UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
School of Professional & Executive Development

Multivariate Description of Data

Tomàs Aluja

Barcelona; January 18th, 2016

Outline

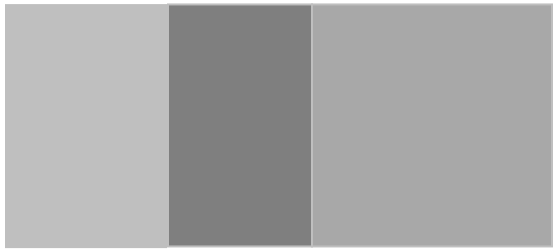
1. Multivariate description of a table
2. PCA: Principal Component Analysis.
3. Finding the Intangibles
4. Application 1: The cars visualisation
5. Application 2: The bank clients visualisation
6. Application 3: The zip decoding problem
7. Extensions of PCA: Correspondence Analysis
8. Extensions of PCA: Multiple Correspondence Analysis

MULTIVARIATE DESCRIPTON OF A TABLE

Why Multivariate Description?

Paradigm:

Data contain information about the generating phenomenon



Goal:
Discovering hidden evidence by multivariate description of data

Data is knowledge

Assumptions: large data files, containing different types of variables, without probabilistic assumptions.

Secondary goal:
Dimensionality reduction

Three steps of Multivariate Description

1.Visualization. The human eye ...

To consent a loss in information by a gain in interpretability.

2.Clustering: Synthesis of reality

Operational simplification of individual's diversity.

3.Profiling: Noticing the difference

What makes a group of individuals "different".

The mess of words

Information Accountable measure of the total variability of data (~ inertia, entropy)

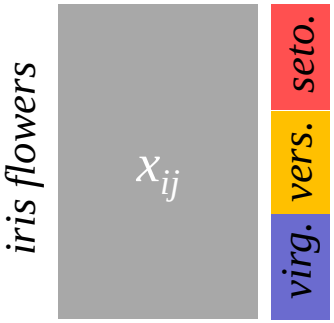
Meaning Subjective interpretation of results

Significance Result not due to chance alone

PRINCIPAL COMPONENT ANALYSIS

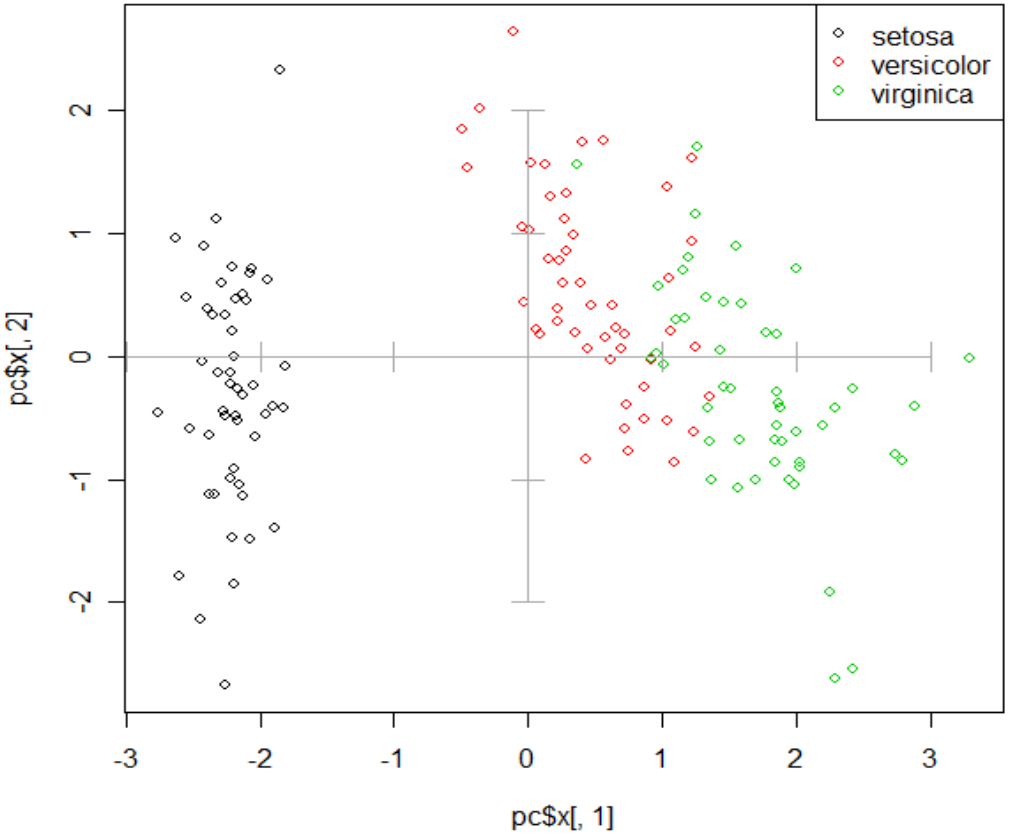
The Principal Component Analysis

sepal & petal measures



Tool to visualize multivariate data

Iris flowers according sepal and petal length and width



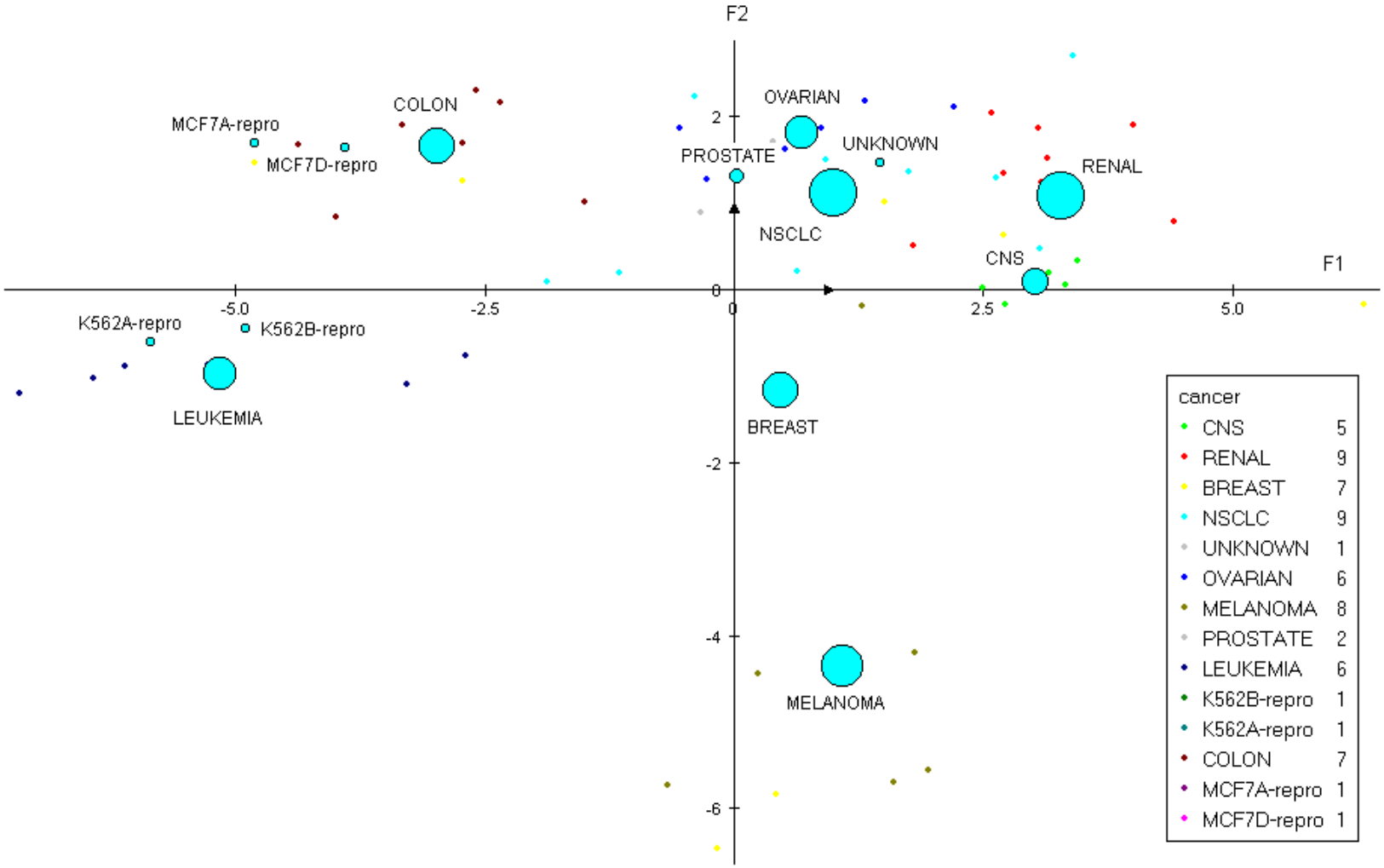
Iris setosa

Iris versicolor

Iris virginica

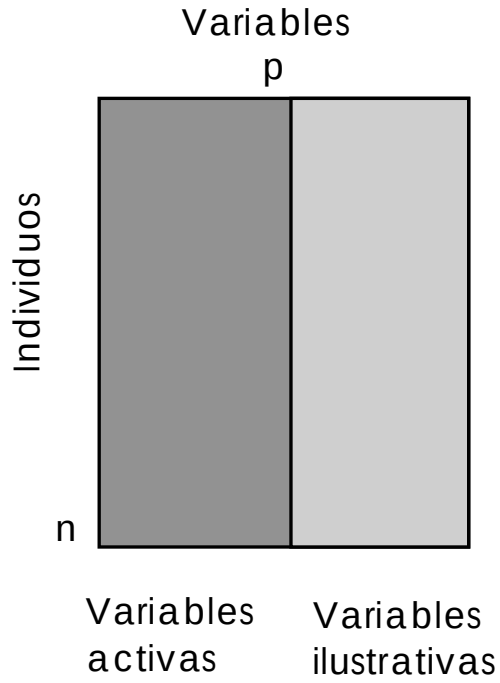
Principal Components Analysis: Visualization of the information contained in a data matrix

Microarray data: 64 cancers 6830 gen cromotografy



1st step: Selection of the active variables

Usually Data can be divided in topics, according the semantics of variables



Active variables
must be of the same type, *all continuous*, and referring to the *same concept*

Principal Component Analysis

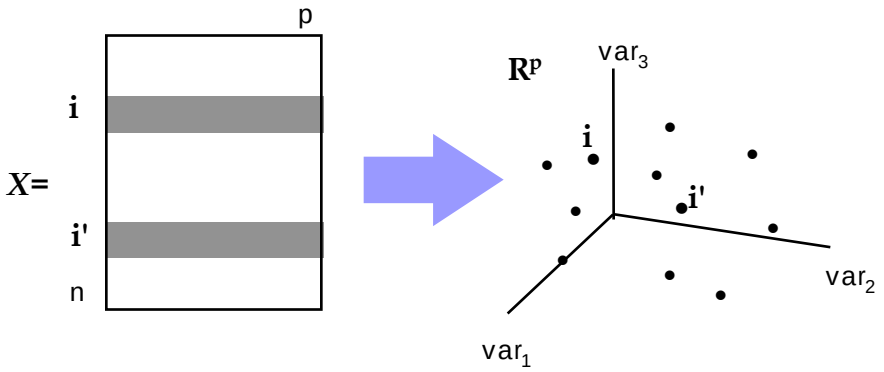
The geometrical idea:

Harold Hotelling, 1895-1973
American statistician

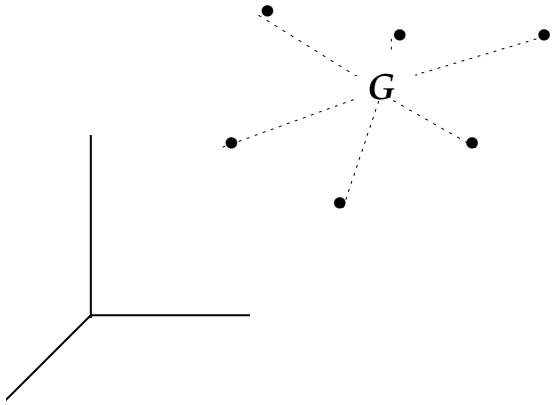
Let's suppose a table X of n individuals with p "active" numeric variables

1. **Cloud of points associated to the rows of the data matrix**

   measuring distances between points

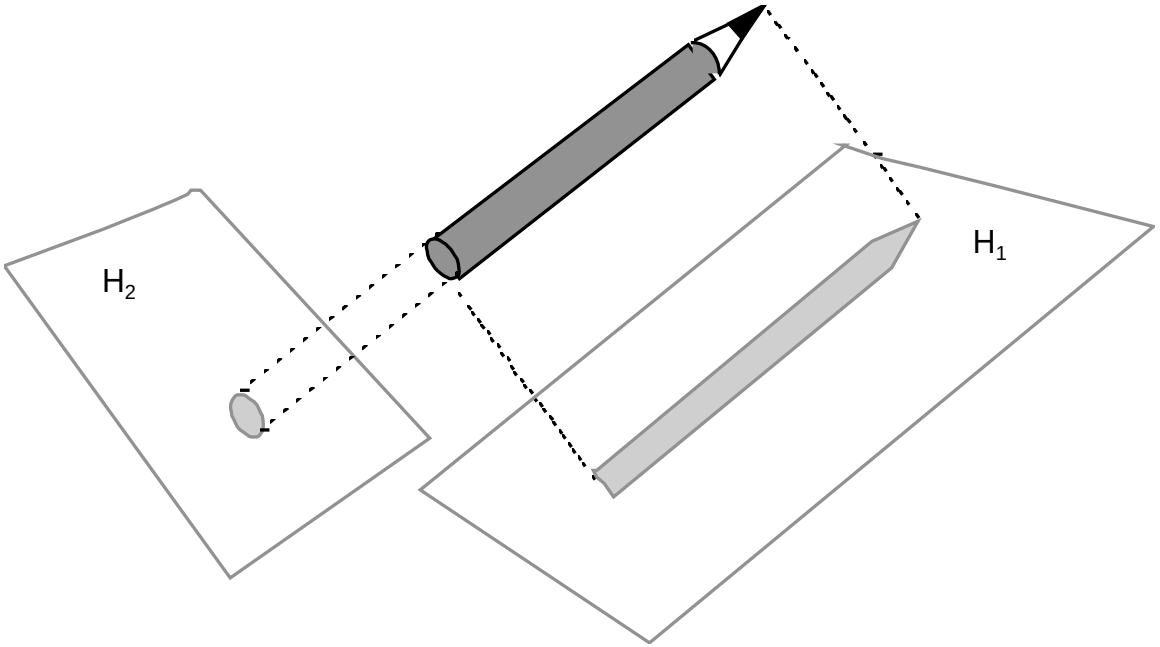


2. **Total information contained in the cloud of points: *the inertia respect G***



PCA: Criterion to optimize

- Purpose:
  - To project the cloud of points upon a subspace (a plan) to retain the maximum of the original cloud **information**.



the PCA mathematics

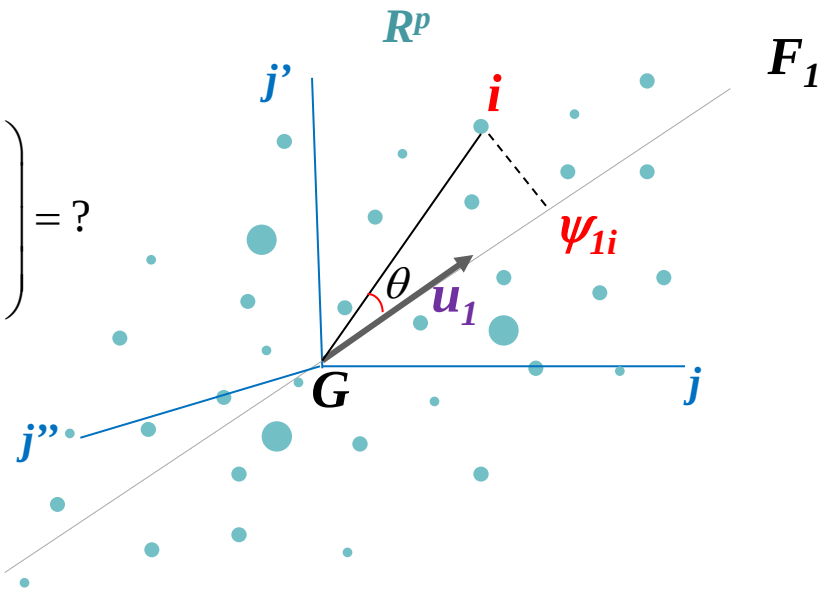
Let X be the standardized data matrix
(→ all variables have the same importance)

$$X = \begin{array}{c} \\ i \\ \\ n \end{array} \begin{array}{c} \begin{array}{cc} j & p \end{array} \\ \begin{bmatrix} & \\ \cdots & \frac{x_{ij} - \bar{x}_j}{s_j} & \cdots \\ & \end{bmatrix} \end{array}$$

$$u_1 = \begin{pmatrix} u_{11} \\ \vdots \\ u_{p1} \end{pmatrix} = ?$$

$$|u_1| = 1$$

Fitness Criterion: Find the subspace maximizing the projected inertia.



We project all points ***i*** on the direction defined by u_1

$$\psi_{1i} = d(i,G) \times \cos(\theta) = x_i' u_1$$

$$\psi_1 = \begin{bmatrix} \psi_{11} \\ \vdots \\ \psi_{1n} \end{bmatrix} = Xu_1$$

We try to find the direction u_1 maximizing the projected inertia

$$\underset{u}{Max} \sum_{i=1}^{n} \frac{1}{n} \psi_{1i}^2$$

the PCA solution

$$\underset{u_1}{Max}\left(u_1' \tfrac{1}{n} X'X\ u_1\right) \rightarrow \underset{u_1}{Max}\left(u_1' R\ u_1\right)$$

solution

X standardized

$$X \leftarrow \begin{pmatrix} \vdots \\ \dfrac{x_{ij} - \bar{x}_j}{s_j} \\ \vdots \end{pmatrix}$$

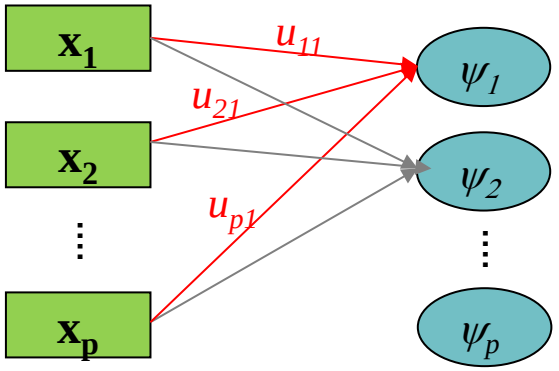
Diagonalization of the correlation matrix *R*

1. → eigenvalues $\rightarrow \lambda_1, \ldots, \lambda_p \quad p = rang(X)$
2. → eigenvectors $\rightarrow u_1, \ldots, u_p$

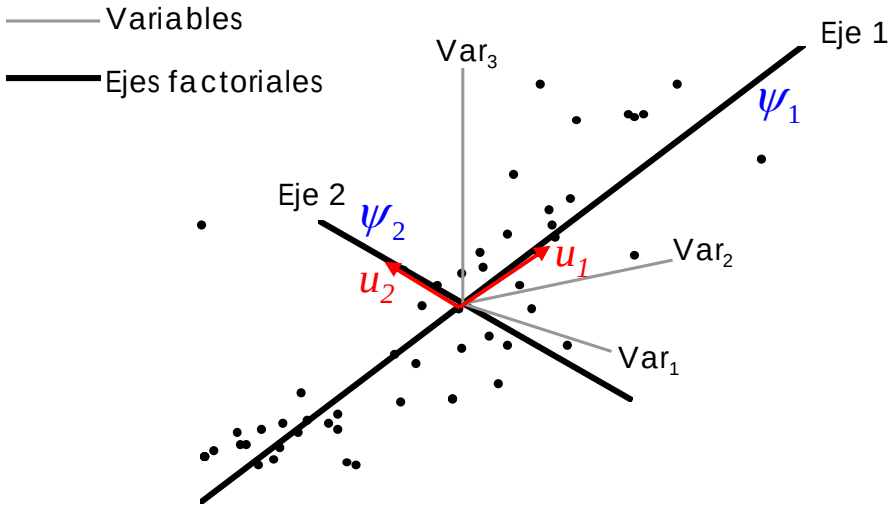
$$\lambda_1 > \lambda_2 > \cdots > \lambda_p \qquad \sum_{\alpha=1}^{p} \lambda_\alpha = p$$

The principal components:

$$\psi_\alpha = Xu_\alpha = u_{1\alpha}\mathbf{x}_1 + u_{2\alpha}\mathbf{x}_2 + \cdots\cdots\cdots + u_{p\alpha}\mathbf{x}_p$$



Interpreting the PCA solution



$$\sum_{i=1}^{n}\frac{1}{n}\psi_{1i}^2=\lambda_1 \qquad \sum_{i=1}^{n}\frac{1}{n}\psi_{2i}^2=\lambda_2 \qquad \cdots\cdots\cdots\cdots\cdots\cdots \qquad \sum_{i=1}^{n}\frac{1}{n}\psi_{pi}^2=\lambda_p$$

$$I_{total}=\mathrm{var}(\psi_1)+\mathrm{var}(\psi_2)+\cdots+\mathrm{var}(\psi_p)=\lambda_1+\lambda_2+\cdots+\lambda_p$$

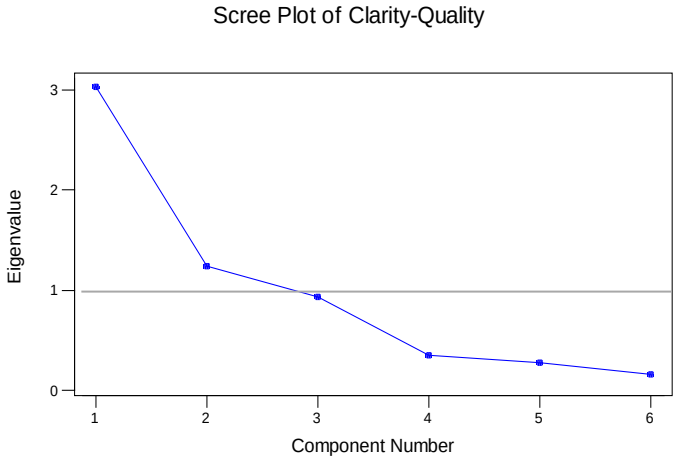
Finding the number of significant components

$$\tau_\alpha = \frac{\text{var}(\psi_\alpha)}{I_T} = \frac{\lambda_\alpha}{\sum_{j=1}^{p} \lambda_j}$$

Importance of a Principal Component:

Proportion of the total variance explained by the α principal component

Screeplot



Ad hoc rules:

1. *Kaiser rule*. Take all principal components with variance greater than the average variance *(=1)*

2. *Last elbow rule* (using a screeplot)

3. *Taking all factorial coordinates up to a fixed percentage of inertia* (≅80%)

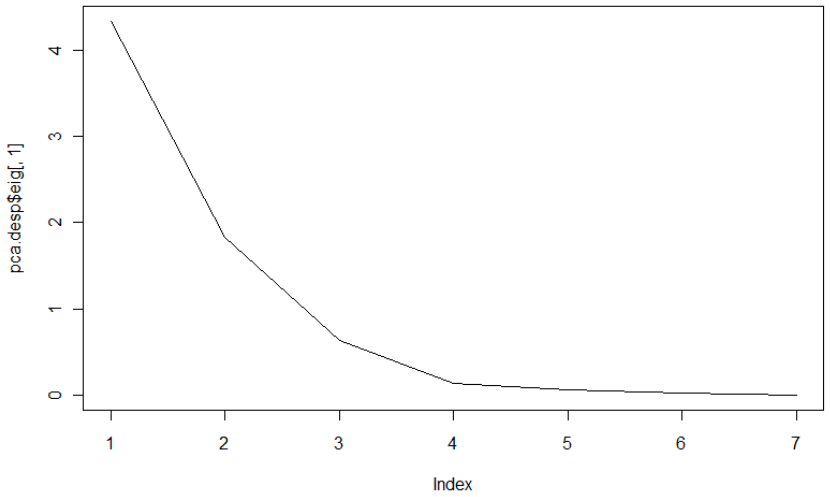
Application of PCA

Expenses by household CSP and size

```
> despeses <- read.table("expenses.txt",header=T,sep="\t")>
> despeses
                        X bread vegetables fruits meat poultry milk wine STATUS CHILDREN
1           Manual 2 niños   332        428    354 1437     526  247  427      1        2
2         Empleado 2 niños   293        559    388 1527     567  239  258      2        2
3       Ejecutivo 2 niños   372        767    562 1948     927  235  433      3        2
4           Manual 3 niños   406        563    341 1507     544  324  407      1        3
5         Empleado 3 niños   386        608    396 1501     558  319  363      2        3
6       Ejecutivo 3 niños   438        843    689 2345    1148  243  341      3        3
7           Manual 4 niños   534        660    367 1620     638  414  407      1        4
8         Empleado 4 niños   460        699    484 1856     762  400  416      2        4
9       Ejecutivo 4 niños   385        789    621 2366    1149  304  282      3        4
10    Manual 5 o más niños   655        776    423 1848     759  495  486      1        5
11  Empleado 5 o más niños   584        995    548 2056     893  518  319      2        5
12 Ejecutivo 5 o más niños   515       1097    887 2630    1167  561  284      3        5

> despeses$STATUS <- as.factor(despeses$STATUS)
> despeses$CHILDREN <- as.factor(despeses$CHILDREN)

> library(FactoMineR)

> pca.desp <- PCA(despeses,quali.sup=c(8,9))
> attributes(pca.desp)
$names
[1] "eig"       "var"       "ind"       "svd"       "quali.sup" "call"

$class
[1] "PCA"   "list "
```

Detecting the significant components?

```
> pca.desp$eig
        eigenvalue percentage of variance cumulative percentage of variance
comp 1 4.3332373164             61.90339023                          61.90339
comp 2 1.8302901700             26.14700243                          88.05039
comp 3 0.6308364243              9.01194892                          97.06234
comp 4 0.1283275007              1.83325001                          98.89559
comp 5 0.0575561897              0.82223128                          99.71782
comp 6 0.0188486021              0.26926574                          99.98709
comp 7 0.0009037968              0.01291138                         100.00000

> plot(pca.desp$eig$eigenvalue,type="l")
```



How many components are significant?

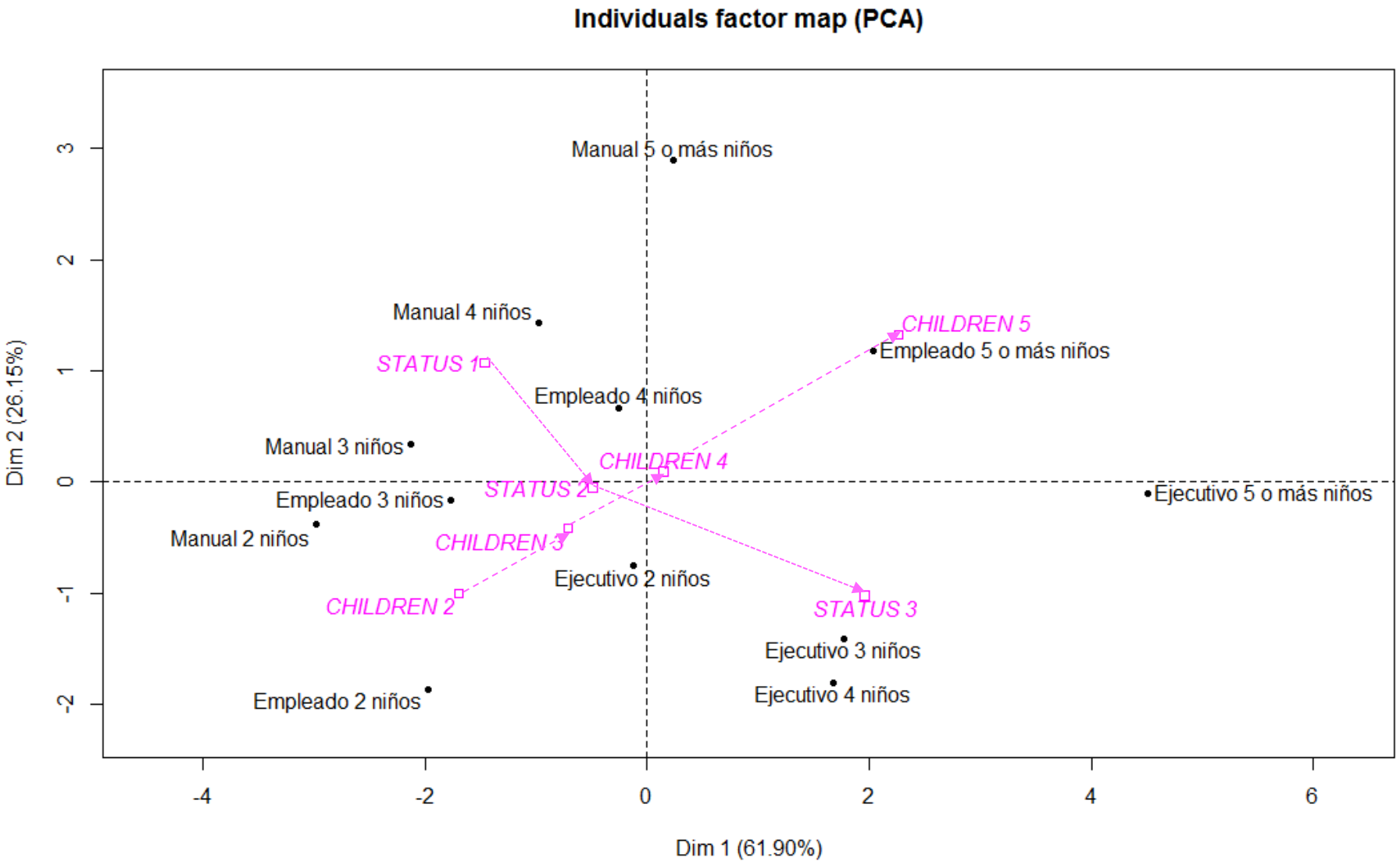
The projection of individuals

$$\psi_\alpha = \mathbf{X}u_\alpha \quad \alpha = 1,\ldots,p$$

```
> pca.desp$ind$coord
```

	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5
	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
Manual 2 niños	-2.9858103	-0.3797425	0.4216600	0.3770988	0.23609399
Empleado 2 niños	-1.9732459	-1.8719298	-1.3641374	-0.1687554	-0.09597354
Ejecutivo 2 niños	-0.1231418	-0.7577698	1.4849255	0.2049872	-0.46387873
Manual 3 niños	-2.1315044	0.3376824	-0.1108927	0.1080235	0.01212832
Empleado 3 niños	-1.7696521	-0.1699392	-0.5356943	0.1645030	-0.18367198
Ejecutivo 3 niños	1.7696361	-1.4151333	1.0392345	-0.4483579	-0.07933595
Manual 4 niños	-0.9743211	1.4326959	-0.2879256	-0.2752643	0.09780071
Empleado 4 niños	-0.2633801	0.6606214	0.2855169	0.2997326	0.16852778
Ejecutivo 4 niños	1.6716869	-1.8132061	0.1028634	-0.4189169	0.44016323
Manual 5 o más niños	0.2308559	2.9002954	0.5922819	-0.2567171	0.12699170
Empleado 5 o más niños	2.0379247	1.1814920	-1.0334300	-0.3385603	-0.34040283
Ejecutivo 5 o más niños	4.5109523	-0.1050664	-0.5944022	0.7522267	0.08155730

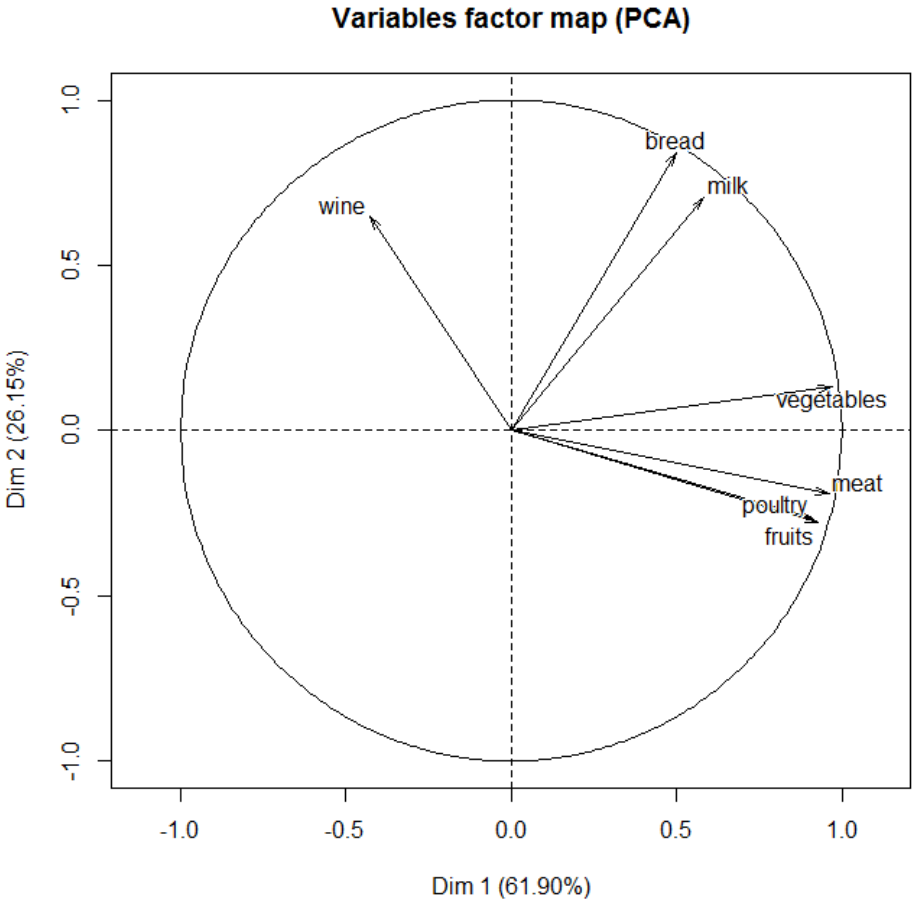
The first map of individuals with the suplementary categories



Interpreting the principal components: The correlation map

$$cor(X,\Psi)=\begin{bmatrix} cor(x_1,\psi_1)\ cor(x_1,\psi_2)\cdots \\ \vdots \qquad\qquad \vdots \qquad \cdots \\ cor(x_p,\psi_1)\,cor(x_p,\psi_2)\cdots \end{bmatrix}$$

```
> pca.desp$var$cor
             Dim.1  Dim.2  Dim.3  Dim.4  Dim.5
bread        0.499  0.842 -0.009 -0.195  0.009
vegetables   0.970  0.133 -0.049 -0.008 -0.194
fruits       0.929 -0.278  0.115  0.196 -0.016
meat         0.962 -0.191  0.165 -0.019  0.099
poultry      0.911 -0.266  0.283 -0.116  0.054
milk         0.584  0.707 -0.352  0.161  0.082
wine        -0.428  0.648  0.619  0.110 -0.016
```



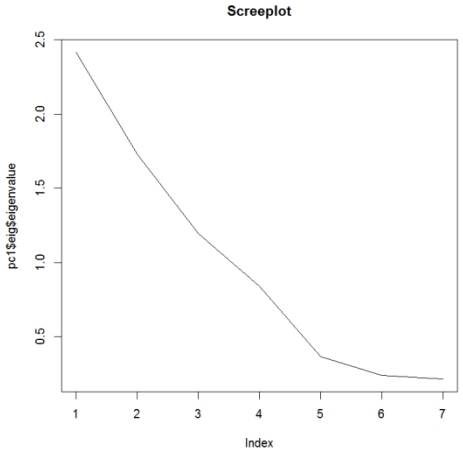
The first component is related with…
The second with …
And the third with …

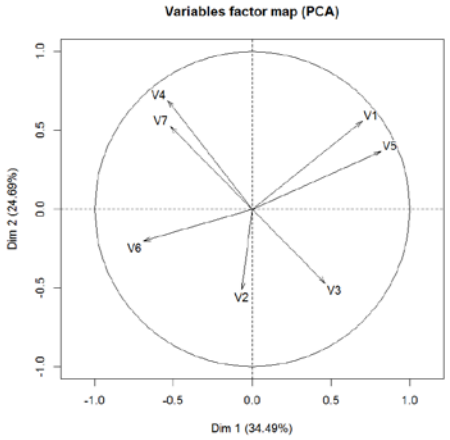
FINDING THE LATENT (INTANGIBLE) CONCEPTS

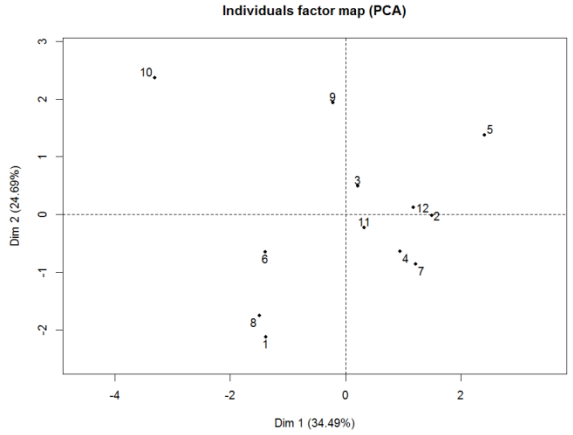
The two extreme cases

random generated data n=12, p=7

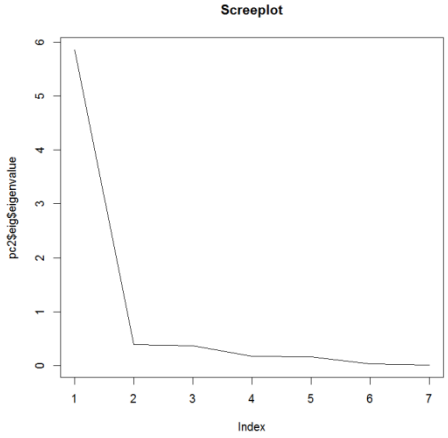
The independence case, ρ_{jk}=0

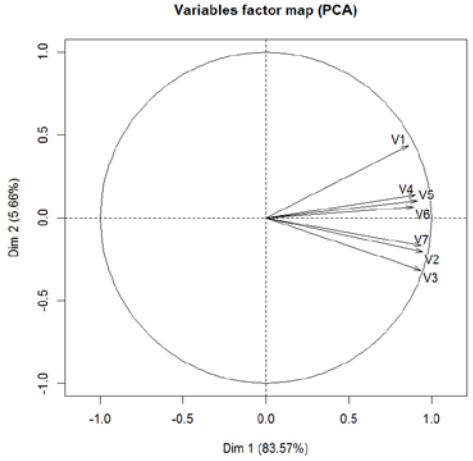


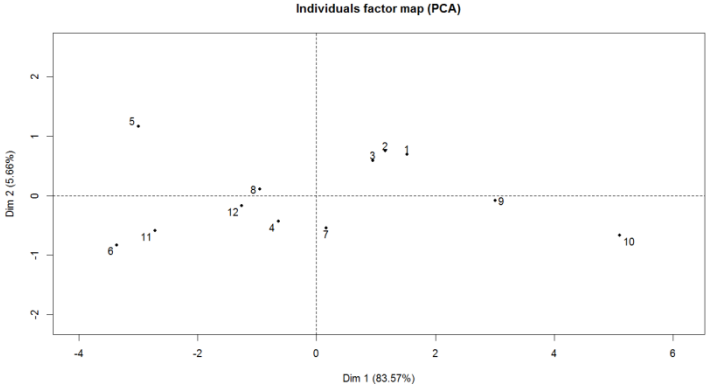




The unidimensional case, ρ_{jk}=0.9

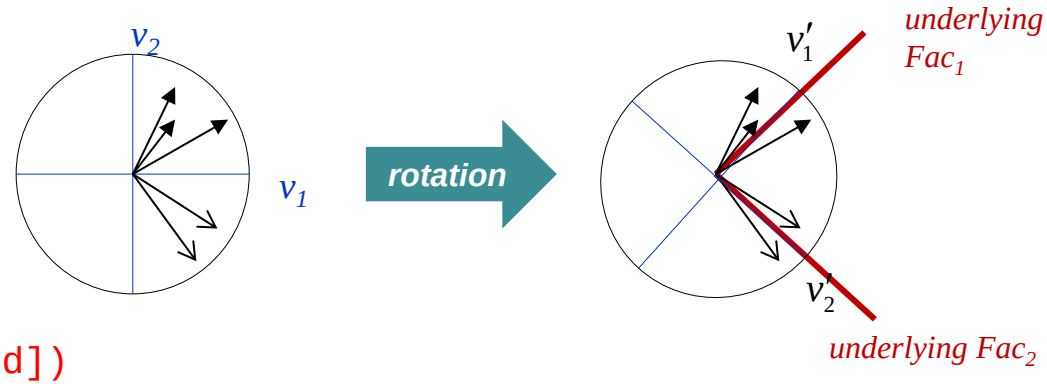






Finding the latent concepts (intangibles)

VARIMAX rotation:



nd = number of significant dimensions

```
> pc.rot = varimax(pc$var$cor[,1:nd])
$loadings
```

Loadings: (= correlations)

	Dim.1	Dim.2	Dim.3
bread	0.184	0.914	0.296
vegetables	0.781	0.559	-0.196
fruits	0.932	0.144	-0.253
meat	0.956	0.212	-0.176
poultry	0.981		-0.103
milk	0.157	0.967	
wine	-0.307	0.115	0.938

	Dim.1	Dim.2	Dim.3
SS loadings	3.508	2.168	1.118
Proportion Var	0.501	0.310	0.160
Cumulative Var	0.501	0.811	0.971

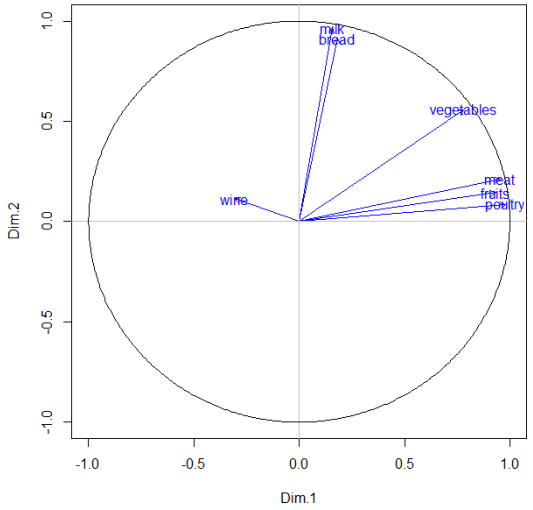
Rotated dimensions are such that variables tend to be very correlated with one of them and zero correlated with the others.

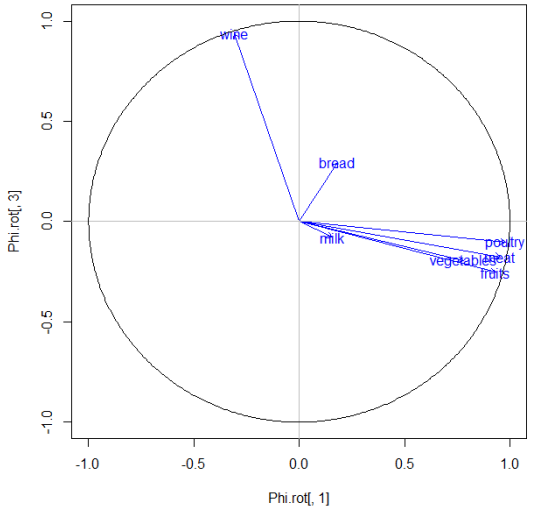
Rotated dimensions highlight the hidden latent concepts present in our observed data

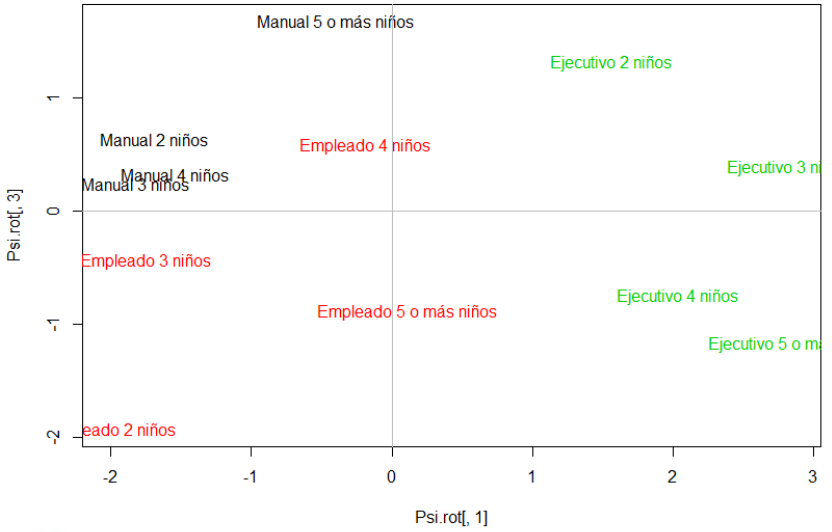
The first rotated component corresponds to …
The second corresponds to …
And the third corresponds to …

Rotation allows an automatic feature extraction/variable selection operation

Mapping the rotated PCA of expenses







$$\psi_{rot}^{stan} = XR^{-1}\varphi_{rot}$$

APPLICATION 1: TYPOLOGY OF CARS

1st. PCA applicattion
Defining a typology of cars

R commands

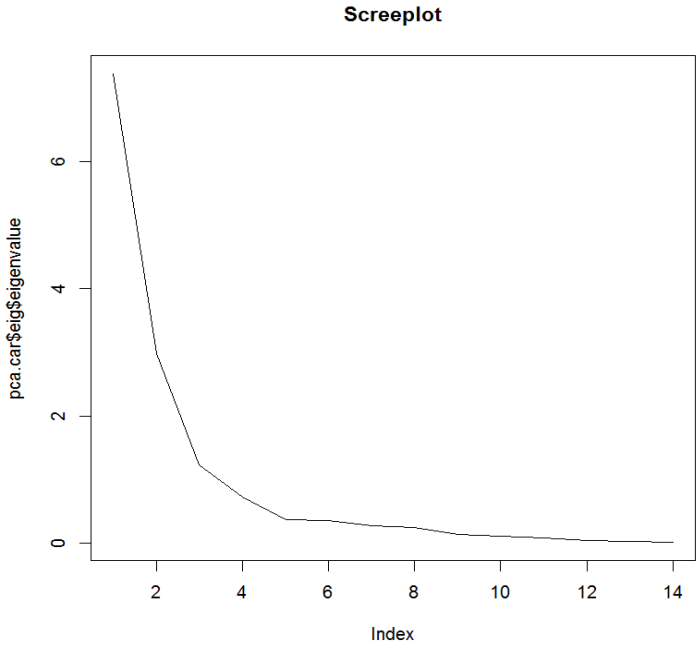
```
setwd("C:/path/to/directory")        #ESPECIFICAR EL ARCHIVO DE TRABAJO
library(FactoMineR)
car <- read.csv("car.csv",header=T,sep=";",dec=".")
```

0 step: A first overview of the data

```
> summary(car)
   cilindrada      potencia        combustible   revoluciones    cilindros        longitud         ancho
 Min.   : 903   Min.   : 37.00   Diesel  : 84   Min.   :3900   Min.   : 2.000   Min.   :313.0   Min.   :148.0
 1st Qu.:1596   1st Qu.: 83.25   Gasolina:406   1st Qu.:5200   1st Qu.: 4.000   1st Qu.:404.0   1st Qu.:165.0
 Median :1955   Median :117.50                  Median :5600   Median : 4.000   Median :435.5   Median :169.0
 Mean   :2147   Mean   :129.79                  Mean   :5507   Mean   : 4.651   Mean   :430.8   Mean   :169.2
 3rd Qu.:2314   3rd Qu.:160.00                  3rd Qu.:6000   3rd Qu.: 5.000   3rd Qu.:461.5   3rd Qu.:174.0
 Max.   :6750   Max.   :478.00                  Max.   :7200   Max.   :12.000   Max.   :527.0   Max.   :198.0

     altura         maletero          peso           plazas        velocidad     poca_aceleracion
 Min.   :113.0   Min.   :  0.0   Min.   : 645.0   Min.   :2.000   Min.   :118.0   Min.   : 4.70
 1st Qu.:137.0   1st Qu.:300.8   1st Qu.: 971.2   1st Qu.:5.000   1st Qu.:173.0   1st Qu.: 8.50
 Median :140.0   Median :420.0   Median :1155.0   Median :5.000   Median :192.0   Median :10.40
 Mean   :139.1   Mean   :402.1   Mean   :1174.8   Mean   :4.896   Mean   :192.9   Mean   :11.01
 3rd Qu.:142.0   3rd Qu.:500.0   3rd Qu.:1325.0   3rd Qu.:5.000   3rd Qu.:212.0   3rd Qu.:12.50
 Max.   :188.0   Max.   :675.0   Max.   :2430.0   Max.   :7.000   Max.   :324.0   Max.   :23.00

      traccion      consumo        coste.Km         precio             marca
 4x4      : 38   Min.   : 4.70   Min.   : 6.90   Min.   :  865   MERCEDES  : 43
 Delantera:312   1st Qu.: 7.90   1st Qu.:11.80   1st Qu.: 1803   RENAULT   : 41
 Trasera  :140   Median : 8.90   Median :14.20   Median : 2794   VOLKSWAGEN: 39
                 Mean   : 9.45   Mean   :14.39   Mean   : 4104   PEUGEOT   : 35
                 3rd Qu.:10.60   3rd Qu.:16.60   3rd Qu.: 4726   OPEL      : 34
                 Max.   :19.60   Max.   :27.40   Max.   :50000   FORD      : 31
                                                                 (Other)   :267
```

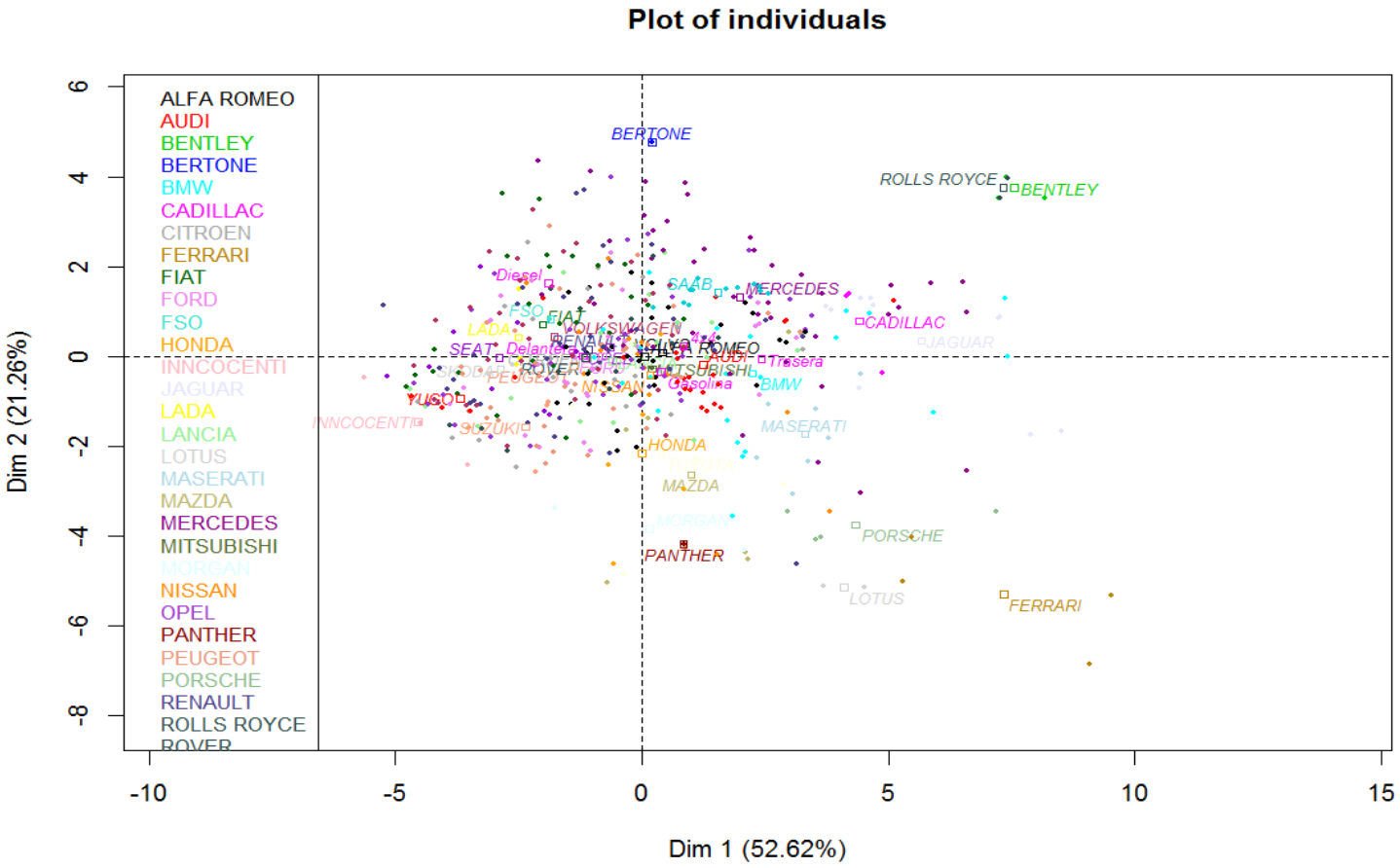
Finding the number of significant dimensions

```
pca.car <- PCA(car,quali.sup=c(3,14,18),quanti.sup=17)

# ELECCION DEL NUMERO DE COMPONENTES SIGNIFICATIVAS
pca.car$eig
        eigenvalue perc. of variance cum. perc,. of variance
comp 1  7.36670576        52.6193269                 52.61933
comp 2  2.97697244        21.2640889                 73.88342
comp 3  1.23531061         8.8236472                 82.70706
comp 4  0.72904964         5.2074974                 87.91456
comp 5  0.36892411         2.6351722                 90.54973
comp 6  0.35392066         2.5280047                 93.07774
comp 7  0.27360660         1.9543328                 95.03207
comp 8  0.24743114         1.7673653                 96.79944
comp 9  0.14386098         1.0275784                 97.82701
comp 10 0.11558728         0.8256234                 98.65264
comp 11 0.08467760         0.6048400                 99.25748
comp 12 0.04116119         0.2940085                 99.55149
comp 13 0.03806771         0.2719122                 99.82340
comp 14 0.02472428         0.1766020                100.00000

plot(pca.car$eig$eigenvalue,type="l",main="Screeplot")

# CUANTAS COMPONENTES SON SIGNFICATIVAS?
```



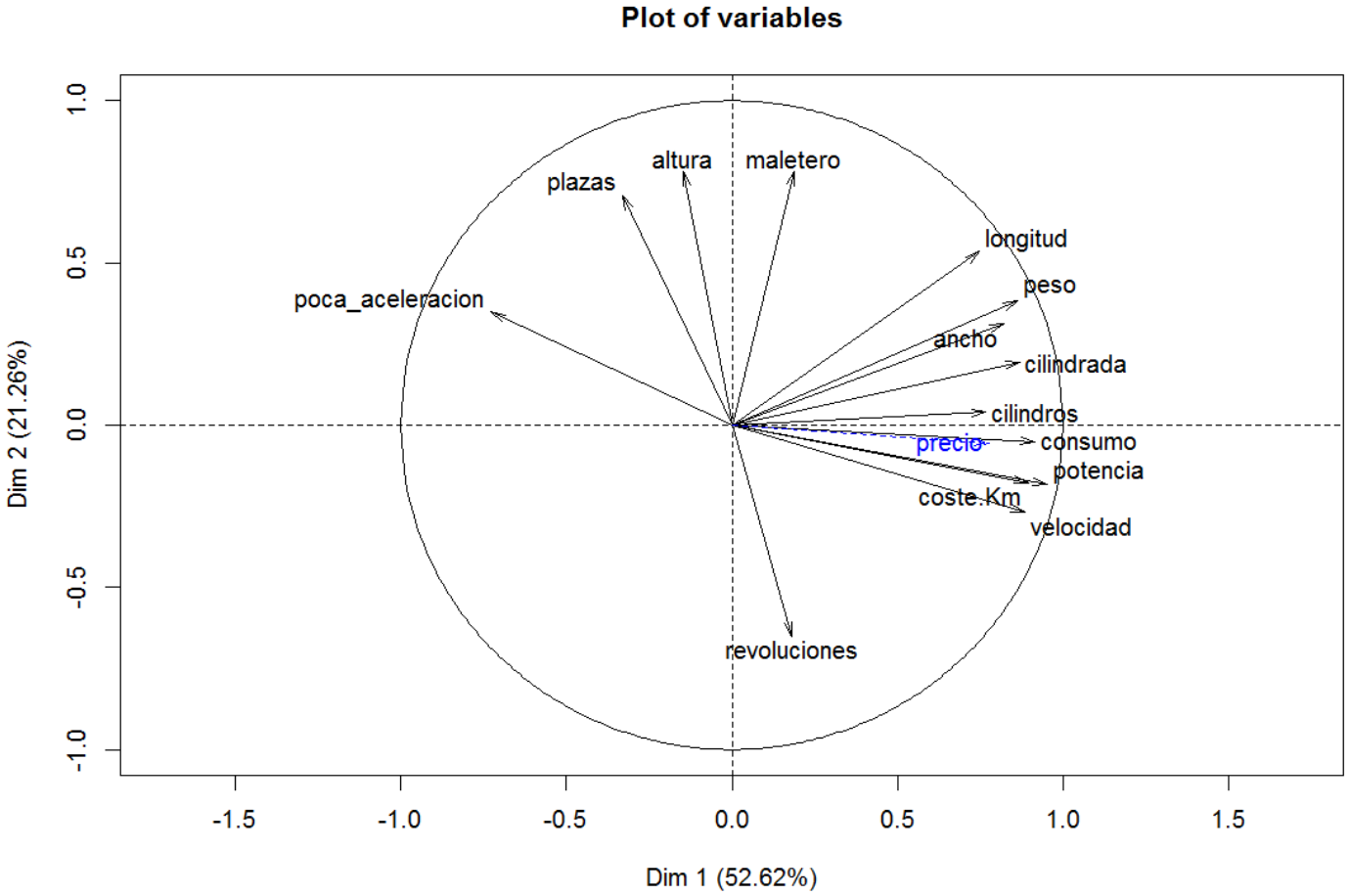
The individuals plot with the brand name

```
# GRAFICO DE LOS INDIVIDUOS EN LAS DIMENSIONES 1 Y 2
plot(pca.car, axes = c(1, 2), choix = c("ind"), habillage=18, label="quali",
title="Plot of individuals", cex=0.7)
```



The variables correlation map

```
# GRAFICO DE LAS VARIABLES EN LAS DIMENSIONES 1 Y 2
plot(pca.car, axes = c(1, 2), choix = c("var"), title="Plot of variables")
```



Looking for the latent concepts

```
> # ROTACION DE LAS COMPONENTES PARA BUSCAR COMPONENTES MAS INTERPRETABLES
> pca.car.rot <- varimax(pca.car$var$cor[,1:nd])
> pca.car.rot
$loadings

Loadings:
                Dim.1  Dim.2  Dim.3  Dim.4
cilindrada       0.914                -0.303
potencia         0.734 -0.251  0.491 -0.313
revoluciones          -0.140  0.875  0.240
cilindros        0.904 -0.105
longitud         0.502  0.168        -0.796
ancho            0.543         0.139 -0.748
altura                 0.894 -0.202
maletero               0.550        -0.728
peso             0.784                -0.536
plazas          -0.226  0.853 -0.153
velocidad        0.531 -0.282  0.648 -0.376
poca_aceleracion -0.348  0.223 -0.757  0.275
consumo          0.848         0.387 -0.179
coste.Km         0.735         0.566 -0.168

                Dim.1  Dim.2  Dim.3  Dim.4
SS loadings     5.077  2.109  2.566  2.556
Proportion Var  0.363  0.151  0.183  0.183
Cumulative Var  0.363  0.513  0.697  0.879
```

APPLICATION 2: TYPOLOGY OF BANK CLIENTS

2nd. PCA applicattion
Finding a typology of clients of a bank

R commands

```
> library(FactoMineR)
> Xtot_act <- read.table(file="Xtot_act.txt", header=TRUE)
> dim(Xtot_act)
[1] 10495    44
```

0 step: A first overview of the data

```
> summary(Xtot_act)
     vista            termini          divises           asseg_apor        asseg_mer          fonds_inv
 Min.   :  0.00   Min.   :  0.00   Min.   :  0.0000   Min.   :  0.000   Min.   :  0.0000   Min.   :  0.000
 1st Qu.: 26.45   1st Qu.:  0.00   1st Qu.:  0.0000   1st Qu.:  0.000   1st Qu.:  0.0000   1st Qu.:  0.000
 Median :100.00   Median :  0.00   Median :  0.0000   Median :  0.000   Median :  0.0000   Median :  0.000
 Mean   : 72.11   Mean   : 12.42   Mean   :  0.0391   Mean   :  5.538   Mean   :  0.1935   Mean   :  1.793
 3rd Qu.:100.00   3rd Qu.:  0.00   3rd Qu.:  0.0000   3rd Qu.:  0.000   3rd Qu.:  0.0000   3rd Qu.:  0.000
 Max.   :100.00   Max.   :100.00   Max.   :100.0000   Max.   :100.000   Max.   :100.0000   Max.   :100.000

   tit_propis        interm_pag         renda_fixa         renda_var          altre_pas          hipoteques
 Min.   :  0.000   Min.   :  0.0000   Min.   :  0.0000   Min.   :  0.0000   Min.   :  0.0000   Min.   :  0.000
 1st Qu.:  0.000   1st Qu.:  0.0000   1st Qu.:  0.0000   1st Qu.:  0.0000   1st Qu.:  0.0000   1st Qu.:  0.000
 Median :  0.000   Median :  0.0000   Median :  0.0000   Median :  0.0000   Median :  0.0000   Median :  0.000
 Mean   :  0.884   Mean   :  0.5051   Mean   :  0.3856   Mean   :  0.5183   Mean   :  0.5469   Mean   :  6.699
 3rd Qu.:  0.000   3rd Qu.:  0.0000   3rd Qu.:  0.0000   3rd Qu.:  0.0000   3rd Qu.:  0.0000   3rd Qu.:  0.000
 Max.   :100.000   Max.   :100.0000   Max.   :100.0000   Max.   :100.0000   Max.   :100.0000   Max.   :100.000

  pres_person       tar_credit        cmp_credit         desc_comer          risc_aval         passiu_total
 Min.   :  0.00   Min.   :  0.00   Min.   :  0.0000   Min.   : 0.00000   Min.   :  0.000   Min.   :    0.0
 1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.0000   1st Qu.: 0.00000   1st Qu.:  0.000   1st Qu.:   27.0
 Median :  0.00   Median :  0.00   Median :  0.0000   Median : 0.00000   Median :  0.000   Median :  155.0
 Mean   : 42.71   Mean   :  2.41   Mean   :  0.4435   Mean   : 0.05664   Mean   :  2.097   Mean   :  833.7
 3rd Qu.:100.00   3rd Qu.:  0.00   3rd Qu.:  0.0000   3rd Qu.: 0.00000   3rd Qu.:  0.000   3rd Qu.:  716.5
 Max.   :100.00   Max.   :100.00   Max.   :100.0000   Max.   :83.22332   Max.   :100.000   Max.   :58793.0
```

A first overview of the data (cont)

```
actiu_total        imp_antic              edat          sexe              ocupacio       forma_ex_ocup
Min.   :    0.0   Min.   :  0.000   30:34_anys:1681   dona   :3662   empl_no_qualif:3626   cta_aliena:5408
1st Qu.:    0.0   1st Qu.:  0.000   35:39_anys:1547   home   :6650   mestresses    :1424   cta_propia:1236
Median :   80.0   Median :  0.000   25:29_anys:1546   sexe_NA: 183   menor-est     :1241   inactius  :3851
Mean   :  833.1   Mean   :  3.285   40:44_anys:1382                  white_collar  : 842
3rd Qu.:  789.5   3rd Qu.:  0.000   45:49_anys:1336                  universitaris : 824
Max.   :59443.0   Max.   :100.000   18:24_anys: 958                  ocup_NA       : 731
                                    (Other)   :2045                  (Other)       :1807

    provincia           primer_rel      nomina      pensionista     car_ab_period     targetes
BARCELONA:6941   PR_1974:1983:4225   Nom_NO:5526   Pens_NO:8305   Car-Ab_NO:1320   Targ_NO:3433
BALEARS  : 762   prim_rel_NA :2285   Nom_SI:4969   Pens_SI:2190   Car-Ab_SI:9175   Targ_SI:7062
GIRONA   : 690   PR_1984:1988:2179
TARRAGONA: 676   PR_1989:1990:1541
LLEIDA   : 523   PR_1991:1992: 212
MADRID   : 174   PR_1964:1973:  34
(Other)  : 729   (Other)     :  19

      seg_vida         tar_debit         visa_master        tar_comer         ambit_rel    ofic_rel
Seg_vid_NO:7119   Tar_deb_NO:6160   Vis_Mast_NO:5831   Tar_com_NO:9835   BCN_ciutat:3154   of_1:5330
Seg_vid_SI:3376   Tar_deb_SI:4335   Vis_Mast_SI:4664   Tar_com_SI: 660   BCN_prov  :3806   of_2:3345
                                                                         Rest_Cat  :2718   of_3:1340
                                                                         Rest_Esp  : 817   of_4: 480

              imp_ab_tot                    imp_car_tot              imp_compr_com      cim         publicitat
Ab_(1.8e+03,3e+03]  :2083   Car_(1.8e+03,3e+03]  :2103   Comç_(-1,0]       :5458   CIM_NO:9423   pub_NO:3579
Ab_(3e+03,4e+03]    :1346   Car_(3e+03,4e+03]    :1271   Comç_(0,30]       : 676   CIM_SI:1072   pub_SI:6916
Ab_(5e+03,8e+03]    :1227   Car_(5e+03,8e+03]    :1210   Comç_(100,200]    : 756
Ab_(1.2e+03,1.8e+03]:1158   Car_(0,400]          :1188   Comç_(1e+03,1e+12]: 993
Ab_(8e+03,1e+07]    :1089   Car_(1.2e+03,1.8e+03]:1171   Comç_(200,500]    :1069
Ab_(0,400]          :1050   Car_(8e+03,1e+07]    :1093   Comç_(30,100]     : 908
(Other)             :2542   (Other)              :2459   Comç_(500,1e+03]  : 635

  preconcedit      contractat          destinacio
precon_NO:6829   contr_NO:5440   dest_NA     :5440
precon_SI:3666   contr_SI:5055   Mobiliari   :1732
                                 Vehicles    :1406
                                 Reste       :1042
                                 Empreses    : 454
                                 No_habitatge: 186
                                 (Other)     : 235
```

1 step: Selection of the active variables

```
> names(Xtot_act)
 [1] "vista"          "termini"
 [3] "divises"        "asseg_apor"
 [5] "asseg_mer"      "fonds_inv"
 [7] "tit_propis"     "interm_pag"
 [9] "renda_fixa"     "renda_var"
[11] "altre_pas"      "hipoteques"
[13] "pres_person"    "tar_credit"
[15] "cmp_credit"     "desc_comer"
[17] "risc_aval"      "passiu_total"
[19] "actiu_total"    "imp_antic"
[21] "edat"           "sexe"
[23] "ocupacio"       "forma_ex_ocup"
[25] "provincia"      "primer_rel"
[27] "nomina"         "pensionista"
[29] "car_ab_period"  "targetes"
[31] "seg_vida"       "tar_debit"
[33] "visa_master"    "tar_comer"
[35] "ambit_rel"      "ofic_rel"
[37] "imp_ab_tot"     "imp_car_tot"
[39] "imp_compr_com"  "cim"
[41] "publicitat"     "preconcedit"
[43] "contractat"     "destinacio"
```

Active variables

Continuous supplementary variables

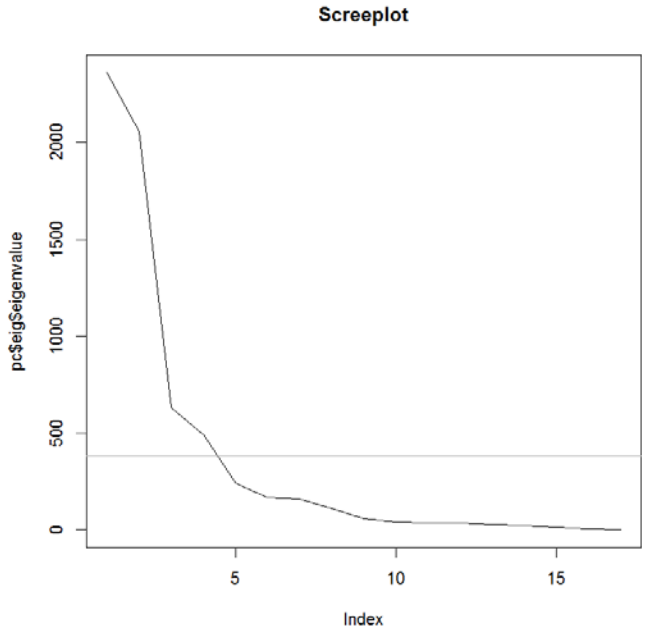
Categorical supplementary variables

Running the PCA

```
> pc <- PCA(Xtot_act, ncp=10, quanti.sup=c(18:20), quali.sup=c(21:44), scale.unit=F)
```

2. step: Detecting the number of significant components

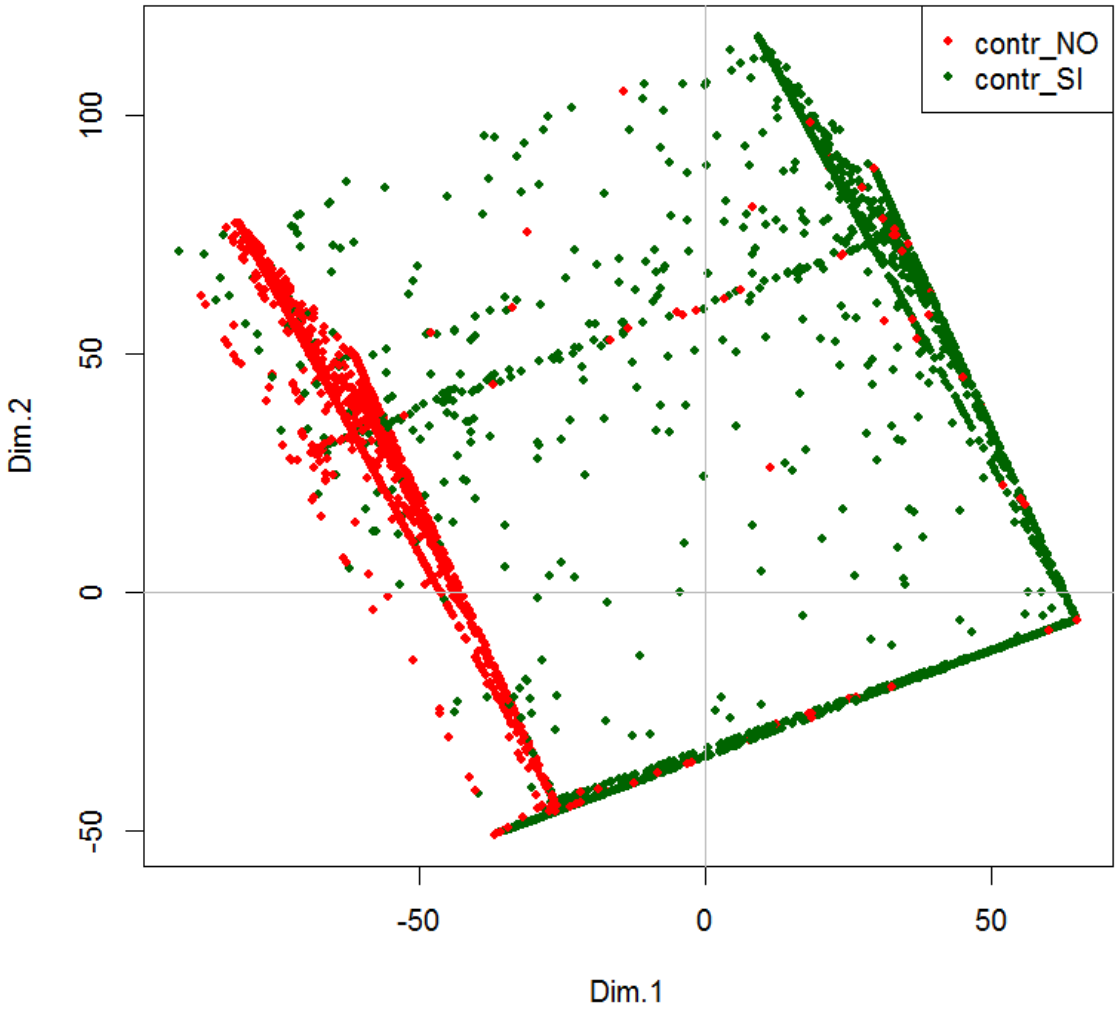
```
> plot(pc$eig$eigenvalue,type="l",main="Screeplot")
> abline(h=mean(pc$eig$eigenvalue),col="gray")
```



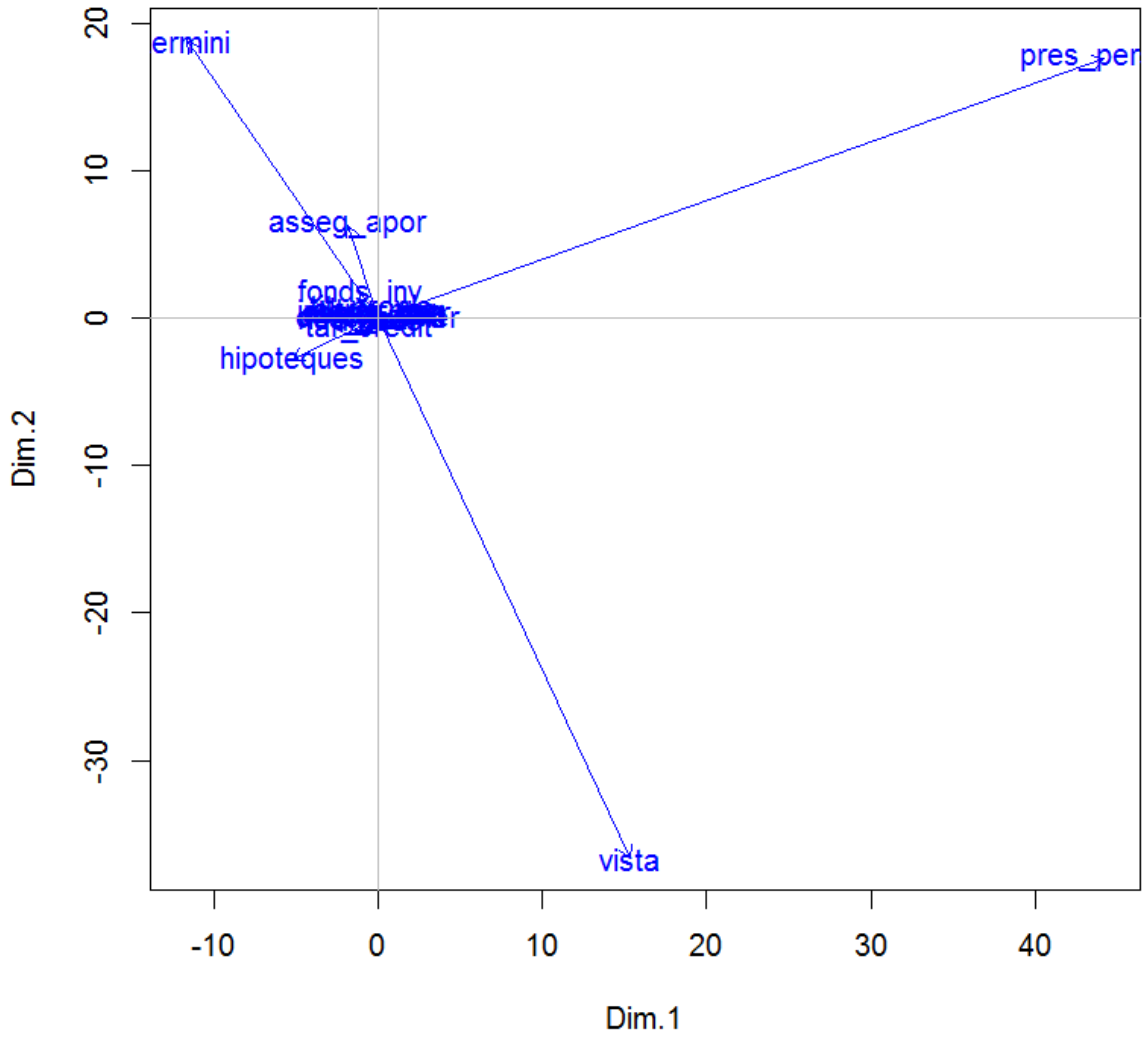

```
> pc$eig
        eigenvalue percentage of variance cumulative % of variance
comp 1  2358.443643             36.48682076                  36.48682
comp 2  2057.587959             31.83236676                  68.31919
comp 3   632.507807              9.78535104                  78.10454
comp 4   492.111352              7.61331683                  85.71786
comp 5   243.941958              3.77395768                  89.49181
comp 6   168.375232              2.60488603                  92.09670
comp 7   158.007198              2.44448509                  94.54118
comp 8   112.029846              1.73318236                  96.27437
comp 9    56.652211              0.87645048                  97.15082
comp 10   40.440213              0.62563920                  97.77646
comp 11   37.214816              0.57574000                  98.35220
comp 12   34.115483              0.52779107                  98.87999
comp 13   29.515129              0.45662030                  99.33661
comp 14   23.020020              0.35613628                  99.69274
comp 15   13.379810              0.20699529                  99.89974
comp 16    3.514459              0.05437121                  99.95411
comp 17    2.966224              0.04588962                 100.00000
```

Significant components = ?

The projection of individuals



The map of variables

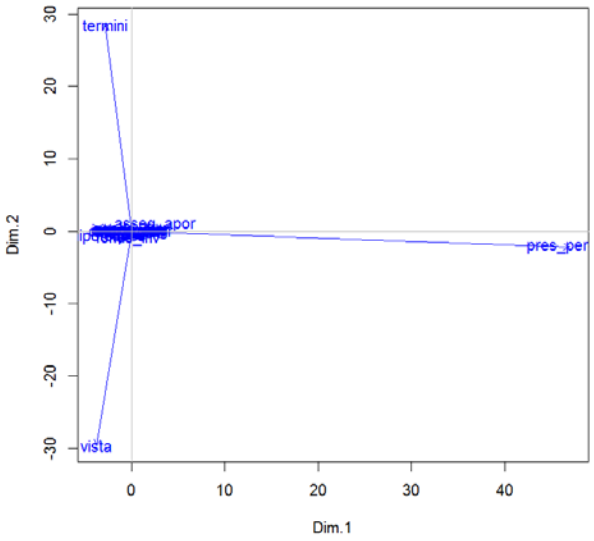


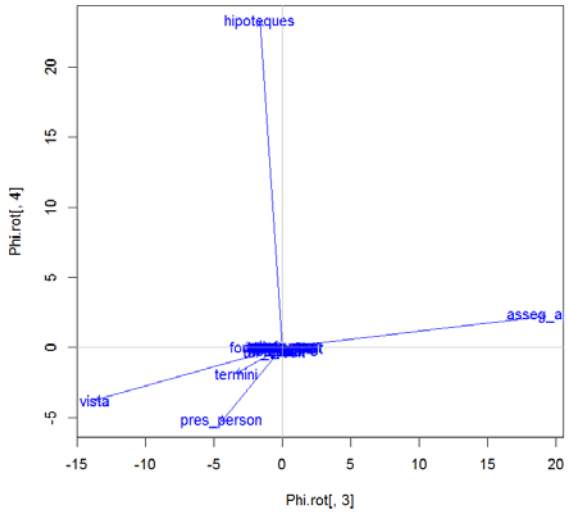
Notice that variables have distinct importance

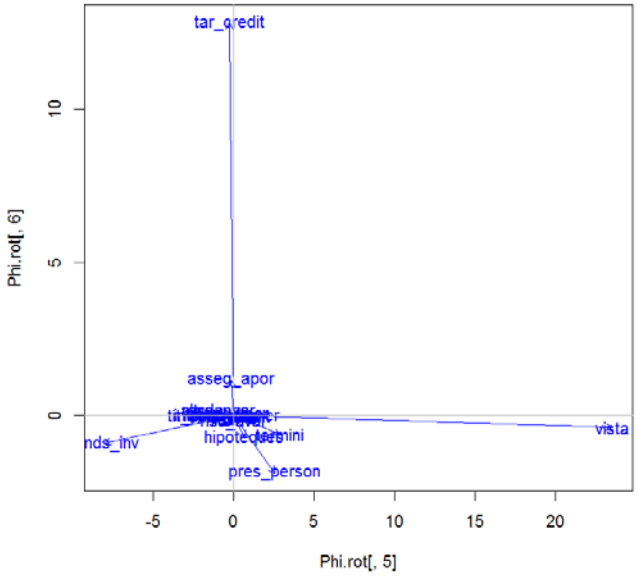
Finding the latent factors

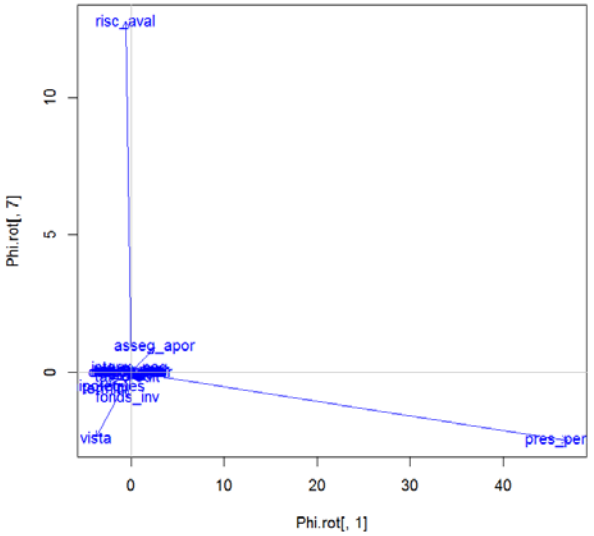
```
> pc.rot = varimax(pc$var$cor[,1:nd])
> pc.rot
$loadings

Loadings:
           Dim.1  Dim.2  Dim.3  Dim.4  Dim.5  Dim.6  Dim.7
vista             -0.726 -0.337         0.579
termini            0.976 -0.116
divises
asseg_apor  0.127         0.969  0.114
asseg_mer
fonds_inv                 -0.133        -0.658
tit_propis                               -0.253
interm_pag                               -0.184
renda_fixa                               -0.149
renda_var                                -0.165
altre_pas                                -0.184
hipoteques -0.102                 0.991
pres_person 0.985                -0.110
tar_credit                                      0.996
cmp_credit
desc_comer
risc_aval                                              0.995
```

The rotated components









Statistical description of the latent factors

```
> for (k in 1:nd) {
+      print(paste("Descripción de la componente rotada",k));
+      print(condes(cbind(Psi.rot[,k],Xtot_act),1,proba=0.000001))}
[1] "Descripción de la componente rotada 1"
$quanti
           correlation      p.value
pres_person   0.98471273 0.000000e+00

$category
             Estimate       p.value
Vehicles    11.3471906  0.000000e+00
Mobiliari    7.6958938  0.000000e+00
contr_SI    41.3190209  0.000000e+00
Seg_vid_SI  32.3850910  0.000000e+00
Reste        6.9154232 7.898976e-204
Empreses     6.2345111  5.951844e-80
Targ_SI      7.6870708  1.392102e-55
CIM_SI      11.7041405  9.945224e-54
Vis_Mast_SI  6.8876614  3.533658e-50

[1] "Descripción de la componente rotada 2"
$quanti
           correlation       p.value
termini      0.9764770  0.000000e+00
passiu_total 0.3040313 2.592341e-223
vista       -0.7264974  0.000000e+00

$category
             Estimate      p.value
65:99_anys    14.98237 4.215381e-22

[1] "Descripción de la componente rotada 3"
$quanti
           correlation       p.value
asseg_apor   0.9689251  0.000000e+00
vista       -0.3365672 2.707002e-276

$category
             Estimate      p.value
dest_NA      3.311964 7.937384e-26
contr_NO     2.496780 7.937384e-26
```

```
[1] "Descripción de la componente rotada 4"
$quanti
           correlation      p.value
hipoteques   0.9906198 0.000000e+00
actiu_total  0.5664826 0.000000e+00

$category
             Estimate      p.value
Car_(8e+03,1e+07]  11.884110 1.117298e-59
Ab_(8e+03,1e+07]   11.726578 3.934611e-57
pub_NO              2.736445 6.137354e-28
contr_SI            2.512615 2.898931e-26
CIM_SI              3.507927 3.110697e-19
Seg_vid_SI          2.164689 1.445415e-17

[1] "Descripción de la componente rotada 5"
$quanti
           correlation       p.value
vista        0.57881922  0.000000e+00
fonds_inv   -0.65813750  0.000000e+00

$category
             Estimate      p.value
pub_SI       3.188917 1.555601e-34
of_1         5.230529 2.701930e-30
Seg_vid_SI   2.529604 1.103594e-21

[1] "Descripción de la componente rotada 6"
$quanti
           correlation      p.value
tar_credit   0.9962458 0.000000e+00

$category
             Estimate      p.value
Vis_Mast_SI  2.663263 1.658722e-97
Targ_SI      1.880440 7.403526e-44
BCN_ciutat   2.060913 4.109685e-23
Comç_(1e+03,1e+12]  2.050259 8.430559e-21
Tar_deb_SI   1.187835 4.158647e-20
Comç_(500,1e+03]    2.644283 1.543908e-17

[1] "Descripción de la componente rotada 7"
$quanti
           correlation      p.value
risc_aval    0.9945908 0.000000e+00

$category
             Estimate      p.value
CIM_SI       1.897942 7.74401e-19
```

APPLICATION 3: THE ZIP DECODING

3rd. PCA application
ZIP decoding problem

Normalized handwritten digits automatically scanned from envelopes by the U.S. Postal Service. The original scanned digits are binary and of different sizes and orientations; the images here have been deslanted and size normalized, resulting in 16 x 16 grayscale images (Le Cun et al., 1990).

The data are in two files, and each line consists of the digit id (0-9) followed by the 256 grayscale values.

There are 7291 training observations and 2007 test observations, distributed as follows:

	0	1	2	3	4	5	6	7	8	9	Total
Train	1194	1005	731	658	652	556	664	645	542	644	7291
Test	359	264	198	166	200	160	170	147	166	177	2007

or as proportions:

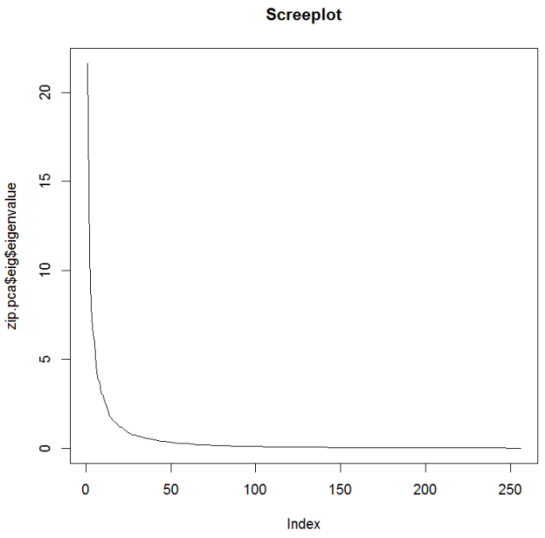
	0	1	2	3	4	5	6	7	8	9
Train	0.16	0.14	0.1	0.09	0.09	0.08	0.09	0.09	0.07	0.09
Test	0.18	0.13	0.1	0.08	0.10	0.08	0.08	0.07	0.08	0.09

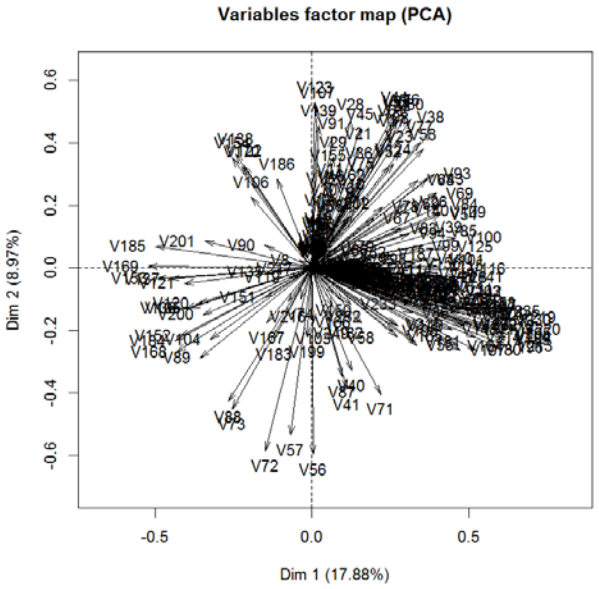
The purpose is to obtain a classification rule of the digits from the pixelized image of them. First we will perform an exploratory analysis by PCA + Clustering.

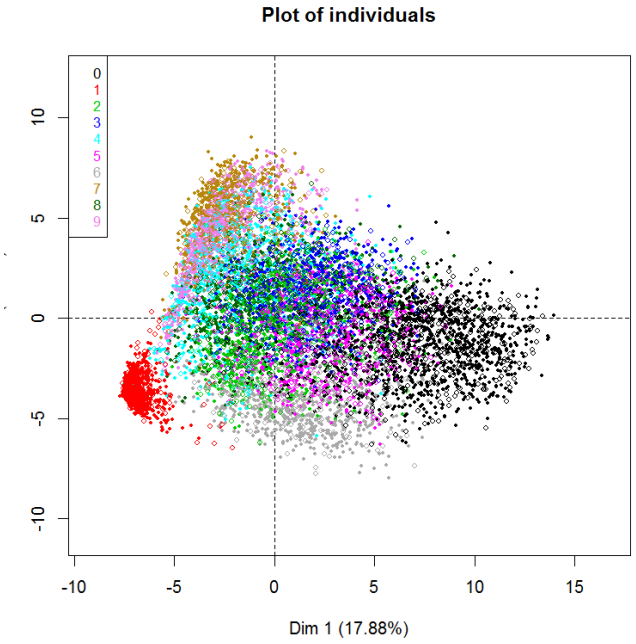
PCA of zip data

```
> head(zip.data)
  digit V1 V2 V3     V4     V5     V6     V7     V8     V9    V10    V11    V12   ....
1     6 -1 -1 -1 -1.000 -1.000 -1.000 -1.000 -0.631  0.862 -0.167 -1.000 -1.000   ....
2     5 -1 -1 -1 -0.813 -0.671 -0.809 -0.887 -0.671 -0.853 -1.000 -1.000 -0.774   ....
3     4 -1 -1 -1 -1.000 -1.000 -1.000 -1.000 -1.000 -1.000 -0.996  0.147  1.000   ....
4     7 -1 -1 -1 -1.000 -1.000 -0.273  0.684  0.960  0.450 -0.067 -0.679 -1.000   ....
5     3 -1 -1 -1 -1.000 -1.000 -0.928 -0.204  0.751  0.466  0.234 -0.809 -1.000   ....
6     6 -1 -1 -1 -1.000 -1.000 -0.397  0.983 -0.535 -1.000 -1.000 -1.000 -1.000   ....
```

```
zip.pca <- PCA(zip.data, quali.sup=1, ind.sup=c((N.train+1):N), scale.unit=F)
```







EXTENSIONS OF PCA: CORRESPONDENCE ANALYSIS

Multivariate exploratory analysis according the (active) data type

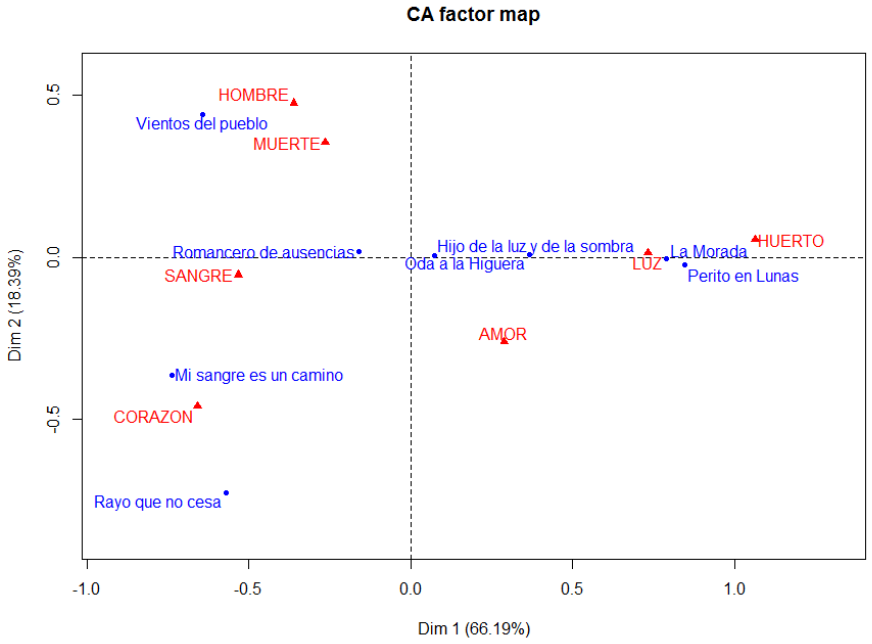
Type of (active) variables	Exploratory Analysis
Individual × continuous variables	Principal Component Analysis (PCA)
Count tables	Correspondence Analysis (CA)
Individual × categorical variables	Multiple Correspondence Analysis (MCA)

Correspondence Analysis

```
> poems_MH <- read.table("poemas_Miguel_Hernandez.txt",header=T,sep="\t")
> poems_MH
                                AMOR CORAZON HUERTO MUERTE SANGRE LUZ HOMBRE
La Morada                         41       3     32     21      8  52      5
Perito en Lunas                    4       1      3      3      1  12      0
Oda a la Higuera                  37       6     11     27     14  35      6
Rayo que no cesa                  17      26      0      8     12   1      1
Mi sangre es un camino             7      16      0      9     26   1      2
Vientos del pueblo                 3      23      2     61     35   3     22
Romancero de ausencias            44      20      2     38     25  19     19
Hijo de la luz y de la sombra     14      11      2     15     13  25      8
```

Miguel Hernandez, 1910-1942

```
> library(FactoMineR)
> ca <- CA(poems_MH)
```



Conditions to apply CA:

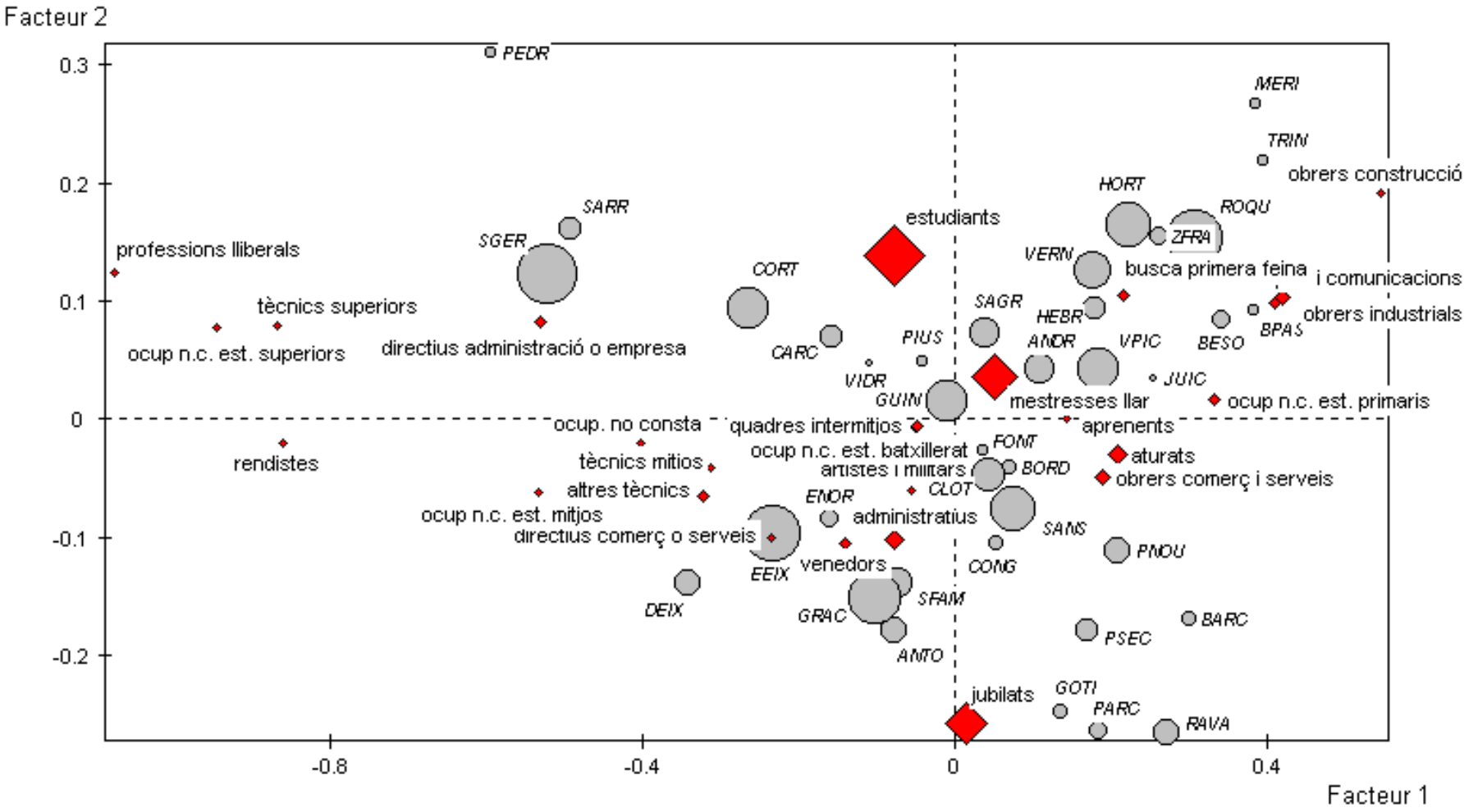
- All cells should have positive numbers
- It makes sense adding rows and colums.
- It makes sense to take this sums as weights for rows and columns.
- It makes sense to compare rows from the row-profiles and columns from the column-profiles.
- It makes sense the Khi-square metric either for row and column profiles comparison.

Barcelona census data

Barri	POBL	DIAD	DICO	PPLL	TESU	TEMI	ALTE	QUIN	ADMI	VENE	ARFF	OBCM	OBTR
BARC	17896	2190	7051	4295	2087	601	323	887	267	195	0	104	4734
PARC	23915	2053	8795	5502	2792	912	546	1760	746	808	1	134	6578
GOTI	17444	1581	5328	4237	2262	649	422	1438	663	864	0	82	5099
RAVA	42009	4738	16013	10509	4541	1348	851	2223	915	869	2	245	10584
ANTO	42159	2521	10497	9476	6661	1628	1259	5076	2502	2538	1	188	13671
EEIX	112174	6692	23186	21646	16290	3923	3363	15868	8644	11097	1465	510	36654
DEIX	47717	2350	8313	8348	6730	1562	1397	7751	4340	6924	2	219	16044
ENOR	31312	2075	7219	6206	4886	1156	980	4229	2308	2248	5	131	10381
SFAM	55150	3709	14216	12156	8524	2209	1635	6453	3271	2976	1	248	17938
PSEC	38770	3333	13414	9664	5251	1537	959	2751	1067	792	2	174	11210
JUIC	1715	297	669	342	178	52	24	100	22	31	0	9	397
ZFRA	28637	3579	8676	7254	3978	1491	995	1790	523	351	0	206	8385
FONT	9637	914	2683	2115	1403	404	279	987	492	357	3	55	3193
BORD	19597	1972	5365	4730	2889	915	664	1716	806	540	0	73	6758
SANS	84894	7555	25192	20299	11822	3808	2623	7395	3336	2859	5	413	27209
CORT	75083	5819	16069	14403	11001	2616	2123	10389	5493	7161	9	332	26232
PEDR	14585	828	1912	1893	2099	301	330	3134	1368	2720	0	76	4740
SGER	113900	6006	16101	15557	15763	2741	2846	23953	10722	20202	9	539	38227
SARR	35996	2068	5816	5212	4702	991	953	6686	3317	6240	11	184	12065
VIDR	1982	220	467	420	229	78	54	213	107	194	0	12	626
GRAC	98590	6249	25248	20810	14545	3719	2799	11969	6362	6837	52	455	32083
CARC	35015	2749	8146	6744	5455	1260	994	4574	2432	2653	8	188	11732
GUIN	77524	5917	19648	17127	12566	3152	2225	8904	4177	3698	110	569	25712
HORT	81739	8209	26822	19497	12580	3572	2026	5381	2070	1568	14	472	24443
HEBR	32759	3165	10344	7536	4701	1582	920	2667	1104	729	11	224	10459
VPIC	76048	5993	24223	18190	11271	3599	2344	6499	2337	1580	12	487	23245
ROQU	105882	2162	37680	25789	15303	4637	2417	5338	1633	921	2	722	29521
MERI	15298	2240	5166	4222	2089	680	280	483	94	44	0	96	3750
SAGR	55807	5045	14264	13392	9015	2637	1764	5442	2596	1652	0	285	18579
CONG	16160	959	4354	4028	2390	699	531	1853	788	558	0	89	4969
ANDR	51947	4697	15034	13220	7529	2465	1548	4182	2023	1249	0	225	16661

(1986)

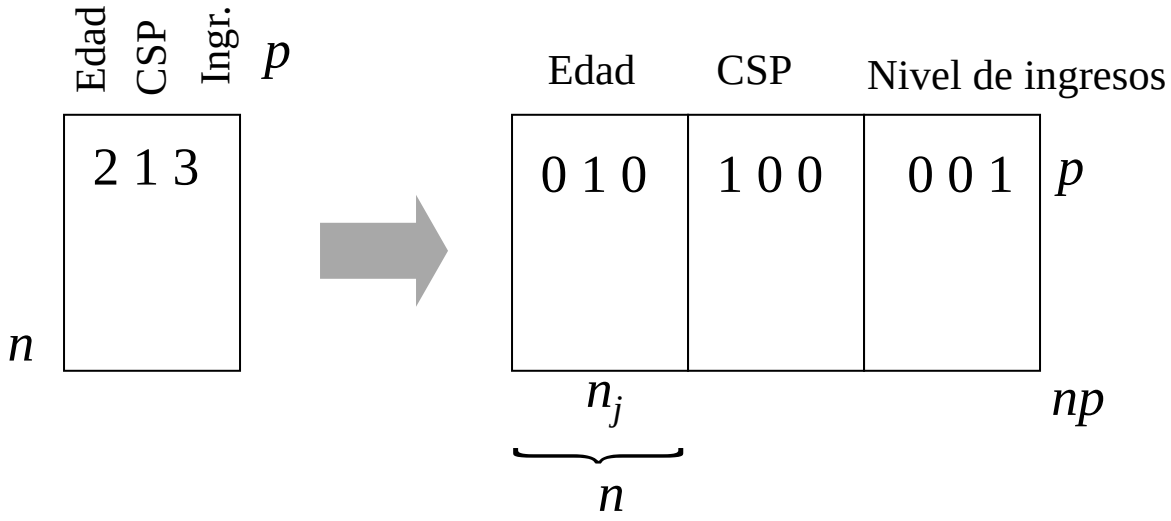
Correspondence Analysis

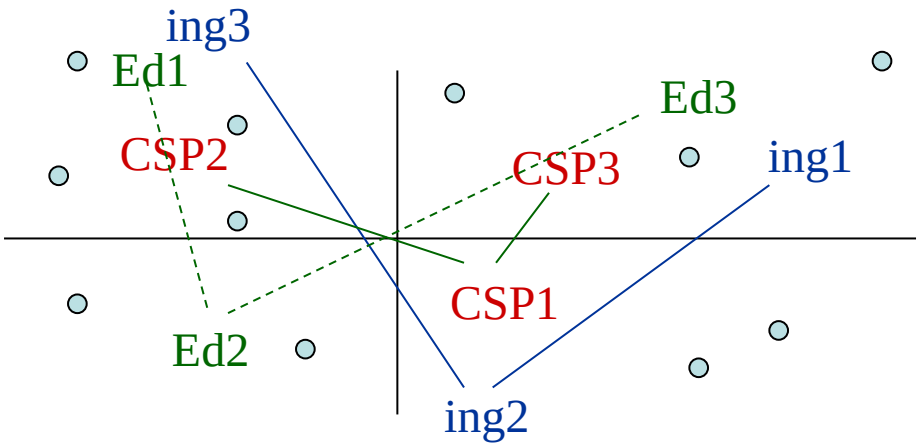


EXTENSIONS OF PCA: MULTIPLE CORRESPONDENCE ANALYSIS

Multiple Correspondence Analysis

Tables of the form individuals × *categorical* variables





Opinion survey. Active variables

```
MULTIPLE CORRESPONDENCE ANALYSIS
ELIMINATION OF ACTIVE CATEGORIES WITH SMALL WEIGHTS
THRESHOLD (PCMIN)  :      2.00 %       WEIGHT:       6.30
BEFORE CLEANING    :      8 ACTIVE QUESTIONS        36 ASSOCIATE CATEGORIES
AFTER CLEANING     :      8 ACTIVE QUESTIONS        31 ASSOCIATE CATEGORIES
TOTAL WEIGHT OF ACTIVE CASES :    315.00
MARGINAL DISTRIBUTIONS OF ACTIVE QUESTIONS
```

CATEGORIES IDENT	LABEL	BEFORE CLEANING COUNT	BEFORE CLEANING WEIGHT	AFTER CLEANING COUNT	AFTER CLEANING WEIGHT	HISTOGRAM OF RELATIVE WEIGHTS,
	8 . The family is the only place where you feel well					
Fam1	Yes	230	230.00	232	232.00	**********
Fam1	No	83	83.00	83	83.00	**********
Fam1	I do not know	2	2.00	==RAND.ASSIGN.==		
	9 . Opinion about marriage					
Mar1	indissoluble	81	81.00	81	81.00	**********
Mar2	dissolved serious pb	108	108.00	108	108.00	**********
Mar3	dissolved if agreem	114	114.00	114	114.00	**********
Mar4	I do not know	12	12.00	12	12.00	***
	12 . Are you satisfied of your daily life					
Cad1	a lot	85	85.00	85	85.00	**********
Cad2	enough	180	180.00	180	180.00	**********
Cad3	a little	37	37.00	37	37.00	********
Cad4	not at all	13	13.00	13	13.00	***
	30 . Your opinion on the evolution of the daily personal life					
Prs1	a lot better	17	17.00	18	18.00	****
Prs2	a little better	71	71.00	71	71.00	**********
Prs3	it is the same	87	87.00	88	88.00	**********
Prs4	a little worse	86	86.00	86	86.00	**********
Prs5	a lot worse	51	51.00	52	52.00	**********
Prs6	I do not know	3	3.00	==RAND.ASSIGN.==		
	32 . Your opinion on the life conditions in the future					
Ftr1	improving a lot	21	21.00	21	21.00	*****
Ftr2	improving a little	69	69.00	69	69.00	**********
Ftr3	the same	95	95.00	95	95.00	**********
Ftr4	a little worse	88	88.00	88	88.00	**********
Ftr5	a lot worse	29	29.00	29	29.00	******
Ftr6	I do not know	13	13.00	13	13.00	***

Active variables (cont.)

34 . Do you think the society needs to change					
Soc1 - yes	218	218.00	218	218.00	**********
Soc1 - no	68	68.00	68	68.00	**********
Soc1 - I do not know	29	29.00	29	29.00	******
35 . The computer science diffusion is...					
Inf1 - desirable	109	109.00	110	110.00	**********
Inf2 - inevitable	170	170.00	173	173.00	**********
Inf3 - dangerous	30	30.00	32	32.00	*******
Inf4 - I do not know	6	6.00	==RAND.ASSIGN.==		
48 . Your opinion on the justice running in 1986					
Jus1 - very well	5	5.00	==RAND.ASSIGN.==		
Jus2 - quite well	99	99.00	103	103.00	**********
Jus3 - quite bad	113	113.00	115	115.00	**********
Jus4 - very bad	81	81.00	82	82.00	**********
Jus5 - I do not know	12	12.00	15	15.00	***
Jus6 - do not answer	5	5.00	==RAND.ASSIGN.==		

MCA. Eigenvalues plot

EIGENVALUES
COMPUTATIONS PRECISION SUMMARY : TRACE BEFORE DIAGONALISATION.. 2.8750
SUM OF EIGENVALUES..... 2.8750
HISTOGRAM OF THE FIRST 23 EIGENVALUES

NUMBER	EIGENVALUE	PERCENTAGE	CUMULATED PERCENTAGE	
1	0.2253	7.84	7.84	**********
2	0.2047	7.12	14.96	*********
3	0.1801	6.27	21.22	********
4	0.1586	5.52	26.74	*******
5	0.1536	5.34	32.08	*******
6	0.1467	5.10	37.18	*******
7	0.1453	5.06	42.24	*******
8	0.1368	4.76	47.00	******
9	0.1321	4.59	51.59	******
10	0.1301	4.53	56.12	******
11	0.1232	4.29	60.40	******
12	0.1183	4.11	64.52	*****
13	0.1169	4.07	68.58	*****
14	0.1106	3.85	72.43	*****
15	0.1085	3.77	76.20	*****
16	0.1050	3.65	79.85	*****
17	0.0994	3.46	83.31	****
18	0.0977	3.40	86.71	****
19	0.0914	3.18	89.89	****
20	0.0827	2.88	92.77	****
21	0.0797	2.77	95.54	****
22	0.0647	2.25	97.79	***
23	0.0636	2.21	100.00	***

Projection of active modalities

LOADINGS, CONTRIBUTIONS AND SQUARED COSINES OF ACTIVE CATEGORIES

CATEGORIES			LOADINGS					CONTRIBUTIONS					SQUARED COSINES				
IDEN - LABEL	REL. WT.	DISTO	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
8 . The family is the only place where you feel well																	
Fam1 - Yes	9.21	0.36	0.14	-0.30	0.13	0.10	-0.08	0.9	4.1	0.9	0.6	0.4	0.06	0.26	0.05	0.03	0.02
Fam1 - No	3.29	2.80	-0.41	0.85	-0.38	-0.27	0.23	2.4	11.6	2.6	1.6	1.1	0.06	0.26	0.05	0.03	0.02
						CUMULATED CONTRIBUTION =		3.3	15.7	3.5	2.1	1.5					
9 . Opinion about marriage																	
Mar1 - indissoluble	3.21	2.89	-0.01	-0.59	0.69	0.39	-0.20	0.0	5.4	8.5	3.1	0.9	0.00	0.12	0.17	0.05	0.01
Mar2 - dissolved serious pb	4.29	1.92	-0.17	-0.11	-0.25	-0.23	-0.14	0.5	0.3	1.5	1.4	0.6	0.01	0.01	0.03	0.03	0.01
Mar3 - dissolved if agreem	4.52	1.76	-0.10	0.57	-0.21	0.10	0.21	0.2	7.1	1.2	0.3	1.4	0.01	0.18	0.03	0.01	0.03
Mar4 - I do not know	0.48	25.25	2.55	-0.42	-0.34	-1.60	0.62	13.7	0.4	0.3	7.7	1.2	0.26	0.01	0.00	0.10	0.02
						CUMULATED CONTRIBUTION =		14.4	13.2	11.5	12.5	4.0					
12 . Are you satisfied of your daily life																	
Cad1 - a lot	3.37	2.71	0.09	-0.28	0.33	-0.72	0.46	0.1	1.3	2.1	10.9	4.6	0.00	0.03	0.04	0.19	0.08
Cad2 - enough	7.14	0.75	-0.15	-0.04	-0.17	0.25	0.00	0.7	0.1	1.2	2.7	0.0	0.03	0.00	0.04	0.08	0.00
Cad3 - a little	1.47	7.51	0.21	0.59	0.29	-0.32	-0.91	0.3	2.5	0.7	0.9	7.8	0.01	0.05	0.01	0.01	0.11
Cad4 - not at all	0.52	23.23	0.86	0.71	-0.57	2.16	-0.38	1.7	1.3	0.9	15.2	0.5	0.03	0.02	0.01	0.20	0.01
						CUMULATED CONTRIBUTION =		2.8	5.2	4.9	29.8	12.9					
30 . Your opinion on the evolution of the daily personal life																	
Prs1 - a lot better	0.71	16.50	-0.66	-0.22	-1.36	0.88	-0.47	1.4	0.2	7.3	3.5	1.0	0.03	0.00	0.11	0.05	0.01
Prs2 - a little better	2.82	3.44	-0.66	-0.20	-0.36	-0.36	-0.18	5.4	0.5	2.0	2.3	0.6	0.13	0.01	0.04	0.04	0.01
Prs3 - it is the same	3.49	2.58	-0.39	0.13	0.56	0.39	0.72	2.3	0.3	6.0	3.4	11.7	0.06	0.01	0.12	0.06	0.20
Prs4 - a little worse	3.41	2.66	0.38	-0.29	0.44	-0.43	-0.66	2.2	1.4	3.6	4.0	9.8	0.05	0.03	0.07	0.07	0.17
Prs5 - a lot worse	2.06	5.06	1.16	0.60	-0.70	0.24	0.29	12.3	3.7	5.7	0.7	1.1	0.26	0.07	0.10	0.01	0.02
						CUMULATED CONTRIBUTION =		23.6	6.1	24.5	13.8	24.2					
32 . Your opinion on the life conditions in the future																	
Ftr1 - improving a lot	0.83	14.00	-0.72	1.42	-0.20	-1.01	0.99	1.9	8.2	0.2	5.4	5.3	0.04	0.14	0.00	0.07	0.07
Ftr2 - improving a little	2.74	3.57	-0.72	0.10	-0.79	0.11	-0.61	6.2	0.1	9.6	0.2	6.5	0.14	0.00	0.18	0.00	0.10
Ftr3 - the same	3.77	2.32	-0.27	-0.49	0.43	0.59	0.60	1.2	4.4	3.9	8.3	8.9	0.03	0.10	0.08	0.15	0.16
Ftr4 - a little worse	3.49	2.58	0.54	0.32	0.57	-0.61	-0.41	4.5	1.7	6.3	8.2	3.9	0.11	0.04	0.13	0.14	0.07
Ftr5 - a lot worse	1.15	9.86	1.56	0.22	-0.87	0.76	-0.04	12.5	0.3	4.8	4.2	0.0	0.25	0.00	0.08	0.06	0.00
Ftr6 - I do not know	0.52	23.23	-0.19	-1.88	-0.55	-0.82	0.11	0.1	8.9	0.9	2.2	0.0	0.00	0.15	0.01	0.03	0.00
						CUMULATED CONTRIBUTION =		26.4	23.6	25.6	28.5	24.7					
34 . Do you think the society needs to change																	
Soc1 - yes	8.65	0.44	0.07	0.15	0.19	0.22	-0.16	0.2	0.9	1.7	2.7	1.5	0.01	0.05	0.08	0.11	0.06
Soc1 - no	2.70	3.63	-0.45	0.18	-0.26	-0.55	-0.04	2.5	0.4	1.0	5.2	0.0	0.06	0.01	0.02	0.08	0.00
Soc1 - I do not know	1.15	9.86	0.57	-1.55	-0.82	-0.39	1.33	1.7	13.5	4.3	1.1	13.2	0.03	0.24	0.07	0.02	0.18
						CUMULATED CONTRIBUTION =		4.3	14.9	7.1	9.0	14.8					
35 . The computer science diffusion is...																	
Inf1 - desirable	4.37	1.86	-0.33	0.44	0.41	-0.14	0.46	2.1	4.2	4.1	0.6	5.9	0.06	0.11	0.09	0.01	0.11
Inf2 - inevitable	6.87	0.82	-0.11	-0.33	-0.30	0.02	-0.36	0.4	3.6	3.4	0.0	5.7	0.01	0.13	0.11	0.00	0.15
Inf3 - dangerous	1.27	8.84	1.74	0.25	0.20	0.39	0.36	17.0	0.4	0.3	1.2	1.1	0.34	0.01	0.00	0.02	0.01
						CUMULATED CONTRIBUTION =		19.5	8.3	7.7	1.8	12.7					
48 . Your opinion on the justice running in 1986																	
Jus2 - quite well	4.09	2.06	-0.35	-0.44	0.27	0.05	-0.16	2.3	3.9	1.6	0.1	0.7	0.06	0.09	0.03	0.00	0.01
Jus3 - quite bad	4.56	1.74	-0.05	0.19	-0.16	0.09	0.22	0.1	0.8	0.6	0.2	1.4	0.00	0.02	0.01	0.00	0.03
Jus4 - very bad	3.25	2.84	0.39	0.51	0.23	-0.06	-0.23	2.2	4.1	1.0	0.1	1.2	0.05	0.09	0.02	0.00	0.02
Jus5 - I do not know	0.60	20.00	0.69	-1.20	-1.90	-0.74	0.73	1.3	4.2	11.9	2.0	2.0	0.02	0.07	0.18	0.03	0.03
						CUMULATED CONTRIBUTION =		5.8	13.0	15.1	2.4	5.3					

Test values of supplementary modalities

CATEGORIES			TEST-VALUES					LOADINGS					
IDEN	LABEL	COUNT	ABS.WT	1	2	3	4	5	1	2	3	4	5	DISTO.
	2 . Urban area size (number of inhabitants)													
Agg1	less than 2000	84	84.00	3.6	-1.7	-1.4	-1.4	1.3	0.33	-0.16	-0.13	-0.13	0.12	2.75
Agg2	2001 to 5000	18	18.00	0.0	-1.7	0.9	-0.8	-0.4	0.00	-0.39	0.21	-0.19	-0.09	16.50
Agg3	5001 to 10000	18	18.00	-0.3	-0.2	0.8	-1.4	-1.4	-0.07	-0.05	0.17	-0.33	-0.33	16.50
Agg4	10001 to 20000	12	12.00	-0.9	-0.2	0.6	0.9	0.2	-0.25	-0.05	0.18	0.27	0.05	25.25
Agg5	20001 to 50000	23	23.00	-0.2	-0.2	0.6	1.6	0.7	-0.04	-0.04	0.12	0.33	0.15	12.70
Agg6	50001 to 100000	18	18.00	-0.3	0.1	0.1	-0.1	0.0	-0.06	0.02	0.02	-0.03	0.00	16.50
Agg7	100001 to 200000	28	28.00	-0.2	0.5	0.1	0.6	0.0	-0.03	0.08	0.02	0.12	-0.01	10.25
Agg8	more than 200000	68	68.00	-2.2	1.6	-0.1	-0.2	-0.2	-0.23	0.17	-0.01	-0.02	-0.02	3.63
Agg9	paris,paris.agglo	46	46.00	-0.8	1.3	-0.1	1.3	-0.8	-0.11	0.18	-0.02	0.18	-0.11	5.85
	3 . Sex of respondent													
Sex1	male	138	138.00	1.7	0.4	1.4	-0.1	-0.5	0.11	0.03	0.09	0.00	-0.03	1.28
Sex2	female	177	177.00	-1.7	-0.4	-1.4	0.1	0.5	-0.08	-0.02	-0.07	0.00	0.03	0.78
	5 . Current situation of the respondent													
Sit1	employed	172	172.00	-1.6	1.2	-1.0	-0.3	-1.6	-0.08	0.06	-0.05	-0.01	-0.08	0.83
Sit2	student	12	12.00	-2.5	2.7	-0.6	-2.0	1.2	-0.72	0.75	-0.16	-0.56	0.35	25.25
Sit3	housewife w/o prof.	52	52.00	1.7	-0.4	-0.2	0.6	-0.2	0.21	-0.05	-0.03	0.07	-0.02	5.06
Sit4	ill, disabled person	2	2.00	0.3	0.9	1.3	0.9	0.7	0.18	0.63	0.90	0.62	0.47	156.50
Sit5	retired people	54	54.00	1.5	-4.8	3.0	1.3	1.8	0.18	-0.60	0.38	0.16	0.22	4.83
Sit6	soldier	0	0.00	0.0	0.0	0.0	0.0	0.0	0.00	0.00	0.00	0.00	0.00	0.00
Sit7	unemployed person	23	23.00	0.3	3.0	-2.2	-1.0	-0.5	0.05	0.61	-0.43	-0.19	-0.09	12.70
	6 . Marital status													
Mat1	single	42	42.00	-2.3	2.5	-1.3	-0.3	2.1	-0.32	0.35	-0.19	-0.04	0.30	6.50
Mat2	married	223	223.00	1.6	-2.1	2.3	1.3	-3.9	0.06	-0.07	0.09	0.05	-0.14	0.41
Mat3	cohabitation	16	16.00	-0.8	2.6	-1.2	-1.5	1.0	-0.20	0.64	-0.29	-0.37	0.25	18.69
Mat4	separated, divorced	15	15.00	-0.2	1.3	-1.1	0.5	0.6	-0.05	0.32	-0.28	0.12	0.16	20.00
Mat5	widower/widow	19	19.00	1.2	-3.1	-0.5	-1.2	2.9	0.26	-0.69	-0.11	-0.27	0.64	15.58
	7 . Educational level of the respondent													
Dip1	no qualifications	54	54.00	3.5	-1.9	0.8	1.1	1.2	0.43	-0.23	0.09	0.14	0.15	4.83
Dip2	primary school	54	54.00	3.0	-1.7	-0.1	0.5	-0.1	0.38	-0.21	-0.02	0.06	-0.01	4.83
Dip3	vocational training	56	56.00	0.6	0.7	0.9	0.2	-0.7	0.07	0.08	0.11	0.02	-0.09	4.63
Dip4	GCSE diploma	28	28.00	-0.8	-0.3	-0.2	-0.8	-0.3	-0.15	-0.06	-0.04	-0.14	-0.05	10.25
Dip5	technical education	9	9.00	0.1	-0.3	-1.1	0.3	0.4	0.03	-0.09	-0.36	0.09	0.12	34.00
Dip6	technical and GCSE	17	17.00	-2.0	0.1	0.1	-1.2	0.4	-0.47	0.03	0.03	-0.29	0.10	17.53
Dip7	high school diploma	45	45.00	-2.9	-0.6	0.0	0.9	0.1	-0.40	-0.08	0.00	0.12	0.01	6.00
Dip8	more high school	43	43.00	-2.6	4.5	-1.0	-1.5	-0.5	-0.36	0.64	-0.15	-0.21	-0.08	6.33
Dip9	other	9	9.00	-0.9	-0.8	0.0	-0.2	-0.5	-0.29	-0.26	0.00	-0.07	-0.17	34.00

Test values of supplementary modalities

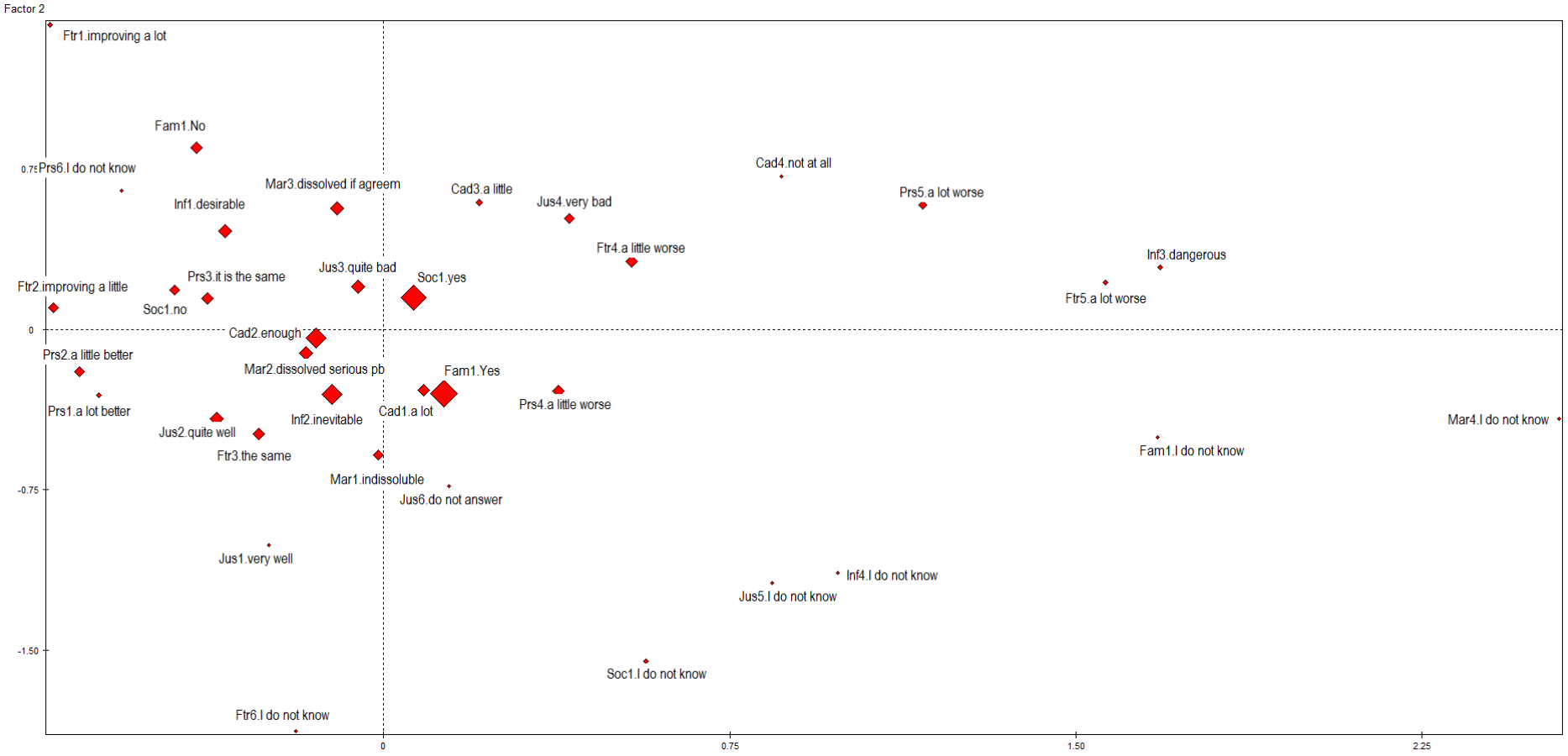
CATEGORIES			TEST-VALUES					LOADINGS					
IDEN - LABEL	COUNT	ABS.WT	1	2	3	4	5	1	2	3	4	5	DISTO.
10 . Housekeeping works, take care of children...													
Mén1 - only women do it	15	15.00	3.7	-2.0	-0.2	0.0	2.5	0.93	-0.51	-0.06	0.01	0.63	20.00
Mén2 - usually the women	85	85.00	-1.6	-2.6	-0.4	-0.6	0.5	-0.15	-0.24	-0.04	-0.05	0.04	2.71
Mén3 - men and women	214	214.00	-0.2	3.6	0.6	0.6	-1.6	-0.01	0.14	0.02	0.03	-0.06	0.47
Mén4 - I do not know	1	1.00	0.5	-1.7	-0.6	-0.9	0.3	0.46	-1.74	-0.64	-0.93	0.31	314.00
18 . At the moment, do you have a professional activity													
Trv1 - yes, full time	143	143.00	-1.3	2.0	-0.2	-0.5	-0.9	-0.08	0.13	-0.01	-0.03	-0.05	1.20
Trv2 - yes, part time	29	29.00	-1.0	-1.6	-1.2	0.5	-0.8	-0.18	-0.28	-0.22	0.09	-0.14	9.86
Trv3 - no	111	111.00	2.2	-1.1	0.8	0.6	0.3	0.17	-0.08	0.06	0.04	0.02	1.84
Trv4 - I have never worked	32	32.00	-0.5	-0.1	0.3	-0.6	1.7	-0.08	-0.01	0.04	-0.11	0.29	8.84
19 . Do you have work-personal life problems													
Cnf1 - yes	76	76.00	-1.9	2.3	-1.1	-0.1	-1.2	-0.19	0.23	-0.11	-0.01	-0.12	3.14
Cnf2 - no	97	97.00	-0.3	-0.8	0.1	-0.1	-0.2	-0.03	-0.07	0.01	-0.01	-0.02	2.25
19_ - missing category	142	142.00	2.0	-1.2	0.8	0.2	1.2	0.12	-0.07	0.05	0.01	0.08	1.22
27 . Do you have children													
Enf1 - yes	243	243.00	2.1	-2.3	1.1	-0.4	-2.6	0.06	-0.07	0.04	-0.01	-0.08	0.30
Enf2 - no	69	69.00	-2.2	2.6	-1.3	0.4	2.8	-0.24	0.28	-0.13	0.04	0.30	3.57
27_ - missing category	3	3.00	0.3	-1.2	0.4	-0.1	-0.5	0.18	-0.70	0.23	-0.08	-0.29	104.00
29 . Do you regularly impose restrictions													
Rst1 - yes	193	193.00	2.4	4.0	-0.3	-0.9	-2.0	0.11	0.18	-0.01	-0.04	-0.09	0.63
Rst2 - no	122	122.00	-2.4	-4.0	0.3	0.9	2.0	-0.17	-0.28	0.02	0.07	0.14	1.58

Correlation of factors of the supplentary continuous variables

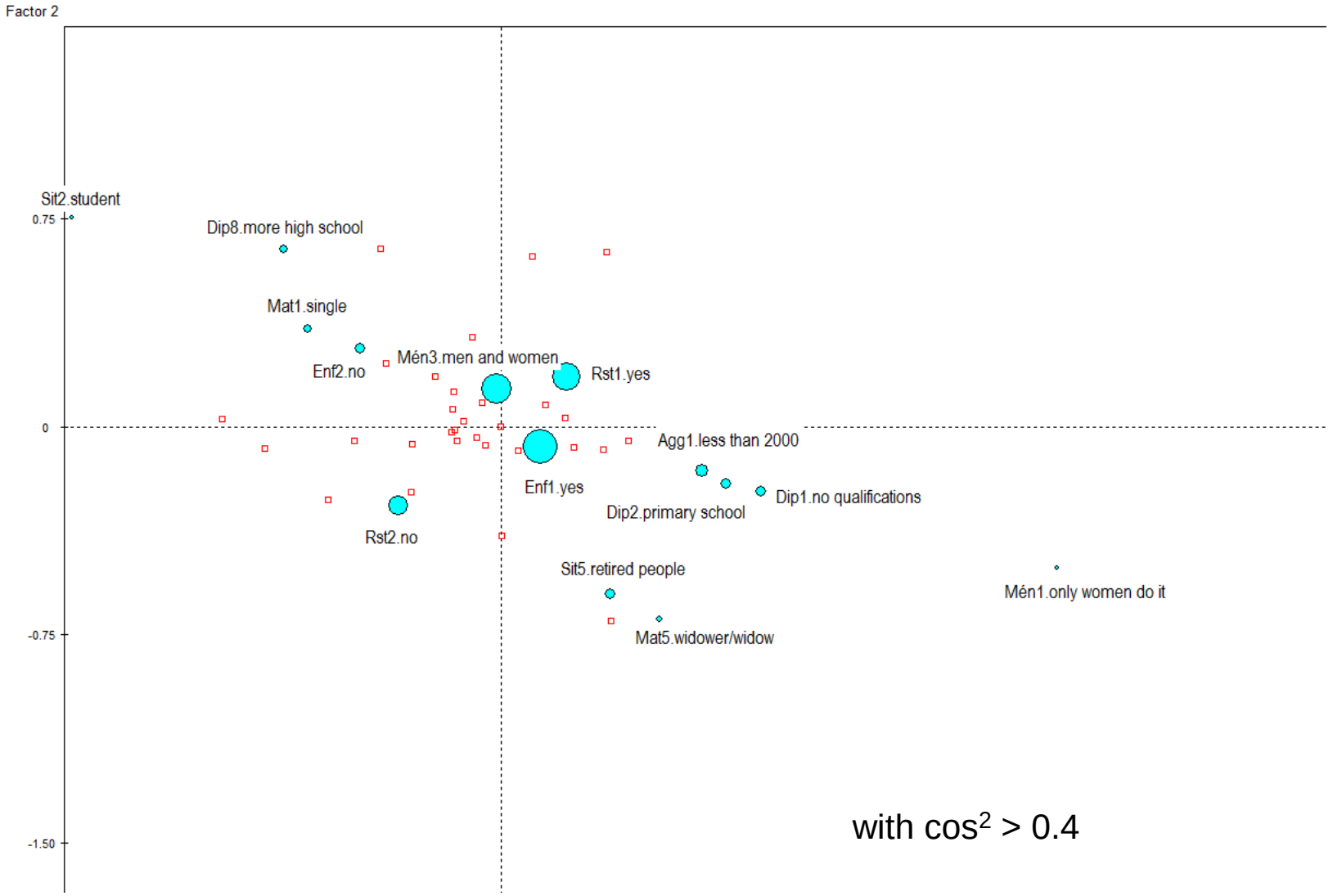
CORRELATIONS BETWEEN CONTINUOUS VARIABLES AND FACTORS
AXES 1 A 5

VARIABLES			SUMMARY STATISTICS				CORRELATIONS				
NUM	(IDEN)	SHORT LABEL	COUNT	ABS.WT	MEAN	ST.DEV.	1	2	3	4	5
4	(Age)	Age of respondent	315	315.00	43.76	16.58	0.21	-0.39	0.15	0.09	0.09
26	(Nbpr)	Number of persons in	315	315.00	3.06	1.41	0.04	0.11	-0.02	-0.09	-0.17
28	(Nbef)	Number of children	315	315.00	1.86	1.67	0.18	-0.15	0.07	-0.04	-0.12
41	(Fami)	Family, children : i	315	315.00	6.65	1.06	-0.06	-0.07	0.08	0.00	-0.03
42	(Trav)	Work, profession : i	315	315.00	5.96	1.54	-0.02	0.12	-0.02	0.01	-0.12
43	(Lois)	Free time, relax: im	315	315.00	5.30	1.45	-0.08	0.10	0.01	-0.10	0.01
44	(Amis)	Friends, acquaintanc	315	315.00	5.19	1.42	-0.11	-0.02	0.02	-0.12	0.06
46	(Reli)	Religion : importanc	315	315.00	3.24	2.02	-0.05	-0.29	0.09	0.06	0.01
50	(PrFm)	State benefits : ave	283	283.00	533.79	926.90	0.06	0.11	-0.01	-0.01	-0.16
51	(Salr)	Salary of the respon	267	267.00	4408.55	4575.34	-0.17	-0.01	-0.01	-0.09	0.01

Active modalities plot (acc. contribution)

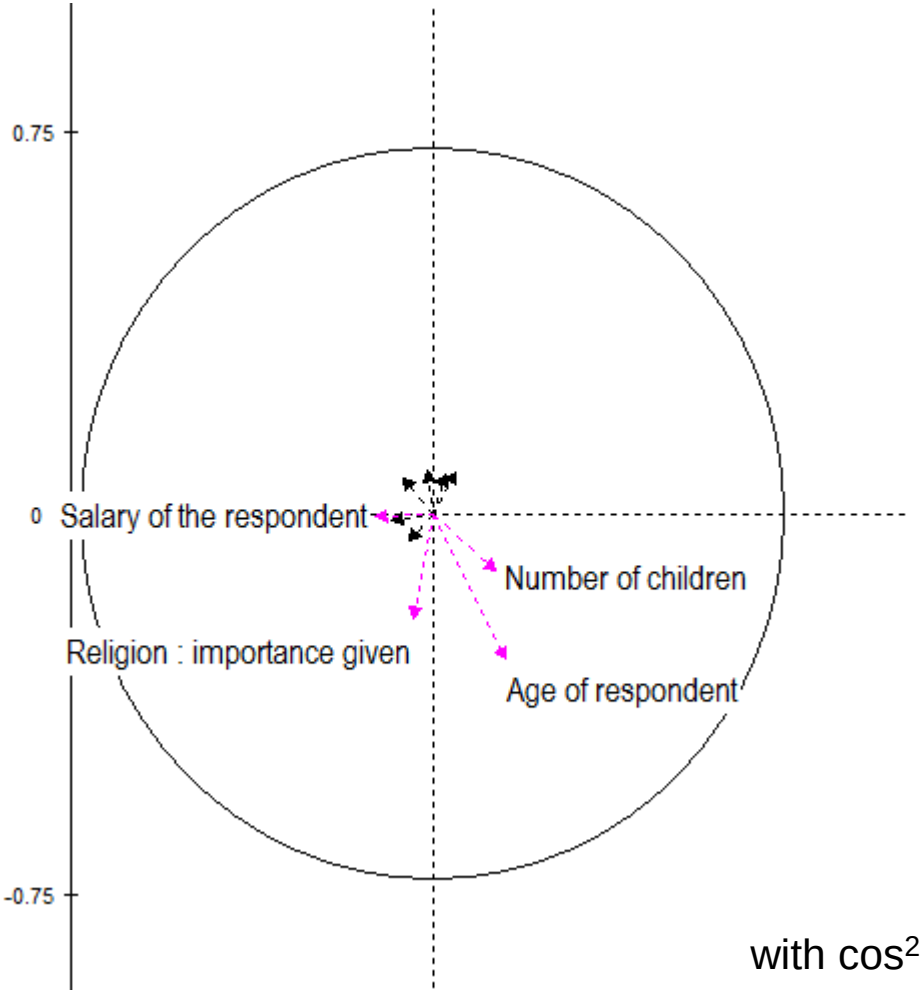


Supplementary categories



with $\cos^2 > 0.4$

Continuous variables. Plot of correlations



with $\cos^2 > 0.4$

References

- Aprender de los Datos: El Análisis de Componentes Principales. Tomàs Aluja, Alain Morineau. EUB, 1999.

- Exploratory Multivariate Analysis by example using R. François Husson, Sébastien Lê, Jérôme Pagès. CHAPMAN and HALL, 2011.