Agrupamiento *(Clustering)*

Cuando agrupamos las observaciones de un conjunto de datos buscamos 1) dividirlas en grupos distintos para que las observaciones dentro de cada grupo sean muy similares entre sí, 2) que las observaciones de diferentes grupos sean muy diferentes entre sí. Por supuesto, para concretar esto, debemos definir qué significa que dos o más observaciones sean similares o diferentes. De hecho, esta es a menudo una consideración específica del dominio debe hacerse en base al conocimiento de los datos que se estudian.

Agrupamiento de k-medias (*k-means Clustering*)

$C_1,\ldots, C_K$ denotan conjuntos de observaciones . Estos conjuntos satisfacen dos propiedades:

1. $C_1 \cup C_2 \cup \ldots \cup C_K = \{1,\ldots, n\}$ cada observación pertenece a al menos uno de los K conjuntos.
2. $C_k \cap C_{k'} = \emptyset$ para todo $k \neq k'$. En otras palabras, los grupos no son superpuestos: ninguna observación pertenece a más de un grupo.

La variación dentro del grupo para el grupo C_k es una medida $\boldsymbol{W}(C_k)$ de cuánto se diferencian entre las observaciones dentro de un grupo. Por eso , para $C_1,\ldots,C_K$, deberemos minimizar el problema:

$$\left\{\sum_{k=1}^{K} W(C_k)\right\} \quad \textbf{(4)}$$

Siendo la $W(C_k)$ la variación dentro de un grupo que definiremos como:

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \qquad \textbf{(5)}$$

donde $|C_k|$ denota el número de observaciones en el grupo *k* . En otras palabras, la variación dentro del grupo para el grupo *k* es la suma de todas las distancias euclidianas al cuadrado por pares entre las observaciones en el grupo *k*, dividida por el número total de observaciones en el grupo *k.*

La combinación de (4) y (5) da el problema de optimización que define el *clustering K-means*

$$\underset{C_1,\ldots,C_K}{\textit{minimizar}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\} \qquad \textbf{(6)}$$

Este es un problema muy difícil de resolver con precisión, ya que hay cerca de K^n formas de particionar *n* observaciones en *K* grupos. ¡Este es un número enorme a menos que *K* y *n* sean pequeños! Afortunadamente, se puede demostrar que un algoritmo muy simple proporciona un óptimo local, una solución bastante buena, al problema de optimización de K-medias.

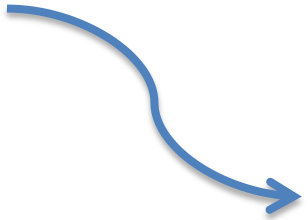
Algoritmo K-medias de agrupamiento

1. Asigne a cada una de las observaciones un número aleatoria de 1 a *K*
   Estos sirven como asignaciones iniciales de grupo para las *n* observaciones.

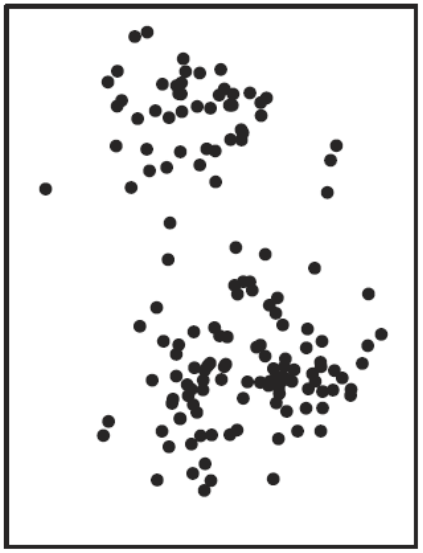
2. Iterar hasta que las asignaciones de grupo **dejen** de cambiar:
   a. Para cada uno de los K grupos, calcule el **centroide** del grupo. El centroide del grupo *k* es el vector de las medias de las p variables predictoras para las observaciones del grupo *k*.
   b. Asigne cada observación al grupo cuyo centroide esté más **cercano** (distancia euclidiana).

Debido a que el algoritmo de K-medias encuentra un óptimo local en lugar de global para la minimización de la fórmula (6), los resultados obtenidos dependerán de la asignación de grupo inicial (aleatoria) de cada observación en el Paso 1 del algoritmo

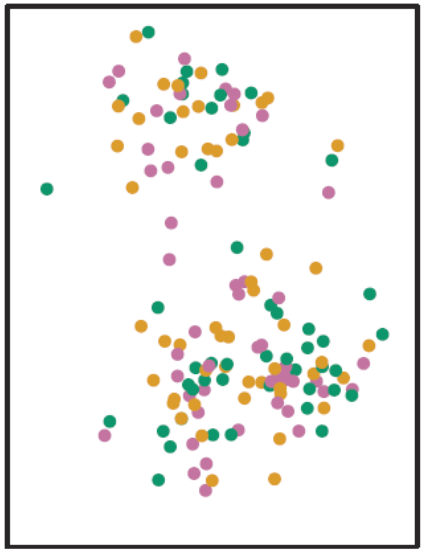
Veámoslo gráficamente



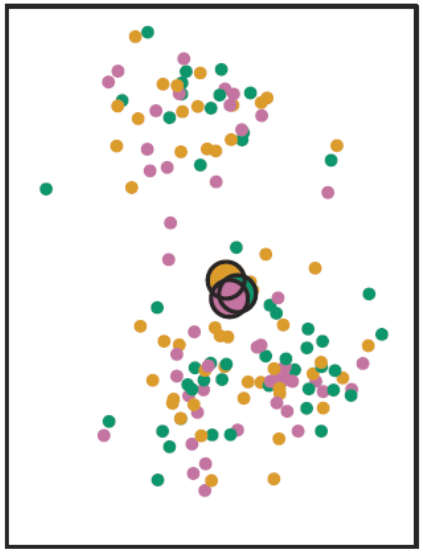
Dataset



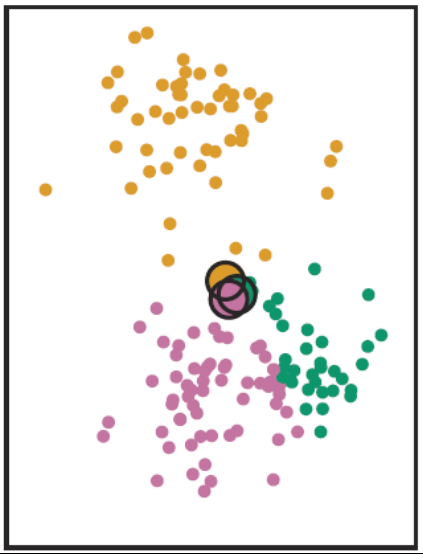
Paso 1



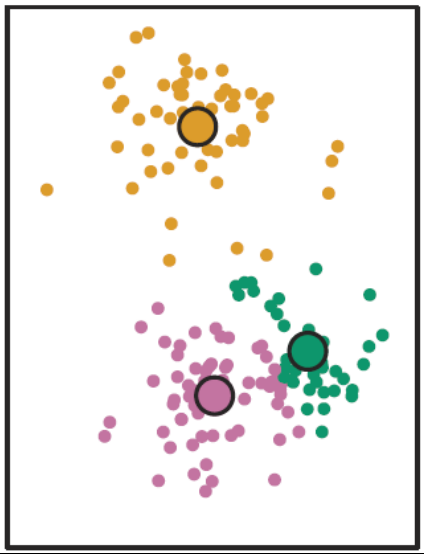
Iteración 1, Paso 2a



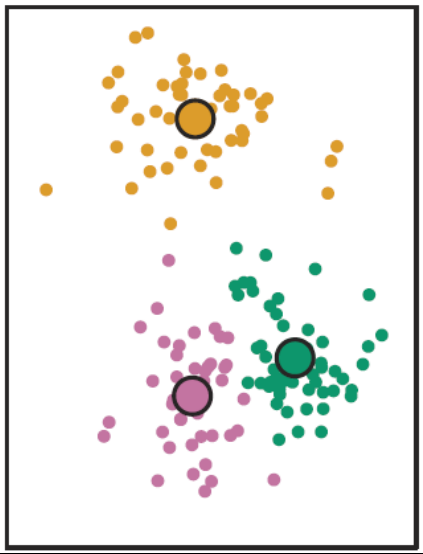
Iteración 1, Paso 2b



Iteración 2, Paso 2a

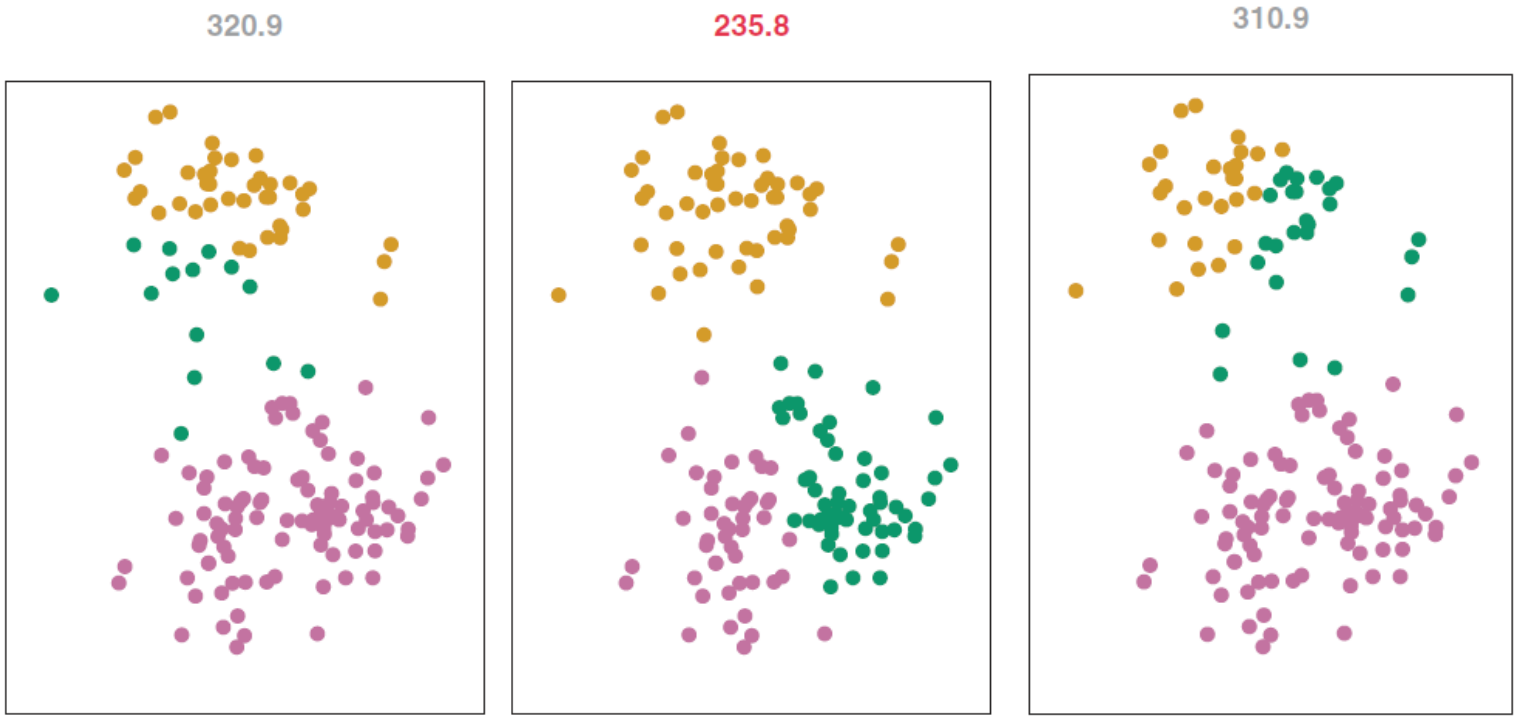


Resultado Final



Centro superior: en el paso 1 del algoritmo, cada observación se asigna al azar a un grupo. Arriba derecha: en el Paso 2 (a), se calculan los centroides del grupo. Estos se muestran como grandes círculos de colores. Inicialmente, los centroides se superponen casi completametamente porque las asignaciones de grupo iniciales se eligieron al azar. Abajo izquierda: en el Paso 2 (b), cada observación se asigna al centroide más cercano. Abajo Centro: el paso 2 (a) es ejecutado otra vez conduciendo a nuevos centroides. Abajo derecha: los resultados después de diez iteraciones.

Veamos el los resultado obtenidos en el mismo dataset con distintas asignaciones aleatorias a grupos en el paso 1 del algoritmo, pero ahora con k=3 (o sea 3 grupos o *clusters*)



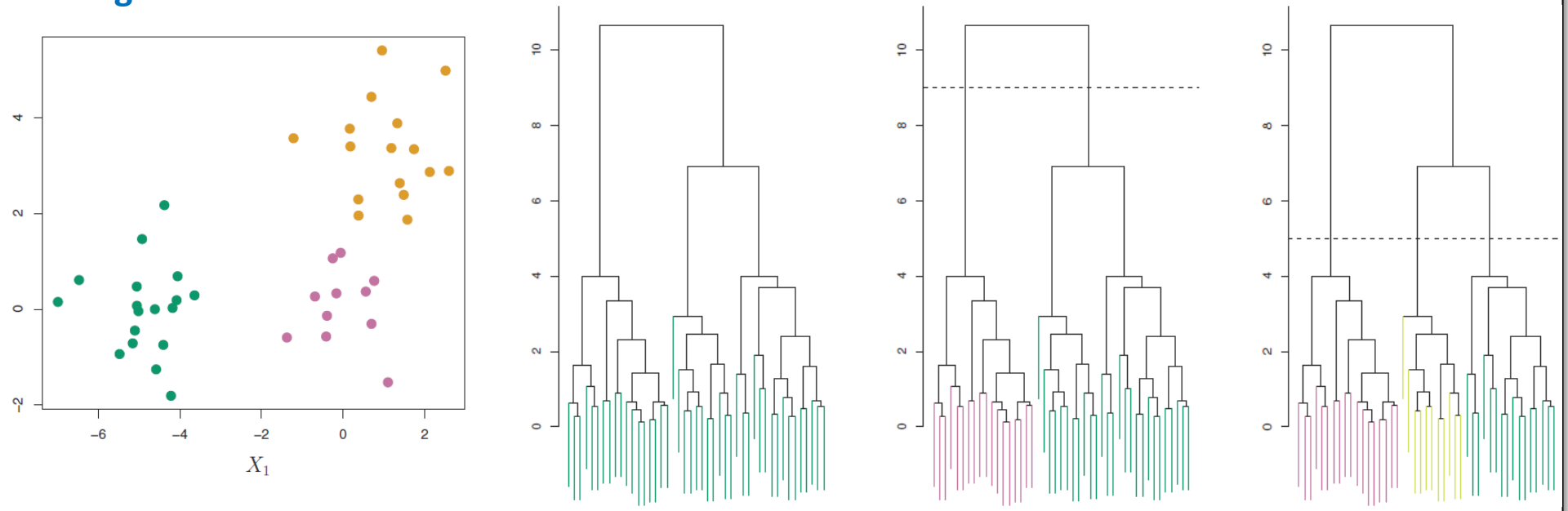
$$\underset{C_1,\ldots,C_K}{\textit{minimizar}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\} \quad \textbf{(6)}$$

Se obtuvieron tres óptimos locales diferentes de la ecuación (6), uno de los cuales dio como resultado un menor valor de ella por lo que proporciona una mejor separación entre los grupos. El etiqueado en rojo indica la mejor solución a la ecuación 6, con un valor objetivo de 235.8

Agrupamiento jerárquico (*Hierarchical Clustering*)

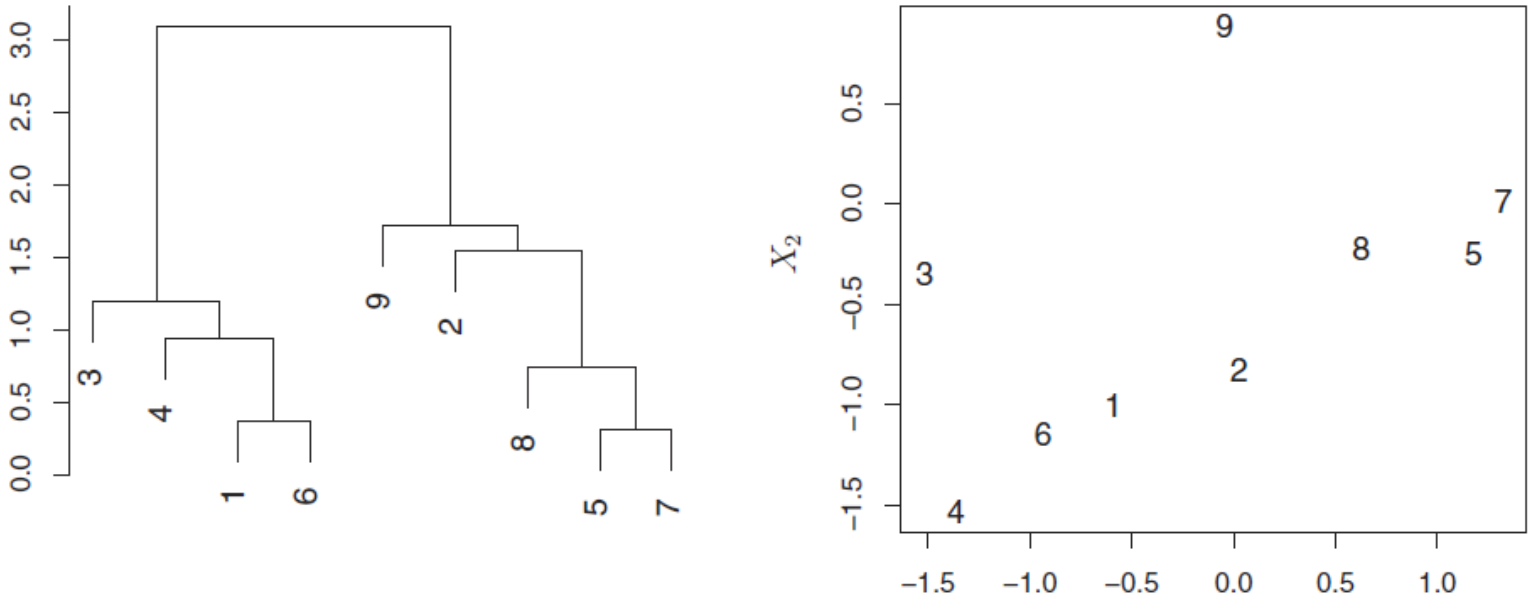
Una desventaja la agrupación de K-medias es que requiere que especifiquemos previamente la cantidad de agrupaciones. La agrupación jerárquica es un enfoque alternativo que no requiere que nos comprometamos con una elección particular de K.

Dendogramas



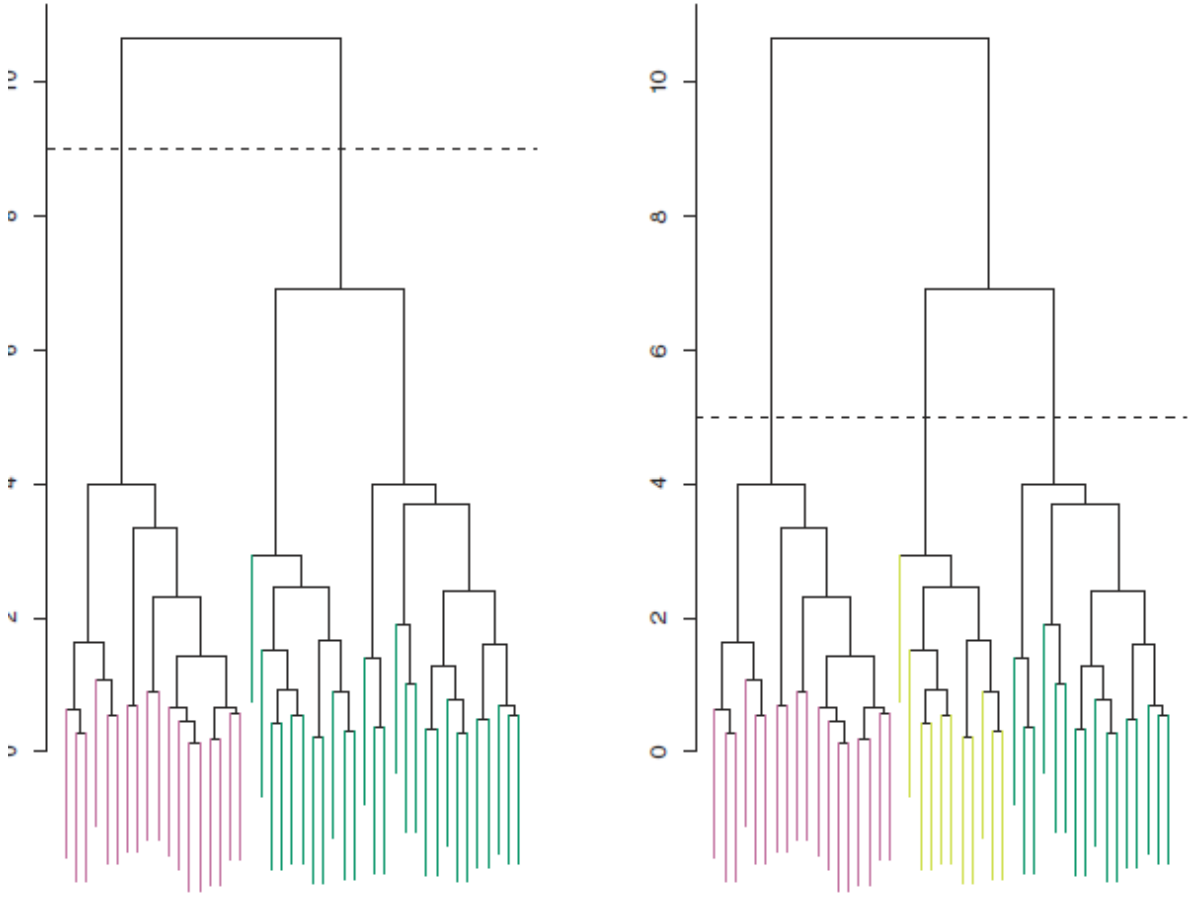
Cada hoja de los dendrogramas de la derecha representa una de las 45 observaciones de la figura izquierda. Sin embargo, a medida que avanzamos por el árbol, algunas hojas comienzan a fundirse en ramas. Estos corresponden a observaciones que son similares entre sí ***medida en distancia euclidiana***

Para cualquier par de observaciones, podemos buscar el punto en el árbol donde las ramas que las contienen se fusionan primero. La altura de esta fusión, medida en el eje vertical, indica qué tan diferentes son las dos observaciones. Por lo tanto, las observaciones que se fusionan en la parte inferior del árbol son bastante similares entre sí , mientras que las observaciones que se fusionan cerca de la parte superior del árbol tenderán a ser bastante diferentes.



Las observaciones 5 y 7 son bastante similares entre sí, al igual que las observaciones 1 y 6. Sin embargo, la observación 9 no es más similar a la observación 2 que a las observaciones 8, 5 y 7, aunque las observaciones 9 y 2 están muy juntas en términos de distancia horizontal en el dendograma. Esto se debe a que las observaciones 2, 8, 5 y 7 se fusionan con la observación 9 a la misma altura, aproximadamente 1.8.

Si hacemos un corte horizontal a través del dendrograma,los distintos conjuntos de observaciones debajo del corte pueden interpretarse como agrupamientos (clusters). En el gráfico de la izq al cortar el dendrograma a una altura de nueve obtenemos dos grupos (k=2), que se muestran en distintos colores . En el gráfico de la derecha, cortar el dendrograma a una altura de cinco da como resultado una agrupación en tres grupos (k=3)



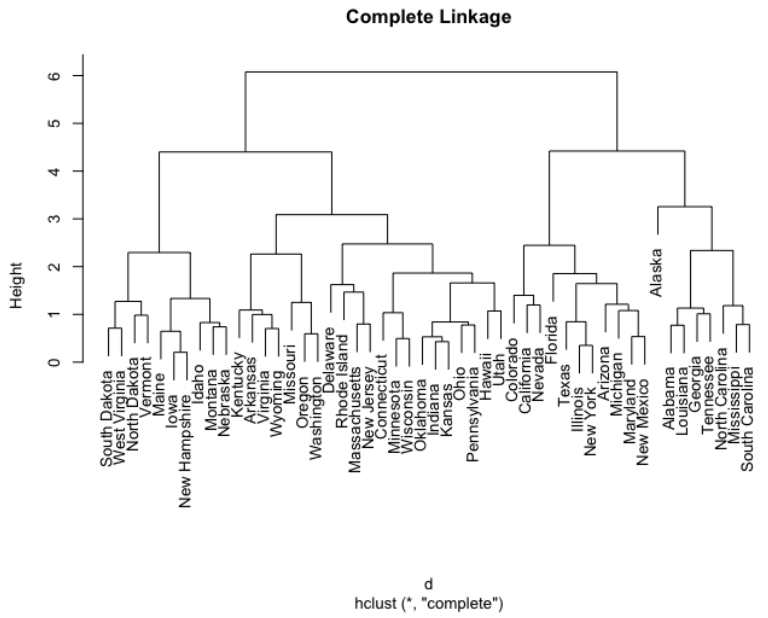
Algoritmo de Agrupamiento Jerárquico

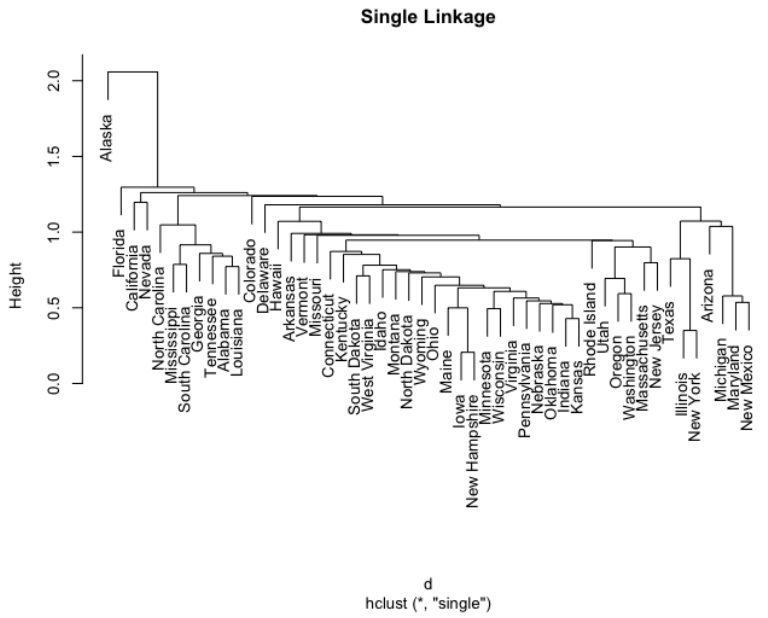
1. Comience con las *n* observaciones y una medida de *disimilitud* (como la distancia euclidiana) entre todos los pares $\binom{n}{2} = n(n-1)/2$de observaciones . Considerar a cada observación como su propio grupo (*cluster)*
2. Para *i = n, n*-1,. . . , 2:
   (a) Examine todas las *disimilitudes* entre pares de grupos (*clusters)* e identifique el par de grupos con mayor disimilitud (o sea el par grupos más similares entre sí). Fusione estos dos grupos. La disimilitud entre estos dos grupos indica la altura en el dendrograma a la que se debe colocar la fusión entre sus ramas.
   (b) Calcule las nuevas disimilitudes entre los i - 1 pares de grupos (*clusters*) restantes.

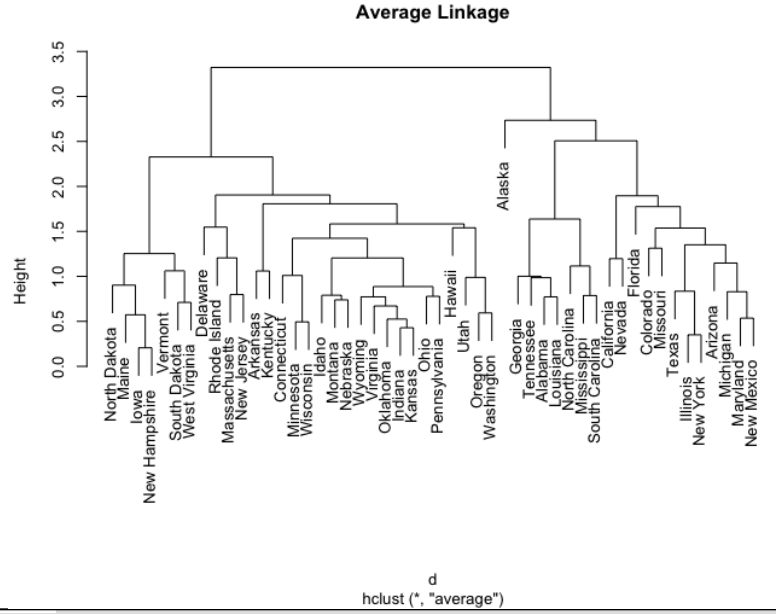
El concepto de *disimilitud* entre un par de observaciones debe extenderse a un par de *clusters*. Esta extensión se logra desarrollando la noción de *enlace*, que define la disimilitud entre dos grupos observaciones (*clusters*). Los cuatro tipos de enlace más comunes: completo, promedio, sencillo y centroide

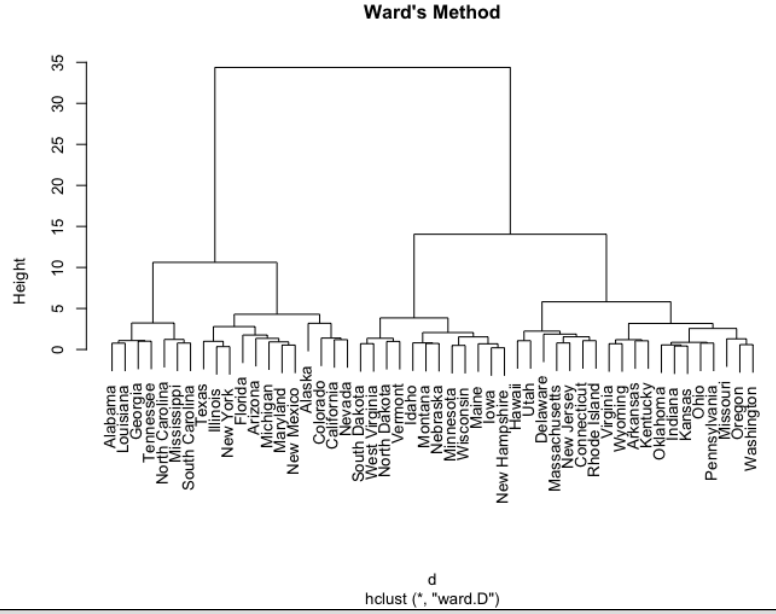
Tipos de vinculaciónes (linkage)

- Completa: Disimilitud máxima entre los grupos. Calcule todas las diferencias de pares entre las observaciones en el grupo A y las observaciones en el grupo B, y registre la mayor de estas diferencias.
- Simple: Disimilitud mínima entre los grupos. Calcule todas las diferencias de pares entre las observaciones en el grupo A y las observaciones en el grupo B, y registre la más pequeña de estas diferencias. El enlace único puede dar lugar a agrupaciones extendidas y finales en las que las observaciones individuales se fusionan de una en una.
- Promedio: Significa disimilitud intercluster. Calcule todas las diferencias de pares entre las observaciones en el grupo A y las observaciones en el grupo B, y registre el promedio de estas diferencias.
- Centroide La disimilitud entre el centroide para el grupo A (un vector medio de longitud p) y el centroide para el grupo B. El enlace del centroides puede dar como resultado inversiones indeseables.
- Varianza mínima de Ward: La disimilitud medida como la varianza total entre dos grupos. En cada paso, se combinan los pares de grupos con una distancia mínima entre grupos.



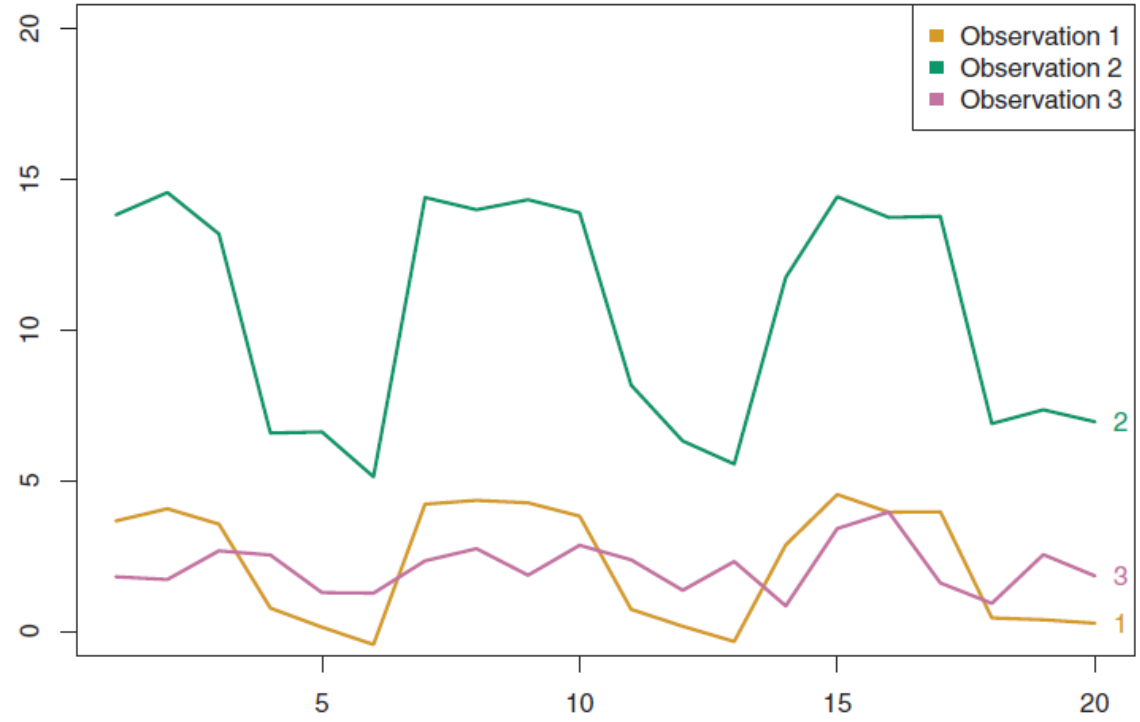






Elección de la Medida de Disimilitud

Hasta ahora hemos usado la distancia euclidiana como medida de disimilitud. Otra medida es la distancia basada en la **correlación** : dos observaciones son similares si sus características están altamente correlacionadas, aunque los valores observados pueden estar muy separados en términos de distancia euclidiana. Aquí la correlación tiene un significado diferente (el clásico mide la variación de una variable respecto a la variación de otra) pues lo que se es similitud, o *correlación,* entre los perfiles de las observaciones de a pares.
Veamos con un ejemplo la diferencia entre distancia euclidiana y distancia por correlación



1 y 3 Distancia Euclidiana pequeña

Cantidad Óptima de Clusters

Los tres métodos más populares para determinar cantidad óptima de clusters son:

1. Metodo del codo (*Elbow Method*)
2. Método de la silueta (*Silhouette Method*)

1. Metodo del codo (*Elbow Method*)

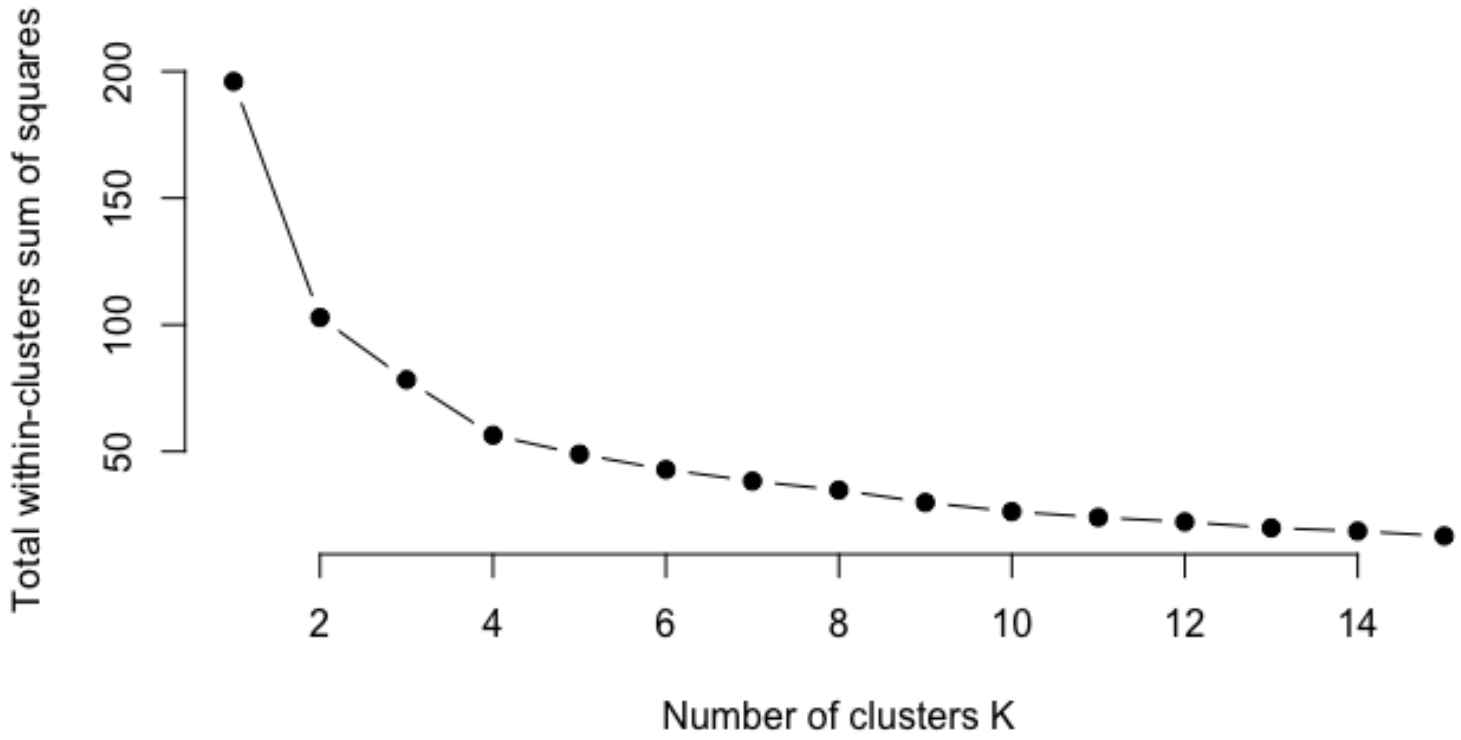
Recuerdemos que la idea básica detrás de los métodos de clústeringes definir clusters de tal manera que se minimice la variación total dentro del cluster (conocida como variación total dentro del cluster o suma total de cuadrados dentro del cluster**)**

WSS = $$\sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

WSS mide la “compactación” de la agrupación y queremos que sea lo más pequeña posible.

Por lo tanto, podemos usar el siguiente algoritmo para definir los clústeres óptimos:

1. Generar modelos de clustering para diferentes valores de k.
2. Para cada *k*, calcule la suma total de cuadrados dentro del grupo (WSS)
3. Grafique la curva de WSS en función del número de clusters *k*.
   Seleccionaremos la cantidad de clusters en el codo (*elbow*)



Método de silueta promedio *(Average Silhouette Method)*

El enfoque de silueta promedio mide la calidad de un agrupamiento. Es decir, determina qué tan bien se encuentra cada observación dentro de su grupo. Un valor cercano a 1 indica una buena agrupación. El método de la silueta promedio calcula la silueta promedio de las observaciones para diferentes valores de k. El número óptimo de grupos k es el que maximiza la silueta promedio en un rango de valores posibles para k.