Nonparametric Markovian Learning of Triggering Kernels for Mutually Exciting and Mutually Inhibiting Multivariate Hawkes Processes

Remi Lemonnier[1,2] and Nicolas Vayatis[1]

[1] CMLA-ENS Cachan,Cachan,France
[2] 1000Mercis,Paris,France

Abstract. In this paper, we address the problem of fitting multivariate Hawkes processes to potentially large-scale data in a setting where series of events are not only mutually-exciting but can also exhibit inhibitive patterns. We focus on nonparametric learning and propose a novel algorithm called MEMIP (Markovian Estimation of Mutually Interacting Processes) that makes use of polynomial approximation theory and self-concordant analysis in order to learn both triggering kernels and base intensities of events. Moreover, considering that N historical observations are available, the algorithm performs log-likelihood maximization in $O(N)$ operations, while the complexity of non-Markovian methods is in $O(N^2)$. Numerical experiments on simulated data, as well as real-world data, show that our method enjoys improved prediction performance when compared to state-of-the art methods like MMEL and exponential kernels.

1 Introduction

Multivariate Hawkes processes are a class of multivariate point processes which are often used to model counting processes where physicals events rate of occurrence usually depend on past occurences of many other events. This is typically the case for earthquakes aftershocks ([1]) and financial trade orders on marketplace ([2], [3],[4], [5]), but also in other fields such as crime prediction ([6]), genome analysis ([7]) and more recently for modeling social interactions ([8], [9]). Multivariate Hawkes processes are fairly well-known from a probabilistic point of view : their Poisson cluster representation was outlined by the seminal paper of Hawkes and Oakes ([10]), stability conditions and sample path large deviations principles were derived in a sequence of papers by Bremaud and Massoulie (see e.g [11]). In the unidimensional case, Ogata [12] showed that the log-likelihood estimator enjoys usual convergence properties under mild regularity conditions. However, in practical applications, estimation of the triggering kernels g_{uv} has always been a difficult task. First, because Hawkes log-likelihood contains the logarithm of the weighted sum of triggering kernels, most of the aforementioned papers made the choice of fixing triggering kernels up to a normalization factor

in order to ensure concavity, that is $g_{uv} = c_{uv} \cdot g$. Secondly, when computational efficiency is an issue, the dependency of the stochastic rate at a given time on all the past occurences implies quadratic complexity in the number of occurences for tasks like log-likelihood computation. This issue has often been tackled by choosing memoryless exponential triggering kernels, but the actual dynamics of kernels strongly depends on the field of application: price impacts of a given trade ([13]) and process of views of Youtube videos ([14]) were shown to be better described by slowly decaying power-law kernels whereas for DNA sequence modelization ([7]) kernels are known to have bounded support. Thus, it is highly desirable to estimate triggering kernels in a data-driven way instead of assuming a given parametric form. Nonparametric estimation has been successfully addressed for unidimensional ([7],[15]), and symmetric bidimensional ([13]) Hawkes processes. In the case where triggering kernels are known to sparsely decompose over a dictionary of basis functions of bounded support (e.g for neuron spikes interactions), a LASSO-based algorithm with provable guarantees was derived in [16].

Recently, combining majorization-minimization techniques with resolution of a Euler-Lagrange equation, Zhou, Zha and Song ([9]) proposed what is to our knowledge the first nonparametric learning algorithm for general multivariate Hawkes processes. But although this work constitutes a significant improvement over existing parametric methods, it still relies on several assumptions. First, interactions between events are assumed to be "mutually-exciting", i.e $g_{uu'}$ are non-negative for all u, u'. We nevertheless argue that in real-world settings, there is no reason to think that interactions beween events are only mutually-exciting. Secondly, the background rates μ_u are assumed to be constant. While this is a common assumption for multivariate Hawkes processes, it was shown by [17] that estimating $\mu_u(t)$ from the data could lead to significant improvement. To address these different issues, we construct a novel algorithm MEMIP (Markovian Estimation of Mutually Interacting Processes) based on polynomial approximation of a mapping of the triggering kernels to $[0, 1]$. Our method does not assume non-negativity on triggering kernels and is able to estimate time-dependent background rate on a data-driven way. Moreover, by constructing a markovian and linear estimator, we carry the more appealing properties of the most widely used parametric setting, where triggering kernels are fixed to exponentials up to a normalization factor : concavity of the log-likelihood that ensures global convergence of the estimator, and $O(N)$ log-likelihood calculation in a single pass through the data. While giving a concave formulation of the exact log-likelihood that can be maximized by multiple optimization techniques, we propose an algorithm based on maximisation of a self-concordant approximation that is shown to outperform state-of-the-art methods on both simulated and real-world data sets.

The paper is organized as follows. In Section 2, we formally define multivariate Hawkes processes as well as the associated log-likelihood maximization problem. In section 3, we decompose the log-likelihood on a basis of memoryless triggering kernels. Through Section 4, we develop two novel algorithms for ex-

act as well as fast approximate maximization of the log-likelihood, analyze their complexity and show numerical convergence results based on the properties of self-concordant functions. In section 5, we show that MEMIP significantly improves over state of the art on both synthetic and real world data sets for the tasks of predicting future events as well as estimating underlying dynamics of the Hawkes process.

2 Setup and Notations

2.1 Model Description and Notation

We consider a multivariate Hawkes process, that is a d-dimensional counting process $N(t) = \{N^u(t) \ : \ u = 1, \dots, d\}$ for which the rate of occurence of each component $N^u(t)$ is defined by:

$$\lambda_u(t) = \left(\mu_u(t) + \sum_{v \in [1 \dots d]} \sum_{t_v < t} g_{uv}(t - t_v) \right)_+ , \quad \forall u = 1, \dots, d \tag{1}$$

where $\mu_u(t)$ is the natural rate of occurence of events along dimension u. Note that the occurence of a given event affects stochastic rates of occurence of every dimension. With an empty history, events of type u will occur as if they were drawn from a non-homogeneous Poisson process of rate $\mu_u(t)$. The kernel function evaluation $g_{uv}(t - t_v)$ quantifies the change in the rate of occurence of event u at time t caused by the realization of event v at time t_v. Following the intuition, we can characterize three situations depending on the values taken by the kernel function at a given time lapse s:

- *Excitation* corresponds to the case where we have $g_{uv}(s) > 0$, *i.e.* an event of type v is more likely to occur if an event of type u has occured at a time distance of s.
- *Independence* is observed when $g_{uv}(s) = 0$, meaning that the realization of an event of type u has no effect on the rate of occurence of an event of type v at time distance s.
- *Inhibition* takes place when $g_{uv}(s) < 0$, *i.e.* an event of type v is less likely to occur if an event of type u occured at time distance s.

Such processes can be seen as a generalization over the common definition of multivariate Hawkes process where the kernels g_{uv} are non-negative and the componentwise background rate μ_u is often taken constant.

2.2 Log-Likelihood of Multivariate Hawkes Processes

Input Observations. We define a *realization* h of a multivariate point process by the triplet $T_h^-, T_h^+, (t_i^h, u_i^h)_{i \in [1 \dots n_h]}$, where T_h^- and T_h^+ are respectively the beginning and the end of the observation period, and (t_i^h, u_i^h), for $i \in [1 \dots n_h]$,

is the sequence of the n_h events occuring during this period. In the rest of the paper, we will assume we are given n i.i.d realizations of a multivariate Hawkes process. Without loss of generality, we will assume $\min_h(T_h^-) = 0$ and take $T = \max_h(T_h^+)$.

Expression of the Log-Likelihood. We first set $\Lambda = \{\lambda_u \ : \ u = 1, \dots, d\}$. For a general multivariate point process, the log-likelihood of the whole dataset $\mathcal{H}$ is given by (*e.g.* [18]):

$$\mathcal{L}(\Lambda, \mathcal{H}) = \sum_{u=1}^{d} \sum_{h \in \mathcal{H}} \int_{T_h^-}^{T_h^+} \ln(\lambda_u(s)) dN_h^u(s) - \sum_{u=1}^{d} \sum_{h \in \mathcal{H}} \int_{T_h^-}^{T_h^+} \lambda_u(s) ds \qquad (2)$$

where $\int f(s)) dN_h^u(s) = \sum_{i=1}^{n_h} f(t_i^h) 1\left\{u_i^h = u\right\}$. In the case of a linear Hawkes process (1), we introduce $\Lambda = (M, G)$ where $M = \{\mu_u \ : \ u = 1, \dots, d\}$ and $G = \{g_{u,v} \ : \ u, v = 1, \dots, d\}$ and the log-likelihood can be rewritten as:

$$\mathcal{L}(M, G, \mathcal{H}) = \sum_{h \in \mathcal{H}} \sum_{i=1}^{n_h} \ln \left(\mu_{u_i^h}(t_i^h) + \sum_{j \ : \ t_j^h < t_i^h} g_{u_j^h, u_i^h}(t_i^h - t_j^h) \right)$$

$$- \sum_{u=1}^{d} \sum_{h \in \mathcal{H}} \int_{T_h^-}^{T_h^+} \left(\mu_u(s) + \sum_{j=1}^{n_h} 1\left\{u_j^h = u\right\} \ g_{u,u_j}(s - t_j) \right)_+ ds \qquad (3)$$

Depending on the parametrization of triggering kernels g_{uv}, this log-likelihood may or may not be concave. For instance, in the widely used setting where the background rates μ_u are constant and the kernels g_{uv} are non-negative and fixed up to the normalization factor ν_{uv}, the log-likelihood is concave and can be relatively easily maximized. However, even for the simple case of nonnegative exponential kernels $g_{uv}(t) = \nu_{uv} \exp(-\alpha_j t)$ where $\nu_{uv} \geq 0$ the product term $\nu_{uv} \exp(-\alpha_v t)$ makes the log-likelihood not concave with respect to α_v. Therefore, global convergence of maximization methods is not guaranteed anymore.

3 Approximations of Multivariate Hawkes Processes on a Basis of Exponential Triggering Kernels

3.1 A K-approximation of the Multivariate Hawkes Process

For a given multivariate Hawkes process $\Lambda = (M, G)$, we consider finite approximations of the components of the rates of occurence μ_u and g_{uv}. We first introduce the following functions:

$$\forall y \in [-\ln(T)/\alpha, 1], \quad \nu_u(y) = \mu_u(-\ln(y)/\alpha) \quad \text{and} \quad f_{uv}(y) = g_{uv}(-\ln(y)/\alpha)$$

and we use Bernstein-type polynomial approximations of order K for ν_u and f_{uv}: there exist coefficients $X_{uv,k}^K$ such that

$$\forall y \in [-\ln(T)/\alpha, 1], \quad \widehat{\nu}^K(y) = \sum_{k=0}^{K} X_{u0,k}^K y^k \quad \text{and} \quad \widehat{f}_{uv}^K(y) = \sum_{k=0}^{K} X_{uv,k}^K y^k \ .$$

These polynomial approximations are known to converge with a polynomial rate for smooth functions (with first r derivatives continuously differentiable) and geometric rate for analytic functions (see below). The K-aproximation considered in this paper relies on a simple change of variable in the Bernstein approximations by setting: $y = \exp(-\alpha t)$. We can now introduce the linear approximation of a multivariate Hawkes process with exponential kernels:

$$\forall t \in [0,T], \quad \widehat{\mu}^K(t) = \sum_{k=0}^{K} X^K_{u0,k} \exp(-k\alpha t) \quad \text{and} \quad \widehat{g}^K_{uv}(t) = \sum_{k=0}^{K} X^K_{uv,k} \exp(-k\alpha t)\,.$$

Classical arguments from approximation theory ([19] and [20]) lead to the following proposition.

Proposition 1. *For any function Ψ defined over $[0,T]$, we consider the supremum norm $||\Psi||_{T,\infty} = \sup_{t\in[0,T]} |\Psi(t)|$. The K-approximations $(\widehat{\mu}^K_u)_{K\geq 1}$ and $(\widehat{g}^K_{uv})_{K\geq 1}$ converge in supremum norm towards true functions μ_u and g_{uv} at the following rates:*

1. *if μ_u is C^r, $\left|\left|\mu_u(t) - \widehat{\mu}^K_u(t)\right|\right|^T_\infty = O(1/K^r)$*
2. *if μ_u is analytic, $\left|\left|\mu_u(t) - \widehat{\mu}^K_u(t)\right|\right|^T_\infty = O(\exp(-K))$*
3. *if g_{uv} is C^r, $\left|\left|g_{uv}(t) - \widehat{g}^K_{uv}(t)\right|\right|^T_\infty = O(1/K^r)$*
4. *if g_{uv} is analytic, $\left|\left|g_{uv}(t) - \widehat{g}^K_{uv}(t)\right|\right|^T_\infty = O(\exp(-K))$.*

Another property of the approximated multivariate Hawkes process is the Markov property of the counting process. We set $\widehat{N}^K(t)$ the d-dimensional Hawkes process uniquely defined by $\widehat{\lambda}^K = (\widehat{\mu}^K_u, \widehat{g}^K_{uv})_{u,v}$.

Proposition 2. *Assume that the empirical estimate $\widehat{N}^K(t)$ of the multivariate Hawkes process is obtained after i.i.d. realizations of $N(t)$ over the time interval $[0,T]$. There exists $(\widehat{\ell}^0, \widehat{\ell}^1, \ldots, \widehat{\ell}^K)$ such that:*

$$\forall u \in \{1,\ldots,d\}\ , \quad \widehat{\lambda}^K(t) = \sum_{k=0}^{K} \left(\widehat{\ell}^k(t)\right)_+$$

and $(\widehat{N}^K(t), \widehat{\ell}^0(t), \widehat{\ell}^1(t), \ldots, \widehat{\ell}^K(t))$ is a Markov Process on $\mathbb{N}^d \times \mathbb{R}^{d(K+1)}$.

The proof results from the following decomposition of each occurrence rate in the approximation: $\forall u \geq 1$,

$$\widehat{\lambda}^K_u(t) = \left(X^K_{u0,0} + \sum_{k=1}^{K} \left(X^K_{u0,k} \exp(-k\alpha t) + \sum_{v\ :\ t_v<t} X^K_{uv,(k-1)} \exp(-k\alpha(t-t_v)) \right) \right.$$
$$\left. + \sum_{v\ :\ t_v<t} X^K_{uv,K} \exp(-(K+1)\alpha(t-t_v)) \right)_+$$

Markov property is then a direct consequence of the dynamics of the functions $\widehat{\ell}_u^k(t)$: they decay at rate $\exp(-k\alpha t)$ and jump by $X^K_{uv,(k-1)}$ whenever an event of type v occurs. As they entirely determine the stochastic rate which determines the conditional probability distribution of $\widehat{N}^K(t)$, the conditional probability distribution of future states of the process $(\widehat{N}^K(t), \widehat{\ell}^0(t), \widehat{\ell}^1(t), ...\widehat{\ell}^K(t))$ is uniquely determined by the present state.

3.2 A New Decomposition of the Log-Likelihood

The algorithms proposed in this paper rely on a novel expression of the log-likelihood over a basis of triggering kernels. We use exponential excitation functions to account for nonlinearity but our algorithms benefit from the properties of linear approximations. Based on the expression of the log-likelihood for general linear multivariate Hawkes process (3), we introduce the following notation to discover the specific expression for the K-approximation based on exponential triggering functions: $\forall u, v = 1, \dots, d$, $\forall k = 1, \dots, K$, $\forall h \in \mathcal{H}$, $\forall i = 1, \dots, n_h$,

$$A^{K,h,i}_{uv,k} = \sum_{j\,:\,t_j^h < t_i^h} 1\left\{u_i^h = v, u_j^h = u\right\} \exp\big(-(k + 1\left\{u > 0\right\})\alpha(t_i^h - t_j^h)\big) \quad (4)$$

$$B^{K,h}_{0v,k}(s) = \exp(-k\alpha s) \quad (5)$$

$$B^{K,h}_{uv,k}(s) = \sum_{j\,:\,t_j^h < s} 1\left\{u_j^h = v\right\} \exp(-(k+1)\alpha(s - t_j^h)) \quad (6)$$

The key expression of the approximate log-likelihood can then be derived by plugging-in the previous notations and replacing the intrinsic parameters (M, G) by the linear coefficients X^K:

$$\mathcal{L}^K(X^K, \mathcal{H}) = \sum_{h\in\mathcal{H}} \sum_{i=1}^{n_h} \ln(A^{K,h,i} X^K) - \sum_{h\in\mathcal{H}} \int_0^{T_h} \left(\sum_{i=1}^{n_h} B^{K,h}(s) X^K\right)_+ ds \quad (7)$$

Note that the dependance of $\mathcal{L}^K$ on the history $\mathcal{H}$ is entirely expressed by vectors $(A^{K,h,i})_{h\in\mathcal{H}, i\in[1\dots n_h]}$ and $(B^{K,h}(s))_{h\in\mathcal{H}, s\in[0,T]}$. An important feature of the approximate log-likelihood expressed in the parameter space defined by linear decompositions onto bases of exponential triggering kernels is given in the following proposition.

Proposition 3. *The function $X \to \mathcal{L}^K(X, \mathcal{H})$ is concave.*

From there, we have a complete roadmap for the design of algorithms estimating the parameters of multidimensional Hawkes processes: the last propostion indicates that a proxy of the log-likelihood (3) can be globally maximized with tools of convex analysis. Moreover, thanks to the approximation rates of convergence (Proposition 1), triggering kernels can be accurately estimated for large K through maximization of the new objective (7). Finally, the Markov property is an important feature that will allow us to construct the vectors $(A^{K,h,i})$ and $(B^{K,h})$ with linear complexity.

4 Markovian Algorithms for the Estimation of Triggering Kernels

Computational tractability of algorithms on large data sets depends on the algorithmic complexity in the dominating dimensions of the problem. For realizations of multivariate Hawkes processes, dominating dimensions are almost always the total number of events $N = \sum_{h \in \mathcal{H}} n_h$ and the time of observation T. Indeed, it would be unrealistic to try to learn d^2 nonparametric functions in an infinite dimensional space with only N observations without the condition $N \gg d^2$. In the rest of the paper, we will therefore focus on constructing two algorithms with no more than linear complexity in N and T.

4.1 Exact Maximization of the Approximated Log-Likelihood

Vectors $(A^{K,h,i})_{h \in \mathcal{H}, i \in [1 \dots n_h]}$ and $(B^{K,h}(s))_{h \in \mathcal{H}, s \in [0,T]}$ can be constructed in a single pass through the data by **Algorithm 1**.

Algorithm 1 Algorithm for construction of vectors $(A^{K,h,i})$ and $(B^{K,h}(s))$

Initialize $i = 0$ and fix a time step dt
for all h **do**
  Initialize $(C^k_{uv} = 0)_{u \geq 1, v \geq 1}$; $t = T_h^-$; $(D^k_{uv}(T_h^-) = 1_{\{u=0\}})_{u \geq 0, v \geq 1}$
  while $t < T_h^+$ **do**
    $t \leftarrow t + \delta t = \min(t + dt, t_i)$
    for all k,u,v **do**
      $C^k_{uv} \leftarrow C^k_{uv} \exp(-(k + 1\{u > 0\}\alpha \delta t)$, $D^k_{uv} \leftarrow D^k_{uv} \exp(-(k + 1\{u > 0\}\alpha \delta t)$
      $B^{K,h}_{uv,k}(t) \leftarrow D^k_{uv}$
    end for
    if $t = t_i$ **then**
      for all k,u **do**
        $A^{K,h,i}_{uv,k} \leftarrow C^k_{uu_i}$
      end for
      for all k,v **do**
        $C^k_{u_i v} \leftarrow C^k_{u_i v} + 1$, $D^k_{u_i v} \leftarrow D^k_{u_i v} + 1$
      end for
      $i \leftarrow i + 1$
    end if
  end while
end for

Complexity of Algorithm 1. With $M = T/dt$ the number of discretizations steps, construction of vectors $(A^{K,h,i})$ and $(B^{K,h}(s))$ has thus a complexity of $O(N+M)$. As each log-likelihood evaluation (7) requires $2N+M$ scalar products computations, various optimization techniques can be used to find the global maximum of $X \rightarrow \mathcal{L}^K(X, \mathcal{H})$ in $O(N + M)$ operations. On the contrary, a

nonmarkovian estimator, even linear, would need at each time t to compute the values of triggering kernels between current time and all preceding occurence times, thus leading to a $O(\sum_h n_h^2)$ complexity. This construction is thus very often the bottleneck of the whole maximization procedure.

4.2 Relaxed Version of the Log-Likelihood

While the previous paragraph exposes a fully tractable method to estimate the triggering kernels for potentially large data sets, we now develop an approximate algorithm called MEMIP, for Markovian Estimation of Mutually Interacting Processes, that leads to a substantial speed-up, as well as theoretical guarantees in terms of efficiency. For this purpose, we approximate the log-likelihood $\mathcal{L}^K(M, G, \mathcal{H})$ by dropping the positive part in log-likelihood (3), *i.e.*

$$\widetilde{\mathcal{L}}^K(M, G, \mathcal{H}) = \sum_{h \in \mathcal{H}} \left(\sum_{i=1}^{n_h} \ln \left(\mu_{u_i^h}(t_i^h) + \sum_{j \,:\, t_j^h < t_i^h} g_{u_j^h, u_i^h}(t_i^h - t_j^h) \right) \right.$$
$$\left. - \sum_{u=1}^{d} \int_{T_h^-}^{T_h^+} \left(\mu_u(s) + \sum_{j=1}^{n_h} 1\left\{u_j^h = u\right\} g_{u,u_j}(s - t_j) \right) ds \right) \tag{8}$$

which can be rewritten:

$$\widehat{\mathcal{L}}^K(X^K, \mathcal{H}) = \sum_{h \in \mathcal{H}} \left(\sum_{i=1}^{n_h} \ln(A^{K,h,i} X^K) \right) - \widehat{B}^K X^K \tag{9}$$

where $\widehat{B}^K_{uv,k} = \sum_{h \in \mathcal{H}} \sum_{j=1}^{n_h} 1\left\{u_j^h = v\right\} \int_{T_h^-}^{T_h^+} \exp(-k\alpha(s - t_j^h))$.

Although $\widehat{\mathcal{L}}^K(X, \mathcal{H})$ is an upper bound of the actual log-likelihood and it is not clear at first sight why its maximization should lead to large values of $\mathcal{L}^K(X, \mathcal{H})$, we point out that the difference $\widehat{\mathcal{L}}^K(X, \mathcal{H}) - \mathcal{L}^K(X, \mathcal{H})$ is only caused by intervals where there exists $u \in [1...d]$ such that $\widehat{\lambda}_u^K(t) = 0$. But maximizers of $\widehat{\mathcal{L}}^K(X, \mathcal{H})$ are very unlikely to exhibit wide range of negative values in their triggering kernels because any single event realization with a predicted nonpositive stochastic rate yields $\widehat{\mathcal{L}}^K(X, \mathcal{H}) = -\infty$. Therefore, we assume we can rely on this approximation in order to construct fast algorithms.

4.3 MEMIP: a Learning Algorithm for Fast Log-Likelihood Estimation

Since the gradient and the hessian matrix of $X \mapsto \widehat{\mathcal{L}}^K(X, \mathcal{H})$ can be computed analytically and their size does not depend on N, we derive the proposed algorithm MEMIP on the base of successive Newton optimizations. In the following, we denote by $\text{NewtonArgMax}(f, x_0)$ the result of a Newton maximization of function f with starting point x_0 using a classical backtracking linesearch method.

The main idea is to construct recursively a sequence $(\widehat{X^1}...\widehat{X^K})$ of maximizers of functions $(\widehat{\mathcal{L}}^k)_{k\in[1...K]}$ by using NewtonArgMax$(\widehat{\mathcal{L}}^{k-1}, \widehat{W}^{k-1})$ as the starting point $\widehat{W}^k$ of maximization of $\widehat{\mathcal{L}}^k$. From the estimated sequence $(\widehat{X}^1...\widehat{X}^K)$, the

Algorithm 2 Algorithm (MEMIP) for learning background rates and triggering kernels of a multivariate Hawkes process

input Mapping parameter $\alpha > 0$, maximal polynomial degree K, starting point $\widehat{W^1} \in \mathbb{R}^{d(d+1)}$
Construct $(A^{K,h,i})$ and B^K according to $O(N)$ modified version of **Algorithm 1**
$\widehat{X}^1 \leftarrow$ NewtonArgMax$(\widehat{\mathcal{L}}^1, \widehat{W}^1)$
for $k \in [2...K]$ **do**
  $\widehat{W}^k = 0$
  for $j \in [1...k-1], u \in [1...d], v \in [0...d]$ **do**
    $\widehat{W}^k_{uv,j} = \widehat{X}^{k-1}_{uv,j}$
  end for
  $\widehat{X}^k \leftarrow$ NewtonArgMax$(\widehat{\mathcal{L}}^k, \widehat{W}^k)$
end for

best value of k can be estimated by cross-validation or various other model selection techniques. Interestingly, $A^{k,h,i} = (A^{K,h,i}_{\bullet,j})_{j\in[1...k]}$ and $B^k = (B^K_{\bullet,j})_{j\in[1...k]}$ such that only $(A^{K,h,i})_{h\in\mathcal{H},i\in[1...n_h]}$ and B^K need to be computed.

Complexity of Algorithm 2. We obtain two substantial computational speed-ups compared to exact log-likelihood maximization. First, time discretization is no longer needed for the construction of B^K. Thus, vectors $(A^{K,h,i})$ and B^K can be constructed with the same procedure than **Algorithm 1** except that updates are made only on time occurence of events. Therefore, construction complexity is $O(N)$. Similarily, approximate log-likelihood evaluations are also of complexity $O(N)$. Secondly, the approximate log-likelihood is separable by type of event u : $\widehat{\mathcal{L}}^K = \sum_{u=1}^d \widehat{\mathcal{L}}^K_u$ where $\widehat{\mathcal{L}}^K_u$ only depends on background rate μ_u and triggering kernels $(g_{uv})_{v\in[1...d]}$. Maximization can thus be parallelized across the different dimensions. Note that because of the Hessian inversion at each Newton step, complexity in d of maximization of $\widehat{\mathcal{L}}^K_u$ is $O(d^3)$ for any u, which yields a $O(d^4)$ overall complexity. In cases where $N \gg d^2$ but $d^4 > N$, the use of quasi-Newton methods might therefore be preferable.

4.4 Self-Concordance Property and Numerical Convergence of MEMIP

Problem (9) can be solved by various optimisation techniques. **Algorithm 2** is actually based on the concept of *self-concordance* ([21]) that we apply to function $X \mapsto -\widehat{\mathcal{L}}^k(X, \mathcal{H})$. Self-concordant functions are, along with strongly-convex functions with Lipschitz-continuous Hessian matrices, a very important class of functions for which nonasymptotic upper bounds of the number of Newton steps

necessary to reach precision ϵ is known. More specifically, the following property holds:

Proposition 4. *Starting from a d(d+1)-dimensional vector $\widehat{W}^1$, MEMIP constructs a sequence of K estimates $(\widehat{X}^1 ... \widehat{X}^K)$ verifying for any $k \in [1...K]$, $|\widehat{\mathcal{L}}^k(\widehat{X}^k, \mathcal{H}) - \sup_X(\widehat{\mathcal{L}}_k(X, \mathcal{H}))| \leq \epsilon$ in at most $C\big(\sup_X(\widehat{\mathcal{L}}_K(X, \mathcal{H})) - \widehat{\mathcal{L}}_1(\widehat{W}^1, \mathcal{H})\big) + K(\log_2\log_2(1/\epsilon) + C\epsilon)$ Newton iterations.*

Lemma 1. *Using Newton method with backtracking line search from a starting point $x_0 \in \mathbf{R}^d$, there exists $C > 0$ depending only on the line search parameters such that the total number of Newton iterations needed to minimize a self-concordant function f up to a precision ϵ is upper bounded by $C(\sup(f) - f(x_0)) + \log_2\log_2(\frac{1}{\epsilon})$.*

Proof of Proposition 4. Self-concordance of functions $(-\widehat{\mathcal{L}}_k)_{k\in[1...K]}$ is a direct consequence of self-concordance on $\mathbf{R}_+^*$ of $f : x \mapsto -\ln(x)$ and affine invariance properties of self-concordant functions. By applying the aforementioned lemma to function $-\widehat{\mathcal{L}}_k$ and starting point $\widehat{W}^k$ at each Newton optimization, we get the bound

$$C\sum_k \big(\sup_X(\widehat{\mathcal{L}}_k(X, \mathcal{H})) - \widehat{\mathcal{L}}^k(\widehat{W}^k, \mathcal{H})\big) + K\log_2\log_2(1/\epsilon) \tag{10}$$

By construction of MEMIP iterates, we also have $\widehat{\mathcal{L}}^k(\widehat{W}^k, \mathcal{H}) = \widehat{\mathcal{L}}^{(k-1)}(\widehat{W}^k, \mathcal{H}) = \widehat{\mathcal{L}}^{(k-1)}(\widehat{X}^{k-1}, \mathcal{H})$ where the first equality holds because for any u, v, $\widehat{W}^k_{uv,k} = 0$ and the second because for any u, v, $j \leq k-1$, $\widehat{W}^{k-1}_{uv,j} = \widehat{X}^{k-1}_{uv,j}$. But for any $k \geq 2$, $\widehat{\mathcal{L}}^{k-1}(\widehat{X}^{k-1}, \mathcal{H}) \geq \sup_X(\widehat{\mathcal{L}}_{k-1}(X, \mathcal{H})) - \epsilon$. Therefore the bound reformulates as

$$C\sum_{k=1}^{K} \big(\sup_X(\widehat{\mathcal{L}}_k(X, \mathcal{H})) - \sup_X(\widehat{\mathcal{L}}_{k-1}(X, \mathcal{H}))\big) + K(\log_2\log_2(1/\epsilon) + C\epsilon) \tag{11}$$

which proves Proposition 4, using the notation $\sup_X(\widehat{\mathcal{L}}_0(X, \mathcal{H})) = \widehat{\mathcal{L}}_1(\widehat{W}^1, \mathcal{H})$. □

Remark. The previous proposition emphasizes the key role played by the starting point $\widehat{W}^1$ in the speed of convergence of Newton-like methods. In our case, a good choice is for instance to select it by classical non-negative maximization techniques for objectives of type (9) (see *e.g* [22]). Because these methods are quite fast, they can also be used for steps $k \in [2...K]$ in order to provide an alternative starting point $\widehat{W}^k_+$. The update $\widehat{X}^k$ is then given by either NewtonArgMax$(\widehat{\mathcal{L}}^k, \widehat{W}^k)$ or NewtonArgMax$(\widehat{\mathcal{L}}^k, \widehat{W}^k_+)$ depending on the most succesful maximization.

5 Experimental Results

We first evaluate MEMIP on realistic synthetic data sets. We compare it to MMEL ([9]) and fixed exponential kernels and show that MEMIP performs significantly better in terms of prediction and triggering kernels recovery.

5.1 Synthetic Data Sets: Experiment Setup and Results

Data Generation We simulate multivariate Hawkes processes by *Ogata modified thinning algorithm* (see e.g. [23]). Since each occurence can potentially increase stochastic rates of all events, special attention has to be paid to avoid *explosion*, *i.e* the occurence of an infinite number of events on a finite time window. In order to avoid such behavior, our simulated data sets verify the sufficient non-explosion condition $\rho(\Gamma) < 1$ where $\rho(\Gamma)$ denotes the spectral radius of the matrix $\Gamma = (\int_0^\infty |g_{uv}(t)dt|)_{uv}$ (see *e.g* [18]). We perform experiments on three different simulated data sets where triggering kernels are taken as

$$g_{uv}(t) = \nu_{uv} \frac{\sin\left(\frac{2\pi t}{\omega_{uv}} + \frac{\pi}{2}((u+v) \bmod 2)\right) + 2}{3(t+1)^2} \tag{12}$$

We sample the periods ω_{uv} from an uniform distribution over $[1, 10]$. Absolute values of normalization factors ν_{uv} are sampled uniformally from $[0, 1/d[$ and their sign is sampled from a Bernoulli law of parameter p. Except for the toy data set, background rates μ_v are taken constant and sampled in $[0, 0.001]$. An important feature of this choice of triggering kernels and parameters is that resulting Hawkes processes respect the aforementioned sufficient non-explosion condition. For quantitative evaluation, we simulate two quite large data sets (1) $d = 300$,$p = 1$ (2) $d = 300$,$p = 0.9$. Thus, data set (1) contains realizations of purely mutually-exciting processes whereas data set (2) has 10% of inhibitive kernels. For each data set, we sample 10 sets of parameters $(\omega_{uv}, \nu_{uv})_{u\geq 1, v\geq 1}$,$(\mu_v)_{v\geq 1}$ and simulate 400,000 i.i.d realizations of the resulting Hawkes process over $[0, 20]$. The first 200,000 are taken as training set and the remaining 200,000 as test set.

Evaluation Metrics We evaluate the different algorithms by two metrics: (a) *Diff* a normalized L^2 distance between the true and estimated triggering kernels, defined by

$$\text{Diff} = \frac{1}{d^2} \sum_{u=1}^{d} \sum_{v=1}^{d} \frac{\int (\widehat{g}_{uv} - g_{uv})^2}{\int \widehat{g}_{uv}^2 + \int g_{uv}^2} \tag{13}$$

, (b) *Pred* a prediction score on the test data set defined as follows. For each dimension $u \in [1...d]$ and occurence i in the test set, probability for that occurence to be of type u is given by $P_i^{true}(u) = \frac{\lambda_u(t_i)}{\sum_{v=1}^d \lambda_v(t_i)}$. Thus, defining $AUC(d, P)$ the area under ROC curve for binary task of predicting $(1_{\{u_i=u\}})_i$ with scores $(P_i^{true}(d))_i$ and $(P_i^{model}(d))_i$ the probabilities estimated by the evaluated model, we set

$$\text{Pred} = \frac{\sum_{u=1}^d \left(AUC(d, P^{model}) - 0.5\right)}{\sum_{u=1}^d \left(AUC(d, P^{true}) - 0.5\right)} \tag{14}$$

Baselines We compare MEMIP to (a) **MMEL** for which we try various sets of number of base kernels, total number of iterations and smoothing hyperparameter, (b) **Exp** the widely used setting where $g_{uv}(t) = \nu_{uv} \exp(-\alpha t)$ and only

ν_{uv} are estimated from the data. In order to give this baseline more flexibility and prediction power, we allow negative values of ν_{uv}. We train three different versions with $\alpha \in \{0.1, 1.0, 10.0\}$.

Results Part 1: Visualization on a Toy Dataset In order to demonstrate the ability of MEMIP to discover the underlying dynamics of Hawkes processes even in presence of inhibition and varying background rates, we construct the following toy bidimensional data set. Amongst the four triggering kernels, g_{11} is taken negative and background rates are defined by $\mu_0 = \frac{cos(\frac{2\pi t}{\omega_0})+2}{1+t}$ and $\mu_1 = \frac{sin(\frac{2\pi t}{\omega_1})+2}{1+t}$ with parameters ω_0 and ω_1 sampled in $[5, 15]$. We sample a set of parameters $(\omega_{uv}, \nu_{uv})_{u\geq 1, v\geq 1}$,$(\mu_v)_{v\geq 1}$ and simulate 200,000 i.i.d realizations of the resulting Hawkes process. From Fig. **1**, we observe that both compared methods MEMIP and MMEL accurately recover nonnegative triggering kernels g_{00}, g_{01} and g_{10}. However, MEMIP is also able to estimate the inhibitive g_{11} whereas MMEL predicts $g_{11} = 0$. Varying background rates μ_0 and μ_1 are also well estimated by MEMIP, whereas by construction MMEL and Exp only return constant values $\bar{\mu}_0$ and $\bar{\mu}_1$.

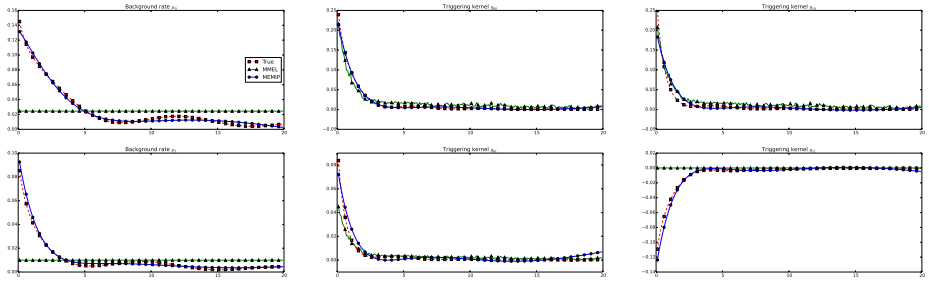

Fig. 1. Triggering kernels and background rates for toy data set estimated by MEMIP and MMEL algorithms vs true triggering kernels and background rate

Results Part 2: Prediction Score In order to evaluate **Pred** score of the competing methods on the generated data sets, we remove for each model the best and worst perfomance over the ten simulated processes, and average **Pred** over the eight remaining one. Empirical 10% confidence intervals are also indicated to assess significativity of the experimental results. From Table **1**, we observe that MEMIP significantly outperforms the competing baselines for both data sets. Prediction rates are quite low for all competing methods which indicates a rather difficult prediction problem, as $90,000$ nonparametric functions are indeed to be estimated from the data. In Fig. **2** , we study the sensitivity of **Pred** score to α and K for simulated data sets (1)(above) and (2)(below). Left plots show MEMIP and Exp **Pred** score with respect to α, as well as best MMEL average score across a broad range of hyperparameters. Empirical 10%

confidence intervals are also plotted in dashed line. We see that MEMIP gives good results in a wide range of values of α, and outperforms the exponential baseline for all values of α. Right plots show MEMIP **Pred** score with respect to K for $\alpha = 0.1$, as well as best Exp and MMEL average score. We see that MEMIP achieves good prediction results for low values of K, and that taking $K > 10$ is not necessary. For very large values of α, we also note that MEMIP and Exp baseline are the same, because the optimal choice of K for MEMIP is $K = 1$.

Table 1. Pred score for prediction of the type of next event on simulated data sets

Dataset	MEMIP	MMEL	Exp
(1) d=300,p=1	$0.288 \in [0.258, 0.310]$	$0.261 \in [0.250, 0.281]$	$0.255 \in [0.236; 0.278]$
(2) d=300,p=0.9	$0.287 \in [0.266, 0.312]$	$0.261 \in [0.241, 0.280]$	$0.256 \in [0.242, 0.280]$

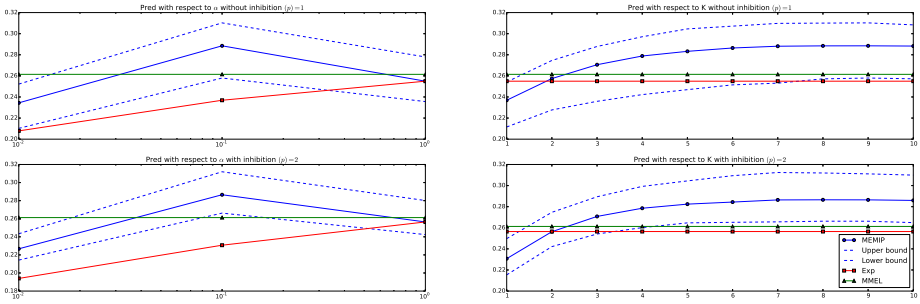

Fig. 2. Sensitivity to hyperparameters α (left) and K(right) for **Pred** score of MEMIP algorithm, compared to Exp and MMEL baselines on non-inhibitive simulated data set (above) and simulated data set with 10 % inhibitive kernels (below)

Results Part 3: Accuracy of Kernel Estimation Besides having a greater prediction power, we observe in Table **2** that MEMIP is also able to estimate the true values of triggering kernels more accurately on both data sets. In Fig. **3**, we study the sensitivity of **Diff** score to α and K for simulated data sets (1)(above) and (2)(below). We see that the variance of **Diff** score is very low for MEMIP, and its fitting error is significatively lower than those of other baselines at level 10%.

Discussion The fact that the proposed algorithm MEMIP outperforms MMEL on a non-inhibitive data set may seem surprising. Actually, even for purely mutually-exciting settings, these two algorithms can exhibit different behaviors. MMEL decomposes the triggering kernels on a low-rank set of basis functions,

Table 2. Diff score for triggering kernels recovery on simulated data sets

Dataset	MEMIP	MMEL	Exp
(1) d=300,p=1	$0.759 \in [0.755, 0.768]$	$0.807 \in [0.803, 0.814]$	$0.791 \in [0.788, 0.800]$
(2) d=300,p=0.9	$0.803 \in [0.793, 0.810]$	$0.839 \in [0.833, 0.844]$	$0.830 \in [0.818, 0.836]$

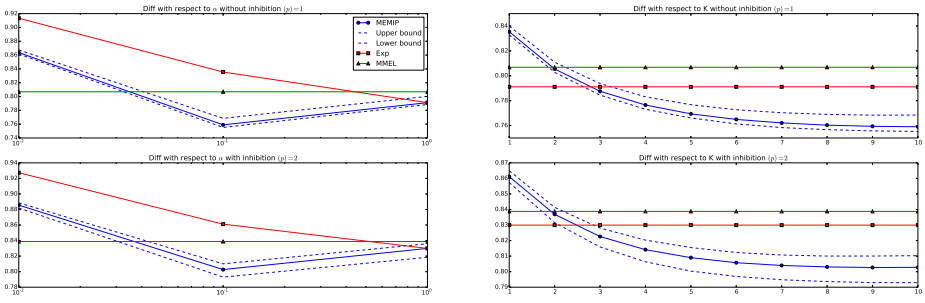

Fig. 3. Sensitivity to hyperparameters α (left) and K(right) for **Diff** score of MEMIP algorithm, compared to Exp and MMEL baselines on non-inhibitive simulated data set (above) and simulated data set with 10 % inhibitive kernels (below)

whereas we fix our basis functions as exponentials, in order to enjoy fast global convergence and ability to learn negative projection coefficients $X_{uv,k}$. Smoothing strategy also plays a key role in experimental results. Indeed, because the log-likelihood (1) can be made arbitrarily high by the sequence of functions $(g^n_{uv})_{n\in N}$ defined by $g^n_{uv}(t) = n1_{\{t\in T_{uv}\}}$ where $T_{uv} = \{t_v - t_u \mid (t_u < t_v \wedge (\exists h \in \mathcal{H} \mid (t_v, v) \in h \wedge (t_u, u) \in h))\}$, smoothing is mandatory when learning triggering kernels by means of log-likelihood maximization. Using a L^2 roughness norm penalization $\alpha \int_0^T g'^2$, MMEL can face difficult dilemmas when fitting power-laws fastly decaying around 0 : either under-estimating the rate when it is at its peak or lowering the smoothness parameter and being vulnerable to overfitting. On the contrary, MEMIP would face difficulties to perfectly fit periodic functions with a very small period, as the derivative of its order K estimates can only vanish $K - 1$ times.

5.2 Experiment on the MemeTracker Data Set

In order to show that the ability to estimate inhibitive triggering kenels and varying background rates yields better accuracy on real-world data sets, we compare the proposed method MEMIP to different baselines on the MemeTracker data set, following the experience plan exposed in [9]. MemeTracker contains links creation between some of the most popular websites between August 2008 and April 2009. We extract link creations between the top 100 popular websites and define the occurence of an event for the i^{th} website as a link creation on this website to one the 99 other websites. We then use half of the data set as training

data and the other half at test data on which each baseline is evaluated by average area under ROC curve for predicting future events. From Fig. **4**, we observe that the proposed method MEMIP achieves a better prediction score than both baselines. Left plot shows MEMIP and Exp prediction score with respect to α, as well as best MMEL score across a broad range of hyperparameters. We see that MEMIP gives good results in a very broad range of values of α, and significantly outperforms the exponential baseline for all values of α. Right plot shows MEMIP prediction score with respect to K for $\alpha = 0.01$, as well as best Exp and MMEL score. For $K = 10$, MEMIP achieves a prediction score of 0.8021 whereas best MMEL and Exp score are respectively 0.6928 and 0.7716. We note that, even for K as low as 3, MEMIP performs the prediction task quite accurately.

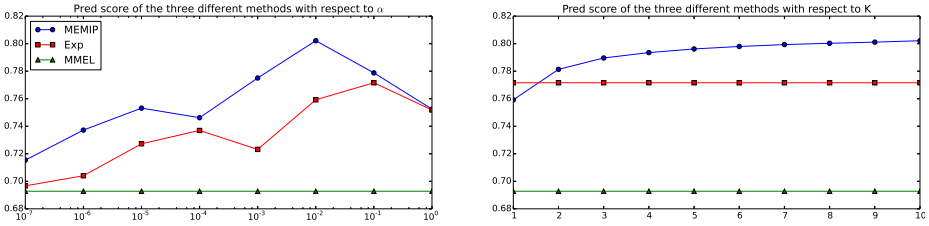

Fig. 4. Sensitivity to hyperparameters α (left) and K(right) for prediction score of MEMIP algorithm, compared to Exp and MMEL baselines on MemeTracker data set

6 Conclusions

In this paper, we propose MEMIP, which is to our knowledge the first method to learn nonparametrically triggering kernels of multivariate Hawkes processes in presence of inhibition and varying background rates. By relying on results of approximation theory, the triggering kernels are decomposed on a basis on memoryless exponential kernels. This maximization of the log-likelihood is then shown to reformulate as a concave maximization problem, that can be solved in linear complexity thanks to the Markov property verified by the proposed estimates. Experimental results on both synthetic and real-world data sets show that the proposed model is able to learn more accurately the underlying dynamics of Hawkes processes and therefore has a greater prediction power.

References

1. Ogata, Y.: Statistical models for earthquake occurrences and residual analysis for point processes. Journal of the American Statistical Association **83**(401) (1988) 9–27

2. Errais, E., Giesecke, K., Goldberg, L.R.: Pricing credit from the top down with affine point processes. Numerical Methods for Finance (2007) 195–201
3. Bauwens, L., Hautsch, N.: Modelling financial high frequency data using point processes. Springer (2009)
4. Bacry, E., Delattre, S., Hoffmann, M., Muzy, J.F.: Modelling microstructure noise with mutually exciting point processes. Quantitative Finance **13**(1) (2013) 65–77
5. Alfonsi, A., Blanc, P.: Dynamic optimal execution in a mixed-market-impact Hawkes price model. arXiv preprint arXiv:1404.0648 (2014)
6. Mohler, G.O., Short, M.B., Brantingham, P.J., Schoenberg, F.P., Tita, G.E.: Self-exciting point process modeling of crime. Journal of the American Statistical Association **106**(493) (2011) 100–108
7. Reynaud-Bouret, P., Schbath, S.: Adaptive estimation for Hawkes processes; application to genome analysis. The Annals of Statistics **38**(5) (2010) 2781–2822
8. Blundell, C., Beck, J., Heller, K.A.: Modelling reciprocating relationships with Hawkes processes. In: Advances in Neural Information Processing Systems. (2012) 2609–2617
9. Zhou, K., Zha, H., Song, L.: Learning triggering kernels for multi-dimensional Hawkes processes. In: Proceedings of the 30th International Conference on Machine Learning (ICML-13). (2013) 1301–1309
10. Hawkes, A.G., Oakes, D.: A cluster process representation of a self-exciting process. Journal of Applied Probability (1974) 493–503
11. Brémaud, P., Massoulié, L.: Stability of nonlinear Hawkes processes. The Annals of Probability (1996) 1563–1588
12. Ogata, Y.: The asymptotic behaviour of maximum likelihood estimators for stationary point processes. Annals of the Institute of Statistical Mathematics **30**(1) (1978) 243–261
13. Bacry, E., Dayri, K., Muzy, J.F.: Non-parametric kernel estimation for symmetric Hawkes processes. Application to high frequency financial data. The European Physical Journal B **85**(5) (2012) 1–12
14. Crane, R., Sornette, D.: Robust dynamic classes revealed by measuring the response function of a social system. Proceedings of the National Academy of Sciences **105**(41) (2008) 15649–15653
15. Lewis, E., Mohler, G.: A nonparametric EM algorithm for multiscale Hawkes processes. Joint Statisticals Meetings 2011 (2011)
16. Hansen, N.R., Reynaud-Bouret, P., Rivoirard, V.: Lasso and probabilistic inequalities for multivariate point processes. arXiv preprint arXiv:1208.0570 (2012)
17. Lewis, E., Mohler, G., Brantingham, P.J., Bertozzi, A.L.: Self-exciting point process models of civilian deaths in Iraq. Security Journal **25**(3) (2011) 244–264
18. Daley, D.J., Vere-Jones, D.: An introduction to the theory of point processes. Springer (2007)
19. Bernstein, S.: Sur l'ordre de la meilleure approximation des fonctions continues par des polynômes de degré donné. Volume 4. Hayez, imprimeur des académies royales (1912)
20. Cheney, E.W., Cheney, E.W.: Introduction to approximation theory. Volume 3. McGraw-Hill New York (1966)
21. Nesterov, Y., Nemirovskii, A.S., Ye, Y.: Interior-point polynomial algorithms in convex programming. Volume 13. SIAM (1994)
22. Seung, D., Lee, L.: Algorithms for non-negative matrix factorization. Advances in neural information processing systems **13** (2001) 556–562
23. Liniger, T.J.: Multivariate Hawkes processes. PhD thesis, Diss., Eidgenössische Technische Hochschule ETH Zürich, Nr. 18403 (2009)