Journal of Machine Learning Research 21 (2020) 1-32 Submitted 3/15; Revised 2/20; Published 4/20

# Sparse and low-rank multivariate Hawkes processes

**Emmanuel Bacry** EMMANUEL.BACRY@POLYTECHNIQUE.EDU
*CEREMADE, CNRS UMR 7534, Université Paris-Dauphine, Paris, France*

**Martin Bompaire** M.BOMPAIRE@CRITEO.COM
*Criteo, Paris, France*

**Stéphane Gaïffas** STEPHANE.GAIFFAS@LPSM.PARIS
*LPSM, CNRS UMR 8001, Université Paris-Diderot, Paris, France*
*DMA, CNRS UMR 8553, Ecole Normale Supérieure, Paris, France*

**Jean-Francois Muzy** MUZY@UNIV-CORSE.FR
*Laboratoire Sciences Pour l'Environnement, CNRS UMR 6134, Université de Corse, Corté, France*

**Editor:** Nicolas Vayatis

## Abstract

We consider the problem of unveiling the implicit network structure of node interactions (such as user interactions in a social network), based only on high-frequency timestamps. Our inference is based on the minimization of the least-squares loss associated with a multivariate Hawkes model, penalized by $\ell_1$ and trace norm of the interaction tensor. We provide a first theoretical analysis for this problem, that includes sparsity and low-rank inducing penalizations. This result involves a new data-driven concentration inequality for matrix martingales in continuous time with observable variance, which is a result of independent interest and a broad range of possible applications since it extends to matrix martingales former results restricted to the scalar case. A consequence of our analysis is the construction of sharply tuned $\ell_1$ and trace-norm penalizations, that leads to a data-driven scaling of the variability of information available for each users. Numerical experiments illustrate the significant improvements achieved by the use of such data-driven penalizations.

**Keywords.** Hawkes processes; Sparsity; Low-Rank; Random matrices; Data-driven concentration

## 1. Introduction

Understanding the dynamics of social interactions is a challenging problem of rapidly growing interest (de Menezes and Barabási, 2004; Leskovec, 2008; Crane and Sornette, 2008; Leskovec et al., 2009) because of the large number of applications in web-advertisement and e-commerce, where large-scale logs of event history are available. A common supervised approach consists in the prediction of labels based on declared interactions (friendship, like, follower, etc.). However such supervision is not always available, and it does not always describe accurately the level of interactions between users. Labels are often only binary while a quantification of the interaction is more interesting, declared interactions are often deprecated, and more generally a supervised approach is not enough to infer the latent communities of users, as temporal patterns of actions of users are much more informative.

For latent social groups recovering, several recent papers (Rodriguez et al., 2011; Gomez-Rodriguez et al., 2013; Daneshmand et al., 2014) consider an approach directly based on the real *actions* or *events* of users (referred to as *nodes* in the following) that are fully identified through their corresponding user id and timestamp. These models assume a structure of data consisting in a sequence of independent cascades, containing the timestamp of each node. In these works, techniques coming from survival analysis are used to derive a tractable convex likelihood, that allows one to infer the latent community structure. However, they require that data are already segmented into sets of independent cascades, which is often unrealistic. Moreover, it does not allow for recurrent events, namely a node can be infected only once, and it cannot incorporate exogenous factors, i.e., influence from the world outside the network.

Another approach is based on self-exciting point processes, such as the Hawkes process (Hawkes, 1971). Previously used for geophysics (Ogata, 1998), high-frequency finance (Bacry et al., 2013, 2015), crime activity (Mohler et al., 2011), these processes have been recently used for the modelization of users activity in social networks, see for instance Crane and Sornette (2008); Blundell et al. (2012); Zhou et al. (2013); Yang and Zha (2013). The structure of the Hawkes model allows us to capture the direct influence of a specific user's action on all the future actions of all the users (including himself). It encompasses in a single likelihood the decay of the influence over time, the levels of interaction between nodes, which can be seen as a weighted asymmetrical adjacency matrix, and a baseline intensity, that measures the level of exogeneity of a user, namely the spontaneous apparition of an action, with no influence from other nodes of the network.

In this paper, we consider such a multivariate Hawkes process (MHP), and we combine convex proxies for sparsity and low-rank of the adjacency tensor and the baseline intensities, that are now of common use in low-rank modeling in collaborative filtering problems (Candès and Tao, 2004, 2009). Note that this approach is also considered in (Zhou et al., 2013). We provide a first theoretical analysis of the generalization error for this problem, see Hansen et al. (2012) for an analysis including only entrywise $\ell_1$ penalization. Namely, we prove a sharp oracle inequality for our procedure, that includes sparsity and low-rank inducing priors, see Theorem 6 in Section 5. This result involves a new data-driven concentration inequality for matrix martingales in continuous time, see Theorems 3 and 4 in Section 3.3, that are results of independent interest, that extends previous non-commutative versions of concentration inequalities for martingales in discrete time, see Tropp (2012). A consequence of our analysis is the construction of sharply tuned $\ell_1$ and trace-norm penalizations, that leads to a data-driven scaling of the variability of information available for each node. We give empirical evidence of the improvements of our data-driven penalizations, by conducting in Section 6 numerical experiments on simulated data. Since the objectives involved are convex with a smooth component, our algorithms build upon standard batch proximal gradient descent algorithms.

## 2. The multivariate Hawkes model and the least-squares functional

Consider a finite network with $d$ nodes (each node corresponding to a user in a social network for instance). For each node $j \in \{1, \ldots, d\}$, we observe the timestamps $\{t_{j,1}, t_{j,2}, \ldots\}$ of actions of node $j$ on the network (a message, a click, etc.). With each node $j$ is associated a counting process $N_j(t) = \sum_{i \geq 1} \mathbf{1}_{t_{j,i} \leq t}$ and we consider the $d$-dimensional counting process $N_t = [N_1(t) \ \cdots \ N_d(t)]^\top$, for $t \geq 0$. We observe this process for $t \in [0, T]$. Each $N_j$ has an intensity $\lambda_j$, meaning that

$$\mathbb{P}\big(N_j \text{ has a jump in } [t, t+dt] \mid \mathcal{F}_t\big) = \lambda_j(t)dt, \quad j = 1, \ldots, d,$$

where $\mathcal{F}_t$ is the $\sigma$-field generated by $N$ up to time $t$. The multivariate Hawkes model assumes that each $N_j$ has an intensity $\lambda_{j,\theta}$ given by

$$\lambda_{j,\theta}(t) = \mu_j + \sum_{j'=1}^{d} \int_{(0,t)} \varphi_{j,j'}(t-s) dN_{j'}(s), \tag{1}$$

where $\mu_j \geq 0$ is the baseline intensity of $j$ (i.e., the intensity of exogenous events of node $j$) and where the functions $\varphi_{j,j'} : \mathbb{R}^+ \to \mathbb{R}$ for $j = 1, \ \ldots, d$, called *kernels*, allow to quantify the impact of node $j'$ on node $j$. Note that the integral used in Equation (1) is a Stieljes integral, namely it simply stands for

$$\int_{(0,t)} \varphi(t-s) dN_{j'}(s) = \sum_{i \,:\, t_{j',i} \in [0,t)} \varphi(t - t_{j',i}).$$

In the paper, we consider general kernel functions $\varphi_{j,j'}(t)$ that can be written as:

$$\varphi_{j,j'}(t) = \sum_{k=1}^{K} a_{j,j',k} h_{j,j',k}(t). \tag{2}$$

where the coefficients $a_{j,j',k}$ are the entries of a $d \times d \times K$ tensor $\mathbb{A}$ (i.e., $(\mathbb{A})_{j,j',k} = a_{j,j',k}$) and the kernels $h_{j,j',k}(t)$ are elements of a fixed dictionnary of non negative and causal functions ($h_{j,j',k} : \mathbb{R}^+ \to \mathbb{R}^+$) such that $\|h_{j,j',k}\|_1 = 1$. In that respect, the weights $a_{j,j',1}, \ldots, a_{j,j',K}$ all quantify the influence of $j'$ on $j$, but the particular weight $a_{j,j',k}$ quantifies it for the $k$-th *decay function* $h_{j,j',k}$. A standard choice is a dictionnary of exponential kernels, $h_{j,j',k}(t) = \alpha_k e^{-\alpha_k t}$ with varying memory parameters $\alpha_1, \ldots, \alpha_K$. This leads to the following standard parametrization of the kernel functions, called *exponential kernels*:

$$\varphi_{j,j'}(t) = \sum_{k=1}^{K} a_{j,j',k} \alpha_k \exp(-\alpha_k t). \tag{3}$$

The main advantage of exponential kernels with fixed memory parameters $\alpha_1, \ldots, \alpha_K$, is that it allows one to handle a convex problem. In the general case or when the memory parameters are unknown, the problem becomes non-convex, more challenging and is beyond the scope of the paper.

The parameter of interest is the *self-excitement* tensor $\mathbb{A}$, which can be viewed as a cross-scale (for $k = 1, \ldots, K$) weighted adjacency matrix of connectivity between nodes, as illustrated in Figure 1 below.

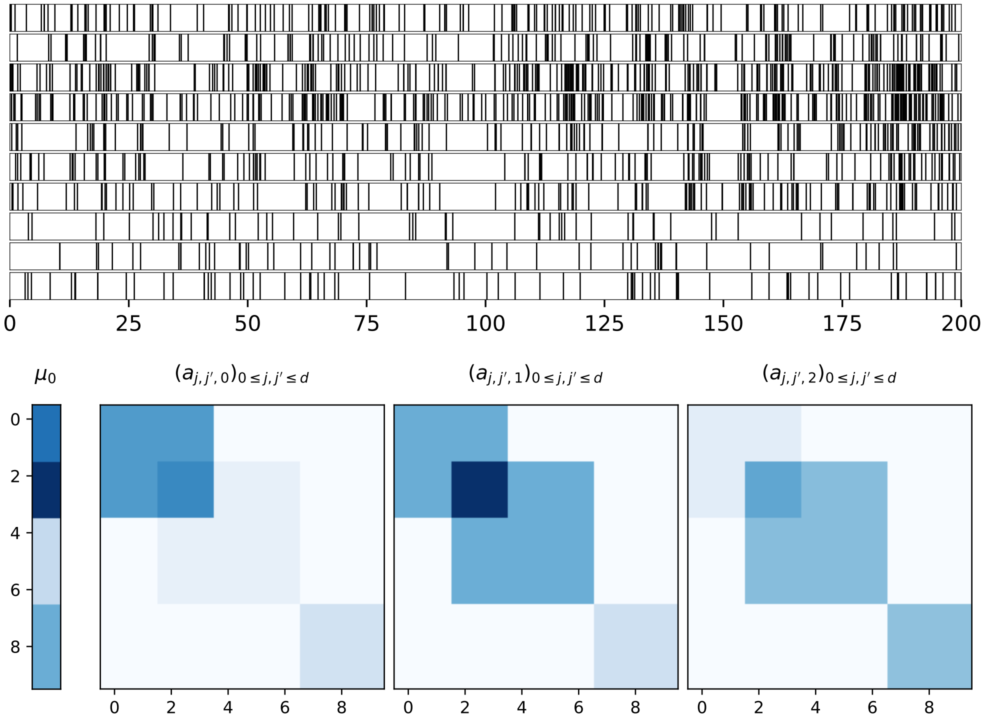

Figure 1: Toy example with $d = 10$ nodes. Based on actions' timestamps of the nodes, represented by vertical bars (top), we aim at recovering the vector $\mu_0$ and the tensor $\mathbb{A}$ of implicit influence between nodes (bottom).

The Hawkes model is particularly relevant for the modelization of the "microscopic" activity of social networks and has attracted a lot of interest in the recent literature (see Crane and Sornette (2008); Blundell et al. (2012); Zhou et al. (2013); Yang and Zha (2013); Linderman and Adams (2014); DuBois et al. (2013); Blundell et al. (2012); Iwata et al. (2013), among others) for this kind of application, with a particular emphasis on Hansen et al. (2012) that gives first theoretical results for the Lasso used with Hawkes processes with an application to neurobiology. The main point is that this simple autoregressive structure of the intensity allows us to capture the direct influence of a user, based on the recurrence and the patterns of his actions, by separating the intensity into a baseline and a self-exciting component, hence allowing to filter exogeneity in the estimation of users' influences on each others.

We introduce in this paper an estimation procedure of $\theta = (\mu, \mathbb{A})$ based on data $\{N_t : t \in [0, T]\}$. The hidden structure underlying the observed actions of nodes is contained in $\mathbb{A}$. Our strategy is based on the least-squares functional given by

$$R_T(\theta) = \|\lambda_\theta\|_T^2 - \frac{2}{T} \sum_{j=1}^{d} \int_{[0,T]} \lambda_{j,\theta}(t) dN_j(t), \tag{4}$$

with respect to $\theta$, where $\|\lambda_\theta\|_T^2 = \frac{1}{T} \sum_{j=1}^{d} \int_{[0,T]} \lambda_{j,\theta}(t)^2 dt$ is the norm associated with the inner product

$$\langle \lambda_\theta, \lambda_{\theta'} \rangle_T = \frac{1}{T} \sum_{j=1}^{d} \int_{[0,T]} \lambda_{j,\theta}(t) \lambda_{j,\theta'}(t) dt. \tag{5}$$

This least-squares function is very natural, and comes from the empirical risk minimization principle (Van De Geer, 2000; Massart, 2007; Koltchinskii, 2011; Bartlett and Mendelson, 2006): assuming that $N_j$ has an unknown ground truth intensity $\lambda_j$ (not necessarily following the Hawkes model), the Doob-Meyer's decomposition gives

$$\int_{[0,T]} \lambda_{j,\theta}(t) dN_j(t) = \int_{[0,T]} \lambda_{j,\theta}(t) \lambda_j(t) dt + \int_{[0,T]} \lambda_{j,\theta}(t) dM_j(t),$$

where $M_j(t) = N_j(t) - \int_0^t \lambda_j(s) ds$ is a continuous-time martingale with upwards jumps of +1. Since the "noise" term $\int_{[0,T]} \lambda_{j,\theta}(t) dM_j(t)$ is centered, we obtain

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

so that we expect a minimum $\hat{\theta}$ of $R_T(\theta)$ to lead to a good estimation $\lambda_{\hat{\theta}}$ of $\lambda$, following the empirical risk minimization principle. As explained in Section 8 below, the noise terms can be written as

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

for a specific tensor $\mathbb{T}_t$ and matrix martingale $\boldsymbol{M}_t$, where $\mathbb{T}_s \circ \boldsymbol{M}_s$ stands for a tensor-matrix product defined in Section 3.1 below. The next Section introduces new results, of independent interest, providing *data-driven* deviation inequalities for the operator norm of a matrix martingale defined as the stochastic integral $\int_0^t \mathbb{T}_s \circ d\boldsymbol{M}_s$. These results allow us, as a by-product, to control the noise terms arising in the application considered in this paper, and lead to a sharp data-driven tuning of the penalizations used on $\mathbb{A}$, as explained in Section 4 below.

## 3. A new data-driven matrix martingale Bernstein's inequality

An important ingredient for the theoretical results proposed in this paper is an observable deviation inequality for continuous time matrix martingales. We first recall previous results obtained in Bacry et al. (2016b) about non-observable deviation inequalities for such objects.

### 3.1. Notations

Let $\mathbb{T}$ be a tensor of shape $m \times n \times p \times q$. It can be considered as a linear mapping from $\mathbb{R}^{p \times q}$ to $\mathbb{R}^{m \times n}$ according to the following "tensor-matrix" product:

$$(\mathbb{T} \circ \boldsymbol{A})_{i,j} = \sum_{k=1}^{p} \sum_{l=1}^{q} \mathbb{T}_{i,j;k,l} \boldsymbol{A}_{k,l}.$$

We will denote by $\mathbb{T}^\top$ the tensor such that $\mathbb{T}^\top \circ \boldsymbol{A} = (\mathbb{T} \circ \boldsymbol{A})^\top$ (i.e., $\mathbb{T}^\top_{i,j;k,l} = \mathbb{T}_{j,i;k,l}$) and by $\mathbb{T}_{\bullet,\bullet;k,l}$ and $\mathbb{T}_{i,j;\bullet,\bullet}$ the matrices obtained when fixing the indices $k,l$ and $i,j$ respectively. Note that $(\mathbb{T} \circ \boldsymbol{A})_{i,j} = \operatorname{tr}(\mathbb{T}_{i,j;\bullet,\bullet} \boldsymbol{A}^\top)$. If $\mathbb{T}$ and $\mathbb{T}'$ are two tensors of dimensions $m \times n \times p \times q$ and $n \times r \times p \times q$ respectively, $\mathbb{T}\mathbb{T}'$ stands for the $m \times r \times p \times q$ tensor defined as $(\mathbb{T}\mathbb{T}')_{i,j;k,l} = (\mathbb{T}_{\bullet,\bullet;k,l}\mathbb{T}'_{\bullet,\bullet;k,l})_{i,j}$. Accordingly, for an integer $r \geq 1$, if $\mathbb{T}_{\bullet,\bullet;a,b}$ are square matrices, we will denote by $\mathbb{T}^r$ the tensor such that $(\mathbb{T}^r)_{i,j;k,l} = (\mathbb{T}^r_{\bullet,\bullet;k,l})_{i,j}$. We also introduce $\|\mathbb{T}\|_{\mathrm{op};\infty} = \max_{k,l} \|\mathbb{T}_{\bullet,\bullet;k,l}\|_{\mathrm{op}}$, the maximum operator norm of all matrices formed by the first two dimensions of tensor $\mathbb{T}$.

In this paper we shall consider the class of $m \times n$ matrix martingales that can be written as

$$\boldsymbol{Z}_{\mathbb{T}}(t) = \int_0^t \mathbb{T}_s \circ d\boldsymbol{M}_s, \tag{6}$$

where $\mathbb{T}_s$ is a tensor with dimensions $m \times n \times p \times q$, whose components are assumed to be locally bounded predictable random functions. The process $\boldsymbol{M}_t$ is a $p \times q$ is matrix with entries that are square integrable martingales with a diagonal quadratic covariation matrix. More explicitly, the entries of $\boldsymbol{Z}_{\mathbb{T}}(t)$ are given by

$$(\boldsymbol{Z}_{\mathbb{T}}(t))_{i,j} = \sum_{k=1}^p \sum_{l=1}^q \int_0^t (\mathbb{T}_s)_{i,j;k,l} (d\boldsymbol{M}_s)_{k,l},$$

where the martingale $\boldsymbol{M}_t$ is a matrix of compensated counting processes $\boldsymbol{M}_t = \boldsymbol{N}_t - \boldsymbol{\lambda}_t$ where $\boldsymbol{N}_t$ is a $p \times q$ matrix counting process (i.e., each component is a counting process) with an intensity process $\boldsymbol{\lambda}_t$ which is predictable, continuous and with finite variations (FV).

### 3.2. A non-observable matrix martingale Bernstein's inequality

The next Theorem (which is a small variation of Theorem 2 in Bacry et al. (2016b)) provides a concentration inequality for $\|\boldsymbol{Z}_{\mathbb{T}}(t)\|_{\mathrm{op}}$, the operator norm of $\boldsymbol{Z}_{\mathbb{T}}(t)$. Before stating the Theorem, let us introduce some more notations. We define

$$b_{\mathbb{T}}(t) = \sup_{0 \leq s \leq t} \max\big(\|\mathbb{T}_s\|_{\mathrm{op};\infty}, \|\mathbb{T}_s^\top\|_{\mathrm{op};\infty}\big), \tag{7}$$

and depending on whether the tensor $\mathbb{T}_s$ is symmetric (i.e., $\mathbb{T}_s^\top = \mathbb{T}_s$ and $m = n$) or not, we define the following.

- If $\mathbb{T}_s$ is symmetric, we define
$$\boldsymbol{W}_{\mathbb{T}}(s) = \mathbb{T}_s^2 \circ \boldsymbol{\lambda}_s \tag{8}$$
and $K_{m,n} = m$
- If $\mathbb{T}_s$ is not symmetric, we define
$$\boldsymbol{W}_{\mathbb{T}}(s) = \begin{bmatrix} \mathbb{T}_s\mathbb{T}_s^\top \circ \boldsymbol{\lambda}_s & \boldsymbol{0} \\ \boldsymbol{0} & \mathbb{T}_s^\top\mathbb{T}_s \circ \boldsymbol{\lambda}_s \end{bmatrix}, \tag{9}$$
and $K_{m,n} = m + n$.

In both cases, we define

$$\boldsymbol{V}_{\mathbb{T}}(t) = \int_0^t \boldsymbol{W}_{\mathbb{T}}(s)\, ds. \tag{10}$$

Finally, all along the paper we denote $\phi(x) = e^x - 1 - x$ for $x \in \mathbb{R}$. The following concentration inequality is an easy consequence of Theorem 1 from Bacry et al. (2016b).

**Theorem 1** *Let $\boldsymbol{Z}_{\mathbb{T}}(t)$ be the $m \times n$ matrix martingale given by Equation* (6)*. Moreover, assume that*

$$\mathbb{E}\Big[\int_0^t \frac{\phi\big(3\max(\|\mathbb{T}_s\|_{\mathrm{op};\infty}, \|\mathbb{T}_s^\top\|_{\mathrm{op};\infty})\big)}{\max(\|\mathbb{T}_s\|_{\mathrm{op};\infty}^2, \|\mathbb{T}_s^\top\|_{\mathrm{op};\infty}^2)}(\boldsymbol{W}_{\mathbb{T}}(s))_{i,j} ds\Big] < +\infty, \tag{11}$$

*for any $1 \leq i, j \leq m+n$. Then for any $\xi \in (0,3)$, $t, b, x > 0$, the following holds*:

$$\mathbb{P}\Big[\|\boldsymbol{Z}_{\mathbb{T}}(t)\|_{\mathrm{op}} \geq \frac{\phi(\xi)}{\xi b}\lambda_{\max}\big(\boldsymbol{V}_{\mathbb{T}}(t)\big) + \frac{xb}{\xi}, \quad b_{\mathbb{T}}(t) \leq b\Big] \leq K_{m,n}e^{-x}. \tag{12}$$

*Optimizing this last inequality on $\xi$ gives*

$$\mathbb{P}\Big[\|\boldsymbol{Z}_{\mathbb{T}}(t)\|_{\mathrm{op}} \geq \sqrt{2vx} + \frac{bx}{3}, \lambda_{\max}(\boldsymbol{V}_{\mathbb{T}}(t)) \leq v, \quad b_{\mathbb{T}}(t) \leq b\Big] \leq K_{m,n}e^{-x}. \tag{13}$$

The proof of Theorem 1 is given in Section 8.1 below. This result is a Freedman (or Bernstein) inequality for the operator norm of $\boldsymbol{Z}_{\mathbb{T}}(t)$, that provides a deviation based on a variance term $\boldsymbol{V}_{\mathbb{T}}(t)$ and a $L^\infty$ term $b_{\mathbb{T}}(t)$. It is a strong generalization of the scalar Freedman inequality for continuous time martingales, and this result match exactly the scalar case whenever $\boldsymbol{Z}_{\mathbb{T}}(t)$ is scalar. A more thorough discussion about the consequences of this result is provided in Bacry et al. (2016b).

### 3.3. Data-driven matrix martingale Bernstein's inequalities

Inequality (13) is of poor practical interest in situations where one observes only the jumping times of the $\boldsymbol{Z}_t$ components (namely $\boldsymbol{N}_t$) and not the stochastic intensity $\boldsymbol{\lambda}_t$. In that respect, one needs a "data driven" inequality where $\boldsymbol{V}_{\mathbb{T}}(t)$ is replaced by its empirical version $\widehat{\boldsymbol{V}}_{\mathbb{T}}(t)$.

- If $\mathbb{T}_s$ is symmetric, we define
$$\widehat{\boldsymbol{V}}_{\mathbb{T}}(t) = \int_0^t \mathbb{T}_s^2 \circ d\boldsymbol{N}_s,$$
- while if $\mathbb{T}_s$ is not symmetric, we define
$$\widehat{\boldsymbol{V}}_{\mathbb{T}}(t) = \begin{bmatrix} \int_0^t \mathbb{T}_s\mathbb{T}_s^\top \circ d\boldsymbol{N}_s & \mathbf{0} \\ \mathbf{0} & \int_0^t \mathbb{T}_s^\top\mathbb{T}_s \circ d\boldsymbol{N}_s \end{bmatrix}.$$

The next Proposition allows us to control $\lambda_{\max}(\boldsymbol{V}_{\mathbb{T}}(t))$ using its observable counterpart $\lambda_{\max}(\widehat{\boldsymbol{V}}_{\mathbb{T}}(t))$ with a large probability. This result is a generalization to arbitrary matrices of dimensions $m \times n$ of an analog inequality originally proven by Hansen et al. (2012) for scalar martingales.

**Proposition 2** *For any $x, b > 0$ and $\xi \in (0,3)$ such that $\xi > \phi(\xi)$, we have*

$$\mathbb{P}\Big[\lambda_{\max}(\boldsymbol{V}_{\mathbb{T}}(t)) \geq \frac{\xi}{\xi - \phi(\xi)}\lambda_{\max}(\widehat{\boldsymbol{V}}_{\mathbb{T}}(t)) + \frac{xb^2}{\xi - \phi(\xi)}, \quad b_{\mathbb{T}}(t) \leq b\Big] \leq K_{m,n}e^{-x},$$

*where $K_{m,n}$ is defined as in Theorem 1. Moreover, choosing $\xi = -W_{-1}(-\frac{2}{3}e^{-2/3}) - 2/3$ (note that $\xi \approx 0.762$), where $W_{-1}$ is the second branch of the Lambert W function, leads to*

$$\mathbb{P}\Big[\lambda_{\max}(\boldsymbol{V}_{\mathbb{T}}(t)) \geq 2\lambda_{\max}(\widehat{\boldsymbol{V}}_{\mathbb{T}}(t)) + cb^2x, \quad b_{\mathbb{T}}(t) \leq b\Big] \leq K_{m,n}e^{-x}$$

*for any $x, b > 0$, with $c = 2.62$.*

Thanks to Proposition 2, we can establish an analog of Theorem 1 where $\lambda_{\max}(\boldsymbol{V}_{\mathbb{T}}(t))$ is replaced by its data-driven version $\lambda_{\max}(\widehat{\boldsymbol{V}}_{\mathbb{T}}(t))$, up to a slight loss in values of the numerical constants.

**Theorem 3** *With the same notations and assumptions as in Theorem 1 one has*

$$\mathbb{P}\bigg[\|\boldsymbol{Z}_{\mathbb{T}}(t)\|_{\mathrm{op}} \geq 2\sqrt{vx} + cbx, \ \ \lambda_{\max}(\widehat{\boldsymbol{V}}_{\mathbb{T}}(t)) \leq v, \ \ b_{\mathbb{T}}(t) \leq b\bigg] \leq 2K_{m,n}e^{-x} \tag{14}$$

*for any $x, b > 0$ with $c = 14.39$.*

The proof of Theorem 3 is given in Section 8.3 below. It follows simple arguments that combine Theorem 1 and Proposition 2. However, this inequality is stated on the events $\{\lambda_{\max}(\widehat{\boldsymbol{V}}_{\mathbb{T}}(t)) \leq v\}$ and $\{b_{\mathbb{T}}(t) \leq b\}$, while an unconditional deviation inequality is more practical. Such a result, which involves some extra technicalities, is stated in the next Theorem.

**Theorem 4** *With the same conditions and notations as in Theorem 3, one has*

$$\mathbb{P}\bigg[\|\boldsymbol{Z}_{\mathbb{T}}(t)\|_{\mathrm{op}} \geq 2\sqrt{\lambda_{\max}(\widehat{\boldsymbol{V}}_{\mathbb{T}}(t))(x + \ell_x(t))} + c(x + \ell_x(t))(1 + b_{\mathbb{T}}(t))\bigg] \leq C_{m,n}e^{-x} \tag{15}$$

*where $C_{m,n} = \frac{\pi^4}{18\log(2)^4}K_{m,n} \leq 23.45K_{m,n}$, where $c = 14.39$ and*

$$\ell_x(t) = 2\log\log\Big(\frac{4\lambda_{\max}(\widehat{\boldsymbol{V}}_{\mathbb{T}}(t))}{x} \vee 2\Big) + 2\log\log(4b_{\mathbb{T}}(t) \vee 2).$$

The proof of this Theorem is given in Section 8.4. It is a result of independent interest, that gives a control on the operator norm of a matrix martingale in continuous time (with jumps at most 1), using only observable quantities. Along with Bacry et al. (2016b), it provides a first deviation inequality for such objects, and it can be understood as a data-driven version of the results given in Bacry et al. (2016b).

## 4. The procedure

We want to produce an estimation procedure of $\theta = (\mu, \mathbb{A})$ based on data from $\{N_t : t \in [0, T]\}$. Following the empirical risk minimization principle, the estimation procedure uses the least-squares functional (4) as a goodness-of-fit. In addition to this goodness-of-fit criterion, we need to use a penalization that allows us to reduce the dimensionality of the model, namely we consider

$$\hat{\theta} \in \operatorname*{argmin}_{\theta = (\mu, \mathbb{A}) \in \mathbb{R}_+^d \times \mathbb{R}_+^{d \times d \times K}} \big\{R_T(\theta) + \mathrm{pen}(\theta)\big\}, \tag{16}$$

for a specific penalization function $\mathrm{pen}(\theta)$ described below. In particular, we want to reduce the dimensionality of $\mathbb{A}$, based on the prior assumption that latent factors explain the connectivity of users in the network. This leads to a low-rank assumption on $\mathbb{A}$, which is commonly used in collaborative filtering and matrix completion techniques (Ricci et al., 2011). Our prior assumptions on $\mu$ and $\mathbb{A}$ are the following.

**Sparsity of $\mu$.** Some nodes are basically inactive and react only if stimulated. Hence, we assume that the baseline intensity vector $\mu$ is sparse.

**Sparsity of $\mathbb{A}$.** A node interacts only with a fraction of other nodes, meaning that for a fixed node $j$, only a few $a_{j,j',k}$ are non-zero. Moreover, a node might react at specific time scales only, namely $a_{j,j',k}$ is non-zero for some $k$ only for fixed $j, j'$. Hence, we assume that $\mathbb{A}$ is an entrywise sparse tensor.

**Low-rank of $\mathbb{A}$.** Using together Equations (1) and (2), one can write

$$\begin{aligned}\lambda_{j,\theta}(t) &= \mu_j + \sum_{j'=1}^{d}\sum_{k=1}^{K} a_{j,j',k}\int_{(0,t)} h_{j,j',k}(t-s)dN_{j'}(s) \\ &= \mu_j + \big(\mathrm{hstack}(\mathbb{A})_{j,\bullet}\big)^\top \mathrm{hstack}(\mathbb{H}(t))_{j,\bullet},\end{aligned} \tag{17}$$

where $\mathbb{H}(t)$ is the $d \times d \times K$ tensor with entries

$$\mathbb{H}_{j,j',k}(t) = \int_{(0,t)} h_{j,j',k}(t-s) dN_{j'}(s), \tag{18}$$

where $(\boldsymbol{X})_{j,\bullet}$ stands for the $j$-th row of a matrix $\boldsymbol{X}$ and where hstack stands for the horizontally stacking operator defined by

$$\mathrm{hstack} : \mathbb{R}^{d \times d \times K} \to \mathbb{R}^{d \times Kd} \quad \text{such that} \quad \mathrm{hstack}(\mathbb{A}) = \begin{bmatrix} \mathbb{A}_{\bullet,\bullet,1} & \cdots & \mathbb{A}_{\bullet,\bullet,K} \end{bmatrix}, \tag{19}$$

where $\mathbb{A}_{\bullet,\bullet,k}$ stands for the $d \times d$ matrix with entries $(\mathbb{A}_{\bullet,\bullet,k})_{j,j'} = \mathbb{A}_{j,j',k}$. In view of Equation (17), all the impacts of nodes $j'$ at time scale $k$ on node $j$ is encoded in the $j$-th row of the $d \times Kd$ matrix $\mathrm{hstack}(\mathbb{A})$. Therefore, a natural assumption is that the matrix $\mathrm{hstack}(\mathbb{A})$ has a low-rank: we assume that there exist latent factors that explain the way nodes impact other nodes through the different scales $k = 1, \ldots, K$.

To induce these prior assumptions on the parameters, we use a penalization based on a mixture of the $\ell_1$ and trace-norms. These norms are respectively the tightest convex relaxations for sparsity and low-rank, see for instance Candès and Tao (2004, 2009). They provide state-of-the art results in compressed sensing and collaborative filtering problems, among many other problems. These two norms have been previously combined for the estimation of sparse and low-rank matrices, see for instance Richard et al. (2014) and Zhou et al. (2013) in the context of MHP. Therefore, we consider the following penalization on the parameter $\theta = (\mu, \mathbb{A})$:

$$\mathrm{pen}(\theta) = \|\mu\|_{1,\hat{w}} + \|\mathbb{A}\|_{1,\hat{\mathbb{W}}} + \hat{\tau} \| \mathrm{hstack}(\mathbb{A})\|_*, \tag{20}$$

where each terms are entry-wise weighted $\ell_1$ and trace-norm penalizations given by

$$\|\mu\|_{1,\hat{w}} = \sum_{j=1}^d \hat{w}_j |\mu_j|, \quad \|\mathbb{A}\|_{1,\hat{\mathbb{W}}} = \sum_{1 \le j,j' \le d, 1 \le k \le K} \hat{\mathbb{W}}_{j,j',k} |\mathbb{A}_{j,j',k}|, \quad \|\boldsymbol{A}\|_* = \sum_{j=1}^d \sigma_j(\boldsymbol{A}),$$

where the $\sigma_1(\boldsymbol{A}) \ge \cdots \ge \sigma_d(\boldsymbol{A})$ are the singular values of a matrix $\boldsymbol{A}$ (we take $\boldsymbol{A} = \mathrm{hstack}(\mathbb{A})$ in the penalization). The weights $\hat{w}$, $\hat{\mathbb{W}}$, and coefficients $\hat{\tau}$ are data-driven tuning parameters described below. The choice of these weights comes from a sharp analysis of the noise terms and lead to a data-driven scaling of the variability of information available for each nodes.

From now on, we fix some confidence level $x > 0$, which corresponds to the probability that the oracle inequality from Theorem 6 holds (see Section 5 below). This can be safely chosen as $x = \log T$ for instance, as described in our numerical experiments (see Section 6 below).

**Weight $\hat{\tau}$ for the trace-norm penalization of** $\mathrm{hstack}(\mathbb{A})$**.** This weight comes from Corollary 7 (see Section 8.5). Let us introduce the $d \times Kd$ matrix $\boldsymbol{H}(t) = \mathrm{hstack}(\mathbb{H}(t))$ where $\mathbb{H}(t)$ is the $d \times d \times K$ tensor defined by (18) and hstack is the horizontally stacking operator defined by (19). Let us also recall that $\| \cdot \|_2$ is the $\ell_2$-norm, and define $\|\boldsymbol{H}\|_{\infty,2} = \max_{1 \le j \le d} \|\boldsymbol{H}_{j,\bullet}\|_2$ where $\boldsymbol{H}_{j,\bullet}$ stands for the $j$-th row of $\boldsymbol{H}$. We define

$$\begin{aligned} \hat{\tau} = 4 & \sqrt{\frac{\lambda_{\max}(\widehat{\boldsymbol{V}}(T)/T)(x + \log(2d) + \ell_\tau(T))}{T}} \\ & + 28.78 \frac{x + \log(2d) + \ell_\tau(T))(1 + \sup_{0 \le t \le T} \|\boldsymbol{H}(t)\|_{\infty,2})}{T} \end{aligned} \tag{21}$$

where

$$\lambda_{\max}(\widehat{\boldsymbol{V}}(T)) = \lambda_{\max}\Big( \int_0^T \boldsymbol{H}^\top(s) \boldsymbol{H}(s) \operatorname{diag}(dN(s)) \Big) \bigvee \max_{j=1,\ldots,d} \int_0^T \|\boldsymbol{H}_{j,\bullet}(t)\|_2^2 dN_j(s),$$

and where

$$\ell_\tau(T) = 2 \log\log\Big( \frac{4 \lambda_{\max}(\widehat{\boldsymbol{V}}(T))}{x} \vee 2 \Big) + 2 \log\log \Big( 4 \sup_{0 \le t \le T} \|\boldsymbol{H}(t)\|_{\infty,2} \vee 2 \Big),$$

where we used the notation $a \vee b = \max(a, b)$ for $a, b \in \mathbb{R}$.

**Weights $\hat{w}_j$ for $\ell_1$-penalization of $\mu$.** These weights are given by

$$\hat{w}_j = 6\sqrt{\frac{(N_j(T)/T)(x + \log d + \ell_j(T))}{T}} + 86.34 \frac{x + \log d + \ell_j(T)}{T} \tag{22}$$

with $\ell_j(T) = 2 \log \log(\frac{4N_j(T)}{x} \vee 2)) + 2 \log \log 4$. The weighting of each coordinate $j$ in the penalization of $\mu$ is natural: it is roughly proportional to the square-root of $N_j(T)/T$, which is the average intensity of events on coordinate $j$. The term $\ell_j(T)$ is a technical term, that can be neglected in practice, see Section 6.

**Weights $\hat{\mathbb{W}}_{j,j'k}$ for $\ell_1$-penalization of $\mathbb{A}$.** Recall that the tensor $\mathbb{H}$ is given by (18). The weights $\hat{\mathbb{W}}_{j,j'k}$ are given by

$$\begin{aligned} \hat{\mathbb{W}}_{j,j',k} = 4\sqrt{\frac{\frac{1}{T}\int_0^T \mathbb{H}_{j,j',k}(t)^2 dN_j(t)(x + \log(Kd^2) + \mathbb{L}_{j,j',k}(T))}{T}} \\ + 28.78 \frac{(x + \log(Kd^2) + \mathbb{L}_{j,j',k}(T))(1 + \sup_{0 \le t \le T} |\mathbb{H}_{j,j',k}(t)|)}{T} \end{aligned} \tag{23}$$

where $\mathbb{L}_{j,j',k}(T) = 2 \log \log \big(\frac{4 \int_0^T \mathbb{H}_{j,j',k}(t)^2 dN_j(t)}{x} \vee 2\big) + 2 \log \log(4 \sup_{0 \le t \le T} |\mathbb{H}_{j,j',k}(t)| \vee 2)$. Once again, this is natural: the variance term $\int_0^T \mathbb{H}_{j,j',k}(t)^2 dN_j(t)$ is, roughly, an estimation of the variance of the self-excitements between coordinates $j$ and $j'$ at time scale $k$. The term $\mathbb{L}_{j,j',k}(T)$ is a technical term that can be neglected in practice.

These weights are actually quite natural: the terms $\lambda_{\max}(\widehat{\boldsymbol{V}}(T))$ and $\int_0^T \mathbb{H}_{j,j',k}(t)^2 dN_j(t)$ correspond to estimations of the noise variance, that are the $L^2$ terms appearing in the empirical Bernstein's inequalities given in Section 3.3. The terms $\sup_{0 \le t \le T} \|\boldsymbol{H}(t)\|_{\infty,2}$ and $\sup_{0 \le t \le T} |\mathbb{H}_{j,j',k}(t)|$ correspond to the $L^\infty$ terms from these Bernstein's inequalities. Once again, these data-driven weights lead to a sharp tuning of the penalizations, as illustrated numerically in Section 6 below.

## 5. A sharp oracle inequality

Recall that the inner product $\langle \lambda_1, \lambda_2 \rangle_T$ is given by (5) and recall that $\| \cdot \|_T$ stands for the corresponding norm. Theorem 6 below is a sharp oracle inequality on the prediction error measured by $\|\lambda_{\hat{\theta}} - \lambda\|_T^2$. For the proof of oracle inequalities with a fast rate, one needs a restricted eigenvalue condition on the Gram matrix of the problem (Bickel et al., 2009; Koltchinskii, 2011). One of the weakest assumptions considered in literature is the Restricted Eigenvalue (RE) condition. In our setting, a natural RE assumption is given in Definition 5 below. First, we need to introduce some simple notations and definitions.

**Some notations and definitions.** If $a, b$ (resp. $\boldsymbol{A}, \boldsymbol{B}$ and $\mathbb{A}, \mathbb{B}$) are vectors (resp. matrices and tensors) of the same size, we always denote by $\langle a, b \rangle$ (resp. $\langle \boldsymbol{A}, \boldsymbol{B} \rangle$ and $\langle \mathbb{A}, \mathbb{B} \rangle$) their inner products. For matrices this can be written as $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \sum_{i,j} \boldsymbol{A}_{i,j} \boldsymbol{B}_{i,j} = \operatorname{tr}(\boldsymbol{A}^\top \boldsymbol{B})$, where tr stands for the trace, while for (say, three dimensional) tensors we write similarly $\langle \mathbb{A}, \mathbb{B} \rangle = \sum_{i,j,k} \mathbb{A}_{i,j,k} \mathbb{B}_{i,j,k}$. We define the Euclidean norm (Frobenius) for tensors and matrices simply as $\|\boldsymbol{A}\|_F = \sqrt{\langle \boldsymbol{A}, \boldsymbol{A} \rangle}$ and $\|\mathbb{A}\|_F = \sqrt{\langle \mathbb{A}, \mathbb{A} \rangle}$. If $\boldsymbol{W}$ (resp. $\mathbb{W}$) is a matrix (resp. tensor) with positive entries, we introduce the weighted entrywise $\ell_1$-norm given by $\|\boldsymbol{A}\|_{1,\boldsymbol{W}} = \langle \boldsymbol{W}, |\boldsymbol{A}| \rangle$, (resp. $\|\mathbb{A}\|_{1,\mathbb{W}} = \langle \mathbb{W}, |\mathbb{A}| \rangle$) where $|\boldsymbol{A}|$ (resp. $|\mathbb{A}|$) contains the absolute values of the entries of $\boldsymbol{A}$ (resp. $\mathbb{A}$). If $A$ is a vector, matrix or tensor then $\|A\|_0$ is the number of non-zero entries of $A$, while $\operatorname{supp}(A)$ stands for the support of $A$ (indices of non-zero entries) For another vector, matrix or tensor $A'$ with the same shape, the notation $[A']_{\operatorname{supp}(A)}$ stands for the vector, matrix or tensor with the same coordinates as $A'$ where we put 0 at indices outside of $\operatorname{supp}(A)$. We also use the notation $u \vee v = \max(u, v)$ for $a, b \in \mathbb{R}$.

If $\boldsymbol{A} = \boldsymbol{U} \boldsymbol{\Sigma} \boldsymbol{V}^\top$ is the SVD of a $m \ \times n$ matrix $\boldsymbol{A}$, with the columns $u_j$ of $\boldsymbol{U}$ and $v_k$ of $\boldsymbol{V}$ being, respectively, the orthonormal left and right singular vectors of $\boldsymbol{A}$, the projection matrix onto the space spanned by the columns (resp. rows) of $\boldsymbol{A}$ is given by $\boldsymbol{P}_{\boldsymbol{U}} = \boldsymbol{U} \boldsymbol{U}^\top$ (resp. $\boldsymbol{P}_{\boldsymbol{V}} = \boldsymbol{V} \boldsymbol{V}^\top$). The operator $\mathcal{P}_{\boldsymbol{A}}$ :

$\mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ given by $\mathcal{P}_{\boldsymbol{A}}(\boldsymbol{B}) = \boldsymbol{P_U B} + \boldsymbol{B P_V} - \boldsymbol{P_U B P_V}$ is the projector onto the linear space spanned by the matrices $u_j x^\top$ and $y v_k^\top$ for all $1 \le j, k \le \operatorname{rank}(\boldsymbol{A})$ and $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$. The projector onto the orthogonal space is given by $\mathcal{P}_{\boldsymbol{A}}^\perp(\boldsymbol{B}) = (\boldsymbol{I} - \boldsymbol{P_U})\boldsymbol{B}(\boldsymbol{I} - \boldsymbol{P_V})$.

**Definition 5** *Fix $\theta = (\mu, \mathbb{A})$ where $\mu \in \mathbb{R}^d$ and $\mathbb{A} \in \mathbb{R}_+^{d\times d\times K}$ and define $\boldsymbol{A} = \operatorname{hstack}(\mathbb{A})$. We define the constant $\kappa(\theta) \in (0, +\infty]$ such that, for any $\theta' = (\mu', \mathbb{A}')$ and $\boldsymbol{A}' = \operatorname{hstack}(\mathbb{A}')$ satisfying*

$$\frac{1}{3}\|(\mu')_{\operatorname{supp}(\mu)^\perp}\|_{1,\hat{w}} + \frac{1}{2}\|(\mathbb{A}')_{\operatorname{supp}(\mathbb{A})^\perp}\|_{1,\hat{\mathbb{W}}} + \frac{1}{2}\hat{\tau}\|\mathcal{P}_{\boldsymbol{A}}^\perp(\boldsymbol{A}')\|_* \\ \le \frac{5}{3}\|(\mu')_{\operatorname{supp}(\mu)}\|_{1,\hat{w}} + \frac{3}{2}\|(\mathbb{A}')_{\operatorname{supp}(\mathbb{A})}\|_{1,\hat{\mathbb{W}}} + \frac{3}{2}\hat{\tau}\|\mathcal{P}_{\boldsymbol{A}}(\boldsymbol{A}')\|_*,$$

*we have*

$$\|(\mu')_{\operatorname{supp}(\mu)}\|_2 \vee \|(\mathbb{A}')_{\operatorname{supp}(\mathbb{A})}\|_F \vee \|\mathcal{P}_{\boldsymbol{A}}(\boldsymbol{A}')\|_F \le \kappa(\theta)\|\lambda_{\theta'}\|_T.$$

The constant $1/\kappa(\theta)$ is a restricted eigenvalue depending on the "support" of $\theta$, which is naturally associated with the problem considered here. Roughly, it requires that for any parameter $\theta'$ that has a support close to the one of $\theta$ (measured by domination of the $\ell_1$ norms outside the support of $\theta$ by the $\ell_1$ norm inside it), we have that the $L^2$ norm of the intensity given by $\|\lambda_{\theta'}\|_T$ can be compared with the $L^2$ norm of $\theta'$ in the support of $\theta$. Note that for a given $\theta$, we simply allow $\kappa(\theta) = +\infty$, so the restricted eigenvalue is zero, whenever the inequality is not met (which makes in such as case the statement of Theorem 6 trivial).

**Theorem 6** *Fix $x > 0$, and let $\hat{\theta}$ be given by* (16) *and* (20) *with tuning parameters given by* (21)*,* (22) *and* (23)*. Then, the inequality*

$$\|\lambda_{\hat{\theta}} - \lambda\|_T^2 \le \inf_{\theta=(\mu,\mathbb{A})} \Big\{ \|\lambda_\theta - \lambda\|_T^2 + 1.25\kappa(\theta)^2 \Big( \|(\hat{w})_{\operatorname{supp}(\mu)}\|_2^2 \\ + \|(\hat{\mathbb{W}})_{\operatorname{supp}(\mathbb{A})}\|_F^2 + \hat{\tau}^2 \operatorname{rank}(\operatorname{hstack}(\mathbb{A})) \Big) \Big\} \tag{24}$$

*holds with a probability larger than $1 - 70.35e^{-x}$.*

The proof of Theorem 6 is given in Section 8.5 below. Note that no assumption is required on the ground truth intensity $\lambda$ of the multivariate counting process $N$ in Theorem 6. Moreover, if one forgets in Section 4 about the negligible terms $\ell_\tau(T), \ell_j(T)$ and $\mathbb{L}_{j,j',k}(T)$ and if one keeps only the dominating $L^2$ terms in $O(1/T)$ (while $L^\infty$ terms are $O(1/T^2)$ in the large $T$ regime), we obtain upper bounds, up to numerical constants (denoted $\lesssim$), for the terms involved in Theorem 5:

$$\|(\hat{w})_{\operatorname{supp}(\mu)}\|_2^2 \lesssim \|\mu\|_0 \max_{j\in\operatorname{supp}(\mu)} \frac{\frac{1}{T}N_j(T)(x + \log d)}{T},$$

where $\|\mu\|_0$ stands for the sparsity of $\mu$,

$$\|(\hat{\mathbb{W}})_{\operatorname{supp}(\mathbb{A})}\|_F^2 \lesssim \|\mathbb{A}\|_0 \max_{(j,j',k)\in\operatorname{supp}(\mathbb{A})} \frac{\frac{1}{T}\int_0^T \mathbb{H}_{j,j',k}(t)^2 dN_j(t)(x + \log(Kd^2))}{T},$$

where $\|\mathbb{A}\|_0$ stands for the sparsity of $\mathbb{A}$, and finally

$$\hat{\tau}^2 \lesssim \operatorname{rank}(\operatorname{hstack}(\mathbb{A})) \frac{\frac{1}{T}\lambda_{\max}(\widehat{\boldsymbol{V}}(T))(x + \log(2d))}{T}.$$

Hence, Theorem 6 proves that $\hat{\theta}$ achieves an optimal trade-off between approximation and complexity, where the complexity is, roughly, measured by

$$\frac{\|\mu\|_0(x + \log d)}{T} \max_j \frac{N_j(T)}{T} + \frac{\|\mathbb{A}\|_0(x + \log(Kd^2))}{T} \max_{j,j',k} \frac{1}{T}\int_0^T \mathbb{H}_{j,j',k}(t)^2 dN_j(t) \\ + \frac{\operatorname{rank}(\operatorname{hstack}(\mathbb{A}))(x + \log(2d))}{T} \frac{1}{T}\lambda_{\max}(\widehat{\boldsymbol{V}}(T)).$$

Note that typically $K \leq d$ so that $\log(Kd^2) \leq 3 \log d$, which means that $\log(Kd^2)$ scales as $\log d$. The complexity term depends on both the sparsity of $\mathbb{A}$ and the rank of $\mathrm{hstack}(\mathbb{A})$. The rate of convergence has the "expected" shape $(\log d)/T$, recalling that $T$ is the length of the observation interval of the process, and these terms are balanced by the empirical variance terms coming out of the new concentration results given in Section 3.3 above.

## 6. Numerical experiments

In this Section we conduct experiments on synthetic datasets to evaluate the performance of our method, based on the proposed data-driven weighting of the penalizations, compared to unweighted penalizations (Zhou et al., 2013). Throughout this Section, we consider the most widely used sum of exponentials kernel, defined in Equation (3).

### 6.1. Simulation setting

We generate Hawkes processes using Ogata's thinning algorithm (Ogata, 1981) with $d = 30$ nodes. Baseline intensities $\mu_j$ are constant on blocks, we use $K = 3$ basis kernels $h_{j,j',k}(t) = \alpha_k e^{-\alpha_k t}$ with $\alpha_1 = 0.5$, $\alpha_1 = 2$ and $\alpha_3 = 5$. We consider three examples for the slices $\mathbb{A}_{\bullet,\bullet,1}$, $\mathbb{A}_{\bullet,\bullet,2}$ and $\mathbb{A}_{\bullet,\bullet,3}$ of the adjacency tensor $\mathbb{A}$, including settings with overlapping boxes, and noisy entries over the block structure, as illustrated in Figure 2. These blocks correspond to the overlapping communities reacting at different time scales. The tensor $\mathbb{A}$ is rescaled so that the operator norm of the matrix $\sum_{k=1}^{3} \mathbb{A}_{\bullet,\bullet,k}$ is equal to $0.8$, guaranteeing to obtain a stationary process. For each simulated data, we increase the length of the time interval $T = 5000, 7000, 10000, 15000, 20000$, and fit each time the procedures. An overall averaging of the results is computed on 100 separate simulations.

### 6.2. Procedures and metrics

We consider a procedure based on the minimization of the least-squares functional (4). This objective is convex, with a goodness-of-fit term which is gradient-Lipschitz: we use first-order optimization algorithms, based on proximal gradient descent. Namely, we use Fista (Beck and Teboulle, 2009) for problems with a single penalization on $\mathbb{A}$ ($\ell_1$-norm or trace norm penalization of $\mathrm{hstack}(\mathbb{A})$) and GFB (generalized forward backward, see Pino et al. (1999)) for mixed $\ell_1$ penalization of $\mathbb{A}$ and trace-norm penalization of $\mathrm{hstack}(\mathbb{A})$. For both procedures we choose a fixed gradient step equal to $1/L$ where $L$ is the Lipschitz constant of the loss, namely the largest singular value of the Hessian (which is constant for this least-squares functional). We limit our algorithms to $25,000$ iterations and stop when the objective relative decrease is less than $10^{-10}$ for Fista and $10^{-7}$ for GFB. We only penalize $\mathbb{A}$ and consider the following procedures:

- L1: non-weighted L1 penalization;
- wL1: weighted L1 penalization;
- Nuclear: non-weighted trace-norm penalization;
- L1Nuclear: non-weighted L1 penalization and trace-norm penalization;
- wL1Nuclear: weighted L1 penalization and trace-norm penalization.

Note that L1Nuclear is the same as the procedure considered in Zhou et al. (2013), however, we use a different optimization algorithm, based on an proximal gradient descent (a first-order method, which is typically faster than an algorithm based on ADMM, as proposed in Zhou et al. (2013)). The data-driven weights used in our procedures are the ones derived from our analysis, see (21) and (23), where we simply put $x = \log T$. For each metric, we tune the constant in front the $\ell_1$ penalization, and the constant in front of the trace-norm penalization in order to obtain the best possible metrics for each procedure, on average over all separate simulations. Namely, there is no test set, we simply display the best metrics obtained by each procedure for a

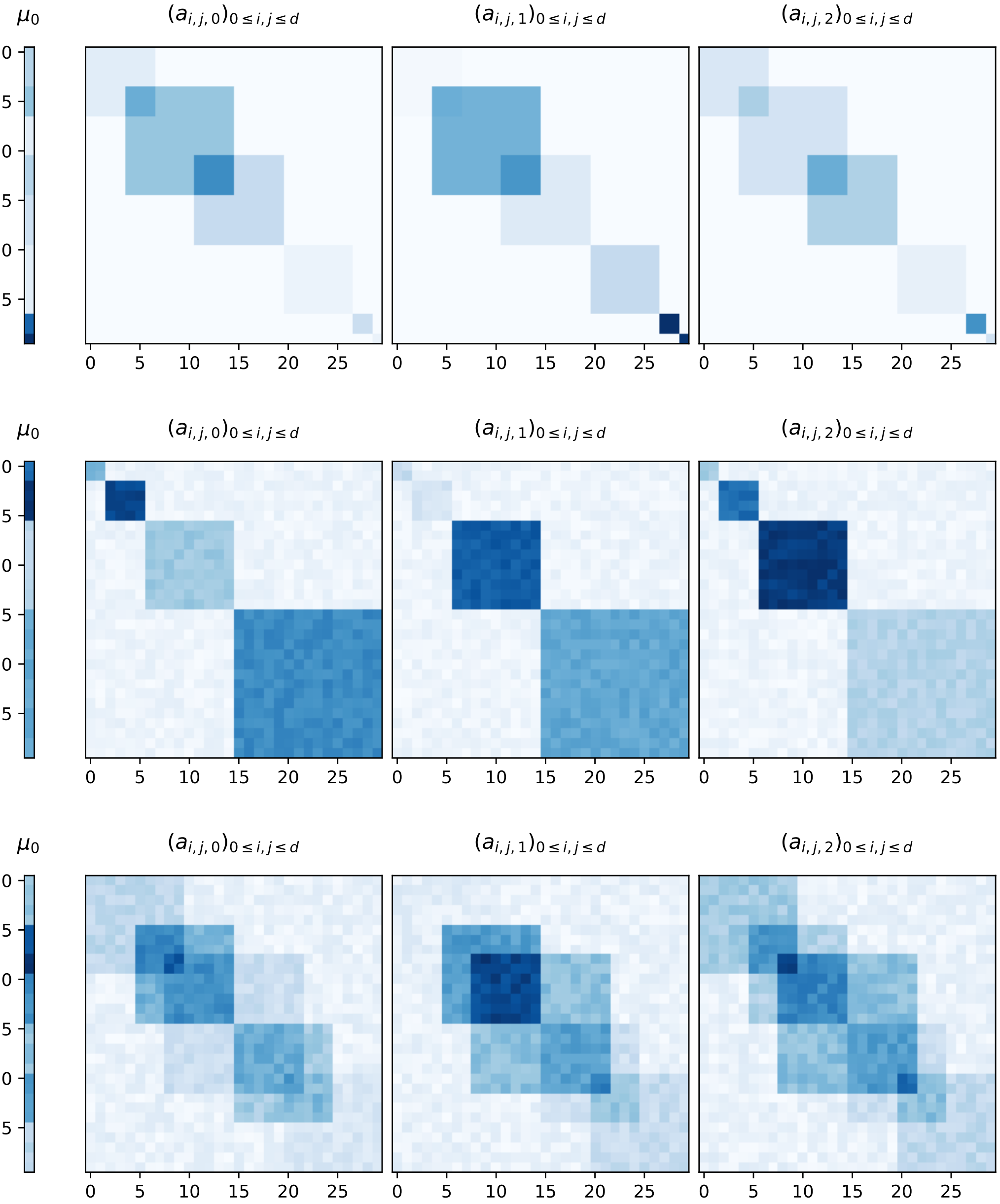

Figure 2: Ground truth vector $\mu$ and tensor $\mathbb{A}$ in dimension 30. Each row corresponds to a different example used in our experiments.

fair comparison. All experiments are done using our `tick` library for `Python3`, see Bacry et al. (2018), its GitHub page is `https://github.com/X-DataInitiative/tick` and documentation is available here `https://x-datainitiative.github.io/tick/`. The following metrics are considered in order to assess the procedures.

**Estimation error:** the relative $\ell_2$ estimation error of $\mathbb{A}$, given by $\|\hat{\mathbb{A}} - \mathbb{A}\|_2^2 / \|\mathbb{A}\|_2^2$

**AUC:** we compute the AUC (area under the ROC curve) between the binarized ground truth matrix $\mathbb{A}$ and the solution $\hat{\mathbb{A}}$ with entries scaled in $[0, 1]$. This allows us to quantify the ability of the procedure to detect the support of the connectivity structure between nodes.

**Kendall:** we compute Kendall's tau-b between all entries of the ground truth matrix $\mathbb{A}$ and the solution $\hat{\mathbb{A}}$. This correlation coefficient takes value between $-1$ and $1$ and compare the number of concordant and discordant pairs. This allows us to quantify the ability of the procedure to rank correctly the intensity of the connectivity between nodes.

### 6.3. Results

In Figure 3 we observe, on an instance of the problem, the strong improvements of wL1 and wL1Nuclear over L1, Nuclear and L1Nuclear respectively. We observe in particular that a sharp tuning of the penalizations, using data-driven weights, leads to a much smaller number of false positives outside the node communities (better viewed on a computer). In Figure 4, we compare all the procedures in terms of estimation error, AUC and Kendall coefficient and confirm the fact that weighted penalizations systematically lead to an improvement, both over unweighted L1, Nuclear and L1Nuclear.

### 6.4. A comparison of the least-squares and likelihood functionals

This paper considers, mostly for theoretical reasons, least-squares as a goodness-of-fit for the Hawkes process. However, estimation in this model is usually achieved by minimizing the goodness-of-fit given by the negative log-likelihood. In what follows, we provide some numerical insights in order to compare objectively both approaches.

First, one can precompute for both functionals some weights in order to accelerate future gradient and value computations. In both cases, the precomputations have similar complexities, unless the number of kernels $K$ is large (see Table 1 below). However, given such precomputations, a remarkable property of the least-squares versus the log likelihood is that value and gradient computation is independent of the total number of observed events (denoted $n$): complexity is $O(K^2 d^3)$ for least-squares, while it is $O(nKd)$ for log likelihood, which means that such computations for least-squares can be orders of magnitude faster whenever $n \gg Kd^2$, which is the case in the setting considered in our experiments. For instance, experiments used to produce Figures 3 and 4 for $T = 20,000$ use about $n \approx 500,000$ events, and $d = 30, K = 3$. Note that, however, the least-squares approach considered here does not scale with respect to $d$ because of its $O(d^3)$ complexity, we recommend to use instead the negative log-likelihood whenever $d$ is large (larger than 1000, say). The complexity of each operation is described in Table 1 below and a numerical illustration of this complexity is displayed in Figure 5, which confirms that computations with least-squares are orders of magnitude faster than with log-likelihood in the considered setting. We don't provide proofs for these complexities, since it follows straightforward arguments, however details about this can be found in Chapter 2 of Bompaire (2018).

Another important point is related to smoothness properties: the negative log-likelihood does not satisfy the gradient-Lipschitz assumption, while this property is required by most first order optimization algorithms to obtain convergence guarantees and an easy tuning of the step-size used in gradient descent. Therefore, for the negative log-likelihood, convergence can be very unstable, while on the contrary, least-squares is gradient Lipschitz and is easy to optimize since it is a quadratic function. Note that in Bompaire et al. (2018) is proposed an alternative approach based on duality, in particular for the negative log-likelihood of the Hawkes

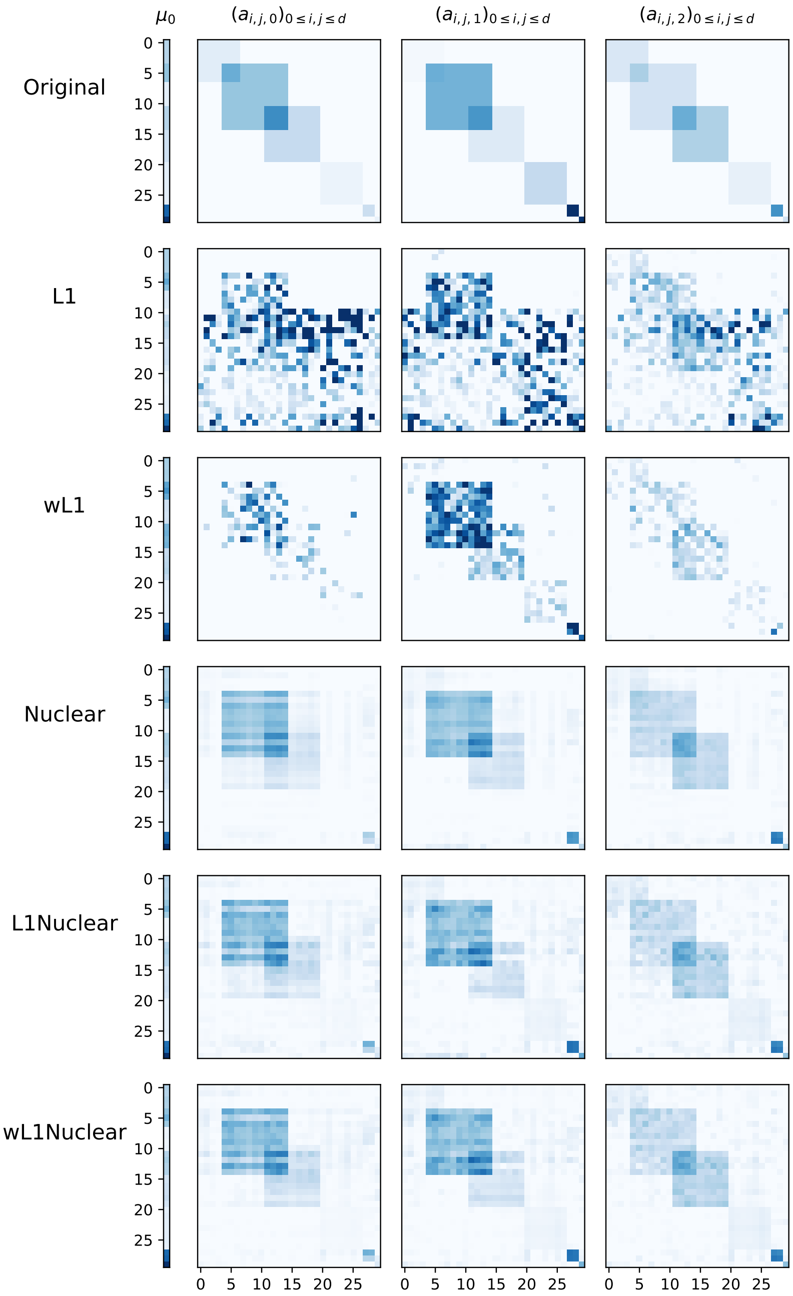

Figure 3: Ground truth tensor $\mathbb{A}$ and recovered tensors using all procedures. We observe that wL1 and wL1Nuclear leads to a much better support recovery, since we observe less false positives outside of the node communities.

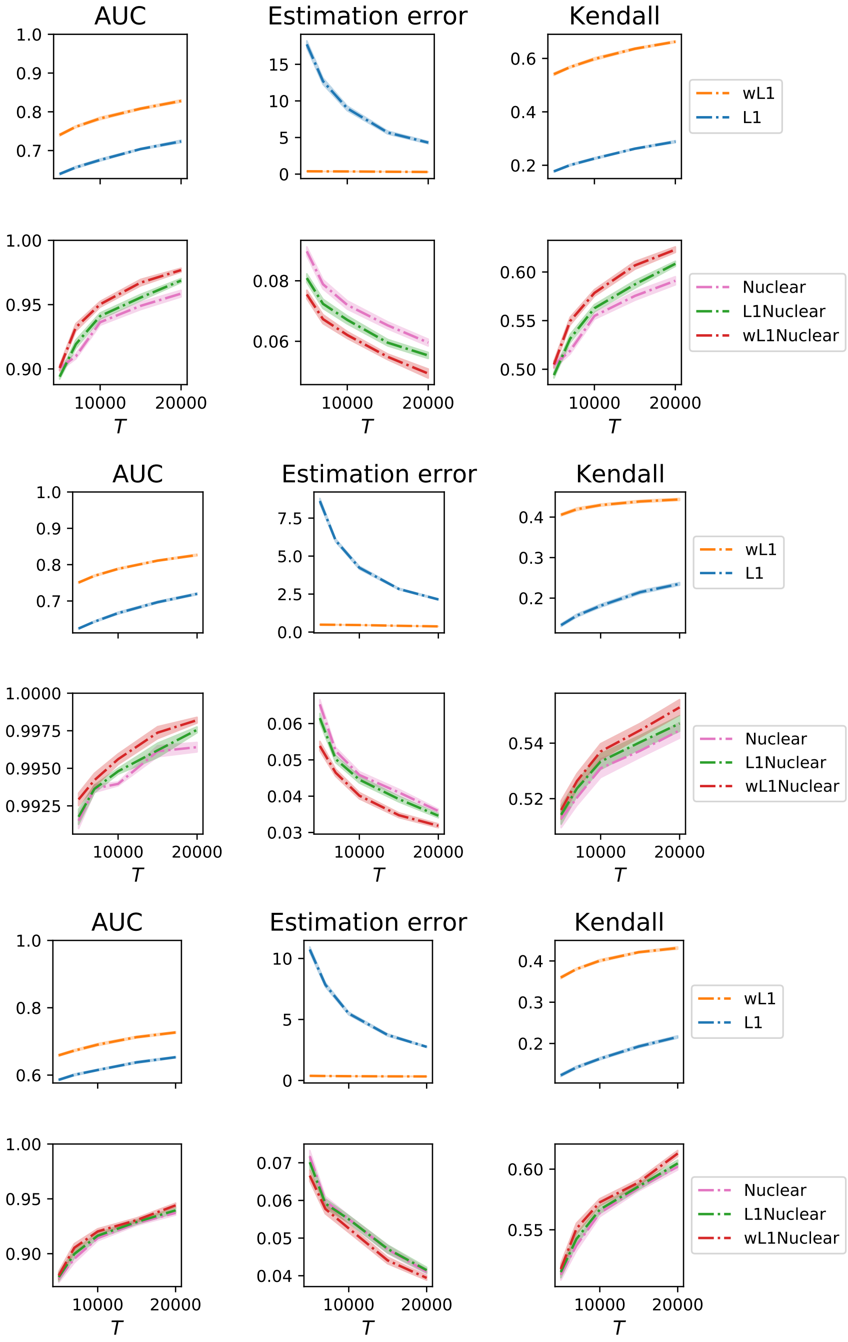

Figure 4: Average metrics achieved by all procedures on the three considered examples of $\mathbb{A}$ (in the same order as the display from Figure 2), and 95% confidence bands, with increasing observation length $T$ over repeated simulations. Weighted penalizations systematically lead to improvements over L1, Nuclear and L1 + Nuclear penalization.

| | pre-computation | memory | value | gradient |
|---|---|---|---|---|
| Least squares | $O(nK^2d)$ | $O(K^2d^3)$ | $O(K^2d^3)$ | $O(K^2d^3)$ |
| Likelihood | $O(nKd)$ | $O(nKd)$ | $O(nKd)$ | $O(nKd)$ |

Table 1: From left to right: Weights precomputation complexity, memory storage, value and gradient complexity for both functionals. Note that for least-squares, the complexity of the value and the gradient with precomputed weights is independent on the number of events $n$.

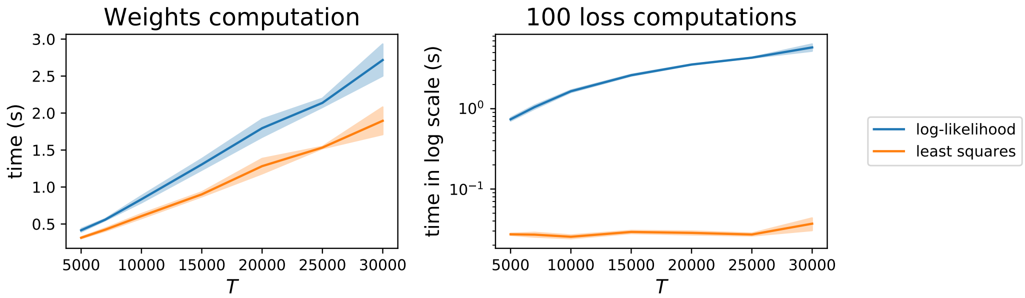

Figure 5: Average time needed for weights (left) and value computation (right) (and 95% confidence bands) for least squares and log-likelihood with precomputations, over repeated simulations. We observe that value computations are order of magnitude faster for least-squares ($y$-scale is logarithmic on the right hand side) and constant with an increasing observation length, while it is strongly increasing for the log-likelihood.

process. Herein one can observe the strong instability of standard first order algorithms (such as the one considered here) for the negative log-likelihood.

In Figure 6 below, we compare the performances of ISTA and FISTA with linesearch for automatic step-size tuning, both for least-squares and negative log-likelihood. This figure confirms that the number of iterations required for least-squares is much smaller than for the negative log-likelihood. This gap is even stronger if we look at the computation times, since each iteration is computationally faster with least squares, and even more so when the observation length increases.

In this Section, we compared least-squares and log-likelihood for the Hawkes process through a computational perspective only, and concluded that least-squares is typically order of magnitude faster. Now, let us compare the statistical performances of both approaches on the same simulation setting as before, with $T = 20,000$, using the metrics defined above, namely Estimation Error, AUC and Kendall. We simply use for this L1 penalization on $\mathbb{A}$, with a strength parameter tuned for each metric and for each goodness-of-fit.

In Figure 7, we observe that both functionals roughly achieve the same performance measured by the Kendall coefficient, but that the negative log-likelihood achieves a slightly better AUC and estimation error than least-squares, at a stronger computational cost. The slightly better statistical performance of maximum likelihood is not surprising, since vanilla maximum likelihood is known to be statistically efficient asymptotically for Hawkes processes, see Ogata (1978), while up to our knowledge, vanilla least-squares estimator is not. This leads to the conclusion that least squares are a very good alternative to maximum likelihood when dealing with a large number of events: statistical accuracy is only slightly deteriorated, but the computational cost is order of magnitudes smaller, and convergence is much more stable.

In Figure 8, we observe the performances achieved by $\ell_1$ versus weighted-$\ell_1$ for the estimators based on the log-likelihood functional. The point here is that we use the weights $\hat{\mathbb{W}}$ from Equation (23) that are derived for the least-squares functional. We observe that, however, these data-driven weights allow to strongly improve over the vanilla $\ell_1$-penalization for the negative log-likelihood estimator as well. This behavior is

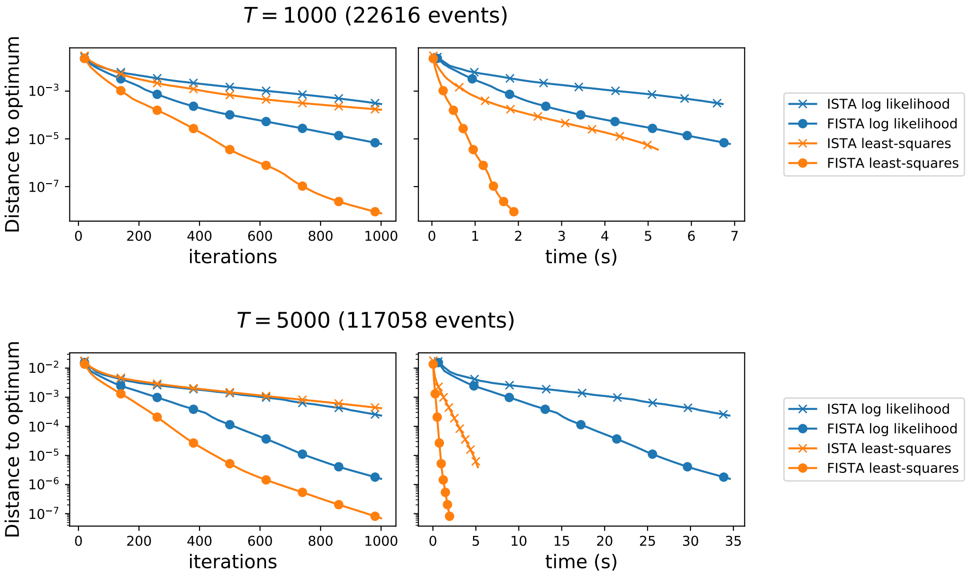

Figure 6: Convergence speed of least squares and likelihood losses with ISTA and FISTA optimization algorithms on two simulations of a Hawkes process with parameters from Figure 2 with observation length $T = 1000$ (top) and $T = 5000$ (bottom). Once again, we observe that the computations are much faster with least-squares, in particular with a large observation length.

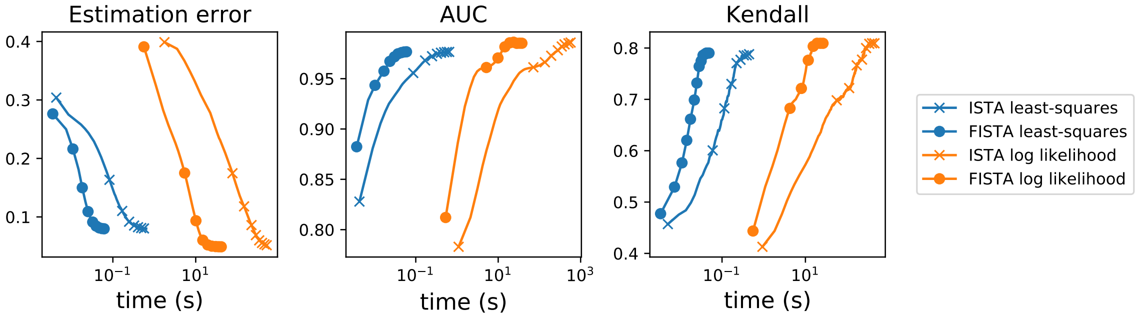

Figure 7: Metrics achieved by least squares and log-likelihood estimators after precomputations. We observe that log-likelihood achieves a slightly better AUC and Estimation Error, but at a stronger computational cost ($x$-axis are on a logarithmic scale).

actually expected, since both functionals are actually close to each other, and the least-squares functional can even be understood as an approximation of the negative log-likelihood one, see Bacry et al. (2016a).

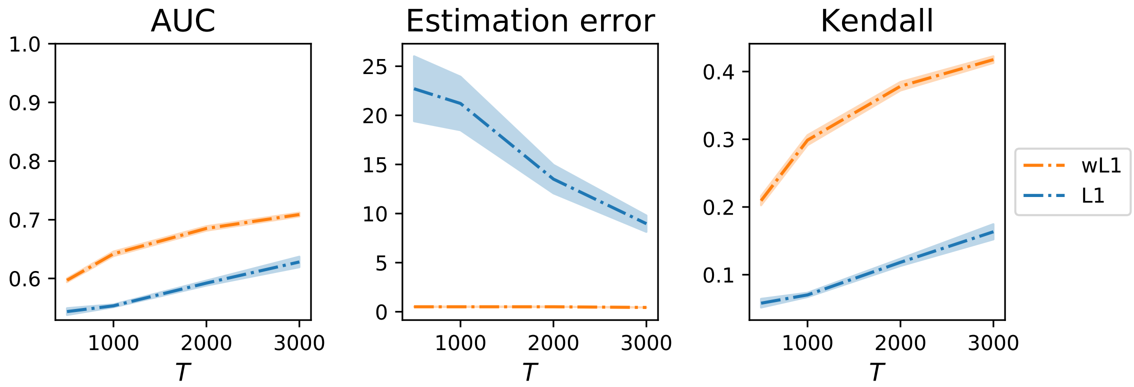

Figure 8: Performances of $\ell_1$ versus weighted-$\ell_1$ for estimators based on the negative log-likelihood functional, where the data-driven weights used in the $\ell_1$ penalization are the ones derived for the least-squares functional. We observe that these weights allow to improve significantly the performances of $\ell_1$-penalized estimators based on the log-likelihood functional, for all the considered metrics. This is expected, since both functionals are actually close to each other.

### 6.5. Sensitivity to the penalization level and weights

In Figure 9, we display the values of the metrics as a function of the penalization level used, both for unweighted and weighted $\ell_1$ penalization. We observe that the weighted $\ell_1$-penalization is more sensitive to its unweighted counterpart, but leads anyway to much better performances even if the penalization level is not perfectly tuned.

In Figure 10 we display the weights $\hat{\mathbb{W}}$ from Equation (23) used in the weighted-$\ell_1$ penalization for a single simulation from the first setting (corresponding to tensor $\mathbb{A}$ displayed in the first row of Figure 2). We observe that these weights are far from being uniform, and effectively induce a strongly varying scaling across kernels $k = 1, 2, 3$ and between nodes. Although this display is hard to interpret, it can be better understood when looked together with the first row of Figure 2: we observe a similarly looking block structure, which means that these weights scale the penalization level roughly following the block structure of the adjacency matrix $\mathbb{A}$ and the intensity of the baseline vector $\mu$.

## 7. Conclusion

In this paper we proposed a careful analysis of the generalization error of the multivariate Hawkes process. Our theoretical analysis required a new concentration inequality for matrix-martingales in continuous time, with an observable variance term, which is a result of independent interest. This analysis led to a new data-driven tuning of sparsity-inducing penalizations, that we assessed on a numerical example. Future works will focus on other theoretical results for non-convex matrix factorization techniques applied to this problem.

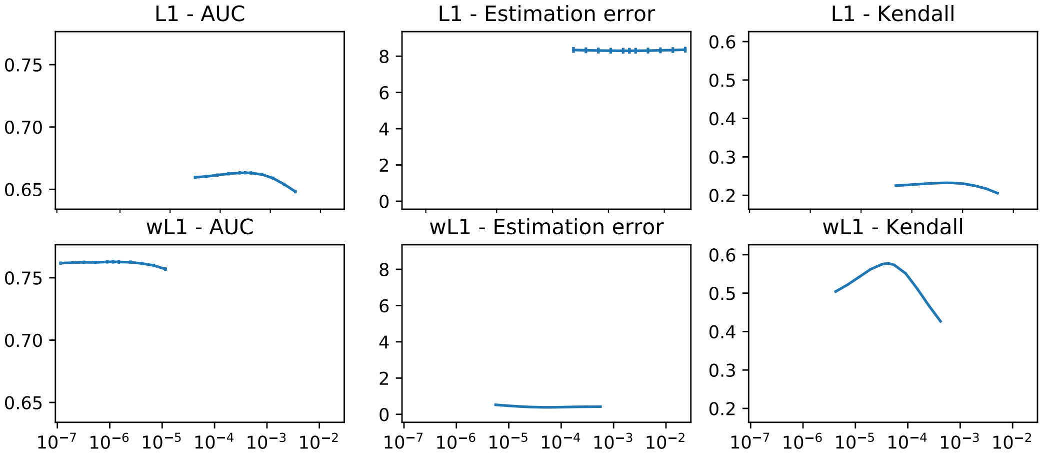

Figure 9: Sensitivity of the metrics (top: AUC, middle: Estimation error, bottom: Kendall) with respect to the penalization level both for unweighted (left-hand side) and weighted (right-hand side) $\ell_1$ penalizations. Weighted $\ell_1$-penalization is more sensitive to its unweighted counterpart, but leads to much better performances even if not perfectly tuned.

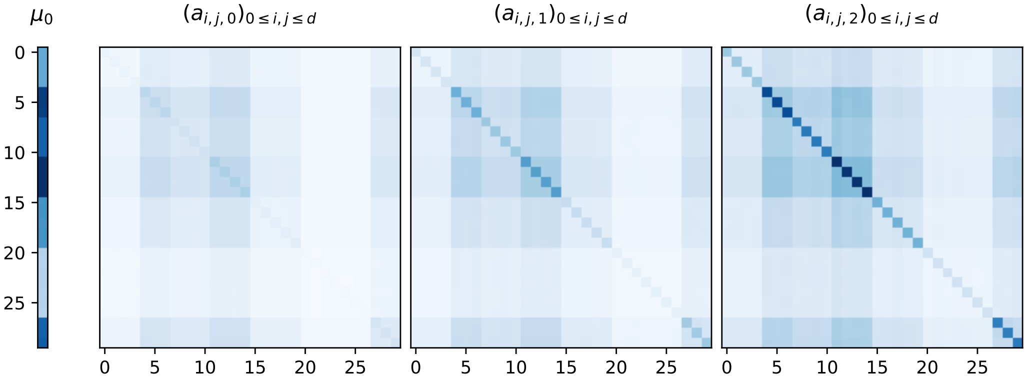

Figure 10: Visualization of the weights used in the weighted-$\ell_1$ penalization for a single simulation from the first setting (corresponding to tensor $\mathbb{A}$ displayed in the first row of Figure 2). This corresponds to the weights from Equation (23), namely $\hat{\mathbb{W}}_{\bullet,\bullet,1}$ (left), $\hat{\mathbb{W}}_{\bullet,\bullet,2}$ (middle) and $\hat{\mathbb{W}}_{\bullet,\bullet,3}$ (right).

## 8. Proofs

This Section contains the proofs of all the results given in the paper. First, we prove the statements concerned with deviation inequalities, namely Theorems 1, 3, Proposition 2 and Theorem 4. Then, we give the proof of Theorem 6, concerning the oracle inequality for the procedure.

### 8.1. Proof of Theorem 1

In Bacry et al. (2016b), a deviation inequality is proven in a slightly more general setting than the one considered in this paper. There are mainly two differences.

- This paper considers only counting processes with uniform jumps of size 1 whereas in Bacry et al. (2016b), jump sizes are controlled by a predictable process $\boldsymbol{J}$. Therefore, it suffices to set $\boldsymbol{J} = \mathbf{1}$ and $\boldsymbol{C}_s = \mathbf{1}$ in Equations (2) and (3) of Bacry et al. (2016b), where $\mathbf{1}$ stands for the all-ones matrices with relevant shapes.
- In Bacry et al. (2016b), the deviation inequality is proved in a general context where no symmetry is assumed on $\mathbb{T}_s$. It forces to consider a symmetric version of $\boldsymbol{W}_{\mathbb{T}}(s)$ as in Eq. (9) increasing the dimension of the working space by a factor of 2, which leads to less precise deviation inequality. In this paper we consider both cases, symmetric and non symmetric, in order to obtain slightly better constants (see the definition of $K_{m,n}$).

With those two differences in mind, following carefully the proof of the concentration inequality in Bacry et al. (2016b) (see the beginning of Appendix B.1 herein) one gets

$$\mathbb{P}\Big[\frac{\lambda_{\max}(\mathscr{S}(\boldsymbol{Z}_t))}{b} \geq \frac{1}{\xi}\lambda_{\max}\Big(\int_0^t \frac{\phi\big(\xi J_{\max}\|\boldsymbol{C}_s\|_\infty \max(\|\mathbb{T}_s\|_{\mathrm{op};\infty}, \|\mathbb{T}_s^\top\|_{\mathrm{op};\infty})b^{-1}\big)}{J_{\max}^2\|\boldsymbol{C}_s\|_\infty^2 \max(\|\mathbb{T}_s\|_{\mathrm{op};\infty}^2, \|\mathbb{T}_s^\top\|_{\mathrm{op};\infty}^2)}\boldsymbol{W}_s ds\Big) + \frac{x}{\xi},$$
$$b_{\mathbb{T}}(t) \leq b\Big] \leq (m+n)e^{-x},$$

where $\xi \in (0,3)$ and $\lambda_{\max}(\mathscr{S}(\boldsymbol{Z}_t)) = \|\boldsymbol{Z}\|_{\mathrm{op}}$ (see the beginning of Appendix B.1 in Bacry et al. (2016b)). Setting $\boldsymbol{J} = \mathbf{1}$, $\boldsymbol{C} = \mathbf{1}$ and taking care of the symmetric case at the same time as the non symmetric one, one gets:

$$\mathbb{P}\Big[\frac{\|\boldsymbol{Z}_t\|_{\mathrm{op}}}{b} \geq \frac{1}{\xi}\lambda_{\max}\Big(\int_0^t \frac{\phi\big(\xi \max(\|\mathbb{T}_s\|_{\mathrm{op};\infty}, \|\mathbb{T}_s^\top\|_{\mathrm{op};\infty})b^{-1}\big)}{\max(\|\mathbb{T}_s\|_{\mathrm{op};\infty}^2, \|\mathbb{T}_s^\top\|_{\mathrm{op};\infty}^2)}\boldsymbol{W}_s ds\Big) + \frac{x}{\xi},$$
$$b_{\mathbb{T}}(t) \leq b\Big] \leq K_{m,n}e^{-x},$$

using the definitions $K_{m,n}$ and $\boldsymbol{W}_s$ introduced previously (depending on the symmetric properties of the tensor $\mathbb{T}_s$). Let us note that on $\{b_{\mathbb{T}}(t) \leq b\}$ one has $\max(\|\mathbb{T}_s\|_{\mathrm{op};\infty}, \|\mathbb{T}_s^\top\|_{\mathrm{op};\infty})b^{-1} \leq 1$ for any $s \in [0,t]$. Thus, since $\phi(xh) \leq h^2\phi(x)$ for any $h \in [0,1]$ and $x > 0$, one gets

$$\mathbb{P}\Big[\frac{\|\boldsymbol{Z}_t\|_{\mathrm{op}}}{b} \geq \frac{\phi(\xi)}{\xi b^2}\lambda_{\max}\Big(\int_0^t \boldsymbol{W}_s ds\Big) + \frac{x}{\xi}\ ,\quad b_{\mathbb{T}}(t) \leq b\Big] \leq K_{m,n}e^{-x}$$

and finally

$$\mathbb{P}\Big[\|\boldsymbol{Z}_t\|_{\mathrm{op}} \geq \frac{\phi(\xi)}{\xi b}\lambda_{\max}(\boldsymbol{V}_t) + \frac{xb}{\xi}\ ,\quad b_{\mathbb{T}}(t) \leq b\Big] \leq K_{m,n}e^{-x}$$

which proves the first part of the Theorem. The second part (i.e., Inequality (13)) can be obtained following some standard tricks (see e.g. Massart (2007)):

(i) on $(0,3)$, $\phi(\xi) \leq \frac{\xi^2}{2(1-\xi/3)}$ and

(ii) $\min_{\xi \in (0,1/c)} \left( \frac{a\xi}{1-c\xi} + \frac{x}{\xi} \right) = 2\sqrt{ax} + cx$ for any $a, c, x > 0$.

Thus applying (i) leads to

$$\mathbb{P}\bigg[ \|\boldsymbol{Z}_t\|_{\text{op}} \geq \frac{\xi}{2b(1-\xi/3)} \lambda_{\max}(\boldsymbol{V}_t) + \frac{xb}{\xi} \ , \quad b_{\mathbb{T}}(t) \leq b \bigg] \leq K_{m,n} e^{-x}$$

or equivalently

$$\mathbb{P}\bigg[ \|\boldsymbol{Z}_t\|_{\text{op}} \geq \frac{\xi}{2b(1-\xi/3)} v + \frac{xb}{\xi} \ , \quad \lambda_{\max}(\boldsymbol{V}_t) \leq v, \ b_{\mathbb{T}}(t) \leq b \bigg] \leq K_{m,n} e^{-x}.$$

Then optimizing on $\xi$ using (ii) with $c = 1/3$ and $a = v/2b^2$, one gets

$$\mathbb{P}\bigg[ \|\boldsymbol{Z}_t\|_{\text{op}} \geq \sqrt{2vx} + \frac{xb}{3} \ , \quad \lambda_{\max}(\boldsymbol{V}_t) \leq v, \ b_{\mathbb{T}}(t) \leq b \bigg] \leq K_{m,n} e^{-x}$$

which concludes the proof of Theorem 1.

### 8.2. Proof of Proposition 2

This Proposition provides a deviation between $\lambda_{\max}(\boldsymbol{V}(t))$ and $\lambda_{\max}(\widehat{\boldsymbol{V}}(t))$. Let us notice that it is a generalization to arbitrary matrices of dimensions $m \times n$ of an analog inequality originally proven by Hansen et al. (2012) for scalar martingales (i.e., in dimension 1). The proof below follows the same lines as these authors. The proof is based on the observation that the difference $\boldsymbol{V}_{\mathbb{T}}(t) - \widehat{\boldsymbol{V}}_{\mathbb{T}}(t)$ can be written as a martingale $\boldsymbol{Z}_{\mathbb{H}}(t)$

$$\boldsymbol{V}_{\mathbb{T}}(t) - \widehat{\boldsymbol{V}}_{\mathbb{T}}(t) = \boldsymbol{Z}_{\mathbb{H}}(t) = \int_0^t \mathbb{H}_s \circ d\boldsymbol{M}_s,$$

where

$$\mathbb{H}_s = \mathbb{T}_s^2 \tag{25}$$

when $\mathbb{T}_s$ is symmetric, while

$$\mathbb{H}_s = \begin{bmatrix} \mathbb{T}_s \mathbb{T}_s^\top & \boldsymbol{0} \\ \boldsymbol{0} & \mathbb{T}_s^\top \mathbb{T}_s \end{bmatrix} \tag{26}$$

if $\mathbb{T}_s$ is not symmetric. Then applying Eq. (12) of Theorem 1 to the martingale $\boldsymbol{Z}_{\mathbb{H}}(t)$ (we are in the symmetric case of the Theorem since $\mathbb{H}_s^\top = \mathbb{H}_s$), one gets

$$\mathbb{P}\bigg[ \|\boldsymbol{Z}_{\mathbb{H}}(t)\|_{\text{op}} \geq \frac{\phi(\xi)}{\xi b} \lambda_{\max}\big(\boldsymbol{V}_{\mathbb{H}}(t)\big) + \frac{xb}{\xi}, \quad b_{\mathbb{H}}(t) \leq b \bigg] \leq K_{m,n} e^{-x}, \tag{27}$$

with

$$\boldsymbol{V}_{\mathbb{H}}(t) = \int_0^t \mathbb{H}_s^2 \circ \boldsymbol{\lambda}_s ds \ . \tag{28}$$

Since

$$\|\boldsymbol{Z}_{\mathbb{H}}(t)\|_{\text{op}} \geq \lambda_{\max}(\boldsymbol{V}_{\mathbb{T}}(t)) - \lambda_{\max}(\widehat{\boldsymbol{V}}_{\mathbb{T}}(t)),$$

we have

$$\mathbb{P}\bigg[ \lambda_{\max}(\boldsymbol{V}_{\mathbb{T}}(t)) \geq \lambda_{\max}(\widehat{\boldsymbol{V}}_{\mathbb{T}}(t)) + \frac{\phi(\xi)}{\xi b} \lambda_{\max}\big(\boldsymbol{V}_{\mathbb{H}}(t)\big) + \frac{xb}{\xi}, \quad b_{\mathbb{H}}(t) \leq b \bigg] \leq K_{m,n} e^{-x}, \tag{29}$$

One can first notice that, from the definitions of $\mathbb{H}$ and $b_{\mathbb{T}}(t)$, one has $b_{\mathbb{H}}(t) \leq b_{\mathbb{T}}^2(t)$. Moreover, since

$$\mathbb{T}_s \mathbb{T}_s^\top \preccurlyeq b_{\mathbb{T}}^2(s) \boldsymbol{I}_m \quad \text{and} \quad \mathbb{T}_s^\top \mathbb{T}_s \preccurlyeq b_{\mathbb{T}}^2(s) \boldsymbol{I}_n$$

for all $s$, we have from Eq. (28),

$$\boldsymbol{V}_{\mathbb{H}}(t) \preccurlyeq b_{\mathbb{T}}^2(t)\boldsymbol{V}_{\mathbb{T}}(t) \tag{30}$$

and therefore

$$\lambda_{\max}(\boldsymbol{V}_{\mathbb{H}}(t)) \le b_{\mathbb{T}}^2(t)\lambda_{\max}(\boldsymbol{V}_{\mathbb{T}}(t)).$$

Inequality (29) then gives:

$$\mathbb{P}\bigg[\lambda_{\max}(\boldsymbol{V}_{\mathbb{T}}(t)) \ge \lambda_{\max}(\widehat{\boldsymbol{V}}_{\mathbb{T}}(t)) + \frac{\phi(\xi)}{\xi}\lambda_{\max}(\boldsymbol{V}_{\mathbb{T}}(t)) + \frac{xb^2}{\xi}, \quad b_{\mathbb{T}}(t) \le b\bigg] \le K_{m,n}e^{-x}, \tag{31}$$

and thus

$$\mathbb{P}\bigg[\lambda_{\max}(\boldsymbol{V}_{\mathbb{T}}(t)) \ge \frac{\xi\lambda_{\max}(\widehat{\boldsymbol{V}}_{\mathbb{T}}(t))}{\xi - \phi(\xi)} + \frac{xb^2}{\xi - \phi(\xi)}, \quad b_{\mathbb{T}}(t) \le b\bigg] \le K_{m,n}e^{-x}, \tag{32}$$

which proves the first inequality stated in Proposition 2. Now, an easy computation proves that the choice $\xi = -W_{-1}(-\frac{2}{3}e^{-2/3}) - 2/3 \approx 0.762$ provides the second desired inequality. □

### 8.3. Proof of Theorem 3

Introduce the set

$$E_t = \{\lambda_{\max}(\boldsymbol{V}_{\mathbb{T}}(t)) \le 2\lambda_{\max}(\widehat{\boldsymbol{V}}_{\mathbb{T}}(t)) + 2.62b^2x\}.$$

We know from Proposition 2 that $\mathbb{P}[E_t^{\complement}, b_{\mathbb{T}}(t) \le b] \le K_{m,n}e^{-x}$. Now, on the set

$$E_t \cap \{\lambda_{\max}(\widehat{\boldsymbol{V}}_{\mathbb{T}}(t)) \le v\} \cap \{b_{\mathbb{T}}(t) \le b\}$$

we have

$$\frac{\phi(\xi)}{\xi b}\lambda_{\max}(\boldsymbol{V}(t)) + \frac{xb}{\xi} \le \frac{\phi(\xi)}{\xi b}2v + \frac{bx}{\xi} + \frac{2.62\phi(3)}{3}bx$$

for any $\xi \in (0,3)$, since $\xi \mapsto \phi(\xi)/\xi$ is increasing. Using again points (i) and (ii) from Section 8.1 proves that the minimum for $\xi \in (0,3)$ of the right hand size of this last inequality is equal to

$$2\sqrt{vx} + \frac{2.62\phi(3)+1}{3}xb \le 2\sqrt{vx} + cxb$$

with $c = 14.39$. Now, the conclusion easily follows from the following decomposition:

$$\begin{aligned}
&\mathbb{P}\bigg[\|\boldsymbol{Z}_{\mathbb{T}}(t)\|_{\mathrm{op}} \ge 2\sqrt{vx} + cbx, \ \ \lambda_{\max}(\widehat{\boldsymbol{V}}_{\mathbb{T}}(t)) \le v, \ \ b_{\mathbb{T}}(t) \le b\bigg] \\
&\quad \le \mathbb{P}[E_t^{\complement}, b_{\mathbb{T}}(t) \le b] + \mathbb{P}\bigg[\|\boldsymbol{Z}_{\mathbb{T}}(t)\|_{\mathrm{op}} \ge 2\sqrt{vx} + cbx, \ \ E_t, \ \ \lambda_{\max}(\widehat{\boldsymbol{V}}_{\mathbb{T}}(t)) \le v, \ \ b_{\mathbb{T}}(t) \le b\bigg] \\
&\quad \le K_{m,n}e^{-x} + \mathbb{P}\bigg[\|\boldsymbol{Z}_t\|_{\mathrm{op}} \ge \frac{\xi}{2b(1-\xi/3)}\lambda_{\max}(\boldsymbol{V}_t) + \frac{xb}{\xi} \ , \ \ b_{\mathbb{T}}(t) \le b\bigg] \\
&\quad \le 2K_{m,n}e^{-x},
\end{aligned}$$

where we used Equation (12) from Theorem 1 in the last inequality.

### 8.4. Proof of Theorem 4

In order to prove this theorem, we are going to use peeling arguments. For any $\epsilon > 0$ and $z > 0$ we define the interval

$$\mathcal{I}_{z,\varepsilon} = [z, z(1+\varepsilon)].$$

Let, $v_0, b_0, \epsilon > 0$ and let us define $v_j = v_0(1+\varepsilon)^j$, $b_j = b_0(1+\varepsilon)^j$. Let us define also the events

$$V_{-1} = \{\lambda_{\max}(\widehat{\boldsymbol{V}}_{\mathbb{T}}(t)) \leq v_0\}, \quad B_{-1} = \{b_{\mathbb{T}}(t) \leq b_0\},$$

and

$$V_j = \{\lambda_{\max}(\widehat{\boldsymbol{V}}_{\mathbb{T}}(t)) \in \mathcal{I}_{v_j,\varepsilon}\}, \quad B_j = \{b_{\mathbb{T}}(t) \in \mathcal{I}_{b_j,\varepsilon}\}$$

for any $j \in \mathbb{N}$. We set $v_0 = w_0 x$, then, from Equation (14), one gets successively

$$\mathbb{P}\Big[\|\boldsymbol{Z}_{\mathbb{T}}(t)\|_{\mathrm{op}} \geq x\big(2\sqrt{w_0} + cb_0\big), V_{-1} \cap B_{-1}\Big] \leq 2K_{m,n}e^{-x}$$

$$\mathbb{P}\Big[\|\boldsymbol{Z}_{\mathbb{T}}(t)\|_{\mathrm{op}} \geq x\big(2\sqrt{w_0} + c(1+\varepsilon)b_{\mathbb{T}}(t)\big), V_{-1} \cap B_j\Big] \leq 2K_{m,n}e^{-x}$$

$$\mathbb{P}\Big[\|\boldsymbol{Z}_{\mathbb{T}}(t)\|_{\mathrm{op}} \geq 2\sqrt{\lambda_{\max}(\widehat{\boldsymbol{V}}_{\mathbb{T}}(t))(1+\varepsilon)x} + cxb_0, V_i \cap B_{-1}\Big] \leq 2K_{m,n}e^{-x}$$

$$\mathbb{P}\Big[\|\boldsymbol{Z}_{\mathbb{T}}(t)\|_{\mathrm{op}} \geq 2\sqrt{\lambda_{\max}(\widehat{\boldsymbol{V}}_{\mathbb{T}}(t))(1+\varepsilon)x} + c(1+\varepsilon)xb_{\mathbb{T}}(t), V_i \cap B_j\Big] \leq 2K_{m,n}e^{-x}$$

for all $i, j \geq 0$. If one denotes $A = 2\sqrt{w_0}/c + b_0$, previous inequalities entail, for any $i, j \geq -1$:

$$\mathbb{P}\Big[\|\boldsymbol{Z}_{\mathbb{T}}(t)\|_{\mathrm{op}} \geq 2\sqrt{\lambda_{\max}(\widehat{\boldsymbol{V}}_{\mathbb{T}}(t))(1+\varepsilon)x} + c(1+\varepsilon)(A + b_{\mathbb{T}}(t))x, V_i \cap B_j\Big] \leq 2K_{m,n}e^{-x}. \tag{33}$$

Let $\alpha > 0$ and define

$$\ell_x(t) = \alpha \log\bigg(\log\Big(\frac{\lambda_{\max}(\widehat{\boldsymbol{V}}_{\mathbb{T}}(t))}{w_0 x}(1+\epsilon)^2 \vee (1+\epsilon)\Big)\bigg) + \alpha\log\bigg(\log\Big(\frac{b_{\mathbb{T}}(t)}{b_0}(1+\epsilon)^2 \vee (1+\epsilon)\Big)\bigg). \tag{34}$$

Since, $\forall i, j \geq -1$, $\lambda_{\max}(\widehat{\boldsymbol{V}}_{\mathbb{T}}(t)) \geq xw_0(1+\varepsilon)^i(1-\delta_{-1,i})$ and $b_{\mathbb{T}}(t) \geq b_0(1+\varepsilon)^j(1-\delta_{-1,j})$ on $V_i \cap B_j$, then one has

$$\ell_x(t) \geq \ell_{i,j} = \log\Big((i+2)^\alpha (j+2)^\alpha (\log(1+\epsilon))^{2\alpha}\Big) \quad \text{on} \quad V_i \cap B_j$$

for any $i, j \geq -1$. Then making the change of variable $x \leftarrow x + \ell_{i,j}$ in (33) gives

$$\begin{aligned}\mathbb{P}\Big[\|\boldsymbol{Z}_{\mathbb{T}}(t)\|_{\mathrm{op}} \geq 2\sqrt{\lambda_{\max}(\widehat{\boldsymbol{V}}_{\mathbb{T}}(t))(1+\varepsilon)(x+\ell_{i,j})} + c(1+\varepsilon)(A + b_{\mathbb{T}}(t))(x+\ell_{i,j}), \quad V_i \cap B_j\Big]\\ \leq 2K_{m,n}e^{-x}e^{-\ell_{i,j}}\end{aligned}$$

and then

$$\begin{aligned}\mathbb{P}\Big[\|\boldsymbol{Z}_{\mathbb{T}}(t)\|_{\mathrm{op}} \geq 2\sqrt{\lambda_{\max}(\widehat{\boldsymbol{V}}_{\mathbb{T}}(t))(1+\varepsilon)(x+\ell_x(t))} + c(1+\varepsilon)(x+\ell_x(t))(A + b_{\mathbb{T}}(t)), \quad V_i \cap B_j\Big]\\ \leq 2K_{m,n}\big[\log(1+\varepsilon)\big]^{-2\alpha}e^{-x}\big[(i+2)(j+2)\big]^{-\alpha}\end{aligned}$$

for any $i, j \geq -1$. Since the whole probability space can be partitioned as $\bigcup_{i,j \in \geq -1} V_i \cap B_j$, one has finally

$$\begin{aligned}&\mathbb{P}\Big[\|\boldsymbol{Z}_{\mathbb{T}}(t)\|_{\mathrm{op}} \geq 2\sqrt{\lambda_{\max}(\widehat{\boldsymbol{V}}_{\mathbb{T}}(t))(1+\varepsilon)(x+\ell_x(t))} + c(1+\varepsilon)(x+\ell_x(t))(A + b_{\mathbb{T}}(t))\Big]\\ &= \sum_{i,j=-1}^{\infty}\mathbb{P}\Big[\|\boldsymbol{Z}_{\mathbb{T}}(t)\|_{\mathrm{op}} \geq 2\sqrt{\lambda_{\max}(\widehat{\boldsymbol{V}}_{\mathbb{T}}(t))(1+\varepsilon)(x+\ell_x(t))}\\ &\qquad + c(1+\varepsilon)(x+\ell_x(t))(A + b_{\mathbb{T}}(t)), \quad V_i \cap B_j\Big]\\ &\leq 2K_{m,n}\big[\log(1+\varepsilon)\big]^{-2\alpha}\big(\sum_{i=1}^{\infty} i^{-\alpha}\big)^2 e^{-x}.\end{aligned}$$

Finally, choosing $\epsilon = b_0 = w_0 = 1$ and $\alpha = 2$ leads to Equation (15) and concludes the proof of the Theorem.

### 8.5. Proof of Theorem 6

If $A, B$ are vectors, matrices or tensors of matching dimensions, we denote by $A \odot B$ their entrywise product (Hadamard product). We recall also that $\boldsymbol{A}_{j,\bullet}$ the $j$-th row of a matrix $\boldsymbol{A}$ and recall that $\|\boldsymbol{A}\|_{\infty,2} = \max_j \|\boldsymbol{A}_{j,\bullet}\|_2$. The proof is based on the proof of a sharp oracle inequality for trace norm penalization, see Koltchinskii et al. (2011) and Koltchinskii (2011). We endow the space $\mathbb{R}^d \times \mathbb{R}^{d\times d\times K}$ with the inner product

$$\langle \theta, \theta' \rangle = \langle \mu, \mu' \rangle + \langle \mathbb{A}, \mathbb{A}' \rangle,$$

where $\theta = (\mu, \mathbb{A})$ and $\theta' = (\mu', \mathbb{A}')$ with $\langle \mu, \mu' \rangle = \mu^\top \mu'$ and

$$\langle \mathbb{A}, \mathbb{A}' \rangle = \sum_{\substack{1 \le j,j' \le d \\ 1 \le k \le K}} \mathbb{A}_{j,j',k} \mathbb{A}'_{j,j',k}.$$

We denote for short $a_{j,j',k} = \mathbb{A}_{j,j',k}$. For any $\theta$, one has

$$\langle \nabla R_T(\hat\theta), \hat\theta - \theta \rangle = 2 \sum_{1 \le j \le d} (\hat\mu_j - \mu_j) \frac{\partial R_T(\hat\theta)}{\partial \hat\mu_j} + \sum_{\substack{1 \le j,j' \le d \\ 1 \le k \le K}} (\hat a_{j,j',k} - a_{j,j',k}) \frac{\partial R_T(\hat\theta)}{\partial \hat a_{j,j',k}}.$$

Let us recall that $H_{j,j',k}(t) = \int_{(0,t)} h_{j,j',k}(t-s) dN_{j'}(s)$. Since

$$\frac{\partial \lambda_{j,\theta}(t)}{\partial \mu_j} = 1 \quad \text{and} \quad \frac{\partial \lambda_{j,\theta}(t)}{\partial a_{j,j',k}} = H_{j,j',k}(t),$$

we have that the derivatives of the empirical risk are given by

$$\frac{\partial R_T(\hat\theta)}{\partial \mu_j} = \frac{2}{T}\Big( \int_0^T \lambda_{j,\theta}(t) dt - \int_0^T dN_j(t) \Big)$$

and

$$\frac{\partial R_T(\hat\theta)}{\partial a_{j,j',k}} = \frac{2}{T}\Big( \int_0^T H_{j,j',k}(t) \lambda_{j,\theta}(t) dt - \int_0^T H_{j,j',k}(t) dN_j(t) \Big).$$

It leads to

$$\begin{aligned}
\langle \nabla R_T(\hat\theta), \hat\theta - \theta \rangle &= \frac{2}{T} \sum_{j=1}^d \int_0^T (\lambda_{j,\hat\theta}(t) - dN_j(t))(\hat\mu_j - \mu_j) \\
&\quad + \frac{2}{T} \sum_{\substack{1 \le j,j' \le d \\ 1 \le k \le K}} \int_0^T H_{j,j',k}(t)(\lambda_{j,\hat\theta}(t) - dN_j(t))(\hat a_{j,j',k} - a_{j,j',k}) \\
&= \frac{2}{T} \sum_{j=1}^d \int_0^T (\lambda_{j,\hat\theta}(t) - \lambda_{j,\theta}(t))(\lambda_{j,\hat\theta}(t) dt - dN_j(t)).
\end{aligned}$$

Let us remind that $M_j(t) = N_j(t) - \int_0^t \lambda_j(s) ds$ are martingales coming from the Doob-Meyer decomposition, so that $dM_j(t) = dN_j(t) - \lambda_j(t) dt$. So, recalling that

$$\langle f, g \rangle_T = \frac{1}{T} \sum_{1 \le j \le d} \int_{[0,T]} f_j(t) g_j(t) dt,$$

we obtain the decomposition

$$\langle \nabla R_T(\hat\theta), \hat\theta - \theta \rangle = 2\langle \lambda_{\hat\theta} - \lambda_\theta, \lambda_{\hat\theta} - \lambda \rangle_T - \frac{2}{T} \sum_{j=1}^d \int_0^T (\lambda_{j,\hat\theta}(t) - \lambda_{j,\theta}(t)) dM_j(t).$$

Namely, we end up with

$$2\langle \lambda_{\hat\theta} - \lambda_\theta, \lambda_{\hat\theta} - \lambda \rangle_T = \langle \nabla R_T(\hat\theta), \hat\theta - \theta \rangle + \frac{2}{T} \sum_{j=1}^d \int_0^T (\lambda_{j,\hat\theta}(t) - \lambda_{j,\theta}(t)) dM_j(t). \tag{35}$$

The parallelogram identity gives

$$2\langle \lambda_{\hat\theta} - \lambda_\theta, \lambda_{\hat\theta} - \lambda \rangle_T = \|\lambda_{\hat\theta} - \lambda\|_T^2 + \|\lambda_{\hat\theta} - \lambda_\theta\|_T^2 - \|\lambda_\theta - \lambda\|_T^2,$$

where we put $\|f\|_T^2 = \langle f, f \rangle_T$. Let us point out that, in the case $\langle \lambda_{\hat\theta} - \lambda_\theta, \lambda_{\hat\theta} - \lambda \rangle_T < 0$, one obtains

$$\|\lambda_{\hat\theta} - \lambda\|_T^2 \le \|\lambda_\theta - \lambda\|_T^2,$$

which directly implies the inequality of the Theorem. Thus, from now on, let us assume that

$$\langle \lambda_{\hat\theta} - \lambda_\theta, \lambda_{\hat\theta} - \lambda \rangle_T \ge 0. \tag{36}$$

The first order condition for $\hat\theta \in \operatorname{argmin}_\theta \{R_T(\theta) + \operatorname{pen}(\theta)\}$ gives

$$-\nabla R_T(\hat\theta) \in \partial \operatorname{pen}(\hat\theta).$$

Let $\hat\theta_\partial = -\nabla R_T(\hat\theta)$. Since the subdifferential is a monotone mapping, we have $\langle \hat\theta - \theta, \hat\theta_\partial - \theta_\partial \rangle \ge 0$ for any $\theta_\partial \in \partial \operatorname{pen}(\theta)$. Thus from (35), one gets $\forall \theta_\partial \in \partial \operatorname{pen}(\theta)$,

$$2\langle \lambda_{\hat\theta} - \lambda_\theta, \lambda_{\hat\theta} - \lambda \rangle_T \le -\langle \theta_\partial, \hat\theta - \theta \rangle + \frac{2}{T} \sum_{j=1}^d \int_0^T (\lambda_{j,\hat\theta}(t) - \lambda_{j,\theta}(t)) dM_j(t). \tag{37}$$

We need now to characterize the structure of the subdifferentials involved in $\operatorname{pen}(\theta)$, to describe $\theta_\partial$. If $g_1(\mu) = \sum_{j=1}^d \hat w_j |\mu_j|$, for $\hat w_j \ge 0$, we have

$$\partial g_1(\mu) = \Big\{ \hat w \odot \operatorname{sign}(\mu) + \hat w \odot f : \|f\|_\infty \le 1, \mu \odot f = 0 \Big\}. \tag{38}$$

If $g_2(\mathbb{A}) = \sum_{1 \le j,j' \le d, 1 \le k \le K} \hat{\mathbb{W}}_{j,j',k} |\mathbb{A}_{j,j',k}|$, for $\hat{\mathbb{W}}_{j,j',k} \ge 0$, we have

$$\partial g_2(\mathbb{A}) = \Big\{ \hat{\mathbb{W}} \odot \operatorname{sign}(\mathbb{A}) + \hat{\mathbb{W}} \odot \mathbb{F} : \|\mathbb{F}\|_\infty \le 1, \mathbb{A} \odot \mathbb{F} = 0 \Big\}. \tag{39}$$

Now let $\boldsymbol{A} = \operatorname{hstack}(\mathbb{A})$ and $\hat{\boldsymbol{A}} = \operatorname{hstack}(\hat{\mathbb{A}})$. Let us recall that if $\boldsymbol{A} = \boldsymbol{U\Sigma V}^\top$ is the SVD of $\boldsymbol{A}$, we have $\mathcal{P}_{\boldsymbol{A}}(\boldsymbol{B}) = \boldsymbol{P_U B} + \boldsymbol{B P_V} - \boldsymbol{P_U B P_V}$ and $\mathcal{P}_{\boldsymbol{A}}^\perp(\boldsymbol{B}) = (\boldsymbol{I} - \boldsymbol{P_U})\boldsymbol{B}(\boldsymbol{I} - \boldsymbol{P_V})$ (projection onto the column and row space of $\boldsymbol{A}$ and projection onto its orthogonal space). Now, for $g_3(\boldsymbol{A}) = \hat\tau \|\boldsymbol{A}\|_*$, we have

$$\partial g_3(\boldsymbol{A}) = \Big\{ \hat\tau \boldsymbol{U V}^\top + \hat\tau \mathcal{P}_{\boldsymbol{A}}^\perp(\boldsymbol{F}) : \|\boldsymbol{F}\|_{\mathrm{op}} \le 1 \Big\}, \tag{40}$$

see for instance (Lewis, 1995). Now, write

$$-\langle \theta_\partial, \hat\theta - \theta \rangle = -\langle \mu_\partial, \hat\mu - \mu \rangle - \langle \mathbb{A}_{\partial,1}, \hat{\mathbb{A}} - \mathbb{A} \rangle - \langle \boldsymbol{A}_{\partial,*}, \hat{\boldsymbol{A}} - \boldsymbol{A} \rangle$$

with $\mu_\partial \in \partial g_1(\mu)$, $\mathbb{A}_{\partial,1} \in \partial g_2(\mathbb{A})$ and $\boldsymbol{A}_{\partial,*} \in \partial g_3(\boldsymbol{A})$. Using Equation (38), (39) and (40), we can write

$$\begin{aligned}-\langle \theta_\partial, \hat\theta - \theta\rangle = &-\langle \hat w \odot \operatorname{sign}(\mu), \hat\mu - \mu\rangle - \langle \hat w \odot f, \hat\mu - \mu\rangle \\ &- \langle \hat{\mathbb{W}} \odot \operatorname{sign}(\mathbb{A}), \hat{\mathbb{A}} - \mathbb{A}\rangle - \langle \hat{\mathbb{W}} \odot \mathbb{F}_1, \hat{\mathbb{A}} - \mathbb{A}\rangle \\ &- \hat\tau\langle \boldsymbol{U}\boldsymbol{V}^\top, \hat{\boldsymbol{A}} - \boldsymbol{A}\rangle - \hat\tau\langle \boldsymbol{F}_*, \mathcal{P}_{\boldsymbol{A}}^\perp(\hat{\boldsymbol{A}} - \boldsymbol{A})\rangle,\end{aligned}$$

where by duality between the norms $\|\cdot\|_1$ and $\|\cdot\|_\infty$, and between $\|\cdot\|_*$ and $\|\cdot\|_{\mathrm{op}}$, we can choose $f, \mathbb{F}_1$ and $\boldsymbol{F}_*$ such that

$$\langle \hat w \odot f, \hat\mu - \mu\rangle = \|(\hat\mu - \mu)_{\operatorname{supp}(\mu)^\perp}\|_{1,\hat w}, \quad \langle \hat{\mathbb{W}} \odot \mathbb{F}_1, \hat{\mathbb{A}} - \mathbb{A}\rangle = \|(\hat{\mathbb{A}} - \mathbb{A})_{\operatorname{supp}(\mathbb{A})^\perp}\|_{1,\hat{\mathbb{W}}}$$

and

$$\langle \boldsymbol{F}_*, \mathcal{P}_{\boldsymbol{A}}^\perp(\hat{\boldsymbol{A}} - \boldsymbol{A})\rangle = \|\mathcal{P}_{\boldsymbol{A}}^\perp(\hat{\boldsymbol{A}} - \boldsymbol{A})\|_*,$$

which leads to

$$\begin{aligned}-\langle \theta_\partial, \hat\theta - \theta\rangle \le &\|(\hat\mu - \mu)_{\operatorname{supp}(\mu)}\|_{1,\hat w} - \|(\hat\mu - \mu)_{\operatorname{supp}(\mu)^\perp}\|_{1,\hat w} \\ &+ \|(\hat{\mathbb{A}} - \mathbb{A})_{\operatorname{supp}(\mathbb{A})}\|_{1,\hat{\mathbb{W}}} - \|(\hat{\mathbb{A}} - \mathbb{A})_{\operatorname{supp}(\mathbb{A})^\perp}\|_{1,\hat{\mathbb{W}}} \\ &+ \hat\tau\|\mathcal{P}_{\boldsymbol{A}}(\hat{\boldsymbol{A}} - \boldsymbol{A})\|_* - \hat\tau\|\mathcal{P}_{\boldsymbol{A}}^\perp(\hat{\boldsymbol{A}} - \boldsymbol{A})\|_*.\end{aligned}$$

Now, we decompose the noise term of (37):

$$\begin{aligned}&\frac{2}{T}\sum_{j=1}^d \int_0^T (\lambda_{j,\hat\theta}(t) - \lambda_{j,\theta}(t)) dM_j(t) \\ &= \frac{2}{T}\sum_{j=1}^d (\hat\mu_j - \mu_j)\int_0^T dM_j(t) + \frac{2}{T}\sum_{\substack{1\le j,j'\le d\\ 1\le k\le K}} (\hat a_{j,j',k} - a_{j,j',k})\int_0^T H_{j,j',k}(t) dM_j(t) \\ &= \frac{2}{T}\langle \hat\mu - \mu, M(T)\rangle + \frac{2}{T}\langle \hat{\mathbb{A}} - \mathbb{A}, \mathbb{Z}(T)\rangle,\end{aligned}$$

where $M(T) = [M_1(T) \cdots M_d(T)]^\top$ and where $\mathbb{Z}(T)$ is the $d \times d \times K$ tensor with entries

$$\mathbb{Z}_{j,j',k}(T) = \int_0^T H_{j,j',k}(t) dM_j(t).$$

Recall that $\operatorname{hstack}$ is the horizontally stacking operator defined by (19). The following upper bounds

$$\begin{aligned}&|\langle \hat\mu - \mu, M(T)\rangle| \le \sum_{j=1}^d |\hat\mu_j - \mu_j||M_j(T)| \\ &|\langle \hat{\mathbb{A}} - \mathbb{A}, \mathbb{Z}(T)\rangle| \le \sum_{\substack{1\le j,j'\le d\\ 1\le k\le K}} |\hat{\mathbb{A}}_{j,j',k} - \mathbb{A}_{j,j',k}||\mathbb{Z}_{j,j',k}(T)| \\ &|\langle \hat{\mathbb{A}} - \mathbb{A}, \mathbb{Z}(T)\rangle| = \langle \operatorname{hstack}(\hat{\mathbb{A}} - \mathbb{A}), \operatorname{hstack}(\mathbb{Z}(T))\rangle \le \|\operatorname{hstack}(\mathbb{Z}(T))\|_{\mathrm{op}}\|\operatorname{hstack}(\hat{\mathbb{A}} - \mathbb{A})\|_*,\end{aligned}$$

entail that we need to upper bound the three terms

$$|M_j(T)|, \quad |\mathbb{Z}_{j,j',k}(T)| \quad \text{and} \quad \|\operatorname{hstack}(\mathbb{Z}(T))\|_{\mathrm{op}}$$

by data-driven quantities. Let us start with $\|\operatorname{hstack}(\mathbb{Z}(T))\|_{\mathrm{op}}$. Denote for short $\boldsymbol{Z}(t) = \operatorname{hstack}(\mathbb{Z}(t))$ and $\boldsymbol{H}(t) = \operatorname{hstack}(\mathbb{H}(t))$ where $\mathbb{H}(t)$ is defined by (18). We note that

$$\boldsymbol{Z}(t) = \int_0^t \operatorname{diag}(dM(s))\boldsymbol{H}(s),$$

namely

$$(\boldsymbol{Z}(t))_{j,j'+(k-1)d} = \int_0^t (\mathbb{H}(t-s))_{j,j',k} dM_j(s)$$

for any $1 \leq j, j' \leq d$ and $1 \leq k \leq K$. We need the following corollary.

**Corollary 7** *The following deviation inequality holds*

$$\mathbb{P}\Big[\|\boldsymbol{Z}(t)\|_{\mathrm{op}} \geq 2\sqrt{\lambda_{\max}(\widehat{\boldsymbol{V}}(t))(x + \log(2d) + \ell(t))} \\ + 14.39(x + \log(2d) + \ell(t))(1 + \sup_{0 \leq s \leq t} \|\boldsymbol{H}(s)\|_{\infty,2})\Big] \leq 23.45 e^{-x}, \tag{41}$$

*where*

$$\lambda_{\max}(\widehat{\boldsymbol{V}}(t)) = \lambda_{\max}\Big(\int_0^t \boldsymbol{H}^\top(s)\boldsymbol{H}(s) \operatorname{diag}(dN(s))\Big) \bigvee \max_{j=1,\ldots,d} \int_0^t \|\boldsymbol{H}_{j,\bullet}(s)\|_2^2 dN_j(s),$$

*and where*

$$\ell(t) = 2\log\log\Big(\frac{4\lambda_{\max}(\widehat{\boldsymbol{V}}(t))}{x} \vee 2\Big) + 2\log\log\Big(4 \sup_{0 \leq s \leq t} \|\boldsymbol{H}(s)\|_{\infty,2} \vee 2\Big).$$

The proof of Corollary 7 is given in Section 8.6 below. Corollary 7 proves that $\frac{1}{T}\|\boldsymbol{Z}(t)\|_{\mathrm{op}} \leq \frac{\hat{\tau}}{2}$ holds with probability $1 - 23.45e^{-x}$, with

$$\hat{\tau} = 4\sqrt{\frac{\lambda_{\max}(\widehat{\boldsymbol{V}}(T)/T)(x + \log(2d) + \ell(T))}{T}} \\ + 28.78\frac{x + \log(2d) + \ell(T))(1 + \sup_{0 \leq t \leq T} \|\boldsymbol{H}(t)\|_{\infty,2})}{T},$$

which leads to the choice of $\hat{\tau}$ given in Section 4. This entails that, on an event of probability larger than $1 - 23.45e^{-x}$, we have

$$\frac{1}{T}|\langle \hat{\mathbb{A}} - \mathbb{A}, \mathbb{Z}(T)\rangle| \leq \frac{\hat{\tau}}{2}\|\operatorname{hstack}(\hat{\mathbb{A}} - \mathbb{A})\|_*.$$

Using again Corollary 7 with $\boldsymbol{H}(t) \equiv 1$ (constant number equal to 1) and $M = M_j$ gives that $\frac{1}{T}|M_j(T)| \leq \frac{\hat{w}_j}{3}$ for all $j = 1, \ldots, d$ with probability $1 - 23.45e^{-x}$ with

$$\hat{w}_j = 6\sqrt{\frac{(N_j(T)/T)(x + \log d + \ell_j(T))}{T}} + 86.34\frac{x + \log d + \ell_j(T)}{T},$$

with $\ell_j(T) = 2\log\log(\frac{4N_j(T)}{x} \vee 2) + 2\log\log 4$. This entails that, on an event of probability larger than $1 - 23.45e^{-x}$, we have

$$\frac{2}{T}|\langle \hat{\mu} - \mu, M(T)\rangle| \leq \frac{2}{3}\|\hat{\mu} - \mu\|_{1,\hat{w}}.$$

Using a last time Corollary 7 with $\boldsymbol{H}(t) = H_{j,j',k}(t)$ and $M = M_j$ gives $\frac{1}{T}|\mathbb{Z}_{j,j',k}(T)| \leq \frac{\hat{\mathbb{W}}_{j,j',k}}{2}$ uniformly for $j, j', k$ for

$$\hat{\mathbb{W}}_{j,j',k} = 4\sqrt{\frac{\frac{1}{T}\int_0^T H_{j,j',k}(t)^2 dN_j(t)(x + \log(Kd^2) + \mathbb{L}_{j,j',k}(T))}{T}} \\ + 28.78\frac{(x + \log(Kd^2) + \mathbb{L}_{j,j',k}(T))(1 + \sup_{0 \leq t \leq T}|H_{j,j',k}(t)|)}{T},$$

where

$$\mathbb{L}_{j,j',k}(T) = 2\log\log\Big(\frac{4\int_0^T H_{j,j',k}(t)^2 dN_j(t)}{x} \vee 2\Big) + 2\log\log\Big(4\sup_{0\le t\le T}|H_{j,j',k}(t)| \vee 2\Big),$$

which entails that on an event of probability larger than $1 - 23.45e^{-x}$, we have

$$\frac{1}{T}|\langle \hat{\mathbb{A}} - \mathbb{A}, \mathbb{Z}(T)\rangle| \le \frac{1}{2}\|\hat{\mathbb{A}} - \mathbb{A}\|_{1,\hat{\mathbb{W}}}.$$

This entails that, with a probability larger than $1 - 3 \times 23.45e^{-x}$, one has

$$\begin{aligned} 0 &\le -\langle \theta_\partial, \hat\theta - \theta\rangle + \frac{2}{T}\sum_{j=1}^d \int_0^T (\lambda_{j,\hat\theta}(t) - \lambda_{j,\theta}(t)) dM_j(t) \\ &\le \frac{5}{3}\|(\hat\mu - \mu)_{\mathrm{supp}(\mu)}\|_{1,\hat w} - \frac{1}{3}\|(\hat\mu - \mu)_{\mathrm{supp}(\mu)^\perp}\|_{1,\hat w} \\ &\quad + \frac{3}{2}\|(\hat{\mathbb{A}} - \mathbb{A})_{\mathrm{supp}(\mathbb{A})}\|_{1,\hat{\mathbb{W}}} - \frac{1}{2}\|(\hat{\mathbb{A}} - \mathbb{A})_{\mathrm{supp}(\mathbb{A})^\perp}\|_{1,\hat{\mathbb{W}}} \\ &\quad + \frac{3}{2}\hat\tau\|\mathcal{P}_{\boldsymbol{A}}(\hat{\boldsymbol{A}} - \boldsymbol{A})\|_* - \frac{1}{2}\hat\tau\|\mathcal{P}_{\boldsymbol{A}}^\perp(\hat{\boldsymbol{A}} - \boldsymbol{A})\|_*, \end{aligned}$$

where we recall once again that $\boldsymbol{A} = \mathrm{hstack}(\mathbb{A})$ and $\hat{\boldsymbol{A}} = \mathrm{hstack}(\hat{\mathbb{A}})$. This matches the constraint of Definition 5 with $\mu' = \hat\mu - \mu$ and $\mathbb{A}' = \hat{\mathbb{A}} - \mathbb{A}$, so that it entails

$$\|(\hat\mu - \mu)_{\mathrm{supp}(\mu)}\|_2 \vee \|(\hat{\mathbb{A}} - \mathbb{A})_{\mathrm{supp}(\mathbb{A})}\|_F \vee \|\mathcal{P}_{\boldsymbol{A}}(\hat{\boldsymbol{A}} - \boldsymbol{A})\|_F \le \kappa(\theta)\|\lambda_{\hat\theta} - \lambda_\theta\|_T. \tag{42}$$

Putting all this together gives

$$\begin{aligned} -\langle \theta_\partial, &\hat\theta - \theta\rangle + \frac{2}{T}\langle \hat\mu - \mu, M(T)\rangle + \frac{2}{T}\langle \hat{\mathbb{A}} - \mathbb{A}, \mathbb{Z}(T)\rangle \\ &\le \frac{5}{3}\|(\hat\mu - \mu)_{\mathrm{supp}(\mu)}\|_{1,\hat w} - \frac{1}{3}\|(\hat\mu - \mu)_{\mathrm{supp}(\mu)^\perp}\|_{1,\hat w} \\ &\quad + \frac{3}{2}\|(\hat{\mathbb{A}} - \mathbb{A})_{\mathrm{supp}(\mathbb{A})}\|_{1,\hat{\mathbb{W}}} - \frac{1}{2}\|(\hat{\mathbb{A}} - \mathbb{A})_{\mathrm{supp}(\mathbb{A})^\perp}\|_{1,\hat{\mathbb{W}}} \\ &\quad + \frac{3}{2}\hat\tau\|\mathcal{P}_{\boldsymbol{A}}(\hat{\boldsymbol{A}} - \boldsymbol{A})\|_* - \frac{1}{2}\hat\tau\|\mathcal{P}_{\boldsymbol{A}}^\perp(\hat{\boldsymbol{A}} - \boldsymbol{A})\|_* \\ &\le \frac{5}{3}\|(\hat w)_{\mathrm{supp}(\mu)}\|_2\|(\hat\mu - \mu)_{\mathrm{supp}(\mu)}\|_2 + \frac{3}{2}\|(\hat{\mathbb{W}})_{\mathrm{supp}(\mathbb{A})}\|_F\|(\hat{\mathbb{A}} - \mathbb{A})_{\mathrm{supp}(\mathbb{A})}\|_F \\ &\quad + \frac{3}{2}\hat\tau\sqrt{\mathrm{rank}(\boldsymbol{A})}\|\mathcal{P}_{\boldsymbol{A}}(\hat{\boldsymbol{A}} - \boldsymbol{A})\|_F, \end{aligned}$$

where we used Cauchy-Schwarz's inequality. This finally gives

$$\begin{aligned} \|\lambda_{\hat\theta} - \lambda\|_T^2 &\le \|\lambda_\theta - \lambda\|_T^2 - \|\lambda_{\hat\theta} - \lambda_\theta\|_T^2 \\ &\quad + \kappa(\theta)\Big(\frac{5}{3}\|(\hat w)_{\mathrm{supp}(\mu)}\|_2 + \frac{3}{2}\|(\hat{\mathbb{W}})_{\mathrm{supp}(\mathbb{A})}\|_F + \frac{3}{2}\hat\tau\sqrt{\mathrm{rank}(\boldsymbol{A})}\Big)\|\lambda_{\hat\theta} - \lambda_\theta\|_T \end{aligned}$$

where we used (42). The conclusion of the proof of Theorem 6 follows from the fact that $ax - x^2 \le a^2/4$ for any $a, x > 0$.

### 8.6. Proof of Corollary 7

We simply use Theorem 4. First, we remark that $\boldsymbol{Z}(t) = \int_0^t \mathbb{T}(s) \circ \mathrm{diag}(dM(s))$ for the tensor $\mathbb{T}(t)$ of size $d \times Kd \times d \times d$ given by

$$(\mathbb{T}(t))_{i,j;k,l} = (\boldsymbol{I})_{i,k}(\boldsymbol{H}(t))_{l,j} \tag{43}$$

for $1 \leq i, k, l \leq d$ and $1 \leq j \leq Kd$. Note that we have

$$\mathbb{T}_{\bullet,\bullet;k,l}(t) = e_k \boldsymbol{H}_{l,\bullet}(t)^\top \quad \text{and} \quad \mathbb{T}_{\bullet,\bullet;k,l}(t)^\top = \boldsymbol{H}_{l,\bullet}(t) e_k^\top \tag{44}$$

where $e_k \in \mathbb{R}^d$ stands for the $k$-th element of the canonical basis of $\mathbb{R}^d$ and where $\boldsymbol{H}_{l,\bullet}(t) \in \mathbb{R}^{Kd}$ stands for the vector corresponding to the $l$-th row of the matrix $\boldsymbol{H}(t)$. Therefore, we have

$$\mathbb{T}_{\bullet,\bullet;k,l}(t)\mathbb{T}_{\bullet,\bullet;k,l}^\top(t) = \|\boldsymbol{H}_{l,\bullet}(t)\|_2^2 e_k e_k^\top \quad \text{and} \quad \mathbb{T}_{\bullet,\bullet;k,l}^\top(t)\mathbb{T}_{\bullet,\bullet;k,l}(t) = \boldsymbol{H}_{l,\bullet}(t)\boldsymbol{H}_{l,\bullet}(t)^\top$$

and therefore

$$\|\mathbb{T}_{\bullet,\bullet;k,l}(t)\|_{\text{op}} = \sqrt{\lambda_{\max}(\mathbb{T}_{\bullet,\bullet;k,l}(t)\mathbb{T}_{\bullet,\bullet;k,l}^\top(t))} = \|\boldsymbol{H}_{l,\bullet}(t)\|_2$$

and

$$\|\mathbb{T}(t)\|_{\text{op};\infty} = \max_{1 \leq l \leq d} \|\boldsymbol{H}_{l,\bullet}(t)\|_2 = \|\boldsymbol{H}(t)\|_{\infty,2}.$$

One can prove in the same way that $\|\mathbb{T}^\top(t)\|_{\text{op};\infty} = \|\boldsymbol{H}(t)\|_{\infty,2}$, so that for this choice of tensor $\mathbb{T}(t)$, we have $b_{\mathbb{T}}(t) = \|\boldsymbol{H}(t)\|_{\infty,2}$. Now, let us explicit what $\widehat{\boldsymbol{V}}_{\mathbb{T}}(t)$ is for the tensor (43). First, let us remind that

$$\widehat{\boldsymbol{V}}_{\mathbb{T}}(t) = \begin{bmatrix} \int_0^t \mathbb{T}(s)\mathbb{T}^\top(s) \circ \operatorname{diag}(dN(s)) & \mathbf{0} \\ \mathbf{0} & \int_0^t \mathbb{T}^\top(s)\mathbb{T}(s) \circ \operatorname{diag}(dN(s)) \end{bmatrix}.$$

Using (44) we get

$$(\mathbb{T}(t)\mathbb{T}(t)^\top)_{\bullet,\bullet;,k,l} = e_k \boldsymbol{H}_{l,\bullet}(t)^\top \boldsymbol{H}_{l,\bullet}(t) e_k^\top = \|\boldsymbol{H}_{l,\bullet}(t)\|_2^2 e_k e_k^\top$$

so that $\int_0^t (\mathbb{T}(s)\mathbb{T}^\top(s)) \circ \operatorname{diag}(dN(s))$ is the diagonal matrix with entries

$$\Big( \int_0^t (\mathbb{T}(s)\mathbb{T}^\top(s)) \circ \operatorname{diag}(dN(s)) \Big)_{j,j} = \int_0^t \|\boldsymbol{H}_{j,\bullet}(s)\|_2^2 dN_j(s),$$

or equivalently

$$\int_0^t (\mathbb{T}(s)\mathbb{T}^\top(s)) \circ \operatorname{diag}(dN(s)) = \int_0^t \operatorname{diag}(\boldsymbol{H}^\top(s)\boldsymbol{H}(s)) \operatorname{diag}(dN(s)).$$

Using again (44) we get

$$(\mathbb{T}^\top(t)\mathbb{T}(t))_{\bullet,\bullet;,k,l} = \boldsymbol{H}_{l,\bullet}(t) e_k^\top e_k \boldsymbol{H}_{l,\bullet}(t)^\top = \boldsymbol{H}_{l,\bullet}(t)\boldsymbol{H}_{l,\bullet}(t)^\top$$

so that $\int_0^t (\mathbb{T}^\top(s)\mathbb{T}(s)) \circ \operatorname{diag}(dN(s))$ is the matrix with entries

$$\Big( \int_0^t (\mathbb{T}^\top(s)\mathbb{T}(s)) \circ \operatorname{diag}(dN(s)) \Big)_{i,j} = \sum_{l=1}^d \int_0^t \boldsymbol{H}_{l,i}(s)\boldsymbol{H}_{l,j}(s) dN_l(s)$$

or equivalently

$$\int_0^t (\mathbb{T}^\top(s)\mathbb{T}(s)) \circ \operatorname{diag}(dN(s)) = \int_0^t \boldsymbol{H}^\top(s)\boldsymbol{H}(s) \operatorname{diag}(dN(s)).$$

Finally, we obtain that

$$\lambda_{\max}(\widehat{\boldsymbol{V}}_t) = \lambda_{\max}\Big( \int_0^t \boldsymbol{H}^\top(s)\boldsymbol{H}(s) \operatorname{diag}(dN(s)) \Big) \bigvee \max_{j=1,\ldots,d} \int_0^t \|\boldsymbol{H}_{j,\bullet}(t)\|_2^2 dN_j(s).$$

This concludes the proof of the corollary. □

## Acknowledgments

This work was funded in part by the French government under management of Agence Nationale de la Recherche as part of the ”Investissements d’avenir” program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute).

## References

E. Bacry, S. Delattre, M. Hoffmann, and J.-F. Muzy. Modelling microstructure noise with mutually exciting point processes. *Quantitative Finance*, 13(1):65–77, 2013.

E. Bacry, I. Mastromatteo, and J.-F. Muzy. Hawkes processes in finance. *Market Microstructure and Liquidity*, 01(01):1550005, 2015.

E. Bacry, S. Gaïffas, I. Mastromatteo, and J.-F. Muzy. Mean-field inference of hawkes point processes. *Journal of Physics A: Mathematical and Theoretical*, 49(17):174006, 2016a.

E. Bacry, S. Gaïffas, and J.-F. Muzy. Concentration inequalities for matrix martingales in continuous time. *Probability Theory and Related Fields*, 170:525–553, 2016b.

E. Bacry, M. Bompaire, P. Deegan, S. Gaïffas, and S. V. Poulsen. tick: a python library for statistical learning, with an emphasis on hawkes processes and time-dependent models. *Journal of Machine Learning Research*, 18(214):1–5, 2018.

P. L. Bartlett and S. Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135(3): 311–334, 2006.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal of Imaging Sciences*, 2(1):183–202, 2009.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.

C. Blundell, K. A Heller, and J. M. Beck. Modelling reciprocating relationships with hawkes processes. In *NIPS*, pages 2609–2617, 2012.

M. Bompaire. *Machine Learning based on Hawkes processes and Stochastic Optimization*. PhD thesis, CMAP, Ecole polytechique, EDMH, 2018.

M. Bompaire, E. Bacry, and S. Gaïffas. Dual optimization for convex constrained objectives without the gradient-lipschitz assumption. *arXiv preprint arXiv:1807.03545*, 2018.

E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 12 (51):4203–4215, 2004.

E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5), 2009.

R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41), 2008.

N. Daneshmand, M. Rodriguez, L. Song, and B. Schölkpof. Estimating diffusion network structure: Recovery conditions, sample complexity, and a soft-thresholding algorithm. *ICML*, 2014.

M. Argollo de Menezes and A.-L. Barabási. Fluctuations in network dynamics. *Phys. Rev. Lett.*, 92:028701, Jan 2004. doi: 10.1103/PhysRevLett.92.028701. URL `http://link.aps.org/doi/10.1103/PhysRevLett.92.028701`.

C. DuBois, C. Butts, and P. Smyth. Stochastic blockmodeling of relational event dynamics. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pages 238–246, 2013.

M. Gomez-Rodriguez, J. Leskovec, and B. Schölkopf. Modeling information propagation with survival theory. *ICML*, 2013.

N. R. Hansen, P. Reynaud-Bouret, and V. Rivoirard. Lasso and probabilistic inequalities for multivariate point processes. Technical report, Arvix preprint, 2012.

A. G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.

T. Iwata, A. Shah, and Z. Ghahramani. Discovering latent influence in online social activities via shared cascade poisson processes. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 266–274. ACM, 2013.

V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Saint-Flour XXXVIII-2008*, volume 2033. Springer, 2011.

V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

J. Leskovec. *Dynamics of large networks*. PhD thesis, Machine Learning Department, Carnegie Mellon University, 2008.

J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD*. ACM, 2009.

A. S. Lewis. The convex analysis of unitarily invariant matrix functions. *Journal of Convex Analysis*, 2(1): 173–183, 1995.

S. W. Linderman and R. P. Adams. Discovering latent network structure in point process data. *arXiv preprint arXiv:1402.0914*, 2014.

P. Massart. *Concentration inequalities and model selection*, volume 1896. Springer, 2007.

G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 2011.

Y. Ogata. The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Annals of the Institute of Statistical Mathematics*, 30(1):243–261, 1978.

Y. Ogata. On lewis' simulation method for point processes. *Information Theory, IEEE Transactions on*, 27 (1):23–31, 1981.

Y. Ogata. Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, 50(2):379–402, 1998.

M. R. Pino, L. Landesa, J. L. Rodriguez, F. Obelleiro, and R. J. Burkholder. The generalized forward-backward method for analyzing the scattering from targets on ocean-like rough surfaces. *IEEE Transactions on Antennas and Propagation*, 47(6):961–969, 1999.

F. Ricci, L. Rokach, and B. Shapira. *Introduction to recommender systems handbook*. Springer, 2011.

E. Richard, S. Gaïffas, and N. Vayatis. Link prediction in graphs with autoregressive features. *Journal of Machine Learning Research*, 2014.

M. Rodriguez, D. Balduzzi, and B. Schölkopf. Uncovering the temporal dynamics of diffusion networks. *ICML*, 2011.

J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

S. Van De Geer. *Empirical Processes in M-estimation*, volume 105. Cambridge university press Cambridge, 2000.

S.-H. Yang and H. Zha. Mixture of mutually exciting processes for viral diffusion. In *ICML*, 2013.

K. Zhou, H. Zha, and L. Song. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *AISTATS*, volume 31, pages 641–649, 2013.