Fiche TD avec le logiciel R : tdr1104

Quelques exemples d'analyses en composantes principales

A.B. Dufour & D. Clot

Table des matières

1 La sécurité routière dans les départements français **1**

2 Le classement de vins par des juges **2**

3 L'évolution des crimes aux USA de 1965 à 2005 **3**

4 Annexe **4**

1 La sécurité routière dans les départements français

Les données proviennent du site officiel `http://www.securite-routiere.gouv.fr/`. Le fichier contient 10 colonnes : le nom du département français, le numéro du département, la région auquel appartient le département, le numéro de la région, le nombre d'accidents corporels en 2010, le nombre d'accidents corporels en 2009, le nombre de tués sur les routes en 2010, le nombre de tués sur les routes en 2009, le nombre de blessés en 2010, le nombre de blessés en 2009, le nombre d'habitants estimé au 1er janvier 2009 (INSEE) et le nombre de tués par million d'habitants en 2010.

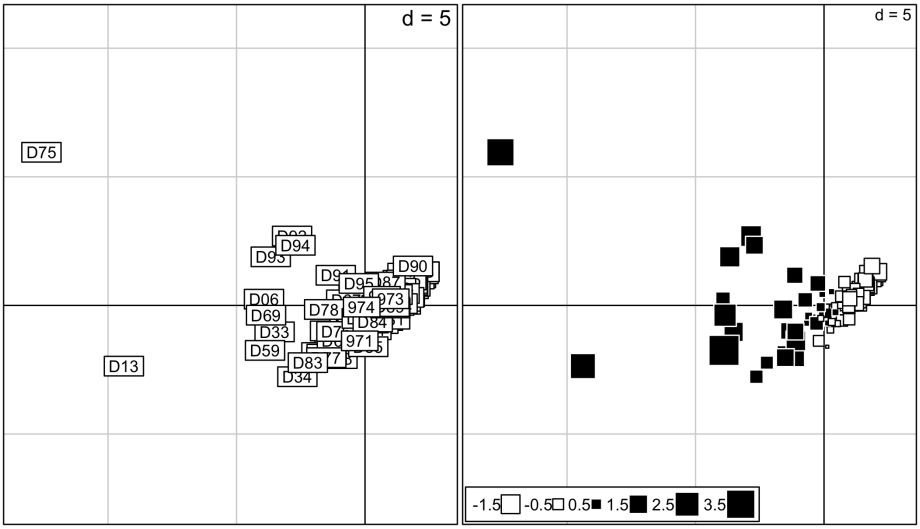
```
 [1] "departement"  "numdep"       "region"       "numregion"    "acc.corps.10"
 [6] "acc.corps.09" "tues.10"      "tues.09"      "blesses.10"   "blesses.09"
[11] "population"   "ratio"
```

On réalise une analyse en composantes principales normée sur les variables nombre d'accidents corporels, nombre de tués sur les routes et nombre de blessés en 2009 et 2010.

```
library(ade4)
acpSR <- dudi.pca(SR0910[,5:10],center=TRUE, scale=TRUE, scannf=FALSE, nf=2)
```

1. Représenter le graphe des valeurs propres à l'aide d'une des deux fonctions `barplot` ou `screeplot`, selon la version d'`ade4`. Le choix de conserver deux facteurs est-il justifié ?
2. Interpréter le cercle des corrélations.
3. Peut-on faire une typologie des départements ?
4. Commenter le lien entre le cercle des corrélations et la représentation des inividus-départements (fonction `scatter`).
5. Calculer la corrélation entre la population et les coordonnées des départements sur le premier axe principal. Commenter.
6. On propose la représentation ci-dessous où l'information sur la population a été rajoutée sur le premier plan factoriel. Commenter.

```
par(mfrow=c(1,2))
s.label(acpSR$li,label=SR0910$numdep,clabel=0.75)
popu <- scale(SR0910$population,center=TRUE,scale=TRUE)
s.value(acpSR$li[,1:2],popu)
```



2 Le classement de vins par des juges

Vingt-cinq juges ont gouté et classé 8 vins du Beaujolais. Deux questions peuvent alors être soulevées et résolues par des analyses en composantes principales.

```
data(macon)
macon
```

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y
A	5	5	4	3	3	4	7	2	1	3	5	4	4	5	4	8	5	7	8	5	4	6	7	2	8
B	4	8	2	4	1	5	2	7	8	8	1	6	3	7	8	5	7	8	1	4	1	5	4	4	6
C	2	6	1	1	6	2	1	5	5	4	3	7	2	2	6	2	1	6	2	1	2	1	2	5	1
D	6	7	5	8	2	6	8	8	6	6	6	5	6	6	3	6	8	1	7	6	7	4	1	6	7
E	1	4	3	2	7	1	6	4	3	1	2	8	1	1	1	3	2	2	6	2	8	2	8	1	2
F	3	2	8	6	5	8	3	3	4	7	8	1	5	8	7	4	4	3	3	8	6	8	6	7	3
G	7	1	6	5	4	7	4	1	7	5	7	3	8	3	2	7	3	5	4	7	3	7	3	8	5
H	8	3	7	7	8	3	5	6	2	2	4	2	7	4	5	1	6	4	5	3	5	3	5	3	4

- ⋆ Quelle est la cohérence du jury ? Peut-on exprimer un compromis entre jugements, un choix collectif ? Peut-on mettre en évidence la ressemblance entre juges ?
  Les juges sont en colonnes dans une ACP normée. Les moyennes sont toutes égales, les variances aussi, la normalisation n'est ni nécessaire, ni nuisible. On peut l'utiliser pour tracer les cercles de corrélation.

- ⋆ Peut-on faire une typologie des juges ? mettre en évidence ce qui les opposent, montrer qu'il existe plusieurs types de jugements ?
  Les juges sont en lignes dans une ACP centrée. On laissera ainsi dominer dans l'analyse les produits qui ont reçu les appréciations les plus variables.

Les deux approches sont antinomiques.
Réaliser les analyses associées à ces deux approches et discuter les résultats.

3 L'évolution des crimes aux USA de 1965 à 2005

Les données sont réparties dans quatre colonnes : l'état, l'année de recueil de l'information, le nombre d'habitants et le nombre de crimes violents.

```
[1] "Etat"          "Date"          "Population"    "Crime_Violent"
```

La table de contingence associant les états et les années (les dates posées en variable qualitative) montre que l'état de New-York n'a pas été échantillonné de 1960 à 1964. C'est pourquoi on étudiera l'évolution de la criminalité de 1965 à 2005. Les années sont passées en colonnes et les états en lignes.

```
rapport <- CSD$Crime_Violent/CSD$Population
rapport <- rapport*1000
annees <- as.character(1965:2005)
#
crimes <- matrix(0,ncol=length(annees),nrow=51)
for (i in 1:length(annees)) crimes[,i] <- rapport[CSD$Date==annees[i]]
crimes <- as.data.frame(crimes)
rownames(crimes) <- unique(CSD$Etat)
colnames(crimes) <- paste("A",annees,sep="")
```

L'objectif est de réaliser une typologie des Etats à partir des courbes des profils de criminalité au cours du temps.

```
tprofils <- apply(crimes,1,function(x) x/sum(x))
profils <- t(tprofils)
```

1. Calculer les sommes en lignes du data frame `profils`
2. Réaliser une analyse en composantes principales centrée sur les profils.
3. Combien y-a-t-il de valeurs propres ? Ce résultat était-il attendu ? Pourquoi ?
4. Représenter graphiquement les états sur le premier plan factoriel. En vous aidant des courbes de l'annexe, tenter une interprétation.

4 Annexe

Dans l'optique de bien interpréter la position des états dans l'ACP sur les profils, on a représenté les 51 courbes.