TP: ACP élémentaire d'un fichier de températures

Résumé

Pratique élémentaire de l'Analyse en composantes principales avec les logiciels SAS et R.

1 Avertissement

- Les différents travaux et analyses proposés tout au long de ces documents sont largement explicités. Les commandes en R ou SAS sont toutes fournies. L'important n'est pas de trouver la bonne syntaxe des commandes ou de finir au plus vite mais de réfléchir sur les méthodes, leurs conditions d'applications, les résultats obtenus. L'apprentissage de ces logiciels et de leur programmation est un autre cours.
- Il est possible de directement copier les commandes de l'affichage de ce texte vers une fenêtre d'édition mais **attention** certains caractères, "modifiés" par les normes d'affichage réservent des surprises, notamment le caractère " ′ " (quote) n'a pas la même fonction que le caractère " ' " (apostrophe). Il est vivement conseillé d'ouvrir une fenêtre d'édition de texte (xemacs, kile, notepad...), pas de traitement de texte (ni word, ni open office), qui contiendra les différentes commandes a faire exécuter.
- Il est aussi possible de prendre le temps d'entrer les commandes au clavier, cela laisse le temps de réfléchir !

2 Objectif

L'objectif de cette séance est d'aborder l'analyse en composantes principales sur un exemple simple afin d'évaluer les différentes possibilités offertes par les logiciels disponibles. Les données étudiées sont celles du fichier `temp.dat`. 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).

Remarque : tous les fichiers de données ainsi que les macro commandes SAS sont disponibles sur le site wikistat/data. Télécharger le fichier `temp.dat`.

3 Avec SAS/Insight

3.1 Lecture des données

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

Lance le module d'analyse interactive des données, ouvrir la table SAS qui vient d'être créée puis déclarer la variable `ville` comme `label`.

3.2 Exploration élémentaire

1. Etudier les distributions de chaque variable (diagramme boîte parallèle de toutes les variables avec `Box plot/ Mosaic plot`), vérifier le comportement correct de ces distributions, l'absence de valeur atypique, l'homogénéité des variances.
2. Tracer la matrice des nuages de points (scatterplot) : commentaire sur la structure particulière des corrélations.

3.3 ACP

1. Le menu `multivariate` d'Insight propose une ACP. Lister les options proposées.
2. Choisir une ACP non réduite (pourquoi ?), demander le calcul de toutes les composantes principales.
3. Tracer les boîtes parallèles de ces composantes. Commentaire sur le choix du nombre de dimension c'est-à-dire le nombre de composantes à retenir.
4. Identifier la ville atypique.
5. Exclure puis ré-inclure cette ville des calculs. Les résultats en sont-ils modifiés ? Que dire de la stabilité du 2ème axe ?
6. Interprétation des axes.

7. Sauver le graphe obtenu (biplot) afin d'en comparer les coordonnées avec les autres approches.

4 Avec des macros ad'hoc de SAS

Télécharger et exécuter dans SAS les macros contenus dans les fichiers : `acp.sas, gacpix.sas, gacpvx.sas, gacpsx.sas, gacpbx.sas`
de `http://www.lsp.ups-tlse.fr/Besse/pub/sas`

puis exécuter successivement en prenant le temps de comprendre les sorties (`output`) ainsi que les graphes :

```
%acp(tempville, ville, janv fevr mars avri mai juin
     juil aout sept octo nove dece,red=cov);
%gacpsx;
%gacpbx;
%gacpix;
%gacpvx;
```

1. L'acp est-elle réduite ?
2. Comparer avec les sorties de SAS/Insight
3. Avec quelles coordonnées sont représentées les villes ?
4. Avec quelles coordonnées sont représentées les variables ?

5 Avec SAS/Stat de base et ODS

Exécuter les commandes ci-dessous si ça marche sinon sauter cette section.

```
ods html;
ods graphics on;
proc princomp data=sasuser.tempville cov;
var janv--dece;
run;
ods graphics off;
ods html close;
```

1. Apprécier le temps d'exécution.
2. Ouvrir le fichier `sashtml.htm`
3. L'ACP est-elle réduite ?
4. Tous les graphiques sont-ils pertinents ? Lequel manque-t-il ?
5. Comparer avec les résultats précédents.
6. Interpréter : "principal components pattern profiles"

6 ACP avec R

6.1 Fonction de base

Editer le fichier `temp.dat` pour introduire une première ligne contenant le nom des 12 variables séparées par un espace. Sauver ce fichier sous un nouveau nom : `temp-r.dat`. Lancer R dans une fenêtre "console" puis exécuter successivement les commandes suivantes :

```
temp=read.table("temp-r.dat")
summary(temp)
plot(temp)
acp=princomp(temp)
summary(acp)
plot(acp)
attributes(acp)
boxplot(data.frame(acp$scores)) # Que contient
                                # l'attribut "scores" ?
biplot(acp) # analyser les échelles des axes
acp$loadings # Que sont les "loadings" ?
```

Comparer avec les sorties précédentes. Est-il simple de bien déterminer quelle matrice de coordonnées est utilisée pour représenter les variables ?

Commenter la position d'Embrun sur les graphiques.

6.2 Librairie FactoMineR

Développée à l'Agrocampus de Rennes (http ://factominer.free.fr) cette librairie est principalement dédiée aux méthodes statistiques factorielles. Elle

apporte des compléments intéressants (qualité et options des graphiques, gestion des variables manquantes) et 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.

```
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)
```

Cette librairie ajoute dans les techniques exploratoires des éléments : p-valeurs de test, ellipse de confiance... supposant implicitement un modèle probabiliste ; ils sont à utiliser avec prudence, plus comme des indicateurs que comme des aides formels à la décision.