Analyse en Composantes Principales (ACP) : Travaux Pratiques avec le logiciel R

Frédéric Santos
CNRS, UMR 5199 PACEA
Courriel : frederic.santos@u-bordeaux.fr

24 février 2015

Table des matières

1. Détails techniques pour R 1
§1.1. *Installations préalables*, 1. — **§1.2**. *Packages à charger*, 2.
2. Brève introduction « théorique » 2
3. Étude de cas : Budgets de l'État de 1872 à 1971 3
§3.1. *Charger les données*, 3. — **§3.2**. *Matrice des corrélations*, 3. — **§3.3**. *Éboulis des valeurs propres*, 4. — **§3.4**. *Procéder à l'ACP*, 4. — **§3.5**. *Analyser le nuage des variables*, 4. — **§3.6**. *Analyser le nuage des individus*, 4. — **§3.7**. *Classification non supervisée*, 5. — **§3.8**. *Représenter des groupes déjà connus*, 5.
4. Gestion des valeurs manquantes. 5
5. Le problème de l'effet taille 6
§5.1. *Définition*, 6. — **§5.2**. *Éliminer l'effet taille avec un tableur*, 7. — **§5.3**. *Éliminer l'effet taille avec R*, 7.
Références. 8

1. Détails techniques pour R

§1.1. Installations préalables. — On commencera par se procurer la dernière version du logiciel R à l'adresse suivante : `http://cran.r-project.org/`

Cas spécifique des utilisateurs de Mac OS. — Le bon fonctionnement des interfaces graphiques de R peut nécessiter l'installation du logiciel Apple X11, disponible à l'adresse suivante : `http://apple-x11.softonic.fr/mac`

Installation de l'interface graphique et des packages spécifiques. — Il n'existe pas *nativement* dans R de fonction réellement efficace et élégante pour réaliser des ACP. En revanche, deux packages très complets écrits par des mathématiciens français contiennent toutes les options nécessaires :

— le package `FactoMineR` (adossé à l'université de Rennes) ;
— le package `ade4` (adossé à l'université de Lyon).

Ce document décrira l'utilisation de FactoMineR, qui est probablement le plus couramment utilisé — et le mieux documenté — des deux. De plus, le site web de FactoMineR contient de nombreux détails et des tutoriels vidéo : `http://factominer.free.fr/`.

D'autre part, R ne fonctionne nativement qu'en ligne de commandes, mais une interface graphique assez intuitive et en constante amélioration existe pour réaliser la plupart des opérations courantes « à la souris » : l'interface R Commander. La commande suivante [1] :

```
install.packages("RcmdrPlugin.FactoMineR", dep=TRUE)
```

permet d'installer en une seule fois l'interface graphique R Commander et tout le nécessaire pour réaliser des ACP « à la souris » en utilisant le package `FactoMineR`.

§1.2. Packages à charger. — Lors du lancement d'une session R, taper la commande `library(Rcmdr)` pour charger l'interface graphique. On peut aussi (sous Windows ou Mac OS) le faire à la souris, *via* le menu Packages > Charger le package... de la console R. Une fois l'interface chargée, utiliser le menu Outils > Charger des plugins Rcmdr pour ajouter les menus relatifs aux ACP.

2. Brève introduction « théorique »

On n'entrera pas ici dans les détails mathématiques ou techniques de l'analyse en composantes principales, mais de nombreux documents disponibles à la BUST ou sur Internet en offrent une présentation exhaustive et néanmoins très digeste [Bac10, C$^+$12, HLP09, Pag10].

L'ACP est une méthode exploratoire (i.e., descriptive) multivariée de réduction dimensionnelle. Lorsque les données sont constituées de n individus décrits par $p > 3$ variables *numériques*, il devient impossible d'effectuer une représentation graphique, comme cela peut être le cas pour $p = 2$ (nuage de points classique en 2D) ou $p = 3$ (nuage de points 3D).

Souvent, et surtout en biologie, les p variables sont fortement inter-corrélées : il y a une grande redondance dans l'information qu'elles délivrent. Il n'y a donc probablement pas besoin d'un espace à p dimensions pour restituer toute l'information, sachant que la plupart des dimensions apportent peu ou prou la même information. Une ACP consiste donc en la transformation des p variables originelles, fortement liées entre elles, en nouvelles variables décorrélées les unes des autres par construction. Ainsi, chaque variable nouvellement créée fournit toujours une information totalement nouvelle par rapport aux variables précédentes.

Ces nouvelles variables sont nommées *composantes principales*, ou plus simplement *axes*. La donnée d'un petit nombre (idéalement, deux ou trois) de composantes principales permet généralement d'épuiser la quasi-totalité de l'information initialement fournie par les p variables originelles inter-corrélées. En réalisant une ACP, on accepte toutefois de perdre un petit pourcentage d'information, ce qui n'est pas gênant, voire avantageux : la *structure* principale des données est conservée tandis que le *bruit* est éliminé.

En définitive, l'ACP permet de restituer en deux ou trois dimensions la structure et les proximités observées dans le nuage d'origine en p dimensions, avec la déformation la plus faible possible.

La construction effective des axes dépasse de très loin l'objectif de ce document, et repose essentiellement sur l'algèbre linéaire (diagonalisation des matrices symétriques réelles, projecteurs orthogonaux, etc.). Le lecteur déjà familier avec ces notions mathématiques, ou simplement curieux, pourra consulter des ouvrages plus techniques pour comprendre la construction effective des axes [EP08, Sap11].

On retiendra que le logiciel renvoie p composantes principales :

1. Ouvrir R et copier-coller la commande dans la fenêtre de script, tout simplement !

— ordonnées par pourcentage d'information restituée (la première composante est la plus importante, la seconde un peu moins, et les autres sont généralement très vite anecdotiques et inintéressantes : c'est le but !) ;
— mutuellement indépendantes.

En résumé, l'ACP s'effectue sur un tableau de n individus décrits par p variables intercorrélées $X_1, \cdots, X_p$; et crée à partir d'elles p composantes principales indépendantes :

$$\begin{aligned} C_1 &= \alpha_{1,1}X_1 + \cdots + \alpha_{1,p}X_p \\ &\vdots \\ C_p &= \alpha_{p,1}X_1 + \cdots + \alpha_{p,p}X_p \end{aligned}$$

Le nuage de points construit à partir des premières composantes contient généralement une information relativement fidèle du nuage de points originel à p dimensions.

3. Étude de cas : Budgets de l'État de 1872 à 1971

Télécharger le jeu de données disponible à l'adresse suivante :

`http://tinyurl.com/budgets-etat`

Ces données détaillent les différents postes de dépense de l'État français sur un siècle. Les valeurs sont données en pourcentage du budget global pour éliminer l'effet de l'inflation et de l'évolution de la valeur nominale du franc sur cette période.

Les postes sont notés PVP (pouvoirs publics), AGR (agriculture), CMI (commerce et industrie), TRA (travail), LOG (logement), EDU (éducation), ACS (action sociale), ANC (anciens combattants), DEF (défense), DET (remboursement de la dette), DIV (divers).

Ici, les « individus » sont les années, et les variables sont les postes de dépense. Les années sont donc des points dans un espace de dimension 11 (égal au nombre de variables).

Durant le siècle considéré, il y a eu deux guerres mondiales, deux phases de forte croissance économique et une phase de forte récession. On se propose d'établir une typologie des budgets de l'État, que l'on interprétera à l'aide de nos connaissances sur le contexte historique.

§3.1. Charger les données. — Ouvrir l'interface graphique de R en chargeant le package `Rcmdr`. Importer le fichier de données *via* le menu Données > Importer des données > depuis un fichier texte. Pour ce fichier CSV, le séparateur de champs est le point-virgule, le séparateur décimal est la virgule, et les données manquantes sont signalées par une cellule vide.

Comme toujours sous R, un excellent réflexe pour voir si le jeu de données a été correctement chargé est de cliquer sur le bouton Visualiser de l'interface R Commander. On peut également afficher un résumé du jeu de données afin de vérifier qu'il ne contient pas d'anomalie : menu Statistiques > Résumés > Jeu de données actif.

Enfin, pour indiquer que la première colonne (`NOM`) contient le nom des individus, aller dans le menu Données > Jeu de données actif > Nom des cas.

§3.2. Matrice des corrélations. — Avant de procéder à l'ACP, on peut éventuellement calculer la matrice de corrélation des variables initiales *via* le menu Statistiques > Résumé > Matrice de corrélations.

Identifier rapidement quelques paires de variables fortement corrélées, et quelques paires de variables quasiment décorrélées.

§3.3. Éboulis des valeurs propres. — Pour connaître le pourcentage d'information (*i.e.* de variance, ou encore d'inertie) porté par chaque axe (ou composante principale) de l'ACP, afficher l'éboulis des valeurs propres, par exemple *via* le menu Statistiques > Analyse multivariée > Analyse en composantes principales. Sélectionner toutes les variables (sauf la variable `ANNEE`), cocher seulement la case Graphique des éboulis dans l'onglet des Options, puis cliquer sur OK.

Commenter l'allure de ce graphique. Étant donnée l'information portée par les deux premiers axes, paraît-il pertinent de considérer également les axes suivants [2] ?

Question subsidiaire : pourquoi le dernier axe porte-t-il une inertie nulle ?

§3.4. Procéder à l'ACP. — Procéder maintenant à l'ACP proprement dite. Aller dans le menu FactoMineR > PCA, puis sélectionner toutes les variables comme variables actives, sauf la variable `ANNEE`. Indiquer qu'il s'agit d'une variable supplémentaire *via* le bouton Variables quantitatives illustratives. *Via* le bouton Sorties, demander des résultats précis pour les variables actives et les individus actifs. Cliquer sur le bouton Appliquer pour lancer les calculs.

§3.5. Analyser le nuage des variables. — Sur le cercle des corrélations (1,2), les principes de lecture sont les suivants :

- plus une variable possède une *qualité de représentation* élevée dans l'ACP, plus sa flèche est longue ;
- plus deux variables sont corrélées, plus leurs flèches pointent dans la même direction (dans le cercle de corrélation, le coefficient de corrélation est symbolisé par les angles géométriques entre les flèches) ;
- plus une variable est proche d'un axe principal de l'ACP, plus elle est liée à lui. Cette dernière règle permet généralement de donner un sens concret aux axes de l'ACP.

Quelles sont les variables qui semblent bien représentées ? Quelles sont les variables moins bien représentées ?

Quelles sont les variables paraissant assez fortement corrélées à chacun des axes ? Tenter d'expliquer qualitativement l'information portée par chacun des deux premiers axes de l'ACP.

Commenter le positionnement de la variable illustrative (ou supplémentaire) Année. Que peut-on en déduire qualitativement ?

Question subsidiaire : comment expliquer la proximité des variables DEF et DIV (et donc leur corrélation) sur le cercle des corrélations, alors qu'elles ont une corrélation assez faible faible en réalité [3] ?

§3.6. Analyser le nuage des individus. — Regarder à présent le nuage des individus. Quel est son aspect général ? Des groupes semblent-ils se former ?

Regarder (dans la fenêtre d'affichage des résultats numériques) la qualité de représentation [4] des individus. Un individu mal représenté se situe généralement (à tort !) près du centre du repère, et sa spécificité est mal prise en compte par l'ACP, pour les composantes

2. Pour répondre à cette question, on peut utliser ou bien la « règle du coude » ou bien la règle de Kaiser. L'idéal étant bien sûr que les deux résultats concordent.

3. Pour éviter ce problème, on peut demander de n'afficher sur le cercle de corrélation que les variables dont la qualité de représentation est supérieure à un seuil donné, via le bouton `Graphical options` de l'interface graphique de `FactoMineR`.

4. Cette dernière, comprise entre 0 et 1, est l'entrée `cos2` de la liste des résultats s'affchant sur la console.

principales considérées. Peut-on identifier des individus mal représentés ? Des individus qui paraissent proches sur le plan factoriel (1,2) mais qui n'étaient pas forcément très proches dans le nuage initial [5] ?

§3.7. Classification non supervisée. — L'impression visuelle ne suffit pas toujours, et il peut être utile, afin de rendre l'interprétation totalement objective, d'utiliser des méthodes de classification non-supervisée [6] pour attester de l'existence de groupes bien définis et bien homogènes.

Dans l'interface graphique de `FactoMineR`, cliquer sur le bouton Réaliser une classification après l'ACP. Normalement, les réglages par défaut conviennent. Cliquer sur OK.

Dans le dendrogramme qui s'affiche, choisir l'endroit où établir la « coupure » entre vos groupes en déplaçant la barre horizontale et en cliquant pour confirmer. Observer les groupes obtenus.

Peut-on trouver un sens aux groupes qui se forment par la méthode de classification non-supervisée, en regard du contexte historique associé à chacun des groupes ? Donner la typologie de chaque groupe de budgets (pour dissocier par exemple les budgets de période de guerre, les budgets de période de récession, etc.) et une idée de leurs caractéristiques globales.

§3.8. Représenter des groupes déjà connus. — Parfois, un découpage naturel du jeu de données en différents groupes existe avant même l'ACP : dans un contexte d'études archéologiques, cela peut être le sexe des individus, ou bien leur site d'origine, leur espèce, etc. On peut souhaiter voir si ces groupes connus *ex-ante* sont bien séparés sur le plan factoriel.

Directement dans le fichier CSV, ajouter une nouvelle variable `CONTEXTE` qui sera un facteur à 3 modalités : Avant-guerre, Entre-deux-guerres, Trente Glorieuses. Donner le contexte de chaque budget, puis enregistrer le fichier, et le charger à nouveau dans R en répétant l'étape précédemment décrite.

Procéder à nouveau à l'ACP, mais en indiquant que les budgets doivent être coloriés en fonction de leur contexte historique : dans la fenêtre ACP de `FactoMineR`, cliquer sur Variables qualitatives illustratives, et choisir la variable de contexte historique. Enfin, visiter le menu Options graphiques et choisir la variable `CONTEXTE` dans la liste Coloration des individus. Cliquer sur OK, et lancer à nouveau l'ACP.

Que constatez-vous ?

4. Gestion des valeurs manquantes

Les calculs mathématiques nécessaires à la réalisation d'une ACP nécessitent bien sûr de disposer de données complètes. En présence de valeurs manquantes, trois choix sont possibles :

— éliminer purement et simplement les individus contenant des valeurs manquantes : par exemple *via* le menu Données > Jeu de données actif > Éliminer les cas contenant des valeurs manquantes ;

5. Il sera généralement illusoire de pouvoir répondre à cette question dès que le nombre de variables de l'ACP sera trop élevé.

6. En statistique, on distingue usuellement les méthodes de *classification supervisée* qui permettent d'expliquer la nature des différences entre des groupes déjà définis *a priori* par l'utilisateur, et les méthodes de classification *non-supervisée* qui procèdent « à l'aveugle » et permettent de savoir si des regroupements peuvent se former parmi un ensemble d'individus donnés.

— remplacer chaque donnée manquante par la moyenne globale obtenue sur tous les individus renseignés pour la variable considérée (c'est le choix par défaut sous R, effectué automatiquement) ;
— si les individus appartiennent à des groupes définis *ex-ante* (sexe, site, ...), remplacer chaque donnée manquante par la moyenne obtenue sur tous les individus renseignés *d'un même groupe* pour la variable considérée. Cette option, n'étant pas disponible sous R Commander, est à développer en lignes de commandes.

5. Le problème de l'effet taille

§5.1. Définition. — L'effet taille est le nom donné à un artefact mathématique très fréquemment (si ce n'est systématiquement) rencontré en analyse en composantes principales sur des données biologiques. Lorsque les variables sont toutes positivement corrélées entre elles (ce qui est le cas dans n'importe quelle étude anthropométrique !), alors l'effet taille se manifeste sur le cercle des corrélations par une situation analogue à celle de la Figure 1.

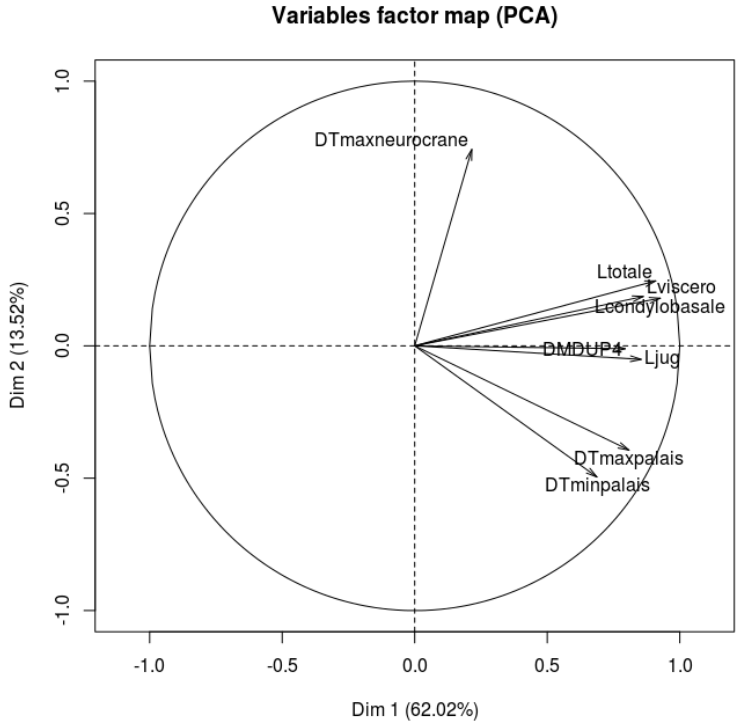

FIGURE 1 – Effet taille en ACP (cas d'une étude métrique)

Dans un tel cas, l'ACP effectue un tri très simple sur l'axe 1 : les « petits » sont à gauche, les « grands » sont à droite. Cela peut parfois avoir un intérêt, mais il s'agit en général d'un effet très indésirable, dont les études métriques cherchent à s'affranchir.

Le principe est alors de ramener tous les individus à la même taille, afin de n'observer sur l'ACP que des différences de forme. Cela permet par exemple de comparer adultes et immatures, hommes et femmes, ou bien sapiens modernes et archaïques au sein d'une même ACP, seulement du point de vue de leur conformation géométrique.

On utilise alors une méthode dite de *double-centrage*, dont [CFVM99] fournit un exemple au cours d'une étude sur la morphologie des carpes — on pourra se référer à [Mey13] pour un exemple d'application en anthropologie biologique. Cette méthode se retrouve également parfois sous le nom de *log-shape ratio* dans la littérature anglophone [DM85].

§5.2. Éliminer l'effet taille avec un tableur. — Cette manipulation peut se réaliser avec un tableur, de préférence libre et gratuit tels que LibreOffice Calc ou Gnumeric. La procédure est la suivante :

(i) Log-transformation des données : si le tableau de données initial est constitué des variables $X_1, X_2, \cdots, X_p$, créer un nouveau tableau de données constitué des variables $\log(X_1), \log(X_2), \cdots, \log(X_p)$.

(ii) Pour chaque individu, calculer la moyenne qu'il obtient sur l'ensemble des variables log-transformées[7]. On considère que ce score moyen constitue une bonne idée de sa « taille ».

(iii) Enfin, pour chaque individu, retrancher à chaque variable la taille moyenne.

§5.3. Éliminer l'effet taille avec R. — Ces manipulations sont bien plus aisées à réaliser avec R, mais doivent impérativement s'effectuer en lignes de commande, ce qui nécessite de connaître quelques rudiments de langage R...

Pour coder le moins possible, on peut par exemple regarder du côté de la fonction `bicenter.wt` du package `ade4` — taper `help(bicenter.wt)` dans une console R pour plus d'informations.

Voici un court exemple issu de données de [Mey13] : il s'agit d'une ACP sur des mesures pelviennes, intégrant 42 humains modernes avec un sex-ratio parfaitement équilibré, et deux néandertaliens. Les données sont disponible à l'adresse `http://tinyurl.com/DataDcACP`.

Effectuer une ACP « classique » avec R et la commenter : le résultat obtenu est-il probant, intéressant ? La série de commandes suivante permet d'effectuer une ACP avec double centrage : observer la différence avec l'ACP « classique ».

```
# Chargement des données et affichage d'un résumé :
> library(FactoMineR)
> DataACP = read.csv("http://tinyurl.com/DataDcACP", header=TRUE,
+ row.names=1, sep=";", dec=",")
> summary(DataACP)

# Centrage en lignes du tableau de données :
> Tab = log(DataACP[ ,-1]) # Tab = seulement les données métriques
> Tab = t(scale(t(Tab), center=TRUE, scale=FALSE)) # centrage en lignes
> Tab = data.frame(DataACP[,1], Tab)
> colnames(Tab)[1] = "Sexe"

# Procéder à l'ACP :
> res.PCA = PCA(Tab, quali.sup=1, graph=FALSE)
```

7. Attention : cela suppose qu'il n'y ait pas de données manquantes dans le tableau, ou qu'elles aient été préalablement estimées !

```
> plot.PCA(res.PCA, habillage=1)
> plot.PCA(res.PCA, choix="var")
```

Références

[Bac10] A. Baccini : Statistique multidimensionnelle (pour les nuls). Cours en ligne, disponible à l'adresse `http://www.math.univ-toulouse.fr/~baccini/zpedago/asdm.pdf`, 2010.

[C+12] P.-A. Cornillon *et al.* : *Statistiques avec R.* Presses Universitaires de Rennes, 3e édition, 2012.

[CFVM99] C. Cibert, Y. Fermon, D. Vallod et F. J. Meunier : Morphological screening of carp *Cyprinus carpio* : relationship between morphology and fillet yield. *Aquatic Living Resources*, 12(1), 1999.

[DM85] J. N. Darroch et J. E. Mosimann : Canonical and Principal Components of Shape. *Biometrika*, 72(2):241–252, Aug. 1985.

[EP08] B. Escofier et J. Pagès : *Analyses factorielles simples et multiples : Objectifs, méthodes et interprétation.* Dunod, 4e édition, 2008.

[HLP09] F. Husson, S. Lê et J. Pagès : *Analyse de données avec R.* Presses Universitaires de Rennes, 2009.

[Mey13] V. Meyer : *Apport de la reconstruction virtuelle du bassin Regourdou 1 (Dordogne, France) à la connaissance de l'obstétrique néandertalienne.* Thèse de doctorat, Université de Bordeaux, 2013.

[Pag10] J. Pagès : *Statistiques générales pour utilisateurs.* Presses Universitaires de Rennes, 2e édition, 2010.

[Sap11] G. Saporta : *Probabilités, Analyse des données et Statistique.* Technip, 3e édition, 2011.