Code application PLS-DA

Lucile Riaboff

14/09/2020

```r
knitr::opts_chunk$set(echo = TRUE)
```

```r
rm(list = ls())
```

CHARGEMENT PACKAGE

```r
library(plsdepot)
library(ggplot2)
library(ggrepel)
```

CHARGEMENT FONCTION CIRCLE FUN

```r
setwd("~/Presentations/PLS-DA") ## Ouverture session avec chemin fonction
source("CircleFun.r")
```

CHARGEMENT FICHIER DE DONNEES

```r
setwd("~/Presentations/PLS-DA") ## Ouverture session avec chemin jeu de donnees
load(file = "data_PLSDA.RData")
## Jeu de donnees :
## 3 premieres colonnes = description observations
## colonne 4 a 44 = 41 variables explicatives
## colonne 45 = variable a expliquer (lame_score)
```

MISE EN FORME DES VARIABLES

```r
## CREATION DE LA MATRICE DE VARIABLES X
Xvar <- data_select[,4:(ncol(data_select)-1)]
## suppression variables qui decrivent les observations et variable a predire

## CREATION DE LA MATRICE DES VARIABLES Y DICHOTOMISEE
no_lame <- ifelse(data_select$lame_score=="no_lame",1,0)
## si observation non boiteuse, alors 1 sinon 0
lame = ifelse(data_select$lame_score=="lame",1,0)
```

```r
## si observation boiteuse, alors 1 sinon 0

Yvar <- data.frame(no_lame = no_lame, lame = lame)

rm(no_lame)
rm(lame)
```

APPLICATION DE LA PLS-DA (PLS-2 SUR VARIABLE Y DICHOTOMISEE)

```r
res.plsreg2 <- plsreg2(Xvar, Yvar, comps = 10, crosval = TRUE)
## remarque : par defaut,centrage et reduction variables par fonction plsreg2
```

CHOIX DU NOMBRE DE COMPOSANTES A RETENIR

```r
## Selection a partir evolution erreur classification
## obtenue par cross validation selon nombre de composantes
## ou selection variables avec Q2 :
## critere de Tenenhaus : composante h retenue si Q2h > 0.0975
## (ou Q2h > 0.05 selon reference)

res.plsreg2$Q2 ## ici seulement premiere composante significative
```

```
##      Q2.no_lame     Q2.lame          Q2
## t1   0.11424141  0.11424141  0.11424141
## t2   0.02533965  0.02533965  0.02533965
## t3  -0.01819481 -0.01819481 -0.01819481
## t4  -0.02544873 -0.02544873 -0.02544873
## t5  -0.05124956 -0.05124956 -0.05124956
## t6  -0.04139387 -0.04139387 -0.04139387
## t7  -0.03947761 -0.03947761 -0.03947761
## t8  -0.04640724 -0.04640724 -0.04640724
## t9  -0.04751747 -0.04751747 -0.04751747
## t10 -0.04995937 -0.04995937 -0.04995937
```

```r
res.plsreg2$Q2cum ## mais amelioration Q2cum deuxieme composante :
```

```
##     Q2cum.no_lame  Q2cum.lame       Q2cum
## t1     0.11424141  0.11424141  0.11424141
## t2     0.13668622  0.13668622  0.13668622
## t3     0.12097839  0.12097839  0.12097839
## t4     0.09860841  0.09860841  0.09860841
## t5     0.05241248  0.05241248  0.05241248
## t6     0.01318817  0.01318817  0.01318817
## t7    -0.02576881 -0.02576881 -0.02576881
## t8    -0.07337190 -0.07337190 -0.07337190
## t9    -0.12437583 -0.12437583 -0.12437583
## t10   -0.18054894 -0.18054894 -0.18102653
```

```
## 1ere et 2eme composantes retenues --> projection plan (t1, t2) OK, avec
## information essentiellement restituee sur t1
```

PERFORMANCE DU MODELE DE LA PLS-DA

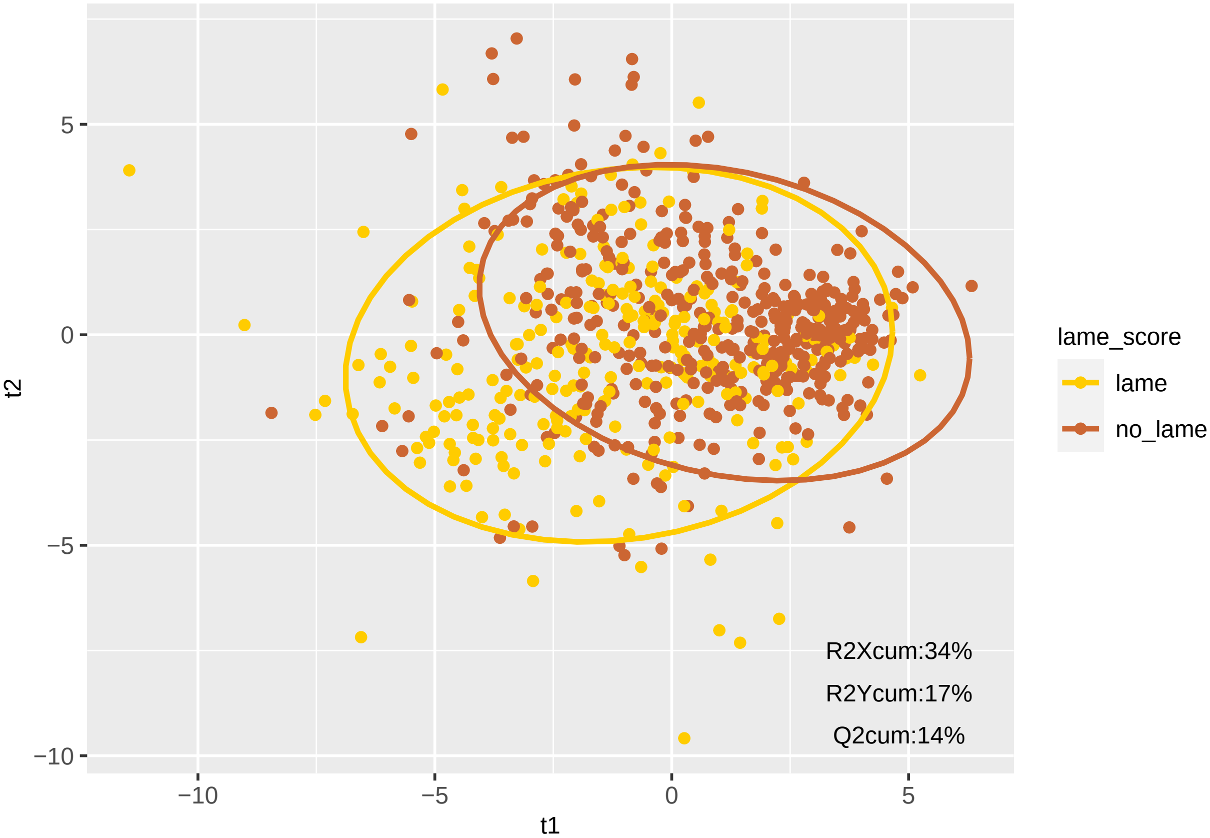
```
Q2cum <- round(res.plsreg2$Q2cum[2]*100) ## Q2 cum deux premieres compo
R2xcum <- round(res.plsreg2$expvar[2,2]*100) ## R2xcum deux premieres compo
R2ycum <- round(res.plsreg2$expvar[2,4]*100) ## R2ycum cum deux premieres compo
## Remarque : R2 et Q2 assez faibles
## --> modele developpe non predictif, mais 1ere composante
## significative --> une partie de l'information de Y (17% >> 0%)
## effectivement expliquee par X
```

IDENTIFICATION DES VARIABLES LES PLUS DISCRIMINANTES AVEC LES VIP

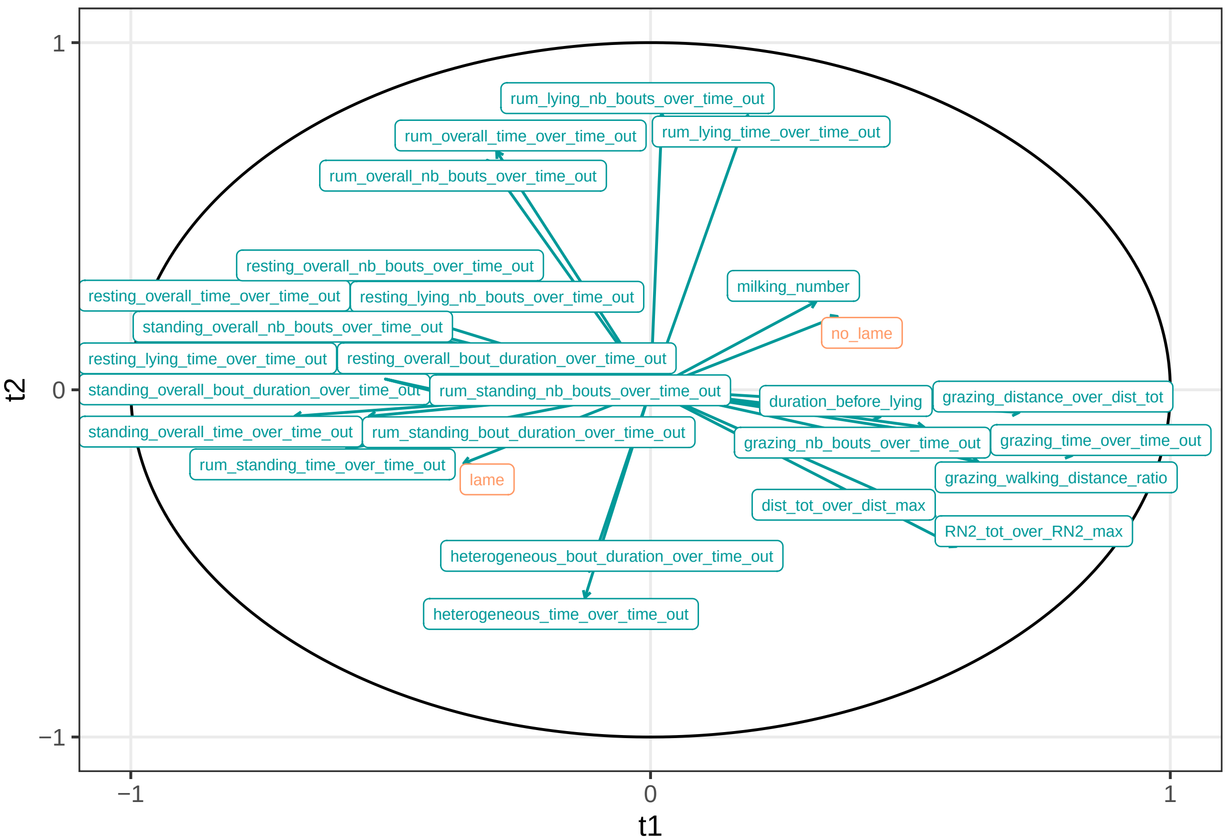
```
## Variables retenues --> VIP > 0.8 sur les deux premieres composantes
var_imp <-res.plsreg2$VIP[which(res.plsreg2$VIP[,2]>=0.8),2]
var_imp_names <-rownames(res.plsreg2$VIP[which(res.plsreg2$VIP[,2]>=0.8),])

var_imp_sort <- sort(var_imp,decreasing = TRUE) ## Variables triees par ordre
## d'importance decroissant a partir des valeurs des VIP
## sur les deux premieres composantes
```

VISUALISATION DES INDIVIDUS ET VARIABLES SUR LES DEUX PREMIERES COMPOSANTES PLS

```
## Plot des individus
df_pda_ind <- data.frame(score_t1 = res.plsreg2$x.scores[,1],
                         score_t2 = res.plsreg2$x.scores[,2],
                         lame_score = data_select$lame_score)

p <- ggplot(data = df_pda_ind, aes(x = score_t1, y = score_t2, col = lame_score))+
geom_point() +  annotate(geom="text", x=4.8, y=-7.5,
label=paste("R2Xcum:", R2xcum, "%", sep = ""), color="black", size = 3) +
annotate(geom="text", x=4.8, y=-8.5,
label=paste("R2Ycum:", R2ycum, "%", sep = ""), color="black", size = 3) +
annotate(geom="text", x=4.8, y=-9.5,
label=paste("Q2cum:", Q2cum, "%", sep = ""), color="black", size = 3) +
scale_color_manual(values=c("#FFCC00", "#CC6633"))+ stat_ellipse(size = 1) +
theme(axis.title.x = element_text(size = 9),
axis.title.y = element_text(size = 9),legend.title = element_text(size = 9),
legend.text = element_text(size = 9)) +
xlab("t1") + ylab("t2")

p
```




```r
## Plot du cercle des correlations avec les variables selectionnees

df_pda_var <-
  res.plsreg2$cor.xt[which(rownames(res.plsreg2$cor.xt)%in%var_imp_names),1:2]

df_pda_var<- data.frame(rbind(df_pda_var, res.plsreg2$cor.yt[,1:2]))
df_pda_var$type <- c(rep("variable", (nrow(df_pda_var)-2)), rep("lame_score",2))
df_pda_var$type <- as.factor(df_pda_var$type)

df_pda_var=cbind(x1=rep(0,times=dim(df_pda_var)[1]),
                 x2=rep(0,times=dim(df_pda_var)[1]),
                 df_pda_var)

colnames(df_pda_var)=c("x1","y1","xend","yend","type")

dat <- circleFun(npoints = 1000)

my_color <- ifelse(df_pda_var$type == "variable","#009999","#FF9966")

g=ggplot(dat,aes(x,y)) + geom_path()+
scale_x_continuous(breaks=seq(-1, 1, by=1))+
scale_y_continuous(breaks=seq(-1, 1, by=1))
g=g + theme_bw()+theme(panel.grid.minor = element_blank())
g=g + geom_segment(aes(x = x1, y = y1, xend = xend, yend = yend),color="#009999",
data = df_pda_var,
arrow = arrow(length = unit(0.01, "npc"))) + xlab("t1") + ylab("t2") +
geom_label_repel(data = df_pda_var,
aes(xend,yend, label = rownames(df_pda_var)),size = 2,
inherit.aes=FALSE,color = my_color,fill ="white",arrow=arrow(length(unit(0,'inches'))),
box.padding = 0)
g
```



1
0
−1
t2
−1
0
1
t1
rum_lying_nb_bouts_over_time_out
rum_overall_time_over_time_out
rum_lying_time_over_time_out
rum_overall_nb_bouts_over_time_out
resting_overall_nb_bouts_over_time_out
resting_overall_time_over_time_out
resting_lying_nb_bouts_over_time_out
milking_number
standing_overall_nb_bouts_over_time_out
no_lame
resting_lying_time_over_time_out
resting_overall_bout_duration_over_time_out
standing_overall_bout_duration_over_time_out
rum_standing_nb_bouts_over_time_out
duration_before_lying
grazing_distance_over_dist_tot
standing_overall_time_over_time_out
rum_standing_bout_duration_over_time_out
grazing_nb_bouts_over_time_out
grazing_time_over_time_out
rum_standing_time_over_time_out
lame
grazing_walking_distance_ratio
dist_tot_over_dist_max
RN2_tot_over_RN2_max
heterogeneous_bout_duration_over_time_out
heterogeneous_time_over_time_out