Análisis multivariante

00R Team

Marzo 2017

Análisis multivariante

Introducción

- Métodos estadísticos para analizar múltiples medidas
- Reducir la dimensión de los datos
- Facilitar interpretación y representación
- Clasificar a los individuos en grupos internamente homogéneos

Matriz de datos

$$X = \begin{bmatrix} x_{11} & x_{12} & ... & x_{1p} \\ x_{21} & ... & ... & x_{22} \\ ... & ... & ... & ... \\ x_{n1} & x_{n2} & ... & x_{np} \end{bmatrix}$$

Donde la variable V_1

$$V_{.1} = [x_{11}, x_{21}, ..., x_{n1}]$$

y la observación w_1.

$$w_{1.} = [x_{11}, x_{12}, ..., x_{1p}]$$

Ejemplo de matriz de datos

	Age	Head.L	Head.W	Neck.G	Length
1	70	15.0	6.5	28	78.0
2	8	10.0	4.5	10	43.5
3	19	10.0	5.0	15	45.0
4	45	13.0	6.5	21	60.0
5	19	11.0	6.5	20	47.5
6	21	14.5	5.5	20	61.0

Distancias. Distancia Euclídea

$$\alpha\gamma\epsilon\omega\mu\epsilon\tau\rho\eta\tau o\zeta\ \mu\eta\delta\epsilon\iota\zeta\ \epsilon\iota\sigma\iota\tau\omega$$

Distancias. Distancia Euclídea

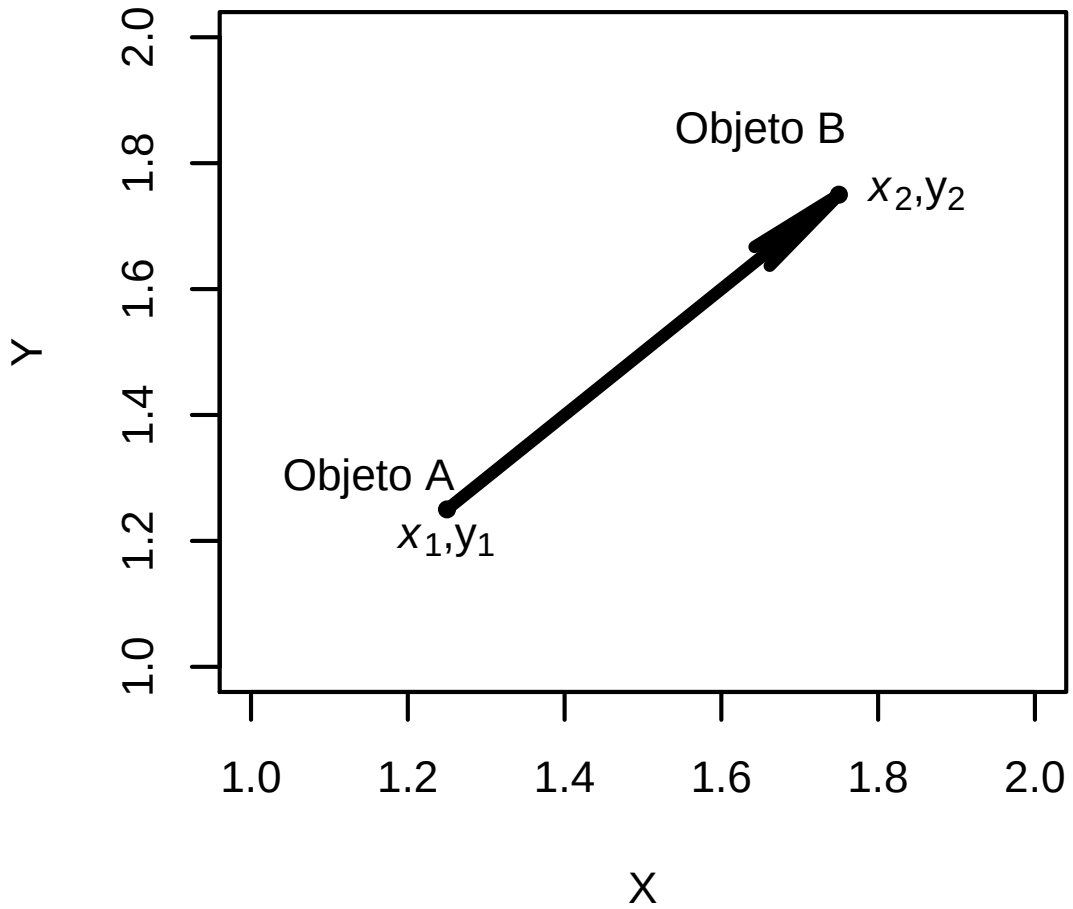
$$\alpha\gamma\epsilon\omega\mu\epsilon\tau\rho\eta\tau o\zeta\ \mu\eta\delta\epsilon\iota\zeta\ \epsilon\iota\sigma\iota\tau\omega$$

`Aquí no entra nadie que no sepa geometría`[1]

[1] Esta inscripción figuraba en la escuela de filosofía de Atenas y refleja la importancia que en la Grecia clásica se le daba a las matemáticas.

Distancias. Distancia Euclídea

En matemáticas la distancia entre dos puntos en un espacio euclídeo equivale a la longitud del segmento que une ambos puntos.



Distancias. Otras distancias y disimilaridades

- Distancia euclídea entre dos puntos

$$d_E(i,j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

- Distancia de Manhattan

$$d_M(i,j) = \sum_{k=1}^{n} |i_k - j_k|$$

- Distancia de Mahalanobis

$$d_m(i,j) = \sqrt{(i-j)^T \sum {}^{-}1(i-j)}$$

Medidas de disimilaridades

- Covarianza

$$\text{cov}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

- Correlación

$$r_{xy} = \frac{\text{cov}(x, y)}{s_x \ s_y}$$

- Correlación convertido a distancia

$$\text{dist}(x, y) = 1 - \frac{r_{xy} + 1}{2}$$

- Índice de Jaccard

$$I_J = \frac{a}{a + b + c}$$

Matriz de covarianzas

```
          Head.W Neck.G  Length Chest.G   Weight
Head.W     2.10   6.34   10.49    9.86   125.60
Neck.G     6.34  29.94   50.01   46.11   582.97
Length    10.49  50.01  112.75   86.57  1063.51
Chest.G    9.86  46.11   86.57   84.01  1011.26
Weight   125.60 582.97 1063.51 1011.26 13077.67
```

```
[1] "Varianza total:  13306.4762662338"
```

Suma de los elementos (varianzas) de la diagonal, representa la información contenida en la matriz.

Las covarianzas representan la información redundate. Si $covarianza \neq 0$ hay redundancia de información.

Normalización/Tipificación

La más común es $\frac{x_i - \overline{x}}{\sigma}$

```
      Head.L Head.W Neck.G Length Chest.G Weight
[1,]   0.95   0.21   1.36   1.79    1.04   1.34
[2,]  -1.42  -1.17  -1.93  -1.46   -1.25  -1.32
[3,]  -1.42  -0.83  -1.02  -1.31   -1.36  -1.01
[4,]   0.00   0.21   0.08   0.10   -0.10   0.01
[5,]  -0.95   0.21  -0.11  -1.08   -1.25  -0.97
```

Matriz de varianzas de datos tipificados

```
round( var( dfN[ , 1:5 ] ), 2 )
```

```
         Head.L Head.W Neck.G Length Chest.G
Head.L     1.00   0.71   0.88   0.91    0.87
Head.W     0.71   1.00   0.80   0.68    0.74
Neck.G     0.88   0.80   1.00   0.86    0.92
Length     0.91   0.68   0.86   1.00    0.89
Chest.G    0.87   0.74   0.92   0.89    1.00
```

Matriz de correlaciones

```
round( cor( osos[ ,c( 3:7 ) ] ), 2 )
```

```
         Head.L Head.W Neck.G Length Chest.G
Head.L     1.00   0.71   0.88   0.91    0.87
Head.W     0.71   1.00   0.80   0.68    0.74
Neck.G     0.88   0.80   1.00   0.86    0.92
Length     0.91   0.68   0.86   1.00    0.89
Chest.G    0.87   0.74   0.92   0.89    1.00
```

Ordenación

Análisis de componentes principales

PCA

- Punto de partida -> matriz de n (observaciones) x p (variables)
- Objetivo y características:
  - Representar la información con menos variables
  - Variables son cuantitativas
  - Se obtienen nuevas variables (componentes principales)
  - Combinación lineal de las originales
  - Incorreladas entre sí.

Análisis de correspondencias

CA

- Punto de partida una tabla de contingencia
- Adecuado para variables categóricas
- La técnica permite visualizar y analizar patrones de asociación entre variables categóricas

CA: tabla de contingencia

	Género	Color Pelo
1	M	Ca
2	M	Ca
3	M	Mo
4	M	Ca
5	H	Ru
6	M	Pe

	Ca	Mo	Pe	Ru	Sum
H	12	14	8	13	47
M	16	8	15	14	53
Sum	28	22	23	27	100

CA: tabla de contingencia II

Tabla de contingencia con frecuencias marginales

	Ca	Mo	Pe	Ru	F_abs	F_rel
H	12	14	8	13	47	0.47
M	16	8	15	14	53	0.53
F_abs	28	22	23	27	100	1
F_rel	0.28	0.22	0.23	0.27	1	

CA: tablas de distribuciones condicionadas

- Distribuciones condicionadas por filas y por columnas
- Se calculan dividiendo cada elemento por el total de su fila (o columna)
- Sirven para comprobar la independencia entre las variables mediante χ^2
  - Variables serán independientes cuando los perfiles sean iguales
  - Ho: No hay diferencia entre perfiles filas/columnas (independencia $p.valor > 0.05$)

CA: tablas de distribuciones condicionadas II

Table 4: Dist. condicionadas columnas

	Ca	Mo	Pe	Ru
H	12/28	14/22	8/23	13/27
M	16/28	8/22	15/23	14/27

Table 5: Dist. condicionadas filas

	Ca	Mo	Pe	Ru
H	12/47	14/47	8/47	13/47
M	16/53	8/53	15/53	14/53

CA: tablas de distribuciones condicionadas III

```
      Color Pelo
Género        Ca        Mo        Pe        Ru
     H 0.4285714 0.6363636 0.3478261 0.4814815
     M 0.5714286 0.3636364 0.6521739 0.5185185

      Color Pelo
Género        Ca        Mo        Pe        Ru
     H 0.2553191 0.2978723 0.1702128 0.2765957
     M 0.3018868 0.1509434 0.2830189 0.2641509


    Pearson's Chi-squared test

data:  tabla
X-squared = 4.0298, df = 3, p-value = 0.2583
```

Escalado multidimensional

MDS

- Aplicación del PCA cuando la matriz de datos es una matriz de distancia o similitudes.
- Es una representación espacial de las relaciones entre individuos en función de sus variables
- Coordenadas principales $\approx$ Componentes principales
- Si distancias euclídeas: escalado métrico
- Distancias no euclídeas, similitudes: escalado no métrico

MDS: La matriz de datos

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5.0	3.6	1.4	0.2
6	5.4	3.9	1.7	0.4
7	4.6	3.4	1.4	0.3
8	5.0	3.4	1.5	0.2

MDS: La matriz de distancia

```
      1     2     3     4     5     6     7
2 0.539
3 0.510 0.300
4 0.648 0.332 0.245
5 0.141 0.608 0.510 0.648
6 0.616 1.091 1.086 1.166 0.616
7 0.520 0.510 0.265 0.332 0.458 0.995
8 0.173 0.424 0.412 0.500 0.224 0.700 0.424
```

Análisis factorial

FA: ideas

- Partimos de la idea de la existencia de fuertes correlaciones entre variables.
- PCA busca factores que explique la mayor parte de la varianza total.
- FA busca factores que expliquen la mayor parte de la varianza común (covarianza)
- Se suele utilizar, para analizar variables de naturaleza abstracta (factores), sólo medibles de forma indirecta.
- Matriz de datos es similar a la empleada en PCA

Clasificación

Características

- Agrupar individuos en grupos internamente homogéneos
- Individuos del mismo grupo sean similares según el criterios de clasificación
- El resultado será una partición de los individuos en k grupos (iterativa)
- O bien se establecerá una estructura jerárquica de los datos (jerárquica)
- Se parte de una matriz de distancias o similitudes

Clasificación iterativa

Kmeans cómo funciona

Vídeo Explicativo Kmeans

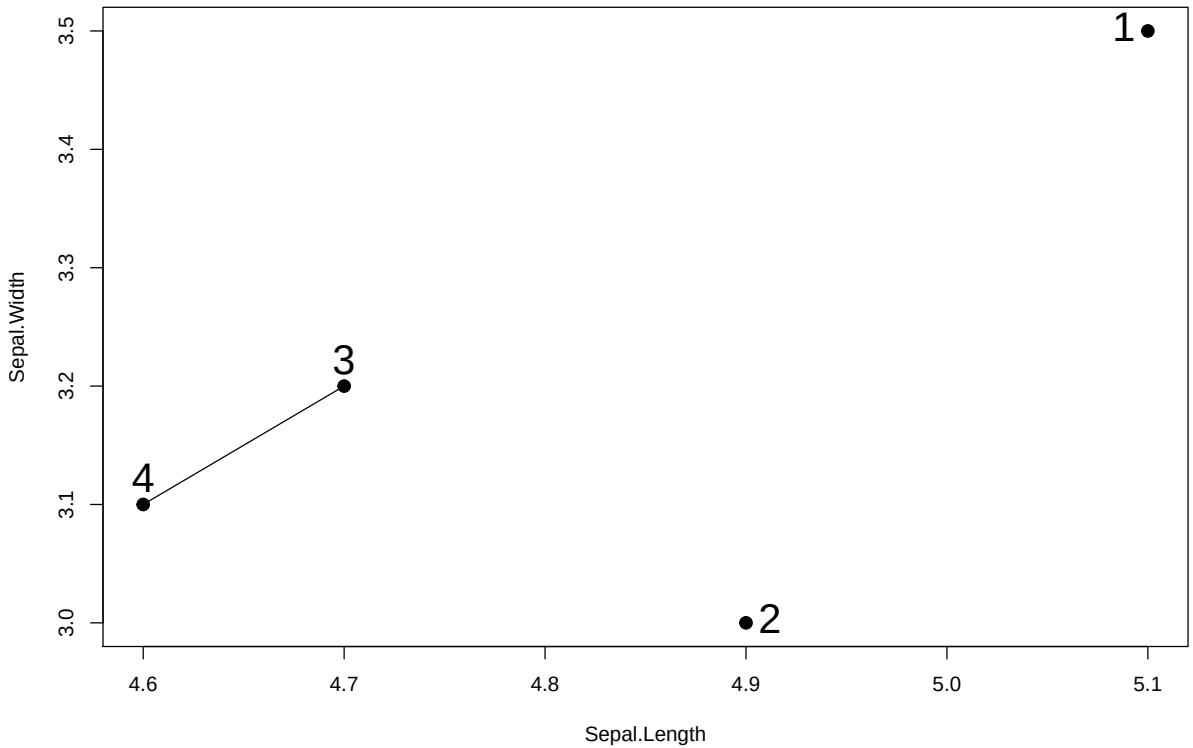
Clasificación jerárquica

Características

- Se crea una estructura jerárquica basada en las distancias entre individuos
- Aglomerativa: cada observación es su propio grupo y los grupos se van mezclando
- Divisiva: todas las observaciones están en el mismo grupo y en cada iteración se van dividiendo los grupos.

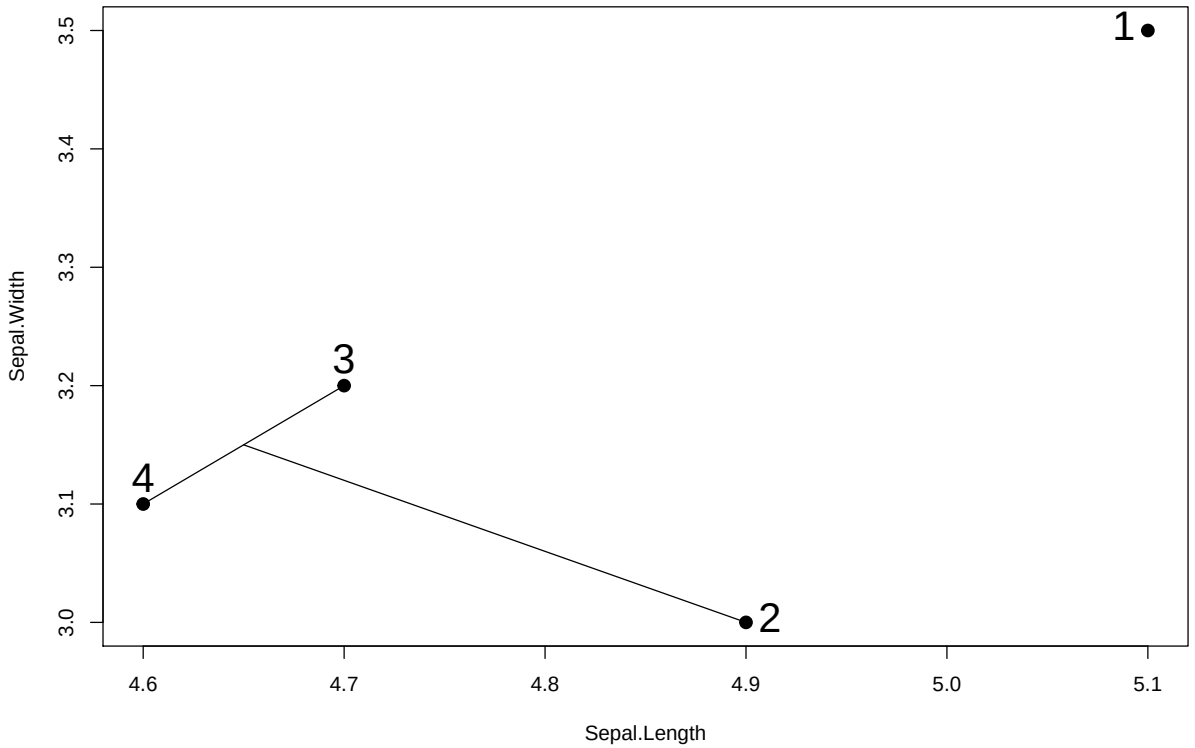
Cómo funciona: iter 1

```
          1         2         3
2 0.5385165
3 0.5000000 0.2828427
4 0.6403124 0.3162278 0.1414214
```



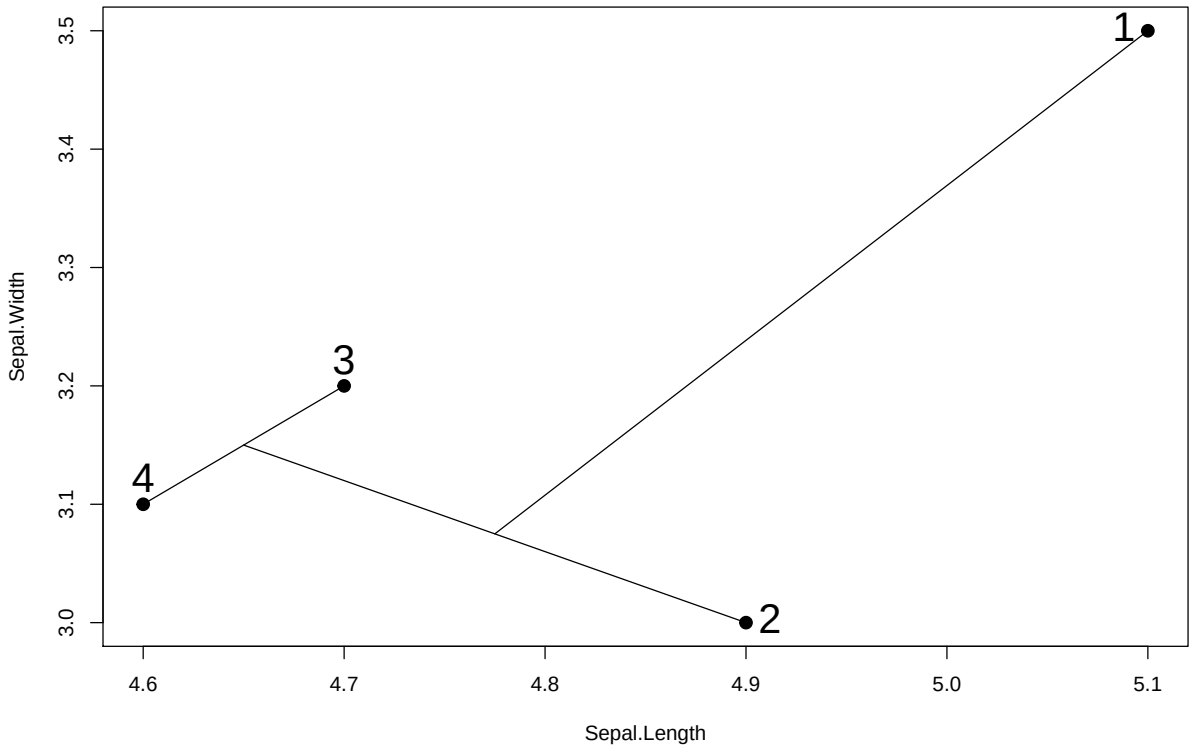
Cómo funciona: iter 2

	1	2
2	0.5385165	
3-4	0.5700877	0.2915476

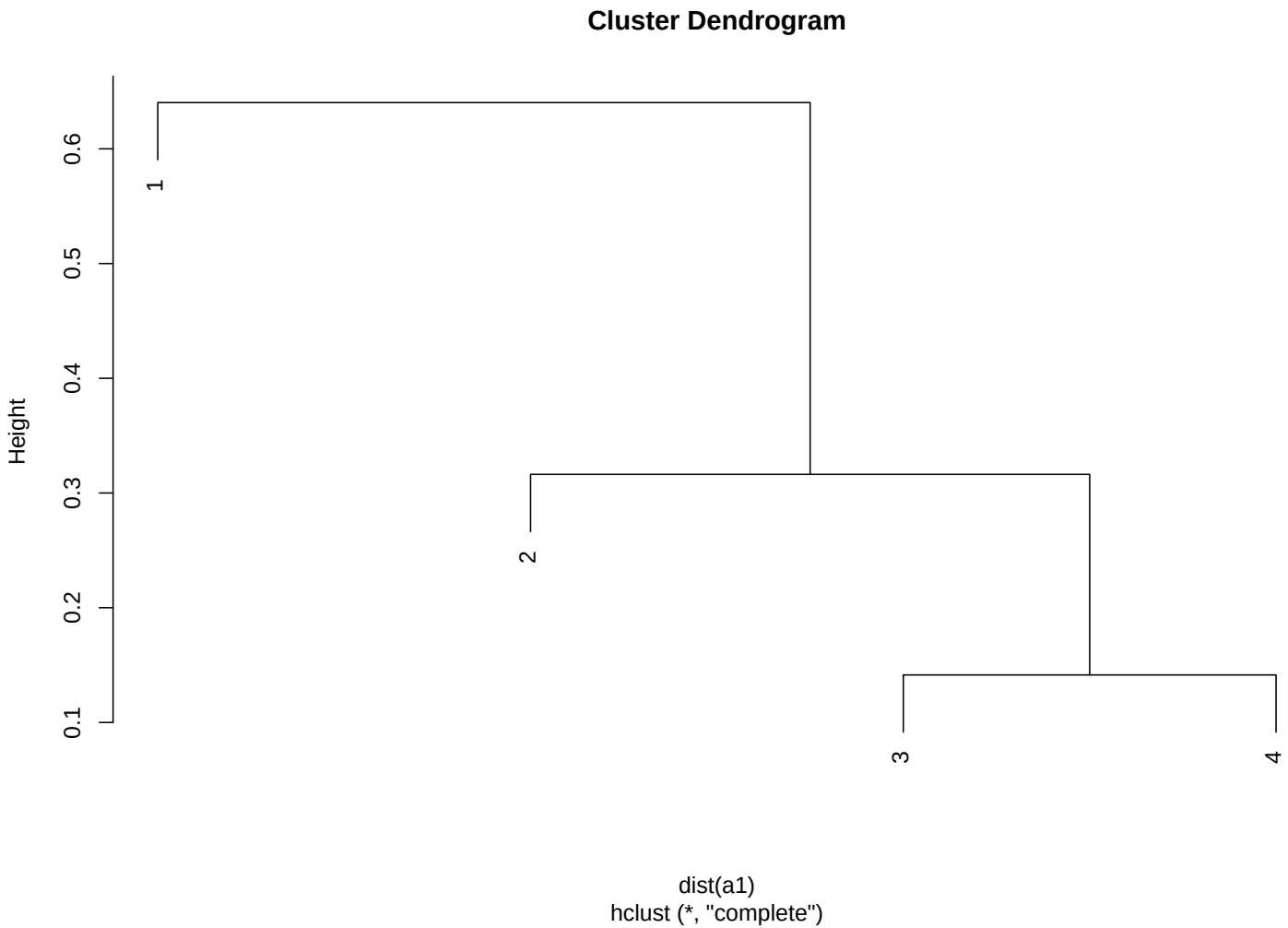


Cómo funciona: iter 3

```
                    1
3-4-2 0.5350234
```



Resultado


Cluster Dendrogram
Height
0.1
0.2
0.3
0.4
0.5
0.6
1
2
3
4
dist(a1)
hclust (*, "complete")