Cluster Analysis

Graciela Boente

Datos de Planetas Enanos

Nombre	Nodo[1]	Inclinación[2]	Eje[3]
1935RF	130.916	4.659	2.2562
1941FD	132.200	4.700	2.1300
195QT	130.070	4.790	2.1893
1940YL	338.333	16.773	2.7465
1953NH	339.625	16.067	2.7335
1930SY	80.804	4.622	2.1890
1949HM	80.804	4.622	2.1906
1929EC	115.072	2.666	3.1676
1948R0	89.900	2.100	3.3500
1951AM	115.072	2.666	3.1676

- [1]: ángulo, en el plano de la órbita terrestre, en el cual el planeta cruza la órbita terrestre
- [2]: ángulo entre el plano de la órbita terrestre y el plano de la órbita del planeta
- [3]: Maxima distancia del planeta al sol dividida por la distancia de la tierra al sol.

Datos de Planetas Enanos

- Existen varios planetas enanos entre Marte y Jupiter.
- En una fotografía versus estrellas fijas, un planeta menor se ve como una línea curva a partir de la cual elementos orbitales pueden ser calculados.
- Muchos astrónomos ven los planetas menores como ruido que oscurece la observación de otros movimientos que ellos consideran interesantes.
- Hay mas de 700.000 planetas enanos de los cuales 518.420 están numerados.

Datos de Planetas Enanos

- Es importante decidir que visiones corresponden al mismo planeta.

- En particular si se dice haber descubierto un nuevo planeta hay que chequear que las observaciones hechas de él no corresponden a ningún otro planeta identificado con un nombre.

Datos de Planetas Enanos

- Es importante decidir que visiones corresponden al mismo planeta.

- En particular si se dice haber descubierto un nuevo planeta hay que chequear que las observaciones hechas de él no corresponden a ningún otro planeta identificado con un nombre.

El dar nombre a los planetas menores y la clasificación de las observaciones hechas es un típico problema de agrupamiento. Los objetos son las visiones y dos objetos son considerados similares si, teniendo en cuenta el error de medición, las observaciones puedan ser posiblemente del mismo planeta.

Un grupo es entonces un conjunto de observaciones del mismo planeta.

Comentarios

- Hemos visto como asignar una nueva observación a grupos ya conocidos.
- Un problema más complejo es el de descubrir cuáles son esos grupos, si no hay un criterio claro.
- La validez de los clusters obtenidos por muchos de los métodos existentes es cuestionable debido a la falta de desarrollo de aspecto probabilísticos y estadísticos que justifiquen esa metodología.

Definiciones

Cluster Analysis es el proceso a través del cuál objetivamente agrupamos juntas entidades en base a sus semejanzas o diferencias.

Estos métodos se conocen también como

- Métodos de clasificación automática o no supervisada
- Reconocimiento de patrones sin supervisión
- Métodos de conglomerados

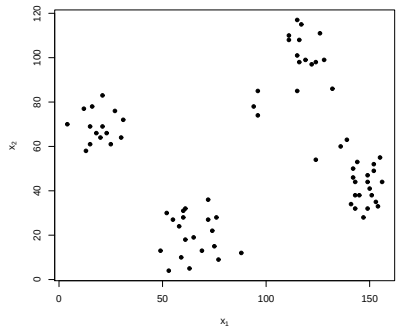
Los métodos de cluster pueden ser

- jerárquicos
- particionantes
- Clusters superpuestos

Análisis de clusters

Dados $\mathbf{x}_1, \ldots, \mathbf{x}_n$, $\mathbf{x}_i \in \mathbb{R}^p$, el propósito del cluster analysis es

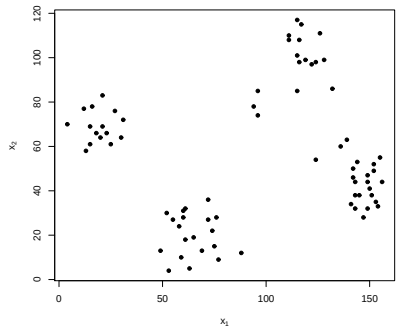
- Dar un esquema de clasificación para agrupar los objetos $\mathbf{x}_1, \ldots, \mathbf{x}_n$ en k grupos



Análisis de clusters

Dados $\mathbf{x}_1, \ldots, \mathbf{x}_n$, $\mathbf{x}_i \in \mathbb{R}^p$, el propósito del cluster analysis es

- Dar un esquema de clasificación para agrupar los objetos $\mathbf{x}_1, \ldots, \mathbf{x}_n$ en k grupos

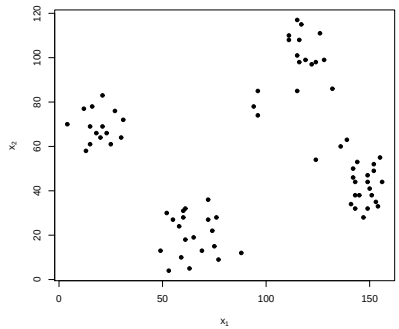


- Hallar, para un número de clusters dado una segmentación adecuada, de modo tal que los grupos sean homogéneos pero separados entre sí.

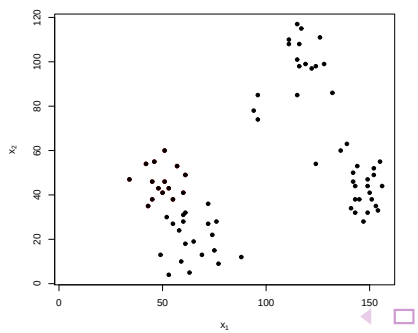
Análisis de clusters

Dados $\mathbf{x}_1, \ldots, \mathbf{x}_n$, $\mathbf{x}_i \in \mathbb{R}^p$, el propósito del cluster analysis es

- Dar un esquema de clasificación para agrupar los objetos $\mathbf{x}_1, \ldots, \mathbf{x}_n$ en k grupos



- Hallar, para un número de clusters dado una segmentación adecuada, de modo tal que los grupos sean homogéneos pero separados entre sí.



Clusters particionantes

Los clusters particionantes tratan el siguiente problema.

Dados $\mathbf{x}_1, \ldots, \mathbf{x}_n$ que sospechamos son heterogéneos, se desea dividirlos en k grupos de modo tal que

- cada elemento pertenezca a uno y sólo uno de los grupos

Clusters particionantes

Los clusters particionantes tratan el siguiente problema.

Dados $\mathbf{x}_1, \ldots, \mathbf{x}_n$ que sospechamos son heterogéneos, se desea dividirlos en k grupos de modo tal que

- cada elemento pertenezca a uno y sólo uno de los grupos
- cada individuo quede clasificado

Clusters particionantes

Los clusters particionantes tratan el siguiente problema.

Dados $\mathbf{x}_1, \ldots, \mathbf{x}_n$ que sospechamos son heterogéneos, se desea dividirlos en k grupos de modo tal que

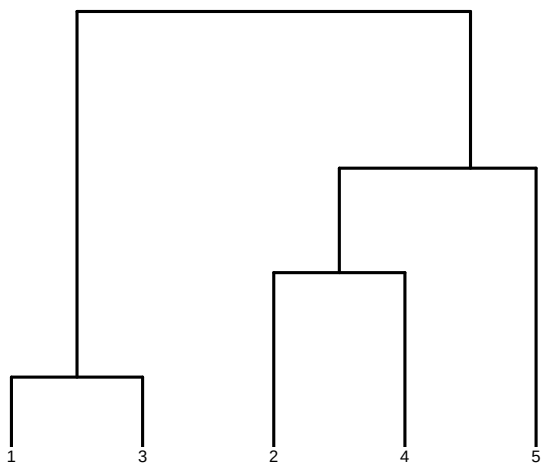
- cada elemento pertenezca a uno y sólo uno de los grupos
- cada individuo quede clasificado
- cada grupo sea internamente homogéneo.

Esto métodos usan la matriz $\mathbf{X}$ de datos.

Clusters jerárquicos

Los clusters jerárquicos desean estructurar los datos de acuerdo a su similitud pero en forma jerárquica.

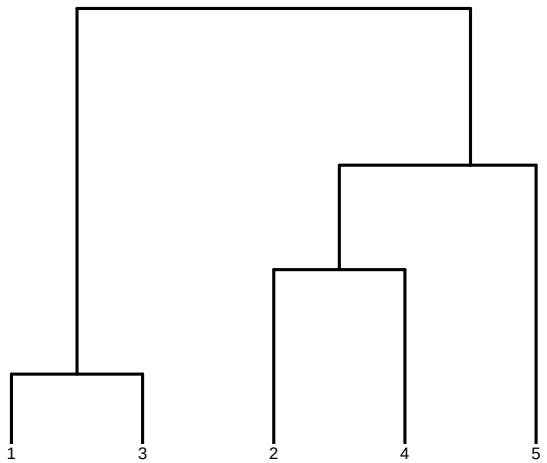
- En lugar de tener una partición se tienen niveles cada vez más finos, de modo que los niveles superiores contengan a los inferiores. Es usual al clasificar plantas o animales.



Clusters jerárquicos

Los clusters jerárquicos desean estructurar los datos de acuerdo a su similitud pero en forma jerárquica.

- En lugar de tener una partición se tienen niveles cada vez más finos, de modo que los niveles superiores contengan a los inferiores. Es usual al clasificar plantas o animales.



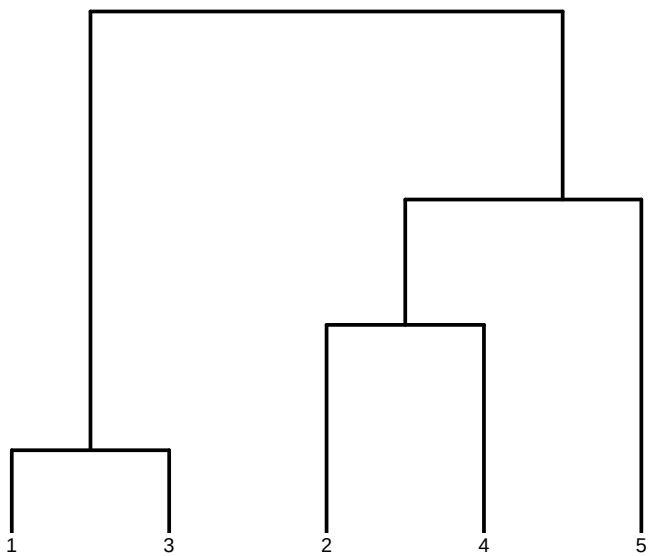
- Estrictamente hablando estos métodos no definen grupos, sino la estructura de asociación en cadena que pueda existir entre los elementos.

Clusters jerárquicos

- Sin embargo, la jerarquía obtenida permite también una partición en grupos.

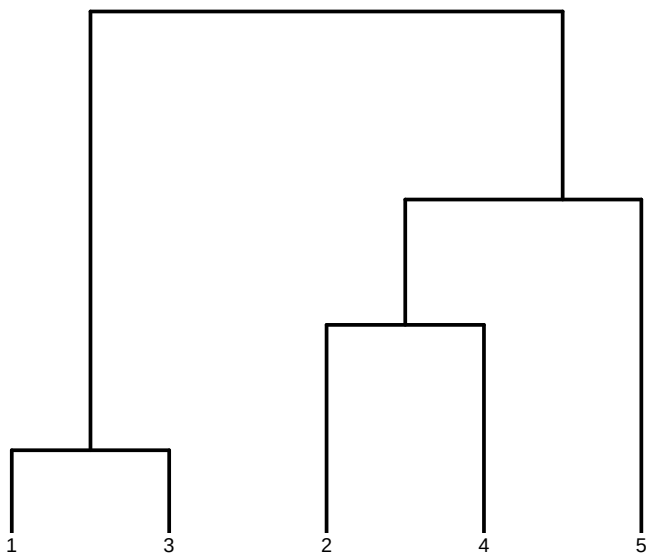
Clusters jerárquicos

- Sin embargo, la jerarquía obtenida permite también una partición en grupos.



Clusters jerárquicos

- Sin embargo, la jerarquía obtenida permite también una partición en grupos.



Esto métodos pueden usar la matriz **X** de datos para construir semejanzas o la matriz de similaridad.

Clasificación de variables

- En problemas con muchas variables (p grande o $p >> n$) es interesante hace un estudio exploratorio inicial para dividir las variables en grupos.
- Este estudio puede orientarnos para plantear modelos formales de reducción de dimensión como los vistos.
- Podemos usar agrupamientos particionantes o jerárquicos.

Para agrupar variables, se parte de

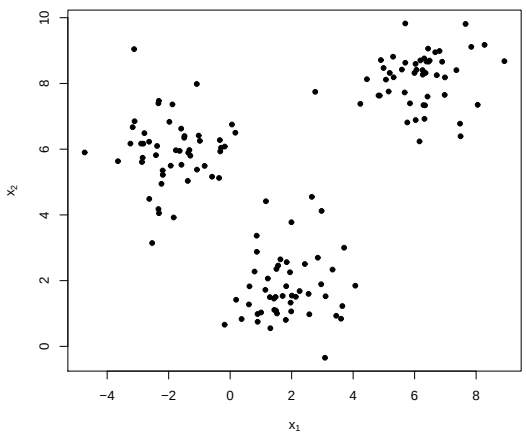
- la matriz de correlación para variables continuas y

- para variables discretas la matriz se construye a partir del coeficiente de contingencia.

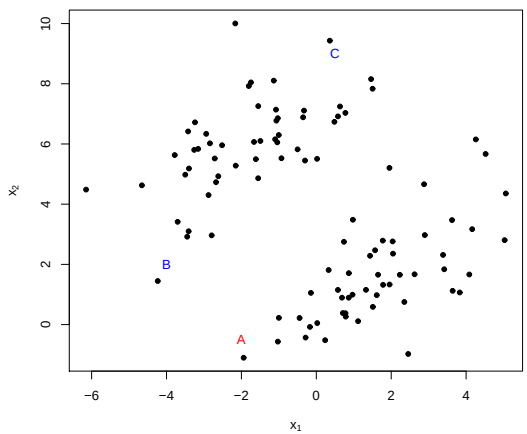
Tipos de cluster

Los clusters pueden ser de varias formas y tamaños

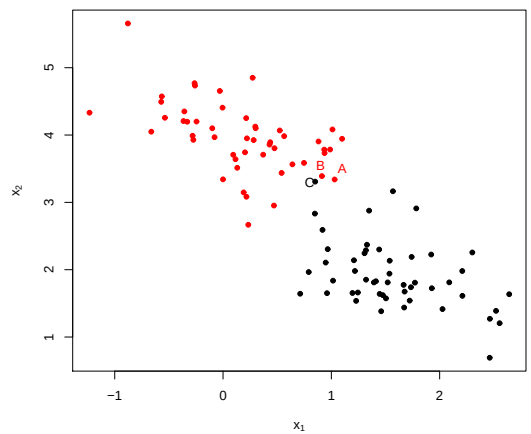
Esféricos



Elípticos

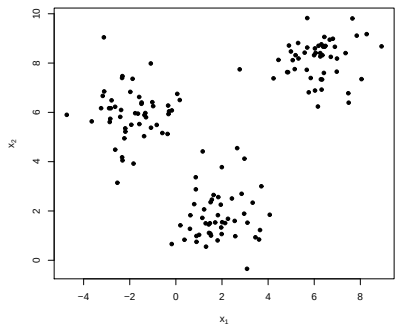


Encandenados



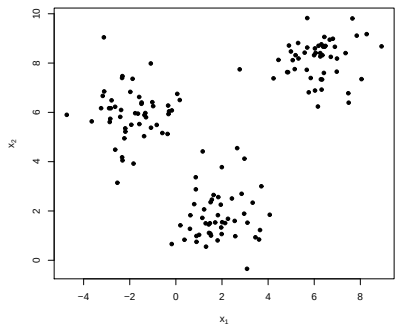
Tipos de cluster

- Para clusters esféricos la mayoría de los métodos llevan a una descripción adecuada.

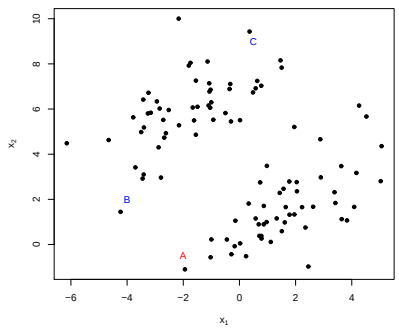


Tipos de cluster

- Para clusters esféricos la mayoría de los métodos llevan a una descripción adecuada.

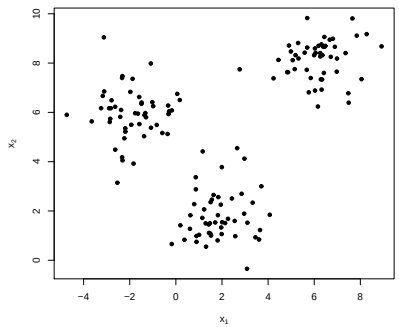


- Para clusters elípticos, un método que usa la distancia entre puntos puede llevar a resultados erroneos.

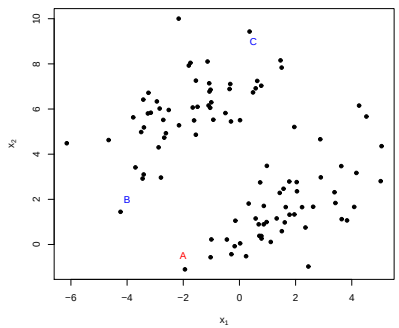


Tipos de cluster

- Para clusters esféricos la mayoría de los métodos llevan a una descripción adecuada.



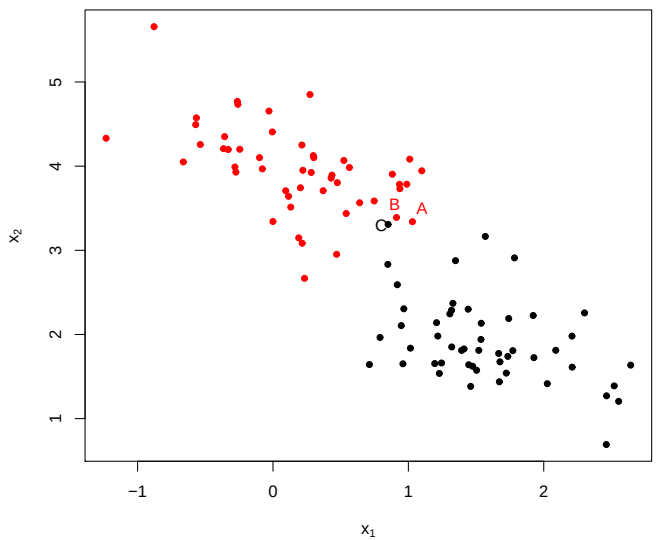
- Para clusters elípticos, un método que usa la distancia entre puntos puede llevar a resultados erroneos.



Un algoritmo que tiende a formar clusters compactos va a formar 4 clusters en lugar de 2 pues la distancia de B a C es mayor que la de B a A.

Tipos de cluster

Algunos algoritmos usan el concepto de vecino más cercano y por lo tanto producen un efecto cadena en datos como en la figura

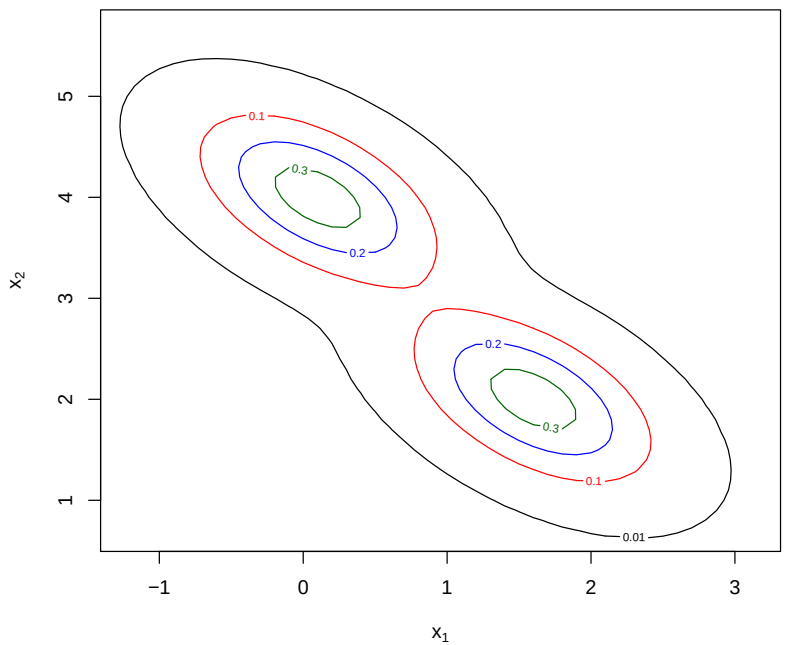


ya que C es cercano a B, B a A y si trabajamos con pares de puntos, obtendremos un solo cluster.

Cluster poblacional: Hartigan (1975)

- Un cluster es una región de alta densidad.
- Dado $f_0 > 0$, un cluster de alta densidad de nivel f_0 para $\mathbf{x} \sim f$ es el **mayor conjunto convexo conexo** de la forma

$$\{\mathbf{x} : f(\mathbf{x}) \geq f_0\}$$



Cluster poblacional: Hartigan (1975)

- Una familia de tales clusters $\mathcal{T}$ forma un árbol en el sentido que

  Dados $A, B \in \mathcal{T}$, se cumple una de las tres
  - $A \subset B$
  - $B \subset A$
  - $A \cap B = \emptyset$

Para árboles jerárquicos es deseable que la sucesión $\mathcal{T}_n$ de dendogramas definida por la muestra $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim \mathbf{x}$ converja en algún sentido a $\mathcal{T}$.

Si $A \cap B = \emptyset$, y si A_n y B_n son los menores clusters de $\mathcal{T}_n$ tales que $A \subset A_n$ y $B \subset B_n$ entonces $\mathbb{P}(A_n \cap B_n = \emptyset) \to 1$.

Método noparamétrico

Esta noción permite definir métodos de cluster a partir de estimadores de la densidad como, por ejemplo,

$$f_n(\mathbf{x}) = \frac{k_n}{nH_k^p \lambda(\mathcal{V}_1)}$$

donde

- $\lambda(\mathcal{V}_1)$ es el volumen de la bola unidad $\mathcal{V}_1 = \{\mathbf{u} : \|\mathbf{u}\| = 1\}$
- $H_k = H_k(\mathbf{x})$ es la distancia de $\mathbf{x}$ a su $k-$ésimo vecino más cercano entre $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

Método noparamétrico

Esta noción permite definir métodos de cluster a partir de estimadores de la densidad como, por ejemplo,

$$f_n(\mathbf{x}) = \frac{k_n}{nH_k^p \lambda(\mathcal{V}_1)}$$

donde

- $\lambda(\mathcal{V}_1)$ es el volumen de la bola unidad $\mathcal{V}_1 = \{\mathbf{u} : \|\mathbf{u}\| = 1\}$
- $H_k = H_k(\mathbf{x})$ es la distancia de $\mathbf{x}$ a su $k-$ésimo vecino más cercano entre $\mathbf{x}_1, \ldots, \mathbf{x}_n$.
- Si f es continua, tenemos que
  - si $k_n \to \infty$ y $k_n/n \to 0$ $f_n(\mathbf{x}) \xrightarrow{p} f(\mathbf{x})$
  - Si además $k_n/\log(n) \to \infty$, $f_n(\mathbf{x}) \xrightarrow{a.s.} f(\mathbf{x})$ y la convergencia es uniforme sobre compactos.

Método noparamétrico

- Más aún, tasas óptimas de convergencia se obtienen si $k_n = an^{\frac{2}{p}}$.

Método noparamétrico

- Más aún, tasas óptimas de convergencia se obtienen si $k_n = an^{\frac{2}{p}}$.

Los clusters quedan definidos por

$$\{\mathbf{x} : f_n(\mathbf{x}) \geq f_0\}.$$

De esta forma obtenemos un árbol $\mathcal{T}_n$ en el sentido antes mencionado que va a converger a $\mathcal{T}$ si se cumple $k_n \to \infty$, $k_n/n \to 0$ y $k_n/\log(n) \to \infty$.

Método noparamétrico

- Más aún, tasas óptimas de convergencia se obtienen si $k_n = an^{\frac{2}{p}}$.

Los clusters quedan definidos por

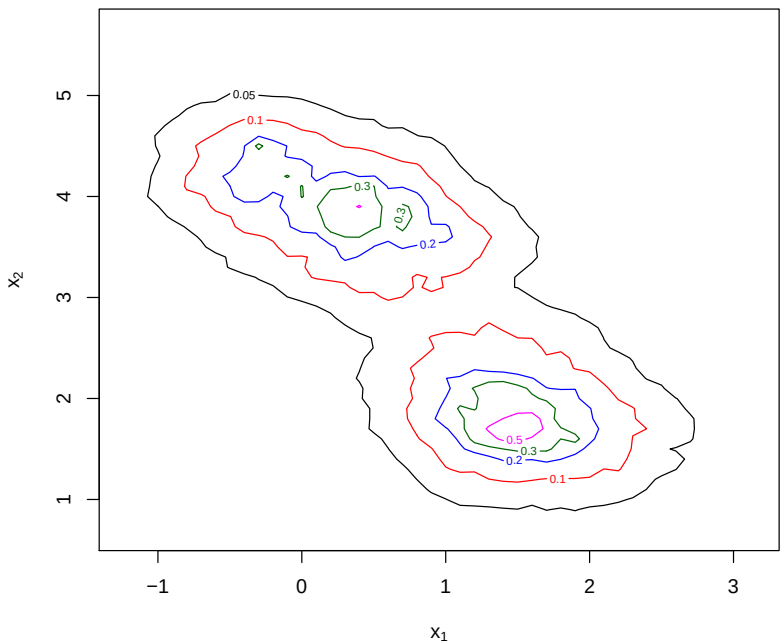
$$\{\mathbf{x} : f_n(\mathbf{x}) \geq f_0\}.$$

De esta forma obtenemos un árbol $\mathcal{T}_n$ en el sentido antes mencionado que va a converger a $\mathcal{T}$ si se cumple $k_n \to \infty$, $k_n/n \to 0$ y $k_n/\log(n) \to \infty$.

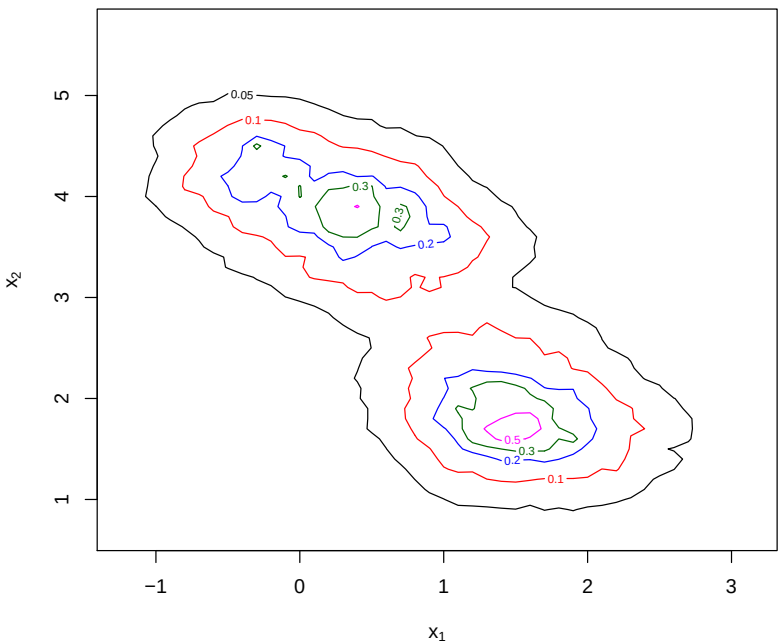
El método de **single linkage** corresponde a la elección $k_n = 1$. Por lo tanto, no se obtiene un estimador consistente de la densidad, lo que explica su tendencia excesiva a encadenar.

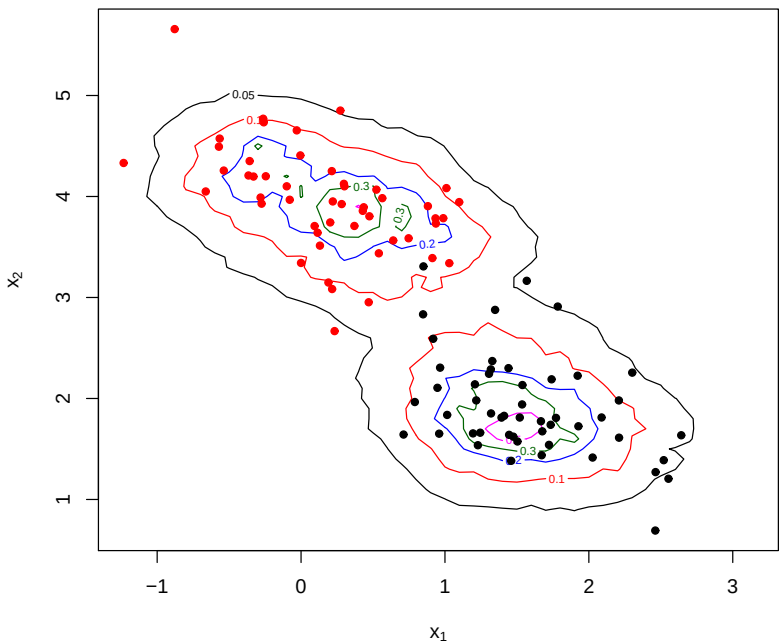
Sin embargo, si $p = 1$ el método de **single linkage** es consistente, pero no lo es para $p > 1$.

Ejemplo: $k-$vecinos con $k = 10$



Ejemplo: $k-$vecinos con $k = 10$





Como elegir k

- En la práctica no existe un método satisfactorio para elegir k
- Una forma es plotear para cada $\mathbf{x}_i$ de

$$\log(k) \text{ versus } H_k(\mathbf{x}_i)$$

- En la frontera de un cluster deberían producirse cortes en el plot.
- Muy costoso si hay muchos datos.

Distancias o disimilaridades

Sea $\mathcal{N} = \{1, \cdots, n\}$ identificaremos a $\mathbf{x}_i$ con su índice i.

Los métodos jerárquicos parten de una matriz de distancias o de similaridad entre elementos de la muestra.

Definición. Diremos que $d : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ es una disimilaridad si

- $d(\mathbf{x}, \mathbf{y}) \geq 0$
- $d(\mathbf{x}, \mathbf{y}) = 0$ si y sólo si $\mathbf{x} = \mathbf{y}$
- $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$

Se dirá una métrica si además $d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y})$.

Disimilaridades

Sea $\mathcal{N} = \{1, \cdots, n\}$ identificaremos a $\mathbf{x}_i$ con su índice i.

Dada una matriz $\mathbf{D} = (d_{ij})_{1 \leq i,j \leq n}$ diremos que $\mathbf{D}$ es una matriz de disimilaridad si

- $d_{ij} \geq 0$
- $d_{ii} = 0$
- $d_{ij} = d_{ji}$

Es métrica si además $d_{ij} \leq d_{i\ell} + d_{\ell j}$.

Disimilaridades

Si las variables son continuas, las disimilaridades más usadas son

a) la distancia euclídea $d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{j=1}^{p}(x_j - y_j)^2}$

Inconvenientes:

- Los cambios de escala afectan el ordenamiento de las distancias
- Depende de las variables con valores más grandes

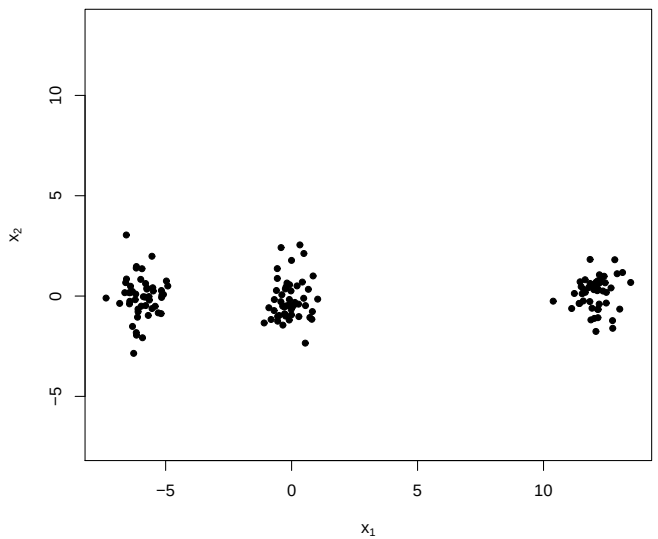
b) la distancia euclídea de las variables estandarizadas univariadamente

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{j=1}^{p}\left(\frac{x_j - y_j}{s_j}\right)^2}$$

Disimilaridades

b) la distancia euclídea de las variables estandarizadas univariadamente

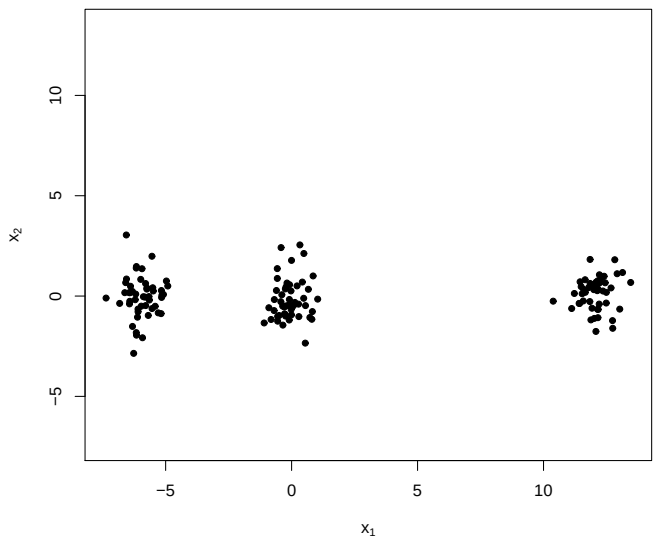
$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{j=1}^{p} \left(\frac{x_j - y_j}{s_j} \right)^2}$$

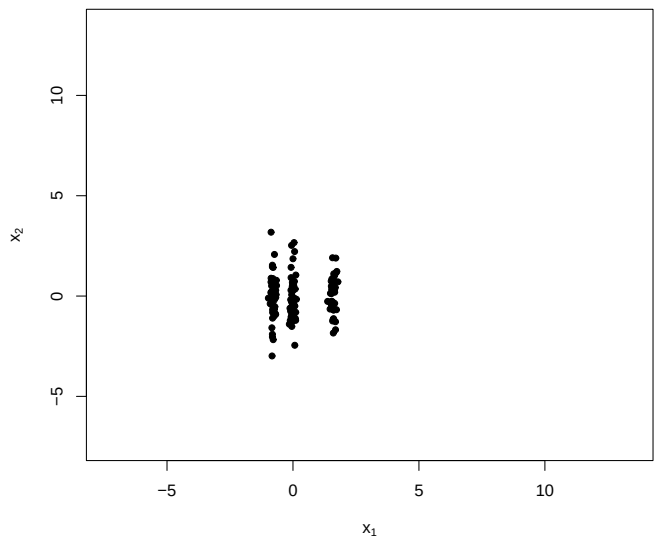


Disimilaridades

b) la distancia euclídea de las variables estandarizadas univariadamente

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{j=1}^{p} \left(\frac{x_j - y_j}{s_j} \right)^2}$$





Disimilaridades

b) la distancia euclídea de las variables estandarizadas univariadamente

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{j=1}^{p} \left(\frac{x_j - y_j}{s_j} \right)^2}$$

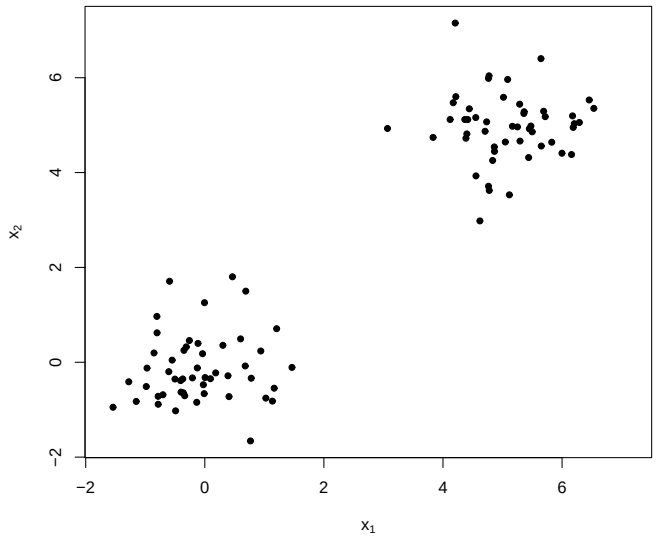
Inconvenientes:

- Estandarizar las variables puede diluir las diferencias entre clusters con respecto a las variables más discriminatorias.

- Las distancias entre puntos dentro de los grupos pueden resultar mayores respecto de las distancias de puntos entre clusters y los grupos resultan menos diferenciados.

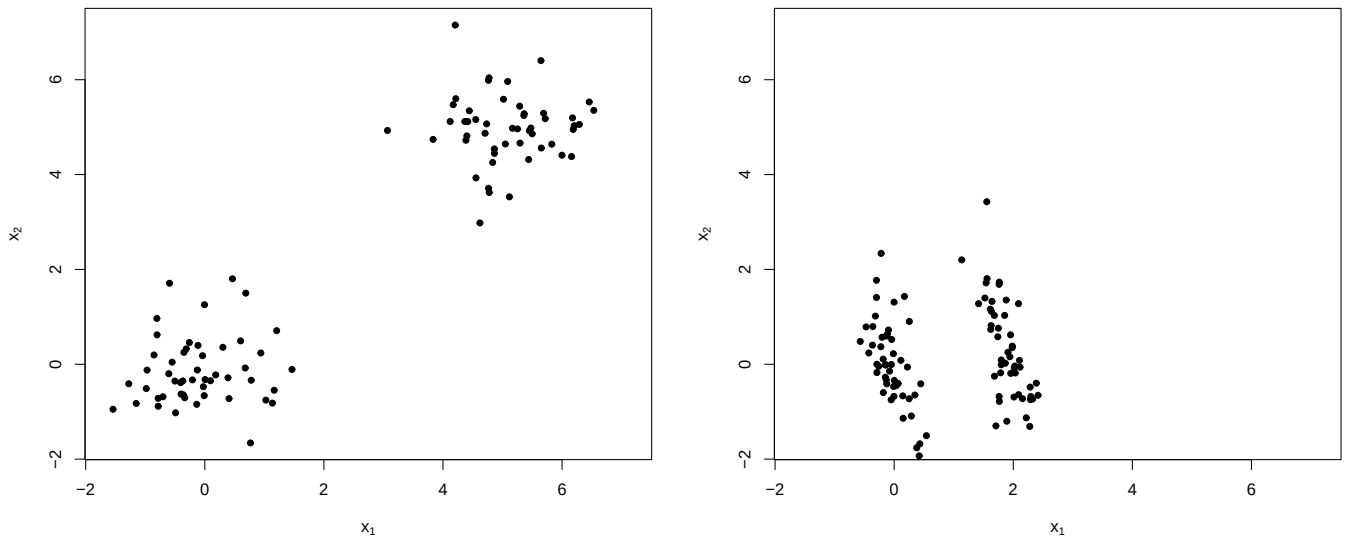
Disimilaridades

En general, no se usa la distancia de Mahalanobis, ya que la matriz de covarianza muestral de toda la muestra puede deformar el análisis



Disimilaridades

En general, no se usa la distancia de Mahalanobis, ya que la matriz de covarianza muestral de toda la muestra puede deformar el análisis



En este caso cada uno los dos grupos fue generado con matriz decovarianza $\mathbf{I}_2$ pero el aspecto del conjunto de todos los puntos hace pensar en una correlación positiva. De hecho la correlación estimada con todos los datos es 0.36.

Inconvenientes:

Las distancias entre puntos dentro de los grupos creció respecto de las distancias de puntos entre clusters y los grupos resultan menos diferenciados.

Similaridades

Este problema se agrava si hay variables binarias. Por eso, se suelen usar matrices de similaridad.

Definición. Diremos que $\mathbf{C} = (c_{ij})_{1 \leq i,j \leq n}$ diremos que $\mathbf{C}$ es una matriz de similaridad si

- $c_{ij} \leq 1$
- $c_{ii} = 1$
- $c_{ij} = c_{ji}$

Podemos crear una disimilaridad a partir de $\mathbf{C}$ como

- $d_{ij} = 1 - c_{ij}$
  o
- $d_{ij} = \sqrt{2(1 - c_{ij})}$ que es métrica si $\mathbf{C}$ es definida positiva

Similaridades

Definición. El coeficiente de similaridad de la variable ℓ entre dos elementos i y j se define como

- $c_{ij,\ell} \leq 1$
- $c_{ii,\ell} = 1$
- $c_{ij,\ell} = c_{ji,\ell}$

Similaridades

Definición. El coeficiente de similaridad de la variable ℓ entre dos elementos i y j se define como

- $c_{ij,\ell} \leq 1$
- $c_{ii,\ell} = 1$
- $c_{ij,\ell} = c_{ji,\ell}$

Gower (1971) propone contruir una similaridad entre los individuos i y j como

$$c_{ij} = \frac{\sum_{\ell=1}^{p} w_{ij,\ell}\, c_{ij,\ell}}{\sum_{\ell=1}^{p} w_{ij,\ell}}$$

$$w_{ij,\ell} \begin{cases} 0 & \text{si la comparación no es posible o no se quiere incluir a la variable } \ell \\ 1 & \text{caso contario} \end{cases}$$

Similaridades: Cómo definirlas?

- Para **variables continuas** $w_{ij,\ell} = 1$ y

$$c_{ij,\ell} = 1 - \frac{|x_{i\ell} - x_{j\ell}|}{R_\ell}$$

donde R_ℓ es el rango de la variable ℓ. Si todas son continuas corresponde a la llamada *métrica de Gower*.

- Para **variables binarias**, o sea, si $x_{i\ell} = 0$ o 1 para todo individuo i,

$$c_{ij,\ell} = \begin{cases} 1 & \text{si } x_{i\ell} = x_{j\ell} \\ 0 & \text{caso contrario} \end{cases}$$

Similaridades: Cómo definirlas?

- Se pueden agrupar las variables binarias en grupos homogéneos y tratarlas conjuntamente. Supongamos que $\mathbf{x}$ tiene todas sus variables binarias entonces podemos definir la tabla

		$\mathbf{x}_i$	
		1	0
$\mathbf{x}_j$	1	α	β
	0	γ	δ

- **Criterio de proporción de coincidencias**

$$c_{ij,\ell} = \frac{\alpha + \delta}{p} \qquad \text{pues} \qquad \alpha + \beta + \gamma + \delta = p$$

- **Criterio de proporción de apariciones**

$$c_{ij,\ell} = \frac{\alpha}{\alpha + \beta + \gamma}$$

Similaridades: Cómo definirlas?

- **Otras propuestas**

$$c_{ij,\ell}(\lambda,\mu) = 1 - \frac{\lambda(\beta+\gamma)}{\alpha+\lambda(\beta+\gamma)+\mu\delta} = \frac{\alpha+\mu\delta}{\alpha+\lambda(\beta+\gamma)+\mu\delta}$$

es métrica si $\lambda \geq 1$ y $\mu = 0$ o 1.

  - $\mu = 0$ se usa para el caso antisimétrico
  - $\mu = 1$ se usa para el caso simétrico.
    Observemos que $\beta + \gamma = \sum_{\ell=1}^{p} |x_{i\ell} - x_{j\ell}|$
  - Si $\lambda = 2$ se da el doble de peso a las coincidencias.

- Para variables discretas con más de dos estados, Gower propone tomar $w_{ij,\ell} = 1$ y

$$c_{ij,\ell} = \begin{cases} 1 & \text{si } x_{i\ell} = x_{j\ell} \\ 0 & \text{caso contrario} \end{cases}$$

Similaridades: Cómo definirlas?

Si $\mathbf{x}$ tiene p_1 variables continuas, p_2 variables binarias y p_3 variables discretas de más de dos valores, podemos definir una similaridad entre los individuos i y j como

$$c_{ij} = \frac{\sum_{\ell=1}^{p_1} \left\{ 1 - \frac{|x_{i\ell} - x_{j\ell}|}{R_\ell} \right\} + \alpha_2 + m_3}{p_1 + (p_2 - \delta_2) + p_3}$$

con

- α_2 y δ_2 los empates en 1 y 0 para las p_2 variables binarias, respectivamente
- m_3 el número de empates para las p_3 variables discretas no dicotómicas

Tipos de Clusters

Dada una matriz de disimilaridades $\mathbf{D} = (d_{ij})_{1 \leq i,j \leq n}$, se desea clasificar cada punto en un grupo.

Los algoritmos jerárquicos son de dos tipos:

- **Algomerativos:** Parten de elementos individuales y los van uniendo en grupos
- **de división:** Parten de todo el conjunto de datos y los dividen sucesivamente hasta llegar a los individuos.

Los métodos aglomerativos son más rápidos y son los más usados.

Clusters Algomerativos: Algoritmo

1. Empiece con $\mathcal{C}_j = \{j\}$, $1 \leq j \leq n$. La distancia $d(\mathcal{C}_\ell, \mathcal{C}_r)$ entre dos clases $\mathcal{C}_\ell$ y $\mathcal{C}_r$ es $d(\mathcal{C}_\ell, \mathcal{C}_r) = d_{\ell r}$

Clusters Algomerativos: Algoritmo

1. Empiece con $\mathcal{C}_j = \{j\}$, $1 \leq j \leq n$. La distancia $d(\mathcal{C}_\ell, \mathcal{C}_r)$ entre dos clases $\mathcal{C}_\ell$ y $\mathcal{C}_r$ es $d(\mathcal{C}_\ell, \mathcal{C}_r) = d_{\ell r}$

2. Seleccione los elementos más próximos en la matriz de distancias y forme con ellos una clase.

Clusters Algomerativos: Algoritmo

1. Empiece con $\mathcal{C}_j = \{j\}$, $1 \leq j \leq n$. La distancia $d(\mathcal{C}_\ell, \mathcal{C}_r)$ entre dos clases $\mathcal{C}_\ell$ y $\mathcal{C}_r$ es $d(\mathcal{C}_\ell, \mathcal{C}_r) = d_{\ell r}$

2. Seleccione los elementos más próximos en la matriz de distancias y forme con ellos una clase.

   Es decir, si $\mathcal{C}_\ell$ y $\mathcal{C}_r$ son tales que

   $$d(\mathcal{C}_\ell, \mathcal{C}_r) = \min_{i \neq j} d(\mathcal{C}_i, \mathcal{C}_j)$$

   defina

   $$\mathcal{C}_\ell^{(new)} = \mathcal{C}_\ell \cup \mathcal{C}_r$$

Clusters Algomerativos: Algoritmo

1. Empiece con $\mathcal{C}_j = \{j\}$, $1 \leq j \leq n$. La distancia $d(\mathcal{C}_\ell, \mathcal{C}_r)$ entre dos clases $\mathcal{C}_\ell$ y $\mathcal{C}_r$ es $d(\mathcal{C}_\ell, \mathcal{C}_r) = d_{\ell r}$

2. Seleccione los elementos más próximos en la matriz de distancias y forme con ellos una clase.

   Es decir, si $\mathcal{C}_\ell$ y $\mathcal{C}_r$ son tales que

   $$d(\mathcal{C}_\ell, \mathcal{C}_r) = \min_{i \neq j} d(\mathcal{C}_i, \mathcal{C}_j)$$

   defina

   $$\mathcal{C}_\ell^{(new)} = \mathcal{C}_\ell \cup \mathcal{C}_r$$

3. Tire la columna y fila r de la matriz de distancias usada en 2. y cambie todas las distancias en las que interviene el cluster ℓ de acuerdo a uno de los criterios que daremos a continuación.

Clusters Algomerativos: Algoritmo

1. Empiece con $\mathcal{C}_j = \{j\}$, $1 \leq j \leq n$. La distancia $d(\mathcal{C}_\ell, \mathcal{C}_r)$ entre dos clases $\mathcal{C}_\ell$ y $\mathcal{C}_r$ es $d(\mathcal{C}_\ell, \mathcal{C}_r) = d_{\ell r}$

2. Seleccione los elementos más próximos en la matriz de distancias y forme con ellos una clase.

   Es decir, si $\mathcal{C}_\ell$ y $\mathcal{C}_r$ son tales que

   $$d(\mathcal{C}_\ell, \mathcal{C}_r) = \min_{i \neq j} d(\mathcal{C}_i, \mathcal{C}_j)$$

   defina

   $$\mathcal{C}_\ell^{(new)} = \mathcal{C}_\ell \cup \mathcal{C}_r$$

3. Tire la columna y fila r de la matriz de distancias usada en 2. y cambie todas las distancias en las que interviene el cluster ℓ de acuerdo a uno de los criterios que daremos a continuación.
4. Volver a 2.

Clusters Algomerativos: Algoritmo

Lance y Williams (1967) proponen tomar como distancia en el punto 3. de acuerdo al siguiente procedimiento.

Sean $\mathcal{C}_1$, $\mathcal{C}_2$ y $\mathcal{C}_3$ tres clusters y sea $\mathcal{C}^{(new)} = \mathcal{C}_1 \cup \mathcal{C}_2$, etonces

$$d(\mathcal{C}^{(new)}, \mathcal{C}_3) = \sum_{\ell=1}^{2} \alpha_j d(\mathcal{C}_\ell, \mathcal{C}_3) + \beta d(\mathcal{C}_1, \mathcal{C}_2) + \gamma \, |d(\mathcal{C}_1, \mathcal{C}_3) - d(\mathcal{C}_2, \mathcal{C}_3)|$$

Clusters Algomerativos: Algoritmo

Lance y Williams (1967) proponen tomar como distancia en el punto 3. de acuerdo al siguiente procedimiento.

Sean $\mathcal{C}_1$, $\mathcal{C}_2$ y $\mathcal{C}_3$ tres clusters y sea $\mathcal{C}^{(new)} = \mathcal{C}_1 \cup \mathcal{C}_2$, etonces

$$d(\mathcal{C}^{(new)}, \mathcal{C}_3) = \sum_{\ell=1}^{2} \alpha_j d(\mathcal{C}_\ell, \mathcal{C}_3) + \beta d(\mathcal{C}_1, \mathcal{C}_2) + \gamma \, |d(\mathcal{C}_1, \mathcal{C}_3) - d(\mathcal{C}_2, \mathcal{C}_3)|$$

	α_ℓ	β	γ
Single linkage	$\frac{1}{2}$	0	$-\frac{1}{2}$
Complete linkage	$\frac{1}{2}$	0	$\frac{1}{2}$
Average linkage	$\frac{n_\ell}{n_1+n_2}$	0	0
Centroide	$\frac{n_\ell}{n_1+n_2}$	$-\frac{n_1 n_2}{(n_1+n_2)^2}$	0
Ward	$\frac{n_\ell+n_3}{n_1+n_2+n_3}$	$-\frac{n_3}{n_1+n_2+n_3}$	0
Mediana	$\frac{1}{2}$	$-\frac{1}{4}$	0

Single Linkage

$$d(\mathcal{C}^{(new)}, \mathcal{C}_3) = \min\{d(\mathcal{C}_1, \mathcal{C}_3), d(\mathcal{C}_2, \mathcal{C}_3)\}$$

Dados dos clusters $\mathcal{C}_1$ y $\mathcal{C}_2$

$$d(\mathcal{C}_1, \mathcal{C}_2) = \min\{d_{rs} : r \in \mathcal{C}_1, s \in \mathcal{C}_2\}$$

Single Linkage

$$d(\mathcal{C}^{(new)}, \mathcal{C}_3) = \min\{d(\mathcal{C}_1, \mathcal{C}_3), d(\mathcal{C}_2, \mathcal{C}_3)\}$$

Dados dos clusters $\mathcal{C}_1$ y $\mathcal{C}_2$

$$d(\mathcal{C}_1, \mathcal{C}_2) = \min\{d_{rs} : r \in \mathcal{C}_1, s \in \mathcal{C}_2\}$$

- Este criterio sólo depende del orden entre las distancias, por lo que es invariante por tranformaciones monótonas de las mismas.

Single Linkage

$$d(\mathcal{C}^{(new)}, \mathcal{C}_3) = \min\{d(\mathcal{C}_1, \mathcal{C}_3), d(\mathcal{C}_2, \mathcal{C}_3)\}$$

Dados dos clusters $\mathcal{C}_1$ y $\mathcal{C}_2$

$$d(\mathcal{C}_1, \mathcal{C}_2) = \min\{d_{rs} : r \in \mathcal{C}_1, s \in \mathcal{C}_2\}$$

- Este criterio sólo depende del orden entre las distancias, por lo que es invariante por tranformaciones monótonas de las mismas.
- El criterio no recobra clusters compactos

Single Linkage

$$d(\mathcal{C}^{(new)}, \mathcal{C}_3) = \min\{d(\mathcal{C}_1, \mathcal{C}_3), d(\mathcal{C}_2, \mathcal{C}_3)\}$$

Dados dos clusters $\mathcal{C}_1$ y $\mathcal{C}_2$

$$d(\mathcal{C}_1, \mathcal{C}_2) = \min\{d_{rs} : r \in \mathcal{C}_1, s \in \mathcal{C}_2\}$$

- Este criterio sólo depende del orden entre las distancias, por lo que es invariante por tranformaciones monótonas de las mismas.
- El criterio no recobra clusters compactos
- Tiende a formar clusters alargados por efecto de encadenamiento.

Complete Linkage

$$d(C^{(new)}, C_3) = \max\{d(C_1, C_3), d(C_2, C_3)\}$$

Dados dos clusters C_1 y C_2

$$d(C_1, C_2) = \max\{d_{rs} : r \in C_1, s \in C_2\}$$

Complete Linkage

$$d(\mathcal{C}^{(new)}, \mathcal{C}_3) = \max\{d(\mathcal{C}_1, \mathcal{C}_3), d(\mathcal{C}_2, \mathcal{C}_3)\}$$

Dados dos clusters $\mathcal{C}_1$ y $\mathcal{C}_2$

$$d(\mathcal{C}_1, \mathcal{C}_2) = \max\{d_{rs} : r \in \mathcal{C}_1, s \in \mathcal{C}_2\}$$

- Este criterio sólo depende del orden entre las distancias, por lo que es invariante por tranformaciones monótonas de las mismas.

Complete Linkage

$$d(\mathcal{C}^{(new)}, \mathcal{C}_3) = \max\{d(\mathcal{C}_1, \mathcal{C}_3), d(\mathcal{C}_2, \mathcal{C}_3)\}$$

Dados dos clusters $\mathcal{C}_1$ y $\mathcal{C}_2$

$$d(\mathcal{C}_1, \mathcal{C}_2) = \max\{d_{rs} : r \in \mathcal{C}_1, s \in \mathcal{C}_2\}$$

- Este criterio sólo depende del orden entre las distancias, por lo que es invariante por tranformaciones monótonas de las mismas.
- Tiende a formar cluster esféricos.

Complete Linkage

$$d(C^{(new)}, C_3) = \max\{d(C_1, C_3), d(C_2, C_3)\}$$

Dados dos clusters C_1 y C_2

$$d(C_1, C_2) = \max\{d_{rs} : r \in C_1, s \in C_2\}$$

- Este criterio sólo depende del orden entre las distancias, por lo que es invariante por tranformaciones monótonas de las mismas.
- Tiende a formar cluster esféricos.
- Puede verse distorsionado por outliers.

Average Linkage

$$d(C^{(new)}, C_3) = \frac{n_1}{n_1 + n_2} d(C_1, C_3) + \frac{n_2}{n_1 + n_2} d(C_2, C_3)$$

Dados dos clusters C_1 y C_2

$$d(C_1, C_2) = \frac{1}{n_1\, n_2} \sum_{r \in C_1} \sum_{s \in C_2} d_{rs}$$

Average Linkage

$$d(\mathcal{C}^{(new)}, \mathcal{C}_3) = \frac{n_1}{n_1 + n_2} d(\mathcal{C}_1, \mathcal{C}_3) + \frac{n_2}{n_1 + n_2} d(\mathcal{C}_2, \mathcal{C}_3)$$

Dados dos clusters $\mathcal{C}_1$ y $\mathcal{C}_2$

$$d(\mathcal{C}_1, \mathcal{C}_2) = \frac{1}{n_1\, n_2} \sum_{r \in \mathcal{C}_1} \sum_{s \in \mathcal{C}_2} d_{rs}$$

- Este criterio **no** es invariante por tranformaciones monótonas.

Average Linkage

$$d(\mathcal{C}^{(new)}, \mathcal{C}_3) = \frac{n_1}{n_1 + n_2} d(\mathcal{C}_1, \mathcal{C}_3) + \frac{n_2}{n_1 + n_2} d(\mathcal{C}_2, \mathcal{C}_3)$$

Dados dos clusters $\mathcal{C}_1$ y $\mathcal{C}_2$

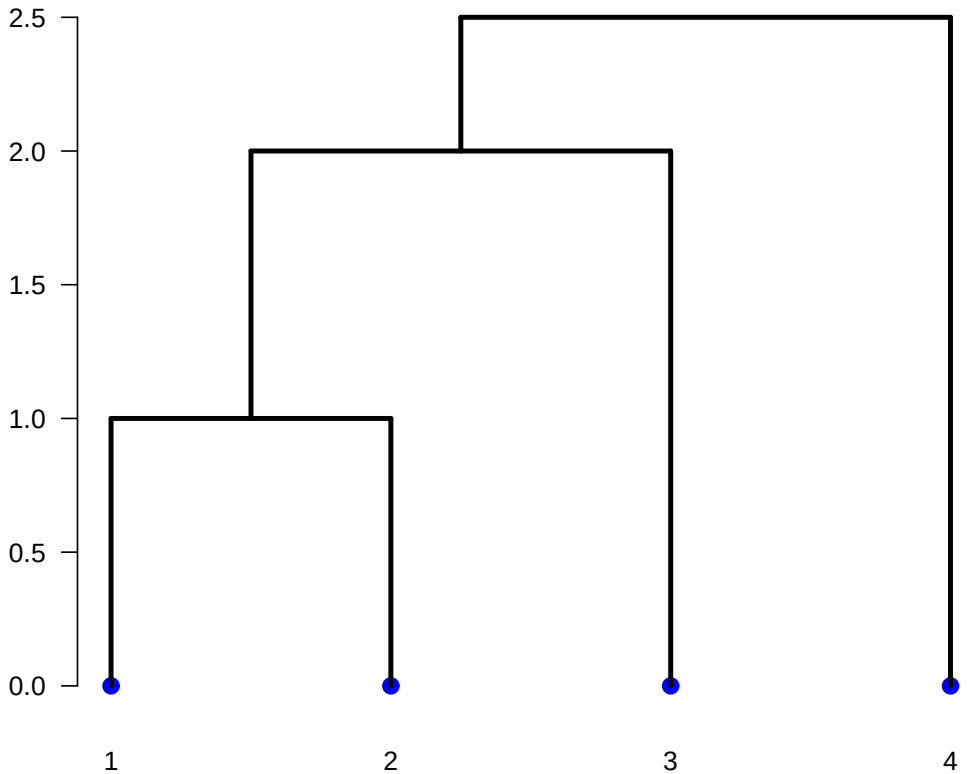
$$d(\mathcal{C}_1, \mathcal{C}_2) = \frac{1}{n_1\, n_2} \sum_{r \in \mathcal{C}_1} \sum_{s \in \mathcal{C}_2} d_{rs}$$

- Este criterio **no** es invariante por tranformaciones monótonas.
- Tiende a formar clusters con poca variabilidad.
- Tiende a formar grupos con igual variabilidad.

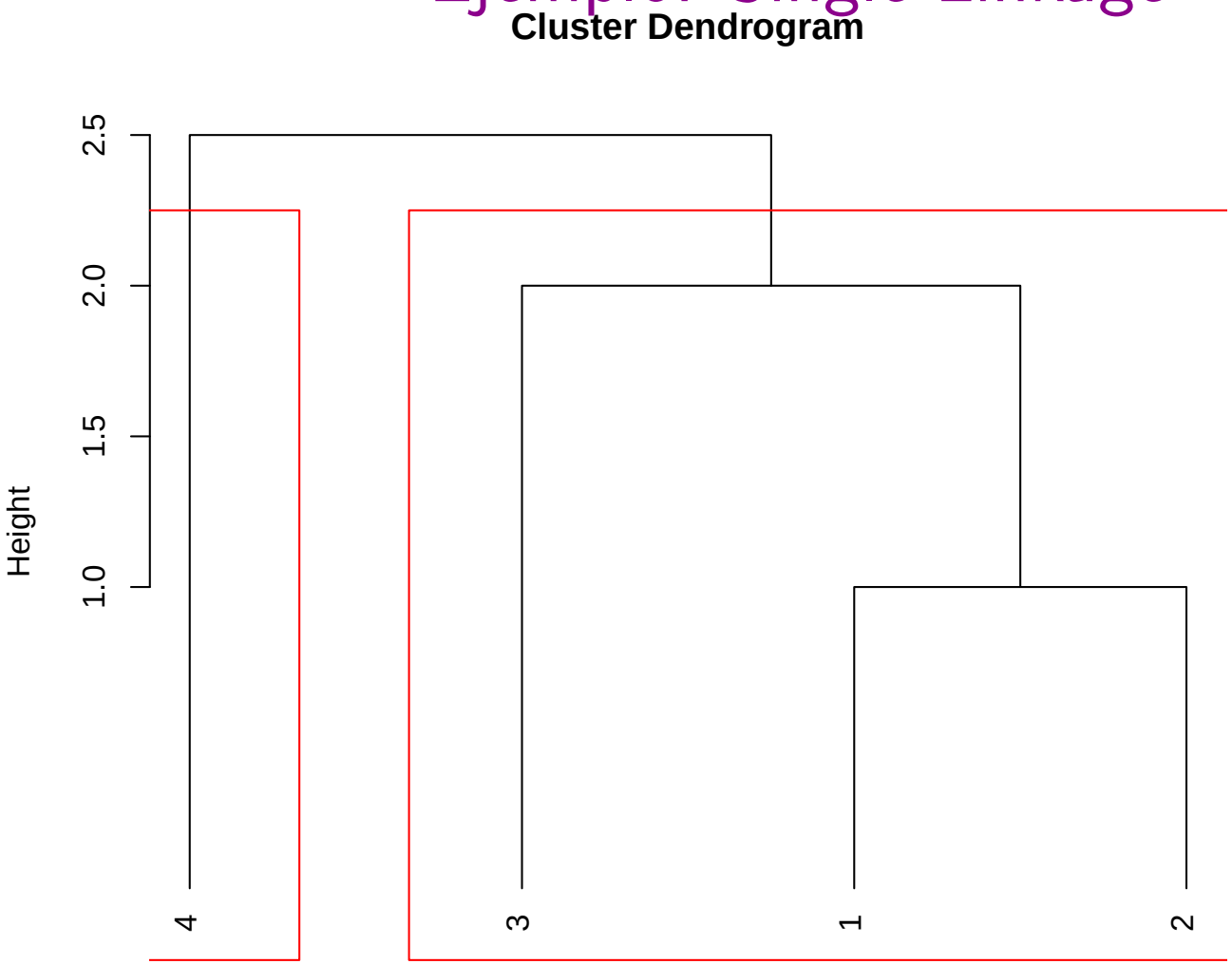
Ejemplo

$$\mathbf{D} = \begin{pmatrix} 0 & 1 & 4 & 2.5 \\ 1 & 0 & 2 & 3 \\ 2 & 2 & 0 & 4 \\ 2.5 & 3 & 4 & 0 \end{pmatrix}$$

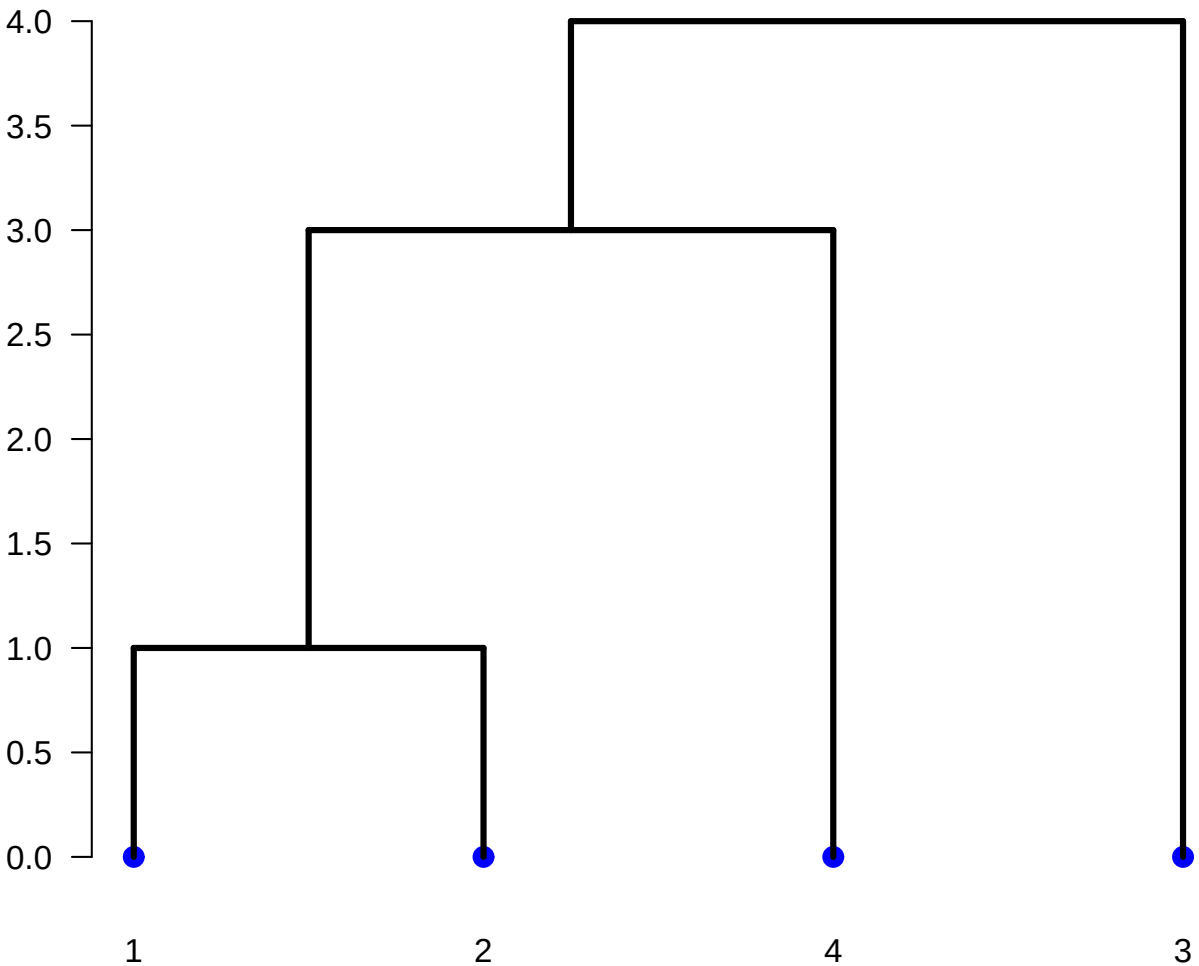
Ejemplo: Single Linkage



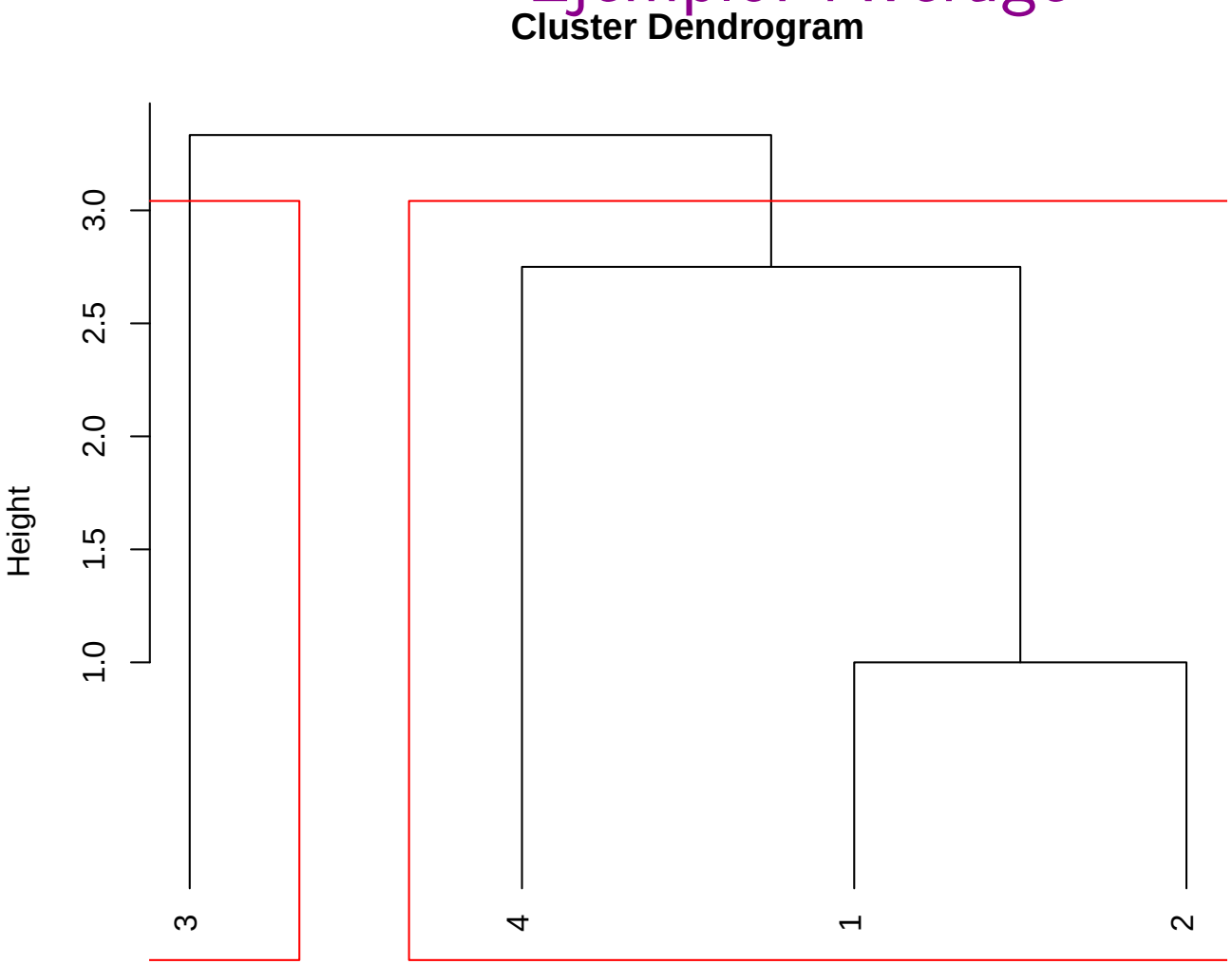
Ejemplo: Single Linkage



Ejemplo: Complete Linkage



Ejemplo: Average



Método del Centroide

Se aplica generalmente con variables continuas cuando se usa como disimilaridad el cuadrado de la distancia euclídea, es decir,

$$d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|^2 .$$

Método del Centroide

Se aplica generalmente con variables continuas cuando se usa como disimilaridad el cuadrado de la distancia euclídea, es decir,

$$d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|^2 .$$

Dados dos clusters $\mathcal{C}_1$ y $\mathcal{C}_2$, la distancia entre ellos es la distancia al cuadrado entre sus centroides

$$d(\mathcal{C}_1, \mathcal{C}_2) = \|\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\|^2 \qquad \bar{\mathbf{x}}_i = \frac{1}{n_i} \sum_{\ell \in \mathcal{C}_i} \mathbf{x}_\ell .$$

Luego, si $\mathcal{C}^{(new)} = \mathcal{C}_1 \cup \mathcal{C}_2$, el centroide de $\mathcal{C}^{(new)}$ es

$$\bar{\mathbf{x}} = \frac{n_1 \bar{\mathbf{x}}_1 + n_2 \bar{\mathbf{x}}_2}{n_1 + n_2} .$$

Método del Centroide

Si $\mathcal{C}^{(new)} = \mathcal{C}_1 \cup \mathcal{C}_2$

$$d(\mathcal{C}^{(new)}, \mathcal{C}_3) = \frac{n_1}{n_1 + n_2} d(\mathcal{C}_1, \mathcal{C}_3) + \frac{n_2}{n_1 + n_2} d(\mathcal{C}_2, \mathcal{C}_3) - \frac{n_1\, n_2}{(n_1 + n_2)^2} d(\mathcal{C}_1, \mathcal{C}_2)$$

Método del Centroide

Si $\mathcal{C}^{(new)} = \mathcal{C}_1 \cup \mathcal{C}_2$

$$d(\mathcal{C}^{(new)}, \mathcal{C}_3) = \frac{n_1}{n_1 + n_2} d(\mathcal{C}_1, \mathcal{C}_3) + \frac{n_2}{n_1 + n_2} d(\mathcal{C}_2, \mathcal{C}_3) - \frac{n_1\, n_2}{(n_1 + n_2)^2} d(\mathcal{C}_1, \mathcal{C}_2)$$

- Puede no dar funciones monótonas de la distancia en cada paso
- Es menos sensible a datos atípicos
- Grupos pequeños pierden identidad al fundirse con los grandes

Método de la mediana

Como el del centroide se aplica a variables continuas y

$$d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|^2 .$$

Si $\mathcal{C}^{(new)} = \mathcal{C}_1 \cup \mathcal{C}_2$, el centro de $\mathcal{C}^{(new)}$ es

$$\overline{\mathbf{x}} = \frac{\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2}{2} .$$

Método de la mediana

Como el del centroide se aplica a variables continuas y

$$d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|^2 .$$

Si $\mathcal{C}^{(new)} = \mathcal{C}_1 \cup \mathcal{C}_2$, el centro de $\mathcal{C}^{(new)}$ es

$$\overline{\mathbf{x}} = \frac{\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2}{2} .$$

$$d(\mathcal{C}^{(new)}, \mathcal{C}_3) = \frac{1}{2}\{d(\mathcal{C}_1, \mathcal{C}_3) + d(\mathcal{C}_2, \mathcal{C}_3)\} - \frac{1}{4} d(\mathcal{C}_1, \mathcal{C}_2)$$

Método de la mediana

Como el del centroide se aplica a variables continuas y

$$d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|^2 .$$

Si $\mathcal{C}^{(new)} = \mathcal{C}_1 \cup \mathcal{C}_2$, el centro de $\mathcal{C}^{(new)}$ es

$$\overline{\mathbf{x}} = \frac{\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2}{2} .$$

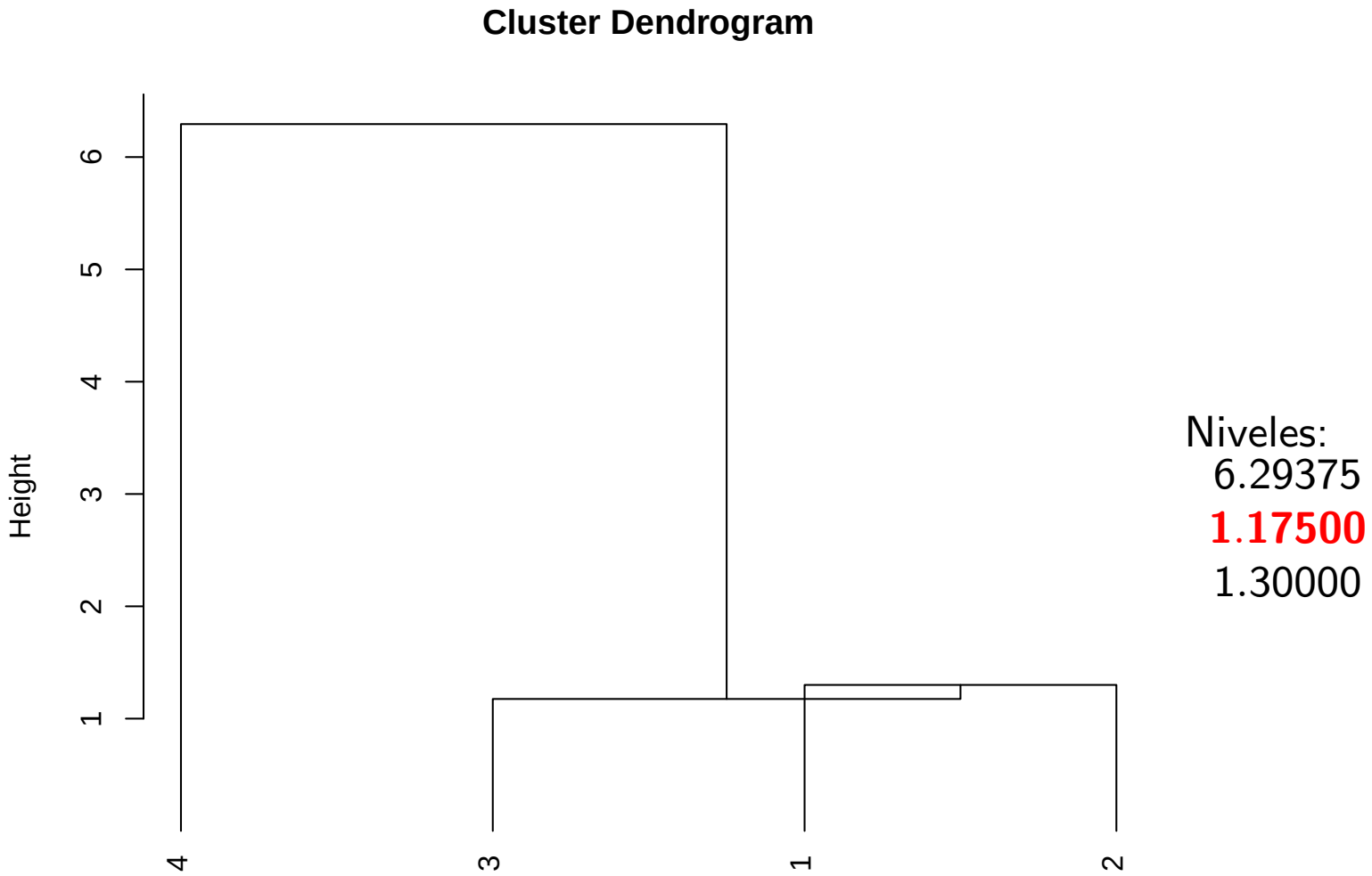
$$d(\mathcal{C}^{(new)}, \mathcal{C}_3) = \frac{1}{2}\{d(\mathcal{C}_1, \mathcal{C}_3) + d(\mathcal{C}_2, \mathcal{C}_3)\} - \frac{1}{4} d(\mathcal{C}_1, \mathcal{C}_2)$$

- Puede no dar funciones monótonas de la distancia en cada paso
- Se introdujo para solucionar la tendencia del método del centroide a que grupos pequeños pierdan su identidad al fundirse con los grandes.

Ejemplo

$$\mathbf{D} = \begin{pmatrix} 0 & 1.3 & 1.4 & 6 \\ & 0 & 1.6 & 5 \\ & & 0 & 8 \\ & & & 0 \end{pmatrix}$$

Ejemplo: Método de la Mediana



Método de Ward

Supongamos tener dividido en K clusters y definamos

$$\mathbf{W} = \sum_{i=1}^{K} \sum_{\ell \in \mathcal{C}_i} (\mathbf{x}_\ell - \overline{\mathbf{x}}_i)(\mathbf{x}_\ell - \overline{\mathbf{x}}_i)^{\mathrm{T}} \qquad \overline{\mathbf{x}}_i = \frac{1}{n_i} \sum_{\ell \in \mathcal{C}_i} \mathbf{x}_\ell$$

- El procedimiento empieza con $K = n$ y por lo tanto, $\mathbf{W} = \mathbf{0}$.
- Luego, se unen los elementos que producen el incremento mínimo de $\mathrm{TR}(\mathbf{W})$, lo que corresponde a buscar los dos puntos más proximos en distancia euclídea.
- Se sigue el proceso sucesivamente

Método de Ward

El incremento al unir $\mathcal{C}_1$ con $\mathcal{C}_2$ en la suma de cuadrados total ($\mathrm{TR}(\mathbf{W})$) es

$$I_{\mathcal{C}_1,\mathcal{C}_2} = \sum_{\ell \in \mathcal{C}_1 \cup \mathcal{C}_2} \|\mathbf{x}_\ell - \bar{\mathbf{x}}\|^2 - \sum_{i=1}^{2} \sum_{\ell \in \mathcal{C}_i} \|\mathbf{x}_\ell - \bar{\mathbf{x}}_i\|^2$$

donde

$$\bar{\mathbf{x}} = \frac{n_1 \bar{\mathbf{x}}_1 + n_2 \bar{\mathbf{x}}_2}{n_1 + n_2}$$

Método de Ward

El incremento al unir $\mathcal{C}_1$ con $\mathcal{C}_2$ en la suma de cuadrados total ($\mathrm{TR}(\mathbf{W})$) es

$$I_{\mathcal{C}_1,\mathcal{C}_2} = \sum_{\ell \in \mathcal{C}_1 \cup \mathcal{C}_2} \|\mathbf{x}_\ell - \bar{\mathbf{x}}\|^2 - \sum_{i=1}^{2} \sum_{\ell \in \mathcal{C}_i} \|\mathbf{x}_\ell - \bar{\mathbf{x}}_i\|^2$$

donde

$$\bar{\mathbf{x}} = \frac{n_1 \bar{\mathbf{x}}_1 + n_2 \bar{\mathbf{x}}_2}{n_1 + n_2}$$

Se unen los clusters que minimizan $I_{\mathcal{C}_1,\mathcal{C}_2}$.

Método de Ward

El incremento al unir $\mathcal{C}_1$ con $\mathcal{C}_2$ en la suma de cuadrados total ($\mathrm{TR}(\mathbf{W})$) es

$$I_{\mathcal{C}_1,\mathcal{C}_2} = \sum_{\ell \in \mathcal{C}_1 \cup \mathcal{C}_2} \|\mathbf{x}_\ell - \overline{\mathbf{x}}\|^2 - \sum_{i=1}^{2} \sum_{\ell \in \mathcal{C}_i} \|\mathbf{x}_\ell - \overline{\mathbf{x}}_i\|^2$$

donde

$$\overline{\mathbf{x}} = \frac{n_1 \overline{\mathbf{x}}_1 + n_2 \overline{\mathbf{x}}_2}{n_1 + n_2}$$

Se unen los clusters que minimizan $I_{\mathcal{C}_1,\mathcal{C}_2}$.

Los grupos que se unen son tales que minimizan la distancia entre sus centros

$$\frac{n_1 n_2}{n_1 + n_2} \|\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2\|^2$$

Método de Ward

- El procedimiento supone que se trabaja con una mezcla de normales con matrices de covarianza $\lambda \mathbf{I}$

- Sensible a outliers.

Arrestos en USA

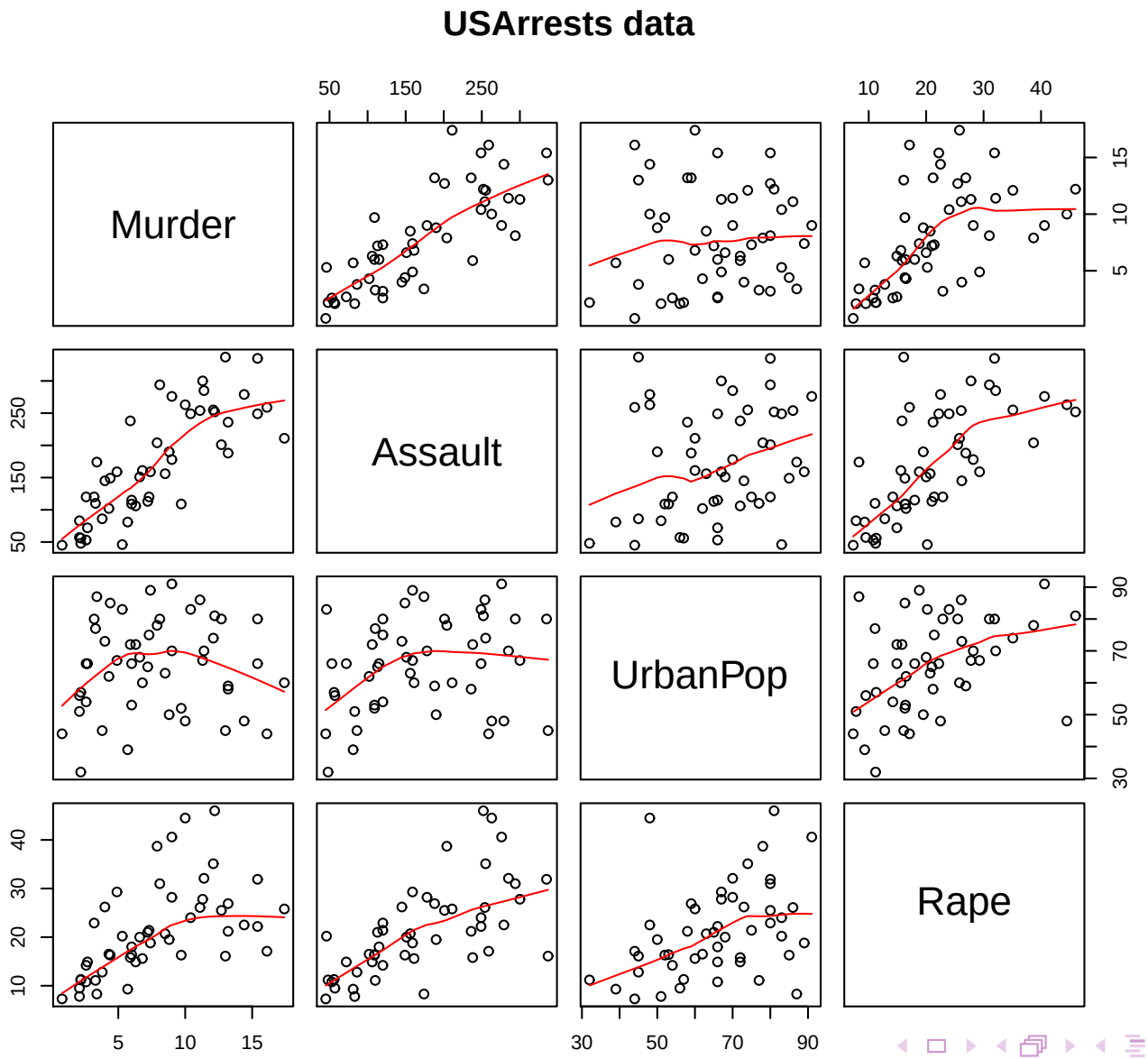
Arrestos por 100.000 residentes por

- asalto
- Asesinato
- Violación

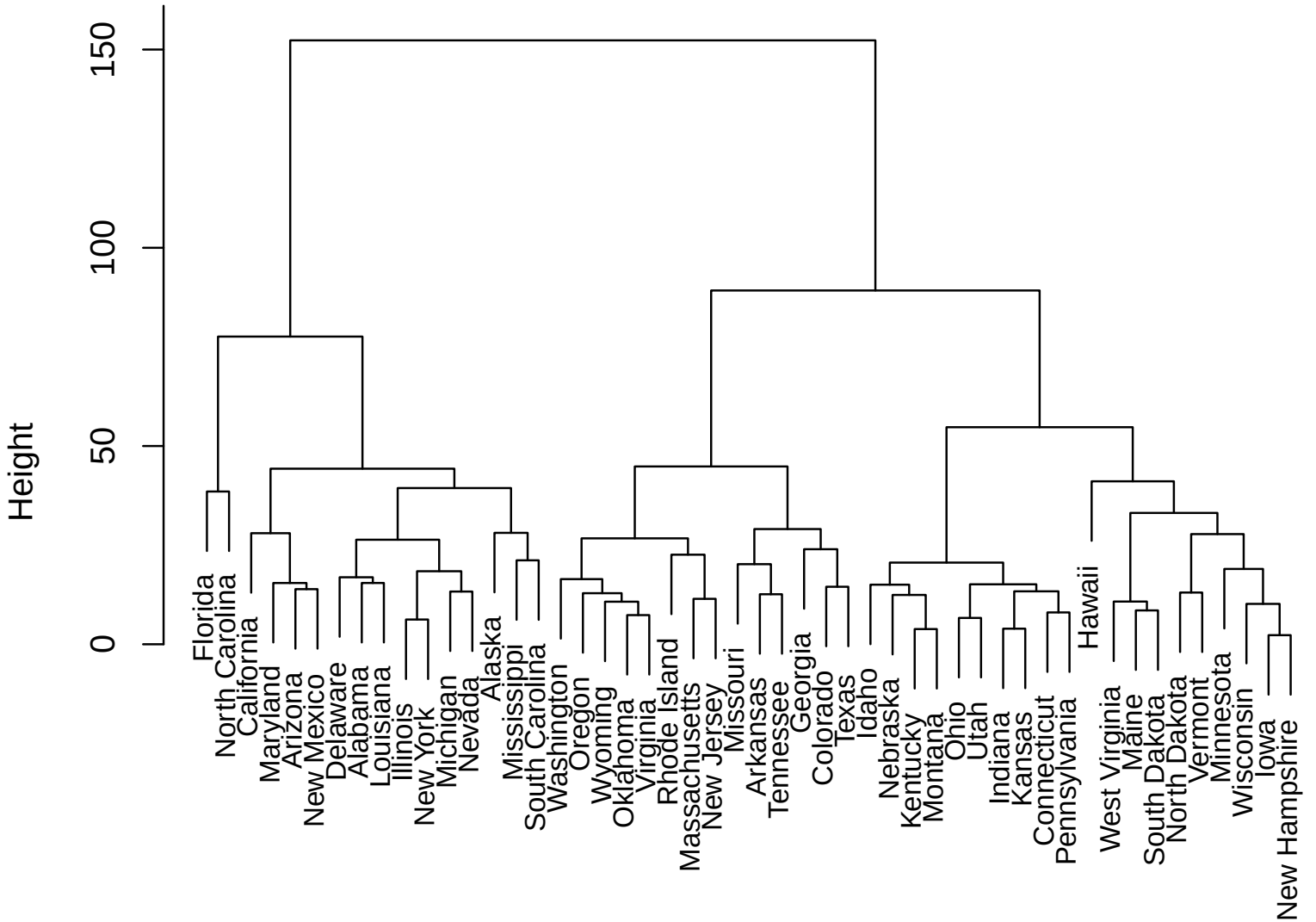
en los 50 estados de USA en 1973.

Se da también la variable población urbana.

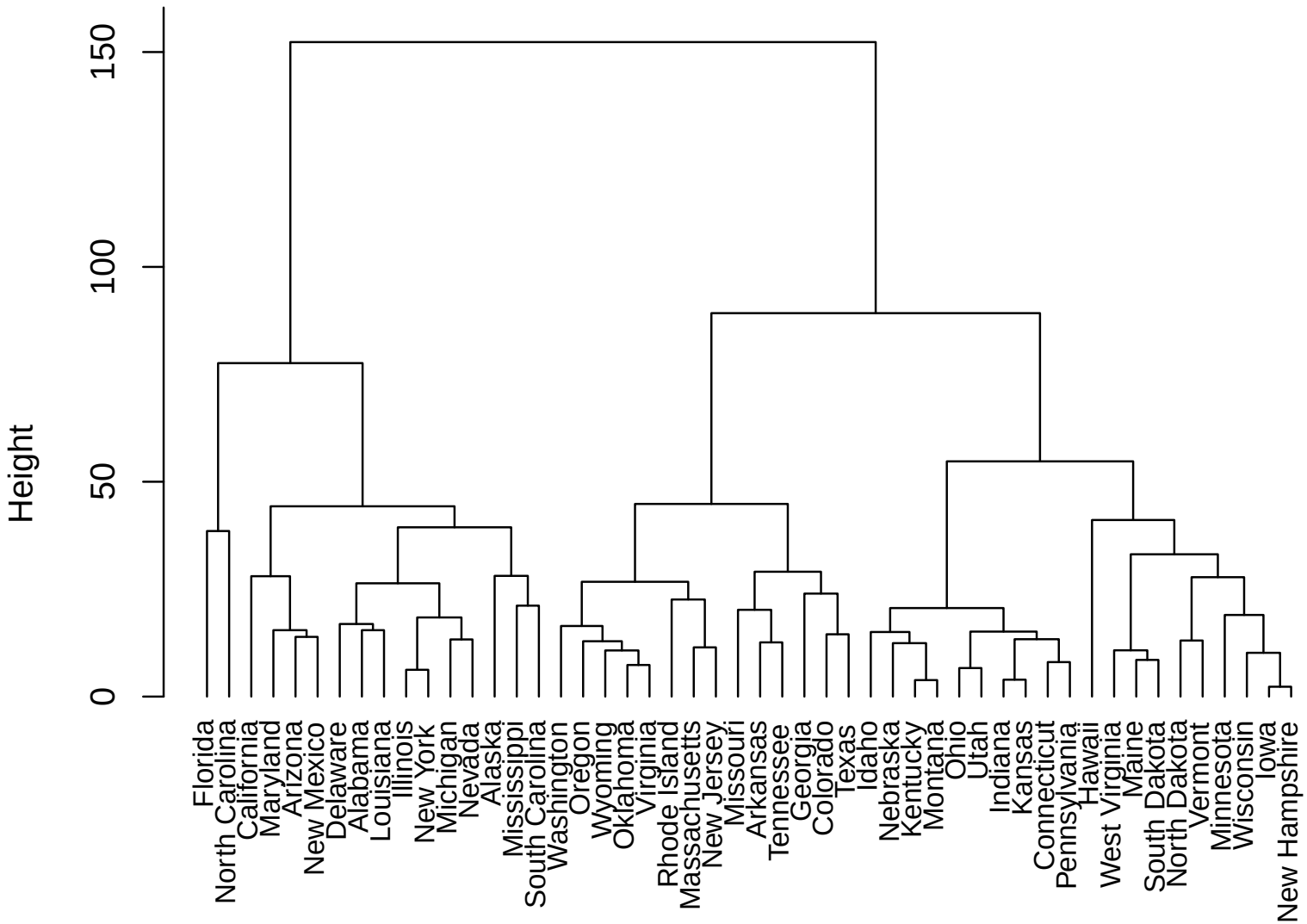
Arrestos en USA



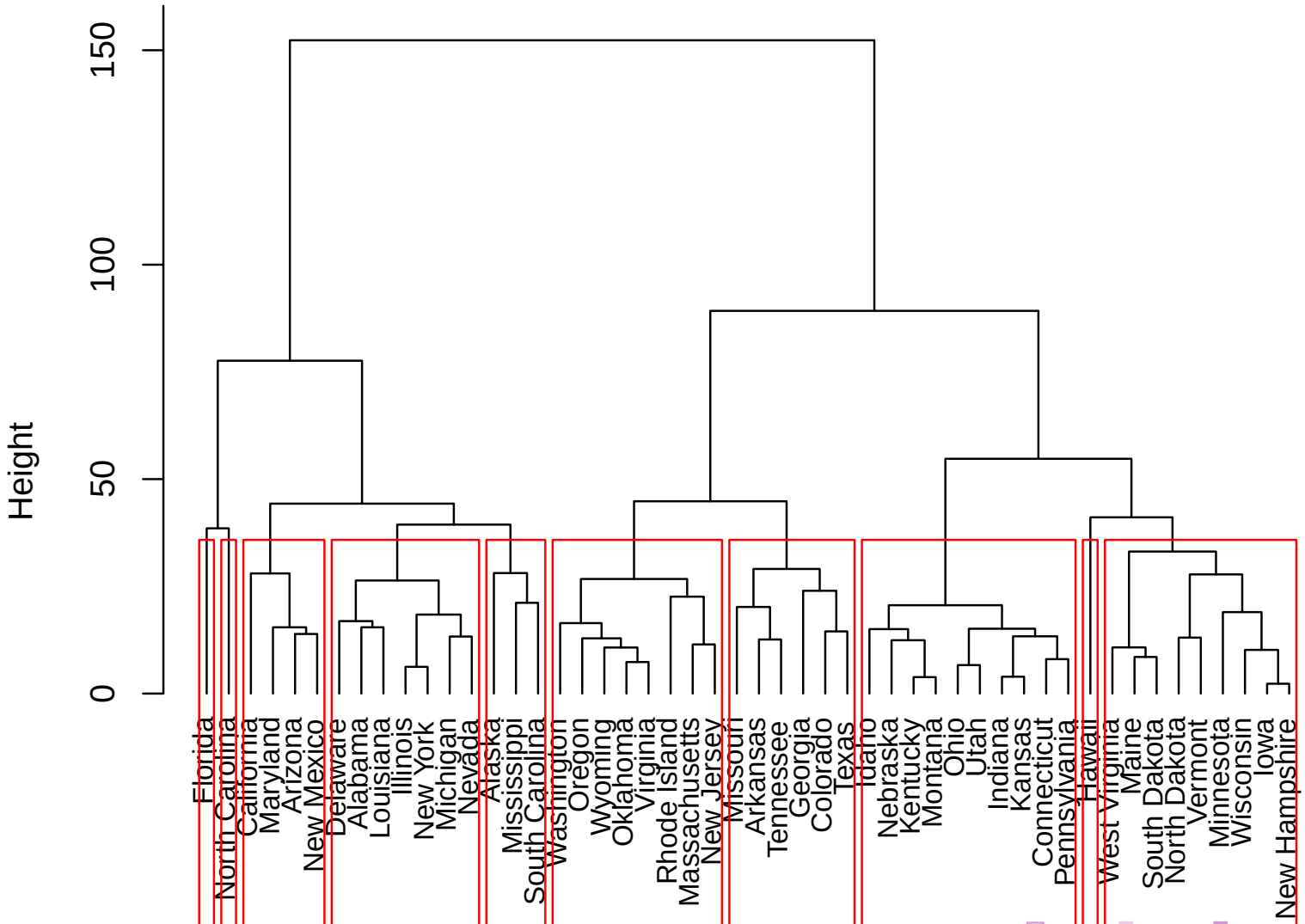
Arrestos en USA: Average Linkage



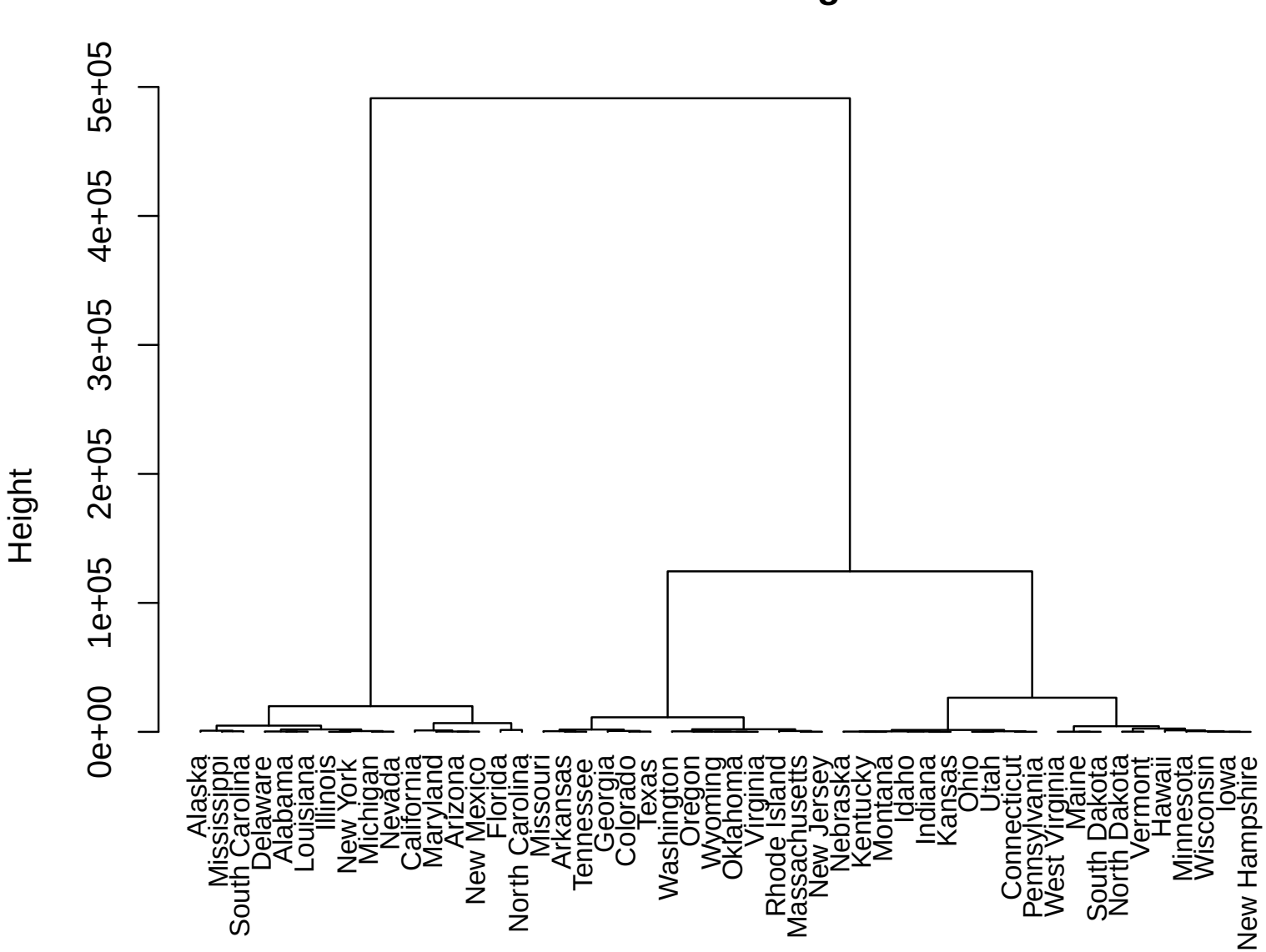
Arrestos en USA: Average Linkage



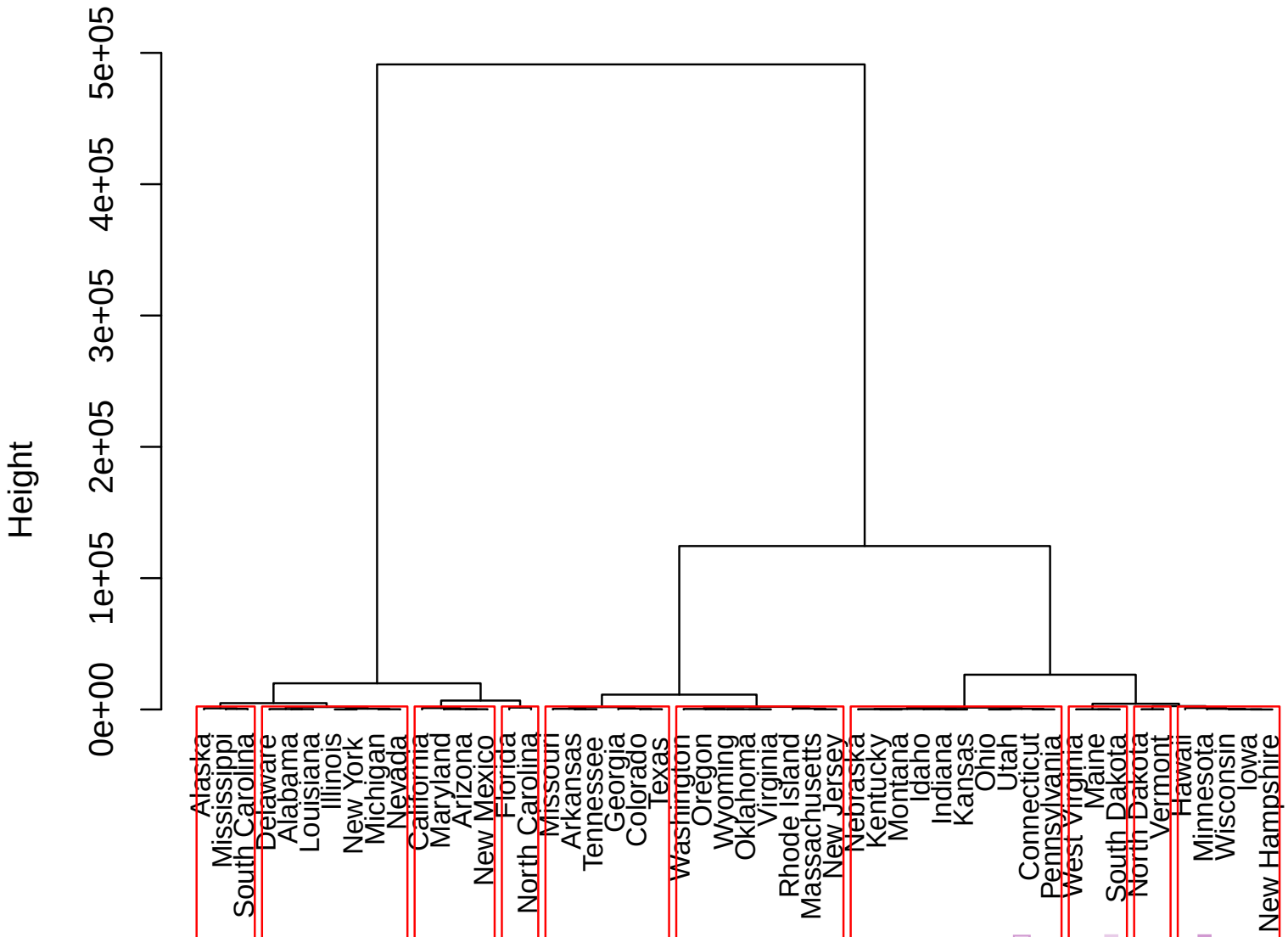
Arrestos en USA: Average Linkage



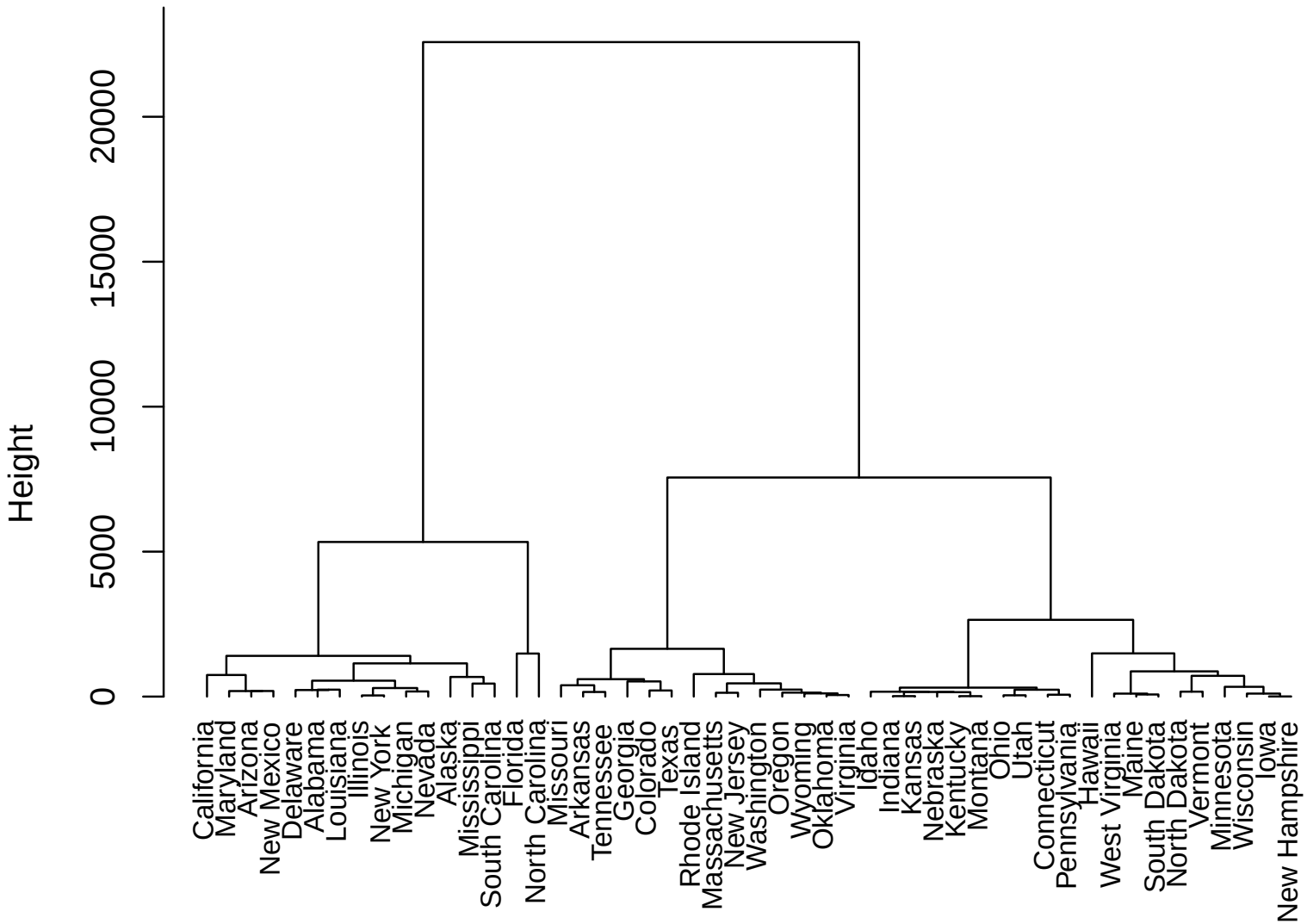
Arrestos en USA: Ward



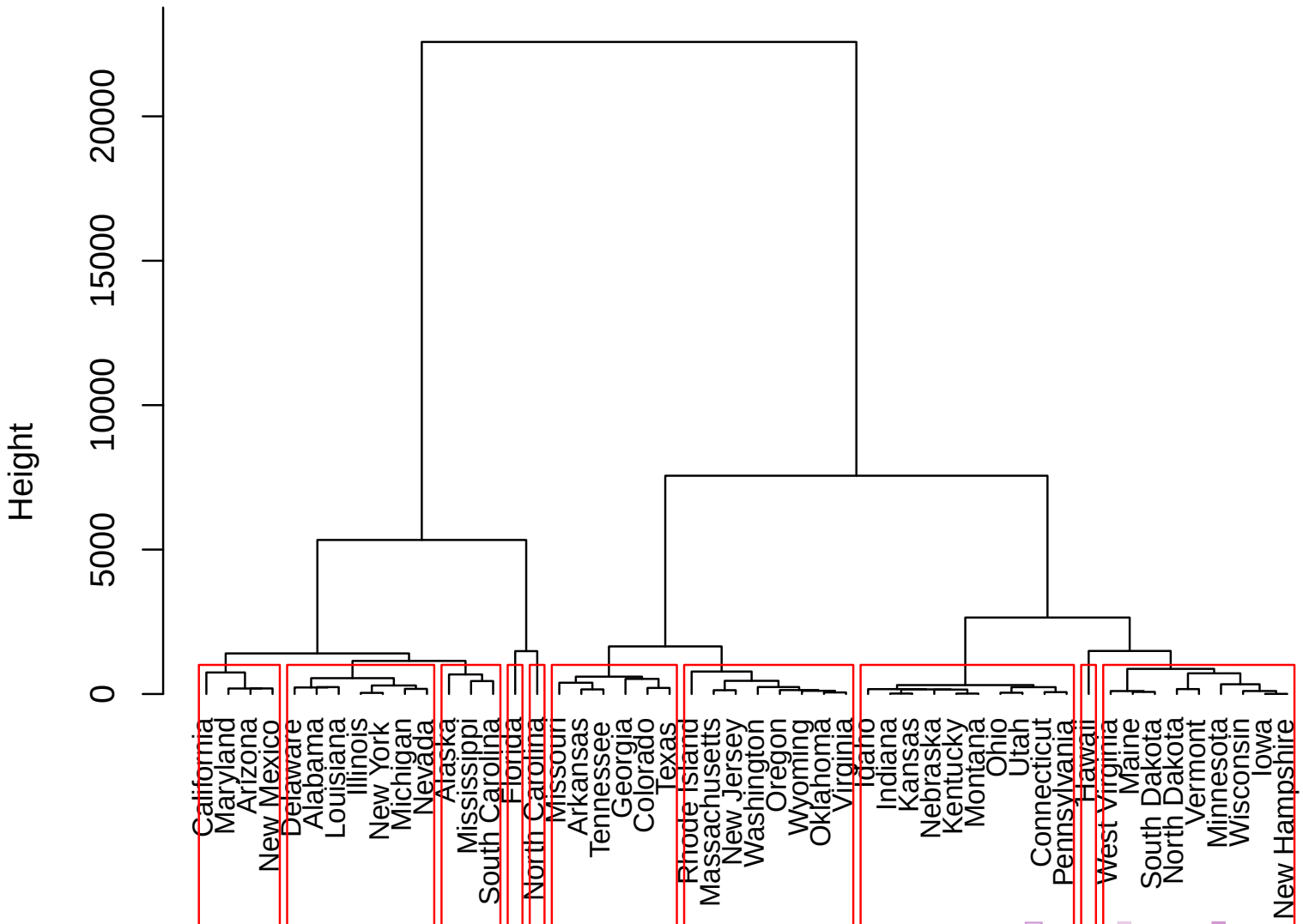
Arrestos en USA: Ward



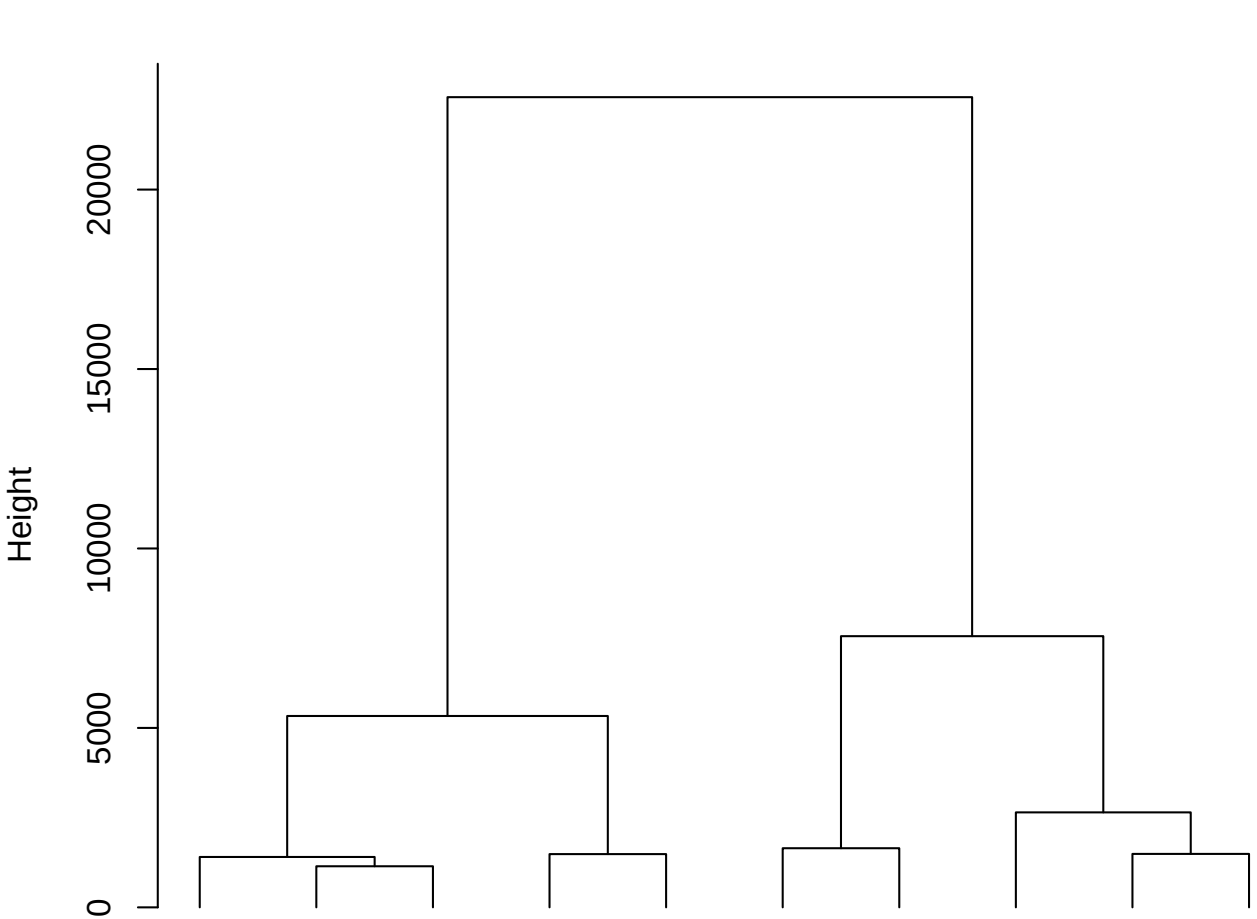
Arrestos en USA: Centroide



Arrestos en USA: Centroide



Arrestos en USA: Centroide



Objetivos

Tenemos $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^p$. Supongamos querer formar k grupos.

Los métodos particionantes buscan $\mathcal{C}_1, \ldots, \mathcal{C}_k$ tales que

- $\#\{\mathcal{C}_j\} > 0$
- $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$

- $\cup_{i=1}^{k} \mathcal{C}_i = \{1, \ldots, n\}$

Objetivos

Tenemos $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^p$. Supongamos querer formar k grupos.

Los métodos particionantes buscan $\mathcal{C}_1, \ldots, \mathcal{C}_k$ tales que

- $\#\{\mathcal{C}_j\} > 0$
- $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$
- $\cup_{i=1}^{k} \mathcal{C}_i = \{1, \ldots, n\}$

El número de posibles particiones es

$$S(n,k) = \frac{1}{k!} \sum_{j=1}^{k} (-1)^j \left(\begin{array}{c} k \\ j \end{array} \right) (k-j)^n \approx \frac{k^n}{k!}$$

Objetivos

Tenemos $\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^p$. Supongamos querer formar k grupos.

Los métodos particionantes buscan $\mathcal{C}_1, \ldots, \mathcal{C}_k$ tales que

- $\#\{\mathcal{C}_j\} > 0$
- $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$
- $\cup_{i=1}^{k} \mathcal{C}_i = \{1, \ldots, n\}$

El número de posibles particiones es

$$S(n,k) = \frac{1}{k!} \sum_{j=1}^{k} (-1)^j \left(\begin{array}{c} k \\ j \end{array} \right) (k-j)^n \approx \frac{k^n}{k!}$$

Por ejemplo, $S(19,3) = 1.9 \times 10^8$.

Si k no se especifica tenemos $T = \sum_{k=1}^{n} S(n,k)$ configuraciones. Para $n = 25$, $T > 4 \times 10^{18}$.

Fundamentos

Supongamos tener k grupos $\mathcal{C}_1, \ldots, \mathcal{C}_k$ y que usamos la distancia euclídea.

Definamos $n_j = \#\mathcal{C}_j$

- El centro del grupo $\mathcal{C}_j$ como

$$\bar{\mathbf{x}}_j = \frac{1}{n_j} \sum_{\ell \in \mathcal{C}_j} \mathbf{x}_\ell$$

- La varianza del grupo $\mathcal{C}_j$

$$e(\mathcal{C}_j) = \sum_{\ell \in \mathcal{C}_j} \|\mathbf{x}_\ell - \bar{\mathbf{x}}_j\|^2$$

- El cuadrado de la distancia de $\mathbf{x}_i$ al centro más próximo

$$m(\mathbf{x}_i) = \min_{1 \leq \ell \leq k} \|\mathbf{x}_i - \bar{\mathbf{x}}_\ell\|^2$$

Fundamentos

El algoritmo de $k-$medias require las siguientes etapas

1. Seleccione K puntos $\mathbf{x}_j^{(0)}$, $1 \leq j \leq K$ como centros iniciales.
   Puede hacerse de alguna de las siguientes formas:
   1.1 Divida aleatoriamente las observaciones en K grupos y tome $\mathbf{x}_j^{(0)}$ como el centro del grupo j.

Fundamentos

El algoritmo de $k-$medias require las siguientes etapas

1. Seleccione K puntos $\mathbf{x}_j^{(0)}$, $1 \leq j \leq K$ como centros iniciales.
   Puede hacerse de alguna de las siguientes formas:
   - 1.1 Divida aleatoriamente las observaciones en K grupos y tome $\mathbf{x}_j^{(0)}$ como el centro del grupo j.
   - 1.2 Tome como centros los puntos más alejados entre sí.

Fundamentos

El algoritmo de $k-$medias require las siguientes etapas

1. Seleccione K puntos $\mathbf{x}_j^{(0)}$, $1 \leq j \leq K$ como centros iniciales.
   Puede hacerse de alguna de las siguientes formas:
   - 1.1 Divida aleatoriamente las observaciones en K grupos y tome $\mathbf{x}_j^{(0)}$ como el centro del grupo j.
   - 1.2 Tome como centros los puntos más alejados entre sí.
   - 1.3 Contruya grupos iniciales *con información a priori* y calcule sus centros.

Fundamentos

El algoritmo de $k-$medias require las siguientes etapas

1. Seleccione K puntos $\mathbf{x}_j^{(0)}$, $1 \leq j \leq K$ como centros iniciales.
   Puede hacerse de alguna de las siguientes formas:
   - 1.1 Divida aleatoriamente las observaciones en K grupos y tome $\mathbf{x}_j^{(0)}$ como el centro del grupo j.
   - 1.2 Tome como centros los puntos más alejados entre sí.
   - 1.3 Contruya grupos iniciales *con información a priori* y calcule sus centros.
   - 1.4 Seleccione centros iniciales *con información a priori*.

Fundamentos

El algoritmo de $k-$medias require las siguientes etapas

1. Seleccione K puntos $\mathbf{x}_j^{(0)}$, $1 \leq j \leq K$ como centros iniciales.
   Puede hacerse de alguna de las siguientes formas:
   - 1.1 Divida aleatoriamente las observaciones en K grupos y tome $\mathbf{x}_j^{(0)}$ como el centro del grupo j.
   - 1.2 Tome como centros los puntos más alejados entre sí.
   - 1.3 Contruya grupos iniciales *con información a priori* y calcule sus centros.
   - 1.4 Seleccione centros iniciales *con información a priori*.
2. Calcule $\|\mathbf{x}_i - \overline{\mathbf{x}}_\ell\|^2$ y asigne $\mathbf{x}_i$ al grupo grupo ℓ cuyo centro es más cercano, o sea, si

$$m(\mathbf{x}_i) = \|\mathbf{x}_i - \overline{\mathbf{x}}_\ell\|^2$$

   La asignación es secuencial y al incorporar un nuevo elemento al grupo se recalculan el nuevo centro.

Fundamentos

3. Defina como criterio de optimalidad el que minimiza

$$\mathcal{V}_K = \sum_{j=1}^{K} e(\mathcal{C}_j) = \sum_{i=1}^{n} m(\mathbf{x}_i)$$

4. Verifique si reasignando alguno de los $\mathbf{x}_i$ mejora el criterio.
5. Repita 2,3,4 hasta que no haya mas cambios.

Fundamentos

3. Defina como criterio de optimalidad el que minimiza

$$\mathcal{V}_K = \sum_{j=1}^{K} e(\mathcal{C}_j) = \sum_{i=1}^{n} m(\mathbf{x}_i)$$

4. Verifique si reasignando alguno de los $\mathbf{x}_i$ mejora el criterio.
5. Repita 2,3,4 hasta que no haya mas cambios.

Observemos que si

$$\mathbf{W} = \sum_{i=1}^{K} \sum_{\ell \in \mathcal{C}_i} (\mathbf{x}_\ell - \overline{\mathbf{x}}_i)(\mathbf{x}_\ell - \overline{\mathbf{x}}_i)^{\mathrm{T}} \qquad \overline{\mathbf{x}}_i = \frac{1}{n_i} \sum_{\ell \in \mathcal{C}_i} \mathbf{x}_\ell$$

$$\mathbf{B} = \sum_{i=1}^{K} n_i (\overline{\mathbf{x}}_i - \overline{\mathbf{x}})(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})^{\mathrm{T}} \qquad \overline{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{K} n_i \overline{\mathbf{x}}_i$$

Criterio de la traza

$$\mathcal{V}_K = \mathrm{TR}(\mathbf{W})$$

- El criterio equivale a minimizar la traza de $\mathbf{W}$ y se llama *criterio de la traza*.

- Equivalentemente, el criterio maximiza $\mathrm{TR}(\mathbf{B})$.

Criterio de la traza

$$\mathcal{V}_K = \mathrm{TR}(\mathbf{W})$$

- El criterio equivale a minimizar la traza de $\mathbf{W}$ y se llama *criterio de la traza*.
- Equivalentemente, el criterio maximiza $\mathrm{TR}(\mathbf{B})$.
- El *criterio de la traza mínima* forma grupos con medias separadas y funciona bien si las matrices de covarianza son de la forma $\lambda_j \mathbf{I}_p$.

Hay otros criterios que minimizan $\det(\mathbf{W})$ o maximizan $\mathrm{TR}(\mathbf{W}^{-1}\mathbf{B})$.

Elijo K

$$\mathcal{V}_K = \mathrm{TR}(\mathbf{W})$$

Una opción es estudiar la reducción de variabilidad al considerar $K+1$ grupos en lugar de K.

Sea

$$F = \frac{\mathcal{V}_K - \mathcal{V}_{K+1}}{\mathcal{V}_{K+1}} (n-K-1)$$

- Se compara F con el percentil de una $\mathcal{F}_{p,p(n-K-1)}$.

Elijo K

$$\mathcal{V}_K = \mathrm{TR}(\mathbf{W})$$

Una opción es estudiar la reducción de variabilidad al considerar $K+1$ grupos en lugar de K.

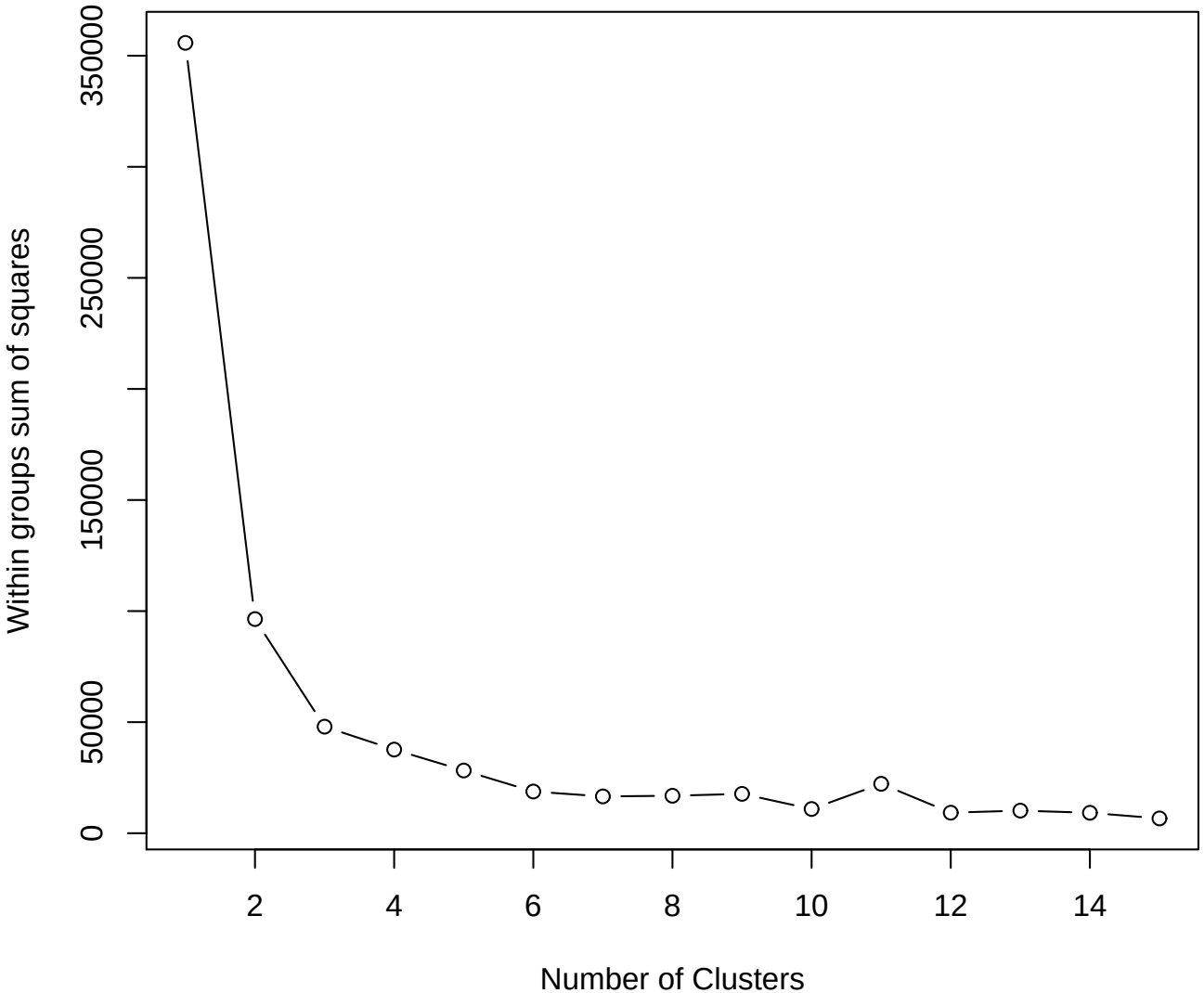
Sea

$$F = \frac{\mathcal{V}_K - \mathcal{V}_{K+1}}{\mathcal{V}_{K+1}} (n-K-1)$$

- Se compara F con el percentil de una $\mathcal{F}_{p,p(n-K-1)}$. **No tiene buena justificación ya que los grupos no tienen porque ser gaussianos**

- Hartigan (1975) sugiere considerar $K+1$ grupos en lugar de K si $F > 10$

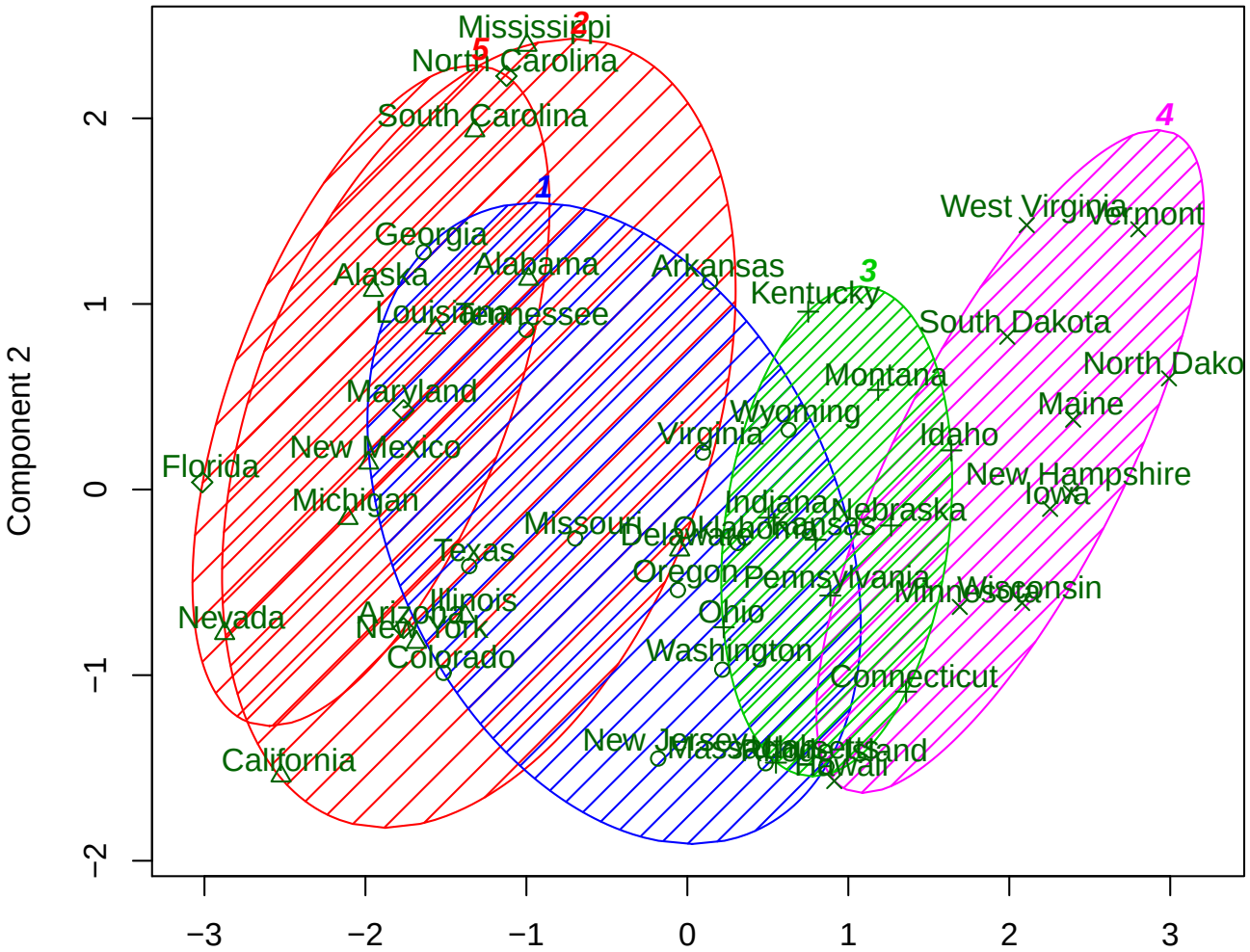
Arrestos en USA: $k-$medias



Arrestos en USA: $k-$medias

Grupo	Asesinato	Asalto	Población	Violacón
1	8.2143	173.2857	70.6429	22.8429
2	11.7667	257.9167	68.4167	28.9333
3	5.5900	112.4000	65.6000	17.2700
4	2.9500	62.7000	53.9000	11.5100
5	11.9500	316.5000	68.0000	26.7000

Arrestos en USA: $k-$medias



Arrestos en USA: $k-$medias