Ladies-BRC

Taller Análisis de componentes principales: FactoMineR

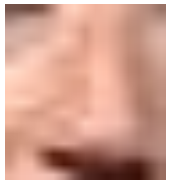
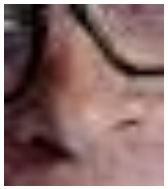
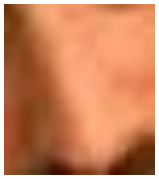
Dra. Patricia E. García

GESAP-INIBIOMA

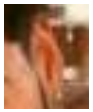
https://patagonianlimnology.weebly.com/

GESAP

¿Cómo funciona nuestro cerebro para reconocer caras?

Nariz como única variable


Las orejas como única variable

Algunos sistemas no pueden describirse como características individuales, sino que deben verse en conjunto



INTRODUCIÓN: Análisis de componentes principales (PCA)

Es una técnica **estadística no paramétrica** utilizada para describir un conjunto de datos.

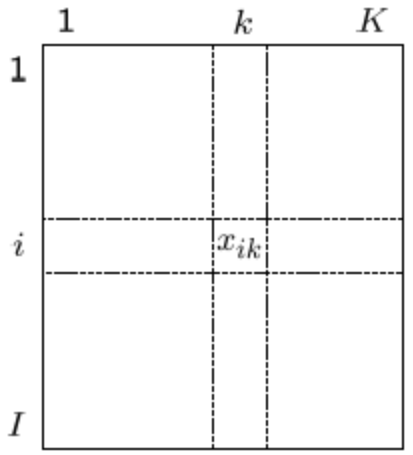
Es un método de reducción de la "dimensionalidad" de datos de una manera *no supervisada*.

Su objetivo es proyectar los datos en las direcciones de máxima varianza y así poder eliminar aquellas *direcciones o planos* que aporten menos información.

Es una técnica de gran versatilidad y muy ampliamente utilizada.

¿Qué tipos de datos utiliza?

Utiliza datos en donde las ***filas*** pueden ser considerados ***individuos*** y las ***columnas variables***.



Podemos usar "Análisis de Componentes Principales" en una gran variedad disciplinas tales como la *ecología, economía, genética, marketing, sociología* etc.

Individuos	Variable 1	Variable 2	Variable 3
1	15	0.6	70
2	25	0.1	50
3	24	0.2	30
4	13	0.3	55
5	23	0.5	54
6	22	0.4	31
7	19	0.1	87

Base de datos

Como la base de datos es una tabla de doble entrada es posible verla desde el lado de los individuos, entonces dos ***individuos*** que tienen características similares *estarán cerca*.

Desde el punto de las variables, podemos visualizar las **relaciones** entre las variables.

Las variables pueden ser cuantitativas continuas o discretas y cualitativas. Pero solo las **cuantitativas continuas** entran en el análisis.

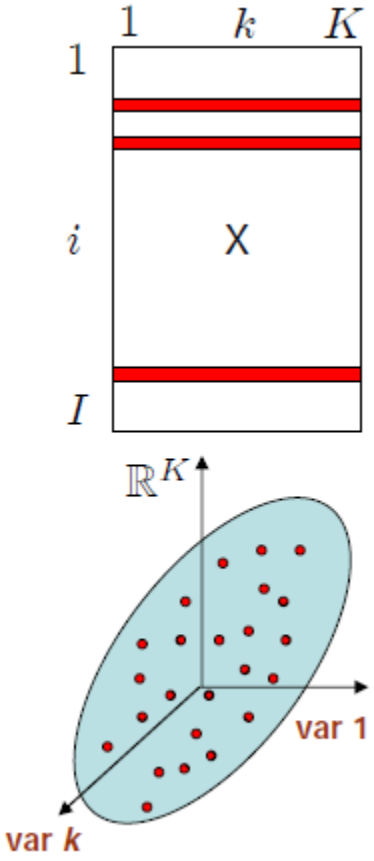
Esta técnica considera que las relaciones son lineares. Por eso, usa coeficientes de correlación.

Es importante usar matrices de correlación para visualizar las relaciones más importantes entre variables.

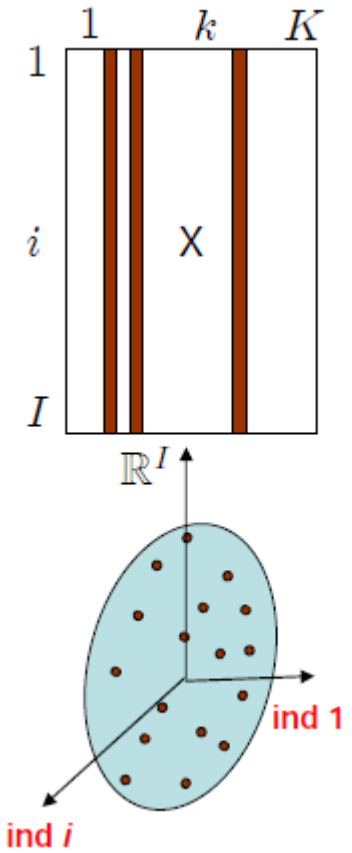
Desde el punto de vista de los individuos se observará una **nube de puntos de los individuos**. Mientras que del punto de vista de las variables una **nube de variables**.

DOS NUBLES DE PUNTOS

El punto de vista desde los individuos

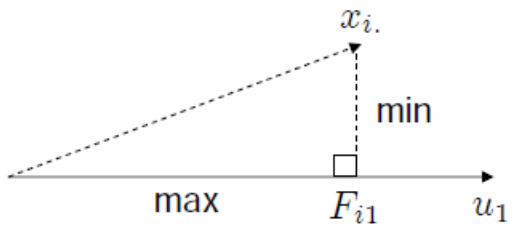


El punto de vista desde las variables



Individuos

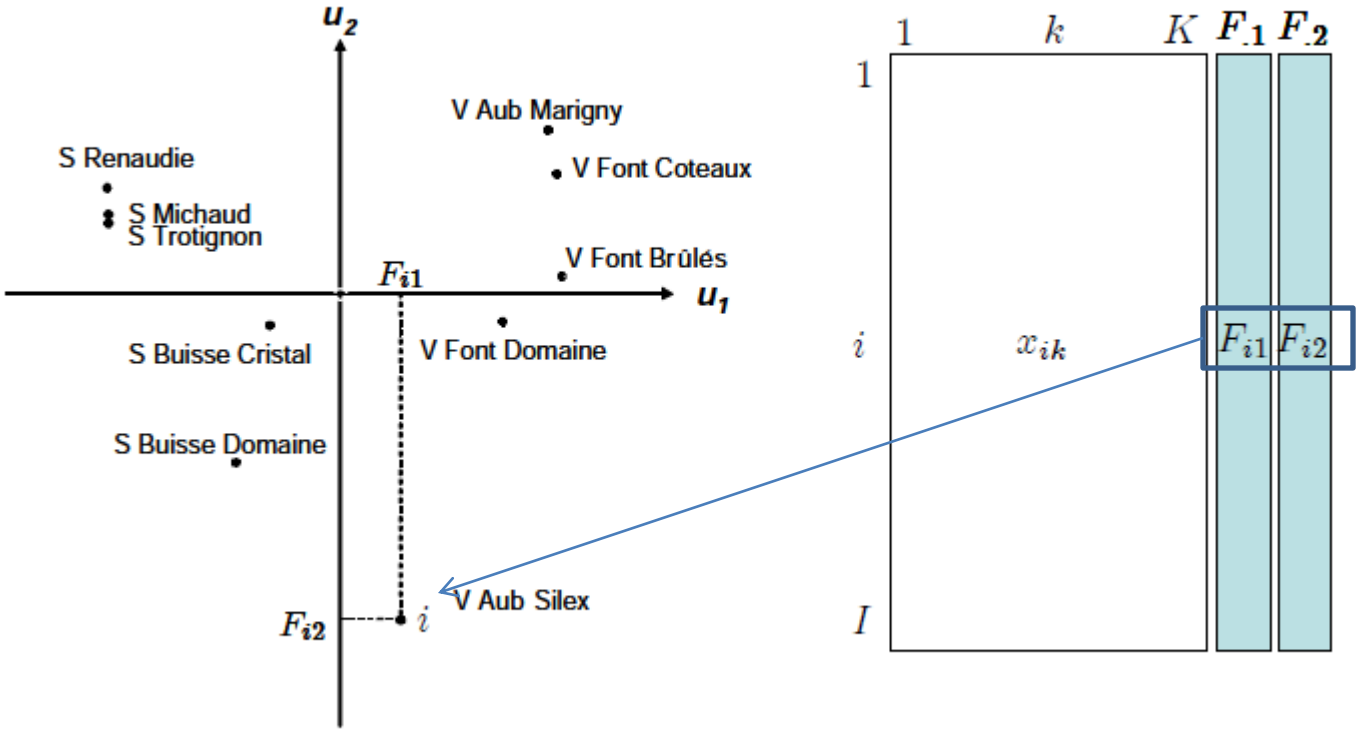
La noción de similitud viene dada por la distancia entre dos individuos al cuadrado. Como hay distancias se puede estudiar de manera geométrica.



A pesar de la alta “dimensionalidad”, la reducción a dos dimensiones permite igual detectar diferencias

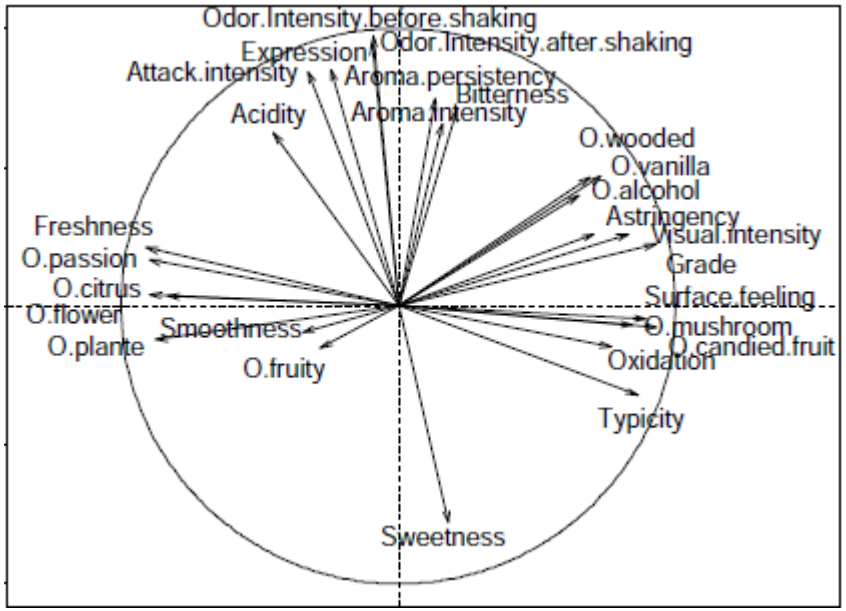
En este ejemplo, las abejas existen en 3 dimensiones. La foto está en 2 dimensiones, sin embargo nos da una buena idea de la distancia entre abejas.

Punto de vista desde los individuos



Punto de vista desde las variables

La nube de puntos de las variables se representa con flechas

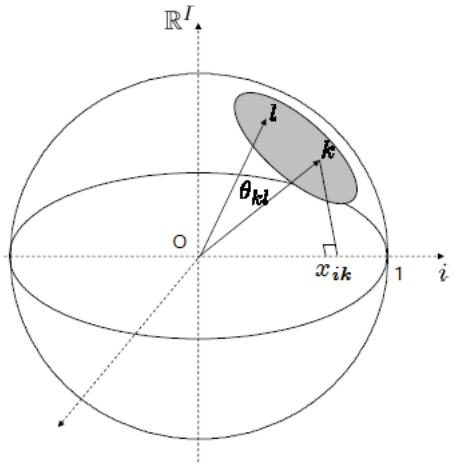


El circulo corresponde a un circulo ***correlación*** que indica la calidad de la representación de las variables

Propiedades de los datos

Los datos pueden tener distintas escalas, debido a que las variables pueden ser muy distintas. Es necesario entonces **centrar** y **estandarizar** los datos.

Centrar: Significa que se traslada la nube de puntos de los individuos para que su centro de gravedad concuerde con el origen. Este procedimiento no modifica la forma de la nube.



Estandarizar: cuando las variables siempre no tienen la misma unidad y para poder compararlas entre si es necesario estandarizar. Se lleva a cabo restando a cada observación la media de la variable dividida por la desviación estándar de cada variable

El procedimiento del análisis

En la primer etapa el análisis de componentes principales halla el primer vector de proyección y por lo tanto el primer eje o componente.

El primer eje o componente maximiza la varianza de los datos.

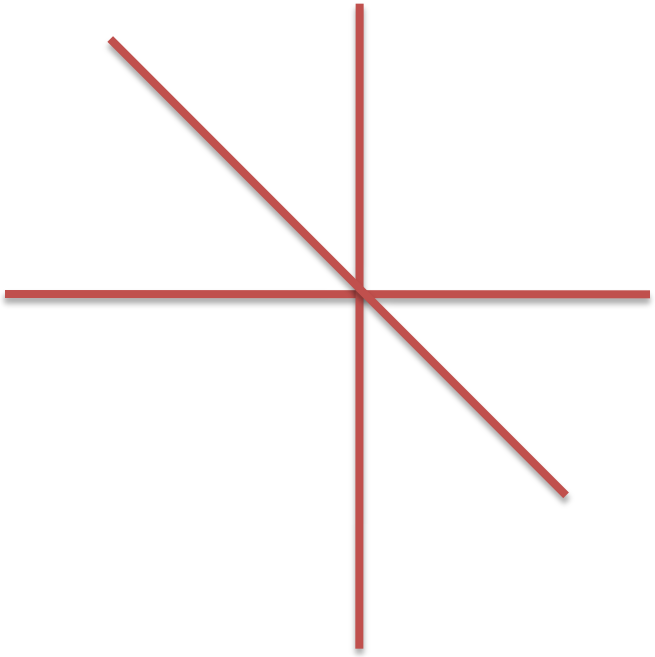
La segunda foto, en dos planos, refleja la **máxima la variación** y es la mejor representación de lo que es un elefante.

Los ejes o componentes

El primer componente distorsiona la nube de punto de los lo menor posible. Queremos que la **distancia al cuadrado** entre un *individuo* y ***su proyección*** en el eje sea la menor posible.

Desde un punto de vista técnico, el análisis de componentes principales, diagonaliza la matriz de correlación para extraer los **eigenvalues** y **eigenvector**.

Una vez definido el primer eje, el resto de los ejes son ortogonales y maximiza la inercia.



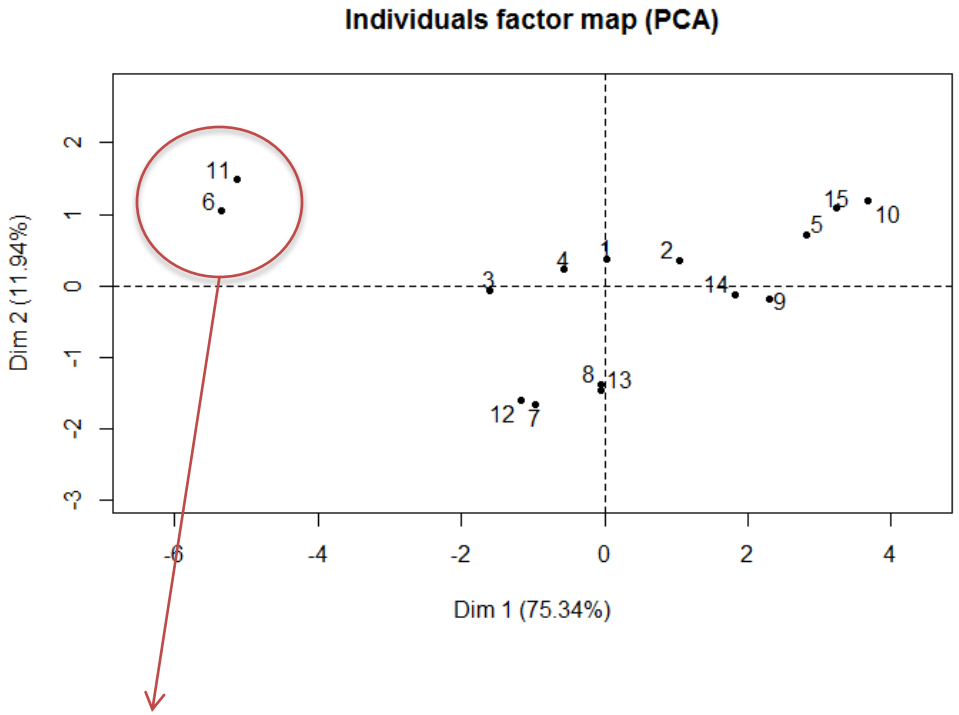
¿Cómo elegir el numero de **dimensiones** en un análisis?

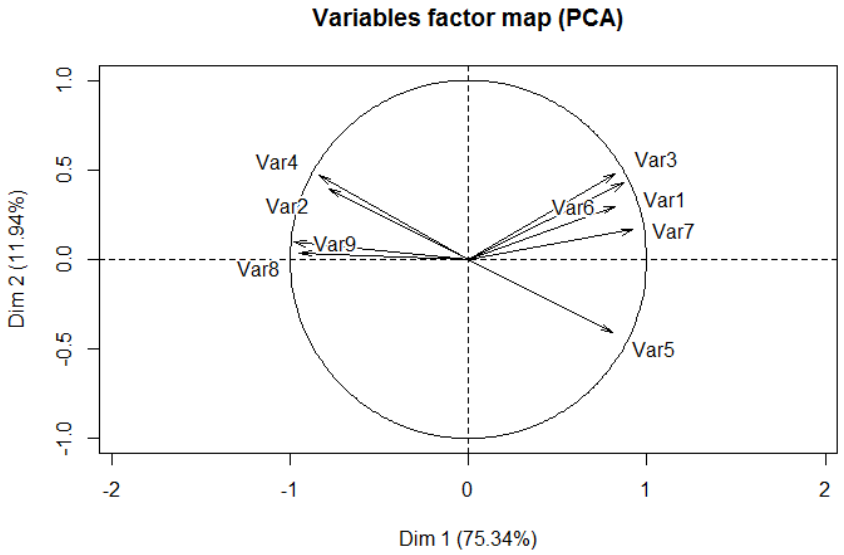
Utilizando los eigenvalue:

Regla del codo ó

eigenvalue > 1

PCA es bastante sensible a valores extremos y/ó outliers





Estos dos puntos tienen altos valores de la variable 4 y 2

En este ejemplo en particular corresponden al tratamiento control o blanco

La eliminación de esos valores va a depender si son realmente outliers y de nuestra pregunta principal

FactoMineR

Es un paquete de R dedicado al análisis multivariado exploratorio de datos.

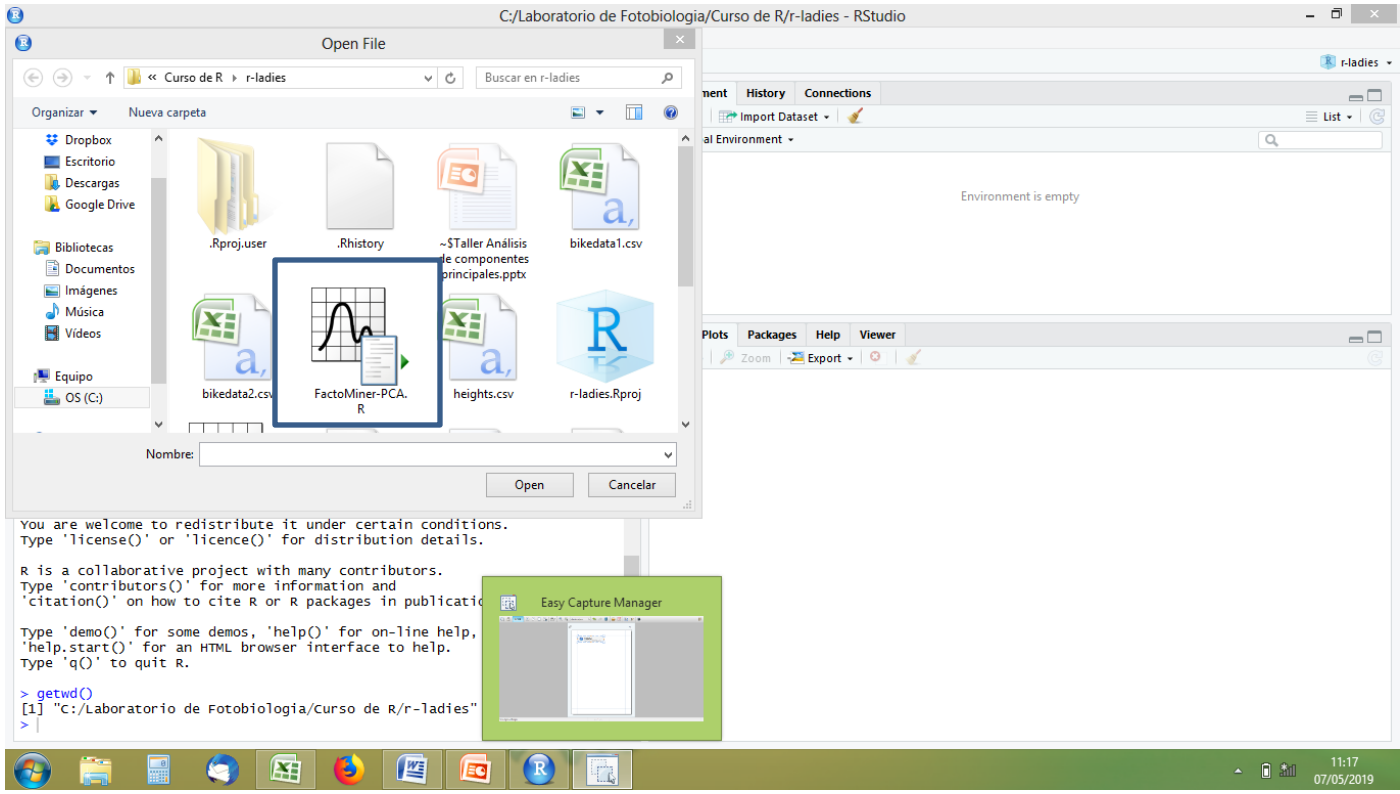
Fue creado y es mantenido por **Francois Husson.**

http://factominer.free.fr/

Pasamos a Rstudio y hacemos un ejemplo

RStudio

Version 1.1.456 – © 2009-2018 RStudio, Inc.



Paquetes a usar: **FactoMiner, factoextra, corrplot, Hmisc**

Base datos: Decathon

Presenta *10 variables* cuantitativas continuas

Dos variables continuas discretas y una variable cualitativa

#data.stan<-scale((data), center=T) ***Esta es la secuencia para centrar y estandarizar***

```
res2 <- rcorr(as.matrix(data[1:10]))
```

Esta sección seleccionamos las variables cuantitativas continuas

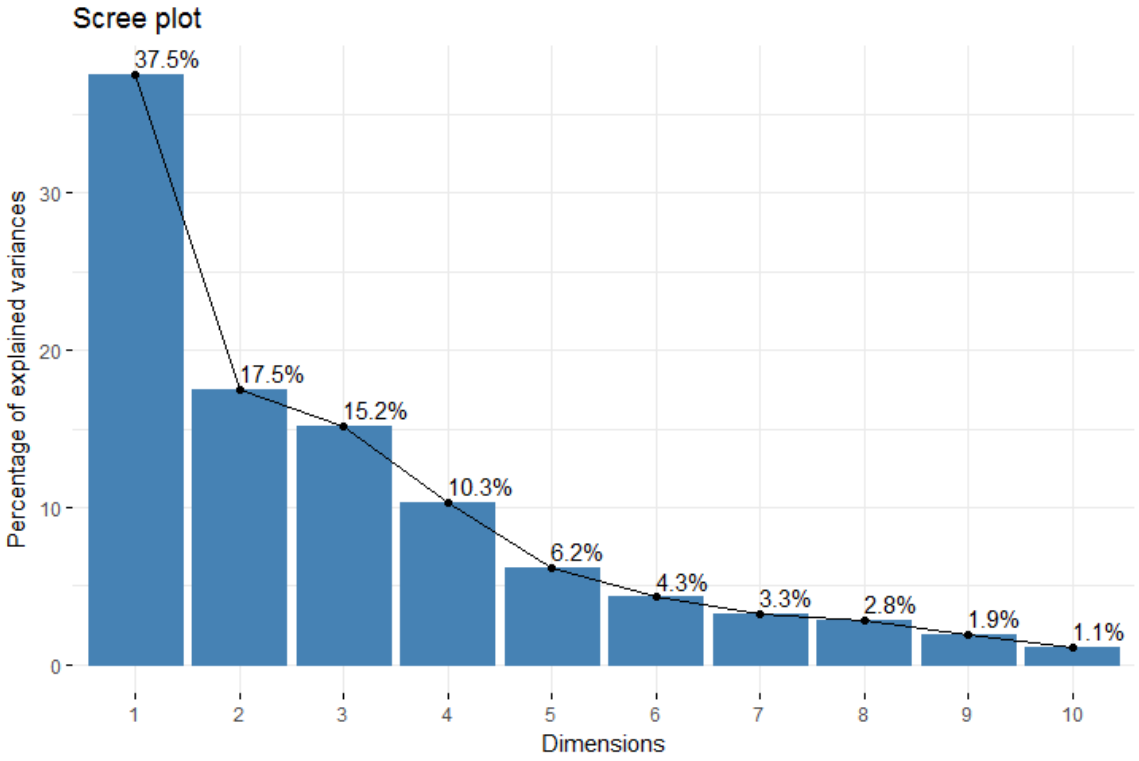
Correlación

```
corrplot(res2$r, type="upper", order="hclust",  p.mat = res2$P,
sig.level = 0.05, insig = "blank", tl.col="black", tl.srt=45)
```

PCA

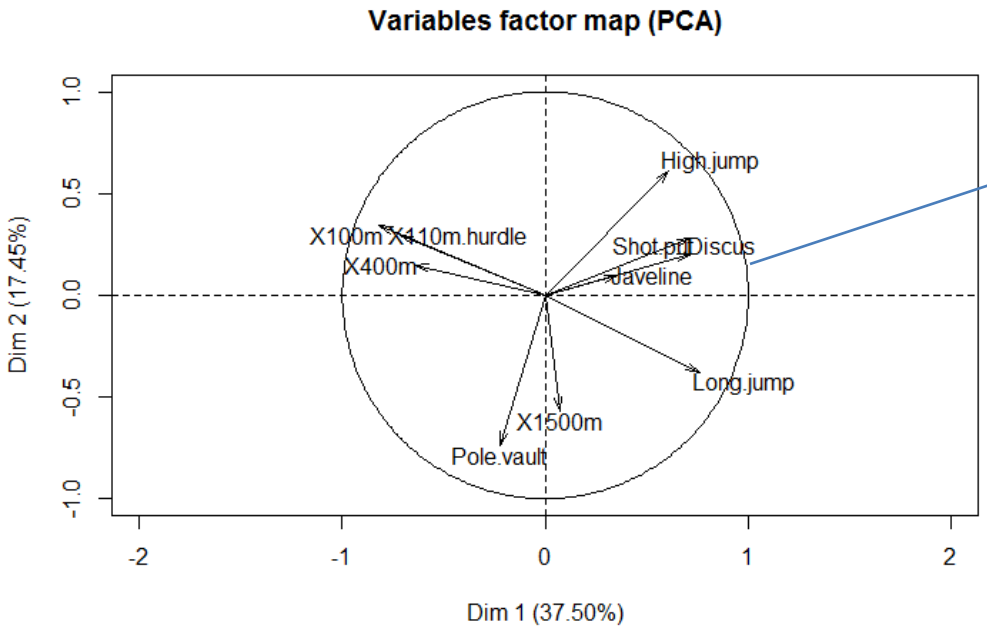
```
res.pca <- PCA(data[1:10], graph = FALSE)
```

El análisis es sobre las variables cuantitativas continuas

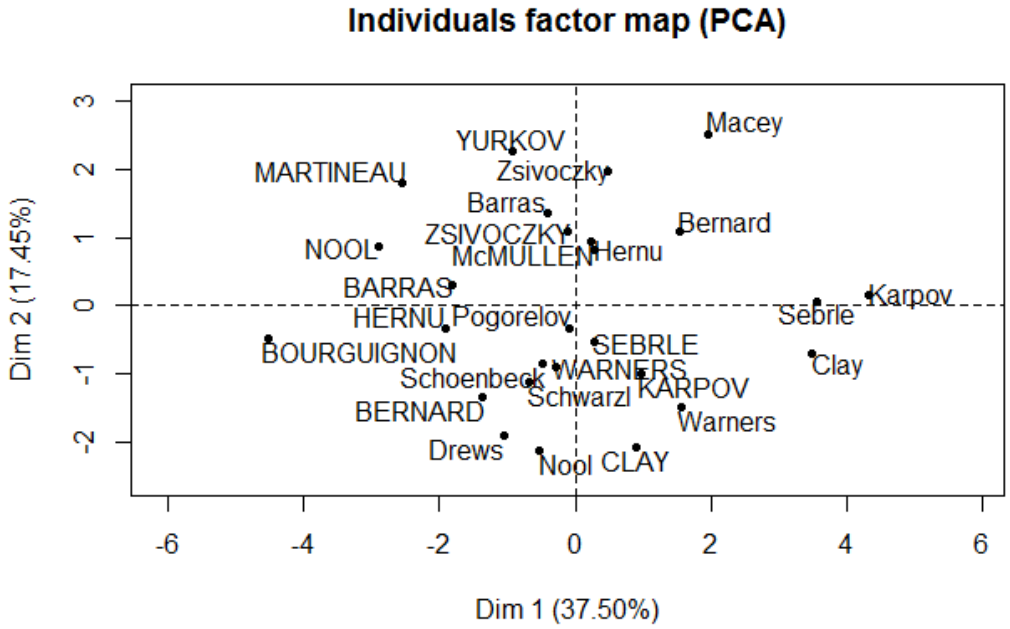


Regla del codo

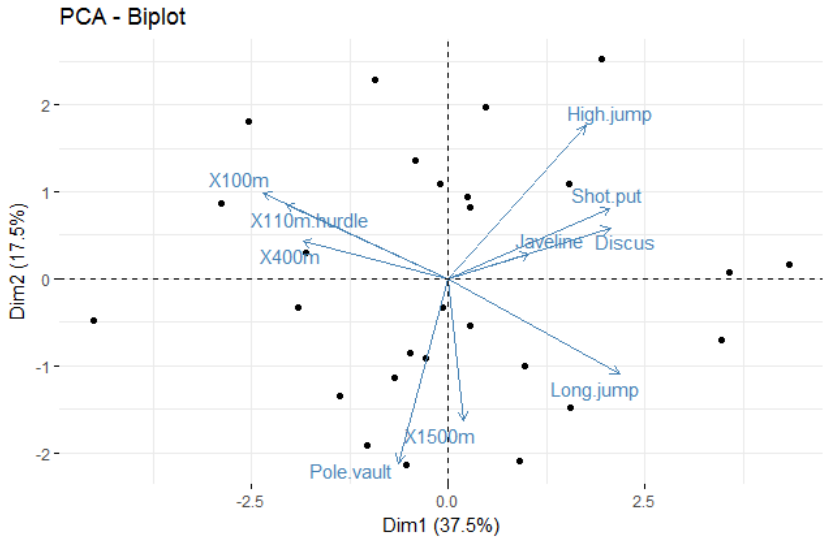
Usando los primeros
3 ejes: 70.11%



Circulo de correlación



Biplot de los individuos y las variables

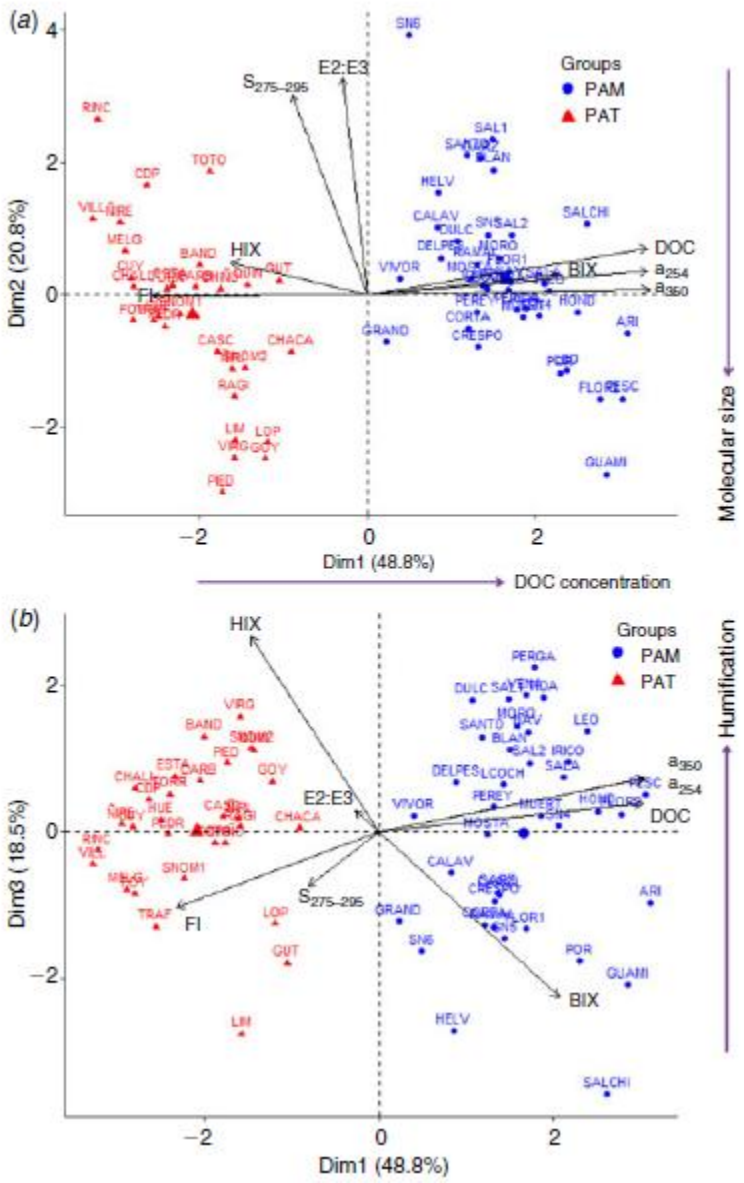


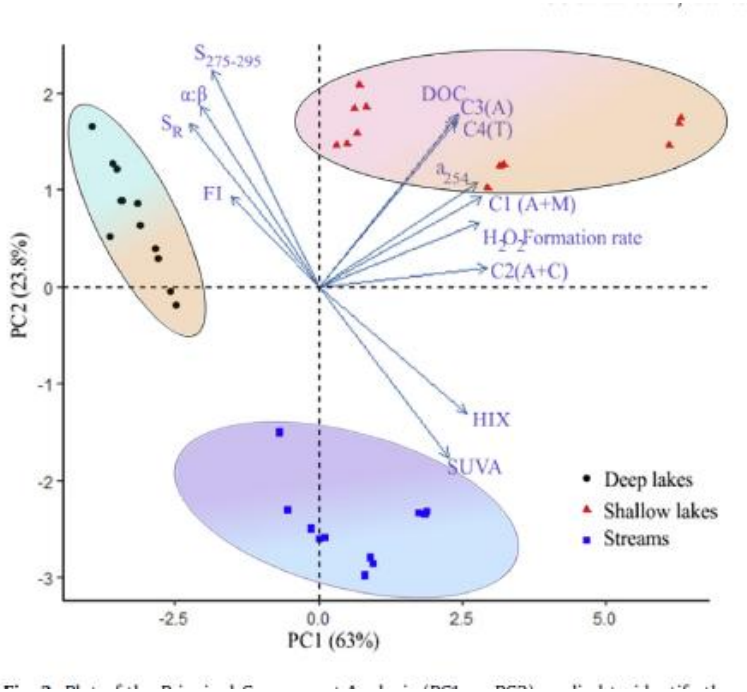
Ambos gráficos son válidos, depende del usuario

```
fviz_pca_biplot(res.pca, label="var", repel=T,
habillage=data$Competition,invisible = "quali")
```

Habillage permite crear un vector para colorear los individuos

Ejemplos de usos de Habillage





PCA usando variables cuantitativas y cualitativas

res.pcaC <- PCA(data, quanti.sup = 11:12, quali.sup = 13)