Statistical learning with Hawkes processes

Stéphane Gaïffas

ÉCOLE POLYTECHNIQUE
UNIVERSITÉ PARIS-SACLAY

You have users of a system

You want to quantify their level of interactions

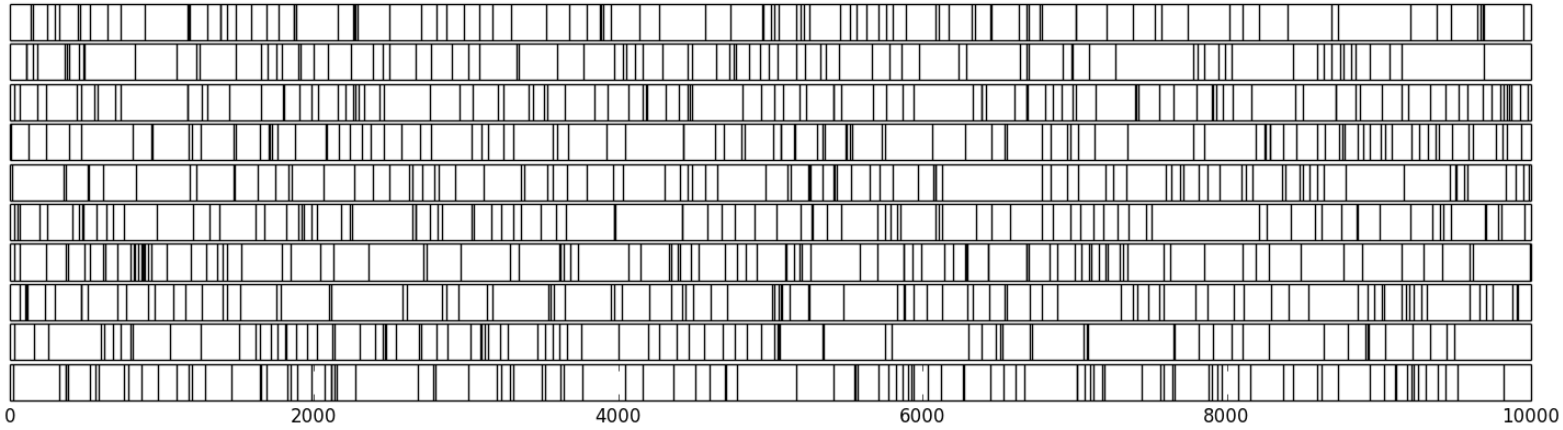
You don't want to use only *declared* interactions: deprecated, not related to the users' activity

You really want levels of interaction driven by user's actions, using their timestamps' patterns

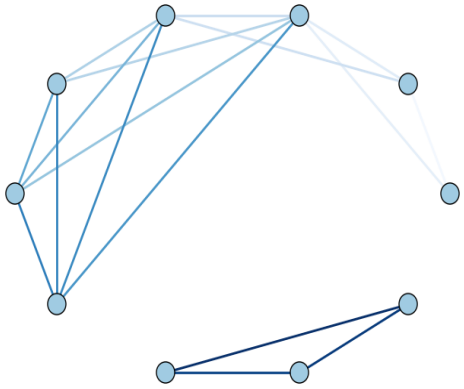
Example 1: Twitter. Timestamps of users' messages. Find something better than the graph given by links of type "user 1 follows user 2"

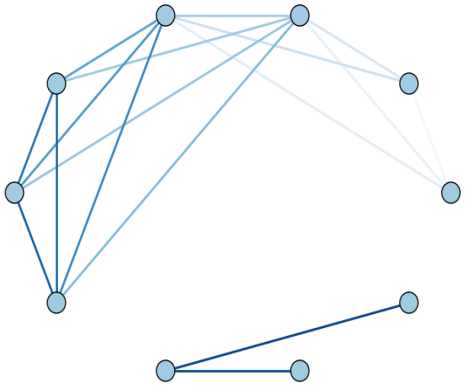
Example 2: MemeTracker. Publications times of articles on websites/blogs, with hyperlinks. Quantify the influence of the publication activity of websites on the others.

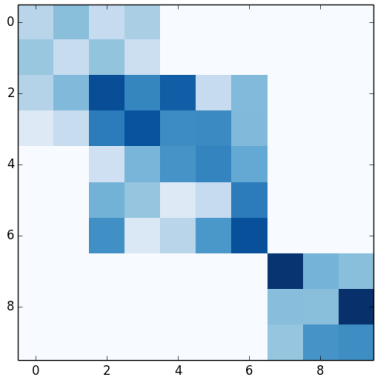
From:



Build:







Data: large number of irregular timestamped events recorded in continuous time

- Activity of users on a social network [DARPA Twitter Bot Challenge 2016, etc.]
- High-frequency variations of signals in finance [Bacry et al. 2013]
- Earthquakes and aftershocks in geophysics [Ogata 1998]
- Crime activity [Mohler 2011 and the `PrePol` startup]
- Genomics, Neurobiology [Reynaud-Bouret et al. 2010]

Methods: in the context of social networks, survival analysis and modeling based on *counting processes* [Gomez et al. 2013, 2015], [Xu et al. 2016]

Setting

- For each node $i \in I = \{1, \ldots, d\}$ 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

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

- $\mu_i \in \mathbb{R}^+$ 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]

Stability condition

- Introduce the matrix with entries

$$\boldsymbol{G}^{ij} = \int_0^{+\infty} \varphi^{ij}(t)dt$$

- Its spectral norm $\|\boldsymbol{G}\|$ must satisfy $\|\boldsymbol{G}\| < 1$ to ensure stability of the process (and stationarity)

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

Estimation for MHP

Parametric estimation (Maximum likelihood)

- First work [Ogata 1978]
- and [Simma and Jordan 2010], [Zhou et al. 2013]
  $\rightarrow$ Expected Maximization (EM) algorithms, with priors

Non parametric estimation

- [Marsan Lengliné 2008], generalized by [Lewis, Mohler 2010]
  $\rightarrow$ EM for penalized likelihood function
  $\rightarrow$ Monovariate Hawkes processes
- [Reynaud-Bouret et al. 2011]
  $\rightarrow$ ℓ_1-penalization over a dictionary
- [Bacry and Muzy 2014]
  $\rightarrow$ Another approach: Weiner-Hopf equations, larger datasets

What for?

- Infer influence and causality directly from *actions* of users
- Exploit the hidden lower-dimensional structure of model parameters for inference/prediction

#Events and dimension d is large. We want:

- a simple parametric model on $\mu = [\mu_i]$ and $\boldsymbol{\varphi} = [\varphi^{ij}]$
- a *tractable* and *scalable* optimization problem
- to encode some prior assumptions using (convex) penalization

Simple parametrization

- Consider
$$\varphi^{ij}(t) = a_{ij} \times \alpha_{ij} e^{-\alpha_{ij} t}$$
- a_{ij} = level of interaction between nodes i and j
- α_{ij} = lifetime of instantaneous excitation of node i by node j

The matrix
$$\boldsymbol{A} = [a_{ij}]_{1 \leq i,j \leq d}$$
is understood has a **weighted adjacency matrix** of mutual influence of nodes $\{1, \ldots, d\}$

- $\boldsymbol{A}$ is non-symmetric: "oriented graph"

A simple parametrization of the MHP

We end up with intensities

$$\lambda^i_{\theta,t} = \mu_i + \int_{(0,t)} \sum_{j=1}^{d} a_{ij}\alpha_{ij} e^{-\alpha_{ij}(t-s)} dN^j_s$$

for $i \in \{1, \ldots, d\}$ where

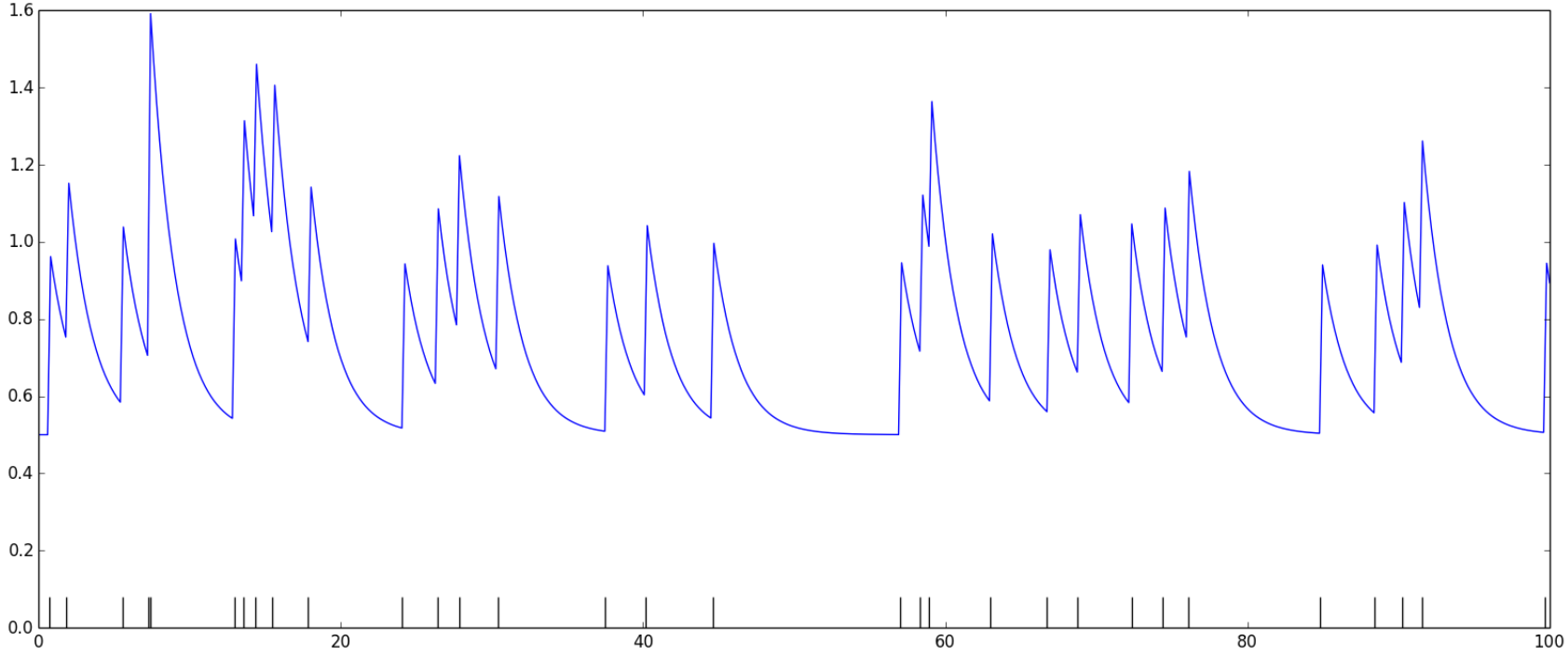
$$\theta = [\mu, \boldsymbol{A}, \boldsymbol{\alpha}]$$

with

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

A simple parametrization of the MHP

For $d = 1$ the intensity $\lambda_{\theta,t}$ looks like this:



Minus log-**likelihood**

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

Least-squares

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

with

$$\lambda^i_{\theta,t} = \mu_i + \sum_{j=1}^{d} a_{ij}\alpha_{ij}\int_{(0,t)} e^{-\alpha_{ij}(t-s)}dN^j_s$$

where $\theta = [\mu, \boldsymbol{A}, \boldsymbol{\alpha}]$ with $\mu = [\mu_i]$, $\boldsymbol{A} = [a_{ij}]$, $\boldsymbol{\alpha} = [\alpha_{ij}]$

A simple framework

Put $\|\lambda_\theta\|_T^2 = \langle \lambda_\theta, \lambda_\theta \rangle_T$ with

$$\langle \lambda_\theta, \lambda_{\theta'} \rangle_T = \frac{1}{T} \sum_{i=1}^{d} \int_{[0,T]} \lambda^i_{\theta,t} \, \lambda^i_{\theta',t} \, dt$$

so that least-squares writes

$$R_T(\theta) = \|\lambda_\theta\|_T^2 - \frac{2}{T} \sum_{i=1}^{d} \int_{[0,T]} \lambda^i_{\theta,t} dN^i_t$$

It is natural: if N has ground truth intensity λ^* then

$$\mathbb{E}[R_T(\theta)] = \mathbb{E}\|\lambda_\theta\|_T^2 - 2\mathbb{E}\langle \lambda_\theta, \lambda^* \rangle_T = \mathbb{E}\|\lambda_\theta - \lambda^*\|_T^2 - \|\lambda^*\|_T,$$

where we used "signal + noise" decomposition (Doob-Meyer):

$$dN^i_t = \lambda^*_t dt + dM^i_t$$

with M^i martingale

A simple framework

A strong assumption: assume that

$$\varphi^{ij}(t) = a_{ij} h_{ij}(t)$$

for **known** h_{ij} meaning that

$$\lambda^i_{\theta,t} = \mu_i + \int_{(0,t)} \sum_{j=1}^{d} a_{ij} h_{ij}(t-s) dN^j_s,$$

where $\theta = [\mu, \boldsymbol{A}]$ with $\mu = [\mu_1, \ldots, \mu_d]^\top$ and $\boldsymbol{A} = [a_{ij}]_{1 \le i,j \le d}$

However

- Most papers using high-dimensional MHP assume $h_{ij}(t) = \alpha e^{-\alpha t}$ for a **known** α!
- e.g. [Yang and Zha 2013], [Zhou et al. 2013], [Farajtabar et al. 2015]
- More on this problem later

Prior assumptions

- Some users are basically inactive and react only if stimulated:

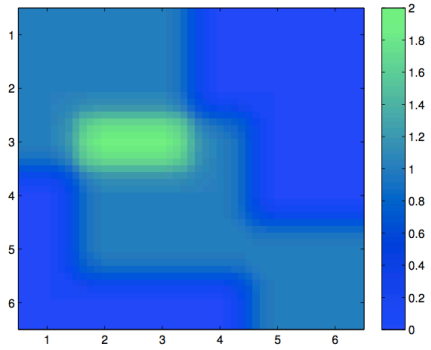
$$\mu \text{ is sparse}$$

- Everybody does not interact with everybody:

$$\boldsymbol{A} \text{ is sparse}$$

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

$$\boldsymbol{A} \text{ is low-rank}$$



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 $\boldsymbol{A}$

So, we use the following penalizations

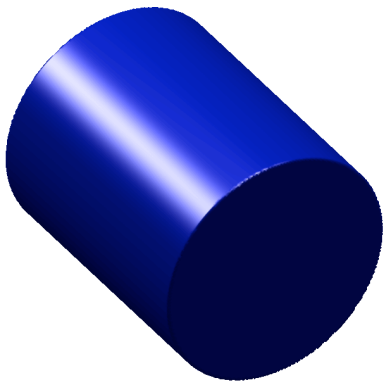
- Use ℓ_1 penalization on μ
- Use ℓ_1 penalization on $\boldsymbol{A}$
- Use trace-norm penalization on $\boldsymbol{A}$

[but other choices might be interesting...]

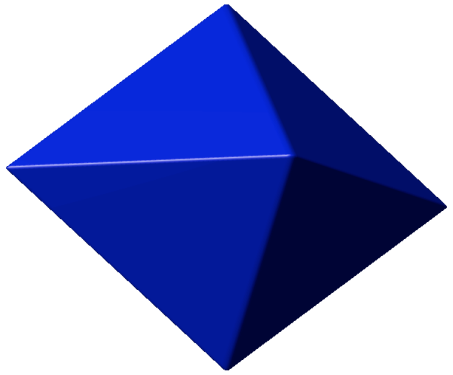
NB1: to induce **sparsity AND low-rank** on $\boldsymbol{A}$, we use the mixed penalization

$$\boldsymbol{A} \mapsto \gamma_* \|\boldsymbol{A}\|_* + \gamma_1 \|\boldsymbol{A}\|_1$$

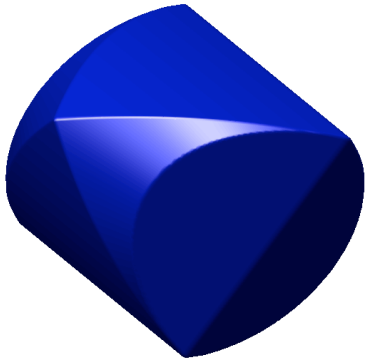
NB2: there exist better ways to induce sparsity and low-rank than this, cf Richard et al (2013) but much harder to minimize



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



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



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

The balls are computed on the set of 2×2 symmetric matrices, which is identified with $\mathbb{R}^3$.

Algorithm

We end up with the problem

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

with mixed penalizations

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

A problem: the "**features scaling**" problem

- Features scaling is necessary for "linear approaches" in supervised learning
- No features and labels here!

$\Rightarrow$ Can be solved here by fine tuning of the penalization terms

Algorithm

Consider instead

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

where this time

$$\operatorname{pen}(\theta) = \|\mu\|_{1,\hat{w}} + \|\boldsymbol{A}\|_{1,\hat{\boldsymbol{W}}} + \hat{w}_*\|\boldsymbol{A}\|_*$$

- Penalization tuned by data-driven weights $\hat{w}$, $\hat{\boldsymbol{W}}$ and $\hat{w}_*$ to solve the "scaling" problem
- Comes from sharp controls of the noise terms, using new probabilistic tools
- Ugly (but computationally easy) formulas

mlpp Home Browse Search

tick

tick is yet another machine learning library for Python 3. The focus is for now on point processes (Poisson processes, Hawkes processes, Cox regression) and Generalized linear models (GLM).

The core of the library is an optimization module providing model computational classes, solvers and proximal operators for regularization. It comes also with inference and simulation tools intended for end-users.

Show me »

Tutorials

Examples of how to simulate models, use the optimization toolbox, or use user-friendly inference tools.

Inference

User-friendly classes for inference of models

Simulation

User-friendly classes for simulation of data

Optimization

The core module of the library: an optimization toolbox consisting of models, solvers and prox (penalization) classes. Almost all of them can be combined together.

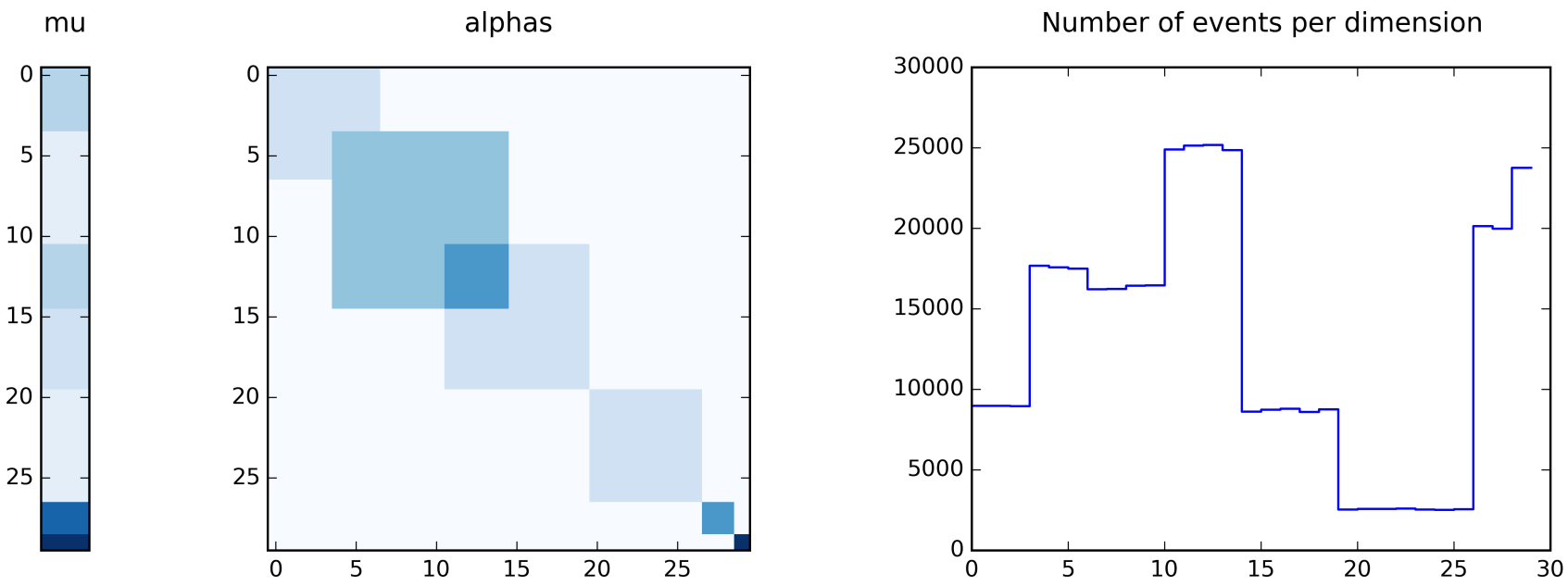
API reference

The full tick's API

Development

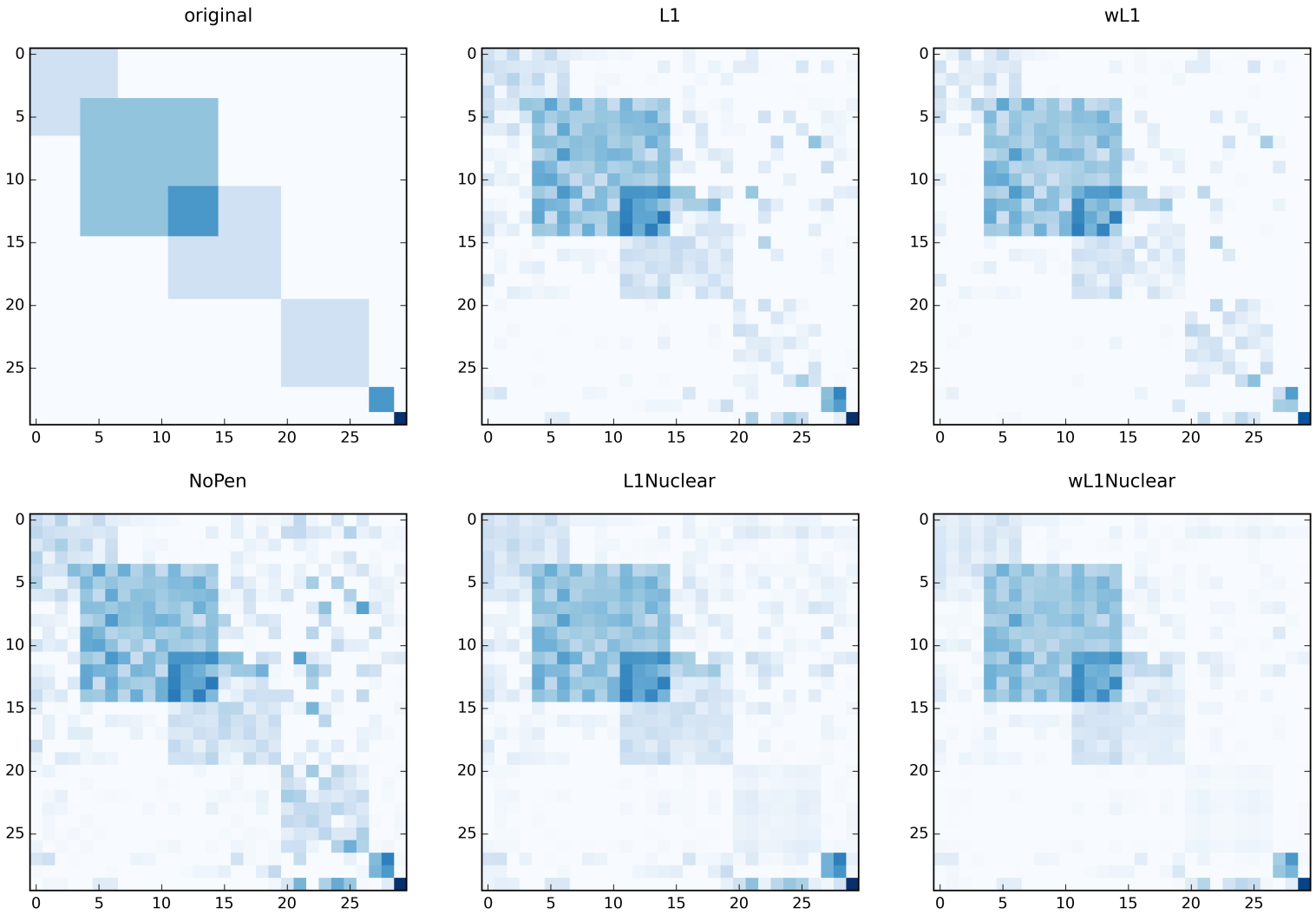
You would like to contribute? Here you will find many tips.

Toy example



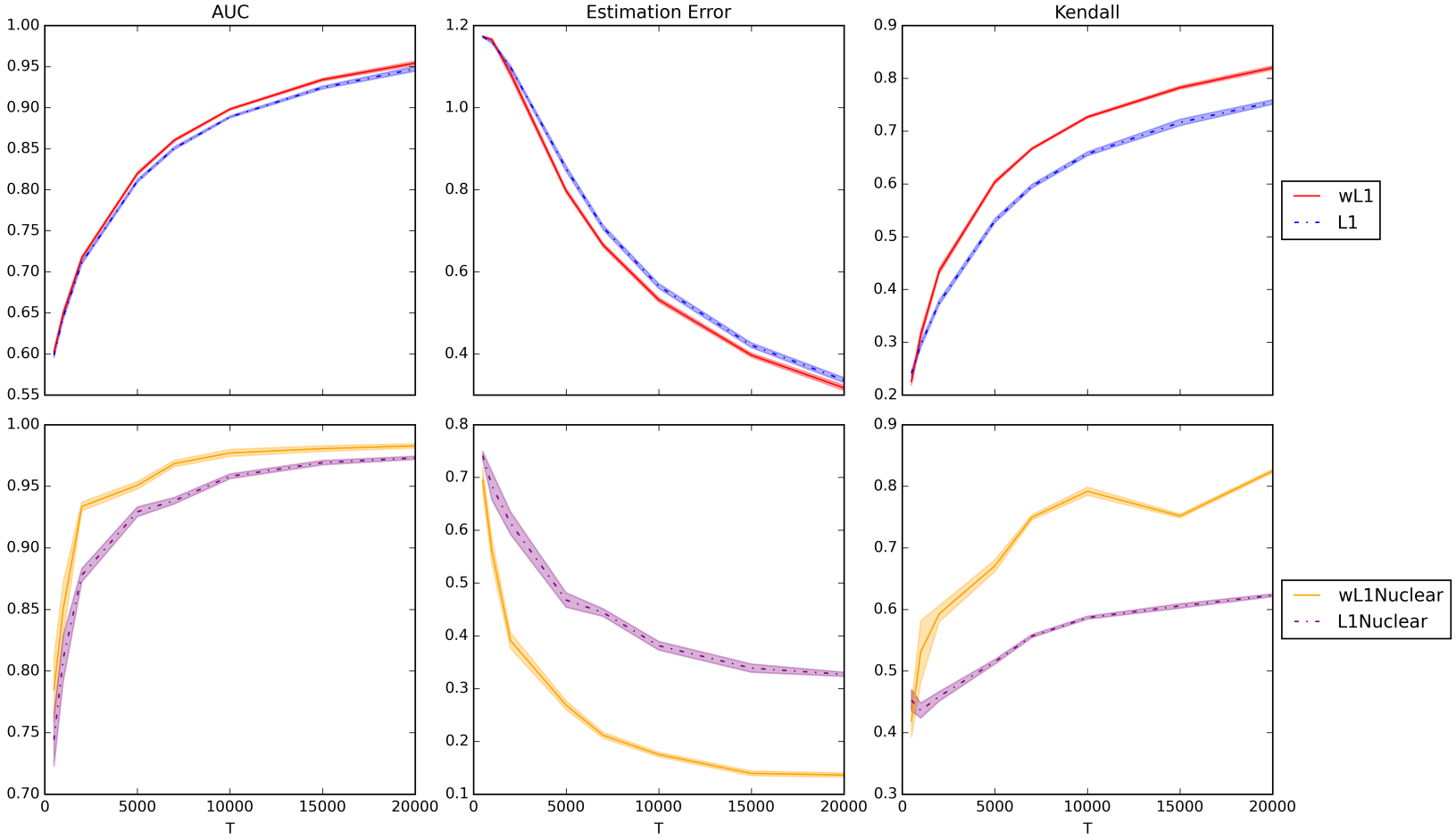
Ground truth parameters μ and $\boldsymbol{A}$, with $d = 30$ and $T = 2000$

Numerical experiment



Ground truth $\boldsymbol{A}$ and instances of recoveries using 6 procedures

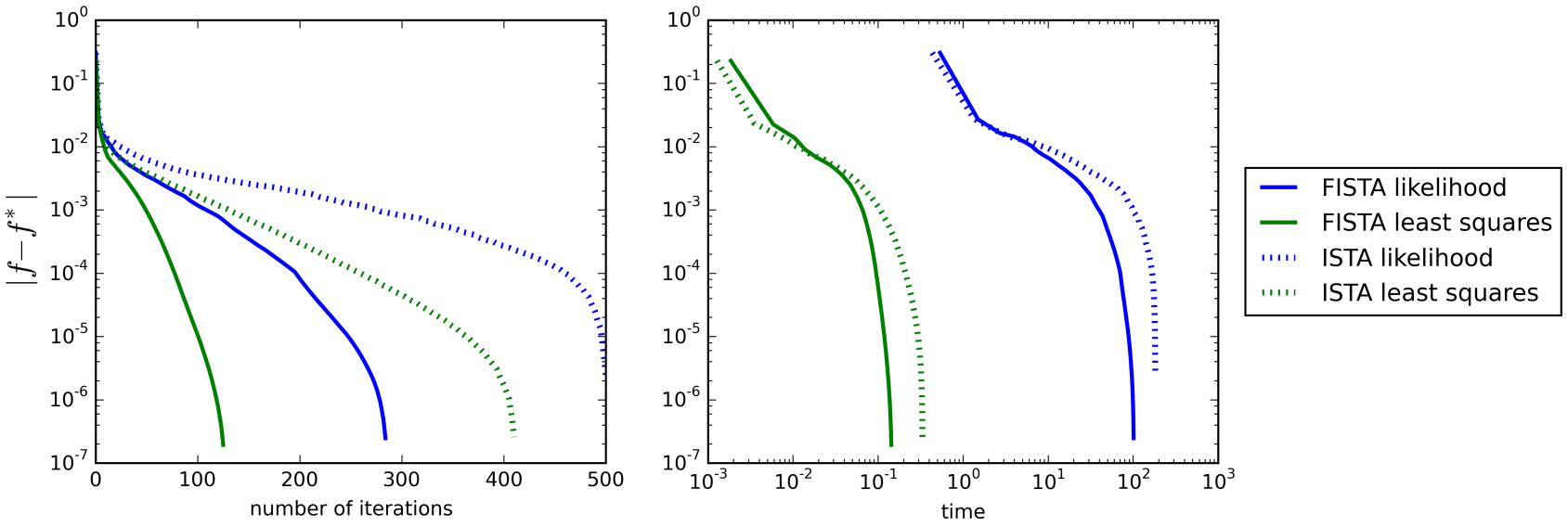
Numerical experiment



averaged AUC, Estimation error and Kandall rank on 100 simulations

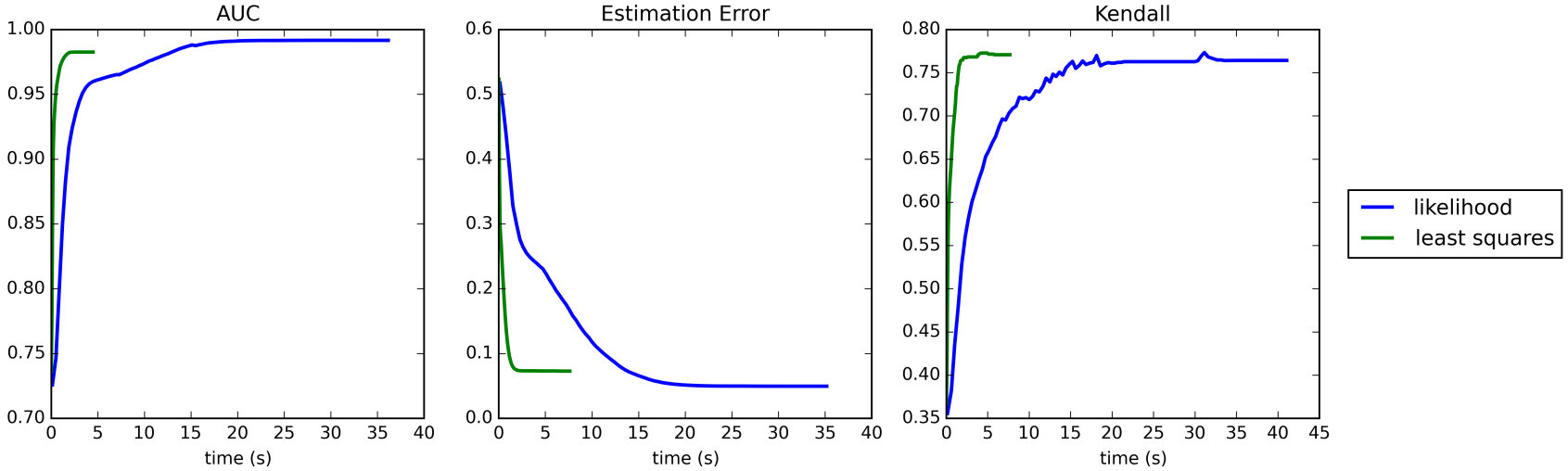
top: non-weighted VS weighted L1
bottom: non-weighted VS weighted L1 + trace norm

Numerical experiment: likelihood VS least-squares



Convergence speed of least squares VS likelihood with (proximal gradient descent with/without acceleration

Numerical experiment: likelihood VS least-squares



Performance achieved by least squares VS likelihood

A sharp oracle inequality

- 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), mandatory assumption to obtain fast rates for convex-relaxation based procedures

Theorem. We have

$$\|\lambda_{\hat{\theta}} - \lambda^*\|_T^2 \leq \inf_{\theta} \Big\{ \|\lambda_\theta - \lambda^*\|_T^2 + \kappa(\theta)^2 \Big(\frac{5}{4} \|(\hat{w})_{\mathrm{supp}(\mu)}\|_2^2 + \frac{9}{8} \|(\hat{\boldsymbol{W}})_{\mathrm{supp}(\boldsymbol{A})}\|_F^2 + \frac{9}{8} \hat{w}_*^2 \,\mathrm{rank}(\boldsymbol{A}) \Big) \Big\}$$

with a probability larger than $1 - 146e^{-x}$.

Theoretical results

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

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

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

Theoretical results

- Data-driven weights come from new "empirical" Bernstein's inequalities, entrywise and for operator norm of the noise $\boldsymbol{Z}_T$ (a matrix martingale)
- Leads to a data-driven scaling of penalization: deals correctly with the inhomogeneity of information over nodes

Noise term is

$$\boldsymbol{Z}_t = \int_0^t \mathrm{diag}[dM_s]\boldsymbol{H}_s,$$

with $\boldsymbol{H}_t$ predictable process with entries

$$(\boldsymbol{H}_t)_{ij} = \int_{(0,t)} h_{ij}(t-s)dN_s^j$$

We need to control $\frac{1}{T}\|\boldsymbol{Z}_T\|_{\mathrm{op}}$

Theoretical results

A consequence of our new concentration inequalities (more after):

$$\mathbb{P}\Bigg[\frac{\|\boldsymbol{Z}_t\|_{\mathrm{op}}}{t} \geq \sqrt{\frac{2v(x+\log(2d))}{t}} + \frac{b(x+\log(2d))}{3t}, \\ b_t \leq b, \quad \lambda_{\max}(\boldsymbol{V}_t) \leq v\Bigg] \leq e^{-x},$$

for any $v, x, b > 0$, where

$$\boldsymbol{V}_t = \frac{1}{t}\int_0^t \|\boldsymbol{H}_s\|_{2,\infty}^2 \begin{bmatrix} \mathrm{diag}[\lambda_s^*] & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{H}_s^\top \mathrm{diag}[\boldsymbol{H}_s\boldsymbol{H}_s^\top]^{-1}\mathrm{diag}[\lambda_s^*]\boldsymbol{H}_s \end{bmatrix} ds$$

and $b_t = \sup_{s\in[0,t]} \|\boldsymbol{H}_s\|_{2,\infty}$ ($\|\cdot\|_{2,\infty}$ = maximum ℓ_2 row norm)

Useless for statistical learning! Event $\lambda_{\max}(\boldsymbol{V}_t) \leq v$ is annoying and $\boldsymbol{V}_t$ is **not observable** (depends on λ^*)!

Theorem [Something better]. For any $x > 0$, we have

$$\frac{\|\boldsymbol{Z}_t\|_{\mathrm{op}}}{t} \leq 8\sqrt{\frac{(x + \log d + \hat{\ell}_{x,t})\lambda_{\max}(\hat{\boldsymbol{V}}_t)}{t}} + \frac{(x + \log d + \hat{\ell}_{x,t})(10.34 + 2.65 b_t)}{t}$$

with a probability larger than $1 - 84.9e^{-x}$, where

$$\hat{\boldsymbol{V}}_t = \frac{1}{t}\int_0^t \|\boldsymbol{H}_s\|_{2,\infty}^2 \begin{bmatrix} \mathrm{diag}[dN_s] & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{H}_s^\top \mathrm{diag}[\boldsymbol{H}_s\boldsymbol{H}_s^\top]^{-1}\mathrm{diag}[dN_s]\boldsymbol{H}_s \end{bmatrix} ds$$

and small ugly term:

$$\hat{\ell}_{x,t} = 4\log\log\left(\frac{2\lambda_{\max}(\hat{\boldsymbol{V}}_t) + 2(4 + b_t^2/3)x}{x} \vee e\right) + 2\log\log\left(b_t^2 \vee e\right).$$

This is a non-commutative deviation inequality with **observable** variance

Main tool: new concentration inequalities for matrix martingales in continuous time

Introduce

$$\boldsymbol{Z}_t = \int_0^t \boldsymbol{A}_s(\boldsymbol{C}_s \odot d\boldsymbol{M}_s)\boldsymbol{B}_s,$$

where $\{\boldsymbol{A}_t\}$, $\{\boldsymbol{C}_t\}$ and $\{\boldsymbol{B}_t\}$ predictable and where $\{\boldsymbol{M}_t\}_{t\geq 0}$ is a "white" matrix martingale, in the sense that $[\mathrm{vec}\boldsymbol{M}]_t$ is diagonal

NB: entries of $\boldsymbol{Z}_t$ are given by

$$(\boldsymbol{Z}_t)_{i,j} = \sum_{k=1}^{p}\sum_{l=1}^{q}\int_0^t (\boldsymbol{A}_s)_{i,k}(\boldsymbol{C}_s)_{k,l}(\boldsymbol{B}_s)_{l,j}(d\boldsymbol{M}_s)_{k,l}.$$

Concentration for purely discountinuous matrix martingale:

- $\boldsymbol{M}_t$ is purely discountinuous and we have

$$\langle \boldsymbol{M} \rangle_t = \int_0^t \boldsymbol{\lambda}_s ds$$

  for a non-negative and predictable intensity process $\{\boldsymbol{\lambda}_t\}_{t \geq 0}$.
- Standard moment assumptions (subexponential tails)

Introduce

$$\boldsymbol{V}_t = \int_0^t \|\boldsymbol{A}_s\|_{\infty,2}^2 \|\boldsymbol{B}_s\|_{2,\infty}^2 \boldsymbol{W}_s ds$$

where

$$\boldsymbol{W}_t = \begin{bmatrix} \boldsymbol{W}_t^1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{W}_t^2 \end{bmatrix}, \tag{1}$$

$$\boldsymbol{W}_t^1 = \boldsymbol{A}_t \operatorname{diag}[\boldsymbol{A}_t^\top \boldsymbol{A}_t]^{-1} \operatorname{diag}\left[(\boldsymbol{C}_t^{\odot 2} \odot \boldsymbol{\lambda}_t)\mathbb{1}\right] \boldsymbol{A}_t^\top$$
$$\boldsymbol{W}_t^2 = \boldsymbol{B}_t^\top \operatorname{diag}[\boldsymbol{B}_t \boldsymbol{B}_t^\top]^{-1} \operatorname{diag}\left[(\boldsymbol{C}_t^{\odot 2} \odot \boldsymbol{\lambda}_t)^\top \mathbb{1}\right] \boldsymbol{B}_t$$

Introduce also

$$b_t = \sup_{s \in [0,t]} \|\boldsymbol{A}_s\|_{\infty,2} \|\boldsymbol{B}_s\|_{2,\infty} \|\boldsymbol{C}_s\|_{\infty}.$$

Theorem.

$$\mathbb{P}\Bigg[\|\boldsymbol{Z}_t\|_{\mathrm{op}} \geq \sqrt{2v(x + \log(m+n))} + \frac{b(x + \log(m+n))}{3}, \\ b_t \leq b, \quad \lambda_{\max}(\boldsymbol{V}_t) \leq v\Bigg] \leq e^{-x},$$

- First result of this type for matrix-martingale in continuous time

Corollary. $\{\boldsymbol{N}_t\}$ a $p \times q$ matrix, each $(\boldsymbol{N}_t)_{i,j}$ is an independent inhomogeneous Poisson processes with intensity $(\boldsymbol{\lambda}_t)_{i,j}$. Consider the martingale $\boldsymbol{M}_t = \boldsymbol{N}_t - \boldsymbol{\Lambda}_t$, where $\boldsymbol{\Lambda}_t = \int_0^t \boldsymbol{\lambda}_s ds$ and let $\{\boldsymbol{C}_t\}$ be deterministic and bounded. We have

$$\begin{aligned}
&\Big\| \int_0^t \boldsymbol{C}_s \odot d(\boldsymbol{N}_t - \boldsymbol{\Lambda}_t) \Big\|_{\mathrm{op}} \\
&\quad \leq \sqrt{2\Big(\Big\| \int_0^t \boldsymbol{C}_s^{\odot 2} \odot \boldsymbol{\lambda}_s ds \Big\|_{1,\infty} \vee \Big\| \int_0^t \boldsymbol{C}_s^{\odot 2} \odot \boldsymbol{\lambda}_s ds \Big\|_{\infty,1} \Big)(x + \log(p+q))} \\
&\qquad + \frac{\sup_{s\in[0,t]} \|\boldsymbol{C}_s\|_\infty (x + \log(p+q))}{3}
\end{aligned}$$

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

Corollary. Even more particualar: $\boldsymbol{N}$ random matrix where $\boldsymbol{N}_{i,j}$ are independent Poisson variables with intensity $\boldsymbol{\lambda}_{i,j}$. We have

$$\|\boldsymbol{N} - \boldsymbol{\lambda}\|_{\mathrm{op}} \leq \sqrt{2(\|\boldsymbol{\lambda}\|_{1,\infty} \vee \|\boldsymbol{\lambda}\|_{\infty,1})(x + \log(p+q))} + \frac{x + \log(p+q)}{3}.$$

- Up to our knowledge, not previously stated in literature
- NB: In the Gaussian case: variance depends on maximum ℓ_2 norm of rows and columns (cf. Tropp (2011))

Some remarks

A non-commutative Hoeffding's inequality when $\boldsymbol{M}_t$ has Brownian motion entries (allowing Itô's formula...), with a similar variance term

A mix of tools from stochastic calculus and random matrix theory

A family of results leading to generalization to continuous-time martingale of matrix deviation inequalities (papers by J. Trop et al.)

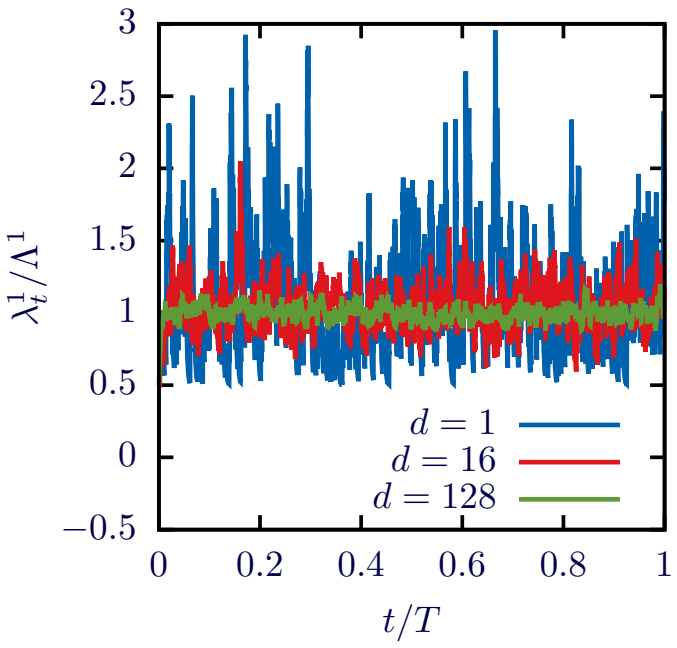
[For experts: as a by-product have give a proof in the discrete-time case that does not require Lieb's convavity Theorem about the trace exponential]

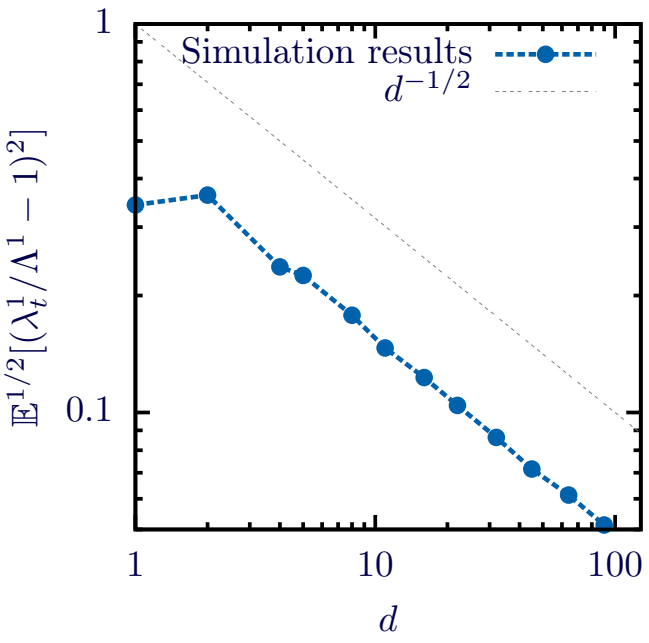
Mean-field inference for Hawkes

Going back to maximum-likelihood estimation, with d very large

- For inference, exploit the fact that d is large

$\Rightarrow$ use a Mean-Field approximation! (from Delattre et al. 2015)





When d is large, we have

$$\lambda_t^i \approx \Lambda^i \quad \text{with} \quad \Lambda_t^i = \mathbb{E}[dN_t^i]/dt$$

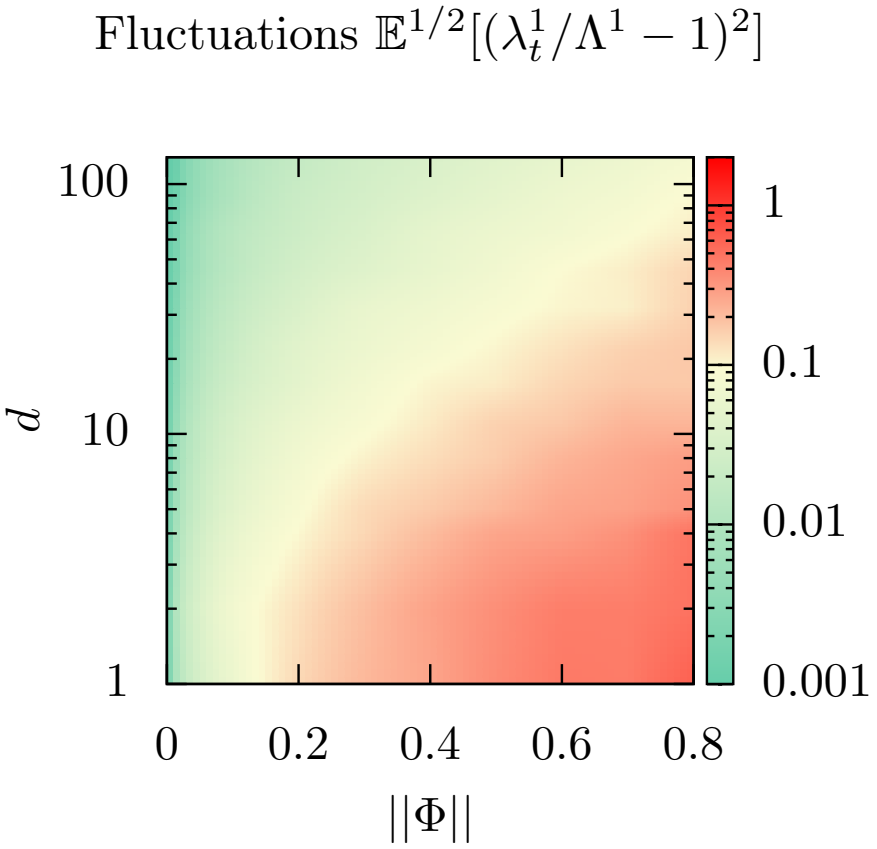
Mean-field inference for Hawkes

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

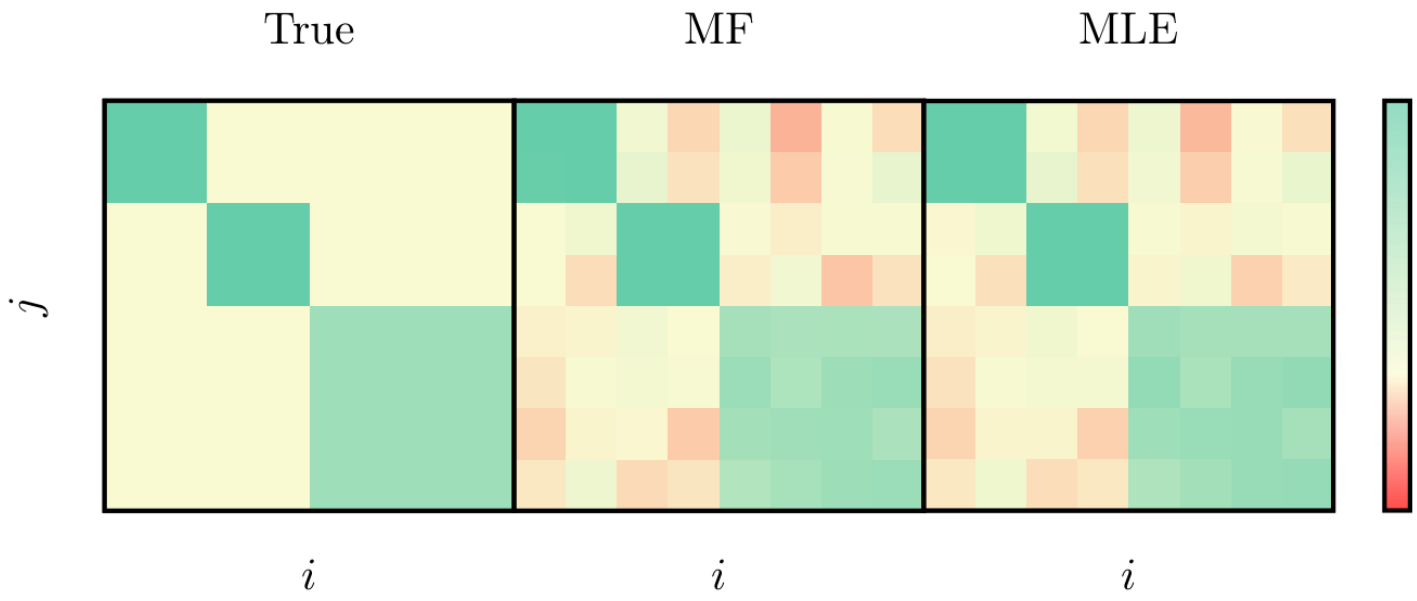
$\Rightarrow$Reduces inference to linear systems



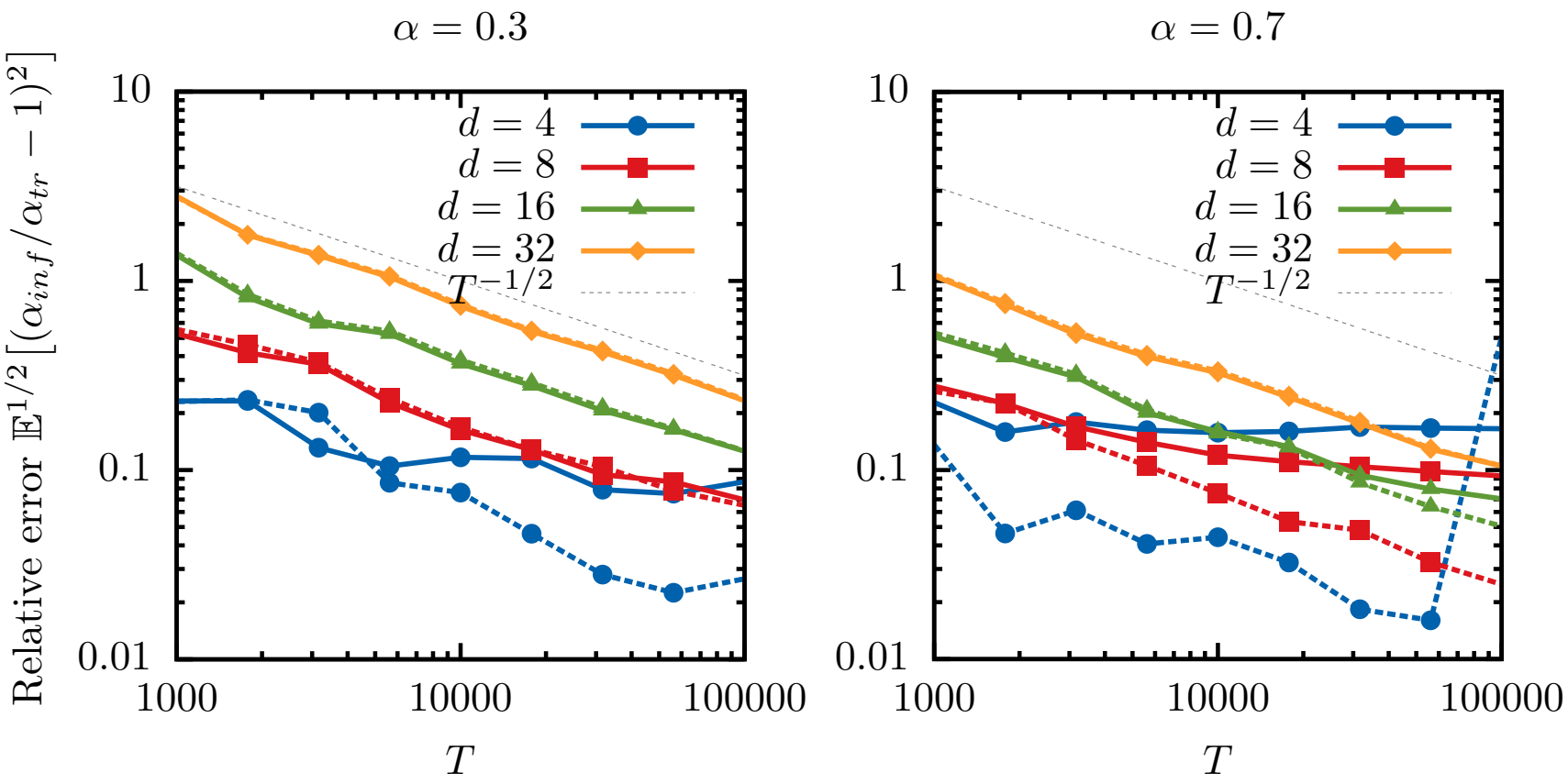
Mean-field inference for Hawkes

No clean proof yet (only on toy example)

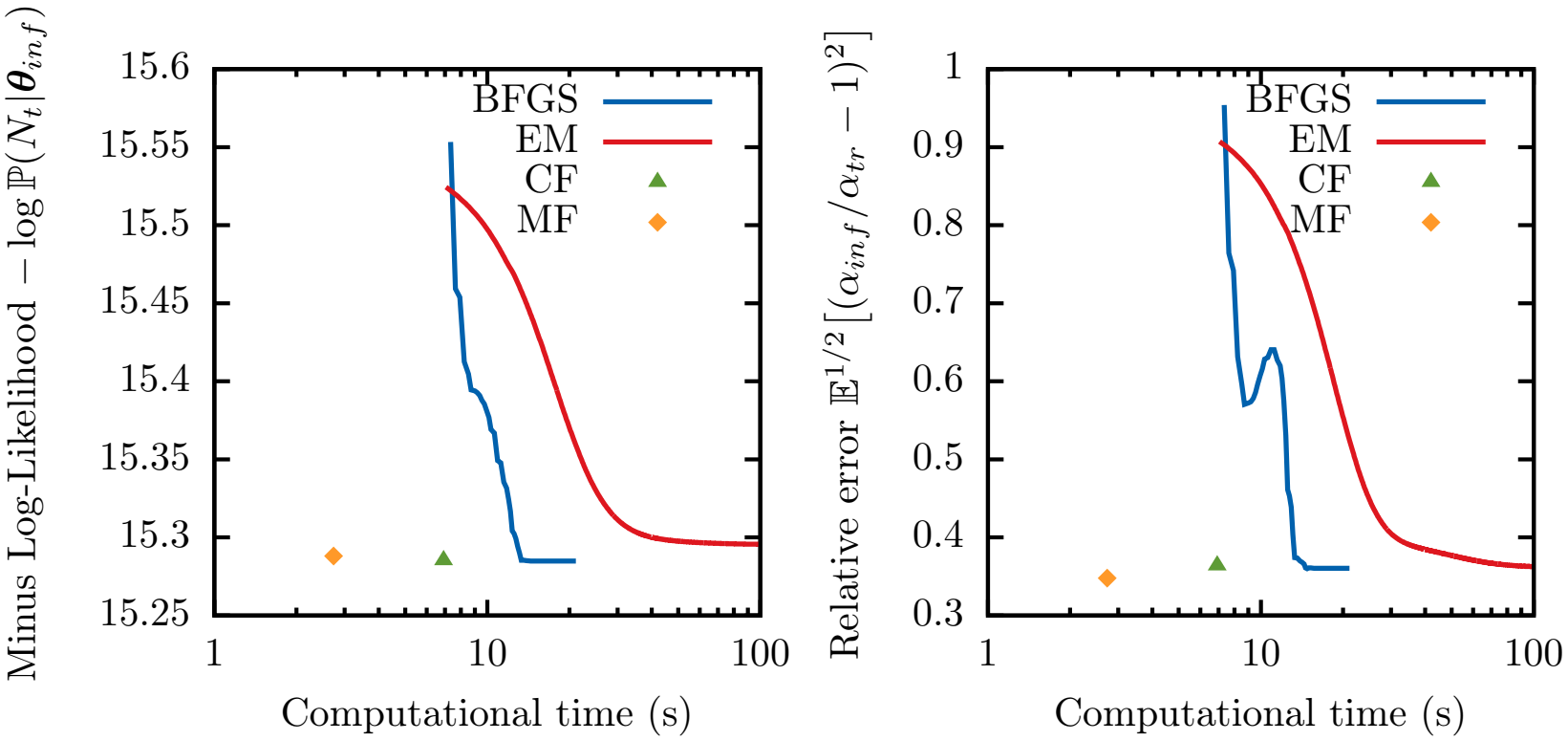
But it works very well empirically



Mean-field inference for Hawkes



It is faster by *several order of magnitude* than state-of-the-art solvers



Some thoughts

- Our original motivation for MHP is influence and causality recovery of nodes
- Knowledge of the full parametrization of MHP is of little interest by itself

A reminder

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

Idea

- Let's not estimate the kernels φ^{ij}, but their *integrals* only!
- Nonparametric approach, no structure imposed on the kernels φ^{ij}
- Let's not use a dictionary either (over-parametrization)
- A way more direct approach

Cumulants matching for MHP

We want to estimate $\boldsymbol{G} = [g^{ij}]$ where

$$g^{ij} = \int_0^{+\infty} \varphi^{ij}(u)\, du \geq 0 \;\text{ for }\; 1 \leq i, j \leq d$$

Remark

- g^{ij} = average total number of events of node i whose *direct* ancestor is an event of node j
- introducing $N_t^{i \leftarrow j}$ that counts the number of events of i whose direct ancestor is an event of j we can prove that

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

Consequence

- $\boldsymbol{G}$ describes mutual influences between nodes
- We know from Eichler et al (2010) that N_t^j does not Granger cause N_t^i iff $g^{ij} = 0$
- Recall stability condition $\|\boldsymbol{G}\| < 1$, which entails that $\boldsymbol{I} - \boldsymbol{G}$ is invertible

Cumulants matching for MHP

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

- Compute estimates of the third order cumulants of the process
- Find $\boldsymbol{G}$ that matches these empirical cumulants
- Highly non-convex problem: polynomial or order 10 with respect to the entries of $(\boldsymbol{I} - \boldsymbol{G})^{-1}$
- Actually not so hard, local minima turns out to be good (deep learning literature)
- Cumulant matching quite powerful for latent topics models, such as Latent Dirichlet Allocation [Bach et al. 2015]

Cumulants matching for MHP

First order-three cumulants can be estimated as

$$\widehat{\Lambda}^i = \frac{1}{T}\sum_{\tau\in Z^i} 1 = \frac{N_T^i}{T}$$

$$\widehat{C}^{ij} = \frac{1}{T}\sum_{\tau\in Z^i}\left(N^j_{\tau+H} - N^j_{\tau-H} - 2H\widehat{\Lambda}^j\right)$$

$$\begin{aligned}\widehat{K}^{ijk} = &\frac{1}{T}\sum_{\tau\in Z^i}\left(N^j_{\tau+H} - N^j_{\tau-H} - 2H\widehat{\Lambda}^j\right)\left(N^k_{\tau+H} - N^k_{\tau-H} - 2H\widehat{\Lambda}^k\right)\\ &-\frac{\widehat{\Lambda}^i}{T}\sum_{\tau\in Z^j}\left(N^k_{\tau+2H} - N^k_{\tau-2H} - 4H\widehat{\Lambda}^k\right)\\ &+2\frac{\widehat{\Lambda}^i}{T}\sum_{\tau\in Z^j}\sum_{\substack{\tau'\in Z^k\\ \tau-2H\le\tau'<\tau}}(\tau-\tau') - 4H^2\widehat{\Lambda}^i\widehat{\Lambda}^j\widehat{\Lambda}^k.\end{aligned}$$

Defining

$$\boldsymbol{R} = (\boldsymbol{I} - \boldsymbol{G})^{-1}$$

we can make a link between the cumulants and $\boldsymbol{G}$

$$\begin{aligned}
\Lambda^i dt &= \mathbb{E}(dN_t^i) \\
C^{ij} dt &= \int_{\tau \in (-\infty, +\infty)} \Big(\mathbb{E}(dN_t^i dN_{t+\tau}^j) - \mathbb{E}(dN_t^i)\mathbb{E}(dN_{t+\tau}^j) \Big) \\
K^{ijk} dt &= \iint_{\tau, \tau' \in (-\infty, +\infty)} \Big(\mathbb{E}(dN_t^i dN_{t+\tau}^j dN_{t+\tau'}^k) \\
&\quad + 2\mathbb{E}(dN_t^i)\mathbb{E}(dN_{t+\tau}^j)\mathbb{E}(dN_{t+\tau'}^k) \\
&\qquad - \mathbb{E}(dN_t^i dN_{t+\tau}^j)\mathbb{E}(dN_{t+\tau'}^k) - \mathbb{E}(dN_t^i dN_{t+\tau'}^k)\mathbb{E}(dN_{t+\tau}^j) \\
&\qquad - \mathbb{E}(dN_{t+\tau}^j dN_{t+\tau'}^k)\mathbb{E}(dN_t^i) \Big),
\end{aligned}$$

and

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

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

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

Why order **three** and not two?

- integrated covariance (order two) contains only symmetric information, and is thus unable to provide causal information
- the skewness of the process breaks the symmetry between past and future so to uniquely fix $\boldsymbol{G}$

Our algorithm [NPHC: Non Parametric Hawkes Cumulant]

- Compute estimators of d^2 third-order cumulant components $\{K^{iij}\}_{1 \leq i,j \leq d}$ (not d^3 !). Put it in $\widehat{\boldsymbol{K^c}}$
- Find
$$\widehat{\boldsymbol{R}} \in \mathrm{argmin}_{\boldsymbol{R}} \|\boldsymbol{K^c}(\boldsymbol{R}) - \widehat{\boldsymbol{K^c}}\|_2^2$$
using a first-order stochastic gradient descent algorithm (AdaGrad in our case)
- Set
$$\widehat{\boldsymbol{G}} = \boldsymbol{I} - \widehat{\boldsymbol{R}}^{-1}$$

Metrics

Relative Error

$$RelErr(\boldsymbol{A}, \boldsymbol{B}) = \frac{1}{d^2} \sum_{i,j} \left(|a^{ij} - b^{ij}| / |a^{ij}| \mathbf{1}_{a^{ij} \neq 0} + |b^{ij}| \mathbf{1}_{a^{ij} = 0} \right)$$

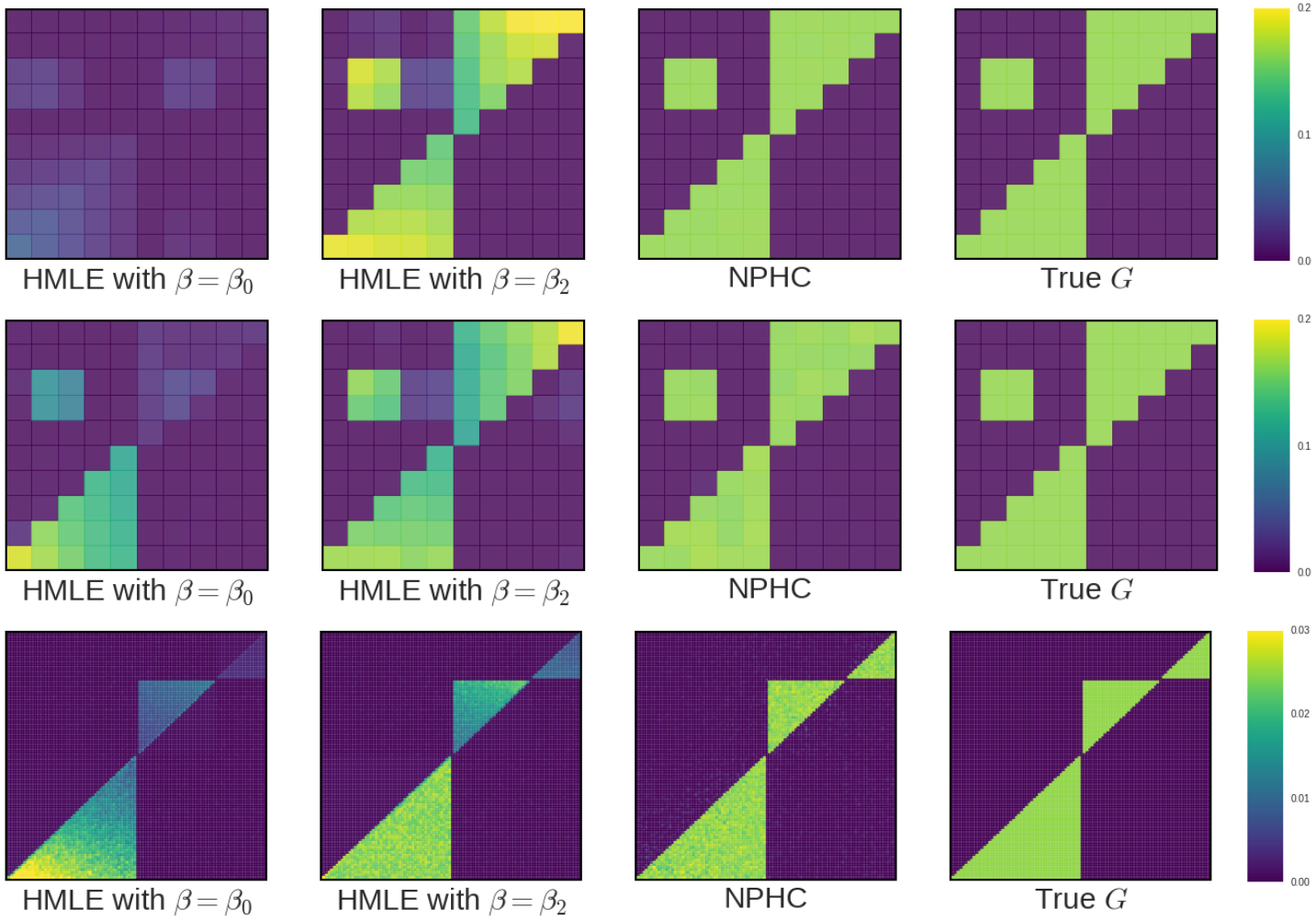
Mean Kendall Rank Correlation

$$MRankCorr(\boldsymbol{A}, \boldsymbol{B}) = \frac{1}{d} \sum_{i=1}^{d} RankCorr([a^{i\bullet}], [b^{i\bullet}]),$$

where

$$RankCorr(x, y) = \frac{1}{d(d-1)/2} (N_{concordant}(x, y) - N_{discordant}(x, y))$$

with $N_{concordant}(x, y) =$ number of pairs (i, j) s.t $x_i > x_j$ and $y_i > y_j$ or $x_i < x_j$ and $y_i < y_j$ and $N_{discordant}(x, y)$ defined conversely



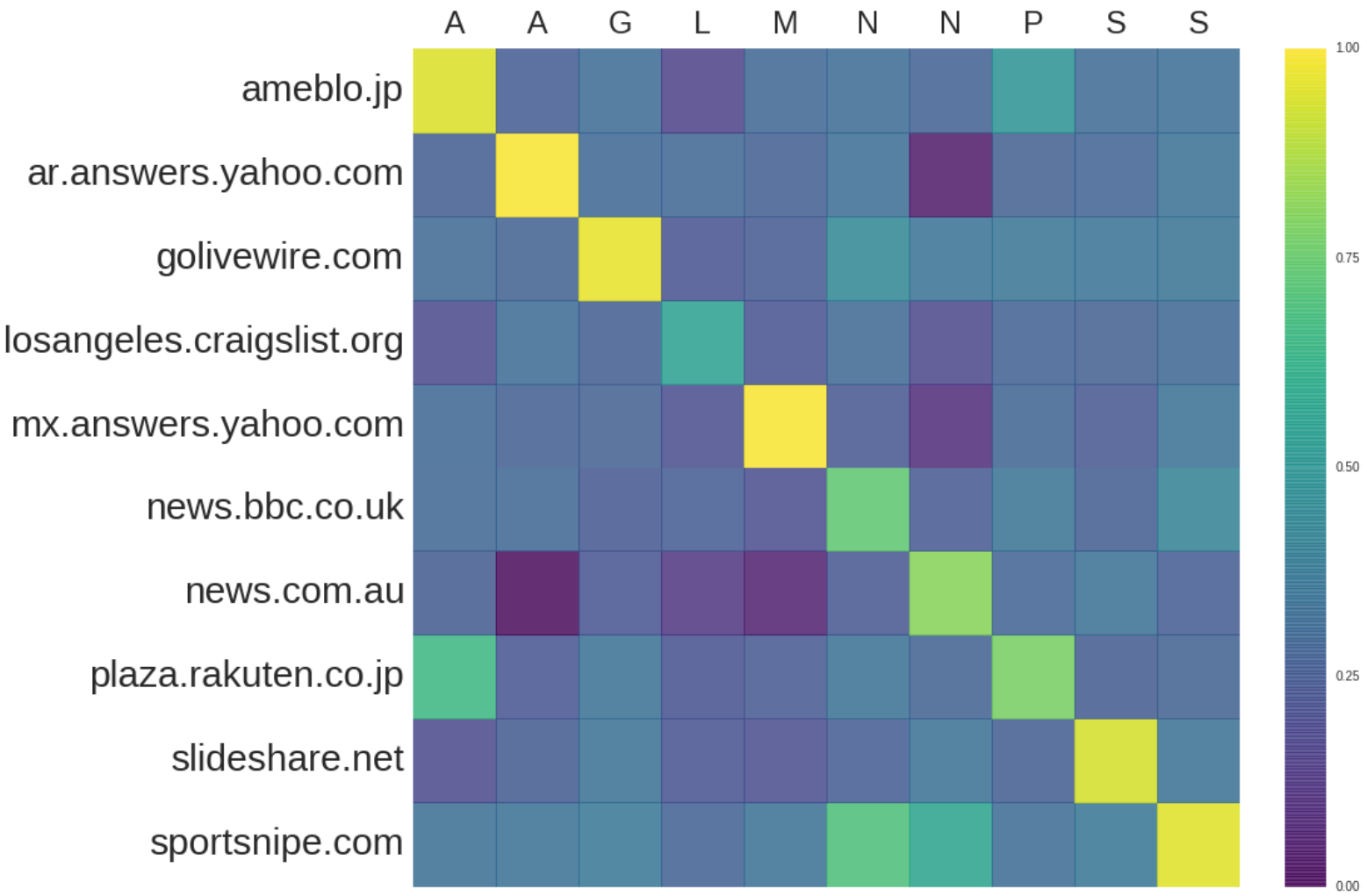
bad param. MLE — best param. MLE — NPHC — ground truth $\boldsymbol{G}$

- *top row* : rectangular kernel ($d = 10$)
- *middle row* : power-low kernel ($d = 10$) (usually very hard...)
- *bottom row* : exponential kernel ($d = 100$)

Experiments with MemeTracker dataset

- keep the 100 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 $\boldsymbol{G}$

NPHC on MemeTracker



$\widehat{G}$

Method	Best HMLE (for RelErr)	Best HMLE (for RankCorr)	NPHC
RelErr	0.153	0.154	**0.064**
MRankCorr	0.035	0.032	**0.175**

Results on the MemeTracker dataset

Conclusion

Take-home message

- Hawkes Process for "time-oriented" machine learning
- Surprisingly relevant to reproduce real-word phenomena (auto-excitation, user influence)

Main contributions

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

Bibliography

- A bound for generalization error for sparse and low-rank multivariate Hawkes processes, with E. Bacry and J-F Bacry [in revision in JMLR]
- Concentration inequalities for matrix martingales in continuous time, with E. Bacry and J-F Bacry [in revision in PTRF]
- Mean-field inference of Hawkes point processes, with E. Bacry, J-F Muzy and I. Mastromatteo [Journal of Physics A]
- Uncovering causality from multivariate Hawkes integrated cumulants, (with M. Achab, E. Bacry, S.G., I. Mastromatteo, J-F Muzy) [submitted to JMLR]

Thank you!