TP: AFD de données transcriptomiques

Résumé

Travaux pratiques sur l'analyse de données transcriptomiques par Analyse factorielle discriminante avec le logiciels SAS.

1 Introduction

Les données sont celles concernant les régimes de 40 souris 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'analyse en composantes principales a facilement mis en évidence les gènes permettant une séparation nette des deux génotypes sans que les classes soient considérées connues a priori. En revanche les régimes ne se regroupaient pas sur les graphiques. L'objectif est alors d'utiliser une analyse factorielle discriminante pour cherche à discriminer, toujours de façon linéaire, ces régimes.

Quelle question préalable pose cette approche ? pourquoi n'est-elle pas calculable directement ?

2 Gestion des données

Deux problèmes apparaissent avec R : il n'y a pas de fonction d'AFD en standard dans R pas plus que de procédure de sélection de variables dans le cadre de l'analyse discriminante. Ces fonctions seraient faciles à programmer mais le choix est fait ici de passer à SAS afin de bénéficier des outils disponibles. Il s'agit d'ailleurs souvent de la bonne stratégie : aller chercher le bon outil dans le logiciel le plus adapté. Les données pourraient être directement lues dans SAS mais la gestion des noms de variables étant déjà traités avec R, le transfert des données entre les deux logiciels est préféré.

2.1 Lecture des données

```
# 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=",")
# 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]
expreg=data.frame(regime,Exprs)
```

2.2 Transfert des données dans SAS

Une bibliothèque (`foreign`) permet de transférer directement des données binaires entre logiciels. Néanmoins ces fonctions ayant des comportement parfois étranges, une approche directe est préférée en deux étapes :

1. Création à partir de R de fichiers textes au format ASCII contenant les données :

```
write.table(expreg,"expreg.dat")
```

2. Lecture des données sous SAS (SAS et R doivent avoir été lancés dans le même répertoire).

```
proc import datafile="expreg.dat"
  out=sasuser.expreg
  dbms=dlm
  replace;
getnames=yes;
run;
```

Vérifier la bonne lecture des données.

3 Sélection des variables

Les gènes ou variables les plus discriminants, en un sens *linéaire*, sont recherchés par la procédure `stepdisc`. Celle-ci fonctionne sur la même principe que celui de la sélection de variables en régression. Les variables les plus (resp. les moins) significatives au sens d'un test sont successivement sélectionnées (resp. éliminées) en suivant l'un des algorithmes : `forward`, `backward`, `stepwise`.

```
proc stepdisc data=sasuser.expreg method=stepwise;
class regime;
var _numeric_;
run;
```

Pourquoi l'algorithme s'arrête à 40 variables avec une corrélation canonique à 1 ? Noter la liste des disons 15 premières variables sélectionnées.

4 Analyse discriminante

Exécuter les macro commandes en introduisant la liste des variables sélectionnées.

```
%afd(expreg,numobs,regime,Lpin1 GSTmu PLTP CYP4A14
       BIEN mHMGCoAS MTHFR COX1 THIOL CYP3A11 CYP26
       LXRa CYP2b10 HPNCL apoC3,4);
%epsf(afdexpr.eps,l=6,h=6);
%gafdvx;
%gafdix;
```

Interpréter succinctement les graphiques (axes discriminants) et la qualité de cette discrimination. Il est clair que la présence d'un biologiste est nécessaire afin d'apprécier la pertinence de la représentation. Comparer la liste des gènes les plus influents avec ceux de l'ACP. Le point suivant à traiter serait de considérer les 10 classes obtenues en croisant les deux facteurs : génotype et régime.