6 Clasificación Jerárquica

- En C.A. *jerárquico* con un método **agregativo** se comienza con n clusters, uno para cada observación, y se finaliza con un único cluster que contiene las n observaciones. En cada paso una observación o cluster de ellas es absorbida en otro cluster.
- Los métodos *jerárquicos* **divisivos** comienzan con un cluster con n observaciones y se divide un cluster en cada paso.
- Son mas comunes los métodos agregativos que los divisivos.
- En cada paso de un método jerárquico los dos clusters "mas cercanos" se unen en uno nuevo. Este proceso es irreversible, una vez unidas dos observaciones no se separan durante el resto del procedimiento.

6.1 Clasificación Jerárquica Agregativa

Proceso secuencial, por el cual en cada etapa una observación o grupo de observaciones se une a otro formando uno nuevo.

El proceso comienza con n grupos (cada uno con un único objeto) y termina co un grupo conteniendo todo el conjunto de datos.

En cada etapa los dos grupos mas cercanos se unen en uno nuevo y por esto **debe considerarse la forma de medir distancias entre grupos**.

6.2 Clasificación Jerárquica Divisiva

Trabaja en la dirección opuesta, partiendo de un único grupo inicial de objetos (todo el conjunto de datos) se divide en dos subgrupos en cada etapa. Estos subgrupos se construyen buscando la división en grupos mas disimiles.

El resultado de ambos metodos puede desplegarse en forma de diagrama bidimensional llamado dendograma, que ilustra las divisiones o uniones que se dan en los sucesivos niveles de la jerarquia

6.3 Pasos principales del algoritmo

1. Con una medida de Disimilaridad (distancia) se evalúan los objetos a clasificar.

   ▶ Se crea una tabla $D_{I\times I}$ simétrica cuyo contenido son las distancias dos a dos entre los I objetos a clasificar($\{d_{ij}\}$).

2. En la tabla $D_{I\times I}$ se busca el término mínimo fuera de la diagonal. mín$\{d(y_i, y_j)\}$ con $y_i, y_j \in I$

   ▶ Se forma un nuevo objeto que contiene a los dos objetos que cumplen esto.

3. Se construye una nueva tabla de distancias $D'_{(I-1)\times(I-1)}$

Para calcular D' debe definirse una forma de calculo entre un grupo con dos objetos y los restantes objetos.

6.4 Métodos

- Vecino más Cercano
- Vecino más Lejano
- Centroide
- Ward

6.4.1 single linkage(nearest neighbour)

En este método la distancia entre dos grupos A y B se define como el mínimo entre un punto de A y otro de B

$$D(A, B) = \text{mín}\{d(y_i, y_j), y_i \in A, y_j \in B\}$$

6.4.2 complete linkage(farthest neighbour)

En este método la distancia entre dos grupos A y B se define como la máxima entre un punto en A y otro en B

$$D(A, B) = \text{máx}\{d(y_i, y_j), y_i \in A, y_j \in B\}$$

6.4.3 Ejemplo

Se considera un conjunto de cinco unidades estadísticas: a, b, c, d, e. Sea la siguiente matriz de distancias

	a	b	c	d	e
a	0	2	6	10	9
b		0	5	9	8
c			0	4	5
d				0	3
e					0

Se observa que la mínima distancia es 2, o, sea, las unidades más cercanas son a y b y por lo tanto son las primeras en *unirse*. La nueva matriz de distancias es

	(a,b)	c	d	e
(a,b)	0	5	9	8
c		0	4	5
d			0	3
e				0

En la siguiente fase se unen e y d formándose la siguiente partición :$(a, b)c(d, e)$ y finalmente se une c con (d, e).

	(a,b)	c	(d,e)
(a,b)	0	5	8
c		0	4
(d,e)			0

	(a,b)	(c,d,e)
(a,b)	0	5
(c,d,e)		0

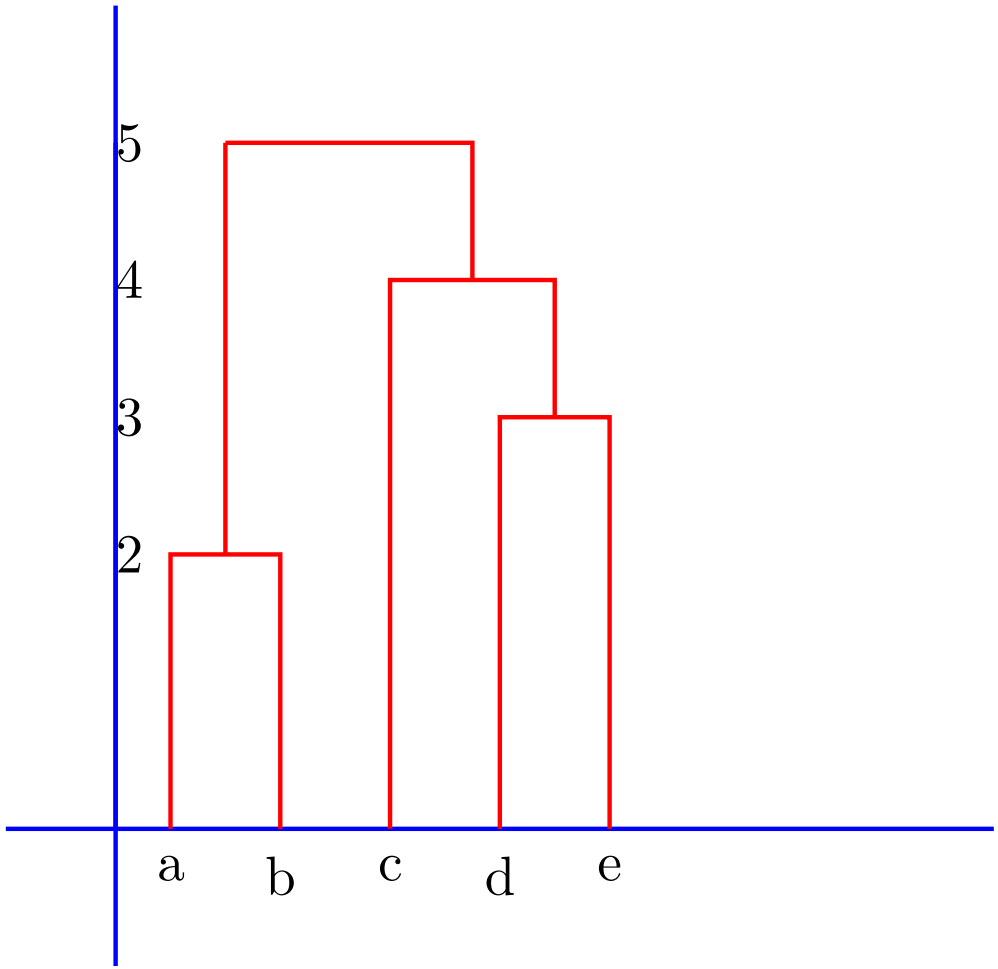


Figura 1: Dendrograma Vecino más Cercano

6.4.4 average linkage

La distancia entre dos grupos A y B se define como el promedio de las $n_A n_B$ distancias entre los n_A puntos de A y n_B puntos de B

$$D(A,B) = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d(y_i, y_j)$$

6.4.5 punto medio

Punto medio de la linea que une A y B

$$m_{AB} = \frac{1}{2}(\bar{y}_A + \bar{y}_B)$$

6.4.6 centroid

La distancia entre dos grupos A y B se define como la distancia euclidea entre los vectores medios de los dos grupos.

$$D(A, B) = d(\bar{y}_A, \bar{y}_B)$$

con $\bar{y}_A = \sum_{i=1}^{n_A} \frac{y_i}{n_A}$

Despues de la unión de los dos grupos, el centroide del nuevo grupo AB esta dado por el promedio ponderado: $\bar{y}_{AB} = \frac{n_A \bar{y}_A + n_B \bar{y}_B}{n_A + n_B}$

Este método es robusto frente a outliers, pero en otros aspectos su performance no es de las mejores.

6.5 Ward

Utiliza las distancias *Dentro*(W) y *Entre* (B) grupos.

$$SCT = SCResidual + SCExplicada \quad (T = W + B)$$

Nombres:(variación dentro=SCResiduales = **W** =SSE(error))

Si AB es el grupo obtenido por combinar los grupos A y B, la suma de las **distancias dentro** de los grupos son:

$$SSE_A = \sum_{i=1}^{n_A} (y_i - \bar{y}_A)'(y_i - \bar{y}_A)$$

$$SSE_B = \sum_{i=1}^{n_B} (y_i - \bar{y}_B)'(y_i - \bar{y}_B)$$

$$SSE_{AB} = \sum_{i=1}^{n_{AB}} (y_i - \bar{y}_{AB})'(y_i - \bar{y}_{AB})$$

- El método de Ward une dos cluster A y B que minimicen el incremento en SSE definido como

$$I_{AB} = SSE_{AB} - (SSE_A + SSE_B)$$

- El criterio de agregación consiste en minimizar el crecimiento de la variación intra-grupos resultante de la agregación de dos grupos en una nueva clase.
- La variación en los grupos de la clase que resulta de la unión $\{K \cup L\}$es más elevada que la suma de las variaciones de las clases K y L que la componen. Por lo que el criterio de agregación es el de reunir las clases (grupos) que minimicen el incremento de la variación intra-grupos provocado por la unión de elementos más heterogéneos.
- El crecimiento de la inercia intraclases de la nueva clase puede ser calculado por la expresión: $\Delta_{(K \cup L)} = \frac{n_K n_L}{n_K + n_L} d^2_{(G_K, G_L)}$

- En cada paso del algoritmo se calcula la expresión planteada anteriormente y se construye la tabla de disimilaridades. Se va uniendo de a 2 grupos teniendo en cuenta cual unión implica un menor incremento de la variación intragrupos.

6.6 Observaciones

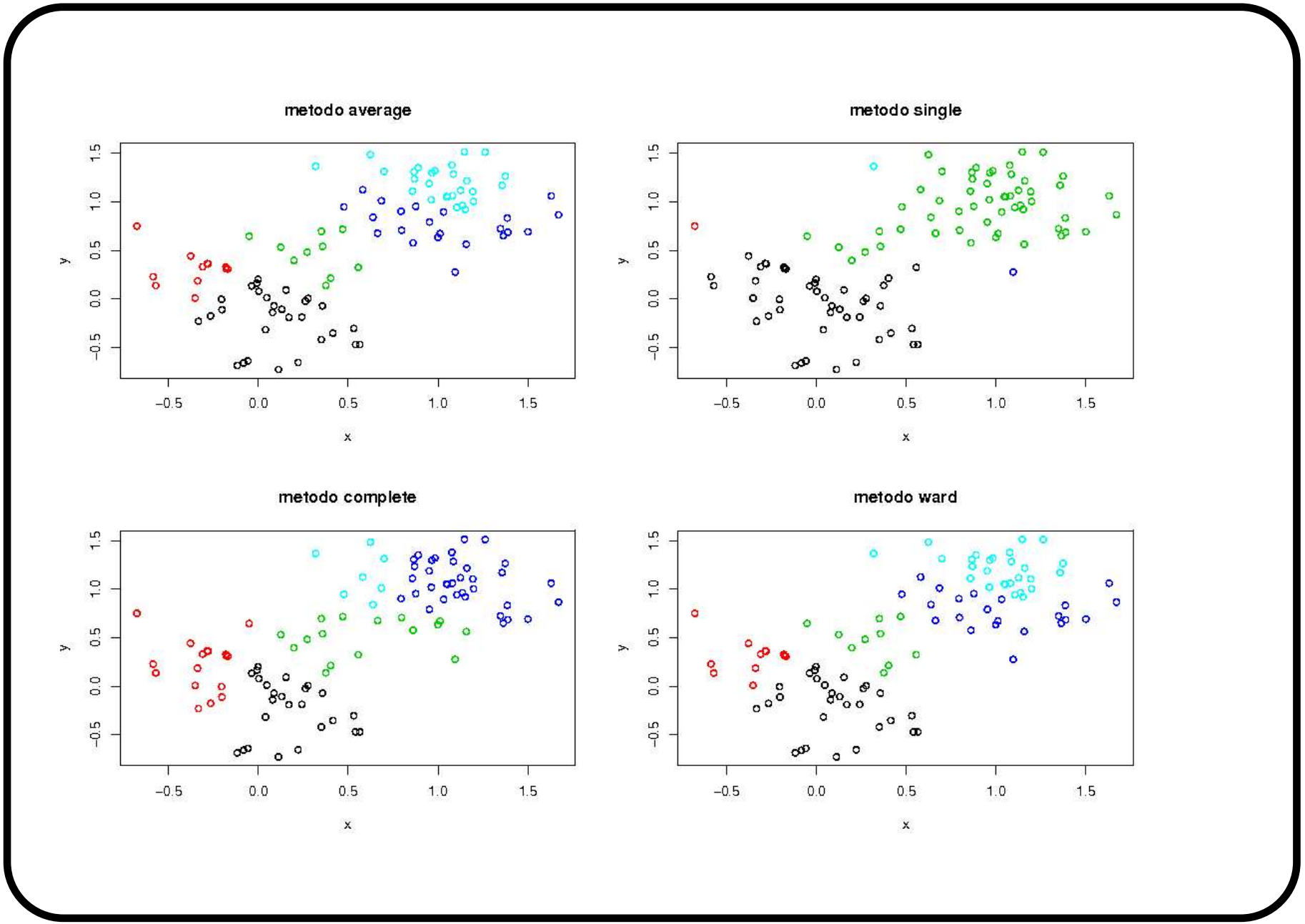
- El método tiende a juntar grupos con pequeño número de observaciones y es fuertemente sesgado hacia la formación de grupos con el mismo número de observaciones.
- El Método Ward aparece relacionado con espacios probabilísticos determinados, pues cuando se cumplen ciertas hipótesis (distribución Normal multivariada, matrices de covarianza esféricas iguales) es cuando el método funciona mejor.

6.6.1 Método beta-flexible(Lance, Williams)

Los métodos aglomerativos mas utilizados pueden describirse en terminos de la siguiente relación de recurrencia, en la cual la relación entre un nuevo grupo AB y otro C se define como:

$$d(C, AB) = \alpha_A d(C, A) + \alpha_B d(C, B) + \beta d(A, B) + \gamma |d(C, A) - d(C, B)|$$

Método	α_i	β	γ
single	$1/2$	0	$-1/2$
complete	$1/2$	0	$1/2$
average	$\frac{n_i}{n_i+n_j}$	0	0
centroide	$\frac{n_i}{n_i+n_j}$	$\frac{-n_i n_j}{(n_i+n_j)^2}$	0
ward			
...			


metodo average
metodo single
metodo complete
metodo ward
x
y
−0.5
0.0
0.5
1.0
1.5

Características generales de los algoritmos

1. El método de *W*ard y el del *C*entroide tienden a producir grupos más esféricos.
2. El método del *V*ecino más Lejano tiende a producir grupos esféricos de diámetros muy parecidos.
3. El método del *C*entroide y el de *W*ard son menos sensibles a la presencia de valores atípicos.
4. El método del *V*ecino más Cercano tiende a producir grupos más enlongados debido al "efecto cadena".
5. El método del *V*ecino más Cercano tiende a separar valores extremos dejándolos en grupos unitarios.

7 Reglas de Detención

La mayoría de los métodos de Análisis de Cluster requieren que el usuario especifique (no jerárquicos) o determine el número de grupos en la solución final (jerárquicos).

Se llaman reglas de detención a distintos métodos de decisión que ayudan a la elección del número de grupos en Análisis de Cluster.

Las reglas o *indices* que se verán han sido mayormente probadas el caso de los métodos jerárquicos y básicamente consisten en determinar el mejor nivel para cortar el dendrograma.

- Informales (representaciones gráficas)
- Formales (*stopping rules*)

  Reglas Globales Evaluan una medida de la bondad de la partición en k cluster usualmente basandose en las variaciones **dentro** y **entre** clusters e identificando el valor de k que optimiza la medida.

  Reglas Locales Examinan si deben unirse un par de cluster (o si un único debe subdividirse). Se basan solamente en una parte de los datos y trabajan solamente en los métodos jerárquicos.

7.1 R cuadrado (R^2)

Establece la relación entre la variación explicada y la variación total, donde la variación explicada la representa la estructura de grupos hallada en cada nivel.

$$R^2 = 1 - \frac{\sum_{k=1}^{K} \sum_{i=1}^{n_k} \sum_{j=1}^{J} (x_{ij(k)} - \bar{x}_{kj})^2}{\sum_{i=1}^{I} \sum_{j=1}^{J} (x_{ij} - \bar{x}_j)^2} \tag{2}$$

- Cuando se tienen I individuos el $R^2 = 1$ en la medida que la variación es toda explicada por los I grupos (individuos).
- Cuando el número de grupos es uno el $R^2 = 0$ en la medida que toda la variación es residual, no existe variación explicada por la estructura de grupos. No existe la variación entre grupos, toda la variación se da en el grupo existente(la nube original).
- En cada etapa de particiones encajadas se observa el valor del indicador y el incremento que se produce en el mismo al pasar de k grupos a $k + 1$ grupos. Si el aporte deja de ser "significativo" nos quedaremos con k grupos.

7.2 Pseudo F

$$\frac{\frac{trB}{k-1}}{\frac{trW}{n-k}} = F_{p(k-1),p(n-k)} \tag{3}$$

trB	suma de las variaciones entre los grupos (variación explicada)
trW	suma de las variaciones en los grupos (variación residual)
k	número de clusters
n	número de observaciones
p	número de variables

- Empíricamente se han determinado algunas reglas que contribuyen a su utilización:
  - Si el indicador crece monótonamente al crecer el número de grupos $k \Rightarrow$ no se puede determinar estructura clara.
  - Si el indicador disminuye monótonamente al crecer el número de grupos $k \Rightarrow$ no se puede determinar claramente estructura de grupos, pero se puede decir que existe una estructura jerárquica.
  - Si el indicador crece, llega a un máximo y luego decrece $\Rightarrow$ la población presenta un número definido de grupos en ese máximo.
- Se comporta "muy bien" en estructura de grupos muy compactos y en presencia de distribuciones normales multivariadas.

Si el análisis se realiza mirando de los I individuos a un único grupo final (de arriba hacia abajo):

Si en el paso $k+1$ el indicador tiene un valor muy chico con respecto al paso del nivel k entonces es conveniente quedarse con $k+1$ grupos en lugar de unirlos.

Esto es, el incremento en la heterogeneidad de unir esos grupos es muy grande y por tanto no es conveniente unirlos.

7.3 pseudo t^2

El *pseudo* t^2 al igual que el *pseudo* F es un indicador útil para determinar el número de grupos, pero no se distribuye *t-student* variable aleatoria.

Está relacionado con el indicador planteado por Duda - Hart, el que compara las trazas de las matrices de variaciones intragrupos G y L con la traza de la matriz de variaciones que surge al unir los grupos G y L.

$$Duda - Hart = DH = \frac{tr\ W_G + tr\ W_L}{tr\ W_{GL}} \tag{4}$$

$$pseudot^2 = \frac{tr\ W_{GL} - (tr\ W_G + tr\ W_L)}{(tr\ W_G + tr\ W_L)/(n_G + n_L - 2)} = \left[\frac{1}{DH} - 1\right](n_G + n_L - 2) \quad ($$

La idea que está detrás de estos indicadores es la de determinar la significación de fusionar dos grupos. En cada paso considera los candidatos a unirse.

Se trata de determinar en cada paso si la disminución en la suma de cuadrados residuales (variación intragrupos, o variación en los grupos) como resultado de pasar de k a $k+1$ grupos es significativa o no. Dicho de otra forma, se trata de determinar si la fusión implica un incremento tal en las variaciones en los grupos que la misma deba ser desechada.

Si el análisis se realiza mirando de los I individuos a un único grupo final (de arriba hacia abajo):

Si en el paso $k+1$ el indicador tiene un valor muy chico con respecto al paso del nivel k entonces es conveniente quedarse con $k+1$ grupos en lugar de unirlos.

Esto es, el incremento en la heterogeneidad de unir esos grupos es muy grande y por tanto no es conveniente unirlos.