Statistical Learning with Hawkes Processes

5th NUS-USPC Workshop

Stéphane Gaïffas

1 Motivation and examples

2 Hawkes processes

3 Dimension reduction for MHP

4 Random matrix theory

5 Causality maps

6 Accelerating training time

7 Software

Motivation and examples

Example 1. Social networks. Understand who is influencing **twitter**: based on the timestamps patterns of messages, **web-data**: publication activity of websites/blogs

Example 2. High Frequency Finance.
From zoomed financial signals ($\Delta t \approx 1$ms, upward / downward price proves and other order book features), build a "causality map"

Example 3. Health-care. Impact of some health events to other health events (all being timestamped, longitudinal data)

1 Motivation and examples

2 Hawkes processes

3 Dimension reduction for MHP

4 Random matrix theory

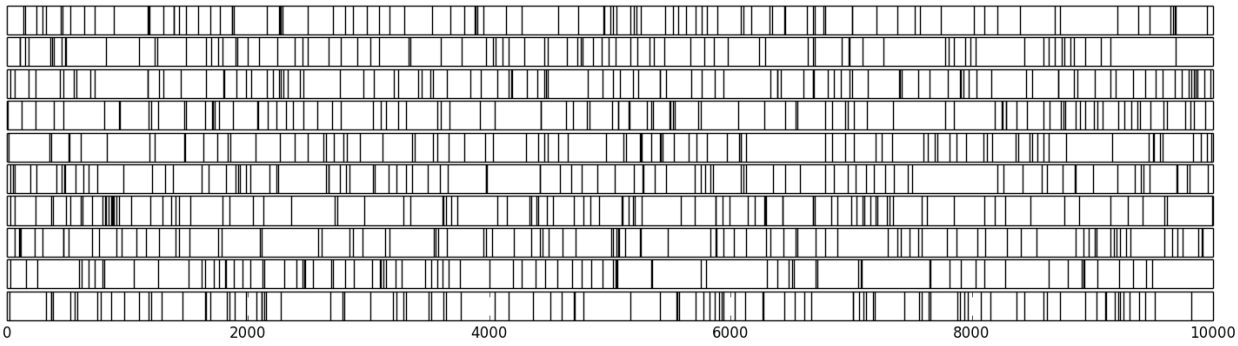
5 Causality maps

6 Accelerating training time

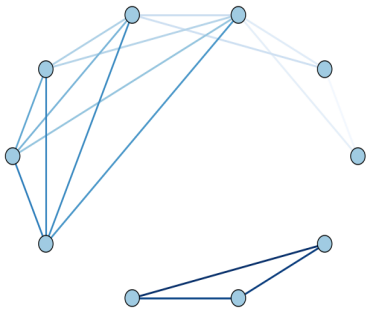
7 Software

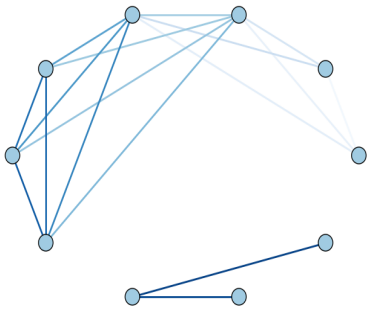
Introduction

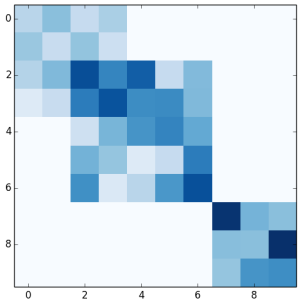
From:



Build:







Introduction

Setting

- For each node $i \in I = \{1, \dots, d\}$ we have a set Z^i of *events*
- Any $\tau \in Z^i$ is the occurence time of an event related to i

Counting process

- Put $N_t = [N_t^1 \cdots N_t^d]^\top$
- $N_t^i = \sum_{\tau \in Z^i} \mathbf{1}_{\tau \leq t}$

Intensity

- Stochastic intensities $\lambda_t = [\lambda_t^1 \cdots \lambda_t^d]^\top$, $\lambda_t^i =$ intensity of N_t^i

$$\lambda_t^i = \lim_{dt \to 0} \frac{\mathbb{P}(N_{t+dt}^i - N_t^i = 1 | \mathcal{F}_t)}{dt}$$

- $\lambda_t^i =$ instantaneous rate of event occurence at time t for node i
- λ_t characterizes the distribution of N_t [Daley et al. 2007]
- Patterns can be captured by *putting structure* on λ_t

The Multivariate Hawkes Process (MHP)

Scaling

- We observe N_t on $[0, T]$. "Asymptotics" in $T \to +\infty$. d is "large"

The Hawkes process

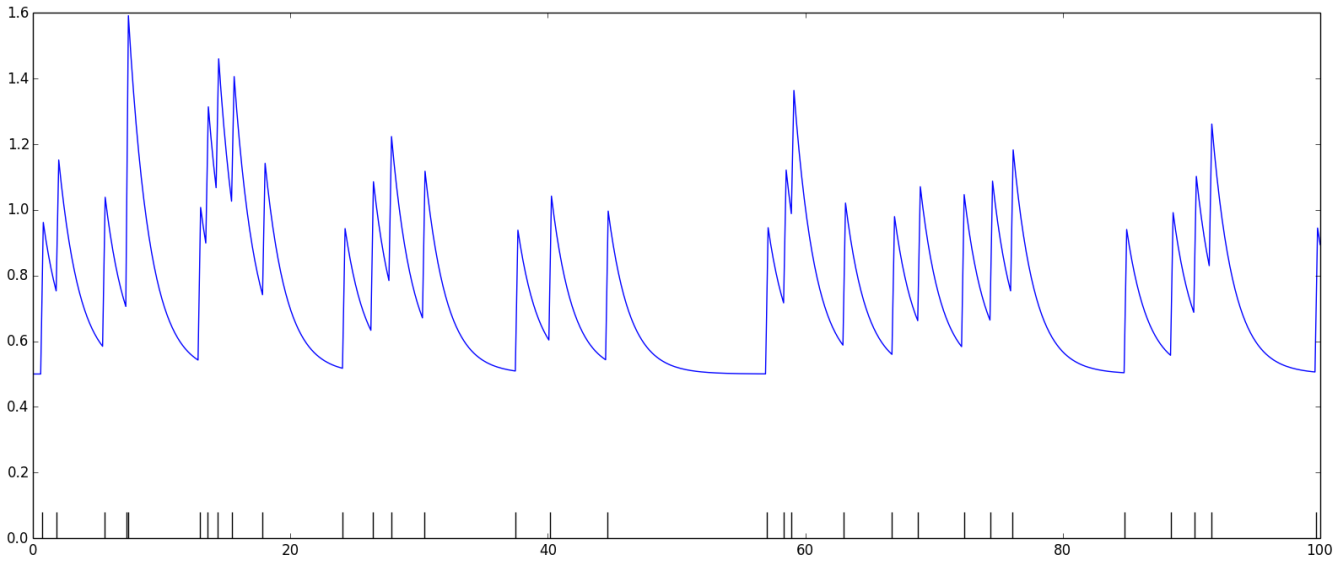
- A particular structure for λ_t: auto-regression
- N_t is called a *Hawkes* process [Hawkes 1971] if

$$\lambda_t^i = \mu_i + \sum_{j=1}^{d} \int_0^t \varphi^{ij}(t - t') dN_{t'}^j = \mu_i + \sum_{j=1}^{d} \sum_{t' \in Z_j : t' < t} \varphi^{ij}(t - t')$$

- $\mu_i \in \mathbb{R}^+$ exogenous intensity
- φ^{ij} non-negative integrable and causal (support $\mathbb{R}_+$) functions
- φ^{ij} are called *kernels*. Encodes the impact of an action by node j on the activity of node i
- Captures *auto-excitation* and *cross-excitation* across nodes, a phenomenon observed in social networks [Crane et al. 2008]

A simple parametrization of the MHP

For $d = 1$, $K = 1$ and $\varphi^{11}(t) = e^{-1}$, intensity $\lambda_{\theta,t}$ looks like:



Stability condition of the MHP

Stability condition

- Introduce
$$\boldsymbol{G}^{ij} = \int_0^{+\infty} \varphi^{ij}(t)dt$$
- Spectral norm must satisfy $\|\boldsymbol{G}\| < 1$ to ensure stability and stationarity of the process

Sum of exponentials **parametric** model:

$$\lambda^i_{\theta,t} = \mu_i + \int_{(0,t)} \sum_{j=1}^{d} \sum_{k=1}^{K} a^k_{ij} \times \alpha_k e^{-\alpha_k (t-s)} dN^j_s$$

for $i \in \{1, \dots, d\}$ with $\alpha_1, \dots, \alpha_K > 0$ given and parameters to infer are $\theta = [\mu, \boldsymbol{A}]$ with

- baselines $\mu = [\mu_1 \cdots \mu_d]^\top \in \mathbb{R}^d_+$
- interactions $\boldsymbol{A} = [a_{ij}]_{1 \le i,j \le d} \in \mathbb{R}^{d \times d}_+ =$ "adjacency matrix"

A brief history of MHP

Brief history

- Introduced in Hawkes 1971
- Earthquakes and geophysics [Kagan and Knopoff 1981], [Zhuang et al. 2012]
- Genomics [Reynaud-Bouret and Schbath 2010]
- High-frequency Finance [Bacry et al. 2013]
- Terrorist activity [Mohler et al. 2011, Porter and White 2012]
- Neurobiology [Hansen et al. 2012]
- Social networks [Carne and Sornette 2008], [Zhou et al.2013]
- And even FPGA-based implementation [Guo and Luk 2013]

A brief history of MHP

THE GENESIS BLOCK

Digital currency research and data

HOME | NEWS | MINING | TRADING | ECONOMICS | REGULATION | BUSINESSES | BI

Home / Bitcoin 201 / Analyzing Trade Clustering To Predict Price Movement In Bitcoin Trading

Analyzing Trade Clustering To Predict Price Movement In Bitcoin Trading

Sep 19, 2013 Posted By Jonathan Heusser In Bitcoin 201, Economics, Featured, News, Trading Tagged Analysis, Bitcoin Trading, Hawkes Process, Jonathan Heusser, London, Price, Trading

MHP in large dimension

What do we want to do?

- Deal with large number of events and large dimension d (number of nodes)
- End up with a *tractable* and *scalable* optimization problem

Goodness-of-fit functionals. Two choices: **minus** log-**likelihood**

$$-\ell_T(\theta) = \frac{1}{T}\sum_{i=1}^{d}\Big\{\int_0^T \lambda^i_{\theta,t}dt - \int_0^T \log\lambda^i_{\theta,t}dN^i_t\Big\}$$

or **least-squares**

$$R_T(\theta) = \frac{1}{T}\sum_{i=1}^{d}\Big\{\int_0^T (\lambda^i_{\theta,t})^2dt - 2\int_0^T \lambda^i_{\theta,t}dN^i_t\Big\}$$

1 Motivation and examples

2 Hawkes processes

3 Dimension reduction for MHP

4 Random matrix theory

5 Causality maps

6 Accelerating training time

7 Software

Contribution 1. Dimension reduction for MHP

Paper. E. Bacry, S. G., J.-F. Muzy, *A generalization error bound for sparse and low-rank multivariate Hawkes processes*, in revision in JMLR

- Parametric setting $\varphi^{ij}(t) = (\boldsymbol{A})_{ij} \times h(t)$
- Low-rank and sparsity inducing penalization on $\boldsymbol{A}$
- Introduces a sharp tuning of the penalizations using data-driven weights
- Leads to optimal error bounds for penalized least-squares (sharp sparse oracle inequality)

Contribution 1. Dimension reduction for MHP

Prior assumptions

- Users are basically inactive and react mostly if stimulated:

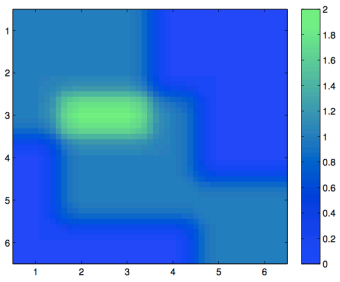
  μ is sparse

- Everybody does not interact with everybody:

  $\boldsymbol{A}$ is sparse

- Interactions have community structure, possibly overlapping, a small number of factors explain interactions:

  $\boldsymbol{A}$ is low-rank



Contribution 1. Dimension reduction for MHP

Standard convex relaxations
(Tibshirani (01), Srebro et al. (05), Bach (08), Candès & Tao (09), etc.)

- Convex relaxation of $\|\boldsymbol{A}\|_0 = \sum_{ij} \mathbf{1}_{\boldsymbol{A}_{ij}>0}$ is ℓ_1-norm:

$$\|\boldsymbol{A}\|_1 = \sum_{ij} |\boldsymbol{A}_{ij}|$$

- Convex relaxation of rank is trace-norm:

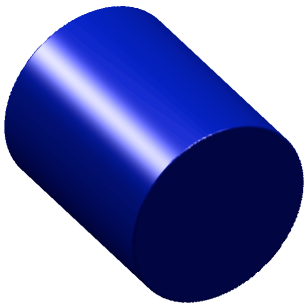
$$\|\boldsymbol{A}\|_* = \sum_{j} \sigma_j(\boldsymbol{A}) = \|\sigma(\boldsymbol{A})\|_1$$

where $\sigma_1(\boldsymbol{A}) \geq \cdots \geq \sigma_d(\boldsymbol{A})$ singular values of $\boldsymbol{A}$

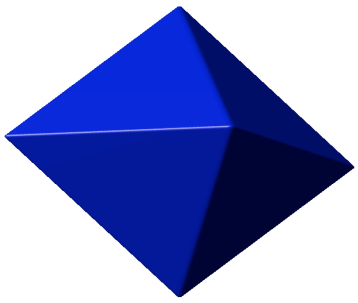
Contribution 1. Dimension reduction for MHP

We use the following penalizations

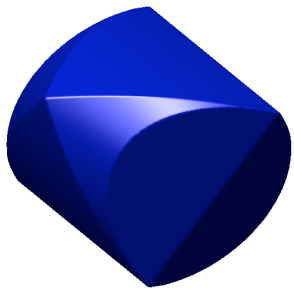
- Use ℓ_1 penalization on μ
- Use ℓ_1 penalization on $\boldsymbol{A}$
- Use trace-norm penalization on $\boldsymbol{A}$



$\{\boldsymbol{A} : \|A\|_* \leq 1\}$



$\{\boldsymbol{A} : \|\boldsymbol{A}\|_1 \leq 1\}$



$\{\boldsymbol{A} : \|\boldsymbol{A}\|_1 + \|\boldsymbol{A}\|_* \leq 1\}$

Balls are on the set of 2×2 symmetric matrices identified with $\mathbb{R}^3$.

Contribution 1. Dimension reduction for MHP

Leads to

$$\hat{\theta} = (\hat{\mu}, \hat{\boldsymbol{A}}) \in \underset{\theta=(\mu,\boldsymbol{A})\in\mathbb{R}_+^d\times\mathbb{R}_+^{d\times d}}{\operatorname{argmin}} \big\{R_T(\theta) + \operatorname{pen}(\theta)\big\},$$

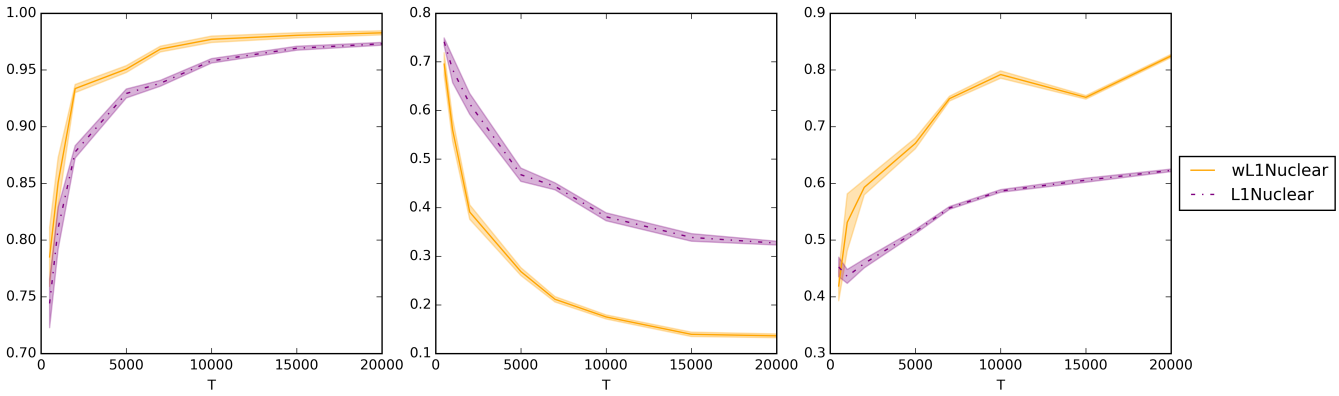
with penalization

$$\operatorname{pen}(\theta) = \tau_1\|\mu\|_1 + \gamma_1\|\boldsymbol{A}\|_1 + \gamma_*\|\boldsymbol{A}\|_*$$

The **features scaling problem**

- Features scaling is necessary for "linear approaches" in supervised learning
- No features and labels here!
- We solve this by sharp data-driven tuning of the penalization terms
- Required a new theory for random matrices with entries that are continuous-time martingales

Contribution 1. Dimension reduction for MHP



Left: AUC; *Middle*: Estimation error; *Right*: Kandall rank correlation

Contribution 1. Dimension reduction for MHP

A strong theoretical guarantee

- Recall $\langle \lambda_1, \lambda_2 \rangle_T = \frac{1}{T} \sum_{i=1}^{d} \int_0^T \lambda_{1,t}^i \lambda_{2,t}^i dt$ and $\|\lambda\|_T^2 = \langle \lambda, \lambda \rangle_T$
- Assume RE in our setting (Restricted Eigenvalues, Compressed Sensing literature)

Theorem. We have

$$\|\lambda_{\hat{\theta}} - \lambda^*\|_T^2 \leq \inf_{\theta} \left\{ \|\lambda_\theta - \lambda^*\|_T^2 + c_1 \kappa(\theta)^2 \Big(\|(\hat{w})_{\mathrm{supp}(\mu)}\|_2^2 + \|(\hat{\boldsymbol{W}})_{\mathrm{supp}(\boldsymbol{A})}\|_F^2 + \hat{w}_*^2 \, \mathrm{rank}(\boldsymbol{A}) \Big) \right\}$$

with a probability larger than $1 - c_2 e^{-x}$.

Contribution 1. Dimension reduction for MHP

Roughly, $\hat{\theta}$ achieves an optimal tradeoff between approximation and complexity given by

$$\frac{\|\mu\|_0 \log d}{T} \max_i \bar{N}^i([0,T]) + \frac{\|\boldsymbol{A}\|_0 \log d}{T} \max_{ij} \hat{v}_T^{ij} + \frac{\text{rank}(A) \log d}{T} \lambda_{\max}(\hat{\boldsymbol{V}}_T)$$

- Complexity measured both by sparsity and rank
- Convergence has shape $(\log d)/T$, where T = length of the observation interval
- Terms are balanced by "empirical variance" terms

1 Motivation and examples

2 Hawkes processes

3 Dimension reduction for MHP

4 Random matrix theory

5 Causality maps

6 Accelerating training time

7 Software

Contribution 2. Random matrix theory

Paper. E. Bacry, S. G. and J-F Muzy, *Concentration inequalities for matrix martingales in continuous time*, PTRF (2017)

Consider a $m \times n$ matrix-martingale given by

$$\boldsymbol{Z}_t = \int_0^t \mathbb{T}_s \circ d\boldsymbol{M}_s,$$

with $(\boldsymbol{M}_t)_{t\geq 0}$ "white" Brownian random matrix and $(\mathbb{T}_t)_{t\geq 0}$ rank-4 preditable tensor. Then

$$\mathbb{P}\left[\|\boldsymbol{Z}_t\|_{\mathrm{op}} \geq \sqrt{2v(x+\log(m+n))}\ ,\ \sigma^2(\boldsymbol{Z}_t) \leq v\right] \leq e^{-x}$$

for any $v, x > 0$ with

$$\sigma^2(\boldsymbol{Z}_t) = \max\left(\left\|\sum_{j=1}^n \langle \boldsymbol{Z}_{\bullet,j}\rangle_t\right\|_{\mathrm{op}}, \left\|\sum_{j=1}^m \langle \boldsymbol{Z}_{j,\bullet}\rangle_t\right\|_{\mathrm{op}}\right).$$

Contribution 2. Random matrix theory

- Strong generalization of previously known inequalities to continuous time (Tropp 2011)
- Very different approach (random matrix tools + stochastic calculus)
- Also the "Poissonian" case: martingale with sub-exponential jumps (counting process, Hawkes processes)

Interesting particular case (previously unknown!). Consider $\boldsymbol{P} = [P_{ij}]$ a $n \times m$ random matrix where P_{ij} is Poisson(λ_{ij}) and put $\boldsymbol{\lambda} = [\lambda_{ij}]$. Then

$$\mathbb{P}\left(\|\boldsymbol{N} - \boldsymbol{\lambda}\|_{\mathrm{op}} \geq \sqrt{2(\|\boldsymbol{\lambda}\|_{1,\infty} \vee \|\boldsymbol{\lambda}\|_{\infty,1})x} + \frac{x}{3} \right) \leq (n+m)e^{-x}$$

for any $x > 0$, where $\|\boldsymbol{\lambda}\|_{1,\infty}$ (resp. $\|\boldsymbol{\lambda}\|_{\infty,1}$) stands for the maximum ℓ_1-norm of rows (resp. columns)

1 Motivation and examples

2 Hawkes processes

3 Dimension reduction for MHP

4 Random matrix theory

5 Causality maps

6 Accelerating training time

7 Software

Contribution 3. Causality maps from MHP without parametric modeling

Paper. Achab et al. *Uncovering Causality from Multivariate Hawkes Integrated Cumulants*, ICML (2017) and JMLR (2017)

Reminder.

$$\lambda_t^i = \mu_i + \sum_{j=1}^{d} \int_0^t \varphi^{ij}(t - t') dN_{t'}^j,$$

Idea

- Don't estimate φ^{ij} (parametric, non-parametric), but only

$$g^{ij} = (\boldsymbol{G})_{ij} = \int_0^{+\infty} \varphi^{ij}$$

- Introducing the (unobserved) counting process $N^{i \leftarrow j}$ = of events from i with direct ancestor from j, we have

$$\mathbb{E}[dN_t^{i \leftarrow j}] = g^{ij} \mathbb{E}[dN_t^j]$$

- g^{ij} = average number of events from i triggered by one event from j

Contribution 3. Causality maps from MHP without parametric modeling

- Actually, if $\varphi^{ij} \geq 0$ then $g^{ij} = 0$ iff N^i does not Granger-cause N^j
- The matrix $\boldsymbol{G}$ encodes **causality**

How to estimate $\boldsymbol{G}$ directly?

- We know (Jovanovic 2014) how to relate integrated cumulants of N_t to $\boldsymbol{R} = (\boldsymbol{I} - \boldsymbol{G})^{-1}$

$$\Lambda^i = \sum_{m=1}^{d} R^{im} \mu^m$$

$$C^{ij} = \sum_{m=1}^{d} \Lambda^m R^{im} R^{jm}$$

$$K^{ijk} = \sum_{m=1}^{d} (R^{im} R^{jm} C^{km} + R^{im} C^{jm} R^{km} + C^{im} R^{jm} R^{km} - 2\Lambda^m R^{im} R^{jm} R^{km}).$$

Contribution 3. Causality maps from MHP without parametric modeling

NPHC. Cumulant matching method for estimation of $\boldsymbol{G}$

- Compute estimates $\widehat{\boldsymbol{C}}$ and $\widehat{\boldsymbol{K^c}}$ of the third order cumulants of the process
- Find $\widehat{\boldsymbol{R}}$ that matches these empirical cumulants

$$\mathcal{L}(\boldsymbol{R}) = (1-\kappa)\|\boldsymbol{K^c}(\boldsymbol{R}) - \widehat{\boldsymbol{K^c}}\|_2^2 + \kappa\|\boldsymbol{C}(\boldsymbol{R}) - \widehat{\boldsymbol{C}}\|_2^2,$$

- Put

$$\widehat{\boldsymbol{R}} = \boldsymbol{I} - \left(\arg\min_{\boldsymbol{R}\in\Theta} \mathcal{L}_T(\boldsymbol{R})\right)^{-1}$$

Theorem.

- It is consistent! (under some assumptions... quite technical)

Contribution 3. Causality maps from MHP without parametric modeling

Remarks.

- Highly non-convex problem: polynomial or order 10 with respect to the entries of $\boldsymbol{R}$
- Not so hard, local minima turns out to be good (deep learning literature), we simply use AdaGrad
- Using order **three** is important (two is not enough): integrated covariance contains only symmetric information: unable to provide causal information
- NPHC **scales better** than state-of-the-art methods, is **robust** towards the kernel shape and directly outputs the **kernel integral**
- Simple `tensorflow` code

Contribution 3. Causality maps from MHP without parametric modeling

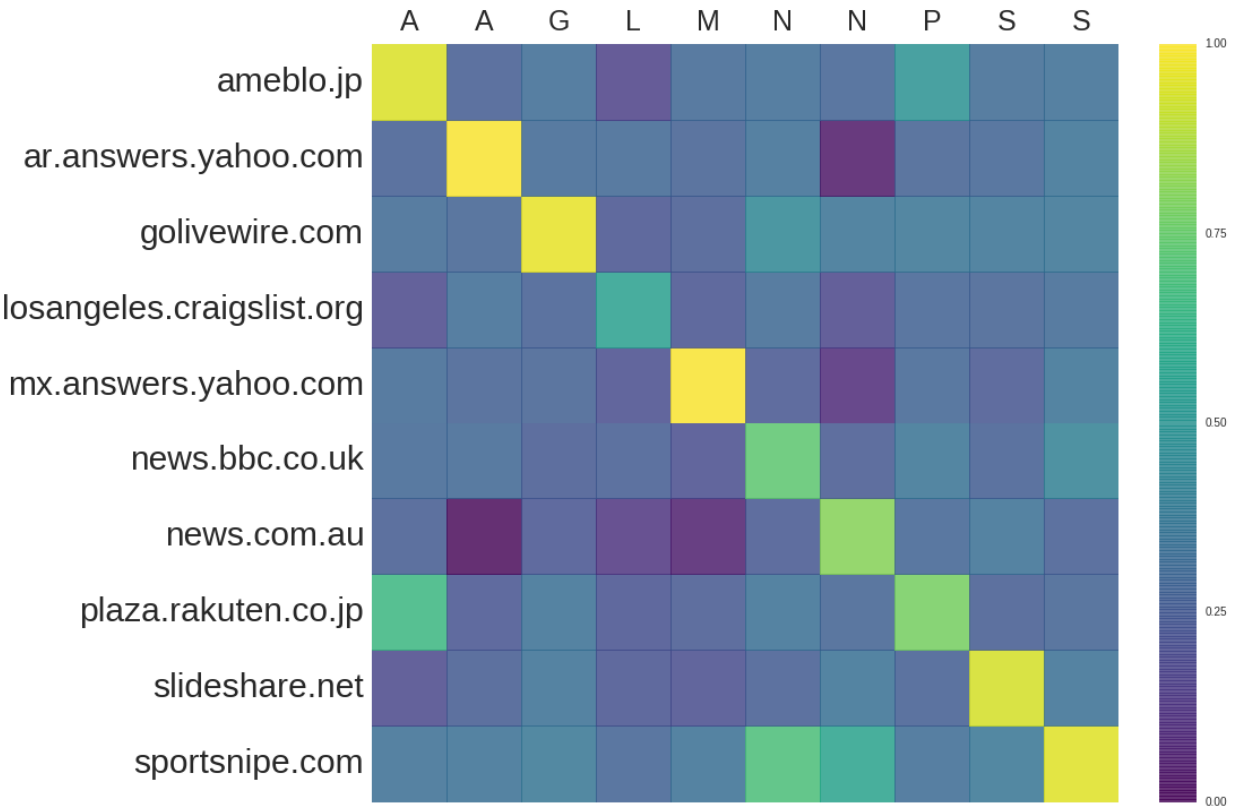
Experiment with MemeTracker dataset

- keep the 200 most active sites
- contains publication times of articles in many websites/blogs, with hyperlinks
- $\approx$ 8 millions events
- Use hyperlinks to establish an estimated ground truth for the matrix $\boldsymbol{G}$

Method	ODE	GC	ADM4	NPHC
RelErr	0.162	0.19	0.092	**0.071**
MRankCorr	0.07	0.053	0.081	**0.095**
Time (s)	2944	2780	2217	**38**

Contribution 3. Causality maps from MHP without parametric modeling

Experiment with MemeTracker dataset



$\widehat{G}$

Contribution 3. Causality maps from MHP without parametric modeling

Order book dynamics.

- Order book: a list of buy and sell orders for a specific financial instrument, the list being updated in real-time throughout the day
- Understand the self and cross-influencing dynamics of all event types in an order book

Introduce

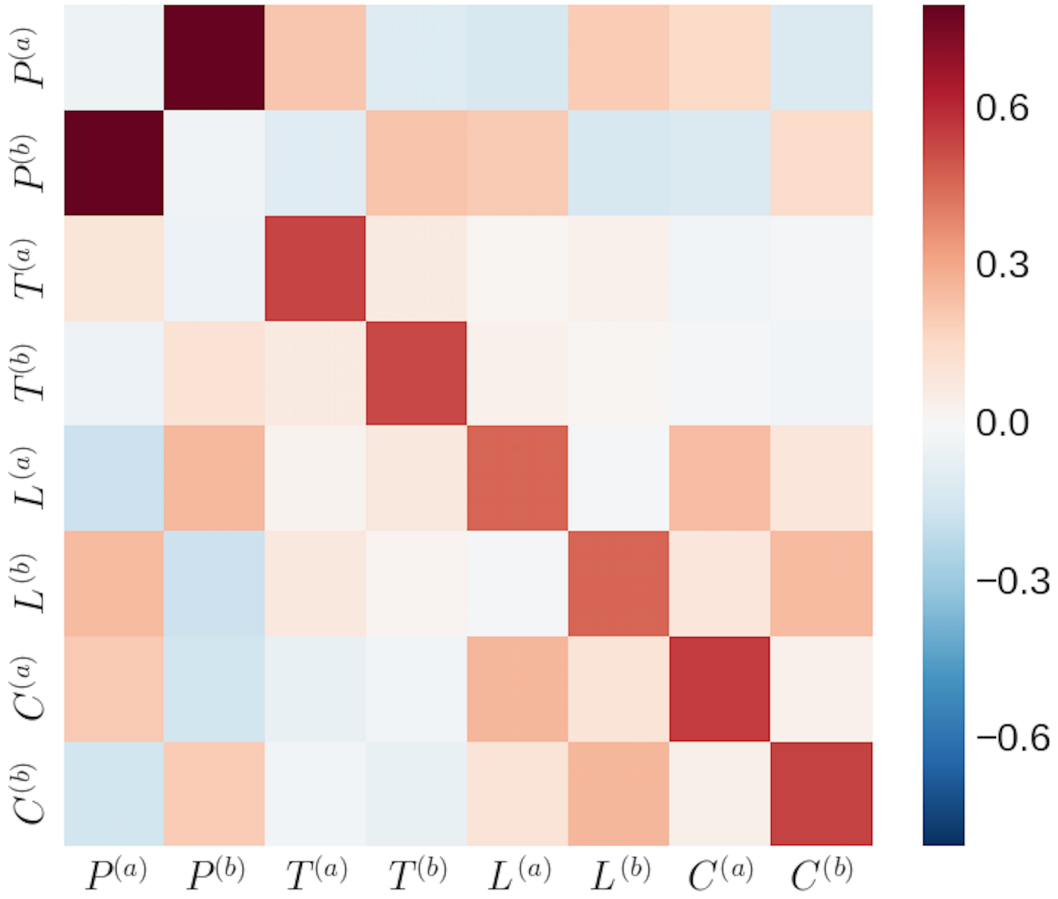
$$N_t = (P_t^{(a)}, P_t^{(b)}, T_t^{(a)}, T_t^{(b)}, L_t^{(a)}, L_t^{(b)}, C_t^{(a)}, C_t^{(b)})$$

where

- $P^{(a)}$ (resp. $P^{(b)}$): upward (resp. downward) price moves;
- $T^{(a)}$ (resp. $T^{(b)}$): market orders at the ask (resp. at the bid) that do not move the price;
- $L^{(a)}$ (resp. $L^{(b)}$): limit orders at the ask (resp. at the bid) that do not move the price;
- $C^{(a)}$ (resp. $C^{(b)}$): cancel orders at the ask (resp. at the bid), that do not move the price.

Data: DAX future contracts between 01/01/2014 and 03/01/2014.

Contribution 3. Causality maps from MHP without parametric modeling



Contribution 3. Causality maps from MHP without parametric modeling

Interpretable results

- Any 2×2 sub-matrix with same kind of inputs (i.e. Prices changes, Trades, Limits or Cancels) is symmetric: ask and bid have symmetric roles;
- Prices are mostly cross-excited: price increase is most likely followed by a price decrease, and conversely;
- Market, limit and cancel orders are strongly self-excited: persistence of order flows, and splitting of meta-orders into sequences of smaller orders.

1 Motivation and examples

2 Hawkes processes

3 Dimension reduction for MHP

4 Random matrix theory

5 Causality maps

6 Accelerating training time

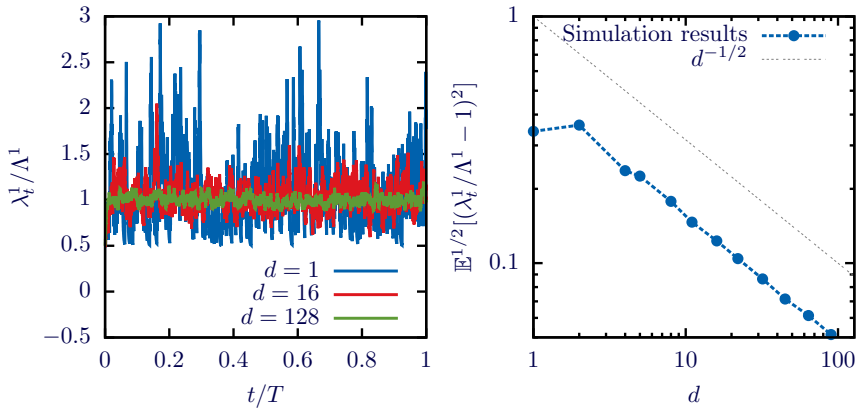
7 Software

Contribution 4. Accelerating training time of MHP

Paper. E. Bacry, S. G., J.-F. Muzy, I. Mastromatteo, *Mean-field inference of Hawkes point processes*, Journal of Physics A, 2016

- Dedicated optimization algorithm for the Hawkes MLE with large number of nodes
- Based on a mean-field approximation
- Partially understood (proof on toy cases)
- Improves state-of-the-art by orders of magnitude

Mean-Field approximation (large number of nodes d helps!)



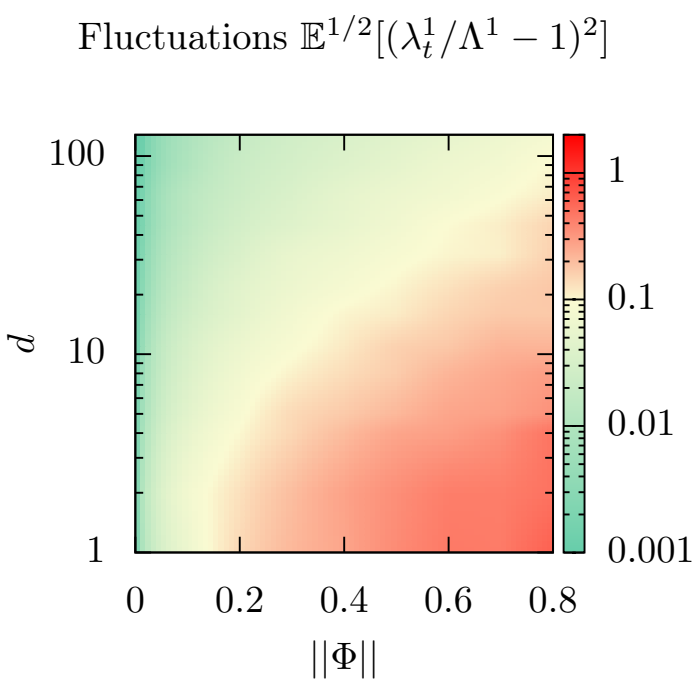
Contribution 4. Accelerating training time of MHP

Use the quadratic approximation

$$\log \lambda_t^i \approx \log \Lambda^i + \frac{\lambda_t^i - \Lambda^i}{\Lambda^i} - \frac{(\lambda_t^i - \Lambda^i)^2}{2(\Lambda^i)^2}$$

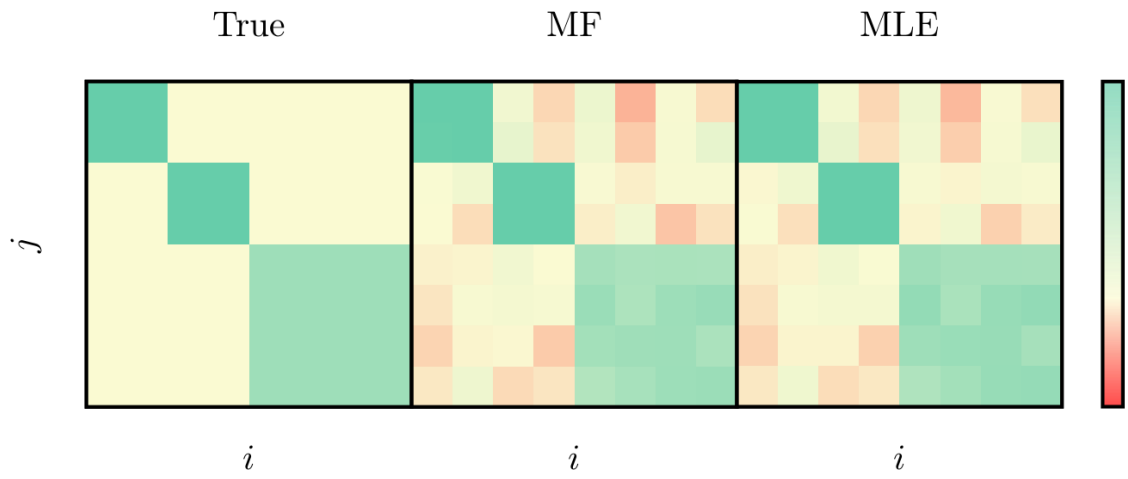
in the log-likelihood

$\Rightarrow$Reduces inference to linear systems



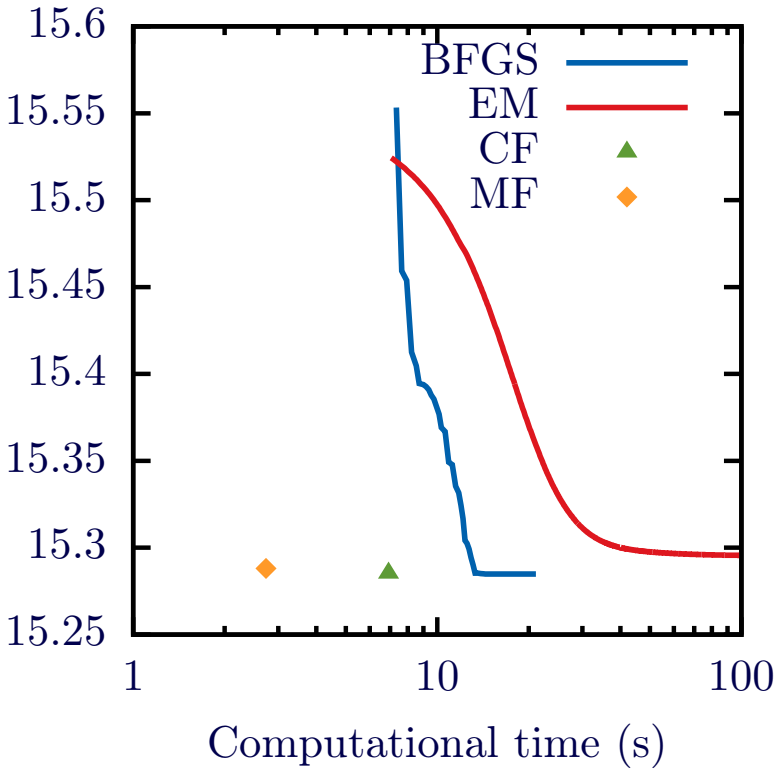
Contribution 4. Accelerating training time of MHP

No clean proof yet (only on toy example) but works very well empirically



Contribution 4. Accelerating training time of MHP

Faster by *several order of magnitude* than state-of-the-art solvers



Conclusion

Take-home message

- Hawkes Process for "time-oriented" machine learning
- Surprisingly relevant to fit real-word phenomena (auto-excitation, user influence)
- Very flexible: intensity can depend on features, other processes, etc.

Main contributions

- Sharp **theoretical guarantees for low-rank inducing penalization** for Hawkes models
- New results about **concentration of matrix-martingales** in continuous time
- Go beyond the parametric approach: **unveil causality using integrated cumulants** matching
- **Improved training time** of the Hawkes model using a "mean-field" approximation

1 Motivation and examples

2 Hawkes processes

3 Dimension reduction for MHP

4 Random matrix theory

5 Causality maps

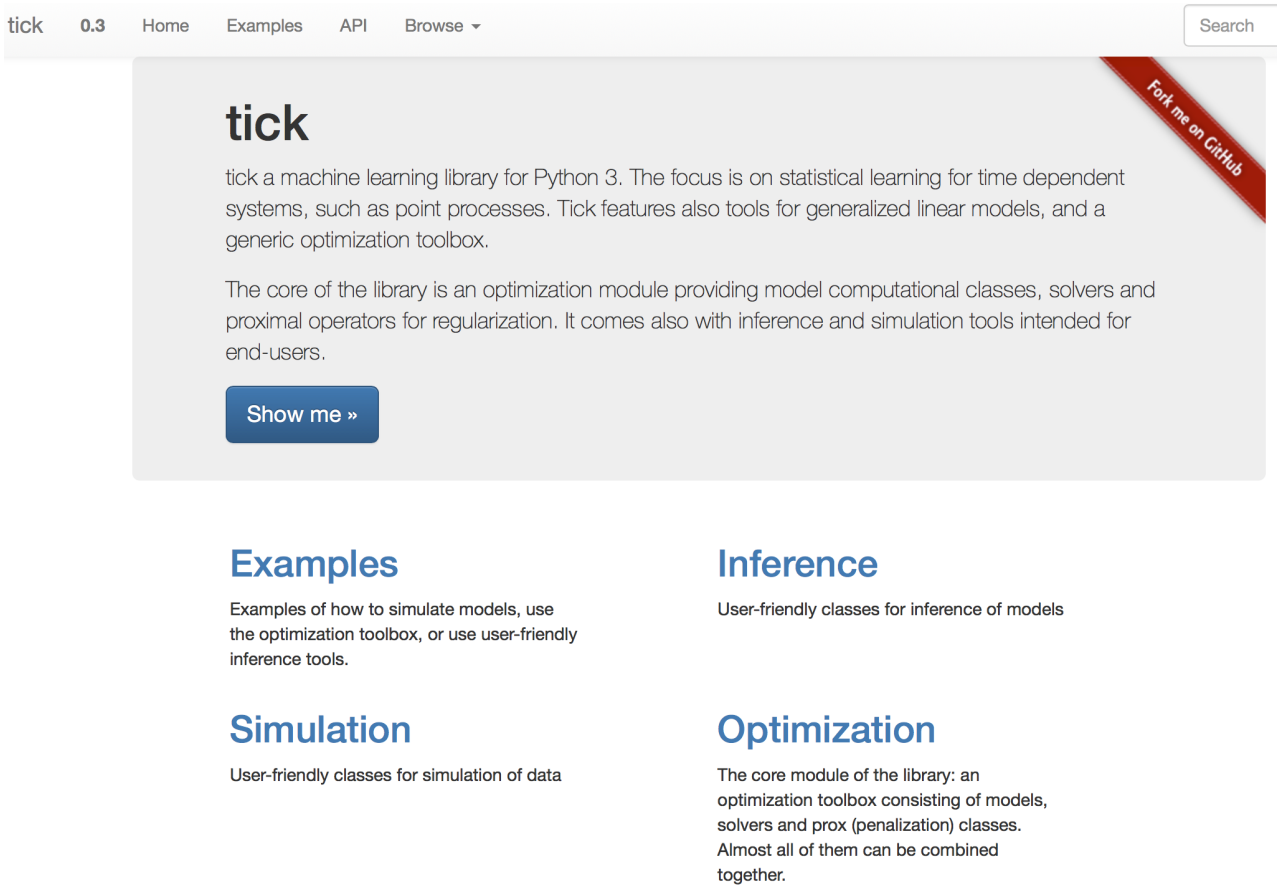
6 Accelerating training time

7 Software

Software: `tick` library

- Python 3 et C++11
- Open-source (BSD-3 License)
- `pip install tick` (on MacOS and Linux...)
- `https://x-datainitiative.github.io/tick`
- Statistical learning for time-dependent models
- Point processes (Poisson, Hawkes), Survival analysis, GLMs (parallelized, sparse, etc.)
- A strong simulation and optimization toolbox
- Partnership with Intel (use-case for new processors with 256 cores)
- Contributors welcome!

Software: `tick` library



Thank you!