Minería de datos y Modelización predictiva I

CAPÍTULO I. Técnicas de Reducción de la dimensión

CAPÍTULO II. Análisis Clúster

CAPÍTULO I. Técnicas de Reducción de la Dimensión

I.1.- Introducción

I.1.1.- Técnicas para el Análisis de Datos Multivariantes

I.1.2.-Variables aleatorias multivariantes. Conceptos básicos

I.2.-Análisis de Componentes Principales

I.2.1.- Objetivos y metodología del A.C. P.

I.2.2.- Obtención y determinación del número de C. P.

I.2.3.- Análisis y representación de las variables en el nuevo espacio formado por las C. P.

I.2.4.- Análisis y representación de los individuos en el nuevo espacio formado por las C. P.

I.2.5.- Caso práctico analizado con R

I.1.- Introducción

El Análisis Multivariante es «**la rama de la estadística que estudia las relaciones entre conjuntos de variables dependientes y los individuos para los cuales se han medido dichas variables**» (Kendall).

Sus métodos analizan conjuntamente ***p* variables, medidas sobre un conjunto de *n* individuos.**

CLASIFICACIÓN:

- ❖**Simplificación estructural**: Análisis de Componentes Principales, Factorial y de correspondencias.
- ❖**Clasificación o agrupación**: Análisis Cluster y técnicas de segmentación.
- ❖**Análisis de interdependencia**: Análisis de correspondencias múltiple, Correlaciones Canónicas.
- ❖**Análisis de dependencia:** Métodos de regresión múltiple, Análisis Discriminante y regresión logística.

I.1.2.-Variables aleatorias multivariantes.

VARIABLE ALEATORIA P-DIMENSIONAL.

$$X = \left(X_1, \ldots, X_p\right)$$

Vector de medias

$$\vec{\mu} = E(X) \quad \text{donde} \quad \mu_i = E(X_i)$$

Matriz de Varianzas- Covarianzas

$$\Sigma = \begin{pmatrix} \sigma_{1,1} & \sigma_{1,2} & \cdots & \sigma_{1,p} \\ \sigma_{2,1} & \sigma_{2,2} & \cdots & \sigma_{2,p} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{p,1} & \sigma_{p,2} & \cdots & \sigma_{p,p} \end{pmatrix}$$

$\sigma_{i,i}$ es la varianza de X_i

$$V\left(X_i\right) = \sigma_{ii} = E\left[(X_i - \mu_i)^2\right]$$

$\sigma_{i,j}$ es la covarianza (X_i, X_j)

$$COV\left(X_i, X_j\right) = \sigma_{ij} = E\left[(X_i - \mu_i)(X_j - \mu_j)\right]$$

Observaciones

Individuos

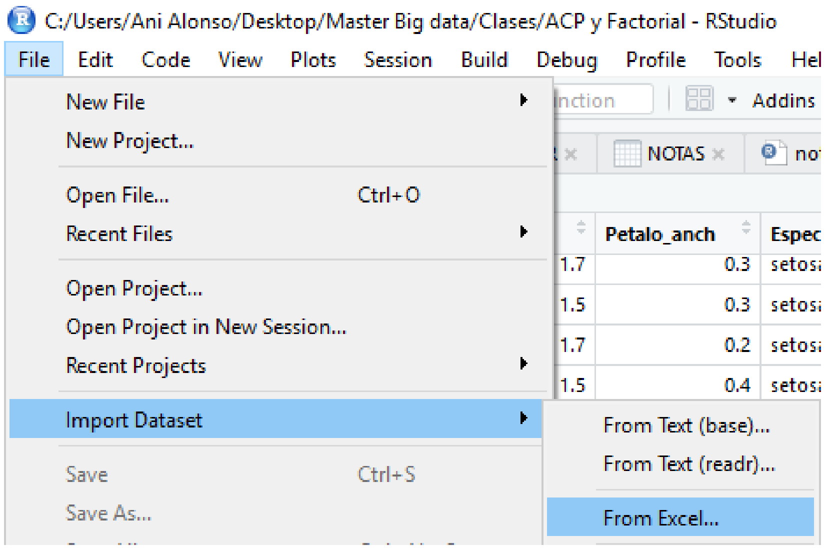
$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdot & \cdot & x_{1p} \\ x_{21} & x_{22} & \cdot & \cdot & x_{2p} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{n1} & x_{n2} & \cdot & \cdot & x_{np} \end{pmatrix}$$

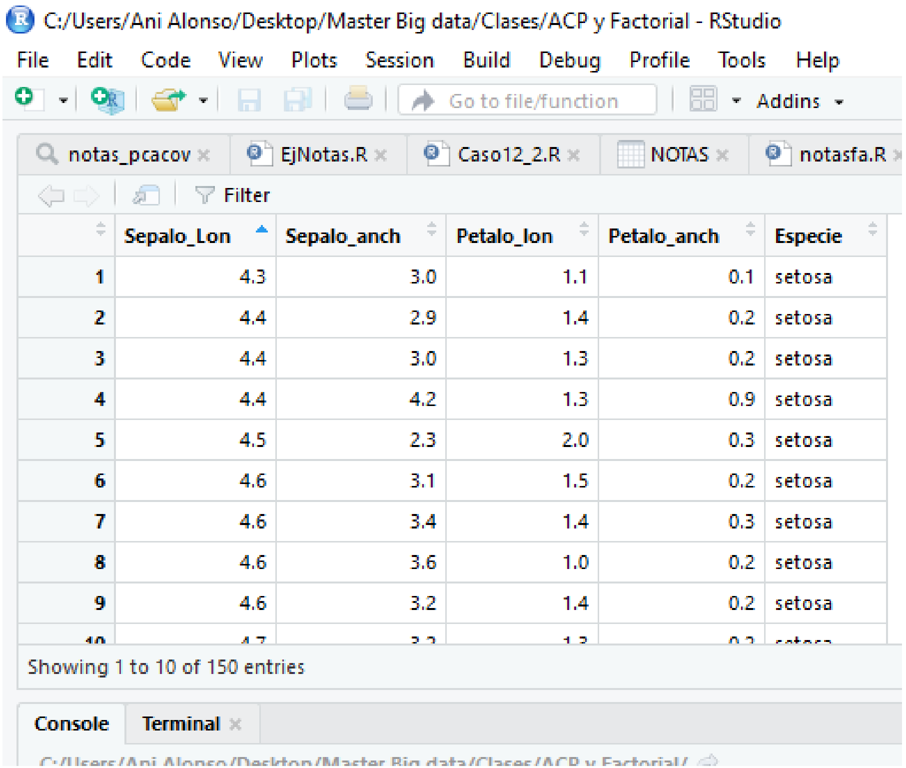
Variables

Estimador

$$\bar{\mathbf{X}} = \left(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p\right)^T$$

Importamos los datos desde Excel






```
datos <- DATOS_IRIS
head(datos)
```

Estimadores de la matriz de Varianzas-Covarianzas y de la matriz de correlaciones

$$\mathbf{S} = \begin{pmatrix} S_{11} & S_{12} & . & . & S_{1p} \\ S_{21} & S_{22} & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ S_{p1} & S_{p2} & . & . & S_{pp} \end{pmatrix}$$

$$S_{ii} = \frac{\sum_{k=1}^{n} (x_{ki} - \bar{x}_i)^2}{n-1}$$

$$S_{ij} = \frac{\sum_{k=1}^{n} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{n-1}$$

$$\mathbf{S} = (\mathbf{X} - \mathbf{1}\bar{\mathbf{X}})'(\mathbf{X} - \mathbf{1}\bar{\mathbf{X}}) / (n-1)$$

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & . & . & r_{1p} \\ r_{21} & 1 & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ r_{p1} & r_{p2} & . & . & 1 \end{pmatrix}$$

$$r_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}}\sqrt{S_{jj}}}$$

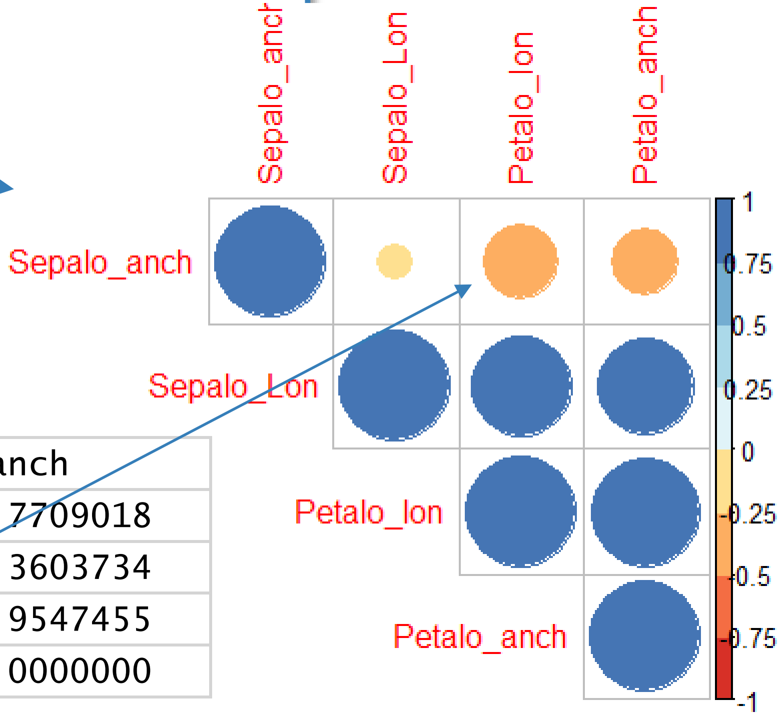
$$\mathbf{R} = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}$$

$$\mathbf{D} = Diag(\mathbf{S})$$

```r
library(corrplot)
library(RColorBrewer)
#Creamos un matriz con las variables numéricas
dat_n<-datos[, c("Sepalo_Lon", "Sepalo_anch", "Petalo_lon", "Petalo_anch")]
R <-cor(dat_n)
Print(R)
corrplot(R, type="upper", order="hclust",
      col=brewer.pal(n=8, name="RdYlBu"))
```

$$\mathbf{R} = \mathbf{D}^{-\frac{1}{2}}\mathbf{S}\mathbf{D}^{-\frac{1}{2}}$$

	Sepalo_Lon	Sepalo_anch	Petalo_lon	Petalo_anch
Sepalo_Lon	1.0000000	-0.1002911	0.8268914	0.7709018
Sepalo_anch	-0.1002911	1.0000000	-0.4457940	-0.3603734
Petalo_lon	0.8268914	-0.4457940	1.0000000	0.9547455
Petalo_anch	0.7709018	-0.3603734	0.9547455	1.0000000



Instalamos y cargamos las librerías que vamos a necesitar

```
install. Packages("lattice")
install.packages(“pastecs")
install.packages("corrplot")
install.packages("ggplot2")
install.packages("factoextra")
install.packages("FactoMineR")
```

```
#===========================
# Cargamos librerías
#===========================
library("lattice")
library(pastecs)
library(corrplot)
library(ggplot2)
library(factoextra)
library(FactoMineR)
```

FactoMineR & factoextra

Analyzing & Visualizing Multivariate Data

- Performs PCA, (M)CA, FAMD, MFA, HCPC & more
- Provides the coordinates, the quality of representation and the contribution of individuals & variables
- Predicts the results for supplementary individuals & variables

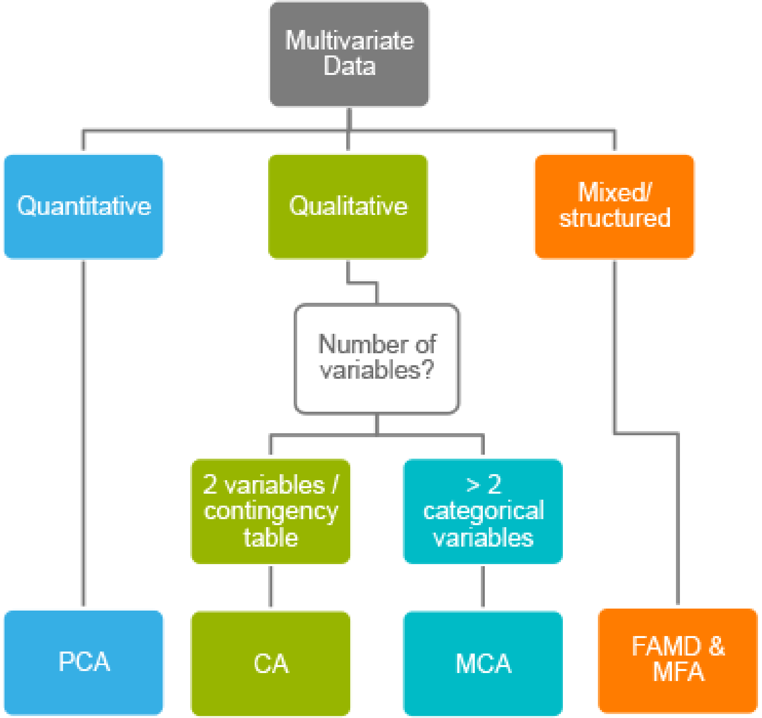
- Produces ggplot2-based elegant data visualization and facilitates the interpretation
- Creates human readable outputs
- Simplifies cluster analysis and visualization

FactoMineR

1. Analyze

Principal Component Methods

Methods to Summarize & Visualize Multivariate Data



- PCA: Principal Component Analysis
- (M) CA: (Multiple) Correspondence Analysis
- FAMD: Factor Analysis of Mixed Data
- MFA: Multiple Factor Analysis

factoextra

2. Visualize

3. Interpret

Functions	Description
fviz_eig (or fviz_eigenvalue)	Visualize eigenvalues.
fviz_pca	Graph of PCA results.
fviz_ca	Graph of CA results.
fviz_mca	Graph of MCA results.
fviz_mfa	Graph of MFA results.
fviz_famd	Graph of FAMD results.
fviz_hmfa	Graph of HMFA results.
fviz_ellipses	Plot ellipses around groups.
fviz_cos2	Visualize element cos2.[1]
fviz_contrib	Visualize element contributions.[2]

Functions	Description
get_eigenvalue	Access to the dimension eigenvalues.
get_pca	Access to PCA outputs.
get_ca	Access to CA outputs.
get_mca	Access to MCA outputs.
get_mfa	Access to MFA outputs.
get_famd	Access to MFA outputs.
get_hmfa	Access to HMFA outputs.
facto_summarize	Summarize the analysis.

I.2.- Análisis de Componentes Principales

Si tomamos demasiadas variables **es difícil visualizar relaciones** entre ellas.

Otro problema que se presenta es la **fuerte correlación**.

Se hace necesario, pues, reducir el número de variables sin perder información.

Es importante resaltar el hecho de que el concepto de **mayor información se relaciona con el de mayor variabilidad o varianza**.

OBJETIVO →

Se buscan **$m<p$** variables que **sean combinaciones lineales de las originales** y que estén incorreladas, recogiendo la mayor parte de la información o variabilidad de los datos.

Ejemplo.

Alumno	Mats	Francés	Inglés	Física
1	1	4	5	3
2	5	5	4	4
3	6	10	9	6
4	3	6	6	4
5	2	4	6	1
6	6	8	7	8
7	6	6	6	7
8	3	5	8	4
9	8	5	5	8
10	2	5	8	4
11	9	7	7	9

¿Existen relaciones entre las calificaciones de algunas asignaturas (variables) que nos permita agruparlas?

¿Podemos calcular uno o dos índices que resuman el comportamiento de los individuos (alumnos) teniendo en cuenta sus calificaciones?

I.2.2.- Obtención y determinación del número de C. P.

¿CÓMO SE OBTIENEN LAS COMPONENTES PRINCIPALES?.

Partiendo bien de la **matriz de Correlaciones** ó de la **matriz de Varianzas-Covarianzas** estimadas, se hallará su descomposición, en función de sus valores propios y la matriz formada por sus autovectores correspondientes.

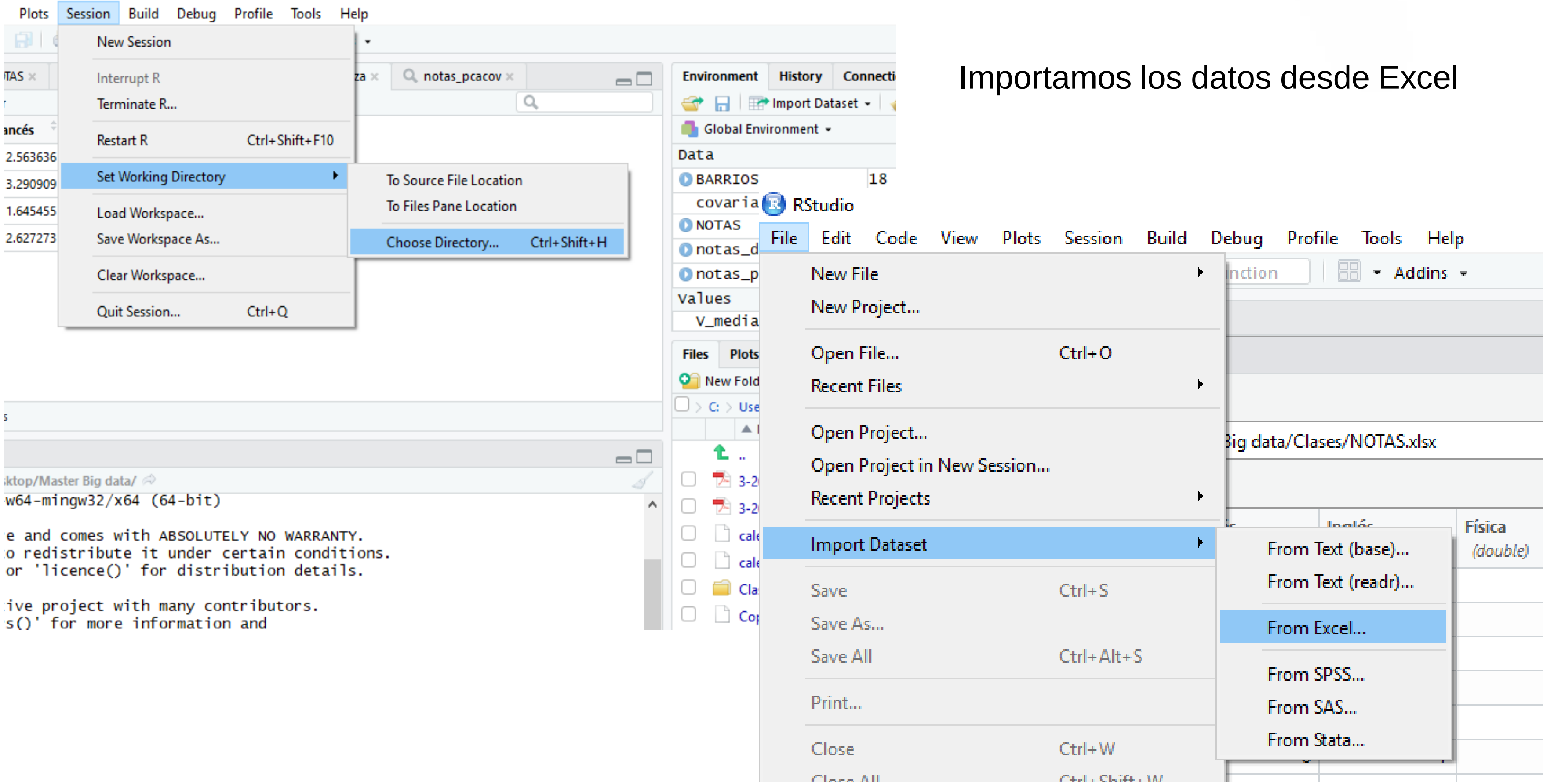
$$\mathbf{S} = \mathbf{P}\boldsymbol{\Lambda}\ \mathbf{P}^{-1} \qquad \mathbf{R} = \mathbf{P}_R\boldsymbol{\Lambda}_R\ \mathbf{P}_R^{-1}$$

Matriz de contiene los autovectores

Matriz diagonal formada por los autovalores

Importamos el fichero Excel con las calificaciones de los 11 alumnos

Importamos los datos desde Excel



¿Cómo realizar un análisis de Componentes Principales con R?

Primero creamos una matriz con los datos de las variables numéricas:

Alumno	Mats	Francés	Inglés	Física
1	1	4	5	3
2	5	5	4	4
3	6	10	9	6
4	3	6	6	4
5	2	4	6	1
6	6	8	7	8
7	6	6	6	7
8	3	5	8	4
9	8	5	5	8
10	2	5	8	4
11	9	7	7	9

```
datos<- as.data.frame(NOTAS)
rownames(datos)<-datos[,1]
notas_dat<-datos[,-1]
```

Guardamos la primera columna que suele ser categórica en el vector rownames para luego utilizarlas como etiquetas

```
#================================
# Descriptivos de las variables
#================================
#Descriptivos
library(pastecs)
stat.desc(notas_dat,basic=FALSE)
```

```
                  Mats   Francés    Inglés    Física
median       5.0000000 5.0000000 6.0000000 4.0000000
mean         4.6363636 5.9090909 6.4545455 5.2727273
SE.mean      0.7893925 0.5469676 0.4545455 0.7518572
CI.mean.0.95 1.7588761 1.2187198 1.0127904 1.6752422
var          6.8545455 3.2909091 2.2727273 6.2181818
std.dev      2.6181187 1.8140863 1.5075567 2.4936282
coef.var     0.5646923 0.3069992 0.2335651 0.4729295
```

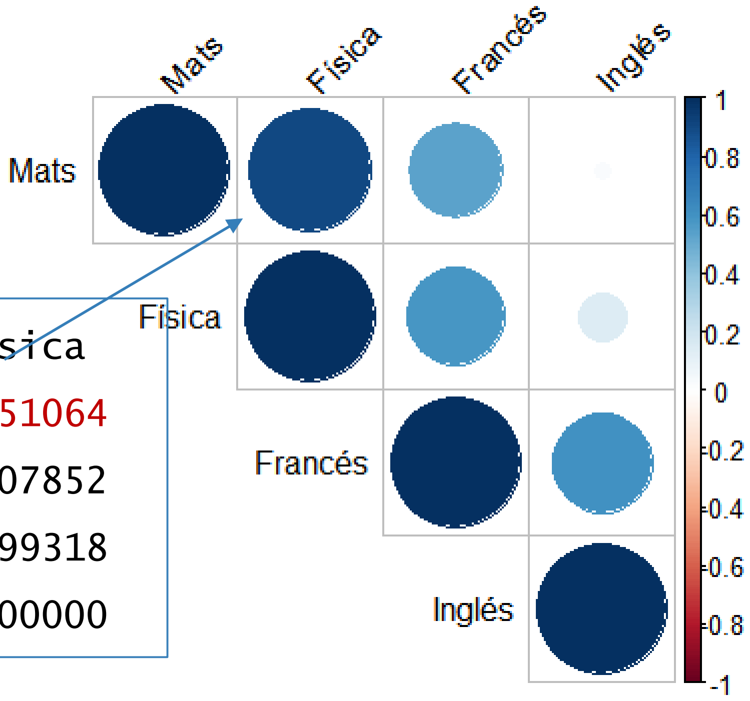
```
#Matriz correlaciones

Library(corrplot)

R<-cor(notas_dat, method="pearson")

print(R, digits = 2)
```

	Mats	Francés	Inglés	Física
Mats	1.000	0.54	0.021	0.91
Francés	0.540	1.00	0.602	0.58
Inglés	0.021	0.60	1.000	0.15
Física	0.905	0.58	0.150	1.00

Cuando tenemos muchas variables es muy útil poder analizar las correlaciones mediante una salida gráfica. La siguiente sintaxis representa las correlaciones según su valor en el gráfico anterior.

```
corrplot(R, type="upper", order="hclust",tl.col="black", tl.srt=45)
```

	Mats	Francés	Inglés	Física
Mats	1.0000000	0.5397705	0.0207294	0.9051064
Francés	0.5397705	1.0000000	0.6016644	0.5807852
Inglés	0.0207294	0.6016644	1.0000000	0.1499318
Física	0.9051064	0.5807852	0.1499318	1.0000000



La sintaxis básica del procedimiento de R para realizar un análisis de componentes principales

```
fit<-PCA(notas_dat, scale.unit=TRUE, ncp=4,graph=TRUE)
```

$$X_{ij}^{*} = \frac{X_{ij} - \bar{X}_{j}}{\sqrt{S_{jj}}}$$

- **Fichero de datos:** Las variables deben ser todas númericas
- **scale.unit:** valor lógico. *TRUE*,indica que los datos serán estandarizados, pasan a tener media 0 y desviación típica 1. Es decir la matriz a diagonalizar es la matriz de correlaciones entre las variables. **Es conveniente utilizarlo siempre.**
- **ncp:** Número de componentes a retener en el resultado final.
- **graph:** valor lógico. TRUE indica que se mostrarán los gráficos.

El resultado del análisis de componentes principales se guardará en el fichero de **datos** fit `head(fit)`

Esta función nos muestra todo el contenido del objeto fit que contiene listas y matrices

Resultados del PCA guardados en fit:

1 "$eig" Autovalores

2 "$var“ Resultados para las variables, que son los siguientes:

3 "varcoord“: Coordenadas de las variables en las Componentes"

4 "varcor" : Correlaciones entre las variables y las Componentes

Coinciden

5 "varcos2" "cosenos al cuadrado de las variables“

6 "varcontrib" "contributions of the variables"

7 "$ind" Resultados para los individuos, que son los siguientes:

8 "indcoord" "coord. for the individuals"

9 "indcos2" "cos2 for the individuals“

10 "indcontrib" "contributions of the individuals“

11 "$call" "summary statistics" : Medidas estadísticas

12 "$call$centre" "mean of the variables“

13 "$call$ecart.type" "standard error of the variables"

Functions	Description
get_eigenvalue	Access to the dimension eigenvalues.
get_pca	Access to PCA outputs.

1 "$eig" Autovalores de la matriz R

Proporción de variabilidad total explicada por la componente 1

```
$`eig`
       eigenvalue percentage of variance cumulative percentage of variance
comp 1 2.48576274    62.144068          62.14407
comp 2 1.19280496    29.820124          91.96419
comp 3 0.23856665     5.964166          97.92836
comp 4 0.08286565     2.071641         100.00000
```

$$\mathbf{R} = \mathbf{P}_R \mathbf{\Lambda}_R \; \mathbf{P}_R^{-1}$$

Proporción de variabilidad total explicada por las componentes 1 y 2.

Los autovalores suman p (diagonal de la matriz R), es decir el número de variables

¿Cómo se construyen las Componentes Principales?

Los autovectores asociados a dichos autovalores nos dan los coeficientes de la combinación lineal con la que se construyen las componentes.

```
$svd$v
          [,1]       [,2]       [,3]       [,4]
[1,] 0.5487816 -0.4158181  0.1001327  0.7182670
[2,] 0.5382892  0.3426941 -0.7625266 -0.1065772
[3,] 0.2834952  0.7829499  0.5292340  0.1628840
[4,] 0.5733358 -0.3108784  0.3583825 -0.6679840
```

$$\mathbf{R} = \mathbf{P}_R \boldsymbol{\Lambda}_R \; \mathbf{P}_R^{-1}$$

Las Componentes Principales, tienen, por lo tanto, vector de dirección, e_1, e_2, e_3 y e_4.

Por ejemplo, la componente 1 será combinación lineal de las variables estandarizadas:

$$CP_1 = 0.548X_1^* + 0.538X_2^* + 0.283X_3^* + 0.573X_4^*$$

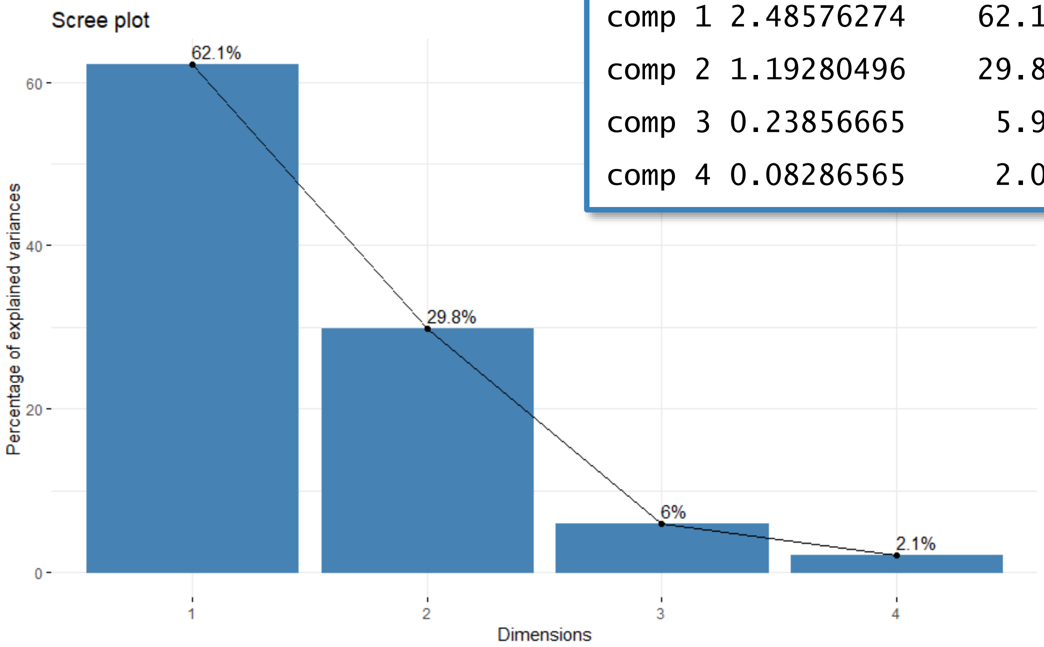
¿Cuántas Componentes principales son suficientes?

Lo ideal es explicar aproximadamente el 90% de la variabilidad total. Al menos que la proporción de **variabilidad explicada sea > 0.7**

```
$`eig`
       eigenvalue percentage of variance cumulative percentage of variance
comp 1 2.48576274    62.144068          62.14407
comp 2 1.19280496    29.820124          91.96419
comp 3 0.23856665     5.964166          97.92836
comp 4 0.08286565     2.071641         100.00000
```

Representación gráfica de la proporción de varianza explicada por cada Componente Principal

```
# Scree plot
Library(factoextra)
fviz_eig(fit,addlabels=TRUE)
```




```
$`eig`
eigenvalue percentage of variance cumulative percentage of variance
comp 1 2.48576274     62.144068      62.14407
comp 2 1.19280496     29.820124      91.96419
comp 3 0.23856665      5.964166      97.92836
comp 4 0.08286565      2.071641     100.00000
```

Las **covarianzas entre cada componente principal y las variables** vienen dadas por el producto de las coordenadas del vector propio que define el componente por su valor propio:

$$Cov(Y_i, X_j) = \lambda_i e_{ij}$$

La **correlación entre una componente principal y una variable X** es proporcional al coeficiente de esa variable en la definición del componente.

$$Corr(Y_i, X_j) = \frac{Cov(Y_i, X_j)}{\sqrt{V(Y_i)V(X_j)}} = \frac{\lambda_i e_{ij}}{\sqrt{\lambda_i s_j^2}} = e_{ij}\frac{\sqrt{\lambda_i}}{s_j}$$

Por ejemplo la correlación entre la variable X_1 y la componente 1 es:

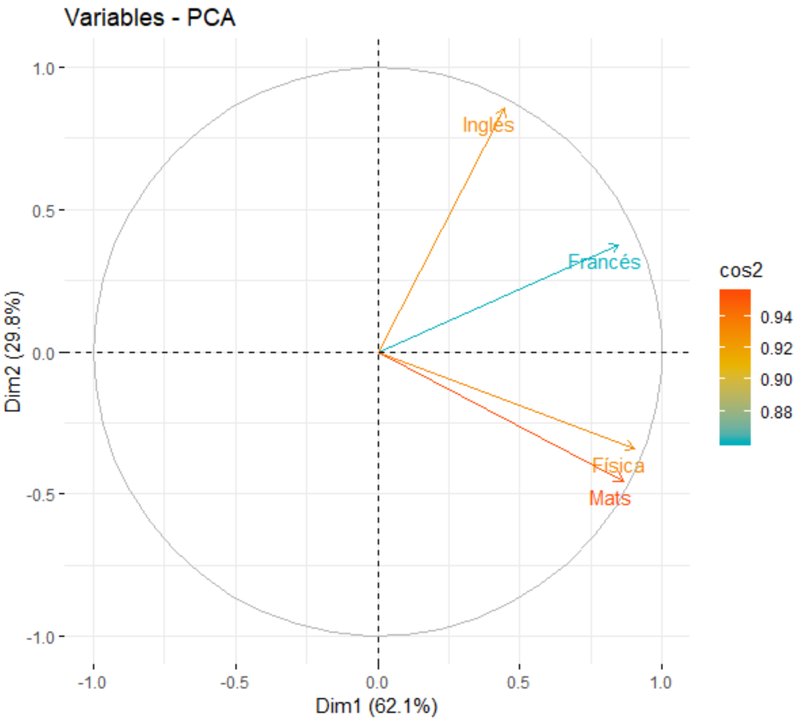
$$Corr(Y_1, X_1) = 0.548\frac{\sqrt{2.486}}{\sqrt{6.854}} = 0.865$$

```
$var$cor
             Dim.1      Dim.2
Mats     0.8652256 -0.4541383
Francés  0.8486830  0.3742755
Inglés   0.4469671  0.8551036
Física   0.9039386 -0.3395278
```

```
#======================================
# Representación gráfica variables
#======================================
fviz_pca_var(fit, col.var="cos2", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
repel = TRUE # Avoid text overlapping (slow if many points))
```

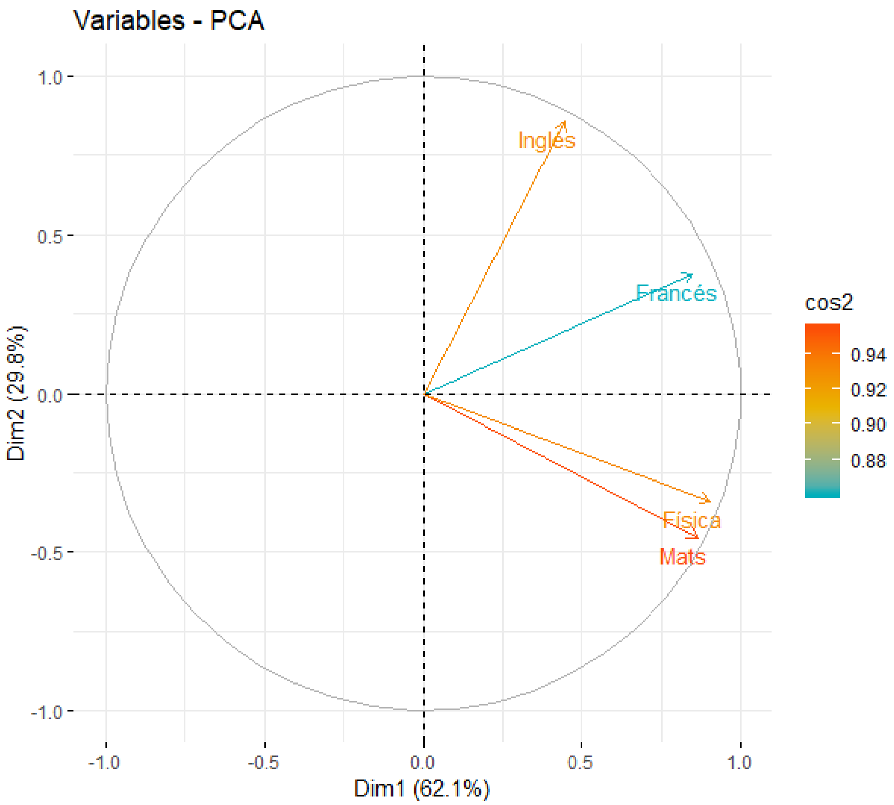
$$Corr(Y_i, X_j^*) = e_{ij}\sqrt{\lambda_i}$$

```
$var$cor
            Dim.1      Dim.2
Mats    0.8652256 -0.4541383
Francés 0.8486830  0.3742755
Inglés  0.4469671  0.8551036
Física  0.9039386 -0.3395278
```



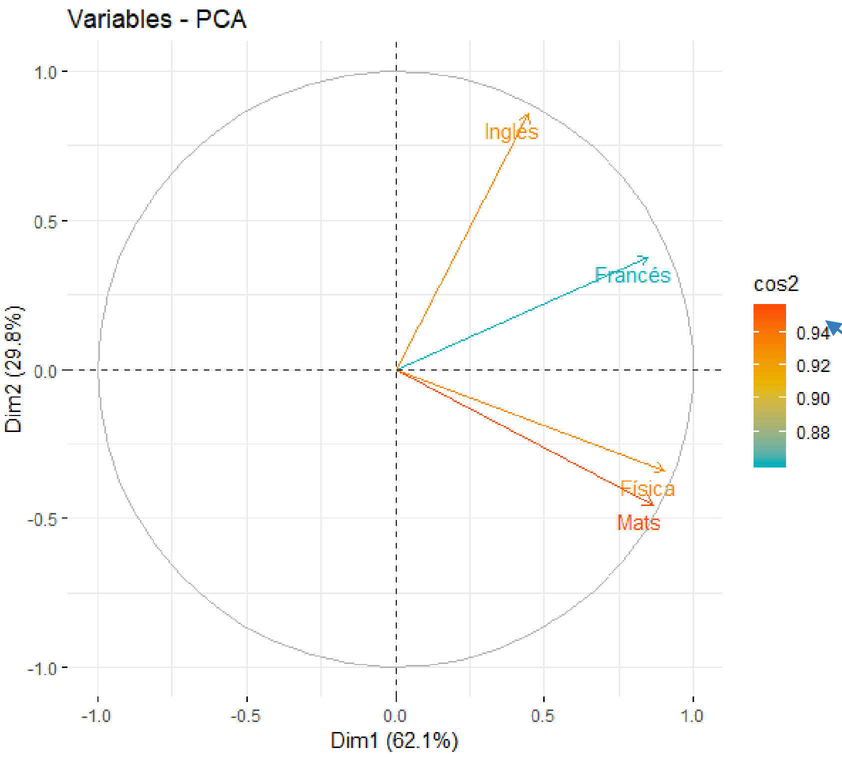
Las coordenadas de cada variable son el coeficiente de correlación entre la variable y las nuevas componentes

La primera componente recoge sobre todo los valores de Física y Matemáticas, por lo que podríamos identificar dicha componente como la que representa las calificaciones de las asignaturas de ciencias.



La segunda componente recoge sobre todo los valores de Francés e Inglés, por lo que podríamos identificar dicha componente como la que recoge las calificaciones en idiomas.

Los cosenos al cuadrado son las correlaciones al cuadrado que expresan la proporción de la varianza de cada variable que es explicada por cada componente:




```
fit[["var"]][["cos2"]]
```

	Dim.1	Dim.2
Mats	0.7486154	0.2062416
Francés	0.7202628	0.1400821
Inglés	0.1997796	0.7312021
Física	0.8171050	0.1152791

$$\mathrm{Cos}^2(Y_1, X_1) = Corr(Y_1, X_1)^2 = 0.865^2$$

$$\mathrm{Cos}^2(Y_2, X_1) = Corr(Y_2, X_1)^2 = -0.454^2$$

$$\boxed{\mathrm{Cos}^2(X_1) = 0.748 + 0.206 = 0.954}$$

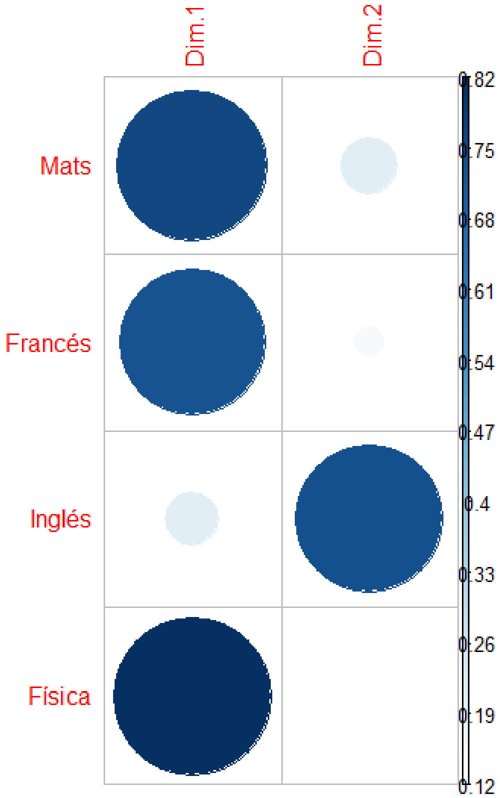
```
$var$cor
```

	Dim.1	Dim.2
Mats	0.8652256	-0.4541383
Francés	0.8486830	0.3742755
Inglés	0.4469671	0.8551036
Física	0.9039386	-0.3395278

```
#================================
# Representación gráfica de los cosenos
#Guardamos los estadísticos asociados a las variables en el objeto var
var<-get_pca_var(fit)
 corrplot(var$cos2,is.corr=FALSE)
```

Gráficamente se muestra que la varianza de las variables Matemáticas, Física y Fránces es explicada por la Componente 1 mientras que la Componente 2 explica Inglés

```
fit[["var"]][["cos2"]]
            Dim.1     Dim.2
Mats    0.7486154 0.2062416
Francés 0.7202628 0.1400821
Inglés  0.1997796 0.7312021
Física  0.8171050 0.1152791
```

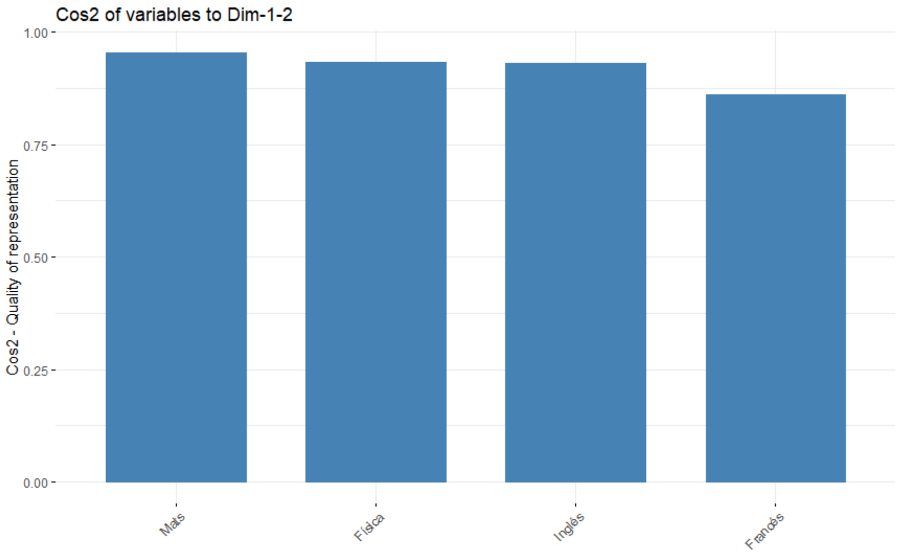


```
#Porcentaje de variabilidad explicada por las dos CP
fviz_cos2(fit, choice="var", axes=1:2)
```

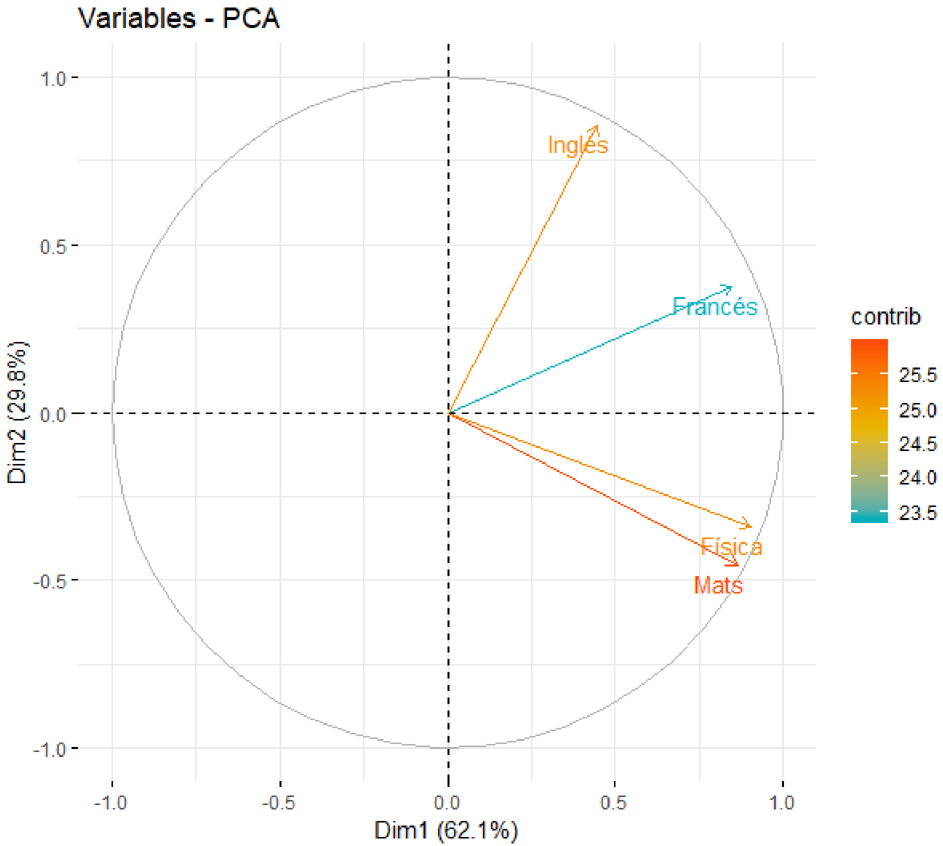
Mostramos el porcentaje de la varianza de las variables que es explicada por las dos Componentes en total

```
fit[["var"]][["cos2"]]
```

	Dim.1	Dim.2
Mats	0.7486154	0.2062416
Francés	0.7202628	0.1400821
Inglés	0.1997796	0.7312021
Física	0.8171050	0.1152791



La contribución de una variable a una Componente es el **porcentaje de varianza de la Componente** que es proviene de esa variable




```
$var$contrib
```

	Dim.1	Dim.2
Mats	30.116124	17.290471
Francés	28.975525	11.743927
Inglés	8.036952	61.301061
Física	32.871399	9.664541

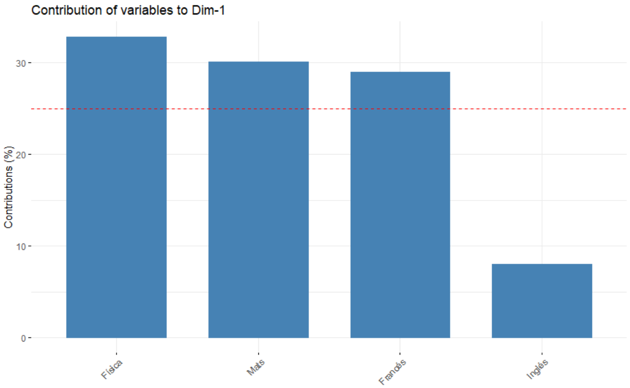
$$Contr(X_1,Y_1)=\frac{\mathrm{Cos}^2(X_1,Y_1)}{\sum_j^p \mathrm{Cos}^2(X_j,Y_1)}\cdot 100$$

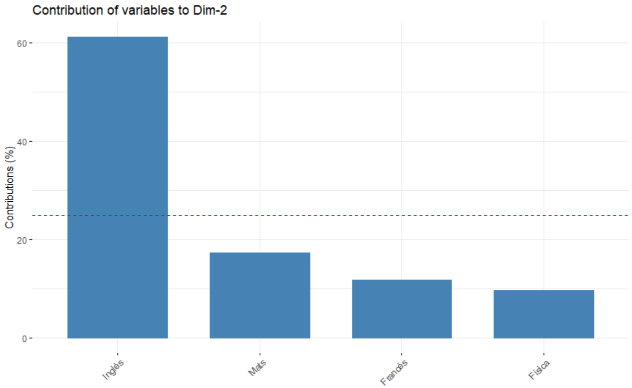
```
fit[["var"]][["cos2"]]
```

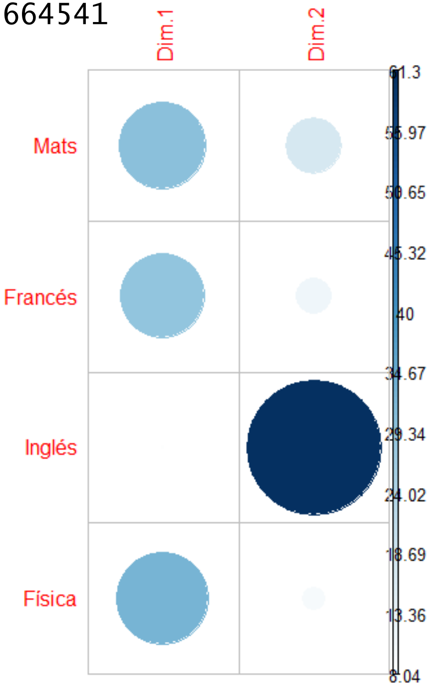
	Dim.1	Dim.2
Mats	0.7486154	0.2062416
Francés	0.7202628	0.1400821
Inglés	0.1997796	0.7312021
Física	0.8171050	0.1152791

```
corrplot(var$contrib,is.corr=FALSE)
#Contribución de las variables a la Componente 1
fviz_contrib(fit,choice="var",axes=1,top=10)
fviz_contrib(fit,choice="var",axes=2,top=10)
```

```
$var$contrib
            Dim.1     Dim.2
Mats     30.116124 17.290471
Francés  28.975525 11.743927
Inglés    8.036952 61.301061
Física   32.871399  9.664541
```







Cuando con dos Componentes no explicamos lo suficiente y necesitamos 3 o más, es necesario representar tanto las variables como los individuos en los planos formados por todas las componentes (1,3), (2,3), etc..

```
fviz_pca_ind(fit, axes = c(2, 3), col.ind = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE # Avoid text overlapping (slow if many points))
```

1.2.4. COORDENADAS DE LOS INDIVIDUOS SOBRE LOS NUEVOS EJES.

```
$ind$`coord`
         Dim.1      Dim.2
1   -2.2284748 -0.2676368
2   -0.9939902 -1.4112633
3    2.2503261  1.8747871
4   -0.7280039  0.2094130
5   -2.2936920  0.3719952
6    1.7157345  0.1276254
7    0.6549433 -0.6825749
8   -0.6447584  1.1006836
9    0.8273274 -1.8893062
10  -0.8645983  1.2672589
11   2.3051863 -0.7009820
```

$$\Psi_{\alpha i} = \sum_{j}^{p} \frac{\left(x_{ij} - \overline{x}_j\right)}{S_j} e_{\alpha j}$$

Observemos, que la coordenada del primer alumno en la Primera Componente, viene dada por

Media de la CP1

$$\sum_{i=1}^{11} y_{i1} = 0$$

Varianza de la CP1

$$\frac{\sum_{i=1}^{11} y_{i1}^2}{10} = \lambda_1 = 2.485$$

$$y_{11} = \frac{(1-4.63)}{2.61} \cdot 0.548 + \frac{(4-5.9)}{1.81} \cdot 0.538 + \frac{(5-6.45)}{1.51} \cdot 0.283 + \frac{(3-5.27)}{2.49} \cdot 0.573 = -2.228$$

En R además podemos representar los individuos en el plano de componentes, porque dichas representaciones pueden ofrecernos información adicional sobre el comportamiento de los individuos.

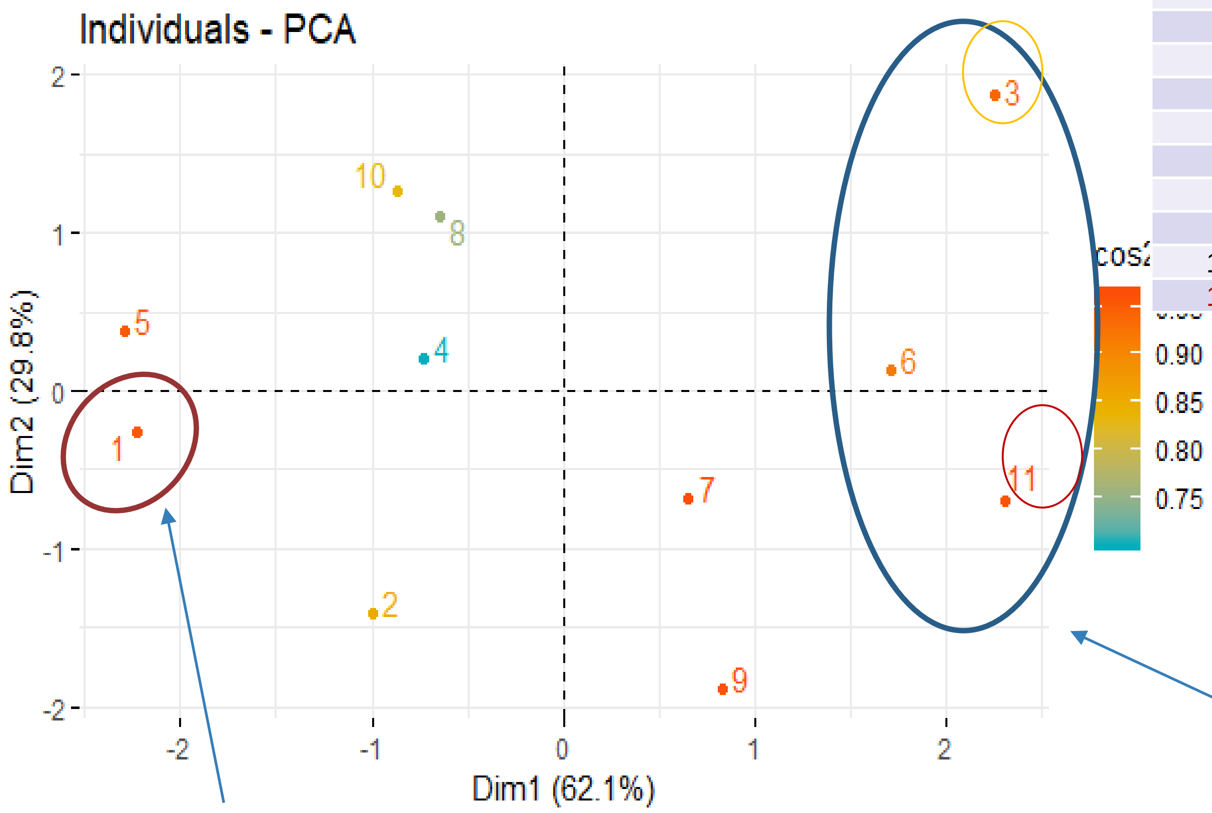
```
#==================================
# Representación gráfica individuos
#==================================
 fviz_pca_ind(fit, col.ind = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE # Avoid text overlapping (slow if many points) )
```

Alumno	Mats	Francés	Inglés	Física
1	1	4	5	3
2	5	5	4	4
3	6	10	9	6
4	3	6	6	4
5	2	4	6	1
6	6	8	7	8
7	6	6	6	7
8	3	5	8	4
9	8	5	5	8
10	2	5	8	4
11	9	7	7	9

```
$ind$`coord`
        Dim.1      Dim.2
1  -2.2284748 -0.2676368
2  -0.9939902 -1.4112633
3   2.2503261  1.8747871
4  -0.7280039  0.2094130
5  -2.2936920  0.3719952
6   1.7157345  0.1276254
7   0.6549433 -0.6825749
8  -0.6447584  1.1006836
9   0.8273274 -1.8893062
10 -0.8645983  1.2672589
11  2.3051863 -0.7009820
```

Representación de los individuos en el espacio de las componentes 1 y 2

Buenos en Letras



Alumno	Mats	Francés	Inglés	Física
1	1	4	5	3
2	5	5	4	4
3	6	10	9	6
4	3	6	6	4
5	2	4	6	1
6	6	8	7	8
7	6	6	6	7
8	3	5	8	4
9	8	5	5	8
10	2	5	8	4
11	9	7	7	9

Buenos en ciencias

1 -2.2284748 -0.2676368

RESUMEN: SISTEMÁTICA DEL ANÁLISIS DE COMPONENTES PRINCIPALES.

1) Determinación, a la vista de los datos, de cual va a ser la matriz de partida: la matriz de Covarianzas o **la de Correlaciones**.
2) Determinación del **número de Componentes**.
3) Interpretación, si procede, de la **relación entre las Componentes y variables originales** a través de los coeficientes de correlación.
4) Representación de las variables en el espacio de las Componentes. Esto permitirá **crear asociaciones entre variables.**
5) Obtención de **las coordenadas de los individuos en el espacio de las Componentes**.
6) Representación los individuos en el espacio de las Componentes. Mediante dichas representaciones, se **detectarán grupos de individuos, observaciones atípicas**, etc.

Bibliografía

- Análisis de Datos Multivariantes. Peña D. 2002.
- Nuevos Métodos de Análisis Multivariante. Cuadras C.M. 2014
- Análisis Multivariante de Datos. Pérez, C. Ed. Garceta. 2013
- Practical Guide to Principal Component Methods in R. A. Kassambara. Ed STHDA.com. 2017
- http://www.sthda.com/english/

ntic
master
revolucionamos la comunicación

UNIVERSIDAD
COMPLUTENSE
DE MADRID