# Índice

# Capítulo 1

# Arqueología y Paleontología

## 1.1. Introducción

Una de las disciplinas que más debería utilizar técnicas estadísticas, especialmente avanzadas, es la Arqueología y la Paleontología. No obstante, mi experiencia al respecto en los trabajos conjuntos que he publicado con grandes profesionales de estas áreas, es lo poco habitual que resuelta su uso, cosa que no pasa en otras áreas como las Ciencias de la Salud, por ejemplo. En este capítulo tratamos de aportar algunos problemas resueltos que pueden ayudar al lector a cambiar esa situación.

En este capítulo se resuelven muchos problemas avanzados de Estadística, por lo que en casi todos ellos se hará referencia al capítulo en el que se estudió la técnica correspondiente en el texto denominado EAA, como dijimos en el Prólogo.

## 1.2. Problemas Avanzados Resueltos

**Problema 1.1**

**Se dispone de información (Bolviken at al., 1982) procedente de la excavación de varios yacimientos mesolíticos cerca de Iversfjord en Finnmark, en el Ártico Noruego. Se trata de 37 tipos de objetos líticos procedentes de 14 cabañas costeras de entre el 3000 y el 600 antes de Cristo, que se cree eran de pesca invernales (de hecho el propósito del estudio era analizar su uso). Tras un análisis preliminar se decidió agrupar los 37 tipos de objetos líticos en sólo 9 categorías según su función: Puntas, Raspadores/Buriles, Útiles sobre Núcleo, Cuchillos, Pesos de Red, Manufactura de Herramientas, Fragmentos de Pizarra, Láminas Manipuladas y Piedras Perforadas. La tabla de frecuencias absolutas es la siguiente:**

| | Tipos de objetos líticos | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Punt** | **Rasp** | **Util** | **Cuchi** | **Pes** | **Manu** | **Frag** | **Lami** | **Pied** |
| **Cabaña** | | | | | | | | | |
| **1** | **19** | **22** | **1** | **17** | **19** | **56** | **10** | **16** | **1** |
| **2** | **8** | **7** | **1** | **3** | **5** | **26** | **8** | **5** | **1** |
| **3** | **5** | **0** | **0** | **0** | **3** | **20** | **6** | **2** | **0** |
| **4** | **15** | **136** | **4** | **2** | **0** | **70** | **4** | **53** | **0** |
| **5** | **8** | **53** | **2** | **7** | **5** | **47** | **11** | **19** | **0** |
| **6** | **36** | **99** | **2** | **7** | **2** | **62** | **3** | **49** | **0** |
| **7** | **3** | **47** | **0** | **0** | **0** | **15** | **3** | **24** | **0** |
| **8** | **35** | **59** | **0** | **4** | **2** | **73** | **11** | **24** | **0** |
| **9** | **5** | **5** | **0** | **0** | **0** | **6** | **8** | **2** | **0** |
| **10** | **9** | **2** | **0** | **0** | **0** | **7** | **5** | **0** | **0** |
| **11** | **9** | **1** | **2** | **0** | **1** | **3** | **1** | **1** | **0** |
| **12** | **1** | **6** | **0** | **0** | **0** | **1** | **0** | **0** | **0** |
| **13** | **0** | **6** | **1** | **0** | **0** | **1** | **0** | **2** | **0** |
| **14** | **1** | **9** | **0** | **0** | **0** | **16** | **0** | **1** | **0** |

**El propósito del estudio es analizar, mediante un Análisis de Correspondencias, qué cabañas pueden ser consideradas como similares, desde el punto de vista de los útiles empleados, qué útiles eran similares y, finalmente, cómo se relacionaban estas dos variables.**

---

Observamos que, al ser el Análisis de Correspondencias (EAA-capítulo 3) una técnica de tipo descriptivo, no necesitamos una distribución normal multivariante para poder ser aplicada, eso sí, a cambio de que las conclusiones finales no se basarán en p-valores.

Entrando ya en la resolución del ejercicio con R, primero vamos a incorporar los datos, denominados `X`, a R con las siguientes sentencias:

```
> x1<-c(19,8,5,15,8,36,3,35,5,9,9,1,0,1)
> x2<-c(22,7,0,136,53,99,47,59,5,2,1,6,6,9)
> x3<-c(1,1,0,4,2,2,0,0,0,0,2,0,1,0)
> x4<-c(17,3,0,2,7,7,0,4,0,0,0,0,0,0)
> x5<-c(19,5,3,0,5,2,0,2,0,0,1,0,0,0)
> x6<-c(56,26,20,70,47,62,15,73,6,7,3,1,1,16)
> x7<-c(10,8,6,4,11,3,3,11,8,5,1,0,0,0)
> x8<-c(16,5,2,53,19,49,24,24,2,0,1,0,2,1)
> x9<-c(1,1,0,0,0,0,0,0,0,0,0,0,0,0)
> X<-matrix(c(x1,x2,x3,x4,x5,x6,x7,x8,x9),ncol=9)
> X
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
 [1,]   19   22    1   17   19   56   10   16    1
 [2,]    8    7    1    3    5   26    8    5    1
 [3,]    5    0    0    0    3   20    6    2    0
 [4,]   15  136    4    2    0   70    4   53    0
 [5,]    8   53    2    7    5   47   11   19    0
 [6,]   36   99    2    7    2   62    3   49    0
 [7,]    3   47    0    0    0   15    3   24    0
 [8,]   35   59    0    4    2   73   11   24    0
```

```
 [9,]    5    5    0    0    0    6    8    2    0
[10,]    9    2    0    0    0    7    5    0    0
[11,]    9    1    2    0    1    3    1    1    0
[12,]    1    6    0    0    0    1    0    0    0
[13,]    0    6    1    0    0    1    0    2    0
[14,]    1    9    0    0    0   16    0    1    0
```

Un sencillo test de la $\chi^2$-cuadrado, obtenido ejecutando

```
> chisq.test(X)

        Pearson's Chi-square test

data:  X
X-squared = 481.8703, df = 104, p-value = < 2.2e-16
```

permite comprobar que no son independientes las Cabañas de los Tipos de utensilios ya que el p-valor del test, casi cero, permite rechazar con gran seguridad la hipótesis nula de independencia entre ambas variables.

Para ejecutar ahora el Análisis de Correspondencias con R (EAA-sección 3.2.1) primero construimos en `(1)` el data frame necesario, asignándole nombres a los "valores" de las dos variables para identificarlos en el gráfico de correspondencias final; luego abrimos la librería que permite ejecutar el Análisis de Correspondencia en `(2)`, ejecutando dicho análisis con `(3)`.

```
> X<-data.frame(x1,x2,x3,x4,x5,x6,x7,x8,x9)                                    (1)
> rownames(X)<-c("1","2","3","4","5","6","7","8","9","10","11","12","13","14")
> colnames(X)<-c("Puntas","Raspadores","Útiles sobre Núcleo","Cuchillos",
+ "Pesos de Red","Manufactura","Pizarra","Láminas","Piedras")
> library(ca)                                                                  (2)
> solu<- ca(X)                                                                 (3)
> solu

 Principal inertias (eigenvalues):
           1       2        3        4        5        6        7        8
Value      0.18713 0.069649 0.041111 0.023965 0.017257 0.006479 0.004229 0.002168
Percentage 53.16%  19.79%   11.68%   6.81%    4.9%     1.84%    1.2%     0.62%
                     (4)
```

En `(4)` vemos que sumando las Inercias de las dos primeras dimensiones, $53'16\,\% + 19'79\,\% = 72{,}95\,\%$ acumulan un $72'195\,\%$ suficientemente grande como para concluir que un gráfico con las dos primeras dimensiones explica bastante bien el resultado final.

El gráfico con el que obtendremos las conclusiones finales lo conseguimos ejecutando `(5)`.

```
> plot(solu)                                                                   (5)
```

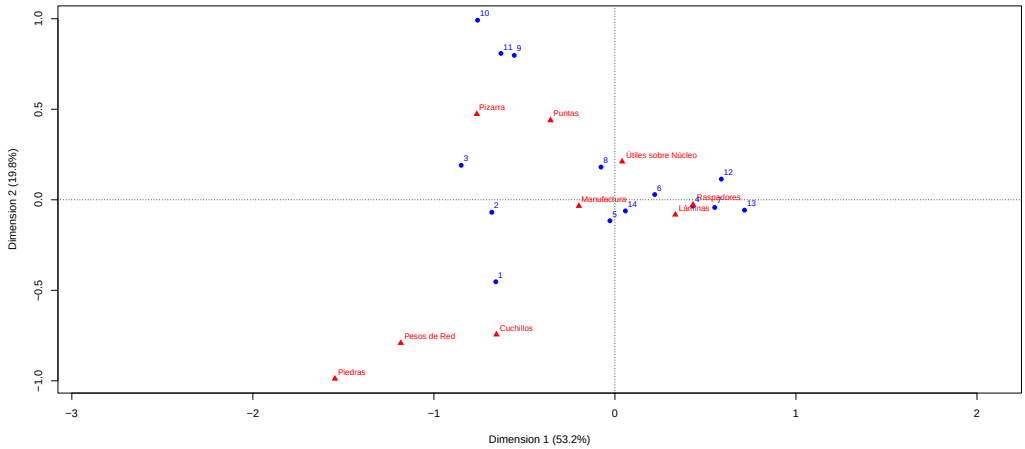



Figura 1.1 : Gráfico de Correspondencias

Dicho gráfico de correspondencias es la la Figura 1.1. De este gráfico podemos concluir que los útiles que aparecen en su lado derecho (Raspadores/buriles, Láminas manipuladas, Útiles sobre núcleo y Manufactura de herramientas) son útiles de mantenimiento, que se efectuaba en las Cabañas 8, 5, 14, 6, 12, 4, 7 y 13, mientras que los útiles de la izquierda del gráfico, por un lado Piedras, Cuchillos, Pesos de red, son de pesca, asociada a las Cabañas 1 y 2; y que los Fragmentos de pizarra y Puntas son útiles de caza, asociada a las Cabañas 9, 10, 11 y 3.

Se observa, por tanto, que el primer eje (el horizontal) indica diferencia entre caza y pesca y el segundo eje (el vertical), tipo de actividad (mantenimiento y caza/pesca).

La cuestión que también se investigaba en el estudio de si las cabañas podían ser consideradas sólo como cabañas de pesca invernales se desecho puesto que, claramente, aparecen tres grupos de cabañas con diferentes propósitos: mantenimiento, caza y pesca.

El mencionado estudio concluye que los datos muestran una fuerte evidencia de diversidad económica (mantenimiento, caza y pesca) y no sólo pesca. También muestra el estudio diversos grados de permanencia del asentamiento (tanto específicas a corto plazo, la caza y la pesca, como actividades a largo plazo, mantenimiento), cuestión que no se creía pudiera ocurrir en yacimientos prehistóricos costeros.

## Problema 1.2

**Los datos de la siguiente tabla (Mathiesen at al., 1981)**

| | Tipos de restos | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** | **I** | **J** | **K** | **L** | **M** | **N** | **P** | **Q** |
| **Es.** | | | | | | | | | | | | | | | | |
| **1** | **27** | **22** | **33** | **1** | **3** | **0** | **272** | **5** | **40** | **31** | **3** | **0** | **0** | **0** | **17** | **0** |
| **2** | **54** | **122** | **35** | **0** | **4** | **0** | **1080** | **36** | **15** | **73** | **11** | **1** | **0** | **0** | **47** | **0** |
| **3** | **44** | **83** | **54** | **3** | **4** | **0** | **842** | **24** | **71** | **81** | **35** | **12** | **0** | **0** | **32** | **0** |
| **4** | **101** | **151** | **90** | **2** | **6** | **0** | **3247** | **14** | **128** | **81** | **20** | **23** | **3** | **0** | **34** | **1** |
| **5** | **101** | **202** | **58** | **4** | **0** | **4** | **3204** | **95** | **99** | **216** | **24** | **92** | **33** | **0** | **22** | **4** |
| **6** | **43** | **61** | **33** | **6** | **13** | **1** | **1082** | **37** | **170** | **138** | **17** | **1** | **5** | **0** | **4** | **4** |
| **7** | **24** | **40** | **17** | **0** | **23** | **4** | **545** | **23** | **3** | **88** | **3** | **0** | **1** | **3** | **2** | **1** |
| **8** | **17** | **24** | **14** | **1** | **30** | **3** | **597** | **22** | **4** | **46** | **4** | **1** | **0** | **0** | **2** | **0** |
| **9** | **27** | **42** | **10** | **2** | **14** | **2** | **294** | **8** | **7** | **33** | **2** | **11** | **0** | **0** | **9** | **0** |
| **10** | **24** | **53** | **20** | **0** | **6** | **0** | **100** | **6** | **3** | **22** | **0** | **28** | **0** | **0** | **14** | **1** |
| **11** | **45** | **78** | **35** | **0** | **30** | **1** | **128** | **4** | **0** | **20** | **0** | **17** | **1** | **0** | **2** | **9** |
| **12** | **109** | **367** | **167** | **1** | **142** | **8** | **348** | **1** | **13** | **38** | **1** | **80** | **1** | **0** | **13** | **6** |
| **13** | **15** | **18** | **17** | **0** | **7** | **8** | **25** | **0** | **0** | **4** | **0** | **14** | **0** | **0** | **0** | **1** |
| **14** | **42** | **41** | **34** | **0** | **15** | **3** | **93** | **0** | **0** | **0** | **0** | **14** | **0** | **0** | **6** | **0** |

**también en el fichero ‘‘Helgoy.txt’’, corresponden a las frecuencias absolutas de restos osteológicos de animales procedentes de excavaciones de asentamientos en el norte de Noruega (corte 1 del túmulo de la isla de Helgoy) efectuadas en 14 estratos (el estrato 1 era el más moderno y el 14 el más antiguo). Los restos que aparecieron era de A = Bóvidos (Bos taurus), B = Ovicápridos (Ovis aries/Carpa hircus), C = Porcino (Sus scrofa dom.), D = Reno (Rangifer tarandus), E = Foca (Fócidos), F = Lagópodo (Lagopus), G = Abadejo (Gadus morrhua), H = Arenque (Melanogrammus aeglefinus), I = Merluza (Pollachius virens), J = Gádidos (Molva molva), K = (Brosme brosme), L = Halibut (Hippoglossus hippoglossus), M = (Sebastes marinus), N = (Anarchichas lupus), P = Pingüino (Álcidos) y Q = Gallo (Gallus gallus f. dom.)**
**Se pretende analizar si un determinado tipo de animal estuvo más presente en una época que en otra, es decir, en un estrato más que en otro.**

Para analizar esta dependencia, ejecutaremos un Análisis de Correspondencias. Podemos ir incorporando los datos por columnas, como en el ejemplo anterior,

```
> x1<-c(19,8,5,15,8,36,3,35,5,9,9,1,0,1)
.....................................
```

para luego formar el data frame que es el tipo de dato que entenderá R. Alternativamente, los incorporaremos directamente ejecutando `(1)`, dando nombre a las filas con `(2)`.

```
> Helgoy<-read.table("e:\\Helgoy.txt",header=T)                          (1)
> rownames(Helgoy)<-c("1","2","3","4","5","6","7","8","9","10","11","12",  (2)
+ "13","14")                                                              (2)
```

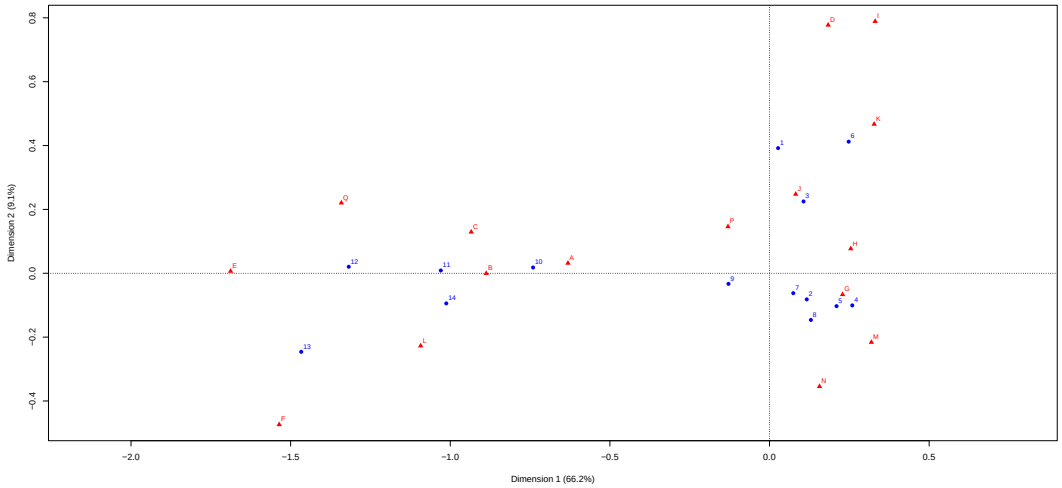



Figura 1.2 : Gráfico de Correspondencias

Para realizar el Análisis de Correspondencias con R (EAA-sección 3.2.1), ejecutamos las siguientes instrucciones

```
> library(ca)
> solu<-ca(Helgoy)
> solu

 Principal inertias (eigenvalues):
           1        2        3        4        5        6        7
Value      0.226063 0.030938 0.020252 0.01883  0.015589 0.01174  0.006824
Percentage 66.25%   9.07%    5.93%    5.52%    4.57%    3.44%    2%
           8        9        10       11        12       13
Value      0.003592 0.002416 0.002307 0.001584 0.000822 0.000279
Percentage 1.05%    0.71%    0.68%    0.46%    0.24%    0.08%
```

Con un gráfico en dos dimensiones, vemos en la tabla anterior, que recogemos una Inercia del $66'25\,\% + 9'07\,\% = 75'32\,\%$ suficientemente grande como para quedar satisfechos.

El Gráfico de Correspondencias se obtiene ejecutando

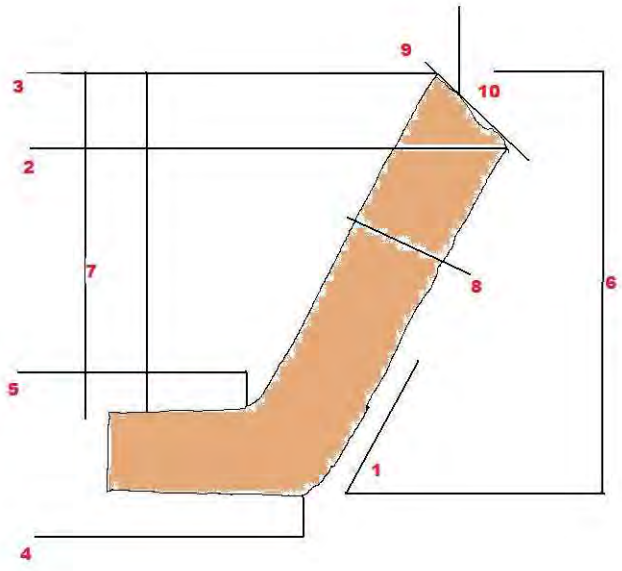
```
> plot(solu)
```

con lo que obtendríamos la Figura 1.2, que nos indica cuál era la mayor presencia de los diferentes tipos de animales en los diferentes estratos. Así, se observa que los animales A, B, C, L, Q, E y F aparecían más en los estratos más antiguos (10, 11, 12, 13 y 14) y el resto de los animales en los estratos más modernos; pero el propio gráfico nos indica cuáles de los animales están

más relacionados con cada estrato. Por ejemplo, el animal K en el estrato 6, el animal J en el estrato 3; o que los animales I y D no están relacionados especialmente con ningún estrato, es decir, que aparecen de una forma similar en todos los estratos.

## Problema 1.3

**Los datos de la tabla del enunciado corresponden a medidas de 41 cuencos de borde oblicuo del período Uruk en Mesopotamia (Johnson, 1973) como los del siguiente dibujo.**




**El dibujo representa la mitad de un cuenco y en él aparecen los números de las variables de la tabla. Éstas son: 1 = Ángulo de la base, 2 = Diámetro del borde (estimado en unidades de $0'5$ cm.), 3 = Diámetro interior del borde (en unidades de $0'5$ cm.), 4 = Diámetro de la base (en $0'5$ cm.), 5 = Diámetro interior de la base (en $0'5$ cm.), 6 = Altura de la pared (medida en $0'1$ cm.), 7 = Altura interior de la pared (en $0'1$ cm.), 8 = Grosor de la pared, 9 = Grosor del borde, 10 = Ángulo del borde.**

| | Medidas | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| **Cuenco** | | | | | | | | | | |
| **1** | **58** | **160** | **150** | **80** | **70** | **73** | **65** | **108** | **145** | **128** |
| **2** | **57** | **140** | **130** | **70** | **65** | **67** | **62** | **94** | **111** | **137** |
| **3** | **55** | **175** | **155** | **70** | **70** | **71** | **61** | **107** | **110** | **137** |
| **4** | **58** | **180** | **170** | **70** | **65** | **84** | **80** | **106** | **121** | **154** |
| **5** | **62** | **195** | **180** | **80** | **70** | **86** | **72** | **108** | **135** | **150** |
| **6** | **60** | **165** | **160** | **70** | **65** | **85** | **78** | **111** | **130** | **159** |
| **7** | **53** | **180** | **170** | **80** | **65** | **85** | **75** | **120** | **123** | **148** |
| **8** | **68** | **130** | **120** | **60** | **50** | **71** | **65** | **108** | **104** | **150** |
| **9** | **48** | **150** | **140** | **70** | **60** | **70** | **55** | **133** | **129** | **165** |
| **10** | **58** | **200** | **190** | **80** | **75** | **96** | **84** | **159** | **141** | **147** |
| **11** | **47** | **210** | **200** | **85** | **75** | **79** | **74** | **114** | **135** | **163** |
| **12** | **60** | **160** | **150** | **80** | **70** | **87** | **80** | **110** | **121** | **136** |
| **13** | **55** | **180** | **170** | **80** | **80** | **88** | **83** | **109** | **118** | **160** |
| **14** | **65** | **190** | **165** | **80** | **75** | **91** | **79** | **132** | **169** | **150** |
| **15** | **63** | **190** | **170** | **75** | **70** | **89** | **85** | **137** | **129** | **155** |
| **16** | **67** | **220** | **210** | **80** | **75** | **118** | **105** | **145** | **138** | **170** |
| **17** | **44** | **170** | **150** | **80** | **70** | **58** | **44** | **103** | **123** | **154** |
| **18** | **63** | **185** | **170** | **75** | **80** | **80** | **74** | **117** | **139** | **148** |
| **19** | **52** | **160** | **150** | **60** | **55** | **75** | **69** | **109** | **126** | **148** |
| **20** | **62** | **215** | **200** | **90** | **85** | **97** | **81** | **138** | **128** | **133** |
| **21** | **41** | **175** | **160** | **65** | **60** | **70** | **62** | **110** | **137** | **151** |
| **22** | **47** | **190** | **170** | **75** | **80** | **69** | **58** | **120** | **129** | **148** |
| **23** | **50** | **185** | **160** | **70** | **65** | **94** | **80** | **126** | **143** | **152** |
| **24** | **55** | **195** | **180** | **70** | **65** | **85** | **80** | **130** | **129** | **151** |
| **25** | **49** | **195** | **180** | **70** | **65** | **77** | **69** | **124** | **102** | **148** |
| **26** | **58** | **140** | **120** | **65** | **60** | **66** | **54** | **113** | **143** | **130** |
| **27** | **62** | **170** | **160** | **65** | **60** | **90** | **70** | **94** | **131** | **137** |
| **28** | **55** | **135** | **120** | **70** | **65** | **73** | **64** | **109** | **102** | **136** |
| **29** | **53** | **170** | **160** | **70** | **65** | **78** | **64** | **123** | **124** | **135** |
| **30** | **60** | **175** | **160** | **70** | **60** | **83** | **70** | **112** | **142** | **155** |
| **31** | **52** | **140** | **120** | **70** | **65** | **73** | **62** | **116** | **126** | **145** |
| **32** | **59** | **150** | **140** | **75** | **70** | **88** | **76** | **101** | **126** | **135** |
| **33** | **61** | **140** | **130** | **70** | **60** | **92** | **85** | **116** | **103** | **152** |
| **34** | **56** | **145** | **130** | **65** | **60** | **72** | **65** | **125** | **134** | **136** |
| **35** | **60** | **175** | **160** | **75** | **65** | **93** | **78** | **111** | **160** | **130** |
| **36** | **53** | **165** | **160** | **70** | **60** | **74** | **65** | **111** | **62** | **160** |
| **37** | **49** | **165** | **150** | **80** | **75** | **75** | **62** | **129** | **147** | **154** |
| **38** | **60** | **160** | **140** | **70** | **65** | **78** | **66** | **114** | **146** | **143** |
| **39** | **59** | **170** | **160** | **70** | **60** | **91** | **77** | **138** | **119** | **146** |
| **40** | **57** | **165** | **160** | **80** | **63** | **77** | **60** | **91** | **124** | **170** |
| **41** | **55** | **170** | **160** | **80** | **65** | **70** | **66** | **140** | **121** | **149** |

**Clasificar estos cuencos mediante un Análisis Cluster.**

---

Primero incorporaremos los datos. Lo mejor es incorporarlos como data frame. Luego, a partir de ahí ya podemos cambiarles el formato si es necesario. Para ello ejecutamos

```
> Cuencos<-read.table("e:\\Cuencos.txt",header=T)
```

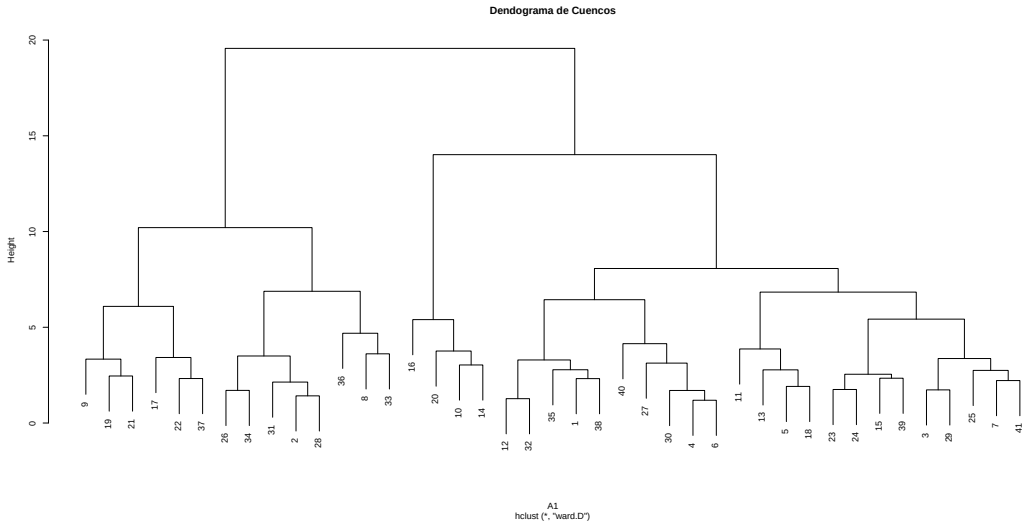



Figura 1.3 : Dendograma con la distancia Euclídea

Realizaremos un Análisis Cluster Jerárquico Aglomerativo (EAA-sección 5.3.1). Según EAA-sección 5.2.2, al ser el tipo de datos cuantitativos en una escala lineal, es decir, una longitud doble de larga aparecería multiplicada por 2, lo razonable es utilizar una distancia para la matriz de datos, Euclídea o Manhattan. Calculemos las dos utilizando R aunque, como decimos en EAA-sección 5.3, para no distorsionar el análisis porque los datos grandes tiendan a dominar el análisis, estandarizamos las observaciones ejecutano,

```
> NCuencos<- scale(Cuencos,center=T,scale=T)
```

```
> A1<-dist(NCuencos,method="euclidean")
> A2<-dist(NCuencos,method="manhattan")
```

De los diferentes tipos de agrupamiento posibles, el que genera clusters más claros es el de Ward, por lo que aplicaremos a ambas matrices de distancia este tipo de agrupamiento, representando después el dendograma con la función `plot`. Para la distancia Euclídea ejecutaríamos

```
> plot(hclust(A1,method="ward"),main="Dendograma de Cuencos")
```

obteniendo la Figura 1.3. Con la distancia de Manhattan ejecutamos

```
> plot(hclust(A2,method="ward"),main="Dendograma de Cuencos")
```

y obtenemos la Figura 1.4.

Aunque de cada una de las dos figuras se obtienen resultados un poco distintos, podemos obtener las siguientes conclusiones: Observamos en ambos

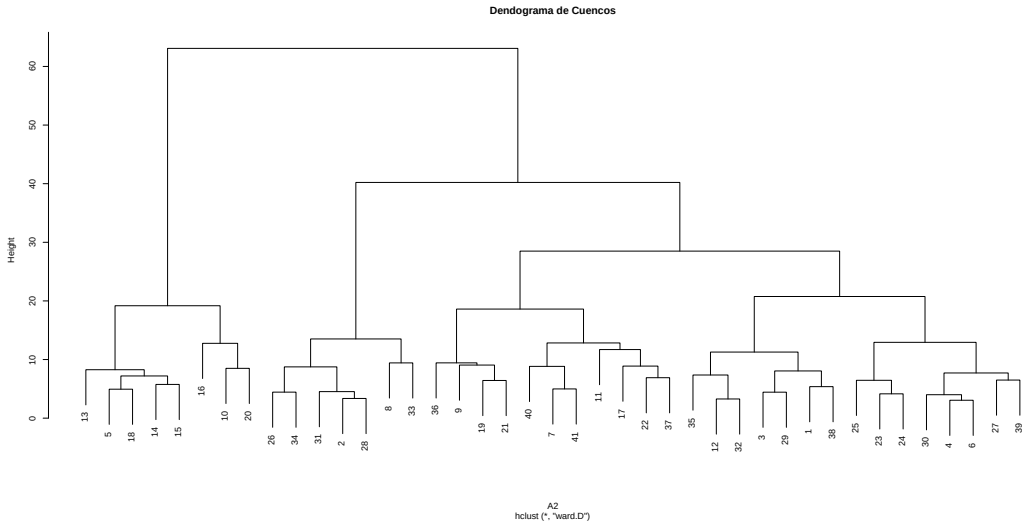



Figura 1.4 : Dendograma con la distancia de Manhattan

dendogramas que la última unión (en la que ya sólo obtenemos un grupo), se produce a una distancia muy grande (más de 20 y más de 60 respectivamente) en comparación con las otras uniones, lo que indica que quizás, lo más razonable sería hacer sólo dos grupos: con la distancia Euclídea, un grupo con los cuencos que van desde el Cuenco 9 hasta el Cuenco 33 (véase la base de la Figura 1.3) y el otro grupo con el resto de cuencos, mientras que con la distancia de Manhattan, un grupo con los que van desde el Cuenco 13 hasta el 20 y otro grupo desde el 26 al 39.

El Criterio del Codo, obtenido ejecutando

```
> codo(NCuencos,numero=40)
```

parece indicarnos que el descenso de mayor pendiente se da para 2 ó 3 clusters.

No obstante, si queremos aumentar un poco más el número de grupos, sólo tenemos que bajar un poco la recta de corte del dendograma paralela al eje de abscisas. A la vista de los gráficos, como mucho, se pueden considerar 3, si trazamos la paralela, por ejemplo, a la altura 12 en el primer gráfico ó 35 en el segundo gráfico.

Considerando 3 clusters, con la distancia Euclídea (primer dendograma) obtenemos los grupos: $\{9, 19, 21, 17, 22, 37, 26, 34, 31, 2, 28, 36, 8, 33\}$, $\{16, 20, 10, 14\}$, y $\{12, 32, 35, 1, 38, 40, 27, 30, 4, 6, 11, 13, 5, 18, 23, 24, 15, 39, 3, 29, 25, 7, 41\}$.

Y con la distancia de Manhattan (segundo dendograma) obtenemos los grupos: $\{13, 5, 18, 14, 15, 16, 10, 20\}$, $\{26, 34, 31, 2, 28, 8, 33\}$, y $\{36, 9, 19, 21, 40, 7, 41, 11, 17, 22, 37, 35, 12, 32, 3, 29, 1, 38, 25, 23, 24, 30, 4, 6, 27, 39\}$.

**Problema 1.4**

**Los datos recogidos por Shumway y Verosub (1992), y que están en el fichero "`sedimentos.txt`", corresponden al espesor de capas de sedimento depositadas por glaciares cerca de Massachusetts en los meses de deshielo de 634 años, desde el año -9835 al año -9202. Ajustar una modelo ARIMA a la serie temporal dada.**

Este tipo de datos aporta mucha información paleoclimática sobre otras variables muy relacionadas, tales como la temperatura de la época porque, en un año cálido, se deposita más tierra y cieno en el fondo del glaciar y, como la disminución del espesor implica una mayor cantidad de depósitos, un aumento de las capas de sedimento en un momento temporal implica un aumento de la temperatura en esa época.

Para ajustar el modelo ARIMA utilizaremos R. Primero incorporaremos los datos a R ejecutando `(1)`. Como el periodo (`frequency`) de los valores de la serie es 1, no hace falta utilizar la función `stl` para analizar la normalidad de los residuos. Su gráfica, obtenida ejecutando `(2)` y que aparece en el lado izquierdo de la Figura 1.5 muestra claramente una falta de normalidad. El histograma de los datos transformados por logaritmos, obtenidos ejecutando `(3)`, sí parece indicar normalidad.

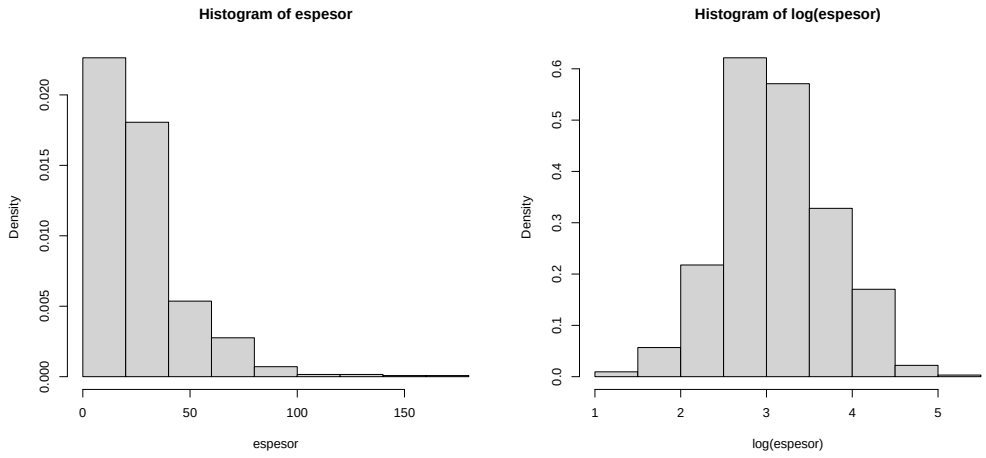



Figura 1.5 : Histogramas para los datos y los logaritmos de los datos

```
> espesor<-ts(scan("e:\\sedimentos.txt"),start=-9835,frequency=1)          (1)
> par(mfrow=c(1,2))
> hist(espesor,prob=T)                                                     (2)
> hist(log(espesor),prob=T)                                                (3)
```

Ahora, lo más simple es ajustar un modelo arima ejecutando `(4)`, lo que

nos sugiere en (5) un modelo ARIMA(1,1,1).

```
> library(forecast)
> auto.arima(log(espesor))                                    (4)
Series: log(espesor)
ARIMA(1,1,1)                                                  (5)

Coefficients:
         ar1      ma1
      0.2330  -0.8858
s.e.  0.0518   0.0292

sigma^2 = 0.2292:  log likelihood = -431.44
AIC=868.88   AICc=868.91   BIC=882.23
```

De hecho, si representamos la serie ejecutando (6) y obteniendo el gráfico de la izquierda de la Figura 1.6, vemos que la serie no parece estacionaria. La representación de la serie diferenciada, obtenida ejecutando (7) y que aparece en el gráfico de la derecha de la misma Figura 1.6, sí que muestra una serie estacionaria. El orden 1 de la segunda componente del ARIMA parece adecuado.

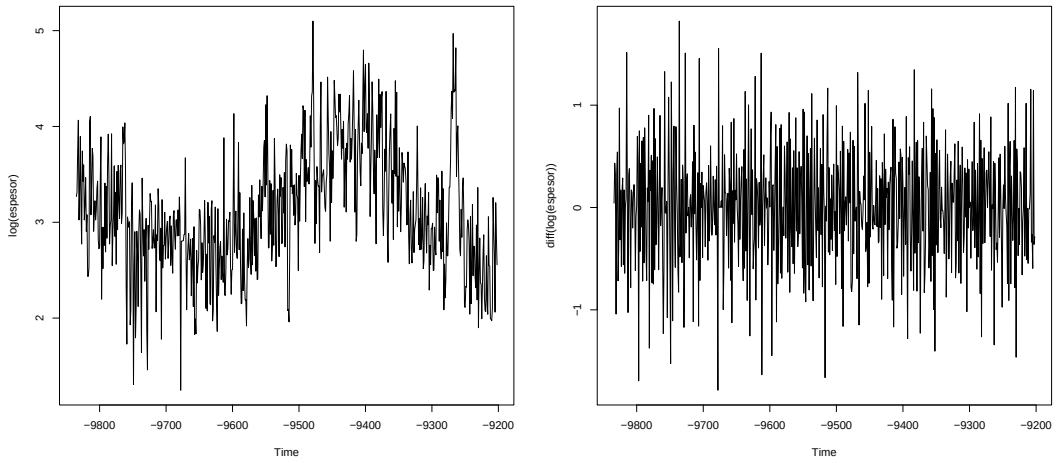



Figura 1.6 : Serie sin diferenciar y diferenciada

```
> par(mfrow=c(1,2))
> plot(log(espesor))                                          (6)
> plot(diff(log(espesor)))                                    (7)
```

Si resumimos las indicaciones dadas en EAA-sección 13.6.1 sobre la identificación del modelo ARMA (ya estacionario) en base a las representaciones de

las funciones de correlación parcial y auto-correlación parcial en la siguiente tabla, en donde decrecer rápidamente significa que queda dentro de las bandas de confianza del dibujo,

| | AR($p$) | MA($q$) | ARMA($p, q$) |
|---|---|---|---|
| ACF | No decrece | Decrece a cero después de $q$ retardos | No decrece |
| PACF | Decrece a cero después de $p$ retardos | No decrece | No decrece |

la representación de las funciones de correlación parcial y auto-correlación parcial de la serie diferenciada, obtenidas ejecutando la siguiente secuencia, la cual da como resultado la Figura 1.7 parece indicarnos un modelo ARMA(0,1,1).

```
> par(mfrow=c(1,2))
> acf(diff(log(espesor)))
> pacf(diff(log(espesor)))
```

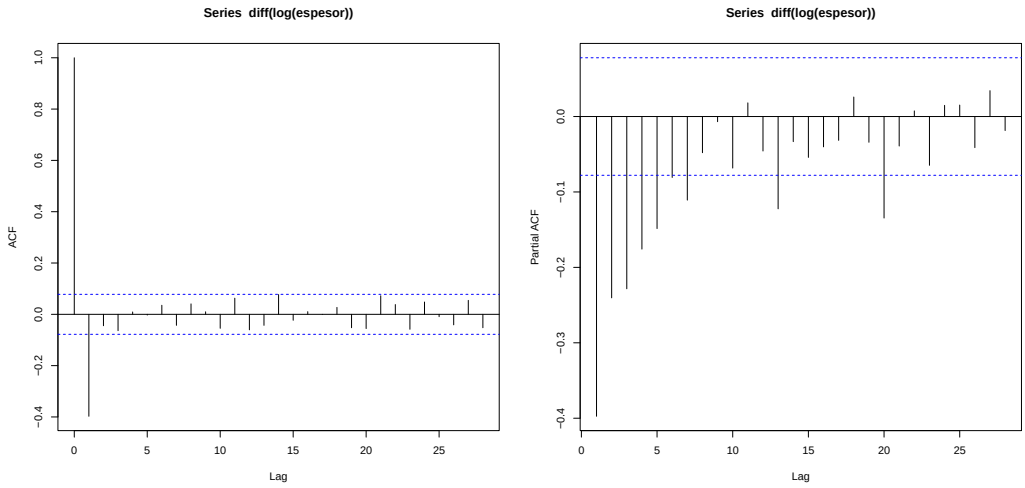



Figura 1.7 : Funciones de correlación y auto-correlación de la serie diferenciada

Si observamos la verosimilitud (y el valor del AIC) de este modelo ejecutando

```
> arima(log(espesor),order=c(0,1,1))

Call:
arima(x = log(espesor), order = c(0, 1, 1))
```

```
Coefficients:
          ma1
      -0.7705
s.e.   0.0341

sigma^2 estimated as 0.2353:  log likelihood = -440.72,  aic = 885.44
```

vemos que apenas se reduce el logaritmo de la verosimilitud, que pasa de $-431'44$ a $-440'88$ o que tampoco aumenta mucho el AIC, que pasa de $868'88$ a $885'44$. No obstante, si realizamos la diagnosis de ambos modelos ejecutando la función `tsdiag`, vemos en la Figura 1.8 que el modelo ARIMA(0,1,1) no pasa el test de Ljung-Box, mientras que el modelo ARIMA(1,1,1) sí lo pasa, según el gráfico de la Figura 1.9.

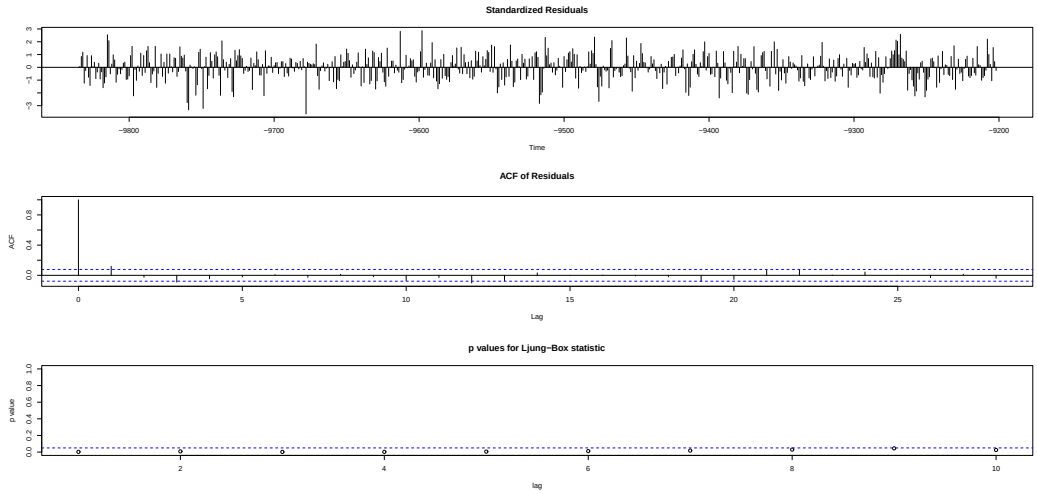



Figura 1.8 : Tests diagnóstico del modelo ARIMA(0,1,1)

```
> resultado1<-arima(log(espesor),order=c(0,1,1))
> resultado2<-arima(log(espesor),order=c(1,1,1))
> tsdiag(resultado1)
> tsdiag(resultado2)
```

Nos quedamos, por tanto, con el modelo ARIMA(1,1,1), con polinomios asociados

$$\delta_p(L) = 1 - 0'233\, L$$

$$\Delta_P(L^s) = 1$$

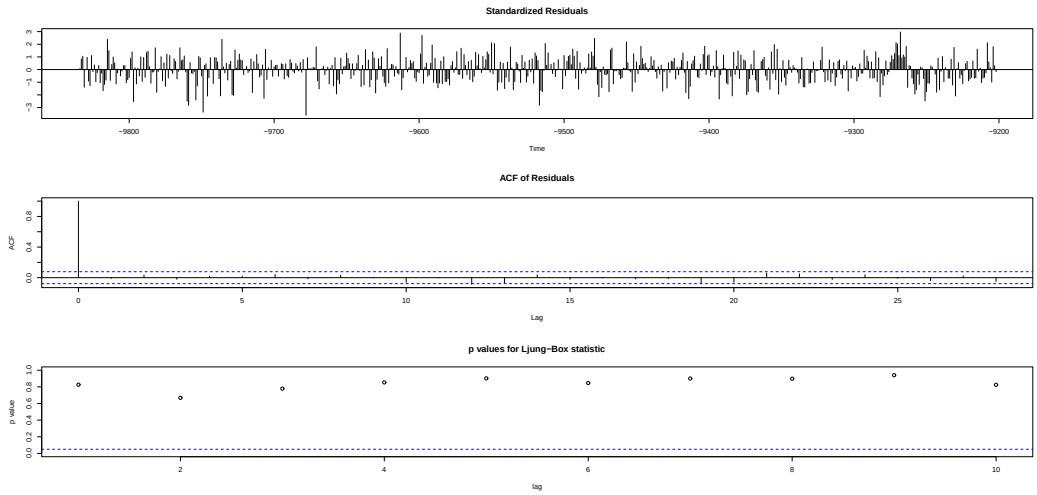



Figura 1.9 : Tests diagnóstico del modelo ARIMA(1,1,1)

$\theta_q(L) = 1 - 0'8858\, L$

$\Theta_Q(L^s) = 1$

por lo que la ecuación

$$\delta_p(L)\Delta_P(L^s)Z_t = \theta_q(L)\Theta_Q(L^s)e_t$$

que define el ARIMA genérico, quedará igual a

$$(1 - 0'233\, L)Z_t = (1 - 0'8858\, L)e_t$$

con $Z_t = \nabla^1 Y_t = Y_t - Y_{t-1}$ . Con lo que, haciendo operaciones, quedará,

$$Y_t = 1'233\, Y_{t-1} - 0'233\, Y_{t-2} + e_t - 0'8858\, e_{t-1}$$

**Problema 1.5**

**Janssen et al (1998) analizaron 180 vasijas de cristal del siglo 15 al siglo 17 mediante rayos X, para determinar las concentraciones de 13 elementos presentes en las vasijas. Los resultados están en el fichero "`Vasijas.txt`".**
**El objetivo es formar grupos de vasijas por la concentración de estos elementos, mediante la utilización de las dos primeras Componentes Principales.**

Mediante el Análisis de Componentes Principales, EAA-capítulo 2, también se pueden formar grupos de individuos, como ocurre con el Análisis de Conglomerados. La manera de hacerlo es representar los datos transformados (los scores) en el nuevo sistema de coordenadas formado por las dos primeras Componentes Principales. La calidad de esta representación vendrá medida por el porcentaje de varianza recogido por esas dos primeras componentes.

Primero incorporamos los datos a R ejecutando

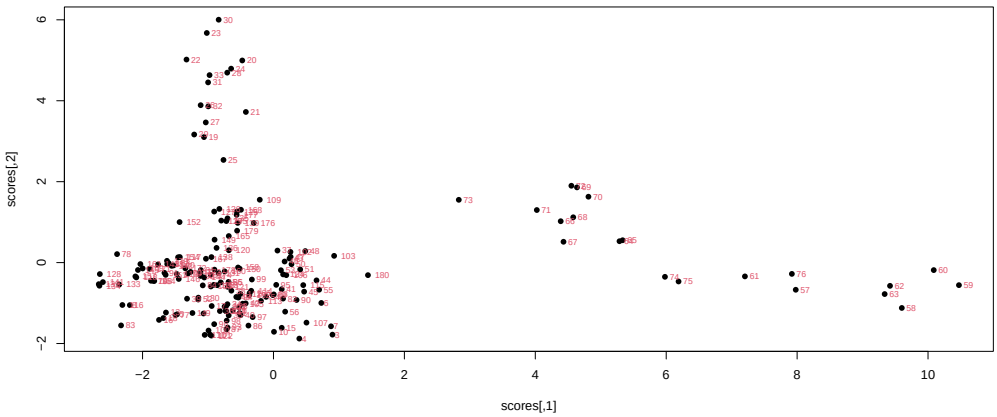



Figura 1.10 : Scores en los ejes de las dos primeras Componentes Principales

```
> Vasijas<-read.table("e:\\Vasijas.txt",header=T)
```

Ahora obtenemos las Componentes Principales y su contribución ejecutando las dos siguientes sentencias. Apuntamos que la función `prcomp` admite como datos, tanto a matrices como a data frames.

```
> resul<-prcomp(Vasijas,scale=T)
> summary(resul)
Importance of components:
                          PC1    PC2    PC3    PC4    PC5    PC6    PC7    PC8
Standard deviation     2.528 1.542 1.1314 1.069 0.7751 0.5781 0.5367 0.4475
Proportion of Variance 0.492 0.183 0.0985 0.088 0.0462 0.0257 0.0222 0.0154
Cumulative Proportion  0.492 0.675 0.7732 0.861 0.9073 0.9330 0.9552 0.9706
                              (1)
                           PC9    PC10   PC11    PC12    PC13
Standard deviation     0.3885 0.33798 0.2624 0.21961 0.00517
Proportion of Variance 0.0116 0.00879 0.0053 0.00371 0.00000
Cumulative Proportion  0.9822 0.99099 0.9963 1.00000 1.00000
```

Como vemos en `(1)`, con las dos primeras Componentes Principales recogemos sólo el 67′5 % de la variación total en los datos, lo que implica que

la clasificación con estas dos primeras Componentes Principales tampoco es óptima.

Los scores los obtenemos ejecutando

```
> scores<-matrix(c(resul$x[,1],resul$x[,2]),ncol=2)
```

y su representación obtenida ejecutando

```
> plot(scores,pch=16)
> text(scores[,1],scores[,2],1:180,adj=-0.5,cex=0.7,col=2)
```

es la dada por la Figura 1.10.

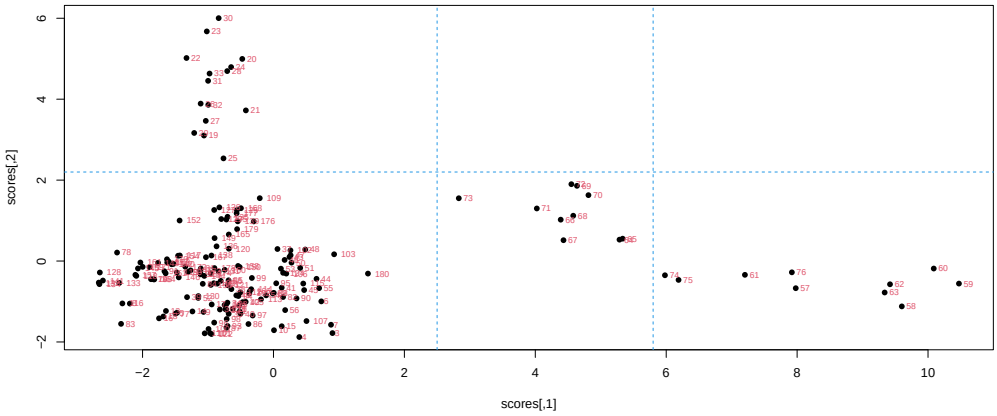



Figura 1.11 : Grupos de Vasijas

Como hay 180 datos, algunas observaciones no se identifican correctamente. Si, por ejemplo, queremos ver si las observaciones 26 y 32 están cerca, además de subir la resolución del fichero pdf, podemos ejecutar los scores, o más concreto

```
> c(scores[26,1],scores[26,2])
[1] -1.112300  3.890444
> c(scores[32,1],scores[32,2])
[1] -0.9957682  3.8610167
```

que verifica que estos dos individuos estás próximos.

Como conclusión podemos hacer cuatro grupos. Uno, con la mayoría de las observaciones, en el lado inferior izquierdo; otro, sobre este primer grupo y, los dos últimos en el lado derecho, divididos por la línea vertical, como puede verse en la Figura 1.11, rectas añadidas a la Figura 1.10 ejecutando

```
> abline(h=2.2,lty=2,col=4)
> abline(v=2.5,lty=2,col=4)
> abline(v=5.8,lty=2,col=4)
```

Los 4 grupos serán, por tanto, Grupo II= $\{19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33\}$ , Grupo III= $\{64, 65, 66, 67, 68, 69, 70, 71, 72, 73\}$ , Grupo IV= $\{57, 58, 59, 60, 61, 62, 63, 74, 75, 76\}$ , y Grupo I = resto de observaciones = $\{1$ a $18, 34$ a $56, 77$ a $180\}$ . Aunque también podrían juntarse los Grupos III y IV. Un estudio detallado posterior de las vasijas descubrió que había cuatro tipos diferentes de vasijas que se correspondían con los cuatro grupos formados por el Análisis de Componentes Principales.

Este ejemplo pone de manifiesto que este Método es útil en la clasificación de observaciones.

### Problema 1.6

**Para comprobar experimentalmente la conocida como "Ley de covariación de Buckman", que expresa una fuerte relación lineal entre varias partes de los amonites y otros moluscos, Hammer y Bucher (2005) recogieron las medidas de 14 especímenes de ammonite Pseudodanubites halli, animal del Jurásico que también se extinguió como los dinosaurios hace 65 millones de años. En concreto se midió la Anchura de sus espirales, $W$; la Altura de sus espirales, $H$; la altura de sus costillas laterales, $LH$, y la altura de sus costillas ventrales $VH$, obteniéndose los siguientes datos:**

| $W$ | $H$ | $LH$ | $VH$ |
|---|---|---|---|
| 14′27 | 11′89 | 1′69 | 0′61 |
| 17′19 | 14′89 | 2′59 | 1′33 |
| 19′22 | 12′95 | 2′11 | 0′65 |
| 21′16 | 11′67 | 3′19 | 0′41 |
| 16′00 | 14′61 | 1′76 | 0′51 |
| 15′28 | 13′29 | 1′99 | 0′40 |
| 16′45 | 11′40 | 1′03 | 0′30 |
| 16′28 | 9′80 | 2′37 | 0′20 |
| 11′29 | 9′36 | 1′30 | 0′19 |
| 15′94 | 9′79 | 1′90 | 0′36 |
| 12′26 | 9′03 | 1′17 | 0′19 |
| 12′05 | 12′00 | 0′84 | 0′18 |
| 9′20 | 8′50 | 0′64 | 0′32 |
| 14′34 | 9′88 | 1′42 | 0′15 |

**también en el fichero "`ammonite.txt`". En el trabajo se analizaron las regresiones lineales simples de $VH$ sobre $H$, de $VH$ sobre $W$, de $LH$ sobre $H$ y de $LH$ sobre $W$ con objeto de determinar una constante de proporcionalidad que permitiera predecir las variables dependientes en función de las independientes. En unos casos llegaron a conclusiones de fuertes relaciones lineales y en otros no, debido a la presencia de datos anómalos. Comparar, en cada caso, la recta de mínimos cuadrados con las rectas de regresión robustas.**

Primero incorporamos los datos a R ejecutando

```
> ammonite<-read.table("e:\\ammonite.txt",header=T)
> ammonite
       W     H   LH   VH
1  14.27 11.89 1.69 0.61
2  17.19 14.89 2.59 1.33
3  19.22 12.95 2.11 0.65
4  21.16 11.67 3.19 0.41
5  16.00 14.61 1.76 0.51
6  15.28 13.29 1.99 0.40
7  16.45 11.40 1.03 0.30
8  16.28  9.80 2.37 0.20
9  11.29  9.36 1.30 0.19
10 15.94  9.79 1.90 0.36
11 12.26  9.03 1.17 0.19
12 12.05 12.00 0.84 0.18
13  9.20  8.50 0.64 0.32
14 14.34  9.88 1.42 0.15
```

Vamos a calcular, para cada par de variables solicitadas, la recta de mínimos cuadrados y tres rectas robustas que permiten analizar la significación del test sobre la igualdad a cero del coeficiente de regresión: la de Regresión de Huber (ejecutada con la función `rlm` de la librería MASS), la recta LTS (ejecutada con la función `ltsReg` de la librería robustbase) y la recta MM (ejecutada con la función `lmrob` de la librería robustbase) (ver EARR-capítulo 7).

Regresión de $VH$ sobre $H$:

Recta de Mínimos Cuadrados=`recta11`

```
> recta11<-lm(VH~H,data=ammonite)
> recta11

Call:
lm(formula = VH ~ H, data = ammonite)

Coefficients:
(Intercept)            H
    -0.8077       0.1076

> summary(recta11)

Call:
lm(formula = VH ~ H, data = ammonite)

Residuals:
     Min       1Q   Median       3Q      Max
-0.30297 -0.11506 -0.02325  0.10226  0.53619

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.80773    0.34974  -2.309  0.03951 *
H            0.10756    0.03033   3.547  0.00402 **                          (1)
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2244 on 12 degrees of freedom
Multiple R-squared: 0.5118,     Adjusted R-squared: 0.4711
F-statistic: 12.58 on 1 and 12 DF,  p-value: 0.004022
```

### Recta Robusta de Huber=recta12

```
> library(MASS)
> recta12<-rlm(VH~H,data=ammonite)
> recta12
Call:
rlm(formula = VH ~ H, data = ammonite)
Converged in 6 iterations

Coefficients:
(Intercept)           H
 -0.5389272   0.0813886

Degrees of freedom: 14 total; 12 residual
Scale estimate: 0.185
> summary(recta12)

Call: rlm(formula = VH ~ H, data = ammonite)
Residuals:
     Min       1Q   Median       3Q      Max
-0.25774 -0.10862 -0.01944  0.12674  0.65705

Coefficients:
            Value   Std. Error t value
(Intercept) -0.5389  0.2846    -1.8935
H            0.0814  0.0247     3.2978

Residual standard error: 0.1854 on 12 degrees of freedom
> 2*(1-pt(3.2978,12))
[1] 0.006366644                                                             (2)
```

### Recta LTS=recta13

```
> library(robustbase)
> recta13<-ltsReg(VH~H,data=ammonite)
> recta13

Call:
ltsReg.formula(formula = VH ~ H, data = ammonite)
```