

TÉCNICAS BILOT PARA EL ANÁLISIS AVANZADO DE DATOS MULTIVARIANTES

CAPACITADORES:



PhD. Purificación Vicente Galindo.
Universidad de Salamanca.

- Directora del Departamento de Estadística
- Coordinadora del Programa de Doctorado en Estadística Multivariante Aplicada.



PhD. Giovany Babativa Márquez.
Consultor e Investigador

- Master en Análisis de Datos y Big Data
- Doctor en Estadística Multivariante Aplicada.
- Consultor Estadístico en el Sector Público y Privado en Colombia



<http://jgbabativam.rbind.io/>



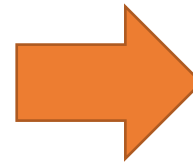
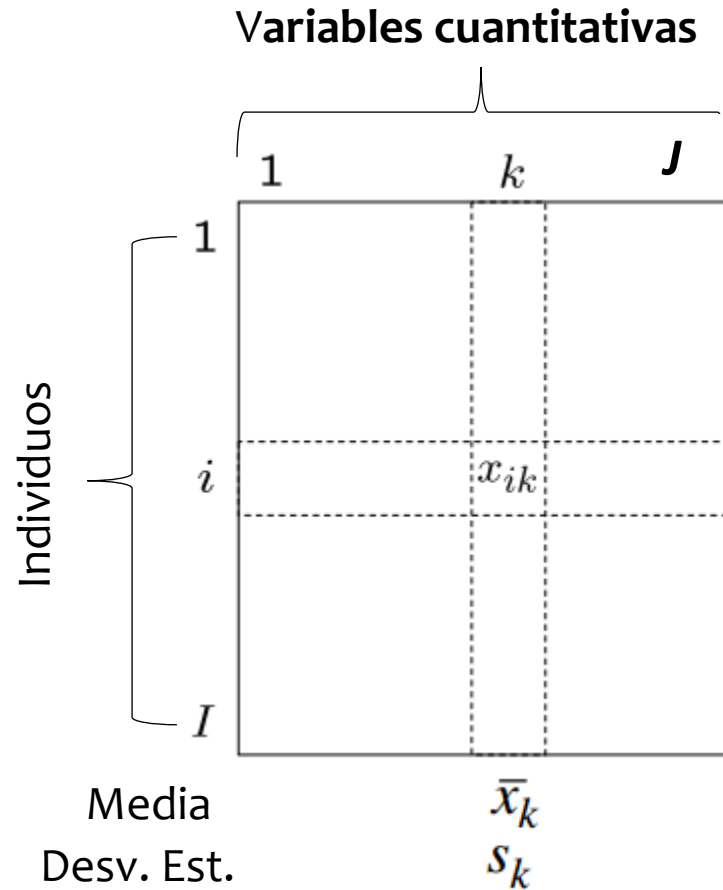
<https://scholar.google.es/citations?user=2NJRN8A8AAJ&hl=es>



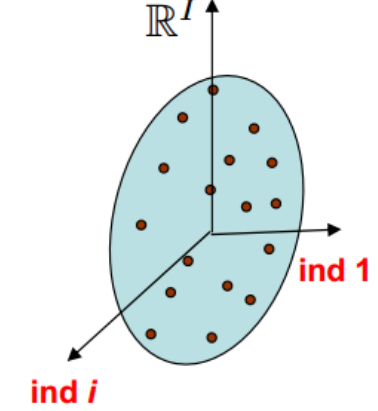
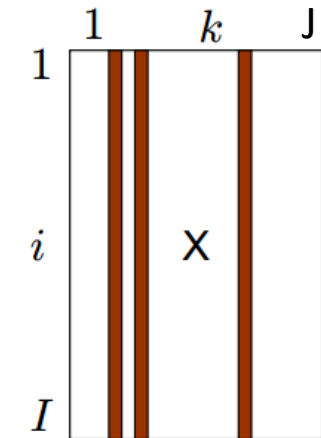
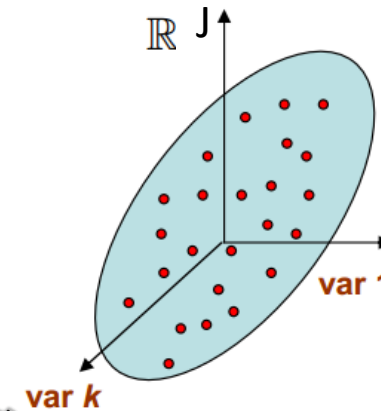
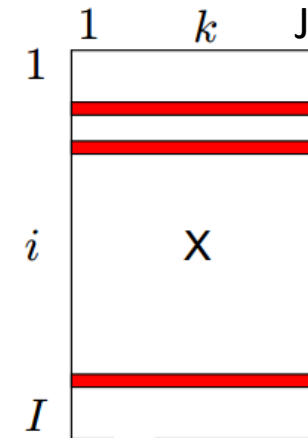
<https://github.com/jgbabativam>



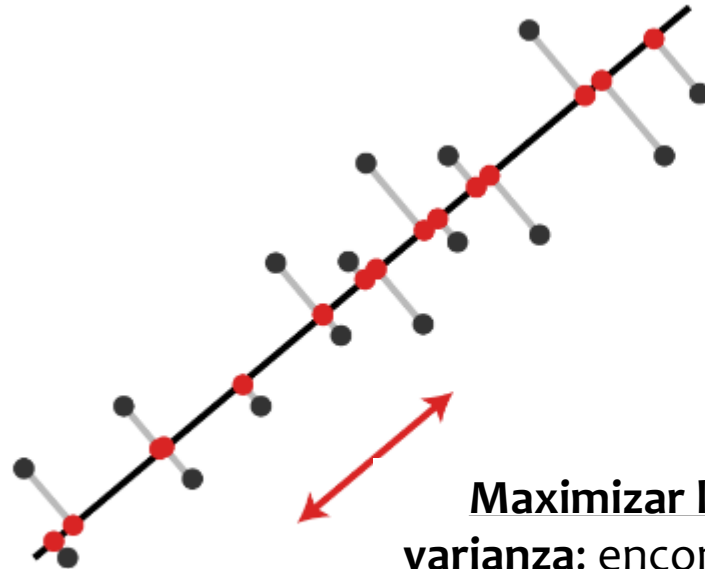
jgbabativam@unal.edu.co



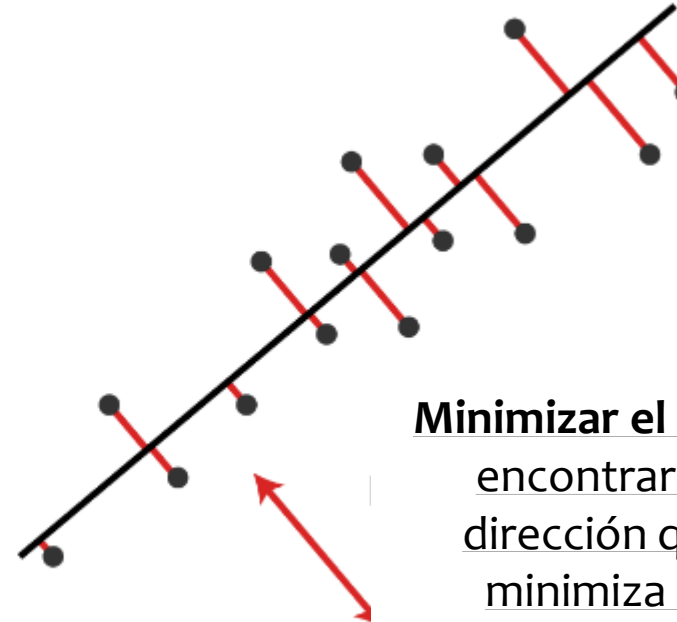
Individuos



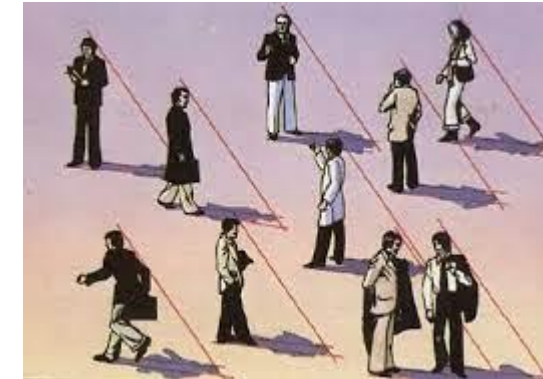
Problema de optimización



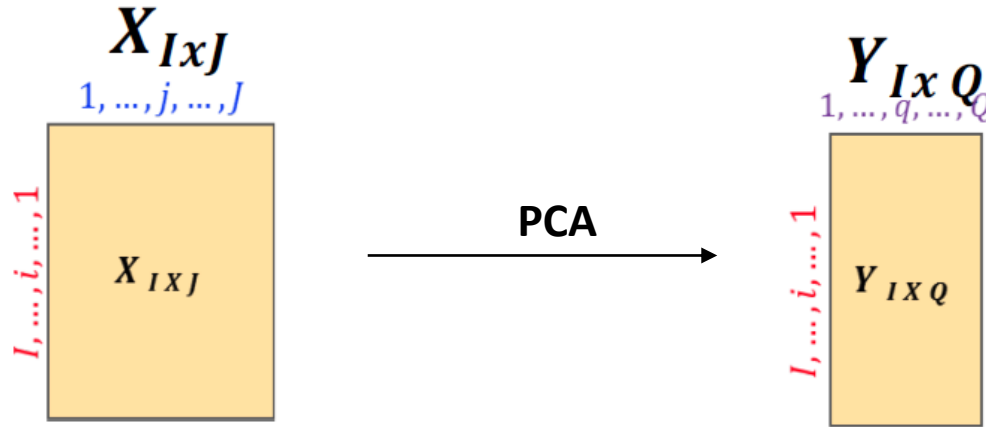
Maximizar la varianza: encontrar la dirección donde los puntos rojos tienen la mayor varianza.



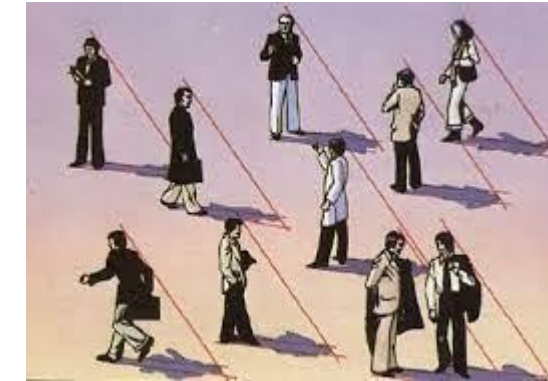
Minimizar el ECM: encontrar la dirección que minimiza la proyección de los puntos en un subespacio de menor dimensión.



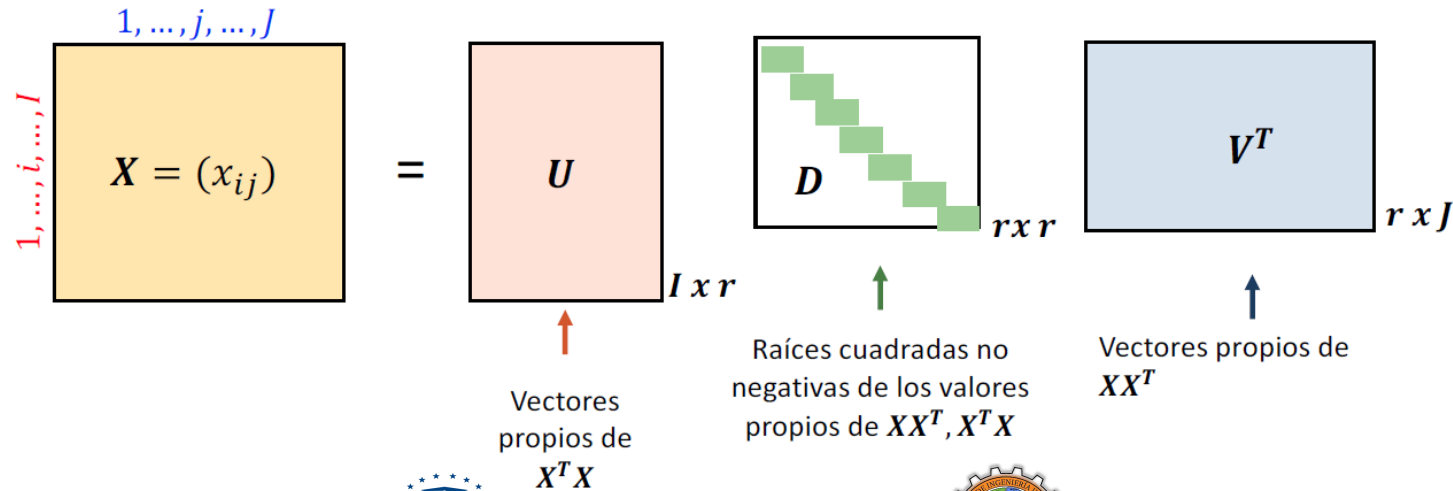
Reproducir la matriz original con menos dimensiones



Direcciones de máxima variabilidad



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

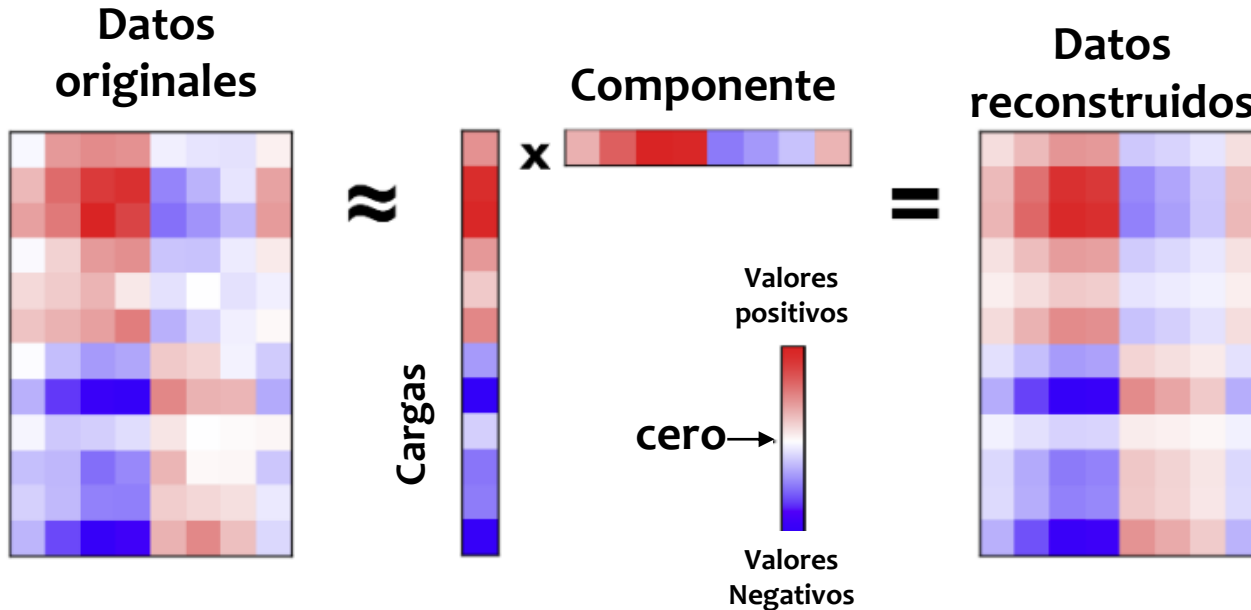


Minimizar el ECM... SDV

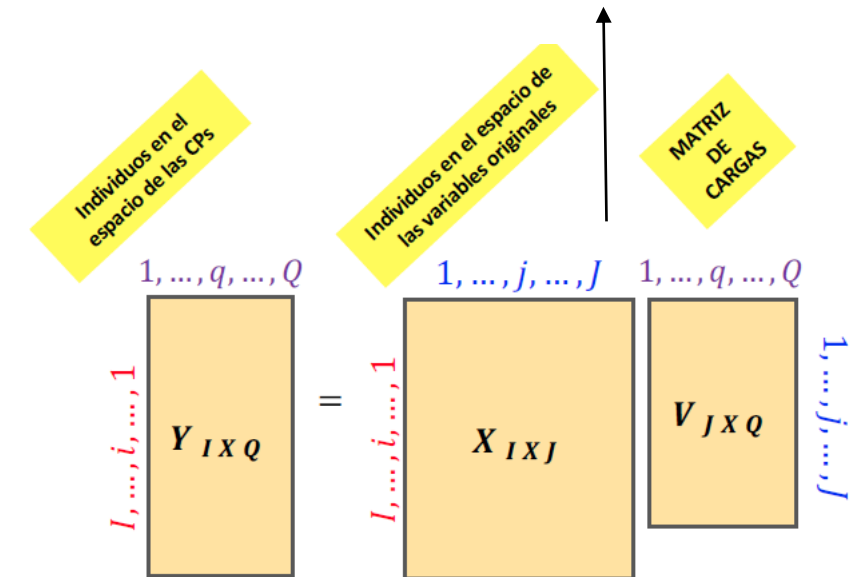
Descomposición en Valores Singulares (Eckart&Young, 1936)

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

Reconstrucción con una CP



Cada componente es una combinación lineal de las variables originales

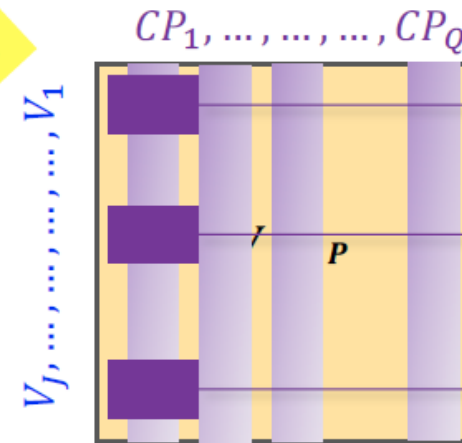


Descomposición en Valores Singulares (Eckart&Young, 1936)

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

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

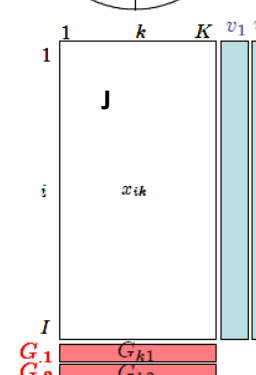
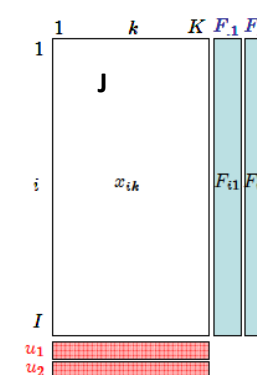
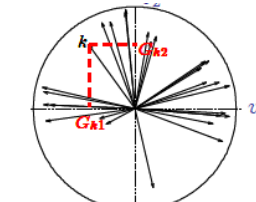
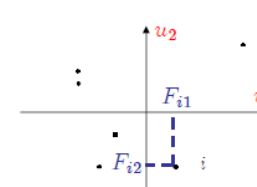
MATRIZ DE CARGAS



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

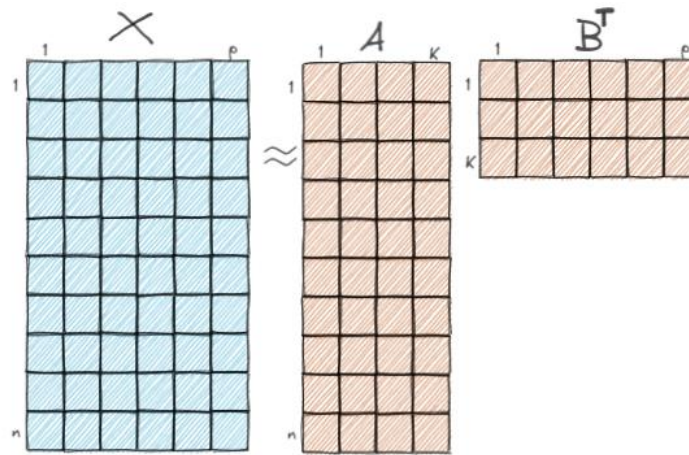


El BILOT aproxima la distribución de una muestra multivariante en un espacio de dimensión reducida, normalmente de dimensión dos, y superpone sobre la misma representaciones de las variables sobre las que se mide la muestra (GOWER 1996). Las representaciones de las variables son normalmente vectores, y coinciden con las direcciones en las que mejor se muestra el cambio individual de cada variable.

El prefijo "bi" se refiere a la superposición, en la misma representación, de individuos y variables.

[Carl Sagan: La cuarta dimensión](#)

Sea \mathbf{X} una matriz de datos (centrada y podría estar estandarizada) que contiene las medidas de n individuos en p variables. Un biplot en dimensión q es una representación gráfica mediante marcadores $\mathbf{A}_{n \times k}$ y $\mathbf{B}_{k \times q}$ (puntos o vectores) para las filas y las columnas respectivamente, de forma que el producto \mathbf{AB}' aproxime \mathbf{X} tan bien como sea posible.



\mathbf{A} contiene un conjunto de n vectores k -dimensionales que representan a las filas y \mathbf{B} contiene un conjunto de p vectores k -dimensionales que representan a las columnas

*Para que la representación sea útil necesitamos imponer una métrica de forma que la descomposición y el biplot resultantes sean únicos. La métrica equivale a imponer restricciones sobre \mathbf{A} o \mathbf{B} , por ejemplo, que sean ortonormales ($\mathbf{B}'\mathbf{B}=\mathbf{I}$)

RESUMEN

$$X_{n \times p} = \begin{matrix} & 1 & \dots & p \\ \begin{matrix} 1 \\ \vdots \\ n \end{matrix} & \begin{matrix} \square & \dots & \square \\ \square & \dots & \square \\ \square & \dots & \square \end{matrix} \end{matrix}$$

Eckart y Young (1936)

$$X \approx U \Lambda V^T$$

Diagram showing the decomposition of matrix X into matrices U , Λ , and V^T .

$$X = AB' + E$$

Calidad de representación

Filas

JK-Biplot
Gabriel (1971)

$$A \approx U \Lambda$$

$$B^T \approx V^T$$

Diagram showing the decomposition of matrix A into matrices U and Λ , and matrix B^T into matrix V^T .

Columnas

GH-Biplot
Gabriel (1971)

$$A \approx U \Lambda$$

$$B^T \approx \Lambda V^T$$

Diagram showing the decomposition of matrix A into matrices U and Λ , and matrix B^T into matrices Λ and V^T .

Filas y Columnas

HJ-Biplot
Galindo (1986)

$$A \approx U \Lambda$$

$$B^T \approx \Lambda V^T$$

Diagram showing the decomposition of matrix A into matrices U and Λ , and matrix B^T into matrices Λ and V^T .

Mismo sistema de referencia

En el **HJ-BIPLLOT** los marcadores para las filas y para las columnas pueden ser representados en el mismo sistema de referencia con máxima calidad de representación (Galindo, 1985; Galindo y Cuadras, 1986)

Partimos de la descomposición en valores singulares de la matriz X

$$X = U\Lambda V^T$$

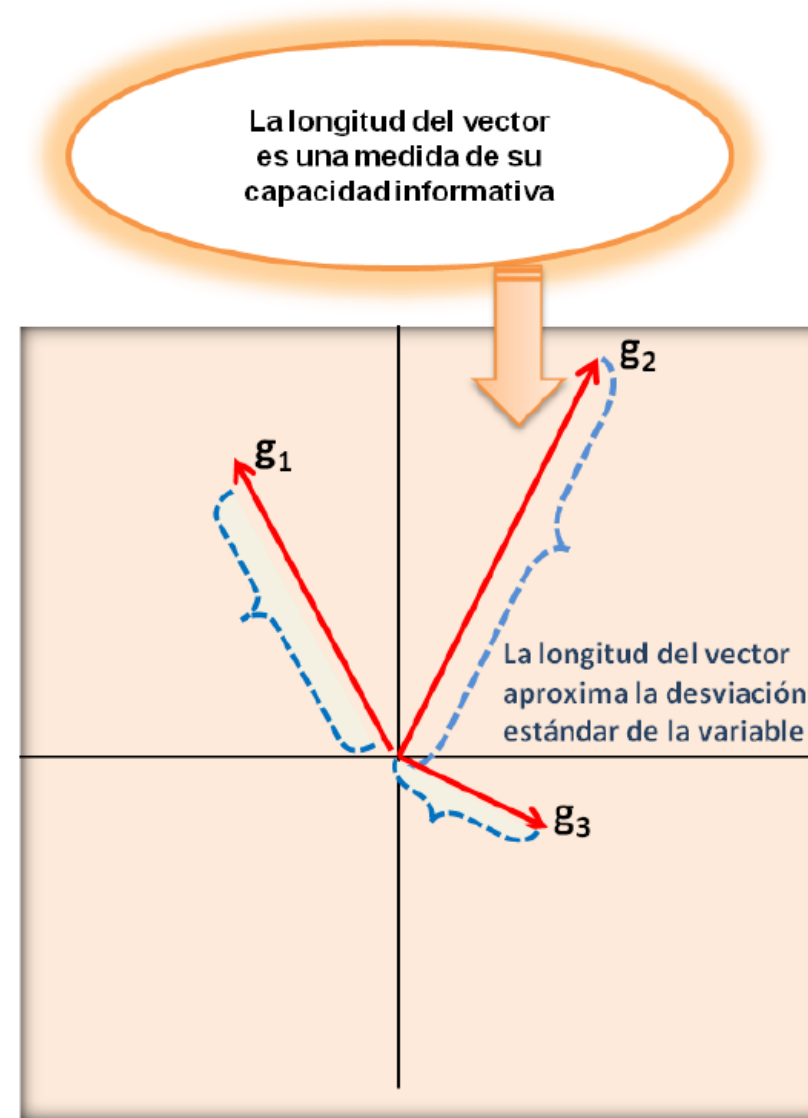
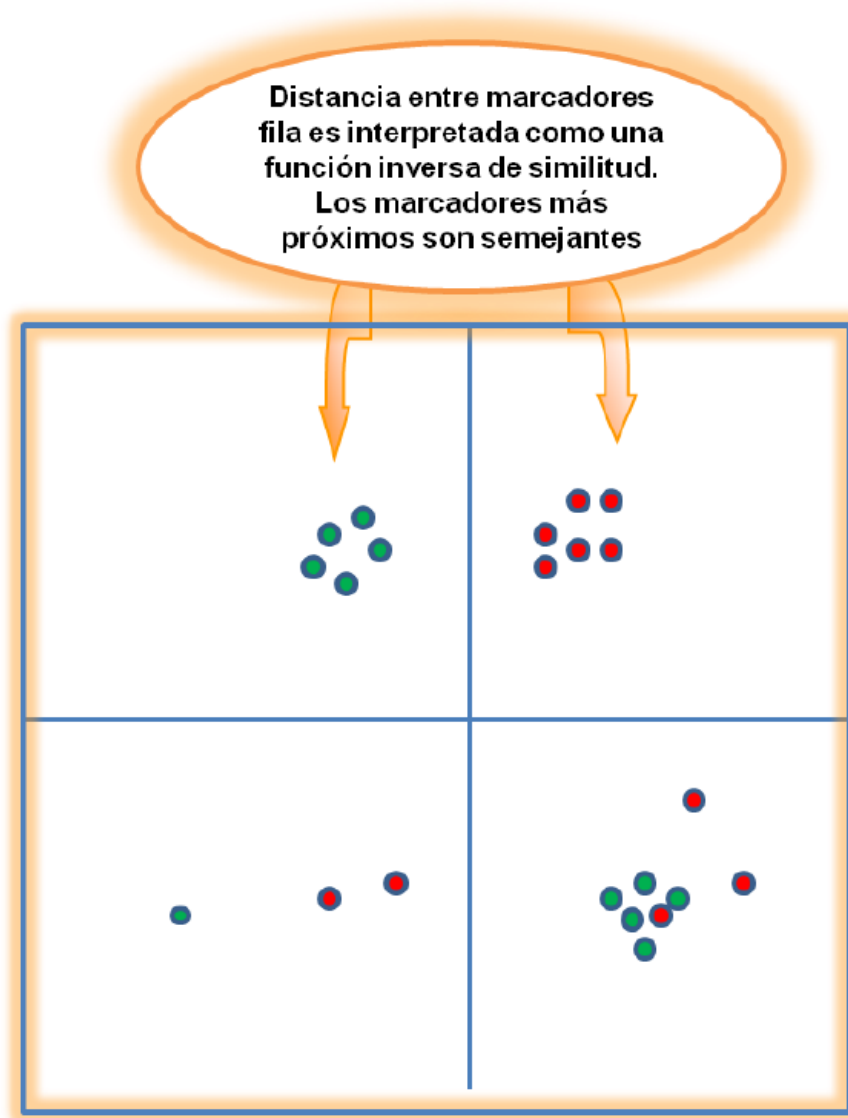


$$A = U_{(k)}\Lambda_{(k)}$$

$$B = V_{(k)}\Lambda_{(k)}$$

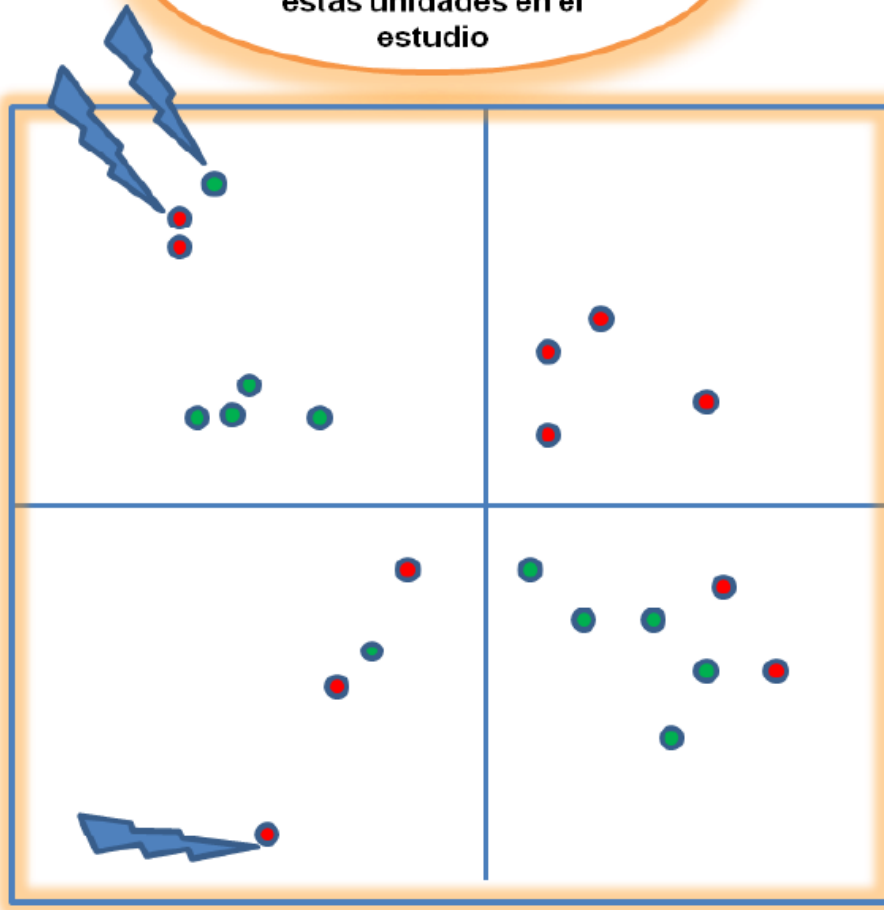
El HJ-Biplot no reproduce los elementos de la matriz X , pero tiene la ventaja de que es una representación simultánea que alcanza la máxima representación para las filas y columnas.

REPRESENTACIÓN SIMULTÁNEA	COORDENADAS FILAS	COORDENADAS COLUMNAS	BONDAD AJUSTE PARA FILAS	BONDAD AJUSTE PARA COLUMNAS
GH-BIPLLOT	U	VΛ	$\frac{2}{\bar{r}}$	$\frac{l_1^2 + l_2^2}{\sum_{a=1}^r \dot{a} l_a^2}$
JK-BIPLLOT	UΛ	V	$\frac{l_1^2 + l_2^2}{\sum_{a=1}^r \dot{a} l_a^2}$	$\frac{2}{\bar{r}}$
HJ-BIPLLOT	U Λ	VΛ	$\frac{l_1^2 + l_2^2}{\sum_{a=1}^r \dot{a} l_a^2}$	$\frac{l_1^2 + l_2^2}{\sum_{a=1}^r \dot{a} l_a^2}$

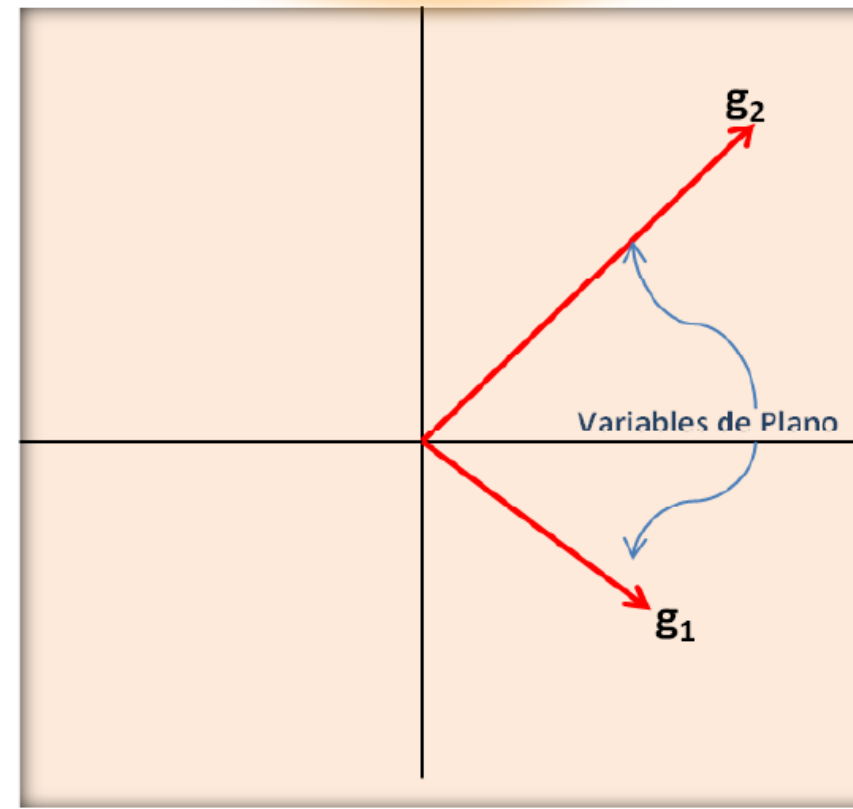


(Cubilla-Montilla, 2021)

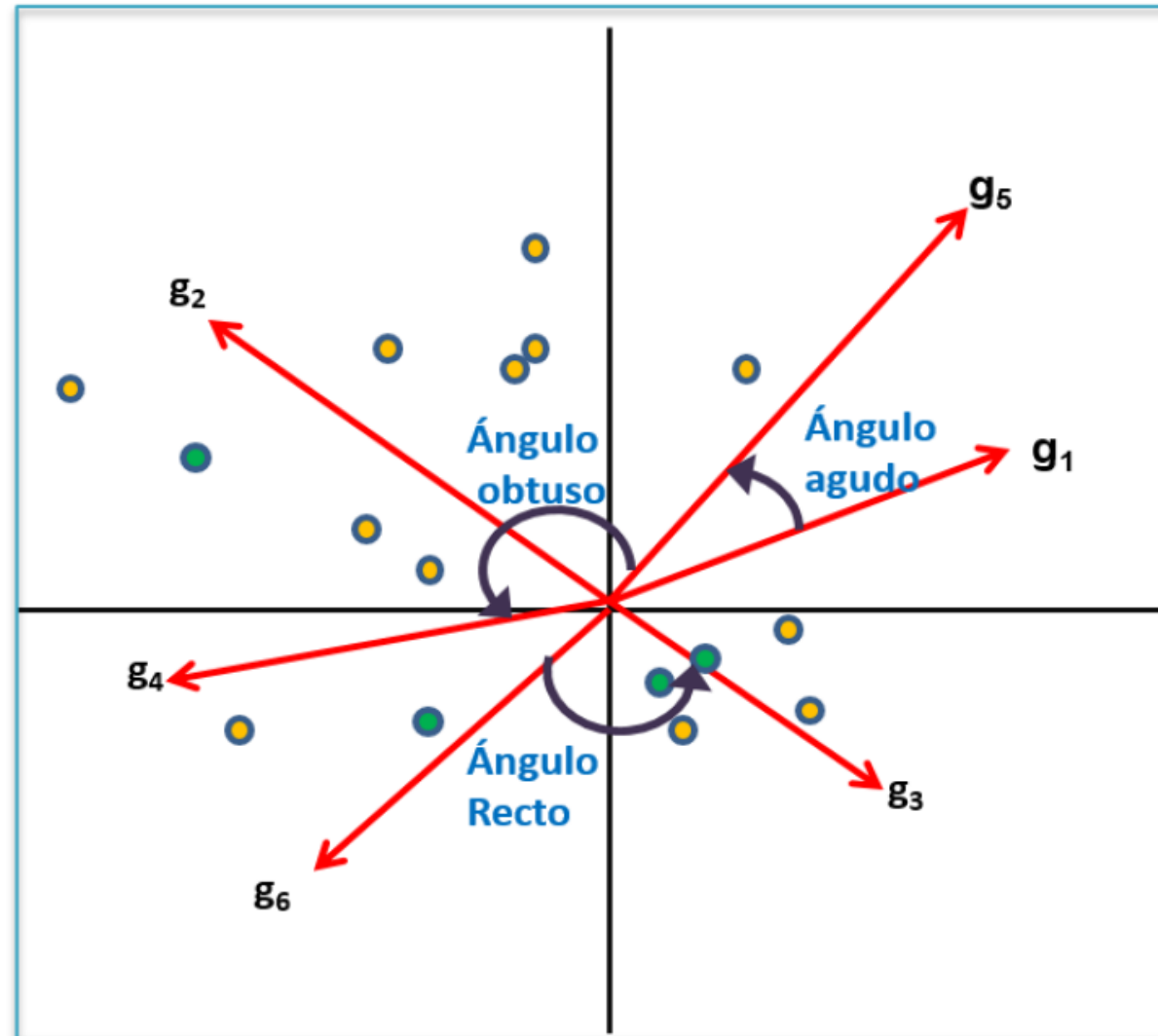
Mientras más distantes
aparecen los puntos del
centro de gravedad, más
representatividad tienen
estas unidades en el
estudio



Variables que aparecen en
las diagonales de los
cuadrantes; éstas no tienen
capacidad discriminante

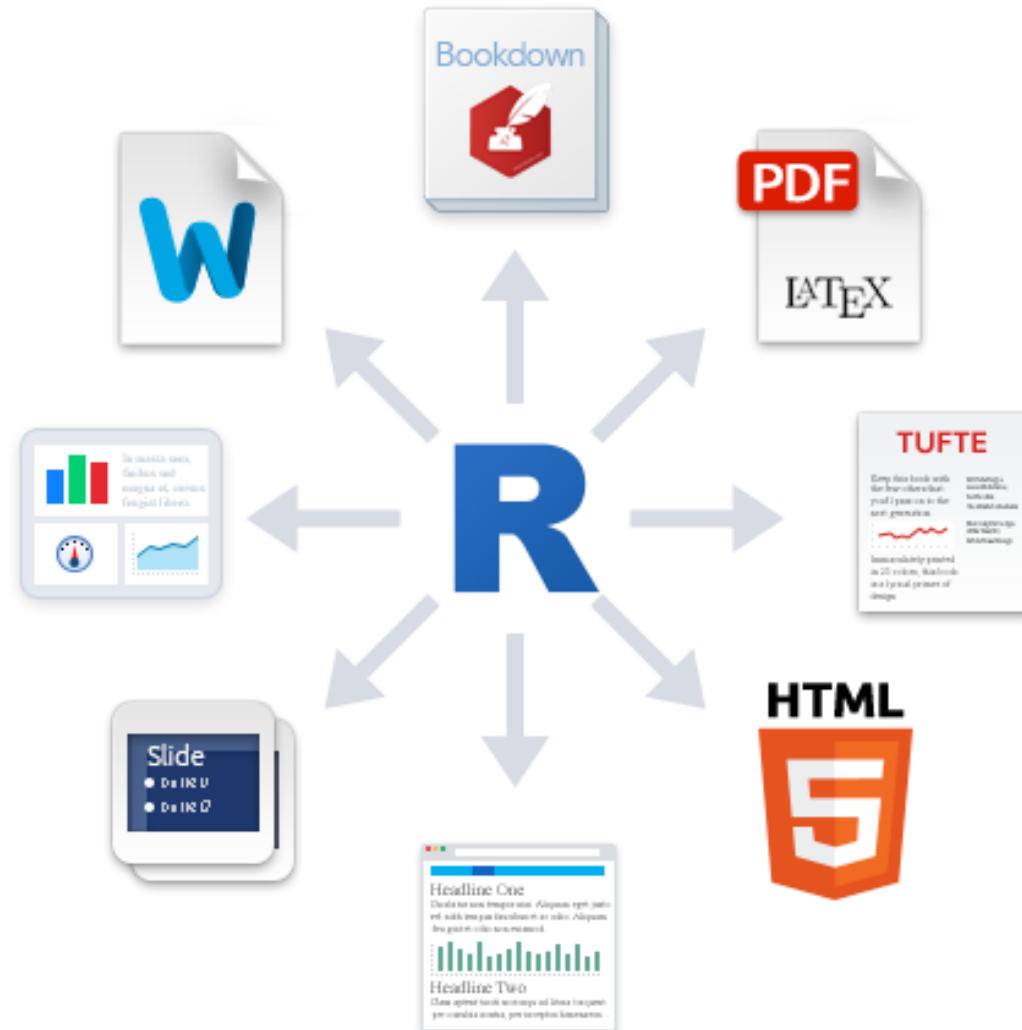


(Cubilla-Montilla, 2021)



(Cubilla-Montilla, 2021)





[pagedown](#)

DIFERENCIA ENTRE R Y RSTUDIO

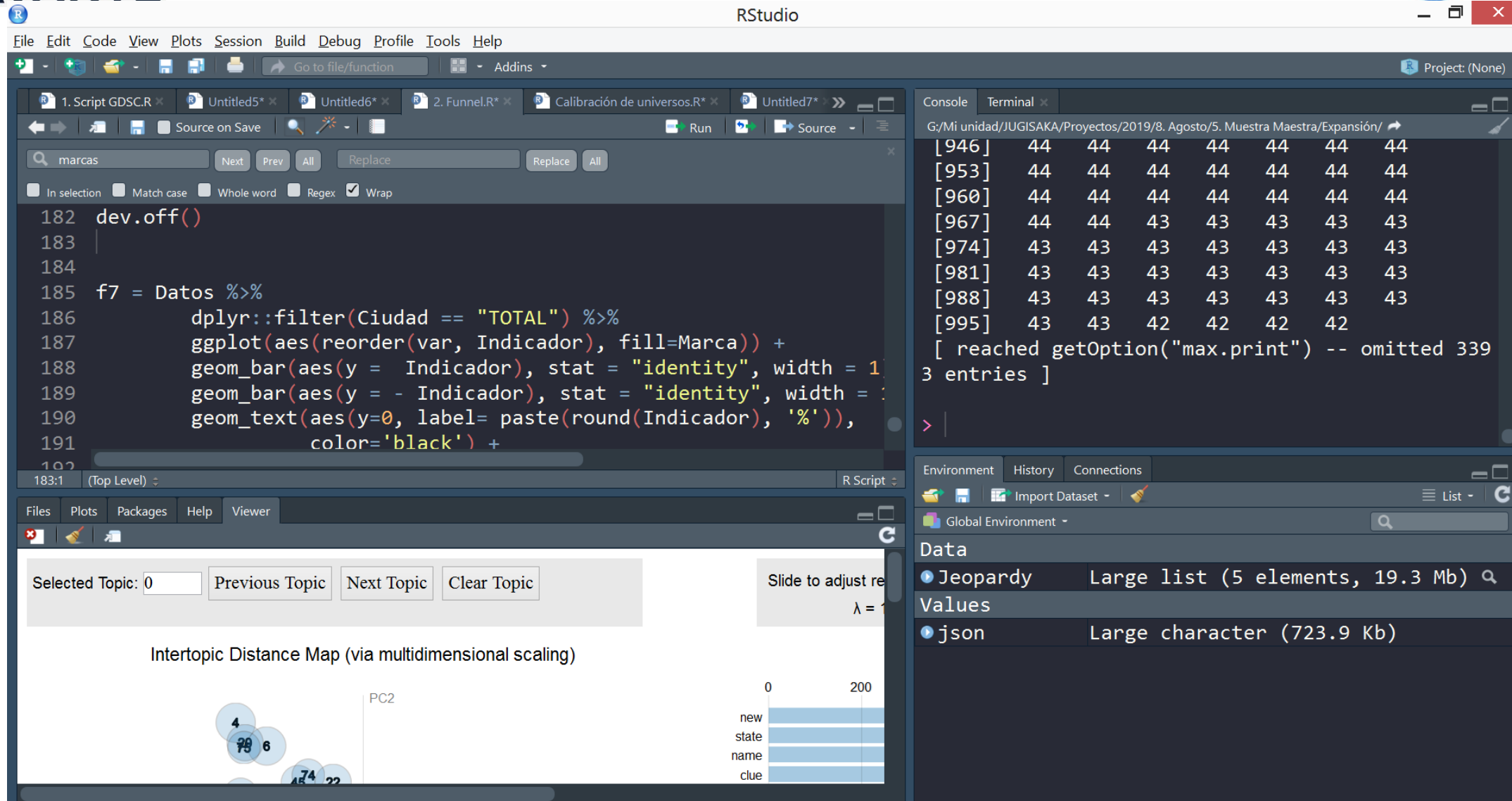




Arte por @allison_horst



Arte por @allison_horst



The screenshot displays the RStudio interface with the following components:

- Script Editor:** Contains R code for data manipulation and plotting.

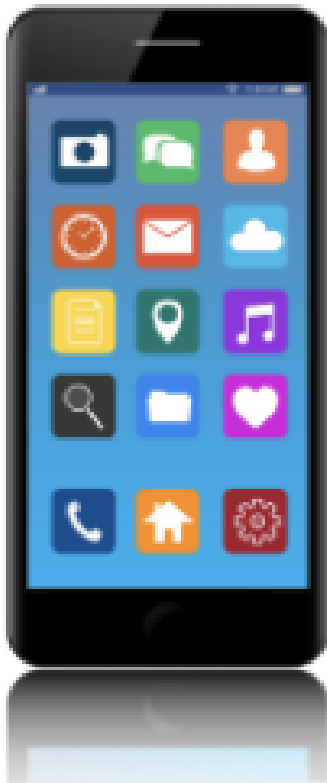

```

182 dev.off()
183
184
185 f7 = Datos %>%
186     dplyr::filter(Ciudad == "TOTAL") %>%
187     ggplot(aes(reorder(var, Indicador), fill=Marca)) +
188     geom_bar(aes(y = Indicador), stat = "identity", width = 1)
189     geom_bar(aes(y = - Indicador), stat = "identity", width = 1)
190     geom_text(aes(y=0, label= paste(round(Indicador), '%')),
191               color='black') +
192
      
```
- Console:** Shows the output of the executed code, displaying a matrix of values and a message about reaching the maximum print limit.


```

[946] 44 44 44 44 44 44 44
[953] 44 44 44 44 44 44 44
[960] 44 44 44 44 44 44 44
[967] 44 44 43 43 43 43 43
[974] 43 43 43 43 43 43 43
[981] 43 43 43 43 43 43 43
[988] 43 43 43 43 43 43 43
[995] 43 43 42 42 42 42
[ reached getOption("max.print") -- omitted 339
3 entries ]
      
```
- Environment:** Shows the Global Environment with two objects:
 - Jeopardy:** Large list (5 elements, 19.3 Mb)
 - json:** Large character (723.9 Kb)
- Viewer:** Displays a web application interface with the following elements:
 - Buttons: Selected Topic: 0, Previous Topic, Next Topic, Clear Topic
 - Text: Intertopic Distance Map (via multidimensional scaling)
 - Figure: A scatter plot showing the relationship between variables, with a legend indicating 'new', 'state', 'name', and 'clue'.

R: Nuevo teléfono



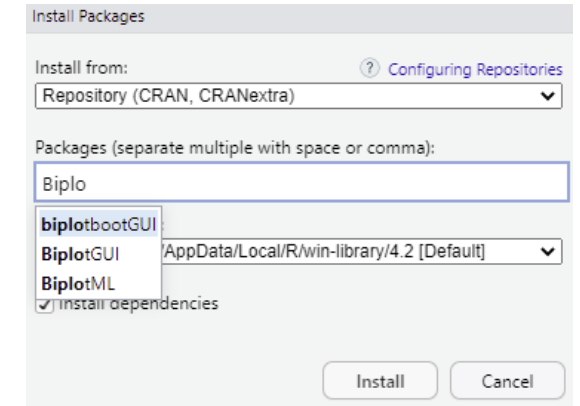
Paquetes: Aplicaciones que se pueden descargar



¿CÓMO SE TRABAJA EN R?

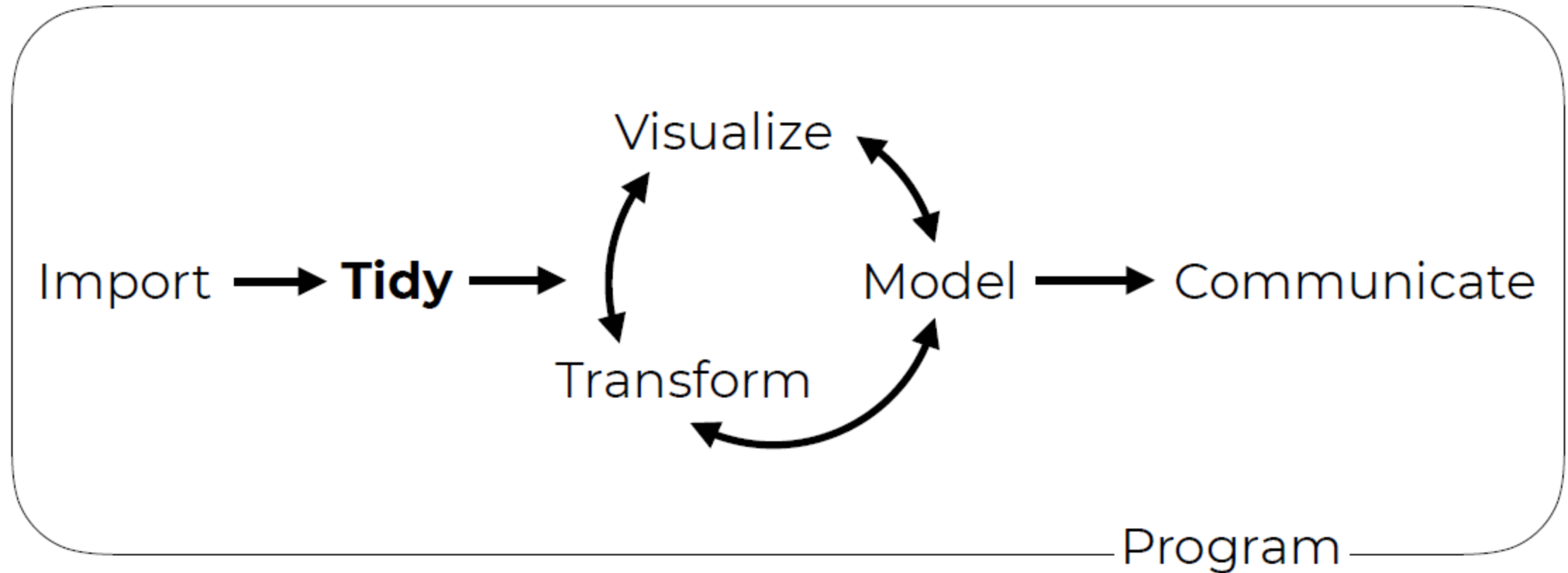
```
install.packages("packagename")
```

```
library(packagename)
```



MultBiplotR
BiplotML
tidyverse
readxl
skimr

citation("package")





```
library(tidyverse)
```



```
library(readr)  
library(dplyr)  
library(tidyr)  
library(ggplot2)  
library(purrr)  
library(tibble)  
library(stringr)  
library(forcats)
```



```
ggplot(data, aes(x = __, y = __)) +  
  geom_point()
```

```
data %>%  
  filter(interesting_variable > z) %>%  
  ggplot(aes(x = __, y = __, colour = condition))  
  geom_point() +  
  facet_wrap(~ group)
```

TU TURNO. MANOS A LA OBRA.

Crea un proyecto para el flujo de trabajo

- 1 Use el conjunto de datos sobre consumo de proteínas en varios países para realizar un análisis HJ-Biplot, Utilice los paquetes
 1. MultBiplotR
 2. BiplotGUI
- 2 Use el conjunto de datos sobre el rendimiento de los jugadores profesionales de fútbol para realizar un análisis multivariante y un análisis clúster. Utilice los paquetes
 1. FactoMineR y factoextra
 2. explor
 3. Factoshiny
- 3 Como ejercicio realizar un análisis multivariante de los indicadores macroeconómicos de los países. Al finalizar la interpretación general, ¿qué puede decir específicamente de los resultados para Ecuador?

- 4 El conjunto de datos de marcas y atributos de los carros contiene la percepción de 1000 personas mayores de 25 años propietarias de vehículos. Realice un análisis multivariante que permita identificar el posicionamiento de las marcas.
1. FactoMineR
 2. explor
 3. Factoshiny
- 5 Use el conjunto de datos sobre hobbies incluido en el paquete FactoMineR y realice un análisis multivariante para los primeros 8 hobbies. Use las variables de estado civil, profesión y cantidad de hobbies como suplementarias.
1. FactoMineR y factoextra
 2. explor
 3. Factoshiny

BIPLOT LOGÍSTICO

El biplot clásico requiere que la matriz X esté conformada por variables de naturaleza cuantitativa y continua, análogo a la regresión lineal. Considere ahora una matriz con datos binarios:

$$X_{n \times p} = \begin{matrix} & \begin{matrix} 1 & \dots & p \end{matrix} \\ \begin{matrix} 1 \\ \vdots \\ i \\ \vdots \\ n \end{matrix} & \begin{bmatrix} 1 & 0 & \dots & 1 \\ 0 & 1 & \dots & 1 \\ \vdots & \vdots & \dots & \vdots \\ 1 & 1 & \dots & 0 \end{bmatrix} \end{matrix}$$



Sea $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T$, con $rank(\mathbf{X}) = r$ y $\mathbf{x}_i \in \{0, 1\}^p$, $i = 1, \dots, n$, $x_{ij} \sim Ber(\pi(\theta_{ij}))$, donde $\pi(\cdot)$ es la inversa de la función de enlace. Usando $\pi(\theta_{ij}) = \{1 + \exp(-\theta_{ij})\}^{-1}$, que representa la probabilidad de que la característica j se encuentre presente en el individuo i .

Teniendo en cuenta que:

$$P(X_{ij} = x_{ij}) = \pi(\theta_{ij})^{x_{ij}} (1 - \pi(\theta_{ij}))^{1-x_{ij}}.$$

La función de verosimilitud es

$$L(\mathbf{X}; \Theta) = \prod_{i=1}^n \prod_{j=1}^p \pi(\theta_{ij})^{x_{ij}} (1 - \pi(\theta_{ij}))^{1-x_{ij}}.$$

Y el negativo del log-verosimilitud se escribe como

$$\mathcal{L}(\Theta) = - \sum_{i=1}^n \sum_{j=1}^p [x_{ij} \log(\pi(\theta_{ij})) + (1 - x_{ij}) \log(1 - \pi(\theta_{ij}))].$$

En este caso **no es apropiado centrar las columnas** porque la matriz centrada ya no está formada por unos y ceros, entonces se extiende la especificación del espacio de parámetros al introducir el vector de desplazamiento μ y obtener un centrado basado en el modelo.

La matriz canónica de parámetros $\Theta = (\theta_1, \dots, \theta_n)^T$ puede ser representada en una estructura de baja dimensión por algún entero $k \leq r$ que satisface

$$\theta_i = \mu + \sum_{s=1}^k a_{is} \mathbf{b}_s, \quad i = 1, \dots, n.$$

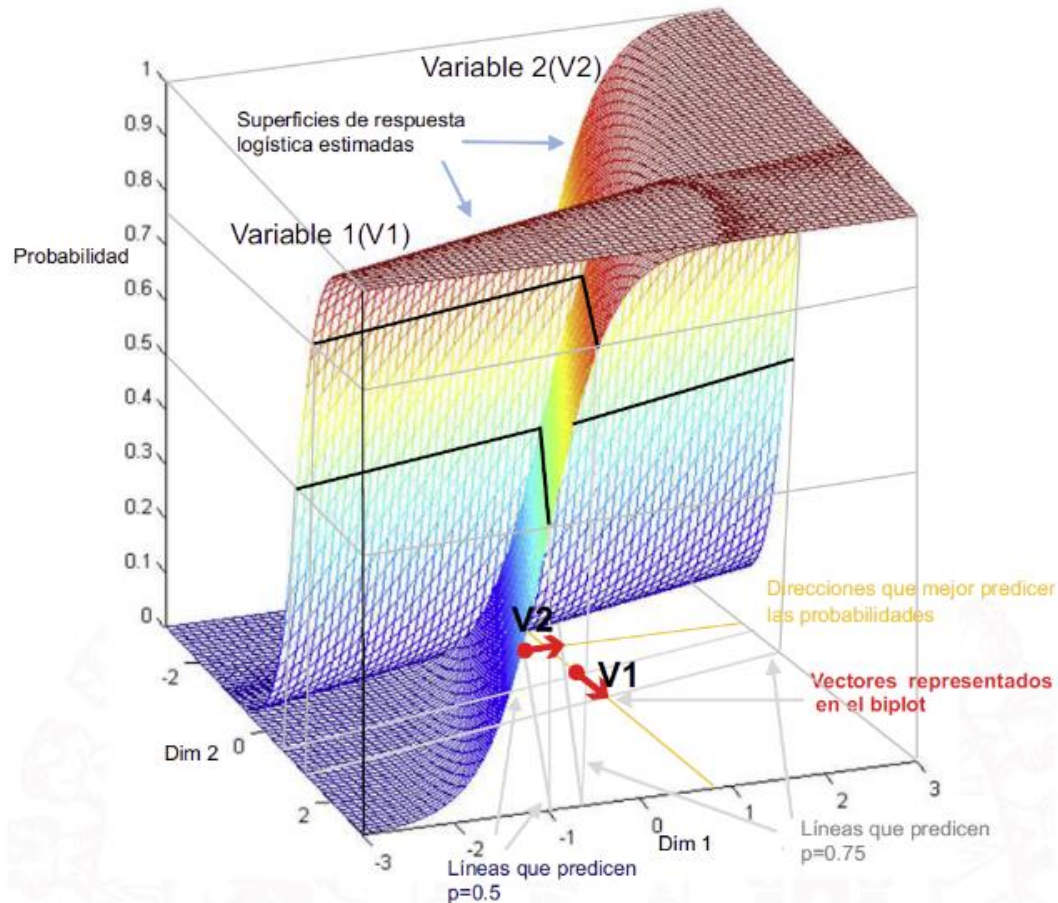
Que expresado en forma matricial se escribe como

$$\Theta = \text{logit}(\Pi) = \mathbf{1}_n \mu^T + \mathbf{A} \mathbf{B}^T,$$

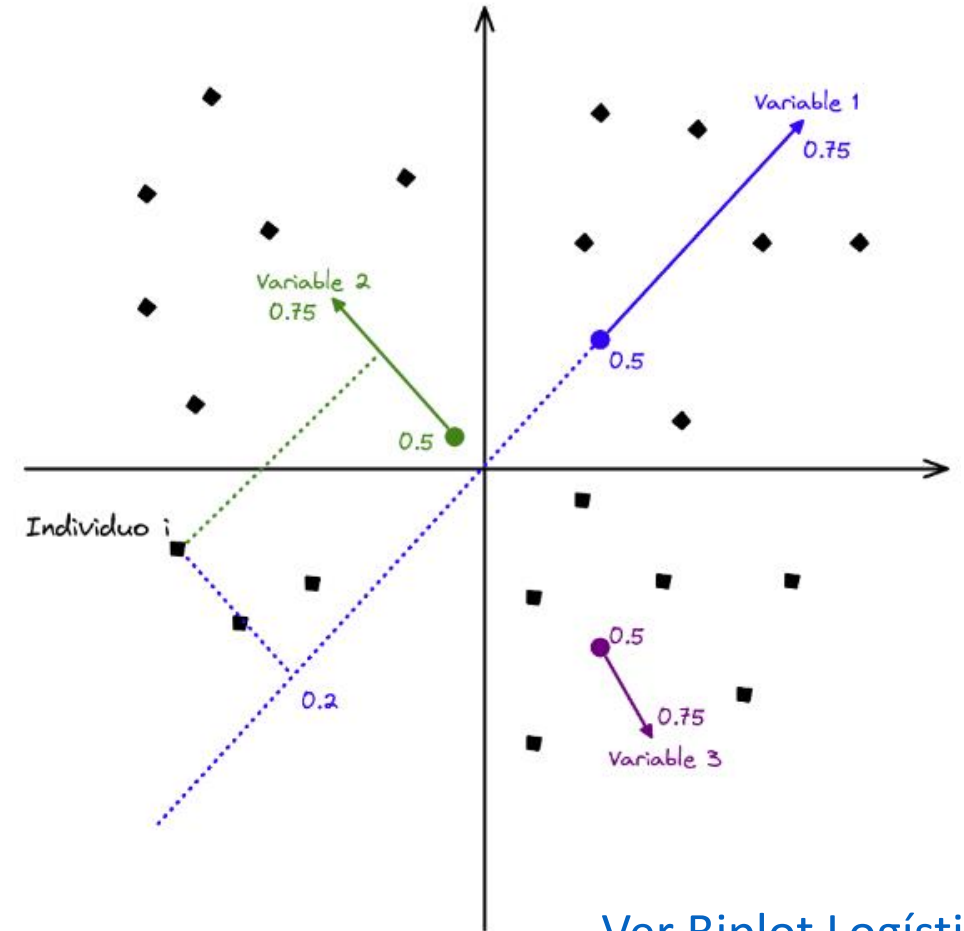
donde $\mathbf{1}_n$ es un vector n -dimensional de unos; $\mu = (\mu_1, \dots, \mu_p)^T$; $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_n)^T$ con $\mathbf{a}_i \in \mathbb{R}^k, i = 1, \dots, n$; $\mathbf{B} = (\mathbf{b}_1, \dots, \mathbf{b}_k)$ con $\mathbf{b}_j \in \mathbb{R}^p, j = 1, \dots, k$; y $\Pi = \pi(\Theta)$ es la matriz de probabilidades esperada cuyo ij -ésimo elemento es igual a $\pi(\theta_{ij})$.

Entonces, $\Theta = \text{logit}(\Pi)$ es un biplot en escala logit y el log-odds es $\theta_{ij} = \mu_j + \mathbf{a}_i^T \mathbf{b}_j$.

Al fijar los marcadores fila **A** y ajustar el modelo logístico para $k = 2$, se obtienen las superficies de respuesta.



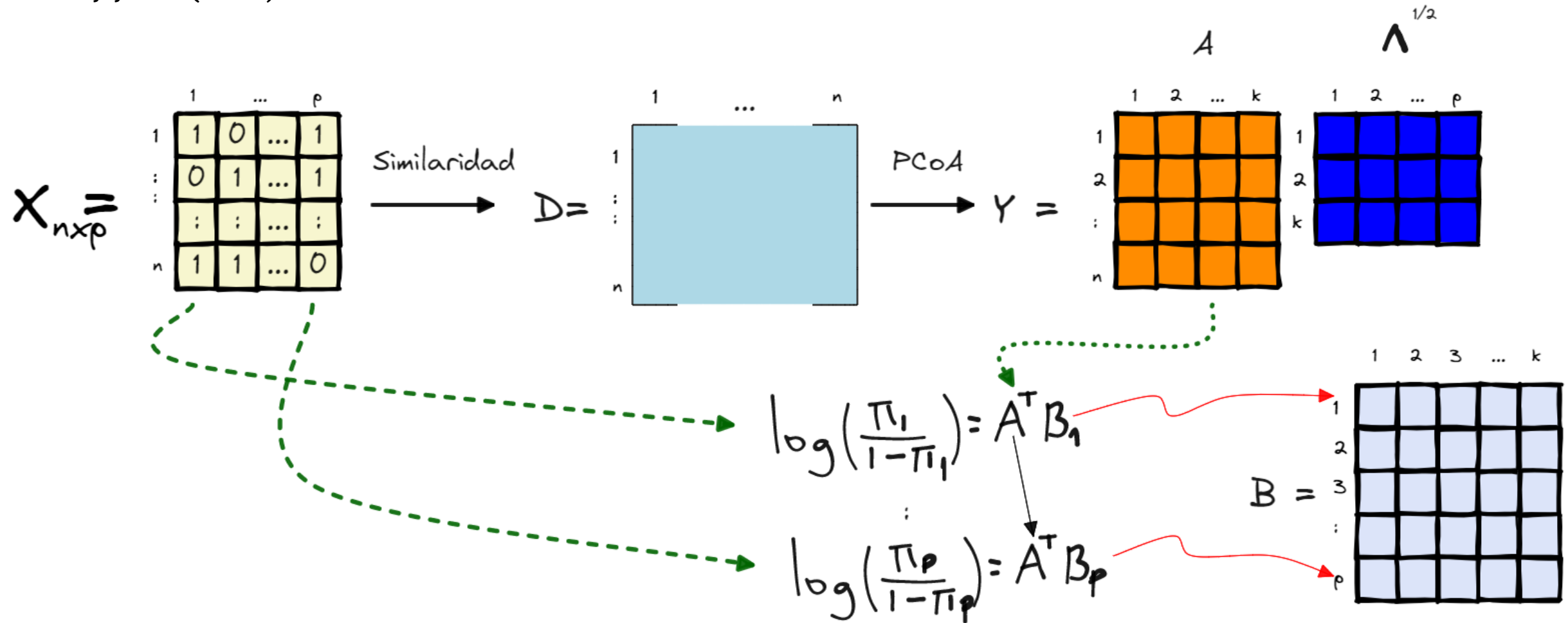
Tomado de Hernández (2016)



[Ver Biplot Logístico 3D](#)

Elaboración propia

Demey y Col. (2008)



Babativa-Márquez, Vicente-Villardón (2021)

La idea es **sustituir el problema de optimización por otro más simple** y que conduzca a la misma solución. El método MM es iterativo y funciona en dos pasos, uno de Mayorización y otro de Minimización.

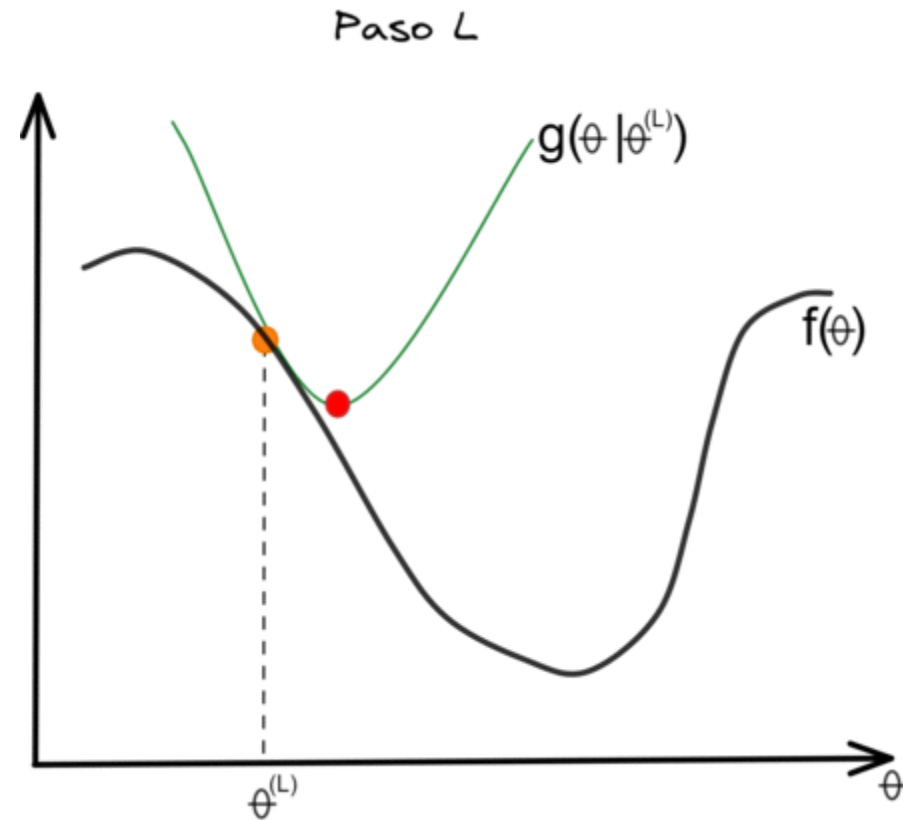
1. La función $g(\theta|\theta^{(l)})$ es una función mayorizada o sustituta de $f(\theta)$ en el punto $\theta^{(l)}$ si

$$\begin{aligned}f(\theta^{(l)}) &= g(\theta^{(l)}|\theta^{(l)}) \\ f(\theta) &\leq g(\theta|\theta^{(l)}) \text{ para todo } \theta\end{aligned}$$

2. El algoritmo de minimización se aplica sobre la función mayorizada sustituta $g(\theta|\theta^{(l)})$, en lugar de la función objetivo inicial. Esto produce el siguiente punto a evaluar $\theta^{(l+1)}$.
3. Si $\theta^{(l+1)}$ representa el mínimo de la función sustituta $g(\theta|\theta^{(l)})$, entonces el método MM lleva a $f(\theta)$ en **dirección descendente** con cada iteración. De esta forma, se cumplen las desigualdades

$$f(\theta^{(l+1)}) \leq g(\theta^{(l+1)}|\theta^{(l)}) \leq g(\theta^{(l)}|\theta^{(l)}) = f(\theta^{(l)}).$$

Babativa-Márquez, Vicente-Villardón (2021)



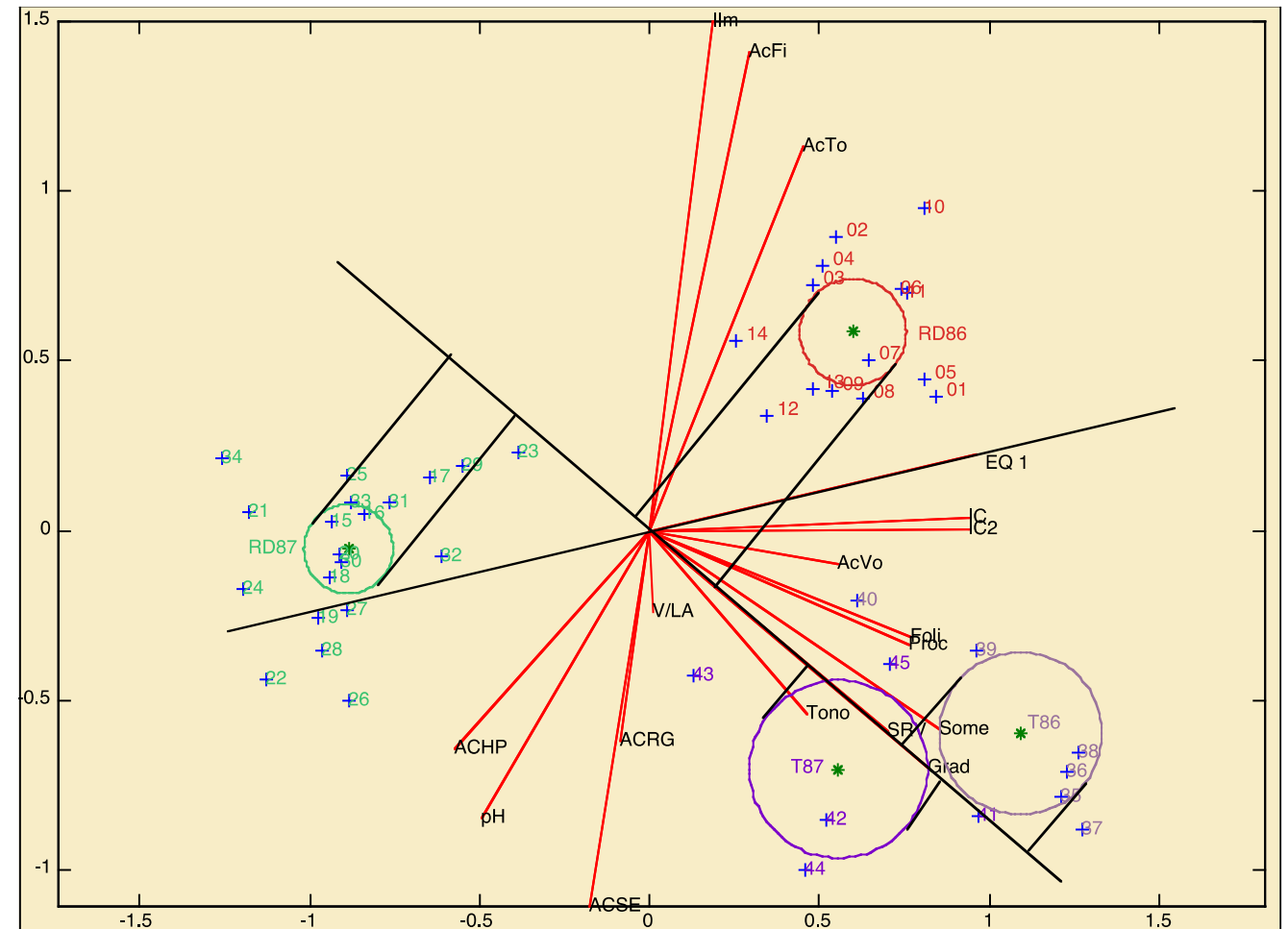
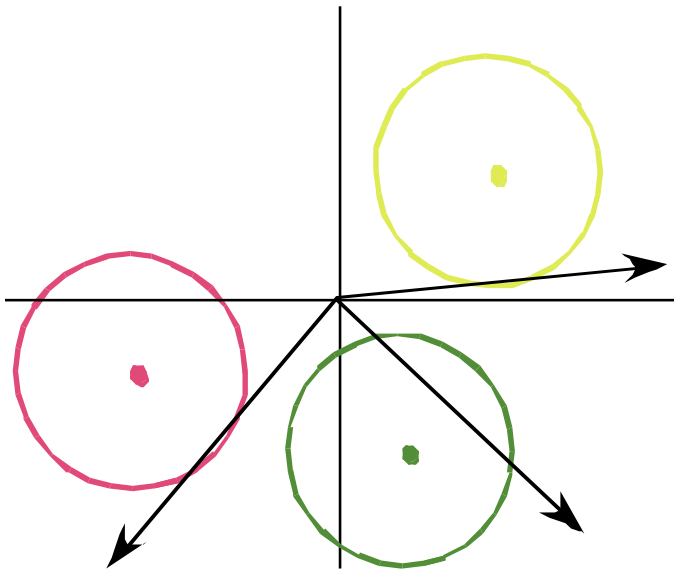
El paquete MultBiplotR contiene el conjunto de datos “spiders” el cual contiene 28 sitios de muestreo donde se identifica la presencia o ausencia de algunas especies de arañas. Realice un biplot logístico para identificar las especies que son más probables en los mismos sitios y concluya sobre las diferencias que se observen.

lorio y col. (2016) realizaron una investigación en el marco del Genomic Determinants of Sensitivity in Cancer 1000 (GDSC1000). De la investigación se pueden extraer diferentes tipos de información sobre líneas celulares de cáncer provenientes de más de 11 mil tumores para 30 tipos de cáncer que integran mutaciones somáticas, copia del número de alteraciones (CNA), metilaciones del ADN y cambios de expresión de genes. Las primeras tres son obtenidas como datos binarios mientras que la expresión genética está medida con variables cuantitativas que son continuas.

El archivo contiene todos los datos unidos en una sola matriz en un formato diferente al requerido y por esta razón fue necesario realizar un preprocesamiento que permitió organizar los datos y adecuarlos para aplicar los métodos.

Para facilitar los análisis obtenidos se incluyeron solo tres tipos de cáncer: carcinoma invasivo de mama (BRCA), adenocarcinoma de pulmón (LUAD) y melanoma cutáneo de piel (SKCM). Realice un análisis a partir de un biplot logístico para los datos de metilación del ADN.

Método que permite visualizar un análisis discriminante o análisis multivariante de la varianza.



Los vinos elaborados en áreas específicas y reconocidos con denominación de origen (DO) son de importancia significativa en las diferentes regiones productoras de vinos. La DO reconoce y garantiza calidad de los vinos fabricados. Consecuentemente, son necesarios una serie de parámetros específicos que permitan a los analistas clasificar distintos vinos en sus correspondientes denominaciones de origen. Entre las características que pueden usarse están la composición en ciertos metales, ácidos orgánicos, ciertos componentes polifenólicos, etc... Los valores de estas características dependen de diversos factores, tales como las variedades de uva empleadas en el proceso de elaboración, o la edad del vino.

Se ha realizado un estudio sobre las dos denominaciones de origen de vinos castellanos (Ribera de Duero y Toro) en dos años diferentes (1986, 1987), con el fin de **distinguir las características diferenciales** entre las dos denominaciones, mediante medidas objetivas obtenidas en laboratorio, de forma que pueda evitarse el fraude en las etiquetas de la denominación sustituyendo ambos vinos debido a su proximidad espacial.

Se han considerado 4 grupos diferentes procedentes de la combinación de denominaciones y años (RD1986, RD1987, T1986, T1987). Se ha considerado el año como posible factor de confusión en la clasificación de los vinos de las dos denominaciones.

Objetivo: Caracterizar las diferencias entre los vinos de dos denominaciones de origen (Ribera del Duero y Toro) según algunas variables objetivo.

- ¿Cuántas variables son de interés en el análisis?
- ¿Cuántas observaciones se tienen en total?
- ¿Cuántos grupos se tienen?
- Indique cuántas observaciones se tienen para cada DO de cada año

Variables medidas:

Grad: Grado alcohólico,

AcVo: Acidez Volatil

AcTo: Acidez Total

AcFi: Acid. Fija

pH

Foli: Fenoles tot (Folin)

Some: Fenoles (Sommers)

SRV: Sust. reactivas a la vanilina

Proc: Procianidoles

ACRG: Antocianos1

ACSE: Antocianos2

ACHP: Antocianos 3

IC : Indice de color 1

IC2 : Indice de color 2

Tono: de color

IIm : Indice de ionización.

EQ1: Edad química

V/LA

PRÁCTICA

18 variables

$$n = 45$$

Ribera de Duero 1986

$$n_1 = 14$$

Ribera de Duero 1987

$$n_2 = 20$$

Toro 1986

$$n_3 = 6$$

Toro 1987

$$n_4 = 5$$

