Machine learning and time: “time accounting” learning

Stéphane Gaïffas[1]

14 mars 2014

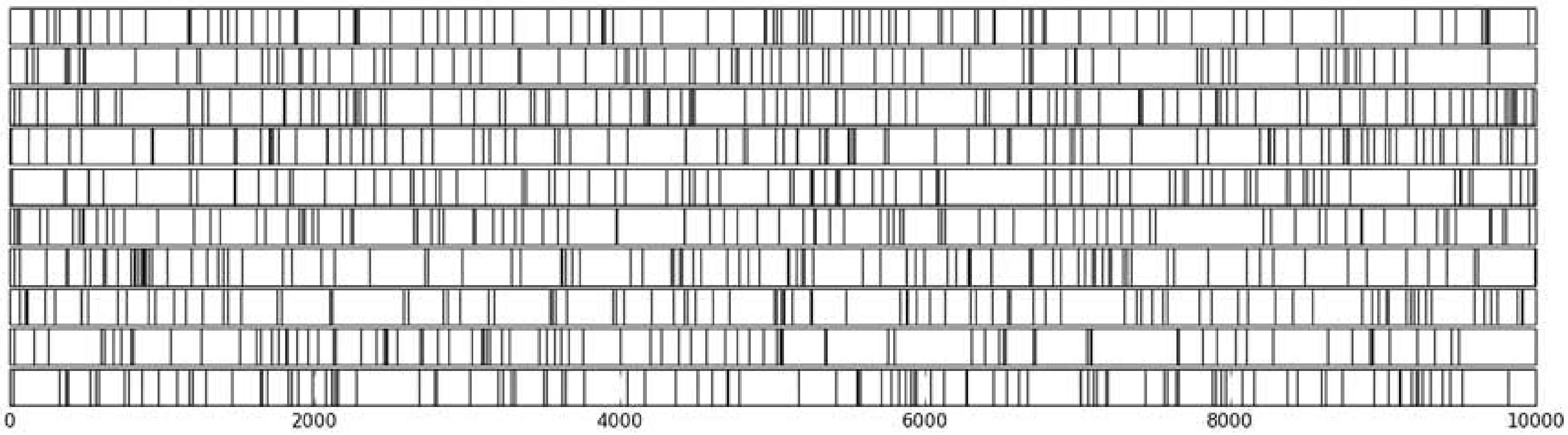
[1]CMAP – Ecole Polytechnique

Introduction

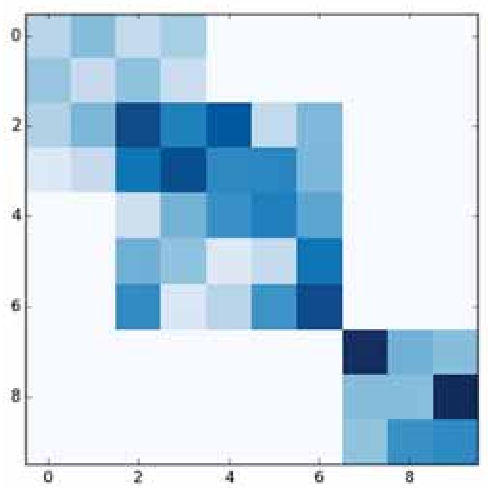
- Finite network with nodes $\{1, \ldots, d\}$: users of a social network, of an e-commerce platform, etc.
- For each node $j \in \{1, \ldots, d\}$ we observe the timestamps $\{t_{j,1}, t_{j,2}, \ldots\}$ of nodes' actions
- Goal: recover **levels of interactions** between users based on the timestamps patterns

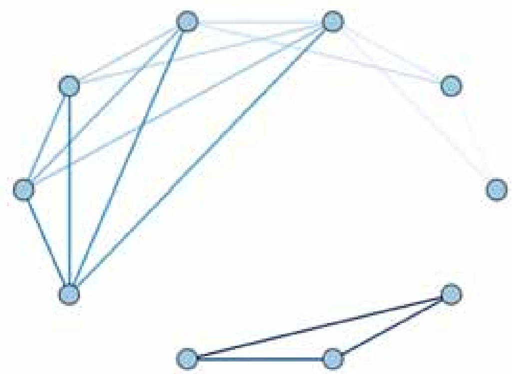
Introduction

From



Quantify interactions between users





Introduction

- Do inference directly from **actions** of users
- Understand the community structure of users underlying the actions
- Exploit the hidden lower-dimensional structure of the network for inference/prediction

Model: Multivariate Hawkes Process (MHP)

- Counting process $N_j(t) = \sum_{i \geq 1} \mathbf{1}_{t_{j,i} \leq t}$
- Data: a d-dimensional counting process $N = [N_1, \ldots, N_d]^\top$
- d is large
- Observed on $[0, T]$. “Asymptotics” in $T \to +\infty$
- N_j has intensity λ_j, namely

$$\begin{aligned}&\mathbb{P}\big(j \text{ does something at time } t \text{ knowning the past}\big)\\&\quad= \mathbb{P}\big(N_j \text{ has a jump in } [t, t+dt] \mid \mathcal{F}_t\big) = \lambda_j(t)dt\end{aligned}$$

for $j = 1, \ldots, d$ where $\mathcal{F}_t$ some filtration

Model: Multivariate Hawkes Process (MHP)

- MHP assumes an autoregressive structure on the intensities:

$$\lambda_j(t) = \mu_j(t) + \int_{(0,t)} \sum_{k=1}^{d} \varphi_{j,k}(t-s)dN_k(s),$$

- $\mu_j(t) \geq 0$ baseline intensity of the j-th coordinate
- $\varphi_{j,k} : \mathbb{R}^+ \to \mathbb{R}^+$ self-exciting component: influence of $k \to j$
- Write this in matrix form

$$\lambda(t) = \boldsymbol{\mu} + \int_{(0,t)} \boldsymbol{\varphi}(t-s)dN(s),$$

with $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_d]^\top$ and $\boldsymbol{\varphi}(t) = [\varphi_{j,k}(t)]_{1 \leq j,k \leq d}$.
- Notation:

$$\int_{(0,t)} \varphi(t-s)dN_j(s) = \sum_{i : t_{j,i} < t} \varphi(t - t_{j,i})$$

A brief history of MHP

Introduced by Hawkes in 1971

- **Earthquakes and geophysics** : Kagan and Knopoff (1981), Zhuang, Harte, Werner, Hainzl and Zhou (2012)
- **Genomics** : Reynaud-Bouret and Schbath (2010)
- **High-frequency Finance** : Bacry Delattre Hoffmann and Muzy (2013)
- **Terrorist activity** : Porter and White (2012)
- **Neurobiology** : Hansen, Reynaud-Bouret and Rivoirard (2012)
- **Social networks** : Carne and Sornette (2008), Simma and Jordan (2010), Zhou Song and Zha (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: some references

Parametric estimation (Maximum likelihood)

- First work : Ogata 78
- Simma and Jordan (2010), Zhou Song and Zha (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, Small amount of data, No theoretical results
- Reynaud-Bouret and Schbath (2010)
  $\rightarrow$ Developed for small amount of data (Sparse penalization)
- Bacry and Muzy (2014)
  $\rightarrow$ Larger amount of data

Dimension d is large:

- Need a simple parametric model on μ and φ
- We want a **convex** optimization problem with smooth loss
- We want to encode some prior assumptions by penalizing this loss

A simple parametrization of the MHP

Simple parametrization:

- Constant baselines $\mu_j(\cdot) \equiv \mu_j$
- Take

$$\varphi_{j,k}(t) = a_{j,k} e^{-\alpha_{j,k} t}$$

- $a_{j,k}$ = level of interaction between nodes j and k
- $\alpha_{j,k}$ = lifetime of instantaneous excitation of node j by node k

The matrix

$$\boldsymbol{A} = [a_{j,k}]_{1 \le j,k \le d}$$

is understood has a **weighted adjacency matrix** of mutual excitement of th nodes $\{1, \ldots, d\}$

- $\boldsymbol{A}$ is non-symmetric

A simple parametrization of the MHP

We end up with intensities

$$\lambda_{j,\theta}(t) = \mu_j + \int_{(0,t)} \sum_{k=1}^{d} a_{j,k} e^{-\alpha_{j,k}(t-s)} dN_k(s)$$

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

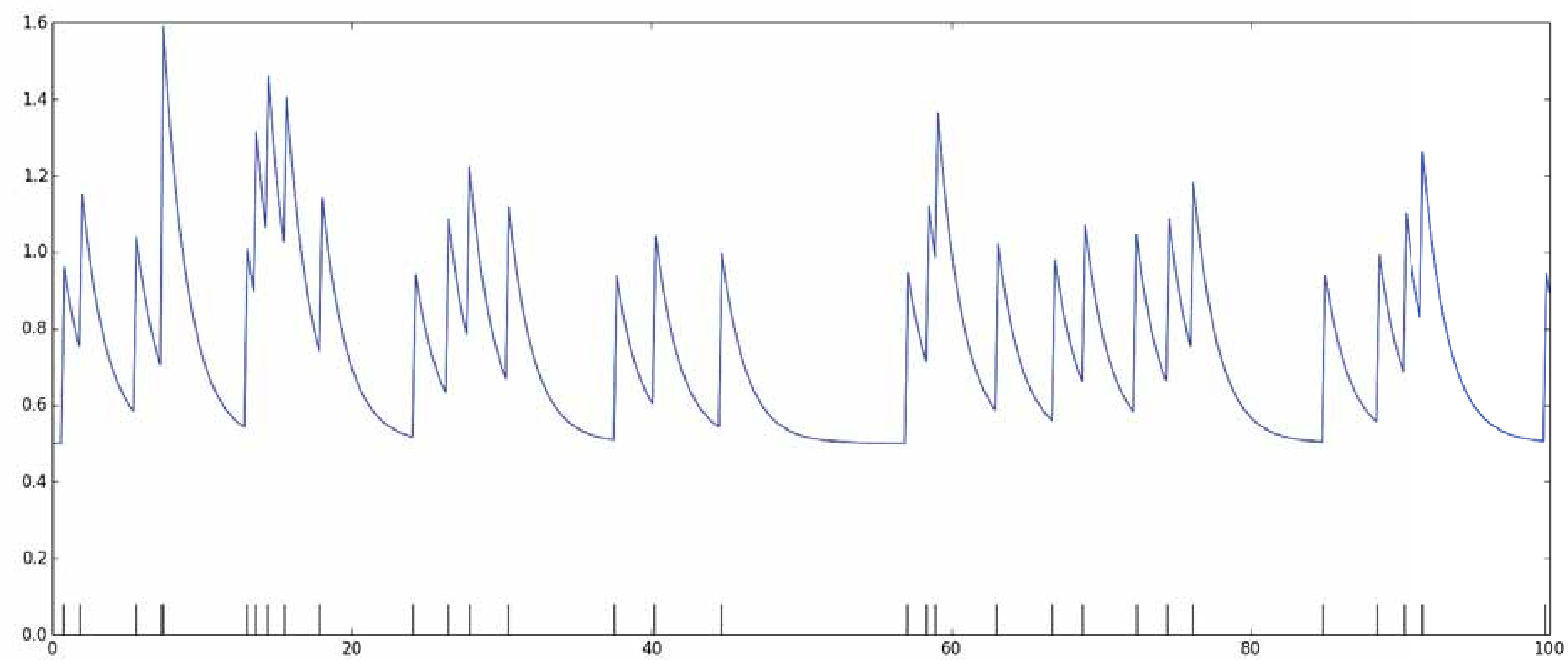
$$\theta = [\mu, \boldsymbol{A}, \boldsymbol{\alpha}]$$

with

- baselines $\mu = [\mu_1, \ldots, \mu_d]^\top \in \mathbb{R}_+^d$
- adjacencies $\boldsymbol{A} = [a_{j,k}]_{1 \le j,k \le d} \in \mathbb{R}_+^{d \times d}$
- decays $\boldsymbol{\alpha} = [\alpha_{j,k}]_{1 \le j,k \le d} \in \mathbb{R}_+^{d \times d}$

A simple parametrization of the MHP

For $d = 1$, intensity λ_θ looks like this:



Goodness-of-fit $= -$ log-likelihood is given by:

$$-\ell_T(\theta) = \sum_{j=1}^{d} \Big\{ \int_0^T (\lambda_{j,\theta}(t) - 1)dt - \int_0^T \log \lambda_{j,\theta}(t) dN_j(t) \Big\}$$

with

$$\lambda_{j,\theta}(t) = \mu_j + \sum_{k=1}^{d} a_{j,k} \int_{(0,t)} \exp\big(- \alpha_{j,k}(t-s)\big) dN_k(s)$$

where $\theta = (\boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{\alpha})$ with $\boldsymbol{\mu} = [\mu_j]$, $\boldsymbol{A} = [A_{j,k}]$, $\boldsymbol{\alpha} = [\alpha_{j,k}]$

Prior assumptions

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

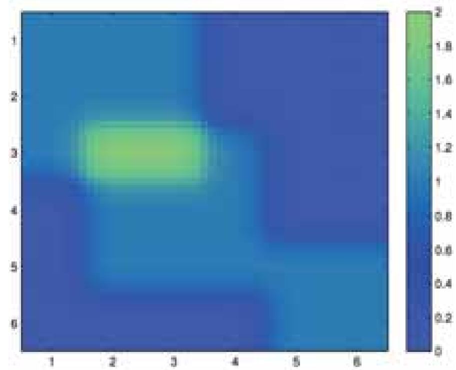
$$\boldsymbol{\mu} \text{ is sparse}$$

- Everybody does not interact with everybody:

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

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

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



- Decays $\boldsymbol{\alpha}$ not sparse, but $\alpha_{j,k}$ should be regularized proportionaly to $a_{j,k}$

Prior encoding by penalization

Standard convex relaxations [Tibshirani 01, ..., Srebro et al. 05, Bach 08, Candès & Recht 08, ...]

- Tightest convex relaxation of $\|\boldsymbol{A}\|_0 = \sum_{j,k} \mathbf{1}_{\boldsymbol{A}_{j,k}>0}$ is ℓ_1-norm:

$$\|\boldsymbol{A}\|_1 = \sum_{j,k} |\boldsymbol{A}_{j,k}|$$

- Tightest convex relaxation of rank is trace-norm:

$$\|A\|_* = \sum_j \sigma_j(A) = \|\sigma(A)\|_1$$

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

Prior encoding by penalization

So, we use the following penalizations

- Use ℓ_1 penalization on $\boldsymbol{\mu}$
- Use ℓ_1 penalization on $\boldsymbol{A}$
- Use trace-norm penalization on $\boldsymbol{A}$
- Use ℓ_2^2 penalization on $\boldsymbol{\alpha}$, weighted by $\boldsymbol{A}$

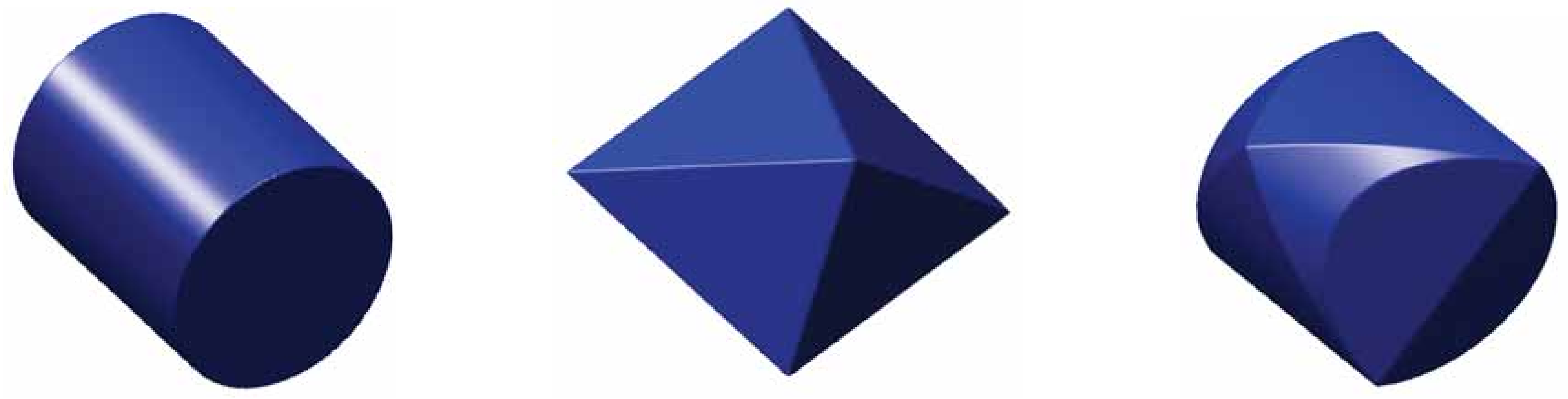
[but other choices might be interesting...]

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

$$\boldsymbol{A} \mapsto w_* \|\boldsymbol{A}\|_* + w_1 \|\boldsymbol{A}\|_1$$

NB2: recent works by Richard et al (2013, 2014): better way to induce sparsity and low-rank than the sum, but not-scalable / non-convex

Sparse and low-rank matrices



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

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

[show video]

Penalized maximum likelihood

Finally, consider

$$\hat{\theta} \in \underset{\theta=(\boldsymbol{\mu},\boldsymbol{A},\boldsymbol{\alpha})}{\operatorname{argmin}} \Big\{ -\frac{1}{T}\ell_T(\theta) + \tau\|\boldsymbol{\mu}\|_1 + \gamma_1\|\boldsymbol{A}\|_1 + \gamma_*\|\boldsymbol{A}\|_* + \frac{\kappa}{2}\|\boldsymbol{A}\odot\boldsymbol{\alpha}\|_F^2 \Big\}$$

where we recall

$$-\frac{1}{T}\ell_T(\theta) = \frac{1}{T}\sum_{j=1}^{d}\Big\{\int_0^T \lambda_{j,\theta}(t)dt - \int_0^T \log\lambda_{j,\theta}(t)dN_j(t)\Big\}$$

with

$$\lambda_{j,\theta}(t) = \mu_j + \sum_{k=1}^{d} a_{j,k}\int_{(0,t)} \exp\big(-\alpha_{j,k}(t-s)\big)dN_k(s)$$

Penalized maximum likelihood: a problem

Problem: $\theta \mapsto \lambda_{j,\theta}(t)$ not convex! Indeed

$$(a, \alpha) \mapsto a h_\alpha(t)$$

never convex when $\alpha \mapsto h_\alpha(t)$ is convex

We **want** convexity for:

- Convergence to a global optimum
- Plethora of optimization algorithms

Generic in the chosen penalization [if proximal operator easy to compute]

Penalized maximum likelihood: reparametrization

A solution: the **perspective function** trick:

- If $\alpha \mapsto h_\alpha(t)$ is convex, then

$$(a, \alpha) \mapsto a h_{\alpha/a}(t)$$

  is **convex**

- Reparametrization $\boldsymbol{\beta} = \boldsymbol{A} \circ \boldsymbol{\alpha}$, leading to

$$\lambda_{j,\theta}(t) = \mu_j + \sum_{k=1}^{d} a_{j,k} \int_{(0,t)} \exp\Big(- \frac{\beta_{j,k}}{a_{j,k}} (t - s) \Big) dN_k(s)$$

  with $\theta = (\boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{\beta})$ for $\boldsymbol{\beta} = [\beta_{j,k}]_{1 \leq j,k \leq d}$

- With this reparametrization

$$\theta \mapsto \lambda_{j,\theta}(t)$$

  is convex

Penalized maximum likelihood: reparametrization

The reparametrization $\boldsymbol{\beta} = \boldsymbol{A} \odot \boldsymbol{\alpha}$ leads to

$$\hat{\theta} \in \operatorname*{argmin}_{\theta = (\boldsymbol{\mu}, \boldsymbol{A}, \boldsymbol{\beta})} \Big\{ -\frac{1}{T}\ell_T(\theta) + \tau \|\boldsymbol{\mu}\|_1 + \gamma_1 \|\boldsymbol{A}\|_1 + \gamma_* \|\boldsymbol{A}\|_* + \frac{\kappa}{2}\|\boldsymbol{\beta}\|_F^2 \Big\} \tag{1}$$

where

$$-\frac{1}{T}\ell_T(\theta) = \frac{1}{T}\sum_{j=1}^{d}\Big\{ \int_0^T \lambda_{j,\theta}(t)dt - \int_0^T \log \lambda_{j,\theta}(t) dN_j(t) \Big\}$$

with

$$\lambda_{j,\theta}(t) = \mu_j + \sum_{k=1}^{d} a_{j,k} \int_{(0,t)} \exp\Big(-\frac{\beta_{j,k}}{a_{j,k}}(t-s)\Big) dN_k(s)$$

Convex optimization – numerical aspects

- Can be solved using first-order routines:

  Fista [Beck Teboulle (2009)], Prisma [Orabona et al (2012)], GFB [Peyre et al. (2011)], Primal-Dual [Chambolle et al. (2009), Condat et al. (2013)], ADMM [Boyd (2012)], etc...

- Gradient of $-\ell_T(\theta)$ using a recursion formula

  $\rightarrow$ Naively $O(n^2 d)$ with $n =$ number of events (very large) but $O(nd)$ when careful (using recursion formulas)

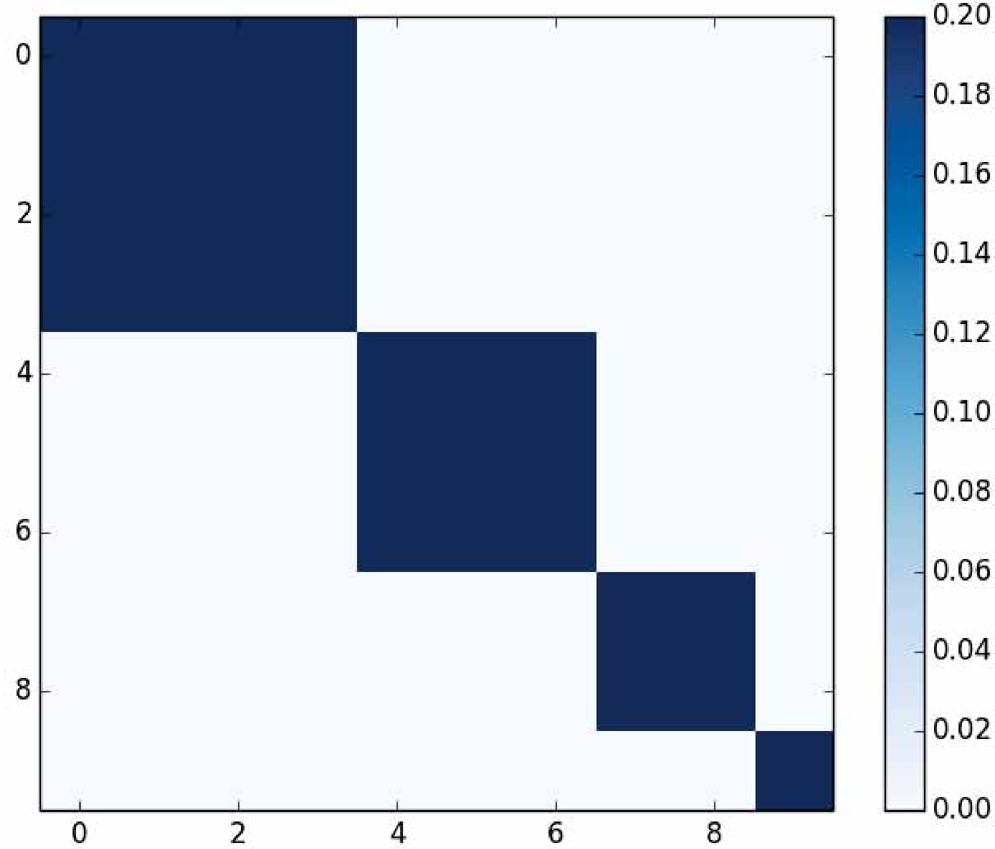
  $\rightarrow$ Parallelized code for this: gradient of each node $j \in \{1, \ldots, d\}$ computed **in parallel**

- Computation bootleneck: exp and log, accelerated using ugly hacking

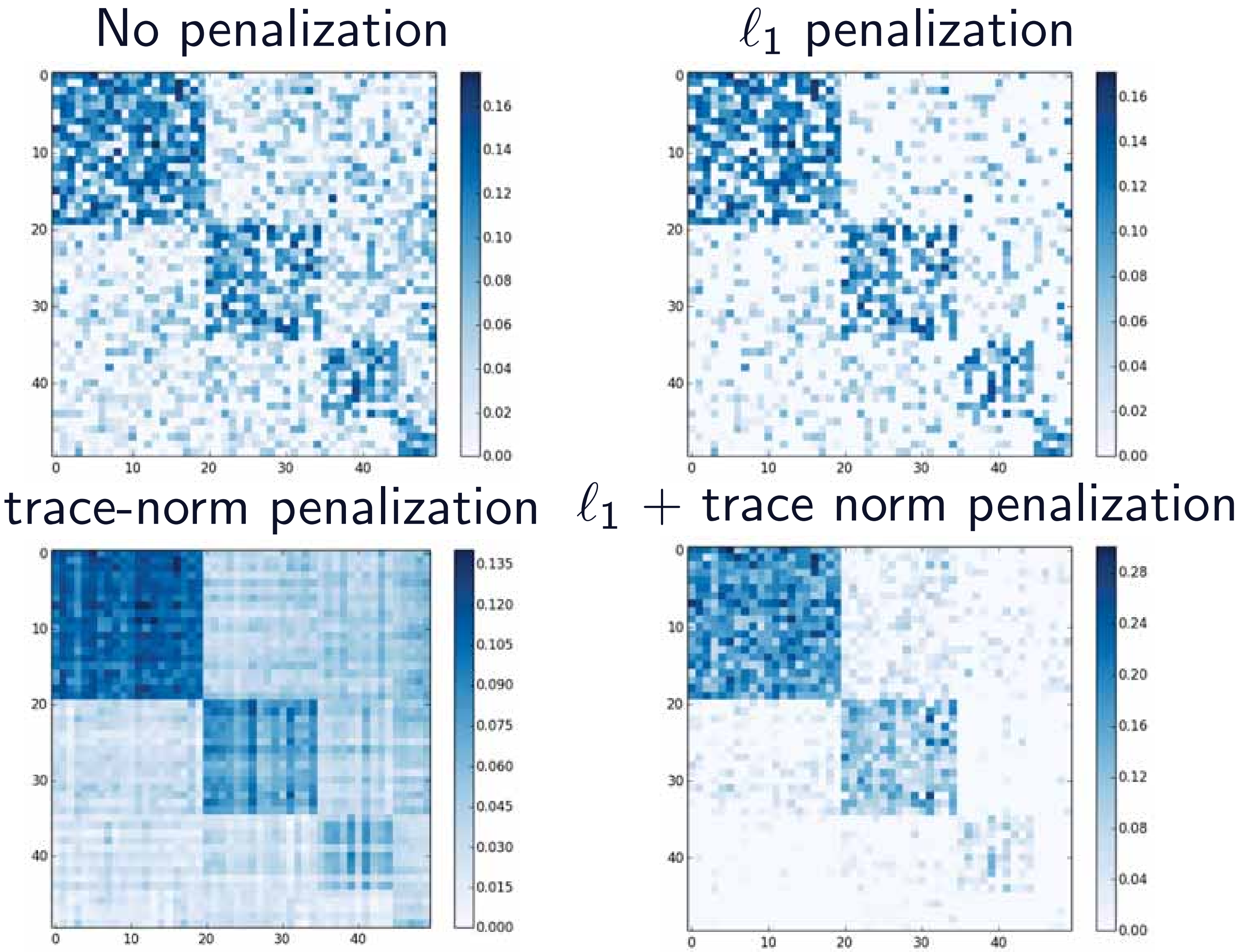
- Trace norm penalization, truncated SVD: default's Lanczos's implementation of Python is fast enough for $d \approx 1K$, use a non-convex factorized formulation $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{V}^\top$ for $d \gg 1K$

Numerical experiment

Toy example: take matrix $\boldsymbol{A}$ as



Numerical experiment: dimension 100, 20100 parameters



We consider a simplified framework

- Fix a set $\{h_{j,k} : 1 \leq j, k \leq d\}$ and intensities

$$\lambda_{j,\theta}(t) = \mu_j + \int_{(0,t)} \sum_{k=1}^{d} a_{j,k} h_{j,k}(t-s) dN_k(s),$$

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

- Instead of $-\log$ likelihood, consider least squares

$$R_T(\theta) = \frac{1}{T} \sum_{j=1}^{d} \Big\{ \int_0^T \lambda_{j,\theta}(t)^2 dt - 2 \int_0^T \lambda_{j,\theta}(t) dN_j(t) \Big\}$$

Some theory: sharp oracle inequalities

Introduce

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

with

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

- Penalization tuned by data-driven weights $\hat{w}$, $\hat{\boldsymbol{W}}$ and $\hat{w}_*$
- Comes from sharp controls of the noise terms
- Solves the **scaling** problem for this model (e.g. feature scaling)

Solving the "feature scaling" problem

ℓ_1-penalization of μ

$$\|\mu\|_{1,\hat{w}} = \sum_{j=1}^{d} \hat{w}_j |\mu_j|$$

with

$$\hat{w}_j \approx \sqrt{\frac{(x + \log d) N_j([0, T])/T}{T}}$$

where $N_j([0, T]) = \#$ events for node j

- Each μ_j penalized by its average events intensity

Solving the "feature scaling" problem

ℓ_1-penalization of $\boldsymbol{A}$

$$\|\boldsymbol{A}\|_{1,\hat{\boldsymbol{W}}} = \sum_{1 \le j,k \le d} \hat{\boldsymbol{W}}_{j,k} |\boldsymbol{A}_{j,k}|$$

with

$$\hat{\boldsymbol{W}}_{j,k} \approx \sqrt{\frac{(x + \log d) \hat{\boldsymbol{V}}_{j,k}(T)}{T}}$$

where

$$\begin{aligned} \hat{\boldsymbol{V}}_{j,k}(t) &= \frac{1}{t} \int_0^t \Big(\int_{(0,s)} h_{j,k}(s-u) dN_k(u) \Big)^2 dN_j(s) \\ &= \text{ variance estimation of the self-excitement for } k \to j \end{aligned}$$

Solving the "feature scaling" problem

Trace-norm penalization of $\boldsymbol{A}$ [difficult]

$$\hat{w}_* \|\boldsymbol{A}\|_* = \hat{w}_* \sum_{j=1}^{d} \sigma_j(\boldsymbol{A})$$

with

$$\hat{w}_* \approx \sqrt{\frac{(x + \log d)(\|\hat{\boldsymbol{V}}_1(T)\|_{\mathrm{op}} \vee \|\hat{\boldsymbol{V}}_2(T)\|_{\mathrm{op}})}{T}}$$

where $\|\cdot\|_{\mathrm{op}}$ = operator norm and where $\hat{\boldsymbol{V}}_1(t)$ diagonal matrix with entries

$$(\hat{\boldsymbol{V}}_1(t))_{j,j} = \frac{1}{t}\int_0^t \|\boldsymbol{H}(s)\|_{2,\infty}^2 dN_j(s),$$

$\hat{\boldsymbol{V}}_2(t)$ matrix with entries

$$(\hat{\boldsymbol{V}}_2(t))_{j,k} = \frac{1}{t}\int_0^t \|\boldsymbol{H}(s)\|_{2,\infty}^2 \sum_{l=1}^{d} \frac{H_{j,l}(s)H_{k,l}(s)}{\|\boldsymbol{H}_{l,\bullet}(s)\|_2^2} dN_l(s),$$

with $\|\cdot\|_{2,\infty}$ = maximum ℓ_2 row norm and $\boldsymbol{H}(t)$ matrix with entries

$$\boldsymbol{H}_{j,k}(t) = \int_{(0,t)} h_{j,k}(t-s)dN_k(s)$$

Solving the "feature scaling" problem

- $\hat{\boldsymbol{V}}_{j,k}(t)$, $\|\hat{\boldsymbol{V}}_1(t)\|_{\mathrm{op}}$ and $\|\hat{\boldsymbol{V}}_2(t)\|_{\mathrm{op}}$ are estimations (based on optional variation) of non-observable variance terms
- It comes from **new Bernstein's concentration inequalities**, used on the noise term
- We develop a new probabilistic tool: **non-commutative concentration inequality for random matrix martingales in continuous time** (theory given by Tropp (2011) applies to discrete time only, and depend on unobserved variance terms)

These tools give a **sharp data-driven tuning** of the penalizations, solving the scaling problem

Some theory: sharp oracle inequalities

Define

$$\langle \lambda_1, \lambda_2 \rangle_T = \frac{1}{T} \sum_{j=1}^{d} \int_0^T \lambda_{1,j}(t) \lambda_{2,j}(t) dt$$

and $\|\lambda\|_T^2 = \langle \lambda, \lambda \rangle_T$

- We use a standard assumption to obtain fast rates for the Lasso: the RE (Restricted Eigenvalue) Assumption [Bickel et al. (2009), Koltchinkii (2011), ...]

A sharp oracle inequality with fast rates

Theorem 1

We have

$$\|\lambda_{\hat{\theta}} - \lambda_0\|_T^2 \leq \inf_{\theta} \Big\{ \|\lambda_\theta - \lambda_0\|_T^2 + \kappa(\theta)^2 \Big(\frac{5}{4} \|(\hat{w})_{\mathrm{supp}(\mu)}\|_2^2 + \frac{9}{8} \|(\hat{\boldsymbol{W}})_{\mathrm{supp}(\boldsymbol{A})}\|_F^2 + \frac{9}{8} \hat{w}_*^2 \operatorname{rank}(\boldsymbol{A}) \Big) \Big\}$$

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

- $\kappa(\theta)$: RE constant
- Sharp: leading constant 1

Some theory: sharp oracle inequalities

Take-home message: $\hat{\theta}$ achieves an optimal tradeoff between approximation and complexity given by

$$\frac{\|\mu\|_0 (x + \log d)}{T} \max_j N_j([0, T])/T + \frac{\|\boldsymbol{A}\|_0 (x + \log d)}{T} \max_{j,k} \hat{\boldsymbol{V}}_{j,k}(T) + \frac{\operatorname{rank}(A)(x + \log d)}{T} \|\hat{\boldsymbol{V}}_1(T)\|_{\mathrm{op}} \vee \|\hat{\boldsymbol{V}}_2(T)\|_{\mathrm{op}}.$$

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

A new concentration inequality

- New Bernstein's empirical concentration inequality for continuous-time matrix martingale
- Consider the random matrix $\boldsymbol{Z}(t)$ with entries

$$\boldsymbol{Z}_{j,k}(t) = \int_0^t \int_{(0,s)} h_{j,k}(s-u)dN_k(u)dM_j(s)$$

where $M_j(t) = N_j(t) - \int_0^t \lambda_j(s)ds$ are martingales obtained by compensation

- This is the noise term in our problem

A new concentration inequality

A classical concentration inequality for $\boldsymbol{Z}_{j,k}$ [Lipster Shiryayev 1986] is

$$\frac{1}{t}(\boldsymbol{Z}(t))_{j,k} \leq \sqrt{\frac{2vx}{t}} + \frac{bx}{3t}$$

for any $x > 0$, with a probability $\geq 1 - e^{-x}$ whenever

$$\frac{1}{t}\langle \boldsymbol{Z}_{j,k} \rangle_t = \frac{1}{t}\int_0^t \Big(\int_{(0,s)} h_{j,k}(s-u)dN_k(u) \Big)^2 \lambda_j(s)ds \leq v$$

and

$$\sup_{s\in[0,t]} \int_{(0,s)} h_{j,k}(s-u)dN_k(u) \leq b$$

A new concentration inequality

- Predictable variation $\langle \boldsymbol{Z}_{j,k} \rangle_t$ depends on non-observed λ_j: this concentration is **useless** for statistics
- Need an **empirical Bernstein's inequality**, with a variance term using the **optional variation**

$$\frac{1}{t}[\boldsymbol{Z}_{j,k}]_t = \frac{1}{t}\int_0^t \Big(\int_{(0,s)} h_{j,k}(s-u)dN_k(u)\Big)^2 dN_j(s)$$

- We need also to remove the event $\{\langle \boldsymbol{Z}_{j,k} \rangle_t \le tv\}$ from this inequality

We provide:

- A control of all the entries $\boldsymbol{Z}_{j,k}$ of $\boldsymbol{Z}$
- A control of $\|\boldsymbol{Z}_t\|_{\mathrm{op}}$

A new concentration inequality

Theorem 2

We have

$$\frac{1}{t}|\boldsymbol{Z}_{j,k}(t)| \leq 2\sqrt{2}\sqrt{\frac{(x + 2\log d + \hat{\boldsymbol{L}}_{j,k}(t))\hat{\boldsymbol{V}}_{j,k}(t)}{t}} + 9.31\frac{(x + 2\log d + \hat{\boldsymbol{L}}_{j,k}(t))\boldsymbol{B}_{j,k}(t)}{t}$$

for any $1 \leq j, k \leq d$, with a probability larger than $1 - 30.55e^{-x}$.

- Based on a previous result by G. and Guilloux (2011), see also Hansen et al (2012)
- Reminiscent of previous works by Audibert (2008)

A new concentration inequality

Theorem 3

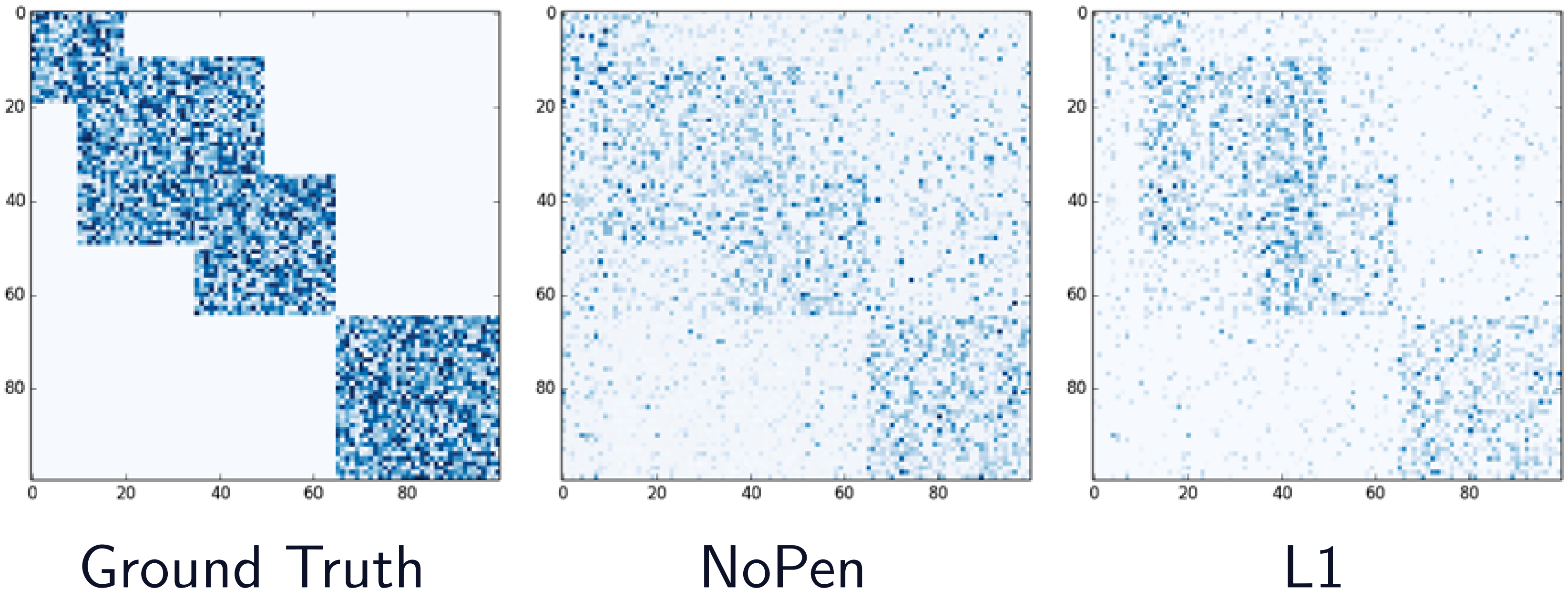
For any $x > 0$, we have

$$\frac{\|\boldsymbol{Z}(t)\|_{\mathrm{op}}}{t} \leq 4\sqrt{\frac{(x + \log d + \hat{\ell}_x(t))\|\hat{\boldsymbol{V}}_1(t)\|_{\mathrm{op}} \vee \|\hat{\boldsymbol{V}}_2(t)\|_{\mathrm{op}}}{t}} + \frac{(x + \log d + \hat{\ell}_x(t))(10.34 + 2.65 \sup_{t\in[0,T]} \|\boldsymbol{H}(t)\|_{2,\infty})}{t}$$

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

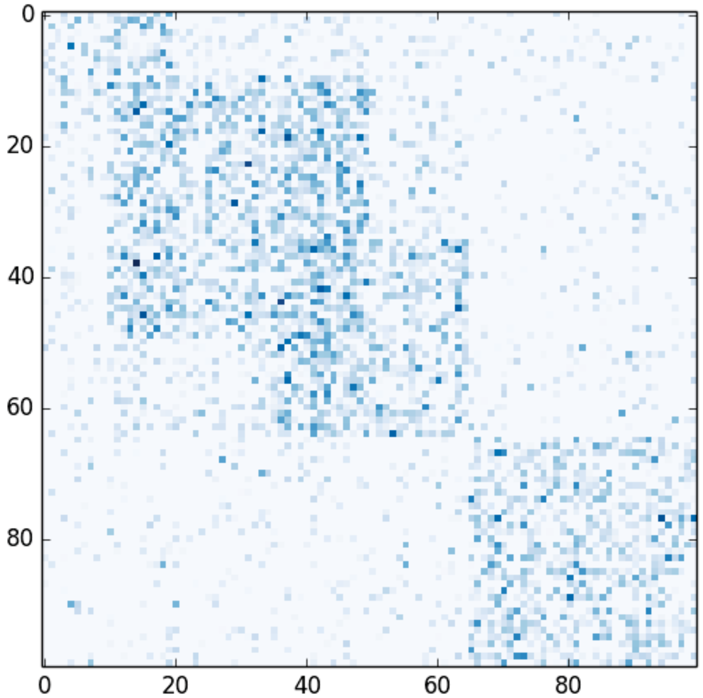
- First non-commutative Bernstein's inequality for countinuous time martingales
- Can be extended to a wider class of martingales
- Extension of [Tropp (2012)] results

Consequence: a sharp scaling of penalizations

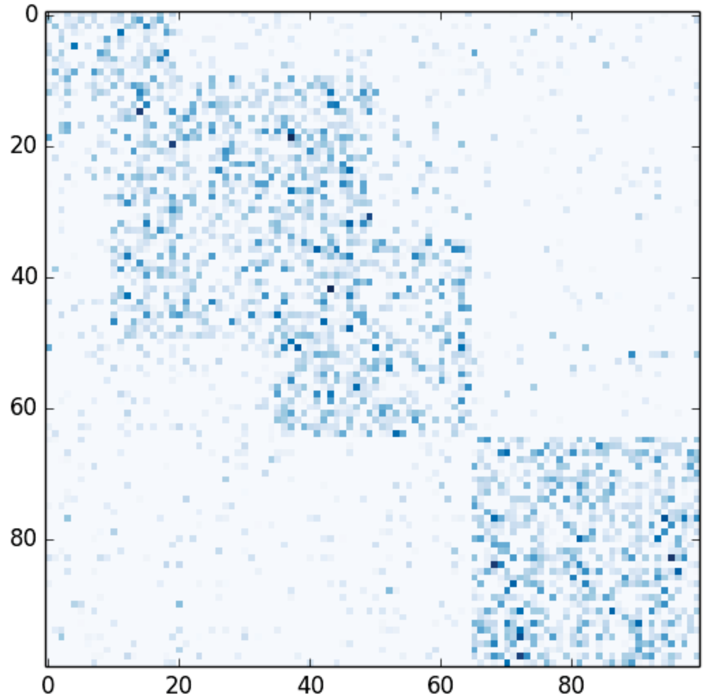


Consequence: a sharp scaling of penalizations

L1 vs wL1



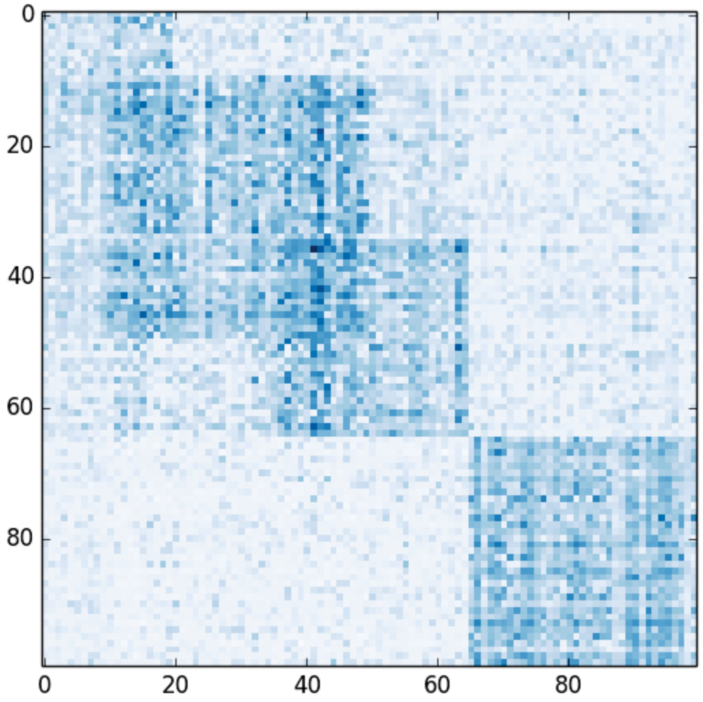
L1



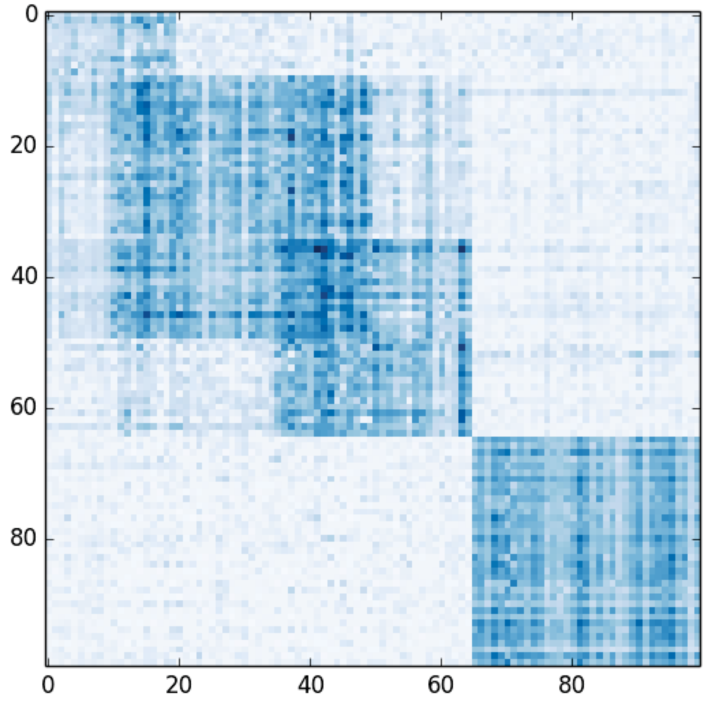
wL1

Consequence: a sharp scaling of penalizations

Nuclear vs wNuclear

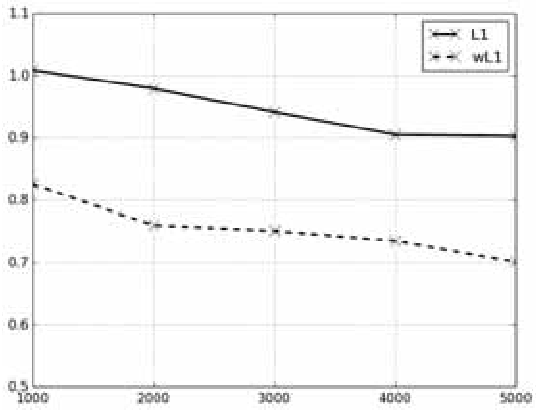


L1Nuclear

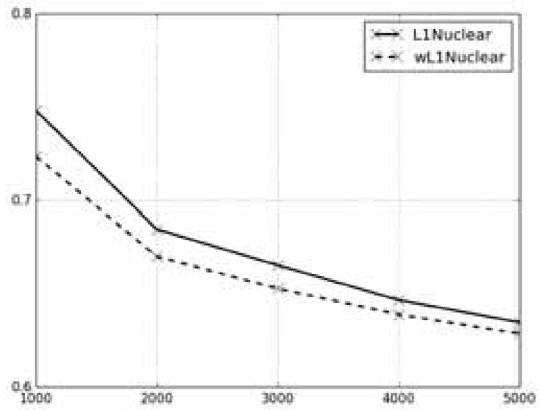


wL1Nuclear

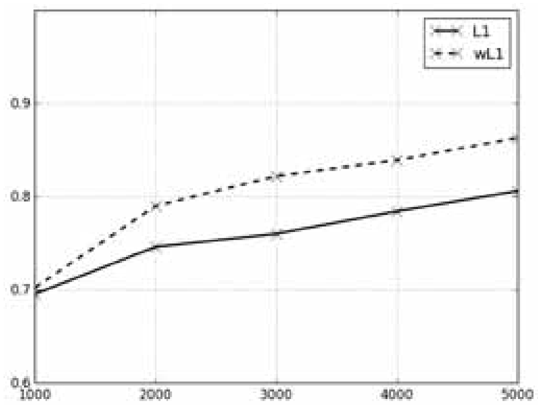
Consequence: a sharp scaling of penalizations



Error for L1 and wL1

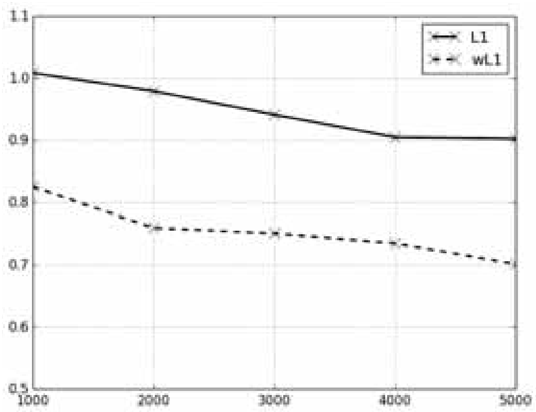


Error for L1Nuclear and wL1Nuclear

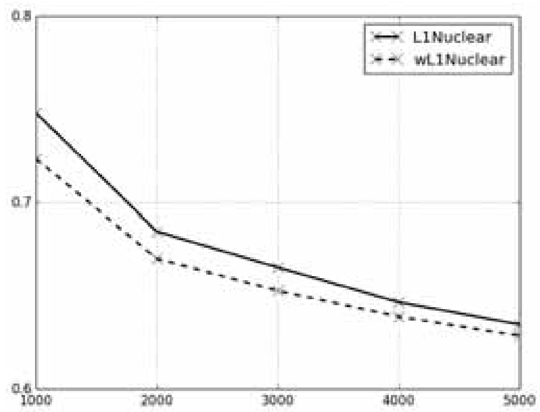


AUC for L1 and wL1

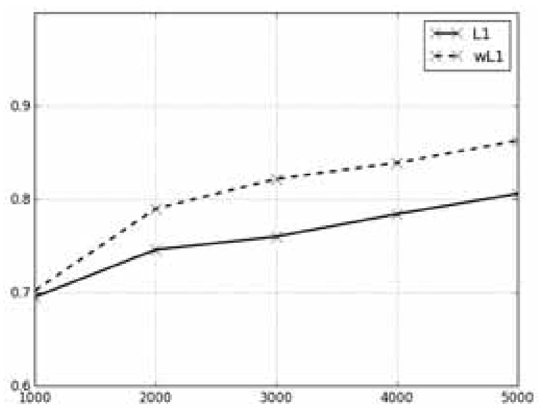
Consequence: a sharp scaling of penalizations



Error for L1 and wL1



Error for L1Nuclear and wL1Nuclear



AUC for L1 and wL1

Take-home message

- Reparametrization of the problem
- Theoretical analysis gives insight to choose the correct scaling of the penalizations
- First oracle inequality for this problem
- This required new probabilistic tools for matrix martingales in continuous time

Perpectives

- Larger scale: factorized form $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{V}^\top$
- Incorporation of features (text, time-varying graph-features, etc.)
- Time varying baseline $\mu(t)$ for non-stationarity

Thank you!