多元数据分析第四次作业

应数 2101　杨嘉昱　2216113458

2024 年 5 月 21 日

4.1

Solution. 计算得 Σ 的特征值以及特征向量为

$$\sigma^2(1+\sqrt{2}\rho) \qquad \frac{1}{2}(1,\sqrt{2},1)^{\mathrm{T}}$$

$$\sigma^2 \qquad \frac{1}{\sqrt{2}}(1,0,-1)^{\mathrm{T}}$$

$$\sigma^2(1-\sqrt{2}\rho) \qquad \frac{1}{2}(1,-\sqrt{2},1)^{\mathrm{T}}$$

从而三个主成分以及主成分贡献率为

$$Y_1 = \frac{X_1+\sqrt{2}X_2+X_3}{2}, \qquad \frac{1+\sqrt{2}\rho}{3}$$

$$Y_2 = \frac{X_1-X_3}{\sqrt{2}}, \qquad \frac{1}{3}$$

$$Y_3 = \frac{X_1-\sqrt{2}X_2+X_3}{2}, \qquad \frac{1-\sqrt{2}\rho}{3}$$

□

4.2

Solution.

1. 由于 $\sigma_{ii}=1$，因此标准化变量的矩阵依然为 ρ。计算得特征值和特征向量为

$$1+2\rho, \qquad \frac{1}{\sqrt{3}}(1,1,1)^{\mathrm{T}}$$

$$1-\rho, \qquad \frac{1}{\sqrt{2}}(-1,1,0)^{\mathrm{T}}$$

$$1-\rho, \qquad \frac{1}{\sqrt{6}}(-1,-1,2)^{\mathrm{T}}$$

从而三个主成分以及主成分贡献率为

$$Y_1 = \frac{X_1+X_2+X_3}{\sqrt{3},} \qquad \frac{1+2\rho}{3}$$

$$Y_2 = \frac{-X_1+X_2}{\sqrt{2}} \qquad \frac{1-\rho}{3}$$

$$Y_3 = \frac{-X_1-X_2+2X_3}{\sqrt{6}} \qquad \frac{1-\rho}{3}$$

2. 当 p 维时，特征值为 $1+(p-1)\rho, 1-\rho(p-1$重$)$。$\lambda=1+(p-1)\rho$ 对应的特征向量为

$$\frac{1}{\sqrt{p}}(1,\cdots,1)^{\mathrm{T}}.$$

$\lambda=1-\rho$ 对应的特征向量为

$$\frac{1}{\sqrt{2}}(1,-1,0,\cdots,0)^{\mathrm{T}}$$
$$\frac{1}{\sqrt{6}}(1,1,-2,0,\cdots,0)^{\mathrm{T}}$$
$$\frac{1}{\sqrt{12}}(1,1,1,-3,0,\cdots,0)^{\mathrm{T}}$$
$$\vdots$$
$$\frac{1}{\sqrt{p(p-1)}}(1,1,\cdots,1,-p+1)^{\mathrm{T}}$$

从而主成分以及主成分贡献率为

$$Y_1=\frac{X_1+\cdots+X_p}{\sqrt{p}}, \qquad \frac{1+(p-1)\rho}{p}$$
$$Y_2=\frac{X_1-X_2}{\sqrt{2}} \qquad \frac{1-\rho}{p}$$
$$Y_3=\frac{X_1+X_2-2X_3}{\sqrt{6}} \qquad \frac{1-\rho}{p}$$
$$\vdots$$
$$Y_j=\frac{(\sum_{i=1}^{j-1}X_i)-X_j}{\sqrt{j(j-1)}}, \qquad \frac{1-\rho}{p}$$
$$\vdots$$
$$Y_p=\frac{(\sum_{i=1}^{p-1}X_i)-X_p}{\sqrt{p(p-1)}}, \qquad \frac{1-\rho}{p}$$

□

4.5

Solution.

1. 相关系数矩阵如下：

	V1	V2	V3	V4	V5	V6	V7	V8
V1	2.6073471	6.7013655	-0.2688379	-0.3852425	-1.4007138	0.9816586	0.4794828	1.8767000
V2	6.7013655	154.7227076	-0.8696283	19.3275752	-5.8285076	27.0508766	4.3810221	37.8776145
V3	-0.2688379	-0.8696283	9.3190783	6.3413886	4.5461748	0.3064821	-0.3615703	-2.8764231
V4	-0.3852425	19.3275752	6.3413886	15.1706806	8.1551610	5.5734545	-0.5680910	4.4427734
V5	-1.4007138	-5.8285076	4.5461748	8.1551610	8.9871334	2.5681579	-0.5317117	-0.8881645
V6	0.9816586	27.0508766	0.3064821	5.5734545	2.5681579	9.3520248	1.0854910	7.8248834
V7	0.4794828	4.3810221	-0.3615703	-0.5680910	-0.5317117	1.0854910	0.7251821	1.2387876
V8	1.8767000	37.8776145	-2.8764231	4.4427734	-0.8881645	7.8248834	1.2387876	13.3011266

2. 得到数据如下：

可见前两个主成分的积累贡献率为

0.9801 0.9965

3. 衣着商品，日用品与粮食支出是居民平均消费的主要原因。排序结果为

30 29 23 22 21 26 27 25 24 18 17 14 28 13 15 12 16 20 19 11 3 9 10 8 4 7 6 5 2 1

```
Importance of components:
                          Comp.1     Comp.2     Comp.3     Comp.4      Comp.5      Comp.6
Standard deviation     51.4938486 6.66314278 2.710830485 1.165623798 0.5941717918 4.632449e-01
Proportion of Variance  0.9801278 0.01641082 0.002716294 0.000502214 0.0001304957 7.932195e-05
Cumulative Proportion   0.9801278 0.99653864 0.999254932 0.999757146 0.9998876418 9.999670e-01
                             Comp.7 Comp.8
Standard deviation     2.989574e-01      0
Proportion of Variance 3.303623e-05      0
Cumulative Proportion  1.000000e+00      1

Loadings:
   Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
V1         0.125  0.133  0.121  0.537  0.713  0.374  0.116
V2  0.951         0.188               -0.190  0.123
V3        -0.467  0.633 -0.250 -0.450  0.337
V4        -0.694 -0.153  0.459  0.281        -0.411  0.162
V5        -0.505 -0.378 -0.173        -0.166  0.732
V6  0.158        -0.400 -0.748  0.152  0.294 -0.368
V7                       -0.153        -0.161         0.970
V8  0.237  0.120 -0.469  0.315 -0.633  0.448
```

所用代码如下：

```r
D <- read.table('exercise4_5.txt')
S <- cov(D)
pr <- princomp(S)
summary(pr, loadings = TRUE)
pc1.coef <- pr$loadings[,1]
score <- as.matrix(D) %*% pc1.coef
sort.obj <- sort(score, decreasing = TRUE, index.return = TRUE)
sort.obj

```

□

4.7

Solution.

1. 设

$$U_1 = a^{\mathrm{T}}X, \qquad V_1 = b^{\mathrm{T}}Y,$$

那么 a, b 满足方程

$$\begin{cases} \max a^{\mathrm{T}}\Sigma_{12}b \\ a^{\mathrm{T}}\Sigma_{11}a = b^{\mathrm{T}}\Sigma_{22}b = 1 \end{cases} \iff \begin{cases} \max 0.95 a_2 b_1 \\ 100a_1^2 + a_2^2 = b_1^2 + 100b_2^2 = 1 \end{cases}$$

解得

$$a = (0,1)^{\mathrm{T}}, \qquad b = (1,0)^{\mathrm{T}}$$

即

$$U_1 = X_2, \qquad V_1 = Y_1, \qquad \rho_{U_1,V_1} = 0.95.$$

2.

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

特征值为 $0, 0.95^2$。

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

特征值为 $0, 0.95^2$。从而特征值相同。

□

4.8

Solution.

1. 第一对典型变量为

$$U_1 = -1.2874680X_1 + 0.7932944X_2, \qquad V_1 = -1.0300605Y_1 + 0.7877876Y_2$$

相关系数为 0.37385370。

第二对典型变量为

$$U_2 = 0.02293401X_1 - 1.01428981X_2, \qquad V_2 = -0.3913529Y_1 - 0.7704365Y_2$$

相关系数为 0.07742401。

2. 两对典型变量的 T 值以及 p 值分别为

$$\begin{matrix} 21.3709908 & 0.8207084 \\ 0.0002672978 & 0.3649731165 \end{matrix}$$

可见第一对典型变量显著相关。

所用代码如下

```r
R11 = matrix(c(1,0.63,0.63,1), nrow = 2, ncol = 2)
R12 = matrix(c(0.24,0.06,-0.06,0.07), nrow = 2, ncol = 2)
R21 = t(R12)
R22 = matrix(c(1,0.42,0.42,1), nrow = 2, ncol = 2)
eig.obj <- eigen(R11)
eigen.mat <- eig.obj$vectors
eigen.val <- eig.obj$values
R11.minus.05 <- eigen.mat %*% diag(1/sqrt(eigen.val)) %*% t(eigen.mat)

A <- R11.minus.05 %*% R12 %*% solve(R22) %*% R21 %*% R11.minus.05
A.eig.obj <- eigen(A)
A.coef <- matrix(0,2,2)
for (i in 1:2) {
    A.coef[,i] <- t(A.eig.obj$vectors[,i]) %*% R11.minus.05
}
A.coef
sqrt(A.eig.obj$values)

eig.obj <- eigen(R22)
eigen.mat <- eig.obj$vectors
eigen.val <- eig.obj$values
R22.minus.05 <- eigen.mat %*% diag(1/sqrt(eigen.val)) %*% t(eigen.mat)

B <- R22.minus.05 %*% R21 %*% solve(R11) %*% R12 %*% R22.minus.05
B.eig.obj <- eigen(B)
B.coef <- matrix(0,2,2)
for (i in 1:2) {
    B.coef[,i] <- t(B.eig.obj$vectors[,i]) %*% R22.minus.05
}
B.coef
sqrt(B.eig.obj$values)

r <- sqrt(B.eig.obj$values)
n <- 140
```

```
q <- 2
p <- 2
m <- length(r)
T.stat <- rep(0,m)
fac <- n - 0.5 * (p + q + 3)
p.value <- rep(0,m)
for (k in 1:m) {
    vector <- 1 - r ^ 2
    Lambda <- prod(vector[k : m])
    T.stat[k] <- -fac * log(Lambda)
    df <- (p-k+1)*(q-k+1)
    p.value[k] <- 1 - pchisq(T.stat[k], df)
}
T.stat
p.value
```

□

4.9

Solution.

1. 第一对典型变量为

$$U_1 = -0.5521896X_1 - 0.5215372X_2 \qquad V_1 = 0.5044484Y_1 + 0.5382877Y_2$$

相关系数为 0.7885079。

第二对典型变量为

$$U_2 = 1.366374X_1 - 1.378365X_2, \qquad V_2 = -1.768570Y_1 + 1.758566Y_2$$

相关系数为 0.0537397。

2. 两对典型变量的 T 值以及 p 值分别为

$$\begin{matrix} 133.0981659 & 0.3947763 \\ 0.0000000 & 0.5297994 \end{matrix}$$

可见第一对典型变量显著相关。

所用代码如下

```
D <- read.table('exercise4_9.txt')
S <- cor(D)
R11 <- S[1:2,1:2]
R12 <- S[1:2,3:4]
R21 <- t(R12)
R22 <- S[3:4,3:4]

eig.obj <- eigen(R11)
eigen.mat <- eig.obj$vectors
eigen.val <- eig.obj$values
R11.minus.05 <- eigen.mat %*% diag(1/sqrt(eigen.val)) %*% t(eigen.mat)

A <- R11.minus.05 %*% R12 %*% solve(R22) %*% R21 %*% R11.minus.05
A.eig.obj <- eigen(A)
A.coef <- matrix(0,2,2)
```

```r
for (i in 1:2) {
    A.coef[,i] <- t(A.eig.obj$vectors[,i]) %*% R11.minus.05
}
A.coef
sqrt(A.eig.obj$values)

eig.obj <- eigen(R22)
eigen.mat <- eig.obj$vectors
eigen.val <- eig.obj$values
R22.minus.05 <- eigen.mat %*% diag(1/sqrt(eigen.val)) %*% t(eigen.mat)

B <- R22.minus.05 %*% R21 %*% solve(R11) %*% R12 %*% R22.minus.05
B.eig.obj <- eigen(B)
B.coef <- matrix(0,2,2)
for (i in 1:2) {
    B.coef[,i] <- t(B.eig.obj$vectors[,i]) %*% R22.minus.05
}
B.coef
sqrt(B.eig.obj$values)

r <- sqrt(B.eig.obj$values)
n <- 140
q <- 2
p <- 2
m <- length(r)
T.stat <- rep(0,m)
fac <- n - 0.5 * (p + q + 3)
p.value <- rep(0,m)
for (k in 1:m) {
    vector <- 1 - r ^ 2
    Lambda <- prod(vector[k : m])
    T.stat[k] <- -fac * log(Lambda)
    df <- (p-k+1)*(q-k+1)
    p.value[k] <- 1 - pchisq(T.stat[k], df)
}
T.stat
p.value
```

□