FactoMineR est un *paquet* R qui implémente les principales méthodes d'analyse de données. Si le paquet n'est pas déjà installé sur le système sur le lequel vous travaillez, vous pouvez l'installer en lançant, sous R, la commande : install.packages("FactoMineR"). Il faut ensuite le charger à chaque utilisation avec la commande library(FactoMineR).

Une documentation de FactoMineR se trouve à l'adresse factominer.free.fr. Le paquet est aussi décrit dans le livre suivant:

> Analyse de données avec R. François Husson, Sébastien Lê et Jérôme Pagès. *Presses Universitaires de Rennes*, 2009.

Analyse en composantes principales

La fonction PCA réalise une ACP sur un tableau de données (un *data.frame*) préalablement chargé dans R : elle retourne un objet qui contient entre autres les coordonnées des individus sur les nouveaux axes, les corrélations entre ces axes et les variables initiales, les qualités de représentations, les valeurs propres,...

Exercice 1 On étudie les résultats d'une enquête réalisée par l'INSEE en 2006 sur la répartition des dépenses des ménages français entre différents postes de consommation. Le tableau base_conso.csv, disponible à l'adresse http://factominer.free.fr/livre/, donne les dépenses moyennes sur certains postes de consommation pour certaines catégories de foyers : d'abord des classes définies par l'âge de la personne de référence, ensuite la dépense moyenne sur chaque poste pour l'ensemble des foyers français, enfin pour les déciles définis par les revenus des foyers : D1 est la classe des 10% de foyers ayant les revenus les plus faibles, alors que D10 est la classe des 10% de foyers ayant les revenus les plus élevés. Pour chacune de ces classes de foyers (les « individus ») le tableau donne les valeurs de 30 variables quantitatives : 26 correspondent à des postes de dépenses, 3 à des totaux partiels, et une est la dépense totale.

Question 1.1 Télécharger la table, l'inspecter, et la charger sous R. Vérifier avec summary et head.

Question 1.2 Quels individus et quelles variables vont être utilisés pour le calcul des axes de projection ?

On lance la fonction de FactoMineR qui réalise l'ACP avec les commandes suivantes :

$$\text{conso} \leftarrow \text{read.table}(\ldots) \;;\; \text{conso.pca} \leftarrow \text{PCA}(\text{conso}, \text{quanti.sup} = 27:30, \text{ind.sup} = 8:18)$$

Question 1.3 À quoi servent les paramètres quali.sup et quanti.sup dans l'appel à la fonction PCA ?

Inspecter l'ensemble des attributs de l'objet conso.pca (en tapant tout simplement le nom de l'objet),

Question 1.4 Quels sont les individus et les variables qui contribuent le plus à la formation de l'axe 1 ? Comment peut-on interpréter l'axe 1 ? De même, que peut-on dire de l'axe 2 ?

Question 1.5 Réaliser un diagramme affichant les valeurs propres, ainsi qu'un diagramme affichant leurs sommes cumulées.

Question 1.6 Afficher les projections des individus et des variables sur les plans 2 et 3. D'après les contributions à la formation de l'axe 3, que peut-on dire de cet axe ?

Question 1.7 Que peut-on dire de l'axe 4 ?

Question 1.8 Comparer avec une ACP non normée (paramètre scale.unit de la fonction PCA).

Analyse factorielle des correspondances

Exercice 2 Le ficher regionales04_res_idf.txt donne les résultats des élections régionales 2004 en Ile-de-France, par départements et par candidats.

Question 2.1 Calculer les matrices des profils-lignes et des profils-colonnes, et les écarts à l'indépendance. Quelles cases du tableau portent le plus d'écart à l'indépendance ?

Question 2.2 Calculer les distances du χ^2 entre les départements, puis entre les candidats. Que peut-on en dire ?

La fonction CA de FactoMineR permet de réaliser une analyse factorielle d'une telle table de contingences.

Question 2.3 En observant les projections sur le plan défini par les axes 1 et 2, ainsi que les contributions et les qualités de représentations, quels corrélations entre les départements et candidats font apparaître chacun de ces axes ? Interprétez ces résultats au vu d'une carte des départements de l'Ile de France (par exemple à l'adresse www.iledefrance.fr/territoire/carte-identite).

Question 2.4 Réaliser des diagrammes des valeurs propres.

Question 2.5 Quelles informations apportent les projections sur les axes 3 et 4 ?

Exercice 3 Le tableau maladies.txt donne les résultats d'une enquête au cours de laquelle on a demandé à des médecins quels médicaments ils prescrivent pour traiter certaines maladies. Un élément T_{ij} du tableau indique donc le nombre de médecins qui prescrivent le médicament i pour la maladie j – sachant qu'un médecin peut prescrire plusieurs médicaments pour une même maladie (par exemple en fonction des caractéristiques des patients).

7 maladies ont été retenues : Typhoïde (TFD), Salmonellose digestive (SAL), Affection ORL (ORL), Pneumopathie (PNE), Méningite (MEN), Affection des voies urinaires (URI), Staphylococcie (STA), et 6 médicaments : Péniciline (peni), Tifomycine (tifo), Tétracyline (tetr), Erythromycine (eryt), Tiophénicol (tiop), Gentalline (gent).

Question 3.1 Le tableau précédent est-il une table de contingence ?

Question 3.2 Combien y a-t-il au minimum d'individus dans l'enquête statistique ? (Qui sont les individus ?)

Question 3.3 Rappeler les définitions des matrices d'individus utilisés pour l'analyse des profils-lignes et des profils-colonnes, ainsi que celle de la matrice à diagonaliser.

Question 3.4 Combien d'axes retient-on pour l'analyse ? (Quel pourcentage d'inertie expliquent-ils ?)

Question 3.5 Les maladies/médicaments sont-ils bien représentés par le plan formé par les deux premiers axes ?

Question 3.6 Décrire les contributions des maladies/médicaments à la formation des deux premiers axes.

Question 3.7 Interpréter graphiquement les résultats de l'afc sur les deux premiers axes.

Exercice 4 Cette étude de cas est extraite de l'ouvrage de Jean-Pierre Benzécri "Analyse des données. Tome 2: l'analyse des correspondances". Parmi les questions, certaines ne concernent pas spécifiquement l'analyse statistique ; elles sont destinées à faciliter l'interprétation des résultats de l'AFC. Une enquête a été effectuée auprès de cent fumeurs afin de choisir les noms de deux nouvelles marques de cigarettes. La première marque est destinée à une clientèle masculine : l'homme ciblé est un connaisseur distingué, raffiné mais viril, de niveau socio-économique élevé. La seconde symbolise un public féminin, élégant, assuré, dynamique. Douze marques ont été retenues : Orly (Orl), Alezan (Ale), Corsaire (Cor), Directoire (Dir), Ducat (Duc), Fontenoy (Fon), Icare (Ica), Zodiac (Zod), Pavois (Pav), Cocker (Coc), Escale (Esc), Hôtesse (Hot).

Pour évaluer leur image auprès du public, onze attributs ont été proposés aux fumeurs : vieillot - désuet (VD), nouveau riche (NR), sobre, élégant (SE), cocasse - ridicule (CR), racé (RA), mièvre (MI), distingué (DI), vulgaire - commun (VC), pour un homme (HO), pour une femme (FE), pour une petite nature (PN).

Les données se trouvent dans le fichier `fume.txt`.

Question 4.1 Le tableau de données est-il une table de contingence ? Quelles sont les marques les plus fréquemment citées ?

Question 4.2 Les publics visés ont-ils des points communs ? Quels sont les attributs correspondants ? Pourquoi a-t-on proposé des attributs ne leur correspondant pas ?

Question 4.3 Regarder les distances du χ^2 entre les marques. Quelles sont les marques dont les distances sont les plus faibles ? Les plus élevées ? Comment interpréter ces résultats ?

Question 4.4 En examinant les valeurs propres et leur diagramme, déterminer les axes principaux qu'il faut garder dans les analyses.

Question 4.5 A l'aide du plan 1x2, expliquer la différence entre les marques les plus distantes les unes des autres, et la ressemblance entre les marques les plus proches. Quelles sont les marques bien représentées sur le plan 1x2 ? Quelles sont les marques importantes dans l'interprétation des axes 1 et 2 ?

Question 4.6 Que peut-on dire des attributs ? De quel côté de l'axe 1 se trouvent les marques qui correspondent aux publics visés ?

Question 4.7 Quelle interprétation peut-on proposer à l'attribut "Pour une femme" tel qu'il apparaît le long de l'axe 2 ? De l'axe 4 ?

Question 4.8 Que peut-on dire de l'axe 5 ?

Analyse des correspondances multiples

On parle d'« analyse des correspondances multiples » lorsqu'on analyse un tableau T donnant les valeurs de Q variables nominales pour I individus. L'ACM est une AFC réalisée sur un tableau dérivé d'un tel tableau de facteurs. On peut réaliser l'ACM soit sur un tableau disjonctif complet, soit sur un tableau de Burt.

Tableau disjonctif complet C'est un tableau qui a une colonne pour chaque modalité de chaque variable : si $T_{iq} = j$, alors dans le tableau disjonctif D associé à T, la colonne correspondant à la modalité j de la variable q contiendra 1 pour l'individu i, sinon elle contiendra 0. Formellement, si j est une modalité de la variable q, alors $D_{ij} = 1$ si $T_{iq} = j$, 0 sinon.

Tableau de Burt C'est le tableau de toutes les tables de contingence de toutes les paires de variables de la table de facteurs initiale. Si j est une modalité de la variable q, et si j' est une modalité de la variable q', le tableau de Burt B correspondant au tableau de facteurs T contient, à l'intersection de la ligne j et de la colonne j', le nombre d'individus qui ont ces deux modalités pour les variables q et q'. B est donc une matrice symétrique, à M lignes et M colonnes, où M est le nombre total de modalités. Si n_j est le nombre d'individus ayant la modalité j, alors la somme des termes de la ligne j de B vaut Qn_j.
On vérifie facilement que $B = D'D$.

Exercice 5 Le fichier pbio.txt[1] contient les réponses de 419 personnes à une enquête réalisée dans des supermarchés angevins et parisiens entre 1996 et 1998 dans le but de connaitre l'avis de consommateurs quant aux produits biologiques et aux produits diététiques – le questionnaire et les codes des réponses aux 11 réponses sont indiqués à la fin du fichier.

Question 5.1 Importer dans R la table du fichier pbio.txt, et inspecter la structure de la table à l'aide des commandes `summary` et `head`. L'entier 0 correspond, dans cette table, à une réponse « ne sait pas ». Pour faciliter certains traitements suivants, on remplacera cette valeur 0 par une autre valeur entière > 0 et qui n'apparaît pas déjà, par exemple 10 :
$T \leftarrow \mathsf{lapply}(\mathsf{pbio}, \mathsf{function}(c)\{\mathsf{sapply}(c, \mathsf{function}(x)\{\mathsf{if}(x == "0")\{10\}\mathsf{else}\{x\}\})\})$

Question 5.2 Calculer avec la fonction tab.disjonctif de FactoMineR le tableau disjonctif correspondant, puis le tableau de Burt.

Question 5.3 Vérifier que les AFC sur ces deux tableaux donnent bien des résultats identiques.

Question 5.4 La fonction MCA permet de réaliser directement l'analyse des correspondances multiples sur le tableau d'origine. Réaliser l'ACM sur pbio avec la fonction MCA. À quoi correspond le point SITPROF_0 sur le graphique faisant apparaître les modalités des variables ? Comparer les résultats obtenus ainsi avec ceux obtenus précédemment.

Question 5.5 Faire un graphique des valeurs propres. Qu'observe-t-on ?

[1] Trouvé sur le site web de Gilles Henault à l'Université d'Angers : `http://forge.info.univ-angers.fr/ gh/Datasets/datasets.htm`

Classification ascendante hiérarchique

On parle de classification quand on cherche à répartir des objets dans des ensembles homogènes d'un certain point de vue. On cherche donc à construire une partition d'un ensemble E en K parties disjointes $E_1, \ldots, E_K$ dont l'union est E tout entier. On se base sur une notion de distance, ou mesure de dissimilarité, sur l'ensemble des objets : on voudrait alors que tous les éléments d'une classe E_i soient proches les uns des autres, et que deux éléments appartenant à deux classes distinctes soient plus éloignés l'un de l'autre.

La classification ascendante hiérarchique est une méthode de classification qui consiste, partant des p classes élémentaires constituée chacune d'un unique objet, à augmenter petit à petit la taille des classes en opérant des unions entre deux classes « proches ». Cela fait intervenir une seconde notion de distance, entre ensembles d'objets cette fois-ci, nous l'appelerons *indice d'agrégation*.

Si on note $d(i, i')$ la distance entre deux objets i et i', des indices d'agrégation possibles entre deux sous-ensemble F et F' de E sont:

saut minimum: (ou single pour R) $D(E, F) = \min_{i \in F, i' \in F'} d(i, i')$. Propriété : calculs simples, risque d'« effet de chaîne »

Ward: (ou ward pour R) Propriété : à chaque étape, on effectue le regroupement qui augmente le moins possible l'inertie (somme des carrés des distances entre points) à l'« intérieur » des classes (l'inertie totale étant constante, on diminue le moins possible l'inertie inter-classes).

D'autres indices d'aggrégation sont possibles.

Classification ascendante hiérarchique avec R La commande dist de R permet de calculer la distance euclidienne entre les lignes d'une matrice. Le résultat de la commande dist peut être passé à la commande hclust, qui réalise une classification ascendante hiérarchique. L'un des paramètre de hclust est l'indice d'agrégation à utiliser. La fonction plclust permet d'afficher le dendogramme résultant. La fonction rect.hclust répartit les objets classés hiérarchiquement en un nombre fixe de classes, et affiche le résultat sur le dendogramme dessiné avec plclust. Par exemple :

$H = \mathsf{hclust}(\mathsf{dist}(T_s), \mathsf{method} = \text{"ward"})$
$\mathsf{plclust}(H, \mathsf{hang} = -0.1, \mathsf{axes} = \mathsf{FALSE}, \mathsf{ann} = \mathsf{FALSE})$
$R = \mathsf{rect.hclust}(H, \mathsf{k} = 2)$

Exercice 6 Reprendre le script réalistant une ACP sur la table onu67_budget_temps.txt, et réaliser une CAH sur les individus, en utilisant la distance euclidienne. Répartir les individus en trois classes, et visualiser ces 3 classes en projetant les individus sur les deux premiers axes factoriels.

Exercice 7 Effectuer maintenant deux CAH sur les candidats et les départements selon les résultats des élections régionales en Ile-de-France en 2004. La distance entre les lignes ou les colonnes de la table de contingence est la distancee du χ^2, mais on se ramène à une distance euclidienne en faisant la même transformation que dans le cas d'une AFC. Comparer aux résultats de l'AFC.