TD Analyse de données multivariées : l'ACP avec `ade4`

Magali Berland
magali.berland@uvsq.fr
Option "Introduction à la Modélisation en Biologie" M2 IHPS

27 novembre 2017

Table des matières

1 Introduction à l'ACP avec `ade4`

L'analyse en Composantes Principales (ACP) est un grand classique de l'analyse des données pour l'étude exploratoire ou la compression d'un grand tableau de données quantitatives. L'objectif de l'ACP est de revenir à un espace de dimension réduite (par exemple 2) en déformant le moins possible la réalité. Il s'agit donc d'obtenir le résumé le plus pertinent possible des données initiales pour pouvoir étudier les liens entre les variables ou bien pour constituer des groupes d'unité statistiques.

1.1 Les données

Nous allons utiliser un jeu de données très simple extrait de `survey` du package `MASS`.

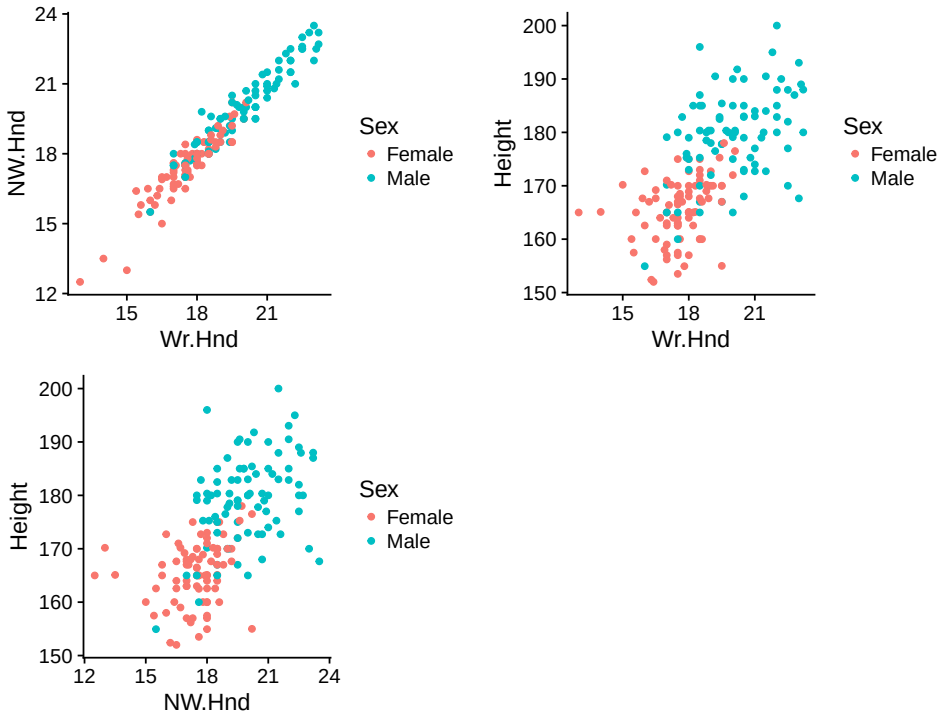
```
library(MASS)
data(survey)
names(survey)
```

```
 [1] "Sex"    "Wr.Hnd" "NW.Hnd" "W.Hnd"  "Fold"   "Pulse"  "Clap"   "Exer"   "Smoke"
[10] "Height" "M.I"    "Age"
```

Pour ne pas nous embêter par la suite avec les données manquantes, nous ne conservons que les individus entièrement documentés :

1. Créez une variable `survey.cc` sans données manquantes en utilisant la fonction `complete.cases()`

2. Que représentent les variables `Wr.Hnd`, `NW.Hnd` et `Height` ? Quelle est l'unité de mesure utilisée ?



3. Reproduisez les graphiques affichés ci-dessus et commentez-les.

1.2 Centrage et réduction

L'ACP est une méthode qui permet de résumer un ensemble de variables numériques en déterminant une variable résumée dite *composante principale* la plus reliée possible aux variables originelles. On va étudier le lien entre les variables `Wr.Hnd`, `NW.Hnd`, `Height`, `Age` et `Pulse`.

4. Représentez ces variables sous forme de boxplot. Quel problème voyez-vous ?
5. L'opération de centrage consiste à enlever la moyenne à chaque variable. Utilisez la fonction scale(), puis faites un nouveau boxplot des variables centrées. Commentez le graphe obtenu.
6. L'opération de centrage et réduction consiste à centrer les données puis diviser les valeurs par l'écart-type. Utilisez la fonction scale(), puis faites un nouveau boxplot des variables centrées-réduites. Commentez le graphe obtenu.

   Quand on fait une ACP normée, on travaille avec les données après centrage et réduction. Il est donc important de bien comprendre à quoi correspondent ces opérations.

1.3 ACP centrée réduite dans `ade4`

7. Utilisez la fonction `dudi.pca()` du package `ade4` pour exécuter une ACP centrée réduite

   Répondre 3 à la question `"Select the number of axes:"`. L'objet renvoyé par la fonction `dudi.pca()` est très riche. Nous allons examiner tous ses composants rapidement.
   - Le data frame `tab` contient les données du tableau initial après centrage et réduction.
   - Le vecteur `cw` donne le poids des colonnes (*column weight*), c'est-à-dire le poids des variables. Par défaut, chaque variable a un poids de 1.
   - Le vecteur `lw` donne le poids des lignes (*line weight*), c'est-à-dire le poids des individus. Par défaut, chaque individu a un poids de 1 .
   - Le vecteur `eig` donne les valeurs propres (*eigen values*). Elles nous renseignent sur l'information (appelée aussi *inertie totale*, *variance* ou *dispertion*) prise en compte par chaque axe :

```
(pve <- 100 * acp$eig/sum(acp$eig))
[1] 50.3828830 22.2681102 18.0332249  8.6358051  0.6799768
cumsum(pve)
[1]  50.38288  72.65099  90.68422  99.32002 100.00000
```

Dans notre exemple, le premier axe factoriel extrait 50.4 % de l'information, le deuxième axe factoriel 22.3 % de l'information. Le premier plan factoriel représente donc 72.7 % de l'inertie initiale. Ceci signifie que lorsque nous projetons le nuage de points initial sur le plan défini par les deux premiers axes factoriels, nous avons perdu relativement peu d'information.
L'analyse est pertinente si, avec un petit nombre d'axe, on explique une part importante de l'inertie.

1.4 Représentations graphiques dans `ade4`

1.4.1 Représantation des individus

La fonction `s.label()` permet de représenter les individus sur les différents plans factoriels.

8. Faites les représentations dans les plans (1,2), (1,3) et (2,3).
   Les graphiques des individus sont interprétés, en tenant compte des qualités de représentation, en termes de regroupement ou dispersions par rapport aux axes factoriels et projections des variables initiales.
   La fonction `s.class()` permet de porter en information supplémentaire une variable qualitative définissant des groupes d'individus, par exemple :

```
s.class(dfxy = acp$li, fac = sexe, col = c("red", "blue"), xax = 1, yax = 2)
```

9. Faites les représentations dans les trois plans factoriels.

1.4.2 Représantation des variables

10. Utilisez la fonction `s.corcircle()` pour représenter les variables initiales dans le nouvel espace. Cette représentation est appelée cercle des corrélations
    Le cercle des corrélations permet de voir, parmi les anciennes variables, les groupes de variables très corrélées entre elles. Par projection sur un plan factoriel les points-variables s'inscrivent dans un cercle de rayon 1 - le cercle des corrélations - et sont d'autant plus proche du bord du cercle que le point-variable est bien représenté par le plan factoriel, c'est-à-dire que la variable est bien corrélée avec les deux facteurs constituant ce plan.
    Attention ! Les variables qui ne sont pas situées au bord du cercle dans un plan factoriel ne sont pas corrélées avec les deux facteurs représentées. Elles ne servent pas à l'interprétation et l'effet de perspective empêche d'interpréter la proximité de deux variables (voir d'autres plans factoriels, où la corrélation sera plus forte).
    L'angle entre 2 point-variables, mesuré par son cosinus est égal au coefficient de corrélation linéaire entre les 2 variables. Ainsi :
    — si les points sont très proches (angle peu différent de 0), X1 et X2 sont très fortement corrélés positivement
    — si l'angle est égal à 90° alors pas de corrélation linéaire entre X1 et X2
    — si les points sont opposés, l'angle vaut 180°, X1 et X2 sont très fortement corrélés négativement

11. Commentez et interprétez le cercle des corrélations que vous avez tracé.

1.4.3 Représantation simultanée des individus et des variables

La fonction `scatter()` permet de représenter simultanément les individus et les variables. C'est une fonction générique associée à un objet de la classe dudi.

```
scatter(acp, posieig="bottomleft")
```

Enrichir le graphique en portant l'information sur les groupes et certains individus particuliers :

```
scatter(acp, clab.row = 0, posieig = "none")
s.class(acp$li, sexe, col = c("red", "blue"), add.plot = TRUE, cstar = 0,
        clabel = 0, cellipse = 0)
s.label(acp$li[106,], clab=1, add.p=T)
```

1.5 Contributions à l'inertie

Les composantes principales sont construites sur la base des variables originelles (et des unités statistiques). On peut estimer l'importance de chaque variable comme un pourcentage. Ces pourcentages sont appelés contributions à l'inertie. Les contributions permettent d'identifier les individus très influents pouvant déterminer à eux seuls l'orientation de certains axes ; ces points sont à vérifier et à caractériser.

```
inertia.dudi(acp)
inertia.dudi(acp, col.inertia = TRUE)
inertia.dudi(acp, row.inertia = TRUE)
```

12. Comment interprétez vous le résultat de ces commandes ?

2 Exercice en autonomie

Le jeu de données "deug" du package ade4 contient les résultats aux examens de 104 étudiants de 2ème année d'université dans 9 disciplines dans sa composante `deug$tab`. Réalisez une ACP de ces données et interprétez le résultat.

1. Quelle est l'échelle des notes à l'examen final ?
2. Combien d'élèves ont réussi leur année ?
3. Quels sont les coefficients de chaque matière ? Représentez le poids des diciplines sous forme d'un graphique.
4. À votre avis, dans quel type de filière universitaire étaient inscrits ces étudiants ?
5. Analysez la distribution des notes dans chaque discipline. Quelles sont les matières difficiles ? Si vous n'aviez qu'une discipline à réviser, laquelle choisiriez-vous ?
6. Que remarquez-vous à propos de la distribution des notes de sport ?
7. Réalisez une ACP centrée réduite, en centrant par rapport aux coefficients.
8. À l'aide de différentes représentations graphiques, proposez une interprétation pour les deux premières composantes principales (= 1er et 2ème axe).
9. Quelles différences voyez-vous avec une ACP centrée réduite qui ne tient pas compte des coefficients des matières ?