Bayesian estimation in High dimensional Hawkes processes

Judith Rousseau

Université Paris-Dauphine and University of Oxford

Co-authors

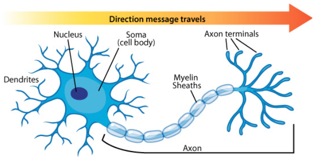
- DÉBORAH SULEM
  Lugano University

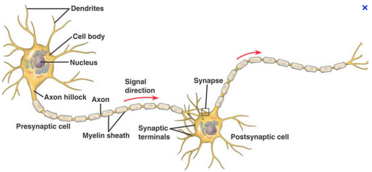
- VINCENT RIVOIRARD
  Université Paris Dauphine

- MAEL DUVERGER
  Université Paris Dauphine

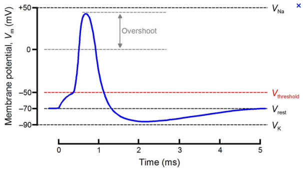
Functional connectivity graph of neurons

A neuron is an electrically excitable cell that processes and transmits information through electrical signals





If upstream signal is strong enough, this cell produces an action potential (also called spike), which is a spiky (electric) signal. Then, this signal is propagated to downstream neurons.

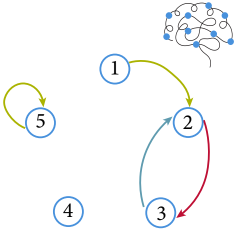


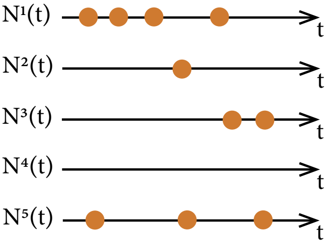
© PhysiologyWeb at www.physiologyweb.com

Action potentials can be recorded and the excitations times can be seen as a point process, each point corresponding to the peak of one action potential of this neuron.

Functional connectivity graph of neurons

Observations: spike trains on 5 neurons on a time window $[0, T]$





Graph of interactions:

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

<u>Goal:</u> Using activity recordings of K neurons, we wish to infer the graph between them. For this purpose, we use probabilistic models based on Hawkes processes.

Point process on $\mathbb{R}$

- N is a point process $\Leftrightarrow$ N is a random set of points on $\mathbb{R}$.

$$N(A) = number\ of\ points\ in\ A$$

- Intensity function

$$\lambda(t)dt = P(point\ in\ [t, t+dt)|N_s, s < t)$$

- example PP : $\lambda(t)$ is deterministic.
- Univariate linear Hawkes process

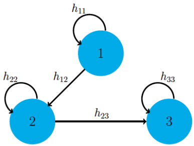
$$\lambda(t) = \nu + \int^{t^-} h(t-s)dN_s = \nu + \sum_{t_i<t} h(t-t_i), \quad h \geq 0$$

h : excitation function, ν background rate

Multidimensional Hawkes: linear and nonlinear

- K neurones interacting: K Point Proc **non independent.** $N = (N^{(1)}, \cdots, N^{(K)})$

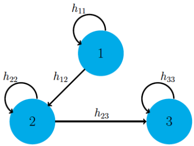
$$\lambda_t^{(j)} = \Phi\left(\nu_j + \sum_{\ell=1}^{K} \int_{-\infty}^{t-} h_{\ell j}(t-u)\mathrm{d}N^{(\ell)}(u)\right)$$
$$= \Phi\left(\nu_j + \sum_{\ell=1}^{K} \sum_{T_\ell \in N^{(\ell)},\, T_\ell < t} h_{\ell j}(t - T_\ell)\right)$$



Multidimensional Hawkes: linear and nonlinear

- K neurones interacting: K Point Proc **non independent.** $N = (N^{(1)}, \cdots, N^{(K)})$

$$\lambda_t^{(j)} = \Phi\left(\nu_j + \sum_{\ell=1}^{K} \int_{-\infty}^{t-} h_{\ell j}(t-u)\mathrm{d}N^{(\ell)}(u)\right)$$
$$= \Phi\left(\nu_j + \sum_{\ell=1}^{K} \sum_{T_\ell \in N^{(\ell)},\, T_\ell < t} h_{\ell j}(t - T_\ell)\right)$$

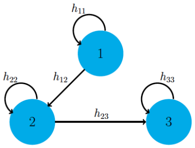


- Non linear Φ to allow for inhibition effects: $h_{\ell j}(x) \leq 0$ for some x.
- Linear Hawkes $\Phi(x) = x$

Multidimensional Hawkes: linear and nonlinear

- K neurones interacting: K Point Proc **non independent.** $N = (N^{(1)}, \cdots, N^{(K)})$

$$\lambda_t^{(j)} = \Phi\left(\nu_j + \sum_{\ell=1}^{K} \int_{-\infty}^{t-} h_{\ell j}(t-u)\mathrm{d}N^{(\ell)}(u)\right)$$
$$= \Phi\left(\nu_j + \sum_{\ell=1}^{K} \sum_{T_\ell \in N^{(\ell)},\, T_\ell < t} h_{\ell j}(t - T_\ell)\right)$$



- Non linear Φ to allow for inhibition effects: $h_{\ell j}(x) \le 0$ for some x.
- Linear Hawkes $\Phi(x) = x$

N is stationary iff spectral radius $sp(\rho) < 1$

$$\text{where} \quad \rho = (\rho_{\ell k})_{\ell,k \le K}, \quad \rho_{\ell k} = \int_0^A h_{\ell k}(x)dx$$

Here we study only the linear case but
(i) high dimensional: $K >> 1$ and/or (ii) BvM results

Multivariate Hawkes processes- Statistical Goal

Estimation of $f = (\nu_j, (h_{\ell j})_{\ell \in [\![1;K]\!]})_{k \in [\![1;K]\!]}$ based on observations of $N = (N^{(j)})_{j \in [\![1;K]\!]}$ on $[0, T]$ with intensity process $(\lambda^{(j)})_{j \in [\![1;K]\!]}$.

- Non parametric aspect :

$$\Pi(\|f - f_0\|_1 \leq \epsilon_T | N) \to 1$$

- Semi parametric

$$\Pi(\sqrt{T}(\Psi(f) - \Psi(f_0) \leq z | N) \to \Phi_{V_0}(z)$$

with $\Psi(f) = \nu_\ell$ or $\Psi(f) = \rho_{\ell k}$ etc .

State of the art for theoretical results [Nonparametric estimation]

$$\lambda_t^{(k)} = \psi_j\Big(\nu_j + \sum_{\ell=1}^{K} \int_{-\infty}^{t-} h_{\ell j}(t-u)\mathrm{d}N^{(\ell)}(u)\Big).$$

- Linear case in the nonparametric setting with wixed K: $\psi_k(x) = x$
  - Lasso-type estimation: Hansen, Reynaud-Bouret and R. (2015)
  - Bayesian estimation: Donnet, Rivoirard and Rousseau (2020)
- Nonlinear case and fixed K:
  - Sulem, Rivoirard, Rousseau [1]& [2] : Bayes & Variational Bayes
  - Parametric approaches: Bonnet, Martinez Herrera and Sangnier (2021a,b) : MLE for exponential kernel functions. Lemonnier and Vayatis (2014): Linear approximation with exponential kernels. Deutsch and Ross (2022) : Bayesian modelling
- Large K Chen, Witten and Shojaie (2017) in non linear (bounded ϕ) estimation of covariances and Bacry, Bompaire, Gaïffas and Muzy (2020) estimation of λ using covariances

Bayesian inference framework

- We assume that we observe over $[-A, T]$ a stationary Hawkes process $N = (N^{(1)}, \cdots, N^{(K)})$.
- log-likelihood at $f = (\nu, h) \in \mathcal{F} \subset \mathbb{R}^K \times \mathcal{H}^{K^2}$:

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

- Π: prior on $(\nu, h) \in \mathcal{F}$. Then posterior distribution:

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

Inference for high dimensional Hawkes models: $K >> 1$

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

$$\lambda_t^{(j)} = \nu_j + \sum_{\ell=1}^{K} \int_{-\infty}^{t-} h_{\ell j}(t-u)\mathrm{d}N^{(\ell)}(u), \quad A_0(j) = \{\ell; h_{\ell j}^0 \neq 0\}$$

- Assumptions:
  - $\nu_j \in \mathbb{R}$, $j \leq K$ and $\mathrm{supp}(h_{\ell j}) \subset [0, A]$, $A > 0$ fixed.
  - $K >> 1$ and $A_0(j)$ unknown but $\max_j |A_0(j)| \leq a_0$ unknown
  - stationary assumption

- Statistical goals: Bayesian estimation of

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

by using observations of $N = (N^{(j)})_{j\in[\![1;K]\!]}$ on $[-A, T]$ with $T \to +\infty$

Example of priors $f = (\nu_j, \mathbf{h} = (h_{\ell j})_{\ell \in [\![1;K]\!]})_{j \in [\![1;K]\!]}$: selection prior

$$d\Pi(f) = \prod_{j=1}^{K} d\Pi_h(\mathbf{h}_{.j}) \prod_j d\Pi_\nu(\nu_j), \quad \mathbf{h}_{.j} = (h_{\ell j}, \ell \leq K)$$

with

1. $\nu_j \overset{iid}{\sim} \pi_\nu$ e.g. π_ν is a Gamma distribution.
2. Inducing sparsity on $\mathbf{h}$: Selection priors ; $\forall j \leq K \;\; A(j) = \{\ell; h_{\ell j} \neq 0\}$

$$|A(j)| = a_j \sim \pi_a, \quad [A(j)|a_j] \sim \mathcal{U}; \quad \delta_{\ell j} = 1 \quad \Leftrightarrow \ell \in A(j)$$

$$\forall \ell \in A(j) \quad h_{\ell j} \overset{iid}{\sim} \Pi_h; \quad \delta = (\delta_{\ell j})_{\ell j} = \text{connectivity graph.}$$

Examples of priors for Π_h or $\Pi_{\bar{h}}$

- Random histogram if $\ell \in A(j)$: $I \sim \mathcal{P}(a),; (w_i^{\ell,j}, i \leq I) \sim G$

$$h_{\ell j}(x) = \sum_{i=1}^{I} 1_{]t_i, t_{i+1}]}(x) w_i^{\ell,j}, \quad \sum_i w_i^{\ell,j}(t_{i+1} - t_i) < 1$$

- mixture of Betas if $\ell \in A(j)$:

$$h_{\ell j}(x) = \frac{1}{A} \int_0^1 g_{\alpha,u}(x/A) dQ_{\ell,j}(u), \quad (Q_{\ell,j}, \alpha_{\ell,j}) \sim \Pi_Q \otimes \pi_\alpha$$

$$Q_{\ell,j} = \rho_{\ell,j} \times \text{Dirichlet Process}, \quad g_{\alpha,u} = \text{Beta}(\alpha/(1-u), \alpha/u)$$

Independent priors: computation and theory

$$d\Pi(f) = \prod_{j=1}^{K} d\Pi_f(f_j), \quad f_j = (\nu_j, h_{\ell j})$$

- Parallel computation : important

$$\pi(df_1, \cdots, df_K|N) = \prod_{j=1}^{K} \pi(df_j|N), \quad f_j = (\nu_k, h_{\ell k}, \ell \leq K)$$

But for each j K functions + 1 scalar to estimate + intractable likelihood

- Selection prior : faster with strong sparsity $k_0 << K$:

$$\text{Nb of possibilities} \quad \asymp K^{k_0 - 1/2} << 2^K$$

- Theory: study $\pi(df_j|N)$ separately.

Posterior contraction rates

- Aim : Posterior contraction rates : Find for all $j \leq K$, $u_T = o(1)$ s.t.

$$\Pi(d(f_j, f_j^0) \leq u_T | N) \xrightarrow{P_{f_0}} 1, \quad d(f_j, f_j^0) = |\nu_j - \nu_j^0| + \sum_{\ell \leq K} \|h_{\ell j} - h_{\ell j}^0\|_1$$

In particular • Fixed K theory [Donnet et al. ,Sulem et al.] If

- If $\Pi(\max_{\ell,j} \|h_{\ell j} - h_{\ell j}^0\|_\infty \leq \epsilon_T) \gtrsim e^{-c_1 T \epsilon_T^2}$
- If $N(\epsilon_T, \mathcal{H}_T, \|\cdot\|_1) \leq CT\epsilon_T^2$ and $\Pi(\mathcal{H}_T^c) \leq e^{-[c_1 + C(K)\log T] T \epsilon_T^2}$

$$\mathcal{H}_T \subset \{h_{\ell j}; \text{s.t.} \|\rho\|_1 \leq 1\}, \quad \rho = (\rho_{\ell j})_{\ell, j \leq K}, \quad \rho_{\ell j} = \int h_{\ell j}$$

Then

$$\Pi(\sum_{\ell,k} \|h_{\ell k} - h_{\ell,k}\| + \|\nu - \nu_0\|_1 \leq \tilde{C}(K)\sqrt{\log T}\epsilon_T | N) = 1 + o_{P_0}(1).$$

$C(K)$?? exponential dependence

Some empirical results: $K = 8$ $N^j[0, T] \approx 300$

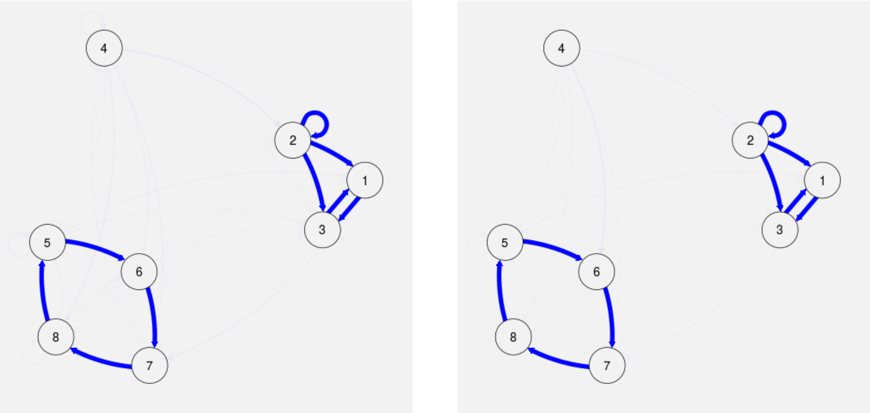


Figure: **Results for scenario 2 for one given dataset**. Posterior estimation of the interaction graph for $T = 10$ *on the left* and $T = 20$ *on the right*, for one randomly chosen dataset. Level of grey and width of the edges proportional to the posterior estimated probability of $\widehat{\mathbb{P}}(\delta_{\ell,k} = 1 | (N_t^{sim})_{t \ in[0,T]})$.

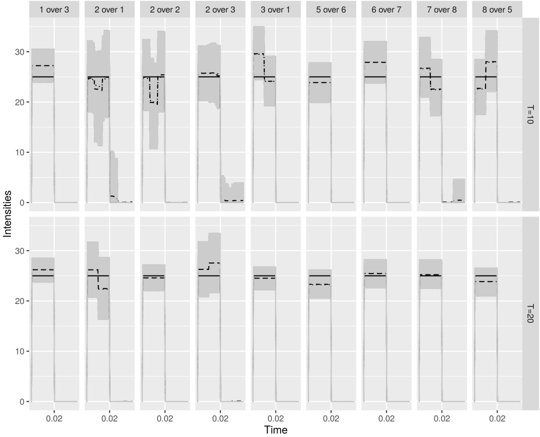


Figure: **Results for scenario 2 for one given dataset**. Estimation of the non null interaction functions $(h_{\ell,k})_{\ell,k=1,\dots,8}$ using the regular prior for $T = 10$ (upper panel) and $T = 20$ (bottom). The gray region indicates the credible region for $h_{\ell,k}(t)$ (delimited by the 5% and 95% percentiles of the posterior distribution). The true $h_{\ell,k}$ is in plain line, the posterior expectation and posterior median for $h_{\ell,k}(t)$ are in dotted and dashed lines respectively.

New posterior concentration rates: selection priors

- true Parameter $f_0 = (\nu^0, \mathbf{h}^0)$. $\rho^0_{\ell j} = \int_0^A h^0_{\ell j}(t)dt$, $A_0(j) = \{\ell;\ h^0_{\ell j} \neq 0\}$.
- Stronger stationary and sparsity assumptions:
  - $|||\rho_0|||_\infty = \max_\ell \sum_j \rho^0_{\ell j} \leq c < 1$ & $\max_j |A_0(j)| \leq a_0$ & $\forall j\ c_0 \leq \nu^0_j \leq 1/c_0$
  - $\max_{j,\ell} \|h^0_{\ell j}\|_\infty \leq C_0$ & $\max_j \mu^0_j \leq C_0$, $\mu^0_j = \mathbb{E}_0[\lambda_t(f_0)]$.
- Selection prior : $A(j) = \{\ell; h_{\ell j} \neq 0\}$ $a(j) = |A(j)| \sim \pi_a$, $h_{\ell j} \in \mathcal{H}$

Theorem (L_1 Posterior contraction rates)

If $\forall \ell \in A_0(j)$ $\Pi(\|h_{\ell j} - h^0_{\ell j}\|_\infty \leq \epsilon_T) \geq e^{-c_1 T \epsilon_T^2}$ & $\pi(\sum_{\ell \notin A_0(j)} \rho_{\ell j} \leq \epsilon_T^2) \geq e^{-c_1 T \epsilon_T^2}$

$$\exists \mathcal{H}_T \subset \mathcal{H} \quad \Pi_h(\mathcal{H}_T^c) = o(e^{-(2c_1+\kappa \log T) T {\epsilon_T}^2}), \quad \log \mathcal{N}(\epsilon_T, \mathcal{H}_T, \|\cdot\|_1) \lesssim T {\epsilon_T}^2.$$

Then, if $K = o(T)$ *and* $\Pi(a(j) \geq L_T) \leq e^{-(2c_1+\kappa \log T) T \epsilon_T^2}$, $L_T \leq \delta \log T$

$$\forall j; \mathbb{E}_0 \left[\Pi\left(|\nu_j - \nu^0_j| + \sum_\ell \|h_{\ell j} - h^0_{\ell j}\|_1 > e^{5 C_0 L_T} \epsilon_T \log T \middle| N \right) \right] = o(1).$$

Some comments

$$v_T = e^{5C_0 L_T} \epsilon_T \log T$$

- $\epsilon_T \approx$ rate if graph of interactions $A_0(j)$ was known
- Prior induced sparsity L_T: has an exponential effect on rate
- If π_a has support $\{0, \cdots, M_T\}$ with $M_T = O(\log\log T)$ Then

$$v_T = \epsilon_T \text{polylog } T.$$

Up to $\log T$ terms v_T is independent of K as soon as $K = o(T)$

- Result is $\forall j, \quad \Pi(\|f_j - f_j^0\|_1 \leq v_T(j)|N) = 1 + o_p(1)$, not $\Pi(\forall j, \|f_j - f_j^0\|_1 \leq v_T(j)|N) = 1 + o_p(1)$

What is ϵ_T?

ϵ_T depends on π_h and smoothness of $\mathbf{h}^0$

- Random histogram prior : $h_{\ell j}(x) = \sum_{i=0}^{I-1} 1_{]i/I,(i+1)/I]}(x) w_i^{\ell,j}/K$ if $\ell \in A(j)$

$$I \sim \mathcal{P}(M), \quad (w_1, \cdots, w_I, w_\emptyset) \sim \mathcal{D}(\alpha),$$

Under sparsity assumptions on f^0 and in Π, if $\mathbf{h}^0$ is Holder $\beta \leq 1$

$$\epsilon_T = T^{-\beta/(2\beta+1)}$$

- Mixture of Betas . Under same assumptions + MFM for Q; if $\mathbf{h}^0$ is Holder β, $\beta > 0$

$$\epsilon_T = T^{-\beta/(2\beta+1)}$$

Some comments on the selection prior

Prior on $A(j)$

- $a_j = |A(j)| \sim \pi_a$
- $[A(j)|a_j] \sim \mathcal{U}$; $h_{\ell j} = 0$ if $\ell \notin A(j)$
- $h_{\ell j} \sim \pi_h$

Spike and Slab prior $h_{\ell k} \sim (1-p)\delta_{(0)} + p\pi_{1,h}$ To have $\pi_a(|A(j)| > \sqrt{\log T}) \leq e^{-cT\epsilon_T^2 \log T}$ one needs

$$p \lesssim \frac{1}{KT\epsilon_T^2} = o(1/K) \quad \text{not desirable}$$

Examples of π_a truncated prior or

$$\pi_a(a(j) \geq x) \lesssim e^{-\gamma_2 e^{\gamma_1 x^r}}, \quad r > 1$$

A weaker intermediate result in empirical loss

Theorem (L_1 Posterior contraction rates)

If $\forall \ell \in A_0(j)$ $\Pi(\|h_{\ell j} - h^0_{\ell j}\|_\infty \leq \epsilon_T) \geq e^{-c_1 T \epsilon_T^2}$ & $\pi(\sum_{\ell \notin A_0(j)} \rho_{\ell j} \leq \epsilon_T^2) \geq e^{-c_1 T \epsilon_T^2}$

$$\exists \mathcal{H}_T \subset \mathcal{H} \quad \Pi_h(\mathcal{H}_T^c) = o(e^{-(2c_1+\kappa \log T) T {\epsilon_T}^2}), \quad \log \mathcal{N}(\epsilon_T, \mathcal{H}_T, \|\cdot\|_1) \lesssim T {\epsilon_T}^2.$$

Then, if $K = o(T)$ *and* $\Pi(a(j) \geq L_T) \leq e^{-(2c_1+\kappa \log T) T \epsilon_T^2}$

$$\forall j; \mathbb{E}_0 \left[\Pi\left(\int_0^T \frac{|\lambda_t(f_j) - \lambda_t(f_j^0)|}{T} dt > \sqrt{L_T} \epsilon_T \log T \Big| N \right) \right] = o(1).$$

L_T does not need to be so small

The loss $\int_0^T \frac{|\lambda_t(f_j) - \lambda_t(f_j^0)|}{T} dt$: prediction but not for parameter estimation

Insights on the results

- For the $d_{1,T} = \int_0^T |\lambda_t^j(f_j) - \lambda_t^j(f_j^0)|dt/T$ rate, $B_T(j)^c = \{d_{1,T}(f_j, f_j^0) \gtrsim u_T\}$

$$\Pi_j(B_{j,T}^c|N) = \frac{\int_{B_{j,T}^c} e^{L_T(f_j;N^j) - L_T(f_j^0;N^j)} d\Pi(f_j)}{\int_{\mathcal{F}} e^{L_T(f_j;N^j) - L_T(f_j^0;N^j)} d\Pi(f_j)} := \frac{\mathcal{N}_T(j)}{D_T(j)}$$

  - $\Pi(\max_{\ell \in A_0(j)} \|h_{\ell,j} - h_{\ell,j}^0\|_\infty \leq \epsilon_T;\ \sum_{\ell \notin A_0(j)} \rho_{\ell j} \leq \epsilon_T^2) \gtrsim e^{-c_1 T \epsilon_T^2}$ implies

$$D_T \gtrsim e^{-C_1 T \epsilon_T^2 \log T} \quad \text{w.h.p}$$

  - L_1 Entropy of $\{h_{\ell j}, \ell \leq S\}$: $E_T(S) \approx |S| E_T(\mathcal{H}_T)$ and $|S| \leq L_T$

$$E_T \lesssim L_T T \epsilon_T^2 \quad \Rightarrow N_T(j) = o(D_T) \quad \text{if} \quad u_T = \sqrt{L_T}\epsilon_T \log T$$

Insights on the results

- For the $d_{1,T} = \int_0^T |\lambda_t^j(f_j) - \lambda_t^j(f_j^0)| dt/T$ rate, $B_T(j)^c = \{d_{1,T}(f_j, f_j^0) \gtrsim u_T\}$

$$\Pi_j(B_{j,T}^c|N) = \frac{\int_{B_{j,T}^c} e^{L_T(f_j;N^j) - L_T(f_j^0;N^j)} d\Pi(f_j)}{\int_{\mathcal{F}} e^{L_T(f_j;N^j) - L_T(f_j^0;N^j)} d\Pi(f_j)} := \frac{\mathcal{N}_T(j)}{D_T(j)}$$

  - $\Pi(\max_{\ell \in A_0(j)} \|h_{\ell,j} - h_{\ell,j}^0\|_\infty \leq \epsilon_T;\ \sum_{\ell \notin A_0(j)} \rho_{\ell j} \leq \epsilon_T^2) \gtrsim e^{-c_1 T \epsilon_T^2}$ implies

$$D_T \gtrsim e^{-C_1 T \epsilon_T^2 \log T} \quad \text{w.h.p}$$

  - L_1 Entropy of $\{h_{\ell j}, \ell \leq S\}$: $E_T(S) \approx |S| E_T(\mathcal{H}_T)$ and $|S| \leq L_T$

$$E_T \lesssim L_T T \epsilon_T^2 \quad \Rightarrow N_T(j) = o(D_T) \quad \text{if} \quad u_T = \sqrt{L_T}\epsilon_T \log T$$

Difficulty : going from $d_{1,T}$ to L_1

Going from $d_{1,T}$ to L_1

$$|\lambda_t(f_j)-\lambda_t(f_j^0)| = |\nu_j-\nu_j^0+\underbrace{\sum_{\ell\in A_0(j)^c\cap A(j)}\int_{t-A}^{t^-} h_{\ell j}(t-s)dN_s^\ell}_{T_2}+\underbrace{\sum_{\ell\in A_0(j)}\int_{t-A}^{t^-}(h_{\ell j}-h_{\ell j}^0)(t-s))dN}_{=0 \text{ if } N^{A_0}[t-A,t)=0}$$

To separate $\nu_j - \nu_j^0$ from h terms: we need $N^{A(j)}[t-A,t)=0$:

- probability $P(N^{A(j)}[t-A,t)=0) \asymp e^{-c_1|A(j)|}$: curse of dimensionality

Bernstein von Mises for functionals of f

- Functionals of interest : $\Psi(f) = \nu_j$ or $\Psi(f) = \int b(x) h_{\ell j}(x) dx$
- Questions
  - Can we estimate $\theta = \Psi(f)$ at a faster rate ?
  - Can we derive en efficient theory ? BvM ?

Bernstein von Mises theorem: general theory [Castillo, R. & Rivoirard, R.]

Existing theory on BvM for $\theta = \Psi(f)$

- If LAN : $\ell_T(f) - \ell_T(f_0) = \sqrt{T} W_T(f - f_0) - \frac{T\|f-f_0\|_L^2}{2} +$ Rest
- If smooth Ψ:

$$\Psi(f) = \Psi(f_0) + < \tilde{\psi}_0, f - f_0 >_L + rest$$

- If $\Pi(A_T|N) = 1 + o_p(1)$ &

$$\frac{\int_{A_n} e^{\ell_T(f_t)} d\Pi(f)}{\int_{A_T} e^{\ell_T(f)} d\Pi(f)} = 1 + o_{P_0}(1), \quad f_t = f - \frac{t\tilde{\psi}_0}{\sqrt{T}}$$

$\tilde{\psi}_0 =$ *least favorable direction* for $\Psi(f)$ at f_0

Bernstein von Mises theorem: general theory [Castillo, R. & Rivoirard, R.]

Existing theory on BvM for $\theta = \Psi(f)$

- If LAN : $\ell_T(f) - \ell_T(f_0) = \sqrt{T} W_T(f - f_0) - \frac{T\|f-f_0\|_L^2}{2} +$ Rest
- If smooth Ψ:

$$\Psi(f) = \Psi(f_0) + < \tilde{\psi}_0, f - f_0 >_L + rest$$

- If $\Pi(A_T|N) = 1 + o_p(1)$ &

$$\frac{\int_{A_n} e^{\ell_T(f_t)} d\Pi(f)}{\int_{A_T} e^{\ell_T(f)} d\Pi(f)} = 1 + o_{P_0}(1), \quad f_t = f - \frac{t\tilde{\psi}_0}{\sqrt{T}}$$

$\tilde{\psi}_0 =$ *least favorable direction* for $\Psi(f)$ at f_0

Then $d_{BL}(\Pi(\sqrt{T}(\theta - \hat{\theta})|N), \mathcal{N}(0, \|\tilde{\psi}_0\|_L^2)) = o_{P_0}(1)$: BvM

LAN expansion: univariate case (for simplicity)

- True parameter: $f_0 = (\nu_0, h_0)$, $\int_0^A h_0(x)dx < 1$
- LAN expansion $a \in \mathbb{R}, g \in L_\infty(0, A)$

$$\ell_T(\nu_0 + a/\sqrt{T}, h_0 + g/\sqrt{T}) - \ell_T(\nu_0, h_0) = W_T(a, g) - \frac{\|(a,g)\|_L^2}{2} + \text{Rest}$$

- LAN scalar product : $\lambda_A(a, g) = a + \int_0^{A^-} g(A - s)dN_s$

$$< (a_1, g_1), (a_2, g_2) >_L = \mathbb{E}_0\left(\frac{\lambda_A((a_1, g_1))\lambda_A((a_2, g_2))}{\lambda_A(\eta_0)}\right),$$

$$W_T(a, g) = \frac{1}{\sqrt{T}} \int_0^T \frac{\lambda_t((a, g))}{\lambda_t(\eta_0)}[dN_t - \lambda_t(\eta_0)dt] \approx \mathcal{N}(0, \|(a, g)\|_L^2)$$

Comment on boundary effect

$$\ell_T(\nu_0 + a/\sqrt{T}, h_0 + g/\sqrt{T}) - \ell_T(\nu_0, h_0) = W_T(a,g) - \frac{\|(a,g)\|_L^2}{2} + \text{Rest}$$

- We need $h_0 + g/\sqrt{T} \in \mathcal{H}$, i.e. $h_0 + g/\sqrt{T} \geq 0$.
  - If $h_0 \geq c_0 > 0$ then $h_0 + g/\sqrt{T} \geq c_0 - \|g\|_\infty/\sqrt{T} > 0$ for T large: No problem
  - If $\inf_x h_0(x) = 0$, then tangent set :

$$\dot{\mathcal{H}} = \{g \in L_\infty(0,A); g(x) \geq 0 \ \forall x \in I_0\}, \quad I_0 = \{x; h_0(x) = 0\}$$

. This case is important for Hawkes processes

least favorable direction $(r_0, g_0) \in \mathbb{R} \times L_\infty$, for $\Psi(\eta) = \nu$

$$r_0 = (1 - \int g_0 f(x)dx)/a_0, \quad g_0 + \frac{\gamma(g_0)}{f} = -r_0$$

where

$$f(x) = \mathbb{E}_0\left(\frac{dN_x}{\lambda_A(\eta_0)}\right), \quad a_0 = \mathbb{E}_0(1/\lambda_A(\eta_0))$$

$$\gamma(g_0)(x) = \int_0^A g_0(s)\mathbb{E}_0\left(\frac{dN_s dN_x}{\lambda_A(\eta_0)}\right) ds$$

g_0 non explicit :
(1) makes it harder to study the change of variable condition $h - ug_0/\sqrt{T}$ in BvM
(2) Not clear that $h - ug_0/\sqrt{T} \geq 0$ if $\inf_x h(x) \approx 0$

Bayesian and frequentist consequences

- Frequentist : $\Psi(\eta) = a\nu + \int b(x)h(x)dx$ can be estimated at the rate $\sqrt{T}$ + efficient theory . Efficient estimator in the form

$$\hat{\Psi} = \Psi(\eta_0) + \frac{W_T(r_0, g_0)}{\sqrt{T}} + o_{P_0}(1)$$

- Bayesian BvM if for all $|u|$ small

$$\frac{\int_{B_n} e^{\ell_T(\eta_u)} d\Pi(\eta)}{\int_{B_n} e^{\ell_T(\eta)} d\Pi(\eta)} = 1 + o_{P_0}(1), \quad \eta_u = \eta - u\frac{(r_0, g_0)}{\sqrt{T}}$$

$$\sup_{\eta \in B_n} \left| \frac{d\Pi^{(u)}(\eta)}{d\Pi(\eta)} - 1 \right| = o(1)?$$

Two reasons to fail: smoothness and positivity

example : random histogram prior

$$h(x) = \sum_{j=1}^{J} h_j 1_{I_j}(x), \quad [h_j|J] \overset{iid}{\sim} \pi_h, \quad J \sim \Pi_J$$

- smoothness (usual story) : Then $h_u = h - \frac{ug_0}{\sqrt{T}} \neq \sum_{j=1}^{J} \tilde{h}_j 1_{I_j}$
- Positivity (unusual): if $h_0 = 0$, then $h \approx 0$ and $h - ug_0/\sqrt{T}$ may by ≤ 0

Two reasons to fail: smoothness and positivity

example : random histogram prior

$$h(x) = \sum_{j=1}^{J} h_j 1_{I_j}(x), \quad [h_j|J] \overset{iid}{\sim} \pi_h, \quad J \sim \Pi_J$$

- smoothness (usual story) : Then $h_u = h - \frac{ug_0}{\sqrt{T}} \neq \sum_{j=1}^{J} \tilde{h}_j 1_{I_j}$
- Positivity (unusual): if $h_0 = 0$, then $h \approx 0$ and $h - ug_0/\sqrt{T}$ may by ≤ 0

$\Pi^{(u)}$ is not absolutely continuous wrt $\quad \Pi$

Enlarging and undersmoothing

$$h(x) = \sum_{j=1} h_j \mathbf{1}_{I_j}(x), \quad [h_j|J] \overset{iid}{\sim} \pi_h, \quad J \sim \Pi_J$$

- Enlarging

$$\Pi(dh) = \frac{\mathbf{1}_{h \geq -c_T} \bar{\Pi}(dh)}{\bar{\Pi}(h \geq -c_T)} \quad c_T \lesssim \frac{\nu}{1 + \log T^2}$$

or non linear ReLU model

$$\lambda_t(\eta) = \left(\nu + \int^{t^-} h(t-s) dN_s\right)_+, \quad h : [0, A] \to \mathbb{R}$$

•Undersmoothing

$$[\sqrt{T}(\Psi(\eta) - \hat{\Psi})|N] \approx \sum_J \Pi(J|N)[\sqrt{T} b_0(J) + \mathcal{N}(0, v_J)]$$

$$b_0(j) =< (0, h_0 - h_{0,J}), (0, g_0 - g_{0,J}) >_{LAN}, \quad v_J \underbrace{=}_{J \to \infty} \|(r_0, g_0)\|_L^2$$

Some comments and open questions

- Prior sparcity L_T : Influence $\sqrt{L_T}\epsilon_T$ for *empirical loss* but $e^{5L_T C_0}\epsilon_T$ for the $\|f - f_0\|_1$ loss. Is it sharp?
- Provides a minimax estimation rate [up to log T]: $T^{-\beta/(2\beta+1)}$ for $K = o(T)$. What happens if $T^B \gtrsim K >> T$?
- Prior does not depend on $|A_0(j)|$, a_0, $c = \max_l \sum_k \rho^0_{lk} < 1$, β : Fully adaptive.
- Selection priors better than spike and slab priors. Both lead to (fairly) computationally intensive algos (necessity to search the space) : Derive families of Bayesian algos that scale with K but remain *statistically optimal*
- Semi parametric inference: do we really have to enlarge the model?
- Need to target the prior for the functional ? Not satisfying (like plug in estimate).

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

European Research Council
Established by the European Commission