

Détection de structure dans des réseaux bipartites

Séminaire des stagiaires

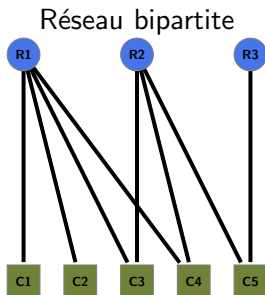
Louis Lacoste

29 juin 2023

Contexte écologique

- Faire de la détection de structure sur un réseau (SBM, LBM) mais intérêt à le faire sur plusieurs
- De nombreux réseaux disponibles (« Web of Life : Ecological Networks Database », s. d.) et décrivant des interactions similaires
- Re-grouper les réseaux selon leur similarité (*clustering* de réseaux)
- Transférer de l'information grâce à la collection (par exemple reconstitution de données manquantes)
- Déterminer des structures d'interactions fines de manière agnostique

Réseaux bipartites¹



Matrice d'incidence

$$X = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Permet de décrire des interactions impliquant deux agents dont les rôles sont de natures différentes.

Par exemple : hôtes-parasites, plantes-pollinisateurs, graines-disperseurs ...

1. Ou *bipartis*. Voir Larousse, s. d.

Latent Block Model (LBM²)

Proposé par Govaert et Nadif, 2005.

Pour

- $Q_1 = |\{\bullet, \bullet, \bullet\}|$ blocs fixés en ligne
- $Q_2 = |\{\bullet, \bullet, \bullet\}|$ blocs fixés en colonne

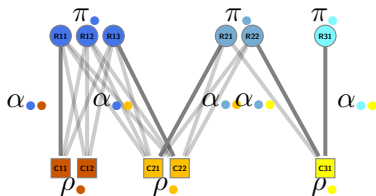


Figure – Exemple de LBM²

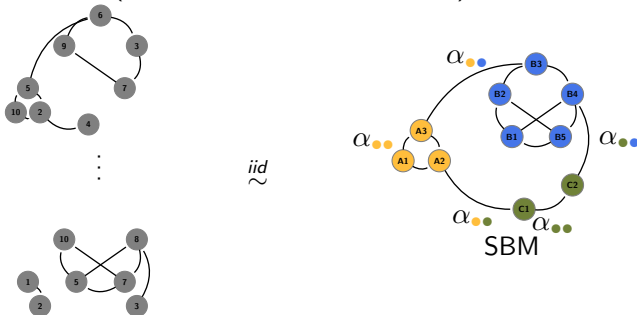
Paramètres

- $\pi_{\bullet} = \mathbb{P}(Z_i = \bullet)$ en ligne et $\rho_{\bullet} = \mathbb{P}(W_j = \bullet)$ en colonne
- $\alpha_{\bullet, \bullet} = \mathbb{P}(X_{ij} = 1 | Z_i = \bullet, W_j = \bullet)$

2. Que j'appellerai par la suite BiSBM

colSBM

Le modèle *colSBM* (Chabert-Liddell et al., 2023).

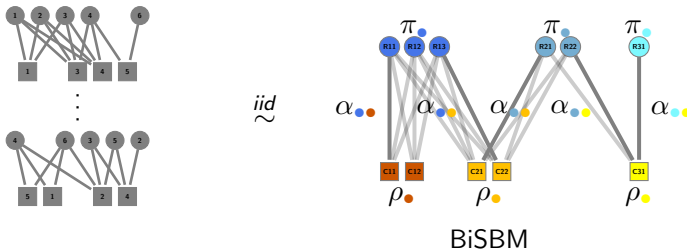


Pour $Q = |\{\bullet, \bullet, \bullet\}|$ blocs fixés :

Paramètres

- $\pi_{\bullet} = \mathbb{P}(Z_i = \bullet)$
- $\alpha_{\bullet, \bullet} = \mathbb{P}(X_{ij} = 1 | Z_i = \bullet, Z_j = \bullet)$

Collections bipartites



Pour

- $Q_1 = |\{\bullet, \bullet, \bullet\}|$ blocs fixés en ligne
- $Q_2 = |\{\bullet, \bullet, \bullet\}|$ blocs fixés en colonne

Paramètres

- $\pi_\bullet = \mathbb{P}(Z_i = \bullet)$ en ligne et $\rho_\bullet = \mathbb{P}(W_j = \bullet)$ en colonne
- $\alpha_{\bullet, \bullet} = \mathbb{P}(X_{ij} = 1 | Z_i = \bullet, W_j = \bullet)$

Différents modèles

iid-colBiSBM

$\pi = (\pi_1, \dots, \pi_{Q_1})$ et $\rho = (\rho_1, \dots, \rho_{Q_2})$, tous les réseaux partagent les mêmes paramètres³

$\pi\rho$ -colBiSBM

$\pi = ((\pi_1^m, \dots, \pi_{Q_1}^m))_{m=1, \dots, M}$ et $\rho = ((\rho_1^m, \dots, \rho_{Q_2}^m))_{m=1, \dots, M}$
avec $\forall q, m \in \llbracket 1, Q_1 \rrbracket \times \llbracket 1, M \rrbracket, \pi_q^m \in [0, 1]$ et $\forall r, m \in \llbracket 1, Q_2 \rrbracket \times \llbracket 1, M \rrbracket, \rho_r^m \in [0, 1]$

Et également deux autres modèles (π -colBiSBM et ρ -colBiSBM) où seulement une des deux dimensions est libre.

3. Dans tous les modèles la structure de connectivité est supposée identique au sein de la collection.

Estimation des paramètres

Maximisation d'une borne inférieure de la log-vraisemblance des données observées.

$$\begin{aligned}
 \ell(\mathbf{X}; \boldsymbol{\theta}) \geq & \sum_{m=1}^M \left(\sum_{i=1}^{n_1^m} \sum_{j=1}^{n_2^m} \sum_{q \in \mathcal{Q}_{1,m}} \sum_{r \in \mathcal{Q}_{2,m}} \tau_{i,q}^{1,m} \tau_{j,r}^{2,m} \log f(X_{ij}^m; \alpha_{qr}) \right. \\
 & + \sum_{i=1}^{n_1^m} \sum_{q \in \mathcal{Q}_{1,m}} \tau_{i,q}^{1,m} \log \pi_q^m + \sum_{j=1}^{n_2^m} \sum_{r \in \mathcal{Q}_{2,m}} \tau_{j,r}^{2,m} \log \rho_r^m \\
 & \left. - \sum_{i=1}^{n_1} \tau_{i,q}^{1,m} \log \tau_{i,q}^{1,m} - \sum_{j=1}^{n_2} \tau_{j,r}^{2,m} \log \tau_{j,r}^{2,m} \right) =: J(\boldsymbol{\tau}; \boldsymbol{\theta})
 \end{aligned}$$

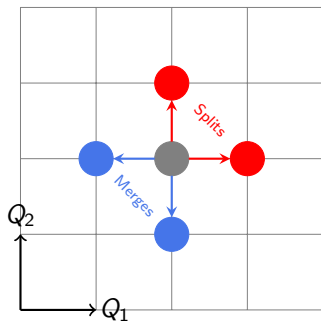
Approximation variationnelle

$\tau_{i,q}^{1,m} = P(Z_i = q | X_{ij}^m)$ et $\tau_{j,r}^{2,m} = P(W_j = r | X_{ij}^m)$ tels que
 $P(Z_i = q, W_j = r | X_{ij}^m) = \tau_{i,q}^{1,m} \times \tau_{j,r}^{2,m}$

Sélection de modèle : choix de (Q_1, Q_2) - Approche gloutonne

Le VEM se fait à Q_1, Q_2 fixés, il faut donc déterminer les “meilleures” coordonnées. Nous maximisons un BIC-L⁴.

Détermination d'un premier mode par approche *gloutonne*



Exploration gloutonne

- Initialisation sur $(1, 2)$ et $(2, 1)$
- Exploration des 4 voisins et déplacement sur le meilleur des 4
- Arrêt après 2 étapes successives sans augmentation du BIC-L

4. *Bayesian Information Criterion - Like*, en adaptant les formules de Chabert-Liddell et al., 2023

Sélection de modèle : choix de (Q_1, Q_2) - Fenêtre glissante

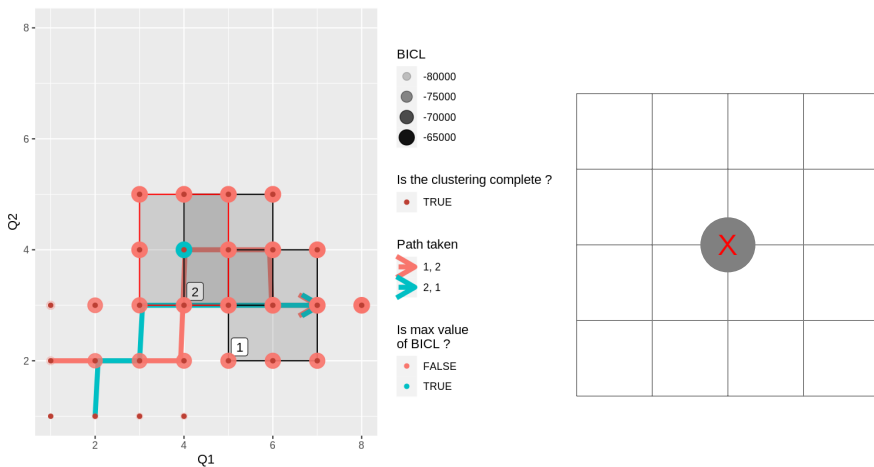


Figure – Exemple de parcours de fenêtre glissante

Sélection de modèle : choix de (Q_1, Q_2) - Fenêtre glissante

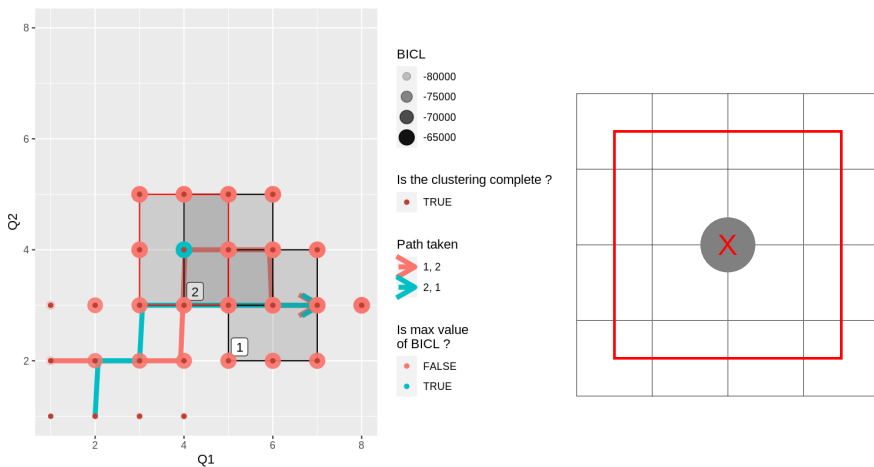


Figure – Exemple de parcours de fenêtre glissante

Sélection de modèle : choix de (Q_1, Q_2) - Fenêtre glissante

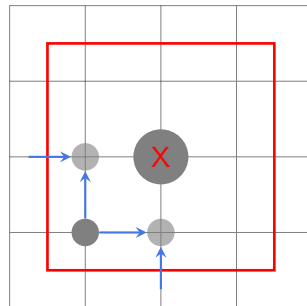
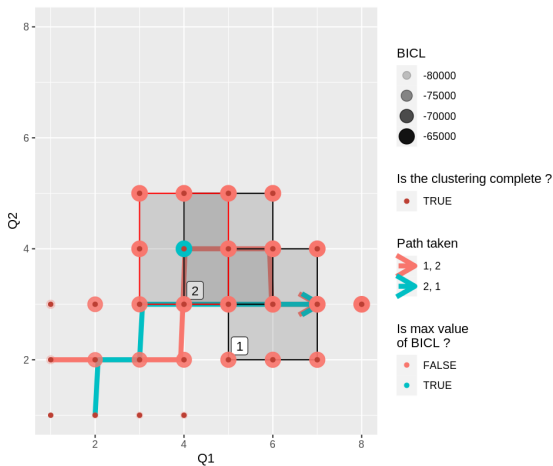


Figure – Exemple de parcours de fenêtre glissante

Sélection de modèle : choix de (Q_1, Q_2) - Fenêtre glissante

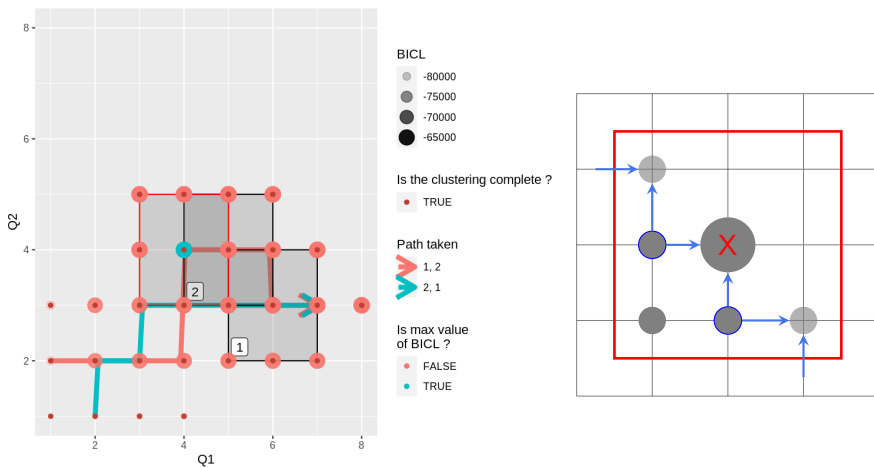


Figure – Exemple de parcours de fenêtre glissante

Sélection de modèle : choix de (Q_1, Q_2) - Fenêtre glissante

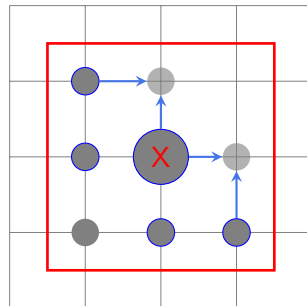
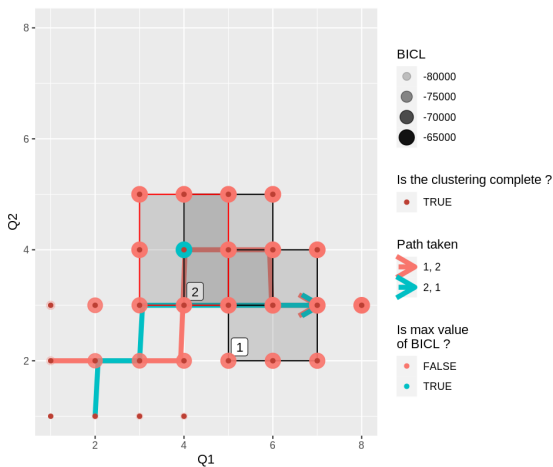
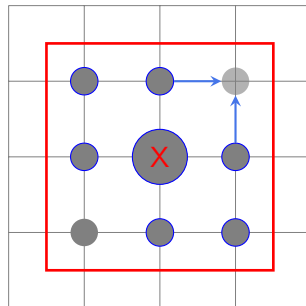


Figure – Exemple de parcours de fenêtre glissante

The figure is a 2D plot with the horizontal axis labeled Q_1 and the vertical axis labeled Q_2 . Both axes range from 0 to 8, with major grid lines every 2 units. A path of colored lines connects several points on the grid. The path starts at $(1, 1)$, goes to $(2, 1)$, then to $(2, 2)$, then to $(3, 2)$, then to $(4, 2)$, then to $(4, 4)$, then to $(5, 4)$, then to $(6, 4)$, then to $(7, 4)$, then to $(7, 2)$, then to $(6, 2)$, then to $(5, 2)$, then to $(4, 2)$, then to $(3, 2)$, then to $(3, 4)$, then to $(3, 5)$, then to $(4, 5)$, then to $(5, 5)$, then to $(6, 5)$, then to $(7, 5)$, then to $(8, 5)$. The path is colored red, blue, and green. A shaded gray region is labeled '1' and '2'. A red arrow points from the shaded region towards the right.

• TRUE



9 / 12

Sélection de modèle : choix de (Q_1, Q_2) - Fenêtre glissante

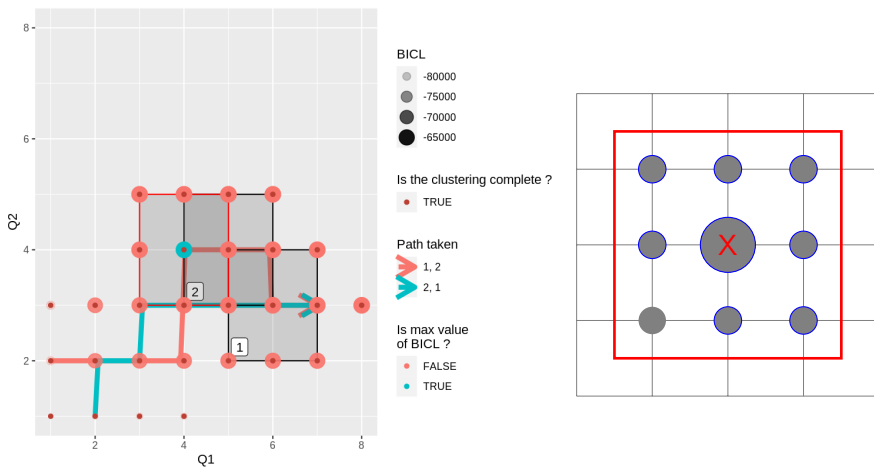


Figure – Exemple de parcours de fenêtre glissante

Sélection de modèle : choix de (Q_1, Q_2) - Fenêtre glissante

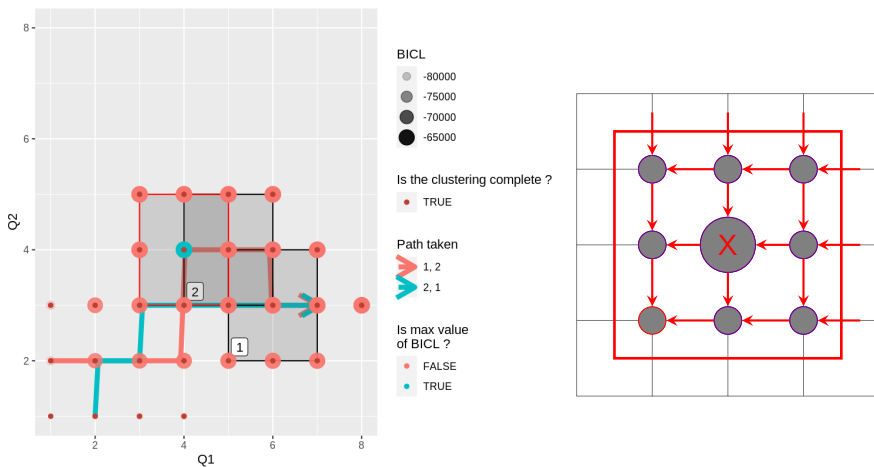
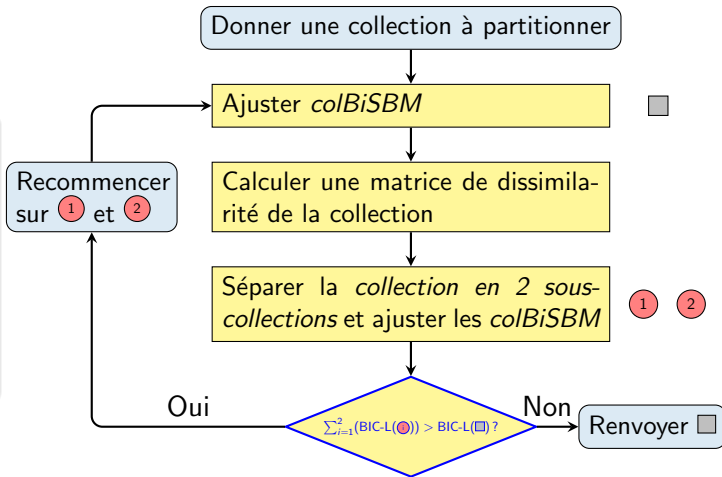


Figure – Exemple de parcours de fenêtre glissante

Clustering de réseaux

Objectif

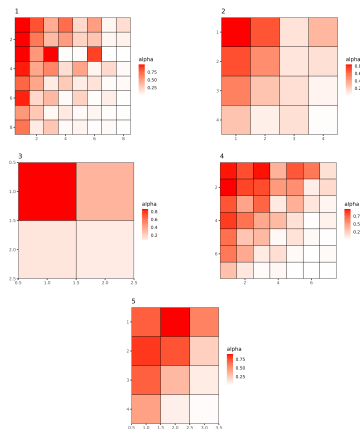
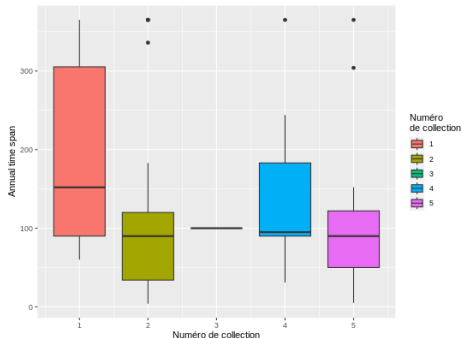
Déterminer une partition qui maximise la somme du BICL de ses sous-collections.



Même approche que Chabert-Liddell et al., 2023

Application, données plantes pollinisateurs

Voici des résultats du modèle *iid-colBiSBM* sur des données plantes-pollinisateurs (Doré et al., 2021 et Thébault et Fontaine, 2020)




N° de collection	1	2	3	4	5
Nombre de réseaux	38	45	1	20	19

Figure – Connectivités de la partition

Conclusion et perspectives

- 4 modèles dont 3 qui ont une flexibilité sur au moins une des dimensions (adaptabilité aux données)
- Partitionner un ensemble de réseaux selon leurs structures
- Comparer les *clusterings* de réseaux obtenus entre données brutes et données corrigées (par exemple par la méthode *CoOPLBM*⁵)

Le package est disponible sur GitHub :

 <https://github.com/Chabert-Liddell/colSBM>

Merci pour votre attention !

5. Anakok et al., 2022

Bibliographie I

Web of Life : Ecological Networks Database. (s. d.).

Larousse, É. (s. d.). Définitions : biparti, bipartite - Dictionnaire de français Larousse.

Govaert, G., & Nadif, M. (2005). An EM Algorithm for the Block Mixture Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(4), 643-647.
<https://doi.org/10.1109/TPAMI.2005.69>

Chabert-Liddell, S.-C., Barbillon, P., & Donnet, S. (2023). Learning Common Structures in a Collection of Networks. An Application to Food Webs. <https://doi.org/10.48550/arXiv.2206.00560>

Doré, M., Fontaine, C., & Thébault, E. (2021). Relative effects of anthropogenic pressures, climate, and sampling design on the structure of pollination networks at the global scale. *Global Change Biology*, 27(6), 1266-1280. <https://doi.org/10.1111/gcb.15474>

Thébault, E., & Fontaine, C. (2020). A Database of Plant-Pollinator Networks. <https://doi.org/10.5281/zenodo.4300427>

Bibliographie II

Anakok, E., Barbillon, P., Fontaine, C., & Thebault, E. (2022).

Disentangling the structure of ecological bipartite networks from observation processes.