TD3 : Réduction de dimension par ACP et PLS

M2 MMA - Université de Paris

2020-2021

Exercice 1 : Interprétation des graphiques de l'ACP

Importation des données

Pour cet exercice, nous travaillerons sur les données `body_full.csv` qui contiennent des informations concernant le corps de $n = 507$ personnes.

```
body <- read.table("body_full.csv",header=TRUE, sep=";",dec=",")
dim(body)
```

```
## [1] 507  25
```

Installer le package `FactoMineR`, charger la librairie correspondante puis passer aux questions.

Questions

1. Tracer les boxplots permettant d'indiquer la distribution des variables du jeu de données. Indiquer si les données doivent être normalisées.
2. Utiliser la fonction `PCA()` du package `FactoMineR` pour effectuer une ACP sur les données. *Remarque* : il ne faut pas oublier de retirer la variable qualitative sous peine de ne pas réussir à lancer la fonction `PCA()`.
3. Interpréter le premier graphique (groupe d'individus). Pour une meilleure visualisation, on pourra représenter les individus de façon différenciée suivant leur genre.
4. Interpréter le second graphique (groupe de variables) pour donner un sens aux deux composantes principales.

Exercice 2 : Problème de normalisation

Chargement des données

Pour cet exercice, nous travaillerons sur les données `athle_records.csv`, qui portent sur les performances de 26 pays à 9 épreuves d'athlétisme.

```
athlete <- read.csv("athle_records.csv",sep = "\t",header=TRUE,dec=",")
rownames(athlete) <- athlete$X
athlete <- athlete[,-1]
rownames(athlete)[3] <- "Bresil"
rownames(athlete)[14] <- "Jamaique"
rownames(athlete)[18] <- "NouvelleZelande"
```

```r
rownames(athlete)[23] <- "Suede"
dim(athlete)
```

```
## [1] 26  9
```

Questions

1. Tracer les boxplots permettant d'indiquer la distribution des variables du jeu de données.
2. Utiliser la fonction `prcomp()` pour effectuer une ACP sur le jeu de données. Identifier les contributions des différentes variables pour chacune des composantes principales.
3. Utiliser la fonction `biplot()` pour représenter le diagramme des individus et variables de l'ACP. Confirmer les résultats observés à la question précédente.
4. Tracer la part de variance expliquée par chacune des composantes à l'aide de la fonction `plot()`.
5. Le phénomène observé aux questions précédentes est dû au fait que les données ne sont pas normées. Reprendre ces questions après avoir log-transformé les données.
6. Tracer les composantes principales 2 et 3 pour étudier les différences entre les groupes.
7. Utiliser le package `FactoMineR` pour effectuer l'ACP sur le jeu de données. Que peut-on constater?

Exercice 3 : Application à un jeu de données biologiques

Importation des données

Pour cet exercice, on reprend le jeu de données `Prostate` déjà utilisé sur lors des TPs précédents.

```r
load("Prostate.Rdata")
```

Questions

1. Tracer les boxplots donnant la distribution des variables du jeu de données puis normaliser les données à l'aide de la fonction `scale()`.
2. Utiliser le package `FactoMineR` pour effectuer une ACP sur le jeu de données.
3. Déterminer le nombre de composantes principales à retenir à l'aide de l'affichage des valeurs propres et leurs cumulées.
4. Reprendre les graphiques de l'ACP pour identifier des caractéristiques caractérisant ces composantes.
5. Utiliser le package `factoextra` et la fonction `fviz_contrib()` pour afficher la contribution des individus puis des variables sur chaque composante.

Exercice 4 : ACP-PLS sur un jeu de données génomique

Chargement des données

Pour cet exercice, on considère le jeu de données `Colon` du package `plsgenomics`. Télécharger le package, charger la librairie et le jeu de données correspondant.

```r
install.packages("plsgenomics")
library(plsgenomics)
data(Colon)
```

Ce jeu de données contient des données d'expression de $p = 2000$ gènes (matrice `X`) pour un ensemble de $n = 62$ tissus, parmi lesquels 40 sont tumoraux (`2` dans la variable `Y`) et 22 sont normaux (`1` dans la variable `Y`). Le but de ce TD est de construire des règles pour déterminer à l'aide de l'expression génomique le type de tissu de provenance.

Questions

1. Créer un jeu d'apprentissage et un jeu test en gardant chaque échantillon dans le jeu d'apprentissage avec probabilité $\frac{2}{3}$. Ecrire ensuite un modèle logistique pour expliquer le tissu d'origine en fonction des expressions de tous les gènes.
2. Procéder à une ACP en utilisant la fonction `pcr()` du package `pls`. Combien faut-il utiliser de composantes pour expliquer 75% de la variance du nuage?
3. Récupérer la matrice de réduction de dimension à partir de l'élément `loadings` du résultat de la fonction `pcr()` puis récupérer le jeu de données de dimension réduite.
4. A l'aide de la fonction `glm()`, utiliser une régression logistique pour apprendre une classification dans ce nouveau jeu de données réduit. Prédire alors le type de tissue d'origine et commenter.
5. Reprendre les questions précédentes en utilisant une PLS cette fois plutôt qu'une ACP. Indiquer l'avantage de cette méthode.