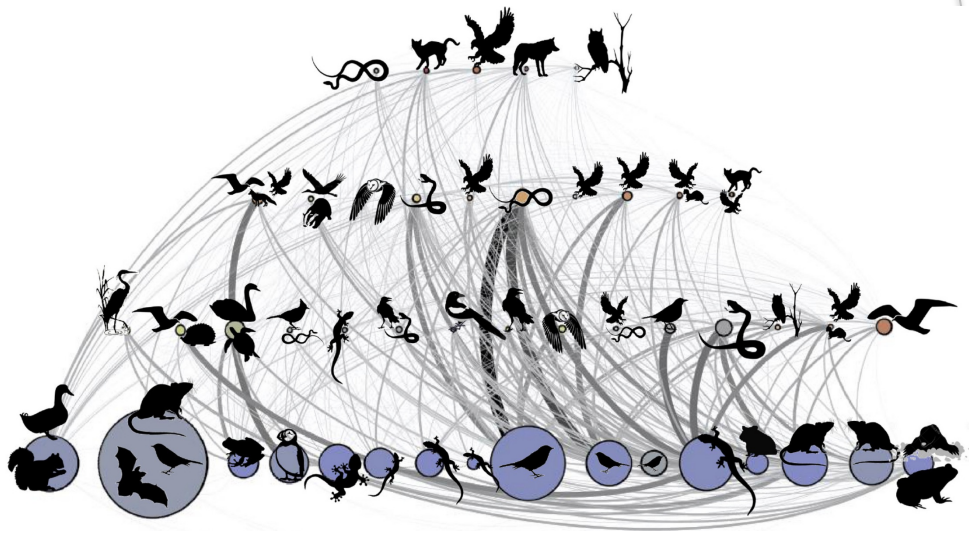
# PLN-Block : inférence de réseaux d'association à l'échelle de groupes d'espèces

Rochebrune – 25 mars 2024

*Jeanne Tous*

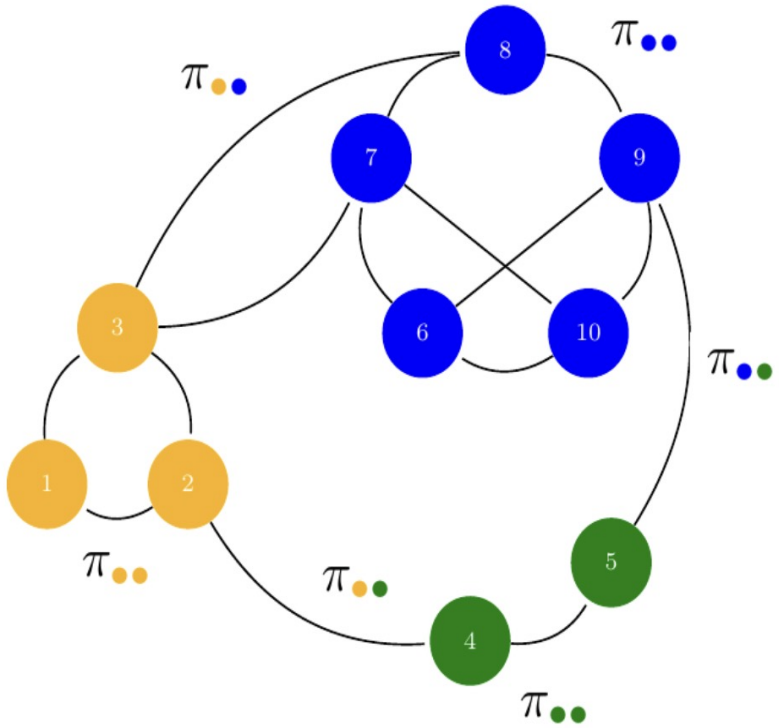
# Motivation : les réseaux d'association d'espèces

- Réseaux qui décrivent des associations entre espèces
- Association : « quelle corrélation reste-t-il entre les abondances de 2 espèces une fois qu'on a tenu compte des covariables ? »
- Les effets des changements environnementaux peuvent être plus importants sur les interactions que sur les espèces (Valiente-Banuet et al., 2015)
- Objets complexes à analyser → que faire :
  - Indicateurs agrégés ?
  - Les étudier à une échelle intermédiaire ?



Source : O'Connor et al., 2019

# Motivation : des groupes d'espèces ?




- Analyser les réseaux à une échelle intermédiaire de groupes d'espèces
- 1 groupe = ensemble d'espèces qui ont un positionnement similaire vis-à-vis des autres espèces dans le réseau (principe des SBM)
- Finalités imaginables :
  - Extraire des informations sur la résilience des écosystèmes
  - Identifier des espèces « indicatrices »

# Inférence de réseaux à partir de données d'abondance : PLN-network

Offset (efforts d'échantillonnage)

Matrice de covariance de Z

$$Y_{i,j}|Z_{i,j} \sim \mathcal{P}(\exp(x_i^T B_j + o_{i,j} + Z_{i,j}))$$

$$Z_i \sim \mathcal{N}(0, \Sigma)$$

Abondance de l'espèce j sur le site i

Covariables environnementales

Effets des covariables environnementales

Dépendances résiduelles entre espèces

- Reconstruction de réseaux :
  - On ajoute une contrainte de sparsité à la fonction de coût : - $|| \Sigma^{-1} ||_{l1,\ off}$
  - Association entre les espèces j et k ssi $\Sigma^{-1}_{j,k} \neq 0$

# Inférence de réseaux à partir de données d'abondance : PLN-network

$$Y_{i,j}|Z_{i,j} \sim \mathcal{P}(\exp(x_i^T B_j + o_{i,j} + Z_{i,j}))$$

Une valeur de Z par (espèce, site)

Traduire une structure latente de dépendance à l'échelle des ***groupes***

Une valeur de Z par (**groupe**, site)

# PLN-Block : ajout d'une structure latente de groupes

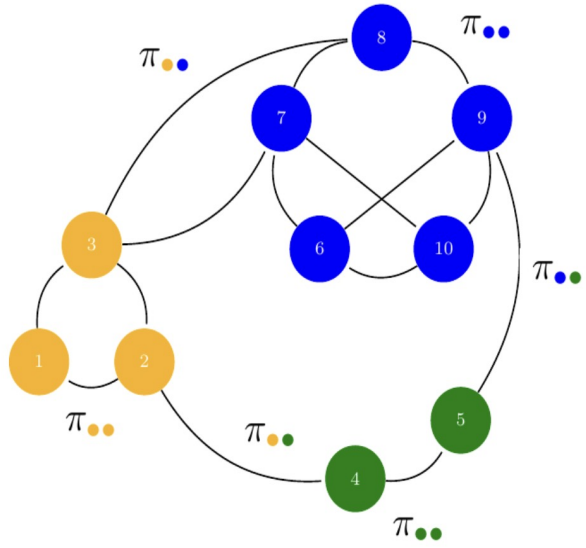
On a un modèle de mélange

Groupe de l'espèce j

Vaut 1 si l'espèce j appartient au groupe k, 0 sinon

$$Y_{i,j}|Z_{i,j}, C_j \sim \mathcal{P}(\exp(x_i^T B_j + o_{i,j} + \sum_{k=1}^{K} Z_{i,k} C_{jk}))$$

Dépendances résiduelles (on garde seulement le Zik qui correspond au groupe k de l'espèce j)




$$\mathbf{Z}_i \in \mathbb{R}^K \sim \mathcal{N}(0, \underbrace{\Sigma}_{K \times K})$$

Concerne maintenant les groupes (et non plus les espèces)

- De nouveau, on peut ajouter une contrainte de sparsité sur $\Sigma^{-1}$ pour retrouver un réseau.

# PLN-Block : ajout d'une structure latente de groupes

$$Y_{i,j}|Z_{i,j}, C_j \sim \mathcal{P}(\exp(x_i^T B_j + o_{i,j} + \sum_{k=1}^{K} Z_{i,k} C_{jk}))$$

- Ce modèle **ne tient pas compte des effets espèces-spécifiques dans la variance.**
- On rajoute une variable latente pour les intégrer.

# PLN-Block : ajout d'une structure latente de groupes

- La variable $W_i$ permet d'intégrer les effets espèces-spécifiques dans la variance.

$$Y_{i,j}|W_i, Z_i, C_j \sim \mathbf{P}(exp(o_{i,j} + W_i + \sum_{q=1}^{K} Z_{i,k} C_{jk}))$$

$$C_j \sim \mathcal{M}(1, (\alpha_k)_{1 \leq k \leq K})$$

$$Z_i \sim \mathcal{N}(0, \Sigma)$$

- Effets des groupes

$$W_i \sim \mathcal{N}(B^T X_i^T, D)$$

- Effets des covariables
- Effets espèces-spécifiques dans la variance : **D est diagonale**

# Optimisation : VEM

$$Y_{i,j}|W_i, Z_i, C_j \sim \mathbf{P}(exp(o_{i,j} + W_i + \sum_{q=1}^{K} Z_{i,k} C_{jk}))$$

$$C_j \sim \mathcal{M}(1, (\alpha_k)_{1 \leq k \leq K}) \quad Z_i \sim \mathcal{N}(0, \Sigma) \quad W_i \sim \mathcal{N}(B^T X_i^T, D)$$

- Approximation variationnelle : on cherche une approximation de $p_\theta(Z, W, C|Y)$ dans $\mathcal{Q} = \{\pi_{\psi_1}(Z)\pi_{\psi_2}(W)\pi_{\psi_3}(C)\}$ avec :
  - $\pi_{\psi_1}(Z_i) \sim \mathcal{N}(\mu_i, \Delta_i^2)$, avec $\Delta_i^2$ diagonale.
  - $\pi_{\psi_2}(W_i) \sim \mathcal{N}(M_i, S_i^2)$, avec $S_i^2$ diagonale.
  - $C_j \sim \mathcal{M}(1, (\tau_{q,j}))$ , avec : $\forall j, \sum_{q=1}^{Q} \tau_{q,j} = 1.$

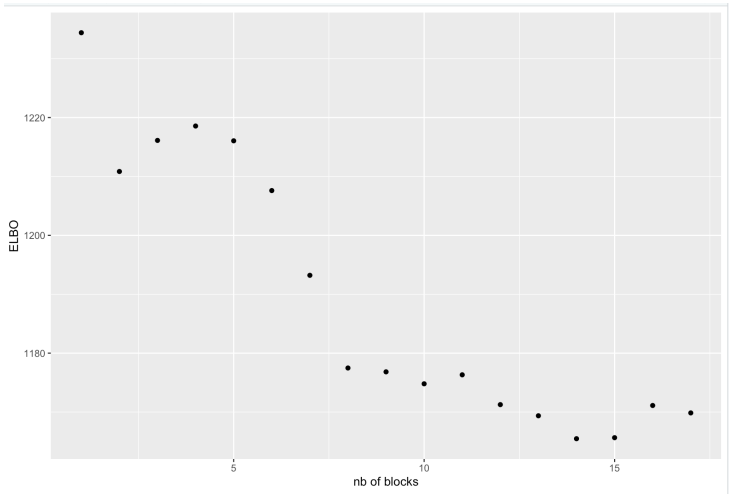
# Optimisation : VEM

$$Y_{i,j}|W_i, Z_i, C_j \sim \mathbf{P}(exp(o_{i,j} + W_i + \sum_{q=1}^{K} Z_{i,k}C_{jk}))$$

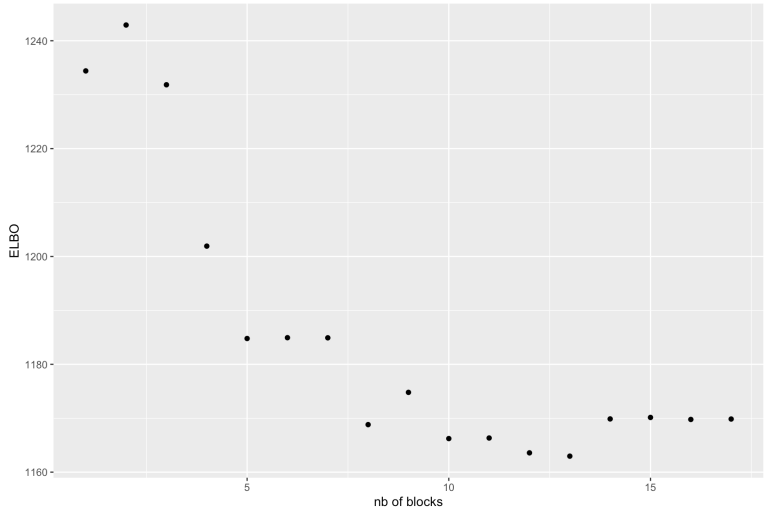
- On a des estimateurs explicites pour :
  - $B$: $\hat{B} = (X^T X)^{-1} X^T \mu$
  - La diagonale de D : $\hat{d} = \mu^T \odot \mu^T + \Delta^T \odot \Delta^T + (XB)^T \odot (XB)^T - 2\mu^T \odot (XB)^T$
  - $\Sigma$ : $\hat{\Sigma} = M^T M + S^{2^+}$ avec $S^{2^+} = \sum_{i=1}^{n} S_i^2$

# Optimisation : VEM

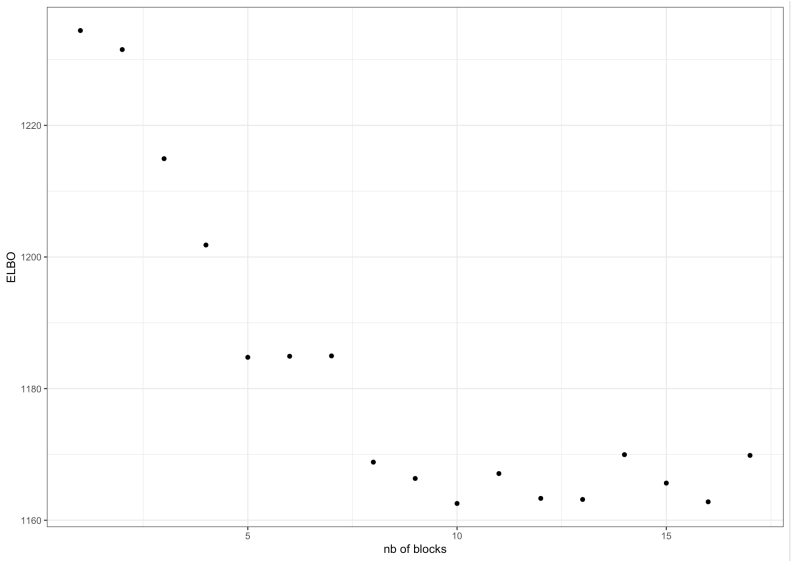
- Initialisation :
  - On part de PLN-network.
  - Pour les blocs : clustering sur les paramètres variationnels M
    - Ce n'est pas toujours le même clustering qui donne les meilleurs résultats.




**Initialisation clust of var**

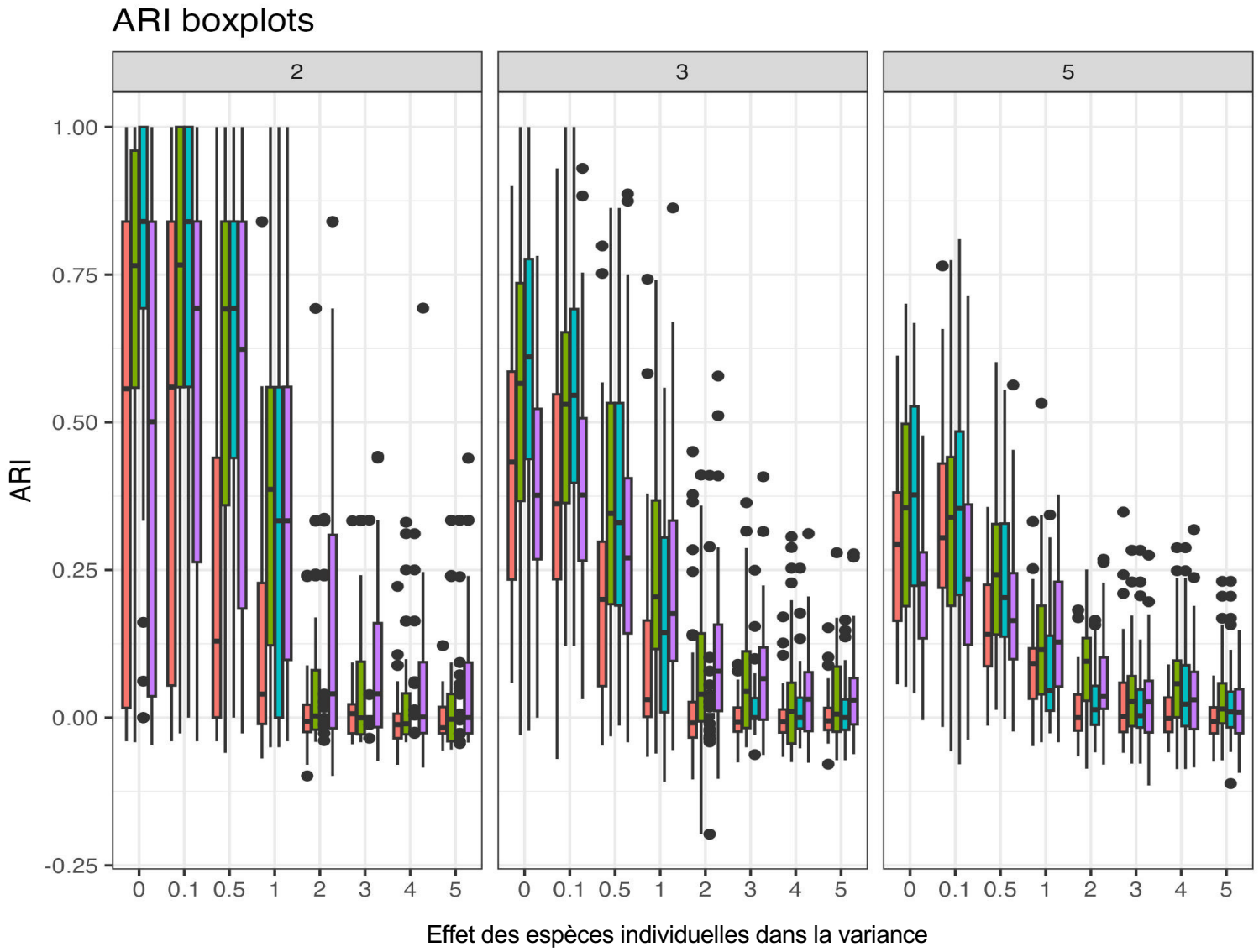



**Initialisation kmeans**

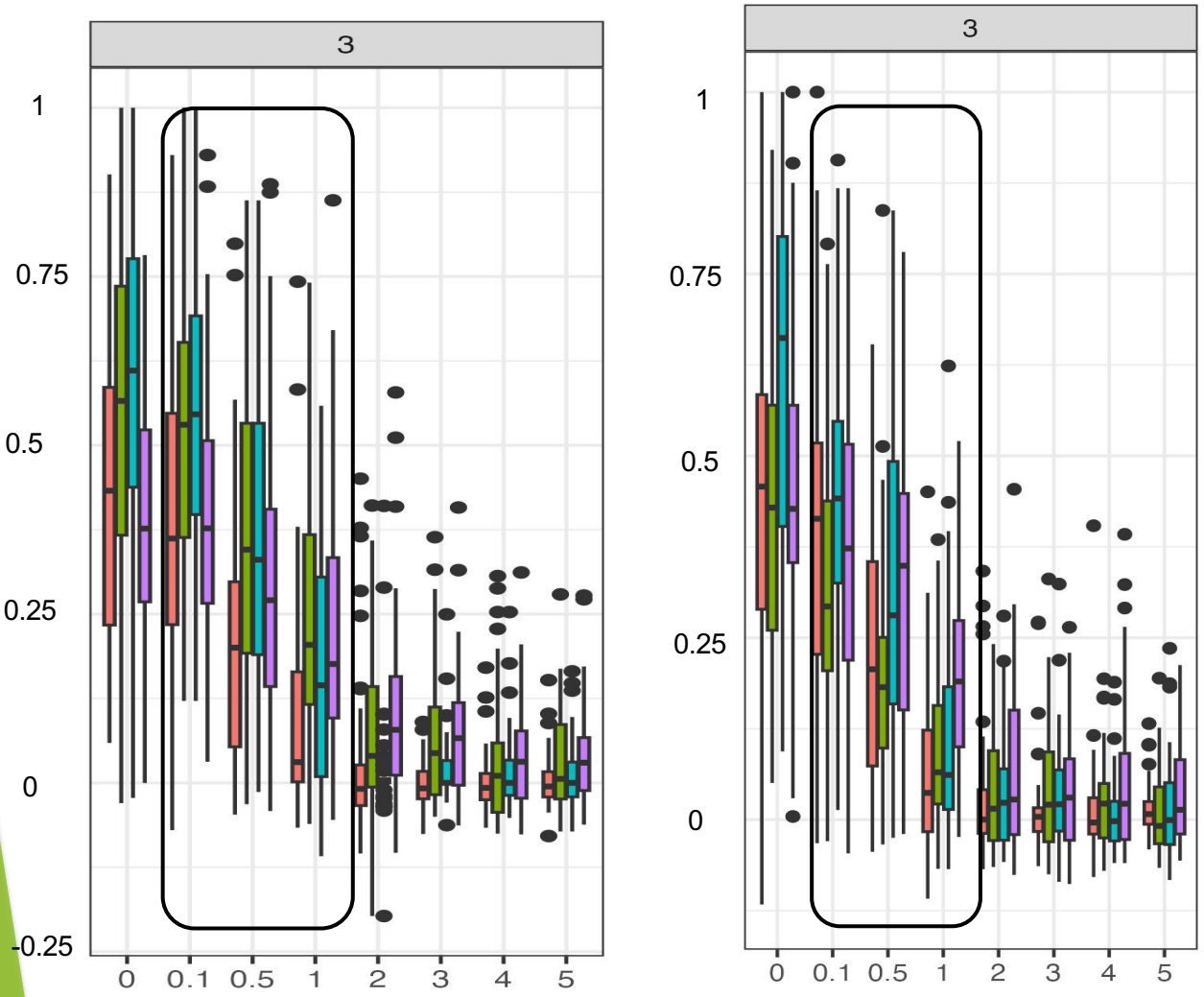



**Choix du meilleur clustering**

# Résultats : capacités de clustering du modèle



# Résultats : capacités de clustering du modèle




Effet des espèces individuelles dans la variance

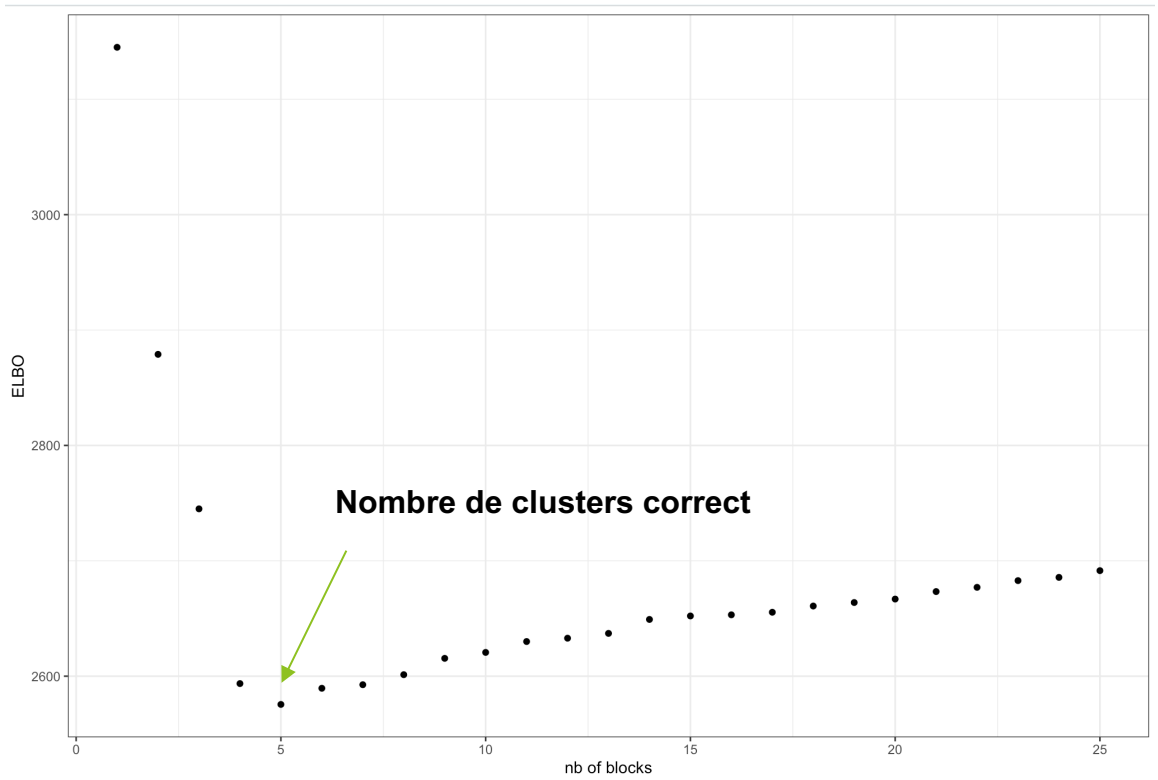
Pris en compte dans le modèle

Non pris en compte dans le modèle

- La prise en compte de l'effet espèce-spécifique dans la variance améliore sensiblement l'ARI dans une partie des cas simulés, quand cet effet n'est « ni trop faible, ni trop important ».

# Question du choix du nombre de groupes

- ELBO sur données simulées en fonction du nombre de blocs demandé.



# La suite (1) : tester un modèle « normal-block »

- Repartir d'un modèle plus simple, **normal-block**, pour des données continues

$$Y_{i,j} = x_i^T B_j + \sum_{k=1}^{K} C_{j,k} W_{i,k} + \epsilon_{i,j}$$

Avec

$$C_{j,k} = \mathbb{1}(j \in k) \qquad W_i \sim \mathcal{N}(0, \Sigma_K)$$

$$C_j \sim \mathcal{M}(\pi) \qquad \epsilon_i \sim \mathcal{N}(0, D)$$

- Inférence d'une structure de groupes pour des données continues
- Un modèle plus simple à optimiser / améliorer / modifier
- Tester sur des données continues réelles avant de repasser à PLN-block

# La suite (2)

- Trouver des données réelles dans lesquelles on peut espérer trouver des effets des groupes (en lien avec la taxonomie ?)

- Intégration d'*a priori*
  - Ajouts de poids différents selon les espèces pour la sparsité
  - Imposer des contraintes sur les effets des covariables selon les espèces
  - Clustering semi-supervisé : placer d'office certaines espèces dans des groupes spécifiques et regarder comment le clustering s'effectue sur les autres.

- Intégration d'une dimension temporelle ?