

Análisis de conglomerados: los arqueros persas

Francesc Carmona

27 de agosto de 2021

```
if(!require(qgraph)) install.packages("qgraph")
if(!require(smacof)) install.packages("smacof")
```

Introducción

En este documento vamos a trabajar el análisis de conglomerados con los atributos de 24 antiguas tallas de arqueros en Persépolis. Este es un ejemplo del libro de D.J. Bartholomew et al. (2008) con datos de Roaf (1978).

Hay veinticuatro arqueros esculpidos en bajorrelieve subiendo la escalera sur en el lado oeste de la cara este de la Apadana en Persépolis, en el sur de Irán. Todos los arqueros tienen un aspecto similar, pero difieren en detalles menores, como la forma en que se riza la barba o la manera en que se decora el tocado. La figura 1 es una imagen del noveno arquero (desde lo alto de la escalera) e identifica 21 rasgos o atributos que pueden diferir entre los arqueros. Cada atributo tiene sólo un pequeño número de variantes, normalmente dos o tres. En Roaf (1983) hallaremos detalles al respecto.

La tabla 1 presenta la matriz de datos que muestra qué arquero tiene qué variante de cada atributo. Los atributos están etiquetados de la A a la U y las variantes de un atributo están etiquetadas con números enteros entre 1 y 6. Un cero en la tabla indica que ese atributo falta en el bajorrelieve de ese arquero, debido a los daños sufridos con el paso del tiempo.

Los arqueólogos e historiadores del arte están interesados en saber si los arqueros fueron esculpidos por un solo escultor o por varios y en saber si estos datos sugieren alguna agrupación que pueda influir en esta cuestión. Roaf (1978, 1983) ha realizado varios análisis de agrupación para identificar grupos de arqueros que podrían haber sido tallados por el mismo escultor o equipo de escultores. Utilizamos estos datos para ilustrar el análisis de conglomerados del vecino más lejano, y le invitamos a probar otros métodos.



Figura 1: Arquero número 9 en Persépolis rodeado de las variantes de los 21 atributos.

Archer	Attribute																				
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	2	2	1	2	3	2	1	1	2	1	1	3	0	0	2	3	4	2	2	2	0
2	2	2	1	2	2	2	1	1	2	1	1	3	1	1	2	0	4	2	2	1	2
3	2	2	1	2	2	2	1	1	2	1	2	3	2	2	2	3	0	3	2	1	3
4	2	2	1	2	3	1	1	1	2	1	2	3	2	2	2	3	4	2	2	1	3
5	2	3	1	3	2	2	1	1	2	1	2	0	2	2	4	3	4	3	2	1	3
6	2	3	1	2	2	2	1	1	2	1	2	3	2	2	4	3	4	3	2	1	3
7	2	3	1	3	3	2	1	1	2	2	2	3	2	2	3	3	4	3	2	1	3
8	2	3	1	3	3	2	2	1	2	2	1	2	2	2	3	3	4	3	2	1	3
9	3	1	1	3	2	2	1	2	2	1	1	2	1	1	3	2	2	4	3	2	2
10	3	1	1	3	2	2	1	2	2	2	1	2	2	2	3	2	2	2	2	2	2
11	3	1	1	3	2	2	1	2	2	1	2	2	2	2	0	2	2	4	3	2	2
12	3	1	2	2	2	2	1	2	2	1	2	2	1	1	1	5	2	2	3	1	3
13	3	1	1	3	2	2	1	2	2	2	2	2	1	1	0	0	3	2	2	2	3
14	3	1	1	3	2	2	1	2	2	2	2	2	2	2	2	5	3	2	2	2	3
15	3	1	1	3	2	2	1	2	2	1	2	2	2	2	3	2	2	4	3	2	2
16	3	1	1	3	2	2	1	2	1	1	2	2	1	1	3	2	4	4	3	2	2
17	3	1	2	3	2	3	1	2	2	2	2	2	1	2	2	3	4	2	2	2	2
18	3	1	1	3	2	2	1	2	2	1	2	2	2	2	6	1	1	2	2	2	2
19	2	1	1	3	2	2	1	2	2	1	2	2	2	2	3	1	4	2	2	2	2
20	2	1	2	2	2	2	3	1	2	1	2	2	2	2	1	3	4	2	2	1	3
21	2	1	2	2	2	2	3	1	2	1	2	2	2	2	0	3	4	2	2	2	3
22	2	1	2	2	2	2	3	1	1	1	2	2	2	1	0	3	4	2	2	2	3
23	2	1	2	2	2	2	3	1	2	1	2	2	2	1	1	3	4	2	2	2	3
24	2	1	1	2	2	2	3	1	2	1	2	2	2	1	2	3	4	2	2	2	3

Tabla 1: Matriz de datos que muestra para cada uno de los 24 arqueros qué variante de cada uno de los 21 atributos de A a U posee (un cero indica que ese atributo para ese arquero no existe).

Análisis de conglomerados

En el artículo de Roaf (1978) se muestra un gráfico resumen (ver figura 2) de las asociaciones de cada arquero con los otros según el número de atributos coincidentes.

A simple vista parece que hay tres grupos. Vamos a estudiar el caso con algunas técnicas multivariantes.

Ejercicio 1

Construir una similaridad entre los arqueros con máximo 21 (el número de atributos) parecida a la que ofrece el libro de Bartholomew et al. (2008) en la tabla 2.13, pero a partir de la distancia de Gower entre atributos.

Dibujar con esta similaridad un diagrama de red que muestre las conexiones con valor igual o superior a 15. Comentar el resultado respecto a la figura 2.

Nota: La función `qgraph()` del paquete del mismo nombre nos servirá. Mejor si fijamos como límite inferior 14.999.

Solución:

En primer lugar cargamos los datos de la tabla 1 y procedemos a considerarlos como factores. Un punto importante es fijar como *missings* a los valores cero.

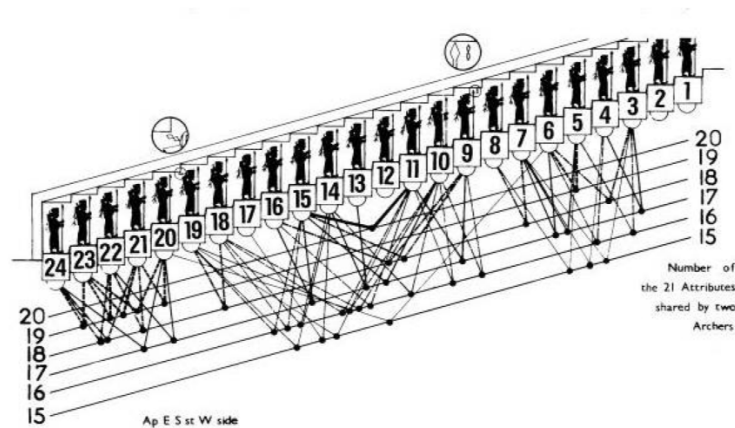


Figura 2: Diagrama de los arqueros en las escaleras con las uniones entre ellos si tienen 15 o más atributos en común.

```
archer_data <- read.csv("archer_data.txt", sep=" ", na.strings="0")
archer_data <- archer_data[,-1]
archer_data <- as.data.frame(lapply(archer_data, factor))
```

La distancia de Gower se puede calcular con la función `daisy()` del paquete `cluster`. Además esta función trata razonablemente los valores faltantes sin eliminar toda las observaciones de la fila.

```
library(cluster)
dd <- daisy(archer_data, metric="gower")
```

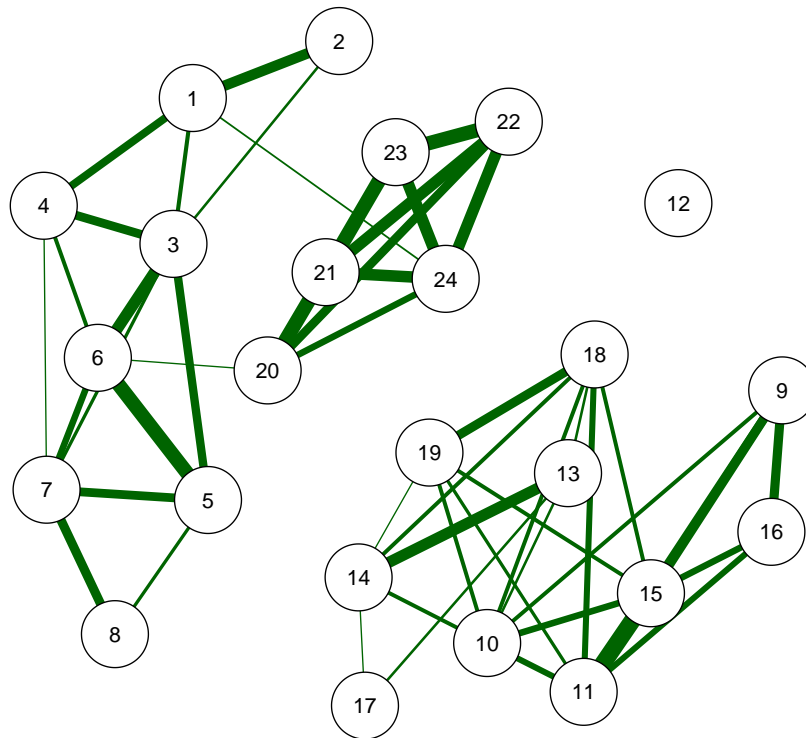
La similitud que proponemos es

```
sdd <- 21 * (1 - dd)
round(as.matrix(sdd),2)[1:6,1:6]
```

```
##      1      2      3      4      5      6
## 1  0.00 18.53 16.06 17.50 12.35 14.00
## 2 18.53  0.00 15.47 14.70 12.16 13.65
## 3 16.06 15.47  0.00 17.85 17.68 18.90
## 4 17.50 14.70 17.85  0.00 14.70 16.00
## 5 12.35 12.16 17.68 14.70  0.00 19.95
## 6 14.00 13.65 18.90 16.00 19.95  0.00
```

Ahora dibujamos el diagrama de red:

```
library(qgraph)
qgraph(sdd, layout="spring", minimum=14.999)
```



Observamos los mismos 3 grupos que en el diagrama de Roaf, aunque hay una débil conexión entre los dos grupos extremos de la escalera. También vemos que el arquero 12 es realmente distinto y que el arquero 17 tiene una débil conexión para el nivel 15 que hemos fijado. También observamos que los arqueros más parecidos son los del grupo final de la escalera.

Ejercicio 2

Comprobar que la distancia de Gower entre estos atributos no es euclídea, de forma que no procede un escalado multidimensional (MDS) clásico.

Representar los arqueros con un MDS no métrico (u ordinal) en dos dimensiones con el método de la regresión isotónica y valorar el resultado. Realizar un *scree plot* con el *stress* del escalado de 1 hasta 5 dimensiones. Valorar el ajuste entre distancias en dimensión 2 con un diagrama de Shepard.

Utilizar también el algoritmo SMACOF (*Scaling by MAyorizing a COmplicated Function*) para ver si hay alguna diferencia con el proceso iterativo habitual de Kruskal-Shepard basado en el algoritmo ALSCAL (*Alternating Least Squares Scaling*).

Nota: Para evitar problemas con los algoritmos, podemos substituir las distancias cero entre arqueros diferentes por un valor bajo como 0.001.

Solución:

En primer lugar comprobamos que la distancia propuesta no es euclídea.

```
library(ade4)
is.euclid(dd)
```

```
## Warning in is.euclid(dd): Zero distance(s)
```

```
## [1] FALSE
```

Además vemos que hay alguna distancia con valor 0 lo que nos puede traer problemas. Sustituimos ese valor 0 por otro muy bajo pero no nulo.

```
which(dd == 0)
```

```
## [1] 189
```

```
dd[189] <- 0.001
```

Vamos a realizar un escalado multidimensional no métrico con la versión del método de Kruskal y Shepard que implementa la función `isoMDS()` basado en la regresión isotónica.

```
library(MASS)
iso.mds <- isoMDS(dd)
```

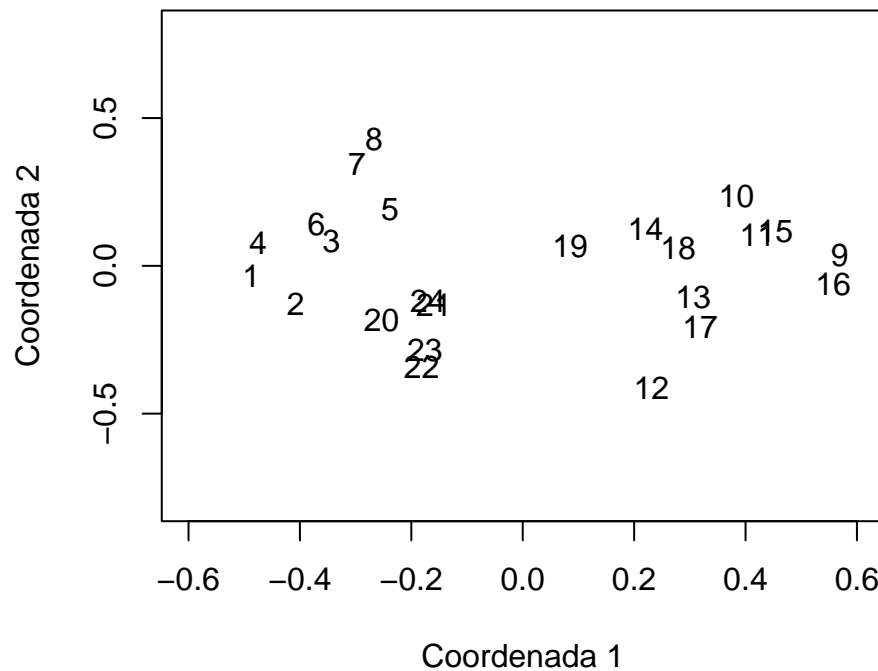
```
## initial value 16.419259
```

```
## iter 5 value 13.562414
```

```
## final value 13.376280
```

```
## converged
```

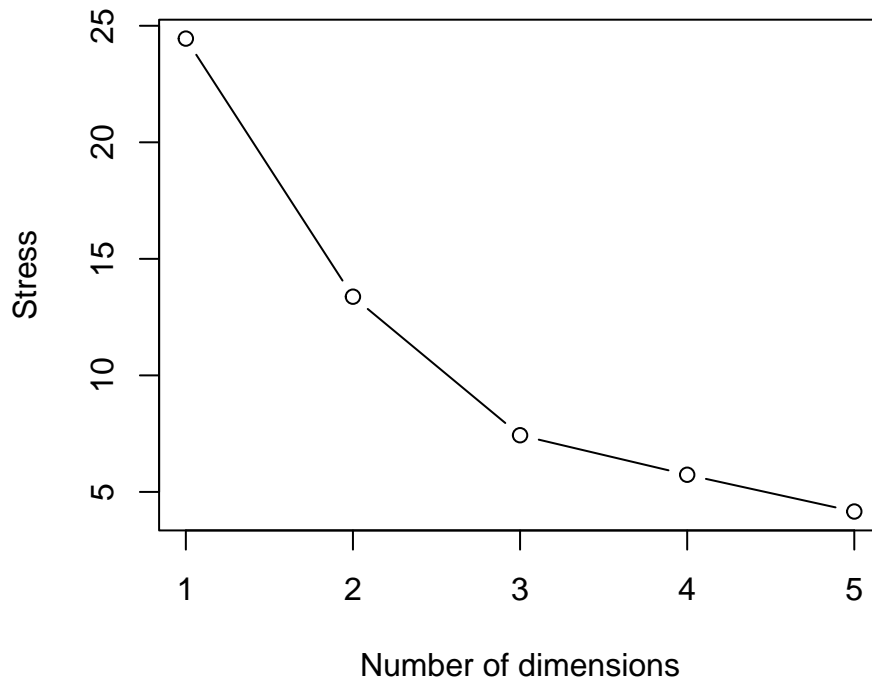
```
plot(iso.mds$points, type = "n", xlim=c(-0.6,0.6), ylim=c(-0.8,0.8),
      xlab="Coordenada 1", ylab="Coordenada 2")
text(iso.mds$points, labels=as.character(1:24))
```



El gráfico muestra dos o tres grupos al estilo de lo que hemos visto en el diagrama de red del ejercicio anterior. Las agrupaciones son las mismas. Es posible que hubiera tres grupos de escultores, uno trabajando en la parte alta (1 a 8), otro en el centro (9 a 19) y un tercero en la parte baja (20 a 24). Estos últimos son muy similares de forma que podría ser la obra de un único escultor. El arquero 12 aparece en el grupo central, aunque ligeramente separado en la representación imperfecta en dimensión 2.

Para valorar si la reducción a dimensión 2 es apropiada se puede realizar un *scree plot*.

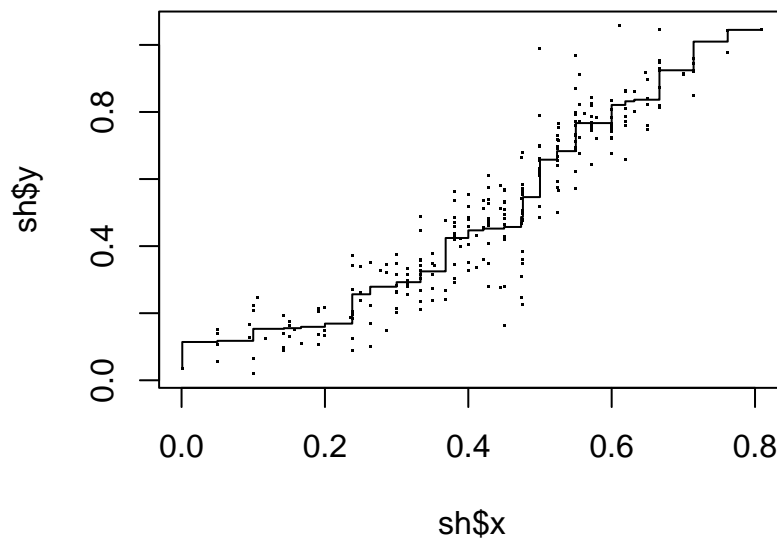
```
scree.plot = function(d, k) {  
  stresses=isoMDS(d, k=k, trace = FALSE)$stress  
  for(i in rev(seq(k-1)))  
    stresses=append(stresses,isoMDS(d, k=i, trace = FALSE)$stress)  
  plot(seq(k), rev(stresses), type="b", xaxp=c(1,k, k-1),  
        ylab="Stress", xlab="Number of dimensions")  
}  
scree.plot(dd, k=5)
```



Parece que *el codo* está sobre la dimensión 3 o 4.

Para valorar el ajuste entre la distancia real y su aproximación en dimensión 2 dibujamos un diagrama de Shepard.

```
sh <- Shepard(dd, iso.mds$points, p=2)  
plot(sh, pch=".")  
lines(sh$x, sh$yf, type = "S")
```

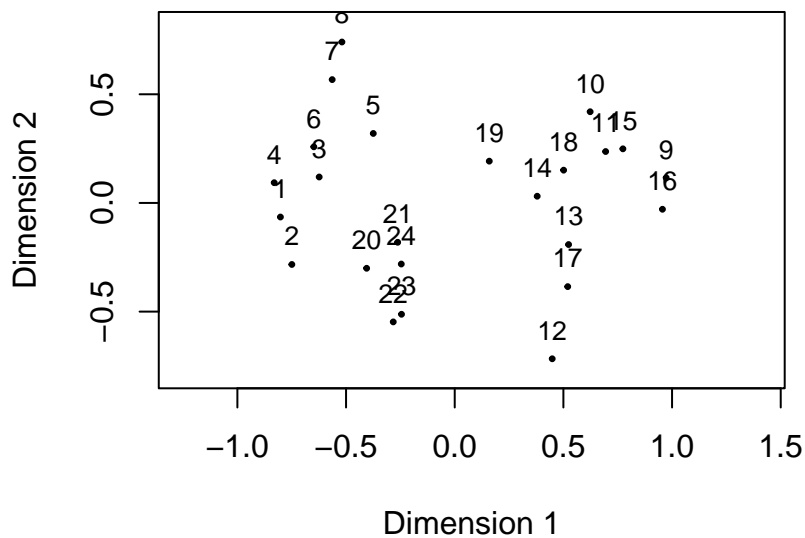


De acuerdo con lo que hemos visto en el *scree plot*, la representación en dimensión 2 no es muy buena.

Veamos finalmente si la representación con el algoritmo SMACOF es distinta. Para calcular dicho algoritmo utilizaremos la función `smacofSym()` del paquete `smacof`.

```
library(smacof)
res.smacof <- smacofSym(dd, type="ordinal", verbose=FALSE)
plot(res.smacof)
```

Configuration Plot



Podemos observar pequeñas diferencias respecto a la configuración obtenida antes, pero las conclusiones generales son las mismas.

Ejercicio 3

Tanto el diagrama de red como el análisis de proximidades que hemos hecho nos han dado una imagen de las posibles agrupaciones. Sin embargo, se trata de apreciaciones visuales. Ha llegado la hora de realizar un análisis de conglomerados.

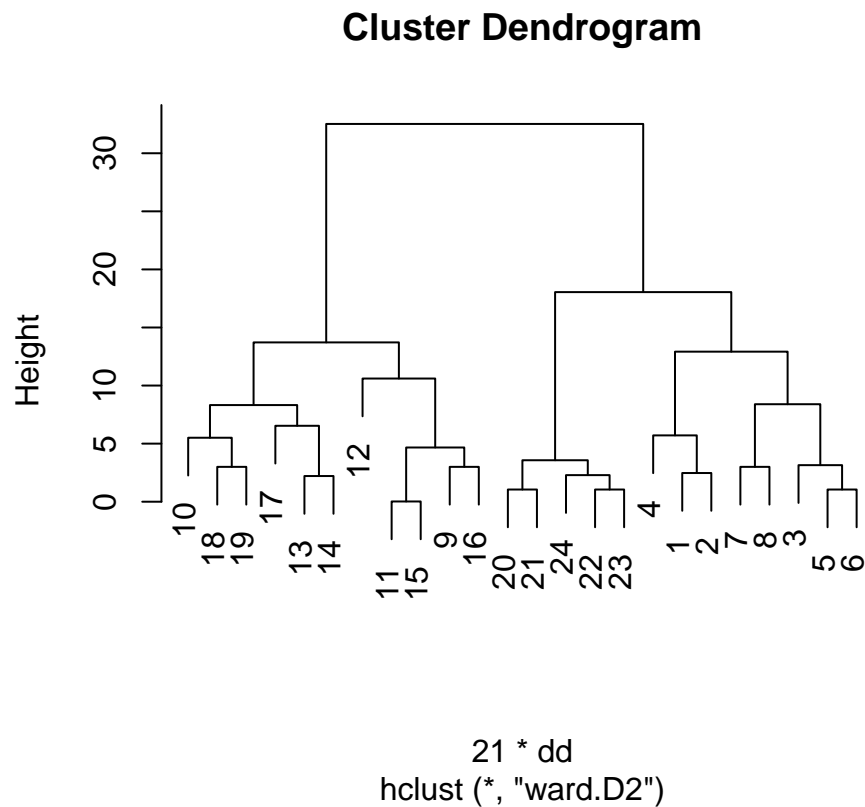
Realizar un análisis de conglomerados jerárquico y valorar el resultado.

Realizar un análisis de conglomerados con el método PAM y decidir el mejor número de grupos.

Nota: Tomaremos como distancia $21 * dd$ para comparar los resultados con el libro de Bartholomew et al. (2008).

Solución:

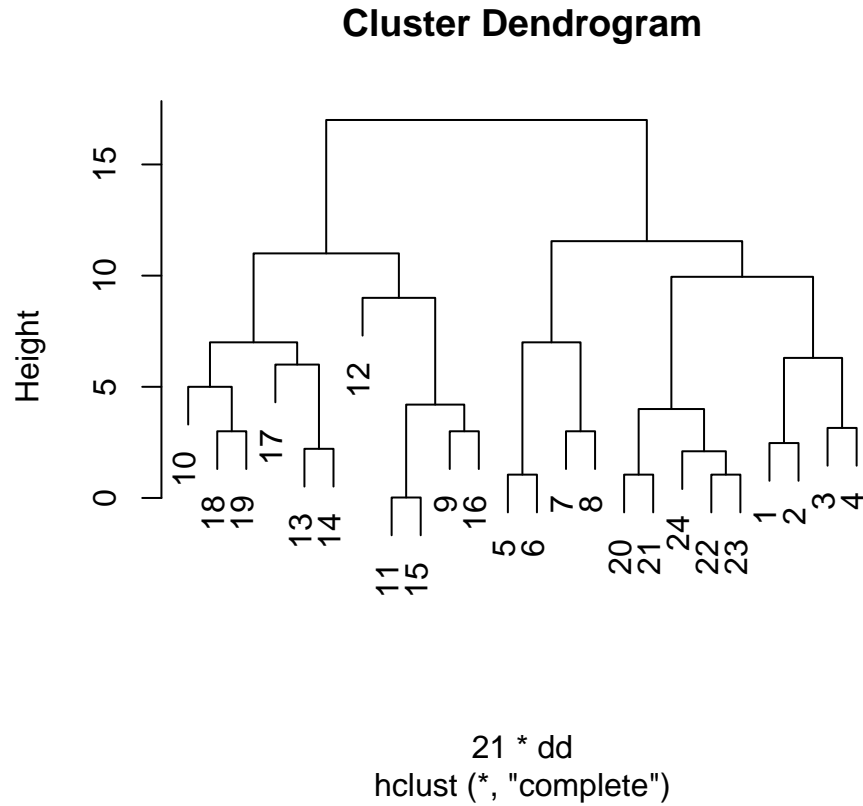
```
arqueros.hc <- hclust(21*dd, method="ward.D2")  
plot(arqueros.hc)
```



Vemos la formación de dos o tres grupos como ya sabemos de los ejercicios anteriores.

En la figura 2.13 del libro de Bartholomew et al. (2008) se ve el dendrograma con el método *complete linkage* (y una distancia ligeramente distinta).

```
plot(hclust(21*dd, method="complete"))
```

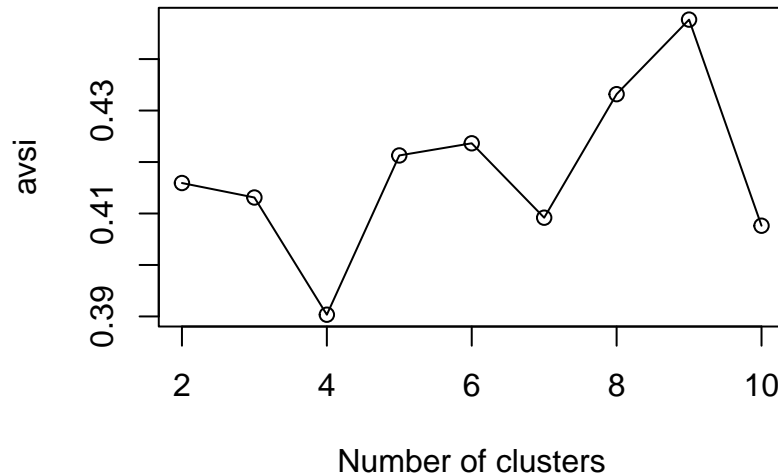


Este dendrograma es casi idéntico al de la Figura 2.13 del libro y varía la forma de considerar 3 grupos.

Nos queda pues la duda de si formar dos o tres grupos y también la composición de esos grupos. Lo mejor será utilizar un método más robusto como el PAM y las siluetas para evaluar la calidad de los grupos.

```
avsi <- c(0)
for(i in 2:10){
  si <- silhouette(pam(21*dd, k=i, diss=TRUE))
  avsi[i-1] <- mean(si[, "sil_width"])
}
plot(2:10, avsi, type="o", xlab="Number of clusters")
title(main="Average silhouette")
```

Average silhouette

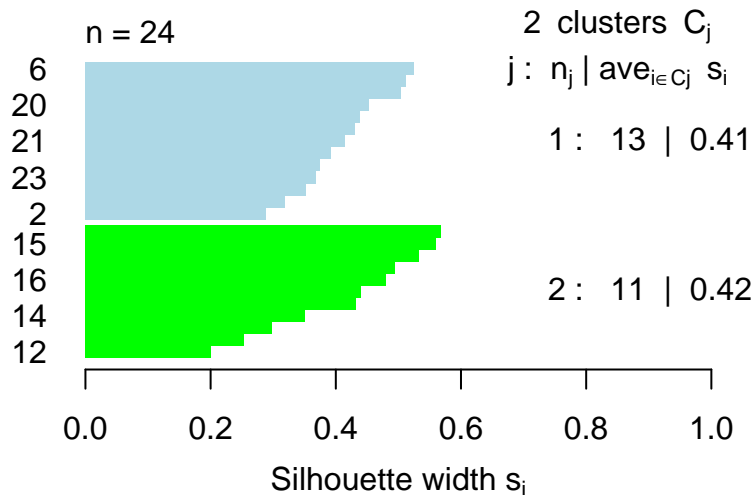


Si se trata de formar grupos a partir de la calidad de las siluetas, parece que el mejor número es 9. Demasiado alto.

Entre los números más bajos, el 2 gana al 3, aunque no al 5 y al 6. En resumen, si hay que decidir entre dos grupos o tres, la silueta nos dice 2.

```
pam2 <- pam(21*dd, 2, diss=TRUE)
gcol <- c("lightblue","green")
plot(silhouette(pam2), col=gcol)
```

Silhouette plot of pam(x = 21 * dd, k = 2,

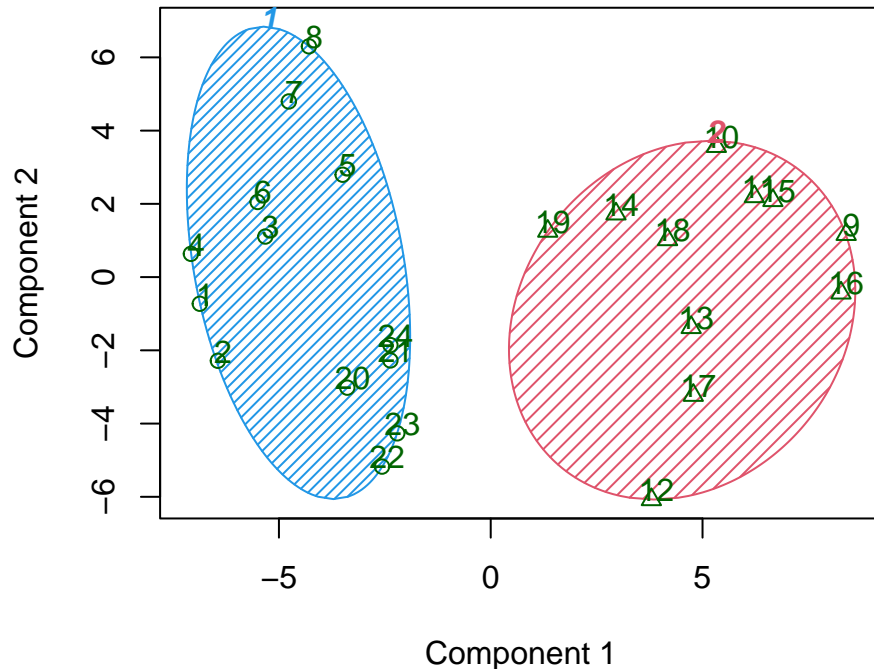


Average silhouette width : 0.42

Finalmente una representación en dos dimensiones de los dos grupos de arqueros.

```
iso.mds <- isoMDS(21 * dd, trace = FALSE)
clusplot(iso.mds$points, pam2$clustering, color=TRUE, shade=TRUE, labels=2, lines=0)
```

CLUSPLOT(iso.mds\$points)



These two components explain 100 % of the point variab

Referencias

David J. Bartholomew, Fiona Steele, Irini Moustaki and Jane I. Galbraith, *Analysis of Multivariate Social Science Data*, Second Edition, Taylor and Francis, 2008.

Micheal Roaf (1983), Sculptures and Sculptors at Persepolis, *Iran*, 21, I-164. doi:10.2307/4299731

Emelina López-González y Ramón Hidalgo Sánchez, Escalamiento Multidimensional No Métrico. Un ejemplo con R empleando el algoritmo SMACOF. *Estudios sobre educación*, vol. 18, 9-35, 2010.