Analyse de données
M1 Statistique et économétrie
V. Monbet

Analyse en composantes principales

Les objectifs de ce TD sont

1. de revoir le cours sur l'analyse en composantes principales,
2. d'apprendre à utiliser l'analyse en composantes principales sous SAS et sous R,
3. d'apprendre à interpréter les résultats.

1 Questions de cours

1. Vérifier que $n^{-1}ZZ^T$ est la covariance de la matrice de données centrées, avec Z la matrice formée par les colonnes $z_k = Xu_k$ où X représente la matrice de données et u_k le $k$ième axe factoriel.
2. Supposons que nous avons p variables aléatoires indépendantes et identiquement distribuées. De quelle forme seront les vecteurs propres correspondant au premier facteur ? Donner les valeurs propres.
3. Supposons que l'on a deux variables aléatoires, X_1 et $X_2 = 2X_1$. Donner la forme des valeurs propres et des vecteurs propres de leur matrice de corrélation ? Combien de valeurs propres sont non nulles ?
4. Quel pourcentage d'inertie est expliqué par le premier facteur dans la question précédente ?

2 Les données

Les données (fichier `depart.dat` sur le forum) proviennent du Groupe d'Etude et de Reflexion Inter-régional (GERI). Elles portent sur 4 grands thèmes : la démographie, l'emploi, la fiscalité directe locale, la criminalité. Les indicateurs sont mesurés sur l'ensemble des départements français métropolitains ainsi que la Corse pendant l'année 1990, ils sont, pour la plupart, des taux calculés relativement à la population totale du département concerné. On observe les variables suivantes :

— numéro de département,
— code du département,

— code de la région,
— `URBR` indicateur de concentration de la population mesurant le caractère urbain ou rural du département,
— `TXCR` taux de croissance de la population sur la période intercensitaire 1983-1990,
— `JEUN` part des 0-19 ans dans la population totale,
— `AGE` part des plus de 65 ans dans la population totale,
— `FE90` taux de fécondité (pour 1000) égal au nombre de naissances rapporté au nombre de femmes fécondes (15 à 49 ans) en moyenne triennale,
— `ETRA` part des étrangers dans la population totale,
— `CHOM` taux de chomage,
— `CRIM` taux de criminalité : nombre de délits par habitant,
— `FISC` produit, en francs constants 1990 et par habitant des quatre taxes locales (professionelle, habitation, foncier bati, foncier non bati).

On observe aussi les parts de chaque profession en catégorie socioprofessionnelle (PCS) dans la population active occupée du département :

— `AGRI` : agriculteurs,
— `ARTI` : artisans,
— `CADR` : cadres supérieurs,
— `EMPL` : employés,
— `OUVR` : ouvriers,
— `PROF` : professions intermédiaires.

3 Analyse en composantes principales sous R

3.1 Les différentes étapes de l'ACP

Nous allons décomposer les différentes étapes de l'ACP.

1. Centrer les données : `dep.c <- scale(dep[,4:18],center=TRUE,scale=FALSE)`
2. Estimer la covariance : `C = cov(dep.c)` .
   Vous pouvez afficher le résultat. Quelles sont les variables qui ont une forte variance ? Certaines variables sont-elles fortement dépendantes ?
   Une telle covariance est difficile à interpréter. Il est alors utile de tracer le cercle des corrélation dans les premiers plans factoriel (voir plus loin).
3. Calculer les valeurs propres et vecteurs propres de la matrice C : `eigen(C)`.
   Que peut-on dire des composantes principales ? Pourquoi le premier vecteur propre a t'il une première composante qui domine largement les autres ? Pour obtenir les variance de tous les variables, on peut faire `apply(dep.c,2,var)`
4. Reprendre les questions 2 et 3 ci dessus avec les données centrées et réduites : `dep.red <- scale(dep[,4:18],center=TRUE,scale=TRUE)`.
   Quelles sont les deux variables contribuent le plus au premier axe factoriel ? Avec quel signe ?

3.2 L'ACP en utilisant les routines existantes

Répondre aux mêmes questions que dans la partie SAS en utilisant le logiciel R et en vous aidant des commandes ci-dessous. On utilise soit la fonction `princomp` soit la fonction `prcomp`. On peut, bien sûr, arriver aux mêmes résultats et produire les mêmes graphiques en utilisant l'une ou l'autre de ces fonctions.

```
# Analyse en composantes principales
dep.red <- scale(dep[,4:18],center=TRUE,scale=TRUE)
acp.dep <- princomp(dep.red)
summary(acp.dep)
str(acp.dep) # donne un aperçu exhaustif des éléments de l'objet acp.dep
biplot(acp.dep)
loadings(acp.dep)
```

Comparer la sortie de `loadings` avec les vecteurs propres obtenus à la question 4 de la section **??**.

Expliquer ce que font les commandes ci-dessous.

```
# Une autre fonction
pc.s = prcomp(dep[,4:18],scale=TRUE,center =TRUE)
summary(pc.s)
plot(pc.s$x[, 1], pc.s$x[, 2], pch = 20, xlab = "Standardised PC 1",
+ ylab = "Standardised PC 2")
abline(h = 0)
abline(v = 0)
abline(h = 3*sqrt(pc.s$sdev[2], col = "red", lty = 2)
abline(v = 3*sqrt(pc.s$sdev[1], col = "red", lty = 2)
abline(h = -3*sqrt(pc.s$sdev[2], col = "red", lty = 2)
abline(v = -3*sqrt(pc.s$sdev[1], col = "red", lty = 2)
(pts <- which((abs(pc.s$x[, 1]) >= 3*sqrt(pc.s$sdev[1]) |
       (abs(pc.s$x[,2]*sqrt(pc.s$sdev[2]) >= 3)))
points(pc.s$x[pts, 1], pc.s$x[pts, 2], pch = 21, col = "red", bg = "blue")
text(pc.s$x[pts, 1], pc.s$x[pts, 2], pts, pos = 4, col = "red")
pc.s$x[pts, c(1, 2)]
```

Il est intéressant de regarder de plus près les résidus, c'est à dire la différence entre les variables d'origine (éventuellement centrées et réduites) et les variables reconstruites à partir de l'ACP. En effet, $XE = S$ avec X la matrice des observations, E la matrice des vecteurs propres et S les composantes principales. On a donc aussi (si toutes les valeurs propres sont non nulles) $X = SE^{-1}$. On peut alors comparer X aux variables reconstruites quand on projette dans un espace de dimension réduite. Que font les commandes suivantes ? Interpréter les résultats.

```
res.reconstr =  dep.red-acp.dep$score%*%solve(acp.dep$loadings)
apply(res.reconstr,2,var)
```

```
res.reconstr.1 =  dep.red
     -matrix(acp.dep$score[,1],95,1)%*%matrix(solve(acp.dep$loadings)[1,],1,15)
apply(res.reconstr.1,2,var)
res.reconstr.2 =  dep.red-acp.dep$score[,1:2]%*%solve(acp.dep$loadings)[1:2,]
apply(res.reconstr.2,2,var)
```

On peut aussi utiliser le package FactoMineR.

```
library(FactoMineR)
# Analyse en composantes principales
pca.dep <- PCA(dep.red)
summary(pca.dep)
str(pca.dep) # donne un aperçu exhaustif des éléments de l'objet acp.dep
```

1. Combien d'axes choisit-on selon la règle de l'ébouli des valeurs propres ? Selon la règle de Kaiser ?
2. Qui sont les individus qui contribuent le plus au premier axe factoriel ? Au second ? Au trois premiers conjointement ?
3. Quelles sont les variables qui contribuent le plus au premier axe factoriel ? Au second ? Au trois premiers conjointement ?
4. Interpréter les axes factoriels interprétables.
5. Que peut-on dire de l'individu 35 ?
6. Y a t'il des individus qui contribuent fortement aux deux premiers axes factoriels mais qui sont mal représentés ?
7. Peut-on interpréter la proximité entre le 06 et le 75 ? 35 et 14 ?
8. Que se passe-t-il si on fait l'ACP sur des données non normées ? Pourquoi ?

4 Analyse en composantes principales sous SAS

Compiler les macros SAS du fichier `SASMACRO/toutacp.sas`. Nous allons utiliser ces macros pour réaliser une analyse en composantes principales. Exécutez les commandes suivantes et/ou consulter le code des différentes macros pour répondre aux questions ci-dessous.

```
%let listev = txcr etra urbr jeun age
      chom agri arti empl ouvr prof fisc crim fe90 ;
%acp(TPexplor.depart,num,&listev) ;
%gacpsx ;
%gacpvx(x=1,y=2,nc=4) ;
%gacpix(x=1,y=2,nc=3) ;
/* En utilisant la région comme identifiant */
%acp(TPexplor.depart,region,&listev) ;
%gacpix(x=1,y=2,nc=3) ;
```

1. Que font les commandes ci-dessus ?
2. Quand on réalise l'ACP, travaille-t-on avec le tableau des données réduites ou non ? A quoi sert le fait de réduire les données ? Que représente l'origine dans le graphe des individus ?
3. Quel est le pourcentage d'inertie restitué par le premier plan factoriel ? Combien de facteurs pensez-vous qu'il faut conserver pour bien résumer l'information ? Justifier la réponse.
4. Commentez le cercle des corrélations puis le graphique des individus identifiés par les codes de régions.
5. Projeter les régions sur le deuxième plan factoriel. Interpréter.