Minería de datos

Sesión 2: Análisis de componentes principales

Paloma Botella Rocamora
Paloma.Botella@gmail.com

Estructura de la sesión.

Estructura de la sesión.

- 1. Anexo metodológico previo (algunas herramientas matemáticas que vamos a utilizar)
  - 1.a. Multiplicadores de Lagrange
  - 1.b. Descomposición de una matriz en valores y vectores propios
- 2. Análisis de componentes principales (ACP)
  - 2.a. Motivación
  - 2.b. Desarrollo teórico
  - 2.c. Ejemplos y ACP con R.

1. Anexo metodológico previo.

Anexo metodológico previo

- **Multiplicadores de Lagrange**
- **Descomposición** de una matriz en **valores** y **vectores propios**

Multiplicadores de Lagrange

- Suele ser habitual encontrarnos con la **necesidad** de **maximizar** o **minimizar funciones**. Para ello solemos hacer:

$$f'(x) = 0$$

- Sin embargo, **en ocasiones** querremos **maximizar una función** pero no para cualquier valor de la coordenada x, **sino para ciertos valores** concretos que cumplan una (o varias) **restricción(es)**.

El **problema que nos planteamos** ahora es:

- Maximizar (o minimizar) la función $f(x)$, donde x puede ser un vector de valores.
- Sujeto a que x ha de cumplir las siguientes condiciones: $g_1(x) = k_1, \ldots, g_r(x) = k_r$.

Por ejemplo:

Maximizar $f(x, y, z) = xyz$ *sujeto a* $g(x, y, z) = x + y + z = 1$

Multiplicadores de Lagrange

Resolución:

- Consideramos la siguiente función:

$$H(x,\lambda) = f(x) - \lambda_1(g_1(x) - k_1) - ... - \lambda_r(g_r(x) - k_r)$$

- La solución al problema de maximización restringido que hemos planteado habrá de cumplir necesariamente el siguiente sistema de ecuaciones:

$$\begin{gathered} \frac{\partial H}{\partial x} = 0 \\ \frac{\partial H}{\partial \lambda_1} = 0 \\ ... \\ \frac{\partial H}{\partial \lambda_r} = 0 \end{gathered}$$

De esta forma el método de los multiplicadores de Lagrange transforma el problema de maximización restringida que teníamos anteriormente en la resolución de un sistema de ecuaciones.

Multiplicadores de Lagrange (III)

Ejemplo: Maximizar $f(x, y, z) = xyz$ sujeto a $g(x, y, z) = x + y + z = 1$

Considremos la funcion $H(x, y, z) = xyz - \lambda(x + y + z - 1)$

- $\frac{\partial H}{\partial x} = yz - \lambda = 0 \Longrightarrow \lambda = yz$
- $\frac{\partial H}{\partial y} = xz - \lambda = 0 \Longrightarrow \lambda = xz$
- $\frac{\partial H}{\partial z} = xy - \lambda = 0 \Longrightarrow \lambda = xy$
- $\frac{\partial H}{\partial \lambda} = x + y + z - 1 = 0 \Longrightarrow x + y + z = 1$

Resulta $x = y = z = \frac{1}{3}$

Descomposición de una matriz en vectores y valores propios.

- Sea A una *matriz cuadrada* de dimensión k.
- v se dice vector propio de A si existe un **escalar** λ tal que :

$$Av = \lambda v$$

- A λ se le conoce como **valor propio** de A **asociado** al vector propio v.

Descomposición de una matriz en vectores y valores propios.

- Es decir, los **vectores propios** son aquellos vectores que multiplicados por una matriz obtienen como resultado el mismo vector o un múltiplo del mismo.
- Todos los vectores propios de una matriz son ortogonales (*perpendiculares*) entre ellos.
- Como la dirección de un vector no cambia al multiplicarlo por un escalar, se suelen escalar a longitud 1.

$$\begin{pmatrix} 2 \\ 1 \end{pmatrix} \rightarrow longitud = \sqrt{(2^2 + 1^2)} = \sqrt{5}$$

$$\begin{pmatrix} 2/\sqrt{5} \\ 1/\sqrt{5} \end{pmatrix} \rightarrow longitud = 1$$

Descomposición de una matriz en vectores y valores propios.

- Por tanto, los **vectores propios de A** cumplen que son todos **ortonormales** entre sí, es decir:

$$i \neq j \implies v_i v_j = 0$$

$$v_i v_i = 1$$

- Está demostrado que para toda matriz A **cuadrada y simétrica** de dimensión k existen k pares $(\lambda_1, v_1), ..., (\lambda_k, v_k)$ de **valores y vectores propios**, respectivamente.

Descomposición de una matriz en vectores y valores propios.

- En general tenemos:

$$Av_1 = \lambda_1 v_1 \ldots Av_k = \lambda_k v_k$$

- Así, podemos expresar matricialmente

$$AV = VD$$

donde $V = (v_1, ..., v_k)$ y $D = diag(\lambda_1, ..., \lambda_k)$.

Descomposición de una matriz en vectores y valores propios.

- Y como V está formado por vectores ortonormales, tenemos:

$$A = AI = A(VV^{-1}) = (AV)V^{-1} = VDV^{-1} = VDV'$$

A esta **expresión** se le conoce como **descomposición en valores y vectores propios** de A.

- Además, si A es **definida positiva** sus valores propios serán positivos. Es decir, **toda matriz de varianzas-covarianzas** se puede descomponer como la expresión anterior con todos los elementos de D positivos.

Descomposición de una matriz en vectores y valores propios.

Cálculo de la descomposición en valores y vectores propios de una matriz:

- Resolver la **ecuación característica de la matriz** para obtener el conjunto de valores propios de la matriz:

$$f(\lambda) = |A - \lambda I| = 0 \Longrightarrow \lambda_1, ..., \lambda_r$$

- Para cada uno de los valores propios determinados anteriormente resolver:

$$Av_i = \lambda_i v_i$$

en función de v_i para determinar el vector propio correspondiente a cada valor propio λ_i.

2. Análisis de componentes principales

Análisis de componentes principales: Motivación

Consideramos el siguiente banco de datos, llamado *MundoDes* sobre indicadores de desarrollo de distintos paises (91 registros):

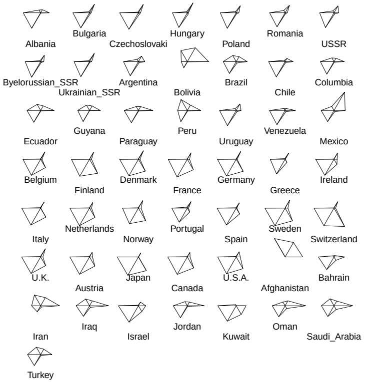
```
##                Tasa Nat Tasa Mort Mort Inf Esp Hom Esp Muj  PNB
## Albania            24.7       5.7     30.8    69.6    75.5  600
## Bulgaria           12.5      11.9     14.4    68.3    74.7 2250
## Czechoslovaki      13.4      11.7     11.3    71.8    77.7 2980
## Hungary            11.6      13.4     14.8    65.4    73.8 2780
## Poland             14.3      10.2     16.0    67.2    75.7 1690
## Romania            13.6      10.7     26.9    66.5    72.4 1640
##                Continente
## Albania                 1
## Bulgaria                1
## Czechoslovaki           1
## Hungary                 1
## Poland                  1
## Romania                 1
```

Este banco de datos contiene las variables cuantitativas:

- **Tasa Nat**: Tasa de Natalidad por cada 1000 habitantes
- **Tasa Mort**: Tasa de Mortalidad por cada 1000 habitantes
- **Mort Inf**: Mortalidad en niños (menores de 1 año) por cada 1000 nacimientos
- **Esp Hom**: Esperanza de vida al nacer en hombres
- **Esp Muj**: Esperanza de vida al nacer en mujeres
- **PNB**: Producto Nacional Bruto

Análisis exploratorio de los datos

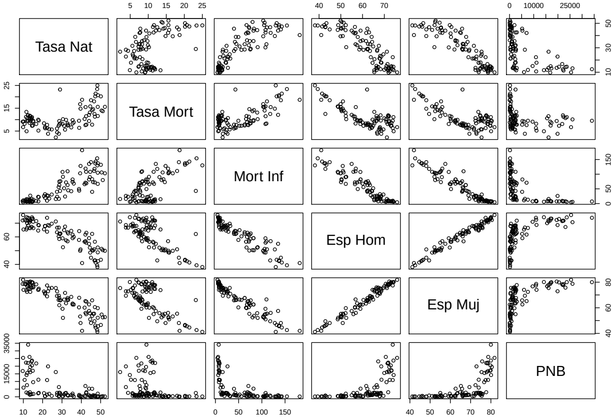
```
stars(MundoDes[1:50, -7])
```



Este tipo de gráficas nos ayudan a describir los individuos (con 91 todavía es asequible explorar los tipos de comportamiento)

Análisis exploratorio de los datos

```
pairs(MundoDes[, -7])
```



Con este tipo de gráficos podemos explorar la relación entre las variables.

Análisis exploratorio de los datos

```
round(cor(MundoDes[, -7]), 2)
```

```
##           Tasa Nat Tasa Mort Mort Inf Esp Hom Esp Muj   PNB
## Tasa Nat      1.00      0.51     0.86   -0.87   -0.89 -0.63
## Tasa Mort     0.51      1.00     0.68   -0.75   -0.71 -0.30
## Mort Inf      0.86      0.68     1.00   -0.94   -0.95 -0.60
## Esp Hom      -0.87     -0.75    -0.94    1.00    0.98  0.64
## Esp Muj      -0.89     -0.71    -0.95    0.98    1.00  0.65
## PNB          -0.63     -0.30    -0.60    0.64    0.65  1.00
```

Pero ¿y cuando dispongamos de 20, o de 50, o de 100 variables. . . ?

Análisis de componentes principales: Motivación

- **Excesivo número de variables a analizar:**
  - Cuando se recoge información de una muestra de datos, se **intenta recoger el mayor número de variables** posible.
  - Por ejemplo, con **20 variables** tendríamos que considerar **190 posibles coeficientes de correlación**. Con **40 variables** tendríamos **780 posibles coeficientes de correlación**.
- **Fuerte correlación entre variables**
  - En la mayoría de ocasiones se presenta una **fuerte correlación entre variables**, algunas de ellas están relacionadas o miden lo mismo bajo distintos puntos de vista (*Por ejemplo: en estudios médicos, han podido medir la presión sanguínea a la salida del corazón y a la salida de los pulmones y ambas están fuertemente relacionadas).*

Análisis de componentes principales: Motivación

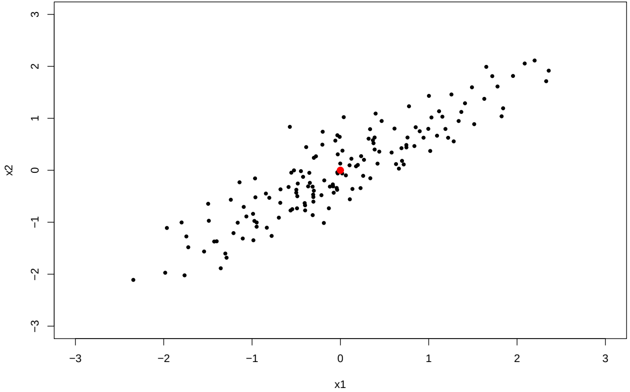
¿Qué podemos hacer?

- **Reducir el número de variables** con la **menor pérdida de de información**
- El concepto de **mayor información** se relaciona con el de **mayor** variabilidad o **varianza**.
- Cuanto mayor sea la variabilidad de los datos (varianza) se considera que existe mayor información.

Análisis de componentes principales: Motivación

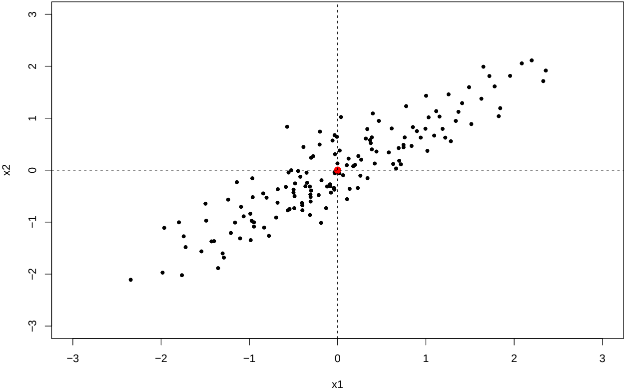
- Además, también resulta de gran utilidad disponer de un **método de reducción de variables** para:
  - Poder **resumir en unas pocas variables las características principales** de los individuos de la muestra.
  - **Conocer qué combinación de variables de las variables originales resumen mejor la información** de las variables que componen el banco de datos .
- El disponer de una técnica de reducción de variables que resuma de forma eficaz la información de una banco de datos motiva el **Análisis de componentes principales (ACP)**.

Supongamos el siguiente banco de datos con dos variables, x_1 y x_2:



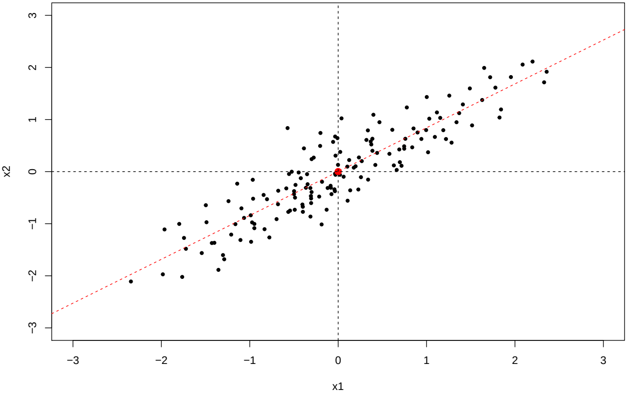
Análisis de componentes principales: Motivación

Nuestros ejes de referencia son:



Análisis de componentes principales: Motivación

Pensemos en construir una variable *ficticia* que combine x_1 y x_2:



Esta variable recogería más información que la que recogen x_1 o x_2.

Análisis de componentes principales: Notación

- Denotaremos la matriz de datos como X.
- Consideraremos que dicha matriz de datos consta de **n individuos** y **p variables**.
- Podemos ver la matriz X como la matrix formada por las p columnas:

$$X = [x_1, x_2, ..., x_p]$$

es decir, $x_1, x_2, ..., x_p$ denotan las **columnas (variables)** de la matriz X.

Análisis de componentes principales: Notación

- Sin pérdida de generalidad, en adelante supondremos que $x_1, x_2, ..., x_p$ tienen *media 0*, es decir, consideraremos las variables centradas en 0.
- La matriz de datos X tendrá asociada una matriz de varianzas-covarianzas Σ (que se podrá calcular como $\Sigma = \frac{1}{n}X'X$, puesto que suponemos media 0 para todas las variables).
- Es deseable que el banco de datos tenga un comportamiento *Normal* (aunque esta técnica es bastante robusta frente a la violación de este supuesto cuando el objetivo es la reducción de la dimensionalidad)

Análisis de componentes principales: Objetivos

- Buscamos transformar los datos originales $X_{n \times p}$ en una matriz $Y_{n \times m}$ con $m < p$, de columnas (**variables**) **incorreladas** que resumen la máxima información original de X.
- Las nuevas variables (columnas de Y) se definen de tal forma que $var(y_1) > var(y_2) > ... > var(y_m)$. Así, y_1 será la variable que **mejor resuma** (explique en mayor medida) X, seguida de y_2,y_3,. . .
- Cada una de las columnas de Y será necesariamente combinación lineal de las columnas de X, es decir, cada **nueva variable** será **combinación lineal** de las **variables originales**.

Análisis de componentes prncipales: cómo se construyen

- Como han de ser combinaciones lineales de variables originales, han de ser de la forma:

$$y_1 = Xa_1, \;\; y_2 = Xa_2, ..., \;\; y_m = Xa_m$$

para ciertos **vectores** $a_1, a_2, ..., a_m$ que habremos de determinar.

Es decir, la nueva variable y_j será de la forma:

$$y_j = Xa_j = a_j^1 x_1 + a_j^2 x_2 + ... + a_j^p x_p$$

donde $a_j = (a_j^1, \ldots, a_j^p)$.

Análisis de componentes principales: construcción

- Como queremos que la **primera componente** y_1, tenga la **máxima varianza**, debemos **restringir los valores de los vectores** $a_1, a_2, ..., a_m$ puesto que podríamos aumentar la varianza de $y_1, y_2, ..., y_m$ aumentando la magnitud de los vectores $a_1, a_2, ..., a_m$.
- Así, suele ser habitual imponer la siguiente restricción a estos vectores:

$$a_j' a_j = \sum_{i=1}^{p} a_{ij}^2 = 1$$

Análisis de componentes principales: Extracción de la primera componente principal

Formulación

- Tenemos como objetivo hallar la combinación lineal Xa_1 de las variables originales de máxima varianza tal que a_1 cumple $a_1' a_1 = 1$.
- La varianza que queremos maximizar sería $Var(Xa_1) = \frac{1}{n}(Xa_1)'(Xa_1)$.
- Matemáticamente este problema se traduce a calcular el vector a_1 que maximice la expresión:

$$\frac{1}{n}(Xa_1)'(Xa_1) = a_1'(\frac{1}{n}X'X)a_1 = a_1'\Sigma a_1$$

sujeto a la **restricción**: $a_1' a_1 = 1$.

Análisis de componentes principales: Extracción de la primera componente principal

Formulación

- Es decir, queremos calcular a_1 que **maximice** la expresión $a_1'\Sigma a_1$ sujeto a la **restricción**: $a_1'a_1 = 1$.
- La maximización de funciones sujeta a restricciones se suele resolver matemáticamente mediante el método de los **multiplicadores de Lagrange**.

Análisis de componentes principales: Extracción de la primera componente principal

Resolución

- Función objetivo : $L(a_1, \lambda) = a_1' \Sigma a_1 - \lambda(a_1' a_1 - 1)$

$$\frac{\partial L}{\partial a_1} = 2\Sigma a_1 - 2\lambda a_1 = 0 \Longrightarrow \Sigma a_1 = \lambda a_1$$
$$\frac{\partial L}{\partial \lambda} = a_1' a_1 - 1 = 0 \Longrightarrow a_1' a_1 = 1$$

- Por tanto, a_1 será necesariamente un vector propio de Σ asociado al valor propio λ.
- Pero además: $var(Xa1) = a_1' \Sigma a_1 = a_1'(\lambda a_1) = \lambda$, que queremos que tome al **mayor valor posible**.

Análisis de componentes principales: Extracción de la primera componente principal

Resolución

- Por tanto, el vector que nos da la primera variable o componente principal es a_1, que se corresponde con el **vector propio** de Σ asociado a su **mayor valor propio**.
- Primera nueva variable o componente principal se calcula como: $y_1 = Xa_1$ y tendrá varianza λ_1.

Análisis de componentes principales: Extracción de la segunda componente principal.

Formulación

- Una vez hemos determinado la combinación lineal de las variables originales que recoge mayor varianza, nos podemos plantear extraer una segunda componente principal que explique información que no explicaba la componente anterior.
- La nueva componente $y_2 = Xa_2$ será aquella combinación lineal de las variables originales que explique una mayor varianza y sujeta a las restricciones: $a_2' a_2 = 1$ y $Cov(y2, y1) = 0$ (*Si no imponemos esta restricción obtendríamos simplemente* $a_2 = a_1$).

Análisis de componentes principales: Extracción de la segunda componente principal.

Formulación

- La segunda restricción se puede expresar como:

$$Cov(y_2, y_1) = Cov(Xa_2, Xa_1) = a_2' \Sigma a_1 = a_2' \lambda a_1 = 0 \Longrightarrow a_2' a_1 = 0$$

- De nuevo tenemos un problema de maximización sujeto en esta ocasión a dos restricciones, por lo que podemos resolverlo mediante el método de *multiplicadores de Lagrange.*
- Queremos calcular a_2 que **maximice** la expresión $a_2' \Sigma a_2$ (para maximizar la varianza de y_2) sujeto a la **restricciones** $a_2' a_2 = 1$ y $a_2' a_1 = 0$.

Análisis de componentes principales: Extracción de la segunda componente principal.

Resolución

- Función objetivo:
$L(a_2, \lambda_2, \delta) = a_2' \Sigma a_2 - \lambda_2(a_2' a_2 - 1) - \delta(a_2' a_1)$

$$\frac{\partial L}{\partial a_2} = 2\Sigma a_2 - 2\lambda_2 a_2 - \delta a_1 = 0$$

- Si multiplicamos la ecuación anterior por a_1':

$$2a_1'\Sigma a_2 - 2\lambda_2 a_1' a_2 - \delta a_1' a_1 = 2a_1'\Sigma a_2 - \delta = 0 \Longrightarrow$$
$$2a_1'\Sigma a_2 - \delta = 2(\Sigma a_1)' a_2 - \delta = 2\lambda_2 a_1' a_2 - \delta = 0 \Longrightarrow \delta = 0$$

- Así, la derivada anterior resulta: $\Sigma a_2 = \lambda_2 a_2$
- Es decir, nuevamente a_2 será un vector propio de la matriz Σ asociado al vector propio λ_2.

Análisis de componentes principales: Extracción de la segunda componente principal.

Resolución (II)

- λ_2 coincide con la varianza de $y_2 = Xa_2$, por tanto habrá de ser tan grande como sea posible.
- Por tanto, λ_2 será el segundo mayor valor propio de Σ y a_2 su vector propio asociado. De esta forma se cumplen las dos restricciones que habíamos impuesto:

$$a_2' a_2 = 1$$
$$a_1' a_2 = 0$$

- Procediendo de manera similar podremos extraer todas las componentes principales que deseemos del banco de datos (p a lo sumo).

Análisis de componentes principales: Proceso de extracción de las componentes principales.

Resumen del proceso

- Partimos de un banco de datos X con n individuos y p variables que tiene matriz de varianzas covarianzas Σ (supongamos $n > p$).
- Calculamos los **valores propios** de la matrix Σ y los ordenamos de mayor (λ_1) a menor (λ_p).
- Obtenemos los **vectores propios** asociados a esos valores propios: $v_1, v_2, ..., v_p$ (respectivamente)

Análisis de componentes principales: Proceso de extracción de las componentes principales.

Resumen del proceso

- Cada vector v_i nos proporciona los coeficientes de la combinación lineal de variables originales $x_1, x_2, ..., x_p$ para forma la componente principal i-ésima.
- Es decir, $y_i = Xv_i = v_i^1 x_1 + v_i^2 x_2 + ... + v_i^p x_p$.
- La componente principal y_i tendrá varianza λ_i, y por tanto y_1 será la nueva variable con más varianza, seguida de y_2, y así sucesivamente hasta y_p, que será la nueva variable con menos varianza.
- Si queremos reducir la dimensionalidad podemos coger las m primeras componentes, con $m < p$ (más adelante veremos un criterio para seleccionar cuántas podemos seleccionar).

Ejemplo

Supongamos que el banco de datos representado por la matrix X consta únicamente de dos variables, con matriz de varianzas-covarianzas:

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

con $\rho > 0$.

En este banco de datos tanto la primera variable como la segunda tienen varianza 1, por lo que la varianza total de este banco de datos es 2.

Veamos qué forma tienen las 2 componentes principales de X.

Análisis de componentes principales: *ejemplo teórico*.

- Vamos a hallar los vectores y valores propios de la matriz Σ.

$$|\Sigma - \lambda I_2| = \left| \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right| = (1-\lambda)^2 - \rho^2 = 0$$

entonces $\lambda = 1 - \rho$ o $\lambda = 1 + \rho$

- Por tanto, los valores propios de Σ serán (por orden de magnitud) $\lambda_1 = 1 + \rho$ y $\lambda_2 = 1 - \rho$.

- Hallamos el vector propio asociado al valor propio $1 + \rho$:

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = (1+\rho) \begin{pmatrix} a \\ b \end{pmatrix} \Longrightarrow \begin{pmatrix} a + \rho b \\ \rho a + b \end{pmatrix} = \begin{pmatrix} a + \rho a \\ b + \rho b \end{pmatrix}$$

- Por tanto $a = b$ y el vector propio asociado a $\lambda_1 = 1 + \rho$ es de la forma:

$$v_1 = (a, a)$$

- Como además, como v_1 debe cumplir que $v_1' v_1 = 1$ resulta:

$$v_1 = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right)$$

Análisis de componentes principales: *ejemplo teórico*.

- Por tanto la primera componente principal de X será:

$$y_1 = Xv_1 = \frac{1}{\sqrt{2}}(x_1 + x_2)$$

- La varianza de esta primera componente es $\lambda_1 = 1 + \rho$, pero realizamos la comprobación:

$$var(y_1) = var(Xv_1) = v_1' \Sigma v_1 = (1 + \rho) v_1' v_1 = (1 + \rho)$$

Análisis de componentes principales: *ejemplo teórico*.

Por tanto, la **primera componente principal es**:

$$y_1 = \frac{1}{\sqrt{2}}(x_1 + x_2)$$

y tiene varianza $1 + \rho$.

- Mediante el mismo procedimiento obtenemos que la **segunda componente principal es**:

$$y_1 = \frac{1}{\sqrt{2}}(x_1 - x_2)$$

y tiene varianza $1 - \rho$.

Análisis de componentes principales: *ejemplo teórico*.

- Así, el nuevo banco de datos Y formado por las dos componentes principales tiene matriz de varianzas covarianzas:

$$\begin{pmatrix} 1+\rho & 0 \\ 0 & 1-\rho \end{pmatrix}$$

- La varianza de la primera componente principal es $1+\rho$ y la de la segunda $1-\rho$ (la varianza total es 2, como la del banco de datos original)
- El valor de la varianza de la primera componente será mayor cuanto mayor sea el valor de ρ, es decir, cuanto mayor sea la correlación entre las variables originales. Y será menor (más cercana a 1) cuanto menor sea el valor de esta correlación.

Análisis de componentes principales: varianza explicada

- La **varianza total** de una banco de datos X con matriz de varianzas-covarianzas Σ será simplemente la traza de Σ (la suma de los elementos de su diagonal).
- La **varianza** de cada **componente principal** es simplemente el **valor propio** que le corresponde.
- Se cumple que $traza(\Sigma) = \lambda_1 + \lambda_2 + ... + \lambda_p$. Es decir, la varianza total del banco de datos original coincide con la varianza total de las p componentes principales.

Análisis de componentes principales: varianza explicada

- Además de que la **varianza** de cada **componente principal** es el **valor propio** que le corresponde, como **las componentes principales son ortogonales** la **covarianza entre ellas vale 0**.
- Así, la matriz de varianzas-covarianzas de la nueva matriz de datos que forma las componentes principales, Y, quedará:

$$Var(Y) = Var(y_1, y_2, ..., y_p) = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \lambda_p \end{pmatrix}$$

Análisis de componentes principales: varianza explicada

- La **proporción de varianza explicada** por las m **primeras componentes principales** será:

$$\frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{p} \lambda_i} = \frac{\sum_{i=1}^{m} \lambda_i}{traza(\Sigma)}$$

- El denominador de este indicador es la **varianza total de todas las componentes principales**, que coincide con la **varianza total del banco de datos original**.

- Este **indicador** es muy útil, ya que nos informa sobre **cuántas componentes principales son necesarias para explicar una proporción determinada de la varianza total**.

- Este valor nos guiará en el **proceso de reducción de variables** para determinar con cuántas componentes podemos reducir los datos originales (conservando una varianza razonable).

Análisis de componentes principales: enfoques

- El ACP que hemos presentado parte de la **descomposición de la matriz de varianzas-covarianzas (o correlaciones) en valores y vectores propios**. Es decir, aprovecha la relación entre las variables para reducir la dimensionalidad
- Existe otra posibilidad prácticamente equivalente. Se trata de la **descomposición de la matriz de datos en valores singulares (SVD)**. Este otro enfoque examina la covarianza/correlación entre individuos del banco de datos.

Los resultados de ambos métodos son muy similares.

Análisis de componentes principales: en R (I)

- Se puede llevar a cabo un análisis de componentes principales en R con diferentes funciones de distintas librerías. Entre ellas se encuentran *princomp* y *prcomp*,
  - *princomp*: Calcula el ACP mediante la descomposición en valores y vectores propios de la matriz de varinazas-covarianzas (o matriz de correlaciones) de las variables consideradas.
  - *prcomp*: Calcula el ACP mediante la descomposición en valores singulares de la matriz de datos.

  Ambos métodos obtienen, en general, resultados muy similares. Si bien es verdad que el segundo puede resultar más estable numéricamente.

Análisis de componentes principales: en R (II)

En adelante nos centraremos en la función *princomp*, pues se corresponde con el procedimiento que hemos descrito en esta sesión. Su *sintaxis en R*:

```
princomp(x, cor = FALSE, scores = TRUE, ...)
```

- *x* representa la matriz de datos
- *cor* recoge si se realizará a partir de correlaciones o no (en ese caso se realizará mediante la matriz de varianzas-covarianzas)
- ...

Análisis de componentes principales: en R (III)

Princomp devuelve un objeto con las siguientes componentes:

- *Sdev*: desviaciones estándar de cada una de las dimensiones del ACP (raiz cuadrada de $\lambda_1, \lambda_2, \ldots$).
- *Loadings*: pesos de las variables en cada una de las variables del ACP (vectores $a_1, a_2, \ldots$ de teoría).
- *Center*: Las medias sustraídas a cada una de las variables.
- *Scale*: En caso de que la escala de las variables haya sido modificada (en breve veremos para qué) este vector contendrá los factores correctores.
- *N.obs*: número de observaciones.
- *Scores*: puntuaciones de los individuos en cada dimensión del ACP ($Xa_1, Xa_2, \ldots$).
- *Call*: Llamada a la función princomp que ha generado el objeto.
- *Na.action*: Tratamiento que se le ha dado a los valores

- Volvamos a nuestro banco de datos inicial de indicadores de desarrollo por países, *MundoDes*.
- Vamos a realizar un **análisis de componentes principales** de (por ahora) las **5 primeras variables del banco de datos**, para explorar estas variables y resumir su contenido.

```
ACP1 <- princomp(MundoDes[, 1:5])
```

- Varianza explicada por cada una de las dimensiones del ACP:

```
summary(ACP1)
```

```
## Importance of components:
##                            Comp.1     Comp.2      Comp.3      Comp.4
## Standard deviation     49.6631000 7.06172925 4.522072696 2.344799529
## Proportion of Variance  0.9695957 0.01960402 0.008038927 0.002161397
## Cumulative Proportion   0.9695957 0.98919969 0.997238616 0.999400013
##                              Comp.5
## Standard deviation     1.2354054430
## Proportion of Variance 0.0005999873
## Cumulative Proportion  1.0000000000
```

Análisis de componentes principales: *ejemplo práctico en R*

- Varianza explicada por cada una de las dimensiones del ACP:

```
summary(ACP1)
```

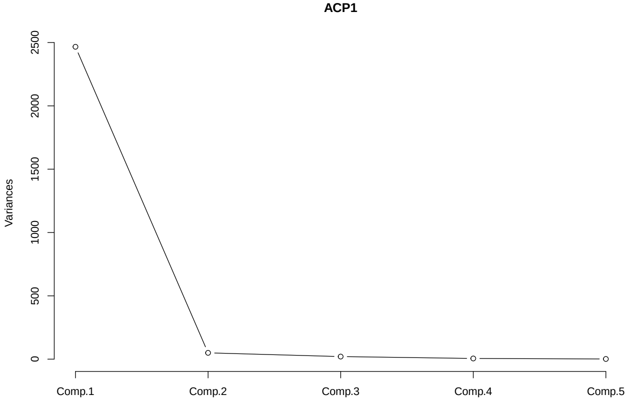
```
## Importance of components:
##                           Comp.1     Comp.2      Comp.3      Comp.4
## Standard deviation     49.6631000 7.06172925 4.522072696 2.344799529
## Proportion of Variance  0.9695957 0.01960402 0.008038927 0.002161397
## Cumulative Proportion   0.9695957 0.98919969 0.997238616 0.999400013
##                              Comp.5
## Standard deviation     1.2354054430
## Proportion of Variance 0.0005999873
## Cumulative Proportion  1.0000000000
```

- La **primera dimensión** del ACP resume perfectamente el contenido de las 5 variables del banco de datos, explica el 97.0% de la varianza original.
- La **segunda dimensión** del ACP explica únicamente el 2% aprox.

Análisis de componentes principales: *ejemplo práctico en R*

Podemos observar cómo disminuye la proporción de varianza que explican las componentes 2,3,... respecto de la primera.

```
plot(ACP1, type = "lines")
```



Pesos de cada componente principal

```
ACP1$loadings
```

```
##
## Loadings:
##           Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## Tasa Nat   0.241  0.907  0.211  0.271
## Tasa Mort               -0.619  0.770 -0.105
## Mort Inf   0.926 -0.311  0.206
## Esp Hom   -0.185 -0.163  0.574  0.361 -0.692
## Esp Muj   -0.215 -0.211  0.448  0.448  0.712
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## SS loadings       1.0    1.0    1.0    1.0    1.0
## Proportion Var    0.2    0.2    0.2    0.2    0.2
## Cumulative Var    0.2    0.4    0.6    0.8    1.0
```

Primera componente principal (97% de varianza aprox.)

```
round(ACP1$loadings[, 1], 3)
```

```
##  Tasa Nat Tasa Mort  Mort Inf   Esp Hom   Esp Muj
##     0.241     0.064     0.926    -0.185    -0.215
```

- La primera dimensión del ACP sitúa en un extremo del eje aquellos países con tasas de natalidad y mortalidad infantil altas y esperanzas de vida bajas tanto en hombres como en mujeres, y en el otro extremo del eje se sitúan los países con comportamiento opuesto.
- ¿Eje **desarrollo**?.

Análisis de componentes principales: *ejemplo práctico en R*

Segunda componente principal (2% de varianza aprox.)

```
round(ACP1$loadings[, 2], 3)
```

```
##  Tasa Nat Tasa Mort  Mort Inf   Esp Hom   Esp Muj
##     0.907    -0.098    -0.311    -0.163    -0.211
```

- La interpretación de la segunda dimensión del ACP ya no es tan trivial, domina la separación de paises de alta natalidad y baja mortalidad infantil y esperanza de vida frente a paises con baja natalidad pero alta mortalidad infantil y esperanza de vida en hombres y mujeres.
- ¿Eje **natalidad vs mortalidad**?

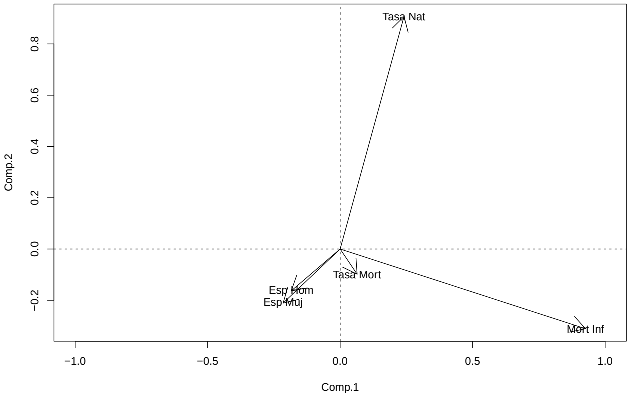
(*No estudiamos el resto de componentes porque son irrelevantes*)

Representación gráfica de los pesos de las componentes principales

```
plot(ACP1$loadings[, 1:2], type = "n", xlim = c(-1, 1))
for (i in 1:nrow(ACP1$loadings)) {
    arrows(0, 0, ACP1$loadings[i, 1], ACP1$loadings[i, 2])
}
text(ACP1$loadings[, 1:2], dimnames(ACP1$loadings)[[1]])
abline(h = 0, lty = 2)
abline(v = 0, lty = 2)
```

Análisis de componentes principales: *ejemplo práctico en R*

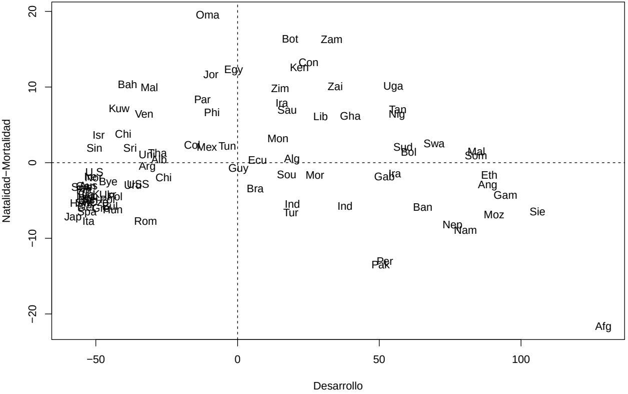
Representación gráfica de los pesos de las componentes principales



Representación gráfica de los individuos según las dos primeras componentes principales

```
plot(ACP1$scores[, 1:2], xlab = "Desarrollo", ylab = "Natalidad-Mortalidad",
    type = "n")
text(ACP1$scores[, 1], ACP1$scores[, 2], substr(MundoDes[, 1], 1,
    3))
abline(h = 0, lty = 2)
abline(v = 0, lty = 2)
```

Representación gráfica de los individuos según las dos primeras componentes principales

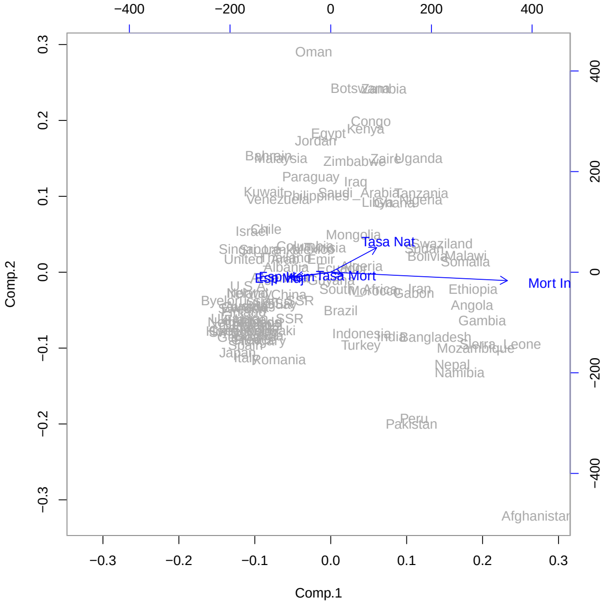


Análisis de componentes principales: *ejemplo práctico en R*

Representación conjunta de individuos y variables que definen las dos primeras componentes principales: función biplot

```
biplot(ACP1, col = c("darkgrey", "blue"))
```

Análisis de componentes principales: *ejemplo práctico en R*



Individuos en los extremos (3) mayores y (3) menores de cada componente principal

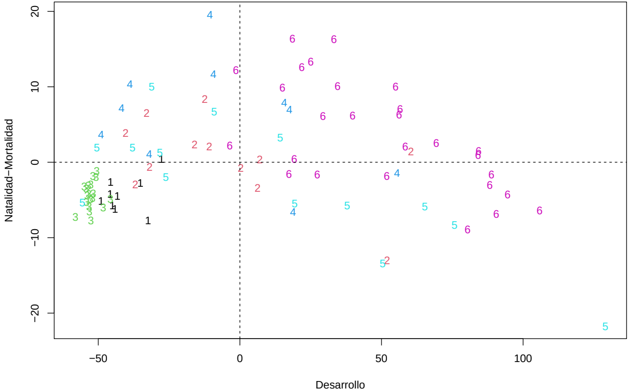
```
rownames(MundoDes)[order(ACP1$scores[, 1])[c(1, 2, 3, 89, 90, 91)]]
```

```
## [1] "Japan"        "Hong_Kong"    "Sweden"       "Gambia"
## [5] "Sierra_Leone" "Afghanistan"
```

```
rownames(MundoDes)[order(ACP1$scores[, 2])[c(1, 2, 3, 89, 90, 91)]]
```

```
## [1] "Afghanistan" "Pakistan"    "Peru"        "Zambia"
## [5] "Botswana"    "Oman"
```

Representamos los paises por su continente



Encontramos gran número de países de zonas 6 (África) y 4 (Medio oriente) en la zona superior del segundo eje. Estos paises tienen una natalidad elevada en relación a su mortalidad. Cuando nos fijamos con atención el segundo eje parece discriminar a los paises de cultura islámica (con mayor natalidad) del resto de países.

Incluyamos ahora la variable **PNB** en el análisis y veamos cómo cambian los resultados.

```
ACP2 <- princomp(MundoDes[, 1:6])
ACP2$loadings
```

```
##
## Loadings:
##           Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
## Tasa Nat          0.216  0.926  0.196  0.236
## Tasa Mort                      -0.625  0.764 -0.117
## Mort Inf          0.936 -0.277  0.207
## Esp Hom          -0.173 -0.146  0.572  0.364 -0.699
## Esp Muj          -0.202 -0.197  0.447  0.474  0.704
## PNB       -1.000
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
## SS loadings     1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var  0.167  0.167  0.167  0.167  0.167  0.167
## Cumulative Var  0.167  0.333  0.500  0.667  0.833  1.000
```

Análisis de componentes principales: *ejemplo práctico en R*

Incluyamos ahora la variable **PNB** en el análisis y veamos cómo cambian los resultados.

- La primera dimensión del ACP cambia dramáticamente.
- La combinación de variables que explica una mayor proporción de varianza del banco de datos es simplemente el producto nacional bruto.

```
summary(ACP2)
```

```
## Importance of components:
##                             Comp.1       Comp.2       Comp.3
## Standard deviation     8049.1442208 3.921148e+01 6.782552e+00
## Proportion of Variance    0.9999751 2.373101e-05 7.100289e-07
## Cumulative Proportion     0.9999751 9.999989e-01 9.999996e-01
##                             Comp.4       Comp.5       Comp.6
## Standard deviation     4.518376e+00 2.200316e+00 1.233306e+00
## Proportion of Variance 3.151044e-07 7.472394e-08 2.347641e-08
## Cumulative Proportion  9.999999e-01 1.000000e+00 1.000000e+00
```

Incluyamos ahora la variable **PNB** en el análisis y veamos cómo cambian los resultados.

- Efectivamente, la primera dimensión explica una variabilidad muy superior al resto.

```
head(MundoDes[, 1:6])
```

```
##                 Tasa Nat Tasa Mort Mort Inf Esp Hom Esp Muj  PNB
## Albania             24.7       5.7     30.8    69.6    75.5  600
## Bulgaria            12.5      11.9     14.4    68.3    74.7 2250
## Czechoslovaki       13.4      11.7     11.3    71.8    77.7 2980
## Hungary             11.6      13.4     14.8    65.4    73.8 2780
## Poland              14.3      10.2     16.0    67.2    75.7 1690
## Romania             13.6      10.7     26.9    66.5    72.4 1640
```

- Por tanto la escala de las variables analizadas tiene una gran influencia sobre el resultado del ACP.
- Sería conveniente evitar esta influencia arbitraria.

Análisis de componentes principales: a partir de la matriz de correlaciones

- Para evitar el efecto de la escala de las variables disponemos de dos alternativas obvias:
  - Llevar a cabo el ACP de las variables estandarizadas en lugar de las variables originales.
  - Llevar a cabo el ACP mediante la descomposición de valores y vectores propios de la matriz de correlación entre variables, en lugar de la matriz de varianzas-covarianzas.
- Ambas propuestas son equivalentes y conducen exactamente a la misma solución.
- En la función *princomp* ponemos *cor=TRUE*.
- Aún así, cuando las variables estén medidas en unidades comparables es preferible realizar el ACP de la matriz de covarianzas ya que las diferencias de escala entre variables pueden ser informativas.

Análisis de componentes principales: a partir de la matriz de correlaciones

Volviendo al ejemplo

```
ACP3 <- princomp(MundoDes[, 1:6], cor = TRUE)
ACP3$loadings
```

```
##
## Loadings:
##           Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
## Tasa Nat   0.416  0.196  0.513  0.683  0.233
## Tasa Mort  0.341 -0.680 -0.524  0.307  0.225
## Mort Inf   0.440         0.222 -0.632  0.578  0.145
## Esp Hom   -0.452                0.114  0.639 -0.605
## Esp Muj   -0.454        -0.130  0.159  0.378  0.780
## PNB       -0.326 -0.699  0.628
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
## SS loadings     1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var  0.167  0.167  0.167  0.167  0.167  0.167
## Cumulative Var  0.167  0.333  0.500  0.667  0.833  1.000
```

- Ahora el primer eje nuevamente mide el desarrollo de los paises.

Análisis de componentes principales: **tipos de componentes**

- Cuando existe una **alta correlación positiva entre las variables** la primera componente suele tener todas su coordenadas del mismo signo y puede interpretarse como un promedio ponderado de todas las variables o un factor global **de tamaño**.
- Los restantes componentes se interpretan como factores **de forma**, y típicamente tienen coordenadas positivas y negativas que implica que contraponen unos grupos de variables frente a otros.
- El signo de las coordenadas es intercambiable, pues marcan una *dirección*.

Análisis de componentes principales: **conclusiones** (I)

- El ACP nos ayuda a **reducir el número de variables creando** unas **variables ficticias** que explican en orden descendente la mayor parte de variabilidad de los sujetos.
- **No siempre es fácil dotar de significado** a cada componente principal.
- No tiene sentido plantear un ACP **cuando no existe correlación entre las variables** (pues en ese caso cada variable explica una parte de variabilidad que no puede ser explicada por otra).

Análisis de componentes principales: **conclusiones** (II)

- El **ACP no es invariante frente a cambios de escala**, por lo que es importante trabajar con variables medidas en escalas comparables o en caso de no ser así trabajar sobre la matriz de correlaciones o estandarizar las variables.
- Si las variables tienen escalas comparables **no se recomienda ni estandarizar ni trabajar sobre la matriz de correlaciones**, pues en ese caso se puede perder parte de la información.

Anexo práctico: PCA en R

Anexo práctico: **PCA en R**

Funciones básicas

- La función *princomp* sobre un banco de datos es equivalente a la función *eigen* sobre su matriz de varianzas-covarianzas.
- La función *prcomp* sobre un banco de datos es equivalente a la función *svd* sobre la misma matriz de datos.

Ejemplo

```
pca.princomp <- princomp(MundoDes[, 1:6], cor = TRUE)
pca.eigen <- eigen(cor(MundoDes[, 1:6]))

pca.prcomp <- prcomp(MundoDes[, 1:6], scale. = TRUE)
pca.svd <- svd(scale(MundoDes[, 1:6]))
```

Ejemplo: Comparamos *princomp* con *prcomp*

```
pca.princomp$sdev
```

```
##    Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6
## 2.1743457 0.8521130 0.6007915 0.3424949 0.2324781 0.1175800
```

```
pca.prcomp$sdev
```

```
## [1] 2.1743457 0.8521130 0.6007915 0.3424949 0.2324781 0.1175800
```

Ejemplo: Comparamos *princomp* con *prcomp*

```
round(pca.princomp$loadings[, 1:3], 3)
```

```
##            Comp.1 Comp.2 Comp.3
## Tasa Nat    0.416  0.196  0.513
## Tasa Mort   0.341 -0.680 -0.524
## Mort Inf    0.440 -0.052  0.222
## Esp Hom    -0.452  0.085 -0.029
## Esp Muj    -0.454  0.034 -0.130
## PNB        -0.326 -0.699  0.628
```

```
round(pca.prcomp$rotation[, 1:3], 3)
```

```
##               PC1    PC2    PC3
## Tasa Nat    0.416 -0.196  0.513
## Tasa Mort   0.341  0.680 -0.524
## Mort Inf    0.440  0.052  0.222
## Esp Hom    -0.452 -0.085 -0.029
## Esp Muj    -0.454 -0.034 -0.130
## PNB        -0.326  0.699  0.628
```

Ejemplo: Comparamos *princomp* con *prcomp*

```
round(pca.princomp$scores[1:3, ], 3)
```

```
##              Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
## Albania      -1.312  1.242 -0.268  0.024  0.299  0.084
## Bulgaria     -1.361  0.023 -1.363  0.012  0.050 -0.043
## Czechoslovaki -1.694  0.046 -1.310  0.167  0.341 -0.054
```

```
round(pca.prcomp$x[1:3, ], 3)
```

```
##                  PC1    PC2    PC3   PC4   PC5    PC6
## Albania       -1.305 -1.236 -0.266 0.024 0.297  0.084
## Bulgaria      -1.353 -0.023 -1.355 0.011 0.049 -0.043
## Czechoslovaki -1.684 -0.045 -1.303 0.166 0.339 -0.054
```

Ejemplo: Comparamos *princomp* con *eigen*

```
pca.princomp$sdev
```

```
##    Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6
## 2.1743457 0.8521130 0.6007915 0.3424949 0.2324781 0.1175800
```

```
sqrt(pca.eigen$values)
```

```
## [1] 2.1743457 0.8521130 0.6007915 0.3424949 0.2324781 0.1175800
```

Ejemplo: Comparamos *princomp* con *eigen*

```
round(pca.princomp$loadings[, 1:3], 3)
```

```
##            Comp.1 Comp.2 Comp.3
## Tasa Nat    0.416  0.196  0.513
## Tasa Mort   0.341 -0.680 -0.524
## Mort Inf    0.440 -0.052  0.222
## Esp Hom    -0.452  0.085 -0.029
## Esp Muj    -0.454  0.034 -0.130
## PNB        -0.326 -0.699  0.628
```

```
round(pca.eigen$vectors[, 1:3], 3)
```

```
##        [,1]   [,2]   [,3]
## [1,] -0.416 -0.196  0.513
## [2,] -0.341  0.680 -0.524
## [3,] -0.440  0.052  0.222
## [4,]  0.452 -0.085 -0.029
## [5,]  0.454 -0.034 -0.130
## [6,]  0.326  0.699  0.628
```

Ejemplo: Comparamos *princomp* **con** *eigen*

```
round(pca.princomp$scores[1:3, ], 3)
```

```
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
## Albania        -1.312  1.242 -0.268  0.024  0.299  0.084
## Bulgaria       -1.361  0.023 -1.363  0.012  0.050 -0.043
## Czechoslovaki -1.694  0.046 -1.310  0.167  0.341 -0.054
```

```
round((scale(MundoDes[, 1:6]) %*% pca.eigen$vectors[, 1:3])[1:3, ],
    3)
```

```
##                [,1]   [,2]   [,3]
## Albania        1.305 -1.236 -0.266
## Bulgaria       1.353 -0.023 -1.355
## Czechoslovaki 1.684 -0.045 -1.303
```

Ejemplo: Comparamos *prcomp* **con** *svd*

```
pca.prcomp$sdev
```

```
## [1] 2.1743457 0.8521130 0.6007915 0.3424949 0.2324781 0.1175800
```

```
(pca.svd$d)/sqrt(dim(MundoDes)[1] - 1)
```

```
## [1] 2.1743457 0.8521130 0.6007915 0.3424949 0.2324781 0.1175800
```

Ejemplo: Comparamos *prcomp* con *svd*

```
round(pca.prcomp$rotation[, 1:3], 3)
```

```
##                PC1    PC2    PC3
## Tasa Nat    0.416 -0.196  0.513
## Tasa Mort   0.341  0.680 -0.524
## Mort Inf    0.440  0.052  0.222
## Esp Hom    -0.452 -0.085 -0.029
## Esp Muj    -0.454 -0.034 -0.130
## PNB        -0.326  0.699  0.628
```

```
round(pca.svd$v[, 1:3], 3)
```

```
##        [,1]   [,2]   [,3]
## [1,]  0.416 -0.196  0.513
## [2,]  0.341  0.680 -0.524
## [3,]  0.440  0.052  0.222
## [4,] -0.452 -0.085 -0.029
## [5,] -0.454 -0.034 -0.130
## [6,] -0.326  0.699  0.628
```

Ejemplo: Comparamos *prcomp* con *svd*

```
round(pca.prcomp$x[1:3, ], 3)
```

```
##                   PC1    PC2    PC3   PC4   PC5    PC6
## Albania        -1.305 -1.236 -0.266 0.024 0.297  0.084
## Bulgaria       -1.353 -0.023 -1.355 0.011 0.049 -0.043
## Czechoslovaki -1.684 -0.045 -1.303 0.166 0.339 -0.054
```

```
round((pca.svd$u %*% diag(pca.svd$d))[1:3, ], 3)
```

```
##        [,1]   [,2]   [,3]  [,4]  [,5]   [,6]
## [1,] -1.305 -1.236 -0.266 0.024 0.297  0.084
## [2,] -1.353 -0.023 -1.355 0.011 0.049 -0.043
## [3,] -1.684 -0.045 -1.303 0.166 0.339 -0.054
```