Some latent variable models in ecology

Stéphane Robin

Based on joint works with Pierre Latouche, Sarah Ouadah [OLR22],
Anna Bonnet [BR25] and Julien Stoehr [SR24]

LPSM, Sorbonne université

StatMathAppli, Fréjus, Sep. 2025

Latent variable models in ecology

Latent ('hidden', 'unobserved', ...) variables are widely used in statistical ecology [PG22] to

- account for heterogeneity
- encode dependency
- represent a 'true' signal observed with noise
- ...

Latent variable models in ecology

Latent ('hidden', 'unobserved', ...) variables are widely used in statistical ecology [PG22] to

- account for heterogeneity
- encode dependency
- represent a 'true' signal observed with noise
- ...

Statistical perspective.

- Nb model parameters $\ll$ Nb latent variables $\simeq$ Nb observed variables.
- Inference of the model parameters much easier if the latent variables were observed.

Latent variable models

Notations.

- Y = observed variables (response),
- Z = unobserved (latent) variables,
- θ = unknown parameter (to be inferred),
- X = covariates (given).

Latent variable models

Notations.

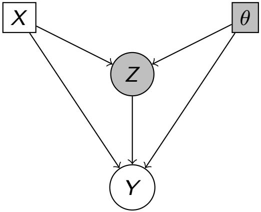
- Y = observed variables (response),
- Z = unobserved (latent) variables,
- θ = unknown parameter (to be inferred),
- X = covariates (given).

General model. (frequentist setting)

- Hidden layer: $Z \sim p_\theta(Z; X)$,
- Observed layer: $Y \mid Z \sim p_\theta(Y \mid Z; X)$.

	observed	unobserved
fix	X	θ
random	Y	Z

Graphical model.



Inference specificity

Obviously:

$$p_\theta(Y) = \int_{\mathcal{Z}} p_\theta(Z) p_\theta(Y \mid Z) \, \mathrm{d}Z$$

[1] $\mathcal{H}(q) = -\mathbb{E}_q[\log q(X)]$

Inference specificity

Obviously:

$$p_\theta(Y) = \int_{\mathcal{Z}} p_\theta(Z) p_\theta(Y \mid Z) \, \mathrm{d}Z$$

EM decomposition [DLR77]:

$$\log p_\theta(Y) = \mathbb{E}[\log p_\theta(Y, Z) \mid Y] + \mathcal{H}\left[p_\theta(Z \mid Y)\right]$$

where $\mathcal{H}$ = entropy[1].

1. Still: $p_\theta(Z \mid Y) = p_\theta(Y, Z) / p_\theta(Y)$.

[1] $\mathcal{H}(q) = -\mathbb{E}_q[\log q(X)]$

Inference specificity

Obviously:

$$p_\theta(Y) = \int_{\mathcal{Z}} p_\theta(Z) p_\theta(Y \mid Z) \, \mathrm{d}Z$$

EM decomposition [DLR77]:

$$\log p_\theta(Y) = \mathbb{E}[\log p_\theta(Y, Z) \mid Y] + \mathcal{H}\left[p_\theta(Z \mid Y)\right]$$

where $\mathcal{H}$ = entropy[1].

1. Still: $p_\theta(Z \mid Y) = p_\theta(Y, Z)/p_\theta(Y)$.

Three typical situations:

[1] $\mathcal{H}(q) = -\mathbb{E}_q[\log q(X)]$

Inference specificity

Obviously:

$$p_\theta(Y) = \int_{\mathcal{Z}} p_\theta(Z) p_\theta(Y \mid Z) \, \mathrm{d}Z$$

EM decomposition [DLR77]:

$$\log p_\theta(Y) = \mathbb{E}[\log p_\theta(Y, Z) \mid Y] + \mathcal{H}\left[p_\theta(Z \mid Y)\right]$$

where $\mathcal{H}$ = entropy[1].

1. Still: $p_\theta(Z \mid Y) = p_\theta(Y, Z)/p_\theta(Y)$.

Three typical situations:

1. Integration wrt Z can be done for free,

[1] $\mathcal{H}(q) = -\mathbb{E}_q[\log q(X)]$

Inference specificity

Obviously:

$$p_\theta(Y) = \int_{\mathcal{Z}} p_\theta(Z) p_\theta(Y \mid Z) \, \mathrm{d}Z$$

EM decomposition [DLR77]:

$$\log p_\theta(Y) = \mathbb{E}[\log p_\theta(Y, Z) \mid Y] + \mathcal{H}\left[p_\theta(Z \mid Y)\right]$$

where $\mathcal{H}$ = entropy[1].

1. Still: $p_\theta(Z \mid Y) = p_\theta(Y, Z) / p_\theta(Y)$.

Three typical situations:

1. Integration wrt Z can be done for free,
2. Integration wrt Z is intractable, but $\mathbb{E}_\theta[f(Z) \mid Y]$ can be dealt with,

[1] $\mathcal{H}(q) = -\mathbb{E}_q[\log q(X)]$

Inference specificity

Obviously:

$$p_\theta(Y) = \int_{\mathcal{Z}} p_\theta(Z) p_\theta(Y \mid Z) \, \mathrm{d}Z$$

EM decomposition [DLR77]:

$$\log p_\theta(Y) = \mathbb{E}[\log p_\theta(Y, Z) \mid Y] + \mathcal{H}\left[p_\theta(Z \mid Y)\right]$$

where $\mathcal{H}$ = entropy[1].

1. Still: $p_\theta(Z \mid Y) = p_\theta(Y, Z)/p_\theta(Y)$.

Three typical situations:

1. Integration wrt Z can be done for free,
2. Integration wrt Z is intractable, but $\mathbb{E}_\theta[f(Z) \mid Y]$ can be dealt with,
3. Integration wrt Z is intractable, and $\mathbb{E}_\theta[f(Z) \mid Y]$ is inaccessible.

[1] $\mathcal{H}(q) = -\mathbb{E}_q[\log q(X)]$

Model 1: Plant pollinator networks

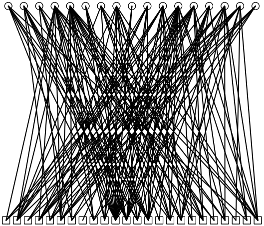
Model 1: Plant pollinator networks

Species.

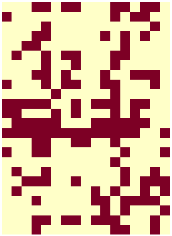
- $i = 1, \dots m$ pollinators
  = bottom nodes = rows
- $j = 1, \dots n$ plants
  = top nodes = columns
- Y_{ij} existence of an interaction between pollinator i and plant j

$$Y_{ij} = \mathbb{I}\{i \sim j\}$$

Network



Adjacency matrix



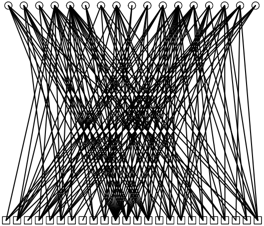
Model 1: Plant pollinator networks

Species.

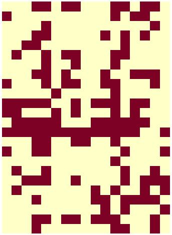
- $i = 1, \dots m$ pollinators
  $=$ bottom nodes $=$ rows
- $j = 1, \dots n$ plants
  $=$ top nodes $=$ columns
- Y_{ij} existence of an interaction between pollinator i and plant j

$$Y_{ij} = \mathbb{I}\{i \sim j\}$$

Network



Adjacency matrix



Network comparison. Many plant-pollinator networks are collected, to be compared across time, space, environmental conditions, ...

- They each involve different sets of species
- And networks are complex objects

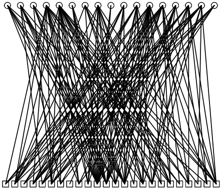
Model 1: Motif-based network embedding

Motif based network embedding: Replace a network with a vector of motif counts
[SROB16,SCB+19] [#48]

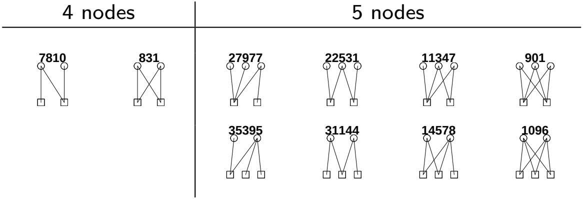
Model 1: Motif-based network embedding

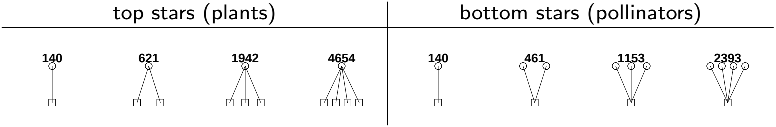
Motif based network embedding: Replace a network with a vector of motif counts [SROB16,SCB+19] [#48]

Network. (24 × 17)



Motif counts. (nodes = species) [#49]





Model 1: Bipartite expected degree distribution

Need for a null model. Motif counts depend on

- the network's dimensions (m pollinators $\times n$ plants),
- the network density (number of edges),
- the existence of generalist and specialist species.

Model 1: Bipartite expected degree distribution

Need for a null model. Motif counts depend on

- the network's dimensions (m pollinators $\times n$ plants),
- the network density (number of edges),
- the existence of generalist and specialist species.

Bipartite expected degree distribution (BEDD) [OLR22]. [#13]

Model 1: Bipartite expected degree distribution

Need for a null model. Motif counts depend on

- the network's dimensions (m pollinators $\times n$ plants),
- the network density (number of edges),
- the existence of generalist and specialist species.

Bipartite expected degree distribution (BEDD) [OLR22]. [#13]

- Latent layer: $Z = (U, V)$:

$$(U_i)_{i=1,\dots m}, (V_j)_{j=1,\dots n} \text{ iid } \sim \mathcal{U}[0,1]$$

Model 1: Bipartite expected degree distribution

Need for a null model. Motif counts depend on

- the network's dimensions (m pollinators $\times n$ plants),
- the network density (number of edges),
- the existence of generalist and specialist species.

Bipartite expected degree distribution (BEDD) [OLR22]. [#13]

- Latent layer: $Z = (U, V)$:

$$(U_i)_{i=1,\dots m}, (V_j)_{j=1,\dots n} \text{ iid } \sim \mathcal{U}[0,1]$$

- Observed layer: Y = network

$$(Y_{ij})_{1 \leqslant i \leqslant m, 1 \leqslant j \leqslant n} \text{ indep.} \mid Z: \quad Y_{ij} \mid U_i, V_i \sim \mathcal{B}(\rho \, g(U_i) \, h(V_j))$$

Model 1: Bipartite expected degree distribution

Need for a null model. Motif counts depend on

- the network's dimensions (m pollinators $\times n$ plants),
- the network density (number of edges),
- the existence of generalist and specialist species.

Bipartite expected degree distribution (BEDD) [OLR22]. [#13]

- Latent layer: $Z = (U, V)$:

$$(U_i)_{i=1,\dots m}, (V_j)_{j=1,\dots n} \text{ iid } \sim \mathcal{U}[0,1]$$

- Observed layer: Y = network

$$(Y_{ij})_{1 \leqslant i \leqslant m, 1 \leqslant j \leqslant n} \text{ indep. } \mid Z : \quad Y_{ij} \mid U_i, V_i \sim \mathcal{B}(\rho \, g(U_i) \, h(V_j))$$

- Parameters: $\theta = (\rho, g, h)$, $\rho \in (0,1)$, $g, h : (0,1) \mapsto (0,1)$

$$\rho = \text{network density}, \qquad \begin{array}{l} g = \text{top node degree imbalance } \left(\int g = 1\right), \\ h = \text{bottom node degree imbalance } \left(\int h = 1\right) \end{array}$$

Model 1: Bipartite expected degree distribution

Need for a null model. Motif counts depend on

- the network's dimensions (m pollinators $\times n$ plants),
- the network density (number of edges),
- the existence of generalist and specialist species.

Bipartite expected degree distribution (BEDD) [OLR22]. [#13]

- Latent layer: $Z = (U, V)$:

$$(U_i)_{i=1,\dots m}, (V_j)_{j=1,\dots n} \text{ iid } \sim \mathcal{U}[0,1]$$

- Observed layer: Y = network

$$(Y_{ij})_{1 \leqslant i \leqslant m, 1 \leqslant j \leqslant n} \text{ indep.} \mid Z : \quad Y_{ij} \mid U_i, V_i \sim \mathcal{B}(\rho\, g(U_i)\, h(V_j))$$

- Parameters: $\theta = (\rho, g, h)$, $\rho \in (0,1)$, $g, h : (0,1) \mapsto (0,1)$

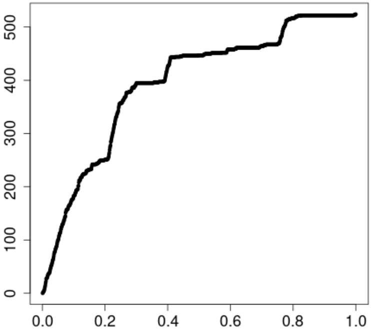
$$\rho = \text{network density}, \qquad g = \text{top node degree imbalance } \left(\textstyle\int g = 1\right),$$
$$h = \text{bottom node degree imbalance } \left(\textstyle\int h = 1\right)$$

Latent variable. $Z = (U, V)$: Accounts for an heterogeneity, which is known to exist.

Model 2: Bat calls

Model 2: Bat calls

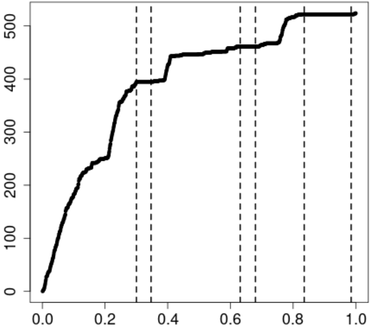
Data [#22].
Overnight recording of bat calls in continuous time



Model 2: Bat calls

Data [#22].
Overnight recording of bat calls in continuous time

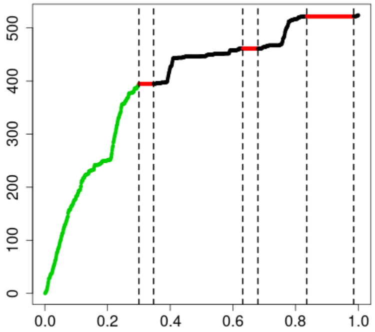
- Can we detect changes in the distribution of events (calls)?



Model 2: Bat calls

Data [#22].
Overnight recording of bat calls in continuous time

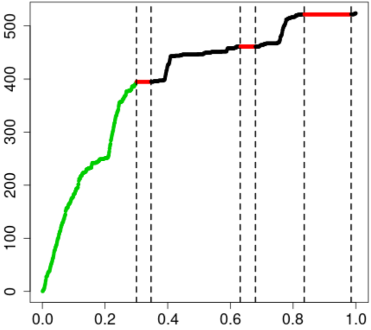
- Can we detect changes in the distribution of events (calls)?
- Can we associate each time period with some underlying behavior?



Model 2: Bat calls

Data [#22].
Overnight recording of bat calls in continuous time

- Can we detect changes in the distribution of events (calls)?
- Can we associate each time period with some underlying behavior?



Specificity.

- Bat calls are emitted in bursts (clusters).

Model 2: Markov-switching Hawkes process

Discrete-time Hawkes process $(Y_k)_{k \geqslant 1}$. Y_k = number of events in the k-th time bin:

$$Y_k \mid (Y_\ell)_{\ell \leqslant k-1} \sim \mathcal{P}\left(\mu + \alpha \sum_{\ell=1}^{k-1} \beta^{\ell-1} Y_{k-\ell}\right)$$

- μ = immigration rate, α, β = influence of the past events (self-exciting).
- InAR process [Kir16], which converges to Hawkes process with exponential kernel. [#53]

Model 2: Markov-switching Hawkes process

Discrete-time Hawkes process $(Y_k)_{k \geqslant 1}$. Y_k = number of events in the k-th time bin:

$$Y_k \mid (Y_\ell)_{\ell \leqslant k-1} \sim \mathcal{P}\left(\mu + \alpha \sum_{\ell=1}^{k-1} \beta^{\ell-1} Y_{k-\ell}\right)$$

- μ = immigration rate, α, β = influence of the past events (self-exciting).
- InAR process [Kir16], which converges to Hawkes process with exponential kernel. [#53]

Markov switching Hawkes process [BR25].

- Hidden path $(Z_k)_{k \geqslant 1}$ = homogeneous Markov chain with Q states

$$(Z_k)_{k \geqslant 1} \sim MC_Q(\nu, \pi);$$

Model 2: Markov-switching Hawkes process

Discrete-time Hawkes process $(Y_k)_{k \geqslant 1}$. Y_k = number of events in the k-th time bin:

$$Y_k \mid (Y_\ell)_{\ell \leqslant k-1} \sim \mathcal{P}\left(\mu + \alpha \sum_{\ell=1}^{k-1} \beta^{\ell-1} Y_{k-\ell}\right)$$

- μ = immigration rate, α, β = influence of the past events (self-exciting).
- InAR process [Kir16], which converges to Hawkes process with exponential kernel. [#53]

Markov switching Hawkes process [BR25].

- Hidden path $(Z_k)_{k \geqslant 1}$ = homogeneous Markov chain with Q states

$$(Z_k)_{k \geqslant 1} \sim MC_Q(\nu, \pi);$$

- Observed counts: for $k \geqslant 1$ and

$$(Y_k \mid (Y_\ell)_{\ell \leqslant k-1}, Z_k = q) \sim \mathcal{P}\left(\mu_q + \alpha \sum_{\ell=1}^{k-1} \beta^{\ell-1} Y_{k-\ell}\right);$$

Model 2: Markov-switching Hawkes process

Discrete-time Hawkes process $(Y_k)_{k \geqslant 1}$. Y_k = number of events in the k-th time bin:

$$Y_k \mid (Y_\ell)_{\ell \leqslant k-1} \sim \mathcal{P}\left(\mu + \alpha \sum_{\ell=1}^{k-1} \beta^{\ell-1} Y_{k-\ell}\right)$$

- μ = immigration rate, α, β = influence of the past events (self-exciting).
- InAR process [Kir16], which converges to Hawkes process with exponential kernel. [#53]

Markov switching Hawkes process [BR25].

- Hidden path $(Z_k)_{k \geqslant 1}$ = homogeneous Markov chain with Q states

$$(Z_k)_{k \geqslant 1} \sim MC_Q(\nu, \pi);$$

- Observed counts: for $k \geqslant 1$ and

$$(Y_k \mid (Y_\ell)_{\ell \leqslant k-1}, Z_k = q) \sim \mathcal{P}\left(\mu_q + \alpha \sum_{\ell=1}^{k-1} \beta^{\ell-1} Y_{k-\ell}\right);$$

Latent variable. Encodes the behavior of the animal(s).

Model 3: Joint species distribution model

Model 3: Joint species distribution model

Species distribution model (SDM). Which conditions favour or hinder a given species?

- $i = 1 \dots n$ sites
- x_i = covariates describing the environment in site i
- Y_i = abundance (ie number of individual) of the species of interest in site i
- SDM = univariate generalized (mixed) (linear) model:

$$Y_i \sim \mathcal{F}(\cdot; x_i, \theta).$$

Model 3: Joint species distribution model

Species distribution model (SDM). Which conditions favour or hinder a given species?

- $i = 1 \dots n$ sites
- x_i = covariates describing the environment in site i
- Y_i = abundance (ie number of individual) of the species of interest in site i
- SDM = univariate generalized (mixed) (linear) model:

$$Y_i \sim \mathcal{F}(\cdot; x_i, \theta).$$

Joint species distribution model (JSDM). Which condition favour or hinder a set of species and how do they 'interact'?

- $j = 1 \dots p$ species
- Y_{ij} = abundance of species j in site i, $Y_i = (Y_{i1}, \dots Y_{ip})$ abundance vector in site i
- JSDM = multivariate generalized (mixed) (linear) model:

$$Y_i \sim \mathcal{F}(\cdot; x_i, \theta).$$

Model 3: Joint species distribution model

Species distribution model (SDM). Which conditions favour or hinder a given species?

- $i = 1 \dots n$ sites
- x_i = covariates describing the environment in site i
- Y_i = abundance (ie number of individual) of the species of interest in site i
- SDM = univariate generalized (mixed) (linear) model:

$$Y_i \sim \mathcal{F}(\cdot; x_i, \theta).$$

Joint species distribution model (JSDM). Which condition favour or hinder a set of species and how do they 'interact'?

- $j = 1 \dots p$ species
- Y_{ij} = abundance of species j in site i, $Y_i = (Y_{i1}, \dots Y_{ip})$ abundance vector in site i
- JSDM = multivariate generalized (mixed) (linear) model:

$$Y_i \sim \mathcal{F}(\cdot; x_i, \theta).$$

Specificity.

- Y_i is a count vector.
- Not that many flexible multivariate distributions for counts on the shelf [IYAR17].

Model 3: Poisson log-normal distribution

Most JSDM resort to a Gaussian latent structure [WBO+15,OA20] to encode the dependence between species.

Model 3: Poisson log-normal distribution

Most JSDM resort to a Gaussian latent structure [WBO+15,OA20] to encode the dependence between species.

Poisson log-normal model [AH89,CMR21].

Model 3: Poisson log-normal distribution

Most JSDM resort to a Gaussian latent structure [WBO+15,OA20] to encode the dependence between species.

Poisson log-normal model [AH89,CMR21].

- Latent layer:

$$(Z_i)_{1 \leqslant i \leqslant n} \text{ iid } \sim \mathcal{N}_p(0, \Sigma)$$

Model 3: Poisson log-normal distribution

Most JSDM resort to a Gaussian latent structure [WBO+15,OA20] to encode the dependence between species.

Poisson log-normal model [AH89,CMR21].

- Latent layer:
$$(Z_i)_{1 \leqslant i \leqslant n} \text{ iid } \sim \mathcal{N}_p(0, \Sigma)$$
- Observed layer: counts $(Y_{ij})_{1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p}$ indep $\mid Z$
$$Y_{ij} \mid Z \sim \mathcal{P}\left(\exp(x_i^\top \beta_j + Z_{ij})\right)$$

Model 3: Poisson log-normal distribution

Most JSDM resort to a Gaussian latent structure [WBO+15,OA20] to encode the dependence between species.

Poisson log-normal model [AH89,CMR21].

- Latent layer:

$$(Z_i)_{1 \leqslant i \leqslant n} \text{ iid } \sim \mathcal{N}_p(0, \Sigma)$$

- Observed layer: counts $(Y_{ij})_{1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p}$ indep $\mid Z$

$$Y_{ij} \mid Z \sim \mathcal{P}\left(\exp(x_i^\top \beta_j + Z_{ij})\right)$$

- Parameters $\theta = (\beta, \Sigma)$:

β_j = effects of the environmental covariates on species j (*abiotic* interactions)

Σ = between-species latent covariance matrix (*biotic* interactions)

Model 3: Poisson log-normal distribution

Most JSDM resort to a Gaussian latent structure [WBO+15,OA20] to encode the dependence between species.

Poisson log-normal model [AH89,CMR21].

- Latent layer:

$$(Z_i)_{1 \leqslant i \leqslant n} \text{ iid } \sim \mathcal{N}_p(0, \Sigma)$$

- Observed layer: counts $(Y_{ij})_{1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p}$ indep $\mid Z$

$$Y_{ij} \mid Z \sim \mathcal{P}\left(\exp(x_i^\top \beta_j + Z_{ij})\right)$$

- Parameters $\theta = (\beta, \Sigma)$:

$$\beta_j = \text{effects of the environmental covariates on species } j \text{ (\textit{abiotic} interactions)}$$
$$\Sigma = \text{between-species latent covariance matrix (\textit{biotic} interactions)}$$

Latent variable. Encodes between-species dependencies in a mathematically convenient way.

Outline

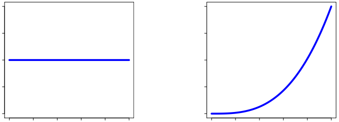
Bipartite expected degree distribution

$h_0(v) =$ $h(v) =$

$$U_i, V_j \sim \mathcal{U}[0,1]$$

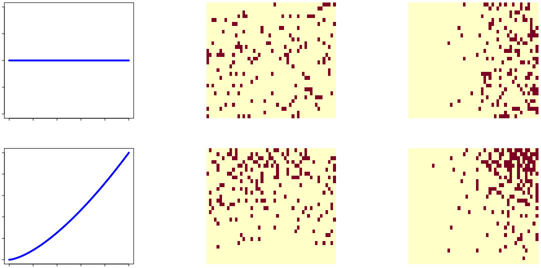
$$Y_{ij} \sim \mathcal{B}(\rho \, g(U_i) \, h(V_j))$$

$$\int g = \int h = 1$$



$g_0(u) =$

$g(u) =$



Bipartite expected degree distribution

$$U_i, V_j \sim \mathcal{U}[0,1]$$

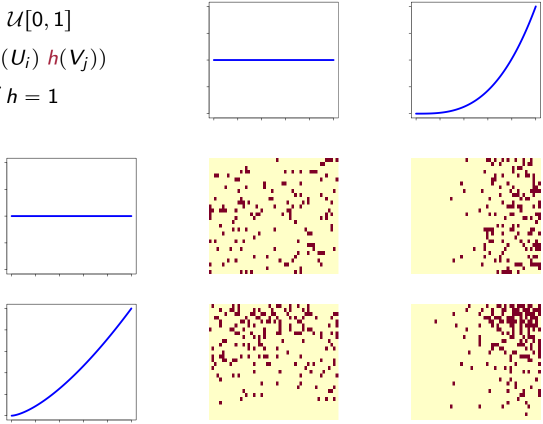
$$Y_{ij} \sim \mathcal{B}(\rho\, g(U_i)\, h(V_j))$$

$$\int g = \int h = 1$$

$h_0(v) =$ $h(v) =$

$g_0(u) =$

$g(u) =$

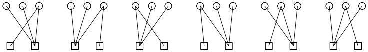


- No preferred or avoided specific connexion
- Graph-exchangeable model: pollinators (and plants) can be permuted
- Bipartite version of the expected degree distribution [CL02]
- Expected degrees: $\mathbb{E}(Y_{i+} \mid U_i) = n\rho g(U_i)$, $\mathbb{E}(Y_{+j} \mid V_j) = m\rho h(V_j)$. [#7]

Motif count

Couting motifs[2]. For a given motif s with p_s top nodes and q_s bottom nodes:

- Determine the r_s *automorphisms* = non-redundant permutations

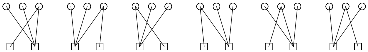


[2]Not in the way of the `bmotif` package [SSS+19]

Motif count

Couting motifs[2]. For a given motif s with p_s top nodes and q_s bottom nodes:

- Determine the r_s *automorphisms* = non-redundant permutations



- Choose p_s nodes among m and q_s nodes among n;
- The number of possible 'positions' is then

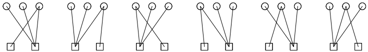
$$c_s := \binom{m}{p_s} \times \binom{n}{q_s} \times r_s;$$

[2]Not in the way of the `bmotif` package [SSS+19]

Motif count

Couting motifs[2]. For a given motif s with p_s top nodes and q_s bottom nodes:

- Determine the r_s *automorphisms* = non-redundant permutations



- Choose p_s nodes among m and q_s nodes among n;
- The number of possible 'positions' is then

$$c_s := \begin{pmatrix} m \\ p_s \end{pmatrix} \times \begin{pmatrix} n \\ q_s \end{pmatrix} \times r_s;$$

- Try all positions $\alpha = 1, \dots c_s$, and count the number of matches:

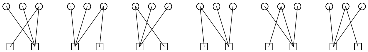
$$N_s = \sum_{\alpha=1}^{c_s} \mathbb{I}\{\text{motif } s \text{ matches at position } \alpha\}.$$

[2]Not in the way of the `bmotif` package [SSS+19]

Motif count

Couting motifs[2]. For a given motif s with p_s top nodes and q_s bottom nodes:

- Determine the r_s *automorphisms* = non-redundant permutations



- Choose p_s nodes among m and q_s nodes among n;
- The number of possible 'positions' is then

$$c_s := \binom{m}{p_s} \times \binom{n}{q_s} \times r_s;$$

- Try all positions $\alpha = 1, \dots c_s$, and count the number of matches:

$$N_s = \sum_{\alpha=1}^{c_s} \mathbb{I}\{\text{motif } s \text{ matches at position } \alpha\}.$$

Expected count. $\mathbb{E}(N_s) = c_s \phi_s$, with

$$\phi_s = \text{matching probability} = \text{'motif probability'}$$

[2]Not in the way of the `bmotif` package [SSS+19]

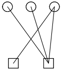
Motif probability

Motif probability $\overline{\phi}_s$ under BEDD[3]. Need to integrate wrt $Z = (U, V)$.

[3] Consider here induced motifs (only the presence of the prescribed edges is required) $\neq$ exact motif

Motif probability

Motif probability $\overline{\phi}_s$ under BEDD[3]. Need to integrate wrt $Z = (U, V)$.

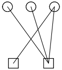
An example. Consider the motif $s =$  with $p_s = 2$ and $q_s = 3$, we have

$$\begin{aligned}\overline{\phi}_s &= \int \cdots \int \rho^4 g(u_1) g(u_2)^3 h(v_1) h(v_2) h(v_3)^2 \; \mathrm{d}u_1 \; \mathrm{d}u_2 \; \mathrm{d}v_1 \; \mathrm{d}v_2 \; \mathrm{d}v_3 \\ &= \left(\int \rho^3 g(u_2)^3 \; \mathrm{d}u_2 \right) \left(\int \rho^2 h(v_3)^2 \; \mathrm{d}v_3 \right) \Big/ \rho \qquad [\#50] \\ &= \text{(bottom 3-star probability)} \times \text{(top 2-star probability)} \,/\, \text{(edge probability)}\end{aligned}$$

[3] Consider here induced motifs (only the presence of the prescribed edges is required) $\neq$ exact motif

Motif probability

Motif probability $\overline{\phi}_s$ under BEDD[3]. Need to integrate wrt $Z = (U, V)$.

An example. Consider the motif $s =$  with $p_s = 2$ and $q_s = 3$, we have

$$\begin{aligned}\overline{\phi}_s &= \int \cdots \int \rho^4 g(u_1) g(u_2)^3 h(v_1) h(v_2) h(v_3)^2 \, \mathrm{d}u_1 \, \mathrm{d}u_2 \, \mathrm{d}v_1 \, \mathrm{d}v_2 \, \mathrm{d}v_3 \\ &= \left(\int \rho^3 g(u_2)^3 \, \mathrm{d}u_2 \right) \left(\int \rho^2 h(v_3)^2 \, \mathrm{d}v_3 \right) \Big/ \rho \qquad [\#50] \\ &= (\text{bottom 3-star probability}) \times (\text{top 2-star probability}) \,/\, (\text{edge probability})\end{aligned}$$

A favourable configuration.

- Edge and star probabilities contain all information.
- Unbiased estimates are given by their respective empirical frequencies $F = N/c$ (sufficient statistics of the BEDD model).
- The integration wrt $Z = (U, V)$ is implicitly achieved (without estimating g and h).

[3] Consider here induced motifs (only the presence of the prescribed edges is required) $\neq$ exact motif

Some more results

Moments of the count.

- Mean: $\mathbb{E}(N_s) = c_s \times \overline{\phi}_s$

[4] Motif counts are also network U-statistics [LM23,LMDMR25]

Some more results

Moments of the count.

- Mean: $\mathbb{E}(N_s) = c_s \times \overline{\phi}_s$
- Variance: Same game, requires to evaluate $\mathbb{E}(N_s^2) = \mathbb{E}\left(\sum_\alpha \text{match}_{s\alpha}\right)^2$
  $\rightarrow$ Need to consider overlaps between positions (*super-motifs*: [PDK+08] [#51])

  $\rightarrow$ Compute the respective expected count in the way as for other motifs
- Covariance: Same game to compute $\mathbb{C}\text{ov}(N_s, N_{s'})$

[4] Motif counts are also network U-statistics [LM23,LMDMR25]

Some more results

Moments of the count.

- Mean: $\mathbb{E}(N_s) = c_s \times \overline{\phi}_s$
- Variance: Same game, requires to evaluate $\mathbb{E}(N_s^2) = \mathbb{E}\left(\sum_\alpha \text{match}_{s\alpha}\right)^2$

  $\rightarrow$ Need to consider overlaps between positions (*super-motifs*: [PDK+08] [#51])

  $\rightarrow$ Compute the respective expected count in the way as for other motifs

- Covariance: Same game to compute $\mathbb{C}\text{ov}(N_s, N_{s'})$

Proposition: Asymptotic normality [OLR22].[4] Under BEDD, for non-star motifs,

- Under sparsity conditions ($\rho \propto m^{-a}n^{-b}$):

$$\left(N_s - \widehat{\mathbb{E}}(N_s)\right) \Big/ \sqrt{\widehat{\mathbb{V}}(N_s)} \quad \xrightarrow{m,n\rightarrow\infty} \quad \mathcal{N}(0,1)$$

- Account for plug-in when moderate network size (Δ-method):

$$\left(N_s - \widehat{\mathbb{E}}(N_s) + \widehat{\mathbb{B}}\left(\widehat{\mathbb{E}}(N_s)\right)\right) \Big/ \sqrt{\widehat{\mathbb{V}}(N_s - \widehat{\mathbb{E}}(N_s))} \quad \xrightarrow{m,n\rightarrow\infty} \quad \mathcal{N}(0,1)$$

[4] Motif counts are also network U-statistics [LM23,LMDMR25]

Model 1: Networks comparison in space and time

French plant-pollinator networks

Joint work with Natasha de Manincor et François Massol

Question. Does the structure of plant-pollinator network vary in space and time?

French plant-pollinator networks

Joint work with Natasha de Manincor et François Massol

Question. Does the structure of plant-pollinator network vary in space and time?

Design.

- 3 French regions (Hauts-de-France, Normandie and Occitanie), 2 sites / region
- 2 years, 7 months / year
- $3 \times 2 \times 2 \times 7 \simeq 82$ networks

French plant-pollinator networks

Joint work with Natasha de Manincor et François Massol

Question. Does the structure of plant-pollinator network vary in space and time?

Design.

- 3 French regions (Hauts-de-France, Normandie and Occitanie), 2 sites / region
- 2 years, 7 months / year
- $3 \times 2 \times 2 \times 7 \simeq 82$ networks

Approach. Distance-based embedding:

- Define a network distance (gathering all motifs)
- Use (permutation-based) multivariate analysis of variance to test spatial or temporal effects ('Adonis', [MA01,ZS06])

Comparing network imbalances

Question. Do network A and B share the same imbalance for pollinators?

Comparing network imbalances

Question. Do network A and B share the same imbalance for pollinators?

Test statistic.

- Assume $A \sim BEDD(\rho^A, g^A, h^A)$ and $B \sim BEDD(\rho^B, g^B, h^B)$

$$H_0 = \{g^A = g^B\}$$

- For motif s, with

$$\widehat{\mathbb{E}}_0(N_s^A) = \widehat{\mathbb{E}}_{\widehat{\rho}^A, \widehat{g}^B, \widehat{h}^A}(N_s^A), \qquad \widehat{\mathbb{E}}_0(N_s^B)) = \widehat{\mathbb{E}}_{\widehat{\rho}^B, \widehat{g}^A, \widehat{h}^B}(N_s^B)$$

we have

$$W_s^{(g)}(A, B) = \frac{(N_s^A - \widehat{\mathbb{E}}_0(N_s^A)) - (N_s^B - \widehat{\mathbb{E}}_0(N_s^B))}{\sqrt{\widehat{\mathbb{V}}_0(N_s^A) + \widehat{\mathbb{V}}_0(N_s^B)}} \overset{H_0}{\sim} \mathcal{N}(0, 1)$$

Comparing network imbalances

Question. Do network A and B share the same imbalance for pollinators?

Test statistic.

- Assume $A \sim BEDD(\rho^A, g^A, h^A)$ and $B \sim BEDD(\rho^B, g^B, h^B)$

$$H_0 = \{g^A = g^B\}$$

- For motif s, with

$$\widehat{\mathbb{E}}_0(N_s^A) = \widehat{\mathbb{E}}_{\widehat{\rho}^A, \widehat{g}^B, \widehat{h}^A}(N_s^A), \qquad \widehat{\mathbb{E}}_0(N_s^B)) = \widehat{\mathbb{E}}_{\widehat{\rho}^B, \widehat{g}^A, \widehat{h}^B}(N_s^B)$$

we have

$$W_s^{(g)}(A, B) = \frac{(N_s^A - \widehat{\mathbb{E}}_0(N_s^A)) - (N_s^B - \widehat{\mathbb{E}}_0(N_s^B))}{\sqrt{\widehat{\mathbb{V}}_0(N_s^A) + \widehat{\mathbb{V}}_0(N_s^B)}} \overset{H_0}{\sim} \mathcal{N}(0, 1)$$

Network 'distance' for pollinator imbalance

$$D^{(g)}(A, B) = \sqrt{\sum_s W_s^{(g)}(A, B)^2}$$

Results

Pollinator imbalance $D^{(g)}$. Adonis anova table

	Df	Sum Of Sqs	R^2	F	Pr(F)
InsectNb	1	69.9	0.2595	42.69	1e-05
PlantNb	1	31.17	0.1157	19.04	1e-05
Year	1	2.66	0.0099	1.62	0.22212
Month	6	24.8	0.092	2.52	0.00959
Region	2	8.67	0.0322	2.65	0.04531
Year:Month	6	4.81	0.0179	0.49	0.88756
Year:Region	2	5.51	0.0204	1.68	0.1787
Month:Region	12	32.41	0.1203	1.65	**0.06346**
Year:Month:Region	12	27.26	0.1012	1.39	0.15884
Residual	38	62.22	0.2309		
Total	81	269.42	1		

Results

Pollinator imbalance $D^{(g)}$. Adonis anova table

	Df	Sum Of Sqs	R^2	F	Pr(F)
InsectNb	1	69.9	0.2595	42.69	1e-05
PlantNb	1	31.17	0.1157	19.04	1e-05
Year	1	2.66	0.0099	1.62	0.22212
Month	6	24.8	0.092	2.52	0.00959
Region	2	8.67	0.0322	2.65	0.04531
Year:Month	6	4.81	0.0179	0.49	0.88756
Year:Region	2	5.51	0.0204	1.68	0.1787
Month:Region	12	32.41	0.1203	1.65	**0.06346**
Year:Month:Region	12	27.26	0.1012	1.39	0.15884
Residual	38	62.22	0.2309		
Total	81	269.42	1		

- Because of small network sizes, need to correct for the number of insects and plants

Results

Pollinator imbalance $D^{(g)}$. Adonis anova table

	Df	Sum Of Sqs	R^2	F	Pr(F)
InsectNb	1	69.9	0.2595	42.69	1e-05
PlantNb	1	31.17	0.1157	19.04	1e-05
Year	1	2.66	0.0099	1.62	0.22212
Month	6	24.8	0.092	2.52	0.00959
Region	2	8.67	0.0322	2.65	0.04531
Year:Month	6	4.81	0.0179	0.49	0.88756
Year:Region	2	5.51	0.0204	1.68	0.1787
Month:Region	12	32.41	0.1203	1.65	**0.06346**
Year:Month:Region	12	27.26	0.1012	1.39	0.15884
Residual	38	62.22	0.2309		
Total	81	269.42	1		

- Because of small network sizes, need to correct for the number of insects and plants
- Significant effect of the region and the month, indicating changes of the insect imbalance both in space and time

Results

Pollinator imbalance $D^{(g)}$. Adonis anova table

	Df	Sum Of Sqs	R^2	F	Pr(F)
InsectNb	1	69.9	0.2595	42.69	1e-05
PlantNb	1	31.17	0.1157	19.04	1e-05
Year	1	2.66	0.0099	1.62	0.22212
Month	6	24.8	0.092	2.52	0.00959
Region	2	8.67	0.0322	2.65	0.04531
Year:Month	6	4.81	0.0179	0.49	0.88756
Year:Region	2	5.51	0.0204	1.68	0.1787
Month:Region	12	32.41	0.1203	1.65	**0.06346**
Year:Month:Region	12	27.26	0.1012	1.39	0.15884
Residual	38	62.22	0.2309		
Total	81	269.42	1		

- Because of small network sizes, need to correct for the number of insects and plants
- Significant effect of the region and the month, indicating changes of the insect imbalance both in space and time
- The pattern is conserved from year to the next (not year effect)

Results

Pollinator imbalance $D^{(g)}$. Adonis anova table

	Df	Sum Of Sqs	R^2	F	Pr(F)
InsectNb	1	69.9	0.2595	42.69	1e-05
PlantNb	1	31.17	0.1157	19.04	1e-05
Year	1	2.66	0.0099	1.62	0.22212
Month	6	24.8	0.092	2.52	0.00959
Region	2	8.67	0.0322	2.65	0.04531
Year:Month	6	4.81	0.0179	0.49	0.88756
Year:Region	2	5.51	0.0204	1.68	0.1787
Month:Region	12	32.41	0.1203	1.65	**0.06346**
Year:Month:Region	12	27.26	0.1012	1.39	0.15884
Residual	38	62.22	0.2309		
Total	81	269.42	1		

- Because of small network sizes, need to correct for the number of insects and plants
- Significant effect of the region and the month, indicating changes of the insect imbalance both in space and time
- The pattern is conserved from year to the next (not year effect)
- No significant effect found for the plant imbalance distance $D^{(h)}$

Outline

Discrete time Markov-switching Hawkes process

Data [#8]. Y_k = number of bat calls during the k-th time bin.

[5] The proof does not rely on [AMR09]

Discrete time Markov-switching Hawkes process

Data [#8]. Y_k = number of bat calls during the k-th time bin.

Markov switching Hawkes process model. In discrete time:

- Hidden path $(Z_k)_{k \geqslant 1}$ = homogeneous Markov chain with Q states

$$(Z_k)_{k \geqslant 1} \sim MC_Q(\nu, \pi)$$

  ν = intial distribution, π = transition matrix;

- Observed counts: for $k \geqslant 1$ and

$$(Y_k \mid (Y_\ell)_{\ell \leqslant k-1}, Z_k = q) \sim \mathcal{P}\left(\mu_q + \alpha \sum_{\ell=1}^{k-1} \beta^{\ell-1} Y_{k-\ell}\right);$$

- Model parameters: $\theta = (\nu, \pi, \mu, \alpha, \beta)$

[5] The proof does not rely on [AMR09]

Discrete time Markov-switching Hawkes process

Data [#8]. Y_k = number of bat calls during the k-th time bin.

Markov switching Hawkes process model. In discrete time:

- Hidden path $(Z_k)_{k \geqslant 1}$ = homogeneous Markov chain with Q states

$$(Z_k)_{k \geqslant 1} \sim MC_Q(\nu, \pi)$$

ν = intial distribution, π = transition matrix;

- Observed counts: for $k \geqslant 1$ and

$$(Y_k \mid (Y_\ell)_{\ell \leqslant k-1}, Z_k = q) \sim \mathcal{P}\left(\mu_q + \alpha \sum_{\ell=1}^{k-1} \beta^{\ell-1} Y_{k-\ell}\right);$$

- Model parameters: $\theta = (\nu, \pi, \mu, \alpha, \beta)$

Proposition: Identifiability [BR25][5].

- The model parameter θ is identifiable from the joint distribution of (Y_1, Y_2, Y_3).
  $(\theta \neq \theta' \Rightarrow p_\theta(\cdot,\cdot,\cdot) \neq p_{\theta'}(\cdot,\cdot,\cdot))$

[5] The proof does not rely on [AMR09]

Markovian representation (homogeneous case)

Homogeneous discrete-time Hawkes process $Y = \{Y_k\}_{k \geqslant 1}$.

$$Y_k \mid (Y_\ell)_{\ell \leqslant k-1} \sim \mathcal{P}\left(\mu + \alpha \sum_{\ell=1}^{k-1} \beta^{\ell-1} Y_{k-\ell}\right)$$

$(Y_k)_{k \geqslant 1}$ is not a Markov chain (because of infinite memory).

Markovian representation (homogeneous case)

Homogeneous discrete-time Hawkes process $Y = \{Y_k\}_{k \geqslant 1}$.

$$Y_k \mid (Y_\ell)_{\ell \leqslant k-1} \sim \mathcal{P}\left(\mu + \alpha \sum_{\ell=1}^{k-1} \beta^{\ell-1} Y_{k-\ell}\right)$$

$(Y_k)_{k \geqslant 1}$ is not a Markov chain (because of infinite memory).

Markovian representation.

- Define

$$U_1 = 0, \qquad\qquad U_k = \alpha \sum_{\ell=1}^{k} \beta^{\ell-1} Y_{k-\ell},$$

Markovian representation (homogeneous case)

Homogeneous discrete-time Hawkes process $Y = \{Y_k\}_{k \geqslant 1}$.

$$Y_k \mid (Y_\ell)_{\ell \leqslant k-1} \sim \mathcal{P}\left(\mu + \alpha \sum_{\ell=1}^{k-1} \beta^{\ell-1} Y_{k-\ell}\right)$$

$(Y_k)_{k \geqslant 1}$ is not a Markov chain (because of infinite memory).

Markovian representation.

- Define

$$U_1 = 0, \qquad U_k = \alpha \sum_{\ell=1}^{k} \beta^{\ell-1} Y_{k-\ell},$$

- then, for $k \geqslant 1$ (with $U_0 = Y_0 = 0$)

$$U_k = \alpha Y_{k-1} + \beta U_{k-1}, \qquad Y_k \mid U_k \sim \mathcal{P}(\mu + U_k).$$

Markovian representation (homogeneous case)

Homogeneous discrete-time Hawkes process $Y = \{Y_k\}_{k \geqslant 1}$.

$$Y_k \mid (Y_\ell)_{\ell \leqslant k-1} \sim \mathcal{P}\left(\mu + \alpha \sum_{\ell=1}^{k-1} \beta^{\ell-1} Y_{k-\ell}\right)$$

$(Y_k)_{k \geqslant 1}$ is not a Markov chain (because of infinite memory).

Markovian representation.

- Define

$$U_1 = 0, \qquad U_k = \alpha \sum_{\ell=1}^{k} \beta^{\ell-1} Y_{k-\ell},$$

- then, for $k \geqslant 1$ (with $U_0 = Y_0 = 0$)

$$U_k = \alpha Y_{k-1} + \beta U_{k-1}, \qquad Y_k \mid U_k \sim \mathcal{P}(\mu + U_k).$$

so $((Y_k, U_k))_{k \geqslant 1}$ forms a Markov chain.

Markovian representation (Markov switching case)

Markov switching Hawkes process model. Can be rephrased as

$$(Y_k \mid (Y_\ell)_{\ell \leqslant k-1}, Z_k = q) \sim \mathcal{P}(\mu_q + U_k)$$

with

$$U_1 = 0, \qquad U_k = \alpha \sum_{\ell=1}^{k} \beta^{\ell-1} Y_{k-\ell},$$

Markovian representation (Markov switching case)

Markov switching Hawkes process model. Can be rephrased as

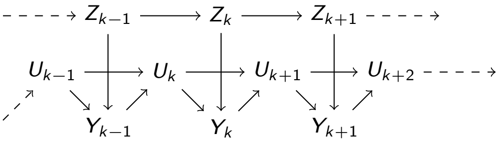
$$(Y_k \mid (Y_\ell)_{\ell \leqslant k-1}, Z_k = q) \sim \mathcal{P}(\mu_q + U_k)$$

with

$$U_1 = 0, \qquad U_k = \alpha \sum_{\ell=1}^{k} \beta^{\ell-1} Y_{k-\ell},$$

Consequence.

- The model is a regular Hidden Markov Model (HMM) with graphical model



$(Z_k)_{k\geqslant 1}$ = hidden path, $(U_k)_{k\geqslant 1}$ = memory, $(Y_k)_{k\geqslant 1}$ = observed process.

Inference

Maximum likelihood inference: $\widehat{\theta} = \arg\max_\theta \log p_\theta(Y)$

Inference

Maximum likelihood inference: $\widehat{\theta} = \arg\max_\theta \log p_\theta(Y)$

EM algorithm for HMM: [DLR77,CMR05]

$$\theta^{(h+1)} = \underbrace{\arg\max_\theta}_{\text{M step}} \underbrace{\mathbb{E}_{\theta^{(h)}}}_{\text{E step}} [\log p_\theta(Y, Z) \mid Y]$$

- E step: Evaluate $Q(\theta \mid \theta^{(h)}) = \mathbb{E}_{\theta^{(h)}}[\log p_\theta(Y, Z) \mid Y]$ (forward-backward recursion)
- M step: Gradient ascent, computing $\nabla_\theta Q(\theta \mid \theta^{(h)})$ by recursion

Inference

Maximum likelihood inference: $\widehat{\theta} = \arg\max_\theta \log p_\theta(Y)$

EM algorithm for HMM: [DLR77,CMR05]

$$\theta^{(h+1)} = \underbrace{\arg\max_\theta}_{\text{M step}} \underbrace{\mathbb{E}_{\theta^{(h)}}}_{\text{E step}} [\log p_\theta(Y, Z) \mid Y]$$

- E step: Evaluate $Q(\theta \mid \theta^{(h)}) = \mathbb{E}_{\theta^{(h)}}[\log p_\theta(Y, Z) \mid Y]$ (forward-backward recursion)
- M step: Gradient ascent, computing $\nabla_\theta Q(\theta \mid \theta^{(h)})$ by recursion

Model selection. Penalized likelihood

$$AIC_Q = \log p_{\widehat{\theta}_Q}(Y) - D_Q, \qquad BIC_Q = \log p_{\widehat{\theta}_Q}(Y) - D_Q \frac{\log(N)}{2}$$

with D_Q = number of parameters $= 2 + Q^2$ and N = number of time bins.

Simulation study (not shown)

Design.

1. Simulate a continuous time Markov-switching Hawkes process
2. With more or less events (control parameter λ)
3. Then discretise with more or less bins (control parameter $N \propto$ nb events)

Simulation study (not shown)

Design.

1. Simulate a continuous time Markov-switching Hawkes process
2. With more or less events (control parameter λ)
3. Then discretise with more or less bins (control parameter $N \propto$ nb events)

Conclusions.

- Inference more accurate when more signal (large λ)!!! [#54]
- Inference more accurate with thinner discretization step (large N)
  But at the price of a higher computational cost [#56]
- BIC does not capture the right number of states
  Sequences not simulated according to the model
- AIC does, when enough signal (λ) and discretization (N)
  Blind to the simulation shift from the model? [#55]

Simulation study (not shown)

Design.

1. Simulate a continuous time Markov-switching Hawkes process
2. With more or less events (control parameter λ)
3. Then discretise with more or less bins (control parameter $N \propto$ nb events)

Conclusions.

- Inference more accurate when more signal (large λ)!!! [#54]
- Inference more accurate with thinner discretization step (large N)
  But at the price of a higher computational cost [#56]
- BIC does not capture the right number of states
  Sequences not simulated according to the model
- AIC does, when enough signal (λ) and discretization (N)
  Blind to the simulation shift from the model? [#55]

Practical recommendations: Take $N = 2n$ and use AIC to choose Q.

Model 2: Bats calls sequences

Vigie-chiro project

Data set.

- Vigie-chiro project French participatory project to monitor bats echolocation calls (`https://www.vigienature.fr/fr/chauves-souris`).
- 2354 overnight recordings collected between October 2010 and January 2020 in 755 locations.
- Restricted to sequences with at least 50 calls $\rightarrow$ 1555 time sequences.

[6] BIC: Poisson $=$ 153 (homo $=$ 132, HMM $=$ 21), Hawkes $=$ 1402 (homo $=$ 775, HMM $=$ 627).

Vigie-chiro project

Data set.

- Vigie-chiro project French participatory project to monitor bats echolocation calls (`https://www.vigienature.fr/fr/chauves-souris`).
- 2354 overnight recordings collected between October 2010 and January 2020 in 755 locations.
- Restricted to sequences with at least 50 calls → 1555 time sequences.

Poisson vs Hawkes / Homogeneous vs HMM. Best model according to AIC[6]

	Poisson	Hawkes	Total
Homogeneous	34	353	387
Hidden Markov	24	1144	1168
Total	58	1497	1555

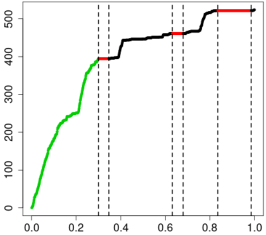
- Memory (95%) and heterogeneity (75%) are present in most sequences
- Hawkes-HMM best fits almost 3 sequences out of 4.

[6] BIC: Poisson = 153 (homo = 132, HMM = 21), Hawkes = 1402 (homo = 775, HMM = 627).

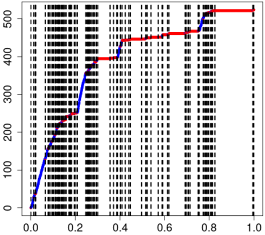
An example

Conditionally most probable states. (MAP)

Hawkes HMM ($\widehat{Q} = 3$)



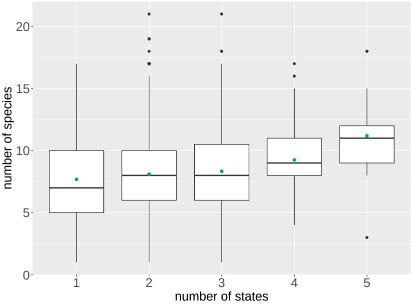
Poisson HMM ($\widehat{Q} = 4$)



- Interpretation of the states: absence of calls, transit and foraging (high call frequency)
- Hawkes-HMM state changes do not correspond to slope changes
- Poisson-HMM needs many state changes to account for self-excitation

States and species

The number of bat species was also recorded each night in each site.



- The number of states does not match the number of species
- More discussion to come with members of the Vigie-chiro project

Outline

Joint species distribution model

Data. n sites, p species,

- x_i = vector of covariates for site i,
- $Y_i = (Y_{i1}, \dots Y_{ip})$ = abundance vector in site i.

Joint species distribution model

Data. n sites, p species,

- x_i = vector of covariates for site i,
- $Y_i = (Y_{i1}, \dots Y_{ip})$ = abundance vector in site i.

Poisson log-normal (PLN) model.

- Latent layer:

$$(Z_i)_{1 \leqslant i \leqslant n} \text{ iid } \sim \mathcal{N}_p(0, \Sigma);$$

- Observed layer: counts $(Y_{ij})_{1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p}$ indep $\mid Z$

$$Y_{ij} \mid Z \sim \mathcal{P}\left(\exp(o_{ij} + x_i^\top \beta_j + Z_{ij})\right),$$

o_{ij} = given 'offset' term, accounting for the sampling effort;

- Parameters $\theta = (\beta, \Sigma)$:

$$\beta_j = \text{abiotic interactions}, \qquad \Sigma = \text{biotic interactions}.$$

An example

A typical dataset.

- Fish species from the Barents sea [FNA06],
- $n = 89$ sites, $p = 30$ species, $d = 4$ covariates (latitude, longitude, temperature, depth).

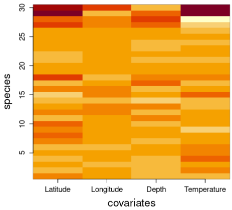
An example

A typical dataset.

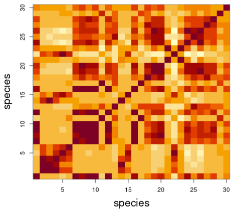
- Fish species from the Barents sea [FNA06],
- $n = 89$ sites, $p = 30$ species, $d = 4$ covariates (latitude, longitude, temperature, depth).

Importan aim of JSDM: Distinguish between abiotic and biotic effects:

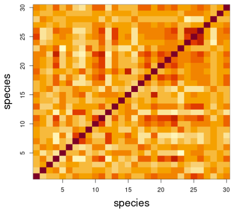
Covariate effects ($\widehat{B}$)



Correlation induced by the environment



Between species correlation ($\widehat{\Sigma}$)



Inference

Maximum likelihood inference via EM. [DLR77]

$$\theta^{(h+1)} = \underbrace{\arg\max_{\theta}}_{\text{M step}} \underbrace{\mathbb{E}_{\theta^{(h)}}}_{\text{E step}} [\log p_\theta(Y, Z) \mid Y]$$

- The E step requires some knowledge about $p_\theta(Z \mid Y)$, which turns out to be intractable for the PLN model.

Inference

Maximum likelihood inference via EM. [DLR77]

$$\theta^{(h+1)} = \underbrace{\arg\max_\theta}_{\text{M step}} \underbrace{\mathbb{E}_{\theta^{(h)}}}_{\text{E step}} [\log p_\theta(Y, Z) \mid Y]$$

- The E step requires some knowledge about $p_\theta(Z \mid Y)$, which turns out to be intractable for the PLN model.

Variational EM [WJ08,BKM17].

Inference

Maximum likelihood inference via EM. [DLR77]

$$\theta^{(h+1)} = \underbrace{\arg\max_{\theta}}_{\text{M step}} \underbrace{\mathbb{E}_{\theta^{(h)}}}_{\text{E step}} [\log p_\theta(Y, Z) \mid Y]$$

- The E step requires some knowledge about $p_\theta(Z \mid Y)$, which turns out to be intractable for the PLN model.

Variational EM [WJ08,BKM17].

- Choose a class $\mathcal{Q}$ of approximate (parametric) distributions and a divergence measure $D[q\|p]$ (e.g. $KL[q\|p]$)

Inference

Maximum likelihood inference via EM. [DLR77]

$$\theta^{(h+1)} = \underbrace{\arg\max_{\theta}}_{\text{M step}} \underbrace{\mathbb{E}_{\theta^{(h)}}}_{\text{E step}} [\log p_\theta(Y, Z) \mid Y]$$

- The E step requires some knowledge about $p_\theta(Z \mid Y)$, which turns out to be intractable for the PLN model.

Variational EM [WJ08,BKM17].

- Choose a class $\mathcal{Q}$ of approximate (parametric) distributions and a divergence measure $D[q \| p]$ (e.g. $KL[q \| p]$)
- VE step (approximation): $q^{(h+1)} = \arg\min_{q \in \mathcal{Q}} D\left[q(Z) \| p_{\theta^{(h)}}(Z \mid Y)\right]$

Inference

Maximum likelihood inference via EM. [DLR77]

$$\theta^{(h+1)} = \underbrace{\arg\max_{\theta}}_{\text{M step}} \underbrace{\mathbb{E}_{\theta^{(h)}}}_{\text{E step}} [\log p_\theta(Y, Z) \mid Y]$$

- The E step requires some knowledge about $p_\theta(Z \mid Y)$, which turns out to be intractable for the PLN model.

Variational EM [WJ08,BKM17].

- Choose a class $\mathcal{Q}$ of approximate (parametric) distributions and a divergence measure $D[q\|p]$ (e.g. $KL[q\|p]$)
- VE step (approximation): $q^{(h+1)} = \arg\min_{q \in \mathcal{Q}} D\left[q(Z) \| p_{\theta^{(h)}}(Z \mid Y)\right]$
- M step (update): $\theta^{(h+1)} = \arg\max_\theta \mathbb{E}_{q^{(h+1)}} [\log p_\theta(Y, Z)]$

Inference

Maximum likelihood inference via EM. [DLR77]

$$\theta^{(h+1)} = \underbrace{\arg\max_{\theta}}_{\text{M step}} \underbrace{\mathbb{E}_{\theta^{(h)}}}_{\text{E step}} [\log p_\theta(Y, Z) \mid Y]$$

- The E step requires some knowledge about $p_\theta(Z \mid Y)$, which turns out to be intractable for the PLN model.

Variational EM [WJ08,BKM17].

- Choose a class $\mathcal{Q}$ of approximate (parametric) distributions and a divergence measure $D[q\|p]$ (e.g. $KL[q\|p]$)
- VE step (approximation): $q^{(h+1)} = \arg\min_{q \in \mathcal{Q}} D\left[q(Z) \| p_{\theta^{(h)}}(Z \mid Y)\right]$
- M step (update): $\theta^{(h+1)} = \arg\max_\theta \mathbb{E}_{q^{(h+1)}} [\log p_\theta(Y, Z)]$
- If $D = KL$, a lower bound of $\log p_\theta(Y)$ ('ELBO') increases at each step

VEM for the Poisson log-normal model

Approximation class. Gaussian approximation [CMR18,CMR19]

$$q(Z) = \prod_{i=1}^{n} \mathcal{N}(Z_i; m_i, S_i)$$

- Parameter estimate $\widehat{\theta} = (\widehat{\Sigma}, \widehat{\beta})$,
- Approximate conditional distribution $Z_i \mid Y_i \approx \mathcal{N}(\widetilde{m}_i, \widetilde{S}_i)$,
- Lower bound $ELBO(\widehat{\theta}, \widetilde{m}, \widetilde{S})$ (R package `PLNmodels`)

VEM for the Poisson log-normal model

Approximation class. Gaussian approximation [CMR18,CMR19]

$$q(Z) = \prod_{i=1}^{n} \mathcal{N}(Z_i; m_i, S_i)$$

- Parameter estimate $\widehat{\theta} = (\widehat{\Sigma}, \widehat{\beta})$,
- Approximate conditional distribution $Z_i \mid Y_i \approx \mathcal{N}(\widetilde{m}_i, \widetilde{S}_i)$,
- Lower bound $ELBO(\widehat{\theta}, \widetilde{m}, \widetilde{S})$ (R package `PLNmodels`)

Variational inference.

- Reasonably easy to implement, fast, empirically accurate
- Very few theoretical guaranties: no general properties as for maximum likelihood (consistency, asymptotic normality, etc.)

  $\rightarrow$ No measure of uncertainty (no test, no confidence interval)
- Can we build upon variational inference to achieve 'genuine' statistical inference?

Toward genuine maximum likelihood inference [SR24]

Monte Carlo EM (MCEM). [CD85] When $p(Z \mid Y)$ can be sampled from:

- MCE step: Sample $(Z^m)_{m=1\dots M} \overset{iid}{\sim} p_{\theta^{(h)}}(Z \mid Y)$, then estimate

$$\widehat{Q}(\theta \mid \theta^{(h)}) := \frac{1}{M} \sum_{m=1}^{M} \log p_\theta(Y, Z^m)$$

- M step: Update

$$\theta^{(h+1)} = \arg\max_\theta \widehat{Q}(\theta \mid \theta^{(h)})$$

Toward genuine maximum likelihood inference [SR24]

Monte Carlo EM (MCEM). [CD85] When $p(Z \mid Y)$ can be sampled from:

- MCE step: Sample $(Z^m)_{m=1\dots M} \overset{iid}{\sim} p_{\theta^{(h)}}(Z \mid Y)$, then estimate

$$\widehat{Q}(\theta \mid \theta^{(h)}) := \frac{1}{M} \sum_{m=1}^{M} \log p_\theta(Y, Z^m)$$

- M step: Update

$$\theta^{(h+1)} = \arg\max_\theta \widehat{Q}(\theta \mid \theta^{(h)})$$

Importance sampling. When $p(Z \mid Y)$ can not be sampled from:

- Sample $(Z^m)_{m=1\dots M} \overset{iid}{\sim} q^{(h)}(Z)$, $q^{(h)}$ = proposal,
- Compute the non-normalized weights $\rho_m^{(h)} = p_{\theta^{(h)}}(Y, Z^m) \Big/ q^{(h)}(Z^m)$,

Toward genuine maximum likelihood inference [SR24]

Monte Carlo EM (MCEM). [CD85] When $p(Z \mid Y)$ can be sampled from:

- MCE step: Sample $(Z^m)_{m=1\dots M} \overset{iid}{\sim} p_{\theta^{(h)}}(Z \mid Y)$, then estimate

$$\widehat{Q}(\theta \mid \theta^{(h)}) := \frac{1}{M} \sum_{m=1}^{M} \log p_\theta(Y, Z^m)$$

- M step: Update

$$\theta^{(h+1)} = \arg\max_\theta \widehat{Q}(\theta \mid \theta^{(h)})$$

Importance sampling. When $p(Z \mid Y)$ can not be sampled from:

- Sample $(Z^m)_{m=1\dots M} \overset{iid}{\sim} q^{(h)}(Z)$, $q^{(h)}$ = proposal,
- Compute the non-normalized weights $\rho_m^{(h)} = p_{\theta^{(h)}}(Y, Z^m) \big/ q^{(h)}(Z^m)$,
- Estimate

$$\widehat{Q}(\theta \mid \theta^{(h)}) := \sum_{m=1}^{M} \rho_m^{(h)} \log p_\theta(Y, Z^m) \Bigg/ \sum_{m=1}^{M} \rho_m^{(h)} ,$$

Composite likelihood

Importance sampling has a poor efficiency[7] in 'large' dimension (say $p \geqslant 10, 15$)

$\rightarrow$ Need to reduce the sampling dimension

[7] Measured in terms of ESS $\simeq$ variance of the weights

Composite likelihood

Importance sampling has a poor efficiency[7] in 'large' dimension (say $p \geqslant 10, 15$)

$\rightarrow$ Need to reduce the sampling dimension

Composite likelihood.

- Build B overlapping blocks $\mathcal{C}_1, \dots \mathcal{C}_B$, each containing k species,
- Define the composite log-likelihood as

$$c\ell_\theta(Y) = \sum_{b=1}^{B} \log p_\theta(Y^b), \qquad \text{where} \quad Y^b = [Y_{ij}]_{i=1,\dots n, j \in \mathcal{C}_b},$$

[7] Measured in terms of ESS $\simeq$ variance of the weights

Composite likelihood

Importance sampling has a poor efficiency[7] in 'large' dimension (say $p \geqslant 10, 15$)

→ Need to reduce the sampling dimension

Composite likelihood.

- Build B overlapping blocks $\mathcal{C}_1, \dots \mathcal{C}_B$, each containing k species,
- Define the composite log-likelihood as

$$c\ell_\theta(Y) = \sum_{b=1}^{B} \log p_\theta(Y^b), \qquad \text{where} \quad Y^b = [Y_{ij}]_{i=1,\dots n, j \in \mathcal{C}_b},$$

- Then, the maximum composite likelihood estimator [VRF11]

$$\widehat{\theta}_{CL} = \arg\max_\theta c\ell_\theta(Y)$$

is consistent, asymptotically Gaussian with asymptotic variance given by

$$J(\theta) = \mathbb{V}_\theta[\nabla_\theta c\ell_\theta(Y)], \qquad H(\theta) = -\mathbb{E}_\theta[\nabla^2_\theta c\ell_\theta(Y)],$$
$$\mathbb{V}_\infty(\widehat{\theta}_{CL}) = H^{-1}(\theta) J(\theta) H^{-1}(\theta).$$

[7] Measured in terms of ESS $\simeq$ variance of the weights

Proposed composite likelihood algorithm

Proposition: EM applies to composite likelihood [SR24].

- Because the latent variables Z can be split in the same way as the observed abundances Y:

$$Z^b = [Z_{ij}]_{i=1,\dots n, j \in \mathcal{C}_b},$$

- The EM decomposition applies within each block.

[8] Not always possible: e.g., need to have $p \mid Bk$ and $p(p-1) \mid Bk(k-1)$

Proposed composite likelihood algorithm

Proposition: EM applies to composite likelihood [SR24].

- Because the latent variables Z can be split in the same way as the observed abundances Y:

$$Z^b = [Z_{ij}]_{i=1,\dots n, j \in \mathcal{C}_b},$$

- The EM decomposition applies within each block.

Proposal for importance sampling.

- Start with $q_b^{(1)}(Z^b) = \widetilde{q}_{VEM}(Z^b)$,
- Then update $q_b^{(h+1)}(Z^b)$ with the estimated mean and variance of $p_{\theta^{(h)}}(Z^b \mid Y^b)$.

[8] Not always possible: e.g., need to have $p \mid Bk$ and $p(p-1) \mid Bk(k-1)$

Proposed composite likelihood algorithm

Proposition: EM applies to composite likelihood [SR24].

- Because the latent variables Z can be split in the same way as the observed abundances Y:

$$Z^b = [Z_{ij}]_{i=1,\dots n, j \in \mathcal{C}_b},$$

- The EM decomposition applies within each block.

Proposal for importance sampling.

- Start with $q_b^{(1)}(Z^b) = \widetilde{q}_{VEM}(Z^b)$,
- Then update $q_b^{(h+1)}(Z^b)$ with the estimated mean and variance of $p_{\theta^{(h)}}(Z^b \mid Y^b)$.

Building the blocks. To guaranty the same precision for all estimates, one would ideally want that

β_j: each species j belongs to the same number of blocks $\mathcal{C}_1, \dots \mathcal{C}_B$,

$\sigma_{jj'}$: each pair of species (j, j') appears in the same number of blocks.

- Same problem as the construction of a incomplete balanced block design[8] [#58]

[8] Not always possible: e.g., need to have $p \mid Bk$ and $p(p-1) \mid Bk(k-1)$

Simulation study

Main aim. Assess normality

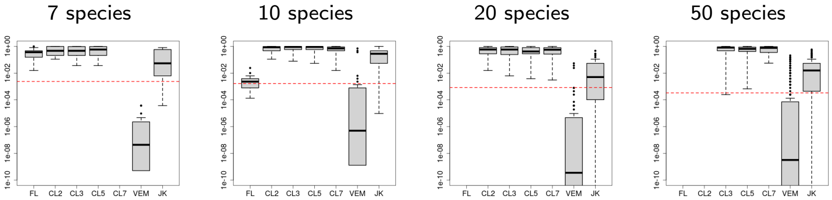
- Test statistic $(\widehat{\theta} - \theta^*) \Big/ \sqrt{\widehat{\mathbb{V}}_\infty(\widehat{\theta})}$ for the regression coefficients
- Criterion = p-value of the Kolmogorov-Smirnov test for normality
- Effect of the block size on the variance of the estimates (actually small [#59])

Simulation study

Main aim. Assess normality

- Test statistic $(\widehat{\theta} - \theta^*) \Big/ \sqrt{\widehat{\mathbb{V}}_\infty(\widehat{\theta})}$ for the regression coefficients
- Criterion = p-value of the Kolmogorov-Smirnov test for normality
- Effect of the block size on the variance of the estimates (actually small [#59])

Results. 100 sites, 3 covariates, 100 simulations



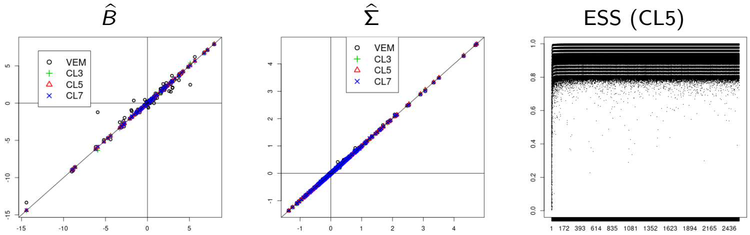
FL = full likelihood, CLk = composite likelihood($k = 2, 3, 5, 7$),

VEM = pseudo Fisher information matrix based on the ELBO,
JK = jackknife variance estimate of $\mathbb{V}(\widehat{\theta}_{VEM})$

Model 3: Fish species from the Barents sea

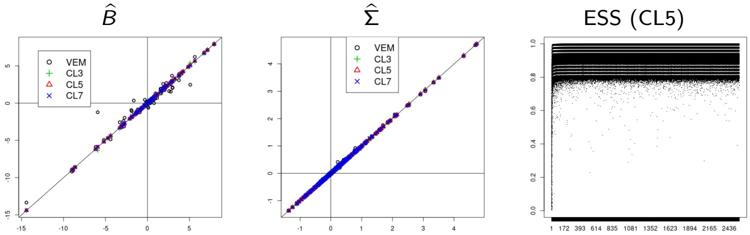
Fish species in the Barents sea

Comparison of the estimates.

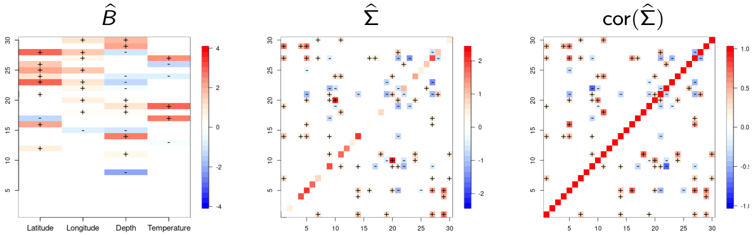


Fish species in the Barents sea

Comparison of the estimates.



Significance. Test statistics $\widehat{\theta} \Big/ \sqrt{\widehat{\mathbb{V}}_\infty(\widehat{\theta})}$



To conclude

Summary

Latent variable models.

- They are ubiquite in statistiscal ecology,
- For various modelling purposes (inferring Z is critical in Model 2, not in Models 1 and 3),
- Latent variables mays play different roles, from almost mechanistic to purely instrumental.

Summary

Latent variable models.

- They are ubiquite in statistiscal ecology,
- For various modelling purposes (inferring Z is critical in Model 2, not in Models 1 and 3),
- Latent variables mays play different roles, from almost mechanistic to purely instrumental.

Inference: No big picture.

- Dealing with latent variable yields specific difficulties, ranging from trivial to intractable,
- Often model-dependant, requiring specific developments,
- Still some generic questions (e.g. safely replace EM with gradient ascent?).

Discussion (some home works ?)

1 - Network motifs (plant-pollinator)

a - No clear understanding of the information brought by each motif;

b - (Asymptotic) normality does not hold for the networks at hand [#52]
(Could explain the poor power of the tests?).

c - BEDD is not consistent with the actual sampling process of the network;

Discussion (some home works ?)

1 - Network motifs (plant-pollinator)

a - No clear understanding of the information brought by each motif;

b - (Asymptotic) normality does not hold for the networks at hand [#52]
(Could explain the poor power of the tests?).

c - BEDD is not consistent with the actual sampling process of the network;

2 - Hawkes hidden Markov model (bats calls)

a - The Markovian representation also holds for non-exponential kernels [#57];

b - No theoretical problem to define a continuous time version of the proposed model
(but many practical ones);

c - A proper model selection criterion accounting for the discretization step is still needed.

Discussion (some home works ?)

1 - Network motifs (plant-pollinator)

a - No clear understanding of the information brought by each motif;

b - (Asymptotic) normality does not hold for the networks at hand [#52]
(Could explain the poor power of the tests?).

c - BEDD is not consistent with the actual sampling process of the network;

2 - Hawkes hidden Markov model (bats calls)

a - The Markovian representation also holds for non-exponential kernels [#57];

b - No theoretical problem to define a continuous time version of the proposed model
(but many practical ones);

c - A proper model selection criterion accounting for the discretization step is still needed.

3 - Poisson log-normal (species abundances)

a - Account for the 'excess' of null abundances [BCGM24];

b - Could we say more about the properties of VEM estimates?

c - The expression of $p_\theta(Z \mid Y)$ is hugly, but the function is actually very regular
$\rightarrow$ Could we 'learn' a deterministic transformation allowing, say, to sample from it?

References I

J. Aitchison and C.H Ho. The multivariate Poisson-log normal distribution. *Biometrika*, 76(4):643–653, 1989.

E.A. Allman, C. Matias, and J.A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, pages 3099–3132, 2009.

B. Batardière, J. Chiquet, F. Gindraud, and M. Mariadassou. Zero-inflation in the multivariate poisson lognormal family. Technical Report 2405.14711, arXiv, 2024.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

A. Bonnet and S. Robin. A markov switching discrete-time hawkes process: application to the monitoring of bats behavior. Technical Report 2507.20153, arXiv, 2025.

G. Celeux and J. Diebolt. The sem algorithm: a probabilistic teacher algorithm derived from the em algorithm for the mixture problem. *Computational statistics quarterly*, 2:73–82, 1985.

F. Chung and L. Lu. Connected components in random graphs with given expected degree sequences. *Annals of Combinatorics*, 6(2):125–145, 2002.

O. Cappé, E. Moulines, and T. Rydén. *Inference in Hidden Markov Models*. Springer, 2005.

J. Chiquet, M. Mariadassou, and S. Robin. Variational inference for probabilistic Poisson PCA. *The Annals of Applied Statistics*, 12(4):2674–2698, 2018.

J. Chiquet, M. Mariadassou, and S. Robin. Variational inference for sparse network reconstruction from count data. In *International Conference on Machine Learning*, pages 1162–1171, 2019.

J. Chiquet, M. Mariadassou, and S. Robin. The Poisson-lognormal model as a versatile framework for the joint analysis of species abundances. *Frontiers in Ecology and Evolution*, 9:188, 2021.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39:1–38, 1977.

References II

M. Fossheim, E. M Nilssen, and M. Aschan. Fish assemblages in the Barents Sea. *Marine Biology Research*, 2(4):260–269, 2006.

D. Inouye, E. Yang, G. I Allen, and P. Ravikumar. A review of multivariate distributions for count data derived from the Poisson distribution. *Computational Statistics*, 9(3), 2017.

M. Kirchner. Hawkes and INAR(∞) processes. *Stochastic Processes and their Applications*, 126(8):2494–2525, 2016.

T. Le Minh. u-statistics on bipartite exchangeable networks. *ESAIM: Probability and Statistics*, 27:576–620, 2023.

T. Le Minh, S. Donnet, F. Massol, and S. Robin. Hoeffding-type decomposition for u-statistics on bipartite networks. *Electronic Journal of Statistics*, 19(1):2829–2875, 2025.

B. H McArdle and M. J Anderson. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology*, 82(1):290–297, 2001.

Otso Ovaskainen and Nerea Abrego. *Joint species distribution modelling: With applications in R*. Cambridge University Press, 2020.

S. Ouadah, P. Latouche, and S. Robin. Motif-based tests for bipartite networks. *Electronic Journal of Statistics*, 16(1):293 – 330, 2022.

F. Picard, J.-J. Daudin, M. Koskas, S. Schbath, and S. Robin. Assessing the exceptionality of network motifs,. *J. Comp. Biol.*, 15(1):1–20, 2008.

N. Peyrard and O. Gimenez, editors. *Statistical Approaches for Hidden Variables in Ecology*. ISTE / Wiley, 2022.

B.I. Simmons, A. Cirtwill, N. Baker, L.V. Dicks, D.B. Stouffer, and W.J. Sutherland. Motifs in bipartite ecological networks: uncovering indirect interactions. *Oikos*, 128(2):154–170, 2019.

J. Stoehr and S. Robin. Composite likelihood inference for the poisson log-normal model. Technical Report 2402.14390, arXiv, 2024.

References III

S. Saavedra, R. P Rohr, J. M Olesen, and J. Bascompte. Nested species interactions promote feasibility over stability during the assembly of a pollinator community. *Ecology and evolution*, 6(4):997–1007, 2016.

B. I Simmons, M. JM Sweering, M. Schillinger, L. V Dicks, W. J Sutherland, and R. Di Clemente. bmotif: A package for motif analyses of bipartite networks. *Methods in Ecology and Evolution*, 10(5):695–701, 2019.

D. Stark. Compound Poisson approximations of subgraph counts in random graphs. *Random Structures & Algorithms*, 18(1):39–60, 2001.

C. Varin, N. Reid, and D. Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21:5–42, 2011.

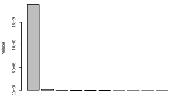
D. I Warton, F. G. Blanchet, R. B. O'Hara, O. Ovaskainen, S. Taskinen, S. C Walker, and F. KC. Hui. So many variables: joint modeling in community ecology. *Trends in Ecology & Evolution*, 30(12):766–779, 2015.

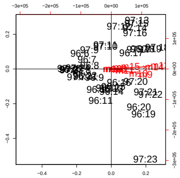
M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1–2):1–305, 2008.

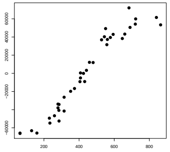
M. A Zapala and N. J Schork. Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. *Proceedings of the national academy of sciences*, 103(51):19430–19435, 2006.

Network embedding: Zackenberg's data [SROB16]

Raw counts			

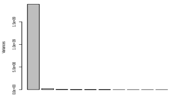
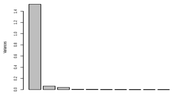
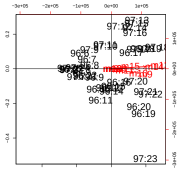
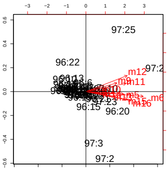
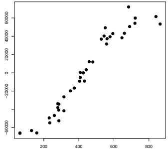
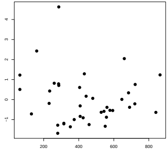






[#6]

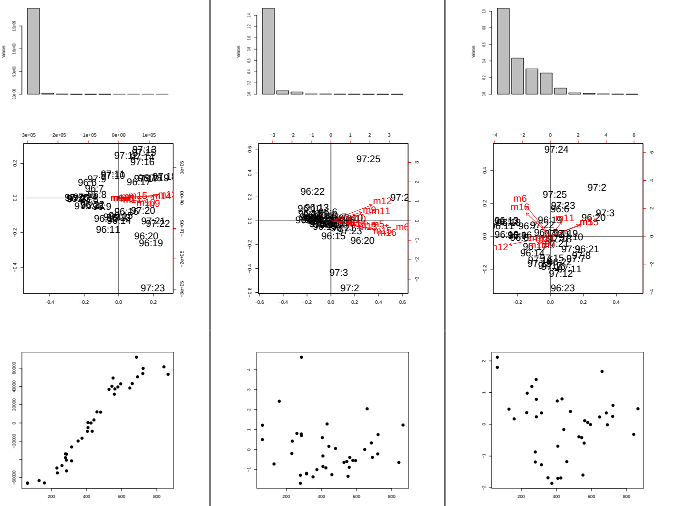
Network embedding: Zackenberg's data [SROB16]

Raw counts	Normalized counts		
			
			
			

[#6]

Network embedding: Zackenberg's data [SROB16]

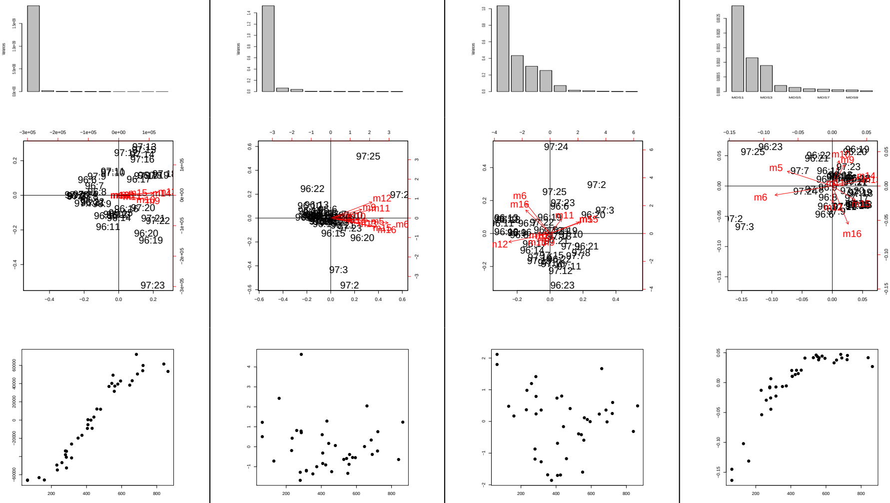
Raw counts	Normalized counts	Choleski



[#6]

Network embedding: Zackenberg's data [SROB16]

Raw counts	Normalized counts	Choleski	Bray-Curtis MDS



[#6]

Bipartite motifs

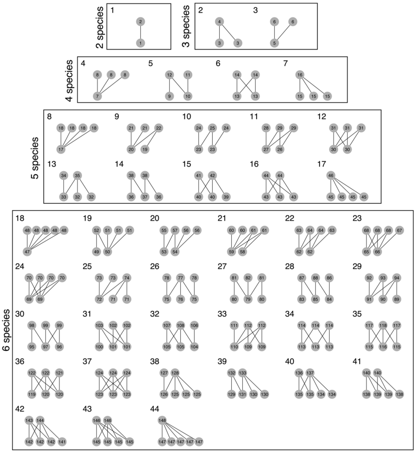
'Meso-scale' analysis. [SCB+19]

- Motifs ='building-blocks'
- between local (several nodes) and global (sub-graph)

Interest.

- Generic description of a network
- Enables network comparison
- Even when the nodes are different

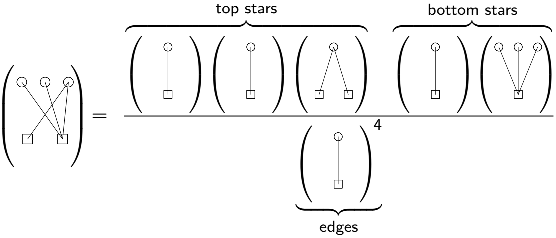
(+ 'species-role': out of the scope here) [#6]

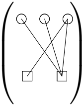


Existing tool. `bmotif` package [SSS+19]: counts motif occurrences (Not an easy task!)

Motif probability

Occurrence probability $\overline{\phi}_s = \mathbb{P}\{Y_{s\alpha} = 1\}$. Under the B-EDD model [OLR22]:



$\overline{\phi}_s = \mathbb{P}$  $= \dfrac{\left(\phi_1^2\phi_2\right)\left(\phi_1\phi_4\right)}{\left(\phi_1\right)^4} \qquad = \dfrac{\phi_2\phi_4}{\phi_1}$ [#49]

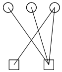
Estimated probability $\overline{F}_s$. [#15]

$$\overline{\phi}_s := \frac{\phi_2\phi_4}{\phi_1} \qquad \rightarrow \qquad \overline{F}_s := \frac{F_2 F_4}{F_1}$$

where F_1, F_2, F_4 = observed frequencies of edges, top stars and bottom stars.

Super-motifs

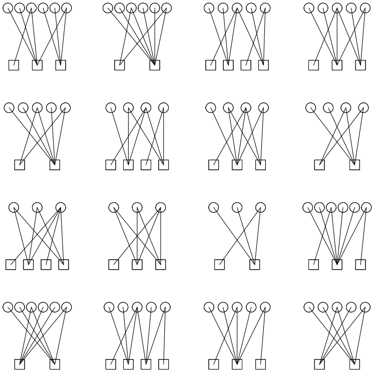
Motif:



Variance:

$$N_s^2 = \left(\sum_\alpha Y_{s\alpha}\right)^2$$
$$= \sum_{\alpha,\beta:\alpha\cap\beta=\emptyset} Y_{s\alpha} Y_{s\beta}$$
$$+ \sum_{\alpha,\beta:\alpha\cap\beta\neq\emptyset} \underbrace{Y_{s\alpha} Y_{s\beta}}_{\text{occurrence of a super-motif}}$$

Some super-motifs:

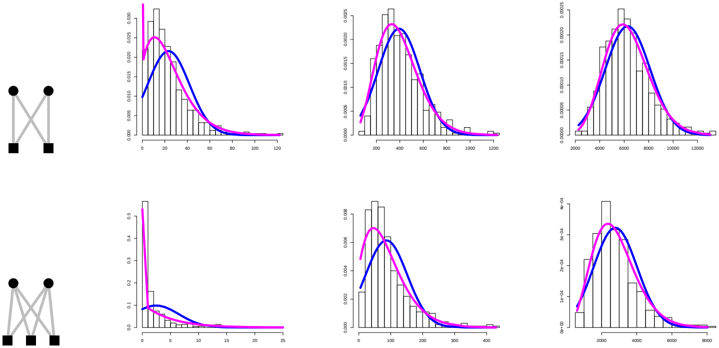


... 396 super-motifs

Covariance: same game, for $Y_{s\alpha} Y_{s'\beta}$ with $s \neq s'$ [#16]

In practice: Asymptotic normality

$(n = 2m/3)$ $m = 50$ $m = 100$ $m = 200$



Normal distribution, Poisson-geometric distribution with same mean and variance [Sta01,PDK$^+$08] [#16] [#44]

Self-exciting exponential Hawkes process

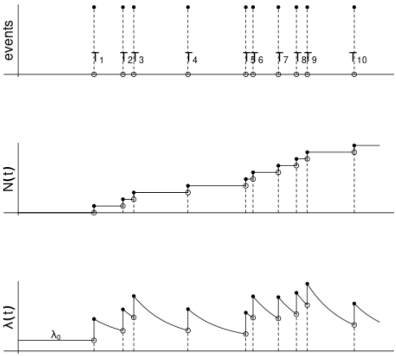
$$\lambda(t) = \lambda_0 + a \sum_{T_k < t} e^{-b(t - T_k)}$$

Self exciting: Each event increases the probability of observing another event

Self-exciting exponential Hawkes process

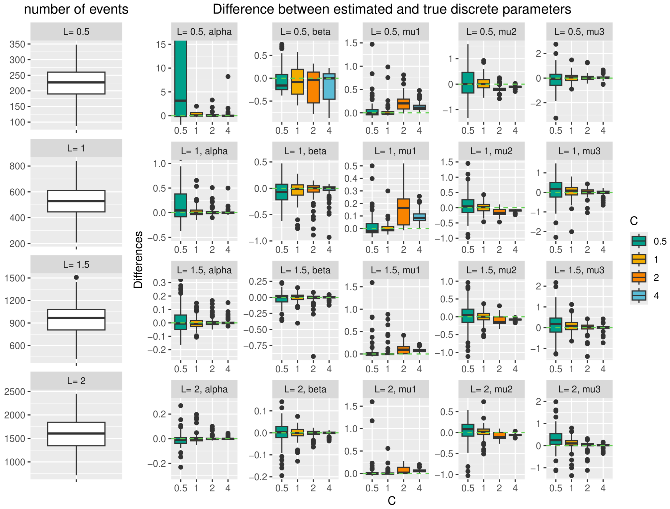
$$\lambda(t) = \lambda_0 + a \sum_{T_k < t} e^{-b(t-T_k)}$$

Self exciting: Each event increases the probability of observing another event



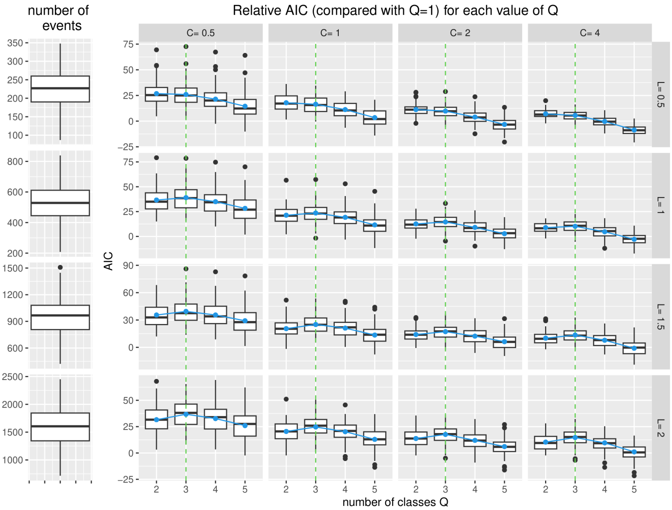
- Exponential kernel function $h(t) = ae^{-bt}$
- $a \geqslant 0$ to ensure that λ is non negative
- $a/b < 1$ to ensure stationarity
- Applications: sismology, epidemiology, vulcanology, neurosciences, ecology, ... [#9]

Simulations: estimation



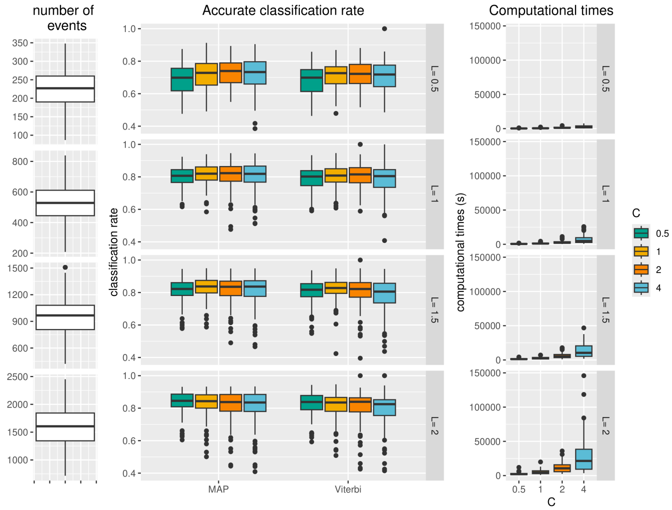
[#26]

Simulations: model selection



[#26]

Simulations: classification and computational time



[#26]

Non exponential kernel function h

Compact support. Suppose that h has no exponential form, but

$$t > L\Delta \qquad \Rightarrow \qquad h(t) = 0.$$

Homogeneous discrete Hawkes process.

$$\left(Y_k \mid Y_{1:(k-1)}\right) \sim \mathcal{P}\left(\mu + \sum_{\ell=1}^{k-1} \alpha_\ell Y_{k-\ell}\right) \qquad \text{with} \quad \alpha_\ell = \int_{(\ell-1)\Delta}^{\ell\Delta} h(t) \, \mathrm{d}t.$$

Markovian representation. Define $U_k = \sum_{\ell \geqslant 1} \alpha_\ell Y_{k-\ell}$, then

$$\left(Y_k \mid Y_{1:(k-1)}\right) = (Y_k \mid U_k) \sim \mathcal{P}\left(\mu + U_k\right)$$

and $((Y_k, U_k))_{k \geqslant 1}$ forms a Markov chain (of order L). [#44]

Number of block for composite likelihood

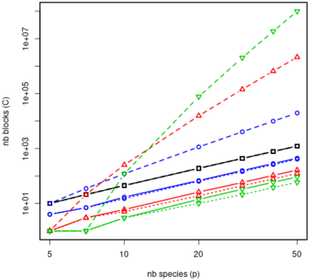


Figure 1: Number of blocks C as a function of the number of species p (in log-log-scale) for blocks of size $k = 2$ (black squares ■), $k = 3$ (blue circles ○), $k = 5$ (red triangles up △) and $k = 7$ (green triangles down ▽). Solid line: number of blocks actually used, dashed line: upper bound $\binom{p}{k}$, dotted line: lower bound $p(p-1)/[k(k-1)]$.

[#38]

Effect of the block size on the variance

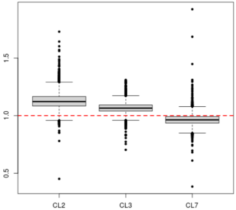


Figure 4: Boxplot of the relative variance of the estimates $\widehat{\beta}_{\ell j}$ of the regression coefficients obtained with the CL2, CL3 and CL7 algorithms, as compared to the CL5 algorithm for $p = 30$ species. Each boxplot is built across the $d \times p = 90$ normalised coefficients $\widetilde{\beta}_{\ell j}$.

[#39]