实验七：因子分析

林泽钦 3160104013 统计 1601

目录

1 实验目的和要求

1.1 实验目的

通过本试验项目，能够理解并掌握如下内容：

- 熟悉潜在因子模型载荷矩阵的不同估计方法；
- 熟悉潜在因子个数的确定方法，因子得分的计算；
- 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；

1.2 实验内容

1.2.1 城镇居民消费数据分析

我国 2010 年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 (x_1)，衣着 (x_2)，居住 (x_3)，医疗 (x_4)，交通通

信 (x_5)，教育 (x_6)，家政 (x_7)，和耐用消费品 (x_8) 来描述消费。试对该数据进行因子分析。

1.2.2 体检数据分析

采用体检数据进行分析。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，请考虑下面的问题：

- 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；
- 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。

1.3 实验环境

- R-3.5.2
- RStudio

1.4 实验所用程序包

```
library(tidyverse)
library(VIM)
library(Hmisc)
library(psycho)
library(GPArotation)
library(factoextra)
```

2 实验过程与结果

2.1 城镇居民消费数据分析

2.1.1 数据导入与预处理

实验中选择以 csv 格式导入文件。

```
mydata1 <- read_csv("Pro7Data1.csv")
```

```
## Parsed with column specification:
## cols(
##   City = col_character(),
##   X1 = col_double(),
##   X2 = col_double(),
##   X3 = col_double(),
##   X4 = col_double(),
##   X5 = col_double(),
##   X6 = col_double(),
##   X7 = col_double(),
##   X8 = col_double()
## )
```

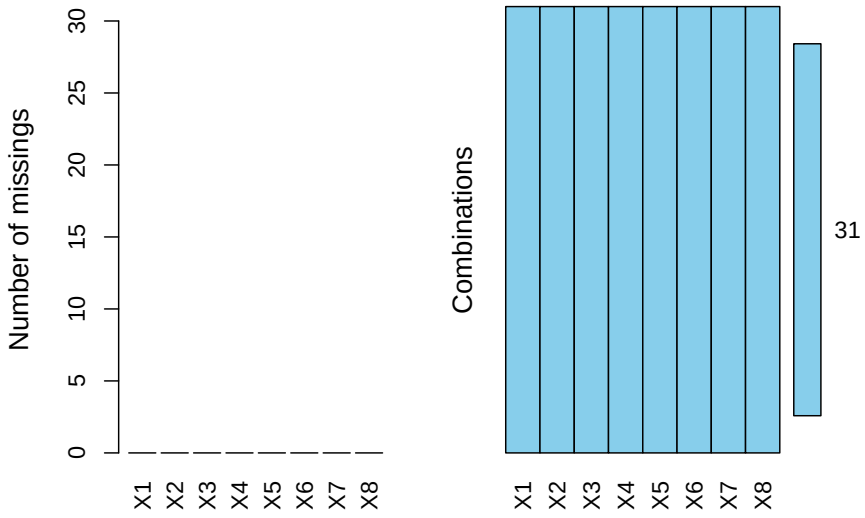
数据的第一列为城市名称，在后续分析中没有太多用处，此处将其转换为行名。

```
mydata1 <- column_to_rownames(mydata1, var = "City")
head(mydata1)
```

```
## # A tibble: 6 x 8
##      X1    X2    X3    X4    X5    X6    X7    X8
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 5562. 1572. 1286. 1563. 2293.  809.  84.7  549.
## 2 5005. 1154. 1528. 1221. 1568.  715.  45.5  468.
## 3 3155. 1137. 1097.  809. 1062.  387.  28.8  306.
## 4 2975. 1138. 1251.  770.  931.  571.  35.4  259.
## 5 3553. 1617. 1028.  870. 1192.  568.  30.5  308.
## 6 4378. 1187. 1271.  913. 1296.  670.  30.4  235.
```

接下来检查数据的完整性。这里使用的是 mice 包的 md.pattern() 函数。经检验，数据完整。

```
aggr(mydata1, prop = FALSE, numbers = TRUE)
```



2.1.2 因子数目确定

为了计算出适合的因子模型，我们需要先确定最佳的因子数目。选择最佳因子数目的方法有很多，例如 Parallel Analysis 等。不过目前哪一种方法更适合并没有定论。实验中选择使用 psycho 包中的 n_factor() 函数来解决这个问题。这个函数使用多种方法计算出相应的最佳因子数目，然后投票决定出最终的因子数目。这个函数的使用参考了博客：

https://www.r-bloggers.com/how-many-factors-to-retain-in-factor-analysis/

计算的代码如下，这里指定因子分析方法为极大似然估计，旋转方法为方差最大旋转。

```
res1 <- n_factors(mydata1, fm = "mle", rotate = "varimax")
print(res1)
```

```
## The choice of 1 factor is supported by 4 (out of 9; 44.44%) methods (Optimal Coordin
## The choice of 2 factor is supported by 4 (out of 9; 44.44%) methods (Eigenvalues (Ka
```

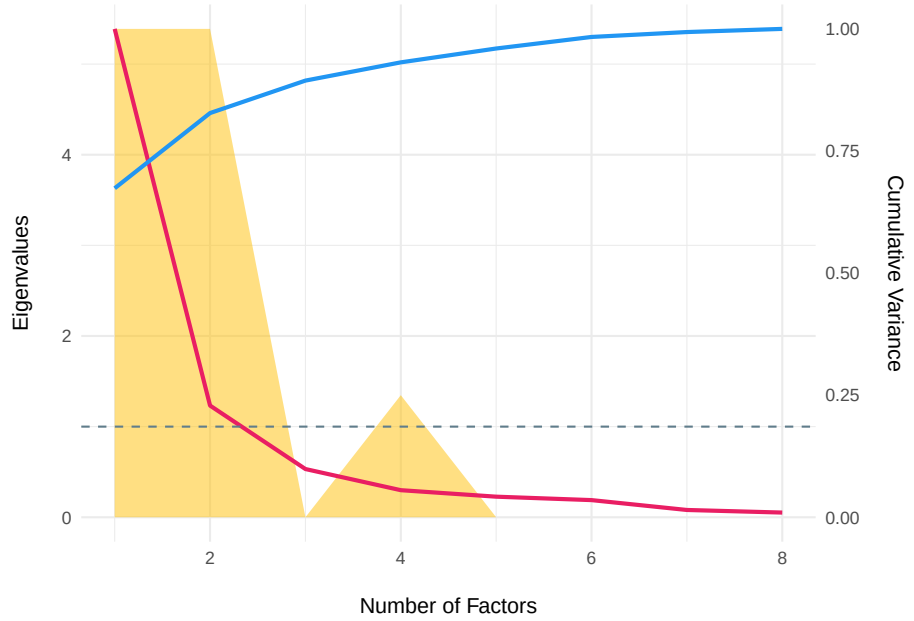
可以看到，有 4 种方法支持因子数目为 1，另有 4 种方法支持因子数目为 2，因此 n_factor() 函数建议使用这两个的其中一个。可以对返回值使用 summary() 函数，来查看各个因子的特征值以及累计贡献率。

```
summary(res1)
```

```
## # A tibble: 8 x 4
##   n.Factors n.Methods Eigenvalues Cum.Variance
##       <int>     <dbl>       <dbl>        <dbl>
## 1         1         4      5.39          0.674
## 2         2         4      1.23          0.828
## 3         3         0      0.532         0.894
## 4         4         1      0.299         0.931
## 5         5         0      0.227         0.960
## 6         6         0      0.189         0.983
## 7         7         0      0.0803        0.994
## 8         8         0      0.0518        1.
```

绘制成较为直观的图像如下。橙色阴影表示有几种方法认为该因子数量最佳，红色折线为特征值，蓝色折线为累计贡献率。

```
plot(res1)
```



另外具体的各种方法选出的最佳因子数如下：

```
res1$values$methods
```

```
##                            Method n_optimal
## 1             Optimal Coordinates         1
## 2             Acceleration Factor         1
## 3               Parallel Analysis         1
## 4 Eigenvalues (Kaiser Criterion)         2
## 5                     Velicer MAP         2
## 6                             BIC         2
## 7       Sample Size Adjusted BIC         4
## 8               VSS Complexity 1         1
## 9               VSS Complexity 2         2
```

2.1.3 因子模型

实验最终选择因子数目为 2，得到的模型为：

```
fa1 <- factanal(mydata1, factor = 2,
                fm = "mle", rotation = "varimax", scores = "Bartlett")
fa1
```

```
##
## Call:
## factanal(x = mydata1, factors = 2, scores = "Bartlett", rotation = "varimax",     fm
##
## Uniquenesses:
##    X1    X2    X3    X4    X5    X6    X7    X8
## 0.123 0.544 0.281 0.191 0.043 0.256 0.106 0.272
##
## Loadings:
##    Factor1 Factor2
## X1 0.925   0.147
## X2 0.232   0.634
## X3 0.659   0.533
## X4 0.126   0.891
```

```
## X5 0.941   0.268
## X6 0.699   0.505
## X7 0.888   0.324
## X8 0.515   0.680
##
##                Factor1 Factor2
## SS loadings      3.788   2.396
## Proportion Var   0.473   0.300
## Cumulative Var   0.473   0.773
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 14.64 on 13 degrees of freedom.
## The p-value is 0.331
```

可以看到，因子 1 中载荷较高的有食品 (x_1)，交通通信 (x_5)，家政 (x_7)，中等的有居住 (x_3)，教育 (x_6)，而因子 2 中较重要的有衣着 (x_2)，医疗 (x_4) 以及耐用消费品 (x_8)。尽管已经经过了旋转，因子所代表的意义解释起来还是具有一定的困难的。

因子得分如下：

```
fa1$scores
```

```
##             Factor1      Factor2
## 北京     0.741087533  2.948117209
## 天津     0.009591327  1.954436960
## 河北    -0.685241873  0.435826001
## 山西    -0.760620289  0.533388513
## 内蒙古 -0.571081866  0.889116990
## 辽宁    -0.128219177  0.489609341
## 吉林    -0.885486455  0.867359267
## 黑龙江 -1.121370234  0.532820921
## 上海     3.617674506  0.001231463
## 江苏     0.335221600  0.309580949
## 浙江     1.615432379  0.601382362
## 安徽    -0.342449419 -0.360772629
```

```
## 福建    1.287197023 -0.860279225
## 江西   -0.334559010 -1.080829734
## 山东   -0.218405404  0.852521723
## 河南   -0.895704096  0.520270829
## 湖北   -0.444660327 -0.232517232
## 湖南   -0.410419874  0.065313010
## 广东    2.417499233 -0.257987812
## 广西    0.320323974 -1.158415844
## 海南    0.315967392 -1.323972133
## 重庆   -0.352251575  0.819269805
## 四川    0.044540766 -1.106787241
## 贵州   -0.299486199 -1.285142622
## 云南    0.011214946 -1.330191591
## 西藏    0.051514138 -2.571861893
## 陕西   -0.595585733  0.595144543
## 甘肃   -0.779127385 -0.324423753
## 青海   -0.777838603 -0.583095755
## 宁夏   -0.607803855  0.456989350
## 新疆   -0.556953442 -0.396101770
```

2.2 体检数据分析

2.2.1 数据导入与预处理

读入数据，并把编号列去除。为了方便之后的分析，这里把 Gender 一列转化为 0-1 变量。

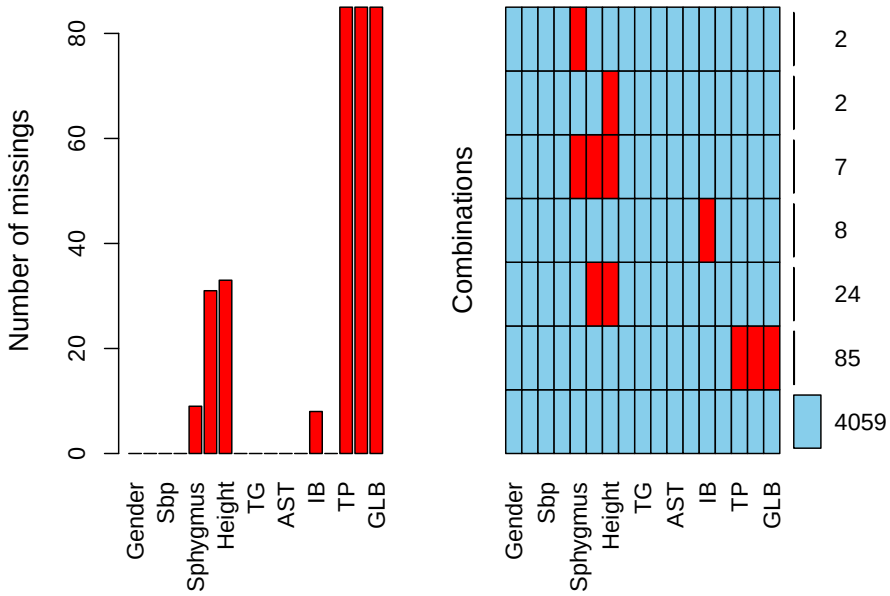
```
mydata2 <- read_csv("Pro7Data2.csv")
mydata2 <- select(mydata2, -No)
mydata2$Gender <- sapply(1:nrow(mydata2), function(i) {
  if(mydata2$Gender[i] == "男") 1 else 0
})
head(mydata2)
```

```
## # A tibble: 6 x 17
```

```
##   Gender   Age   Sbp   Dbp Sphygmus Weight Height    TC    TG   ALT   AST
##    <dbl> <dbl> <dbl> <dbl>    <dbl>  <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
## 1      1    38   115    80       78   63.7    162  5.66  2.74    38    17
## 2      1    40   124    77       97   78.2    175  5.87 10.3     51    21
## 3      0    59   136    80       82   59.5    155  4.44  2.88    20    14
## 4      1    30   126    78      112   67.2    174  3.88  1.32    50    26
## 5      1    40   121    74       86   55.1    159  5.99  1.79    43    27
## 6      1    56   131    73       86   62.6    176  3.42  0.6     28    20
## # ... with 6 more variables: `T-BIL` <dbl>, IB <dbl>, ALP <dbl>, TP <dbl>,
## #   Alb <dbl>, GLB <dbl>
```

然后检查数据的缺失情况。

```
aggr(mydata2, prop = FALSE, numbers = TRUE)
```



大约有 100 组数据存在缺失，由于占比不是很高，简单起见，这里只通过平均值来代替。

```
mydata2$Sphygmus <- impute(mydata2$Sphygmus, mean)
mydata2$IB <- impute(mydata2$IB, mean)
mydata2$TP <- impute(mydata2$TP, mean)
mydata2$Alb <- impute(mydata2$Alb, mean)
mydata2$GLB <- impute(mydata2$GLB, mean)
mydata2$Weight <- impute(mydata2$Weight, mean)
mydata2$Height <- impute(mydata2$Height, mean)
```

2.2.2 主成分降维

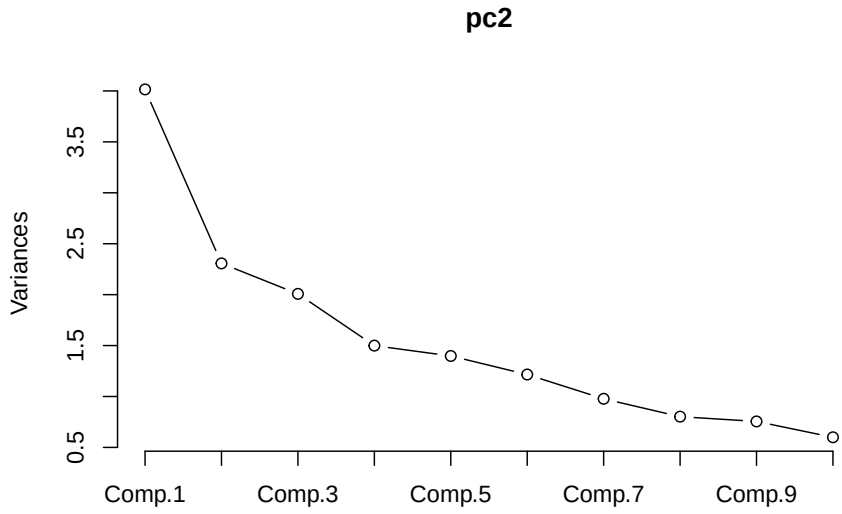
计算样本的主成分如下：

```
pc2 <- princomp(mydata2, cor = TRUE)
summary(pc2)
```

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
## Standard deviation     2.0037704 1.5188177 1.4168738 1.22454273 1.18233282
## Proportion of Variance 0.2361821 0.1356945 0.1180901 0.08820617 0.08223005
## Cumulative Proportion  0.2361821 0.3718767 0.4899667 0.57817290 0.66040296
##                            Comp.6    Comp.7     Comp.8     Comp.9
## Standard deviation     1.10260314 0.9885983 0.89554559 0.86891731
## Proportion of Variance 0.07151375 0.0574898 0.04717658 0.04441278
## Cumulative Proportion  0.73191670 0.7894065 0.83658308 0.88099586
##                           Comp.10    Comp.11    Comp.12    Comp.13
## Standard deviation     0.77428855 0.72787200 0.57644215 0.48685917
## Proportion of Variance 0.03526604 0.03116457 0.01954621 0.01394305
## Cumulative Proportion  0.91626191 0.94742648 0.96697269 0.98091574
##                           Comp.14     Comp.15     Comp.16      Comp.17
## Standard deviation     0.45130070 0.315590038 0.145475422 4.470348e-08
## Proportion of Variance 0.01198072 0.005858651 0.001244888 1.175530e-16
## Cumulative Proportion  0.99289646 0.998755112 1.000000000 1.000000e+00
```

主成分的碎石图如下。

```
screeplot(pc2, type = "lines")
```



为使主成分累积贡献率达到 85% 以上，需要前 9 个主成分。

```
pc2.score <- predict(pc2)[, 1:9]
```

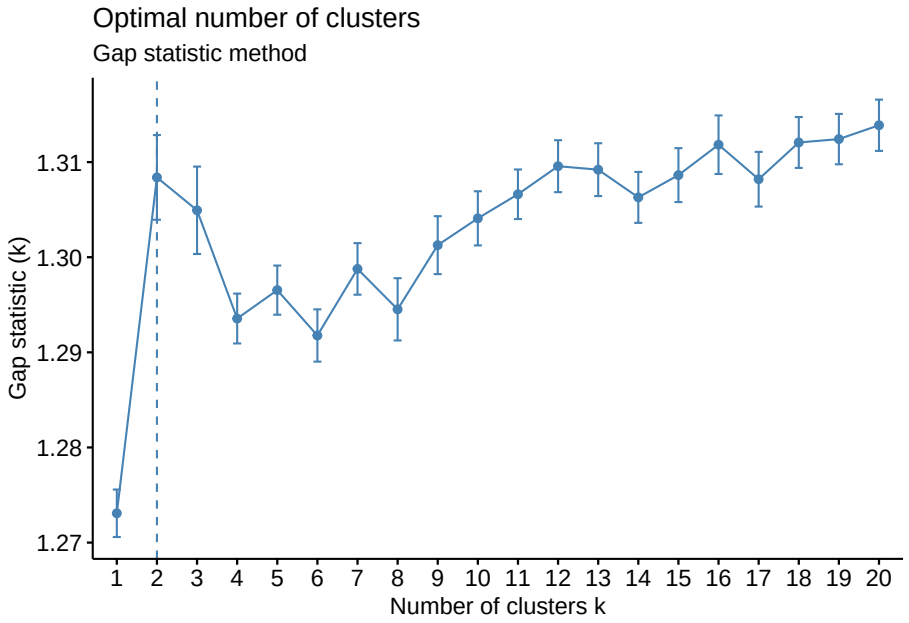
2.2.3 聚类分析

实验中使用 Gap Statistics 方法来确定聚类的数目，其具体方法以及理论支持参见 Robert Tibshirani, Guenther Walther, Trevor Hastie 在 2001 年发表的论文：“Estimating the Number of Clusters in a Data Set Via the Gap Statistic”。

factoextra 包中的 fviz_nbclust() 函数实现了这个假设检验。这里使用最大迭代次数为 30 的 kmeans 算法来进行聚类。下图展示了聚类数目为 1 到 20 时计算出的 Gap Statistic。根据算法建议，我们将样品分为 2 类。

```
fviz_nbclust(pc2.score, kmeans, k.max = 20,
             method = "gap_stat", nboot = 500,
             verbose = FALSE, iter.max = 30) +
  labs(subtitle = "Gap statistic method")
```

```
## Warning: Quick-TRANSfer stage steps exceeded maximum (= 209350)
```



kmeans 聚类得到的中心如下。

```
kmeans.res2 <- kmeans(pc2.score, centers = 2, iter.max = 30)
kmeans.res2$centers
```

```
##       Comp.1     Comp.2      Comp.3     Comp.4     Comp.5      Comp.6
## 1  1.345726  0.2918427  0.03085808  0.1183847  0.1579139  0.05017794
## 2 -1.922577 -0.4169424 -0.04408552 -0.1691308 -0.2256044 -0.07168693
##         Comp.7       Comp.8       Comp.9
## 1  0.04488807  0.003103473  0.009104332
## 2 -0.06412954 -0.004433790 -0.013006943
```

2.2.4 因子分析

这次的数据量较大，所以直接使用累积贡献率来确定公共因子数量。经计算，满足累积贡献率大于 70% 的最少的因子数量为 8。计算得到的因子模型如下：

```
fa2 <- factanal(mydata2, factor = 8,
                fm = "mle", rotation = "varimax")
fa2
```

```
##
## Call:
## factanal(x = mydata2, factors = 8, rotation = "varimax", fm = "mle")
##
## Uniquenesses:
##   Gender      Age      Sbp      Dbp Sphygmus   Weight   Height       TC
##    0.300    0.518    0.268    0.270    0.670    0.283    0.290    0.604
##       TG      ALT      AST    T-BIL       IB      ALP       TP      Alb
##    0.559    0.138    0.153    0.026    0.027    0.640    0.005    0.005
##      GLB
##    0.005
##
## Loadings:
##          Factor1 Factor2 Factor3 Factor4 Factor5 Factor6 Factor7 Factor8
## Gender    0.727   0.136   0.139  -0.103   0.195   0.153           0.229
## Age                                      0.340  -0.210   0.194   0.523
## Sbp       0.150           0.108           0.842           0.116
## Dbp       0.252           0.128           0.804           0.153
## Sphygmus -0.145                           0.191                  -0.416
## Weight    0.744           0.207           0.240           0.231
## Height    0.845
## TC                                       0.130           0.599
## TG        0.197           0.128   0.146                   0.576
## ALT       0.200           0.897           0.102           0.150
## AST                       0.908           0.114           0.115
## T-BIL     0.109   0.977
## IB        0.109   0.978
## ALP       0.137           0.216   0.108   0.179           0.102   0.325
## TP                                0.849           0.490   0.150
## Alb       0.154   0.126           0.121           0.956          -0.132
```

```
## GLB      -0.139                  0.972                  0.138
##
##                Factor1 Factor2 Factor3 Factor4 Factor5 Factor6 Factor7
## SS loadings      2.083   1.970   1.803   1.745   1.693   1.263   0.921
## Proportion Var   0.123   0.116   0.106   0.103   0.100   0.074   0.054
## Cumulative Var   0.123   0.238   0.344   0.447   0.547   0.621   0.675
##                Factor8
## SS loadings      0.648
## Proportion Var   0.038
## Cumulative Var   0.713
##
## Test of the hypothesis that 8 factors are sufficient.
## The chi square statistic is 115201.3 on 28 degrees of freedom.
## The p-value is 0
```

各因子的解释：

- 第 1 公因子：对性别，身高以及体重的影响大，可认为是基础体征因子。
- 第 2 公因子：对 T-BIL (总胆红素), IB (间接胆红素) 的影响大，可认为是胆红素因子。
- 第 3 公因子：对 ALT (谷丙转氨酶) 以及 AST (谷草转氨酶) 影响大，认为是转氨酶因子。
- 第 4 公因子：对 TB (总蛋白) 以及 GLB (球蛋白) 影响大，认为是球蛋白因子。
- 第 5 公因子：对 sbp (收缩压), dbp (舒张压) 影响大，可看作血压因子。
- 第 6 公因子：Alb (白蛋白) 因子。
- 第 7 公因子：对 TC (总胆固醇), TG (甘油三酯) 影响大，可看作脂肪因子。
- 第 8 公因子：对年龄以及脉搏影响大，认为是年龄因子。