TP: ACP de données “cubiques”

Résumé

Analyse de données "cubiques" par Analyse en composantes principales.

1 Objectif

L’objectif de cette séance est la mise en œuvre de l’analyse en composantes principales avec SAS sur un exemple plus réaliste de données socio-économiques se présentant sous la forme d’un cube de données, c’est-à-dire dépendant de trois indices : le numéro de ligne, le numéro de variable et l’année d’observation de cette variable. La spécificité de ces données nécessitent une adaptation des représentations graphiques. Les procédures et graphiques proposés en standard (insight, ODS) sont en effet vite limitées.

2 Pays de l’OCDE

2.1 Les données

Les données sont issues de l’Observatoire de l’OCDE. Pour chaque pays membre et pour chacune des années 75, 77, 79, 81, on connaît les valeurs prises par les variables suivantes qui sont toutes des taux :

- Taux brut de natalité,
- Taux de chômage,
- Pourcentage d’actifs dans le secteur primaire,
- Pourcentage d’actifs dans le secteur secondaire,
- produit intérieur brut (par habitant),
- Formation brute de capital fixe (par habitant),
- Hausse des prix,
- Recettes courantes (par habitant),
- Mortalité infantile,
- Consommation de protéines animales (par habitant),
- Consommation d’énergie (par habitant).

Elles sont disponibles dans le fichier : `ocde.dat`.

2.2 Lecture

Exécuter le programme suivant afin de lire les données et créer la table SAS correspondante :

```
data sasuser.ocde;
infile 'ocde.dat';
input pays $ natal chomage a_prim a_sec pib fbcf
         infl recc m_inf prot nrj;
run;
```

2.3 Description élémentaire

- Distributions, normalité, points atypiques...
- Corrélations, matrice des nuages de points.

Commentaire sur les distributions, les natures des liaisons.

3 ACP des pays de l’OCDE

Les mêmes variables sont observées, sur les mêmes pays ou individus à quatre dates différentes. Plusieurs stratégies d’analyse sont possibles (tableau moyen, tableaux concaténés, meilleur compromis). La plus adaptée pour ces données est de considérer les observations des variables pour chacun des “individus” pays$\times$années. Chaque pays est donc observé 4 fois et cette structure chronologique doit apparaître dans le graphique afin d’illustrer la dynamique économique de la période considérée.

3.1 ACP de base

Calculer l’ACP de ces données avec SAS/Insight : diagrammes en boîtes des composantes, biplot. Commentaire sur la dimension à retenir, les valeurs atypiques. L’axe 1 est-il sensible à la présence des valeurs atypiques ? Interpréter les axes. La représentation des individus est-elle simple à appréhender ?

3.2 Prise en compte du temps

Compte tenu de la structure particulière des données, une approche spécifique est nécessaire en utilisant les macros ci-dessous.

`%acp(ocde,pays,natal--nrj);`

Explorer rapidement les résultats fournis puis tracer les graphiques.

`%gacpbx;%gacpsx;`

Retrouver le nombre de composantes, les individus atypiques.

`%gacpvx;`

Interpréter les axes factoriels. Les deux premiers, le troisième ?

Le graphique des individus demande une adaptation de la macro `gacpix` afin de relier les pays dans l'ordre chronologique. Exécuter le programme ci-dessous.

```
%let x=1;
%let y=2;
 data anno;
 retain  xsys ysys '2';
 set coorindq nobs=nind;
 style='swiss';
 if mod(_n_, 4) ne 1 then delete ;
 y= prin&y;
 x= prin&x;
 text=ident;
 size=1;
run;
proc gplot data= coorindq;
 title;
 footnote ;
 axis1  length=14cm; /* attention taille */
 axis2  length=8cm;
 symbol v=dot i=join r=13 height=.5;
 plot prin&y*prin&x=ident / annotate=anno frame
href=0
      vref=0  nolegend haxis=axis1 vaxis=axis2;
run;
goptions reset=all;
quit;
```

Commenter le programme. Interpréter les résultats obtenus en prenant plus particulièrement en compte les profils des différents pays. Noter bien que tout logiciel, toute librairie, aussi rafinés soient-ils, trouvent leur limite dans la construction de graphiques spécifiques à un exemple particulier. Ainsi FactoMineR propose de nombreuses options graphiques pour la représentation des plans factoriels mais pas celle permettant de relier le spoint entre eux ; il faut revenir, comme avec SAS, aux commandes de base du logiciel R.