# Introduction à l'analyse multivariée (factorielle) sous R

Stéphane CHAMPELY

7 septembre 2005

# Table des matières

# Chapitre 1

# Introduction

## 1.1 Les données multivariées

Les données sont des mesures effectuées sur des unités statistiques (individus, animaux, objets, organisations sportives). En première approche, la nature de ces mesures est soit qualitative (les unités appartiennent à des catégories) soit quantitative (nombres). On n'utilise pas le même type de technique statistique suivant la nature des données.

Lorsqu'une seule mesure est prise sur chaque unité on parle de contexte univarié, s'il y en a deux on parle de contexte bivarié, au-delà, il s'agit d'un contexte multivarié.

Nous verrons tout d'abord des méthodes dévolues à des données multivariées uniquement quantitatives, puis à des données uniquement qualitatives, nous aborderons rapidement le problème mixte et nous finirons par un problème multivarié spécifique, celui de l'étude d'un tableau croisé.

## 1.2 L'approche factorielle des données multivariées

Plusieurs méthodes statistiques multivariées existent mais nous nous concentrerons dans cette introduction sur l'approche dite factorielle. En deux mots, elle consiste à résumer les différentes mesures par un nombre faible de variables de synthèse (idéalement une ou deux) qui retiennent l'essentiel de l'information.

On peut alors étudier ces variables de synthèse pour elles-mêmes, c'est-à-dire les employer par la suite dans des régressions par exemple ou le plus souvent les utiliser pour représenter graphiquement (1) les relations entre les variables originelles et (2) les proximités entre les unités statistiques. Cette possibilité d'avoir une représentation graphique de la structure d'un grand jeu de données autorisant une interprétation relativement intuitive a fait le succès de la méthode.

## 1.3 Le logiciel R et le package ade4

Impossible d'imaginer réaliser des analyses multivariées factorielles à la main ou même à la calculatrice ou avec un tableur. La plupart des logiciels statistiques professionnels possèdent de telles options mais (1) ils coûtent chers, (2) ils sont parfois peu flexibles, en particulier en ce qui concerne les représentations graphiques et (3) il existe des approches assez différentes de l'analyse multivariée factorielle et en particulier il y a une nette différente entre l'approche française et l'approche anglo-saxonne (qui elle-même se subdivise en de nombreuses tendances).

Le choix du logiciel R, malgré sa complexité de prise en main, permet un grande richesse technique et le package ade4 est sans doute ce qui se fait de mieux en analyse multivariée factorielle sur le marché, tout en étant gratuit ! Ce package ade4 a été réalisé par D. Chessel et son équipe du laboratoire de Biométrie de l'université Lyon1. Ce package permet de réaliser de nombreuses analyses, dont nous ne verrons qu'une petite partie, il est en particulier l'un des rares logiciels à permettre de réaliser des analyses multi-tableaux et des analyses de données multivariées spatialisées. Les outils de représentation graphiques sont sans équivalent et la théorie mathématique sous-jacente a été profondemment unifiée par les auteurs.

Il convient donc, pour suivre cette introduction à l'analyse multivariée en apprenant simultanément à réaliser concrètement les propositions, d'installer le logiciel R et le package ade4. On trouvera dans l'annexe A un guide rapide d'installation de R (pour windows, désolé les filles) et de ade4.

# Chapitre 2

# Analyse en Composantes Principales

## 2.1 Les données multivariées quantitatives

Le jeu de données monde84 est fourni avec le package ade4 sous la forme d'une structure de données (une "class") qu'on appelle un "dataframe", qui correspond à la notion de tableau. Pour le rendre accessible [1], l'afficher et obtenir des informations :

```
data(monde84)
monde84
?monde84
```

Ce dataframe comprend donc 5 mesures démographiques (PIB, croissance de la population...) en colonnes concernant 48 pays en lignes. Ces mesures étant toutes numériques, nous sommes dans un contexte dit *multivarié quantitatif*.

On peut lancer une première vague d'analyses statistiques, indispensable mais parfois malheureusement négligée, en s'intéressant dans un premier temps séparément à chaque variable :

```
attach(monde84)
summary(monde84)
hist(pib)
hist(log(pib))
hist(croipop)
hist(morta)
hist(log(morta))
hist(anal)
hist(log(anal+1))
```

[1] Attention, il faut préalablement avoir chargé le package ade4

```
hist(scol)
```

La première conclusion serait donc d'effectuer quelques transformations afin de rendre les distributions sinon normales du moins plus symétriques. D'où la création d'un second data.frame monde2.

```
monde2<-data.frame(log(pib),croipop,log(morta),log(anal+1),scol)
dimnames(monde2)<-list(dimnames(monde84)[[1]], c("lpib","croipop","lmorta","lanal","scol"))
```

Dans un second temps nous regardons les relations deux à deux entre ces nouvelles variables mais aussi entre les variables non transformées.

```
cor(monde84)
cor(monde2)
plot(monde84)
plot(monde2)
```

On voit que les transformations on permis d'obtenir des relations plus linéaires entre les variables (figure 2.1). La statistique multivariée se prête en effet mieux à l'analyse de relations linéaires et de variables de distribution symétrique. On va donc privilégier le dataframe monde2.

Mais comment étudier globalement les relations entre cinq variables et non plus deux à deux ? Comment connaître les ressemblances et dissemblances entre les pays ? Ce qui les fonde ? L'analyse multivariée dite *Analyse en Composante Principales (ACP)* permet de répondre à de telles questions !

## 2.2 Une première théorie de l'ACP : trouver une variable de synthèse

La première façon d'expliquer l'ACP est de la présenter comme une méthode qui permet de construire une variable ressemblant le plus possible à toutes les variables du tableau étudié. On appellera cette variable de synthèse une composante principale.

**Définition 1 (composante principale)** *La composante principale est une nouvelle variable qui a pour propriété d'être de corrélation maximum avec l'ensemble des variables du tableau étudié. Plus précisément, la somme des carrés de corrélations* [2] *de cette variable avec les variables originelles est maximisée. Ce maximum est le pouvoir de synthèse de cette variable appelé* valeur propre.

Suivant les logiciels, cette composante principale, qui a toujours une moyenne nulle, peut être de variance 1 ou de variance égale à la valeur propre. Notons que cela ne change rien à la propriété de maximisation.

[2] Le carré du coefficient de corrélation est appelé *coefficient de détermination*

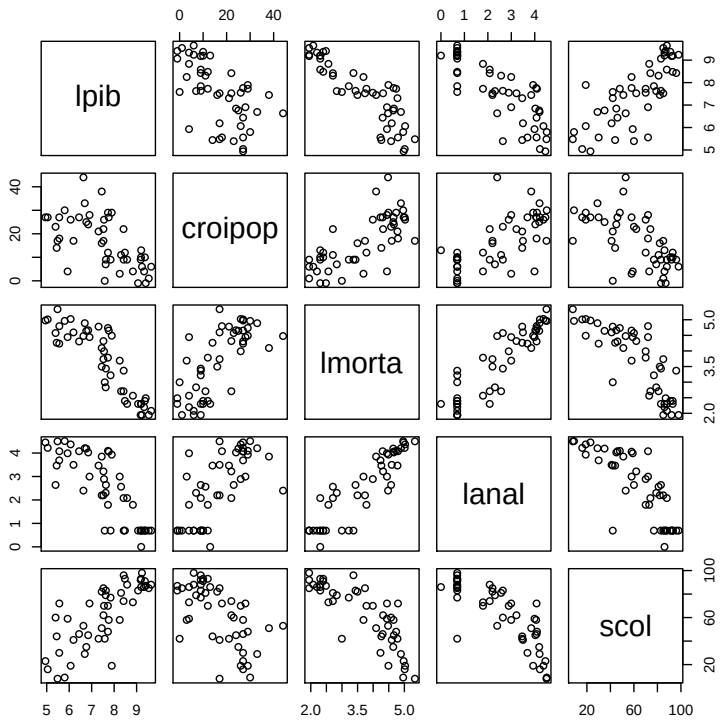


FIG. 2.1 – Nuages de points entre les variables du dataframe monde2

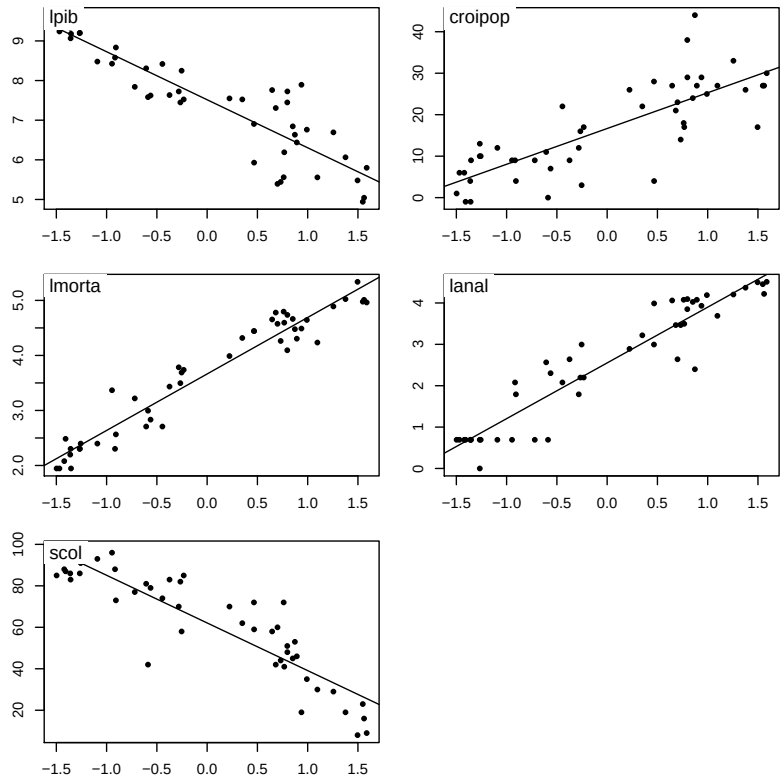


FIG. 2.2 – Nuages de points entre les variables du dataframe monde2 et la composante principale issues de son ACP

On voit dans la figure 2.2 une représentation graphique de cette propriété : chaque variable originelle est couplée dans un nuage de points à la composante principale. Toutes les variables sont très bien reliées linéairement, parfois positivement (croipop, lmorta, lanal) parfois négativement (lpib, scol) avec la première composante principale. On peut interpréter cette variable de synthèse comme une mesure de (non-)développement des pays.

## 2.3 Réaliser l'ACP avec ade4

La programmation du package ade4 est basée sur une théorie [1] dite du *schéma de dualité* (en anglais duality diagram, abrégé dans ade4 en *dudi*). La fonction qui permet de réaliser une ACP dans R s'appelle dudi.pca. Comme toutes les fonctions R, dudi.pca crée un objet, que l'on peut afficher, représenter graphiquement ou ré-utiliser dans d'autres analyses. On demande lors de l'exécution de la fonction dudi.pca de chosir le nombre d'axes à sélectionner, nous en

parlerons plus tard, mais il s'agit ici de simplement répondre : 1 et de valider (taper la touche Entrée).

```
pca.monde2<-dudi.pca(monde2,center=T,scale=T)
score.pca(pca.monde2)
```

Tout objet de "class" dudi a pour valeur les composants suivants :
- $eig qui donne une indication sur le pouvoir de synthèse de l'analyse de ce dataframe,
- $c1 et $co [3] qui donnent des informations sur la structure des colonnes du dataframe et
- $l1 et $li [4] qui donnent des informations sur la structure des lignes du dataframe.

Afin d'obtenir la composante principale, de variance 1 (dans $l1) ou de variance égale à la valeur propre (dans $li) et la représenter graphiquement, on emploie :

```
pca.monde2$l1
pca.monde2$li
dotchart(pca.monde2$li[,1],labels=dimnames(pca.monde2$li)[[1]])
```

et pour lire les corrélations de cette variable de synthèse avec les autres variables et la valeur propre on utilise

```
pca.monde2$co
pca.monde2$eig
```

La variable du tableau la plus reliée à la composante principale est lmorta ($r = 0.96$) et la moins reliée croipop (avec quand même $r = 0.79$). La somme des carrés de ces corrélations est de 4.07377747, sur un total au maximum de 5, on dit que la première composante principale explique 4.07/5=81% de l'information [5] du tableau monde2, ce qui est remarquable.

## 2.4 Solution du second ordre de l'ACP

L'objet de "class" "list" olympic fourni par le package ade4 comprend dans sa composante olympic$tab les performances de 33 décathloniens lors des (10) épreuves des jeux olympiques de 1988. Commençons par un rapide coup d'oeil sur les corrélations entre ces mesures :

```
data(olympic)
```

[3] ces deux vecteurs sont proportionnels, suivant l'image que l'on souhaite avoir de la structure des colonnes, l'un convient mieux que l'autre (voir plus loin)

[4] Mêmes remarques que ci-dessus, ces deux vecteurs sont aussi proportionnels

[5] On parle aussi d'inertie

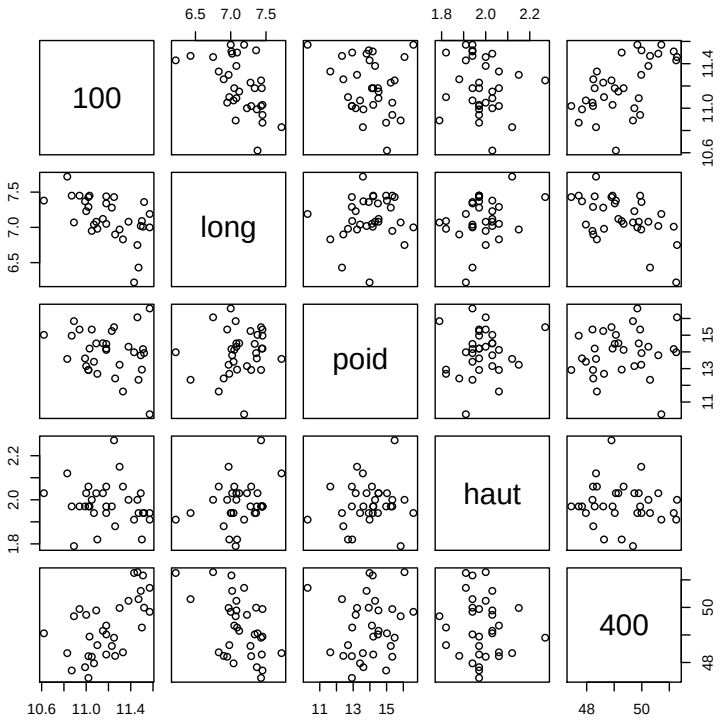


FIG. 2.3 – Extraits des Nuages de points entre les variables du dataframe olympic$tab

```
cor(olympic$tab)
plot(olympic$tab)
```

Notons dans la figure 2.3 que les performances aux 100m et 110m haies sont liées positivement, que le lancer du poids est également relié positivement au lancer du disque et que le saut en longueur est relié négativement au résultat du 100m [6].

Réalisons l'ACP de ce dataframe et déterminons la composante principale. Son pouvoir de synthèse est de 3.4182381 sur un maximum possible de 10 soit 34%. On est loin d'avoir épuisé toute l'information disponible!

[6] Interprétons cette corrélation a priori surprenante : une forte valeur de la variable longueur correspond généralement à une faible valeur du 100m. Donc si on saute loin, on met généralement peu de temps au 100m, les deux performances sont en définitive liées positivement bien que les variables le soient négativement. Si la performance au 100m était mesurée en termes de vitesse plutôt que de temps cela serait plus simple ...

```
pca.olympic<-dudi.pca(olympic$tab,center=T,scale=T)
pca.olympic$eig
```

On va donc rechercher une autre variable de synthèse nous permettant d'affiner notre compréhension de l'épreuve du décathlon. La nouvelle variable de synthèse doit nous offrir un résumé radicalement neuf, c'est pourquoi on décide d'employer une contrainte pour la définir.

**Définition 2** *La nouvelle variable de synthèse doit être non corrélée à la première - ce qui garantit une information nouvelle - mais doit aussi viser le même objectif, c'est-à-dire être la plus liée possible aux variables originelles. On parle de deuxième composante principale.*

Bien sûr, elle ne peut être aussi performante en termes de pouvoir de synthèse que la première dont la recherche se faisait sans contrainte, ce qui explique que la deuxième valeur propre sera toujours plus faible. Le processus peut être itéré, on peut rechercher une troisième variable de synthèse, non corrélée aux deux premières, optimisant le même critère.

Deux obstacles s'opposent à aller trop loin en ce sens :

1. notre objectif étant de résumer le tableau, on ne multipliera pas les variables de synthèse, car remplacer les variables originelles par un grand nombre de variables de synthèse (dont l'interprétation est moins immédiate) est contreproductif et
2. il est simple de proposer une représentation graphique de la structure sur la base de deux variables de synthèse, avec trois cela devient ardu, au delà, personnellement, je n'y arrive pas.

La représentation graphique à deux dimensions de la structure des lignes se fait en réalisant un nuage de points croisant les deux variables de synthèse (fichiers $li), ce qui n'est possible que si on a sélectionné deux dimensions lors de la réalisations de l'ACP [7] !!! On parle de *plan factoriel* des unités statistiques.

```
pca.olympic<-dudi.pca(olympic$tab,center=T,scale=T)
s.label(pca.olympic$li)
```

On voit sur la figure 2.4 les 33 unités statistiques selon les deux composantes principales dont le pouvoir de synthèse est de $(3.4182381+2.6063931)/10$ soit 60% de l'information. La position des individus sur cette image exprime donc une large part de leur proximité originale dans le dataframe olympic$tab. Les concurrents sont ici numérotés en fonction du résultat final ; on voit ainsi que les meilleurs concurrents sont situés du même coté (à droite) et les moins bons à gauche [8]. Une propriété de l'ACP est que l'individu qui est situé à l'origine du graphique est l'individu moyen, c'est-à-dire celui qui réaliserait le résultat moyen à chacune des 10 épreuves.

[7] Si ce n'est pas le cas, recommencer l'ACP

[8] Le numéro 1 est caché sous le numéro 11

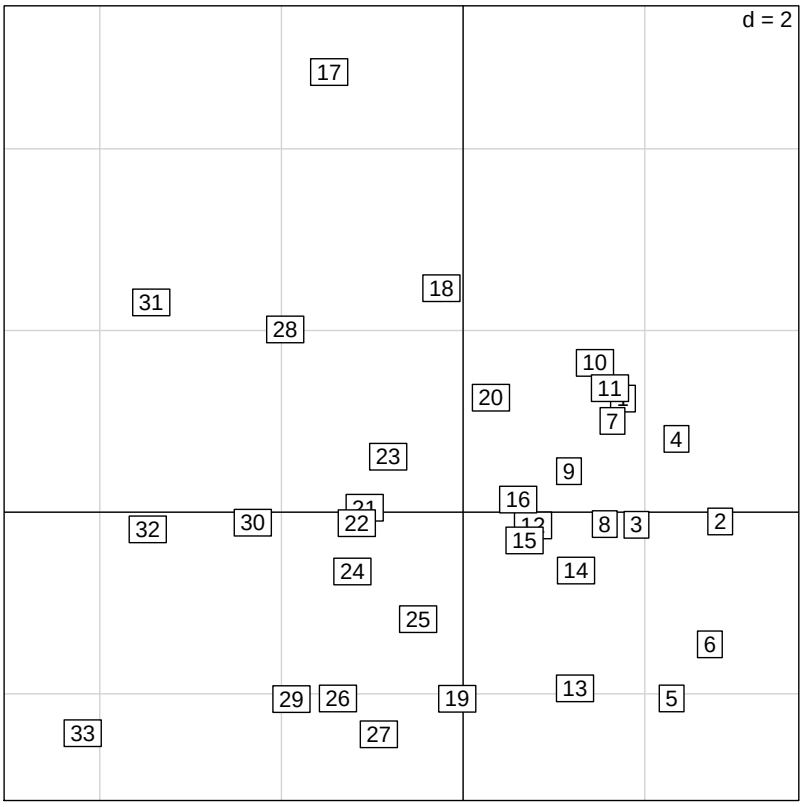


FIG. 2.4 – Plan factoriel : nuage de points des deux premières composantes principales issues de l'ACP du dataframe olympic$tab

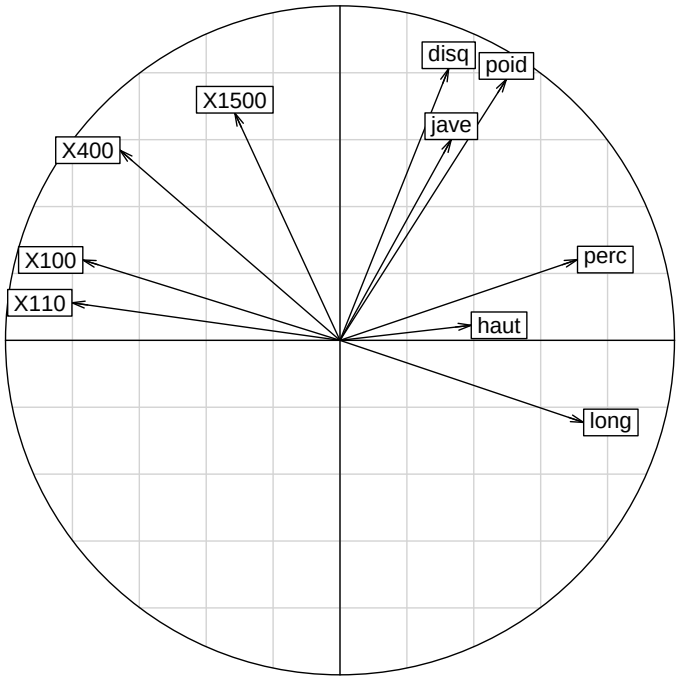


FIG. 2.5 – Cercle des corrélations de l'ACP du dataframe olympic$tab

Bien sûr, il faut aussi se souvenir que 40% de l'information n'est pas exprimée sur cette image et donc que la position de certains points peut se révéler trompeuse. On verra dans la section 2.5 comment se prémunir contre ce type d'ennuis.

En ce qui concerne les relations entre les variables originelles, elles s'obtiennent indirectement en regardant les relations entre ces variables et les deux variables de synthèse. Ainsi, dans le *cercle des corrélations* (cf figure 2.5), chaque variable originelle est située suivant sa corrélation sur l'axe des X avec la première composante principale et sur l'axe des Y avec la deuxième.

```
s.corcircle(pca.olympic$co)
```

Les relations entre les variables originelles sont interprétables en termes de directions, c'est pourquoi ces variables sont souvent représentées non comme des points mais en tant que flèches. On utilisera la grille de lecture suivante :

– si deux variables vont dans la même direction, elles sont corrélées positivement, par exemple ici le 100m et le 110m haies,

- si deux variables sont dans des directions opposées, elles sont corrélées négativbement, par exemple le 100m et le saut en longueur et
- si deux flèches sont perpendiculaires, les deux variables sont non corrélées, ici le lancer du poids et le 100m

Globalement on voit ici ressortir qu'il existe deux groupes de performances, celles qui sont liées à la vitesse de l'individu et celles liées à sa force. Á nouveau, il faut se souvenir que seulement 60% de l'information est présente sur cette image et donc qu'elle n'est pas un reflet exact de la situation. Il est bon de se reporter à la matrice des corrélations pour vérifier les structures repérées et à d'autres outils proposés dans la section 2.5.

**Remarque 1** *En ACP, la longueur des flèches sur le cercle des corrélations a aussi du sens, elle indique la qualité de représentation de la variable concernée sur l'image. Si la flèche est de longueur 1, la mesure est représentée à 100%. On concentrera donc l'interprétation sur les flèches les plus longues.*

**Remarque 2** *Si la solution d'ordre 3 est choisie, il est possible soit d'essayer de lire des représentations 3D (mais c'est toujours difficile) ou de réaliser les trois représentations 2D possibles (mais c'est toujours difficile). Au delà de trois, c'est toujours très difficile ...*

**Exercice 1** *Le fichier courses.csv comprend les résultats de 51 nations aux jeux olympiques uniquement en ce qui concerne les épreuves de courses à pieds (100m jusqu'au marathon). Le temps donné est le meilleur temps réussi par cette nation jusqu'aux JO de 1984.*

- *Réaliser l'ACP de ce jeu de données. Á quel pourcentage d'information correspond la première valeur propre ? La deuxième ?*
- *Représenter le cercle des corrélations. Les flèches vont dans la même direction, qu'est-ce que cela signifie (on parle d'*effet taille*) ?*
- *Représenter le nuage des points des unités statistiques. Expliquer la position de Singapour, des États-Unis et du Kenya.*

## 2.5 Les diagnostics

### 2.5.1 Choix de la dimension de représentation

Il faut choisir le nombre de variables de synthèse à retenir dans l'interprétation. Plusieurs méthodes existent dans la littérature. On peut noter dès à présent que s'il existait une méthode infaillible, je ne donnerai que celle-ci ...

L'information contenue dans le tableau de l'ACP est égale au nombre de variables [9]. Les valeurs propres expriment la qualité du résumé offert par les variables de synthèse en s'ajoutant. On peut donc considérer que si l'information dépasse un certain seuil, on s'arrêtera. Si le seuil est fixé à 69% [10], les deux

[9] On parle souvent d'*inertie* dans la littérature de l'analyse factorielle pour désigner la quantité d'information

[10] Un nombre renversant

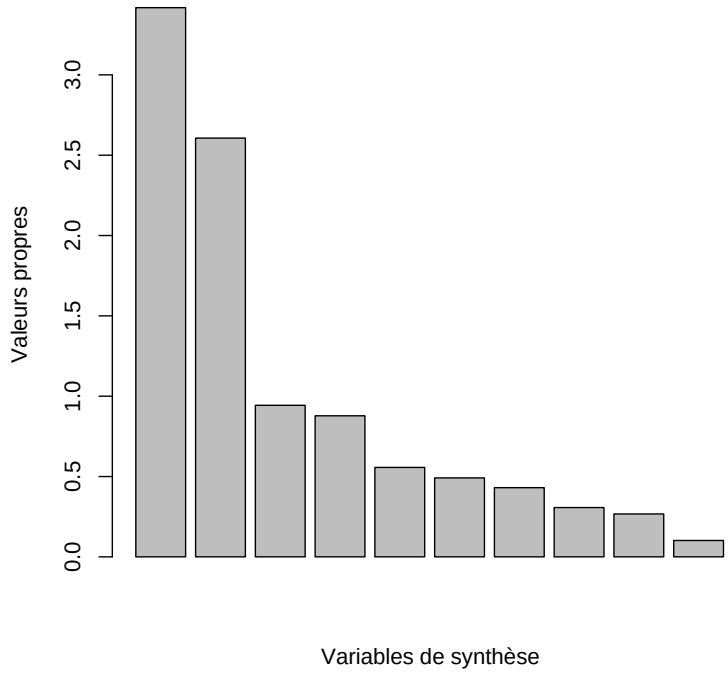


FIG. 2.6 – Éboulis des valeurs propres de l'ACP du dataframe olympic$tab

premières variables de synthèse dans l'exemple des marathoniens ne constituant que 60% ne sont pas suffisantes.

La *règle de Kaiser* repose sur le fait que dans l'ACP, toute variable originelle apporte 1 à l'information. Donc toute valeur propre supérieure à 1 peut être considérée comme supérieure aux variables originelles. On retiendrait deux valeurs propres dans l'exemple des décathloniens.

Le diagramme des valeurs propres successives (voir figure 2.6) appelé *éboulis des valeurs propres* permet de décider suivant sa forme de la dimension retenue. Parfois un coude apparaît très nettement (et parfois non . . .), ici nous conduisant à contenir deux variables de synthèse (attention à l'effet taille . . .).

```
barplot(pca.olympic$eig,xlab="Variables de synthèse",ylab="Valeurs propres")
```

Les statisticiens anglo-saxons ont développé dans un contexte probabiliste de normalité des tests du choix du nombre de valeurs propres (voir [2] p. 235) Faut-il y croire ?

Besse [3] propose de réaliser des boîtes à moustaches des variables de synthèse

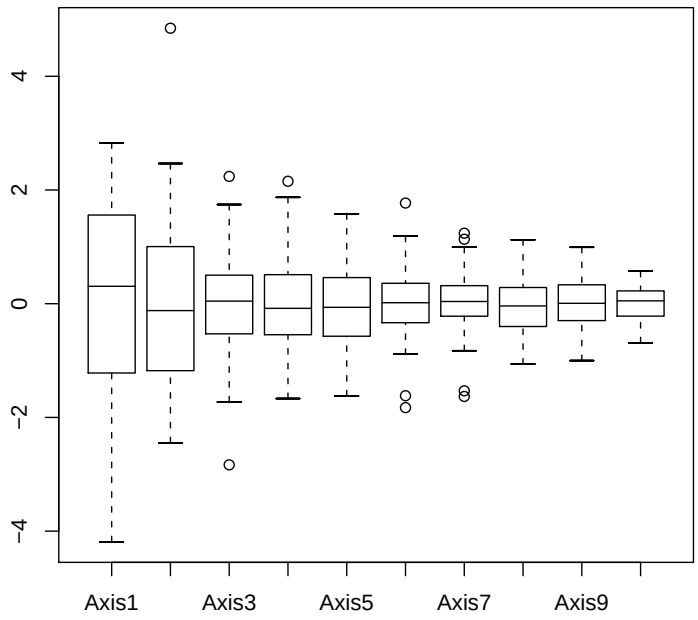


FIG. 2.7 – Boîtes à moustaches des composantes principales de l'ACP du data-frame olympic$tab

successives (voir figure 2.7). On voit clairement se dessiner les différences de variance et les variables instables.

```
boxplot(dudi.pca(olympic$tab,center=T,scale=T,nf=10,scannf=FALSE)$li)
```

### 2.5.2 La qualité de représentation

En ne retenant que les premières variables de synthèse, nous résumons le tableau mais perdons bien entendu une partie de l'information. Il faut souligner que celle-ci n'est pas perdue de façon homogène, certaines parties du tableau peuvent en être plus victimes ! Ainsi certaines unités statistiques seraient moins bien représentées que d'autres sur le plan factoriel ou certaines variables sur le cercle des corrélations. L'interprétation en ce qui les concerne ne peut par conséquent être basée uniquement sur ces images déformées (il faudra revenir au tableau de départ et à la matrice des corrélations).

Heureusement, il est possible de calculer pour chaque unité statistique (resp. chaque variable) le pourcentage d'information contenu dans les premières variables de sythèse par rapport à l'information disponible dans le tableau général.

**Définition 3** *Les* cosinus carrés - *dits aussi contributions relatives - sont la part d'information retenue par les k premières variables de synthèse en ce qui concerne les unités statistiques (resp. les variables).*

On les obtient grâce à la fonction inertia.dudi que l'on applique à l'objet obtenu suite à l'ACP, c'est-à-dire "en sortie" de la fonction dudi.pca, par exemple pca.olympic.

```
inertia.dudi(pca.olympic,row.inertia=TRUE,col.inertia=TRUE)
```

En ce qui concerne les unités statistiques - les lignes du tableau - la composante $row.rel [11] donne ces cosinus carrés pour chaque variable de synthèse. On peut l'obtenir de façon cumulée dans la composante $row.cum . Pour le cas des décathloniens où l'on a décidé de retenir deux axes, on voit que le premier décathlonien sur deux axes n'a que 28.9% de son information représentée, il n'est donc pas "à sa place", en revanche le deuxième qui a 90.1% d'information représentée est peu déformé par l'image offerte.

En ce qui concerne les variables, la composante $col.cum nous donne le pourcentage d'information représenté sur le cercle des corrélations dans la colonne 2. Si le lancer du disque est bien représenté avec 76.6%, le saut en hauteur l'est très mal à 15.6%. Une troisième variable de synthèse permettrait sans doute d'en savoir plus sur cette mesure.

**Remarque 3** *Observer comme la longueur des flèches sur le cercle des corrélations (Figure 2.5) est directement proportionnelle aux valeurs de la deuxième colonne de la composante $col.cum comme cela a déjà été évoqué.*

**Exercice 2** *Utiliser la fonction inertia.dudi sur le fichier de l'état du monde. Il y a un problème, cette fonction "plante" lorsqu'on a sélectionné un seul axe.*

- *Réaliser à nouveau l'ACP en conservant deux axes, mais l'analyse des représentations se fera sur un seul.*
- *Quels sont les pays qui sont les mieux représentés ?*
- *Quels sont les pays les moins bien représentés ?*
- *En ce qui concerne la Chine, quel est son cosinus carré ? Aller voir dans le tableau monde2 et expliquer pourquoi*
- *Les variables sont-elles toutes bien représentées ?*

## 2.6 Diverses remarques sur l'ACP

L'ACP produit des variables de synthèse non corrélées. Elles sont parfois employées dans des régressions multiples, dans des classifications automatiques,

[11] avec le signe correspondant à la variable de synthèse, la dernière colonne faisant référence à un autre type de diagnostic

dans des analyses de variance ou pour résumer un grand groupe de variables en quelques unes [12].

L'ACP est une méthode basée sur les corrélations et donc sensible aux points extrêmes. Les résultats peuvent être profondément influencés par leur présence. Des versions robustes existent (voir package amap de R).

D'aucuns utilisent l'ACP pour repérer des groupes d'individus. Il existe des méthodes spécifiques : classification automatique, analyse de mélanges, voire même des méthodes factorielles spécifiques (voir package amap de R).

L'ACP est une méthode basée sur des combinaisons linéaires des variables originelles, certaines propositions théoriques permettent de relâcher cette contrainte (citation Ferraty ou Durand).

[12] quoique dans ce cas une *analyse en facteurs*, voir [2] soit probablement plus indiquée

# Chapitre 3

# Une variante : l'ACP centrée

## 3.1 Les données multivariées homogènes

Le fichier piscine.csv correspond à un sondage effectué en 1999 sur la clientèle des piscines lyonnaises en vue de connaître ses souhaits sur des investissements futurs concernant ces établissements. Divers équipements/installations étant proposés, les sondés pouvaient noter ces propositions sur une échelle de Likert de 1 (pas du tout souhaité) à 5 (vraiment très souhaité).

Dans ce fichier, il y a 85 individus qui présentent des données manquantes, ils seront exclus de l'analyse. Le package ade4 ne fournit pas en effet de méthodes pour traiter ce type de problème.

```
piscine<-read.table("piscine.csv",header=TRUE,sep=";",dec=",")
print(piscine)
dim(piscine)
piscine2<-na.omit(piscine)
dim(piscine2)
```

Avec ces données, non seulement les corrélations entre les variables ont un sens, mais les variables sont directement comparables entre elles puisque de même nature. Ainsi les moyennes indiquent des niveaux de souhaits différents et les variances l'homogénéité des réponses des sondés (plus la variance est grande, plus leurs réponses diffèrent).

L'ACP vue jusqu'à présent élimine ces deux types d'information en se concentrant uniquement sur les corrélations. Nous allons à présent envisager une méthode d'analyse qui retient l'information concernant les variances c'est-à-dire qui donne plus d'importance aux variables les plus dispersées.

## 3.2 La géométrie de l'ACP centrée

Afin de développer cette nouvelle méthode ilest nécessaire de reconsidérer les choses sous un angle géométrique. Chaque ligne du tableau piscine peut être vu un point dans un espace de dimension 13 (certes difficilement visualisable). Il faut donc imaginer un nuage de 615 points dans cet espace. Existe-t-il une *direction d'allongement* de ce nuage qui indiquerait donc une *structure* ? Pour le savoir, on peut choisir un *axe* et projeter chaque point perpendiculairement sur cet axe. Si les points projetés sont éloignés, cet axe est direction d'allongement. On va identifier la notion d'allongement à la notion d'éloignement des points et plus présicément à la notion statistique de variance de ces projections. Il s'agit donc, parmi toutes les directions possibles, de trouver l'axe qui maximise cette variance c'est-à-dire l'allongement.

On démontre que la direction maximisant la variance passe par le point moyen et qu'elle peut s'obtenir par une méthode mathématique dite diagnolisation en valeurs propres [1].

Lorsque les variables sont mesurées dans des unités différentes comme c'était le cas dans le précédent chapitre, on commence par standardiser les mesures, c'est-à-dire qu'à chaque variable est retirée sa moyenne et qu'elle est divisée par son écart-type. La recherche de la direction d'allongement est alors strictement équivalente à l'ACP, c'est-à-dire que les points projetés sont les valeurs de la variable de synthèse, la variance maximum atteinte est la valeur propre et la direction d'allongement est proportionnelle aux corrélations. Mais lorque les unités sont comparables, l'étape initiale de standardisation est inutile et gomme en particulier l'information sur les différences de variances. On doit appliquer la méthode de recherche de la direction d'allongmenet sans tranformation du tableau initial, on parle alors d'*ACP centrée*.

**Définition 4** *L'ACP centrée ou* ACP sur matrice de covariances *d'un tableau à n lignes et p colonnes consiste à chercher dans un espace de dimension p une direction dit* axe principal *sur laquelle les n points correspondant aux lignes du tableau puissent être projetés, ces projections présentant une variance maximale.*

*Une seconde direction orthogonale à la première maximisant le même critère consitue la solution du second ordre de ce problème.*

## 3.3 Réaliser l'ACP centrée avec ade4

En modifiant l'argument "scale" de la fonction dudi.pca, on réalise cette analyse.

```
pca.piscine<-dudi.pca(piscine2,center=TRUE,scale=FALSE)
```

La représentation des unités statistiques en deux dimensions - le plan factoriel - s'obtient comme précédemment, l'origine représente à nouveau le point

[1] D'où le nom des pouvoirs de synthèse

moyen et les proximités sur la représentation graphique s'interprètent en termes de proximités dans le tableau original. On ne peut ici tirer grand chose du graphique (non présenté) car dans le cadre d'un sondage la position d'un individu ne nous intéresse pas.

```
s.label(pca.piscine$li)
```

En revanche, la représentation des relations entre les variables est un peu différente car les variables originelles ne sont pas situées en fonction de leurs corrélations avec les variables de synthèse mais de leurs covariances. Globalement, les interprétations en termes de directions restent valables, mais les flèches ne sont plus contenues dans un cercle de longueur 1 et leur longueur n'est plus forcément gage de leur qualité de représention. On ne peut plus faire l'économie de la lecture des cosinus carrés.

On voit sur la figure **??** se dégager un groupe de variables reliées à des équipements utilisant l'eau pour le bien-être des clients et un second groupe constituant un "pôle social".

```
s.arrow(pca.piscine$co)
```

## 3.4 Diagnostics : les contributions à l'inertie

Nous allons maintenant souligner que la longueur des flèches, si elle ne suffit plus à exprimer la qualité de représentation donne quand même une indication sur l'importance de la variable dans la construction de l'image obtenue, ce qu'on appelle sa *contribution à l'inertie*.

En effet, il existe une troisième présentation de l'ACP où les $p$ variables originelles $X_1$, $X_2$ $\ldots X_p$ sont combinées par un système de coefficients $a_1$, $a_2$ $\ldots a_p$ pour donner une variable de synthèse $a_1X_1 + a_2X_2 + \cdots + a_pX_p$.

**Définition 5** *L'ACP peut être vue comme la recherche d'une combinaison linéaire de variance maximum des variables originelles soit centrées (dans l'ACP sur matrice de covariances) soit standardisées (dans l'ACP sur matrice de corrélations).*

*On peut ensuite rechercher une deuxième combinaison linéaire maximisant le même critère avec une contrainte d'orthogonalité entre les systèmes de coefficients de la première et la deuxième.*

**Remarque 4** *On utilise une contrainte de taille pour ces coefficients* $a_1^2 + a_2^2 + \cdots + a_p^2 = 1$*, sinon, il suffirait de multiplier ces coefficients par 2 (par exemple) pour que la variance de la variable de synthèse soit multipliées par* $2^2 = 4$*.*

On peut donc calculer les variables qui ont le plus participé à la construction de chaque variable de synthèse par l'intermédiaire des quantités $a_j^2$. Ces quanti-

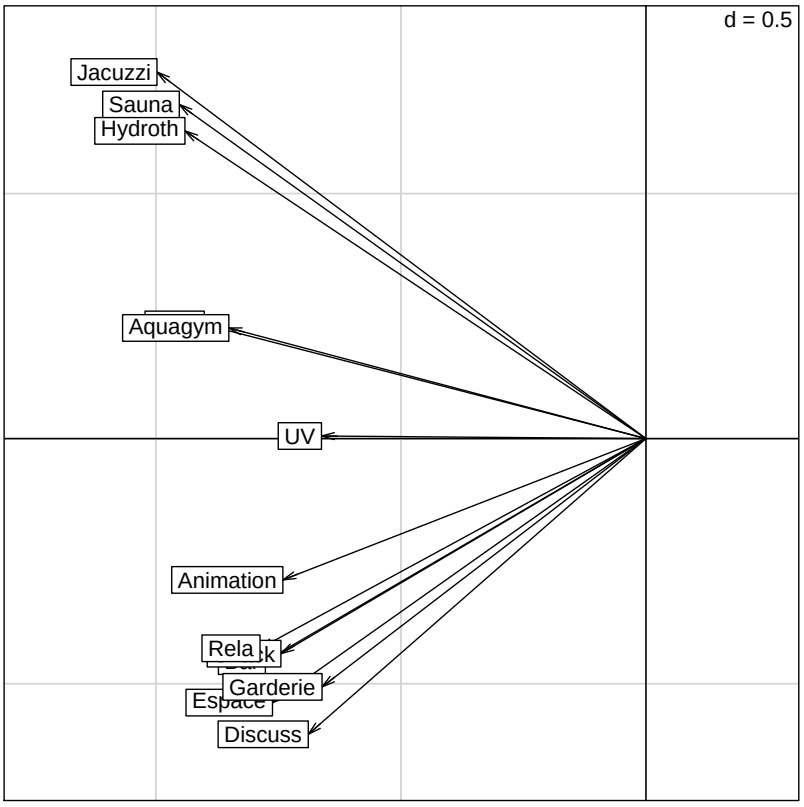


FIG. 3.1 – Représentation des variables suite à l'ACP centrée du dataframe piscine2

tés dites contributions à l'inertie sont calculées par la fonction inertia.dudi, leurs valeurs résidant dans la composante $col.abs [2].

**Définition 6** *Les contributions à l'inertie des variables (resp. des lignes) déterminent l'importance de chacune dans la construction des variables de synthèse, et donc la signification à leur accorder.*

On voit ainsi en ce qui concerne l'exemple des piscines que la contribution à l'inertie de la variable UV à la construction de la première variable de synthèse est de 5.11% et à la deuxième est de 0%. La première variable a été contruite par l'ensemble des variables (caractéristique d'un effet taille), en revanche, la seconde l'a été essentiellement par les variables Jacuzzi, Sauna, Hydroth, Discuss.

```
inertia.dudi(pca.piscine,col.inertia=TRUE,row.inertia=FALSE)
```

La longueur des flèches sur le graphique des covariances reflète ces contributions à l'inertie, donc une variable signifiante pour la construction du graphique et son interprétation possède une flèche longue. En ACP on peut donc se passer de la lecture fastidieuse des contributions, elles sont visibles [3].

## 3.5 Le biplot

Un laboratoire pharmaceutique (exemple tiré de l'ouvrage très accessible de Foucart [5]) a réalisé une étude du choix de dentifrices auprès de 100 personnes sur les critères suivants : HALE (haleine agréable), TART (élimine le tartre), PHAR (vendu en pharmacie), GEN (protège les gencives), BLAN (blanchit les dents), CAR (empêche les caries), GOUT (a bon goût) et PRIX (prix d'un tube). Les notes données varient de 1 (pas important) à 5 (très important). Le fichier dentifrices.csv comporte donc 100 lignes et 8 colonnes.

**Exercice 3** *Réaliser l'ACP de ce jeu de données*
- *Que choisir : ACP sur matrice de covariances ou de corrélations ?*
- *Quelle dimension nous incline à retenir l'éboulis des valeurs propres ?*
- *Produire le plan factoriel des individus. Quel est son intérêt en l'espèce ?*
- *Produire le graphique des covariances. Quelles sont les liaisons entre les variables ?*
- *Produire les diagnostics. Quelles sont les variables importantes pour la construction du graphique. Quelles sont les variables les plus mal représentées sur ce graphique ?*

Nous allons maintenant étudier un graphique qui réunit les résultats concernant les lignes et les colonnes : le *biplot* (figure 3.2). Il est basé sur une propriété

[2] On ne va pas s'intéresser aux contributions à l'inertie des lignes pour l'exemple des piscines car dans un sondage les individus ne nous intéressent pas personnellement

[3] C'est même plus intéressant car la plupart des logiciels donnent les contributions à l'inertie axe par axe or ce qui nous intéresse c'est généralement la solution classique de dimension 2, qui se traduit dans la longueur de la flèche!

de l'ACP dite de *reconstitution des données*. on peut démontrer que le tableau de départ peut être reconstitué, au moins de façon approximative, en utilisant les premières variables de synthèse et les coefficients correspondant.

Pour reconstituer la valeur de l'individu 55 (par exemple) par rapport à la variable CARI (par exemple), on projette orthogonalement l'unité 55 sur la droite correspondant à la variable sur le graphique. On constate que, par rapport au point origine, cette projection est à l'opposé de la direction CARI, l'unité dans le tableau a alors très probablement [4] une valeur plus faible (c'est 1) que la moyenne (3.43) pour cette variable. En revanche, l'unité 59 qui est projetée dans la direction de CARI a probablement une valeur (c'est 5) supérieure à la même moyenne. Si un individu comme le 56 a une projection proche de la moyenne, il doit avoir une valeur moyenne, (en fait oui et non, puisque c'est 2).

```
dentifrice<-read.table("dentifrice.csv",header=TRUE,sep=";",dec=",")
pca.dentifrice<-dudi.pca(dentifrice,center=TRUE,scale=FALSE)
scatter(pca.dentifrice)
```

## 3.6 Pour en finir avec les données homogènes

L'analyse de données homogènes a donc montré que l'ACP est une méthode de décomposition de la structure d'un tableau qui pouvait être centré ou standardisé. En allant plus loin, le tableau à décomposer peut être transformé de bien des façons.

Ainsi, on peut imaginer que prendre le point moyen comme point origine obligé n'est pas toujours judicieux. Dans le package ade4, le jeu de données deug2 constitue un exemple très intéressant où des notes d'étudiants doivent plutôt être rapportées à 10, ce qui est la pratique usuelle des examinateurs plutôt qu'à la moyenne du groupe comme le fait auutomatiquement l'ACP (centrée ou standardisée).

Il est également possible de ne pas centrer le tableau, voir dans le package un exemple avec le célèbre jeu de données tortues. Le point origine est alors le point zéro.

On peut parfois retirer du tableau une information déjà connue. Ainsi en régressant chacune des variables originelles sur un groupe de variables considérées comme explicatives, il est possible d'analyser les résidus ce qui donne toute la famille dites des ACP sur variables instrumentales (voir les fonctions pcaviortho ou pcavi dans ade4).

Enfin, lorsque les variables sont la même mesure qui est répétée dans le temps, il existe des méthodes factorielles qui sont spécifiques mais d'une grande complexité (voir [6] si très motivé).

[4] probablement signifie ici que le résultat de l'approximation dépend de la qualité du graphique. Si la représentation à deux dimensions constitue une grande part de l'information, la reconstitution est bonne

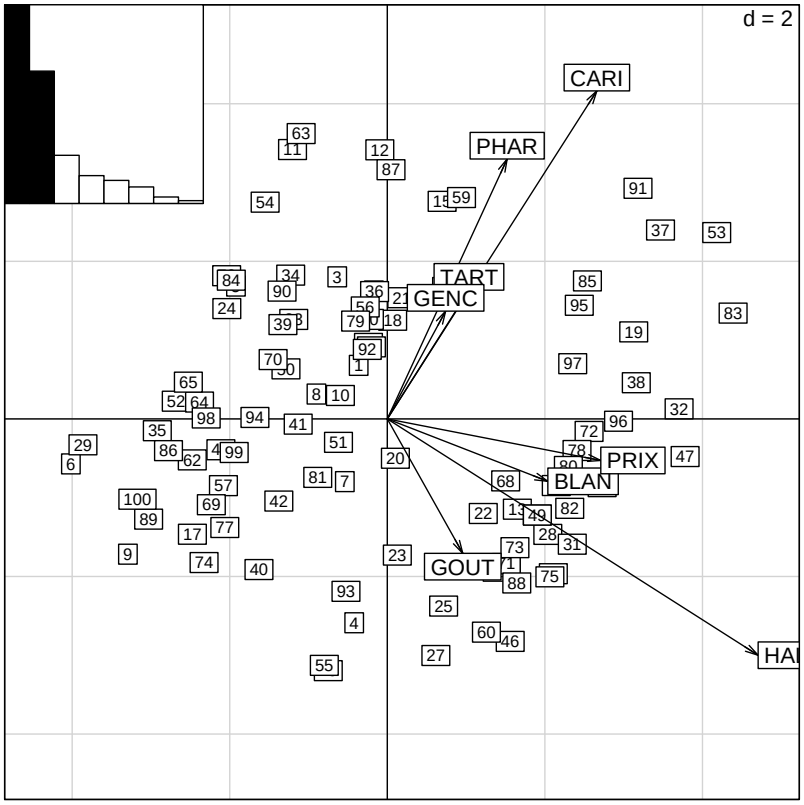


FIG. 3.2 – Biplot de l'ACP centrée du dataframe dentifrice

# Chapitre 4

# L'Analyse des Correspondances Multiples

## 4.1 Les données multivariées qualitatives

Le fichier boubouille.csv [1] contient en ce qui concerne 134 chattes, une mesure d'âge répartie en classes d'âges, une mesure de fécondité répartie en classes (nombre de chatons conçus dans l'année) et une mesure, considérée elle aussi comme qualitative, du nombre de portées dans l'année. Remarquons que toutes ces mesures initialement quantitatives ont été transformées en variables qualitatives, ce qui correspond pour le logiciel R à la "class" d'objets dite factor [2].

```
boubouille<-read.table("boubouille.csv",header=TRUE,sep=";",dec=",")
class(boubouille[,"age"])
class(boubouille[,"fecond"])
class(boubouille[,"nbport"])
```

Ces trois variables doivent bien sûr être initialement étudiées de façon univariée, afin de voir pour chacune si les catégories ont des fréquences observées similaires ou si certaines sont très rares [3] et quel est le nombre de catégories.

Puis, il faut également étudier, si le nombre de variables n'est pas trop grand, les relations entre ces variables deux à deux, ce qui se fait classiquement à l'aide de la statistique du chi-carré, avec le test afférent. On s'aperçoit en particulier que les trois variables considérées sont très liées.

[1] qui est une version particulière du fichier chatcat du package ade4

[2] Pour que la conversion des données du fichier en factor se fasse de façon automatique, les mesures sont des chaînes de caractères. Par exemple 1 devient un. Sinon, le logiciel confronté à des nombres considère à raison qu'il s'agit d'une mesure numérique. Dans ce cas, il faut forcer la mesure à devenir qualitative grâce à la fontion factor

[3] ce qui aura un impact sur l'efficacité de l'analyse

```
plot(boubouille[,"age"])
plot(boubouille[,"fecond"])
plot(boubouille[,"nbport"])
chisq.test(table(boubouille[,"age"],boubouille[,"fecond"]))
chisq.test(table(boubouille[,"age"],boubouille[,"nbport"]))
chisq.test(table(boubouille[,"fecond"],boubouille[,"nbport"]))
```

## 4.2 Une vision théorique proche de l'ACP

L'*Analyse des Correspondances Multiples* (ACM) permet d'étudier des données multivariées qualitatives. Elle est susceptible de nombreuses approches théoriques différentes (voir [7] pour une synthèse). Je vais privilégier une vision très proche de la présentation de l'ACP.

Dans l'ACP, on cherche une variable de synthèse (numérique) liée avec les variables originelles (numériques elles aussi) ce qu'on formalise à l'aide du coefficient de corrélation linéaire. Lorsque les variables de départ sont qualitatives, la liaison avec la variable de synthèse (numérique) se quantifie en termes de *rapport de corrélation*.

**Remarque 5** *Qu'est-ce que le rapport de corrélation (voir [8] pour un développement détaillé de la notion) ? Lorsqu'on cherche si une variable numérique est reliée à une variable qualitative - mettons que cette dernière ne comprenne que deux catégories - on calcule la moyenne de la variable numérique dans chacun des deux groupes d'unités repérées par les catégories. Plus ces deux moyennes diffèrent, plus le comportement de la variable numérique est modifiée suivant le groupe considéré c'est-à-dire la variable qualitative, donc plus les deux variables sont liées. Le rapport de corrélation correspond à la variance de ces deux moyennes (qui quantifie leurs différences et se généralise au cas de plus de deux groupes) rapportée à la variance générale. Le rapport de corrélation est donc une quantité évoluant de 0 (lorsque les deux variables ne sont pas liées) à 1 lorsqu'elles le sont parfaitement.*

L'ACM va donc reprendre l'objectif de l'ACP mais à l'aide d'un indicateur de liaison - le rapport de corrélation - adapté à la nature des variables originelles.

**Définition 7** *L'ACM est la recherche d'une variable de synthèse telle que la somme* [4] *de ces rapports de corrélation avec les variables originelles soit maximum.*

On obtiendra donc une variable de synthèse, qui a un pouvoir de synthèse, appelé là encore valeur propre, et qui permet de séparer au mieux les catégories pour chaque variable au sens où les moyennes dans chaque groupe sont les plus différentes possible.

[4] ou la moyenne ce qui est strictement équivalent

## 4.3 Réaliser l'ACM avec ade4

La fonction dudi.mca permet d'obtenir un objet de class dudi dont les composantes peuvent être utilisées de la façon suivante :

- la variable de synthèse $l1 grâce à laquelle on peut, pour chaque variable originelle, calculer
- les moyennes des catégories dans $co, ces moyennes donnant
- les rapports de corrélation dans $cr,
- la moyenne de ces rapports constituant la valeur propre donnée dans $eig.

```
acm.boubouille<-dudi.acm(boubouille)
acm.boubouille$eig
acm.boubouille$cr
acm.boubouille$co
```

En l'espèce, la valeur propre est de 0.7011 ce qui signifie qu'en moyenne la relation est forte entre la variable de synthèse et les variables originelles, relations qu'on peut décomposer grâce aux rapports de corrélation qui sont respectivement de 0.56 (age), 0.79 (fecond) et 0.76 (nbport).

Les moyennes des catégories sont par exemple pour la variable nbport -1.06 pour le groupe 1p et 0.76 pour le groupe 2p. Il est préférable de représenter graphiquement l'ensemble de ces moyennes grâce à la fonction score [5].

```
score(acm.boubouille)
```

On a donc un graphique (figure 4.1) par variable dans l'ACM. Voyons comment on peut les relier. Le graphique du haut correspond à la variable age. Chacune des cinq lignes horizontales représente une des catégories de la variable et sur chaque ligne on voit des traits qui correspondent aux valeurs prises par la variable de synthèse amis uniquement pour les unités statistiques qui appartiennent à la catégorie en jeu. La moyenne de ces valeurs est calculé et le carré portant le nom de la catégorie est positionné à l'emplacement de cette moyenne.

La forte séparation de ces carrés/moyennes indique un fort rapport de corrélation entre la variable de synthèse et la variable considérée, donc une forte liaison.

Sur le graphique du haut, on voit donc que les chattes qui ont un an correspondent à des valeurs négatives de la variable de synthèse. En même temps, le graphique du milieu, qui concerne la fécondité, montre que les chattes qui ont des valeurs négatives sont celles qui ont eu 1-2 chatons. Parallèlement, dans le graphique du bas, la catégorie placée à cet endroit est celle d'une portée

[5] On constate ici que la fonction score n'a pas le même effet que dans le chapitre précédent. En effet, R est un langage de programmation par objet, ce qui signifie qu'une fonction peut avoir des comportement différents suivant la nature de l'objet auquel elle s'applique. En fait derrière la fonction générique, se cache en réalité deux fonctions différentes score.pca et score.acm

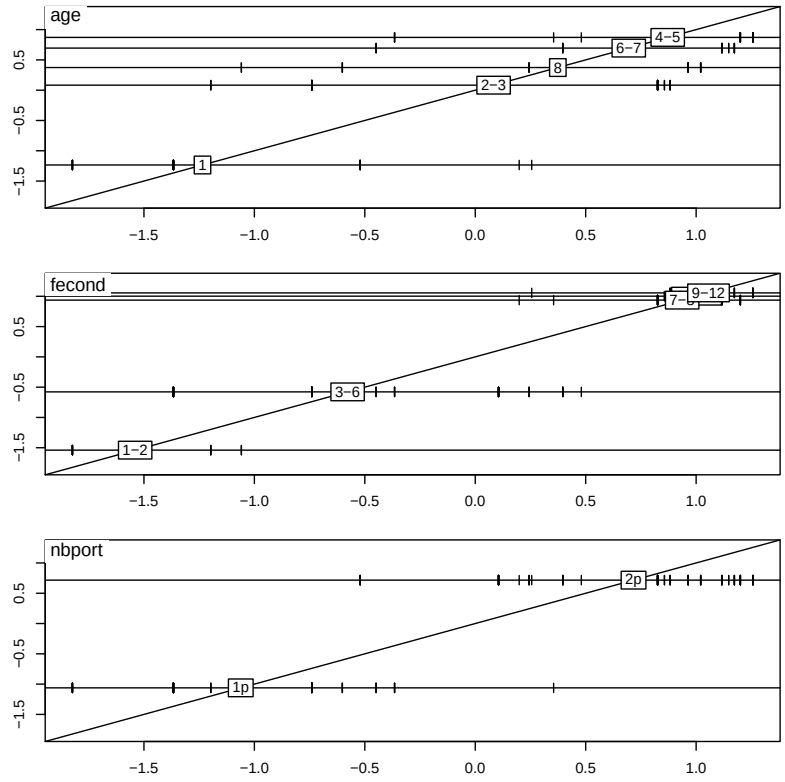


FIG. 4.1 – Graphique de l'ACM du dataframe boubouille. La première variable de synthèse permet de calculer pour chaque variable (age, fecond, nbport) les moyennes des catégories les portant et de les représenter sous forme de carrés

unique. La variable de synthèse sert donc à faire le lien entre les trois variables originelles.

On constate en résumé que plus les chattes vieillissent plus elles ont tendance à avoir deux portées et donc plus de chatons. Toutefois, vers 8 ans, leur fécondité diminue (et le nombre de portées aussi).

## 4.4 Solution du second ordre de l'ACM

Le fichier chiens.csv est décrit par exemple dans saporta. Il correspond pour 27 races de chiens (unités statistiques) à une mise en catégorie de leur taille, leur poids, leur vélocité . . .une catégorie plus élevée indiquant que cette caractéristique est plus prononcée dans cette race.

L'ACM donne une première valeur de synthèse dont le pouvoir de synthèse est de 0.4876. On peut estimer qu'il est intéressant de rechercher d'autres variables de synthèse comme en ACP, c'est-à-dire non corrélées à la première et maximisant le même critère de somme des rapports de corrélation maximum. Le processus peut s'itérer si cette deuxième variable de synthèse ne suffit pas.

```
chiens<-read.table("chiens.csv",header=TRUE,sep=" ;",dec=",")
for(j in 1 :6)chiens[,j]<-factor(chiens[,j]) [6]
acm.chiens<-dudi.acm(chiens)
acm.chiens$eig
```

L'éboulis des valeurs propres montre ici que deux dimensions ressortent nettement. Graphiquement, on va donc représenter la situation des unités statistiques par les deux variables de synthèse (ce qui donne un plan factoriel de ces unités) et pour chaque variable qualitative, les catégories sont placées au centre de gravité des unités qui lui correspondent. Dans le logiciel ade4, chaque variable est représentée séparément.

```
scatter(acm.chiens)
```

Comme précédemment, le lien se fait par la position des catégories. On commence d'abord par lire les rapports de corrélation des variables de synthèse avec les variables originelles (composante $cr) afin de concentrer l'étude sur les variables les plus structurantes. Ici, on voit sur la première variable de synthèse qu'il s'agit de la taille, du poids et de l'affection et pour la deuxième variable de synthèse du poids, de la vélocité et de la taille.

Le graphique 4.2 montre par exemple que les chiens de petite taille (catégorie 1) sont aussi les chiens de petit poids (catégorie 1).

**Exercice 4** *Interpréter l'ensemble de ces graphiques, en particulier :*

[6] Cette ligne est indispensable, sinon ade4 refuse de réaliser l'ACM. En effet, les valeurs du fichier sont toutes des chiffres de 1 à 3 que le logiciel prend pour des données numériques. Il faut donc préciser que ce sont en fait des catégories, ce que permet la fonction factor().

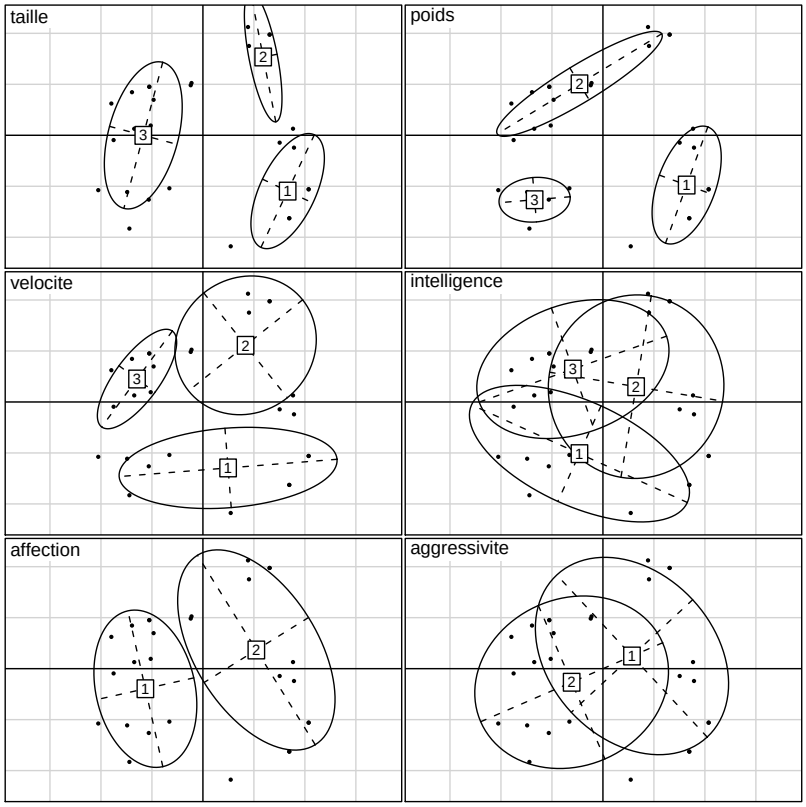


FIG. 4.2 – Pour chaque variable, les catégories sont situées au centre de gravité des unités les portant dont les coordonnées sont celles des deux variables de synthèses issues de l'ACM du dataframe chiens

- *que pensez-vous des tailles des chiens de poids de la deuxième catégorie ? et de la troisième ?*
- *Quel est à votre avis le sens des ellipses sur ces graphiques ?*
- *Que peut-on dire de l'aggressivité et de l'intelligence des chiens ?*

**Exercice 5** *Le package ade4 contient le fichier* ours *qui peut être étudié par une analyse des correspondances multiples.*

**Exercice 6** *Le package MASS contient le fichier* farms *qui peut aussi faire l'objet d'une ACM. Ce package contient par ailleurs des fonctions très utiles dont on peut avoir une description dans le remarquable ouvrage* **??**.

## 4.5 Quelques remarques sur l'ACM

L'ACM reste une analyse assez délicate à employer.

- Très souvent, les utilisateurs sont déçus car les éboulis des valeurs propres sont moins spectaculaires que ceux des ACP. En effet, la nature des variables qualitatives rend plus difficile l'existence de liens intenses.
- Les graphiques restent assez difficiles à lire, on n'a pas seulement une représentation de chaque variable dans un unique graphique comme en ACP, mais de leurs catégories ce qui multiplie les points et dans des graphique différents, sinon l'empilement est souvent illisible. Il faut plus d'expériences pour une utilisation efficace de ce type d'analyse.
- Il faut aussi savoir ne pas rentrer des dizaines de variables [7] dans une ACM car sinon l'exploitation en est impossible ; on se contente généralement des questions relatives au même thème.
- Il vaut mieux ne pas faire entrer n'importe quelle variable qualitative dans une ACM. On conseille (voir référence **??** pour l'explication théorique) de ne pas avoir des catégories de très faible effectif pour une variable qualitative et de ne pas non plus avoir de variable avec un trop grand nombre de catégories. Sinon, ces variables déséquilibrent l'analyse.

[7] par exemple toutes les quesntions d'un sondage

# Chapitre 5

# Une généralisation : l'analyse de Hill et Smith

## 5.1 Les données multivariées mixtes

Le fichier vie.csv contient pour 105 individus les résultats d'une enquête sur les conditions de vie des français (cf ouvrage de [?]). Les mesures sont le sexe, l'âge ainsi que des questions de comportement et d'opinion. Elles mêlent six variables qualitatives comme le sexe, et trois variables quantitatives comme l'âge [1].

Ce jeu de données est représentatif des situations habituelles de prises de mesures car il est rare que les données recueillies soient toutes quantitatives ou toutes qualitatives. On a habituellement affaire à des textsfdonnées multivariées mixtes. On ne peut donc directement employer l'ACP ou l'ACM sur les données. Il est dommage, pour des raisons purement techniques, de séparer des mesures potentiellement reliées et de réaliser deux analyses.

```
vie<-read.table("vie.csv",header=TRUE,sep=";",dec=",")
print(vie)
```

## 5.2 Analyser des données mixtes

Pour s'en sortir avec des moyens connus, les données quantitatives (âge, revenu) sont réparties en classes et on les considère comme qualitatives afin de les injecter aussi dans une ACM.

**Exercice 7** *Transformer les variables age, logement et television en variables*

[1] J'ai considéré que les variables concernant le logement et la télévision, mesurées sur une échelle de likert de 4 niveaux étaient quantitatives

*qualitatives (utiliser la fonction R cut pour la variable age* [2] *et la fonction textsf-factor pour les deux autres. Réaliser une ACM des neuf variables qualitatives.*

Une deuxième solution (voir par exemple dans [?]) est d'utiliser la technique dite des variables supplémentaires, une ACM est pratiquée sur les mesures qualitatives, puis les corrélations entre les variables de synthèse obtenues et les mesures numériques originelles permettent de situer ces dernières graphiquement sur le même plan que les autres. L'ennui est que les mesures numériques ne sont pas partie prenante de l'analyse mais sont uniquement utilisées dans l'interprétation.

**Définition 8** *L'*analyse de Hill et Smith *consiste à rechercher une variable de synthèse la plus reliée possible aux variables originelles, au sens du coefficient de détermination si la variable est numérique ou du rapport de corrélation si elle est qualtative. La somme (ou la moyenne) de ces indicateurs de liaison est maximisée.*

*Le processus peut être itéré en cherchant une deuxième variable de synthèse non corrélée qui optimise le même critère.*

## 5.3 Réaliser l'analyse mixte avec ade4

La fonction dudi.mix [3] permet de réaliser l'analyse de Hill et Smith. Elle génère un objet dont les composantes les plus intéressantes sont les suivantes : $cr contient les coefficients de détermination ou les rapports de corrélation, $eig contient les pouvoirs de synthèse, $l1 contient les variables de synthèse et $co permet de positionner les variables numériques et les catégories des variables qualitatives.

L'analyse de l'éboulis des valeurs propres devrait nous conduire ici à retenir trois variables de synthèse. Notons que les trois valeurs propres correspondantes sont les seules à être supérieures à 1.

```
mix.vie<-dudi.mix(vie)
```

## 5.4 Les représentations graphiques dans l'analyse mixte

En termes de représentations graphiques, on peut réaliser un plan factoriel des unités statistiques à l'aide des variables de synthèse. On peut "par dessus", comme dans un biplot, rajouter les variables numériques et les centres de gravité correspondant aux catégories pour les variables qualitatives. C'est ce que fait la fonction scatter.mix du package ade4 (voir figure 5.1).

[2] avec un découpage de type 20-30, 30-40 ...

[3] Il existe aussi une fonction dudi.hillsmith quasiment équivalente

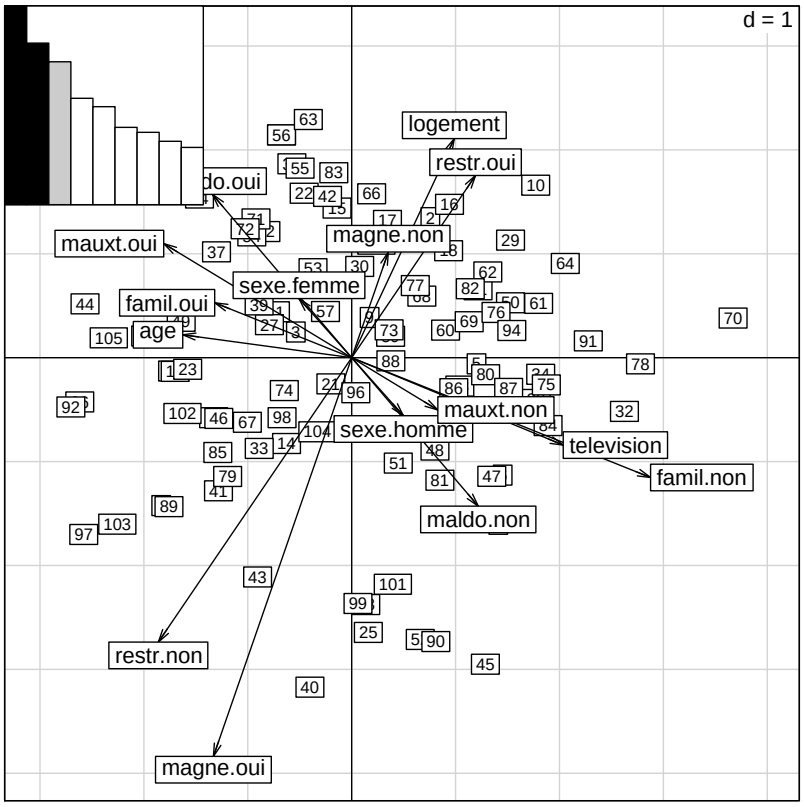


FIG. 5.1 – Les deux premières variables de synthèse issues de son analyse de Hill et Smith permettent de positionner les individus, les variables numériques sont placées dans l'esprit d'un biplot et les catégories de variables qualitatives comme des centres de gravité

scatter(mix.vie)

Je préfère employer trois fonctions que j'ai créées et qui sont présentées en annexe de ce document : scatter.mix.cr, scatter.mix.numeric et scatter.mix.categorical. Un premier graphique (figure 5.2) consiste à représenter les indicateurs de liaison (coefficients de détermination ou rapports de corrélation) entre les variables originelles et les variables de synthèse. On voit que les variables les plus liées à la première variable sont television et famille (et age) et pour la seconde logement, restrictions et magneto (pour n'abordons ici que les deux premières variables de synthèse).

Ensuite, on peut donc détailler les relations à l'aide d'un classique cercle des corrélations pour les variables numériques (figure 5.3) et une représentation de type ACM par centres de gravité pour les variables qualitatives (figure 5.4).

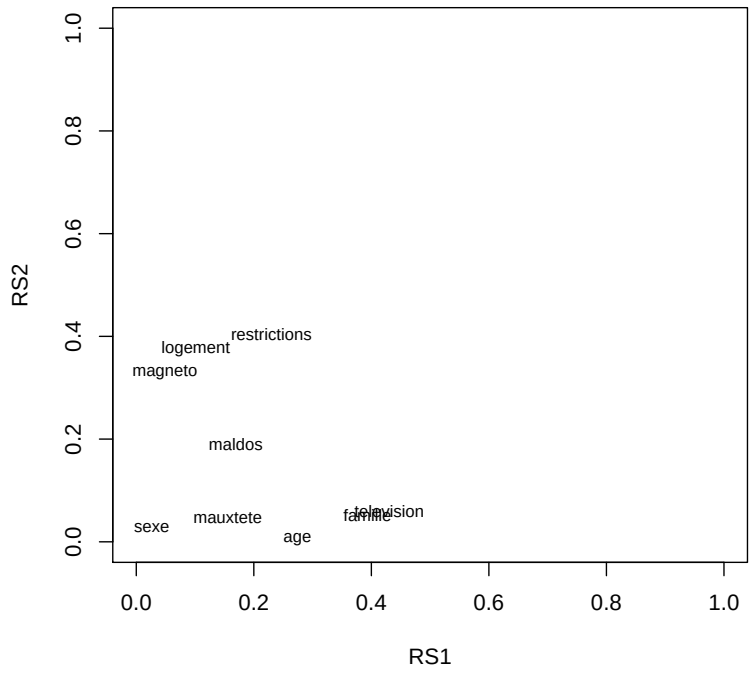


FIG. 5.2 – Coefficients de détermination ou rapports de corrélation entre les variables du fichier vie et les deux premières variables de synthèse issues de son analyse de Hill et Smith

On constate ainsi sur le premier axe, que plus on est âgé et plus on regarde la télévision (voir sur le cercle des corrélations) et qu'en même temps on a tendance à penser que la famille est le seul endroit où l'on se sente bien (oui dans les graphiques de type ACM pour la variable famille du côté de la flèche âge)

```
scatter.mix.cr(mix.vie)
scatter.mix.numeric(mix.vie)
scatter.mix.categorical(mix.vie)
```

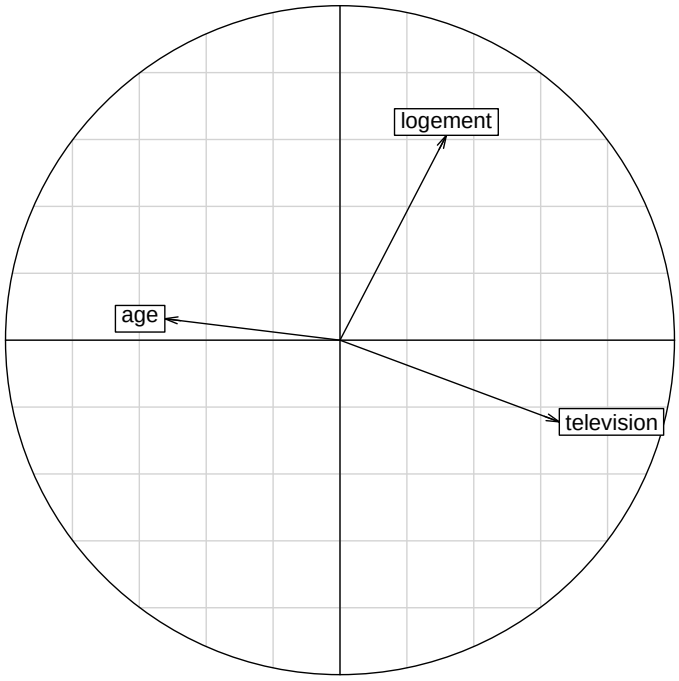


FIG. 5.3 – Cercle des corrélations des variables numériques du fichier vie avec les variables de synthèse issues de son analyse de Hill et Smith

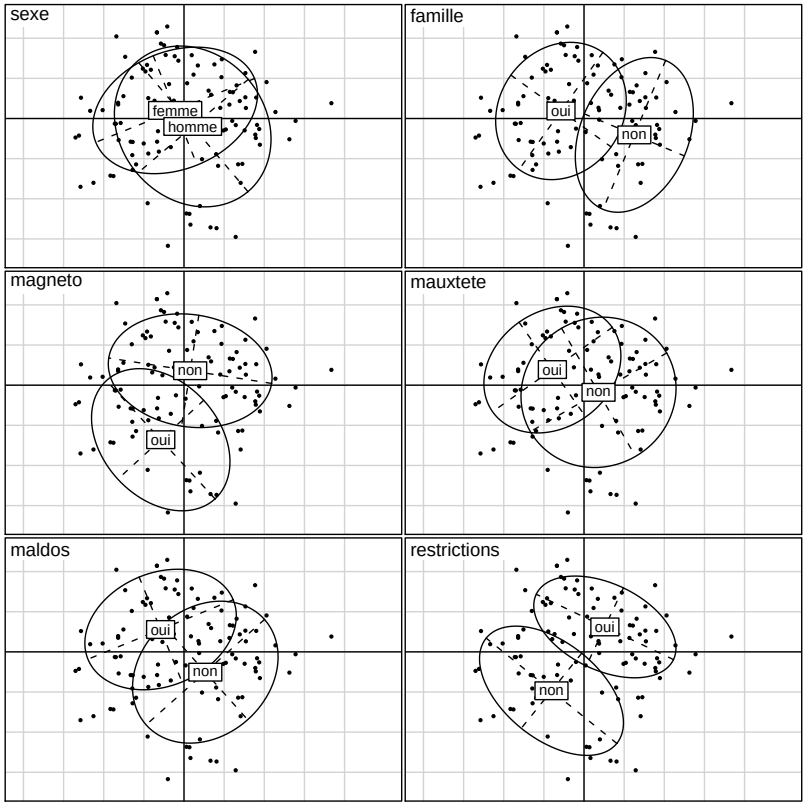


FIG. 5.4 – Représentation de type ACM pour les variables qualitatives du fichier vie et les deux premières variables de synthèse issues de son analyse de Hill et Smith

# Chapitre 6

# L'analyse des Correspondances Simples

Cette méthode, assez spécifique, a été utilisée dans des contextes variés par l'école d'analyse des données "à la française" et reste très employée par les praticiens des sciences sociales et humaines. Á ce titre, nous lui accordons un développement.

## 6.1 Le tableau croisé

Stricto sensu, l'*Analyse factorielle des correspondances (AFC)* s'applique à des tableaux croisés, dits aussi *tables de contingence*.

Ce type de tableau est obtenu en croisant deux variables qualitatives à $I$ et $J$ catégories. Chaque cellule du tableau correspond alors au nombre d'unités statistiques appartenant simultanément aux deux catégories. Les lignes ne sont donc pas des unités statistiques et les colonnes des variables mais il s'agit dans les deux cas de catégories.

Ainsi, le fichier couleurs.csv (tiré de [**?**]) provient d'une population de 592 femmes sur lesquelles ont été enregistrés deux caractères : (1) la couleur de leurs yeux (4 catégories) et celle de leur cheveux (4 catégories). L'information initiale correspond donc à un tableau à 592 lignes et deux colonnes qu'on pourrait soumettre à une ACM [1] mais qui en termes de croisement fournit une table $4 \times 4$. L'analyse multivariée est donc ici une analyse bivariée.

## 6.2 L'analyse habituelle d'un tableau croisé

Elle consiste à calculer la statistique du chi-carré afin de mesurer la liaison entre les deux variables. Si cette liaison est statistiquement significative (comme

[1] cela aurait d'ailleurs exactement le même effet !

TAB. 6.1 – Tableau croisé des couleurs yeux et cheveux

| | Cbrun | Cchatain | Croux | Cblond |
|---|---|---|---|---|
| Ymarron | 68 | 119 | 26 | 7 |
| Ynoisette | 15 | 54 | 14 | 10 |
| Yvert | 5 | 29 | 14 | 16 |
| Ybleu | 20 | 84 | 17 | 94 |

c'est le cas avec le fichier couleur), on explore la structure du tableau en comparant ses lignes en termes de fréquences. Pour de petits tableaux tels que celui qui nous sert d'exemple, c'est largement suffisant, mais lorsque lignes et colonnes sont nombreuses, cette étude devient complexe.

**Exercice 8** *Les commandes R ci-dessous permettent d'importer le tableau couleur, de calculer la statistique du chi-carré et les fréquences correspondant à chaque ligne du tableau. Interpréter ces résultats.*

```
couleur<-read.table("couleur.csv",header=T,sep=";",dec=",")
chisq.test(couleur)
sweep(couleur,1,apply(couleur,1,sum),"/")
```

## 6.3 Une vision théorique basée sur le "scoring"

La technique du *scoring* permet d'étudier des tableaux croisés de grande dimension. On peut en effet affecter *a priori* un score à chacune des colonnes, par exemple (-1,-1,1,1) [2] qui opère une opposition entre cheveux foncés (brun, chatain) et clairs (roux, blond).

Á chaque couleur des yeux - chaque ligne - correspond une fréquence observée pour les cheveux (voir tableau 6.2). Ainsi pour les yeux marrons, on obtient (0.3091,0.5409,0.1182,0.0318). Il est possible de calculer le score moyen pour les yeux marrons qui est de : $0.3091 \times (-1) + 0.5409 \times (-1) + 0.1182 \times (1) + 0.0318 \times (1) = -0.7$. Ce score moyen négatif montre que ces individus ont des cheveux plutôt foncés.

Pour les yeux bleus, on obtient un score moyen de $0.0930 \times (-1) + 0.3907 \times (-1) + 0791 \times (1) + 0.4372 \times (1) = 0.0326$ qui est positif indiquant que les cheveux clairs dominent dans cette sous-population.

On pourrait donc assez bien discriminer les quatre couleurs des yeux sur la base du scoring proposé pour la couleur des cheveux. Cependant, nous pouvons nous poser deux questions : (1) existe-t-il un scoring des cheveux permettant de discriminer encore mieux les couleur des yeux et (2) lorsque nous connaissons moins le sujet - l'opposition clair/foncé était attendue en l'espèce - ou ne voulons

[2] Ces scores sony généralement centrés et standardisés d'une manière ou d'une autre

TAB. 6.2 – Tableau de profils des couleurs yeux et cheveux % en lignes

| | Cbrun | Cchatain | Croux | Cblond |
|---|---|---|---|---|
| Ymarron | 0.3091 | 0.5409 | 0.1182 | 0.0318 |
| Ynoisette | 0.1613 | 0.5806 | 0.1505 | 0.1075 |
| Yvert | 0.0781 | 0.4531 | 0.2188 | 0.2500 |
| Ybleu | 0.0930 | 0.3907 | 0.0791 | 0.4372 |

pas utiliser d'a priori, pouvons-nous définir un scoring optimal, qui nous aiderait alors à mieux comprendre la structure du tableau de données ?

**Définition 9** *L'AFC est une méthode permettant de définir pour un tableau croisé un scoring sur les colonnes* [3] *tel que les scores moyens des lignes (obtenus en utilisant les fréquences du tableau des profils) soient les plus discriminants possible, au sens de la variance de ces scores moyens.*

## 6.4 Réaliser l'AFC avec ade4

La fonction dudi.coa permet d'obtenir un objet de "class" dudi.coa. Les scores optimaux standardisés des colonnes sont disponibles à partir de la composante $c1, les scores moyens des lignes correspondant dans son argument $li et la variance maximum obtenue dans $eig.

```
coa.couleur<-dudi.ca(couleur)
coa.couleur$eig
coa.couleur$c1
coa.couleur$li
score(coa.couleur)
```

On peut constater que la méthode a choisi comme scoring optimal pour les colonnes (-1.1042,-0.3244,-0.2834,1.8282) qui reflète la structure majeure de ce jeu de données, l'opposition claire/foncé. On peut représenter ces scores [4] graphiquement comme dans la figure 6.1.

**Remarque 6** *Si on décide de chercher à l'inverse un scoring des lignes optimal dans le sens qu'il fournisse les scores moyens les plus discriminants possibles pour les colonnes, on retrouve, à une dilatation près, les mêmes résultats. L'AFC est une méthode symétrique pour les lignes et les colonnes du tableau.*

[3] soumis à des contraintes de standardisation particulières

[4] En fait la figure présentée ne donne pas exactement ces scores, mais une version "dilatée" afin de les considérer de façon symétrique, voir le détail dans l'aide à la fonction score.coa

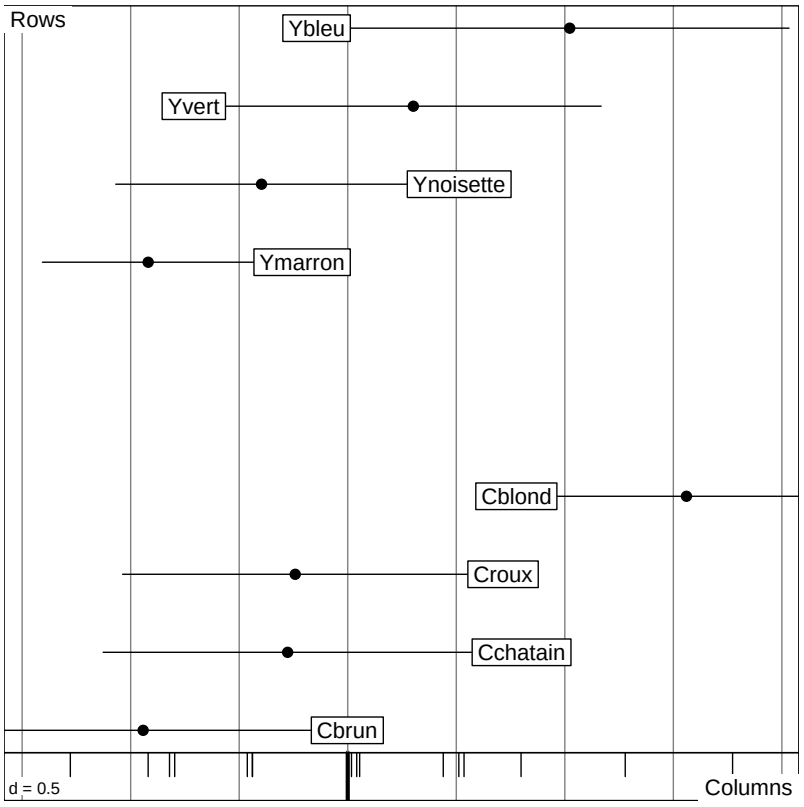


FIG. 6.1 – Représentation des scores optimaux des colonnes et des scores moyens de lignes pour l'AFC du fichier couleur

## 6.5 Solution du second ordre de l'AFC

Le jeu de données housetasks fourni avec le package ade4 décrit des tâches diverses accomplies dans un ménage (faire les papiers, laver le linge, la vaisselle, conduire, bricoler ...) croisées par la personne qui les accomplit (homme, femme, les deux ou chacun son tour).

L'AFC de ce tableau montre que le pouvoir de synthèse du premier score est de 0.5428 qui est à comparer avec le maximum possible de 1.1149 (Il s'agit de $fracX^2n$ où $n = 1744$ est le nombre d'unités dans le tableau et $X^2 = 1944.456$ la valeur observée de la statistique du chi-carré). La première variable ne contient donc que 49% de l'information du tableau croisé.

Comme les autres analyses factorielles, l'AFC peut être itérée. On recherche alors un scoring des colonnes, orthogonal au premier, qui optimise le même critère de discrimination des scores moyens des lignes.

Des représentations graphiques des lignes et des colonnes en dimension deux sont alors envisageables. Il est possible :

- de représenter les colonnes par leur scorings optimaux et les lignes par les scores moyens correspondant
- à l'inverse, les scorings optimaux des lignes et les scores moyens des colonnes ou
- de réaliser un graphique de type biplot où lignes et colonnes sont représentées à la même échelle. C'est la voie la plus classique choisie par la fonction scatter.coa (voir figure 6.2) mais l'argument met permet d'obtenir les deux autres variantes.

```
data(housetasks) housetasks
chisq.test(housetasks)
sum(housetasks)
coa.house<-dudi.coa(housetasks)
scatter.coa(coa.house)
```

**Exercice 9** *Analyser précisément la répartion des tâches à l'aide de la figure 6.2.*

**Exercice 10** *Le fichier television.csv correspond au nombre d'heures de diffusion par les principales chaînes de télévision française (en colonnes) des sports majeurs (en lignes).*

- *Réaliser l'AFC de ce jeu de données*
- *Quelle dimension de représentation choisir ?*
- *Représenter uniquement le premier score lignes et colonnes. Quel est leur signification ?*
- *Faire une représenttion graphique à deux dimensions des scores 2 et 3. Qu'en pensez-vous ?*

**Exercice 11** *Le jeu de données sarcelles (plus exactement sa composante sarcelles$tab) fourni par le package ade4 croise une information temporelle et une*

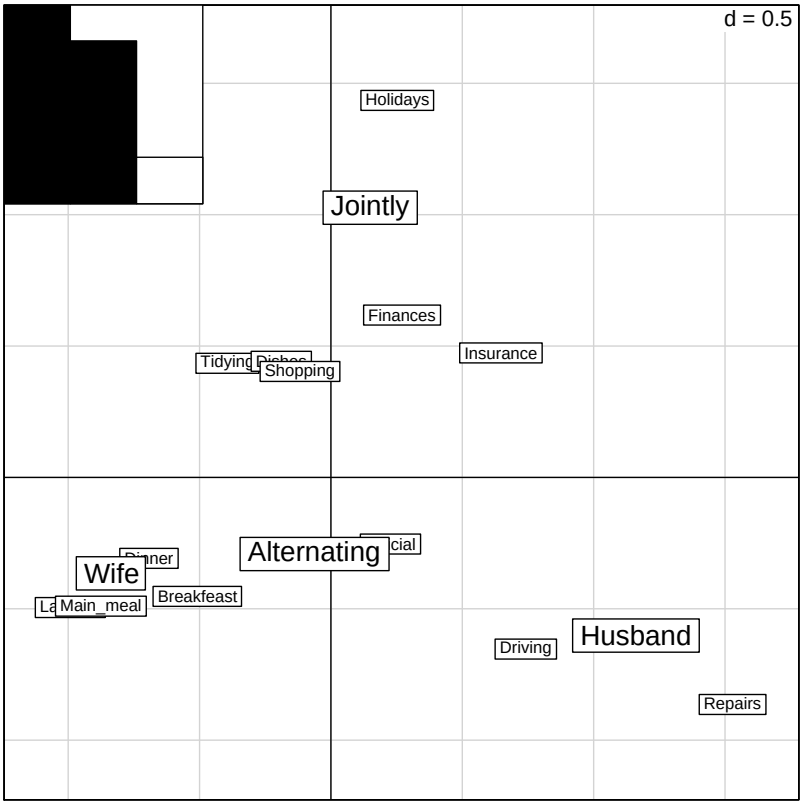


FIG. 6.2 – Représentation simultanée des scores lignes et colonnes (à la même échelle) de l'AFC du fichier housetasks

*information spatiale quant à la présence de ces oiseaux en Europe.*

*Réaliser l'AFC de ce jeu de données. Interpréter ces résultats. Pourquoi serait-il intéressant de disposer d'une carte et quelle représentation graphique pourrait-on envisager ?*

## 6.6 Diverses remarques sur l'AFC

L'AFC est une analyse parfaitement symétrique. Or parfois, les lignes et les colonnes ne jouent pas le même rôle dans la table de contingence. Il existe des variantes de l'AFC dites *AFC non symétriques* très intéressantes (voir la fiche d'aide de la fonction dudi.nsc dans le package ade4). En particulier la version non symétrique n'est pas perturbée par la présence de colonnes qui contiendraient un très faible effectif, au contraire de l'AFC qui y est très sensible (voir exercice 10).

L'AFC est une méthode dérivée de l'ACP qui emploie des pondérations très spéciales. L'une des conséquences est que la contribution des lignes et des colonnes ne peut se lire uniquement d'après leur éloignement sur les graphiques comme c'était le cas en ACP mais qu'il est absolument nécessaire de lire les fichiers de diagnostics afin de savoir à quoi s'en tenir. Il est tout à fait possible en AFC qu'une colonne ait une très forte contribution à l'inertie d'un axe en étant peu éloignée du point origine sur le graphe. A contrario, une colonne peut être loin du point origine mais n'avoir que peu de contribution. Tout cela est lié en fait à l'effectif de cette colonne.

# Annexe A

# Installation du logiciel R et du package ade4

## A.1 Installation de R pour Windows

1. Aller sur le site http ://www.r-project.org/ afin de récupérer le logiciel d'installation de R. Cliquer sur Download afin de trouver un site miroir proche pour limiter le temps de chargement, par exemple le site http ://cran.univ-lyon1.fr.
2. Charger une distribution précompilée (Precompiled Binary Distributions) pour Windows (95 and later). Cliquer sur Base, puis charger sur votre disque dur (C :) le logiciel d'installation rw2011.exe (la dernière version à l'heure où j'écris ce document, mais depuis ...).
3. Double-cliquer sur le logiciel rw2011.exe afin de procéder à l'installation de R. Choisir les options par défaut afin de l'installer sur le disque C.

## A.2 Utilisation de R

Utiliser le menu démarrer, puis Tous les programmes, puis R, puis R 2.0.0. Le logiciel s'ouvre alors et une fenêtre "Rconsole" apparaît.

Il faut indiquer au logiciel R dans quel répertoire windows il doit aller chercher les fichiers (en particulier les jeux de données) dont nous avons besoin et où les sauvegarder également ; ce répertoire est dit répertoire de travail. Dans le menu déroulant "File", existe une option "Change dir..." qui par l'intermédiaire d'une arborescence permet de choisir le répertoire qui nous convient (par défaut c'est C :\Program Files\R\rw2000).

On peut alors taper des commandes en face du "prompt >" qui s'inscriront en rouge, alors que les réponses de R apparaitront en bleu.

Lorsque la séquence de travail est terminée, il est bon de sauvegarder les objets R créés pour un usage ultérieur ce qui se fait par l'intermédiaire du menu

"File" avec l'option "Save Workspace". Lors d'une prochaine séance, on utilisera l'option "Load Workspace" du menu "File" afin de charger ces objets.

## A.3 Installation du package ade4

1. Recommencer la procédure d'installation comme pour R : prendre une distribution précompilée pour Windows, mais au lieu de choisir l'option "Base", choisir l'option "Contrib". Aller dans le répertoire "2.2/" et enregistrer ade4_1.3-3.zip (dermier répertoire et dernière version ade4 à l'heure où j'écris ces lignes) sur votre disque dur (C :).
2. Dézipper ce fichier dans le répertoire C :\Program Files\R\rw2000\library (si vous n'avez pas l'utilitaire de dézippage aller le récupérer sur le site http ://www.winzip.com/)

## A.4 Utilisation d'ade4

Une fois le logiciel R lancé, il suffit d'aller dans le menu déroulant "Packages" et de choisir l'option "Load Package". Une fenêtre apparaît alors où l'on peut sélectionner ade4 et valider ce choix en cliquant sur OK.

# Annexe B

# Quelques fonctions R

```
scatter.mix.cr<-
function(obj.dudi,xax=1,yax=2,...)
plot(obj.dudi$cr[,c(xax,yax)],type="n",xlim=c(0,1),ylim=c(0,1),...)
text(obj.dudi$cr[,c(xax,yax)],lab=row.names(obj.dudi$cr),cex=0.75)
```

```
scatter.mix.numeric<-
function(obj.dudi,xax=1,yax=2,...)
indexation<-obj.dudi$index=="q"
numero<-match(seq(1,length(obj.dudi$index))[indexation],obj.dudi$assign)
noms<-row.names(obj.dudi$co[numero,c(xax,yax)])
s.corcircle(obj.dudi$co[numero,c(xax,yax)],label=noms,...)
```

```
scatter.mix.categorical<-
function(dudi.obj,xax=1,yax=2,csub=2,possub="topleft",...)
def.par<-par(no.readonly=TRUE)
on.exit(par(def.par))
tabcomplet<-eval(as.list(dudi.obj$call)[[2]],sys.frame(0))
indexation<-dudi.obj$index=="f"
oritab<-tabcomplet[,indexation]
nvar<-ncol(oritab)
par(mfrow=n2mfrow(nvar))
for(i in 1 :nvar)s.class(dudi.obj$li,oritab[,i],
xax=xax,yax=yax,clab=1.5,sub=names(oritab[i],csub=csub,possub=possub,cgrid=0,csta=0)
```

# Annexe C

# Quelques jeux de données d'ade4

Les jeux de données d'ade4 permettent de mettre en oeuvre les techniques vues dans ce document. Afin de pouvoir analyser ces jeux de données, on doit faire appel au package ade4 avec

```
library(ade4)
```

les déclarer disponible par exemple pour le jeu de données mariages avec

```
data(mariages)
```

voir les informations détaillées avec

```
?mariages
```

Les jeux de données suivants sont à mon sens les plus simples (c'est-à-dire qu'ils ne réclament pas de connaissances bio-écologiques pointues :

- mariages
- atlas
- rhone
- veuvage
- chats
- skulls
- syndicats
- morphosport
- fruits
- ecomor
- chazeb
- clementines

- aviurba
- doubs
- bordeaux

# Bibliographie


[1] Escoufier, Y. (1987) The duality diagram : a means of better practical applications In Development in numerical ecology, Legendre, P. & Legendre, L. (Eds.) NATO advanced Institute, Serie G. Springer Verlag, Berlin, 139-156.

[2] Mardia K., Kent J. et Bibby J. (1979) Multivariate analysis. Academic Press.

[3] Baccini A. et Besse Ph. (1999) Statistique descriptive multidimensionnelle Publications du Laboratoire de Statistique et Probabilités, Université Paul Sabatier Toulouse .

[4] Lebart L., Morineau A. et Piron M. (1995) Statistique exploratoire multidimensionnelle. Dunod.

[5] Foucart T. (1997) L'analyse des données mode d'emploi. Presses Universitaires de Rennes.

[6] Ramsay J.O. et Silverman (19 ? ?) Functional data analysis. Springer-Verlag.

[7] Tenenhous M. et Young F.W. (1985) An analysis and synthesis of multiple correspondence analysis ; optimalsclaing, dual scaling homogeneity analysis and other methods for quantifying categoricalmultivariate data. Psychometrika, 50, p.91-119.

[8] Champely S. (2003) Statistique vraiment appliquée au sport. Editions de Boeck.

[9] Venables et Ripley (1999) Applied statistics with S+.