Jimenez-Ruiz-Julia-PEC1

Análisis de Datos Ómicos

Tabla de contenido

Abstract

La cachexia es un síndrome metabólico complejo (Evans et al, 2008) que presenta dificultades de diagnostico debido a la falta de consenso clínico sobre su diagnóstico. Esto ha hecho que no se haya podido desarrollar tratamiento específico para esta patología.

Hasta ahora el diagnóstico de Cachexia se ha realizado normalmente en pacientes con IMC inferior a 20.0 kg/m^2. Otros criterios utilizados han sido la pérdida de masa muscular acelerada, fuerza muscular disminuida, fatiga, anorexia, masa muscular libre de grasa disminuida, y diferencias bioquímicas como inflamación anemia o hipoalbuminemia.

Tras el diagnóstico, los pacientes tienen dianas terapéuticas limitadas, la nutrición parenteral no es una buena línea, ya que, aunque mantengan el peso, siguen perdiendo masa muscular esquelética. Se ha probado la utilización de andrógenos con escasos resultados.

En este trabajo en el que se compara la orina de pacientes con personas sanas, se intenta buscar diferencias en algunos metabolitos con el objetivo de encontrar dianas terapéuticas o posibles líneas de investigación adicionales.

A lo largo de este trabajo se analizan los datos en una primera aproximación, para luego entrar en un análisis más profundo en el que se analizan las principales componentes del dataset. También se intenta buscar una jerarquía de los datos para ver si hay metabolitos agrupados que nos puedan sugerir alguna ruta metabólica que podamos estudiar.

En definitiva, este trabajo supone una exploración inicial de los datos metabólicos de pacientes de cachexia, una patología a todos los efectos, desconocida.

Objetivos del estudio

El principal objetivo es identificar las diferencias de los metabolitos de los pacientes de cachexia para intentar averiguar de dónde procede la pérdida de masa muscular esquelética. Una vez identificadas las diferencias metabólicas entre los dos grupos, se podrán identificar potenciales dianas terapéuticas.

Los autores del estudio en el que se basa la base de datos alertan sobre la ausencia de información sobre la información y los intentos infructuosos de abordajes terapéuticos de la enfermedad, aunque sugieren que la suplementación con hormonas andrógenos pueden tener cierto efecto preventivo del deterioro de la masa muscular esquelética.

Es por ello, por lo que este trabajo, en el que se estudia otros metabolitos pueda sugerir el estudio de algunas rutas metabólicas hasta ahora no contempladas o incluso el desarrollo de alguna línea terapéutica.

Materiales y métodos

El diseño experimental de este estudio ha sido la obtención de datos en crudo de pacientes de cachexia (n=47) y la comparación con datos de pacientes sanos (n=30) de la base de datos "specimen.datasets" de R. En total, se analizaron 65 variables metabólicas de 77 muestras. Los datos obtenidos por tanto corresponden a la cuantificación de los metabolitos (unidades desconocidas).

Algunos de los metabolitos cuantificados son aminoácidos presentes en el tejido muscular esquelético como la lisina. También se ha analizado la glucosa, o metabolitos intermedios como la X1. Methylnicotinamide un producto de la metabolización de la Nicotinamida

Estos datos se encontraban dispuestos en un repositorio de GitHub, se ha elegido uno aleatoriamente. Posteriormente, se ha procedido a la lectura del artículo de referencia y a la prospección inicial de los datos. El análisis de los datos se ha realizado en el programa Rstudio.

En la primera prospección de los datos, se han extraído los metadatos que, en este caso, han sido: el identificador del paciente y el diagnóstico o ausencia de él (cachexic o sano).

Posteriormente se ha realizado un Análisis de Componentes principales (PCA) para intentar simplificar el análisis de las 65 variables. A continuación, y debido a la naturaleza de los datos se realizó también en R studio, un análisis de proximidades mediante un "Cluster Analysis" con el fin de obtener grupos de variables dentro del conjunto de observaciones.

Para crear el repositorio, con la cuenta en GitHub previamente creada en actividades anteriores de la asignatura. He creado un repositorio según las instrucciones proporcionadas en el enunciado de la actividad. He configurado el repositorio, configurando los permisos de acceso y he subido los elementos requeridos. El link al repositorio se encuentra al final del presente informe.

Resultados

En primer lugar, se hizo un análisis preliminar de los datos para ver qué datos contenía nuestro dataset y sobre todo qué tipo de datos. El código completo se encuentra en el repositorio, pero con el fin de evitar la carga de datos que impida la lectura fluida del documento, inserto la imagen que refleja que en nuestro dataset hay 77 observaciones y 65 variables analizadas.

```
str(data_cach)

## 'data.frame':    77 obs. of  65 variables:
##  $ Patient.ID                : chr  "PIF_178" "PIF_087" "PIF_090" "NETL_005_V1" ...
##  $ Muscle.loss               : chr  "cachexic" "cachexic" "cachexic" "cachexic" ...
##  $ X1.6.Anhydro.beta.D.glucose: num  40.9 62.2 270.4 154.5 22.2 ...
##  $ X1.Methylnicotinamide     : num  65.4 340.4 64.7 53 73.7 ...
##  $ X2.Aminobutyrate          : num  18.7 24.3 12.2 172.4 15.6 ...
##  $ X2.Hydroxyisobutyrate     : num  26.1 41.7 65.4 74.4 83.9 ...
##  $ X2.Oxoglutarate           : num  71.5 67.4 23.8 1199.9 33.1 ...
##  $ X3.Aminoisobutyrate       : num  1480.3 116.8 14.3 555.6 29.7 ...
##  $ X3.Hydroxybutyrate        : num  56.83 43.82 5.64 175.91 76.71 ...
##  $ X3.Hydroxyisovalerate     : num  10.1 79.8 23.3 25 69.4 ...
##  $ X3.Indoxylsulfate         : num  567 369 665 412 166 ...
##  $ X4.Hydroxyphenylacetate   : num  120.3 432.7 292.9 214.9 97.5 ...
##  $ Acetate                   : num  126.5 212.7 314.2 37.3 407.5 ...
##  $ Acetone                   : num  9.49 11.82 4.44 206.44 44.26 ...
##  $ Adipate                   : num  38.1 327 131.6 144 15 ...
##  $ Alanine                   : num  314 871 464 590 1119 ...
##  $ Asparagine                : num  159.2 157.6 89.1 273.1 42.5 ...
##  $ Betaine                   : num  110 245 117 279 392 ...
##  $ Carnitine                 : num  265.1 120.3 25 200.3 84.8 ...
```

Con la función str conseguimos saber qué tipo de datos tenemos. De las 65 variables, todas son numéricas, muy convenientes para nuestro estudio posterior y naturalmente identificación del paciente (“Patient ID”) y la variable (“Muscle.loss”) que indica si el paciente es sano o enfermo, también es de tipo *character*.

Una vez hecha la primera visual de los datos, decido convertir la variable “Muscle.loss” a factor mediante la función as.factor. También, sabiendo qué datos componen el dataset, se opta por considerar estas dos primeras variables como metadatos ya que nos aportan información adicional sobre el sujeto del estudio.

Por tanto, nuestro SummarizedExperiment lo compondrán las variables numéricas que cuantifican los metabolitos de la orina de los pacientes de cachexia y los pacientes sanos.

```
cach_Num<-data_cach[,3:65]
```

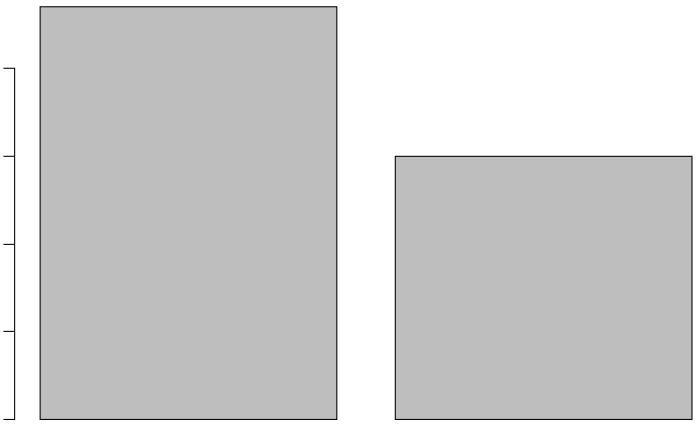
Compruebo que el dataset está bien.

```
str(cach_Num)

## 'data.frame':    77 obs. of  63 variables:
##  $ X1.6.Anhydro.beta.D.glucose: num  40.9 62.2 270.4 154.5 22.2 ...
##  $ X1.Methylnicotinamide     : num  65.4 340.4 64.7 53 73.7 ...
##  $ X2.Aminobutyrate          : num  18.7 24.3 12.2 172.4 15.6 ...
##  $ X2.Hydroxyisobutyrate     : num  26.1 41.7 65.4 74.4 83.9 ...
##  $ X2.Oxoglutarate           : num  71.5 67.4 23.8 1199.9 33.1 ...
##  $ X3.Aminoisobutyrate       : num  1480.3 116.8 14.3 555.6 29.7 ...
##  $ X3.Hydroxybutyrate        : num  56.83 43.82 5.64 175.91 76.71 ...
```

Tal y como se observa en la imagen nuestro SummarizedExperiment (cach_Num) ya lo componen 63 variables y todo de ellas numéricas. Se mantienen todas nuestras observaciones. En algún caso, podría ser beneficioso en este punto conocer si nuestra base de datos tiene algún valor nulo (na) y eliminar esa observación. En nuestro caso, nuestro dataset no tiene ningún na.

Tras la exploración de datos inicial, el siguiente paso es visualizar los datos mediante histogramas. Al tener tal cantidad de variable no es muy intuitivo la visualización. A veces este paso puede sugerir tendencias o comportamientos de alguna variable. En todo caso, el histograma claro es que indica las muestras de pacientes vs las muestras de pacientes sanos. El resto de variable tienen un comportamiento similar en la mayoría de los pacientes excepto alguna anomalía puntual.



A continuación, tal y como indican los ejemplos proporcionados [2] al ser variables del mismo tipo, son todo mediciones de metabolitos urinarios, pero en escalas heterogéneas, basaremos nuestro análisis de las componentes principales en la matriz de covarianzas de los datos centrado.

Posteriormente obtendremos nuestra matriz de correlaciones, debido a la magnitud de nuestros datos, me limitaré a incluir en este informe una pequeña muestra de la matriz con el objetivo de facilitar la lectura de este.

```{r}
R<-cor(cach_Num)
show(R)
```

	X1.6.Anhydro.beta.D.glucose	X1.Methylnicotinamide	X2.Aminobutyrate	X2.Hydroxyisobutyrate
X1.6.Anhydro.beta.D.glucose	1.000000000	0.058737462	0.261133385	0.50200025
X1.Methylnicotinamide	0.058737462	1.000000000	0.001473031	0.31919954
X2.Aminobutyrate	0.261133385	0.001473031	1.000000000	0.38620663
X2.Hydroxyisobutyrate	0.502000254	0.319199543	0.386206630	1.00000000
X2.Oxoglutarate	-0.011638061	0.070344207	0.267917147	0.39089812
X3.Aminoisobutyrate	0.066443616	0.020079514	0.312870120	0.13761371
X3.Hydroxybutyrate	0.213140748	0.143886414	0.602726922	0.52360916
X3.Hydroxyisovalerate	0.315202900	0.353413911	0.111229773	0.42380791
X3.Indoxylsulfate	0.284075941	0.350952439	0.318235800	0.38780848
X4.Hydroxyphenylacetate	0.362211859	0.193748381	0.288713988	0.44297043
Acetate	0.259746510	0.173594249	0.081460919	0.40408094
Acetone	0.037500937	-0.006823968	0.630615394	0.21207715
Adipate	0.151231292	0.270729654	0.285638282	0.18129022
Alanine	0.301051249	0.290684760	0.319832448	0.60718560
Asparagine	0.308768847	0.285769035	0.528665638	0.62123530

	X2.Oxoglutarate	X3.Aminoisobutyrate	X3.Hydroxybutyrate	X3.Hydroxyisovalerate	X3.Indoxylsulfate
X1.6.Anhydro.beta.D.glucose	-0.011638061	0.066443616	0.21314075	0.31520290	0.28407594
X1.Methylnicotinamide	0.070344207	0.020079514	0.14388641	0.35341391	0.35095244
X2.Aminobutyrate	0.267917147	0.312870120	0.60272692	0.11122977	0.31823580
X2.Hydroxyisobutyrate	0.390898116	0.137613706	0.52360916	0.42380791	0.38780848
X2.Oxoglutarate	1.000000000	0.107301792	0.46386373	0.08194807	0.08812611
X3.Aminoisobutyrate	0.107301792	1.000000000	0.46891499	0.03406502	0.30286355
X3.Hydroxybutyrate	0.463863728	0.468914987	1.00000000	0.38301574	0.41694158
X3.Hydroxyisovalerate	0.081948070	0.034065017	0.38301574	1.00000000	0.44725745

Con la cantidad de variables y combinaciones posibles, resulta imposible discernir cuáles son las más relevantes. Pasamos ahora a calcular las componentes principales. En este caso he escogido la función princomp y prcomp al ser las más elegantes, siendo consciente que otras formas de cálculo pueden resultar en pequeñas diferencias en los resultados. Este método calcula las componentes principales mediante la descomposición en valores singulares de la matriz de datos.

Los dos métodos escogidos indican que los dos principales componentes serían X1.6.Anhydro.beta.D.glucose y X1.Methylnicotinamide

```
Loadings:
                       Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10 Comp.11 Comp.12 Comp.13
X1.6.Anhydro.beta.D.glucose
X1.Methylnicotinamide                                                                            0.125   0.104
X2.Aminobutyrate
X2.Hydroxyisobutyrate
X2.Oxoglutarate                                          -0.129 -0.161              -0.611          -0.506
X3.Aminoisobutyrate                                                                       0.109  -0.373  -0.203
X3.Hydroxybutyrate
X3.Hydroxyisovalerate
```

```
```{r}
PCAS3$rotation[,1]
```
```

```
X1.6.Anhydro.beta.D.glucose       X1.Methylnicotinamide           X2.Aminobutyrate       X2.Hydroxyisobutyrate
               0.0073705650                0.0071888723                0.0015869038                0.0023680187
              X2.Oxoglutarate         X3.Aminoisobutyrate         X3.Hydroxybutyrate         X3.Hydroxyisovalerate
               0.0212316043                0.0096481157                0.0024423997                0.0023581616
           X3.Indoxylsulfate     X4.Hydroxyphenylacetate                     Acetate                     Acetone
               0.0200924404                0.0091624317                0.0062654655                0.0008158222
                     Adipate                     Alanine                   Asparagine                     Betaine
               0.0030140039                0.0277339653                0.0056489585                0.0056298088
                  Carnitine                     Citrate                    Creatine                  Creatinine
               0.0024653825                0.2194561979                0.0071002685                0.9416806687
                Dimethylamine                 Ethanolamine                     Formate                     Fucose
               0.0388025987                0.0303773866                0.0181786141                0.0095168598
                     Fumarate                     Glucose                    Glutamine                     Glycine
               0.0010202294                0.0320750612                0.0278575074                0.0791026073
                   Glycolate               Guanidoacetate                   Hippurate                   Histidine
               0.0145728062                0.0051942710                0.1975957249                0.0321503967
                  Hypoxanthine                  Isoleucine                     Lactate                     Leucine
               0.0066106507                0.0004928095                0.0105073404                0.0021044991
                        Lysine                  Methylamine               Methylguanidine          N.N.Dimethylglycine
               0.0052488167                0.0012301660                0.0011894641                0.0025904768
           O.Acetylcarnitine                  Pantothenate                  Pyroglutamate                     Pyruvate
```

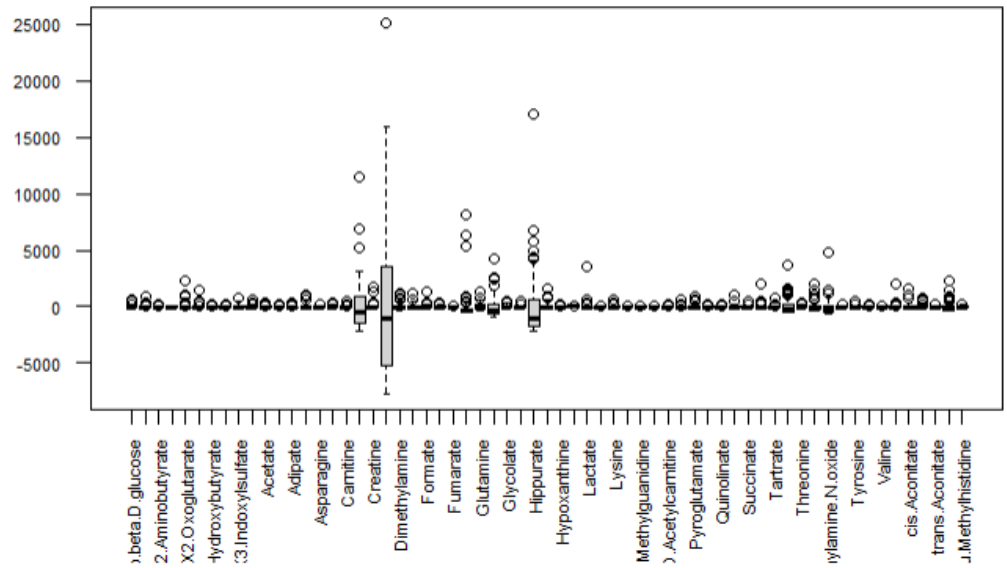
Fijándonos en el tercer método prcomp que es el más fácil de interpretar, vemos que la ecuación que podría explicar nuestro dataset sería:

Y= 0.94x Creatinina+0.22xCitrato+0.197Hippurate

He añadido las tres componentes principales, porque explican la mayoría de la variabilidad y porque las siguientes componentes principales que le siguen, aportan mucho menos a la explicación de la variabilidad.

Lo siguiente que he hecho ha sido ver un posible cluster de las observaciones. Este proceso tiene que hacerse una vez se haya hecho el preprocesamiento y filtrado de los datos.

Para asegurarnos de que es así, representamos los datos mediante boxplot. Esperamos obtener valores de expresión relativa normalizados.



Para nuestra sorpresa, se observa que la variable que sobresale es la creatina, metabolito diferente a la creatinina. Esto indica que no nuestros datos no están normalizados. Otra de las variables que destaca es la carnitina. Los metabolitos que sobresalen en el gráfico son los que presentan desviaciones estándar más altas.

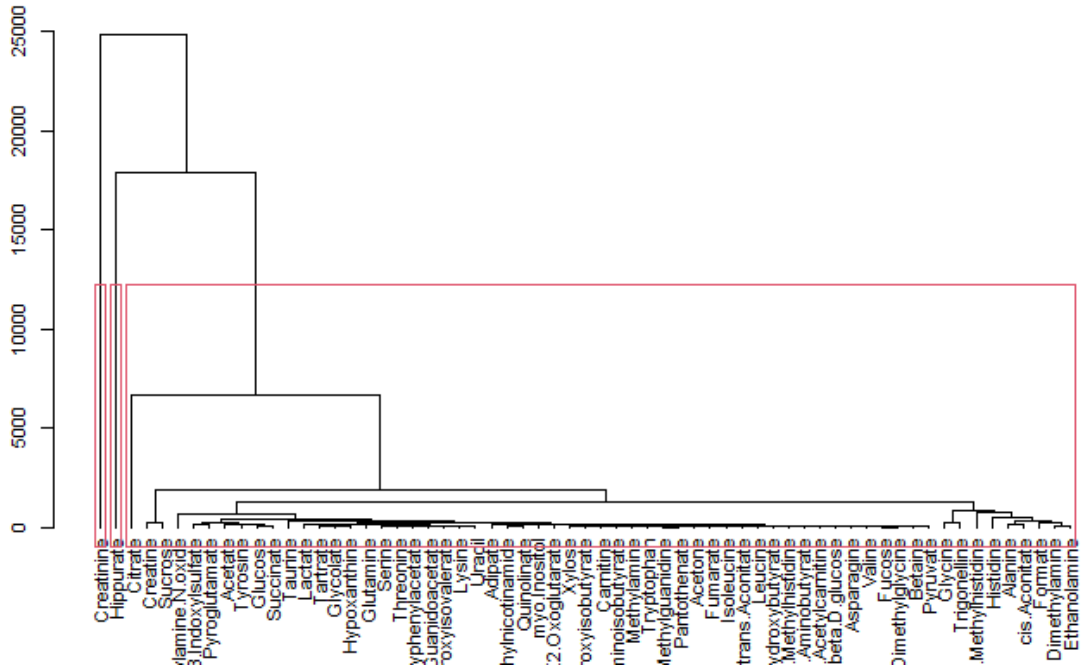
A continuación, seleccionaremos aquellos metabolitos que tengan el 1% de las desviaciones estándares más altas. Esto se hace así porque se considera que solo los metabolitos que tienen mas variabilidad son útiles para la construcción no la distinción de grupos.

En nuestro caso, corresponden a las variables 2 y 63.

```{r}
percentage <- c(0.975)
sds <- apply(cach_Num, MARGIN=1, FUN="sd")
sel <- (sds>quantile(sds,percentage))
cach.sel <- cach_Num[sel, ]
dim(cach.sel)
```

```
[1]  2 63
```

Con estos datos, para hacerlo lo más visual posible, elaboramos un dendograma.



Tal y como esperábamos, la creatinina aparece en un brazo del dendograma separada del resto de variables. Tras ella, se abre otro brazo del dendograma que se divide en el hippurato y otro brazo con el resto de citrato y la creatina y sucrosa en otro brazo conjunto.

Discusión y limitaciones y conclusiones del estudio

En este trabajo se ha hecho una somera exploración de los datos en las que se ha averiguado mediante la técnica de Análisis de Componentes Principales que las variables que explican la mayoría de la variabilidad de nuestros datos son la Creatinina, el Citrato y el Hippurato, en ese orden.

Posteriormente se ha llevado a cabo un Análisis de Cluster con el fin de intentar dislucidar si había alguna agrupación de las observaciones. Para una mejor visualización de la información se ha elaborado un dendograma.

El dendograma ha mostrado lo que el PCA sugería, hay una agrupación por creatinina y luego por hipurato y citrato, en el caso del dendograma se altera el orden.

En cualquier caso, los dos análisis llevados a cabo sugieren que estos tres metabolitos explican la variabilidad de los datos. Si bien es cierto que en ningún caso se sabe si están aumentados o disminuidos respecto a los pacientes sanos y tampoco se ha demostrado que la diferencia sea de los pacientes de cachexia frente a los sanos. Eso es materia de un análisis posterior más exhaustivo.

En relación con las limitaciones del estudio está la ausencia de información sobre la selección de pacientes y sobre el proceso de obtención de la muestra.

En este caso, la selección de pacientes es importante, ya en la revisión de 2008, Evans indica que para los futuros estudios es recomendable hacer una clasificación de los pacientes según el tiempo desde la aparición de los síntomas debido al carácter progresivo de los síntomas.

En relación con los pacientes, también es importante conocer si el estudio es mono o multicéntrico ya que hay una falta de consenso en los criterios diagnósticos de los pacientes. Esto podría provocar que algunos pacientes del estudio indicados como pacientes de cachexia no lo fueran, por ejemplo.

Tampoco tenemos una tabla de las características principales de los pacientes, por ejemplo, la edad, el sexo, el peso o si están tomando algún tratamiento en el presente. Todas estas variables podrían estar influyendo en la composición de la orina obtenida.

En el caso de la edad, se ha visto que los síntomas no son iguales en niños que en adultos, por ejemplo, por lo que sería interesante conocer la edad de los pacientes.

Además, no sabemos en qué momento se ha obtenido la orina, o si es la orina de todo un día. Esto podría también influir en la composición de la misma.

Con el análisis de los datos se concluye que hay una agrupación de las observaciones con la creatinina. Aunque no sabemos si está aumentada o disminuida. También se observa una diferencia por el hippurato y citrato. Se sugiere el estudio de vías metabólicas de degradación o endogénesis de estos metabolitos con el fin de identificar rutas alteradas en pacientes de Cachexia.

Link al repositorio

https://github.com/juliajrruiz95/Jimenez-Ruiz-Julia-PEC1

Bibliografía

[1]Evans WJ, Morley JE, Argilés J, Bales C, Baracos V, Guttridge D, Jatoi A, Kalantar-Zadeh K, Lochs H, Mantovani G, Marks D, Mitch WE, Muscaritoli M, Najand A, Ponikowski P, Rossi Fanelli F, Schambelan M, Schols A, Schuster M, Thomas D, Wolfe R, Anker SD. Cachexia: a new definition. Clin Nutr. 2008 Dec;27(6):793-9. doi: 10.1016/j.clnu.2008.06.013. Epub 2008 Aug 21. PMID: 18718696.

[2]https://aspteaching.github.io/AMVCasos/#an%C3%A1lisis-de-componentes-principales