Scénario: AFCM et exploration de p variables qualitatives

Résumé

Comparaison entre des analyses des correspondances appliquées à une table de contingence, un tableau disjonctif complet et un tableau de Burt comme généralisation de l'Analyse des correspondances à l'analyse des correspondances multiple. Exemple d'AFCM élémentaire et AFCM de variables avec interactions.

1 Généralisation de l'AFC

1.1 Avec R

A développer.

1.2 Avec SAS

Le premier objectif est d'illustrer les différentes façons d'obtenir une généralisation de l'AFC. Comparer les résultats (valeurs propres, graphiques) obtenus par AFC simple ou multiple de différents tableaux construits à partir des deux mêmes variables fictives.

Création des données

```
data afcfic;
input ident $ csp $ sport $ effectif;
cards;
quatre csp3 S 1
un     csp1 R 1
deux   csp2 R 1
trois  csp2 S 1
cinq   csp3 T 1
six    csp4 T 1
;
run;
```

AFC de la table de contingence

```
proc corresp data=afcfic observed out=resul;
tables csp, sport; /*attention à la virgule~!*/
weight effectif;
run;
%gafcx;
```

AFC du tableau de Burt

```
proc corresp data=afcfic observed out=resul;
tables csp  sport; /*attention à la virgule~!*/
weight effectif;
run;
%gafcx;
```

AFC du tableau disjonctif complet

```
proc corresp data=afcfic observed out=resul;
tables ident,csp sport; /*attention à la virgule~!*/
weight effectif;
run;
%gafcx;
```

2 Exemple simple d'AFCM

Nous nous proposons de rechercher si les regroupements de races de chiens issus d'une AFCM sont compatibles avec la fonction de ces chiens considérée comme variable supplémentaire.

2.1 Gestion des données en SAS

Les données sont extraites de Bréfort (1982). Elles contiennent le descriptif des qualités de 27 races de chien : tailles, poids, vélocité, intelligence, codées sur trois modalités (1 faible, 2 moyen, 3 fort), affection et agressivité sur deux modalités (1 faible, 2 forte), enfin la fonction sur trois modalités (1 compagnie,

2 chasse, 3 utilité). L'obtention de graphes des modalités explicites nécessite un recodage précis des modalités avec des libellés facilement identifiables.

Charger le fichier chiens.dat.

```
data sasuser.chiens;
infile 'chiens.dat';
input race $ taille $ poids $ velocite $
 intellig $ affect $ agress $ fonction $;
select(taille);
when('1')taille='T-';
when('2')taille='T+';
when('3')taille='T++';
otherwise;end;
select(poids);
when('1')poids='P-';
when('2')poids='P+';
when('3')poids='P++';
otherwise;end;
select(velocite);
when('1')velocite='V-';
when('2')velocite='V+';
when('3')velocite='V++';
otherwise;end;
select(intellig);
when('1')intellig='I-';
when('2')intellig='I+';
when('3')intellig='I++';
otherwise;end;
select(affect);
when('1')affect='Af-';
when('2')affect='Af+';
otherwise;end;
select(agress);
when('1')agress='Ag-';
when('2')agress='Ag+';
otherwise;end;
select(fonction);
when('1')fonction='Com';
when('2')fonction='Cha';
when('3')fonction='Uti';
otherwise;end;
run;
```

2.2 AFCM avec SAS

Graphiques par défaut

Les graphiques de cette première approche ne sont pas très explicites.

```
proc corresp data=sasuser.chiens mca out=resul;
tables taille--fonction;
run;
```

La même analyse en ajoutant les identificateurs des races des chiens et la variable `fonction` en supplémentaire.

```
proc corresp data=sasuser.chiens  out=resul;
tables race,taille--fonction;
supplementary fonction;
run;
```

Graphiques détaillés

L'édition d'un graphique plus explicite représentant chaque race de chien par une couleur dépendant de sa fonction nécessite quelques manipulations. Donner une interprétation.

```
proc sort data=resul out=result;
by _name_;
run;
proc sort data=sasuser.chiens out=chienst;
by race;
run;
data fusion;
merge result chienst (rename=(race=_name_)
      keep=race fonction);
by _name_;
```

```
run;
%let ident=_name_;
%let x=1;
%let y=2;
%let nc=4;
data anno;
 set fusion;
 retain  xsys ysys '2';
 style='swiss';
 y= dim&y;
 x= dim&x;
 select(fonction);
 when('Cha') color='green';
 when('Com') color='blue';
 when('Uti') color='red';
 otherwise color='black';
 end;
 text=substr(&ident,1,&nc);
 size=0.6;
 label y = "Axe &y"
       x = "Axe &x";
keep x y text xsys ysys size color;
proc gplot data=anno;
 title;
 axis1  length=8cm; /* attention taille */
 axis2  length=8cm;
 symbol1 v=none;
 plot y*x=1 / annotate=anno frame  href=0 vref=0
  haxis=axis1 vaxis=axis2 ;
run;
goptions reset=all;
quit;
```

2.3 AFCM avec R

Transfert des données

Pour éviter de recoder les données dans R, celles-ci sont exportées en l'état.

```
/* Exportation d'un fichier SAS en format .csv */
proc export data=sasuser.chiens
     outfile= "chiens.csv"
     DBMS=CSV REPLACE;
run;
```

Rechercher où SAS a "rangé" le ficher `chiens.csv`, sans doute dans le répertoire duquel l'exécution a été lancée. Éventuellement le déplacer dans le répertoire courant de R.

```
# retour à R
chiens=read.csv("chiens.csv")
 # Vérifier que tout s'est bien passé
#Attention, la première colonne identifie la race du
#chien et donc chaque ``individu''
dimnames(chiens)[[1]]=as.character(chiens[,1])
chiens=chiens[,-1]
summary(chiens)
```

AFCM avec `FactoMineR`

La librairie `FactoMineR` est particulièrement conçue pour exécuter et représenter des analyses factorielles des correspondances.

```
library(FactoMineR)
#afcm avec la fonction en supplémentaire
afcm=MCA(chiens,quali.sup=7,graph=F)
plot(afcm, choix="ind",habillage="quali")
#sans les individus
plot(afcm,habillage="quali",invisible="ind")
```

Seule "erreur" dans la représentation : fournir des parts de "variance" ou "chi2" expliquées par les axes qui n'ont pas de signification statistique à cause de la présence de valeurs propres "artificielles" non nulles issues de la construction du tableau de Burt.

3 Variables avec interactions

L'objet de cet exemple est de mettre en évidence les limitations de l'AFCM ou les précautions à prendre pour traiter des données complexes. L'AFCM

analyse le seul tableau de Burt croisant (tables de contingence) les variables deux à deux. Par construction, cette méthode ne prend donc pas en compte de possibles interactions d'ordre supérieur à deux entre les variables. Ce peut être trompeur.

3.1 Les données

Les données relatives à plusieurs variables qualitatives sont représentées habituellement sous la forme d'une table de contingence *complète*. L'exemple ci-dessous est extrait de Bishop et al. (1976). Il décrit les résultats partiels d'une enquête réalisée dans trois centres hospitaliers (Boston, Glamorgan, Tokio) sur des patientes atteintes d'un cancer du sein. On se propose d'étudier la survie de ces patientes trois ans après le diagnostic. En plus de cette information, quatre autres variables sont documentées pour chacune des patientes :

- le centre de diagnostic,
- la tranche d'âge,
- le degré d'inflammation chronique,
- l'apparence relative (bénigne ou maligne).

L'objectif de cette étude est une analyse descriptive (AFCM) de cette table en recherchant à mettre en évidence les facteurs de décès.

TABLE 1 – *Données sous la forme d'une table de contingence complète*

			Histologie			
			Inflamation minime		Grande inflamation	
Centre	*Age*	*Survie*	Maligne	Bénigne	Maligne	Bénigne
Tokio	< 50	non	9	7	4	3
		oui	26	68	25	9
	$50 - 69$	non	9	9	11	2
		oui	20	46	18	5
	> 70	non	2	3	1	0
		oui	1	6	5	1
Boston	< 50	non	6	7	6	0
		oui	11	24	4	0
	$50 - 69$	non	8	20	3	2
		oui	18	58	10	3
	> 70	non	9	18	3	0
		oui	15	26	1	1
Glamorgan	< 50	non	16	7	3	0
		oui	16	20	8	1
	$50 - 69$	non	14	12	3	0
		oui	27	39	10	4
	> 70	non	3	7	3	0
		oui	12	11	4	1

3.2 Gestion des données dans SAS

Les données sont structurées sous la forme d'un fichier de 72 lignes et 6 colonnes. Chaque ligne décrit le contenu d'une cellule de la table de contingence complète (effectif, modalité de chaque variable) avec le découpage suivant :

```
Effectif | centre | age | survie | inflammation | apparence
```

La table est lue ligne par ligne. Les premières lignes du fichier contiennent donc :

```
     9   1   1    1   1   1
     7   1   1    1   1   2
     4   1   1    1   2   1
     3   1   1    1   2   2
    26   1   1    2   1   1
```

```
    68   1   1    2   1   2
    25   1   1    2   2   1
    ...
```

Lire les données disponibles dans le fichier : diagnos.dat du répertoire usuel.

```
data sasuser.diagnos;
  infile 'diagnos.dat';
  input eff c  ag  m  i  a ;
run;
```

Puis le programme suivant recode les modalités avec des libellés explicites avec la convention suivante : les modalités d'une même variable commence avec la même lettre majuscule afin de les identifier plus facilement sur les graphes.

```
data sasuser.diagnos2 (keep = eff centre age
                      survie inflam appar);
set sasuser.diagnos ;
select (c);
when (1) centre='Ctoki';
when (2) centre='Cbost';
when (3) centre='Cglam';
otherwise;
end;
select(ag);
when (1) age='A<50';
when (2) age='A>-<';
when (3) age='A>70';
otherwise;
end;

select (m);
when(1) survie='Snon';
when(2) survie='Soui';
otherwise;
end;
select (i);
when(1) inflam='Ipet';
when(2) inflam='Igra';
otherwise;
end;
select (a);
when(1) appar='Tmal';
when(2) appar='Tben';
otherwise;
end;
drop c ag  m i a ;
run;
```

Différents traitements uni ou bi-variés (graphes, tables de contingence, tests peuvent alors être entrepris en particulier pour analyser la liaison de la variable survie avec les autres. Ils sont laissés de côté.

3.3 Première AFCM

Par défaut SAS calcule les coordonnées des modalités en dimension 2 et crée une table contenant divers résultats d'aide à l'interprétation ; d'autres options sont possibles.

Calculs

```
proc corresp  data=sasuser.diagnos2 observed
              out=resul mca;
     tables centre age survie inflam appar;
     weight eff;
     run;
%gafcx;
%gafcix; /* le même en couleur */
```

Explorer le code de cette dernière macro pour comprendre sur quel critère les couleurs sont définies et donc sous quelle forme doivent se présenter les données.

Plan factoriel

Interpréter le deuxième axe ; à la lumière de ce graphique, quels sont les facteurs de décès ?

3.4 Prise en compte des interactions

Variable croisée

Le graphique de l'analyse précédente suggère l'influence de l'âge mais aussi celle du centre de diagnostic dans les risques de décès avant trois ans. Pour expliciter ces liaisons, les données sont reconsidérées de la façon suivante :

- les variables `centre` et `age` sont croisées pour construire une variable `agecent` à 9 modalités,
- les variables `inflam` et `appar` sont croisées également pour définir la variable `histo` à 4 modalités,

```
data sasuser.diagnos3;
set  sasuser.diagnos2;
if centre='Ctoki' then
           if age='A<50' then agecent='XT<50';
           else if age='A>-<' then agecent='XT>-<';
           else agecent='XT>70';
if centre='Cbost' then
           if age='A<50' then agecent='XB<50';
           else if age='A>-<' then agecent='XB>-<';
           else agecent='XB>70';
if centre='Cglam' then
           if age='A<50' then agecent='XG<50';
           else if age='A>-<' then agecent='XG>-<';
           else agecent='XG>70';
if inflam='Igra' then
           if appar='Tmal' then histo='Hg-m';
           else if appar='Tben' then histo='Hg-b';
if inflam='Ipet' then
           if appar='Tmal' then histo='Hp-m';
           else if appar='Tben' then histo='Hp-b';
run;
```

Analyse et graphique

Une nouvelle analyse est calculée en considérant, comme actives, les deux variables nouvellement créées ainsi que la variable `survie` et, comme illustratives, les variables initiales : `centre, age, inflam, appar`.

```
proc corresp data=sasuser.diagnos3 observed
                out=resul mca;
tables  survie agecent histo centre age inflam appar;
sup centre age inflam appar;
weight eff;
run;
%gafcix;
```

Apprécier l'importance des couleurs pour interpréter ce type de graphique dès que le nombre de modalités est élevé.

Remarquer les positions particulières des modalités des variables supplémentaires par rapport à celles qui ont été créées. Interpréter les effets respectifs de l'histologie, de l'âge et du centre sur les risques de décès. Comment expliqueriez-vous le taux de mortalité important des patientes de Glamorgan de moins de 50 ans ?

3.5 Analyse avec R

Exercice : écrire le programme R exécutant la préparation des données à partir du fichier initial plutôt que de transférer directement le fichier déjà transformer comme ci-dessous.

Transfert

```
/* Exportation d'un fichier SAS en format .csv */
proc export data=sasuser.diagnos3
     outfile= "diagnos3.csv"
     DBMS=CSV REPLACE;
run;
```

```
# retour à R
diagnos3=read.csv("diagnos3.csv")
# Vérifier le bon transfert
summary(diagnos3)
```

AFCM avec `FactoMineR`

```
library(FactoMineR)
```

```
# fréquences "biaisées" car le programme
# ne supporte pas des fréquences nulles
freq=diagnos3[,1]+.0001
afcm=MCA(diagnos3[,2:8],quali.sup=c(1,2,4,5),
   row.w=freq,graph=F)
plot(afcm, choix="ind",invisible="ind",
   habillage="quali")
```

Comparer les graphiques.