应用多元统计分析习题参考解答

刘辰昂

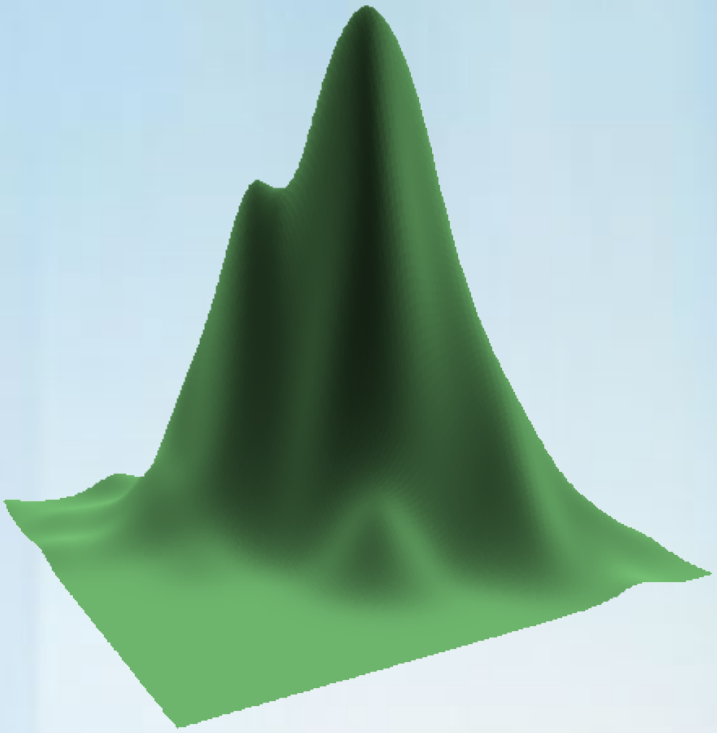
浙江大学

Email: liuchenang@gmail.com

Weibo: weibo.com/liuchenang

Blog: chenangliu.info

LinkedIn: Chenang Liu



2014 年 8 月 24 日

序言

书中对应的教材是北京大学出版社的《应用多元统计分析》一书，高惠璇老师主编。因为之前在论坛上多次看到寻课后习题答案的求助帖，而现在网上也没有一份完整的习题解答，故结合网上一些现有的资料以及去年在学该课程时的作业，再经过修改加工整理成册。解答覆盖了全书除第四章回归分析外其余十章的所有课后习题(更新完毕后)，由于并未附上原题故请配合教材使用。另外无心之作难免会出现各种疏漏，由此对使用者可能造成的任何后果恕不负责，但如若发现，烦请尽快联系我，我会及时更正，以使受害最小化，谢谢！

关于上机题的几点说明：(1)所有上机题均通过R语言实现，原书作者已经提供了SAS 版本的解答，可在网上直接下载；(2) 相关的数据均采用txt文件，由于不同版本的数据可能存在细微不同，为防止尴尬建议从笔者博客上下载，下载地址与本书下载地址相同；(3) 若无特殊说明，本书代码运行前待读取的数据均已存入工作目录下，若出现数据无法读取请先检查数据是否存放正确；(4)采用版本为R3.0.1，由于R 语言更新速度很快，尤其是用到扩展包的地方，如果出现变动，会尽可能及时的更新。也烦请及时指正。

目录

第一章 绪论

1-1

(1)参考代码

```
data<-read.table("0101.txt",col.name=c("number","age",
"weight","time","spulse","rpulse","mpulse","OXY"));
attach(data);
par(font.lab=4,font.axis=2,font.main=4,cex.main=2,
bg="lightgray");
plot.default(OXY,time,col="blue",lwd=3,cex=1.1,
main="Time")
dev.new();
par(font.lab=4,font.axis=2,font.main=4,cex.main=2,
bg="lightgray");
plot.default(OXY,age,col="red",lwd=3,cex=1.1,
main="Age")
```

散布图如下:

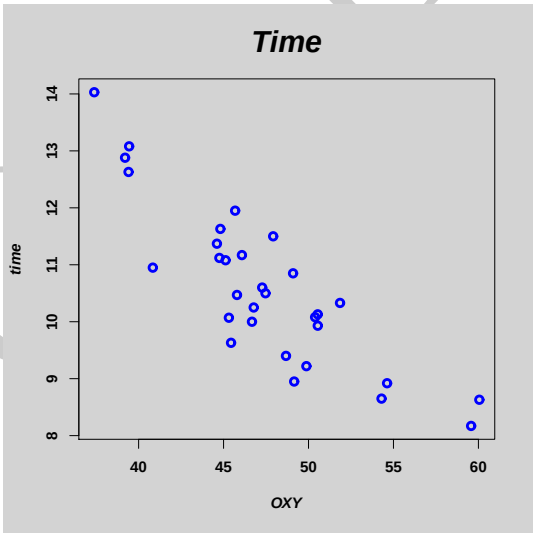


图 1:

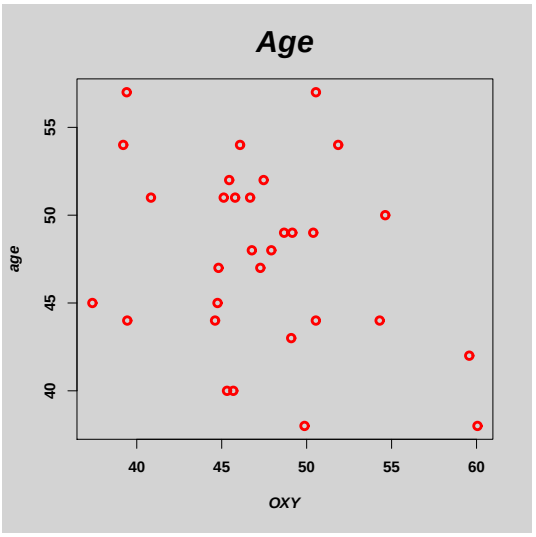


图 2:

由图我们可以看到肺活量与跑步时间有明显的负相关关系，即肺活量越大，跑步时间越短。但肺活量和年龄却从直观上看不出明显的相关关系。

此外值得一提的是本题也可以采用ggplot2作图系统实现。

```
library(ggplot2);
data<-read.table("0101.txt",col.name=c("number","age",
"weight","time","spulse","rpulse","mpulse","OXY"));
attach(data);
dev.new();
qplot(OXY,time,data=data,colour=age);
dev.new();
qplot(OXY,age,data=data,colour=time);
```

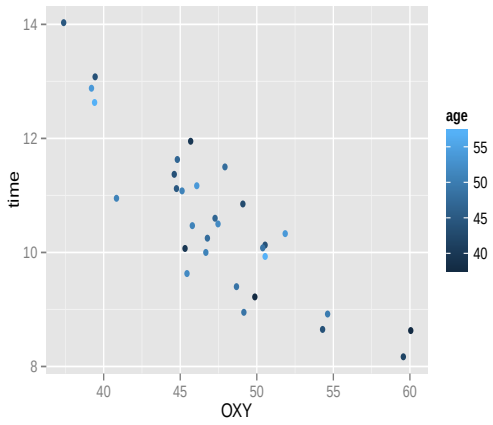


图 3:

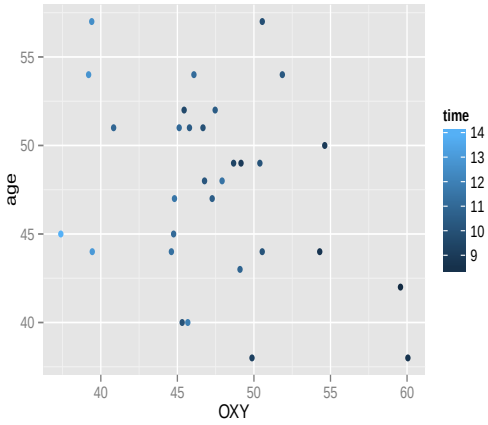


图 4:

(2)绘制散布图矩阵可以通过最基本的plot()函数实现，自带的graphics包还有专门用于散布图矩阵绘制的pairs() 函数，具体的用法可以参考谢益辉的《现代统计图形》。car包中的scatterplot.matrix()函数(调用时可采用简写spm())给出了一种较pairs在定义对角线上更为简洁的方式，参考代码如下：

```
data<-read.table("0101.txt",col.name=c("number",
```

```
"age","weight","time","spulse","rpulse","mpulse","OXY"));
data3<-data[c(1,2,21,22),];
library(car);
spm(data[,-1]);
```

在ggplot2作图系统中则是提供了ggpairs()函数来实现(原先的plotmatrix()函数已经被弃用,ggpairs 被收录在GGally包中)，参考代码如下：

```
library(GGally);
ggpairs(data[,-1]);
```

效果图如下：

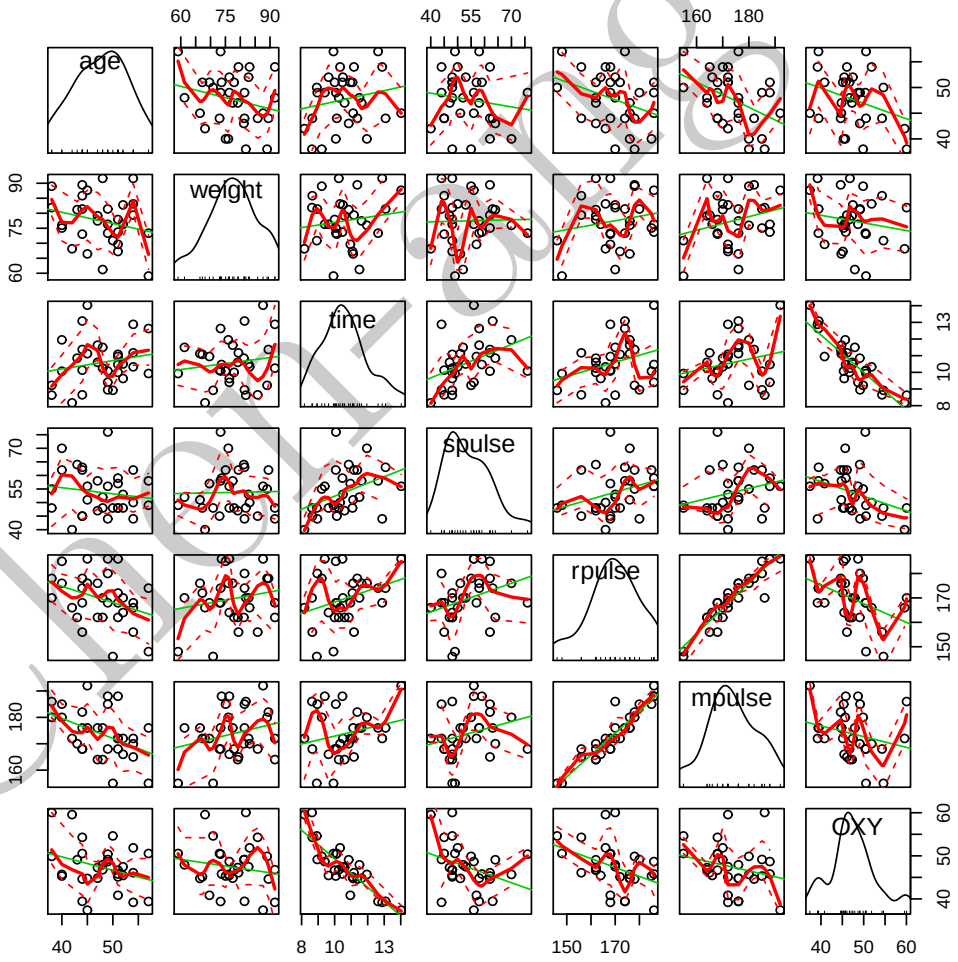


图 5:

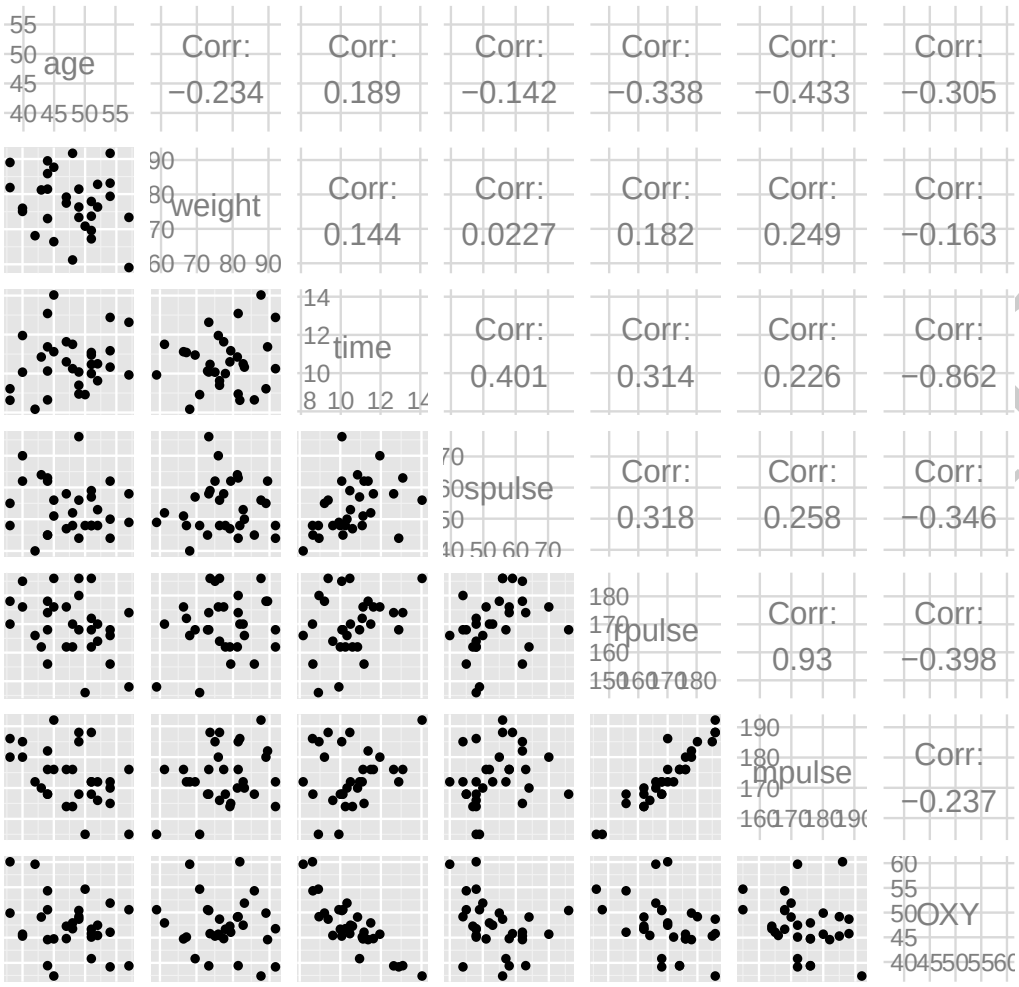


图 6:

由于变量较多，所以效果并不理想(可以进一步调整参数改善效果)。根据上图我们可以看到某些变量之间具有明显的相关性，例如rpulse与mpulse两者存在很强的正相关性，以及之前提到的time和OXY存在明显负相关，其余变量间则相关性并不强。

(3)轮廓图的绘制方法众多，lattice和ggplot2等不同的作图系统均有不错的支持，在基础作图系统中通过MASS包中的parcoord()即能很方便的实现，参考代码如下：

```
library(MASS);
data<-read.table("0101.txt",col.name=c("number",
"age","weight","time","spulse","rpulse","mpulse","OXY"));
```

```
data3<-data[c(1,2,21,22),];
par(bg="lightgrey",font.axis=2,mar=rep(3,4));
parcoord(data3[,-1],lty=1:4,col=1:4,lwd=2);
```

同样在ggplot2作图系统中也同样可以实现，但值得一提的是原先的ggpcp() 函数同样也已经被弃用，现在轮廓图的绘制函数也已被移至GGally 包中，参考代码如下(数据同上)：

```
library(GGally);
ggparcoord(data3,columns=c(2:8),groupColumn=4);
```

效果图如下：

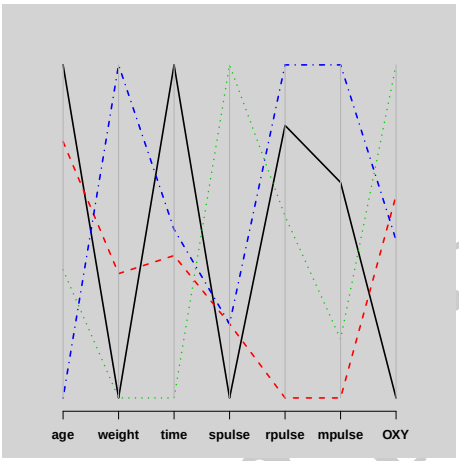


图 7:

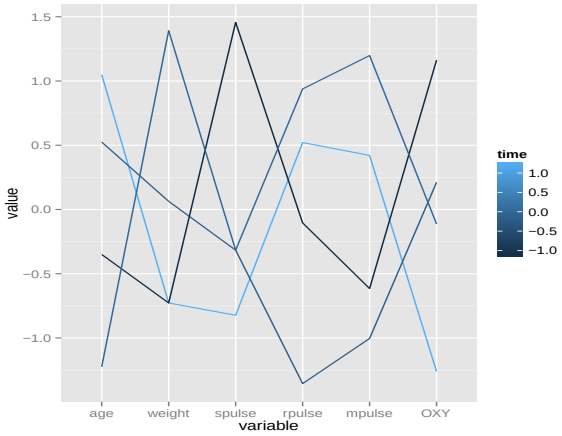


图 8:

此外，lattice中的parallel()函数以及iplots包中的ipcp()函数也均可以实现轮廓图的绘制，读者可自行参考帮助文档。

雷达图比较常用的绘制方法主要有两种，一是可以利用R中graphics包自带的stars函数，参考代码如下：

```
data<-read.table("0101.txt",col.name=c("number",
"age","weight","time","spulse","rpulse","mpulse","OXY"));
data3<-data[c(1,2,21,22),];
par(bg="grey",bty="n",xaxt="n", yaxt="n",mar=rep(2,4),
pch=19,font=2);
```

```
stars(data3[,-1],locations=c(0,0),radius=F,key.loc=c(0,0),
main="",frame.plot=TRUE,axes=TRUE,lwd=2,
col.lines=rainbow(10));
```

为取得更好的视觉效果，也可以采用更为灵活的fmsb包中的radarchart函数，参考代码如下：

```
library(fmsb);
data<-read.table("0101.txt",col.name=c("number",
"age","weight","time","spulse","rpulse","mpulse","OXY"));
data3<-data[c(1,2,21,22),];
data3t<-as.data.frame(data3[,-1]);
par(bg="grey",bty="n",xaxt="n", yaxt="n",
mar=rep(0,4),pch=19,font=2);
radarchart(data3t,centerzero=T,maxmin=F);
```

效果图如下

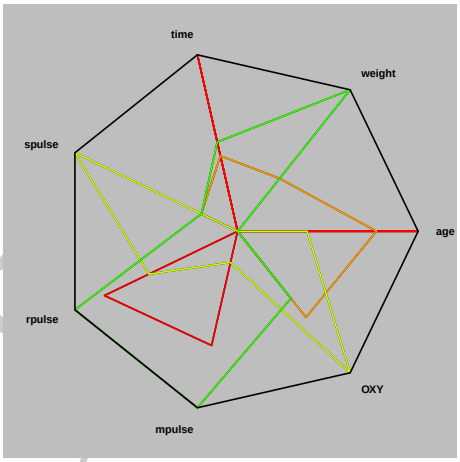


图 9:

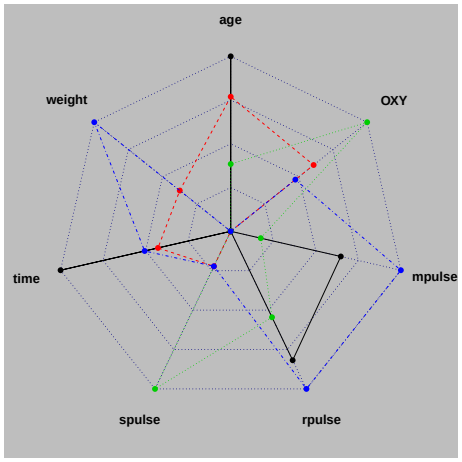


图 10:

如果想通过ggplot2实现的话可以参考stackoverflow上的这个帖子。

(4)调和曲线图的绘制函数可以在MSG包中以及andrews包中分别找到，因效果类似以MSG为例，参考代码如下：

```
data<-read.table("0101.txt",col.name=c("number",
"age","weight","time","spulse","rpulse","mpulse","OXY"));
data3<-data[c(1,2,21,22),];
library(MSG);
par(bg="lightgrey",font.lab=2,font.axis=2)
andrews_curve(data3[,-1]);
```

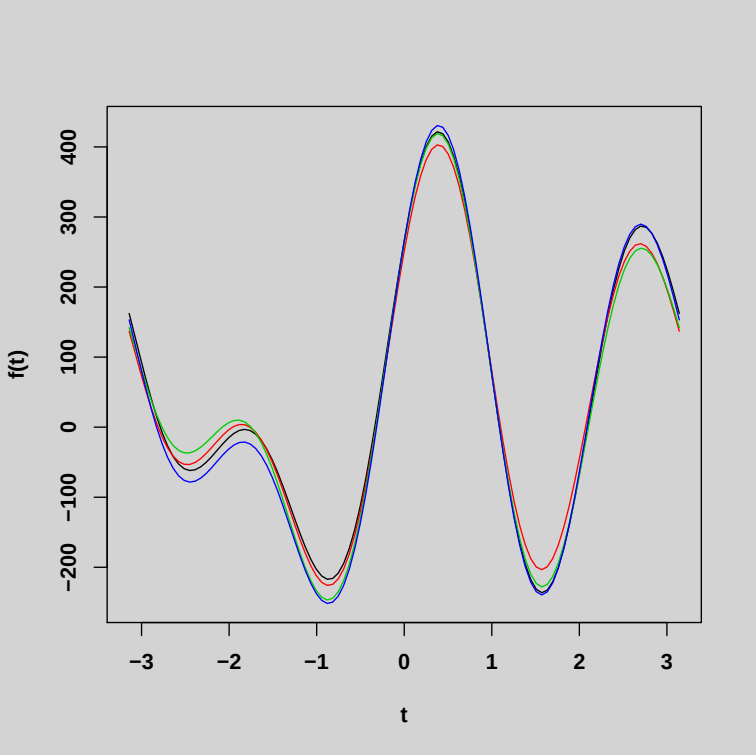


图 11:

第二章　多元正态分布及参数的估计

2-1

根据2.2节中的性质二即可知$Y \sim N_2(A\mu + d, A\Sigma A')$，代入数据即可得

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

$$= \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

$$A\Sigma A' = 2 \begin{pmatrix} 0.5 & -1 & 0.5 \\ -0.5 & 0 & -0.5 \end{pmatrix} \begin{pmatrix} 0.5 & -0.5 \\ -1 & 0 \\ 0.5 & -0.5 \end{pmatrix}$$

$$= \begin{pmatrix} 3 & -1 \\ -1 & 1 \end{pmatrix}$$

故由此知$Y \sim N_2(\begin{pmatrix} 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 & -1 \\ -1 & 1 \end{pmatrix})$

2-2

(1)不妨令$Y_1 = X_1 + X_2$, $Y_2 = X_1 - X_2$, 且

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

故同样由2.2节性质二知$Y \sim N_2(C\mu, C\Sigma C')$, 再令$C\Sigma C' = \Sigma_Y$, 由定理2.3.1 即可知若$\Sigma_Y$为对角阵，独立性即成立，代入数据计算得

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

故命题得证

(2)由2.2节性质4知X_1+X_2, X_1-X_2即Y_1, Y_2均为一维正态随机变量，且由均值的性质可知$\mu_{Y_1}=\mu_1+\mu_2, \mu_{Y_2}=\mu_1-\mu_2$, 再结合(1)中结论即可得

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

2-3

(1)本题可视为前一小题的进一步推广，采用类似的思路，由题设可知

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

故同样由2.2节性质二可知，$Y \sim N_{2p}(C\mu, C\Sigma C')$, 代入数据即有

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

同样由定理2.3.1可知独立性成立，故得证

(2)由前一小题结论可知

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

故由性质2推论可得两者分布为

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

2-4

(1)这里不妨令$X^{(1)}=(X_1, X_2)', X^{(2)}=X_3$，由定理2.3.2可知

$$(X^{(1)}|X^{(2)}) \sim N_2(\mu_{1\cdot 2}, \Sigma_{11\cdot 2})$$

代入数据即有

$$\mu_{1\cdot 2} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \rho \\ \rho \end{pmatrix} (x_3 - \mu_3)$$
$$= \begin{pmatrix} \mu_1 + \rho(x_3 - \mu_3) \\ \mu_2 + \rho(x_3 - \mu_3) \end{pmatrix}$$

$$\Sigma_{11\cdot 2} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} - \begin{pmatrix} \rho \\ \rho \end{pmatrix} \begin{pmatrix} \rho & \rho \end{pmatrix}$$
$$= \begin{pmatrix} 1-\rho^2 & \rho-\rho^2 \\ \rho-\rho^2 & 1-\rho^2 \end{pmatrix}$$

同理可得

$$(X_1|X_2, X_3) \sim N(\mu_1 + \frac{\rho}{1+\rho}(x_2 + x_3 - \mu_2 - \mu_3), \frac{1+\rho-2\rho^2}{1+\rho})$$

(2)由(1)的结论即可知条件协方差均为$\rho - \rho^2$

2-5

先考虑X_1与$X_1 + X_2$的联合分布，由X_1与X_2的独立性知，X_1与$X_1 + X_2$的任意线性和均为一元正态随机变量，故由2.2节性质4其联合分布为二元正态分布，且形式为

$$\begin{pmatrix} X_1 \\ X_1 + X_2 \end{pmatrix} \sim N_2(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix})$$

故由定理2.3.2即可得

$$\mu = \frac{1}{2}(x_1 + x_2)$$
$$\Sigma = \frac{1}{2}$$

2-6

(1)不妨记$Y = 3X_1 - 2X_2 + X_3$, 同样由性质4知Y为正态随机变量，故只需确定均值与方差即可，代入数据有

$$\mu_y = 13$$
$$\sigma_y = 9$$

(2)令$Z = X_3 - a'(X_1, X_2)'$, 由联合分布的正态性知Z为正态随机变量，先求协方差有

$$\begin{aligned} Cov(X_3, Z) &= D(X_3) - a_1 Cov(X_1, X_3) - a_2 Cov(X_2, X_3) \\ &= 2 - a_1 - 2a_2 \end{aligned}$$

由正态性知只需协方差为0即独立性成立，故对于满足$a_1 + 2a_2 = 2$的向量a均满足题意。

2-7

(1)不独立。$Cov(X_1, 2X_2) = 2Cov(X_1, X_2) = -4$。

(2)独立。因为由协方差矩阵知$Cov(X_2, X_3) = 0$。

(3)独立。结合协方差阵由定理2.3.1即可知两者独立。

(4)独立。$Cov(\frac{1}{2}(X_1 + X_2, X_3) = \frac{1}{2}(Cov(X_1, X_2), Cov(X_2, X_3)) = 0$。

(5)不独立。$Cov(X_2, X_2 - \frac{5}{2}X_1 - X_3) \neq 0$。

2-8

先由2.2节性质2知Y与Z均服从多元正态分布，故两者的独立等

价于协方差阵为全零阵，再结合2.1 节性质一第四条，有

$$\begin{aligned} COV(Y,Z) &= COV(AX+d, BX+c) \\ &= COV(AX, BX) \\ &= A \cdot COV(X,X) \cdot B' \\ &= A\Sigma B' \end{aligned}$$

故$A\Sigma B' = O$等价于协方差阵全零，故命题得证

2-9

(1)代入数据计算即有$AA' = I_4$

(2)**命题一：** 由题设可知$Y_1 = 2\bar{X}$，且

$$\begin{aligned} \sum_{i=1}^{4} Y_i^2 = Y'Y = X'A'AX = X'X = \sum_{i=1}^{4} X_i^2 \\ = \sum_{i=1}^{4}(X_i - \bar{X})^2 + n\bar{X}^2 \\ = \sum_{i=1}^{4}(X_i - \bar{X})^2 + Y_1^2 \end{aligned}$$

故命题一得证。

命题二： 由协方差阵性质知

$$D(Y) = AD(X)A' = A(\sigma^2 I_4)A' = \sigma^2 I_4$$

由正态分布独立与不相关的等价性可得Y_1, Y_2, Y_3, Y_4相互独立。命题二得证。

命题三： 结合命题二的结论，只需再求均值向量即可，由题设知

$$E(Y) = A \cdot E(X) = (2\mu, 0, 0, 0)'$$

故命题三得证。

2-10

事实上同样由2.2节性质二可知矩阵A只需满足$AI_2A' = AA' = \Sigma$即可。根据题意即只需给出任一符合该条件的矩阵A即可，不妨令

$$A = \begin{pmatrix} 2 & 0 \\ 1 & 0 \end{pmatrix}$$

即满足题意。

2-11

采用比较系数法，即加原式与所对应的标准形式对比，利用恒等关系求解。解得

$$\begin{cases} \mu_1 = 4 \\ \mu_2 = 3 \\ \sigma_1 = 1 \sigma_2 = \sqrt{2} \\ \rho = -\frac{\sqrt{2}}{2} \end{cases}$$

所以矩阵向量和协方差矩阵分别为

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

另外本题也可采用求边缘分布、协方差以及配方等方法求解。

2-12

(1)通过定义证明，分段讨论，当$x < -1$时，有

$$P(X_2 \leq x) = P(X_1 \leq x) = \Phi(x)$$

当$-1 \leq x \leq 1$时，有

$$\begin{aligned} P(X_2 \leq x) &= P(X_2 \leq -1) + P(-1 < X_2 \leq x) \\ &= P(X_1 \leq -1) + P(-x \leq X_1 < 1) \\ &= P(X_1 \leq -1) + P(-1 < X_1 \leq x) \\ &= P(X_1 \leq x) = \Phi(x) \end{aligned}$$

当$x > 1$时，有

$$\begin{aligned} P(X_2 \leq x) &= P(X_2 \leq -1) + P(-1 < X_2 \leq 1) + P(1 < X_2 \leq x) \\ &= P(X_1 \leq -1) + P(-1 < -X_1 \leq 1) + P(1 < X_1 \leq x) \\ &= P(X_1 \leq x) = \Phi(x) \end{aligned}$$

故由此可证$X_2 \sim N(0,1)$

(2)由2.2节性质4所给出的充要条件可知只需证明存在某个(X_1, X_2)的线性函数不是一维正态随机变量即可。令$Y = X_1 - X_2$，考虑其在0处的概率取值情况，由题设可知

$$\begin{aligned} P(Y = 0) &= 1 - P(-1 \leq X_1 \leq 1) \\ &= 2\Phi(-1) \neq 0 \end{aligned}$$

而事实上若$Y\ N(0,1)$，则$P(Y = 0) = 0$, 即Y不服从一元正态分布，故(X_1, X_2)不服从二元正态分布。

2-13

(1)由协方差阵定义知

$$\begin{aligned} \Sigma &= E[(X - \mu)(X - \mu)'] \\ &= E[XX' + \mu\mu' - X\mu' - \mu X'] \\ &= E[XX'] + \mu\mu' - E(X)\mu' - \mu E(X') \\ &= E[XX'] - \mu\mu' \end{aligned}$$

移项后即可得证。

(2)利用第一小题结论并结合迹的性质，由题设可知

$$\begin{aligned} E(X'AX) &= E(\mathrm{tr}(X'AX)) = E(\mathrm{tr}(AXX')) \\ &= \mathrm{tr}(AE(XX')) = \mathrm{tr}(A(\Sigma + \mu\mu')) \\ &= tr(\Sigma A) + \mu' A\mu \end{aligned}$$

(3)代入数据计算即可得$E(X'AX) = \sigma^2(p-1)$。第二部分的证明题目似乎有些问题。

2-14

本题毫无意义，略过。答案与定理2.5.1相同。

2-15

由定理2.5.1, 并结合其推导过程，即可得

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(X_{(i)} - \mu_0)(X_{(i)} - \mu_0)'$$

2-16

本题即为34页最佳预测结论的证明。不妨令$X = (x_1, x_2, \cdot, x_m)$, 再令$h(X)$为任意的$m$元函数，则有

$$\begin{aligned} E(Y - h(X))^2 &= E(Y - g(X) + g(X) - h(X))^2 \\ &= E(Y - g(X))^2 + E(g(X) - h(X))^2 \\ &\quad + 2E((Y - g(X))(g(X) - h(X))) \end{aligned}$$

事实上可以证明第三项为0，证明如下

$$\begin{aligned}E((Y-g(X))(g(X)-h(X))) &= E(E((Y-g(X))(g(X)-h(X)|X))) \\ &= E((g(X)-h(X))E((Y-g(X))|X)) \\ &= E((g(X)-h(X))(E(Y|X)-g(X))) \\ &= 0\end{aligned}$$

故对任意的$h(X)$都有

$$E(Y-g(X))^2 \leq E(Y-h(X))^2$$

由此故原命题得证。

2-17

(1)不妨记Σ的特征值从大到小分别为$\lambda_1, \lambda_2, \cdots, \lambda_p$,对应的特征向量为$l_1, l_2, \cdots, l_p$，又$\Sigma^{-1}$为对称阵，故对其进行谱分解有

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

再根据题设可知$f = a$等价于

$$(x-\mu)'\Sigma^{-1}(x-\mu) = b^2$$

这里$b^2 = -2\ln[a(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}]$, 且要求$a$能使$b^2 > 0$。再令$y_i = (x-\mu)'l_i (i = 1, 2, \cdots, p)$, 则有

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

化简后可得

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

方程即为椭圆方程，故命题得证。

(2)根据题设条件可求得协方差阵的特征值为

$$\lambda_1 = \sigma^2(1+\rho)$$
$$\lambda_2 = \sigma^2(1-\rho)$$

特征值对应的特征向量为

$$l_1 = (\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})'$$
$$l_2 = (\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2})'$$

再由第一小题结论可知椭圆方程为

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

故长轴半径为$b\sigma\sqrt{(1+\rho)}$, 短轴半径为$b\sigma\sqrt{(1-\rho)}$, 方向分别沿着$(1,1)$方向与$(1,-1)$方向。

2-18

(1)即证$E(Z)=\mu$，由题设可知对任意满足题意的i, 都有$EX_{(i)}=\mu$，结合$\sum_{i=1}^{n} c_i = 1$ 可得

$$E(Z) = \sum_{i=1}^{n} c_i EX(i) = \mu \sum_{i=1}^{n} c_i = \mu$$

故原命题得证。

(2)显然Z为p元正态随机向量，故只需证$\Sigma_Z = c'c\Sigma$即可，考虑任意的$Z_i = c_i X_{(i)}$ 都有$D(Z_i) = c_i^2\Sigma$, 所以

$$D(Z) = \sum_{i=1}^{n} D(Z_i) = c'c\Sigma$$

结合第一小题关于均值向量的结论原命题得证。

(3)由协方差阵性质可知Σ非负定，故由题设可知只需使$c'c$取到最小值即可，由柯西不等式取等号的条件即可知等权重情况下$c'c$ 极小，故命题得证。

2-19

R语言代码及输出结果如下

```
> data<-read.table("0219.txt",header=F);
> mu<-apply(data,2,mean)
> mu
   V1    V2    V3    V4
 5.50 68.10 46.50 32.29
> mu<-apply(data[,-1],2,mean)
> mu
   V2    V3    V4
68.10 46.50 32.29
> cov(data[,-1])
          V2         V3       V4
V2  4.766667 -1.9444444 1.9344444
V3 -1.944444  3.8333333 0.6166667
V4  1.934444  0.6166667 6.1898889
> cormat<-cor(data[,-1])
> cormat
           V2         V3        V4
V2  1.0000000 -0.4548832 0.3561291
V3 -0.4548832  1.0000000 0.1265962
V4  0.3561291  0.1265962 1.0000000
```

第三章　多元正态总体参数的假设检验

3-1

与教材52页结论三证明思路相同，由于A为对称幂等矩阵，而对称幂等矩阵的特征值非0即1，且只有r 个非0 特征值，即存在正交矩阵Γ，记为$(\gamma_1, \gamma_2, \cdots, \gamma_n)$, 使得

$$\Gamma' A \Gamma = \begin{pmatrix} I_r & O \\ O & O \end{pmatrix}$$

令$Y = (Y_1, Y_2, \cdots, Y_n)' = \Gamma' X$(即$X = \Gamma Y$), 则

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

进一步可以得到

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

显然$Y_i (i = 1, 2, \cdots, r)$间两两相互独立，且$Y_i \sim N(r_i'\mu, \sigma^2)$, 故由非中心卡方分布定义知，$\frac{1}{\sigma^2} X'AX \sim \chi^2(r, \delta)$, 其中非中心参数

$$\begin{aligned} \delta &= \frac{1}{\sigma^2} \sum_{i=1}^{r} (\gamma_i'\mu)^2 = \frac{1}{\sigma^2} [\mu'(\gamma_1\gamma_1' + \cdots + \gamma_r\gamma_r')\mu] \\ &= \frac{1}{\sigma^2} \mu'(\gamma_1, \cdots, \gamma_r)(\gamma_1', \cdots, \gamma_r')'\mu \\ &= \frac{1}{\sigma^2} \mu' \Gamma \begin{pmatrix} I_r & O \\ O & O \end{pmatrix} \Gamma'\mu \\ &= \frac{1}{\sigma^2} \mu' A \mu \end{aligned}$$

故命题得证。

3-2

即教材54页结论6的充分性证明，证明思路与结论5部分类似。设$\text{rank}(A) = r > 0$, 且$r < n$, 因为当$r = 0$ 时，则$A = O$，独立性显

然成立，同样当$r=0$时，由$AB=O$知$B=O$, 独立性也显然成立。故这里只需考虑一般情况即可，因为A为对称矩阵，故存在正交矩阵Γ 使得

$$\Gamma' A\Gamma = \begin{pmatrix} D_r & O \\ O & O \end{pmatrix}$$

且这里

$$D_r = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_r \end{pmatrix}$$

$\lambda_i (i=1,\cdots,r)$为A的特征值。由此可以得到

$$AB = \Gamma \begin{pmatrix} D_r & O \\ O & O \end{pmatrix} \Gamma' \cdot B\Gamma\Gamma'$$

不妨令$H = \Gamma' B\Gamma = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix}$, 其中$H_{11}$为r阶方阵，$H$为n阶方阵，由此进一步可得

$$AB = \Gamma \begin{pmatrix} D_r & O \\ O & O \end{pmatrix} \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix} \Gamma' = \Gamma \begin{pmatrix} D_r H_{11} & D_r H_{12} \\ O & O \end{pmatrix} \Gamma'$$

由于$AB=O$故$D_r H_{11} = O, D_r H_{12} = O$, 又根据之前的假设可知$D_r$满秩，故有$H_{11} = O_{r\times r}, H_{12} = O_{r\times(n-r)}$, 再由对称性知$H_{21} = O_{(n-r)\times r}$。因此

$$H = \Gamma' B\Gamma = \begin{pmatrix} O & O \\ O & H_{22} \end{pmatrix}$$

再令$Y = \Gamma' X$(即$X = \Gamma Y$), 则$Y \sim N_n(\Gamma'\mu, \sigma^2 I_n)$, 且结合之前的结论可知

$$\xi = X'AX = T'\Gamma' A\Gamma Y = \sum_{i=1}^{r} \lambda_i Y_i^2$$

$$\eta = X'BX = T'\Gamma' B\Gamma Y = Y'HY = (Y_{r+1},\cdots,Y_n) H_{22} (Y_{r+1},\cdots,Y_n)'$$

由于Y_i两两相互独立，故$X'AX$与$X'BX$相互独立。

3-3

本题即教材55页结论3的证明。参考结论2的证明思路，由协方差阵的性质, 有$\text{rank}(\Sigma)=p$, 且存在正交阵Γ和$\lambda_i>0(i=1,2,\cdots,p)$, 满足

$$\Sigma=\Sigma^{\frac{1}{2}}\cdot\Sigma^{\frac{1}{2}}$$

这里$\Sigma^{\frac{1}{2}}=\Gamma\text{diag}(\sqrt{\lambda_1},\cdots,\sqrt{\lambda_p})\Gamma'$ 为Σ的平方根阵，并且记

$$\Sigma^{-\frac{1}{2}}=\Gamma\text{diag}(\frac{1}{\sqrt{\lambda_1}},\cdots,\frac{1}{\sqrt{\lambda_p}})\Gamma'$$

显然有$\Sigma^{\frac{1}{2}}\Sigma^{-\frac{1}{2}}=I_p$, 之后考虑如下变换:

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

于是可以得到

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

由上一小题对应的结论6可知，两者独立等价于

$$\Sigma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}}\cdot\Sigma^{\frac{1}{2}}B\Sigma^{\frac{1}{2}}=O_{p\times p}$$
$$\Longleftrightarrow\Sigma A\Sigma B\Sigma=O_{p\times p}$$

故命题得证。

3-4

Wishart分布性质4证明:

先令$X_{(\alpha)}=(X_{(\alpha)}^{(1)},X_{(\alpha)}^{(2)})'$，通过调整分量的相对比例可以得到

$$X_{(\alpha)}^{(1)}\sim N_r(0,\Sigma_{11}),X_{(\alpha)}^{(2)}\sim N_{p-r}(0,\Sigma_{22})$$

再记 $\underset{n\times p}{X}=(x_{ij})=(\underset{n\times r}{X(1)}|\ \underset{n\times(p-r)}{X(2)}\ )$, 则有

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

根据对应关系即有

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

由Wishart分布定义即定义3.1.4可知，

$$W_{11}=X(1)'X(1)=\sum_{\alpha=1}^{n}(X_{(\alpha)}^{(1)})'X_{(\alpha)}^{(1)}\sim W_r(n,\Sigma_{11})$$
$$W_{22}=X(2)'X(2)=\sum_{\alpha=1}^{n}(X_{(\alpha)}^{(2)})'X_{(\alpha)}^{(2)}\sim W_{p-r}(n,\Sigma_{22})$$

即第一条得证。此外，当$\Sigma_{12}=O$时，对任意$\alpha=1,2,\cdots,n$ 有$X_{(\alpha)}^{(1)}$和$X_{(\alpha)}^{(2)}$ 独立，故可得W_{11}与W_{22}独立，即第二条得证。

Hotelling分布性质5:

设$X_(\alpha)(\alpha=1,\cdots,n)$是来自总体分布为$N_p(\mu,\Sigma)$的简单随机样本，$\bar{X}$和$A_x$ 则分别为样本均值和样本离差阵，根据教材61页性质1有

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

再令$Y_{(i)}=CX_{(i)}+d(i=1,\cdots,n)$, 这里$C$为$p\times p$非退化常数阵，$d$为$p\times 1$ 常数向量。则由教材23页性质2可得对于任意$i=1,\cdots,n$均有

$$Y_{(i)}\sim N_p(C\mu+d,C\sigma C')$$

由此再考虑样本均值向量与离差阵有

$$\begin{aligned}\bar{Y} &= C\bar{X} + d \\ A_y &= \sum_{i=1}^{n}(Y_{(i)} - \bar{Y})(Y_{(i)} - \bar{Y})' \\ &= C[\sum_{i=1}^{n}(X_{(i)} - \bar{X})(X_{(i)} - \bar{X})']C' \\ &= CA_xC'\end{aligned}$$

再记$\mu_y = C\mu + d$,则有

$$\begin{aligned}T_y^2 &= n(n-1)(\bar{Y} - \mu_y)'A_y^{-1}(\bar{Y} - \mu_y) \\ &= n(n-1)(\bar{X} - \mu)'C'[CA_xC']^{-1}C(\bar{X} - \mu) \\ &= n(n-1)(\bar{X} - \mu)'A_x^{-1}(\bar{X} - \mu)\end{aligned}$$

即有$T_x^2 = T_y^2$, 性质得证。

3-5

记总体为X，设$X_{(\alpha)}(\alpha = 1, \cdots, n)$为来自总体$X$的样本，且这里$n > p$, 由定义可知似然比统计量为

$$\lambda = \frac{\max\limits_{\mu=\mu_0} L(\mu_0, \Sigma_0)}{\max\limits_{\mu} L(\mu, \Sigma_0)}$$

对于分子有

$$\begin{aligned}\max_{\mu=\mu_0} L(\mu_0, \Sigma_0) &= \frac{1}{|2\pi\Sigma_0|^{\frac{n}{2}}}\exp[-\frac{1}{2}\sum_{\alpha=1}^{n}(X_{(\alpha)} - \mu_0)'\Sigma_0^{-1}(X_{(\alpha)} - \mu_0)] \\ &= \frac{1}{|2\pi\Sigma_0|^{\frac{n}{2}}}\exp[-\frac{1}{2}\mathrm{tr}[\Sigma_0^{-1}\sum_{\alpha=1}^{n}(X_{(\alpha)} - \mu_0)(X_{(\alpha)} - \mu_0)']] \\ &= \frac{1}{|2\pi\Sigma_0|^{\frac{n}{2}}}\exp[-\frac{1}{2}\mathrm{tr}[\Sigma_0^{-1}A_0]]\end{aligned}$$

对于分母，结合第二章中结论可知$\mu = \bar{X}$，同样的方法计算可得

$$\max_{\mu} L(\mu, \Sigma_0) = \frac{1}{|2\pi\Sigma_0|^{\frac{n}{2}}}\exp[-\frac{1}{2}\mathrm{tr}[\Sigma_0^{-1}A]]$$

将上述结果代入λ中，有

$$\begin{aligned}\lambda &= \exp[\frac{1}{2}\mathrm{tr}[\Sigma_0^{-1}A] - \frac{1}{2}\mathrm{tr}[\Sigma_0^{-1}A_0]] \\ &= \exp[\frac{1}{2}\mathrm{tr}[\Sigma_0^{-1}A] - \frac{1}{2}\mathrm{tr}[\Sigma_0^{-1}(A + n(\bar{X}-\mu_0)(\bar{X}-\mu_0)')]] \\ &= \exp[-\frac{n}{2}\mathrm{tr}[(\bar{X}-\mu_0)'\Sigma_0^{-1}(\bar{X}-\mu_0)]] \\ &= \exp[-\frac{n}{2}[(\bar{X}-\mu_0)'\Sigma_0^{-1}(\bar{X}-\mu_0)]]\end{aligned}$$

对数变换后有

$$-2\ln\lambda = n(\bar{X}-\mu_0)'\Sigma_0^{-1}(\bar{X}-\mu_0) \triangleq \xi$$

当原假设成立时有$\sqrt{n}(\bar{X}-\mu_0) \sim N_p(0,\Sigma_0)$, 再由教材54页结论1即可得

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

3-6

不妨令$Y_{(\alpha)} = CX_\alpha(\alpha = 1,2,\cdots,n)$, 则由题设可知

$$Y_\alpha \sim N_k(C\mu, C\Sigma C')$$

由此问题即转化为了协方差阵未知时均值向量的检验，故由教材68页的结论可知检验统计两可取为

$$\begin{aligned}F &= \frac{n-k}{(n-1)k}T^2 \\ T^2 &= (n-1)n(C\bar{X}-r)'[CAC']^{-1}(C\bar{X}-r)\end{aligned}$$

且当原假设成立时有$F \sim F(k,n-k)$, 这里$A = \sum_{i=1}^{n}(X_i-\bar{X})(X_i-\bar{X})'$为样本离差阵。

3-7

本题与上一小题类似，并利用教材72页所给结论可知，检验H_0的似然比统计量为

$$F = \frac{n-p+1}{(n-1)(p-1)}T^2$$
$$T^2 = (n-1)n(C\bar{X})'[CAC']^{-1}(C\bar{X})$$

这里A样本离差阵，且当原假设成立时有检验统计量的分布为$F \sim F(p-1, n-p+1)$。

3-8

由题设可知，原假设即为$\mu_1 : \mu_2 : \mu_3 = 6 : 4 : 1$，将其转化为$C\mu = O$的形式，即解方程组

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

解得$C = \begin{pmatrix} 1 & 0 & -6 \\ 0 & 1 & -4 \end{pmatrix}$, 当然解不唯一。故检验的假设为

$$H_0 : C\mu = O, H_1 : C\mu \neq O$$

同样与3-6类似，参考其结论可取检验统计量为

$$F = \frac{n-2}{2(n-1)}T^2$$
$$T^2 = (n-1)n(C\bar{X})'[CAC']^{-1}C\bar{X}$$

且当原假设成立时，$F \sim F(2, n-2)$, 再结合题中所给数据得$p = 3, n = 6$, 代入数据计算得

$$T^2 = 47.1434, F = 18.8574, p - value = 0.009195 < 0.05$$

故在显著性水平0.05条件下拒绝原假设，即认为这组数据与人类的一般规律不一致。此外本题也可采用R语言直接求解(借助ICSNP包), 参考代码及输出结果如下

```
> data<-read.table("308.txt",head=F);
> data<-as.matrix(data);
> data1<-data[1:6,];
> library("ICSNP");
> C<-t(matrix(c(1,0,0,1,-6,-4),2,3));
> X<-data1%*%C;
> HotellingsT2(X,mu=c(0,0));

        Hotelling's one sample T2-test

data:  X
T.2 = 18.8574, df1 = 2, df2 = 4, p-value = 0.009195
alternative hypothesis: true location is not equal to c(0,0)
```

3-9

记总体为X，设$X_{(\alpha)}(\alpha = 1, \cdots, n)$为来自总体$X$的样本，且这里$n > p$, 由定义可知似然比统计量为

$$\lambda = \frac{\max\limits_{\mu} L(\mu, \Sigma_0)}{\max\limits_{\mu,\Sigma} L(\mu, \Sigma)}$$

对于分子，由第二章中相关结论可知$\mu = \bar{X}$, 故

$$\begin{aligned}
\max_{\mu} L(\mu, \Sigma_0) &= \frac{1}{|2\pi\Sigma_0|^{\frac{n}{2}}} \exp[-\frac{1}{2}\sum_{\alpha=1}^{n}(X_{(\alpha)} - \bar{X})'\Sigma_0^{-1}(X_{(\alpha)} - \bar{X})] \\
&= \frac{1}{|2\pi\Sigma_0|^{\frac{n}{2}}} \exp[-\frac{1}{2}\mathrm{tr}[\Sigma_0^{-1}\sum_{\alpha=1}^{n}(X_{(\alpha)} - \bar{X})(X_{(\alpha)} - \bar{X})']] \\
&= (2\pi)^{\frac{np}{2}}|\Sigma_0|^{-\frac{n}{2}} \exp[-\frac{1}{2}\mathrm{tr}[\Sigma_0^{-1}A]]
\end{aligned}$$

对于分母有

$$\max_{\mu,\Sigma} L(\mu,\Sigma) = L(\bar{X},\frac{1}{n}A) = (\frac{n}{2\pi e})^{\frac{np}{2}}|A|^{-\frac{n}{2}}$$
$$= (2\pi)^{-\frac{np}{2}}(\frac{n}{e})^{\frac{np}{2}}|A|^{-\frac{n}{2}}$$

由此即可得到

$$\lambda = \frac{|\Sigma_0|^{-\frac{n}{2}}\exp[-\frac{1}{2}\mathrm{tr}[\Sigma_0^{-1}A]]}{(\frac{n}{2})^{\frac{np}{2}}|A|^{-\frac{n}{2}}}$$
$$= (\frac{e}{n})^{\frac{np}{2}}\exp(-\frac{1}{2}\mathrm{tr}(\Sigma_0^{-1}A))|\Sigma_0^{-1}A|^{\frac{n}{2}}$$

再由定理3.2.1可知，当样本量充分大时，有

$$-2\ln\lambda \sim \chi^2(\frac{p(p+1)}{2})$$

3-10

设$X_{(\alpha)}^{(1)}$,$X_{(\alpha)}^{(2)}$分别为来自两p维正态总体的简单随机样本，则由似然比检验的定义可知检验统计量为

$$\lambda = \frac{\max\limits_{\mu,\Sigma>0} L(\mu,\Sigma)}{\max\limits_{\mu^{(1)},\mu^{(2)},\Sigma>0} L(\mu^{(1)},\mu^{(2)},\Sigma)}$$

再令$\bar{X}^{(1)}$,$\bar{X}^{(2)}$分别为两总体样本均值向量，A_1, A_2 分别为两总体的样本离差阵，$\bar{X}, T$为全部样本的均值向量与样本离差阵

$$T = A_1 + A_2 + \sum_{i=1}^{2} n_i(\bar{X}^{(i)} - \bar{X})(\bar{X}^{(i)} - \bar{X})' \triangleq A + B$$

这里A即为组内离差阵，B即为组间离差阵。考虑分子的取值，显然当$\mu = \bar{X}, \Sigma = \frac{T}{n}$ 时取到极大值，故代入整理得

$$\max_{\mu,\Sigma>0} L(\mu,\Sigma) = (2\pi)^{-\frac{np}{2}}(\frac{n}{e})^{\frac{np}{2}}|T|^{-\frac{n}{2}}$$

再考虑分母，当$\mu^{(1)} = \bar{X}^{(1)}, \mu^{(2)} = \bar{X}^{(2)}, \Sigma = \frac{A}{n}$ 时取到极大值，故同样代入整理得

$$\max_{\mu^{(1)},\mu^{(2)},\Sigma>0} L(\mu^{(1)},\mu^{(2)},\Sigma) = (2\pi)^{-\frac{np}{2}}(\frac{n}{e})^{\frac{np}{2}}|A|^{-\frac{n}{2}}$$

故似然比统计量整理为

$$\lambda = (\frac{|A|}{|A+B|})^{\frac{n}{2}} = \Lambda^{\frac{n}{2}}$$

可以考虑进一步化简(也可直接利用Wilks分布临界值表)，利用教材64页结论1可将Λ 分布转化为F 分布(因为只有两总体故这里对应$n_2 = 1$)，即

$$F = \frac{n-p-1}{p}\frac{1-\Lambda}{\Lambda}$$

且当原假设成立时有$F \sim F(p, pn-p-1)$。

3-11

本题为两总体均值向量(协方差阵相等但未知)的检验，由教材77页的结论可知这里检验统计两可取为

$$F = \frac{n+m-p-1}{(n+m-2)}T^2$$
$$T^2 = (n+m-2)\frac{mn}{m+n}(\bar{X}-\bar{Y})'[A_1+A_2]^{-1}(\bar{X}-\bar{Y})$$

且在原假设成立时有$F \sim F(p, n+m-p-1)$, 再结合题中所给数据中这里$p = 3, n = 6, m = 9$，最后代入数据计算得

$$T^2 = 5.3117, F = 1.4982, p-value = 0.2693 > 0.05$$

故无法拒绝原假设，即H_0相容。同样本题也可直接采用R语言求解，参考代码及输出结果如下：

```
> data<-read.table("308.txt",head=F);
> data<-as.matrix(data);
> library("ICSNP");
> X1<-data[1:6,];
> X2<-data[7:15,];
> HotellingsT2(X1,X2);
```

```
        Hotelling's two sample T2-test

data:  X1 and X2
T.2 = 1.4982, df1 = 3, df2 = 11, p-value = 0.2693
alternative hypothesis:
true location difference is not equal to c(0,0,0)
```

3-12

(1)由于协方差检验的函数已经被移出CRAN，因此函数的源代码可以R-bloggers 上下载，同时在附录二中也已附上函数源代码，检验的参考代码及输出结果如下：

```
> data<-read.table("312.txt",head=F);
> data<-as.matrix(data);
> data1<-data[1:5,];
> data2<-data[6:9,];
> data3<-data[10:13,];
> s1<-cov(data1);
> s2<-cov(data2);
> s3<-cov(data3);
> covmat<-list(s1,s2,s3);
> varcomp(covmat,n=c(5,4,4));

        Equality of Covariances Matrices Test

data:  covmat
corrected lambda* = 13.6755, df = 12, p-value = 0.4185
```

由于p值明显大于显著性水平0.05，故无法拒绝原假设，即认为H_0相容。

(2)由第一小题的结论可知，可作协方差阵相等假设，参考代码及输出结果如下

```
> data<-read.table("312.txt",head=F);
> data<-as.matrix(data);
> library("ICSNP");
> X1<-data[1:5,];
> X2<-data[6:9,];
> HotellingsT2(X1,X2);

        Hotelling's two sample T2-test

data:  X1 and X2
T.2 = 32.0989, df1 = 3, df2 = 5, p-value = 0.001083
alternative hypothesis:
true location difference is not equal to c(0,0,0)
```

由于p值远远小于显著性水平0.05，故拒绝原假设，接受备择假设即A和B两地区岩石的化学成分有显著差异。

(3)即多元方差分析，可通过R中stats包自带的manova函数实现，参考代码及输出结果如下

```
> data<-read.table("312.txt",head=F);
> manova.data <- data.frame(group = as.factor(rep(1:3,c(5,4,4))),
+ y1=data[,1], y2=data[,2],y3=data[,3]);
> with(manova.data, tapply(y1, group, mean));
     1      2      3
47.472 54.380 42.110
> with(manova.data, tapply(y2, group, mean));
    1     2     3
```

```
5.604 4.470 9.825
> with(manova.data, tapply(y3, group, mean));
     1      2      3
0.1440 0.1525 0.0475
> m1 <- manova(cbind(y1,y2,y3) ~ group, manova.data);
> summary(m1, test = "Wilks");
          Df    Wilks approx F num Df den Df    Pr(>F)
group      2 0.016038    18.39      6     16 2.345e-06 ***
Residuals 10
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
0.1 ‘ ’
```

根据输出结果可以看到$\Lambda = 0.016038, F = 18.39$，且p值远远小于显著性水平0.05，因此拒绝原假设，接受备择假设即A,B,C三个地区的岩石存在显著差异。

(4)可利用教材93页给出的方法编写脚本求解，参考代码及输出结果如下

```
> data<-read.table("312.txt",head=F);
> data<-as.matrix(data);
> A1<-4*cov(data[1:5,]);
> A2<-3*cov(data[6:9,]);
> A3<-3*cov(data[10:13,]);
> A<-A1+A2+A3;
> B<-diag(A);
> b<-13-1.5-(3^3-3)/(3*(3^2-3));
> f<-0.5*(3*4-2*3);
> xi<--b*log(det(A)/prod(B));
> p<-pchisq(xi,f,lower.tail=F);
```

```
> b;f;xi;p;
[1] 10.16667
[1] 3
[1] 3.265038
[1] 0.3525384
```

根据结果可知p值为0.3525明显大于显著性水平0.05，故无法拒绝原假设，即H_0相容(但不能得出相互独立的结论，其余不拒绝原假设的题同理)。

3-13

(1)在R自带的stats包中的shapiro.test()可用于做Shapiro-Wilk检验，其余大多用于一元正态性检验的R函数可在nortest包中获取，参考代码及输出结果如下(仅第一组第一个分量)

```
> data<-read.table("313.txt",head=F);
> data<-as.matrix(data);
> data1<-data[,1];
> shapiro.test(data1);#Shapiro-Wilk检验
        Shapiro-Wilk normality test
data:  data1
W = 0.9444, p-value = 0.2898
> library("nortest");
> lillie.test(data1)#进行Kolmogorov-Smirnov检验。
        Lilliefors (Kolmogorov-Smirnov) normality test
data:  data1
D = 0.1498, p-value = 0.2822
> ad.test(data1)#进行Anderson-Darling正态性检验。
        Anderson-Darling normality test
data:  data1
```

```
A = 0.4888, p-value = 0.1974
> cvm.test(data1)#进行Cramer-von Mises正态性检验。
        Cramer-von Mises normality test
data:  data1
W = 0.0869, p-value = 0.1584
> pearson.test(data1)#进行Pearson卡方正态性检验。
        Pearson chi-square normality test
data:  data1
P = 8, p-value = 0.09158
> sf.test(data1)#进行Shapiro-Francia正态性检验
        Shapiro-Francia normality test
data:  data1
W = 0.9538, p-value = 0.3639
```

根据检验p值(> 0.05)可以认为第一组第一分量无法拒绝原假设(正态假设)。其余各组各分量均同理可得。

(2)多元正态分布的检验可借助MVN包实现，包中提供了三种正态检验函数，函数中均提供了plot参数，选择TRUE(缺省)即在提供检验结果的同时绘制卡方(qq)图。参考代码及输出结果如下(仅第一组，其余组方法完全相同)：

```
data<-read.table("313.txt",head=F);
data<-as.matrix(data);
data1<-data[,1:4];
library("MVN");
par(bg="lightgrey",font.lab=2,font.axis=2)
mardia.test(data1);
#HZ.test(data1);
#ston.test(data1);
```

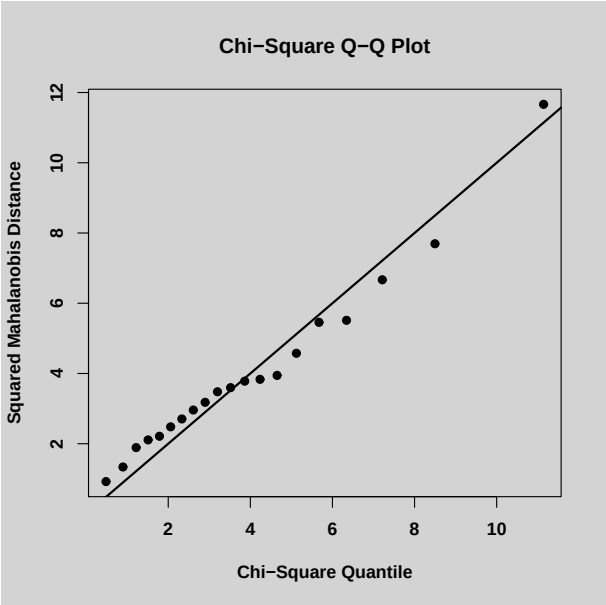


图 12:

此外与多元正态性检验相关的包还有mvnormtest、mvsf、ICS、mvShapiroTest等。

第五章 判别分析

5-1

(1)根据题设可得

$$\begin{aligned}
P(2|1) &= P(x \le \mu^*|x \sim N(\mu^{(1)}, \sigma_1^2)) + P(x \le \mu_*|x \sim N(\mu^{(1)}, \sigma_1^2)) \\
&= P(\frac{x-\mu^{(1)}}{\sigma_1} \le \frac{\mu^*-\mu^{(1)}}{\sigma_1}) + P(\frac{x-\mu^{(1)}}{\sigma_1} \ge \frac{\mu_*-\mu^{(1)}}{\sigma_1}) \\
&= \Phi(\frac{\mu^*-\mu^{(1)}}{\sigma_1}) + [1-\Phi(\frac{\mu_*-\mu^1}{\sigma_1})] \\
&= \Phi(\frac{\mu^{(2)}-\mu^{(1)}}{\sigma_1}) + [1-\Phi(\frac{\mu_{(1)}-\mu^{(2)}}{\sigma_1})] \\
&= \Phi(\frac{\mu^{(2)}-\mu^{(1)}}{\sigma_1+\sigma_2}) + \Phi(\frac{\mu_{(2)}-\mu^{(1)}}{\sigma_2-\sigma_1})
\end{aligned}$$

(2)与第一小题类似，可以得到

$$\begin{aligned}
P(1|2) &= P(\mu^* < X < \mu_*|X \sim N(\mu^{(2)}, \sigma_2^2)) \\
&= P(\frac{X-\mu^{(2)}}{\sigma_2} < \frac{\mu_*-\mu^{(2)}}{\sigma_2}) - P(\frac{X-\mu^{(2)}}{\sigma_2} \le \frac{\mu^*-\mu^{(2)}}{\sigma_2}) \\
&= P(\frac{X-\mu^{(2)}}{\sigma_2} < -\frac{\mu^{(1)}-\mu^{(2)}}{\sigma_1-\sigma_2}) - P(\frac{X-\mu^{(2)}}{\sigma_2} \le \frac{\mu^{(1)}-\mu^{(2)}}{\sigma_1+\sigma_2}) \\
&= \Phi(\frac{\mu^{(1)}-\mu^{(2)}}{\sigma_2-\sigma_1}) - \Phi(\frac{\mu^{(1)}-\mu^{(2)}}{\sigma_1+\sigma_2})
\end{aligned}$$

5-2

(1)按距离判别准则，根据题中所给数据计算可得

$$\begin{aligned}
d_1^2(x) &= \frac{(x-\mu^{(1)})^2}{\sigma_1^2} = 1 \\
d_2^2(x) &= \frac{(x-\mu^{(2)})^2}{\sigma_2^2} = 1.5625 \\
d_3^2(x) &= \frac{(x-\mu^{(3)})^2}{\sigma_3^2} = 0.25
\end{aligned}$$

由于$d_3^2(x) < d_1^2(x) < d_2^2(x)$，故$x \in G_3$

(2)由题设即可知此事贝叶斯判别即为广义平方距离判别，故由题设可得

$$D_1^2(x) = d_1^2(x) + g_1(1) + g_2(1) = d_1^2(x) + \ln \sigma_1^2 = -0.3863$$
$$D_2^2(x) = d_2^2(x) + g_1(2) + g_2(2) = d_2^2(x) + \ln \sigma_2^2 = 2.9488$$
$$D_3^2(x) = d_3^2(x) + g_1(3) + g_2(3) = d_3^2(x) + \ln \sigma_3^2 = 0.25$$

由于$D_1^2(x) < D_3^2(x) < D_2^2(x)$，故$x \in G_1$

5-3

(1)即证$E(a'X|G_1) - \bar{\mu} > 0$，有题设可知

$$\begin{aligned} E(a'X|G_1) - \bar{\mu} &= a'\mu^{(1)} - \frac{1}{2}(a'\mu^{(1)} + a'\mu^{(2)}) \\ &= \frac{1}{2}(a'\mu^{(1)} - a'\mu^{(2)}) \\ &= \frac{1}{2}(\mu^{(1)} - \mu^{(2)})'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) \end{aligned}$$

显然$\Sigma > 0$，故$E(a'X|G_1) - \bar{\mu} > 0$，即$E(a'X|G_1) > \bar{\mu}$得证。

(2)与第一小题类似，有

$$\begin{aligned} E(a'X|G_2) - \bar{\mu} &= a'\mu^{(2)} - \frac{1}{2}(a'\mu^{(1)} + a'\mu^{(2)}) \\ &= -\frac{1}{2}(a'\mu^{(1)} - a'\mu^{(2)}) \\ &= -\frac{1}{2}(\mu^{(1)} - \mu^{(2)})'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) \end{aligned}$$

同样由$\Sigma > 0$，得$E(a'X|G_1) - \bar{\mu} > 0$，即$E(a'X|G_2) < \bar{\mu}$得证。

5-4

(1)按Fisher准则判别，由题设组内离差阵与组间离差阵分别为

$$A = \Sigma_1 + \Sigma_2 = \begin{pmatrix} 18 & 12 \\ 12 & 32 \end{pmatrix} + \begin{pmatrix} 20 & -7 \\ -7 & 5 \end{pmatrix} = \begin{pmatrix} 38 & 5 \\ 5 & 37 \end{pmatrix}$$

$$B = (\mu^{(1)} - \mu^{(2)})(\mu^{(1)} - \mu^{(2)})' = \begin{pmatrix} 100 & 100 \\ 100 & 100 \end{pmatrix}$$

考虑$A^{-1}B$的特征值，结合代数知识可知其非零特征值等价于

$$\begin{aligned} d^2 &= (\bar{X}^{(1)} - \bar{X}^{(2)})' A^{-1} (\bar{X}^{(1)} - \bar{X}^{(2)}) \\ &= \frac{6500}{1381} = 4.7067 \end{aligned}$$

取$a = \frac{1}{d} A^{-1}(\mu^{(1)} - \mu^{(2)}) =$, 满足方程$a'Aa = 1$，故可得线性判别函数为

$$u(X) = a'X = \frac{-1}{89765}(32X_1 + 33X_2)$$

又$Ba = d^2 Aa$, 故判别效率为

$$\Delta(a) = \frac{a'Ba}{a'Aa} = \lambda = d^2 = 4.7067$$

判别准则为

$$\begin{cases} X \in G_1 & u(X) > u^* \\ X \in G_2 & u(X) \leq u^* \end{cases}$$

计算阈值u^*，由于

$$\begin{aligned} \sigma_1^2 &= a'\Sigma_1 a = \frac{78624}{89765} = 0.8759 \\ \sigma_2^2 &= a'\Sigma_2 a = \frac{11141}{89765} = 0.1241 \\ \bar{u}^{(1)} &= a'\mu^{(1)} = \frac{-815}{\sqrt{89765}} = -2.7202 \\ \bar{u}^{(2)} &= a'\mu^{(2)} = \frac{-1465}{\sqrt{89765}} = -4.8897 \end{aligned}$$

故

$$\mu^* = \frac{\sigma_2\bar{\mu}^{(1)} + \sigma_1\bar{\mu}^{(2)}}{\sigma_1 + \sigma_2} = -4.2964$$

代入题中所给数据，当$X_{(1)} = (20, 20)'$时，有

$$u(X_{(1)}) = -4.3390 < u^*$$

当$X_{(2)} = (15, 20)'$时，有

$$u(X_{(2)}) = -3.8050 > u^*$$

即$X_{(1)} \in G_2,\ X_{(2)} \in G_1$

(2)按Bayes准则，且此时假设$\Sigma = \Sigma_2 = \Sigma_1$，根据题设有

$$h_1(X) = q_2 L(1|2) f_2(X)$$
$$h_2(X) = q_1 L(2|1) f_1(X)$$

考虑两者大小关系，因先验概率相等，故即

$$\begin{aligned}\frac{h_1(X)}{h_2(X)} &= \frac{L(1|2)f_2(X)}{L(2|1)f_1(X)} \\ &= 7.5 frac f_2(X) f_1(X) \\ &= 7.5\exp(\frac{10}{216}(X-\bar{\mu})'(10,3)')\end{aligned}$$

代入题中所给数据，当$X_{(1)} = (20, 20)'$时，有

$$\frac{h_1(X_{(1)})}{h_2(X_{(1)})} = 75.9229 > 1$$

当$X_{(2)} = (15, 20)'$时，有

$$\frac{h_1(X_{(2)})}{h_2(X_{(2)})} = 7.5 > 1$$

即$X_{(1)} \in G_2,\ X_{(2)} \in G_2$

(3)当$x=(20,20)'$时，有

$$P(G_1|x)=\frac{q_1f_1(x)}{\sum_{i=1}^{2}q_if_i(x)}=\frac{f_1(x)}{f_1(x)+f_2(x)}=0.7306$$

$$P(G_2|x)=\frac{q_2f_2(x)}{\sum_{i=1}^{2}q_if_i(x)}=\frac{f_2(x)}{f_1(x)+f_2(x)}=0.2694$$

5-5

根据题设可知

$$\frac{(a'd)^2}{a'Sa}=\frac{(a'd)(a'd)'}{a'Sa}=\frac{a'(\bar{X}^{(1)}-\bar{X}^{(2)})(\bar{X}^{(1)}-\bar{X}^{(2)})'a}{a'Sa}=\frac{a'Ba}{a'Sa}\leq\lambda_1$$

这里λ_1为$S^{-1}B$的最大特征值，当且仅当a为λ_1对应的特征向量时等号成立。考虑λ_1 的取值，由代数知识可知$S^{-1}B$与$D^2=(\bar{X}^{(1)}-\bar{X}^{(2)})'S^{-1}(\bar{X}^{(1)}-\bar{X}^{(2)})$非零特征值相同。即$\lambda_1=D^2$。故只需证$a$为$D^2$对应的一个特征向量

$$\begin{aligned}S^{-1}Ba&=S^{-1}(\bar{X}^{(1)}-\bar{X}^{(2)})(\bar{X}^{(1)}-\bar{X}^{(2)})'S^{-1}(\bar{X}^{(1)}-\bar{X}^{(2)})\\&=S^{-1}(\bar{X}^{(1)}-\bar{X}^{(2)})\cdot D^2\\&=D^2a\end{aligned}$$

故a为λ_1对应的特征向量，即原命题得证

5-6

(1)考虑线性判别函数$W(X)$的分布，由于$W(X)$是X的线性函数，故当$X\in G_1$时，有$W(X)\sim N_1(\mu_1,\sigma_1^2)$，其中

$$\mu_1=(\mu^{(1)}-\bar{\mu})'a=\frac{1}{2}d^2$$

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

这里d^2为马氏距离，即$d^2=(\mu^{(1)}-\mu^{(2)})'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)})$，故

$$\begin{aligned}P(2|1)&=P(W(X)\le 0|X\in G_1)=P(\frac{W(X)-\mu_1}{\sigma_1}\le\frac{0-\mu_1}{\sigma_1})\\&=P(\frac{W(X)-\mu_1}{\sigma_1}\le-\frac{d^2}{2d})=1-\Phi(\frac{1}{2}d)\end{aligned}$$

(2)方法同第一小题，当$X\in G_2$时，有$W(X)\sim N_1(\mu_2,\sigma_2^2)$，其中$\mu_2=-\frac{1}{2}d^2,\sigma_2^2=d$，故同理可以得到

$$\begin{aligned}P(1|2)&=P(W(X)>0|X\in G_2)=P(\frac{W(X)-\mu_2}{\sigma_2}>\frac{0-\mu_2}{\sigma_2})\\&=P(\frac{W(X)-\mu_2}{\sigma_2}>\frac{d^2}{2d})=1-\Phi(\frac{1}{2}d)\end{aligned}$$

5-7

(1)贝叶斯判别准则，根据教材185页的结论（广义平方距离）即可知判别准则为

$$\begin{cases}X\in G_1, & W(X)>d\\X\in G_2, & W(X)\le d\end{cases}$$

其中这里$W(X)=(X-\bar{\mu})'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)})$，且$\bar{\mu}=\frac{1}{2}(\mu^{(1)}+\mu^{(2)})$，此外这里$d=\ln\frac{q_2L(1|2)}{q_1L(2|1)}$

(2)距离判别准则，根据教材179页结论即可知判别准则为

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

这里$W(X)$同第一小题，可以看到在两总体且协方差阵相等时，若先验概率相等且错判损失相等，则两种判别准则等价。

5-8

由题设可知这里$r=1,k=3,n=30$，参考教材208页中的步骤，计算变量X_1,X_2,X_4在X_3时给定时的判别能力即威尔克斯统计量如

下

$$U_{1|(1)}=\frac{a_{11}^{(1)}}{t_{11}^{(1)}}=\frac{28571.5}{28884.9}=0.989$$
$$U_{2|(1)}=\frac{a_{22}^{(1)}}{t_{22}^{(1)}}=\frac{114.9}{148.3}=0.775$$
$$U_{4|(1)}=\frac{a_{44}^{(1)}}{t_{44}^{(1)}}=\frac{15375.8}{27925.9}=0.551$$

易知$U_{4|(1)}=\min U_{i|(1)}$。再检验X_4在三个总体中是否有显著差异，即检验$H_0:U_{4|(1)}^{(1)}=U_{4|(1)}^{(2)}=U_{4|(1)}^{(3)}$，在$H_0$成立时威尔克斯统计量

$$U_{4|(1)}\sim\Lambda(1,n-k-r,k-1)$$

基于此构造F统计量，当H_0成立时

$$F=\frac{1-U_{4|(1)}}{U_{4|(1)}}\frac{26}{2}\sim F(2,26)$$

代入数据可得此时$F=10.6106$，又$f_{0.05}(2,26)=3.81$(p值小于0.01)，故检验结果显著，即下一步可引入变量X_4

5-9

(1)根据题设首先检验均值差异显著性，采用HotellingT2检验，基于ICSNP包运行如下代码：

```
D509<-read.table("509.txt",head=F);
D5091<-D509[1:7,1:3];#含矿
D5092<-D509[8:14,1:3];#不含矿
D5093<-D509[1:14,];#先剔除待判样本
#均值向量差异性检验
library(ICSNP);
HotellingsT2(D5091,D5092);
```

得到输出

```
        Hotelling's two sample T2-test

data:  D5091 and D5092
T.2 = 3.1089, df1 = 3, df2 = 10, p-value = 0.07557
alternative hypothesis: true location difference
is not equal to c(0,0,0)
```

根据输出结果可知$p = 0.0756 < 0.1$，故在显著性水平$\alpha = 0.10$下两总体的均值向量有显著差异，故此时讨论判别问题是有意义的。再考虑判别，由于此时假设协方差阵相等，故采用线性判别，又基于正态假设，故广义平方距离判别与贝叶斯判别等价，运行如下代码(函数lda函数基于贝叶斯判别)

```
library(MASS);
results<-lda(V4~V1+V2+V3,D5093);
#样本回判
predict(results,D5093)$class;
```

得到输出为

```
 [1] 1 1 1 1 1 2 1 2 2 2 2 2 1 2
Levels: 1 2
```

可知判别结果有两个为错判：含矿的第6号错判为不含矿；而不含矿的第13号矿错判为含矿。

(2)根据第一小题运行结果结合继续运行如下代码

```
D5094<-D509[15,];
predict(results,D5094)$class;
```

得到输出

```
[1] 2
Levels: 1 2
```

根据输出即可知判定结果为不含矿。

5-10

(1)由于类别2的协方差阵奇异，即协方差阵求逆会非常不稳定，故这里认为协方差阵相等(方差分析等步骤可参考5-11，这里略去)，即贝叶斯判别与马氏距离判别一致，运行如下代码(含输出)

```
> D510<-read.table("510.txt",head=F);
> library("MASS");
> results<-lda(V5~V1+V2+V3+V4,D510[1:17,]);
> predict(results,D510[1:17,])$class;
 [1] 1 3 3 1 2 3 3 1 1 1 2 2 1 1 1 3 3
Levels: 1 2 3
> predict(results,D510[18:20,])$class;
[1] 1 3 3
Levels: 1 2 3
> table(D510[1:17,]$V5,predict(results)$class);

   1 2 3
  1 7 0 0
  2 0 3 1
  3 1 0 5
```

根据输出结果可以看到错判了两个观测：第3号原属于类别2的被错判为类别3，第9号原属于类别3的被错判为类别1。待判的3个观测依次被判归为1,3,3类。

由于事实上通过通过检验可以发现三总体协方差阵不相等(采用第三章所给出的方法，这里略过)，故判别结果与教材附录中的参考答案存在出入，读者可自行尝试二次判别(采用SAS软件，或调整R中的tol求逆)。

(2)不妨采用逐步判别法，调用SDDA包运行如下代码

```
library(SDDA);
attach(D510[1:17,]);
A<-as.factor(V5);
MD510<-as.matrix(D510[1:17,1:4]);
s1<-sdda(MD510,A,start =c(FALSE,TRUE,FALSE,TRUE),
         method="lda");
s2<-predict(s1, MD510);
detach(D510);
```

输出结果为

```
> s1;s2;
SDDA using dlda .
n = 3 samples and p = 4 variables.
Group levels are: 1 2 3 .
3 variables are chosen in total.
Variables are V2 V3 V4
 [1] "1" "3" "2" "1" "2" "2" "3" "1" "1" "1"
     "2" "2" "1" "1" "1" "3" "3"
> table(s2,A);
    A
s2   1 2 3
  1 7 0 1
  2 0 4 1
  3 0 0 4
```

根据输出结果可以看到最终选出了$X2, X3, X4$三个变量，共有两个错判，第6号原属类别3被错判为类别2，第9号原属类别3被错判为类别1；又

```
s3<-predict(s1,as.matrix(D510[18:20,1:4]));s3;
```

故待判的三个观测依次被判归为1,2,3类

5-11

(1)同样首先检验均值差异显著性，由于类别大于两种，故采用多元方差分析，运行如下代码：

```
D511<-read.table("511.txt",head=F);
D5111<-D511[D511[,4]==1,1:3];
D5112<-D511[D511[,4]==2,1:3];
D5113<-D511[D511[,4]==3,1:3];
D5110<-rbind(D5111,D5112,D5113);
manova.data<-data.frame(group=as.factor(rep(1:3,c(4,6,4))),
y1=D5110[,1],y2=D5110[,2],y3=D5110[,3]);
m<- manova(cbind(y1,y2,y3) ~ group, manova.data);
```

得到输出结果为

```
> summary(m)
          Df Pillai approx F num Df den Df    Pr(>F)
group      2  1.758   24.212      6     20 3.521e-08 ***
Residuals 11
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
0.1
```

即结果极为显著，故可以认为此时讨论判别是有意义的(严格来说还需要多重比较，考虑到这并非本题重点故略过)，再考虑判别，同样由于假设协方差阵相等，故采用线性判别，先验概率为$q_1 = \frac{2}{7}, q_2 = \frac{3}{7}, q_3 = \frac{2}{7}$，故运行如下代码(含输出结果)

```
> library("MASS");
> results<-lda(V4~V1+V2+V3,D511[1:14,],prior=c(2,3,2)/7);
```

```
> #样本回判
> predict(results,D511[1:14,])$class;
 [1] 2 2 2 3 3 1 2 1 1 2 3 3 1 2
Levels: 1 2 3
> table(D511[1:14,]$V4,predict(results)$class)

   1 2 3
 1 4 0 0
 2 0 6 0
 3 0 0 4
```

根据输出结果可知14个监测点全部判对。事实上还可以根据输出对判别结果进行可视化，为兼顾方便和美观，这里采用ggplot2绘图，运行如下代码

```
library(ggplot2);
Data<-cbind(D5110,V4=matrix(c(rep("A",4),rep("B",6),
      rep("C",4)),ncol=1))
results<-lda(V4~V1+V2+V3,Data,prior=c(2,3,2)/7);
ld=predict(results)$x;
p=ggplot(cbind(Data,as.data.frame(ld)),aes(x=LD1,y=LD2))
p+geom_point(aes(colour=V4),alpha=0.8,size=4)
```

即可得到下图，其思想类似之后章节介绍的主成分分析

(2)由于并没有协方差阵相等的假设，故这里考虑平方判别，运行如下代码

```
> results2<-qda(V4~V1+V2+V3,D511[1:14,]);
> predict(results2,D511[15:16,])$class;
[1] 2 2
Levels: 1 2 3
```

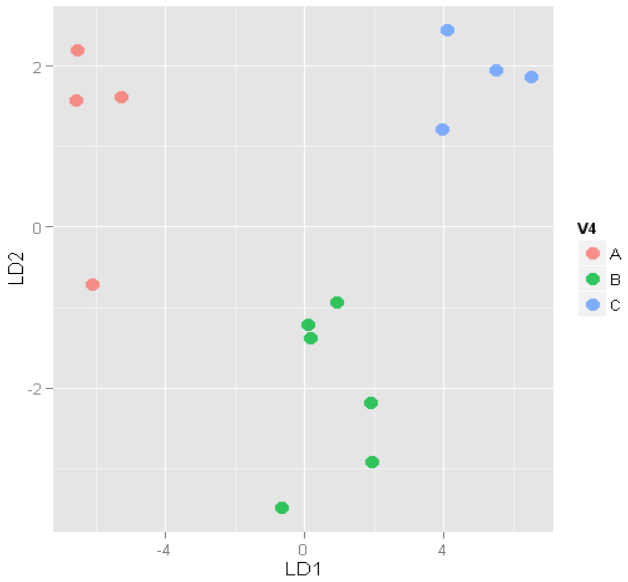


图 13: 判别可视化

可知待判的两个单位均被判归为第2类，教材附录中所给的参考答案应为采用线性判别后的结果，读者可自行尝试。[1]

[1]请务必分清教材中所介绍的各种判别方法间的关系(如何时两种方法等价)，以便使用正确的函数求解

第六章 聚类分析

6-1

(1)只需证明新组成的函数满足的三个要求即可，不妨设$d_{ij}^{(1)},d_{ij}^{(2)}$为两种任意的距离函数，则有

$$
\begin{aligned}
&d_{ij}=d_{ij}^{(1)}+d_{ij}^{(2)}\geq 0 \quad \text{当且仅当}X_i=X_j\text{时}d_{ij}=0\\
&d_{ij}=d_{ij}^{(1)}+d_{ij}^{(2)}=d_{ji}^{(1)}+d_{ji}^{(2)}=d_{ji} \quad \forall i,j\\
&d_{ij}=d_{ij}^{(1)}+d_{ij}^{(2)}\leq d_{ik}^{(1)}+d_{kj}^{(1)}+d_{ik}^{(2)}+d_{kj}^{(2)}=d_{ik}+d_{kj} \quad \forall i,j,k
\end{aligned}
$$

故命题得证

(2)令d_{ij}为任一距离函数，a为任一正常数，令$D_{ij}=ad_{ij}$，则只需证D_{ij}满足距离的三个要求即可，事实上

$$
\begin{aligned}
&D_{ij}=ad_{ij}\geq 0 \quad \text{当且仅当}X_i=X_j\text{时}D_{ij}=0\\
&D_{ij}=ad_{ij}=ad_{ji}=D_{ji} \quad \forall i,j\\
&D_{ij}=ad_{ij}\leq a(d_{ik}+d_{kj})=ad_{ik}+ad_{kj}=aD_{ik}+aD_{kj} \quad \forall i,j,k
\end{aligned}
$$

故命题得证

(3)类似地，只需要d_{ij}^{*}满足距离的三个要求即可，同样容易得到

$$
\begin{aligned}
&d_{ij}^{*}=\frac{d_{ij}}{d_{ij}+c}\geq 0 \quad \text{当且仅当}X_i=X_j\text{时}d_{ij}^{*}=0\\
&d_{ij}^{*}=\frac{d_{ij}}{d_{ij}+c}=\frac{d_{ji}}{d_{ji}+c}=d_{ji}^{*} \quad \forall i,j\\
&d_{ij}^{*}=\frac{1}{1+c/d_{ij}}\leq\frac{1}{1+c/(d_{ik}+d_{kj})}=\frac{d_{ik}+d_{kj}}{d_{ik}+d_{kj}+c}\\
&\quad=\frac{d_{ik}}{d_{ik}+d_{kj}+c}+\frac{d_{kj}}{d_{ik}+d_{kj}+c}\\
&\quad\leq\frac{d_{ik}}{d_{ik}+c}+\frac{d_{kj}}{d_{kj}+c}=d_{ik}^{*}+d_{kj}^{*}
\end{aligned}
$$

故命题得证

(4)此时不一定满足距离的第三个条件，可通过反例说明。

6-2

不妨设变量X_i和X_j为两二值变量(0-1变量)，它们的n次观测值记为$x_{ti}, x_{tj}(t = 1, \cdots, n)$。由二值变量的列联表可知变量$X_i = 1$的观测次数为$a + b$，$X_i = 0$的观测次数为$c + d$; 变量$X_i$和$X_j$取值均为1的观测次数为$a$, 取值均为0的观测次数为$d$。即

	1	0	行和
1	a	b	$a+b$
0	c	d	$c+d$
列和	$a+c$	$b+d$	$a+b+c+d$

(1)利用两定量变量相关系数的公式

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

分别考虑分子分母的化简，对于分子有

$$\begin{aligned}\sum_{t=1}^{n}(x_{ti} - \bar{x}_i)(x_{tj} - \bar{x}_j) &= \sum_{t=1}^{n} x_{ti}x_{tj} - n\bar{x}_i\bar{x}_j \\ &= a - n\frac{a+b}{n}\frac{a+c}{n} = \frac{an - (a+b)(a+c)}{n} \\ &= \frac{a(a+b+c+d) - (a+b)(a+c)}{n} \\ &= \frac{ad - bc}{n}\end{aligned}$$

对于分母有

$$\begin{aligned}\sum_{t=1}^{n}(x_{ti} - \bar{x}_i)^2 &= \sum_{t=1}^{n} x_{ti}^2 - n\bar{x}_i^2 \\ &= a + b - n(\frac{a+b}{n})^2 \\ &= \frac{(a+b)(c+d)}{n}\end{aligned}$$

同理可得$\sum_{t=1}^{n}(x_{tj}-\bar{x}_j)^2=\frac{(a+c)(b+d)}{n}$，由此即可得到

$$C_{ij}(7)=\frac{ad-bc}{\sqrt{(a+b)(c+d)}\sqrt{(a+c)(b+d)}}$$

故命题得证

(2)同样利用两定量变量夹角余弦的定义式

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

可以看到$\sum_{t=1}^{n}x_{ti}x_{tj}=a,\ \sum_{t=1}^{n}x_{ti}=a+b,\ \sum_{t=1}^{n}x_{tj}=a+c$，故代入即有

$$C_{ij}(9)=\cos\alpha_{ij}=\frac{a}{\sqrt{(a+b)(a+c)}}$$

6-3

(1)最长距离法：根据题设所给的距离阵可知先合并$\{X_{(1)},X_{(4)}\}=CL4$，并类距离$D_1=1$，并类后的平方距离矩阵为

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

再合并$\{X_{(2)},X_{(5)}\}=CL3$，并类距离$D_2=3$，并类后新的平方距离矩阵为

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

继续合并$\{CL3CL4\} = CL2$，并类距离$D_3 = 8$，并类后新的平方距离矩阵为

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

最后将所有样品合并为一类$CL1$，并类距离$D_4 = 10$

(2)类平均法，方法完全一致，先合并$\{X_{(1)}, X_{(4)}\} = CL4$，并类距离$D_1 = 1$，并类后的平方距离矩阵为

$$D^{(2)} = \begin{pmatrix} 0 & & & \\ 9^2 & 0 & & \\ 3^2 & 5^2 & 0 & \\ \frac{65}{2} & \frac{136}{2} & 50 & 0 \end{pmatrix} \begin{matrix} X_{(2)} \\ X_{(3)} \\ X_{(5)} \\ CL4 \end{matrix}$$

再合并$\{X_{(2)}, X_{(5)}\} = CL3$，并类距离$D_2 = 3$，并类后新的平方距离矩阵为

$$D^{(3)} = \begin{pmatrix} 0 & & \\ \frac{136}{2} & 0 & \\ \frac{106}{2} & \frac{165}{4} & 0 \end{pmatrix} \begin{matrix} X_{(3)} \\ CL4 \\ CL3 \end{matrix}$$

继续合并$\{CL3CL4\} = CL2$，并类距离$D_3 = \sqrt{\frac{165}{4}}$，并类后新的平方距离矩阵为

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

最后将所有样品合并为一类$CL1$，并类距离$D_4 = \sqrt{\frac{121}{2}}$

6-4

第L次合并G_P和G_q为新类G_r，并类距离$D_L = D_{pq}$，并且必有$D_{pq}^2 \leq D_{ij}^2$，考虑题中所给的递推公式，事实上当$\gamma = 0, \alpha_p \geq$

$0, \alpha_q \geq 0, \alpha_p + \alpha_q + \beta \geq 1$时，容易得到

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

即聚类方法具有单调性，故这里只需考虑各聚类方法的$\gamma, \alpha_q, \alpha_q, \alpha_p + \alpha_q + \beta$的取值情况即可。

(1)类平均法

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

故类平均法具有单调性，得证

(2)可变类平均法

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

故可变类平均法具有单调性，得证

(3)可变法

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

故可变法也具有单调性，得证

(4)Ward法

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

故Ward法同样具有单调性，得证

6-5

(1)考虑最短距离法的单调性，设第L步从类间距离矩阵$D^{(L-1)}=(D_{ij}^{(L-1)})$出发，假设

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

故G_p与G_q合并为新类G_r，此时第L步的并类距离为

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

且由递推公式可知新类与其它类的距离为

$$D_{rk}^{(L)}=\min(D_{pk}^{(L-1)},D_{qk}^{(L-1)})\geq D_{pq}^{(L-1)}=D_{(L)}$$

这里$k=\neq p,q$。同样地设第$L+1$步从类间距离阵$D^{(L)}=(D_{ij}^{(L)})$出发，由于

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

同样$i,j\neq r,p,q$，故第$L+1$步的并类距离为

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

即单调性得证

(2)同理可证最长距离也具有单调性

6-6

(1)中间距离法：不妨取$\beta=\frac{1}{4}$，此时递推公式为

$$D_{rk}^2=\frac{1}{2}(D_{pk}^2+D_{qk}^2)-\frac{1}{4}D_{pq}^2$$

根据题设可知此时最小距离为$D_1=1$，故合并B,C记为CL_2，此时A与新类的类间距离为

$$D^2=\frac{1}{2}(D_{AB}^2+D_{AC}^2)-\frac{1}{4}D_{BC}^2=0.85$$

即$D_2=\sqrt{0.85}=0.92<1=D_1$，不满足单调性

(2)重心法：事实上由于重心法的递推公式为

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

这里$k\neq p,q$，可以看到当$n_p=1,n_q=1,n_r=2$时，上式等价于第一小题中的中间距离法递推公式，故采用相同的方法即可证明。

6-7

暂缺，可参考网上版本

6-8

暂缺，可参考网上版本

6-9

(1)由题设条件计算样品间的欧式平方距离矩阵得

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

(2)故合并$\{1,2\}=CL4$，并类聚类$D_1=\sqrt{0.5}=0.707$，进一步利用Ward法递推公式计算并类后的平方距离矩阵得

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

(3)根据上一步的结果进一步合并$\{5,7\}=CL3$，并类距离$D_2=\sqrt{2}=1.414$，同样利用Ward法递推公式计算并类后的平方距离矩阵得

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

(4)同样根据上一步的结果进一步合并$\{CL3,10\}=CL2$，并类距离$D_3=\sqrt{\frac{32}{3}}=3.266$，并类后平方距离矩阵为

$$D^{(4)}=\begin{pmatrix} 0 & \\ \frac{245}{6} & 0 \end{pmatrix}\begin{matrix} CL2 \\ CL4 \end{matrix}$$

(5)合并$\{CL4,CL2\}=\{1,2,5,7,10\}=CL1$，并类距离$D_4=\sqrt{\frac{245}{6}}=6.39$

(6)故综上所述可知各分类法b_k及相应的总离差平方和W_k如下表所示

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

6-10

(1)分别采用类平均法和Ward法进行系统聚类并绘制谱系聚类图，系统聚类可通过hclust()函数实现[2]，运行如下代码

```
D610<-read.table("610.txt",header=F);
name<-c("Ag","Al","Cu","Ca","Sb","Bi","Sn");
colnames(D610)<-name;
```

[2]系统聚类法还可通过cluster包中的agnes()函数实现，相比而言输出更丰富也更为灵活

```
dist1<-dist(D610);
hcaverage<-hclust(dist1,"average");
hcward<-hclust(dist1,"ward");
#dendrogram
plot(hcaverage,hang=-1);#The simplest way
#make the dendrogram much more beautiful
dev.new()
par(bg="#DDE3CA")
plot(hcward,col="#487AA1",col.main="#45ADA8",
     col.lab="#7C8071", col.axis="#F38630",lwd=3,
     lty=3,sub="",hang=-1,axes=FALSE)
#add axis
axis(side=2, at=seq(0,15000,5000),col="#F38630",
     labels=FALSE,lwd=2)
#add text in margin
mtext(seq(0,15000,5000),side=2,at=seq(0,15000,5000),
     line=1,col="#A38630",las=2)
```

绘制的谱系聚类图(对右图做了适当的美化，供参考)如下图所示

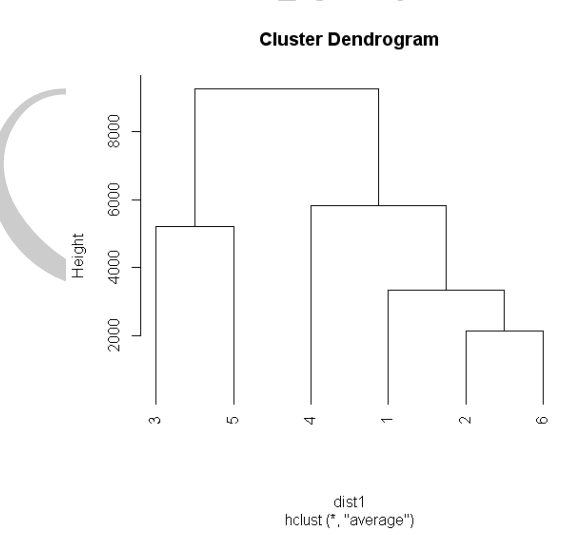


图 14: 类平均法

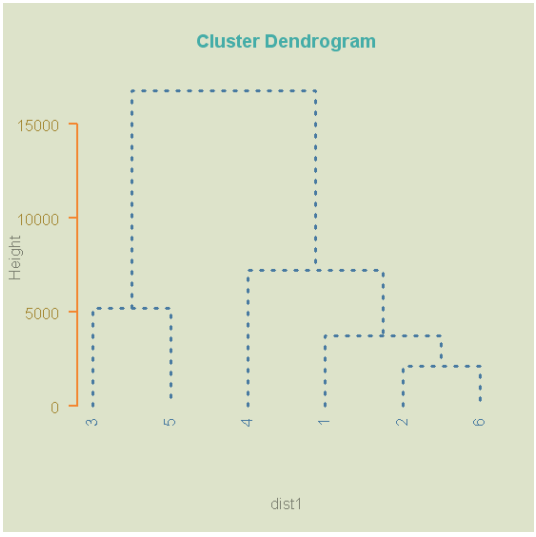


图 15: Ward法

由图可以看到两种方法的聚类结果基本一致，若将弹头分为三类，则结果均为第1,2,6号为一类，第4号单独为1类，第3,5号为一类。

(2)将$d_{ij}^2 = 1 - c_{ij}^2$定义为变量间距离，则可通过如下代码实现

```
D610<-read.table("610.txt",header=F);
name<-c("Ag","Al","Cu","Ca","Sb","Bi","Sn");
colnames(D610)<-name;
C610<-cor(D610);
D6102<-(1-(abs(C610))^2)^0.5;
dvar<-dist(D6102);
hcvaraver<-hclust(dvar,"average");
hcvarward<-hclust(dvar,"ward");
#dendrogram objects
hcdaver<-as.dendrogram(hcvaraver);
hcdward<-as.dendrogram(hcvarward);
par(lwd=3,lty=3);
plot(hcdaver,type="triangle",axes=F);
axis(side=2,at=seq(0,1.5,0.5),col="#F38630",
     labels=FALSE,lwd=2);
dev.new()
par(lwd=3);
plot(hcdward,type="triangle",axes=F);
axis(side=2,at=seq(0,1.5,0.5),col="#F38630",
     labels=FALSE,lwd=2);
```

绘制的谱系聚类图如下图所示，将对象的类型修改为dendrogram可以在调用泛型函数plot得到更灵活的图表类型

根据上图可以看到两种方法的分类结果稍有不同，如同样将变量分为三类，则类平均法将Ag和Bi分为第一类，将Sb 单独分为一

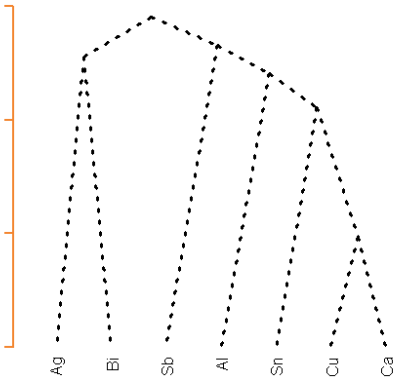


图 16: 类平均法

图 17: Ward法

类，将Al、Sn、Cu、Ca分为另一类；Ward法则将Ag，Bi，Sn分为第一类，Sb单独一类，Al、Cu、Ca分为另一类。

此外也可以通过热图的形式直观的观察样本与变量的聚类情况，运行如下代码[3]

```
library(RColorBrewer)
heatmap(as.matrix(D610),col=brewer.pal(9,"RdYlGn"),
        scale="column",margins=c(4,8));
```

6-11

分别采用类平均法和Ward法进行系统聚类并绘制谱系聚类图，运行如下代码

```
D611<-read.table("611.txt",header=F);
name<-c("Cu","Ag","Bi");
observation<-as.character(1:14);
rownames(D611)<-observation;
colnames(D611)<-name;
dist2<-dist(D611);
```

[3]热图效果的好坏一定程度上取决于调色盘的选取合适与否，读者可自行尝试

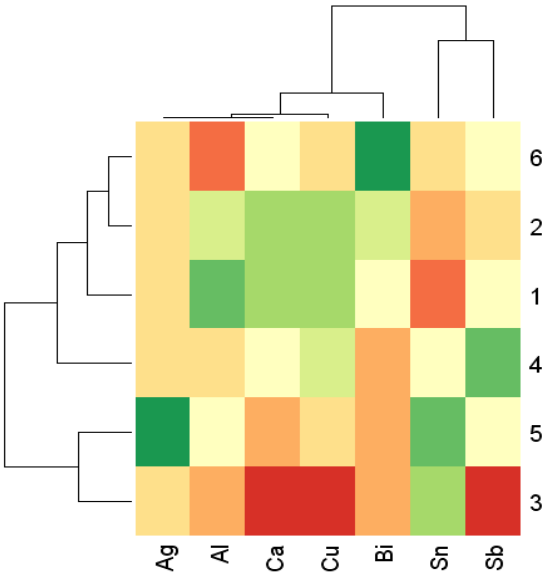


图 18: 热图

```
hcaverage<-hclust(dist2,"average");
hcward<-hclust(dist2,"ward");
labelColors<-c("#556270","#4ECDC4","#1B676B","#FF6B6B");
clusMember<-cutree(hcaverage,4);
hcdend<-as.dendrogram(hcaverage);
colLab<-function(n){
   if(is.leaf(n)){
      a<-attributes(n)
      labCol<-labelColors[clusMember
            [which(names(clusMember)==a$label)]]
      attr(n,"nodePar")<-c(a$nodePar,lab.col=labCol)
   }
   n
}
clusDendro<-dendrapply(hcdend,colLab);
par(lty=2,bg="#DDE3CA");
plot(clusDendro,axes=F);
```

```
axis(side=2,at=seq(0,1.2,0.2),col="#F38630",
     labels=FALSE,lwd=2);
dev.new();
library(ape)
mypal<-c("#556270","#4ECDC4","#1B676B","#FF6B6B");
clus4<-cutree(hcward,4);
par(bg="#E8DDCB",mar=rep(2.4,4),cex=1.3);
plot(as.phylo(hcward),type="fan",edge.width=2,
     edge.color="darkgrey",tip.color=mypal[clus4]);
```

由于本题样本点相对较多，为使谱系聚类图更为美观，在代码中对谱系图的绘制借助于ape包也做了一定的改进 根据谱系聚类图可

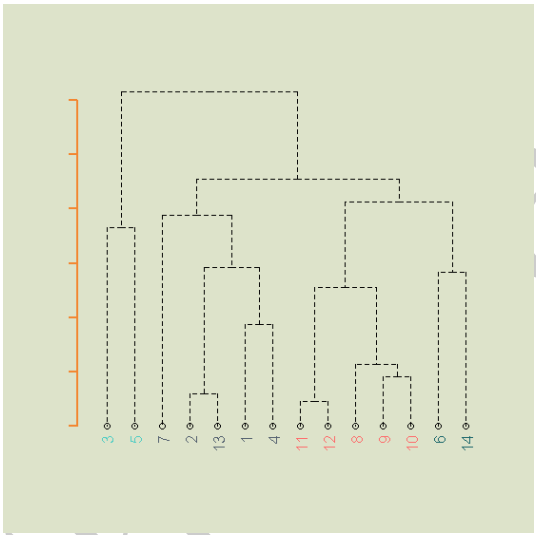


图 19: 类平均法

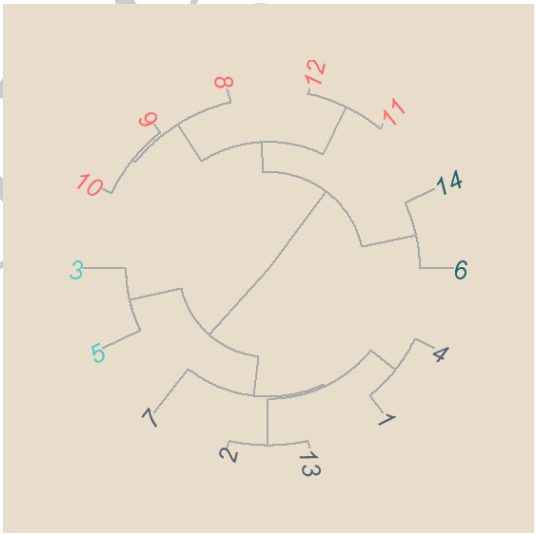


图 20: Ward法

以看到类平均法可把14个岩石比标本分为四类：{1,2,4,7,13}, {3,5}, {8,9,10,11,12 },{6,14}. 而用Ward法同样也可以分为四类：{1,4}, {2,3,5,7,13}, {8,9,10,11,12}, {6,14}。可以看到上述分类结果中前两类属含矿而后两类属不含矿。当然系统聚类的分类结果也常常会与实际类别有差别。

此外我们还可以对图表做进一步改进，借助于经典多维标度(主坐标分析)，我们也可以通过散点图来直观的观察聚类的效果，利

用ggplot2 作图系统，在之前代码的基础上运行如下代码

```
hc<-hclust(dist(D611),"ward")
cluster<-cutree(hc, k=4)
xy<-data.frame(cmdscale(dist(D611)),factor(cluster))
names(xy)<-c("x","y","cluster")
xy$model<-rownames(xy)
library(ggplot2)
p=ggplot(xy,aes(x,y,label=rownames(D611)))
p+geom_point(aes(colour=cluster),size=3)
```

输出散点图如下

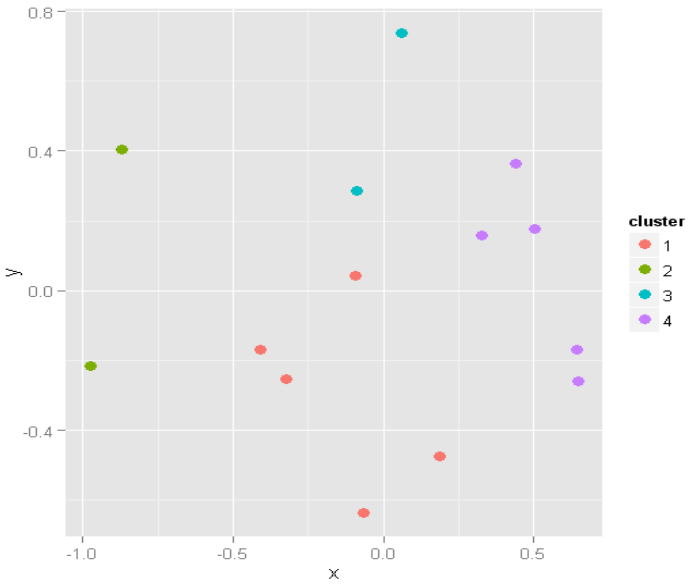


图 21: 聚类效果图

6-12

方法与6-10, 6-11完全一致，请读者自行尝试。类平均法与Ward法分类结果一致，均可以把16个观测点分为4类：{1,2,7}, {3,10,14,15}, {6,8,9,13},{4,5,11,12,16}.

第七章 主成分分析

7-1

从协方差阵出发得总体主成分为

$$Z_1 = 0.040X_1 + 0.999X_2, \quad Var(Z_1) = \lambda_1 = 100.1614$$
$$Z_2 = 0.999X_1 - 0.040X_2, \quad Var(Z_2) = \lambda_2 = 0.8386$$

从相关阵出发得总体主成分为

$$Z_1^* = 0.707X_1^* + 0.707X_2^*, \quad Var(Z_1^*) = \lambda_1^* = 1.4$$
$$Z_2^* = 0.707X_1^* - 0.707X_2^*, \quad Var(Z_2^*) = \lambda_2^* = 0.6$$

这里带$*$即为标准化变量。对比之下不难发现以下三个主要不同点

(1)两者所得到的总体主成分不同且有较大差异

(2)各主成分的方差贡献率也有显著不同

(3)会得出截止迥异的分析结论，但相对来说标准化后分析更为合理[4]

7-2

(1)先对协方差阵Σ求特征值与特征向量得到

$$\lambda_1 = 1 + \rho$$
$$\lambda_2 = 1 - \rho$$
$$\alpha_1 = (\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})'$$
$$\alpha_2 = (\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2})'$$

故根据定义即可得到总体主成分为

$$Z_1 = \frac{\sqrt{2}}{2}X_1 + \frac{\sqrt{2}}{2}X_2 (Var(Z_1) = 1 + \rho)$$
$$Z_2 = \frac{\sqrt{2}}{2}X_1 + \frac{\sqrt{2}}{2}X_2 (Var(Z_2) = 1 - \rho)$$

[4]对比仅供参考

(2)根据第一题结论即可得到长轴方向为$\alpha_1 = (\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})'$, 短轴方向为$\alpha_2 = (\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2})'$

(3)第一主成分贡献率达95%以上即$\frac{\lambda_1}{\lambda_1+\lambda_2} \geq 95\%$, 即$\frac{1+\rho}{2} \geq 0.95$, 故$\rho$的取值应为$\rho \geq 0.9$

7-3

(1)不妨先从相关阵出发求其特征值及其对应的特征向量，由题设$\det(\Sigma/\sigma^2 - \lambda I_P) = [1-\lambda+(p-1)\rho](1-\lambda-\rho)^{p-1}$ 故$\lambda_1 = 1+(p-1)\rho, \lambda_2 = \lambda_3 = \cdots = \lambda_p = 1-\rho$, 进一步得到$\lambda_1$对应的特征向量为$(\frac{1}{\sqrt{p}}, \cdots, \frac{1}{\sqrt{p}})'$, 故由定义即可得总体第一主成分为$Z_1 = \frac{1}{\sqrt{p}}(X_1 + X_2 + \cdots + X_p)$, 即得证。

(2)由于$\lambda_1 = \sigma^2[1+(p-1)\rho]$, 故第一主成分的贡献率即为

$$\frac{\lambda_1}{\sum\limits_i \lambda_i} = \frac{\sigma^2[1+(p-1)\rho]}{p\sigma^2} = \rho + \frac{1-\rho}{p}$$

7-4

设协方差阵Σ的特征值为$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, 对应的单位正交向量为$a_1, a_2, \cdots, a_p$, 则$\Sigma^{-1}$的谱分解式为

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

代入题设即有

$$\begin{aligned}(X-\mu)' \cdot \sum_{i=1}^{p} \frac{1}{\lambda_i} a_i a_i' \cdot (X-\mu) &= \sum_{i=1}^{p} \frac{1}{\lambda_i} [(X-\mu)' a_i]^2 \\ &= \frac{Z_1^2}{\lambda_1} + \frac{Z_2^2}{\lambda_2} + \cdots + \frac{Z_p^2}{\lambda_p} = C^2\end{aligned}$$

这里$Z_i = (X-\mu)' a_i$, 上式等价于

$$\frac{Z_1^2}{\lambda_1 C^2} + \frac{Z_2^2}{\lambda_2 C^2} + \cdots + \frac{Z_p^2}{\lambda_p C^2} = 1$$

即为等概率密度椭圆的方程，由此可以看出椭圆的第i个主轴的方向即为第i个主成分的方向。

7-5

根据题设容易得到协方差阵Σ的特征值与对应的特征向量为

$$\lambda_1 = 4, \quad \alpha_1 = (1,0,0)'$$
$$\lambda_2 = 4, \quad \alpha_2 = (0,1,0)'$$
$$\lambda_3 = 2, \quad \alpha_3 = (0,0,1)'$$

故总体主成分为$Z_i = X_i(i = 1,2,3)$, 三主成分对应的方差分别为4,4,2. [5]

7-6

当$0 < \rho \leq \frac{1}{\sqrt{2}}$时，从协方差阵出发，计算得协方差阵$\Sigma$的特征值及对应的特征向量为

$$\lambda_1 = \sigma^2(1+\sqrt{2}\rho), \quad \alpha_1 = (\frac{1}{2}, \frac{\sqrt{2}}{2}, \frac{1}{2})'$$
$$\lambda_2 = \sigma^2, \qquad\qquad \alpha_2 = (\frac{\sqrt{2}}{2}, 0, -\frac{\sqrt{2}}{2})'$$
$$\lambda_3 = \sigma^2(1-\sqrt{2}\rho), \quad \alpha_3 = (\frac{1}{2}, -\frac{\sqrt{2}}{2}, \frac{1}{2})'$$

由此即可得到此时总体主成分为

$$Z_1 = \frac{1}{2}X_1 + \frac{\sqrt{2}}{2}X_2 + \frac{1}{2}X_3$$
$$Z_2 = \frac{\sqrt{2}}{2}X_1 - \frac{\sqrt{2}}{2}X_3$$
$$Z_3 = \frac{1}{2}X_1 - \frac{\sqrt{2}}{2}X_2 + \frac{1}{2}X_3$$

对应的方差为

$$Var(Z_1) = \lambda_1 = \sigma^2(1+\sqrt{2}\rho)$$
$$Var(Z_2) = \lambda_2 = \sigma^2$$
$$Var(Z_3) = \lambda_3 = \sigma^2(1-\sqrt{2}\rho)$$

[5]事实上由题设即可得到变量两两不相关，故总体主成分等价于原始变量

方差贡献率分别为

$$\frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3}=\frac{1+\sqrt{2}\rho}{3}$$
$$\frac{\lambda_2}{\lambda_1+\lambda_2+\lambda_3}=\frac{1}{3}$$
$$\frac{\lambda_3}{\lambda_1+\lambda_2+\lambda_3}=\frac{1-\sqrt{2}\rho}{3}$$

同理也可以得到$\rho=0$与$-\frac{1}{\sqrt{2}}\leq\rho<0$时的情形.

7-7

采用与7-3类似方法可以求得协方差阵Σ的特征值与对应的特征向量，从而得到总体主成分为

$$Z_1=\frac{1}{2}(X_1+X_2+X_3+X_4),\quad Var(Z_1)=\lambda_1=\sigma^2+\sigma_{12}+\sigma_{13}+\sigma_{14}$$
$$Z_2=\frac{1}{2}(X_1+X_2-X_3-X_4),\quad Var(Z_2)=\lambda_2=\sigma^2+\sigma_{12}-\sigma_{13}-\sigma_{14}$$
$$Z_3=\frac{1}{2}(X_1-X_2+X_3-X_4),\quad Var(Z_3)=\lambda_3=\sigma^2-\sigma_{12}+\sigma_{13}-\sigma_{14}$$
$$Z_4=\frac{1}{2}(X_1-X_2-X_3+X_4),\quad Var(Z_4)=\lambda_4=\sigma^2-\sigma_{12}-\sigma_{13}+\sigma_{14}$$

7-8

见教材276-277页(1)$\hat{b}_j=(a_{1j},a_{2j},\cdots,a_{mj})'$

(2)X_j的回归平方和为

$$\begin{aligned}U_j&=(n-1)\sum_{t=1}^{m}\lambda_t a_{jt}^2\\&=(n-1)\sum_{t=1}^{m}\rho^2(X_j,Z_t)=(n-1)\nu_j\end{aligned}$$

X_j的残差平方和为

$$Q_j=(n-1)(1-\nu_j)$$

X_j的决定系数为

$$R_j^2=\nu_j=\sum_{t=1}^{m}\rho^2(X_j,Z_t)$$

7-9

(1)利用样本的似然函数可以得到λ_1的极大似然估计量

(2)通过定义说明即可

7-10

(1)先证充分性。设Y的协方差阵的特征值分别为$\nu_1 \geq \nu_2 \geq \cdots \geq \nu_p \geq 0$, 则可以得到

$$L'(\Sigma + \sigma^2 I_p)L = diag(\nu_1, \nu_2, \cdots, \nu_p)$$

从而可以得到

$$\begin{aligned} L'\Sigma L &= diag(\nu_1, \nu_2, \cdots, \nu_p) - L'\sigma^2 I_p L \\ &= diag(\nu_1, \nu_2, \cdots, \nu_p) - \sigma^2 I_p \\ &= diag(\nu_1 - \sigma^2, \nu_2 - \sigma^2, \cdots, \nu_p - \sigma^2) \end{aligned}$$

结合协方差阵的非负性，可知$\nu_1 - \sigma^2 \geq \nu_2 - \sigma^2 \geq \cdots \geq \nu_p - \sigma^2$为$\Sigma$的特征值，且$L$的列向量为对应的特征向量。由主成分定义可知$L'X$是$X$的主成分。

(2)必要性。设Σ的特征值为$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, 则可以得到

$$L'\sigma L = diag(\lambda_1, \lambda_2, \cdots, \lambda_p)$$

从而可以得到

$$\begin{aligned} L'(\Sigma + \sigma^2 I_p)L &= L'\Sigma L + L'\sigma^2 I_p L \\ &= diag(\lambda_1, \lambda_2, \cdots, \lambda_p) + \sigma^2 I_p \\ &= diag(\lambda_1 + \sigma^2, \lambda_2 + \sigma^2, \cdots, \lambda_p + \sigma^2) \end{aligned}$$

即$\lambda_1 + \sigma^2 \geq \lambda_2 + \sigma^2 \geq \cdots \geq \lambda_p + \sigma^2$为$\Sigma + \sigma^2 I_p$的特征值，且$L$ 的列向量为对应的特征向量。由主成分定义可知$L'Y$是Y的主成分。故命题得证。

7-11

(1)简单的主成分分析可通过pricomp函数实现，[6] 参考代码如下

```
D711<-read.table("711.txt",head=F);
D711.pr<-princomp(D711,cor=T);
summary(D711.pr,loadings=T);
```

部分输出如下

```
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4
Standard deviation     1.7620762 1.7021873 0.9644768 0.80132532
Proportion of Variance 0.3881141 0.3621802 0.1162769 0.08026528
Cumulative Proportion  0.3881141 0.7502943 0.8665712 0.94683649
```

由输出可知，综合前三个主成分可解释原变量信息的86.66%，而综合前四个主成分可解释原变量信息的94.68%。

还可以采用更为灵活的factomineR包进行分析, 参考代码如下

```
library(FactoMineR);
res.pca<-PCA(data, scale.unit=TRUE, ncp=3, graph=T);
```

(2)由第一小题知，可选取前三个主成分进行分析，因此利用前三个主成分得分对行业进行聚类，并对第一主成分进行排序，参考代码如下

```
data<-read.table("711.txt",head=F);
data.pr<-princomp(data,cor=T);
summary(data.pr,loadings=T);
data.pred<-predict(data.pr);
dimnames(data.pred)<-list(1:13,1:8);
```

[6]函数prcomp与princomp类似均用于主成分分析，但两者区别在于prcomp是通过奇异值分解的方法来求解主成分而princomp则采用的是特征值

```
sort(data.pred[,1],decreasing=TRUE);
d<-dist(data.pred[,1:3]);
hc<-hclust(d,"ward");
plot(hc,hang=-1);
rect.hclust(hc,k=3,border="red");
```

根据输出结果得，按第一主成分得分由小到大可排序如下：

$$8 < 10 < 12 < 7 < 9 < 11 < 13 < 6 < 4 < 2 < 3 < 1 < 5$$

此外通过Ward法可将行业分成4类，分类结果如下：

- G_1={2，3，6}
- G_2={7，11，12}
- G_3={8，9，10}
- G_4={1，4，5，13}

7-12

与上一小题基本类似，具体参考代码如下

```
D712<-read.table("712.txt",head=F);
D7120<-D712[,2:7];
D712.pr<-princomp(D7120,cor=T);
summary(D712.pr,loadings=T);
screeplot(D712.pr,type="lines");
D712.pred<-predict(D712.pr);
dimnames(D712.pred)<-list(D712[,1],1:6);
sort(D712.pred[,1],decreasing=TRUE);
d<-dist(D712.pred[,1:2]);
hc<-hclust(d,"ward");
plot(hc,hang=-1);
```

根据输出可知，若综合为两个主成分，可解释原变量信息的81.24%，故这里取前两个主成分分析即可。根据输出可以得到按第一主成分得分由小到大可排序如下：山西>河北>河南>江西>内蒙古>黑龙江>福建>安徽>山东>吉林>江苏>辽宁>天津>浙江>北京>上海.

利用前两主成分得分进行聚类分析，用Ward法可将地区分为四类，分类结果如下：

- G_1={北京，上海}
- G_2={河北，山西，内蒙古，河南}
- G_3={黑龙江，吉林，江西，安徽，福建}
- G_4={浙江，辽宁，江苏，天津，山东}

第八章 因子分析

8-1

事实上根据相关阵R我们不难发现以下关系

$$R=\begin{pmatrix}0.9\\0.7\\0.5\end{pmatrix}\begin{pmatrix}0.9&0.7&0.5\end{pmatrix}+\begin{pmatrix}0.19&0&0\\0&0.51&0\\0&0&0.75\end{pmatrix}$$

由此即可得到$m=1$时的正交因子模型为

$$\begin{cases}X_1&=0.9F_1+\epsilon_1\\X_2&=0.7F_2+\epsilon_2\\X_3&=0.5F_3+\epsilon_3\end{cases}\qquad D=\begin{pmatrix}0.19&0&0\\0&0.51&0\\0&0&0.75\end{pmatrix}$$

8-2

(1)当$m=1$时，$A=\sqrt{\lambda_1}l_1=(0.8758,0.8312,0.7111)'$, 故

$$\sigma_1^2=1-0.8758^2=0.2330$$
$$\sigma_2^2=1-0.8312^2=0.3901$$
$$\sigma_3^2=1-0.7111^2=0.4943$$

从而可以得到

$$\begin{cases}X_1=0.8758F_1+\epsilon_1\\X_2=0.8312F_2+\epsilon_2\\X_3=0.7111F_3+\epsilon_3\end{cases}$$

$$D=\begin{pmatrix}0.2330&0&0\\0&0.3091&0\\0&0&0.4943\end{pmatrix}$$

$$
\begin{aligned}
\epsilon &= R - (AA' + D) \\
&= \begin{pmatrix} 1.00 & 0.63 & 0.45 \\ 0.63 & 1.00 & 0.35 \\ 0.45 & 0.35 & 1.00 \end{pmatrix} - \begin{pmatrix} 1.00 & 0.7280 & 0.6228 \\ 0.7280 & 1.00 & 0.5911 \\ 0.6228 & 0.5911 & 1.00 \end{pmatrix} \\
&= \begin{pmatrix} 0 & -0.0980 & -0.1728 \\ -0.0980 & 0 & -0.2411 \\ -0.1728 & -0.2411 & 0 \end{pmatrix}
\end{aligned}
$$

因此$Q(1) = 2(0.0980^2 + 0.1728^2 + 0.2411^2) = 0.1951$

(2)当$m = 2$时,

$$
A = (\sqrt{\lambda_1}l_1, \sqrt{\lambda_2}l_2) = \begin{pmatrix} 0.8758 & -0.1802 \\ 0.8312 & -0.4048 \\ 0.7111 & 0.6950 \end{pmatrix}
$$

故而

$$
\begin{aligned}
\sigma_1^2 &= 1 - 0.8758^2 - 0.1802^2 = 0.2006 \\
\sigma_2^2 &= 1 - 0.8312^2 - 0.4048^2 = 0.1453 \\
\sigma_3^2 &= 1 - 0.7111^2 - 0.6950^2 = 0.01122
\end{aligned}
$$

$$
\begin{cases}
X_1 &= 0.8758F_1 - 0.1802F_2 + \epsilon_1 \\
X_2 &= 0.8312F_1 - 0.4048F_2 + \epsilon_2 \\
X_3 &= 0.7111F_1 + 0.6950F_2 + \epsilon_3
\end{cases}
$$

$$
D = \begin{pmatrix} 0.2006 & 0 & 0 \\ 0 & 0.1453 & 0 \\ 0 & 0 & 0.01122 \end{pmatrix}
$$

$$\epsilon = R - (AA' + D) =$$
$$= \begin{pmatrix} 1.00 & 0.63 & 0.45 \\ 0.63 & 1.00 & 0.35 \\ 0.45 & 0.35 & 1.00 \end{pmatrix} - \begin{pmatrix} 1.00 & 0.8009 & 0.4975 \\ 0.8009 & 1.00 & 0.3097 \\ 0.4975 & 0.3097 & 1.00 \end{pmatrix}$$
$$= \begin{pmatrix} 0 & -0.1709 & -0.0475 \\ -0.1709 & 0 & 0.0403 \\ -0.0475 & 0.0403 & 0 \end{pmatrix}$$

故$Q(2) = 2(0.1709^2 + 0.0475^2 + 0.0403^2) = 0.06611$

$(3)Q(2) = 0.0662 < 0.1$, 故$m = 2$时的主成分解符合要求。

8-3

暂略

8-4

由于

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

故根据题设可以得到

$$\epsilon = S - (AA' + D) = \sum_{j=m+1}^{p} \lambda_j l_j l_j' - D$$

令$BB' = \sum_{j=m+1}^{p} \lambda_j l_j l_j'$, 则有

$$\begin{aligned}\sum_{j=m+1}^{p} \lambda_j^2 &= tr(BB' \cdot BB') = tr[(\epsilon + D)(\epsilon + D)'] \\ &= tr(\epsilon\epsilon') + tr(DD') \\ &= Q() + \sum_{j=1}^{p} (\sigma_j^2)^2\end{aligned}$$

由此即可得到$Q(m) \leq \sum_{j=m+1}^{p} \lambda_j^2$, 故得证

8-5

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

(2)主成分分析中主成分的个数和变量个数p相同，它是将一组具有相关关系的变量变换为一组互不相关的变量，而因子分析的目的是要用尽可能少的公因子，以便构造一个结构简单的因子模型

(3)主成分分析是将主成分表示为原始变量的线性组合，而因子分析是将原始变量表示为公因子和特殊因子的线性组合，用假设的公因子来“解释”相关阵的内部依赖关系[7]

8-6

(1)主成分分析，参考上一章习题。

(2)因子分析：stats包中自带的factanal()函数采用的是极大似然法估计参数，但由于自由度的非负约束，这里采用极大似然法估计最多只能取3个公因子(解释比例不到70%)，参考代码如下

```
D806<-read.table("806.txt",header=F);
library(MASS);
```

[7]引自教材294页，更多的异同点均可在网上搜到，这里不再赘述，但事实上在真正实践中并没有必要去留意这些

```
D806<-as.matrix(D806);
factanal(D806,factors=3,scores="regression",rotation="varimax");
```

此外题中所给数据的样本量太小显然也并不适合极大似然法估计。故这里采用主成分法估计，基于《统计建模与R 软件》一书中的相应代码实现[8]，函数代码见附录，估计求解参考代码如下

```
D806<-read.table("806.txt",header=F);
R<-cor(D806);
source("factorPCAlca.R");
fa<-factorPCAlca(R,m=4);
vm<-varimax(fa$loadings,normalize=F);
```

根据输出可以看到前四个公因子可反映原始变量的89.26%，方差最大正交旋转后载荷阵如下

```
   Factor1 Factor2 Factor3 Factor4
X1  0.836   0.268  -0.246  -0.310
X2 -0.801   0.183           0.237
X3 -0.271  -0.921          -0.123
X4                 -0.971
X5 -0.254                   0.951
X6  0.387  -0.786  -0.193
```

可以看到第一公因子主要代表X_1和X_2, 即氯和硫化氢的浓度；第二公因子主要代表X_3和X_6，即二氧化硫和环己烷的浓度；第三公因子主要代表X_4, 即碳4的浓度；第四公因子主要代表X_5, 即环氧氯丙烷的浓度。[9]

8-7

同样根据数据情况这里继续采用主成分法分析，参考代码如下

[8]代码可改进之处颇多，读者可自行尝试

[9]对于主成分分析和因子分析，读者也可尝试更为灵活的psych包

```
D807<-read.table("807.txt",header=F);
princomp(D807);
R<-cor(D807);
source("factorPCAlca.R");
fa<-factorPCAlca(R,m=2);
```

部分输出结构如下

```
$loadings
      Factor1     Factor2
X1 -0.3504853  0.92792025
X2 -0.9760372 -0.09685944
X3 -0.9397083  0.23600842
X4 -0.9395850 -0.23402584
X5 -0.9546112 -0.24363420

$var
      common    spcific
X1 0.9838760 0.01612404
X2 0.9620303 0.03796969
X3 0.9387517 0.06124830
X4 0.9375880 0.06241198
X5 0.9706402 0.02935975

$B
                 Factor1   Factor2
SS loadings    3.7526428 1.0402434
Proportion Var 0.7505286 0.2080487
Cumulative Var 0.7505286 0.9585772
```

可以看到取前两个公因子即可反映原始变量信息的95.86%，根据输出可以得到正交因子模型如下

$$\begin{cases} X_1 &= 0.35049F_1 + 0.92792F_2 + \epsilon_1, \quad h_1^2 = 0.9838760 \\ X_2 &= 0.97604F_1 - 0.09686F_2 + \epsilon_2, \quad h_2^2 = 0.9620303 \\ X_3 &= 0.93971F_1 + 0.23601F_2 + \epsilon_3, \quad h_3^2 = 0.9387517 \\ X_4 &= 0.93958F_1 - 0.23403F_2 + \epsilon_4, \quad h_4^2 = 0.9375880 \\ X_5 &= 0.95461F_1 - 0.24363F_2 + \epsilon_5, \quad h_5^2 = 0.9706402 \end{cases}$$

8-8

本题即可采用主成分法也可采用极大似然法估计，两者结果略有不同，但相对来说主成分法更容易解释，极大似然法估计的参考代码如下

```
D808<-read.table("808.txt",header=F);
factanal(D808,factors=2,scores="regression",
         rotation="varimax");
```

部分输出如下

```
Loadings:
   Factor1 Factor2
V1         0.691
V2 0.135   0.633
V3 0.476   0.640
V4 0.545   0.470
V5 0.997

                Factor1 Factor2
SS loadings       1.537   1.510
```

```
Proportion Var   0.307   0.302
Cumulative Var   0.307   0.609

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 0.11 on 1 degree of freedom.
The p-value is 0.739
```

可以看到前两个公因子可反映原始变量信息的61%, 并且通过检验也可以看到取两个公因子已经足够。根据方差最大正交旋转后的载荷阵可以看到第一公因子主要反映了总成绩和分析的成绩，而第二公因子则主要反映了力学、物理、代数三门课的成绩。此外主成分法求解的结果可直接参考教材所给的参考答案。

第九章 对应分析方法

9-1

(1)对应分析可采用anacor包求解，参考代码如下

```
D901<-read.table("901.txt",header=F);
library(anacor);
colnames(D901)<-c("氯","硫化氢","二氧化硫","碳4","环氧氯丙
烷","环己烷");
rownames(D901)<-c("1","2","3","4","5","6","7","8");
A901<-anacor(D901,scaling=c("centroid","centroid"));
plot(A901,plot.type="jointplot",xlim=c(-1.5,1.2),ylim=c(-1,1.2),
     asp=1,font=2,font.lab=2);
```

部分输出结果如下

```
> A901
CA fit:
Sum of eigenvalues:  0.6288095
Total chi-square value: 3.589

Chi-Square decomposition:
            Chisq Proportion Cumulative Proportion
Component 1 2.512      0.700                 0.700
Component 2 0.606      0.169                 0.869
Component 3 0.281      0.078                 0.947
Component 4 0.144      0.040                 0.987
Component 5 0.047      0.013                 1.000
```

可以看到总χ^2统计量的86.9%可用前两维说明，故可以认为样品点和变量点用二维表示即可，所绘制的散点图如下

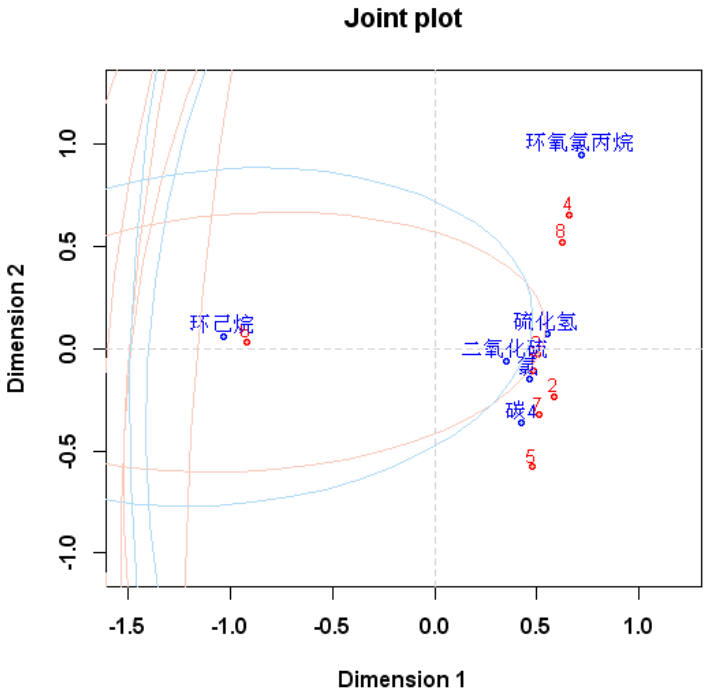


图 22: 对应分析

根据散点图我们可以粗略地将变量点和样本点分为三类:

- G_1={环己烷和6}，表示6号样品含环己烷浓度较高
- G_2={环氧氯丙烷和4,8}，表示4,8这两个样本点中环氧氯丙烷的浓度较高
- G_3={氯,硫化氢,二氧化硫,碳4和1,2,3,5,7}，表示1,2,3,5,7这五个样本点中前四种气体的浓度较高

此外也可采用MASS包中的corresp函数，参考代码如下读者可自行加工完善

```
D901<-read.table("901.txt",header=F);
library(MASS);
D901<-as.matrix(D901);
colnames(D901)<-c("氯","硫化氢","二氧化硫","碳4","环氧氯丙烷","环己烷");
```

```
result<-corresp(D901,nf=2);
biplot(result,font=2);
```

(2)由习题8.6即可知，取前四个公因子，第一个公因子主要代表氯和硫化氢的浓度；第二公因子主要代表二氧化硫和环已烷的浓度；第三公因子主要代表碳4的浓度；第四公因子主要代表环氧氯丙烷的浓度。利用第一、第二因子得分的散点图，可将8个样本点分为三类：G_1={2,3,4,7}，G_1={1,5,8}，G_1={6}

(3)同样由习题8.6的分析结果可知，8个样本点可分为三类：G_1={2,3,5,7}，G_1={1,4,8}，G_1={6}

9-2

(1)相对于前一题，事实上对应同样可以采用factomineR包分析，操作更为简洁且可以得到更为丰富的输出结果，参考代码如下

```
D902<-read.table("902.txt",header=F);
library(FactoMineR);
colnames(D902)<-c("食品","衣着","燃料","住房","生活用品及其他","文化生活服务支出");
result<-CA(D902);
```

由输出结果可知，总χ^2统计量的92.10%用前两维即可说明，故行点和列点之间的关系用二维表示就已经足够。行点和列点的散点图如下[10]

根据散点图我们可以粗略地将行点和列点分为5类[11]

- G_1={江苏，浙江和X_1(食品)}，说明这两个地区农民用于食品的消费比例比较大
- G_2={河北，河南，辽宁，内蒙古，黑龙江，山西，吉林和X_2(衣着)}，说明这些地区农民用于衣着的消费比例较大

[10]值得一提的是，不同于经典的对应分析，这里还可以用补充元素作为测试集

[11]本题结果并不那么明显，分类也较为牵强，这里尊重教材所给的结果，仅供参考

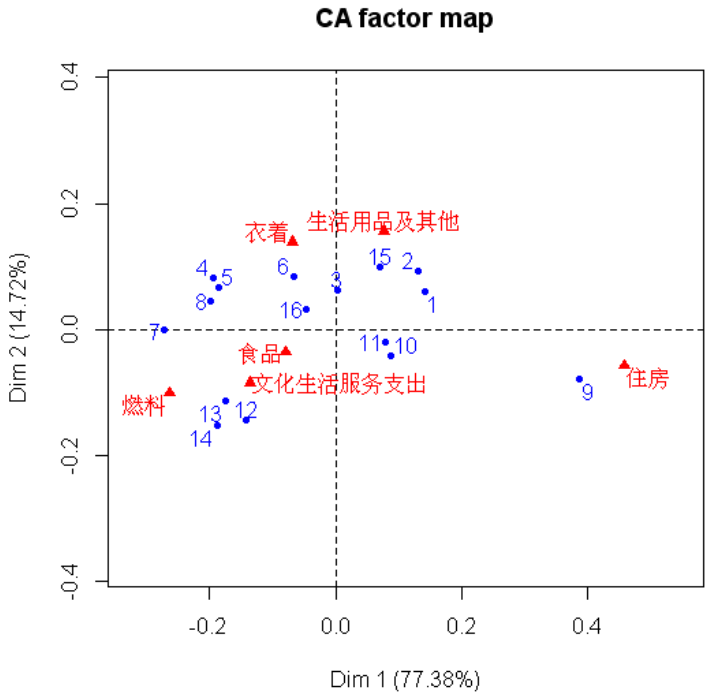


图 23: 对应分析

- G_3={安徽，福建，江西和X_3(燃料)，X_6(文化生活服务支出)}，表示这三个地区农民用于燃料和文化生活服务的消费比例较大
- G_4={上海，X_4(住房) }，说明上海农民用于住房的消费比例比较大
- G_5={北京，天津，山东和X_5(生活用品及其他)}，说明这三个地区的农民用于生活用品及其他的消费比例比较大

(2)因子代码可直接参考第八章习题，根据分析结果只需取前三个公共因子便可反应原始变量信息的91.38%，其中第一公因子主要表示X_1(食品)、X_2(衣着)、X_4(住房)、X_5(生活用品及其他)；第二公因子主要表示X_6(文化生活服务支出)；第三公因子主要表示X_3.利用第二公因子得分对第一公因子得分的散点图，可将16个地区分为4类即

- G_1={北京，上海}
- G_2={福建，安徽，吉林，江西}

- G_3={山东，天津，河南，山西，内蒙古，河北}
- G_4={浙江，江苏，黑龙江，辽宁}

(3)见教材243页

9-3

暂略

第十章 典型相关分析

10-1

根据题设可以得到

$$R_{11}=\begin{pmatrix}1.0 & 0.5\\ 0.5 & 1.0\end{pmatrix}, R_{11}^{-1}=\frac{4}{3}\begin{pmatrix}1.0 & -0.5\\ -0.5 & 1.0\end{pmatrix}$$
$$R_{22}=\begin{pmatrix}1.0 & 0.6\\ 0.6 & 1.0\end{pmatrix} R_{22}^{-1}=\frac{25}{16}\begin{pmatrix}1.0 & 0.6\\ 0.6 & 1.0\end{pmatrix}$$

故由此可得

$$\begin{aligned}M_1^*&=R_{11}^{-1}R_{12}R_{22}^{-1}R_{21}\\ &=\frac{25}{12}\begin{pmatrix}1.0 & -0.5\\ -0.5 & 1.0\end{pmatrix}\begin{pmatrix}0.7 & 0.7\\ 0.7 & 0.7\end{pmatrix}\begin{pmatrix}1.0 & -0.6\\ -0.6 & 1.0\end{pmatrix}\begin{pmatrix}0.7 & 0.7\\ 0.7 & 0.7\end{pmatrix}\\ &=\frac{49}{120}\begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix}\end{aligned}$$

计算得M_1^*的特征值$\lambda_1^2=\frac{49}{60}, \lambda_2^2=0$. 且$M_1^*$对应$\lambda_1^2$, 且满足$a'R_{11}a=1$的特征向量为$a=(\frac{1}{\sqrt{3}},\frac{1}{\sqrt{3}})'$, 类似地也可以得到$b=(\frac{\sqrt{5}}{4},\frac{\sqrt{5}}{4})'$, 故根据定义可知第一对典型相关变量

$$V_1=a'X=\frac{\sqrt{3}}{3}(X_1+X_2)$$
$$W_1=b'Y=\frac{\sqrt{5}}{4}(Y_1+Y_2)$$

第一典型相关系数$\rho_1=\lambda_1=\sqrt{\frac{49}{60}}=0.9037$ [12]

10-2

10-3

函数cancor

[12]本题结果也可通过R语言直接求解得到

10-4

10-5

第十一章 偏最小二乘回归分析

11-1

11-2

附录一 封面图片代码

```
library(mvtnorm);
library(MASS);
set.seed(5)
sigma <-matrix(c(4,2,2,3), ncol=2)
x<-rmvnorm(n=500, mean=c(1,2), sigma=sigma,
           method="chol")
z<-kde2d(x[,1],x[,2],n=200);
par(mar=rep(0,4))
persp(z, theta = 60, phi = 5, col = "lightgreen",
shade = 0.4, border = NA,box = FALSE)
```

附录二 协方差检验函数代码

```
varcomp <- function(covmat,n) {
   if (is.list(covmat)) {
    if (length(covmat)<2)
        stop("covmat must be a list
              with at least 2 elements")
    ps <- as.vector(sapply(covmat,dim))
    if (sum(ps[1] == ps) != length(ps))
        stop("all covariance matrices must
        have the same dimension")
    p <- ps[1]
        q <- length(covmat)
        if (length(n) == 1)
        Ng <- rep(n,q)
    else if (length(n) == q)
        Ng <- n
    else
        stop("n must be equal length(covmat)
             or 1")
    DNAME <- deparse(substitute(covmat))
   }
   else
    stop("covmat must be a list")
   ng <- Ng - 1
   Ag <- lapply(1:length(covmat),function(i,mat,n)
    { n[i] * mat[[i]] },mat=covmat,n=ng)
   A <- matrix(colSums(matrix(unlist(Ag),
```

```
      ncol=p^2,byrow=T)),ncol=p)
  detAg <- sapply(Ag,det)
  detA <- det(A)
  V1 <- prod(detAg^(ng/2))/(detA^(sum(ng)/2))
  kg <- ng/sum(ng)
  l1 <- prod((1/kg)^kg)^(p*sum(ng)/2) * V1
  rho <- 1 - (sum(1/ng) -
      1/sum(ng))*(2*p^2+3*p-1)/(6*(p+1)*(q-1))
  w2 <- p*(p+1) * ((p-1)*(p+2) * (sum(1/ng^2)
        - 1/(sum(ng)^2))
       - 6*(q-1)*(1-rho)^2) / (48*rho^2)
  f <- 0.5 * (q-1)*p*(p+1)
  STATISTIC <- -2*rho*log(l1)
  PVAL <- 1 - (pchisq(STATISTIC,f) +
      w2*(pchisq(STATISTIC,f+4)-pchisq(STATISTIC,f)))
  names(STATISTIC)<-"corrected lambda*"
  names(f)<-"df"
  RVAL <- structure(list(statistic = STATISTIC,
  parameter = f,p.value = PVAL, data.name = DNAME,
  method = "Equality of Covariances
          Matrices Test"),class="htest")
  return(RVAL)
}
```

附录三　因子分析主成分法代码

```
factorPCAlca<-function(S, m){
   p<-nrow(S); diag_S<-diag(S);
   sum_rank<-sum(diag_S)
   rowname<-paste("X", 1:p, sep="")
   colname<-paste("Factor", 1:m, sep="")
   A<-matrix(0, nrow=p, ncol=m,
   dimnames=list(rowname,colname))
   eig<-eigen(S)
   for(i in 1:m){
      A[,i]<-sqrt(eig$values[i])*eig$vectors[,i]
   }
   h<-diag(A%*%t(A))
   rowname<-c("SS loadings","Proportion Var",
              "Cumulative Var")
   B<-matrix(0,nrow=3,ncol=m,
   dimnames=list(rowname,colname))
   for(i in 1:m){
      B[1,i]<-sum(A[,i]^2)
      B[2,i]<-B[1,i]/sum_rank
      B[3,i]<-sum(B[1,1:i])/sum_rank
   }
   method<-c("Principal Component Method")
   list(method=method,loadings=A,
   var=cbind(common=h,spcific=diag_S-h),B=B)
}
```