UNIVERSIDAD DE MURCIA

PRÁCTICA: ANÁLISIS DISCRIMINANTE.
ESTADÍSTICA MULTIVARIANTE. GRADO MATEMÁTICAS.
RStudio 2021.09.1

Sumario: En esta práctica mostramos cómo clasificar individuos entre diversos grupos a partir de sus medidas (numéricas) en algunas variables. Para ello necesitaremos disponer de una muestra de las variables en estudio en individuos de cada grupo (al menos dos individuos por cada grupo).

1. Estudio inicial de los datos.

Siempre que se aplique una técnica de análisis multivariante es conveniente hacer un análisis inicial de los datos. En este caso trataremos de estudiar las diferencias de las variables en cada uno de los grupos para dilucidar si serán de utilidad a la hora de clasificar a los individuos de los distintos grupos.

Comenzaremos leyendo los datos del objeto *d* del fichero `escarabajos.rda` (Aula virtual). Para leer este archivo debemos teclear

```
load('g:/.../escarabajos.rda')
```

indicando la ruta completa en dónde se encuentra el archivo o cambiando el directorio de trabajo[1]. Para ver los datos basta teclear `d` o `View(d)`. Comprobaremos así que dicho fichero contiene una muestra de 40 escarabajos de dos especies diferentes (Haltica Oleracea y Haltica Carduorum) a los que se les han medido 4 variables: distancia desde el tórax al surco transversal X_1 (micras), longitud X_2 (0.01mm.), longitud de la base de las antenas secundarias X_3 y terciarias X_4 (en micras). La variable código indica la especie a la que pertenece cada individuo (HO=1, HC=2). Puede observarse que hay un escarabajo (40) del que se desconoce la especie lo que en R se escribe como `NA` (Not Available).

Podemos comenzar estudiando las variables por separado. Si queremos ver solo los datos de la variable `surco` haremos: `d$surco` o `d[,1]`. Por ejemplo, para estudiar esta variable, podemos comenzar calculando sus estadísticos básicos (medias, cuartiles y valores extremos) en cada grupo haciendo:

```
tapply(d$surco,d$especie,summary)
```

[1]Para cambiar el directorio de trabajo de RStudio elegir Session>Set Working Directory.

De esta forma observamos que la media de la variable `surco` es más grande en la especie HO (194.5) que en la HC (179.6) y que su valor en el escarabajo 40 (182.2) está más cerca de la media de la especie HC. También podemos representarla gráficamente tecleando:

```
plot(d$surco,d$codigo)
```

Si queremos que aparezca el escarabajo 40 debemos añadir:

```
text(d$surco[40],1.5,labels='e40')
```

obteniendo el gráfico de la Figura 1, izquierda. En esta gráfica podemos observar que la variable `surco` parece un poco mayor en el grupo 1 (HO) pero que no discrimina (separa) bien a los grupos. Con esta variable no es sencillo clasificar al escarabajo 40 pero si tenemos que elegir un grupo lo incluiríamos en el grupo 2 (HC) ya que está más cerca de su media. Se obtiene una gráfica similar (derecha) haciendo `plot(d$surco,d$especie)`. En este caso, R etiqueta los datos por orden alfabético (ASCII) con *=1, HC=2 y HO=3. Estudie las restantes variables. ¿Cuál es la que mejor discrimina? ¿Dónde clasifica al escarabajo 40?

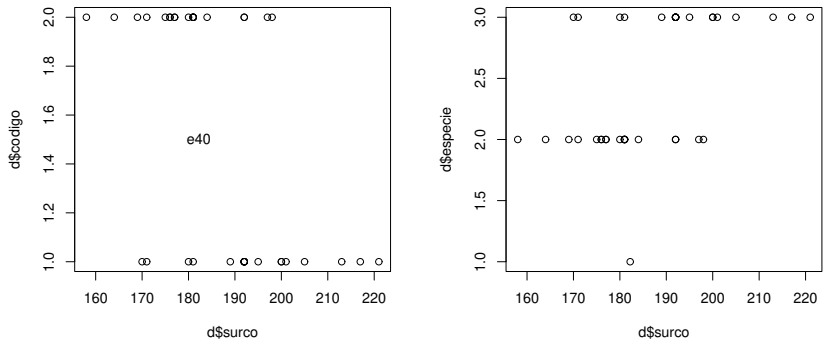

Figura 1: Gráficos de la variable `surco` por grupos.

Los gráficos caja-bigote por grupos para `surco` se pueden hacer con:

```
boxplot(d$surco~ d$especie)
```

(el símbolo ~ se puede escribir pulsando simultáneamente Alt y 126) obteniendo el gráfico de la Figura 2. En este gráfico apreciamos que si usáramos solo la variable `surco` para clasificar, como las cajas no se solapan, más del 75 %

de los individuos se clasificarían bien. También observamos que el escarabajo 40 estaría en la caja de la especie HC (por poco) pero que no sería un valor atípico en la HO. Estudie las restantes variables.

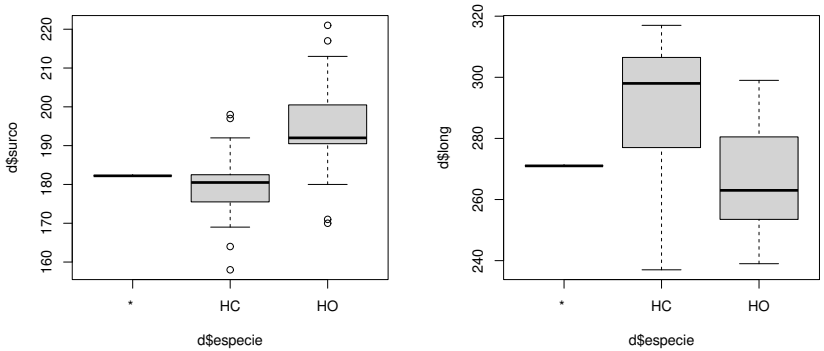

Figura 2: Gráficos caja-bigote de las variables `surco` y `long` por grupos.

En segundo lugar podemos estudiar las variables por parejas. Por ejemplo, para analizar `surco` y `long`, podemos hacer:

```
plot(d$surco,d$long,pch=as.integer(d$especie))
legend('bottomright',legend=c('e40','HC','HO'),pch=1:3)
```

obteniendo el gráfico de la Figura 3 (izquierda) en el que se observa que, con estas dos variables, los dos grupos están bastante separados, pero que el escarabajo 40 estaría entre ambos grupos por lo que no es sencillo clasificarlo. Estudie las otras dos variables. Se obtiene un gráfico similar haciendo:

```
plot(d$surco, d$long)
text(d$surco,d$long,d$especie,cex=0.7,pos=4,col='red')
```

(`cex` indica el tamaño y `pos` la posición de la etiqueta).

Finalmente, podemos hacer un gráfico similar al de la Figura 3 pero usando las dos primeras componentes principales (ver práctica 3) que contienen información sobre todas las variables. Recordemos que las componentes principales (basadas en la matriz de correlaciones) se calculan con:

```
pca<-princomp(d[,1:4],cor=TRUE)
```

y se pueden representar las dos primeras componentes por grupos con:

```
biplot(pca,pc.biplot=TRUE,xlabs=d$especie)
```

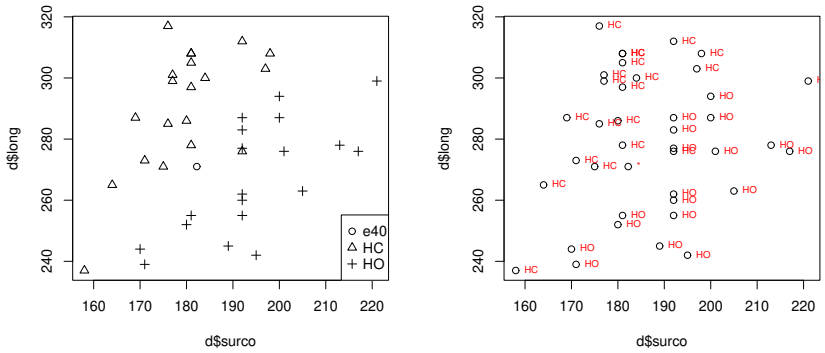

Figura 3: Gráfico conjunto de las variables `surco` y `long` por grupos.

El resultado puede verse en la Figura 4. En este gráfico también se aprecia que los grupos se pueden separar bastante bien. Señalar no obstante que las dos primeras componentes principales no son necesariamente las mejores variables para clasificar a estos individuos (como veremos en las secciones siguientes).

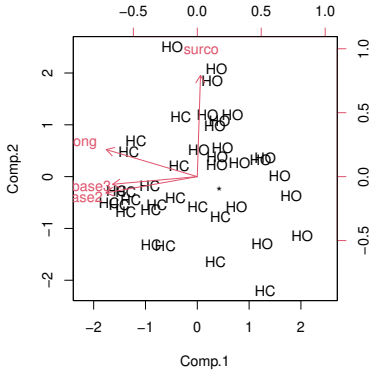

Figura 4: Gráfico de las dos primeras componentes principales por grupos.

2. Análisis Discriminante Lineal (LDA).

Para calcular la función discriminate lineal (FDL) de Fisher para distinguir entre dos grupos debemos suponer que sus matrices de covarianzas (teóricas) son iguales. Entonces la FDL valdrá $L = L(Z) = \mathbf{a}'Z$ donde $(Z_1, \ldots, Z_k)'$ son las medidas del individuo a clasificar y los coeficientes se calculan como

$$\mathbf{a}' = \lambda(\mu_X - \mu_Y)'V^{-1}, \tag{1}$$

donde λ es un número real cualquiera distinto de cero, V es la matriz de covarianzas común y μ_X y μ_Y son los vectores de medias en cada grupo de las variables usadas para clasificar. En la práctica estas medias teóricas se sustituyen por sus estimaciones $\overline{X}$ e $\overline{Y}$ y V se estima mediante:

$$S = \frac{1}{n_1 + n_2 - 2}[(n_1 - 1)S_1 + (n_2 - 1)S_2]$$

siendo n_1 y n_2 los tamaños muestrales de cada grupo y S_1 y S_2 las matrices de cuasicovarianzas muestrales de cada grupo.

Para calcular (estimar) $\mathbf{a}$ en R debemos cargar primero el "paquete" denominado **MASS** tecleando `library('MASS')`. Una vez cargado, debemos hacer:

```
LDA<-lda(d[1:39,1:4],d[1:39,6],prior=c(0.5,0.5))
```

Tecleando `LDA` comprobamos que las probabilidades de pertenencia a priori asignadas a cada grupo valen 0.5 (si no se especifica nada se computan como si los individuos fuesen una muestra, es decir, como $19/39 = 0.4871795$ (HO) y $20/39 = 0.5128205$ (HC)), los vectores de medias de los grupos son:

$$\overline{X} = (194.4737, 267.0526, 137.3684, 185.9474)$$

y

$$\overline{Y} = (179.5500, 290.8000, 157.2000, 209.2500)$$

y que los (unos) coeficientes estimados de la FDL son

$$\mathbf{a} = (-0.09327642, 0.03522706, 0.02875538, 0.03872998).$$

Si queremos guardar estos coeficientes en el objeto a haremos:

```
a<-LDA$scaling
```

Para clasificar a un individuo con medidas z calcularemos su proyección $L(z)$ y las proyecciones de las medias de los grupos $L(\overline{X})$ y $L(\overline{Y})$, clasificándolo en el grupo que tenga la media más cercana a su proyección. La frontera de las regiones de clasificación vendrá dada por la media de las proyecciones de las medias: $K = (L(\overline{X}) + L(\overline{Y}))/2$. Para calcular L podemos definir la función:

```
L<-function(z) sum(a*z)
```

De esta forma, podemos calcular la proyección de la media $L(\overline{X})$ de la especie HO haciendo:

```
mHO<- L(LDA$means[1,])
```

obteniendo $mHO = L(\overline{X}) = 2.419488$. Análogamente, podemos calcular $mHC = L(\overline{Y}) = 6.120841$. De esta forma, haciendo `K<-(mHC+mHO)/2`, obtenemos $K = 4.270164$. Por lo tanto, la regla de decisión óptima según este criterio sería: Si $L(z) > K$ se clasifica como HC (grupo 2) y si no como HO (grupo 1).

Podemos calcular las proyecciones de los 40 escarabajos haciendo:

```
D<-1:40
for (i in 1:40) D[i]<-L(d[i,1:4])
```

Tecleando D comprobamos que para el escarabajo 1 se obtiene $D[1] = 1.253859$ que, como es menor que $K = 4.270164$, nos conduciría a clasificarlo como del grupo HO (correctamente). Análogamente, para el escarabajo 40, obtenemos $D[40] = 3.968782$ que, como es menor que K, nos conduciría a clasificarlo como del grupo HO (con un margen pequeño). Podemos representar estas "puntuaciones discriminates" haciendo:

```
plot(D,d$codigo,ylim=c(1,2.1))
text(D,d$codigo,cex=0.7,pos=1,col='red')
```

Podemos incluir la puntuación del escarabajo 40 y la constante K en el gráfico haciendo:

```
text(D[40],1.5,labels='*')
text(D[40],1.5,labels='e40',cex=0.7, pos=3,col='red')
abline(v=K)
text(K+0.1,1.5,labels='K',cex=0.7,pos=3,col='red')
```

De esta forma se obtiene el gráfico de la Figura 5. En este gráfico se observa que el escarabajo 27 se clasificaría erróneamente y que el 40 se clasificaría en el grupo 1 (HO) pero con un margen pequeño (cerca de K).

Otros autores prefieren calcular las puntuaciones como $D - K$ con lo que la regla de decisión dependerá de si las puntuaciones son positivas o negativas. La puntuación $D - K$ se puede obtener de forma automática haciendo:

```
predict(LDA,d[,1:4])->P
```

Las puntuaciones se obtienen tecleando `P` o `P$x`. Compruebe que coinciden con los valores de `D-K`. Estos valores se pueden representar como en la Figura 5 o en forma de histograma haciendo:

```
ldahist(P$x,g=d$especie)
```

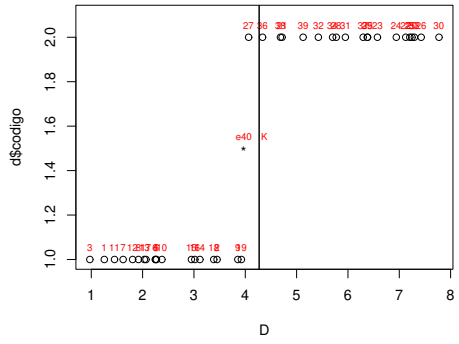

Figura 5: Gráfico de las puntuaciones discriminantes.

Haciendo `P$class` podemos ver en qué grupo se clasifican los 40 escarabajos. Tecleando:

```
P$class==d[,6]->Resumen
```

podemos ver si la clasificación es correcta para los 39 escarabajos de los que se conoce su grupo. Podemos hacer un recuento de estos resultados con:

```
table(P$class,d[,6])
```

Estos valores se pueden resumir con los valores de la Tabla 1. Esta tabla sirve para comprobar si este procedimiento de clasificación es adecuado. En este caso, obtenemos buenos resultados ya que todos los individuos del primer grupo se clasifican correctamente y sólo uno del grupo 2 (el escarabajo 27) se clasifica erróneamente como del grupo 1. Análogamente, comprobamos que todos los individuos clasificados como del grupo 2 se han clasificado correctamente pero que uno clasificado como del grupo 1, en realidad pertenecía al grupo 2 (de nuevo el 27).

Finalmente, haciendo:

```
P$posterior
```

podemos ver las "probabilidades" a posteriori (verosimilitudes normalizadas) de pertenencia a cada grupo bajo normalidad dadas por:

$$\Pr(i|z) = \frac{\pi_i f_i(z)}{\pi_1 f_1(z) + \pi_2 f_2(z)},$$

Tabla 1: Resumen de los resultados de clasificación usando LDA

Grupo verdadero	1 (HO)	2 (HC)	Total
Clasificados en el grupo 1	19	1	20
Clasificados en el grupo 2	0	19	19
Total	19	20	39

donde π_1 y π_2 son las probabilidades a priori (0.5 en este ejemplo) y f_1 y f_2 son las funciones de densidad normales estimadas de cada grupo. Aquí podemos ver que las probabilidades de pertenencia para el escarabajo 40 valen $\Pr(1|z = e40) = 0.7531572$ y $\Pr(2|z = e40) = 0.2468428$, que nos muestran que para un individuo de estas medidas la clasificación no es muy fiable. Evidentemente, los individuos se clasifican usando LDA en el grupo en el que resultan más verosímiles (ambos métodos son equivalentes).

La función `predict` también se puede usar para clasificar a nuevos individuos. Por ejemplo, para clasificar a un escarabajo con medidas $z = (185, 280, 150, 200)$, haremos:

```
z<-c(185,280,150,200)
predict(LDA,z)
```

con lo que se obtiene que z se clasifica en el grupo 2, con una puntuación $D - K = 0.3965766$ y una probabilidad a posteriori de pertenencia al grupo 2 de 0.8127334. Compruebe que la puntuación coincide con $L(z) - K$.

Los valores de la Tabla 1 se pueden usar para estimar las proporciones de acierto en cada caso. Por ejemplo, la probabilidad de acierto global estimada es $38/39 = 0.974359$. Estas estimaciones suelen dar valores ligeramente mayores que los reales ya que al clasificar a un individuo, se ha usado la información proporcionada por el propio individuo. Sin embargo, cuando se clasifica a un individuo nuevo (e40), éste no se usa en el procedimiento de clasificación. Para evitar ésto, podemos usar la técnica denominada *validación cruzada* (*cross validation* o CV) que consiste en que, al clasificar a los individuos de los que se conoce su grupo, el individuo a clasificar no se usa en el procedimiento de clasificación (se tacha). Para hacer esto en R debemos teclear:

```
LDACV<-lda(d[1:39,1:4],d[1:39,6],prior=c(0.5,0.5),CV=TRUE)
table(LDACV$class,d[1:39,6])
LDACV$class==d[1:39,6])
```

De esta forma, podemos comprobar que hay 3 escarabajos del grupo 2 que se clasifican mal (21, 27 y 36) y el resumen correcto de clasificación se-

ría el dado en la Tabla 2. En ella comprobamos, por ejemplo, que la estimación (no sesgada) de la probabilidad global de acierto en este LDA es: $p = (19 + 17)/39 = 0.9230769$ (ligeramente menor que la calculada anteriormente sin CV). Al usar validación cruzada las probabilidades a posteriori de los individuos con grupos conocidos también cambian (ya que no se usan). Por ejemplo, para el escarabajo 21 obtenemos 0.5291374 y 0.4708626, mientras que antes eran 0.1606631 y 0.8393369. La validación cruzada no afecta a la clasificación de los individuos de los que se desconoce el grupo.

Tabla 2: Resumen de los resultados de clasificación usando LDA y validación cruzada.

Grupo verdadero	1 (HO)	2 (HC)	Total
Clasificados en el grupo 1	19	3	22
Clasificados en el grupo 2	0	17	17
Total	19	20	39

Tanto las probabilidades de pertenencia, como las puntuaciones (la constante K) y las clasificaciones finales se verán influenciadas por las probabilidades a priori. Por ejemplo, si no indicamos las probabilidades a priori (es decir, asumimos que éstas se calculen a partir de la muestra), para el escarabajo 40 se obtiene una puntuación en $D - K$ de -0.34883551 y probabilidades a posteriori de $\Pr(1|e40) = 0.7434978$ y $\Pr(2|e40) = 0.2565022$, por lo que se sigue clasificando en el grupo 1. Los aciertos con estas probabilidades a priori son los mismos. Sin embargo, compruebe que con las probabilidades a priori 0.2 y 0.8, existe un escarabajo (19) del grupo 1 que se clasifica en el 2 y que el escarabajo 40 se clasifica en el grupo 2. La clasificación será óptima cuando se usen las probabilidades de pertenencia reales en cada grupo (que suelen ser desconocidas).

Por último señalar que cuando se dispongan de $m > 2$ grupos, necesitaremos una función discriminante para distinguir entre cada pareja de grupos. Practique con los ficheros de datos propuestos al final de la práctica.

3. Análisis Discriminante Cuadrático (QDA).

Cuando las variables usadas para clasificar sean normales (multivariantes) en cada grupo pero sus matrices de convarianzas (teóricas) no sean iguales, el procedimiento óptimo de clasificación consiste en comparar sus funciones de densidad (verosimilitudes o probabilidades a posteriori) estimadas mediante:

$$f_1(z) = c \mid S_1 \mid^{-1/2} \exp(-\frac{1}{2}(z - \overline{X})' S_1^{-1}(z - \overline{X})$$
$$f_2(z) = c \mid S_2 \mid^{-1/2} \exp(-\frac{1}{2}(z - \overline{Y})' S_2^{-1}(z - \overline{Y})$$

En la sección anterior se estimaban usando la estimación de la matriz de varianzas común S. Note que ahora las matrices de covarianzas de cada grupo se estiman usando solo los datos de ese grupo. Esto es equivalente a comparar las funciones discriminantes cuadráticas:

$$QDF_1(z) = (z - \overline{X})' S_1^{-1}(z - \overline{X}) + \log \mid S_1 \mid \quad (2)$$
$$QDF_2(z) = (z - \overline{Y})' S_2^{-1}(z - \overline{Y}) + \log \mid S_2 \mid \quad (3)$$

clasificando a un individuo en donde QDF sea mínima. Note que las funciones QDF son iguales a las distancias de Mahalanobis al cuadrado de cada grupo más una constante que depende del grupo. Cuando los determinantes sean iguales, el método será equivalente al de distancia de Mahalanobis mínima.

Para realizar un QDA en R con los datos de los escarabajos incluidos en el objeto `d` debemos hacer:

```
QDA<-qda(d[1:39, 1:4], d[1:39, 6],prior=c(0.5,0.5))
```

Tecleando QDA comprobamos que en este procedimiento no aparecen los coeficientes de las QDF. Para obtener los coeficientes que convierten a los datos en esféricos y las constantes debemos teclear:

```
QDA$scaling
QDA$ldet
```

respectivamente. Compruebe que con la segunda opción se obtienen $\log \mid S_1 \mid = 19.41635$ y $\log \mid S_2 \mid = 19.56726$. La primera opción nos proporciona matrices triangulares U_i tales que $U_i U_i' = S_i^{-1}$. De esta forma las funciones discriminantes cuadráticas se pueden calcular como:

$$QDF_1(z) = (U_1'z - U_1'\overline{X})'(U_1'z - U_1'\overline{X}) + \log \mid S_1 \mid \quad (4)$$
$$QDF_2(z) = (U_2'z - U_2'\overline{Y})'(U_2'z - U_2'\overline{Y}) + \log \mid S_2 \mid, \quad (5)$$

es decir, la transformación $U_i'z$ convierte a los datos del grupo i en esféricos ya que $Cov(U_i'z) = U_i' S_i U_i$ y como $U_i U_i' = S_i^{-1}$, entonces $S_i = (U_i')^{-1} U_i^{-1}$ y

$$Cov(U_i'z) = U_i' S_i U_i = U_i (U_i')^{-1} U_i^{-1} U_i = I.$$

Para obtener las predicciones basadas en estas funciones o, equivalentemente, en las probabilidades a posteriori, podemos hacer:

```
predict(QDA,d[,1:4])->P
```
Tecleando `P` comprobamos que solo hay un escarabajo mal clasificado (el 27) y que el escarabajo 40 se clasifica en el grupo 1 (como en el LDA). En este caso, las probabilidades de pertenencia valen 0.5817418 y 0.4182582 por lo que esta clasificación no es fiable.

De nuevo podemos obtener una tabla resumen de las clasificaciones con:
```
table(P$class,d$codigo)
```
Para que esta tabla sea más realista debemos usar validación cruzada haciendo:
```
QDACV<-qda(d[1:39,1:4],d[1:39,6],prior=c(0.5,0.5),CV=TRUE)
table(QDACV$class,d[1:39,6])
```
obteniendo los resultados de la Tabla 3. Los resultados son similares a los obtenidos con el LDA (aquí hay un error más) con una probabilidad global de acierto estimada de $p_{QDA} = 35/39 = 0.8974359$.

Tabla 3: Resumen de los resultados de clasificación usando QDA con validación cruzada.

Grupo verdadero	1 (HO)	2 (HC)	Total
Clasificados en el grupo 1	17	2	19
Clasificados en el grupo 2	2	18	20
Total	19	20	39

La función `predict` también se puede usar para clasificar a nuevos individuos. Por ejemplo, para clasificar a un escarabajo con medidas $z = (185, 280, 150, 200)$, haremos:
```
z<-c(185,280,150,200)
predict(QDA,z)
```
con lo que se obtiene que z se clasifica en el grupo 2, con probabilidad a posteriori de pertenencia al grupo 2 de 0.9636754. Esta clasificación sí es fiable (bajo la hipótesis de normalidad).

4. Comprobaciones del modelo.

En primer lugar podemos tener la duda de si es mejor aplicar LDA o QDA. El primer método funciona bien si las matrices de covarianzas teóricas son iguales y el segundo si los datos son normales en cada grupo. Se cumplan o no esas hipótesis, el método de validación cruzada nos proporciona estimaciones de las probabilidades de acierto en cada caso y nos permite la comparación global de las técnicas LDA y QDA. También tenemos la opción de usar ambas técnicas y comprobar si los resultados coinciden.

Si queremos estudiar las hipótesis del LDA, la matriz de cuasicovarianzas del primer grupo se puede calcular con:
```
S1<-cov(d[1:19,1:4])
```
También se pueden separar los datos del grupo 1 con:
```
d1<-d[d$especie=='HO',1:4]
```
y su matriz de cuasicovarianzas se calcula como `cov(d1)`. Análogamente, se calcula la del segundo grupo obteniéndose:

$$S_1 = \begin{pmatrix} 187.596 & 176.863 & 48.371 & 113.582 \\ 176.863 & 345.386 & 75.980 & 118.781 \\ 48.371 & 75.980 & 66.357 & 16.243 \\ 113.582 & 118.781 & 16.243 & 239.942 \end{pmatrix}$$

$$S_2 = \begin{pmatrix} 101.839 & 128.063 & 36.989 & 32.592 \\ 128.063 & 389.011 & 165.358 & 94.368 \\ 36.989 & 165.358 & 167.537 & 66.526 \\ 32.592 & 94.368 & 66.526 & 177.882 \end{pmatrix}$$

De esta forma, comprobamos que las matrices de covarianzas de los grupos son bastante diferentes (no parecen estimaciones de una misma matriz).

Para comprobar que las computaciones de R para los coeficientes del LDA dados en (1) son correctas podemos calcular la estimación de la matriz de covarianzas común V con $S = \frac{1}{n+m-2}[(n-1)S_1 + (m-1)S_2]$. Haciendo: `S<-(18*S1+19*S2)/37`, obtenemos

$$S = \begin{pmatrix} 143.559 & 151.803 & 42.527 & 71.993 \\ 151.803 & 367.788 & 121.877 & 106.245 \\ 42.527 & 121.877 & 118.314 & 42.064 \\ 71.993 & 106.245 & 42.064 & 208.073 \end{pmatrix}.$$

Su inversa se calcula con: `solve(S)->In`. Las medias de los grupos se calculan con:
```
LDA$means[1,]->m1
LDA$means[2,]->m2
```
(o con `colMeans(d1)`) y los coeficientes como
```
(m1-m2)%*%In->a
```
(donde $\% * \%$ denota el producto de matrices en R) obteniendo $\mathbf{a} = (0.345249, -0.1303878, -0.1064338, -0.1433533)$. Para comprobar que son proporcionales a los obtenidos por R haremos:

```
LDA$scaling/t(a)
```
donde `t(a)` es el traspuesto de `a`. Así obtenemos que R usa $\lambda = -0.2701715$.

Si queremos estudiar qué variables influyen más en los procedimientos de clasificación LDA, como las variables originales pueden tener escalas diferentes (como ocurre en nuestro ejemplo), no podemos comparar directamente los coeficientes obtenidos con ellas. Sin embargo, si estandarizamos las variables originales, como éstas tendrán valores similares, los coeficientes obtenidos con ellas en el LDA sí se podrán usar para estudiar la influencia de las variables en la clasificación. Al contrario de lo que ocurría en el PCA, los procedimientos de clasificación LDA y QDA dan el mismo resultado si se usan las variables estandarizadas (no se ven afectados por cambios de escala y/o localización). Para estandarizar los datos haremos:
```
ds<-scale(d[,1:4])
```
y calculando los coeficientes con:
```
lda(ds[1:39,1:4],d[1:39,6],prior=c(0.5,0.5))
```
obtenemos: -1.2937164 (surco), 0.7809833 (long), 0.4182667 (base2) y 0.7084167 (base3). Por lo tanto, la variable que más influye (mejor discrimina) es surco y la que menos base2.

También nos podemos plantear si podemos eliminar alguna variable, cuál sería la más adecuada. Para esto podemos usar los procedimientos de validación cruzada y estudiar qué opción proporciona los mejores resultados teniendo claro que la mejor opción es siempre usarlas todas. Por ejemplo, si eliminamos *surco* haciendo:
```
lda(d[1:39,2:4],d[1:39,6],prior=c(0.5,0.5),CV=TRUE)
```
comprobamos que hay 7 escarabajos que se clasifican mal. Eliminado las otras variables comprobamos que las mejores opciones son eliminar la variable *long* o la variable *base*2 (en ambos casos solo hay 2 escarabajos que se clasifiquen mal). Análogamente, podemos estudiar cuál es la mejor pareja de variables (o la variable) que mejor discriminan. Se puede aplicar un procedimiento similar en el QDA.

También podemos comprobar cómo se calculan las probabilidades a posteriori. Para ello debemos cargar el "paquete" **mvtnorm** y teclear:
```
dmvnorm(d[40,1:4],m1,S)->f1
dmvnorm(d[40,1:4],m2,S)->f2
f1/(f1+f2)
```
De esta forma se obtiene la probabilidad a posteriori del escarabajo 40 en el grupo 1, $\Pr(1|z = e40) = 0.7531572$, en el caso de probabilidades a priori iguales. Para obtener la que se obtiene con las probabilidades a priori proporcionadas por los grupos debemos hacer:
```
19*f1/(19*f1+20*f2)
```
obteniendo $\Pr(1|e40) = 0.743497$ (como en la sección 2). Compruebe usando un procedimiento análogo (pero sustituyendo S por $S1$ y $S2$) las probabilidades a posteriori calculadas en el QDA.

Por último, señalar que para que estas "probabilidades" (verosimilitudes) sean correctas, las variables deben ser normales en cada grupo. Esta hipótesis también se usa en el QDA. Para hacer un test de normalidad multivariante (Shapiro-Wilk) debemos cargar el paquete: **mvnormtest** y hacer:
```
mshapiro.test(t(d[1:19,1:4]))
```
obteniendo un p-valor de 0.2013, por lo que el primer grupo pasaría el test de normalidad. Análogamente, para el segundo se obtiene un p-valor de 0.05769, que nos conduciría a aceptar la normalidad con $\alpha = 0.05$ por muy poco. Esto se puede deber al escarabajo 27 que, como hemos visto durante toda la práctica tiene unas medidas raras para ser del grupo 2. Los datos para este grupo se pueden ver haciendo: `plot(d[20:39,1:4])`.

Cuando en un LDA hay más de dos grupos, algunos autores prefieren calcular las funciones discriminantes lineales por grupos dadas por:

$$L_i(z) = z'S^{-1}m_i - m_i'S^{-1}m_i/2,$$

donde S es la matriz de covarianzas ponderada (calculada anteriormente) y m_i son las medias muestrales de los grupos. Para calcularlas en R haremos:
```
solve(S)%*%m1
solve(S)%*%m2
-0.5*t(m1)%*%solve(S)%*%m1
-0.5*t(m2)%*%solve(S)%*%m2
```
obteniendo:

$$L_1(z) = 0.9557217z_1 - 0.0208622z_2 + 0.6842504z_3 + 0.4353125z_4 - 177.6155$$

$$L_2(z) = 0.6104728z_1 + 0.1095255z_2 + 0.7906842z_3 + 0.5786658z_4 - 193.4209$$

Los individuos se clasificarán en el grupo con valor máximo de estas funciones. Este método es equivalente al de máxima verosimilitud (probabilidad a posteriori) con matrices de covarianzas iguales por lo que se obtendrán los mismos resultados que en la sección 2. También es equivalente a usar las funciones discriminantes de Fisher paso a paso. De hecho, éstas se obtienen restando las

funciones discriminates de los grupos, es decir: $L_1(z) - L_2(z) = a'z - K$. Por ejemplo, para el escarabajo 40 obtenemos:

$$L_1(182.22, 271.01, 140.99, 190.15) = 170.1294$$

$$L_2(182.22, 271.01, 140.99, 190.15) = 169.0138$$

por lo que se clasificaría en el grupo 1 (HO).

De forma análoga, en el QDA se pueden calcular las funciones cuadráticas definidas por (3). En este caso, los individuos se incluyen en el grupo con el valor mínimo para esas funciones. Esto es equivalente a usar el método se máxima verosimilitud (probabilidad a posteriori) con matrices de covarianzas distintas por lo que se obtendrán las mismas clasificaciones que en la sección 3. Para el escarabajo 40 se obtiene:

$$QDF_1(z) = 22.76789$$

$$QDF_2(z) = 23.42774$$

por lo que se clasificaría (de nuevo) en el grupo 1.

De forma similar se pueden calcular las distancias de Mahalanobis (al cuadrado) dadas en el QDA por:

$$D_1^2(z) = (z - \overline{X})' S_1^{-1} (z - \overline{X})$$
$$D_2^2(z) = (z - \overline{Y})' S_2^{-1} (z - \overline{Y})$$

obteniendo para el escarabajo 40: $D_1^2(z) = 3.351539$ y $D_2^2(z) = 3.860477$ por lo que se clasificaría en el grupo 1 (en el más cercano). También se puede usar el comando `mahalanobis(x,y,V)`. Los métodos de clasificación son equivalentes si los determinantes de las matrices de covarianzas de los grupos son iguales. Por lo tanto, se pueden obtener resultados diferentes de los obtenidos en la sección 3. Estas distancias también se pueden calcular usando las transformaciones proporcionadas por las matrices U_i incluidas en `QDA$scaling`. Por ejemplo, los transformados en el grupo 2 de la media del grupo 2 y el escarabajo 40 son

$$U_2'\overline{Y} = (-17.792105, 4.306052, -6.051622, 9.862908)$$

$$U_2'z = (-18.056683, 2.772973, -5.519009, 8.787518),$$

respectivamente, y su distancia Euclídea al cuadrado es 3.860477.

Para calcular estas distancias en el LDA debemos reemplazar S_1 y S_2 por S obteniendo para el escarabajo 40: $D_1^2(z) = 2.801345$ y $D_2^2(z) = 5.03239$ por lo que se clasificaría en el grupo 1 (en el más cercano). En este caso, los métodos son equivalentes por lo que se obtendrán los mismos resultados que en la sección 2 (con probabilidades a priori iguales). Cuando hay más de dos grupos, `LDA$scaling` proporciona la matriz U tal que $UU' = S^{-1}$, es decir, la transformación $U'z$ es esférica en todos los grupos (ver sección 3). Con `predict(LDA)` podemos ver los transformados de los individuos que se pueden representar con `plot`. Si sólo hay dos grupos, los transformados esféricos se proyectan sobre la recta formada por los transformados de las dos medias (función de Fisher), etc. Si hay tres grupos, los transformados esféricos se proyectan sobre el plano formado por los transformados de las tres medias, etc. En estos transformados, las regiones de clasificación vendrán determinadas por las mediatrices ya que se usa la distancia Euclídea.

5. Ejercicios.

1. Aplicar un DA a los datos de las columnas 5-10 del objeto d del fichero `bears.rda`[2]) para estudiar si esas medidas sirven para determinar el sexo del oso. Las variables son: Head.L= longitud de la cabeza (pulgadas), Head.W=anchura de la cabeza (pulgadas), Neck.G=perímetro cuello (pulgadas), Length=altura (pulgadas), Chest.G=perímetro pecho (pulgadas), Weight=peso (libras). Fuente: Minitab15.

2. Aplicar un DA a los datos del fichero `pottery` del paquete MVA[3] que contiene resultados de análisis químicos de cerámica británica de la época romana de diversas regiones y hornos (kiln). La región 1 corresponde al horno 1, la región 2 a los hornos 2 y 3, y la región 3 a los hornos 4 y 5. ¿Podemos usar estas medidas para determinar el origen de la cerámica?

3. Aplicar un DA a los datos del objeto d del fichero `pulgas.rda`[2].

4. Aplicar un DA a los datos `wine.R` que se obtienen al teclear en R: `source('F:/Multivariante/wine.R')` o

```
wine<-read.table('http://archive.ics.uci.edu
/ml/machine-learning-databases/wine/wine.data',sep=',')
```

Los datos contienen concentraciones de 13 diferentes análisis químicos en vinos de la misma región de Italia producidos por tres diferentes bodegas (primera columna). Fuente: http://little-book-of-r-for-multivariate-analysis.readthedocs.org/en/latest/src/multivariateanalysis.html

[2]Para este tipo de archivos teclear `load('f:/name.rda')` indicando la ruta completa en donde se encuentra el archivo y reemplazando `name` por el nombre del archivo.

[3]Para leer este conjunto de datos hay que instalar el paquete MVA pinchando en el menú: Instalar>Paquete seleccionando MVA y tecleando en R: `library('MVA')`