

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

SOBRE MÍ



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



<https://scholar.google.es/citations?user=2NJRNg8A AAAJ&hl=es>

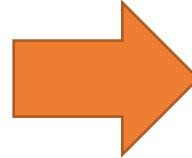
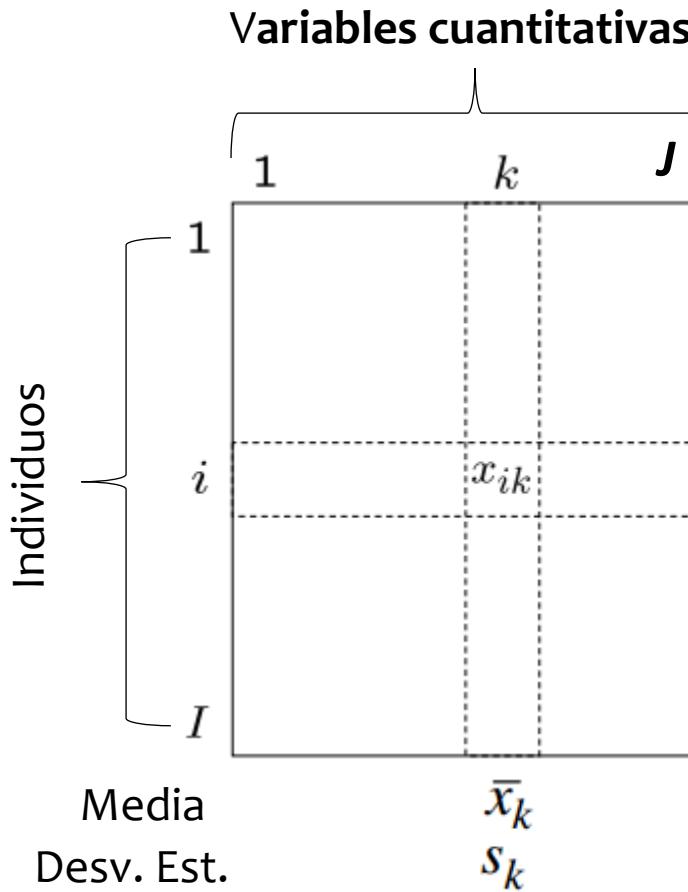
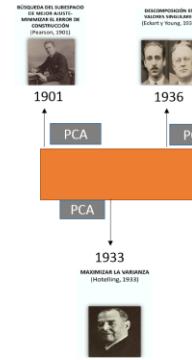


<https://github.com/jgbabativam>

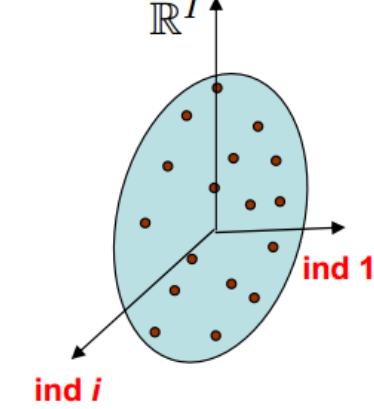
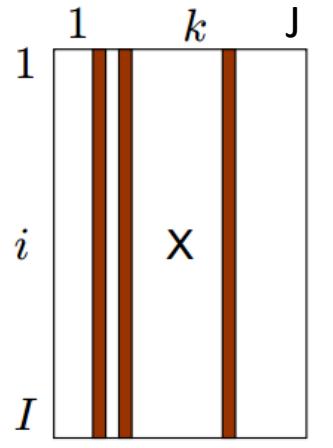
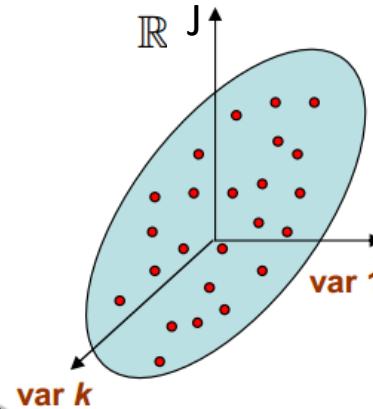
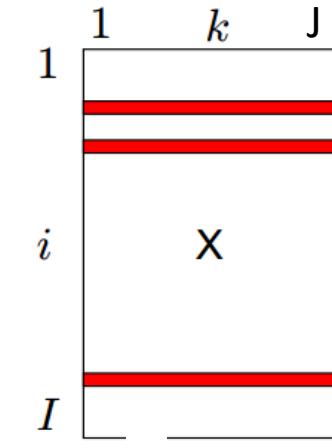


jgbabativam@unal.edu.co

RESUMEN: ACP

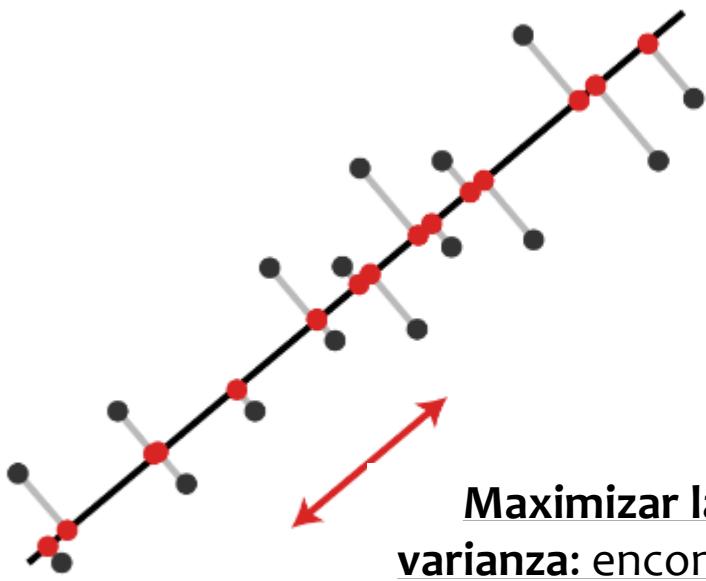


Individuos

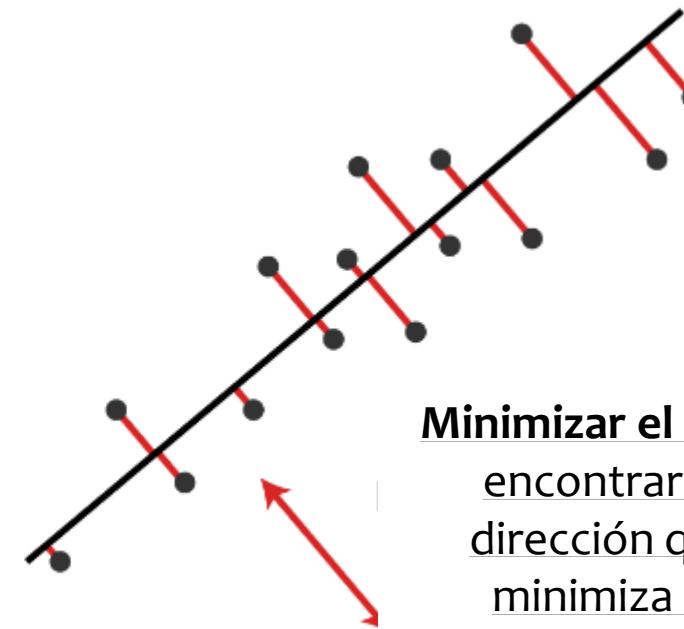


RESUMEN: ACP

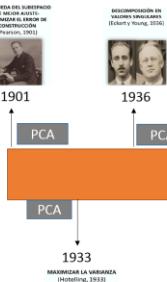
Problema de optimización



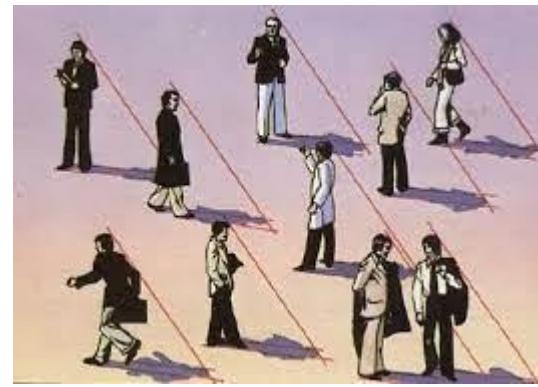
Maximizar la varianza: encontrar la dirección donde los puntos rojos tienen la mayor varianza.



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

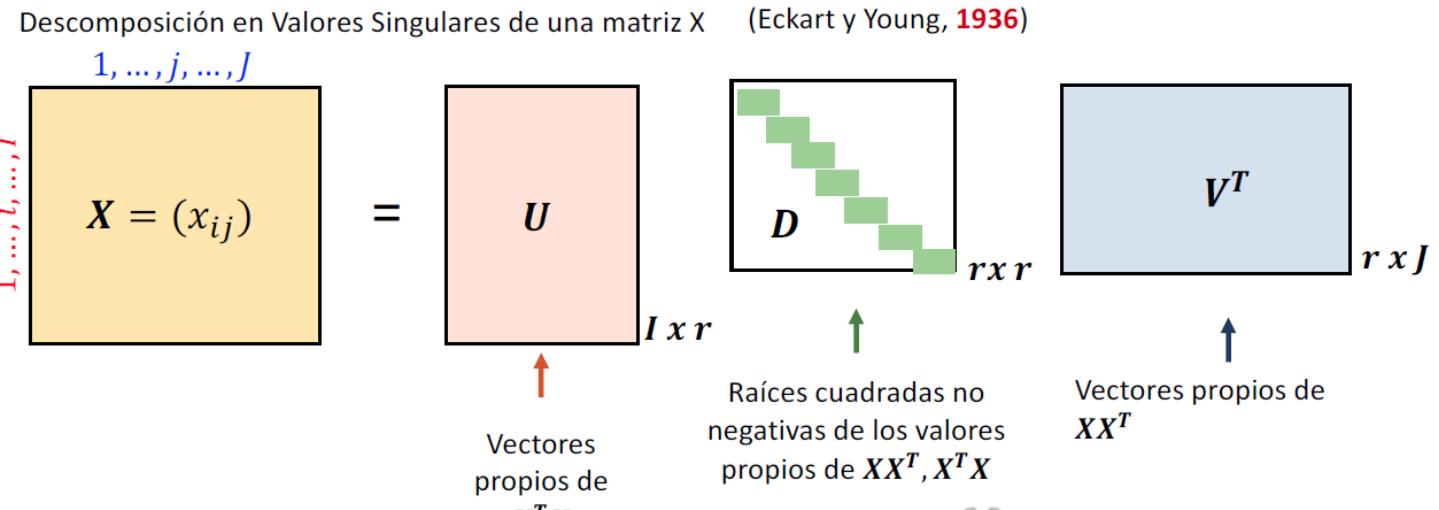
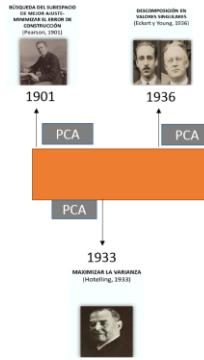
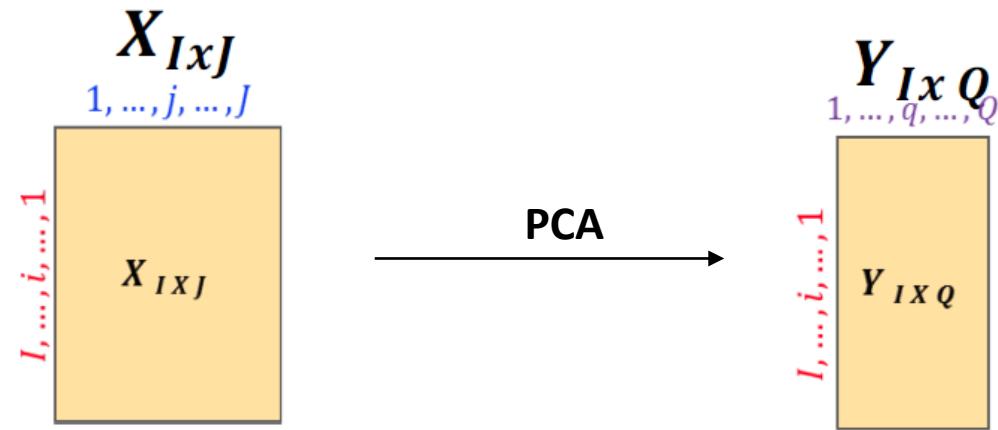


Evolución de las técnicas multivariantes



RESUMEN: ACP

Reproducir la matriz original con menos dimensiones



RESUMEN: ACP

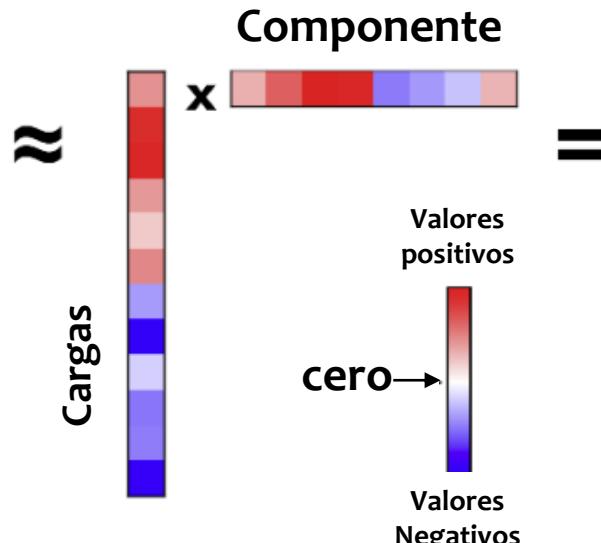
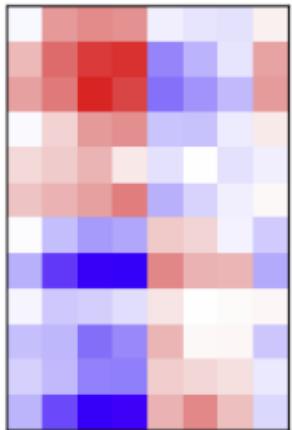
Descomposición en Valores Singulares

(Eckart&Young, 1936)

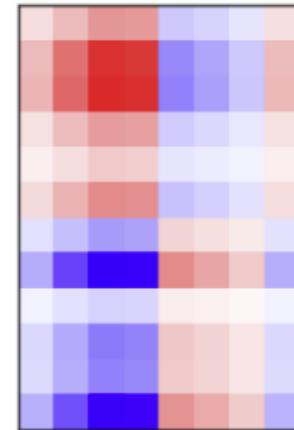
$$X = UDV^T = YV^T$$

Reconstrucción con una CP

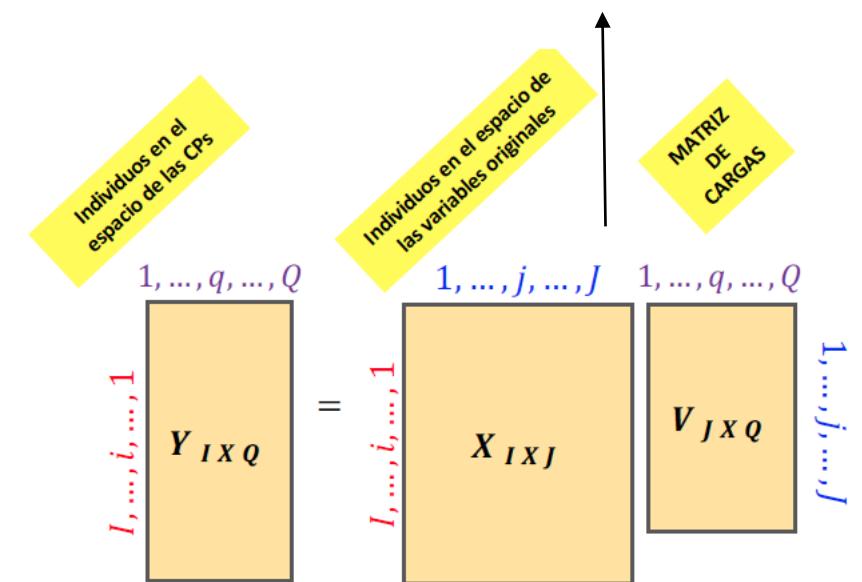
Datos originales



Datos
reconstruidos



Cada componente es una combinación lineal de las variables originales



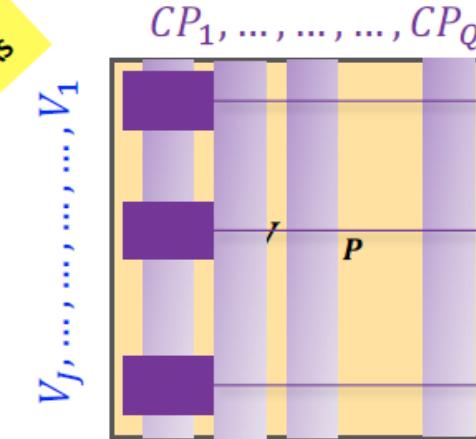
RESUMEN: ACP

Descomposición en Valores Singulares
(Eckart&Young, 1936)

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

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

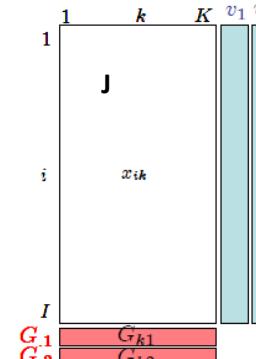
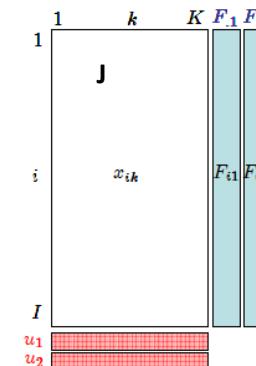
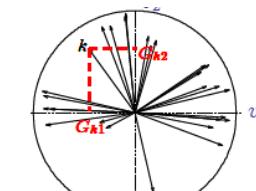
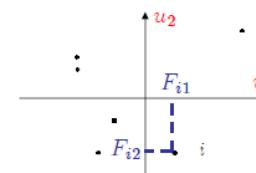
MATRIZ
DE
CARGAS



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



RESUMEN BIPILOT

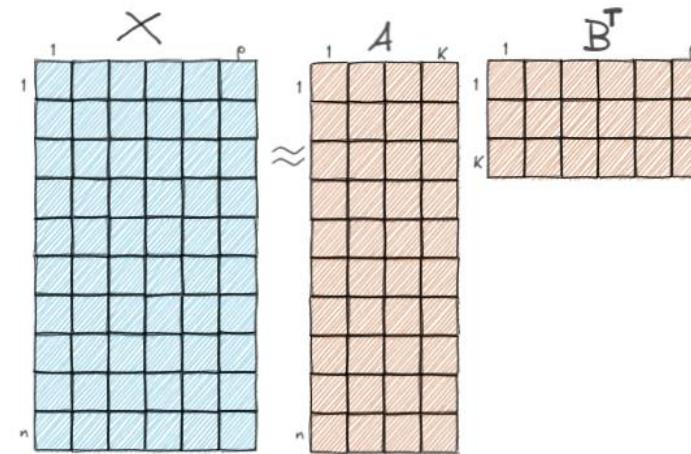
El BIPILOT aproxima la distribución de una muestra multivariante en un espacio de dimensión reducida, normalmente de dimensión dos, y superpone sobre la misma representaciones de las variables sobre las que se mide la muestra (GOWER 1996). Las representaciones de las variables son normalmente vectores, 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](#)

MÉTODOS BIPLOT

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}_{nxk} y \mathbf{B}_{kxq} (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

$$X_{n \times p} = \begin{array}{|c|c|c|c|c|} \hline & 1 & \cdots & p \\ \hline 1 & & & & \\ \vdots & & & & \\ n & & & & \\ \hline \end{array}$$

↓
Eckart y Young (1936)

$$X \approx \begin{array}{|c|c|c|c|c|} \hline & 1 & \cdots & k & \Lambda & 1 & \cdots & p \\ \hline 1 & & & & & & & \\ \vdots & & & & & & & \\ n & & & & & & & \\ \hline \end{array}$$

$$X = AB' + E$$

Calidad de representación

Filas

JK-Biplot
Gabriel (1971)

$$\begin{array}{|c|c|c|c|c|} \hline & 1 & \cdots & k & \Lambda \\ \hline 1 & & & & \\ \vdots & & & & \\ n & & & & \\ \hline \end{array} = \begin{array}{|c|c|c|c|c|} \hline & 1 & \cdots & k & \Lambda \\ \hline 1 & & & & \\ \vdots & & & & \\ n & & & & \\ \hline \end{array}$$

$$\begin{array}{|c|c|c|c|c|} \hline & 1 & \cdots & p & V^T \\ \hline 1 & & & & \\ \vdots & & & & \\ k & & & & \\ \hline \end{array} = \begin{array}{|c|c|c|c|c|} \hline & 1 & \cdots & p & V^T \\ \hline 1 & & & & \\ \vdots & & & & \\ k & & & & \\ \hline \end{array}$$

Columnas

GH-Biplot
Gabriel (1971)

$$\begin{array}{|c|c|c|c|c|} \hline & 1 & \cdots & k & \Lambda \\ \hline 1 & & & & \\ \vdots & & & & \\ n & & & & \\ \hline \end{array} = \begin{array}{|c|c|c|c|c|} \hline & 1 & \cdots & k & \Lambda \\ \hline 1 & & & & \\ \vdots & & & & \\ n & & & & \\ \hline \end{array}$$

$$\begin{array}{|c|c|c|c|c|} \hline & 1 & \cdots & p & V^T \\ \hline 1 & & & & \\ \vdots & & & & \\ k & & & & \\ \hline \end{array} = \begin{array}{|c|c|c|c|c|} \hline & 1 & \cdots & k & V^T \\ \hline 1 & & & & \\ \vdots & & & & \\ k & & & & \\ \hline \end{array}$$

Filas y Columnas

HJ-Biplot
Galindo (1986)

$$\begin{array}{|c|c|c|c|c|} \hline & 1 & \cdots & k & \Lambda \\ \hline 1 & & & & \\ \vdots & & & & \\ n & & & & \\ \hline \end{array} = \begin{array}{|c|c|c|c|c|} \hline & 1 & \cdots & k & \Lambda \\ \hline 1 & & & & \\ \vdots & & & & \\ n & & & & \\ \hline \end{array}$$

$$\begin{array}{|c|c|c|c|c|} \hline & 1 & \cdots & p & V^T \\ \hline 1 & & & & \\ \vdots & & & & \\ k & & & & \\ \hline \end{array} = \begin{array}{|c|c|c|c|c|} \hline & 1 & \cdots & k & V^T \\ \hline 1 & & & & \\ \vdots & & & & \\ k & & & & \\ \hline \end{array}$$

Mismo sistema de
referencia

En el **HJ-BIPLOT** 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

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



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

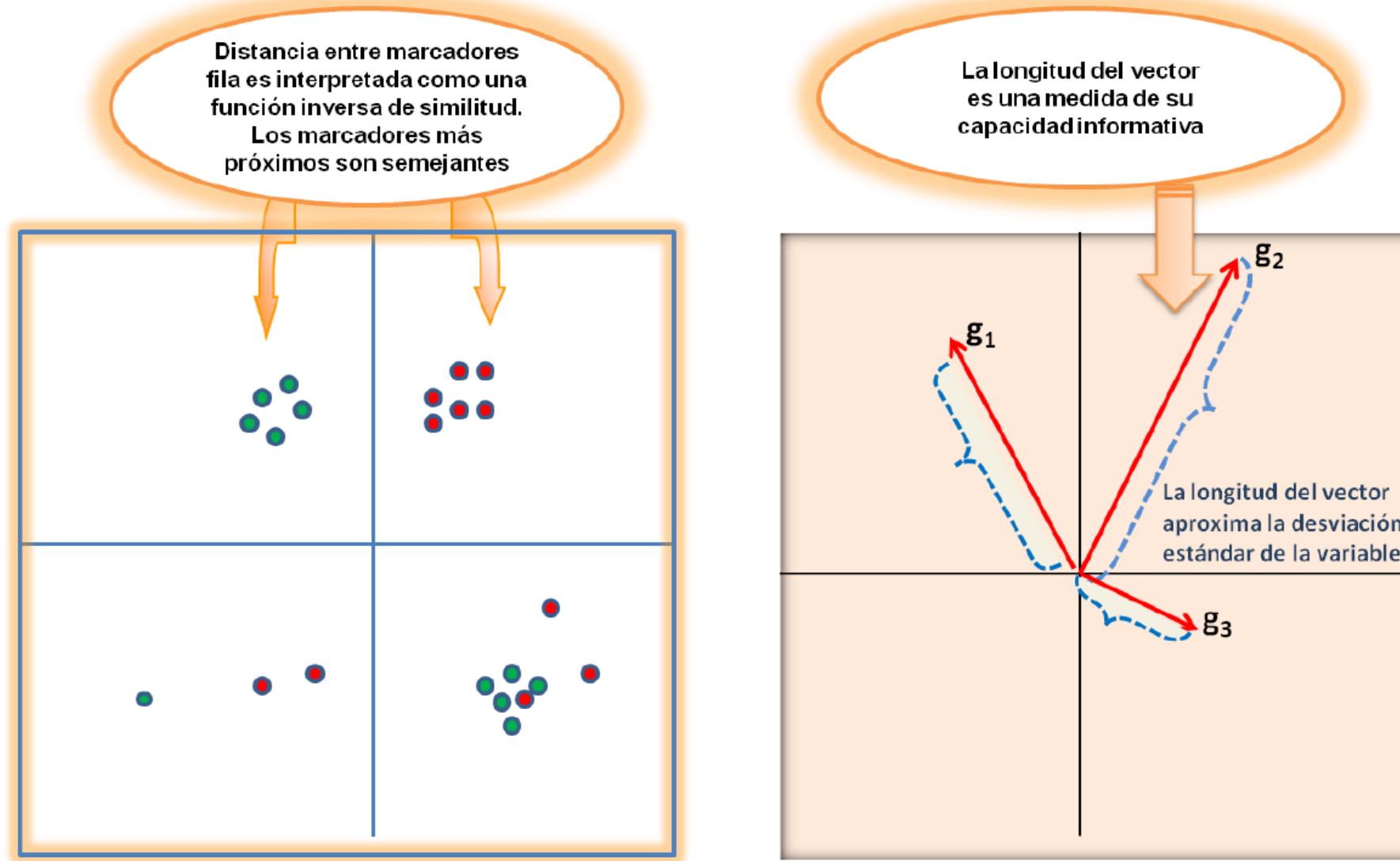
$$B = V_{(k)} \Lambda_{(k)}$$

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.

CALIDAD DE LA REPRESENTACIÓN

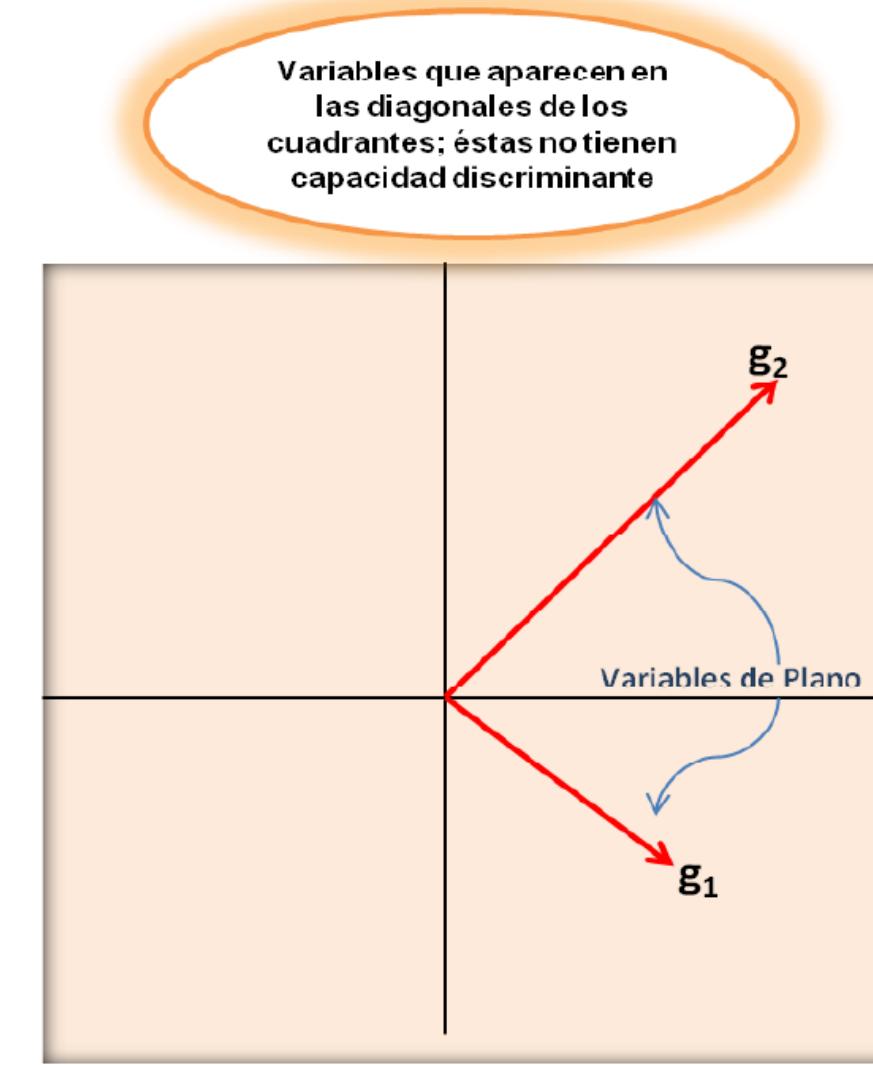
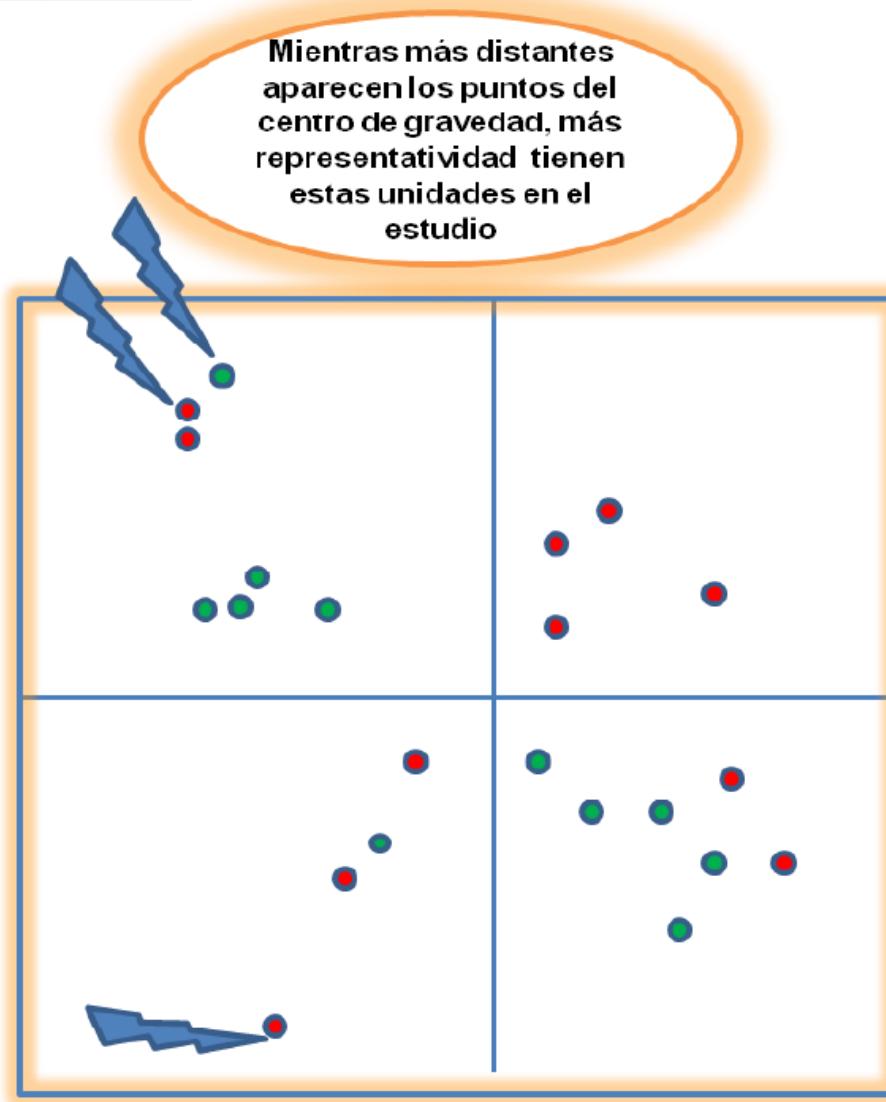
REPRESENTACIÓN SIMULTÁNEA	COORDENADAS FILAS	COORDENADAS COLUMNAS	BONDAD AJUSTE PARA FILAS	BONDAD AJUSTE PARA COLUMNAS
GH-BIPILOT	U	VΛ	$\frac{2}{r}$	$\frac{l_1^2 + l_2^2}{\sum_{a=1}^r l_a^2}$
JK-BIPILOT	UΛ	V	$\frac{l_1^2 + l_2^2}{\sum_{a=1}^r l_a^2}$	$\frac{2}{r}$
HJ-BIPILOT	U Λ	VΛ	$\frac{l_1^2 + l_2^2}{\sum_{a=1}^r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

INTERPRETACIÓN



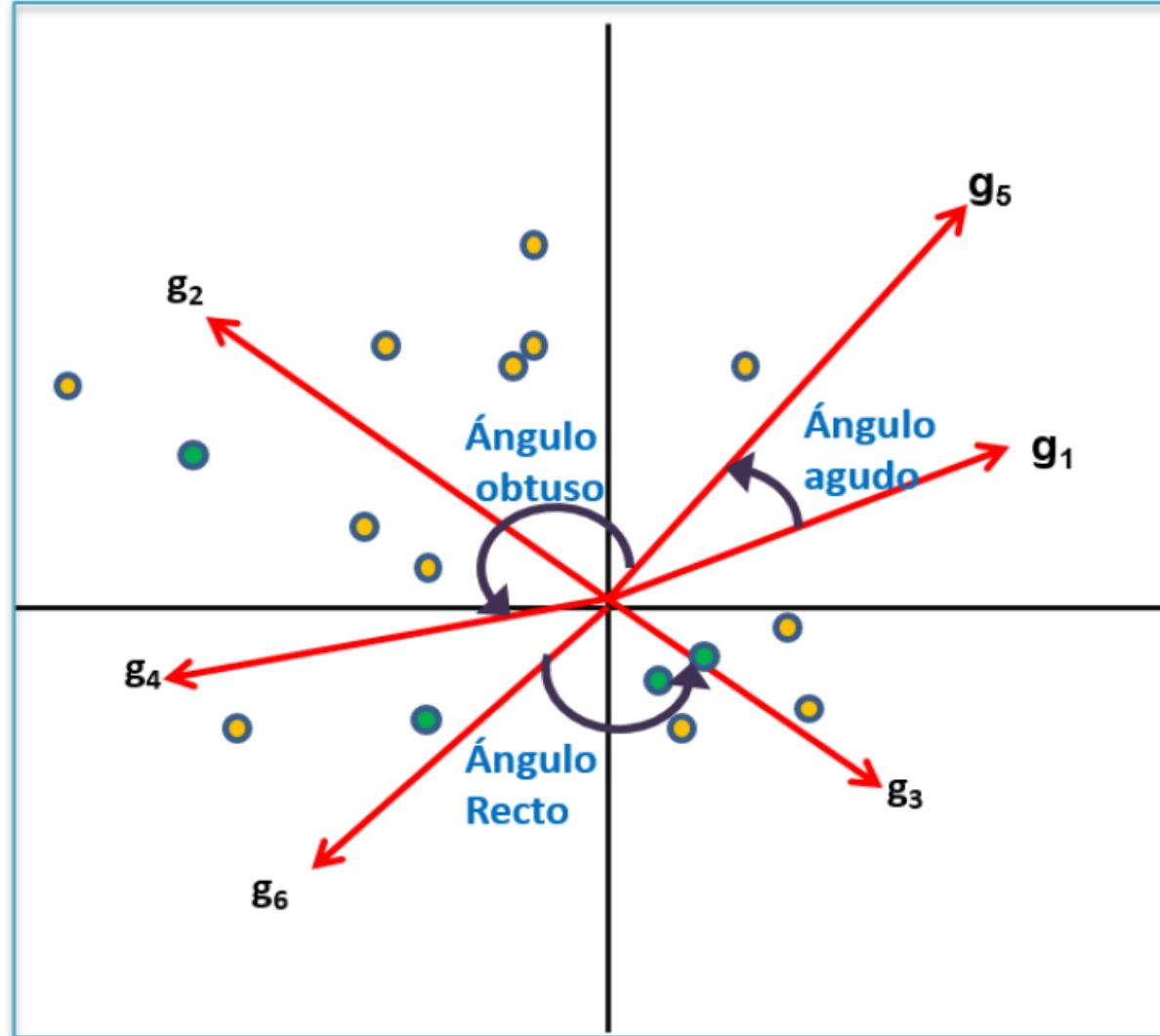
(Cubilla-Montilla, 2021)

INTERPRETACIÓN



(Cubilla-Montilla, 2021)

INTERPRETACIÓN

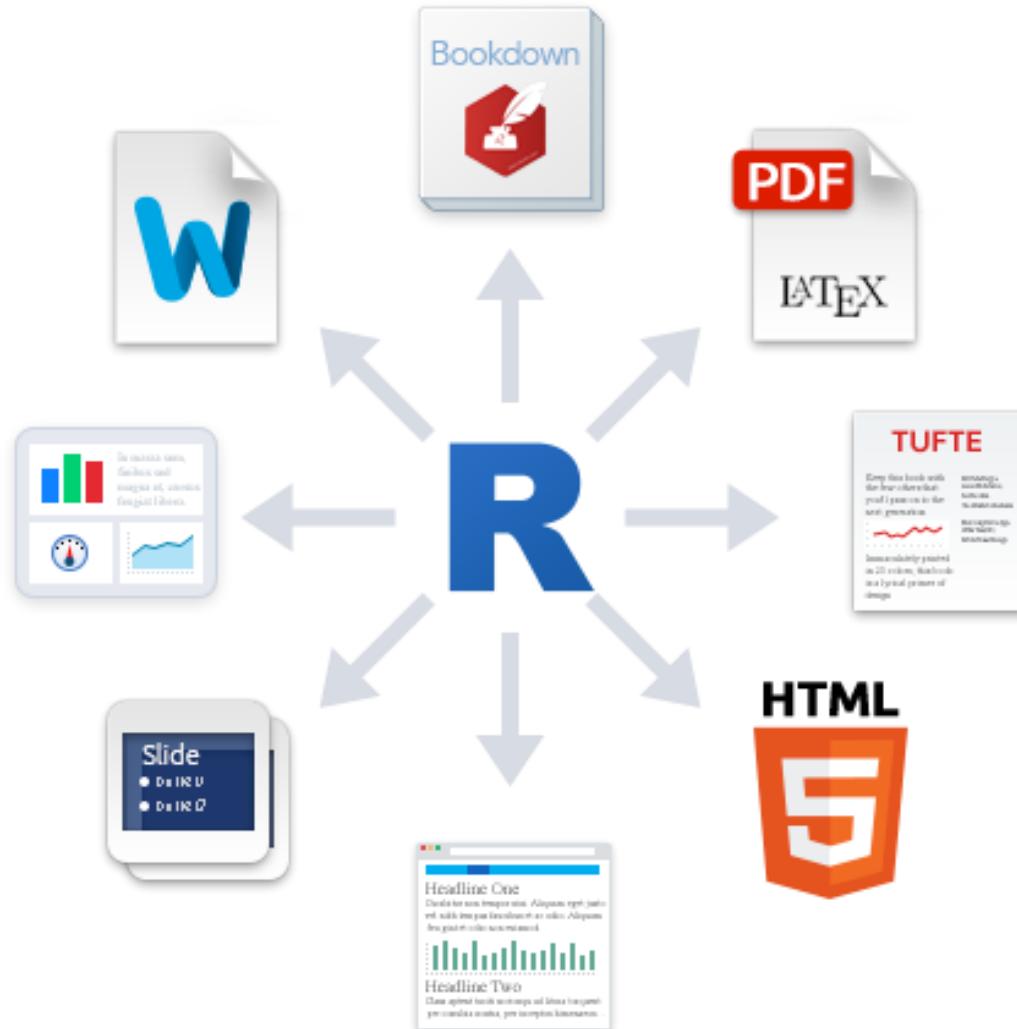


(Cubilla-Montilla, 2021)

PRÁCTICA



PRÁCTICA



[pagedown](#)

DIFERENCIA ENTRE R Y RSTUDIO



DIFERENCIA ENTRE R Y RSTUDIO



Arte por @allison_horst



Arte por @allison_horst

ENTORNO DE RSTUDIO

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Addins

Project: (None)

Console Terminal

```
G:/Mi unidad/JUGISAKA/Proyectos/2019/8. Agosto/5. Muestra Maestra/Expansión/
[946] 44 44 44 44 44 44 44 44
[953] 44 44 44 44 44 44 44 44
[960] 44 44 44 44 44 44 44 44
[967] 44 44 43 43 43 43 43 43
[974] 43 43 43 43 43 43 43 43
[981] 43 43 43 43 43 43 43 43
[988] 43 43 43 43 43 43 43 43
[995] 43 43 42 42 42 42 42
[ reached getOption("max.print") -- omitted 339
3 entries ]
```

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), '%')), color='black') +
191

183:1 (Top Level) R Script

Files Plots Packages Help Viewer

Selected Topic: 0 Previous Topic Next Topic Clear Topic

Slide to adjust re

Intertopic Distance Map (via multidimensional scaling)

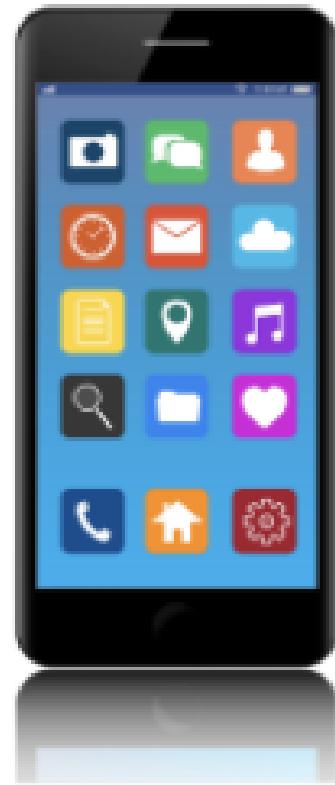
PC2

0 200

new
state
name
clue

¿CÓMO SE TRABAJA EN R?

R: Nuevo teléfono móvil

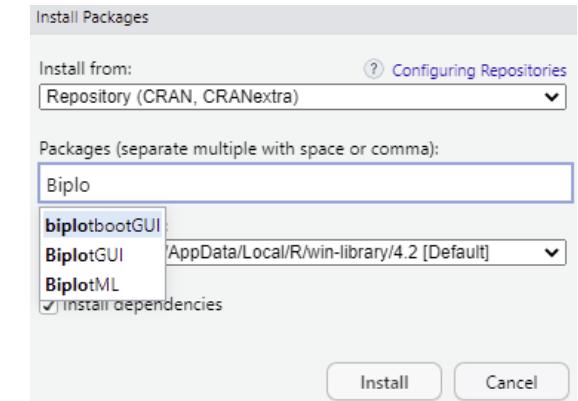


Paquetes: Aplicaciones que se pueden descargar



¿CÓMO SE TRABAJA EN R?

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

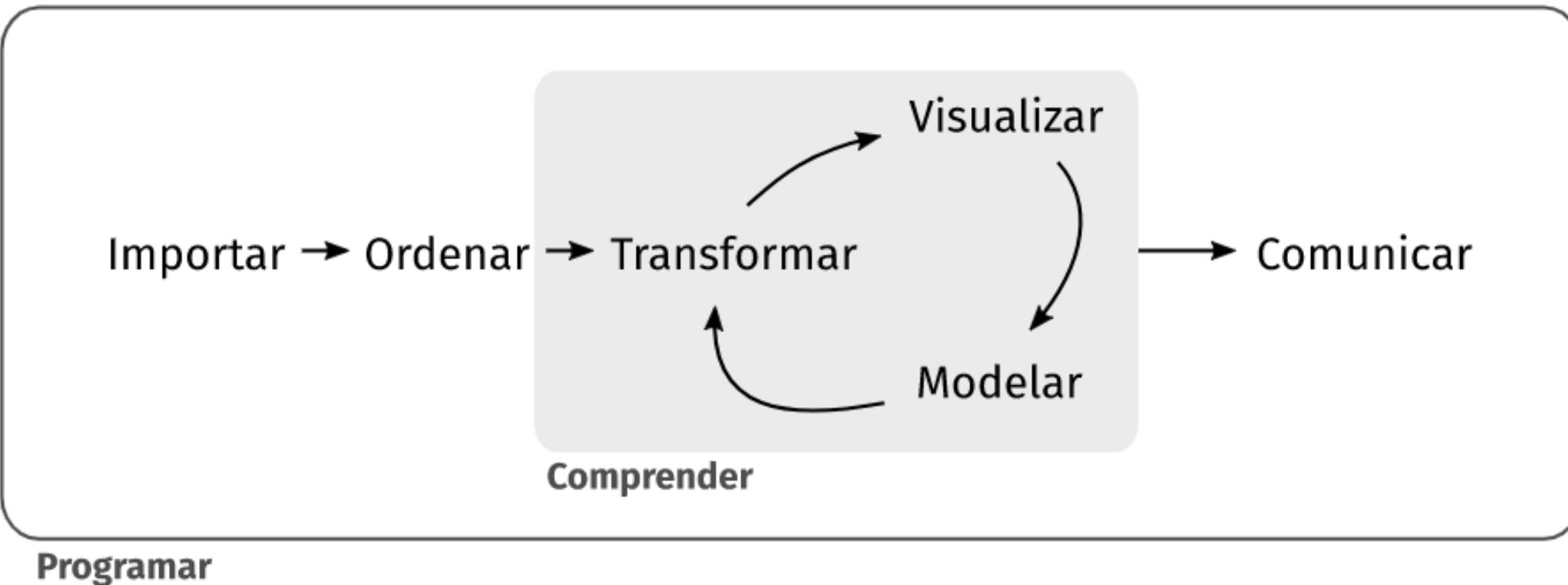


```
library(packagename)
```

MultBiplotR
BiplotML
tidyverse
readxl
skimr

citation("package")

¿CÓMO SE TRABAJA EN R?



[Wickham & Golemund \(2003\).](#)

ENTORNO tidyverse



```
library(tidyverse)
```



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

**TU TURNO.
MANOS A LA OBRA.**

Crea un proyecto para el flujo de trabajo

Crea una carpeta para el proyecto



Curso Multivariante



data



images



R



slides

Descarga los datos de Moodle o de <https://github.com/jgbabativam/Multivariante3Way>

main ▾ 1 branch 0 tags

jgbabativam update datasets

📁 R	update datasets
📁 data	update datasets
📁 images	update datasets
📁 slides	update datasets
📄 .gitignore	update datasets
📄 LICENSE	Initial commit
📄 Multivariate3Way.Rproj	update datasets
📄 README.md	Initial commit

Local Codespaces New

Clone

HTTPS SSH GitHub CLI

<https://github.com/jgbabativam/Multivariate3Way>

Use Git or checkout with SVN using the web URL.

Open with GitHub Desktop

Download ZIP

3 weeks ago

Descarga los datos de Moodle o de <https://github.com/jgbabativam/Multivariate3Way>

- 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**

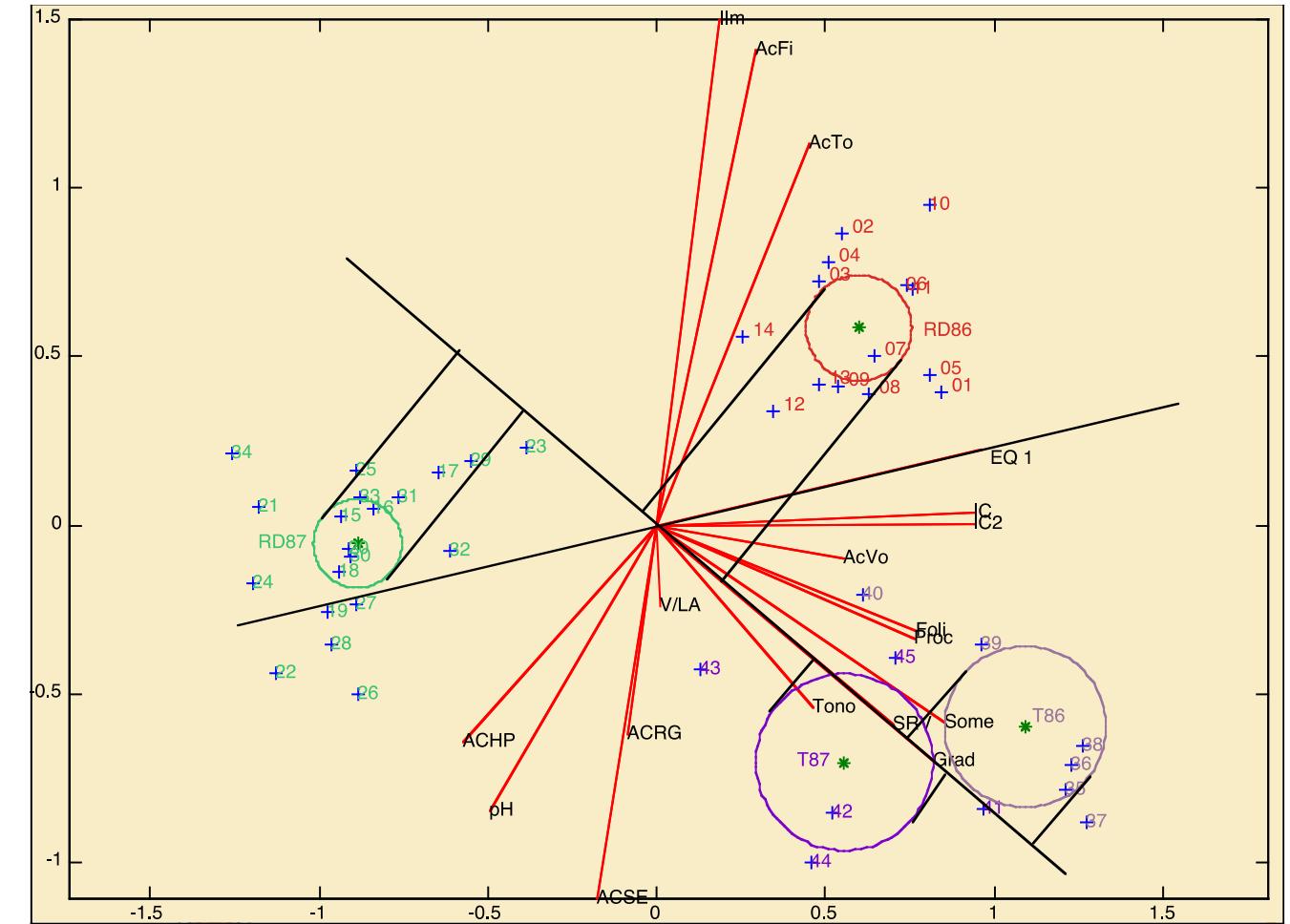
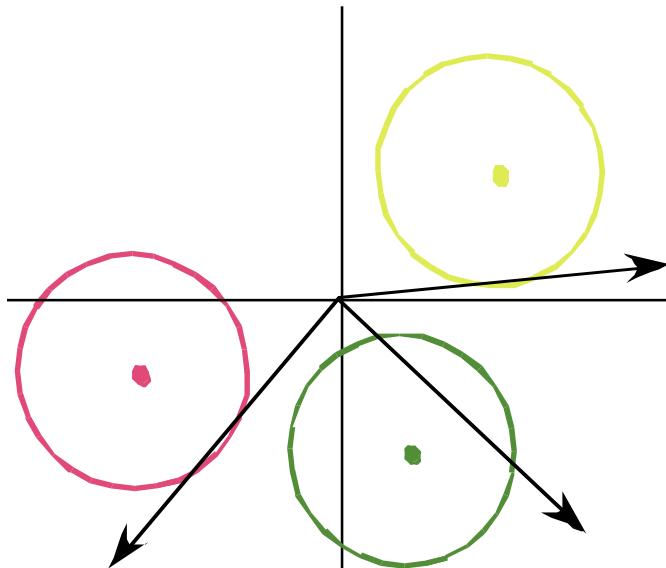
[Use el código cargado en moodle](#)
- 3 Como ejercicio realizar un análisis multivariante de los indicadores macroeconómicos de los países. Al finalizar la interpretación general, ¿qué puede decir específicamente de los resultados para Ecuador?

PRÁCTICA: DATOS CATEGÓRICOS

- 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

MANOVA BI PLOT – BI PLOT CANÓNICO

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

$n_2 = 20$

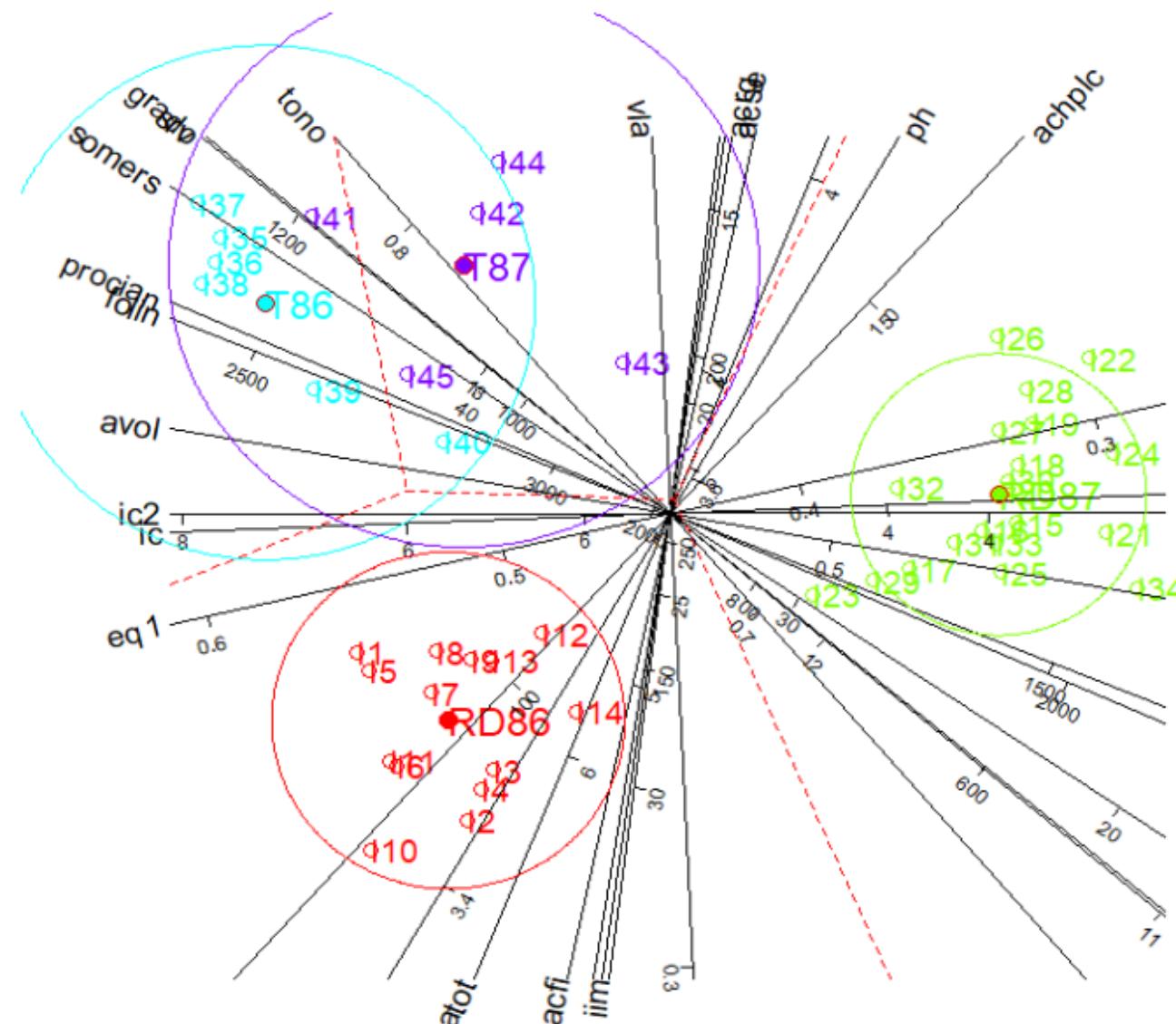
Toro 1986

$n_3 = 6$

Toro 1987

$n_4 = 5$

Canonical/MANOVA Biplot / 1 - 2 (94.37 %)



BI PLOT 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:

$$X_{n \times p} = \begin{array}{c|ccc} & 1 & \dots & p \\ \hline 1 & 1 & 0 & \dots & 1 \\ 2 & 0 & 1 & \dots & 1 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ n & 1 & 1 & \dots & 0 \end{array}$$



BI PLOT LOGÍSTICO

Sea $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T$, con $\text{rank}(\mathbf{X}) = r$ y $\mathbf{x}_i \in \{0, 1\}^p$, $i = 1, \dots, n$, $x_{ij} \sim \text{Ber}(\pi(\theta_{ij}))$, 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

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

Y el negativo del log-verosimilitud se escribe como

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

BI PLOT LOGÍSTICO

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

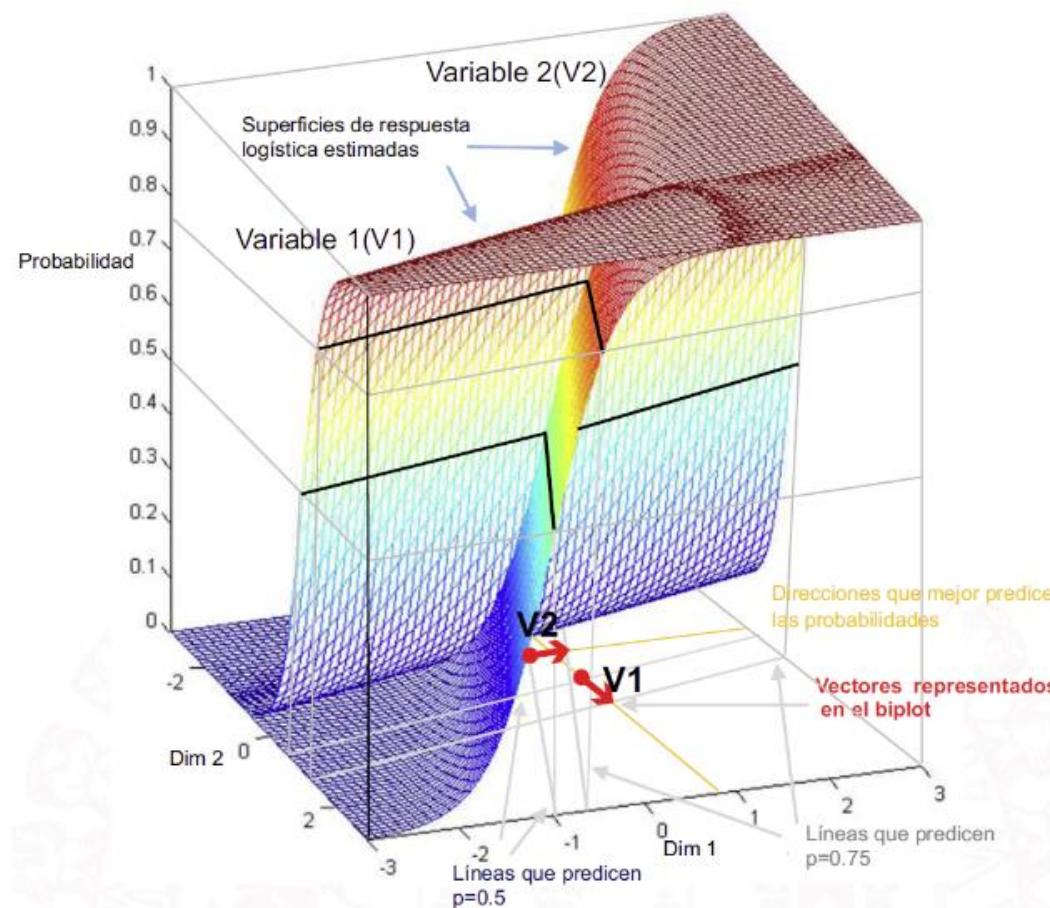
$$\Theta = \text{logit}(\Pi) = \mathbf{1}_n \mu^T + \mathbf{A} \mathbf{B}^T,$$

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 $\pi(\theta_{ij})$.

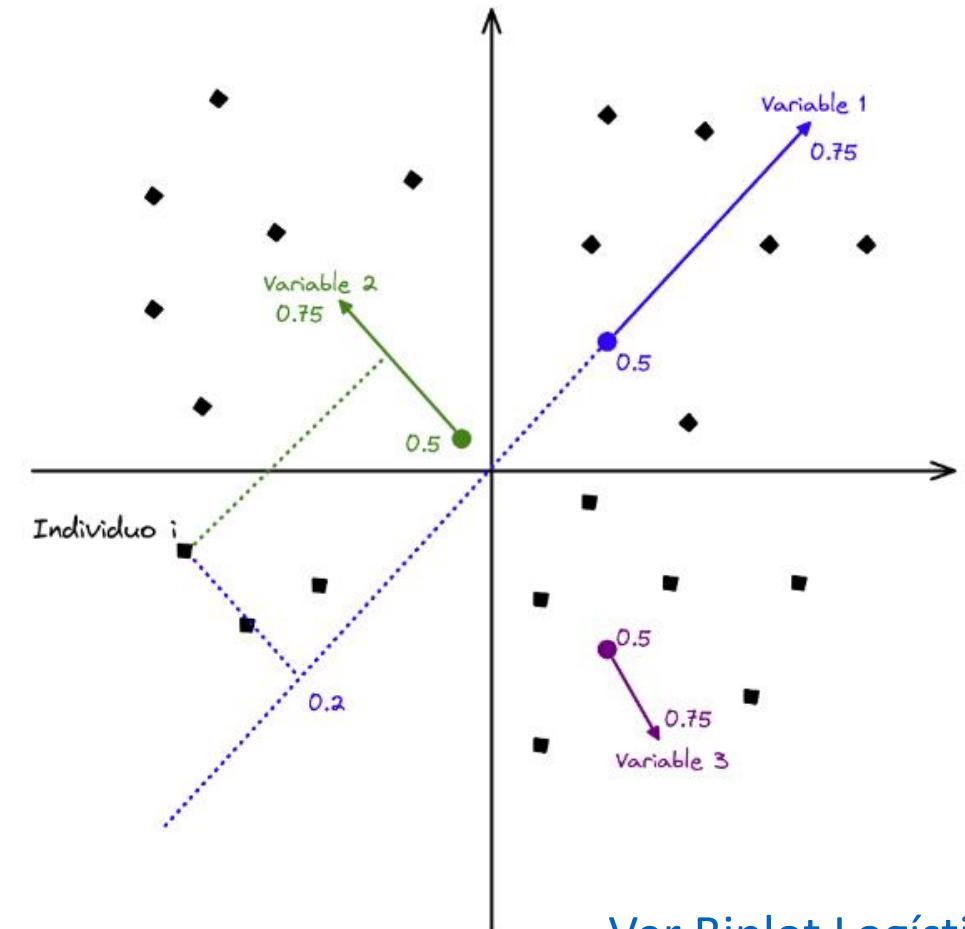
Entonces, $\Theta = \text{logit}(\Pi)$ es un biplot en escala logit y el log-odds es $\theta_{ij} = \mu_j + \mathbf{a}_i^T \mathbf{b}_j$.

BI PLOT LOGÍSTICO

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

BI PLOT LOGÍSTICO EXTERNO

Demey y Col. (2008)

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

Similaridad

$$D = \begin{matrix} 1 & \dots & n \\ \begin{matrix} 1 & & & \\ & \ddots & & \\ n & & & \end{matrix} \end{matrix}$$

PCoA

$$Y = \begin{matrix} A & \Lambda^{1/2} \\ \begin{matrix} 1 & 2 & \dots & k \\ 1 & 2 & \dots & p \\ \vdots & \vdots & \dots & \vdots \\ n & k & \dots & k \end{matrix} \end{matrix}$$

$$\log\left(\frac{\pi_1}{1-\pi_1}\right) = A^T \beta_1$$

$$\log\left(\frac{\pi_p}{1-\pi_p}\right) = A^T \beta_p$$

$$B = \begin{matrix} 1 & 2 & 3 & \dots & k \\ \begin{matrix} 1 & & & & \\ 2 & & & & \\ 3 & & & & \\ \vdots & & & & \\ p & & & & \end{matrix} \end{matrix}$$

BI PLOT LOGÍSTICO ML

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

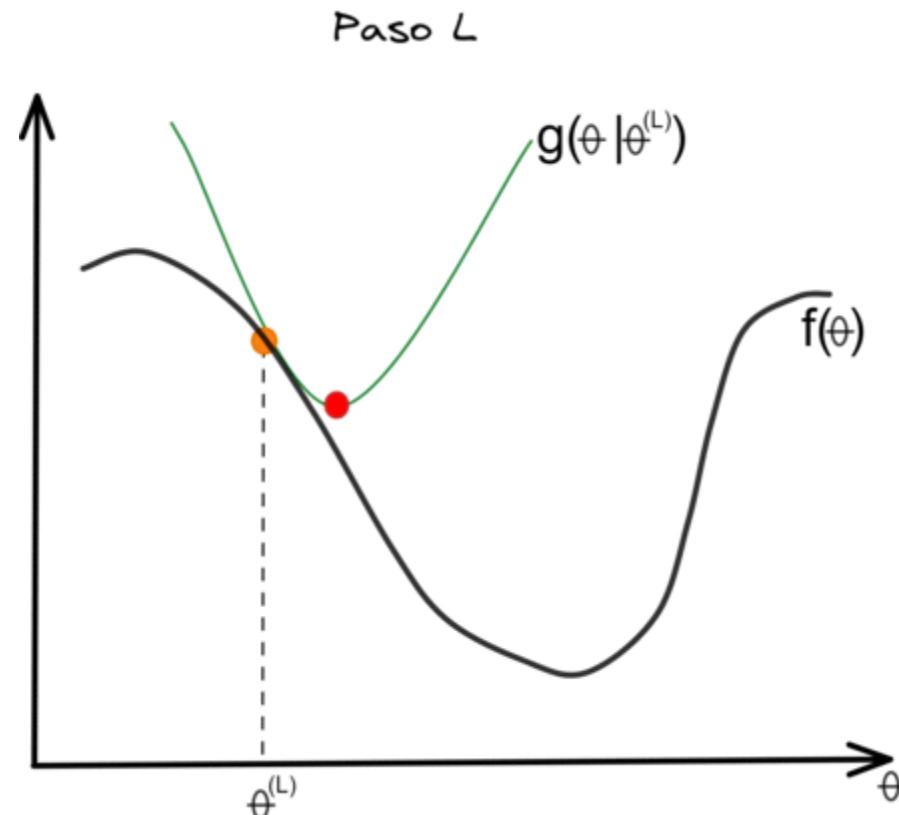
$$f(\theta^{(l)}) = g(\theta^{(l)}|\theta^{(l)})$$
$$f(\theta) \leq g(\theta|\theta^{(l)}) \text{ para todo } \theta$$

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

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

BI PLOT LOGÍSTICO ML

Babativa-Márquez, Vicente-Villardón (2021)



PRÁCTICA 1

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.

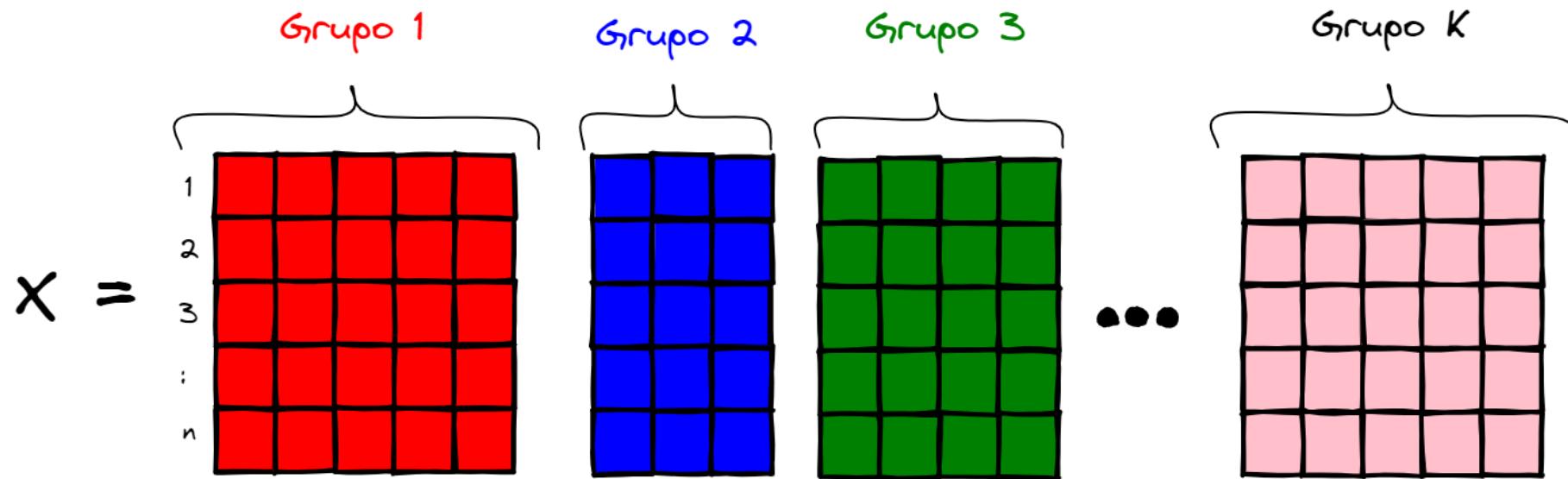
PRÁCTICA 2

Iorio 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.

CASO 1. Se mide a los mismos individuos y se tienen bloques de variables.

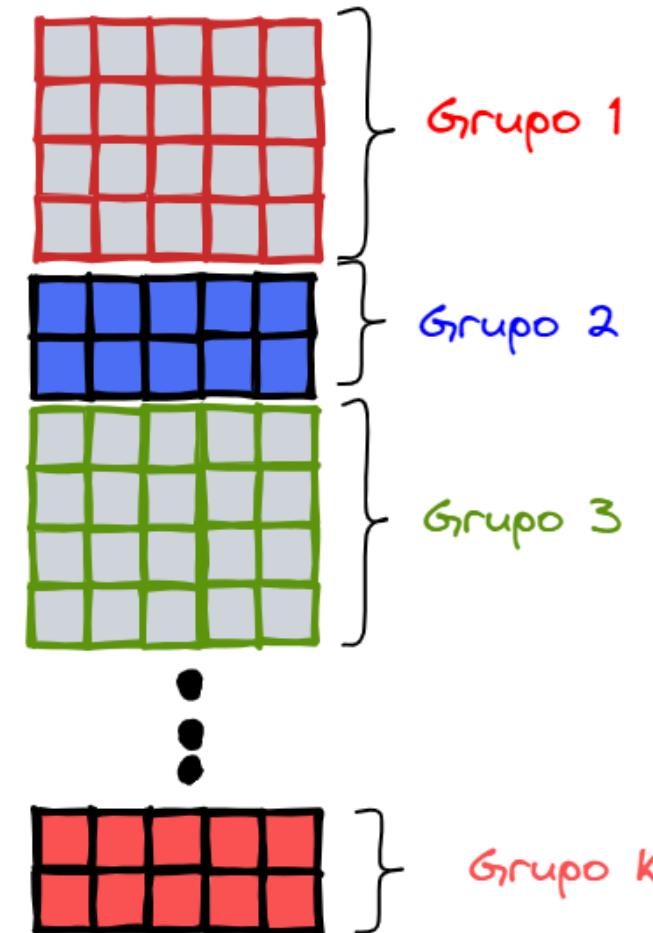


Estudios de corte transversal: cada tema puede ser un grupo de variables.

Integración de datos: características de la víctima, salud, educación, ...

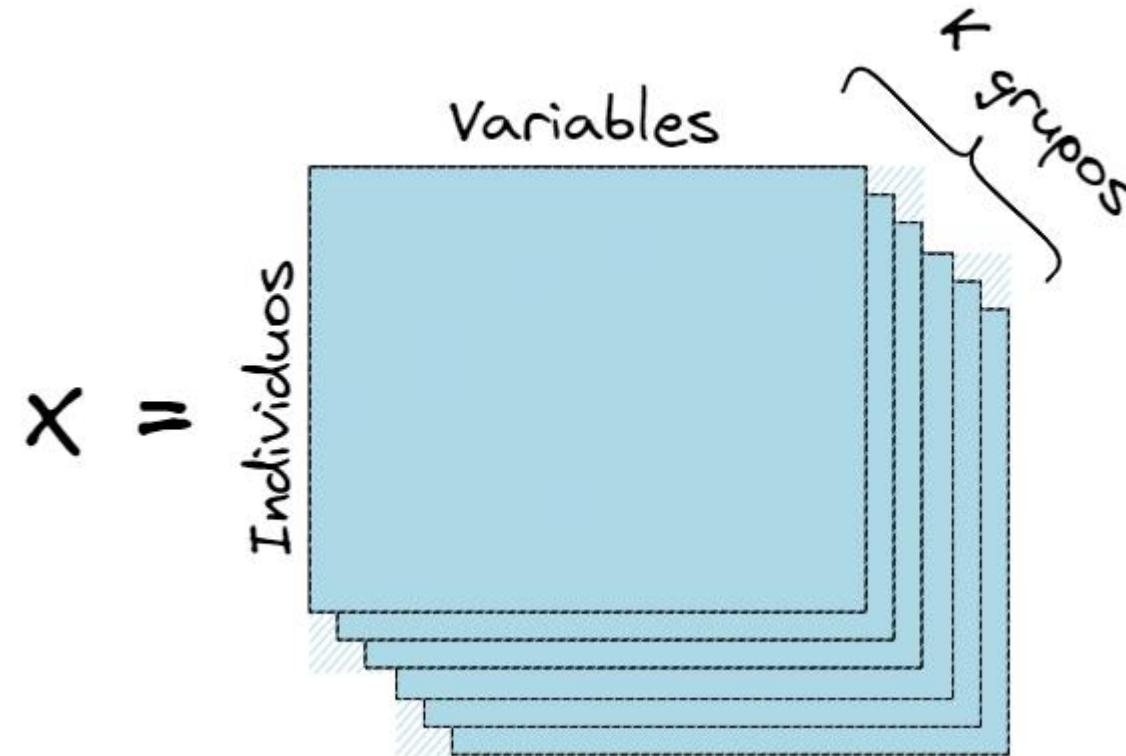
CASO 2. Diferentes individuos (bloques). Se miden las mismas variables

$X =$



Estudios de seguimiento (tracking)
Análisis de los cambios en las dinámicas del conflicto

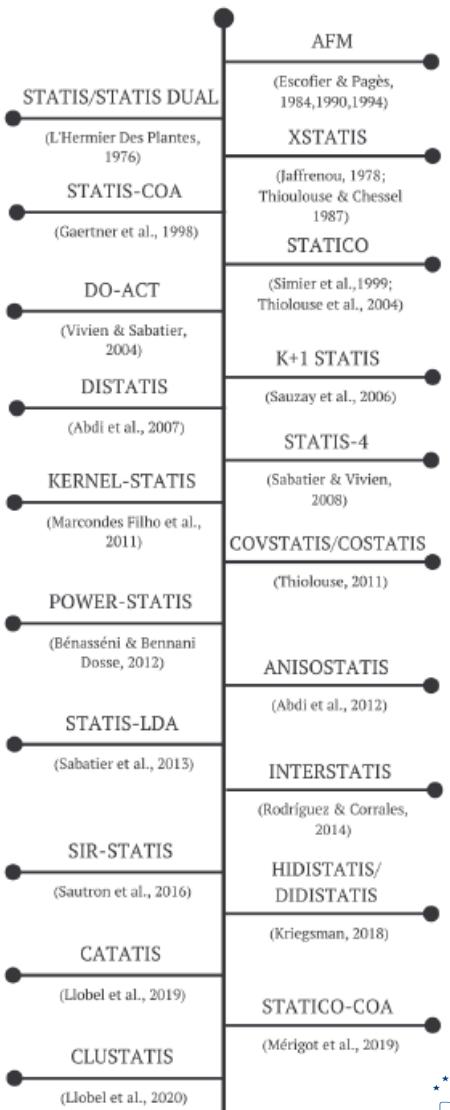
CASO 3. Se miden los mismos individuos y las mismas variables en repetidas ocasiones.



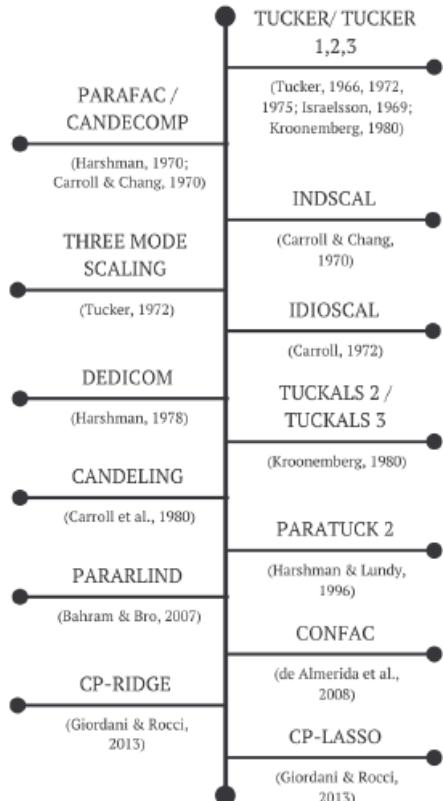
Indicadores económicos de los países en diferentes años.

*En los casos anteriores las matrices podrían representar distancias.

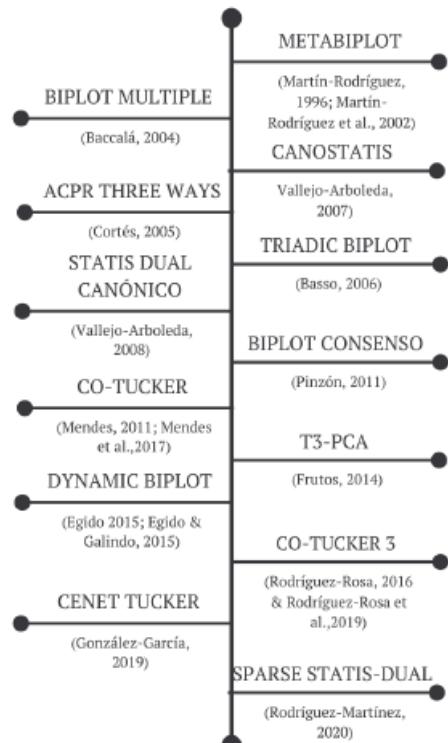
FRANCESAS



ANGLOSAJONA



SALMANTINA



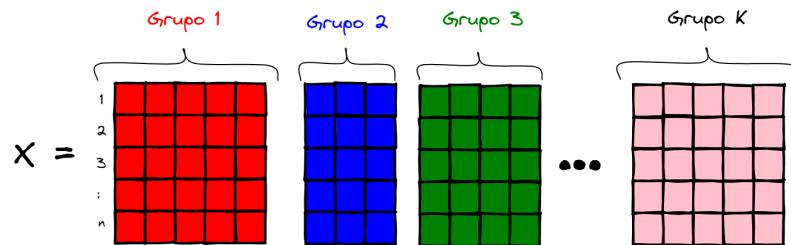
EL MÉTODOS STATIS

(L'HERMIER des PLANTES, 1976). Structuration de Tableaux A Trois Indices de la Statistique

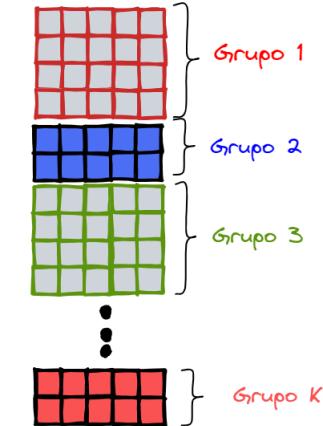
Los métodos STATIS buscan una estructura común a todas las matrices, denominada estructura CONSENSO

Se determinará si la distancia entre individuos, o la estructura de covariación es estable entre una tabla y otra. Además, permite conocer la estructura de asociación entre las tablas.

EL MÉTODOS STATIS



STATIS



STATIS - DUAL

Captura de la información de cada grupo

$$W_k = X_k X_k^\top$$

$$W_k = X_k M_k X_k^\top$$

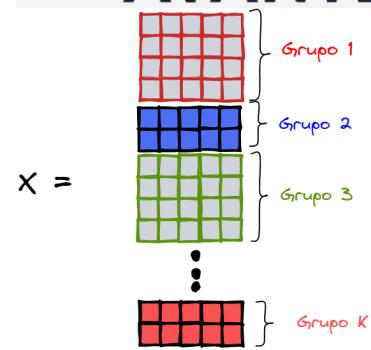
Producto interno de individuos. M es una matriz de ponderaciones de los individuos. Dimensión nxn.

$$C_k = X_k^\top X_k$$

$$C_k = X_k^\top D_k X$$

Asociación entre variables. D es una matriz de ponderaciones de las variables. Dimensión pxp.

EL MÉTODOS STATIS - DUAL



p-Variables

$$C_k = X_k^T X_k$$

$$\begin{bmatrix} X_1 \end{bmatrix}$$



$$\begin{bmatrix} C_1 \end{bmatrix}_{p \times p}$$

$$\begin{bmatrix} X_2 \end{bmatrix}$$



$$\begin{bmatrix} C_2 \end{bmatrix}_{p \times p}$$



$$\begin{bmatrix} X_k \end{bmatrix}$$

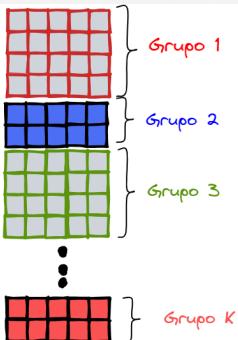


$$\begin{bmatrix} C_K \end{bmatrix}_{p \times p}$$

Matrices con las asociaciones
entre variables

EL MÉTODOS STATIS - DUAL

$X =$



p-Variables

$$\begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}$$

$$\begin{bmatrix} X_2 \\ \vdots \\ X_p \end{bmatrix}$$

⋮

$$\begin{bmatrix} X_k \\ \vdots \\ X_p \end{bmatrix}$$

$$C_k = X_k^T X_k$$

$$\begin{bmatrix} C_1 \\ \vdots \\ C_p \end{bmatrix}$$

$$\begin{bmatrix} C_2 \\ \vdots \\ C_p \end{bmatrix}$$

$$\vdots$$

$$\begin{bmatrix} C_K \\ \vdots \\ C_p \end{bmatrix}$$

Matrices con las asociaciones entre variables

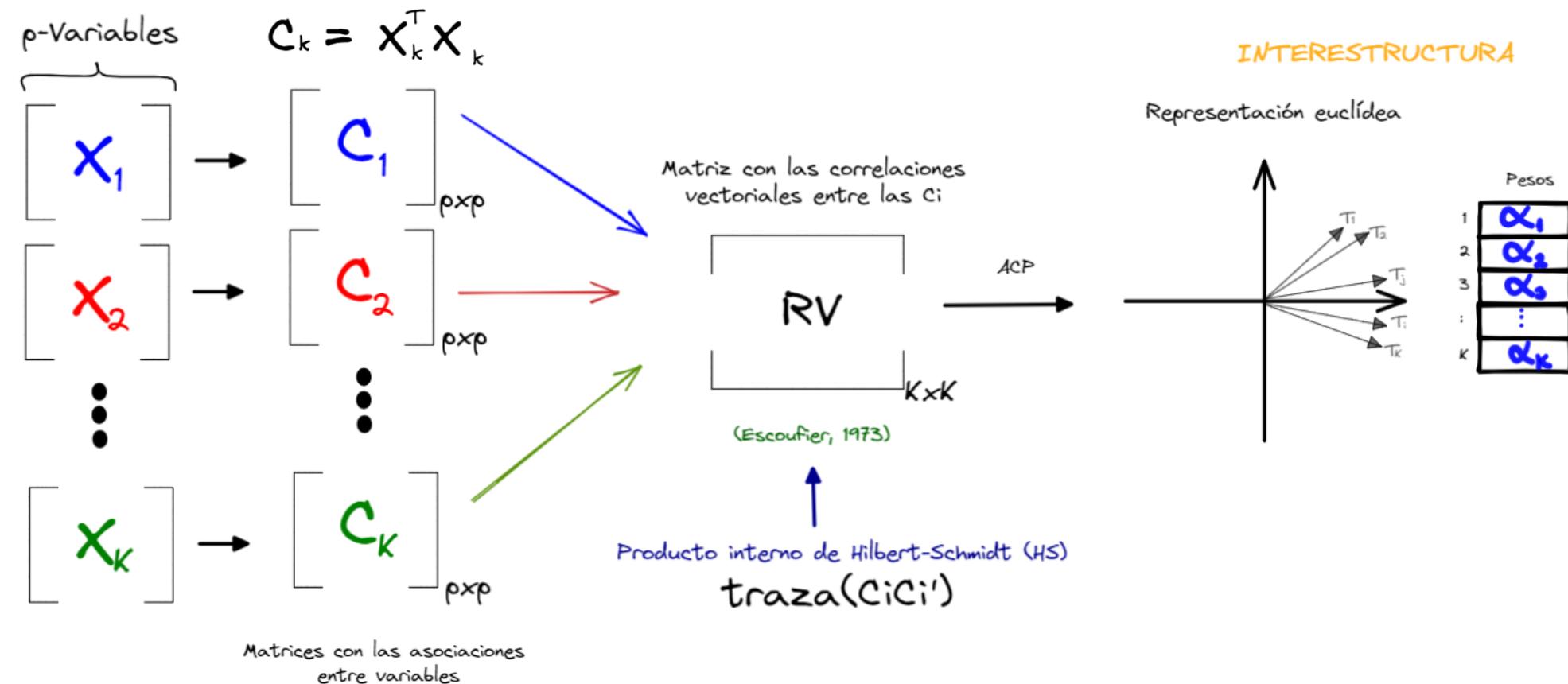
Matriz con las correlaciones vectoriales entre las C_i

$$\begin{bmatrix} RV \\ \vdots \\ RV \end{bmatrix}_{K \times K}$$

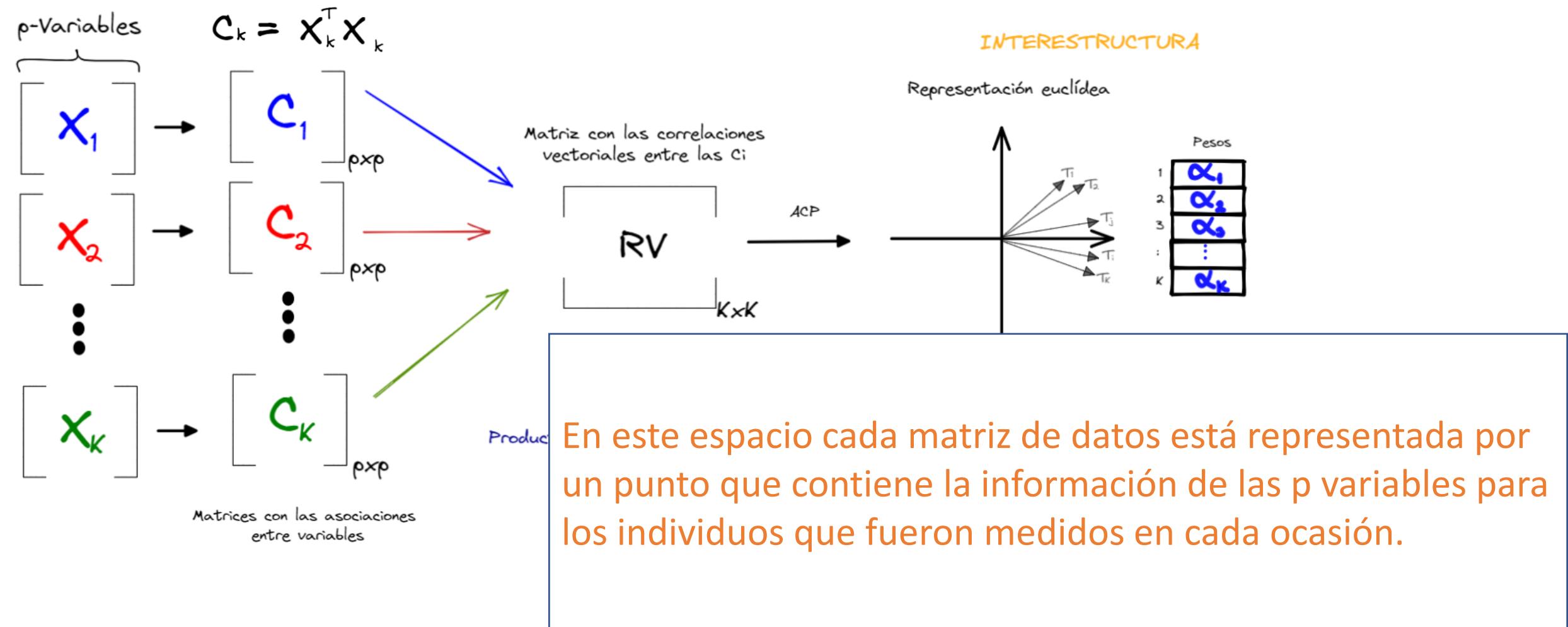
(Escoufier, 1973)

Producto interno de Hilbert-Schmidt (HS)
 $\text{traza}(C_i C_i')$

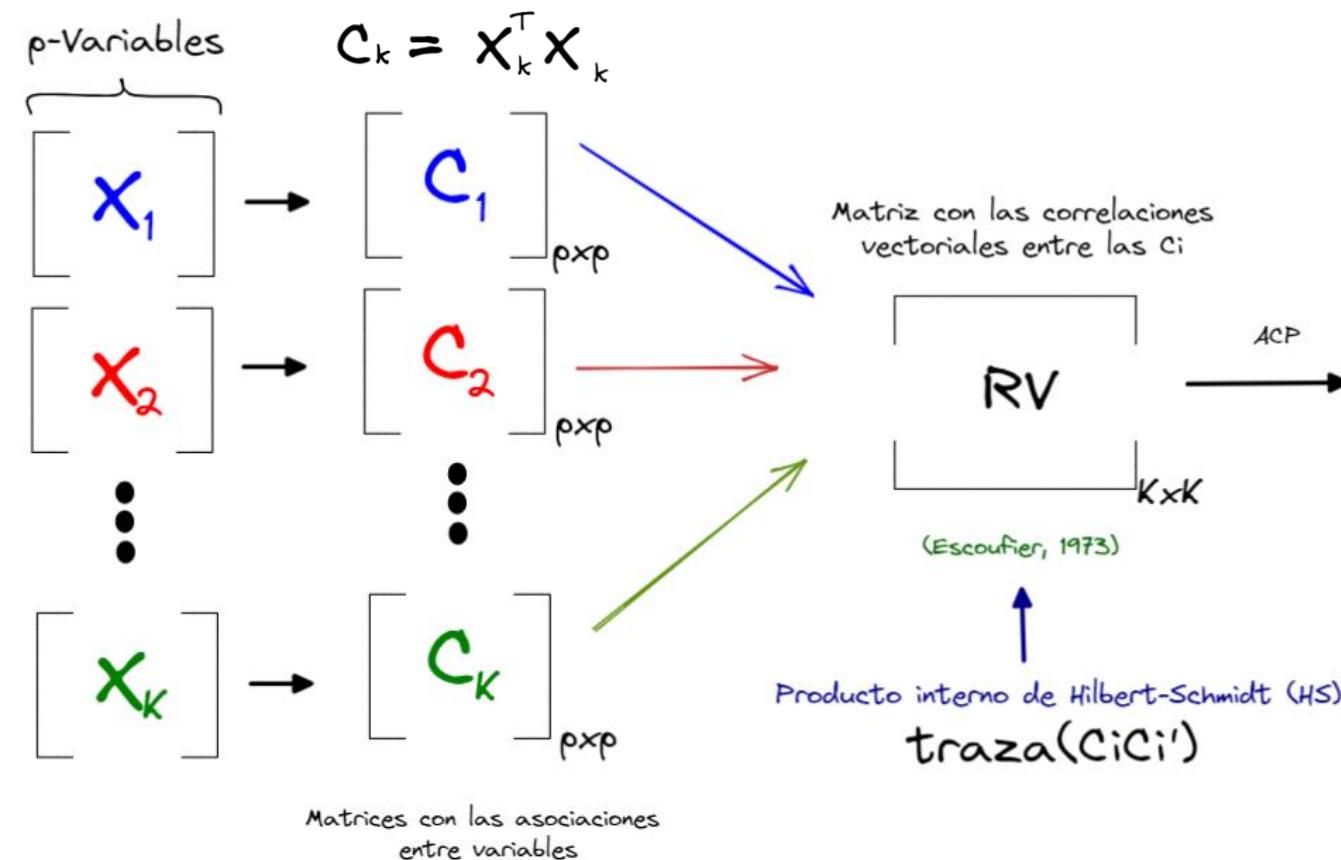
EL MÉTODOS STATIS - DUAL



EL MÉTODOS STATIS - DUAL

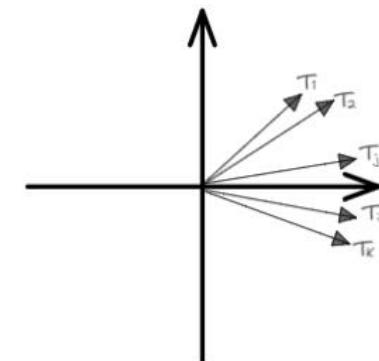


EL MÉTODOS STATIS - DUAL



INTERESTRUCTURA

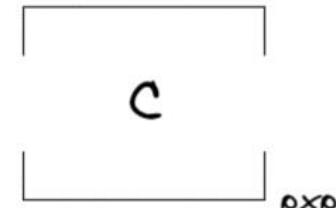
Representación euclídea



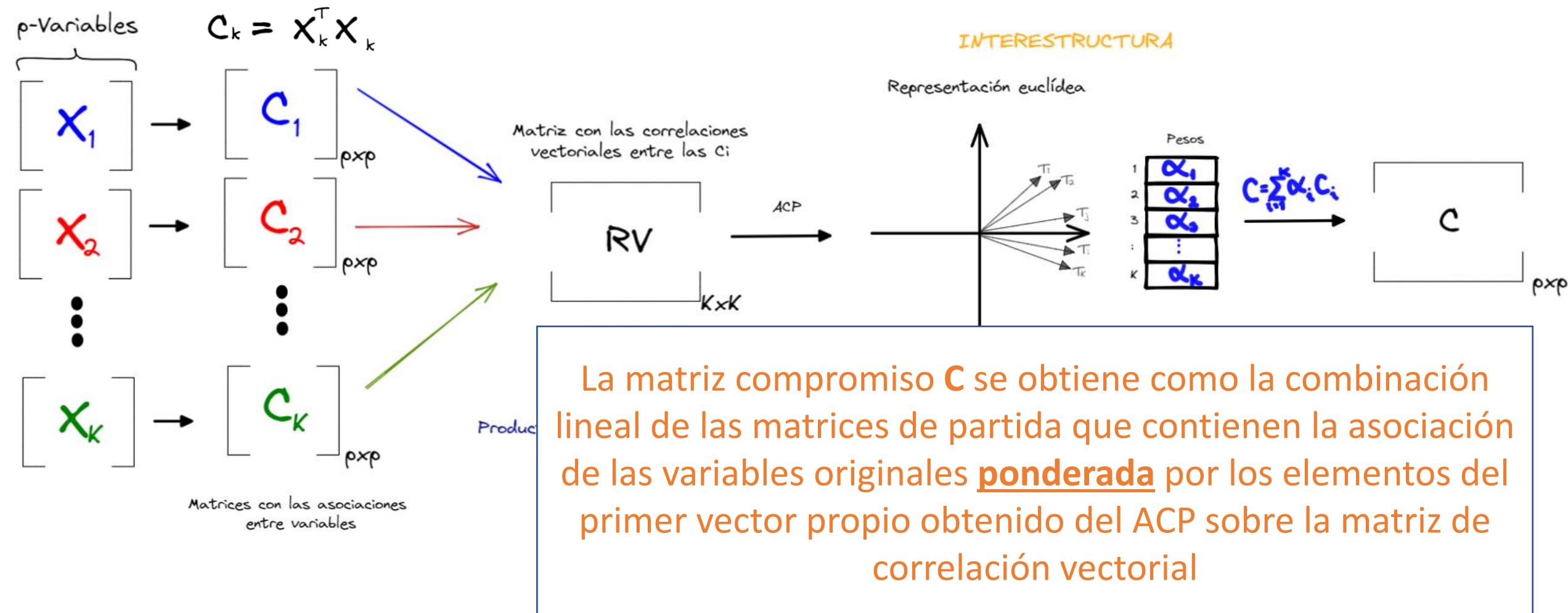
Pesos

1	α_1
2	α_2
3	α_3
:	
K	α_K

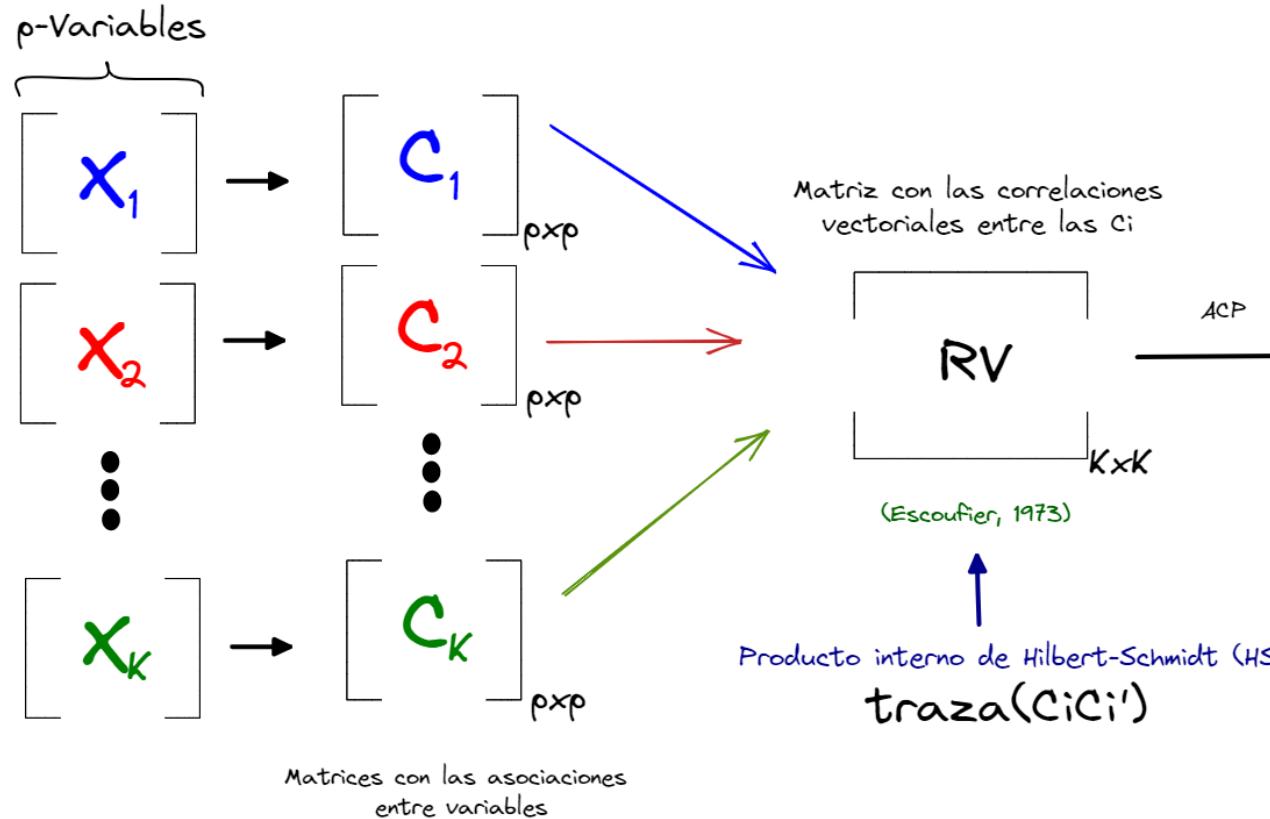
$$C = \sum_{i=1}^K \alpha_i C_i$$



EL MÉTODOS STATIS - DUAL

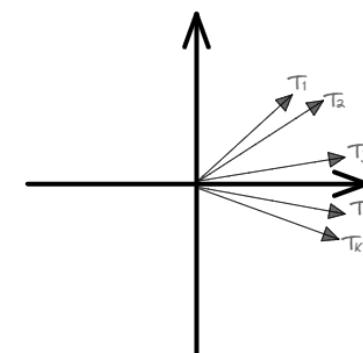


EL MÉTODOS STATIS - DUAL

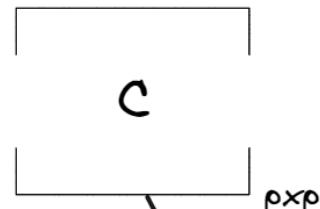


INTERESTRUCTURA

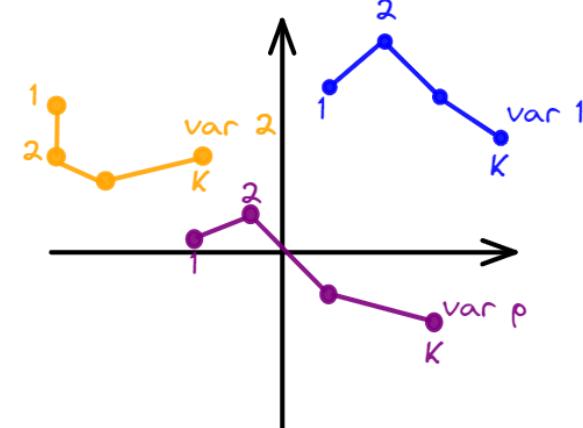
Representación euclídea



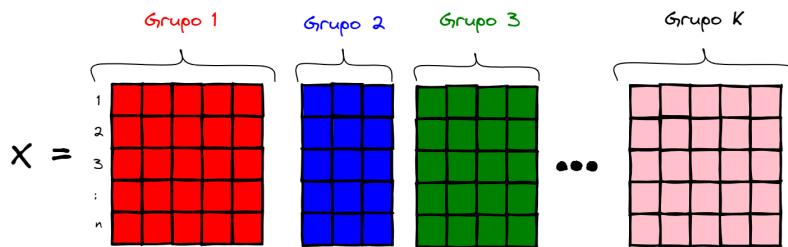
$$C = \sum_{i=1}^K \alpha_i C_i$$



INTRAESTRUCTURA



EL MÉTODOS STATIS

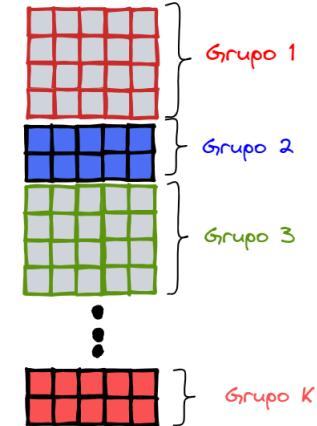


STATIS

$$W_k = X_k X_k^\top$$

Objeto compromiso

$$W = \sum_{i=1}^K \alpha'_i W_i$$



STATIS - DUAL

$$C_k = X_k^\top X_k$$

Objeto compromiso

$$C = \sum_{i=1}^K \alpha_i C_i$$

La matriz compromiso (consenso) es la media ponderada de las configuraciones en cada ocasión, dichas ponderaciones son obtenidas a partir de un ACP sobre la matriz RV.

1. Resumir la información de cada ocasión al nivel de análisis que se requiere (individuos $W=XX'$, variables $C=X'X$).
2. Calcular la matriz de correlación vectorial usando el producto interno de Hilbert-Schmidt (HS).
3. Representar la inter-estructura en un subespacio de dimensión reducida (ACP sobre RV)
4. Analizar la inter-estructura, se espera que la primera componente explique una proporción alta de la información.
5. Configurar la matriz compromiso o consenso, usando como ponderadores al primer vector propio de la matriz RV.
6. Analizar la intra-estructura.
7. Analizar las trayectorias.

EJEMPLO 1

El conflicto armado interno en Colombia lleva más de 5 décadas. El Registro Único de Víctimas reportaba en el año 2021 un total de 9.134.347 víctimas entre desaparición forzada, amenazas, secuestros, homicidios, reclutamiento de menores, desplazamiento, entre otras modalidades de violencia que han ocurrido en más de 10 millones de eventos.

En el año 2016 el Gobierno de Colombia y la guerrilla de las FARC firmaron el acuerdo de paz, con lo cual se creó la Comisión para el Esclarecimiento de la Verdad, la Convivencia y la No Repetición (CEV) y, la Jurisdicción Especial para la Paz (JEP), con el fin de tener un proceso de justicia transicional.

La CEV recopiló datos de más de 500 fuentes buscando explicar las dinámicas del conflicto armado.

EJEMPLO 2

El artículo de Abdi et al. ([2012](#)) se presentan diferentes conjuntos de datos que permiten realizar análisis multivariante para tablas múltiples.

Para este primer ejemplo, los autores seleccionaron doce (12) vinos elaborados con uvas Sauvignon Blanc provenientes de tres (3) regiones (Nueva Zelanda, Francia y Canadá) y se eligieron cuatro (4) vinos de cada región.

Se le pidió a 10 asesores expertos que evaluaran estos vinos, para ello se usaron escalas de calificación de 9 puntos, utilizando cuatro variables consideradas estándar para la evaluación de estos vinos (cat-peel, maracuyá, pimiento verde y mineral). Y segundo, si sentían la necesidad, los evaluadores tenían la libertad de añadir variables propias (algunos evaluadores no eligen ninguna, algunos eligen una, dos o más variables). Realizar un análisis basado en la estructura de los datos.

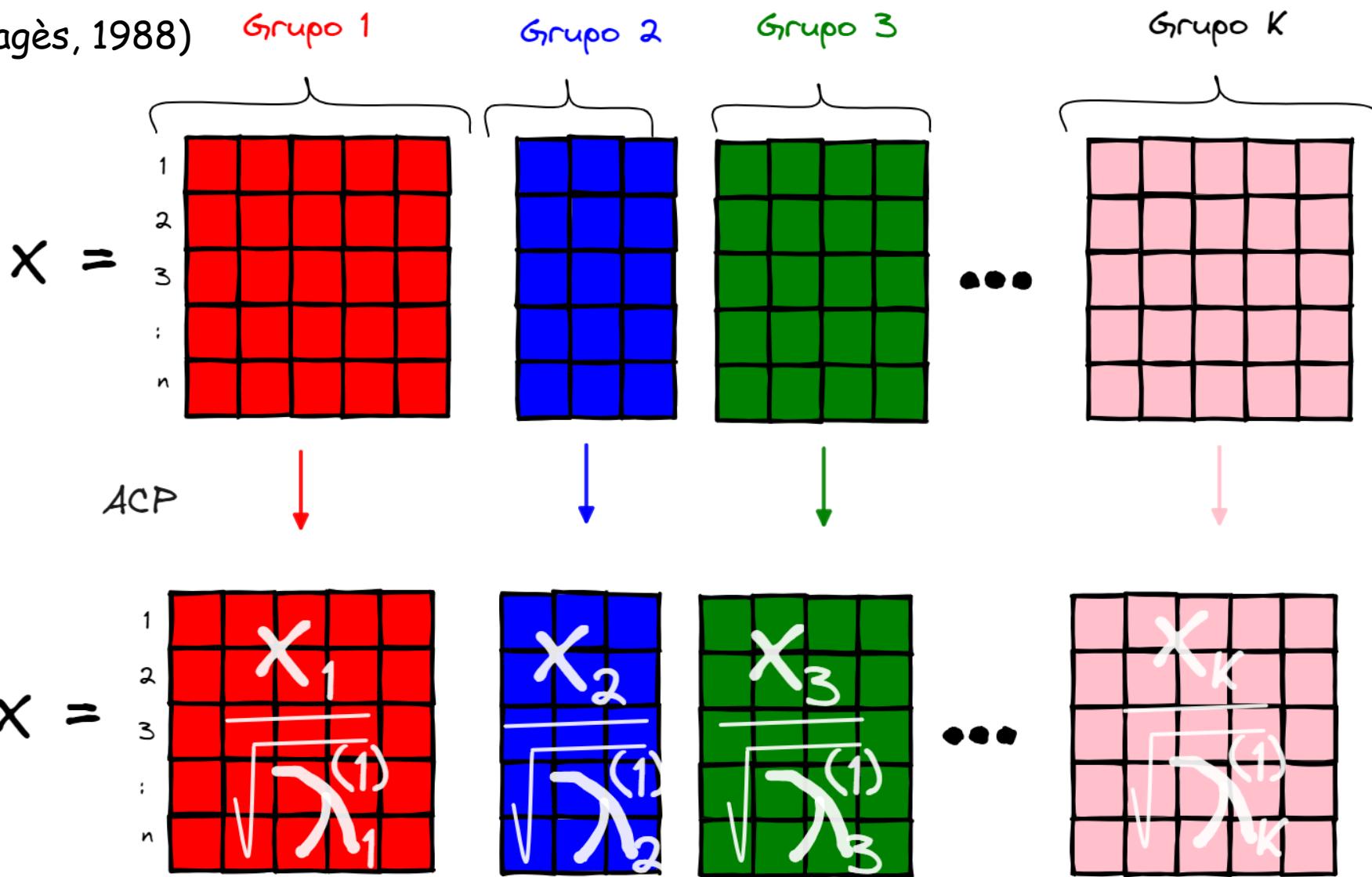
EJEMPLO 3

La famosa tesis de Jean Verneaux (1973), parte de los datos revisados por el autor están disponible por **data(JV73)** en la biblioteca **ade4**.

El objeto es una lista de 6 componentes: morpho, phychi, poi, xy, contour y fac.riv. Originalmente había 111 estaciones repartidas a lo largo de 12 ríos. Se eliminan las algunas estaciones colocadas en los embalses del Alto Doubs y que no están relacionados con el problema planteado. El conjunto de datos queda compuesto por 92 estaciones.

ANÁLISIS FACTORIAL MÚLTIPLE

(Escoufier & Pagès, 1988)



Propiedades

- El número de variables en cada grupo puede ser diferente.
- La naturaleza de las variables en cada grupo debe ser la misma.
- La naturaleza de las variables entre grupos no necesariamente debe ser la misma.

Se le asignará un peso a cada variable teniendo en cuenta:

- Dentro de un grupo, el peso para cada variable debe ser el mismo con el fin de no distorsionar la inercia dentro del grupo.
- Lograr que la máxima inercia de un eje sea igual a 1. Como esto depende del valor propio más grande, se atribuirá a cada variable del grupo j un peso de $w_{ij} = \frac{1}{\sqrt{\lambda_1^j}}$ donde λ_1^j es el primer valor propio de un ACP separado para el grupo j .

$$\left[\frac{x_1}{\sqrt{\lambda_1^1}}; \frac{x_2}{\sqrt{\lambda_1^2}}; \dots; \frac{x_j}{\sqrt{\lambda_1^j}} \right]$$

APLICACIÓN: ANÁLISIS SENSORIAL



Las variables que se miden regularmente están asociadas a los sentidos del olfato, visión, gusto, etc. medidas en escalas hedonistas o numéricas

Olfato

- Intensidad
- Calidad
- Tipo de aroma

Gusto

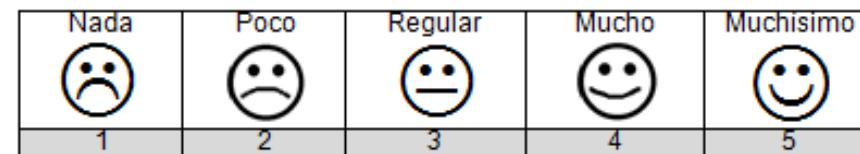
- Amargo
- Sal
- Dulce
- Astringencia

Visión

- Empaque
- Color
- ...

Escalas hedonistas

P12.(MOSTRAR TARJETA DE GUSTO) ¿QUE TANTO TE GUSTÓ EL SABOR EN GENERAL DE ESTA GALLETITA QUE ACABASTE DE PROBAR? RU



Escala numérica discreta

No me gustó para nada						Me gusto Demasiado
1	2	3	4	5	6	7

Escala numérica continua

No me gustó para nada



Me gusto demasiado

APLICACIÓN: ANÁLISIS SENSORIAL

Se cuenta con los datos para 21 tipos de vinos Val de Loire a los que se les midió 31 variables (2 categóricas). La primera variable corresponde a la etiqueta de origen, la segunda corresponde al suelo y las demás corresponden a descriptores sensoriales.

	Label	Soil	Odor.intensity. Aroma.quality. before.shaking before.shaking			Visual. intensity Nuance ...			Odor. Intensity of.odour ...	Quality. of odour ...	Attack. intensity	Acidity ...	Overall. quality	Typical
2EL	Saumur	Env1	3.074	3	...	4.321	4	...	3.407	3.308	...	2.963	2.107	...
1CHA	Saumur	Env1	2.964	2.821	...	3.222	3	...	3.37	3	...	3.036	2.107	...
1FON	Bourgueuil	Env1	2.857	2.929	...	3.536	3.393	...	3.25	2.929	...	3.222	2.179	...
1VAU	Chinon	Env2	2.808	2.593	...	2.893	2.786	...	3.16	2.88	...	2.704	3.179	...
1DAM	Saumur	Reference	3.607	3.429	...	4.393	4.036	...	3.536	3.36	...	3.464	2.571	...
2BOU	Bourgueuil	Reference	2.857	3.111	...	4.464	4.259	...	3.179	3.385	...	3.286	2.393	...
1BOI	Bourgueuil	Reference	3.214	3.222	...	4.143	3.929	...	3.429	3.5	...	3.393	2.607	...
3EL	Saumur	Env1	3.12	2.852	...	4.214	3.857	...	3.654	3.077	...	3.25	2.179	...
DOM1	Chinon	Env1	2.857	2.815	...	4.037	3.893	...	3.357	3.346	...	3.286	2.286	...
1TUR	Saumur	Env2	2.893	3	...	3.704	3.407	...	3.222	3.259	...	2.893	2.357	...
4EL	Saumur	Env2	3.25	3.286	...	3.857	3.643	...	3.607	3.385	...	3.321	2.429	...
PER1	Saumur	Env2	3.393	3.179	...	4.714	4.5	...	3.481	3.385	...	3.357	2.429	...
2DAM	Saumur	Reference	3.179	3.286	...	4.222	4.071	...	3.481	3.423	...	3.393	2.286	...
1POY	Saumur	Reference	3.071	3.107	...	4.714	4.536	...	3.357	3.444	...	3.519	2.111	...
1ING	Bourgueuil	Env1	3.107	3.143	...	4.071	3.893	...	3.357	3.37	...	3.185	2.286	...
1BEN	Bourgueuil	Reference	2.929	3.179	...	3.889	3.429	...	3.286	3.308	...	3.393	2.393	...
2BEA	Chinon	Reference	3.036	3.179	...	3.786	3.607	...	3.444	3.5	...	3.071	2.571	...
1ROC	Chinon	Env2	3.071	2.926	...	3.679	3.393	...	3.37	3.36	...	3.071	2.393	...
2ING	Bourgueuil	Env1	2.643	2.786	...	2.607	2.536	...	2.889	2.8	...	2.179	2.25	...
T1	Saumur	Env4	3.696	3.192	...	4.321	4	...	3.737	3.08	...	2.963	2.407	...
T2	Saumur	Env4	3.708	2.926	...	4.321	4.107	...	3.727	2.885	...	3.333	2.571	...

Considere los siguientes grupos de variables para el análisis:

Grupo 1: 2 variables que representan el origen del vino – c(1:2)

Grupo 2: 5 variables para el olor de los vinos antes de agitar -- c(3:7)

Grupo 3: 3 variables para los aspectos visuales

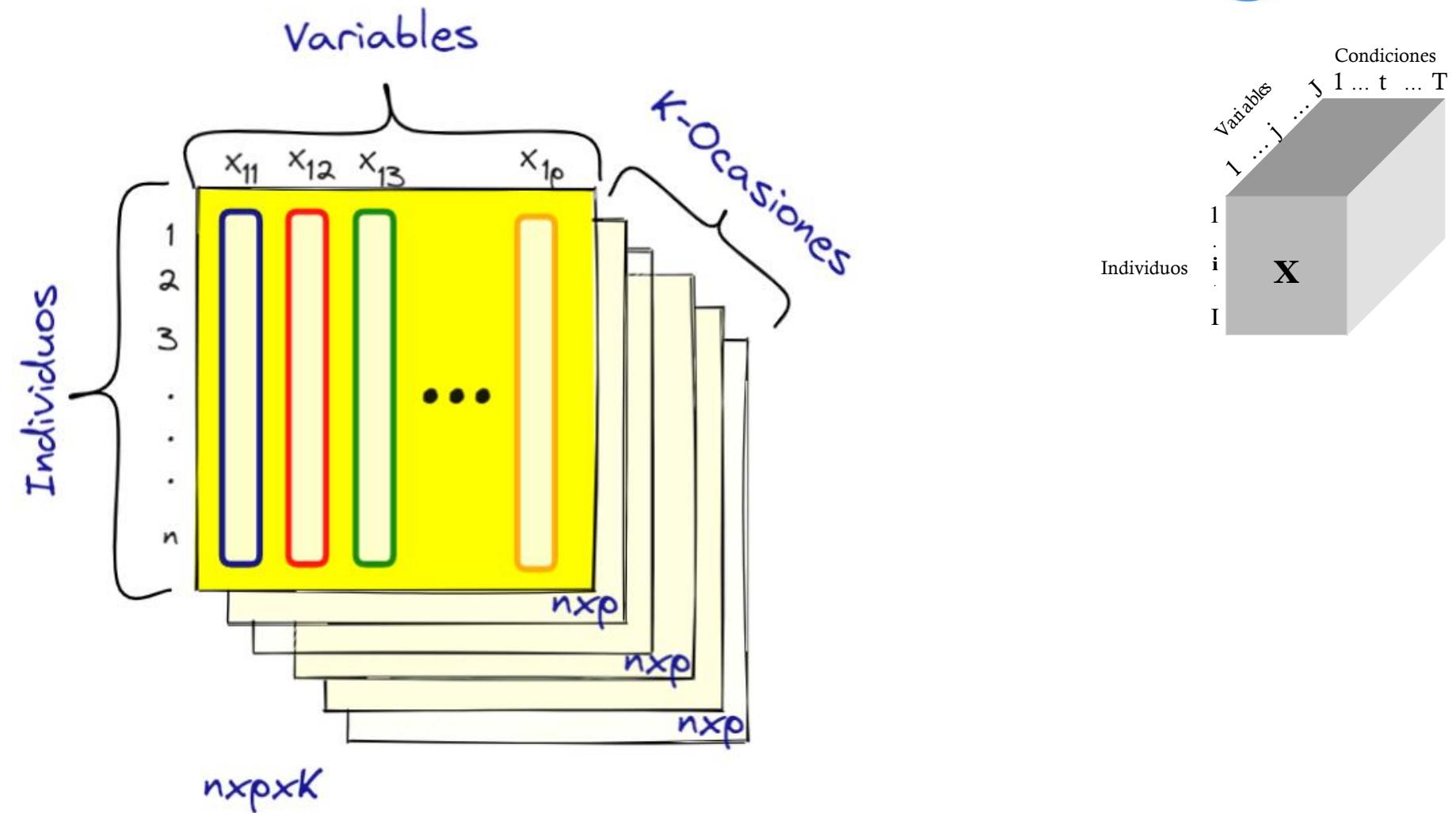
Grupo 4: 10 variables para el olor después de agitar

Grupo 5: 9 variables que representan el sabor de los vinos.

Grupo 6: 2 variables para el desempeño general

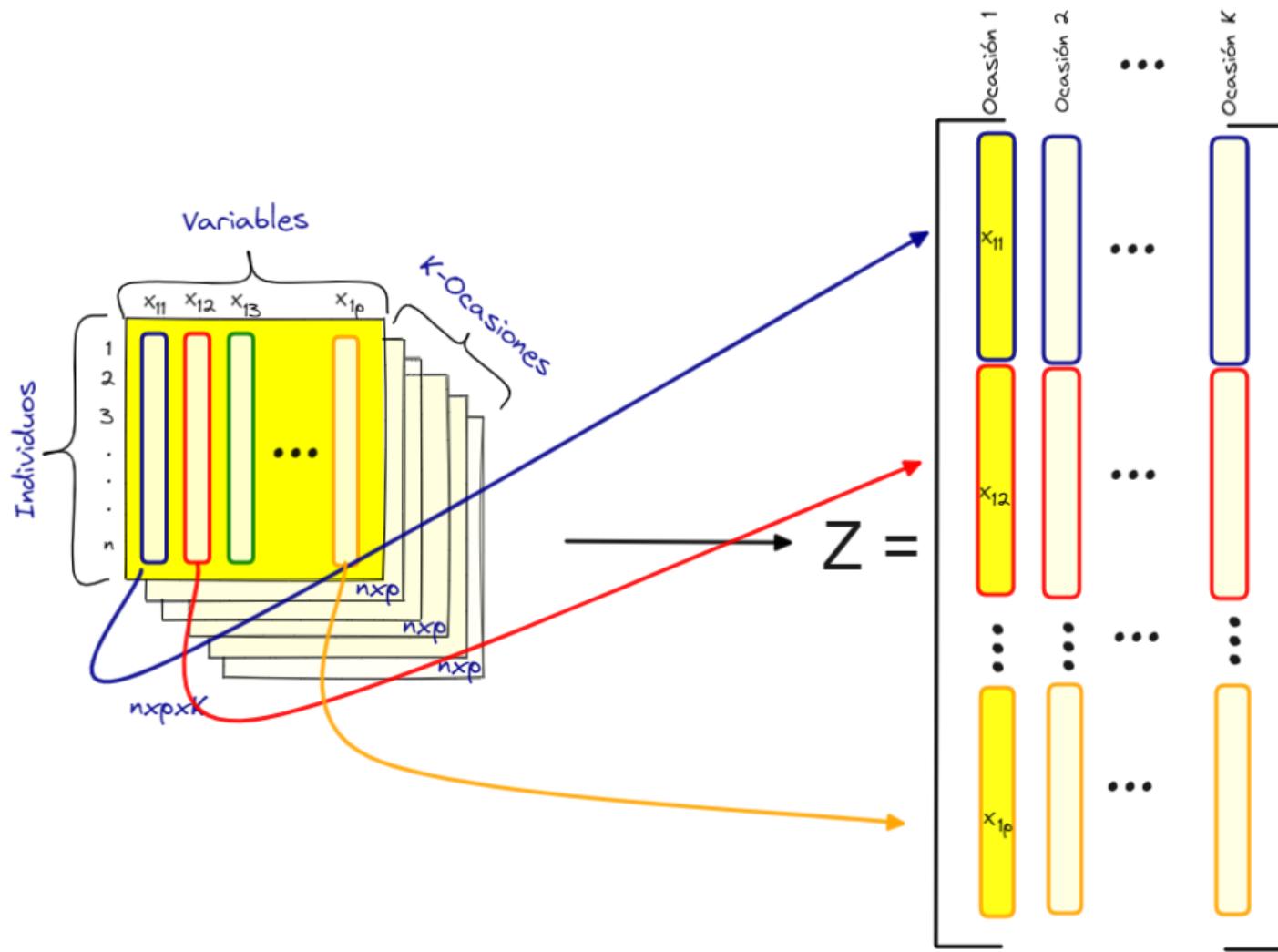
```
> data(wine)
> colnames(wine)
[1] "Label"                               "Soil"
[3] "Odor.Intensity.before.shaking"      "Aroma.quality.before.shaking"
[5] "Fruity.before.shaking"                "Flower.before.shaking"
[7] "Spice.before.shaking"                 "Visual.intensity"
[9] "Nuance"                               "Surface.feeling"
[11] "Odor.Intensity"                     "Quality.of.odour"
[13] "Fruity"                                "Flower"
[15] "Spice"                                 "Plante"
[17] "Phenolic"                             "Aroma.intensity"
[19] "Aroma.persistency"                   "Aroma.quality"
[21] "Attack.intensity"                    "Acidity"
[23] "Astringency"                          "Alcohol"
[25] "Balance"                               "Smooth"
[27] "Bitterness"                           "Intensity"
[29] "Harmony"                              "Overall.quality"
[31] "Typical"
```

ANÁLISIS TRIADICO (PTA, X-STATIS)

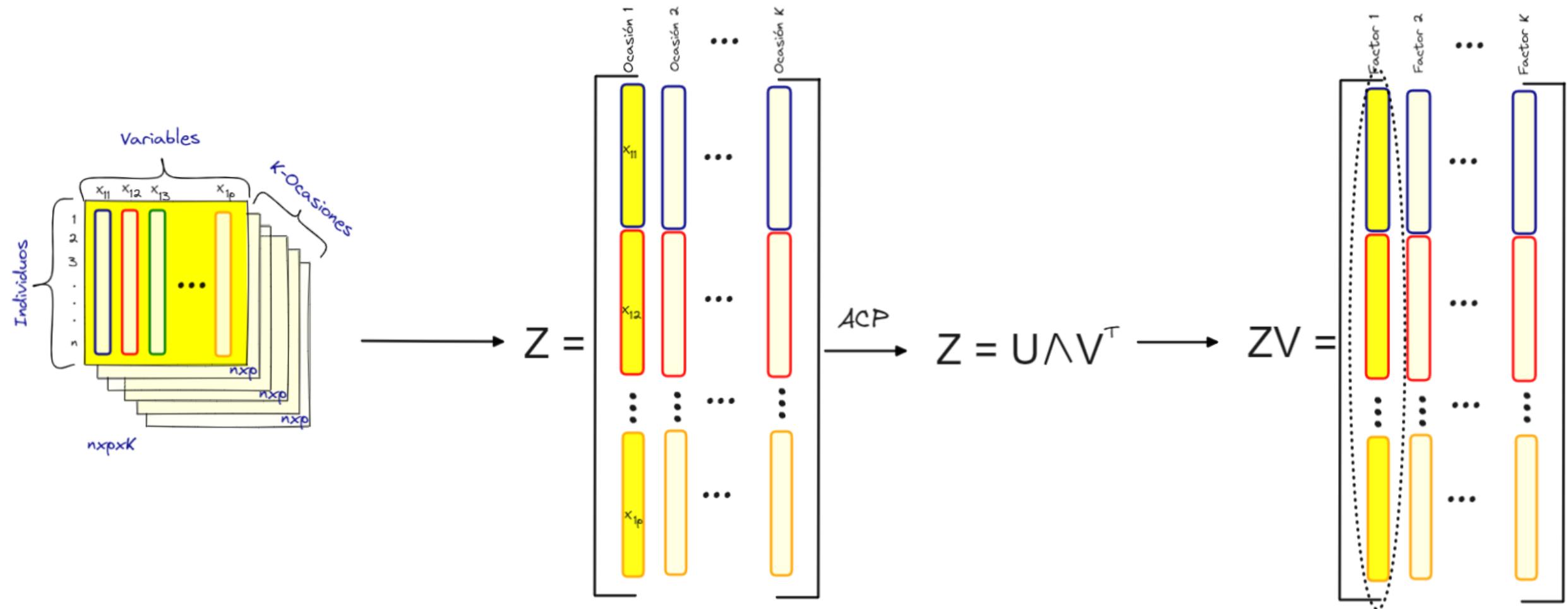


En este caso se han medido las mismas p variables a los mismos n individuos durante K - ocasiones

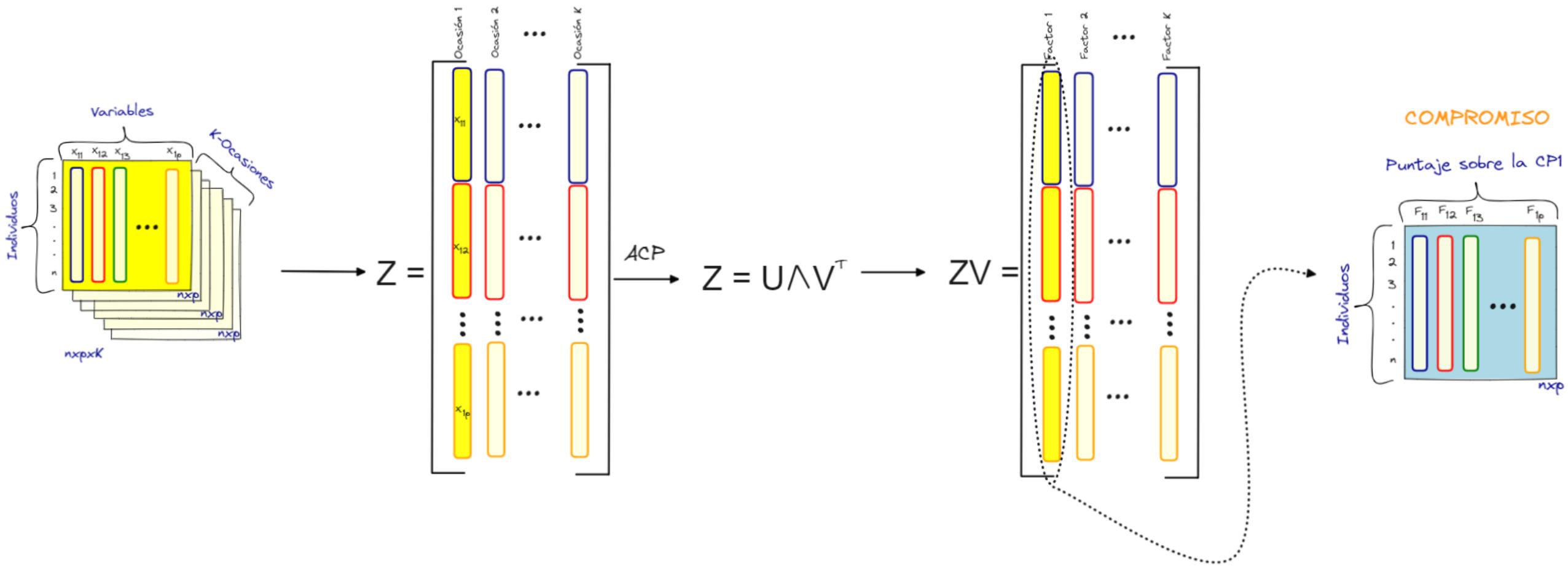
ANÁLISIS PTA (X-STATIS)



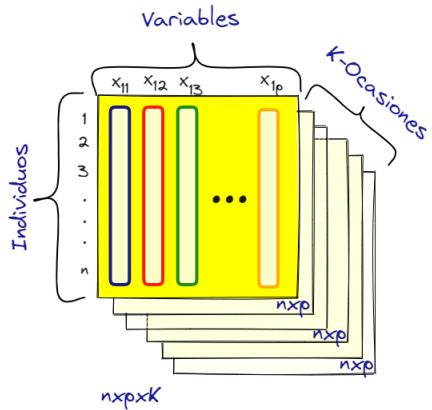
ANÁLISIS PTA (X-STATIS)



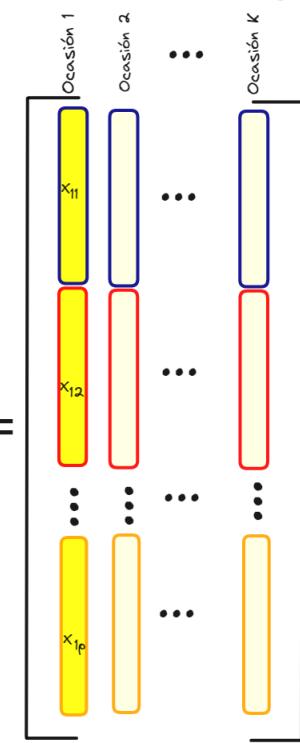
ANÁLISIS PTA (X-STATIS)



ANÁLISIS PTA (X-STATIS)

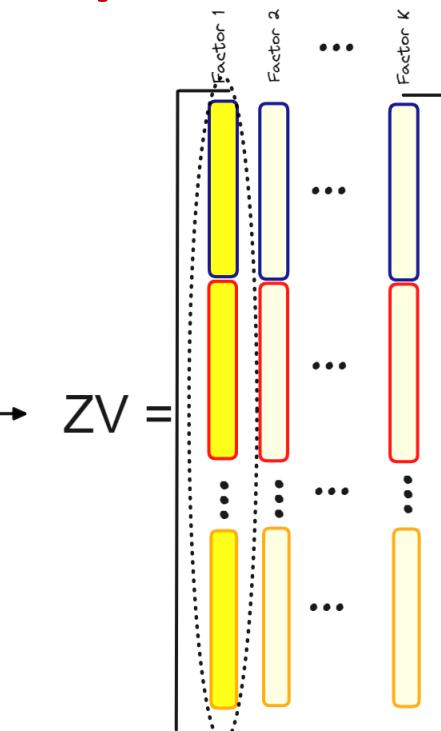


$$Z =$$



ACP

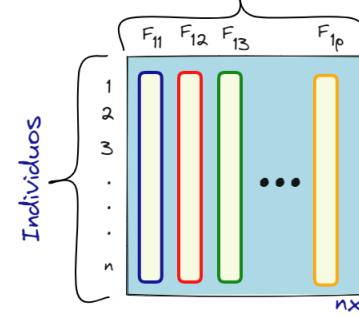
$$Z = U \Lambda V^T$$



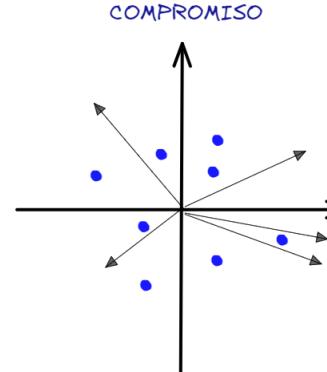
COMPROMISO

Puntaje sobre la CPI

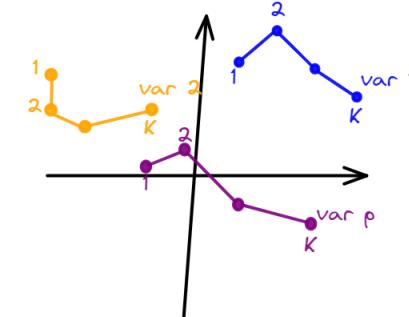
Individuos



ACP



Proyección de filas y columnas del compromiso

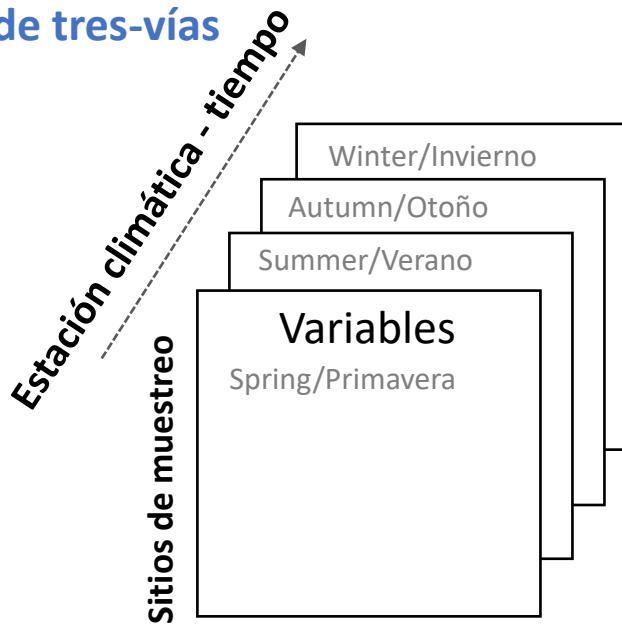


Proyección de filas y columnas de cada tabla en el espacio del compromiso

APLICACIÓN: DATOS ECOLÓGICOS

Se tienen los datos sobre cinco lugares de muestreo a lo largo de un pequeño arroyo francés (el Meaudret), los cuales fueron recogidos cuatro veces, en primavera, verano, otoño e invierno y se midieron variables ambientales (químicas) y biológicas.

Estructura de la matriz de tres-vías



id	estacion	sitio	Temp	Flow	pH	Cond	Bdo5	Oxyd	Ammo	Nitr	Phos
sp_1	spring	S1	10	41	8,5	295	2,3	1,4	0,12	3,4	0,11
sp_2	spring	S2	11	158	8,3	315	7,6	3,3	2,85	2,7	1,5
sp_3	spring	S3	11	198	8,5	290	3,3	1,5	0,4	4	0,1
sp_4	spring	S4	12	280	8,6	290	3,5	1,5	0,45	4	0,73
sp_5	spring	S5	13	322	8,5	285	3,6	1,6	0,48	4,6	0,84
su_1	summer	S1	13	62	8,3	325	2,3	1,8	0,11	3	0,13
su_2	summer	S2	13	80	7,6	380	21	5,7	9,8	0,8	3,65
su_3	summer	S3	15	100	7,8	385	15	2,5	7,9	7,7	4,5
su_4	summer	S4	16	140	8	360	12	2,6	4,9	8,4	3,45
su_5	summer	S5	15	160	8,4	345	1,7	1,9	0,22	10	1,74
au_1	autumn	S1	1	25	8,4	315	1,6	0,5	0,07	6,4	0,03
au_2	autumn	S2	3	63	8	425	36	8	12,5	2,2	6,5
au_3	autumn	S3	2	79	8,1	350	7,1	1,9	2,7	13,2	3,7
au_4	autumn	S4	3	85	8,3	330	2	1,4	0,42	12	1,6
au_5	autumn	S5	2	72	8,6	305	1,6	0,9	0,1	9,5	1,25
wi_1	winter	S1	3	118	8	325	1,6	1,2	0,17	1,8	0,19
wi_2	winter	S2	3	252	8,3	360	9,5	2,9	2,52	4,6	1,6
wi_3	winter	S3	3	315	8,3	370	8,7	2,8	2,8	4,8	2,85
wi_4	winter	S4	3	498	8,3	330	4,8	1,6	1,04	4,4	0,82
wi_5	winter	S5	2	390	8,2	330	1,7	1,2	0,56	5	0,6

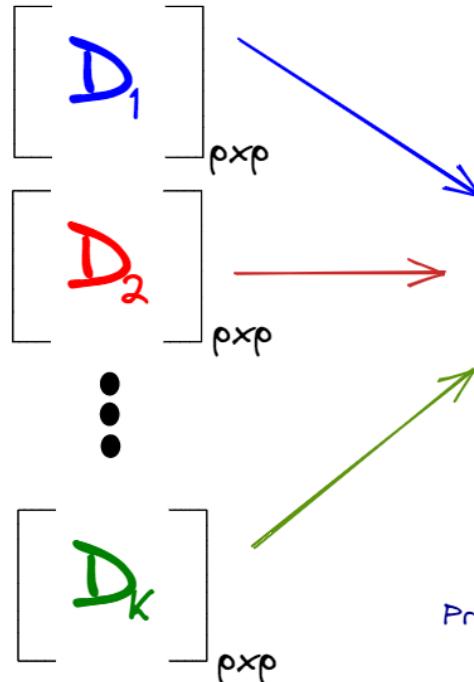
K= 4 estaciones climáticas

p= 9 mismas variables ambientales (físico-químicas)

n = 5 mismos individuos – sitios de muestreo

Mismas variables y mismas filas en todas las ocasiones

p -Variables



Matriz con las correlaciones vectoriales entre las D_i

RV $K \times K$

(Escoufier, 1973)

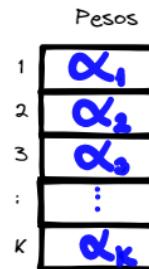
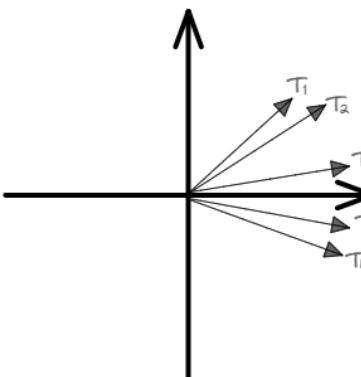
Producto interno de Hilbert-Schmidt (HS)
 $\text{traza}(D_i D_i')$

(Abdi y Col, 2007).

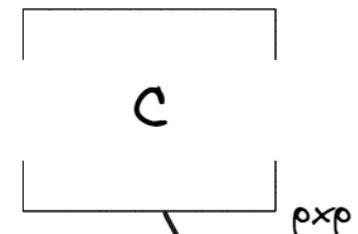
Se presenta como una generalización del MDS para el caso de tablas con 3 entradas.

INTERESTRUCTURA

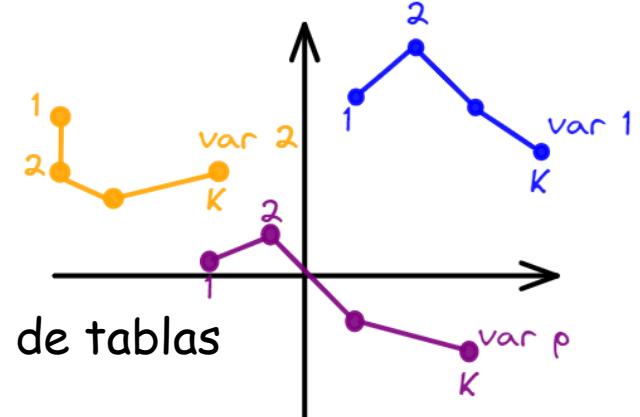
Representación euclídea



$$D = \sum_{i=1}^K \alpha_i D_i$$



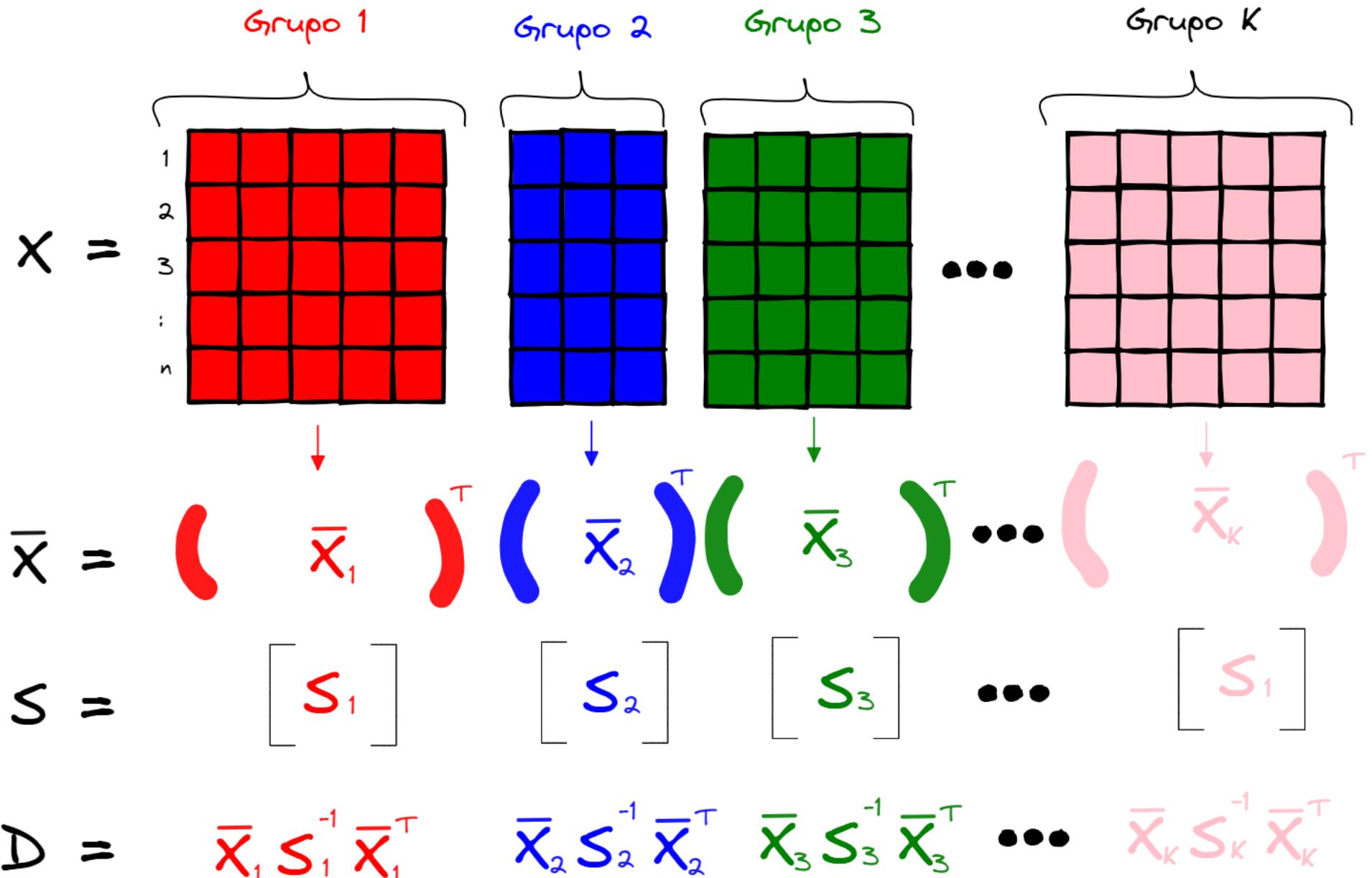
INTRAESTRUCTURA



1. Este método busca diferencias entre los grupos.
2. Se calcula la distancia de Mahalanobis para cada tabla (similar a un análisis discriminante lineal).
3. Con las matrices de distancia se aplica un análisis DISTATIS para integrarlos y representarlos en un compromiso.

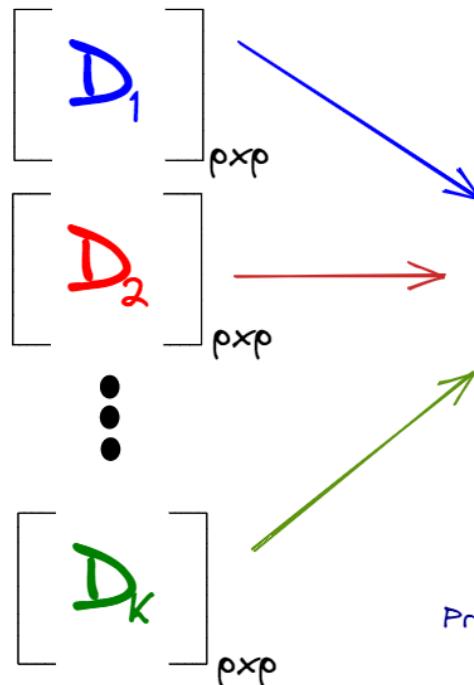
Vallejo-Arboleda A, Vicente-Villardón JL, Galindo- Villardón MP. Canonical STATIS: Biplot analysis of multi-table group structured data based on STATISACT methodology. *Comp Stat Data Anal* 2007, 51:4193-4205.

STATIS CANÓNICO - CANOSTATIS

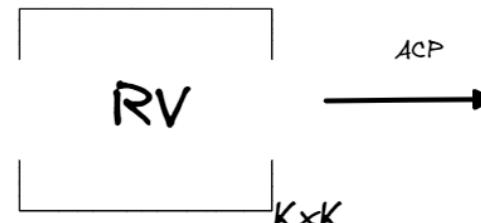


STATIS CANÓNICO - CANOSTATIS

p -Variables



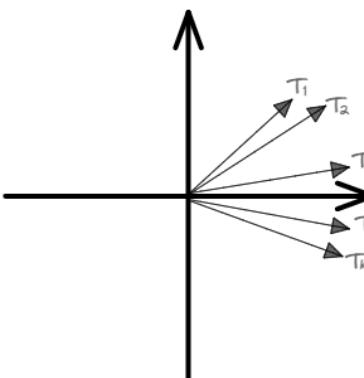
Matriz con las correlaciones vectoriales entre las D_i



Producto interno de Hilbert-Schmidt (HS)
 $\text{traza}(D_i D_i')$

INTERESTRUCTURA

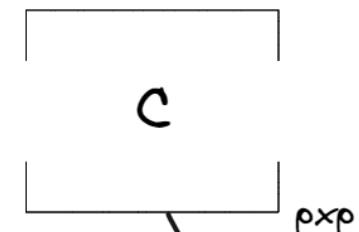
Representación euclídea



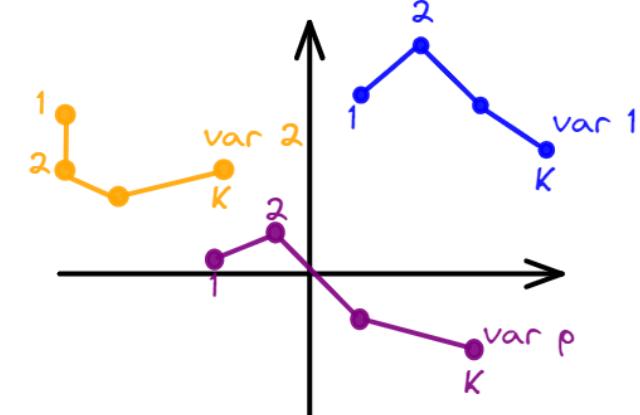
Pesos

1	α_1
2	α_2
3	α_3
:	
K	α_K

$$D = \sum_{i=1}^K \alpha_i D_i$$

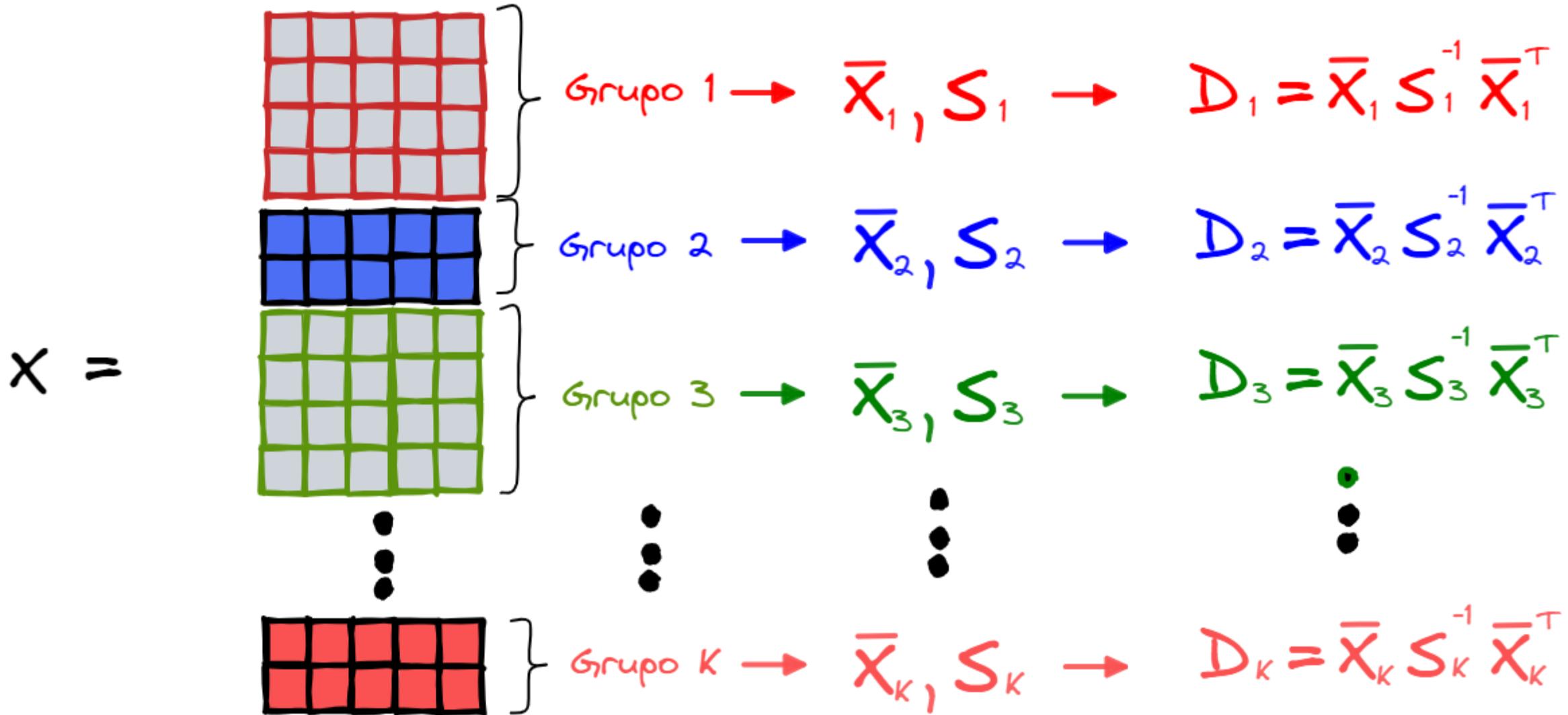


INTRAESTRUCTURA



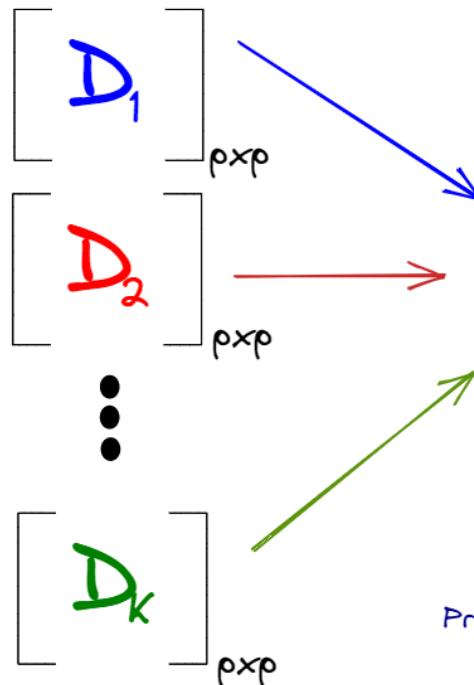
Trayectorias

STATIS CANÓNICO – CANOSTATIS (DUAL)

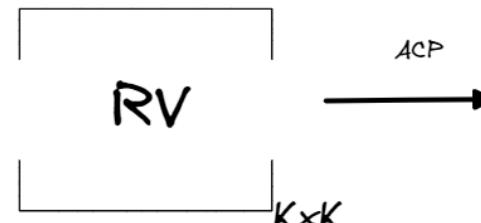


STATIS CANÓNICO - CANOSTATIS

p -Variables



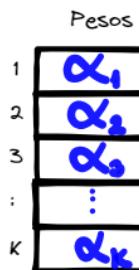
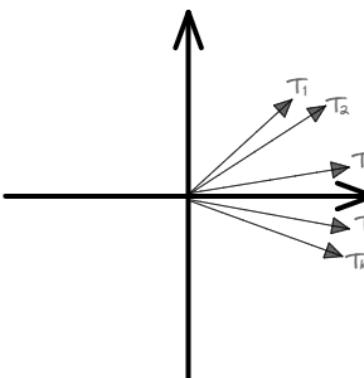
Matriz con las correlaciones vectoriales entre las D_i



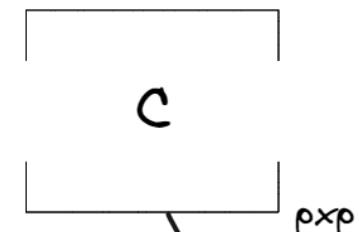
Producto interno de Hilbert-Schmidt (HS)
 $\text{traza}(D_i D_i')$

INTERESTRUCTURA

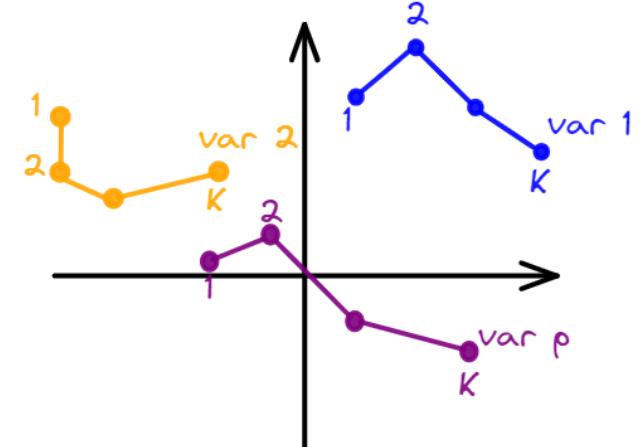
Representación euclídea



$$D = \sum_{i=1}^k \alpha_i D_i$$



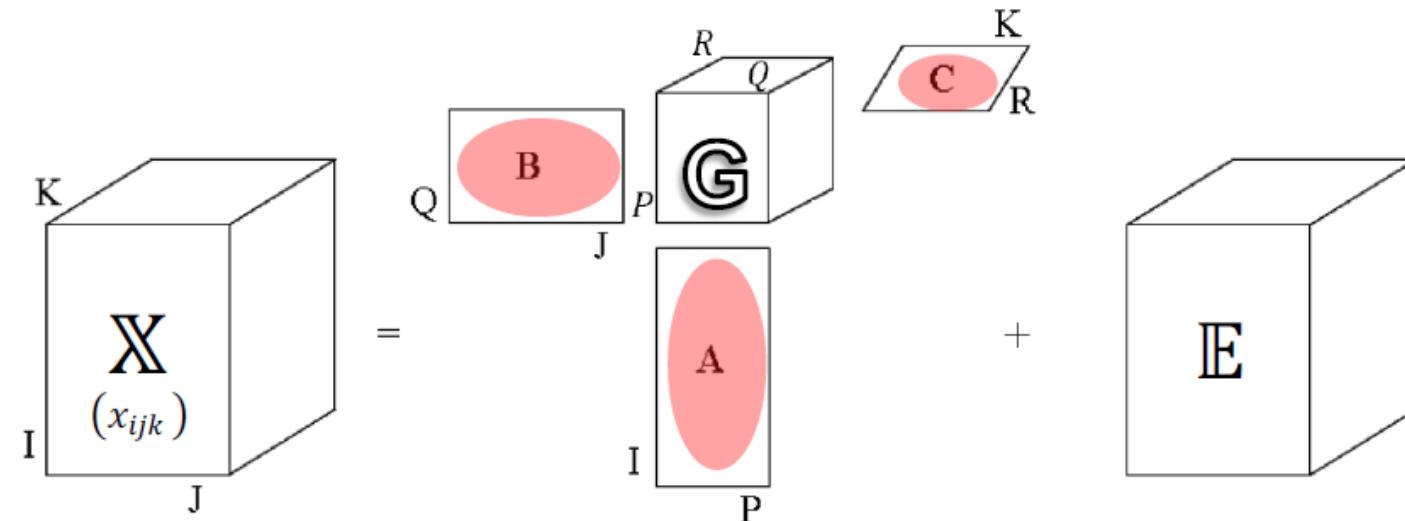
INTRAESTRUCTURA



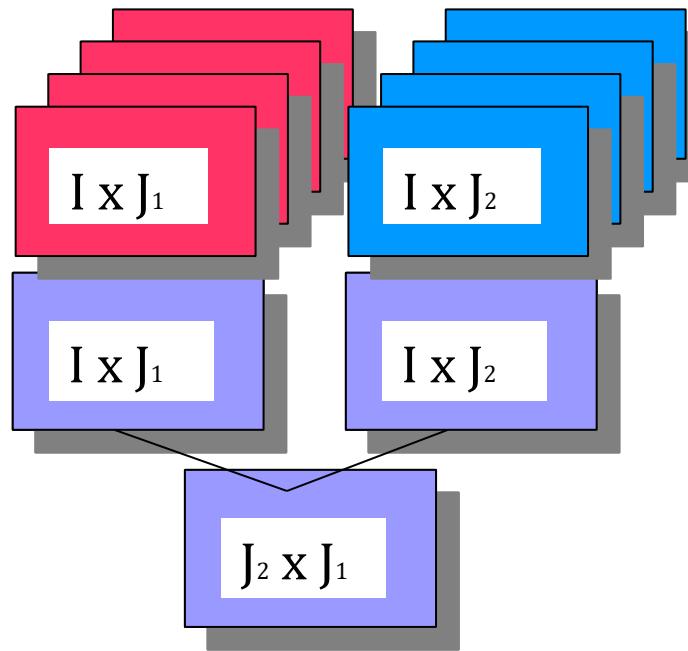
Matrices de distancias



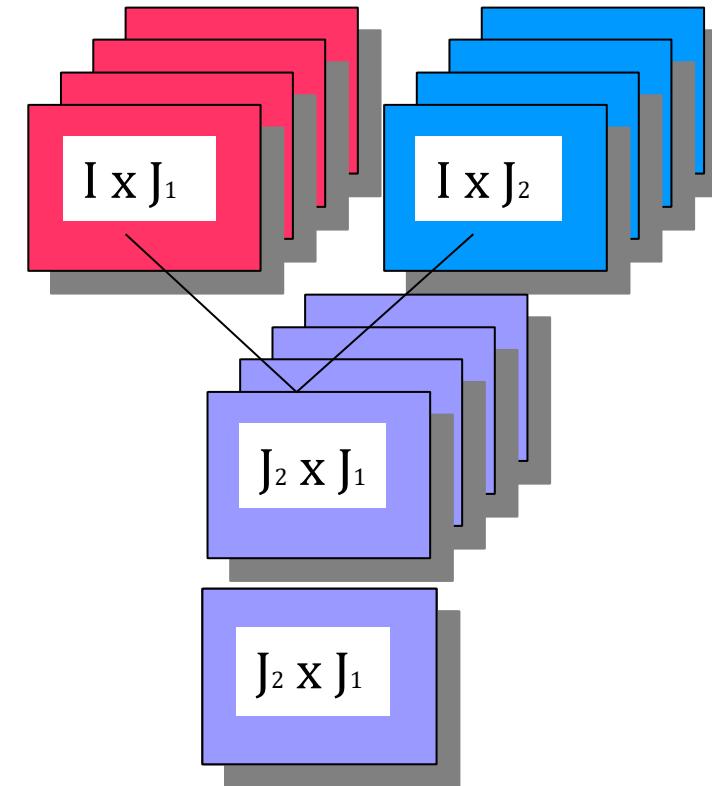
Tucker



Los métodos STATIS permiten capturar la parte estable, mientras que los métodos Tucker permiten encontrar interacciones



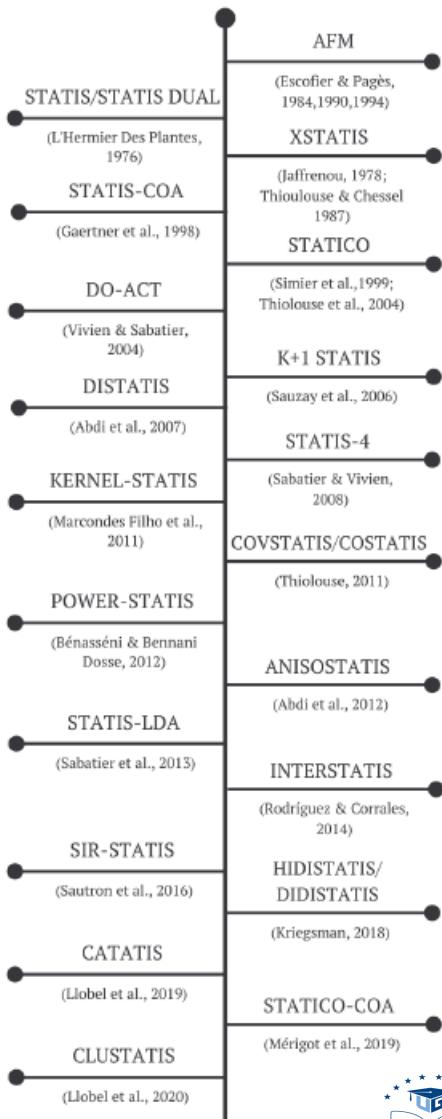
COSTATIS



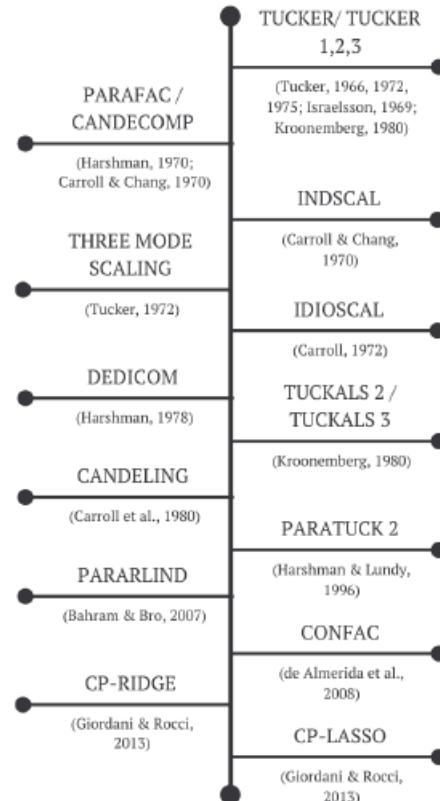
STATICO

MÉTODOS PARA DATOS CON 3 ENTRADAS

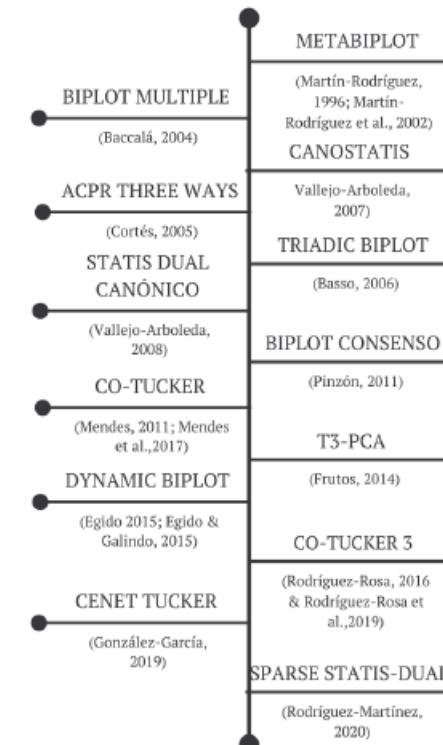
FRANCESAS



ANGLOSAJONA



SALMANTINA



1. MultBIPLOT.

2. Entorno [ade4](#) contiene otros paquetes que extienden las funcionalidades

- [adegraphics](#): Mejora la representación gráfica.
- [ade4TkGUI](#): Interfaz gráfica para el usuario.
- [adespatial](#): Análisis Multivariante Espacial.
- [adegenet](#): Análisis Exploratorio Multivariante de datos genéticos.
- [adehabitat](#), [adiv](#), [adephylo](#)

3. Entorno FactoMineR

- [factoextra](#)
- [Factoshiny](#)
- [FactoInvestigate](#)
- [SensoMineR](#)

4. Escuela Salmantina

- MultBiplotR
- BiplotML
- SparseBiplots
- biplotbootGUI
- :::

¡GRACIAS!