TP: ACP de données transcriptomiques

Résumé

Travaux pratiques sur l'analyse de données transcriptomiques par Analyse en composantes principales avec le logiciels R.

Suivant : AFD de données soci-économiques

1 Objectif

Les données ont été fournies par Thierry Pineau et Pascal Martin du laboratoire de pharmacologie et toxicologie de l'INRA de Toulouse (site de Saint-Martin). Elles proviennent d'une étude de nutrition chez la souris. Pour 40 souris, nous disposons :

- des données d'expression de 120 gènes recueillies sur membrane nylon avec marquage radioactif ; une "normalisation" et une transformation (log) ont déjà été appliquées aux données ;
- des mesures de 21 acides gras hépatiques. Ces variables ne seront pas étudiées dans cette première approche mais le fichier contient des informations relatives au régime ainsi qu'au génotype des souris.

Par ailleurs, les 40 souris sont réparties selon deux facteurs :

- Génotype (2 modalités) : les souris sont soit de type sauvage (wt) soit génétiquement modifiées (PPAR) ; 20 souris dans chaque cas.
- Régime (5 modalités) : les 5 régimes alimentaires sont notés `ref`, `efad`, `dha`, `lin`, `tournesol` ; 4 souris de chaque génotype sont soumises à chaque régime alimentaire.

L'objectif majeur sur ce genre de données est le recherche de l'effet des différents facteurs, ici les génotypes et les régimes, sur les expressions des gènes. Mais, de façon préalable il est indispensable d'aborder les données par une étude descriptive afin d'en étudier les distributions ainsi que la structure de corrélation. Les aspects plus biologiques ne seront pas abordés au cours du traitement statistique de ces données ; des éléments d'interprétation se trouvent dans le cours et les articles cités en référence.

2 Prise en compte des données

Exécuter les commandes ci-dessous :

```
# lecture du fichier contenant les concentrations
# d'acides gras, les génotypes et les régimes
lipides=read.table("lipides.csv",sep=",")
# lecture du fichier contenant les expressions
Exprs=read.table("genes.csv",sep=",")
# vérification de la bonne lecture des fichiers
dim(Exprs)
dim(lipides)
summary(Exprs)
summary(lipides)
# extraction du facteur génotype
genotype=as.factor(lipides[,23])
# extraction du facteur régime
regime=as.factor(lipides[,22])
# extraction des concentrations d'acides gras
lipides=lipides[,1 :21]
# suppression d'une colonne (un gène)
# présentant des erreurs de mesure
Exprs=Exprs[,-3]
```

3 Étude uni variée

Exécuter les commandes et commenter les résultats. Noter en particulier l'effet du centrage et de la réduction.

```
class(Exprs)
class(t(Exprs))
boxplot(Exprs)
boxplot(data.frame(t(Exprs)),horizontal=TRUE)
boxplot(data.frame(scale(Exprs,scale=FALSE)))
boxplot(data.frame(scale(Exprs)))
boxplot(data.frame(scale(t(Exprs),scale=FALSE)),
        horizontal=TRUE)
```

```
genes=dimnames(Exprs)[[2]]
namsour=dimnames(Exprs)[[1]]
```

4 Différentes analyses en composantes principales

Exécuter les commandes et commenter les résultats. Noter les différentes stratégies (réduction, transposition) et en particulier la différence importante entre les fonctions `princomp` et `prcomp` : `prcomp`, qui calcule directement la décomposition en valeurs singulières, peut traiter une matrice avec plus de colonnes que de lignes.

```
souracp=princomp(t(Exprs))
plot(souracp)
biplot(souracp)

souracp=princomp(t(scale(Exprs, scale=FALSE)))
plot(souracp)
biplot(souracp)

souracp=prcomp(Exprs)
boxplot((data.frame(souracp$x)))
plot(souracp)
biplot(souracp)

souracp=prcomp(Exprs,scale=TRUE)
boxplot(data.frame(souracp$x))
biplot(souracp)
```

Les commandes ci-dessous permettent de distinguer régimes et génotypes sur les représentations.

```
pt=23+as.integer(genotype)

plot(souracp$x,type="p",pch=pt,cex=2)
text(30*souracp$rotation,genes,col="blue")
```

```
plot(souracp$x,type="p",pch=pt,cex=2,
     col=as.integer(regime))
text(30*souracp$rotation,genes,col="blue")

souracp=prcomp(t(scale(Exprs,scale=FALSE)))
plot(souracp)
biplot(souracp)
```

Du fait du nombre important de variables/gènes, les graphiques restent assez illisibles. Les commandes ci-dessous proposent une sélection des gènes pour la représentation. Sur quel critère se base cette sélection ? Commenter les commandes.

```
coord=souracp$x[,1:2]
coord2=coord^2
contrib=apply(coord2,1,sum)/sum(souracp[[1]][1:2]^2)
hist(contrib)
selec=contrib>0.5
sum(selec)
genes[selec]

plot(souracp$rotation,type="p",pch=pt,cex=2,
     col=as.integer(regime))
text(0.2*souracp$x[selec,],genes[selec],col="blue")
biplot(souracp$x[selec,],souracp$rotation)
```

Commentaire sur la facilité de distinguer les génotypes, l'importance de certains gènes dans ce rôle. Que dire de la distinction des régimes ?

Commenter les graphiques obtenus ci-après.

```
boxplot(CYP4A14~genotype,data=Exprs)
boxplot(CYP3A11~genotype,data=Exprs)
boxplot(THIOL~genotype,data=Exprs)
boxplot(PMDCI~genotype,data=Exprs)
```