Non-parametric estimation for non-linear Hawkes processes

Vincent Rivoirard

Université Paris-Dauphine

Co-authors

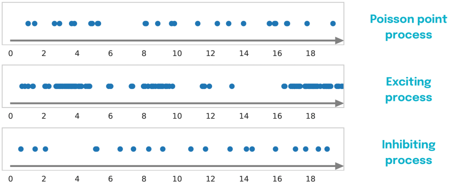
- DÉBORAH SULEM
  Pompeu Fabra University

- JUDITH ROUSSEAU
  Oxford University

Temporal point processes for event data

We wish to consider a probabilistic framework to model

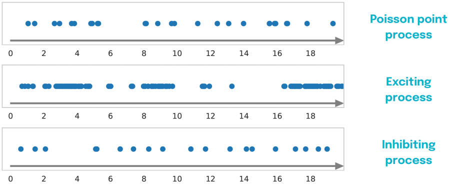
1. multivariate sequences of temporal events
2. with interactions between past and future occurrences
3. that can be positive (excitation) or negative (inhibition)



Temporal point processes for event data

We wish to consider a probabilistic framework to model

1. multivariate sequences of temporal events
2. with interactions between past and future occurrences
3. that can be positive (excitation) or negative (inhibition)

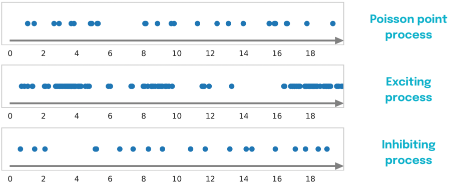


Alan G. Hawkes (1971a, 1971b, 1972) introduced a family of models for self-exciting and mutually exciting point processes. The "Hawkes process" terminology is due to Brillinger (1975) and Ogata (1978) and popularized by Daley and Vere-Jones (1988).

Temporal point processes for event data

We wish to consider a probabilistic framework to model

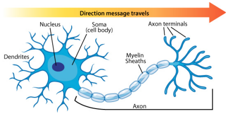
1. multivariate sequences of temporal events
2. with interactions between past and future occurrences
3. that can be positive (excitation) or negative (inhibition)

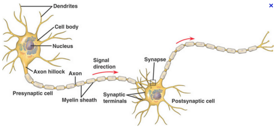


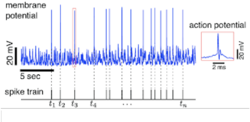
Alan G. Hawkes (1971a, 1971b, 1972) introduced a family of models for self-exciting and mutually exciting point processes. The "Hawkes process" terminology is due to Brillinger (1975) and Ogata (1978) and popularized by Daley and Vere-Jones (1988).

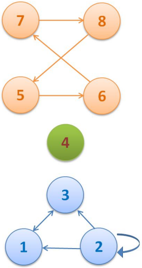
In the sequel, we shall consider non-linear versions of Hawkes processes.

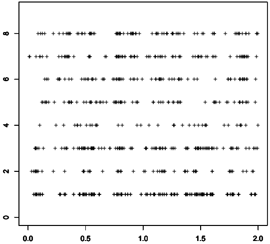
Functional connectivity graph of neurons











Our goal is to propose a scalable and adaptive statistical procedure to estimate parameters of the Hawkes model allowing to detect independence or exciting/inhibiting interactions between pairs of dimensions.

From linear to non-linear Hawkes processes

- A point process $N = (N_t)_{t \in \mathbb{R}}$ is a random countable set of points of $\mathbb{R}$ or equivalently a non-decreasing integer-valued process.
- The intensity λ_t of N represents the probability to observe a point at the time t conditionally on the past before t:

$$\lambda_t dt = \mathbb{P}(N \text{ has a jump} \in [t, t + dt] \mid N_s, s < t)$$

- Examples:
  - Poisson processes correspond to the case where $(\lambda_t)_t$ is not random. And the Poisson process is homogeneous if, in addition, λ_t does not depend on t.

From linear to non-linear Hawkes processes

- A point process $N = (N_t)_{t\in\mathbb{R}}$ is a random countable set of points of $\mathbb{R}$ or equivalently a non-decreasing integer-valued process.
- The intensity λ_t of N represents the probability to observe a point at the time t conditionally on the past before t:

$$\lambda_t dt = \mathbb{P}(N \text{ has a jump} \in [t, t+dt] \mid N_s, s < t)$$

- Examples:
  - Poisson processes correspond to the case where $(\lambda_t)_t$ is not random. And the Poisson process is homogeneous if, in addition, λ_t does not depend on t.
  - Linear univariate Hawkes process: with $\nu > 0$ and $h \geq 0$ supported by $\mathbb{R}_+$:

$$\lambda_t = \nu + \int_{-\infty}^{t-} h(t-u)dN_u = \nu + \sum_{X\in N,\, X<t} h(t-X)$$

ν is called the background rate and h the self-exciting function.

From linear to non-linear Hawkes processes

- A point process $N = (N_t)_{t\in\mathbb{R}}$ is a random countable set of points of $\mathbb{R}$ or equivalently a non-decreasing integer-valued process.
- The intensity λ_t of N represents the probability to observe a point at the time t conditionally on the past before t:

$$\lambda_t dt = \mathbb{P}(N \text{ has a jump} \in [t, t+dt] \mid N_s, s < t)$$

- Examples:
  - Poisson processes correspond to the case where $(\lambda_t)_t$ is not random. And the Poisson process is homogeneous if, in addition, λ_t does not depend on t.
  - Linear univariate Hawkes process: with $\nu > 0$ and $h \geq 0$ supported by $\mathbb{R}_+$:

$$\lambda_t = \nu + \int_{-\infty}^{t-} h(t-u)dN_u = \nu + \sum_{X\in N,\, X<t} h(t-X)$$

ν is called the background rate and h the self-exciting function.
Cluster representation (Hawkes and Oakes (1974)): A univariate linear Hawkes process can be viewed as a branching process over an homogeneous Poisson process

From linear to non-linear Hawkes processes

- A point process $N = (N_t)_{t \in \mathbb{R}}$ is a random countable set of points of $\mathbb{R}$ or equivalently a non-decreasing integer-valued process.
- The intensity λ_t of N represents the probability to observe a point at the time t conditionally on the past before t:

$$\lambda_t dt = \mathbb{P}(N \text{ has a jump} \in [t, t + dt] \mid N_s, s < t)$$

- Examples:

  - Poisson processes correspond to the case where $(\lambda_t)_t$ is not random. And the Poisson process is homogeneous if, in addition, λ_t does not depend on t.

  - Linear univariate Hawkes process: with $\nu > 0$ and $h \geq 0$ supported by $\mathbb{R}_+$:

$$\lambda_t = \nu + \int_{-\infty}^{t-} h(t-u) dN_u = \nu + \sum_{X \in N, X < t} h(t - X)$$

  ν is called the background rate and h the self-exciting function.
  Cluster representation (Hawkes and Oakes (1974)): A univariate linear Hawkes process can be viewed as a branching process over an homogeneous Poisson process

  - Non-linear univariate Hawkes process: with $\Phi \geq 0$,

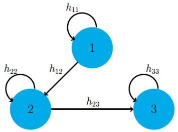
$$\lambda_t = \Phi\left(\int_{-\infty}^{t-} h(t-u) dN_u\right) = \Phi\Big(\sum_{X \in N, X < t} h(t - X)\Big)$$

Multivariate non-linear Hawkes processes

To model interactions between K neurons, we extend the previous expression. For a neuron $k \in [\![1; K]\!]$, we model its activity by a point process $N^{(k)}$ whose intensity is

$$\lambda_t^{(k)} = \psi_k\Big(\nu_k + \sum_{\ell=1}^{K} \int_{-\infty}^{t-} h_{\ell k}(t-u)dN^{(\ell)}(u)\Big)$$

$$= \psi_k\Big(\nu_k + \sum_{\ell=1}^{K} \sum_{X_\ell \in N^{(\ell)},\, X_\ell < t} h_{\ell k}(t - X_\ell)\Big)$$



- ψ_k: nonnegative and nondecreasing link function
  - linear link function: $\psi_k(x) = x$ but requires $h_{\ell k} \geq 0$ for all ℓ
  - example of non-linear link function: $\psi_k(x) = x_+ = \max(x, 0)$ (ReLU)
- $\nu_k > 0$: background rates
- $h_{\ell k}$: interaction functions
  - If $h_{\ell k} = 0$: $N^{(k)}$ is locally independent of $N^{(\ell)}$
  - If $h_{\ell k}$ is positive: $N^{(\ell)}$ excites $N^{(k)}$
  - If $h_{\ell k}$ is negative: $N^{(\ell)}$ inhibits $N^{(k)}$
  - If $h_{\ell k}$ is signed: excitation and inhibition

Multivariate Hawkes processes

Definition

A K-dimensional continuous time process $N = (N_t)_t = (N_t^{(1)}, \dots, N_t^{(K)})_t$ is a multivariate non-linear Hawkes process if

(i) almost surely, for $k \neq \ell$, $(N_t^{(k)})_t$ and $(N_t^{(\ell)})_t$ never jump simultaneously

(ii) for all k, the intensity of $(N_t^{(k)})_t$ is given by

$$\lambda_t^{(k)} = \psi_k\Big(\nu_k + \sum_{\ell=1}^{K} \int_{-\infty}^{t-} h_{\ell k}(t-u)dN^{(\ell)}(u)\Big).$$

- Existence and uniqueness of a stationary distribution for N established by Brémaud and Massoulié (1996, 2001).
- See also Delattre, Fournier and Hoffmann (2016) and Costa, Graham, Marsalle and Tran (2020) for other relevant probabilistic results.
- Statistical Goal: Estimation of $f = (\nu_k, (h_{\ell k})_{\ell \in [\![1;K]\!]})_{k \in [\![1;K]\!]}$ based on observations of $N = (N^{(k)})_{k \in [\![1;K]\!]}$ on $[0, T]$ with intensity process $(\lambda^{(k)})_{k \in [\![1;K]\!]}$.

Nonlinear Hawkes processes: State of the art and our contribution

Hawkes (2018) claimed : "*Some function of the intensity gives us a non-linear Hawkes process. These are more difficult to deal with, and therefore not frequently used.*"

State of the art for non-linear Hawkes processes:

- Asymptotic analysis of second order statistics (cross-covariance): Chen, Shojaie, Shea-Brown and Witten (2019) extended by Cai, Zhang and Guan (2022)
- Parametric approaches for exponential interaction functions: Lemonnier and Vayatis (2014), Bonnet, Martinez Herrera and Sangnier (2021, 2023) and Deutsch and Ross (2022).
- Variational Bayes algorithms: For very specific link functions, Zhou, Kong, Zhang, Feng and Zhu (2021) and Malem-Shinitski, Ojeda and Opper (2021) developed efficient Bayesian algorithms based on mean-field approximations and augmented likelihood. However, these methods do not consider the high-dimensional nonparametric setting.

Our contribution: Scalable nonparametric Bayesian estimation in the multivariate setting for the non-linear case

Inference for non-linear Hawkes models

- We observe $N = (N^{(k)})_{k \in [\![1;K]\!]}$ on $[0, T]$ with intensity process $(\lambda^{(k)})_{k \in [\![1;K]\!]}$ given by

$$\lambda_t^{(k)} = \psi\Big(\nu_k + \sum_{\ell=1}^{K} \int_{-\infty}^{t-} h_{\ell k}(t-u)dN^{(\ell)}(u)\Big)$$

  where $\psi : \mathbb{R} \longmapsto \mathbb{R}_+$ is known and non-decreasing

- Assumptions:
  - the ν_k's are positive
  - the $h_{\ell k}$'s are bounded
  - the support of the $h_{\ell k}$'s is included into $[0, A]$, with $A < \infty$ known

  We do not assume that the $h_{\ell k}$'s are non-negative, so inhibition is possible.

- Statistical goals: Bayesian estimation of

$$f = (\nu_k, (h_{\ell k})_{\ell \in [\![1;K]\!]})_{k \in [\![1;K]\!]}$$

  with in mind $T \to +\infty$

Stationarity

Intensity process of $N = (N^{(k)})_{k \in [\![1;K]\!]}$:

$$\lambda_t^{(k)} = \psi\Big(\nu_k + \sum_{\ell=1}^{K} \int_{-\infty}^{t-} h_{\ell k}(t-u) dN^{(\ell)}(u)\Big) = \psi\Big(\nu_k + \sum_{\ell=1}^{K} \sum_{\substack{X_\ell \in N^{(\ell)} \\ X_\ell < t}} h_{\ell k}(t - X_\ell)\Big)$$

with $\psi : \mathbb{R} \longmapsto \mathbb{R}_+$ known and non-decreasing. Extension of results by Brémaud and Massoulié (1996, 2001):

Proposition

If one of the following conditions is satisfied:

(S1) ψ *is bounded:* $\exists \Lambda > 0, \forall x \in \mathbb{R}, \psi(x) \leq \Lambda$

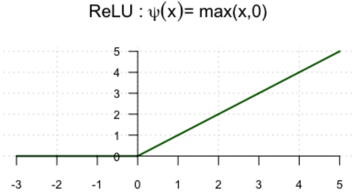
(S2) ψ *is L-Lipschitz, with* $L > 0$ *and* $\|S^+\|$, *the spectral norm of the matrix* S^+ *with entries* $S^+_{\ell k} = L\|h^+_{\ell k}\|_1$ *satisfies* $\|S^+\| < 1$

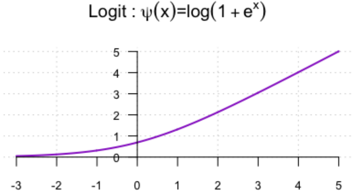
then there exists a unique stationary version of the process N with finite average

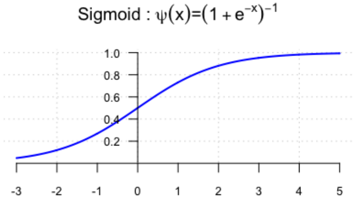
Notation:

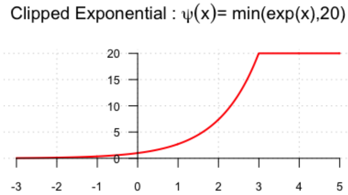
$$h^+_{\ell k}(x) = \max(h_{\ell k}(x), 0), \quad h^-_{\ell k}(x) = \max(-h_{\ell k}(x), 0)$$

Typical link functions









Identifiability

Intensity process of $N = (N^{(k)})_{k \in [\![1;K]\!]}$:

$$\lambda_t^{(k)} = \psi\Big(\nu_k + \sum_{\ell=1}^{K} \int_{-\infty}^{t-} h_{\ell k}(t-u)dN^{(\ell)}(u)\Big) = \psi\Big(\nu_k + \sum_{\ell=1}^{K} \sum_{\substack{X_\ell \in N^{(\ell)} \\ X_\ell < t}} h_{\ell k}(t - X_\ell)\Big)$$

with $\psi : \mathbb{R} \longmapsto \mathbb{R}_+$ known, non-decreasing and L-Lipschitz.

Proposition

If ψ is bijective on an open interval I so that for any k

$$[\nu_k - \max_\ell \|h_{\ell k}^-\|_\infty ; \nu_k + \max_\ell \|h_{\ell k}^+\|_\infty] \subset I,$$

then the distribution of N is identifiable for T large enough.

Remark: Identifiability is satisfied

- for logit $\psi(x) = \log(1 + e^x)$ and sigmoid $\psi(x) = (1 + e^{-x})^{-1}$ link functions
- for the ReLU function, $\psi(x) = \max(x, 0)$, we assume for any k,

$$\max_\ell \|h_{\ell k}^-\|_\infty < \nu_k$$

- for the clipped exponential function, $\psi(x) = \min(e^x, \Lambda)$, we assume for any k,

$$\max_\ell \|h_{\ell k}^+\|_\infty + \nu_k < \log \Lambda$$

Bayesian inference framework

- We observe over a time window $[-A, T]$ a stationary K-dimensional Hawkes process N with unknown parameter $f_0 = (\nu^0, h^0) = (\nu_k^0, (h_{\ell k}^0)_{\ell \in [\![1;K]\!]})_{k \in [\![1;K]\!]}$.
- The log-likelihood for a parameter $f = (\nu, h) = (\nu_k, (h_{\ell k})_{\ell \in [\![1;K]\!]})_{k \in [\![1;K]\!]}$ is

$$L_T(f) := \sum_{k=1}^{K} L_T^k(f), \quad L_T^k(f) = \int_0^T \log(\lambda_t^k(f)) dN_t^k - \int_0^T \lambda_t^k(f) dt.$$

- Let Π a prior distribution on the parameter space $\mathcal{F}$. The posterior distribution is:

$$\Pi(B|N) = \frac{\int_B \exp(L_T(f)) d\Pi(f)}{\int_{\mathcal{F}} \exp(L_T(f)) d\Pi(f)}, \quad B \subset \mathcal{F}.$$

  Remark: The posterior distribution is doubly intractable.
- Questions:
  - When $T \to +\infty$, does $\Pi(\cdot|N)$ concentrate around f_0?
  - If yes, at which rate?
- We shall consider the $\mathbb{L}_1$-loss:

$$\|f - f_0\|_1 := \|\nu - \nu^0\|_{\ell_1} + \sum_{k=1}^{K} \sum_{\ell=1}^{K} \|h_{\ell k} - h_{\ell k}^0\|_1$$

Posterior concentration rates

We assume previous conditions to obtain stationarity and identifiability are satisfied.

Theorem

Assume

$$\inf_x \psi(x) > 0. \tag{1}$$

Let $\epsilon_T = o(1)$ *be a positive sequence verifying* $\log^3 T = O(T\epsilon_T^2)$. *We set for* $B > 0$

$$B(\epsilon_T, B) = \Big\{ f \in \mathcal{F}; \quad \|\nu - \nu^0\|_{\ell_\infty} \le \epsilon_T,\ \max_{\ell,k} \|h_{\ell k} - h^0_{\ell k}\|_\infty \le \epsilon_T,\ \max_{\ell,k} \|h_{\ell k}\|_\infty < B \Big\}.$$

Let Π *be a prior distribution on* $\mathcal{F}$. *We assume that for* T *large enough:*

- $\exists\, c_1 > 0$ *s.t.* $\Pi(B(\epsilon_T, B)) \ge e^{-c_1 T \epsilon_T^2}$
- $\exists\, \mathcal{F}_T \subset \mathcal{F}$, $\zeta_0 > 0$ *and* $x_0 > 0$ *such that*

$$\Pi(\mathcal{F}_T^c) = o(e^{-c_1 T\epsilon_T^2}), \quad \log \mathcal{N}\left(\zeta_0 \epsilon_T, \mathcal{F}_T, \|\cdot\|_1\right) \le x_0 T \epsilon_T^2$$

Then, for $M > 0$ *large enough, we have*

$$\mathbb{E}_0\left[\Pi\left(\|f - f_0\|_1 > M\epsilon_T \middle| N\right)\right] = o(1).$$

Posterior concentration rates

- Assumption $\inf_x \psi(x) = 0$ is strong.
- The result of the theorem holds by replacing ϵ_T with $\epsilon_T\sqrt{\log T}$ if we only assume that $\psi(x) > 0$ for any $x \in \mathbb{R}$, and $\sqrt{\psi}$ and $\log(\psi)$ are Lipschitz functions. This is satisfied by logit, sigmoid and clipped exponential functions.

Posterior concentration rates

- Assumption $\inf_x \psi(x) = 0$ is strong.
- The result of the theorem holds by replacing ϵ_T with $\epsilon_T \sqrt{\log T}$ if we only assume that $\psi(x) > 0$ for any $x \in \mathbb{R}$, and $\sqrt{\psi}$ and $\log(\psi)$ are Lipschitz functions. This is satisfied by logit, sigmoid and clipped exponential functions.
- The result of the theorem holds by replacing ϵ_T with $\epsilon_T \log T$ if

$$\psi(x) = \max(x, 0)$$

and if we further assume that

$$\limsup_{T \to +\infty} \frac{1}{T} \mathbb{E}_0 \left[\int_0^T \frac{1_{\{\lambda_t^{(k)}(f_0) > 0\}}}{\lambda_t^{(k)}(f_0)} dt \right] < +\infty, \quad \forall k \in [\![1; K]\!].$$

This is satisfied if for instance for any ℓ $h^0_{\ell k}$ is an histogram and for all t, $h^0_{\ell k}(t) \in \mathbb{Q}$

Posterior concentration rates

- Assumption $\inf_x \psi(x) = 0$ is strong.
- The result of the theorem holds by replacing ϵ_T with $\epsilon_T \sqrt{\log T}$ if we only assume that $\psi(x) > 0$ for any $x \in \mathbb{R}$, and $\sqrt{\psi}$ and $\log(\psi)$ are Lipschitz functions. This is satisfied by logit, sigmoid and clipped exponential functions.
- The result of the theorem holds by replacing ϵ_T with $\epsilon_T \log T$ if

$$\psi(x) = \max(x, 0)$$

and if we further assume that

$$\limsup_{T \to +\infty} \frac{1}{T} \mathbb{E}_0 \left[\int_0^T \frac{1_{\{\lambda_t^{(k)}(f_0) > 0\}}}{\lambda_t^{(k)}(f_0)} dt \right] < +\infty, \quad \forall k \in [\![1; K]\!].$$

This is satisfied if for instance for any ℓ $h^0_{\ell k}$ is an histogram and for all t, $h^0_{\ell k}(t) \in \mathbb{Q}$

- The case

$$\psi(x) = \theta + \max(x, 0), \quad x \in \mathbb{R},$$

with θ unknown and positive can be dealt with. Under the same assumptions of the theorem, we also achieve the rate ϵ_T.

Spike and slab prior distribution

We define a prior distribution on $f = (\nu_k, (h_{\ell k})_{\ell \in [\![1;K]\!]})_{k \in [\![1;K]\!]}$ of the form

$$d\Pi(f) = d\Pi_h(h) \prod_k d\Pi_\nu(\nu_k),$$

with

1. Π_ν having a positive and continuous density on $\mathbb{R}_+^*$, e.g. a Gamma distribution.
2. For $h = (h_{\ell k})_{\ell,k}$, we write

$$h_{\ell k} = \delta_{\ell k} \bar{h}_{\ell k}, \quad \delta_{\ell k} \in \{0,1\}, \quad \delta_{\ell k} \neq 0 \iff \bar{h}_{\ell k} \neq 0$$

so that $\delta = (\delta_{\ell k})_{\ell k}$ is the connectivity graph. We then consider

(a) $\delta \sim \Pi_\delta$, where Π_δ is a prior on $\{0,1\}^{K^2}$, e.g. $\delta_{\ell k} \overset{i.i.d.}{\sim} \mathcal{B}er(p)$

(b) Given δ, we use a truncated distribution on $h|\delta$ of the form

$$d\Pi_h(h|\delta) \propto \left(\prod_{\ell,k} d\tilde{\Pi}_{h|\delta}(h_{\ell k}) \right) \times 1_{\|S^+\|<1}(h),$$

with

$$\tilde{\Pi}_{h|\delta}(h_{\ell k}) = \delta_{\ell k} \tilde{\Pi}_h(\bar{h}_{\ell k}) + (1 - \delta_{\ell k}) \delta_{\{0\}}(\bar{h}_{\ell k}),$$

and $\tilde{\Pi}_h$ is a nonparametric prior, e.g. a random histogram, or a spline prior

Minimax rate on Hölder classes

Corollary

Assume all interaction functions are Hölderian functions:

$$h^0_{\ell k} \in \mathcal{H}(\beta, L_0), \quad 1 \leq \ell, k \leq K,$$

with $\beta > 0$ and $L_0 > 0$. Then, under the previous prior,

$$\mathbb{E}_0 \left[\Pi \left(\|f - f_0\|_1 \gtrsim \epsilon_T \middle| N \right) \right] = o(1),$$

with

$$\epsilon_T = T^{-\frac{\beta}{2\beta+1}} (\log T)^{\square},$$

which is optimal up to the logarithmic term. Furthermore, with

$$(\hat{\nu}, \hat{h}) = \mathbb{E}^{\Pi}[f|N] = \int_{\mathcal{F}} f d\Pi(f|N),$$

$\hat{f}$ converging to f_0 at the rate ϵ_T for the $\mathbb{L}_1$-norm:

$$\mathbb{P}_0 \left(\|\hat{f} - f_0\|_1 \gtrsim \epsilon_T \right) = o(1).$$

Numerical results - Histogram case

- We sample one observation of a Hawkes process with K neurons, link function ψ and parameter $f_0 = (\nu^0, h^0)$ on $[0, T]$. We take $A = 0.1$.
- We assume $h^0 \in \mathcal{H}^D_{histo}$ for some $D \geq 1$, with

$$\mathcal{H}^D_{histo} = \left\{ h = (h_{\ell k})_{\ell,k};\ h_{\ell k}(x) = \sum_{j=1}^{2^D} w^j_{\ell k} e_j(x),\ x \in [0, A] \right\}, \quad e_j = \frac{2^D}{A} 1_{\left[\frac{A(j-1)}{2^D}, \frac{Aj}{2^D}\right)}$$

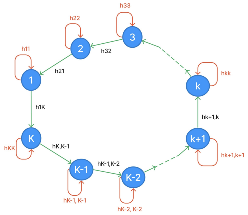


Figure: True graph: $2K - 1$ non-zero interaction functions. Scenario 1 corresponds to self-excitation and Scenario 2 corresponds to self-inhibition

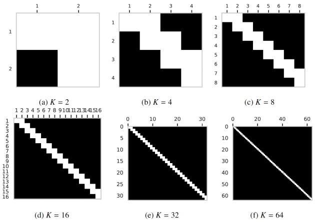


Figure: True graph for different dimensions: $2K - 1$ non-zero interaction functions for $K \in \{2, 4, 8, 16, 32, 64\}$ (correspond to white squares)

Variational Bayesian estimation

- Difficulty of computing the nonparametric posterior distribution since

$$\Pi(B|N) = \frac{\int_B e^{L_T(f)} d\Pi(f)}{\int_{\mathcal{F}} e^{L_T(f)} d\Pi(f)} \quad e^{L_T(f)} = \prod_{k=1}^{K} \left[e^{-\int_0^T \lambda_t^k(f) dt} \prod_{\substack{X_k \in N^{(k)} \\ X_k \le T}} \lambda_{X_k}^{(k)}(f) \right]$$

and $\lambda_t^{(k)}(f) = \psi\Big(\nu_k + \sum_{\ell=1}^K \sum_{\substack{X_\ell \in N^{(\ell)} \\ X_\ell < t}} h_{\ell k}(t - X_\ell)\Big)$

- Instead, we approximate the posterior distribution and use Variational Bayes methods. Let $\mathcal{V}$ be an approximating family of distributions on $\mathcal{F}$.

$$\hat{Q} := \arg\min_{Q \in \mathcal{V}} KL\left(Q || \Pi(.|N)\right), \quad KL(Q||Q') := \begin{cases} \int \log\left(\frac{dQ}{dQ'}\right) dQ & \text{if } Q \ll Q' \\ +\infty & \text{otherwise} \end{cases}$$

- Standard assumptions + $\min_{Q \in \mathcal{V}} KL(Q||\Pi(.|N)) = O(T\epsilon_T^2)$ gives

$$\mathbb{E}_0\left[\hat{Q}\left(\|f - f_0\|_1 > \epsilon_T\right)\right] = o(1)$$

- A common choice of variational class is a mean-field family:

$$\mathcal{V}_{MF} = \left\{ Q : \ dQ(\vartheta) = \prod_{d=1}^{D} dQ_d(\vartheta_d) \right\}.$$

Augmented mean-field variational inference

- The log-likelihood function of the non-linear Hawkes model is augmented with some latent variable $z \in \mathcal{Z}$, with $\mathcal{Z}$ the latent parameter space. We denote $L_T^A(f, z)$ the augmented log-likelihood and define the augmented posterior distribution as

$$\Pi_A(B|N) = \frac{\int_B e^{L_T^A(f,z)} d(\Pi(f) \times \mathbb{P}_A(z))}{\int_{\mathcal{F}\times\mathcal{Z}} e^{L_T^A(f,z)} d(\Pi(f) \times \mathbb{P}_A(z))}, \quad B \subset \mathcal{F} \times \mathcal{Z},$$

where $\mathbb{P}_A$ is a prior distribution on z and we consider

$$\mathcal{V}_{AMF} = \left\{ Q : \mathcal{F} \times \mathcal{Z} \to [0,1];\ Q(f,z) = Q_1(f)Q_2(z) \right\}.$$

- The augmented mean-field variational posterior is defined as

$$\hat{Q}_{AMF}(f,z) := \arg\min_{Q \in \mathcal{V}_{AMF}} KL\left(Q(f,z)||\Pi_A(f,z|N)\right) =: \hat{Q}_1(f)\hat{Q}_2(z)$$

and verifies

$$\hat{Q}_1(f) \propto \exp\left(\mathbb{E}_{\hat{Q}_2}[\log p(f,z,N)]\right), \quad \hat{Q}_2(z) \propto \exp\left(\mathbb{E}_{\hat{Q}_1}[\log p(f,z,N)]\right),$$

where $p(f,z,N)$ is the joint density of the parameter, the latent variable, and the observations $\Rightarrow$ Iterative algorithm that updates each factor alternatively

Adaptive variational Bayes algorithm in the sigmoid model

- We consider the sigmoid case

$$\psi(x) = (1 + e^{-x})^{-1}$$

and follow the augmentation strategy proposed by Zhou, Kong, Zhang, Feng and Zhu (2021) and Malem-Shinitski, Ojeda and Opper (2021) based on a Gaussian representation of ψ in terms of Pólya-Gamma variables.

- For certain families of Gaussian priors, $\hat{Q}_1$ and $\hat{Q}_2$ are conjugate to the priors, which allows to design iterative algorithms with closed-forms updates.
- More precisely, in the following parametrization for the prior model:

$$d\Pi(f) = d\Pi_h(h) \prod_k d\Pi_\nu(\nu_k),$$

we write

$$h_{\ell k} = \delta_{\ell k} \bar{h}_{\ell k}, \quad \delta_{\ell k} \in \{0, 1\}, \quad \delta_{\ell k} \neq 0 \iff \bar{h}_{\ell k} \neq 0,$$

and

$$\bar{h}_{\ell k}(x) = \sum_j w_{\ell k}^j e_j(x), \quad w_{\ell k}^j \sim \mathcal{N}(0, \sigma^2)$$

For fixed δ, the previous strategy is tractable.

Augmented mean-field variational for the sigmoid case

- Strategy proposed by Zhou, Kong, Zhang, Feng and Zhu (2021) and Malem-Shinitski, Ojeda and Opper (2021) for the sigmoid case

$$\psi(x) = (1 + e^{-x})^{-1}.$$

1. If p_{PG} is the Polya-Gamma density

$$\psi(x) = \mathbb{E}_{\omega \sim p_{PG}}\left[e^{g(\omega,x)}\right], \quad g(\omega, x) = -\frac{\omega x^2}{2} + \frac{x}{2} - \log 2$$

2. Campbell's theorem: For a Poisson point process $\bar{N}$ on a space $\mathcal{X}$ with intensity measure $\Lambda : \mathcal{X} \to \mathbb{R}^+$, and for any function $\zeta : \mathcal{X} \to \mathbb{R}$

$$\exp\left(\int (e^{\zeta(x)} - 1)\Lambda(dx)\right) = \mathbb{E}\left[\prod_{x \in \bar{N}} e^{\zeta(x)}\right].$$

- Using these ideas, we obtain the doubly augmented log-likelihood:

$$L_T^A(f, \omega, \bar{Z}; N) = \sum_{k \in [K]} \left\{ \sum_{i \in [N_k]} \left[g(\omega_i^k, \tilde{\lambda}_{T_i^k}(f)) + \log p_{PG}(\omega_i^k; 1, 0)\right] + \sum_{j \in [\bar{N}_k]} \left[g(\bar{\omega}_j^k, -\tilde{\lambda}_{\bar{T}_j}(f)) + \log p_{PG}(\bar{\omega}_j^k; 1, 0)\right] \right\}.$$

Model selection

- How to estimate δ?
- Model selection for Variational Bayes: Compute VB posterior $\hat{Q}_\delta$ and

$$\mathrm{ELBO}(\hat{Q}_\delta) = \mathbb{E}_{\hat{Q}_\delta}\left[\log \frac{p(f,z,N)}{\hat{Q}_\delta(f,z)}\right]$$

Choose

$$\hat{\delta} = \mathrm{argmax}_\delta \mathrm{ELBO}(\hat{Q}_\delta)$$

With $\delta = (\delta_{\ell k})_{1\le \ell,k\le K} \in \{0,1\}^{K^2}$, we have 2^{K^2} models: intractable as soon as K is moderately large.

- We propose the following alternative:
  1. We apply the previous strategy with $\delta_{\ell k} = 1$ for any ℓ, k.
  2. We order the obtained $\mathbb{L}_1$-norm of the interaction functions $\|\hat{h}_{\ell k}\|_1$
  3. We determine the largest jump providing a threshold η and set

$$\hat{\delta}_{\ell k} = 0 \iff \|\hat{h}_{\ell k}\|_1 \le \eta.$$

  4. We apply the previous strategy with $\hat{\delta}$.

Numerical experiments

We investigate the behavior of our procedure with respect to:
- the dimension K
- the graph sparsity
- model mis-specification
- the support of interaction functions: A

Numerical performances

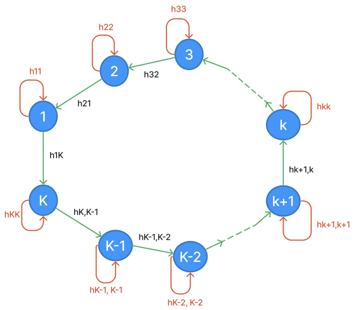


Figure: True (sparse) graph: $2K - 1$ non-zero interaction functions.
- Green edges: excitation
- Red edges: self-excitation (scenario 1) or self-inhibation (scenario 2)

K	Scenario	# observations
2	Self-excitation ($T = 500$) Self-inhibition ($T = 700$)	5 680 4 800
4	Self-excitation ($T = 500$) Self-inhibition $T = 700$)	11 338 9 895
8	Self-excitation ($T = 500$) Self-inhibition $T = 700$)	22 514 19 746
16	Self-excitation ($T = 500$) Self-inhibition $T = 700$)	51 246 37 166
32	Self-excitation ($T = 500$) Self-inhibition $T = 700$)	96 803 76 106
64	Self-excitation ($T = 200$) Self-inhibition ($T = 300$)	117 862 133 200

Numerical performances

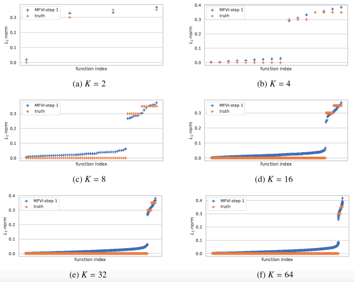


Figure: Estimated $\mathbb{L}_1$-norms of interaction functions plotted in increasing order in the Self-excitation scenario for $K \in \{2, 4, 8, 16, 32, 64\}$

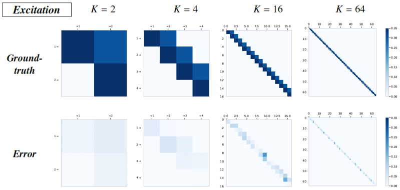


Figure: Heatmaps of the entries of the matrix $(\|h^0_{\ell k}\|_1)_{\ell,k}$ (top) and $(\mathbb{E}[\|h^0_{\ell k} - h_{\ell k}\|_1])_{\ell,k}$ (bottom) in the Self-excitation scenario.

Numerical performances

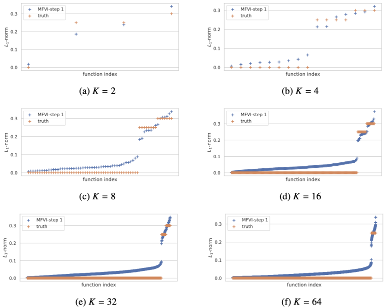


Figure: Estimated $\mathbb{L}_1$-norms of interaction functions plotted in increasing order in the Self-inhibition scenario for $K \in \{2, 4, 8, 16, 32, 64\}$

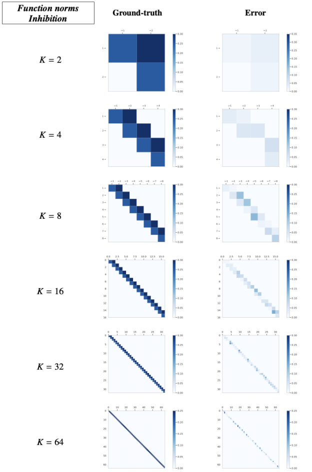


Figure: Heatmaps of the entries of the matrix $(\|h^0_{\ell k}\|_1)_{\ell,k}$ (left) and $(\mathbb{E}[\|h^0_{\ell k} - h_{\ell k}\|_1])_{\ell,k}$ (right) in the Self-inhibition scenario.

Numerical performances

K	Scenario	$\hat{\delta} = \delta_0$	Risk
2	Self-excitation Self-inhibition	Yes Yes	0.79 0.35
4	Self-excitation Self-inhibition	Yes Yes	1.01 0.92
8	Self-excitation Self-inhibition	Yes Yes	2.10 2.12
16	Self-excitation Self-inhibition	Yes Yes	5.77 4.48
32	Self-excitation Self-inhibition	Yes Yes	10.57 8.53
64	Self-excitation Self-inhibition	Yes Yes	23.74 18.33

Table: Performance of Algorithm. We report the $\mathbb{L}_1$-risk and if the model with largest marginal probability corresponds to the true one.

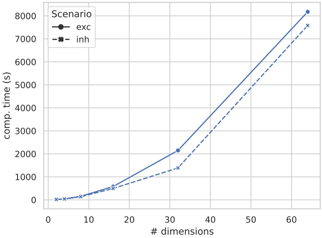


Figure: Computational times of our two-step mean-field variational algorithm in the Excitation (exc) and Self-inhibition (inh) scenarios for $K = 2, 4, 8, 16, 32, 64$.

$$\|f - f_0\|_1 := \|\nu - \nu^0\|_{\ell_1} + \sum_{k=1}^{K} \sum_{\ell=1}^{K} \|h_{\ell k} - h^0_{\ell k}\|_1$$

Numerical performances

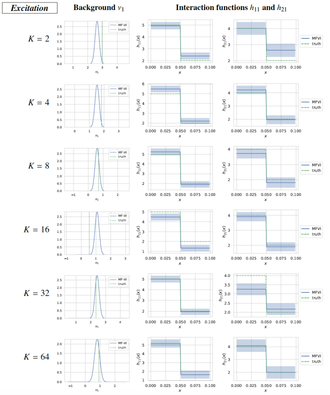


Figure: Mode variational posterior distributions on ν_1 (left column) and interaction functions h_{11} and h_{21} (second and third columns) in the excitation scenario.

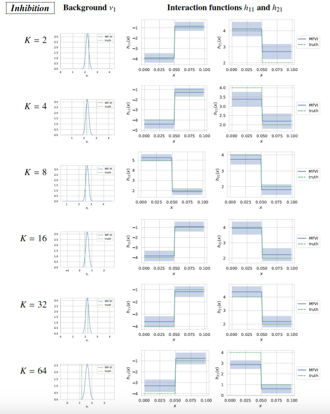
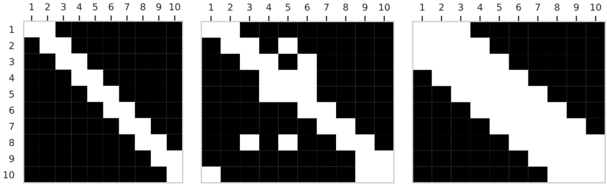


Figure: Mode variational posterior distributions on ν_1 (left column) and interaction functions h_{11} and h_{21} (second and third columns) in the self-inhibition scenario.

Numerical performances - Graph sparsity

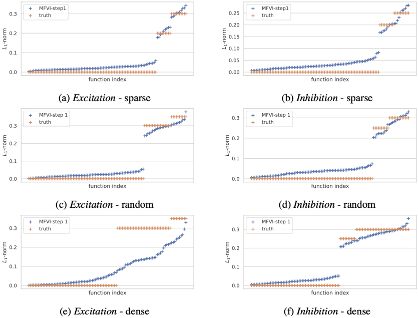
We test the performances of our procedure with respect to the graph sparsity ($K = 10$).

Scenario	Graph	# Edges	# Events	# Excursions
Self-excitation	Sparse	$2K - 1$	24638	431
	Random	$3K - 1$	27475	398
	Dense	$5K - 6$	90788	2
Self-inhibition	Sparse	$2K - 1$	22683	911
	Random	$3K - 1$	24031	884
	Dense	$5K - 6$	35291	547



(a) sparse (b) random (c) dense

Numerical performances - Graph sparsity



(a) *Excitation* - sparse (b) *Inhibition* - sparse
(c) *Excitation* - random (d) *Inhibition* - random
(e) *Excitation* - dense (f) *Inhibition* - dense

Figure: Estimated $\mathbb{L}_1$-norms of interaction functions plotted in increasing order

Scenario	Graph	Graph accuracy	Risk
Self-Exc.	Sparse	1.00	2.91
	Random	1.00	4.00
	Dense	0.5	17.67
Self-Inh.	Sparse	1.00	2.62
	Random	0.99	3.44
	Dense	1.00	2.67

Numerical performances - Mis-specification

We set $T = 300$ and $K = 10$ and construct synthetic mis-specified data by simulating a Hawkes process where the link function ψ is chosen as:

- ReLU: $\psi(x) = \max(x, 0)$;
- Logit: $\psi(x) = \log(1 + e^x)$;
- Mis-specified sigmoid, with unknown multiplicative parameter.

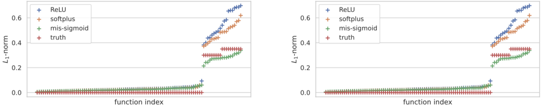


Figure: Estimated $\mathbb{L}_1$-norms of interaction functions plotted in increasing order in the Self-excitation and Self-inhibition scenarios

Scenario	Link	Graph acc.
Self-exc.	ReLU	1.00
	Softplus	1.00
	MS sigmoid	1.00
Self-inh.	ReLU	1.00
	Softplus	1.00
	MS sigmoid	0.99

The gaps allow to estimate well the connectivity graph but the other parameters cannot be well estimated. Nonetheless, the sign of the interaction functions is well recovered in all settings.

Numerical performances - Robustness with respect to A

We test the robustness of our variational method to mis-specification of the memory parameter A. We generate data from the sigmoid Hawkes process with $K = 10$ and with ground-truth parameter $A_0 = 0.1$, $T = 500$ and apply our variational method with $A \in \{0.05, 0.1, 0.2, 0.4\}$.

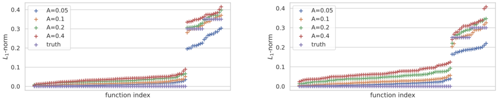
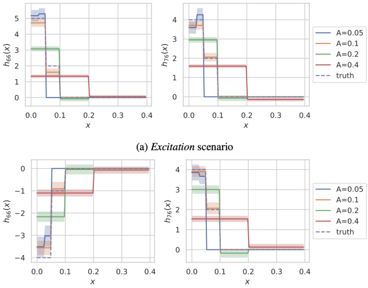


Figure: Estimated $\mathbb{L}_1$-norms of interaction functions plotted in increasing order in the Self-excitation and Self-inhibition scenarios



(a) *Excitation* scenario

(b) *Inhibition* scenario

The graph is well estimated with the gap heuristics.

Thank you for your attention.
Questions and remarks are welcomed!

References:

- SULEM D., RIVOIRARD V. AND ROUSSEAU J. (2023) *Bayesian estimation of non-linear Hawkes processes.* To appear in Bernoulli

- SULEM D., RIVOIRARD V. AND ROUSSEAU J. (2023) *Scalable and adaptive variational Bayes methods for Hawkes processes.* Submitted.