Modelos estadísticos multivariantes. Trabajo final.

Rafael Requena Garrido

June 21, 2019

1 Introducción

Este trabajo pretende ser una aplicación práctica de los conocimientos adquiridos en la asignatura ***Modelos estadísticos multivariantes*** del grado de Matemáticas de la Universidad de Málaga.

Disponemos de un conjunto de datos de 210 observaciones de 3 tipos de semillas de trigo distintas, para las cuales medimos las siguientes características: ***area, perimeter, compactness*** $(C = 4\pi A/P^2)$***, length of kernel, width of kernel, asymmetry coefficient, length of kernel groove, type***. Con el desarrollo del análisis perseguimos dos objetivos claros:

- **Predecir.** Como en la mayoría de los casos en los que se elabora un modelo matemático que *explique* y represente a una serie de datos, el objetivo principal va a ser encontrar un método que nos permita, dada una nueva observación de la cual desconocemos el tipo de semilla que es a priori, clasificarla y decididir a cuál de los 3 tipos pertenece con la mayor exactitud posible. Por tanto, queremos encontrar uno de los posibles modelos que mejor sea capaz de realizar estas predicciones.

- **Explicar.** Este objetivo va implícito en el análisis. No solo queremos poder desarrollar un modelo que nos permita explicar los datos, clasificarlos y que sea capaz de realizar buenas predicciones, sino que también queremos que este modelo sea lo más simple posible, es decir, queremos que tenga el menor número de variables mediante las cuales podamos explicar la mayor cantidad de información. Este punto es muy importante y no hay que perderlo nunca de vista, ya que a pesar de que en el ámbito académico no juegue un papel muy importante por no trabajar con cantidades de datos muy grandes, sí que es fundamental en los casos prácticos en los que se realiza cualquier análisis de datos con cantidades masivas de datos, ya que serían en muchos casos inviables de llevar a cabo sin la búsqueda de esta eficiencia.

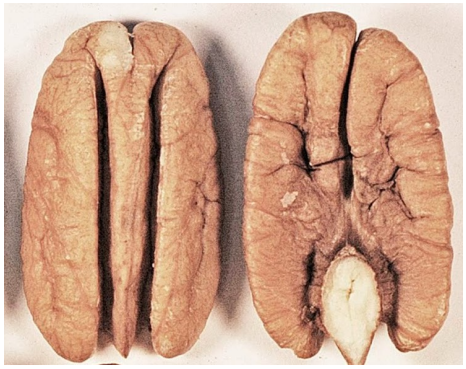

Figure 1: Semilla de trigo donde podemos apreciar la endidura del grano, cuya longitud estamos midiendo.

2 Metodología

Dado que en cada observación no estamos considerando un gran número de variables, el método que vamos a emplear en el estudio va a ser el *análisis discriminante.* Para ello, verificaremos al principio que se cumplen las hipótesis necesarias para poder utilizarlo, que son normalidad e igualdad de varianza.

Para comprobarlo, lo primero que haremos será ver cómo es la matriz de correlación de los datos. Esta, junto con los histogramas de cada variable, nos indicará no solo la normalidad de las mismas, sino también como se explican las unas con las otras. En esta parte utilizaremos varios métodos gráficos para ilustrar esta información, que no será más que representar en primera instancia la matriz de correlación y los histogramas por separado y posteriormente de forma conjunta.

Una vez realizadas estas comprobaciones comenzaremos con el análisis discriminante *lineal.* El grueso del trabajo consistirá en el desarrollo del mismo. Para ello, lo que haremos en primer lugar será desarrollarlo con todos los datos y ver cómo de bien es capaz de predecirlos. Esto sería la versión más simple y la menos realista, ya que estas predicciones no nos arrojarían gran información sobre lo bien que clasifica nuestro modelo nuevas observaciones. Una vez finalizada esta parte, lo que haremos será una versión "mejorada" y más realista en cuanto a lo que podemos encontrarnos en la práctica. Por tanto, lo que haremos en esta nueva situación será elaborar el mismo modelo pero *entrenándolo* con una partición de los datos, para así ver luego cómo se comporta a la hora de realizar las predicciones con las observaciones restantes. Cabe notar que, evidentemente, a mayor número de observaciones mejor será nuestro modelo. No obstante, como ya se ha mencionado en la introducción, en cualquier análisis de datos es fundamental poder explicar la mayor cantidad de información posible con el menor número de datos. Además, una vez extraída toda la información, veremos de forma rápida si uno cuadrático pudiera mejorarlo.

Finalmente, como tenemos suficientes variables, haremos el estudio de los datos mediante un ***análisis por componentes principales (PCA)*** de manera breve para corroborar que las conclusiones obtenidas en el método son correctas y que, por lo tanto, tenemos unos buenos datos, en referencia a que habrá una serie de variables que nos aporten suficiente información como para poder llevar a cabo la clasificación de las semillas.[1]

Comencemos con el análisis discriminante lineal. En este método estamos interesados en encontrar combinaciones lineales de las variables que separen de la mejor forma posible los distintos grupos de los datos. En nuestro caso, tratamos de encontrar las combinaciones lineales de las distintas variables medidas en las semillas que mejor nos dividan a las mismas en los 3 tipos posibles.

[1] Se podría plantear el análisis de los datos mediante el análisis clúster pero, en este caso, carece de sentido, pues este método tiene relevancia cuando desconocemos en qué grupos se distribuyen los datos.

Supongamos k poblaciones distintas en el espacio de dimensión. Nuestro objetivo es representar la separación de grupos en menos dimensiones, es decir, buscamos regiones

$$A_k = \{x : P(C_k)f_k(x) > P(C_i)f_i(x), \forall i, i \neq k\},$$

donde $P(C_k)$ es la probabilidad a priori de pertenecer al grupo k-ésimo y f_k es la función discriminante del mismo grupo. Si además suponemos que la probabilidad a priori de pertenencia a cada grupo es la misma, entonces la función discriminante es la regla de decisión mediante la cual clasificaremos nuestras observaciones en una población u otra. En este modelo, además, estamos suponiendo que los costes de equivocarnos a la hora de clasificar una observación en un grupo al que no pertenece son iguales, por eso el hecho de que pertenezca a una región o a otra solo depende de las probabilidades y las funciones discriminantes. Por tanto, si además consideramos que las funciones f_k sean distribuciones normales con distintos vectores de media pero igual matriz de covarianzas, llegamos a

$$P(C_k)f_k(x) > P(C_i)f_i(x) \iff D_i^2 > D_k^2, \ \forall i, i \neq k,$$

donde $D_i = (x-\mu_i)^T\Sigma^{-1}(x-\mu_i)$ es la distancia de Mahalanobis entre el punto observado, x, y la población i. Es decir, nuestro problema de clasificación se basa en encontrar la población cuya media esté más próxima a nuestra observación, usando esta distancia. Si simplificamos un poco más las expresiones anteriores y reordenamos, llegamos a definir el discriminador lineal, L_k,

$$L_k(x) := -2w_k^t x + w_k^t \mu_k,$$

que es la función que queremos minimizar, donde $w_k := \Sigma^{-1}\mu_k$. Lo siguiente a determinar son las fronteras entre las regiones i, k. Para ello, imponemos que $A_{ik} := L_i(x) - L_k(x) = 0$, donde A_{ik} denota la frontera entre la región i y la región k. Trabajando las expresiones anteriores, llegamos a que la frontera viene determinada por

$$w_{ik}^t x = w_{ik}^t(\frac{\mu_k + \mu_i}{2}),$$

que es la ecuación de un hiperplano, donde $w_{ik} := \Sigma^{-1}(\mu_k - \mu_i)$.

Por último, como hemos comentado que haremos un PCA para verificar los resultados obtenidos en el análisis discriminante, expongamos de forma escueta cuál es la idea de este método. Con esta técnica, lo que buscamos principalmente es reducir la dimensionalidad de nuestros datos, es decir, buscamos poder expresar la misma información que nos aporta cada una de las variables con un número menor de las mismas. Para ello, construimos nuevas variables que sean combinaciones lineales de las originales. Por eso, este método es muy recomendable cuando trabajamos con observaciones en las que medimos un gran número de variables, ya que muchas veces no todas las variables arrojan mucha información al análisis por ser poco explicativas o tener un alto grado de dependencia, con lo que las agrupamos en nuevas variables "artificiales" que nos ayudan a simplificar el problema y a realizar el estudio.

3 Resultados

Para realizar el análisis de los datos vamos a usar el lenguaje de programación R, integrado en el entorno RStudio. Lo primero que hacemos es instalar los paquetes que vamos a necesitar y cargar las librerías que vamos a emplear:[2]

```
install.packages("readr")
install.packages("corrplot")
install.packages("klaR")
install.packages("pander")
install.packages("GGally")
install.packages("devtools")
install.packages("psych")
install.packages("ggfortify")

library("readr")
library(MASS)
library(ggplot2)
library(dplyr)
library(klaR)
library(pander)
library(corrplot)
library(car)
library(GGally)
library(psych)
library(devtools)
library(ggfortify)

install_github("vqv/ggbiplot")
library(ggbiplot)
```

A continuación cargamos los datos y renombramos la cabecera para darle a las variables nombres con los que nos resulte más cómodo trabajar y hacemos una visualización de los mismos para ver que todo va bien:

```
data <- read.table(file.choose(), sep = "" , header = F)
names (data) = c("Ar", "Pe", "Co", "LofK", "WofK"
                 , "AsC", "LofKG", "Ty")

data
```

```
> data
    Ar    Pe     Co    LofK  WofK  AsC     LofKG  Ty
1   15.26 14.84  0.8710 5.763 3.312 2.2210 5.220  1
2   14.88 14.57  0.8811 5.554 3.333 1.0180 4.956  1
3   14.29 14.09  0.9050 5.291 3.337 2.6990 4.825  1
4   13.84 13.94  0.8955 5.324 3.379 2.2590 4.805  1
5   16.14 14.99  0.9034 5.658 3.562 1.3550 5.175  1
6   14.38 14.21  0.8951 5.386 3.312 2.4620 4.956  1
```

[2] En las representaciones no estamos teniendo en cuenta los tipos de las semillas, sino que solo tenemos en cuenta el resto de variables.

```
7    14.69  14.49  0.8799  5.563  3.259  3.5860  5.219   1
8    14.11  14.10  0.8911  5.420  3.302  2.7000  5.000   1
9    16.63  15.46  0.8747  6.053  3.465  2.0400  5.877   1
10   16.44  15.25  0.8880  5.884  3.505  1.9690  5.533   1
[ reached 'max' / getOption("max.print") --
omitted 200 rows ]
```

Como dijimos al comienzo de la metodología, antes de aplicar cualquier método, comprobamos que se verifiquen las hipótesis de normalidad multivariante e igualdad de varianzas. Para ello exponemos dos formas de ver la matriz de correlación: la primera un mapa de calor, y la segunda, no tan visual pero con mayor información a simple vista:

```
corrplot(cor(data), method = "color")
```

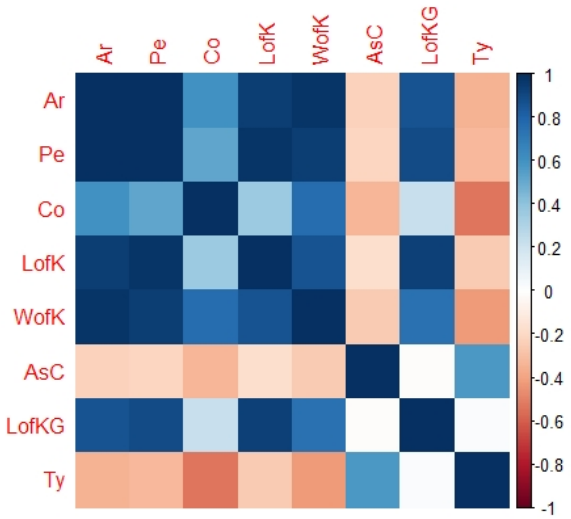

Figure 2: Mapa de calor

Escribimos ahora las instrucciones para elaborar nuestra versión más visual de la matriz de correlación:

```
pairs.panels(data[1:7],
gap = 0,
bg = c("blue", "green",
"red")[data$Ty],pch = 21)
mtext("Tipos de semillas: blue = 1; green = 2;
        red = 3", 1, line=3.7,cex=.8)
```

Para obtener la matriz de correlación:

```
> round(cor(data), 2)
```

	Ar	Pe	Co	LofK	WofK	AsC	LofKG	Ty
Ar	1.00	0.99	0.61	0.95	0.97	−0.23	0.86	−0.35

Pe	0.99	1.00	0.53	0.97	0.94	−0.22	0.89	−0.33
Co	0.61	0.53	1.00	0.37	0.76	−0.33	0.23	−0.53
LofK	0.95	0.97	0.37	1.00	0.86	−0.17	0.93	−0.26
WofK	0.97	0.94	0.76	0.86	1.00	−0.26	0.75	−0.42
AsC	−0.23	−0.22	−0.33	−0.17	−0.26	1.00	−0.01	0.58
LofKG	0.86	0.89	0.23	0.93	0.75	−0.01	1.00	0.02
Ty	−0.35	−0.33	−0.53	−0.26	−0.42	0.58	0.02	1.00

Y ahora, para visualizar mejor la información que nos aporta la matriz de correlación:

```
pairs(data[1:7],pch = 19,
col = my_cols[data$Ty],diag.panel=panel.hist,
          main="Semillas de trigo",
upper.panel=panel.cor)
mtext("Tipos de semillas: blue= 1; green= 2; red= 3"
          , 1, line=3.7,cex=.8)
```

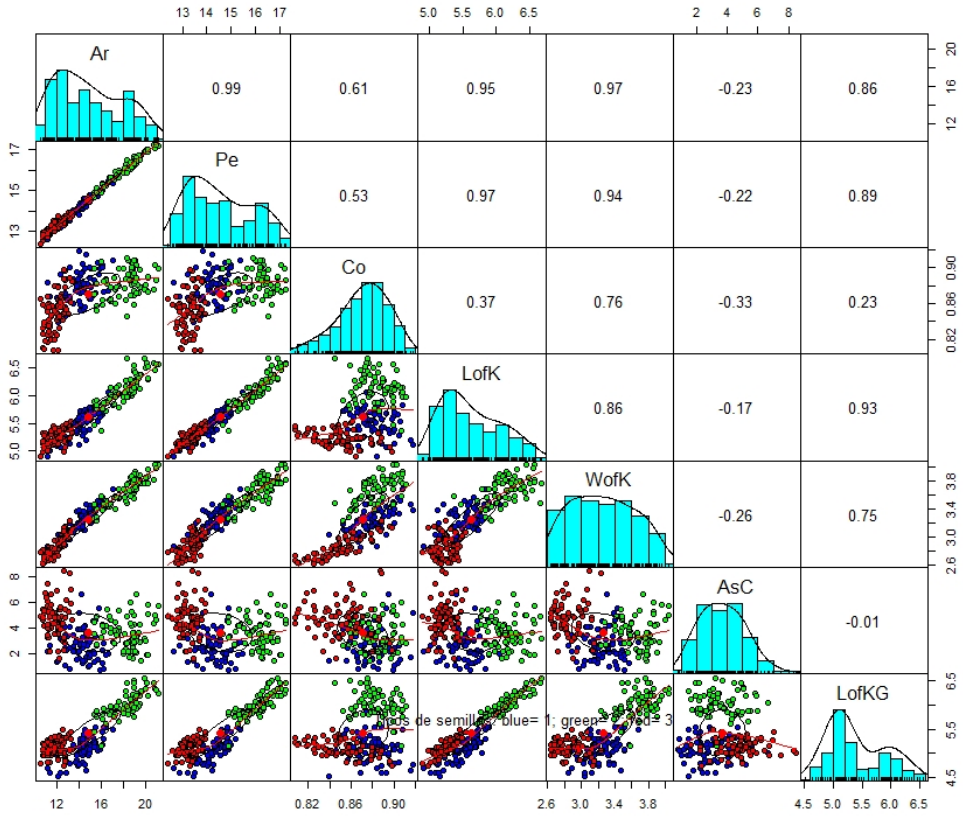

Figure 3:

Si queremos ver los histogramas de cada variable por separado para visualizarlos mejor, escribimos lo siguiente:

```
for (j in 1:7) {
hist(data[,j], xlab=colnames(data)[j],
main=paste("Histograma de",colnames(data[j])),
col="cyan", labels = TRUE)
}
```

Y obtenemos, por ejemplo

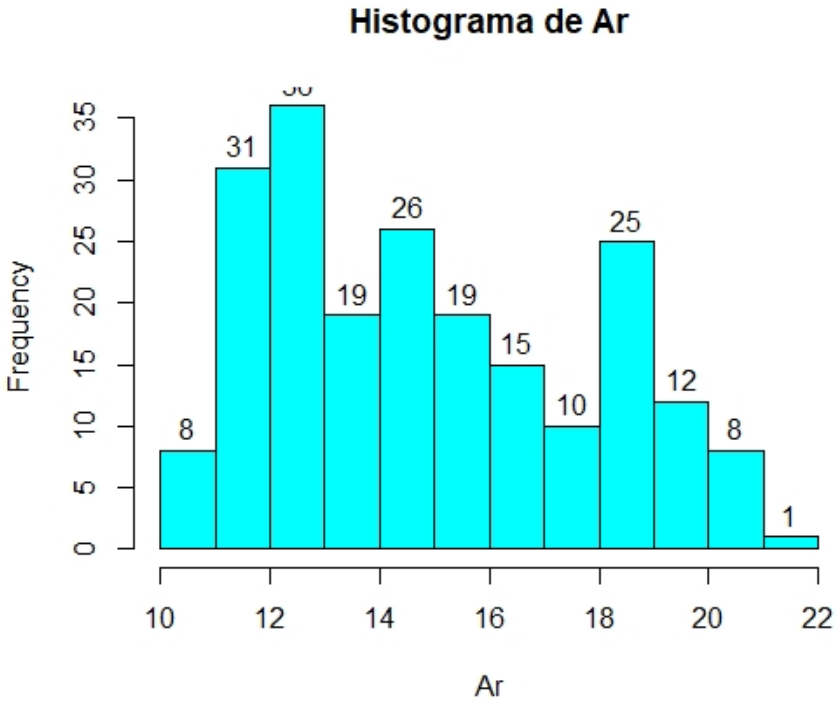

Figure 4: Histograma de la variable Área.

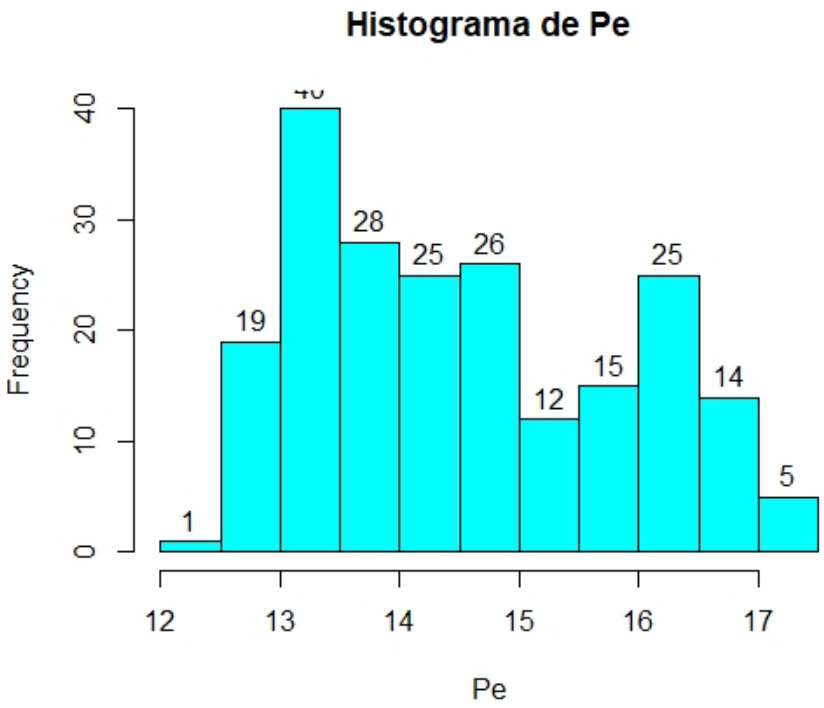

Figure 5: Histograma de la variable Perímetro.

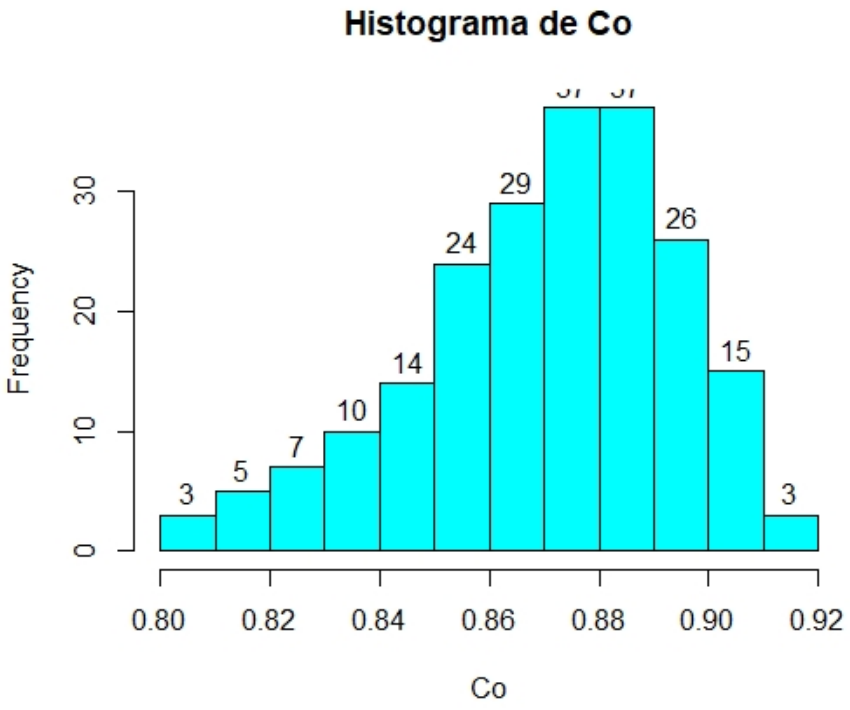

Figure 6: Histograma de la variable Compactibilidad.

Una vez hemos comprobado que efectivamente se verifican las hipótesis necesarias para el análisis discriminante, comencemos a realizarlo. En primer lugar hagamos el análisis discriminante lineal con todos los datos.

```
mod=lda(Ty~., data=data)
```

```
> mod
Call:
lda(Ty ~ ., data = data)

Prior probabilities of groups:
1         2         3
0.3333333 0.3333333 0.3333333

Group means:
  Ar       Pe       Co          LofK     WofK     AsC
1 14.33443 14.29429 0.8800700 5.508057 3.244629 2.667403
2 18.33429 16.13571 0.8835171 6.148029 3.677414 3.644800
3 11.87386 13.24786 0.8494086 5.229514 2.853771 4.788400
  LofKG
1 5.087214
2 6.020600
3 5.116400

Coefficients of linear discriminants:
          LD1          LD2
Ar    -0.42377861   4.1953167
Pe     3.79919995  -8.5057958
Co     5.92772810 -86.9823024
LofK  -5.98819597  -7.8306747
WofK   0.03704822   0.7141043
AsC   -0.04504722   0.3212538
LofKG  3.11807592   6.9138493

Proportion of trace:
LD1    LD2
0.6814 0.3186
```

Si ahora queremos tener una representación gráfica del método LDA

```
ggplotLDA <- function(data){
 if (!is.null(Terms <- data$terms)) {
   datas <- model.frame(data)
   X <- model.matrix(delete.response(Terms), datas)
   f <- model.response(datas)
   xint <- match("(Intercept)", colnames(X), nomatch
                                                = 0L)
   if (xint > 0L)
   X <- X[, -xint, drop = FALSE]
 }
means <- colMeans(data$means)
```

```
X <- scale(X, center = means, scale =
                              FALSE) %*% data$scaling
a <- as.data.frame(cbind(X, labels=as.character(f)))
a <- data.frame(X, labels=as.character(f))
return(a)
}

Graph <- ggplotLDA(mod)
ggplot(Graph, aes(LD1,LD2, color=labels))+geom_point()
+stat_ellipse(aes(x=LD1, y=LD2, fill = labels), alpha
                                        =0.2, geom = "polygon")
```

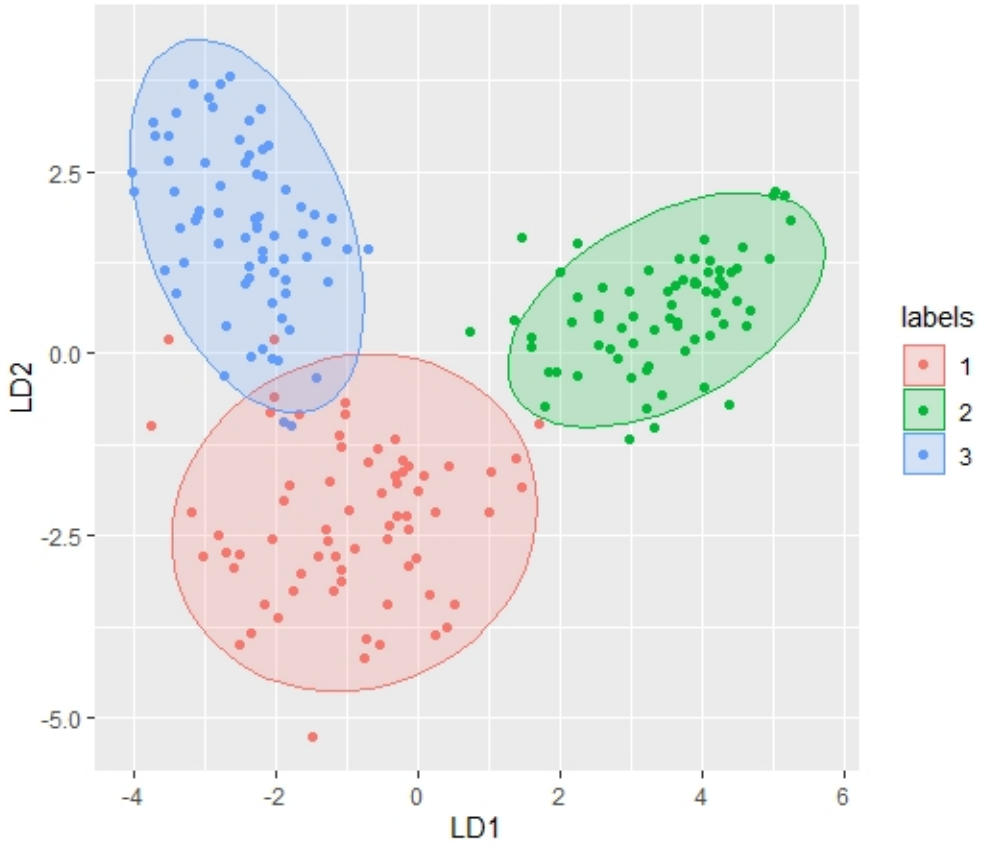

Figure 7:

Ahora veamos como este modelo clasificaría las observaciones a partir de las cuales lo hemos elaborado.

```
p <- predict(mod)
freq <- table(p$class, data$Ty)

> freq
```

```
   1  2  3
1 66  0  3
2  1 70  0
3  3  0 67

> (66+70+67)/210*100
[1] 96.66667

> (1-(66+70+67)/210)*100
[1] 3.333333
```

Con estas últimas entradas de consola, lo que hemos comprobado es qué tanto por ciento de los datos están bien clasificados (la primera entrada) y cuáles mal (la segunda).
También representaremos los histogramas del LDA:

```
ldahist(data=mod.values$x[,1],g=data$Ty)
ldahist(data=mod.values$x[,2],g=data$Ty)
```

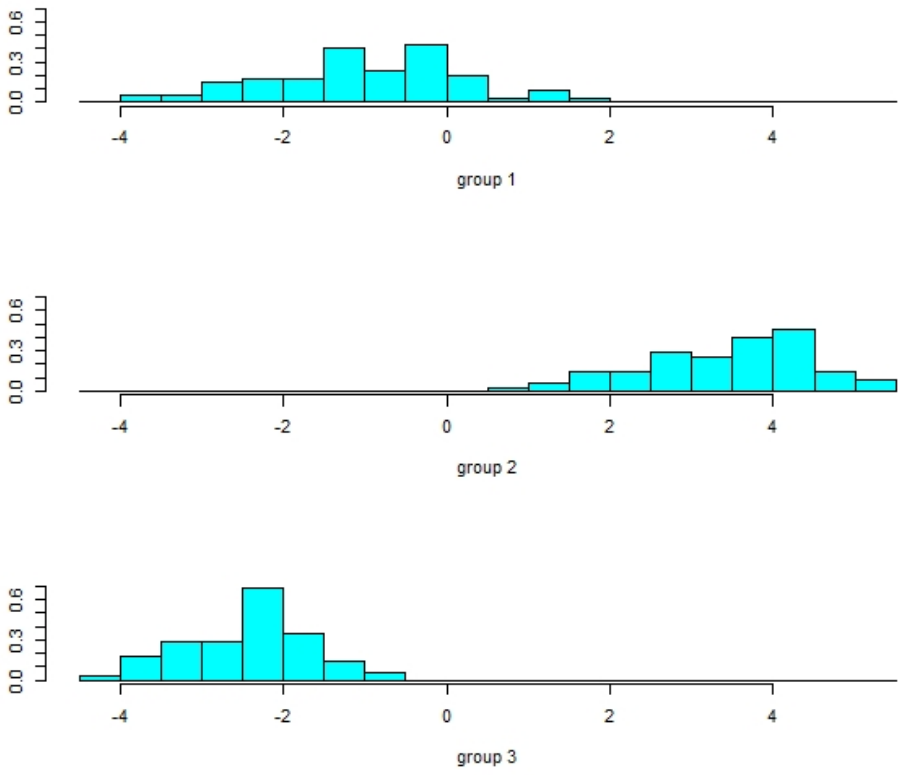

Figure 8: LD1

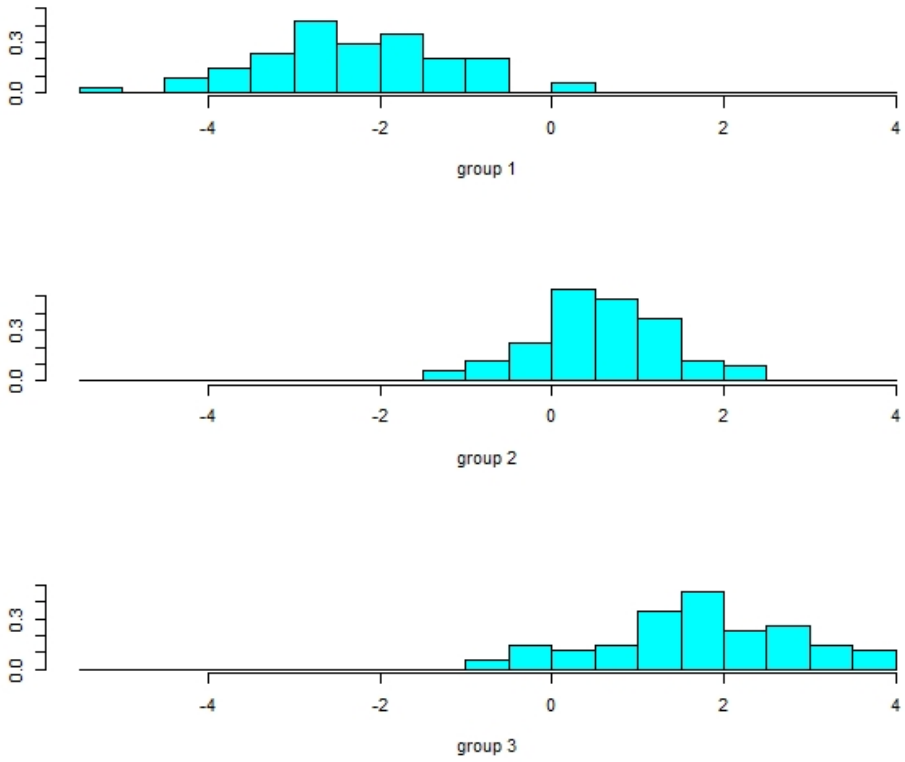

Figure 9: LD2

Con esto completaríamos el primer caso de análisis discriminante. Ahora pasaríamos al segundo caso, en el que haremos un modelo con la mitad de los datos.

Para ello creemos dos particiones de forma aleatoria de los datos: una con la que entrenaremos al modelo y otra con la que veremos como se comporta.

```
set.seed(555)
ind <- sample(2, nrow(data),
              replace = TRUE,
              prob = c(0.6, 0.4))
training <- data[ind==1,]
testing <- data[ind==2,]
```

A continuación, elaboremos de la misma forma que anteriormente el modelo.

```
mod2 <- lda(Ty~., data=training)
```

```
> mod2
Call:
lda(Ty ~ ., data = training)
```

```
Prior probabilities of groups:
1         2         3
0.3983051 0.2881356 0.3135593

Group means:
  Ar       Pe       Co          LofK     WofK     AsC
1 14.39149 14.31830 0.8806085 5.520319 3.250149 2.621749
2 18.25618 16.10882 0.8827500 6.146118 3.664353 3.632794
3 11.96838 13.28162 0.8521135 5.224595 2.871189 4.694432
  LofKG
1 5.111128
2 6.011824
3 5.118811

Coefficients of linear discriminants:
              LD1          LD2
Ar      -0.81924133 -4.22373705
Pe      -2.54531844  7.83518799
Co     -13.33864493 90.92105315
LofK     5.57709087  8.85792896
WofK     4.91253235  0.08498834
AsC     -0.01346304 -0.40528018
LofKG   -2.16132829 -6.51978059

Proportion of trace:
LD1    LD2
0.6274 0.3726
```

Si ahora representamos gráficamente el modelo resultante con la función que hemos utilizado, tan solo tenemos que hacer el siguiente cambio:

```
Graph2 <- ggplotLDA(mod3)
ggplot(Graph2, aes(LD1,LD2, color=labels))
        +geom_point() +
        stat_ellipse(aes(x=LD1, y=LD2, fill = labels),
        alpha=0.2, geom = "polygon")
```

Y si también queremos representar los histogramas de LD1 y LD2 de este nuevo modelo:

```
ldahist(data=mod2.values$x[,1],g=training$Ty)
ldahist(data=mod2.values$x[,2],g=training$Ty)
```

Una vez hechos estos cambios, obtenemos los siguientes resultados:

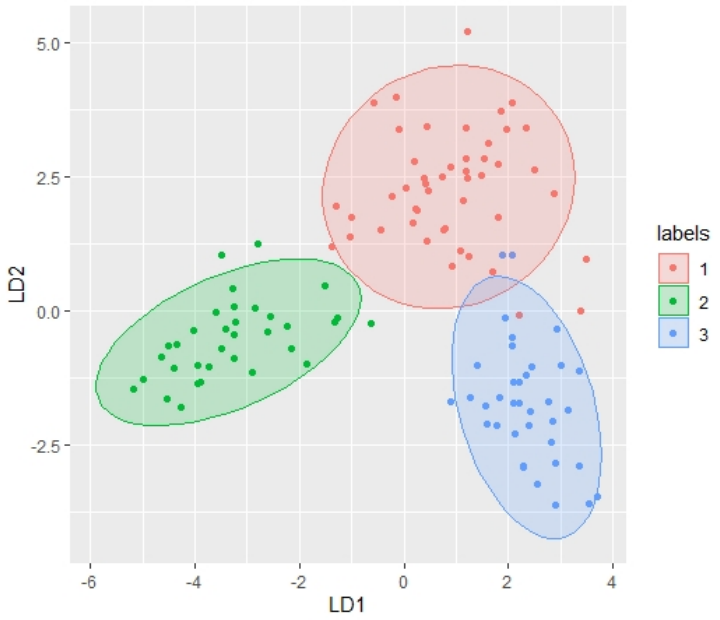

Figure 10:

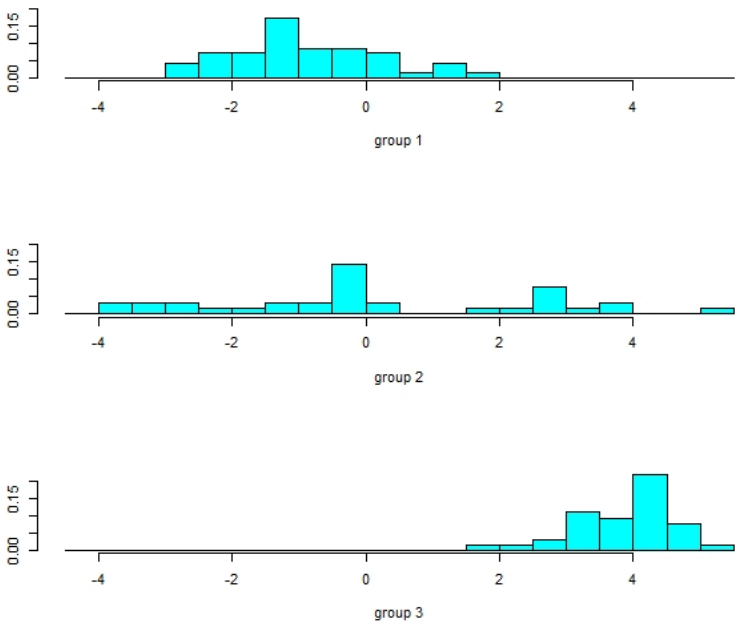

Figure 11: LD1

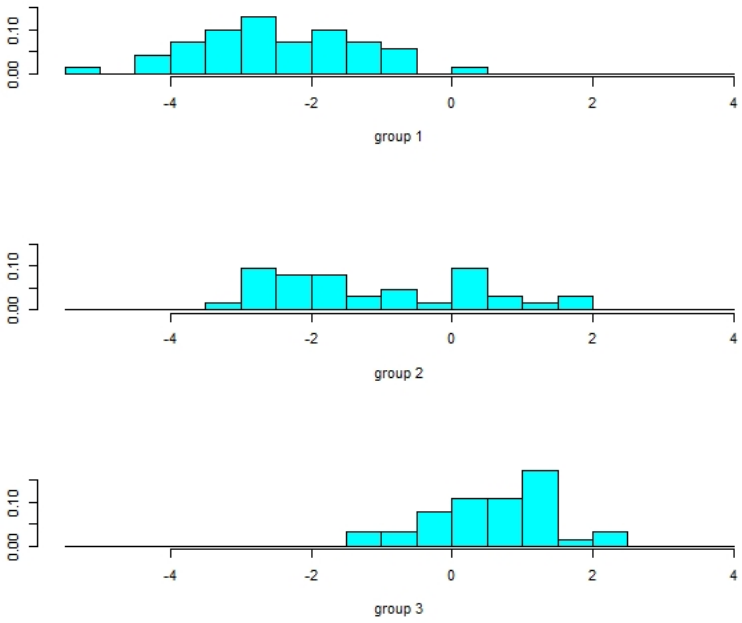

Figure 12: LD2

Y para concluir este segundo caso, veamos como clasifica el modelo los datos con los que ha entrenado y las nuevas observaciones:

```
p2 <- predict(mod2)
freq2 <- table(p2$class, training$Ty)

> freq2

   1  2  3
1 45  0  2
2  0 34  0
3  2  0 35

> (45+34+35)/118*100
[1] 96.61017
> (1-(45+34+35)/118)*100
[1] 3.38983

p3 <- predict(mod2, testing)
freq3 <- table(p3$class, testing$Ty)
```

```
> freq3

   1  2  3
1 23  0  1
2  0 36  0
3  0  0 32

> (23+36+32)/92*100
[1] 98.91304
> (1-(23+36+32)/92)*100
[1] 1.086957
```

Para acabar con el análisis discriminante, realicemos uno cuadrático para ver si este modelo mejorase los anteriores:

```
mod3 = qda(Ty~., data=data)

p4 <- predict(mod3)
freq4 <- table(p4$class, data$Ty)

> freq4

   1  2  3
1 64  1  2
2  2 69  0
3  4  0 68

> (64+69+68)/210*100
[1] 95.71429
> (1-(64+69+68)/210)*100
[1] 4.285714
```

Finalmente, apliquemos un PCA a los datos parar comparar la eficacia de estos dos métodos y poder saber con mayor certeza si disponemos de unos datos que resulten relevantes en el estudio, en el sentido de que las variables que estamos considerando realmente nos ayuden a clasificar los distintos tipos de semillas.

Comenzamos con las siguientes instrucciones para obtener los primeros datos de las componentes principales:

```
> pca <- prcomp(data)
> summary(pca)

Importance of components:
                          PC1    PC2     PC3     PC4     PC5
Standard deviation     3.2997 1.5268 0.63004 0.23346 0.09215
Proportion of Variance 0.7957 0.1704 0.02901 0.00398 0.00062
Cumulative Proportion  0.7957 0.9661 0.99509 0.99908 0.99970
```

	PC6	PC7	PC8
Standard deviation	0.05152	0.03847	0.005037
Proportion of Variance	0.00019	0.00011	0.000000
Cumulative Proportion	0.99989	1.00000	1.000000

Buscamos ahora la manera de decidir gráficamente con qué componentes principales nos quedamos. Para ello

```
> screeplot(pca)
```

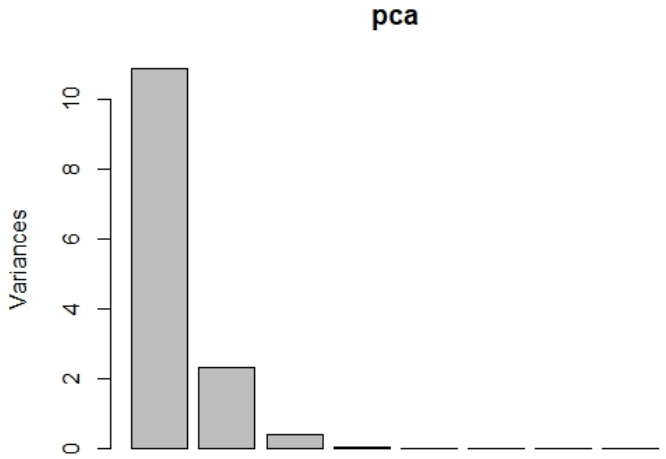

Figure 13:

Con las tablas anteriores y la gráfica que acabamos de obtener ya tenemos información suficiente para poder decidir con qué componentes principales nos quedamos. Hay distintos criterios para decidir con cuáles nos quedamos:[3]

1. Seleccionar componentes hasta cubrir una proporción determinada de varianza, como el 80% o el 90%. Este criterio es bastante arbitrario, con lo cual debe aplicarse con cuidado.

2. Desechar las componentes asociadas a valores propios inferiores a una cota. En particular, cuando se trabaja con la matriz de correlación, este criterio nos lleva a seleccionar los valores propios mayores que 1. También puede ser una regla arbitraria en determinados casos.

[3]Hay que tener cuidado cuando utilizamos esta técnica y aplicamos estos criterios, ya que datos con mucha variabilidad o atípicos pueden llevarnos a sacar conclusiones erróneas. Una solución es estandarizarlos, lo cual puede llegar a ser la forma más correcta de trabajar con ellos. En este caso las observaciones no nos plantean estos problemas.

3. Realizar un gráfico de λ_i frente a i. Comenzamos seleccionando componentes hasta que las restantes tengan aproximadamente el mismo valor de λ_i.

Si además queremos representar a las componentes principales la nube de puntos, lo único que tenemos que hacer es escribir la siguiente instrucción en R:

```
> autoplot(pca, loadings = TRUE, loadings.label = TRUE)
```

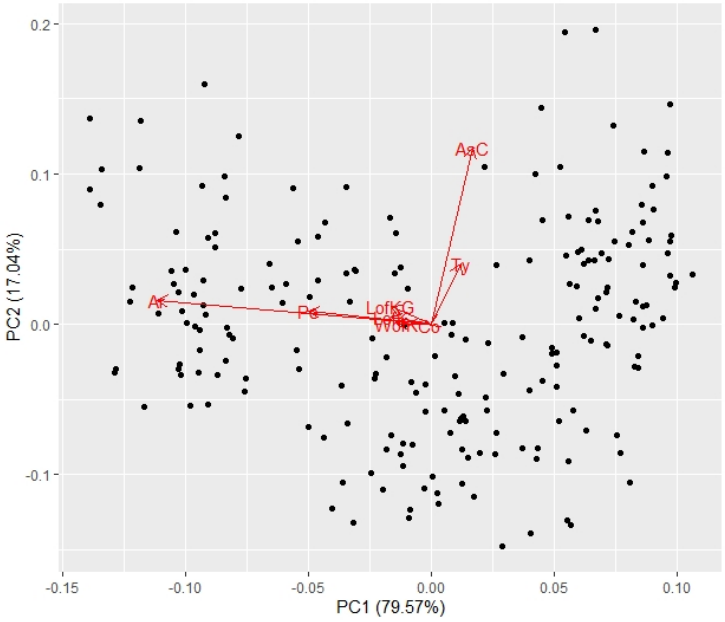

Figure 14:

Por tanto, con la información obtenida, decidimos quedarnos con las dos primeras componentes principales, ya que con ellas tenemos explicada un 96,61% de la varianza.

Para calcular los coeficientes de los vectores de las dos primeras componentes principales sobre las distintas variables (ya que el resto a priori no nos interesan):

```
> data.pca$rotation[,1] #PC1
Ar          Pe          Co        LofK        WofK
-0.4444735  -0.4415715  -0.2770174  -0.4235633  -0.4328187
AsC       LofKG
0.1186925  -0.3871608

> data.pca$rotation[,2] #PC2
Ar          Pe          Co        LofK        WofK
0.02656355  0.08400282  -0.52915125  0.20597518  -0.11668963
AsC       LofKG
0.71688203  0.37719327
```

4 Conclusiones

Como se ha recalcado desde el principio, los métodos se pueden aplicar porque se dan las hipótesis necesarias para los mismos, las cuales se pueden ver tanto en el mapa de calor, en la matriz de correlación, los histogramas y la visualización conjunta de todos ellos. Si observamos la Figura 3, vemos que hay correlación lineal entre la variable correspondiente al área con las variables correspondientes al perímetro, la longitud y el ancho de la semilla y, en menor medida, en la longitud de la ranura de la semilla. Esto parece bastante lógico, ya que el área de cualquier figura está relacionada íntimamente con sus dimensiones.

El primer análisis discriminante nos da unos buenos resultados. Esto se debe a que tenemos bastantes simplificaciones en nuestro modelo, como por ejemplo suponer que las probabilidades a priori son las mismas. Por otra parte, vemos que el primer discriminante clasifica los datos en una proporción del 68,14%, mientras que el segundo discriminante solo lo hace en un 31,86%. En la Figura 7 podemos observar como el método clasifica los datos que le hemos proporcionado. Por otra parte, si observamos los histogramas de cada uno de los discriminantes, también aquí podemos comprobar como el LD1 clasifica mejor los datos, ya que los histogramas de cada uno de los grupos se "pisan" menos que los del LD2. Además, al final del método hemos hecho una simulación de como el modelo es capaz de predecir que tipo de semilla se corresponde con cada una de las observaciones, obteniendo un acierto de un 96,67% y un 3,33% de errores.

En el segundo caso para el análisis discriminante lineal el modelo resulta ligeramente más complejo, pues disponemos de menos observaciones y más dispares, lo cual produce que, por ejemplo, las probabilidades a priori de pertenecer a un tipo de semilla o a otro sean diferentes. En este caso el primer discriminante clasifica en una proporción menor que en la situación anterior. Esto podemos comprobarlo tanto numéricamente en los datos que nos proporciona R al realizar el modelo como en los histogramas de cada uno de los discriminantes. Con respecto a las predicciones, vemos que las que hemos usado para construir el modelo las realiza con un acierto de un 96,61% y un error de 3,39%. Por otro lado, las que no hemos usado las predice con un acierto del 98,91% y un 1,09% de error. Esto último no significa que el modelo sea mejor que el anterior, ya que podría deberse a que estuviésemos utilizando muchas más semillas de un cierto tipo para entrenar el modelo que del resto, lo cual nos llevaría a hacer mejores predicciones de este tipo de semillas. Para poder afirmar que este modelo es mejor que el anterior tendríamos que realizar un gran número de simulaciones con distintas particiones cada una. No obstante lo más probable es que los dos modelos sean más o menos igual de buenos, ya que no estamos utilizando más información para construirlos.

Para acabar con el análisis discriminante, comparamos las versiones lineales con el caso cuadrático. De forma breve, observamos que con este modelo tenemos una tasa de acierto de un 95,71% y de 4,28% de error. Entonces, a la vista de estos resultados, vemos que los modelos lineales explican clasifican mejor los tipos de semillas según las variables de las que disponemos que el cuadrático, ya que estos tienen un mayor porcentaje de acierto. Además, como estas tasas de

acierto son considerablemente elevadas y a falta de más información, podemos admitir que el tipo de semilla de un grano de trigo sigue una relación lineal con las variables que tenemos (ya que el análisis discriminante lineal está muy relacionado con los modelos de regresión lineales).

Por último vemos que el PCA nos corrobora los datos analizados con el LDA y el QDA. Con respecto a este método, observemos que, como decíamos al principio, el área guardaba una correlación lineal frente al perímetro, la longitud y el ancho de la semilla y la longitud de la ranura de la semilla. Esta relación se ve reflejada en el PCA en el hecho de que los coeficientes asociados a los vectores de la primera componente principal para estas variables son prácticamente los mismos.

Concluimos entonces que tanto las observaciones de las que disponemos como las variables medidas en cada una de ellas son muy explicativas para clasificar una nueva semilla en función de estas variables. Por tanto, si usamos cualquiera de estos métodos, tendremos altas probabilidades de realizar clasificaciones acertadas para nuevas observaciones.

Referencias

[1] http://ciberconta.unizar.es/LECCION/discri/INICIO.HTML

[2] Apuntes del profesor Ruiz Galacho de la asignatura *Modelos estadísticos multivariantes*, Universidad de Málaga.

[3] http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/AMult/tema6am.pdf

[4] https://www.youtube.com/watch?v=WUCnHx0QDSI (para el modelo LDA con partición de las observaciones).

[5] https://www.statmethods.net/advstats/discriminant.html

[6] https://www.statmethods.net/advstats/factor.html

[7] https://stackoverflow.com/questions/46810988/r-how-to-visualize-pca

[8] http://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-format-and-visualize-a-correlation-matrix-using-r-software# use-corrplot-function-draw-a-correlogram