



Clustering de données moléculaires par modèle SBM

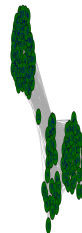
Mohamed Anwar ABOUABDALLAH, 5.A. Lyon 1

Dirigé par Nathalie Peyrard, Alain Franc, Olivier Coulaud



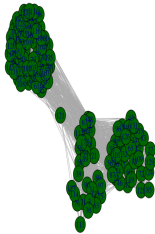


$$\Rightarrow \begin{pmatrix} 0 & 2 & \dots & 3 \\ \vdots & 0 & \ddots & 11 \\ \vdots & \ddots & \ddots & 0 \\ 7 & \dots & \dots & 0 \end{pmatrix} \Rightarrow \begin{pmatrix} 1 & 0 & \dots & 1 \\ \vdots & \ddots & \ddots & 1 \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 1 \end{pmatrix} \Rightarrow$$

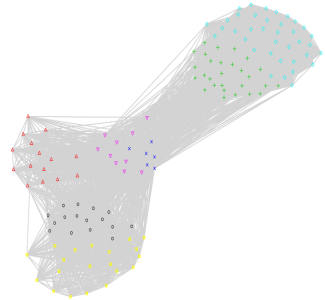


- **Comment caractériser la diversité de cette communauté à partir de ces données ?**

- **Q₁ : Comment trouver les OTU ?**



Graph représentant toutes les classes avec layout FL



- **Q₂ : Comment caractériser les relations entre les OTU ?**



Comment faire mieux qu'un indice scalaire pour calculer la biodiversité?

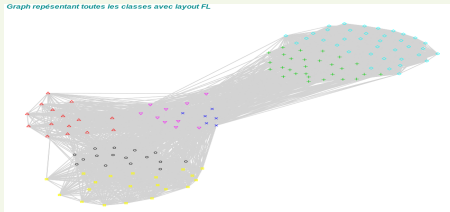
Modèles à Blocs Stochastiques :

Proposition

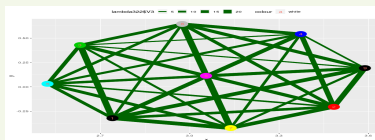
Nous proposons d'utiliser les modèles SBM car :

- Il reconstitue les classes (classe \simeq OTU) (Q_1).

Graph représentant toutes les classes avec layout FL



- Il fournit une structure de connection entre les classes (Q_2)



SBM présence/absence d'Arêtes :

$$n, A, B, \pi, Z, \alpha$$

- $n \in \mathbb{N}$: Le nombre de d'individus (connue).
- $A \in \mathbb{M}_n([0, 1])$: Matrice d'adjacence (connue).
- $B \in \mathbb{N}$: Le nombre de classes.
- $\pi : \forall q, q' = 1, \dots, B, \pi_{q,q'}$ la probabilité de connection entre deux classes (inconnue). Exemple :

$$\pi = \begin{pmatrix} 0.2450382 & 0.5319291 & 0.1892362 & 0.6480529 \\ 0.5319291 & 0.9958117 & 0.9285523 & 0.9984116 \\ 0.1892362 & 0.9285523 & 0.8878484 & 0.3832446 \\ 0.6480529 & 0.9984116 & 0.3832446 & 0.9960055 \end{pmatrix}$$

- $Z \in [0, 1]^{n \times B}$: Indique la classe des espèces (inconnue).
- $\alpha \in [0, 1]^B, \alpha_q = P(Z_{i,q} = 1)$: La probabilité à priori de chaque classe (inconnue).


SBM étendu au cas d'arêtes pondérés :

$$n, D, B, \lambda, Z, \alpha$$

- $D \in \mathbb{M}_n(\mathbb{N})$, $D_{i,j}$ représente la distance entre les sommets.
- $\lambda \in \mathbb{M}_n(\mathbb{R})$, $\forall q, q' = 1, \dots, B$, $\lambda_{q,q'}$ le paramètre d'une loi de Poisson pour tirer la distance entre deux séquences en fonction des classes auxquelles elles appartiennent. Exemple :

$$\lambda = \begin{pmatrix} 8.817153 & 13.972507 & 18.156499 & 16.979863 \\ 13.972507 & 5.433602 & 16.235190 & 14.654911 \\ 18.156499 & 16.235190 & 3.037915 & 5.265202 \\ 16.979863 & 14.654911 & 5.265202 & 3.787646 \end{pmatrix}$$

- Les autres termes définis de manière similaire.

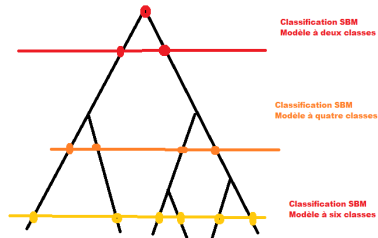
- **Comment estimer les paramètres de les modèles ?**
 - **Algorithme VEM** : Version approchée de l'EM où on fait une hypothèse d'indépendance des classes sachant le graphe.
- **Comment sélectionner le modèle ?**
 - Le critère ICL serait utilisé pour choisir notre B (nombre de classes)
- Sur , il est possible d'utiliser la librairie **BlockModels**.

Questions étudiées :

- À nombre égal de classes :
Est-ce que les classifications des deux méthodes SBM_bin et SBM_dis se ressemblent ?

- À nombre croissants de classes :

Est-ce qu'à l'intérieur d'un modèle (SBM_bin soit SBM_dis) les classifications emboîtées ?



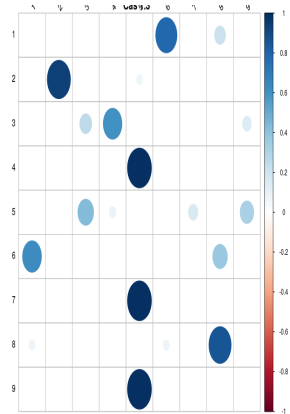
Les tests sont effectués sur les données portant sur les diatomées (Data set de 132 individus.)

Comparaison des classifications SBM_bin SBM_dis :

Matrice de comparaison entre les deux méthodes :

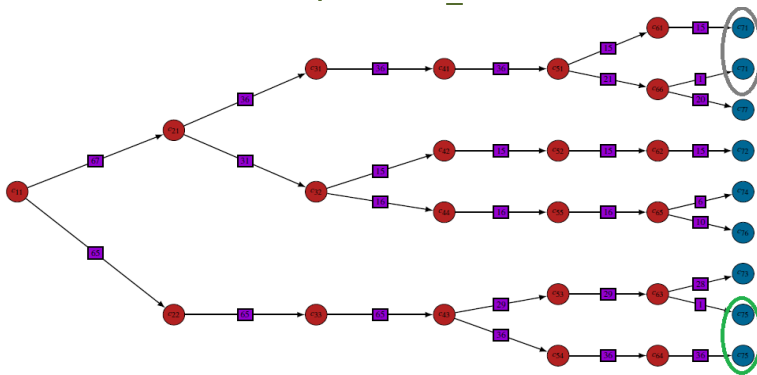
$$M = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 11 & 0 & 3 & 0 \\ 0 & 14 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 7 & 17 & 0 & 0 & 0 & 0 & 4 \\ 0 & 0 & 0 & 0 & 5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 16 & 3 & 0 & 0 & 6 & 0 & 12 \\ 5 & 0 & 0 & 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 7 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 12 & 0 \\ 0 & 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 \end{pmatrix}$$

- Simulation avec $B_{\text{dis}}^{\text{opt}} = 9$
- On applique cette transformation :
 $M_{i,j} \longrightarrow \frac{M_{i,j}}{\sum_j M_{i,j}}$ et on trace la heatmap associée.



ement des

- **Emboitement des classes pour SBM bin :**



Bilan : Sans l'imposer, on retrouve quasiment une structure en arbre.

tion :

●

Les modèles SBM sont très intéressants pour le clustering des données moléculaires :

- La différence entre les modèles n'est pas très importante.
- On arrive à retrouver une structure en arbre sans l'imposer.

●

- Tester ces méthodes sur de grands jeux de données.
- Les quelques différences entre les deux modèles sont trop faibles ce qui impose de continuer la phase d'exploration.
- Tester sur des jeux de données où on connaît à la fois les espèces, les genres et les familles.

ement des

- **Emboitement des classes pour SBM** dis :

