Bayesian Nonparametric inference for Nonlinear Hawkes processes

Déborah Sulem

UNIVERSITY OF OXFORD
DEPARTMENT OF STATISTICS

International Society for Bayesian Analysis World Meeting
Montreal, July, 1st 2022

Bayesian nonparametrics for nonlinear Hawkes

Joint work with

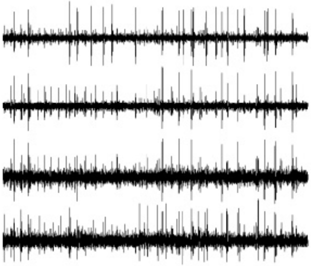
- Judith Rousseau
  University of Oxford

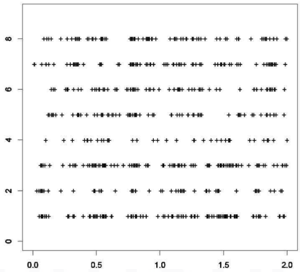
- Vincent Rivoirard
  Université Paris-Dauphine

Presentation Outline

Neuronal data modelling

- Spike = electric impulse emitted by a neuron
- "Event-wise" data = times of occurrences at each location (neuron)
- *Excitation* (clustering behaviour) and *inhibition* ("cancellation")





Which neurons are **functionally connected**?
What is the **type** and **strength** of their interactions?

Point and Intensity processes

Definition 1.1 (Temporal point process)

$N = (N_t^1, \dots N_t^K)_{t>0}$ is a K-dimensional TPP if N_t^k counts the number of points until t at component k.

Point and Intensity processes

Definition 1.1 (Temporal point process)

$N = (N_t^1, \dots N_t^K)_{t>0}$ is a K-dimensional TPP if N_t^k counts the number of points until t at component k.

Definition 1.2 (Intensity process)

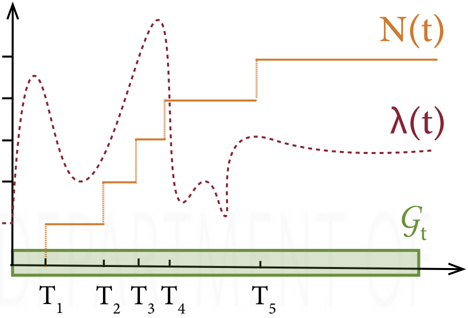
The intensity process $\lambda(t|\mathcal{G}_t) = (\lambda^1(t), \dots, \lambda^K(t))$ is the probability of observing a point at time t conditionally on the past of the process:

$$\lambda^k(t|\mathcal{G}_t)\mathrm{d}t = \mathbb{P}\left[N^k \text{ jumps in } [t, t+\mathrm{d}t] \,\middle|\, N_s, s < t \right], \quad 1 \leqslant k \leqslant K$$

Point and intensity processes

Examples:

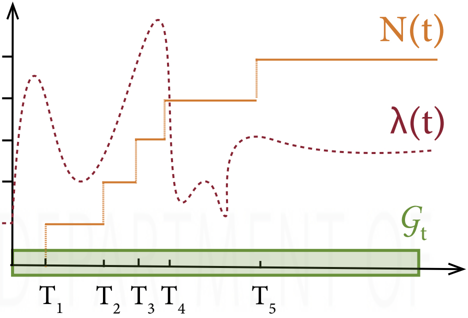
- Poisson point process: $\lambda(t)$ is deterministic (independent of $\mathcal{G}_t$)
- **Hawkes process:** $\lambda(t)$ is a stochastic process ($\mathcal{G}_t$-predictable)



Point and intensity processes

Examples:

- Poisson point process: $\lambda(t)$ is deterministic (independent of $\mathcal{G}_t$)
- **Hawkes process:** $\lambda(t)$ is a stochastic process ($\mathcal{G}_t$-predictable)

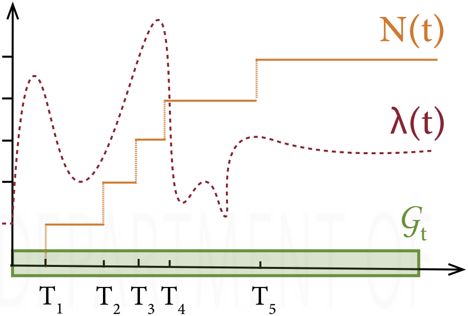


Applications: neuroscience Gerhard et al. (2017), genomics Gusto & Schbat (2005), finance Bacry & Muzy (2013), criminology Mohler (2013), epidemiology Browning et al. (2021), tweet popularity Zhao et al. 2015 ...

Point and intensity processes

Examples:

- Poisson point process: $\lambda(t)$ is deterministic (independent of $\mathcal{G}_t$)
- **Hawkes process:** $\lambda(t)$ is a stochastic process ($\mathcal{G}_t$-predictable)



Applications: neuroscience Gerhard et al. (2017), genomics Gusto & Schbat (2005), finance Bacry & Muzy (2013), criminology Mohler (2013), epidemiology Browning et al. (2021), tweet popularity Zhao et al. 2015 ...

Temporal dependencies and interactions: **excitation / contagion / diffusion, clustering / cascades, branching / causality**, **inhibition**, time decay, memory loss, nonlinear effects, state-switching, ...

Linear univariate Hawkes processes

Definition 2.1 (Hawkes, 1971)

Let $\nu > 0$ and $h : \mathbb{R}_+ \to \mathbb{R}_+$ with $\|h\|_1 < 1$. If

$$\lambda(t) = \nu + \int_{-\infty}^{t^-} h(t-u)\mathrm{d}N_u = \nu + \sum_{T_i \in N,\, T_i < t} h(t - T_i),$$

N is a linear univariate Hawkes process with **spontaneous rate** ν and **self-exciting function** h.

Linear univariate Hawkes processes

Definition 2.1 (Hawkes, 1971)

Let $\nu > 0$ and $h : \mathbb{R}_+ \to \mathbb{R}_+$ with $\|h\|_1 < 1$. If

$$\lambda(t) = \nu + \int_{-\infty}^{t^-} h(t-u)\mathrm{d}N_u = \nu + \sum_{T_i \in N,\, T_i < t} h(t - T_i),$$

N is a linear univariate Hawkes process with **spontaneous rate** ν and **self-exciting function** h.

Remarks:

- $Im(h) \subset \mathbb{R}_+ \implies$ **(self) excitation:** events cause events

Linear univariate Hawkes processes

Definition 2.1 (Hawkes, 1971)

Let $\nu > 0$ and $h : \mathbb{R}_+ \to \mathbb{R}_+$ with $\|h\|_1 < 1$. If

$$\lambda(t) = \nu + \int_{-\infty}^{t^-} h(t-u)\mathrm{d}N_u = \nu + \sum_{T_i \in N,\, T_i < t} h(t - T_i),$$

N is a linear univariate Hawkes process with **spontaneous rate** ν and **self-exciting function** h.

Remarks:

- $Im(h) \subset \mathbb{R}_+ \implies$ **(self) excitation:** events cause events
- **Branching (and causal) representation:** an event can
  - appear spontaneously at rate ν
  - be caused by a previous event at the rate $h(x)$

Multivariate nonlinear Hawkes process

Definition 2.2 (Generalized Hawkes process)

A K-dimensional continuous point process N is a Hawkes process if

- almost surely N^k and N^l never jump simultaneously
- N^k has intensity

$$\lambda^k(t) = \phi_k \left(\nu_k + \sum_{l=1}^{K} \sum_{T_i \in N^l, T_i < t} h_{lk}(t - T_i) \right) \tag{1}$$

Multivariate nonlinear Hawkes process

Definition 2.2 (Generalized Hawkes process)

A K-dimensional continuous point process N is a Hawkes process if

- almost surely N^k and N^l never jump simultaneously
- N^k has intensity

$$\lambda^k(t) = \phi_k \left(\nu_k + \sum_{l=1}^{K} \sum_{T_i \in N^l, T_i < t} h_{lk}(t - T_i) \right) \tag{1}$$

- $h_{lk} : \mathbb{R}^+ \to \mathbb{R}$: **interaction function $N^l \Rightarrow N^k$**
  - for x s.t. $h_{lk}(x) > 0$: **excitation**
  - for x s.t. $h_{lk}(x) < 0$: **inhibition**

Multivariate nonlinear Hawkes process

Definition 2.2 (Generalized Hawkes process)

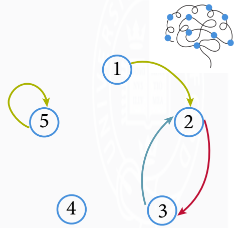
A K-dimensional continuous point process N is a Hawkes process if

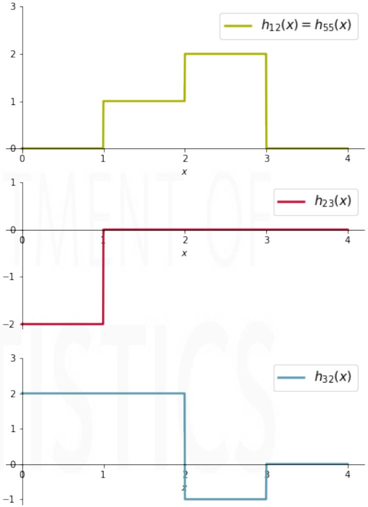
- almost surely N^k and N^l never jump simultaneously
- N^k has intensity

$$\lambda^k(t) = \phi_k \left(\nu_k + \sum_{l=1}^{K} \sum_{T_i \in N^l, T_i < t} h_{lk}(t - T_i) \right) \quad (1)$$

- $h_{lk} : \mathbb{R}^+ \to \mathbb{R}$: **interaction function $N^l \Rightarrow N^k$**
  - for x s.t. $h_{lk}(x) > 0$: **excitation**
  - for x s.t. $h_{lk}(x) < 0$: **inhibition**
- Link functions:
  - ReLU: $\phi_k(x) = (x)_+$
  - Sigmoid: $\phi_k(x) = (1 + e^{-x})^{-1}$
  - Softplus: $\phi_k(x) = \log(1 + e^x)$
  - ...
- Existence and uniqueness of a stationary distribution and

Example with 5 neurons





- **Interaction functions:**
  - Excitating $h_{12} = h_{55}$
  - Inhibiting h_{23}
  - Mixed h_{32}
- **Connectivity graph:**
  $\Delta = (\delta_{lk})_{l,k} \in \{0,1\}^{K \times K}$,
  $\delta_{lk} = 0 \iff h_{lk} = 0$

Estimation methods

Linear parametric model $h(t) = \alpha\beta e^{-\beta t} \implies f = (\nu, \alpha, \beta)$:

- EM: Veen & Schoenberg (2008)
- Metropolis-within-Gibbs: Rasmussen (2013)

Estimation methods

Linear parametric model $h(t) = \alpha\beta e^{-\beta t} \implies f = (\nu, \alpha, \beta)$:

- EM: Veen & Schoenberg (2008)
- Metropolis-within-Gibbs: Rasmussen (2013)

... and nonparametric model

- (Penalised) MLE with B-splines Gusto et al. (2010) or KDE Lewis et al. (2011)
- (Penalised) contrast with random histograms Hansen et al. (2015)

Estimation methods

Linear parametric model $h(t) = \alpha\beta e^{-\beta t} \implies f = (\nu, \alpha, \beta)$:

- EM: Veen & Schoenberg (2008)
- Metropolis-within-Gibbs: Rasmussen (2013)

... and nonparametric model

- (Penalised) MLE with B-splines Gusto et al. (2010) or KDE Lewis et al. (2011)
- (Penalised) contrast with random histograms Hansen et al. (2015)
- Reversible-jump MCMC using histogram priors Donnet et al. (2018)

Estimation methods

Linear parametric model $h(t) = \alpha\beta e^{-\beta t} \implies f = (\nu, \alpha, \beta)$:

- EM: Veen & Schoenberg (2008)
- Metropolis-within-Gibbs: Rasmussen (2013)

... and nonparametric model

- (Penalised) MLE with B-splines Gusto et al. (2010) or KDE Lewis et al. (2011)
- (Penalised) contrast with random histograms Hansen et al. (2015)
- Reversible-jump MCMC using histogram priors Donnet et al. (2018)

Nonlinear model

- Linear approximation over a dictionary Cai et al. (2021)
- MCMC for ReLU link with $h_{lk}(t) = K_{lk} e^{-\beta_{lk} t}$ Deutsch et al. (2022)
- Gibbs sampler in the sigmoidal Hawkes with GP prior Malem-Shinitski et al. (2022) or (fixed) basis decomposition Zhou et al. (2021)

Bayesian inference problem

- Assume that we observe a Hawkes process N on $[0, T]$ with link functions $(\phi_k^0)_k$ and parameter $f_0 = (\nu_0, h_0)$ with intensity

$$\lambda^k(t; f_0, \phi_0) = \phi_k^0 \left(\nu_k^0 + \sum_{l=1}^{K} \sum_{T_i \in N^l, T_i < t} h_{lk}^0(t - T_i) \right)$$

- Assume the model is **stationary** and **identifiable**

Bayesian inference problem

- Assume that we observe a Hawkes process N on $[0, T]$ with link functions $(\phi_k^0)_k$ and parameter $f_0 = (\nu_0, h_0)$ with intensity

$$\lambda^k(t; f_0, \phi_0) = \phi_k^0 \left(\nu_k^0 + \sum_{l=1}^{K} \sum_{T_i \in N^l, T_i < t} h_{lk}^0(t - T_i) \right)$$

- Assume the model is **stationary** and **identifiable**
- Given the likelihood function at $f = (\nu, h)$ and ϕ,

$$L_T(N; f, \phi) = \exp\left\{ \sum_{k=1}^{K} \left[\sum_{i=1}^{n_k} \log(\lambda^k(T_i^k; f, \phi)) - \int_0^T \lambda^k(t; f, \phi) dt \right] \right\}$$

Bayesian inference problem

- Assume that we observe a Hawkes process N on $[0, T]$ with link functions $(\phi_k^0)_k$ and parameter $f_0 = (\nu_0, h_0)$ with intensity

$$\lambda^k(t; f_0, \phi_0) = \phi_k^0 \left(\nu_k^0 + \sum_{l=1}^{K} \sum_{T_i \in N^l, T_i < t} h_{lk}^0(t - T_i) \right)$$

- Assume the model is **stationary** and **identifiable**
- Given the likelihood function at $f = (\nu, h)$ and ϕ,

$$L_T(N; f, \phi) = \exp\left\{ \sum_{k=1}^{K} \left[\sum_{i=1}^{n_k} \log(\lambda^k(T_i^k; f, \phi)) - \int_0^T \lambda^k(t; f, \phi) dt \right] \right\}$$

and a prior distribution Π on f (e.g., Gaussian processes, mixtures, splines,..), the posterior distribution is given by

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

Our results: model assumptions

- When is the model **stationary**?

Our results: model assumptions

- When is the model **stationary**?
  - If $\forall k, \|\phi_k\|_\infty < \infty$ **or** ϕ_k is L-Lipschitz and $S = (L \|h_{lk}\|_1)_{l,k}$ has a spectral radius $\rho(S) < 1$ Brémaud et al. (1996)

Our results: model assumptions

- When is the model **stationary**?
  - If $\forall k, \|\phi_k\|_\infty < \infty$ **or** ϕ_k is L-Lipschitz and $S = (L \|h_{lk}\|_1)_{l,k}$ has a spectral radius $\rho(S) < 1$ Brémaud et al. (1996)
  - If $h_{lk} = K_{lk} e^{-\beta_{lk}}$ and $\rho(K^+) < 1$ with $K^+ = (\max(K_{lk}, 0))_{l,k}$ Deutsch et al. (2022)

Our results: model assumptions

- When is the model **stationary**?
  - If $\forall k, \|\phi_k\|_\infty < \infty$ **or** ϕ_k is L-Lipschitz and $S = (L \|h_{lk}\|_1)_{l,k}$ has a spectral radius $\rho(S) < 1$ Brémaud et al. (1996)
  - If $h_{lk} = K_{lk} e^{-\beta_{lk}}$ and $\rho(K^+) < 1$ with $K^+ = (\max(K_{lk}, 0))_{l,k}$ Deutsch et al. (2022)
  - If $S^+ = (L \left\|h_{lk}^+\right\|_1)_{l,k}$ has a sup norm $\|S^+\|_\infty < 1$ (extension of Costa et al. (2019))

Our results: model assumptions

- When is the model **stationary**?
  - If $\forall k, \|\phi_k\|_\infty < \infty$ **or** ϕ_k is L-Lipschitz and $S = (L\|h_{lk}\|_1)_{l,k}$ has a spectral radius $\rho(S) < 1$ Brémaud et al. (1996)
  - If $h_{lk} = K_{lk}e^{-\beta_{lk}}$ and $\rho(K^+) < 1$ with $K^+ = (\max(K_{lk}, 0))_{l,k}$ Deutsch et al. (2022)
  - If $S^+ = (L\|h_{lk}^+\|_1)_{l,k}$ has a sup norm $\|S^+\|_\infty < 1$ (extension of Costa et al. (2019))
- When is the model **identifiable**?

Our results: model assumptions

- When is the model **stationary**?
  - If $\forall k, \|\phi_k\|_\infty < \infty$ **or** ϕ_k is L-Lipschitz and $S = (L \|h_{lk}\|_1)_{l,k}$ has a spectral radius $\rho(S) < 1$ Brémaud et al. (1996)
  - If $h_{lk} = K_{lk} e^{-\beta_{lk}}$ and $\rho(K^+) < 1$ with $K^+ = (\max(K_{lk}, 0))_{l,k}$ Deutsch et al. (2022)
  - If $S^+ = (L \left\|h^+_{lk}\right\|_1)_{l,k}$ has a sup norm $\|S^+\|_\infty < 1$ (extension of Costa et al. (2019))
- When is the model **identifiable**?

  If there exists $\varepsilon > 0$ such that $\forall k$, ϕ_k injective on

$$I_k = (\nu_k - \max_{1 \leqslant l \leqslant K} \left\|h^-_{lk}\right\|_\infty - \varepsilon, \nu_k + \max_{1 \leqslant l \leqslant K} \left\|h^+_{lk}\right\|_\infty + \varepsilon)$$

Our results: posterior asymptotic properties

- Does the posterior $\Pi(.|N)$ **concentrates** around the truth f_0 when $T \to \infty$? i.e., with $\epsilon_T = o(1)$ and $d = L_1$-distance,

$$\mathbb{E}_{f_0}\left[\Pi\left(d(f, f_0) < \epsilon_T \middle| N\right)\right] \xrightarrow[T\to\infty]{} 1.$$

Our results: posterior asymptotic properties

- Does the posterior $\Pi(.|N)$ **concentrates** around the truth f_0 when $T \to \infty$? i.e., with $\epsilon_T = o(1)$ and $d = L_1$-distance,

$$\mathbb{E}_{f_0}\left[\Pi\left(d(f, f_0) < \epsilon_T \middle| N\right)\right] \xrightarrow[T\to\infty]{} 1.$$

Yes if

- Model assumptions:
  - ϕ_k^{-1} on $J_k = \phi_k(I_k)$ Lipschitz
  - $\inf_x \phi_k(x) > 0$ **or** $\sqrt{\phi_k}$ and $\log \phi_k$ L-Lispchitz
- Prior assumptions: standard ones for regression or density estimation (Ghosal & van der Vaart 2007) (prior mass, sieve & entropy conditions)

Our results: renewal and choice of excursions

- **Finite-memory** process: $\forall l, k,\ supp(h_{lk}) \subset [0, A]$ with $A > 0$

Our results: renewal and choice of excursions

- **Finite-memory** process: $\forall l, k,\ supp(h_{lk}) \subset [0, A]$ with $A > 0$
- **Regeneration** times: $\tau_0 = 0, \tau_1, \tau_2, \dots$ defined as

$$\tau_j = \inf \left\{ t > \tau_{j-1};\ N|_{[t-A,t)} \neq \emptyset,\ N|_{(t-A,t]} = \emptyset \right\}, \quad j \geqslant 1$$

Our results: renewal and choice of excursions

- **Finite-memory** process: $\forall l, k,\ supp(h_{lk}) \subset [0, A]$ with $A > 0$
- **Regeneration** times: $\tau_0 = 0, \tau_1, \tau_2, \dots$ defined as

$$\tau_j = \inf\left\{t > \tau_{j-1};\ N|_{[t-A,t)} \neq \emptyset,\ N|_{(t-A,t]} = \emptyset\right\}, \quad j \geqslant 1$$

- Independent and identically distributed **excursions**

$$N|_{[\tau_j,\tau_{j+1})}, \quad 0 \leqslant j \leqslant J_T - 1$$

Our results: renewal and choice of excursions

- **Finite-memory** process: $\forall l, k,\ supp(h_{lk}) \subset [0, A]$ with $A > 0$
- **Regeneration** times: $\tau_0 = 0, \tau_1, \tau_2, \dots$ defined as

$$\tau_j = \inf\left\{t > \tau_{j-1};\ N|_{[t-A,t)} \neq \emptyset,\ N|_{(t-A,t]} = \emptyset\right\}, \quad j \geqslant 1$$

- Independent and identically distributed **excursions**

$$N|_{[\tau_j,\tau_{j+1})}, \quad 0 \leqslant j \leqslant J_T - 1$$

- Concentration inequalities for the number of excursions J_T, finite-moments of the number of points per excursions, ...

Our results: variational posterior asymptotics

- If T and/or K large, can we **approximate** the posterior?

Our results: variational posterior asymptotics

- If T and/or K large, can we **approximate** the posterior?
  Yes, using **Variational inference**: with $\mathcal{V}$ a mean-field variational family of distributions on $\mathcal{F}$, i.e.,

$$\mathcal{V} = \left\{ Q : dQ(f) = \prod_k dQ_{k1}(\nu_k) \prod_l dQ_{k2}(h_{lk}) \right\},$$

  where Q_{2k} includes the nonparametric prior family (Gaussian processes, basis decomposition, etc).

Our results: variational posterior asymptotics

- If T and/or K large, can we **approximate** the posterior? Yes, using **Variational inference**: with $\mathcal{V}$ a mean-field variational family of distributions on $\mathcal{F}$, i.e.,

$$\mathcal{V} = \left\{ Q : dQ(f) = \prod_k dQ_{k1}(\nu_k) \prod_l dQ_{k2}(h_{lk}) \right\},$$

where Q_{2k} includes the nonparametric prior family (Gaussian processes, basis decomposition, etc). The variational posterior is defined as

$$\hat{Q} = \arg\min_{Q \in \mathcal{V}} KL(Q || \Pi(.|N))$$

and under similar conditions $\mathbb{E}_0 \left[\hat{Q}(d(f, f_0) \lesssim \epsilon_T) \right] \xrightarrow[T\to\infty]{} 1.$

Our work in progress

- Can we **compute** the posterior and variational posterior distributions? (Work in progress)

Our work in progress

- Can we **compute** the posterior and variational posterior distributions? (Work in progress)
    - With data augmentation, Gibbs sampler and CAVI algorithm in the sigmoid model Zhou et al. (2021)
    - HMC in semi-parametric and low dimensional model ($K < 8$)
    - Stochastic Variational Inference?

Discussion

- Bayesian inference in the nonlinear Hawkes model can be done with standard nonparametric priors

Discussion

- Bayesian inference in the nonlinear Hawkes model can be done with standard nonparametric priors
- Computational bottleneck for point processes estimation

Frequentists $1 - 0$ Bayesians

  - Development of approximate and empirical Bayes methods
  - Parallelisation

Discussion

- Bayesian inference in the nonlinear Hawkes model can be done with standard nonparametric priors
- Computational bottleneck for point processes estimation

Parametric Frequentists $2 - 0$ Nonparametric Bayesians

  - Development of approximate and empirical Bayes methods
  - Parallelisation

Discussion

- Bayesian inference in the nonlinear Hawkes model can be done with standard nonparametric priors
- Computational bottleneck for point processes estimation

  Parametric Frequentists $2 - 0$ Nonparametric Bayesians

  - Development of approximate and empirical Bayes methods
  - Parallelisation
- Theory for
  - High-dimensional model $K \to \infty$
  - Time-varying models: time-dependent background rate X. Miscouridou (work in progress), change-points R. Browning (2021), Hidden Markov model Zhou et al. (2021)
  - Semi-parametric inference?

References

Fengshuo Zhang and Chao Gao. *Convergence rates of variational posterior distributions*, 2017.

Sophie Donnet, Vincent Rivoirard, and Judith Rousseau. *Nonparametric Bayesian estimation for multivariate Hawkes processes*, 2020.

Feng Zhou, Quyu Kong, Yixuan Zhang, Cheng Feng, and Jun Zhu. *Non-linear hawkes processes in time-varying system*, 2021.

Noa Malem-Shinitski, Cesar Ojeda, and Manfred Opper. *Nonlinear hawkes process with gaussian process and self effects*, 2021.

D.S., Vincent Rivoirard, and Judith Rousseau. *Bayesian estimation of nonlinear Hawkes process*, 2021.

Thank you for your attention!

Contact me at deborah.sulem@stats.ox.ac.uk