Détection de structures et *clustering* dans des réseaux bipartites

Séminaire des stagiaires

Louis Lacoste, encadré par Pierre Barbillon et Sophie Donnet

4 juillet 2024

Contexte écologique

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

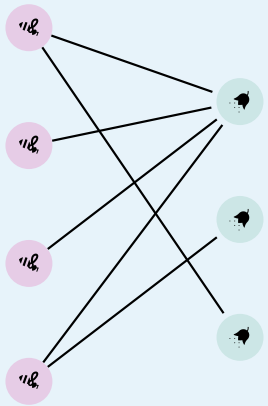


Figure 1 – Exemple d'un réseau plantes-pollinisateurs

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

Matrice d'adjacence associée

Contexte écologique

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

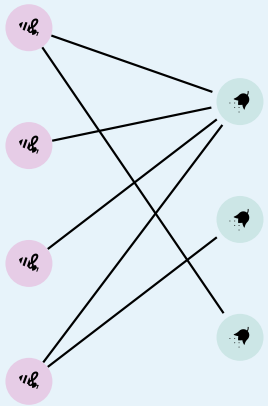


Figure 1 – Exemple d'un réseau plantes-pollinisateurs

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

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.

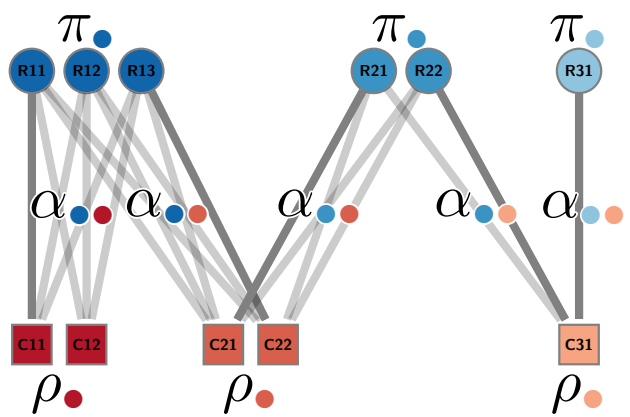


Figure 2 – Exemple de LBM [1]

Pour

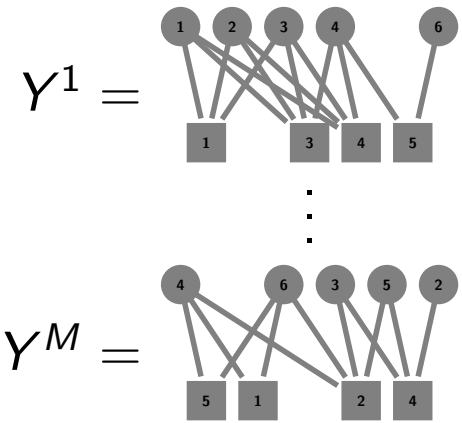
- $Q_1 = |\{\bullet, \bullet, \bullet\}|$ blocs fixés en ligne
- $Q_2 = |\{\bullet, \bullet, \bullet\}|$ blocs fixés en colonne

Paramètres

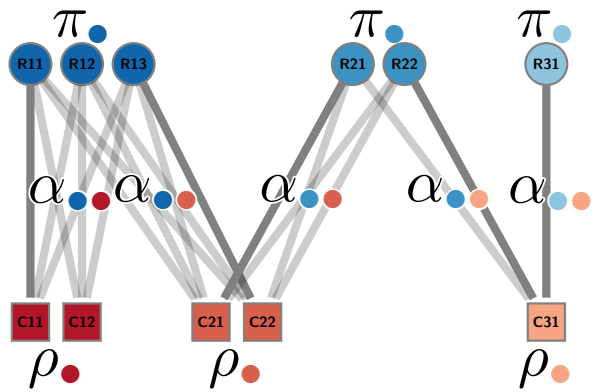
- $\pi_\bullet = \mathbb{P}(Z_i = \bullet)$ en ligne et $\rho_\bullet = \mathbb{P}(W_j = \bullet)$ en colonne
- $\alpha_{\bullet\bullet} = \mathbb{P}(X_{ij} = 1 | Z_i = \bullet, W_j = \bullet)$

1. Que j'appellerai par la suite BiSBM

Collections bipartites



$\overset{iid}{\sim}$

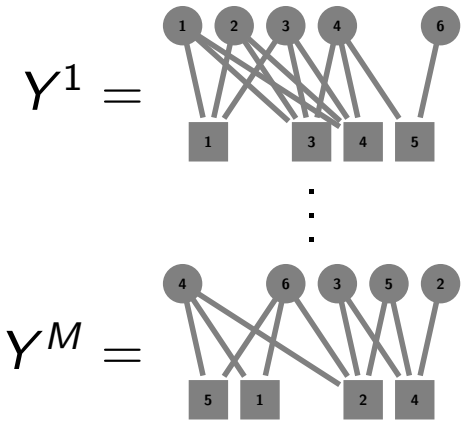


- $Q_1 = |\{\bullet, \bullet, \bullet\}|$ blocs fixés en ligne
- $Q_2 = |\{\bullet, \bullet, \bullet\}|$ blocs fixés en colonne

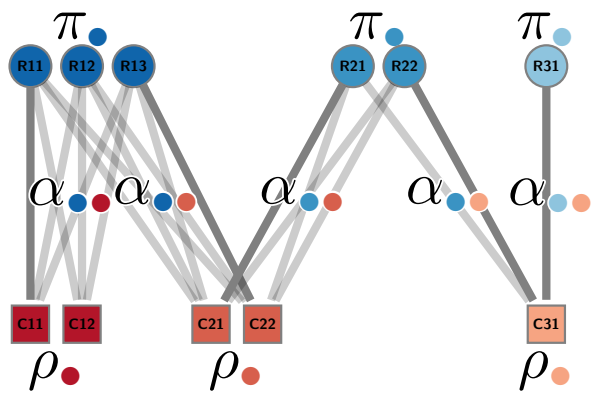
Paramètres

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



$\overset{iid}{\sim}$

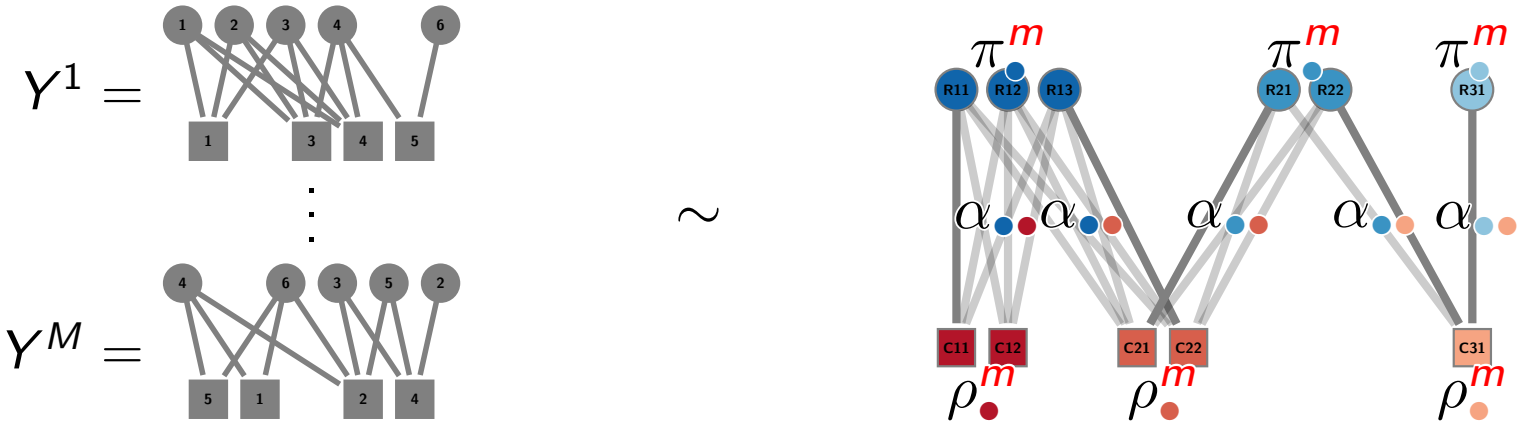


iid-colBiSBM

$\boldsymbol{\pi} = (\pi_1, \dots \pi_{Q_1})$ 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

En adaptant Chabert-Liddell et al., 2024 qui se base sur la méthode proposée par Daudin et al., 2008 utilisant l'algorithme *Variational EM*.

$$\ell(\mathbf{X};\boldsymbol{\theta}) \geq \sum_{m=1}^{M} \left(Q^m(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) + \mathcal{H}(\mathcal{R}_{\mathbf{X}^m,\boldsymbol{\theta}^{(t)}}(\mathbf{Z}^m,\mathbf{W}^m)) \right) =: J(\boldsymbol{\tau};\boldsymbol{\theta})$$

où $Q^m(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) = \mathrm{E}_{\mathbf{Z}^m,\mathbf{W}^m \sim \mathcal{R}_{\mathbf{X}^m,\boldsymbol{\theta}^{(t)}}(.)} [\log p(\mathbf{X}^m,\mathbf{Z}^m,\mathbf{W}^m|\boldsymbol{\theta})]$

▸ Formule développée EM variationnel

Approximation variationnelle

$\mathcal{R}_{\mathbf{X}^m,\boldsymbol{\theta}^{(t)}}(\mathbf{Z}^m,\mathbf{W}^m) = P(\mathbf{Z}^m|\mathbf{X}^m,\boldsymbol{\theta}^{(t)})P(\mathbf{W}^m|\mathbf{X}^m,\boldsymbol{\theta}^{(t)})$, c'est à dire avoir une indépendance lignes, colonnes.

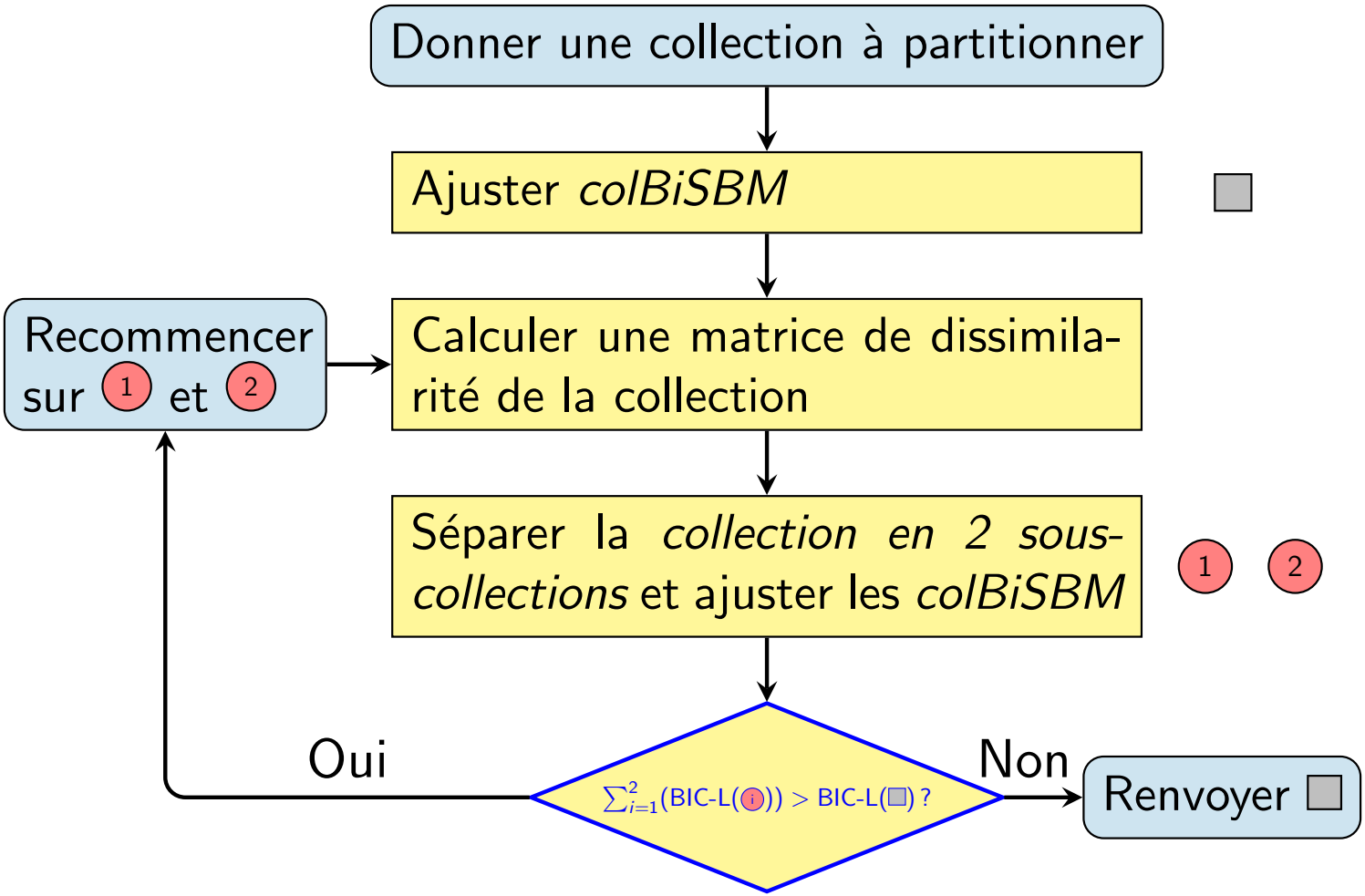
Problème de choix de (Q_1, Q_2)

L'estimation de paramètres se fait à Q_1, Q_2 blocs fixés, il faut donc déterminer les "meilleures" coordonnées. Nous maximisons un critère, le *Bayesian Information Criterion - Like* (BIC-L), de vraisemblance pénalisée en adaptant les formules de Chabert-Liddell et al., 2024.

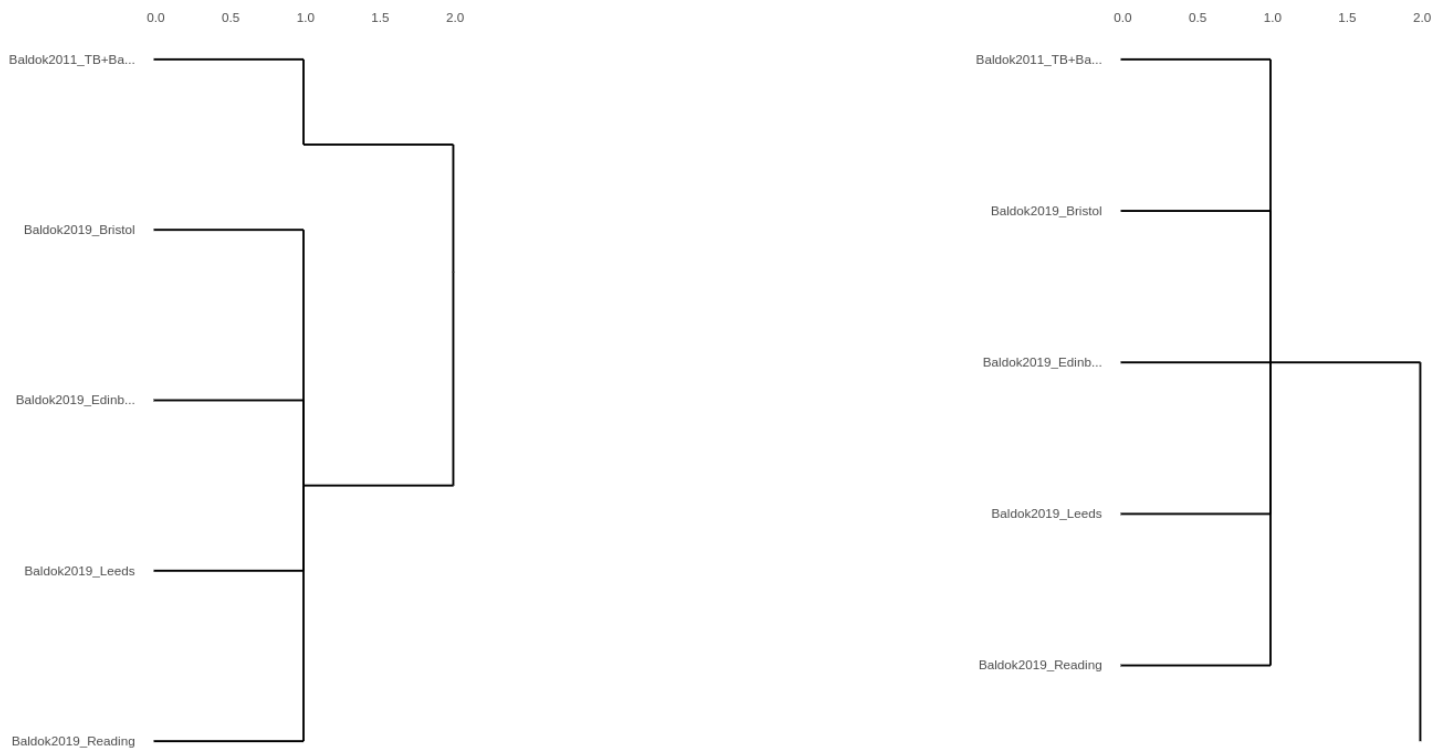
Problèmes de l'exploration

- Exploration de l'espace $\mathbb{N}^2$ coûteux, besoin d'une stratégie.
- Sensibilité aux initialisations et à l'aléatoire.

Clustering de réseaux



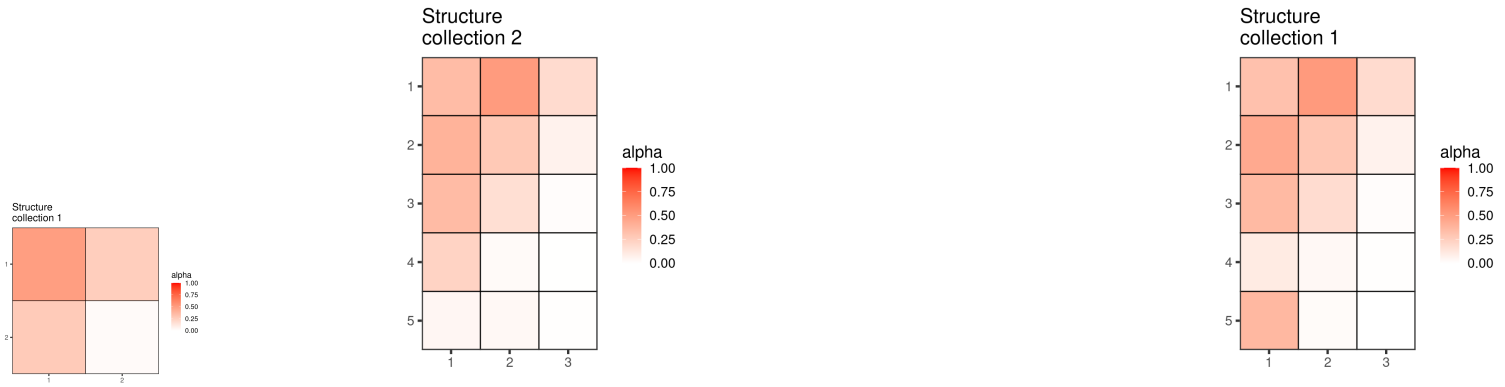
Application à Baldock et al., 2011, 2019 I



(a) Modèle *iid* (b) Modèle $\pi\rho$

Figure 3 – Partitionnement des réseaux de Baldock et al., 2011, 2019

Application à Baldock et al., 2011, 2019 II



(a) Modèle *iid* (b) Modèle $\pi\rho$

Figure 4 – Structures détectées pour les réseaux de Baldock et al., 2011, 2019

Conclusion et perspectives

Capacités

- 4 modèles dont 3 qui ont une flexibilité sur au moins une des dimensions (adaptabilité aux données).
- Détecter structures classiques et moins classique de façon agnostique.
- Partitionner un ensemble de réseaux selon leurs structures.

Perspectives

- Investiguer stabilité à la *graine*.
- Intégration au package `colSBM` et publication CRAN
- 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

Daudin, J.-J., Picard, F., & Robin, S. (2008).A mixture model for random graphs. *Stat Comput*, *18*(2), 173-183. https://doi.org/10.1007/s11222-007-9046-7

Baldock, K. C. R., Memmott, J., Ruiz-Guajardo, J. C., Roze, D., & Stone, G. N. (2011).Daily temporal structure in African savanna flower visitation networks and consequences for network sampling. *Ecology*, *92*(3), 687-698. https://doi.org/10.1890/10-1110.1

Bibliographie II

Baldock, K. C. R., Goddard, M. A., Hicks, D. M., Kunin, W. E., Mitschunas, N., Morse, H., Osgathorpe, L. M., Potts, S. G., Robertson, K. M., Scott, A. V., Staniczenko, P. P. A., Stone, G. N., Vaughan, I. P., & Memmott, J. (2019).A systems approach reveals urban pollinator hotspots and conservation opportunities. *Nat Ecol Evol*, *3*(3), 363-373. https://doi.org/10.1038/s41559-018-0769-y

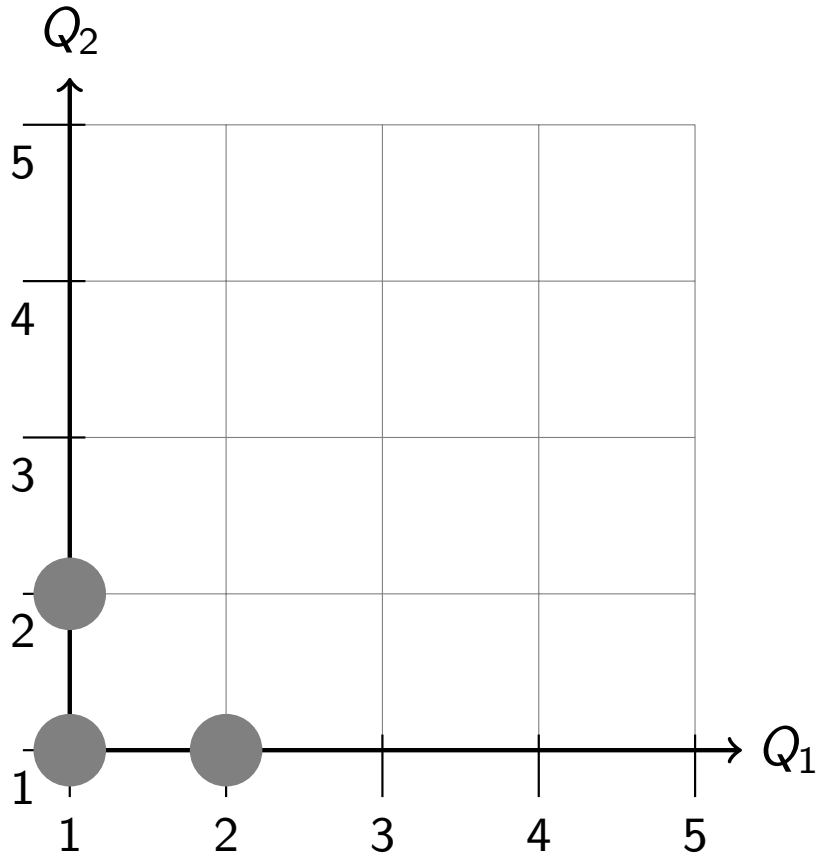
Formule développée EM variationnel

Back

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

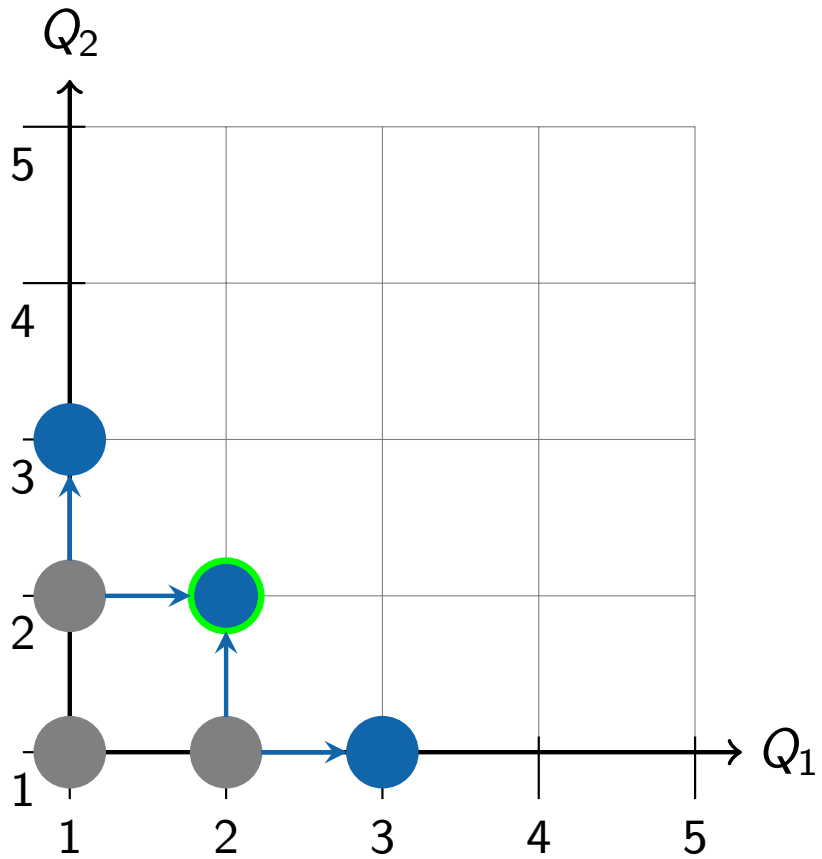
où ℓ désigne la log vraisemblance.

Choix de (Q_1, Q_2) - Approche gloutonne



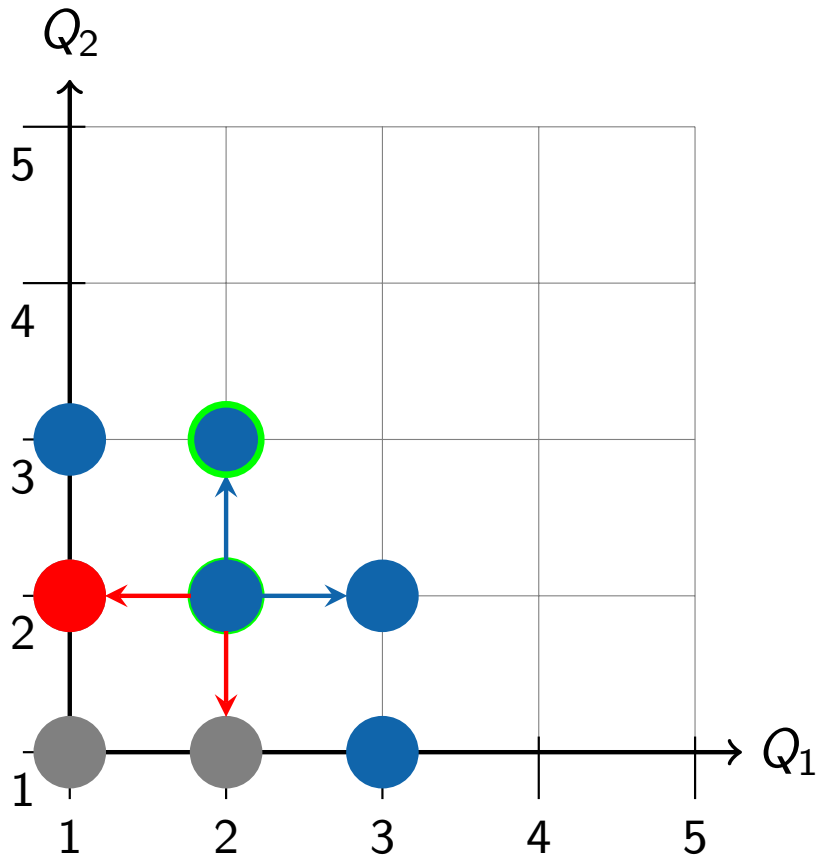
- Modèle initialisé :

Choix de (Q_1, Q_2) - Approche gloutonne



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

Choix de (Q_1, Q_2) - Approche gloutonne



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

Choix de (Q_1, Q_2) - Fenêtre glissante

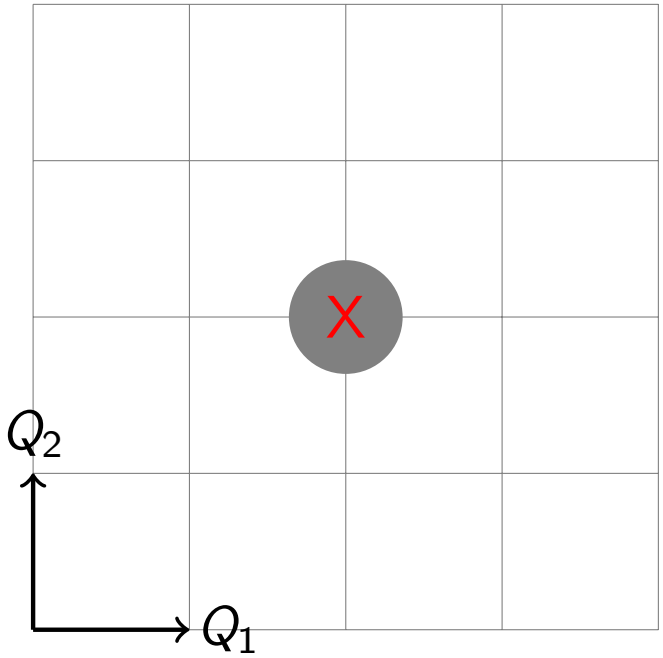


Figure 5 – Fenêtre glissante

Choix de (Q_1, Q_2) - Fenêtre glissante

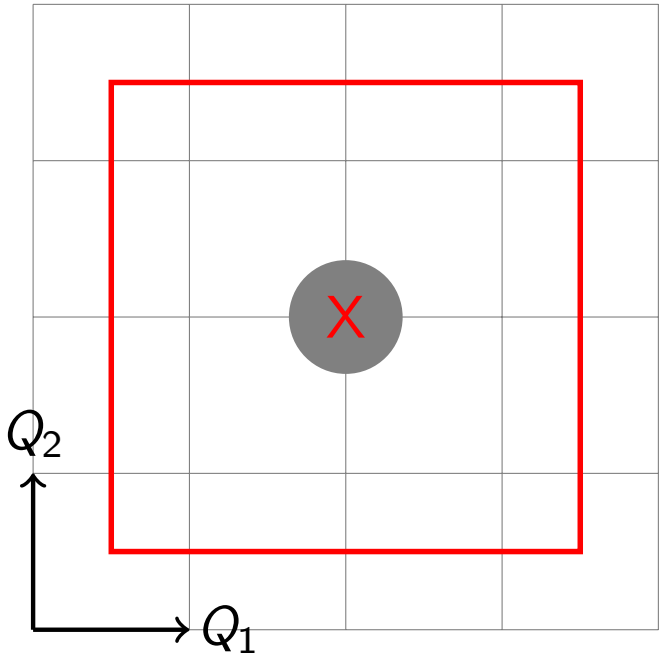


Figure 5 – Fenêtre glissante

Choix de (Q_1, Q_2) - Fenêtre glissante

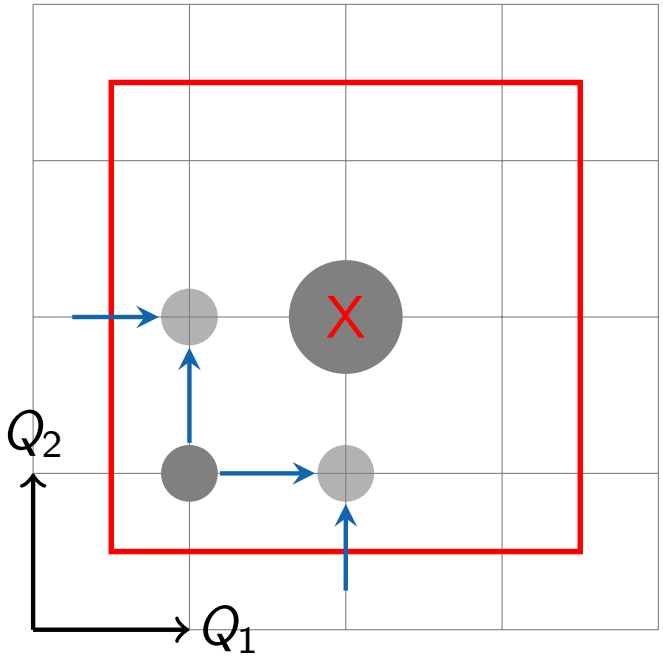


Figure 5 – Fenêtre glissante

Initialisation du modèle si nécessaire

Choix de (Q_1, Q_2) - Fenêtre glissante

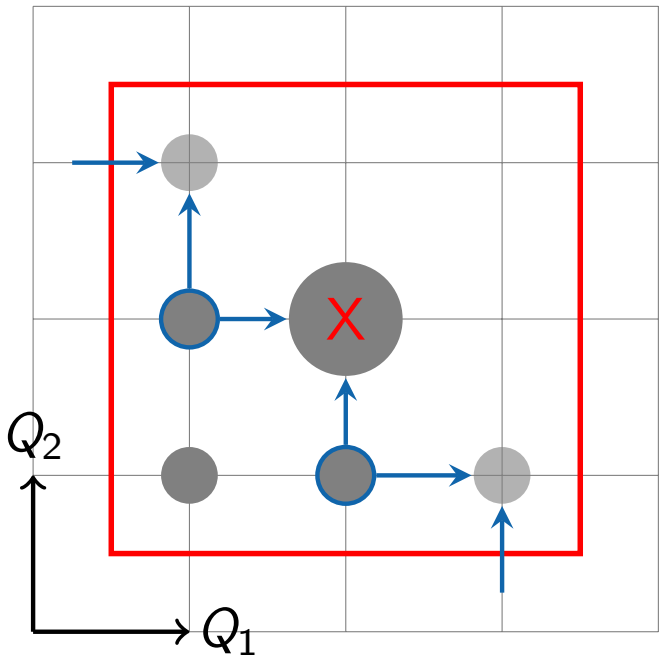


Figure 5 – Fenêtre glissante

Choix de (Q_1, Q_2) - Fenêtre glissante

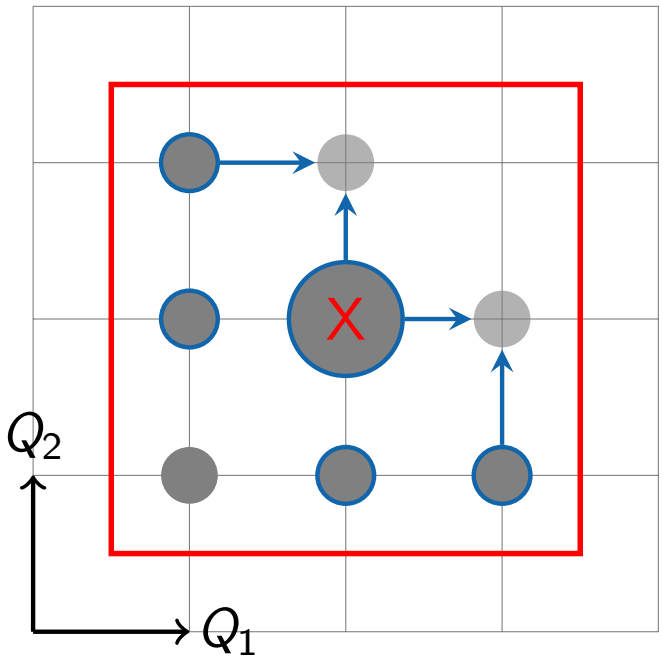


Figure 5 – Fenêtre glissante

Choix de (Q_1, Q_2) - Fenêtre glissante

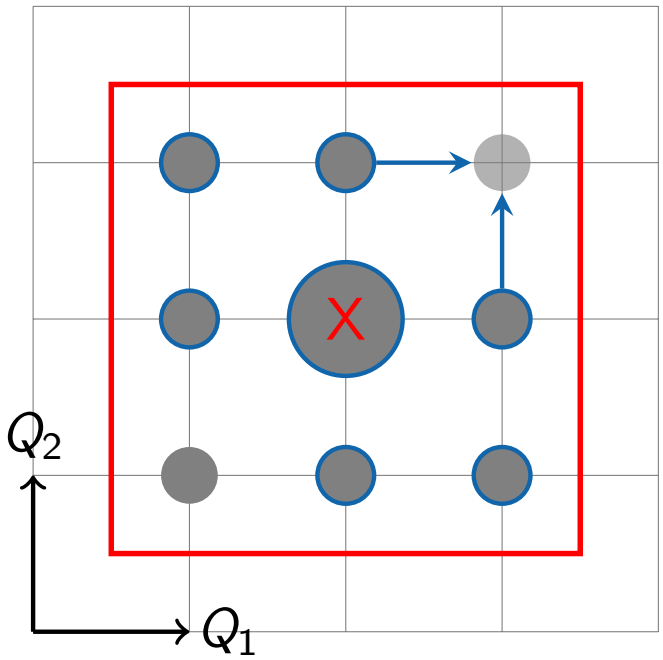


Figure 5 – Fenêtre glissante

Choix de (Q_1, Q_2) - Fenêtre glissante

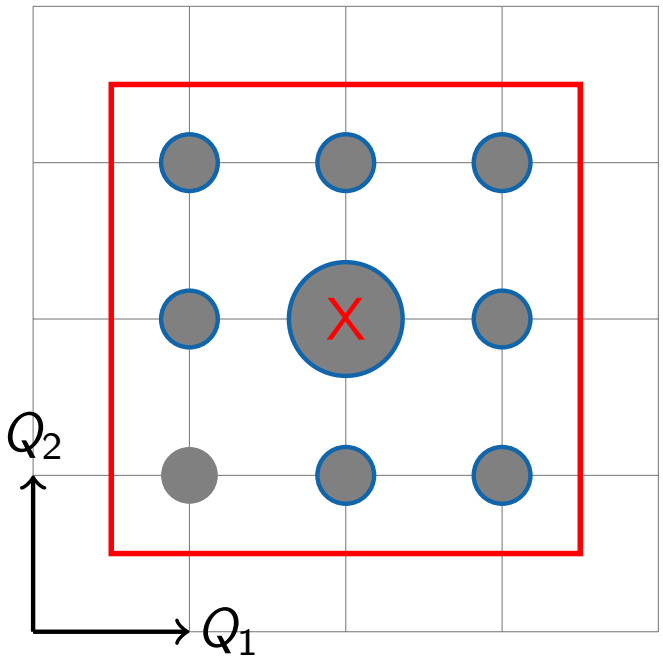


Figure 5 – Fenêtre glissante

Choix de (Q_1, Q_2) - Fenêtre glissante

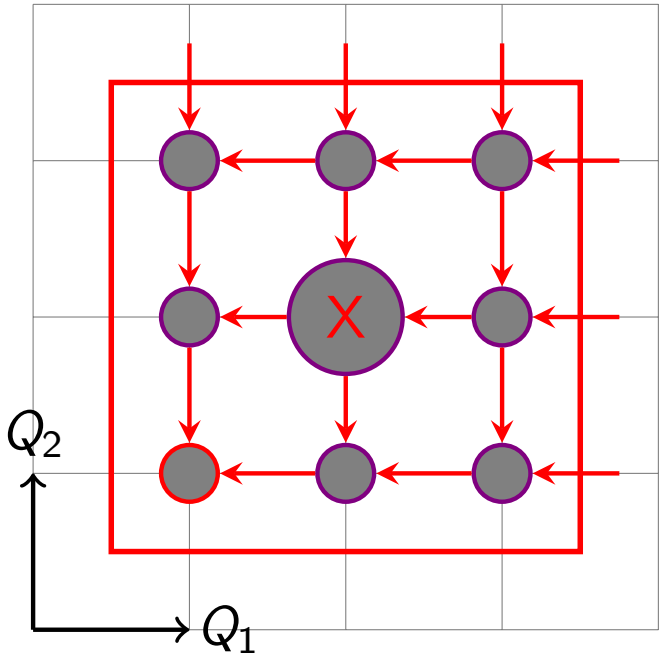


Figure 5 – Fenêtre glissante

Choix de (Q_1, Q_2) - Fenêtre glissante

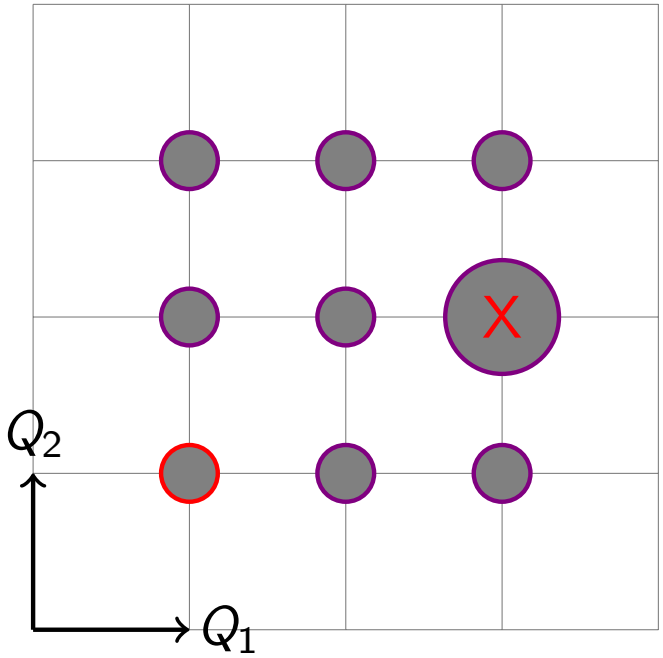


Figure 5 – Fenêtre glissante

Localisation du nouveau mode

Choix de (Q_1, Q_2) - Fenêtre glissante

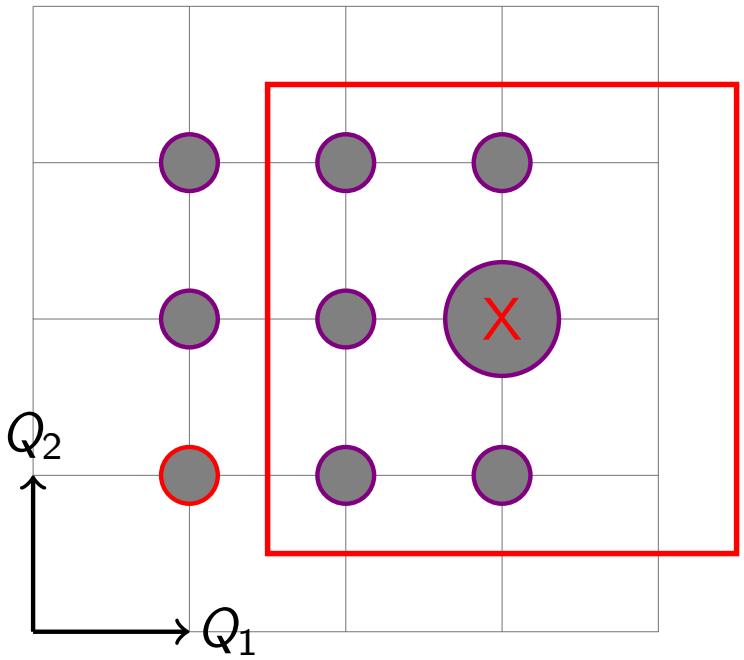


Figure 5 – Fenêtre glissante

Déplacement sur le nouveau mode puis itération

Bibliographie des annexes I