Aprendizaje no supervisado

Clustering jerárquico y no jerárquico

Máster en Data Science

Mario Encinar, PhD – **MAPFRE**
encinar@ucm.es

Contenidos

1 | Introducción

Aprendizaje No Supervisado

- Se trata de herramientas cuyo objetivo es entender los datos, **sin una variable objetivo** (o supervisor), y **organizarlos en grupos de manera natural**.

- Mientras que, por otro lado, el aprendizaje supervisado consiste en predecir o estimar una variable objetivo (o respuesta) desde varios inputs (predictores).

- El Clustering es una las técnicas más utilizadas en Aprendizaje No Supervisado, otra es el PCA (Principal Component Analysis).
  - k-means clustering aquí.

- En general, es difícil determinar el número de clusters óptimo, o grupos naturales en los que se organizan los datos.

Cluster Analysis

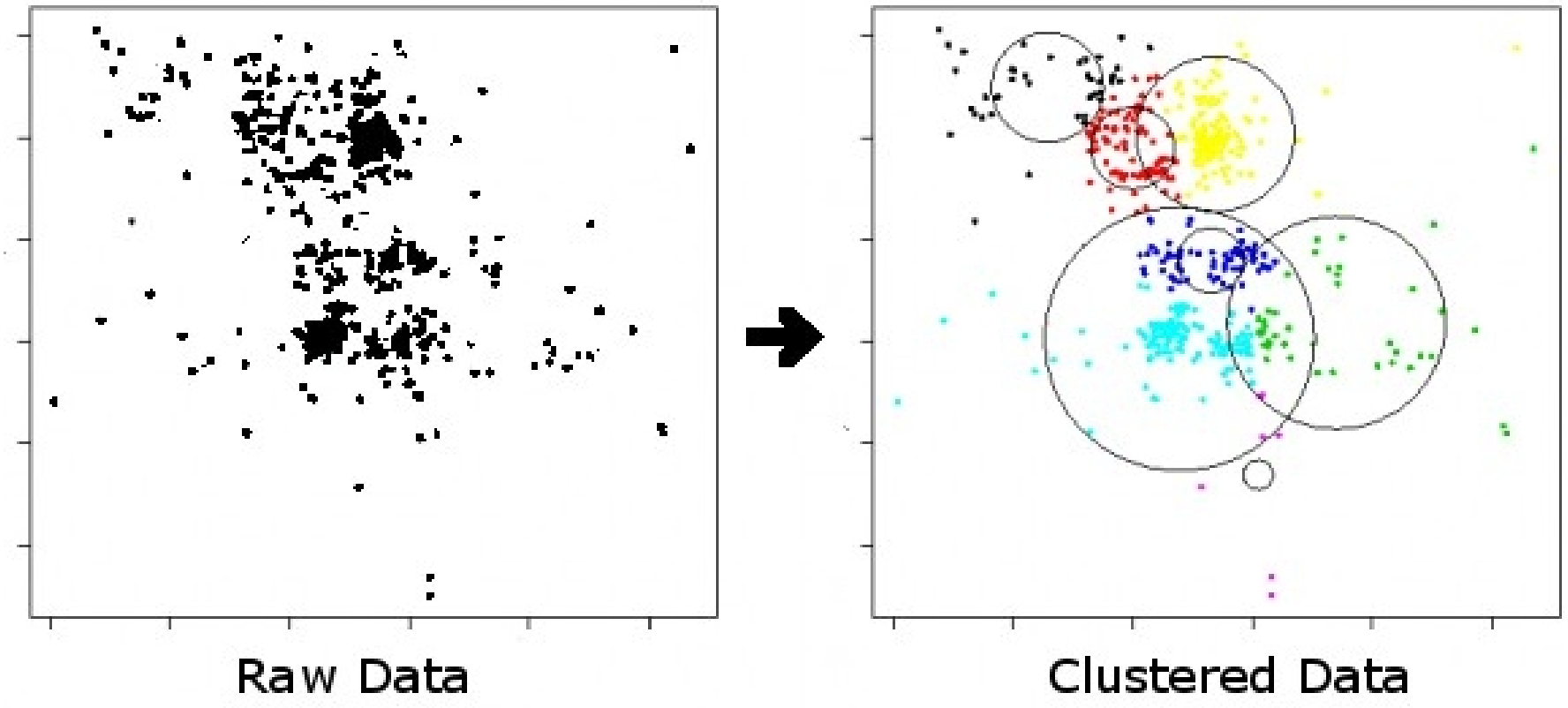
Clasificar las **observaciones** de una matriz de datos X (p.e. un DataFrame de R) en grupos homogéneos:

- Las observaciones dentro de cada grupo deben ser similares.
- Las observaciones de diferentes grupos deben ser diferentes.
- Se mide la **similaridad** o **proximidad** a través de distancias.
- No conocemos a priori el grupo al que pertenecen las observaciones.
- Tampoco conocemos a priori el número de grupos.

Aplicaciones del Clustering

- Marketing: Organizar los clientes en diferentes perfiles (según su consumo, edad, etc.) para ofrecer campañas de publicidad dirigidas.
- Finanzas: Clasificar compañías según su rendimiento en los mercados financieros.
- Salud: Organizar pacientes en grupos de tratamiento.
- Seguros: Identificar grupos de asegurados con altos costes de siniestros.
- Redes Sociales: Identificar comunidades.
- Genética: Selección genética.
- Otras: Sistemas de recomendación, detección de anomalías, análisis de imagen, grupos de resultados de búsqueda, etc.

Idea principal del Clustering



Ideas Básicas

- El Clustering estudia datos para los cuales el número de grupos es desconocido e indefinido.

- Necesitamos poner el foco en distancias **intra-cluster**, para incrementar la similaridad.

  - Cómo de cerca están los datos unos de otros.
  - Se denomina coloquialmente distancia o medida de similaridad.

- Y también en la distancia **inter-cluster**, para disminuir la similaridad.

  - Cómo de cerca están los clusters entre sí
  - Se denomina comúnmente la función linkage.

- Por tanto, la única información que usa el clustering son **similaridades**.

Ejemplo

ID	Gender	Age	Salary	Balance
1	F	27	21000	550
2	M	51	64000	900
3	M	53	75000	825
4	F	32	55000	1100
5	M	45	50000	875
6	F	37	45000	650

¿Qué clientes son más similares?

Ejemplo

	Term1	Term2	Term3	Term4	Term5	Term6
Doc1	0	2	0	0	1	0
Doc2	3	5	4	3	0	0
Doc3	3	0	0	0	3	4
Doc4	0	0	0	3	0	4
Doc5	2	1	2	3	0	1
Doc6	1	4	2	1	2	0

¿Qué documentos son más similares?

Donde Term1, Term2, … son variables que expresan la frecuencia de aparición de diferentes términos en los diferentes documentos (Doc1, Doc2,…).

Cluster Analysis: Distancias

- La elección de una medida de la distancia (o **métrica**) influye en la forma de los clusters.
- Similaridad se relaciona inversamente con distancia.
- Un buen clustering consigue una alta similaridad intra-cluster y una baja similaridad inter-cluster
- Por tanto, la elección de la medida de similaridad es crucial
- Propiedades de distancia:
  - $\text{dist}(\text{obs}_i, \text{obs}_j) \geq 0$ and $\text{dist}(\text{obs}_i, \text{obs}_i) = 0$
  - $\text{dist}(\text{obs}_i, \text{obs}_j) = \text{dist}(\text{obs}_j, \text{obs}_i)$
  - $\text{dist}(\text{obs}_i, \text{obs}_j) \leq \text{dist}(\text{obs}_i, \text{obs}_k) + \text{dist}(\text{obs}_k, \text{obs}_j)$

Cluster Analysis: Distancias

Existen muchas medidas de similaridad, con diferente grado de subjetividad (según la aplicación):

- Distancia Euclídea: $d(x,y) = \|x-y\|_2 = \sqrt{(x-y)'(x-y)}$
- Mahalanobis (o distancia estadística): $d(x,y) = \sqrt{(x-y)'S^{-1}(x-y)}$
- Distancia Minkowski (L^p): $d(x,y) = \|x-y\|_p$
  If $p = 1$, Manhattan o *city block distance*.
  If $p \to \infty$, distancia Chebychev.
- Muchas otras, como las distancias *kernel*: $d(x,y) = \|\phi(x) - \phi(y)\|$ (para datos altamente no lineales).
- Distancia Gower, que puede manejar datos mixtos (ordinales, nominales, etc.)

Cluster Analysis: Distancias

- Incluso cuando las observaciones no son numéricas, necesitamos métricas de distancias

- Pero las distancias pueden definirse incluso para datos no numéricos:

  Ejemplo: si tenemos nombres de personas, la distancia puede ser 0 cuando las dos personas tienen el mismo nombre (o similar) y uno en caso contrario.

- En general, aplicaciones complejas implica definir de manera creativa métricas de distancia entre observaciones

Cluster Analysis: Variables/Features

- Las variables o features que incluímos en el análisis pueden tener gran impacto en la solución.

- Es necesario recurrir al conocimiento de la posible aplicación del analista, su creatividad y expertise. Un buen análisis exploratorio suele ser de ayuda.

- En la práctica, algunas variables se utilizan para el clustering y otras para el perfilado de los grupos de observaciones.

  Por ejemplo: en un dataset de banca, las variables asociadas con los atributos socio-económicos (como income/balance/risk) se pueden usar para la segmentación; y el resto de variables, como las características de los clientes (profesión, región en la que viven, etc) pueden ayudar a describir los grupos que devuelve el clustering.

- Podemos calcular estadísticos resumen (medias, desviaciones, etc) de los atributos de perfilado y ver cómo se comportan en diferentes segmentos.

Cluster Analysis: Métodos

Métodos de Partición: Dividen las observaciones en un número de grupos pre-especificado.

Objetivo: Agrupar observaciones similares basándose en alguna distancia (similaridad) entre observaciones:

- K -means: Los objetos dentro de cada cluster se encuentran lo más cerca posible entre sí, y lo más lejos posible de los objetos de otros clústeres. Cada clúster se caracteriza por un *centroide*.
- Métodos basados en modelos: Mezclas de distribuciones estadísticas.

Métodos Jerárquicos: Comienzan con clústeres con una sola observación y unen los clústeres en pasos iterativos posteriores.

Dendrograma: Un árbol de jerarquía multi-nivel, donde los clusteres a un nivel se engloban en un cluster de un nivel más alto. Esto permite decidir qué nivel de agrupamiento es más apropiado.

2 | Métodos de Partición

Cluster Analysis: Métodos de Partición

- Es necesario fijar previamente el número de grupos, K.
- Al final del algoritmo de clustering cada observación va a uno de estos grupos.
- La herramienta más popular de partición: K –means.
- La métrica de distancia es la Euclídea (la distancia natural en nuestro espacio tridimensional).
- Proporciona una solución razonable.
- Es muy rápido.

Cluster Analysis: K -means

- Asigna aleatoriamente cada observación a uno de los K grupos.
- Calcula las K medias de cada muestra (centroide) de las observaciones de cada grupo.
- Asigna cada observación al grupo con la media más cercana (usando la norma euclídea)
- Repite los dos pasos anteriores hasta que no cambian los grupos.

Ilustración animada: Link

Ilustración animada: Link

Cluster Analysis: K -means

- El algoritmo proporciona una solución local: trata de minimizar la suma de distancias de cada objeto a su centroide de grupo (variabilidad intra-grupos).
- Por lo tanto, los clústeres finales dependen de la asignación aleatoria inicial.
- Es recomendable ejecutar el algoritmo varias veces con diferentes asignaciones iniciales.
- La mejor solución será aquella con la menor variabilidad intra-grupo.
- Cuando las variables están en unidades diferentes, hay que estandarizar las variables.
- K –means solo trata con variables numéricas.

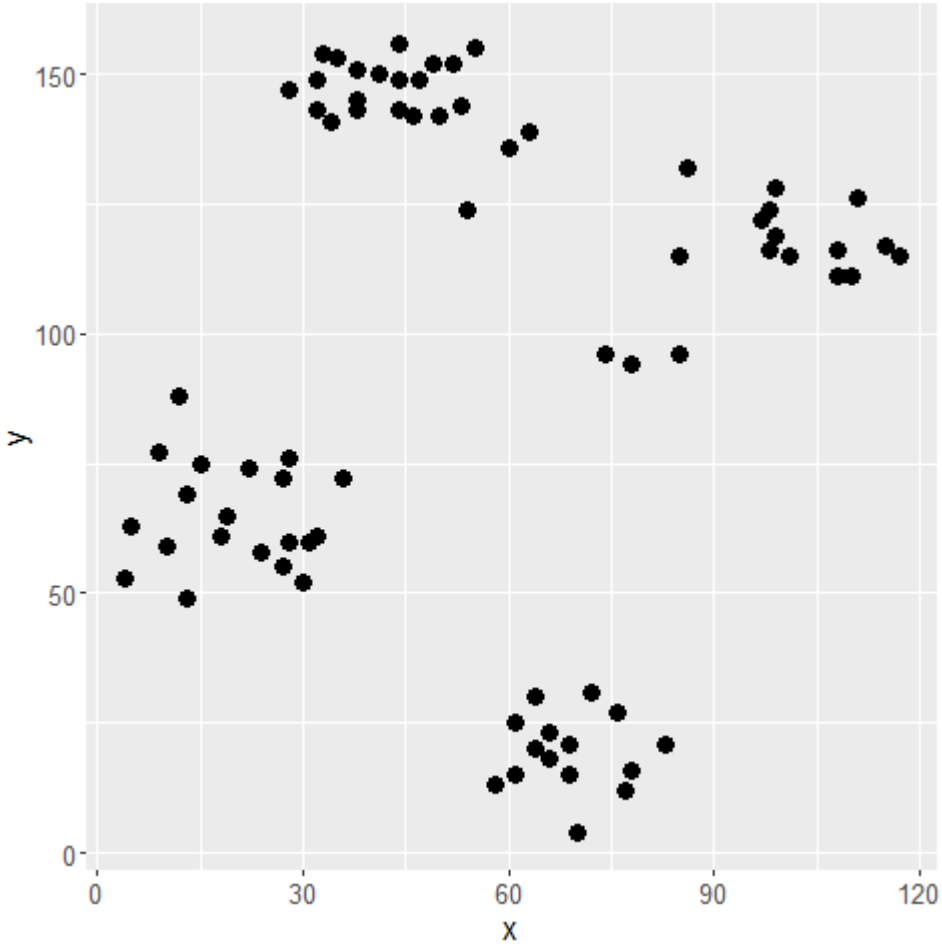
Cluster Analysis: K -means

Consideraciones prácticas:

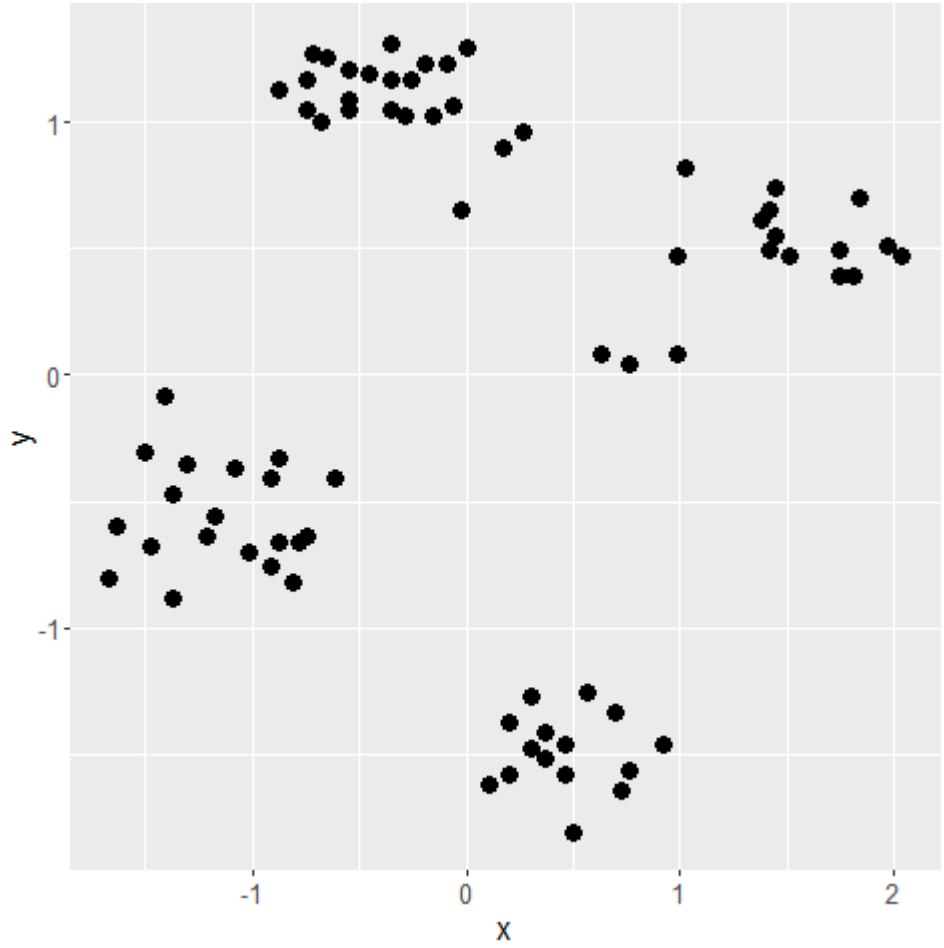
- Para encontrar un número apropiado de grupos, se aplica el algoritmo con K = 2, 3,. . . y se observa la variabilidad intra grupo y también la interpretabilidad: como un gráfico de codo en PCA.
- Aplicar el algoritmo con datos originales y también con datos estandarizados (o cambiar la distancia, por defecto Euclídea).
- Cuidado con los valores atípicos: pueden alejar los centroides de su posición verdadera.
- No sabemos qué variable contribuye más a la agrupación (asumimos que cada variable tiene el mismo peso).
- La forma de los clústeres suele ser circular (distancia euclídea) o elipsoides (distancia Mahalanobis).

Cluster Analysis: Ejemplo

Dataset sencillo: 75 puntos en 2D

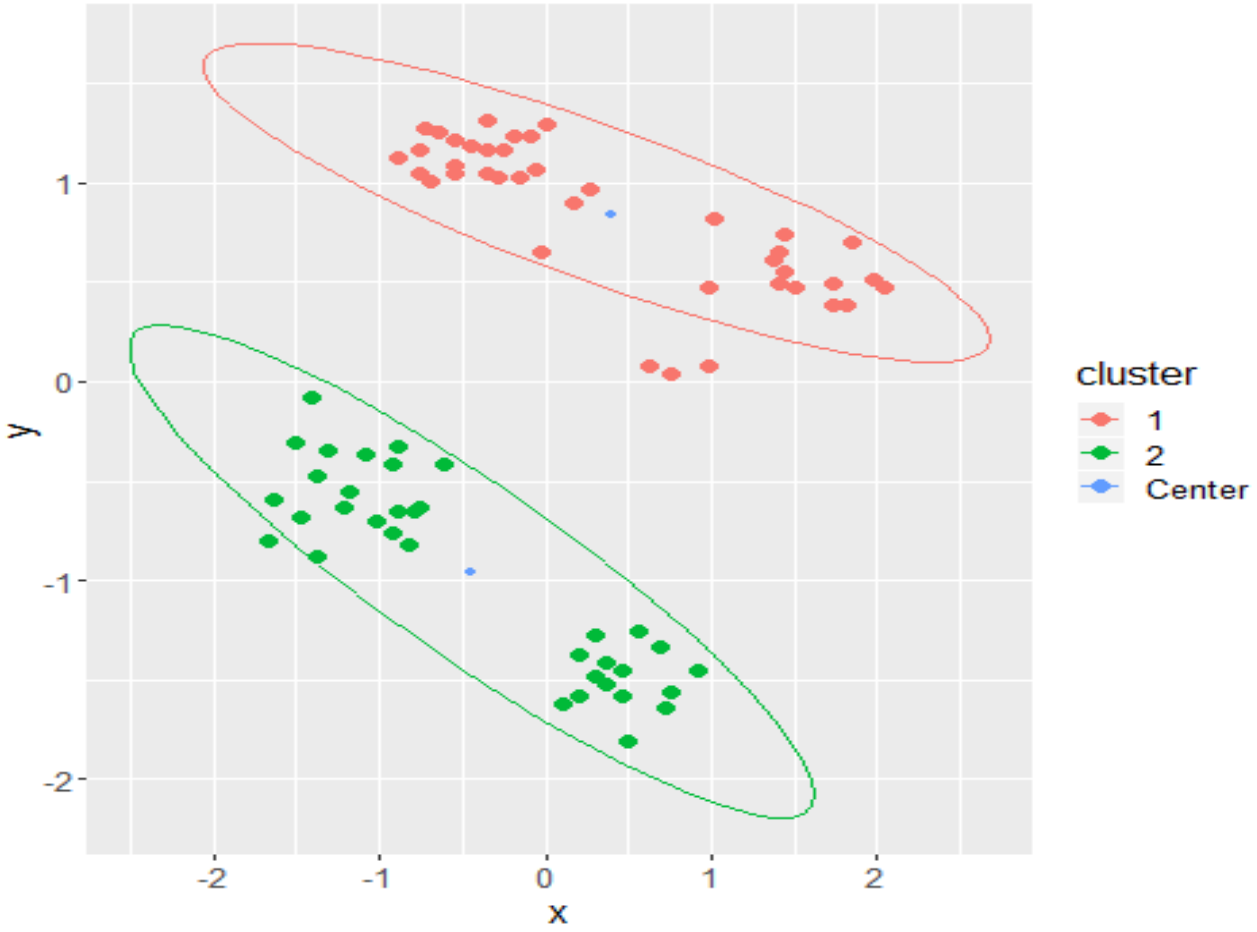


Estandarizado



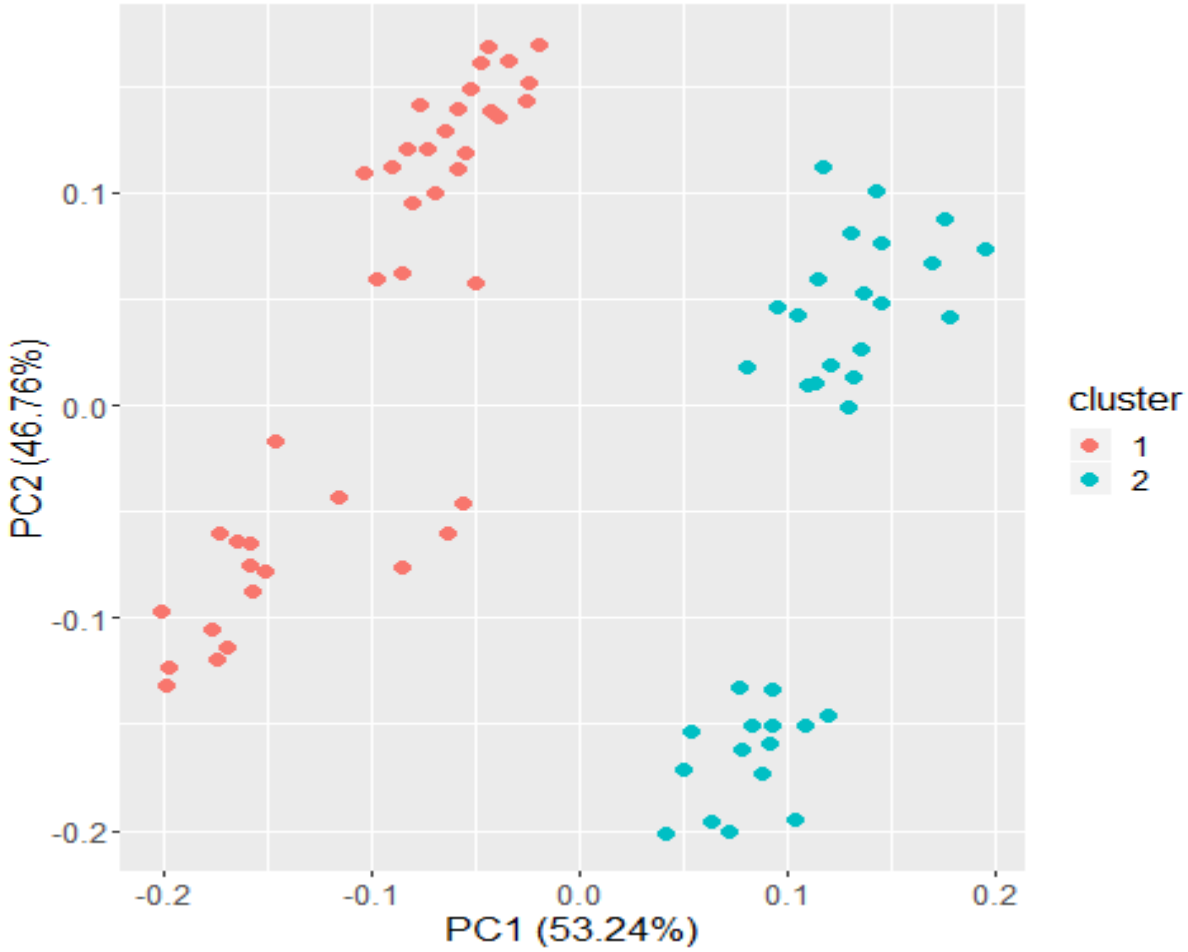
Cluster Analysis: Ejemplo

Consideremos que hay dos grupos



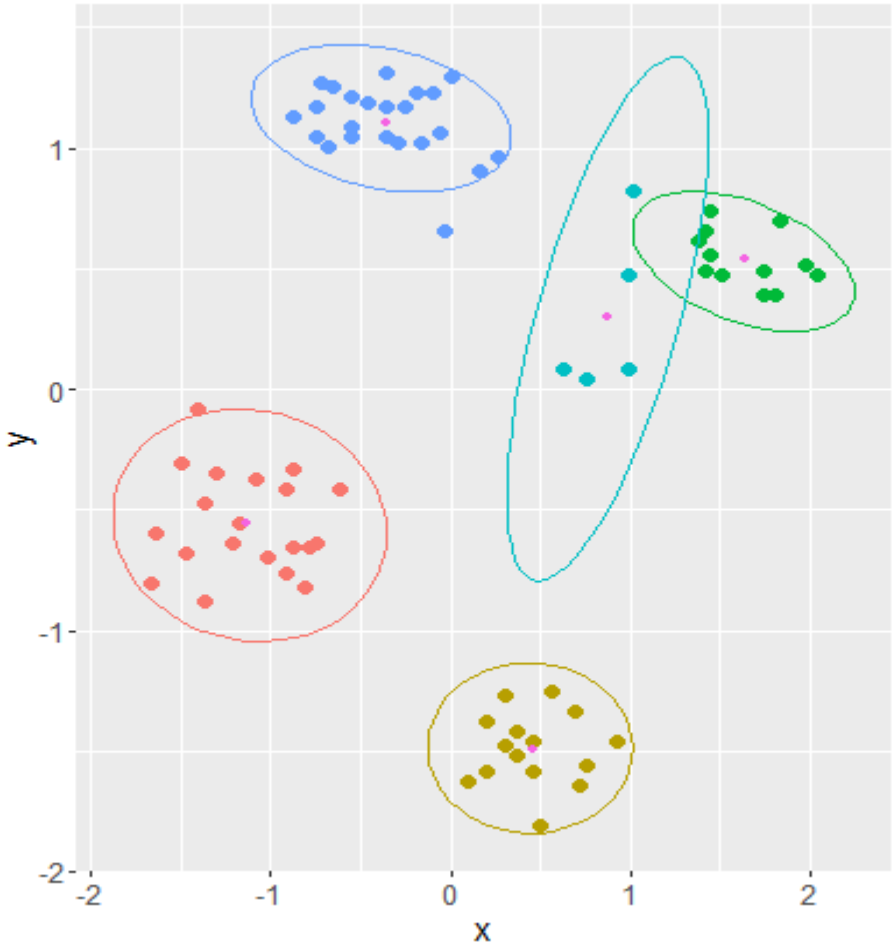
Cluster Analysis: Ejemplo

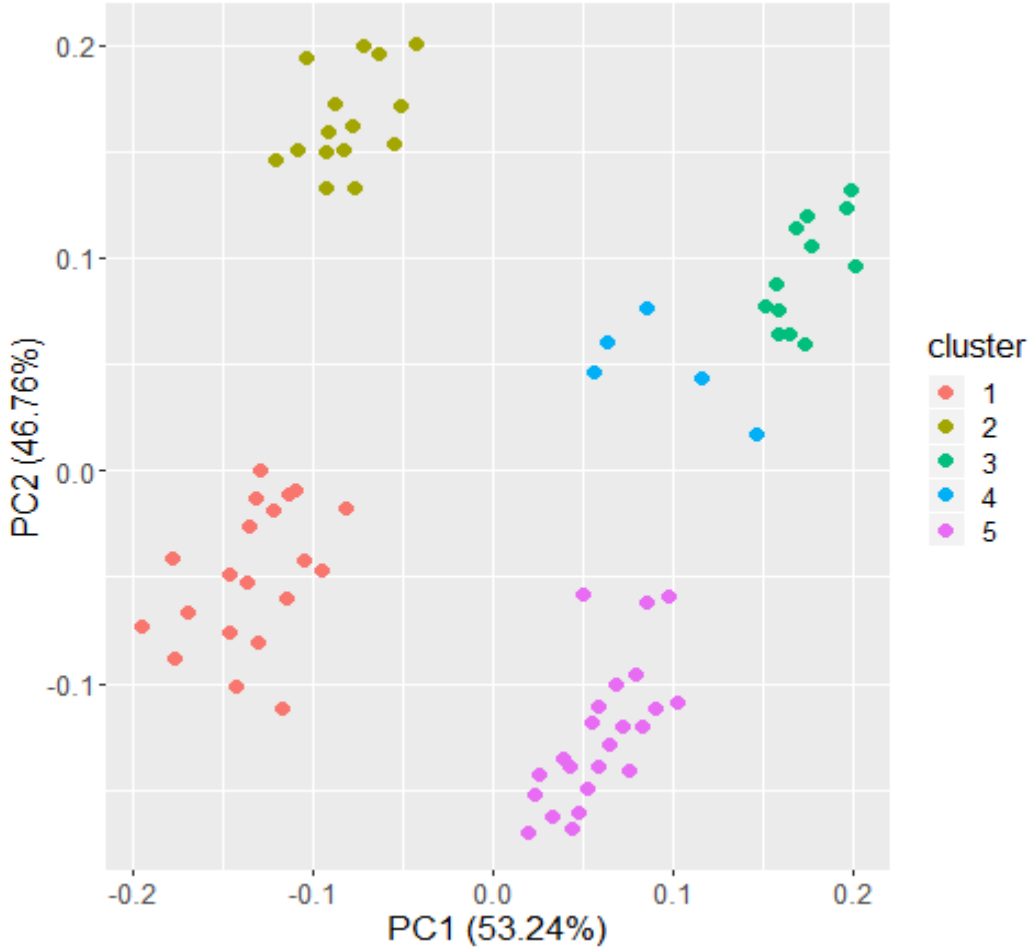
En alta dimensionalidad, es útil dibujar los clusters en las primeras componentes principales



Cluster Analysis: Ejemplo

Consideremos ahora que existen 5 grupos



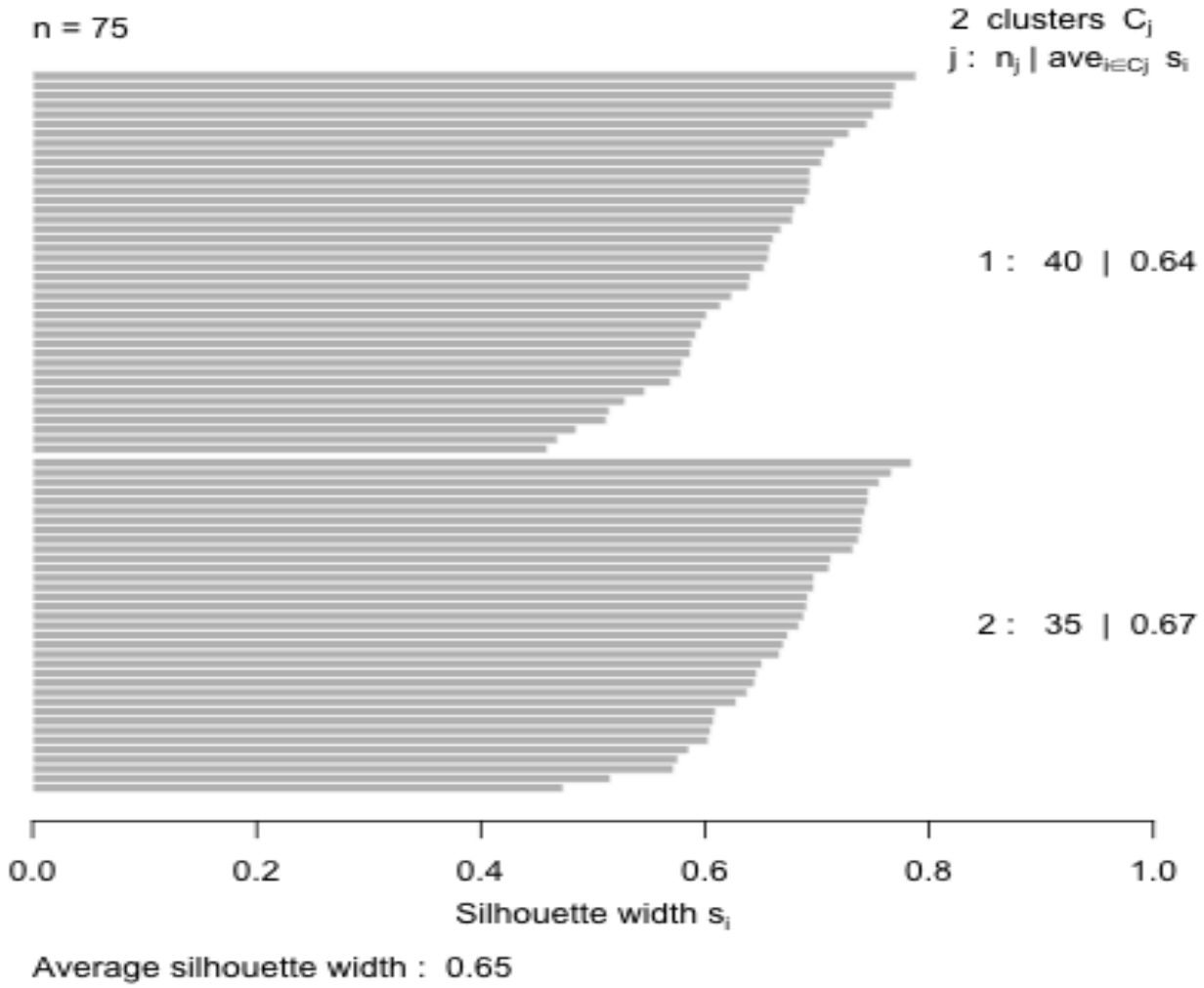


Cluster Analysis: Métricas de Evaluación

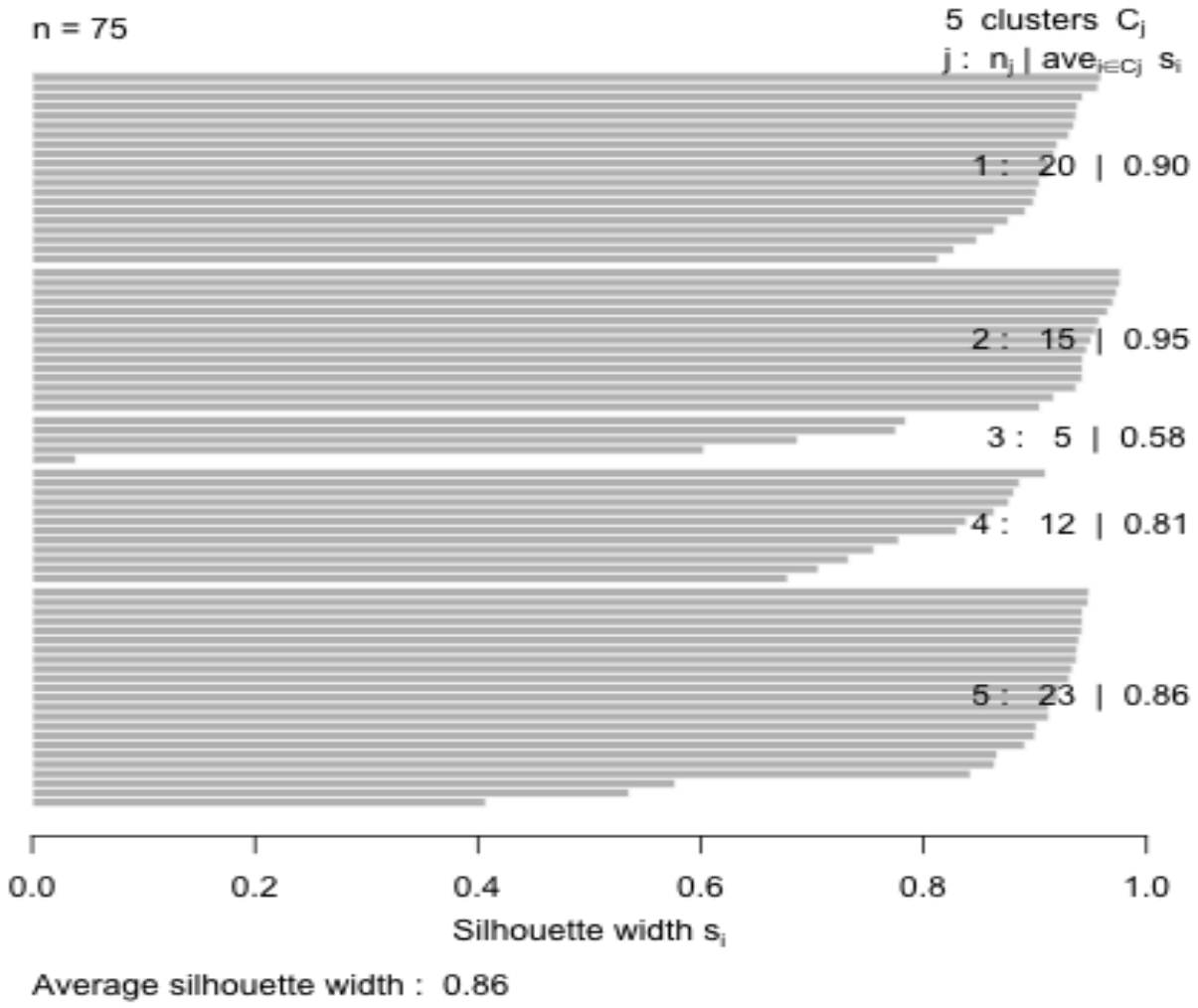
Silhouette plot

- Proporciona una representación gráfica de qué tan bien se encuentra cada objeto dentro de su grupo.
- Los puntos con un gran coeficiente de silueta (cerca de 1) están bien agrupados, aquellos con un coeficiente bajo (cerca de 0) tienden a estar entre los grupos.
- Los valores de silueta negativos indican que la observación probablemente se ubica en el grupo incorrecto.
- El gráfico también es útil para adivinar la cantidad de clústeres.

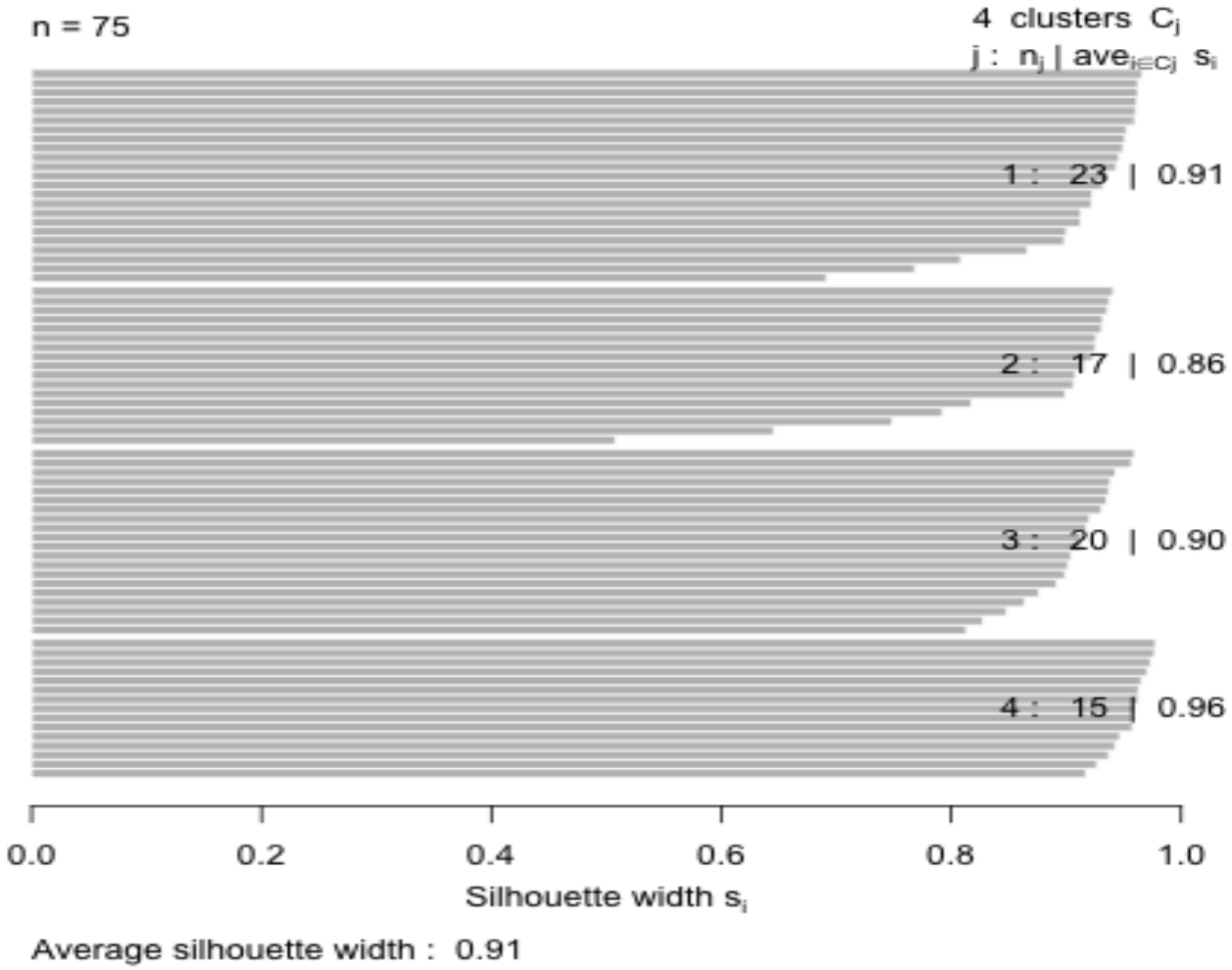
Silhouette plot: Ejemplo con k = 2



Silhouette plot: Ejemplo con k = 5



Silhouette plot: Ejemplo con k = 4

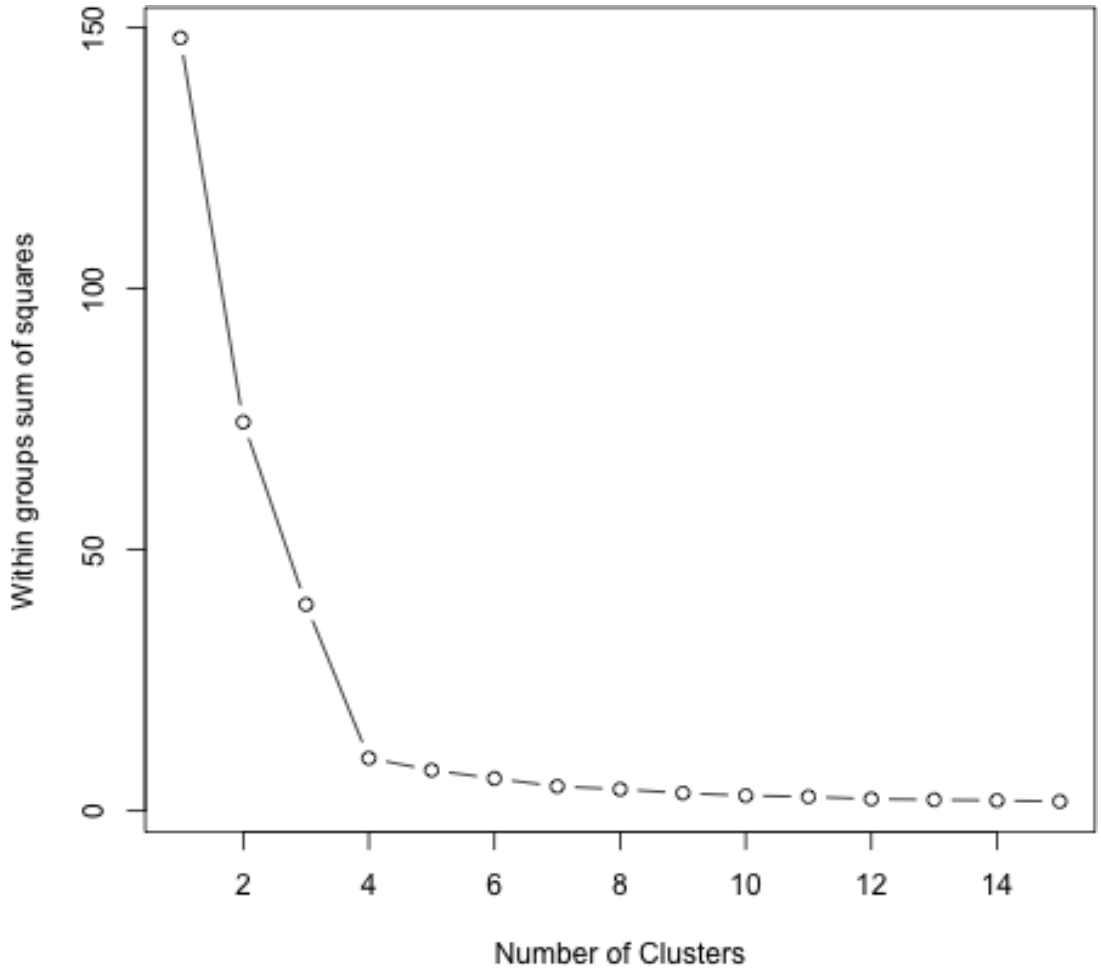


Cómo determinar el número de clústeres

- Dado que esto es aprendizaje no supervisado (sin supervisor), no tenemos una manera formal de determinar el número de clústeres.
- Es necesario equilibrar la cantidad de clústeres y la varianza promedio dentro de los clústeres.
- Cuanto mayor sea el número de grupos, menor será la varianza: no hay manera de minimizar ambas medidas.
- La experiencia / conocimiento es una ventaja: use su mejor juicio.
- Herramientas que ayudan en la decisión: diagrama de silueta (Silhouette plot), gráfico de codo (Scree plot), etc.

Cómo determinar el número de clústeres

Scree plot



Otros Métodos de Partición: K-Medoids

- Similar a K-means, pero ahora los centros son observaciones *de facto* (medoids) en lugar de promedios de los datos.
- Fácil de entender y más robusto a los valores atípicos.
- Puede funcionar mejor que K-means para pequeños conjuntos de datos.
- Pero no escala bien en alta dimensión.
- En PAM, se considera la distancia euclidea o de Manhattan a menos que definamos una matriz de disimilitud (en lugar de una matriz de datos).

Otros Métodos de Partición: Basados en modelos

- Partición de clústeres basado en la hipótesis de que los datos son una mezcla de K distribuciones estadísticas.
- Es una generalización estadística de k-means.
- **Gaussian mixture models**: supongamos que las distribuciones estadísticas K son normales multivariantes.
- La estimación de los parámetros se realiza mediante el algoritmo de maximización de expectativas (EM).
- Una vez que se realiza la estimación, cada observación se asigna a la componente (clúster) con mayor probabilidad (obtenida por el Teorema de Bayes).

3 | Métodos Jerárquicos

Cluster Analysis: Hierarchical Clustering

- Con este método no necesitamos preseleccionar la cantidad de grupos.
- Representación visual: **Dendrograma.**
- Dos tipos: aglomerativo y divisivo

  Nos enfocamos en aglomerativo: fusionar de abajo hacia arriba (*bottom-up*).

Cluster Analysis:
Agglomerative hierarchical clustering

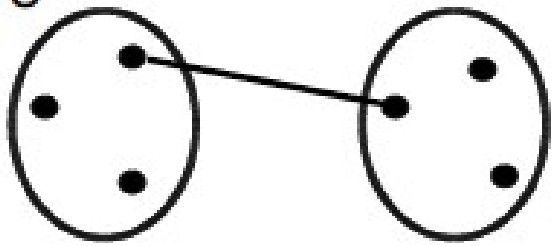
En los métodos aglomerativos, hay muchas maneras de unir los grupos, o lo que es lo mismo medir distancias entre dos grupos

Esto se llama **linkage** o dissimilarity:

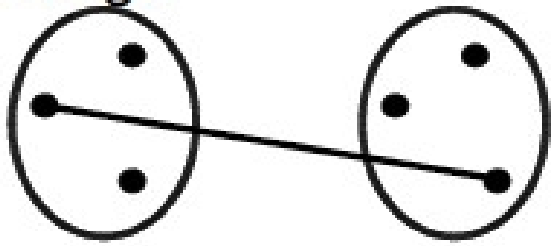
- Single Linkage (vecino más cercano).
- Complete Linkage (vecino más lejano).
- Average Linkage (un compromiso entre los métodos anteriores).
- El criterio de Ward (otro compromiso).

Cluster Analysis: Linkage

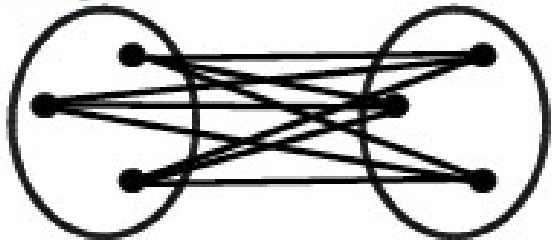
Single Linkage



Complete Linkage



Average Linkage



Cluster Analysis: Agglomerative methods

- 1. Cada observación forma un grupo individual
- 2. Se calcula la matriz de distancias D como $\mathrm{d}(x_i, x_j)$ para $i, j = 1, \ldots, n$
  (la distancia predeterminada es la Euclídea)
- 3. Se determina la distancia más corta en D, digamos d_{IJ} en $D = D_1$

  Se fusionan los clusters I y J para formar un nuevo cluster IJ

Cluster Analysis: Agglomerative methods

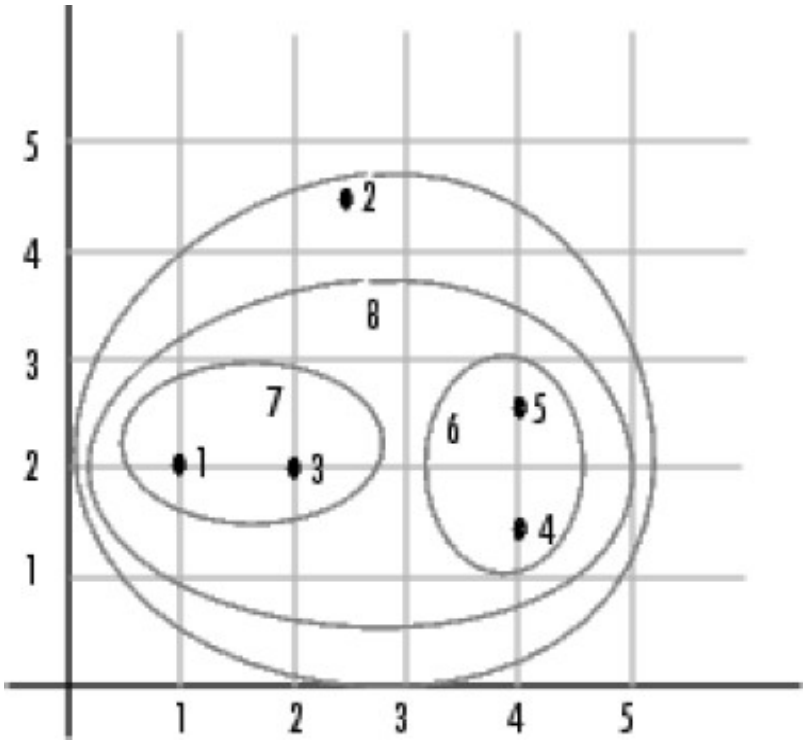
- 4. Se calculan las distancias, $d_{IJ,K}$ entre el nuevo cluster IJ y el resto de clusters $K \neq IJ$. Estas distancias se basan en:
  - Single Linkage: $d_{IJ,K} = \min\{d_{IK}, d_{JK}\}$
  - Complete Linkage: $d_{IJ,K} = \max\{d_{IK}, d_{JK}\}$
  - Single Linkage: $d_{IJ,K} = \frac{1}{N_{IJ}N_K} \sum_{i \in IJ} \sum_{k \in K} d_{ik}$

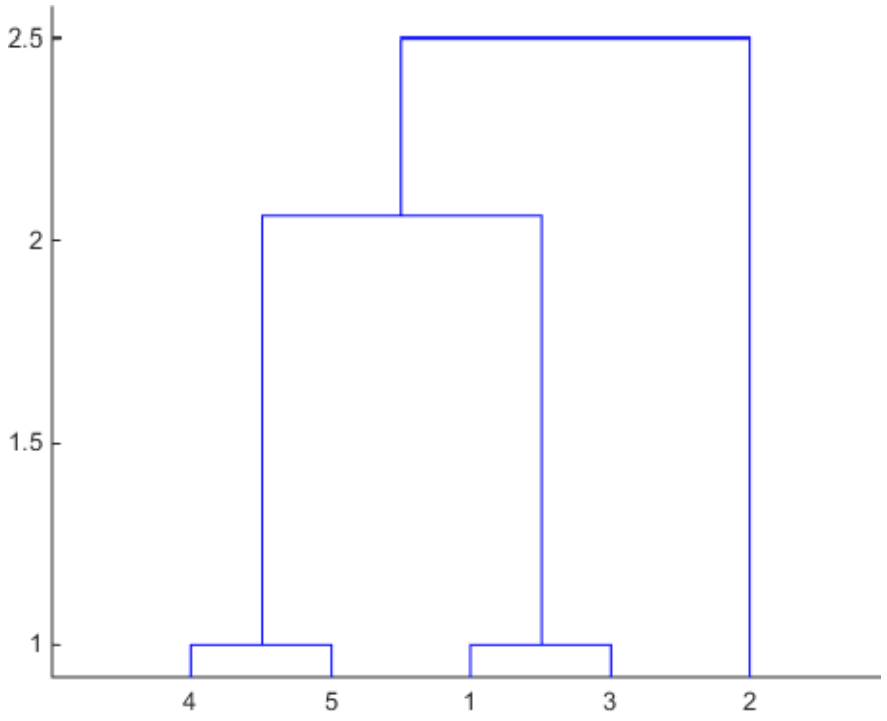
Cluster Analysis: Agglomerative methods

- 5. Se forma una nueva matriz D_2 eliminando la fila y la columna I, y también la fila y la columna J; y luego se agrega una nueva fila y columna IJ con las similitudes calculadas en el paso anterior.
- 6. Repítase n - 1 veces los pasos 3, 4 y 5. En el último paso, $D_n = \emptyset$ y todas las observaciones forman un solo grupo.
- Al final, tenemos una lista con todos los clústeres unidos en cada paso con las disimilaridades asociadas.

Animated illustration: Link

Cluster Analysis: Linkage and Dendrogram

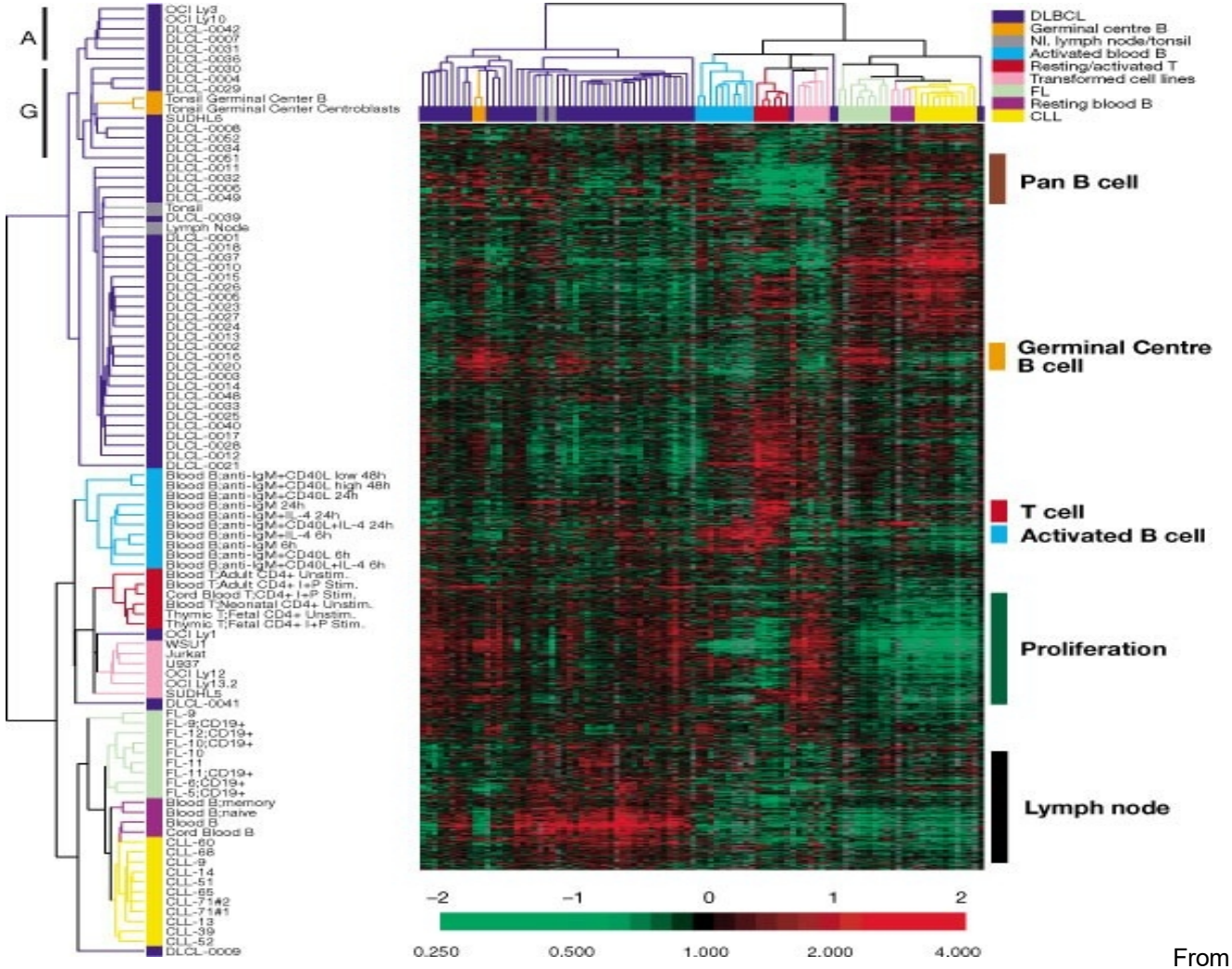




Cluster Analysis: Dendrograma

- La información final generalmente se presenta como un **dendrograma**:
  - En la parte inferior del árbol tenemos las n observaciones.
  - De abajo hacia arriba, se representa la fusión en cada paso.
  - En el eje vertical encontramos las distancias que representan cuán cerca / lejos están los diferentes grupos.
- Si cortamos el dendrograma en un nivel determinado, entonces tenemos el clustering con un número de grupos correspondiente.
- Conveniente estandarizar los datos, aunque la distancia Mahalanobis se puede utilizar en su lugar

Genetic Application: clustering both samples and genes



From Nature (2000); 403(6769):503-11

Comparativa final: Partición vs Jerárquico

- k -means se adapta bien a la dimensión (complejidad lineal, O(n)), pero la solución es local.
- Las agrupaciones jerárquicas escalan mal, $O(n^2)$, pero los resultados son más intuitivos.
- El particionado es óptimo bajo ciertas condiciones y es fácil definir conglomerados. Pero requiere seleccionar una cantidad de clústeres y los resultados no son reproducibles
- Jerárquico es bastante visual y reproducible, pero más difícil de asignar clusters

Clustering: Robustness Analysis

- Los clústeres deben ser relativamente robustos respecto a: los modelos utilizados, el subconjunto de observaciones, la distancia, etc.
- Esto significa que la mayoría de las observaciones deberían pertenecer a los mismos grupos bajo algunos cambios. De lo contrario, la agrupación puede no ser válida o confiable.
- La interpretación de los clústeres también debería ser robusta.

Clustering: Robustness Analysis

Necesitamos verificar los cambios usando:

- Diferentes subconjuntos de los datos originales.
- Diferentes variables / características.
- Diferentes medidas de distancia.
- Diferentes métodos de agrupamiento.
- Número de grupos.
- Estandarización y / o transformación de variables.

Esto significa que el clustering es un proceso iterativo con muchas variaciones (tal vez millones) hasta que logremos una solución confiable / conveniente / interpretable.

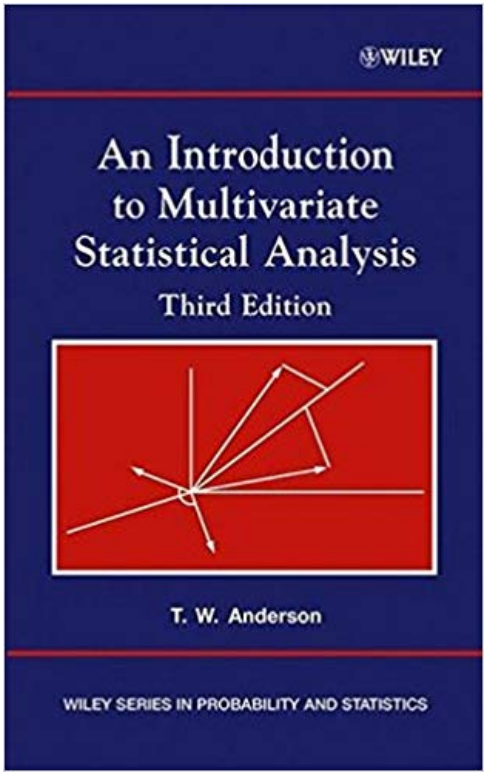
Cluster Analysis: Observaciones finales

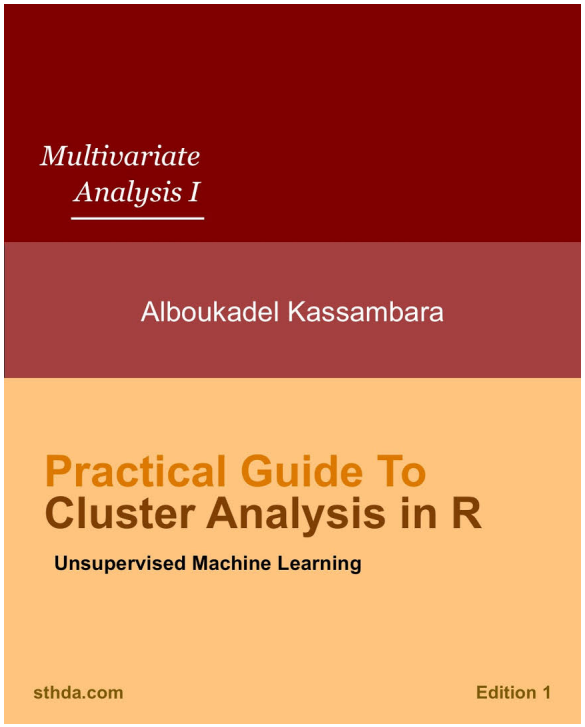
Algunas decisiones a considerar:

- Si el objetivo es que los grupos tengan significado, entonces el clustering debería capturar la estructura natural de los datos.
- Otras veces, el análisis de clustering se usa como punto de partida para otros objetivos.
- El análisis de conglomerados proporciona una abstracción desde los individuos a los grupos a los que pertenecen.
- En cierto sentido, clustering es una forma de clasificación en la que las etiquetas se derivan de los datos.
- Tenga cuidado con la maldición de la dimensionalidad: intente eliminar variables que sean muy ruidosas o que no sean interesantes, intente diferentes subconjuntos de variables (clique), reduzca la dimensión (PCA / SVD)
- Cuidado con los valores atípicos.
- La validación de clusters es difícil y algunas veces frustrante

4 | Materiales

An Introduction to Multivariate Statistical Analysis
Practical Guide to Cluster Analysis in R





[Otras medidas de validación (*must read*)]

[Librería y visualizaciones más aparentes (factoextra)]