Tutorial to PCA and PLS

Benoit Liquet[*1]

[1]Macquarie University

*benoit.liquet-weiland@mq.edu.au

Contents

1 PCA

1.1 Data

The dataset includes gene expression data for 6830 genes from 64 cancer samples (from different cancer subtypes).

Data can be downloaded form http://www-stat.stanford.edu/~tibs/ElemStatLearn/

For this analysis (and to simplify the plots), 5 subtypes with only 1 samples have been removed: UNKNOWN, K562B-repro, K562A-repro, MCF7A-repro, MCF7D-repro

1.2 Data Management

- Read the data from the txt file `nci.data.txt`

```
## [1] 6830   64
```

- What is the dimension of your Data Frame ?
- The names of the 64 samples are stored in the file `subtypes_names.Rdata`. Load the names and removed from the data frame the subtypes with only 1 samples:

UNKNOWN, K562B-repro, K562A-repro, MCF7A-repro, MCF7D-repro

Here an example:

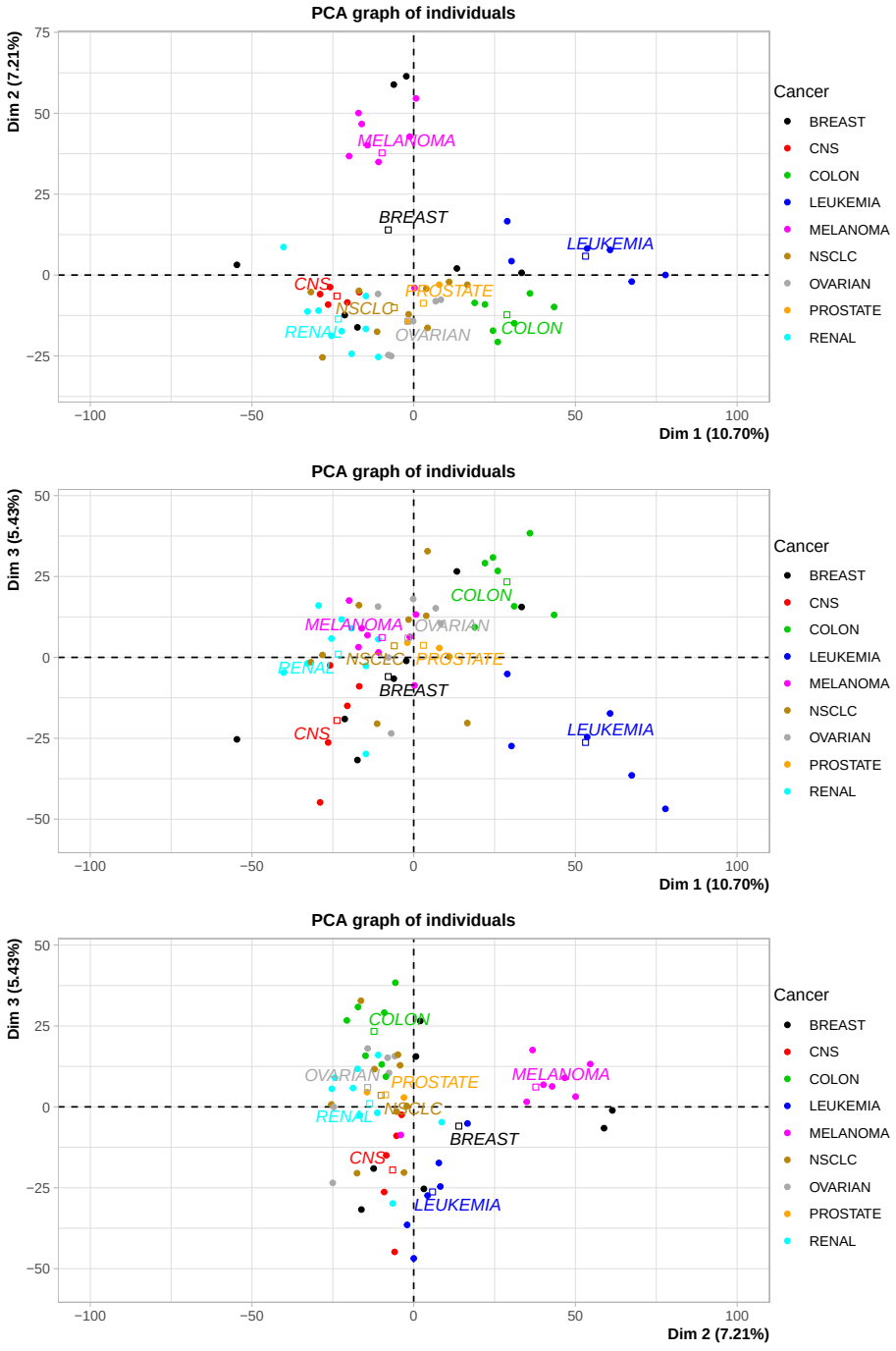
```
## Remove subtype with only 1 sample
one.sample <- c("UNKNOWN", "K562B-repro", "K562A-repro", "MCF7A-repro", "MCF7D-repro")
ind <- which(names.data%in%one.sample)
names.final <- names.data[-ind]
dat.1 <- dat.1[,-ind]
dim(dat.1)
## [1] 6830   59
dat.2 <- t(dat.1)
dim(dat.2)
## [1]   59 6830
```

- So now you are working with 59 samples.

1.3 Run a PCA

- Run a PCA
- Choice of number of component
- Project the samples on the first 3 components

You should get some plots like the following



1.4 Sparse PCA

Find a way to select the most relevant genes by using a sparse PCA.

2 PLS-DA

```
table(names.final)
## names.final
##   BREAST      CNS    COLON LEUKEMIA MELANOMA    NSCLC  OVARIAN PROSTATE
##        7        5        7        6        8        9        6        2
##    RENAL
##        9
```

We will define a binary response variable:

- 1 for subtypes cancer : Colon, Leukemia, Prostate, NSCLC
- 0 for: BREAST, CNS, MELANOMA, OVARIAN, RENAL

```
Y <- rep(0,59)
group1 <- which(names.final%in%c("COLON","LEUKEMIA","PROSTATE","NSCLC"))
Y[group1] <-1
table(Y)
## Y
##  0  1
## 35 24
```

2.1 run a PLS-DA model

2.2 choose the number of components

2.3 Project the samples on the first two components map

2.4 Run a sparse PLD-DA model

2.5 Choose the number of variables to select in each component