UNIVERSITAT DE BARCELONA

Universitat Oberta de Catalunya

Estadística Multivariante
Ejercicio de recuperación de la PEC 1

Alex Sánchez y Francesc Carmona

2 de Enero de 2020
Fecha límite de entrega de la solución: 17-01-2021

Ejercicio para personas que suspendieron la PEC 1 exclusivamente.

El proyecto TCGA (*The Cancer Genome Atlas*) realizó un exhaustivo estudio con el objetivo de proporcionar una base de datos pública y abierta con datos ómicos y clínicos de de más de una docena de tipos de cánceres.

Para este ejercicio hemos descargado y homogeneizado dos archivos de datos, uno con las variables clínicas y otros dos con datos ómicos uno de expresión génica y proteómica. Los datos de expresión génica se han filtrado quedándonos con los 171 genes de mayor variabilidad, mientras que los datos de proteómica no se ha hecho ésto porque el número de muestras era muy inferior.

En total disponemos de la información siguiente:

```
## [1] "clinsSmall"  "expresSmall" "protsSmall"
```

```
## expresSmall:
```

```
## [1] 178 166
```

```
##         TCGA-A2-A0CQ TCGA-A7-A0CD TCGA-A8-A07E TCGA-A8-A09E TCGA-A8-A09K
## AGR3        6.273750     4.618250     4.618250     6.165000       4.9890
## MUCL1       0.854000    -0.817500     5.740000    -0.003500      -0.0965
## NPY1R       7.045833     4.529167     1.713000     2.314833       1.3290
## CEACAM6     3.236300     5.702800     3.127300     6.659100      -3.7394
## SYT13       6.183500     5.604500     1.771750     4.468750       4.4780
## TFAP2B      3.857000     3.956000     3.509375    -1.592875      -2.9685
```

```
##
## protsSmall:
```

```
## [1] 226 166
```

```
##                 TCGA-A2-A0CQ TCGA-A7-A0CD TCGA-A8-A07E TCGA-A8-A09E
## 14-3-3_beta     -0.116435332  -0.07586002    0.4456241  0.033968109
## 14-3-3_epsilon   0.206356661  -0.11724431    0.3711343  0.000572023
## 14-3-3_zeta     -0.216609941   0.22568722   -0.3858968 -0.159102488
## 4E-BP1          -0.099411460  -0.41324482   -0.7331879 -0.339753841
## 4E-BP1_pS65     -0.004717038  -0.49765351   -0.1386842 -0.261990167
## 4E-BP1_pT37_T46 -0.321369575   0.09030310    0.1851034  0.478754014
##                 TCGA-A8-A09K
```

```
## 14-3-3_beta       0.14142004
## 14-3-3_epsilon   -0.11254593
## 14-3-3_zeta      -0.11388628
## 4E-BP1           -0.32189563
## 4E-BP1_pS65       0.09483988
## 4E-BP1_pT37_T46   0.49680972
```

```
##
## clinsSmall:
```

```
## [1] 166  16
```

```
##              years_to_birth vital_status days_to_last_followup
## TCGA-A2-A0CQ             62            0                  2695
## TCGA-A7-A0CD             66            0                  1165
## TCGA-A8-A07E             81            0                   608
## TCGA-A8-A09E             73            0                  1492
## TCGA-A8-A09K             68            0                   912
##              tumor_tissue_site pathologic_stage
## TCGA-A2-A0CQ            breast         stage ia
## TCGA-A7-A0CD            breast          stage i
## TCGA-A8-A07E            breast          stage x
## TCGA-A8-A09E            breast       stage iiib
## TCGA-A8-A09K            breast        stage iia
```

Las variables registradas son

	Descripción
years_to_birth	Edad
vital_status	Ha fallecido?
days_to_last_followup	Dias desde la última revisión
tumor_tissue_site	Tejido en que se localiza el tumor
pathologic_stage	Estadío patológico
pathology_T_stage	Estadío tipo T (extendido)
pathology_N_stage	Estadío tipo N
pathology_M_stage	Estadío tipo M
gender	Género
date_of_initial_pathologic_diagnosis	Fecha (año) del diagnóstico
radiation_therapy	Radioterapia
histological_type	Tipo histoógico
number_of_lymph_nodes	Número de nodos linfáticos
race	Raza
ethnicity	Etnia
Estadio_T	Estadío tipo T (simplificado)

P1. Realizar un resumen numérico y gráfico de los datos clínicos. Observar si hay valores faltantes. ¿Recomendaríais eliminar alguna variable? Podéis usar la siguiente codificación para separar las variables clínicas que son factores de las que son numéricas:

```
cols2Factor <- c("vital_status", "tumor_tissue_site", "pathologic_stage",
                 "pathology_T_stage", "pathology_N_stage", "pathology_M_stage",
                 "gender", "radiation_therapy","histological_type",
                 "race", "ethnicity", "Estadio_T")
cols2Integer <-  c("years_to_birth","days_to_last_followup",
                   "date_of_initial_pathologic_diagnosis","number_of_lymph_nodes")
clinsFactor <- clinsSmall[,cols2Factor]
clinsNum <- clinsSmall[,cols2Integer]
```

P2. Deseamos seleccionar los 10 genes y las 10 proteínas que presenten una mayor diferencia mejor entre el estadio T1 y los demás que llamaremos "TSup".

- Escribir una función que permita seleccionar dichos genes y dichas proteínas y
- Construir una tabla con los estadísticos que permiten diferenciarlos ordenados de mayor a menor, para genes y otra para proteínas.
- Crear dos subconjuntos "top10Expres" y "top10Prots" con las expresiones de dichos genes y proteínas.

P3. Hacer un gráfico multivariante de las correlaciones dos a dos con los cinco primeros genes, las cinco primeras proteínas y con una matriz combinada de las dos anteriores

P4. Crear una matriz de distancias usando como distancia "1-*la correlación*" y visualizad las distancias entre genes y proteínas con un mapa de colores agrupado por similaridades como el que ofrece la función `heatmap` o similares.

P5. Usando las cinco proteínas más diferencialmente expresadas calcular el vector de medias y la matriz de covarianzas por separado para los dos grupos (niveles del factor `Estadio`). ¿Concuerdan los resultados con lo que esperaríais por la forma en que habéis seleccionado los genes/proteínas?

P6. Realizar un análisis de componentes principales basado en la matriz de los 10 genes y la de las 10 proteínas. ¿Cual os parece más adecuada para representar las muestras en dimensión reducida?

P7. *Con las componentes principales que retengan, desde vuestro punto de vista, más variabilidad en las primeras componentes* interpretar los dos primeros ejes principales mediante el gráfico de correlaciones con las variables originales.

Nota: Observad que para interpretar adecuadamente las dos o tres primeras componentes deberíais recoger información de los genes. Esto puede hacerse mediante instrucciones de `R/Bioconductor` o buscándolos en la base de daatos de NCBI, tal como habréis aprendido en asignaturas como *Genómica Computacional* o *Análisis de datos ómicos*).

P8. A partir del análisis realizado representar las muestras en un gráfico de dos dimensiones con puntos distintos según el Estadio. ¿Podemos interpretar las componentes en función de la gravedad del tumor?

Para finalizar realizaremos algunas comparaciones multivariables. Para ello volveremos a utilizar los estadios originales eliminando las muestras de proteómica con NAs así como la que pertenece a un estadio desconocido ("tX").

Además, con el fin de tener más muestras que variables eliminaremos aquellas proteínas con menor variabilidad, quedándonos únicamente con el 33 % de proteínas más variables. El código para preparar los datos es el siguiente:

```r
noNAprotsSmall <- protsSmall[complete.cases(protsSmall),]
dim(noNAprotsSmall)
noNAClins <- clinsSmall[colnames(noNAprotsSmall),]
clins2 <- noNAClins [noNAClins$Estadio_T!="tx",]
clins2$Estadio_T <- droplevels(clins2$Estadio_T)
table(clins2$Estadio_T)
prots2 <- noNAprotsSmall[, rownames(clins2)]

sds <- apply (prots2, 1, sd)
prots2b <- cbind (prots2, sds)
prots2b_sort <- prots2b[order(prots2b[,"sds"], decreasing = TRUE),]
rows2Keep <- round(nrow(prots2b)* 0.333,0)
prots2b_sort  <- prots2b_sort [1:rows2Keep,]
# dim(prots2b_sort)
prots2 <-prots2b_sort[,-166]
# dim(clins2)
```

P9. Si suponemos que se verifican todas las condiciones, realizar un MANOVA de un factor, utilizando la variable `Estadio`, para decidir si hay diferencias en las proteínas seleccionadas. ¿Qué diferencia(s) hay entre este test y el test realizado en la primera parte para seleccionar 10 proteínas?

P10. Aunque los datos normalizados de proteínas pueden considerarse aproximadamente normales, es bueno realizar algún tipo de verificación. Realizar algunos análisis gráficos y numéricos para determinar si existe normalidad multivariante.