Analyse Factorielle des Correspondances : AFC
Analyse des Correspondances Multiples : ACM

L'AFC permet d'analyser la liaison existant entre deux variables qualitatives. Si on dispose de plus de deux variables qualitatives, on aura recours à l'Analyse des Correspondances Multiples (ACM).
Pour plus de détails sur l'AFC et l'ACM que ce qui est présenté ici, voir le site de *François Husson* (cours-vidéos).

Les données : soient X et Y deux variables aléatoires qualitatives prenant respectivement les modalités $\{x_1, \ldots, x_I\}$ et $\{y_1, \ldots, y_J\}$. On les observe simultanément sur n individus. Les données sont présentées sous la forme d'une table de contingence.

	y_1	$\ldots$	y_j	$\ldots$	y_J	Total
x_1	n_{11}	$\ldots$	n_{1j}	$\ldots$	n_{1J}	$n_{1\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	n_{i1}	$\ldots$	n_{ij}	$\ldots$	n_{iJ}	$n_{i\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_I	n_{I1}	$\ldots$	n_{Ij}	$\ldots$	n_{IJ}	$n_{I\bullet}$
Total	$n_{\bullet 1}$	$\ldots$	$n_{\bullet j}$	$\ldots$	$n_{\bullet J}$	n

Exemple : Résultat du premier tour des élections présidentielles de 2007. Nombre de voix obtenus par les 12 candidats dans chacune des 23 régions françaises.

```
> head(data)
          Sarkozy Bayrou  Royal Le.Pen Besanc. Villiers Voynet Laguiller  Bove
Alsace     362391 214259 171282 135730   33310    22492  20382     13821 13758
Aquitaine  532127 417546 557300 168664   78230    34028  28285     22046 27941
Auvergne   238152 169395 225477  78704   41522    18730  12090     12936 13532
Bourgogne  297544 175213 241094 119041   42246    24971  13690     14440 12296
Bretagne   557507 451988 564100 143926   94205    41212  39026     25662 28484
Centre     460425 278175 345352 168912   65347    45720  22655     22279 17395
          Buffet Nihous Schivardi
Alsace      6100   5142      2522
Aquitaine  41791  35300      7572
Auvergne   21920  12474      4207
Bourgogne  18154  12079      3608
Bretagne   31860  21207      5169
Centre     30003  20567      5696
```

Avant de mettre en oeuvre une AFC, il faut s'assurer que cette liaison existe bien. Pour cela, on peut utiliser le test du khi-deux d'indépendance. L'AFC est ensuite un raffinement du test du khi-deux visant à visualiser la nature de la liaison entre les deux variables i.e à apporter des réponses à des questions du type :

- Y a-t-il des lignes du tableau (modalités de X ie régions) qui se "ressemblent", ie telles que les distributions des modalités de Y soient analogues ? Y a-t-il des lignes du tableau qui s'opposent, ie telles que les distributions des modalités de Y soient très différentes.
- Idem en terme des colonnes du tableau (modalités de Y ie candidats).
- Y a-t-il des associations modalité de X - modalité de Y qui s'attirent (effectif conjoint particulièrement élevé, ie une région dans laquelle on vote majoritairement pour un candidat) ou qui se repoussent (effectif conjoint particulièrement faible) ?

Notations :

- Les marges du tableau (totaux en lignes et en colonnes)sont définies par :

$$n_{i\bullet} = \sum_{j=1}^{J} n_{ij}; \quad n_{\bullet j} = \sum_{i=1}^{I} n_{ij}$$

- On peut tout définir en terme de fréquences :

$$f_{ij} = \frac{n_{ij}}{n}; \quad f_{i\bullet} = \sum_{j=1}^{J} f_{ij} = \frac{n_{i\bullet}}{n}; \quad f_{\bullet j} = \sum_{i=1}^{I} f_{ij} = \frac{n_{\bullet j}}{n}$$

- On appelle i-ème profil ligne :

$$\{\frac{f_{i1}}{f_{i\bullet}}, \ldots, \frac{f_{iJ}}{f_{i\bullet}}\}$$

- On appelle j-ème profil-colonne :

$$\{\frac{f_{1j}}{f_{\bullet j}}, \ldots, \frac{f_{Ij}}{f_{\bullet j}}\}$$

Indépendance entre les deux variables :
Il y a indépendance entre deux variables X et Y si $P(X = x_i \text{ et } Y = y_j) = P(X = x_i)P(Y = x_j)$, ie. si

$$\frac{n_{ij}}{n} \approx \frac{n_{i\bullet}}{n}\frac{n_{\bullet j}}{n} \quad \text{i.e} \quad f_{ij} = f_{i\bullet}f_{\bullet j}$$

Autrement dit si

$$\forall i \quad \frac{f_{ij}}{f_{i\bullet}} = f_{\bullet j} \quad \text{ou de façon symétrique} \quad \forall j \quad \frac{f_{ij}}{f_{\bullet j}} = f_{i\bullet}$$

Le test du Khi-deux :
H_0 : indépendance entre X et Y. La statistique de test mesure l'écart à l'indépendance :

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - \frac{n_{i\bullet}n_{\bullet j}}{n})^2}{\frac{n_{i\bullet}n_{\bullet j}}{n}} = n \times \sum_i \sum_j \frac{(f_{ij} - f_{i\bullet}f_{\bullet j})^2}{f_{i\bullet}f_{\bullet j}}$$

D'où vient l'écart à l'indépendance ?
Le principe de l'AFC va être de comparer les I profils ligne au profil ligne moyen et idem pour les colonnes. Cette analyse va se faire au moyen de deux ACP un peu particulières.

1. On réalise l'ACP du tableau des profils-lignes (les individus de cette ACP sont les lignes de la table de contingence, ie les modalités de X) et l'on fait la représentation graphique des individus (dans cette ACP particulière, on ne s'intéresse pas au graphique des variables).
2. On réalise de même l'ACP du tableau des profils-colonnes (les individus de cette ACP sont maintenant les colonnes de la table de contingence) et l'on fait la représentation graphique des individus.
3. On peut montrer que ces deux ACP se correspondent (ce qui est normal, puisque leurs données sont extraites de la même table de contingence) et on peut donc superposer les deux représentations graphiques. C'est l'interprétation graphique qui permet d'étudier la liaison entre les deux variables.

Ce qui change par rapport à l'ACP classique :
Considérons l'ACP du nuage des profils lignes. C'est similaire pour l'ACP du nuage des profils colonnes.

— Les I profils lignes sont dans l'espace $\mathbb{R}^J$, chaque ligne a pour poids $f_{i\bullet}$ (en ACP poids $1/n$ pour chaque individu). Le centre de gravité G_I correspond à

$$\sum_i f_{i\bullet}\frac{f_{ij}}{f_{i\bullet}} = f_{\bullet j}$$

i.e au profil ligne moyen. Le centre de gravité va représenter la situation d'indépendance entre les deux variables.

— La distance entre deux profils ligne i et i' est donnée par :

$$d^2_{\chi^2}(i,i') = \sum_{j=1}^{J} \frac{1}{f_{\bullet j}} (\frac{f_{ij}}{f_{i\bullet}} - \frac{f_{i'j}}{f_{i'\bullet}})^2$$

Ce n'est pas la distance euclidienne usuelle (comme en ACP), mais la distance dite "du khi-deux". Pourquoi cette métrique plutôt que la distance euclienne ?

- La pondération par $\frac{1}{f_{\bullet j}}$ revient à donner des importances comparables aux colonnes. Ainsi dans notre exemple cela évite de donner trop de poids aux "gros" candidats.
- De plus la métrique du χ^2 possède la propriété dite d'équivalence distributionnelle : si on regroupe deux colonnes qui ont les mêmes profils, alors la distance entre les profils-lignes n'est pas modifiée. C'est une notion particulièrement importante pour l'analyse de tableaux lexicaux dans lesquelles on compte le nombre d'occurences du mot j dans un texte i. Les textes correspondent aux lignes, et les mots aux colonnes. Se pose la question de regrouper le même mot lorsqu'il apparait au singulier ou au pluriel. La réponse est que si le mot au pluriel et au singulier ont les mêmes profils, alors peu importe. On peut les regouper ou non. Par contre s'ils ont de profils différents, alors il ne faut pas les regrouper parce que cela signifie qu'ils correspondent à des notions différentes.

— La distance entre le profil-ligne i et le centre de gravité G_I est donnée par :

$$d^2_{\chi^2}(i,G_I) = \sum_{j=1}^{J} \frac{1}{f_{\bullet j}} (\frac{f_{ij}}{f_{i\bullet}} - f_{\bullet j})^2$$

La distance d'un point avec le centre de gravité traduit l'écart à l'indépendance : si un point est très proche de l'origine, alors le profil correspondant est similaire au profil moyen. Au contraire si un point est éloigné du centre de gravité, alors le profil s'écarte du modèle d'indépendance.

— L'inertie est donnée par :

$$I_T = \sum_i \sum_j \frac{(f_{ij} - f_{i\bullet} f_{\bullet j})^2}{f_{i\bullet} f_{\bullet j}} = \frac{\chi^2}{n}$$

où χ^2 est la statistique du test du Khi-deux. La méthode AFC revient à trouver une décomposition pertinente du I_T selon plusieurs axes factoriels.

On peut montrer qu'il y a dualité entre les deux nuages profils-lignes et profils-colonnes. On peut donc (contrairement à l'ACP) utiliser une représentation simultanée en faisant figurer sur un même graphique les individus lignes et les individus colonnes. On utilise pour cela les relations barycentriques qui permettent de calculer la coordonnée sur l'axe s du j-ème profil colonne à partir des coordonnées sur l'axe s des I profils ligne :

$$coord_s(j) = \frac{1}{\sqrt{\lambda_s}} \sum_{i=1}^{I} \frac{f_{ij}}{f_{\bullet j}} coord_s(i)$$

La coordonnée sur l'axe s du j-ème profil colonne est donc au barycentre des coordonnées sur l'axe s des I profils lignes, ce barycentre étant pondéré par les $\frac{f_{ij}}{f_{\bullet j}}$, avec un coefficient de dilatation de $\frac{1}{\sqrt{\lambda_s}}$ (λ_s est la valeur propre). Autrement dit sur le graphe, une modalité colonne j "se trouve près" des profil lignes i pour lesquelles $\frac{f_{ij}}{f_{\bullet j}}$ sont les plus élevées.

Interpréation du graphique simultané :
De même qu'en ACP, pour analyser les proximités entre les points sur les graphiques factoriels, on s'intéresse surtout aux points bien projetés (ayant un cos2 élevé).
La distance d'un point à l'origine (centre de gravité) représente son écart à l'indépendance, ie au profil moyen.
La proximité entre modalités d'une même variable indique que les distributions sont voisines.
On peut également interpréter la proximité entre modalités des deux variables : si un point-ligne et un point-colonne sont proches, cela indique une attirance entre ces modalités, ie un effectif conjoint particulier élévé par rapport à la moyenne qui contribue fortement de manière positive à la dépendance entre ces deux variables.

Exercice 1. AFC sur les résultats du premier tour des élections présidentielles de 2007.

1. Importer les données dans R, avec comme nom de ligne les candidats et comme nom de colonnes les régions. Vérifier qu'elles sont bien chargées.
2. Statistique descriptive :
   — représenter la part de vote de chaque candidat sous forme d'un "camenbert".
   — Comparer les régions en terme d'électeurs. Quelles sont les régions qui comptent le plus en terme d'électeurs ? Le moins ?
3. Montrer, en utilisant un test du Kdi-deux que les deux variables sont liées.
4. Réaliser l'AFC du tableau de contingence des votes au premier tour des élections présidentielles en utilisant le package "FactoMineR" :

   ```
   library(FactoMineR)
   res=CA(data)
   ```

   (a) Combien d'axes garder pour que l'information soit représentée de manière satisfaisante ?
   (b) Commenter les contributions des différentes candidats aux premiers axes principaux, et interpréter les axes.

   ```
   res$col$contrib
   ```

   Attention : en AFC, les points les plus excentrés sur les axes ne sont pas nécessairement ceux qui contribuent le plus (à cause des poids dans la métrique utilisée).

   (c) Faire de même pour les régions.
5. Modalités supplémentaires. On pourrait choisir de ne pas entrer certaines des modalités dans l'inférence des axes. On pense en particulier à l'outremer ainsi qu'à deux "petits" candidats (Nihous et Schivari). On procède de la façon suivante :

   ```
   res1=CA(data, row.sup=23, col.sup=(11,12))
   ```

   Interpréter cette AFC. Vous pourrez par exemple répondre aux questions suivantes :
   (a) Donner une région où on a particulièrement voté Le Pen. Où on a particulièrement peu voté Le Pen.
   (b) En région PACA pour quel(s) candidat(s) a t'on majoritairement voté ?
   (c) Y a-t-il des régions qui se ressemblent, ie dans lesquels les votes pour les différents candidats sont similaires ? Y a-t-il au contraire des régions qui s'opposent (résultats très différents) ?
   (d) Y a-t-il des candidats dont les résultats se ressemblent : ils n'obtiennent pas nécessairement les mêmes scores, mais les régions où ils obtiennent de bons scores sont les mêmes ? Y a-t-il des candidats dont les résultats s'opposent ?
   (e) Y a-t-il des régions dont les résultats sont proches des résultats nationaux (ie des résultats moyens toutes régions confondues) ?

Plus d'options sur le graphique en AFC (notamment pour ne représenter que les points à forte contribution, ou bien représentés)

`https://rdrr.io/cran/FactoMineR/man/plot.CA.html`

Vous pouvez par exemple essayer les commandes suivantes :

```
plot(res1,selectRow="cos2 0.8", selectCol="cos2 0.8")
plot(res1,selectRow="contrib 2", selectCol="contrib 2")
```

Exercice 2. Lorsqu'il y a plus de deux variables qualitatives, on a recours à l'analyse des correspondances multiples ACM. En particulier, l'ACM est la méthode adaptée aux tableaux de type enquête qui donnent les réponses de I individus à J questions à choix multiples (variables qualitatives), chacune de ces questions ayant K_j modalités ($j = 1, \ldots, J$). On note $K = K_1 + \ldots + K_J$ le nombre total de modalités.

1. Charger le fichier de données *chiens* qui donne les caractéristiques de plusieurs races de chiens. Quel est la structure du fichier de données, sa dimension ? Dans la suite on mettra de côté la dernière colonne du fichier de données (celle donnant la "fonction" du chien) On note :

   ```
   data=chiens[,1:6]
   ```

   Dans le fichier data, qui sont les I individus, les J variables qualitatives ? Combien de modalités K_j par variable ? combien de modalités K au total ?

2. L'ACM revient à faire l'AFC du tableau disjontif complet. Sur le jeu de données data, faire les commandes suivantes et comprendre ce qu'est le tableau disjontif complet. Quelle est sa dimension ?

   ```
   library(FactoMineR)
   tab.disjonctif(data)
   ```

   En ACM on traite le tableau disjontif complet comme une table de contingence. On travaille donc sur les profils ligne qui correspondent aux individus, et sur les profils colonnes qui correspondent aux modalités. Avec le package FactomineR, on demande la réalisation d'une ACM avec la commande MCA. En entrée de cette fonction, il faut donner le tableau initial (de taille $I \times J$, et non le tableau disjontif complet). On peut demander à certaines variables d'être supplémentaires, ie on cherchera à les interpréter, mais elles ne sont pas utilisés pour déterminer les axes.

   ```
   res=MCA(chiens, quali.sup=7)
   ```

   Remarque. Dans la littérature anglo-saxone, l'ACM consite à effectuer une AFC sur "le tableau de Burt" qui est un assemblage de tables de contingence, et non sur le tableau disjontif complet. On peut montrer un lien entre les deux façons de faire.

3. Combien d'axes garder ?
   On peut montrer que le pourcentage maximum d'inertie apporté par un axe est de $(\frac{J}{K-J}) \times 100$. *Ainsi il est fréquent d'avoir des pourcentage d'inertie faible lorsque le nombre de modalités est élevée. Lorsque c'est le cas, on interprétera les axes dont l'inertie est supérieure à* $1/J$.

4. Interprétation : on utilise comme en AFC une distance basée sur le Khi-deux. Sans rentrer dans les détails de l'écriture des distances (voir le site de *François Husson* si vous le souhaitez), on peut faire les remarques pratiques suivantes :
   - Le poids de chaque individu est constant égal à $1/I$, mais les modalités ont un poids d'autant plus grand que la modalité est fréquente.
   - Deux individus seront proches s'ils ont beaucoup de modalités en commun, sachant que l'on donne plus de poids dans cette distance au fait que ces deux individus ont en commun une modalité rare.
   - Deux modalités seront proches si elles sont possédées par les mêmes individus.
   - Représentation simultanée : un individu est "placé" du côté des modalités qu'il possède, et une modalités du côté des individus qui la possède.
   - Les cos^2 seront en général petit tout simplement parce qu'il peut y avoir beaucoup d'axes (au maximum $min(I-1, K-J)$).

   Interpréter les deux premiers axes.

   Pour les races que vous connaissez, vérifier qu'elles sont "bien placées". Pour les autres qu'en déduisez-vous ?

   Interpréter la variable supplémantaire "fonction". Par exemple, à quoi correspond un chien de chasse ? un chien de compagnie ?