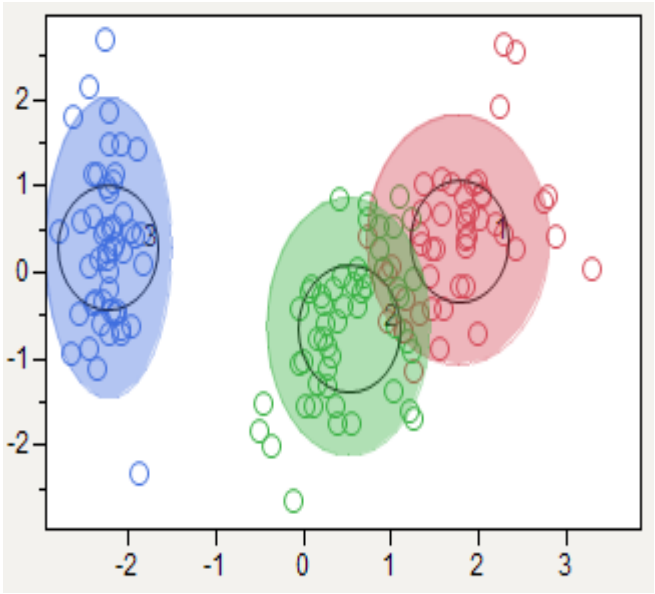
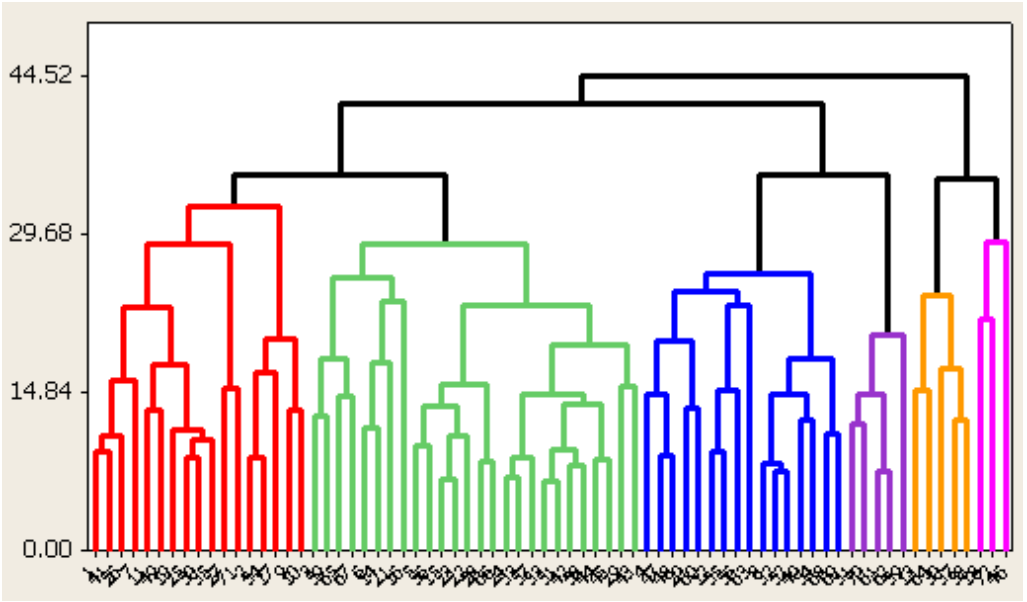
Análisis de agrupamientos





Oscar Centeno Mora

Introducción

- El Análisis Cluster, conocido como Análisis de Conglomerados, es una técnica estadística multivariante que busca agrupar elementos (o variables) tratando de lograr la máxima homogeneidad en cada grupo y la mayor diferencia entre los grupos.

- El Análisis Cluster tiene una importante tradición de aplicación en muchas áreas de investigación. Sin embargo, junto con los beneficios del Análisis Cluster existen algunos inconvenientes. El Análisis Cluster es una técnica descriptiva, ateórica y no inferencial. El Análisis Cluster no tiene bases estadísticas sobre las que deducir inferencias estadísticas para una población a partir de una muestra, es un método basado en criterios geométricos y se utiliza fundamentalmente como una técnica exploratoria, descriptiva pero no explicativa.

Introducción

- Las soluciones no son únicas, en la medida en que la pertenencia al conglomerado para cualquier número de soluciones depende de muchos elementos del procedimiento elegido. Por otra parte, la solución cluster depende totalmente de las variables utilizadas, la adición o destrucción de variables relevantes puede tener un impacto substancial sobre la solución resultante.

- Los algoritmos de formación de conglomerados se agrupan en dos categorías:

- Algoritmos de partición: Método de dividir el conjunto de observaciones en k conglomerados (clusters), en donde k lo define inicialmente el usuario.

Introducción

- Algoritmos jerárquicos: Método que entrega una jerarquía de divisiones del conjunto de elementos en conglomerados.

- Un método jerárquico aglomerativo parte con una situación en que cada observación forma un conglomerado y en sucesivos pasos se van uniendo, hasta que finalmente todas las situaciones están en un único conglomerado.

- Un método jerárquico disociativo sigue el sentido inverso, parte de un gran conglomerado y en pasos sucesivos se va dividiendo hasta que cada observación queda en un conglomerado distinto.

Introducción

- El análisis de conglomerados nos va a permitir contestar a preguntas tales como:

- ¿Es posible identificar cuáles son las empresas en las que sería más deseable invertir?

- ¿Es posible identificar grupos de clientes a los que les pueda interesar un nuevo producto que una empresa va a lanzar al mercado?

- ¿Se pueden clasificar las bodegas de en función de las características químicas y ópticas del vino que producen?

Objetivos

- El análisis de cluster busca particionar un conjunto de objetos en dos o más grupos basados en la similitud de esos objetos según una serie especificada de características.

- Como técnica se utiliza para desarrollar una clasificación objetiva de los objetos.

- El conocimiento de los grupos puede permitir una descripción sintética de un conjunto de datos multidimensional complejo. De ahí su uso en minería de datos.

- Esta descripción sintética se consigue sustituyendo la descripción de todos los elementos de un grupo por la de un representante característico del mismo

Etapas del análisis de conglomerados

- El análisis por conglomerados se suele llevar a cabo de la siguiente forma:

1. Elección de las variables
2. Elección de la medida de asociación
3. Elección de la técnica Cluster
4. Validación de los resultados

Elección de las variables

Elección de las variables

- La selección de variables debe tomar en cuenta consideraciones teórico-conceptuales y prácticas.

- Debe haber un racional que respalde el uso de las variables ya sea basado en teoría explícita, investigación pasada, o supuestos

- Deben incluirse sólo las variables que caracterizan los objetos que van a ser agrupados y están relacionadas con los objetivos del análisis.

- La inclusión de variables irrelevantes incrementa la posibilidad de tener valores extremos
- Se recomienda examinar los resultados y eliminar variables que no se diferencian entre diferentes clústers

Supuestos y restricciones

- Este análisis no es una técnica inferencial sino descriptiva que busca cuantificar las características estructurales de una cantidad de individuos
- No hay requerimientos de distribuciones (normalidad, homoscedasticidad, linealidad) pero la multicolinealidad puede afectar los resultados.
- Debe darse énfasis a la representatividad de una muestra para poder generalizar a la población.
- Según lo que se tome en cuenta para realizar la clasificación puede llegarse a obtener diferentes grupos (personas según clase social vs consumo de alcohol)
- Hay clasificaciones más útiles que otras (libros clasificados por materias vs clasificación por el color de su portada).

A tener en cuenta antes del análisis…

- Hay tres aspectos importantes que pueden afectar los resultados de la investigación:

1. Valores extremos
2. Medición de la similitud
3. Estandarización de los datos

Elección de las medidas de asociación

La similitud

- La similitud entre objetos es una medida de correspondencia o semejanza entre los objetos que se piensa agrupar.
- Las características que definen la similitud son predefinidas y se combinan en alguna medida de similitud calculada para todos los pares de objetos.
- Cada objeto puede ser comparado con todos los demás usando esta medida.
- Se agrupan los objetos considerados similares según la medida que se haya escogido.
- Se pueden usar medidas basadas en correlaciones, distancias y asociaciones.

La similitud

- Un ejemplo de similitud se presenta a continuación.

$$D = \begin{pmatrix} 0 & d_{12} & \cdots & d_{1n} \\ d_{21} & 0 & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \\ d_{n1} & d_{n2} & & 0 \end{pmatrix}$$

- ¿Qué notan? ¿Diferencias con matriz de correlación?

Distancias para variables continuas

- Si TODAS las variables son continuas, las medidas de similitud más conocidas son:

1. Distancia Euclídea o euclidiana.
2. Distancia de Mahalanobis.
3. Distancia de Manhattan.
4. Distancia de Minkowsky
5. Distancia de Chebishev

Distancia Euclídea

- La distancia más común es la distancia euclídea o euclideana. Esta se define como:

$$d_{ij} = \sqrt{\sum_{k=1}^{q} (x_{ik} - x_{jk})^2} = \sqrt{(X_i - X_j)^T (X_i - X_j)}$$

donde x_{ik} es el valor que toma la *i-ésima* unidad en la *k-ésima* variable, y X_i es el vector de *q* valores de la *i-ésima* unidad: $X_i^T = (x_{i1}, \ldots, x_{ik}, \ldots, x_{iq})$

Distancia de Mahalanobis

- Esta distancia incluye un procedimiento de estandarización dentro de la distancia Euclídea y ajusta según las correlaciones entre variables:

$$d_{ij}^{M} = (X_i - X_j)^T S^{-1} (X_i - X_j)$$

donde S es la matriz de covariancias de dimensión *qxq*.

- Es invariante frente a transformaciones lineales de las variables y Toma en cuenta las correlaciones entre variables.

Distancia de Manhattan (City-block)

- En esta medida se usa el valor absoluto de las diferencias en lugar de su cuadrado:

$$d_{ij}^{(1)} = \sum_{k=1}^{q} | x_{ik} - x_{jk} |$$

- Esta medida tiene la restricción de que asume que las variables no están correlacionadas

Distancia de Minkowski

- Es el caso más general de la distancia city-block, pero cuenta con un exponente r:

$$d_{ij}^{(r)} = \left(\sum_{k=1}^{q} | x_{ik} - x_{jk} |^{r} \right)^{1/r} ; r > 0$$

Distancia de Chebyshev

- La distancia entre dos puntos (representados por sus vectores) es la mayor de sus diferencias a lo largo de cualquiera de sus dimensiones coordenadas:

$$d_{ij}^{m} = \max\{| x_{ik} - x_{jk} |; k = 1,..., q\}$$

Distancia para variables binarias

- Para variables binarias existe una serie de coeficientes de similitud que se calculan a partir de las coincidencias entre variables.

- En un conjunto de *q* variables, al comparar el *i-ésimo* y *j-ésimo* individuo se cuenta:
  - a_{ij}: número de variables en las que ambos tienen 1.
  - b_{ij}: número de variables en las que sólo uno de los dos tiene 1.
  - c_{ij}: número de variables en las que ambos tienen 0.
  - $q= a_{ij} +b_{ij} +c_{ij}$

- Las distancias se calculan a partir de los coeficientes de similitud haciendo: $d_{ij}=1-s_{ij}$.

Distancia para variables binarias

Las más conocidas son:

- Coeficiente de Jaccard:

$$s_{ij}^{J} = \frac{a_{ij}}{a_{ij} + b_{ij}}$$

- Coeficiente de Russel y Rao:

$$s_{ij}^{RR} = \frac{a_{ij}}{q}$$

- Coeficiente de Sokal y Michener:

$$s_{ij}^{SM} = \frac{a_{ij} + c_{ij}}{q}$$

- Coeficiente de Dice-Sorensen:

$$s_{ij}^{DS} = \frac{2a_{ij}}{2a_{ij} + b_{ij}}$$

Distancia para varios tipos de variables

- Cuando se tienen tanto variables métricas como variables nominales y/o ordinales se recomienda usar la distancia de Gower:

$$d_{ij}^{G} = \frac{1}{q - d_{00}} \left[\sum_{h=1}^{p_1} \frac{|x_{ih} - x_{jh}|}{R_h} + \beta_{ij} \right]$$

donde:

- R_h = rango de la *h-ésima* variable cuantitativa,
- β_{ij} = número de variables cualitativas que no son idénticas,
- q = número total de variables,
- d_{00} = número de variables binarias asimétricas con ambas respuestas en cero. Este caso es menos común pues una variable es binaria asimétrica si el cero no se puede cambiar, por lo que en la mayoría de los casos d_{00}=0.

La importancia de la estandarización

- Variables con mayor dispersión tienen un mayor impacto en el valor final de similitud
- Hay una ponderación implícita de las variables basada en la dispersión relativa cuando se construyen algunas medidas de distancia.

- Se puede realizar la estandarización de varias formas:
  - Convertir cada variable a puntajes estándar (restar la media y dividir por la desviación estándar de la variable)
  - Dividir los puntajes de cada variable por el rango de la misma
  - Estandarizar cada caso dentro de sí mismo (se resta la media del individuo y se divide por la desviación estándar del individuo) para remover efectos de respuesta alta o baja por individuo. Esto es útil en datos de actitudes.

Distancia entre grupos

- A partir de una matriz de distancias entre individuos se puede derivar el concepto de distancia entre dos grupos.
- Hay varias formas de definir la distancia entre dos grupos:

❑ Vecino más cercano

$$\delta(A,B) = \min\{d(x_i, x_j); x_i \in A, x_j \in B\}$$

❑ Vecino más lejano

$$\delta(A,B) = \max\{d(x_i, x_j); x_i \in A, x_j \in B\}$$

❑ Salto promedio

$$\delta(A,B) = \frac{1}{n_A n_B} \sum_{x_i \in A, x_j \in B} d(x_i, x_j)$$

❑ Distancia de Ward

donde g_A es el centroide del clúster A

$$\delta(A,B) = \frac{n_A n_B}{n_A + n_B} \left\| g_A - g_B \right\|^2$$

Elección de la técnica de Cluster

Técnicas de agrupamiento

Hay varios tipos de métodos de realizar el agrupamiento de objetos. Entre ellos están:

- Técnicas jerárquicas aglomerativas.
- Agrupamiento por k-medias
- Agrupamiento basado en modelos (no lo vamos a ver).

Agrupamiento jerárquico

- Consiste en realizar una partición única de los datos en un número específico de clases o grupos en un único paso. Se realiza una serie de particiones a partir de *n* clústers donde cada uno contiene sólo un individuo.

- Se hacen fusiones sucesivas de los clústers ya formados. Las fusiones son irreversibles, es decir, si dos individuos ya se han puesto en un mismo clúster, éstos no podrán aparecer en diferentes grupos en etapas subsecuentes.

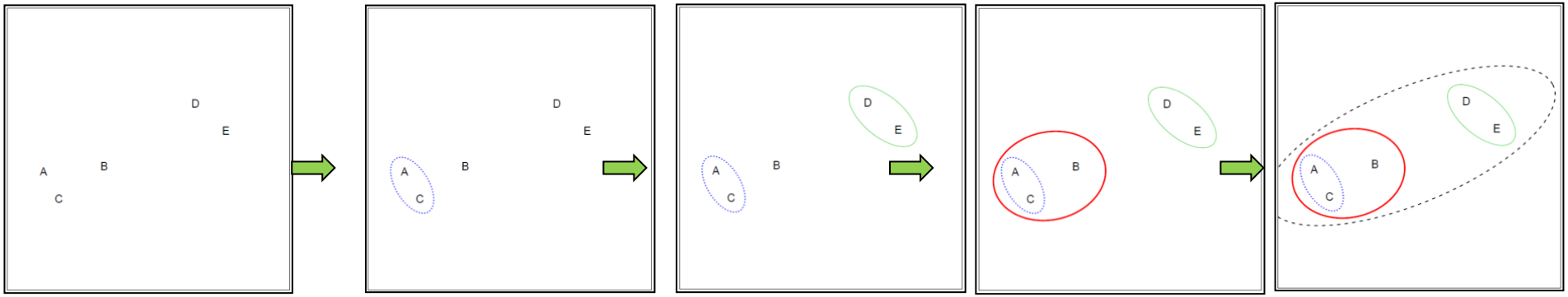
- Al final todos llegan a fusionarse en un solo grupo. Se debe decidir cuál es el número de clústers que da el mejor ajuste.

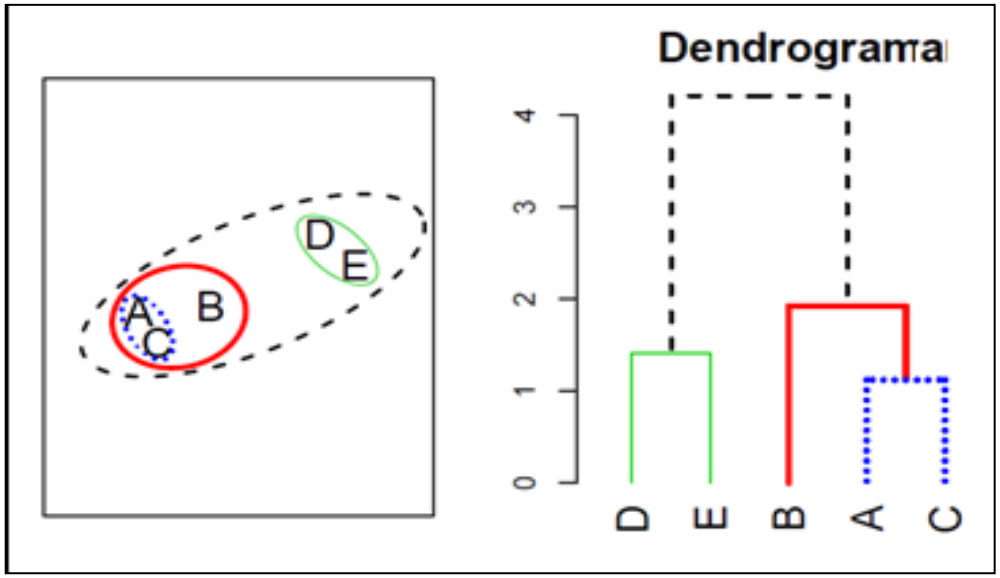
Agrupamiento jerárquico: dendograma

- Un dendrograma es un tipo de representación gráfica o diagrama de datos en forma de árbol que organiza los datos en subcategorías que se van dividiendo hasta llegar al nivel de detalle deseado (asemejándose a las ramas de un árbol que se van dividiendo en otras sucesivamente).

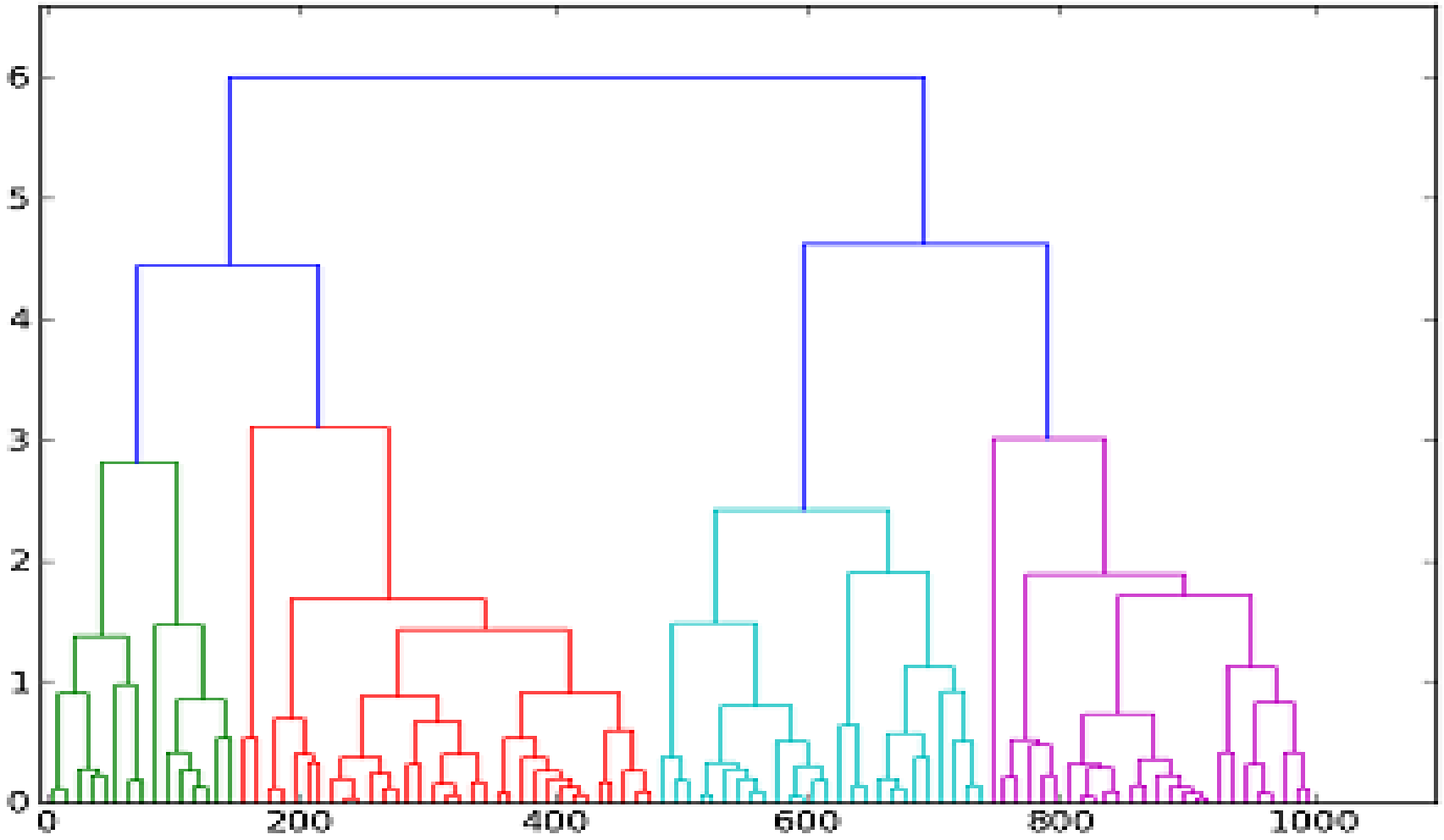
- Permite apreciar las relaciones de agrupación entre los datos e incluso entre grupos de ellos. Observando las sucesivas subdivisiones podemos hacernos una idea sobre los criterios de agrupación de los mismos, la distancia entre los datos según las relaciones establecidas, etc.

Construcción y representación





Dendograma



Poda del dendograma

- Una vez obtenido el dendograma se requiere escoger una partición de los objetos a partir del gráfico.

- Se traza una línea horizontal a cierta altura y de ahí se obtiene un cierto número de ramas separadas.

- De forma informal se puede examinar el tamaño de los cambios en altura en el dendograma y se toma un cambio «grande» para poner ahí la línea horizontal

Representación con el PCA

- Una vez que se ha decidido podar en un punto el dendograma, se forman los clústers sugeridos.
- Se obtienen los dos primeros componentes principales a partir de las variables incluidas en el análisis.
- Se hace un gráfico con las puntuaciones de los dos primeros componentes y se ponen colores según los clústers establecidos.
- El método del vecino más cercano tiende a presentar un problema de encadenamiento: tendencia a incorporar puntos intermedios dentro de un clúster ya existente.

Método de k-medias

- Para un número determinado *k* se busca la partición que minimiza algún criterio de decisión, para el cual tener valores más bajos es una indicación de una buena solución.
- El criterio más común es la suma de cuadrados dentro de grupo:

$$SCDG = \sum_{h=1}^{k} \sum_{i \in G_h} \sum_{j=1}^{q} \left(x_{ij} - \bar{x}_j^{(l)} \right)^2$$

donde los valores se comparan con el promedio de la variable en el clúster a que pertenecen

- Si se intenta considerar todas las posibles particiones del conjunto de datos en *k* grupos se puede llegar a números extremadamente altos de posibilidades, lo cual hace el problema computacionalmente impráctico

Algoritmo del método de k-medias

- Paso inicial: se toma una partición arbitraria en *k* clústers que puede venir del resultado de otro método de agrupamiento.

1) Se calcula el criterio (SCDG u otro) producido al mover cada individuo del clúster en que se encuentra hacia otro clúster.
2) Se realiza el cambio que lleva a la mayor disminución en el criterio.
3) Repetir pasos 1) y 2) hasta que ningún movimiento produzca disminuciones en el criterio.

Desventaja del método k-medias

- Este método no es invariante ante cambios de escala.

- Impone una estructura esférica en los datos aún cuando no se comporten de esa forma.

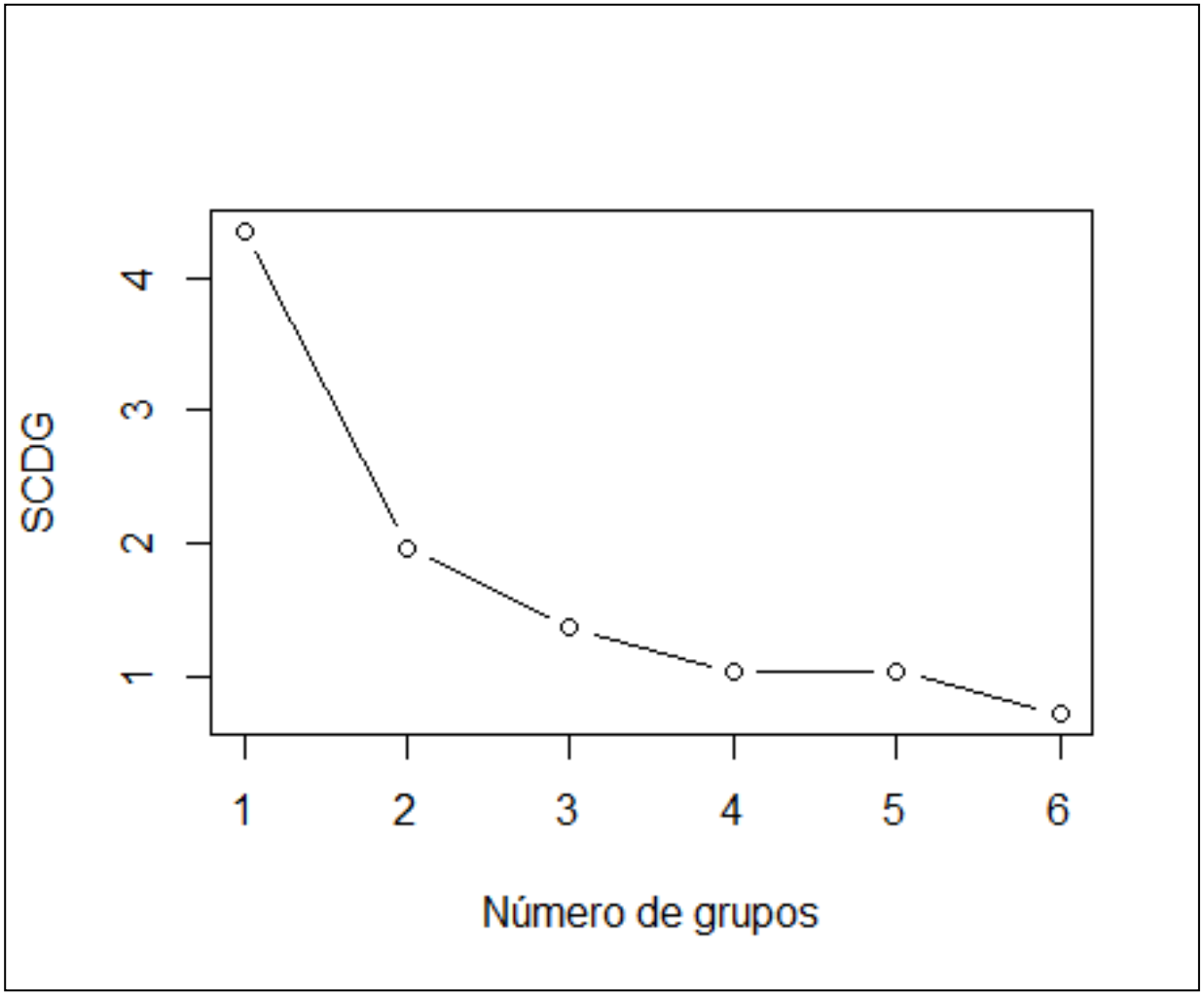
Selección del número de clústers

- Para determinar un número adecuado *k* de clústers se busca el valor mínimo de la SCDG para un *k* fijo, y se va moviendo el valor de *k*.

- Conforme el número *k* crece, la SCDG siempre decrece.

- Se grafican los valores de SCDG contra *k* y se busca una forma de «codo» en el gráfico como un indicador del punto más adecuado.

Selección del número de clústers



Análisis de los clusters finales

- Una vez que se han hecho los agrupamientos, existen varias formas de presentar los resultados:

1. Componentes principales
2. Sombras y vecindarios
3. Rayas
4. Perfiles


The End