Tema 1

Aplicación en R

En R existen varias posibilidades de ejecutar un análisis de componentes principales, nosotros nos vamos a centrar en una de las opciones del paquete **ADE4** (en el tema de Análisis Factorial comentaremos los paquetes *prcomp* y *princomp*). La sintaxis es:

```
dudi.pca(df, row.w = rep(1, nrow(df))/nrow(df),col.w = rep(1, ncol(df)),
center = TRUE, scale = TRUE,scannf = TRUE, nf = 2)
```

donde:

- **df**: es un data frame con n filas (individuos) y p columnas (variables numéricas).
- **row.w**: es opcional y es el peso de las columnas (por defecto uniforme).
- **col.w**: es opcional y es el peso de las filas.
- **center**: es una valor lógico o numérico. Si es True, se centra por la media, si es False no se centra. Si es un vector numérico, la longitud debe ser igual al número de columnas.
- **scale**: es un valor lógico que indica si el vector de columnas debe ser normalizado por los pesos de row.w.
- **scannf**: valor lógico que indica si el screeplot será facilitado.
- **nf**: si scannf es False, nf es un entero que indica el número de ejes.

Los objetos del paquete **pca** son:

- **tab**: es el data frame analizado, dependiendo de la transformación de los datos.
- **cw**: pesos de las columnas.
- **lw**: pesos de las filas.
- **eig**: los autovalores.
- **rank**: rango de la matriz analizada.
- **nf**: numero de factores.
- **c1**: los valores de las columnas normalizados, por ejemplo los ejes principales.
- **l1**: valores de las filas normalizados.
- **co**: columna de las coordenadas
- **li**: fila de las coordenadas
- **call**: función call.
- **cent**: el vector p que contiene la media de las variables.
- **norm**: vector p que contiene las desviaciones de las variables.

1.1. Ejemplo de aplicación

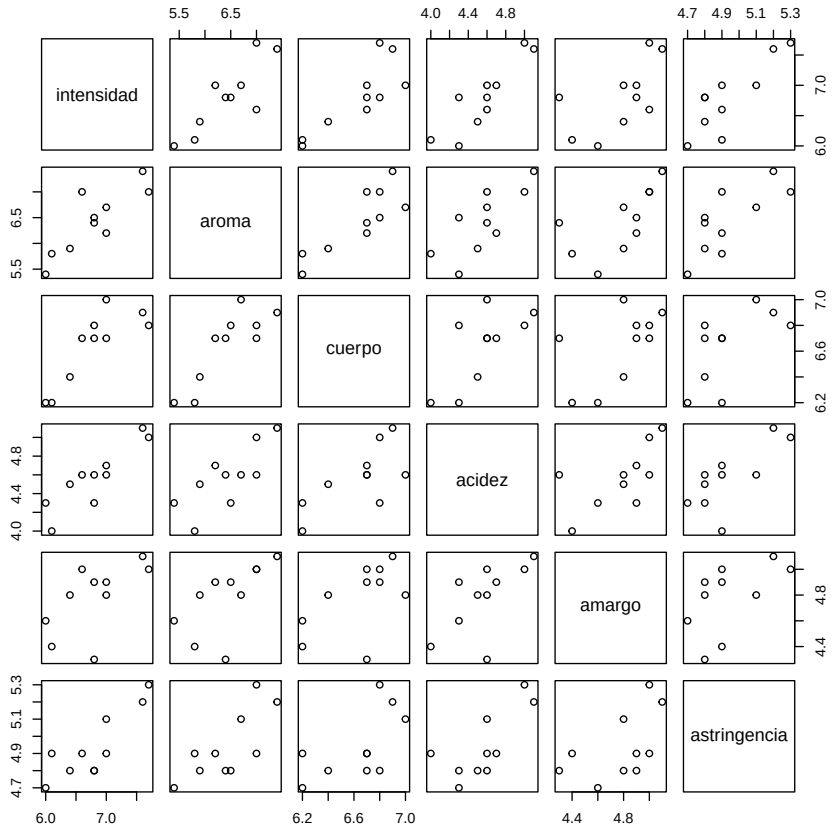
Vamos a utilizar el fichero **comprincipales.txt** que para diez tipos de cafe mide seis cualidades:

```
datos<-read.table("comprincipales.txt",header=T,row.names=1)
attach(datos)

## The following objects are masked from datos (pos = 3):
##
##     acidez, amargo, aroma,
##     astringencia, cuerpo, intensidad
## The following objects are masked from datos (pos = 6):
##
##     acidez, amargo, aroma,
##     astringencia, cuerpo, intensidad

datos

##     intensidad aroma cuerpo acidez amargo
## T1         7.7   7.0    6.8    5.0    5.0
## T2         6.0   5.4    6.2    4.3    4.6
## T3         6.4   5.9    6.4    4.5    4.8
## T4         6.8   6.4    6.7    4.6    4.3
## T5         7.0   6.2    6.7    4.7    4.9
## T6         7.6   7.4    6.9    5.1    5.1
## T7         6.1   5.8    6.2    4.0    4.4
## T8         6.8   6.5    6.8    4.3    4.9
## T9         6.6   7.0    6.7    4.6    5.0
## T10        7.0   6.7    7.0    4.6    4.8
##     astringencia
## T1           5.3
## T2           4.7
## T3           4.8
## T4           4.8
## T5           4.9
## T6           5.2
## T7           4.9
## T8           4.8
## T9           4.9
## T10          5.1
```

Evidentemente, antes de realizar cualquier análisis de componentes principales, podemos realizar un resumen estadístico o gráficos descriptivos bidimensionales, por ejemplo:

```
summary(datos)

##    intensidad       aroma
##  Min.   :6.00   Min.   :5.400
##  1st Qu.:6.45   1st Qu.:5.975
##  Median :6.80   Median :6.450
##  Mean   :6.80   Mean   :6.430
##  3rd Qu.:7.00   3rd Qu.:6.925
##  Max.   :7.70   Max.   :7.400
##      cuerpo          acidez
##  Min.   :6.200   Min.   :4.000
##  1st Qu.:6.475   1st Qu.:4.350
```

```
##  Median :6.700   Median :4.600
##  Mean   :6.640   Mean   :4.570
##  3rd Qu.:6.800   3rd Qu.:4.675
##  Max.   :7.000   Max.   :5.100
##      amargo       astringencia
##  Min.   :4.300   Min.   :4.70
##  1st Qu.:4.650   1st Qu.:4.80
##  Median :4.850   Median :4.90
##  Mean   :4.780   Mean   :4.94
##  3rd Qu.:4.975   3rd Qu.:5.05
##  Max.   :5.100   Max.   :5.30

plot(datos)
```



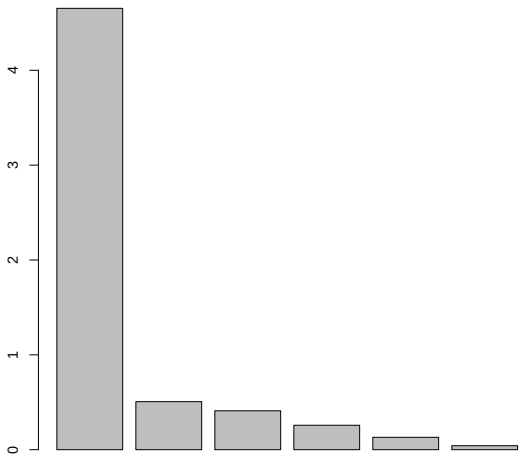
Del mismo modo, es interesante, estudiar la matriz de correlaciones, y ver, que estas sean en general altas, ya que esta es una de las hipótesis para el análisis de componentes principales. Para ello usamos la función *cor*

```
cor(datos)

##              intensidad     aroma    cuerpo
## intensidad    1.0000000 0.8454693 0.8315965
## aroma         0.8454693 1.0000000 0.8507676
## cuerpo        0.8315965 0.8507676 1.0000000
## acidez        0.8927236 0.7725890 0.6954213
## amargo        0.6228274 0.6578403 0.5649069
```

```
## astringencia  0.8468706 0.7588402 0.6177433
##                  acidez    amargo
## intensidad   0.8927236 0.6228274
## aroma        0.7725890 0.6578403
## cuerpo       0.6954213 0.5649069
## acidez       1.0000000 0.6446742
## amargo       0.6446742 1.0000000
## astringencia 0.7339586 0.5515843
##              astringencia
## intensidad      0.8468706
## aroma           0.7588402
## cuerpo          0.6177433
## acidez          0.7339586
## amargo          0.5515843
## astringencia    1.0000000
```

Como ya hemos comentado, la función con la que vamos a realizar el análisis de componentes principales, va a ser la función *dudi.pca*.

```
library(ade4)
acp<-dudi.pca(df=datos,scannf=T,nf=2)

## Select the number of axes:
```



De esta manera generaremos el análisis de componentes principales y a su vez obtenemos la representación de las gráfica de los autovalores, en la que podemos ver que el primero es, con mucha diferencia, el más importante, es decir, el que más contribuye a la explicación de las variables.

Para ver la importancia (contribuciones) absolutas y relativas, vamos a usar la función *inertia.dudi*, que calcula dichas contribuciones:

```
acpi<-inertia.dudi(acp, row.inertia=T, col.inertia=T)
acpi

## Inertia information:
```

```
## Call: inertia.dudi(x = acp, row.inertia = T, col.inertia = T)
##
## Decomposition of total inertia:
##     inertia     cum  cum(%)
## Ax1 4.65309   4.653   77.55
## Ax2 0.50650   5.160   85.99
## Ax3 0.41022   5.570   92.83
## Ax4 0.25765   5.827   97.12
## Ax5 0.13050   5.958   99.30
## Ax6 0.04205   6.000  100.00
##
## Row contributions (%):
##      T1      T2      T3      T4      T5
##  17.645  18.326   4.694   7.093   1.321
##      T6      T7      T8      T9     T10
##  20.691  18.846   3.221   3.244   4.919
##
## Row absolute contributions (%):
##        Axis1   Axis2
## T1  20.18380  1.8697
## T2  22.20579  4.4912
## T3   4.52692  9.0770
## T4   1.57280 49.2957
## T5   0.25566  2.9316
## T6  26.09941  0.2755
## T7  20.53950  7.2574
## T8   0.08928  5.6724
## T9   0.77690 13.5865
## T10  3.74995  5.5432
##
## Signed row relative contributions:
##       Axis1    Axis2
## T1   -88.71  -0.8945
## T2    93.97   2.0688
## T3    74.79  16.3245
## T4    17.20 -58.6724
## T5   -15.01  18.7331
## T6   -97.82   0.1124
## T7    84.52  -3.2507
## T8     2.15  14.8664
## T9   -18.57  35.3510
## T10  -59.12  -9.5132
##
## Cumulative sum of row relative contributions (%):
##       Axis1 Axis1:2 Axis3:6
## T1    88.71   89.60  10.396
## T2    93.97   96.04   3.961
## T3    74.79   91.12   8.881
## T4    17.20   75.87  24.130
## T5    15.01   33.74  66.258
## T6    97.82   97.93   2.066
## T7    84.52   87.77  12.231
## T8     2.15   17.02  82.984
## T9    18.57   53.92  46.079
```

```
## T10   59.12   68.64  31.363
##
## Column contributions (%):
##   intensidad        aroma       cuerpo
##        16.67        16.67        16.67
##       acidez       amargo astringencia
##        16.67        16.67        16.67
##
## Column absolute contributions (%):
##                Axis1    Axis2
## intensidad     19.81  4.99933
## aroma          18.54  0.46594
## cuerpo         16.19  3.49526
## acidez         17.47  0.04528
## amargo         12.21 83.77494
## astringencia   15.79  7.21926
##
## Signed column relative contributions:
##                Axis1    Axis2
## intensidad    -92.17 -2.53214
## aroma         -86.26 -0.23600
## cuerpo        -75.33 -1.77034
## acidez        -81.27 -0.02293
## amargo        -56.80 42.43173
## astringencia  -73.48 -3.65653
##
## Cumulative sum of column relative contributions (%):
##                Axis1 Axis1:2 Axis3:6
## intensidad     92.17   94.70   5.299
## aroma          86.26   86.50  13.504
## cuerpo         75.33   77.10  22.895
## acidez         81.27   81.29  18.709
## amargo         56.80   99.23   0.767
## astringencia   73.48   77.13  22.868
```

aquí podemos ver como el primer eje, explicará un 77,55 % de la inercia, y que el resto de componentes explican, respectivamente un 8.44 % (0.8599-0.7755 %), 6.83 % (0.9282-0.8599 %), 4.3 %, 2.16 % y 0.71 %.

A continuación vamos a ir analizando las salidas que nos proporciona R.

En primer lugar vemos los resultados para las filas. En este caso, obtendremos la representación de cada fila en el espacio bidimensional (normalizado y sin normalizar):

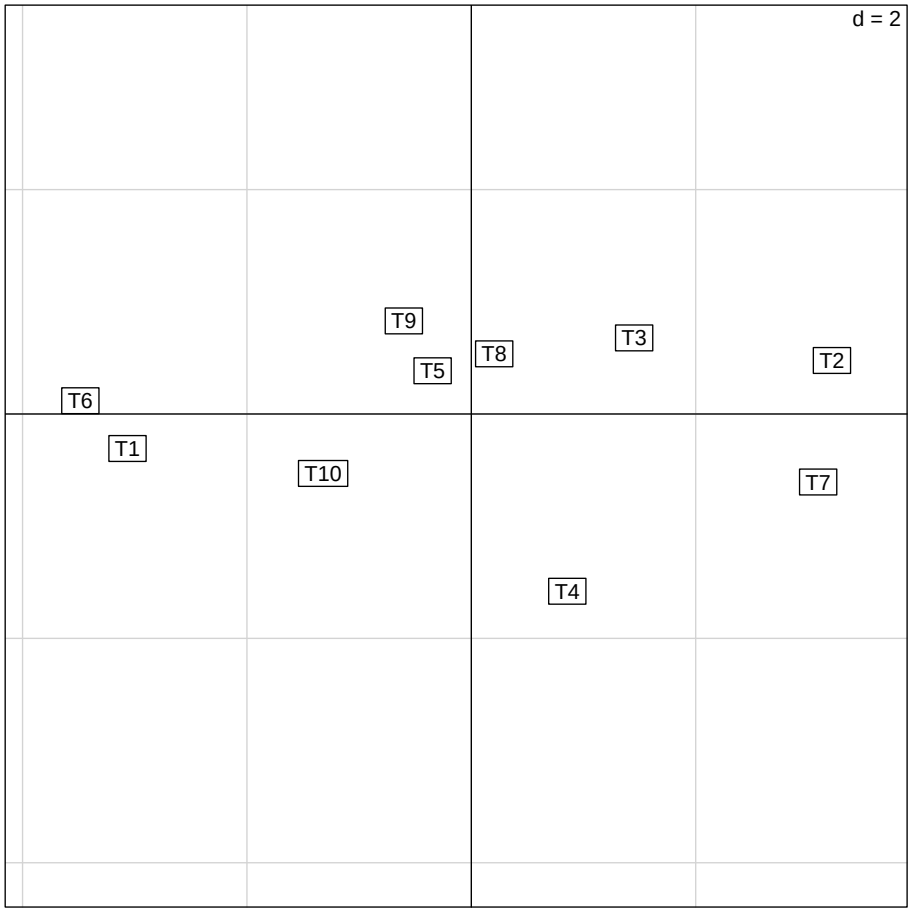
```
acp$l1
```

```
##            RS1        RS2
## T1  -1.4206971 -0.4324037
## T2   1.4901606  0.6701615
## T3   0.6728236  0.9527306
## T4   0.3965848 -2.2202626
## T5  -0.1598943  0.5414411
## T6  -1.6155313  0.1659967
## T7   1.4331609 -0.8519015
## T8   0.0944892  0.7531501
```

```
## T9  -0.2787284  1.1656112
## T10 -0.6123680 -0.7445234

acp$li

##          Axis1      Axis2
## T1  -3.0645876 -0.3077356
## T2   3.2144274  0.4769444
## T3   1.4513486  0.6780448
## T4   0.8554737 -1.5801293
## T5  -0.3449082  0.3853359
## T6  -3.4848647  0.1181375
## T7   3.0914732 -0.6062862
## T8   0.2038228  0.5360062
## T9  -0.6012454  0.8295489
## T10 -1.3209397 -0.5298667
```

Estos puntos los podemos representar con la orden **s.label** de la forma:

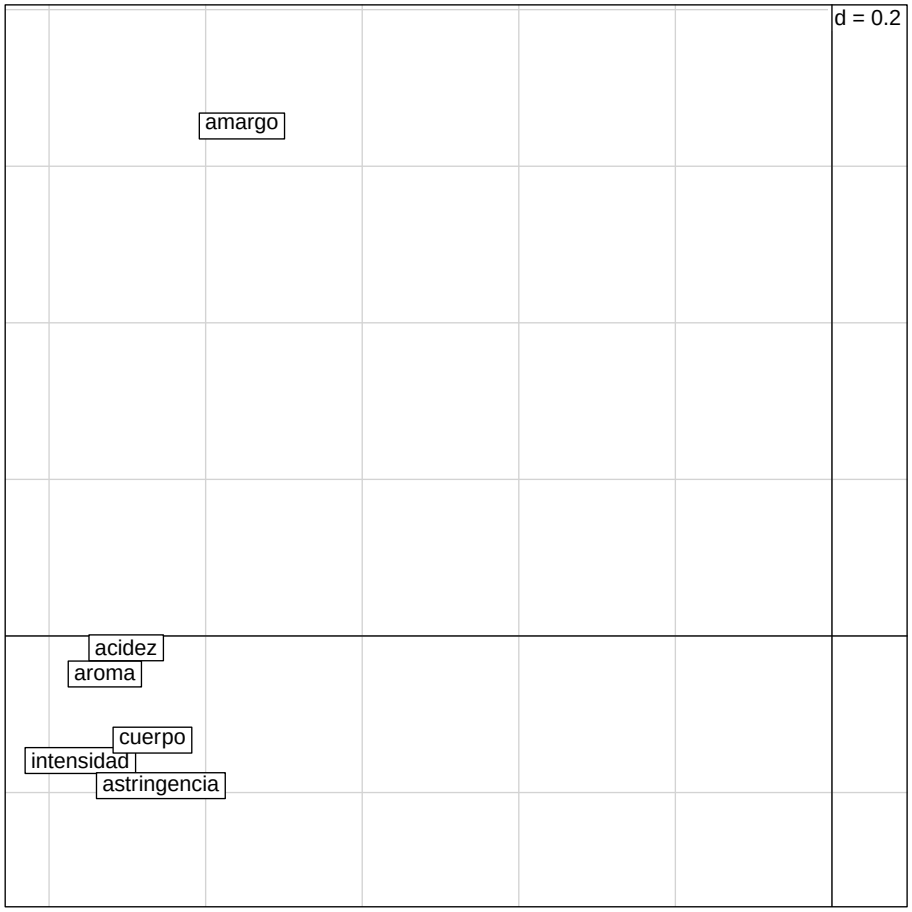
```
s.label(acp$li)
```



En donde podemos ver, como se agrupan los cafes Tipo6 y Tipo10 (a la izquierda); los Tipo9, Tipo5 y Tipo8 (en el centro) y el Tipo2 con el Tipo7 (en la derecha).

Del mismo modo la representación de las columnas será:

```
s.label(acp$co)
```



En este figura, observamos como la amargura se contrapone al resto de cualidades del cafe. Y las ayudas a la interpretación son:

```
acp$co
```

```
##                   Comp1       Comp2
## intensidad   -0.9600465 -0.15912706
## aroma        -0.9287631 -0.04857958
## cuerpo       -0.8679552 -0.13305403
## acidez       -0.9014864 -0.01514345
## amargo       -0.7536662  0.65139639
## astringencia -0.8571809 -0.19122055
```

```
acp$c1
```

```
##                     CS1         CS2
## intensidad   -0.4450633 -0.22359174
## aroma        -0.4305607 -0.06825988
## cuerpo       -0.4023711 -0.18695615
## acidez       -0.4179156 -0.02127828
## amargo       -0.3493884  0.91528650
## astringencia -0.3973763 -0.26868676
```

con estas ayudas, del mismo modo, podemos ver como con respecto a la primera componente, todas las cualidades toman valores similares, sin embargo la segunda componente, contrapone la amargura al resto de

cualidades.

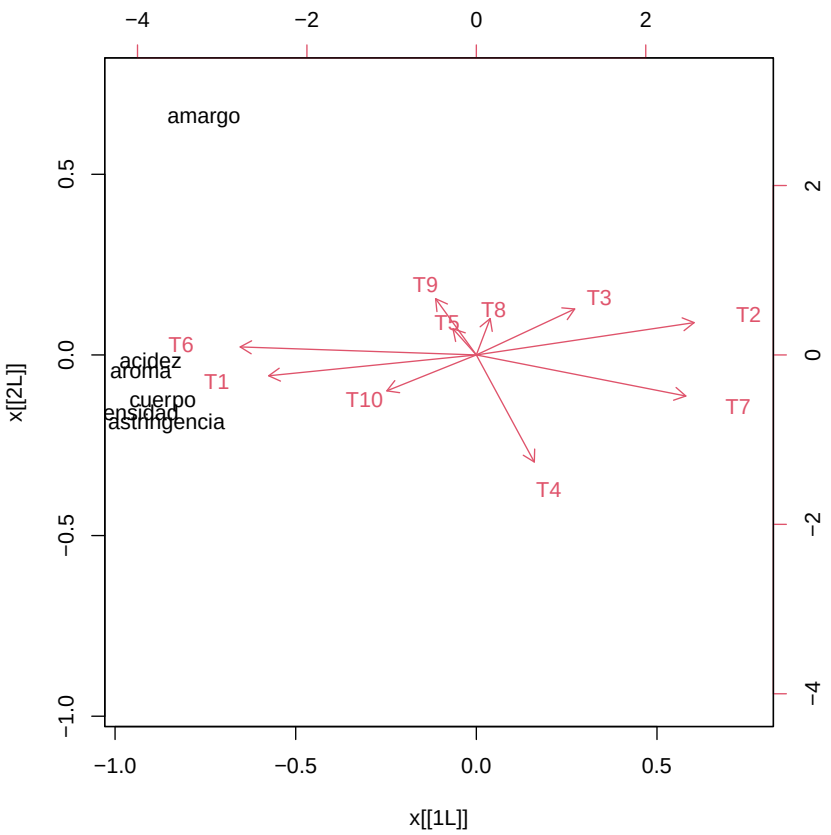
También podemos analizar las contribuciones a la inercia de la filas (o columnas) de modo que:

```
acpi

## Inertia information:
## Call: inertia.dudi(x = acp, row.inertia = T, col.inertia = T)
##
## Decomposition of total inertia:
##     inertia     cum  cum(%)
## Ax1 4.65309   4.653   77.55
## Ax2 0.50650   5.160   85.99
## Ax3 0.41022   5.570   92.83
## Ax4 0.25765   5.827   97.12
## Ax5 0.13050   5.958   99.30
## Ax6 0.04205   6.000  100.00
##
## Row contributions (%):
##      T1      T2      T3      T4      T5
##  17.645  18.326   4.694   7.093   1.321
##      T6      T7      T8      T9     T10
##  20.691  18.846   3.221   3.244   4.919
##
## Row absolute contributions (%):
##        Axis1   Axis2
## T1  20.18380  1.8697
## T2  22.20579  4.4912
## T3   4.52692  9.0770
## T4   1.57280 49.2957
## T5   0.25566  2.9316
## T6  26.09941  0.2755
## T7  20.53950  7.2574
## T8   0.08928  5.6724
## T9   0.77690 13.5865
## T10  3.74995  5.5432
##
## Signed row relative contributions:
##       Axis1    Axis2
## T1   -88.71  -0.8945
## T2    93.97   2.0688
## T3    74.79  16.3245
## T4    17.20 -58.6724
## T5   -15.01  18.7331
## T6   -97.82   0.1124
## T7    84.52  -3.2507
## T8     2.15  14.8664
## T9   -18.57  35.3510
## T10  -59.12  -9.5132
##
## Cumulative sum of row relative contributions (%):
##       Axis1 Axis1:2 Axis3:6
## T1    88.71   89.60  10.396
## T2    93.97   96.04   3.961
## T3    74.79   91.12   8.881
```

```
## T4    17.20   75.87  24.130
## T5    15.01   33.74  66.258
## T6    97.82   97.93   2.066
## T7    84.52   87.77  12.231
## T8     2.15   17.02  82.984
## T9    18.57   53.92  46.079
## T10   59.12   68.64  31.363
##
## Column contributions (%):
##   intensidad        aroma       cuerpo
##        16.67        16.67        16.67
##       acidez       amargo astringencia
##        16.67        16.67        16.67
##
## Column absolute contributions (%):
##                Axis1    Axis2
## intensidad     19.81  4.99933
## aroma          18.54  0.46594
## cuerpo         16.19  3.49526
## acidez         17.47  0.04528
## amargo         12.21 83.77494
## astringencia   15.79  7.21926
##
## Signed column relative contributions:
##                Axis1    Axis2
## intensidad    -92.17 -2.53214
## aroma         -86.26 -0.23600
## cuerpo        -75.33 -1.77034
## acidez        -81.27 -0.02293
## amargo        -56.80 42.43173
## astringencia  -73.48 -3.65653
##
## Cumulative sum of column relative contributions (%):
##                Axis1 Axis1:2 Axis3:6
## intensidad     92.17   94.70   5.299
## aroma          86.26   86.50  13.504
## cuerpo         75.33   77.10  22.895
## acidez         81.27   81.29  18.709
## amargo         56.80   99.23   0.767
## astringencia   73.48   77.13  22.868
```

También podemos obtener la representación conjunta de filas y columnas sin más que:

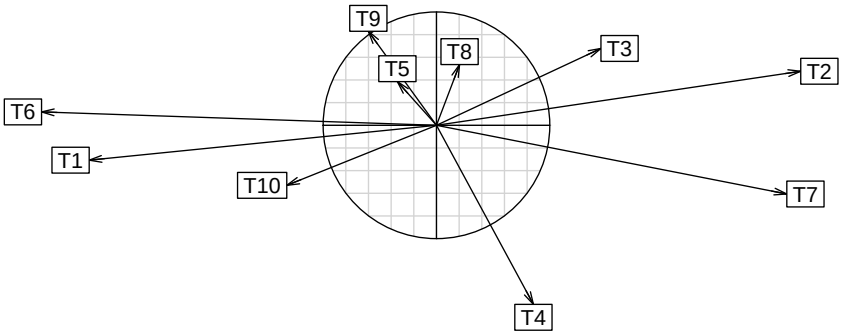
```
biplot(acp$co,acp$li)
```



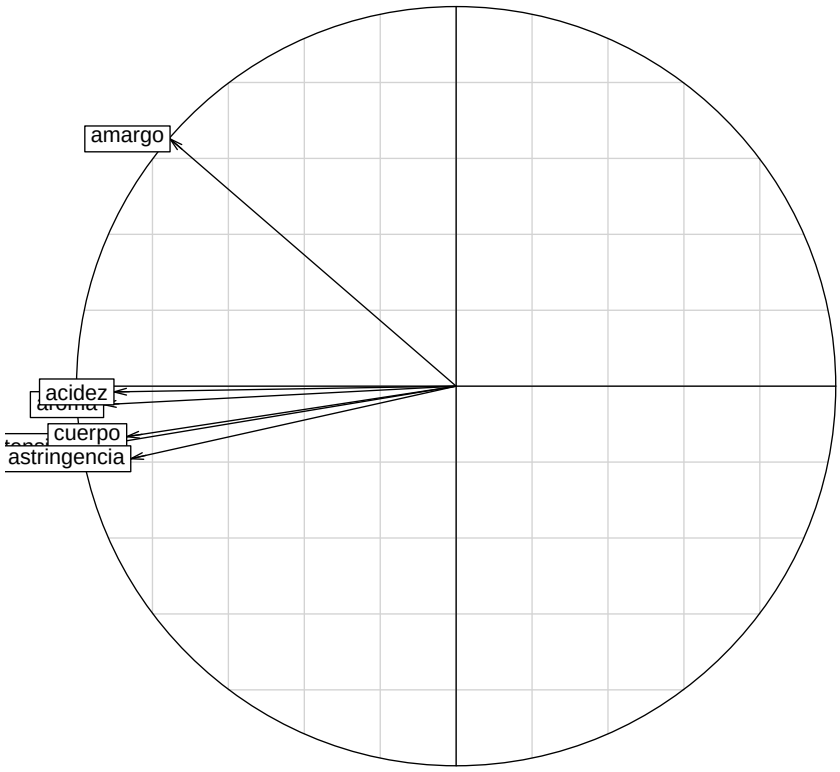
En este gráfico podremos sacar las conclusiones bidimensionales, como que el Tipo9, será cercano a amargo, o el Tipo10 a cuerpo, intensidad y astringencia.

Podemos obtener, finalmente, una representación de las correlaciones de las variables, con la orden

```
s.corcircle(acp$li)
```

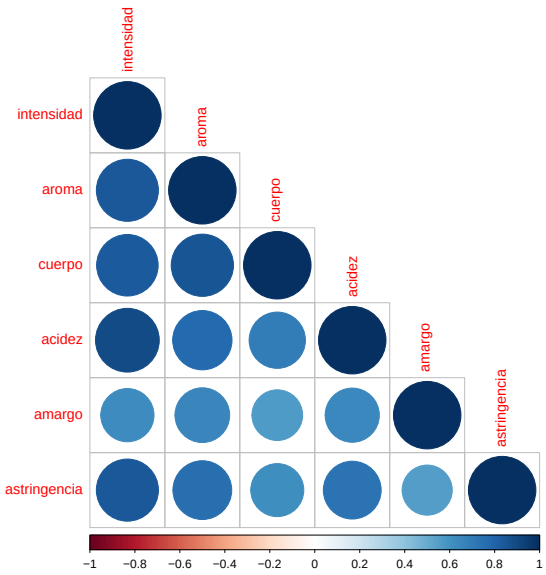



```
s.corcircle(acp$co)
```

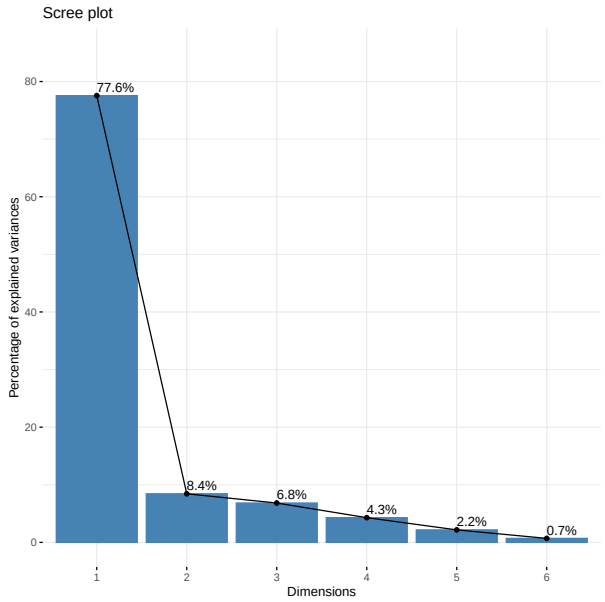


Se puede usar el paquete qqplot o factoextra para obtener unos gráficos de mejor calidad, por ejemplo

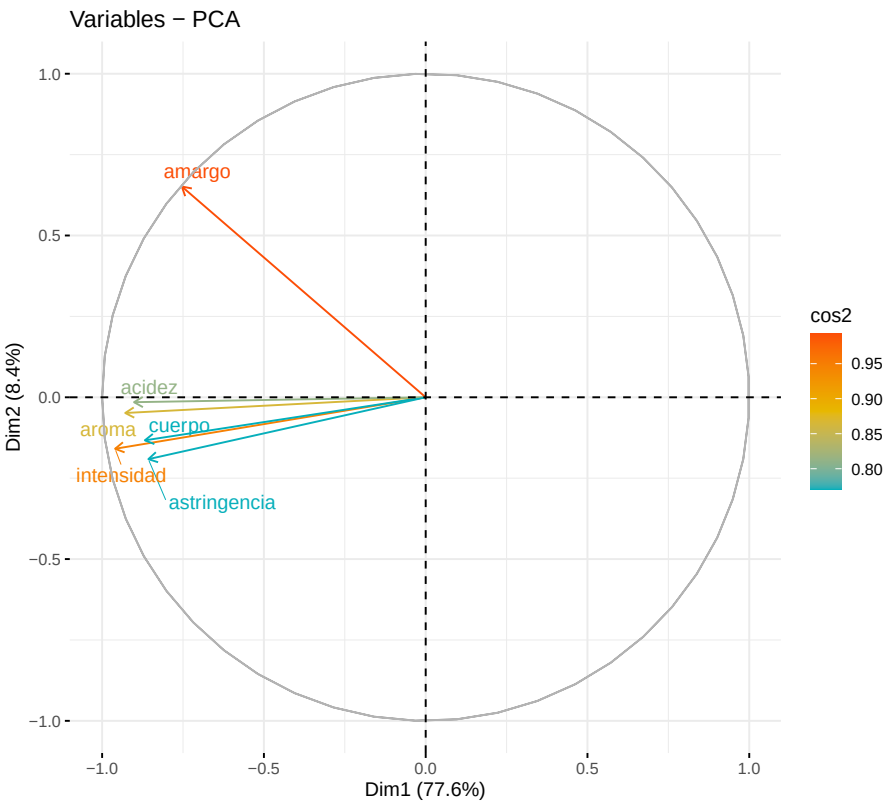
```
library(corrplot)#para ver la correlación de las variables

corrplot(cor(datos),sig.level=0.05,typ="lower")
library(factoextra)
```



```
fviz_eig(acp, addlabels = TRUE, ylim = c(0, 85))
```




```
fviz_pca_var(acp, col.var = "cos2",
gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
repel = TRUE # Avoid text overlapping
)
```


Variables − PCA
amargo
acidez
aroma
cuerpo
intensidad
astringencia
cos2
0.95
0.90
0.85
0.80
Dim1 (77.6%)
Dim2 (8.4%)
1.0
0.5
0.0
−0.5
−1.0