Analyse de données
M1 Statistique et économétrie - 2012-2013
V. Monbet
Analyse en composantes principales (suite)

Ce TD constitue un complément sur l'analyse en composantes principales et plus généralement sur l'analyse descriptive d'un jeu de données mutivarié.

1 Première partie

1.1 Construction de l'ACP et étude des résidus

Nous reprenons les données sur les départements français et nous travaillons sous R.

1. Faire l'analyse en composantes principales à la main (en utilisant la fonction `eigen`) puis en utilisant la fonction `princomp`. Vérifier qu'on obtient les mêmes valeurs propres et les mêmes vecteurs propres.
2. Construire les coordonnées des individus sur le premier axe factoriel en utilisant les résultats de l'ACP que vous avez réalisée à la main. Vérifier qu'on obtient bien les mêmes coordonnées que dans les résultats de `princomp`.
3. Utiliser les résultats de la question précédente pour tracer un graphique représentant la projection des individus sur le premier plan factoriel. On pourra utiliser la fonction `text` pour ajouter une étiquette à chaque point.
4. 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 ie les composantes principales. 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.

```
n = length(dep[,1])
X.res =  X-S%*%solve(E)
summary(X.res)
sd(X.res)
X.res1 =  X-matrix(S[,1],n,1)%*%matrix(solve(E)[1,],1,length(E[1,])
summary(X.res1)
sd(X.res1)
X.res2 = X-matrix(S[,1:2],n,2)%*%matrix(solve(E)[1:2,],2,length(E[1,])
summary(X.res2)
sd(X.res2)
```

1.2 Individus extrêmes

Certains individus peuvent être considérés comme extrêmes (ou outliers).

1. Proposer une méthode utilisant la projection sur le premier plan factoriel et permettant d'indentifier de façon automatique les individus extrêmes.
2. Refaire l'ACP en considérant ces individus comme des individus supplémentaires. Tracer la projection sur le premier plan factoriel des individus ayant servi à l'analyse et des individus supplémentaires. On utilisera une autre couleur pour les individus supplémentaires. Par exemple, dans la figure 1, les départements d'Ile de France sont considérés en individus supplémentaires et la figure est obtenue à l'aide des commandes

```
fd = which(dep$region=="IdF")
pr = princomp(dep.red[-fd,])
plot(pr$score,pch=20,xlim = c(-6,6))
pr.s = dep.red[fd,]%*%pr$loadings
points(pr.s,pch=20,col="red")
text(pr.s[,1],pr.s[,2],labels=dep[fd,1],pos=2,col="red")
```

1.3 Mise en évidence d'information discrète

Pour aller un peu plus loin, on peut matérialiser une variable discrète sur les graphiques de l'ACP. Par exemple, ici nous allons mettre en évidence les régions.

```
library(ade4)
pr = princomp(dep.red)
par(mfrow=c(1,2)
s.chull(pr$scores,fac=dep$region)
points(pr$score,pch=20)
 text(pr$score[,1],pr$score[,2],labels=dep[,1],pos=2,col="red")
```

On observe que l'axe un oppose les régions riches aux régions plus pauvres et que l'axe deux oppose les régions très rurales aux (anciens) bassins industriels.

2 Seconde partie

Considérons 104 étudiants ayant passé 9 matières pour leur examen. La table de données `deug.Rdata` (à charger à l'aide de la fonction `load`) est une liste à trois composantes : `tab`, `result` et `cent`. La composante `deug$tab` contient des données brutes :

1. Algèbre et Analyse des données (sur 100)
2. Analyse (sur 60)
3. Probabilités (sur 80)
4. Informatique (sur 60)
5. Dominante (Sociologie ou Économie sur 120)
6. Options (sur 40)
7. Ouvertures (sur 40)
8. Anglais (sur 40)

9. Education physique et sportive (bonification <=15)

La composante `deug$result`

D- Éliminé après les épreuves écrites
C- Éliminé après l'oral de rattrapage
B- Admis sans mention après l'oral de rattrapage
B Admis avec la mention Passable
A Admis avec la mention Assez Bien.
A+ Admis avec la mention B

C'est une variable qualitative non ordonnée. On décide de la tranformer en une variable qualitative ordonnée pour mettre un peu d'ordre.

```
deug$result # avant d'ordonner
deug$result <- factor(deug$result,
levels = c("D", "C-", "B-", "B","A", "A+"), order = TRUE)
```

La composante `deug$cent` donne donne le nombre de points qu'il faut pour avoir la moyenne dans chaque matière.

2.1 Analyse descriptive univariée

Quand les variables sont peu nombreuses comme ici, on peut se permettre de commencer l'étude des données par une analyse descriptive univariée.

1. Représenter la distribution des notes de chaque matière sous la forme d'un boxplot. On tracera toutes les variables sur la même figure. Commenter.
2. Représenter la distribution de la variable résultat, par exemple en utilisant un diagramme en batons. Donner à votre graphique le titre *Résultat final à l'examen*, l'axe des abscisses devra porter la légende *Résultat* et l'axe des ordonnés *Nombre d'étudiants*.
3. Déduire de la composante `deug$cent` le coefficient de chacune des disciplines. Tracer un diagramme en baton horizontal pour représenter le poids de chaque matière dans la note finale. Vous pourrez ordinner les matières. Le sport est traité à part car il joue un rôle particulier.
4. Utiliser les estimateurs à noyau pour représenter la distribution des notes de chaque matière. On pourra matérialiser la note moyenne par une barre verticale rouge (voir par exemple la figure 1). On mettra tous les graphiques sur la même figure grâce à l'option `par(mfrow=)`.

2.2 Analyse en composantes principales normée

1. Réaliser une analyse en composantes principales normée. Tracer les graphiques utiles à l'interprétation. Interpréter et analyser les résultats.
2. Expliquer le graphique obtenu à l'aide des commandes suivantes. Ce graphique s'appelle *graphe canonique*.

```
pr = princomp(scale(deug$tab))
par(mfrow = c(3, 3), mar = c(2, 4, 2, 0) + 0.1)
for (i in 1:9) {
x <-pr$score[, 1]
```

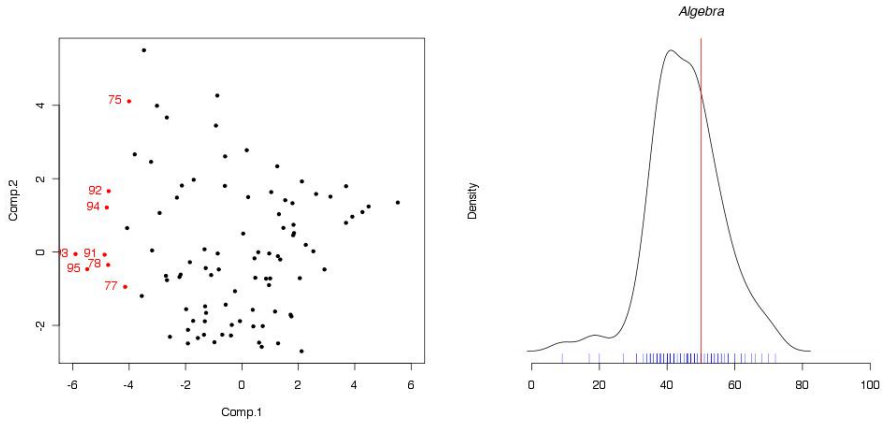

FIGURE 1 –

```
y <- deug$tab[, i]
titre <- paste("r =", round(cor(x, y), digits = 3), sep = "")
plot(x, y, las = 1, ylab = colnames(deug$tab)[i], main = titre,pch=20)
abline(lm(y ~ x))
}
```

3. Analyser les résidus associés au premier plan factoriel.

2.3 Analyse en composantes principales centrée sur les notes moyennes

En général, les étudiants ne s'occupent que de leurs affaires et du nombre de points qui les séparent de la barre fixée a priori, en général 10/20. La composante `deug$cent` donne, pour chaque matière, le nombre de points correspondant à la moyenne de 10/20, toutes diffférences se comptant en points d'avance ou en points de retard.

```
par(mfrow = c(1, 2), mar = c(5, 5, 4, 2) + 0.1)
boxplot(deug$tab, horizontal = T, las = 1,
main = "Avant centrage et réduction",
xlab = "Note", col = "lightblue")
boxplot(as.data.frame(scale(deug$tab, center = deug$cent, scale = FALSE)),
horizontal = T, las = 1, main = "Après centrage sur 10/20",
xlab = "Points d'avance ou de retard", col = "lightblue")
abline(v = 0, col = "red")
```

1. Faire l'ACP non normée des données centrées sur la note correspondant à 10/20. Interpréter et analyser les résultats.
2. Utiliser la fonction `s.chull` pour matérialiser le résultat final des étudiants. Commenter et analyser le graphique obtenu.

2.4 Conclusion

Conclure votre travail.