第四次数据分析作业

张申铎 2176112379

目录


```
library(expm)
library(purrr)
```

1 Problem 1

1. (P153 4.1 题) 设总体 $\mathbf{X}=(X_1,X_2,X_3)^T$ 的**协方差矩阵**为

$$\mathbf{\Sigma}=\begin{pmatrix}\sigma^2 & \sigma^2\rho & 0\\ \sigma^2\rho & \sigma^2 & \sigma^2\rho\\ 0 & \sigma^2\rho & \sigma^2\end{pmatrix},\quad |\rho|<\frac{1}{\sqrt{2}}$$

求 $\mathbf{X}$ 的主成分及各主成分的贡献率。

考虑协方差矩阵,

$$|\Sigma-\lambda I|=\left|\begin{pmatrix}\sigma^2-\lambda & \sigma^2\rho 0 & \\ \sigma^2\rho & \sigma^2-\lambda & \sigma^2\rho\\ 0 & \sigma^2\rho & \sigma^2-\lambda\end{pmatrix}\right| \tag{1}$$

通过计算, 上面的行列式等于

$$(\sigma Q^2-\lambda)(\sigma^2-\lambda-\sqrt{2}\sigma^2\rho)(\sigma^2-\rho+\sqrt{2}\sigma^2\rho) \tag{2}$$

通过观察我们可以看出其三个特征值与其对应的特征向量分别为,

$$\lambda_1=\sigma^2, \qquad e_1=\frac{1}{\sqrt{2}}(-1,0,1)^T \tag{3}$$

$$\lambda_2=\sigma^2-\sqrt{2}\sigma^2\rho, \qquad e_2=\frac{1}{2}(1,-\sqrt{2},1)^T \tag{4}$$

$$\lambda_3=\sigma^2+\sqrt{2}\sigma^2\rho, \qquad e_3=\frac{1}{2}(1,\sqrt{2},1)^T \tag{5}$$

$$\tag{6}$$

故其主成分为

$$Y_1=-\frac{1}{\sqrt{2}}X_1+\frac{1}{\sqrt{2}}X_2 \tag{7}$$

$$Y_2=\frac{1}{2}X_1-\frac{\sqrt{2}}{2}X_2+\frac{1}{2}X_3 \tag{8}$$

$$Y_3=\frac{1}{2}X_1+\frac{\sqrt{2}}{2}X_2+\frac{1}{2}X_3 \tag{9}$$

$$\tag{10}$$

对应的贡献率为 $\frac{1}{3}, \frac{1+\sqrt{2}\rho}{3}, \frac{1-\sqrt{2}\rho}{3}$.

2 Problem 2

2. (P153 4.2 题) 设总体 $\mathbf{X}=(X_1,X_2,X_3)^T$ 的**相关系数矩阵**为

$$\boldsymbol{\rho}=\begin{pmatrix}1 & \rho & \rho\\ \rho & 1 & \rho\\ \rho & \rho & 1\end{pmatrix}, \quad \rho>0$$

(1) 求 $\mathbf{X}$ 的**标准化变量的主成分**及其贡献率.

(2) 将上述结果推广到 p 维情形。

2.1 (1)

考虑相关系数矩阵,

$$|\rho-\lambda I|=\begin{pmatrix}1-\lambda & \rho & \rho\\ \rho & 1-\lambda & \rho\\ \rho & \rho & 1-\lambda\end{pmatrix} \tag{11}$$

通过矩阵的对称形式, 我们发现直接带入 $\lambda=1-\rho$ 可以使得这个行列式三行相等为零, 并且此时其对应的特征子空间为两维. 同时如果我们带入 $\lambda=1+2\rho$, 我们的通过把任意两行加到第三行上面也可以得到一个零行, 故这也是他的一个特征值.

故其相关系数矩阵的特征值与特征向量为

$$\lambda_1=1-\rho, \qquad e_1=\frac{1}{\sqrt{2}}(-1,0,1) \tag{12}$$

$$\lambda_2=1-\rho, \qquad e_2=\frac{1}{\sqrt{2}}(-1,1,0) \tag{13}$$

$$\lambda_3=1+2\rho, \qquad e_2=\frac{1}{\sqrt{3}}(1,1,1) \tag{14}$$

$$\tag{15}$$

故 X 其标准化变量的主成分与贡献率为

$$Y_1 = \frac{1}{\sqrt{2}}(X_3 - X_1) \qquad \frac{1-\rho}{3} \tag{16}$$

$$Y_2 = \frac{1}{\sqrt{2}}(X_2 - X_1) \qquad \frac{1-\rho}{3} \tag{17}$$

$$Y_3 = \frac{1}{\sqrt{3}}(X_1 + X_2 + X_3) \qquad \frac{1+2\rho}{3} \tag{18}$$

$$\tag{19}$$

2.2 (2)

推广到 p 维首先我们观察到 $1-\rho$ 与 $1+(p-1)\rho$ 因为同样的理由依然是系数矩阵的特征值, 并且前者的特征子空间维数为 $p-1$ 后者的特征子空间维数为 1. 并且带入特征值我们可以直接解出特征向量.

$$\lambda_1 = 1-\rho \qquad e_1 = \frac{1}{\sqrt{2}}(-1, 1, 0, \cdots, 0) \tag{20}$$

$$\lambda_2 = 1-\rho \qquad e_2 = \frac{1}{\sqrt{2}}(-1, 0, 1, \cdots, 0) \tag{21}$$

$$\cdots \qquad \cdots \tag{22}$$

$$\lambda_{p-1} = 1-\rho \qquad e_{p-1} = \frac{1}{\sqrt{2}}(-1, 1, 0, \cdots, 1) \tag{23}$$

$$\lambda_p = 1+(p-1)\rho \qquad e_p = \frac{1}{\sqrt{p}}(1, 1, \cdots, 1) \tag{24}$$

对其正交化后得

$$\lambda_1 = 1-\rho \qquad e_1 = \frac{1}{\sqrt{2}}(1, -1, 0, \cdots, 0) \tag{25}$$

$$\lambda_2 = 1-\rho \qquad e_2 = (\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, -\sqrt{\frac{2}{3}}, \cdots, 0) \tag{26}$$

$$\cdots \qquad \cdots \tag{27}$$

$$\lambda_{p-1} = 1-\rho \qquad e_{p-1} = (\frac{1}{\sqrt{p(p-1)}}, \frac{1}{\sqrt{p(p-1)}}, \frac{1}{\sqrt{p(p-1)}}, \cdots, \sqrt{\frac{p-1}{p}}) \tag{28}$$

$$\lambda_p = 1+(p-1)\rho \qquad e_p = \frac{1}{\sqrt{p}}(1, 1, \cdots, 1) \tag{29}$$

故其主成分与其对应的贡献率为为

$$Y_1 = \frac{1}{\sqrt{2}}X_1 - \frac{1}{\sqrt{2}}X_2 \qquad \frac{1-\rho}{p} \tag{30}$$

$$Y_2 = \frac{1}{\sqrt{6}}X_1 + \frac{1}{\sqrt{6}}X_2 - \sqrt{\frac{2}{3}}X_3 \qquad \frac{1-\rho}{p} \tag{31}$$

$$\cdots \qquad \cdots \tag{32}$$

$$Y_p = \frac{1}{\sqrt{p}}(X_1 + X_2 + \cdots + X_p) \qquad \frac{1+(p-1)\rho}{p} \tag{33}$$

$$\tag{34}$$

3 Problem 3

3. (P154 4.5 题)下表 1(见 QQ 群的 exercise4.5.txt 文件)给出了 1991 年我国 30 个省、自治区、直辖市城镇居民的月平均消费数据，所考察的 8 个指标(单位均为元/人)如下：

X_1：人均粮食支出；　　X_2：人均副食支出；

X_3：人均烟酒茶支出；　　X_4：人均其他副食支出；

X_5：人均衣着商品支出；　　X_6：人均日用品支出

X_7：人均燃料支出；　　X_8：人均非商品支出

表 1. 1991 年我国 30 个省、自治区、直辖市城镇居民月均消费数据.

(1) 求样本相关系数矩阵 $\mathbf{R}$.

(2) 从 $\mathbf{R}$ 出发作主成分分析，求各主成分的贡献率和前两个主成分的累计贡献率.

(3) 求出前两个主成分并解释其意义，按第一主成分得分将 30 个省、自治区、直辖市排序，结果如何？

3.1 (1)

```
prob3 <- read.table("./exercise4_5.txt", quote="\"", comment.char="",row.names = 1)
cor(prob3)
```

```
##             V2          V3          V4          V5          V6         V7
## V2  1.00000000  0.33364671 -0.05453868 -0.06125369 -0.28936059 0.19879627
## V3  0.33364671  1.00000000 -0.02290183  0.39893102 -0.15630387 0.71113407
## V4 -0.05453868 -0.02290183  1.00000000  0.53332919  0.49676279 0.03282961
## V5 -0.06125369  0.39893102  0.53332919  1.00000000  0.69842442 0.46791730
## V6 -0.28936059 -0.15630387  0.49676279  0.69842442  1.00000000 0.28012915
## V7  0.19879627  0.71113407  0.03282961  0.46791730  0.28012915 1.00000000
## V8  0.34869847  0.41359462 -0.13908580 -0.17127417 -0.20827738 0.41682128
## V9  0.31867736  0.83495175 -0.25835810  0.31275728 -0.08123414 0.70158588
##            V8          V9
## V2  0.3486985  0.31867736
## V3  0.4135946  0.83495175
## V4 -0.1390858 -0.25835810
## V5 -0.1712742  0.31275728
## V6 -0.2082774 -0.08123414
## V7  0.4168213  0.70158588
## V8  1.0000000  0.39886792
## V9  0.3988679  1.00000000
```

其样本相关系数矩阵如上所示. 其 V2 代表 $x1$, 其余类似.

3.2 (2)

```
analysis3 <- princomp(prob3,cor = TRUE)
summary(analysis3)
```

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
## Standard deviation     1.759627 1.5385783 0.9591597 0.84019364 0.70600448
## Proportion of Variance 0.387036 0.2959029 0.1149984 0.08824067 0.06230529
## Cumulative Proportion  0.387036 0.6829389 0.7979373 0.88617801 0.94848330
```

```
##                            Comp.6     Comp.7      Comp.8
## Standard deviation     0.47946669 0.36162932 0.226868982
## Proportion of Variance 0.02873604 0.01634697 0.006433692
## Cumulative Proportion  0.97721934 0.99356631 1.000000000
```

则其主成分的贡献率如上表 Proportion of Variance 所示.

前两个主成分的累积贡献率为 0.6829, 从表中 Cumulative Proportion 得到.

3.3 (3)

```
analysis3$loadings
```

```
##
## Loadings:
##    Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## V2  0.250  0.241  0.694  0.377  0.502
## V3  0.519                0.225 -0.424         0.282  0.643
## V4        -0.475  0.578        -0.510  0.173 -0.381
## V5  0.254 -0.538         0.231        -0.399  0.472 -0.458
## V6        -0.575        -0.285  0.516 -0.146 -0.159  0.521
## V7  0.493 -0.135 -0.145 -0.224  0.177  0.755        -0.244
## V8  0.317  0.261  0.286 -0.768        -0.355  0.131
## V9  0.509        -0.271  0.177        -0.305 -0.708 -0.181
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var  0.125  0.125  0.125  0.125  0.125  0.125  0.125  0.125
## Cumulative Var  0.125  0.250  0.375  0.500  0.625  0.750  0.875  1.000
```

从表中我们可以看出前两个主成分为

$$Y_1 = 0.250X_1 + 0.519X_2 + 0.254X_4 + 0.493X_6 + 0.317X_7 + 0.509X_8 \tag{35}$$

$$Y_2 = 0.241X_1 - 0.475X_3 - 0.538X_4 - 0.575X_5 - 0.135X_6 + 0.261X_7 \tag{36}$$

其含义为在样本在量纲统一后, 包含特征信息最多的两个方向上面的值.

其从大到小的排序结果为 (前五个)

```
row.names(prob3[order(as.matrix(prob3)%*%analysis3$loadings[,1],decreasing = TRUE),])
```

```
##  [1] "广东"   "上海"   "北京"   "海南"   "浙江"   "广西"   "福建"   "天津"
##  [9] "江苏"   "四川"   "辽宁"   "西藏"   "湖南"   "云南"   "湖北"   "山东"
## [17] "安徽"   "贵州"   "宁夏"   "新疆"   "陕西"   "河北"   "江西"   "吉林"
## [25] "青海"   "甘肃"   "黑龙江" "河南"   "内蒙古" "山西"
```

4 Problem 4

4. (P157 4.7 题) 设 $\mathbf{X}=(X_1,X_2)^T$， $\mathbf{Y}=(Y_1,Y_2)^T$ 的联合协方差矩阵为

$$\mathbf{\Sigma}=\begin{pmatrix}\mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12}\\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22}\end{pmatrix}=\left(\begin{array}{cc:cc} 100 & 0 & 0 & 0\\ 0 & 1 & 0.95 & 0\\ \hdashline 0 & 0.95 & 1 & 0\\ 0 & 0 & 0 & 100\end{array}\right)$$

(1) 证明 $\mathbf{X}$ 和 $\mathbf{Y}$ 的第一对典型相关变量为 $U_2=X_2, V_1=Y_1$，其典型相关系数为 0.95；

(2) 求 $\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$ 的特征值，它们和 $\mathbf{\Sigma}_{11}^{-\frac{1}{2}}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-\frac{1}{2}}$ 的特征值是否相同？

4.1 (1)

$$A=\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2} \tag{37}$$

$$=\begin{pmatrix}\frac{1}{10} & \\ & 1\end{pmatrix}\begin{pmatrix} & \\ 0.95 & \end{pmatrix}\begin{pmatrix}1 & \\ & \frac{1}{100}\end{pmatrix}\begin{pmatrix} & 0.95\\ & \end{pmatrix}\begin{pmatrix}\frac{1}{10} & \\ & 1\end{pmatrix} \tag{38}$$

$$=\begin{pmatrix}0 & 0\\ 0 & 0.95^2\end{pmatrix} \tag{39}$$

故 A 的第一个特征值为 0.95^2. 并且这个特征值对应的特征向量为 $(0,1)^T$. 则典型相关系数为 0.95. 故 $U_1 = (0,1)\Sigma_{11}^{-1/2}X = X_2$

在考虑 B,

$$B = \Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1/2} \tag{40}$$

$$= \begin{pmatrix} 1 & \\ & \frac{1}{10} \end{pmatrix}\begin{pmatrix} & 0.95 \\ & \end{pmatrix}\begin{pmatrix} \frac{1}{100} & \\ & 1 \end{pmatrix}\begin{pmatrix} & \\ 0.95 & \end{pmatrix}\begin{pmatrix} 1 & \\ & \frac{1}{10} \end{pmatrix} \tag{41}$$

$$= \begin{pmatrix} 0.95^2 & 0 \\ 0 & 0 \end{pmatrix} \tag{42}$$

故 B 的特征值 0.95^2 所对应的特征向量为 $(1,0)^T$, 故 $V_1 = (0,1)\Sigma_{22}^{-1/2}Y = Y_2$.

4.2 (2)

$$\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \begin{pmatrix} 1/10 & \\ & 1 \end{pmatrix}\begin{pmatrix} & \\ 0.95 & \end{pmatrix}\begin{pmatrix} 1 & \\ & 1/10 \end{pmatrix}\begin{pmatrix} & 0.95 \\ & \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0.95^2 \end{pmatrix} \tag{43}$$

其与 A 相等, 故其特征值分解也相同, 特征值也相同.

5 Problem 5

5. (P157 4.9 题) 下表 2(见 QQ 群 exercise4.9.txt) 是 25 个家庭的成年长子的头长(X_1)和头宽(X_2) 与成年次子的头长(Y_1) 和头宽(Y_2) 的测量数据. 试分别从样本协方差矩阵 $\mathbf{\Sigma}$ 和样本相关系数矩阵 $\mathbf{R}$ 出发作典型相关分析，求各典型变量对及典型相关系数，检验各典型变量对是否显著相关($\alpha = 0.05$)?两种情况下的结果有何异同?

表 2. 25 个家庭的成年长子与成年次子的头长和头宽数据.

```
prob5 <- read.table("./exercise4_9.txt", row.names=1, quote="\"", comment.char="")
p<-2
q<-2
```

5.1 从协方差矩阵出发

首先计算 A,B 以及其特征值分解

```
M5 <- cov(prob5)
A <- solve(M5[1:p,1:q]) %*% M5[1:p,(q+1):(p+q)] %*% solve(M5[(p+1):(p+q),(q+1):(p+q)]) %*% M5[(p+1
B <- solve(M5[(p+1):(p+q),(q+1):(p+q)]) %*% M5[(p+1):(p+q),1:p] %*% solve(M5[1:p,1:q])  %*% M5[1:p
A.eig <- eigen(A)
B.eig <- eigen(B)
```

```
sqrt(A.eig$values)
```

```
## [1] 0.7885079 0.0537397
```

```
t(A.eig$vectors)%*%solve(M5[1:p,1:q]%^%{1/2})
```

```
##              V2        V3
## [1,]  0.6245384 0.7809941
## [2,] -0.5993391 0.8004953
```

```
t(B.eig$vectors)%*%solve(M5[(p+1):(p+q),(q+1):(p+q)]%^%{1/2})
```

```
##              V4         V5
## [1,] -0.5307857 -0.8475060
## [2,] -0.5578217  0.8299608
```

所以我们可以看到典型相关系数为 0.7785 的第一组样本典型变量为

$$U_1 = 0.6245x_1 + 0.7810x_2 \tag{44}$$

$$V_1 = -0.6838y_1 - 0.7297y_2 \tag{45}$$

典型相关系数为 0.0537 的第二组的样本典型变量为

$$U_2 = -0.5993x_1 + 0.8005x_2 \tag{46}$$

$$V_2 = -0.7091y_1 + 0.7051y_2 \tag{47}$$

```
A.T <- -(25-(p+q+3)/2)*log(accumulate(1-A.eig$values,prod,.dir = "backward"))
B.T <- -(25-(p+q+3)/2)*log(accumulate(1-B.eig$values,prod,.dir = "backward"))
1-as.numeric( map2(A.T,c(4,1),pchisq))# 计算 p-value
```

```
## [1] 0.0003218897 0.8030815916
```

我们可以看到在 0.05 的显著性水平下我们可以拒绝假设 $\rho_1 = 0$, 也就是说第一组典型变量显著相关. 同时第二组显著变量不显著相关.

5.2 从相关系数矩阵出发

首先计算 A,B 以及其特征值分解

```
M5 <- cor(prob5)
A <- solve(M5[1:p,1:q]) %*% M5[1:p,(q+1):(p+q)] %*% solve(M5[(p+1):(p+q),(q+1):(p+q)]) %*% M5[(p+1
B <- solve(M5[(p+1):(p+q),(q+1):(p+q)]) %*% M5[(p+1):(p+q),1:p] %*% solve(M5[1:p,1:q])  %*% M5[1:p
A.eig <- eigen(A)
B.eig <- eigen(B)
```

```
sqrt(A.eig$values)
```

```
## [1] 0.7885079 0.0537397
```

```
t(A.eig$vectors)%*%solve(M5[1:p,1:q]%^%{1/2})
```

```
##              V2        V3
## [1,]  0.7269968 0.6866408
## [2,] -0.7040109 0.7101892
```

```
t(B.eig$vectors)%*%solve(M5[(p+1):(p+q),(q+1):(p+q)]%^%{1/2})
```

```
##              V4         V5
## [1,] -0.6837994 -0.7296700
## [2,] -0.7091095  0.7050984
```

所以我们可以看到典型相关系数为 0.7785 的第一组样本典型变量为

$$U_1 = 0.7270x_1 + 0.6866x_2 \tag{48}$$

$$V_2 = -0.6838y_1 - 0.7297y_2 \tag{49}$$

$$\tag{50}$$

典型相关系数为 0.0537 的第二组的样本典型变量为

$$U_2 = -0.7040x_1 + 0.7102x_2 \tag{51}$$

$$V_2 = -0.7091y_1 + 0.7050y_2 \tag{52}$$

```
A.T <- -(25-(p+q+3)/2)*log(accumulate(1-A.eig$values,prod,.dir = "backward"))
B.T <- -(25-(p+q+3)/2)*log(accumulate(1-B.eig$values,prod,.dir = "backward"))
1-as.numeric( map2(A.T,c(4,1),pchisq))# 计算 p-value
```

```
## [1] 0.0003218897 0.8030815916
```

我们可以看到在 0.05 的显著性水平下我们可以拒绝假设 $\rho_1 = 0$, 也就是说第一组典型变量显著相关. 同时第二组显著变量不显著相关.

我们可以看到从两个不同的起点出发的到的典型相关系数相同但是典型相关变量不同, 显著性也相同.