Scénario: ACP de températures et de données “cubiques”

Résumé

Illustration de l'Analyse en Composantes Principales sur deux jeux de données, l'un de courbes de température et le deuxième de données socio-économique "cubiques".

1 Exploration de courbes de températures

1.1 Données

Les données étudiées sont celles du fichier `tempR.dat` disponibles sur le site wikistat/data. Il contient les moyennes, entre 1931 et 1960, des températures mensuelles moyennes de 36 villes françaises. La première variable correspond au nom de la ville (4 caractères), les 12 suivantes représentent chacune un mois de l'année (source : Mémorial de la Météorologie nationale). Une moyenne journalière est la moyenne des températures min et max permettant ensuite de calculer la moyenne mensuelle. Moyenner ensuite sur 10 ces valeurs conduit à des courbes relativement régulières.

```
# lire les données dans R
temp=read.table("tempR.dat")
# vérification et stat élémentaires
summary(temp)
```

1.2 Exploration uni et bidimensionnelle avec R

```
# Ensemble des courbes
ts.plot(t(temp))
# toutes les distributions
boxplot(temp)
```

Que dire des dispersions, de l'homogénéité des unités et des variances.

```
# Corrélations
cor(temp)
plot(temp$janv,temp$fevr)
plot(temp$janv,temp$juin)
pairs(temp)
```

Que dire de la structure de corrélation très particulière de ces données ?

1.3 Analyse en composantes principales avec R

```
acp=princomp(temp)
summary(acp)
plot(acp)
boxplot(data.frame(acp$scores))
biplot(acp) # analyser les échelles des axes
```

1. Expliciter le choix entre ACP réduite ou non, comparer les différences.
2. Combien d'axes faut-il retenir ? Justifier.
3. Identifier la ville atypique de l'axe 2. Que faire ?
4. Interprétation des axes.
5. Commenter la position d'Embrun sur les graphiques.

Les données sont ici très particulières, des courbes fonction du temps. En conséquence, les vecteurs propres le sont également et les courbes sont décomposées sur cette base de “fonctions” discrétisées. Vecteurs ou plutôt “fonctions propres” de l'ACP.

```
plot.ts(acp$loadings[,1:6],
   main="Fonctions propres")
```

1.4 Librairie FactoMineR

Cette librairie apporte des compléments intéressants (qualité et options des graphiques, gestion des variables manquantes) et surtout elle vient particulièrement compléter les fonctions de base de R pour l'analyse des variables qualitatives. Voici les principaux résultats de l'ACP.

Comparer avec les résultats numériques précédemment obtenus.

```
library(FactoMineR)
acp=PCA(temp, scale.unit=FALSE,ncp=12,graph=T)
barplot(acp$eig[,1])
boxplot(acp$ind$coord)
acp$svd$V
dimdesc(acp,axes=c(1,2))
acp=PCA(temp, scale.unit=TRUE,ncp=12,graph=T)
```

Il s’agit également de résoudre la question concernant l’observation atypique sur l’axe 2. Faut-il la conserver ? Cette question est abordée en considérant deux ACPs, celle avec et sans ce point afin de s’assurer que sa suppression ne perturbe pas trop les premiers axes, notamment le 2ème. Comparer la représentation des variables avec celle obtenue ci-dessous en considérant supplémentaire la ville atypique .

```
acp=PCA(temp, scale.unit=TRUE,ncp=12,graph=T,
   ind.sup=8)
```

Peut-on conserver cette ville dans l’analyse ?

Cette librairie ajoute à ces techniques exploratoires des éléments “inférentiels” : p-valeurs de test, ellipse de confiance... supposant implicitement un modèle probabiliste (distributions gaussiennes multidimensionnelles) ; ils sont à utiliser avec prudence, plus comme des indicateurs que comme des aides formelles à la décision.

1.5 ACP avec SAS

Afin de pouvoir comparer la cohérence des résultats, voici ceux fournis par le logiciel SAS.

Les données

Les données sont cette fois dans le fichier `temp.dat` à télécharger.

```
data sasuser.tempville;
infile "temp.dat" ;
input ville $ janv fevr mars avri mai juin juil aout
      sept octo nove dece;
run;
```

ACP

Les options par défaut fournissent les principaux résultats mais ceux-ci de sont guère modifiables. Les mettre en forme demande nettement plus d’efforts, ce sera entreprit dans d’autres scénarios. L’ACP est-elle réduite ?

```
proc princomp data=sasuser.tempville cov
  plots(ncomp=2)=(pattern(circles=1.0)score);
   var janv--dec;
   id ville;
run;
```

Comparer les graphiques fournis par SAS avec ceux de R.

2 Donées socio-économiques "cubiques"

2.1 Introduction

Objectif

L’objectif de cette section est l’exploration de données socio-économiques plus complexes. La principale spécificité de ces données est de se présenter sous la forme d’un “cube” de données ou tableau à trois entrées : le numéro de ligne, le numéro de variable et l’année d’observation de cette variable. Après une description classique, la mise en œuvre de l’analyse en composantes principales avec SAS nécessitent une adaptation des représentations graphiques. Les procédures et graphiques proposés en standard sont en effet vite limités pour permettre de construire des représentations faciles à interpréter. Plus de compétences en programmation SAS (ou R) s’avèrent donc vite indispensables.

Les données

Les données sont issues de l’Observatoire de l’OCDE. Pour chaque pays membre et pour chacune des années 1975, 1977, 1979, 1981, 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.

Les mêmes variables sont donc 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×années.

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  ;
year = 1975+2*mod(_n_ - 1, 4);
run;
```

2.2 Description élémentaire

Le module d’exploration interactif des données (SAS/Insight) ayant disparu avec la version 9.4, il devient vite lourd d’explorer rapidement un jeu de données de taille raisonnable. L’usage de SAS ne se justifie alors que par le traitement répétitif de jeux de données volumineux. La stratégie de SAS fut d’éviter les doublons notamment avec l’autre logiciel interactif de la marque : JMP. Mais ce dernier nécessite un apprentissage qui s’avère moins intuitif que SAS/Insight.

Il faudra donc se contenter des fonctions rudimentaires de SAS et SAS/Stat pour l’exploration qui ont le défaut de produire beaucoup (trop ?) de résultats avec SAS

Description unidimensionnel

Les années sont tout d’abord regroupées pour produire toutes les statistiques univariées avant de les calculer par année.

```
proc univariate data=sasuser.ocde normal plot;
   var natal--nrj;
run;
proc sort data=sasuser.ocde out=tocde;
by year;
proc univariate data=tocde plot;
   var natal--nrj;
          by year;
run;
```

Consulter “rapidement” ces résultats ; Que dire à propos de la symétrie des distributions, de leur normalité, des valeurs atypiques.

Description bidimensionnel

```
proc corr data=sasuser.ocde  plots=matrix;
   var natal--nrj;
run;
```

Que dire de la structure de corrélation ?

2.3 ACP

Chaque pays étant observé 4 fois, la principale difficulté technique est de faire apparaître cette structure chronologique dans les graphique afin d’illustrer la dynamique économique de la période considérée.

ACP de base

Calculer l’ACP de ces données avec la procédure de base.

```
proc princomp data=sasuser.ocde
  plots(ncomp=2)=(pattern(circles=1.0)score);
  var natal--nrj;
```

```
  id pays;
run;
```

Commentaire sur la dimension à retenir. La représentation des individus est-elle simple à appréhender ? Les observations atypiques sont-elles faciles à identifier ?

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

Macros commandes spécifiques

Pour compléter les résultats, des macros-commandes permettent de détailler les calculs avant de préparer les graphiques souhaités. Pour ce faire il est nécessaire de les télécharger puis de les faire exécuter par SAS afin qu'elles soient connues du système. Télécharger les fichiers :
`acp.sas, gacpix.sas, gacpvx.sas, gacpsx.sas, gacpbx.sas`
du répertoire
`http://www.math.univ-toulouse.fr/~besse/pub/sas`

Une fois chaque fichier exécuté, lancer successivement les commandes suivantes en prenant le temps de comprendre les sorties (`output`) ainsi que les graphes :

```
%acp(ocde,pays,natal--nrj);
%gacpbx;%gacpsx;
%gacpvx;
%gacpix;
```

Vérifier qu'il s'agit bien des mêmes résultats. Choix de la dimension, identification des observations atypiques. Le graphe des individus n'apportent rien de plus si ce n'est une indication de la qualité des représentations proportionnelle à la taille des libellés. Identifier les observations atypiques dont on peut s'inquiéter du rôle possible sur l'orientation du premier axe.

La macro admet une variable supplémentaire permettant de définir un poids. Cette variable est ajoutée à la table avant de ré-exécuter l'ensemble.

```
data sasuser.p_ocde;
set sasuser.ocde;
if pays eq "PO" then poids=0 else ; Poids=1;
run;
%acp(p_ocde,pays,natal--nrj,poids=poids);
%gacpbx;%gacpsx;
%gacpvx;
%gacpix;
```

Les observations de poids nulles sont toujours présentes en supplémentaires mais n'influencent plus les calculs des axes. Conclusion sur leur effet ?

Prise en compte du temps

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

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 raffiné soient-il, trouve ses 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 les point entre eux ; il faut revenir, comme avec SAS, aux commandes de base du logiciel R.