ Kronecker积：设$\mathbf{A},\mathbf{B}$分别为$m\times n,p\times q$的矩阵，定义$mp\times nq$的矩阵$\mathbf{C}=\mathbf{A}\otimes\mathbf{B}=(a_{ij}\mathbf{B})$，称为矩阵$\mathbf{A},\mathbf{B}$的Kronecker积。

第2章 数据可视化与R语言

● 函数plot()对应不同数据类型时绘制的图形

数值/数值,数值——散点图；因子/一维频数分布表——条形图；
因子,因子-脊形图；二维列联表——马赛克图；
数值,因子——箱线图；数据框——散点图矩阵

● 数据可视化的图形

■ 轮廓图（平行坐标图、多线图）：可以比较各样本在多个变量上取值的异同；也可以直观地看出单个变量数值的分散程度等。

■ 雷达图（蜘蛛图）：以二维图表形式展示多变量数据。

■ 星图：利用n个p边形，可以比较n个样本的相似性。

■ 脸谱图：将p个维度的数据用人脸部位的形状或大小来表示。

■ 散点图：可以直观看出两个变量之间相关关系及相关的程度。

■ 气泡图：三维散点图的变化，气泡大小表示第三个变量的大小。

第3章 多元正态分布

● 联合分布函数

$$F(\boldsymbol{x})=F(x_1,\cdots,x_p)=\Pr\{X_1\leqslant x_1,\cdots,X_p\leqslant x_p\}$$

● 概率密度函数的性质

若$f(x_1,\cdots,x_p)$在$(x_1,\cdots,x_p)$处连续，有：$\frac{\partial^pF(x_1,\cdots,x_p)}{\partial x_1\cdots\partial x_p}=f(x_1,\cdots,x_p)$

● 边缘概率密度函数

若p维随机向量X作为整体，有联合分布函数$f(x_1,\cdots,x_p)$，其任意一个分量$\boldsymbol{X}^{(m)}=(X_1,\cdots,X_m)'$也是一个$m$维的随机向量（其中$1\leqslant m<p$），也有自己的联合分布函数，记为：

$$f_{X^{(m)}}(x_1,\cdots,x_m)=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}f(x_1,\cdots,x_p)\mathrm{d}x_{m+1}\cdots\mathrm{d}x_p$$

● 随机向量的相互独立及等价条件

■ 若对一切x和y有下式成立，则称X和Y相互独立：

$$\Pr(X\leqslant x,Y\leqslant y)=\Pr(X\leqslant x)\Pr(Y\leqslant y)$$

■ X和Y相互独立等价于：

$$F(x,y)=F_X(x)F_Y(y)\text{或}f(x,y)=f_X(x)f_Y(y)$$

● 随机向量的条件分布

给定$X_{m+1}=x_{m+1},\cdots,X_p=x_p$条件下，$(X_1,\cdots,X_m)'$的条件分布函数为：$\int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_m}\frac{f(u_1,\cdots,u_m,x_{m+1},\cdots,x_p)}{f_{X^{(-m)}}(x_{m+1},\cdots,x_p)}\mathrm{d}u_1\cdots\mathrm{d}u_m$

条件概率密度函数为：$\frac{f(x_1,\cdots,x_p)}{f_{X^{(-m)}}(x_{m+1},\cdots,x_p)}$

其中$f_{X^{(-m)}}(x_{m+1},\cdots,x_p)$为$X^{-m}=(X_{m+1},\cdots,X_p)'$的边缘密度。

● 协方差

■ 单个随机向量的协方差

设$X=(X_1,\cdots,X_p)'$是p维随机向量，则X的协方差矩阵为：

$$\mathrm{Cov}(\mathbf{X})=E[(X-E(X))(X-E(X))']$$

$$=\begin{pmatrix}\mathrm{Cov}(X_1,X_1)&\mathrm{Cov}(X_1,X_2)&\cdots&\mathrm{Cov}(X_1,X_p)\\ \mathrm{Cov}(X_2,X_1)&\mathrm{Cov}(X_2,X_2)&\cdots&\mathrm{Cov}(X_2,X_p)\\ \vdots&\vdots&&\vdots\\ \mathrm{Cov}(X_p,X_1)&\mathrm{Cov}(X_p,X_2)&\cdots&\mathrm{Cov}(X_p,X_p)\end{pmatrix}$$

记为$\Sigma=\mathrm{Cov}(\boldsymbol{X})=(\sigma_{ij})_{p\times p}$，其中$\sigma_{ij}=Cov(X_i,X_j)$

■ 两个随机向量的协方差

设$X=(X_1,\cdots,X_p)'$和$Y=(Y_1,\cdots,Y_q)'$分别是p维和q维的两个随机向量，则X和Y的协方差矩阵为：

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

$$=\begin{pmatrix}\mathrm{Cov}(X_1,Y_1)&\mathrm{Cov}(X_1,Y_2)&\cdots&\mathrm{Cov}(X_1,Y_q)\\ \mathrm{Cov}(X_2,Y_1)&\mathrm{Cov}(X_2,Y_2)&\cdots&\mathrm{Cov}(X_2,Y_q)\\ \vdots&\vdots&&\vdots\\ \mathrm{Cov}(X_p,Y_1)&\mathrm{Cov}(X_p,Y_2)&\cdots&\mathrm{Cov}(X_p,Y_q)\end{pmatrix}$$

● 相关系数矩阵

$R=(\rho_{ij})_{p\times p}$为$X$的相关系数矩阵，其中：

$$\rho_{ij}=\frac{\mathrm{Cov}(X_i,X_j)}{\sqrt{\mathrm{Var}(X_i)}\sqrt{\mathrm{Var}(X_j)}}=\frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}},\quad i,j=1,\cdots,p$$

其中$\sigma_{ii}=Var(X_i)$为随机变量X_i的方差。

● 均值向量与协方差矩阵的计算

$$E(\mathbf{AXB})=\mathbf{A}E(\mathbf{X})\mathbf{B},\ E(\mathbf{AX}+\mathbf{BY})=\mathbf{A}E(\mathbf{X})+\mathbf{B}E(\mathbf{Y})$$

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

$$\mathrm{Cov}(\mathbf{X},\mathbf{Y}+\mathbf{Z})=\mathrm{Cov}(\mathbf{X},\mathbf{Y})+\mathrm{Cov}(\mathbf{X},\mathbf{Z})$$

$$E(\mathbf{X}'\mathbf{AX})=\mathrm{tr}(\mathbf{A}\mathrm{Cov}(\mathbf{X}))+\mu'\mathbf{A}\mu$$

● 协方差矩阵的性质

■ 当$X=Y$和$p=q$时，$Cov(X,Y)=Cov(X)$

■ 当X和Y独立时，$Cov(X,Y)=\mathbf{0}_{p\times q}$

■ 若$Cov(X,Y)=\mathbf{0}_{p\times q}$，则称随机向量$X$和$Y$不相关

■ 协方差矩阵$\Sigma=Cov(X)$是对称非负定矩阵

■ 记$\mathbf{D}^{1/2}=\mathrm{diag}(\sqrt{\sigma_{11}},\cdots,\sqrt{\sigma_{pp}})$为标准差对角阵，则：

$$\Sigma=\mathbf{D}^{1/2}\mathbf{R}\mathbf{D}^{1/2},\quad \mathbf{R}=\mathbf{D}^{-1/2}\Sigma\mathbf{D}^{-1/2}$$

● 随机向量的变换

定义随机向量$X=(X_1,\cdots,X_p)'$的联合密度函数为$f(x_1,\cdots,x_p)$，p维空间到y空间的一一对应实值函数$y_i=y_i(x_1,\cdots,x_p)$，$i=1,\cdots,p$，随机向量$Y_1,\cdots,Y_p$为$Y_i=y_i(X_1,\cdots,X_p)$，存在逆变换$x_i=x_i(y_1,\cdots,y_p)$，有雅克比矩阵：

$$J(y_1,\cdots,y_p)=\mathrm{mod}\begin{vmatrix}\frac{\partial x_1}{\partial y_1}&\frac{\partial x_1}{\partial y_2}&\cdots&\frac{\partial x_1}{\partial y_p}\\ \vdots&\vdots&\ddots&\vdots\\ \frac{\partial x_p}{\partial y_1}&\frac{\partial x_p}{\partial y_2}&\cdots&\frac{\partial x_p}{\partial y_p}\end{vmatrix}$$

其中mod为绝对值，那么就得到$Y=(Y_1,\cdots,Y_p)'$的概率密度函数：

$$g(y_1,\cdots,y_p)=f[x_1(y_1,\cdots,y_p),\cdots,x_p(y_1,\cdots,y_p)]J(y_1,\cdots,y_p)$$

● 标准多元正态分布的性质

设$Y=(Y_1,\cdots,Y_q)'\sim N_p(0,\mathbf{I}_p)$，$a$为$p$元向量，$A$和$B$为对称矩阵，则有：$\mathrm{Cov}(\boldsymbol{a'Y},\boldsymbol{Y'}A\boldsymbol{Y})=0$，$\mathrm{Cov}(\mathbf{Y'AY},\mathbf{Y'BY})=2\mathrm{tr}(\mathbf{AB})$

● 多元正态分布的概率密度函数

若X是一个p维随机向量，Σ为p阶正定矩阵，则：

$$f(\boldsymbol{x})=\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(x-\boldsymbol{\mu})'\Sigma^{-1}(x-\boldsymbol{\mu})\right\}$$

● 多元正态分布的其他三个等价定义

■ 基于一元正态分布作线性变换：

设$Y_1,\cdots,Y_q$为独立同分布随机变量序列，$Y_i\sim N(0,1)$，$\mathbf{A}$是$p\times q$常数矩阵，$\boldsymbol{\mu}$是$p\times1$常数向量，则称Y的线性组合$X=\mathbf{A}Y+\boldsymbol{\mu}$为$p$元正态分布，且满足$X\sim N_p(\boldsymbol{\mu},\mathbf{AA'})$

■ 特征函数：若X特征函数满足下式，则X服从p元正态分布：

$$\varphi_X(\boldsymbol{t})=E[e^{\mathrm{i}\boldsymbol{t}X}]=\mathrm{e}^{\mathrm{i}\boldsymbol{t'\mu}-\frac{1}{2}\boldsymbol{t}'\Sigma\boldsymbol{t}}$$

■ 基于一元正态分布的线性组合：若$X=(X_1,\cdots,X_p)'$的任意线性组合均服从一元正态分布，则称X为p元正态分布。

● 多元正态分布的运算性质

■ 线性变换：设$\boldsymbol{X}\sim N_p(\boldsymbol{\mu},\Sigma)$，$\mathbf{Z}=\mathbf{BX}+\boldsymbol{\theta}$，则有：

$$\mathbf{Z}\sim N_q(\mathbf{B}\boldsymbol{\mu}+\boldsymbol{\theta},\mathbf{B}\Sigma\mathbf{B}')$$

■ 可加性：设$\mathbf{X}_1,\cdots,\mathbf{X}_k$是相互独立的$k$组$p$维随机向量，$\mathbf{X}_i\sim N_p(\mu_i,\Sigma_i)$，则有：$\sum_{i=1}^kX_i\sim N_p\left(\sum_{i=1}^k\mu_i,\sum_{i=1}^k\Sigma_i\right)$

■ 二次型：设$X\sim N_p(\mu,\Sigma)$，对于p阶矩阵A，二次型$(X-\mu)'A(X-\mu)$服从中心χ_m^2分布的充要条件为$A\Sigma A\Sigma=A\Sigma$，且其自由度为$\mathrm{tr}(A\Sigma)$

● 多元正态分布的边际分布

设$\boldsymbol{X}\sim N_p(\boldsymbol{\mu},\Sigma)$，$p\geqslant2$，将$\boldsymbol{X},\boldsymbol{\mu}$和$\Sigma$剖分为：

$$\boldsymbol{X}=\begin{bmatrix}\boldsymbol{X}^{(1)}\\ \boldsymbol{X}^{(2)}\end{bmatrix},\quad\boldsymbol{\mu}=\begin{bmatrix}\boldsymbol{\mu}^{(1)}\\ \boldsymbol{\mu}^{(2)}\end{bmatrix},\quad\Sigma=\begin{bmatrix}\Sigma_{11}&\Sigma_{12}\\ \Sigma_{21}&\Sigma_{22}\end{bmatrix}$$

则$\boldsymbol{X}^{(1)}\sim N_q(\boldsymbol{\mu}^{(1)},\Sigma_{11})$，$\boldsymbol{X}^{(2)}\sim N_{p-q}(\boldsymbol{\mu}^{(2)},\Sigma_{22})$

● 多元正态分布的条件分布

给定$\boldsymbol{X}^{(2)}=x^{(2)}$时，$\boldsymbol{X}^{(1)}$的条件分布服从$q$元正态分布，即：

$$(\boldsymbol{X}^{(1)}|\boldsymbol{X}^{(2)}=\boldsymbol{x}^{(2)})\sim N_q(\boldsymbol{\mu}_{1.2},\Sigma_{11.2})$$

其中$\boldsymbol{\mu}_{1.2}=\boldsymbol{\mu}^{(1)}+\Sigma_{12}\Sigma_{22}^{-1}(x^{(2)}-\boldsymbol{\mu}^{(2)})$，$\Sigma_{11.2}=\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$

● 多元正态分布的边际、条件分布的独立性

■ $\Sigma_{11}\geq\Sigma_{11.2}$，当且仅当$\Sigma_{12}=\Sigma_{21}'=0$时等号成立，此时$\boldsymbol{X}^{(1)}$与$\boldsymbol{X}^{(2)}$相互独立。

■ $\boldsymbol{X}^{(1)}$与$\boldsymbol{X}^{(2)}-\Sigma_{21}\Sigma^{-1}\boldsymbol{X}^{(1)}$相互独立。

■ 若$\boldsymbol{X}^{(1)}$与$\boldsymbol{X}^{(2)}$相互独立，则存在q维实数向量$\boldsymbol{a}$和$p-q$维实数向量$\boldsymbol{b}$，满足$\xi=\boldsymbol{a}'\mathbf{X}^{(1)}$和$\eta=\boldsymbol{b}'\mathbf{X}^{(2)}$独立。

● 矩阵正态分布

设$\mathbf{X}$为$n\times p$随机矩阵，若$\mathrm{Vec}(\mathbf{X}')\sim N_{np}(1_n\otimes\boldsymbol{\mu},\mathbf{I}_n\otimes\Sigma)$，则称$\mathbf{X}$服从矩阵正态分布,记作$\mathbf{X}\sim N_{n\times p}(\mathbf{M},\mathbf{I}_n\otimes\Sigma)$,其中$\mathbf{M}=\underbrace{[\boldsymbol{\mu},\cdots,\boldsymbol{\mu}]}_{n个}{}^T$

令$\mathbf{Z}=\mathbf{AXB}+\mathbf{D}$，则有$\mathbf{Z}\sim N_{k\times q}(\mathbf{AMB}+\mathbf{D},(\mathbf{AA}')\otimes(\mathbf{B}'\Sigma\mathbf{B}))$

第4章 多元正态总体的抽样分布

● 二次型分布（卡方分布）

■ 定义：设$X_i\sim N_1(\mu_i,\sigma^2)(i=1,\cdots,n)$的随机变量相互独立，令$\boldsymbol{X}=(X_1,\cdots,X_n)'$，$Y=X'X=\sum_{i=1}^nX_i^2$，且$\boldsymbol{X}\sim N_n(\mu,\sigma^2\mathbf{I}_n)$

则$Y/\sigma^2\sim\chi_n^2(\delta)$，其中非中心参数$\delta=\boldsymbol{\mu}'\boldsymbol{\mu}/\sigma^2$，$Y$表示$\boldsymbol{X}$的二次型。

■ 性质4.1.1、4.1.2、4.1.3：设$\mathbf{X}\sim N_n(\boldsymbol{\mu},\sigma^2\mathbf{I}_n)$，$\mathbf{A}=\mathbf{A}'$，则有：

$$\frac{1}{\sigma^2}\mathbf{X'AX}\sim\chi_r^2(\delta)\Leftrightarrow\mathbf{A}=\mathbf{A}^2,$$

其中$\delta=\frac{1}{\sigma^2}\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}$，$\boldsymbol{\mu}=(\mu_1,\cdots,\mu_n)'$，$\mathrm{rank}(\mathbf{A})=r(r\leqslant n)$

■ 性质4.1.4、4.1.5：$\mathbf{A},\mathbf{B}$为n阶对称矩阵，$\mathbf{B_0}$为$m\times n$矩阵，则

$$\mathbf{B_0A}=\mathbf{0}_{m\times n}\Rightarrow\mathbf{B}\boldsymbol{X}\text{与}'\boldsymbol{X}\mathbf{A}\boldsymbol{X}\text{独立}$$

$$\mathbf{AB}=\mathbf{0}_{n\times n}\Longleftrightarrow\boldsymbol{X}'\mathbf{A}\boldsymbol{X}\text{与}\boldsymbol{X}'\mathbf{B}\boldsymbol{X}\text{独立}$$

■ 性质4.1.6、4.1.7：设$\boldsymbol{X}\sim N_p(\boldsymbol{\mu},\Sigma),\Sigma>0$,$\mathbf{A}=\mathbf{A}'$且秩为$r<n$

则有：

$$(\boldsymbol{X}-\boldsymbol{\mu})'\mathbf{A}(\boldsymbol{X}-\boldsymbol{\mu})\sim\chi_r^2\Leftrightarrow\mathbf{\Sigma A\Sigma A\Sigma}=\mathbf{\Sigma A\Sigma}$$

$$\boldsymbol{X}'\mathbf{\Sigma}^{-1}\boldsymbol{X}\sim\chi_p^2(\delta),\ \delta=\boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu}$$

■ 性质4.1.8：$\mathbf{A},\mathbf{B}$为p阶对称矩阵，则有：

$$\mathbf{A\Sigma B}=\mathbf{0}_{p\times p}\Leftrightarrow(\boldsymbol{X}-\boldsymbol{\mu})'\mathbf{A}(\boldsymbol{X}-\boldsymbol{\mu})\text{与}(\boldsymbol{X}-\boldsymbol{\mu})'\mathbf{B}(\boldsymbol{X}-\boldsymbol{\mu})\text{独立}$$

● Wishart分布

■ 定义：设$\boldsymbol{X}_i\sim N_p(\mathbf{0},\mathbf{\Sigma})(i=1,\cdots,n)$的随机变量相互独立，记$\mathbf{X}=(\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n)'$，则称$p$阶矩阵$\mathbf{W}=\mathbf{X}'\mathbf{X}=\sum_{i=1}^n\mathbf{X}_i\mathbf{X}_i'\sim W_p(n,\mathbf{\Sigma})$，即$\mathbf{W}$服从Wishart分布，其中$n$称为自由度。

■ 均值、变换、可加性：若$\mathbf{W}\sim W_p(n,\Sigma)$ $\mathbf{W}_i\sim W_p(n_i,\Sigma)$，则

$$E(\mathbf{W})=n\mathbf{\Sigma},\ \mathbf{CWC'}\sim W_k(n,\mathbf{C\Sigma C'}),\ \sum_{i=1}^k\mathbf{W}_i\sim W_p\left(\sum_{i=1}^kn_i,\mathbf{\Sigma}\right)$$

■ 矩阵二次型：若$\mathbf{X}$为随机样本矩阵，$\mathbf{A}$为幂等矩阵，则矩阵二次型$\mathbf{Q}=\mathbf{X'AX}$服从Wishart分布$W_p(m,\mathbf{\Sigma})$，其中$m=\mathrm{tr}(\mathbf{A})$；$\mathbf{P'X}$与$\mathbf{Q}$独立的充分必要条件是$\mathbf{AP}=0$.

■ 行列式、逆矩阵期望、逆矩阵性质：设$\mathbf{W}\sim W_p(n,\mathbf{\Sigma}),\mathbf{\Sigma}>0$，$n\geqslant p$，则$|\mathbf{W}|\sim|\mathbf{\Sigma}|\gamma_1\gamma_2\cdots\gamma_p$，$E(\mathbf{W}^{-1})=\frac{1}{n-p-1}\mathbf{\Sigma}^{-1}$，其中$\gamma_i$独立,且$\gamma_i\sim\chi_{n-i+1}^2$;对于任意非零$p$维向量$\boldsymbol{a}$,总有$\frac{\boldsymbol{a}'\Sigma^{-1}\boldsymbol{a}}{\boldsymbol{a}'\mathrm{W}^{-1}\boldsymbol{a}}\sim\chi_{n-p+1}^2$

■ Bartlett分解：设$\mathbf{W}\sim W_p(n,\mathbf{I}_p)$,$n\geqslant p$，作分解$\mathbf{W}=\mathbf{TT}'$，其中$\mathbf{T}$为正对角线元素的下三角矩阵。令$\mathbf{T}=(t_{ij})_{p\times p}$，则$t_{ij}$相互独立，且$i>j$时$t_{ij}\sim N(0,1)$；$i=j$时$t_{ij}^2\sim\chi_{n-p+1}^2$.

● Hotelling T²分布

■ t分布：设$X\sim N(0,1)$,$Y\sim\chi_n^2$且X,Y独立，则$t=\frac{X}{\sqrt{Y/n}}\sim t_n$

■ Hotelling T²分布：设$\boldsymbol{X}\sim N_p(0,\mathbf{\Sigma})$,$\boldsymbol{W}\sim W_p(n,\mathbf{\Sigma})$,$\mathbf{\Sigma}>0$,$n\geqslant p$且$\boldsymbol{X},\boldsymbol{W}$相互独立，则$T^2=n\boldsymbol{X}'\mathbf{W}^{-1}\mathbf{X}\sim T^2(p,n)$，称为服从自由度为$n$的Hotelling T²分布，该分布只与$n,p$有关，与$\mathbf{\Sigma}$无关。

■ 性质：设$\bar{\boldsymbol{x}},\mathbf{S}$分别是正态总体$N_p(\boldsymbol{\mu},\mathbf{\Sigma})$的样本均值向量和样本协方差矩阵，则$T^2=n(\bar{\boldsymbol{x}}-\boldsymbol{\mu})'\mathbf{S}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})\sim T^2(p,n-1)$

第5章 多元正态分布的参数估计

● 简单随机样本矩阵的统计量

■ 样本均值向量：$\bar{\mathbf{x}}=\frac{1}{n}\sum_{i=1}^n\mathbf{x}_i=\frac{1}{n}\mathbf{X}'\mathbf{1}_n=(\bar{x}_1,\cdots,\bar{x}_p)'$

■ 样本离差阵：

$$\mathbf{V}=\sum_{i=1}^n(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})'=\mathbf{X'X}-n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}'=\mathbf{X}'\left[\mathbf{I}_n-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\right]\mathbf{X}$$

■ 样本协方差阵：$\mathbf{S}=\frac{1}{n-1}\mathbf{V}=(s_{ij})_{p\times p}$

■ 样本相关阵：$\widetilde{\mathbf{R}}=(r_{ij})_{p\times p}$，$r_{ij}=\frac{s_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}}$，$i,j=1,\cdots,p.$

● 样本均值向量与样本离差阵抽样分布的性质

(1) $\bar{\mathbf{x}}\sim N_p(\boldsymbol{\mu},\Sigma/n)$；(2) $\mathbf{V}\sim W_p(n-1,\Sigma)$；(3) $\bar{\boldsymbol{x}},\mathbf{V}$相互独立

● 统计量的性质

■ 无偏性：$E\left[\hat{\boldsymbol{\theta}}(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n)\right]=\boldsymbol{\theta}$

■ 充分性，相合性，完备性，有效性

第9章 主成分分析

● （总体）主成分分析的基本模型

设$X=(X_1,\cdots,X_p)'$，作线性变换得到$Z=(Z_1,\cdots,Z_p)'$：

其中$Z_i=a_i'\mathbf{X}=a_{i1}X_1+a_{i2}X_2+\cdots+a_{ip}X_p$，$i=1,2,\cdots,p$

● 目标函数（信息最大化的准则）

对于第i个主成分，限制条件下：

①$\boldsymbol{a}_i'\boldsymbol{a}_i=1$ ②$\boldsymbol{a}_i'\Sigma\boldsymbol{a}_j=0$，$j=1,\cdots,i-1$

令$\mathrm{Var}(Z_i)=\mathrm{Var}(\boldsymbol{a}'\boldsymbol{X})$最大化。（此时$\mathrm{Var}(Z_i)=\lambda_i$）

● 主成分的计算和性质

令$\mathbf{\Sigma}$是p维随机向量$X=(X_1,\cdots,X_p)'$的协方差矩阵，且有特征值和单位特征向量序列$(\lambda_1,\boldsymbol{a}_1),(\lambda_2,\boldsymbol{a}_2),\cdots,(\lambda_p,\boldsymbol{a}_p)$，其中特征值从大到小排列。令$X$的第$i$个主成分为$Z_i=a_i'\mathbf{X}$，则：

$$\sum_{i=1}^p\sigma_{ii}=\sum_{i=1}^p\mathrm{Var}(X_i)=\sum_{i=1}^p\lambda_i=\sum_{i=1}^p\mathrm{Var}(Z_i)=\mathrm{tr}(\mathbf{\Sigma})$$

● 贡献率和累计贡献率

■ 贡献率：第i个主成分方差在总方差中的占比：$\frac{\lambda_i}{\lambda_1+\lambda_2+\cdots+\lambda_p}$

■ 累计贡献率：前m个主成分方差在总方差中所占的比例。

● 因子载荷量与相关系数的平方和

■ 因子载荷量：即主成分Z_i与原变量X_k的相关系数

$$\rho(Z_i,X_k)=\frac{a_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}},\quad i,k=1,\cdots,p$$

■ 相关系数的平方和：将因子载荷量关于Z_i求平方和，则有：

$$\sum_{i=1}^p\rho^2(Z_i,X_k)=\sum_{i=1}^p\frac{a_{ik}^2\lambda_i}{\sigma_{kk}}=1,\quad k=1,\cdots,p.$$

若只对前m个主成分求平方和，则定义$v_k^{(m)}$为前m个主成分$Z_1,\cdots,Z_m$对原变量X_k的贡献率：$v_k^{(m)}=\sum_{i=1}^m\rho^2(Z_i,X_k)$

● 基于标准化的（总体）主成分分析

对总体进行标准化：$X_i^*=\dfrac{X_i-E(X_i)}{\sqrt{\mathrm{Var}(X_i)}}$后，再进行主成分分析，等价于直接对相关系数矩阵$\mathbf{R}$进行谱分解得到$(\lambda_i^*,\boldsymbol{a}_i^*)$后，基于$X_i^*$的主成分为：$Z_i^*=\boldsymbol{a}_i^{*\prime}\boldsymbol{X}^*=\boldsymbol{a}_i^{*\prime}\boldsymbol{D}^{-1/2}(\boldsymbol{X}-\boldsymbol{\mu})$，$i=1,\cdots,p$，其中$\mathbf{D}^{1/2}=\mathrm{diag}\left(\sqrt{\sigma_{11}},\cdots,\sqrt{\sigma_{pp}}\right)$为标准差对角阵。

此时有：$\sum_{i=1}^{p}\mathrm{Var}\left(Z_i^*\right)=\sum_{i=1}^{p}\mathrm{Var}\left(X_i^*\right)=\sum_{i=1}^{p}\lambda_i^*=p$

	Z_1^*	$\cdots$	Z_i^*	$\cdots$	Z_p^*	$\sum_{i=1}^{p}\rho_{ik}^2$
X_1^*	$a_{11}^*\sqrt{\lambda_1^*}$	$\cdots$	$a_{i1}^*\sqrt{\lambda_i^*}$	$\cdots$	$a_{p1}^*\sqrt{\lambda_p^*}$	1
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
X_k^*	$a_{1k}^*\sqrt{\lambda_1^*}$	$\cdots$	$a_{ik}^*\sqrt{\lambda_i^*}$	$\cdots$	$a_{pk}^*\sqrt{\lambda_p^*}$	1
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
X_p^*	$a_{1p}^*\sqrt{\lambda_1^*}$	$\cdots$	$a_{ip}^*\sqrt{\lambda_i^*}$	$\cdots$	$a_{pp}^*\sqrt{\lambda_p^*}$	1
$\sum_{k=1}^{p}\rho_{ik}^2$	λ_1^*	$\cdots$	λ_i^*	$\cdots$	λ_p^*	$\sum_{i=1}^{p}\sum_{k=1}^{p}\rho_{ik}^2=p$

● 样本主成分分析

将$\mathbf{\Sigma}$用$\mathbf{S}$替代，则λ_i，$\mathbf{a}_i$，$\mathrm{Var}(z_i)$，$\mathrm{Cov}(z_i,z_j)$，$\rho(z_i,x_k)$分别替换成$\hat{\lambda}_i$，$\hat{\mathbf{a}}_i$，$\widehat{\mathrm{Var}}(z_i)$，$\widehat{\mathrm{Cov}}(z_i,z_j)$，$r(z_i,x_k)$，性质与总体主成分相同.

其中$z_{i,k}=\hat{\mathbf{a}}_i'\boldsymbol{x}_k=\hat{a}_{i1}x_{k1}+\cdots+\hat{a}_{ip}x_{kp}$称为$\boldsymbol{x}_k$在第$i$个主成分上的得分，称$\left[\hat{\boldsymbol{a}}_1,\cdots,\hat{\boldsymbol{a}}_p\right]$为载荷。

● 基于标准化的（样本）主成分分析

将$\mathbf{R}$用$\hat{\mathbf{R}}$替代，则λ_i，$\mathbf{a}_i$，$\mathrm{Var}(z_i)$，$\mathrm{Cov}(z_i,z_j)$，$\rho(z_i,x_k^*)$分别替换成$\hat{\lambda}_i$，$\hat{\mathbf{a}}_i$，$\widehat{\mathrm{Var}}(z_i)$，$\widehat{\mathrm{Cov}}(z_i,z_j)$，$r(z_i,x_k^*)$，性质与标准化的总体主成分相同。

● 主成分分析在图像处理中的应用

■ 奇异值分解（SVD）：

对$n\times p$数据矩阵$\mathbf{X}$作分解：$\mathbf{X}=\mathbf{U\Lambda V'}$，其中$\mathbf{U'U}=\mathbf{I}_r$，$\mathbf{V'V}=\mathbf{I}_r$

$\mathbf{\Lambda}$为$\mathbf{X'X}$（或$\mathbf{XX'}$）的r个共同的非零特征值的平方根的对角矩阵。进一步地，$\mathbf{U},\mathbf{V}$的列分别包含了$\mathbf{XX'},\mathbf{X'X}$的特征向量。

■ 基于SVD的主成分分析：

记$\mathbf{M}$为对数据矩阵$\mathbf{X}$中心化后的数据矩阵，则$\mathbf{M}$的样本协方差矩阵为$\mathbf{S}=\dfrac{1}{n-1}\mathbf{M'M}$；再对$\mathbf{M}$作SVD分解$\mathbf{M}=\mathbf{U\Lambda V'}$，则$\mathbf{Z}=\mathbf{MV}$为主成分得分。此时对$\mathbf{M}$进行近似：$\mathbf{M}\approx(\mathbf{MV}_m)\mathbf{V'}_m=\hat{\mathbf{M}}$

■ 人脸识别与特征脸：

记$n\times p$维图像数据矩阵为$\mathbf{X}$，令$\mathbf{Y}=\dfrac{1}{\sqrt{n-1}}(\mathbf{X}-\mathbf{1}\boldsymbol{\mu}')$，其中$\boldsymbol{\mu}'=\mathbf{1'X}/n$。令$\lambda_i$为$\mathbf{YY'}$的第$i$个特征值，对应的特征向量为$\boldsymbol{u}_i$，那么$\lambda_i$对应$\mathbf{Y'Y}$的特征向量为$\boldsymbol{\phi}_i=\boldsymbol{Y}'\boldsymbol{u}_i$

选择m个主成分，记为$\hat{\Phi}_m=[\boldsymbol{\phi}_1,\cdots,\boldsymbol{\phi}_m]$，称为$m$个特征脸，则对任意一个图像$\boldsymbol{x}$，其主成分得分为：$\hat{\boldsymbol{Y}}_{m\times1}=\hat{\Phi}_m'(\boldsymbol{x}-\boldsymbol{\mu})$。因此：$\hat{\boldsymbol{x}}=\hat{\Phi}_m\hat{\boldsymbol{Y}}_{m\times1}+\boldsymbol{\mu}$

第10章 因子分析

● 因子分析模型及假设

设$\boldsymbol{X}$为p维随机向量，$\boldsymbol{\mu}$为其均值向量，若：

$$\boldsymbol{X}=\mathrm{A}\boldsymbol{F}+\boldsymbol{\mu}+\boldsymbol{\varepsilon}$$

其中因子载荷矩阵A是一个$p\times k$的常数矩阵($k<p$)，公共因子向量$\boldsymbol{F}$为k维随机向量，特殊因子向量$\boldsymbol{\varepsilon}$为p维随机向量。且要求满足：

$E(\boldsymbol{F})=0$, $\mathrm{Cov}(\boldsymbol{F})=\boldsymbol{I}_k$, $E(\boldsymbol{\varepsilon})=0$, $\mathrm{Cov}(\boldsymbol{\varepsilon})=\Psi$, $\mathrm{Cov}(\boldsymbol{F},\boldsymbol{\varepsilon})=0$

此时为正交因子分析模型，且$\mathrm{cov}(X)=\Sigma=\mathbf{AA'}+\Psi$

● 因子分析的统计意义

■ 载荷矩阵的统计意义：

可以证明$\mathrm{Cov}(\boldsymbol{X},\boldsymbol{F})=\mathrm{A}$，即$\mathrm{Cov}(X_i,F_j)=a_{ij}$，说明每个载荷$a_{ij}$反映了第$i$个变量与第$j$个公共因子的相关重要性。绝对值越大，相关的密切程度越高。

■ 变量共同度的统计意义：

记$h_i=\sum_{j=1}^{k}a_{ij}^2$，称为共同度，反映了$X_i$对公共因子的依赖程度。

注意到$\mathrm{Var}(X_i)=\sigma_{ii}=\sum_{j=1}^{k}a_{ij}^2+\psi_i=h_i+\psi_i$，因此如果$h_i$靠近$\sigma_{ii}$，则$\psi_i$非常小，表明因子分析的效果好。

■ 公共因子方差贡献的统计意义：

记$\mathrm{SS\ loadings}(F_j)=\sum_{i=1}^{p}a_{ij}^2$为公共因子$F_j$对$\boldsymbol{X}$的各个分量的总方差贡献，$\mathrm{SS\ loadings}(F_j)/\sum\mathrm{Var}(X_i)$为衡量公共因子$F_j$的贡献率。

● 因子分析模型的性质

■ 因子载荷矩阵不唯一：

令$\mathbf{Q}$为任意$k\times k$正交矩阵，令$\boldsymbol{X}=(\mathbf{AQ})(\mathbf{Q'F})+\boldsymbol{\mu}+\boldsymbol{\varepsilon}$，则新的因子载荷矩阵与公共因子向量仍然满足模型假设。

■ 因子分析具有尺度不变性：

令$\mathbf{C}=\mathrm{diag}(c_1,\cdots,c_p)$，令$\mathbf{C}\boldsymbol{X}-\mathbf{C}\boldsymbol{\mu}=\mathbf{CA}\boldsymbol{F}+\mathbf{C}\boldsymbol{\varepsilon}$，则新的因子载荷矩阵与公共因子向量仍然满足模型假设。

● 因子载荷矩阵的估计方法

■ 主成分法：

1. 求出$\mathbf{X}$的协方差矩阵Σ或无偏估计$\mathbf{S}$，并作谱分解：

$$\Sigma=\mathrm{AA'}+\Psi=\mathrm{U\Lambda U'}=\sum_{i=1}^{p}\lambda_iu_iu_i'$$

2. 取前k个较大的特征值及其对应的特征向量：

$$\Lambda_1=\mathrm{diag}(\lambda_1,\cdots,\lambda_k),\ \mathbf{U}_1=(u_1,\cdots,u_k)$$

3. 因子载荷矩阵与特殊因子方差估计为：

$$\hat{\mathrm{A}}=\mathrm{U}_1\Lambda_1^{1/2},\quad \hat{\psi}_i=\sigma_{ii}-\hat{h}_i$$

■ 主因子法：

1. 求出$\mathbf{X}$的相关系数矩阵$\mathbf{R}$，并假定已知特殊因子的方差Ψ，并对约相关矩阵$\mathbf{R}^*=\mathbf{R}-\Psi$作近似谱分解：

$$\mathbf{R}^*=\mathbf{AA'}\approx\sum_{i=1}^{p}d_i\mathbf{q}_i\mathbf{q}_i'=\mathbf{Q}_1\mathbf{D}_1\mathbf{Q'}_1$$

2. 因子载荷矩阵与特殊因子的方差估计为：

$$\hat{\mathbf{A}}=\mathbf{Q}_1\mathbf{D}_1^{1/2},\quad \hat{\psi}_i=1-\sum_{j=1}^{k}\hat{a}_{ij}^2$$

3. 给定初始值$\hat{h}_i$，替换$\hat{\mathbf{R}}$的对角线元素得到$\hat{\mathbf{R}}^*$，以此开始迭代更新

注：(1)要求最终求出的$\hat{\psi}_i\geqslant0$；

(2)初始$\hat{h}_i=1$时，主因子法与主成分法等价

■ 极大似然法（要求假定X服从多元正态分布）：

1. 给定Ψ的初始值（参考主因子法）

2. 对$\Psi^{-1/2}\mathbf{S}\Psi^{-1/2}$作谱分解，求其前$k$个特征值以及特征向量，并令

$$\mathbf{Q}=(\boldsymbol{q}_1,\cdots,\boldsymbol{q}_k),\quad \hat{\Delta}=\mathrm{diag}(\lambda_1-1,\cdots,\lambda_k-1)$$

3. 令$\hat{\mathbf{A}}=\Psi_0^{1/2}\mathbf{Q}\hat{\Delta}^{1/2}$，$\hat{\Psi}=\mathrm{diag}\left(\mathbf{S}-\hat{\mathbf{A}}\hat{\mathbf{A}}'\right)$，重复迭代。

● 因子旋转（正交旋转方差最大法）

记$\mathbf{A}$为未旋转的$p\times k$的载荷矩阵，$\mathbf{G}$为$k\times k$的正交矩阵，旋转后的因子载荷矩阵为$\mathbf{B}=\mathbf{AG}=(b_{ij})_{p\times k}$，方差最大准则要求最大化

$\phi=\sum_{j=1}^{k}\sum_{i=1}^{p}\left(d_{ij}^2-\bar{d}_j\right)^2$，其中：

$$d_{ij}=\frac{b_{ij}}{\sqrt{h_i}},\ \bar{d}_j=\frac{1}{p}\sum_{i=1}^{p}d_{ij}^2,\ \sqrt{h_i}=\left(\sum_{j=1}^{k}b_{ij}^2\right)^{1/2}$$

另外还有其他方法，如四次方最大法和等量最大法。

● 因子得分

■ Thomson因子得分：

假设X服从p元正态分布，F的先验分布为$N_k(0,\mathbf{I}_k)$，考虑对$X-\mu,F$的联合分布求条件分布的均值，就得到

$$E(\boldsymbol{F}|\boldsymbol{X})=\boldsymbol{F}^*=\mathrm{A}'(\mathrm{AA}'+\Psi)^{-1}(X-\boldsymbol{\mu})$$
$$=(\mathbf{I}_k+\mathbf{A}'\Psi^{-1}\mathbf{A})^{-1}\mathbf{A}'\Psi^{-1}(\boldsymbol{X}-\boldsymbol{\mu})$$

其中两种形式等价，且$\boldsymbol{\mu},\mathbf{A},\Psi,\Sigma$可用估计量代替。

■ Bartlett因子得分：

假设X服从p元正态分布，则Bartlett因子得分为：

$$\hat{\mathbf{F}}=(\mathbf{A}'\Psi^{-1}\mathbf{A})^{-1}\mathbf{A}'\Psi^{-1}(\mathbf{X}-\mu)$$

性质：(1)Bartlett因子得分具有无偏性；(2)当因子载荷采用主成分法估计时，Bartlett因子得分就是X的前k个样本主成分的标准化；(3)在已知$\boldsymbol{\mu},\mathbf{A},\Psi$的情况下，Thomson因子得分的预测均方误差小于Bartlett因子得分的预测均方误差。

● 因子分析与主成分分析的关系

■ 主成分分析对数据来源的总体分布的协方差结构不做任何假设；而因子分析假设数据来源于因子分析模型

■ 假设特殊因子的方差为零时，二者等价。

■ 主成分分析侧重"变异量"，而因子分析更重视相关变量的"共变异量"。

第11章 判别分析

● 判别准则简介

给定以一个体的测量值向量$\boldsymbol{x}=(x_1,\cdots,x_p)'$，判断该个体来自总体$\pi_1$或者总体$\pi_2$，对于该问题，主要是寻找判别方法，把$\mathbb{R}^p$空间分成两个区域$R_1$和$R_2$，记为$R=(R_1,R_2)$. 如果这个个体的观测值向量$\boldsymbol{x}\in R_1$，则把他判断为来自总体$\pi_1$，否则为来自总体$\pi_2$.

● 判别分析的错判损失

统计决策	π_1	π_2
π_1	0	$C(2\|1)$
π_2	$C(1\|2)$	0

● 判别概率

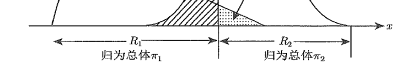
假设总体π_1的密度函数为$f_1(\boldsymbol{x})$，总体π_2的密度函数为$f_2(\boldsymbol{x})$，则：

个体来自总体π_1,被正确判别为π_1的概率为:$\mathrm{Pr}(1|1,R)=\int_{R_1}f_1(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$

个体来自总体π_1,被错判为π_2的概率为:$\mathrm{Pr}(2|1,R)=\int_{R_2}f_1(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$

个体来自总体π_2,被正确判别为π_2的概率为:$\mathrm{Pr}(2|2,R)=\int_{R_2}f_2(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$

个体来自总体π_2,被错判为π_1的概率为:$\mathrm{Pr}(1|2,R)=\int_{R_1}f_2(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$



● Bayes判别方法

令q_1是来自总体π_1的观测值的先验概率，q_2是来自总体π_2的观测值的先验概率，满足$q_1+q_2=1$.Bayes判别方法就是找到一种R_1,R_2的划分，使得错判的平均损失(ECM)最小化：

$$\mathrm{ECM}(R_1,R_2)=C(2|1)\mathrm{Pr}(2|1,R)q_1+C(1|2)\mathrm{Pr}(1|2,R)q_2$$

● 两个总体的判别准则（结合错判损失）

$$R_1=\{\boldsymbol{x}:[C(2|1)q_1]f_1(\boldsymbol{x})\geqslant[C(1|2)q_2]f_2(\boldsymbol{x})\}$$
$$R_2=\{\boldsymbol{x}:[C(2|1)q_1]f_1(\boldsymbol{x})<[C(1|2)q_2]f_2(\boldsymbol{x})\}$$

● 两个已知多元正态分布的判别

■ 协方差不相等的情况：

令$\delta(\boldsymbol{x})=-\dfrac{1}{2}\boldsymbol{x}'(\Sigma_1^{-1}-\Sigma_2^{-1})x+(\boldsymbol{\mu}_1'\Sigma_1^{-1}-\boldsymbol{\mu}_2'\Sigma_2^{-1})\boldsymbol{x}-\xi$，其中

$\xi=\dfrac{1}{2}\ln\left(\dfrac{|\Sigma_1|}{|\Sigma_2|}\right)+\dfrac{1}{2}\left(\boldsymbol{\mu}_1'\Sigma_1^{-1}\boldsymbol{\mu}_1-\boldsymbol{\mu}_2'\Sigma_2^{-1}\boldsymbol{\mu}_2\right)$，$k=\dfrac{q_2C(1|2)}{q_1C(2|1)}$，则最好的判别区域为：$R_1=\{\boldsymbol{x}:\delta(\boldsymbol{x})\geqslant\ln k\}$，$R_2=\{\boldsymbol{x}:\delta(\boldsymbol{x})<\ln k\}$

■ 协方差相等的情况：

若$\mathbf{\Sigma}_1=\mathbf{\Sigma}_2=\mathbf{\Sigma}$，$k$值同上，则最好的判别区域为：

$$R_1=\left\{x:\ x'\Sigma^{-1}(\mu_1-\mu_2)-\frac{1}{2}(\mu_1+\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)\geqslant\ln k\right\}$$
$$R_2=\left\{x:\ x'\Sigma^{-1}(\mu_1-\mu_2)-\frac{1}{2}(\mu_1+\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)<\ln k\right\}$$

此时$\delta(\boldsymbol{x})$退化为关于$\boldsymbol{x}$的一次函数。

● 参数未知时两个正态总体的判别

假设对两个总体存在两组历史样本，$\boldsymbol{x}_1^{(1)},\cdots,\boldsymbol{x}_{n_1}^{(1)}\sim\mathrm{N}_p(\boldsymbol{\mu}_1,\Sigma_1)$，$\boldsymbol{x}_1^{(2)},\cdots,\boldsymbol{x}_{n_2}^{(2)}\sim\mathrm{N}_p(\boldsymbol{\mu}_2,\Sigma_2)$，则可以用这两组样本估计$\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\mathbf{\Sigma}_1,\mathbf{\Sigma}_2$并在判别区域中使用这些估计量替换未知参数，判别准则同上。注意当$\mathbf{\Sigma}_1=\mathbf{\Sigma}_2=\mathbf{\Sigma}$时，$\mathbf{S}=\dfrac{\mathbf{V}^{(1)}+\mathbf{V}^{(2)}}{n_1+n_2-2}$

● 一组样本的整体判别

假设有一个来自总体π_1或π_2新的样本$\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n$，下面将样本当作整体进行判别。首先计算估计量$\bar{\boldsymbol{x}}=\dfrac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i$，$\hat{\Sigma}=\dfrac{\mathbf{V}^{(1)}+\mathbf{V}^{(2)}+\mathbf{V}}{n_1+n_2+n-3}$

其中$\mathbf{V}=\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})'$为新样本的样本离差矩阵。定义Fisher线性判别函数为：$\left[\bar{x}-\dfrac{1}{2}\left(\hat{\mu}_1+\hat{\mu}_2\right)\right]'\hat{\mathbf{\Sigma}}^{-1}\left(\hat{\mu}_1-\hat{\mu}_2\right)$

第12章 聚类分析

● 聚类分析的分类

■ R-型聚类:对变量或指标进行分类；Q-型聚类:对样本进行分类

● 数据标准化

■ 中心化变换：$x_{ij}^*=x_{ij}-\bar{x}_j$；标准化变换：$x_{ij}^*=\dfrac{x_{ij}-\bar{x}_j}{s_j}$

■ 极差标准化变换：$x_{ij}^*=\dfrac{x_{ij}-\bar{x}_j}{R_j}$

■ 极差正规化变换:$x_{ij}^*=\dfrac{x_{ij}-\min\limits_{1\leq t\leq n}x_{tj}}{R_j}$

■ 对数变换:$x_{ij}^*=log(x_{ij})$

● 样本之间的距离

■ 性质：非负性；对称性；三角不等式

■ Minkowski距离：$d_{ij}(q)=\left[\sum_{l=1}^{p}|x_{il}-x_{jl}|^q\right]^{1/q}$

其中$q=1,2,\infty$依次为绝对值、欧式、切比雪夫距离

不足：(1)距离与各变量的量纲有关；(2)没有考虑指标间的相关性；(3)没有考虑各变量方差的不同

■ 马氏距离：$d_{ij}^2(M)=(\boldsymbol{x}_i-\boldsymbol{x}_j)'\mathbf{S}^{-1}(\boldsymbol{x}_i-\boldsymbol{x}_j)$

优点：(1)排除了各指标之间相关性的干扰；(2)不受各个指标量纲的影响；(3)将原始数据作线性变换后，距离不变

■ Canberra（兰氏）距离、斜交空间距离

● 变量之间的相似系数

■ 性质：$|c_{ij}|\leqslant1$；$c_{ij}=c_{ji}$；$c_{ij}=\pm1$当且仅当X_i,X_j共线

■ 夹角余弦：$c_{ij}(1)=\dfrac{\sum_{k=1}^{n}x_{ki}x_{kj}}{\sqrt{\sum_{k=1}^{n}x_{ki}^2}\sqrt{\sum_{k=1}^{n}x_{kj}^2}}$

■ Pearson相关系数$c_{ij}(2)=\dfrac{\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)^2}\sqrt{\sum_{k=1}^{n}(x_{kj}-\bar{x}_j)^2}}$

■ Kendallτ相关系数、Spearman相关系数

● K均值聚类

■ 基本性质：(1)每个观测样本属于K个类中至少一个类；(2)没有一个观测样本同时属于两个类或更多的类中。

■ 目标函数：$\min\limits_{G_1,\cdots,G_K}\left\{\sum_{k=1}^{K}\dfrac{1}{|G_k|}\sum_{i,i'\in G_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2\right\}$

■ 算法步骤：

1. 为每个观测样本随机分配一个从1到K的数字，可以看作是这些观测样本的初始类；

2. 重复以下操作，直到类的分配停止位置：

(1) 分别计算K个类的类中心，第k个类的中心是第k个类中样本的p维观测向量的均值向量；

(2) 将每个观测样本分配到距离其最近的类中心所在的类中。

上述算法保证了下式的单调递减：

$$\frac{1}{|G_k|}\sum_{i,i'\in G_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2=2\sum_{i\in G_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2$$

■ K值的确定：绘制关于类内总平方和的碎石图

类内总平方和：$\mathrm{WSS}_k=\sum_{l=1}^{k}\mathrm{SS}^{(l)}=\sum_{l=1}^{k}\sum_{i=1}^{n_l}\sum_{j=1}^{p}\left(x_{ij}^{(l)}-\bar{x}_j^{(l)}\right)^2$

■ 缺陷：(1)对椭球或椭圆形的凸形区域的数据更适用，对其他非凸的数据集则可能失效；(2)对异常值敏感

● 系统聚类的算法步骤：

1. 首先将每个样本看作一类，计算n个观测样本中所有两两样本间的距离（共$\mathrm{C}_n^2=n(n-1)/2$）个。

2. 令$k=n,n-1,\cdots,2$：

(1) 在第k个类中，比较任意两类间的距离，找到距离最小的那一对类，并将他们合并起来；

(2) 计算剩下的$k-1$个新类中，每两个类之间的距离，同样把距离最小的两个类合并。