PCA

Marc Galland

1/21/2019

Introduction

Library: mixOmics

Install the library. Check lesson instructions if you don't know how to install the library

```
library("mixOmics",verbose = FALSE,quietly = TRUE)
```

```
##
## Loaded mixOmics 6.6.1
##
## Thank you for using mixOmics! Learn how to apply our methods with our tutorials on www.mixOmics.org,
## Questions: email us at mixomics[at]math.univ-toulouse.fr
## Bugs, Issues? https://github.com/mixOmicsTeam/mixOmics/issues
## Cite us:  citation('mixOmics')
```

Dataset used: the liver.toxicity study

The data come from a liver toxicity study (Bushel et al., 2007) in which 64 male rats of the inbred strain Fisher 344 were exposed to non-toxic (50 or 150 mg/kg), moderately toxic (1500 mg/kg) or severely toxic (2000 mg/kg) doses of acetaminophen (paracetamol) in a controlled experiment. Necropsies were performed at 6, 18, 24 and 48 hours after exposure and the mRNA from the liver was extracted. Ten clinical chemistry measurements of variables containing markers for liver injury are available for each subject and the serum enzymes levels are measured numerically. The data were further normalized and pre-processed by Bushel et al.(2007).

The `liver.toxicity` is a list in the package that contains:
- `gene`: a data frame with 64 rows and 3116 columns, corresponding to the expression levels of 3,116 genes measured on 64 rats.
- `clinic`: a data frame with 64 rows and 10 columns, corresponding to the measurements of 10 clinical variables on the same 64 rats.
- `treatment`: data frame with 64 rows and 4 columns, indicating the treatment information of the 64 rats, such as doses of acetaminophen and times of necropsy.
- `gene.ID`: a data frame with 3116 rows and 2 columns, indicating geneBank IDs of the annotated genes.

More details are available in R if you type `?liver.toxicity` after loading the `mixOmics` library.

Vocabulary

Some of the vocabulary used can be tougth sometimes: We list here the main methodological or theoretical concepts you need to know to be able to efficiently apply mixOmics:
- **Individuals, observations or samples**: the experimental units on which information are collected, e.g. patients, cell lines, cells, faecal samples…
- **Variables, predictors**: read-out measured on each sample, e.g. gene (expression), protein or OTU (abundance), weight…
- **Variance**: measures the spread of one variable. In our methods we estimate the variance of components

rather that variable read-outs. A high variance indicates that the data points are very spread out from the mean, and from one another (scattered).
- **Covariance**: measures the strength of the relationship between two variables, i.e whether they co-vary. A high covariance value indicates a strong relationship, e.g weight and height in individuals frequently vary roughly in the same way; roughly, the heaviest are the tallest. A covariance value has no lower or upper bound. - **Correlation**: a standardized version of the covariance that is bounded by -1 and 1.
- **Component**: an artificial variable built from a linear combination of the observed variables in a given data set. Variable coefficients are optimally defined based on some statistical criterion. For example in Principal Component Analysis, the coefficients in the (principal) component is defined so as to maximise the variance of the component.
- **Loadings**: variable coefficients used to define a component.
- **Sample plot**: representation of the samples projected in a small space spanned (defined) by the components. Samples coordinates are determined by their components values, or scores.
- **Correlation circle plot**: representation of the variables in a space spanned by the components. Each variable coordinate is defined as the correlation between the original variable value and each component. A correlation circle plot enables to visualise the correlation between variables - negative or positive correlation, defined by the cosine angle between the centre of the circle and each variable point) and the contribution of each variable to each component - defined by absolute value of the coordinate on each component. For this interpretation, data need to be centred and scaled (by default in most of our methods except PCA).
- **Unsupervised analysis**: the method does not take into account any known sample groups and the analysis is exploratory. Examples of unsupervised methods covered in this vignette are Principal Component Analysis (PCA, Chapter 4), Projection to Latent Structures (PLS, Chapter 6), and also Canonical Correlation Analysis (CCA, not covered here).
- **Supervised analysis**: the method includes a vector indicating the class membership of each sample. The aim is to discriminate sample groups and perform sample class prediction. Examples of supervised methods covered in this vignette are PLS Discriminant Analysis (PLS-DA, Chapter 5), DIABLO (Chapter 7) and also MINT (not covered here (F Rohart et al. 2017)).

PCA hands-on session

Loading the dataset

To illustrate PCA, we focus on the expression levels of the genes in the data frame liver.toxicity$gene.

```
library(mixOmics,quietly = TRUE)
data(liver.toxicity)
X <- liver.toxicity$gene
dim(X)                      # check the number of rows (samples) and columns (genes)
```

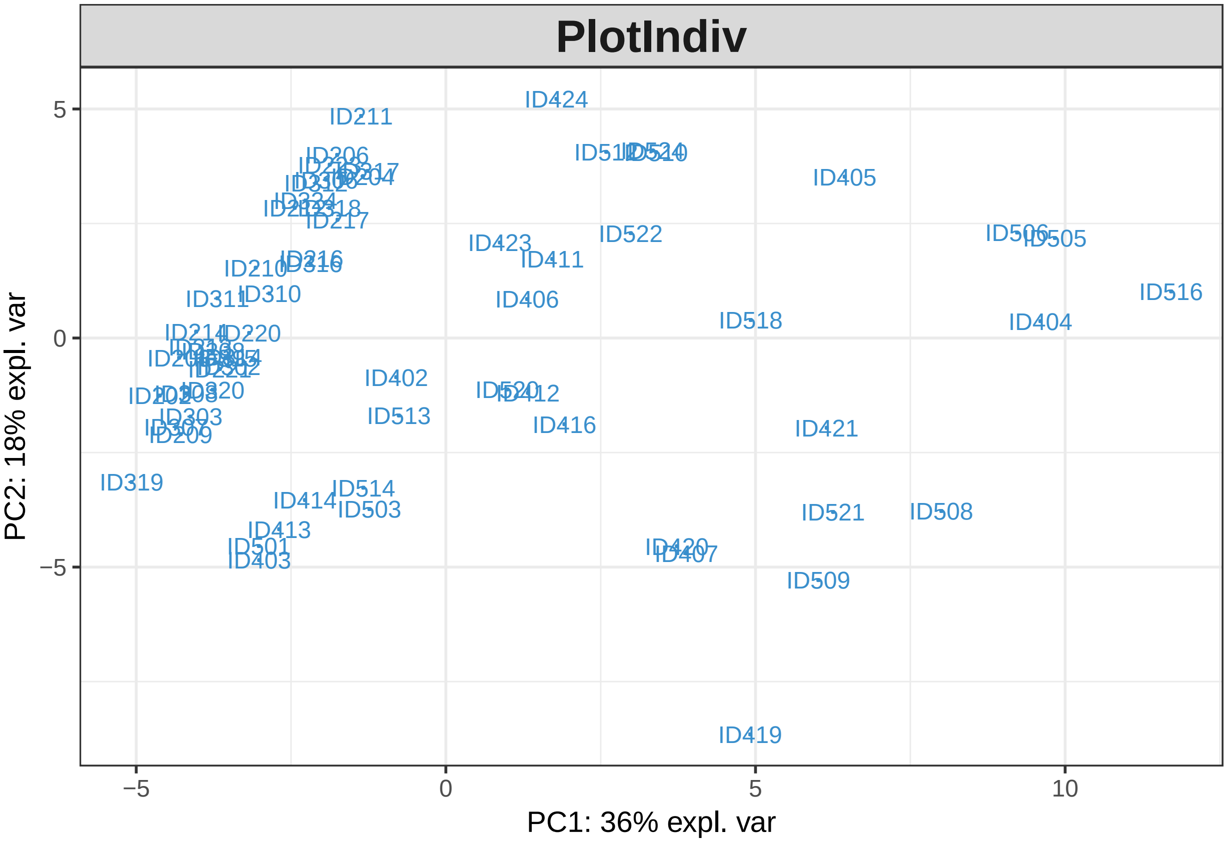
```
## [1]   64 3116
```

```
X[1:5,1:5]                  # takes a glimpse into the dataset
```

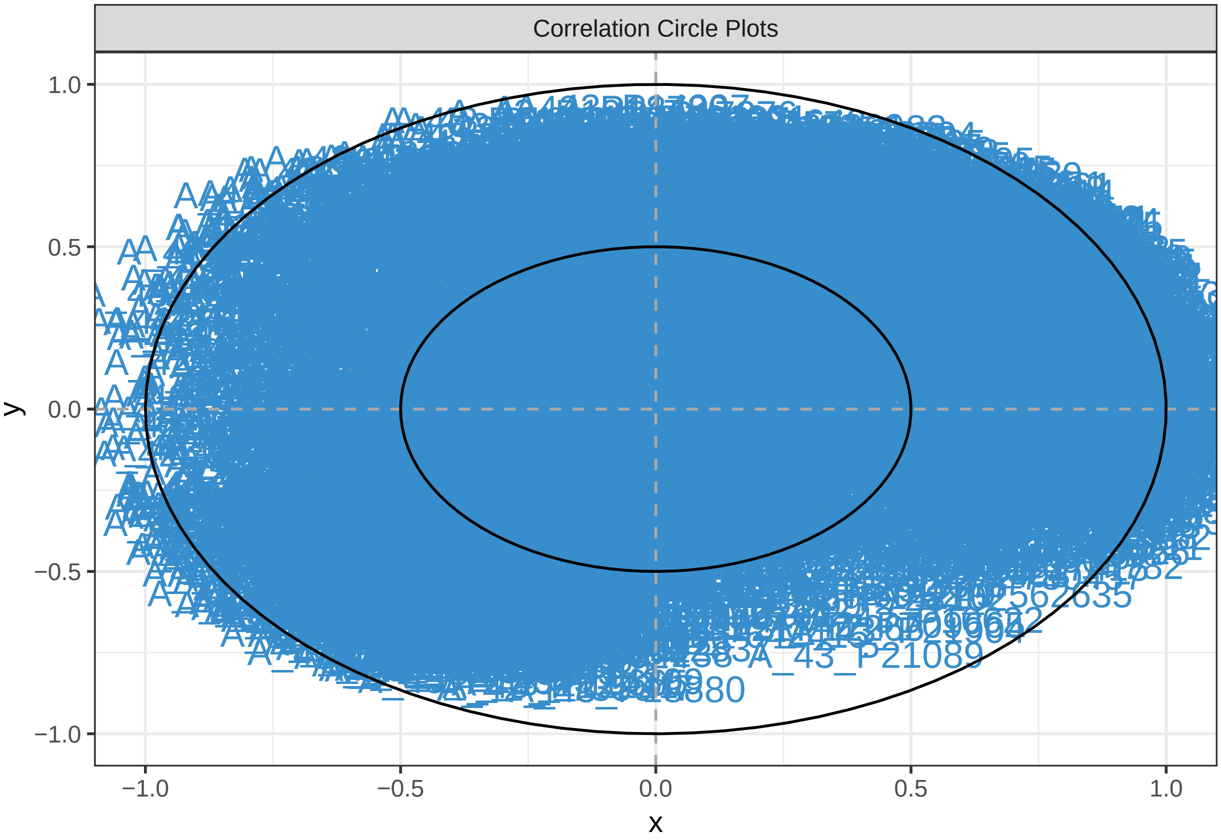
```
##       A_43_P14555 A_43_P22290 A_43_P20792 A_43_P21286 A_43_P12995
## ID202     0.05169    -0.08120     0.00617     0.00003     0.00676
## ID203     0.01548     0.17515     0.04845     0.04730     0.08869
## ID204    -0.01509    -0.04790    -0.02313    -0.02269    -0.07799
## ID206    -0.02654     0.02407    -0.00558    -0.00775    -0.02446
## ID208     0.05005     0.02214    -0.02543    -0.04412    -0.01130
```

Quick start

```
MyResult.pca <- pca(X,ncomp = 10,center = TRUE,scale = FALSE)       # 1 Run the method
plotIndiv(MyResult.pca)                                              # 2 Plot the samples (first two compo
```



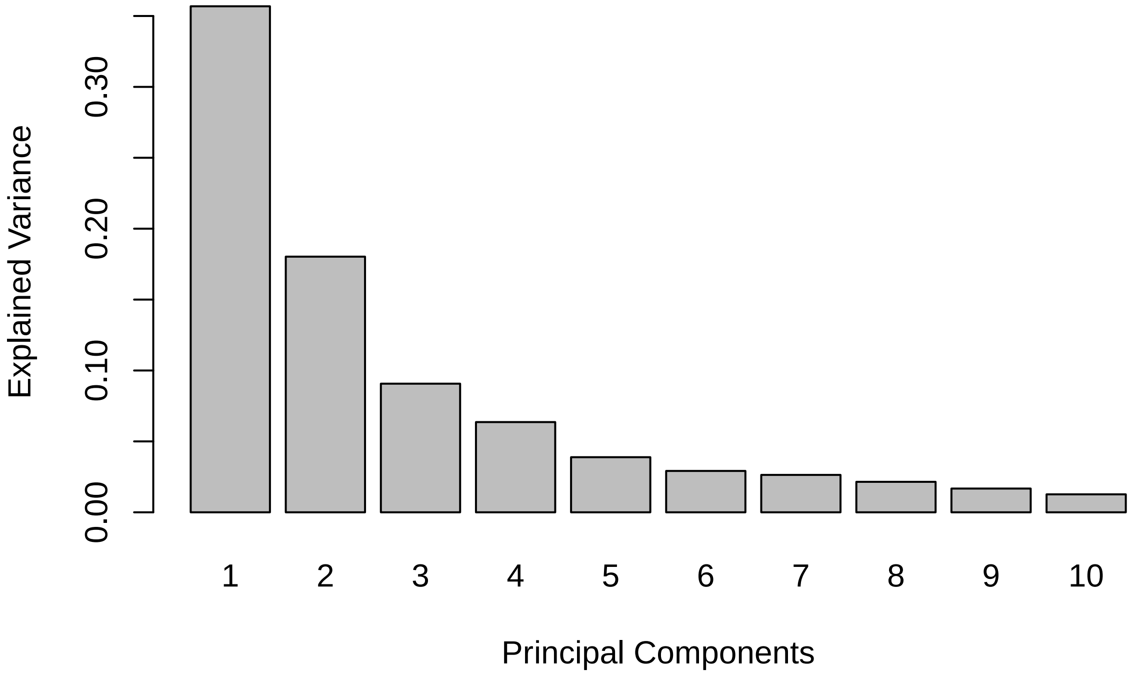
```
plotVar(MyResult.pca)                                                # 3 plot the variables (first two com
```



How many components do I need?

The amount of variance explained by each component can be visualised. This will give us an indication of how many principal components we would need.

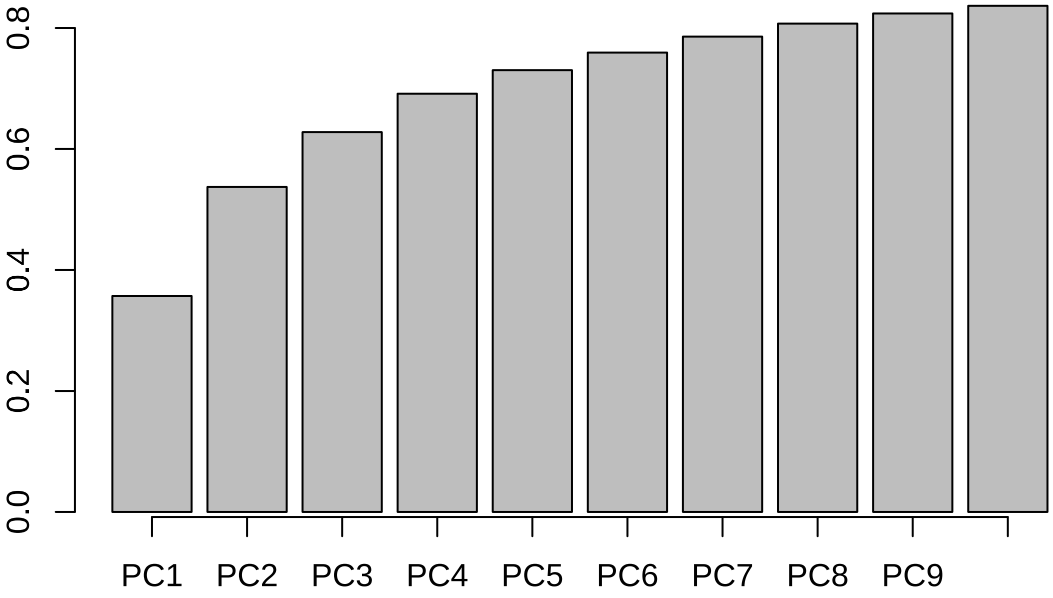
```
# screeplot
plot(MyResult.pca)
```



```
# you can also see print the exact numbers for each component
MyResult.pca$explained_variance
```

```
##        PC1        PC2        PC3        PC4        PC5        PC6
## 0.35684128 0.18027769 0.09069665 0.06362638 0.03885429 0.02916076
##        PC7        PC8        PC9       PC10
## 0.02637122 0.02148534 0.01674690 0.01265538
```

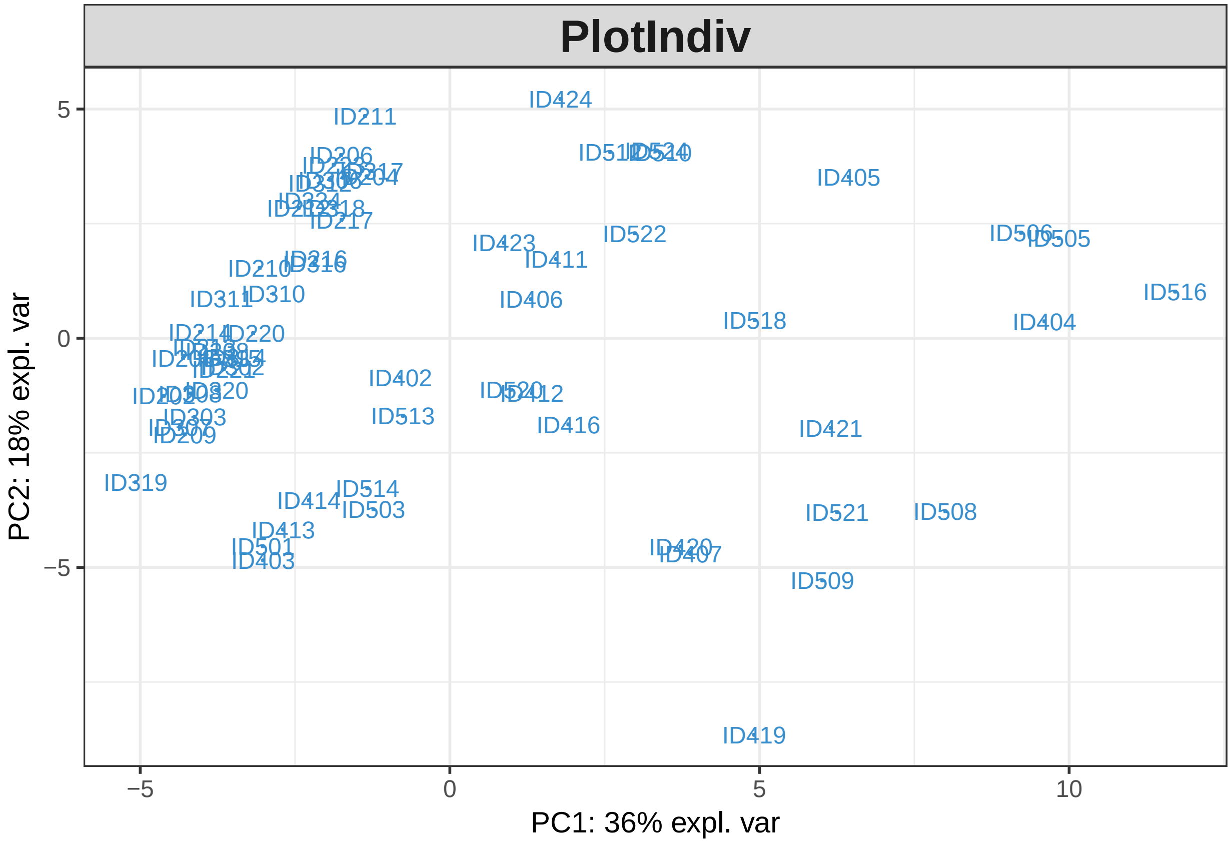
```
# we can also have a look at the cumulative proportion of variance explained
barplot(MyResult.pca$cum.var,axis.lty = 1)
```



Simple and more customised plots

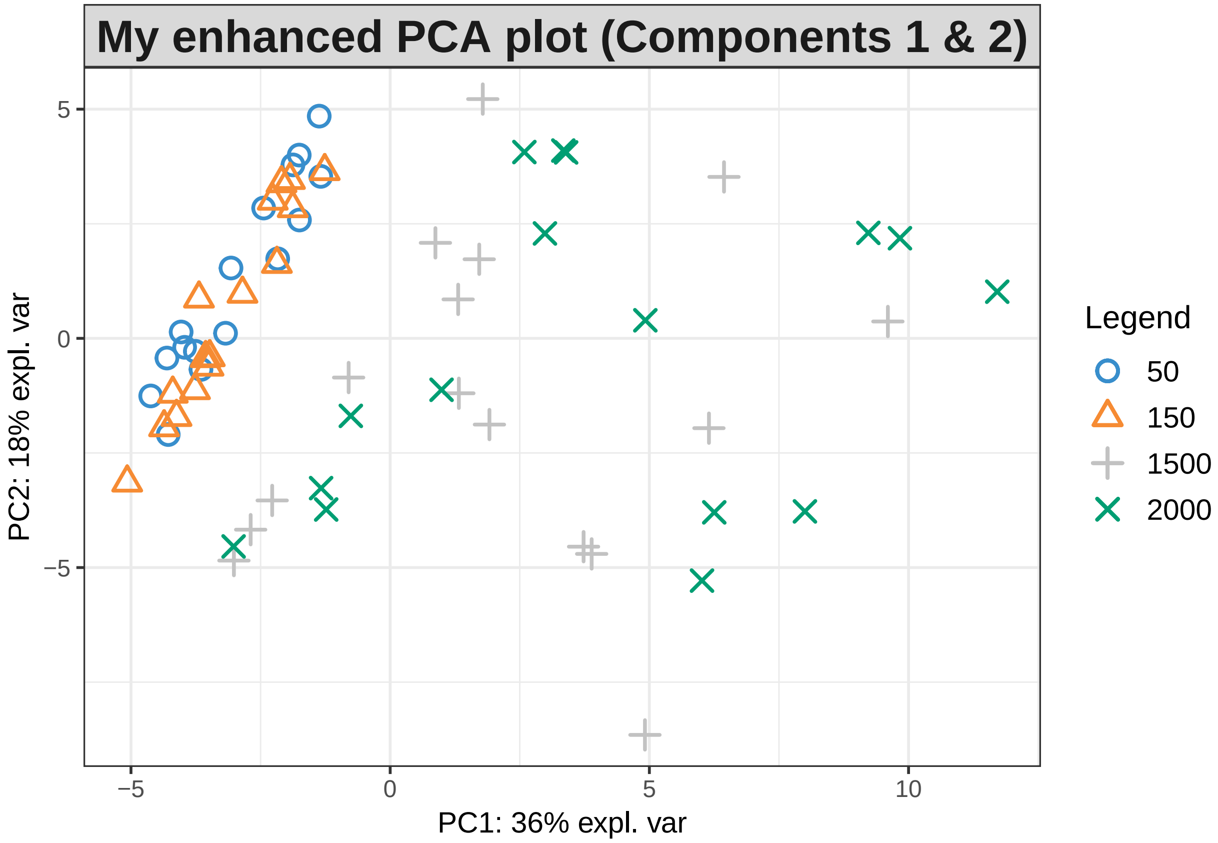
A simple plot first. By default it takes the first two principal components.

```
plotIndiv(MyResult.pca)
```



We can customize this plot a bit more to better display the grouping of samples.

```
# we read the animal applied dose from the initial imported data
grouping_per_dose = liver.toxicity$treatment$Dose.Group

# we plot again with that information in mind
plotIndiv(MyResult.pca,comp = c(1,2),ind.names = FALSE,title = "My enhanced PCA plot (Components 1 & 2)
```

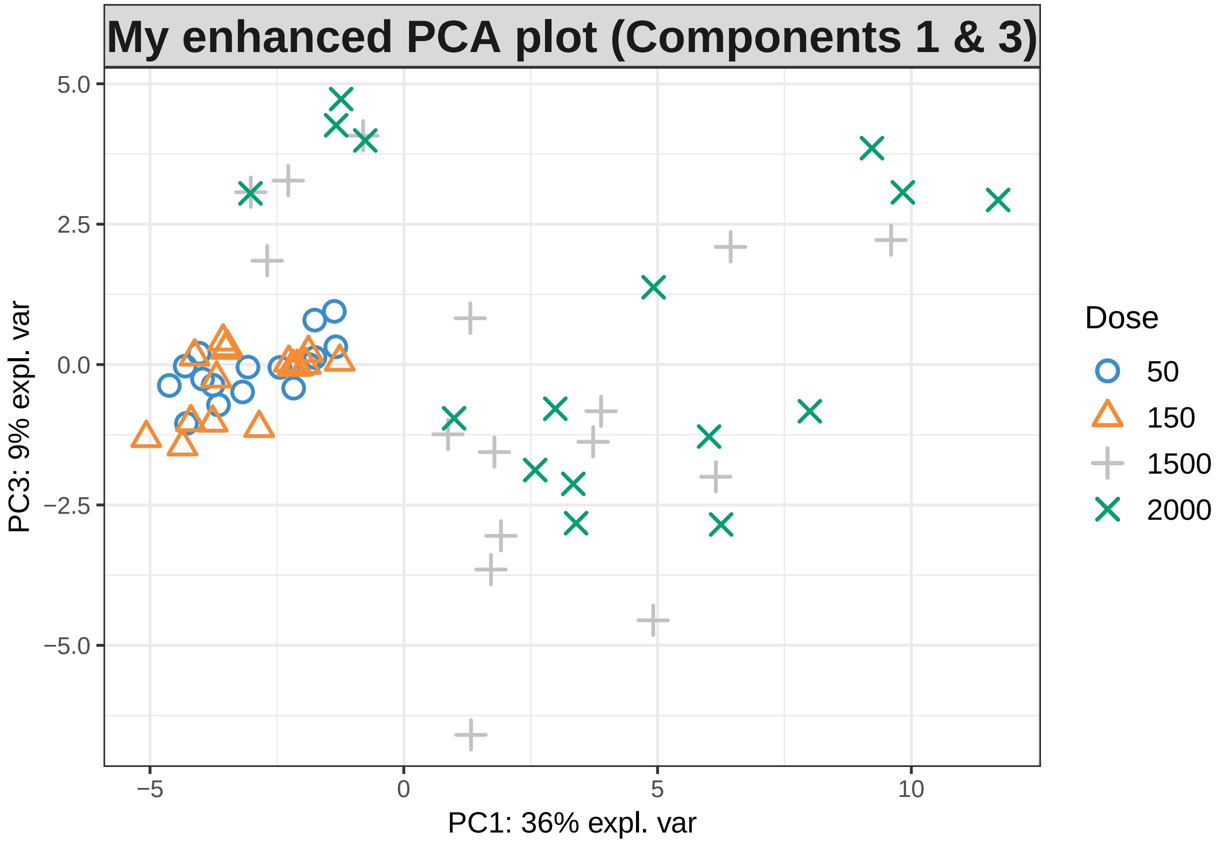


From this plot, we can distinguish the samples exposed to small doses or higher doses of acetaminophen.

Visualising other components

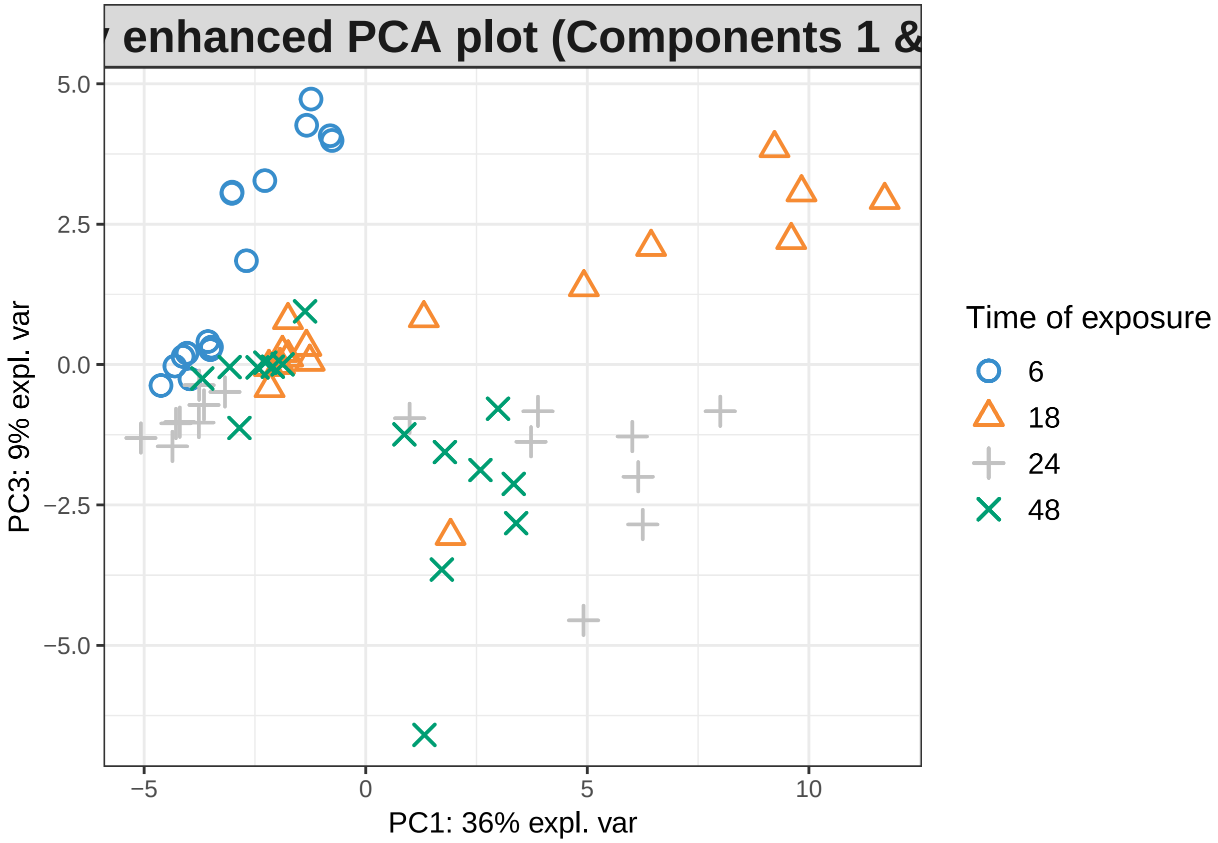
Since the third principal component also explained some variation (question -> how much?), we will also plot samples using this component.

```
#
plotIndiv(MyResult.pca,
          comp = c(1,3),
          legend = TRUE,
          legend.title = "Dose",
          group = grouping_per_dose,
          ind.names=FALSE,
          title = 'My enhanced PCA plot (Components 1 & 3)'
          )
```



It seems that the third component does not help to discriminate samples based on their dose exposure. Can time of necropsies be related to that third component?

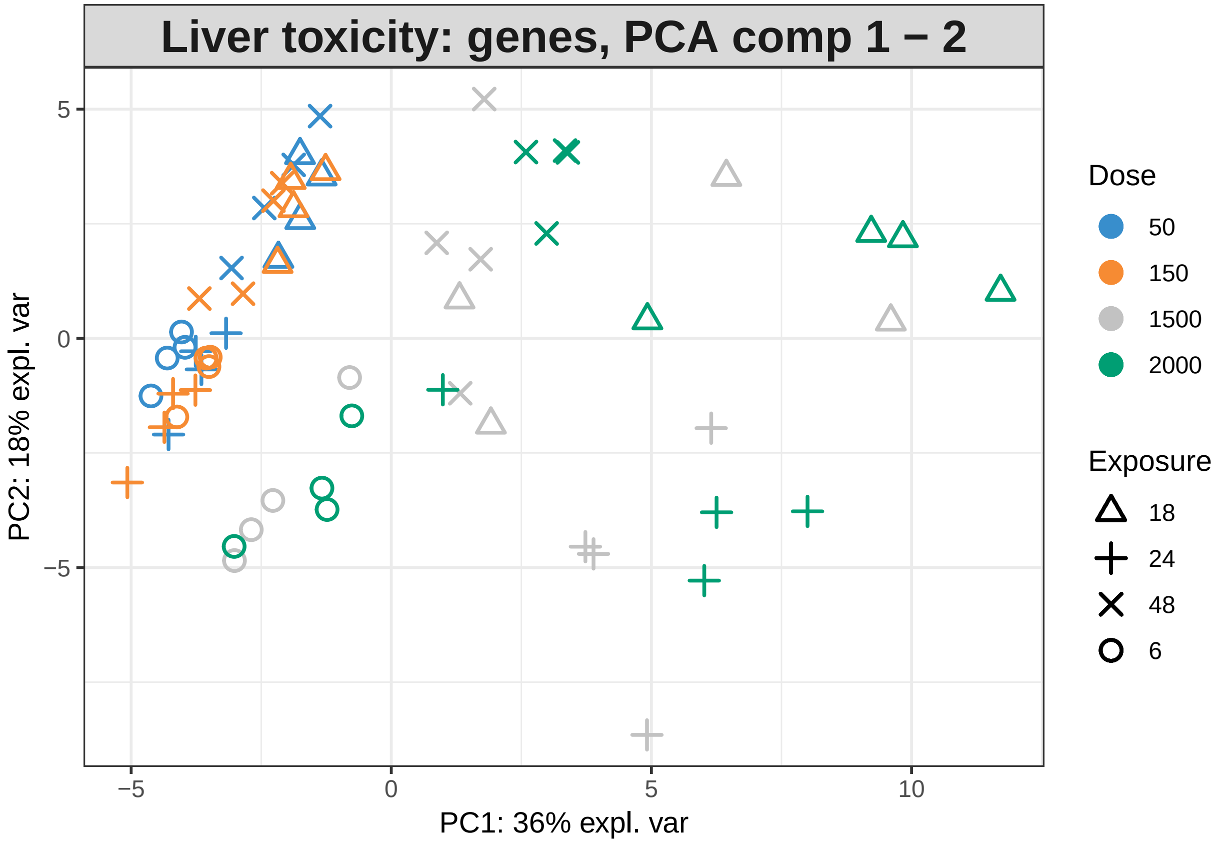
```
# grouping by time of necropsy
grouping_per_time = liver.toxicity$treatment$Time.Group

# plot
plotIndiv(MyResult.pca, comp = c(1,3),
          ind.names = FALSE,
          legend = TRUE,
          legend.title = "Time of exposure",
          group = grouping_per_time,
          title = 'My enhanced PCA plot (Components 1 & 3)')
```



Displaying two experimental groupings on the first two principal components

If we wish to display several experimental info, we can combine both color and symbols on the plot.

```r
plotIndiv(MyResult.pca,
          ind.names = FALSE,
          group = grouping_per_dose,
          pch = as.factor(grouping_per_time),
          legend = TRUE,
          legend.title = 'Dose',
          legend.title.pch = 'Exposure',
          title = 'Liver toxicity: genes, PCA comp 1 - 2')
```


Liver toxicity: genes, PCA comp 1 − 2
PC2: 18% expl. var
PC1: 36% expl. var
5
0
−5
−5
0
5
10
Dose
50
150
1500
2000
Exposure
18
24
48
6