Espinosa_delAlba_Maria_PAC1

Maria Espinosa

2024-11-01

```
library(BiocManager)
library(SummarizedExperiment)

library(readr)
library(dplyr)

library(MASS)

GastricCancer_NMR <- read.csv("C:/Users/MARIA
ESPINOSA/Documents/Universitat/MASTER ESTADISTICA/Analisi dades
omiques/GastricCancer_NMR.csv", sep=";" )
str(GastricCancer_NMR)
```

```
## 'data.frame':    140 obs. of  153 variables:
##  $ Idx       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ SampleID  : chr  "sample_1" "sample_2" "sample_3" "sample_4" ...
##  $ SampleType: chr  "QC" "Sample" "Sample" "Sample" ...
##  $ Class     : chr  "QC" "GC" "BN" "HE" ...
##  $ M1        : num  90.1 43 214.3 31.6 81.9 ...
##  $ M2        : num  491.6 525.7 10703.2 59.7 258.7 ...
##  $ M3        : num  202.9 130.2 104.7 86.4 315.1 ...
##  $ M4        : num  35 NA 46.8 14 8.7 18.7 NA 18.2 8.4 36 ...
##  $ M5        : num  164.2 694.5 483.4 88.6 243.2 ...
##  $ M6        : num  19.7 114.5 152.3 10.3 18.4 ...
##  $ M7        : num  41 37.9 110.1 170.3 349.4 ...
##  $ M8        : num  46.5 125.7 85.1 23.9 61.1 ...
##  $ M9        : num  17.3 57.8 238.3 NA 12.2 ...
##  $ M10       : num  106.8 NA 48 NA 72.9 ...
##  $ M11       : num  61.7 490.6 2441.2 140.7 48.7 ...
##  $ M12       : num  75.3 203.4 100 12.6 57.3 ...
##  $ M13       : num  79.7 330.8 873.3 46.3 140.1 ...
##  $ M14       : num  35.3 NA 29.3 62.9 77.8 52.3 34.6 30.3 61.9 28.4
...
##  $ M15       : num  28.8 210.7 45.4 38.3 51 ...
##  $ M16       : num  245.9 150.3 226.4 49.7 71.3 ...
##  $ M17       : num  5.8 71.5 36.3 0.4 18.8 ...
##  $ M18       : num  122.2 553.1 371.9 31.3 265 ...
##  $ M19       : num  90.9 217.7 98.1 38.6 84.7 ...
##  $ M20       : num  47 207.9 116.5 53.2 78.8 ...
##  $ M21       : num  49 238.8 30.3 45.4 141.3 ...
##  $ M22       : num  37.2 297.2 24.6 3.7 58 ...
##  $ M23       : num  422 591 593 130 722 ...
##  $ M24       : num  155.1 446.4 232.6 1.6 174.7 ...
##  $ M25       : num  21.4 28.3 35.1 26.6 58.9 29.1 57.7 6.6 39.6 26.2
```

```
...
##  $ M26       : num  16.5 47.9 26.8 10.3 65.9 20.3 53.5 29.3 20.8 15.4
...
##  $ M27       : num  NA 126.8 78.4 22.8 2.8 ...
##  $ M28       : num  19.9 12.9 3772.3 14.5 15.3 ...
##  $ M29       : num  65 291.9 144.3 25.5 63.5 ...
##  $ M30       : num  22 58.5 52.8 11.2 38.1 ...
##  $ M31       : num  18.9 1336.6 0.2 4.8 11.2 ...
##  $ M32       : num  97.8 621.2 360.1 111.6 233.6 ...
##  $ M33       : num  274 777 532 133 328 ...
##  $ M34       : num  26.3 324.5 507 37.3 79.4 ...
##  $ M35       : num  110.8 282.8 3207.5 NA 7.1 ...
##  $ M36       : num  13 154.3 161.7 34.5 305.9 ...
##  $ M37       : num  33.7 403.3 21.7 31 12.2 ...
##  $ M38       : num  352.3 11.7 947.7 60.5 327.2 ...
##  $ M39       : num  34.7 66.1 185.6 14.7 23.3 ...
##  $ M40       : num  33.8 50.9 124.3 28.8 87.3 ...
##  $ M41       : num  69.9 NA 80 80.1 84.7 NA 96.1 1.1 444 41.7 ...
##  $ M42       : num  47.7 226.8 129.5 67 116.2 ...
##  $ M43       : num  7.9 46.2 40.9 11 40.7 19.3 10 21 26.9 7.2 ...
##  $ M44       : num  119.1 45.4 79.4 34.8 38.4 ...
##  $ M45       : num  632 5354 1231 1679 2036 ...
##  $ M46       : num  403.7 109.3 382.8 9.7 278.7 ...
##  $ M47       : num  47.4 NA NA 38.6 99.2 ...
##  $ M48       : num  8030 16745 12939 4563 12562 ...
##  $ M49       : num  91.3 775.4 464.5 112.2 390.7 ...
##  $ M50       : num  98.1 563.5 512.4 67.6 197.8 ...
##  $ M51       : num  372 521 NA 263 576 ...
##  $ M52       : num  148 307 198 135 191 ...
##  $ M53       : num  562 1292 6414 131 760 ...
##  $ M54       : num  0.3 4.6 26.7 1 2.3 0.4 18 NA 30.6 2.1 ...
##  $ M55       : num  102.6 604.3 19.7 36.2 70.4 ...
##  $ M56       : num  74.1 140.3 145.8 44.7 148.6 ...
##  $ M57       : num  200 1159 9436 142 612 ...
##  $ M58       : num  196 1631 1155 160 490 ...
##  $ M59       : num  474.5 590.9 6537.7 74.6 619.4 ...
##  $ M60       : num  170.9 798.6 2813.2 38.7 390.7 ...
##  $ M61       : num  565 1216 3118 264 1036 ...
##  $ M62       : num  232 786 NA 196 442 ...
##  $ M63       : num  253 2966 12077 512 824 ...
##  $ M64       : num  116 490 612 310 580 ...
##  $ M65       : num  263 1416 1798 157 663 ...
##  $ M66       : num  3104 3168 689 632 507 ...
##  $ M67       : num  94.5 601 531.3 55.2 121.4 ...
##  $ M68       : num  129 257 125 265 232 ...
##  $ M69       : num  0.6 288 112.7 29.9 73.3 ...
##  $ M70       : num  13.8 68 194.8 14 42.3 ...
##  $ M71       : num  23.8 232.6 51.5 16.2 32.9 ...
##  $ M72       : num  170 428 243 NA 250 ...
##  $ M73       : num  134.3 260 214 18.6 79.2 ...
```

```
##  $ M74       : num  9.9 88.2 46.2 9.2 20 1.6 19.3 6.2 29.1 13.9 ...
##  $ M75       : num  75.7 408.1 395.6 100.1 503.2 ...
##  $ M76       : num  607 1127 2941 73 708 ...
##  $ M77       : num  452 799 1532 NA 775 ...
##  $ M78       : num  0.1 46 NA 1.6 NA 40.7 34 1.3 31.9 NA ...
##  $ M79       : num  20.2 199.6 NA NA NA ...
##  $ M80       : num  456 205 205 142 251 ...
##  $ M81       : num  314 1400 2943 304 426 ...
##  $ M82       : num  NA 121.7 147.8 0.3 38.5 ...
##  $ M83       : num  NA 8708 NA 198 214 ...
##  $ M84       : num  8.8 108.4 18.7 67.7 30.2 ...
##  $ M85       : num  35.2 125.5 85.8 0.7 5.8 ...
##  $ M86       : num  40.2 162 91.4 54.1 695.7 ...
##  $ M87       : num  16.8 28.3 99.3 19.2 23.7 ...
##  $ M88       : num  136.3 319.6 396 75.2 216.2 ...
##  $ M89       : num  384 3712 694 180 486 ...
##  $ M90       : num  28.6 285.6 277.9 26.3 75.8 ...
##  $ M91       : num  60 337.5 61.3 26.2 65.2 ...
##  $ M92       : num  10.1 84.3 NA 12.1 20.4 18.5 36.5 0.1 43.9 10.7 ...
##  $ M93       : num  69.8 120.4 236.4 50.7 111.9 ...
##  $ M94       : num  99.3 2383.3 579 119.6 569.7 ...
##  $ M95       : num  13.8 97.2 NA NA 36.8 NA 80.1 12.5 17.1 37.6 ...
##   [list output truncated]
```

```
View(GastricCancer_NMR)

GastricCancer_NMR$Class<-as.factor(GastricCancer_NMR$Class)


GastricCancer_NMR_metadada <- read.csv("C:/Users/MARIA
ESPINOSA/Documents/Universitat/MASTER ESTADISTICA/Analisi dades
omiques/GastricCancer_NMR_metadada.csv", sep=";" )
View(GastricCancer_NMR_metadada)
str(GastricCancer_NMR_metadada)
```

```
## 'data.frame':    149 obs. of  5 variables:
##  $ Idx         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Name        : chr  "M1" "M2" "M3" "M4" ...
##  $ Label       : chr  "1_3-Dimethylurate" "1_6-Anhydro-?-D-glucose"
"1_7-Dimethylxanthine" "1-Methylnicotinamide" ...
##  $ Perc_missing: chr  "11,42857143" "0,714285714" "5" "8,571428571"
...
##  $ QC_RSD      : chr  "32,20800495" "31,17802801" "34,99060496"
"12,80420087" ...
```

```
save(GastricCancer_NMR, file = "GastricCancer_NMR.Rda")
save(GastricCancer_NMR_metadada, file = "GastricCancer_NMR_metadada.Rda")
```

Entre els datasets de metabolòmica proposats he escollit el “Gastric Cancer”. Després de descarregar-lo, he guardat les dues pestanyes per separat en format .csv, a més he preprocessat aquelles dades que estaven amb el format decimal “,”

i ho he substituït per “.”. Posteriorment, he carregat les diferents dades i metadades en RStudio.

Les dades utilitzades son mostres d’orina corresponents a 43 pacients amb càncer gàstric, 40 amb malalties gàstriques benignes i 40 pacients sans. Utilitzant espectroscòpia de ressonància magnètica nuclear d’hidrogen s’han obtingut variables respostes de metabòlits.

En la base de dades “GastricCancer_NMR” es troben 153 columnes de variables, que inclouen 3 variables categòriques: “SampleID, SampleType i Class”, aquestes conformen les metadades de les mostres. La resta son valors de diferents metabòlits dels pacients estudiats. El segon arxiu carregat “GastricCancer_NMR_metada” conté les metadates dels metabòlits estudiats, conté per exemple el nom (“Name”) que s’utilitza en el primer document i el metabòlit al que correspon (“label”).

```
data<-as.matrix(GastricCancer_NMR[,5:153])#Així guardo les dades dels metabòlits.
metadata_row<-as.data.frame(GastricCancer_NMR[,1:5]) #Així guardo les metadades de les mostres.

metadata_column<-as.data.frame(GastricCancer_NMR_metadada) #Així guardo les metadades dels metabòlits.
se <- SummarizedExperiment(assays = list(counts = data),
                           colData = metadata_column,
                           rowData = metadata_row)
se

## class: SummarizedExperiment
## dim: 140 149
## metadata(0):
## assays(1): counts
## rownames: NULL
## rowData names(5): Idx SampleID SampleType Class M1
## colnames(149): M1 M2 ... M148 M149
## colData names(5): Idx Name Label Perc_missing QC_RSD
```

Primerament he generat el conjunt de dades de variable resposta en format de matriu, seleccionant les columnes que hi corresponen. Seguidament he seleccionat les columnes que corresponen a les metadades de les mostres i ho he guardat com un data frame. Per últim, des del segon fitxer carregat les dades com un data frame, adjudicant-les amb el nom de metadades dels metabòlits estudiats.

Amb la classe SummarizedExperiment he generat un dataframe que conté diferents capes, que corresponen a les dades, les metadades de les mostres i les metadades dels metabòlits.

```
library(FactoMineR)
library(factoextra)
```

```r
library(tibble)
GastricCancer_NMR[is.na(GastricCancer_NMR)] <- 0
pca.GC <- prcomp(GastricCancer_NMR[,5:153], scale = TRUE)
summary(pca.GC)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4    PC5     PC6
PC7
## Standard deviation     6.2554 3.4503 2.90580 2.65042 2.3574 2.26951
2.14152
## Proportion of Variance 0.2626 0.0799 0.05667 0.04715 0.0373 0.03457
0.03078
## Cumulative Proportion  0.2626 0.3425 0.39918 0.44633 0.4836 0.51819
0.54897
##                           PC8     PC9    PC10    PC11    PC12    PC13
PC14
## Standard deviation     1.930 1.86193 1.78120 1.72056 1.63905 1.55116
1.53874
## Proportion of Variance 0.025 0.02327 0.02129 0.01987 0.01803 0.01615
0.01589
## Cumulative Proportion  0.574 0.59724 0.61854 0.63840 0.65643 0.67258
0.68847
##                           PC15    PC16    PC17    PC18    PC19    PC20
PC21
## Standard deviation     1.50157 1.43960 1.40321 1.36820 1.33592 1.32287
1.27819
## Proportion of Variance 0.01513 0.01391 0.01321 0.01256 0.01198 0.01174
0.01096
## Cumulative Proportion  0.70361 0.71751 0.73073 0.74329 0.75527 0.76702
0.77798
```

```r
valors<-pca.GC$rotation
valors1 <- rownames_to_column(as.data.frame(pca.GC$rotation), var =
"metabòlits")
valors1<- as.data.frame(valors1)
contribucio_pc1 <- valors1[, 1:2]
arrange(contribucio_pc1, PC1)
```

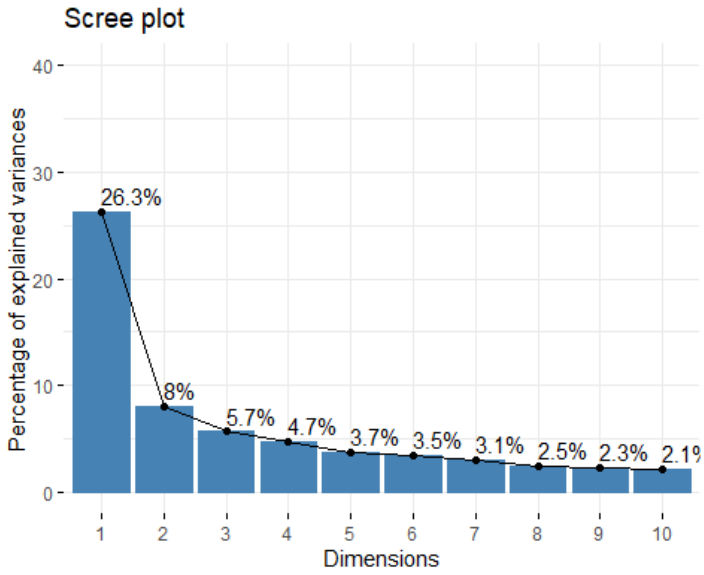
```
##     metabòlits          PC1
## 1          M65 -0.143226547
## 2          M88 -0.141860168
## 3         M104 -0.138650512
## 4          M74 -0.138443961
## 5         M107 -0.133071344
## 6          M90 -0.132440312
## 7          M48 -0.126972639
## 8         M114 -0.125598222
## 9          M12 -0.123862879
## 10          M5 -0.122523391
## 11        M122 -0.122449471
```

```
## 12          M62 -0.120512125
.

.

.

## 143         M96 -0.017578644
## 144         M30 -0.016910734
## 145         M60 -0.010976697
## 146        M144 -0.006322825
## 147         M29 -0.003426590
## 148         M80  0.003525438
## 149         M13  0.004844785
```

Un PCA és un mètode que ens permet reduir les dimensions de la nostra base de dades. Duu a terme una transformació de les variables de la nostra matriu de dades que reflexa les diferents fonts de variabilitat de les dades però no estan correlacionades. Aquestes noves components es construeixen de tal forma que tenen una capacitat explicativa decreixent, és a dir, cada component explica més que la següent. Això ens permet quedar-nos amb les primeres components i descartar la resta. El resultat és que podem utilitzar els valors de les primeres components per obtenir un representació de les dades en dimensions reduïdes.

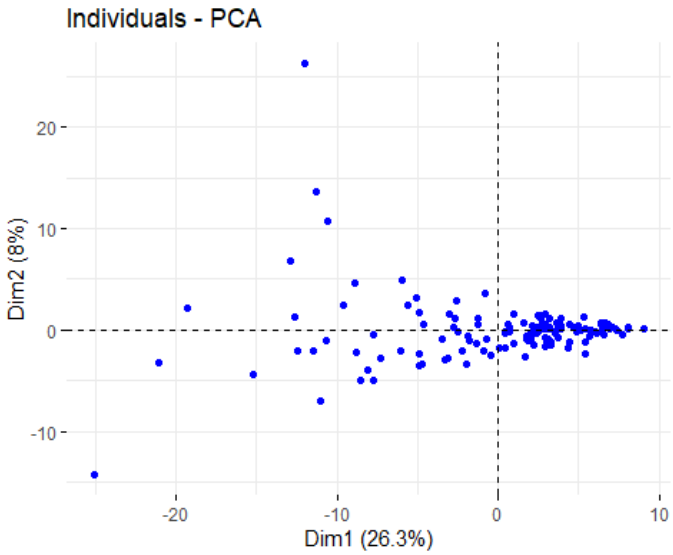
Abans de realitzar la nostra PCA, substituïm els “NAs” que puguin haver en la nostra matriu de dades per 0. Seguidament generem el PCA. El resum ens mostra la desviació estàndard, la proporció de variància explicada per cada component i l’acumulada. Com s’observa, la primera component ens explica un 26,26% de la variabilitat de les dades però la segona només un 7,99% per tant, les dues primeres components principals expliquen només un 34,25% de la variabilitat, és bastant baix. A més, degut a que les components principals son una variable formada per una combinació lineal, amb diferents pesos, de les variables originals, és interessant saber quines són les variables més influents. Les variables més influents son les que tinguin un pes, en valor absolut, més alt. En la primera component principal són els metabòlits: Glycylproline, N-Acetylglutamine, Proline i Isoleucine.

```r
fviz_screeplot(pca.GC, addlabels = TRUE, ylim = c(0, 40))
```



El gràfic ens pot servir per decidir amb quantes components principals quedar-nos al reduir les dimensions de les nostres dades. Veiem que la primera component és la que explica un percentatge major però la resta expliquen un percentatge molt inferior i bastant similars. Per tant, podríem quedar-nos amb les dues primeres components.

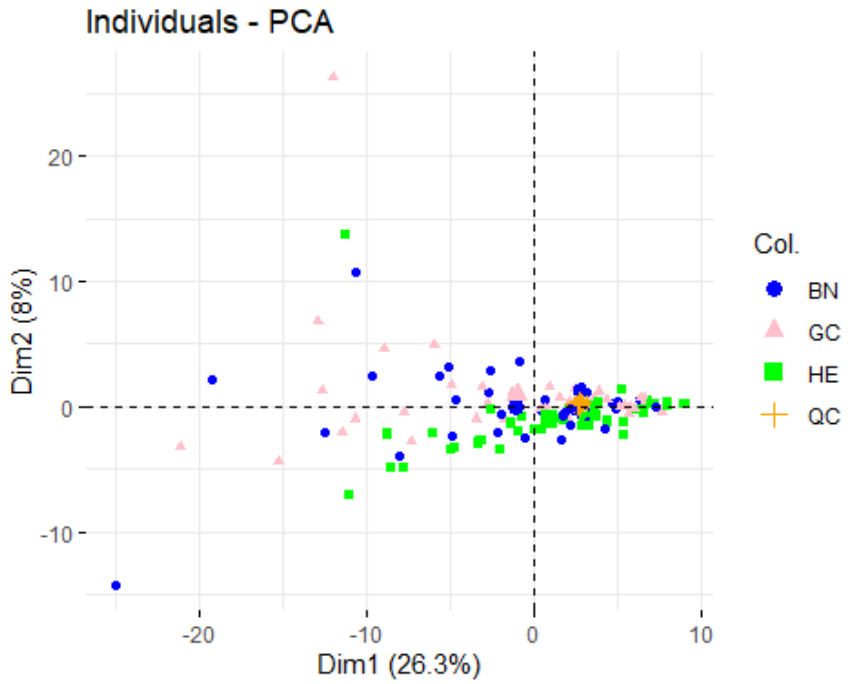
```
fviz_pca_ind(pca.GC, geom.ind = "point",
             col.ind = "blue",
             axes = c(1, 2),
             pointsize = 1.5)
```



El segon gràfic ens mostra com es distribueixen les dades en aquests espai de dues dimensions formades per les dues components principals. Veiem que

bàsicament es distribueixen al llarg de l'eix de les x. Fet que confirma el que mostraven la variabilitat explicada per les components principals.

```
fviz_pca_ind(pca.GC,
             geom.ind = "point", # mostrar només puntss
             col.ind = GastricCancer_NMR$Class  ,    # color per grups
             palette = c("blue", "pink", "green", "orange"),
             addEllipses = FALSE,
)
```



Per últim, al tercer gràfic, veiem el mateix que l'anterior però amb les classes que teníem de pacients sans, amb càncer maligne i amb tumors benignes. No observem grups diferenciats segons la classe de malaltia en funció d'aquestes dues components principals.

En conclusió, hem utilitzat una base de dades que conté un nombre de variables molt gran, tenim informació sobre les mostres que tenim (quin tipus de classe dins de la malaltia son) i tenim dades sobre els metabòlits que son la variable resposta. Ens interessa reduir les dimensions d'aquestes dades, és a dir, a través d'una PCA intentem crear unes noves variables explicatives independents les quals estan formades per una combinació lineal de totes les variables originals. Aquestes noves variables o components principals ens expliquen de manera decreixen la variabilitat de les nostres dades (només un 34% entre les dues primeres components). És interessant veure quins metabòlits tenen un pes major en la combinació lineal de la primera component principal i seria interessant estudiar per què son aquests compostos els que tenen un valor més gran. I finalment, he intentat veure si en aquest nou espai de dues dimensions les dades

es distribuïen en grups segons la classe de malaltia a la que pertanyien, però no es veu cap patró clar.

La direcció del repositori de github2 que conté : l' informe, l'objecte contenidor amb les dades i les metadades en format binari de R (arxiu amb extensió . Rda), el codi R per a l' exploració de les dades I dades en format text i o les metadades sobre el dataset en un arxiu markdown: https://github.com/mariaesdal/Espinosa_delAlba_Maria_PEC1