Aprendizaje no supervisado

Gonzalo Rivero

Big Data en investigación social y opinión pública

22 a 26 de julio, 2019

Aprendizaje no supervisado

- La segunda mitad de este curso se centrará en *aprendizaje supervisado.*
  - Modelos de regresión
  - Modelos de clasificación
- En un problema supervisado, tenemos observaciones de variables X_1, X_2, ... X_p para cada objeto y también tenemos una variable resultado Y. El objetivo es predecir Y usando X_1, X_2, ... X_p
- Ahora nos centraremos en *aprendizaje no supervisado,* en el que solo observamos X_1, X_2, ... X_p. No intentaremos hacer predicción porque no tenemos la variable de respuesta.

¿Para qué es útil el aprendizaje no supervisado?

- El objetivo es descubrir cosas interesantes sobre los datos que hemos obtenido.
  - ¿Existen subgrupos en las observaciones?
  - ¿Podemos agrupar las variables de alguna manera?
  - ¿Cuál es la mejor forma de resumir la información con menos variables?
  - ¿Como podemos visualizar los datos?
- Discutiremos tres modelos:
  - *Análisis de componentes principales,* una herramienta usada para visualización de datos o como pre-procesamiento aplicado antes de modelos supervisados.
  - *Modelos de escalado multidimensional,* unas técnicas para visualizar el nivel de similaridad entre observaciones.
  - *Análisis de conglomerados,* un conjunto de técnicas usadas para descubrir subgrupos en los datos.

Las dificultades del aprendizaje no supervisado

- Es más subjetivo que los modelos supervisados y no hay un objetivo claro del análisis, como en el caso supervisado
- Por tanto es más difícil de evaluar si el modelo ha funcionado bien
- Pero las técnicas supervisadas son útiles como exploración de los datos
- Y también como forma de resolver problemas que naturalmente no tienen variable respuesta:
  - Identificar grupos de votantes definidos por características sociodemográficas, políticas o geográficas.
  - Determinar qué películas son parecidas en función de cómo han sido evaluadas por los consumidores.

Otra ventaja

- Por lo general es más sencillo conseguir *datos no etiquetados* que *etiquetados*, que pueden necesitar participación humana.
- Además, los datos etiquetados no siempre facilitan el problema. Por ejemplo, para evaluar si comentarios de una película son positivos o negativos.

Análisis de Componentes Principales

- PCA produce una representación de los datos en menos dimensiones que las originales.
- Idea fundamental es que algunas dimensiones no son interesantes o presentan redundancia. Encuentra combinación lineal de las variables que tienen mayor varianza y que no están correlacionadas entre sí.
- Puede servir para crear nuevas variables que resumen la base de datos y utilizarlas como input para tareas supervisadas. También útil como herramienta de visualización.

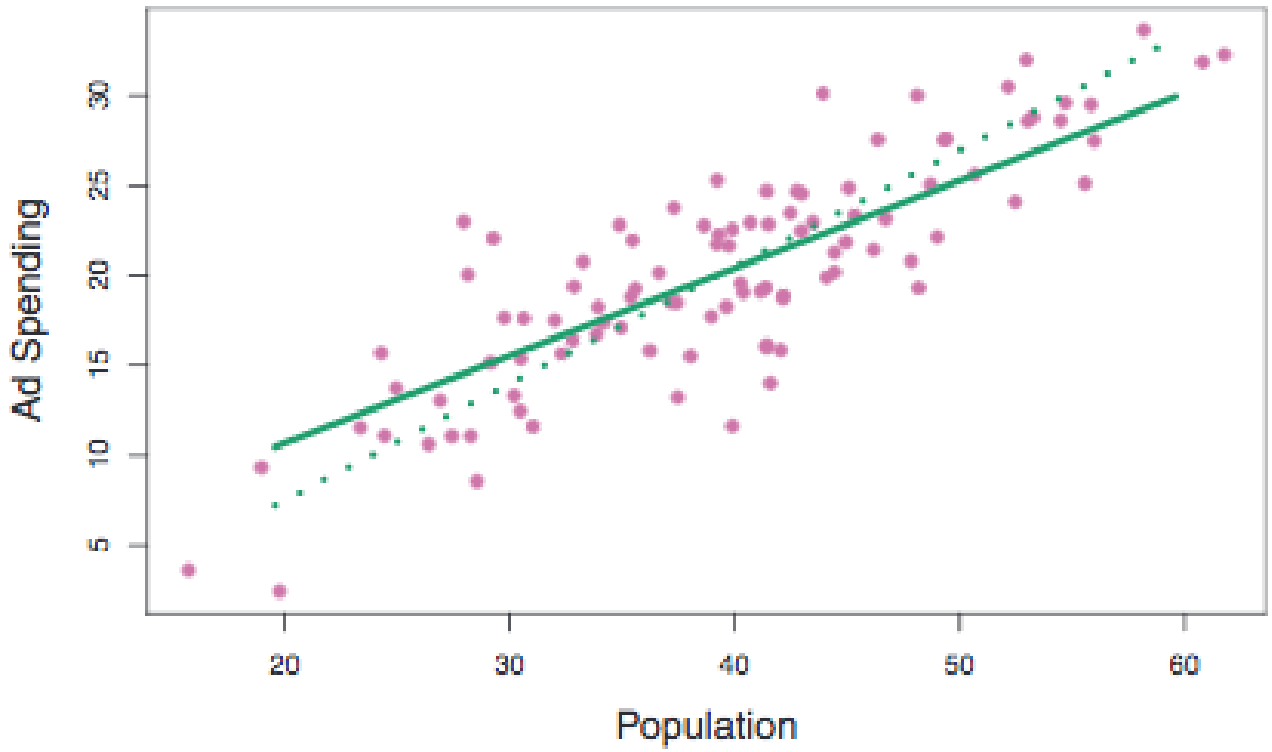
Análisis

- El primer componente de un conjunto de variables X_1, X_2, ... X_p es una combinación *normalizada* de las variables

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p \tag{1}$$

  tal que $\sum_j \phi_{j1}^2 = 1$ (normalización) y tenga la mayor varianza posible.
- Los elementos $\phi_{11}, \ldots, \phi_{p1}$ son las *cargas* del componente principal.
- Imponemos la restricción en las cargas ya que si no podríamos hacer que las cargas fuesen arbitrariamente grandes y por tanto darnos varianzas arbitrariamente grandes.

Ejemplo



Tamaño de la población y gasto en anuncios para cien ciudades. La línea verde indica la dirección del primer componente principal.

Cálculo de los componentes principales

- Tenemos datos X de tamaño $n \times p$. Solo estamos interesados en la varianza, así que asumimos que X está centrada (las medias de las columnas son cero).
- Intentamos calcular la combinación lineal de las variables

$$z_{i1} = \phi_{11} x_{i1} + \phi_{21} x_{i2} + \cdots + \phi_{p1} x_{ip} \tag{2}$$

  para $i = 1, \ldots, n$ tal que z_{i1} tenga la mayor varianza, con la condición de que $\sum_j \phi_{j1}^2 = 1$.
- Ya que $x_i j$ tiene media cero, también tiene media cero $z_i 1$. por tanto, la varianza de z_{i1} es $\frac{1}{n} \sum_i z_{i1}^2$

Cálculo de los componentes principales (continuación)

- Sustituyendo

$$\max_{\phi_{11},\ldots,\phi_{p1}} \frac{1}{n} \sum_i \left(\sum_j \phi_{j1} x_{ij} \right)^2 \tag{3}$$

  sujeto a $\sum_j \phi_{j1}^2 = 1$.
- Puede resolverse mediante descomposición en valores singulares.

Interpretación geométrica

- El vector carga ϕ_1 con elementos $\phi_{11}, \ldots, \phi_{p1}$ define la dirección en el espacio de las variables que tienen la mayor variación.
- Si proyectamos los n datos x_1, x_2, ... x_p en esa dirección, el resultado son la puntuación del componente principal.

Otros componentes principales

- El segundo componente principal es la combinación lineal de variables X_1, X_2, ... X_p que tenga mayor varianza y que no esté correlacionada con Z_1.
- El segundo componente es

$$z_{i2} = \phi_{12}x_{i1} + \phi_{22}x_{i2} + \cdots + \phi_{p2}x_{ip} \tag{4}$$

en donde $\phi_{12}, \ldots, \phi_{p2}$ son las *cargas* del segundo componente principal.

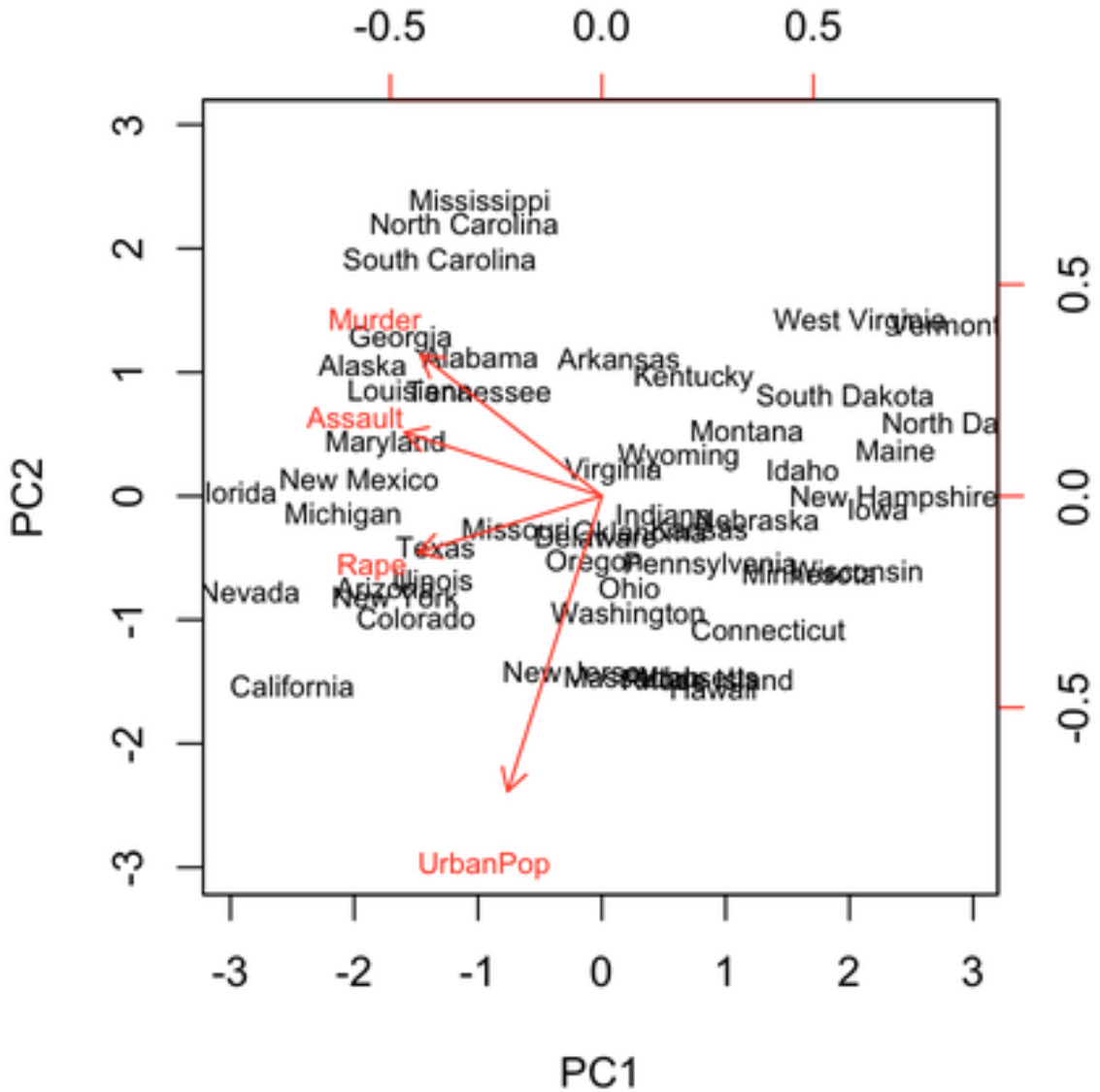
Otros componentes principales

- Constreñir Z_2 a que no esté correlacionado con Z_1 es lo mismo que imponer que ϕ_2 sea ortogonal a ϕ_1. Lo mismo para el resto de componentes.
- Pueden existir como mucho $\text{mín}(n-1, p)$ componentes.

Ejemplo

- Base de datos `USAarrests`: Para cada estados de Estados Unidos, los datos contienen el número de detenciones por cada 100.000 residentes en tres crímenes: `Assault`, `Murder`, y `Rape`. La base incluye también `UrbanPop` (porcentaje de la población en áreas urbanas).
- Los vectores de componentes principales tienen $n = 50$ y los vectores de cargas de los componentes principales tienen $p = 4$.
- Normalizamos los datos de cada variable para que tengan media 0 y desviación típica unidad.

Ejemplo


-0.5
0.0
0.5
3
2
1
0
-1
-2
-3
PC2
Mississippi
North Carolina
South Carolina
Murder
Georgia
Alaska
Alabama
Arkansas
Kentucky
South Dakota
Assault
Maryland
Montana
Wyoming
Maine
New Mexico
Virginia
Idaho
Michigan
Iowa
Texas
Rape
Illinois
Nevada
Ohio
Washington
Connecticut
Colorado
California
Hawaii
UrbanPop
-3
-2
-1
0
1
2
3
PC1
0.5
0.0
-0.5

Ejemplo

- La figura muestra los dos componentes principales de la base de datos.
- Los nombres de los estados representan las puntuacións para cada uno de los dos componentes
- Las líneas rojas indican los dos componentes principales. Por ejemplo, la carga de Rape en el primer componente es -0.54 y la carga en el segundo componente es -0.17.
- La figura es un *biplot* porque representa tanto los pesos de los componentes como las cargas.

Otra interpretación de los componentes principales

- El primer componente principal define la superficie que es *más cercana* a las observaciones usando la distancia Euclídea media cuadrática.
- La misma interpretación se puede aplicar al resto de las dimensiones.

Precauciones en la intepretación

- Cada componente es único, con la excepción de un cambio de signo.
- Si las unidades están en diferentes unidades es recomendable escalarlas para que tengan desviación unitaria.

Proporción de varianza explicada

- Para entender cómo cada componente captura la variación en los datos necesitamos la *proporción de varianza explicada*.
- La *varianza total* en los datos es

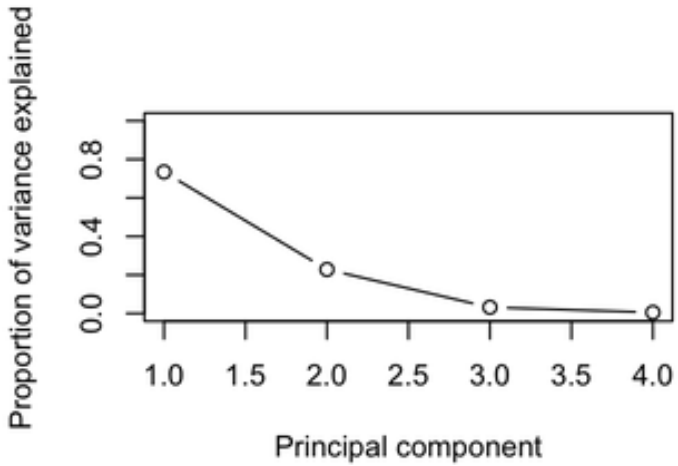
$$\sum_j var(X_j) = \sum_j \frac{1}{n} \sum_i x_{ij}^2 \tag{5}$$

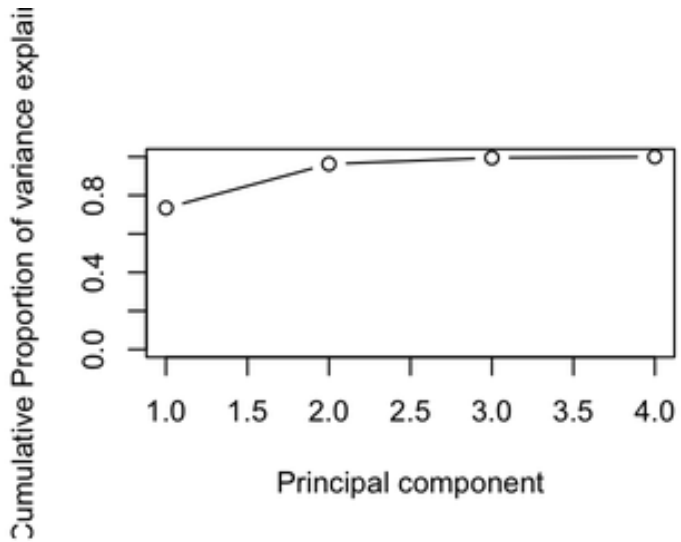
y la varianza explicada por el m-ésimo componente es

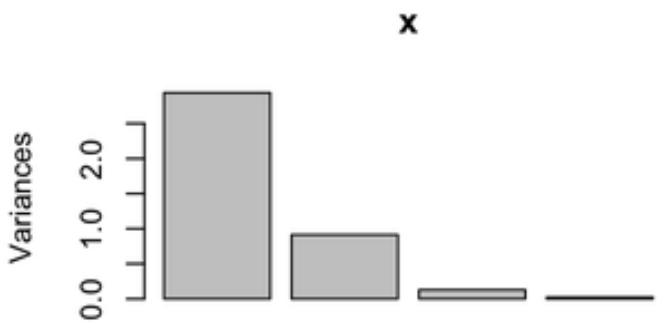
$$var(Z_m) = \frac{1}{n} \sum_i z_{im} \tag{6}$$

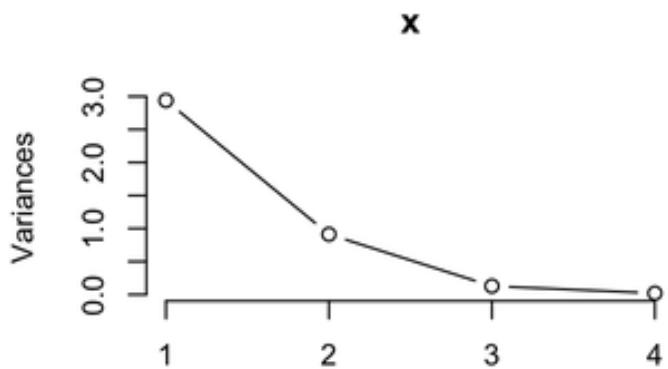
- Además, $\sum_j var(X_j) = \sum_j var(Z_m)$, para el número total m de componentes que puedan calcularse.
- Podemos asociar a cada componente una proporción de la varianza total.

Proporción de varianza explicada









¿Cuántos componentes retener?

- Podemos usar la proporción de varianza explicada asociada a un número de componentes
  - Fijamos una proporción que queremos alcanzar
  - Inspeccionamos descensos marginales en varianza explicada
  - En la práctica, tendemos a retener pocos
  - Buscamos "esquinas" en el *scree plot* anterior.

Análisis de conglomerados

- Ténicas que intentan buscar subgrupos de observaciones en los datos.
- Cada grupo contiene observaciones que son similares entre ellas y diferentes a las demás
- Sujeto a una definición específica de qué significa *diferente* y *similar* en un determinado contexto

Segmentación de mercados

- Una situación habitual es tener que agrupar consumidores de los que tenemos una gran cantidad de información
- Nuestro objetivo es segmentar el mercado de tal manera que podamos identificar los individuos que son más propensos a comprar un producto
- Queremos agrupar a individuos similares entre ellos de tal forma que cada grupo sea lo más homogéneo internamente y distintivo.

Formas de obtener conglomerados

- Empezamos por especificar el número de grupos e intentamos asignar observaciones a cada grupo: *análisis de k-medias*.
- Agrupamos las observaciones por pares que son lo más parecidos posibles y definimos qué agrupación, usando un *dendrograma*, es la que mejor representa los datos: *análisis jerárquico*

Modelo de k-medias

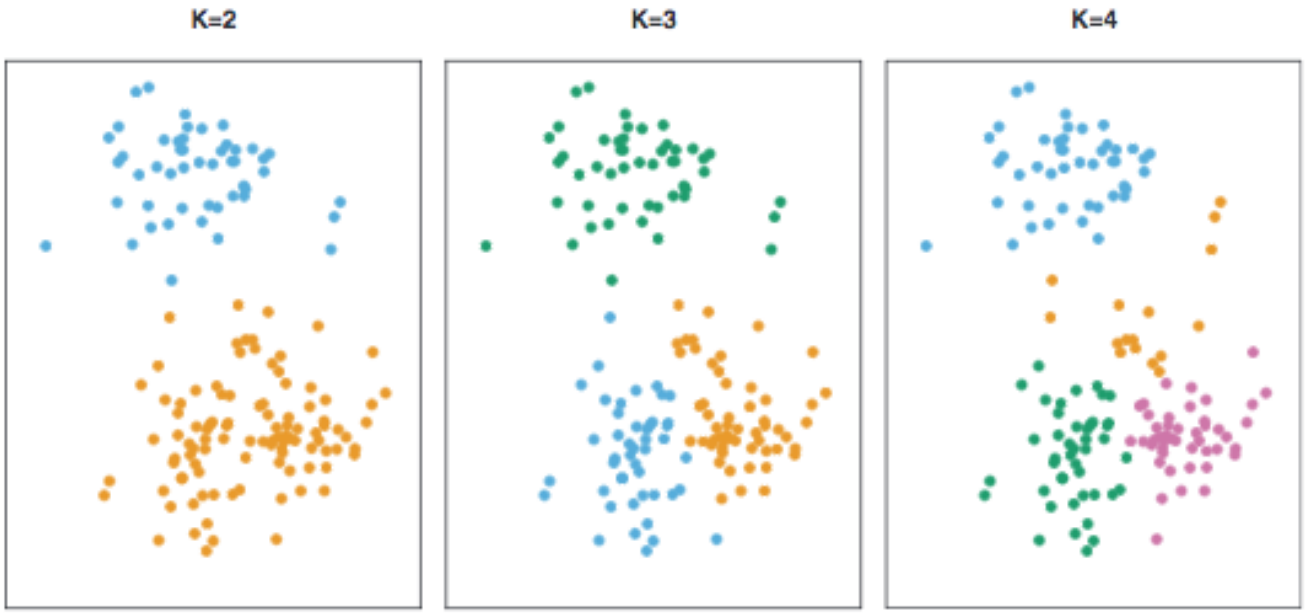
- Usuario especifica número de grupos y algoritmo determina a qué grupo pertenece cada observación.
- Objetivo es crear grupos que tengan menor variación dentro de cada grupo.

$$\min_{C_k} \left\{ \sum_k W(C_k) \right\} \tag{7}$$

con C_k siendo un grupo y W la variación dentro del grupo

- Una medida típica de W es el cuadrado de la distancia euclidea entre todas las observaciones del grupo.
- Hay K^n posibles particiones.

Método de k-medias



150 puntos en dos dimensiones. Las figuras muestran conglomerados para diferentes números de k. El color es arbitrario, así como la identidad de cada grupo.

Un algoritmo para el modelo de k-medias

1. Inicializar aleatoriamente la asignación de las observaciones en grupos 1 a K grupos.
2. Iterar hasta convergencia
   - 2.1 Para cada grupo, calcular el *centroide*, el vector de medias de cada una de las dimensiones.
   - 2.2 Asignar cada observación al grupo con el centroide más cercano (usando distancia Euclídea).
3. El algoritmo produce óptimos locales que dependen de la asignación inicial. Es conveniente repetir con inicializaciones aleatorias.

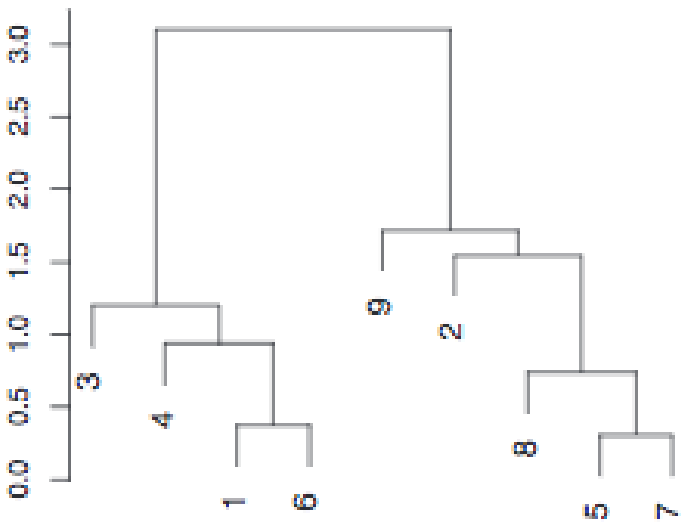
Modelos jerárquicos

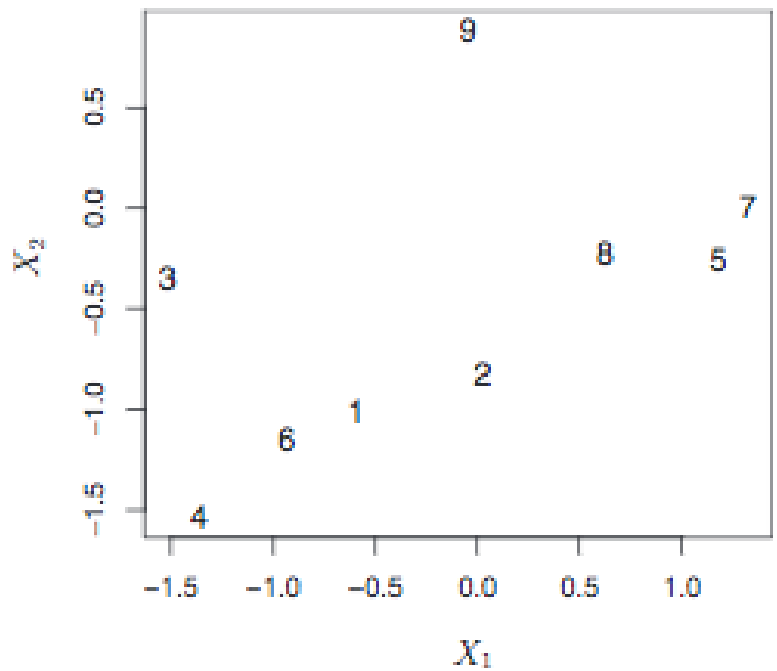
- Una desventaja del método de k-medias es que requiere especificar el número de grupos.
- Los métodos jerárquicos evitan este problema.
- También permite visualizar los datos de forma muy sencilla mediante un *dendrograma*
- La forma más habitual de construir el modelo jerárquico es el método *aglomerativo*: la construcción del dendrograma empieza desde abajo.

Crear las agrupaciones

1. Definir una distancia entre todos los pares de observaciones
2. Examinar las distancias e identificar los pares que son más cercanos un fusionarlos.
3. Computar la nueva distancia entre cada grupo.
4. Terminar cuando exista una única agrupación

El dendrograma





El dendrograma

- A medida que nos movemos hacia arriba, las observaciones se fusionan.
- La fusión ocurre para observaciones/grupos que son similares
  - Si se fusionan abajo son muy similares
  - Si se fusionan arriba son muy diferentes
  - Proximidad horizontal es irrelevante
- Podemos identificar grupos con un corte horizontal
- Observaciones debajo del corte constituyen grupos
- Un dendrograma puede construir cualquier número de grupos

Maneras de fusionar grupos

- Cuatro aproximaciones
  - *Completa:* Mayor disimilitud entre grupos. Usar la mayor disimilitud entre pares de observaciones dentro de cada grupo.
  - *Única:* Menor disimilitud entre grupos. Usar la menor disimilitud entre pares de observaciones dentro de cada grupo.
  - *Media:* Usar la disimilitud media entre pares de observaciones dentro de cada grupo.
  - *Centroide:* Calcular un centroide para grupo (poco usada).
- Podemos usar métricas diferentes a la euclídea, como correlación.
- Escalar las variables es muy importante.

¿Cuántos grupos escoger?

- Mismo tipo de aproximación que con componentes principales:
  - Calcular medida en la que los grupos "explican" los datos
  - Escoger número de grupos que produzca menor incremento en varianza total.
- Ambos modelos son muy sensibles a observaciones que no pertenecen a ningún grupo. Puede ser necesario un preprocesado.