实验七：因子分析

沈雨萱 3180104691

目录

1 实验概况

一．实验目的与要求：通过本试验项目，使学生理解并掌握如下内容（1）熟悉潜在因子模型载荷矩阵的不同估计方法；（2）熟悉潜在因子个数的确定方法，因子得分的计算；（3）能够利用因子模型（或正交旋转）对所考虑问题做出合理的解释；

二．实验内容

（1）我国 2010 你那各地区城镇居民家庭平均每人全年消费数据如 ex6.7 所示，这些数据指标分别从食品（x1）, 衣着（x2），居住 (x3)，医疗 (x4)，交通通信 (x5)，教育 (x6)，家政 (x7) 和耐用消费品 (x8) 来描述消费。试对该数据进行因子分析。

（2）采用 “体检数据”。这是一组 4000 多个样本的体检资料，分别有常规体检的一系列指标，其中，体检数据，请考虑下面的问题：一、利用主成分方法变量进行降维，然后进行相应的主成分方法聚类分析；二、构建因子分析模型，进行因子旋转，分析每个因子的意义及这些潜在的因子与年龄的关系。

2 实验结果

2.1 （1）

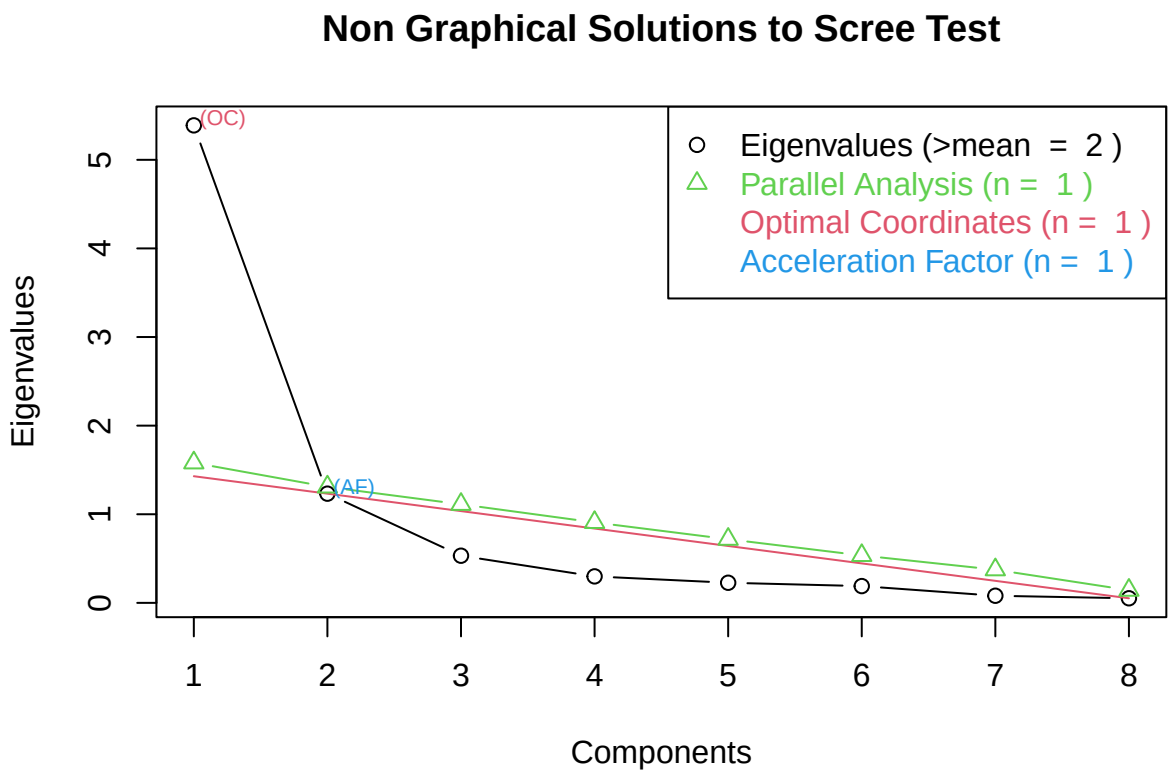
（1）我国 2010 你那各地区城镇居民家庭平均每人全年消费数据如 ex6.7 所示，这些数据指标分别从食品（x1）, 衣着（x2），居住 (x3)，医疗 (x4)，交通通信 (x5)，教育 (x6)，家政 (x7) 和耐用消费品 (x8) 来描述消费。试对该数据进行因子分析。

```
data_0 <- read.csv("ex6.7.csv",encoding = "UTF-8",na.strings=c("","  ","NA"),header=T)
data_0 <- na.omit(data_0)
row.names(data_0) <- data_0[,1]
X<-data_0[,-1]
# 确定应提取的因子个数
library(nFactors)
```

```
##
## 载入程辑包: 'nFactors'
```

```
## The following object is masked from 'package:lattice':
##
##     parallel
```

```
ev <- eigen(cor(X)) # 获取特征值
ap <- parallel(subject=nrow(X),var=ncol(X),
  rep=100,cent=.05) # subject 指样本个数, var 是指变量个数
nS <- nScree(x=ev$values, aparallel=ap$eigen$qevpea) # 确定探索性因子分析中应保留的因子
plotnScree(nS) # 绘制碎石图
```



```r
# 因子分析
fre<-factanal(X, 3, scores="Bartlett", rotation="none")
fre
```

```
##
## Call:
## factanal(x = X, factors = 3, scores = "Bartlett", rotation = "none")
##
## Uniquenesses:
##       食品       衣着       居住       医疗   交通通讯       教育   家庭服务
##      0.108      0.426      0.005      0.200      0.041      0.253      0.108
## 耐用消费品
##      0.292
##
## Loadings:
##            Factor1 Factor2 Factor3
## 食品        0.710           0.621
## 衣着        0.402   0.630   0.124
## 居住        0.993
## 医疗        0.564   0.669  -0.186
## 交通通讯    0.821           0.533
## 教育        0.787   0.225   0.277
## 家庭服务    0.836           0.438
## 耐用消费品  0.730   0.398   0.127
##
##                Factor1 Factor2 Factor3
## SS loadings      4.497   1.057   1.014
## Proportion Var   0.562   0.132   0.127
## Cumulative Var   0.562   0.694   0.821
##
## Test of the hypothesis that 3 factors are sufficient.
## The chi square statistic is 9.15 on 7 degrees of freedom.
## The p-value is 0.242
```

因为 p=0.242 $>$ 0.05，因此这三个因子足够解释这些变量

2.2 （2）

（2）采用“体检数据”。这是一组 4000 多个样本的体检资料，分别有常规体检的一系列指标，其中，体检数据，请考虑下面的问题：一、利用主成分方法变量进行降维，然后进行相应的主成分方法聚类分析；二、构建因子分析模型，进行因子旋转，分析每个因子的意义及这些潜在的因子与年龄的关系。

```
data_0 <- read.csv("exam.csv",encoding = "UTF-8",na.strings=c("","  ","NA","  未检"),header=T,row.nam
# 处理异常值
for (i in colnames(data_0)){
  if (i=='Gender'){
    data_0[,i] <- as.numeric(factor(data_0[,i]))
  }else{
    data_0[,i] <- impute(as.numeric(data_0[,i]),mean)
  }
}
sum(is.na(data_0))
```

```
## [1] 0
```

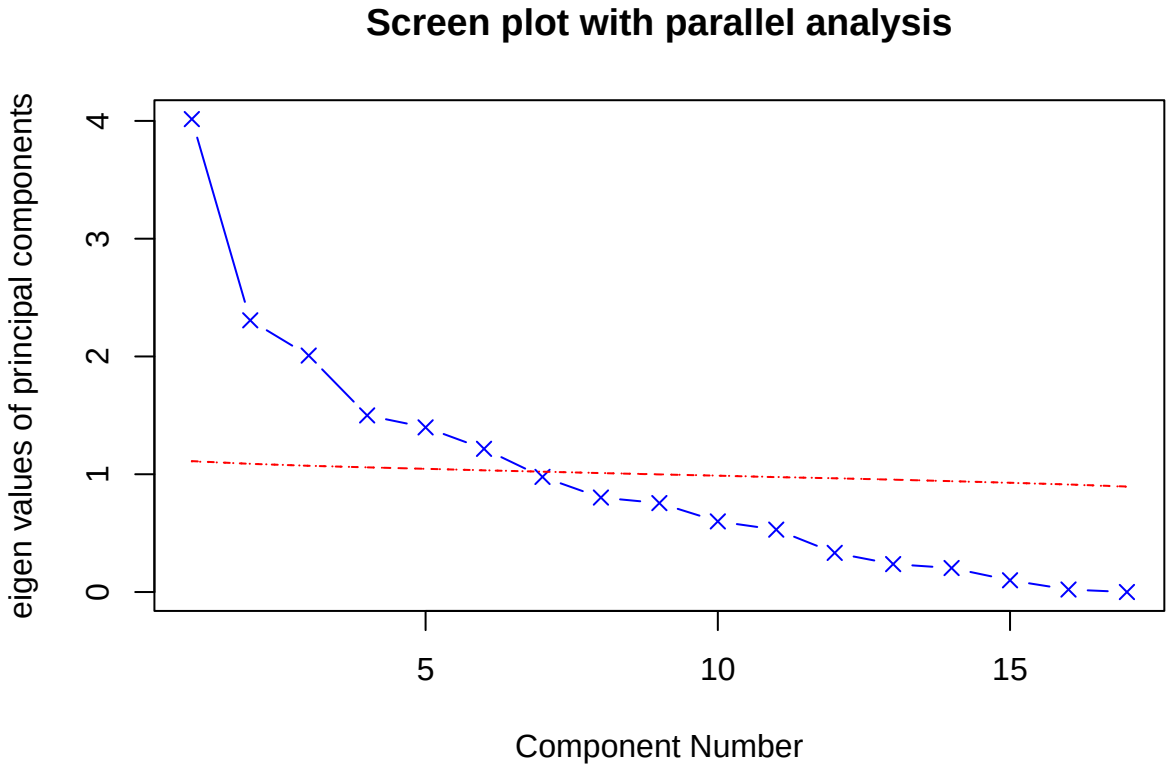
```
X <- data_0
# 主成分个数分析
library(psych)
```

```
##
## 载入程辑包: 'psych'

## The following object is masked from 'package:Hmisc':
##
##     describe

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
plot.new()
fa.parallel(data_0,fa="pc",n.iter=100,show.legend=FALSE,main="Screen plot with parallel analysis")
```

```
## In smc, smcs < 0 were set to .0
```

```
## In smc, smcs < 0 were set to .0
## In smc, smcs < 0 were set to .0

## In factor.scores, the correlation matrix is singular, an approximation is used
```

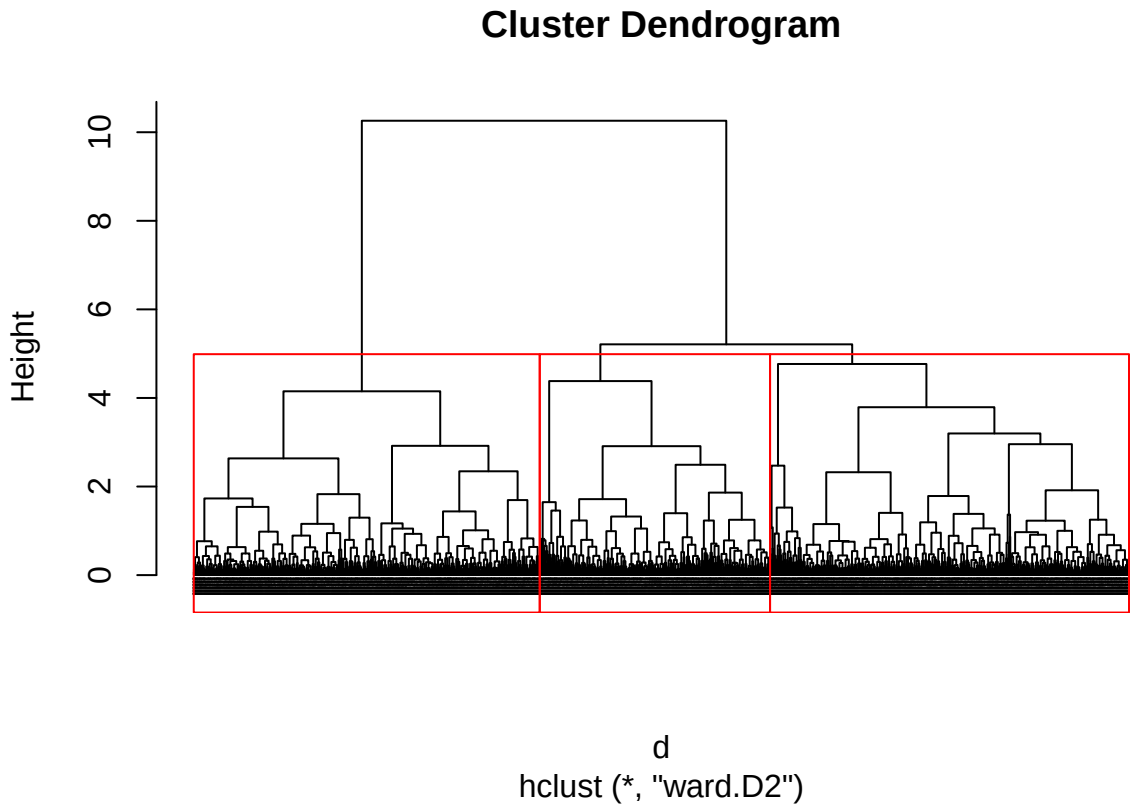



```
## Parallel analysis suggests that the number of factors =  NA  and the number of components =  6
```

选择 6 个主成分，对其进行聚类分析

```
pc6<-principal(data_0,nfactors=6,rotate="none")
X1 <- pc6[["scores"]]
# 聚类分析
#cl_single <- NbClust(X1, distance="euclidean",
#                min.nc=2, max.nc=15, method="ward.D2")
center<-sweep(X1, 2, apply(X1, 2, mean))# 按列中心化
R<-apply(X1, 2, max)-apply(X1, 2, min)# 计算列极差
X_star<-sweep(center, 2, R, "/")# 极差标准化，均值为 0，极差为 1
d<-dist(X_star,method = "euclidean")
```

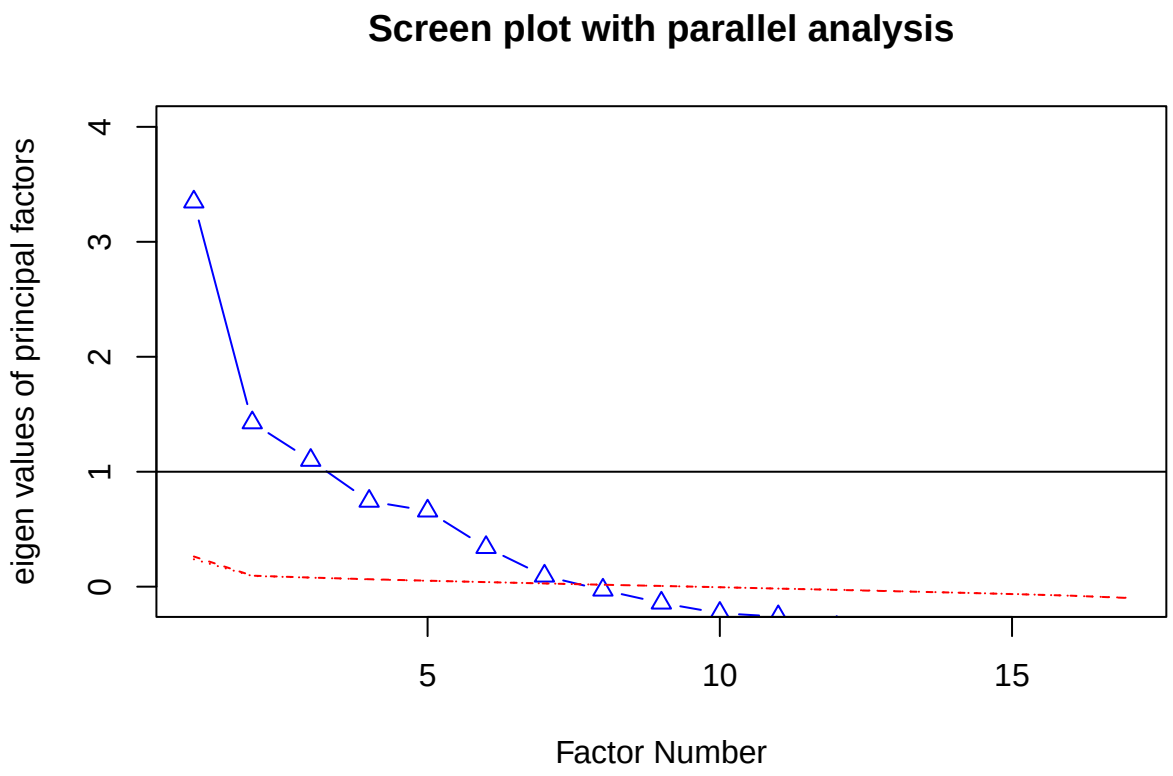
```
model1=hclust(d,method='ward.D2')
result=cutree(model1,k=3)
plot(model1,cex=0.1,hang=-1);re1<-rect.hclust(model1, k=3, border="red")
```



二、构建因子分析模型，进行因子旋转，分析每个因子的意义及这些潜在的因子与年龄的关系。

```
# 分析因子数
plot.new()
fa.parallel(data_0,fa="fa",n.iter=100,show.legend=FALSE,main="Screen plot with parallel analysis")
```

```
## In smc, smcs < 0 were set to .0
## In smc, smcs < 0 were set to .0
## In smc, smcs < 0 were set to .0

## In factor.scores, the correlation matrix is singular, an approximation is used
```




```
## Parallel analysis suggests that the number of factors =  7  and the number of components =  NA
```

得到因子个数为 7

```
# 因子分析
fre<-factanal(data_0, 7, scores="Bartlett", rotation="varimax")
fre
```

```
##
## Call:
## factanal(x = data_0, factors = 7, scores = "Bartlett", rotation = "varimax")
##
## Uniquenesses:
##   Gender      Age      Sbp      Dbp Sphygmus   Weight   Height       TC
##    0.311    0.538    0.279    0.281    0.696    0.293    0.302    0.627
##       TG      ALT      AST    T.BIL       IB      ALP       TP      Alb
##    0.580    0.146    0.161    0.033    0.033    0.664    0.005    0.005
##      GLB
```

```
##    0.005
##
## Loadings:
##          Factor1 Factor2 Factor3 Factor4 Factor5 Factor6 Factor7
## Gender   -0.741  -0.142  -0.124   0.121  -0.134  -0.129  -0.204
## Age                                      0.219  -0.265   0.533
## Sbp       0.178                          0.799           0.277
## Dbp       0.277           0.117          0.774           0.233
## Sphygmus -0.156                   0.124   0.263          -0.193
## Weight    0.756           0.211          0.201           0.194
## Height    0.843
## TC                                                       0.515
## TG        0.220           0.139   0.164                  0.467
## ALT       0.204           0.896          0.101           0.131
## AST                       0.898                          0.178
## T.BIL     0.112   0.976
## IB        0.114   0.976
## ALP       0.155           0.190          0.112           0.317
## TP                                0.849          0.503   0.117
## Alb       0.151   0.125           0.122          0.965
## GLB      -0.142                   0.971                  0.149
##
##                Factor1 Factor2 Factor3 Factor4 Factor5 Factor6 Factor7
## SS loadings      2.159   1.968   1.770   1.756   1.466   1.318   1.201
## Proportion Var   0.127   0.116   0.104   0.103   0.086   0.078   0.071
## Cumulative Var   0.127   0.243   0.347   0.450   0.536   0.614   0.685
##
## Test of the hypothesis that 7 factors are sufficient.
## The chi square statistic is 116345.9 on 38 degrees of freedom.
## The p-value is 0
```

因子意义分析：Factor1 主要与年龄、身高、重量有关，Factor2 主要与 T.BIL、IB 有关，Factor3 主要与 ALT、AST 有关，Factor4 主要与 TP、GLB 有关，Factor5 主要与 Alb 有关，Factor6 主要与 Age、TG、TC 有关。Factor5、6、7 与年龄有潜在关系，其中 Factor7 有较强的正相关，Factor6 有一定负相关，Factor5 有一定正相关。