ntic master
School

UNIVERSIDAD
COMPLUTENSE
DE MADRID

TÉCNICAS DE ANÁLISIS MULTIVARIANTE

CAPÍTULO 1. Técnicas de Reducción de la dimensión

CAPÍTULO 2. Análisis Clúster

UNIVERSIDAD
COMPLUTENSE
DE MADRID

CAPÍTULO I. Técnicas de Reducción de la Dimensión

1.1.- 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, árboles de clasificación.

❖**Análisis de interdependencia:** Análisis de correspondencias múltiple, Correlaciones Canónicas. Modelos loglineales.

❖**Análisis de dependencia:** Métodos de regresión múltiple, Análisis Discriminante y regresión logística.

1.1.2.-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

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}$$

σ_{ii} es la varianza de X_i

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

σ_{ij} 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

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

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

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

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

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

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

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

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

$$\boxed{-1 \leq r_{ij} \leq 1}$$

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

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

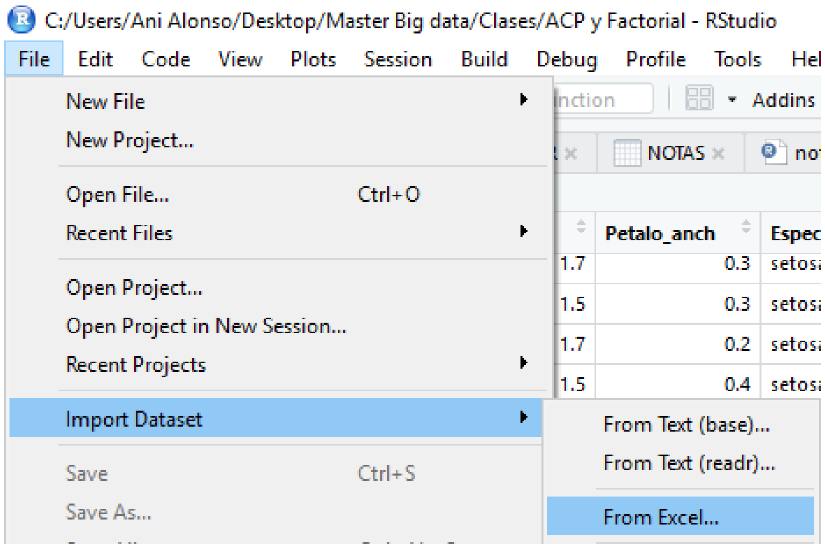
Análisis de datos Multivariantes con R

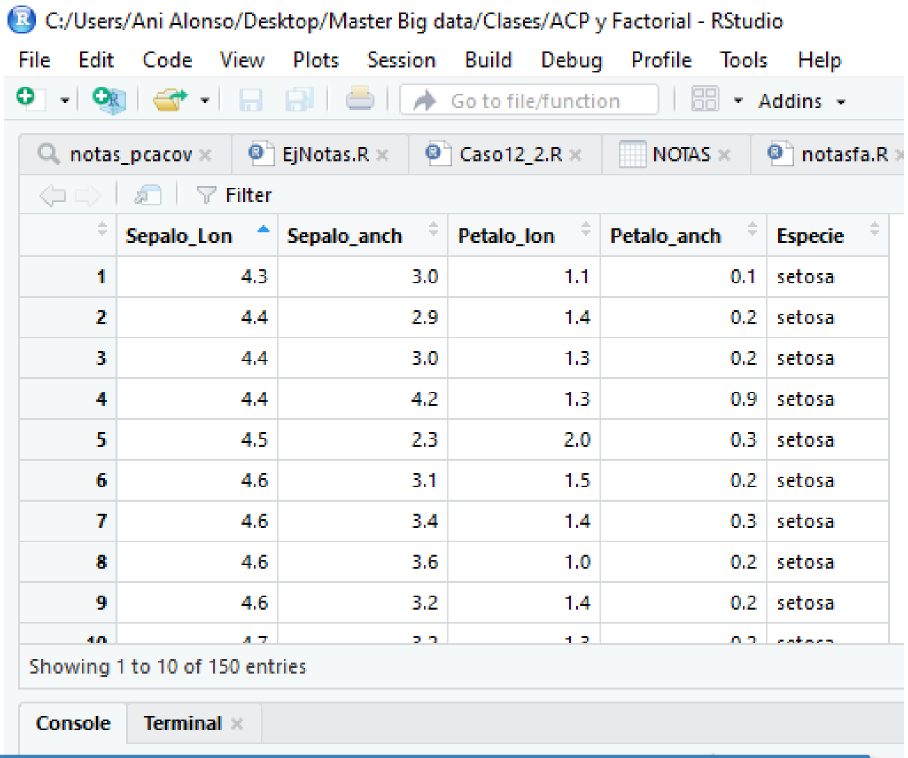
Instalamos y cargamos las librerías que vamos a necesitar

```r
install.packages("readxl")
install.packages("knitr")
install.packages("pastecs")
install.packages("ggplot2")
install.packages("corrplot")
install.packages("factoextra")
install.packages("FactoMineR")
```

```r
# Cargamos librerías
library(readxl)
library(knitr)
library(pastecs)
library(ggplot2)
library(corrplot)
library(factoextra)
library(FactoMineR)
```

Importamos los datos desde Excel






```
DATOS_IRIS<- read_excel("C:/Users/reven/OneDrive/Desktop/Master Big data/Clases/ACP y Factorial/DATOS IRIS.xlsx")
datos<- as.data.frame(DATOS_IRIS)
```

```r
dat_Iris<-datos[-5]

tabla1<-stat.desc(dat_Iris,basic=FALSE)
knitr::kable(tabla1,  digits =2,caption = "Estadísticos descriptivos")
```

	Sepalo_Lon	Sepalo_anch	Petalo_lon	Petalo_anch
median	5.80	3.00	4.35	1.30
mean	5.86	3.07	3.77	1.21
SE.mean	0.07	0.04	0.14	0.06
CI.mean.0.95	0.14	0.08	0.28	0.12
var	0.71	0.22	3.07	0.57
std.dev	0.84	0.47	1.75	0.75
coef.var	0.14	0.15	0.47	0.62

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

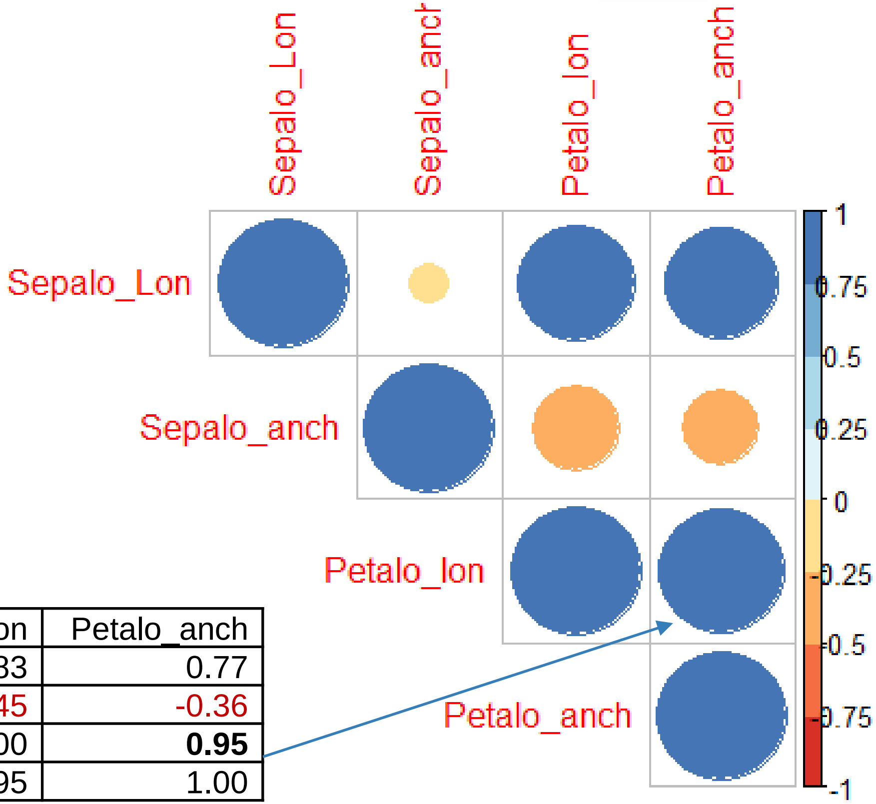
```r
R <-cor(dat_Iris)
knitr::kable(R,  digits =2,caption = "Correlaciones")
```

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

	Sepalo_Lon	Sepalo_anch	Petalo_lon	Petalo_anch
Sepalo_Lon	1.00	-0.10	0.83	0.77
Sepalo_anch	-0.10	1.00	-0.45	-0.36
Petalo_lon	0.83	-0.45	1.00	**0.95**
Petalo_anch	0.77	-0.36	0.95	1.00

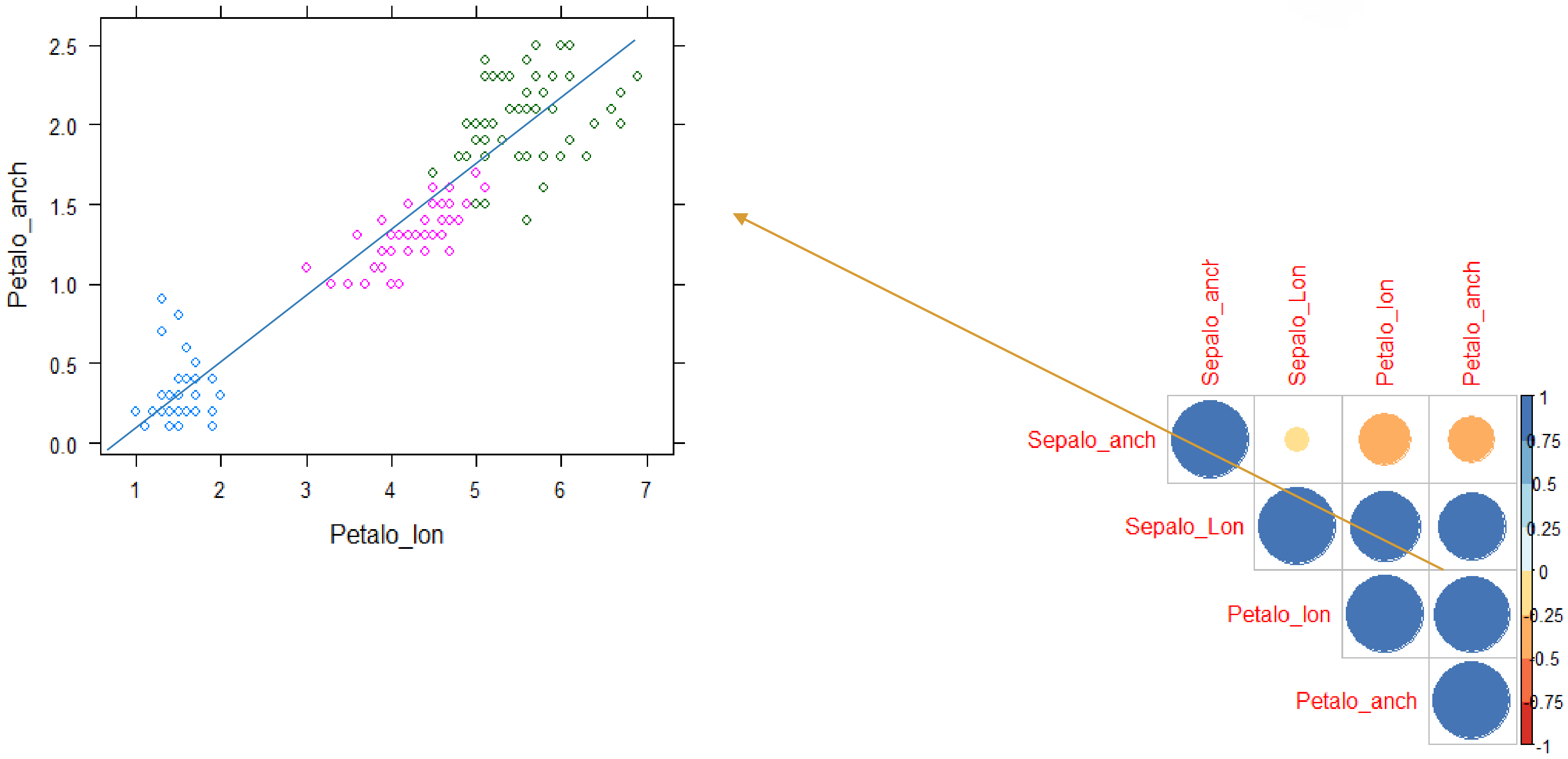
```
corrplot(R, type="upper", col=brewer.pal(n=8, name="RdYlBu"))
```

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

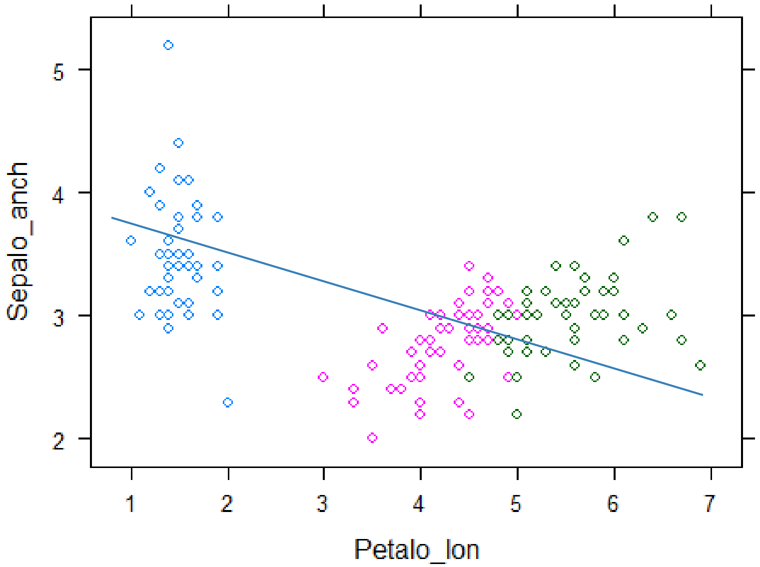


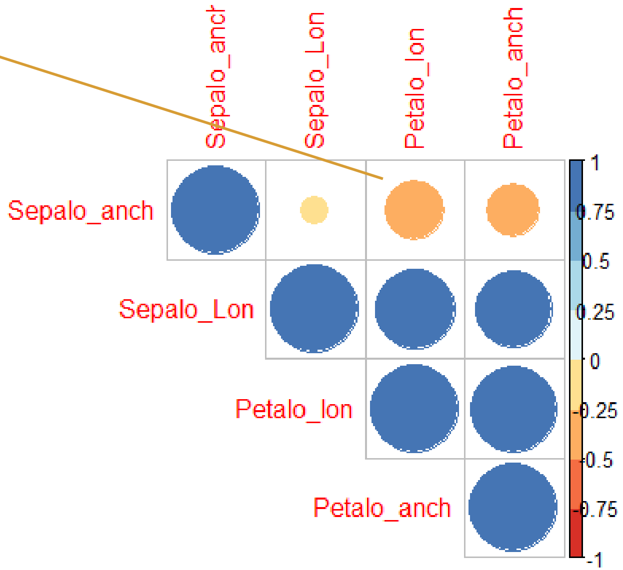
	Sepalo_Lon	Sepalo_anch	Petalo_lon	Petalo_anch
Sepalo_Lon	1.00	-0.10	0.83	0.77
Sepalo_anch	-0.10	1.00	-0.45	-0.36
Petalo_lon	0.83	-0.45	1.00	**0.95**
Petalo_anch	0.77	-0.36	0.95	1.00

```
xyplot(Petalo_anch ~ Petalo_lon, group = Especie, data =DATOS_IRIS)
```



```
xyplot(Sepalo_anch ~ Petalo_lon, group = Especie, data =DATOS_IRIS)
```





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

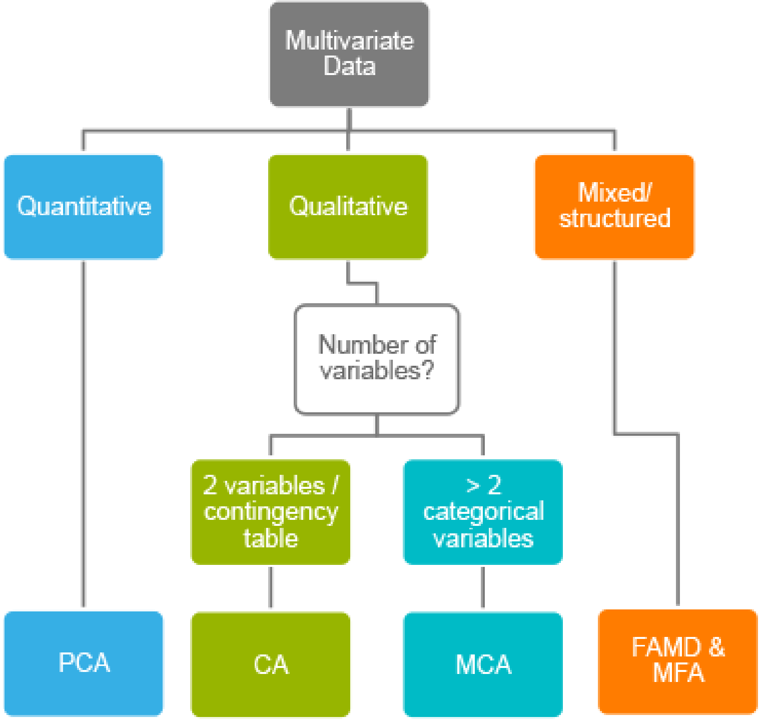
- Produces ggplot2-based elegant data visualization and facilitates the interpretation
- Creates human readable outputs
- Simplifies cluster analysis and visualization

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.

1.2.- 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

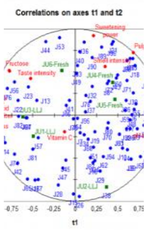
Se buscan **$m<p$** variables que **sean combinaciones lineales de las originales** y que estén incorreladas, recogiendo la mayor parte de la información o variabilidad de los datos.

¿Para que sirven las Componentes Principales?

Para crear índices socieconómicos que resuman la información de muchas variables. Índices psicológicos.

Para visualizar relaciones entre individuos

Para diversificar el riesgo en gestión de carteras de inversión, Portfolios



Como herramienta para un mejor funcionamiento de modelos de predicción y clasificación. Regresión PCR, PLS, Cluster sobre componentes

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?

1.2.2.- Obtención y determinación del número de C. P.

¿CÓMO SE OBTIENEN LAS COMPONENTES PRINCIPALES?.

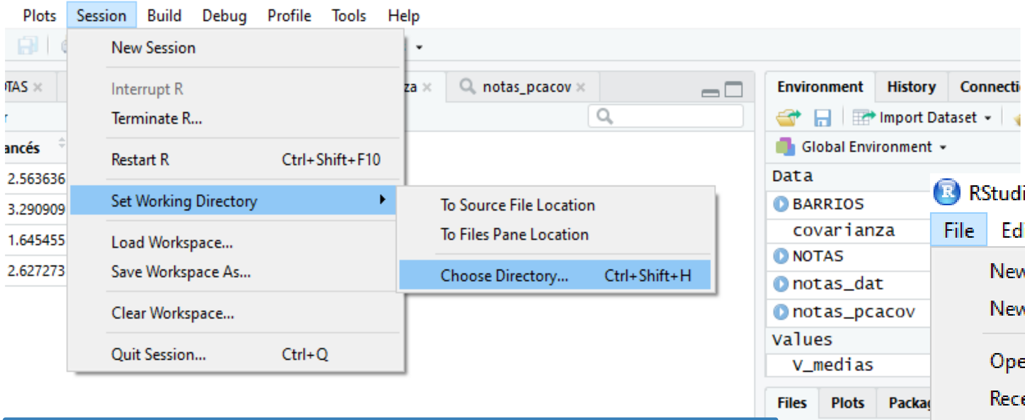
Partiendo bien de la **matriz de Correlaciones** ó de la **matriz de Varianzas-Covarianzas** estimadas, se hallará su descomposición, en función de sus valores propios y la matriz formada por sus autovectores correspondientes.

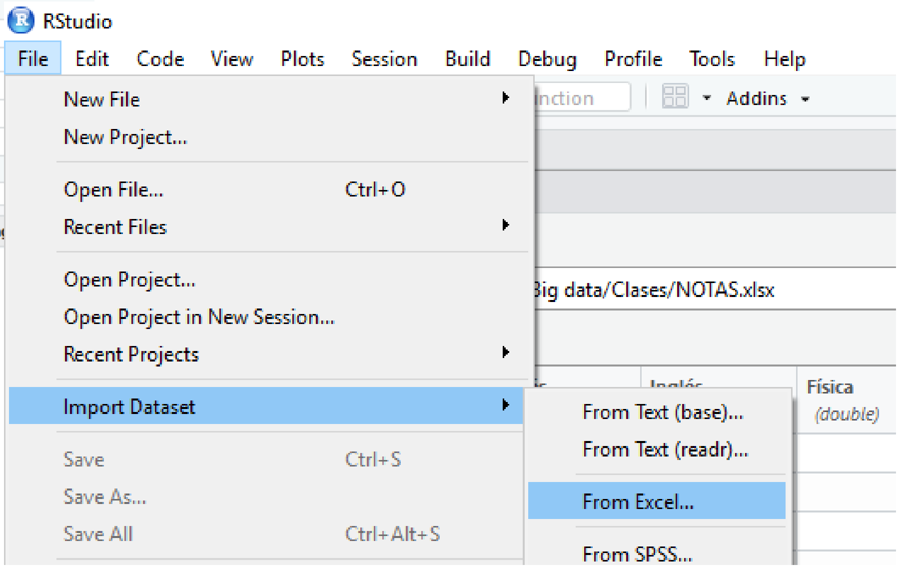
$$\mathbf{S} = \mathbf{P}\boldsymbol{\Lambda}\ \mathbf{P}^{-1} \qquad\qquad \mathbf{R} = \mathbf{P}_R\boldsymbol{\Lambda}_R\ \mathbf{P}_R^{\ -1}$$

Matriz de contiene los autovectores

Matriz diagonal formada por los autovalores

Importamos el fichero Excel con las calificaciones de los 11 alumnos





Para importar los datos desde Excel

tenemos dos opciones:

Si utilizamos Rmarkdown es obligatorio

incluir la sintaxis en el Notebook

```
# Importar la base de datos de NOTAS
NOTAS<- read_excel("C:/Users/reven/OneDrive/Desktop/Master Big data/Clases/ACP y Factorial/NOTAS.xlsx")
```

¿Cómo realizar un análisis de Componentes Principales con R?

Primero creamos un dataframe con todos los datos (datos) y posteriormente extraemos la primera columna que suele ser cualitativa y contiene la identificación de de las observaciones. De esta forma nos quedamos con una matriz con sólo las variables numéricas (notas_dat):

```
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

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

```
#Descriptivos
tabla1<-stat.desc(notas_dat,basic=FALSE)
knitr::kable(tabla1,  digits =2,caption = "Estadísticos descriptivos")
```

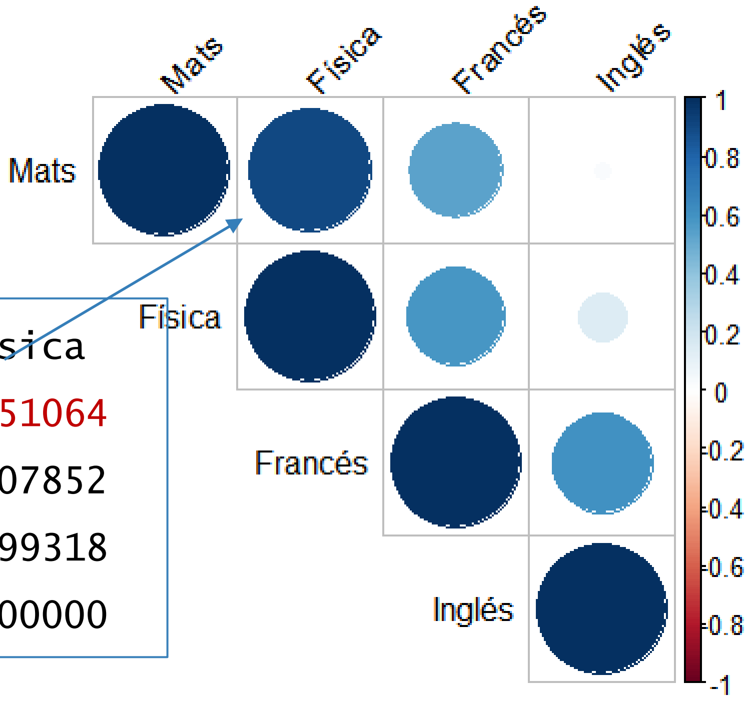
	Mats	Francés	Inglés	Física
median	5.00	5.00	6.00	4.00
mean	4.64	5.91	6.45	5.27
SE.mean	0.79	0.55	0.45	0.75
CI.mean.0.95	1.76	1.22	1.01	1.68
var	6.85	3.29	2.27	6.22
std.dev	2.62	1.81	1.51	2.49
coef.var	0.56	0.31	0.23	0.47

```
#Matriz correlacioenes
R<-cor(notas_dat, method="pearson")
knitr::kable(R,  digits =2,caption = "Correlaciones")
```

	Mats	Francés	Inglés	Física
Mats	1.00	0.54	0.02	0.91
Francés	0.54	1.00	0.60	0.58
Inglés	0.02	0.60	1.00	0.15
Física	0.91	0.58	0.15	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:

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" "contribuciones de las variables a cada Componente"

7 "$ind" Resultados para los individuos:

8 "indcoord" “Valores de los individuos en las CP"

9 "indcos2" " Calidad de la representación del ind“

10 "indcontrib" " Contribución de los individuos a la variabilidad de cada Componente“

11 "$call" "summary statistics" : Medidas estadísticas utilizadas para estandarizar

12 "$call$centre" " Media de las variables“

13 "$call$ecart.type" “Error estandar de las 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
```

λ_1

λ_2

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

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
```

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

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:

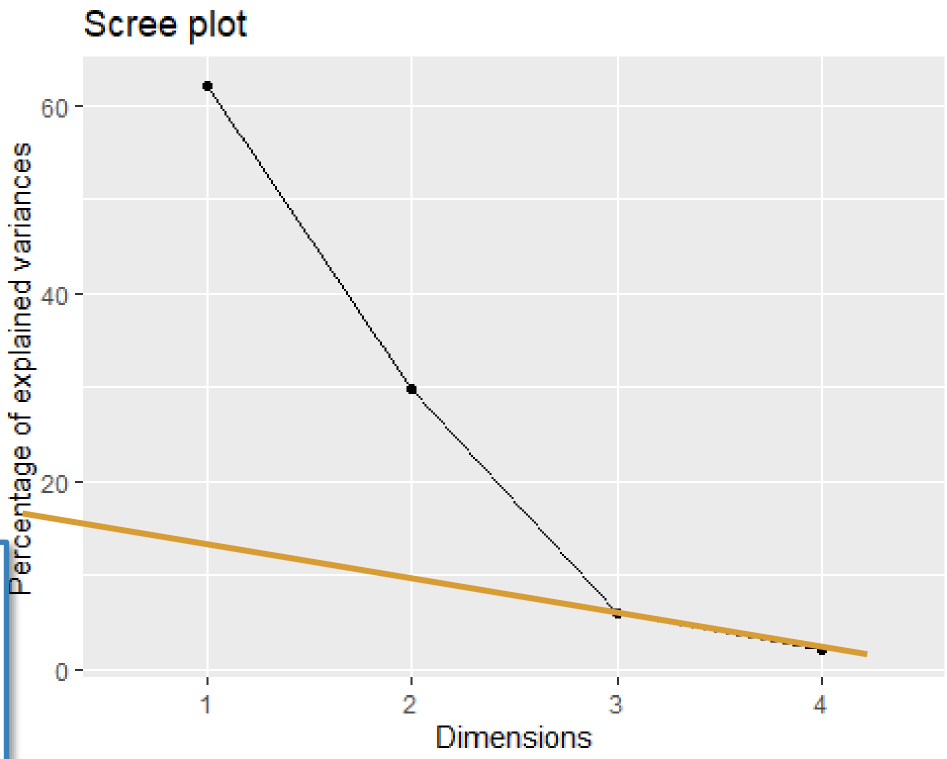
$$CP_1 = 0.548X_1^* + 0.538X_2^* + 0.283X_3^* + 0.573X_4^*$$

¿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
```

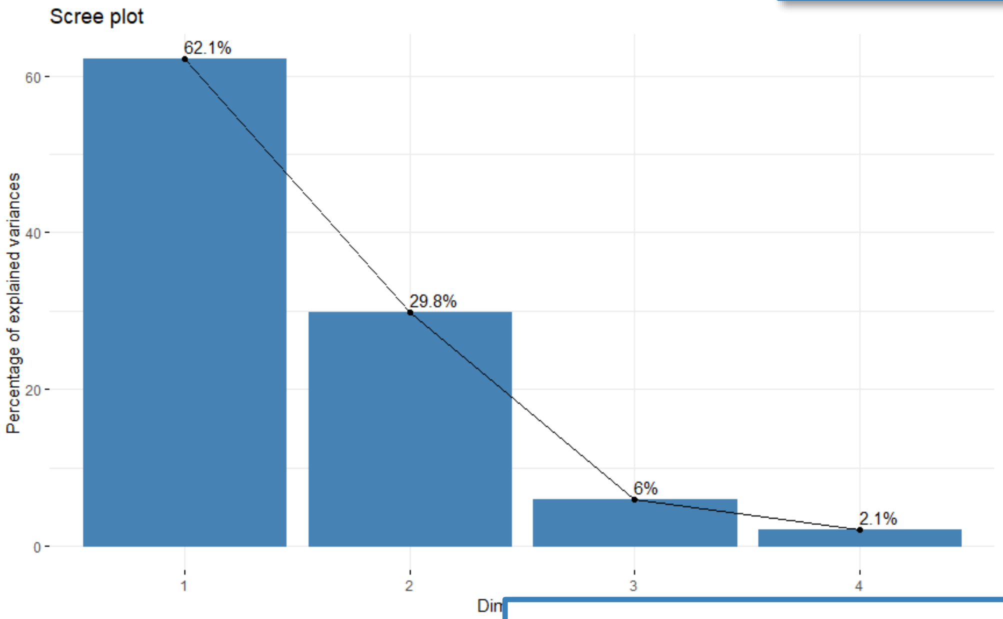
```
# Scree plot1
fviz_eig(fit, geom="line")+ theme_grey()
```



	eigen	p variance	cumulative
comp1	2.48576274	62.144068	62.14407
comp2	1.19280496	29.820124	91.96419
comp3	0.23856665	5.964166	97.92836
comp4	0.08286565	2.071641	100.00000

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

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




```
#====================================
# Decidimos hacer el análisis con 2 componentes
#====================================
fit<-PCA(notas_dat,scale.unit=TRUE, ncp=2, graph=TRUE)
```

1.2.3 RELACIÓN ENTRE LAS COMPONENTES PRINCIPALES Y LAS VARIABLES

Las **covarianzas entre cada componente principal y las variables** son el producto de los valores del autovector por el autovalor correspondiente:

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

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

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

```r
#Guardamos los estadísticos asociados a las variables en el objeto var
var<-get_pca_var(fit)
knitr::kable(var$cor,  digits =3,caption = "Correlaciones de la CP con las variables")
```

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

$$Corr(CP_1, X_1) = 0.548\frac{\sqrt{2.486}}{\sqrt{6.854}} = 0.865$$

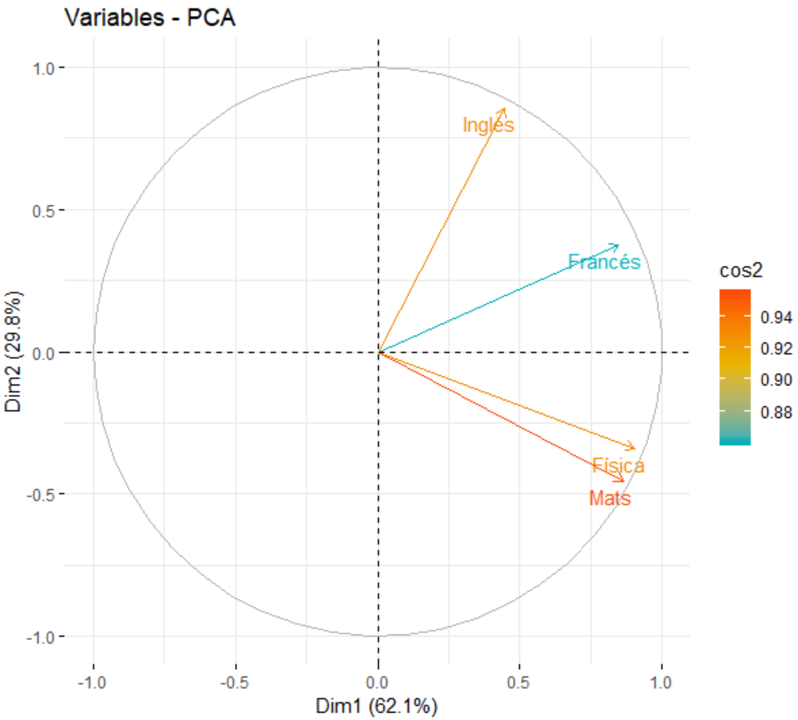
Correlaciones de la CP con las variables

	Dim.1	Dim.2
Mats	0.865	-0.454
Francés	0.849	0.374
Inglés	0.447	0.855
Física	0.904	-0.340

```r
#===================================
# Representación gráfica variables
#===================================
fviz_pca_var(fit, col.var="cos2",gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
repel = TRUE)
```

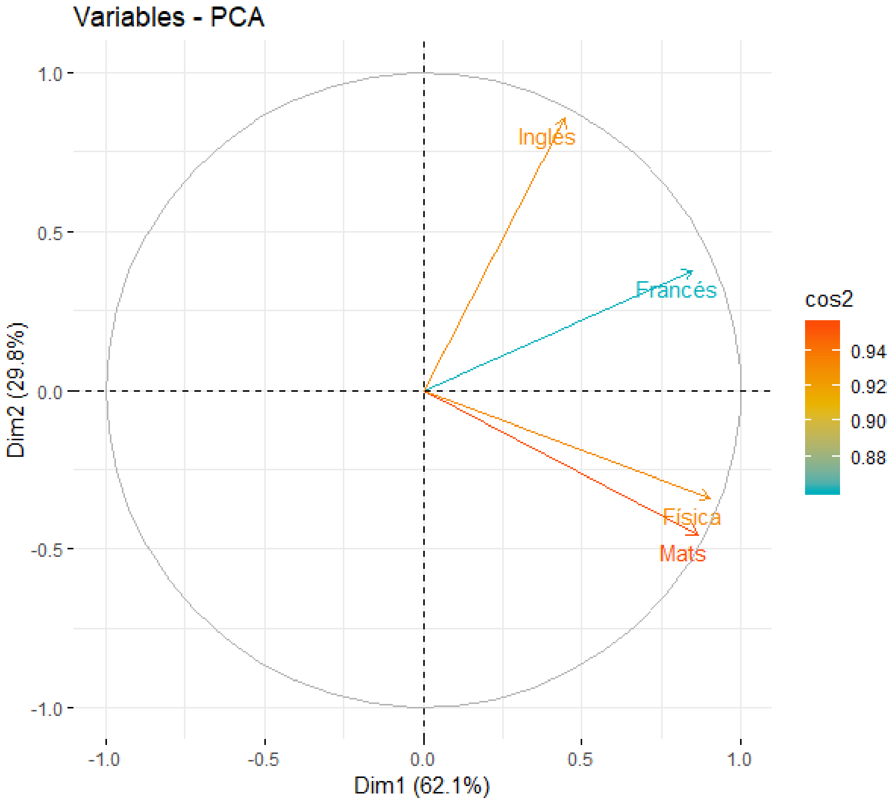
$$Corr(CP_i, X_j^*) = e_{ij}\sqrt{\lambda_i}$$

```
$var$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, Matemáticas y Francés, por lo que podríamos identificar dicha componente como la que representa las calificaciones de las asignaturas de ciencias.

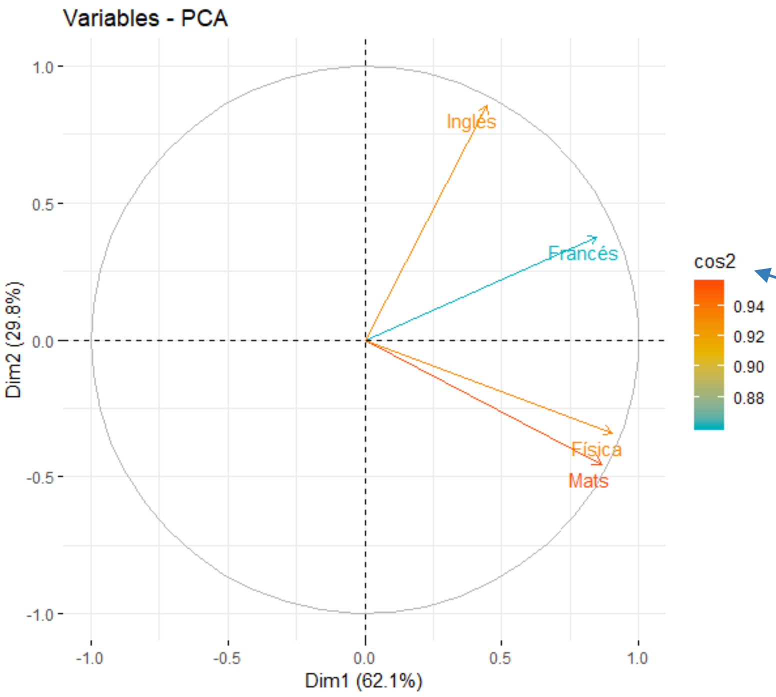


La segunda componente recoge sobre todo la información de la variable Inglés.

Fránces es la variable peor representada en estas dos componentes.

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:

```
knitr::kable(var$cos2,  digits =2,caption = "Cosenos al cuadrado")
```



	Dim.1	Dim.2
Mats	0.75	0.21
Francés	0.72	0.14
Inglés	0.20	0.73
Física	0.82	0.12

$$\mathrm{Cos}^2(CP_1, X_1) = Corr(CP_1, X_1)^2 = 0.865^2 = 0.75$$

$$\mathrm{Cos}^2(CP_2, X_1) = Corr(CP_2, X_1)^2 = -0.454^2 = 0.21$$

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

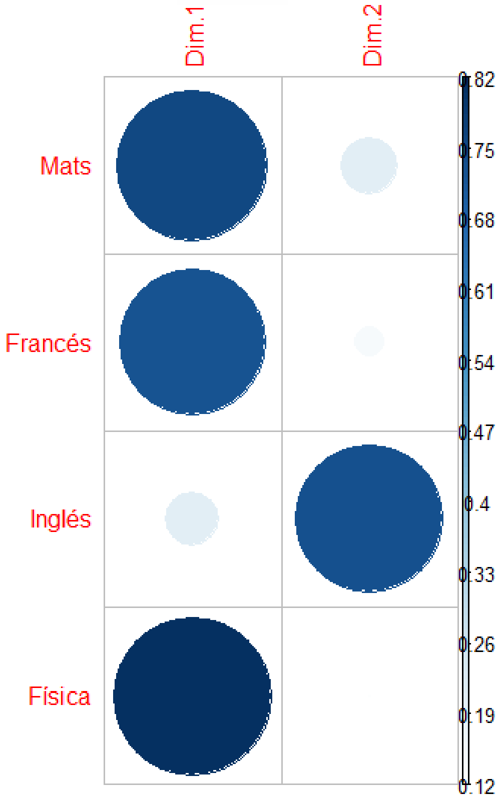
Correlacciones de la CP con las variables

	Dim.1	Dim.2
Mats	0.865	-0.454
Francés	0.849	0.374
Inglés	0.447	0.855
Física	0.904	-0.340

```
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

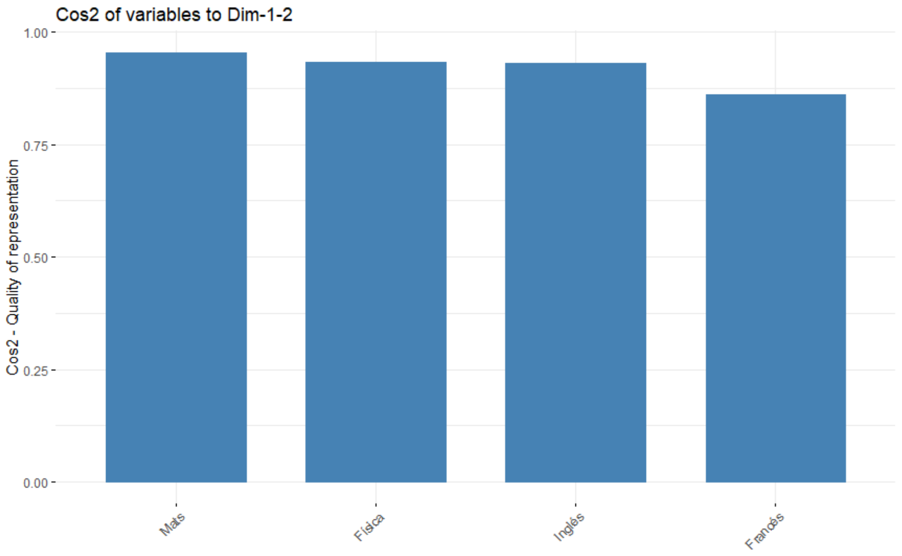
	Dim.1	Dim.2
Mats	0.75	0.21
Francés	0.72	0.14
Inglés	0.20	0.73
Física	0.82	0.12



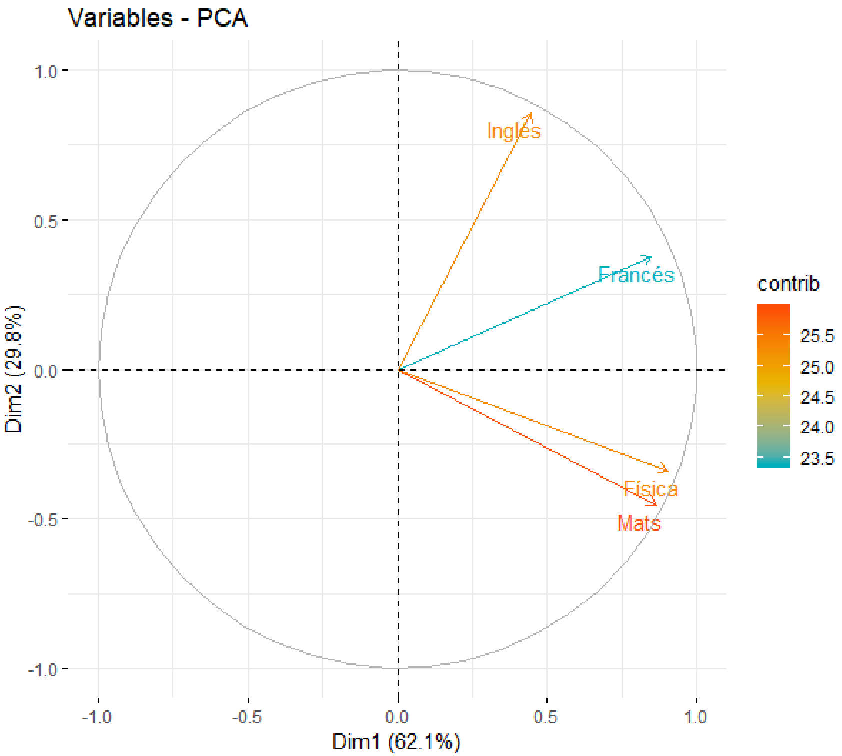
```
#Porcentaje de variabilidad explicada por las dos CP
fviz_cos2(fit,choice="var",axes=1:2)
```

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

	Dim.1	Dim.2
Mats	0.75	0.21
Francés	0.72	0.14
Inglés	0.20	0.73
Física	0.82	0.12



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
```

$$Contr(X_1, CP_1) = \frac{\mathrm{Cos}^2(X_1, CP_1)}{\sum_j^p \mathrm{Cos}^2(X_j, CP_1)} \cdot 100$$

	Dim.1	Dim.2
Mats	0.75	0.21
Francés	0.72	0.14
Inglés	0.20	0.73
Física	0.82	0.12

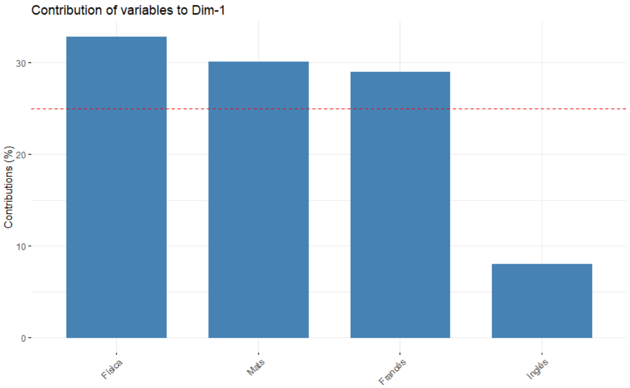
```r
#Contribución de las variables a la Componente 1
fviz_contrib(fit,choice="var",axes=1,top=10)
```

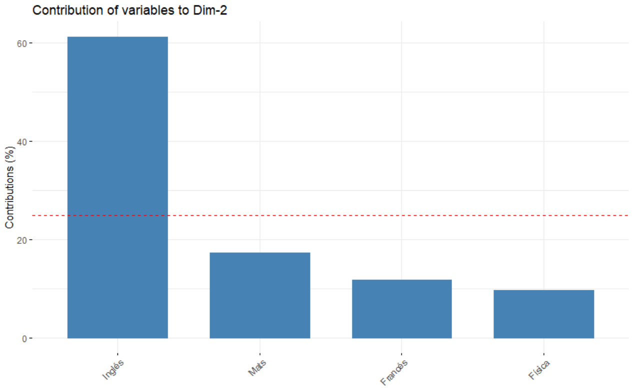
```r
fviz_contrib(fit,choice="var",axes=2,top=10)
```

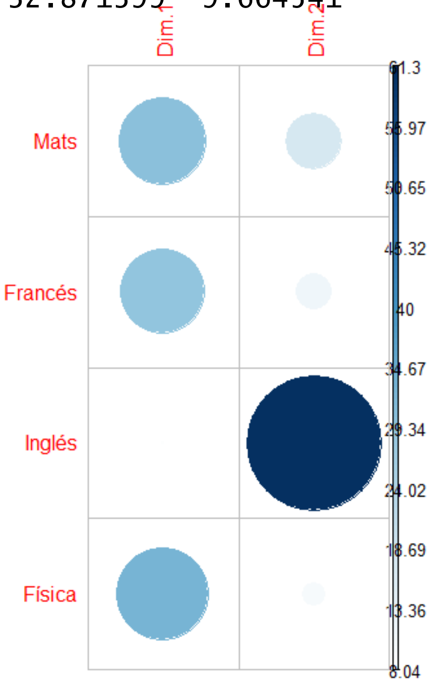
```r
corrplot(var$contrib,is.corr=FALSE)
```

varcontrib

	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)
```

1.2.4. COORDENADAS DE LOS INDIVIDUOS SOBRE LOS NUEVOS EJES.

¿Qué valores toman los individuos, observaciones, en las nuevas variables, CP?

```
ind<-get_pca_ind(fit)
knitr::kable(ind$coord,  digits =3,caption = "Valores de los individuos en las Cp")
```

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

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

Dim.1	Dim.2
-2.228	-0.268
-0.994	-1.411
2.250	1.875
-0.728	0.209
-2.294	0.372
1.716	0.128
0.655	-0.683
-0.645	1.101
0.827	-1.889
-0.865	1.267
2.305	-0.701

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

Puesto que las CP están construidas sobre las variables estandarizadas, los valores que toman las observaciones en las nuevas variables son puntuaciones alrededor de cero.

Dim.1	Dim.2
-2.228	-0.268
-0.994	-1.411
2.250	1.875
-0.728	0.209
-2.294	0.372
1.716	0.128
0.655	-0.683
-0.645	1.101
0.827	-1.889
-0.865	1.267
2.305	-0.701

Media de la CP1

$$\overline{cp_1} = \frac{\sum_{i=1}^{11} cp_{i1}}{11} = 0$$

Varianza de la CP1

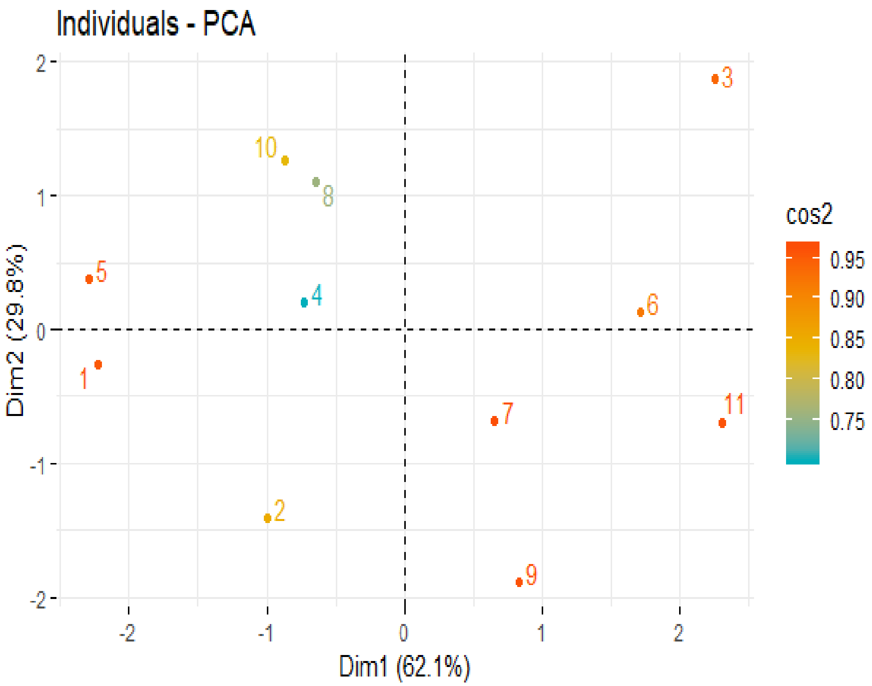
$$\frac{\sum_{i=1}^{11} cp_{i1}^2}{10} = \lambda_1 = 2.485$$

Valores negativos indican observaciones que están por debajo de la media.

El intervalo de valores que puede tomar la Componente 1 será aproximadamente $\left[-3\sqrt{\lambda_1}, 3\sqrt{\lambda_1}\right]$

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.

```r
fviz_pca_ind(fit, col.ind = "cos2",
              gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), repel = TRUE)
```

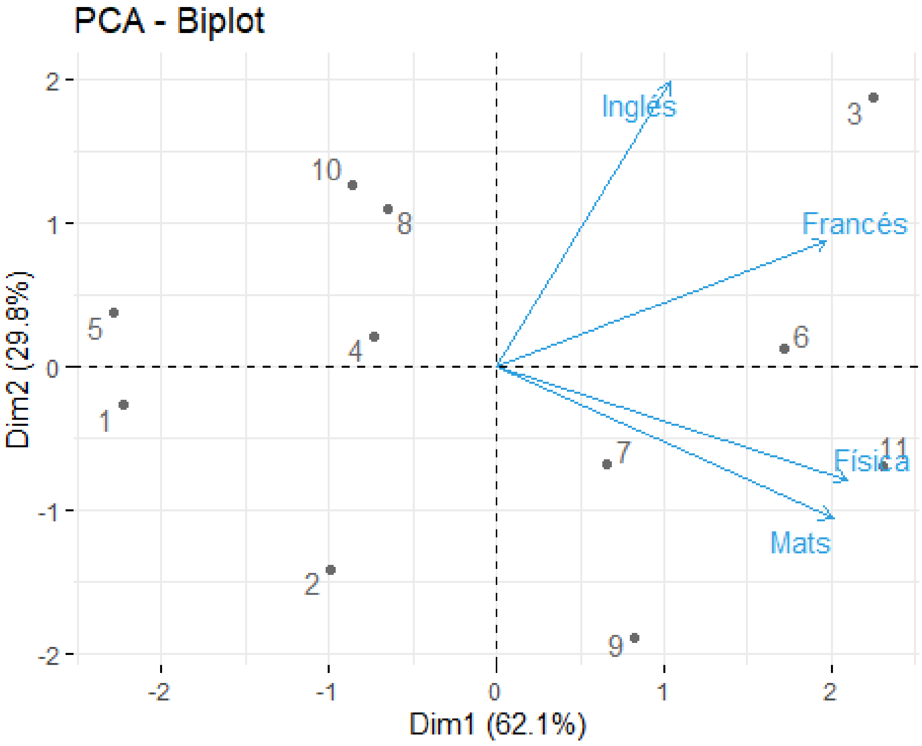


Dim.1	Dim.2
-2.228	-0.268
-0.994	-1.411
2.250	1.875
-0.728	0.209
-2.294	0.372
1.716	0.128
0.655	-0.683
-0.645	1.101
0.827	-1.889
-0.865	1.267
2.305	-0.701

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

Representación conjunta de las variables y los individuos en los planos de las componentes: Biplot

```
fviz_pca_biplot(fit, repel = TRUE, col.var = "#2E9FDF", # Variables color
                col.ind = "#696969")  # Individuals color
```



- ✓ Cuidado: Para representar de forma conjunta las variables y los individuos, la longitud de las flecha se ha modificado (no son las correlaciones) pero se mantienen los ángulos
- ✓ Los individuos próximos a una variable tienen valor alto de esa variable
- ✓ Los individuos en la misma dirección pero sentido opuesto tendrán un valor bajo de dicha variables.
- ✓ Ángulos agudos entre variables indican correlación positiva y cuanto menor ángulo mayor correlación. Si el ángulo es 90º las variables están incorreladas

Bibliografía

- An Introduction of Applied Multivariate Analysis with R. Everitt B, Hothorn T. Ed Wiley. 2011. Libro completo con explicaciones teóricas y ejemplos resueltos en R aunque con librerías básicas. Además aparecen otros temas relacionados como el análisis factorial
- Principal Components Analisys. Jolliffe, I.T. Ed Springer 2002 . Libro completo con explicaciones teóricas muy amplias y algunos ejemplos. El más citado por artículos de investigación.
- Nuevos Métodos de Análisis Multivariante. Cuadras C.M. 2014 . Libro completo con explicaciones teóricas y otras técnicas multivariantes
- Practical Guide to Principal Component Methods in R. A. Kassambara. Ed STHDA.com. 2017 . Libro completo que explica las librerías factominer y factoextra mediante ejemplos. Sin explicaciones teóricas
- Package ‘factoextra’ . Explicación del funcionamiento de la librería y la sintaxis detallada
- Package ‘factominer’ . Explicación del funcionamiento de la librería y la sintaxis detallada
- http://www.sthda.com/english/

UNIVERSIDAD
COMPLUTENSE
DE MADRID