MÉTODOS PARA REDUCIR LA DIMENSIÓN: ANÁLISIS DE COMPONENTES PRNCIPALES

Autor: Juana María Alonso Revenga

OBJETIVOS Y COMPETENCIAS A ALCANZAR.

- Extraer las Componentes Principales de una matriz de datos.
- Conocer las reglas que justifiquen el número de Componentes Principales a retener.
- Interpretar, cuando ello sea posible, las Componentes Principales.
- Obtener la proyección de las observaciones sobre cada componente principal
- Conocer la utilidad que tienen las Componentes Principales para utilizarse como entrada en otras técnicas estadísticas.
- Conocer las hipótesis del modelo Factorial.
- Conocer el grado de satisfacción obtenido desde el punto de vista estadístico del Análisis Factorial llevado a cabo.
- Conocer los fundamentos de la rotación de los Factores originales.
- Proyectar las observaciones sobre el plano Factorial.
- Representar gráficamente tanto las variables como las observaciones sobre el espacio Factorial obtenido

.1.1. INTRODUCCIÓN.

Cuando se recoge la información de una muestra de datos, lo más frecuente es tomar el mayor número posible de variables. Sin embargo, si tomamos demasiadas variables **es difícil visualizar relaciones** entre ellas.

Otro problema que se presenta es la **fuerte correlación** que muchas veces se presenta entre las variables: si tomamos demasiadas variables (cosa que en general sucede cuando no se sabe demasiado sobre los datos o sólo se tiene ánimo exploratorio), lo normal es que estén relacionadas o que midan lo mismo bajo distintos puntos de vista. Por ejemplo, en estudios médicos, la presión sanguínea a la salida del corazón y a la salida de los pulmones están fuertemente relacionadas.

Se hace necesario, pues, reducir el número de variables. Es importante resaltar el hecho de que el concepto de **mayor información se relaciona con el de mayor variabilidad o varianza**. Cuanto mayor sea la variabilidad de los datos (varianza) se considera que existe mayor información, lo cual está relacionado con el concepto de entropía.

Estas técnicas fueron inicialmente desarrolladas por Pearson a finales del siglo XIX y

posteriormente fueron estudiadas por Hotelling en los años 30 del siglo XX. Sin embargo, hasta la aparición de los ordenadores no se empezaron a popularizar.

Para estudiar las relaciones que se presentan entre p variables correlacionadas (que miden información común) se puede transformar el conjunto original de variables en otro (que no tenga repetición o redundancia en la información) **llamado conjunto de componentes principales.** Las nuevas variables son combinaciones lineales de las anteriores y se van construyendo según el orden de importancia en cuanto a la variabilidad total que recogen de la muestra. De modo ideal, se buscan m < p variables que sean combinaciones lineales de las p originales y que estén incorreladas, recogiendo la mayor parte de la información o variabilidad de los datos.

Si las variables originales están incorreladas de partida, entonces no tiene sentido realizar un análisis de componentes principales.

El análisis de componentes principales es una técnica matemática que no requiere la suposición de normalidad multivariante de los datos, aunque si esto último se cumple se puede dar una interpretación más profunda de dichos componentes.

El análisis factorial tiene por objeto explicar si un conjunto de variables observadas por un pequeño número de variables latentes, o no observadas, que llamaremos factores. Por ejemplo, supongamos que hemos tomado veinte medidas físicas del cuerpo de una persona: estatura, longitud del tronco y de las extremidades, anchura de hombros, peso, etc. Es intuitivo que todas estas medidas no son independientes entre sí, y que conocidas algunas de ellas podemos prever con poco error las restantes. Una explicación de este hecho es que las dimensiones del cuerpo humano dependen de ciertos factores, y si estos fuesen conocidos podríamos prever con pequeño error los valores de las variables observadas. Como segundo ejemplo, supongamos que estamos interesados en estudiar el desarrollo humano en los países del mundo, y que disponemos de muchas variables económicas, sociales y demográficas, en general dependientes entre sí, que están relacionadas con el desarrollo. Podemos preguntarnos si el desarrollo de un país depende de un pequeño número de factores tales que, conocidos sus valores, podríamos prever el conjunto de las variables de cada País. Como tercer ejemplo, supongamos que medimos con distintas pruebas la capacidad mental de un individuo para procesar información y resolver problemas. **Podemos preguntarnos si existen unos factores, no directamente observables, que explican el conjunto de resultados observados**. El conjunto de estos factores será lo que llamamos inteligencia y es importante conocer cuántas dimensiones distintas tiene este concepto y cómo caracterizarlas y medirlas. El análisis factorial surge impulsado por el interés de Karl Pearson y Charles Spearman

en comprender las dimensiones de la inteligencia humana en los años 30, y muchos de sus avances se han producido en el área de la psicometría.

El análisis factorial está relacionado con los componentes principales, pero existen ciertas diferencias. En primer lugar, **las componentes principales se construyen para explicar las varianzas, mientras que los factores se construyen para explicar las covarianzas o correlaciones entre las variables.** En segundo lugar, **componentes principales es un herramienta descriptiva, mientras que el análisis factorial presupone un modelo estadístico formal** de generación de la muestra dada.

Según nuestro objetivo final y las características de los datos de que dispongamos, será adecuado aplicar el Análisis de Componentes Principales, que es en realidad un caso particular del Análisis Factorial, y que por motivos pedagógicos estudiamos en primer lugar, o el Análisis Factorial , de más amplios objetivos y de mayor exigencia con respecto a los datos de partida.

1.2. ANÁLISIS DE COMPONENTES PRINCIPALES.

La idea básica del método de Análisis de Componentes Principales es describir la variación de un conjunto de datos multivariante en términos de un conjunto incorrelado de variables, cada una de las cuales es una combinación lineal particular de las variables originales. Estas nuevas variables, denominadas **Componentes Principales**, se obtienen en orden decreciente a su importancia, es decir:

- **la primera Componente Principal recoge la máxima información** (variación) de los datos originales,
- **la segunda Componente se elige de forma que recoja la mayor cantidad de información, que no haya sido recogida por la primera Componente, es decir, recogerá información de las variables iniciales que sean incorreladas con la primera Componente Principal**, y así sucesivamente.

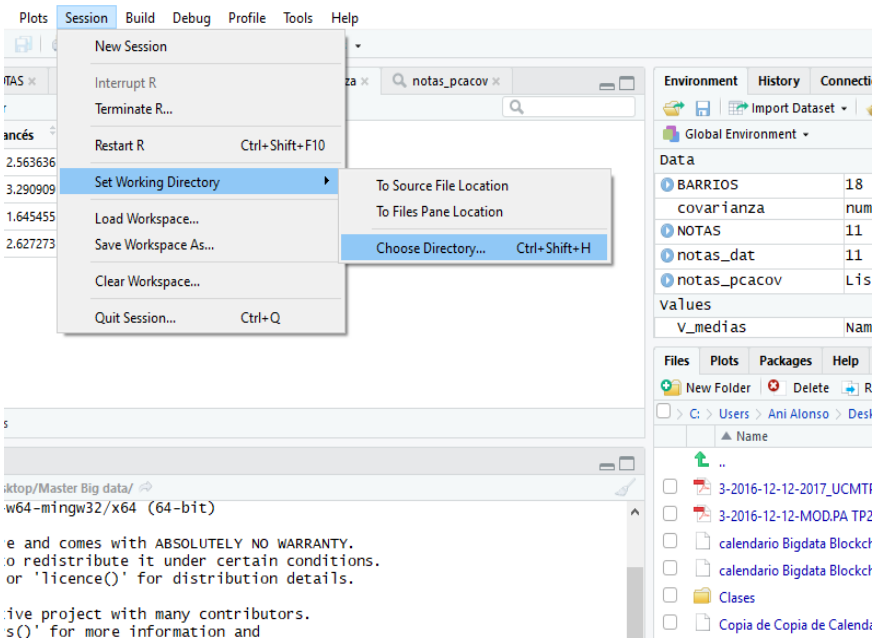
El objetivo de este análisis es ver si las primeras Componentes Principales recogen la mayor parte de la variación de los datos originales. Si esto es así, dichas Componentes se pueden utilizar para resumir los datos con la mínima pérdida de información. Esto dará lugar a importantes simplificaciones en los análisis posteriores que se puedan llevar a cabo. Al examinar este nuevo conjunto de variables se pueden obtener respuestas a preguntas importantes dentro del análisis de datos como: existencia de datos atípicos, multicolinealidad entre las variables, agrupación de observaciones en subgrupos y comprobar hipótesis de distribución normal multivariante

EJEMPLO:

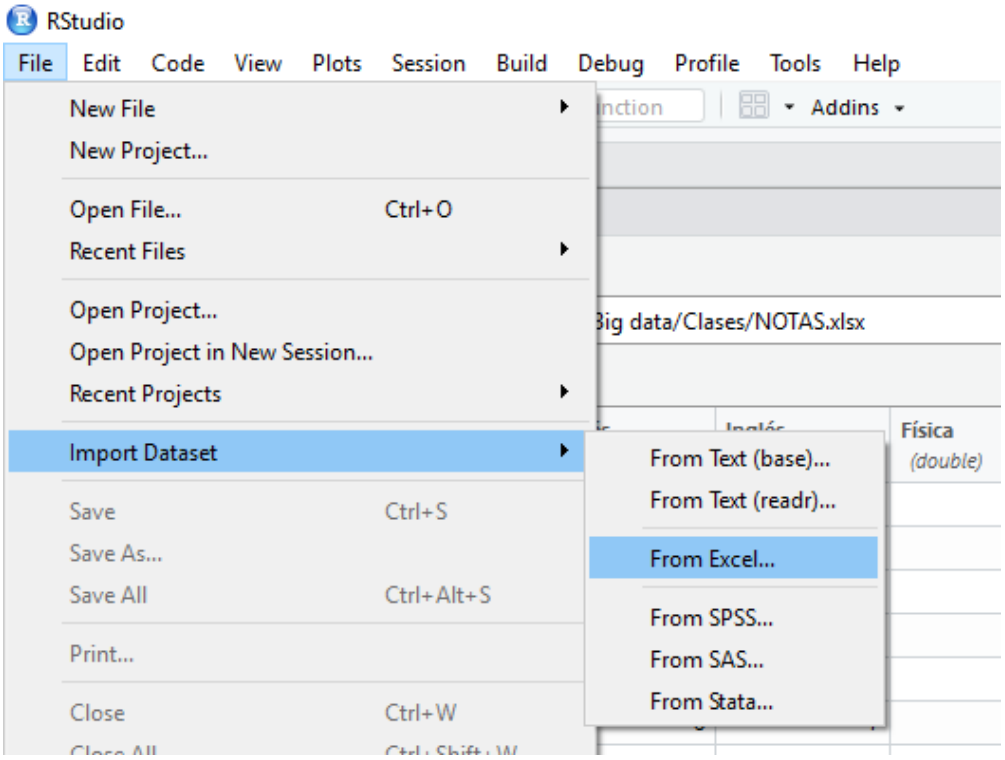
Se desea analizar las características de un grupo de alumnos con el objetivo de asignarles un índice que les oriente en cuanto a sus facultades para proseguir sus estudios. Para ello, se consideran las calificaciones obtenidas en Matemáticas, Francés , Inglés y Física. Los datos de partida son los siguientes:

Obs	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

Para comenzar la sesión en R elegimos la carpeta en donde vamos a crear nuestros conjuntos de datos



Importamos los datos desde Excel



A continuación cargamos las librerías que vamos a necesitar.

```
install.packages("ggplot2")
install.packages("factoextra")
install.packages("pastecs")
install.packages("psych")
install.packages("FactoMineR")
install.packages("corrplot")
install.packages("paran")
install.packages("nFactors")
#============================
# Cargamos librerías
#============================
library(ggplot2)
library(factoextra)
library(psych)
library(FactoMineR)
library(pastecs)
library(corrplot)
library(paran)
library(nFactors)
```

1.2.1. ¿CÓMO SE OBTIENEN LAS COMPONENTES PRINCIPALES?.

Desde un punto de vista algebraico las Componentes Principales son una combinación particular de p variables aleatorias. Geométricamente, estas combinaciones lineales representan la selección de un nuevo sistema de coordenadas obtenido mediante la rotación del sistema original. Los nuevos ejes representan las direcciones con máxima variabilidad y producen una descripción más simple y ordenada de la estructura de la matriz de Varianzas-Covarianzas.

Sea la variable $X = \left(X_1, X_2, ..., X_p\right)^T$ cuyo vector de medias es $\mu_X = E\left[X\right] = \left(\mu_1, \mu_2, ..., \mu_p\right)^T$ y matriz de varianzas-covarianzas $\Sigma = E\left[(X - \mu_X)(X - \mu_X)^T\right]$

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

Luego $V\left(X_i\right) = \sigma_{ii} = E\left[(X_i - \mu_i)^2\right]$ y $COV\left(X_i, X_j\right) = \sigma_{ij} = E\left[(X_i - \mu_i)(X_j - \mu_j)\right]$.

Supongamos que disponemos de un conjunto de datos, observaciones de la variable X, dispuesto de forma matricial (con n observaciones y p variables):

$$\mathrm{X} = \begin{pmatrix} x_{11} & x_{12} & . & . & x_{1p} \\ x_{21} & x_{22} & . & . & x_{2p} \\ . & . & . & . & . \\ . & . & . & . & . \\ x_{n1} & x_{n2} & . & . & x_{xp} \end{pmatrix}$$

El estimador del vector de medias será $\overline{X} = \left(\overline{x}_1, \overline{x}_2, ..., \overline{x}_p\right)^T$ donde

$$\overline{x}_i = \frac{\sum_{k=1}^{n} x_{ik}}{n}$$

y el estimador de la matriz de varianzas-covarianzas

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

Donde $S_{ii} = \dfrac{\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)^2}{n}$ y $S_{ij} = \dfrac{\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{n}$.

A partir de la matriz S se puede calcular la matriz de correlaciones muestrales

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

Donde $r_{ij} = \dfrac{S_{ij}}{\sqrt{S_{ii}}\sqrt{S_{jj}}}$.

Estos estimadores se pueden calcular utilizando notación matricial, como indicamos a continuación.

Expresiones matriciales.

Los estadísticos anteriores, $\overline{\mathbf{X}}$, S, R, se pueden escribir en función de la matriz de datos, X, según las siguientes expresiones matriciales**:**

Vector de medias muestrales: $\overline{\mathbf{X}} = \dfrac{1}{n}\mathbf{X}'\mathbf{1}_n$

Matriz de Covarianzas muestral: $\mathrm{S} = \dfrac{1}{n-1}\mathrm{X}'\left(\mathrm{I}_n - \dfrac{1}{n}1_n 1_n'\right)\mathrm{X}$

Matriz de Correlaciones muestrales: $\mathbf{R} = \mathbf{D}^{-1/2}\,\mathbf{S}\,\mathbf{D}^{-1/2}$

donde : $\mathbf{D} = \begin{pmatrix} s_{11} & 0 & \dots & 0 \\ 0 & s_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & s_{pp} \end{pmatrix}$

Existe la siguiente relación entre ellas:

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

Luego, partiendo bien de la matriz de Correlaciones o 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.

$$\mathbf{S} = \mathbf{P}\mathbf{\Lambda}\ \ \mathbf{P}^{-1} \quad \text{o bien} \quad \mathbf{R} = \mathbf{P}_R\mathbf{\Lambda}_R\ \ \mathbf{P}_R^{-1} \quad .$$

Donde $\mathbf{P}$ es la matriz de contiene los autovectores y $\mathbf{\Lambda}$ una matriz diagonal formada por los autovalores Según se ha explicado en el apartado anterior, la primera Componente Principal, será aquella variable cuyo vector dirección coincide con el del autovector correspondiente al mayor autovalor de la matriz de partida, sea la de Correlaciones o la de Varianzas-Covarianzas. **Designaremos a partir de este momento por $\mathbf{e}_\alpha$ a los autovectores.** **Así las Componentes principales se construyen como combinación lineal de las variables originales estandarizadas con coeficientes determinados por los autovectores.**

$$CP_j = e_{1j}X_1^* + e_{2j}X_2^* + ... + e_{pj}X_p^*$$

Para realizar un análisis de componentes principales con R seguiremos los siguientes pasos:

Utilizaremos las librerias Factoextra y factominer, que se resumen en los siguientes esquemas.

FactoMineR & factoextra

Analyzing & Visualizing Multivariate Data

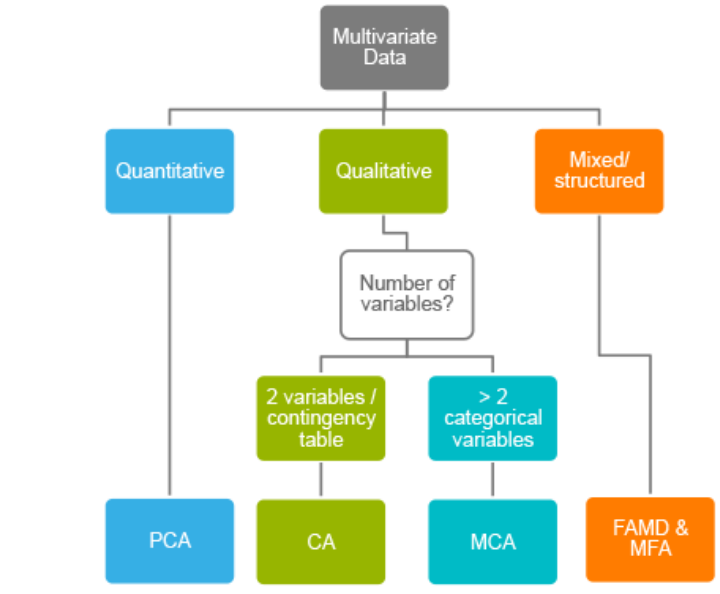
- Performs PCA, (M)CA, FAMD, MFA, HCPC & more
- Provides the coordinates, the quality of representation and the contribution of individuals & variables
- Predicts the results for supplementary individuals & variables

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

Dentro de la librería factoextra se encuentra el análisis de Componentes Principakes, PCA.

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

Además existen funciones que nos permiten visualizar mediante gráficos el resultado obtenido con el PCA.

Functions	Description
fviz_eig (or fviz_eigenvalue)	Visualize eigenvalues.
fviz_pca	Graph of PCA results.
fviz_cos2	Visualize element cos2.[1]
fviz_contrib	Visualize element contributions.[2]

Además también dispone de funciones que nos permiten extraer información del análisis de PCA realizado.

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

Continuando con el EJEMPLO anterior, sobre el fichero con las notas de 11 alumnos que habíamos importado de Excel.

Primero creamos un Data Frame con los datos de las variables:

```
datos<- as.data.frame(NOTAS)

rownames(datos)<-datos[,1]

notas  dat<-datos[,-1]
```

Es importante crear así el Data Frame para tener la primera columna alfanumérica con el identificador del individuo que más tarde utilizaremos para las representaciones gráficas. Además creamos la matriz notas_dat con solo los valores numéricos de las variables necesaria como entrada para el procedimiento PRINCOMP

Calculamos los descriptivos unidimensionales

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

Tenemos que los valores de los estadísticos descriptivos de cada una de las las variables son:

	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

Recordamos que siempre es conveniente realizar el análisis de Componentes principales sobre la matriz de correlaciones en lugar de sobre la matriz de varianzas-covarianzas.

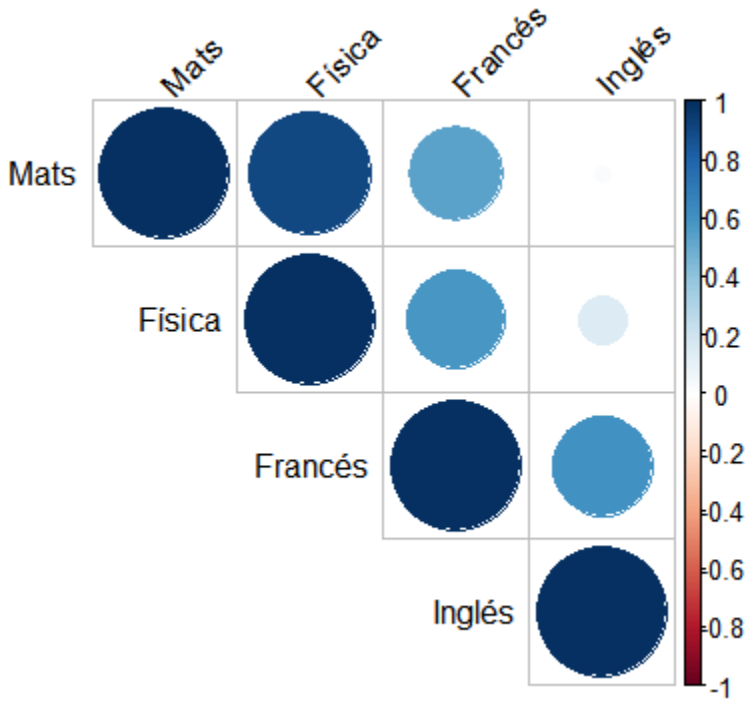
Para comprobar si las variables están relacionadas de forma lineal entre ellas calculamos la matriz de correlaciones y la guardamos en la matriz R

```
#Matriz correlaciones
Library(corrplot)
R<-cor(notas_dat, method="pearson")
```

	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

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)
```



La sintaxis básica del procedimiento de R para realizar un análisis de componentes principa*les*

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

La opción `scale. unit=TRUE` , significa que calculamos los autovalores de la matriz d e correlaciones, porque las variables son estandarizadas

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

Recordamos que las variables estandarizadas tienen media cero y varianza 1. Así la suma de las varianzas de esta nuevas variables es igual al número de variables p

`Ncp=4` nos indica el número de componentes que se van a calcular, normalmente un número inferior al de variables que tenemos.

El resultado del análisis de componentes principales se guardará en el fichero de **datos** fit. Si queremos ver el contenido de este fichero haremos

```
head(fit)
```

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

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"

Veamos cada uno de los elementos desglosados

1 "$eig" Autovalores

```
$`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
```

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

Los autovectores asociados a dichos autovalores nos dan los coeficientes de la combi nación 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
```

Así la Componente 1 será:

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

1.2.3. DETERMINACIÓN DEL NÚMERO DE COMPONENTES PRINCIPALES

Como se indicó con anterioridad uno de los objetivos del Análisis de Componentes Principales consiste en detectar la verdadera dimensión del problema. Esto provoca la necesidad de tomar una decisión ¿cuántas Componentes debemos retener?. Realmente no existe una respuesta que mecanice de algún modo la decisión. Existen varios métodos que ayudan a elegir el número de Componentes Principales a retener (m)

1. Se elige una proporción de variabilidad a explicar que se considere suficiente (cuando son encuestas reales puede ser suficiente el 70%, si son pruebas de laboratorio se debería ser más exigentes –en torno al 90-95%).
2. *Método de Cattell*. Realización de una gráfica para los autovalores. Esta gráfica se construye al representar los puntos: $(1,\lambda_1)$, $(2,\lambda_2)$,...,(p,λ_p). Cuando los puntos de la gráfica se nivelan, estos autovalores suelen estar suficientemente cercanos a cero como para poder despreciarlos. Se puede trazar una recta que aglutine en su entorno a los autovalores más pequeños y todos los que queden por encima corresponderían a las Componentes Principales retenidas.

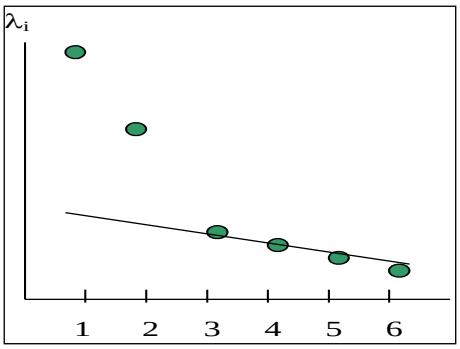
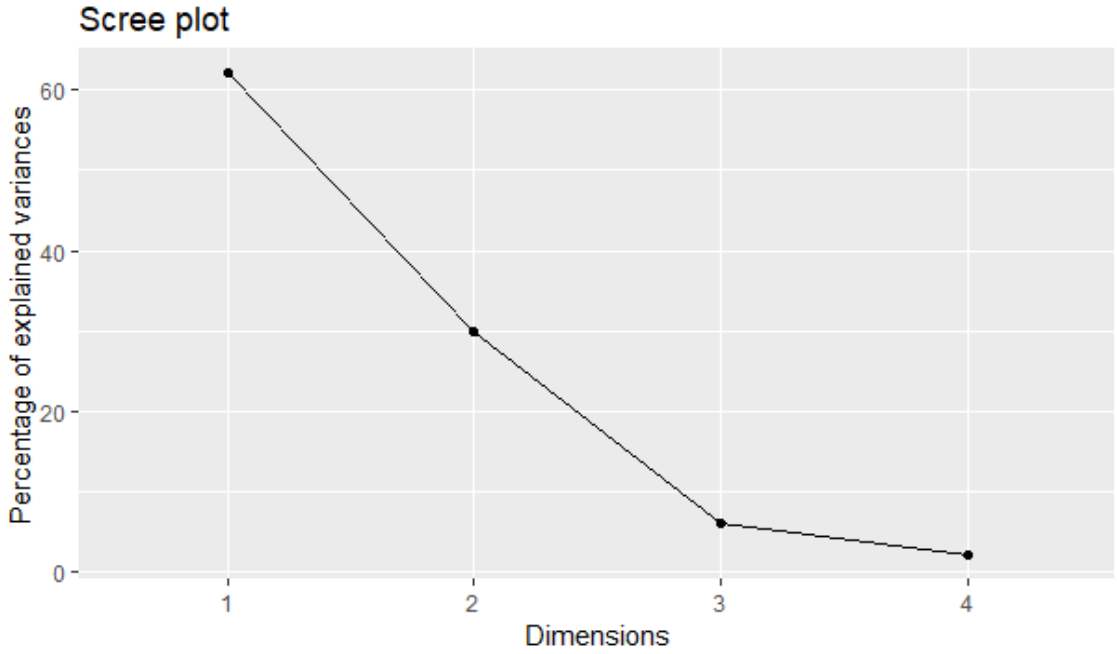


Figura 2.2. Ejemplo de Scree Plot de Catell

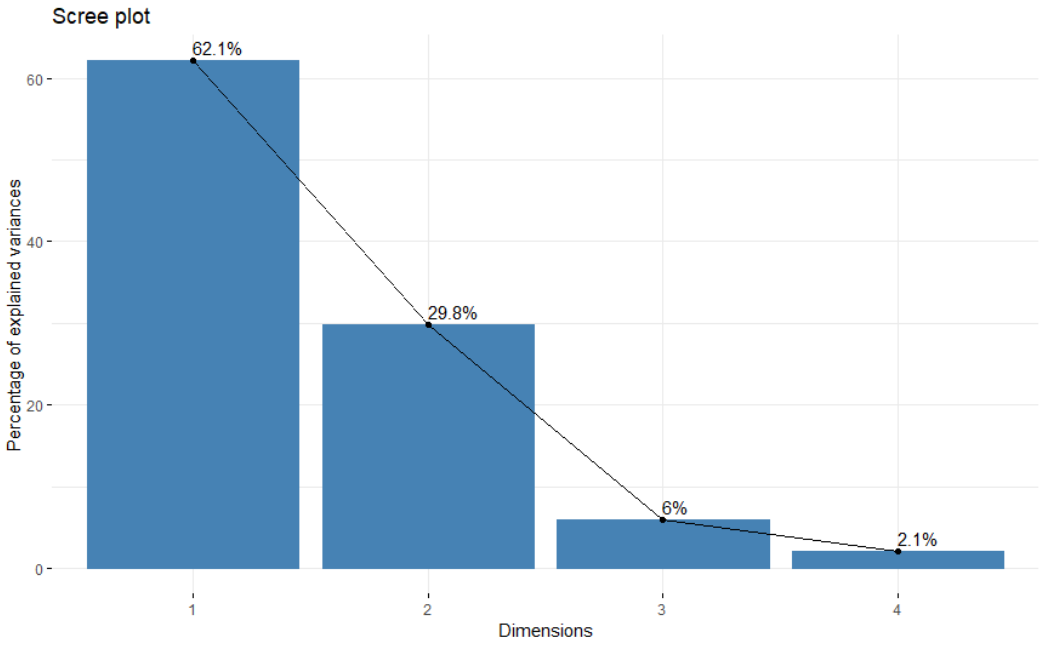
Por ejemplo en la Tabla que recoge la información relativa a los autovalores parece adecuado retener las 2 primeras Componentes Principales, debido a que los dos primeros autovalores explican el 91.96% de la variabilidad.

```
$`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 plot
Library(factoextra)
fviz_eig(fit, geom="line")+
 theme_grey()
```




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



Por lo que decidiríamos quedarnos con 2 de las cuatro componentes principales que explicarían el 91,96% de la varianza de las variables iniciales. Los dos autovalores asociados a las componentes seleccionadas valen 2.485 y 1.192, y en términos de porcentaje explican el 62.14% de la variabilidad y el 29.82%.

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 el valor propio:

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

La correlación entre un componente principal y una variable X es proporcional al coeficiente de esa variable en la definición del componente, y el coeficiente de proporcionalidad es el cociente entre la desviación típica del componente y la desviación típica de la variable.

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

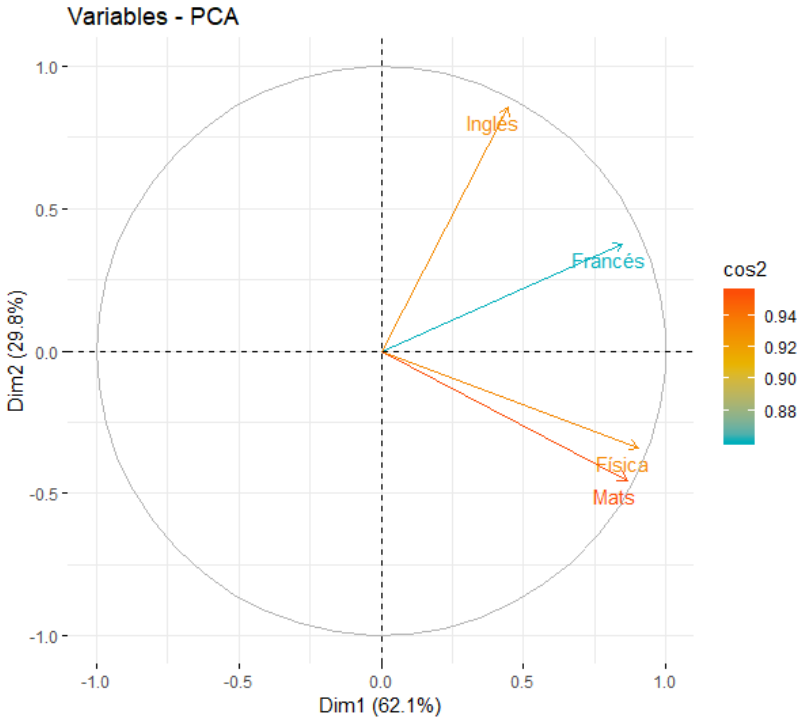
```
$var$cor
            Dim.1      Dim.2       Dim.3       Dim.4
Mats     0.8652256 -0.4541383  0.04890811  0.20676317
Francés  0.8486830  0.3742755 -0.37244303 -0.03067974
Inglés   0.4469671  0.8551036  0.25849529  0.04688842
Física   0.9039386 -0.3395278  0.17504577 -0.19228851
```

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

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

La siguiente representación gráfica nos puede ayudar a entender como están recogidas nuestras variables iniciales en las nuevas componentes, puesto que representa el coeficiente de correlación entre las variables y las nuevas componentes.

```
#==================================
# 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))
```



Por ejemplo 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

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

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

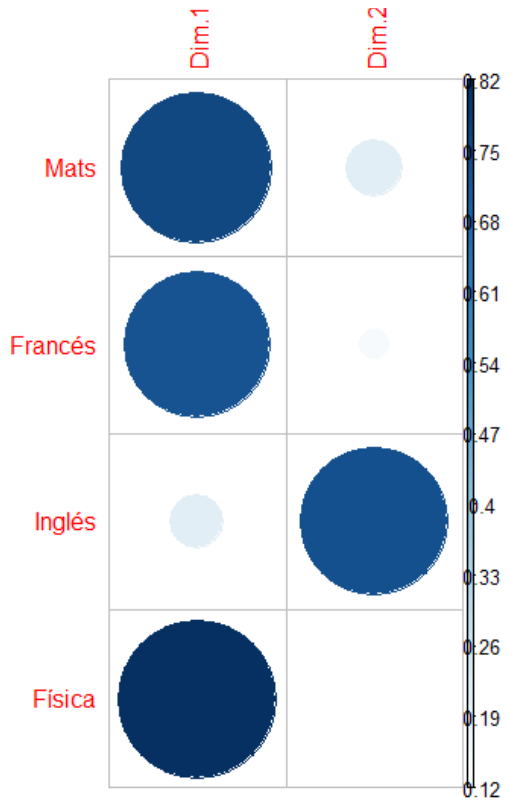
```
$var$cor
            Dim.1      Dim.2
Mats    0.8652256 -0.4541383
Francés 0.8486830  0.3742755
Inglés  0.4469671  0.8551036
Física  0.9039386 -0.3395278
```

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

```
#==================================
# 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



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

```
$var$contrib
            Dim.1     Dim.2     Dim.3     Dim.4
Mats    30.116124 17.290471  1.002656 51.590749
Francés 28.975525 11.743927 58.144677  1.135871
Inglés   8.036952 61.301061 28.008868  2.653119
Física  32.871399  9.664541 12.843799 44.620261
```

1.2.4. COORDENADAS DE LOS INDIVIDUOS SOBRE LOS NUEVOS EJES.

Una vez seleccionadas las direcciones de los ejes que explican la máxima variabilidad, pasamos a situar las observaciones sobre los nuevos ejes luego tenemos que calcular sus nuevas coordenadas. Las coordenadas de los n puntos individuos sobre el eje correspondiente a $\mathbf{e}_\alpha$ son las n componentes del vector

$$\psi_\alpha = \mathbf{X}\mathbf{e}_\alpha,$$

OBSERVACIÓN: A la hora de calcular las coordenadas, en realidad se emplea la matriz X, en la que $x_{ij} - \bar{x}_j$, si el Análisis se realiza a partir de la Matriz de Varianzas-Covarianzas. Si el A.C.P. parte de la matriz de Correlaciones, para obtener las coordenadas de los individuos la matriz X es aquella en que $\dfrac{x_{ij} - \bar{x}_j}{s_j}$.

Como la matriz de datos de partida está centrada (hemos restado su media a cada componente), es sencillo comprobar que la suma de las coordenadas de los n individuos sobre cada eje es nula.

$$\sum_{i=1}^{n} \Psi_{\alpha i} = \sum_{i=1}^{n}\sum_{j=1}^{p} \frac{\left(x_{ij} - \bar{x}_j\right)}{S_j} e_{\alpha j} = \sum_{j=1}^{n} e_{\alpha j} \sum_{i=1}^{p} \frac{\left(x_{ij} - \bar{x}_j\right)}{S_j} = 0 \quad \forall \alpha$$

Por lo que la media de las n coordenadas es cero y su varianza será:

$$\Psi_\alpha' \Psi_\alpha = \mathbf{e}_\alpha' \mathbf{X}' \mathbf{X} \mathbf{e}_\alpha = \lambda_\alpha$$

*Continuando con el **EJEMPLO** de las notas de los alumnos en las cuatro asignaturas, tenemos en el siguiente cuadro, las coordenadas de los 11 alumnos en las Componentes Principales.*

```
$ind
$ind$`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
```

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

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

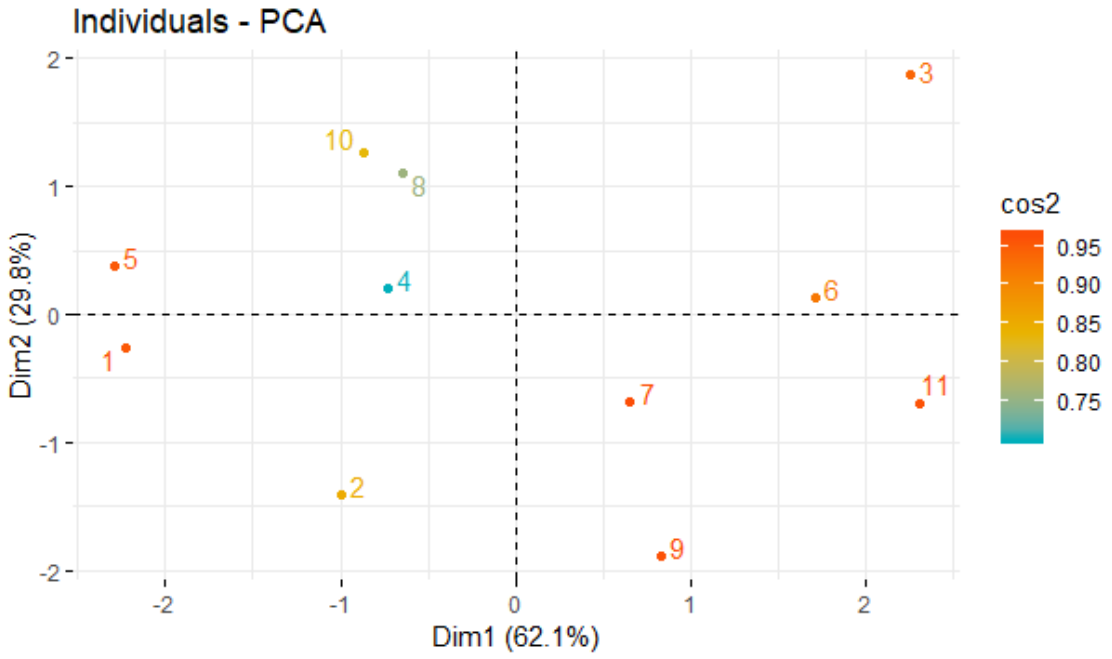
Además, la media de las variables Ψ_1, Ψ_2, Ψ_3 *y* Ψ_4*, vale 0. Se puede comprobar que:*

$$\text{Var}(\Psi_1) = \frac{\sum_{i=1}^{11} y_{i1}^2}{10} = \lambda_1 = 2.485.$$

Análogamente, para el resto de variables Ψ_2, Ψ_3 y Ψ_4.

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))
```




```
$ind
```

```
$ind$`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
```

1.2.5. COMENTARIOS FINALES AL ANÁLISIS DE COMPONENTES PRINCIPALES.

- El Análisis de Componentes Principales sobre Σ es apropiado en aquellos casos en donde todas las variables están medidas por lo menos en unidades comparables teniendo varianzas semejantes. Si las variables no se miden en las mismas unidades cualquier cambio en la escala de medición en alguna de las variables tendrá efecto sobre las Componentes Principales. Este cambio de escala puede invertir los papeles de las variables importantes.
- Por otro lado si una de las variables tiene una varianza mucho mayor que las demás dominará la primera componente principal, sin importar la estructura de las covarianzas de las variables, y, en este caso no tiene sentido práctico la realización de un Análisis de Componentes Principales diagonalizando la citada matriz de Varianzas-Covarianzas.
- Cuando no parezca que las variables se estén midiendo en unidades semejantes, se aplica el análisis a la matriz de correlaciones. En muchos casos se aplica el análisis sobre esta última matriz pero no siempre está justificado.
- Se puede dar la circunstancia de que alguno de los autovalores (bien de la matriz de Varianzas-Covarianzas, bien de la de Correlaciones) sea 0. Esto significa que algunas de las variables originales serán linealmente dependientes de las otras. Así si retenemos m Componentes Principales, en esencia estamos suponiendo que hay p-m autovalores iguales a 0 (realmente no significativamente diferentes de 0), y sus autovectores correspondientes definen p-m restricciones lineales, linealmente independientes, sobre las variables observables. Las ecuaciones derivadas de las restricciones lineales (relaciones estructurales) proporcionan

información acerca de cómo están relacionadas las variables entre sí. Nos ayudan a detectar relaciones de multicolinealidad entre un conjunto de variables predictoras en los problemas de regresión.

- A veces los investigadores buscan un subconjunto de variables originales que contengan, en algún sentido, virtualmente toda la información disponible en el conjunto de datos completo (con todas las variables). Existen muchos métodos de elección de las citadas variables y todos ellos se encuentran relacionados con las Componentes Principales. Uno de tales métodos consiste en primer lugar en determinar las Componentes que se retienen por alguna de las técnicas enumeradas en el punto 2.2.4. A continuación se selecciona la variable cuyo coeficiente sea el mayor en valor absoluto.
- Aunque cuando se usa el Análisis de Componentes Principales en la práctica el mayor interés suele centrarse en las Componentes con mayor varianza, hay ocasiones donde el examen de las dos últimas Componentes elegidas puede ser de utilidad. Por ejemplo estas Componentes pueden identificar observaciones atípicas que pueden provocar el incremento ficticio de la dimensión. Estos valores atípicos se pueden obtener representando las observaciones en el espacio generado por las dos últimas Componentes y observando la existencia de puntos aislados, los cuales serán identificados como atípicos.
- Por último, hay que tener en cuenta que las Componentes Principales no tienen que ser "variables interpretables". El objetivo verdadero del A. C. P. no es conseguir variables interpretables, sino detectar la auténtica dimensión del problema.

1.2.5 SISTEMÁTICA DEL ANÁLISIS DE COMPONENTES PRINCIPALES.

En este apartado se presenta el esquema de los pasos a seguir para realizar un Análisis de Componentes Principales. Estos pasos son los siguientes:

1) Elección de las variables que se pretende estudiar.
2) Depuración de los datos, lo que conlleva tratamiento de valores faltantes y eliminación de atípicos.
3) Estudiar la matriz de Correlaciones para comprobar que existen muchas variables con correlacciones grandes, próximas a 1 en valor absoluto.
4) Obtención de autovalores y autovectores.
5) Determinación del número de Componentes a retener. Se pueden tener los siguientes criterios:

Como se diagonaliza la matriz de Correlaciones:

- autovalores mayores que 1, autovalores mayores que 0.7, utilizar la gráfica de Catell, proporción de variabilidad explicada > 0.7.

6) Interpretación, si procede, de las Componentes a través de los coeficientes de los autovectores respectivos.
7) Obtención de las puntuaciones de los individuos en el espacio de las Componentes.
8) Representar las variables en el espacio de las Componentes. Esto permitirá crear asociaciones entre variables.
9) Representar 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/