

Practica 2

Analisis de Componentes principales

01 de abril de 2017

Contents

1	Introducción	1
2	Ejercicio [1]	2
2.1	Carga de datos	2
2.2	Descripción de los datos	3
2.3	Gráfico I	4
2.4	Componentes principales	5
2.5	Gráfico II	8
2.6	Caso tipificado	9
2.7	Conclusiones	10
3	Ejercicio [2]	10
3.1	Carga y descripción de los datos	11
3.2	Cambio de unidades	13
3.3	Caso tipificado	14
3.4	Reducción de la dimensión	14
3.5	Otros gráficos	17
3.6	Conclusiones	22
4	Ejercicio [3]	23
4.1	Carga y descripción de los datos	23
4.2	Componentes principales	26
4.3	Gráfico de pendiente	27
4.4	Autovectores	28
4.5	Importancia de componentes	31
4.6	Diagrama de barras	34
4.7	Conclusiones	35

1 Introducción

Las prácticas pueden verse en formato web en la dirección:

xvilan.github.io/practicasADM

El código puede descargarse en formato Rmarkdown y volver a replicarse en:

<https://github.com/xvilan/practicasADM>

Se compone de tres ejercicios en los que se realiza un análisis de componentes principales con tres conjuntos de datos (**ventas**, **scktp**, **dowjones**).

En estadística, el **Análisis de Componentes Principales** (en español *ACP*, en inglés, *PCA*) es una técnica utilizada para reducir la dimensionalidad de un conjunto de datos.

Técnicamente, el ACP busca la proyección según la cual los datos queden mejor representados en términos de mínimos cuadrados. Convierte un conjunto de observaciones de variables posiblemente correlacionadas en un conjunto de valores de variables sin correlación lineal llamadas componentes principales.

Los objetivos del Análisis de Componentes Principales son:

- Resumir la variación total en una **dimensión menor**.
- Identificar fuentes de variación **interpretables**.
- Explorar características **inesperadas** de los datos.

El ACP se emplea sobre todo en análisis exploratorio de datos y como herramienta de procesado inicial de los datos. El ACP comporta el cálculo de la descomposición en autovalores y autovectores de la matriz de covarianza, normalmente tras centrar los datos en la media de cada atributo. Las componentes principales están basadas en combinaciones lineales de los datos originales centrados $X - \bar{x}$

Nota

Esta web es un R Notebook. Se ha desarrollado con el lenguaje Rmarkdown en **R-Studio**, que utiliza *knitr* y *pandoc* para exportar el código a varios formatos: *html*, *pdf*, *doc* y *md*. Es posible descargar el archivo *.Rmd* desde github y ejecutar los scripts mediante R-Studio. Para ello es necesario tener instalados los paquetes:

```
install.packages("rmarkdown")
install.packages("installr")
install.packages("pandoc")
```

2 Ejercicio [1]

En este ejercicio se hace una descripción pormenorizada del proceso de ACP. En los sucesivos las explicaciones se simplificarán empleando sólo las funciones directas. Los resultados se redondearán, de manera general, a cuatro decimales.

2.1 Carga de datos

Se carga el archivo **ventas.txt** a R-Studio, indicando previamente a R-Studio cuál es el directorio de trabajo en el que estará localizado el archivo.

```
ventas = read.table('ventas.txt');
show(ventas);

##           ventas beneficios activos
## G.M.        26.7      3.3    15.8
## Exxon       38.4      2.4    19.5
## Ford        19.2      1.7     8.4
## Mobil       20.6      1.0     8.2
## Texaco      18.9      0.9     9.4
## Std.Oil     14.8      1.0     7.6
## IBM         19.0      2.7    12.6
## Gulf        14.2      0.8     7.3
## G.E.        13.7      1.1     5.9
## Chrysler    7.7      0.2     2.9
```

A continuación, se realiza una exploración del contenido del dataset, como por ejemplo, consultar el número de filas y columnas o el tipo de tabla de datos creada. El número de filas, se consulta empleando la función “*nrow*”, y el tipo de archivo con la función “*class*”.

```

n = nrow(ventas);
tipo_de_tabla = class(ventas);
n;

## [1] 10

tipo_de_tabla;
## [1] "data.frame"

También pueden utilizarse otras funciones como “str”, que devuelve una descripción de la estructura de la tabla,

```

```

str(ventas)

## 'data.frame':   10 obs. of  3 variables:
## $ ventas      : num  26.7 38.4 19.2 20.6 18.9 14.8 19 14.2 13.7 7.7
## $ beneficios: num  3.3 2.4 1.7 1 0.9 1 2.7 0.8 1.1 0.2
## $ activos    : num  15.8 19.5 8.4 8.2 9.4 7.6 12.6 7.3 5.9 2.9

```

El archivo **ventas.txt** contiene valores de tres variables: ventas, beneficios y activos, para 10 empresas de los sectores petrolero, automóvil y tecnológico.

El ejercicio nos indica que se utilizarán sólo las dos primeras columnas, por tanto se trata de un **análisis de dos variables**.

```

(ventas = read.table('ventas.txt')[,1:2]);

##      ventas beneficios
## G.M.      26.7      3.3
## Exxon     38.4      2.4
## Ford      19.2      1.7
## Mobil     20.6      1.0
## Texaco    18.9      0.9
## Std.Oil   14.8      1.0
## IBM       19.0      2.7
## Gulf      14.2      0.8
## G.E.      13.7      1.1
## Chrysler  7.7      0.2

```

2.2 Descripción de los datos

El primer paso en el ACP es crear el vector de **medias muestrales m** . El vector de medias muestrales, dada una matriz de datos $X_{n \times p} = (x_{ij})$, se describe como:

$$\bar{x} = \frac{1}{n} X^T 1_n = (\bar{x}_j : j = 1, \dots, p)$$

En R puede utilizarse la función “*colMeans*” para calcular la media de las columnas.

```

m = round(colMeans(ventas), digits = 4);
round(m, digits = 4);

##      ventas beneficios
##      19.32      1.51

```

La matriz de **covarianzas muestrales S** , se define como:

$$S = \frac{1}{n} X^T X - \bar{x} \bar{x}^T = (s_{jk} : j = 1, \dots, p)$$

y en R se calcula con la función “*cov*”,

```

S = cov(ventas);
round(S, digits = 4);

##           ventas beneficios
## ventas      70.4107     5.8731
## beneficios  5.8731     0.9699

```

A continuación, se calcula la **Variación Total**, utilizando la función $VT = \text{traza}(S)$, que es la suma la diagonal principal de la matriz de covarianzas muestrales S ,

```

(vt.ventas = round(sum(diag(S)), digits = 4));
## [1] 71.3806

```

Nota

Para el cálculo posterior de las variables tipificadas, es interesante calcular la raíz de la diagonal principal de matriz de covarianzas, de forma que se obtienen además las desviaciones típicas de las variables.

```

round(sqrt(diag(S)), digits = 4);

##           ventas beneficios
##           8.3911     0.9848

```

El siguiente paso es calcular la **matriz de correlación R** , cuya diagonal principal será la unidad.

Si $D = \text{diag}(S)$ y $d = \text{diag}(D^{-1/2})$ entonces la matriz de correlaciones se define como:

$$R = D^{-1/2} \times S \times D^{-1/2} = S \times dd^T = (r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}s_{kk}}})$$

y en R se calcula mediante la función “*cor*”,

```

R = cor(ventas);
round(R, digits = 4);

##           ventas beneficios
## ventas      1.0000     0.7107
## beneficios  0.7107     1.0000

```

Se puede comprobar que existe una correlación positiva de 0.7 entre ambas variables.

2.3 Gráfico I

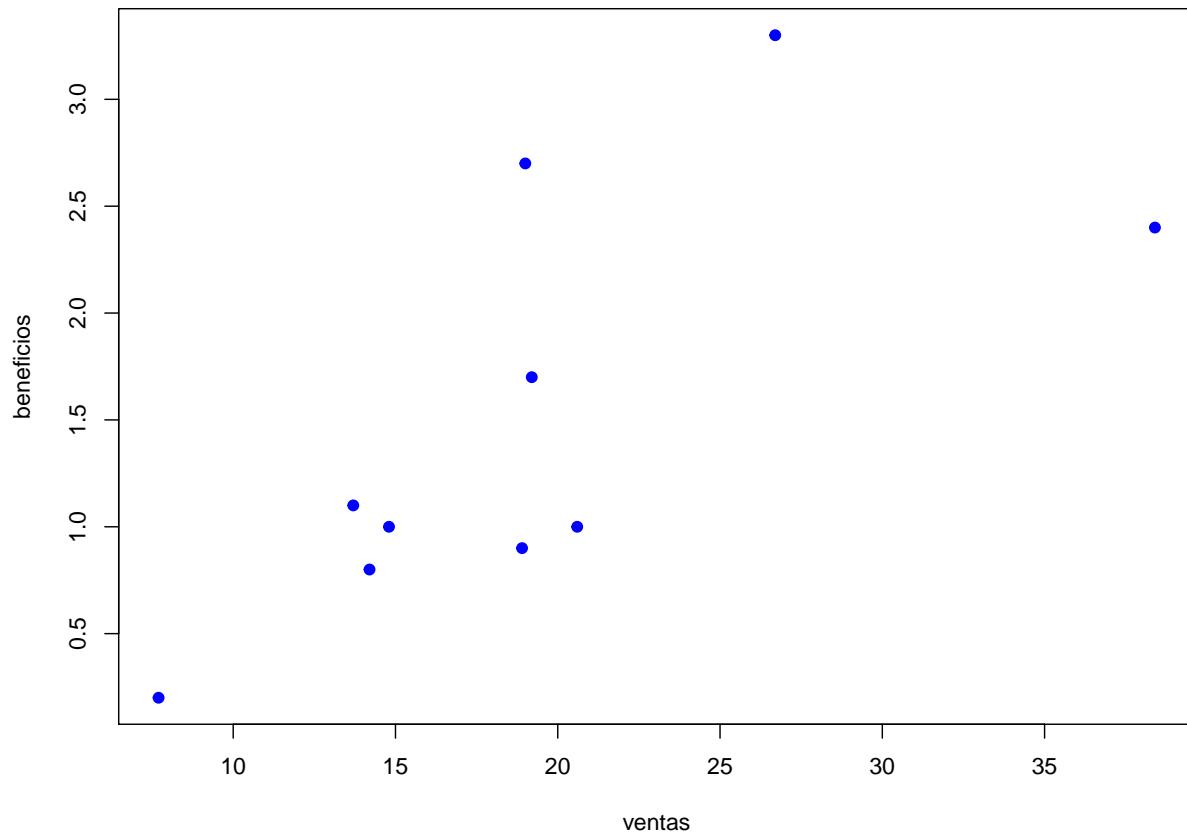
En este apartado se genera un gráfico de tipo **Matriz de Nube Puntos** o *scatter plot matrix* usando la función “*plot*” con una serie de parámetros opcionales, que configuran la salida gráfica.

```

plot(ventas,pch = 19,cex = 1,upper.panel = NULL, col = 4,
main = 'Datos de Ventas',cex.main = 1.4);

```

Datos de Ventas



2.4 Componentes principales

El software R-Studio incluye en el paquete *stats* la función “*princomp*”, que se utiliza en ACP. Toma como argumento la matriz de datos y devuelve un vector con siete elementos como respuesta:

Elemento	Descripción
\$sdev	Las desviaciones estándar de las componentes principales.
\$loadings	Matriz de autovectores y matriz de proporciones.
\$center	Vector de medias muestrales de la matriz datos.

Elemento	Descripción
\$scale	Desv. típicas o las escalas aplicadas a cada variable (por defecto 1).
\$n.obs	El número de observaciones.
\$scores	Matriz de componentes principales (si es calculable).
\$call	Llamada al nombre de la matriz de datos.

Puede consultarse la ayuda de esta función usando:

```
help("princomp");
```

Alinvocarse la función de “*princomp*”, muestra todos los elementos en la respuesta:

```
pc = princomp(ventas);
pc[1:7];

## $sdev
##      Comp.1    Comp.2
## 7.9883356 0.6549767
##
## $loadings
##
## Loadings:
##           Comp.1   Comp.2
## ventas      -0.996
## beneficios     -0.996
##
##           Comp.1   Comp.2
## SS loadings     1.0    1.0
## Proportion Var  0.5    0.5
## Cumulative Var 0.5    1.0
##
## $center
##      ventas beneficios
##      19.32      1.51
##
## $scale
##      ventas beneficios
##      1            1
##
## $n.obs
## [1] 10
##
## $scores
##           Comp.1    Comp.2
```

```

## G.M.      -7.5039104 -1.16611703
## Exxon    -19.0875511  0.70985412
## Ford      0.1036787 -0.19937585
## Mobil     -1.2328300  0.61532930
## Texaco    0.4695753  0.57271203
## Std.Oil   4.5468245  0.12994951
## IBM       0.2192909 -1.21260524
## Gulf      5.1614570  0.27903624
## G.E.      5.6345973 -0.06175449
## Chrysler 11.6888678  0.33297143
##
## $call
## princomp(x = ventas)

```

Puede accederse a cada elemento utilizando el operador (\$). Por ejemplo, para el cálculo de **varianzas** (σ^2) puede utilizarse el resultado de las desviaciones estandard elevado al cuadrado:

```

var= pc$sdev^2;
round(var, digits = 4);

## Comp.1  Comp.2
## 63.8135  0.4290

```

Generando la variable **Sc** puede utilizarse en un análisis equivalente de la Covarianza, donde,

$$S_c = \frac{n}{n-1} * S.$$

```

n=4
Sc = 1 - (1/n);

```

de forma que se obtiene la varianza equivalente y, en consecuencia, el vector de **Autovalores** ($L_{p \times 1}$):

```

var.eq = (1/Sc)*var;
round(var.eq, digits = 4);

## Comp.1  Comp.2
## 85.0847  0.5720

```

La matriz de coeficientes de los **Autovectores** **G** puede extraerse del objeto *pc*, donde, $G = (g_1, g_2, \dots, g_p)$

```

G = with(pc, loadings)[,];
round(G, digits = 4);

##           Comp.1  Comp.2
## ventas      -0.9965  0.0837
## beneficios -0.0837 -0.9965

```

Y para obtener la **matriz de componentes principales** o *scores*,

$$Y = (X - 1_n \bar{x}^T) \times G,$$

```

scores = data.frame(with(pc, scores));
show(scores);

##           Comp.1  Comp.2
## G.M.      -7.5039104 -1.16611703
## Exxon    -19.0875511  0.70985412
## Ford      0.1036787 -0.19937585
## Mobil     -1.2328300  0.61532930
## Texaco    0.4695753  0.57271203
## Std.Oil   4.5468245  0.12994951
## IBM       0.2192909 -1.21260524

```

```

## Gulf      5.1614570  0.27903624
## G.E.      5.6345973 -0.06175449
## Chrysler 11.6888678  0.33297143

```

Es necesario cambiar al signo tanto de G como de $scores$, ya que este es devuelto con el signo cambiado. Es habitual en la práctica mostrar los *pesos* de estos coeficientes como positivos en la primera componente,

```

G = -G;
round(G, digits = 4);

##          Comp.1  Comp.2
## ventas    0.9965 -0.0837
## beneficios 0.0837  0.9965

```

El significado de estos *pesos* o coeficientes sería:

```

Comp.1 = 0.9965 x ventas + 0.0837 x beneficio
Comp.2 = -0.0837 x ventas + 0.9965 x beneficio

```

Estos coeficientes son, además, los senos y los cosenos del ángulo de rotación entre los ejes de los CP y los ejes de las variables.

Nota

Un ejercicio adicional, sería comprobar que la matriz de componentes principales cumple con las propiedades del teorema de ACP,

- El vector de **medias muestrales** $m.cp$ (\bar{y}) es igual a cero.
- La matriz de **covarianzas muestrales** $S.cp$ (S_y) tiene como diagonal el vector de autovalores.
- Las componentes son **incorreladas de media cero y decreciente de variabilidad**.

```

(m.cp = round(colMeans(scores), digits = 4));

## Comp.1 Comp.2
##      0      0

(S.cp = round(cov(scores), digits = 4));

##          Comp.1  Comp.2
## Comp.1 70.9039  0.0000
## Comp.2  0.0000  0.4767

```

Otra comprobación a realizar sobre la matriz de componentes principales es, calcular su *Variación Total*, que también debe coincidir con la VT de los datos:

```

(VT.cp = round(sum(diag(cov(scores))), digits = 4));
## [1] 71.3806

```

2.5 Gráfico II

Se genera un gráfico de **Matriz de nube puntos** (*scatter plot matrix*) de la matriz de componentes principales. Para ello, se crea una variable r , que es el rango o *dimensión* de la matriz, y se utiliza la función *plot*.

```

r = with(scores, range(Comp.1));
plot(scores,

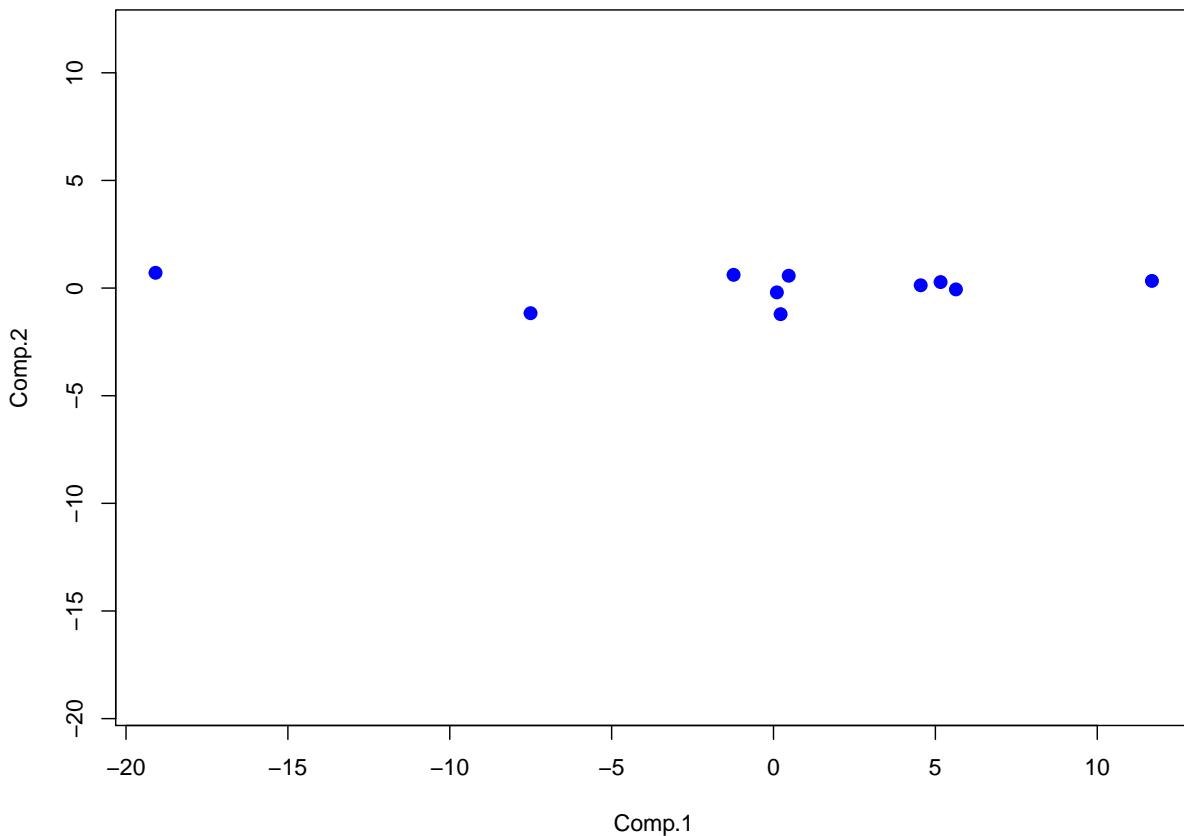
```

```

pch = 19,
cex = 1.2,
upper.panel = NULL,
col = 4 ,
main = 'ACP de Ventas',
cex.main = 1.3,
xlim = r,
ylim = r);

```

ACP de Ventas



Puede observarse en el gráfico, que prácticamente todos los valores se encuentran sobre una línea horizontal.

2.6 Caso tipificado

En este apartado se realiza una tipificación. Con esta transformación se intenta evitar problemas derivados de las diferentes escalas de los datos.

En el caso tipificado, se utiliza como matriz de covarianzas muestrales la **matriz de correlación muestral R** .

$$R = Z_i = (X_i - \bar{x}) / \sqrt{s_{ii}}, \quad i = 1, \dots, p$$

En este apartado se utiliza, alternativamente, la función “*eigen*” del paquete *basis*, con la que se obtiene directamente los **autovalores** y **autovectores** de cualquier matriz cuadrada,

```

eigen(R);

## $values
## [1] 1.7107028 0.2892972
##
## $vectors
##          [,1]      [,2]
## [1,] 0.7071068 -0.7071068
## [2,] 0.7071068  0.7071068

Se crea el vector de autovalores  $k$  y la matriz de autovectores  $H$  para la nueva matriz de equicorrelación,
K = with(eigen(R), values);
round(K,4);

## [1] 1.7107 0.2893

H = with(eigen(R),vectors);
round(H,4);

##          [,1]      [,2]
## [1,] 0.7071 -0.7071
## [2,] 0.7071  0.7071

se calculan las proporciones acumuladas,
round(cumsum(K/sum(K)),digits = 4);

## [1] 0.8554 1.0000

y se calculan las correlaciones de la primera componente principal con las variables ventas y beneficios,
sqrt(K[1]*H[,1]);

## [1] 1.099841 1.099841

```

2.7 Conclusiones

Como las proporciones son 0.855 y 1, las correlaciones entre la CP y las dos variables son iguales, ya que el autovector es constante.

Por otro lado, el valor de las componentes sería,

```

cp.1 = 0.7071 x ventas + 0.7071 x beneficio
cp.2 =-0.7071 x ventas + 0.7071 x beneficio

```

Estos coeficientes son los senos y los cosenos del ángulo de rotación entre los ejes de los CP y los ejes de las variables tipificadas.

En este caso lo correcto es utilizar la tipificación de los datos para evitar el problema de la escala. Se puede deducir comprobando que el valor de la columna *ventas* que es mayor que el de la columna *beneficios*.

3 Ejercicio [2]

En este ejercicio se realiza un análisis de componentes principales para el conjunto de datos *sckpr.txt*. Este dataset contiene los datos semanales de retornos sobre opciones de cinco compañías:

- AC: Allied Chemical.(Química)
- DP: Du Pont. (Química)
- UC: Union Carbide.(Química)
- EX: Exxon. (Petrolera)
- TX: Texaco. (Petrolera)

3.1 Carga y descripción de los datos

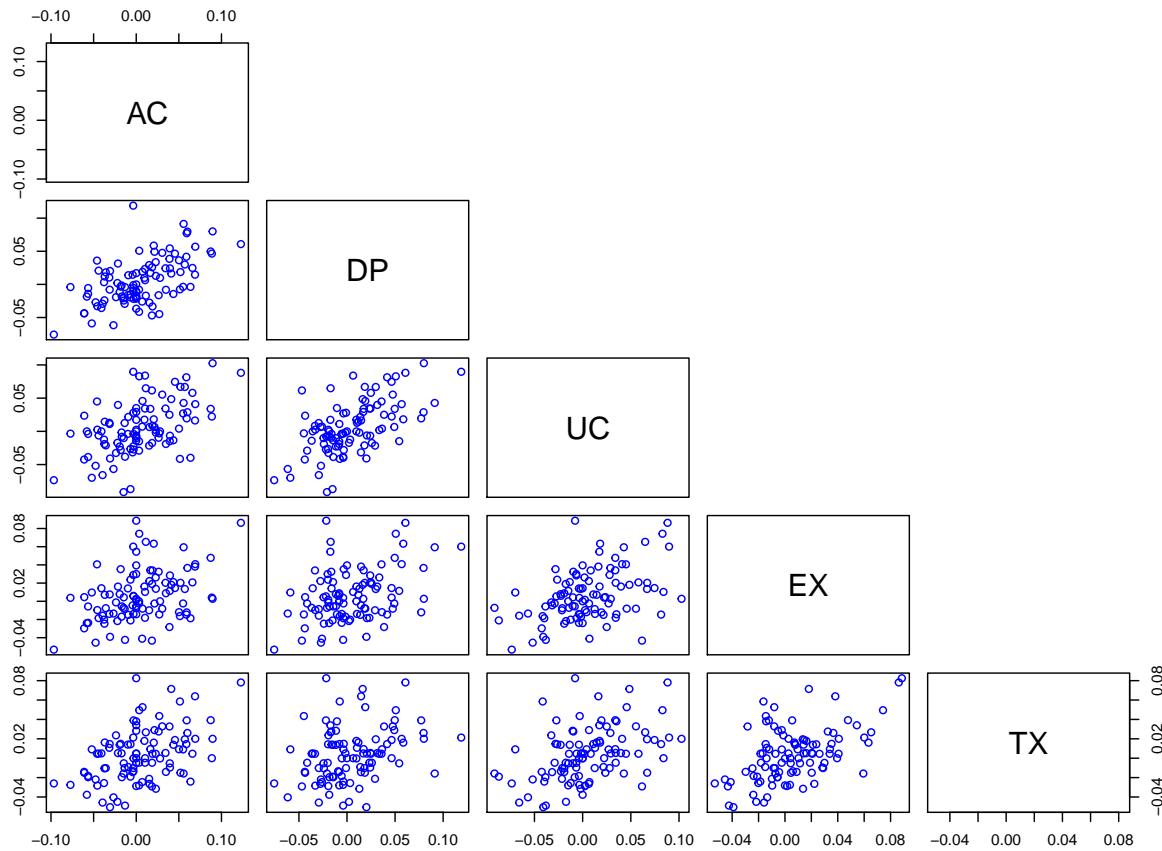
```
data.1 = read.table('stckpr.txt', header = TRUE);
head(data.1)

##          AC         DP         UC         EX         TX
## 1 0.000000 0.000000 0.000000 0.039473 0.000000
## 2 0.027027 -0.044855 -0.003030 -0.014466 0.043478
## 3 0.122807  0.060773  0.088146  0.086238 0.078124
## 4 0.057031  0.029948  0.066808  0.013513 0.019512
## 5 0.063670 -0.003793 -0.039788 -0.018644 -0.024154
## 6 0.003521  0.050761  0.082873  0.074265 0.049504
```

Diagrama de dispersión

```
plot(data.1, upper.panel = NULL,
      main= "Diagrama de dispersión",
      col = 4)
```

Diagrama de dispersión



Matriz de covarianza

```
S = cov(data.1);
round(S, digits = 4);

##      AC      DP      UC      EX      TX
## AC 0.0016 0.0008 0.0008 4e-04 5e-04
## DP 0.0008 0.0012 0.0008 4e-04 3e-04
## UC 0.0008 0.0008 0.0016 5e-04 5e-04
## EX 0.0004 0.0004 0.0005 8e-04 4e-04
## TX 0.0005 0.0003 0.0005 4e-04 8e-04
```

Autovalores y autovectores

```
eigen(S);

## $values
## [1] 0.0035953867 0.0007921798 0.0007364426 0.0005086686 0.0003437707
##
## $vectors
##      [,1]      [,2]      [,3]      [,4]      [,5]
## [1,] -0.5605914  0.73884565 -0.1260222  0.28373183 -0.20846832
## [2,] -0.4698673 -0.09286987 -0.4675066 -0.68793190  0.28069055
## [3,] -0.5473322 -0.65401929 -0.1140581  0.50045312 -0.09603973
## [4,] -0.2908932 -0.11267353  0.6099196 -0.43808002 -0.58203935
```

```

## [5,] -0.2842017 0.07103332 0.6168831 0.06227778 0.72784638

Variación Total

L = with(eigen(S), values);
L/sum(L);

## [1] 0.60159252 0.13255027 0.12322412 0.08511218 0.05752091

```

Ante estos resultados es destacable que los autovalores son bajos, debido a un problema de unidades de media. Es decir, no hay problema de varianzas dominantes, sino un problema de unidades, para solventarlo se multiplica por 100 para expresar los retornos semanales en forma de porcentaje.

3.2 Cambio de unidades

Se multiplican los datos por 100 y se recalcula la **matriz de covarianzas muestrales**,

```

S = cov(100*data.1);
round(S, digits = 4);

##          AC      DP      UC      EX      TX
## AC 16.2993 8.1667 8.1007 4.4224 5.1397
## DP  8.1667 12.2938 8.2763 3.8686 3.1094
## UC  8.1007 8.2763 15.5608 4.8728 4.6248
## EX  4.4224 3.8686 4.8728 8.0233 4.0847
## TX  5.1397 3.1094 4.6248 4.0847 7.5874

```

Calculo de las **desviaciones típicas** como la raíz de la diagonal principal de la matriz de covarianzas,

```

round(sqrt(diag(S)), digits = 4);

##          AC      DP      UC      EX      TX
## 4.0372 3.5062 3.9447 2.8325 2.7545

los autovalores y autovectores

eigen(S);

## $values
## [1] 35.953867 7.921798 7.364426 5.086686 3.437707
##
## $vectors
##           [,1]      [,2]      [,3]      [,4]      [,5]
## [1,] -0.5605914 0.73884565 -0.1260222 0.28373183 -0.20846832
## [2,] -0.4698673 -0.09286987 -0.4675066 -0.68793190  0.28069055
## [3,] -0.5473322 -0.65401929 -0.1140581  0.50045312 -0.09603973
## [4,] -0.2908932 -0.11267353  0.6099196 -0.43808002 -0.58203935
## [5,] -0.2842017  0.07103332  0.6168831  0.06227778  0.72784638

```

Los pesos de la primera componente principal son prácticamente iguales.

Variación Total

```

L = with(eigen(S), values);
L/sum(L);

## [1] 0.60159252 0.13255027 0.12322412 0.08511218 0.05752091

```

Matriz de correlaciones

```

R = cor(data.1);
round(R, digits = 4);

```

```

##      AC      DP      UC      EX      TX
## AC 1.0000 0.5769 0.5087 0.3867 0.4622
## DP 0.5769 1.0000 0.5984 0.3895 0.3220
## UC 0.5087 0.5984 1.0000 0.4361 0.4256
## EX 0.3867 0.3895 0.4361 1.0000 0.5235
## TX 0.4622 0.3220 0.4256 0.5235 1.0000

```

En este punto pueden calcularse las *correlaciones mínima y máxima*. Para la máxima, previamente se resta una matriz identidad de tamaño n, para que no se utilicen los valores iguales a uno,

```

round(c(min(R),max(R - diag(ncol(data.1)))), digits = 4);
## [1] 0.3220 0.5984

```

La correlación entre todas las variables es positiva y se encuentra entre los valores 0.3 y 0.6.

3.3 Caso tipificado

Para calcular el caso tipificado, se calculan los **autovalores** y **autovectores** sobre la matriz de correlaciones muestrales,

```

eigen(R);
## $values
## [1] 2.8564869 0.8091185 0.5400440 0.4513468 0.3430038
##
## $vectors
##      [,1]      [,2]      [,3]      [,4]      [,5]
## [1,] -0.4635405  0.2408499  0.6133570 -0.3813727 -0.4532876
## [2,] -0.4570764  0.5090997 -0.1778996 -0.2113068  0.6749814
## [3,] -0.4699804  0.2605774 -0.3370355  0.6640985 -0.3957247
## [4,] -0.4216770 -0.5252647 -0.5390181 -0.4728036 -0.1794482
## [5,] -0.4213291 -0.5822416  0.4336029  0.3812273  0.3874672

```

Y se obtienen las **proporciones de variación explicada** de la matriz tipificada y comparada con las proporciones sin tipificar,

```

K = with(eigen(R), values);
K/sum(K);

## [1] 0.57129738 0.16182370 0.10800880 0.09026936 0.06860076
L/sum(L);

## [1] 0.60159252 0.13255027 0.12322412 0.08511218 0.05752091

```

Prácticamente son iguales, ya que los valores de retornos (logaritmos de las diferencias) se encuentran en la misma escala.

3.4 Reducción de la dimensión

En este apartado se trabaja con el objetivo del ACP, la reducción de la dimensión para realizar un resumen de los datos. En primer lugar, se calcula la proporción de variación y proporción acumulada, mostrándose luego en una única matriz de variación explicada.

```

prop = K/sum(K);
acum = cumsum(prop);
table = round(data.frame(S,prop,acum, row.names = NULL ),digits = 4);

```

Proporciones

```
prop;  
## [1] 0.57129738 0.16182370 0.10800880 0.09026936 0.06860076
```

Proporciones acumuladas

```
acum;  
## [1] 0.5712974 0.7331211 0.8411299 0.9313992 1.0000000
```

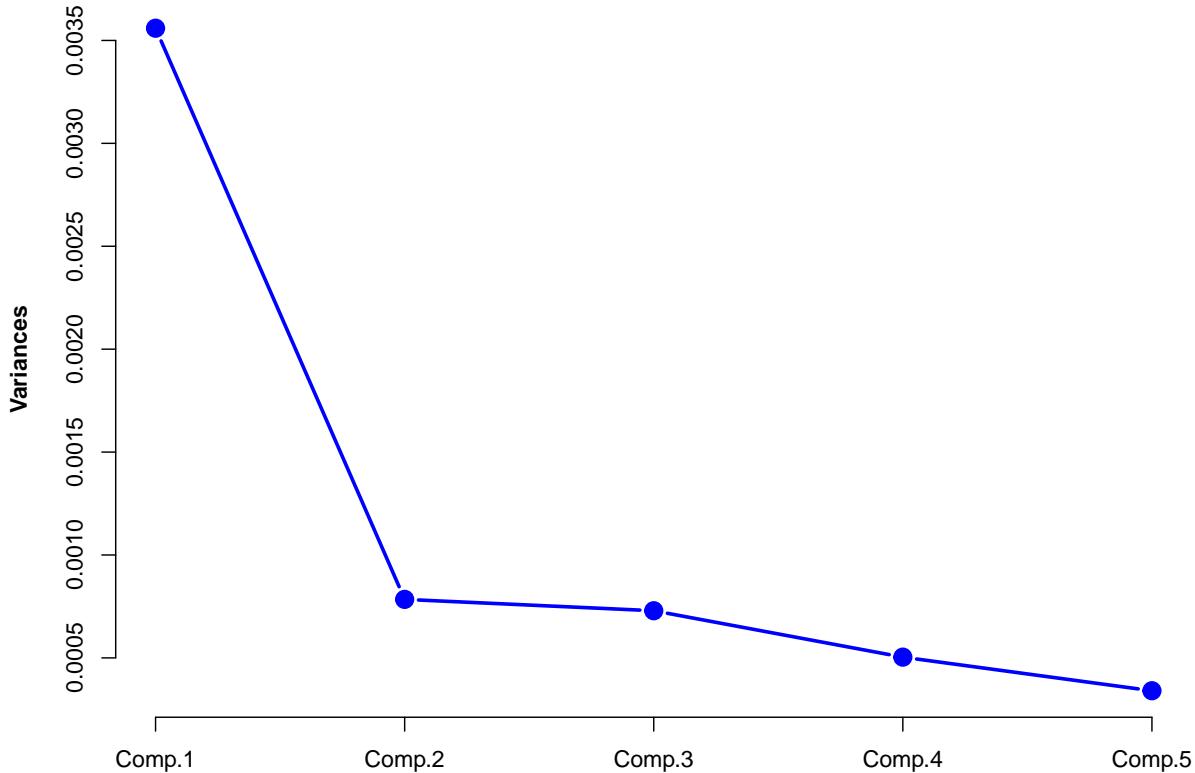
Matriz de proporción explicada

```
table;  
  
##          AC      DP      UC      EX      TX    prop    acum  
## 1 16.2993 8.1667 8.1007 4.4224 5.1397 0.5713 0.5713  
## 2 8.1667 12.2938 8.2763 3.8686 3.1094 0.1618 0.7331  
## 3 8.1007 8.2763 15.5608 4.8728 4.6248 0.1080 0.8411  
## 4 4.4224 3.8686 4.8728 8.0233 4.0847 0.0903 0.9314  
## 5 5.1397 3.1094 4.6248 4.0847 7.5874 0.0686 1.0000
```

Gráfico de sedimentación

```
pc = princomp(data.1);  
plot(pc, type = 'l', col = 4,  
pch = 19, lwd = 2.5, cex = 1.5,  
main = 'Gráfico de pendiente de autovalores',  
font.lab = 2,  
cex.main = 1.3);
```

Gráfico de pendiente de autovalores



A partir de la variación acumulada y el gráfico de sedimentación, podríamos concluir que la **reducción de la dimensión** sería de $q = 2$, ya que Y_2 explica el 85,74% de la variación.

La determinación del número de componentes a retener es, en parte, arbitraria y queda a juicio del investigador. Un criterio, por ejemplo, sería retener los factores con autovalor superior a 1. Con ese criterio la segunda componente no sería elegida.

Otra operación en el ACP, es reducir la dimensionalidad de las variables, desechando las variables (compañías) menos informativas entre las originales de nuestro estudio. Joliffe (1972, 1973) ha propuesto un método para ello que consiste en reducir directamente las variables originales al número de ellas que cumplan con los siguientes requisitos:

- Elegir los componentes principales cuyos autovalores sean mayores que 0.7. En este caso se elegirían las 3 primeras componentes
- De las componentes principales escogidas, seleccionar la variable con mayor valor absoluto (que no haya sido seleccionada previamente).

Según lo cual, se trabajaría exclusivamente con las compañías UC, TX y Ac.

En el siguiente apartado se exploran otro tipo de gráficos para estudiar la carga de las componentes y la proximidad entre observaciones y variables, de forma de averiguar si arrojan más información sobre la reducción de la dimensión.

3.5 Otros gráficos

En esta sección, y a modo ilustrativo, se utiliza la función “*prcomp*” para el ACP, también incluida en el paquete *stats*, creando el objeto *acp* que dota con la información a los gráficos.

Se obtienen las desviaciones standard de las componentes principales y el elemento *Rotation* que son los **autovectores**, el equivalente a *loadings* de la función “*princomp*”.

```
(acp = prcomp(data.1));  
  
## Standard deviations:  
## [1] 0.05996154 0.02814569 0.02713748 0.02255368 0.01854106  
##  
## Rotation:  
##          PC1         PC2         PC3         PC4         PC5  
## AC 0.5605914 0.73884565 -0.1260222 0.28373183 -0.20846832  
## DP 0.4698673 -0.09286987 -0.4675066 -0.68793190 0.28069055  
## UC 0.5473322 -0.65401929 -0.1140581 0.50045312 -0.09603973  
## EX 0.2908932 -0.11267353 0.6099196 -0.43808002 -0.58203935  
## TX 0.2842017 0.07103332 0.6168831 0.06227778 0.72784638
```

Se puede trabajar usando datos reescalados, indicando a la función el valor lógico *scale*, para tener una varianza unitaria antes de que se realice el análisis.

```
(acp.r = prcomp(data.1, scale = TRUE));  
  
## Standard deviations:  
## [1] 1.6901145 0.8995101 0.7348768 0.6718235 0.5856653  
##  
## Rotation:  
##          PC1         PC2         PC3         PC4         PC5  
## AC 0.4635405 -0.2408499 0.6133570 -0.3813727 0.4532876  
## DP 0.4570764 -0.5090997 -0.1778996 -0.2113068 -0.6749814  
## UC 0.4699804 -0.2605774 -0.3370355 0.6640985 0.3957247  
## EX 0.4216770 0.5252647 -0.5390181 -0.4728036 0.1794482  
## TX 0.4213291 0.5822416 0.4336029 0.3812273 -0.3874672
```

Los **autovalores** se obtienen como:

```
acp$sdev^2;  
  
## [1] 0.0035953867 0.0007921798 0.0007364426 0.0005086686 0.0003437707
```

Mediante un resumen del objeto *acp* es sencillo obtener la **proporción de la varianza explicada** y su **acumulación** como en los apartados anteriores,

```
summary(acp);  
  
## Importance of components:  
##          PC1         PC2         PC3         PC4         PC5  
## Standard deviation 0.05996 0.02815 0.02714 0.02255 0.01854  
## Proportion of Variance 0.60159 0.13255 0.12322 0.08511 0.05752  
## Cumulative Proportion 0.60159 0.73414 0.85737 0.94248 1.00000
```

Es interesante en el análisis gráfico, calcular las **correlaciones** entre las variables y las componentes,

```
(corvar = acp.r$rotation %*% diag(acp.r$sdev));  
  
##          [,1]      [,2]      [,3]      [,4]      [,5]  
## AC 0.7834366 -0.2166469 0.4507418 -0.2562151 0.2654748  
## DP 0.7725114 -0.4579403 -0.1307343 -0.1419609 -0.3953132
```

```

## UC 0.7943207 -0.2343920 -0.2476796  0.4461570  0.2317622
## EX 0.7126824  0.4724809 -0.3961119 -0.3176406  0.1050966
## TX 0.7120945  0.5237322  0.3186447  0.2561175 -0.2269261

```

La función *prcomp* devuelve, además, el valor de los datos rotados x , es decir, los datos centrados multiplicados por la matriz de rotación,

```

head(acp.r$x);

##          PC1        PC2        PC3        PC4        PC5
## [1,] 0.2445649 0.6767803 -0.6995557 -0.6199562 0.2375746
## [2,] -0.2038989 1.1056309  1.6753642  0.8461270 0.4208211
## [3,]  5.3881808 0.9980347  0.4446003 -0.3616913 0.5280689
## [4,]  1.9977323 -0.6085663  0.2452845  0.4889131 0.5326489
## [5,] -0.7825238 -0.9734296  1.3526574 -1.2326002 0.5978952
## [6,]  3.2092324  1.0628780 -1.4944836  0.5404445 -0.3446589

```

Gráfico de la proporción de la varianza

```
barplot(summary(acp)$importance[2, ], main = "Proporción de varianza");
```

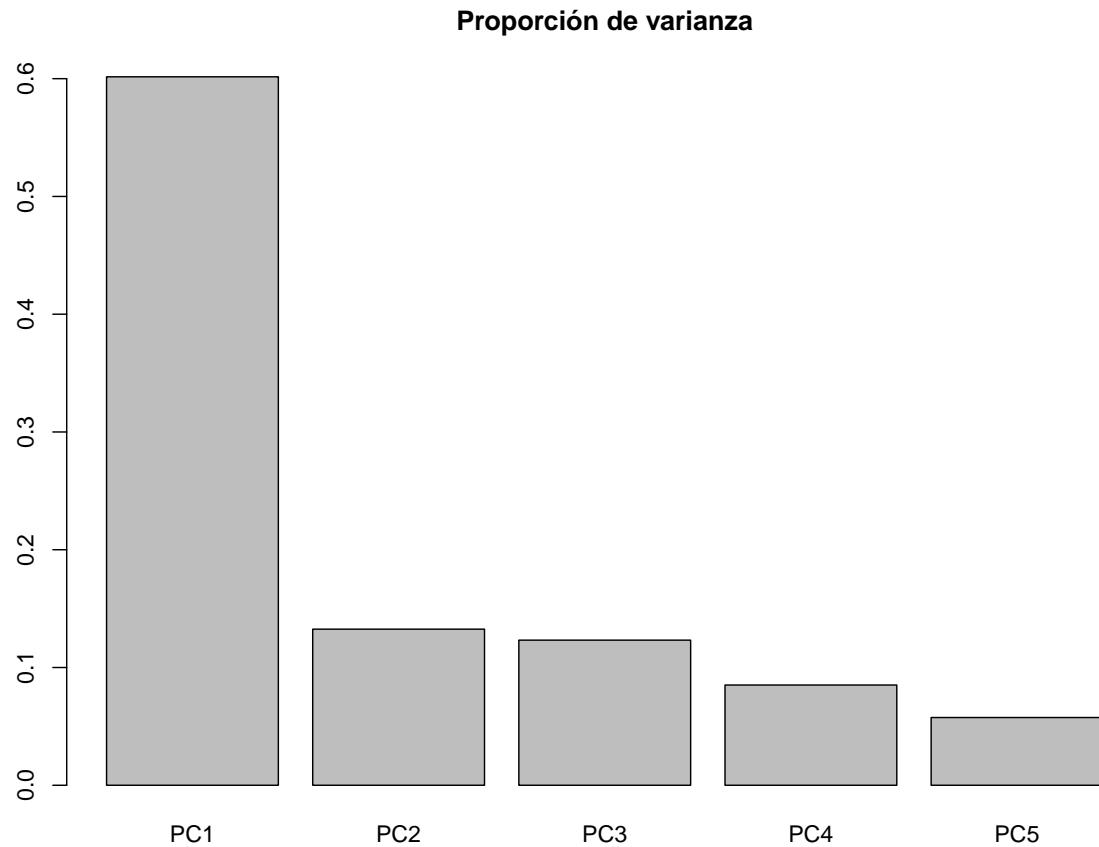


Gráfico de correlación entre las variables y las CP1 y CP2

```

plot(-1:1, -1:1, main = "Correlación CP1-CP2",
     type ='n',
     asp = 1,

```

```

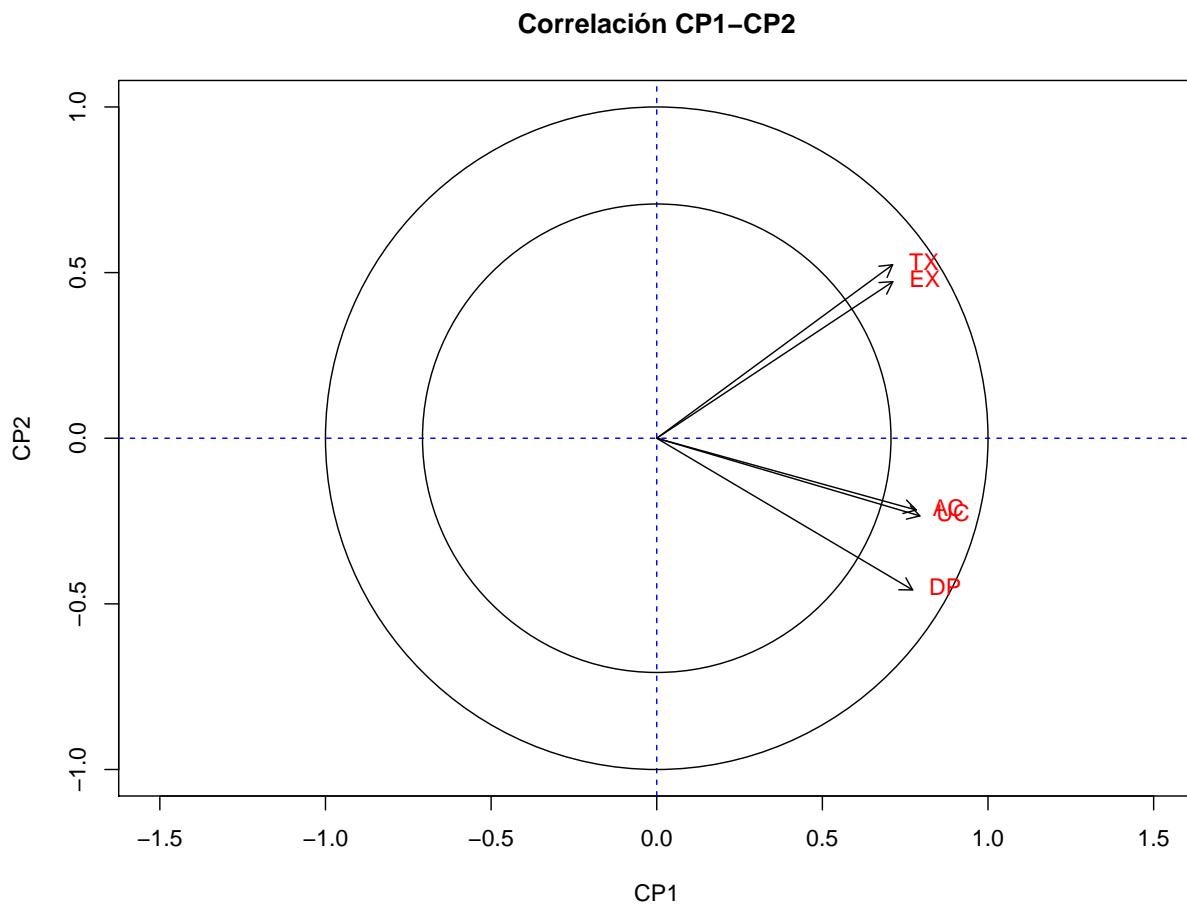
xlab = 'CP1',
ylab = 'CP2');

abline(h = 0, v = 0, lty = 2, col = 4);

## Dibuja un círculo de centro (0,0) y radio 1
symbols(0, 0, 1, inches = F, add = T);
symbols(0, 0, sqrt(.5), inches = F, add = T);

## Dibuja los vectores y coloca los nombres
arrows(0, 0, corvar[,1], corvar[,2], length = .1);
text(corvar[,1], corvar[,2], colnames(data.1), pos=4, offset = .6, col = 2, font = 0.3);

```



En el primer cuadrante se observa que TX y EX (petroleras) son mayores que cero, mientras que AC, DP y UC (químicas) son negativas para la CP2. Todas las flechas tienen el mismo tamaño, por tanto los coeficientes son similares.

Gráfico de los individuos

```

plot(acp$x[, 1:2],
      main = "Individuos",
      pch = 19,
      col= 4);
abline(h = 0, v = 0, lty = 2, col = 4);

```

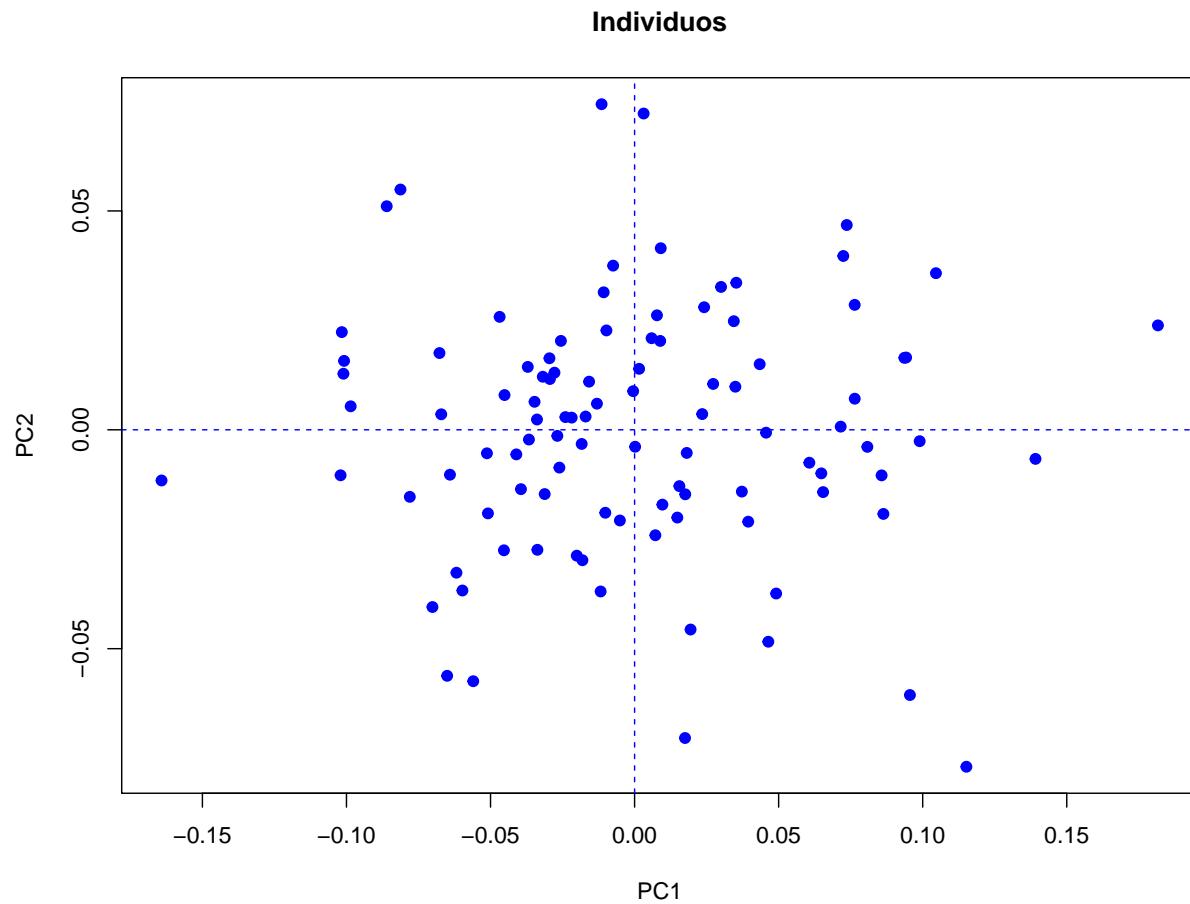


Gráfico Bliplot

Los **Biplots** son un tipo de gráfico exploratorio usado en Estadística. Se trata de una generalización multivariante de un diagrama de dispersión de dos variables. De la misma manera que un diagrama de dispersión muestra la distribución conjunta de dos variables, un Biplot representa tres o más variables. El biplot aproxima la distribución de una muestra multivariante en un espacio de dimensión reducida, normalmente de dimensión dos, y superpone sobre la misma representaciones de las variables sobre las que se mide la muestra.

```
biplot(acp);
abline(h = 0, v = 0, lty = 2, col = 4);
```

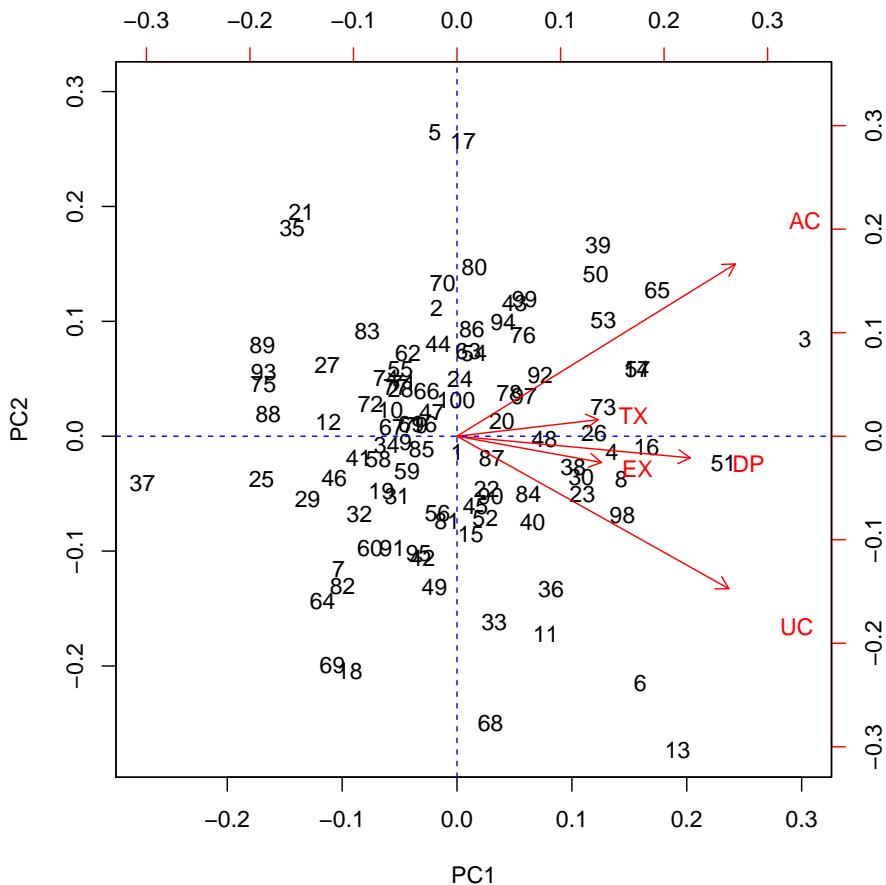
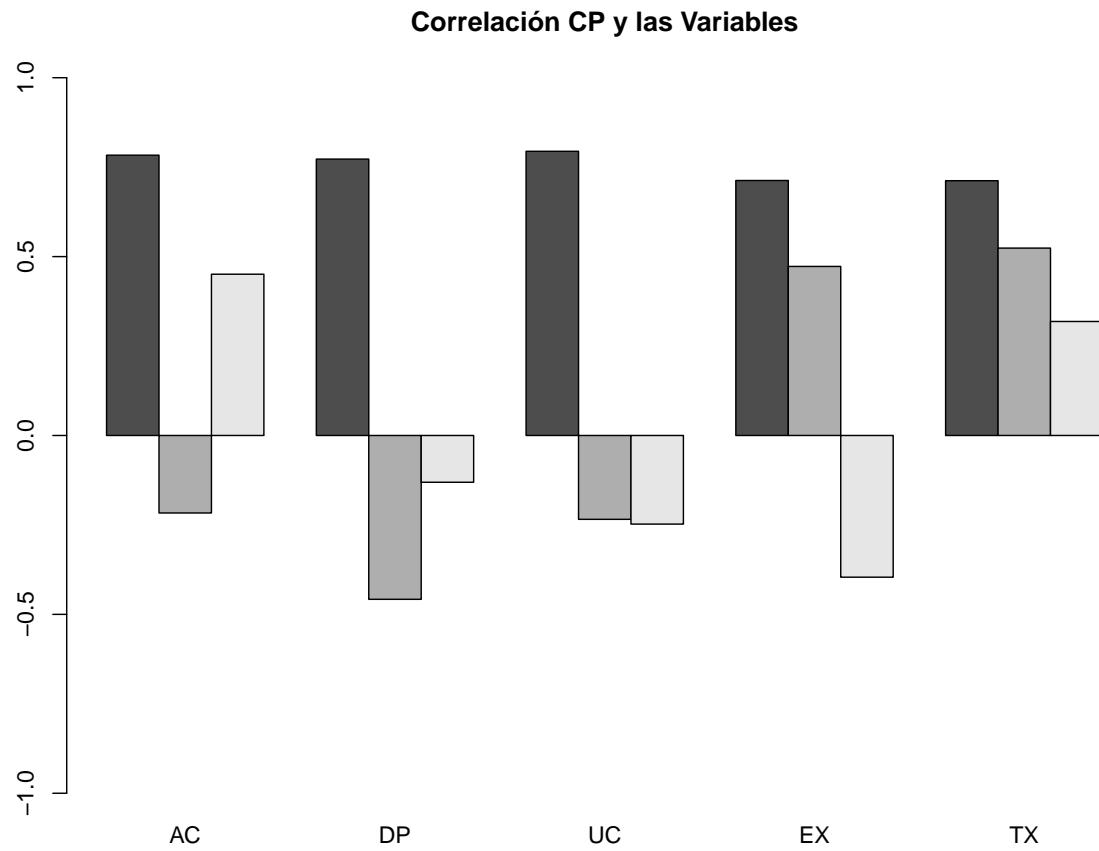


Gráfico de barras para la Correlación CP y las Variables

```
data.cor = acp.r$rotation %*% diag(acp.r$sdev);
barplot(t(data.cor[,1:3]),beside = TRUE, main= "Correlación CP y las Variables",ylim = c(-1,1));
```



3.6 Conclusiones

El objeto de este análisis es averiguar si existe algún tipo de relación entre los valores de los retornos de las opciones de las cinco compañías, tres dedicadas a la industria química y dos petroleras.

- Las componentes principales acumulan el 84,11% de la variación.
- La CP1 explica un 57% de la varación y sus coeficientes son prácticamente iguales, entorno a 0.44. Además, tiene todas sus coordenadas del mismo signo, hecho que indica la existencia de una alta correlación positiva entre todas las variables. La CP1 puede interpretarse como un promedio ponderado de todas las variables, o un factor global de “tamaño”, aunque no lo sea *exactamente* para el caso de las opciones sobre retornos.
- La CP2 explica un 16% de la varación y tiene coordenadas positivas y negativas, lo que incida que existe una contraposición entre el grupo de compañías químicas y el grupo de petroleras.
- La CP3 explica un 16% de la varación, depende de las dos primeras componentes, y se observa que las compañías petroleras están en negativo y las químicas en positivo, por lo tanto compara su comportamiento y es una componente importante.
- En la reducción de la dimensión deben mantenerse las 3 primeras componentes principales CP1, CP2 y CP3.

Estos resultados coinciden con Benjamin F.King, 1966 donde, mediante diversos métodos de análisis factorial

aplicados a los datos dan un apoyo notable a la hipótesis de que el movimiento de un grupo de cambios de precios de materias primas puede desglosarse en componentes del mercado y de la industria. Intenta identificar el agrupamiento de la industria mediante un control de los signos de carga de los componentes principales.

4 Ejercicio [3]

En este ejercicio se realiza un análisis de componentes principales para el conjunto de datos *DJ*. Del cual queremos considerar los retornos asociados a los valores diarios de cierre del índice bursátil Dow Jones durante los años 1992-1998 para las empresas:

- MO: Philip Morris
- KO: Coca Cola
- EK: Kodak
- HWP: Hewlett Packard
- INTC: Intel
- MSFT: Microsoft
- MCD: McDonald's
- WMT: Walmart
- DIS: Disney

4.1 Carga y descripción de los datos

En primer lugar, es necesario instalar los paquetes y las dependencias para trabajar con el conjunto de datos **DJ**,

```
install.packages("QRM");
library(QRM);

#paquete para graficar:
library (zoo)
#para comprobar los paquetes instalados:
search();
```

y en segundo lugar, cargar el conjunto de datos,

```
library(QRM);
library (zoo);
data(DJ);
```

se crea una variable vacía con los datos del índice dowjones. Para obtener los datos se deben llamar desde la memoria de ejecución. Por ejemplo, se crea una variable **data.ini** con todos los datos de **DJ** y se comprueba su tamaño (*filas x columnas*):

```
dim(data.ini);
```

Se genera un vector con las cabeceras o nombres de las variables de interés.

```
I = c('MO', 'KO', 'EK', 'HWP', 'INTC', 'MSFT', 'IBM', 'MCD', 'WMT', 'DIS');
```

Se utiliza la función “*window*” del paquete *timeseries* que permite extraer los **valores** para unas fechas dadas por el ejercicio.

```
data.S = window(DJ[,I], '1992-01-01', '1998-12-31');
```

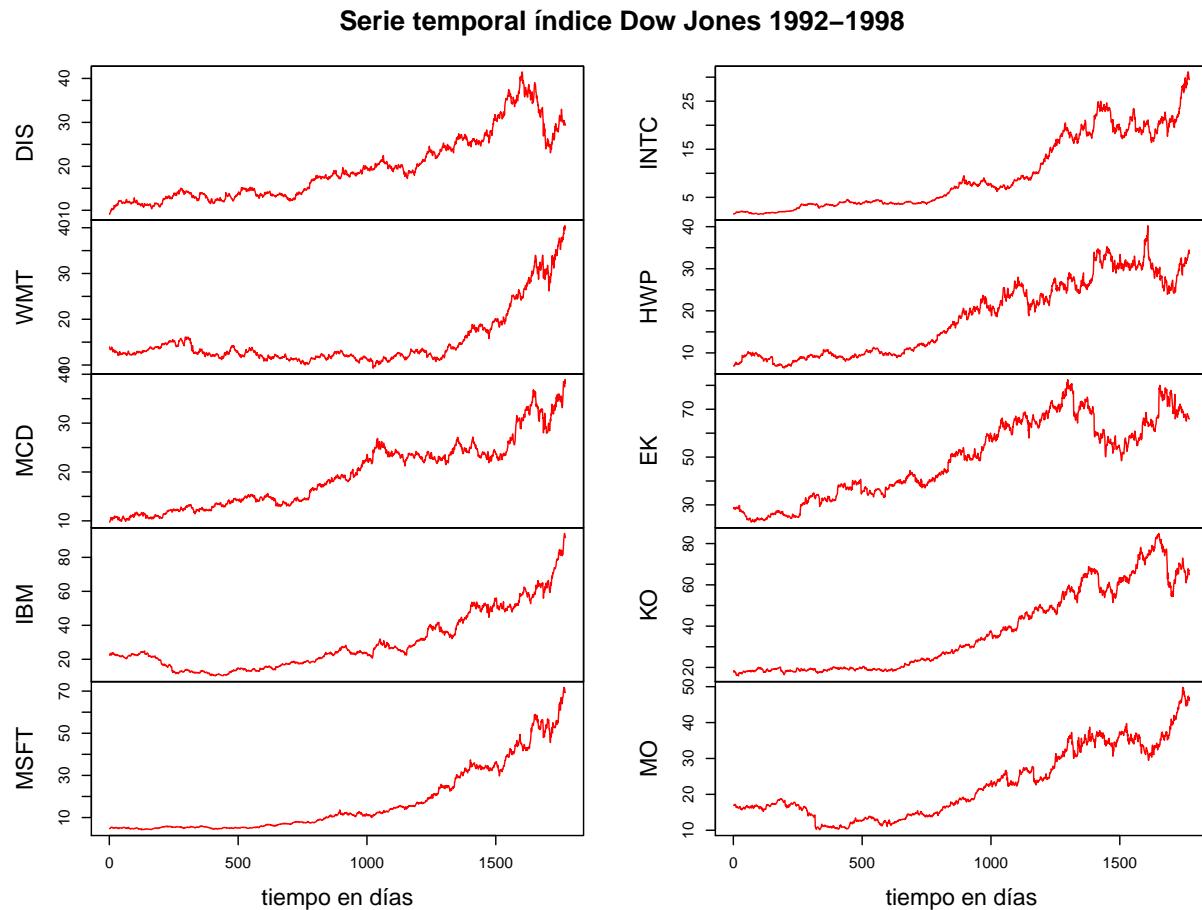
Mediante la función “*returs*” se calculan los **retornos**, que pueden ser continuos o discretos. En este caso se trabaja sobre datos discretos.

```
data.0 = returns(data.S, method = 'discrete');
```

Gráficos

Se crea una matriz de gráficos para mostrar la serie temporal de cada compañía.

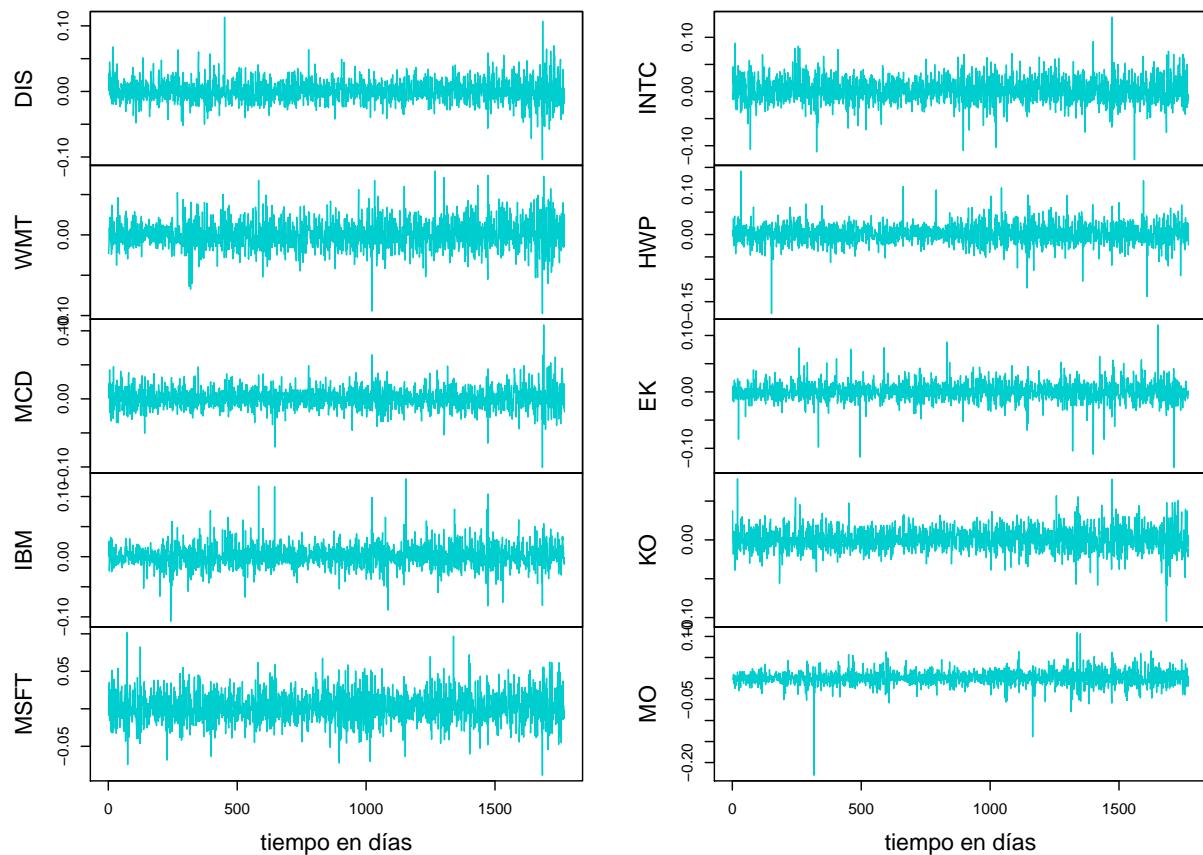
```
plot.zoo(data.S[,10:1],  
         main = 'Serie temporal índice Dow Jones 1992-1998',  
         xlab = 'tiempo en días',  
         col = 'red',  
         cex.main = 1.75);
```



Y para mostrar la matriz de gráficos de los retornos.

```
plot.zoo(data.0[,10:1],  
         main = 'Retornos diarios índice Dow Jones 1992-1998',  
         xlab = 'tiempo en días',  
         col = 'cyan3',  
         cex.main = 1.75);
```

Retornos diarios índice Dow Jones 1992–1998



Se realiza una transformación de los datos, multiplicándolos por 100 para pasarlo a %.

```
data.1 = 100*data.frame(data.0);
str(data.1);

## 'data.frame':    1769 obs. of  10 variables:
## $ MO : num -0.157 -0.157 0 0.945 0.936 ...
## $ KO : num 3.739 -0.301 -0.903 -1.824 -0.928 ...
## $ EK : num -1.523 0.773 1.023 -0.506 -1.781 ...
## $ HWP : num 0.895 -0.222 -0.445 3.572 3.448 ...
## $ INTC: num -0.493 -1.499 4.566 1.455 2.395 ...
## $ MSFT: num -0.878 3.099 3.004 3.542 2.414 ...
## $ IBM : num 0.138 2.075 2.574 -2.378 -1.218 ...
## $ MCD : num 0.318 2.222 2.174 0 4.256 ...
## $ WMT : num -2.326 -1.732 -0.44 0.885 0 ...
## $ DIS : num 2.697 0.884 0.104 2.083 4.489 ...
```

Se realiza un diagrama de nube de puntos. Una característica curiosa de este tipo de datos, retornos diarios, es la “asimetría de las colas” que es observable gráficamente

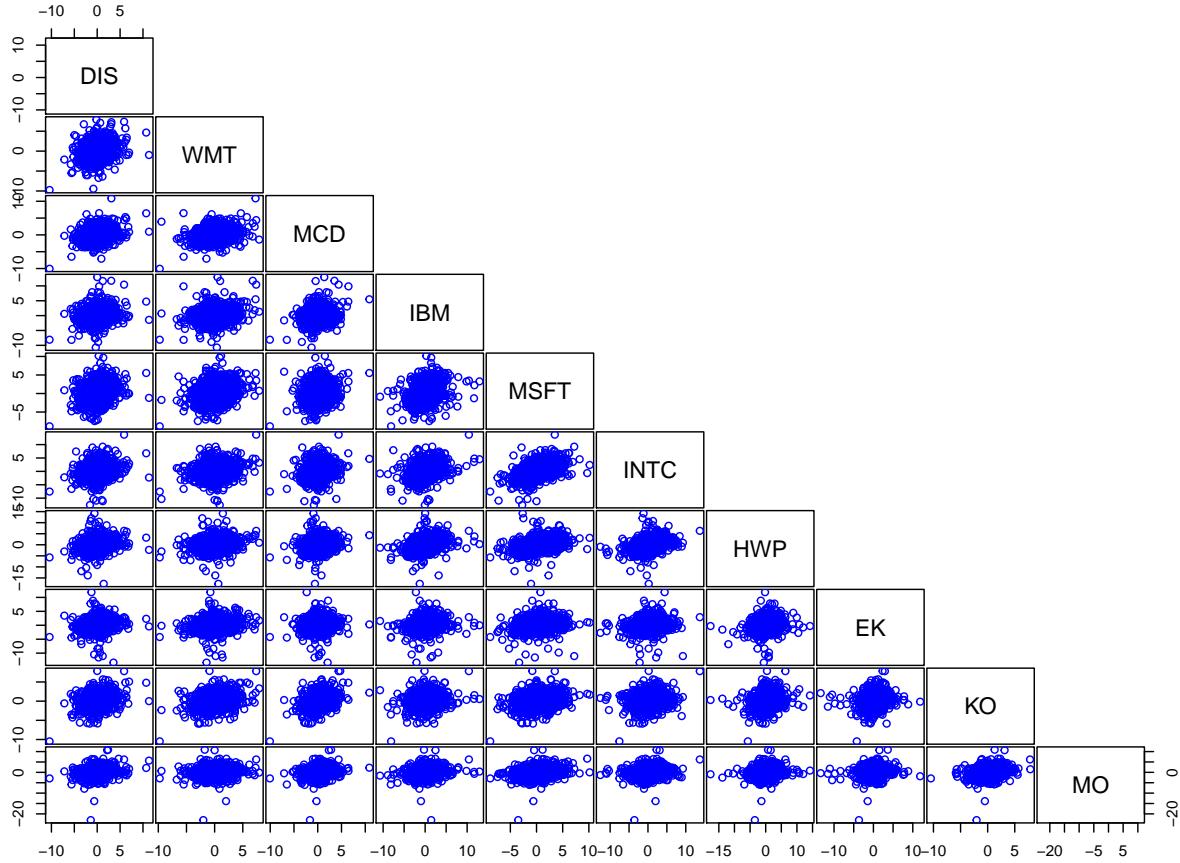
```
plot(data.1[10:1], upper.panel = NULL,
  main = 'Retornos diarios índice Dow Jones 1992–1998',
  col = 4,
```

```

labels = colnames(data.1[10:1]),
cex.labels = 1.5,
gap = .15);

```

Retornos diarios índice Dow Jones 1992–1998



Con la raíz de la diagonal principal de matriz de covarianzas, se obtienen las desviaciones típicas de los **retornos**,

```

round(sqrt(diag(cov(data.1))), digits = 4);

##      MO      KO      EK      HWP     INTC     MSFT      IBM      MCD      WMT      DIS
## 1.7738 1.4801 1.7146 2.2456 2.3315 2.0620 1.9161 1.5213 1.7756 1.7033

```

4.2 Componentes principales

Se crea el objeto *pc* para hacer la matriz de componentes principales sobre la matriz *data.1* y se comprueba su estructura y elementos,

```

pc = princomp(data.1);
#Estructura: str(pc);
#Elementos:
ls(pc);

## [1] "call"      "center"    "loadings"   "n.obs"      "scale"      "scores"
## [7] "sdev"

```

Las descripciones de la función “*princomp*” se desarrolla en el apartado 2.4.

Por ejemplo, para comprobar las coordenadas de los autovectores:pc\$loadings, se utiliza el operador de extracción [:]:

```
round(pc$loadings[,1:6],digits = 4);  
##          Comp.1   Comp.2   Comp.3   Comp.4   Comp.5   Comp.6  
## MO    -0.1975 -0.3949  0.0885  0.6219  0.4807 -0.1005  
## KO    -0.1913 -0.3429 -0.0004 -0.0813  0.0539  0.0894  
## EK    -0.1639 -0.2260 -0.0313 -0.0621 -0.3389 -0.8949  
## HWP   -0.4528  0.2542 -0.7607 -0.1753  0.3302 -0.0683  
## INTC  -0.5115  0.4547  0.4582 -0.0944 -0.0125 -0.0278  
## MSFT  -0.4356  0.1014  0.3808  0.0130  0.1630 -0.0170  
## IBM   -0.3203  0.0656 -0.2328  0.5561 -0.6703  0.2633  
## MCD   -0.1782 -0.3104 -0.0151 -0.0140  0.0008  0.1477  
## WMT   -0.2381 -0.3927  0.0537 -0.4746 -0.2525  0.2571  
## DIS   -0.2239 -0.3727 -0.0158 -0.1681  0.0483  0.1342
```

Y para comprobar la matriz de componentes principales, se extraen las primeras filas y columnas evitando mostrar todos los datos,

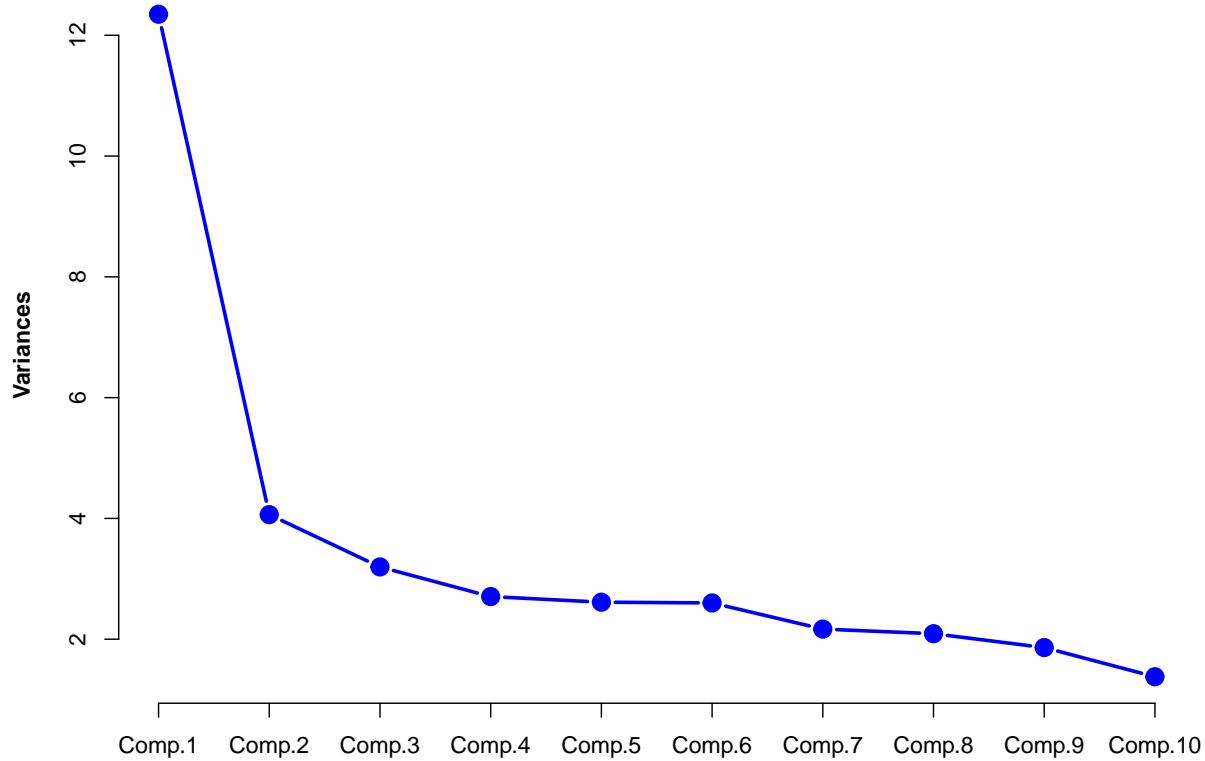
```
round(pc$scores[1:10,1:6],digits = 3)  
##          Comp.1   Comp.2   Comp.3   Comp.4   Comp.5   Comp.6  
## 1992-01-03 -0.006 -1.125 -1.463  0.292  1.424  1.530  
## 1992-01-06 -1.017 -0.619 -0.050  1.893 -0.814 -0.149  
## 1992-01-07 -4.217  1.992  2.835  1.286 -1.719 -0.242  
## 1992-01-08 -3.225  1.034 -0.084 -2.044  3.736 -0.267  
## 1992-01-09 -4.577 -0.495 -0.369 -1.524  3.541  1.986  
## 1992-01-10  4.089  0.465  0.608  0.050 -1.202 -0.121  
## 1992-01-13 -1.205  0.816  0.910 -1.992 -0.002  0.470  
## 1992-01-14 -5.492 -1.920 -1.805 -0.632 -1.010 -0.232  
## 1992-01-15 -4.113  4.610 -1.916  0.039 -1.376 -0.300  
## 1992-01-16 -0.209  7.917  3.355 -1.051 -1.724 -0.203
```

4.3 Gráfico de pendiente

Se genera un gráfico de pendiente o de *sedimentación*

```
plot(pc, type = 'l', col = 4,  
     pch = 19, lwd = 2.5, cex = 1.5,  
     main = 'Gráfico de pendiente de autovalores',  
     font.lab = 2,  
     cex.main = 1.3);
```

Gráfico de pendiente de autovalores



En el gráfico de pendiente de autovalores, se observa un “*codo*” en la componente 3, a partir de la cual se estabiliza.

4.4 Autovectores

Se crea el objeto G que contiene a la matriz de autovectores,

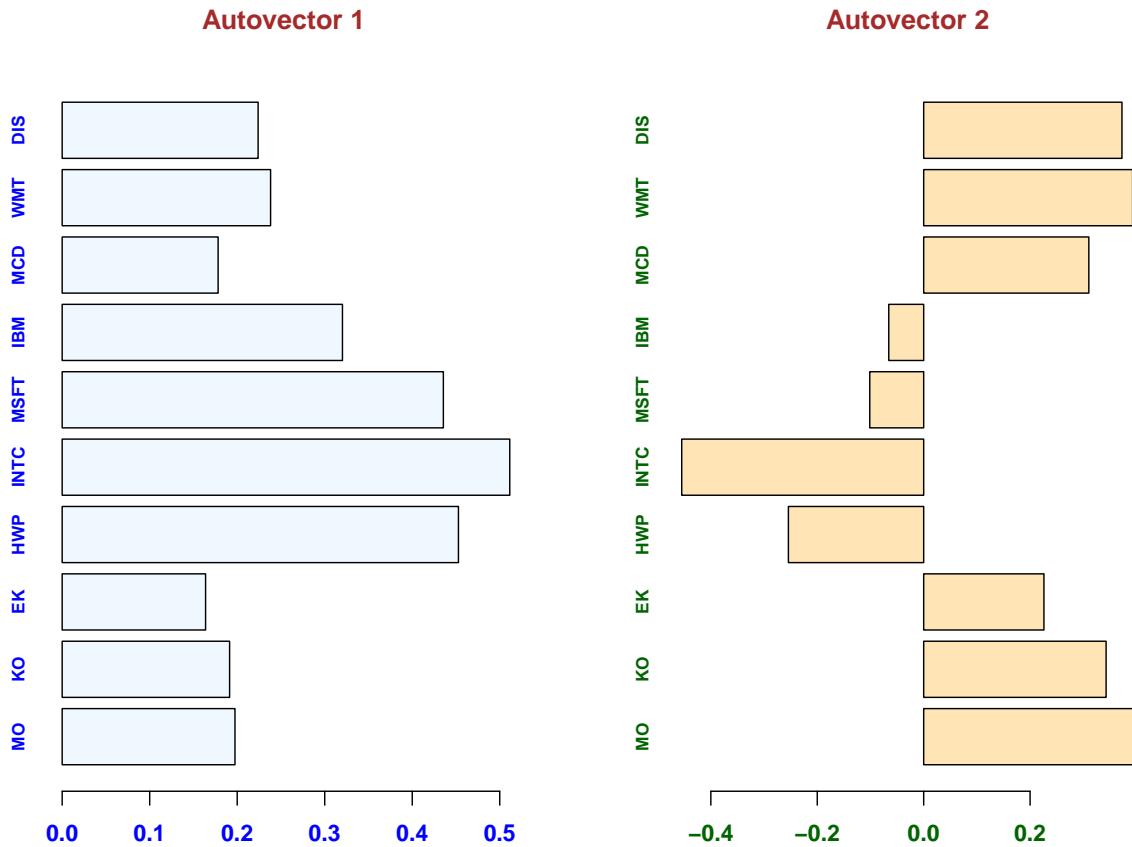
```
G = loadings(pc) [,];
```

Se crea un gráfico de barras resumiendo los valores de los autovectores que definen las dos primeras componentes principales

```
par(mfrow = c(1,2));
```

```
barplot(-G[,1], horiz = TRUE, cex.names = 0.75, font = 2, col.axis = 'blue',
main = 'Autovector 1', col.main = 'brown', col = 'aliceblue');
```

```
barplot(-G[,2], horiz = TRUE, cex.names = 0.75, font = 2, col.axis = 'darkgreen',
main = 'Autovector 2', col.main = 'brown', col = 'moccasin');
```



A continuación se realiza el mismo gráfico, pero dividiendo el primer autovector por la suma de coordenadas.

```

round(-G[,1], digits = 4);
##      MO      KO      EK      HWP      INTC      MSFT      IBM      MCD      WMT      DIS
## 0.1975 0.1913 0.1639 0.4528 0.5115 0.4356 0.3203 0.1782 0.2381 0.2239

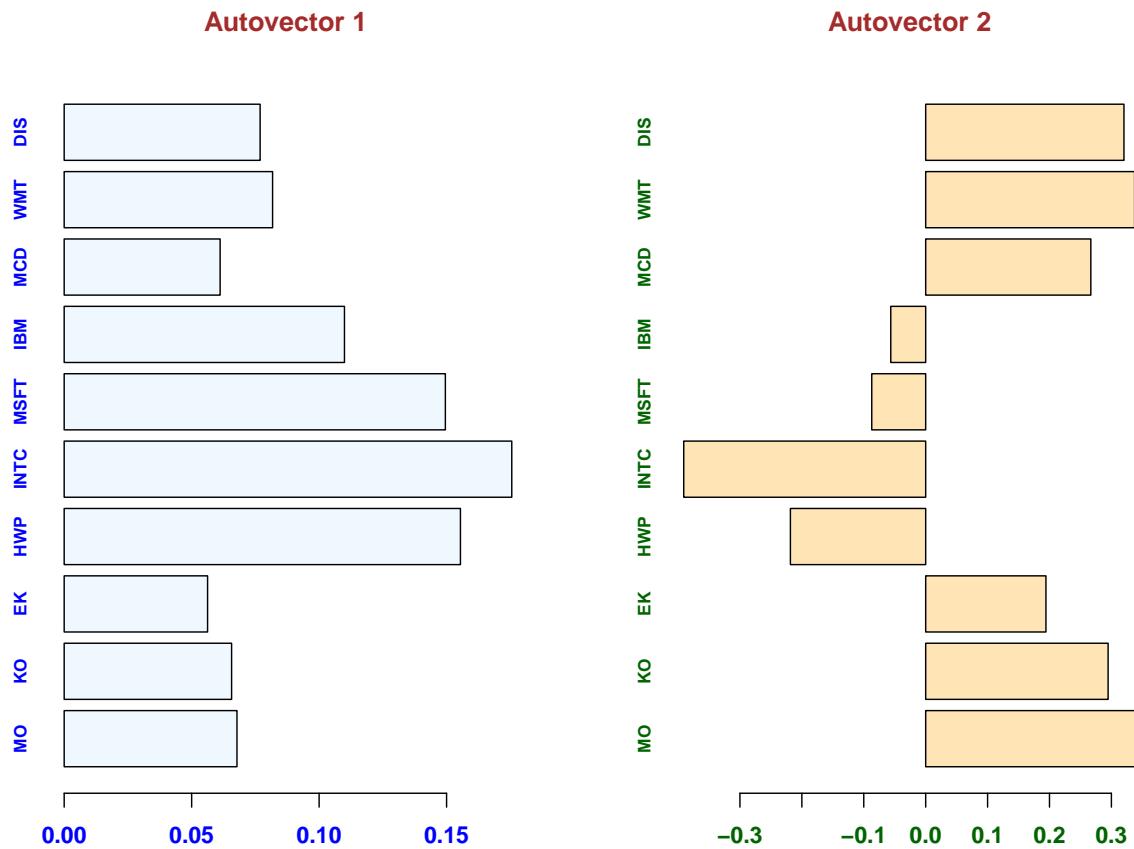
round(-G[,1]/sum(-G[,1]), digits = 4);
##      MO      KO      EK      HWP      INTC      MSFT      IBM      MCD      WMT      DIS
## 0.0678 0.0657 0.0563 0.1554 0.1756 0.1495 0.1099 0.0612 0.0817 0.0769

par(mfrow = c(1,2));

barplot((-G[,1]/sum(-G[,1])), horiz = TRUE, cex.names = 0.75, font = 2, col.axis = 'blue',
main = 'Autovector 1', col.main = 'brown', col = 'aliceblue');

barplot(-G[,2]/sum(-G[,2]), horiz = TRUE, cex.names = 0.75, font = 2, col.axis = 'darkgreen',
main = 'Autovector 2', col.main = 'brown', col = 'moccasin');

```

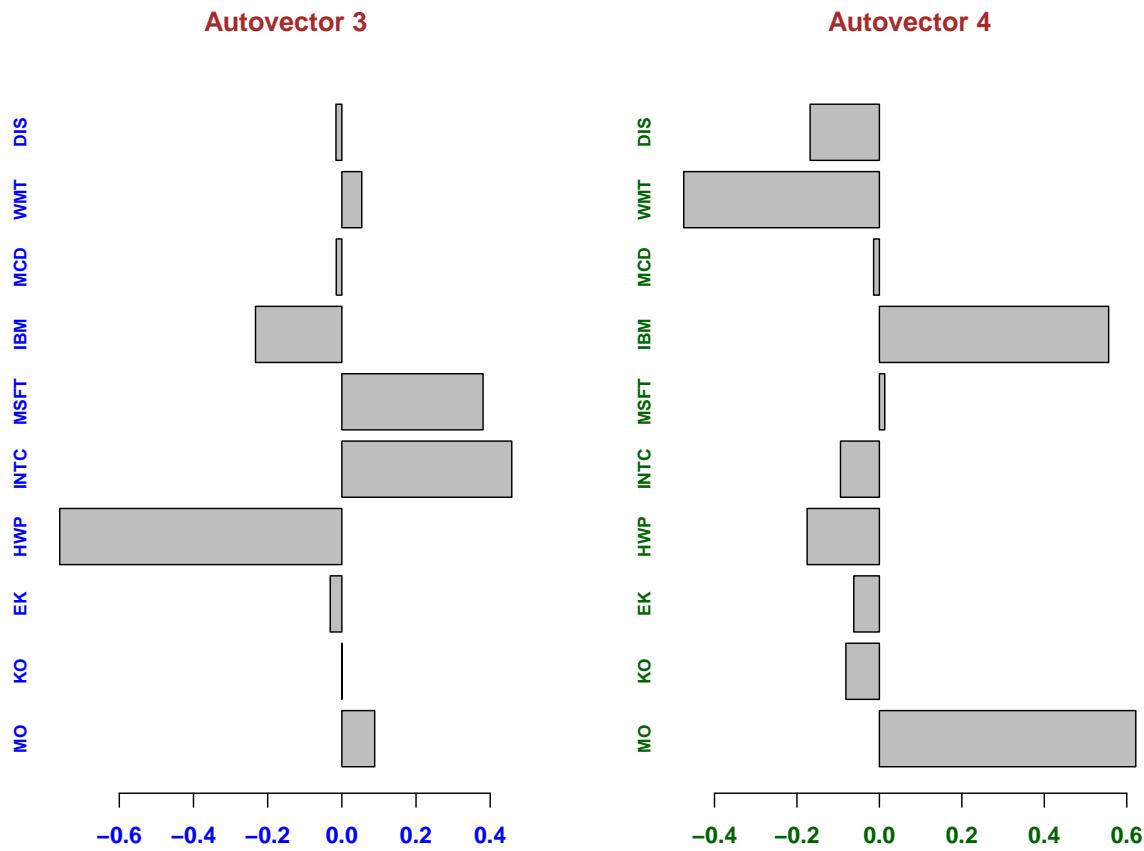


Y, a continuación, para los autovectores 3 y 4.

```
par(mfrow = c(1,2));

barplot(G[,3], horiz = TRUE, cex.names = 0.75, font = 2, col.axis = 'blue',
main = 'Autovector 3', col.main = 'brown');

barplot(G[,4], horiz = TRUE, cex.names = 0.75, font = 2, col.axis = 'darkgreen',
main = 'Autovector 4', col.main = 'brown');
```



4.5 Importancia de componentes

En esta sección se genera una tabla que muestra la varianza, su proporción y su proporción acumulada, creando previamente los objetos correspondientes:

```
var = with(summary(pc), sdev^2);
table = data.frame(var, var/sum(var), cumsum(var/sum(var)));
colnames(table) = c('var', 'prop', 'acum');
round(table, digits = 4);

##           var     prop    acum
## Comp.1   12.3485 0.3525 0.3525
## Comp.2    4.0631 0.1160 0.4685
## Comp.3    3.1968 0.0913 0.5598
## Comp.4    2.7061 0.0773 0.6370
## Comp.5    2.6130 0.0746 0.7116
## Comp.6    2.6009 0.0743 0.7859
## Comp.7    2.1679 0.0619 0.8478
## Comp.8    2.0912 0.0597 0.9075
## Comp.9    1.8619 0.0532 0.9606
## Comp.10   1.3786 0.0394 1.0000
```

Gráfico de correlación entre las variables y las CP1 y CP2

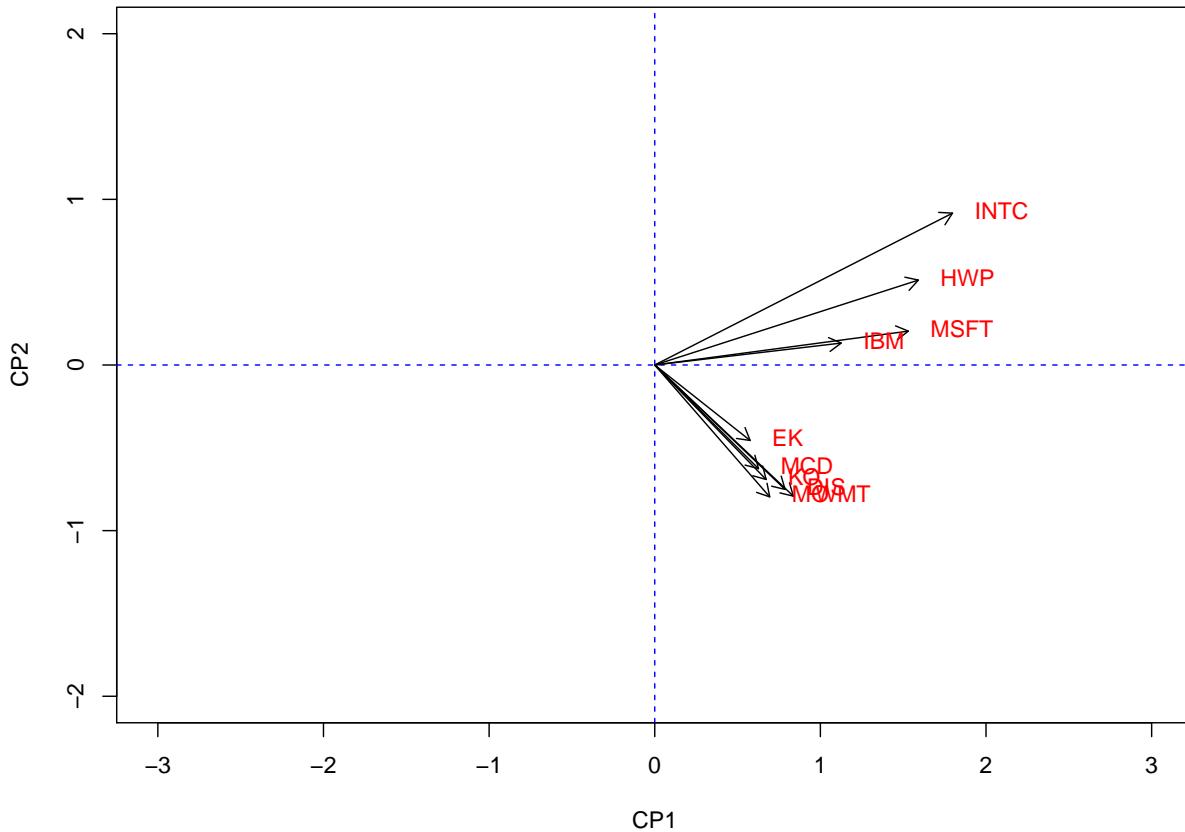
```

#valores sin escalar
ancp = prcomp(data.1);
corvar_ancp = ancp$rotation %*% diag(ancp$sdev);

#gráfico
plot(-2:2, -2:2,
      type = 'n',
      asp = 1,
      xlab = 'CP1',
      ylab = 'CP2');
abline(h = 0, v = 0, lty = 2, col = 4);

arrows(0, 0, corvar_ancp[,1], corvar_ancp[,2], length = .1);
text(corvar_ancp[,1], corvar_ancp[,2], colnames(data.1), pos = 4, offset = .8, col = 2, font = 0.3);

```



Existe una relación positiva entre las tecnológicas y las dos primeras componentes principales,

Gráfico de correlación entre las variables y las CP2 y CP3

```

#valores sin escalar
ancp = prcomp(data.1);
corvar_ancp = ancp$rotation %*% diag(ancp$sdev);

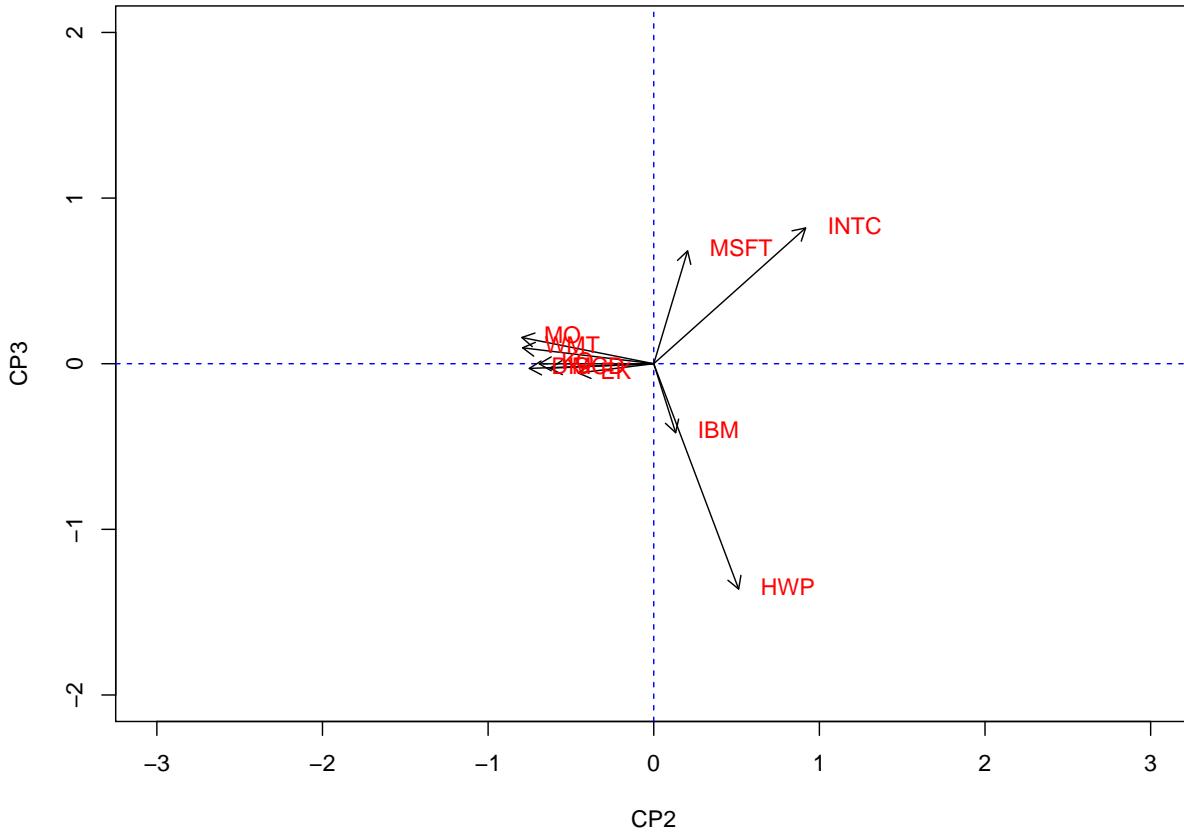
#gráfico

```

```

plot(-2:2, -2:2,
  type = 'n',
  asp = 1,
  xlab = 'CP2',
  ylab = 'CP3');
abline(h = 0, v = 0, lty = 2, col = 4);
arrows(0, 0, corvar_ancp[,2], corvar_ancp[,3], length = .1);
text(corvar_ancp[,2], corvar_ancp[,3], colnames(data.1), pos = 4, offset = .8, col = 2, font = 0.3);

```



Las direcciones se mantienen positivas en la CP3 para Microsoft e Intel.

Gráfico de correlación entre las variables y las CP3 y CP4

```

#valores sin escalar
ancp = prcomp(data.1);
corvar_ancp = ancp$rotation %*% diag(ancp$sdev);

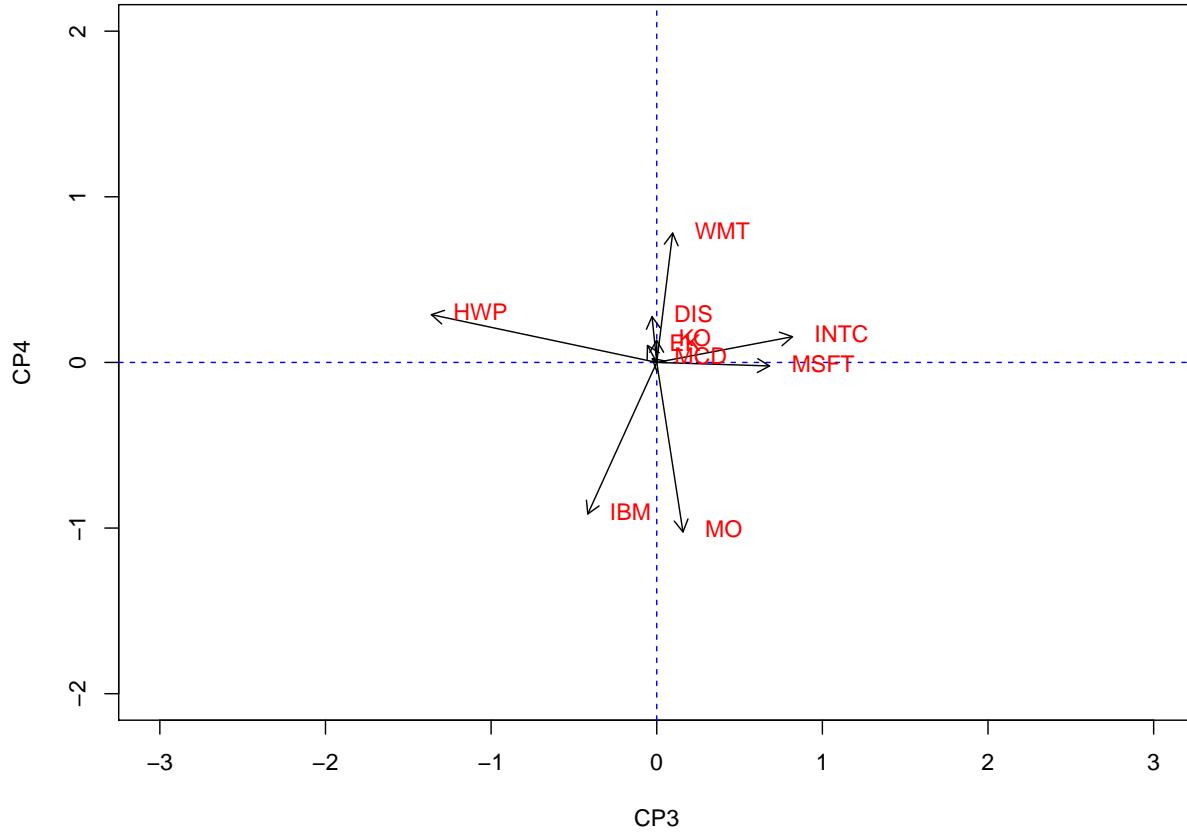
#gráfico
plot(-2:2, -2:2,
  type = 'n',
  asp = 1,
  xlab = 'CP3',
  ylab='CP4');
abline(h = 0, v = 0, lty = 2, col = 4);

```

```

arrows(0, 0, corvar_ancp[,3], corvar_ancp[,4], length=.1);
text(corvar_ancp[,3], corvar_ancp[,4], colnames(data.1), pos = 4, offset = .8, col = 2, font = 0.3);

```



No se observa una interpretación clara para las CP3 y CP4.

4.6 Diagrama de barras

Finalmente se muestra un gráfico de barras con las varianzas de todas las componentes principales. Para ello se crean dos matrices con los datos de la tabla de proporciones acumuladas,

```

z = matrix(as.matrix(table[1]), nrow = 1);
u = matrix(as.matrix(table[3]), nrow = 1);

```

Y se crea un objeto gráfico “bp”

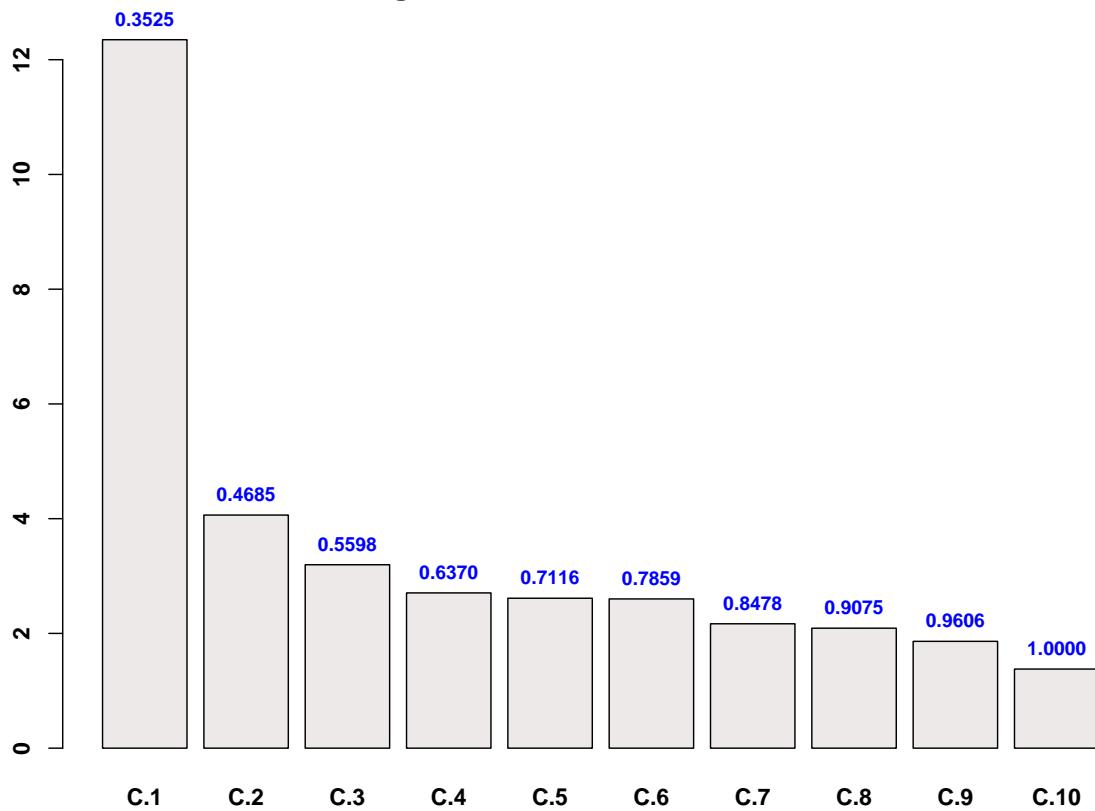
```

bp = barplot(z,
  col = 'snow2',
  main = 'Diagrama de barras de autovalores',
  font = 2,
  cex.main = 1.3,
  names.arg = paste0('C.',c(1:10)));

text(x = bp, y = z, labels = format(round(u, digits = 4)),
cex = 0.85, pos = 3, xpd = NA, font = 2, col = 4);

```

Diagrama de barras de autovalores



Las alturas de las barras son las varianzas y las los valores sobre las barras, representan la la **varianza acumulada**.

4.7 Conclusiones

- La CP1 es ponderada positivamente para todas las variables y puede considerarse que describe un tipo de cartera de índices. No suman uno pero, mediante un escalado podría conseguirse.
- La CP2 tiene pesos positivos para las compañías de consumo y pesos negativos para las compañías de tecnología.
- Podrían utilizarse las dos primeras CP para definir los factores de variación aunque sólo consiguen explicar el 46% de la varación y hasta la quinta componente principal no se recoge el 71% de variación.
- La mayor variabilidad es de una cartera promediada, donde se producen coordenadas negativas para las compañías tecnológicas y positivas para las de consumo. Ello llevaría a proponer de venta corta de la tecnología para comprar los títulos de consumo.
- La CP3 parecen mostrar algún tipo de relación o competencia interna entre Microsoft e Intel.