Détection de structure dans des réseaux bipartites

Séminaire des stagiaires

Louis Lacoste

4 juillet 2024

Contexte écologique

- Nombreux réseaux disponibles pour interactions similaires.
- Suivi biodiversité, robustesse et risque d'effondrement ...

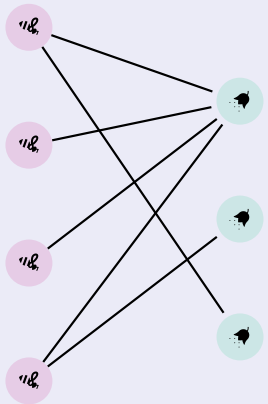


Figure – Exemple d'un réseau plantes-pollinisateurs

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

Matrice d'adjacence associée

Contexte écologique

- Nombreux réseaux disponibles pour interactions similaires.
- Suivi biodiversité, robustesse et risque d'effondrement ...

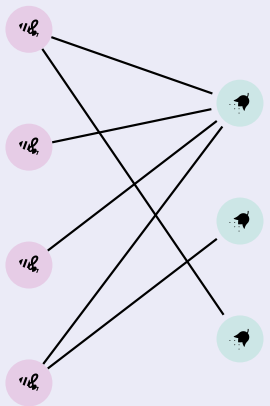


Figure – Exemple d'un réseau plantes-pollinisateurs

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

Matrice d'adjacence associée

Contexte mathématique

Pour un unique réseau : variables latentes, *embedding*, ...
Motivations pour proposer des méthodes adaptées aux collections de réseaux :

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

Latent Block Model (LBM [1])

Proposé par Govaert et Nadif, 2005.

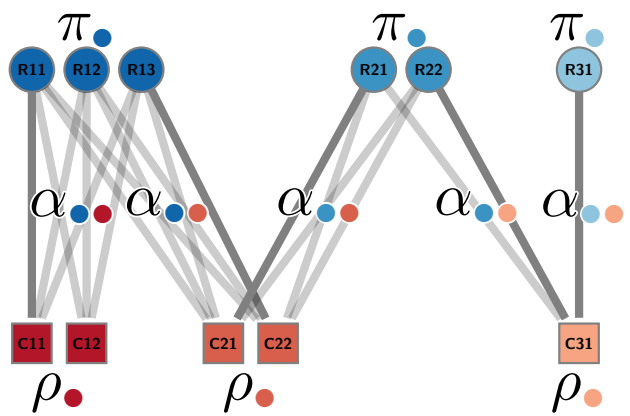


Figure – Exemple de LBM [1]

Pour

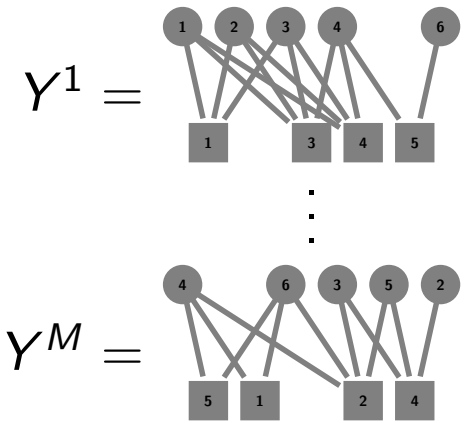
- $Q_1 = |\{\bullet, \bullet, \bullet\}|$ blocs fixés en ligne
- $Q_2 = |\{\bullet, \bullet, \bullet\}|$ blocs fixés en colonne

Paramètres

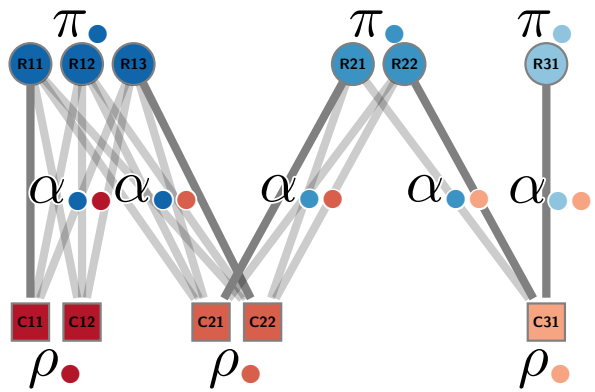
- $\pi_\bullet = \mathbb{P}(Z_i = \bullet)$ en ligne et $\rho_\bullet = \mathbb{P}(W_j = \bullet)$ en colonne
- $\alpha_{\bullet\bullet} = \mathbb{P}(X_{ij} = 1 | Z_i = \bullet, W_j = \bullet)$

1. Que j'appellerai par la suite BiSBM

Collections bipartites



$\overset{iid}{\sim}$

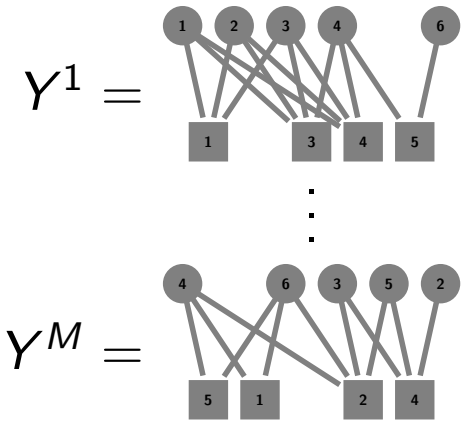


- $Q_1 = |\{\bullet, \bullet, \bullet\}|$ blocs fixés en ligne
- $Q_2 = |\{\bullet, \bullet, \bullet\}|$ blocs fixés en colonne

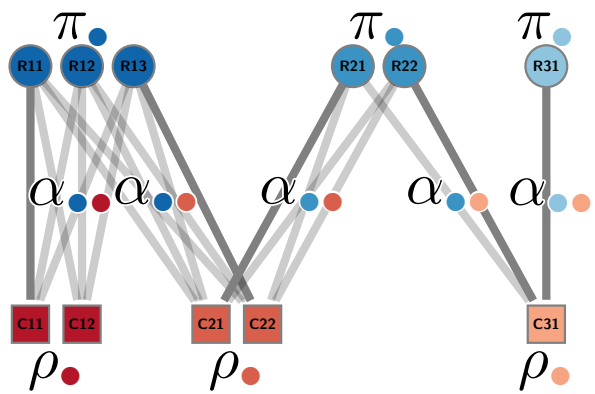
Paramètres

- $\pi_\bullet = \mathbb{P}(Z_i = \bullet)$ 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



$\overset{iid}{\sim}$

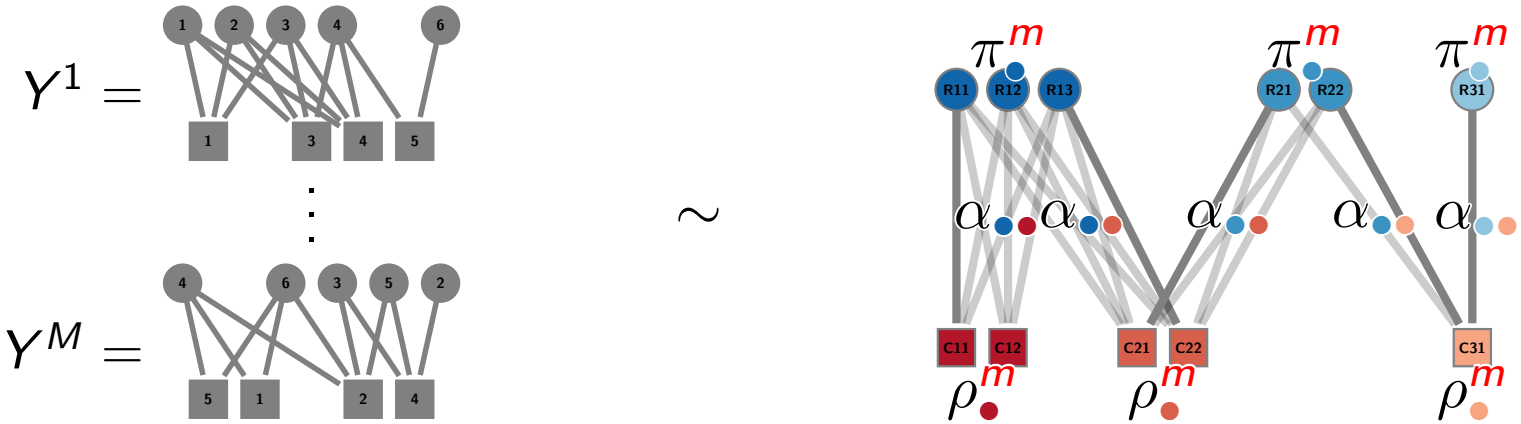


iid-colBiSBM

$\boldsymbol{\pi} = (\pi_1, \dots \pi_{Q_1})$ et $\boldsymbol{\rho} = (\rho_1, \dots \rho_{Q_2})$

Dans tous les modèles la structure de connectivité $(\boldsymbol{\alpha})$ est supposée identique au sein de la collection.

Différents modèles



$\pi\rho$-colBiSBM

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

Dans tous les modèles la structure de connectivité ($\boldsymbol{\alpha}$) est supposée identique au sein de la collection.

Estimation des paramètres

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

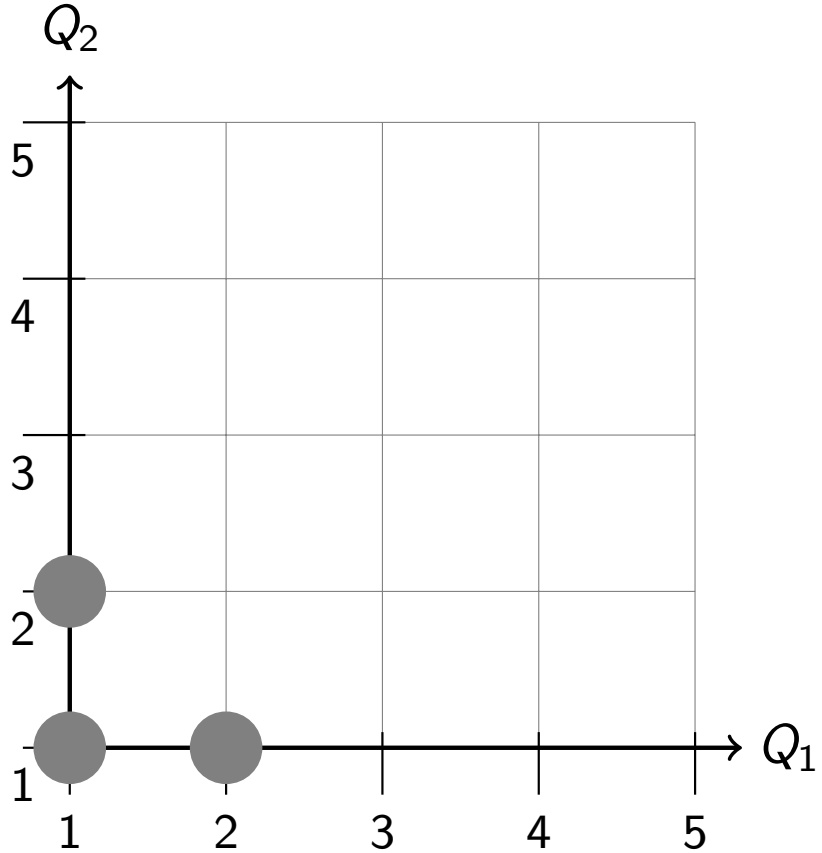
$$\ell(\mathbf{X};\boldsymbol{\theta}) \geq \sum_{m=1}^{M} \Bigg(\sum_{i=1}^{n_1^m} \sum_{j=1}^{n_2^m} \sum_{q\in\mathcal{Q}_{1,m}} \sum_{r\in\mathcal{Q}_{2,m}} \tau_{i,q}^{1,m}\tau_{j,r}^{2,m} \log f(X_{ij}^m;\alpha_{qr}) + \sum_{i=1}^{n_1^m} \sum_{q\in\mathcal{Q}_{1,m}} \tau_{i,q}^{1,m} \log {\pi_q}^m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

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

Choix de (Q_1, Q_2) - Approche gloutonne

L'estimation de paramètres se fait à Q_1, Q_2 fixés, il faut donc déterminer les "meilleures" coordonnées. Nous maximisons un critère [2].

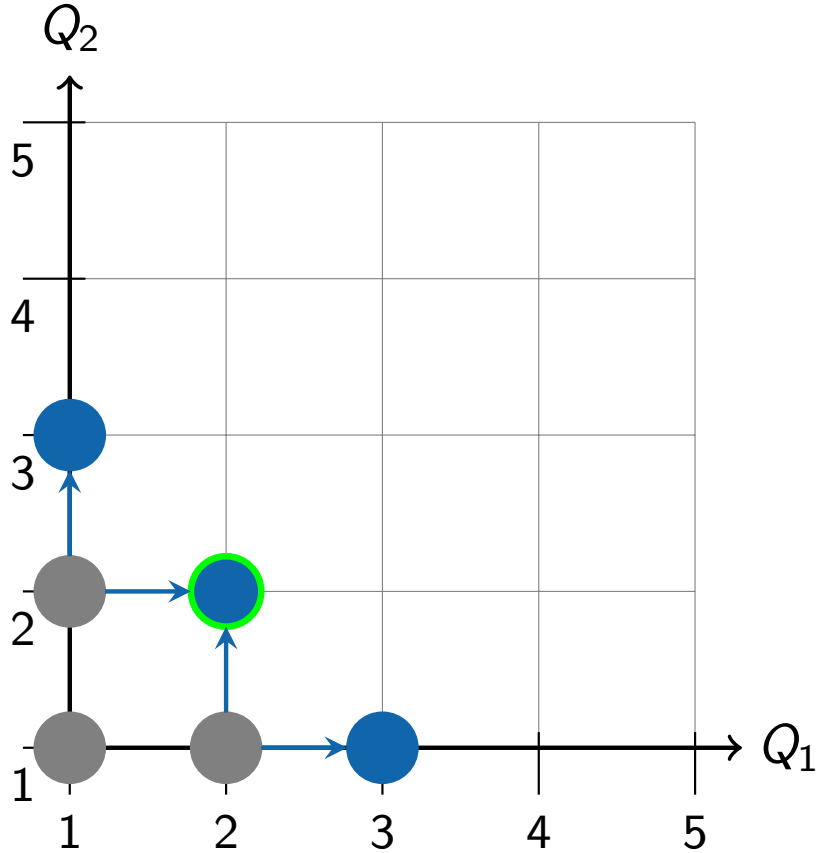


- Modèle initialisé :

2. *Bayesian Information Criterion - Like*, vraisemblance pénalisée en adaptant les formules de Chabert-Liddell et al., 2024

Choix de (Q_1, Q_2) - Approche gloutonne

L'estimation de paramètres se fait à Q_1, Q_2 fixés, il faut donc déterminer les "meilleures" coordonnées. Nous maximisons un critère [2].

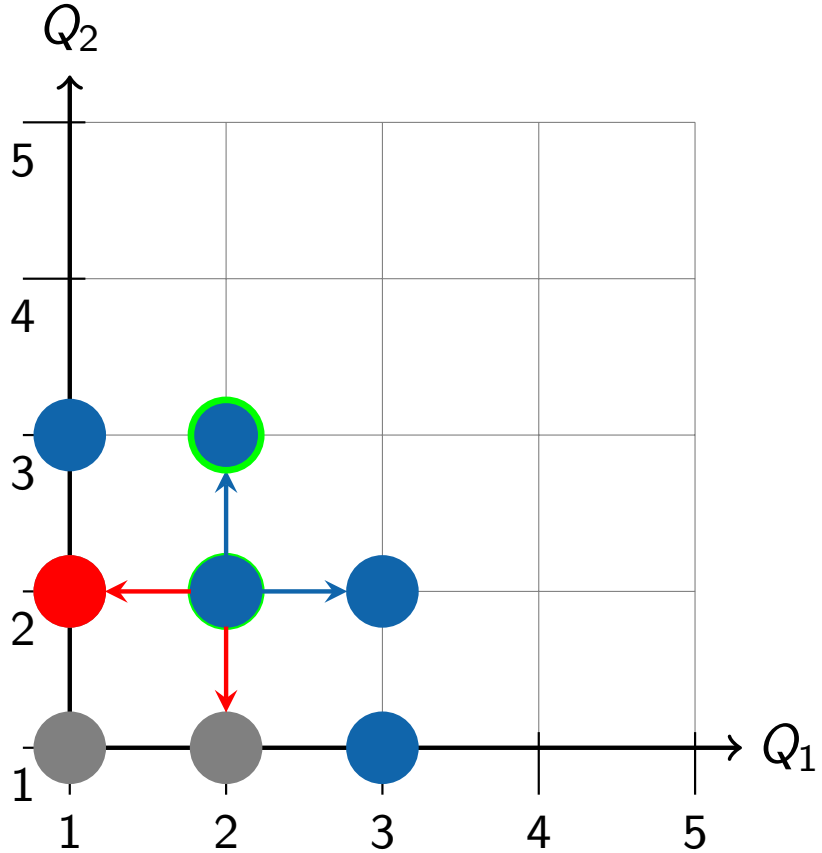


- Modèle initialisé :
- Modèle après *split* :
- Modèle maximisant le critère :

2. *Bayesian Information Criterion - Like*, vraisemblance pénalisée en adaptant les formules de Chabert-Liddell et al., 2024

Choix de (Q_1, Q_2) - Approche gloutonne

L'estimation de paramètres se fait à Q_1, Q_2 fixés, il faut donc déterminer les "meilleures" coordonnées. Nous maximisons un critère [2].



- Modèle initialisé :
- Modèle après *split* :
- Modèle maximisant le critère :
- Modèle après *merge* :

2. *Bayesian Information Criterion - Like*, vraisemblance pénalisée en adaptant les formules de Chabert-Liddell et al., 2024

Choix de (Q_1, Q_2) - Fenêtre glissante

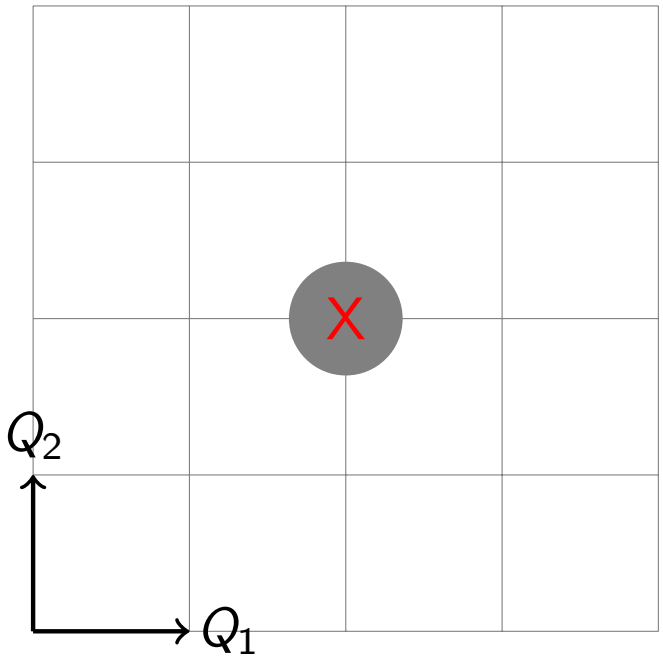


Figure – Fenêtre glissante

Choix de (Q_1, Q_2) - Fenêtre glissante

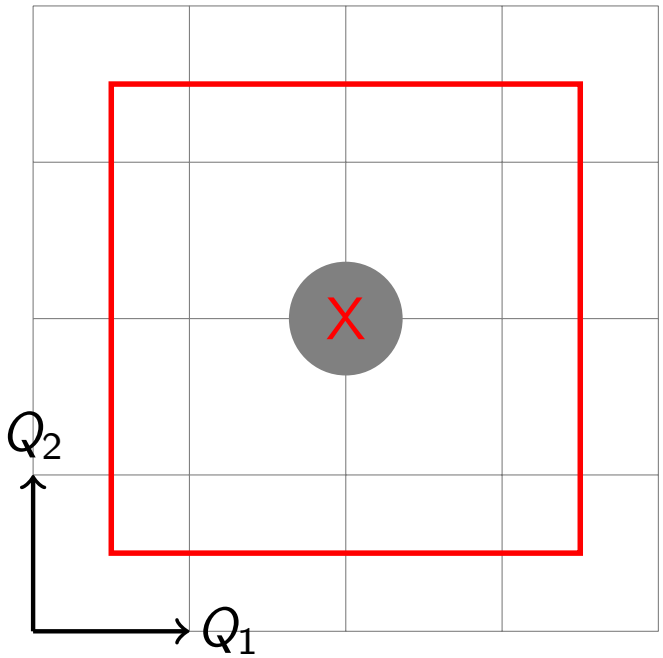


Figure – Fenêtre glissante

Choix de (Q_1, Q_2) - Fenêtre glissante

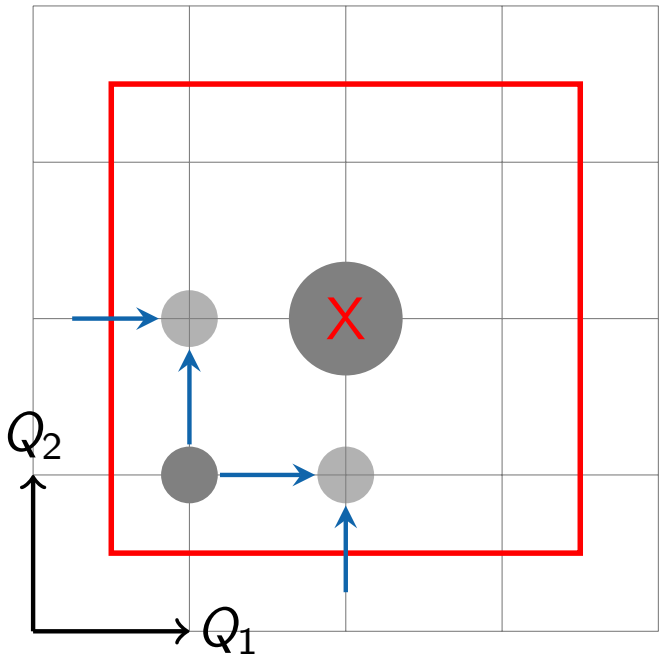


Figure – Fenêtre glissante

Initialisation du modèle si nécessaire

Choix de (Q_1, Q_2) - Fenêtre glissante

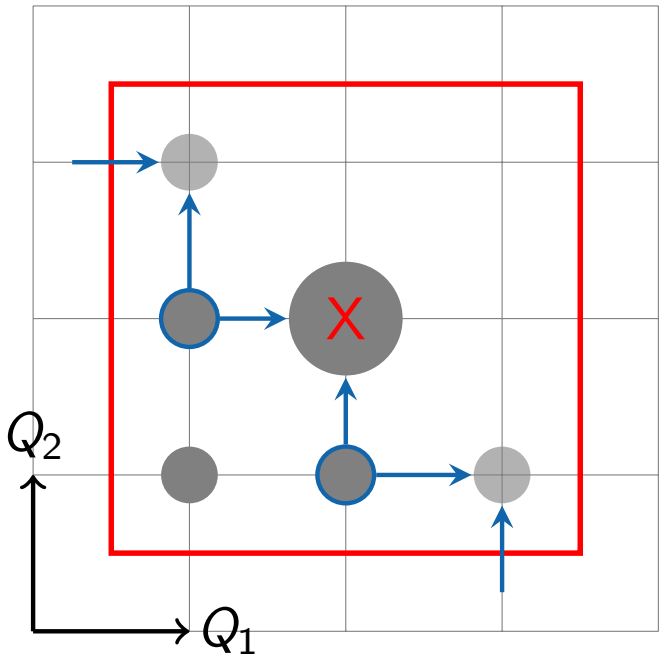


Figure – Fenêtre glissante

Choix de (Q_1, Q_2) - Fenêtre glissante

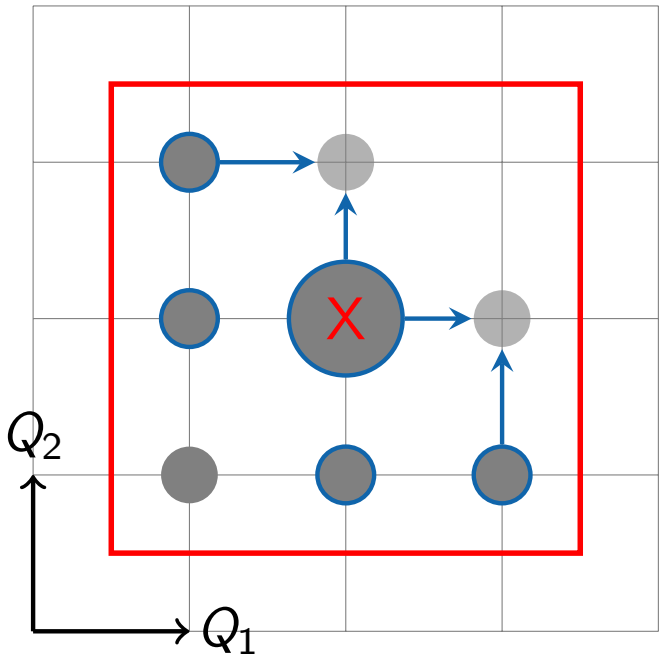


Figure – Fenêtre glissante

Choix de (Q_1, Q_2) - Fenêtre glissante

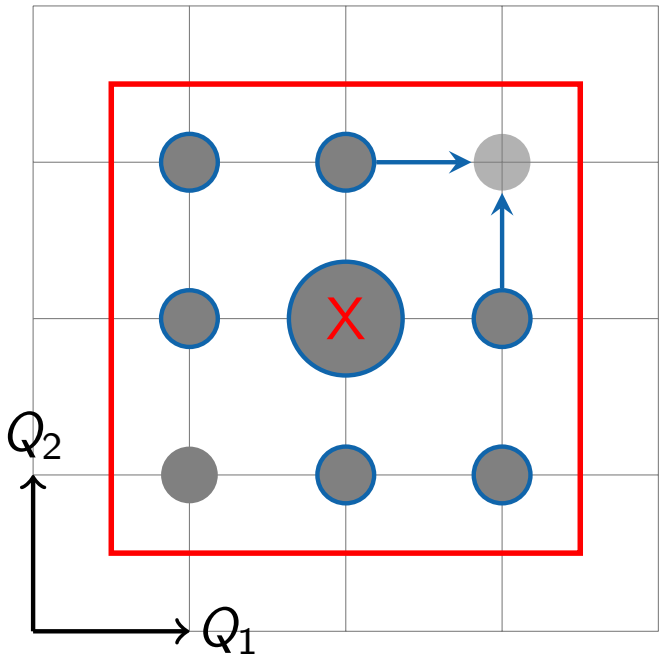


Figure – Fenêtre glissante

Choix de (Q_1, Q_2) - Fenêtre glissante

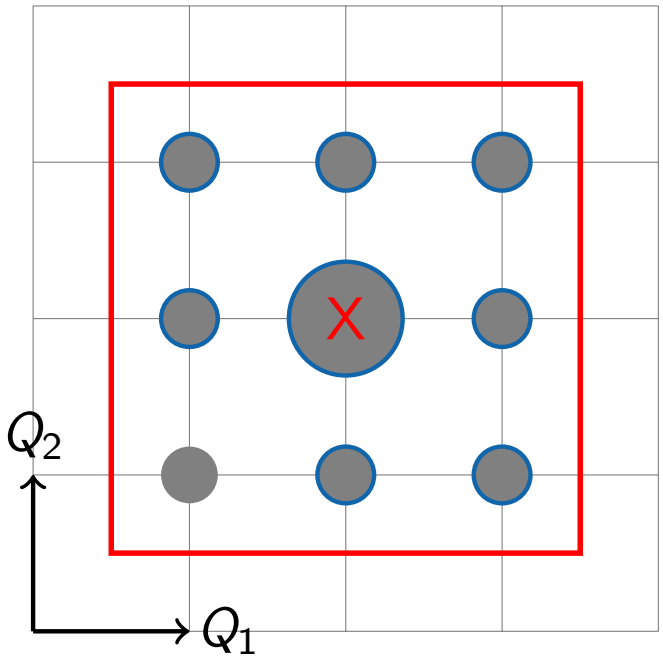


Figure – Fenêtre glissante

Choix de (Q_1, Q_2) - Fenêtre glissante

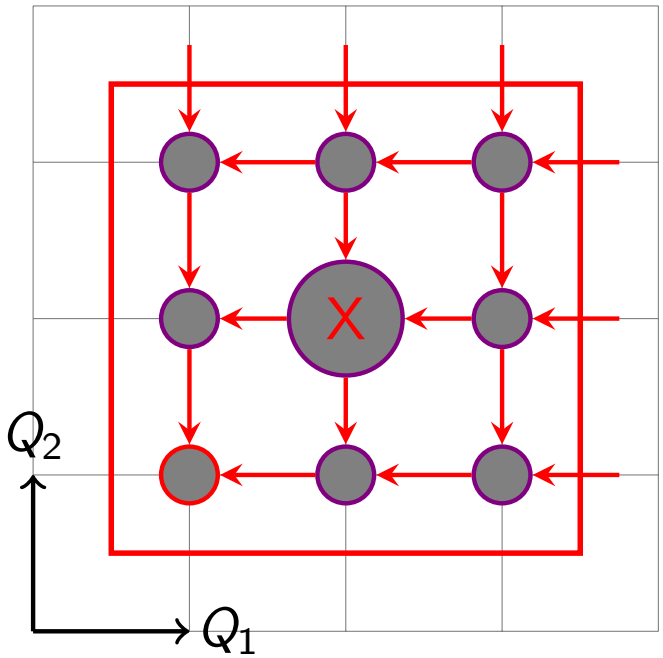


Figure – Fenêtre glissante

Choix de (Q_1, Q_2) - Fenêtre glissante

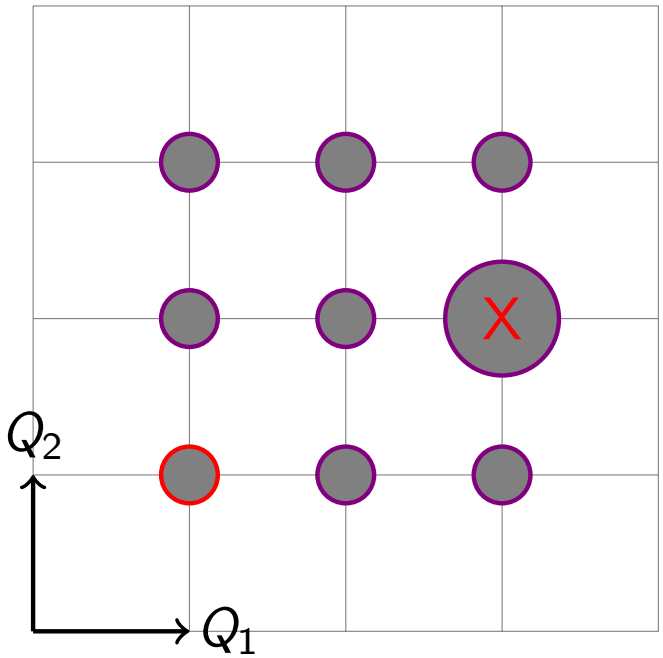


Figure – Fenêtre glissante

Localisation du nouveau mode

Choix de (Q_1, Q_2) - Fenêtre glissante

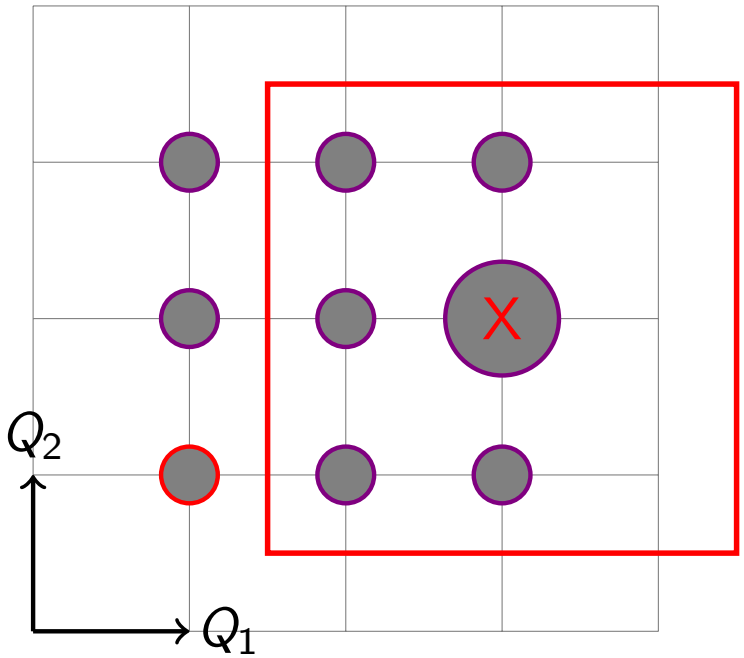
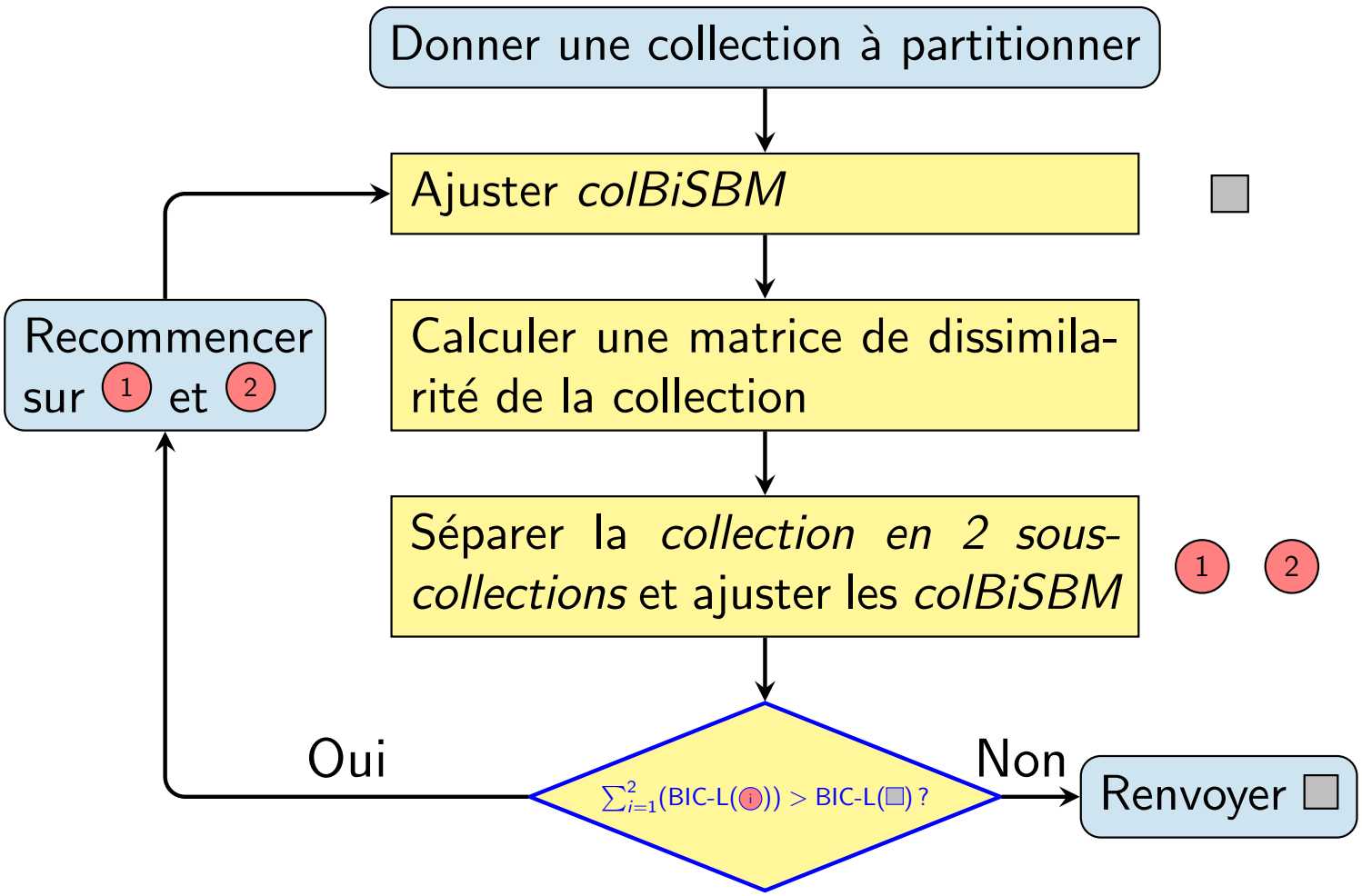


Figure – Fenêtre glissante

Déplacement sur le nouveau mode puis itération

Clustering de réseaux



Conclusion et perspectives

Capacités

- 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.

Perspectives

- Investiguer stabilité à la *graine*.
- Preuve d'identifiabilité du modèle $\pi\rho$.

Merci pour votre attention !

Bibliographie 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. (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