Analizando datos

Alfredo Cortell-Nicolau

```
##
## Attaching package: 'Momocs'
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

Hoy en día, el campo de la Morfometría Geométrica está relativamente consolidado, y la mayoría de técnicas que se utilizan para el análisis de los datos que produce se han demostrado como solventes y útiles. Ahora bien, esta misma consolidación puede, en ocasiones, producir protocolos cerrados que limiten el alcance de las respuestas posibles que se pueden alcanzar de acuerdo con las soluciones metodológicas prestablecidas. Dicho esto, es imposible abarcar en un único workshop toda la serie de técnicas que podrían aplicarse, y la serie de cuestiones que podría abordarse, con la calidad de los datos producidos por la Morfometría Geométrica. Por lo tanto, en esta última sesión se abordará principalmente el concepto de la Morfometría Geométrica y las técnicas más comunes que suelen aplicarse. Aún así, se ofrecerá un ejemplo de técnicas adicionales que podrían aplicarse, con la intención de que consideréis posibles nuevas cuestiones e investiguéis cuales serían las formas más útiles de responder a esas cuestiones.

Datos métricos

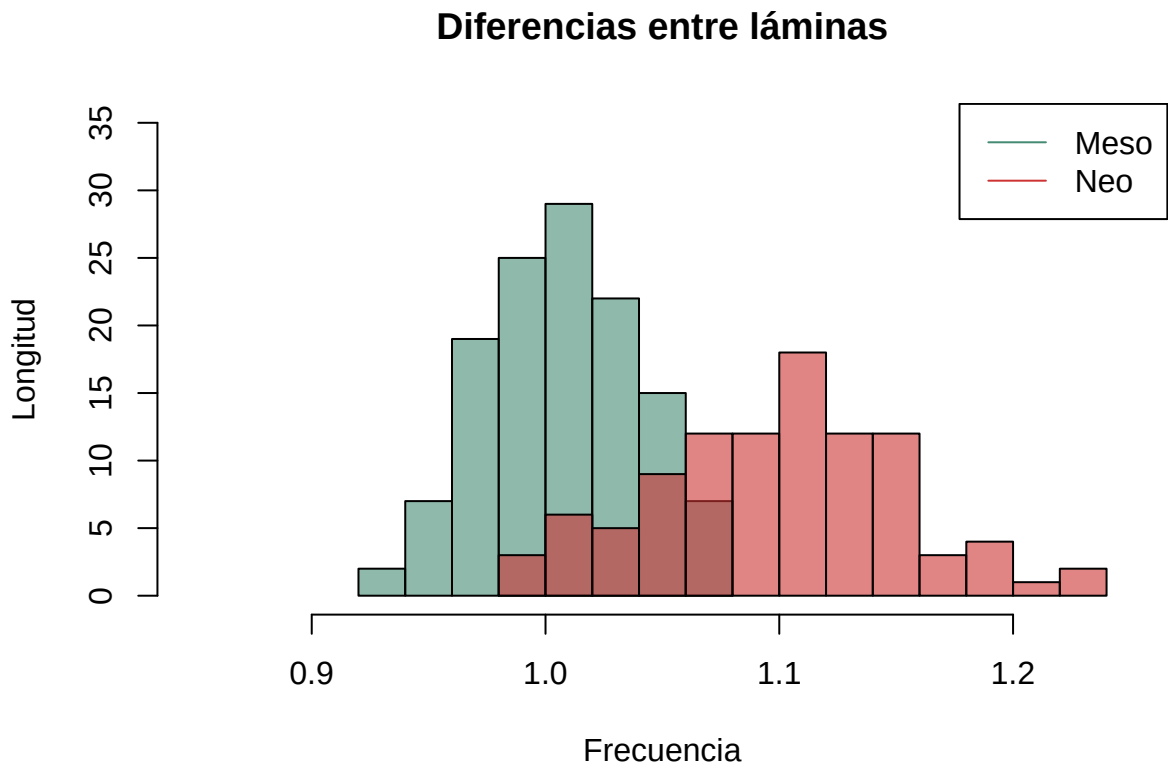
En realidad, y como se ha comentado al principio, es solo una forma muy sofisticada de entender la morfometría (es decir, las medidas) de un objeto o artefacto. Pensándolo bien, esto lleva haciéndose en Arqueología desde diferentes perspectivas prácticamente desde el inicio de la disciplina. Las arqueólogas y arqueólogos siempre han intentado saber si podían captar diferencias en poblaciones de acuerdo, por ejemplo, con la amplitud de las láminas de sílex, con la circunferencia de las bocas de vasos cerámicos o con la longitud de húmeros de animales (aplíquese el concepto a cualquier contexto donde sea relevante). Muchos de estos análisis se hacían, al principio, de forma visual y las conclusiones eran fruto de la experiencia personal de la investigadora o investigador que estudiara las colecciones específicas. Por suerte, y desde hace algún tiempo, la inclusión de soluciones estadísticas para captar estas diferencias, al menos en sus aspectos más básicos, no es totalmente ajena a la arqueología. Este workshop ni intenta ni puede ser una clase de estadística, pero sí podría ser pertinente introducir algunos conceptos básicos que puedan ser útiles en un futuro, y que son relevantes para entender cómo funciona la Morfometría Geométrica. Por ejemplo, consideremos el caso donde hemos registrado la longitud de distintas láminas mesolíticas y neolíticas. Esto nos ha dejado con el siguiente data frame (*los datos son simulados*).

```
laminas <- read.csv("./Datos/blades.csv")[,-1]
```

Podemos querer ver cómo se distribuyen las longitudes de cada uno de los dos grupos del siguietne modo

```
## Plot datos
hist(laminas$Length[laminas$Affiliation == "HG"], xlim = c(0.85,1.25),
     ylim = c(0,35), breaks = 10, col = adjustcolor("aquamarine4", alpha.f = 0.6),
```

```
     ylab = "Longitud", xlab = "Frecuencia", main = "Diferencias entre láminas")
hist(laminas$Length[laminas$Affiliation == "F"], add = TRUE, breaks = 10,
     col = adjustcolor("brown3", alpha.f = 0.6))
legend("topright", legend = c("Meso","Neo"), lty = c(1,1),
       col =c("aquamarine4","brown3"))
```



Solo esta visión debería darnos una idea de que, en efecto, existen diferencias entre las dos poblaciones pero ¿Cómo podemos llegar más lejos? Hay toda una serie de test estadísticos que se destinan a resolver estar preguntas, y cada uno viene con una especie de 'condiciones de uso'. No voy a entrar en cada uno de ellos, pero si queréis explorar este camino, es algo que deberíais de tener en cuenta. En este caso, vamos a aplicar un test-t, que se utiliza normalmente para establecer si dos grupos de población pertenecen a la misma distribución (*e.g* son la misma población) de acuerdo con la asunción de que los datos son numéricos, provienen de una distribución normal y son homocedásticos (tienen igualdad de varianzas). Aplicar el test en R es muy sencillo.

```
## Preparamos los datos
HG <- laminas$Length[laminas$Affiliation == "HG"] # caza-recolectores
Fa <- laminas$Length[laminas$Affiliation == "F"] # agricultores

t.test(HG,Fa)
```

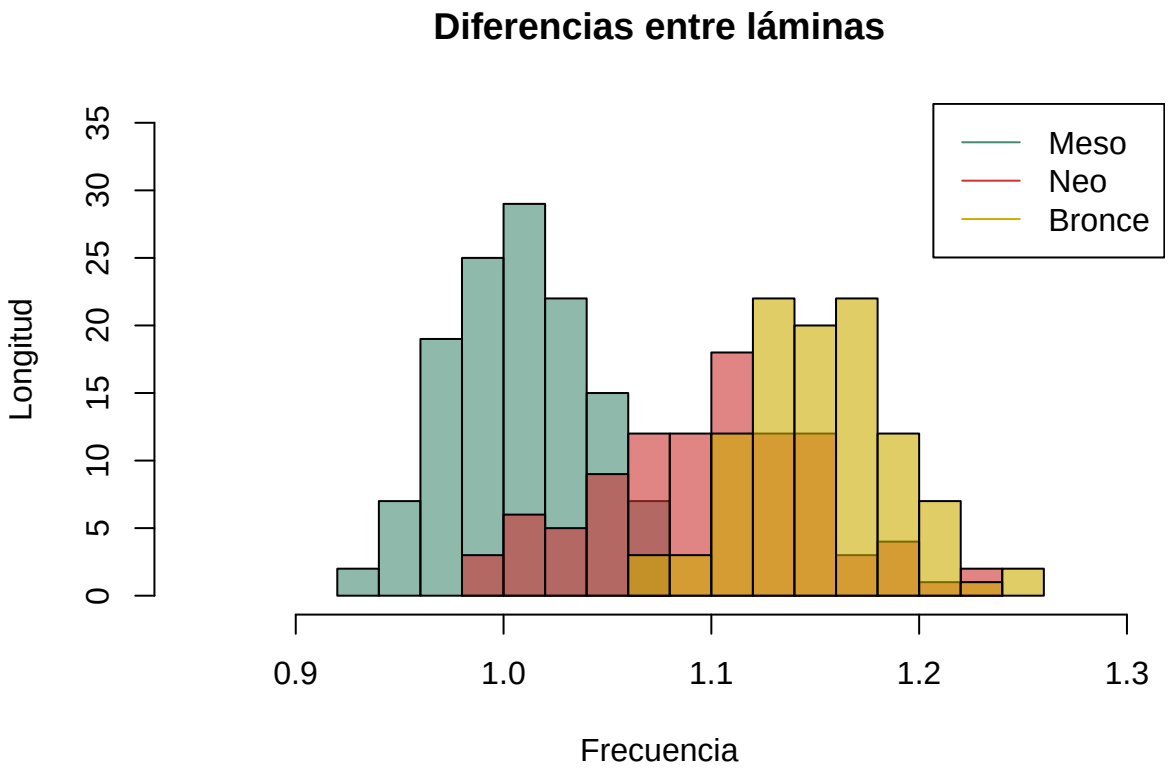
```
##
##  Welch Two Sample t-test
##
## data:  HG and Fa
## t = -15.364, df = 156.49, p-value < 2.2e-16
```

```
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.10395922 -0.08027359
## sample estimates:
## mean of x mean of y
##  1.006827  1.098944
```

El resultado del test confirma dos cosas. La primera es que no existe igualdad de varianzas (y de ahí el uso del test de Welch) y la segunda es que, efectivamente, las dos poblaciones son distintas.

Pero puede ser el caso de que dispongamos más información. Por ejemplo, puede que también, y afortunadamente, se haya registrado la longitud de láminas de la Edad del Bronce, y que también nos interese meter estas en la comparación. De nuevo, la primero cosa que debemos hacer (ahora y siempre), es explorar nuestros datos visualmente.

```
## Plot datos
hist(laminas$Length[laminas$Affiliation == "HG"], xlim = c(0.85,1.3),
     ylim = c(0,35), breaks = 10, col = adjustcolor("aquamarine4", alpha.f = 0.6),
     ylab = "Longitud", xlab = "Frecuencia", main = "Diferencias entre láminas")
hist(laminas$Length[laminas$Affiliation == "F"], add = TRUE, breaks = 10,
     col = adjustcolor("brown3", alpha.f = 0.6))
hist(laminas$Length[laminas$Affiliation == "BR"], add = TRUE, breaks = 10,
     col = adjustcolor("gold3", alpha.f = 0.6))
legend("topright", legend = c("Meso","Neo","Bronce"), lty = c(1,1,1),
       col =c("aquamarine4","brown3","gold3"))
```



En este caso, y porque tenemos tres poblaciones, en vez de dos grupos, será más apropiado utilizar un análisis de la varianza, también conocido como ANOVA, asumiendo que nuestras poblaciones están normalmente distribuidas, son independientes y son homocedásticas.

Existen tests para comprobar cada una de estas asunciones, pero no entraremos en ellos, al no ser este un workshop de estadística.

```
## Test ANOVA
aov.res <- aov(Length ~ Affiliation, laminas)
summary(aov.res)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Affiliation   2  1.258  0.6289   379.7 <2e-16 ***
## Residuals   326  0.540  0.0017
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Estos resultados nos indican que al menos una de las tres distribuciones proviene de una población diferente. Para saber exactamente cual, habría que hacer test adicionales, los conocidos como test *post-hoc* como, por ejemplo, el test de Tukey.

Hay una infinidad más de técnicas y enfoques que pueden adaptarse a casi cualquier pregunta y datos que se os puedan ocurrir, pero por ahora lo dejamos aquí para que tengáis un idea básica.

Técnicas comunes en GM

Y ¿Cómo se relaciona todo esto con la Morfometría Geométrica? Pues, como se comentaba anteriormente, la Morfometría Geométrica es una versión más compleja de lo anterior. En primer lugar ¿Qué medidas tenemos cuando realizamos una extracción de datos con morfometría geométrica?

En nuestro caso anterior teníamos una única medida por observación o artefacto (la longitud de la lámina), pero en el caso de la Morfometría Geométrica tendremos varias medidas por observación, y esto variará dependiendo de si utilizamos landmarks o outlines. En el caso de los landmarks, el número de medidas será igual al número de landmarks $m = lm$, donde m es el número de medidas y lm es el número de landmarks. En cambio, en el caso de los outlines, el número de medidas es $m = 4 \times h$, donde h es el número de armónicos utilizados. Se deduce de esto que la información que recuperamos es significativamente más detallada que en el caso anterior, pero también, y por lo mismo, es más compleja de tratar.

Por ejemplo, en el caso del ANOVA anterior, habíamos considerado como una variable dependiente o repuesta (longitud) depende de otra variable independiente o predictora (afiliación). En este caso, nuestra morfometría no está definida por esa única variable, sino por todas las variables morfométricas que ha producido nuestra extracción de datos mediante Morfometría Geométrica. Por lo tanto, ANOVA (que requiere una única variable respuesta) ya no es aplicable. Del mismo modo, una gran parte de análisis estadísticos requerirán una única variable respuesta, y ya no son aplicables (porque nuestra variable respuesta son todas las que produce la morfometría). En este punto, cobran importancia clave dos conceptos, que son los que nos llevan al desarrollo metodológico predominante en Morfometría Geométrica.

- **Análisis supervisados vs análisis no supervisados:** En general, hablamos de análisis supervisados cuando disponemos de una variable respuesta. Es decir, hay una variable sobre la que queremos investigar, y que depende de toda una seria de co-variables (u otros elementos) que le influyen y que pueden producir cambios en aquella. Por el contrario, los análisis no supervisados refieren a análisis en los cuales no tenemos una variable respuesta. En este caso, no asumimos una variable dependiente de las demás y simplemente nos centramos en comprender como función la relación de las variables consideradas en un plano no jerárquico.

- **Análisis univariante vs análisis multivariante:** En este caso, cuando nos referimos a análisis univariante, consideramos una única variable como respuesta. Sin embargo, en el análisis podemos considerar una serie variables (más de una) como respuesta. La mayoría de técnicas que se aplican al análisis univariante pueden aplicarse también al multivariante, pero la interpretación de las mismas siempre va a ser más compleja.

En el caso de la Morfometría Geométrica suele plantearse como que la morfometría depende de algo más (por ejemplo, del lugar de procedencia de las muestras, de la fase cronológica o de la atribución cultural). Por lo tanto, debemos recurrir a análisis multivariantes y/o no supervisados. Vamos a ver en algo de detalle alguna de las técnicas más utilizadas. Para ello, utilizaremos el objeto de outlines que hemos creado en la sesión anterior. Si no lo tienes en el environment cárgalo. Para ello, puedes hacer lo siguiente (aunque esto te cargará el que he preparado yo, que puede ser ligeramente diferente del tuyo):

```
out_PCA <- readRDS("./Datos/out_PCA.rds")
geo_out_F <- readRDS("./Datos/geo_out_F.rds")
geos <- read.csv("./Datos/mom_geos.csv")[,-1]
```

Análisis de Componentes Principales

Ya hemos visto anteriormente una versión rápida del análisis de componentes principales, pero ¿Qué es exactamente? Esta técnica, en general, se utiliza para una reducción de dimensionalidad. En esencia, se intenta expresar elementos complejos (con gran cantidad de variables) en elementos más simples (con menos variables), donde se pierda la menor cantidad posible de información. El objetivo de esto es una mejor comprensión de los datos para entender cómo funcionan estos en ulteriores análisis. Ahora bien, los datos extraídos en Morfometría Geométrica no tienen sentido de por sí (a excepción de algunos landmarks) y, por lo tanto, intentar adjudicarles un significado deja de tener sentido. Como consecuencia, muchas de las aplicaciones de componentes principales en morfometría geométrica han virado hacia una especie de análisis de agrupamiento (ver de acuerdo con qué se agrupan morfométricamente o no los distintos artefactos estudiados) que, aunque aplicable, no es el objetivo último de esta técnica, ni esta esta técnica es la más recomendable para este objetivo. Abundaremos más en esto pero, por ahora, vamos a ver cómo funciona conceptualmente (os ahorraré las mates!).

En esencia, se van a producir una serie componentes principales (PC), uno por variable analizada, donde cada componente explica un porcentaje de la variabilidad total de los datos. Todos los componentes explicarían toda la variabilidad, pero si los utilizáramos todos no reduciríamos en nada la dimensionalidad así que no tiene demasiado sentido utilizarlos todos. El primer componente principal (PC1) es el que más variabilidad explica, y la cantidad de variabilidad explicada se va reduciendo conforme avanzamos en componentes (normalmente exponencialmete). En un mundo ideal, querríamos que los dos primeros componentes principales explicaran el máximo de la variabilidad posible, y así podríamos enseñarlo todo en una única figura donde PC1 = x y PC2 = y, pero en la vida real esto no siempre es así. Como regla básica, suele considerarse que un 70-80% de variabilidad explicada es suficiente, aunque ese valor depende de muchas cosas. Y ¿Cómo sé cuántos componentes debo utilizar? La mayoría de funciones que realizan análisis de componentes principales incluyen los valores de variabilidad explicada por componentes y cumulativos. Para el objeto resultante del análisis de componentes principales en `Momocs`, podemos ver la variabilidad explicada cumulativa y por componente del siguiente modo

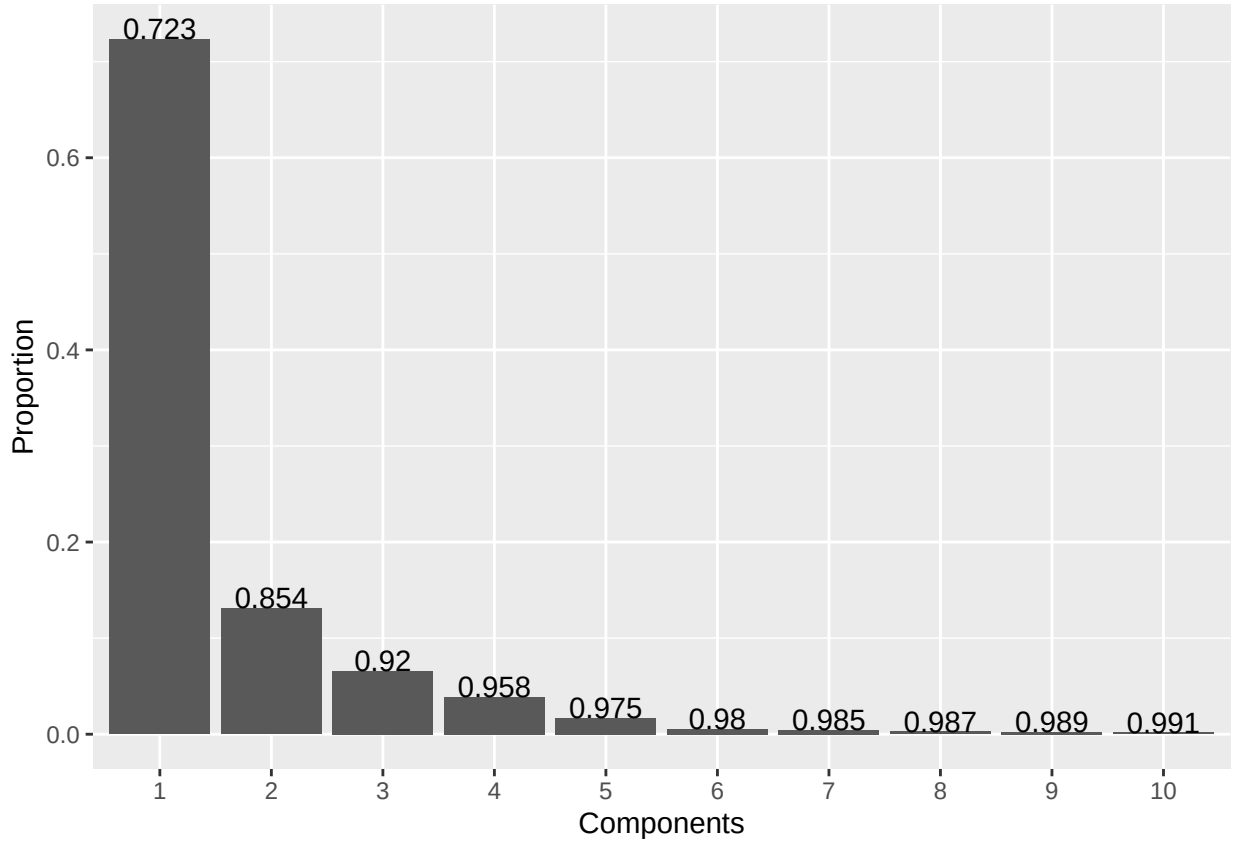
```
scree(out_PCA)
```

```
## # A tibble: 28 x 3
##     axis proportion cumsum
##    <int>      <dbl>  <dbl>
##  1     1    0.723    0.723
##  2     2    0.131    0.854
```

```
##  3     3    0.0657   0.920
##  4     4    0.0385   0.958
##  5     5    0.0167   0.975
##  6     6    0.00539  0.980
##  7     7    0.00421  0.985
##  8     8    0.00268  0.987
##  9     9    0.00214  0.989
## 10    10    0.00174  0.991
## # i 18 more rows
```

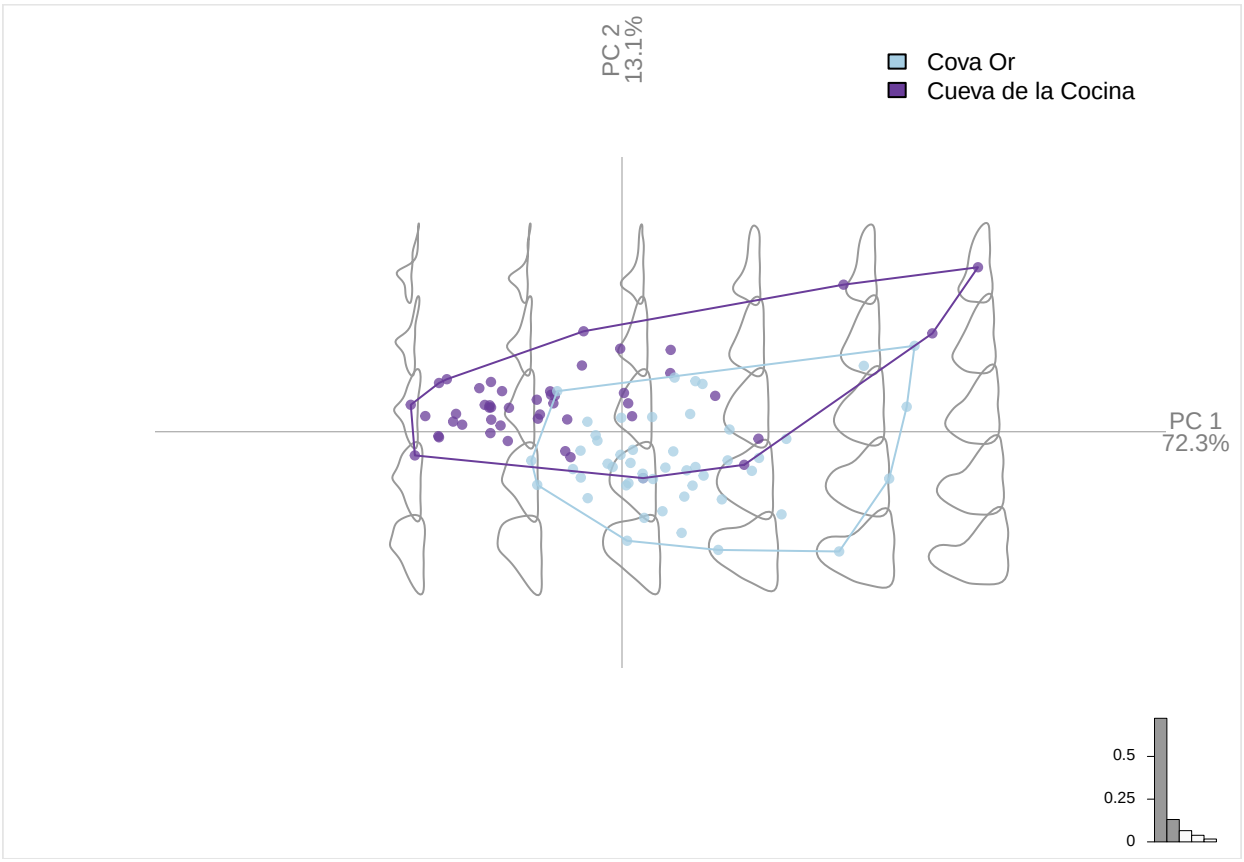
Pero es más fácil observar esto en un plot, y los hay específicos para ello. Más en concreto, el llamado scree plot.

```
scree_plot(out_PCA)
```

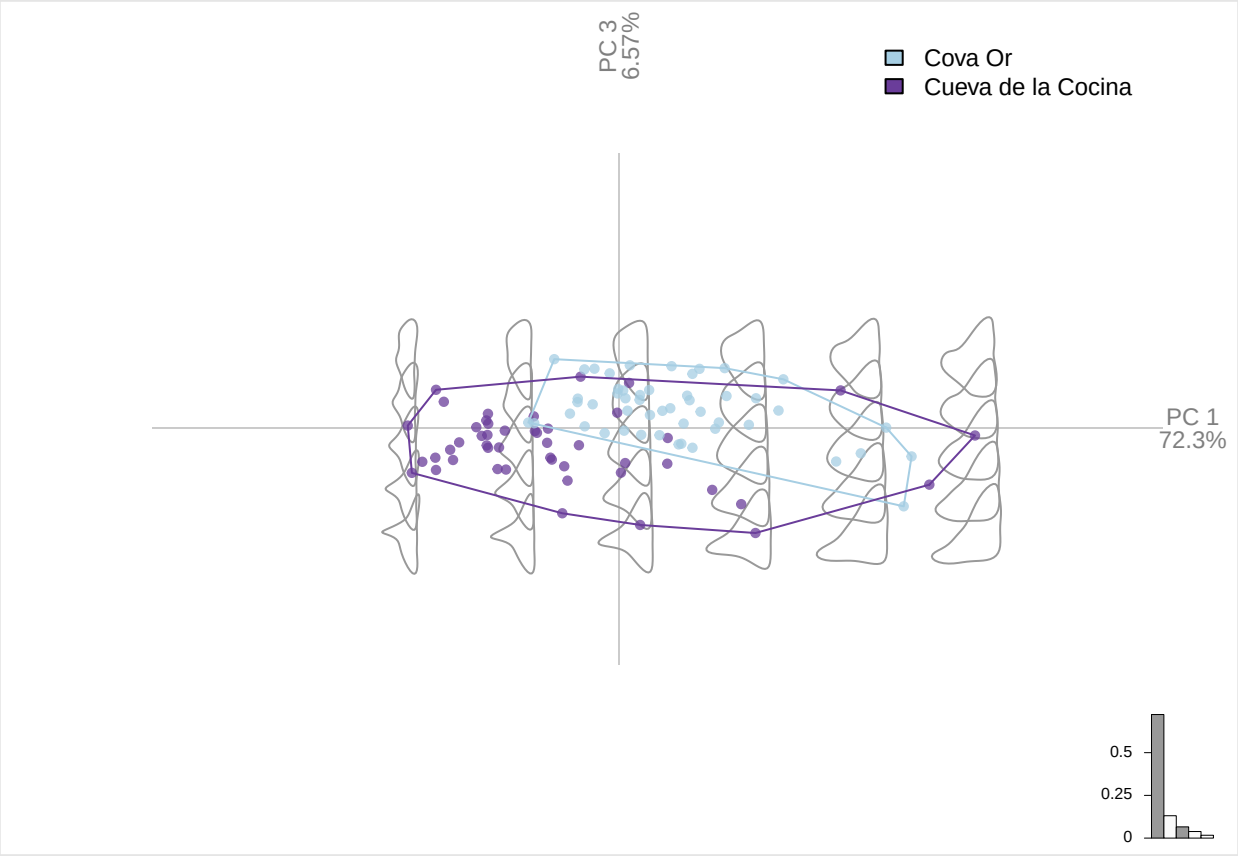


Como podéis observar, este plot nos indica gráficamente la variabilidad acumulada por cada componente y es una buena referencia para decidir cuántos componentes debería de incluir en mi posterior análisis. En este caso, los dos primeros componentes justifican el 85% de la varianza, así que sería suficiente para desarrollar nuestro análisis, y lo tenéis representado en los gráficos de la sesión previa. Pero ¿Y si necesitáramos más componentes? ¿Cómo podemos hacer un gráfico que represente más componentes? Un gráfico 3D, por más que os parezca atractivo, no suele ser una buena, ya que normalmente es bastante difícil de entender. Supongamos que es nuestro caso y que queremos representar los primeros tres componentes principales. En ese caso, la forma más ilustrativa sería presentar tres plots donde vemos cómo se relacionan, respectivamente, el PC1 y el PC2, el PC1 y el PC3 y el PC2 y el PC3. Con la función `plot_PCA` de `Momocs` podemos hacer esto, simplemente cambiando el argumento `axes`.

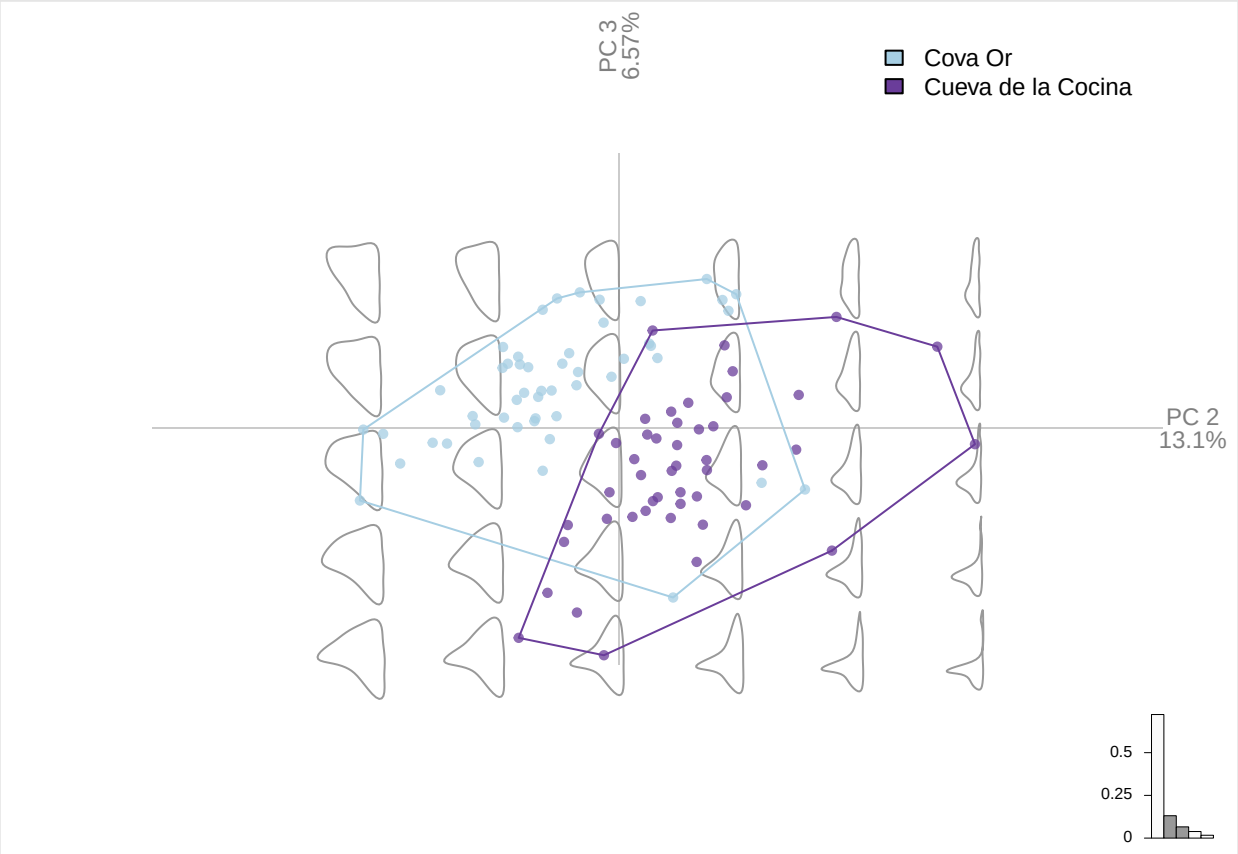
```
plot_PCA(out_PCA, 1, axes = c(1,2), palette = col_qual,morphospace = TRUE)
```




```
plot_PCA(out_PCA, 1, axes = c(1,3), palette = col_qual,morphospace = TRUE)
```



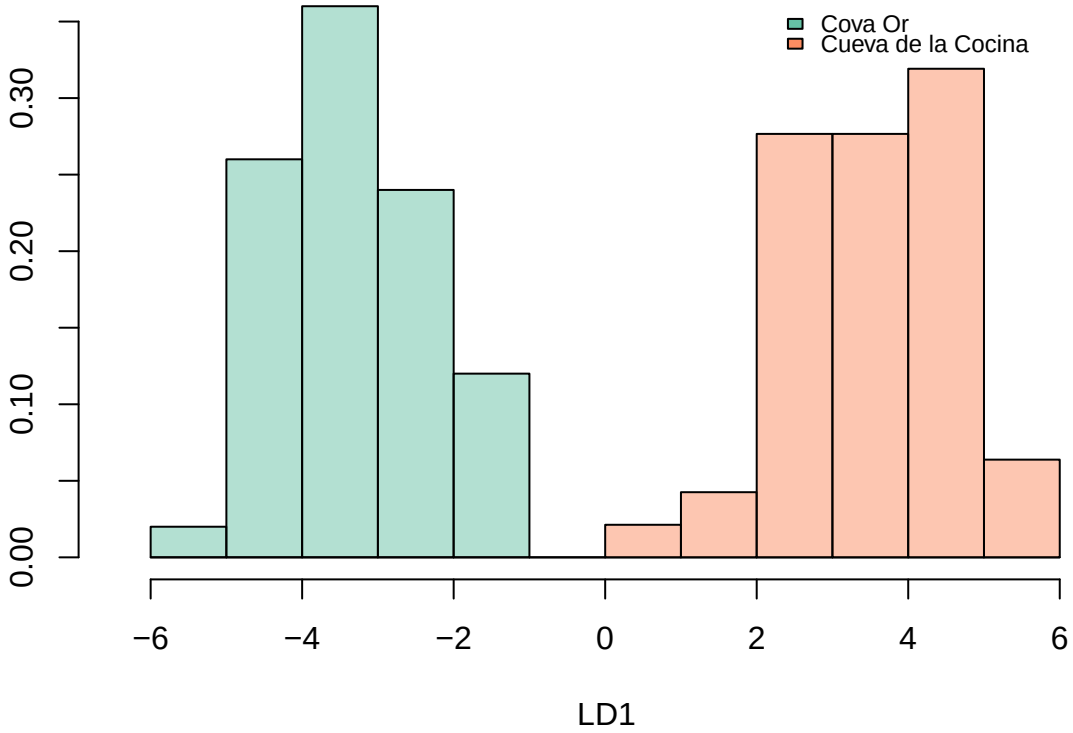

```
plot_PCA(out_PCA, 1, axes = c(2,3), palette = col_qual,morphospace = TRUE)
```



Análisis Linear Discriminante

En cualquier caso, los análisis de componentes principales no están pensando para realizar agrupamientos, y existen mejores técnicas para ello. Más en concreto, los análisis de componentes principales se centran en la variabilidad intragrupal mientras que, si lo que queremos es entender las diferencias entre grupos, deberíamos de fijarnos en la maximización de la variabilidad intergrupal. Ese es el caso del análisis linear discriminante. El concepto y la forma de entenderlo es muy similar a cómo entendemos el análisis de componentes principales. Vemos un plot de un LDA para comprenderlo.

```
out_LDA <- LDA(geo_out_F, fac = as.factor(geos$Site))
plot_LDA(out_LDA) ## Como solo tenemos dos niveles, prepara un histograma
```



Aún así ¿Es suficiente una inspcección visual? Viendo este histograma, y a partir de este 'resumen' de los datos, ya podríamos aplicar alguna de las técnicas de la primera sección de esta sesión ¿Se os ocurre cual?

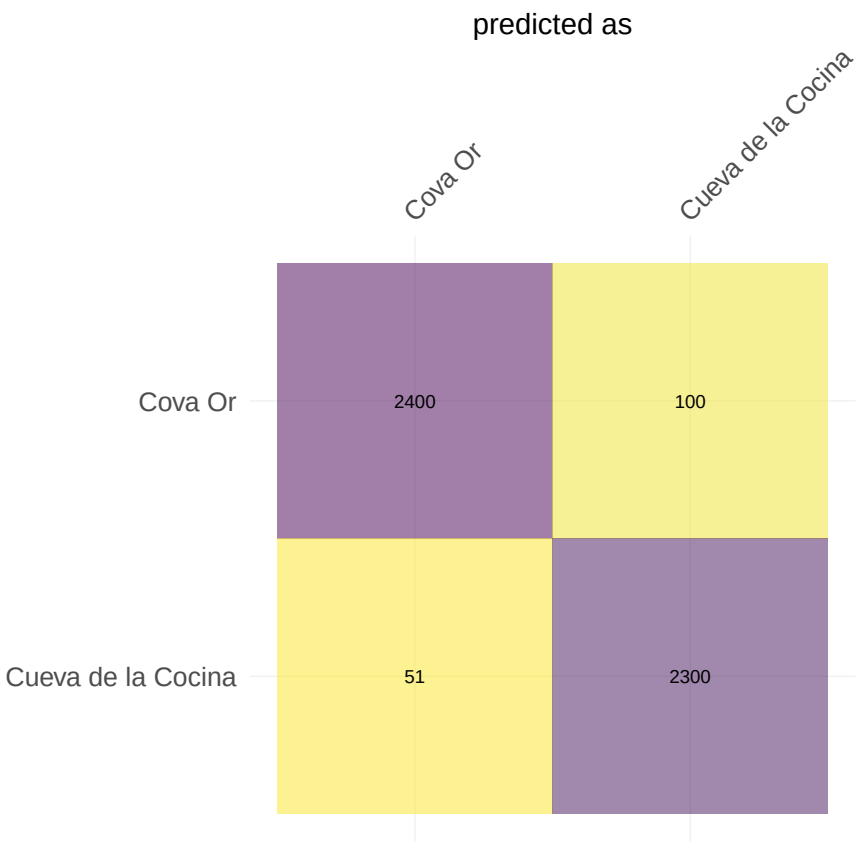
Más que visualizaciones

En todo caso, la inspección visual no es suficiente, y la mayoría de software y paquetes de análisis de morfometría geométrica son conscientes de ello e implementan soluciones. En primer lugar, una de las cosas que nos va a interesar es saber cómo de bueno es nuestro modelo. Una técnica muy utilizada en este sentido es la **validación cruzada**. Esta técnica se usa para testear el modelo y, esencialmente, consiste en 'preguntarle' al modelo a qué grupo diría que pertenece un artefacto específico (del cual ya sabemos su procedencia) y ver 'si acierta' o no.

```
## Creamos el primer objeto
cvn <- out_LDA$CV.tab

## Repetimos el proceso 50 veces
for(i in 1:50){
  provcvn <- LDA(geo_out_F, fac = as.factor(geos$Site))
  provcvn <- provcvn$CV.tab
  cvn <- provcvn+cvn
}

plot_CV(cvn)
```



Vemos que nuestro modelo tiene un gran porcentaje de acierto (*e.g.* puede diferenciar bien), con muy pocos fallos en la reclasificación, acertando más de un 95% de veces (índice de acierto del 0.958).

Ahora que estamos tranquilos de que nuestro modelo funciona, podemos querer calcular las diferencias entre los dos yacimientos teniendo en cuenta toda la morfometría y de una forma más precisa. Como se comentaba anteriormente, si se tiene en cuenta toda la morfometría tendremos varias variables respuesta, así que, en este caso no podremos utilizar el análisis de varianza (ANOVA) que exije una única respuesta, pero sí podemos utilizar el análisis de varianza multivariado (MANOVA), que se incluye por defecto entre las funciones de Momocs. Su aplicación es fácil, y se debe desarrollar sobre los PCs obtenidos en nuestro PCA anterior.

```
geo_man <- MANOVA(out_PCA, fac = as.factor(geos$Site))
```

```
## PC axes 1 to 10 were retained
```

```
geo_man
```

```
##           Df Hotelling-Lawley approx F num Df den Df    Pr(>F)
## fac        1           7.4525   64.091     10     86 < 2.2e-16 ***
## Residuals 95
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Podemos ver como, efectivamente, el lugar de procedencia de estos geométricos definitivamente tiene un impacto en su morfometría final.

Propuestas adicionales

Las técnicas vistas más arriba son algunas de las más comunes (aunque no las únicas) aplicadas de modo estándar en la Morfometría Geométrica. Sin embargo, y como último elemento del workshop, me gustaría llamar la atención sobre la unión entre las dos secciones anteriores. En primer lugar, pensar que, como ya se ha mencionado, la Morfometría Geométrica solo es una forma sofisticada de conseguir medidas de varios arteactos y, por lo tanto, todo aquello que quisiéramos hacer con medidas más simples también lo podemos hacer con estas más complejas.

Hemos visto cómo GM produce una gran cantidad de información, pero también hemos visto que esa gran cantidad de variables puede reducirse a una o dos (PC1 y PC2) sin prácticamente perder variabilidad. En nuestro caso, los PC1 y PC2 conforman aproximadamente el 85% de nuestra variabilidad y, por lo tanto, pueden utilizarse como proxies de la morfometría general del objeto. Esencialmente, hemos reducido toda esa morfometría en únicamente dos variables. Con esto hecho, podemos incluir otras variables no morfométricas para un análisis más completo. Por ejemplo, consideremos que tenemos información sobre la dirección del retoque. Una vez hemos extraído las dos variables de la morfometría, podemos también incluir el retoque. Configuremos nuestros datos.

```
# Extraemos PC1 y PC2 del objeto
PC1 <- out_PCA$x[,1]
PC2 <- out_PCA$x[,2]

# Introducimos la información del retoque
retoque <- readRDS("./Datos/retoque.rds")

# Creamos nuestro data frame con los datos
dat <- data.frame("Yacimiento" = geos$Site,
                  "PC1" = PC1,
                  "PC2" = PC2,
                  "Retoque" = retoque)
head(dat)
```

```
##             Yacimiento         PC1         PC2 Retoque
## 114 Cueva de la Cocina -0.52232678 -0.01593079 Abrupto
## 120 Cueva de la Cocina -0.32585872 -0.02642919 Abrupto
## 261 Cueva de la Cocina -0.23997653  0.03733892 Abrupto
## 328 Cueva de la Cocina  0.01772766  0.08120579 Abrupto
## 346 Cueva de la Cocina -0.37865631  0.06888281 Abrupto
## 491 Cueva de la Cocina -0.37754949  0.07556067 Abrupto
```

Supongamos, por ejemplo, que queremos ver cómo cada una de estas tres variables contribuye a las diferencias entre un yacimiento y otro. Tenemos una variable respuesta 'yacimiento' que es binomial (tiene dos categorías, una por yacimiento) y, por lo tanto, el tipo de modelo más apropiado que necesitamos es una regresión logística. Es decir, estamos hablando de un modelo de la siguiente forma

$$logit(\theta_i) = log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \beta_1 PC_1 + \beta_2 PC_2 + \beta_3 retoque$$

No se va a entrar en detalles de por qué este u otro modelo, ya que eso requeriría alguna sesión específica. Lo importante aquí es que nos demos cuenta de que se puede ir más allá (y lo que estamos haciendo ahora no es particularmente sofisticado, hay cosas mucho más molonas!). Aplicar esta regresión en R es muy sencillo.

```r
mod1 <- glm(as.factor(Yacimiento) ~ PC1 + PC2 + retoque, data = dat, family = "binomial")
summary(mod1)
```

```
##
## Call:
## glm(formula = as.factor(Yacimiento) ~ PC1 + PC2 + retoque, family = "binomial",
##     data = dat)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.03897  -0.00008   0.00000   0.31057   2.20507
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)       0.6162     0.4087   1.508 0.131606
## PC1              -4.8710     1.3088  -3.722 0.000198 ***
## PC2              13.3207     3.3932   3.926 8.65e-05 ***
## retoqueOblicuo  -20.6584  2808.0278  -0.007 0.994130
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 134.378  on 96  degrees of freedom
## Residual deviance:  41.878  on 93  degrees of freedom
## AIC: 49.878
##
## Number of Fisher Scoring iterations: 19
```

Sin entrar en elementos y posibles modeizaciones más complejas (este modelo es mejorable), vemos cómo, en principio, la morfometría puede servir como elemento para diferencias entre yacimientos, mientras que el retoque no tanto.

En este workshop se ha hablado de muchas cosas, y no se espera que se entiendan todas a la primera, pero que sirva de un recurso al que volver de vez en cuando y ampliar cuando sea necesario a partir de aquí. Una vez se entienden estos conceptos, lo que queráis hacer realmente va a depender de lo que se os pueda ocurrir.