Análisis de Conglomerados (Cluster Analysis)

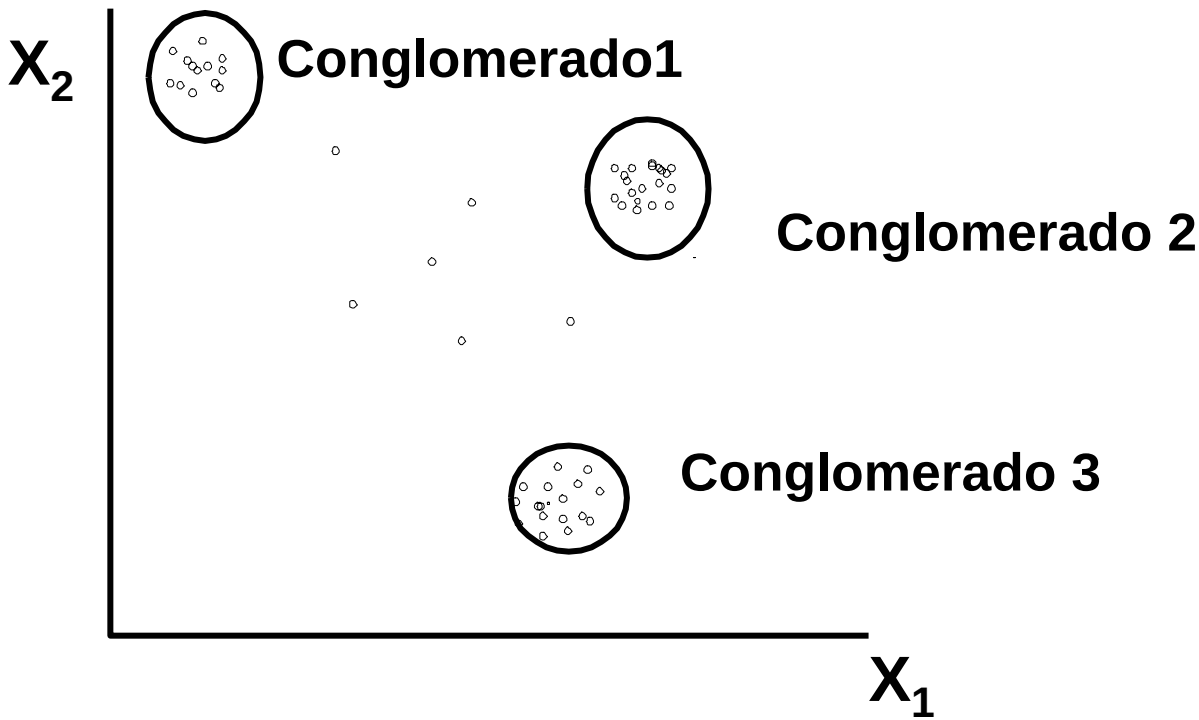
Ana María López

Área de Metodología de las Ciencias del Comportamiento
Departamento de Psicología Experimental

ÍNDICE

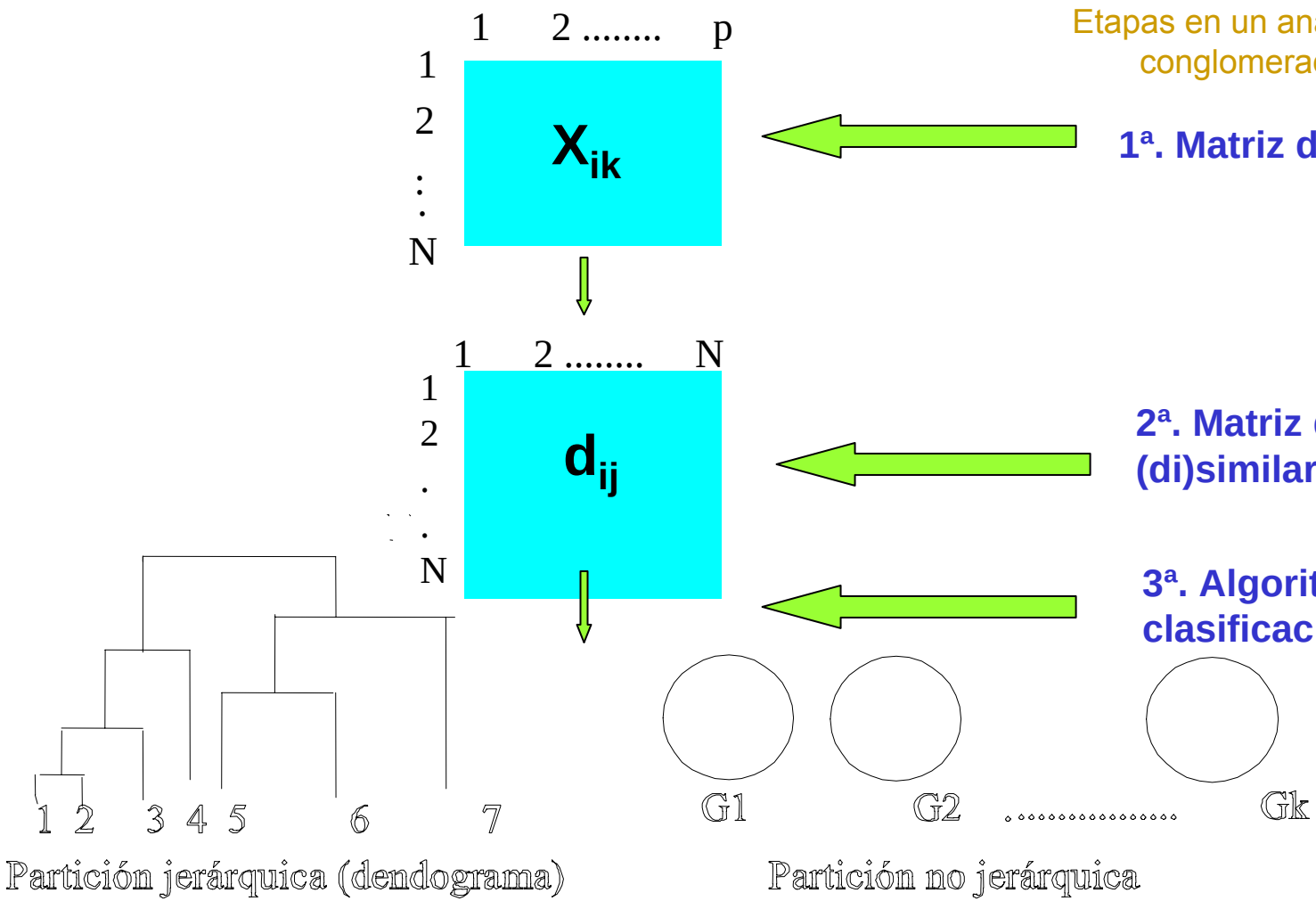
1. Introducción

OBJETIVO: el análisis de conglomerados (cluster analysis, taxonomía numérica o procedimientos de clasificación) es un término genérico para una amplia variedad de procedimientos con un objetivo común: la formación de clases de sujetos o de variables similares. El objetivo es identificar grupos de manera que la variabilidad intraclase sea inferior a la variabilidad entreclases. En el gráfico siguiente hemos representado el resultado de una análisis de conglomerado en base a dos variables.



1. Introducción

Dado que el objetivo fundamental de una análisis de conglomerados es realizar una partición de la muestra en grupos similares, el punto de partida es una matriz de similaridades o de distancias entre los sujetos, objetos o variables que queremos agrupar.



Etapas en un análisis de conglomerados

1ª. Matriz de datos

2ª. Matriz de (di)similaridades

3ª. Algoritmo de clasificación

Partición jerárquica (dendograma)

Partición no jerárquica

2. Estimación Cuantitativa de la (Di)Similaridad entre Objetos

2.1.1. Distancia Euclídea

La Distancia euclídea (d_{ij}): mide el parecido entre unidades de análisis que han sido evaluadas en un conjunto de variables métricas (cuantitativas). La distancia euclídea para dos sujetos viene dada por:

$$d_{ij} = \sqrt{\sum \left(X_{ik} - X_{jk} \right)^2}$$

Ejemplo 1: Calcular el parecido entre tres alumnos/as de psicología a partir de sus notas en la asignaturas de: Análisis de Datos I, Análisis de Datos II, Psicología experimental y Metodología Observacional utilizando la distancia euclídea.

La matriz de datos para los tres sujetos es:

2.1.1. Distancia Euclídea

$$X = \begin{pmatrix} 3 & 5 & 2 & 4 \\ 1 & 0 & 3 & 5 \\ 9 & 10 & 2 & 5 \end{pmatrix}$$

Matriz de datos

$$d_{12} = \sqrt{(3-1)^2 + (5-0)^2 + (2-3)^2 + (4-5)^2} = 5{,}57$$

$$d_{13} = \sqrt{(3-9)^2 + (5-10)^2 + (2-2)^2 + (4-5)^2} = 7{,}75$$

$$d_{23} = \sqrt{(1-9)^2 + (0-10)^2 + (3-2)^2 + (5-5)^2} = 12{,}85$$

Matriz de distancias

	S1	S2	S3
S1	0	5.57	7.75
S2		0	12.85
S3			0

2.1.1. Distancia Euclídea

Un problema de la distancia euclídea, como medida de similaridad, es su dependencia de las diferentes escalas en que estén medidas las variables. Escalas y rangos de variación diferentes pueden afectar al análisis de conglomerados. Este problema se soluciona si en vez de calcular la distancia euclídea con puntuaciones directas se calcula con puntuaciones normalizadas. Estandarizar las puntuaciones de los sujetos en las variables es uno de los procedimientos de normalización más frecuentes en análisis de datos. Un ejemplo aclarará el procedimiento de cálculo de la distancia euclídea y su dependencia de las escalas.

2.1.1. Distancia Euclídea

EJEMPLO 2. Supongamos que estamos interesados en agrupar a una muestra de 5 familias en base al número de hijos, al sueldo en euros al mes y al tamaño de la casa en metros cuadrados. La matriz de datos de la que partimos es:

Hijos	Salario	Metros
1,00	723,00	60,00
1,00	900,00	60,00
4,00	800,00	80,00
,00	1205,00	50,00
2,00	600,00	65,00

Podemos como antes calcular las distancias entre los sujetos a partir de las puntuaciones directas o bien podemos calcularlas a Partir de las variables estandarizadas. Para calcular las puntuaciones típicas utilizamos la expresión:

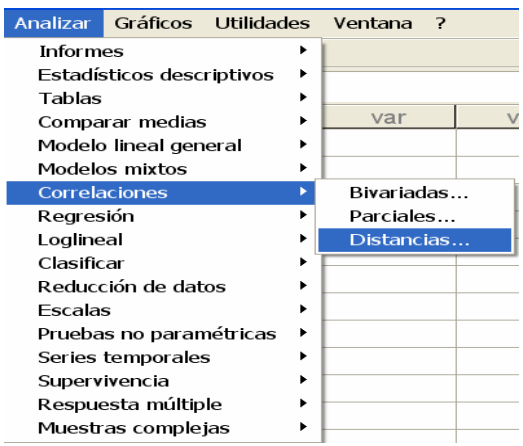
$$Z_{ij} = \frac{X_{ij} - \overline{X}_j}{S_j}$$

Con SPSS directamente en el cuadro de diálogo de distancias o de análisis de conglomerados podemos optar por la matriz de distancias en puntuaciones directas o en puntuaciones estandarizadas.

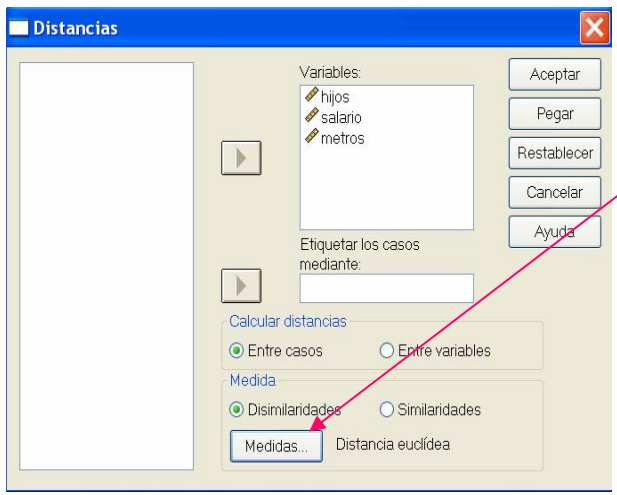
2.1.1. Distancia Euclídea

Los siguientes cuadros muestran cómo obtener la matriz de distancias en puntuaciones estandarizadas.

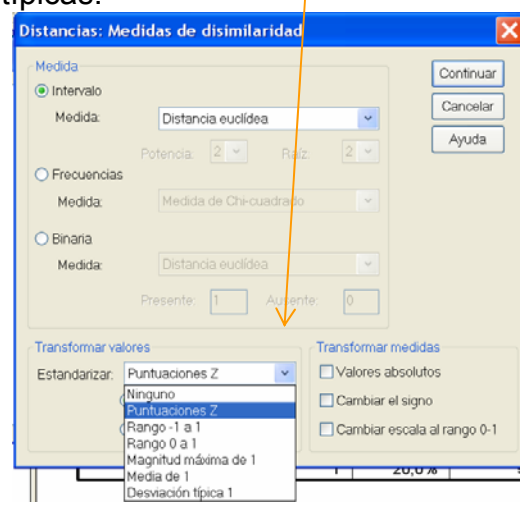
1º.



2º. Insertamos en el cuadro Variables todas las que vamos a utilizar en el cálculo de la distancia



3º. Pulsamos en el botón Medidas y a continuación en la flecha del cuadro Estandarizar. Aparecen diferentes opciones de normalización. La primera es la correspondiente a Puntuaciones típicas.



2.1.1. Distancia Euclídea

Si hubiéramos utilizado la opción de no estandarizar las distancias entre las familias serían:

Matriz de distancias

	Distancia euclídea				
	1	2	3	4	5
1	,000	177,000	79,612	482,105	123,106
2	177,000	,000	102,025	305,166	300,043
3	79,612	102,025	,000	406,129	200,572
4	482,105	305,166	406,129	,000	605,189
5	123,106	300,043	200,572	605,189	,000

Esta es una matriz de disimilaridades

Como puede observarse, las familias más parecidas son la familia primera y la tercera. Sin embargo, son familias que salvo en que tienen un salario similar son diferentes en el resto de las variables. Si por el contrario seleccionamos la opción estandarizar la matriz de distancias que obtenemos es:

Matriz de distancias

	Distancia euclídea				
	1	2	3	4	5
1	,000	,773	2,713	2,388	,965
2	,773	,000	2,727	1,745	1,537
3	2,713	2,727	,000	4,194	2,092
4	2,388	1,745	4,194	,000	3,256
5	,965	1,537	2,092	3,256	,000

Esta es una matriz de disimilaridades

Con las puntuaciones estandarizadas las familias más parecidas son la primera y la segunda. Es evidente que los resultados de un análisis de conglomerados son distintos si se parte de matrices de similaridad o distancia que ordenen a los sujetos de manera distinta. Es por ello que en caso de variables medidas en escalas distintas es necesario normalizar.

2.1.2. Distancia Euclídea Binaria

Versión de la distancia euclídea para variables dicotómicas. Este índice, que representaremos por d_{ij}, se calcula a partir de una tabla de frecuencias 2x2 elaborada para cada par de sujetos o variables a clasificar. Como las variables son dicotómicas podemos codificar con 1 (presencia) a una categoría de la variable y con 0 (ausencia) a la otra categoría. Para cada par de sujetos se obtiene la siguiente tabla.

S_i ↷	S_j ->	1	0
1		a	b
0		c	d

En la tabla:
a: es la frecuencia de acuerdos en el valor 1 para el conjunto de variables
d: es la frecuencia de acuerdos en el valor 0 para el conjunto de variables
b: es la frecuencia de desacuerdos. En las variables en las que el sujeto i tiene un 1 el sujeto j tiene un 0.
c: es la frecuencia de desacuerdos. En las variables en las que el sujeto i tiene un 0 el sujeto j tiene un 1.
La distancia euclídea binaria se obtiene de la tabla de frecuencias con la expresión

$$d_{ij} = \sqrt{b + c} = \sqrt{desacuerdos}$$

La distancia así definida es un índice de la disimilaridad entre sujetos. El valor mínimo de 0 ocurre cuando todo son acuerdos y el valor máximo cuando no hay acuerdos en los sujetos comparados. Un ejemplo sencillo aclarará estos conceptos.

2.1.2. Distancia Euclídea Binaria

EJEMPLO 3. Supongamos que queremos agrupar a los sujetos de una muestra (N = 5) en función de su parecido en un conjunto de variables todas ellas dicotómicas: X_1: Estado civil: soltero(1)-casado(0); X_2: Situación laboral: activo(1)-parado(0); X_3: Nivel de estudios: bajo(1)-alto(0); X_4: Creencias religiosa: creyente(1)-no creyente(0); X_5: Tendencia de voto en las últimas elecciones: izquierda(1)-derecha(0). La matriz de datos de la que partimos es

	X_1	X_2	X_3	X_4	X_5
S_1	1	1	0	1	0
S_2	1	1	1	0	0
S_3	0	0	0	1	1
S_4	0	0	0	0	1
S_5	1	0	0	1	0

Matriz de distancias

	S_2	S_3	S_4	S_5
S_1	$\sqrt{2} = 1.4142$	$\sqrt{3} = 1.7321$	$\sqrt{4} = 2$	$\sqrt{1} = 1$
S_2		$\sqrt{5} = 2.2361$	$\sqrt{4} = 2$	$\sqrt{3} = 1.7321$
S_3			$\sqrt{1} = 1$	$\sqrt{2} = 1.4142$
S_4				$\sqrt{3} = 1.7321$

		S_2		S_3		S_4		S_5	
		1	0	1	0	1	0	1	0
Sujeto 1	1	2	1	1	2	0	3	2	1
	0	1	1	1	1	1	1	0	2
Sujeto 2	1			0	3	0	3	1	2
	0			2	0	1	1	1	1
Sujeto 3	1					1	1	1	1
	0					0	3	1	2
Sujeto 4	1							0	1
	0							2	2

Tablas de frecuencias

2.1.2. Distancia Euclídea Binaria

En muchas ocasiones tenemos que agrupar a sujetos evaluados en variables politómicas (más de dos categorías). En estos casos podemos seguir utilizando la distancia euclídea binaria transformando las variables, previamente, en dicotómicas. Por ejemplo, imaginemos que tenemos que clasificar a una muestra de sujetos en función de su estados civil (soltero-1, casado-2, y viudo-3), el sexo (hombre-1 y mujer-0) y sus creencias religiosas (no creyente-1, católico-2, protestante-3 y testigo de Jehová-4) y que partimos de la siguiente matriz de datos:

	E.Civil	Sexo	C.Religiosas
S_1	1	1	1
S_2	2	0	2
S_3	3	0	4
S_4	2	1	1
S_5	1	1	1

2.1.2. Distancia Euclídea Binaria

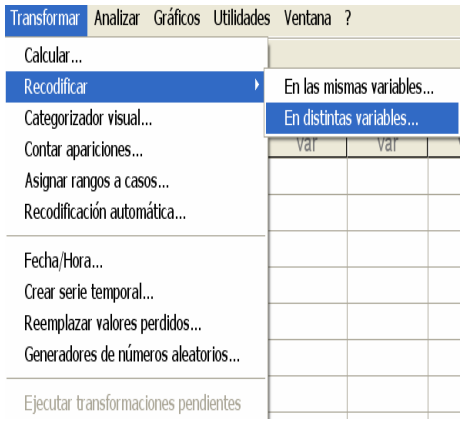
Para poder aplicar la distancia euclídea binaria tendríamos que transformar las variables politómicas de la matriz anterior en variables dicotómicas. Para transformar una variable politómica con k categorías en dicotómicas creamos *k*-1 variables dicotómicas. Para el ejemplo con el que estamos trabajando la nueva matriz sería la siguiente:

	Solt	Casado	Hombre	NoCreye	Catól	Protest
	1	0	1	1	0	0
	0	1	0	0	1	0
	0	0	0	0	0	0
	0	1	1	1	0	0
	1	0	1	1	0	0

2.1.2. Distancia Euclídea Binaria

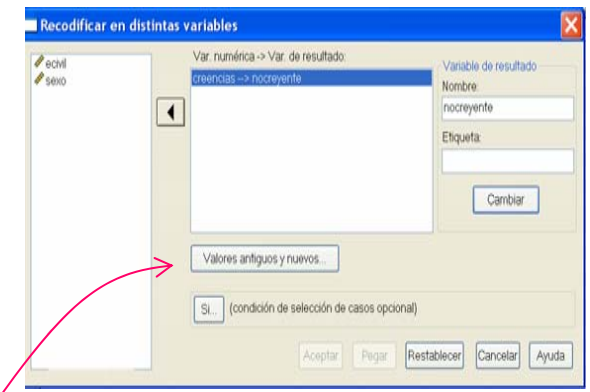
Podemos transformar las variables politómicas en dicotómicas con SPSS. Los pasos a seguir serían:

1º. Seleccionamos Recodificar en distintas variables del menú Transformar

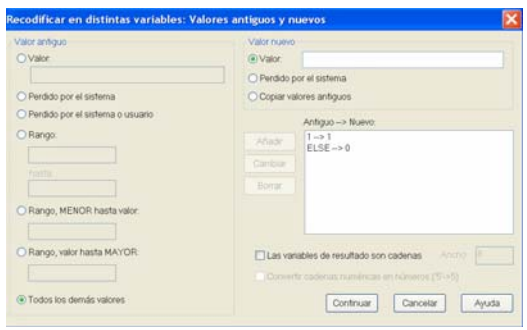


2º. Insertamos en Var.numérica->Var de resultado la variable que queremos transformar: creencias.

3º. Escribimos un Nombre para la primera variable dicotómica: nocreyente porque vamos a convertir en dicotómica la categoría no creyente. A continuación pinchamos en Cambiar.



4º. Pinchamos en Valores antiguos y nuevos y en el botón Valor antiguo escribimos el valor de la variable original que vamos a convertir y en Valor nuevo escribimos 1 y pulsamos en Añadir. A continuación seleccionamos Todos los demás valores y en Valor nuevo escribimos 0 y pulsamos Añadir y Continuar. Con estos pasos hemos creado una variable dicotómica de la categoría no creyente de la variable original. Para crear el resto de variable dicotómicas repetimos estos pasos dos veces.



2.1.3. Distancia para variables categóricas

En los paquetes estadísticos, no obstante, hay implementadas medidas de distancia para variables categóricas que no requieren de la transformación a dicotómicas. Concretamente, en SPSS podemos seleccionar para medir la distancia entre los sujetos evaluados en un conjuntos de variables categóricas las distancias Chi-Cuadrado o la distancia Phi-Cuadrado. Ambas se basan en el modelo de independencia. La distancia Phi-Cuadrado se obtiene dividiendo la Chi-cuadrado entre la raíz cuadrada del total de observaciones.

Además de las medidas de distancia se pueden utilizar multitud de medidas de similaridad. Para variables cuantitativas se puede utilizar el coeficiente de correlación de Pearson y para variables cualitativas existen más de 100 índices derivados de la tabla de contingencia que hemos descrito anteriormente para la distancia euclídea binaria. Muchos de ellos están implementados en SPSS pero dado que las medidas de distancia descritas son las más utilizadas en análisis de conglomerados no describiremos en este documento el resto de los índices.

3. Algoritmos de Clasificación

Una vez que se ha obtenido la matriz de (di)similaridades, el paso siguiente es aplicar una regla que nos permita agrupar a los sujetos o variables similares. Al conjunto de tales reglas se les denomina algoritmos de clasificación.

De dos tipos han sido, fundamentalmente, los algoritmos de clasificación propuestos:

3.1. Algoritmos de Clasificación Jerárquicos.

3.2. Algoritmos de Clasificación No Jerárquicos o basados en la partición en grupos disjuntos de los elementos a clasificar

.

3.1. Algoritmos de Clasificación Jerárquicos

Los algoritmos de clasificación jerárquicos pueden ser de 2 tipos:

Aglomerativos o ascendentes y divisivos o descendentes. Los algoritmos aglomerativos, mediante la utilización de algún criterio, van agrupando unidades de análisis en cada paso hasta llegar a un conglomerado que engloba a la totalidad. En los algoritmos divisivos se parte del conjunto total de elementos considerados como un conglomerado y, según algún criterio, se procede a dividir el grupo en grupos más pequeños llegando, en la última etapa del procedimiento, a considerar a cada elemento del grupo inicial como el conglomerado más simple y de máxima homogeneidad.

En lo que sigue vamos a describir como funcionan los algoritmos de clasificación jerárquicos ascendentes.

3.1.1. Algoritmos de Clasificación Jerárquicos ascendentes

La manera de proceder de los algoritmos jerárquicos ascendentes es como sigue:

1º. La primera etapa de cualquier algoritmo de clasificación ascendente trataría al conjunto de *N* elementos como una primera partición (C_0) en conglomerados de máxima homogeneidad o parecido. En esta primera etapa hay tanto conglomerados como sujetos a agrupar.

2º. En la segunda etapa se agrupan las clases (conglomerados) de la primera partición (C_0) que estén más próximas (o que sean más parecidas) según la medida de similaridad o dismilaridad que se haya definido entre los *N* elementos. La segunda partición (C_1) contiene un conglomerado menos que la primera.

Se recalculan las distancias entre la nueva clase y el resto y nos encontramos en la misma situación que en la etapa 1 pero con N - 1 conglomerados.

El proceso se repite hasta que en la última partición (C_{N-1}) se obtiene un sólo grupo que contiene a todos los elementos. El resultado de este proceso es lo que se llama una jerarquía indexada y su representación gráfica es el dendograma. En el dendograma, o árbol de la jerarquía, se representan en el eje positivo de abscisas los elementos a agrupar y en el eje positivo de ordenadas las distancias correspondientes a los diferentes niveles de agregación denominados también índices de partición o coeficientes de agregación.

3.1.1. Algoritmos de Clasificación Jerárquicos ascendentes

Para entender como proceden este tipo de algoritmos vamos a aplicarlos sobre las matrices de distancias que se dan a continuación. Este ejercicio sólo tiene utilidad didáctica.

Las matrices son distancias entre 5 sujetos, denominados A, B, C, D y E, medidos en un conjunto de variables.

$$Matriz1$$

$$\begin{pmatrix} A & B & C & D & E \\ 0 & 0.1 & 0.3 & 0.4 & 0.5 \\ & 0 & 0.3 & 0.4 & 0.5 \\ & & 0 & 0.4 & 0.5 \\ & & & 0 & 0.5 \\ & & & & 0 \end{pmatrix}$$

$$Matriz2$$

$$\begin{pmatrix} 0 & 2 & 2 & 3 & 6 \\ & 0 & 3 & 4 & 5 \\ & & 0 & 8 & 9 \\ & & & 0 & 7 \\ & & & & 0 \end{pmatrix}$$

3.1.1. Algoritmos de Clasificación Jerárquicos ascendentes

Aplicación de una algoritmo de clasificación ascendente sobre la Matriz 1.

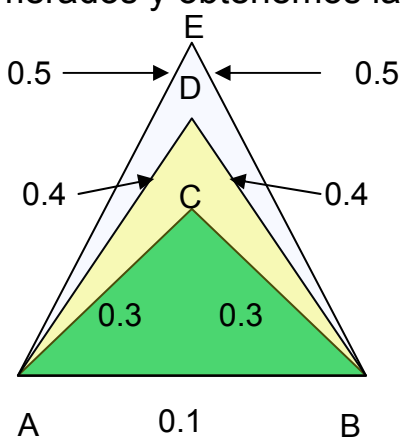
Etapa 0: Se inicia el proceso con la partición:

C_0: (A), (B), (C), (D), (E) Índice de la partición: d(i)=0

Etapa 1: Agrupamos los dos sujetos más parecidos, menor distancia, formando con ellos un conglomerado. En este caso agregamos a los sujetos A y B y obtenemos la partición:

C_1: (A,B), (C), (D), (E) Índice de la partición: d(i)=0.1

En esta etapa recalculamos las distancias que afectan al nuevo conglomerado con el resto es decir: d((A,B), (C)), d((A,B), (D)), d((A,B), (E)) comparando las distancias entre los sujetos que componen los dos conglomerados y obtenemos la siguiente matriz de distancias.



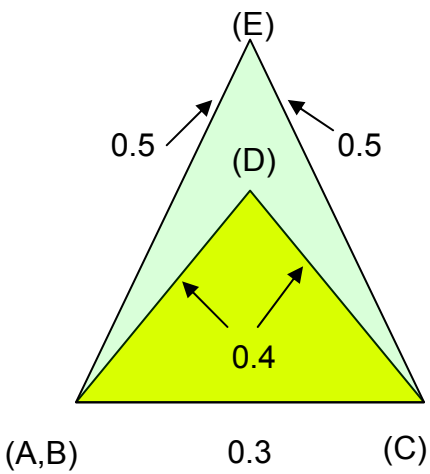
	(A,B)	(C)	(D)	(E)
(A,B)	0	0.3	0.4	0.5
(C)		0	0.4	0.5
(D)			0	0.5

3.1.1. Algoritmos de Clasificación Jerárquicos ascendentes

Etapa 2: Agrupamos los conglomerados más parecidos buscando en la nueva matriz la distancia más pequeña que en este caso corresponde a los conglomerados (A,B) y (C) y obtenemos la siguiente partición:

C_2: (A,B,C), (D), (E) Índice de la partición: d(i)=0.3

Recalculamos, de nuevo, las distancias que afectan al nuevo conglomerado con el resto es decir: d((A,B,C)), (D)), d((A,B,C),(E)) comparando las distancias entre los sujetos que componen los dos conglomerados y obtenemos la siguiente matriz de distancias.



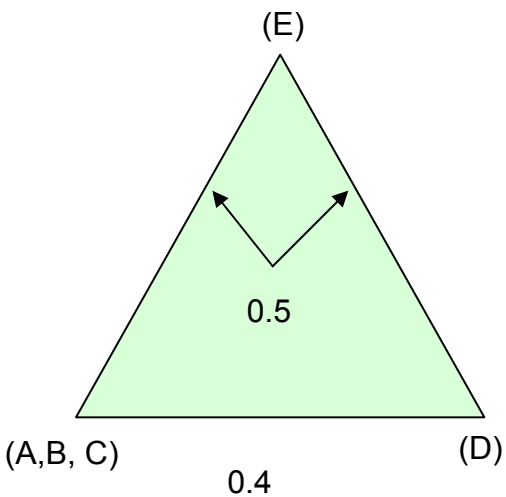
	(A,B,C)	(D)	(E)
(A,B,C)	0	0.4	0.5
(D)		0	0.5

3.1.1. Algoritmos de Clasificación Jerárquicos ascendentes

Etapa 3: Agrupamos los conglomerados más parecidos buscando en la nueva matriz la distancia más pequeña que en este caso corresponde a los conglomerados (A,B,C) y (D). El resultado de la agregación es ahora:

C_3: (A,B,C, D), (E) Índice de la partición: d(i)=0.4

Recalculamos las distancias que afectan al nuevo conglomerado con el resto es decir: d((A,B,C, D),(E)) y obtenemos la siguiente matriz de distancias.



	(A,B,C,D)	(E)
(A,B,C,D)	0	0.5

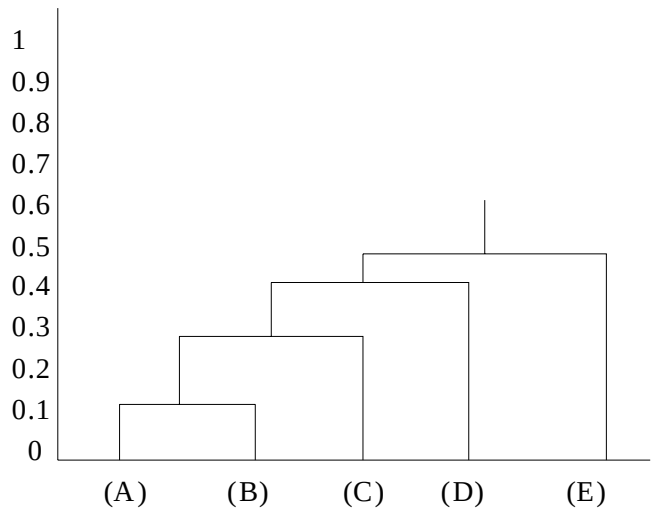
3.1.1. Algoritmos de Clasificación Jerárquicos ascendentes

C_4: (A,B,C, D, E)

Índice de la partición: d(i)=0.5

	(A,B,C,D,E)
(A,B,C,D,E)	0

La representación gráfica del proceso de agregación seguido se denomina dendograma o árbol de la Jerarquía y corresponde a la figura siguiente:



3.1.1. Algoritmos de Clasificación Jerárquicos ascendentes

En cada etapa del procedimiento seguido para formar los grupos hemos recalculado la matriz de distancias entre clases. La distancia de cada elemento del nuevo conglomerado al resto era la misma, cuando esto ocurre para las sucesivas particiones se dice que sobre los elementos a agrupar hay definida una distancia ultramétrica, propiedad necesaria para obtener una clasificación jerárquica aglomerativa. Cuando en la matriz de distancias no se cumpla la propiedad ultramétrica, que por otro lado es lo habitual, hay que transformar la matriz en una matriz de distancias ultramétricas.

Los diferentes procedimientos que permiten transformar una distancia cualquiera en ultramétrica se conocen como *métodos de agregación*. De los diferentes métodos propuestos los más utilizados e implementados en los paquetes estadísticos de uso más frecuentes son:

a) Método de la distancia mínima (vecino más próximo en SPSS)

b) Método de la distancia máxima (vecino más lejano en SPSS)

c) Método de la distancia media (vinculación intergrupos en SPSS)

Describiremos estos tres métodos utilizando como ejemplo la matriz 2.

3.1.1. Algoritmos de Clasificación Jerárquicos ascendentes

El método de la distancia mínima asigna como distancia entre conglomerados el valor mínimo de las distancias entre todos los elementos que componen los conglomerados comparados. Sea (i,j) un conglomerado formado en una determinada etapa del algoritmo de clasificación. Supongamos que al recalcular la distancia entre (i,j) y (k) nos encontramos que:

$$d(i,k) \neq d(j,k)$$

el método de la distancia mínima asigna como distancia entre los conglomerados (i,j) y (k) el valor más pequeño de la siguientes distancias:

$$d((i,j),(k))=\min\{d(i,k), (j,k)\}$$

Aplicando este método vamos a agregar a los sujetos cuyas distancias son las de la matriz 2

3.1.1.1.Método de la Distancia Mínima

Etapa 0: Se inicia el proceso con la partición:

C0: (1), (2), (3), (4), (5) Índice de la partición: d(i)=0

Etapa 1: Agrupamos los dos sujetos más parecidos, menor distancia, formando con ellos un conglomerado. En este caso agregamos a los sujetos 1 y 2 y obtenemos la partición:

C1: (1,2), (3), (4), (5) Índice de la partición: d(i)=2

Antes de seguir agregando, recalculamos las distancias que afectan al nuevo conglomerado. Es decir calculamos: d((1,2), (3)), d((1,2), (4)), d((1,2), (5)) comparando las distancias entre los sujetos que componen los dos conglomerados y obtenemos la siguiente matriz de distancias. Aplicando el método del mínimo las distancia entre los conglomerados anteriores

d((1,2),(3))) = min {d(1,3), d(2,3)} = 2,

d((1,2),(4))) = min {d(1,4), d(2,4)} = 3

d((1,2),(5))) = min {d(1,5), d(2,5)} = 5

y la nueva matriz de distancias sería:

	(1,2)	(3)	(4)	(5)
(1,2)	0	2	3	5
(3)		0	8	9
(4)			0	7
(5)				0

3.1.1.1.Método de la Distancia Mínima

Etapa 2: Agrupamos los conglomerados más próximos buscando el índice más pequeño en la nueva matriz. Este índice corresponde a la distancia entre los conglomerados (1,2) y (3). Agrupamos estos conglomerados y obtenemos la partición

C2 : (1,2,3), (4), (5) Índice de agregación es d((1,2,3)) = 2.

Volvemos a calcular las distancias entre los conglomerados d((1,2,3), (4)) y d((1,2,3), (5)) aplicando el método del mínimo:

d((1,2,3),(4)) = min [d((1,2), (4)), d((3), (4))] = 3;

d((1,2,3),(5)) = min [d((1,2), (5)), d((3), (5))] = 5

La matriz de distancias recalculada es:

	(1,2,3)	**(4)**	**(5)**
(1,2,3)	0	3	5
(4)		0	7
(5)			0

3.1.1.1.Método de la Distancia Mínima

ETAPA 3: Agregamos los conglomerados más próximos tomando como referencia la última matriz de distancias calculada. Los conglomerados más próximos son ahora: (1,2,3) y (4) de manera que la nueva partición es

C3 : (1,2,3,4), (5)

el índice de la partición es d((1,2,3,4)) = 3. La matriz de distancia recalculadas es:

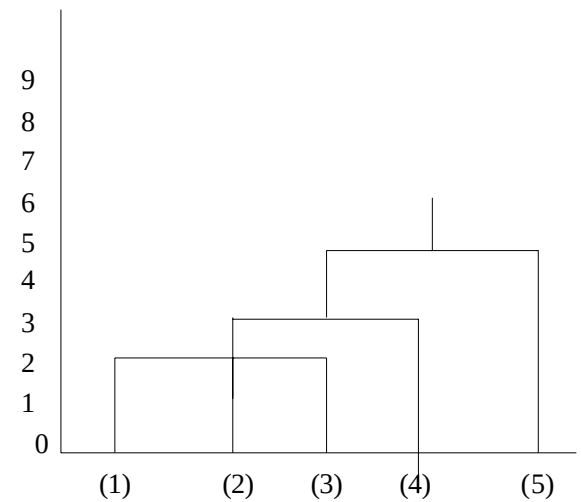
	(1,2,3,4)	(5)
(1,2,3,4)	0	3

ETAPA 4: El último paso consiste en agregar los conglomerados (1,2,3,4) y (5) y obtenemos la partición:

C4 : (1,2,3,4,5)

el índice de esta partición es d((1,2,3,4,5)) = 4 y la nueva matriz de distancia es

	(1,2,3,4,5)
(1,2,3,4,5)	0



3.1.1.2.Método de la Distancia Máxima

El método de la distancia máxima asigna como distancia entre conglomerado el valor máximo de las distancias entre todos los elementos que componen los conglomerados comparados. Sea (i,j) un conglomerado formado en una determinada etapa del algoritmo de clasificación. Supongamos que al recalcular la distancia entre (i,j) y (k) nos encontramos que:

$$d(i,k) \neq d(j,k)$$

el método de la distancia máxima asigna como distancia entre los conglomerados (i,j) y (k) el valor más grande de la siguientes distancias:

$$d((i,j),(k))=\text{máx}\ \{d(i,k),\ (j,k))\}$$

Aplicando este método vamos a agregar a los sujetos cuyas distancias son las de la matriz 2

3.1.1.2.Método de la Distancia Máxima

Etapa 0: Se inicia el proceso con la partición:

C0: (1), (2), (3), (4), (5) Índice de la partición: d(i)=0

Etapa 1: Agrupamos los dos sujetos más parecidos, menor distancia, formando con ellos un conglomerado.
En este caso agregamos a los sujetos 1 y 2 y obtenemos la partición:

C1: (1,2), (3), (4), (5) Índice de la partición: d(i)=2

Antes de seguir agregando, recalculamos las distancias que afectan al nuevo conglomerado. Es decir calculamos: d((1,2), (3)), d((1,2), (4)), d((1,2), (5)) comparando las distancias entre los sujetos que componen los dos conglomerados y obtenemos la siguiente matriz de distancias.

Aplicando el método del mínimo las distancia entre los conglomerados anteriores

d((1,2),(3))) = máx {d(1,3), d(2,3)} = 3,

d((1,2),(4))) = máx {d(1,4), d(2,4)} = 4

d((1,2),(5))) = máx {d(1,5), d(2,5)} = 6

y la nueva matriz de distancias sería:

	(1,2)	**(3)**	**(4)**	**(5)**
(1,2)	0	3	4	6
(3)		0	8	9
(4)			0	7
(5)				0

3.1.1.2.Método de la Distancia Máxima

Etapa 2: Agrupamos los conglomerados más próximos buscando el índice más pequeño en la nueva matriz. Este índice corresponde a la distancia entre los conglomerados (1,2) y (3). Agrupamos estos conglomerados y obtenemos la partición

C2 : (1,2,3), (4), (5) Índice de agregación es d((1,2,3)) = 3.

Volvemos a calcular las distancias entre los conglomerados d((1,2,3), (4)) y d((1,2,3), (5)) aplicando el método del mínimo:

d((1,2,3),(4)) = máx [d((1,2), (4)), d((3), (4))] = 8;

d((1,2,3),(5)) = máx [d((1,2), (5)), d((3), (5))] = 9

La matriz de distancias recalculada es:

	(1,2,3)	**(4)**	**(5)**
(1,2,3)	0	8	9
(4)		0	7
(5)			0

3.1.1.2.Método de la Distancia Máxima

ETAPA 3: Agregamos los conglomerados más próximos tomando como referencia la última matriz de distancias calculada. Los conglomerados más próximos son ahora: (4) y (5) de manera que la nueva partición es

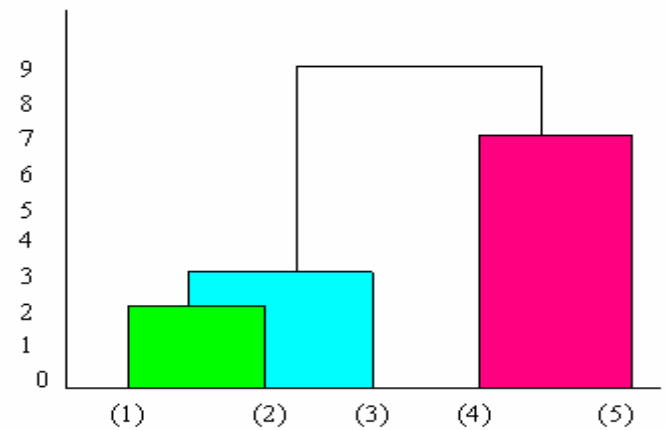
C3 : (1,2,3,), (4,5)

el índice de la partición es d((4,5)) = 7. La matriz de distancia recalculadas es:

	(1,2,3)	**(4,5)**
(1,2,3)	0	9
(4,5)		0

ETAPA 4: El último paso consiste en agregar los conglomerados (1,2,3) y (4,5) y obtenemos la partición: C4 : (1,2,3,4,5) el índice de esta partición es d((1,2,3,4,5)) = 9 y la nueva

	(1,2,3,4,5)
(1,2,3,4,5)	0



3.1.1.2.Método de la Distancia Media

El método de la distancia media asigna como distancia entre conglomerado el valor medio de las distancias entre todos los elementos que componen los conglomerados comparados. Sea (i,j) un conglomerado formado en una determinada etapa del algoritmo de clasificación. Supongamos que al recalcular la distancia entre (i,j) y (k) nos encontramos que:

$$d(i,k) \neq d(j,k)$$

el método de la distancia media asigna como distancia entre los conglomerados (i,j) y (k) el valor

$$d\left[(i,j)(k)\right] = \frac{d(i,k) + d(j,k)}{2}$$

Aplicando este método vamos a agregar a los sujetos cuyas distancias son las de la matriz 2

3.1.1.3.Método de la Distancia Media

Etapa 0: Se inicia el proceso con la partición:

C0: (1), (2), (3), (4), (5) Índice de la partición: d(i)=0

Etapa 1: Agrupamos los dos sujetos más parecidos, menor distancia, formando con ellos un conglomerado. En este caso agregamos a los sujetos 1 y 2 y obtenemos la partición:

C1: (1,2), (3), (4), (5) Índice de la partición: d(i)=2

Antes de seguir agregando, recalculamos las distancias que afectan al nuevo conglomerado. Es decir calculamos: d((1,2), (3)), d((1,2), (4)), d((1,2), (5)) comparando las distancias entre los sujetos que componen los dos conglomerados y obtenemos la siguiente matriz de distancias.

Aplicando el método del mínimo las distancia entre los conglomerados anteriores

$$d\left[(1,2),(3)\right]=\frac{d(1,3)+d(2,3)}{2}=\frac{2+3}{2}=2.5;\ d\left[(1,2),(4)\right]=\frac{d(1,4)+d(2,4)}{2}=\frac{3+4}{2}=3.5$$

$$d\left[(1,2),(5)\right]=\frac{d(1,5)+d(2,5)}{2}=\frac{6+5}{2}=5.5$$

y la nueva matriz de distancias sería:

	(1,2)	**(3)**	**(4)**	**(5)**
(1,2)	0	2.5	3.5	5.5
(3)		0	8	9
(4)			0	7
(5)				0

3.1.1.3.Método de la Distancia Media

Etapa 2: Agrupamos los conglomerados más próximos buscando el índice más pequeño en la nueva matriz. Este índice corresponde a la distancia entre los conglomerados (1,2) y (3). Agrupamos estos conglomerados y obtenemos la partición

C2 : (1,2,3), (4), (5) Índice de agregación es d((1,2,3)) = 2.5.

Volvemos a calcular las distancias entre los conglomerados d((1,2,3), (4)) y d((1,2,3), (5)) aplicando el método de la media:

$$d\left[(1,2,3),(4)\right] = \frac{d(1,4)+d(2,4)+d(3,4)}{3} = \frac{3+4+8}{3} = 5$$

$$d\left[(1,2,3),(5)\right] = \frac{d((1,5)+d(2,5)+d(3,5)}{3} = \frac{6+5+9}{3} = 6.6$$

La matriz de distancias recalculada es:

	(1,2,3)	**(4)**	**(5)**
(123)	0	5	6.6
(4)		0	7
(5)			0

3.1.1.3.Método de la Distancia Media

ETAPA 3: Agregamos los conglomerados más próximos tomando como referencia la última matriz de distancias calculada. Los conglomerados más próximos son ahora: (1,2,3) y (4) de manera que la nueva partición es

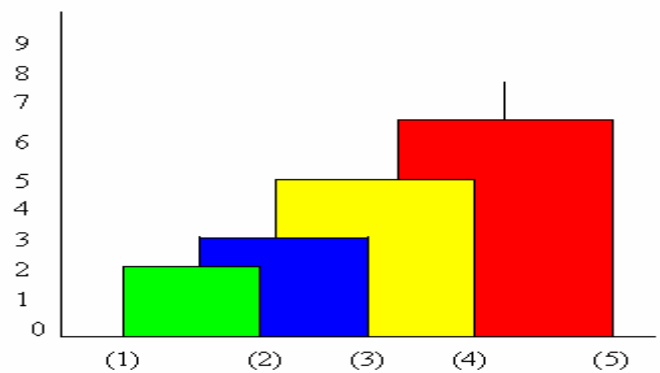
C3 : (1,2,3,4), (5)

el índice de la partición es d((1,2,3,4)) = 5. La matriz de distancia recalculadas es:

	(1,2,3,4)	**(5)**
(1,2,3)	0	6.75
(5)		0

ETAPA 4: El último paso consiste en agregar los conglomerados (1,2,3,4) y (5) y obtenemos la partición: C4 : (1,2,3,4,5) el índice de esta partición es d((1,2,3,4,5)) = 6.75 y la nueva matriz de distancia es

	(1,2,3,4,5)
(1,2,3,4,5)	0



3.1.1.5. Determinación del método y del Número de conglomerados en los Algoritmos Jerárquicos Ascendentes

Los diferentes métodos descritos modifican en mayor o menor medida las distancias iniciales entre sujetos. El método del mínimo aproxima a los sujetos y se denomina *espacio contractivo*, el método del máximo aleja a los sujetos y se denomina *espacio dilatante* y, por último, el método de la media suele conservar las distancias y se denomina *espacio conservativo.*

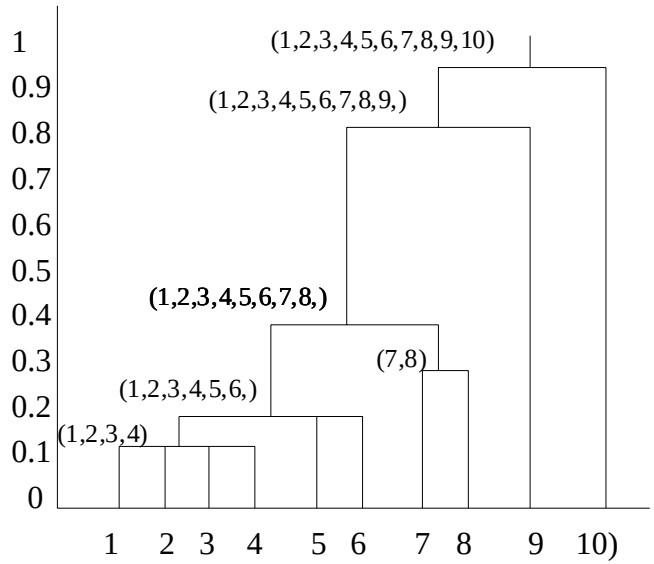
La interpretación del resultado que proporciona un algoritmo ascendente dista mucho de referirse al grupo final sino que se refiere a un número mayor de conglomerados obtenidos en etapas intermedias del proceso de agregación. Si nos centramos en el dendograma la pregunta que cabría hacerse es ¿qué índice de agregación o nivel de corte proporciona una solución en conglomerados óptima?. La respuesta a dicha pregunta, aunque debería constituir una etapa más del análisis de conglomerados, está entre los problemas no resueltos. Las razones para tal deficiencia son por una parte, la ausencia de una hipótesis nula y por otra la compleja naturaleza de las distribuciones muestrales multivariantes.

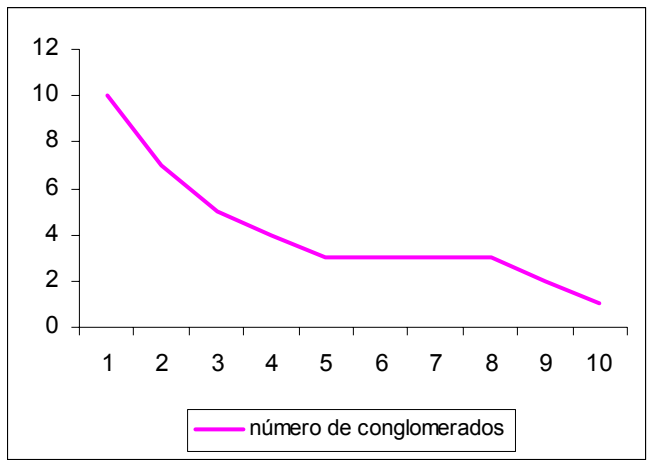
Ante tales dificultades sólo hemos encontrado, en la literatura revisada, procedimientos basados en la inspección visual del dendograma y/o en gráficos derivados de él.

3.1.1.5. Determinación del método y del Número de conglomerados en los Algoritmos Jerárquicos Ascendentes

El procedimiento más simple es decidir el punto de corte por “inspección subjetiva”. Este procedimiento no es muy satisfactorio pues está, generalmente, sesgado por las necesidades y opiniones de los investigadores acerca de la estructura correcta de sus datos. Un procedimiento más formal consiste en utilizar el gráfico de sedimentación (*scree test*) que ya se utiliza en análisis factorial. Para el análisis de conglomerados, este gráfico se construye colocando en el eje de abscisas los valores de los índices de agregación de las distintas etapas del algoritmo de clasificación utilizado y, en el eje de ordenadas, el número de conglomerados correspondiente a cada uno de los índices. Lo esperable es observar un primer tramo con una pendiente negativa grande donde el decaimiento es muy rápido para un intervalo de índices de agregación relativamente pequeño y, un tramo en donde la pendiente es muy pequeña y el decaimiento es muy lento. El número de conglomerados óptimo, que se recomienda como solución final, es el último valor del primer tramo del gráfico de sedimentación.

3.1.1.5. Determinación del método y del Número de conglomerados en los Algoritmos Jerárquicos Ascendentes





De la observación del gráfico de la Figura se deduce que 3 o 4 conglomerados serían una solución óptima.

Otro método sencillo para determinar el número de conglomerados es el denominado *tamaño de paso* que consiste en determinar la mayor diferencia en índices de agregación para dos etapas sucesivas. El número de conglomerados es el de la etapa inmediatamente anterior. Podemos representar la diferencia en índices de agregación en etapas sucesivas y buscar un cambio importante.

Algoritmos de clasificación no jerárquicos: Métodos de partición

de Clasificación No Jerárquicos: Métodos de Parti
s (K-medias)

En este conjunto de métodos, a diferencia de los jerárquicos, se forman *k* grupos siendo *k* un número que el Investigador decide a priori. Para decidir acerca del número de conglomerados se utiliza el conocimiento que se tiene de investigaciones previas o se utiliza un procedimiento de clasificación jerárquica que ayude a identificar el número de clases adecuado. Fijar un número muy pequeño puede llevar a conclusiones pobres, mientras que fijar un número demasiado grande complica la interpretación. Lo ideal es repetir el análisis con distintos valores de *k* y seleccionar el que más satisfaga las expectativas del investigador.

Para llegar a la formación de conglomerados se sigue un proceso iterativo que intenta optimizar una función criterio. Se han propuesto varias funciones y una de las más eficaces consiste en la reasignación de una observación al centro más próximo.

Este tipo de algoritmos comienzan con una selección de tantos sujetos como conglomerados queramos formar. Los sujetos inicialmente seleccionados constituyen los centros de las clases e inducen una primera partición por asignación del resto de los sujetos al centro más próximo.

Las distintas técnicas de partición difieren en cómo se determinan los conglomerados iniciales y como son asignados a las clases. De los distintos algoritmos de clasificación no jerárquicos vamos a describir el método de *k-medias o agregación alrededor de centros móviles* por ser de los más utilizados e implementados en los paquetes estadísticos de uso frecuente en investigación.

3.2. Algoritmos de Clasificación No Jerárquicos: Métodos de Partición. Agregación alrededor de Centros Móviles (K-medias)

Se parte de un conjunto de sujetos a clasificar en base a sus semejanzas en *p* variables. Se decide un número *k* de clases y se procede de la siguiente manera:

Etapa 0: Se determinan *k* centros. La forma de elegir los centros puede ser aleatoria o fija (por ejemplo los *k* primeros sujeto de la matriz de datos).

$$\left\{ I_1^0, I_2^0, \ldots, I_k^0 \right\}$$

estos centros inducen una primera partición del conjunto de sujetos

$$\left\{ C_1^0, C_2^0, \ldots, C_k^0 \right\}$$

3.2. Algoritmos de Clasificación No Jerárquicos: Métodos de Partición. Agregación alrededor de Centros Móviles (K-medias)

De tal manera que un sujeto (i) pertenece a la clase I^0_1 si está más próximo a C_1^0 que al resto de los centros. La proximidad del sujeto (i) a Clase I_1^0 es la distancia euclídea del sujeto al centro de dicha clase.

El centro de la clase -o centroide- es el vector de medias para Las *p* variables medidas en la investigación. Es decir, el centroide de la clase I^0_1 es

$$\left\{ \bar{X}_1^0, \bar{X}_2^0, \ldots, \bar{X}_p^0 \right\}$$

y la distancia del sujeto (i) al centroide viene dada por

$$d = \sqrt{\sum_{j=1}^{p} \left(X_{ij} - \bar{X}_{I,j} \right)^2}$$

3.2. Algoritmos de Clasificación No Jerárquicos: Métodos de Partición. Agregación alrededor de Centros Móviles (K-medias)

Etapa 1: Se determinan *k* nuevos centros

$$\left\{ C_1^1, C_2^1, \ldots, C_k^1 \right\}$$

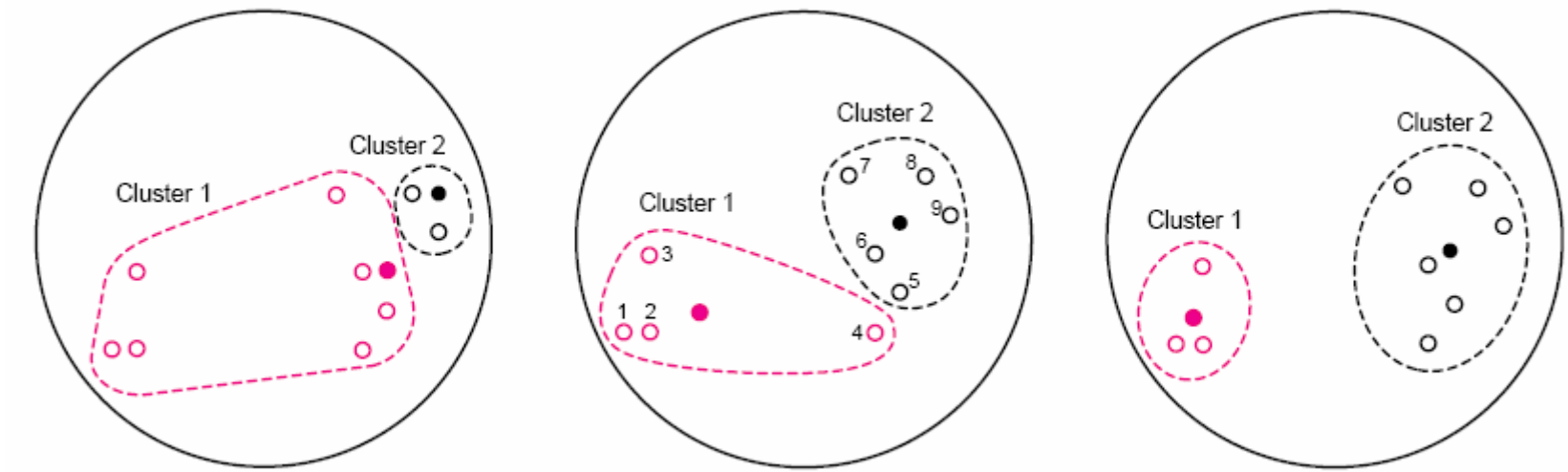
calculando el centroide de las clases de la primera partición. Los nuevos centros inducen una nueva partición

$$\left\{ I_1^1, I_2^1, \ldots, I_k^1 \right\}$$

Esta nueva partición se obtiene recalculando las distancias de cada sujeto a los centros de la partición y si es necesario se reasigna a los sujetos. El criterio para reasignar es que la distancia de un sujeto a otro centro sea menor que la distancia al centro de la clase a la que temporalmente se ha asignado.

3.2. Algoritmos de Clasificación No Jerárquicos: Métodos de Partición. Agregación alrededor de Centros Móviles (K-medias)

Las etapas anteriores se repiten hasta que en dos etapas sucesivas se obtiene la misma partición. Es decir, no hay reasignación de sujetos y por tanto los centroides no cambian. Otro criterio para terminar es fijar, a priori, un número máximo de etapas o un valor mínimo para las distancia entre centros. Gráficamente el algoritmo de k-medias lo hemos representado en la siguiente figura



Inicializa los valores de las semillas para un número k de conglomerados

Se calculan las distancias de cada sujeto a los valores iniciales y se obtiene la primera partición

Se calculan los nuevos centroides y se reasigna a los sujetos

No hay reasignación