

Bioestadística avanzada Machine learning + Análisis multivariado

Adriana Pérez

Machine learning

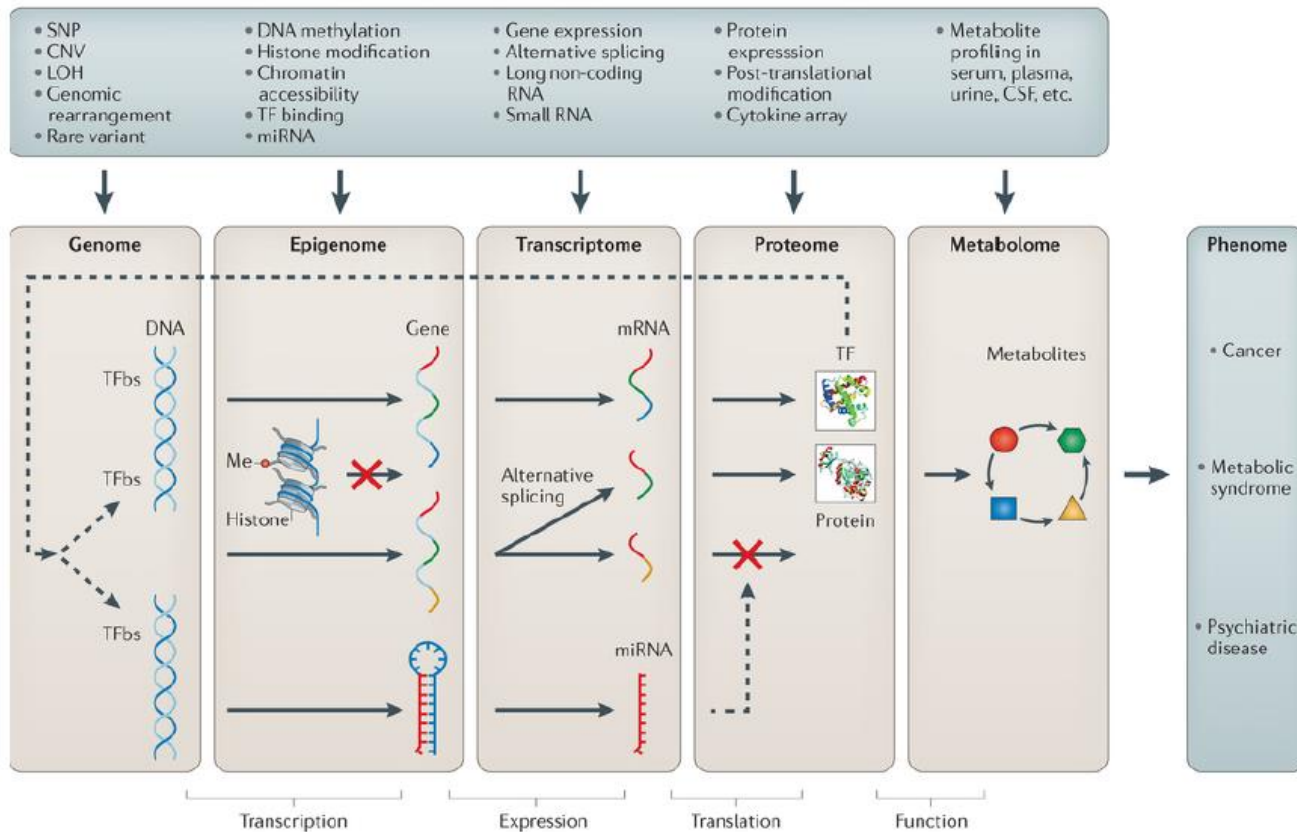
Aprendizaje automático

- Disciplina de ciencias de la computación (ciencia de datos)
- Una máquina (i.e., algoritmo or modelo) es capaz de efectuar nuevas predicciones basadas en los datos
- La máquina usa un **conjunto de entrenamiento (training set)** para aprender cómo las variables independientes (input) se relacionan con las variables respuesta (output)
- Luego que el algoritmo es entrenado, es testeado con un nuevo set de datos independientes (**conjunto de prueba o test set**) y se evalúa la calidad del método

- Aprendizaje supervisado:
 - Variables output: variable respuesta, dependiente (Y)
 - Variables input: predictoras, independientes, "explicatorias", features (X)
 - El objetivo es predecir a la VR en futuras observaciones
 - Problemas de Clasificación: La VR es cualitativa. Regresión logística
 - Problemas de Regresión: La VR es cuantitativa. Regresión lineal
- Aprendizaje no supervisado
 - No hay VR
 - El objetivo es entender la estructura de los datos (agrupamiento, ordenamiento)
 - Estudiar relación entre los individuos. Clustering
 - Estudiar la relación entre las variables. Análisis de componentes principales

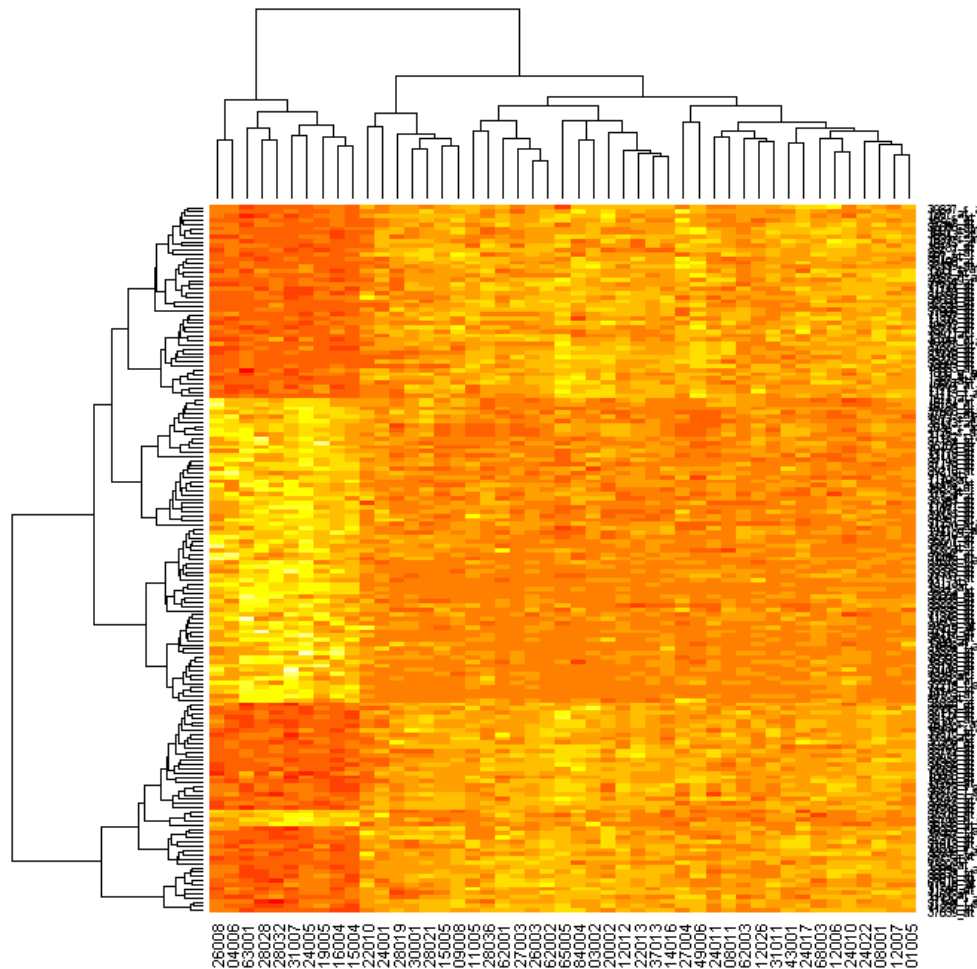
Omicas:

Una visión holística de un sistema biológico



Ritchie et al, 2015, *Nature Reviews Genetics*

Patrones de expresión de genes

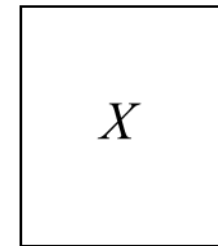


- 47 muestras de pacientes con leukemia linfocítica aguda
- Expresión de 165 genes por microarray

Variables

$x_1 \dots x_p$

n Objetos



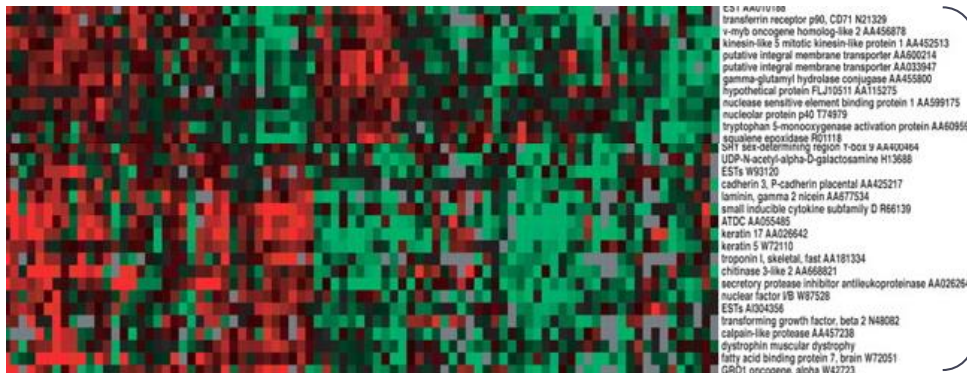
Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications

Sørlie T et al. PNAS 2001;98:10869-10874

Muestras de tejido mamario de 85 pacientes

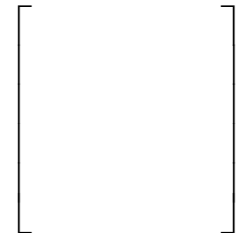
- 78 corresponden a distintos tipos de carcinomas de mama
- 3 tumores benignos
- 4 tejido normal de mama

Se midió la expresión de 476 genes distintos



p variables

n_1 objetos
+
 n_2 objetos
...+...
 n_k objetos



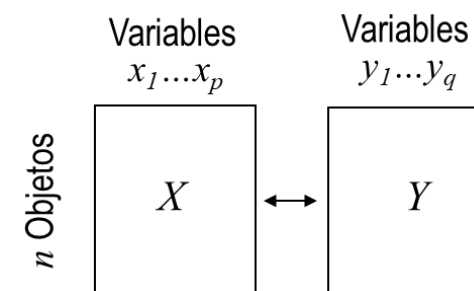
Objetos?
Variables?
Dimensiones de X?

Genetic variability in domesticated *Capsicum* spp as assessed by morphological and agronomic data

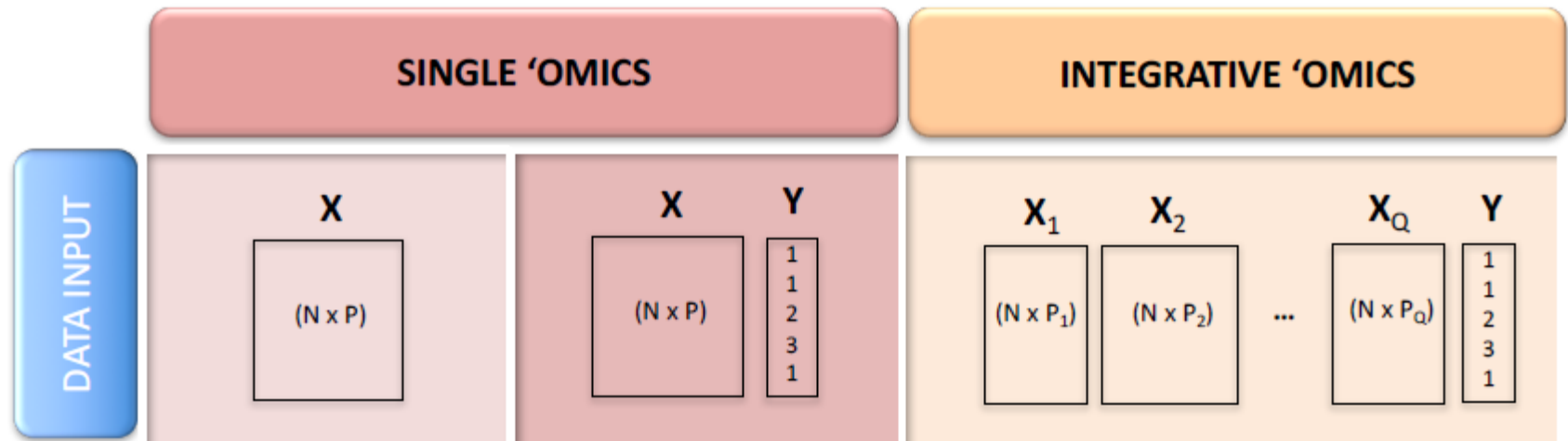


Table 2. Descriptors studied and discrimination of the classes observed for 56 *Capsicum* accessions according to the IPGRI (1995).

Descriptor	Classes observed according to descriptors for <i>Capsicum</i> (IPGRI, 1995)
Morphological traits	
Stem color	1: green; 2: green with purple stripes; 3: purple
Anther color	2: yellow; 3: pale blue; 4: blue; 5: purple
Corolla color	1: white; 8: purple; 9: white with yellow-green spots; 10: white-green; 11: yellow with purple base; 12: purple with yellow base
Number of flowers per axil	1: one; 2: two; 3: three or more
Flower position	3: pendant; 5: intermediate; 7: erect
Plant growth habit	5: intermediate (compact); 7: erect
Immature fruit color	2: yellow; 3: green; 4: orange; 5: purple; 7: other
Fruit color at mature stage	1: white; 3: pale orange-yellow; 4: orange-yellow; 5: pale orange; 6: orange; 7: light red; 8: red; 9: dark red; 10: purple
Fruit shape	1: elongate; 2: almost round; 3: triangular; 4: campanulate; 5: blocky; 6: tomato-pepper; 7: ellipse; 8: Scotch bonnet
Fruit surface	1: smooth; 2: semi-wrinkled; 3: wrinkled
Number of locules	Observed cross-section of 10 cut fruits
Cotyledon color	1: light green; 5: purple; 7: variegated
Calyx annular constriction	0: absent; 1: present
Neck at base of fruit	0: absent; 1: present
Agronomic traits	
Plant height (cm)	Recorded when in 50% of plants the first fruit has begun to ripen
Plant canopy width (cm)	Measured immediately after first harvest, at the widest point
Days to flowering	Number of days from transplanting until 50% of plants have at least one open flower
Days to fruiting	Number of days from transplanting until 50% of plants bear mature fruits
Fruit length (cm)	Measured at the largest point in 10 ripe fruits
Fruit width (cm)	Measured at the widest point, as an average of 10 ripe fruits
Mean fruit weight (g)	Ratio of total fruit weight/total fruit number
1000-seed weight (g)	Estimated counting 250 seeds
Number of seeds per fruit	Counting the total number of seeds per fruit. Average of at least 10 fruits from 10 random plants
Fruit number per plant	Total number of fruits harvested in each plant
Fruit yield per plant (g)	Total weight of all harvested fruits in each plant



Tipos de datos y métodos analíticos



- ✓ Análisis de componentes principales
- ✓ Análisis de agrupamiento o de clusters

No supervisado

- ✓ Regresión lineal
- ✓ Regresión logística
- ✓ Análisis discriminante

Supervisado

- ✓ Análisis canónicos

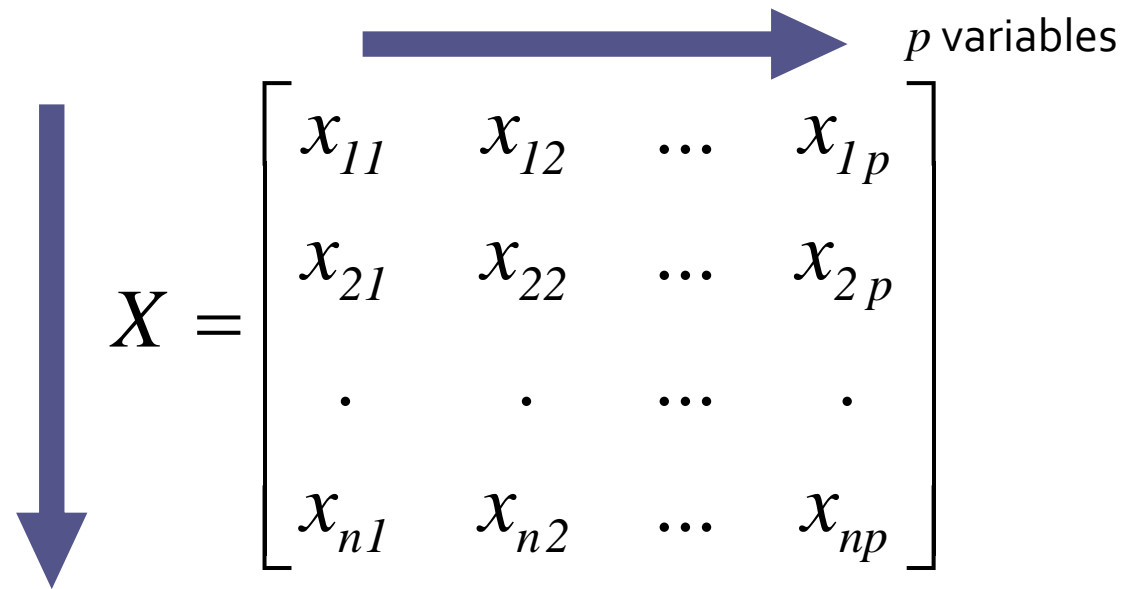
Métodos LASSO aplicados en los modelos de regresión para realizar la selección de variables (métodos lineales "sparse" o dispersos)

Métodos no supervisados

- Para describir la estructura subyacente de un fenómeno de una manera más parsimoniosa (reducción o simplificación)
 - Ordenamiento y agrupación de objetos
 - Reducción o simplificación de variables

Cuando se dispone de un gran número de variables, es natural preguntarse si éstas pueden ser reemplazadas por un número menor de variables o de funciones de estas variables, sin demasiada pérdida de información, para facilitar el análisis y la interpretación de los datos (Rao, 1964)

Matriz de datos X de $n \times p$



$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \cdot & \cdot & \dots & \cdot \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

n unidades experimentales o de observación

- $i=1$ a n objetos con $j=1$ a p variables medidas en cada objeto: Matriz de datos X de $n \times p$: n ue, p variables. Dimensionalidad: $n \times p$
- Se registra más de una variable en n objetos (unidades experimentales, unidades de observación, individuos)
- los datos se deben visualizar como una nube p dimensional

Datos demográficos

- % de población urbana (residente en localidades de al menos 2500 hab)
- % de analfabetismo en individuos mayores de 15 años
- tasa global de fecundidad, \approx promedio de hijos por mujer

$n = 20$

$p = 3$



País	% población urbana	%Analfabetismo	Fecundidad
Argentina	91,8	2,8	2,4
Bolivia	64,2	11,7	4,0
Brasil	83,4	11,1	2,3
Chile	86,6	3,5	2,0
Colombia	76,6	7,1	2,6
Costa Rica	62,6	3,8	2,3
Cuba	76,1	2,7	1,6
Ecuador	62,8	7,0	2,8
El Salvador	57,8	18,9	2,9
Guatemala	50,0	28,2	4,6
Haití	41,8	45,2	4,0
Honduras	47,9	22,0	3,7
México	76,5	7,4	2,5
Nicaragua	56,9	31,9	3,3
Panamá	65,8	7,0	2,7
Paraguay	58,4	5,6	3,8
Perú	72,6	8,4	2,9
Rep. Dominic	65,6	14,5	2,7
Uruguay	91,9	2,0	2,3
Venezuela	92,8	6,0	2,7

Fuente: Anuario estadístico de América Latina y el Caribe 2006 editado por la Comisión Económica para América Latina y el Caribe (CEPAL) (<http://www.eclac.org>).

Análisis de clusters o agrupamiento

- Es una técnica de clasificación
- El objetivo es **agrupar** objetos en grupos o *clusters* homogéneos basándose en su similitud MV
- Generalmente exploratorio: desarrollar una clasificación objetiva de los objetos
- También llamado análisis de conglomerados, análisis de clasificación, taxonomía numérica
- Análisis de similitud-disimilitud entre objetos
- La pertenencia de los objetos a los grupos no está definida *a priori* sino que surge de los datos \Rightarrow el método detecta la **estructura** “**natural**” de agrupamiento

Medidas de similitud-disimilitud entre objetos

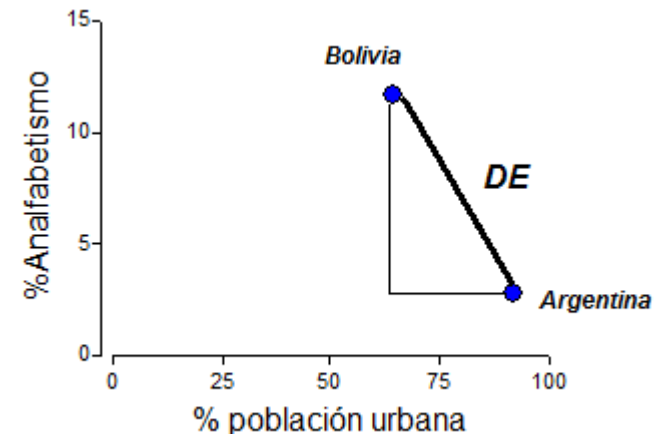
- Los índices de similitud miden cuan similares son dos objetos; los de disimilitud miden cuan distintos son y representan la **distancia multivariada**: cuan alejados están dos objetos en el espacio MV. Obviamente son conceptos complementarios.
- Existen distintas formas de medir distancias (según si las variables son cuali o cuantitativas, según escala de medición, etc)

Distancia euclídea

$$D_{E_{i,k}} = \sqrt{\sum_{j=1}^p (x_{ij} - x_{kj})^2}$$

$$D_{E_{i,k}} = \sqrt{(91,80 - 64,2)^2 + (2,80 - 11,70)^2} = 29$$

País	% población urbana	%Analfabetismo
Argentina	91,8	2,8
Bolivia	64,2	11,7



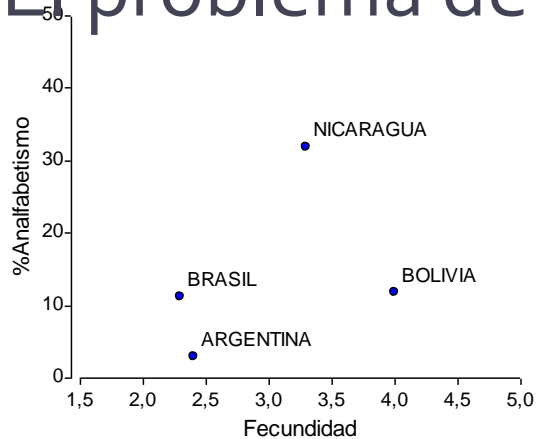
- Para variables cuantitativas es la más usada
- Varía entre 0 e infinito
- Es fuertemente afectada por la escala de las variables
- Si las variables tienen distintas unidades, los datos deben ser **estandarizados** para evitar distorsiones espúreas (que dominen las variables de mayor magnitud o de mayor varianza)

Matriz de disimilitud D de $n \times n$

$$D = \begin{bmatrix} 0 & d_{12} & \dots & d_{1n} \\ d_{21} & 0 & \dots & d_{2n} \\ \cdot & \cdot & \dots & \cdot \\ d_{n1} & d_{n2} & \dots & 0 \end{bmatrix}$$

- Cuadrada, de $n \times n$
- Simétrica
- Ceros en la diagonal
- Puede intercambiarse disimilitud (distancia) con similitud

Distancias: El problema de la escala



	País	% población urbana	%Analfabetismo	Fecundidad
1	Argentina	91,8	2,8	2,4
2	Bolivia	64,2	11,7	4
3	Brasil	83,4	11,1	2,3
4	Nicaragua	56,9	31,9	3,3

Medidas de distancia y similitudes

Euclidea

	1	2	3	14
1	0,00			
2	9,04	0,00		
3	8,30	1,80	0,00	
14	29,11	20,21	20,82	0,00

Euclidea (datos estandarizados)

	1	2	3	14
1	0,00			
2	2,12	0,00		
3	0,68	2,11	0,00	
14	2,61	1,85	2,09	0,00

$$z = \frac{x_i - \bar{x}}{DE}$$

Sin estandarizar:

1. Argentina-Nicaragua
2. Brasil-Nicaragua
3. Bolivia-Nicaragua
4. Argentina-Bolivia
5. Argentina-Brasil
6. Bolivia-Brasil

Estandarizando:

1. Argentina-Nicaragua
2. Argentina-Bolivia
3. Bolivia-Brasil
4. Brasil-Nicaragua
5. Bolivia-Nicaragua
6. Argentina-Brasil

Matrices de distancias

```
#estandarizar la matriz X  
scale(x)
```

```
#calcular las distancias  
library(vegan)  
vegdist(x, method="bray", binary=FALSE, diag=FALSE,  
upper=FALSE, na.rm = FALSE, ...)
```

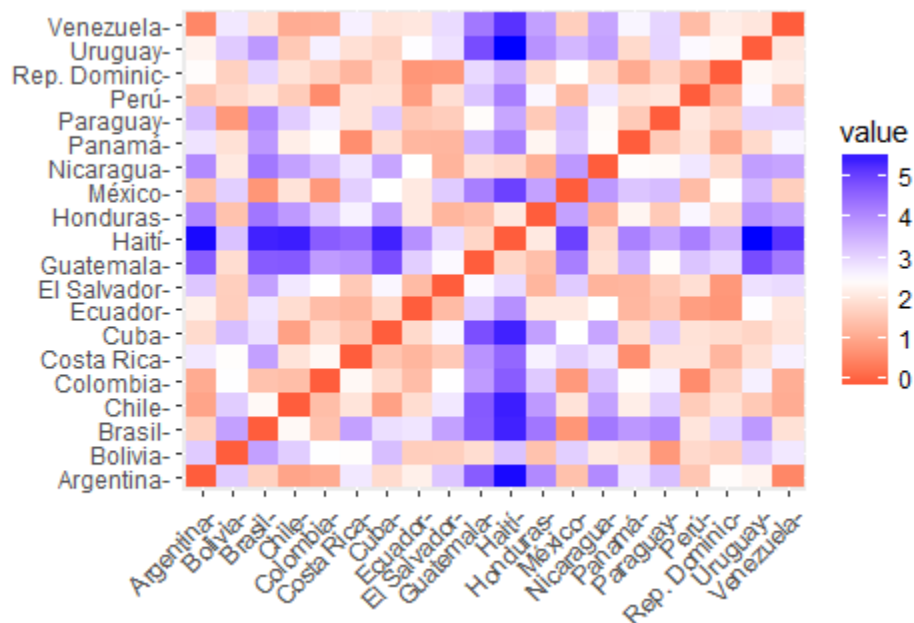
```
method  "manhattan", "euclidean", "canberra", "bray",  
"kulczynski", "jaccard", "gower", "altGower", "morisita",  
"horn", "mountford", "raup" , "binomial", "chao", "cao"  
"mahalanobis"
```

```
#Otra opción  
library(stats)  
dist(x, method = "euclidean", diag = FALSE, upper = FALSE)
```

```
method    "euclidean", "maximum", "manhattan", "canberra",  
          "binary"
```

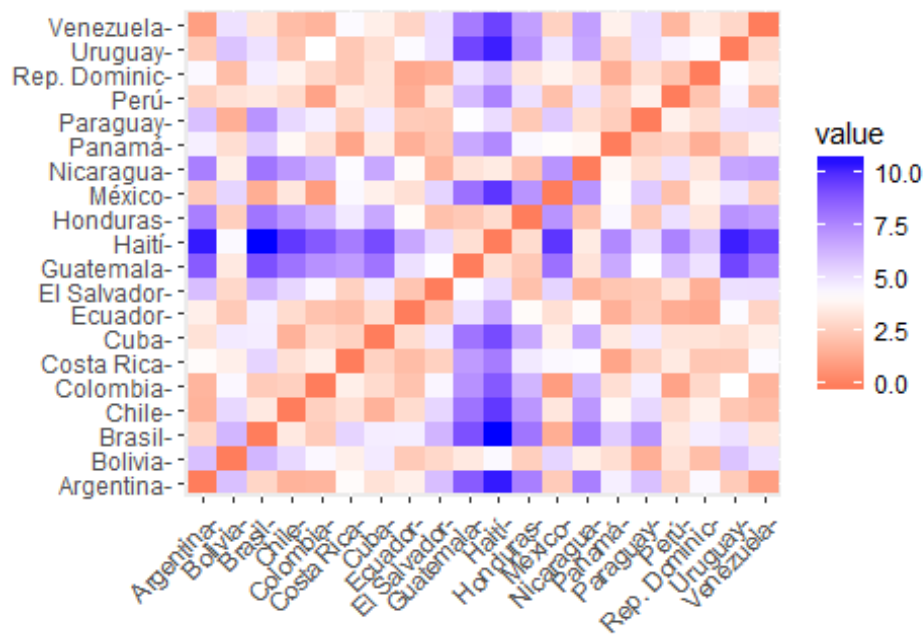
Matriz de distancias euclídeas con datos estandarizados

	Argentina	Bolivia	Brasil	Chile	Colombia
Argentina	0.0				
Bolivia	8.6	0.0			
Brasil	4.2	9.3	0.0		
Chile	2.5	8.2	5.7	0.0	
Colombia	3.2	7.1	5.2	3.5	0.0



Matriz de distancias Manhattan con datos estandarizados

	Argentina	Bolivia	Brasil	Chile	Colombia
Argentina	0.0				
Bolivia	34.3	0.0			
Brasil	16.0	35.8	0.0		
Chile	9.3	31.2	21.0	0.0	
Colombia	13.1	27.6	22.0	13.4	0.0



Medidas de similitud-disimilitud

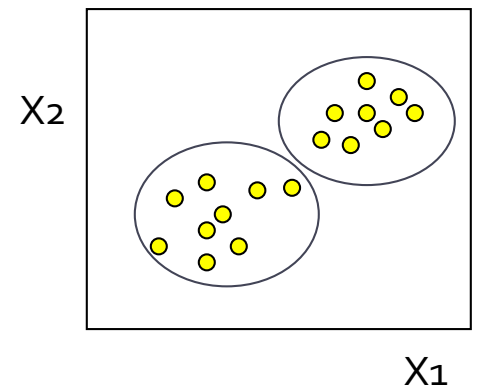
- para v.continuas
 - Distancia euclídea
 - Distancia City block o Manhattan
 - Distancia de Mahalanobis
 - Bray-Curtis o Sørensen
- para v.cualitativas (conteos)
 - Chi-cuadrado
 - Coeficiente de Jaccard (para v-dicotómicas: presencia-ausencia), similitud
- Para v. cuali y cuanti
 - Coeficiente de Gower, similitud

¿Cuál elegir?

- Las medidas de similitud pueden ser convertidas a medidas de distancia y viceversa.
- Si las variables están medidas en la misma escala y no hay ceros: Euclídea o Manhattan
- Si hay valores ausentes se puede utilizar distancias promedio
- Si se estudia abundancia de sp, la distancia debería alcanzar un máximo cuando las unidades no tienen sp en común. Ciertamente para Bray Curtis pero no para Euclídea y Manhattan
- Para datos binarios: Bray Curtis / Sørensen
- Cuando la matriz de datos contiene v. cuali y cuanti: Gower

Clusters: procedimiento

- Los objetos se comparan en función de una medida de (di)similitud
- Se generan grupos o clusters de manera tal que los objetos en un mismo cluster son más similares entre sí que con respecto a los de otros clusters



Tres pasos para el agrupamiento

1. Definir una medida de (di)similitud entre objetos: Medidas de distancia = disimilitud
 - Distancia euclídea
 - City block o Manhattan
 - etc
2. Definir el criterio de agrupamiento
 - Jerárquico
 - No jerárquico
3. Definir el método de ligamiento

```
dist(x, method = "euclidean",  
upper = FALSE)
```

Criterio de agrupamiento

- Distintos algoritmos de agrupamiento
- Todos se basan en maximizar las diferencias entre clusters en relación a la variación dentro de los clusters
- Métodos:
 - Jerárquicos
 - No jerárquicos

Métodos de agrupamiento jerárquicos

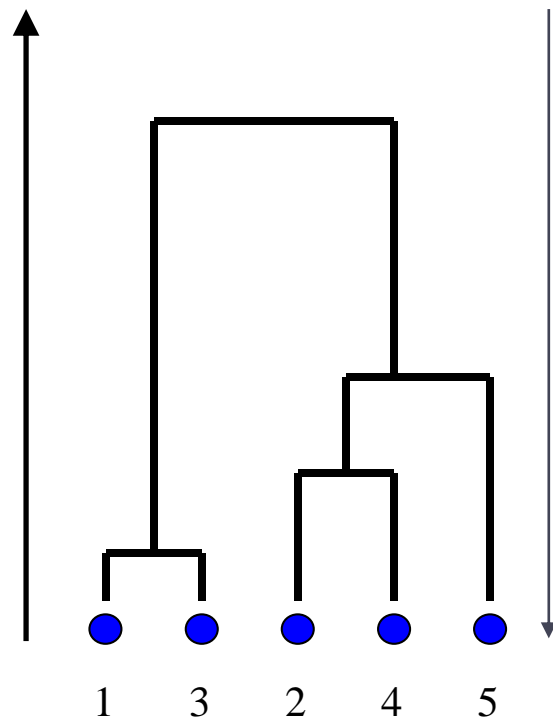
Los objetos se agrupan en una sucesión anidada:

1. Se calcula la **matriz de disimilitud** entre todos los pares de objetos
 2. Se forma el primer cluster con los dos objetos con menor disimilitud
 3. Se recalcula la disimilitud entre este cluster y los restantes objetos
 4. Se forma un 2do cluster entre los dos objetos con menor disimilitud
 5. El proceso continúa hasta que todos los objetos pertenecen a un único cluster
- Aglomerativo, aunque se puede pensar al revés (divisivo)
 - Los objetos no pueden cambiar de cluster una vez que fueron asignados

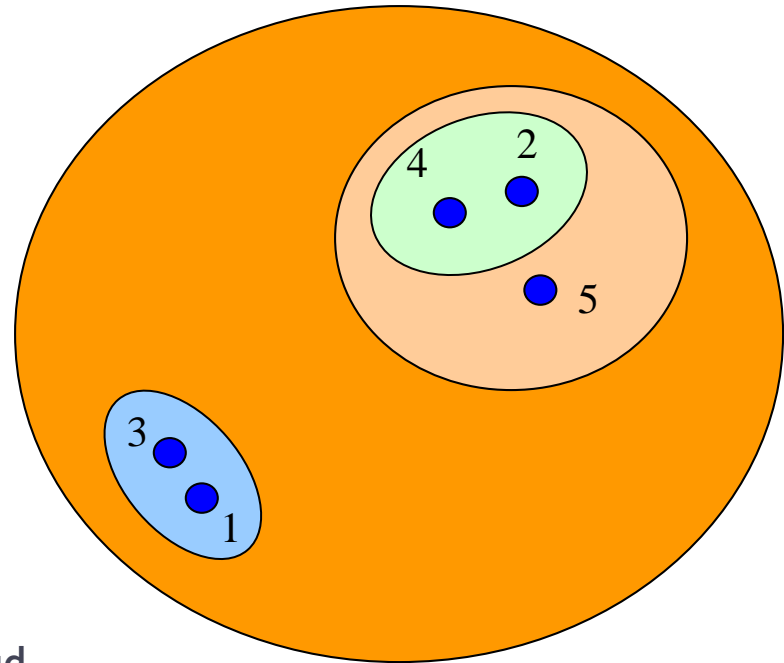
Métodos de agrupamiento jerárquicos

- Dendrograma: Representación gráfica
 - Árbol binario que muestra la estructura de los clusters
 - Provee una medida de la similitud entre clusters

Distancia entre clusters

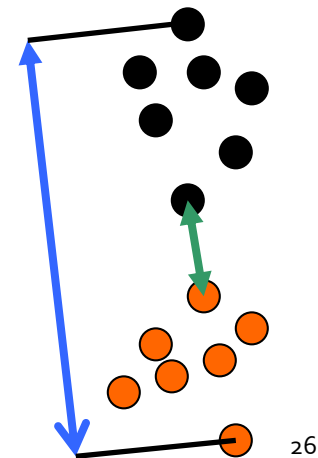


Similitud



¿Y cómo se mide la distancia entre clusters (paso 3)?

- Ligamiento simple, vecino más cercano: distancia mínima entre objetos de cada cluster
 - Tiende a separar los extremos antes que separar los grupos principales \Rightarrow tiende a producir agrupamientos elongados, en cadena
- Ligamiento completo, vecino más lejano: distancia máxima entre objetos de cada cluster
 - Tiende a producir grupos compactos, de igual diámetro. Sensible a valores extremos
- ligamiento promedio: promedio de las distancias de todas las combinaciones de dos objetos, uno de cada cluster
 - Tiende a producir grupos de igual varianza.



¿Y cómo se mide la distancia entre clusters (paso 3)?

- ❑ Método del centroide: distancia entre objetos y centroides. Es el más robusto a valores extremos
- ❑ Método de Ward o de mínima varianza: mín SC dentro de cada cluster para todas las variables. Se recomienda para datos normales y homocedásticos. Sensible a valores extremos

Un ejemplo

Matriz de datos $n \times p$

Classical format

	Variable				
	1	2	3	4	5
Participant 1	1.0	5.0	7.0	7.0	5.0
Participant 2	2.0	6.0	2.0	8.0	6.0
Participant 3	4.0	4.0	5.0	5.0	4.0
Participant 4	5.0	1.0	4.0	4.0	3.0
Participant 5	7.0	2.0	3.0	9.0	2.0

Matriz de distancia euclídea al cuadrado $n \times n$

Object	Object				
	1	2	3	4	5
1	0				
2	29	0			
3	19	30	0		
4	54	63	13	0	
5	74	59	37	32	0

Matrices de distancia según algoritmo de agrupamiento

Single linkage method squared Euclidean distance matrices

Object	Object				Object	Object		
	1	2	(3,4)	5		(1,3,4)	2	5
1	0				(1,3,4)	0		
2	29	0			2	29	0	
(3,4)	19	30	0		5	32	59	0
5	74	59	32	0				

Maximum linkage method squared Euclidean distance matrices

Object	Object				Object	Object		
	1	2	(3,4)	5		(1,2)	(3,4)	5
1	0				(1,2)	0		
2	29	0			(3,4)	63	0	
(3,4)	54	63	0		5	74	37	0
5	74	59	37	0				

Unweighted pair-group method squared Euclidean distance matrices

Object	Object				Object	Object		
	1	2	(3,4)	5		(1,2)	(3,4)	5
1	0				(1,2)	0		
2	29	0			(3,4)	41.5	0	
(3,4)	36.5	46.5	0		5	66.5	34.5	0
5	74	59	34.5	0				

Centroid method squared Euclidean distance matrix

Object	Object			
	1	2	(3,4)	5
1	0			
2	29	0		
(3,4)	33.3	43.3	0	
5	74	59	31.2	0

Gore P.A. 2000.
Handbook of applied
multivariate statistics

Métodos de ligamiento: ¿cuál usar?

- Si existe gran diferencia entre los grupos, los resultados serán similares utilizando cualquiera de los métodos
- Pero en caso contrario, se pueden obtener resultados muy distintos según el método aplicado
- Ligamiento simple y completo usan la información de sólo dos objetos, sin considerar la estructura del cluster \Rightarrow pueden producir resultados irregulares

```
hclust(d, method = "complete", members = NULL)
```

`d` a dissimilarity structure as produced by `dist`.

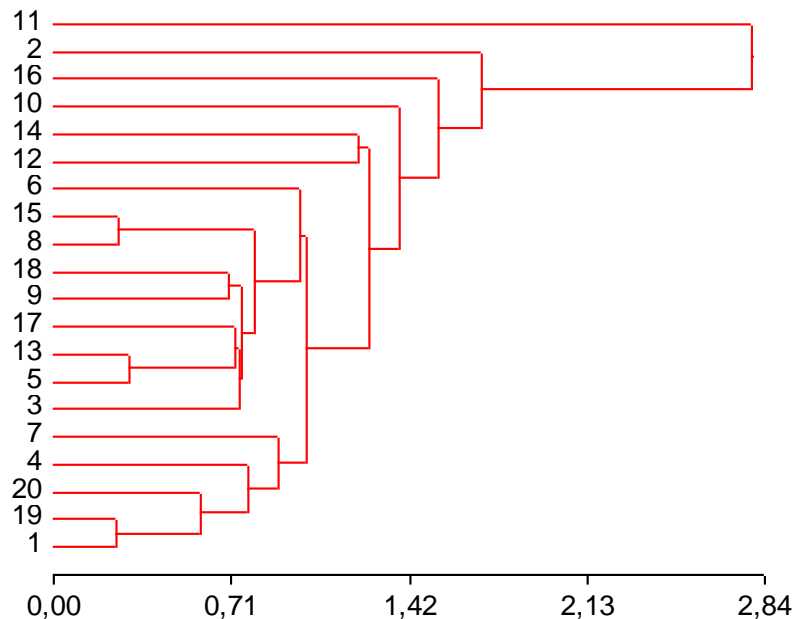
`method` the agglomeration method to be used. This should be (an unambiguous abbreviation of) one of "ward.D", "ward.D2", "single", "complete", "average" (= UPGMA), "mcquitty" (= WPGMA), "median" (= WPGMC) or "centroid" (= UPGMC).

```
plot(x)
```

```
di<-dist(data)
cl<-hclust(di)
plot(cl)
```

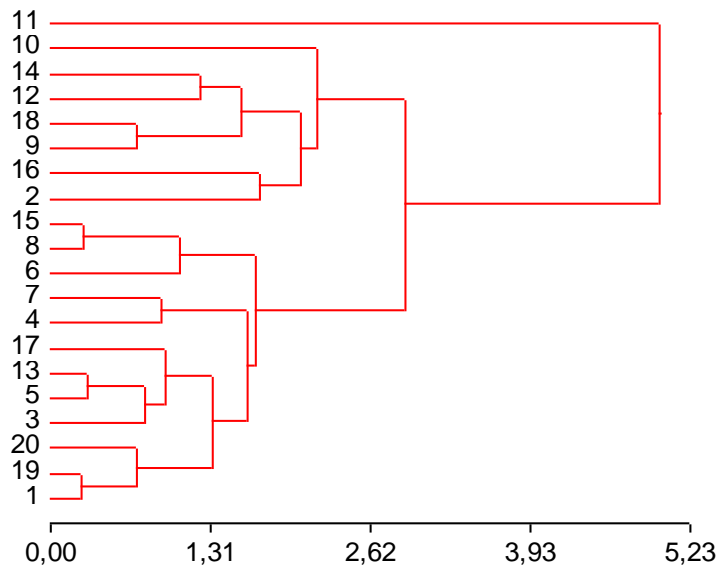
Encadenamiento Simple (Single linkage)

Distancia: (Euclidea)



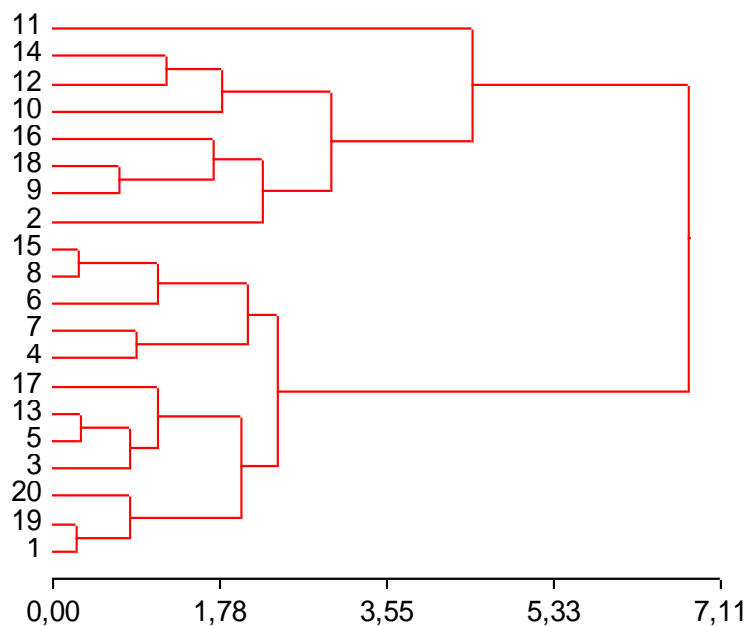
Promedio (Average linkage)

Distancia: (Euclidea)



Encadenamiento Completo (Complete linkage)

Distancia: (Euclidea)



- 1 Argentina
- 2 Bolivia
- 3 Brasil
- 4 Chile
- 5 Colombia
- 6 Costa Rica
- 7 Cuba
- 8 Ecuador
- 9 El Salvador
- 10 Guatemala
- 11 Haití
- 12 Honduras
- 13 México
- 14 Nicaragua
- 15 Panamá
- 16 Paraguay
- 17 Perú
- 18 Rep. Dominic
- 19 Uruguay
- 20 Venezuela

- % población urbana
- %Analfabetismo
- Fecundidad
- Esperanza de vida

Validación

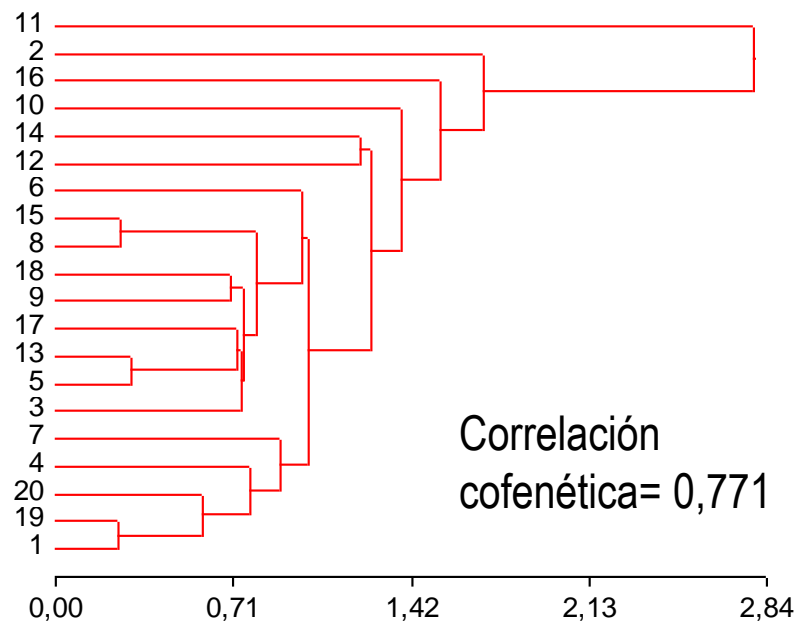
¿Los resultados son estables y representativos de la población de la cual se extrajo la muestra?

- Se aconseja aplicar varios algoritmos de agrupamiento y de medidas de distancia y comparar resultados para diferenciar agrupamientos naturales y artificiales
- Coeficiente de correlación cofenética: mide la correlación entre las distancias en el dendrograma y las distancias originales de los objetos \Rightarrow seleccionar el agrupamiento con $>$ coeficiente

```
d1 <- dist(x)
hc <- hclust(d1, "ave")
d2 <- cophenetic(hc)
cor(d1, d2)
```

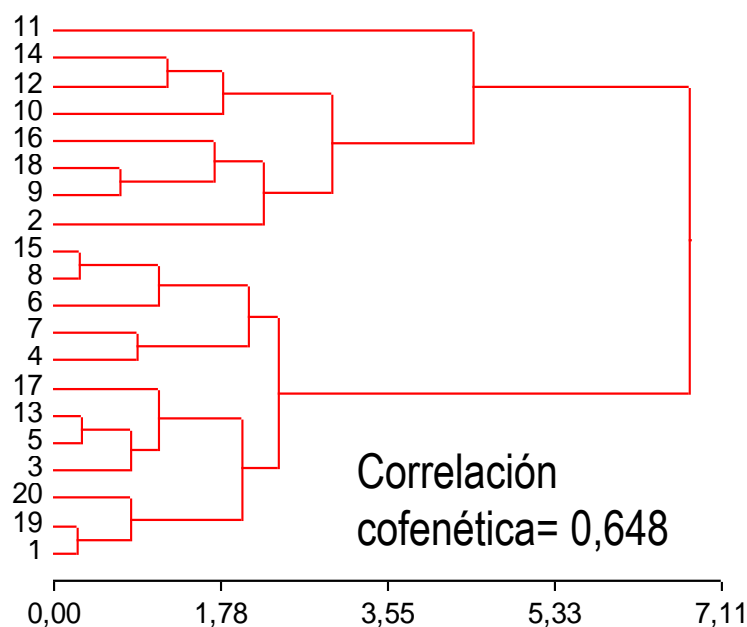

Encadenamiento Simple (Single linkage)

Distancia: (Euclídea)



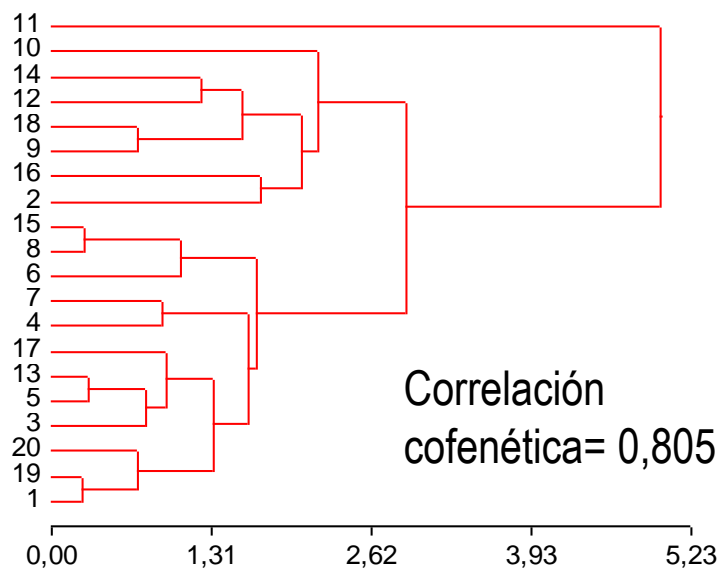
Encadenamiento Completo (Complete linkage)

Distancia: (Euclídea)



Promedio (Average linkage)

Distancia: (Euclídea)



- 1 Argentina
- 2 Bolivia
- 3 Brasil
- 4 Chile
- 5 Colombia
- 6 Costa Rica
- 7 Cuba
- 8 Ecuador
- 9 El Salvador
- 10 Guatemala
- 11 Haití
- 12 Honduras
- 13 México
- 14 Nicaragua
- 15 Panamá
- 16 Paraguay
- 17 Perú
- 18 Rep. Dominic
- 19 Uruguay
- 20 Venezuela

- % población urbana
- %Analfabetismo
- Fecundidad
- Esperanza de vida

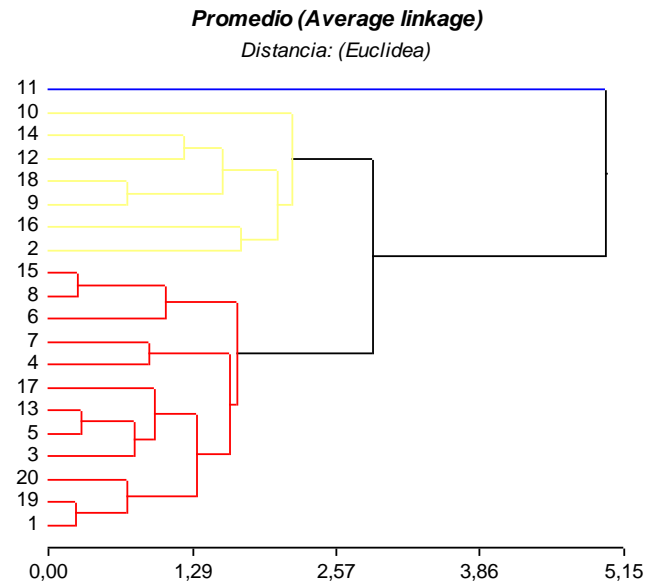
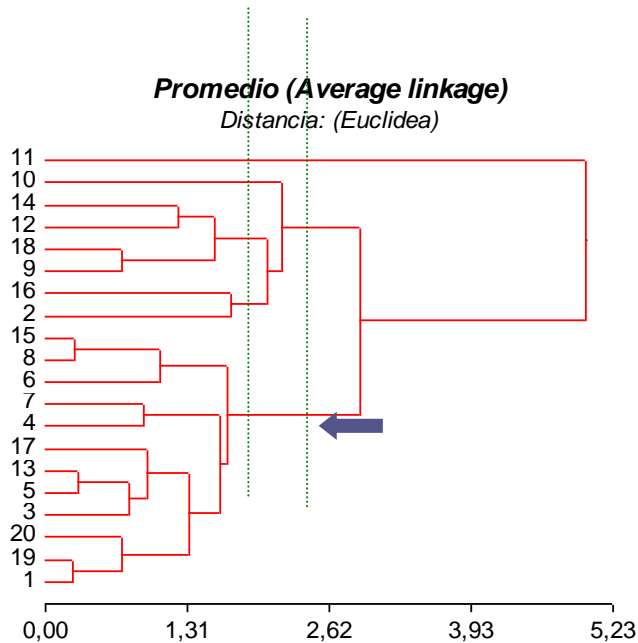
Cantidad óptima de clusters

- No existe un criterio objetivo
- Compromiso entre homogeneidad de los grupos y cantidad de grupos
- Opciones:
 - línea de corte (por ej. 50% de la máxima distancia)
 - Saltos importantes en los agrupamientos
 - Consideraciones teóricas

```
# corta el dendrograma en 5 clusters  
grupo <- cutree(c1, k=5)
```

```
# dibuja el dendrograma con bordes rojos alrededor de  
los 5 clusters  
rect.hclust(c1, k=5, border="red")
```

En el ejemplo

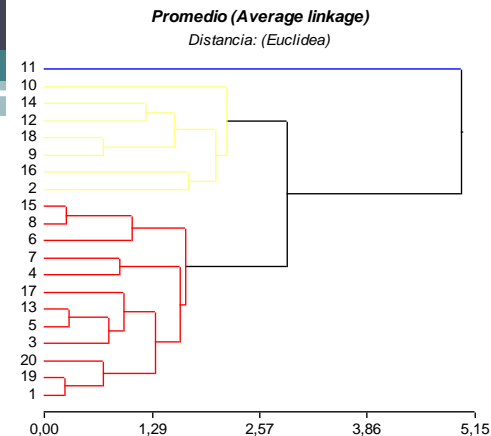


- 1 Argentina
- 2 Bolivia
- 3 Brasil
- 4 Chile
- 5 Colombia
- 6 Costa Rica
- 7 Cuba
- 8 Ecuador
- 9 El Salvador
- 10 Guatemala
- 11 Haití
- 12 Honduras
- 13 México
- 14 Nicaragua
- 15 Panamá
- 16 Paraguay
- 17 Perú
- 18 Rep. Dominic
- 19 Uruguay
- 20 Venezuela

Clasificación e interpretación

- Cada objeto es asignado a un cluster
- Los objetos pueden compararse en relación a las variables originales
- Los clusters pueden ser comparados en relación a sus centroides

En el ejemplo



Caso	País	% población urbana	%Analfabetismo	Fecundidad	Esperanza	Conglomerado
1	Argentina	91,80	2,80	2,40	74,30	1
2	Bolivia	64,20	11,70	4,00	63,80	3
3	Brasil	83,40	11,10	2,30	71,00	1
4	Chile	86,60	3,50	2,00	77,70	1

Argentina
 Bolivia
 Brasil
 Chile
 Colombia
 Costa Rica
 Cuba
 Ecuador
 El Salvador
 Guatemala
 Haití
 Honduras
 México
 Nicaragua
 Panamá
 Paraguay
 Perú
 Rep. Dominic
 Uruguay
 Venezuela

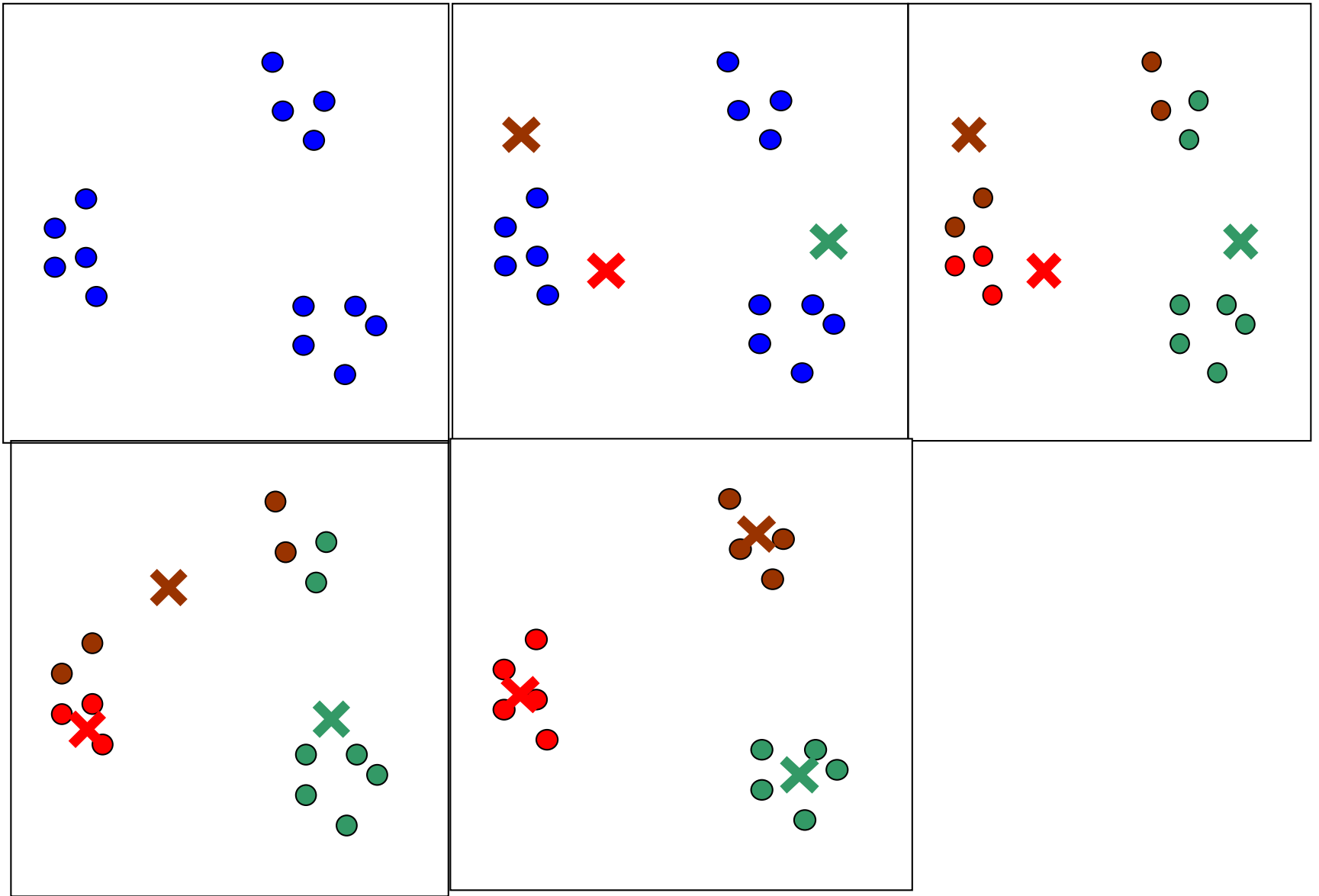
Estadística descriptiva

Conglomerado	Resumen	% población urbana	%Analfabetismo	Fecundidad	Esperanza
1	n	12,00	12,00	12,00	12,00
1	Media	78,29	5,73	2,43	74,18
1	CV	14,20	48,29	15,04	3,43
2	n	1,00	1,00	1,00	1,00
2	Media	41,80	45,20	4,00	59,20
2	CV	0,00	0,00	0,00	0,00
3	n	7,00	7,00	7,00	7,00
3	Media	57,26	18,97	3,57	69,24
3	CV	11,49	48,75	18,41	3,63

Métodos de agrupamiento no jerárquicos

Método de la K medias: la cantidad de k grupos se decide a priori

1. Se eligen k puntos arbitrariamente (semillas), y se los considera como centroides de los k clusters
 2. Se asignan los objetos al centroide más cercano, formando k clusters
 3. Se calculan los centroides de estos clusters arbitrarios
 4. Se realiza una nueva partición de los objetos, asignándolos al centroide más cercano
 5. El proceso se repite hasta que no hay pasajes de individuos de un cluster a otro
- Ojo: Los resultados pueden variar según la posición de los centros iniciales



```
# Análisis d cluster por K-Means
cl<-kmeans(x, centers)

#medias de los clusters
# get cluster means
aggregate(x,by=list(cl$cluster),FUN=mean)

cl
plot(x, col = cl$cluster)

points(cl$centers, col = 1:centers, pch = 8)

# agregar el cluster de pertenencia a cada
observación
x <- data.frame(x, cl$cluster)
```


Comentarios

- No es una técnica de inferencia estadística, por lo que no requiere de los supuestos usuales de normalidad y homocedasticidad.
- Las técnicas MV son muy sensibles a los outliers. Utilizar técnicas univariadas (box plot), bivariadas (graficos de dispersión) o mutivariadas (**distancia de Mahalanobis**) para detectarlos

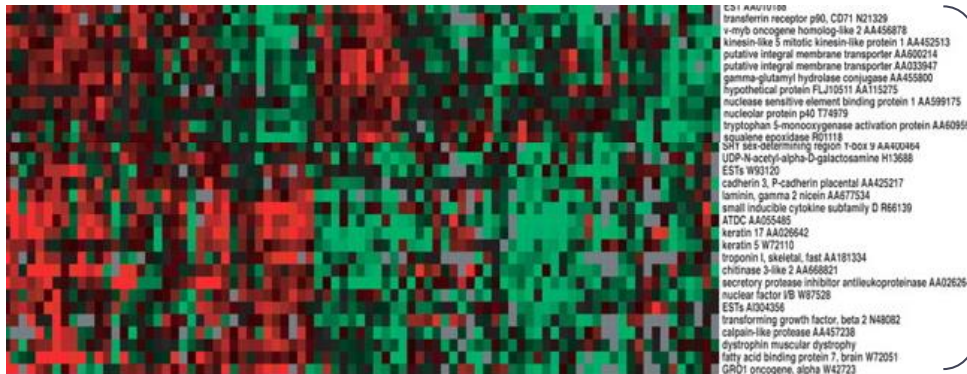
```
s<-cov(data) #matriz de covarianzas
centroide<-colMeans(data)
D2<-mahalanobis(data,centroide,s)
(MV_out<-which(D2>qchisq(.95, df=9))) #p GL
plot(D2)
```

- Ojo con clusters conteniendo sólo uno o dos objetos \Rightarrow outliers? Extraer y repetir el análisis

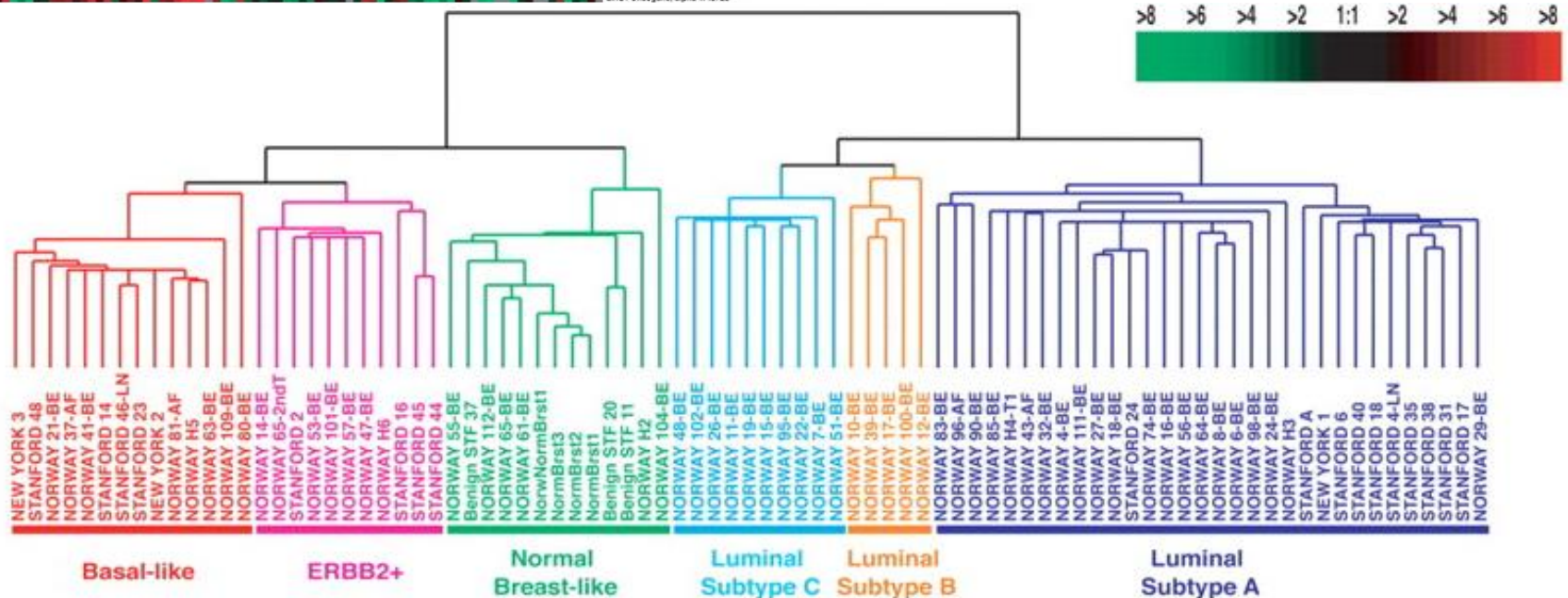
Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications

Sørlie T et al. PNAS 2001;98:10869-10874

85 Muestras de distintos tipos de carcinomas de mama (objetos)



Gene expression patterns of 85 experimental samples representing 78 carcinomas, three benign tumors, and four normal tissues, analyzed by hierarchical clustering using the 476 cDNA intrinsic clone set



Actividad



1. El archivo de Excel *lagunas.natgeo.xls* contiene datos de variables físico-químicas y de la comunidad de picocianobacterias de 52 lagunas bonaerenses.
2. Aplique un análisis de clusters jerárquico sobre las variables físico-químicas e identifique grupos de lagunas
3. Pruebe distintas alternativas de ligamiento/aglomeración
4. Aplique un análisis de clusters no jerárquico

Preparar los datos

```
#omitir los casos con datos faltantes  
data <- na.omit(data)
```

```
# estandarizar las variables  
data <- scale(data)
```