Maria Teresa Casasnovas

PAC1

1. RESUM

Aquest estudi costa de crear un objecte de classe *SummarizedExperiment* amb dades i metadades d'un estudi prèviament realitzat sobre el càncer gàstric. Les dades s'han extret d'un repositori GitHub que contenia una taula Excel amb els resultats i una altra taula amb la informació de cada columna. Els resultats mostren el valor de les 149 modificacions metabòliques estudiades de cada mostra i en l'altra taula s'indica el nom de cada alteració bioquímica, el seu *percentage of missing* i la QC_RSD (Desviació estàndard relativa del control de qualitat). Una vegada carregades les dades, s'han modificat per aconseguir el format correcte per crear l'objecte de classe *SummarizedExperiment* i després seguidament s'han explorat de dades per tal de comprovar la classe, les dimensions, el nom de les columnes i la visualització de les taules. Finalment, s'han visualitzat els histogrames corresponents, les matrius i els components principals.

2. OBJECTIUS

a. Descarregar i exportar un dataset de *metabolòmica*
b. Crear un objecte de classe *SummarizedExperiment* amb les dades i les metadades
c. Explorar les dades de l'objecte *SummarizedExperiment*
    i. Recórrer les classes de *SummarizedExperiment*
    ii. Realitzar els càlculs per explorar les dades

3. MÈTODES

- Origen i naturalesa de les dades: les dades procedeixen del tutorial sobre l'anàlisi bàsic de dades de metabolòmica de CIMBC que utilitza un conjunt de dades provinent d'un estudi publicat prèviament per Chan et al. (2016) a la revista *British Journal of Cancer*. Les dades són arxius processats de RMN d'espectroscòpia de mostres biològiques, amb metabòlits identificats i anotats.

- Metodologia emprada: primer s'han descarregat les dades de GitHub, després s'han exportat a R, s'han modificat perquè siguin adequades per crear l'objecte de classe *SummarizedExperiments* i finalment s'han explorat les dades. Per a realitzar el procediment, s'ha consultat la informació del paquet a la pàgina web de Bioconductor.

- Eines estadístiques i bioinformàtiques utilitzades: s'ha utilitzat el paquet *SummarizedExperiment* de Bioconductor que s'utilitza per emmagatzemar matrius rectangulars de resultats experimentals generats mitjançant experiments de seqüenciació i microarrays. La principal diferència entre *SummarizedExperimentes* i *ExpressionSet* és que el primer més flexible en la informació de files, cosa que permet les basades en GRanges i les descrites per DataFrames arbitràries.

- Procediment general d'anàlisis: primer s'han hagut de modificar les dades per a després poder de crear l'objecte *SummarizedExperimentes*. Seguidament, s'ha extret la informació de cada classe de l'objecte (tipus de document, nombre i nom de les columnes, nombre de files, *summary*…) i finalment s'han realitzats les passes per obtenir els histogrames, les matrius i el components principals.

4. RESULTATS

4.1 Informació del *SummarizedExperimentes*

```
> se
class: SummarizedExperiment
dim: 140 149
metadata(0):
assays(1): Data
rownames: NULL
rowData names(0):
colnames(149): M1 M2 ... M148 M149
colData names(4): Name Label Perc_missing QC_RSD
```

Figura 1

Aquí es pot observar la informació de *se*: les dimensions de 140 files i les 149 columnes, el nom de cada columna i el nom de les columnes de l'arxiu addicional.

```
> class(assays(se)$Data)
[1] "matrix" "array"
> class(colData(se))
[1] "DFrame"
attr(,"package")
[1] "S4Vectors"
```

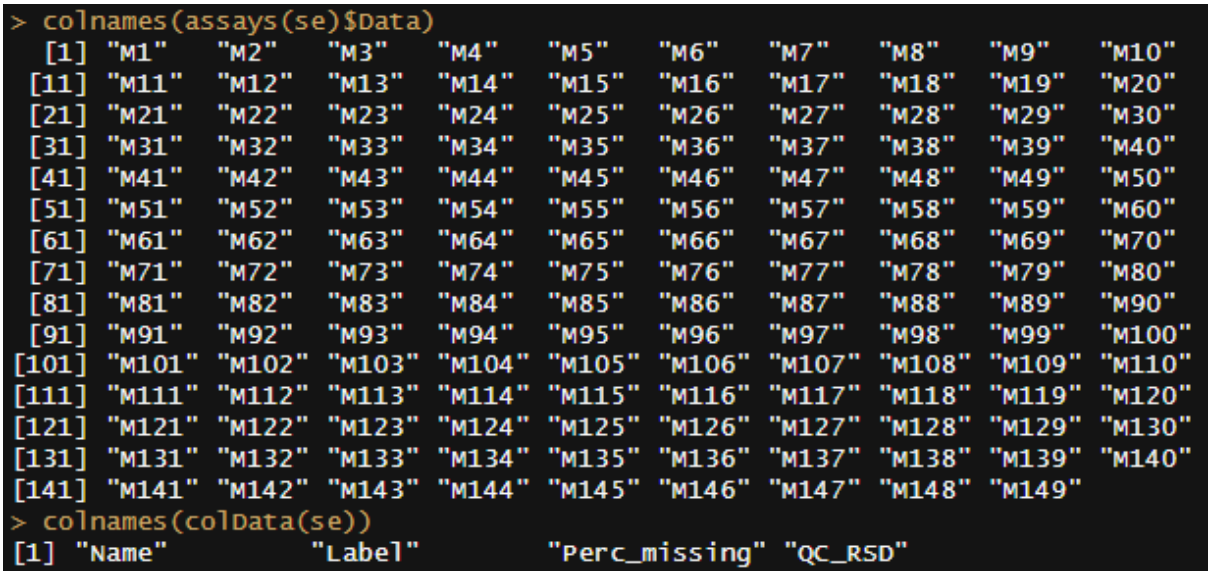
Figura 2

La tipologia de la classe *assay* és de matriu i la de *colData* de *dataframe*.

```
> dim(assays(se)$Data)
[1] 140 149
> dim(colData(se))
[1] 149   4
```

Figura 3

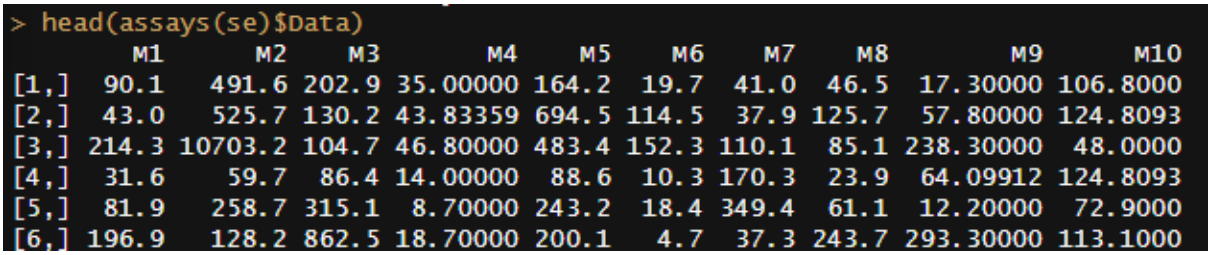
Les dades tenen 140 files i 149 columnes i les metadades 149 files i 4 columnes. Òbviament, el nombre de columnes i files d'un arxiu i de l'altre han de coincidir.



```
> colnames(assays(se)$Data)
  [1] "M1"   "M2"   "M3"   "M4"   "M5"   "M6"   "M7"   "M8"   "M9"   "M10"
 [11] "M11"  "M12"  "M13"  "M14"  "M15"  "M16"  "M17"  "M18"  "M19"  "M20"
 [21] "M21"  "M22"  "M23"  "M24"  "M25"  "M26"  "M27"  "M28"  "M29"  "M30"
 [31] "M31"  "M32"  "M33"  "M34"  "M35"  "M36"  "M37"  "M38"  "M39"  "M40"
 [41] "M41"  "M42"  "M43"  "M44"  "M45"  "M46"  "M47"  "M48"  "M49"  "M50"
 [51] "M51"  "M52"  "M53"  "M54"  "M55"  "M56"  "M57"  "M58"  "M59"  "M60"
 [61] "M61"  "M62"  "M63"  "M64"  "M65"  "M66"  "M67"  "M68"  "M69"  "M70"
 [71] "M71"  "M72"  "M73"  "M74"  "M75"  "M76"  "M77"  "M78"  "M79"  "M80"
 [81] "M81"  "M82"  "M83"  "M84"  "M85"  "M86"  "M87"  "M88"  "M89"  "M90"
 [91] "M91"  "M92"  "M93"  "M94"  "M95"  "M96"  "M97"  "M98"  "M99"  "M100"
[101] "M101" "M102" "M103" "M104" "M105" "M106" "M107" "M108" "M109" "M110"
[111] "M111" "M112" "M113" "M114" "M115" "M116" "M117" "M118" "M119" "M120"
[121] "M121" "M122" "M123" "M124" "M125" "M126" "M127" "M128" "M129" "M130"
[131] "M131" "M132" "M133" "M134" "M135" "M136" "M137" "M138" "M139" "M140"
[141] "M141" "M142" "M143" "M144" "M145" "M146" "M147" "M148" "M149"
> colnames(colData(se))
[1] "Name"         "Label"        "Perc_missing" "QC_RSD"
```

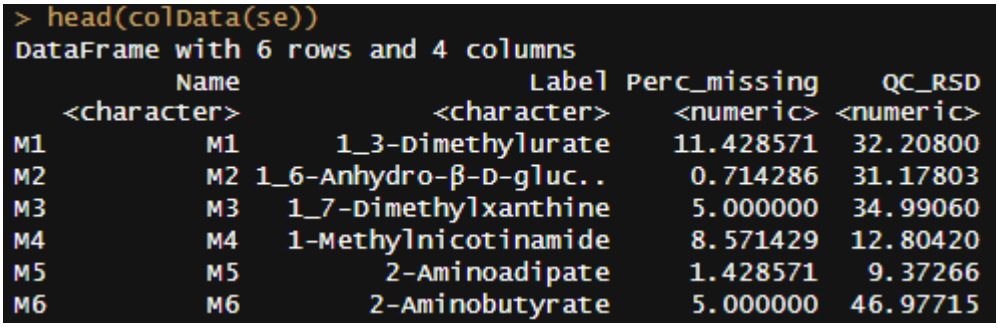
Figura 4

Cada columna de les dades és un metabòlit i cada columna de les metadades és el metabòlit, amb el nom, el seu *percentage of missing* i la QC_RSD.



```
> head(assays(se)$Data)
       M1      M2    M3       M4    M5    M6    M7    M8        M9      M10
[1,]  90.1   491.6 202.9 35.00000 164.2  19.7  41.0  46.5  17.30000 106.8000
[2,]  43.0   525.7 130.2 43.83359 694.5 114.5  37.9 125.7  57.80000 124.8093
[3,] 214.3 10703.2 104.7 46.80000 483.4 152.3 110.1  85.1 238.30000  48.0000
[4,]  31.6    59.7  86.4 14.00000  88.6  10.3 170.3  23.9  64.09912 124.8093
[5,]  81.9   258.7 315.1  8.70000 243.2  18.4 349.4  61.1  12.20000  72.9000
[6,] 196.9   128.2 862.5 18.70000 200.1   4.7  37.3 243.7 293.30000 113.1000
```

Figura 5



```
> head(colData(se))
DataFrame with 6 rows and 4 columns
          Name                  Label Perc_missing    QC_RSD
   <character>            <character>    <numeric> <numeric>
M1          M1      1_3-Dimethylurate    11.428571  32.20800
M2          M2 1_6-Anhydro-β-D-gluc..     0.714286  31.17803
M3          M3   1_7-Dimethylxanthine     5.000000  34.99060
M4          M4   1-Methylnicotinamide     8.571429  12.80420
M5          M5         2-Aminoadipate     1.428571   9.37266
M6          M6        2-Aminobutyrate     5.000000  46.97715
```

Figura 6

Aquí es poden observar les 6 primeres files de les dues classes.

4.2 ESTUDIS ESTADÍSTICS

```
> summary(assays(se)$Data)
       M1                M2                 M3                M4
 Min.   :  0.40   Min.   :    3.1   Min.   :  0.10   Min.   :  0.10
 1st Qu.: 31.82   1st Qu.:  141.4   1st Qu.: 57.05   1st Qu.: 19.80
 Median : 79.20   Median :  270.6   Median :112.85   Median : 38.20
 Mean   :101.07   Mean   :  642.0   Mean   :146.37   Mean   : 43.83
 3rd Qu.:123.75   3rd Qu.:  492.3   3rd Qu.:195.12   3rd Qu.: 50.20
 Max.   :909.90   Max.   :26195.8   Max.   :862.50   Max.   :242.50
```

Figura 7

Resum estadístic de l'*assay*; amb el mínim, màxim, mitjana, mediana i quartils de cada metabòlit.

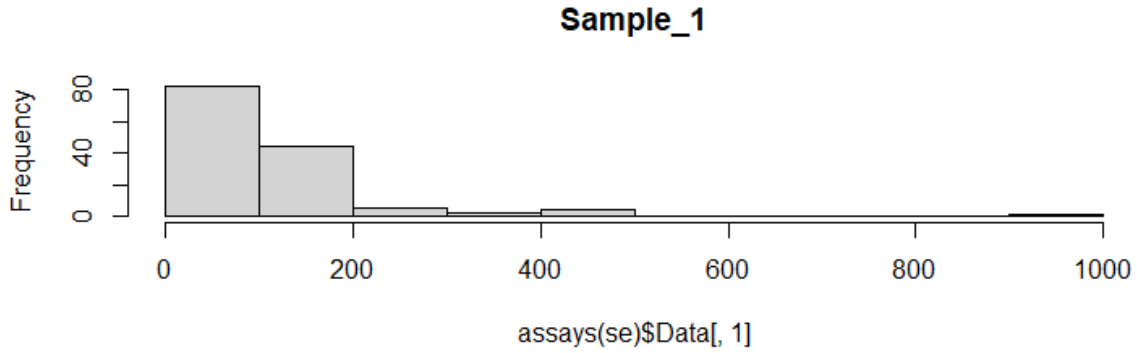


Figura 8

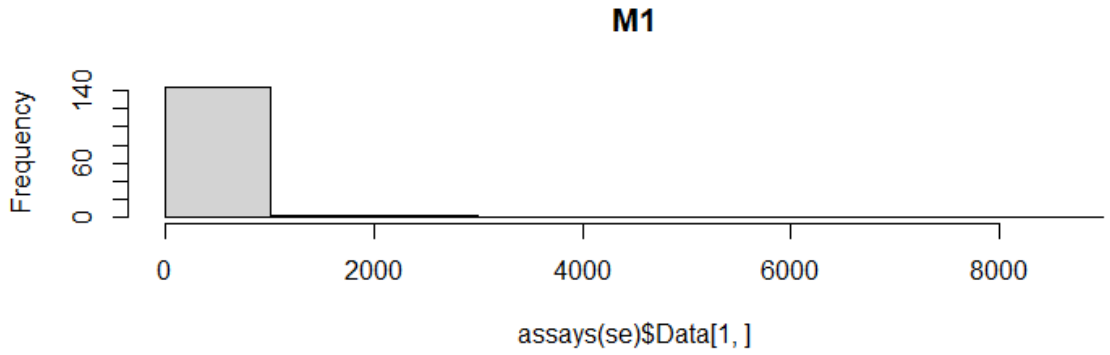


Figura 9

Histogrames de la primera mostra i del primer metabòlit, per canviar de mostra o metabòlit només s'ha de canviar pel seu nombre entre []. La majoria dels metabòlits de la mostra 1 tenen valors d'entre 0 i 200 i la majoria de les mostres del primer metabòlit es mouen per davall dels 1000.

```
> DataNum <- scale(assays(se)$Data, center = TRUE, scale=FALSE)
> apply(DataNum,2, mean)
           M1            M2            M3            M4            M5
-4.820308e-15 -2.892032e-14 -2.237992e-15  1.064594e-15 -6.383584e-15
           M6            M7            M8            M9           M10
 6.320589e-16 -4.312560e-15 -6.497766e-15  4.966724e-15 -4.114938e-15
```

Figura 10

```
> n<- dim(assays(se)$Data)[1]
> S<-cov(DataNum)*(n-1)/n
> show(S)
                M1           M2           M3           M4           M5
M1    13424.894253   50573.4074   5486.10976  -249.914324   8764.7638
M2    50573.407410 5666060.3002  16773.65113  1012.357404 124371.6554
M3     5486.109757   16773.6511  16391.06939   267.512062  14199.5361
M4     -249.914324    1012.3574    267.51206  1383.442325   3529.8925
M5     8764.763788  124371.6554  14199.53609  3529.892470 111493.2353
M6     1050.757972   16051.7069   987.22238   380.293556   9237.1478
```

Figura 11

D'aquesta manera es centren els valors de les dades i es realitza l'estimació de la matriu de covariància.

```
> R<-cor(DataNum)
> show(R)
                M1           M2           M3           M4           M5
M1    1.0000000000  0.183369112  0.3698327875 -0.0579902770  0.226548150
M2    0.1833691123  1.000000000  0.0550406261  0.0114343831  0.156479179
M3    0.3698327875  0.055040626  1.0000000000  0.0561770844  0.332159175
M4   -0.0579902770  0.011434383  0.0561770844  1.0000000000  0.284221343
M5    0.2265481503  0.156479179  0.3321591745  0.2842213432  1.000000000
M6    0.1929620786  0.143484615  0.1640726001  0.2175520343  0.588624515
```

Figura 12

I així s'obté la matriu de correlacions on es pot observar la correlació que hi ha entre un metabòlit i un altre. Òbviament, la diagonal ha de donar sempre 1.

```
> EIG <- eigen(S)
> show(EIG)
eigen() decomposition
$values
  [1]  1.871245e+08  5.911752e+07  4.069585e+07  1.551380e+07  6.189238e+06
```

Figura 13

Càlculs dels valors propis per després calcular les components principals.

```
> eigenVecs1 <- EIG$vectors
> PCAS1 <- DataNum %*% eigenVecs1
> head(PCAS1)
          [,1]        [,2]       [,3]       [,4]       [,5]       [,6]
[1,]  1752.632   3258.2768   633.6683  -176.2037   471.3191  2590.7230
[2,]  1027.116 -10424.3544 -1149.9566  1552.0480   859.8775   110.6202
[3,] -1172.871  -2442.0637 -4621.4276 19132.4826  -390.4971 -1506.2200
[4,]  1860.850   6213.4191  2602.7941  -539.0302   510.4216 -1122.7681
[5,]  1562.798   -261.1025 -2489.3300  -614.6194   261.0626   614.5622
[6,]  1771.452 -12118.3113 13537.8466  -619.0373 -1586.4311   250.5068
```

Figura 14

Després es multiplica la matriu original per la matriu de vectors propis.

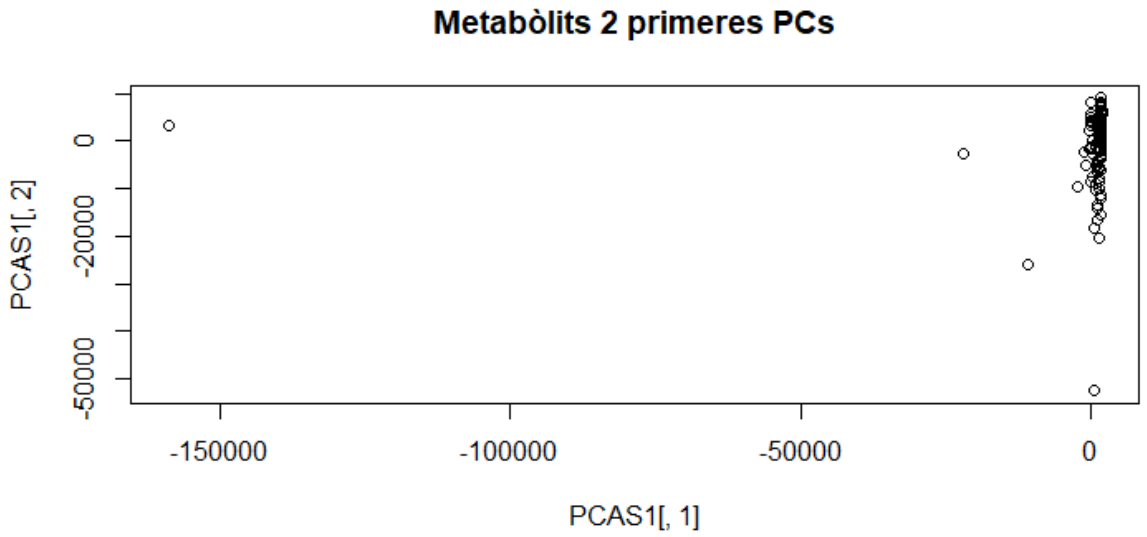


Figura 15

Finalment es grafiquen les dues components principals. La majoria dels valors es resultats es situen agrupats menys 4 que s'allunyen molt.

```
> vars1<- EIG$values/sum(EIG$values)
> round(vars1,3)
 [1] 0.544 0.172 0.118 0.045 0.018 0.017 0.015 0.012 0.010 0.007 0.006 0.006
```

Figura 16

S'obtenen els percentatges dels components principals. La primera component principal explica el 54% de la variabilitat, la segona, un 17% i la tercera, un 12%.

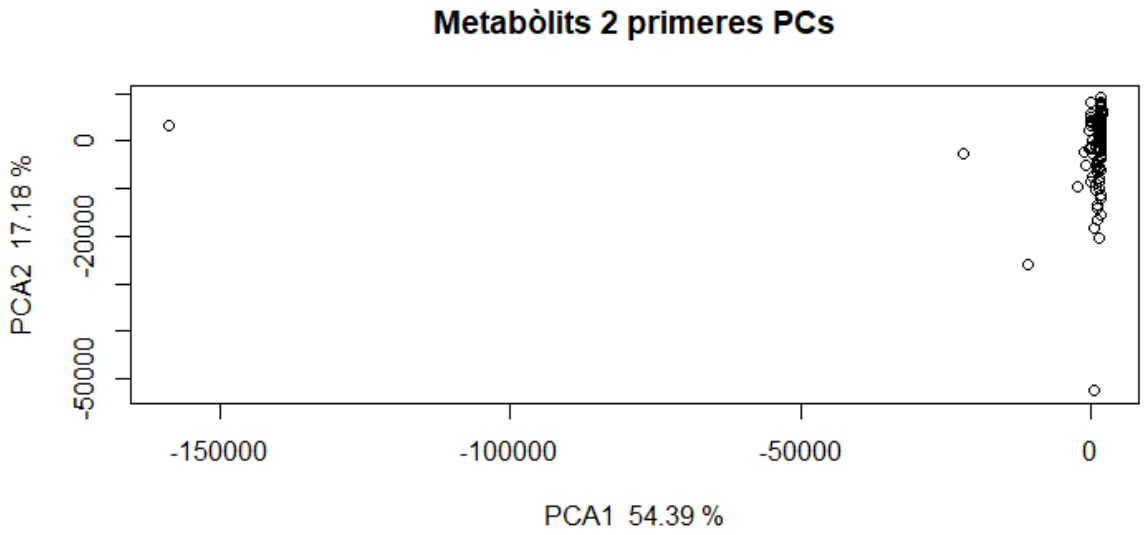


Figura 16

I es grafiquen els percentatges dels dos primers components principals.

5. DISCUSSIÓ

SummarizedExperiment és una bona eina per tractar dades juntament amb les seves metadades, ja que t'assegures del fet que quan modifiquis alguna cosa, es modifiqui en tots els arxius. Per exemple, si decideixes eliminar una columna dels resultats, també s'elimina la fila corresponent de l'arxiu que conté la informació d'aquell paràmetre.

En el meu cas, s'han hagut de modificar una mica les dades abans de crear aquest objecte perquè hi havia molts de valors nuls (NA) i algunes columnes que no ens interessaven en aquest moment.

Una vegada creat l'objecte *SummarizedExperiment* s'ha pogut recórrer correctament a la informació de cada classe, però a l'hora d'extreure el resum estadístic, en haver-hi 149 variables, és complicat poder treure'n conclusions. Pot ser, s'haurien d'ordenar per mitjanes o fer un estudi estadístic per veure aquelles que tenen més importància o que estan correlacionades entre elles i no fa falta que es tenguin en compte.

El mateix passa amb els histogrames, només podem observar uns quants histogrames a la vegada per poder-los comparar. Les matrius de covariància i correlació també ens servirien per observar les relacions entre totes les variables, però són difícils d'analitzar al ser matrius de 149x149.

Pel que fa al PCA, hi ha valors que s'allunyen molt de la gran majoria, s'hauria de revisar quins valors són i veure si són anormals o es poden obviar per treure un altre gràfic en què es pugui veure millor la dispersió que tenen la resta de la multitud. Finalment, s'obté una bona anàlisi de components principals, ja que el primer ja explica més de la meitat de la variabilitat.

6. CONCLUSIONS

- L'estudi consta de 149 metabòlits en 140 mostres
- La majoria de les mostres s'agrupen quan es tenen en compte els dos primers components principals
- Hi ha 4 mostres que s'allunyen de la resta
- Més del 50% de la variabilitat de les mostres es pot explicar amb la primera component principal

7. REFERÈNCIES

https://github.com/MarionaCasasnovas/Casasnovas_MariaTeresa_PAC1