Science des données II : cours 4

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

UMONS

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

UMONS

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

- Table de contingence : représentation de proportions.

Exemple: autopollinisation de fleurs roses d'*Antirrhinum majus* => on s'attend à obtenir les fleurs suivantes selon la génétique Mendélienne :

`H0: 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 H0 ?

UMONS

Rappel - Table de contingence et test du Chi^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 x 234 = 58.5, r: 0.50 x 234 = 117, B: 0.25 x 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 :

χ^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...

UMONS

Rappel – Chi^2 pour un tableau r x k

Le test se généralise pour un tableau de contingence r x k :

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

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

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).

UMONS

Distance Euclidienne *versus* Chi^2

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).

UMONS

AFC dans R

Calcul simple à partir d'un script R:

```
library(MASS)                    # Package contenant la fonction
data(caith)
caith                            # Tableau de contingence exemple
plot(corresp(caith, nf = 2))     # AFC en une seule commande
```

Interprétation du graphique :

- Niveaux de la variable en colonne en rouge,
- Niveaux de la variable en ligne en noir,
- 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.

UMONS

Démonstration de l'AFM

L'**analyse factorielle multiple** (AFM ou MFA en anglais) permet d'analyser plusieurs sets de variables quantitatives (numériques) ou qualitatives (facteurs) mesurées sur les mêmes individus **en une seule analyse synthétique.**

- Pour les sous-tableaux numériques, c'est une ACP qui est réalisée
- Pour les sous-tableaux qualitatifs, c'est une ACM (analyse des correspondances multiples) qui est réalisée. Elle est assez proche de l'AFC, mais n'analyse le tableau que dans un sens.

Pour une démonstration de l'AFM, voyez la vidéo ici :
https://www.dailymotion.com/video/x13gva9

Pour aller plus loin...

Une version plus détaillée de l'AFM est présentée ici :
http://www.sthda.com/french/articles/38-methodes-des-composantes-principales-dans-r-guide-pratique/77-afm-analyse-factorielle-multiple-avec-r-l-essentiel/

UMONS