# CEDEX - Curso de formación estadística

## Técnicas de análisis multivariante

Andrés M. Alonso

Departamento de Estadística
Universidad Carlos III de Madrid

Madrid - 18 de octubre de 2007

# Estructura del curso

1. Técnicas de análisis multivariante - I.
   - Introducción.
   - Técnicas descriptivas numéricas.
   - Técnicas descriptivas gráficas.
   - Análisis de componentes principales.

2. Técnicas de análisis multivariante - II.
   - Análisis factorial.
   - Escalado multidimensional.
   - Análisis de correspondencias.

# Estructura del curso

4. Técnicas de análisis multivariante - III.

   - Problemas de clasificación.
   - Análisis discriminante lineal.
   - Análisis discriminante cuadrático.
   - Análisis discriminante logístico.

5. Técnicas de análisis multivariante - IV.

   - Análisis cluster jerárquico.
   - Análisis cluster por partición.
   - Análisis de correlaciones canónicas.

6. Técnicas de análisis multivariante - V.

   - Sesión práctica con SPSS.

# Introducción

- Variables
- Observaciones.
- Matriz de datos.
- Ejemplos.

Variable (vectorial o multivariante): es un conjunto de características o rasgos de los elementos de una población. Notación: $\mathbf{x}$.

Observación o dato: valor de una variable multivariante en un elemento de la muestra. Notación: $\mathbf{x}_i$ corresponde al elemento $i$.

Matriz de datos: representación de los valores de una muestra de tamaño $n$ de una variable vectorial $\mathbf{x}$.

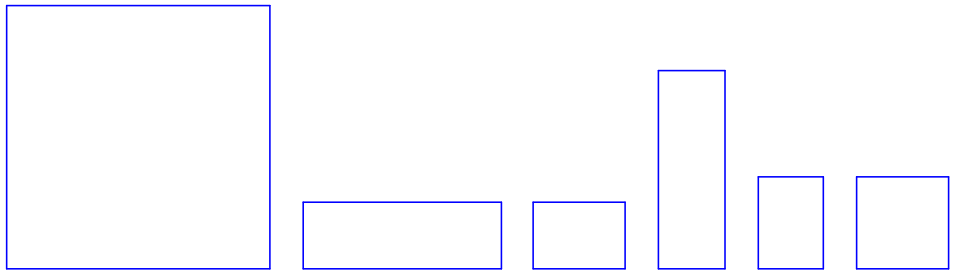
$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix} = \left[\mathbf{x}_{(1)}\mathbf{x}_{(2)} \cdots \mathbf{x}_{(p)}\right],$$

donde:
- $x_{ij}$ es el valor de la variable escalar $j$ en el individuo $i$.
- $\mathbf{x}_i'$ es un vector fila $1 \times p$ que representa los valores de las $p$ variables univariantes en el individuo $i$.
- $\mathbf{x}_{(j)}$ es un vector columna $n \times 1$ que representa los valores de la variable escalar $j$ en las $n$ observaciones.

## Ejemplo 0. Rectángulos.

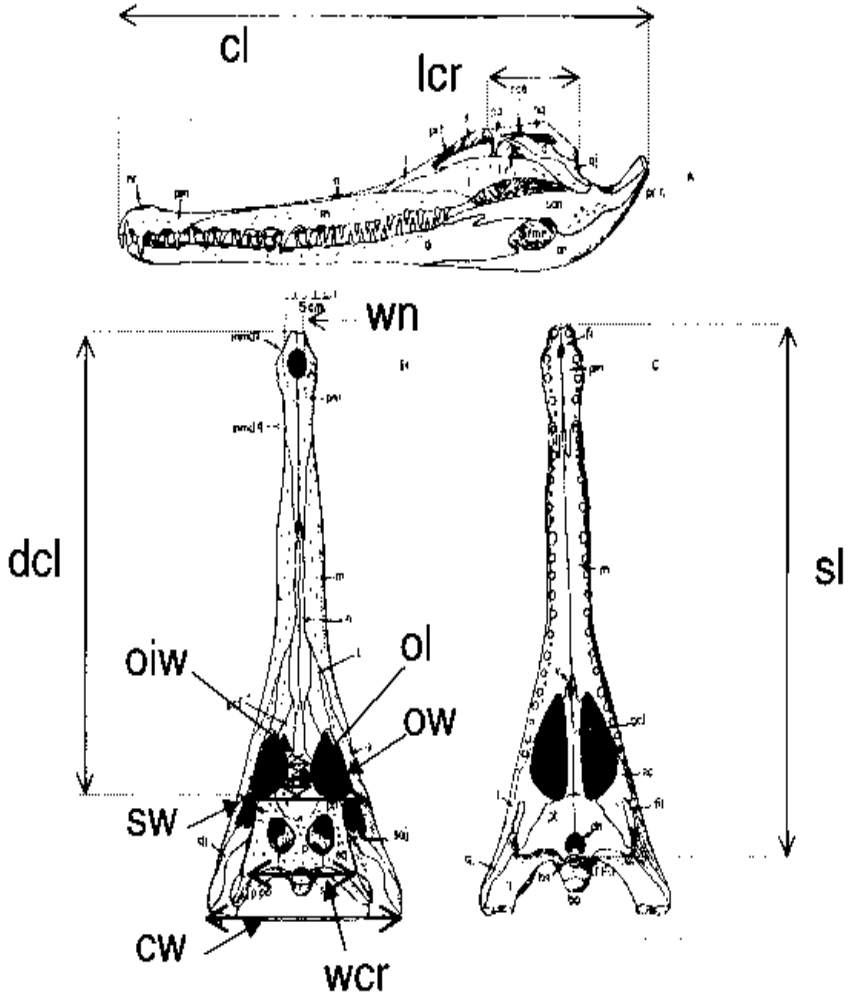
Ejemplo 5.9 del libro *Análisis de Datos Multivariantes* de Daniel Peña. Se tienen 6 observaciones bivariantes, cada observación corresponde con un rectángulo y las variables univariantes son la longitud de la base y la altura del rectángulo. La matriz de datos es:

$$\mathbf{X} = \begin{bmatrix} 2{,}0 & 2{,}0 \\ 1{,}5 & 0{,}5 \\ 0{,}7 & 0{,}5 \\ 0{,}5 & 1{,}5 \\ 0{,}5 & 0{,}7 \\ 0{,}7 & 0{,}7 \end{bmatrix}.$$



# Ejemplo 1. Medidas de cráneos de cocodrilos.

| Código | Descripción |
|---|---|
| cl | Longitud del cráneo |
| cw | Ancho del cráneo |
| sw | Ancho del hocico |
| sl | Longitud del hocico |
| dcl | Longitud dorsal del cráneo |
| ow | Ancho máximo orbital |
| oiw | Ancho mínimo inter–orbital |
| ol | Longitud máxima orbital |
| lcr | Longitud del paladar post–orbital |
| wcr | Ancho posterior del paladar craneal |
| wn | Ancho máximo entre orificios nasales |



Ejemplo 2. Medidas o características de automóviles.

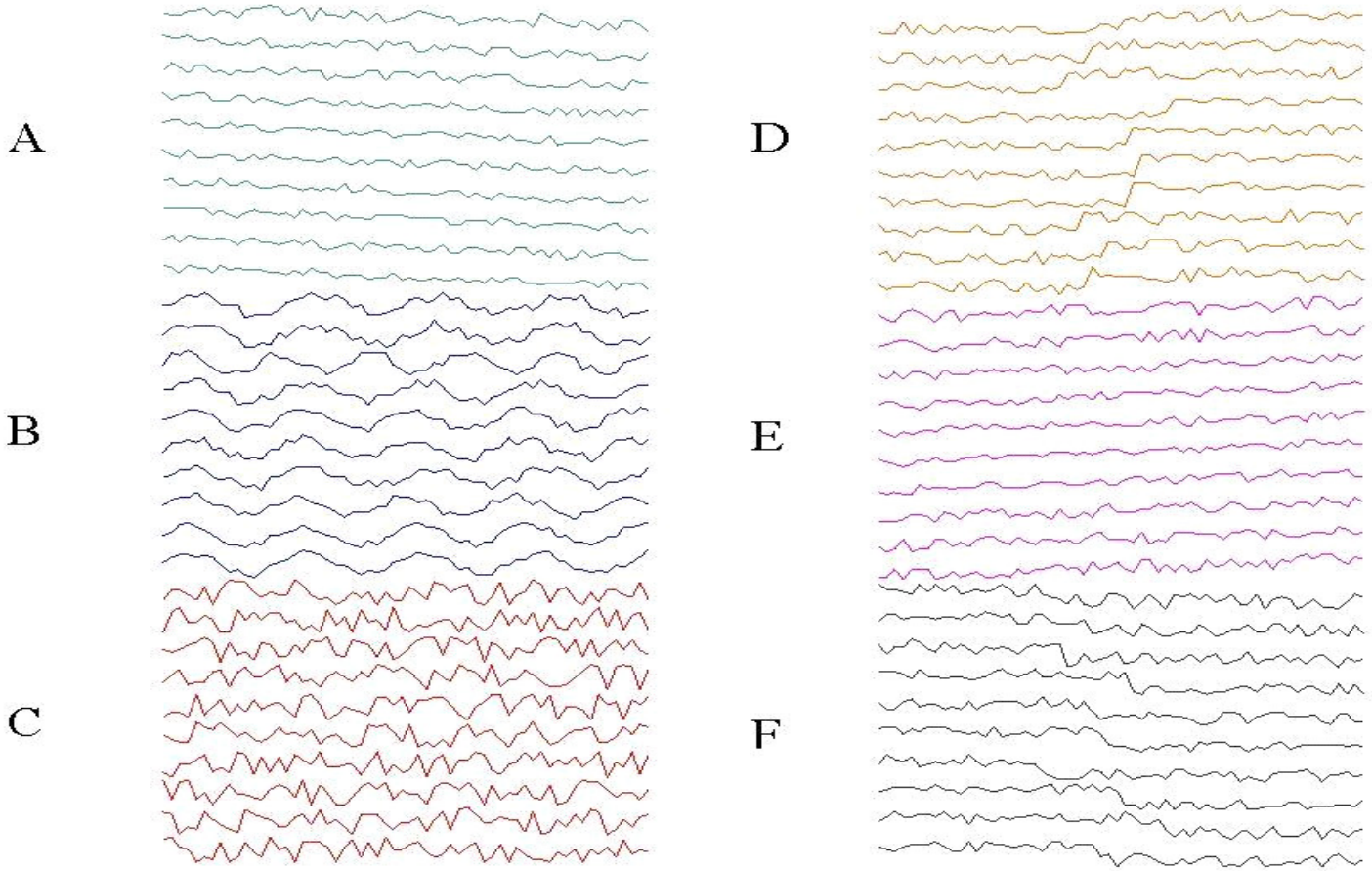
| Código | Descripción |
|---|---|
| consumo | Consumo (l/100Km) |
| motor | Cilindrada en cc |
| cv | Potencia (CV) |
| peso | Peso total (kg) |
| acel | Aceleración 0 a 100 km/h (segundos) |
| año | Año del modelo |
| origen | País de origen |
| cilindr | Número de cilindros |

## Ejemplo 3. Gases contaminantes

En la Tabla siguiente se presentan las 10 primeras observaciones de cinco variables de niveles de gases contaminantes (CO: $X_3$, NO: $X_4$, $NO_2$: $X_5$, $O_3$: $X_6$, y HC: $X_7$) y dos variables relacionadas (Intensidad del viento: $X_1$, y Radiación solar: $X_2$).

| $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|---|---|
| 8 | 98 | 7 | 2 | 12 | 8 | 2 |
| 7 | 107 | 4 | 3 | 9 | 5 | 3 |
| 7 | 103 | 4 | 3 | 5 | 6 | 3 |
| 10 | 88 | 5 | 2 | 8 | 15 | 4 |
| 6 | 91 | 4 | 2 | 8 | 10 | 3 |
| 8 | 90 | 5 | 2 | 12 | 12 | 4 |
| 9 | 84 | 7 | 4 | 12 | 15 | 5 |
| 5 | 72 | 6 | 4 | 21 | 14 | 4 |
| 7 | 82 | 5 | 1 | 11 | 11 | 3 |
| 8 | 64 | 5 | 2 | 13 | 9 | 4 |

Ejemplo 4. Gráficos de control de un proceso industrial.




$\boldsymbol{X}_{n\times 60}$: 60 Mediciones del proceso en $n$ máquinas.

## Ejemplo 5. Esclerosis múltiple.

En un estudio sobre esclerosis múltiple se registran las respuestas del ojo izquierdo $(I)$ y del ojo derecho $(D)$ a dos estímulos visuales diferentes. Se consideran dos grupos, 29 individuos que padecen esclerosis múltiple y un grupo control de 69 individuos que no la padecen. Se registran las siguientes variables: $X_1$: Edad, $X_2 = R1L+R1D$, $X_3 = |R1L-R1D|$, $X_4 = R2L+R2D$, $X_5 = |R2L - R2D|$.

| $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | Paciente/Control |
|---|---|---|---|---|---|
| 23 | 148.0 | 0.8 | 205.4 | 0.6 | 1 |
| 25 | 195.2 | 3.2 | 262.8 | 0.4 | 1 |
| 25 | 158.0 | 8.0 | 209.8 | 12.2 | 1 |
| 28 | 134.4 | 0.0 | 198.4 | 3.2 | 1 |
| 29 | 190.2 | 14.2 | 243.8 | 10.6 | 1 |
| 18 | 152.0 | 1.6 | 198.4 | 0.0 | 0 |
| 19 | 138.0 | 0.4 | 180.8 | 1.6 | 0 |
| 20 | 144.0 | 0.0 | 186.4 | 0.8 | 0 |
| 20 | 143.6 | 3.2 | 194.8 | 0.0 | 0 |
| 20 | 148.8 | 0.0 | 217.6 | 0.0 | 0 |

# Técnicas descriptivas numéricas

- Estadísticos univariantes y bivariantes.
- Vector de medias y matriz de covarianzas.
- Proyecciones y combinaciones lineales.
- Estandarización univariante.
- Estandarización multivariante.

# Estadísticos univariantes y bivariantes

Media muestral de la variable $x_j$:

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}.$$

Varianza muestral de la variable $x_j$:

$$s_j^2 = s_{jj} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2.$$

Covarianza muestral entre las variables $x_j$ y $x_k$:

$$s_{jk} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k).$$

Los estadísticos anteriores dependen de las unidades de medidas y por esto suelen utilizarse, como complemento en el resumen numérico, los siguientes estadísticos:

Coeficiente de variación de la variable $x_j$:

$$CV_j = \sqrt{\frac{s_j^2}{\bar{x}_j^2}},$$

que podrá calcularse siempre que $\bar{x}_j$ sea distinta de cero.

Correlación muestral entre las variables $x_j$ y $x_k$:

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}s_{kk}}} = \frac{s_{jk}}{s_j s_k}.$$

# Estadísticos multivariantes - I

Vector de medias muestral de la variable vectorial $\mathbf{x}$:

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}.$$

$\bar{\mathbf{x}}$ es un vector de dimensión $p \times 1$. También podemos obtener el vector de medias de la siguiente expresión:

$$\bar{\mathbf{x}} = \frac{1}{n}\mathbf{X}'\mathbf{1},$$

donde $\mathbf{1}$ es un vector de unos de dimensión $n \times 1$.

# Estadísticos multivariantes - II

Matriz de varianzas y covarianzas de la variable vectorial $\mathbf{x}$:

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}.$$

$\mathbf{S}$ es una matriz cuadrada simétrica ($s_{jk} = s_{kj}$) de dimensión $p \times p$. También podemos obtener la matriz de varianzas y covarianzas de las siguientes expresiones:

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' = \frac{1}{n}(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')'(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}') = \frac{1}{n}\tilde{\mathbf{X}}'\tilde{\mathbf{X}},$$

donde la matriz $\tilde{\mathbf{X}} = \mathbf{X} - \mathbf{1}\bar{\mathbf{x}}' = \mathbf{X} - \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{X}$ recibe el nombre de matriz de datos centrados.

# Estadísticos multivariantes - Ejemplo - I

**Ejemplo 0.** De las siguientes salidas de SPSS podemos obtener el vector de medias y las matrices de covarianzas y de correlaciones del conjunto de datos de rectángulos:

**Estadísticos descriptivos**

| | N | Media | Desv. típ. | Varianza |
|---|---|---|---|---|
| BASE | 6 | ,9833 | ,62102 | ,386 |
| ALTURA | 6 | ,9833 | ,62102 | ,386 |
| N válido (según lista) | 6 | | | |

Vector de medias:

$$\bar{\mathbf{x}} = \begin{bmatrix} 0{,}9833 \\ 0{,}9833 \end{bmatrix}.$$

# Estadísticos multivariantes - Ejemplo - II

## Ejemplo 0.

**Correlaciones**

| | | BASE | ALTURA |
|---|---|---|---|
| BASE | Correlación de Pearson | 1 | ,461 |
| | Covarianza | ,386 | ,178 |
| | N | 6 | 6 |
| ALTURA | Correlación de Pearson | ,461 | 1 |
| | Covarianza | ,178 | ,386 |
| | N | 6 | 6 |

Matriz de covarianzas: $\mathbf{S} = \begin{bmatrix} 0{,}386 & 0{,}178 \\ 0{,}178 & 0{,}386 \end{bmatrix}$.

Matriz de correlaciones: $\mathbf{R} = \begin{bmatrix} 1{,}000 & 0{,}461 \\ 0{,}461 & 1{,}000 \end{bmatrix}$.

# Estadísticos multivariantes - Ejemplo - III

**Estadísticos descriptivos**

| | N | Mínimo | Máximo | Media | Desv. típ. |
|---|---|---|---|---|---|
| Consumo (l/100Km) | 398 | 5 | 26 | 11,23 | 3,946 |
| Cilindrada en cc | 406 | 66 | 7456 | 3179,73 | 1724,013 |
| Potencia (CV) | 400 | 46 | 230 | 104,83 | 38,522 |
| Peso total (kg) | 406 | 244 | 1713 | 989,51 | 283,277 |
| Aceleración 0 a 100 km/h (segundos) | 406 | 8 | 25 | 15,50 | 2,821 |
| Año del modelo | 406 | 70 | 82 | 75,92 | 3,749 |
| País de origen | 405 | 1 | 3 | 1,57 | ,798 |
| Número de cilindros | 405 | 3 | 8 | 5,47 | 1,710 |
| N válido (según lista) | 391 | | | | |

- La media y la varianza no tienen sentido en la variable "Pais de origen".
- El vector de medias es:

$$\bar{\mathbf{x}} = \left[ \begin{array}{ccccccc} 11,23 & 3179,73 & 104,83 & 989,51 & 15,50 & 75,92 & 5,47 \end{array} \right]'.$$

# Estadísticos multivariantes - Ejemplo - IV

**Correlaciones**

| | | Consumo (l/100Km) | Cilindrada en cc | Potencia (CV) | Peso total (kg) | Aceleración 0 a 100 km/h (segundos) | Año del modelo | Número de cilindros |
|---|---|---|---|---|---|---|---|---|
| Consumo (l/100Km) | Correlación de Pearson | 1 | ,837** | ,836** | ,837** | -,490** | -,554** | ,842** |
| | Sig. (bilateral) | . | ,000 | ,000 | ,000 | ,000 | ,000 | ,000 |
| | N | 398 | 398 | 392 | 398 | 398 | 398 | 397 |
| Cilindrada en cc | Correlación de Pearson | ,837** | 1 | ,897** | ,933** | -,545** | -,370** | ,952** |
| | Sig. (bilateral) | ,000 | . | ,000 | ,000 | ,000 | ,000 | ,000 |
| | N | 398 | 406 | 400 | 406 | 406 | 406 | 405 |
| Potencia (CV) | Correlación de Pearson | ,836** | ,897** | 1 | ,859** | -,701** | -,417** | ,844** |
| | Sig. (bilateral) | ,000 | ,000 | . | ,000 | ,000 | ,000 | ,000 |
| | N | 392 | 400 | 400 | 400 | 400 | 400 | 399 |
| Peso total (kg) | Correlación de Pearson | ,837** | ,933** | ,859** | 1 | -,415** | -,296** | ,895** |
| | Sig. (bilateral) | ,000 | ,000 | ,000 | . | ,000 | ,000 | ,000 |
| | N | 398 | 406 | 400 | 406 | 406 | 406 | 405 |
| Aceleración 0 a 100 km/h (segundos) | Correlación de Pearson | -,490** | -,545** | -,701** | -,415** | 1 | ,314** | -,528** |
| | Sig. (bilateral) | ,000 | ,000 | ,000 | ,000 | . | ,000 | ,000 |
| | N | 398 | 406 | 400 | 406 | 406 | 406 | 405 |
| Año del modelo | Correlación de Pearson | -,554** | -,370** | -,417** | -,296** | ,314** | 1 | -,357** |
| | Sig. (bilateral) | ,000 | ,000 | ,000 | ,000 | ,000 | . | ,000 |
| | N | 398 | 406 | 400 | 406 | 406 | 406 | 405 |
| Número de cilindros | Correlación de Pearson | ,842** | ,952** | ,844** | ,895** | -,528** | -,357** | 1 |
| | Sig. (bilateral) | ,000 | ,000 | ,000 | ,000 | ,000 | ,000 | . |
| | N | 397 | 405 | 399 | 405 | 405 | 405 | 405 |

**. La correlación es significativa al nivel 0,01 (bilateral).

# Estadísticos multivariantes - Ejemplo - V

**Correlaciones**

| | | VIENTO | RADIACIO | CO | NO | NO2 | O3 | HC |
|---|---|---|---|---|---|---|---|---|
| VIENTO | Correlación de Pearson | 1 | -,101 | -,194 | -,270 | -,110 | -,254 | ,156 |
| | Sig. (bilateral) | . | ,523 | ,219 | ,084 | ,489 | ,105 | ,324 |
| | N | 42 | 42 | 42 | 42 | 42 | 42 | 42 |
| RADIACIO | Correlación de Pearson | -,101 | 1 | ,183 | -,074 | ,116 | ,319* | ,052 |
| | Sig. (bilateral) | ,523 | . | ,247 | ,643 | ,465 | ,039 | ,744 |
| | N | 42 | 42 | 42 | 42 | 42 | 42 | 42 |
| CO | Correlación de Pearson | -,194 | ,183 | 1 | ,502** | ,557** | ,411** | ,166 |
| | Sig. (bilateral) | ,219 | ,247 | . | ,001 | ,000 | ,007 | ,293 |
| | N | 42 | 42 | 42 | 42 | 42 | 42 | 42 |
| NO | Correlación de Pearson | -,270 | -,074 | ,502** | 1 | ,297 | -,134 | ,235 |
| | Sig. (bilateral) | ,084 | ,643 | ,001 | . | ,056 | ,398 | ,135 |
| | N | 42 | 42 | 42 | 42 | 42 | 42 | 42 |
| NO2 | Correlación de Pearson | -,110 | ,116 | ,557** | ,297 | 1 | ,167 | ,448** |
| | Sig. (bilateral) | ,489 | ,465 | ,000 | ,056 | . | ,292 | ,003 |
| | N | 42 | 42 | 42 | 42 | 42 | 42 | 42 |
| O3 | Correlación de Pearson | -,254 | ,319* | ,411** | -,134 | ,167 | 1 | ,154 |
| | Sig. (bilateral) | ,105 | ,039 | ,007 | ,398 | ,292 | . | ,329 |
| | N | 42 | 42 | 42 | 42 | 42 | 42 | 42 |
| HC | Correlación de Pearson | ,156 | ,052 | ,166 | ,235 | ,448** | ,154 | 1 |
| | Sig. (bilateral) | ,324 | ,744 | ,293 | ,135 | ,003 | ,329 | . |
| | N | 42 | 42 | 42 | 42 | 42 | 42 | 42 |

*. La correlación es significante al nivel 0,05 (bilateral).

**. La correlación es significativa al nivel 0,01 (bilateral).

# Proyecciones y combinaciones lineales

Una forma simple de resumir una variable vectorial, $\mathbf{x}$, es construir una variable univariante, $y$, que sea el resultado de una combinación lineal de las componentes de $\mathbf{x}$:

$$y = \mathbf{a}'\mathbf{x},$$

donde $\mathbf{a}$ es un vector de constantes de dimensión $p \times 1$.

Si obtenemos las combinaciones lineales de todos los datos tendremos un vector $\mathbf{y}$ de dimensión $n \times 1$. $\mathbf{y}$ puede obtenerse de la siguiente expresión:

$$\mathbf{y} = \mathbf{X}\mathbf{a},$$

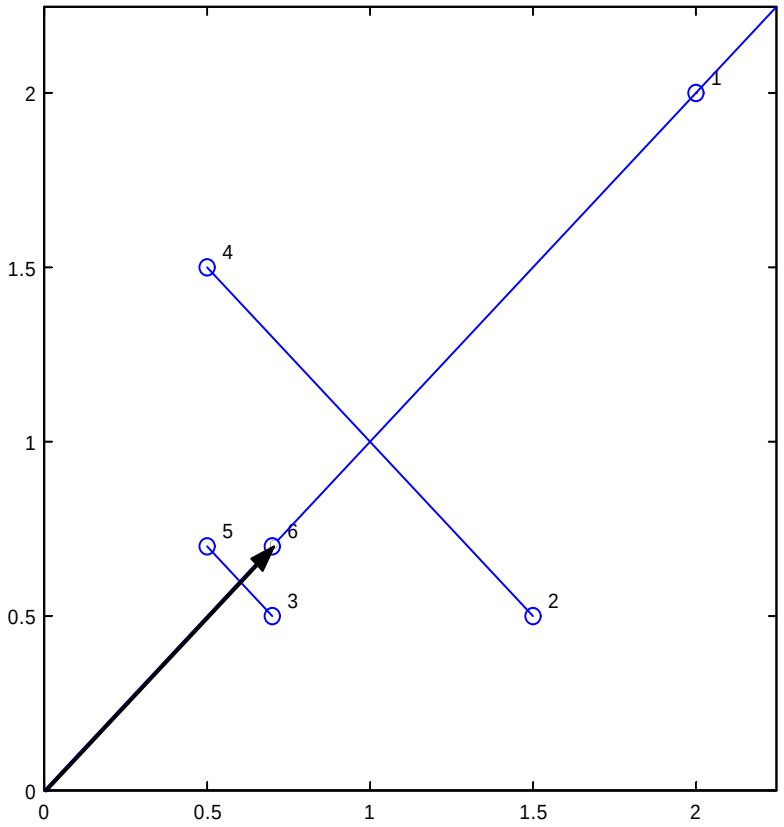
donde $\mathbf{X}$ es la matriz de datos de dimensión $n \times p$.

# Ejemplo de rectángulos

**Ejemplo 0.** En el ejemplo de los rectángulos, una variable de interés es el perímetro del rectángulo, 2(base + altura), que podemos obtener mediante:

$$\mathbf{y} = \mathbf{X}\mathbf{a} = \begin{bmatrix} 2{,}0 & 2{,}0 \\ 1{,}5 & 0{,}5 \\ 0{,}7 & 0{,}5 \\ 0{,}5 & 1{,}5 \\ 0{,}5 & 0{,}7 \\ 0{,}7 & 0{,}7 \end{bmatrix} \begin{bmatrix} 2{,}0 \\ 2{,}0 \end{bmatrix}$$

$$= \begin{bmatrix} 8{,}00 \\ 4{,}00 \\ 2{,}40 \\ 4{,}00 \\ 2{,}40 \\ 2{,}80 \end{bmatrix}$$



# Estandarización univariante

Estandarización univariante:

$$\mathbf{y} = \mathbf{D}^{-1/2}(\mathbf{x} - \bar{\mathbf{x}}),$$

donde $\mathbf{D}^{-1/2}$ es una matriz diagonal de dimensión $p \times p$ con la siguiente expresión:

$$\mathbf{D}^{-1/2} = \begin{bmatrix} s_1^{-1} & 0 & \cdots & 0 \\ 0 & s_2^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_p^{-1} \end{bmatrix}.$$

Propiedades:

- La media de $\mathbf{y}$ es cero, i.e., $\bar{\mathbf{y}} = \mathbf{0}$.
- La matriz de covarianzas de $\mathbf{y}$ es la matriz de correlaciones de $\mathbf{x}$, i.e., $\mathbf{S}_\mathbf{y} = \mathbf{R}_\mathbf{x}$.

# Estandarización multivariante

Estandarización multivariante: Si $\mathbf{S_x}$ es la matriz de covarianzas de $\mathbf{x}$ podemos definir su raíz cuadrada, $\mathbf{S_x^{1/2}}$, por la siguiente condición:

$$\mathbf{S_x} = \mathbf{S_x^{1/2}}(\mathbf{S_x^{1/2}})'.$$

Esto nos permitirá definir la estandarización multivariante mediante la expresión:

$$\mathbf{y} = \mathbf{S_x^{-1/2}}(\mathbf{x} - \bar{\mathbf{x}}).$$

Propiedades:

- La media de $\mathbf{y}$ es cero, i.e., $\bar{\mathbf{y}} = \mathbf{0}$.
- La matriz de covarianzas de $\mathbf{y}$ es la matriz identidad de dimensión $p \times p$, i.e., $\mathbf{S_y} = \mathbf{I}$.

# Técnicas descriptivas gráficas

- Objetivos.

- Ejemplos de representación gráfica de los datos:

  - Matriz de diagramas de dispersión.
  - Diagramas de estrellas.
  - Diagramas de caras de Chernoff.
  - Diagramas de Andrews.

# Representación gráfica de datos

El objetivo que perseguimos con la representación gráfica de datos es identificar:

- Relaciones (¿débil/fuerte ó lineal/no lineal?).
- Grupos (¿los grupos o conglomerados observados corresponden a grupos o categorías conocidas?)
- Atípicos.

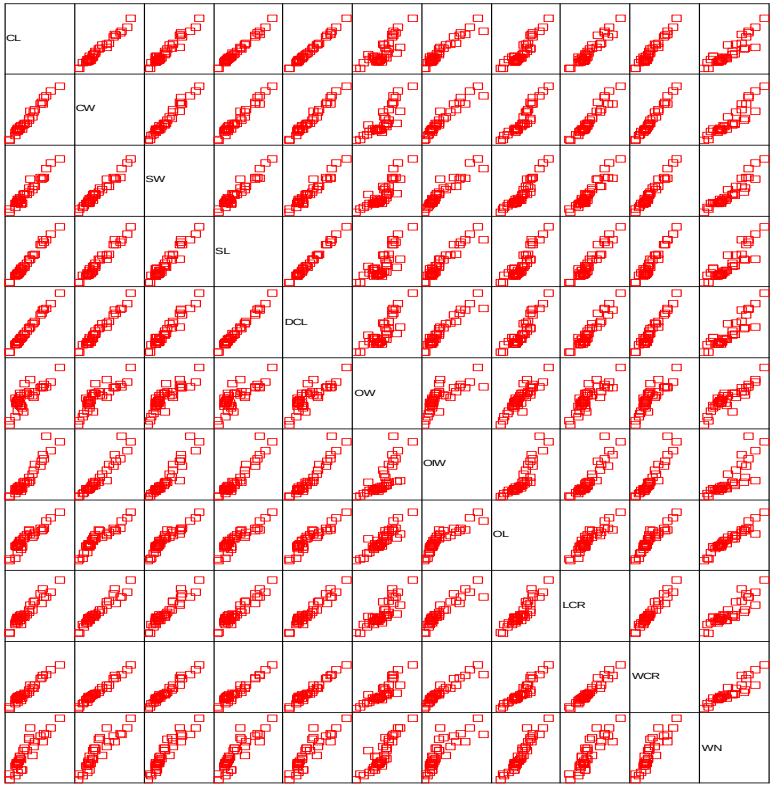
Estudiaremos los siguientes gráficos:

- Matriz de diagramas de dispersión.
- Diagramas de estrellas.
- Diagramas de caras.
- Diagramas de Andrews.

# Matriz de diagramas de dispersión - I

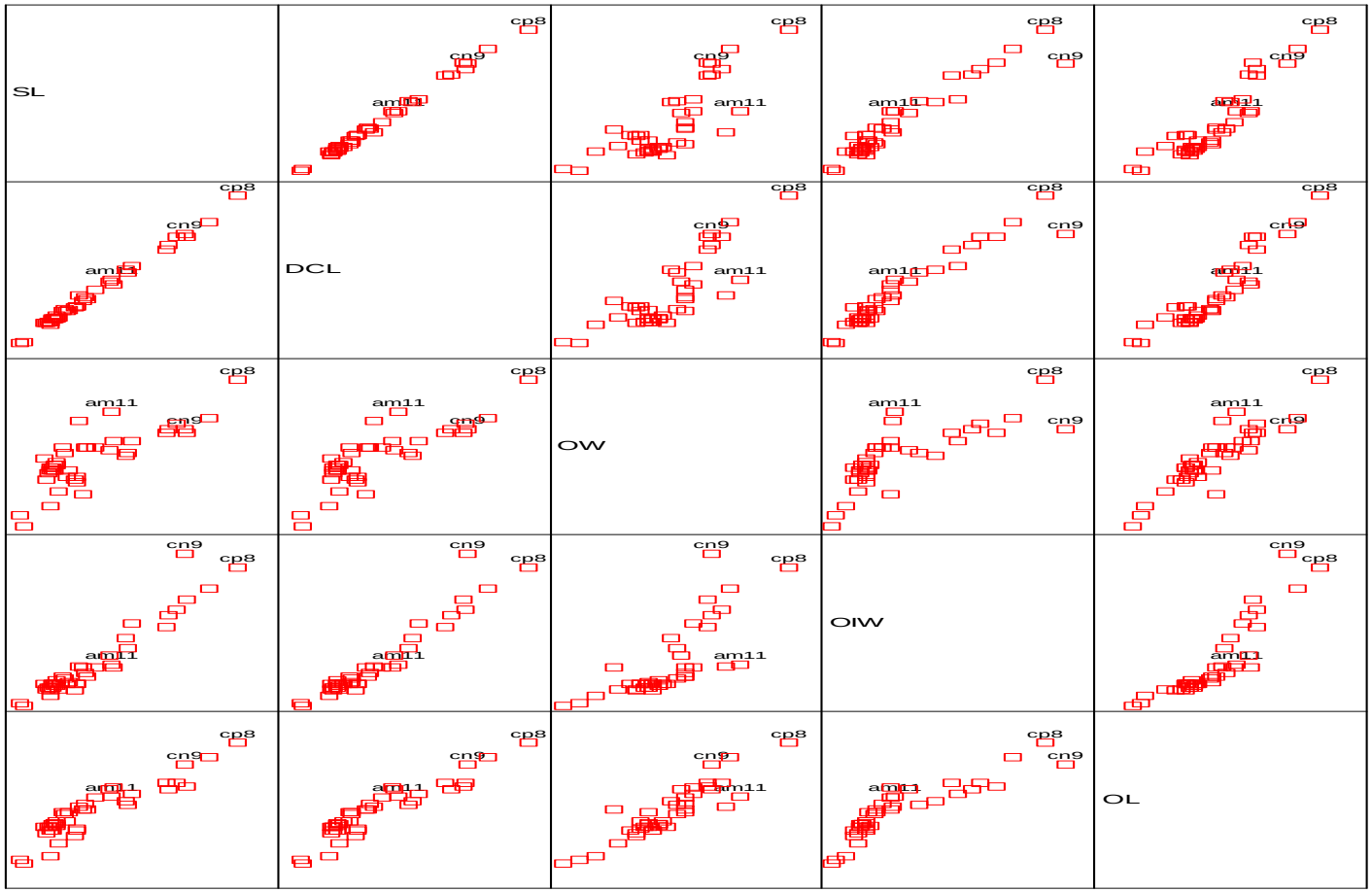
Si tenemos $p$ variables podemos construir $p(p-1)/2$ diagramas de dispersión diferentes tomando las variables por pares. Una manera de presentar estos gráficos es en forma de matriz.

**Ejemplo 1.** La Figura muestra la matriz de diagramas de dispersión en la que observamos, por ejemplo: (*i*) relaciones lineales entre la mayor parte de las variables, (*ii*) posible relación no lineal entre las variables `oiw` y `ow`, y entre `oiw` y `wn`, (*iii*) posibles atípicos en la variable `ow`.



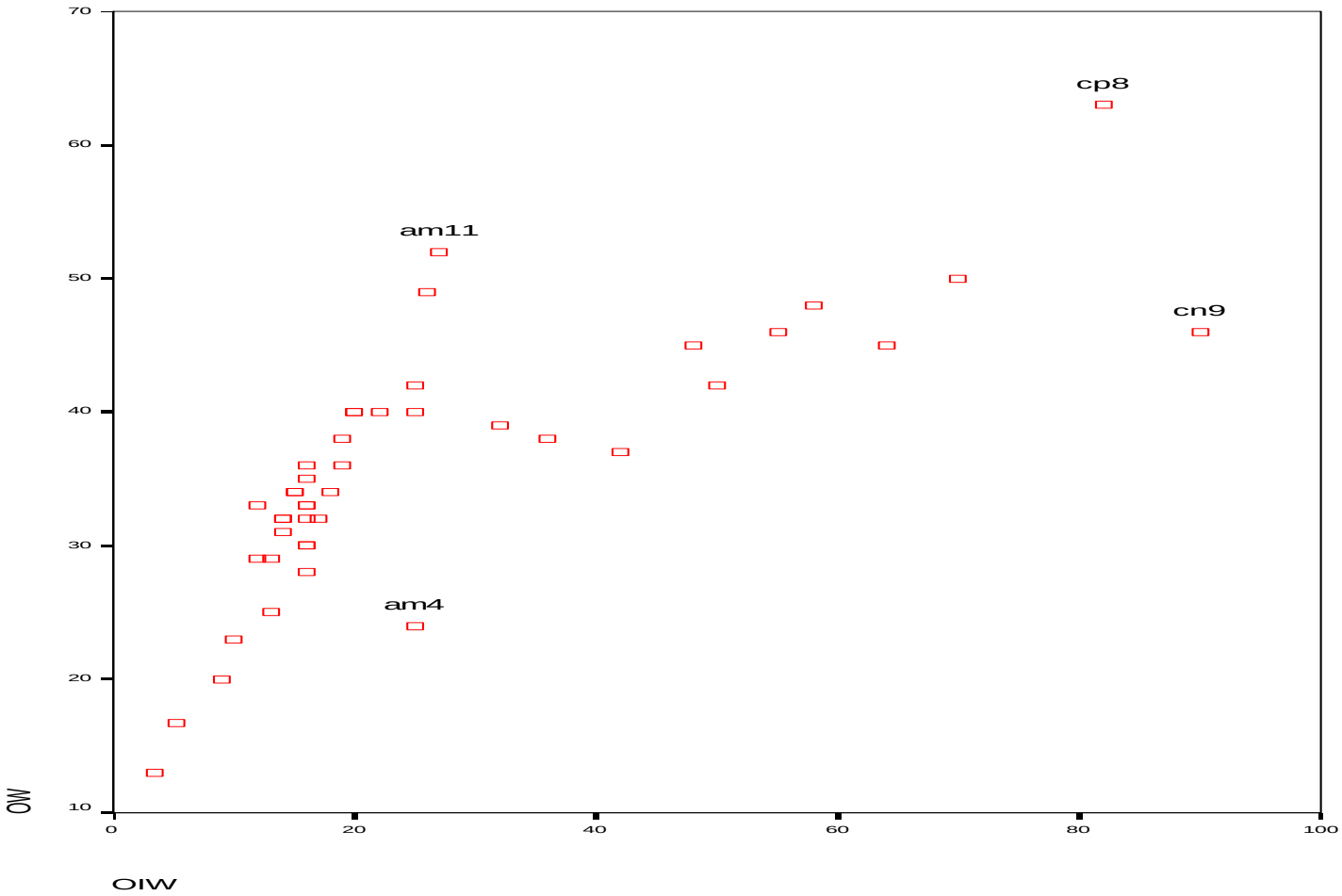
# Matriz de diagramas de dispersión - II

**Ejemplo 1.** (Zoom x2)



# Matriz de diagramas de dispersión - III

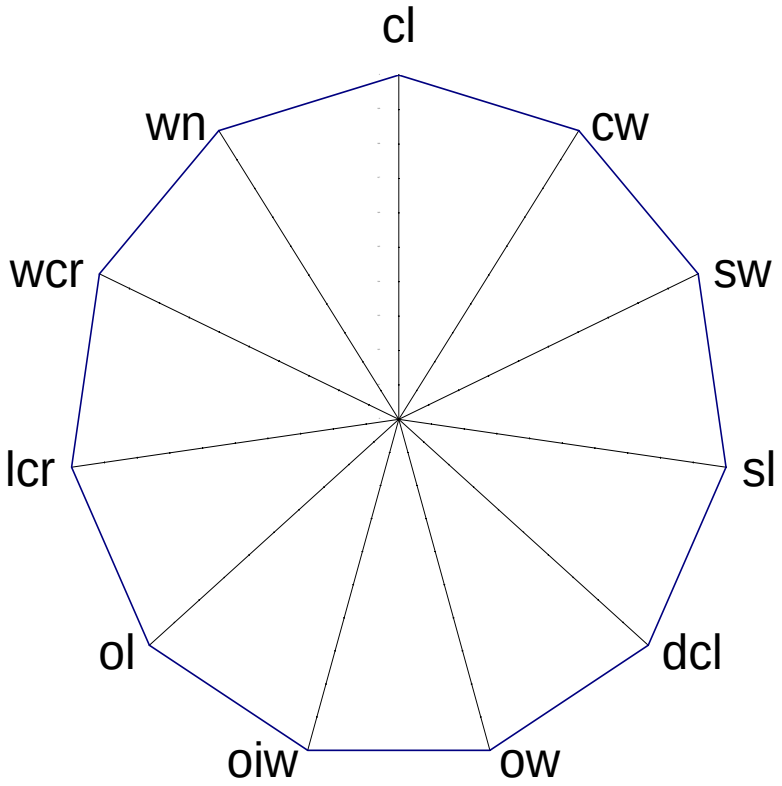
**Ejemplo 1.** (Zoom x8)



# Diagramas de estrellas - I

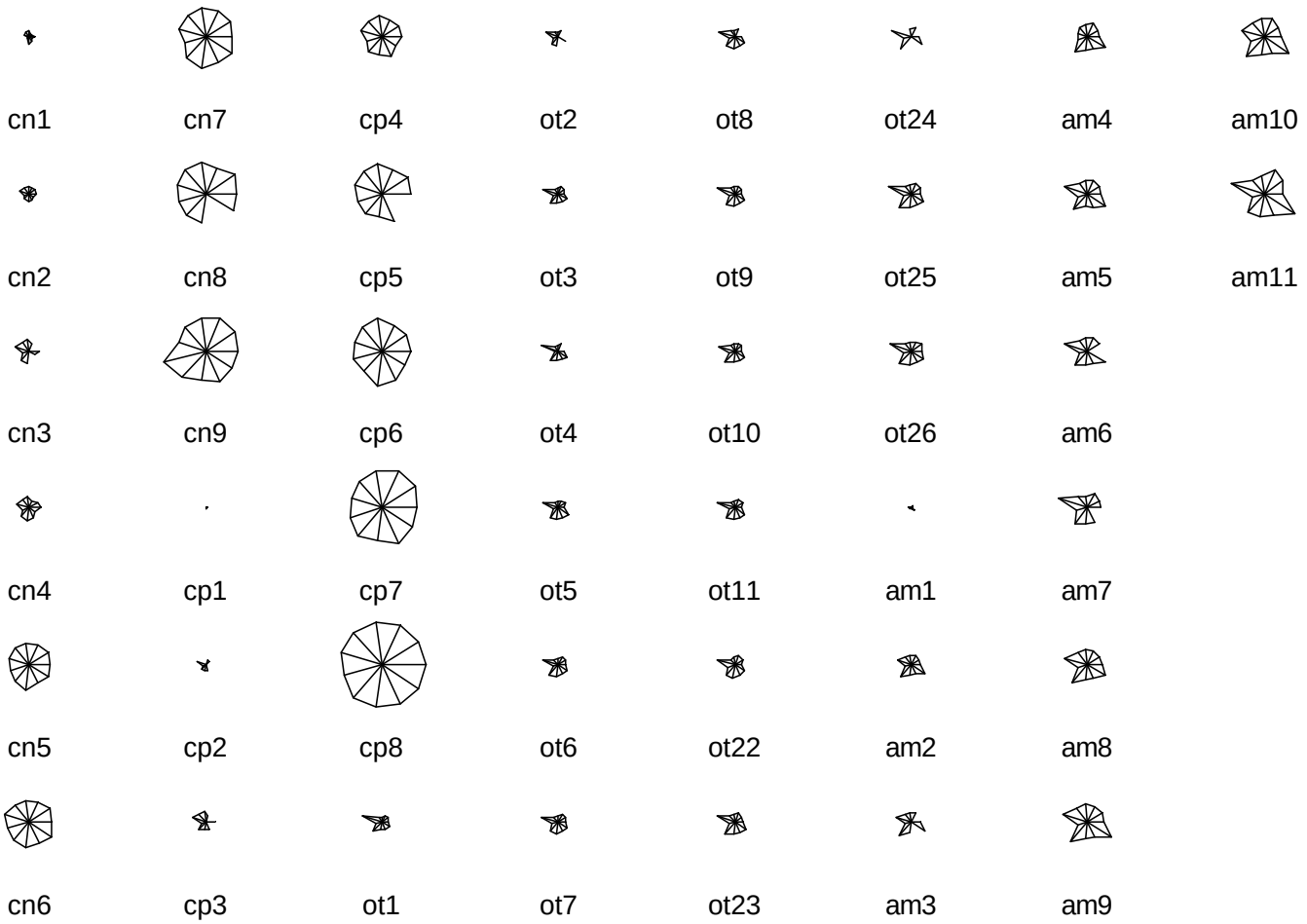
Cada dato se representará mediante una estrella que tendrá tantos rayos o ejes como variables se deseen representar.

La longitud del rayo $j$-ésimo en la estrella que representa al datos $i$ dependerá del valor de la variable $j$ en ese dato, $x_{ij}$.



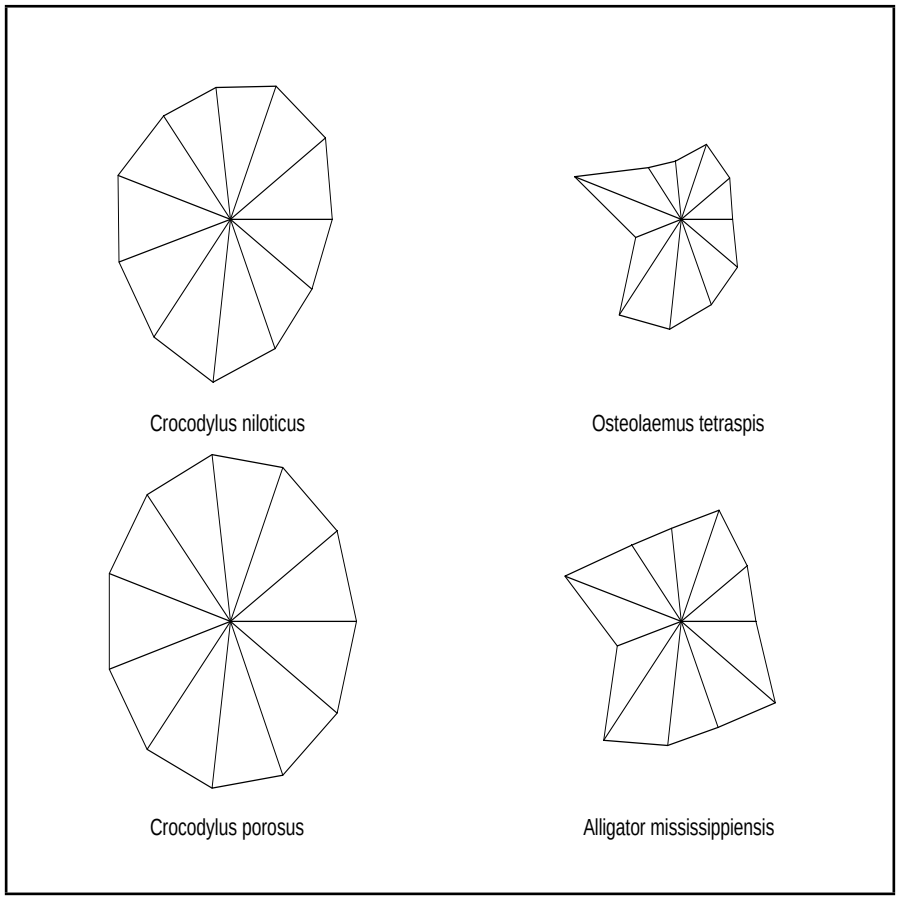
# Diagramas de estrellas - II

**Ejemplo 1.** 44 observaciones.



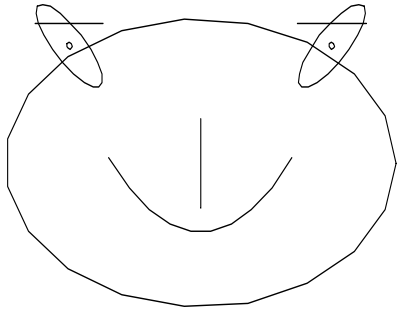
# Diagramas de estrellas - III

**Ejemplo 1.** Medias por especies.

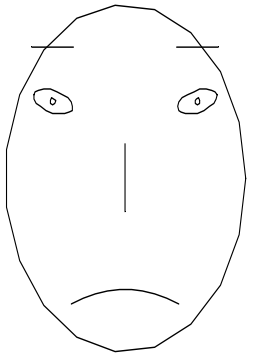


# Diagramas de caras

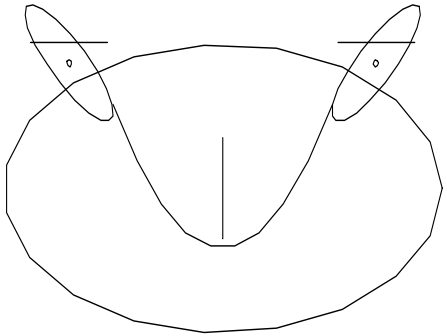
Caras de Chernoff: Cada dato se representará mediante una cara. A cada variable se asocia un rasgo o característica de una cara, por ejemplo: (1) área de la cara, (2) forma de la cara, (3) longitud de la nariz, (4) localización de la boca, (5) curva de la sonrisa (6) grosor de la boca, (7) localización, separación, inclinación, forma y grosor de los ojos, etcétera.



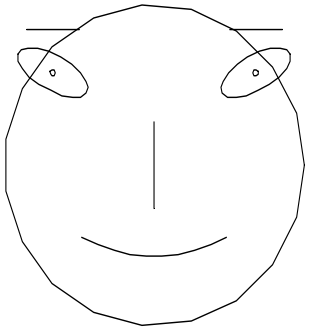
Crocodylus niloticus



Osteolaemus tetraspis



Crocodylus porosus



Alligator mississippiensis

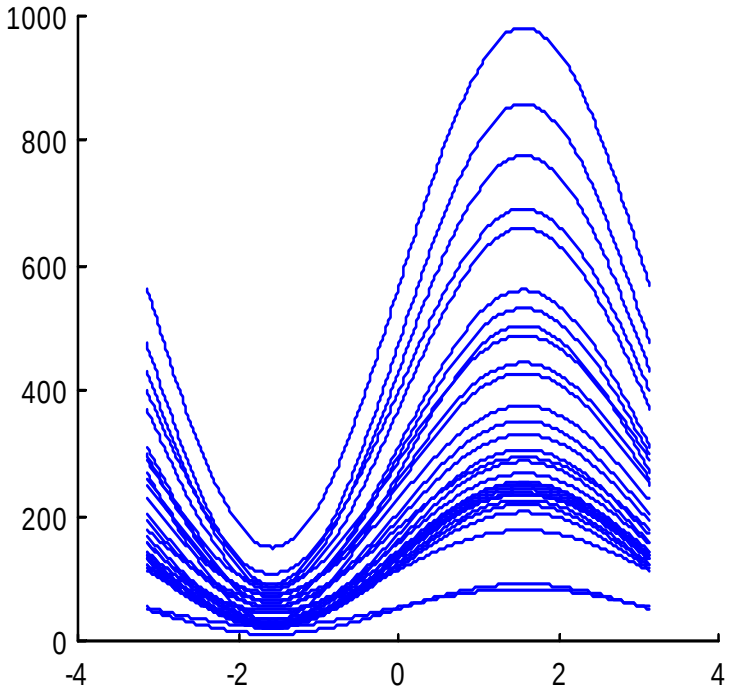
# Diagramas de Andrews - I

Los diagramas de Andrews representan al vector de observaciones $\mathbf{x}_i' = [x_{i1}\, x_{i2} \cdots x_{ip}]$ mediante el gráfico de la siguiente función:

$$f_i(t) = \frac{x_{i1}}{\sqrt{2}} + x_{i2}\sin(t) + x_{i3}\cos(t) + x_{i4}\sin(2t) + x_{i5}\cos(2t) + \cdots$$

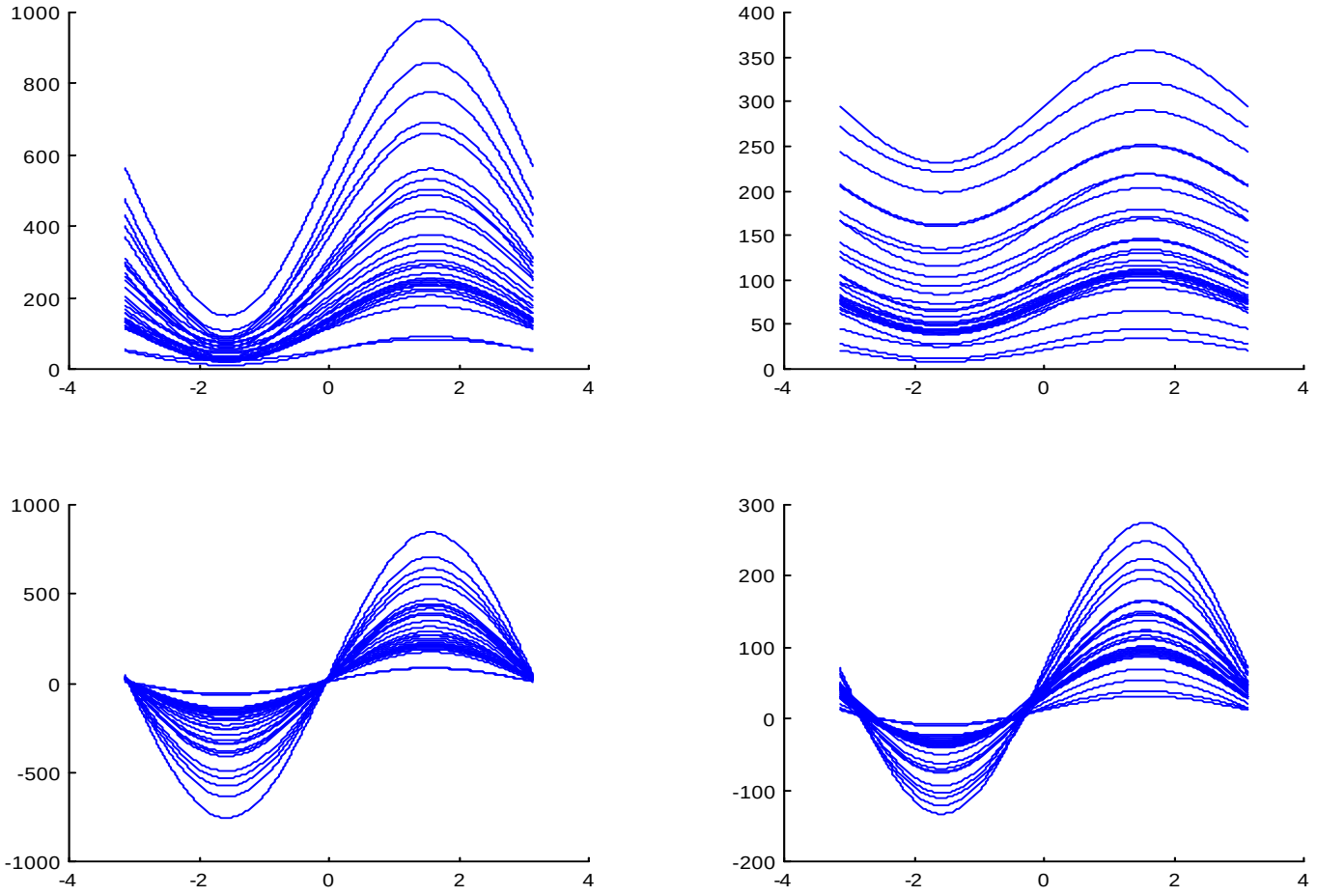
con $-\pi \leq t \leq \pi$.

Es claro que la función anterior cambia si cambiamos el orden de las variables, por lo que se recomienda explorar distintos ordenes para decidir cuál representa mejor los datos.



# Diagramas de Andrews - II

## Ejemplo 1.



# Análisis de componentes principales

- Interpretación geométrica.
- Obtención y propiedades de las componentes principales.
- Criterios para elegir el número de componentes.
- Interpretación de las componentes.

# Análisis de componentes principales

- Al estudiar una matriz de datos $\mathbf{X}$, es posible que encontremos correlaciones altas (en valor absoluto) entre varias variables. El caso más extremo es que una de las variables sea combinación lineal de las restantes. Entonces, el investigador puede preguntarse si no sería más adecuado estudiar un subconjunto de las variables originales o combinaciones lineales de éstas.

- También el número de variables, $p$, puede ser grande, lo que dificulta su análisis conjunto y en tal caso el trabajo del investigador se facilitaría si existiese un conjunto de dimensión menor $(r < p)$ de combinaciones lineales que describiera la matriz de datos $\mathbf{X}$ con una "pequeña pérdida de información".

$$\Downarrow$$

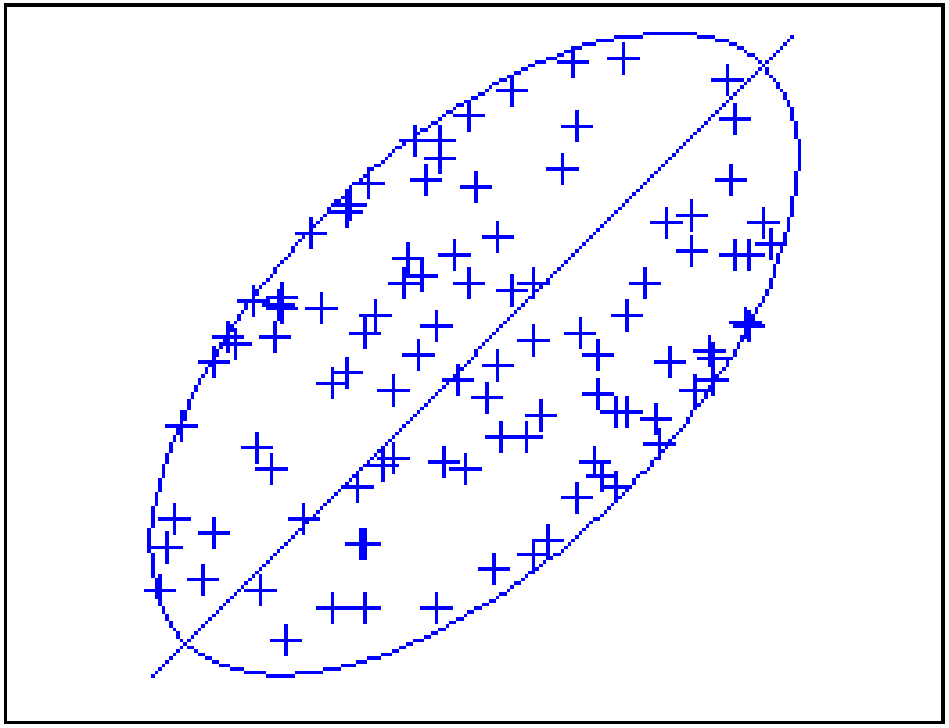
Reducción de la dimensión

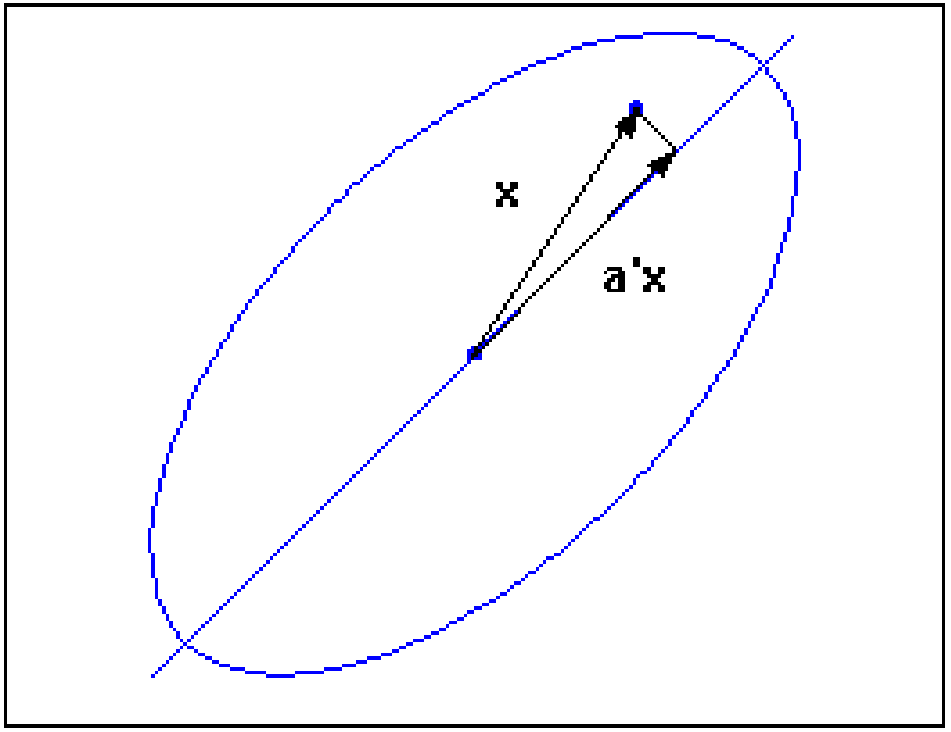
El análisis de componentes principales tiene como objetivo la reducción de la dimensión de $p$ variables preservando en lo posible la estructura de varianzas presente en la matriz $\mathbf{X}$. Se intentará explicar la mayor variabilidad posible con un número $r < p$ de combinaciones lineales de las variables originales. Así:

- La primera componente principal será la combinación lineal $\mathbf{z}_1 = \mathbf{X}\mathbf{a}_1$ que tenga varianza máxima.
- La segunda componente principal será la combinación lineal $\mathbf{z}_2 = \mathbf{X}\mathbf{a}_2$ que tenga varianza máxima y que sea incorrelada con $\mathbf{z}_1$.
- Las siguientes componentes se definen de manera similar, es decir, se intenta obtener la máxima varianza con combinaciones lineales que sean incorreladas con las componentes previamente calculadas.

¿Cuántas componentes se necesitan para explicar el 100 % de la variabilidad?

# Interpretación geométrica





# Obtención de las componentes principales

Supuesto inicial: El vector de medias cumple que $\bar{\mathbf{x}} = \mathbf{0}$.

Obtención de la primera componente principal: $\mathbf{z}_1 = \mathbf{X}\mathbf{a}_1$.

Varianza de $\mathbf{z}_1$: $\sigma^2_{\mathbf{z}_1} = \mathbf{a}_1'\mathbf{S}\mathbf{a}_1$, donde $\mathbf{S} = \frac{1}{n}\mathbf{X}'\mathbf{X}$ es la matriz de covarianzas de $\mathbf{x}$.

¿Qué problema debemos resolver para obtener $\mathbf{z}_1$?

$$\begin{aligned} &\text{Maximizar } \{\mathbf{a}_1'\mathbf{S}\mathbf{a}_1\} \\ &\text{s.a. } ||\mathbf{a}_1|| = 1. \end{aligned}$$

**Solución**:

- Mediante los multiplicadores de Lagrange:

$$L = \mathbf{a}_1'\mathbf{S}\mathbf{a}_1 - \lambda(\mathbf{a}_1'\mathbf{a}_1 - 1).$$

- Derivamos respecto de $\mathbf{a}_1$ e igualamos la derivada a $\mathbf{0}$:

$$\frac{\partial L}{\partial \mathbf{a}_1} = 2\mathbf{S}\mathbf{a}_1 - 2\lambda\mathbf{a}_1 = \mathbf{0}.$$

- La solución cumple que: $\mathbf{S}\mathbf{a}_1 = \lambda\mathbf{a}_1$.

El vector, $\mathbf{a}_1$, que define la primera componente principal es un vector propio de la matriz de covarianzas, $\mathbf{S}$.

PERO, $\sigma^2_{\mathbf{z}_1} = \mathbf{a}_1'\mathbf{S}\mathbf{a}_1 = \lambda\mathbf{a}_1'\mathbf{a}_1 = \lambda$, ENTONCES:

El vector, $\mathbf{a}_1$, que define la primera componente principal es el vector propio asociado al mayor valor **propio** de la matriz de covarianzas, $\mathbf{S}$.

Obtención de la segunda componente principal: $\mathbf{z}_2 = \mathbf{X}\mathbf{a}_2$.

- Problema a resolver:

$$\text{Maximizar} \{\mathbf{a}_2' \mathbf{S} \mathbf{a}_2\}$$

$$\text{s.a.} \begin{cases} ||\mathbf{a}_2|| = 1. \\ \mathbf{a}_1' \mathbf{a}_2 = 0. \end{cases}$$

- Que equivale a:

$$L = \mathbf{a}_2' \mathbf{S} \mathbf{a}_2 - \lambda_1 (\mathbf{a}_2' \mathbf{a}_2 - 1) - \lambda_2 \mathbf{a}_1' \mathbf{a}_2.$$

- Derivamos respecto de $\mathbf{a}_2$ e igualamos la derivada a $\mathbf{0}$:

$$\frac{\partial L}{\partial \mathbf{a}_2} = 2\mathbf{S}\mathbf{a}_2 - 2\lambda_1 \mathbf{a}_2 - \lambda_2 \mathbf{a}_1 = \mathbf{0}.$$

Obtención de la segunda componente principal:

- Premultiplicando la expresión anterior por $\mathbf{a}_1'$ obtenemos:

$$2\mathbf{a}_1'\mathbf{S}\mathbf{a}_2 - 2\lambda_1\mathbf{a}_1'\mathbf{a}_2 - \lambda_2\mathbf{a}_1'\mathbf{a}_1 = 0 + 0 + \lambda_2 = 0,$$

es decir $\lambda_2 = 0$. Por lo tanto:

$$2\mathbf{S}\mathbf{a}_2 = 2\lambda_1\mathbf{a}_2.$$

El vector, $\mathbf{a}_2$, que define la segunda componente principal es el vector propio asociado al **segundo mayor** valor propio de la matriz de covarianzas, $\mathbf{S}$.

# Componentes principales - Ejemplo

**Ejemplo 0.** En el tema anterior calculamos la matriz de covarianzas de este ejemplo:

$$\mathbf{S} = \begin{bmatrix} 0{,}386 & 0{,}178 \\ 0{,}178 & 0{,}386 \end{bmatrix}.$$

y sus valores y vectores propios:

$$\lambda_1 = 0{,}5633, \quad \mathbf{a}_1 = \begin{bmatrix} 0{,}7071 \\ 0{,}7071 \end{bmatrix}, \quad y \quad \lambda_2 = 0{,}2080, \quad \mathbf{a}_2 = \begin{bmatrix} 0{,}7071 \\ -0{,}7071 \end{bmatrix}.$$

De manera que las componentes principales son:

$$\begin{aligned} z_1 &= 0{,}7071\, x_1 + 0{,}7071\, x_2, \\ z_2 &= 0{,}7071\, x_1 - 0{,}7071\, x_2. \end{aligned}$$

# Componentes principales - Ejemplo con SPSS - I

**Ejemplo 0.** Resultados utilizando SPSS:

**Matriz de componentes**

| | Bruta | | Reescalada | |
|---|---|---|---|---|
| | Componente | | Componente | |
| | 1 | 2 | 1 | 2 |
| BASE | ,531 | -,322 | ,855 | -,519 |
| ALTURA | ,531 | ,322 | ,855 | ,519 |

Comp. bruta:
$$\begin{cases} \mathbf{b}_1 = \sqrt{\lambda_1}\mathbf{a}_1 = \sqrt{0{,}5633}\begin{bmatrix} 0{,}7071 \\ 0{,}7071 \end{bmatrix} = \begin{bmatrix} 0{,}5307 \\ 0{,}5307 \end{bmatrix} \\ \mathbf{b}_2 = \sqrt{\lambda_2}\mathbf{a}_2 = \sqrt{0{,}2080}\begin{bmatrix} 0{,}7071 \\ -0{,}7071 \end{bmatrix} = \begin{bmatrix} 0{,}3224 \\ -0{,}3224 \end{bmatrix}. \end{cases}$$

# Componentes principales - Ejemplo con SPSS - II

**Ejemplo 0.**

**Matriz de componentes**

| | Bruta | | Reescalada | |
|---|---|---|---|---|
| | Componente | | Componente | |
| | 1 | 2 | 1 | 2 |
| BASE | ,531 | -,322 | ,855 | -,519 |
| ALTURA | ,531 | ,322 | ,855 | ,519 |

Comp. re-escalada:

$$\begin{cases} \mathbf{c}_1 = \begin{bmatrix} b_{11}/\sigma_1 \\ b_{12}/\sigma_2 \end{bmatrix} = \begin{bmatrix} 0{,}5307/0{,}621 \\ 0{,}5307/0{,}621 \end{bmatrix} = \begin{bmatrix} 0{,}8551 \\ 0{,}8551 \end{bmatrix} \\ \mathbf{c}_2 = \begin{bmatrix} b_{21}/\sigma_1 \\ b_{22}/\sigma_2 \end{bmatrix} = \begin{bmatrix} 0{,}3224/0{,}621 \\ -0{,}3224/0{,}621 \end{bmatrix} = \begin{bmatrix} 0{,}5191 \\ -0{,}5191 \end{bmatrix}. \end{cases}$$

# Componentes principales - Ejemplo - III

**Matriz de componentes[a]**

| | Bruta | | | | | |
|---|---|---|---|---|---|---|
| | Componente | | | | | |
| | 1 | 2 | 3 | 4 | 5 | 6 |
| Consumo (l/100Km) | 3,404 | ,736 | -,523 | -1,137 | ,095 | 1,205 |
| Cilindrada en cc | 1714,500 | -15,221 | ,271 | ,002 | ,001 | ,000 |
| Potencia (CV) | 34,415 | 2,310 | -16,596 | ,175 | ,146 | -,027 |
| Peso total (kg) | 264,193 | 98,475 | ,422 | -,008 | -,016 | -,010 |
| Aceleración 0 a 100 km/h (segundos) | -1,507 | ,676 | 1,445 | -,183 | 1,659 | -,095 |
| Año del modelo | -1,347 | ,419 | ,785 | 3,274 | ,118 | ,415 |
| Número de cilindros | 1,620 | ,029 | ,054 | -,010 | -,036 | ,068 |

Método de extracción: Análisis de componentes principales.

**Matriz de componentes[a]**

| | Reescalada | | | | | |
|---|---|---|---|---|---|---|
| | Componente | | | | | |
| | 1 | 2 | 3 | 4 | 5 | 6 |
| Consumo (l/100Km) | ,874 | ,189 | -,134 | -,292 | ,024 | ,310 |
| Cilindrada en cc | 1,000 | -,009 | ,000 | ,000 | ,000 | ,000 |
| Potencia (CV) | ,899 | ,060 | -,434 | ,005 | ,004 | -,001 |
| Peso total (kg) | ,937 | ,349 | ,001 | ,000 | ,000 | ,000 |
| Aceleración 0 a 100 km/h (segundos) | -,546 | ,245 | ,524 | -,066 | ,601 | -,034 |
| Año del modelo | -,366 | ,114 | ,214 | ,891 | ,032 | ,113 |
| Número de cilindros | ,951 | ,017 | ,031 | -,006 | -,021 | ,040 |

Método de extracción: Análisis de componentes principales.

a. 6 componentes extraídos

# Propiedades de las componentes principales - I

1. Conservan la variabilidad inicial: la suma de las varianzas de las $p$ componentes principales es igual a la de las $p$ variables originales:

$$\sum_{j=1}^{p} \sigma_{x_j}^2 = \sum_{j=1}^{p} \lambda_j = \sum_{j=1}^{p} \sigma_{z_j}^2.$$

2. La proporción de variabilidad explicada por una componente es igual al valor propio asociado dividido por la suma de los valores propios de $\mathbf{S}$:

$$\text{var}(\sigma_{z_h}^2) = \frac{\lambda_h}{\sum_{j=1}^{p} \lambda_j}.$$

3. Las covarianzas entre la componente principal $z_h$ y la variable $\mathbf{x}$ es:

$$\text{Cov}(z_h, \mathbf{x}) = \lambda_h \mathbf{a}_h,$$

donde $\lambda_h$ es el $h$-ésimo valor propio de $\mathbf{S}$ y $\mathbf{a}_h$ su vector propio asociado.

# Propiedades de las componentes principales - II

4. La correlación entre la componente principal $z_h$ y la variable univariante $x_k$ es:

$$\text{Corr}(z_h, x_k) = \frac{\lambda_h a_{kh}}{\sqrt{\lambda_h s_k^2}} = a_{kh}\frac{\sqrt{\lambda_h}}{s_k}.$$

5. La estandarización de las componentes principales, $\mathbf{Z}$, permite obtener la estandarización multivariante de la matriz de datos, $\mathbf{X}$:

$$\mathbf{Z}_u = \mathbf{Z}\mathbf{D}^{-1/2} = \mathbf{X}\mathbf{A}\mathbf{D}^{-1/2},$$

y recordamos que $\mathbf{Y}_m = \mathbf{X}\mathbf{A}\mathbf{D}^{-1/2}\mathbf{A}'$. Por lo tanto, $\mathbf{Z}_u$ y $\mathbf{Y}_m$ son "iguales" salvo rotaciones.

# Análisis normado de componentes principales

¿Cómo es la primera componente de $\mathbf{S} = \begin{bmatrix} 100 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix}$?

Respuesta: $\mathbf{a}_1' = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$.

Problema: Una variable con "mayor" varianza que el resto de las variables tendrá asociada la primera componente principal. $\Leftarrow$ Ejemplo 2

Solución: Obtener las componentes principales de la matriz de correlaciones.

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0{,}5 \\ 0 & 0{,}5 & 1 \end{bmatrix}$$

Cuyos valores y vectores propios son:

$$\begin{aligned} \lambda_1 &= 1{,}5, \quad \mathbf{a}_1' = \begin{bmatrix} 0 & 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}, \\ \lambda_2 &= 1{,}0, \quad \mathbf{a}_2' = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}, \\ \lambda_3 &= 0{,}5, \quad \mathbf{a}_3' = \begin{bmatrix} 0 & 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}. \end{aligned}$$

# Propiedades de las componentes principales - III

6. La proporción de variabilidad explicada por una componente normada $z_h^R$ es:

$$\text{var}(\sigma^2_{z_h^R}) = \frac{\lambda_h^R}{\sum_{j=1}^p \lambda_j^R} = \frac{\lambda_h^R}{p},$$

donde $\lambda_h^R$ es el $h$-ésimo valor propio de la matriz $\mathbf{R}$.

7. Las covarianzas entre la componente principal normada $z_h^R$ y la variable vectorial $\mathbf{y}_u$ (estandarización univariante de $\mathbf{x}$) es:

$$\text{Cov}(z_h^R, \mathbf{y}_u) = \lambda_h^R \mathbf{a}_h^R,$$

donde $\lambda_h^R$ es el $h$-ésimo valor propio de $\mathbf{R}$ y $\mathbf{a}_h^R$ su vector propio asociado.

8. La correlación entre la componente principal $z_h^R$ y la variable univariante $y_k$ (estandarización univariante de $x_k$) es:

$$\text{Corr}(z_h^R, y_k) = a_{kh}^R \sqrt{\lambda_h^R}.$$

# Componentes principales normadas - Ejemplo- I

Observación: En general, los valores y vectores propios de $\mathbf{S}$ y de $\mathbf{R}$ no coinciden. Esto hace que los resultados del análisis de componentes principales y de componentes principales normadas sean, en general, diferentes.

**Ejemplo 0.** Obtenemos los valores y vectores propios de la matriz de correlaciones, $\mathbf{R} = \begin{bmatrix} 1{,}000 & 0{,}461 \\ 0{,}461 & 1{,}000 \end{bmatrix}$:

$$\lambda_1^R = 1{,}4610, \quad \mathbf{a}_1^R = \begin{bmatrix} 0{,}7071 \\ 0{,}7071 \end{bmatrix}, \quad y \quad \lambda_2^R = 0{,}5390, \quad \mathbf{a}_2^R = \begin{bmatrix} 0{,}7071 \\ -0{,}7071 \end{bmatrix}.$$

Entonces, las componentes principales son: $\begin{cases} z_1^R &= 0{,}7071\, y_1 + 0{,}7071\, y_2, \\ z_2^R &= 0{,}7071\, y_1 - 0{,}7071\, y_2. \end{cases}$

En este caso los vectores propios de $\mathbf{S}$ y $\mathbf{R}$ coinciden.

# Componentes principales normadas - Ejemplo - II

**Ejemplo 0.** Resultados utilizando SPSS:

**Matriz de componentes**

| | Componente | |
|---|---|---|
| | 1 | 2 |
| BASE | ,855 | ,519 |
| ALTURA | ,855 | -,519 |

Componentes:

$$\begin{cases} \mathbf{a}_1 = \frac{1}{\sqrt{\lambda_1}}\mathbf{b}_1 = \frac{1}{\sqrt{1{,}4610}} \begin{bmatrix} 0{,}855 \\ 0{,}855 \end{bmatrix} \approx \begin{bmatrix} 0{,}7073 \\ 0{,}7073 \end{bmatrix} \\ \mathbf{a}_2 = \frac{1}{\sqrt{\lambda_2}}\mathbf{b}_2 = \frac{1}{\sqrt{0{,}539}} \begin{bmatrix} 0{,}519 \\ -0{,}519 \end{bmatrix} \approx \begin{bmatrix} 0{,}7069 \\ -0{,}7069 \end{bmatrix} \end{cases}$$

# Componentes principales normadas - Ejemplo - III

**Matriz de componentes[a]**

| | Componente | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| Consumo (l/100Km) | ,936 | -,088 | ,195 | ,186 | -,198 | ,064 |
| Cilindrada en cc | ,964 | ,161 | ,075 | -,115 | ,052 | -,027 |
| Potencia (CV) | ,951 | ,041 | -,150 | ,148 | ,187 | ,114 |
| Peso total (kg) | ,928 | ,233 | ,205 | ,091 | ,032 | -,173 |
| Aceleración 0 a 100 km/h (segundos) | -,648 | ,120 | ,747 | ,018 | ,072 | ,053 |
| Año del modelo | -,499 | ,845 | -,172 | ,063 | -,047 | ,031 |
| Número de cilindros | ,934 | ,184 | ,103 | -,262 | -,054 | ,073 |

Método de extracción: Análisis de componentes principales.

a. 6 componentes extraídos

# Criterios de reducción de la dimensión

- Gráfico de sedimentación o de "codo": Obtener el gráfico de los valores propios, $\lambda_i$, frente a $i$. Buscar un codo en el gráfico, i.e., un punto a partir del cual los valores propios son aproximadamente iguales.

- Criterio de la varianza explicada: Seleccionar el número de componentes necesario para explicar una proporción predeterminada de la varianza, por ejemplo, el 80 % o el 90 %.

- Criterio del valor propio: Seleccionar los componentes principales asociados a valores propios superiores a un valor prefijado, por ejemplo, la varianza media:

$$\sum_{j=1}^{p} \lambda_j / p \qquad \text{en componentes principales,}$$

$$\sum_{j=1}^{p} \lambda_j^R / p = 1 \qquad \text{en componentes principales normadas.}$$

# Reducción de la dimensión - Ejemplo - I

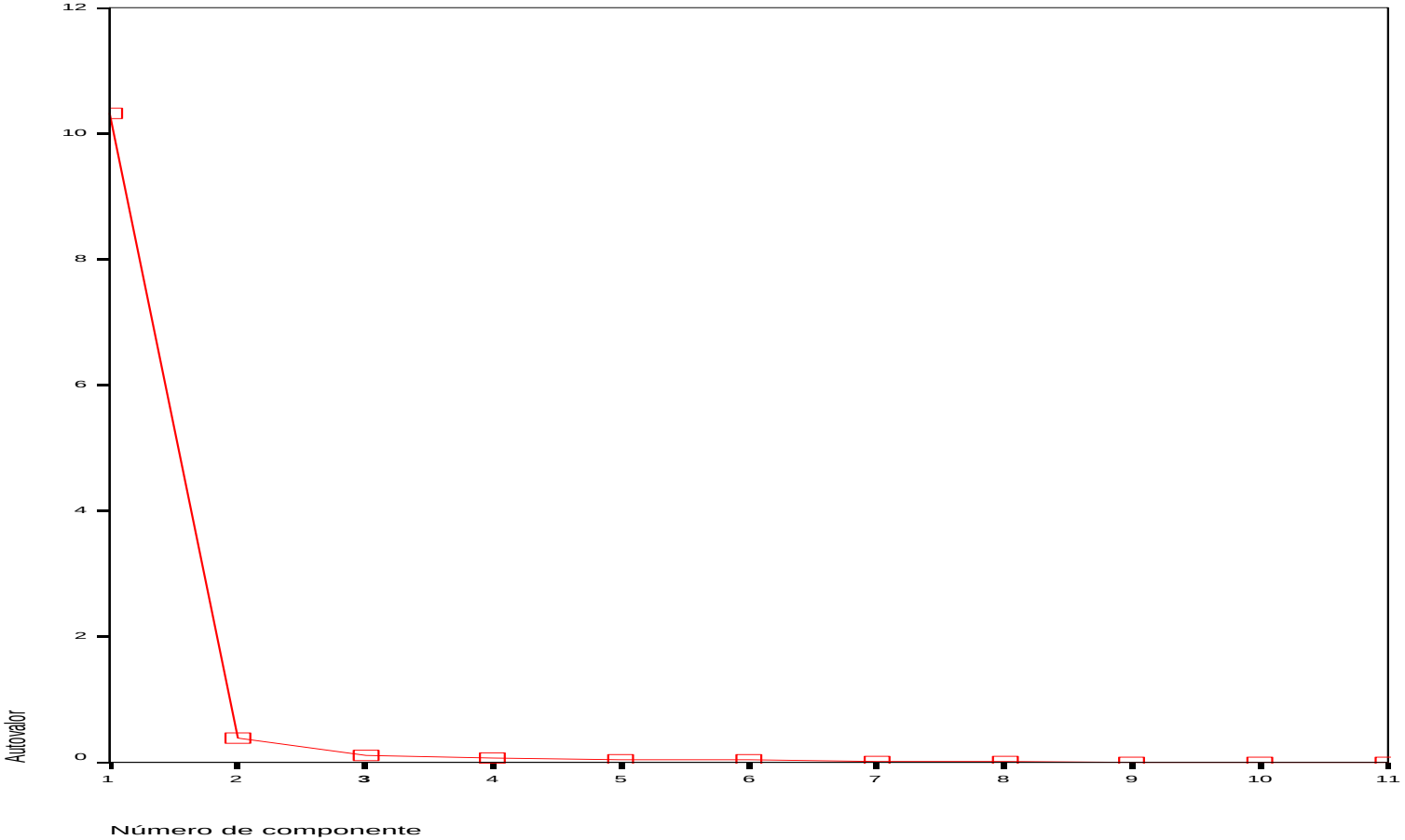
**Ejemplo 1.** Análisis de componentes principales normadas.

- El criterio de la variabilidad explicada ($> 90\,\%$) sugiere utilizar una componente.
- El criterio del valor propio ($> 1$) sugiere utilizar una componente.

| Componente | Autovalores iniciales | | |
|---|---|---|---|
| | Total | % de la varianza | % acumulado |
| 1 | 10,326 | 93,871 | 93,871 |
| 2 | ,383 | 3,480 | 97,352 |
| 3 | ,114 | 1,038 | 98,390 |
| 4 | 6,490E-02 | ,590 | 98,980 |
| 5 | 4,130E-02 | ,375 | 99,355 |
| 6 | 3,910E-02 | ,355 | 99,711 |
| 7 | 1,965E-02 | ,179 | 99,889 |
| 8 | 7,515E-03 | 6,832E-02 | 99,958 |
| 9 | 3,306E-03 | 3,005E-02 | 99,988 |
| 10 | 1,051E-03 | 9,556E-03 | 99,997 |
| 11 | 3,090E-04 | 2,809E-03 | 100,000 |

# Reducción de la dimensión - Ejemplo - II

**Ejemplo 1.**




- El criterio del gráfico de sedimentación sugiere utilizar una componente.

# Reducción de la dimensión - Ejemplo - III

**Ejemplo 2.** Análisis de componentes principales.

**Varianza total explicada**

| Componente | Autovalores iniciales | | | Sumas de las saturaciones al cuadrado de la extracción | | |
|---|---|---|---|---|---|---|
| | Total | % de la varianza | % acumulado | Total | % de la varianza | % acumulado |
| 1 | 3010511,5 | 99,661 | 99,661 | 3010511,5 | 99,661 | 99,661 |
| 2 | 9935,469 | ,329 | 99,990 | | | |
| 3 | 278,648 | ,009 | 99,999 | | | |
| 4 | 12,078 | ,000 | 100,000 | | | |
| 5 | 2,798 | 9,263E-05 | 100,000 | | | |
| 6 | 1,639 | 5,426E-05 | 100,000 | | | |
| 7 | ,268 | 8,878E-06 | 100,000 | | | |

Método de extracción: Análisis de Componentes principales.

# Reducción de la dimensión - Ejemplo - IV

**Ejemplo 2.** Análisis de componentes principales normadas.

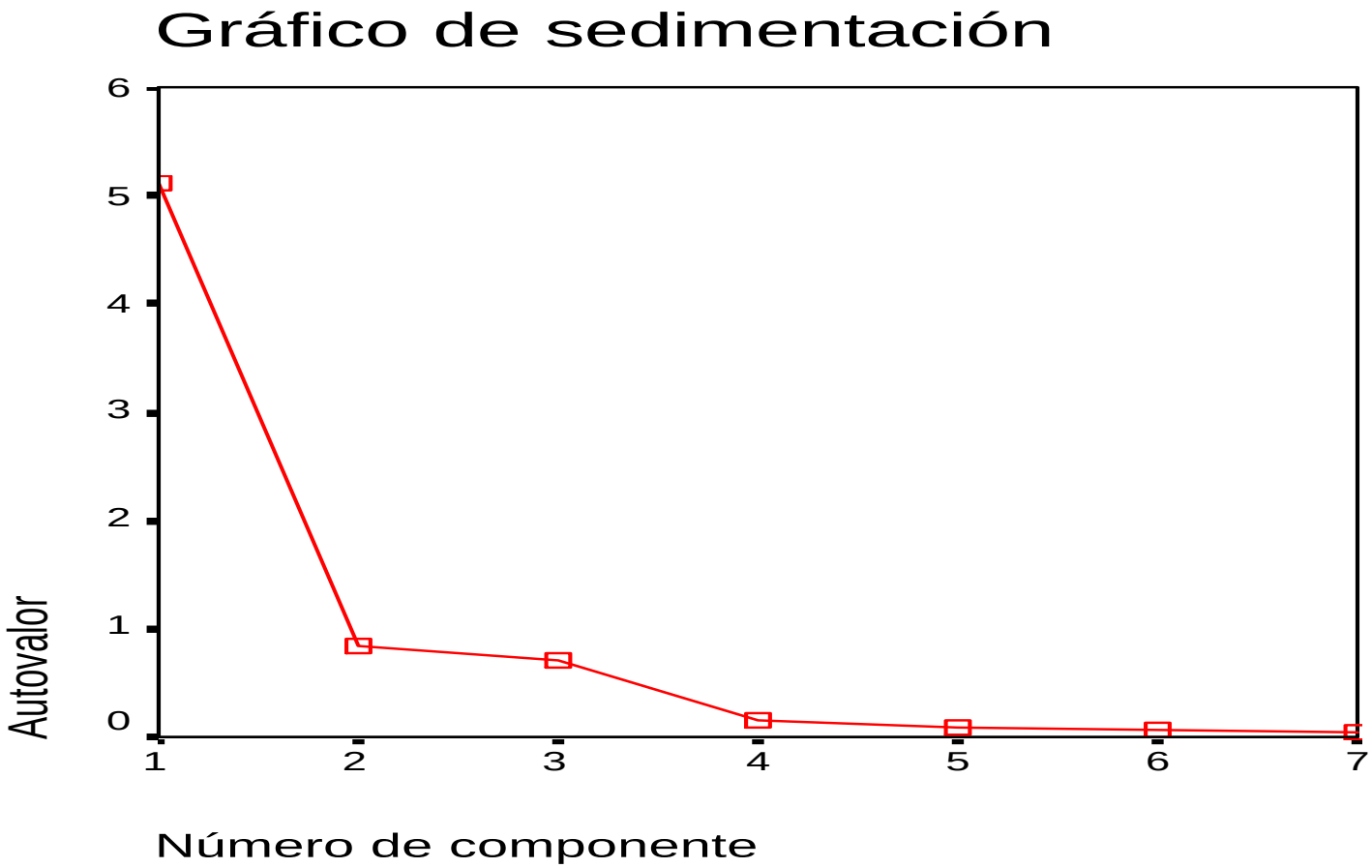
**Varianza total explicada**

| Componente | Autovalores iniciales | | | Sumas de las saturaciones al cuadrado de la extracción | | |
|---|---|---|---|---|---|---|
| | Total | % de la varianza | % acumulado | Total | % de la varianza | % acumulado |
| 1 | 5,112 | 73,024 | 73,024 | 5,112 | 73,024 | 73,024 |
| 2 | ,852 | 12,168 | 85,192 | | | |
| 3 | ,706 | 10,085 | 95,276 | | | |
| 4 | ,151 | 2,158 | 97,434 | | | |
| 5 | ,088 | 1,264 | 98,698 | | | |
| 6 | ,057 | ,813 | 99,511 | | | |
| 7 | ,034 | ,489 | 100,000 | | | |

Método de extracción: Análisis de Componentes principales.

# Reducción de la dimensión - Ejemplo - V

**Ejemplo 2.** Análisis de componentes principales normadas.

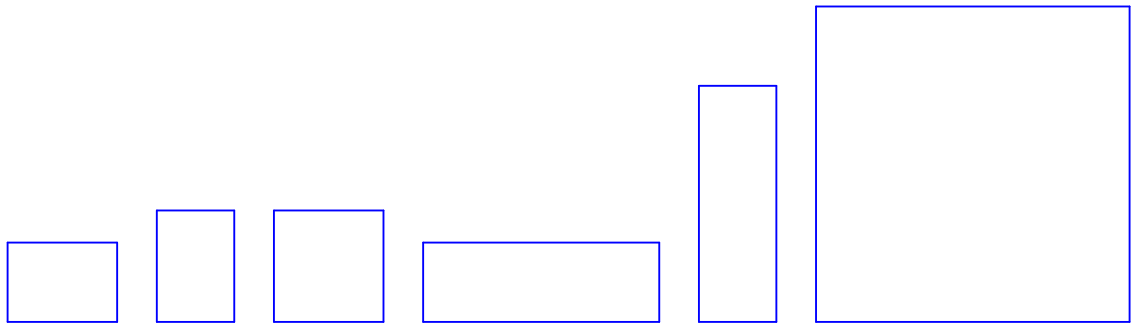


# Interpretación de las componentes - Ejemplo - I

**Ejemplo 0.** Las componentes principales: $\begin{cases} z_1 &= 0{,}7071\, x_1 + 0{,}7071\, x_2, \\ z_2 &= 0{,}7071\, x_1 - 0{,}7071\, x_2. \end{cases}$

- La primera componente, que explica el 73.03 % de la variabilidad total, asigna igual peso a las variables `base` y `altura`, $x_1$ y $x_2$. Si re–escribimos esta componente como: $z_1 = \frac{0{,}7071}{2}(2x_1 + 2x_2)$ podemos interpretarla como una ponderación del perímetro del rectángulo.

  Si ordenamos los datos según esa componente. obtenemos:



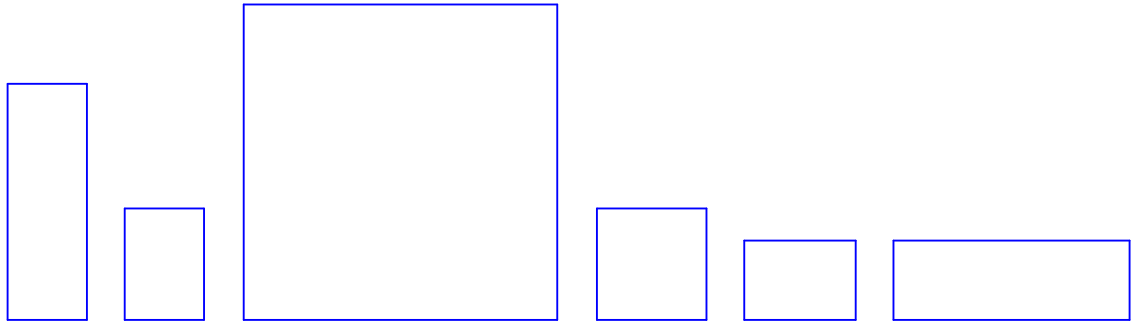
  Es decir, los rectángulos quedan ordenados según su "tamaño".

# Interpretación de las componentes - Ejemplo- II

**Ejemplo 0.** Las componentes principales: $\begin{cases} z_1 &= 0{,}7071\, x_1 + 0{,}7071\, x_2, \\ z_2 &= 0{,}7071\, x_1 - 0{,}7071\, x_2. \end{cases}$

- La segunda componente, que explica el 26.97 % de la variabilidad total, asigna igual peso a la `base` y la `altura` pero con signo diferente. Así, por ejemplo, un valor de $z_2$ positivo corresponderá a un rectángulo con más base que altura.

  Si ordenamos los datos según esa componente, obtenemos:



  Es decir, los rectángulos quedan ordenados según su "forma".

# Interpretación de las componentes - Casos Particulares - I

Componentes principales de una matriz diagonal: $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_p^2 \end{bmatrix}$.

Entonces, los pares valor–vector propio son:

$$\sigma_1^2 \text{ y } \mathbf{a}_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \sigma_2^2 \text{ y } \mathbf{a}_2 = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \quad \cdots, \quad \sigma_p^2 \text{ y } \mathbf{a}_p = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}.$$

- Las componentes principales en matrices diagonales son las variables originales.

- En una matriz de covarianzas no necesariamente diagonal, si existe una variable, $x_k$, incorrelada con el resto de las variables, entonces habrá una componente principal que dará peso 1 a la variable $x_k$ y 0 al resto.

# Interpretación de las componentes - Casos Particulares - II

Componentes principales de una matriz equicorrelada: $\mathbf{R} = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}$

Entonces, los pares de valor–vector propio son:

$$\begin{aligned}
\lambda_1 = 1 + (p-1)\rho \quad & \mathbf{a}_1' = \left[\frac{1}{\sqrt{p}}, \frac{1}{\sqrt{p}}, \frac{1}{\sqrt{p}}, \frac{1}{\sqrt{p}}, \ldots, \frac{1}{\sqrt{p}}\right], \\
\lambda_2 = 1 - \rho \quad & \mathbf{a}_2' = \left[\frac{1}{\sqrt{1\times 2}}, \frac{-1}{\sqrt{1\times 2}}, 0, 0, \ldots, 0\right], \\
\lambda_3 = 1 - \rho \quad & \mathbf{a}_3' = \left[\frac{1}{\sqrt{2\times 3}}, \frac{1}{\sqrt{2\times 3}}, \frac{-2}{\sqrt{2\times 3}}, 0, \ldots, 0\right], \\
\vdots \quad & \vdots \\
\lambda_p = 1 - \rho \quad & \mathbf{a}_p' = \left[\frac{1}{\sqrt{(p-1)p}}, \frac{1}{\sqrt{(p-1)p}}, \frac{1}{\sqrt{(p-1)p}}, \frac{1}{\sqrt{(p-1)p}}, \ldots, \frac{-(p-1)}{\sqrt{(p-1)p}}\right].
\end{aligned}$$

# Interpretación de las componentes - Casos Particulares - III

Componentes principales de una matriz equicorrelada:

- Si $\rho > 0$, entonces el mayor valor propio es $\lambda_1 = 1 + (p-1)\rho$ y su vector propio asociado $\mathbf{a}_1$ define una componente principal que asigna igual peso a todas las variables: $z_1 = \frac{1}{\sqrt{p}} \sum_{j=1}^{p} x_j$.

- Si $\rho > 0$, entonces la primera componente principal explica una proporción $\frac{1+(p-1)\rho}{p} = \rho + \frac{1-\rho}{p}$. Por ejemplo, si $\rho = 0{,}9$ y $p = 10$, entonces la primera componente explica el 90.01 % de la variabilidad total.

- Si $\rho$ es cercano a 1, entonces las restantes $p-1$ componentes, explican una pequeña proporción de la variabilidad total.

# Interpretación de las componentes - Ejemplo - I

**Ejemplo 1.** La matriz de correlaciones de este ejemplo es aproximadamente equicorrelada:

| | CL | CW | SW | SL | DCL | OW | OIW | OL | LCR | WCR | WN |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CL | 1,000 | ,991 | ,976 | ,997 | ,999 | ,821 | ,963 | ,929 | ,962 | ,984 | ,900 |
| CW | ,991 | 1,000 | ,987 | ,986 | ,989 | ,840 | ,965 | ,934 | ,968 | ,993 | ,914 |
| SW | ,976 | ,987 | 1,000 | ,969 | ,974 | ,859 | ,952 | ,950 | ,956 | ,985 | ,941 |
| SL | ,997 | ,986 | ,969 | 1,000 | ,998 | ,796 | ,958 | ,917 | ,958 | ,978 | ,890 |
| DCL | ,999 | ,989 | ,974 | ,998 | 1,000 | ,824 | ,961 | ,930 | ,964 | ,983 | ,900 |
| OW | ,821 | ,840 | ,859 | ,796 | ,824 | 1,000 | ,766 | ,906 | ,858 | ,861 | ,893 |
| OIW | ,963 | ,965 | ,952 | ,958 | ,961 | ,766 | 1,000 | ,895 | ,932 | ,958 | ,833 |
| OL | ,929 | ,934 | ,950 | ,917 | ,930 | ,906 | ,895 | 1,000 | ,922 | ,945 | ,954 |
| LCR | ,962 | ,968 | ,956 | ,958 | ,964 | ,858 | ,932 | ,922 | 1,000 | ,974 | ,886 |
| WCR | ,984 | ,993 | ,985 | ,978 | ,983 | ,861 | ,958 | ,945 | ,974 | 1,000 | ,908 |
| WN | ,900 | ,914 | ,941 | ,890 | ,900 | ,893 | ,833 | ,954 | ,886 | ,908 | 1,000 |

# Interpretación de las componentes - Ejemplo - II

**Ejemplo 1.**

La primera componente principal estará definida por un vector aproximadamente igual a $\mathbf{a}_1' = \left[\frac{1}{\sqrt{11}}, \frac{1}{\sqrt{11}}, \ldots, \frac{1}{\sqrt{11}}\right]$.

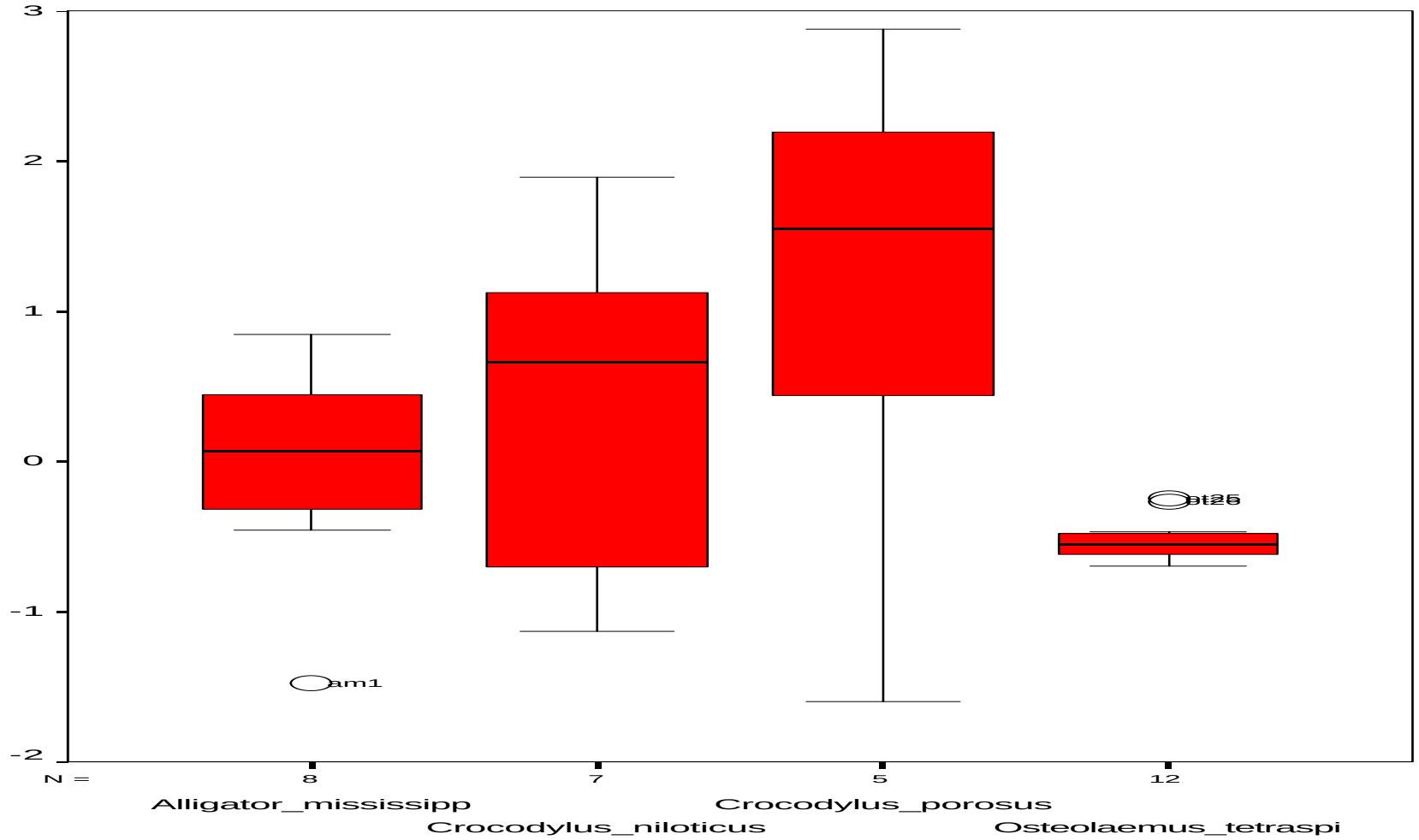
Recordemos que en SPSS aparece $\sqrt{\lambda_1}\mathbf{a}_1$, por tanto los coeficientes serán aproximadamente iguales a $\frac{\sqrt{10{,}326}}{\sqrt{11}} \approx 0{,}969$.

**Matriz de componentes**

| | Componente |
|---|---|
| | 1 |
| CL | ,989 |
| CW | ,992 |
| SW | ,991 |
| SL | ,982 |
| DCL | ,988 |
| OW | ,882 |
| OIW | ,957 |
| OL | ,964 |
| LCR | ,975 |
| WCR | ,993 |
| WN | ,940 |

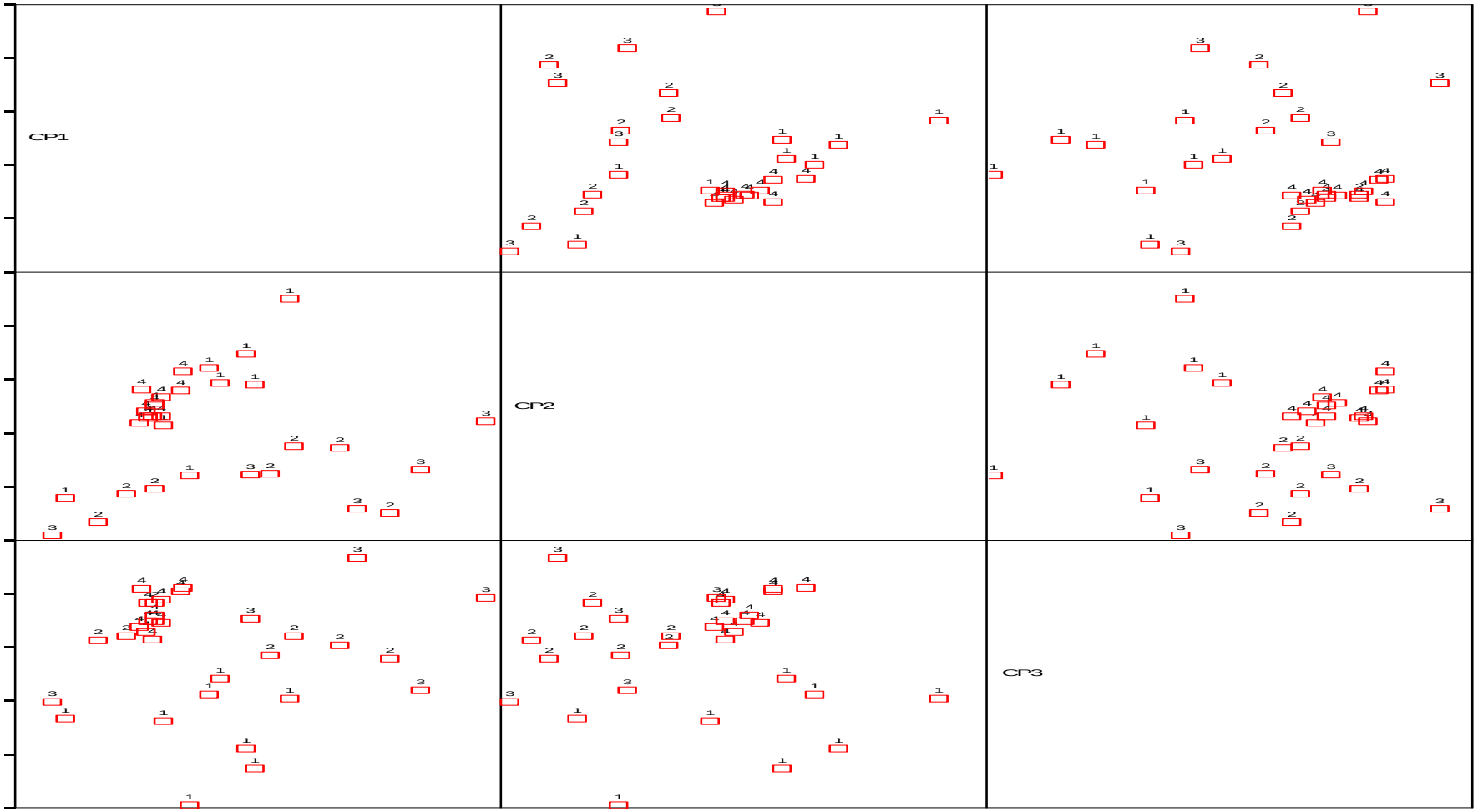
# Interpretación de las componentes - Ejemplo - III

**Ejemplo 1.** Diagrama de caja de la primera componente.



# Interpretación de las componentes - Ejemplo - IV

**Ejemplo 1.** Matriz de diagramas de dispersión de las tres primeras CP.



# Interpretación de las componentes - Ejemplo - V

**Ejemplo 2.** Análisis de componentes principales normadas.

**Matriz de componentes[a]**

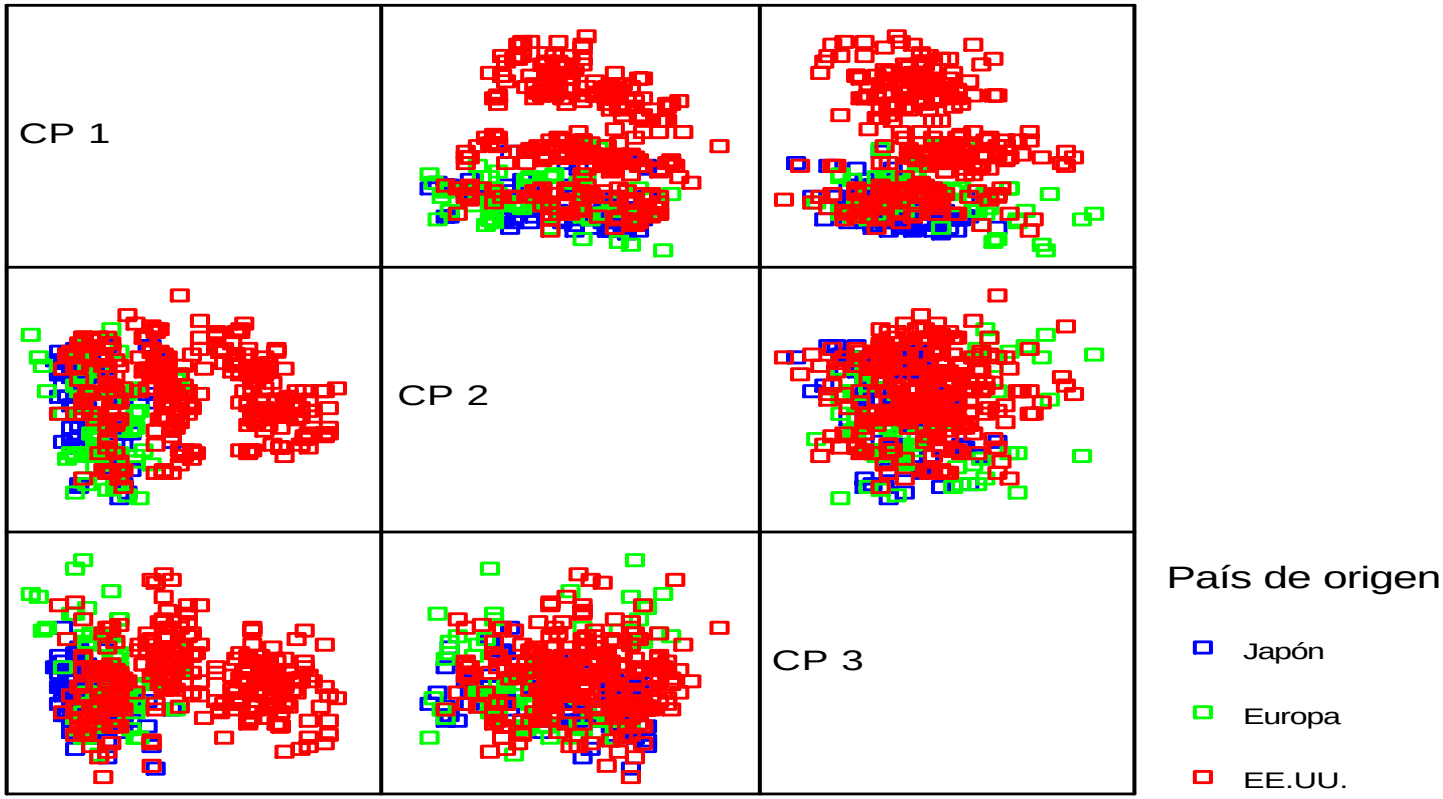
| | Componente | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| Consumo (l/100Km) | ,936 | -,088 | ,195 |
| Cilindrada en cc | ,964 | ,161 | ,075 |
| Potencia (CV) | ,951 | ,041 | -,150 |
| Peso total (kg) | ,928 | ,233 | ,205 |
| Aceleración 0 a 100 km/h (segundos) | -,648 | ,120 | ,747 |
| Año del modelo | -,499 | ,845 | -,172 |
| Número de cilindros | ,934 | ,184 | ,103 |

Método de extracción: Análisis de componentes principales.

a. 3 componentes extraídos

# Interpretación de las componentes - Ejemplo - VI

**Ejemplo 2.** Análisis de componentes principales normadas.



# Interpretación de las componentes - Ejemplo - VII

**Ejemplo 5.** Esclerosis múltiple.

| Componente | Autovalores iniciales | | |
|---|---|---|---|
| | Total | % de la varianza | % acumulado |
| 1 | 2,917 | 58,342 | 58,342 |
| 2 | 1,227 | 24,534 | 82,876 |
| 3 | ,703 | 14,056 | 96,932 |
| 4 | 9,095E-02 | 1,819 | 98,751 |
| 5 | 6,245E-02 | 1,249 | 100,000 |

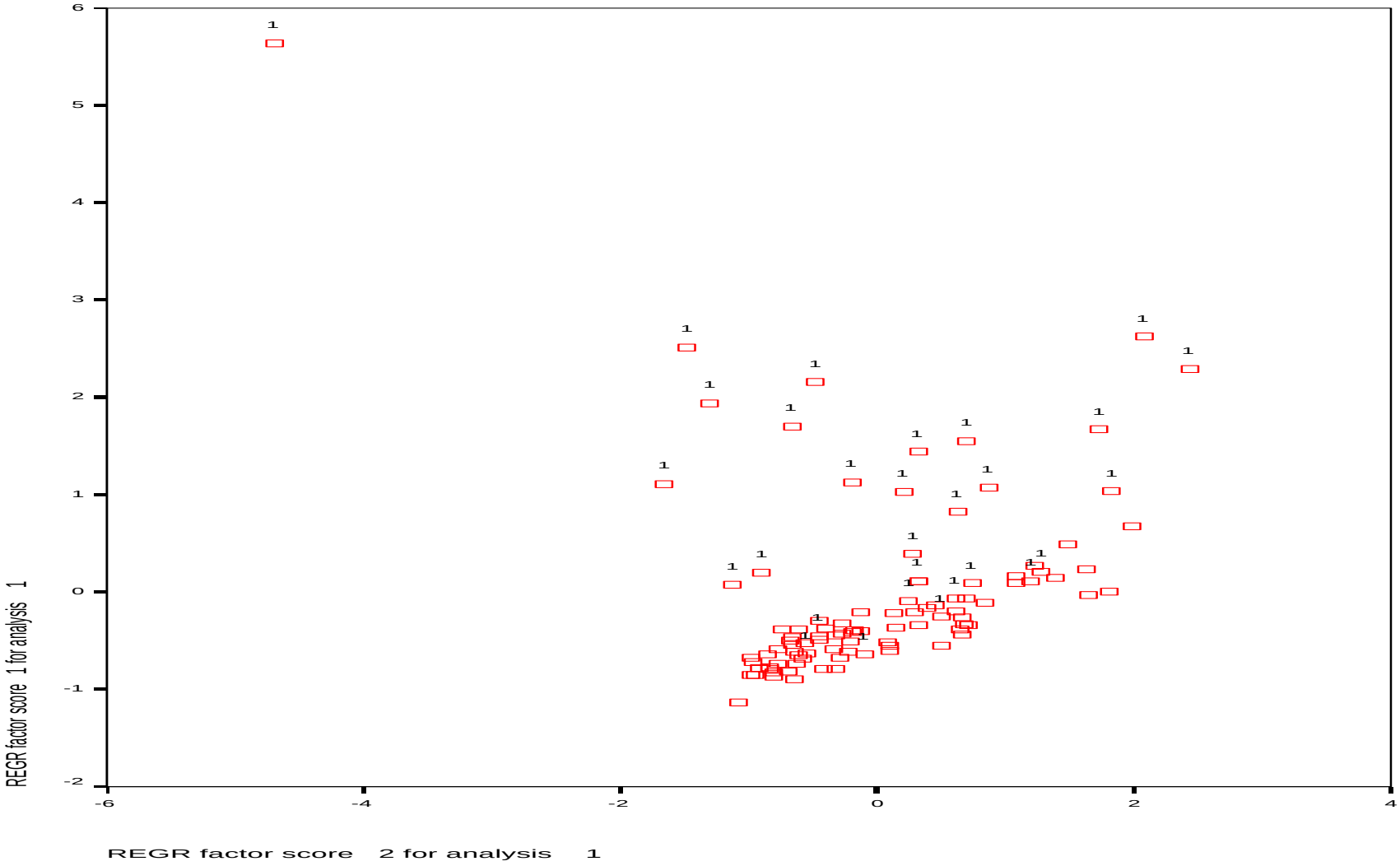
**Matriz de componentes**

| | Componente | |
|---|---|---|
| | 1 | 2 |
| EDAD | ,299 | ,734 |
| R1SUMA | ,878 | ,316 |
| R1DIF | ,862 | -,433 |
| R2SUMA | ,852 | ,336 |
| R2DIF | ,766 | -,535 |

- La primera componente da mayor peso a las variables relacionadas con las respuestas a estímulos visuales, y menor peso a la edad.
- La segunda componente da mayor peso a la edad, y por otra parte contrapone las variables de tipo respuesta conjunta y respuesta diferencial.

# Interpretación de las componentes - Ejemplo - VIIII

**Ejemplo 5.** Esclerosis múltiple.



# Ejemplo con gráficos de control - I

**Ejemplo 4.** Seis tipos de escenarios.

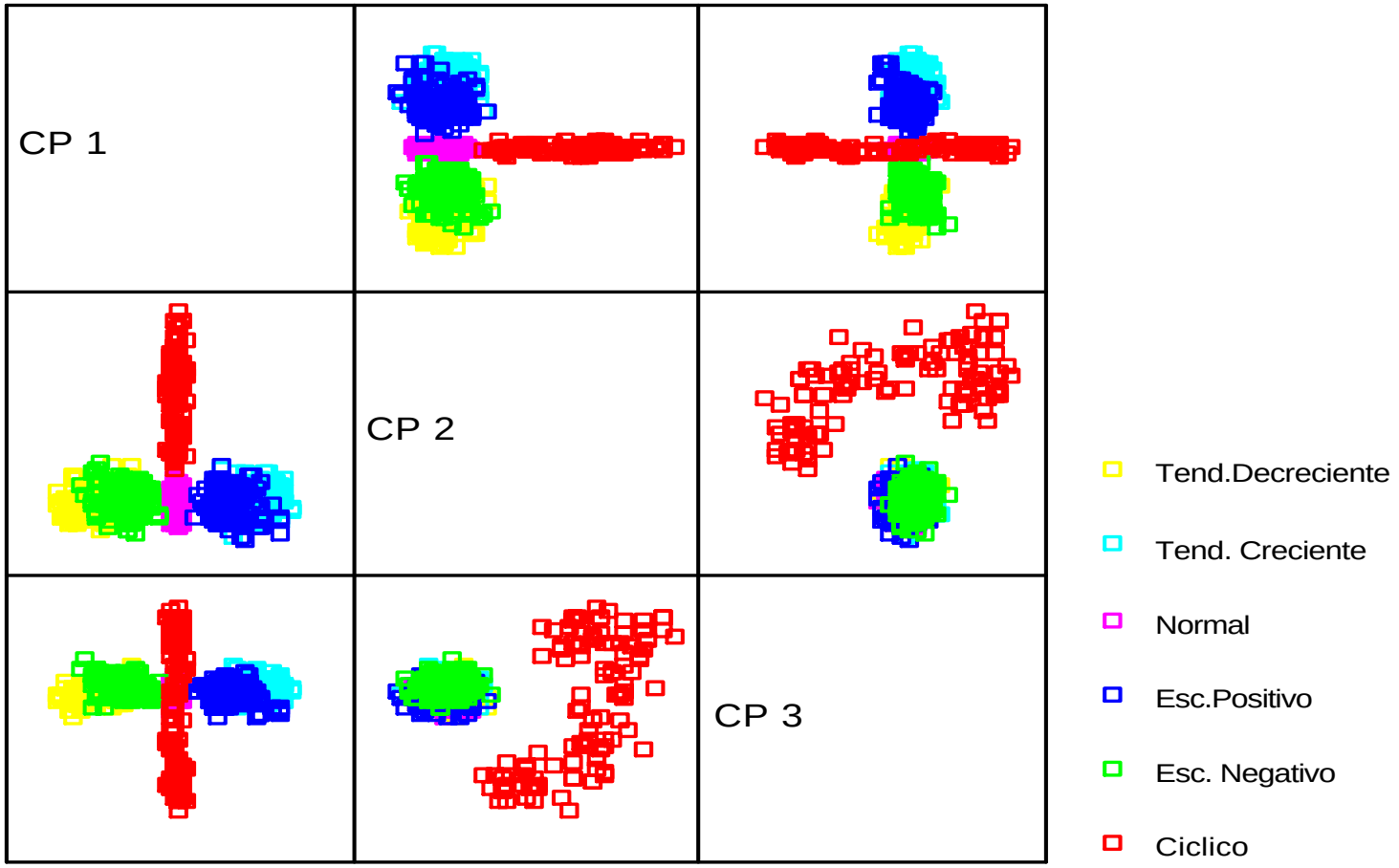
**Varianza total explicada**

| Componente | Autovalores iniciales | | | Sumas de las saturaciones al cuadrado de la extracción | | |
|---|---|---|---|---|---|---|
| | Total | % de la varianza | % acumulado | Total | % de la varianza | % acumulado |
| 1 | 31,479 | 52,465 | 52,465 | 31,479 | 52,465 | 52,465 |
| 2 | 5,930 | 9,884 | 62,348 | 5,930 | 9,884 | 62,348 |
| 3 | 4,184 | 6,973 | 69,322 | 4,184 | 6,973 | 69,322 |
| 4 | 1,989 | 3,314 | 72,636 | 1,989 | 3,314 | 72,636 |
| 5 | 1,846 | 3,077 | 75,712 | 1,846 | 3,077 | 75,712 |
| 6 | 1,254 | 2,090 | 77,803 | 1,254 | 2,090 | 77,803 |
| 7 | 1,011 | 1,685 | 79,488 | 1,011 | 1,685 | 79,488 |
| 8 | ,957 | 1,595 | 81,082 | | | |
| 9 | ,736 | 1,226 | 82,309 | | | |
| 10 | ,715 | 1,192 | 83,501 | | | |
| : | : | : | : | | | |
| 60 | ,045 | ,076 | 99,792 | | | |

Método de extracción: Análisis de Componentes principales.

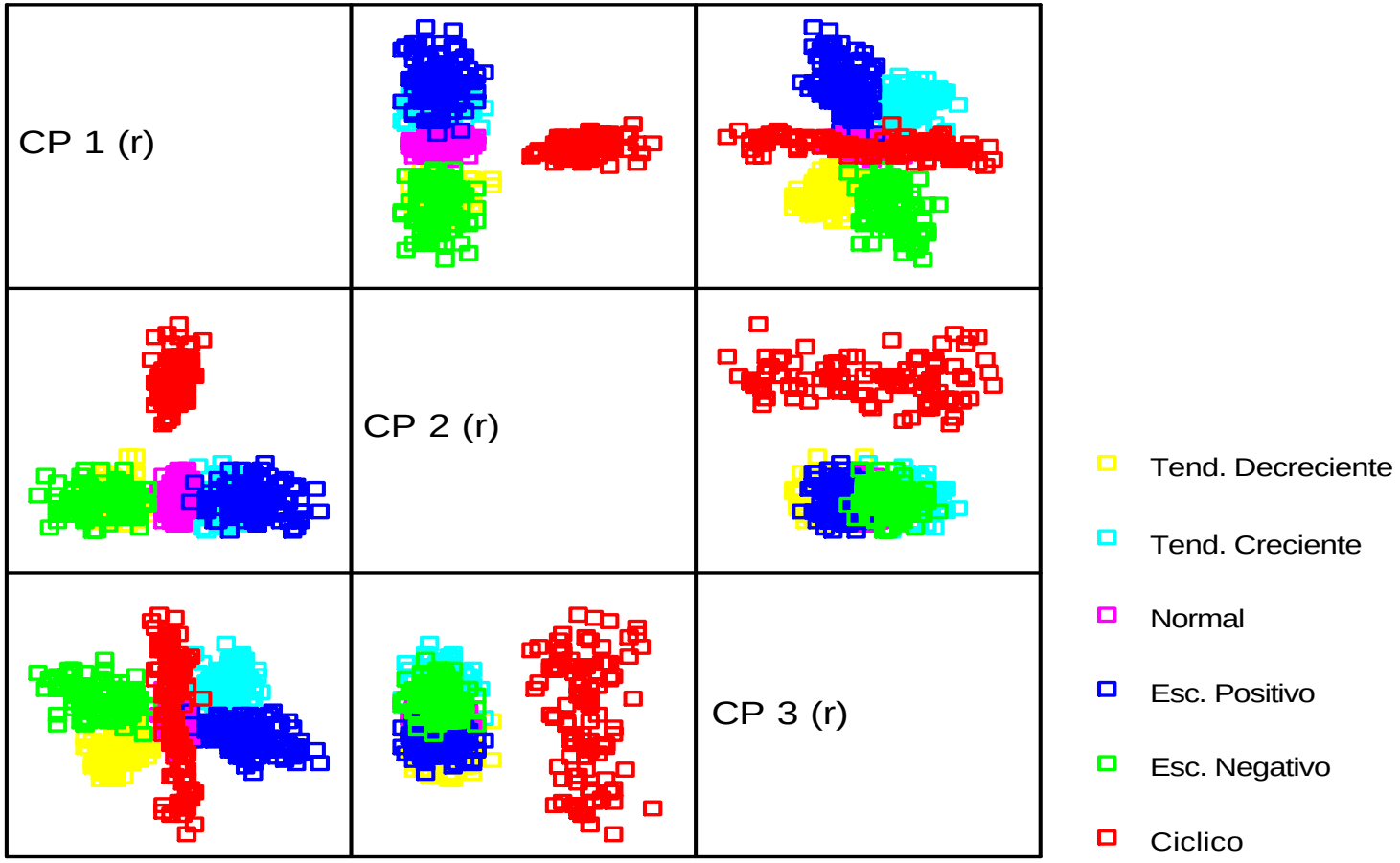
# Ejemplo con gráficos de control - II

**Ejemplo 4.** (AF) Componentes principales normadas.



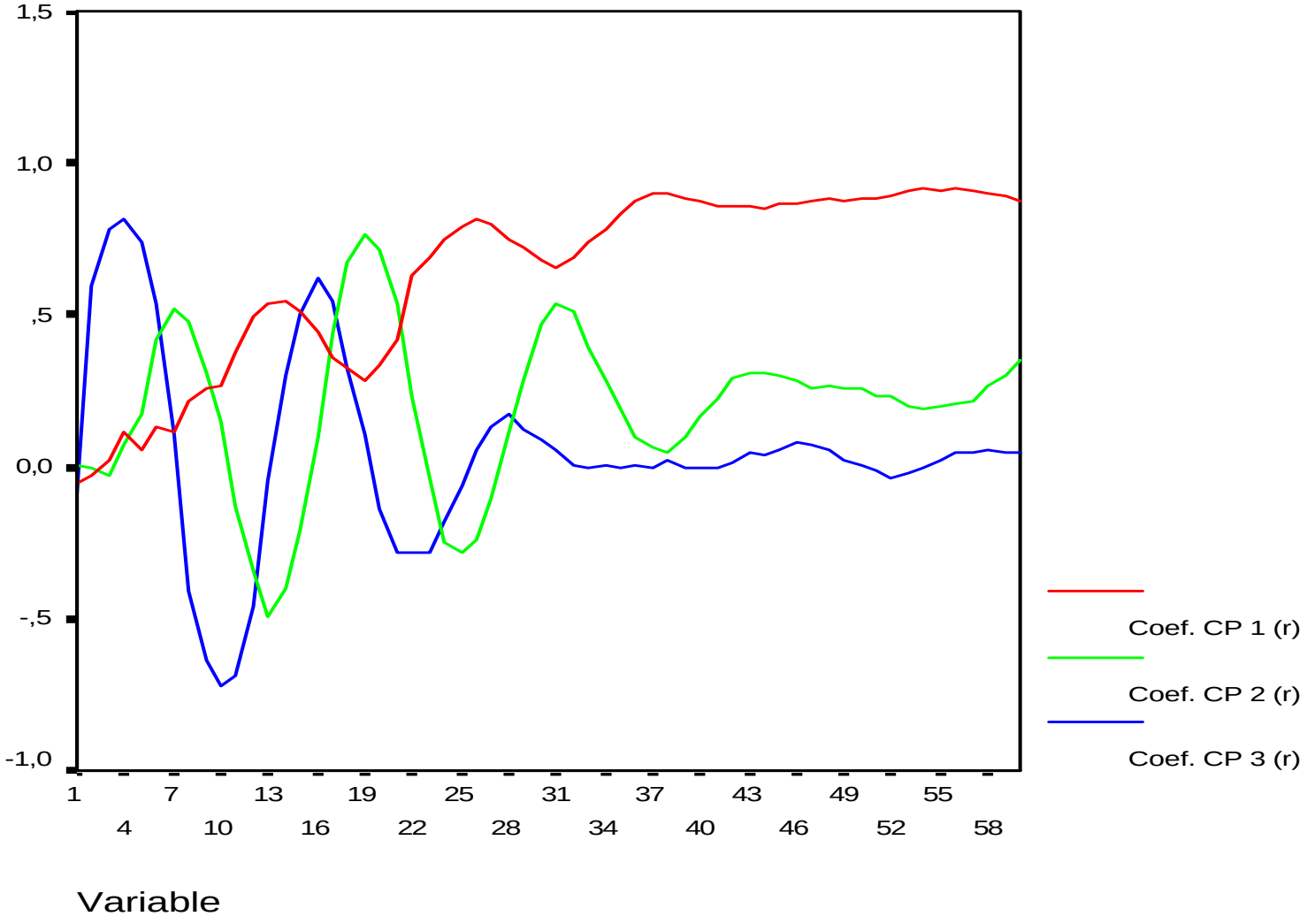
# Ejemplo con gráficos de control - III

**Ejemplo 4.** (AF) Componentes principales normadas rotadas.



# Ejemplo con gráficos de control - IV

**Ejemplo 4.** Interpretación de las CP - factores.



# Lecturas recomendadas

- Análisis descriptivo: Capítulo 2 de Baillo y Grané (2007); Capítulo 1 de Cuadras (2004); Capítulo 1 de Johnson y Wichern (2002); Capítulos 3 y 4 de Peña (2002);
- Análisis de componentes principales: Capítulo 4 de Baillo y Grané (2007); Capítulo 5 de Cuadras (2004); Capítulo 8 de Johnson y Wichern (2002); Capítulo 2 de McGarigal et al (2000); Capítulo 5 de Peña (2002); Capítulo 7 de Selvin (1995).

**Baillo, A. y Grané, A. (2007)** 100 problemas resueltos de estadística multivariante (Implementados en Matlab), Delta Publicaciones.

**Cuadras, C. (2004)** Análisis multivariante, Universidad de Barcelona.

**Johnson, R.A. y Wichern, W.A. (2002)** Applied multivariate statistical analysis, Prentice Hall.

**McGarigal, K., Cushman, S. y Stafford, S. (2000)** Multivariate analysis for wildlife and ecology research, Springer.

**Peña, D. (2002)** Análisis de datos multivariantes, McGraw–Hill.

**Selvin, S. (1995)** Practical biostatistical methods, Duxbury Press.