Bats Monitoring: A Classification Procedure of Bats Behaviors based on Hawkes Processes

École de printemps de la Chaire MMB 2024

Romain Lacoste, PhD student
In collaboration with C. Denis, C. Dion-Blanc, L. Sansonnet and Y. Bas

Aussois

June 12, 2024

LAMA Laboratoire d'Analyse et de Mathématiques Appliquées – UMR 8050
LPSM Laboratoire de probabilités statistique & modélisation
Modélisation Mathématique et Biodiversité

Table of contents

Table of contents

Two behaviors:

- **commuting** mode;
- **foraging** mode.

Goal: predicting the majority behavior of bats at sites throughout France.

- ▶ discriminate the **foraging** behavior from the **commuting** behavior.

Motivations:

- contribute to address spatial ecology issues;
- automate decision-making with few input variables.

Data: time of echolocation calls of **differents species** recorded as part of **Vigie-Chiro** participatory project.

- ▶ we focus on the **Common Pipistrelle**.

VIGIECHIRO

Echolocation: used by bats for **foraging** and **commuting**.

Behavioral characterization: via the way bats emit calls (see Griffin *et al.* (1960)).

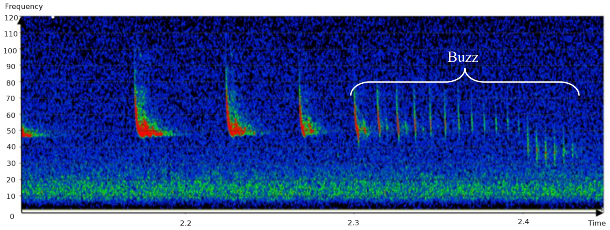


Figure: Sonogram containing a feeding buzz.

- consider the temporal distribution of the calls.
- sequence of calls $(T_\ell)_{\ell \geq 1}$ as a realization of a point process N.

Table of contents

Hawkes processes: family of point processes introduced in Hawkes (1971).

Exponential model: for $Y \in \{0, 1\}$, $\theta_Y \in \Theta$, conditional intensity given for $t \geq 0$ by:

$$\lambda_{\theta_Y}(t) := \mu_Y + \int_0^t \alpha_Y \beta_Y e^{-\beta_Y (t-s)} \, \mathrm{d}N(s) = \mu_Y + \sum_{T_\ell < t} \alpha_Y \beta_Y e^{-\beta_Y (t - T_\ell)},$$

where

- $\Theta = \{\mu > 0, 0 \leq \alpha < 1, \ \beta \geq 0\}$;
- $(T_\ell)_{\ell \geq 1}$ are the **jump times** of the process, Y the label.

Modeling: **the start time** of a call considered as a **jump** of the Hawkes process.

Classification: procedure is based on the likelihood and relies on Empirical Risk Minimization (ERM).

Table of contents

- Calls recorded over one night at 755 sites in France.

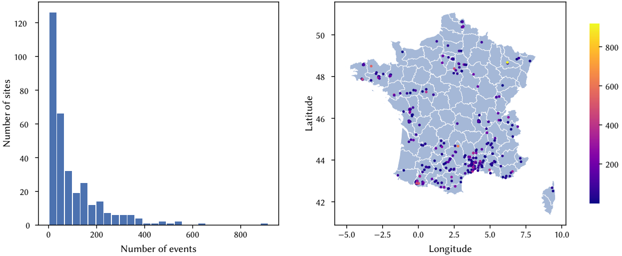


Figure: Each point on the map represents a site and its colour refers to the number of events in the temporal sequences.

- 332 labeled sites.
- 423 unlabeled sites.

Classification on labeled data: testing over 20 Monte-Carlo repetitions.

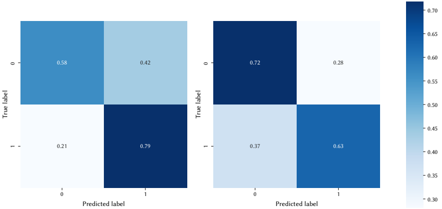


Figure: Confusion matrix of prediction on test data. Score: ERM: 68.13% (4.15), RF: 67.35% (2.21).

Prediction on unlabeled sites: tricky since bats have mixed behavior.

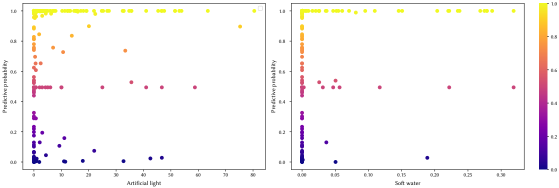


Figure: Predictive probability on unlabeled data as a function of environmental covariates.

Conclusion:

- Hawkes processes revelant for data modeling;
- classification procedure: prediction and behavioral confidence index;
- tool to ecologist for predicting bats behavior.

Bats Monitoring: A Classification Procedure of Bats Behaviors based on Hawkes Processes, C. Denis, C. Dion-Blanc, R.E. Lacoste, L. Sansonnet and Y. Bas (2023), The Journal of the Royal Statistical Society, Series C.

Perspectives:

- look at other species with more marked behavior;
- extension to multivariate Hawkes process.

Table of contents

Multidimensional linear Hawkes process (MHP)

Multivariate Hawkes process: $N = (N_1, \ldots, N_M)$ is defined by M point processes on $\mathbb{R}_+^*$.

- $M > 1$ is the dimension of the network.

***j*-th conditional intensity:** given for $t \geq 0$ by:

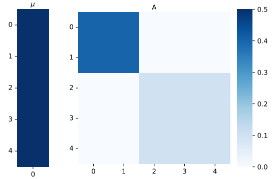
$$\lambda_j(t) := \mu_j + \sum_{j'=1}^{M} a_{j,j'} \int_0^t h(t-s)\, \mathrm{d}N_{j'}(s) = \mu_j + \sum_{j'=1}^{M} a_{j,j'} \sum_{T_{j',\ell} < t} h(t - T_{j',\ell}),$$

where

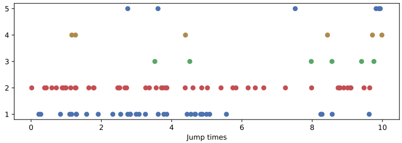
- $\mu = (\mu_1, \ldots, \mu_M) \in (\mathbb{R}_+^*)^M$ is the **exogenous intensity vector**;
- $A = (a_{j,j'})_{j,j'} \in \mathbb{R}_+^{M \times M}$ is the **interaction matrix**;
- $h : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ such that $\int_0^\infty h(t)dt \leq 1$ is the **kernel function**.

Modeling interaction within a network

Example: in a network of dimension $M = 5$.



(a) *exogenous intensity vector and interaction matrix*



(b) *Jump times of the associated MHP*

► a MHP models mutual excitation effects between connected components of a network, which depend on past interactions.

Parametric model

Parametrization: each λ_j depends on an unknown parameter θ^* belonging to:

$$\Theta := \left\{\mu \in (\mathbb{R}_+^*)^M,\ A \in \mathbb{R}_+^{M\times M},\ \rho(A) < 1\right\} \in \mathbb{R}_+^{M\times(M+1)},$$

where $\rho(A)$ is the spectral radius of A.

Assumption: N have **finite** exponential moment.

Modeling hypothesis: h **known**.

Notation: $\theta^* = (\mu^*, A^*) \in \Theta$ the true and **unknown** parameter.

- λ_{j,θ^*} the conditional intensity of the j-th component associated with this parameter.

Goal: **recover the support** of θ^*: $\mathrm{supp}(\theta^*)$.

Statistical setting

Let $T > 0$ be the upper bound of the observation interval.

Notation: $\mathcal{T}_T := \{\{T_{j,\ell}\}_{1 \leq \ell \leq N_j(T)}, 1 \leq j \leq M\}$ the jump times of a MHP $N = (N_1, \ldots, N_M)$ **observed in short time** on $[0, T]$.

Data: training n-sample $D_n := \{\mathcal{T}_T^{(1)}, \ldots, \mathcal{T}_T^{(n)}\}$ which consists of independent copies of $\mathcal{T}_T$.

Asymptotic setting: in $n \to \infty$ the number of trials (not in T as in Bacry, Bompaire, Gaïffas, and Muzy (2020)).

- the path may not have reached stationary regime.

High-dimensional framework

High-dimension: the dimension of the network M may be very large.

- ► in particular $M(M+1)$ may be larger than n.

Sparsity assumption: A^* **sparse**.

► individuals in the network only impacted by a small portion of other individuals.

Motivation:

- reduction of the problem dimension;
- facilitate interpretation;
- often very natural from a modeling standpoint.

Goodness-of-fit functional:

$$R_{T,n}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{T}\sum_{j=1}^{M}\int_0^T \lambda_{j,\theta}^{(i)}(t)^2 \,\mathrm{d}t - 2\int_0^T \lambda_{j,\theta}^{(i)}(t)\,\mathrm{d}N_j^{(i)}(t)\right),$$

where $\lambda_{j,\theta}^{(i)}(t)$ and $N_j^{(i)}(t)$ are defined from the i-th repetition.

Estimator:

$$\hat{\theta} \in \underset{\theta \in \mathbb{R}^{M\times(M+1)}}{\operatorname{argmin}} \left\{ R_{T,n}(\theta) + \kappa \sum_{j=1}^{M}\sum_{j'=1}^{M} |\theta_{j,j'}| \right\},$$

where κ is the regularization constant to be calibrated.

Table of contents

Some additional notations

For each $t \in (0, T]$ the **random matrix** $\mathbb{H}_t \in \mathbb{R}^{n\times(M+1)}$

$$(\mathbb{H}_t)_{i,j} = H_j^{(i)}(t), \quad \text{with} \quad H_j^{(i)}(t) := \int_0^t h(t-s)\,\mathrm{d}N_j^{(i)}(s),\ j \neq 0,\ H_0^{(i)} \equiv 1.$$

$$\mathbb{H} = \frac{1}{T}\int_0^T \mathbb{H}_t'\mathbb{H}_t\,\mathrm{d}t.$$

- $S_j^* := \{\theta_{j,j'}^* \neq 0,\ 0 \leq j' \leq M\}$ the **support of the j-th line** of θ^*
  - ▶ contains at least an element (as μ_j^* is non-zero).
- $\mathbb{H}_{S_j^*,S_j^*} := (\mathbb{H}_{j',j'})_{j'\in S_j^*}$.
  - ▶ submatrix given by deleting the rows and columns belonging to the complementary of the support $S_j^{*c} := \{\theta_{j,j'}^* = 0,\ 0 \leq j' \leq M\}$.

Assumptions on $\mathbb{H}$ (1/3)

Assumption 1: (Mutual incoherence)

There exists some $1 \geq \gamma > 0$ such that

$$\max_{j \in \{1,\dots,M\}} \|\mathbb{H}_{S_j^{*c}, S_j^*} \mathbb{H}^{-1}_{S_j^*, S_j^*}\|_\infty \leq 1 - \gamma \text{ a.s.}$$

- ensures there is not too much correlation between active and non-active variables;
- the incoherence parameter $\gamma \in (0, 1]$ must not be too small.

Assumption 2: (Minimum eigenvalue)

There exists $\Lambda_0 > 0$ such that

$$\min_{j\in\{1,\dots,M\}} \Lambda_{\min}\left(\frac{\mathbb{H}_{S_j^*,S_j^*}}{n}\right) \geq \Lambda_0 \ \text{ a.s.}$$

- imposes each matrix $\mathbb{H}_{S_j^*,S_j^*}$ to be invertible;
- identifiability of the problem restricted to each S_j^*;
- ensures that the submatrix $\mathbb{H}_{S_j^*,S_j^*}$ does not have its columns linearly dependent.

Assumption 3: (Minimum signal condition)

$$\min_{j,j' \in S^*} \left|\theta^*_{j,j'}\right| > \Lambda_0 \max_j \left|S^*_j\right|^2 \frac{\log^4(nM^2)}{\sqrt{n}}$$

- ensures that the non-zero entries of the true parameter are big enough to detect;
- imposes that the minimum value $\theta^*_{\min}$ (non-zero) cannot decay to zero faster than the regularization parameter κ chosen in the next theorem.

Theorem 1

Under assumptions 1, 2, et 3. Let $\kappa = \frac{\log^4(nM^2)}{\sqrt{n}}$. For n large enough, with probability greater than $1 - \frac{C_0}{n}$ with $C_0 > 0$, the penalized least-squares contrast admits a unique solution $\hat{\theta}$ which satisfies the following properties:

1. $\hat{\theta}_{j,j'} \geq 0$
2. $\mathrm{supp}(\hat{\theta}) = \mathrm{supp}(\theta^*)$;
3. $\left\|\hat{\theta} - \theta^*\right\|_\infty \leq \frac{\Lambda_0 \max_j |S_j^*|^2 \log^4(nM^2)}{\sqrt{n}}$

The proof follows the primal-dual-witness method (see Hastie, Tibshirani and Wainwright (2015)).

Implementation details (1/2)

Objective function : written as the sum of two functions.

$$R_{T,n}(\theta) + \kappa \sum_{j=1}^{M} \sum_{j'=1}^{M} |\theta_{j,j'}|$$

▶ use first-order optimization algorithm based on proximal methods with Nesterov's momentum method, namely **FISTA** (see Beck and Teboulle (2009)).

FISTA: new iterate is based on a specific linear combination of the previous two points.

- significantly faster rate of convergence than **ISTA**;
- additional computation cost is marginal (requires only one gradient evaluation per iteration as for **ISTA**);
- descent step used is $1/L$ with L the Lipschitz constant of the gradient

Calibration of κ: use **EBIC** criteria (see Chen (2008)).

EBIC: for some $\gamma \in [0, 1]$, $\kappa \in \Delta$

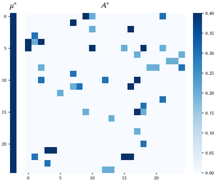
$$\mathrm{EBIC}_{\gamma}(\kappa) := -2L_{T,n}\left(\hat{\theta}(\kappa)\right) + \left|S_{\hat{\theta}(\kappa)}\right| \log(n) + 2\gamma \log\left(\binom{M^2}{\left|S_{\hat{\theta}(\kappa)}\right|}\right)$$

where $\hat{\theta}(\kappa)$ is the LASSO estimates with the tuning parameter κ, $L_{T,n}$ is the log-likelihood of the model, $\left|S_{\hat{\theta}(\kappa)}\right|$ is the number of active coefficients of $\hat{\theta}(\kappa)$.

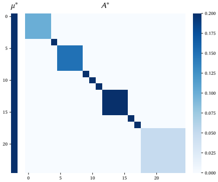
- relevant in a high-dimensional setting with parsimony assumptions;
- we choose the constant $\gamma = 1$
- we explore a grid of size $|\Delta| = 40$

Synthetic data generation: paths simulated by cluster process representation algorithm;

Panel of scenarios: vary the sparsity rate of A^* as well as its structure



(a) *Scenario 1*



(b) *Scenario 2*

Figure: $\theta^* = (\mu^*, A^*)$ in both scenarios. Sparsity rate A^* in *Scenario 1* : 92%, in *Scenario 2* : 85%.

Evaluation: using the following metrics

$$d_H\left(A^*, \hat{A}\right) = \frac{1}{M^2} \sum_{j,j'=1}^{M} \mathbb{1}_{\left\{A^*_{j,j'} \neq \hat{A}_{j,j'}\right\}}, \text{ and } d_{\ell_2}\left(A^*, \hat{A}\right) = \sqrt{\sum_{j,j'=1}^{M} \left|A^*_{j,j'} - \hat{A}_{j,j'}\right|^2};$$

Visual results for one repetition

- $M = 25$, $T = 5$, $h(s) = \beta \exp(-\beta s)$ with $\beta = 3$.

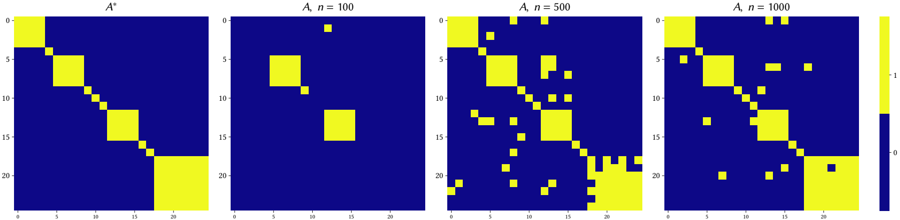


Figure: True support $\mathrm{supp}(\theta^*)$ and recovered support $\mathrm{supp}(\hat{\theta})$ in *Scenario 2*. The impact of n is investigated.

Results for 30 Monte-Carlo repetitions

	d_H			d_{l_2}		
	$n = 100$	$n = 500$	$n = 1000$	$n = 100$	$n = 500$	$n = 1000$
Scenario 1	0.03 (0.01)	0.03 (0.01)	0.03 (0.01)	0.96 (0.11)	0.44 (0.04)	0.32 (0.04)
Scenario 2	0.10 (0.01)	0.05 (0.01)	0.04 (0.01)	1.04 (0.10)	0.47 (0.06)	0.32 (0.03)

Table: Lasso results

▶ larger n is, the better the support is reconstructed, either in terms of Hamming distance or ℓ_2 distance.

	n	# events	time (sec)
Scenario 1	100	9524 (147)	68.64 (0.22)
	500	47651 (354)	334.13 (1.30)
	1000	95737 (632)	670.51 (2.67)

Table: Number of observed events, average execution time for *Scenario 1*.

▶ fast computational time (optimized C++ code).

Conclusion:

- consistency of the support and the convergence of the estimator;
- good numerical results on synthetic data;

ERM-LASSO classification rule for Multivariate Hawkes Processes paths, C. Denis, C. Dion-Blanc, R.E. Lacoste and L. Sansonnet, Soon on Hal.

Sparkle: a statistical learning toolkit for Hawkes process modeling in `Python`, R.E. Lacoste, In progress.

Perspectives:

- include inhibition interactions;
- ecological bat problem: each component of the MHP would model echolocation calls associated with a species;
  - ▶ model the effects of inter-species cooperation and competition between species.

Thank you for your attention!

Any questions?

Point processes: model the occurrence of random events over time.

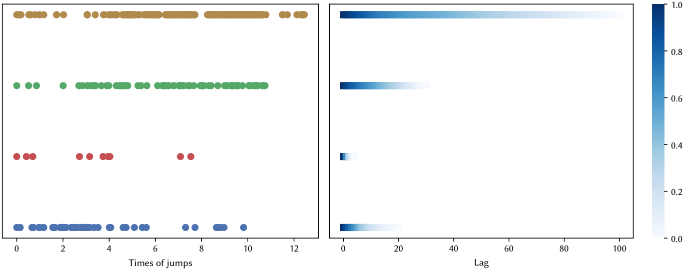


Figure: On the left are represented the start times of echolocation calls sequences, on the right it is the autocorrelation as a function of the lag for four nights.

- presence of strong temporal dependence in data.

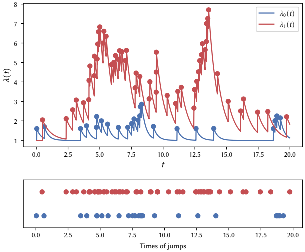
Mixture model

Let $\mathcal{D}_n^L = \left\{(\mathcal{T}_T^1, Y^1), \dots, (\mathcal{T}_T^n, Y^n)\right\}$ be a sample of i.i.d. observations such that:

- **Label:** $Y \sim \mathcal{B}(p^*)$;
- **Feature:** $\mathcal{T}_T = \left(T_1, \dots, T_{N_T}\right)$ of intensity $\lambda_{\theta_Y^*}(t)$ on $[0, T]$ with $\theta_Y^* \in \Theta$.

Goal: learn a decision rule g from $\mathcal{D}_n^L$ such that $g(\mathcal{T}_T)$ is a prediction of the label Y.

▶ given a new unlabeled feature $\mathcal{T}_T^{n+1}$, our guess for Y^{n+1} is $g\left(\mathcal{T}_T^{n+1}\right)$.



Quality of label prediction: measured by its missclassification risk

$$\mathcal{R}(g) := \mathbb{P}\left(g\left(\mathcal{T}_T^{n+1}\right) \neq Y^{n+1}\right).$$

Bayes rule and empirical risk minimization

Bayes rule: characterized by

$$g_{p^*,\boldsymbol{\theta}^*}(\mathcal{T}_T) = \mathbb{1}_{\left\{\eta_{p^*,\boldsymbol{\theta}^*}(\mathcal{T}_T) > \frac{1}{2}\right\}}$$

where $\eta_{p^*,\boldsymbol{\theta}^*}(\mathcal{T}_T) := \mathbb{P}(Y = 1 | \mathcal{T}_T) = \frac{p^* \exp\left(F_{\theta_1^*}(\mathcal{T}_T)\right)}{p^* \exp\left(F_{\theta_1^*}(\mathcal{T}_T)\right) + (1-p^*) \exp\left(F_{\theta_0^*}(\mathcal{T}_T)\right)}$

Empirical risk: based on $\mathcal{D}_n$ estimates $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{Y^i=1\}}$ and solve :

$$\hat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta}\in\Theta^2}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\left\{g_{\hat{p},\boldsymbol{\theta}}\left(\mathcal{T}_T^i\right) \neq Y^i\right\}}$$

- minimize this require to solve a non convex optimization problem.

Convexification: replace the $0-1$ loss by a **convex surrogate** (see Zhang (2004)) and based on $\mathcal{D}_n$ solve instead :

$$\hat{\theta} \in \underset{\theta \in \Theta^2}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(Z^i - f_{\hat{p},\theta}(\mathcal{T}_T^i) \right)^2$$

where $Z^i = 2Y_i - 1$ and $f_{\hat{p},\theta}(\mathcal{T}_T) = 2\eta_{\hat{p},\theta}(\mathcal{T}_T) - 1$.

Model: $\hat{\mathcal{F}} = \{2\eta - 1 : \ \eta \in \hat{H}\}$ where

$$\hat{H} = \left\{ \eta_{\hat{p},\theta}\left(\mathcal{T}_T\right) = \frac{\hat{p} \exp\left(F_{\theta_1}(\mathcal{T}_T)\right)}{\hat{p} \exp\left(F_{\theta_1}(\mathcal{T}_T)\right) + (1-\hat{p}) \exp\left(F_{\theta_0}(\mathcal{T}_T)\right)} \right\}$$

Classifier: $\hat{g}(\mathcal{T}_T) = \mathbb{1}_{\left\{\hat{f}(\mathcal{T}_T) \geq 0\right\}}$.

ERM procedure: provides estimates of (θ_0^*, θ_1^*).

- gives a model for the behavior within each class.

Model evaluation: by performing a goodness-of-fit test.

- using the **Time-Rescaling Theorem** (see Daley and Vere-Jones (2003)).

Theorem

Let $\Lambda(t) = \int_0^t \lambda(s)\, ds$ be the **compensator** of the process N. Then, a.s., the transformed sequence $\{\tau_j = \Lambda(T_j)\}$ is a realization of a unit-rate Poisson process if and only if the original sequence $\{T_j\}$ is a realization from the point process N.

Test H_0: "*the sequence of observations is a realization of the point process with intensity $\lambda_{\hat{\theta}_k}$*".

- test if $\{\Lambda_{\hat{\theta}_k}(T_{j+1}) - \Lambda_{\hat{\theta}_k}(T_j)\} \overset{\text{iid}}{\sim} \mathcal{E}(1)$

Test results

Labeled data:

	$\hat{g}(\mathcal{T})$	
	p-value	Acceptance Rate
Class 0	0.26 (0.06)	0.66 (0.11)
Class 1	0.15 (0.03)	0.45 (0.07)

Table: Mean p-values and reject rate for a 5% significance level test.

Unlabeled data:

	$\hat{g}(\mathcal{T})$	
	p-value	Acceptance Rate
Class 0	0.15	0.43
Class 1	0.21	0.49

Table: Mean p-values and acceptance rate for a 5% significance level test.