Statistical learning

Principal components analysis (PCA)

V. Lefieux

1 Libraries

```
library(FactoMineR) # For PCA
library(ggplot2) # Graphical library
```

2 Data

```
setwd("/Users/vincentlefieux/Documents/Data/")
cars=read.table("data_cars.txt",row.names=1,header=TRUE,sep=";",dec=",")
```

To define price classes (breakpoints: quartiles):

```
cars$PRI_CL=cut(cars$PRI,breaks=quantile(cars$PRI),include.lowest=TRUE)
levels(cars$PRI_CL)=c("P1","P2","P3","P4")
```

P1, P2, P3 et P4 are ranked in ascending order.

3 Univariate exploratory analysis

```
stars(cars[,1:6],key.loc=c(10,1.8),main="Radar diagrams",cex=0.5,flip.labels=TRUE)
```

Radar diagrams

```r
summary(cars)
```

```
##       CYL            POW              LEN             WID
##  Min.   :1166   Min.   : 55.00   Min.   :393.0   Min.   :157.0
##  1st Qu.:1310   1st Qu.: 70.75   1st Qu.:424.0   1st Qu.:162.2
##  Median :1578   Median : 82.00   Median :434.5   Median :167.0
##  Mean   :1632   Mean   : 84.61   Mean   :433.5   Mean   :166.7
##  3rd Qu.:1798   3rd Qu.: 98.00   3rd Qu.:448.0   3rd Qu.:169.8
##  Max.   :2664   Max.   :128.00   Max.   :469.0   Max.   :177.0
##       WGT            SPD        FIN         PRI         PRI_CL
##  Min.   : 815   Min.   :140.0   B :7   Min.   :22100   P1:5
##  1st Qu.:1020   1st Qu.:151.2   M :5   1st Qu.:29842   P2:4
##  Median :1088   Median :160.0   TB:6   Median :33345   P3:4
##  Mean   :1079   Mean   :158.3          Mean   :34159   P4:5
##  3rd Qu.:1127   3rd Qu.:165.0          3rd Qu.:38458
##  Max.   :1370   Max.   :180.0          Max.   :47700
```

```r
boxplot(cars[,1:6],main="Boxplots")
```

```r
apply(cars[,1:6],2,"sd")
```

```
##        CYL        POW        LEN        WID        WGT        SPD
## 373.929846  20.376281  22.107358   5.313689 136.957808  12.140383
```

4 Bivariate exploratory analysis

```r
pairs(cars[,1:6],main="Scatterplot")
```

The correlation matrix is:

```
cor(cars[,1:6])
```

```
##           CYL       POW       LEN       WID       WGT       SPD
## CYL 1.0000000 0.7966277 0.7014619 0.6297572 0.7889520 0.6649340
## POW 0.7966277 1.0000000 0.6413624 0.5208320 0.7652930 0.8443795
## LEN 0.7014619 0.6413624 1.0000000 0.8492664 0.8680903 0.4759285
## WID 0.6297572 0.5208320 0.8492664 1.0000000 0.7168739 0.4729453
## WGT 0.7889520 0.7652930 0.8680903 0.7168739 1.0000000 0.4775956
## SPD 0.6649340 0.8443795 0.4759285 0.4729453 0.4775956 1.0000000
```

5 Principal components analysis

```
out_pca=PCA(cars[,c(1:6,9)],scale.unit=TRUE,ncp=6,quali.sup=7,graph=FALSE)
summary(out_pca)
```

```
##
## Call:
## PCA(X = cars[, c(1:6, 9)], scale.unit = TRUE, ncp = 6, quali.sup = 7,
##      graph = FALSE)
##
##
## Eigenvalues
##                         Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6
## Variance                4.421   0.856   0.373   0.214   0.093   0.043
## % of var.              73.681  14.268   6.218   3.565   1.547   0.722
## Cumulative % of var.   73.681  87.949  94.166  97.732  99.278 100.000
##
```

```
## Individuals (the 10 first)
##                      Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2
## ALFASUD-TI-1350  |  2.868 | -2.139  5.749  0.556 |  1.786 20.693  0.388 |
## AUDI-100-L       |  2.583 |  1.561  3.064  0.365 | -1.527 15.133  0.349 |
## SIMCA-1300-GLS   |  1.469 | -1.119  1.575  0.580 | -0.675  2.953  0.211 |
## CITROEN-GS-CLUB  |  2.604 | -2.574  8.324  0.977 |  0.113  0.083  0.002 |
## FIAT-132-1600GLS |  1.081 |  0.428  0.230  0.157 |  0.696  3.140  0.414 |
## LANCIA-BETA-1300 |  1.065 | -0.304  0.116  0.082 | -0.196  0.250  0.034 |
## PEUGEOT-504      |  1.230 |  0.684  0.588  0.309 | -0.933  5.650  0.575 |
## RENAULT-16-TL    |  2.374 | -1.948  4.771  0.674 | -0.980  6.238  0.171 |
## RENAULT-30-TS    |  4.668 |  4.410 24.437  0.892 |  1.064  7.342  0.052 |
## TOYOTA COROLLA   |  4.036 | -3.986 19.964  0.975 |  0.236  0.362  0.003 |
##                   Dim.3    ctr   cos2
## ALFASUD-TI-1350   0.572  4.870  0.040 |
## AUDI-100-L        1.315 25.762  0.259 |
## SIMCA-1300-GLS    0.457  3.104  0.097 |
## CITROEN-GS-CLUB   0.149  0.329  0.003 |
## FIAT-132-1600GLS -0.193  0.556  0.032 |
## LANCIA-BETA-1300  0.676  6.801  0.402 |
## PEUGEOT-504      -0.257  0.982  0.044 |
## RENAULT-16-TL    -0.620  5.716  0.068 |
## RENAULT-30-TS    -0.594  5.246  0.016 |
## TOYOTA COROLLA   -0.303  1.368  0.006 |
##
## Variables
##                     Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3
## CYL              |  0.893 18.057  0.798 |  0.115  1.542  0.013 | -0.216
## POW              |  0.887 17.791  0.787 |  0.385 17.287  0.148 | -0.113
## LEN              |  0.886 17.763  0.785 | -0.381 16.959  0.145 |  0.041
## WID              |  0.814 14.971  0.662 | -0.413 19.899  0.170 |  0.369
## WGT              |  0.905 18.534  0.819 | -0.225  5.889  0.050 | -0.296
## SPD              |  0.755 12.884  0.570 |  0.574 38.423  0.329 |  0.297
##                     ctr   cos2
## CYL              12.504  0.047 |
## POW               3.420  0.013 |
## LEN               0.457  0.002 |
## WID              36.587  0.136 |
## WGT              23.464  0.088 |
## SPD              23.568  0.088 |
##
## Supplementary categories
##                      Dist    Dim.1   cos2 v.test    Dim.2   cos2 v.test
## P1               |  2.004 | -2.000  0.996 -2.433 | -0.023  0.000 -0.062 |
## P2               |  0.462 | -0.278  0.363 -0.292 | -0.006  0.000 -0.014 |
## P3               |  0.676 |  0.614  0.826  0.644 |  0.036  0.003  0.086 |
## P4               |  1.769 |  1.732  0.958  2.106 | -0.002  0.000 -0.004 |
##                   Dim.3   cos2 v.test
## P1               -0.070  0.001 -0.291 |
## P2                0.123  0.071  0.444 |
## P3                0.245  0.132  0.886 |
## P4               -0.225  0.016 -0.943 |
```

List of generated objects:

```r
out_pca
```

```
## **Results for the Principal Component Analysis (PCA)**
## The analysis was performed on 18 individuals, described by 7 variables
## *The results are available in the following objects:
##
##    name
## 1  "$eig"
## 2  "$var"
## 3  "$var$coord"
## 4  "$var$cor"
## 5  "$var$cos2"
## 6  "$var$contrib"
## 7  "$ind"
## 8  "$ind$coord"
## 9  "$ind$cos2"
## 10 "$ind$contrib"
## 11 "$quali.sup"
## 12 "$quali.sup$coord"
## 13 "$quali.sup$v.test"
## 14 "$call"
## 15 "$call$centre"
## 16 "$call$ecart.type"
## 17 "$call$row.w"
## 18 "$call$col.w"
##    description
## 1  "eigenvalues"
## 2  "results for the variables"
## 3  "coord. for the variables"
## 4  "correlations variables - dimensions"
## 5  "cos2 for the variables"
## 6  "contributions of the variables"
## 7  "results for the individuals"
## 8  "coord. for the individuals"
## 9  "cos2 for the individuals"
## 10 "contributions of the individuals"
## 11 "results for the supplementary categorical variables"
## 12 "coord. for the supplementary categories"
## 13 "v-test of the supplementary categories"
## 14 "summary statistics"
## 15 "mean of the variables"
## 16 "standard error of the variables"
## 17 "weights for the individuals"
## 18 "weights for the variables"
```

To calculate eigenvalues and obtain the screeplot:

```r
eigenv=out_pca$eig[,"eigenvalue"]
eigenv_cum=cumsum(eigenv)/sum(eigenv)
cp=1:length(eigenv)
ev=data.frame(cp=cp,eigenv=eigenv)
ev_cum=data.frame(cp=cp,eigenv_cum=eigenv_cum)

plot(ev$cp,ev$eigenv,type="h",lwd=50,lend="butt",
     xlab="Principal components",ylab="Eigenvalues",
```

```
     main="Screeplot")
```

```
# With ggplot:
ggplot(data=ev,aes(x=cp,y=eigenv))+
  geom_bar(stat="identity",fill="steelblue")+
  theme_minimal()+
  ggtitle("Screeplot")+
  xlab("Principal components")+
  ylab("Eigenvalues")+
  scale_x_continuous(breaks=cp)
```

```r
plot(ev_cum$cp,ev_cum$eigenv_cum,type="h",lwd=50,ylim=c(0,1),lend="butt",
     xlab="Principal components",ylab="Total inertia resumed",
     main="Total inertia resumed / Principal components")
```

```r
# With ggplot:
ggplot(data=ev_cum,aes(x=cp,y=eigenv_cum))+
  geom_bar(stat="identity",fill="steelblue")+
  theme_minimal()+
  ggtitle("Total inertia resumed / Principal components")+
  xlab("Principal components")+
  ylab("Total inertia resumed")+
  scale_x_continuous(breaks=cp)
```

To plot the correlation circle:

```
plot.PCA(out_pca,shadow=TRUE,cex=0.8,axes=c(1,2),choix="var",new.plot=TRUE,
         title="Correlation circle")
```

To project the individuals onto principal components (with or without individuals names, depending or not on price classes):

```
plot.PCA(out_pca,shadow=TRUE,cex=2,axes=c(1,2),choix="ind",label="none",new.plot=TRUE,
         title="Projections of individuals")
```

```r
plot.PCA(out_pca,shadow=TRUE,cex=0.8,axes=c(1,2),choix="ind",label="ind",new.plot=TRUE,
         title="Projections of individuals: with individuals names")
```

```r
plot.PCA(out_pca,shadow=TRUE,cex=2,axes=c(1,2),choix="ind",habillage=7,label="none",new.plot=TRUE,
         title="Projections of individuals: depending on price classes")
```

To study the quality of the individuals representation and the individuals contributions:

```
plot.PCA(out_pca,shadow=TRUE,cex=0.8,select="contrib 10",
         title="The 10 individuals with the biggest contributions")
```

```r
plot.PCA(out_pca,shadow=TRUE,cex=0.8,select="cos2 10",
         title="The 10 individuals with the biggest qualities of representation")
```

```r
plot.PCA(out_pca,shadow=TRUE,cex=0.8,select="cos2 0.5",
         title="The individuals with a quality of representation over 0.5")
```

To obtain the qualities of the individuals representation and the individuals contributions:

```
co2=out_pca$ind$cos2
co2
```

```
##                       Dim.1       Dim.2        Dim.3        Dim.4
## ALFASUD-TI-1350  0.55621843 0.387670355 0.0397592085 4.957299e-03
## AUDI-100-L       0.36533449 0.349406285 0.2592151934 6.693304e-03
## SIMCA-1300-GLS   0.58028427 0.210694005 0.0965452233 1.301258e-02
## CITROEN-GS-CLUB  0.97699183 0.001879434 0.0032555327 4.436297e-05
## FIAT-132-1600GLS 0.15657878 0.413825852 0.0319549851 3.370691e-01
## LANCIA-BETA-1300 0.08155471 0.033899586 0.4024052108 2.723210e-01
## PEUGEOT-504      0.30920173 0.575488494 0.0436002142 2.727640e-02
## RENAULT-16-TL    0.67353923 0.170535407 0.0681009218 7.048333e-02
## RENAULT-30-TS    0.89243070 0.051919840 0.0161671191 3.294356e-02
## TOYOTA COROLLA   0.97521929 0.003425970 0.0056408243 4.314853e-03
## ALFETTA-1.66     0.04297836 0.820652103 0.0001395888 1.292292e-01
## PRINCESS-1800HL  0.53094720 0.362855147 0.0240305669 4.715339e-02
## DATSUN-200L      0.77838982 0.028137029 0.1392566275 5.365503e-02
## TAUNUS-2000-GL   0.70481879 0.096496323 0.0325622996 1.381938e-01
## RANCHO           0.24327262 0.410468952 0.2011310138 6.518147e-02
## MAZDA-9295       0.21733631 0.185337481 0.0083456693 1.546818e-02
## OPEL-REKORD-L    0.86189979 0.001789839 0.1042601945 9.182812e-03
## LADA-1300        0.92605245 0.002606517 0.0416366325 1.167793e-03
##                        Dim.5        Dim.6
## ALFASUD-TI-1350  0.0110412256 3.534820e-04
## AUDI-100-L       0.0033050424 1.604569e-02
## SIMCA-1300-GLS   0.0652509452 3.421298e-02
## CITROEN-GS-CLUB  0.0075901594 1.023868e-02
## FIAT-132-1600GLS 0.0593906426 1.180644e-03
```

```
## LANCIA-BETA-1300 0.1745464048 3.527309e-02
## PEUGEOT-504      0.0287782353 1.565493e-02
## RENAULT-16-TL    0.0152471967 2.093916e-03
## RENAULT-30-TS    0.0064499529 8.883564e-05
## TOYOTA COROLLA   0.0047588536 6.640208e-03
## ALFETTA-1.66     0.0063405794 6.601324e-04
## PRINCESS-1800HL  0.0175051651 1.750853e-02
## DATSUN-200L      0.0002664249 2.950709e-04
## TAUNUS-2000-GL   0.0018155492 2.611325e-02
## RANCHO           0.0724232173 7.522728e-03
## MAZDA-9295       0.4057935637 1.677188e-01
## OPEL-REKORD-L    0.0188317772 4.035588e-03
## LADA-1300        0.0183951365 1.014147e-02
```

```
contrib=out_pca$ind$contrib
contrib
```

```
##                       Dim.1       Dim.2       Dim.3        Dim.4
## ALFASUD-TI-1350   5.7492537 20.69330670  4.86995152  1.058917485
## AUDI-100-L        3.0639513 15.13293304 25.76157771  1.160065973
## SIMCA-1300-GLS    1.5746358  2.95251897  3.10449465  0.729715894
## CITROEN-GS-CLUB   8.3243598  0.08269669  0.32870233  0.007811447
## FIAT-132-1600GLS  0.2300459  3.13978898  0.55634100 10.234143652
## LANCIA-BETA-1300  0.1163180  0.24968553  6.80114726  8.026567168
## PEUGEOT-504       0.5878172  5.64986748  0.98222242  1.071613935
## RENAULT-16-TL     4.7710990  6.23837245  5.71648851 10.317937178
## RENAULT-30-TS    24.4368837  7.34185620  5.24595754 18.642010492
## TOYOTA COROLLA   19.9640247  0.36218506  1.36838781  1.825419683
## ALFETTA-1.66      0.2407083 23.73566876  0.00926430 14.957289357
## PRINCESS-1800HL   1.3027651  4.59779144  0.69871425  2.390994825
## DATSUN-200L      10.8701289  2.02916325 23.04486821 15.484561594
## TAUNUS-2000-GL    2.1726682  1.53612980  1.18946543  8.803500743
## RANCHO            0.6002290  5.23004324  5.88062674  3.323524037
## MAZDA-9295        0.1869563  0.82332696  0.08507263  0.274978001
## OPEL-REKORD-L     6.5887644  0.07065818  9.44466473  1.450687356
## LADA-1300         9.2193906  0.13400728  4.91205295  0.240261181
##                       Dim.5      Dim.6
## ALFASUD-TI-1350   5.4367117  0.3731210
## AUDI-100-L        1.3204475 13.7425014
## SIMCA-1300-GLS    8.4348826  9.4808311
## CITROEN-GS-CLUB   3.0808016  8.9088174
## FIAT-132-1600GLS  4.1567374  0.1771401
## LANCIA-BETA-1300 11.8593696  5.1375672
## PEUGEOT-504       2.6062585  3.0392600
## RENAULT-16-TL     5.1451550  1.5147153
## RENAULT-30-TS     8.4135793  0.2484133
## TOYOTA COROLLA    4.6408880 13.8817357
## ALFETTA-1.66      1.6916988  0.3775622
## PRINCESS-1800HL   2.0461332  4.3871325
## DATSUN-200L       0.1772414  0.4208043
## TAUNUS-2000-GL    0.2666103  8.2204045
## RANCHO            8.5124485  1.8954632
## MAZDA-9295       16.6289932 14.7334981
## OPEL-REKORD-L     6.8579045  3.1504331
## LADA-1300         8.7241389 10.3105995
```