Principal Component Analysis ,Exploratory Factor Analysis and Multidimensinal scaling with zip data

We report the results in this article to conduct Principal Component Analysis and Factor Analysis with the zip dataset. The main parts are as following:

1. Data description and missing values imputation
2. Principal component analysis
3. Factor analysis
4. Multidimensional scaling

Data Description

First, we convert the data format to rda in R with the package 'sas7bdat' and save it as 'zip1.rda' in the working path. Then we can see the details of the description of the data set.

```
setwd("F:/商业数据挖掘_光华/homework/2")

# library(sas7bdat) read data files
# zip1=read.sas7bdat('zipdemo1.sas7bdat') save(zip1,file='zip1.rda')

##### load the data files
load("zip1.rda")
summary(zip1)
```

```
##     ZIPCODE         DMAWLTHT       INCMINDX        WEALTHRT
##  Min.   : 1001   Min.   :0.00   Min.   :  3.0   Min.   :0.00
##  1st Qu.:26071   1st Qu.:2.00   1st Qu.: 69.0   1st Qu.:2.00
##  Median :49052   Median :4.00   Median : 84.0   Median :3.00
##  Mean   :49135   Mean   :4.01   Mean   : 90.3   Mean   :3.63
##  3rd Qu.:71292   3rd Qu.:6.00   3rd Qu.:104.0   3rd Qu.:5.00
##  Max.   :99929   Max.   :9.00   Max.   :409.0   Max.   :9.00
##                                 NA's   :129
##     PRCWHTE         PRCBLCK          PRCHISP          PRCUN18
##  Min.   :  0.0   Min.   : 0.00   Min.   :  0.00   Min.   :  0
##  1st Qu.: 86.0   1st Qu.: 0.00   1st Qu.:  0.00   1st Qu.: 33
##  Median : 97.0   Median : 0.00   Median :  1.00   Median : 38
##  Mean   : 87.9   Mean   : 6.23   Mean   :  3.53   Mean   : 38
##  3rd Qu.: 99.0   3rd Qu.: 4.00   3rd Qu.:  2.00   3rd Qu.: 43
##  Max.   :100.0   Max.   :99.00   Max.   :100.00   Max.   :100
##
##     PRCOWNO         PRCTHRE         PERPERHH        PRCNCD1
##  Min.   : 0.0   Min.   :  0.0   Min.   : 1.00   Min.   :  0.0
##  1st Qu.:69.0   1st Qu.: 39.0   1st Qu.: 2.50   1st Qu.: 81.0
##  Median :77.0   Median : 45.0   Median : 2.70   Median : 92.0
##  Mean   :73.9   Mean   : 44.5   Mean   : 2.76   Mean   : 86.5
##  3rd Qu.:83.0   3rd Qu.: 51.0   3rd Qu.: 2.90   3rd Qu.: 99.0
##  Max.   :99.0   Max.   :100.0   Max.   :25.50   Max.   :100.0
##                                 NA's   :114
##     MEDSCHYR        PRC25BA         PRCNCD3          PRCNC10
##  Min.   : 3.5   Min.   :  0   Min.   :  0.00   Min.   :  0.00
##  1st Qu.:12.3   1st Qu.:  8   1st Qu.:  0.00   1st Qu.:  0.00
```

```
##  Median :12.5   Median : 12   Median :  4.00   Median :  0.00
##  Mean   :12.7   Mean   : 15   Mean   :  9.44   Mean   :  4.58
##  3rd Qu.:12.9   3rd Qu.: 19   3rd Qu.: 12.00   3rd Qu.:  5.00
##  Max.   :19.7   Max.   :100   Max.   :100.00   Max.   :100.00
##  NA's   :124
##     OOMEDHVL         OOHVI          PRCOOHV          ISPSA
##  Min.   :  0.0   Min.   :  0.0   Min.   :  0.0   Min.   :   3
##  1st Qu.: 36.6   1st Qu.: 56.0   1st Qu.: 12.0   1st Qu.:2510
##  Median : 52.2   Median : 77.0   Median : 29.0   Median :2794
##  Mean   : 72.6   Mean   : 85.7   Mean   : 34.6   Mean   :2875
##  3rd Qu.: 83.6   3rd Qu.:104.0   3rd Qu.: 53.0   3rd Qu.:3169
##  Max.   :500.0   Max.   :255.0   Max.   :100.0   Max.   :5779
##                                                  NA's   :126
##     PRCRENT         PRC3544        PRC4554        PRC5564
##  Min.   :  0.0   Min.   : 0.0   Min.   : 0.0   Min.   : 0.0
##  1st Qu.: 17.0   1st Qu.:19.0   1st Qu.:14.0   1st Qu.:13.0
##  Median : 23.0   Median :21.0   Median :16.0   Median :14.0
##  Mean   : 25.8   Mean   :21.6   Mean   :15.9   Mean   :14.5
##  3rd Qu.: 30.0   3rd Qu.:24.0   3rd Qu.:18.0   3rd Qu.:16.0
##  Max.   :100.0   Max.   :59.0   Max.   :50.0   Max.   :58.0
##
##      PRC65P          PRC55P         HHMEDAGE         CEMI
##  Min.   :  0.0   Min.   :  0.0   Min.   :20.0   Min.   :  0.1
##  1st Qu.: 19.0   1st Qu.: 33.0   1st Qu.:45.0   1st Qu.: 32.1
##  Median : 24.0   Median : 39.0   Median :48.0   Median : 39.6
##  Mean   : 24.5   Mean   : 38.9   Mean   :48.8   Mean   : 43.6
##  3rd Qu.: 30.0   3rd Qu.: 45.0   3rd Qu.:52.0   3rd Qu.: 50.3
##  Max.   :100.0   Max.   :100.0   Max.   :80.0   Max.   :235.0
##                                  NA's   :114    NA's   :128
##     PRC500K         PRC200K          PRC100K       PRCHHFM
##  Min.   : 0.00   Min.   :  0.00   Min.   :  0   Min.   :  0
##  1st Qu.: 0.00   1st Qu.:  0.00   1st Qu.:  2   1st Qu.: 70
##  Median : 0.00   Median :  1.00   Median :  7   Median : 76
##  Mean   : 0.88   Mean   :  6.79   Mean   : 22   Mean   : 74
##  3rd Qu.: 0.00   3rd Qu.:  3.00   3rd Qu.: 30   3rd Qu.: 80
##  Max.   :95.00   Max.   :100.00   Max.   :100   Max.   :100
##
##     POPULAT
##  Min.   :     1
##  1st Qu.:   590
##  Median :  2050
##  Mean   :  7366
##  3rd Qu.:  8545
##  Max.   :112047
##  NA's   :866
```

```
sum(!complete.cases(zip1))
```

```
## [1] 908
```

Some NAs are included in some variables. So we simply conduct the multiple imputation with the 'mi' package in R. Since the it takes a long time to do such process, we just save the imputated dataset before. Here we also deplay the codes.

```
##### conduct the multiple imputation require(mi) zip1_com =
mi(zip1[,-1])
##### comp_zip1 = mi.data.frame(zip1_com,m=1) zip1_comp =
##### cbind(zip1[,1],comp_zip1) save(zip1_comp,file='zip1_comp.rda')
load("zip1_comp.rda")
summary(zip1_comp)
```

```
##    zip1[, 1]        DMAWLTHT       INCMINDX        WEALTHRT
##  Min.   : 1001   Min.   :0.00   Min.   :-15.1   Min.   :0.00
##  1st Qu.:26071   1st Qu.:2.00   1st Qu.: 69.0   1st Qu.:2.00
##  Median :49052   Median :4.00   Median : 84.0   Median :3.00
##  Mean   :49135   Mean   :4.01   Mean   : 90.2   Mean   :3.63
##  3rd Qu.:71292   3rd Qu.:6.00   3rd Qu.:103.0   3rd Qu.:5.00
##  Max.   :99929   Max.   :9.00   Max.   :409.0   Max.   :9.00
##     PRCWHTE         PRCBLCK         PRCHISP         PRCUN18
##  Min.   :  0.0   Min.   : 0.00   Min.   :  0.00   Min.   :  0
##  1st Qu.: 86.0   1st Qu.: 0.00   1st Qu.:  0.00   1st Qu.: 33
##  Median : 97.0   Median : 0.00   Median :  1.00   Median : 38
##  Mean   : 87.9   Mean   : 6.23   Mean   :  3.53   Mean   : 38
##  3rd Qu.: 99.0   3rd Qu.: 4.00   3rd Qu.:  2.00   3rd Qu.: 43
##  Max.   :100.0   Max.   :99.00   Max.   :100.00   Max.   :100
##     PRCOWNO         PRCTHRE         PERPERHH        PRCNCD1
##  Min.   : 0.0   Min.   :  0.0   Min.   : 0.75   Min.   :  0.0
##  1st Qu.:69.0   1st Qu.: 39.0   1st Qu.: 2.50   1st Qu.: 81.0
##  Median :77.0   Median : 45.0   Median : 2.70   Median : 92.0
##  Mean   :73.9   Mean   : 44.5   Mean   : 2.76   Mean   : 86.5
##  3rd Qu.:83.0   3rd Qu.: 51.0   3rd Qu.: 2.90   3rd Qu.: 99.0
##  Max.   :99.0   Max.   :100.0   Max.   :25.50   Max.   :100.0
##     MEDSCHYR        PRC25BA         PRCNCD3         PRCNC10
##  Min.   : 3.5   Min.   :  0   Min.   :  0.00   Min.   :  0.00
##  1st Qu.:12.3   1st Qu.:  8   1st Qu.:  0.00   1st Qu.:  0.00
##  Median :12.5   Median : 12   Median :  4.00   Median :  0.00
##  Mean   :12.7   Mean   : 15   Mean   :  9.44   Mean   :  4.58
##  3rd Qu.:12.9   3rd Qu.: 19   3rd Qu.: 12.00   3rd Qu.:  5.00
##  Max.   :19.7   Max.   :100   Max.   :100.00   Max.   :100.00
##     OOMEDHVL         OOHVI          PRCOOHV         ISPSA
##  Min.   :  0.0   Min.   :  0.0   Min.   :  0.0   Min.   :   3
##  1st Qu.: 36.6   1st Qu.: 56.0   1st Qu.: 12.0   1st Qu.:2507
##  Median : 52.2   Median : 77.0   Median : 29.0   Median :2792
##  Mean   : 72.6   Mean   : 85.7   Mean   : 34.6   Mean   :2871
##  3rd Qu.: 83.6   3rd Qu.:104.0   3rd Qu.: 53.0   3rd Qu.:3167
##  Max.   :500.0   Max.   :255.0   Max.   :100.0   Max.   :5779
##     PRCRENT         PRC3544         PRC4554         PRC5564
##  Min.   :  0.0   Min.   : 0.0   Min.   : 0.0   Min.   : 0.0
##  1st Qu.: 17.0   1st Qu.:19.0   1st Qu.:14.0   1st Qu.:13.0
##  Median : 23.0   Median :21.0   Median :16.0   Median :14.0
##  Mean   : 25.8   Mean   :21.6   Mean   :15.9   Mean   :14.5
##  3rd Qu.: 30.0   3rd Qu.:24.0   3rd Qu.:18.0   3rd Qu.:16.0
##  Max.   :100.0   Max.   :59.0   Max.   :50.0   Max.   :58.0
##      PRC65P          PRC55P          HHMEDAGE         CEMI
##  Min.   :  0.0   Min.   :  0.0   Min.   :20.0   Min.   : -4.73
##  1st Qu.: 19.0   1st Qu.: 33.0   1st Qu.:45.0   1st Qu.: 32.10
##  Median : 24.0   Median : 39.0   Median :48.0   Median : 39.60
##  Mean   : 24.5   Mean   : 38.9   Mean   :48.8   Mean   : 43.61
##  3rd Qu.: 30.0   3rd Qu.: 45.0   3rd Qu.:52.0   3rd Qu.: 50.40
##  Max.   :100.0   Max.   :100.0   Max.   :80.0   Max.   :235.00
##     PRC500K         PRC200K         PRC100K         PRCHHFM
##  Min.   : 0.00   Min.   :  0.00   Min.   :  0   Min.   :  0
##  1st Qu.: 0.00   1st Qu.:  0.00   1st Qu.:  2   1st Qu.: 70
##  Median : 0.00   Median :  1.00   Median :  7   Median : 76
##  Mean   : 0.88   Mean   :  6.79   Mean   : 22   Mean   : 74
##  3rd Qu.: 0.00   3rd Qu.:  3.00   3rd Qu.: 30   3rd Qu.: 80
##  Max.   :95.00   Max.   :100.00   Max.   :100   Max.   :100
##     POPULAT
##  Min.   :-23109
##  1st Qu.:   591
##  Median :  2101
##  Mean   :  7424
##  3rd Qu.:  8946
##  Max.   :112047
```

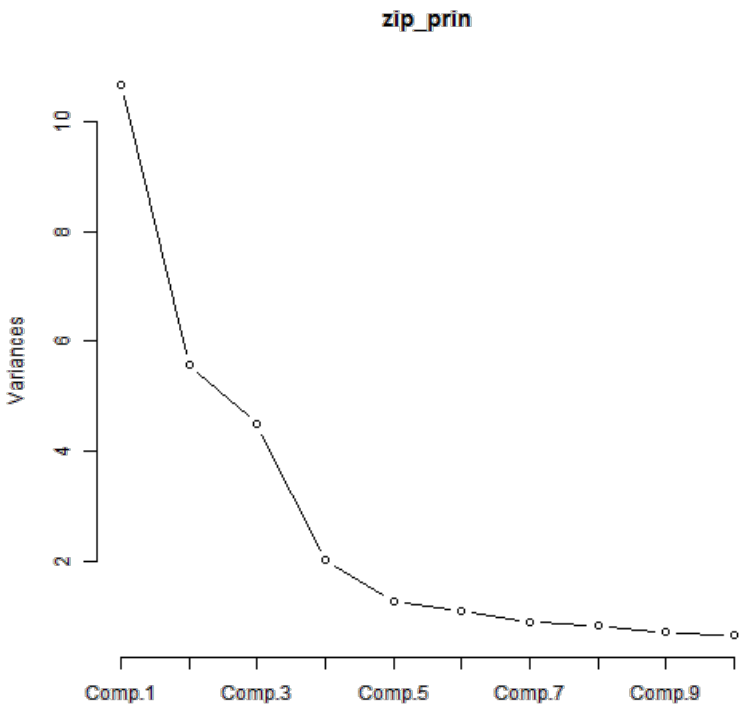
Principal Component Analysis

We conduct the principal component analysis using the function princomp. The correlation matrix is used for the estimation.

```
#### Principal Component Analysis
dd = zip1_comp[, -1]
zip_prin = princomp(dd, cor = T, scores = T)
```

Then we draw the scree plot to select the appropriate component number. There's a turning point at Comp.5 in the picture, so we simply choose 5 as the component number, which explains 74.98 percent of the variance.

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



we summary the results for details of the components. Besides, if we follow the Kaiser Principle, we'd better to choose 6 as the component number, which explains 78.35 percent variance.

```
summary(zip_prin)
```

```
## Importance of components:
##                           Comp.1 Comp.2 Comp.3  Comp.4  Comp.5  Comp.6
## Standard deviation     3.2662 2.3583 2.1195 1.41499 1.12699 1.03750
## Proportion of Variance 0.3334 0.1738 0.1404 0.06257 0.03969 0.03364
## Cumulative Proportion  0.3334 0.5072 0.6476 0.71013 0.74983 0.78346
##                            Comp.7  Comp.8  Comp.9 Comp.10 Comp.11
Comp.12
## Standard deviation     0.93575 0.90199 0.83016 0.79778 0.75660
0.74258
## Proportion of Variance 0.02736 0.02542 0.02154 0.01989 0.01789
0.01723
## Cumulative Proportion  0.81083 0.83625 0.85779 0.87768 0.89556
0.91280
##                           Comp.13 Comp.14 Comp.15 Comp.16  Comp.17
Comp.18
## Standard deviation     0.69760 0.65749 0.63389 0.55628 0.434257
0.403444
## Proportion of Variance 0.01521 0.01351 0.01256 0.00967 0.005893
0.005086
## Cumulative Proportion  0.92800 0.94151 0.95407 0.96374 0.969634
0.974721
##                            Comp.19  Comp.20  Comp.21  Comp.22  Comp.23
## Standard deviation     0.375547 0.358284 0.315746 0.301360 0.263100
## Proportion of Variance 0.004407 0.004011 0.003115 0.002838 0.002163
## Cumulative Proportion  0.979128 0.983140 0.986255 0.989093 0.991256
##                            Comp.24  Comp.25  Comp.26   Comp.27
Comp.28
## Standard deviation     0.256898 0.241009 0.201135 0.1787350
0.163270
## Proportion of Variance 0.002062 0.001815 0.001264 0.0009983
0.000833
## Cumulative Proportion  0.993319 0.995134 0.996398 0.9973964
0.998229
##                              Comp.29   Comp.30   Comp.31   Comp.32
## Standard deviation     0.1534187 0.1320470 0.1186227 4.018e-02
## Proportion of Variance 0.0007355 0.0005449 0.0004397 5.044e-05
## Cumulative Proportion  0.9989649 0.9995098 0.9999496 1.000e+00
```

In addition, we look into the loading matrix to achieve more information.

The first component is about the population age distribution against the wealth indicators.

The second mainly describes the races distribution and wealth ratings against some other indicators like ages and NCDB.

The third contains the comparison about the races structure and wealth indexes, ages, salary structures etc.

Other components can also be analyzed in this way.

To summary, the variables can be clustered into aspects about wealth, salary structures, social positions, human races structures, age structures etc.

```
loadings(zip_prin)
```

```
##
## Loadings:
##          Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
Comp.9
## DMAWLTHT -0.211 -0.125         0.141  0.297                -0.202
-0.262
```

```
## INCMINDX -0.260        -0.116         0.165                -0.129
-0.243
## WEALTHRT -0.248 -0.112                0.290                -0.123
-0.190
## PRCWHTE         -0.122 -0.253  0.422 -0.116 -0.308 -0.276  0.128
## PRCBLCK                 0.155 -0.336  0.315  0.612 -0.165
## PRCHISP                 0.172 -0.329        -0.442  0.601 -0.234
-0.117
## PRCUN18         -0.292  0.288 -0.107
## PRCOWNO         -0.312 -0.231         0.111                 0.110
## PRCTHRE         -0.316  0.245 -0.166        -0.116 -0.108
## PERPERHH        -0.112  0.204 -0.156        -0.139 -0.491 -0.622
## PRCNCD1   0.175 -0.271 -0.122        -0.128  0.181  0.147 -0.120
## MEDSCHYR -0.247        -0.110  0.105 -0.100  0.142        -0.211
0.150
## PRC25BA  -0.262                                            -0.198
0.219
## PRCNCD3  -0.167  0.276  0.124         0.148 -0.242 -0.222  0.141
## PRCNC10  -0.146  0.265                0.166 -0.242 -0.218  0.166
-0.164
## OOMEDHVL -0.258               -0.233 -0.256                0.120
## OOHVI    -0.262                      0.121
## PRCOOHV  -0.250                      0.197
## ISPSA    -0.264                                           -0.167
0.288
## PRCRENT          0.286  0.219                      -0.160
## PRC3544  -0.153 -0.238  0.173
0.194
## PRC4554         -0.263        -0.165                       0.320
-0.186
## PRC5564                 -0.255 -0.311  0.195 -0.132 -0.137  0.146
## PRC65P    0.160  0.156 -0.300 -0.120                       -0.179
0.143
## PRC55P    0.161  0.113 -0.331 -0.197  0.140 -0.107        -0.105
## HHMEDAGE  0.145        -0.343 -0.215  0.139                -0.116
## CEMI     -0.271        -0.113
-0.108
## PRC500K  -0.151        -0.127 -0.247 -0.381        -0.144
-0.490
## PRC200K  -0.223        -0.124 -0.275 -0.321
## PRC100K  -0.244               -0.176 -0.172
0.353
## PRCHHFM         -0.362        -0.151  0.127 -0.210 -0.152
0.140
## POPULAT  -0.140  0.141  0.123 -0.100  0.290 -0.109  0.140  0.194
0.299
##          Comp.10 Comp.11 Comp.12 Comp.13 Comp.14 Comp.15 Comp.16
Comp.17
## DMAWLTHT          0.104           -0.157   0.306            0.171
0.600
## INCMINDX  0.109                                            -0.104
-0.461
## WEALTHRT                                    0.196
## PRCWHTE                          -0.163
## PRCBLCK                   0.121
## PRCHISP  -0.134           0.147
## PRCUN18   0.227   0.180   0.176                            -0.172
## PRCOWNO                           0.432  -0.118            0.116
0.105
## PRCTHRE   0.193   0.117                                    -0.234
## PERPERHH -0.130  -0.278  -0.344   0.165                     0.150
## PRCNCD1                          -0.188          -0.138   0.156
## MEDSCHYR -0.159   0.344                  -0.276           -0.220
## PRC25BA  -0.194   0.248   0.112          -0.171
## PRCNCD3                          0.176
## PRCNC10           0.148   0.136   0.320
```

```
## OOMEDHVL                                          0.117
0.126
## OOHVI            -0.367   0.240          -0.296  -0.152
## PRCOOHV          -0.468   0.223          -0.332  -0.169
## ISPSA    -0.148   0.203                  -0.112
0.218
## PRCRENT                    0.249  -0.551   0.152  -0.212
## PRC3544           0.147   0.229                          0.756
-0.305
## PRC4554  -0.509           -0.120           0.170  -0.629  -0.118
## PRC5564  -0.414                   -0.397  -0.251   0.494
## PRC65P    0.238            0.144   0.104   0.163  -0.234
## PRC55P
## HHMEDAGE                                          0.128  -0.170
-0.145
## CEMI                       -0.150           0.219   0.134  -0.143
-0.384
## PRC500K   0.319   0.245                  -0.298           0.211
## PRC200K                                   0.195
0.130
## PRC100K          -0.249                   0.342   0.160
## PRCHHFM   0.159   0.143   0.149  -0.102                   -0.185
## POPULAT   0.285   0.143  -0.659  -0.179  -0.193  -0.201   0.173
##          Comp.18 Comp.19 Comp.20 Comp.21 Comp.22 Comp.23 Comp.24
Comp.25
## DMAWLTHT         -0.189           -0.288                   0.198
## INCMINDX  0.131                                    0.130   0.176
0.168
## WEALTHRT                            0.301          -0.157  -0.668
0.132
## PRCWHTE                    0.521                   -0.228   0.151
## PRCBLCK           0.106    0.440   0.131           -0.120   0.121
## PRCHISP                    0.343                   -0.114
## PRCUN18                   -0.206          -0.114  -0.224
0.156
## PRCOWNO                    0.168          -0.110  -0.182
0.136
## PRCTHRE                             -0.158  -0.236  -0.350
-0.136
## PERPERHH
## PRCNCD1   0.586   0.214
## MEDSCHYR         -0.263   0.131  -0.215   0.533  -0.275  -0.199
## PRC25BA                    0.151  -0.202  -0.655   0.327  -0.154
-0.162
## PRCNCD3  -0.187                                           -0.106
0.103
## PRCNC10   0.626   0.190  -0.106
-0.114
## OOMEDHVL
## OOHVI                               -0.168                   0.292
0.540
## PRCOOHV                   -0.138                   -0.119
-0.577
## ISPSA    -0.155   0.358  -0.355   0.528           -0.103   0.278
0.107
## PRCRENT                    0.101
## PRC3544          -0.183  -0.117
## PRC4554
## PRC5564
## PRC65P
## PRC55P
## HHMEDAGE                  -0.241
## CEMI              0.135                                   0.343
-0.390
## PRC500K  -0.186   0.260                                   -0.128
## PRC200K   0.277  -0.556            0.411
```

```
## PRC100K           0.405          -0.380                   -0.170
0.128
## PRCHHFM           0.136   0.182   0.128   0.359   0.624  -0.129
## POPULAT   0.109
##          Comp.26 Comp.27 Comp.28 Comp.29 Comp.30 Comp.31 Comp.32
## DMAWLTHT
## INCMINDX          0.221           0.520           0.352
## WEALTHRT         -0.186          -0.212
## PRCWHTE  -0.308                           0.104
## PRCBLCK  -0.183
## PRCHISP  -0.121
## PRCUN18  -0.339   0.429  -0.375  -0.195   0.154
## PRCOWNO   0.282   0.253                  -0.556
## PRCTHRE   0.164  -0.459   0.365   0.216
## PERPERHH
## PRCNCD1   0.150   0.250   0.391           0.220
## MEDSCHYR
## PRC25BA
## PRCNCD3   0.189   0.452   0.517           0.282  -0.101
## PRCNC10          -0.189  -0.158
## OOMEDHVL  0.162                  -0.294   0.230   0.731
## OOHVI            -0.233          -0.238          -0.189
## PRCOOHV           0.201           0.146
## ISPSA
## PRCRENT   0.262   0.131                  -0.500
## PRC3544
## PRC4554
## PRC5564                                                  0.228
## PRC65P    0.244          -0.220   0.184   0.220           0.621
## PRC55P    0.200          -0.195   0.154   0.193          -0.750
## HHMEDAGE -0.489           0.389  -0.322  -0.275   0.136
## CEMI      0.247                  -0.397          -0.316
## PRC500K                                          -0.159
## PRC200K                          0.135           -0.261
## PRC100K  -0.195   0.108           0.232  -0.100  -0.210
## PRCHHFM
## POPULAT
##
##                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  0.031  0.031  0.031  0.031  0.031  0.031  0.031
0.031
## Cumulative Var  0.031  0.063  0.094  0.125  0.156  0.188  0.219
0.250
##                Comp.9 Comp.10 Comp.11 Comp.12 Comp.13 Comp.14
Comp.15
## SS loadings     1.000   1.000   1.000   1.000   1.000   1.000
1.000
## Proportion Var  0.031   0.031   0.031   0.031   0.031   0.031
0.031
## Cumulative Var  0.281   0.312   0.344   0.375   0.406   0.438
0.469
##                Comp.16 Comp.17 Comp.18 Comp.19 Comp.20 Comp.21
Comp.22
## SS loadings      1.000   1.000   1.000   1.000   1.000   1.000
1.000
## Proportion Var   0.031   0.031   0.031   0.031   0.031   0.031
0.031
## Cumulative Var   0.500   0.531   0.562   0.594   0.625   0.656
0.688
##                Comp.23 Comp.24 Comp.25 Comp.26 Comp.27 Comp.28
Comp.29
## SS loadings      1.000   1.000   1.000   1.000   1.000   1.000
1.000
```

```
## Proportion Var   0.031   0.031   0.031   0.031   0.031   0.031
0.031
## Cumulative Var   0.719   0.750   0.781   0.812   0.844   0.875
0.906
##                Comp.30 Comp.31 Comp.32
## SS loadings      1.000   1.000   1.000
## Proportion Var   0.031   0.031   0.031
## Cumulative Var   0.938   0.969   1.000
```

Factor Analysis

We conduct the Factor Analysis with the MLE method. To avoid the scaling problem, we center and scale the variables with zero mean and 1 sd.

Besides, we use the varimax principle to conduct the rotation.

```
#### Factor Analysis
dd_std = scale(dd)
zip_fac = factanal(dd_std, factor = 10, rotation = "varimax", n.obs =
nrow(dd),
    control = list(trace = T))
```

```
## start 1 value: 4.202 uniqs: 0.2922 0.0280 0.1142 0.0050 0.4845
0.6891 0.0350 0.0050 0.0433 0.7544 0.0876 0.1402 0.0659 0.0050 0.1813
0.0050 0.0621 0.0609 0.1355 0.0837 0.3189 0.5247 0.0050 0.0050 0.0050
0.0577 0.0472 0.4756 0.1099 0.1863 0.1081 0.6299
```

```
zip_fac
```

```
##
## Call:
## factanal(x = dd_std, factors = 10, n.obs = nrow(dd), rotation =
"varimax",     control = list(trace = T))
##
## Uniquenesses:
## DMAWLTHT INCMINDX WEALTHRT  PRCWHTE  PRCBLCK  PRCHISP  PRCUN18
PRCOWNO
##    0.292    0.028    0.114    0.005    0.484    0.689    0.035
0.005
##  PRCTHRE PERPERHH  PRCNCD1 MEDSCHYR  PRC25BA  PRCNCD3  PRCNC10
OOMEDHVL
##    0.043    0.754    0.088    0.140    0.066    0.005    0.181
0.005
##    OOHVI  PRCOOHV    ISPSA  PRCRENT  PRC3544  PRC4554  PRC5564
PRC65P
##    0.062    0.061    0.135    0.084    0.319    0.525    0.005
0.005
##   PRC55P HHMEDAGE     CEMI  PRC500K  PRC200K  PRC100K  PRCHHFM
POPULAT
##    0.005    0.058    0.047    0.476    0.110    0.186    0.108
0.630
##
## Loadings:
##          Factor1 Factor2 Factor3 Factor4 Factor5 Factor6 Factor7
Factor8
## DMAWLTHT  0.798   0.133          -0.125           0.146
## INCMINDX  0.918   0.320          -0.111
## WEALTHRT  0.885   0.212   0.105  -0.158
## PRCWHTE   0.130          -0.117   0.117           0.967
```

```
## PRCBLCK                   0.108                 -0.699
## PRCHISP                   0.160          0.167  -0.485
## PRCUN18          -0.137  -0.170  -0.310  0.889  -0.135
## PRCOWNO   0.116          -0.533   0.226  0.257   0.280   0.690
0.144
## PRCTHRE                   -0.164  -0.289  0.903  -0.114   0.103
## PERPERHH                          -0.167  0.426  -0.150  -0.106
## PRCNCD1  -0.181  -0.181  -0.870   0.131  0.184   0.128   0.120
## MEDSCHYR  0.599   0.384   0.168  -0.151 -0.156   0.107
## PRC25BA   0.601   0.431   0.217  -0.113 -0.113
## PRCNCD3   0.142   0.152   0.950         -0.132          -0.101
## PRCNC10   0.145           0.858         -0.189
## OOMEDHVL  0.388   0.883   0.184
## OOHVI     0.703   0.419   0.192  -0.110
## PRCOOHV   0.692   0.311   0.197  -0.153
## ISPSA     0.630   0.393   0.217  -0.145          0.105
## PRCRENT  -0.121           0.609                 -0.202  -0.675
## PRC3544   0.295   0.161          -0.507  0.494          0.112
## PRC4554   0.283   0.216  -0.210  -0.218  0.356          0.267
0.235
## PRC5564                   -0.214   0.427                 0.124
0.864
## PRC65P   -0.235           -0.103   0.905  -0.312
## PRC55P   -0.210           -0.154   0.894  -0.265
0.220
## HHMEDAGE -0.141           -0.203   0.880  -0.243
0.188
## CEMI      0.779   0.534   0.103  -0.144
## PRC500K   0.217   0.674
## PRC200K   0.306   0.877   0.123
## PRC100K   0.368   0.766   0.169  -0.137
## PRCHHFM                   -0.353          0.811   0.126   0.232
0.152
## POPULAT   0.201   0.168   0.471  -0.115          -0.235
##          Factor9 Factor10
## DMAWLTHT
## INCMINDX
## WEALTHRT
## PRCWHTE
## PRCBLCK
## PRCHISP
## PRCUN18
## PRCOWNO
## PRCTHRE
## PERPERHH
## PRCNCD1
## MEDSCHYR  0.509
## PRC25BA   0.550
## PRCNCD3
## PRCNC10
## OOMEDHVL          0.119
## OOHVI             0.462
## PRCOOHV           0.546
## ISPSA     0.470
## PRCRENT
## PRC3544   0.183
## PRC4554
## PRC5564
## PRC65P
## PRC55P
## HHMEDAGE
## CEMI             -0.120
## PRC500K
## PRC200K
## PRC100K   0.146   0.148
## PRCHHFM
```

```
## POPULAT
##
##                Factor1 Factor2 Factor3 Factor4 Factor5 Factor6
Factor7
## SS loadings      5.856   4.023   3.983   3.402   3.291   2.017
1.189
## Proportion Var   0.183   0.126   0.124   0.106   0.103   0.063
0.037
## Cumulative Var   0.183   0.309   0.433   0.539   0.642   0.705
0.743
##                Factor8 Factor9 Factor10
## SS loadings      0.982   0.895    0.614
## Proportion Var   0.031   0.028    0.019
## Cumulative Var   0.773   0.801    0.820
##
## Test of the hypothesis that 10 factors are sufficient.
## The chi square statistic is 143661 on 221 degrees of freedom.
## The p-value is 0
```

Several ways to decide Number of Factors

We use the principle that every factor's variance's ratio should be larger than 1/p. Since the variance of the X (all variance) is p (32), we just need to guarantee that the factor's variance is larger than 1.

```
flag = 1
i = 15
while (flag == 1) {
    zip_fac = factanal(dd_std, factor = i, rotation = "varimax", n.obs
= nrow(dd))
    fac_var = apply(zip_fac$loadings, 2, function(x) return(sum(x^2)))
    cat("nFactor:", i, "\n", "Factor variance:", fac_var, "\n", "\n")
    if (all(fac_var >= 1))
        flag = 0
    i = i - 1
}
```

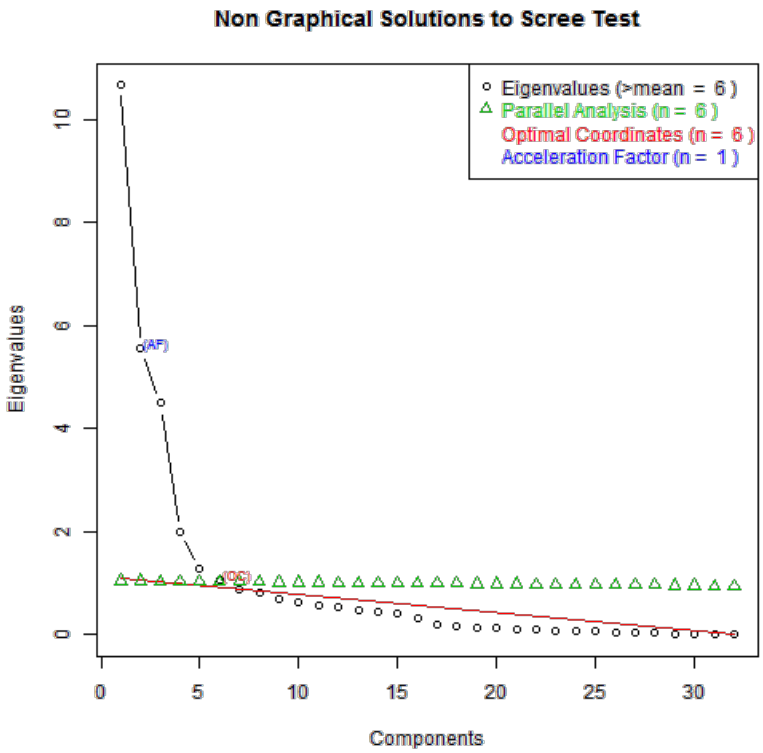
```
## nFactor: 15
##  Factor variance: 7.249 4.138 3.301 3.285 1.95 1.574 1.333 1.176
1.071 0.9774 0.8678 0.5817 0.3446 0.1688 0.1171
##
## nFactor: 14
##  Factor variance: 7.207 4.114 3.369 3.259 1.729 1.598 1.572 1.231
1.073 0.9975 0.8721 0.5801 0.2482 0.1219
##
## nFactor: 13
##  Factor variance: 6.321 4.117 3.366 3.285 3.245 1.552 1.229 1.085
0.9865 0.9407 0.8888 0.5982 0.1863
##
## nFactor: 12
##  Factor variance: 7.454 4.131 3.317 3.313 1.939 1.531 1.265 1.201
1.103 0.9898 0.7484 0.5442
##
## nFactor: 11
##  Factor variance: 7.396 3.988 3.336 3.306 2.012 2.006 1.242 1.195
0.9898 0.7557 0.5494
##
## nFactor: 10
##  Factor variance: 5.856 4.023 3.983 3.402 3.291 2.017 1.189 0.9824
0.8955 0.6141
##
## nFactor: 9
##  Factor variance: 6.563 3.884 3.665 3.411 3.33 2.066 1.107 1.004
0.5823
##
## nFactor: 8
##  Factor variance: 7.224 3.834 3.437 3.27 3.181 2.064 1.101 1.007
##
```

Here 8 is the number of the factors that subject to the constraint.

Besides that, we can also use the eigenvalues to decide the number. According to the principal component analysis before, we can choose 6 as our number of factors, which accounts for 78.34 percent of the sum of eigenvalues. We can also make other choices due to the demanded cumulative proportion in the summary(zip_prin) before.

The parallel analysis is also a way to decide the number of factors, which is developed by Raiche, Riopel, and Blais. We the R package 'nFactors' fot the analysis.

```
# Determine Number of Factors to Extract
library(nFactors)
ev <- eigen(cor(dd_std))  # get eigenvalues
ap <- parallel(subject = nrow(dd_std), var = ncol(dd_std), rep = 100,
cent = 0.05)
nS <- nScree(x = ev$values, aparallel = ap$eigen$qevpea)
plotnScree(nS)
```



The result gives 6 as the best number.

There are other methods provided in this package for the number decision process.

```
nBartlett(x = ev$values, N = nrow(dd_std), alpha = 0.05, details =
TRUE)
```

```
## bartlett anderson   lawley
##       31       31       31
```

```
nBentler(x = ev$values, N = nrow(dd_std), alpha = 0.05, details =
TRUE)
```

```
## [1] 29
```

```
nCng(x = ev$values, model = "factors", details = TRUE)
```

```
## [1] 3
```

Factor Explainations

We choose 6 as the factor numbers.

We then use the varimax and promax rotation method to achieve the loading matrix.

```
#### Factor Analysis
dd_std = scale(dd)
zip_fac = factanal(dd_std, factor = 6, rotation = "varimax", n.obs =
nrow(dd))
loadings(zip_fac)
```

```
##
## Loadings:
##          Factor1 Factor2 Factor3 Factor4 Factor5 Factor6
## DMAWLTHT  0.854
## INCMINDX  0.934                    0.202
## WEALTHRT  0.931                           -0.126
## PRCWHTE   0.252  -0.316  -0.150           0.140
## PRCBLCK  -0.163   0.233   0.113
## PRCHISP  -0.125   0.260   0.223   0.115
## PRCUN18          -0.131   0.900  -0.128  -0.291
## PRCOWNO   0.122  -0.681   0.277           0.259   0.176
## PRCTHRE          -0.144   0.942          -0.252
## PERPERHH                  0.437          -0.161
## PRCNCD1  -0.300  -0.877   0.169  -0.115
## MEDSCHYR  0.689   0.116  -0.197   0.325  -0.153
## PRC25BA   0.700   0.182  -0.149   0.383  -0.114
## PRCNCD3   0.281   0.938  -0.134
## PRCNC10   0.265   0.840  -0.187
## OOMEDHVL  0.526   0.174           0.826
## OOHVI     0.763   0.132           0.382
## PRCOOHV   0.751   0.136           0.281  -0.125
## ISPSA     0.750   0.152           0.326  -0.132
## PRCRENT           0.730  -0.128          -0.114  -0.125
## PRC3544   0.368           0.490   0.125  -0.477
## PRC4554   0.313  -0.268   0.387   0.178  -0.192   0.246
## PRC5564          -0.255                   0.405   0.870
## PRC65P   -0.271  -0.135  -0.351           0.881
## PRC55P   -0.251  -0.193  -0.301           0.868   0.228
## HHMEDAGE -0.187  -0.249  -0.272           0.855   0.198
## CEMI      0.848                    0.411  -0.117
## PRC500K   0.275                    0.636
## PRC200K   0.422   0.127           0.830
## PRC100K   0.507   0.162           0.715  -0.108
## PRCHHFM          -0.414   0.799                   0.154
## POPULAT   0.254   0.500           0.130
##
##                Factor1 Factor2 Factor3 Factor4 Factor5 Factor6
## SS loadings      7.505   4.487   3.561   3.235   3.128   1.025
## Proportion Var   0.235   0.140   0.111   0.101   0.098   0.032
## Cumulative Var   0.235   0.375   0.486   0.587   0.685   0.717
```

```
zip_fac = factanal(dd_std, factor = 6, rotation = "promax", n.obs =
nrow(dd))
loadings(zip_fac)
```

```
##
## Loadings:
##          Factor1 Factor2 Factor3 Factor4 Factor5 Factor6
## DMAWLTHT  0.991           -0.256
## INCMINDX  0.978
## WEALTHRT  1.026           -0.171
## PRCWHTE   0.409  -0.272  -0.175   0.144  -0.157
## PRCBLCK  -0.193   0.299                   0.175
## PRCHISP  -0.232   0.285   0.183           0.292
## PRCUN18                                   0.934  -0.151
## PRCOWNO   0.300  -0.522           0.291   0.273
## PRCTHRE                                   0.972
## PERPERHH          0.114                   0.458
## PRCNCD1  -0.207  -0.930
## MEDSCHYR  0.594            0.197  -0.110  -0.243
## PRC25BA   0.582            0.268          -0.160
## PRCNCD3   0.210   1.045            0.101
## PRCNC10   0.224   0.950  -0.141
## OOMEDHVL  0.153            0.922
## OOHVI     0.665   0.130   0.254
## PRCOOHV   0.694   0.154   0.122
## ISPSA     0.671   0.103   0.183
## PRCRENT  -0.232   0.684
## PRC3544   0.260  -0.119           -0.399   0.377
## PRC4554   0.222  -0.223   0.148  -0.257   0.236   0.281
## PRC5564                                  -0.167   1.015
## PRC65P                             1.089          -0.215
## PRC55P                             0.928           0.132
## HHMEDAGE                           0.940
## CEMI      0.755            0.281
## PRC500K                    0.751   0.117
## PRC200K                    0.963
## PRC100K   0.178            0.782
## PRCHHFM   0.167  -0.171            0.225   0.856
## POPULAT   0.173   0.539                    0.105
##
##                Factor1 Factor2 Factor3 Factor4 Factor5 Factor6
## SS loadings      6.422   4.339   3.500   3.392   3.330   1.261
## Proportion Var   0.201   0.136   0.109   0.106   0.104   0.039
## Cumulative Var   0.201   0.336   0.446   0.552   0.656   0.695
```

From the loading matrix we can see the promax rotation are closely related to the varimax rotation in the top3 factors. We now look into more details of the loading matrix of the promax rotation.

The first factor has a high loading value at WEALTHRT, DMAWLTHT, INCMINDX and CEMI, they are mainly about the income index. And it has relative negative loading on PRCBLCK, PRCNCD1 and PRCRENT compared to the first aspect, which are mainly related to the human races percent.

The second factor are mostly the PRCNCD3, PRCNCD10 aginst PRCNCD1 and PRCOWNO. They are mainly the descriptions of the %NCDB HH.

The third has a high loading on PRC500K, PRC200K, PRC100K and OOMEDHVL. They are mainly about the OOH Value.

The fourth factor has the highest loading on PRC65P, HHMEDAGE, PRC55P and a relative negative value on PRC3544, PRC4554. They are mostly descriptions of the age structure.

The fifth factor is highest loaded on PRCUN18, PRCTHRE and PRCHHFM. They have a closed relationship with the %HH value.

Multidimensional Scaling

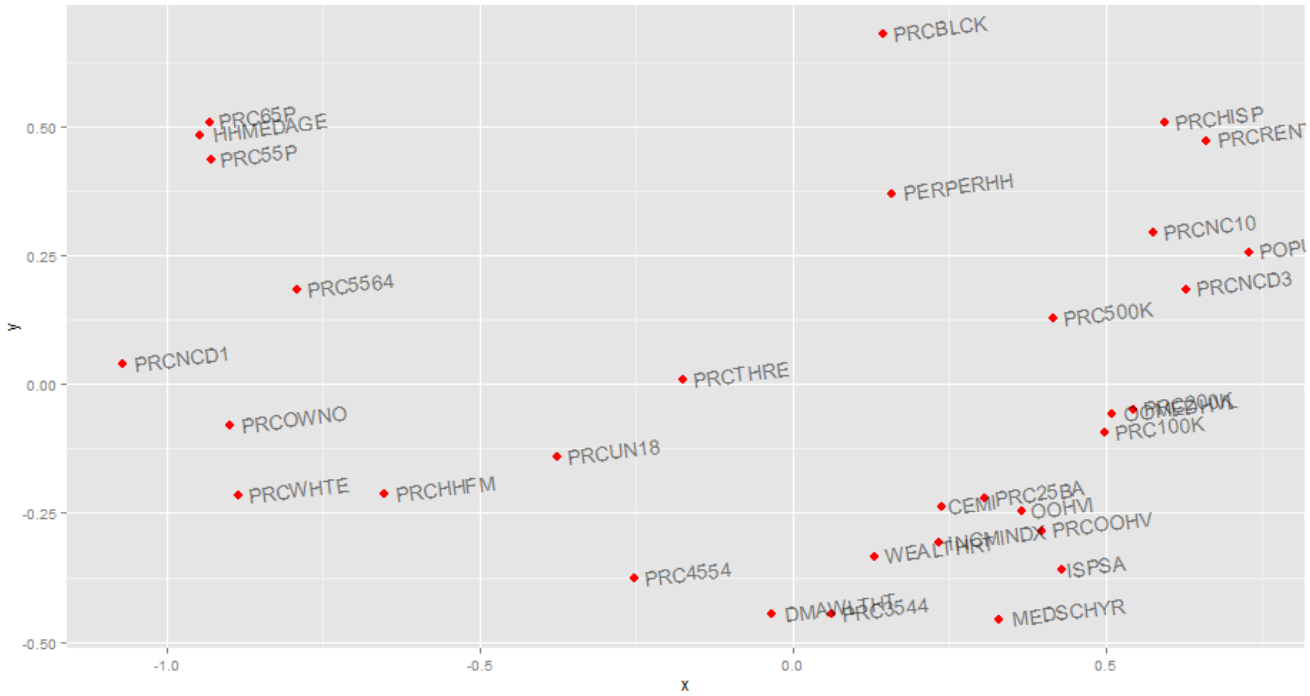
We use the package MASS to do the multidimensional scaling with data zip2. For details, we use 1-correlation as the distance measure.

```
library("MASS")
library(ggplot2)
load("zip2.rda")
dist = 1 - cor(zip2[, -1])

zip2_mds = isoMDS(dist)
```

```
## initial  value 19.343707
## iter   5 value 14.884488
## final  value 14.760496
## converged
```

```
x = zip2_mds$points[, 1]
y = zip2_mds$points[, 2]
g = ggplot(data.frame(x, y), aes(x, y, label = colnames(zip2)[-1]))
g + geom_point(shape = 16, size = 3, colour = "red") + geom_text(hjust
= -0.1,
    vjust = 0.5, alpha = 0.5, angle = 7)
```



From the figure, we can see some features are clustered with each other.

The CEMI, PRC25BA, OOHVI, PRCOOHV, ISPSA, WEALTHRT, DMAWLTHT, MEDSCHYR are very closed to each other. That means the wealth ,the education level and the social position are high correlated.

The PRC65P, HHMEDAGE, PRC55P are very closed. They are all descriptions of ages. We may infer that the householder's median age are upon 55.

The PRC200K, PRC100K, OOHVI are very closed to each other.

There isn't a significant clustering sign of other features.