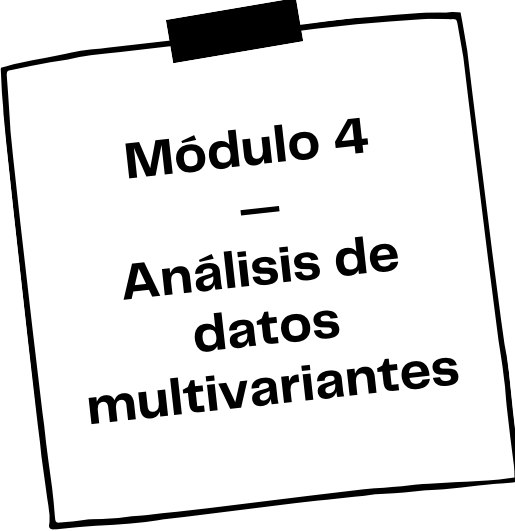


Componentes principales



Módulo 4
—
**Análisis de
datos
multivariantes**

Análisis de datos multivariantes

Uno de los objetivos principales del análisis de datos multivariantes es la reducción de la dimensionalidad que consiste en describir los valores de p variables por un pequeño subconjunto r de ellas. De esta forma, **se reduce la dimensión del problema con una pequeña pérdida de información**

El subconjunto de variables se construye de tal forma que capturen la máxima información posible de las variables originales

Componentes principales

El análisis de componentes principales (PCA) tiene como objetivo representar adecuadamente la información de n observaciones de p variables posiblemente correlacionadas, con un número menor de variables conocidas como **Componentes Principales**, que son combinaciones lineales de las variables originales y que son independientes entre sí

$$Y_1 = a_1' X = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p$$

$$Y_2 = a_2' X = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p$$

$$\vdots \qquad \qquad \qquad \vdots$$

$$Y_p = a_p' X = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p$$

Componentes principales

Cuando queremos analizar una gran cantidad de variables no es posible graficarlas, lo que dificulta tener una idea de las posibles tendencias que se presentan en los datos. Por esta razón, es útil aplicar el análisis de componentes principales ya que al reducir la dimensionalidad de los datos, será posible visualizarlos e identificar si hay grupos de individuos que son muy similares o muy diferentes entre si

Pasos para calcular los componentes principales

1. Estandarizar las variables originales

$$Z_i = \frac{x_i - \bar{X}}{S}$$

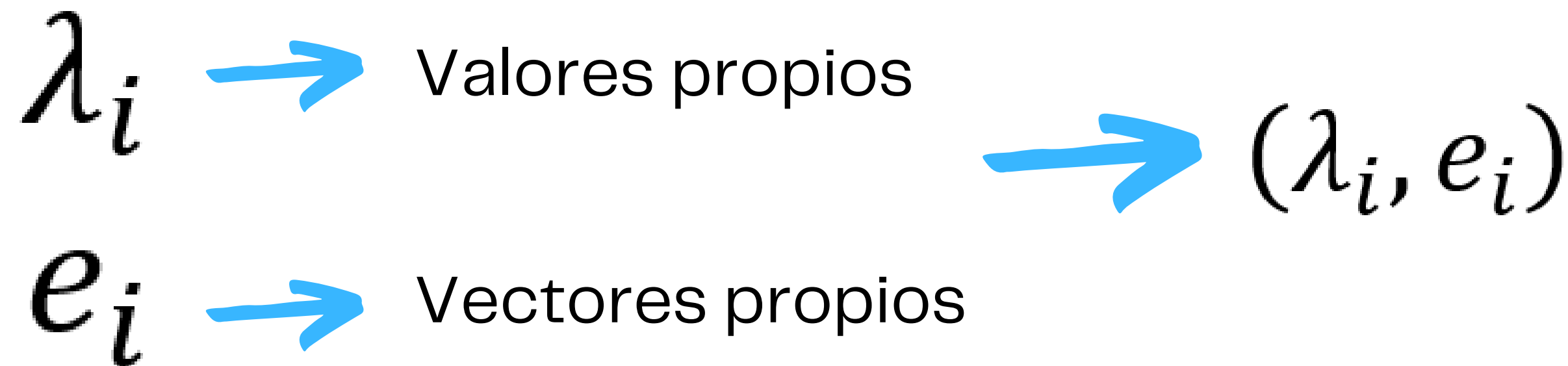
Pasos para calcular los componentes principales

2. Calcular la matriz de varianzas y covarianzas de la matriz de datos estandarizados

$$\mathbf{S} = \begin{bmatrix} s_1^2 & \dots & s_{1p} \\ \vdots & & \vdots \\ s_{p1} & \dots & s_p^2 \end{bmatrix} = \begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \dots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) & \dots & \text{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \text{Cov}(X_n, X_2) & \dots & \text{Var}(X_n) \end{bmatrix}$$

Pasos para calcular los componentes principales

4. Calcular los valores propios y vectores propios de la matriz de varianzas y covarianzas



**La matriz de varianzas y covarianzas tiene tantos valores propios como filas (o columnas).
Cada valor propio tiene un vector propio asociado.**

Pasos para calcular los componentes principales

5. Ordenar las parejas de valores y vectores propios de mayor a menor según el valor del valor propio

$$(\lambda_1, e_1), \quad (\lambda_2, e_2), \quad \dots, \quad (\lambda_p, e_p)$$



$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$$

Pasos para calcular los componentes principales

6. Calcular los componentes principales

La i -ésima componente principal está dada por la siguiente combinación lineal:

$$Y_i = e_i' X = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ip}X_p$$

Calculemos los componentes principales para nuestra base de datos de características físicas de 27 estudiantes. Los coeficientes representan el grado de importancia de cada variable con la respectiva CP:

$$\begin{aligned} Y_1 &= 0.41X_1 + 0.39X_2 + 0.40X_3 + 0.39X_4 + 0.38X_5 + 0.29X_6 + 0.37X_7 \\ Y_2 &= -0.15X_1 + 0.04X_2 - 0.20X_3 - 0.31X_4 + 0.11X_5 + 0.89X_6 - 0.15X_7 \\ Y_3 &= -0.05X_1 + 0.29X_2 - 0.12X_3 + 0.16X_4 + 0.56X_5 - 0.19X_6 - 0.71X_7 \\ Y_4 &= 0.32X_1 - 0.79X_2 - 0.19X_3 + 0.37X_4 + 0.25X_5 + 0.13X_6 - 0.07X_7 \\ Y_5 &= 0.09X_1 - 0.26X_2 + 0.39X_3 - 0.69X_4 + 0.49X_5 - 0.16X_6 + 0.11X_7 \\ Y_6 &= -0.31X_1 + 0.09X_2 - 0.61X_3 + 0.01X_4 + 0.44X_5 - 0.15X_6 + 0.56X_7 \\ Y_7 &= -0.77X_1 - 0.22X_2 + 0.47X_3 + 0.33X_4 + 0.13X_5 + 0.08X_6 + 0.03X_7 \end{aligned} \quad \rightarrow \quad \begin{aligned} X_1 &: est \\ X_2 &: pes \\ X_3 &: lpie \\ X_4 &: lbr \\ X_5 &: aes \\ X_6 &: dcr \\ X_7 &: lrt \end{aligned}$$

Cada una de estas componentes principales explica un porcentaje de la variación total en el conjunto de datos

Ahora debemos obtener el valor de cada observación (estudiante) en la primera y segunda componente principal:

Estudiante 1:

$$x_{11} = 159, x_{12} = 49, x_{13} = 36, x_{14} = 68, x_{15} = 42, x_{16} = 57, x_{17} = 40$$



Escalamos los valores (le restamos su media y la dividimos por su desviación estándar):

$$z_{11} = -0.95, z_{12} = -1.16, z_{13} = -1.04, z_{14} = -1.10, z_{15} = -0.95, z_{16} = -0.13, z_{17} = -0.98$$

Ahora debemos obtener el valor de cada observación (estudiante) en la primera y segunda componente principal:

Estudiante 1:

Reemplazamos estos valores en la primera CP:

$$\begin{aligned} Y_1 &= 0.41X_1 + 0.39X_2 + 0.40X_3 + 0.39X_4 + 0.38X_5 + 0.29X_6 + 0.37X_7 \\ &= 0.41(-0.95) + 0.39(-1.16) + 0.40(-1.04) + 0.39(-1.10) + 0.38(-0.95) + 0.29(-0.13) + 0.37(-0.98) \\ &= -2.45 \end{aligned}$$

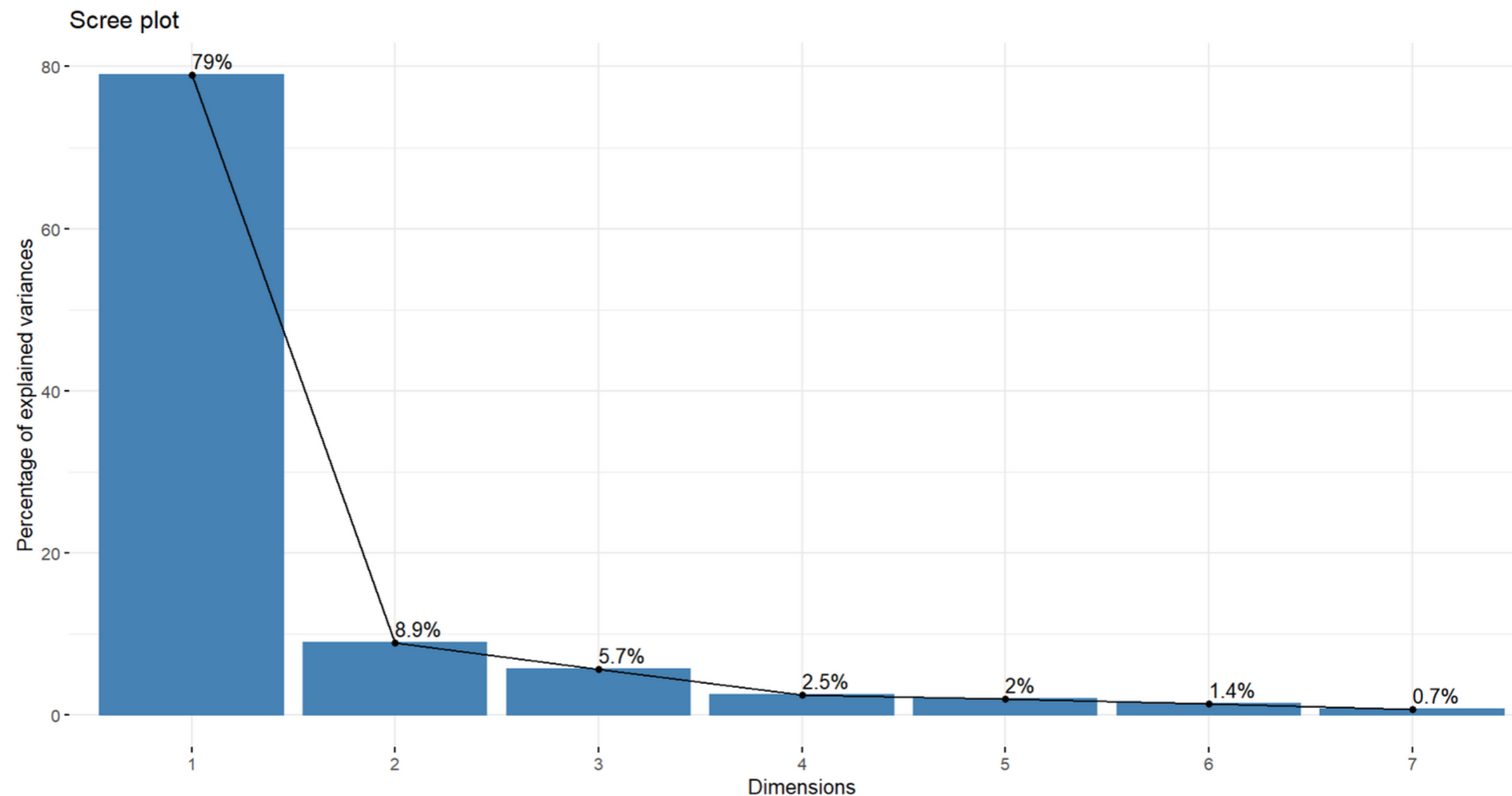
Reemplazamos estos valores en la segunda CP:

$$\begin{aligned} Y_2 &= -0.15X_1 + 0.04X_2 - 0.20X_3 - 0.31X_4 + 0.11X_5 + 0.89X_6 - 0.15X_7 \\ &= -0.15(-0.95) + 0.04(-1.16) - 0.20(-1.04) - 0.31(-1.10) + 0.11(-0.95) + 0.89(-0.13) - 0.15(-0.98) \\ &= 0.57 \end{aligned}$$

Puntajes de cada observación (estudiante) sobre cada componente principal:

Estudiante	PC1	PC2	PC3	PC4	PC5	PC6	PC7
1	-2,458	0,577	-0,158	0,211	0,020	-0,143	-0,021
2	-0,710	-1,092	-0,260	-0,371	0,071	0,268	0,217
3	0,576	0,314	0,118	0,385	-0,119	0,456	-0,209
4	-0,933	0,354	-1,096	0,536	-0,555	0,016	-0,111
5	-2,203	-1,063	0,634	0,415	0,183	0,244	-0,286
6	-1,049	-0,249	0,196	-0,742	-0,073	0,162	0,082
7	-1,496	-1,449	-0,766	0,385	0,105	-0,416	-0,003
8	2,695	0,791	-0,579	-0,094	0,767	-0,226	-0,109
9	2,570	0,276	-0,644	0,284	-0,454	-0,078	-0,358
10	-2,122	0,574	-0,056	0,257	0,201	0,290	0,146

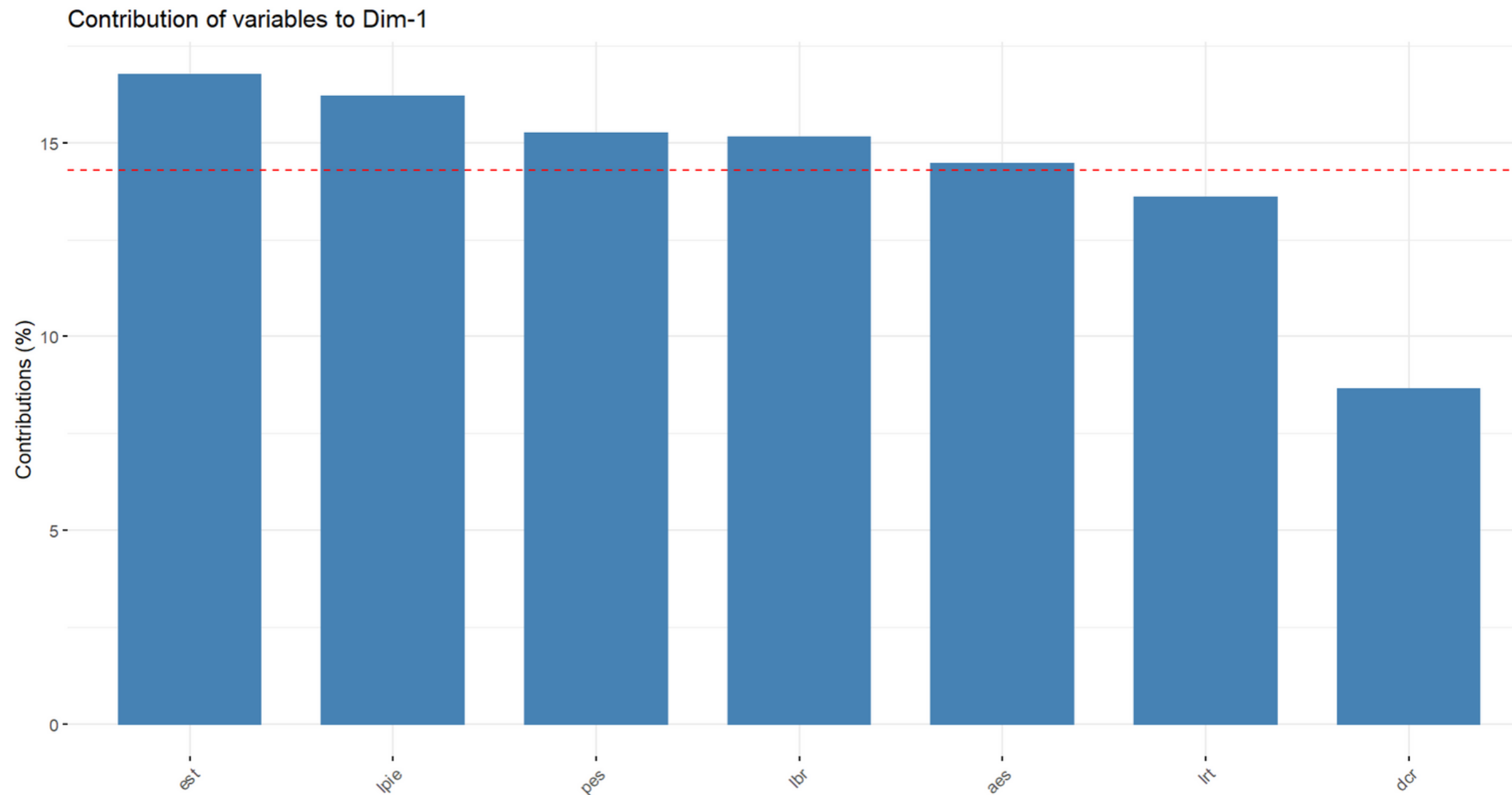
Gráfico de sedimentación: gráfico que muestra el porcentaje de varianza explicado por cada componente principal



- El componente principal 1 explica el 79% de la variabilidad total
- El componente principal 2 explica el 8% de la variabilidad total
- El componente principal 3 explica el 5% de la variabilidad total

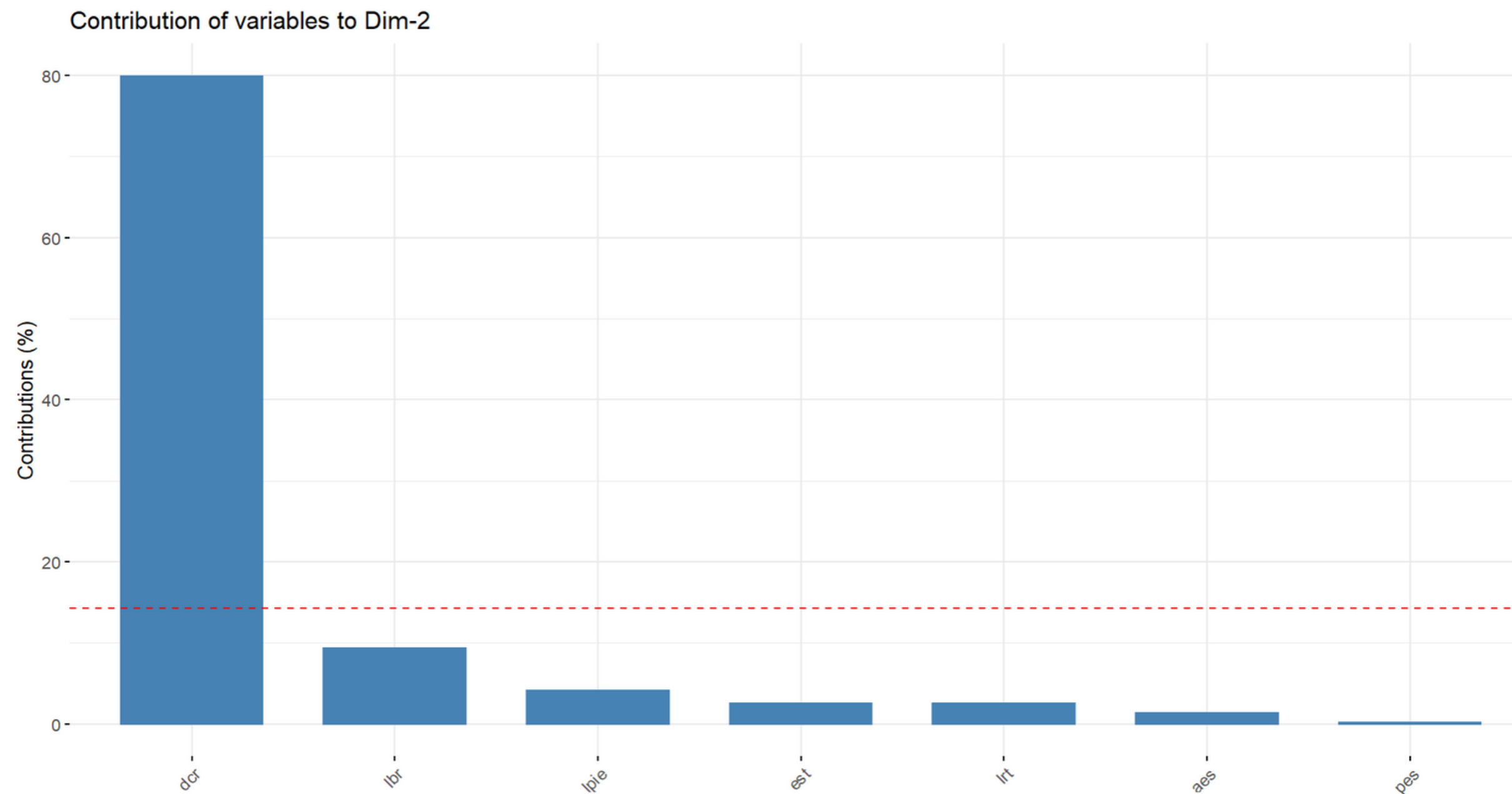
La proporción de variabilidad acumulada para los primeros dos componentes es de aproximadamente 88 %

Gráfico de contribuciones Componente 1



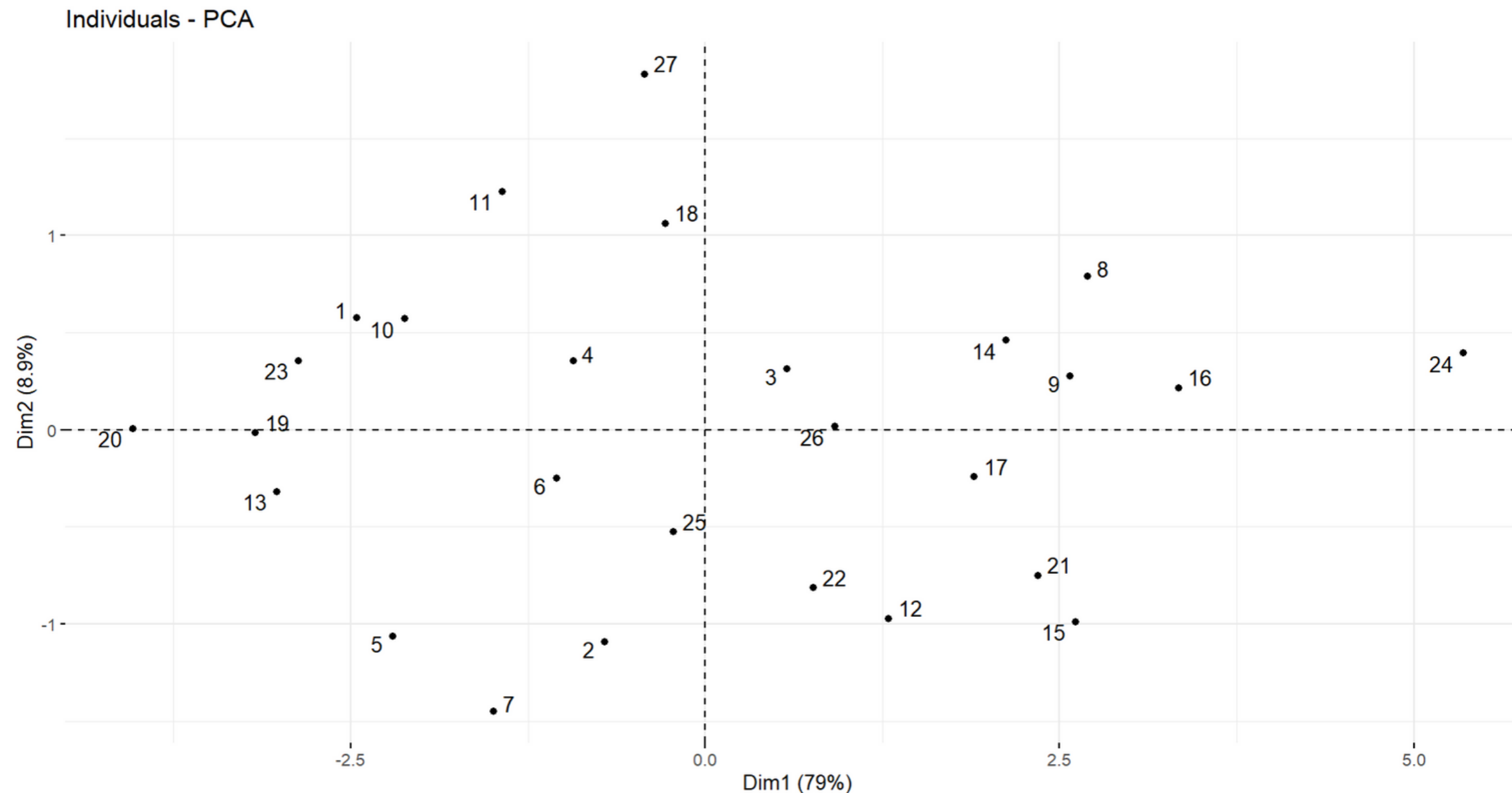
La línea punteada roja expresa el valor que supondría un escenario en el que todas las variables contribuyen por igual. Una variable con una contribución superior a este valor puede considerarse importante para la construcción de la componente 1

Gráfico de contribuciones Componente 2



La línea punteada roja expresa el valor que supondría un escenario en el que todas las variables contribuyen por igual. Una variable con una contribución superior a este valor puede considerarse importante para la construcción de la componente 2

Gráfico de individuos: representa visualmente las puntuaciones del segundo componente principal versus las puntuaciones del primer componente principal

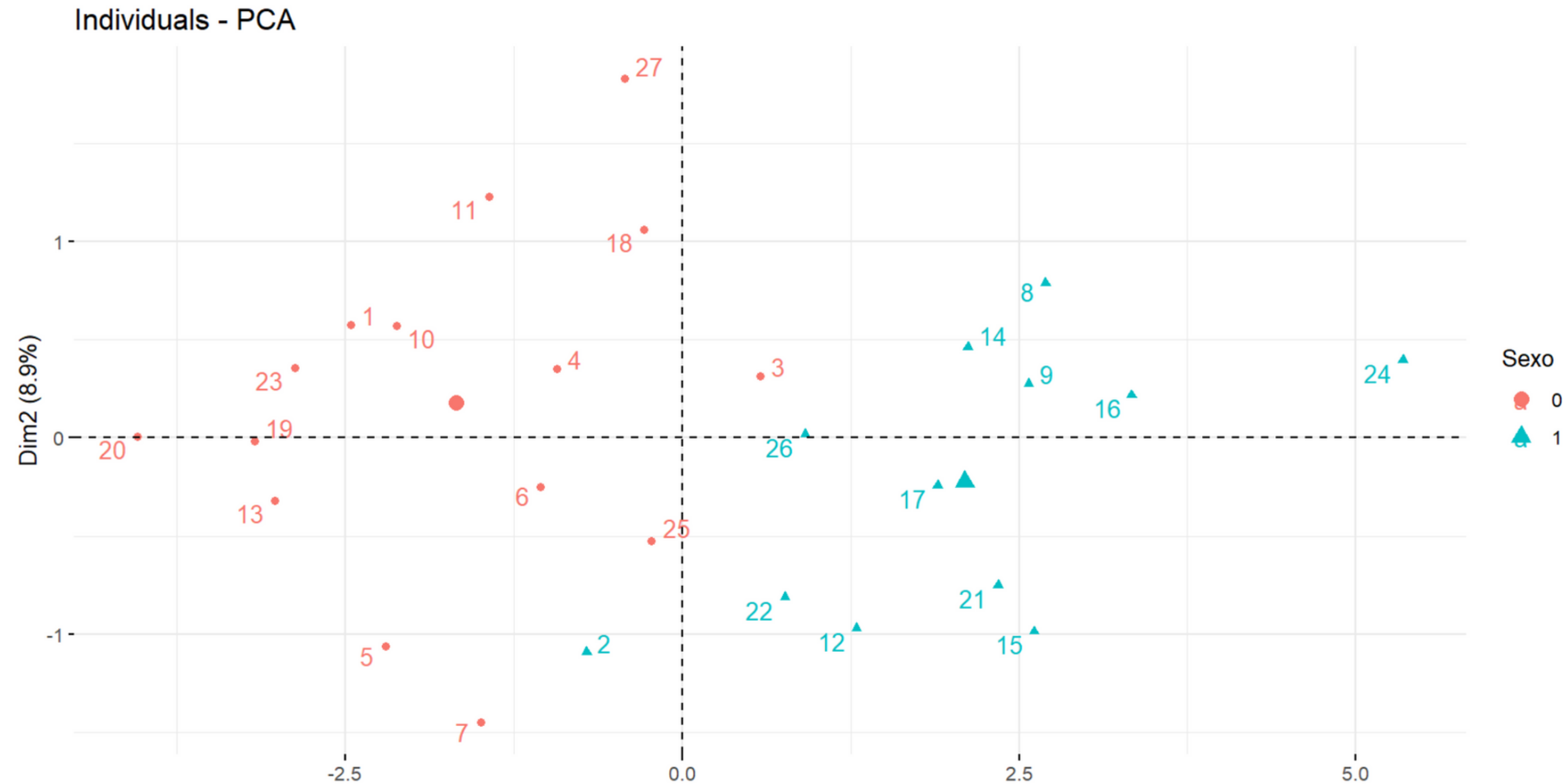


Esta gráfica es útil para evaluar la estructura de los datos y detectar conglomerados, valores atípicos y tendencias

Interpretación gráfico de individuos

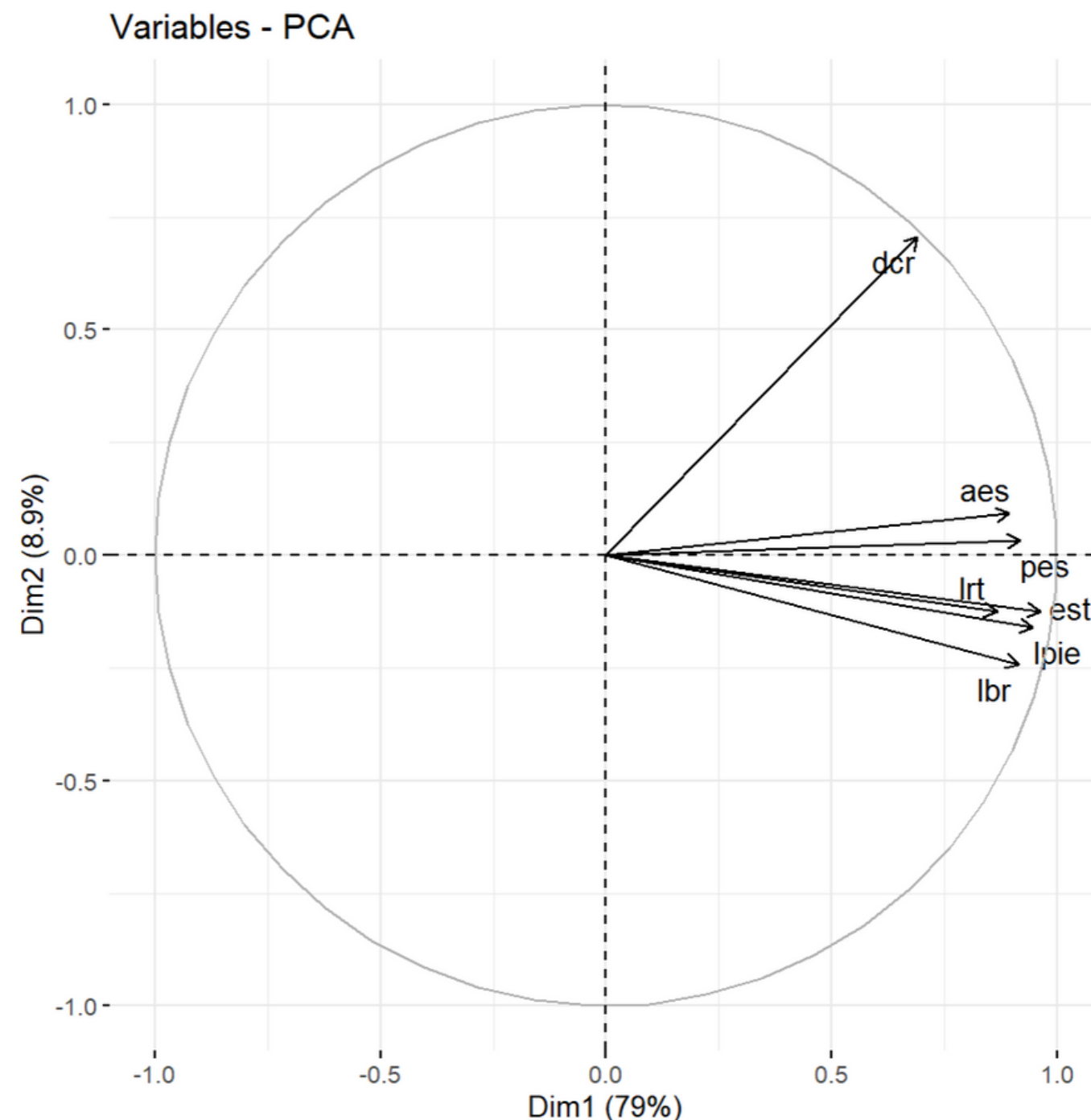
- Las observaciones que son similares entre sí serán similares en el gráfico de individuos, mientras que las observaciones mucho más alejadas son diferentes.
- Las puntuaciones más alejadas son valores atípicos.
- Los puntos cercanos al promedio aparecen cerca del origen del gráfico de individuos.
- El estudiante 27 tiene uno de los valores más altos respecto a la variable **dcr**. La segunda componente está altamente influenciada por esta variable con una ponderación de 0.89.
- El estudiante 24 tiene los valores más altos respecto a las variables **est, lpie, aes, lrt** y uno de los más altos respecto a las variables **pes** y **lbr**. La primera componente está altamente influenciada por estas variables.
- El estudiante 20 tiene los valores más bajos respecto a las variables **est, lpie, lbr** y **lrt**. La primera componente está altamente influenciada por estas variables.
- El componente principal 1 separa los datos en dos grupos.

Gráfico de individuos coloreado por grupos (Sexo)



A la derecha se observan los estudiantes con género masculino y a la izquierda los estudiantes con género femenino

Círculo de correlaciones: La correlación entre una variable y un componente principal se utiliza como las coordenadas de la variable en el componente principal



- Las variables correlacionadas positivamente se agrupan.
- Las variables correlacionadas negativamente se colocan en lados opuestos del origen del gráfico (cuadrantes opuestos).
- Las variables que están alejadas del origen están bien representadas por los componentes principales.
- Entre más paralelo a un eje de una PC es un vector, más contribuye solo a esa PC.