# ¿CÓMO ENFRENTARSE A DATOS COMPLEJOS?

***INTRODUCCIÓN AL ANÁLISIS MULTIVARIANTE: TÉCNICAS BÁSICAS DE ORDENACIÓN Y CLASIFICACIÓN EN R.***


VNiVERSiDAD D SALAMANCA

CAMPUS DE EXCELENCIA INTERNACIONAL

> JULIA VEGA ÁLVAREZ

> JENNIFER MORALES BARBERO

# CONTENIDO

# PRIMEROS PASOS

- CONSIDERACIONES PREVIAS
- ANÁLISIS EXPLORATORIO

# DUDAS EXISTENCIALES

## ¿Qué tengo?

## ¿Qué hago con lo que tengo?

# CONSIDERACIONES PREVIAS

## ¿SON BUENOS MIS DATOS?

- ¿He recogido muestras representativas de la población en estudio?
- ¿Existe algún sesgo en los datos recogidos?

## ¿ESTÁN CORRECTAMENTE EXPRESADOS?

- Revisar las matrices de datos en busca de errores de codificación

## ¿QUÉ TIPO DE VARIABLES TENGO?

- **CUANTITATIVAS (DISCRETA o CONTINUA)**
- **CUALITATIVAS (NOMINAL u ORDINAL)**

## ¿CÓMO PUEDO TRANSFORMAR MIS DATOS PARA AJUSTARLOS A LOS ANÁLISIS QUE VOY A REALIZAR?

- **ELIMINAR DATOS ATÍPICOS o OUTLIERS** que introduzcan ruido en los análisis.
- **DIVIDIR EL ARCHIVO** de los datos en varias partes para facilitar su interpretación.
- **ELIMINAR/AGRUPAR VARIABLES** con el fin de mejorar la interpretabilidad de los datos.
- **TRANSFORMAR** matemáticamente las variables para que se puedan aplicar las técnicas estadísticas elegidas.

# ANÁLISIS EXPLORATORIO DE DATOS

**"Conjunto de técnicas estadísticas cuya finalidad es conseguir un entendimiento básico de los datos y de las relaciones existentes entre las variables analizadas".**

## ➢ ¿ME FALTA INFORMACIÓN?

- Tratamiento y evaluación de **DATOS AUSENTES (MISSING)**

## ➢ ¿HAY CASOS RAROS?

- Identificación de **CASOS ATÍPICOS (OUTLIERS)**

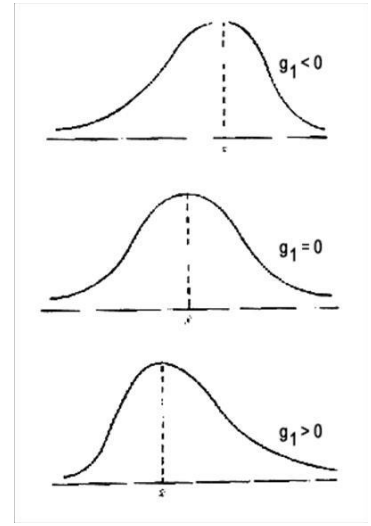
## ➢ ¿CUMPLEN MIS DATOS CON LOS SUPUESTOS TEÓRICOS NECESARIOS PARA APLICAR LAS TÉCNICAS ESTADÍSTICAS QUE QUIERO UTILIZAR?

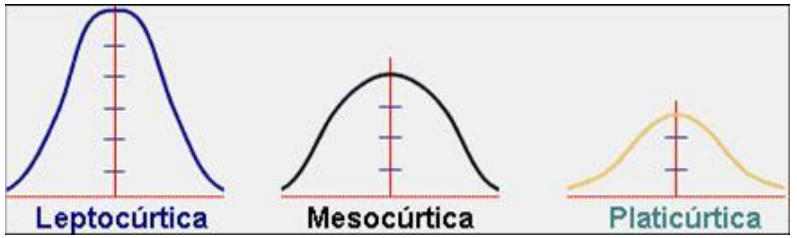
- Comprobación de los **SUPUESTOS de las técnicas MULTIVARIANTES** (normalidad, linealidad, homocedasticidad)

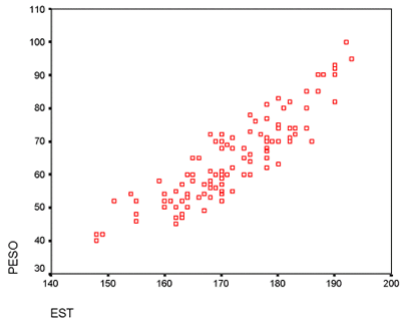
# ESTADÍSTICA DESCRIPTIVA BÁSICA

## ¿QUÉ ESTRUCTURA Y CARACTERÍSTICAS BÁSICAS TIENEN MIS DATOS?

- **MEDIDAS DE POSICIÓN:** CUARTILES, DECILES Y PERCENTILES
- **MEDIDAS DE TENDENCIA CENTRAL:** MEDIA, MEDIANA Y MODA.
- **COEFICIENTES DE ASIMETRÍA (SKEWNESS)**
- **GRADO DE APUNTAMIENTO O CURTOSIS**
- **CORRELACIÓN ENTRE VARIABLES**




| Tipo de variable | Representación gráfica | Medida de tendencia central | Medida de dispersión |
|---|---|---|---|
| NOMINAL | Diagrama de barras, líneas, sectores | Moda | |
| ORDINAL | Boxplot | Mediana | Rango intercuartílico |
| NUMÉRICA | Histograma<br>Polígono de frecuencias | Media | Desviación típica |

# TÉCNICAS DE ORDENACIÓN

- ANÁLISIS INDIRECTO DE GRADIENTE
- ANÁLISIS DIRECTO DE GRADIENTE

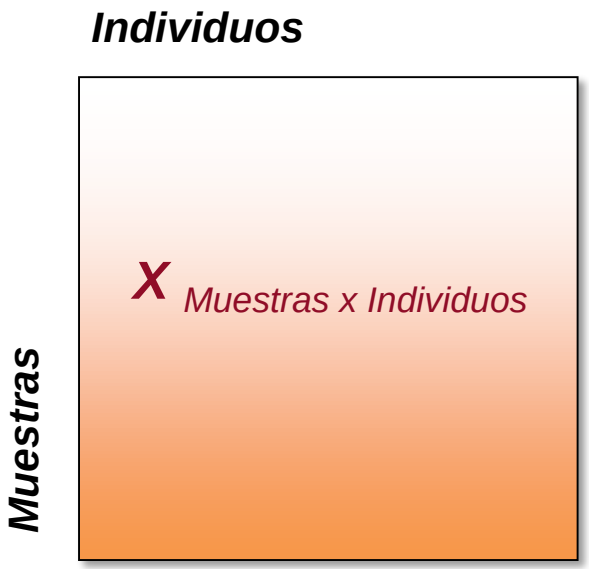
# ANÁLISIS DE GRADIENTE

**"Conjunto de técnicas que permiten estudiar las relaciones existentes entre la composición de las comunidades naturales y las características ambientales de las mismas"**

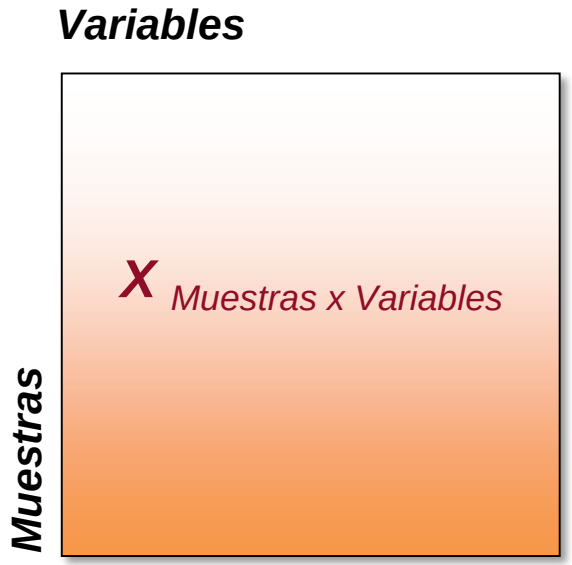
- **PREMISAS**
  - Se **asume una estructura latente** en los datos de composición, que viene **determinada por unas variables ambientales conocidas o desconocidas**.

- **OBJETIVOS**
  - **REDUCIR LA COMPLEJIDAD DE LOS DATOS**, mediante su representación en un diagrama de ordenación de pocas dimensiones.
  - **CONSERVAR la máxima cantidad de INFORMACIÓN posible**, mediante la generación de unos factores latentes que se ordenan según la varianza o inercia explicada.
  - **DESCRIBIR GRÁFICAMENTE** los patrones de composición y las relaciones entre individuos y variables.
  - **FACILITAR LA INTERPRETACIÓN** de la distribución de los individuos en relación a **GRADIENTES** determinados por variables ambientales conocidas (directos) o desconocidas (indirectos).

# ANÁLISIS DE GRADIENTE




**Individuos**
- Especies
- Pacientes
- Consumidores
- Trabajadores
- Empresas
- Industrias
- Genes
- Productos
- Grupos de población
- Elementos químicos




**Variables cuantitativas continuas**
- Altura
- Peso
- Edad
- Diversidad
- Indicadores económicos
- Factores ambientales
- Distancia geográfica
- Tiempo

**Muestras**
- Unidades de muestreo
- Puntos geográficos
- Entidades privadas
- Comunidades ecológicas
- Países
- Sectores económicos

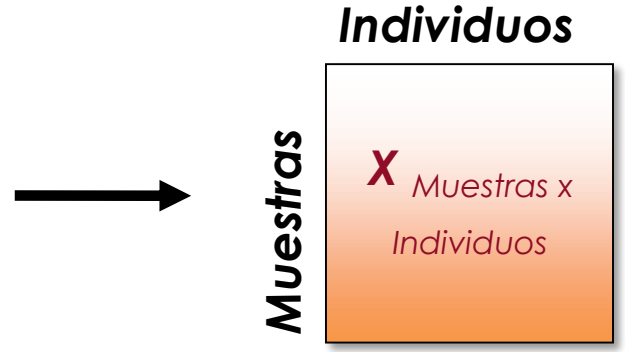
**Medidas**
- Abundancia
- Nº Individuos
- Densidad
- Cobertura
- Producción
- Rendimiento
- Pres/Aus

- ❑ **MUESTRAS = VARIABLES CATEGÓRICAS**
- ❑ **INDIVIDUOS = VARIABLES CUANTITATIVAS**

# ANÁLISIS DE GRADIENTE

➢ **ANÁLISIS INDIRECTO DE GRADIENTE**

(Sin información ambiental)




➢ **ANÁLISIS DIRECTO DE GRADIENTE**

(Ejes constreñidos por información ambiental)

# MÉTODOS INDIRECTOS de GRADIENTE

- **DETECCIÓN INDIRECTA DE GRADIENTES** DE VARIACIÓN DE COMPOSICIÓN ENTRE MUESTRAS.
- CREAR UN CONJUNTO DE **DIMENSIONES LATENTES, NO CONSTREÑIDAS POR FACTORES EXTERNOS,** QUE MEJOR RESUMAN LA VARIABILIDAD DEL CONJUNTO DE DATOS
- **LOS EJES DE ORDENACIÓN INFIEREN GRADIENTES**


| ANÁLISIS INDIRECTO DE GRADIENTE | | RELACIÓN LINEAL | RELACIÓN NO LINEAL |
|---|---|---|---|
| MEDIDA DE DISTANCIA | **EUCLÍDEA** | PCA | |
| | $\chi^2$ | | CA |
| | **CUALQUIERA** | PCoA/ MDS | NMDS |

**¿QUÉ TIPO DE DISTANCIA QUIERO PRESERVAR EN EL PLANO DE ORDENACIÓN?**

**¿QUÉ TIPO DE RELACIÓN EXISTE ENTRE INDIVIDUOS Y MUESTRAS?**

# PROCESO DE LOS MÉTODOS DE ORDENACIÓN INDIRECTOS

**MATRIZ ORIGINAL DE DATOS**

- **COMPOSICIÓN** (PCA, CA)
- **DISTANCIA** (MDS, NMDS)

**ESCALADO DE DATOS**

- **CENTRADO**
- **ESTANDARIZADO**

**MATRIZ DE PARTIDA**

- **CORRELACIÓN/ COVARIANZAS** (PCA, CA)
- **DISTANCIAS** (MDS, NMDS)

**SELECCIÓN DE EJES**

- **MÁXIMA VARIABILIDAD: EIGENVALUES** (PCA, CA, MDS)
- **DISTANCIAS ORIGINALES** (NMDS)

**DISTANCIAS**

*Muestras*

| | | | | |
|---|---|---|---|---|
| 0 | | | | |
| 0,4 | 0 | | | |
| 0,2 | 0,1 | 0 | | |
| 0,3 | 0,8 | 0,1 | 0 | |
| 0,6 | 0,6 | 0,7 | 0,5 | 0 |

*Muestras*

MDS, NMDS

**CORR/COV**

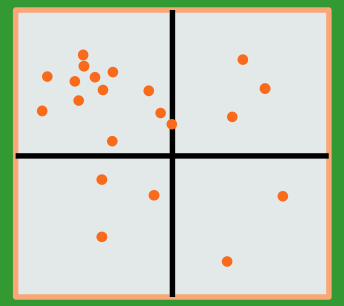
*Individuos*

| | | | | |
|---|---|---|---|---|
| 1 | | | | |
| 0,4 | 1 | | | |
| 0,2 | 0,1 | 1 | | |
| 0,3 | 0,8 | 0,1 | 1 | |
| 0,6 | 0,6 | 0,7 | 0,5 | 1 |

*Individuos*

PCA, CA,

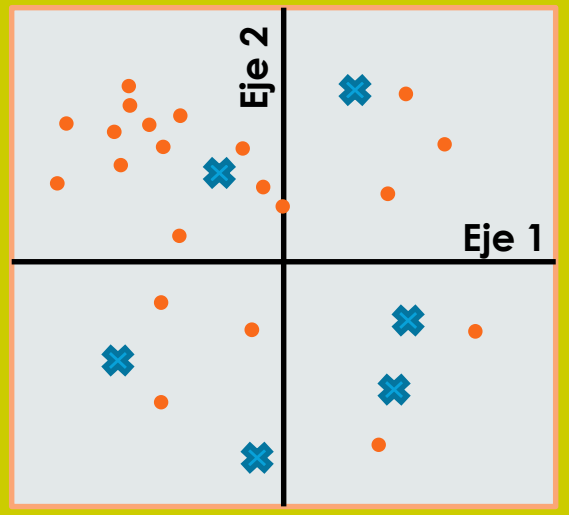
**ORDENACIÓN INICIAL**



**ROTACIÓN DE EJES**

- **VARIMAX**
- **QUARTIMAX**
- **ROTACIÓN DE DIMENSIONES CON VARIABLE EXTERNA**

**ORDENACIÓN FINAL**



# 1) ANÁLISIS DE COMPONENTES PRINCIPALES (PCA)

## A. PREMISAS

- **VARIABLES CUANTITATIVAS**, **CORRELACIONADAS** ENTRE SÍ
- **SENSIBLE AL TIPO DE ESCALADO** DE LAS VARIABLES ORIGINALES, 0.
- **SUPUESTOS DE NORMALIDAD**, LINEALIDAD, CORRELACIÓN Y NO MULTICOLINEALIDAD
- **DISTRIBUCIÓN DE LAS VARIABLES Y RELACIÓN ENTRE VARIABLES LINEAL**

## B. FUNDAMENTO

- **EJES= COMPONENTES PRINCIPALES = COMBINACIÓN LINEAL** DE LAS VARIABLES ORIGINALES.
- Preserva **DISTANCIA EUCLÍDEA** entre variables
- **Busca el espacio que recoge la MÁXIMA VARIABILIDAD ORIGINAL**

## C. OBJETIVO

- **Resumir el CONJUNTO DE DATOS y estudiar las RELACIONES ENTRE VARIABLES.**

## PCA en R

**library(vegan)**

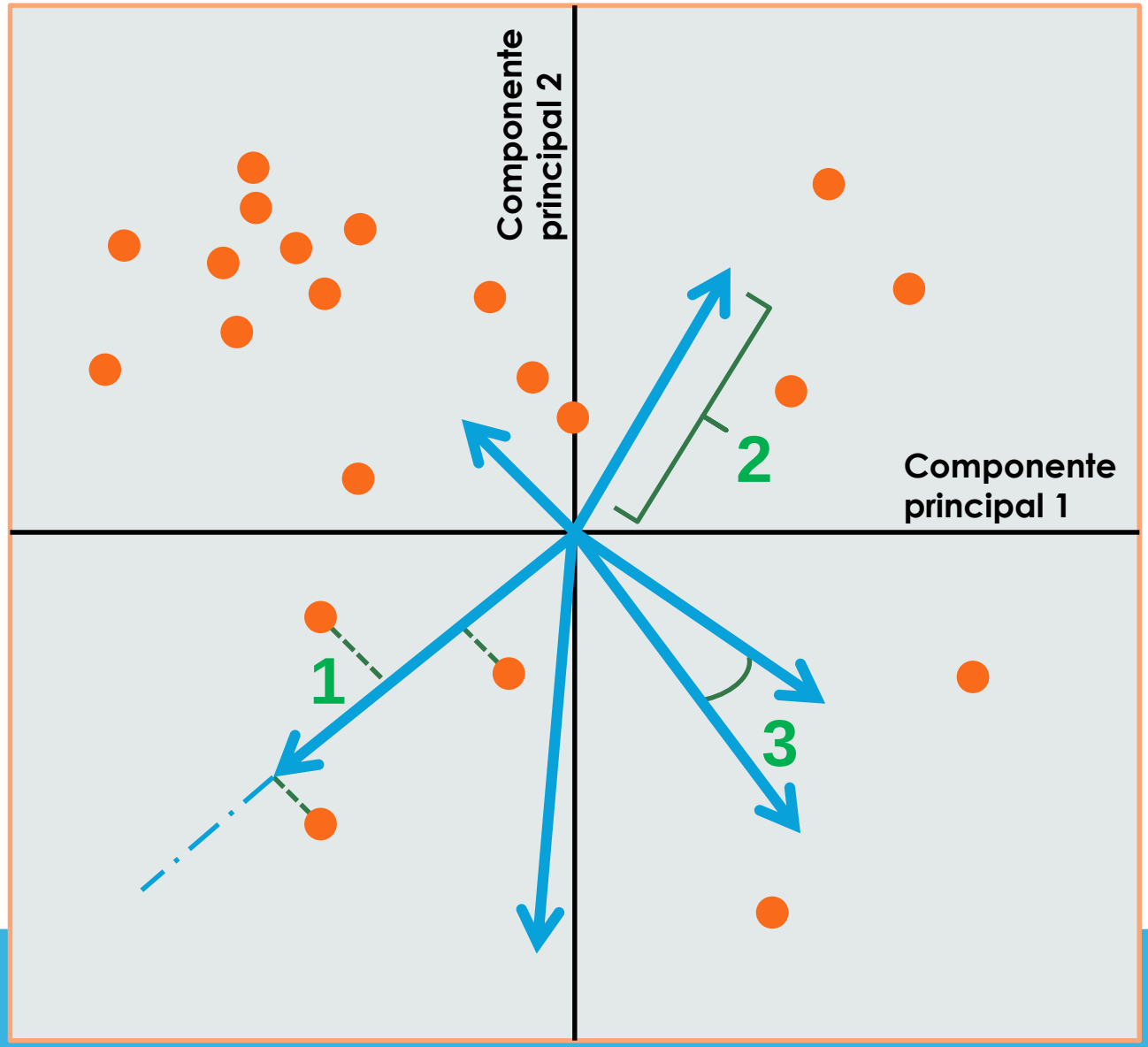
**data(dune)**

```
pca.dune <- rda(dune, scale = F)
summary(pca.dune)
screeplot(pca.dune)
pca <- biplot(pca.dune, choices = c(1,2), scaling = 3, type = c("text", "points"))
goodness(pca.dune, choices = c(1,2), statistic = c("explained"), summarize = TRUE)
```

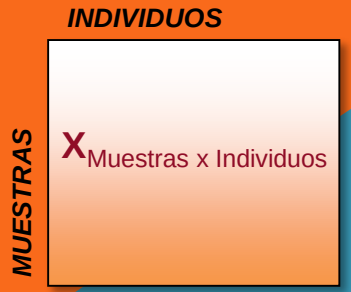
# INTERPRETACIÓN GRÁFICA

1. **ABUNDANCIA DE INDIVIDUOS EN LAS MUESTRAS**
   - *Proyecciones ortogonales de los puntos sobre los vectores*
2. **VARIABILIDAD DE LOS INDIVIDUOS**
   - Longitud de los vectores
3. **CORRELACIÓN/COVARIANZA ENTRE INDIVIDUOS**
   - Ángulos entre vectores

- **ANÁLISIS DE ITEMS EN VALIDACIÓN DE TESTS**




❖ **PUNTUACIONES FACTORIALES (scores)**
Coordenadas en el plano de ordenación
❖ **CARGAS FACTORIALES (loadings)**
Contribución relativa de las variables en los componentes




**Matriz *PCA***

**Supone relaciones LINEALES**
**Preserva distancia EUCLÍDEA**

# 2) ANÁLISIS DE COORDENADAS PRINCIPALES (PcoA) = ESCALAMIENTO MULTIDIMENSIONAL MÉTRICO (MDS)

## A. PREMISAS

- VARIABLES **CUANTITATIVAS/CUALITATIVAS/MIXTAS**
- SENSIBLE AL **TIPO DE MEDIDA** USADO. Si Euclídea MDS=PCA
- RELACIÓN INDIVIDUOS/MUESTRAS **LINEAL**

## B. FUNDAMENTO

- **EJES= COORDENADAS PRINCIPALES = Distancia euclídea entre coordenadas ≈ distancias originales.** Representación euclídea de un conjunto de objetos cuyas relaciones son medidas por cualquier medida de distancia elegida por el usuario.
- **Preserva *CUALQUIER DISTANCIA*** entre muestras: Bray-Curtis, Jaccard, Manhattan...
- **Busca el espacio que REPRODUCE APROXIMADAMENTE LAS DISTANCIAS ORIGINALES y recoge la MÁXIMA VARIABILIDAD de la matriz de distancias**

## C. OBJETIVO

- **RESUMIR LAS DISIMILARIDADES ENTRE MUESTRAS y ENCONTRAR GRUPOS HOMOGÉNEOS DE MUESTRAS.**

## MDS en R

**library(vegan)**
**data (varespec)**

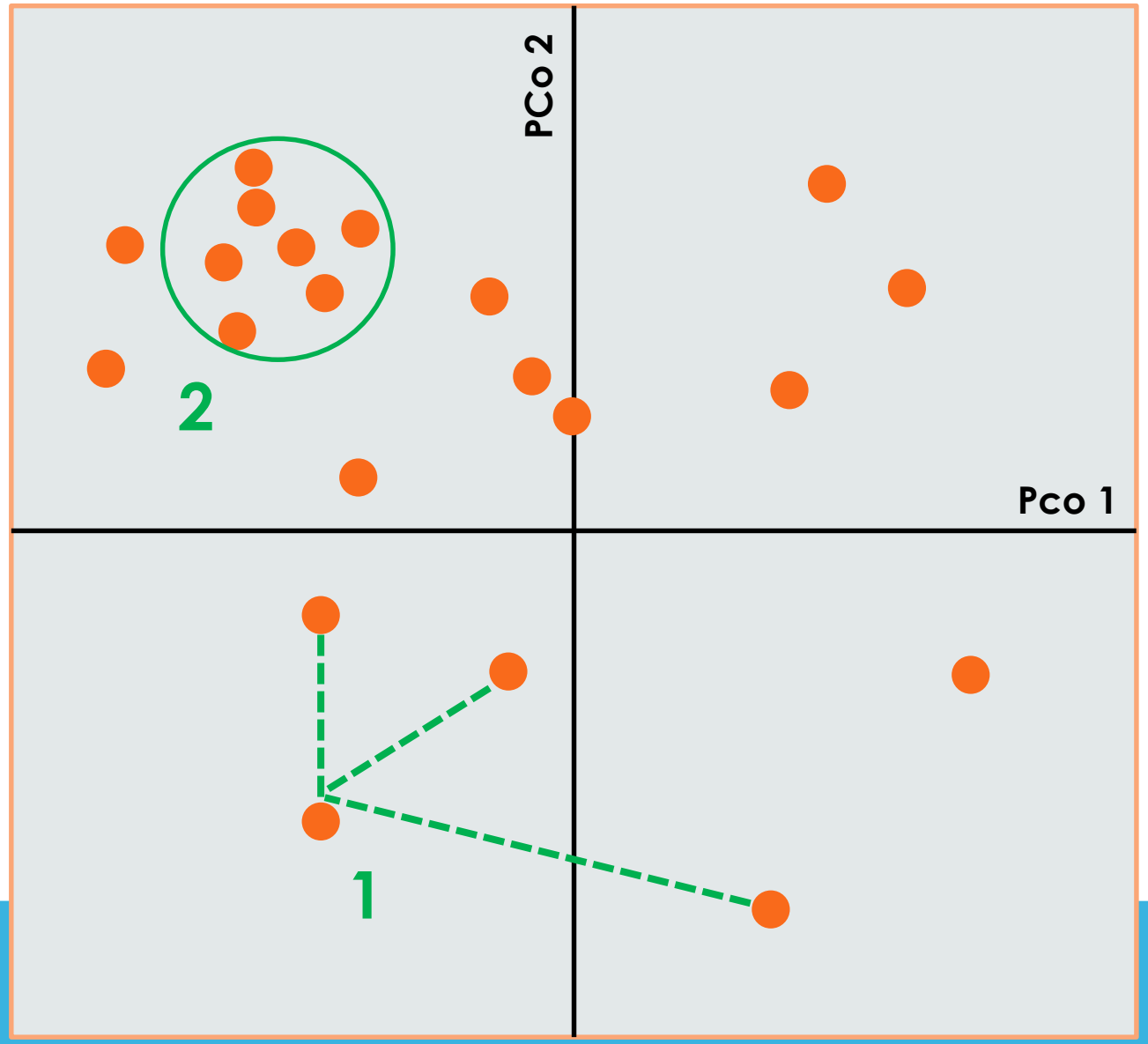
```
> braydist <- vegdist(varespec, method = "bray")
> mds.vares <- cmdscale(braydist, k = 2, eig = F, wascores = T)
> plot.mds <- ordiplot(mds.vares, type = "t")
> abline(h=0, v=0)
```

# INTERPRETACIÓN GRÁFICA

1. ***SIMILARIDAD en términos de medida de distancia ENTRE MUESTRAS***
   - *Distancia entre puntos.*

2. **DETECCIÓN DE GRUPOS HOMOGÉNEOS DE MUESTRAS**
   - Proximidad entre grupos de puntos

- **ESTUDIOS DE ESTRUCTURA GENÉTICA**
- **MAPAS PERCEPTUALES**




❖ **Sites scores = POSICIÓN de las muestras en el plano de ordenación**

***INDIVIDUOS***

***MUESTRAS***

$X_{\text{Muestras x Individuos}}$

**Matriz *PCoA/MDS***

**Supone relaciones LINEALES**
**Preserva CUALQUIER distancia**

# 3) ESCALAMIENTO MULTIDIMENSIONAL NO MÉTRICO (NMDS)

## A. PREMISAS

- VARIABLES **CUANTITATIVAS/CUALITATIVAS/MIXTAS**
- SENSIBLE AL **TIPO DE MEDIDA** USADO
- RELACIÓN INDIVIDUOS/MUESTRAS **NO LINEAL**

## B. FUNDAMENTO

- **EJES = DIMENSIONES** = Recoge **SIMILARIDAD ENTRE PARES DE MUESTRAS**
- ***Preserva CUALQUIER DISTANCIA entre muestras:*** : Bray-Curtis, Jaccard, Manhattan...
- **Busca el espacio que MEJOR REPRESENTE LAS DISTANCIAS ORIGINALES en términos de rangos de orden (Algoritmo iterativo < STRESS)**

## C. OBJETIVO

- **RESUMIR LAS DISIMILARIDADES ENTRE MUESTRAS y ENCONTRAR GRUPOS HOMOGÉNEOS DE MUESTRAS.**

## NMDS en R

**library(vegan)**
**library(MASS)**
**data (varespec)**

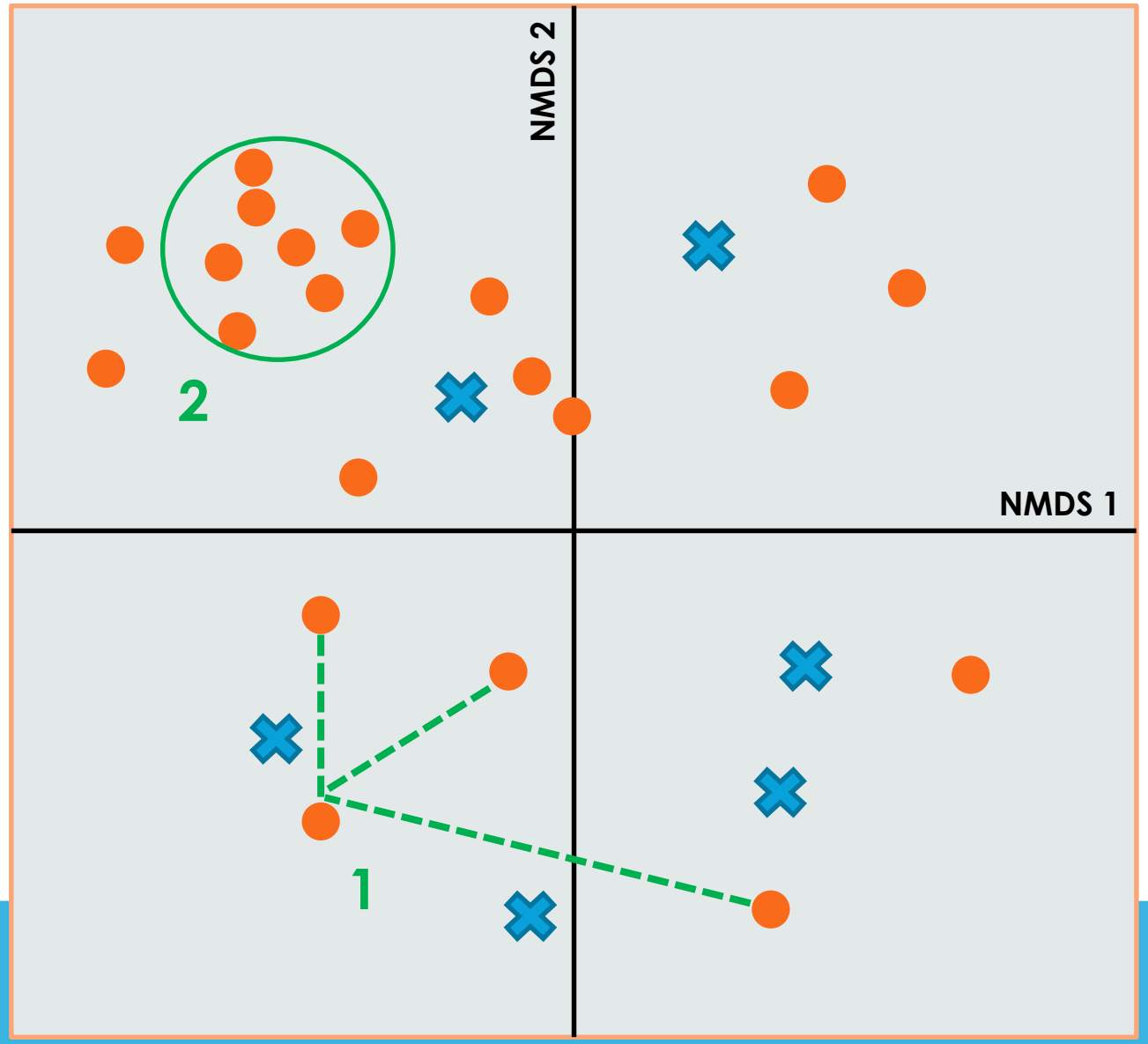
```
> braydist <- vegdist(varespec)
> nmds.vares <- metaMDS(varespec, distance ="bray", wascores = T, autotransform = F)
> plot(nmds.vares, display = "sites", type = 't', xlim = c(-1, 1.5), ylim = c(-1, 1))
> stressplot(nmds.vares)
> goodness(nmds.vares, display = c("sites"), choices = c(1,2), statistic = c("distance"))
```

# INTERPRETACIÓN GRÁFICA

1. ***PROBABILIDAD DE SIMILARIDAD ENTRE MUESTRAS en términos de rangos de orden***
   - *Distancia entre puntos.*

2. **DETECCIÓN DE GRUPOS HOMOGÉNEOS DE MUESTRAS**
   - Proximidad entre grupos definidos de puntos

**Se superponen las coordenadas para especies como medias ponderadas**




**MAPAS PERCEPTUALES:** Estudiar las preferencias o percepciones de Personas/Clientes/Consumidores sobre Políticas/Productos/Servicios...

- **MARKETING**
- **CIENCIAS SOCIALES**

***VARIABLES***

***MUESTRAS***

$X_{\text{Muestras x Variables}}$

**Matriz *NMDS***

**Supone relaciones NO LINEALES CUALQUIER DISTANCIA**

❖ **MEDIDA DE BONDAD DE AJUSTE = STRESS < 0.1**

# 4) ANÁLISIS (FACTORIAL) DE CORRESPONDENCIAS (CA)

## A. PREMISAS

- **VARIABLES NOMINALES, datos de frecuencia, presencia/ausencia...**
- **SENSIBLE A DATOS RELATIVIZADOS**, OUTLIERS,
- DISTRIBUCIÓN DE LAS VARIABLES **UNIMODAL**


**Tabla de contingencia**

## B. FUNDAMENTO

- EJES = FACTORES = **MÁXIMA ABSORCIÓN INERCIA**
- **TRABAJA CON PERFILES**: coordenadas ponderadas por frecuencia relativa
- Preserva **DISTANCIA $\chi^2$** *entre perfiles fila o columna.*
- Busca el subespacio que **MEJOR REPRESENTE LA CORRESPONDENCIA ENTRE MUESTRAS e INDIVIDUOS** (< $\chi^2$ entre perfiles)

## C. OBJETIVO

- **REPRESENTAR LOS PERFILES FILA O COLUMNA con precisión y estudiar las RELACIONES ENTRE ELLOS.**

## CA en R

**library(vegan)**
**data (varespec)**

```
> ca.vares <- cca(varespec)
> plot(ca.vares, scaling = 3)
> screeplot(ca.vares)
> summary(ca.vares)
> goodness(ca.vares, display = c("sites"), choices = c(1,2), statistic = c("explained"),
summarize = TRUE)
```

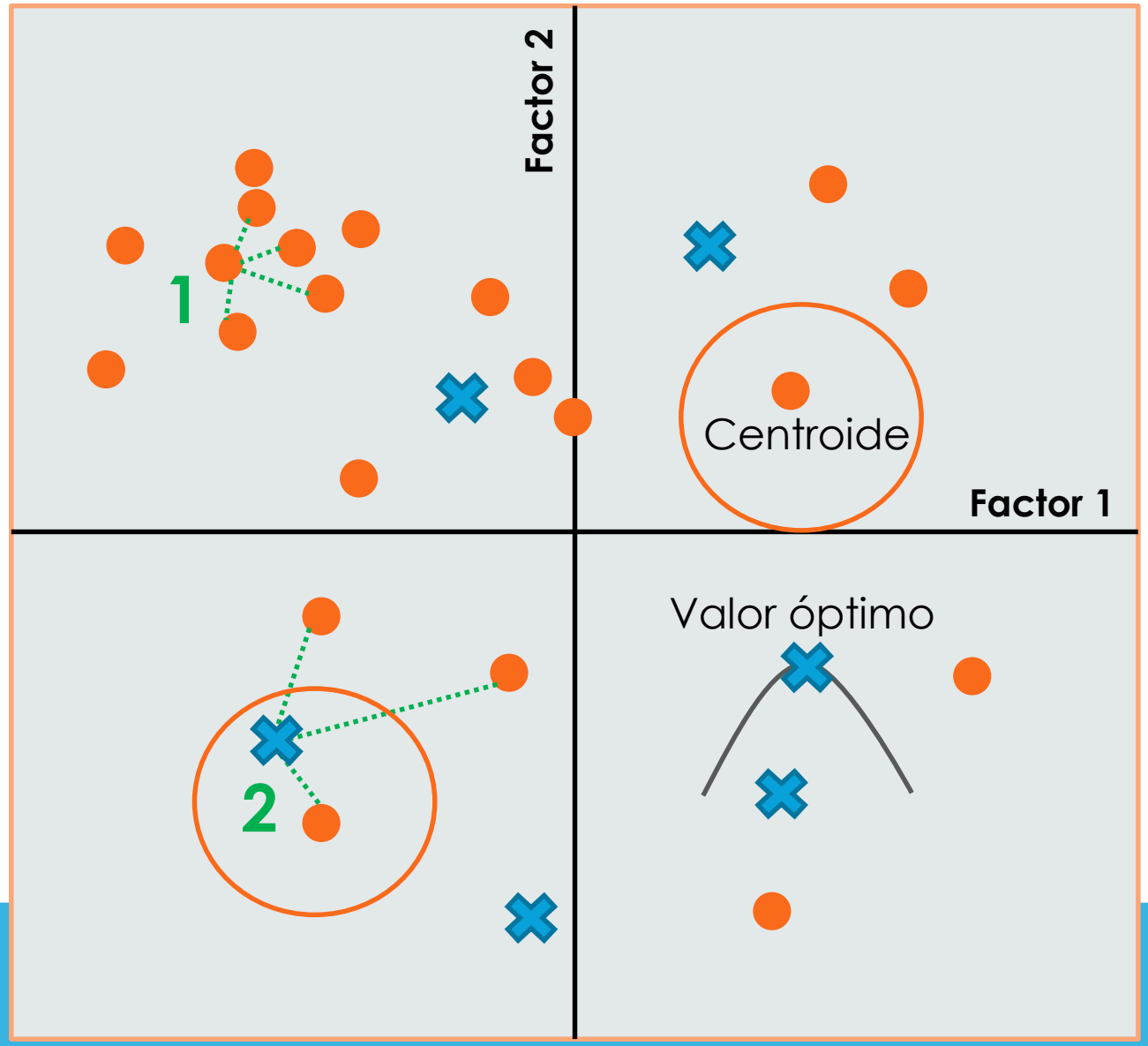
# INTERPRETACIÓN GRÁFICA BIPLOT

1. **SIMILARIDAD DE COMPOSICIÓN ENTRE MUESTRAS**
   - Distancia entre puntos
2. **PROBABILIDAD DE QUE UN INDIVIDUO SEA FRECUENTE EN UNA MUESTRA/PROBABILIDAD DE ASOCIACIÓN ENTRE CATEGORÍAS**
   - Distancia entre puntos individuos y puntos muestras.

❑ Peor representación hacia el origen ≈ perfil promedio

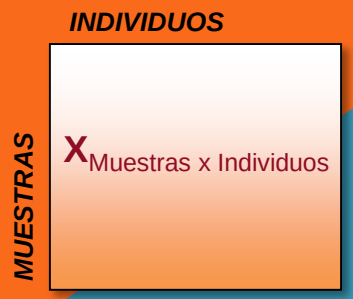
❑ No representa las distancias originales

- **ESTUDIOS ECOLÓGICOS DE GRADIENTE LARGO**
- **ESTUDIOS PSICOLÓGICOS**




❖ **PUNTOS= PERFILES = CENTROIDES**

❖ **INERCIA= VARIABILIDAD EXPLICADA**




**Matriz CA**

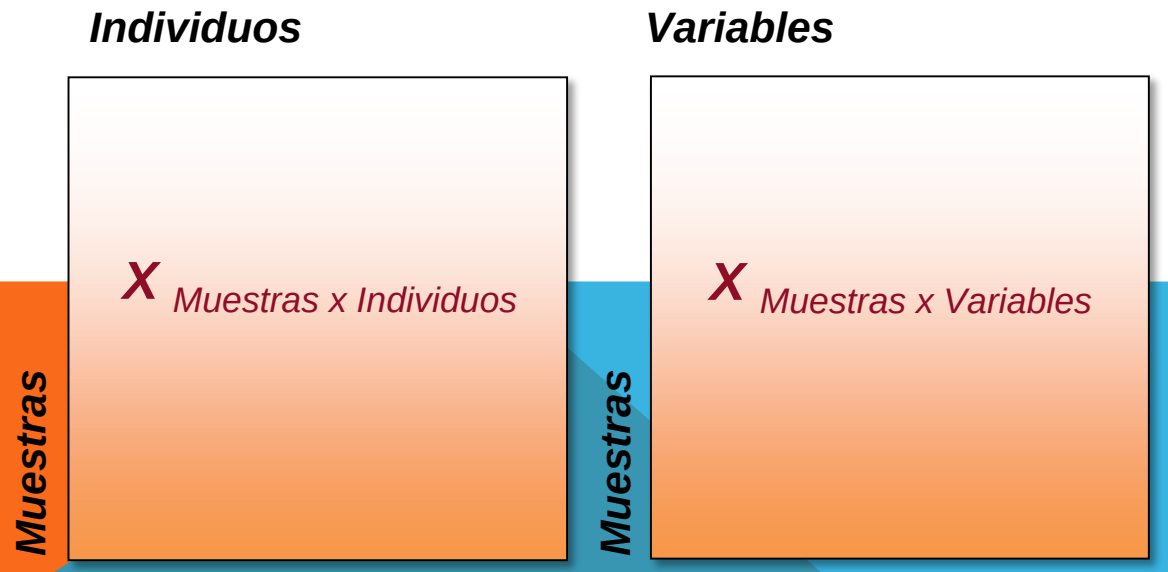
**Supone relaciones UNIMODALES**

**Distancia $\chi^2$**

# ANÁLISIS DIRECTO de GRADIENTE

- **DETECCIÓN DIRECTA DE GRADIENTES** DE VARIACIÓN DE COMPOSICIÓN ENTRE MUESTRAS **EN FUNCIÓN DE UNAS VARIABLES DETERMINADAS A PRIORI**
- CREAR UN CONJUNTO DE **DIMENSIONES LATENTES** DENOMINADAS **FACTORES CONSTREÑIDOS POR VARIABLES EXTERNAS**
- **LOS EJES DE ORDENACIÓN REPRESENTAN GRADIENTES**

| ANÁLISIS DIRECTO DE GRADIENTE | | RELACIÓN LINEAL | RELACIÓN NO LINEAL |
|---|---|---|---|
| MEDIDA DE DISTANCIA | **EUCLÍDEA** | RDA | |
| | $\chi^2$ | | CCA |




**¿QUÉ TIPO DE DISTANCIA QUIERO PRESERVAR EN EL PLANO DE ORDENACIÓN?**

**¿QUÉ TIPO DE RELACIÓN EXISTE ENTRE INDIVIDUOS Y MUESTRAS?**

# PROCESO DE LOS MÉTODOS DE ORDENACIÓN DIRECTOS

**MATRIZ ORIGINAL DE DATOS**

- **COMPOSICIÓN/ VARIABLES DEPENDIENTES** (RDA, CCA)

**ESCALADO DE DATOS**

- **CENTRADO**
- **ESTANDARIZADO**

**MATRIZ DE PARTIDA**

- **CORRELACIÓN/ COVARIANZAS** (RDA, CCA)

**SELECCIÓN DE EJES**

- **COMBINACIÓN LINEAL DE VAR. AMBIENTALES** (RDA, CCA)

**CORR/COV**

*Individuos*

| | | | | |
|---|---|---|---|---|
| 1 | | | | |
| 0,4 | 1 | | | |
| 0,2 | 0,1 | 1 | | |
| 0,3 | 0,8 | 0,1 | 1 | |
| 0,6 | 0,6 | 0,7 | 0,5 | 1 |

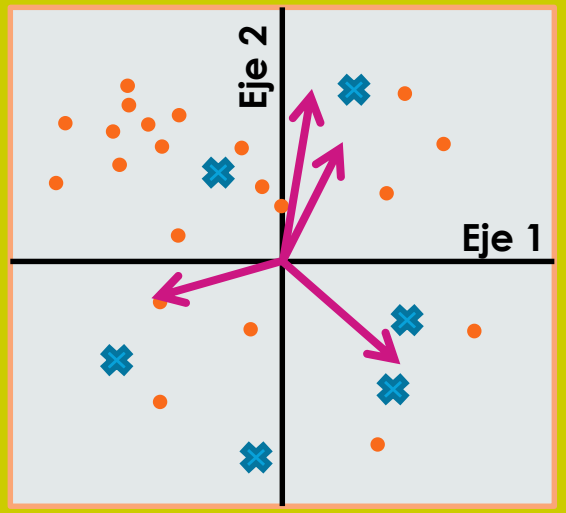
*Individuos*

RDA, CCA

**VARIABLES DEPENDIENTES**

- **Distribución lineal:** RDA
- **Distribución unimodal:** CCA

**ORDENACIÓN FINAL**



# 1) ANÁLISIS DE REDUNDANCIA (RDA)

## A. PREMISAS

- **VARIABLES RESPUESTA (y): CUANTITATIVAS, distribución LINEAL**
- **VARIABLES EXPLICATIVAS (x): CUANTITATIVAS, relación LINEAL con** VARIABLES RESPUESTA.
- **VARIABLES EXPLICATIVAS < Nº DE MUESTRAS**

## B. FUNDAMENTO

- **EJES CANÓNICOS = COMBINACIÓN LINEAL** DE LAS **VARIABLES AMBIENTALES.**
- **Busca el subespacio RESTRINGIDO POR LAS VARIABLES AMBIENTALES que recoge la MÁXIMA VARIABILIDAD ORIGINAL**
- **VARIANZA TOTAL= VARIANZA CONSTREÑIDA + VARIANZA NO CONSTREÑIDA. (V.C > V.N.C)**

## C. OBJETIVO

- **RESUMIR LA VARIACIÓN DE COMPOSICIÓN DE LAS MUESTRAS QUE PUEDE SER EXPLICADA POR VARIABLES AMBIENTALES.**

## RDA en R

**library(vegan)**

**data(dune)**

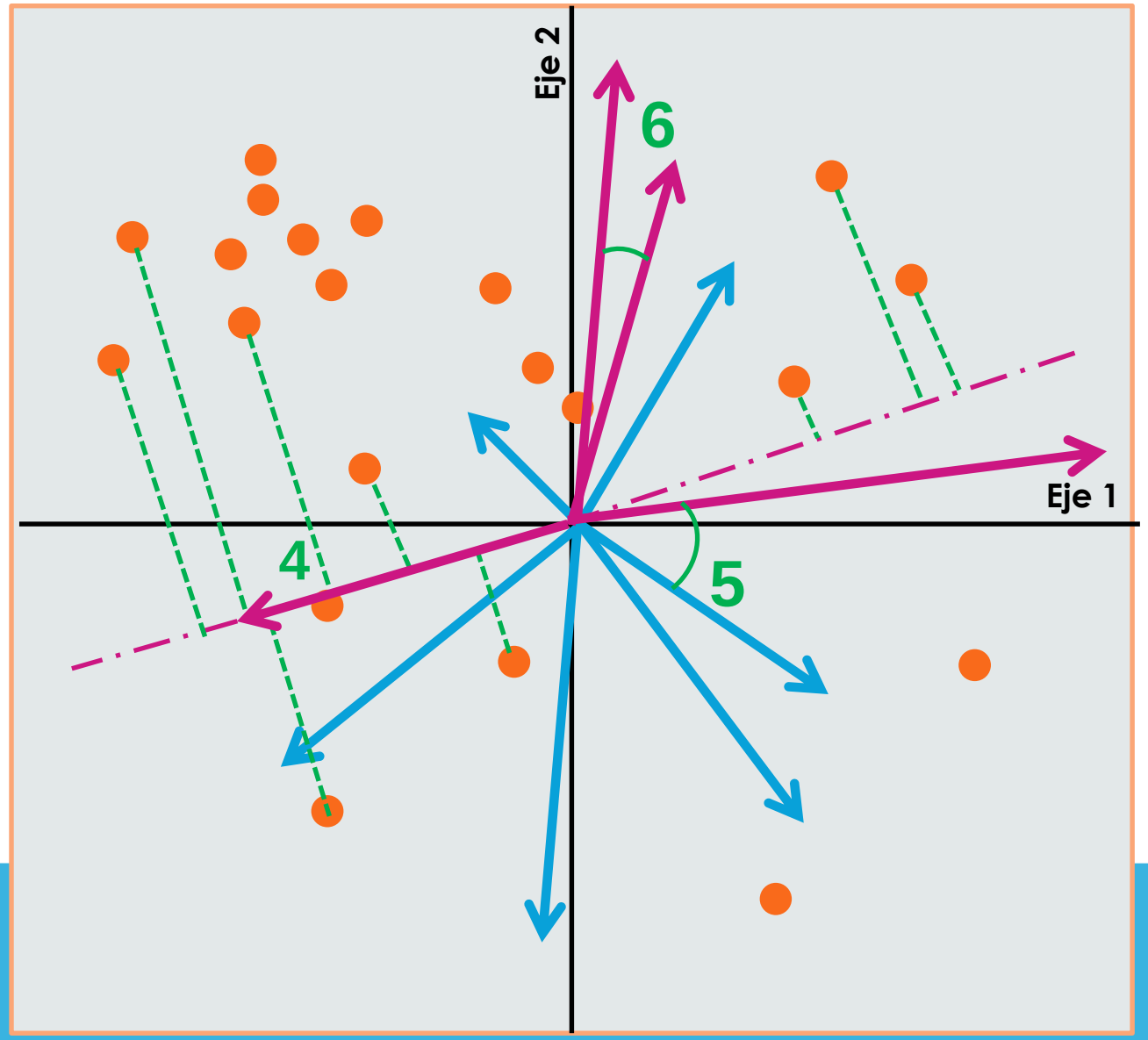
**data(dune.env)**

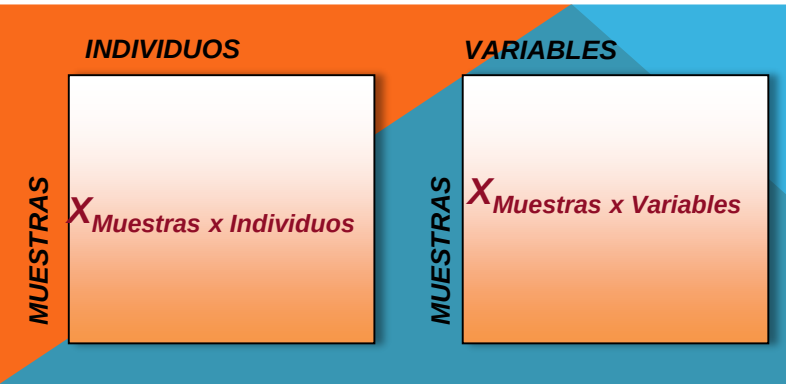
```
> rda.dune <- rda(dune ~ Manure, dune.env, scale= F)
> plot(rda.dune, scaling = 3)
> screeplot(rda.dune)
> summary(rda.dune)
> goodness(rda.dune, display = c("species"), choices = c(1,2), statistic = c("explained"),
summarize = TRUE)
> anova(rda.dune)
```

# INTERPRETACIÓN GRÁFICA TRIPLOT

4. **VALOR DE LAS VARIABLES EN LAS MUESTRAS**
   - Proyecciones ortogonales de los puntos sobre los vectores
5. **CORRELACIÓN/COVARIANZA ENTRE INDIVIDUOS Y VARIABLES AMBIENTALES**
   - Ángulos entre vectores
6. **CORRELACIÓN/COVARIANZA ENTRE VARIABLES AMBIENTALES**

- **ESTUDIOS ECOLÓGICOS DE GRADIENTE CORTO**




**Relacion LINEAL**

**Distancia EUCLÍDEA**

- PUNTOS = MUESTRAS = UNIDADES DE MUESTREO
- VECTORES = INDIVIDUOS = OBJETO DE ESTUDIO
- VECTORES RAYADOS = VARIABLES = VAR. AMBIENTALES

# 2) ANÁLISIS CANÓNICO DE CORRESPONDENCIAS (CCA)

## A. PREMISAS

- **VARIABLES RESPUESTA (y): NOMINALES, distribución UNIMODAL**
- **VARIABLES EXPLICATIVAS (x): CUANTITATIVAS, relación LINEAL con** VARIABLES RESPUESTA.
- **VARIABLES EXPLICATIVAS < Nº DE MUESTRAS**

## B. FUNDAMENTO

- **EJES CANÓNICOS = COMBINACIÓN LINEAL DE LAS VARIABLES AMBIENTALES**
- Busca el **subespacio restringido por las VARIABLES AMBIENTALES** que **MEJOR REPRESENTE LA CORRESPONDENCIA ENTRE MUESTRAS e INDIVIDUOS** (< $\chi^2$ entre perfiles)
- **INERCIA TOTAL= INERCIA CONSTREÑIDA + INERCIA NO CONSTREÑIDA. (I.C > I.N.C)**

## C. OBJETIVO

- **RESUMIR LAS POSICIONES DE LOS PERFILES FILA O COLUMNA con precisión EN RELACIÓN A UN GRADIENTE AMBIENTAL DEFINIDO y estudiar las RELACIONES ENTRE ELLOS.**

## CCA en R

**library(vegan)**
**data (varespec)**
**data(varechem)**

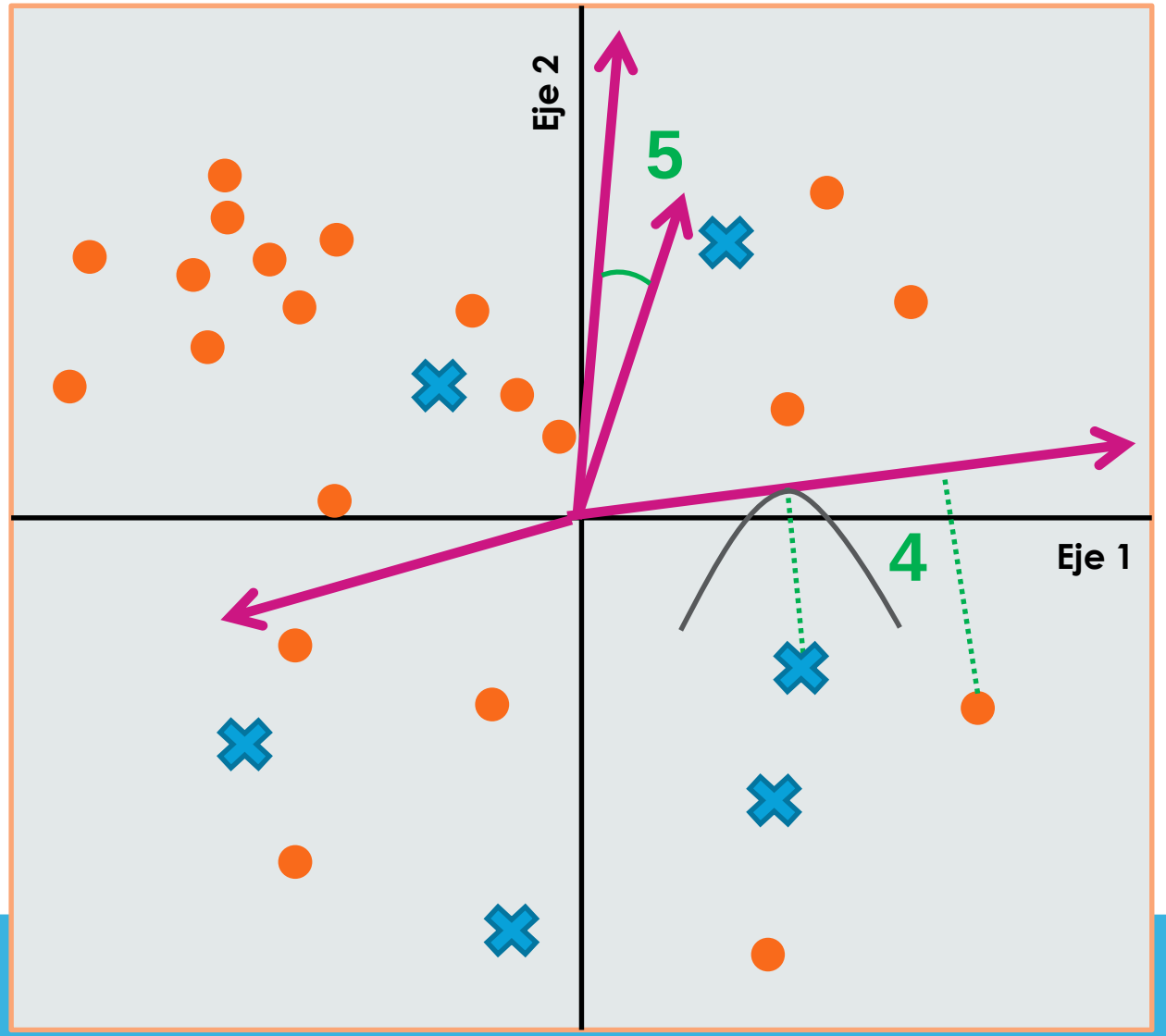
```
> cca.vares <- cca(varespec, varechem)
> plot(cca.vares, scaling = 3)
> screeplot(cca.vares)
> summary(cca.vares)
> goodness(cca.vares, display = c("sites"), choices = c(1,2), statistic = c("explained"), summarize = TRUE)
> anova(cca.vares)
> vif.cca(cca.vares)
```

# INTERPRETACIÓN GRÁFICA TRIPLOT

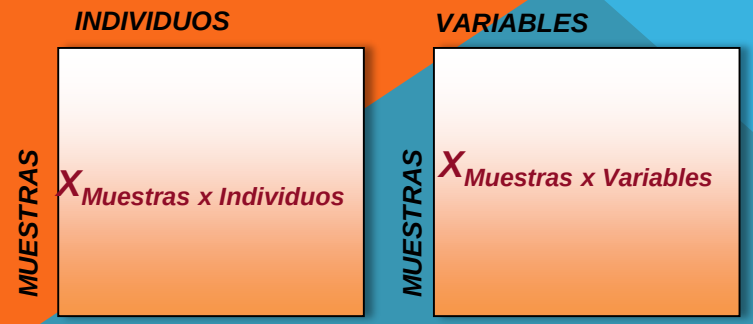
4. **VALOR DE LA VARIABLE EN LA MUESTRA/ ÓPTIMO DEL INDIVIDUO EN LA VARIABLE**
   - Proyecciones ortogonales de los puntos sobre los vectores
5. **CORRELACIÓN/COVARIANZA ENTRE VARIABLES AMBIENTALES**
   - Ángulos entre vectores

**APLICACIONES**

- **ESTUDIOS ECOLÓGICOS DE GRADIENTE LARGO**
- **ANÁLISIS TEXTUAL (TEXT MINING)**




- PUNTOS = PERFILES = CENTROIDES
- VECTORES = VARIABLES AMBIENTALES




**Relación UNIMODAL**

**Distancia $\chi^2$**

MUESTRAS = TIPOS DE RESTAURANTES

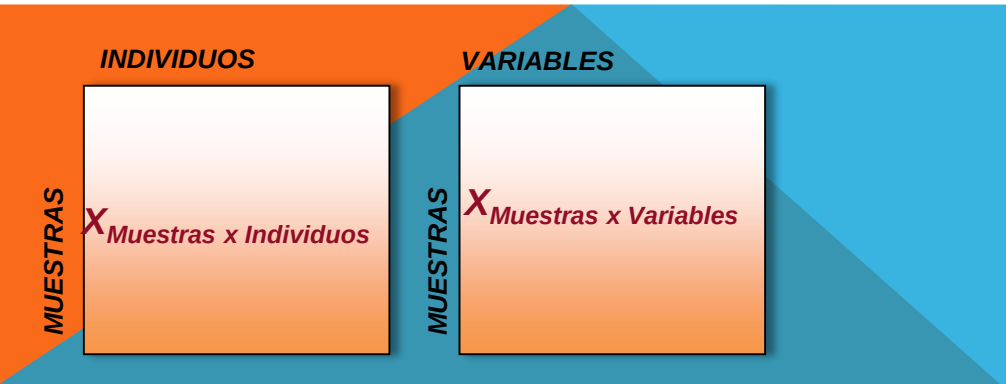
- PIZZERÍAS
- HAMBURGUESERÍAS
- ITALIANO
- COCINA DE AUTOR
- ORIENTAL

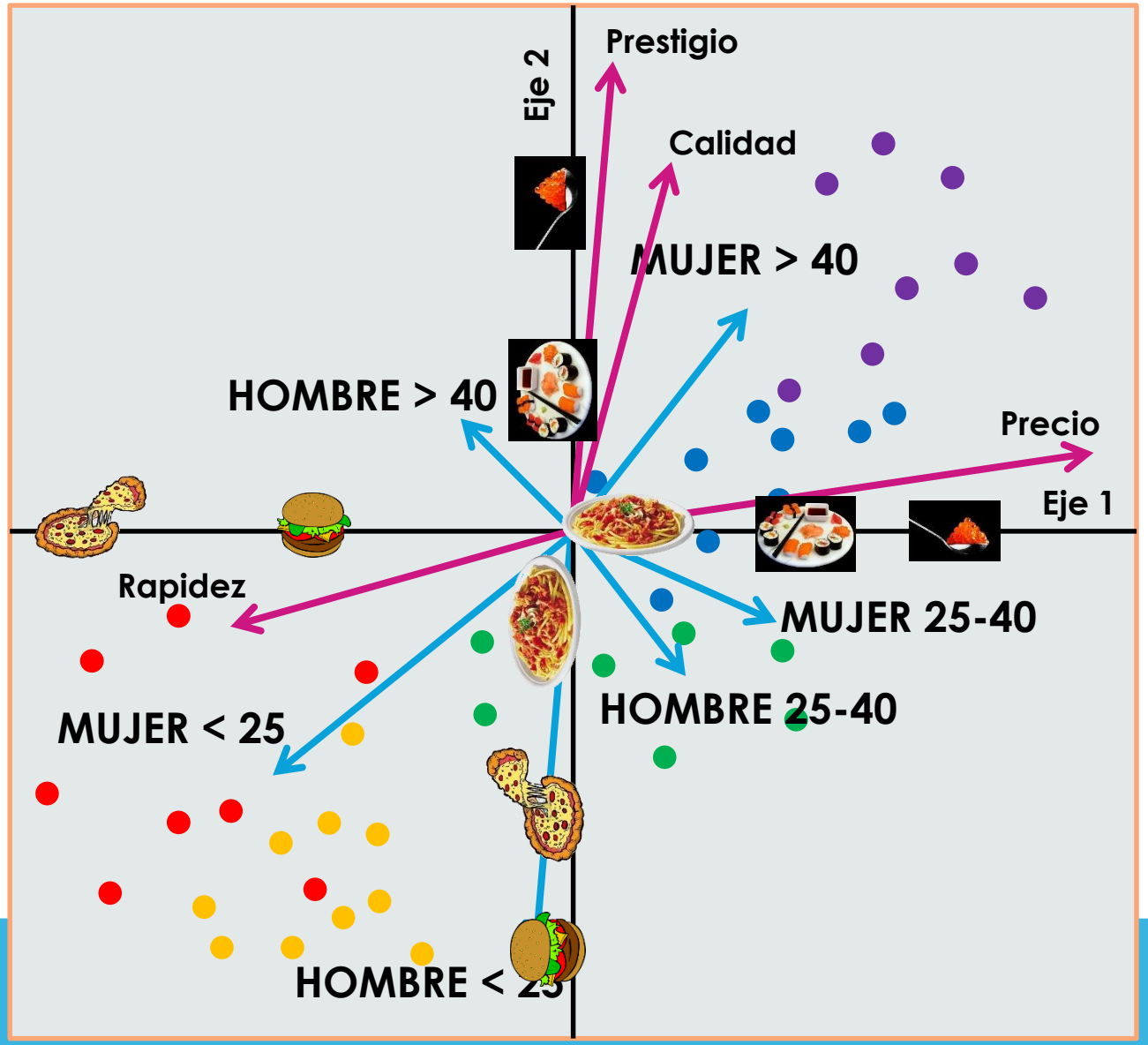
INDIVIDUOS= % Clientes por sexo y edad

- MUJER < 25
- MUJER 25-40
- MUJER > 40
- HOMBRE < 25
- HOMBRE 25-40
- HOMBRE > 40

VARIABLES que constriñen los ejes

- Precio medio del menu
- Rapidez del servicio
- Calidad nutricional
- Prestigio culinario




- PUNTOS = MUESTRAS = UNIDADES DE MUESTREO
- VECTORES AZUL = INDIVIDUOS = VARIABLES RESPUESTA
- VECTORES ROSA = VARIABLES AMBIENTALES = VARIABLES EXPLICATIVAS

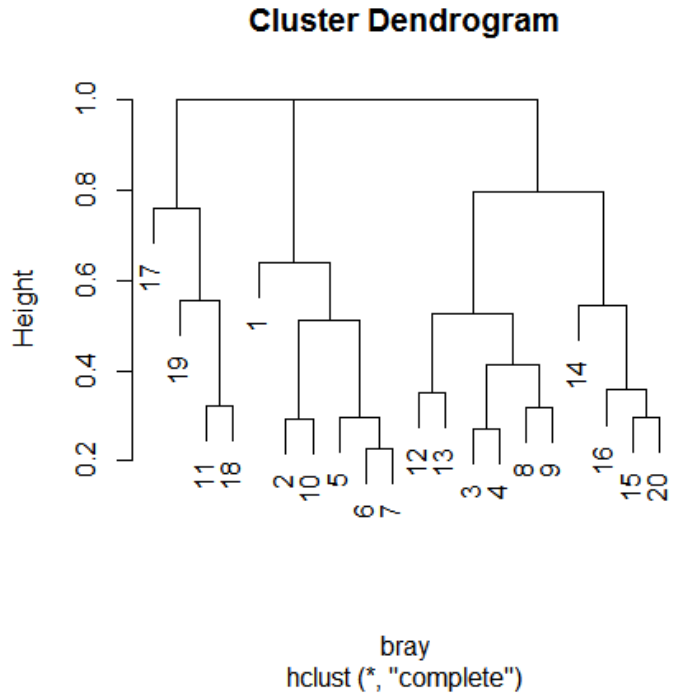
# MÉTODOS DE CLASIFICACIÓN

- ANÁLISIS CLUSTER

# ANÁLISIS CLUSTER

**"Conjunto de técnicas multivariantes que tienen por objetivo la clasificación de las muestras en grupos homogéneos o cluster"**

1. **ELECCIÓN DE DATOS**
   - Estandarización
2. **MEDIDA DE DISTANCIA/SIMILITUD**
   - Euclídea, Manhattan, Mahalanobis...
3. **CRITERIO DE AGRUPACIÓN**
   - **JERÁRQUICO**
     - AGLOMERATIVO
     - DIVISIVO
   - **NO JERÁRQUICO**
4. **ELECCIÓN DEL NÚMERO DE GRUPOS**




**CLUSTER en R**

**library(vegan)**

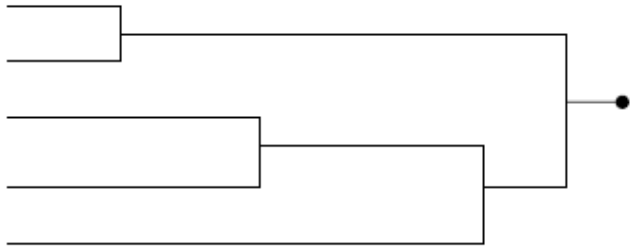
```
## JERÁRQUICO ##
##GENERAR MATRIZ DE DISTANCIAS##
> bray <-vegdist(dune, method = "bray", binary = "FALSE")
##GENERAR DENDROGRAMA##
> cluster <- hclust(bray, method = "complete", members = NULL)
> plot(cluster)
```

# ANÁLISIS CLUSTER

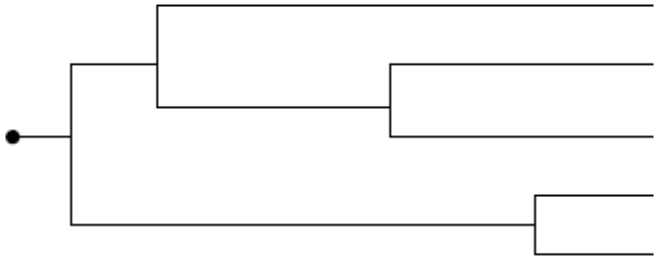
- **TIPOS**

## A. MÉTODOS JERÁRQUICOS

- Dendrograma o árbol de clasificación
- Las muestras se asignan a los cluster por un criterio de DISTANCIA
- AGLOMERATIVOS
  - Nearest Neighbour o Single linkage (distancia mínima)
  - Furthest Neighbour o Complete linkage (Distancia máxima)
  - Distancia entre centroides (centroid)
  - Distancia promedio (UPGMA o average linkage)
  - Distancia mediana (Median)
  - Ward



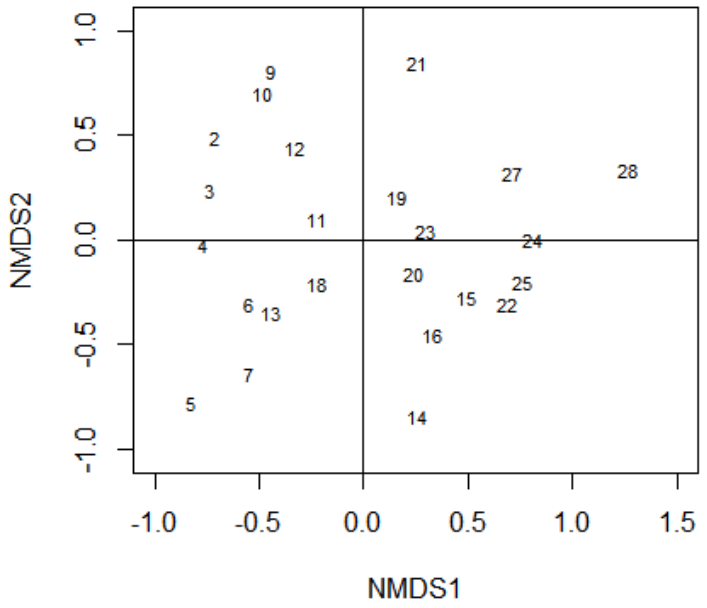
- DIVISIVOS
  - Cálculo iterativo de centros
  - Monothetic
  - Polythetic

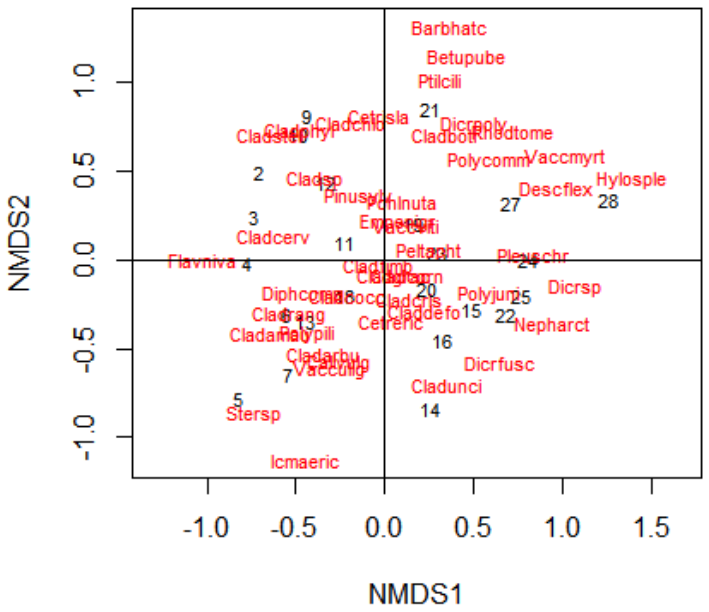


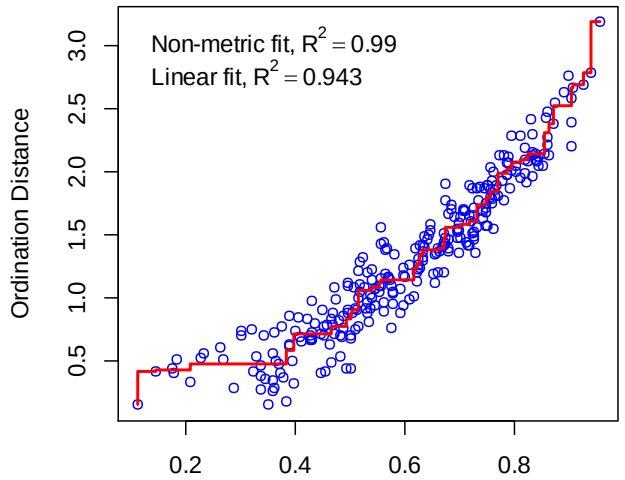
## B. MÉTODOS NO JERÁRQUICOS

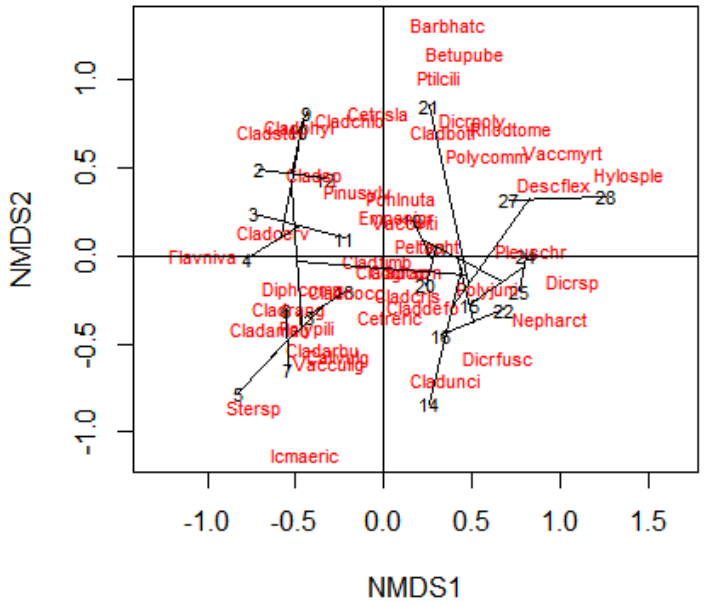
- Nº de cluster definido a priori
- Las muestras se intercambian entre los cluster, sin establecer relaciones entre ellos, según un criterio de optimización
  - K-means

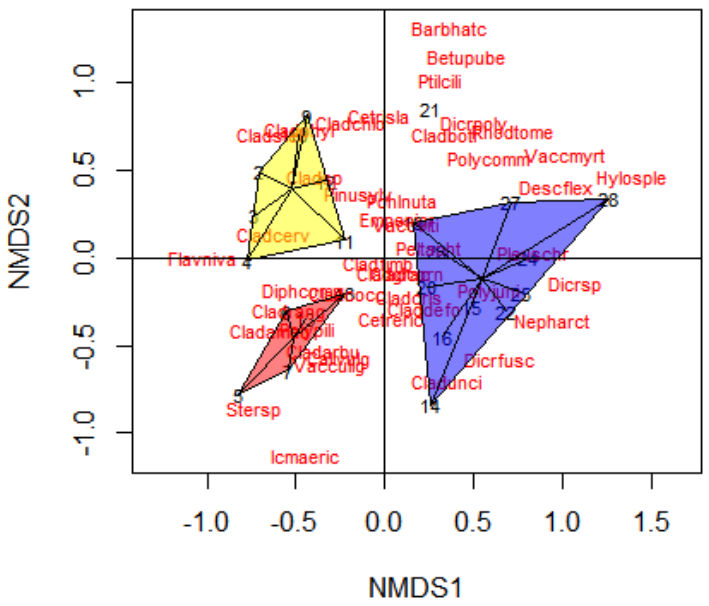
# EDICIÓN DE GRÁFICOS EN R



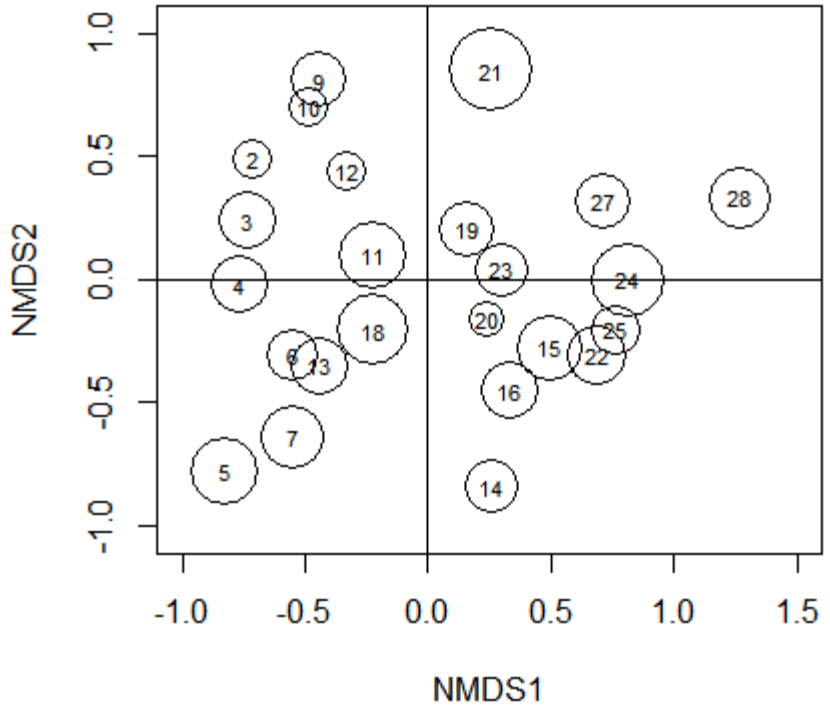



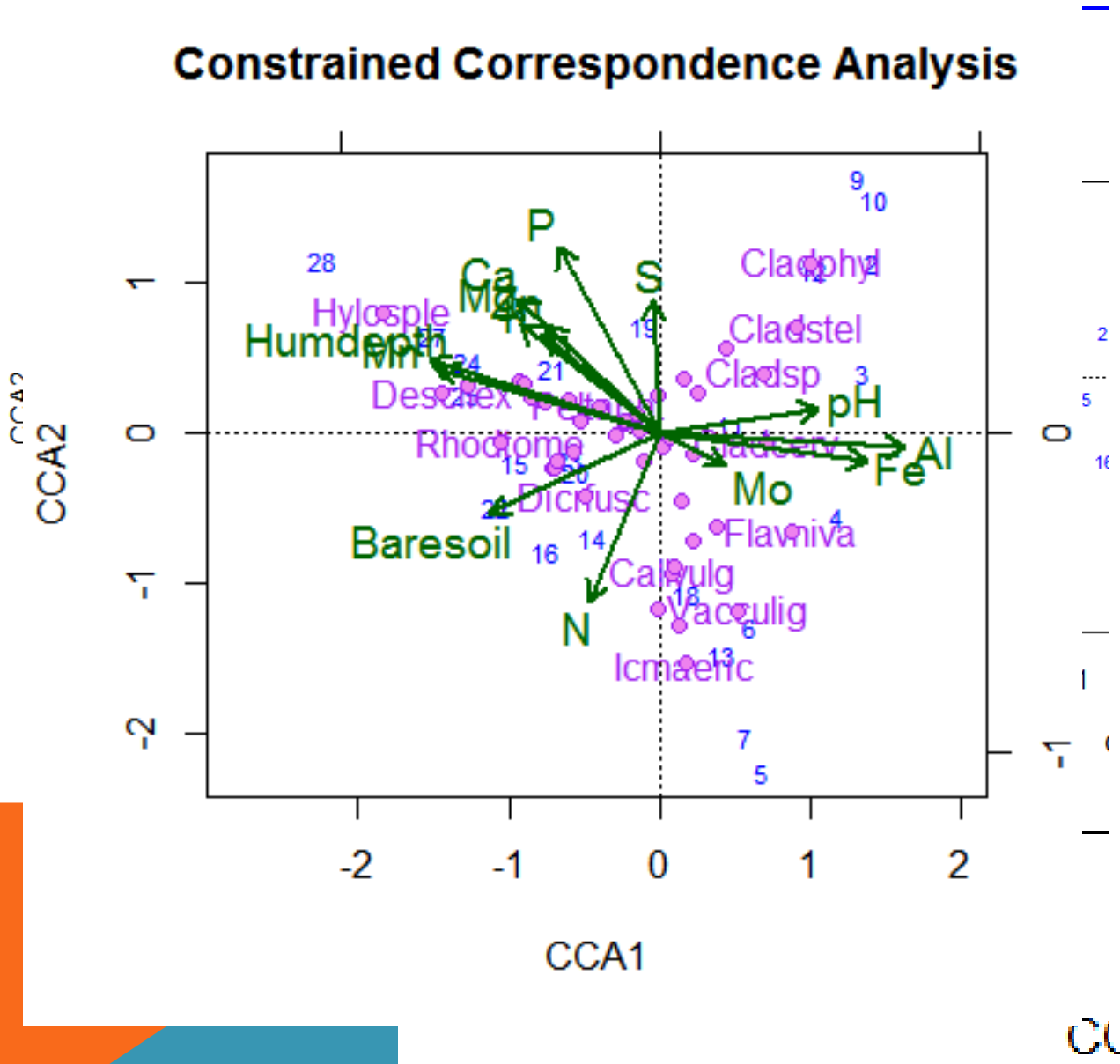



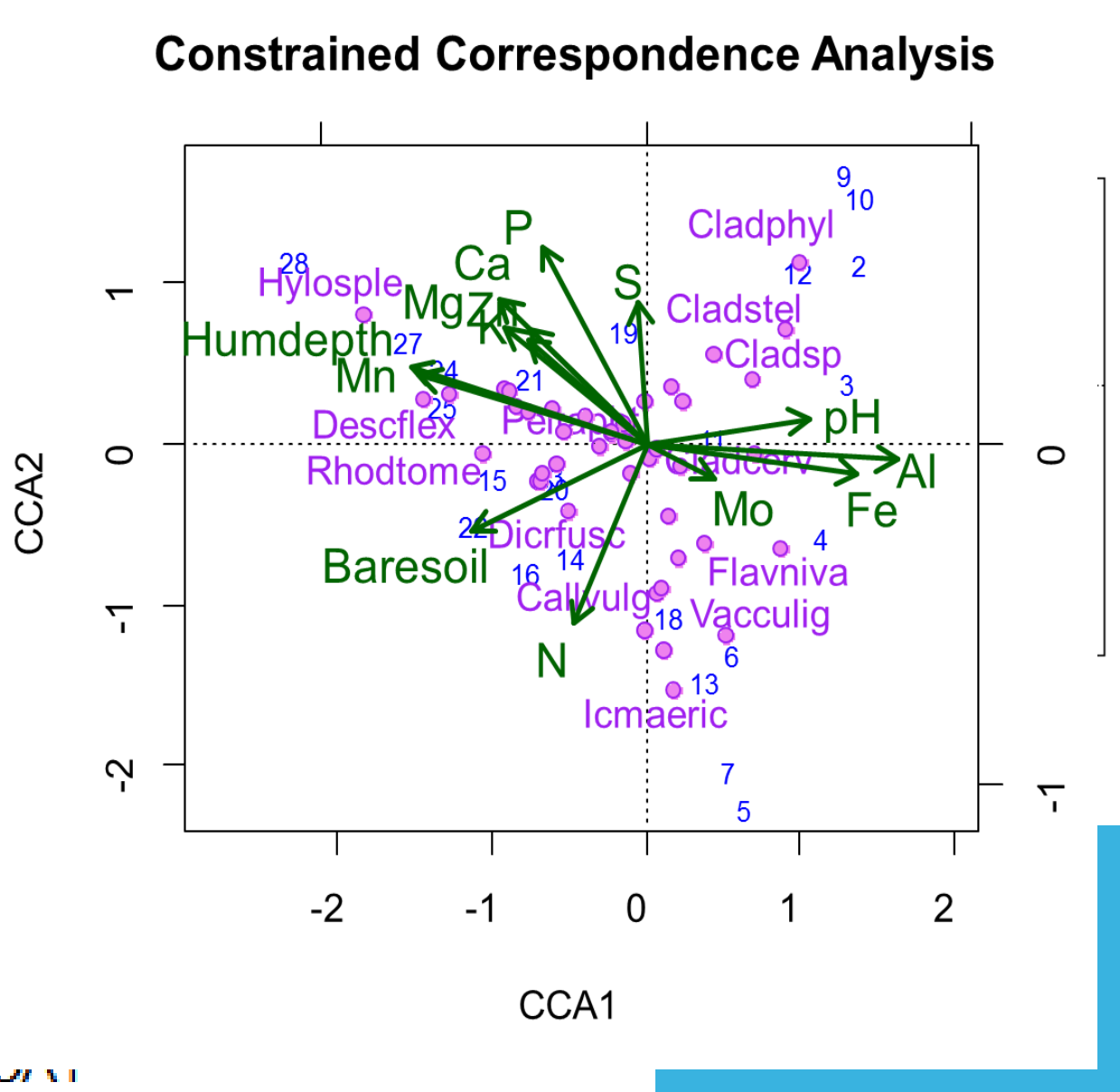




## Goodness of fit




Constrained Correspondence Analysis
P
Ca
Mg
S
Humdepth
Mn
pH
Al
Fe
Mo
Baresoil
N
Hylosple
Cladphyl
Cladstel
Cladsp
Descflex
Rhodtome
Dicrfusc
Flavniva
Callvulg
Vacculig
Icmaeric
CCA1
CCA2


Constrained Correspondence Analysis
P
Ca
Mg
K
S
Humdepth
Mn
pH
Al
Fe
Mo
Baresoil
N
Hylosple
Cladphyl
Cladstel
Cladsp
Descflex
Rhodtome
Dicrfusc
Flavniva
Callvulg
Vacculig
Icmaeric
CCA1
CCA2

orditkplot()

# MUCHAS GRACIAS POR SU ATENCIÓN