TP 3 : ACP et AFD

1 Données `départements` : ACP et AFD

1.1 Présentation

Les données proviennent du Groupe d'Etude et de Réflexion Inter-régional (GERI). Elles décrivent quatre grands thèmes : la démographie, l'emploi, la fiscalité directe locale et la criminalité de chacun des départements français métropolitains et la Corse. Les indicateurs ont été observés pendant l'année 1990, ils sont, pour la plupart des taux calculés relativement à la population totale du département concerné ainsi que la part de chaque profession et catégorie socio-professionnelle dans la population active occupée du département . Voici leur liste :

- Identificateur du numéro du département
- Identificateur du code du département
- Identificateur du code de la région
- TXCR : taux de croissance de la population sur la période intercensitaire 1982-1990
- ETRA : part des étrangers dans la population totale
- URBR : indicateur de concentration de la population mesurant le caractère urbain ou rural d'un département
- JEUN : part des 0-19 ans dans la population totale
- AGE : part des plus de 65 ans dans la population totale
- CHOM : taux de chômage
- AGRI : agriculteurs
- ARTI : artisans
- CADR : cadres supérieurs
- EMPL : employés
- OUVR : ouvriers
- PROF : professions intermédiaires
- FISC : produit en francs constants 1990 et par habitant des quatres taxes directes locales (professionnelle, habitation, foncier bâti et foncier non-bâti)
- CRIM : taux de criminalité (nombre de délits par habitant)
- FE90 : taux de fécondité pour 1000 égal au nombre de naissances rapportés au nombre de femmes fécondes (de 15 à 49 ans) en moyenne triennale

1.2 ACP du tableau de départements

- Lire les données depart.dat et créer la table `dpt`.
- Renommer les variables. Extraire dans la table `dpt0` la table `dpt` décrite par les seules variables quantitatives intéressantes.
- Effectuer une ACP réduite du tableau `département`.
- Donner les représentations graphiques issues de l'acp en utilisant le nombre d'axes que l'on aura pris soin de garder pour la représentation. Justifier ce nombre d'axes.
- Interpréter ces axes.

1.3 AFD du tableau de départements

On se propose de mettre en évidence les plus grandes disparités inter-régionales et donc de rechercher les variables ou combinaisons de variables expliquant au mieux le découpage régional. Pour simplifier, nous procédons à des regroupements afin de construire des régions moins nombreuses comprenant des nombre de départements plus semblables. D'autre part, la région "Ile de France", trop particulière et donc trop facile à discriminer est laissée à part.

- NPC+Pic+HNo+ChA = Nd
- Als+Lor+FrC=Es
- BNo+Bre+PaL=Ws
- Cen+Bou=CN
- PoC+Lim=CW
- Auv+RhA=CE
- Aqu+MiP=SW
- LaR+PAC+Cor=SE

- Créer un vecteur nul de longueur correspondant au nombre de départements. Affecter à chacun des éléments de ce vecteur l'un des 8 groupes associés aux affectations définies précédemment.

```
n<-rep(0,95); n[dpt[,3]=="NPC"]<-1
n[dpt[,3]=="Pic"]<-1; n[dpt[,3]=="HNo"]<-1
n[dpt[,3]=="ChA"]<-1; n[dpt[,3]=="Als"]<-2
n[dpt[,3]=="Lor"]<-2; n[dpt[,3]=="FrC"]<-2
n[dpt[,3]=="BNo"]<-3; n[dpt[,3]=="Bre"]<-3
n[dpt[,3]=="PaL"]<-3; n[dpt[,3]=="Cen"]<-4
n[dpt[,3]=="Bou"]<-4; n[dpt[,3]=="PoC"]<-5
n[dpt[,3]=="Lim"]<-5; n[dpt[,3]=="Auv"]<-6
n[dpt[,3]=="RhA"]<-6; n[dpt[,3]=="Aqu"]<-7
n[dpt[,3]=="MiP"]<-7; n[dpt[,3]=="LaR"]<-8
n[dpt[,3]=="PAC"]<-8; n[dpt[,3]=="Cor"]<-8
indice=n>0
m<-subset(n,n>0)
xc<-1:95
ind<-xc[indice]
departements<-dpt0[ind,]
```

```
lbls<-c("Nord","Est","Ouest","CentreNord","CentreOuest","CentreEst",
              "SudEst","SudOuest")
partition=factor(m,labels=lbls)
```

- En utilisant la fonction `discrimin` de la librairie `MASS`, effectuer l'afd sur le tableau des départements précédents. Commenter.

```
library(ade4)

departements.afd <- discrimin(dudi.pca(departements,scan=FALSE),
                partition,scan=FALSE)
departements.afd
plot(departements.afd)
```

2 Données `Iris` : AFD et AFD décisionnelle

Nous reprenons dans cette section le jeu de données bien connu des `iris`, déjà vu dans un TP précédent.

Dans un premier temps, pour explorer les données, on va chercher à étudier la discrimination qui existe entre les différentes espèces dans des espaces de dimension 2 ou 3 à partir des données initiales.

```
data(iris)
attach(iris)

# Discrimination en 2 dimensions #
plot(iris[,1],iris[,2],col=c("blue","red","green")[Species])

# Discrimination en 3 dimension #
library(rgl)

plot3d(iris[,1],iris[,2],iris[,3],col=c("blue","red","green")[Species],type="s")
```

Nous réalisons maintenant l'AFD sur le jeu de données complet. Commenter.

```
library(ade4)

iris.afd <- discrimin(dudi.pca(iris[,1:4],scan=FALSE),Species,scan=FALSE)
iris.afd
plot(iris.afd)
```

Nous réalisons maintenant une AFD dans un but d'apprentissage et de prédiction. Dans ce cas, on parle d'AFD décisionnelle. On va estimer les paramètres du modèle statistique sous-jacent à la méthode d'AFD sur un échantillon de taille 75 et on cherchera ensuite à prédire les espèces des iris présents dans l'échantillon de test (échantillon restant). On étudiera le pourcentage de mal classés.

```
library(MASS)

train <- sample(1:150, 75)
table(Species[train])

iris.afd <- lda(Species ~ ., iris, prior = c(1,1,1)/3, subset = train)
iris.afd

pred <- predict(iris.afd, iris[-train, ])$class

table(Species[-train],pred)
```