

Comparaison de structures de réseaux. Applications à des réseaux écologiques

Audition de candidature de thèse à l'EDMH

Louis Lacoste

Dirigé par Sophie Donnet et Pierre Barbillon, co-encadré par Julie Aubert

UMR MIA Paris-Saclay

23 mai 2024

- 1 Parcours
- 2 Sujet de thèse
- 3 Organisation de la thèse

Parcours

- 2018–2020, Classe Préparatoire BCPST
- 2020–2022, 1ère et 2ème année en formation Ingénieur AgroParisTech
Cours optionnels suivis : statistiques spatiales, mathématiques pour la santé, ingénierie par la simulation informatique ...
- 2022–2023, Année de césure
- 2023–2024, M2 Mathématiques pour les Sciences du Vivant, Université Paris-Saclay
UC à choix 2^e semestre : modèles à variables latentes, statistiques spatiales et méthodes de statistiques en grandes dimension

Expériences professionnelles

- 2022 Mai–Déc., Stage assistant ingénieur en Qualité chez Eurofins Food France.
- 2023 Janv.–Juillet, Détection de structures dans des collections de réseaux bipartites et écriture du package implémentant la méthode. Stage dans l'UMR MIA Paris-Saclay, supervisé par Pierre Barbillon.
- 2024 Avril–Sept., Détection de structures et clustering de réseaux écologiques. Stage dans l'UMR MIA Paris-Saclay, supervisé par Pierre Barbillon et Sophie Donnet.

Sujet de thèse

Contexte écologique

- Nombreux réseaux disponibles (« [Web of Life : Ecological Networks Database](#) », s. d.) pour interactions similaires. Par exemple, interactions proies-prédateurs, plantes-pollinisateurs ...
- Suivi biodiversité, analyse de robustesse et risque d'effondrement.

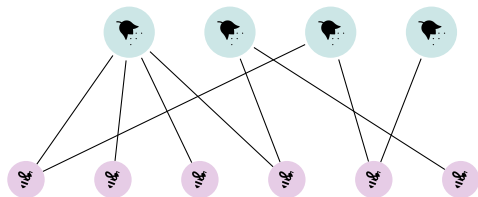


Figure 1 – Exemple d'un réseau plantes-pollinisateurs

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

Matrice d'adjacence associée

- En écologie microbienne réseaux permettent le suivi de la qualité des sols.

Détection de structure¹ pour un unique réseau bien connu avec par exemple :

- Modèles de *clustering* à variables latentes.
- *Embedding* par apprentissage profond.
- Et bien d'autres méthodes.

Mais des motivations pour considérer des collections :

- Espèces différentes, rôles analogues.
- Transfert d'informations grands vers petits réseaux.
- Regrouper les réseaux selon leur similarité (*clustering* de réseaux).

Axe 1 : Modèles à variables latentes pour une collection de réseaux bipartites

Latent Block Model (LBM)

Proposé par Govaert et Nadif, 2005.

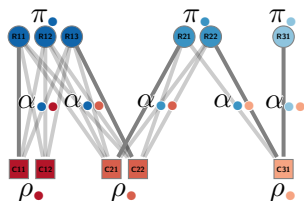


Figure 2 – Exemple de LBM²

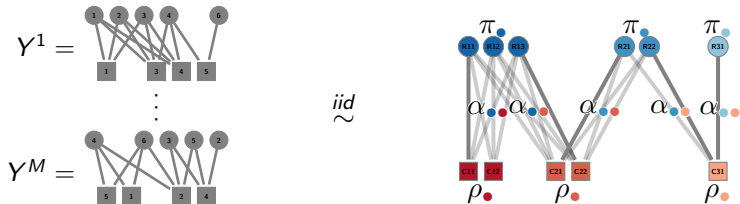
Pour

- $Q_1 = |\{\bullet, \bullet, \bullet\}|$ 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
- $\alpha_{\bullet, \bullet} = \mathbb{P}(X_{ij} = 1 | Z_i = \bullet, W_j = \bullet)$

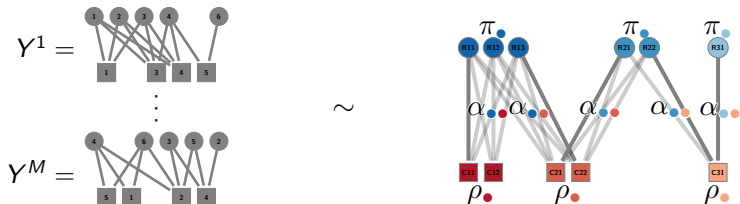
Collections bipartites



Modèle *iid*-colBiSBM

$$\forall m \in \llbracket 1, M \rrbracket, Y^m \sim LBM_{n_1^m, n_2^m}(\pi, \rho, \alpha)$$

Collections bipartites



Modèle $\pi\rho$ -colBiSBM

$$\forall m \in \llbracket 1, M \rrbracket, Y^m \sim LBM_{n_1^m, n_2^m}(\pi^m, \rho^m, \alpha)$$

- Écriture du modèle colBiSBM.
- Dérivation des formules d'inférence et d'un critère de sélection de modèle par vraisemblance pénalisée (choix du nombre de blocs).
- Implémentation des formules et du critère et développement algorithmique pour l'exploration de l'espace de paramètres.
- Développement d'une méthode de partitionnement d'une large collection de réseaux basée sur la maximisation d'un critère de sélection de modèle.
- Implémentation du code intégrée au package³ *colSBM*.

Application du *clustering*, données plantes pollinisateurs

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

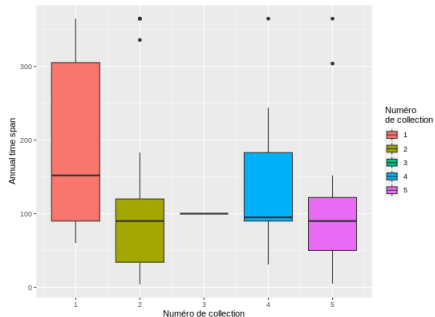
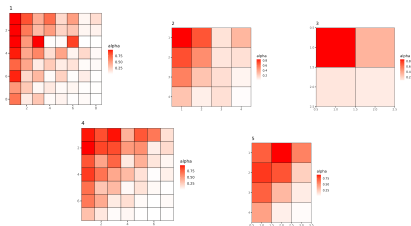


Figure 3 – Connectivités de la partition

N°de collection	1	2	3	4	5	Total
Nombre de réseaux	38	45	1	20	19	123

- Finaliser l'analyse sur données réelles commencée sur [Doré et al., 2021](#) ; [Thébault et Fontaine, 2020](#) avec les interprétations des écologues en vue d'une publication.
- Preuve d'identifiabilité du modèle ([Chabert-Liddell et al., 2024](#) ; [Celisse et al., 2012](#) ; [Keribin et al., 2015](#) ; [Brault & Mariadassou, 2015](#)).

Axe 2 : Embedding de nœuds par apprentissage profond pour comparaison des topologies de réseaux

Graph Neural Networks et Variational AutoEncoder

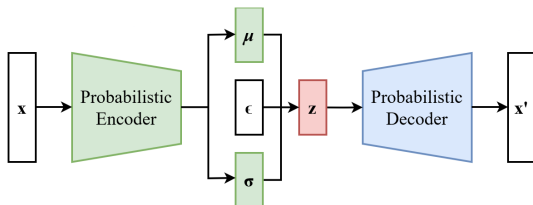


Figure 4 – Schéma d'« Auto-encodeur variationnel », 2024

Problème des graphes pour les réseaux de neurones : **prise en compte des invariances**

- Utilisation des *Graph Convolutional Networks* de [Kipf et Welling, 2017](#) résout ce problème.
- Utiliser des *Variational AutoEncoder* de [Kingma et Welling, 2022](#) et de [Kipf et Welling, 2016](#) pour projection des nœuds dans un espace latent.
- Explorer le *Deep Latent Space Model* de [Yang et al., 2024](#)

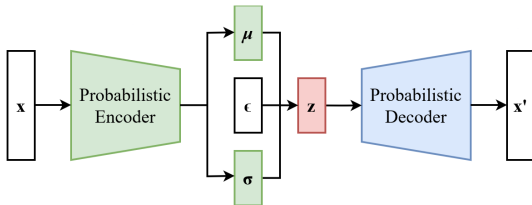


Figure 4 – Schéma d'« Auto-encodeur variationnel », 2024

Idées à explorer pour la comparaison de réseaux

- *Embedding* commun des nœuds des réseaux à comparer sur un même espace latent. Puis comparaison des distributions ainsi obtenues.
- Encodeurs différents mais un décodeur commun pour comparer les représentations obtenues

Axe 3 : Inférence jointe de réseaux

	OTU_1	...	OTU_p
Éch. 1	$X_{1,1}$...	$X_{1,p} = 500$
\vdots	\vdots		\vdots
Éch. n	$X_{n,1} = 10$...	$X_{n,p}$

Table 1 – Table d'OTU synthétique

TODO Ici une image avec des liens de probas proches pour faire commentaire

Figure 5 – Réseau de co-occurrence inféré

- Avènement des techniques de séquençage et donc disponibilité des OTU⁴
- Incertitude d'inférence disponible mais négligée par la suite, important pour les réseaux microbiens seulement inférés
- Autre limites et détails dans [Matchado et al., 2021](#)

Idées à explorer pour l'inférence jointe de réseaux

- Avec M tableaux d'OTU, on peut supposer :

Modèle hiérarchique

$$\forall m \in \llbracket 1, M \rrbracket, X_1^m, \dots, X_p^m \rightsquigarrow \mathcal{M}(Y^m)$$

$$Y^m \rightsquigarrow LBM(\pi, \rho, \alpha) \text{ ou } Y^m \rightsquigarrow DLSM(f_D, f_E)$$

- Réussir à mettre en évidence des bactéries aux rôles fonctionnels proches selon des conditions d'expériences différentes en tenant compte de l'incertitude d'inférence
- Formaliser une méthode pour déterminer si le changement d'unité taxonomique change la structure

Organisation de la thèse

Planning prévisionnel de la thèse

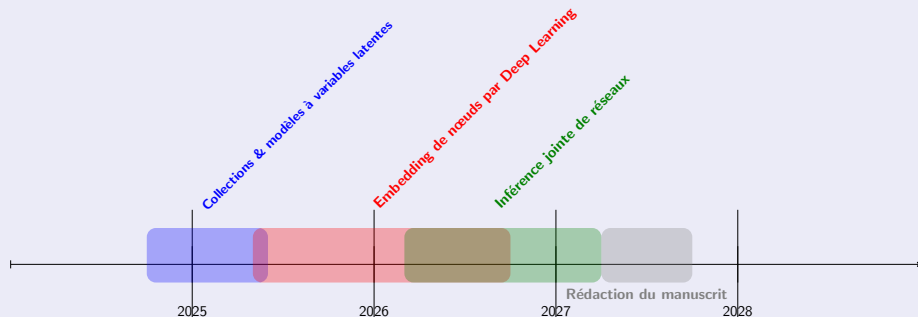


Figure 6 – Chronologie prévue

Financement

L'INRAE, par le département MathNum accorde déjà 50% des financements de la thèse.

Merci pour votre attention.

Bibliographie I

- Web of Life : Ecological Networks Database.* (s. d.). Récupérée juin 17, 2023, à partir de <https://www.web-of-life.es/map.php>
- 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>
- 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, décembre 1). *A Database of Plant-Pollinator Networks* (Version 1). *Zenodo*.
<https://doi.org/10.5281/zenodo.4300427>

- Chabert-Liddell, S.-C., Barbillon, P., & Donnet, S. (2024). Learning Common Structures in a Collection of Networks. An Application to Food Webs. *The Annals of Applied Statistics*, 18(2), 1213-1235.
<https://doi.org/10.1214/23-AOAS1831>
- Celisse, A., Daudin, J.-J., & Pierre, L. (2012). Consistency of Maximum-Likelihood and Variational Estimators in the Stochastic Block Model. *Electronic Journal of Statistics*, 6, 1847-1899.
<https://doi.org/10.1214/12-EJS729>
- Keribin, C., Brault, V., Celeux, G., & Govaert, G. (2015). Estimation and selection for the latent block model on categorical data. *Stat Comput*, 25(6), 1201-1216.
<https://doi.org/10.1007/s11222-014-9472-2>

Bibliographie III

- Brault, V., & Mariadassou, M. (2015). Co-clustering through Latent Bloc Model : a Review. *Journal de la société française de statistique*, 156(3), 120-139. Récupérée mai 15, 2024, à partir de http://www.numdam.org/item/JSFS_2015__156_3_120_0/
- Auto-encodeur variationnel. (2024, mars 13). In *Wikipédia*. Récupérée mai 21, 2024, à partir de https://fr.wikipedia.org/w/index.php?title=Auto-encodeur_variationnel&oldid=213326719
Page Version ID : 213326719.
- Kipf, T. N., & Welling, M. (2017, février 22). *Semi-Supervised Classification with Graph Convolutional Networks*. *arXiv* : 1609.02907 [cs, stat].
<https://doi.org/10.48550/arXiv.1609.02907>

Bibliographie IV

- Kingma, D. P., & Welling, M. (2022, décembre 10). *Auto-Encoding Variational Bayes*. arXiv : 1312.6114 [cs, stat].
<https://doi.org/10.48550/arXiv.1312.6114>
- Kipf, T. N., & Welling, M. (2016, novembre 21). *Variational Graph Auto-Encoders*. arXiv : 1611.07308 [cs, stat].
<https://doi.org/10.48550/arXiv.1611.07308>
- Yang, H., Kong, Q., & Mao, W. (2024). A Deep Latent Space Model for Graph Representation Learning. *Neurocomputing*, 576, 127342.
<https://doi.org/10.1016/j.neucom.2024.127342>
- Matchado, M. S., Lauber, M., Reitmeier, S., Kacprowski, T., Baumbach, J., Haller, D., & List, M. (2021). Network Analysis Methods for Studying Microbial Communities : A Mini Review. *Computational and Structural Biotechnology Journal*, 19, 2687-2698. <https://doi.org/10.1016/j.csbj.2021.05.001>

Annexes

Modèles à variables latentes pour collection de réseaux bipartites

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

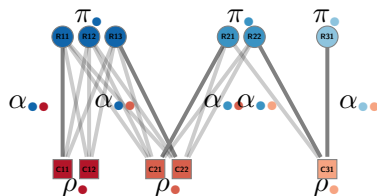
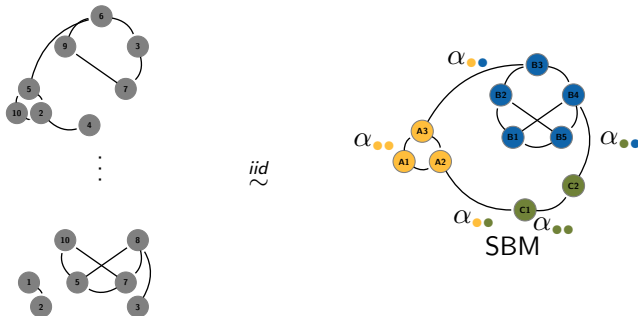


Figure 7 – Exemple de LBM⁵

Paramètres

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

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

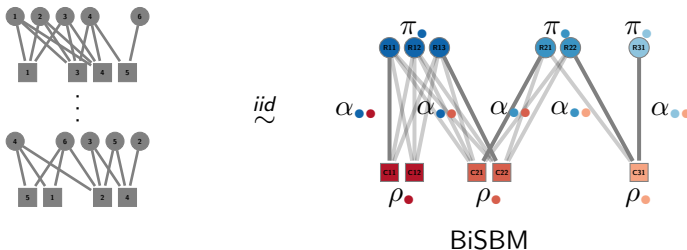


Pour $Q = |\{\bullet, \bullet, \bullet\}|$ blocs fixés :

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

iid-colBiSBM

$\pi = (\pi_1, \dots, \pi_{Q_1})$ et $\rho = (\rho_1, \dots, \rho_{Q_2})$, tous les réseaux partagent les mêmes paramètres⁶

$\pi\rho$ -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.

6. 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.

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

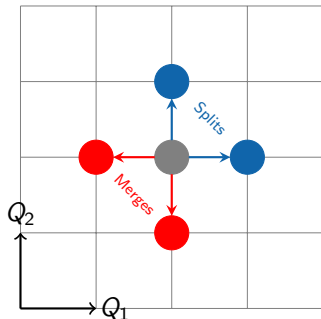
Approximation variationnelle

$\tau_{i,q}^{1,m} = P(Z_i = q | X_{ij}^m)$ 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-L⁷.

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

7. *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

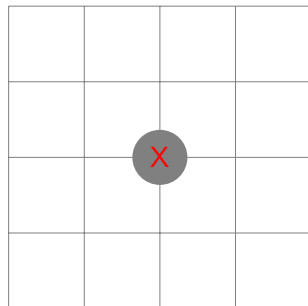
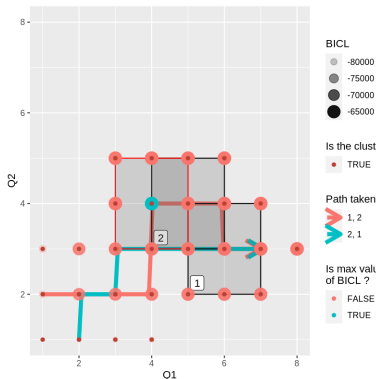


Figure 8 – Exemple de parcours de fenêtre glissante

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

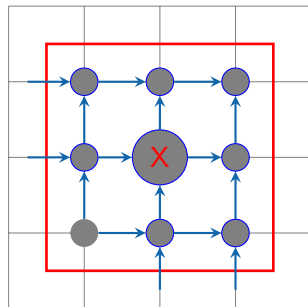
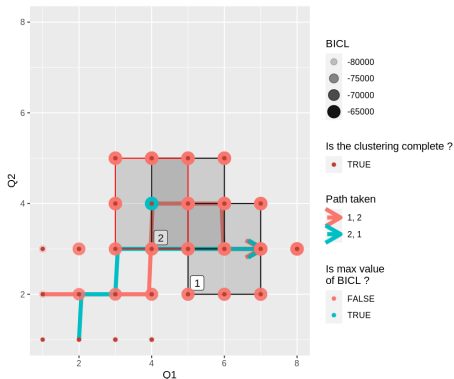


Figure 8 – Exemple de parcours de fenêtre glissante

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

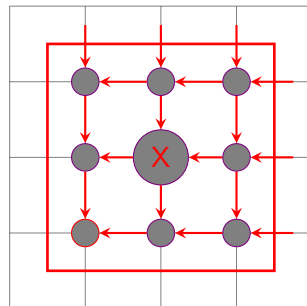
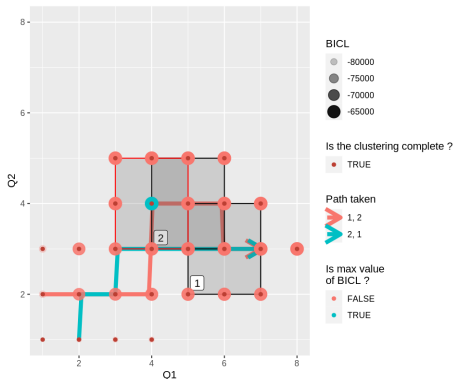
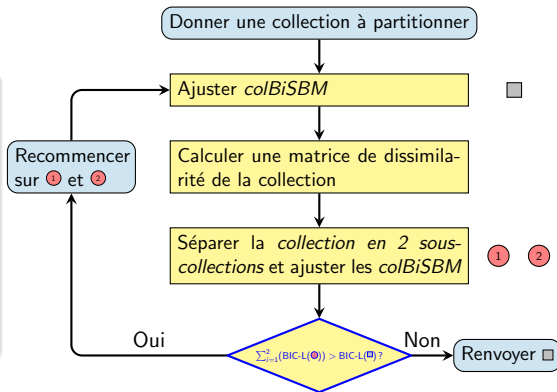


Figure 8 – Exemple de parcours de fenêtre glissante

Clustering de réseaux

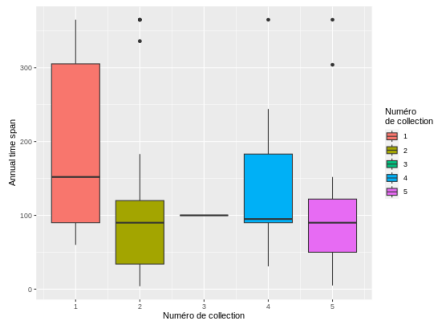
Objectif

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



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

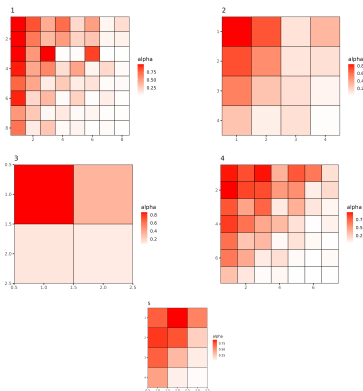



Figure 9 – Connectivités de la partition

- 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*⁸)

Le package est disponible sur GitHub :

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

Bibliographie des annexes I

- 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, mars 27). *Learning Common Structures in a Collection of Networks. An Application to Food Webs*. arXiv : 2206.00560 [stat].
<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, décembre 1). *A Database of Plant-Pollinator Networks (Version 1)*. Zenodo.
<https://doi.org/10.5281/zenodo.4300427>

Anakok, E., Barbillon, P., Fontaine, C., & Thebault, E. (2022, novembre 29). *Disentangling the structure of ecological bipartite networks from observation processes*. arXiv : 2211.16364 [stat]. Récupérée juin 14, 2023, à partir de <http://arxiv.org/abs/2211.16364>