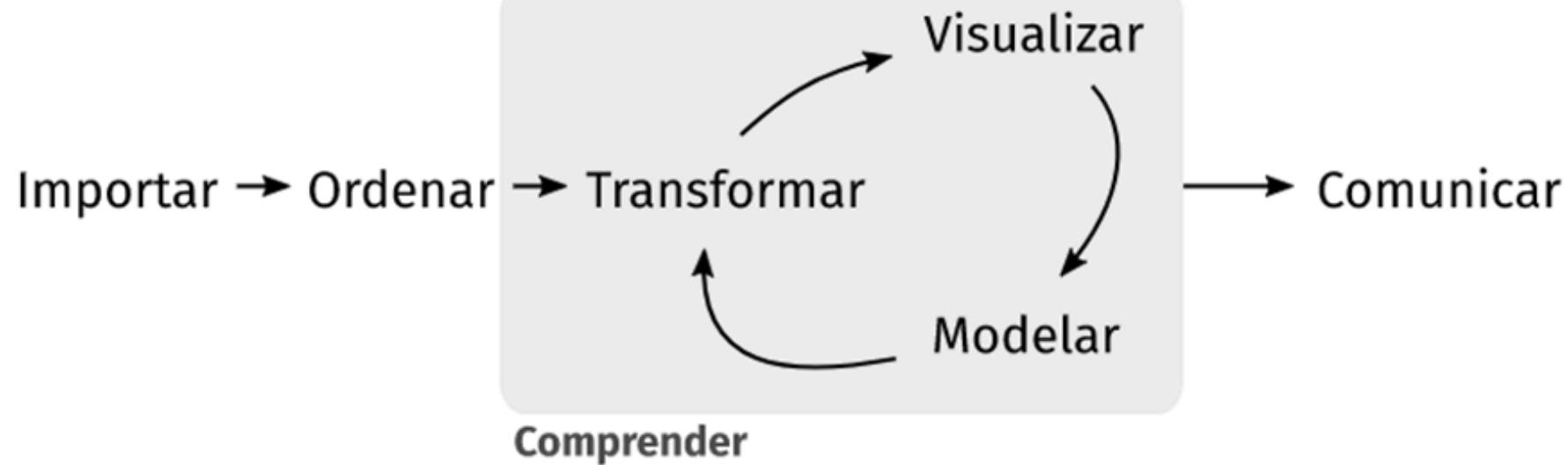


Analítica de datos aplicada a estudios sobre desarrollo

Giovany Babativa, PhD

Proceso de analítica



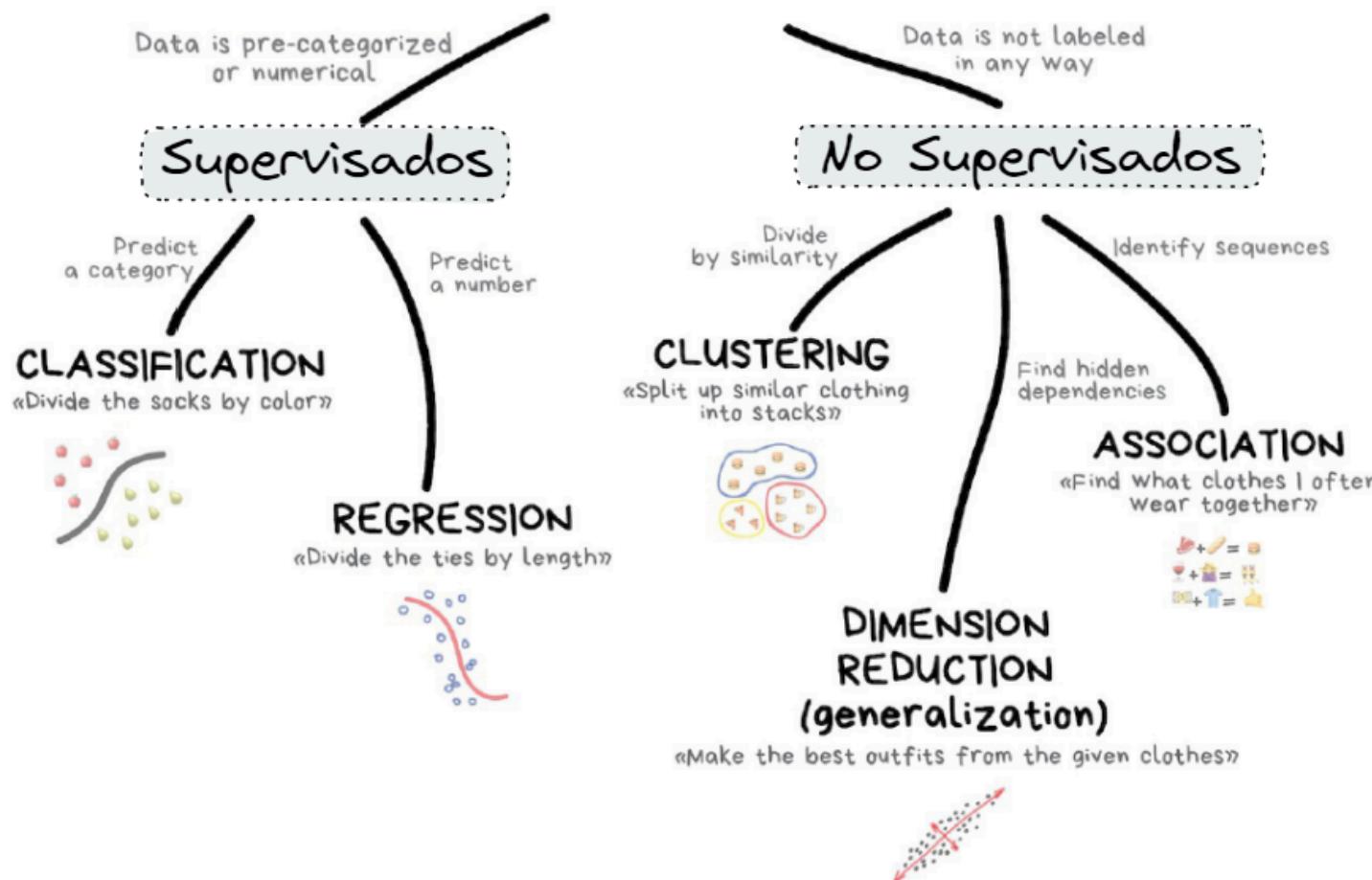
Programar

Wickham, H. y otros (2023)

MÉTODOS MULTIVARIANTES

Modelos de analítica

Modelos Multivariantes

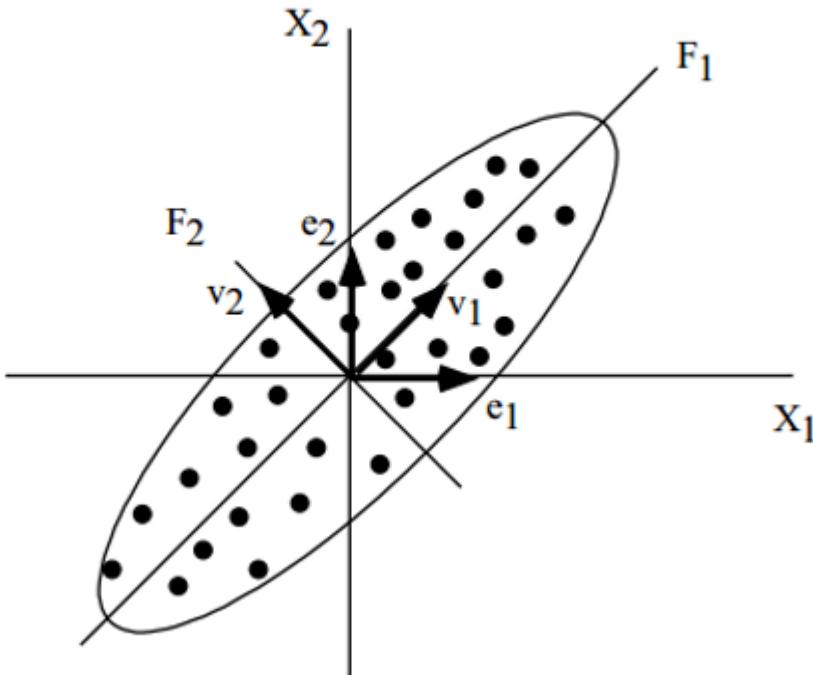


Fuente: Machine Learning for Everyone

ANÁLISIS DE COMPONENTES PRINCIPALES

Objetivo general

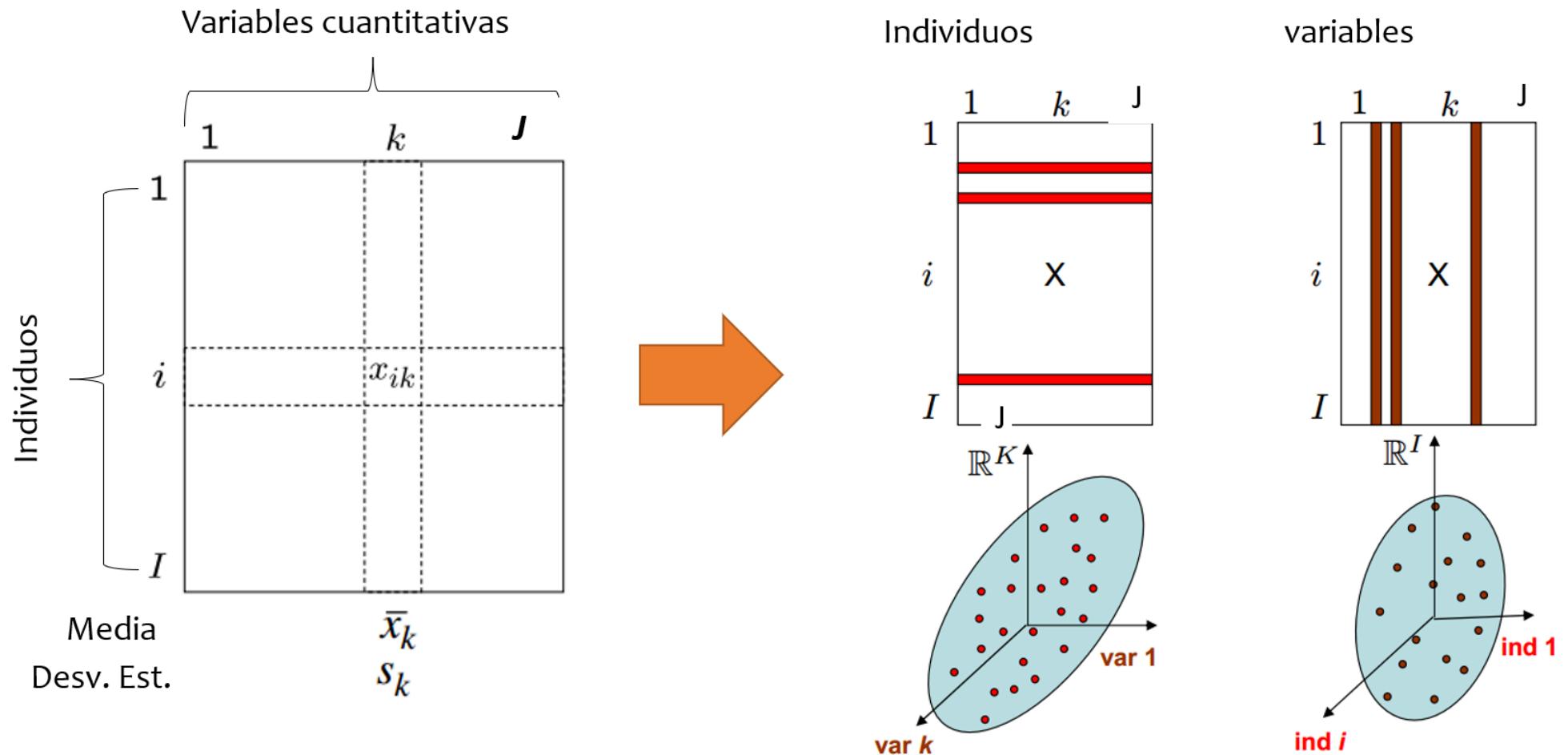
Método para reducir la dimensionalidad de los datos conservando la mayor cantidad de información. El método se debe usar cuando las variables son cuantitativas y existe presencia de correlación



Objetivos específicos

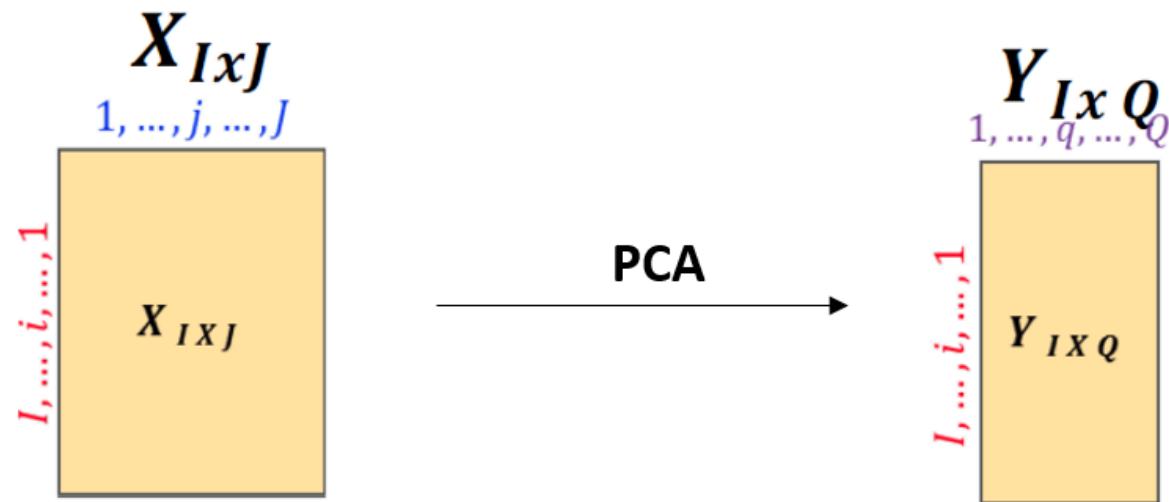
- **Visualizar patrones:** Sirve para visualizar la estructura de los datos y detectar patrones emergentes.
- **Construir de índices sintéticos:** Las variables originales se resumen en un conjunto menor de componentes principales que contienen información de todas las variables.
- **Identificar factores clave:** Identifica los factores principales que explican los cambios relacionados con el tema de interés.
- **Identificar grupos:** Ayuda a identificar grupos de individuos que comparten características similares.

¿Cómo funciona la técnica?



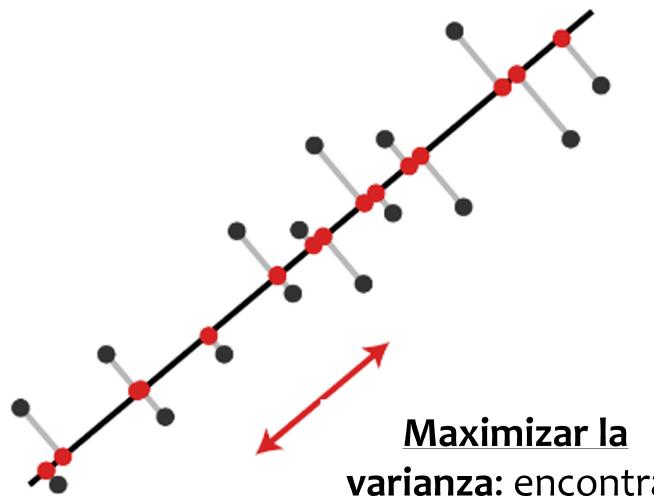
Componentes principales

Reproducir la matriz original usando menos dimensiones

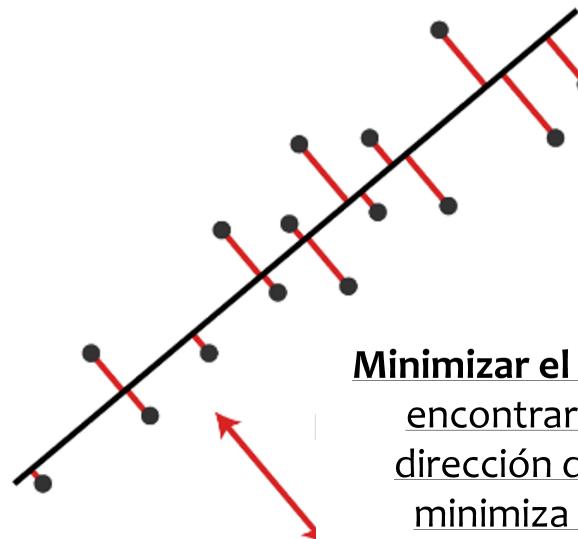


Problema de optimización

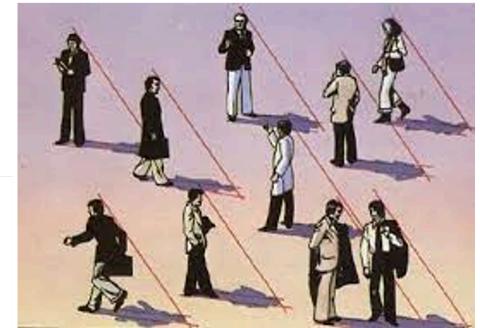
Enfoque de Hotelling (1933) vs Pearson (1901)



Maximizar la varianza: encontrar la dirección donde los puntos rojos tienen la mayor varianza.



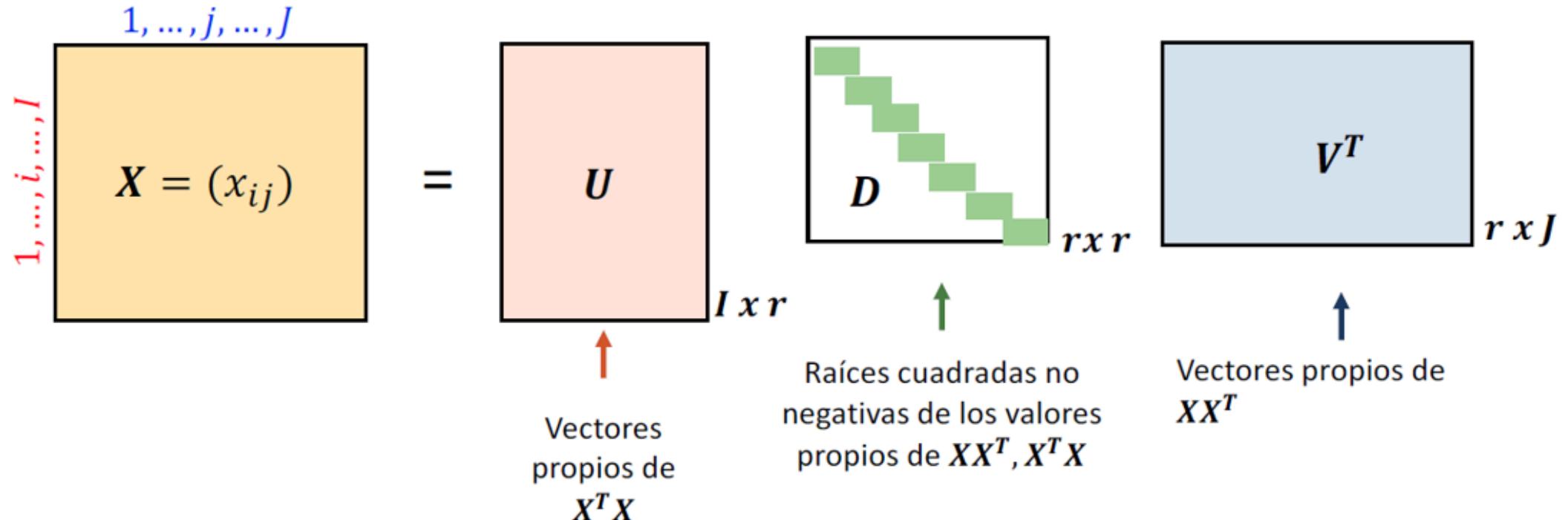
Minimizar el ECM: encontrar la dirección que minimiza la proyección de los puntos en un subespacio de menor dimensión.



Teorema de la factorización - SDV

Reproducir la matriz original con menos dimensiones.

Descomposición en Valores Singulares de una matriz X (Eckart y Young, 1936)



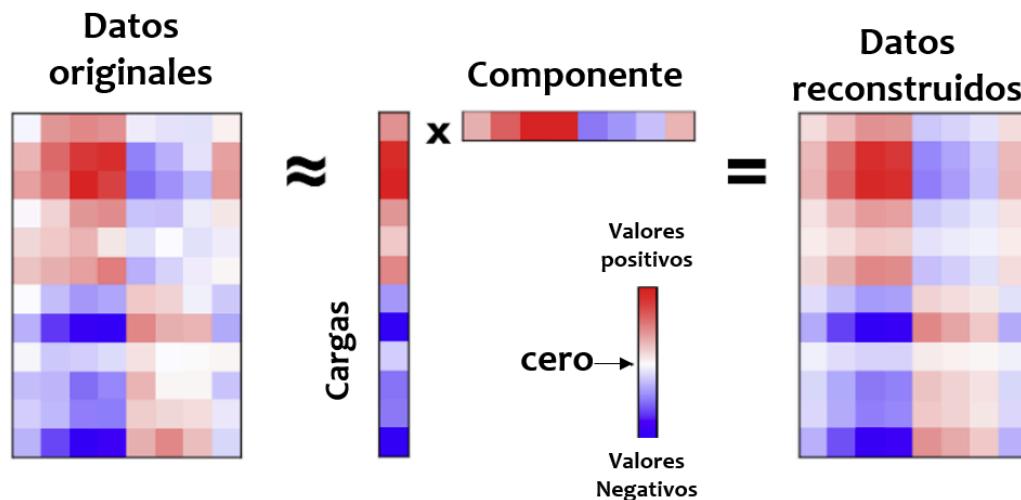
Esquema de las componentes

Descomposición en Valores Singulares

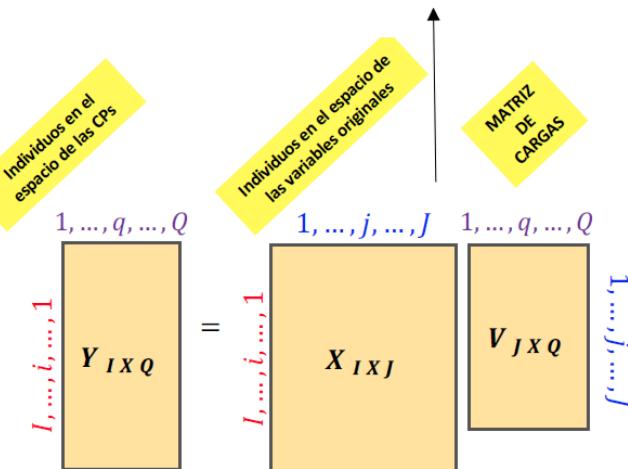
(Eckart&Young, 1936)

$$X = UDV^T = YV^T$$

Reconstrucción con una CP



Cada componente es una combinación lineal de las variables originales



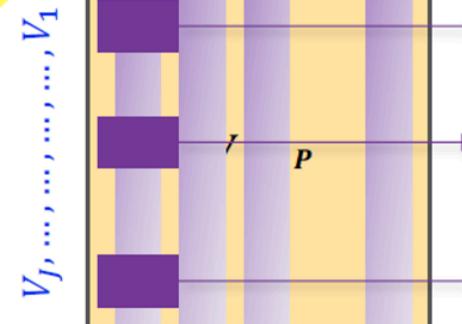
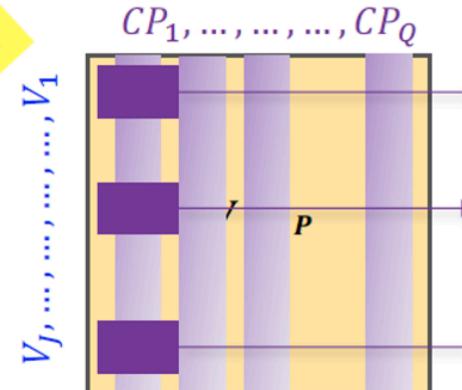
Resumen

Descomposición en Valores Singulares
(Eckart&Young, 1936)

$$X = UDV^T = YV^T$$

$$\begin{matrix} 1, \dots, q, \dots, Q \\ I, \dots, i, \dots, 1 \end{matrix} \quad Y_{I \times Q} = \begin{matrix} 1, \dots, j, \dots, J \\ I, \dots, i, \dots, 1 \end{matrix} \quad X_{I \times J} \quad \begin{matrix} 1, \dots, q, \dots, Q \\ 1, \dots, j, \dots, J \end{matrix} \quad V_{J \times Q}$$

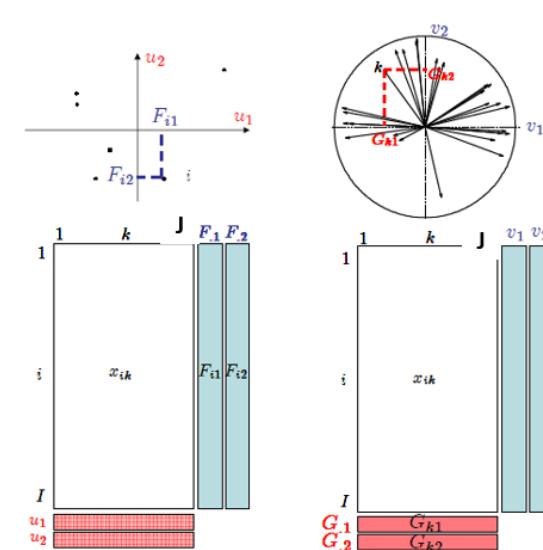
MATRIZ
DE
CARGAS



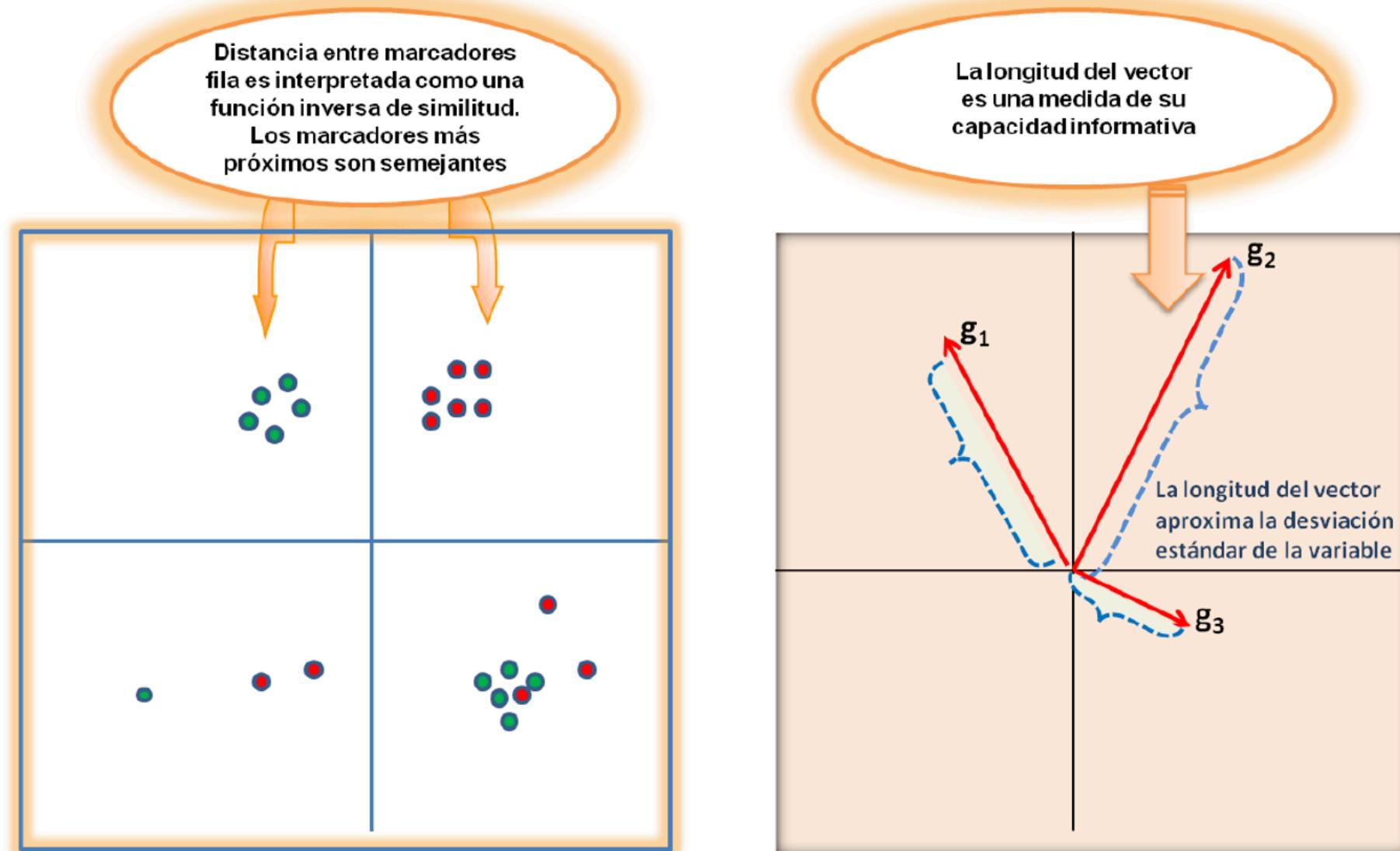
Contribución de la V_1
a la formación de la
PC1

Contribución de la V_j a
la formación de la PC1

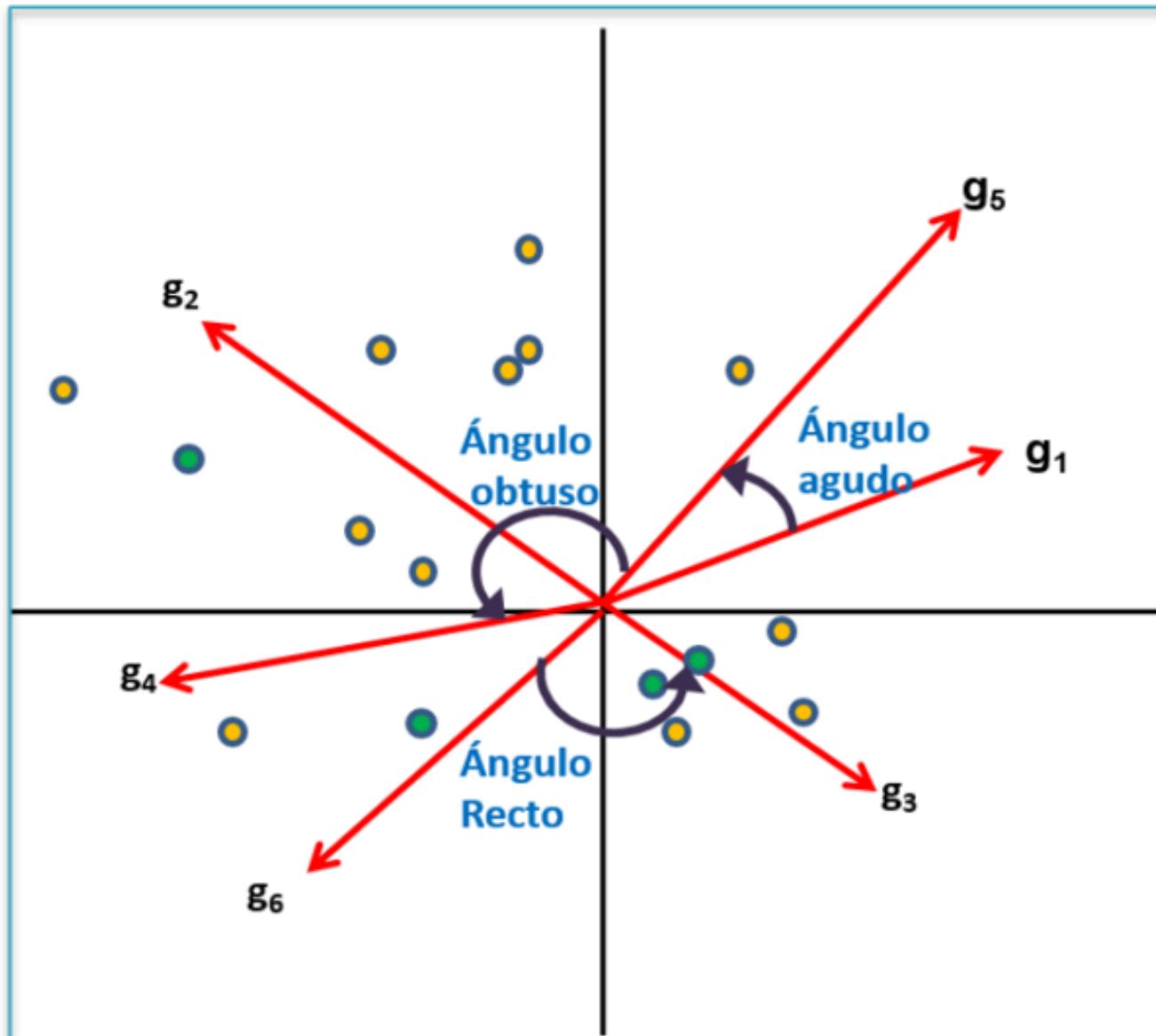
Contribución de la V_J
a la formación de la
PC1



Interpretación en el espacio de las componentes



Interpretación en el espacio de las componentes



Ejemplo

El conjunto de datos `RESUMEN.sav` contiene un preprocesamiento obtenido de la GEIH del DANE a nivel departamental para algunas variables de interés.

```
1 library(pacman)
2 p_load(tidyverse, janitor,
3         FactoMineR, factoextra, Factoshiny,
4         skimr, corrplot, psych, gt, gtsummary, haven)
5
6 url <- "https://github.com/jgbabativam/AnaDatos/raw/main/datos/RESUMEN.sav"
7 datos <- read_sav(url) |> as_factor()
```

Use el comando `glimpse()` y `skim()` para explorar el conjunto de datos.

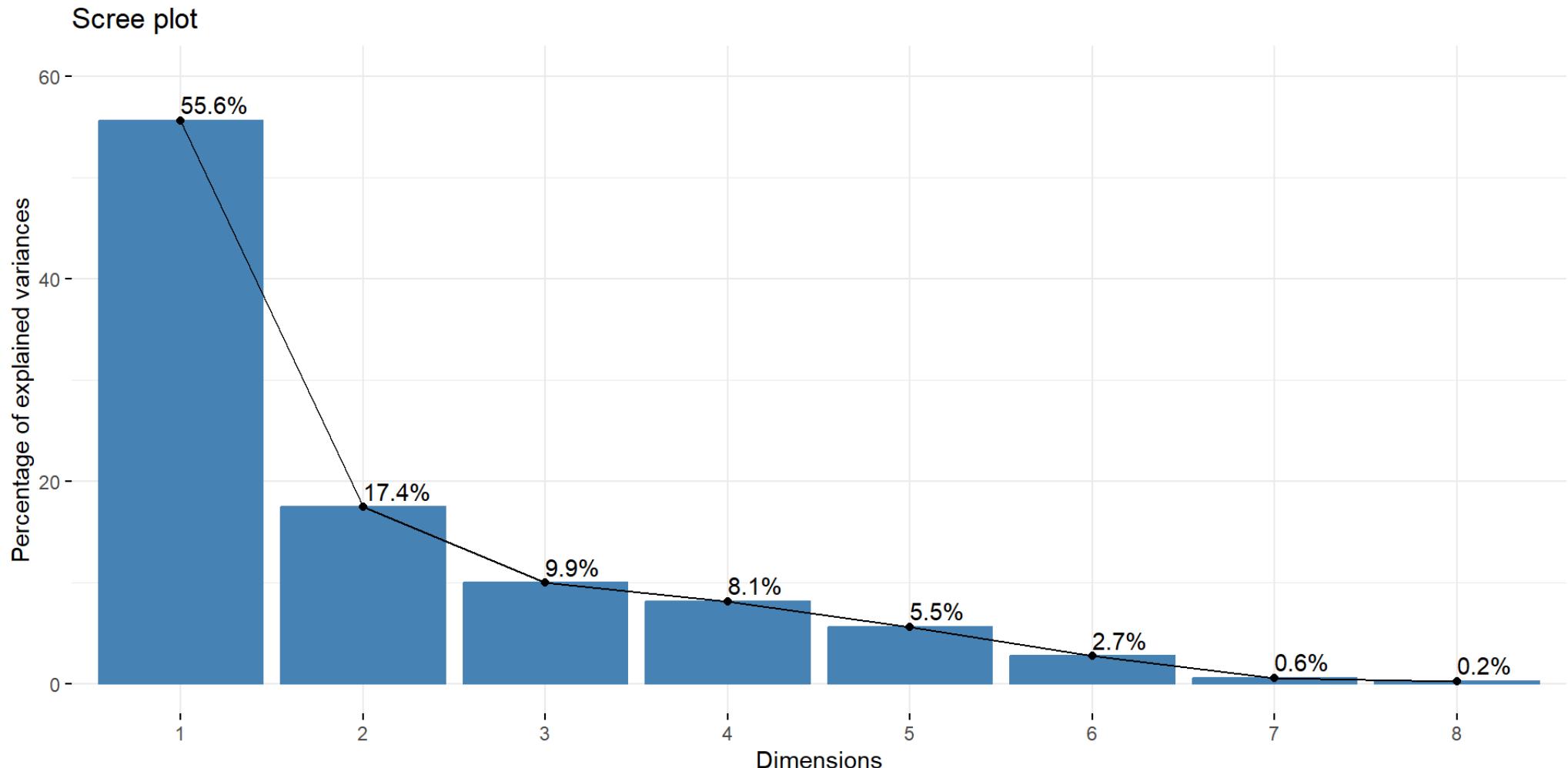
Preparación del conjunto de datos

```
1 datos <- datos |>  
2     column_to_rownames(var = "DPTO")
```

- Use la función `Factoshiny(datos)` y ajuste los parámetros del modelo.
- Explore el peso de las variables mediante la función `PCA(datos)` del paquete `FactoMineR`.

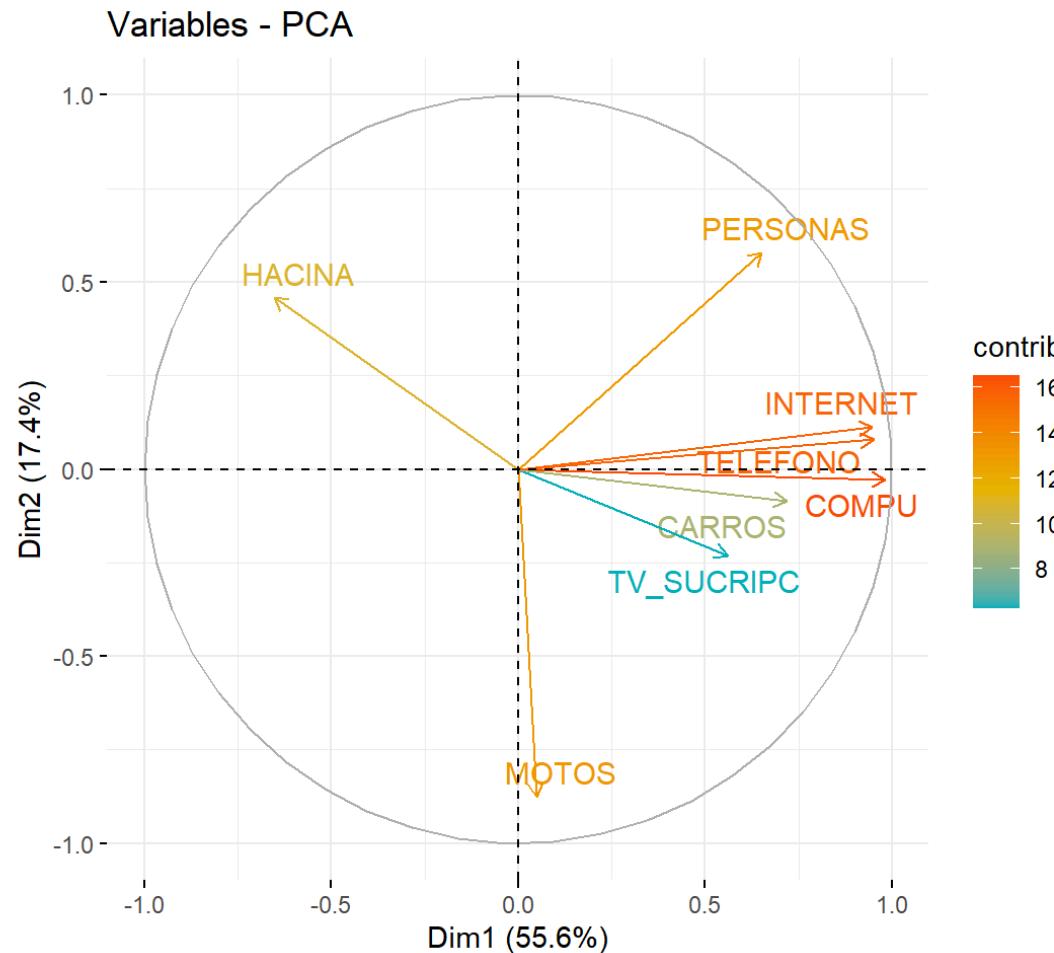
Analice los resultados

```
1 res <- PCA(datos, scale.unit = T, graph = F)
2 fviz_screepplot(res, addlabels = TRUE, ylim = c(0, 60))
```



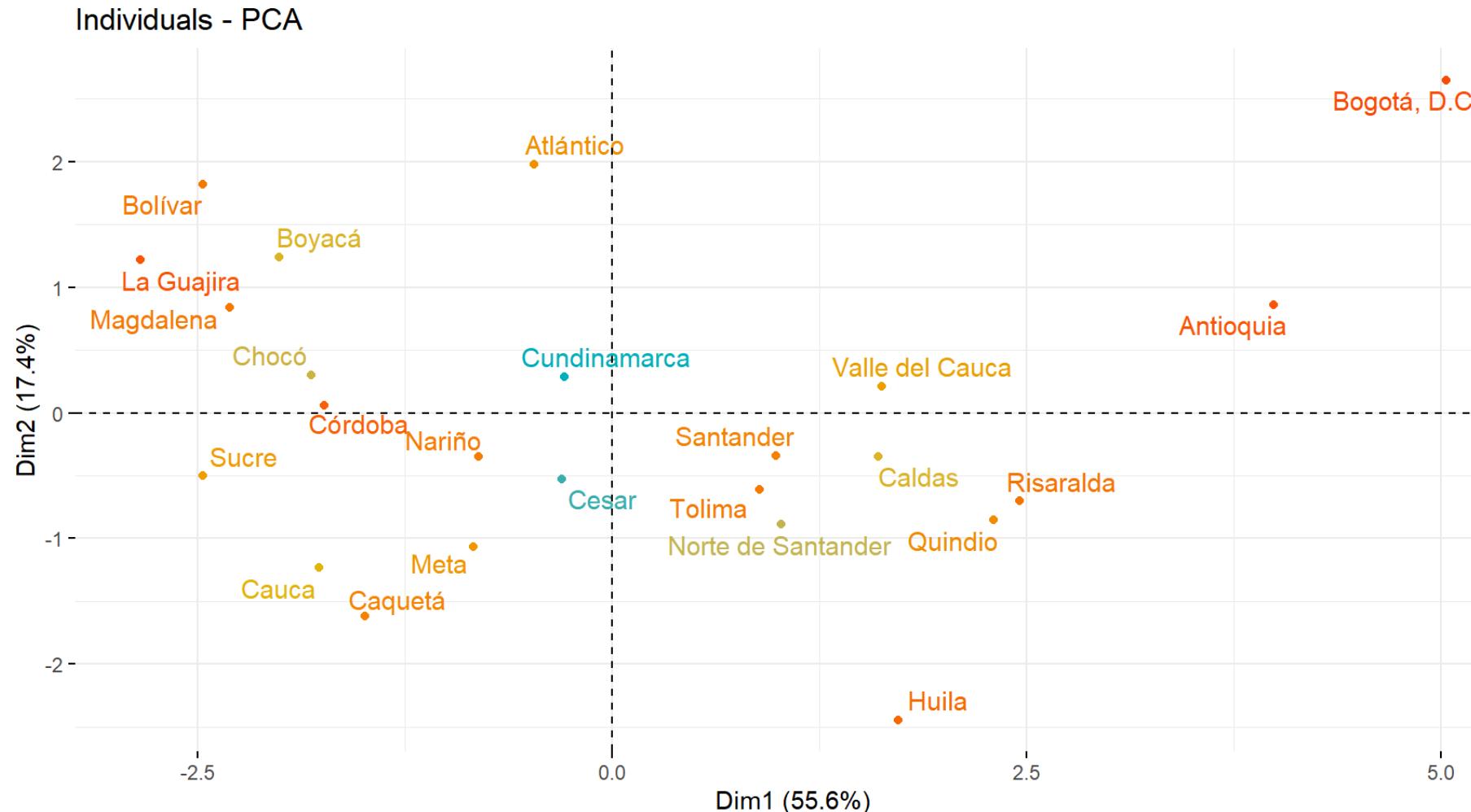
Primer plano factorial para las variables

```
1 fviz_pca_var(res,  
2                 col.var="contrib",  
3                 gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),  
4                 repel = TRUE)
```



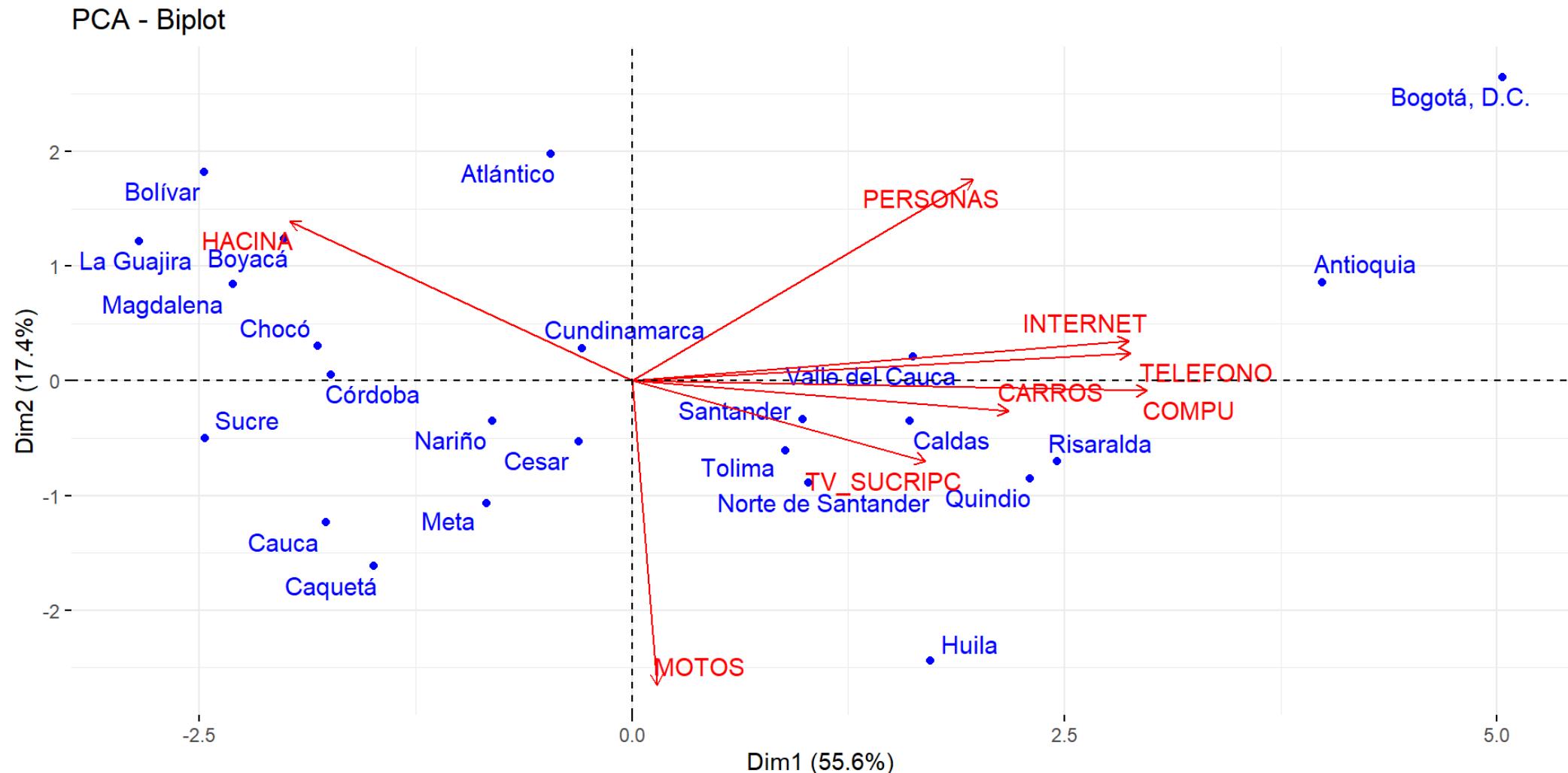
Primer plano factorial para los individuos

```
1 fviz_pca_ind(res, col.ind = "cos2",
2                         gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
3                         repel = TRUE)
```



Biplot

```
1 fviz_pca_biplot(res, repel = TRUE, col.ind = "blue", col.var = "red")
```



Construcción de índices sintéticos

Tenga en cuenta que:

$$\mathbf{Y} = \mathbf{X}\mathbf{V}$$

De manera que la matriz \mathbf{V} son los ponderadores de las variables en la matriz \mathbf{X} , con lo cual \mathbf{Y} es un índice que resume la información contenida en las variables originales.

ANÁLISIS DE CORRESPONDENCIAS

Análisis de correspondencias

Mientras que el PCA se usa para tratar variables cuantitativas que tienen algún grado de asociación lineal, el *análisis de correspondencias* es un método que surge de las tablas de contingencia y permite estudiar las relaciones entre variables nominales. Este análisis permite:

- Identificar patrones de asociación entre variables categóricas.
- Hacer una reducción de la dimensionalidad.
- Observar la proximidad entre individuos y entre variables.
- Hacer un pre-procesamiento para el análisis de clúster.

Estructura del conjunto de datos

encuesta	preocupaciones	estadocivil	sexo
1	El dinero	Soltero	mujer
2	El dinero	Soltero	mujer
3	Armonía fa...	Casado	hombre
4	Armonía fa...	Separado/...	mujer
5	Su salud	Soltero	mujer
6	Su salud	Separado/...	hombre
7	Su salud	Soltero	hombre
8	El dinero	Soltero	mujer
9	El dinero	Separado/...	mujer
10	Su salud	Soltero	mujer
11	Armonía fa...	Soltero	hombre
12	Su salud	Soltero	mujer
13	El dinero	Casado	hombre
14	Su salud	Separado/...	hombre
15	El dinero	Soltero	mujer
16	Su vida afe...	Separado/...	mujer
17	Armonía fa...	Soltero	hombre
18	Su salud	Casado	hombre
19	Su salud	Soltero	mujer
20	Su vida afe...	Soltero	hombre
21	El dinero	Soltero	mujer

Funcionamiento del análisis de correspondencias

$X =$

	variable 1	variable j	variable J
1			
i	0 1 0 0 0	x_{ik}	0 0 1 0
I			

Individuos

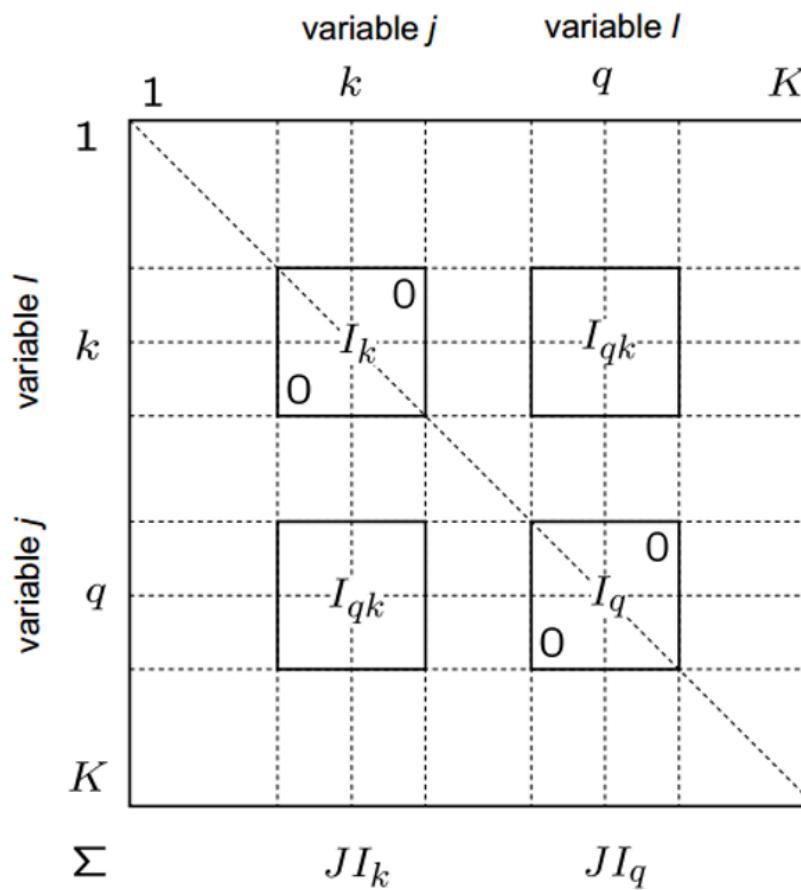
La suma por columna es I_k .
 $\#$ de individuos que contestaron la modalidad k en la variable j

La suma por fila da J .
 $\#$ de preguntas

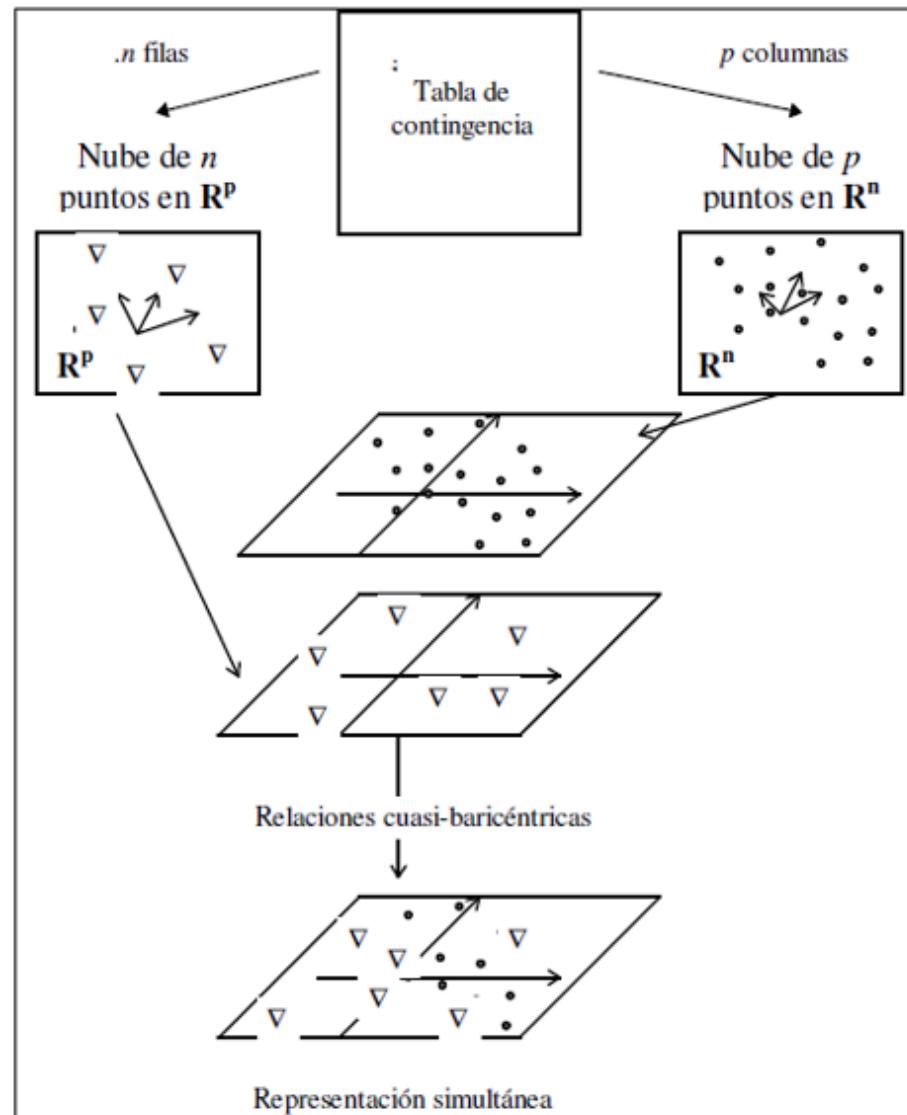
El total es $I * J$

Generación de tablas de contingencia

Al realizar la operación $\mathbf{X}^T \mathbf{X}$ se llega a la matriz que concatena todas las tablas de contingencia entre pares de variables, denominada matriz de *Burt*



Proyección sobre el espacio factorial



Ejemplo

El conjunto de datos `corresp.sav` contiene 50 respuestas de una encuesta.

```
1 library(pacman)
2 p_load(tidyverse, janitor, patchwork,
3         FactoMineR, factoextra, Factoshiny,
4         skimr, corrplot, psych, gt, gtsummary, haven)
5
6 url <- "https://github.com/jgbabativam/AnaDatos/raw/main/datos/corresp.sav"
7 datos <- read_sav(url) |> as_factor()
```

Use `glimpse()` y `skim()` para explorar el conjunto de datos.

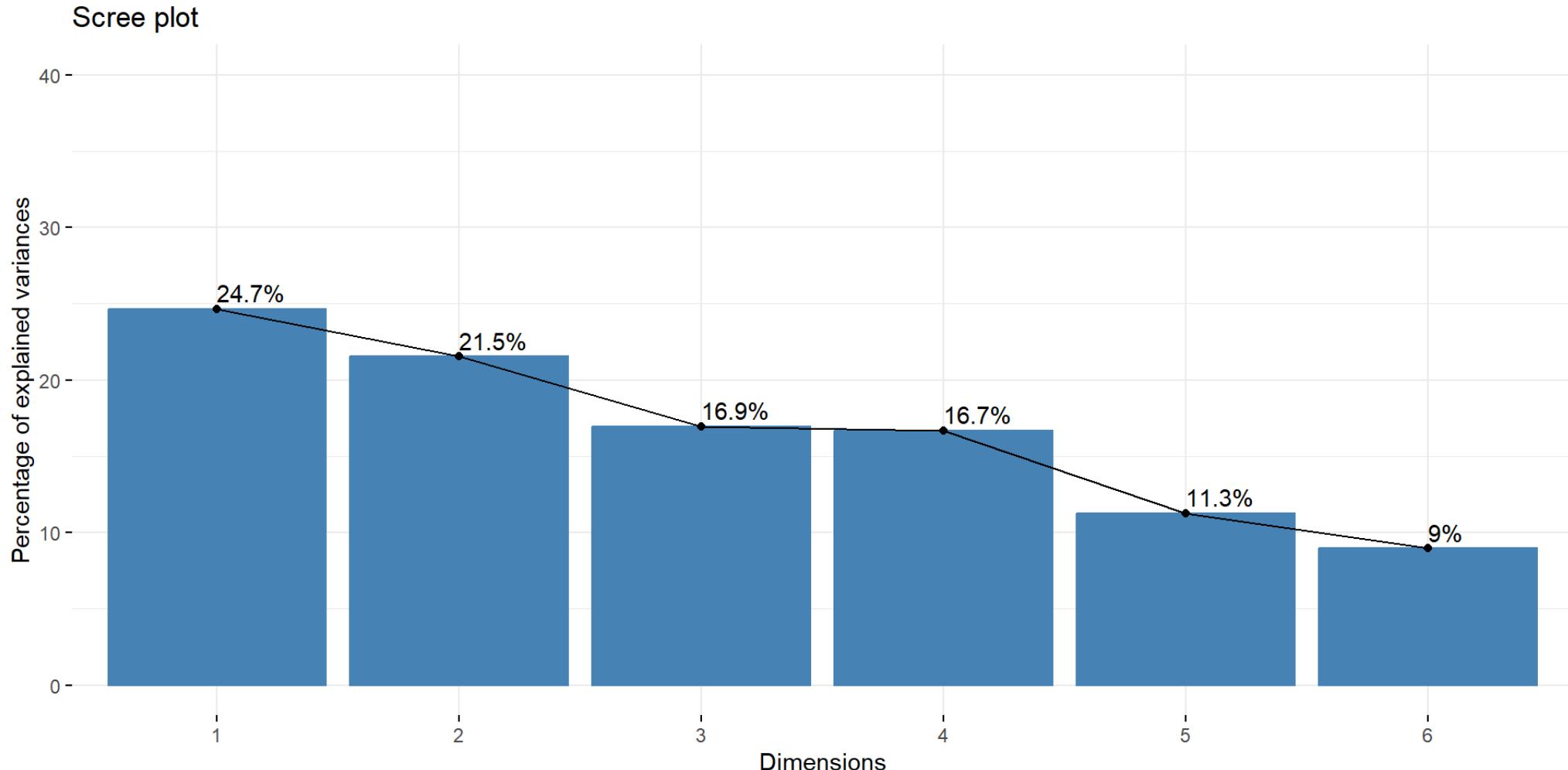
Preparación de los datos

```
1 datos <- datos |>  
2     column_to_rownames(var = "encuesta")
```

- Use la función `Factoshiny(datos)` y ajuste los parámetros del modelo.
- Explore la contribución y el coseno al cuadrado usando `MCA(datos)` del paquete `FactoMineR`.

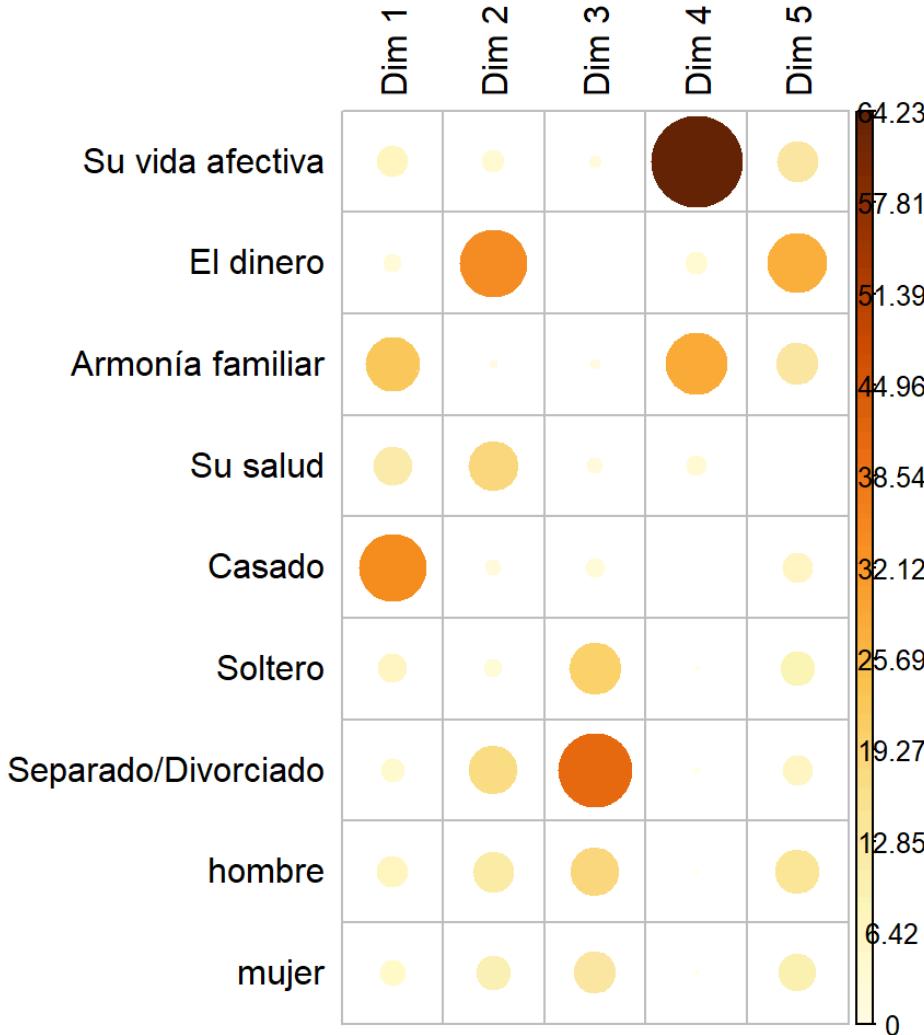
Analice los resultados

```
1 res <- MCA(datos, graph = F)
2 fviz_screepplot(res, addlabels = TRUE, ylim = c(0, 40))
```



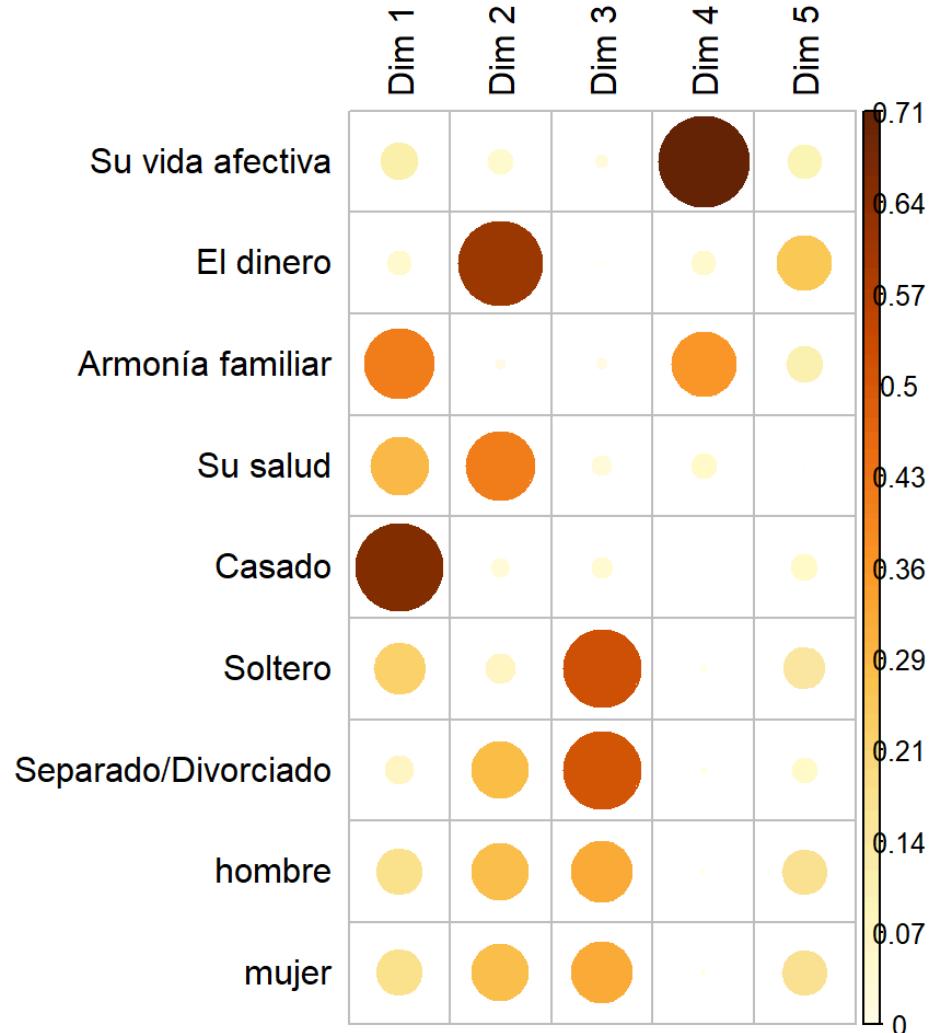
Análisis de las contribuciones

```
1 corrplot(res$var$contrib, is.corr=FALSE, tl.col = "black")
```



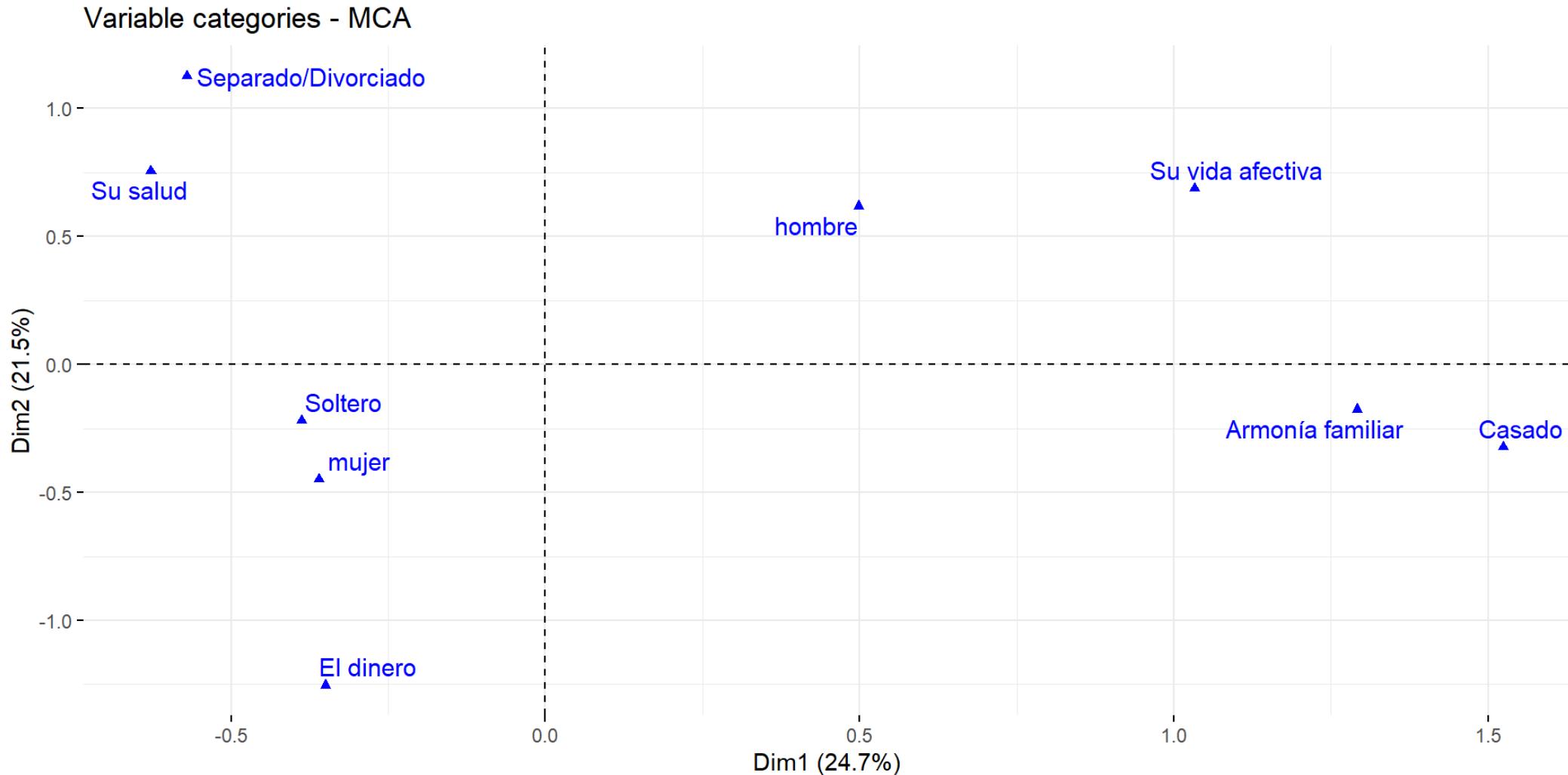
Análisis de los cosenos

```
1 corrplot(res$var$cos2, is.corr=FALSE, tl.col = "black")
```



Primer plano factorial

```
1 fviz_mca_var(res, repel = TRUE, col.var = "blue")
```



GRACIAS!

Referencias

- Husson, F., Lê, S., & Pagès, J. (2017). Exploratory multivariate analysis by example using R. CRC press.
- Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). Multivariate data analysis 6th Edition.
<https://doi.org/10.1201/9780367409913>
- Aldás Manzano, J., & Uriel Jiménez, E. (2017). Análisis multivariante aplicado con R. Ediciones Paraninfo, SA.

