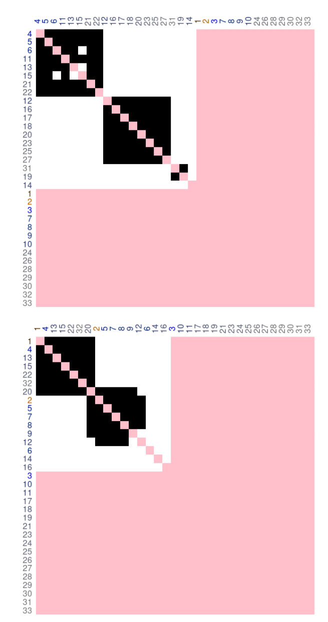
Recovering a persistent underlying interaction network from multiple networks with stochastic block structure

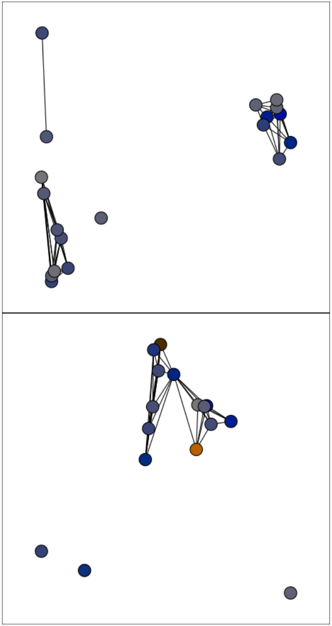
aka Maman les p'tits bateaux qui vont sur l'eaux font-ils du zèle

Saint-Clair Chabert-Liddell
Joint work with S. Mahévas and N. Bez

March 28, 2024
Statistiques au sommet de Rochebrune

Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA-Paris


4 5 6 11 13 15 21 22 12 16 17 18 20 23 25 27 31 19 14 1 2 3 7 8 9 10 24 26 28 29 30 32 33
4 5 6 11 13 15 21 22 12 16 17 18 20 23 25 27 31 19 14 1 2 3 7 8 9 10 24 26 28 29 30 32 33
1 4 13 15 22 32 20 2 5 7 8 9 12 6 14 16 3 10 11 17 18 19 21 23 24 25 26 27 28 29 30 31 33
1 4 13 15 22 32 20 2 5 7 8 9 12 6 14 16 3 10 11 17 18 19 21 23 24 25 26 27 28 29 30 31 33



Motivation

Data

- $(\mathbf{X}^l)_{l \in \mathcal{L} = \{1,\ldots,L\}}$ over a common set $\mathcal{N} = \{1, \ldots, N\}$ of nodes
- $\mathcal{N}^l \subset \mathcal{N}$ observed nodes

$$\mathbf{X}^l_{ij} = \mathbf{X}^l_{ji} = \begin{cases} \omega^l_{ij} & \text{if } (i,j) \in \mathcal{N}^l \times \mathcal{N}^l, i \neq j, \\ \mathtt{NA} & \text{otherwise.} \end{cases}$$

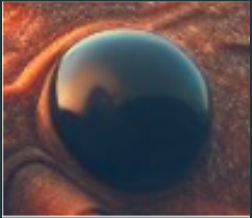
Objectives

- Separate persistent interactions (collaboration) from noise (weather, ressources. . .)
- Networks (fishing day) clustering

Idea

- A mixture of observation processes with stochastic block structures

Modeling



Simple noisy realization model

$(A_{ij})_{(i,j)\subset\mathcal{N}\times\mathcal{N}}$ latent binary network of persistent interactions

- Observing a missing edge

$$(1 - X_{ij}^l)|A_{ij} = 1 \sim \mathcal{B}(1 - \alpha)$$

- Observing a spurious edge

$$X_{ij}^l = 1|A_{ij} = 0 \sim \mathcal{B}(\beta)$$

Two ways to extend this model

Newman (2018)

Network model

Any model assuming (conditional) independence between edges, e.g.

- Erdős-Rényi (ER):

$$A_{ij} \sim \mathcal{B}(\gamma^A)$$

- Stochastic Block Model:
  Each node in one of Q^A blocks:

$$Z_i^A \sim \mathcal{M}\left(1, \pi^A = (\pi_1^A, \ldots, \pi_{Q_A}^A)\right).$$

  Edges are independent given group memberships:

$$A_{ij} | Z_{iq}^A Z_{jr}^A = 1 \sim \mathcal{B}(\gamma_{qr}^A)$$

Le et al. (2018); Young et al. (2021)

A mixture of observation processes

Each network is the realization of one of K observation processes.

$$W_l \sim \mathcal{M}\big(1, \rho = (\rho_1, \ldots, \rho_K)\big).$$

Then,

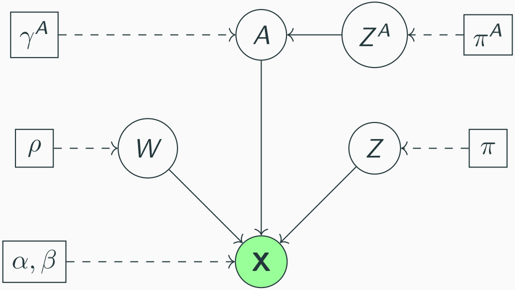
$$X_{ij}^l | W_{lk} = 1, A_{ij} \sim \begin{cases} \mathcal{B}(\alpha^k) & \text{if } A_{ij} = 1, \\ \mathcal{B}(\beta^k) & \text{if } A_{ij} = 0. \end{cases}$$

Newman (2018); Mantziou et al. (2024)

A mixture of observation processes with stochastic block structure

$$Z_i^k \sim \mathcal{M}\big(1, \pi^k = (\pi_1^k, \ldots, \pi_{Q_k}^k)\big).$$

$$\mathbf{X}_{ij}^l | W_{lk} = 1, Z_{iq}^k Z_{jr}^k = 1, A_{ij} \sim \begin{cases} \mathcal{B}(\alpha_{qr}^k) & \text{if } A_{ij} = 1, \\ \mathcal{B}(\beta_{qr}^k) & \text{if } A_{ij} = 0. \end{cases}$$

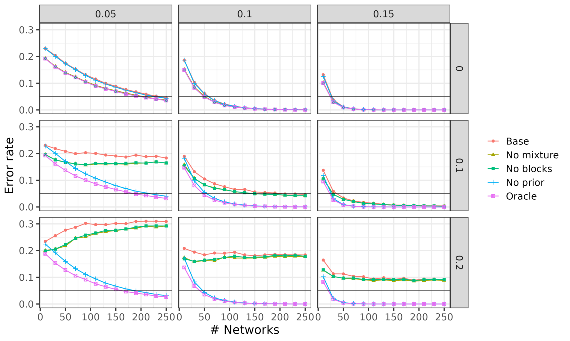


Oracle analysis

Compute MLE $\hat{A}_{ij}$ knowing (Z, W, Z^A) and $\theta = \{\pi^A, \gamma^A, \rho, \pi, \alpha, \beta\}$:

$$\hat{A}_{ij} = \mathbf{1}\left\{\log \frac{\phi_{\mathcal{B}}(X_{ij}^l, \alpha_{ij}^l)\gamma_{ij}^A}{\phi_{\mathcal{B}}(X_{ij}^l, \beta_{ij}^l)(1-\gamma_{ij}^A)} > 0\right\}$$

Monte Carlo estimates of: $err(A, \hat{A}) = \mathbb{E}\left[\frac{\sum_{i>j}(1-A_{ij})\hat{A}_{ij} + A_{ij}(1-\hat{A}_{ij})}{\binom{N}{2}}\right]$



Submodels and two problems with identifiability

Rewrite $X_{ij} = A_{ij}B_{ij} + (1 - A_{ij})C_{ij}$, with $B_{ij} \sim \mathcal{B}(\alpha_{ij})$ and $C_{ij} \sim \mathcal{B}(\beta_{ij})$

Submodels

AND $\beta \equiv 0$, $X_{ij} = A_{ij}B_{ij}$

OR $\alpha \equiv 1$, $X_{ij} = \max(A_{ij}, C_{ij})$

Identifiability

1. $B \leftrightarrow C$ and $A \leftrightarrow 1 - A$
   Solution: collaborating vessels are closer on average $\alpha > \beta$.
2. $A \leftrightarrow (B, C)$
   Inverting persistent interactions and observation process. (Dealt with mixture (B^k, C^k) in practice?).

Vallès-Català et al. (2016)

Model extensions

Missing nodes depends on the observation blocks

$$R_i^l | W_{kl} = 1, Z_{iq}^k = 1 \sim \mathcal{B}(\mu_{lq})$$

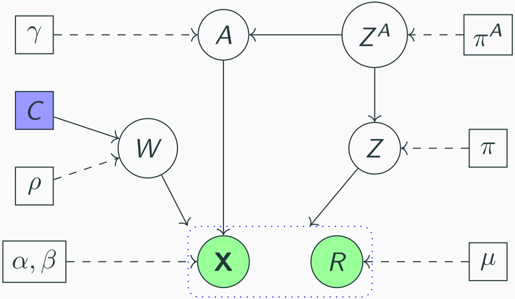
Hierarchical block memberships Blocks of observation processes depend on the persistent blocks

$$Z_i^k | Z_{iq_A}^A = 1 \sim \mathcal{M}(\pi_{q_A 1}^k, \ldots, \pi_{q_A Q_k}^k)$$

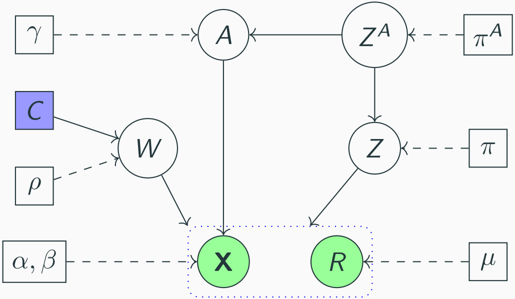
Network covariates Observation process clustering depends on discrete covariates (year, site...)

$$W_l | C_l = c \sim \mathcal{M}(\rho_{c1}, \ldots, \rho_{cK})$$

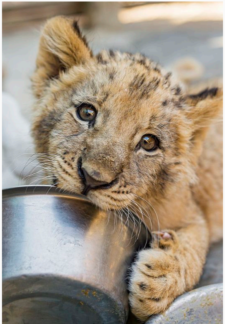
Pourquoi est-ce que le lion n'a pas pu terminer son repas ?



Pourquoi est-ce que le lion n'a pas pu terminer son repas ?



Parce qu'il était rassasié !
(rassasié = dag in Dutch)



Inference

Stochastic variational inference
Minimize $-ELBO(\theta, q) \geq -\log p(X; \theta)$

- Automatic differentiation (`pytorch`)
- Mini-batches of $B_s \leq L$ networks
- Reparameterize in $(\mathbb{R}^{d_\theta}, \mathbb{R}^{d_q})$
- Estimate $q(A)$ alternatively
- Pyramidal training (train the best models after nb_{init} epochs)
- Peak the loss with $Bs = L$

Model selection
With Integrated Classification Likelihood (ICL) criterion

- Select $(\widehat{K}, \widehat{Q})$
- Refine with model extensions and prior network model

Leger (2023)

Application

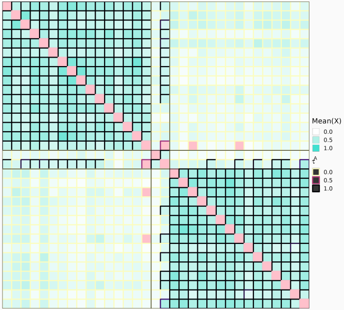
Material and methods

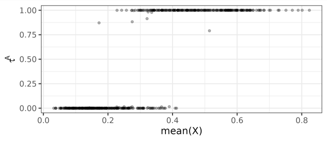
1071 networks with missing nodes ($|\mathcal{N}^I| \in [10, 33)$)

- Set $Q_k = Q$ for all $k \in \{1, \ldots, K\}$
- For selected $(\hat{K}, \hat{Q}, \hat{Q}_A)$:
    - Hierarchical block memberships
    - Network covariates: month, quarter, year, quarter$\times$year

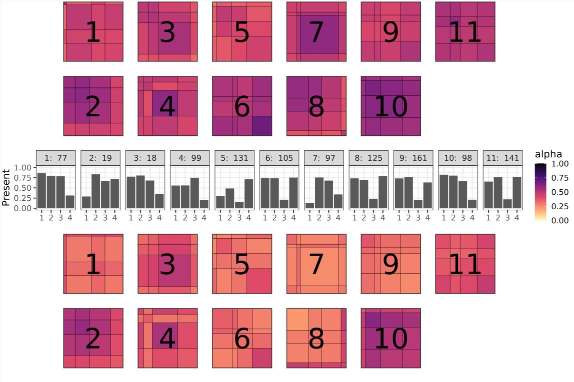
Persistent interactions

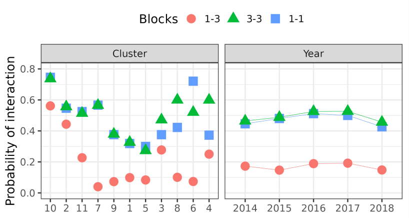
- $\widehat{Q} = 11$, $\widehat{W} = 4$, $\widehat{Q^A} = 3$
- Year covariates and hierarchical block memberships
- 2 fully connected communities, 2 residual vessels



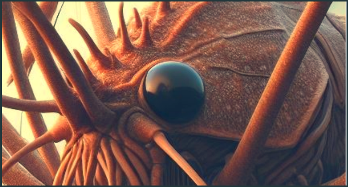


Observation processes





Nephrops LPUE



Landings per unit effort (LPUE)

- C_{atch}PUE proportional to biomass
- LPUE approximates CPUE

Standardized LPUE, GLM with gamma family and log link:

Vessel characteristics size, power, fishing gear

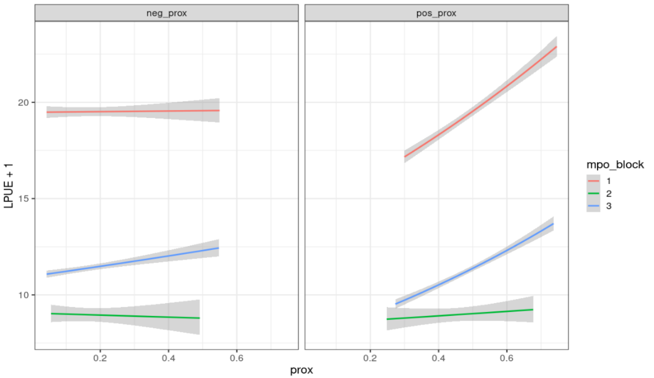
Date year, month

Trajectory Position in the latent space

Collective behaviour Cluster (fishing day), Blocks (persistent interactions), Inter-block proximity, Intra-block proximity

Proximity

- LPUEs increase with intra-community proximity
- Not so much with inter-community proximity



THåMes

- `python` library: `https://github.com/Chabert-Liddell/multiplexobs`
- Handles (0,1) value data with Beta or Continuous Bernoulli
- Dynamic in the process: $W^{l}|W_{lk'=1}\mathcal{M}(1,(\pi_{k'1},\ldots,\pi_{k'K})$

THåMes

- python library: https://github.com/Chabert-Liddell/multiplexobs
- Handles (0,1) value data with Beta or Continuous Bernoulli
- Dynamic in the process: $W_l | W_{lk'} = 1 \sim \mathcal{M}(1, (\pi_{k'1}, \ldots, \pi_{k'K}))$

References

Le, C. M., Levin, K., and Levina, E. (2018). Estimating a network from multiple noisy realizations. *Electronic Journal of Statistics*, 12(2):4697–4740. Publisher: Institute of Mathematical Statistics and Bernoulli Society.

Leger, J.-B. (2023). Parametrization cookbook: A set of bijective parametrizations for using machine learning methods in statistical inference.

Mantziou, A., Lunagómez, S., and Mitra, R. (2024). Bayesian model-based clustering for populations of network data. *The Annals of Applied Statistics*, 18(1):266–302.

Newman, M. E. J. (2018). Estimating network structure from unreliable measurements. *Physical Review E*, 98(6):062321. arXiv:1803.02427 [physics].

Vallès-Català, T., Massucci, F. A., Guimerà, R., and Sales-Pardo, M. (2016). Multilayer Stochastic Block Models Reveal the Multilayer Structure of Complex Networks. *Physical Review X*, 6(1):011036.

Young, J.-G., Cantwell, G. T., and Newman, M. E. J. (2021). Bayesian inference of network structure from unreliable data. *Journal of Complex Networks*, 8(6):cnaa046. arXiv:2008.03334 [physics, stat].