TP: Classification non supervisée avec SAS

Résumé

Travaux pratiques sur la classification non supervisée (CAH, kmeans) avec SAS.

Suivant : Classification non supervisée avec R

1 Introduction

Au cours de cette séance, le problème de la classification de données est abordé de façon particulière. L'objectif est de classer un grand nombre d'individus en un nombre adéquat de classes. Pour cela, une méthode combinant l'utilisation des procédures de classification `fastclus` et `cluster` est mise en œuvre. Cette méthode reste efficace quand les données sont en nombre important, la procédure `fastclus` pouvant traiter jusqu'à 100 000 individus. La démarche consiste tout d'abord à déterminer le nombre K de classes à l'aide d'une classification hiérarchique exécutée sur les barycentres de classes issues d'une classification par nuées dynamiques. Cette première classification recherchant un nombre important de classes. Ensuite, une classification hiérarchique en K classes est opérée sur l'ensemble du fichier, la classification obtenue est enfin améliorée en enchaînant une dernière fois les nuées dynamiques initialisées par les barycentres.

Remarque importante : les données étudiées sont ici toutes quantitatives et peuvent donc directement être prises en compte par les procédures de classification. Dans le cas contraire : variables qualitatives ou mélange de variables quantitatives et qualitatives, d'autres approches sont nécessaires : définir une distance adaptée entre les individus ou calcul de "scores" quantitatifs (composantes principales) par une AFCM pour se ramener à des variables quantitatives.

Enfin, nous insistons sur l'intérêt de pouvoir visualiser les classes obtenues par une méthode factorielle adaptée : ici l'ACP car ce sont des variables qui sont observées. Dans le cas où une matrice de distances ou de dissimilarités serait utilisée pour la classification, le MDS servirait à la représentation.

2 Les données

Les données proviennent de l'Agence d'Urbanisme de l'Agglomération Toulousaine. Elles concernent les horaires de passage du bus N°2 à 9 carrefours du trajet Rangueil-Matabiau (antérieur à la mise en place du métro), au cours d'une journée du printemps 1991 et d'une journée du printemps 1993. Le tableau bus contient 282 observations (correspondant aux 282 départs) et 10 variables. La variable `id` identifie les bus : `A1b6.43` correspond au bus parti de Rangueil à 6.43 heures (les valeurs sont des fractions d'heure) pour l'année 1991. Les autres variables repèrent les horaires de passage aux 9 carrefours : CHUouFAC, Ponsan, Rocade, Récollets, Jules Guesde, Esquirol, Strasbourg, Bonrepos, Matabiau. La classification de ces données a pour but d'identifier les différentes plages horaires pour la circulation des bus en relation avec les difficultés de circulation.

Exécuter le programme de lecture des données.

```
title 'Classification␣des␣donnees␣Bus';
data sasuser.bus;
infile 'busrm.dat';
input  id $ CHUouFAC Ponsan Rocade Recollet JGuesde
       Esquirol Strasbg  Bonrepos  Matabiau ;
run;
```

On s'intéresse plus particulièrement aux temps de parcours c'est-à-dire aux durées écoulées entre deux stations. Le nouveau tableau de données est obtenu de la façon suivante :

```
data sasuser.bus1(drop = CHUouFAC Ponsan Rocade
     Recollet JGuesde Esquirol Strasbg
     Bonrepos Matabiau);
set sasuser.bus;
CHU_Pons = Ponsan-CHUouFAC;
Pons_Roc = Rocade-Ponsan;
Roc_Rec  = Recollet-Rocade;
Rec_JGu  = JGuesde-Recollet;
```

```
JGu_Esq  = Esquirol-JGuesde;
Esq_Stra = Strasbg-Esquirol;
Stra_Bon = Bonrepos-Strasbg;
Bon_Mata = Matabiau-Bonrepos;
run;
```

3 Choix du nombre de classes

Pour ces données, le nombre de lignes (de départs de bus) est relativement faible et il est possible de calculer directement une CAH, ce qui ne serait pas possible avec un fichier plus gros. Par souci de généralité, on adopte ici une stratégie applicable même à des très gros fichiers. Les différentes étapes permettant de choisir un nombre pertinent de classes sont alors les suivantes :

i. *Restriction du nombre des individus à classer* par réallocation itérative (procédure `fastclus`) des individus dans L classes, où L est choisi arbitrairement égal au dixième de l'effectif initial.

ii. *Classification Ascendante Hiérarchique* (procédure `cluster`) des barycentres des L classes obtenues précédemment. Le poids d'un barycentre est égal à la somme des poids des individus de sa classe. Le saut de Ward est utilisé par défaut. On rappelle que le saut de Ward est utilisé quand on veut maximiser l'inertie inter de la partition.

iii. *Représentation graphique du R^2 semi-partiel* (cas du saut de Ward) pour aider l'utilisateur dans le choix du nombre de classes.

Les trois premières étapes sont effectuées à l'aide de la macro-procédure `choixnc`, qui prend en arguments le tableau de données, la liste des variables, le nombre L de classes, et l'indice d'éloignement utilisé dans la CAH (Ward par défaut) :

```
%choixnc(bus1,CHU_Pons--Bon_Mata,28);
```

Le deuxième graphique, est obtenu en exécutant la macro ci-dessous en précisant un nombre max de classes :

```
%critere(10);
```

4 Classification

Une fois déterminé le nombre K de classes de la partition finale au vu des graphiques précédents, on propose de comparer les classifications obtenues par les deux méthodes qui suivent.

4.1 CAH sur les données initiales

On effectue une Classification Ascendante Hiérarchique, mais cette fois à partir du tableau de données initiales (`sasuser.bus1`), et en spécifiant K. Cette classification est effectuée à l'aide de la macro `cah`. Elle produit en sortie une table `sasuser.chclasse` contenant les données initiales et le numéro de classe `cluster` de chaque individu, qui est ensuite utilisé pour représenter la classification, et une table `poles` contenant les barycentres des classes.

```
%cah(bus1,id,CHU_Pons--Bon_Mata,5);
```

On peut visualiser cette première classification, en effectuant la représentation des individus sur les plans factoriels de l'ACP, chaque individu étant identifié par le numéro de sa classe. Après avoir déterminé le nombre de facteurs nécessaires à une bonne représentation, interpréter les axes et commenter la classification des individus sur les différents plans.

```
%acp(chclasse,cluster,CHU_Pons--Bon_Mata);
%gacpixm;
%gacpvx;
```

4.2 Nuées dynamiques

La deuxième méthode consiste à effectuer une classification par nuées dynamiques, en choisissant comme pôles de départ les barycentres des K classes obtenues en sortie de la macro `cah` (cette étape peut être vue comme une amélioration de la partition obtenue après exécution de `cah`). La macro `nudnc` réalise cette partition et fournit en sortie une table SAS `sasuser.ndclasse` contenant : les variables initiales, le numéro de classe `classe` de chaque individu, et une variable `distance` donnant la distance de chaque individu au barycentre de sa classe. Cette procédure peut également être utilisée indépendamment en ne précisant pas le paramètre `seed=pole` d'initialisation des groupes.

```
%nudnc(bus1,id,CHU_Pons--Bon_Mata,5,init=poles);
```

Cette partition finale est représentée comme précédemment sur les plans factoriels de l'ACP. Après exécution des commandes suivantes, comparer avec la partition précédemment obtenue :

```
%acp(ndclasse,classe,CHU_Pons--Bon_Mata);
%gacpixm;
%gacpvx;
```

Malheureusement, SAS ne fournit pas de représentation des enveloppes convexes des classes permettant de comparer plus facilement leur imbrication. Quelle interprétation faire des axes ? Quelles indications cela donne-t-il sur les classes ? Une analyse discriminante peut également être employée pour tenter de mieux représenter les classes et faciliter leur interprétation.

```
%afd(ndclasse,classe,classe,CHU_Pons--Bon_Mata,4);
%gafdix;
%gafdvx;
```

Les résultats sont-ils très différents ?