Corrected Observation Process for Latent Block Model

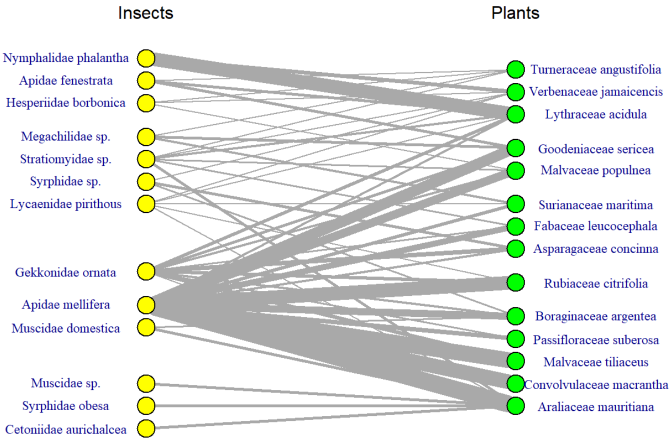
Emre Anakok

directed by Pierre Barbillon, Colin Fontaine & Elisa Thebault

August 29, 2022

Introduction

Bipartite graph, Data from Olesen*et al*, 2002



Introduction

- Contingency table $M_{i,j}$ of size $n_1 \times n_2$
- Structure of the contingency matrix ?

Contingency data from Lara-Romero *et al*, 2016

LBM model

- For rows : a species i is in a group $K_i \in \{1, \ldots, Q_1\}$
- For columns : a species j is in a group $L_j \in \{1, \ldots, Q_2\}$
- First approach : $M_{i,j} | (K_i = k, L_j = l) \sim \mathcal{B}(\pi_{k,l})$

SBM package on R computes the parameters and chooses the best number of groups with the ICL criterion.

Simulation and sampling example

- Assumption : $K = L = 1, \pi_{k,l} = c_0$

Complete network $M_{i,j}$

Subsampled network $R_{i,j}$ (70%)

SBM on example

- Fitting a SBM model on data doesn't yield the same result

SBM fit on $M_{i,j}$

SBM fit on $R_{i,j}$

Table of Contents

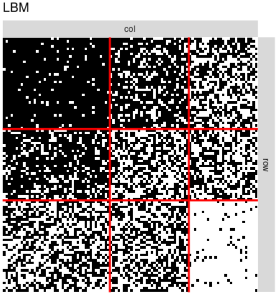
Outline

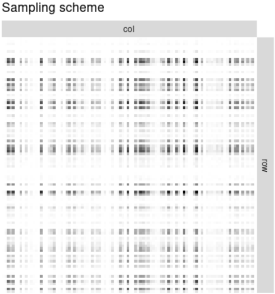
Contingency table format

The following method can't work with binary contingency table.

Frequency data is needed.

CoOP-LBM





$$M \sim \text{LBM}\,(\alpha, \beta, \pi)$$
$$\sum \alpha = \sum \beta = 1, \quad \pi \in]0,1[^{Q_1 \times Q_2}$$
with latent variable Z^1, Z^2

$$N \sim \text{Sampling scheme}$$
$$\mathcal{P}(\lambda_i \mu_j G)$$
$$\lambda_i, \mu_j \in]0,1],\ G > 0$$

$$R_{i,j} = M_{i,j} \times N_{i,j}$$

CoOP-LBM

- $M \sim$ LBM of parameters $\theta_M = (\alpha, \beta, \pi)$
- $N \sim$ Sampling scheme of parameters $\theta_N = (\lambda, \mu, G)$
- $R \sim$ CoOP-LBM of parameters $\theta = (\theta_M, \theta_N)$ if $R_{i,j} = M_{i,j} \times N_{i,j}$

M and N are supposed independent.

$R_{i,j}$ can be equal to 0 for 2 reasons :

- Forbidden interaction : $M_{i,j} = 0$
- Missed interaction : $N_{i,j} = 0$

CoOP-LBM log-likelihood

The log-likelihood given Z^1 and Z^2 can be written as

$$\log \mathcal{L}(R, \theta, Z^1, Z^2) = \log \mathcal{L}(\theta, Z^1) + \log \mathcal{L}(\theta, Z^2) + \log \mathcal{L}(R, \theta | Z^1, Z^2)$$

The observed likelihood is then written :

$$\log \mathcal{L} = \sum_{(Z^1, Z^2) \in (\mathcal{Z}^1, \mathcal{Z}^2)} \log \mathcal{L}(R, \theta, Z^1, Z^2).$$

As the LBM, the sum is intractable.

Outline

Algorithm 1: Stochastic EM for CoOP-LBM inference

Initialisation : $Z^1_{(0)}, Z^2_{(0)}, \pi_{(0)}, \tilde{M}_{(0)}$

repeat

1. M-step a) : update $\alpha_{(n+1)}, \beta_{(n+1)} | Z^1_{(n)}, Z^2_{(n)}$
2. M-step b) : update $\lambda_{(n+1)}, \mu_{(n+1)}, G_{(n+1)} | \tilde{M}_{(n)}$
3. S-step a) : simulate $\tilde{M}_{(n+1)} | Z^1_{(n)}, Z^2_{(n)}, \pi_{(n)}, \lambda_{(n+1)}, \mu_{(n+1)}, G_{(n+1)}$
4. M-step c) : update $\pi_{(n+1)} | \tilde{M}_{(n+1)}, Z^1_{(n)}, Z^2_{(n)}$
5. S-step b) : simulate $Z^1_{(n+1)}, Z^2_{(n+1)} | \alpha_{(n+1)}, \beta_{(n+1)}, \pi_{(n+1)}, \tilde{M}_{(n+1)}$

until Number of iterations reached

$\tilde{M}_{(n)}$ is a matrix where missing interaction are simulated with a Bernoulli variable of probability $\mathbb{P}(M_{i,j} = 1 | R_{i,j} = 0)$.

Particularity of the algorithm

- M-step b) : λ, μ, G are updated with a fixed point algorithm.
- S-step a) :

$$\mathbb{P}(M_{i,j} = 1 | R_{i,j} = 0, \lambda_i, \mu_j, G, \pi, Z^1_{ik} = 1, Z^2_{jl} = 1) = \frac{\pi_{kl} e^{-\lambda_i \mu_j G}}{1 - \pi_{kl}(1 - e^{-\lambda_i \mu_j G})}$$

- S-step b) :

$$\mathbb{P}(Z^1_{ik} = 1 | R, \theta, Z^2) \propto \mathbb{P}(R | \theta, Z^1_{ik} = 1) \mathbb{P}(Z^1_{ik} = 1)$$

Model selection

- ICL criterion :

$$ICL(m_{Q_1,Q_2}) = \max_{\theta} \mathcal{L}(R, \widehat{Z^1}, \widehat{Z^2}|\theta, m_{Q_1,Q_2})$$
$$-\frac{Q_1 - 1}{2}\log(n_1) - \frac{Q_2 - 1}{2}\log(n_2) - \frac{Q_1 Q_2 + n_1 + n_2 - 1}{2}\log(n_1 n_2)$$

Outline

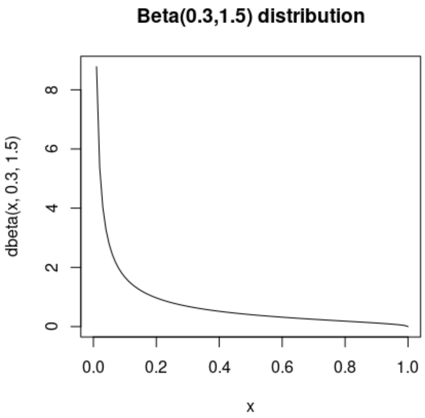
Presentation of CoOP-LBM

Inference

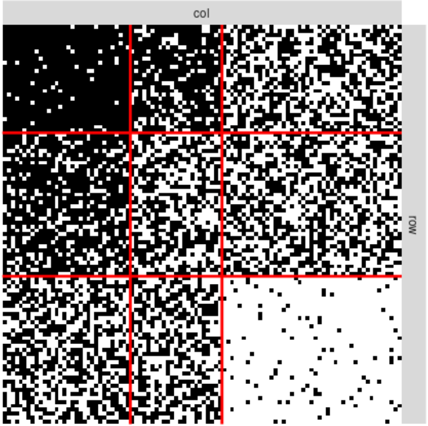
Application on simulated data

Application on real data

Settings



Beta distribution for λ, μ



Contingency matrix M

ARI score

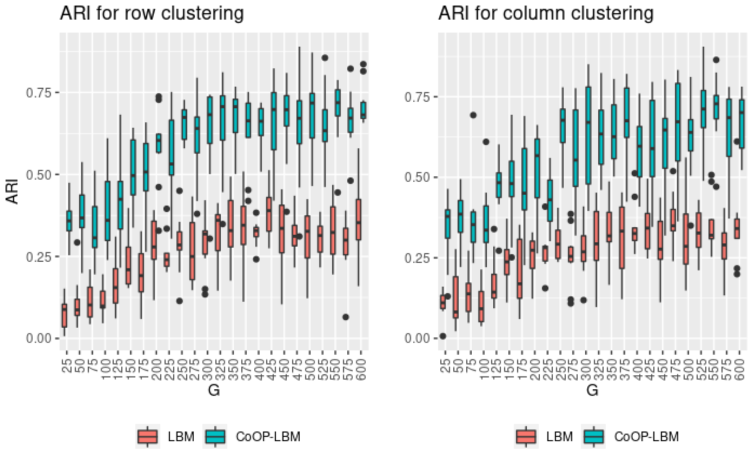
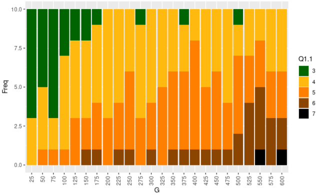
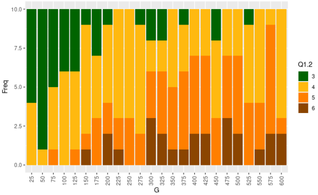


Figure – ARI score for rows and columns when the number of groups is unknown

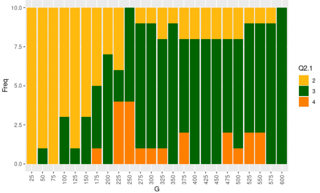
Number of groups estimated



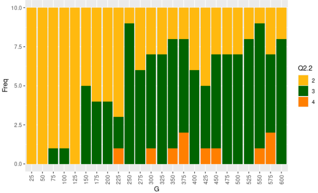
Estimated Q_1 for LBM



Estimated Q_2 for LBM



Estimated Q_1 for CoOP-LBM



Estimated Q_2 for CoOP-LBM

AUC on missing data

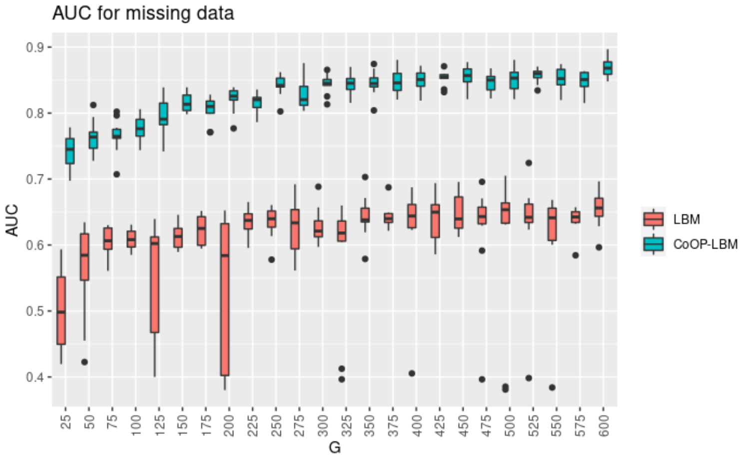


Figure – AUC of the ROC for the missing data

Outline

Presentation of CoOP-LBM

Inference

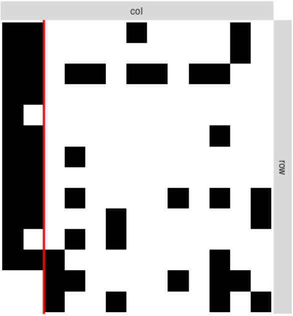
Application on simulated data

Application on real data

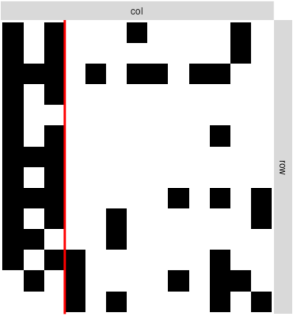
Data presentation

- Olesen et al., 2002 : Invasion of pollination networks on oceanic islands : importance of invader complexes and endemic super generalists
- 14 species of plants, 13 species of insects.
- 1395 interactions observed.

Fitting models on the network



LBM



CoOP-LBM

Only difference is observed for the insect species Lycaenidae pirithous

Estimated coverage

Boraginaceae argentea	0.8493672
Asparagaceae concinna	0.8007462
Araliaceae mauritiana	0.9998364
Malvaceae tiliaceus	0.8137610
Convolvulaceae macrantha	0.6746022
Fabaceae leucocephala	0.8028505
Rubiaceae citrifolia	0.9993632
Passifloraceae suberosa	0.7164068
Lythraceae acidula	0.9952563
Goodeniaceae sericea	0.8680551
Surianaceae maritima	0.9703093
Malvaceae populnea	0.8608773
Verbenaceae jamaicencis	0.9348724
Turneraceae angustifolia	0.7207160

Apidae mellifera	0.9997752
Hesperiidae borbonica	0.9227533
Lycaenidae pirithous	0.4642996
Muscidae sp.	0.8855992
Megachilidae sp.	0.9971414
Muscidae domestica	0.9371763
Syrphidae obesa	0.8847282
Nymphalidae phalantha	1.0000000
Gekkonidae ornata	0.9452383
Cetoniidae aurichalcea	0.8761673
Stratiomyidae sp.	0.9988445
Syrphidae sp.	0.9876197
Apidae fenestrata	0.9964598

Lycaenidae pirithous has been observed only 7 times on 5 different flowers.

Conclusion

- CoOP-LBM has better result than SBM in our simulation settings and on real data.
- It can change our perception of networks by correcting the structure.
- Available soon in a R package.