

Science des données II : module 7



Analyse Factorielle des Correspondances (AFC)

Philippe Grosjean & Guyliann Engels

Université de Mons, Belgique
Laboratoire d'Écologie numérique des Milieux aquatiques



<http://biodatascience-course.sciviews.org>
sdd@sciviews.org

Analyse Factorielle des Correspondances (AFC)

- En anglais: **Correspondence Analysis**
- Tableau multivarié, données qualitatives ou semi-quantitatives L'ACP ne peut être utilisée
- **Tableau de contingence** à double entrée, ou ...
- Tableau de type **dénombrement espèces – stations**
- Utilisation de la statistique **Chi-carré**

Rappel - Table de contingence et test du χ^2 (1)

- **Table de contingence** : représentation de proportions.

Exemple: autopolinisation de fleurs roses d'*Antirrhinum majus* \Rightarrow on s'attend à obtenir les fleurs suivantes selon la génétique Mendélienne :

H_0 : rouge: 25%, rose: 50%, blanche: 25%

Le résultat est le suivant :

rouge: 54, rose:122, blanche: 58

Soit des probabilités estimées respectives de :

rouge: 23.1%, rose:52.1%, blanche: 24.8%

Comment savoir si ces observations confirment H_0 ?

Rappel - Table de contingence et test du χ^2 (2)

- Tester les **proportions observées** a_i (R: 54, r: 122, B: 58).
- Nombre total de fleurs : $54 + 122 + 58 = 234$.
- Comparaison à un **effectif théorique** α_i :

R: $0.25 \times 234 = 58.5$, r: $0.50 \times 234 = 117$, B: $0.25 \times 234 = 58.5$

- Calcul de la statistique **chi carré** » (χ^2) par :

$$\chi^2 = \sum \frac{(a_i - \alpha_i)^2}{\alpha_i}$$

Cela donne :

$$\chi^2 = (54-58.5)^2/58.5 + (122-117)^2/117 + (58-58.5)^2/58.5 = 0.56$$

- Comparaison de cette statistique à la **distribution théorique du χ^2** pour décider si on rejette H_0 ou pas...

Rappel – Chi² pour un tableau $r \times k$

Le test se **généralise** pour un tableau de contingence $r \times k$:

$$\alpha_i = \frac{\text{total ligne} \cdot \text{total colonne}}{\text{total général}}$$

Le **nombre de ddl** = $(r - 1) \cdot (k - 1)$.

Notez ceci:

$$\chi^2 = \sum \frac{(a_i - \alpha_i)^2}{\alpha_i}$$

- Les termes respectifs du χ^2 pour chaque cellule du tableau **quantifient** l'écart entre les observations et un tableau sous H_0 où toutes les observations sont indépendantes.
- On peut appliquer une **ACP** si on remplace les effectifs observé par leur contribution au χ^2 , puisque l'on obtient alors une variable calculée quantitative.
- Le tableau de contingence peut être traité indifféremment dans les deux sens (**pas de distinction cas et variable**).

Distance Euclidienne *versus* Chi²

Distance euclidienne (au carré) :

$$d^2(i, i') = \sum_{j=1}^p (x_{ij} - x_{i'j})^2$$

Distance du χ^2 :

$$d^2(i, i') = \sum_{j=1}^p \frac{(x_{ij}/x_{i.} - x_{i'j}/x_{i'.})^2}{x_{.j}}$$

où :

- $x_{i.}$ est la somme de la ligne i sur toutes les colonnes,
- $x_{.j}$ est la somme de la colonne j sur toutes les lignes.

On peut faire le même calcul en **inversant les lignes et les colonnes**.

Exemple

voir séance d'exercices

Analyse Factorielle des Correspondances

- Projection des correspondances entre les espèces (ligne) et les stations (colonnes) dans un espace simplifié = idem que l'ACP.
- **Interprétation** : plus les points sont proches les uns des autres, plus ils ont un comportement similaire.
- Comme les lignes et les colonnes ont même rôle dans un tableau de contingence, on calcule selon les deux orientations (=2 ACP) et on superpose sur le même graphique =>
 - **Pour les espèces** : elles sont présentes ou absentes simultanément.
 - **Pour les stations** : composition faunistique similaire.
 - **Espèces *versus* stations** : ces stations sont caractérisées essentiellement par les espèces proches sur le graphe (il y a correspondance entre les deux).

AFC dans R

Calcul simple à partir d'un script R:

```
library(MASS)           # Package contenant la fonction
data(caith)             # Tableau de contingence exemple
caith                   # AFC en une seule commande
chart$biplot(ca(caith))
```

Interprétation du graphique :

- Niveaux de la variable en colonne en rouge,
- Niveaux de la variable en ligne en bleu,
- La distance entre les niveaux d'une variable indique leur **similarité ou différences**,
- Le rapprochement entre les points d'une variable et de l'autre indique la **correspondance** entre eux.