Science des données II : cours 1

Introduction & matrices de distances

Philippe Grosjean & Guyliann Engels

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

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

UMONS

Introduction

- Au cours de **biostatistique et probabilités** de Bab2, nous avons abordé :
  - Les statistiques descriptives, qui résument à l'aide de descripteurs ou présentent de manière visuelle (graphique) le contenu de jeux de données,
  - Les statistiques inférentielles, basées sur les tests d'hypothèse pour répondre à des questions « binaires » (est-ce ceci –H_0–, ou son contraire –H_1– ?)

- Dans ce cours, nous aborderons l'analyse des données. Elle permet d'**explorer** des gros jeux de données multivariés et d'y découvrir des **structures** ou des **associations** entre individus. Nous étudierons également la modélisation avec la régression linéaire.

UMONS

Analyse de données

L'analyse de données ou multivariée se subdivise en deux groupes de techniques complémentaires :

- Les méthodes de **classification**. Ces méthodes regroupent les objets en plusieurs sous-ensembles distincts (**groupes**), chaque individu étant plus semblable aux autres au sein de son propre sous-ensemble qu'il ne l'est par rapport aux individus des autres sous-ensembles.
- Les méthodes d'**ordination**. Elles représentent les individus sur des graphes bi- ou tridimensionnels (ou dit, une représentation sur des **cartes**, par analogie à la géographie). La distance des individus entre eux est relative à leur degré de ressemblance ou de différence (on dit : **similarité** ou **dissimilarité**).

UMONS

Analyse de données

L'analyse de données ou multivariée se subdivise en deux groupes de techniques complémentaires :

- Les méthodes de **classification**. Ces méthodes regroupent les objets en plusieurs sous-ensembles distincts (**groupes**), chaque individu étant plus semblable aux autres au sein de son propre sous-ensemble qu'il ne l'est par rapport aux individus des autres sous-ensembles.
- Les méthodes d'**ordination**. Elles représentent les individus sur des graphes bi- ou tridimensionnels (ou dit, une représentation sur des **cartes**, par analogie à la géographie). La distance des individus entre eux est relative à leur degré de ressemblance ou de différence (on dit : **similarité** ou **dissimilarité**).

UMONS

Exemples de jeux de données traités

- Le dénombrement (quasi-)exhaustif des espèces rencontrées en différents lieux (stations).
- L'abondance (nombre d'individus) dans chaque taxon en différentes stations échantillonnées.
- Une étude détaillée des conditions du milieu (paramètres physico-chimiques, terrain, etc.) où vivent les animaux étudiés en différentes stations.
- Un tableau rassemblant la biométrie (taille, poids, poids de certains organes, ...) d'un nombre important d'individus d'une population donnée, mesurés afin de déterminer l'évolution de ces mesures biométriques au cours de la croissance.
- Etc.

UMONS

Références

Pas (encore) de syllabus ; références conseillées :

- **Venables W.N. & B.D. Ripley, 2002**. Modern applied statistics with S-PLUS (4th ed.). Springer, New York, 495 pp.

- **Legendre, P. & L. Legendre, 1998**. Numerical ecology (2nd English ed.). Elsevier, Amsterdam, 853 pp.

UMONS

Matrice de distances

UMONS

Matrice de distances

- La matrice de distances est le point de départ (première étape explicite ou implicite) de nombreuses analyses multivariées.
- Présentation des individus ou des variables aussi bien en ligne qu'en colonne => **deux points de vue possibles**.
- Les éléments de la matrice correspondent à **toutes les paires possibles, prises deux à deux**; on obtient une **matrice carrée**.
- Exemples de matrices de distances déjà abordées :
  - Matrice de **variances/covariances** = distances euclidiennes au carré.
  - Matrice de **corrélation** = distances euclidiennes au carré sur des données standardisées.

UMONS

Indice de similarité/dissimilarité

- Un indice de similarité (similarity index) est une descripteur statistique (nombre unique) de la similitude de deux échantillons ou individus représentés par plusieurs variables (échantillon multivarié).
- Un indice de similarité prend une valeur**comprise entre 0 (différence totale) et 1 ou 100% (similitude totale)**.
- Un indice de dissimilarité est le complément d'un indice de similarité (dis = 1 − sim); **sa valeur est comprise entre 100% (différence totale) et 0 (similitude totale). Attention: dans certains cas, un indice de dissimilarité peut varier de 0 à $+\infty$** (lorsqu'il s'agit d'une distance, euclidienne par exemple).
- Tous les indices de similarité / dissimilarité peuvent servir à construire des matrices de distance.

UMONS

Indice de dissimilarité : Bray-Curtis

- = coefficient de Czecanowski

$$\mathrm{D}_{\mathrm{Bray-Curtis}_{j,k}} = \frac{\sum_{i=1}^{n} |y_{ij} - y_{ik}|}{\sum_{i=1}^{n} (y_{ij} + y_{ik})}$$

- S'utilise pour mesurer la similitude entre échantillon sur base du **dénombrement d'espèces**. Si le **nombre d'individus est très variable** (espèces dominantes *versus* espèces rares), penser à transformer (ex: $log(x+1)$, double racine carrée, ...)

UMONS

Indice de dissimilarité : Canberra

- Distance similaire à Bray-Curtis mais pondérant les espèces en fonction du nombre d'occurrences.

$$\mathrm{D}_{\mathrm{Canberra}\,j,k} = \frac{1}{nz} \sum_{i'=1}^{nz} \frac{|y_{i'j} - y_{i'k}|}{|y_{i'j}| + |y_{i'k}|}$$

où nz est le nombre de valeurs non nulles.

- Toutes les espèces contribuent de manière égale => possibilité de surimportance d'une espèces mesurée une seule fois !
- Toute double absence n'est pas prise en compte => se comporte bien face aux tableaux comportant beaucoup de zéros (idem Bray-Curtis).

UMONS

Utilisation d'indices Bray-Curtis et Canberra

Seuls les indices ne dépendant pas des doubles zéros sont utilisables pour des dénombrements d'espèces ou des présence-absence.

- Bray-Curtis => résultat dominé par les espèces les plus abondantes.
- Canberra => risque de domination des espèces rares.
- Bray-Curtis sur données transformées (log(x+1) ou double racine carrée) = **souvent bon compromis**.
- Si les volumes échantillonnés entre stations ne sont pas comparables, **il faut standardiser.**

UMONS

Indice de dissimilarité : distance Euclidienne

- Distance géométrique entre les points:

$$\mathrm{D}_{\mathrm{Euclidean}\,j,k} = \sqrt{\sum_{i=1}^{n}(y_{ij} - y_{ik})^2}$$

UMONS

Indice de dissimilarité : Manhattan

- ou encore city-block distance.

$$\mathrm{D}_{\mathrm{Manhattan}\,j,k} = \sum_{i=1}^{n} |y_{ij} - y_{ik}|$$

- Les indices de **dissimilarité** de Bray-Curtis et Camberra sont complémentaires aux indices de similarité correspondants (dis = 1 – sim).

UMONS

Utilisation de ces indices

- Les distances euclidienne ou de Manhattan sont à préférer pour les **mesures environnementales**.

- Les distances de Bray-Curtis ou Canberra sont meilleure pour les **dénombrements d'espèces** (nombreux double zéro) !

UMONS

Indices de (dis)similarité - propriétés

- Les indices varient en 0 et 1 (0 et 100%), mais les distances sont utilisées aussi comme indices de dissimilarité et varient entre 0 et ∞.

- Un indice est dit métrique si :
  - **Minimum 0** : $I_{a,b} = 0$ si $a = b$
  - **Positif** : $I_{a,b} > 0$ si $a \neq b$
  - **Symétrique** : $I_{a,b} = I_{b,a}$
  - **Inégalité triangulaire** : $I_{a,b} + I_{b,c} >= I_{a,c}$

- Un indice est semimétrique s'il répond à toutes les conditions sauf la quatrième.

- Un indice est dit non métrique dans les autres cas.

UMONS

Indice de (dis)similarité - métrique

Distance	**Type**
Bray-Curtis	semimétrique
Canberra	métrique
Euclidienne	métrique
Manhattan	métrique
Chi carré	métrique
(correlation)	(non métrique)
(variance/covariance)	(non métrique)

UMONS