Bayesian estimation of nonlinear Hawkes processes

Judith Rousseau, joint work with D. Sulem (Oxford), V. Rivoirard (Paris Dauphine) & S. Donnet (INRA)

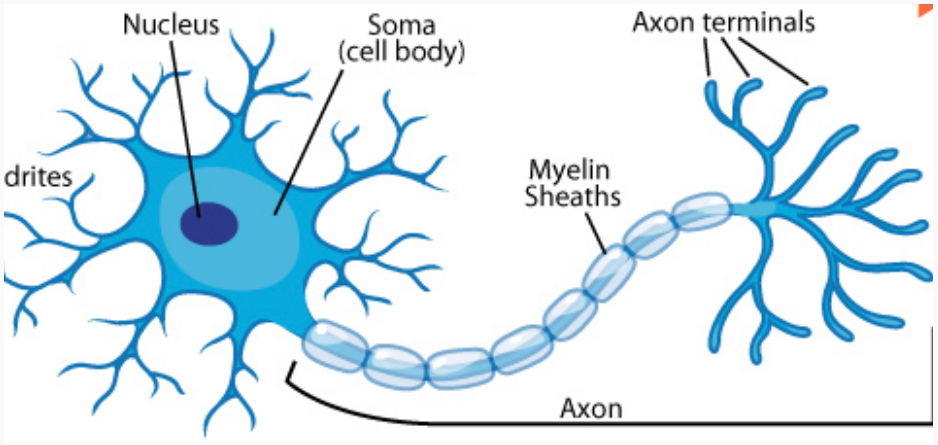
December 11, 2020

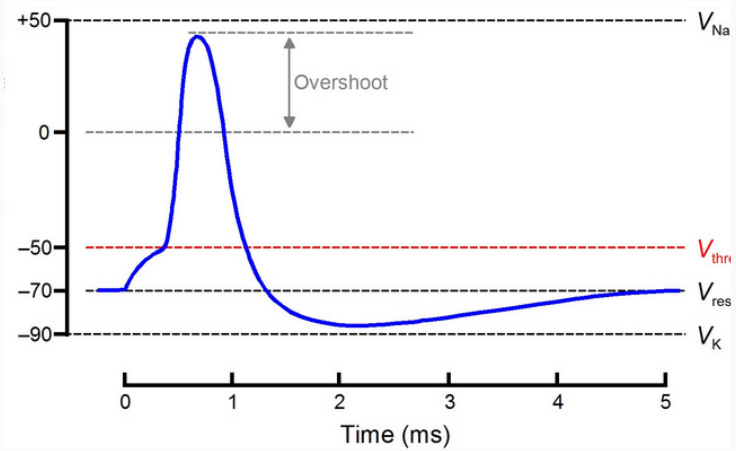
colloquium Berlin-postdam

Motivations and definition of Hawkes processes

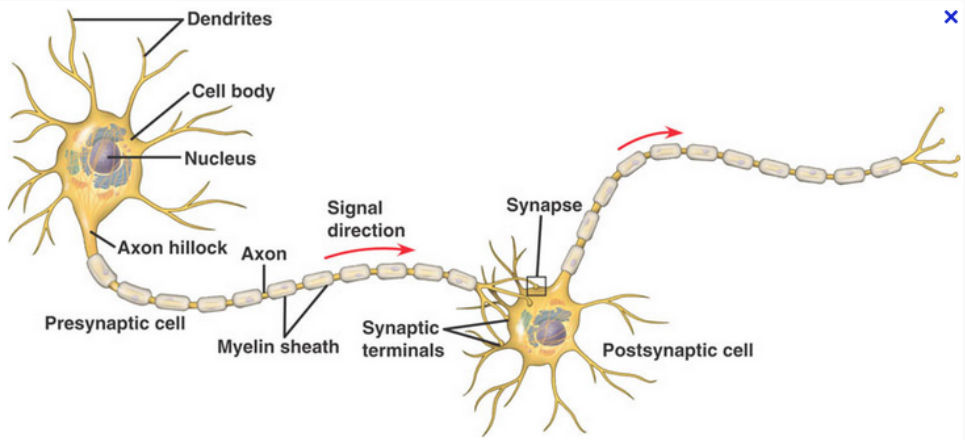
Neurobiological motivations

A neuron transmits information through electrical signals: **action potentials** (also called **spike trains**)



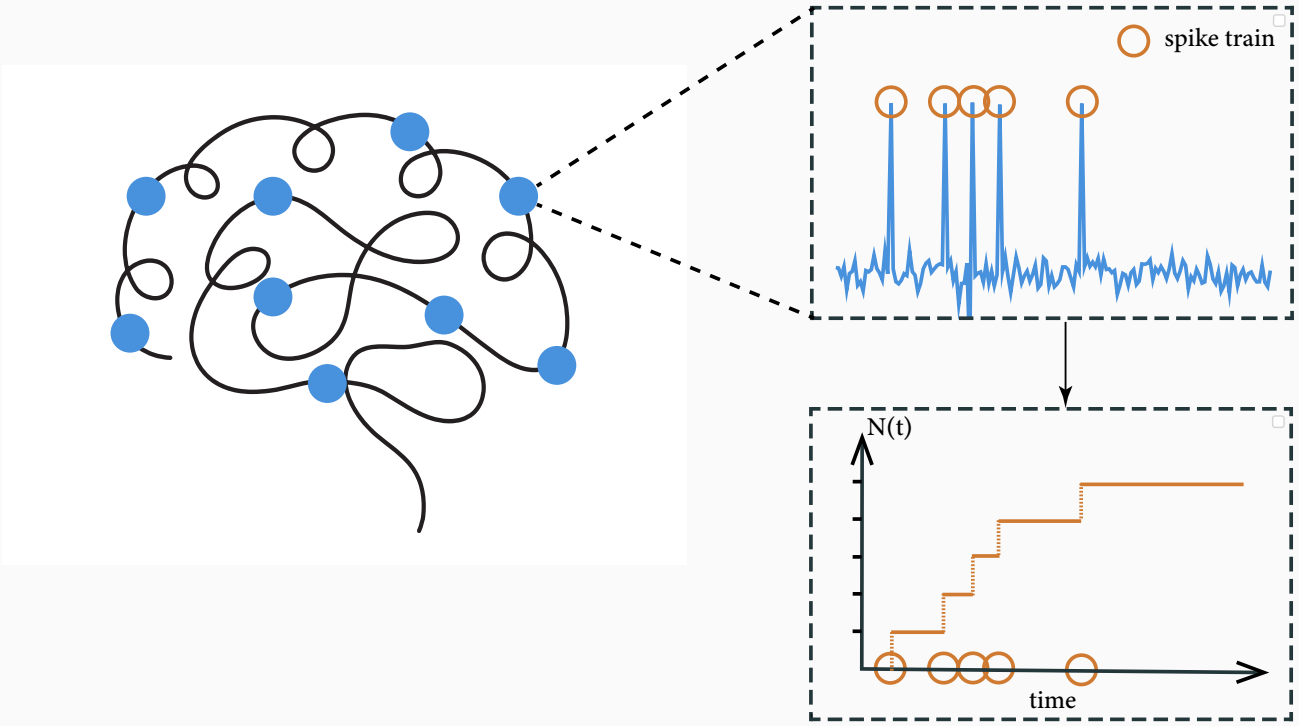


Action potentials can be recorded and the excitation times can be seen as a **point process**.



Neurobiological motivations

- Data = excitation times recorded at K neurons $\Rightarrow$ **multivariate point process**
- Goal: Infer the **graph** of connections and estimate the **properties** of the interactions (**excitation** and/or **inhibition**)
- Model: **nonlinear Hawkes processes**



Temporal point process

Definition (Point process)

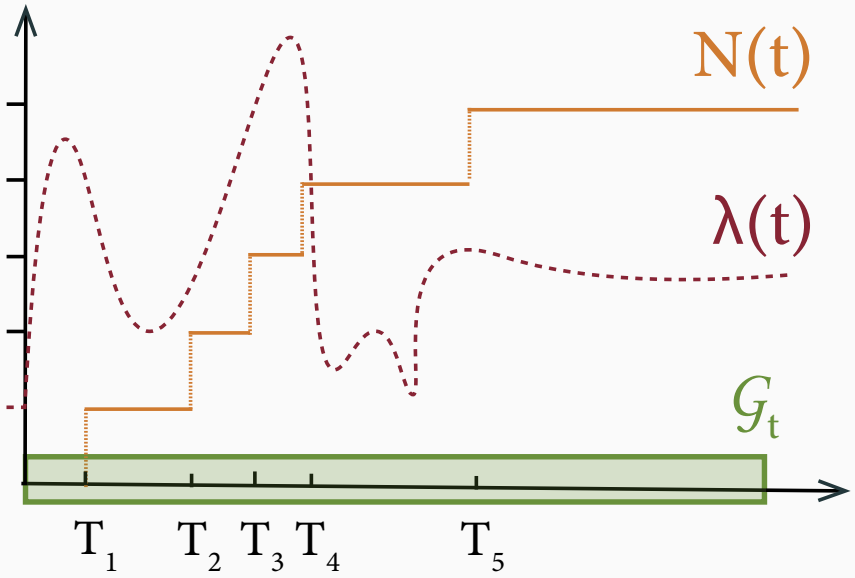
A **point process** N is a random countable set of points of $\mathbb{R}$ or equivalently a non-decreasing integer-valued process $(N_t)_t$.

Definition (Intensity of a point process)

The **intensity** λ_t of N represents the probability to observe a point at the time t conditionally on the past before t:

$\lambda_t \mathrm{d}t = \mathbb{P}[N \text{ has a jump} \in [t, t + \mathrm{d}t] \text{ conditionally on the past before } t)]$

Example: Poisson Process: λ_t is not random. **Homogeneous** if λ_t does not depend on t.



Linear univariate Hawkes processes

$\lambda_t \mathrm{d}t = \mathbb{P}\left[N \text{ has a jump} \in [t, t+\mathrm{d}t] \text{ conditionally on the past before } t\right]$

Definition

Linear univariate Hawkes process (Hawkes (1971)) Let $\nu > 0$ and $h \geqslant 0$ supported by $\mathbb{R}_+$ such that $\int_0^{+\infty} h(t)\mathrm{d}t < 1$. Then any point process N whose intensity is

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

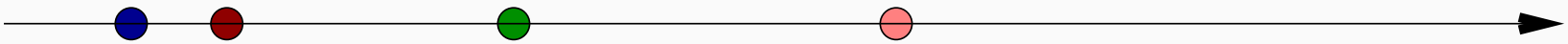
is called a linear univariate Hawkes process with **spontaneous rate** ν and **self-exciting function** h.

Remarks:

- $h \geqslant 0 \Rightarrow$ excitation
- Unique stationary distribution (Brémaud and Massoulié [**?**])
- Representation as a branching process with ancestor rate ν and reproducing function h

Linear univariate Hawkes processes: example

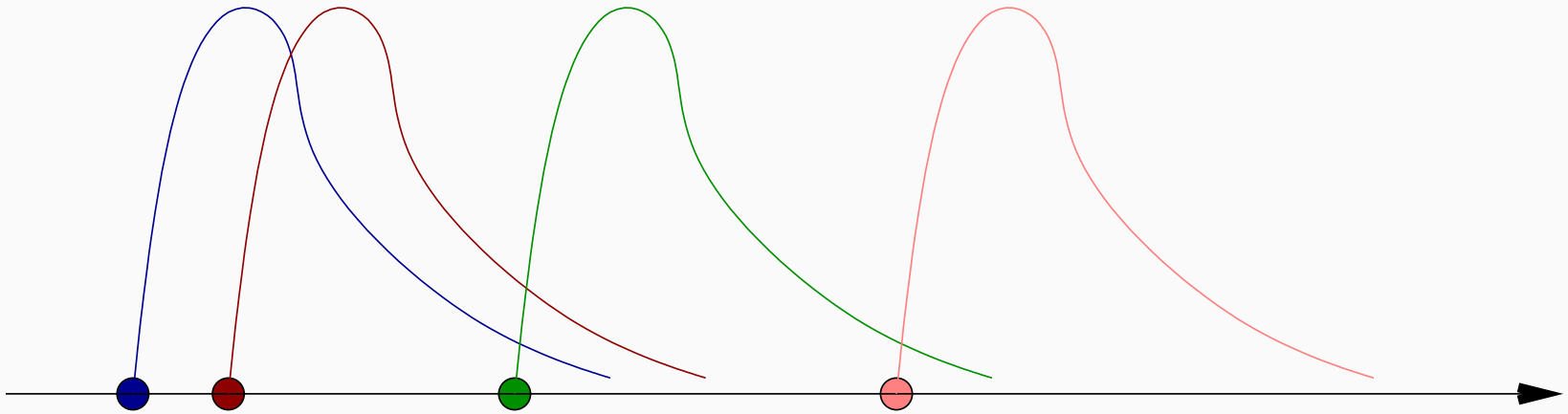
Branching process with ancestor rate $\nu > 0$ and reproduction $h \geqslant 0$ supported by $\mathbb{R}_+$.



- Ancestors: Realizations of a Poisson Process with $\lambda_t = \nu$

Linear univariate Hawkes processes: example

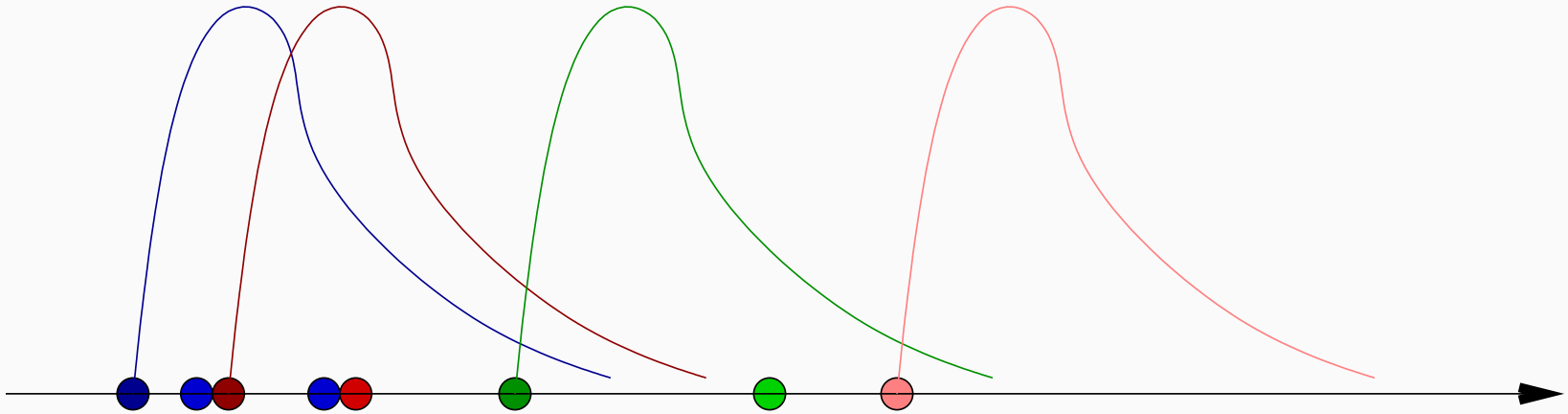
Branching process with ancestor rate $\nu > 0$ and reproduction $h \geqslant 0$ supported by $\mathbb{R}_+$.



- Ancestors: Realizations of a Poisson Process with $\lambda_t = \nu$
- Each ancestor can give birth to children according to a Poisson Process with $\lambda_t = h(t)$

Linear univariate Hawkes processes: example

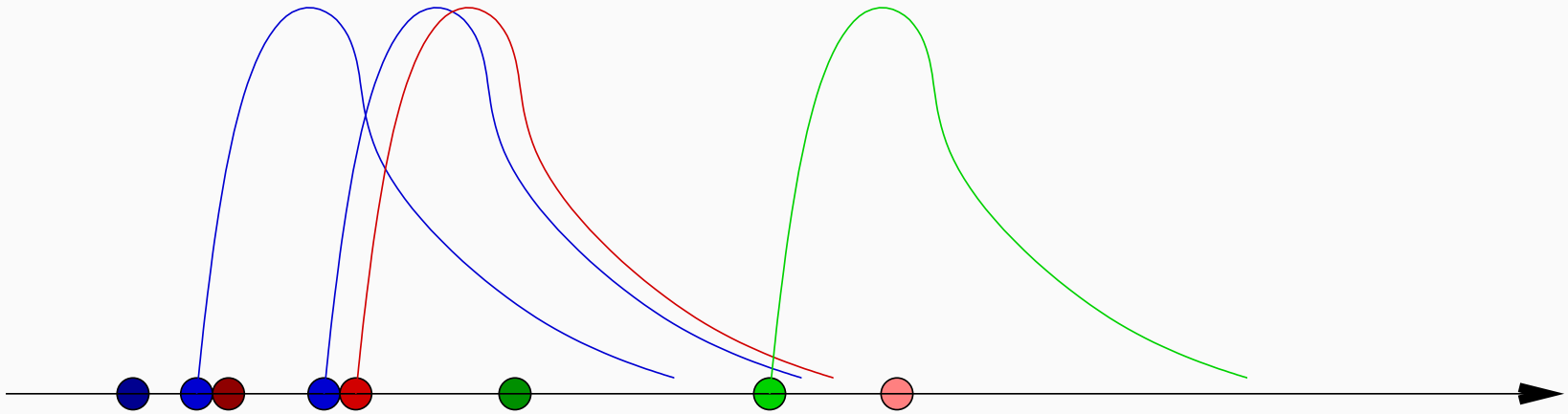
Branching process with ancestor rate $\nu > 0$ and reproduction $h \geqslant 0$ supported by $\mathbb{R}_+$.



- Ancestors: Realizations of a Poisson Process with $\lambda_t = \nu$
- Each ancestor can give birth to children according to a Poisson Process with $\lambda_t = h(t)$

Linear univariate Hawkes processes: example

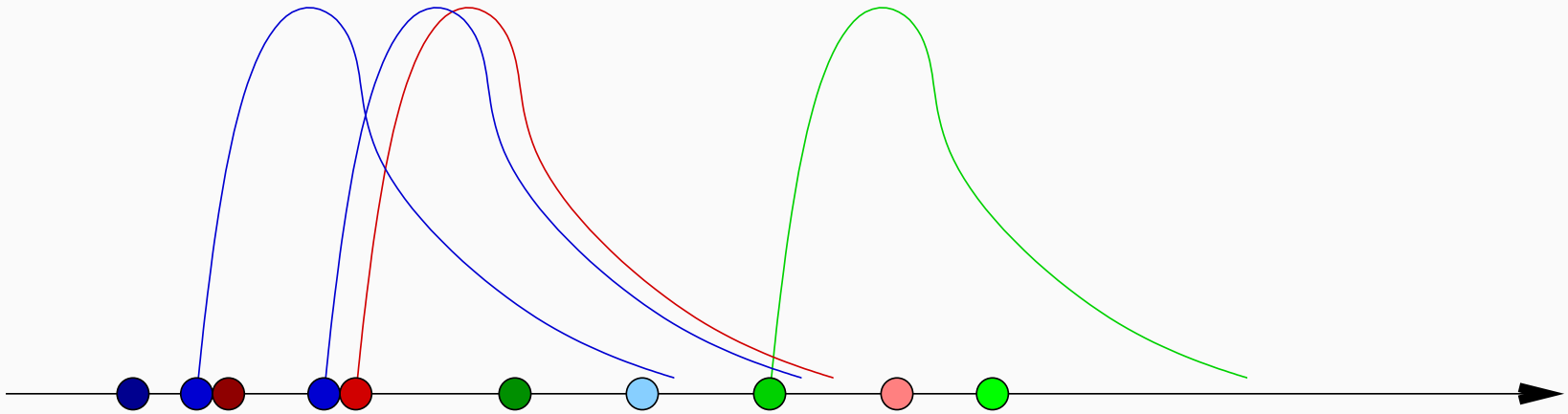
Branching process with ancestor rate $\nu > 0$ and reproduction $h \geqslant 0$ supported by $\mathbb{R}_+$.



- Ancestors: Realizations of a Poisson Process with $\lambda_t = \nu$
- Each ancestor can give birth to children according to a Poisson Process with $\lambda_t = h(t)$
- Each child can give birth to children according to a Poisson Process. with $\lambda_t = h(t)$

Linear univariate Hawkes processes: example

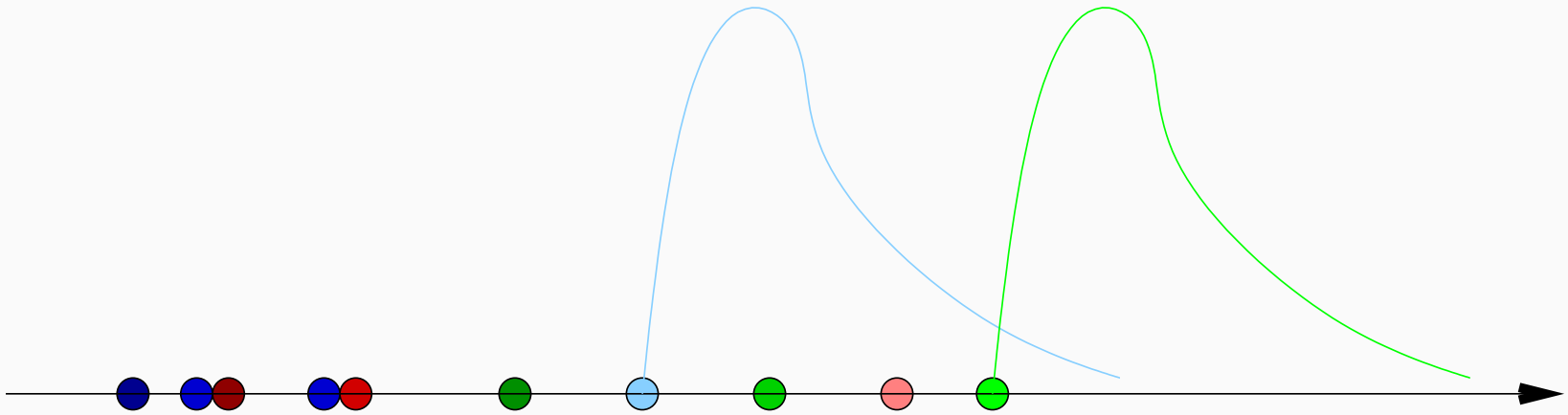
Branching process with ancestor rate $\nu > 0$ and reproduction $h \geqslant 0$ supported by $\mathbb{R}_+$.



- Ancestors: Realizations of a Poisson Process with $\lambda_t = \nu$
- Each ancestor can give birth to children according to a Poisson Process with $\lambda_t = h(t)$
- Each child can give birth to children according to a Poisson Process. with $\lambda_t = h(t)$

Linear univariate Hawkes processes: example

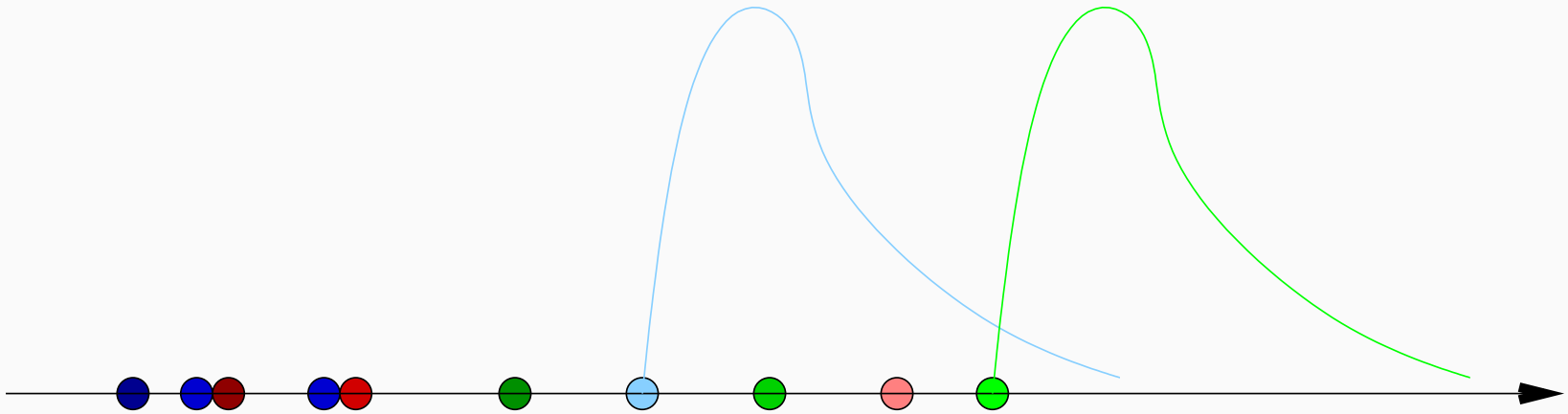
Branching process with ancestor rate $\nu > 0$ and reproduction $h \geqslant 0$ supported by $\mathbb{R}_+$.



- Ancestors: Realizations of a Poisson Process with $\lambda_t = \nu$
- Each ancestor can give birth to children according to a Poisson Process with $\lambda_t = h(t)$
- Each child can give birth to children according to a Poisson Process. with $\lambda_t = h(t)$

Linear univariate Hawkes processes: example

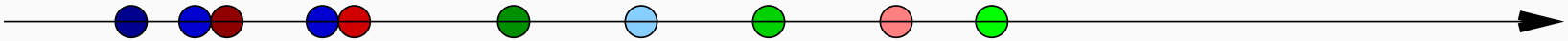
Branching process with ancestor rate $\nu > 0$ and reproduction $h \geqslant 0$ supported by $\mathbb{R}_+$.



- Ancestors: Realizations of a Poisson Process with $\lambda_t = \nu$
- Each ancestor can give birth to children according to a Poisson Process with $\lambda_t = h(t)$
- Each child can give birth to children according to a Poisson Process. with $\lambda_t = h(t)$
- Extinction if $\int_0^{+\infty} h(t)\mathrm{d}t < 1$

Linear univariate Hawkes processes: example

Branching process with ancestor rate $\nu > 0$ and reproduction $h \geqslant 0$ supported by $\mathbb{R}_+$.



- Ancestors: Realizations of a Poisson Process with $\lambda_t = \nu$
- Each ancestor can give birth to children according to a Poisson Process with $\lambda_t = h(t)$
- Each child can give birth to children according to a Poisson Process. with $\lambda_t = h(t)$
- Extinction if $\int_0^{+\infty} h(t)\mathrm{d}t < 1$
- Hawkes process = all the points where colors are not distinguished

Multivariate Hawkes process

- **Multivariate linear Hawkes:** $N = (N^1, \dots, N^K)$ where N^k models the activity of neuron $k \in [K]$ whose intensity is

$$\lambda_t^k = \nu_k + \sum_{l=1}^{K} \delta_{lk} \int_{-\infty}^{t-} h_{lk}(t-s) dN_s^l = \nu_k + \sum_{l=1}^{K} \sum_{T_i \in N^l, T_i < t} \delta_{lk} h_{lk}(t - T_i)$$

  - **Interaction** functions from l to k: $h_{lk} \geqslant 0$ **(excitation)** with support in $[0, A]$ (bounded memory)
  - **graph** of connections: $\delta_{lk} \in \{0, 1\}$

Multivariate Hawkes process

- **Multivariate linear Hawkes:** $N = (N^1, \dots, N^K)$ where N^k models the activity of neuron $k \in [K]$ whose intensity is

$$\lambda_t^k = \nu_k + \sum_{l=1}^{K} \delta_{lk} \int_{-\infty}^{t-} h_{lk}(t-s) dN_s^l = \nu_k + \sum_{l=1}^{K} \sum_{T_i \in N^l, T_i < t} \delta_{lk} h_{lk}(t - T_i)$$

  - **Interaction** functions from l to k: $h_{lk} \geqslant 0$ **(excitation)** with support in $[0, A]$ (bounded memory)
  - **graph** of connections: $\delta_{lk} \in \{0, 1\}$

- **Multivariate nonlinear** Hawkes:

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

  - (Nonlinear) **link** function $\phi^k \geqslant 0$
  - h_{lk} **signed** function $\Rightarrow$ **excitation and inhibition**

Multivariate Hawkes processes

Theorem (Brémaud and Massoulié (1996))

Existence and uniqueness of a stationary distribution for N if $\|\phi^k\|_\infty < \infty$ for any $k \in [K]$ or ϕ^k is 1-Lipschitz and the matrix Γ with entries $\Gamma_{lk} = \|h_{lk}\|_1$ has a spectral radius < 1.

Applications of Hawkes processes

Hawkes process are useful to model many situations where **excitation or inhibition phenomena** play a crucial role.

- **earthquakes**: Ozaki (1979), Ogata and Akaike (1982)
- **neuroscience:** Chornoboy, Schramm and Karr (1988)
- **genome analysis**: Gusto and Schbath (2005), Reynaud-Bouret and Schbath (2010)
- **financial data**: Embrechts, Liniger and Lin (2011), Bacry and Muzy (2013, 2014)
- **diffusion across social networks**: Crane and Sornette (2008)
- **analyze and predict the diffusion of COVID-19**: Mengersen, Paraha, Rivoirard, Rousseau and S. (2020)

Bayesian inference framework

Statistical goal

$$\lambda_t^k(f_0) = \phi^k\left(\nu_k^0 + \sum_{l=1}^{K} \sum_{T_i \in N^l, T_i < t} \delta_{lk} h_{lk}^0(t - T_i)\right), \quad k \in [K]$$

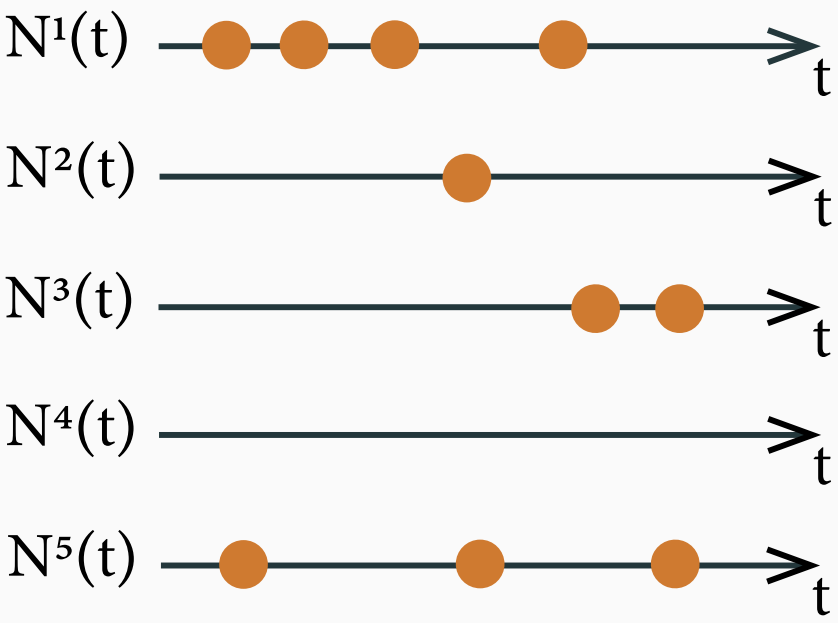
One observes a K-dimensional Hawkes P. $(N^k)_{k=1,\dots,K}$ on $[0, T]$ with **true parameter** $f_0 = ((\nu_k^0)_{k=1,\dots,K}, (h_{\ell k}^0)_{\ell,k=1,\dots,K})$.

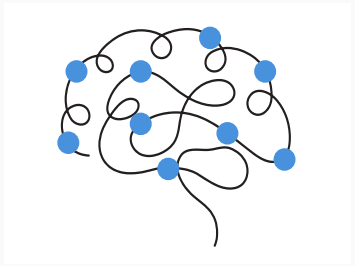
Statistical Goals:

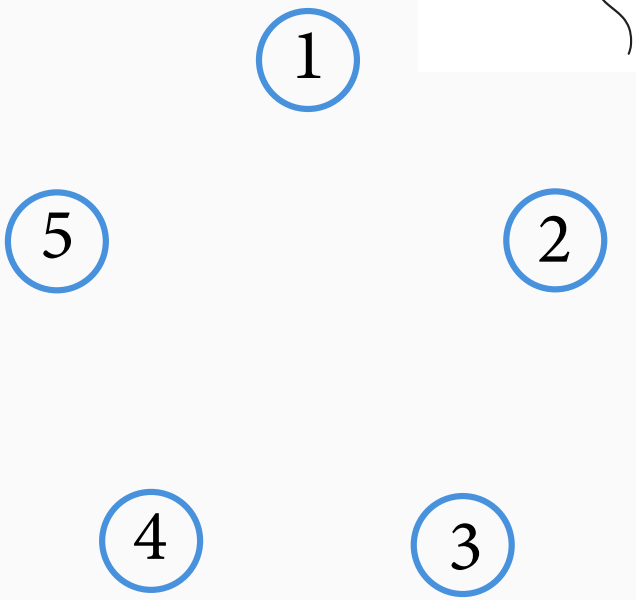
- Estimate the **process parameters** f_0
- Infer the **graph adjacency** $\delta_0 = (\delta_{\ell k}^0)_{\ell,k=1,\dots,K} \in \{0,1\}^{K\times K}$
- Infer some **link parameters of** ϕ^k**:** θ_0

Nonlinear Hawkes process: example with 5 neurons

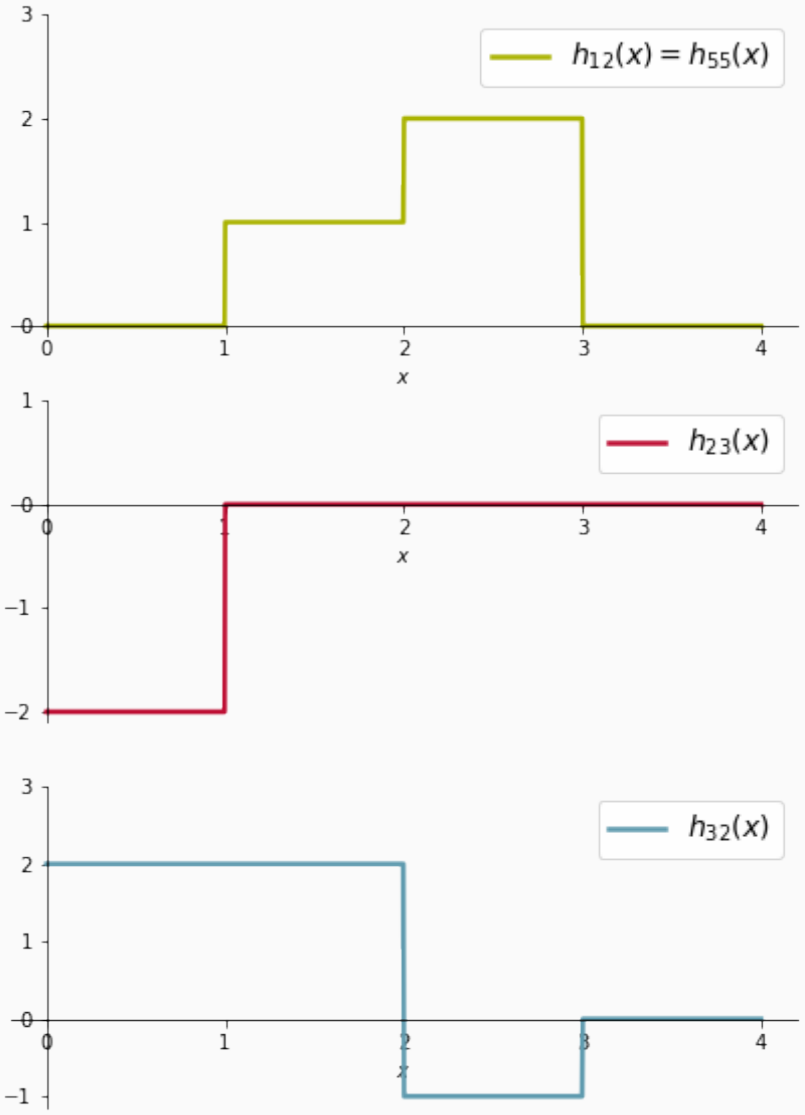
Observations: spike trains emitted by 5 neurons in a time window $[0, T]$ with $T = 5$ secs

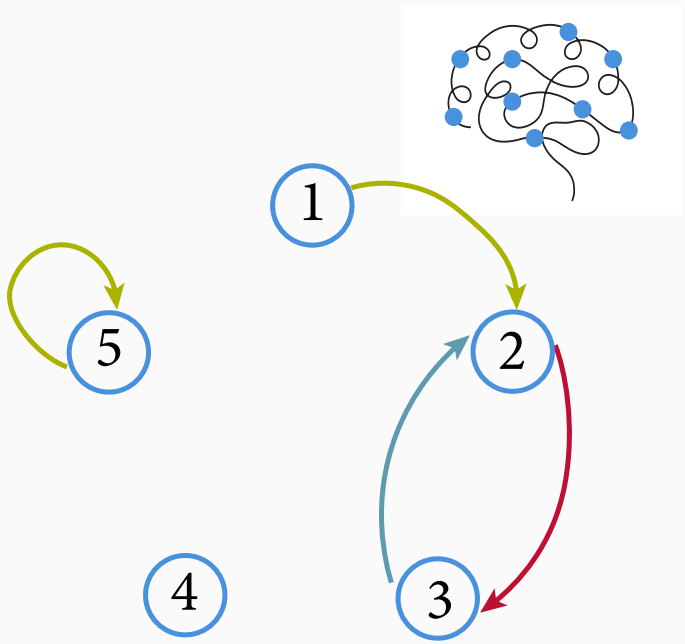






Nonlinear Hawkes process: example with 5 neurons





Graph of interactions:

$$\delta = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Bayesian approach

$$\lambda_t^k(f) = \nu_k + \sum_{l=1}^{k} \delta_{lk} \int^{t^-} h_{lk}(t-s) dN^l(s).$$

- **prior distribution** Π on $\eta = (f, \theta)$
- update via **likelihood**

$$\log \mathcal{L}_T(N; f, \theta) := \sum_{k=1}^{K} \left[\int_0^T \log(\lambda_t^k(f)) dN_t^k - \int_0^T \lambda_t^k(f) dt \right],$$

- Posterior distribution

$$\Pi(B|N) = \frac{\int_B \exp(L_T(N; f)) d\Pi(f, \theta)}{\int_{\mathcal{F}\times\Theta} \exp(L_T(N; f, \theta)) d\Pi(f, \theta)}.$$

- Posterior leads to **estimates, credible sets, tests, prediction**

$$e.g. \quad \hat{h}_{lk}(t) = E^{\pi}(h_{lk}(t)|N),$$

Posterior concentration and consistency

- **posterior concentration rate** $\epsilon_T = o(1)$ s.t. $\eta = (f, \theta)$

$$\mathbb{E}_{\eta_0}\left[\Pi\left(d(\eta_0, \eta) < \epsilon_T | N\right)\right] \xrightarrow[T\to\infty]{} 1, \quad d(\cdot,\cdot) = \text{distance},$$

- intensity

$$\lambda_t^k(f) = \phi_{\theta_k}\left(\nu_k + \sum_{l=1}^{K} \sum_{T_i \in N^l, T_i < t} \delta_{lk} h_{lk}(t - T_i)\right),$$

we consider the $\mathbb{L}_1$-distance

$$d(\eta_1, \eta_2) := \|f_1 - f_2\|_1 = \sum_{k=1}^{K} |\nu_k^1 - \nu_k^2| + \sum_{k=1}^{K}\sum_{\ell=1}^{K} \|h_{k\ell}^1 - h_{k\ell}^2\|_1$$

or

$$d_2(\eta_1, \eta_2) := \|f_1 - f_2\|_1 + \sum_k |\theta_{1k} - \theta_{2k}|$$

Posterior concentration and consistency

- **posterior concentration rate** $\epsilon_T = o(1)$ s.t. $\eta = (f, \theta)$

$$\mathbb{E}_{\eta_0}\left[\Pi\left(d(\eta_0, \eta) < \epsilon_T | N\right)\right] \xrightarrow[T\to\infty]{} 1, \quad d(\cdot,\cdot) = \text{distance},$$

- intensity

$$\lambda_t^k(f) = \phi_{\theta_k}\left(\nu_k + \sum_{l=1}^{K} \sum_{T_i \in N^l, T_i < t} \delta_{lk} h_{lk}(t - T_i)\right),$$

we consider the $\mathbb{L}_1$-distance

$$d(\eta_1, \eta_2) := \|f_1 - f_2\|_1 = \sum_{k=1}^{K} |\nu_k^1 - \nu_k^2| + \sum_{k=1}^{K}\sum_{\ell=1}^{K} \|h_{k\ell}^1 - h_{k\ell}^2\|_1$$

or

$$d_2(\eta_1, \eta_2) := \|f_1 - f_2\|_1 + \sum_k |\theta_{1k} - \theta_{2k}|$$

- **Posterior consistency** on the graph

$$\mathbb{E}_{f_0}\left[\Pi\left(\delta = \delta_0 | N\right)\right] \xrightarrow[T\to\infty]{} 1.$$

Our results

Contributions

- We consider 3 nonlinear models, 2 have additional parameters θ attached to the **non-linearity**
- We obtain Bayesian concentration rates for f and θ
- We obtain the posterior consistency on the graph δ

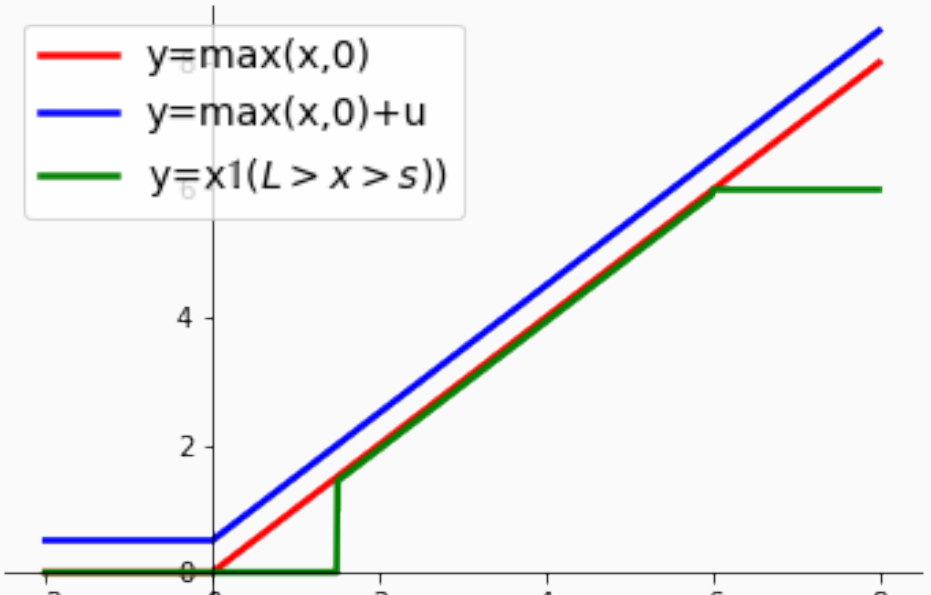
Nonlinear models

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

with 3 non-linear link functions

- **Model 0**: $\phi^k(x) = x$ and $h_{lk} \geqslant 0$ (linear)
- **Model 1:** $\phi^k(x) = \theta_k + \max(x, 0)$, $\theta_k > 0$
- **Model 2:** $\phi^k(x) = \mathbb{1}_{x > \theta_k} \min(x, \Lambda_k)$, $\theta_k > 0$, $\Lambda_k > 0$
- **Model 3:** $\phi^k(x) = \max(x, 0)$

Link parameters in models 1 and 2: $\theta = (\theta_k)_{k=1}^K$



Posterior concentration rates: general theorem

Ghosal and van der Vaart Theory :

- Kullback-Leibler condition :

$$\Pi(P : KL(P_0^T, P^T) \leqslant T\epsilon_T^2, V(P_0^T, P^T) \leqslant T\epsilon_T^2) \geqslant e^{-c_1 T\epsilon_T^2}$$

- Testing condition : $\exists \phi(N) \in [0,1]$ and $\mathcal{F}_T$ s.t. $\Pi(\mathcal{F}_T^c) \leqslant e^{-(c_1+2)T\epsilon_T^2}$

$$E_0(\phi) = o(1), \qquad \sup_{d(\eta,\eta_0)>M\epsilon_T,\mathcal{F}_T} E_\eta(1-\phi) \leqslant e^{-(c_1+2)T\epsilon_T^2},$$

Then

$$E_0\Pi(d(\eta,\eta_0) > M\epsilon_T|N) = o(1)$$

Aim 1 express these [implicit] conditions as *simple* (at least common) conditions

Aim 2 Only useful for *statistical distances* – not enough for estimation of the graph

Posterior concentration rates for Hawkes : case θ known

$$\mathcal{H} = \{h_{lk} : [0, A] \to \mathbb{R},\ \|h_{lk}\|_\infty < \infty,\ r(S) < 1\} \quad S = (\|h_{lk}\|_1)_{lk}$$

- KL bis :

$$\Pi(\max_{l,k} \|h_{lk} - h^0_{lk}\|_\infty \leqslant \epsilon_T, |\nu_k - \nu^0_k| \leqslant \epsilon_T) \gtrsim e^{-c_1 T \epsilon_T^2}$$

- Testing : $\exists \mathcal{H}_T \subset \mathcal{H}$

$$\mathcal{N}(\epsilon_T, \mathcal{H}_T, \|\cdot\|_1) \leqslant x_0 T \epsilon_T^2$$

Theorem (Posterior concentration rate on f and θ)

*Under **KL bis** and **Testing** and $\|h^-_{lk}\|_\infty < \nu_k$ In Models 1 and 2*

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

and in Model 3 : under addditional assumption

$$\frac{1}{T}\mathbb{E}_0\left(\int_0^T \frac{\mathbb{1}_{\lambda_t(f_0)>0}}{\lambda_t(f_0)} dt\right) < +\infty$$

Comments

- KLbis and Testing : very standard in literature - - similar to conditions for density estimation, regression etc . One can us directly results already proved on families of priors
- Model 3 : Case linear is a sub case of Model 3

$$\frac{1}{T}\mathbb{E}_0\left(\int_0^T \frac{\mathbb{1}_{\lambda_t(f_0)>0}}{\lambda_t(f_0)}dt\right) < +\infty$$

If $h_{lk}^- \neq 0$: nasty conditions. If this is not satisfied

$$\Pi(d(f,f_0) \leqslant \sqrt{\epsilon_T}|N) = 1 + o_p(1)$$

- Important tool : renewal times
  $\tau_{j+1} = \inf\{t > \tau_j, N[t-A,t) > 0\, N(t-A,t] = 0\}$ $(N|_{[\tau_j,\tau_{j+1})})_j$ are iid.

Posterior concentration rates for Hawkes : case θ unknown

$$M_1 : \phi_\theta(x) = \theta + x_+, \quad M_2 : \min(x, \Lambda)\mathbf{1}_{x>\theta}, \quad M_3 : x_+$$

Theorem

If in addition : $\forall k, \exists l$ *s.t.* $h_{lk}^- \geqslant c_* > 0$ *on* (x_1, x_2) *then in* M_1

$$\mathbb{E}_{f_0}\left[\Pi(\|\theta - \theta_0\|_1 < \epsilon_T | N)\right] \xrightarrow[T\to\infty]{} 1,$$

and in Model 2, true if h_{lk}^- *Lipschitz on some interval near 0*

$$\mathbb{E}_{f_0}\left[\Pi(\|\theta - \theta_0\|_1 < \sqrt{\epsilon_T} | N)\right] \xrightarrow[T\to\infty]{} 1.$$

comments

- $\exists l$ s.t. $h_{lk}^{-} > 0$: necessary for identifiability. hence $h_{lk}^{-} \geqslant c_* > 0$ on (x_1, x_2) is a weak condition
- in model 2 : stronger condition and slower rate : harder problem . lower bound on the rate is an open problem.
- Estimation of θ a bit harder : it came as a suprise to be able to estimate θ.

Posterior consistency of the graph

$$h_{lk} = \delta_{lk} h_{lk}, \quad \delta_{lk} \in \{0,1\},\ h_{lk} = 0 \text{ if } \delta_{lk} = 0$$

- Prior

$$\delta = (\delta_{lk})_{lk}, \quad (h_{lk}, (l,k) \in \mathcal{I}(\delta) | \delta) \sim \pi_{h||delta}, \quad \mathcal{I}(\delta) = \{(l,k) : \delta_{lk} = 1\}$$

Theorem

Under the same conditions as case θ known

- *If $\delta^0_{lk} = 1$ and $\|h^0_{lk}\|_1 \geqslant M' \epsilon_T$*

$$\mathbb{E}_0 \left[\Pi(\delta_{lk} = \delta^0_{lk} | N) \right] \xrightarrow[T \to \infty]{} 1$$

- *If $\delta^0_{lk} = 0$ and if*

$$\forall (l,k) \in \mathcal{I}(\delta),\ \Pi(\|h_{lk}\|_1 \leqslant \epsilon_T | \delta) \leqslant e^{-C_1 T \epsilon_T^2}$$

then

$$\mathbb{E}_0 \left[\Pi(\delta_{lk} = \delta^0_{lk} | N) \right] \xrightarrow[T \to \infty]{} 1$$

comments

- Easier to detect signal than to see that $h^0_{lk} = 0$
- Extra condition in the null case :

$$\forall (l,k) \in \mathcal{I}(\delta),\ \Pi(\|h_{lk}\|_1 \leqslant \epsilon_T | \delta) \leqslant e^{-C_1 T \epsilon_T^2}$$

  - Simple to verify in the following hierarchical prior

$$h_{lk} = \delta_{lk} S_{lk} \bar{h}_{lk}, \quad \|\bar{h}_{lk}\|_1 = 1, \quad d\Pi(h) = \Pi(\delta)\Pi(S|\delta)d\Pi(\bar{h}|\delta)$$

  - Unpleasant : strong penalisation around 0
$Pi(S_{lk} \leqslant \epsilon_T | \delta_{lk} = 1) \leqslant e^{-C_1 T \epsilon_T^2}$ e.g. Gamma-type prior on some power $p \in \mathbb{N}$ of the norms

$$\Pi(\rho^p) \propto \rho^{-p\alpha - p} \exp(-\beta/\rho^p) \mathbb{1}_{[0,1]}(\rho),$$

  - impact on estimation and detection of small signal

Example of priors : Concentration rate: case of Holder functions

One can choose Histograms, spline, wavelets etc . . . simple conditions $h_{lk} = \delta_{lk} S_{lk} \bar{h}_{lk}$

- Histogram prior

$$\delta_{lk} \overset{iid}{\sim} \mathcal{B}(p), \quad S_{lk}|\delta_{lk} = 1 \sim \Pi_S, \quad \bar{h}_{lk} = \sum_j e_j w_j \mathbf{1}_{I_j}, \quad e_j = -1, 1, \quad (w_1, \cdots, w_J)$$

$$I_j = (t_j, t_{j+1})$$

Corollary

In Models 1 and 2, under the ***random histogram prior*** *satisfying the assumptions, if* $\forall k, l \in [K], h_{kl}^0 \in \mathcal{H}(\beta, L)$ *with* $\beta \in (0, 1]$, *then*

$$\mathbb{E}_0 \left[\Pi \left(d(f_0, f) < \left(\frac{T}{\log T} \right)^{-\beta/(2\beta+1)} | N \right) \right] \xrightarrow[T \to \infty]{} 1.$$

Conclusion

- Theory for $\mathbb{L}_1$-posterior concentration rates under weak assumptions in different non linear models:
  - $\phi^k(x) = \max(x, 0)$: not completely solved...
  - $\phi^k(x) = \theta_k + \max(x, 0)$
  - $\phi^k(x) = \mathbb{1}_{x>\theta_k} \min(x, \Lambda_k)$: estimation of θ harder
- Inference of the graph parameter with realistic priors
- Possible extensions
  - sparse high-dimensional case
  - processes with unbounded memory
  - concentration rates in supremum norms

The end

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