# UNIDAD DIDÁCTICA 2: Reducción de la dimensión
# Tema 1: Análisis de Componentes Principales

**TÉCNICAS ESTADÍSTICAS PARA EL APRENDIZAJE I**

Máster Universitario en Estadística Computacional
y Ciencia de Datos para la Toma de Decisiones

# Índice

# Introducción

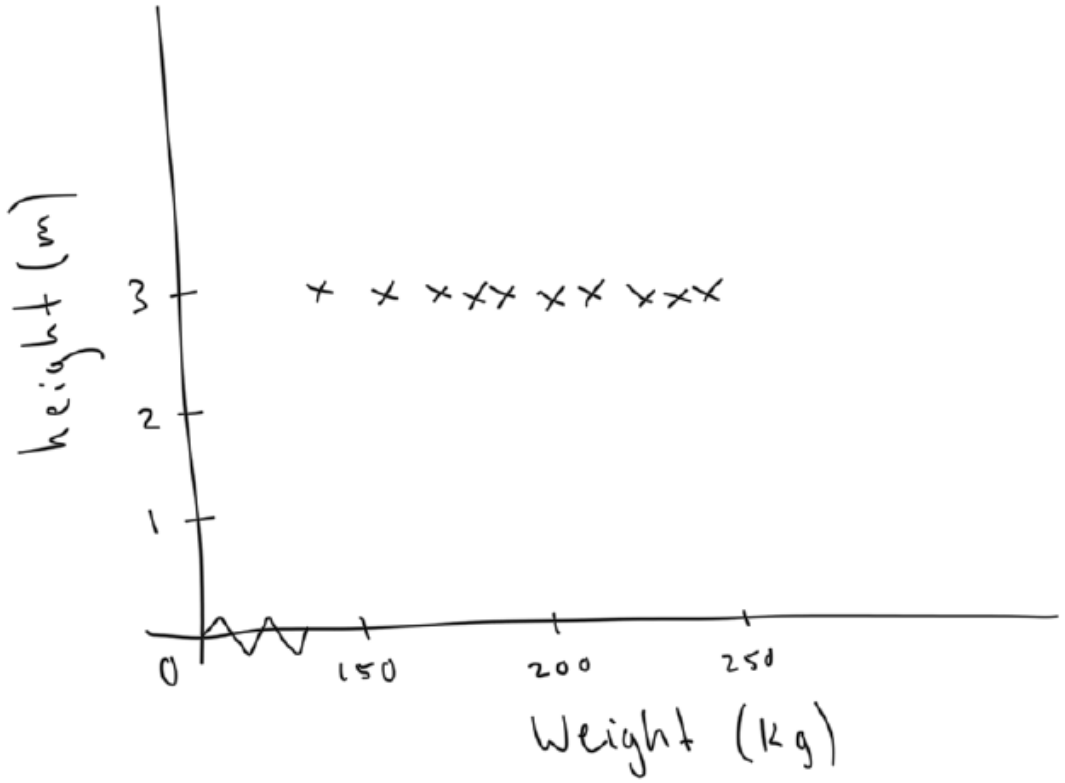
# Introducción

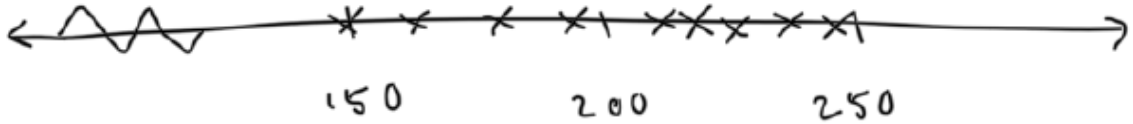
- El análisis de Componentes Principales o PCA es una de las **técnicas de aprendizaje no supervisado**, las cuales suelen aplicarse como parte del **análisis exploratorio** de los datos.

- Método estadístico que permite **simplificar la complejidad de espacios muestrales** con muchas dimensiones a la vez que conserva su información. Supóngase que existe una muestra con n individuos cada uno con p variables ($X_1$, $X_2$,..., $X_p$), es decir, el espacio muestral tiene p dimensiones. PCA permite encontrar un número de variables subyacentes ($r < p$) que explican aproximadamente lo mismo que las p variables originales.

- El método de PCA permite por lo tanto "**condensar**" la información aportada por múltiples variables en unas pocas componentes. Esto lo convierte en un método muy útil de aplicar previa utilización de otras técnicas estadísticas tales como regresión, clustering...

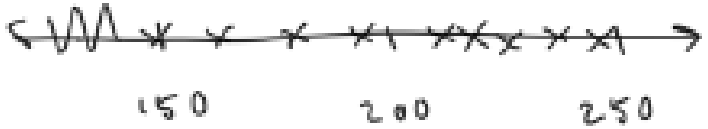
# Ejemplo




Intuitivamente, podemos ver que aquellas variables más interesantes de nuestro dataset son aquella que varían más

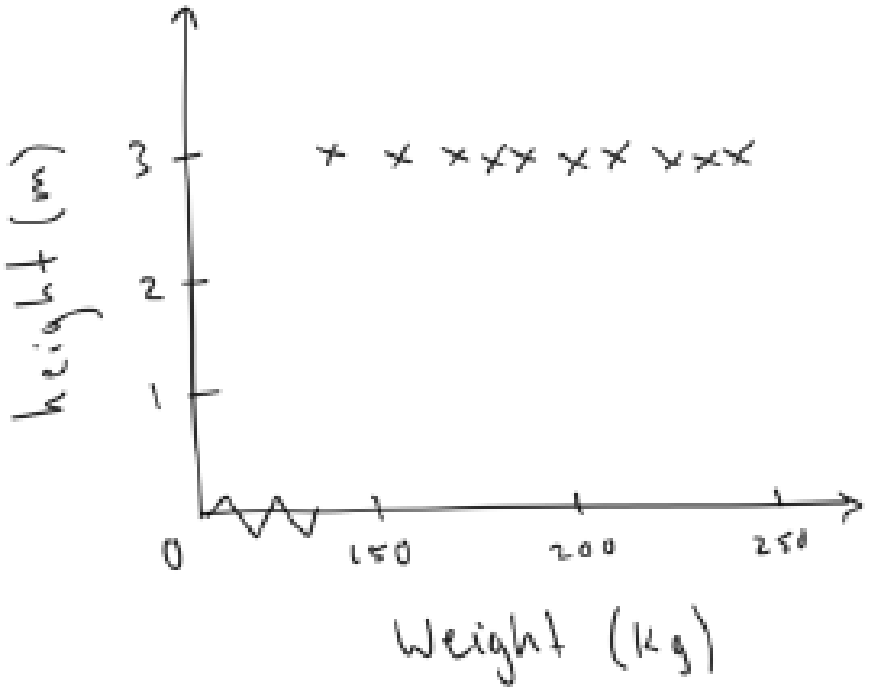
# Ejemplo




$x_2 = m x_1 + b$

$m = 0$

$b = 3$



# Interpretación geométrica

# Interpretación geométrica

Cuando trabajamos con una muestra, la **matriz de covarianzas muestral** la denotamos por $S=[S_{ij}]$, y se definirá como:

$$S = \begin{bmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & s_2^2 & \cdots & s_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_p^2 \end{bmatrix}$$

donde $S_{ij}$ representa la covarianza muestral entre las variables $X_i$ y $X_j$:

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

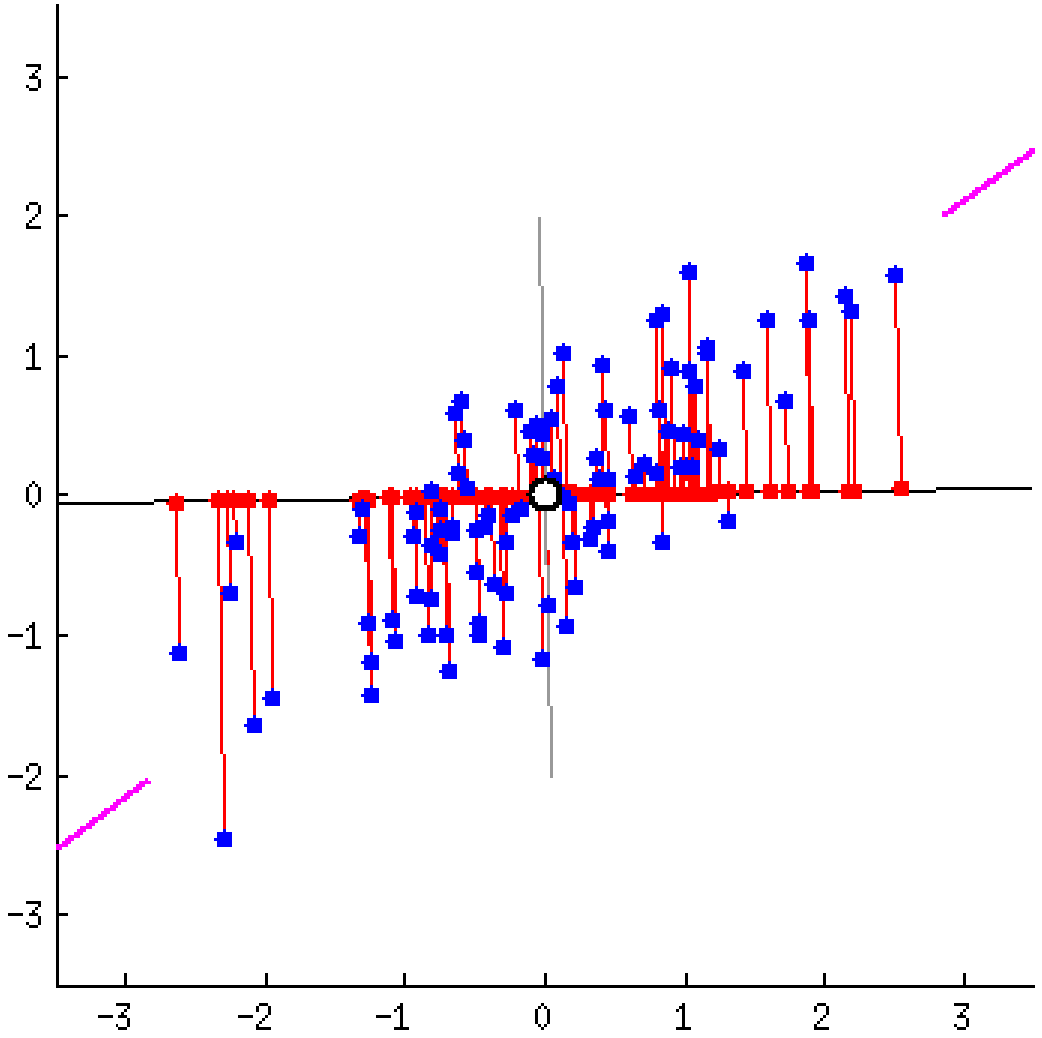
# Interpretación geométrica

• Disponemos de p-variables en n elementos. Supondremos en que previamente hemos restado a cada variable su media, de manera que las variables tienen media cero y su matriz de covarianzas vendrá dada por $\mathbf{S = 1/n\ X'X}$.

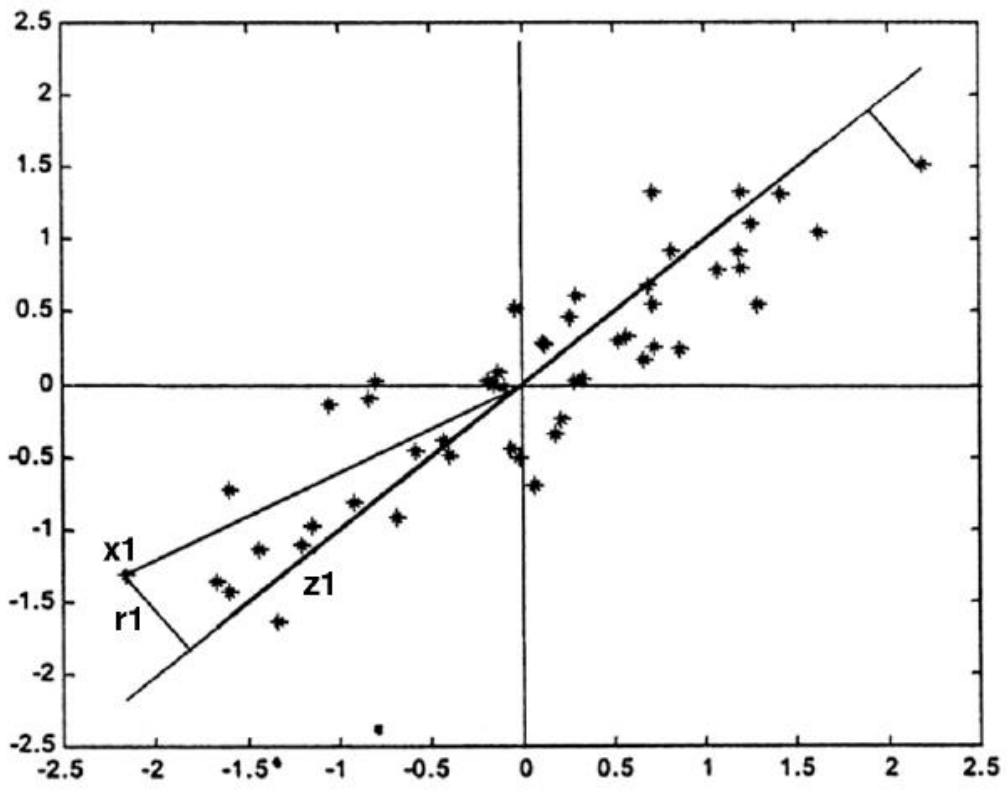
• Se desea **encontrar un subespacio de dimensión menor que p** tal que al proyectar sobre él los puntos conserven su estructura con la menor distorsión posible.

• Consideremos el caso de dos dimensiones (p = 2), y consideremos un subespacio de dimensión uno, una recta. Se desea que las proyecciones de los puntos sobre esta recta mantengan, lo más posible, sus posiciones relativas.

# Interpretación geométrica



# Interpretación geométrica




- En la figura se indica el diagrama de dispersión y una recta que, intuitivamente, proporciona un buen resumen de los datos, ya que la recta pasa cerca de todos los puntos y las distancias entre ellos se mantienen aproximadamente en su proyección sobre la recta.

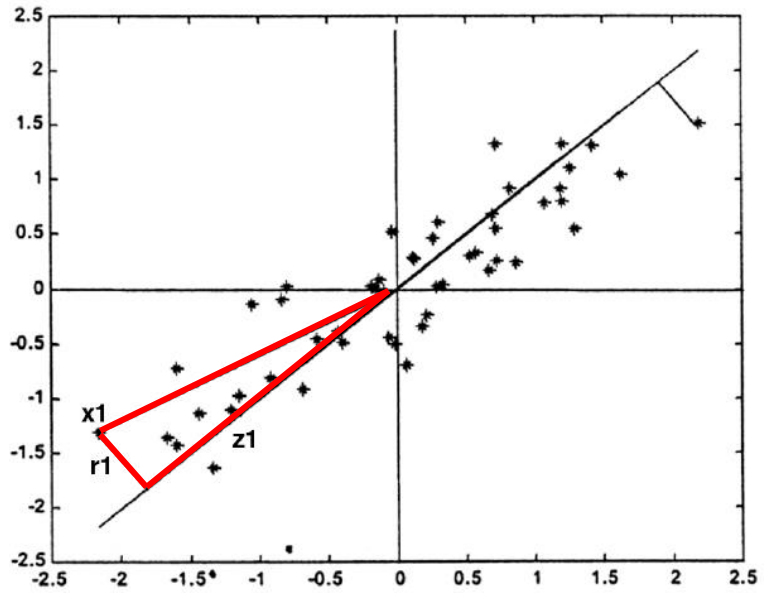
- La condición de que la recta pase cerca de la mayoría de los puntos puede concretarse exigiendo que las distancias entre los puntos originales y sus proyecciones sobre la recta sean lo más pequeñas posibles.

- Si consideramos un punto x1, la proyección del punto es el escalar:

$$z1 = a_1\, x1_1 + a_2\, x1_2$$

- En general: $Z = a_1\, X_1 + a_2\, X_2$

# Interpretación geométrica




Al proyectar cada punto sobre la recta se forma un triangulo rectángulo donde la hipotenusa es la distancia del punto al origen $(\mathbf{x}_i'\mathbf{x}_i)^{1/2}$ y los catetos la proyección del punto sobre la recta **($z_i$)** y la distancia entre el punto y su proyección ($r_i$).
Por el teorema de Pitágoras, podemos escribir:

$$\mathbf{x}_i'\mathbf{x}_i = z_i^2 + r_i^2,$$

Sumando esta expresión para todos los puntos, se obtiene:

$$\sum_{i=1}^{n} \mathbf{x}_i'\mathbf{x}_i = \sum_{i=1}^{n} z_i^2 + \sum_{i=1}^{n} r_i^2.$$

Como el primer miembro es constante, minimizar $\sum_{i=1}^{n} r_i^2$:

(la suma de las distancias a la recta de todos los puntos) es equivalente a maximizar $\sum_{i=1}^{n} z_i^2$:

Como las **proyecciones $z_i$ son variables de media cero** (las variables $x_i$ se han centrado previamente), **maximizar la suma de sus cuadrados equivale a maximizar su varianza**, y obtenemos el criterio de encontrar la dirección de proyección que maximice la varianza de los datos proyectados.

# Cálculo de las componentes

# Cálculo de las componentes

**Primera componente principal:**

La **primera componente principal** se define como la combinación lineal de las variables originales que tiene varianza máxima. Los valores en esta primera componentes de los n individuos se representan por un vector $z_1$, dado por.

$$z_1 = a_{11}x_1 + \ldots + a_{1p}x_p = X\mathrm{a}_1$$

Como las **variables originales** tienen **media cero** también $\mathbf{z_1}$ tendrá **media nula**. Su varianza será:

$$\frac{1}{n}\mathbf{z}_1'\mathbf{z}_1 = \frac{1}{n}\mathbf{a}_1'\mathbf{X}'\mathbf{X}\mathbf{a}_1 = \mathbf{a_1'}\mathbf{S}\mathbf{a_1}$$

Donde S es la matriz de varianzas y covarianzas. Puesto que la varianza de $z_1$ puede ser incrementada sin control, simplemente haciendo más grandes los coeficientes de $\mathbf{a_1}$, se impone una restricción sobre los coeficientes. En este caso:

$$a'_1\, a_1 = 1$$

# Cálculo de las componentes

Para encontrar los coeficientes que definen la primera componente principal necesitamos elegir los elementos de $a_1$ de forma que se maximice la varianza de $z_1$ sujeto a la restricción $a'_1 a_1$.

$$\underset{a_1}{Max} \quad a'_1 S a_1$$

$$s.a. \quad a'_1 a_1 = 1$$

# Cálculo de las componentes

$$\begin{aligned} \underset{a_1}{Max} \quad & a'_1 S a_1 \\ s.a. \quad & a'_1 a_1 = 1 \end{aligned}$$

Mediante el multiplicador de Lagrange: $M = \mathbf{a}_1' \mathbf{S} \mathbf{a}_1 - \lambda(\mathbf{a}_1' \mathbf{a}_1 - 1)$

Derivando respecto de $\mathbf{a_1}$ e igualando a cero:

$$\frac{\partial M}{\partial \mathbf{a}_1} = 2\mathbf{S}\mathbf{a}_1 - 2\lambda \mathbf{a}_1 = 0 \qquad \mathbf{S}\mathbf{a}_1 = \lambda \mathbf{a}_1$$

Que implica que $\mathbf{a_1}$ es un vector propio de la matriz **S** y **λ** su correspondiente valor propio:

Para determinar qué valor propio de S es la solución, multiplicando por la izquierda por $\mathbf{a'_1}$

$$\mathbf{a}_1' \mathbf{S} \mathbf{a}_1 = \lambda \mathbf{a}_1' \mathbf{a}_1 = \lambda$$

Concluimos que **λ es la varianza de** $z_1$, como ésta es la cantidad que queremos maximizar, **λ será el mayor valor propio de la matriz S. Su vector asociado $\mathbf{a_1}$ define los coeficientes de cada variable en la primera componente principal.**

# Cálculo de las componentes

La **segunda componente** se obtiene de la siguiente forma:

$$\begin{aligned} \underset{a_2}{Max} \quad & a'_2 S a_2 \\ s.a. \quad & a'_2 a_2 = 1 \\ & a'_2 a_1 = 0 \end{aligned}$$

$$M = a'_2 S a_2 - \lambda_2 (a'_2 a_2 - 1) - \alpha a'_2 a_1$$

$$\frac{\partial M}{\partial a_2} = 2Sa_2 - 2\lambda_2 a_2 - \alpha a_1 = 0$$

$$premult \ \ por \ a'_1$$

$$2a'_1 Sa_2 - 2\lambda_2 a'_1 a_2 - \alpha a'_1 a_1 = 0 + 0 - \alpha = 0$$

$$2Sa_2 = 2\lambda_2 a_2$$

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 varianzas-covarianzas, S**.

Así ocurre con todas las demás componentes: $a_i$ es el vector propio de S correspondiente al i-ésimo valor propio.

# Cálculo de las componentes

**Ejemplo:**

$$S = \begin{bmatrix} 0.35 & 0.15 & -0.19 \\ 0.15 & 0.13 & -0.03 \\ -0.19 & -0.03 & 0.16 \end{bmatrix}$$

Los valores propios son las raíces de la ecuación $|\mathbf{S} - \lambda\mathbf{I}| = 0$

$$|\mathbf{S} - \lambda\mathbf{I}| =$$

$$= \left| \begin{bmatrix} 0.35 & 0.15 & -0.19 \\ 0.15 & 0.13 & -0.03 \\ -0.19 & -0.03 & 0.16 \end{bmatrix} - \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix} \right| =$$

$$= 0.000382 - 0.0628\lambda + 0.64\lambda^2 - \lambda^3 = 0$$

$$\lambda_1 = 0.521, \quad \lambda_2 = 0.113, \; \lambda_3 = 6.51 \times 10^{-3}.$$

$$\mathbf{S}\mathbf{a}_1 = \lambda_1 \mathbf{a}_1$$

# Cálculo de las componentes

$$\begin{bmatrix} 0.35 & 0.15 & -0.19 \\ 0.15 & 0.13 & -0.03 \\ -0.19 & -0.03 & 0.16 \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{12} \\ a_{13} \end{bmatrix} = 0.521 \times \begin{bmatrix} a_{11} \\ a_{12} \\ a_{13} \end{bmatrix}$$

$$\begin{bmatrix} -0.171a_{11} + 0.15a_{12} - 0.19a_{13} \\ 0.15a_{11} - 0.391a_{12} - 0.03a_{13} \\ -0.19a_{11} - 0.03a_{12} - 0.361a_{13} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Sistema es compatible indeterminado

Para encontrar una de las infinitas soluciones tomemos la primera variable como parámetro, x, y resolvamos el sistema en función de x.
La solución es:

$$\{a_{11} = x,\ a_{12} = 0.427x,\ a_{13} = -0.562x\}$$

El valor de x se obtiene ahora imponiendo que el vector tenga norma unidad

$$\mathbf{a}_1 = \begin{bmatrix} -0.817 \\ -0.349 \\ 0.459 \end{bmatrix}$$

$$Z_1 = -0.817X_1 - 0.349X_2 + 0.459X_3$$

# Cálculo de las componentes

**Generalización**

Análogamente, **el espacio de dimensión r que mejor representa a los puntos viene definido por los vectores propios asociados a los r mayores valores propios de S**. A las nuevas variables componentes principales. En general, la matriz X (y por tanto la S) tiene rango p, existiendo entonces tantas componentes principales como variables que se obtendrán calculando los valores propios, $\lambda_1, \ldots, \lambda_p$, de la matriz de varianzas y covarianzas, S, mediante:

$$|\mathbf{S} - \lambda \mathbf{I}| = 0$$

y sus vectores asociados son: $(\mathbf{S} - \lambda_i \mathbf{I})\mathbf{a}_i = 0$

Los términos $\lambda_i$ son reales, al ser la matriz S simétrica, y positivos, ya que S es semidefinida positiva.

Por ser S simétrica si $\lambda_j$ y $\lambda_h$ son dos raíces distintas, sus vectores asociados son ortogonales.

# Cálculo de las componentes

Llamando Z a la matriz cuyas columnas son los valores de los p componentes en los n individuos, estas nuevas variables están relacionadas con las originales mediante:

$$\mathbf{Z} = \mathbf{X}\mathbf{A}$$

Donde:

$$\mathbf{A}'\mathbf{A} = \mathbf{I}$$

Calcular los componentes principales equivale a aplicar una transformación ortogonal A a las variables X (ejes originales) para obtener unas nuevas variables Z incorreladas entre sí. Esta operación puede interpretarse como elegir unos nuevos ejes coordenados.

# Cálculo de las componentes

**Ejemplo:** Matriz de varianzas y covarianzas entre nueve indicadores económicos medidos en distintas empresas

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ | $x_8$ | $x_9$ |
|---|---|---|---|---|---|---|---|---|
| 177 | 179 | 95 | 96 | 53 | 32 | $-7$ | $-4$ | $-3$ |
| | 419 | 245 | 131 | 181 | 127 | $-2$ | 1 | 4 |
| | | 302 | 60 | 109 | 142 | 4 | 0,4 | 11 |
| | | | 158 | 102 | 42 | 4 | 3 | 2 |
| | | | | 137 | 96 | 4 | 5 | 6 |
| | | | | | 128 | 2 | 2 | 8 |
| | | | | | | 34 | 31 | 33 |
| | | | | | | | 39 | 39 |
| | | | | | | | | 48 |

El rasgo más característico de esta tabla es la distinta magnitud de las seis primeras variables respecto al resto (esto, veremos que esto lo recoge la primera componente principal).

# Cálculo de las componentes

Los valores propios de esta matriz:

| Componente | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| $\lambda_i$ | 878.5 | 196.1 | 128.6 | 103.4 | 81.2 | 37.8 | 7.0 | 5.7 | 3.5 |

| Componente | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ | $x_8$ | $x_9$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.30 | 0.66 | 0.48 | 0.26 | 0.32 | 0.27 | 0.00 | 0.00 | 0.01 |
| 2 | −0.48 | −0.15 | 0.58 | −0.49 | −0.04 | 0.37 | 0.06 | 0.04 | 0.08 |
| 3 | −0.41 | −0.18 | −0.23 | 0.45 | 0.49 | 0.27 | 0.26 | 0.28 | 0.29 |

**Se observa que:**

La primera componente principal es una media ponderada de las primeras seis variables. Observamos la distinta magnitud de las seis primeras variables respecto al resto (esto se observa en la matriz S y lo recoge la primera componente principal).

La tercera componente incorpora, por un lado, las tres ultimas variables y, por otro, contrapone las tres primeras variables frente al resto.

# Propiedades de las componentes

# Propiedades de las componentes

**1.- Conservan la variabilidad inicial:** la suma de las varianzas de los componentes es igual a la suma de las varianzas de las variables originales.

$$Var\ (z_h) = \lambda_h$$

$$\sum_{i=1}^{p} Var(x_i) = \sum \lambda_i = \sum_{i=1}^{p} Var(z_i)$$

Las nuevas variables $z_i$ tienen conjuntamente la misma variabilidad que las variables originales.

# Propiedades de las componentes

**2.-** El cociente entre el valor propio $\lambda_h$ y la suma de todos ellos es la **proporción de variabilidad representada por la componente** $z_h$.

$$\lambda_h / \sum \lambda_i$$

La suma de los k primeros valores propios dividido por la suma de todos ellos es la proporción de variabilidad representada por las k primeras componentes.

$$\frac{\lambda_1 + \lambda_2 + \ldots + \lambda_k}{\sum_{i=1}^{p} \lambda_i}$$

# Propiedades de las componentes

**3.- Las covarianzas entre cada componente principal y las variables X** vienen dadas por el producto de las coordenadas del vector propio que define el componente por su valor propio:.

$$Cov(z_i; x_1, \ldots, x_p) = \lambda_i \mathrm{a}_i = \left(\lambda_i a_{i1}, \ldots, \lambda_i a_{ip}\right)$$

**4.- La correlación entre una componente principal y una variable es:**

$$Corr(z_i, x_j) = \frac{Cov(z_i, x_j)}{\sqrt{Var(z_i)Var(x_j)}} = \frac{\lambda_i a_{ij}}{\sqrt{\lambda_i s_j^2}} = a_{ij}\frac{\sqrt{\lambda_i}}{s_j}$$

# Análisis con correlaciones (análisis normado)

# Análisis normado o con correlaciones

Las componentes principales se obtienen maximizando la varianza de la proyección. La maximización dependerá decisivamente de estas escalas de medida y **las variables con valores más grandes tendrán más peso en el análisis**.

Si queremos evitar este problema, **conviene estandarizar las variables antes de calcular los componentes, de manera que las magnitudes de los valores numéricos de las variables X sean similares**, las varianzas son la unidad, y las covarianzas son los coeficientes de correlación.

**Las componentes principales normados se obtienen calculando los vectores y valores propios de la matriz R**, de coeficientes de correlación. Llamando $\lambda_p^R$ a las raíces características de esa matriz, que suponemos no singular, se verifica que:

$$\sum_{i=1}^{p} \lambda_i^R = traza(R) = p$$

# Análisis normado o con correlaciones

**Las propiedades de los componentes extraídos de R son:**

1. La proporción de variación explicada por $\lambda_i^R$: $\frac{\lambda_i^R}{p}$

2. Las correlaciones entre cada componente $z_i$ y las variables X originales: $a_i \sqrt{\lambda_i^R}$

- Cuando las variables X originales están en distintas unidades conviene aplicar el análisis de la matriz de correlaciones.
- Cuando las variables tienen las mismas unidades, ambas alternativas son posibles.

# Análisis normado o con correlaciones

**Ejemplo:** Matriz de correlaciones entre nueve indicadores económicos medidos en distintas empresas

$$\mathbf{R} = \begin{bmatrix} 1 & 0.66 & 0.41 & 0.57 & 0.34 & 0.21 & -0.09 & -0.05 & -0.03 \\ & 1 & 0.69 & 0.51 & 0.76 & 0.55 & -0.01 & 0.01 & 0.03 \\ & & 1 & 0.28 & 0.54 & 0.72 & 0.04 & 0.00 & 0.09 \\ & & & 1 & 0.69 & 0.30 & 0.05 & 0.03 & 0.02 \\ & & & & 1 & 0.73 & 0.06 & 0.07 & 0.07 \\ & & & & & 1 & 0.03 & 0.03 & 0.10 \\ & & & & & & 1 & 0.85 & 0.82 \\ & & & & & & & 1 & 0.90 \\ & & & & & & & & 1 \end{bmatrix}$$

# Cálculo de las componentes

Los valores propios de esta matriz:

| $\lambda_i$ | 3.70 | 2.72 | 1.06 | 0.70 | 0.30 | 0.23 | 0.16 | 0.09 | 0.03 |
|---|---|---|---|---|---|---|---|---|---|

Los vectores propios asociados a los tres primeros valores propios son:

| $\lambda$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ | $x_8$ | $x_9$ |
|---|---|---|---|---|---|---|---|---|---|
| 3.7 | 0.34 | 0.46 | 0.41 | 0.36 | 0.46 | 0.40 | 0.06 | 0.06 | 0.08 |
| 2.72 | −0.11 | −0.07 | −0.03 | −0.04 | −0.02 | −0.01 | 0.56 | 0.58 | 0.57 |
| 1.06 | −0.54 | −0.05 | 0.38 | −0.52 | 0.07 | 0.53 | −0.04 | −0.07 | 0.00 |

**Si comparamos estos resultados con los anteriores:**

El primer vector propio cambia apreciablemente. Con la matriz de varianzas las variables con más peso en la componente eran las que tenían una mayor varianza: la 2, luego la 3 y finalmente las 1, 4, 5 y 6 con un peso parecido. Sin embargo, al utilizar la matriz de correlaciones el peso de las variables esta más relacionado con las correlaciones.
La proporción de variabilidad explicada por la primera componente cambia mucho: de 878.5/1441.8 (60.9%) a 3.7/9 (41%).

La segunda componente cambia completamente: ahora está prácticamente asociado a las tres últimas variables. La proporción de variabilidad que explica ha aumentado considerablemente, del 196/1441.8 (13.6%) a 2.72/9 (30%).

El tercer vector propio es también distinto en ambas matrices.

# Selección del número de componentes

# Selección del número de componentes

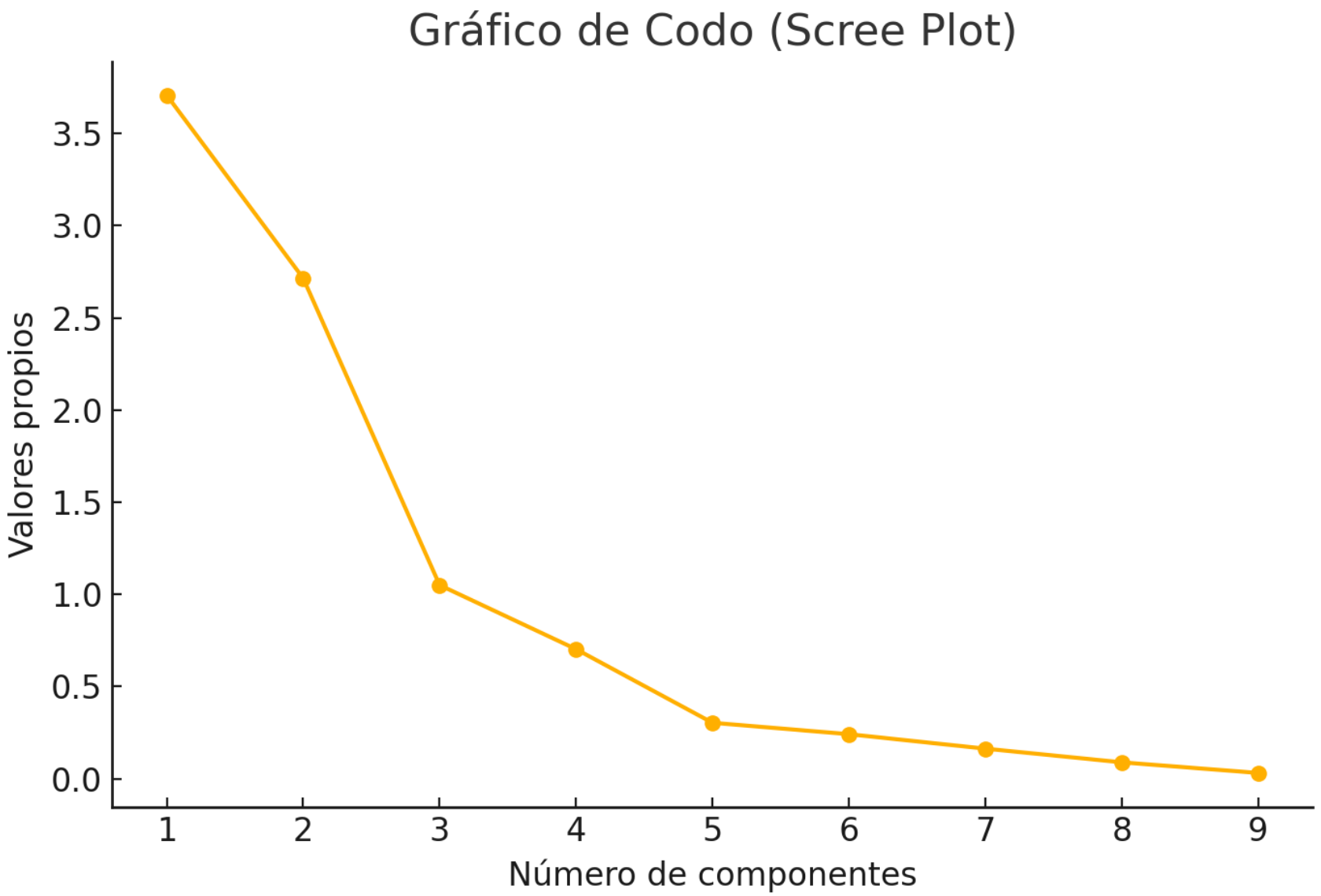
Existen distintas reglas para seleccionar el número de componentes:

1. Realizar un grafico de $\lambda_i$ frente a i. Comenzar seleccionando componentes hasta que los restantes tengan aproximadamente el mismo valor de i. La idea es buscar un "codo" en el gráfico, es decir, un punto a partir del cual los valores propios son aproximadamente iguales. El criterio es quedarse con un número de componentes que excluya los asociados a valores pequeños y aproximadamente del mismo tamaño.

2. Seleccionar componentes hasta cubrir una proporción determinada de varianza (80 o el 90 por 100). Esta regla es arbitraria y debe aplicarse con cierto cuidado. Por ejemplo, es posible que un único componente recoja el 90 por 100 de la variabilidad y, sin embargo, pueden existir otros componentes que sean muy adecuados para explicar otros aspectos de las variables.

3. Desechar aquellos componentes asociados a valores propios inferiores a una cota, que suele fijarse como la varianza media $\sum \lambda_i / p$. Cuando se trabaja con la matriz de correlación, el valor medio de los componentes es 1, y esta regla lleva a seleccionar los valores propios mayores que la unidad. Esta regla es arbitraria: una variable que sea independiente del resto suele llevarse un componente principal y puede tener un valor propio mayor que la unidad. Sin embargo, si está incorrelada con el resto puede ser una variable poco relevante para el análisis, y no aportar mucho a la comprensión del fenómeno global.

# Selección del número de componentes



# Selección del número de componentes

2. Seleccionar componentes hasta cubrir un 80% de la varianza

| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
|---|---|---|---|---|---|---|---|---|---|
| $\lambda_i$ | 3.70 | 2.72 | 1.06 | 0.70 | 0.30 | 0.23 | 0.16 | 0.09 | 0.03 |
| **Varianza explicada (%)** | 41.19 | 30.16 | 11.66 | 7.80 | 3.36 | 2.67 | 1.81 | 0.98 | 0.34 |
| **Varianza acumulada (%)** | 41.19 | 71.36 | 83.02 | 90.82 | 94.19 | 96.65 | 98.67 | 99.65 | 100 |

3. Desechar aquellos componentes asociados a valores propios inferiores a 1

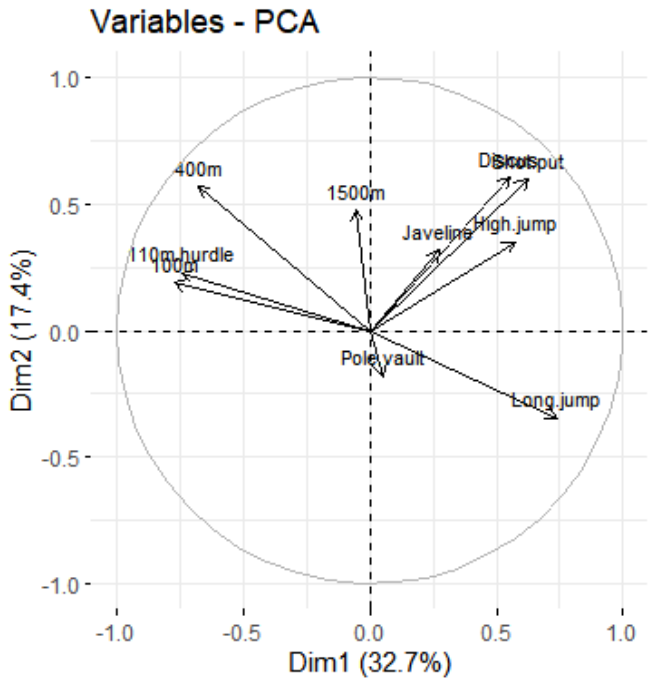
| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
|---|---|---|---|---|---|---|---|---|---|
| $\lambda_i$ | 3.70 | 2.72 | 1.06 | 0.70 | 0.30 | 0.23 | 0.16 | 0.09 | 0.03 |
| **Varianza explicada (%)** | 41.19 | 30.16 | 11.66 | 7.80 | 3.36 | 2.67 | 1.81 | 0.98 | 0.34 |
| **Varianza acumulada (%)** | 41.19 | 71.36 | 83.02 | 90.82 | 94.19 | 96.65 | 98.67 | 99.65 | 100 |

# Representación gráfica

# Interpretación gráfica

**Representación de las variables:**

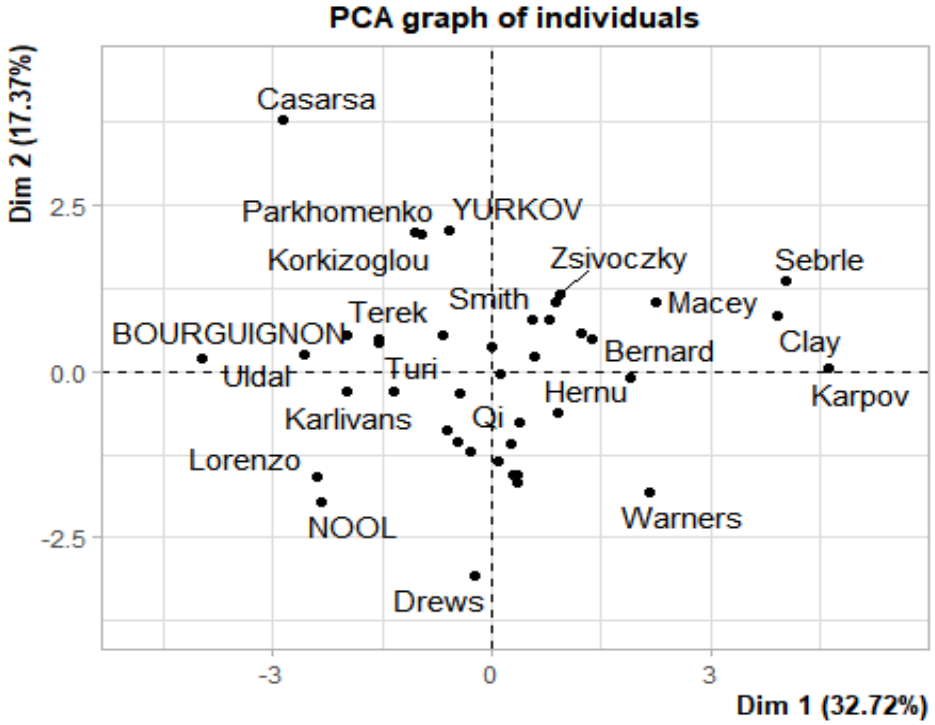
Representar las variables sobre las componentes principales. La correlación entre una variable y una componente principal se utiliza como la coordenada de dicha variable sobre la componente principal. De esta manera podemos obtener un gráfico de correlación de variables. Las variables positivamente correlacionadas se agrupan juntas o próximas, mientras que las negativamente correlacionadas se representan en lados opuestos del origen o cuadrantes opuestos.



# Interpretación gráfica

**Proyecciones de las observaciones:**

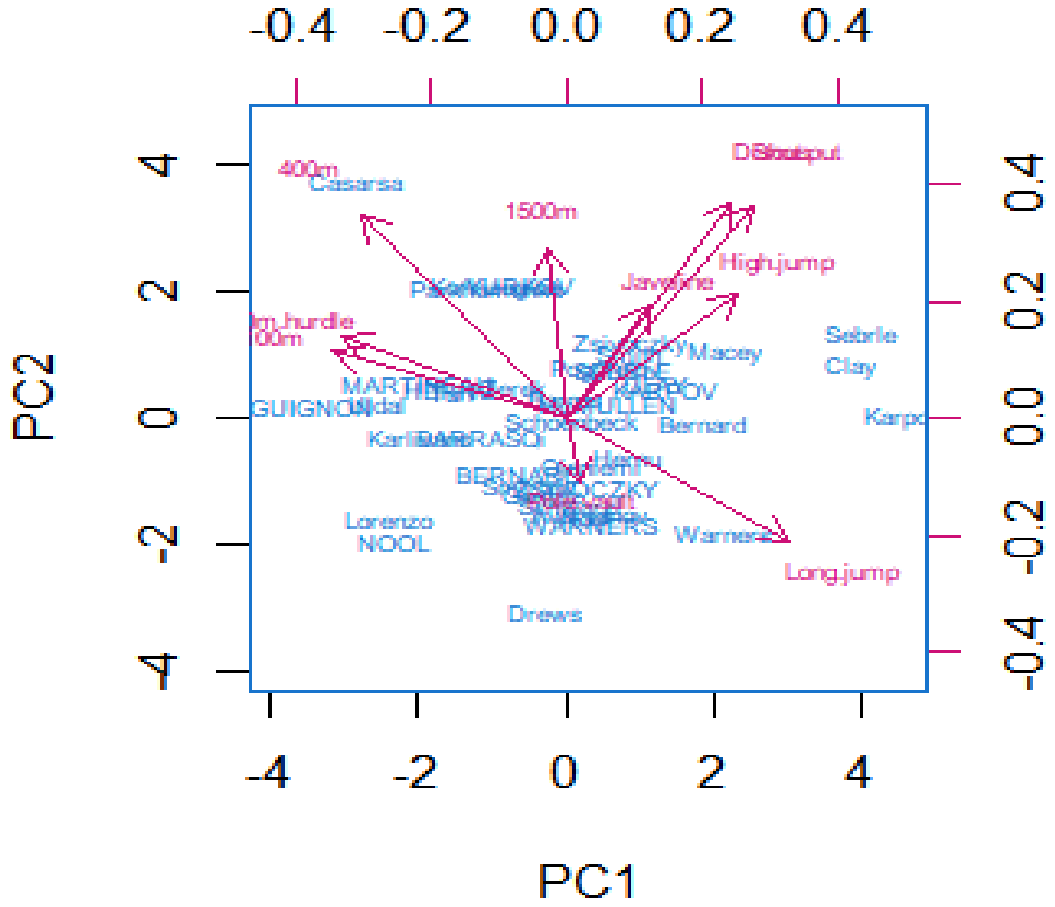
La interpretación de los componentes principales se favorece representando las proyecciones de las observaciones sobre un espacio de dimensión dos, definido por parejas de los componentes principales más importantes. La representación habitual es tomar dos ejes ortogonales que representen dos de las componentes consideradas, y situar cada punto sobre ese plano por sus coordenadas con relación a estos ejes.



# Interpretación gráfica

**Representación sujetos y variables:**

La interpretación se favorece representando, además de las observaciones, las variables originales.



# Pruebas de adecuación de las Componentes Principales

# Prueba de esfericidad de Barlett

La prueba de esfericidad de Bartlett contrasta si la matriz de correlaciones es una matriz identidad. El estadístico de Bartlett se obtiene a partir de una transformación del determinante de la matriz de correlaciones y cuanto mayor sea, más improbable es que la matriz sea una matriz identidad.

$$H_0 : S = I$$
$$H_1 : S \neq I$$

$$\chi^2_{\left(0.5(p^2-p)\right)} = -n\left[n-1-\frac{1}{6}\left(2p+5\right)\right]\ln\left|R\right|$$

# KMO

La medida de la adecuación muestral de **Kaiser-Meyer-Olkin** (*Coeficiente KMO*) contrasta si las correlaciones parciales entre las variables son pequeñas. Toma valores entre 0 y 1, e indica que el análisis de componentes principales es más adecuado cuanto mayor sea su valor.

Para:

KMO ≥ 0.9, el test es muy bueno

KMO ≥ 0.8: Notable

KMO ≥ 0.7: Mediano

KMO ≥ 0.6: Bajo

KMO < 0.5: Muy bajo

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

$R = \left[ r_{ij} \right]$ Matriz de correlaciones

$A = \left[ a_{ij} \right]$ Matriz de correlaciones parciales

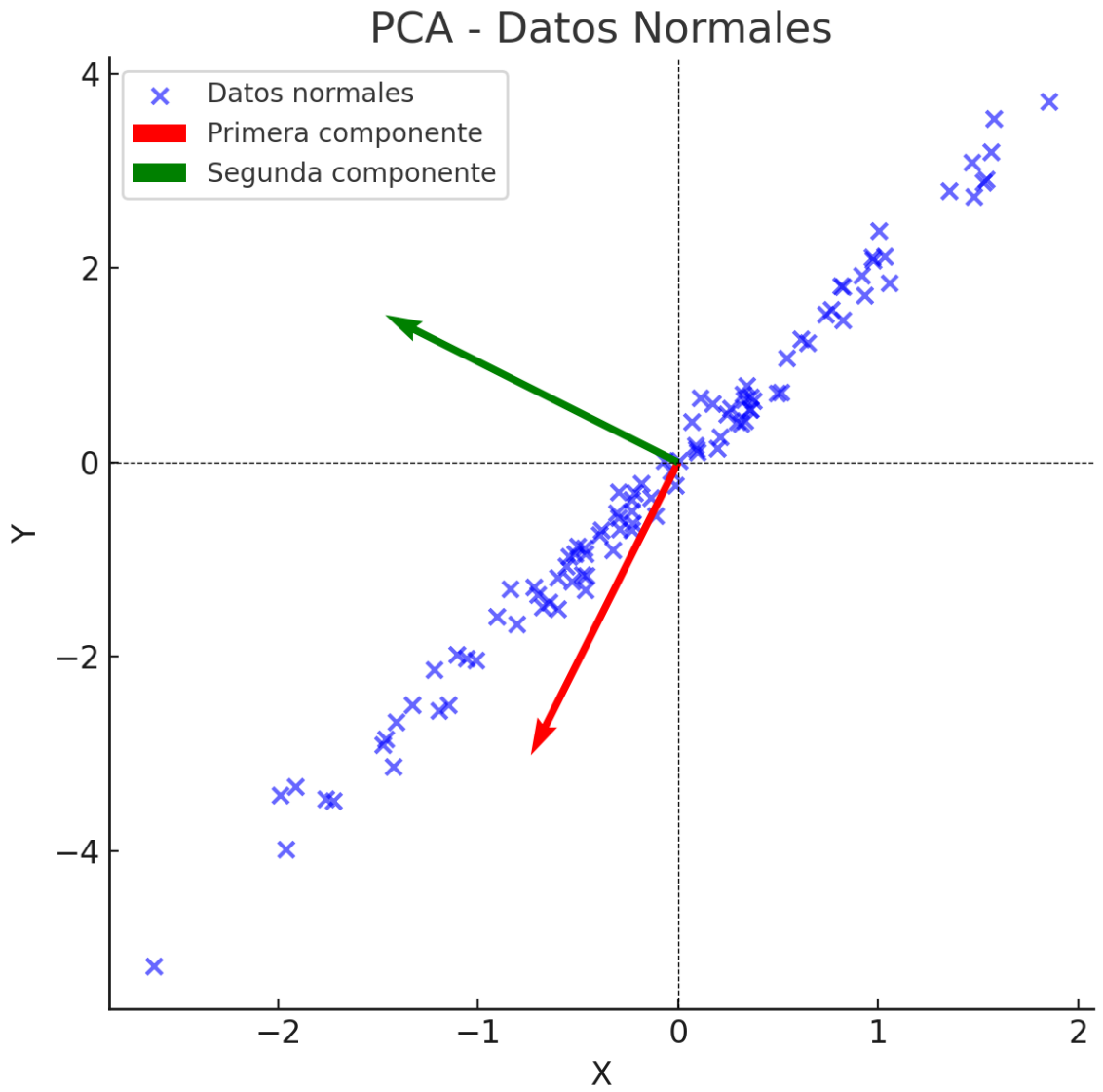
# Datos atípicos

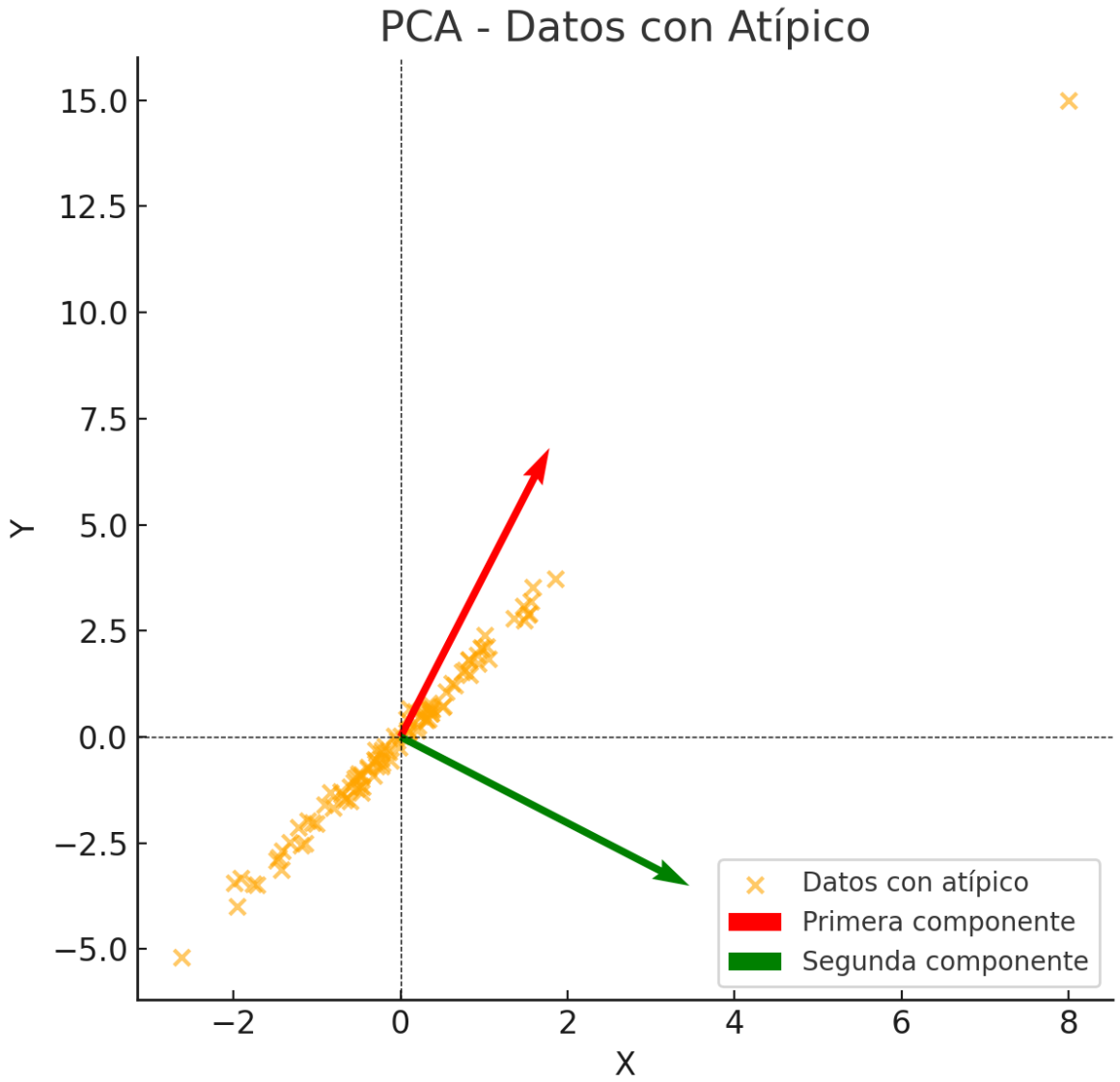
# Datos atípicos

Antes de obtener los componentes principales conviene asegurarse de que no existen datos atípicos, ya que, los atípicos pueden distorsionar totalmente la matriz de covarianzas.

Las componentes principales podrían utilizarse para detectar datos atípicos multivariantes, ya que un valor muy extremo se llevará un componente principal y aparecerá como extremo sobre esta componente. Desgraciadamente, aunque los componentes pueden identificar atípicos aislados, no hay garantía de que funcionen cuando existen grupos de atípicos, debido al problema de enmascaramiento.

# Datos atípicos



# Bibliografía

# Bibliografía


- Aldás Manzano, J., & Uriel Jimenez, E. (2017). Análisis multivariante aplicado con R. Ediciones Paraninfo, SA
- Hair, Anderson, Tatham, Black. (2001). Análisis Multivariante.
- Husson, F., Lê, S., & Pagès, J. (2011). Exploratory multivariate analysis by example using R (Vol. 15). Boca Raton: CRC press.
- Peña, D. (2002). Análisis de datos multivariantes. McGraw-Hill.