

# Reducción de la factorialidad. Análisis de Componentes Principales.

Ricardo Alberich, Juan Gabriel Gomila y Arnau Mir

# Introducción

- Uno de los problemas centrales del análisis de datos es la **reducción de la dimensionalidad**.
- Este concepto consiste en describir con cierta precisión los valores de las  $p$  variables por un pequeño subconjunto  $r < p$  de ellas con una pérdida mínima de información.
- Éste es el objetivo del **análisis de componentes principales**: dadas  $n$  observaciones de  $p$  variables se analiza, **si es razonable**, representar esta información en un **espacio con menos variables**.
- Para alcanzar dicho objetivo, vamos a realizar un **ajuste ortogonal por mínimos cuadrados**.

# Análisis de Componentes Principales

# Introducción: Matriz (tabla) de datos.

| Ind.     | $x_1$     | $x_2$     | $\dots$  | $x_p$     | $v_1$    | $v_2$    |
|----------|-----------|-----------|----------|-----------|----------|----------|
| 1        | $x_{11}$  | $x_{12}$  | $\dots$  | $x_{1p}$  | $v_{11}$ | $v_{12}$ |
| 2        | $x_{21}$  | $x_{22}$  | $\dots$  | $x_{2p}$  | $v_{21}$ | $v_{22}$ |
| 3        | $x_{31}$  | $x_{32}$  | $\dots$  | $x_{3p}$  | $v_{31}$ | $v_{32}$ |
| $\vdots$ | $\vdots$  | $\vdots$  | $\vdots$ | $\ddots$  | $\vdots$ | $\vdots$ |
| $n$      | $x_{n1}$  | $x_{n2}$  | $\dots$  | $x_{np}$  | $v_{n1}$ | $v_{n2}$ |
| $su_1$   | $su_{11}$ | $su_{12}$ | $\dots$  | $su_{1p}$ |          |          |
| $su_2$   | $su_{21}$ | $su_{22}$ | $\dots$  | $su_{2p}$ |          |          |

# Introducción: Matriz (tabla) de datos.

- Donde las variables  $x_1, \dots, x_n$  describen una concepto común de los  $n$  individuos observados.
- Las variables  $v_1, v_2$  son de perfil (o explicativas) y los individuos  $s_1, s_2$  son individuos suplementarios o ilustrativos.
- Tanto los individuos como las variables suplementarias ayudan a interpretar la variabilidad de los datos.

# Objetivos del análisis

## Objetivos del análisis

- Reducción de la dimensionalidad (factorialidad).
- Lo que se busca es un espacio de variables más reducido y fácil de interpretar.
- El problema es que si reducimos el número de variables es posible que perdamos parte toda la variabilidad de los datos originales.
- Así la idea básica es consentir una pérdida de información para lograr una ganancia en la significación.

# Análisis Factorial

- Algunos autores consideran el **ACP como una parte del Análisis Factorial**.
- En las **técnicas de Análisi Factorial** se postula que **la variabilidad total** se puede explicar mediante distintos tipos de factores:
- **factores comunes** subyacentes ( $F_i$ ).
- **factores específicos** de las variables ( $E_i$ ).
- **Error o fluctuaciones aleatorias** ( $A_i$ ).

$$X_1 = \alpha_{11}F_1 + \alpha_{12}F_2 + \cdots + \alpha_{1k}F_k + E_1 + A_1$$

$$X_2 = \alpha_{21}F_1 + \alpha_{22}F_2 + \cdots + \alpha_{2k}F_k + E_2 + A_2$$

.....

$$X_p = \alpha_{p1}F_1 + \alpha_{p2}F_2 + \cdots + \alpha_{pk}F_k + E_p + A_p$$

# Análisis Factorial

- Podríamos decir que en un Análisis Factorial se fija a priori la cantidad de varianza de cada variable que debe quedar interpretada por los factores comunes.
- Este valor recibe el nombre de comunalidad y se suele representar como  $h_i^2$ .

Así tenemos\_\_

- La comunalidad de la variable  $X_i$ ,  $h_i^2$ , es la varianza explicada por  $F_1, F_2, \dots, F_k$ .
- La diferencia  $s_i^2 - h_i^2$  es la varianza de la variable  $X_i$  que explican los factores específicos y aleatorios.

Var. observada = Var. común + Var. específica y aleatorios.



## El problema de los Componentes Principales

# El problema de los Componentes Principales

Todos los factores son comunes

$$X_1 = \alpha_{11}CP_1 + \alpha_{12}CP_2 + \cdots + \alpha_{1p}CP_p$$

$$X_2 = \alpha_{21}CP_1 + \alpha_{22}CP_2 + \cdots + \alpha_{2p}CP_p$$

.....

$$X_p = \alpha_{p1}CP_1 + \alpha_{p2}CP_2 + \cdots + \alpha_{pp}CP_p$$

Se trata de encontrar unas nuevas variables  $CP_1, \dots, CP_p$ , a las que llamaremos componentes principales, de forma que:

# El problema de los Componentes Principales

- Se cumplan las condiciones anteriores.
- El origen de las variables esté situado en el vector de medias o centro de gravedad de las observaciones.
- Sean incorreladas entre si  $Cor(CP_i, CP_j) = 0$  para  $i \neq j, i, j = 1, \dots, p$ .
- Se cumple que  $Var(CP_1) \geq Var(CP_2) \geq \dots \geq Var(CP_p)$  y hagan máximas estas varianzas.
- Se conserva la varianza total (inercia) de la nube de puntos.

## Tipos de A.C.P.

## Tipos de A.C.P:

- Sobre los datos centrados: a cada variable se le resta su media  $x_i - \bar{x}_i$ .
- Sobre los datos tipificados  $\frac{x_i - \bar{x}_i}{s_i}$ .
- En el primer caso las variables centradas tienen media cero y la misma varianza que las variables originales: se le suele llamar ACP de covarianzas.
- En el segundo caso las variables tipificadas tienen media cero y varianza 1: se le suele llamar ACP de correlaciones o normado.

## Tipos de A.C.P:

Recordemos que dada una matriz de datos  $\mathbf{X}$  ( $n \times p$  es decirde  $n$  individuos y  $p$  variables) representábamos por  $\tilde{\mathbf{X}}$  la matriz de datos centrada. Entonces:

- La matriz de covarianzas de  $\mathbf{X}$  viene dada por

$$\mathbf{S} = 1/n \tilde{\mathbf{X}}^T \tilde{\mathbf{X}}$$

- Si llamamos  $\mathbf{Z}$  a la tabla de los datos tipificados, la matriz de correlaciones viene dada por

$$\mathbf{R} = 1/n \mathbf{Z}^T \mathbf{Z}$$

## Propiedades

- Los **componentes principales vienen determinadas por los vectores propios ortonormales** (ordenados de mayor a menor valor propio) de la matriz de covarianzas (para datos centrados) y de la matriz de correlaciones (para los datos tipificados).
- Así en el ACP de covarianzas cada variable interviene con su propia varianza mientras que el ACP de correlaciones todas las variables tienen varianza 1.

ACP covarianzas:



## ACP covarianzas:

- Sea **S** la matriz de covarianzas de orden  $p$ . Calculamos sus valores propios

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$$

y los correspondientes vectores propios ortonormales (perpendiculares y de norma 1)

$$\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_p$$

- Las **direcciones de los componentes principales** quedan determinadas por su respectivo vector propio.
- **Cálculo de las coordenadas de la nueva matriz de datos respecto a las nuevas variables CP:**

# Ejemplo

Vamos a realizar un ACP sobre el ejemplo de la estatura de un niño recién nacido.

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | Sexo |
|-------|-------|-------|-------|------|
| 78    | 48.2  | 2.75  | 29.5  | Niña |
| 69    | 45.5  | 2.15  | 26.3  | Niña |
| 77    | 46.3  | 4.41  | 32.2  | Niña |
| 88    | 49    | 5.52  | 36.5  | Niño |
| 67    | 43    | 3.21  | 27.2  | Niña |
| 80    | 48    | 4.32  | 27.7  | Niña |
| 74    | 48    | 2.31  | 28.3  | Niña |
| 94    | 53    | 4.3   | 30.3  | Niño |
| 102   | 58    | 3.71  | 28.7  | Niño |

# Ejemplo

Donde:

- $x_1$  : edad en días
- $x_2$  : estatura al nacer en cm.
- $x_3$  : peso en Kg. al nacer
- $x_4$  : aumento en tanto por ciento de su peso con respecto de su peso al nacer. \*El sexo es una variable de perfil que intentaremos explicar con nuestro análisis de componentes principales.

## Código para la carga de datos

```
n = 9
p = 4
X = matrix(c(78,48.2,2.75,29.5,69,45.5,2.15,26.3,
77,46.3,4.41,32.2, 88,49,5.52,36.5, 67,43,3.21,27.2,
80,48,4.32,27.7, 74,48,2.31,28.3, 94,53,4.3,30.3,
102,58,3.71,28.7),nrow=n,byrow=T)
Datos= as.data.frame(X)
names(Datos) = paste("x",c(1:p),sep="")
Sexo = as.factor(c("Niña","Niña","Niña","Niño",
"Niña","Niña","Niña","Niño","Niño"))
Datos$Sexo=Sexo
```

# Código diagrama matricial

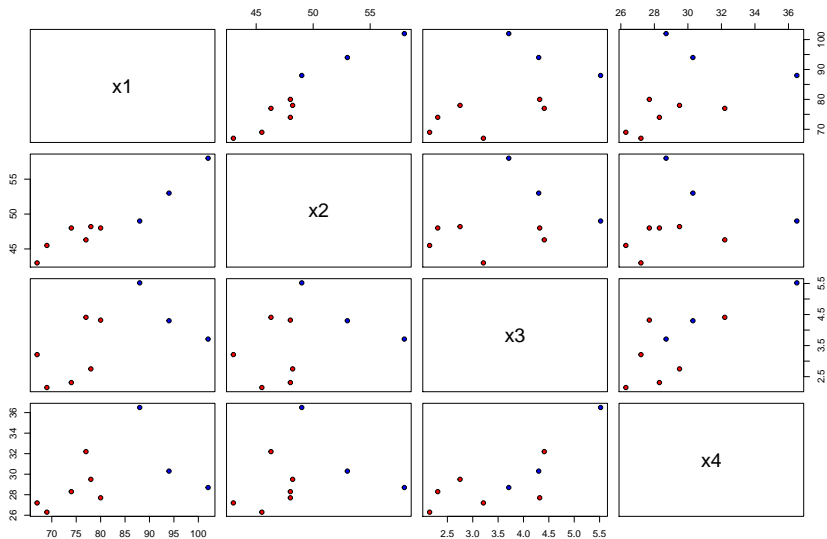
El siguiente código dibuja un diagrama matricial de las variables.

```
pairs(Datos[,1:4],pch=21,  
bg = c("red", "blue")[unclass(Datos$Sexo)],  
main="Diagrama matricial de las variables.  
\n Azul Niño, Rojo Niña")  
legend(15,-4,legend=levels(Datos$Sexo),pch=21,  
pt.bg=c("red", "blue"),title="Sexo")
```

que producen este gráfico...

# Diagrama matricial

Azul Niño, Rojo Niña



# Cálculos básicos

En lo que sigue todos los datos se redondean al tercer decimal.

Daremos el código de R que realiza el cálculo, en el código no se redondea: La matriz centrada de los datos anteriores es:

$$\tilde{\mathbf{X}} = \begin{pmatrix} -3.000 & -0.578 & -0.881 & -0.133 \\ -12.000 & -3.278 & -1.481 & -3.333 \\ -4.000 & -2.478 & 0.779 & 2.567 \\ 7.000 & 0.222 & 1.889 & 6.867 \\ -14.000 & -5.778 & -0.421 & -2.433 \\ -1.000 & -0.778 & 0.689 & -1.933 \\ -7.000 & -0.778 & -1.321 & -1.333 \\ 13.000 & 4.222 & 0.669 & 0.667 \\ 21.000 & 9.222 & 0.079 & -0.933 \end{pmatrix}$$







# Ejemplo

- La matriz de covarianzas de los datos anteriores es:

$$\mathbf{S} = \begin{pmatrix} 119.333 & 43.133 & 6.148 & 12.511 \\ 43.133 & 17.193 & 1.148 & 1.886 \\ 6.148 & 1.148 & 1.111 & 2.428 \\ 12.511 & 1.886 & 2.428 & 8.624 \end{pmatrix}$$

- Los valores propios son:

$$\lambda_1 = 136.615, \quad \lambda_2 = 8.861, \quad \lambda_3 = 0.738, \quad \lambda_4 = 0.047.$$

# Ejemplo

- Los vectores propios ortonormales correspondientes a los valores propios, son las columnas de la siguiente matriz:

$$\begin{pmatrix} 0.934 & -0.022 & 0.256 & 0.247 \\ 0.339 & 0.354 & -0.661 & -0.568 \\ 0.047 & -0.248 & 0.566 & -0.785 \\ 0.097 & -0.902 & -0.421 & -0.013 \end{pmatrix}$$

# Ejemplo

```
##           [,1]      [,2]      [,3]      [,4]
## [1,] 119.333333 43.133333 6.147778 12.511111
## [2,]  43.133333 17.192840 1.147802  1.886296
## [3,]   6.147778  1.147802 1.110610  2.427852
## [4,]  12.511111  1.886296 2.427852  8.624444

## eigen() decomposition
## $values
## [1] 136.61529623   8.86125966   0.73789460   0.04677667
##
## $vectors
##           [,1]      [,2]      [,3]      [,4]
## [1,] 0.93439437 -0.02238785  0.2555755  0.24715806
## [2,] 0.33947477  0.35413519 -0.6610845 -0.56772562
## [3,] 0.04701065 -0.24770838  0.5656945 -0.78512437
## [4,] 0.09723192 -0.90151407 -0.4214714 -0.01342527
```

# Ejemplo

- Las expresiones de las variables nuevas  $CP_i$  en función de las antiguas, notemos que se calculan sobre los datos centrados, son:

$$CP_1 = 0.934 \cdot \tilde{X}_1 + 0.339 \cdot \tilde{X}_2 + 0.047 \cdot \tilde{X}_3 \\ + 0.097 \cdot \tilde{X}_4,$$

$$CP_2 = -0.022 \cdot \tilde{X}_1 + 0.354 \cdot \tilde{X}_2 - 0.248 \cdot \tilde{X}_3 \\ - 0.902 \cdot \tilde{X}_4,$$

$$CP_3 = 0.256 \cdot \tilde{X}_1 - 0.661 \cdot \tilde{X}_2 + 0.566 \cdot \tilde{X}_3 \\ - 0.421 \cdot \tilde{X}_4,$$

$$CP_4 = 0.247 \cdot \tilde{X}_1 - 0.568 \cdot \tilde{X}_2 - 0.785 \cdot \tilde{X}_3 \\ - 0.013 \cdot \tilde{X}_4.$$

# Ejemplo

- La nueva matriz de datos respecto de las nuevas variables será:

$$\mathbf{CP} = \tilde{\mathbf{X}}_{\mathbf{u}} = \begin{pmatrix} -3.054 & 0.201 & -0.827 & 0.280 \\ -12.719 & 2.480 & -0.333 & 0.103 \\ -4.293 & -3.295 & -0.025 & -0.228 \\ 7.373 & -6.736 & -0.183 & 0.029 \\ -15.299 & 0.565 & 1.029 & 0.183 \\ -1.354 & 1.319 & 1.463 & -0.321 \\ -6.997 & 1.411 & -1.460 & -0.233 \\ 13.677 & 0.437 & 0.629 & 0.282 \\ 22.666 & 3.618 & -0.292 & -0.095 \end{pmatrix}$$

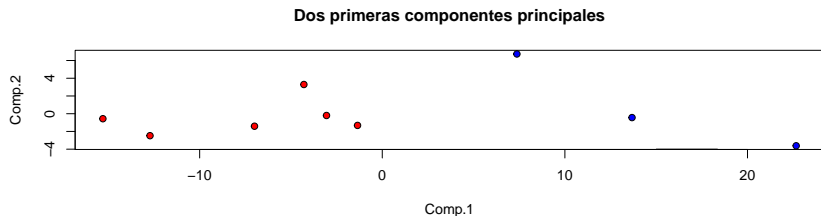
## Ejemplo

- Se puede observar que si se multiplican escalarmente dos columnas cualesquiera, el resultado es nulo. Es decir, las columnas de la nueva matriz de datos son ortogonales dos a dos.

| ## |      | [,1]       | [,2]       | [,3]        | [,4]        |
|----|------|------------|------------|-------------|-------------|
| ## | [1,] | -3.053748  | 0.2006059  | -0.82694086 | 0.28015137  |
| ## | [2,] | -12.719227 | 2.4794016  | -0.33290064 | 0.10262352  |
| ## | [3,] | -4.292580  | -3.2951470 | -0.02537552 | -0.22788270 |
| ## | [4,] | 7.372619   | -6.7367152 | -0.18337900 | 0.02878003  |
| ## | [5,] | -15.299363 | 0.5649058  | 1.02897164  | 0.18330688  |
| ## | [6,] | -1.354065  | 1.3188263  | 1.46321590  | -0.32046716 |
| ## | [7,] | -6.996583  | 1.4101388  | -1.46016599 | -0.23337069 |
| ## | [8,] | 13.676693  | 0.4370899  | 0.62871103  | 0.28191432  |
| ## | [9,] | 22.665914  | 3.6172335  | -0.29151312 | -0.09474550 |

# Ejemplo

Como podemos observar, nuestro análisis ha explicado la variable de perfil sexo ya que distingue entre niños y niñas con las dos primeras componentes.



El siguiente código dibuja todos los componentes

```
pairs(solacp$scores,pch=21,  
bg = c("red", "blue")[unclass(Datos$Sexo)],  
main="Diagrama matricial de  
los componentes principales")
```



ACP correlaciones.

## ACP correlaciones.

Sea  $\mathbf{R}$  la matriz de correlaciones de orden  $p$ . Calcularemos sus valores propios

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$$

y los correspondientes vectores propios ortonormales.

$$\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_p$$

Las direcciones de los componentes principales quedan determinadas por el vector propio correspondiente.

Cálculo de las coordenadas de la nueva matriz de datos respecto de las nuevas variables CP:

$$\mathbf{CP} = \mathbf{Z}\mathbf{u},$$

donde  $\mathbf{Z}$  es la matriz de datos tipificados y  $\mathbf{u}$  es la matriz de los vectores propios.

## Ejemplo

*Realicemos un análisis ACP de correlaciones con el ejemplo anterior.*  
La matriz tipificada de datos es:

$$\mathbf{Z} = \begin{pmatrix} -0.275 & -0.139 & -0.836 & -0.045 \\ -1.099 & -0.791 & -1.405 & -1.135 \\ -0.366 & -0.598 & 0.739 & 0.874 \\ 0.641 & 0.054 & 1.792 & 2.338 \\ -1.282 & -1.393 & -0.400 & -0.829 \\ -0.092 & -0.188 & 0.654 & -0.658 \\ -0.641 & -0.188 & -1.254 & -0.454 \\ 1.190 & 1.018 & 0.635 & 0.227 \\ 1.922 & 2.224 & 0.075 & -0.318 \end{pmatrix}$$

## Ejemplo

La matriz de correlaciones **R** vale, en este caso:

$$\mathbf{R} = \begin{pmatrix} 1.000 & 0.952 & 0.534 & 0.390 \\ 0.952 & 1.000 & 0.263 & 0.155 \\ 0.534 & 0.263 & 1.000 & 0.784 \\ 0.390 & 0.155 & 0.784 & 1.000 \end{pmatrix}$$

Los valores propios de dicha matriz son:

$$2.560, \quad 1.229, \quad 0.208, \quad 0.00325.$$

\*La matriz de los vectores propios es:

$$\begin{pmatrix} 0.573 & 0.359 & -0.038 & 0.736 \\ 0.478 & 0.578 & 0.145 & -0.646 \\ 0.499 & -0.459 & -0.707 & -0.201 \\ 0.442 & -0.572 & 0.691 & -0.029 \end{pmatrix}$$

## Ejemplo

\*Las expresiones de las variables nuevas  $CP_i$  en función de las antiguas  $Z_i$  son:

$$CP_1 = 0.573 \cdot Z_1 + 0.478 \cdot Z_2 + 0.499 \cdot Z_3 \\ + 0.442 \cdot Z_4,$$

$$CP_2 = 0.359 \cdot Z_1 + 0.578 \cdot Z_2 - 0.459 \cdot Z_3 \\ - 0.572 \cdot Z_4,$$

$$CP_3 = -0.038 \cdot Z_1 + 0.145 \cdot Z_2 - 0.707 \cdot Z_3 \\ + 0.691 \cdot Z_4,$$

$$CP_4 = 0.736 \cdot Z_1 - 0.646 \cdot Z_2 - 0.201 \cdot Z_3 \\ - 0.029 \cdot Z_4.$$

## Ejemplo

\*La nueva matriz de datos respecto de las nuevas variables será:

$$\mathbf{CP} = \mathbf{Zu} = \begin{pmatrix} -0.661 & 0.231 & 0.550 & 0.057 \\ -2.209 & 0.443 & 0.137 & 0.018 \\ 0.259 & -1.316 & 0.008 & -0.058 \\ 2.319 & -1.899 & 0.332 & 0.008 \\ -1.965 & -0.608 & -0.444 & 0.061 \\ -0.107 & -0.065 & -0.941 & -0.058 \\ -1.282 & 0.497 & 0.570 & -0.085 \\ 1.585 & 0.594 & -0.189 & 0.084 \\ 2.061 & 2.122 & -0.023 & -0.027 \end{pmatrix}$$

\*Se puede observar que si calculamos el producto escalar de dos columnas cualesquiera, el resultado es nulo. Es decir, las columnas de la nueva matriz de datos son ortogonales dos a dos.

Propiedades ACP covarianzas.

# Propiedades ACP covarianzas.

Sea  $\mathbf{X}$  una matriz de datos  $n \times p$  y sea

$$\mathbf{S} = \begin{pmatrix} s_1^2 & s_{12} & \dots & s_{1p} \\ s_{21} & s_2^2 & \dots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \dots & s_p^2 \end{pmatrix}$$

su matriz de covarianzas.

Recordemos que  $s_i^2$  es la varianza de la variable  $\mathbf{x}_i$  y que  $s_{ij}$  son las covarianzas de la variables  $\mathbf{x}_i$  y  $\mathbf{x}_j$ .

Además la Varianza Total =  $tr(\mathbf{S}) = \sum_{i=1}^p s_i^2$



# Propiedades ACP covarianzas.

- $Var(\mathbf{CP}_i) = \lambda_i$ . La varianza de cada componente principal es su valor propio.
- $\sum_{i=1}^n Var(\mathbf{CP}_i) = \sum_{i=1}^n \lambda_i = tr(\mathbf{S}) = \sum_{i=1}^n s_i^2$ . Por lo tanto los componentes principales reproducen la varianza total
- Los componentes principales tienen correlación cero entre sí (son *incorrelados*) por lo tanto su matriz de covarianzas es

$$\mathbf{S}_{CP} = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \lambda_p \end{pmatrix}$$

## Propiedades ACP covarianzas.

- $\det(\mathbf{S}_{CP}) = \prod_{i=1}^n \lambda_i = \det(\mathbf{S})$ . Luego los componentes principales conservan la varianza generalizada.
- La proporción de varianza explicada por la componente  $j$ -ésima es

$$\frac{\lambda_j}{\sum_{i=1}^n \lambda_i}.$$

Además al ser\* incorrelados\* la proporción de varianza explicada por los  $k$  primeros componentes es

$$\frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^n \lambda_i}.$$

- $\text{Cov}(\tilde{\mathbf{X}}_i, \mathbf{CP}_j) = \lambda_j u_{ji}$ ;  $\text{corr}(\tilde{\mathbf{X}}_i, \mathbf{CP}_j) = \frac{\sqrt{\lambda_j} u_{ji}}{s_i}$  donde  $u_{ji}$  es la  $i$ -ésima componente del vector propio  $\mathbf{u}_j$ .

## Ejemplo

Vamos a comprobar las propiedades anteriores con nuestro ejemplo. Recordemos las matrices de datos de las variables originales **X** (centradas) y de las variables en componentes principales **CP**:

$$\tilde{\mathbf{X}} = \begin{pmatrix} -3.000 & -0.578 & -0.881 & -0.133 \\ -12.000 & -3.278 & -1.481 & -3.333 \\ -4.000 & -2.478 & 0.779 & 2.567 \\ 7.000 & 0.222 & 1.889 & 6.867 \\ -14.000 & -5.778 & -0.421 & -2.433 \\ -1.000 & -0.778 & 0.689 & -1.933 \\ -7.000 & -0.778 & -1.321 & -1.333 \\ 13.000 & 4.222 & 0.669 & 0.667 \\ 21.000 & 9.222 & 0.079 & -0.933 \end{pmatrix},$$
$$\mathbf{CP} = \begin{pmatrix} -3.054 & 0.201 & -0.827 & 0.280 \\ -12.719 & 2.480 & -0.333 & 0.103 \\ -4.293 & -3.295 & -0.025 & -0.228 \\ 7.373 & -6.736 & -0.183 & 0.029 \\ -15.299 & 0.565 & 1.029 & 0.183 \\ -1.354 & 1.319 & 1.463 & -0.321 \\ -6.997 & 1.411 & -1.460 & -0.233 \\ 13.677 & 0.437 & 0.629 & 0.282 \\ 22.666 & 3.618 & -0.292 & -0.095 \end{pmatrix}.$$

# Ejemplo

- La matriz de los vectores propios de la matriz **S** era:

$$\begin{pmatrix} 0.934 & -0.022 & 0.256 & 0.247 \\ 0.339 & 0.354 & -0.661 & -0.568 \\ 0.047 & -0.248 & 0.566 & -0.785 \\ 0.097 & -0.902 & -0.421 & -0.013 \end{pmatrix}.$$

- Las varianzas de las variables *CP* son las siguientes:

$$\begin{aligned} \text{Var}(\mathbf{CP}_1) &= 136.615, & \text{Var}(\mathbf{CP}_2) &= 8.861, \\ \text{Var}(\mathbf{CP}_3) &= 0.738, & \text{Var}(\mathbf{CP}_4) &= 0.0468, \end{aligned}$$

que son los valores propios de la matriz de covarianzas **S**.

- La traza de la matriz **S** vale:  $\text{tr}(\mathbf{S}) = 146.261$ . Si sumamos los 4 valores propios, su valor coincide con la suma de los valores propios:

$$\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 146.261.$$

## Ejemplo

- La **matriz de covarianzas de las variables CP** es:

$$\text{cov}(\mathbf{CP}) = \begin{pmatrix} 136.615 & 0.000 & 0.000 & 0.000 \\ 0.000 & 8.861 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.738 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.047 \end{pmatrix}$$

Podemos observar que es una matriz diagonal con los valores propios de la matriz **S** en la diagonal.

- El **determinante de las matrices de covarianzas de  $\tilde{\mathbf{X}}$  y CP** vale 41.785, valor que **coincide con el producto de los valores propios de la matriz S**:

$$\prod_{i=1}^4 \lambda_i = 136.615 \cdot 8.861 \cdot 0.738 \cdot 0.0468 = 41.785.$$

# Ejemplo

La proporción de varianza explicada por los componentes es:

| Variables                   | Varianza Explicada                          |
|-----------------------------|---|
| <b>CP<sub>1</sub></b>       | $136.615/146.261 = 0.934$                   |
| <b>CP<sub>1,2</sub></b>     | $(136.615 + 8.861)/146.261 = 0.995$         |
| <b>CP<sub>1,2,3</sub></b>   | $(136.615 + 8.861 + 0.738)/146.261 = 0.999$ |
| <b>CP<sub>1,2,3,4</sub></b> | 1   |

# Ejemplo

- La matriz de covarianzas entre las variables  $\tilde{\mathbf{X}}$  y  $\mathbf{CP}$  vale:

$$\text{cov}(\tilde{\mathbf{X}}, \mathbf{CP}) = \begin{pmatrix} 127.653 & -0.198 & 0.189 & 0.012 \\ 46.377 & 3.138 & -0.488 & -0.027 \\ 6.422 & -2.195 & 0.417 & -0.037 \\ 13.283 & -7.989 & -0.311 & -0.001 \end{pmatrix}$$

Recuperemos la matriz de vectores propios de la matriz  $\mathbf{S}$ :

$$\begin{pmatrix} 0.934 & -0.022 & 0.256 & 0.247 \\ 0.339 & 0.354 & -0.661 & -0.568 \\ 0.047 & -0.248 & 0.566 & -0.785 \\ 0.097 & -0.902 & -0.421 & -0.013 \end{pmatrix}.$$

## Ejemplo

Si multiplicamos la primera columna de la matriz anterior

$$\begin{pmatrix} 0.934 \\ 0.339 \\ 0.047 \\ 0.097 \end{pmatrix}$$

por el valor propio 136.615 de la matriz **S** obtenemos la primera columna de la matriz

$\text{Cov}(\tilde{\mathbf{X}}, \mathbf{CP})$ :

$$136.615 \cdot \begin{pmatrix} 0.934 \\ 0.339 \\ 0.047 \\ 0.097 \end{pmatrix} = \begin{pmatrix} 127.652 \\ 46.377 \\ 6.422 \\ 13.283 \end{pmatrix}$$



# Ejemplo

- En general, tenemos que

$$\mathbf{u} \cdot \text{diag}(\lambda) = \text{Cov}(\tilde{\mathbf{X}}, \mathbf{CP}),$$

donde  $\mathbf{u}$  es la matriz formada por los vectores propios de la matriz  $\mathbf{S}$  y  $\text{diag}(\lambda)$  es una matriz diagonal con los valores propios de la matriz  $\mathbf{S}$  en la diagonal.

# Propiedades ACP covarianzas.

- La **primer componente principal** es la recta que **conserva mayor inercia** de la nube de puntos.
- Las **dos primeras componentes** principales forman el **plano** que conserva **mayor inercia** de la nube de puntos.
- Lo mismo sucede con los espacios formados por las  $k$  primeras componentes

Propiedades ACP correlaciones.

## Propiedades ACP correlaciones.

Sea  $\mathbf{X}$  una matriz de datos  $n \times p$  y sea

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{pmatrix}$$

Su matriz de correlaciones. Se verifican las siguientes propiedades:

# Propiedades ACP correlaciones

- Recordemos que la diagonal es 1 pues es la varianza de los datos tipificados y que  $r_{ij}$  son las correlaciones lineales de las variables  $\mathbf{x}_i$  y  $\mathbf{x}_j$ .
- Además la Varianza Total =  $tr(\mathbf{R}) = p$
- $Var(\mathbf{CP}_i) = \lambda_i$ . El valor propio del componente es igual a su varianza
- $\sum_{i=1}^n var(\mathbf{CP}_i) = \sum_{i=1}^n \lambda_i = tr(\mathbf{R}) = p$ . Por lo tanto los componentes principales reproducen la varianza total y ésta es igual al número de variables  $p$ .

## Propiedades ACP correlaciones.

- Los componentes principales tienen correlación cero entre sí (son *incorrelados*) por lo tanto su matriz de covarianzas ( que este caso es igual a la de correlaciones es

$$\mathbf{S}_{CP} = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_p \end{pmatrix}$$

# Propiedades ACP correlaciones.

- $\det(\mathbf{S}_{CP}) = \prod_{i=1}^n \lambda_i = \det(\mathbf{R})$ . Luego los componentes principales conservan la varianza generalizada.
- La proporción de varianza explicada por cada componente es

$$\frac{\lambda_i}{p}.$$

Además al ser *incorreladas* la proporción de varianza explicada por los  $k$  primeros componentes es

$$\frac{\sum_{i=1}^k \lambda_i}{p}.$$

- $\text{corr}(\mathbf{Z}_i, \mathbf{CP}_j) = \sqrt{\lambda_j} \cdot u_{ji}$  donde  $u_{ji}$  es la  $i$ -ésima componente del vector propio  $\mathbf{u}_j$ .

## Propiedades ACP correlaciones.

Vamos a comprobar las propiedades anteriores con nuestro ejemplo. Recordemos las matrices de datos estandarizada **Z** y de las variables en componentes principales **CP**:

$$\mathbf{Z} = \begin{pmatrix} -0.275 & -0.139 & -0.836 & -0.045 \\ -1.099 & -0.791 & -1.405 & -1.135 \\ -0.366 & -0.598 & 0.739 & 0.874 \\ 0.641 & 0.054 & 1.792 & 2.338 \\ -1.282 & -1.393 & -0.400 & -0.829 \\ -0.092 & -0.188 & 0.654 & -0.658 \\ -0.641 & -0.188 & -1.254 & -0.454 \\ 1.190 & 1.018 & 0.635 & 0.227 \\ 1.922 & 2.224 & 0.075 & -0.318 \end{pmatrix}$$



## Ejemplo

$$\mathbf{CP} = \begin{pmatrix} -0.661 & 0.231 & 0.550 & 0.057 \\ -2.209 & 0.443 & 0.137 & 0.018 \\ 0.259 & -1.316 & 0.008 & -0.058 \\ 2.319 & -1.899 & 0.332 & 0.008 \\ -1.965 & -0.608 & -0.444 & 0.061 \\ -0.107 & -0.065 & -0.941 & -0.058 \\ -1.282 & 0.497 & 0.570 & -0.085 \\ 1.585 & 0.594 & -0.189 & 0.084 \\ 2.061 & 2.122 & -0.023 & -0.027 \end{pmatrix}.$$

# Ejemplo

Las varianzas de las variables  $\mathbf{CP}_i = \lambda_i$  son las siguientes:

$$\begin{aligned}\text{Var}(\mathbf{CP}_1) &= 2.560, & \text{Var}(\mathbf{CP}_2) &= 1.229, \\ \text{Var}(\mathbf{CP}_3) &= 0.208, & \text{Var}(\mathbf{CP}_4) &= 0.00325,\end{aligned}$$

Estos valores son los valores propios de la matriz  $\mathbf{R}$ .

## Ejemplo

Se puede comprobar que su suma vale 4, que es el valor de  $p$  en nuestro caso.

Si calculamos la matriz de covarianzas de las variables **CP** obtenemos una matriz diagonal que son los valores propios de la matriz **R** calculados anteriormente:

$$\text{Cov}(\mathbf{CP}) = \mathbf{S}_{\mathbf{CP}} = \begin{pmatrix} 2.560 & 0.000 & 0.000 & 0.000 \\ 0.000 & 1.229 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.208 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.003 \end{pmatrix},$$

# Ejemplo

El determinante de la matriz  $\mathbf{S_{CP}}$  es:

$$\det(\mathbf{S_{CP}}) = 0.00213,$$

que coincide con el producto de los valores propios de la matriz  $\mathbf{R}$ :

$$\prod_{i=1}^4 \lambda_i = 2.560 \cdot 1.229 \cdot 0.208 \cdot 0.00325 = 0.00213.$$

# Ejemplo

La **proporción de varianza explicada por los componentes** es:

| Variables                   | Varianza Explicada                  |
|-----------------------------|-------------------------------------|
| <b>CP<sub>1</sub></b>       | $2.560/4 = 0.640$                   |
| <b>CP<sub>1,2</sub></b>     | $(2.560 + 1.229)/4 = 0.947$         |
| <b>CP<sub>1,2,3</sub></b>   | $(2.560 + 1.229 + 0.208)/4 = 0.999$ |
| <b>CP<sub>1,2,3,4</sub></b> | 1                                   |

## Ejemplo

La matriz de correlaciones entre las variables **Z** y **CP** es:

$$\text{Cor}(\mathbf{Z}, \mathbf{CP}) = \begin{pmatrix} 0.916 & 0.398 & -0.017 & 0.042 \\ 0.764 & 0.641 & 0.066 & -0.037 \\ 0.798 & -0.509 & -0.323 & -0.011 \\ 0.706 & -0.634 & 0.315 & -0.002 \end{pmatrix}.$$

La matriz de vectores propios de la matriz **R** es:

$$\begin{pmatrix} 0.573 & 0.359 & -0.038 & 0.736 \\ 0.478 & 0.578 & 0.145 & -0.646 \\ 0.499 & -0.459 & -0.707 & -0.201 \\ 0.442 & -0.572 & 0.691 & -0.029 \end{pmatrix}.$$

## Ejemplo

Si multiplicamos la primera columna de la matriz anterior

$$\begin{pmatrix} 0.573 \\ 0.478 \\ 0.499 \\ 0.442 \end{pmatrix}$$

por la raíz cuadrada del primer valor propio de la matriz **R**,  $\sqrt{2.560}$ , obtenemos la primera columna de la matriz

$$\text{Cor}(\mathbf{Z}, \mathbf{CP}),$$

efectivamente

## Ejemplo

$$\sqrt{2.560} \cdot \begin{pmatrix} 0.573 \\ 0.478 \\ 0.499 \\ 0.442 \end{pmatrix} = \begin{pmatrix} 0.916 \\ 0.764 \\ 0.798 \\ 0.706 \end{pmatrix}$$

$$\text{Cor}(\mathbf{Z}, \mathbf{CP}) = \begin{pmatrix} 0.916 & 0.398 & -0.017 & 0.042 \\ 0.764 & 0.641 & 0.066 & -0.037 \\ 0.798 & -0.509 & -0.323 & -0.011 \\ 0.706 & -0.634 & 0.315 & -0.002 \end{pmatrix}$$



# Ejemplo

- En general, podemos escribir:

$$\mathbf{u} \cdot \text{diag}(\sqrt{\lambda}) = \text{Cor}(\mathbf{Z}, \mathbf{CP}),$$

donde  $\mathbf{u}$  es la matriz formada por los vectores propios de la matriz  $\mathbf{R}$  y  $\text{diag}(\sqrt{\lambda})$  es una matriz diagonal con la raíz cuadrada de los valores propios de la matriz  $\mathbf{R}$  en la diagonal.

# Propiedades ACP correlaciones

- La primera componente principal es la recta que conserva mayor inercia de la nube de puntos.
- Los dos primeros componentes principales forman el plano que conserva mayor inercia de la nube de puntos.
- Lo mismo sucede con los espacios formados por los  $k$  primeros componentes

## Etapas de un ACP

# Etapas de un ACP

- Determinar las variables e individuos que intervienen en el análisis, las variables de perfil y los individuos ilustrativos.
- Decidir si se realiza el análisis sobre los datos brutos (matriz de covarianzas) o sobre los datos tipificados (matriz de correlaciones).
- Cuando las variables originales  $\mathbf{X}$  están medidas en distintas unidades, conviene aplicar el análisis de correlaciones. Si están en las mismas unidades, ambas alternativas son posibles.
- Si las diferencias entre las varianzas son informativas y queremos tenerlas en cuenta en el análisis, no debemos estandarizar las variables.

# Etapas de un ACP

- Reducción de la dimensionalidad; tenemos que decidir cuántas componente retenemos. La cantidad de varianza retenida es:

| Comp.    | Valor propio | Cantidad retenida  |
|----------|--------------|--|
| $Cp_1$   | $\lambda_1$  | $\lambda_1 / \sum_{i=1}^p \lambda_i$                           |
| $Cp_2$   | $\lambda_2$  | $(\lambda_1 + \lambda_2) / \sum_{i=1}^p \lambda_i$             |
| $Cp_3$   | $\lambda_3$  | $(\lambda_1 + \lambda_2 + \lambda_3) / \sum_{i=1}^p \lambda_i$ |
| $\vdots$ | $\vdots$     | $\vdots$   |
| $Cp_p$   | $\lambda_p$  | $(\lambda_1 + \dots + \lambda_p) / \sum_{i=1}^p \lambda_i = 1$ |

## Retención de componentes

# Retención de componentes

Una vez realizado el ACP tengo que decidir que número de componentes se retienen. Existen diversos métodos: **Seleccionar una proporción fija de varianza**. Seleccionar componentes hasta cubrir una proporción determinada de varianza, como el 80% o el 90%.

- En el ejemplo que hemos desarrollado, tenemos que con un análisis de covarianzas, si sólo elegimos la primera componente, cubrimos el 93.4% de la varianza. Si elegimos, las dos primeras, cubrimos el 99.5% de la varianza. Con las tres primeras, cubrimos el 99.9% de la varianza.
- En cambio, con un análisis de correlaciones, con la primera componente, sólo cubrimos el 64% de la varianza; con las dos primeras, el 94.7% de la varianza y con las tres primeras, el 99.9% de la varianza.

## Técnicas de retención de retención de componentes



# Retención de componentes

## Método de la Media aritmética.

- Se retienen todas las componentes **CP<sub>i</sub>** que cumplan

$$\lambda_i \geq \bar{\lambda} = \frac{\sum_{i=1}^p \lambda_i}{p}$$

- En el caso del análisis de correlaciones, la condición anterior equivale a retener los componentes con valores propios mayores que ~1.

# Retención de componentes

## Método de la Media aritmética.

- En nuestro ejemplo, para el análisis de covarianzas, tenemos que:  $\bar{\lambda} = 36.565$ . Recordemos que los valores propios de la matriz de covarianzas **S** son:

136.615, 8.861, 0.738, 0.0468.

Por tanto, tenemos que retener sólo la componente **CP<sub>1</sub>**.

- Para el análisis de correlaciones, recordemos que los valores propios de la matriz **R** son:

2.560, 1.229, 0.208, 0.00324.

En este caso, tenemos que retener los componentes **CP<sub>1</sub>** y **CP<sub>2</sub>**.

## Gráfico de sedimentación, regla del codo

- Gráfico de sedimentación (*screeplot*) es una técnica gráfica de para la retención de componentes.
- Se representan los vectores propios ordenados de mayor a menor unidos por una poligonal o simplemente un diagrama de barras.
- Se retienen los componente hasta el que *sedimenta*. El códigos el siguiente

## Gráfico de sedimentación, regla del codo, código

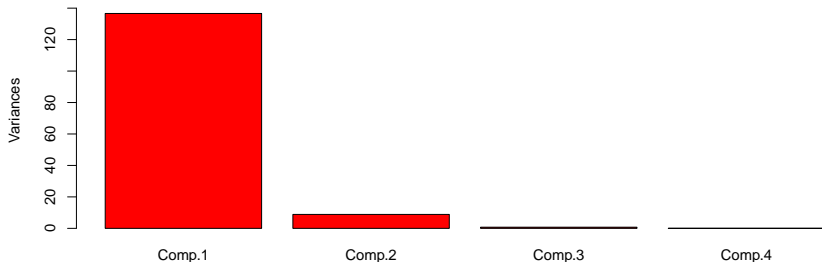
```
screeplot(solacp,type="lines",  
          main="Gráfico de sedimentación")  
screeplot(solacp,type="barplot",  
          main="Gráfico de sedimentación",ylim=c(0,150))
```

# Gráfico de sedimentación, regla del codo, poligonal.



# Gráfico de sedimentación, regla del codo, barras.

Gráfico de sedimentación



Hay muchas otras pruebas más como la pruebas de Hipótesis de Anderson:

$$\begin{cases} H_0 : \lambda_m = \dots = \lambda_p \\ H_1 : \text{no todos iguales} \end{cases}$$

## Adecuación de los datos al ACP

- Coeficiente de adecuación muestral (Kaiser Meyer y Olkin):

$$KMO = \frac{\sum_j \sum_{i \neq j} r_{ij}^2}{\sum_j \sum_{i \neq j} r_{ij}^2 + \sum_j \sum_{i \neq j} a_{ij}^2}$$

donde  $r_{ij}$  son los coef. de correlación entre las variables  $i$  y  $j$ , mientras que los  $a_{ij}$  son los coef. de correlación parcial entre las variables  $i$  y  $j$  (equivalentes a las correlaciones entre los residuos de la regresión de estas dos variables con las restantes).



# Adecuación de los datos al ACP

- Niveles de  $KMO \geq 0.5$  son considerados aceptables.

En nuestro ejemplo, las correlaciones parciales son:

```
library(corpcor)
cor2pcor(cor(X))
```

```
##           [,1]      [,2]      [,3]      [,4]
## [1,] 1.0000000  0.9957943  0.90468746  0.31170534
## [2,] 0.9957943  1.0000000 -0.89188935 -0.31783572
## [3,] 0.9046875 -0.8918894  1.00000000  0.03389341
## [4,] 0.3117053 -0.3178357  0.03389341  1.00000000
```

# Adecuación de los datos al ACP

Crearemos una función que calcule el KMO:

```
kmo.test <- function(df){  
  cor.sq = cor(df)^2  
  cor.sumsq = (sum(cor.sq)-dim(cor.sq)[1])  
  pcor.sq = cor2pcor(cor(df))^2  
  pcor.sumsq = (sum(pcor.sq)-dim(pcor.sq)[1])  
  kmo = cor.sumsq/(cor.sumsq+pcor.sumsq)  
  return(kmo)  
}  
  
kmo.test(X)
```

```
## [1] 0.4225498
```

El test de esfericidad de Barlett contrasta si la matriz de correlaciones es la identidad.

## Representaciones gráficas. Biplots

# Biplots

- El análisis de componentes principales se utiliza para representar gráficamente la tabla de datos. Entre otras representaciones gráficas la denominada **biplot** es uno de los más útiles.
- Este gráfico presenta los componentes principales junto con la proyección de las variables originales *como si fueran vectores* que **expresan la dirección en la que aumentan las variables**.
- Para obtener esta **representación conjunta** las coordenadas de los componentes principales y las variables es **necesario modificarlas multiplicándolas por constante de escala**. Esta constante puede variar según el algoritmo que se utilice.

## Ejemplo

- En el mismo gráfico, se representan como vectores cada una de las variables iniciales  $\mathbf{X}$ . Las coordenadas de dichas variables son las correlaciones de dichas variables con las componentes principales divididas por la raíz cuadrada del valor propio correspondiente a la componente principal.
- Sea  $\tilde{\mathbf{X}}_i$  es una variable inicial. Calculemos la coordenadas de dicha variable inicial como vector de dos componentes representado en el biplot. Dichas coordenadas valen:

$$\frac{\text{Cor}(\tilde{\mathbf{X}}_i, \mathbf{CP}_j)}{\sqrt{\lambda_j}},$$

para  $j = 1, 2$  que son los ejes correspondientes a las dos componentes principales.

# Ejemplo

- Escrito en forma matricial, se calcula:

$$\text{Cor}(\tilde{\mathbf{X}}, \mathbf{CP}) \cdot \text{diag} \left( \frac{1}{\sqrt{\lambda}} \right).$$

- La expresión anterior, usando las propiedades de las componentes principales es equivalente a calcular

$$\mathbf{u} \cdot \text{diag} \left( \sqrt{\lambda} \right).$$

- Los vectores correspondientes a las variables iniciales se multiplican por un factor de escala dependiente de cada paquete estadístico que se use (en el caso de  $\{\mathbf{R}\}$ , dicho factor es 3).

## Ejemplo

- Vamos a realizar un biplot con los datos del ejemplo que hemos desarrollado en este capítulo. Se tendrá en cuenta el caso de ACP usando la matriz de covarianzas.
- Las coordenadas de las componentes principales de nuestros datos eran las siguientes:

$$\mathbf{CP} = \begin{pmatrix} -3.054 & 0.201 & -0.827 & 0.280 \\ -12.719 & 2.480 & -0.333 & 0.103 \\ -4.293 & -3.295 & -0.025 & -0.228 \\ 7.373 & -6.736 & -0.183 & 0.029 \\ -15.299 & 0.565 & 1.029 & 0.183 \\ -1.354 & 1.319 & 1.463 & -0.321 \\ -6.997 & 1.411 & -1.460 & -0.233 \\ 13.677 & 0.437 & 0.629 & 0.282 \\ 22.666 & 3.618 & -0.292 & -0.095 \end{pmatrix}.$$

# Ejemplo

- Sólo representaremos las dos primeras componentes principales que corresponden a las dos primeras columnas de la matriz anterior.
- Vamos a estandarizar los datos. Calculemos el módulo de cada componente principal:

$$|\mathbf{CP}_1| = \sqrt{(-3.054)^2 + \dots + (22.666)^2} = 35.065,$$

$$|\mathbf{CP}_2| = \sqrt{(0.201)^2 + \dots + (3.618)^2} = 8.930.$$

- Dividiendo cada una de las dos componentes principales por su módulo correspondiente, obtenemos las coordenadas del biplot de cada dato:

$$\begin{pmatrix} -3.054/35.065 & 0.201/8.903 \\ -12.719/35.065 & 2.480/8.903 \\ -4.293/35.065 & -3.295/8.903 \\ 7.373/35.065 & -6.736/8.903 \\ -15.299/35.065 & 0.565/8.903 \end{pmatrix} = \begin{pmatrix} -0.087 & 0.023 \\ -0.363 & 0.278 \\ -0.122 & -0.369 \\ 0.210 & -0.754 \\ -0.436 & 0.063 \end{pmatrix}.$$



## Ejemplo

- Ahora, vamos a encontrar las componentes de los vectores que representan las variables. Recordemos que la matriz de vectores propios de la matriz de covarianzas era:

$$\mathbf{u} = \begin{pmatrix} 0.934 & -0.022 & 0.256 & 0.247 \\ 0.339 & 0.354 & -0.661 & -0.568 \\ 0.047 & -0.248 & 0.566 & -0.785 \\ 0.097 & -0.902 & -0.421 & -0.013 \end{pmatrix},$$

cuyos valores propios asociados eran  
136.615, 8.861, 0.738, 0.0468.

## Ejemplo

- Para hallar las coordenadas de los vectores que representarán las variables, hemos de multiplicar cada vector propio por la raíz cuadrada de su valor propio correspondiente y por 3:

$$\begin{aligned} & \mathbf{u} \cdot 3 \cdot \begin{pmatrix} \sqrt{136.615} & 0 & 0 & 0 \\ 0 & \sqrt{8.861} & 0 & 0 \\ 0 & 0 & \sqrt{0.738} & 0 \\ 0 & 0 & 0 & \sqrt{0.0468} \end{pmatrix} \\ &= \begin{pmatrix} 32.764 & -0.200 & 0.659 & 0.160 \\ 11.904 & 3.163 & -1.704 & -0.368 \\ 1.648 & -2.212 & 1.458 & -0.509 \\ 3.409 & -8.051 & -1.086 & -0.009 \end{pmatrix}. \end{aligned}$$

## Ejemplo biplot

Las componentes de los vectores que representarán las variables serán:

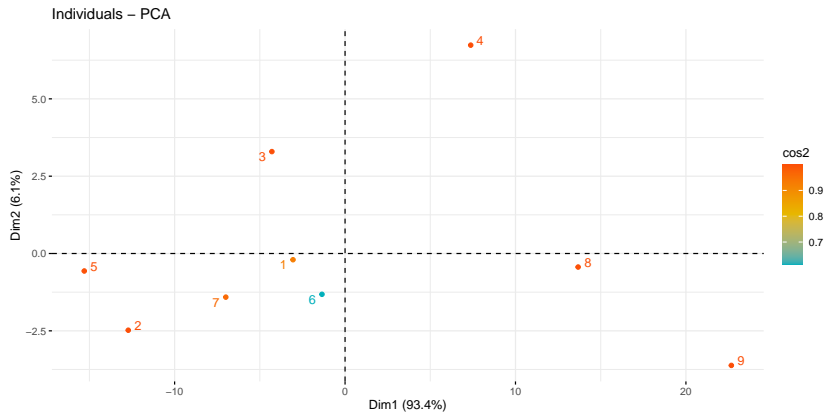
$$\begin{aligned} &CP_1(32.764, -0.200), CP_2(11.904, 3.163), \\ &CP_3(1.648, -2.212), CP_4(3.409, -8.051). \end{aligned}$$

El código `biplot(solacp)` dibuja el biplot de los dos primeros componentes.

## Ejemplo biplot

```
factoextra::fviz_pca_ind(solapc,
  col.ind = "cos2",
  # Color por calidad de
  # la representación
  gradient.cols =
    c("#00AFBB",
      "#E7B800", "#FC4E07"),
  repel = TRUE
  # permite solapar texto
)
```

# Ejemplo biplot



# Interpretación de un biplot

- La representación de las observaciones o los datos en un biplot equivale a proyectar las observaciones sobre el plano de las componentes principales estandarizadas para que tengan varianza unidad.
- La representación de variables mediante vectores de dos coordenadas cumple que la correlación entre dos variables iniciales  $\mathbf{X}_i$  y  $\mathbf{X}_j$  es aproximadamente el coseno del ángulo que forman en el biplot. Por tanto, si dos variables  $\mathbf{X}_i$  y  $\mathbf{X}_j$  están muy correlacionadas, el coseno será grande y el ángulo entre los vectores, pequeño. En caso contrario, si están poco correlacionadas, el coseno será pequeño y el ángulo entre los vectores estará próximo a un ángulo recto.

# Comunalidades.

En un ACP la comunalidad de la variable  $X_j$  retenida por las  $k$  primeras componentes es la proporción de varianza de la variable que queda explicada por esas componentes. Por ejemplo.

- Si retenemos sólo el componente  $CP_1$  la comunalidad de la variable  $X_j$  es:

$$h_j = r_{j1}^2 = \left(u_{1j}\sqrt{\lambda_1}\right)^2$$

\* Si retenemos los componentes  $CP_1$  y  $CP_2$  la comunalidad de la variable  $X_j$  es:

$$h_j = r_{j1}^2 + r_{j2}^2 = \left(u_{1j}\sqrt{\lambda_1}\right)^2 + \left(u_{2j}\sqrt{\lambda_2}\right)^2$$

## Interpretación de las variables y los individuos



# Interpretación de las variables y los individuos

- Las variables también pueden representar de forma simultanea con los individuos en los componentes principales.
- Esta representación se hace mediante las coordenadas que la matriz de componentes que nos explican las correlaciones de cada factor con cada variable.

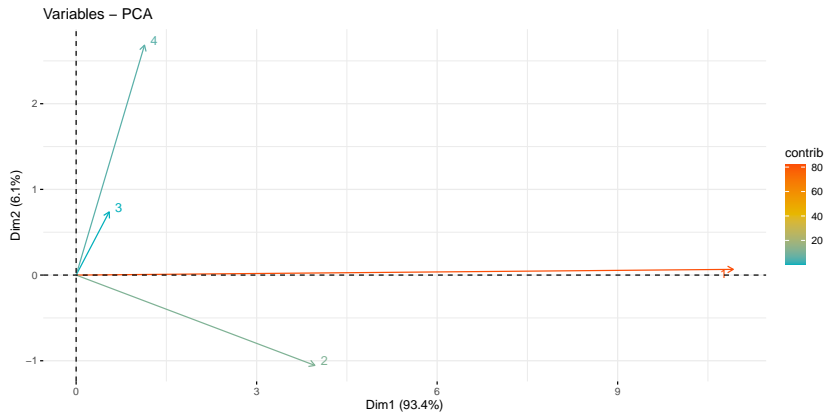
# Círculo de correlación

- Cada variable está representada por el vector que une el origen de coordenadas con el punto.
- Todos están en círculo unidad (círculo de correlación).
- A medida que cada variable se acerca a la circunferencia unidad está mejor representado por los componentes retenidas y viceversa.
- El ángulo entre variables y componentes nos da una idea de su correlación, al nivel de retención de varianza total que tengamos.
- Así variable perpendiculares tenderán a ser *incorreladas*.
- Los valores de una variable crecen en la dirección de ésta.

# Circulo de correlación

```
factoextra::fviz_pca_var(solapc,  
  col.var = "contrib",  
  # Color por contribución a la pca  
  gradient.cols = c("#00AFBB",  
                    "#E7B800",  
                    "#FC4E07"),  
  repel = TRUE  
)
```

# Circulo de correlación



Y muchas cosas más..

# Y muchas cosas más..

Para acabar... **Análisis Factorial Confirmatorio y Exploratorio**

- El Análisis factorial confirmatorio se realiza sobre modelos establecidos de factores y se hacen inferencias sobre sus propiedades.
- El análisis factorial descriptivo ayuda a la descripción de los datos y a la búsqueda de factores.

**Relación del ACP con otras técnicas de análisis de datos**

- Regresión Lineal Múltiple
- Clasificación.
- Análisis de correspondencias simples y múltiples.
- ... y muchas otras más