INTRODUCCION AL ANALISIS DE CLUSTER

José Luis Vicente Villardón
Departamento de Estadística
Universidad de Salamanca

DEFINICION E INTRODUCCION

El Análisis de Clusters (o Análisis de conglomerados) es una técnica de Análisis Exploratorio de Datos para resolver problemas de clasificación. Su objeto consiste en ordenar objetos (personas, cosas, animales, plantas, variables, etc, …) en grupos (conglomerados o clusters) de forma que el grado de asociación/similitud entre miembros del mismo cluster sea más fuerte que el grado de asociación/similitud entre miembros de diferentes clusters. Cada cluster se describe como la clase a la que sus miembros pertenecen.

El análisis de cluster es un método que permite descubrir asociaciones y estructuras en los datos que no son evidentes a priori pero que pueden ser útiles una vez que se han encontrado. Los resultados de un Análisis de Clusters pueden contribuir a la definición formal de un esquema de clasificación tal como una taxonomía para un conjunto de objetos, a sugerir modelos estadísticos para describir poblaciones, a asignar nuevos individuos a las clases para diagnóstico e identificación, etc …

Podemos encontrarnos dos tipos fundamentales de métodos de clasificación: **Jerárquicos** y **No Jerárquicos**. En los primeros, la clasificación resultante tiene un número creciente de clases anidadas mientras que en el segundo las clases no son anidadas.

Los métodos pueden dividirse en **aglomerativos** y **divisivos**. En los primeros se parte de tantas clases como objetos tengamos que clasificar y en pasos sucesivos vamos obteniendo clases de objetos similares, mientras que en los segundos se parte de una única clase formada por todos los objetos que se va dividiendo en clases sucesivamente.

Estudiaremos fundamentalmente métodos jerárquicos aglomerativos.

Los pasos que seguiremos para una clasificación jerárquica son

fundamentalmente los siguientes

1.- Decidir que datos tomamos para cada uno de los casos.

Generalmente tomaremos varias variables todas del mismo tipo (continuas, categóricas, etc.) ya que suele ser difícil mezclar tipos distintos..

2.- Elegimos una medida de la distancia entre los objetos a clasificar, que serán los clusters o clases iniciales.

3.- Buscamos que clusters son más similares.

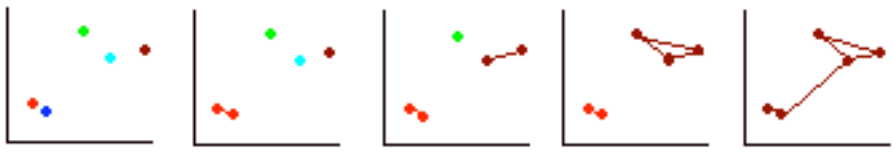
4.- Juntamos estos dos clusters en un nuevo cluster que tenga al menos 2 objetos, de forma que el número de clusters decrece en una unidad.

5.- Calculamos la distancia entre este nuevo cluster y el resto.

No es necesario recalcular todas las distancias, solamente las del nuevo cluster con los anteriores.

6.- Repetimos desde el paso 3 hasta que todos los objetos estén en un único cluster.

Los pasos se resumen en el diagrama siguiente.



Los distintos métodos o algoritmos dependen del método utilizado en el paso 5 para calcular la distancia entre clusters. Es necesario resaltar, que los distintos métodos para el cálculo de las distancias entre clusters producen distintas clasificaciones, por lo que no existe una clasificación correcta única.

LA REPRESENTACION GRÁFICA DE UNA CLASIFICACIÓN JERÁRQUICA: EL DENDOGRAMA

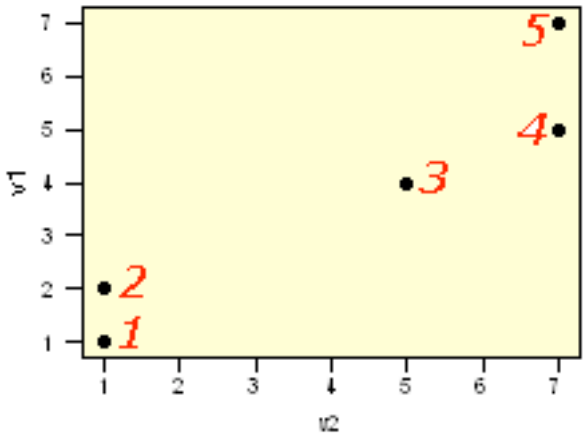
Un dendograma es una representación gráfica en forma de árbol que resume el proceso de agrupación en un análisis de clusters. Los objetos similares se conectan mediante enlaces cuya posición en el diagrama está determinada por el nivel de similitud/disimilitud entre los objetos.

Para entender la construcción de un dendograma y su significado utilizaremos un ejemplo sencillo que lo ilustre.

Consideremos un ejemplo sencillo con solo 5 objetos y dos variables.

objeto	v1	v2
1	1	1
2	2	1
3	4	5
4	7	7
5	5	7

Los puntos representados en el espacio euclídeo bidimensional aparecen en el gráfico siguiente.



A partir de estos datos consideramos la matriz de distancias euclídeas entre los objetos.

	1	2	3	4	5
1	0.0				
2	1.0	0.0			
3	5.0	4.5	0.0		
4	8.5	7.8	3.6	0.0	
5	7.2	6.7	2.2	2.0	0.0

Inicialmente tenemos 5 clusters, uno por cada uno de los objetos a clasificar. De acuerdo con la matriz de distancias, los objetos (clusters) más similares son el 1 y el 2 (con distancia 1), por lo tanto los fusionamos construyendo un nuevo cluster que contiene los objetos 1 y 2. Llamaremos A al cluster.

El problema ahora es medir la distancia de este cluster al resto de los objetos/clusters. Para este ejercicio lo que haremos será tomar como representante del grupo el centroide de los puntos que forman el cluster, es decir, el punto que tiene como coordenadas las medias de los valores de las variables para sus componentes, es decir, las coordenadas de A son A = ((1+2)/2;(1+1)/2)=(1.5; 1).

Tendríamos entonces la siguiente tabla de datos

cluster	v1	v2
A	1.5	1
3	4	5
4	7	7
5	5	7

A partir de estas coordenadas calculamos la nueva matriz de distancias entre los clusters que tenemos en este momento

	A	3	4	5
A	0.0			
3	4.7	0.0		
4	8.1	3.6	0.0	
5	6.9	2.2	2.0	0.0

Ahora los clusters más similares son el 4 y el 5 (con distancia 2) que se fusionan en un nuevo cluster que llamaremos B. El centroide del cluster es el punto (6, 7).

La nueva tabla de datos es

cluster	v1	v2
A	1.5	1
3	4	5
B	6	7

Recalculando como antes la matriz de distancias tenemos.

	A	B	3
A	0.0		
B	7.5	0.0	
3	4.7	2.8	0.0

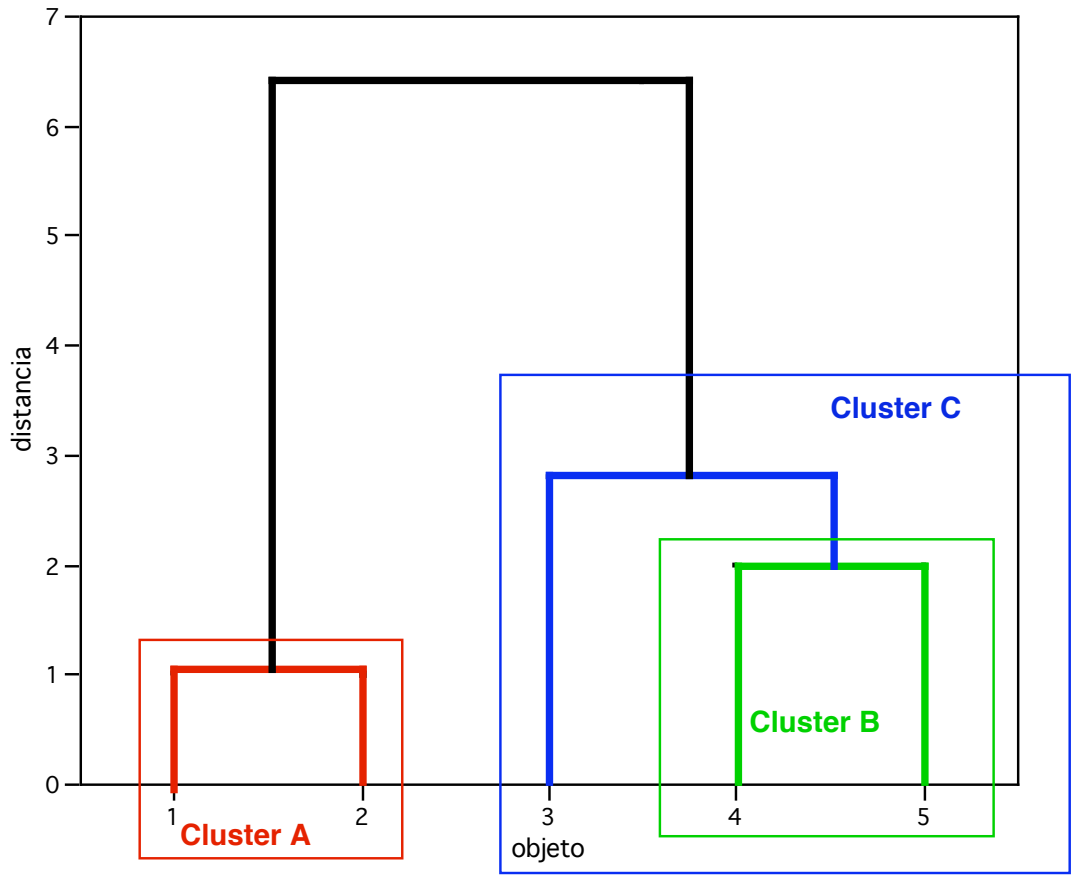
La distancia más pequeña está ahora entre el cluster B y el 3 (distancia 2.8) que se fusionan en uno nuevo que denominamos C. Los valores medios (centroide) son: v1 = (4+5+7)/3 = 5.3, v2 = (5+7+7)/3 = 6.3). La tabla de datos es:

cluster	v1	v2
A	1.5	1
C	5.3	6.3

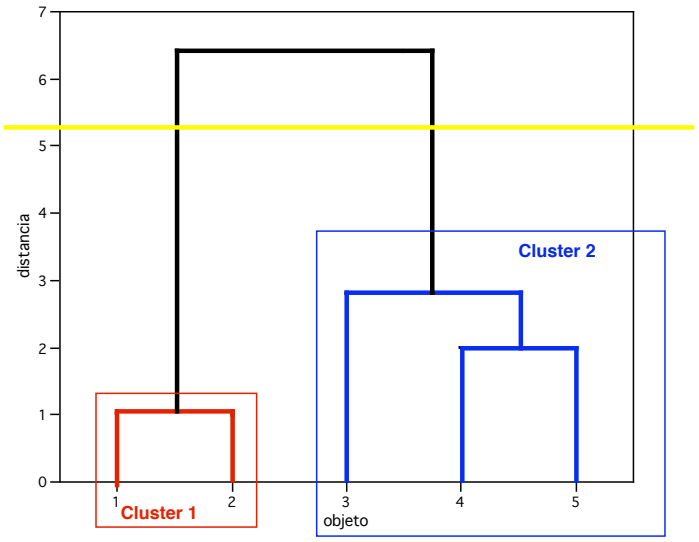
Tenemos ahora solamente dos clusters con distancia 6.4 que se fusionarán en el paso siguiente terminando el proceso.

	A	C
A	0.0	
C	6.4	0.0

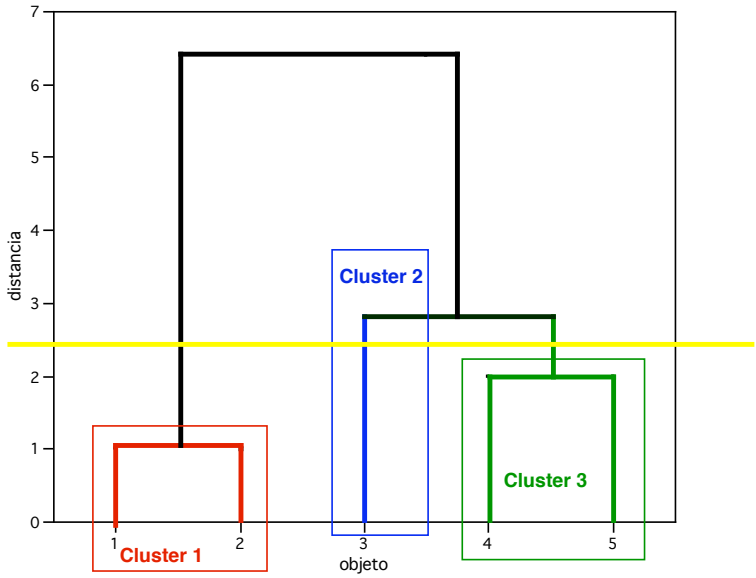
El proceso completo de fusiones puede resumirse mediante un dendograma.



En el gráfico parece evidente que tenemos 2 clusters, el que habíamos denominado A y el que habíamos denominado C. En general, si cortamos el dendograma mediante una línea horizontal como en el gráfico siguiente, determinamos el número de clusters en que dividimos el conjunto de objetos.

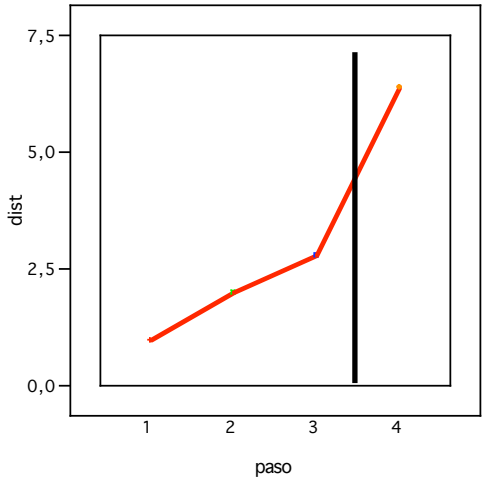


Cortando el dendograma como en la figura anterior obtendríamos 2 clusters. Si lo cortamos como en la figura siguiente obtendríamos 3.



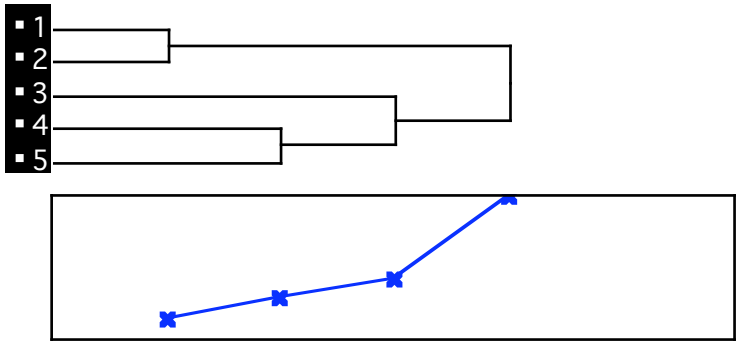
La decisión sobre el número óptimo de clusters es subjetiva, especialmente cuando se incrementa el número de objetos ya que si se seleccionan demasiado pocos, los clusters resultantes son heterogéneos y artificiales, mientras que si se seleccionan demasiados, la interpretación de los mismos suele ser complicada.

Como ayuda a la decisión sobre el número de clusters se suelen representar los distintos pasos del algoritmo y la distancia a la que se produce la fusión. En los primeros pasos el salto en las distancias será pequeño, mientras que en los últimos el salto entre pasos será mayor. El punto de corte será aquel en el que comienzan a producirse saltos bruscos.



En el ejemplo, el salto brusco se produce entre los pasos 3 y 4, luego el punto óptimo es el 3, en el que había 2 clusters.

En algunas ocasiones el dendograma y el gráfico de la evolución de las fusiones.



ALGORITMOS PARA EL ANALISIS DE CLUSTER: DISTINTAS FORMAS DE MEDIR LA DISTANCIA ENTRE CLUSTERS

Como ya indicábamos, existen diversas formas de medir la distancia entre clusters que producen diferentes agrupaciones y diferentes dendogramas. No hay un criterio para seleccionar cual de los algoritmos es el mejor. La decisión es normalmente subjetiva y depende del método que mejor refleje los propósitos de cada estudio particular. Veremos a continuación algunos de los métodos más usuales.

Comenzaremos con una exposición general de los métodos para continuar con expresiones particulares de los mismos.

Si dos objetos o grupos P y Q se han agrupado, la distancia del grupos con otro objeto R puede calcularse como una función de las distancias entre los tres objetos o grupos de la forma siguiente:

$$d(R,P+Q)=\delta_1 d(R,P)+\delta_2 d(R,Q)+\delta_3 d(P,Q)+\delta_4\left|d(R,P)-d(R,Q)\right|$$

donde δ_j son constantes de ponderación. Los distintos métodos dependen de los valores que demos a las ponderaciones. En la tabla siguientes mostramos los pesos para algunos de los métodos más comunes.

Método	δ_1	δ_2	δ_3	δ_4
Salto mínimo	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
Salto máximo	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
Media	$\frac{n_P}{n_P+n_Q}$	$\frac{n_Q}{n_P+n_Q}$	0	0
Centroide	$\frac{n_P}{n_P+n_Q}$	$\frac{n_Q}{n_P+n_Q}$	$-\frac{n_P n_Q}{(n_P+n_Q)^2}$	0
Mediana	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0
Ward	$\frac{n_R+n_P}{n_R+n_P+n_Q}$	$\frac{n_R+n_Q}{n_R+n_P+n_Q}$	$-\frac{n_R}{n_R+n_P+n_Q}$	0
Método Flexible	$\frac{1-\beta}{2}$	$\frac{1-\beta}{2}$	β	0

Donde n_R, n_P, n_Q denotan el número de objetos en cada uno de los grupos y β es un valor arbitrario entre 0 y 1.

MÉTODO DE LA MEDIA (AVERAGE LINKAGE)

En el método de la media, la distancia entre clusters se calcula como la distancia media entre pares de observaciones, una de cada cluster.

$$d(R,P+Q)=\frac{1}{2}d(R,P)+\frac{1}{2}d(R,Q)$$

Para el ejemplo anterior con matriz de distancias

	1	2	3	4	5
1	0.0				
2	1.0	0.0			
3	5.0	4.5	0.0		
4	8.5	7.8	3.6	0.0	
5	7.2	6.7	2.2	2.0	0.0

después de agrupar el 1 y el 2 en el cluster A, calculamos las distancias de A a 3, 4 y 5

	1	2		distan
3	5.0	4.5	(5+4.5)/2	4,75
4	8.5	7.8	(8.5+7.8)/2	8,15
5	7.2	6.7	(7.2+6.7)/2	6,95

la matriz de distancias es entonces

	A	3	4	5
A	0.0			
3	4.75	0.0		
4	8.15	3.6	0.0	
5	6.95	2.2	2.0	0.0

De nuevo, la distancia más pequeña es entre 4 y 5, por lo que los fusionamos en un cluster que denominamos B,

Calculamos la distancia entre B y el resto, es decir, Ay 3.

Entre A y B, buscamos las distancias entre todos los pares de puntos y calculamos la media

		B	
A		4	5
	1	8,5	7,2
	2	7,8	6,7

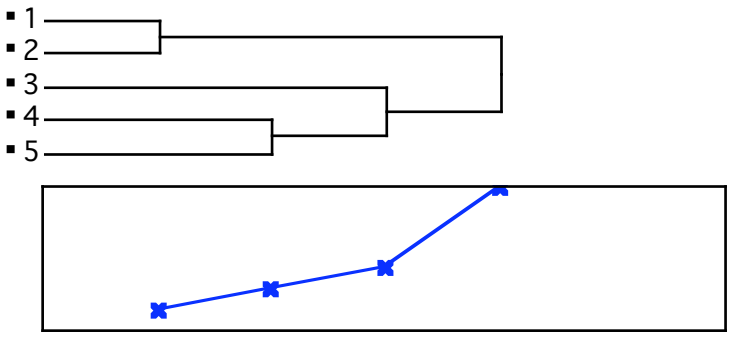
La media de los 4 valores es 7.55

	A	B	3
A	0.0		
B	7.55	0.0	
3	4.75	2.9	0.0

El valor más pequeño es 2.9, luego juntamos B con 3 en C, la distancia es

	A	C
A	0	
C	6,62	0

y el proceso termina. El dendograma obtenido sería muy similar al del ejemplo anterior con la única difiere ligeramente en las distancias a las que se fusionan los clusters.



Obsérvese que en el proceso se han utilizado solamente las distancias, de forma que para este procedimiento no es necesario disponer de los valores originales de las variables, basta con cualquiera de las matrices de distancias que utilizábamos en capítulos anteriores.

CARACTERISTICAS

- Proporciona clusters ni demasiado grandes ni demasiado pequeños.
- Pueden utilizarse medidas de la similitud o de la disimilitud.
- No es invariante por transformaciones monótonas de las distancias.
- Tiende a fusionar clusters con varianzas pequeñas y tiende a proporcionar clusters con la misma varianza.
- Buena representación gráfica de los resultados.

METODO DEL VECINO MÁS PRÓXIMO

En el método del vecino más próximo la distancia entre dos clusters es el mínimo de las distancias entre un objeto de un cluster y un objeto del otro.

$$d(R,P+Q)=\min\big(d(R,P),d(R,Q)\big)$$

Sobre el ejemplo con matriz de distancias

	1	2	3	4	5
1	0.0				
2	1.0	0.0			
3	5.0	4.5	0.0		
4	8.5	7.8	3.6	0.0	
5	7.2	6.7	2.2	2.0	0.0

después de agrupar el 1 y el 2 en el cluster A, calculamos las distancias de A a 3, 4 y 5

	1	2		distan
3	5.0	4.5	min(5; 4.5)	4,5
4	8.5	7.8	min(8.5; 7.8)	7,8
5	7.2	6.7	min(7.2; 6.7)	6,7

la matriz de distancias es entonces

	A	3	4	5
A	0.0			
3	4.5	0.0		
4	7,8	3.6	0.0	
5	6,7	2.2	2.0	0.0

De nuevo, la distancia más pequeña es entre 4 y 5, por lo que los fusionamos en un cluster que denominamos B,

Calculamos la distancia entre B y el resto, es decir, Ay 3.

Entre A y B, buscamos las distancias entre todos los pares de puntos y calculamos el mínimo

		B	
A		4	5
	1	8,5	7,2
	2	7,8	6,7

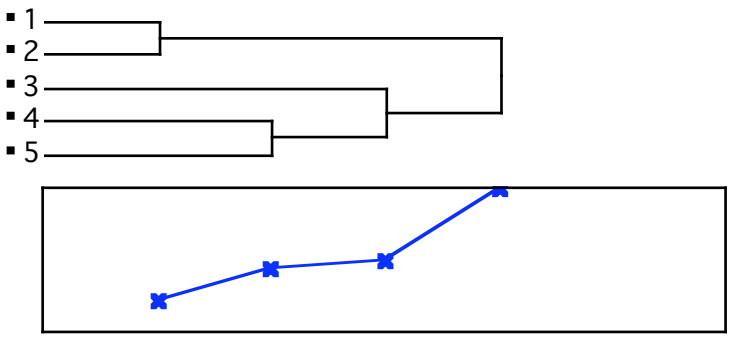
El mínimo de los 4 valores es 7,2. La distancia entre B y 3 es 2,2

	A	B	3
A	0.0		
B	7.2	0.0	
3	4.5	2.2	0.0

El valor más pequeño es 2.2, luego juntamos B con 3 en C, la distancia es

	A	C
A	0	
C	4,5	0

y el proceso termina. El dendograma obtenido sería muy similar al de los ejemplos anteriores con la única difiere ligeramente en las distancias a las que se fusionan los clusters.



Obsérvese que en el proceso se han utilizado solamente las distancias, de forma que para este procedimiento basta con cualquiera de las matrices de distancias que utilizábamos en capítulos anteriores.

CARACTERISTICAS

- No es útil para resumir datos.
- Útil para detectar outliers (estarán entre los últimos en unirse a la jerarquía).
- Pueden usarse medidas de la similitud o de la disimilitud.
- Tiende a construir clusters demasiado grandes y sin sentido.
- Invariante bajo transformaciones monótonas de la matriz de distancias.

METODO DEL VECINO MAS LEJANO (COMPLETE LINKAGE)

En el método del vecino más lejano la distancia entre dos clusters es el máximo de las distancias entre un objeto de un cluster y un objeto del otro.

$$d(R,P+Q)=\max\big(d(R,P),d(R,Q)\big)$$

Sobre el ejemplo con matriz de distancias

	1	2	3	4	5
1	0.0				
2	1.0	0.0			
3	5.0	4.5	0.0		
4	8.5	7.8	3.6	0.0	
5	7.2	6.7	2.2	2.0	0.0

después de agrupar el 1 y el 2 en el cluster A, calculamos las distancias de A a 3, 4 y 5

	1	2		distan
3	5.0	4.5	max(5; 4.5)	5
4	8.5	7.8	max(8.5; 7.8)	8,5
5	7.2	6.7	max(7.2; 6.7)	7,2

la matriz de distancias es entonces

	A	3	4	5
A	0.0			
3	5	0.0		
4	8,5	3.6	0.0	
5	7,2	2.2	2.0	0.0

De nuevo, la distancia más pequeña es entre 4 y 5, por lo que los fusionamos en un cluster que denominamos B,

Calculamos la distancia entre B y el resto, es decir, Ay 3.

Entre A y B, buscamos las distancias entre todos los pares de puntos y calculamos el máximo

		B	
A		4	5
	1	8,5	7,2
	2	7,8	6,7

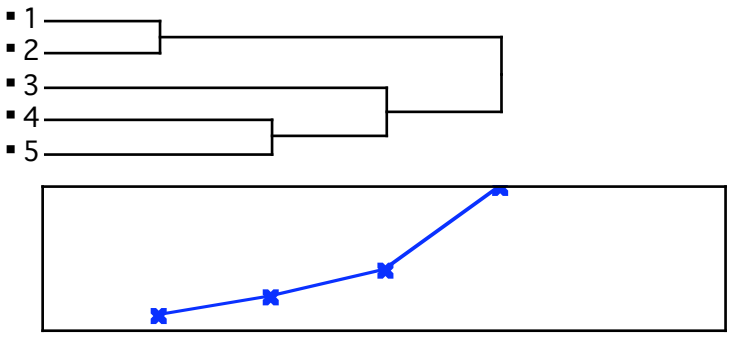
El máximo de los 4 valores es 8,5. La distancia entre B y 3 es 3,6

	A	B	3
A	0.0		
B	8,5	0.0	
3	5	3,6	0.0

El valor más pequeño es 3,6, luego juntamos B con 3 en C, la distancia es

	A	C
A	0	
C	8,5	0

y el proceso termina. El dendograma obtenido sería muy similar al de los ejemplos anteriores con la única difiere ligeramente en las distancias a las que se fusionan los clusters.



CARACTERISTICAS

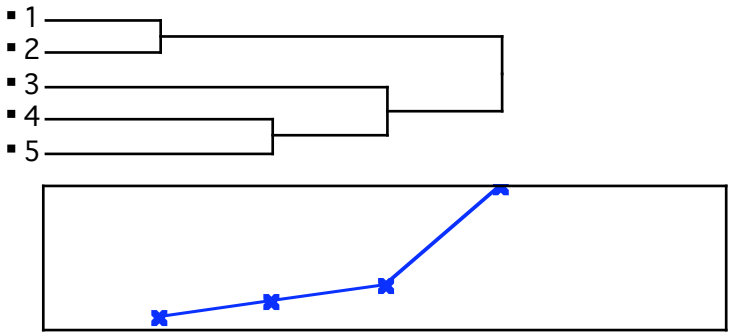
- Útil para detectar outliers.
- Pueden usarse medidas de la similitud o de la disimilitud.
- Tiende a construir clusters pequeños y compactos.
- Invariante bajo transformaciones monótonas de la matriz de distancias.

MÉTODO DE WARD (MÉTODO DE VARIANZA MINIMA)

La distancia entre dos clusters se calcula como la suma de cuadrados entre grupos en el ANOVA sumando para todas las variables. En cada paso se minimiza la suma de cuadrados dentro de los clusters sobre todas la particiones posibles obtenidas fusionando dos clusters del paso anterior. Las sumas de cuadrados son más fáciles de entender cuando se expresan como porcentaje de la suma de cuadrados total.

Los cálculos son más complejos que para los casos anteriores, por lo que no mostraremos ahora todo el proceso.

Para el ejemplo inicial el dendograma obtenido sería el siguiente.



Obsérvese que ahora son necesarios los valores de las variables originales. En la práctica, cuando se dispone solamente de una matriz de distancias, se puede aplicar el método obteniendo primero las coordenadas principales de los objetos o cualquier configuración procedente de cualquiera de los procedimientos de MDS.

CARACTERISTICAS

- El método suele ser muy eficiente.
- Tiende a crear clusters de pequeño tamaño.
- Se puede usar la matriz de distancias así como una tabla de contingencia.
- Invariante bajo transformaciones monótonas de la matriz de distancias.
- Puede ser sensible a los outliers.

MÉTODO DEL CENTROIDE

El método del centroide es el que se utilizó en el ejemplo ilustrativo para la construcción del dendograma. La distancia entre dos clusters se calcula como la distancia entre los centroides de los mismos, por tanto es necesario disponer de los valores originales de las variables.

CARACTERISTICAS

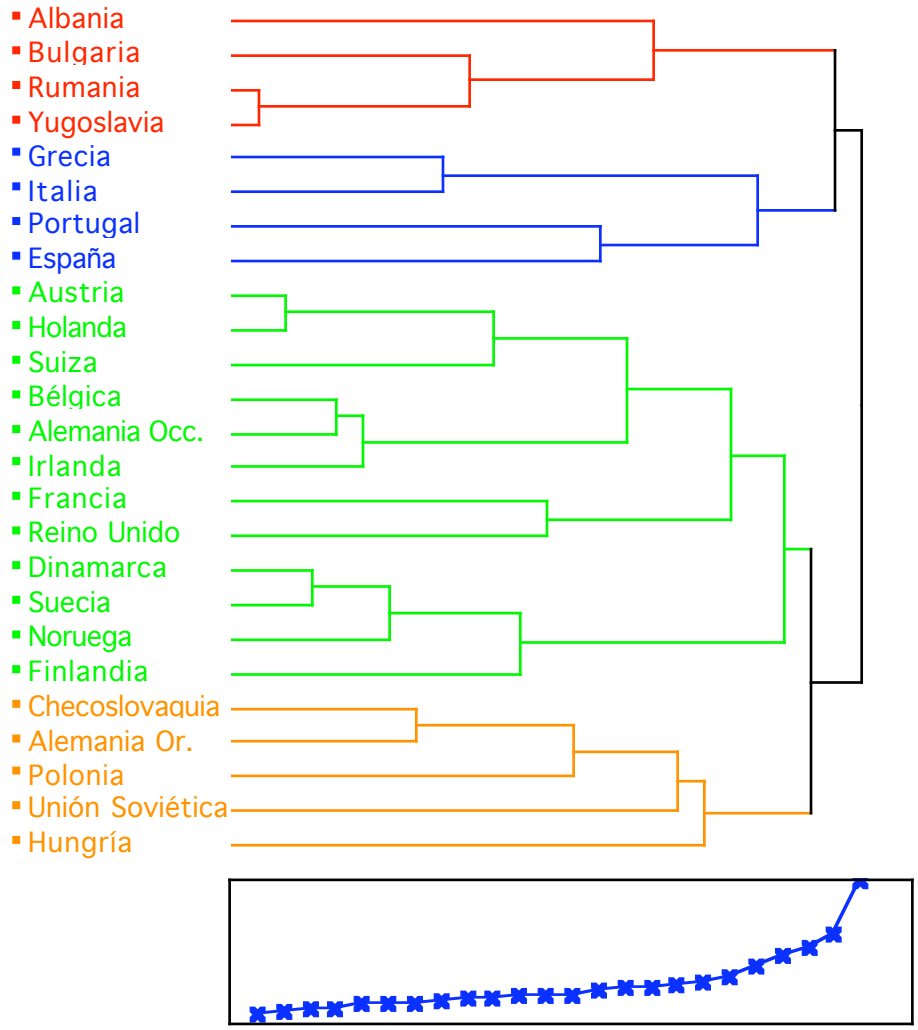
- Las variables deben estar en escala de intervalo.
- Las distancias entre grupos se calculan como las distancias entre los vectores medios.
- Si los tamaños de los dos grupos a mezclar son muy diferentes, entonces el centroide del nuevo grupo será muy próximo al de mayor tamaño y probablente estará dentro de este grupo.

EJEMPLOS

Consumo de proteínas en varios países Europeos.

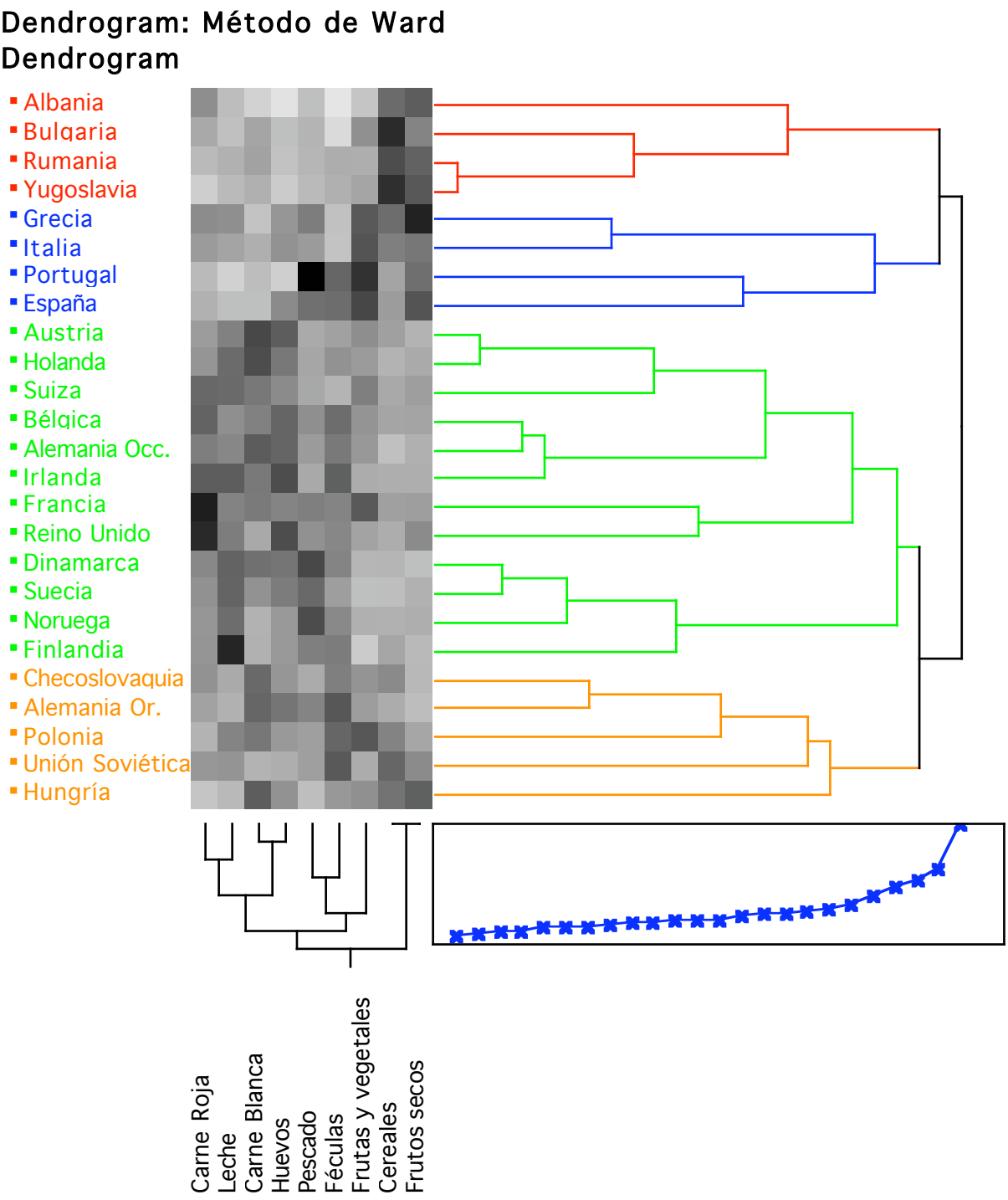
Pais	Carne Roja	Carne Blanca	Huevsos	Leche	Pescado	Cereales	Féculas	Frutos secos	Frutos y vegetales
Albania	10,1	1,4	0,5	8,9	0,2	42,3	0,6	5,5	1,7
Austria	8,9	14	4,3	19,9	2,1	28	3,6	1,3	4,3
Bélgica	13,5	9,3	4,1	17,5	4,5	26,6	5,7	2,1	4
Bulgaria	7,8	6	1,6	8,3	1,2	56,7	1,1	3,7	4,2
Checoslovaquia	9,7	11,4	2,8	12,5	2	34,3	5	1,1	4
Dinamarca	10,6	10,8	3,7	25	9,9	21,9	4,8	0,7	2,4
Alemania Or.	8,4	11,6	3,7	11,1	5,4	24,6	6,5	0,8	3,6
Finlandia	9,5	4,9	2,7	33,7	5,8	26,3	5,1	1	1,4
Francia	18	9,9	3,3	19,5	5,7	28,1	4,8	2,4	6,5
Grecia	10,2	3	2,8	17,6	5,9	41,7	2,2	7,8	6,5
Hungría	5,3	12,4	2,9	9,7	0,3	40,1	4	5,4	4,2
Irlanda	13,9	10	4,7	25,8	2,2	24	6,2	1,6	2,9
Italia	9	5,1	2,9	13,7	3,4	36,8	2,1	4,3	6,7
Holanda	9,5	13,6	3,6	23,4	2,5	22,4	4,2	1,8	3,7
Noruega	9,4	4,7	2,7	23,3	9,7	23	4,6	1,6	2,7
Polonia	6,9	10,2	2,7	19,3	3	36,1	5,9	2	6,6
Portugal	6,2	3,7	1,1	4,9	14,2	27	5,9	4,7	7,9
Rumania	6,2	6,3	1,5	11,1	1	49,6	3,1	5,3	2,8
España	7,1	3,4	3,1	8,6	7	29,2	5,7	5,9	7,2
Suecia	9,9	7,8	3,5	24,7	7,5	19,5	3,7	1,4	2
Suiza	13,1	10,1	3,1	23,8	2,3	25,6	2,8	2,4	4,9
Reino Unido	17,4	5,7	4,7	20,6	4,3	24,3	4,7	3,4	3,3
Unión Sov.	9,3	4,6	2,1	16,6	3	43,6	6,4	3,4	2,9
Alemania Occ.	11,4	12,5	4,1	18,8	3,4	18,6	5,2	1,5	3,8
Yugoslavia	4,4	5	1,2	9,5	0,6	55,9	3	5,7	3,2

Dendrograma : Método de Ward



La partición con 2 clusters agrupa por una parte los países mediterráneos frente al resto de los países europeos, dentro de los dos grupos podemos hacer una subdivisión entre países comunistas y países no comunistas, obteniéndose así una subdivisión en 4 clusters. Si subdividimos los países de Europa central en dos clusters podemos separar los países nórdicos del resto, obteniendo así una subdivisión en 5 clusters. Podemos ver entonces que el consumo de distintos tipos de proteínas y, por tanto, el tipo de alimentación está condicionado por la distribución geográfica y el sistema político-cultural de los países.

Además del agrupamiento de los individuos es posible también realizar una agrupamiento de las variables utilizando, por ejemplo, la correlación entre las mismas. El gráfico siguiente muestra una agrupación de las variables, junto con la de los individuos y además una representación de la matriz completa de datos en la que la magnitud de los valores se representa mediante un código de colores en el que los colores más fuertes representan valores más altos en las variables.



En la agrupación de las variables se muestra como se asocian el consumo de carnes y huevos, que caracterizan a los países centroeuropeos del bloque occidental, cereales y

frutos secos, que caracterizan a los países mediterráneos comunistas, frutas, féculas y pescado, que caracterizan a los países mediterráneos no comunistas. Los países centroeuropeos del antiguo bloque comunista tienen valores menos preponderantes en casi todas las variables, aunque se caracterizan por el consumo de féculas.