

Scénario: AFC de deux variables qualitatives

Résumé

Détail des calculs de l'Analyse Factorielle des Correspondances en R puis application sur un exemple avec la librairie FactoMineR et enfin SAS. Le deuxième exemple met en évidence le rôle de la métrique du χ^2 en comparant ACP et AFC sur les mêmes données issues d'élections.

1 CSP et lieux de vacances

Les données sont une table de contingence croisant une variable de catégorie socioprofessionnelle avec un type d'hébergement en vacances mais sous une forme particulière car archivée en trois colonnes. La première contient l'effectif conjoint des modalités contenus dans les colonnes suivantes. La deuxième colonne est le code de catégorie socioprofessionnelle et la 3ème le code correspondant à un type d'hébergement.

1.1 AFC avec R

Lecture, description des données

Charger le fichier `vaccsp.dat`. Attention, les données ne sont pas sous une forme adaptée à la fonction `table`; il faut utiliser la fonction `xtable` pour définir une table de contingence.

```
vaccsp=read.table("vaccsp.dat",
  col.names=c("effectif", "CSP", "heber"))
summary(vaccsp)
# construire la table de contingence
T=xtable(effectif~CSP+heber,vaccsp)
# opération inverse
as.data.frame.table(T)
mosaicplot(T)
```

```
# test du chi2
chisq.test(T)
```

Conclusion sur la description de ces variables, de leur liaison.

AFC pas à pas

Calcul des marges, métriques et profils.

```
# Fréquences marginales ou barycentres
n=sum(T); gc=apply(T,2,sum)/n; gr=apply(T,1,sum)/n
# Métriques
Dr=diag(gr); Dc=diag(gc)
Dr1=diag(1/gr); Dc1=diag(1/gc)
# Profils lignes et colonnes
A=t(T)%*%Dr1/n; B= T%*%Dc1/n
#
A%*%B%*%gc ; gc
B%*%A%*%gr ; gr
```

Vérifier ci-dessus que les fréquences marginales ou barycentres sont bien les vecteurs propres associés à la valeur propre 1 des produits des matrices de profils.

Décomposition en valeurs singulières de la table de contingence. Attention, cette SVD est celle généralisée car calculée par rapport aux métriques dites du χ^2 . Se reporter au scénario concerné pour plus d'explications. Il n'est pas possible d'utiliser directement une fonction comme par exemple `prcomp` ou `princomp` pour calculer l'ACP des profils lignes et colonnes à cause de ce changement de métrique. Attention dans la construction des matrices, la matrice `T` est carrée mais il ne faut pas confondre le rôle des métriques.

```
# SVD de 1/nT relative à Dr-1 et Dc-1
Dr12=diag(1/sqrt(gr)); Dc12=diag(1/sqrt(gc))
res=svd(Dr12%*%T%*%Dc12/n)
res$d # valeur propre 1 à retirer
L=res$d[-1]
# Rechargement de métrique
U=diag(sqrt(gr))%*%res$u[-1]
V=diag(sqrt(gc))%*%res$v[-1]
```

```
# Vérifications de la chi2-orthonormalité
t(U)%*%Dr1%*%U
t(V)%*%Dc1%*%V
# Composantes principales ACP lignes et colonnes
Cr=Dr1%*%U%*%diag(L)
Cc=Dc1%*%V%*%diag(L)
# représentations graphiques
plot(Cc, type="n")
text(Cc, colnames(T))
abline(h=0,v=0)
plot(Cr, type="n")
text(Cr, rownames(T))
abline(h=0,v=0)
```

Rien de bien compliqué dans ces instructions mais autant utiliser une librairie qui propose des graphiques de meilleure qualité avec une offre détaillée d'options.

Avec FactoMineR

```
# Charger la librairie
library(FactoMineR)
# Analyse des correspondances
res.afc=CA(T)
summary(res.afc)
```

Comparer les résultats avec ceux précédemment obtenus.

Interprétation des graphiques : positions relatives des modalités "hébergement" puis des modalités "CSP" et enfin de celles des modalités des deux variables.

1.2 AFC avec SAS

Les mêmes résultats sont produits avec SAS.

Lecture et description des données

Exécuter le programme :

```
data sasuser.vaccsp;
```

```
infile "vaccsp.dat" ;
input eff csp $ heber $;
run;
```

Effectif des modalités et calcul du test du χ^2 d'indépendance entre les variables.

```
proc freq data=sasuser.vaccsp;
tables heber csp;
tables heber*csp /chisq;
weight eff; /* Effectifs de la table */
run;
```

AFC

L'AFC est obtenue avec les sorties graphiques par défaut.

```
proc corresp data=sasuser.vaccsp observed
out=resul dim=3;
tables csp,heber;
weight eff;
run;
```

Les graphiques fournis par le système ODS ne sont pas toujours très explicites ; ils sont également produits par la macro SAS à télécharger et exécuter avant son utilisation.

```
%gafcx(ident=num);
```

2 les élections présidentielles de 95

2.1 Les données

Pour le premier tour des élections présidentielles de 1995, on connaît, pour chacun des 95 départements métropolitains et la Corse¹ les informations suivantes :

- le nombre d'inscrits,

1. Les résultats des deux départements de Corse sont agrégés.

- le nombre de votants,
- le nombre de suffrages exprimés,
- et, dans l'ordre, le nombre de voix des candidats Villiers, Le Pen, Chirac, Laguillier, Chem., Jospin, Voynet, Balladur, Hue.

Charger le fichier [elec95.dat](#) puis lire les données et calculer le nombre des abstentions, des votes blancs ou nuls.

```
data sasuser.elec95;
infile "elec95.dat" dlm='09'x;
input num $ inscrits votants exprimes villiers
      le_pen chirac laguill cheminad jospin
      voynet balladur hue poids;
abstent=inscrits-votants;
blancs=votants-exprimes; run;
```

2.2 AFC simple

Exécuter le programme suivant :

```
proc corresp data=sasuser.elec95 observed
            out=resul dim=3;
var abstent blancs villiers le_pen chirac laguill
    cheminad jospin voynet balladur hue;
id num;
run;
%gafcx (num);
```

Commentaire à propos du vote “Villiers” en Vendée. Ce département est mis en “supplémentaire” dans l’analyse suivante en modifiant les pondérations :

```
proc corresp data=sasuser.elec95 observed
            out=resul dim=3;
var abstent blancs villiers le_pen chirac laguill
    cheminad jospin voynet balladur hue;
id num;
weight poids; run;
%gafcx (num);
```

Choisir le nombre d’axes, tenter de les interpréter. Que dire des problèmes posés par cette représentation graphique, que fait ressortir en priorité l’AFC ?

2.3 Comparaison avec l’ACP

Faut-il considérer des nombres de suffrages exprimés ou des taux de suffrage exprimés dans l’ACP ? Pourquoi ? Commenter le rôle de la métrique du χ^2 dans le cadre de l’AFC en comparant les résultats avec ceux de l’ACP des taux de suffrage exprimés pour chacun des candidats. Interpréter les résultats.

```
data sasuser.telec95;
set sasuser.elec95;
villiers=villiers/inscrits;
le_pen=le_pen/inscrits;
chirac=chirac/inscrits;
laguill=laguill/inscrits;
cheminad=cheminad/inscrits;
jospin=jospin/inscrits;
voynet=voynet/inscrits;
balladur=balladur/inscrits;
hue=hue/inscrits; run;
%acp (telec95, num, villiers--hue);
%gacpbx (p=9); %gacpsx (p=9); %gacpix; %gacpvx;
```

En conclusion de cette comparaison ACP/AFC, que dire du rôle de la métrique du chi-deux ?