TP: Analyse des correspondances multiple et interactions

Résumé

Exemple d'analyse montrant l'importance de la nécessaire prise en compte d'interactions en analyse des correspondances multiple.

1 introduction

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

2 Gestion des données

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 Analyse des Correspondances Multiple

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.

3.1 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.

3.2 Plan factoriel

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

4 Prise en compte des interactions

4.1 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;
```

4.2 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. SAS/INSIGHT permet également de construire une représentation colorée : ouvrir la table `work.resul` dans sas/insight, colorer les lignes en fonctions des modalités de la variable `_NAME_`, construire un `scatter plot` (ou un `rotating plot` graphe trois D) des variables `DIM1` et `DIM2`.

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 ?

5 Analyse avec R

Exercice : calculer avec R la gestion des données à partir du fichier initial plutôt que de transférer directement le fichier déjà transformer comme ci-dessous.

5.1 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)
```

5.2 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.