Assignment 4

Jose Pérez Cano & Álvaro Ribot Barrado

1.

```
seeds <- read.table("http://www-eio.upc.es/%7Ejan/Data/MVA/seedsdataset.dat", col.names = c("Area", "Pe
head(seeds)
```

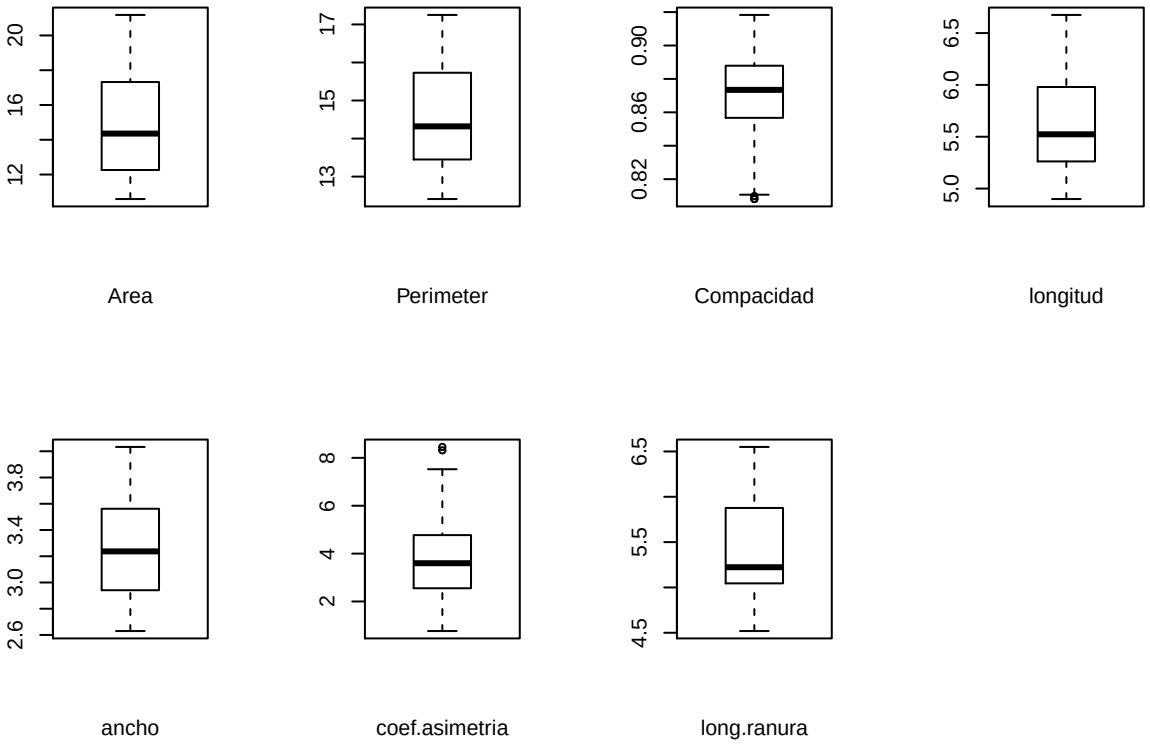
```
##    Area Perimeter Compacidad longitud ancho coef.asimetria long.ranura
## 1 15.26     14.84     0.8710    5.763 3.312          2.221       5.220
## 2 14.88     14.57     0.8811    5.554 3.333          1.018       4.956
## 3 14.29     14.09     0.9050    5.291 3.337          2.699       4.825
## 4 13.84     13.94     0.8955    5.324 3.379          2.259       4.805
## 5 16.14     14.99     0.9034    5.658 3.562          1.355       5.175
## 6 14.38     14.21     0.8951    5.386 3.312          2.462       4.956
##   variedad
## 1        1
## 2        1
## 3        1
## 4        1
## 5        1
## 6        1
```

2.

```
variedad <- seeds[,8]
seeds <- seeds[,-8]
nums <- row.names(seeds)
```

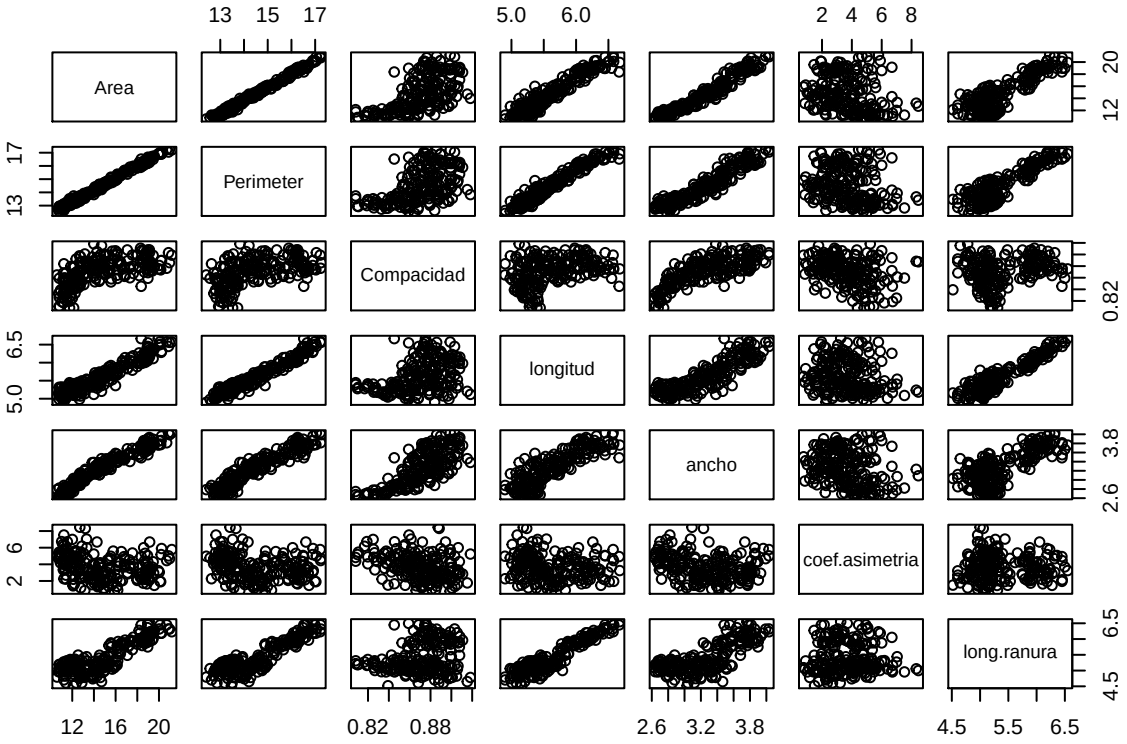
3. No se observan grandes anomalías pero sí se observa que hay dist.stribuciones no simétricas como la longitud de ranura y algunos outliers en el coeficiente de asimetría.

```
par(mfrow = c(2,4))
for(i in 1:ncol(seeds)) {
  boxplot(seeds[,i], xlab = names(seeds)[i])
}
```



4. Al haber relaciones no lineales entre las variables salen formas diversas. El coeficiente de asimetría parece no tener correlación con ninguna variable. Y en algunos plots parecen diferenciarse al menos dos grupos. Por lo demás no parece que haya anomalías destacables. Además, podemos observar que hay pares de variables con correlacion muy alta (por ejemplo, Área y Perímetro).

```
pairs(seeds)
```



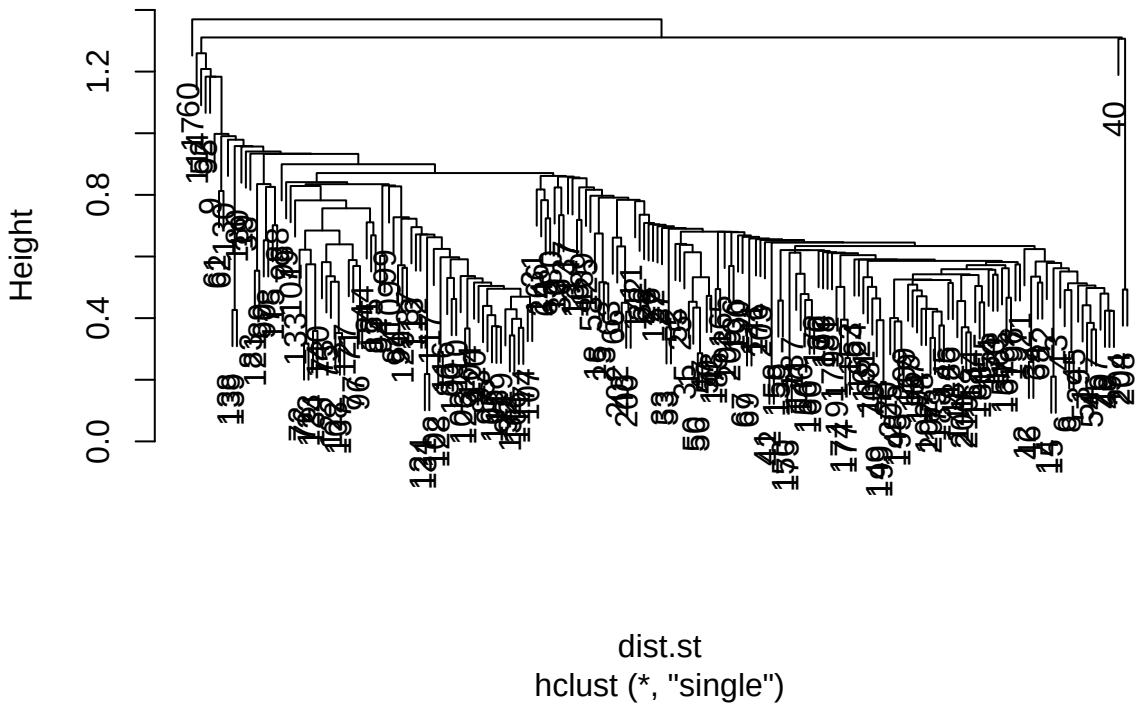
5.

```
non.scaled.seeds <- seeds
seeds <- scale(seeds)
dist.st <- dist(seeds)
as.matrix(dist.st)[1:5,1:5]
```

```
##          1        2         3         4        5
## 1 0.000000 1.181864 2.0970435 1.8748257 1.679029
## 2 1.181864 0.000000 1.6949066 1.3385798 1.361268
## 3 2.097044 1.694907 0.0000000 0.5513196 1.797465
## 4 1.874826 1.338580 0.5513196 0.0000000 1.764733
## 5 1.679029 1.361268 1.7974648 1.7647326 0.000000
```

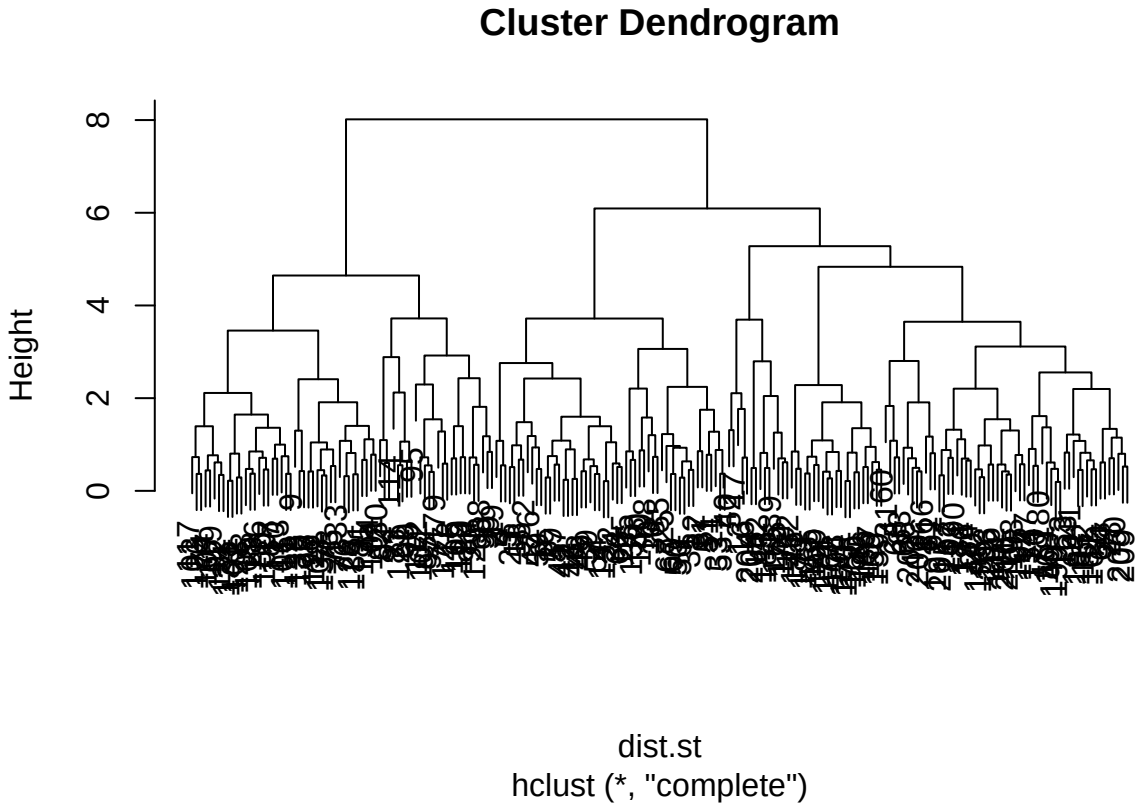
6. Si a partir de este dendograma tuvieramos que decir que hay tres variedades concluiriamos que dos de ellas solo aparecen en tres semillas y la otra tiene el resto de semillas. Este dendograma no separa demasiado y junta muchas semillas en un mismo grupo, por lo que no queda nada claro que haya tres variedades a partir de este dendograma.

```
nearest_neighbour <- hclust(dist.st, method = "single")
plot(nearest_neighbour)
```



7. Con este método obtenemos mejores separaciones y se ve que está más balanceado el reparto de las semillas por grupo. También vemos que los grandes grupos se juntan más arriba.

```
farthest_neighbour <- hclust(dist.st, method = "complete")
plot(farthest_neighbour)
```



8. La dist.stancia aproximada deberia ser 5.7. El análisis ha identificado bastante bien a las variedades. De hecho, un 88% de las semillas se identificarían correctamente si suponemos que los clusters coinciden con las variables.

```
clusters <- cutree(farthest_neighbour, h = 5.7)
(cross.table <- table(variedad, clusters))
```

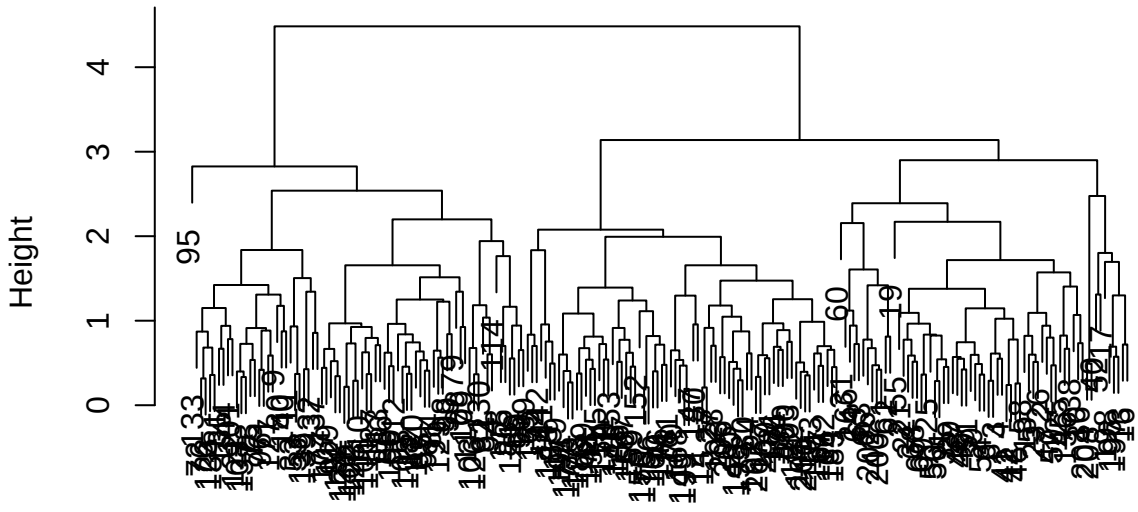
```
##         clusters
## variedad  1  2  3
##        1 48  2 20
##        2  4 66  0
##        3  0  0 70
```

```
sum(diag(cross.table))/sum(cross.table)
```

```
## [1] 0.8761905
```

9. Se observa que la variedad "Kama" y "Canadian" se agrupan en un cluster y la variedad "Rosa" en otro con tan solo 7 semillas de "Kama" en el cluster 2 y 2 de "Rosa" en el cluster 1. Observamos pues que el "average linkage" separa muy bien las semillas de las variedad 2 de las semillas de las variedades 1 y 3.

```
avg_link <- hclust(dist.st, method = "average")
plot(avg_link)
```

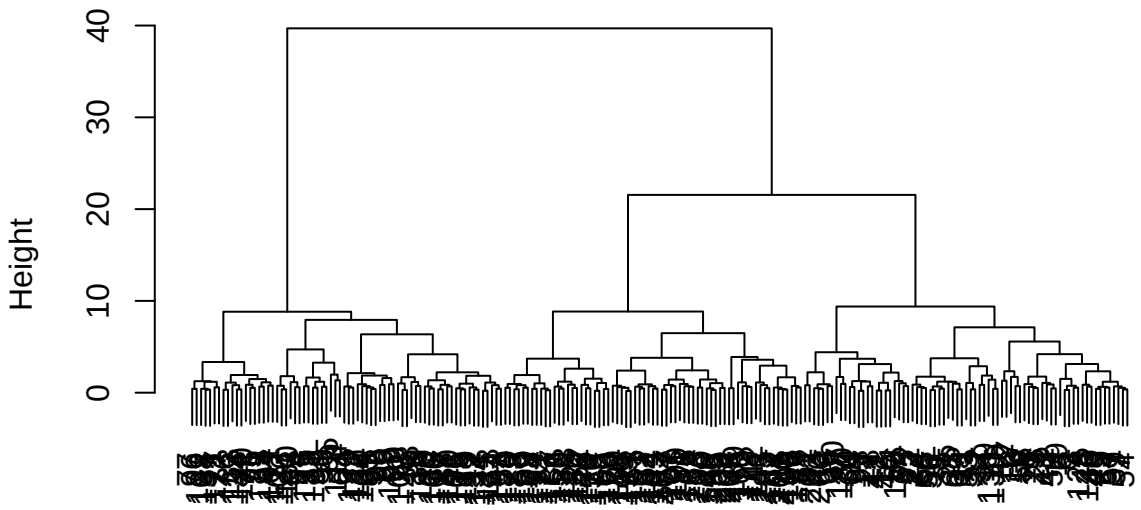



```
clusters2 <- cutree(avg_link, k = 2)
table(variedad, clusters2)
```

```
##         clusters2
## variedad  1  2
##        1 63  7
##        2  2 68
##        3 70  0
```

10. Se clasifican correctamente un 93% de las semillas

```
out.w <- hclust(dist.st, method = "ward.D2")
plot(out.w)
```




```
groups.3.ward <- cutree(out.w, k = 3)
(cross.table.ward <- table(groups.3.ward, variedad))
```

```
##              variedad
## groups.3.ward  1  2  3
##             1 64  4  5
##             2  4 66  0
##             3  2  0 65
```

```
sum(diag(cross.table.ward))/sum(cross.table.ward)
```

```
## [1] 0.9285714
```

11. En primer lugar asocia el segundo grupo al tercero y el tercero al segundo, es decir, cambiando las etiquetas pero separandolos igual. Obviando ese detalle acierta en un 93% de las veces, exactamente igual que con el método de Ward.

```
set.seed(123)
(km <- kmeans(seeds, 3))
```

```
## K-means clustering with 3 clusters of sizes 75, 66, 69
##
## Cluster means:
##         Area  Perimeter Compacidad   longitud       ancho coef.asimetria
## 1 -0.2092509 -0.2443817  0.4209483 -0.3400265 -0.05324506    -0.65890742
## 2 -1.0414310 -1.0111329 -1.0339614 -0.8859528 -1.11844610     0.79756054
## 3  1.2235980  1.2328029  0.5314541  1.2170272  1.12769307    -0.04668027
##   long.ranura
## 1  -0.6627691
```

```
## 2  -0.5820606
## 3   1.2771548
##
## Clustering vector:
##   [1] 1 1 1 1 1 1 1 1 3 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [36] 1 1 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 2
##  [71] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
## [106] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 3 3 3 3 3 3 3 3 1 3 3 1 3
## [141] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2
## [176] 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 1 2 1 2 1 2 2 2 1 2 2 2 2
##
## Within cluster sum of squares by cluster:
## [1] 160.7960 118.2489 149.8787
##  (between_SS / total_SS =  70.7 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

```
km$cluster[km$cluster==2] <- 4
km$cluster[km$cluster==3] <- 2
km$cluster[km$cluster==4] <- 3

(cross.table.km <- table(km$cluster, variedad))
```

```
##    variedad
##      1  2  3
##   1 65  3  7
##   2  2 67  0
##   3  3  0 63
```

```
sum(diag(cross.table.km))/sum(cross.table.km)
```

```
## [1] 0.9285714
```

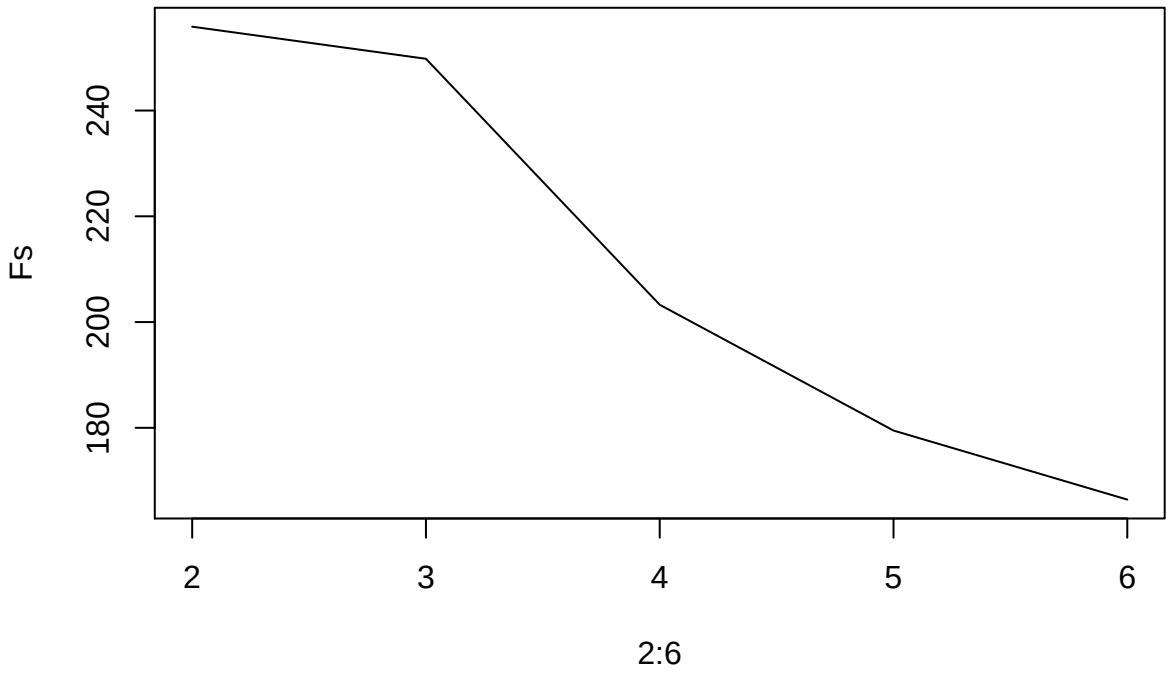
12. Según el criterio de Calinski-Harabasz la cantidad óptima de grupos es 2 para utilizar con kmeans.

```
library(clusterSim)
```

```
## Loading required package: cluster
```

```
## Loading required package: MASS
```

```
Fs <- c()
for (k in 2:6) {
  aux.km <- kmeans(seeds, k)
  Fs <- c(Fs, index.G1(seeds, aux.km$cluster))
}
plot(2:6, Fs, type="l")
```

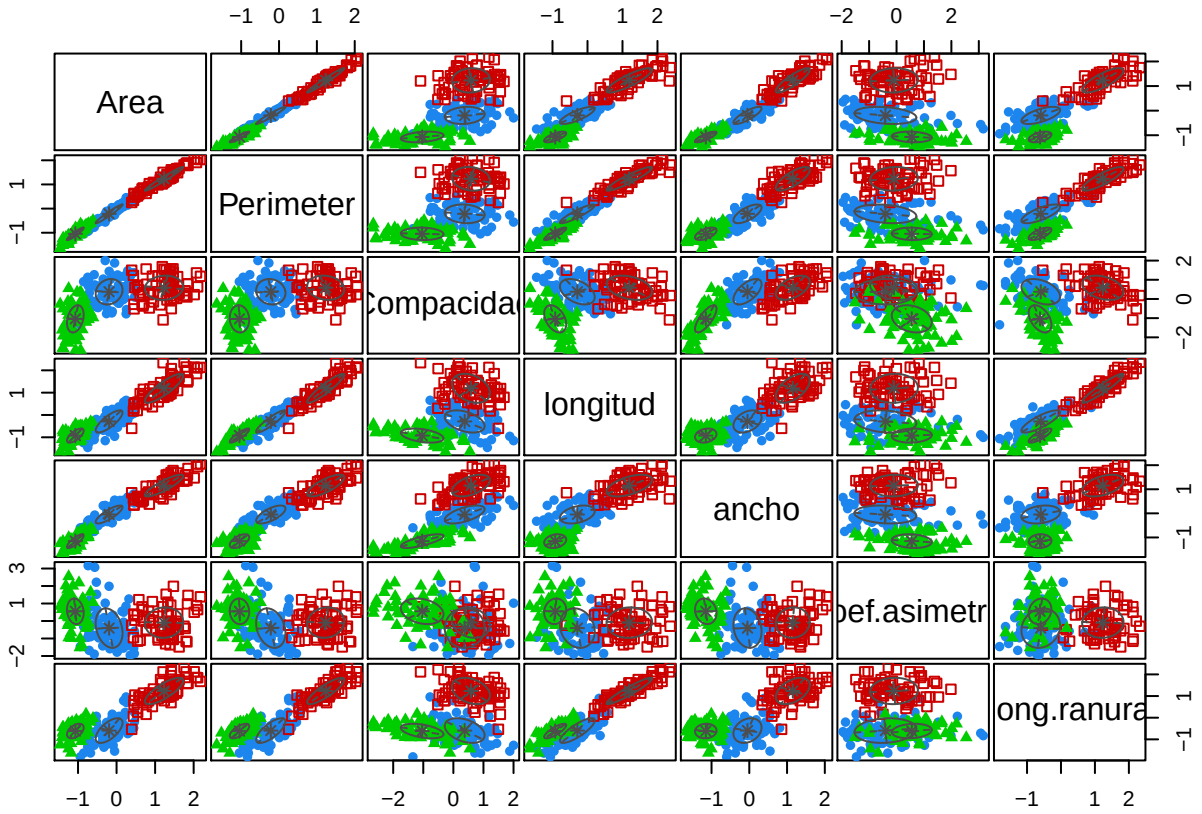


13. Se equivoca en el 10% de los casos, peor que los otros criterios. Si no estandarizásemos los datos se equivocaría un $8/$, mejoraría pero no llegaría a superar los otros criterios.

```
library(mclust)
```

```
## Package 'mclust' version 5.4.5
## Type 'citation("mclust")' for citing this R package in publications.
```

```
model <- Mclust(seeds, G=3)
plot(model, what="classification")
```




```
(cross.table.mix <- table(model$classification, variedad))
```

```
##    variedad
##       1  2  3
##   1 62  5  9
##   2  4 65  0
##   3  4  0 61
```

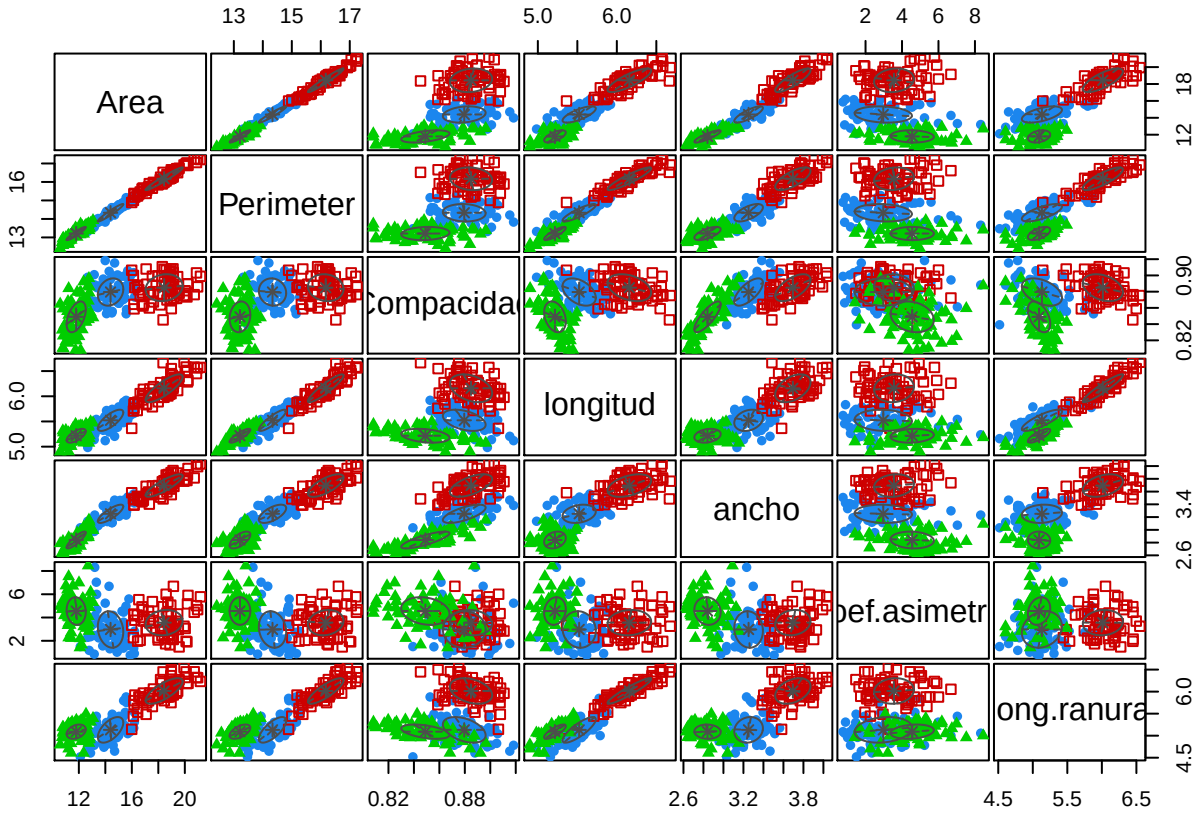
```
sum(diag(cross.table.mix))/sum(cross.table.mix)
```

```
## [1] 0.8952381
```

```
1-sum(diag(cross.table.mix))/sum(cross.table.mix)
```

```
## [1] 0.1047619
```

```
model2 <- Mclust(non.scaled.seeds, G=3)
plot(model2, what="classification")
```




```
(cross.table.mix2 <- table(model2$classification, variedad))
```

```
##    variedad
##       1  2  3
##   1 60  5  5
##   2  4 65  0
##   3  6  0 65
```

```
sum(diag(cross.table.mix2))/sum(cross.table.mix2)
```

```
## [1] 0.9047619
```

```
1-sum(diag(cross.table.mix2))/sum(cross.table.mix2)
```

```
## [1] 0.0952381
```

14. En total, los tres métodos clasifican mal 6 semillas simultáneamente. Que son la 38, 70, 136, 139, 200 y la 202.

```
# Guardé los id de cada semilla (row name) al principio en la variable nums
err.ward <- nums[groups.3.ward != variedad]
err.km <- nums[km$cluster != variedad]
err.km <- err.km[err.km %in% err.ward]
err.gauss <- nums[model$classification != variedad]

err <- err.gauss[err.gauss %in% err.km]
(err <- as.numeric(err))
```

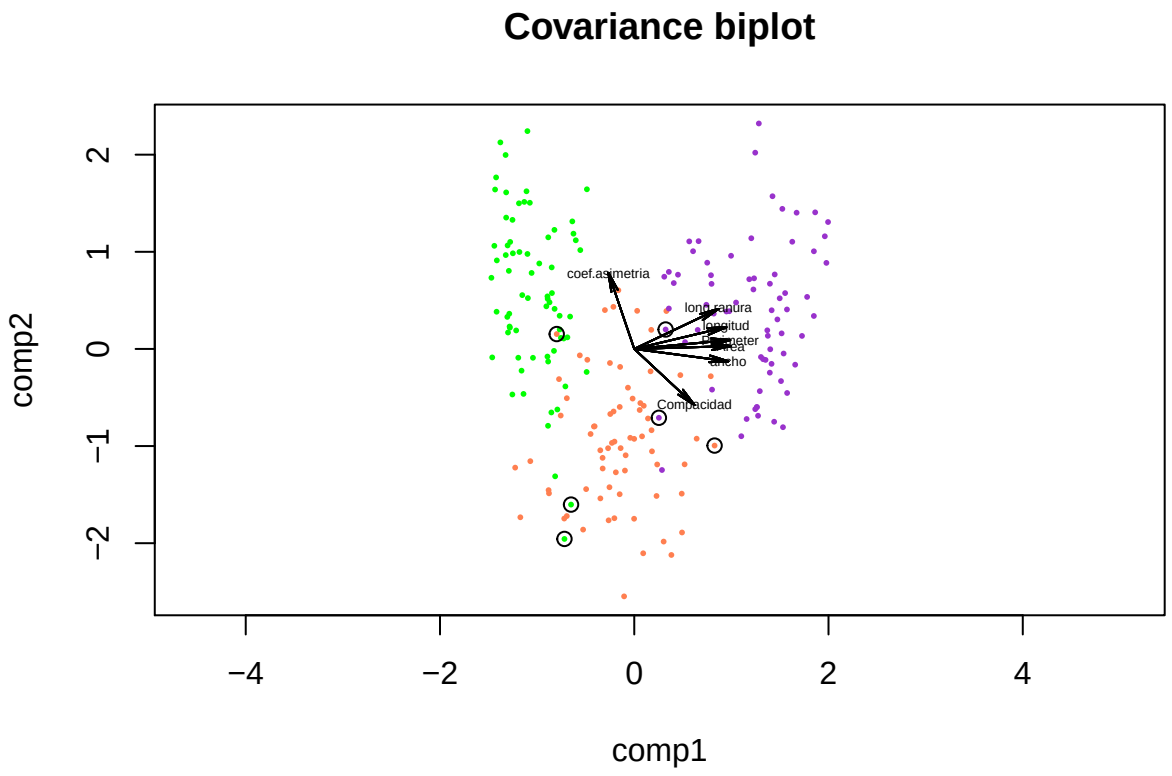
```
## [1]  38  70 136 139 200 202
```

```
length(err)
```

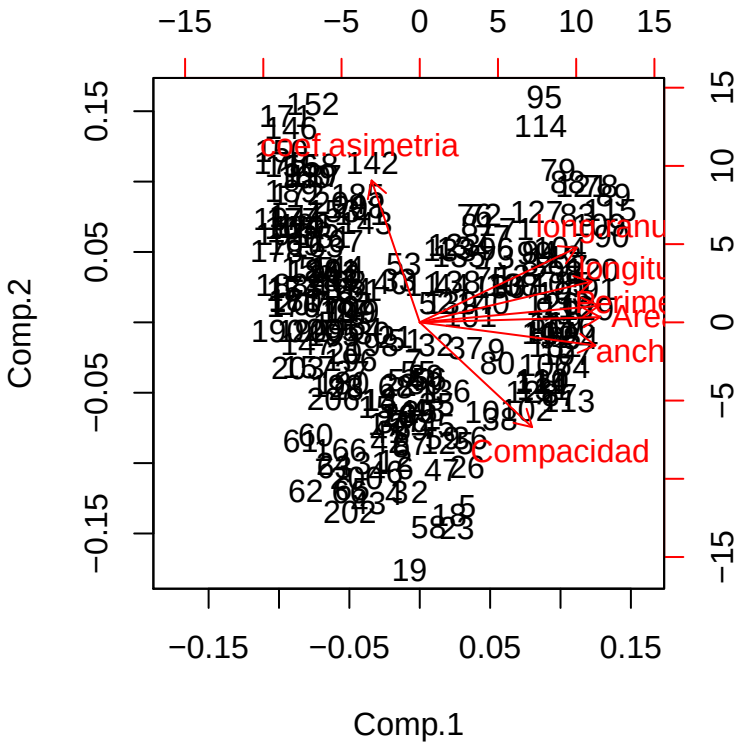
```
## [1] 6
```

15. Hemos representado el biplot tanto con la función biplot como con cálculos a mano para poder discernir los grupos con colores y los puntos que están mal clasificados resaltarlos. Las semillas mal clasificadas están rodeadas con un círculo negro. Observamos que los errores comunes se producen cuando una semilla de un grupo está situada muy cercana al centroide de los otros grupos o en la frontera entre ambos.

```
acp <- princomp(seeds)

cols <- variedad
cols[cols==1] <- "coral"
cols[cols==2] <- "darkorchid"
cols[cols==3] <- "green"

Fp <- acp$scores
Fs <- Fp %*% solve(diag(acp$sdev))

Gp <- acp$loadings
Gs <- Gp %*% diag(acp$sdev)

plot(Fs[,1], Fs[,2], col=cols, pch=16, cex=0.4, asp = 1, main="Covariance biplot", xlab="comp1", ylab="
points(Fs[err,1], Fs[err,2], asp=1, col="black")

arrows(rep(0, length(Gp)), rep(0,length(Gp)), Gs[,1], Gs[,2], length=0.1, angle=10)
text(x = Gs[,1], y = Gs[,2], label=colnames(seeds), cex=0.4)
```



```r
biplot(acp)
```



16.

Para intentar mejorar el resultado se podrían combinar diferentes métodos y usar un criterio democrático para decidir en caso de no haber acuerdo, de este modo los errores se reducirían a aquellas semillas que se clasifican en otro grupo simultáneamente por todos los métodos.

Aun así, esto tendría un límite. Los datos seguirán una dist.stribución determinada (desconocida) y frente a dicha dist.stribución existe un clasificador teórico llamado clasificador de Bayes que tiene una probabilidad de error que no se puede superar. Por tanto, cualquier método tendrá un cierto error al clasificar. Por ejemplo, existe una probabilidad no nula de que una semilla del grupo 2 se encuentre justo en el centroide del grupo 1, y sería imposible discernir que no es de ese grupo porque se parece más a ese grupo aunque sea improbable.