Uncovering Causality from Multivariate Hawkes Integrated Cumulants

Massil Achab [1] **Emmanuel Bacry** [1] **Stéphane Gaïffas** [1] **Iacopo Mastromatteo** [2] **Jean-François Muzy** [1 3]

Abstract

We design a new nonparametric method that allows one to estimate the matrix of integrated kernels of a multivariate Hawkes process. This matrix not only encodes the mutual influences of each node of the process, but also disentangles the causality relationships between them. Our approach is the first that leads to an estimation of this matrix *without any parametric modeling and estimation of the kernels themselves*. As a consequence, it can give an estimation of causality relationships between nodes (or users), based on their activity timestamps (on a social network for instance), without knowing or estimating the shape of the activities lifetime. For that purpose, we introduce a moment matching method that fits the second-order and the third-order integrated cumulants of the process. A theoretical analysis allows us to prove that this new estimation technique is consistent. Moreover, we show on numerical experiments that our approach is indeed very robust to the shape of the kernels, and gives appealing results on the MemeTracker database and on financial order book data.

1. Introduction

In many applications, one needs to deal with data containing a very large number of irregular timestamped events that are recorded in continuous time. These events can reflect, for instance, the activity of users on a social network (Subrahmanian et al., 2016), high-frequency variations of signals in finance (Bacry & Muzy, 2014), earthquakes and aftershocks in geophysics (Ogata, 1998), crime activity (Mohler et al., 2011) or position of genes in genomics (Reynaud-Bouret & Schbath, 2010). In this context, multidimensional counting processes based models play a paramount role. Within this framework, an important task is to recover the mutual influence of the nodes, by leveraging on their timestamp patterns (Gomez-Rodriguez et al., 2013; Farajtabar et al., 2015; Xu et al., 2016).

Consider a set of nodes $I = \{1, \dots, d\}$. For each $i \in I$, we observe a set Z^i of *events*, where any $\tau \in Z^i$ labels the occurrence time of an event related to the activity of i. The events of all nodes can be represented as a vector of counting processes $\boldsymbol{N}_t = [N_t^1 \cdots N_t^d]^\top$, where N_t^i counts the number of events of node i until time $t \in \mathbb{R}^+$, namely $N_t^i = \sum_{\tau \in Z^i} \mathbb{1}_{\{\tau \leq t\}}$. The vector of stochastic intensities $\boldsymbol{\lambda}_t = [\lambda_t^1 \cdots \lambda_t^d]^\top$ associated with the multivariate counting process $\boldsymbol{N}_t$ is defined as

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

for $i \in I$, where the filtration $\mathcal{F}_t$ encodes the information available up to time t. The coordinate λ_t^i gives the expected instantaneous rate of event occurrence at time t for node i. The vector $\boldsymbol{\lambda}_t$ characterizes the distribution of $\boldsymbol{N}_t$, see (Daley & Vere-Jones, 2003), and patterns in the events time-series can be captured by structuring these intensities.

1.1. Hawkes processes

The Hawkes process framework (Hawkes, 1971) corresponds to an autoregressive structure of the intensities in order to capture self-excitation and cross-excitation of nodes, which is a phenomenon typically observed in social networks (Crane & Sornette, 2008). Namely, $\boldsymbol{N}_t$ is called a *Hawkes point process* if the stochastic intensities can be written as

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

where $\mu^i \in \mathbb{R}^+$ is an exogenous intensity and ϕ^{ij} are positive, integrable and causal (with support in $\mathbb{R}_+$) functions called *kernels* encoding the impact of an action by node j on the activity of node i. Note that when all kernels are zero, the process is a simple homogeneous multivariate Poisson process.

1.2. Related works

Most papers use very simple parameterizations of the kernels (Yang & Zha, 2013; Zhou et al., 2013b; Farajtabar

[1] Ecole Polytechnique, Palaiseau, France [2] Capital Fund Management, Paris, France [3] Université de Corse, Corte, France. Correspondence to: Massil Achab <massil.achab@m4x.org>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

et al., 2015), they are of the form $\phi^{ij}(t) = \alpha_{ij} h(t)$ with $\alpha_{ij} \in \mathbb{R}^+$ quantifying the intensity of the influence of j on i and $h(t)$ a (normalized) function that characterizes the time-profile of this influence and that is *shared* by all couples of nodes (i, j) (most often, it is chosen to be either exponential $h(t) = \beta e^{-\beta t}$ or power-law $h(t) = \beta t^{-(\beta+1)}$). This is highly non-realistic: there is a priori no reason for assuming that the time-profile of the influence of a node j on a node i does not depend on the pair (i, j). Moreover, assuming an exponential shape or a power-law shape for $h(t)$ arbitrarily imposes an event impact that is always instantly maximal, and that can only decrease with time, while in practice, there may exist a latency between an event and its impact.

In order to improve this and have more flexibility on the shape of the kernels, nonparametric estimation is considered in (Lewis & Mohler, 2011) and extended to the multi-dimensional case in (Zhou et al., 2013a). An alternative method is proposed in (Bacry & Muzy, 2016) where non-parametric estimation is formulated as a Wiener-Hopf equation. Another nonparametric strategy considers a decomposition of kernels on a dictionary of function $h_1, \ldots, h_K$, namely $\phi^{ij}(t) = \sum_{k=1}^{K} a_k^{ij} h_k(t)$, where the coefficients a_k^{ij} are estimated, see (Hansen et al., 2015; Lemonnier & Vayatis, 2014) and (Xu et al., 2016), where group-lasso is used to induce a sparsity pattern on the coefficients a_k^{ij} that is shared across $k = 1, \ldots, K$.

Such methods are heavy when d is large, since they rely on likelihood maximization or least squares minimization within an over-parametrized space in order to gain flexibility on the shape of the kernels. This is problematic, since the original motivation for the use of Hawkes processes is to estimate the influence and causality of nodes, the knowledge of the full parametrization of the model being of little interest by itself.

1.3. Granger Causality

Since the question of a *real causality* is too complex in general, most econometricians agreed on the simpler definition of Granger causality (Granger, 1969). Its mathematical formulation is a statistical hypothesis test: X causes Y *in the sense of Granger causality* if forecasting future values of Y is more successful while taking X past values into account. In (Eichler et al., 2016), it is shown that for $\boldsymbol{N}_t$ a multivariate Hawkes process, N_t^j does not Granger-cause N_t^i w.r.t $\boldsymbol{N}_t$ if and only if $\phi^{ij}(u) = 0$ for $u \in \mathbb{R}_+$. Since the kernels take positive values, the latter condition is equivalent to $\int_0^\infty \phi^{ij}(u) du = 0$.

In the following, we'll refer to *learning the kernels' integrals* as *uncovering causality* since each integral encodes the notion of Granger causality, and is also linked to the number of events directly caused from a node to another node, as described below at Eq. (2).

1.4. Our contribution: cumulants matching

Our paper solves this problem with a different and more direct approach. Instead of trying to estimate the kernels ϕ^{ij}, we focus on the direct estimation of their *integrals*. Namely, we want to estimate the matrix $\boldsymbol{G} = [g^{ij}]$ where

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

From the definition of Hawkes process as a Poisson cluster process (Jovanović et al., 2015), g^{ij} can be simply interpreted as the average total number of events of node i whose *direct* ancestor is a given event of node j (by direct we mean that interactions mediated by any other intermediate event are not counted). In that respect, $\boldsymbol{G}$ not only describes the mutual influences between nodes, but it also quantifies their *direct causal* relationships. Namely, introducing the counting function $N_t^{i \leftarrow j}$ that counts the number of events of i whose direct ancestor is an event of j, we know from (Bacry et al., 2015) that

$$\mathbb{E}[dN_t^{i \leftarrow j}] = g^{ij} \mathbb{E}[dN_t^j] = g^{ij} \Lambda^j dt, \quad (2)$$

where we introduced Λ^i as the intensity expectation, namely satisfying $\mathbb{E}[dN_t^i] = \Lambda^i dt$. Note that Λ^i does not depend on time by stationarity of $\boldsymbol{N}_t$, which is known to hold under the *stability condition* $\|\boldsymbol{G}\| < 1$, where $\|\boldsymbol{G}\|$ stands for the spectral norm of $\boldsymbol{G}$. In particular, this condition implies the non-singularity of $\boldsymbol{I_d} - \boldsymbol{G}$.

The main idea is to estimate the matrix $\boldsymbol{G}$ directly using a matching cumulants (or moments) method. Indeed the cumulants write as centered moments, up to the third order. For higher order, they are computable using the cumulant generating function. First, we compute an estimation $\widehat{\boldsymbol{M}}$ of some moments $M(\boldsymbol{G})$, that are uniquely defined by $\boldsymbol{G}$. Second, we look for a matrix $\widehat{\boldsymbol{G}}$ that minimizes the L^2 error $\|M(\widehat{\boldsymbol{G}}) - \widehat{\boldsymbol{M}}\|^2$. This approach turns out to be particularly robust to the kernel shapes, which is not the case of all previous approaches for causality recovery with the Hawkes model. We call this method NPHC (Non Parametric Hawkes Cumulant), since our approach is of nonparametric nature. This new approach is confirmed by a theoretical analysis allowing to prove the consistency of the NPHC estimator, by using tools from Generalized Method of Moments, see (Hall, 2005), and a technical original proof that is detailed in the supplementary material. Note that moment and cumulant matching techniques proved particularly powerful for latent topic models, in particular Latent Dirichlet Allocation, see (Podosinnikova et al., 2015). Previous works (Da Fonseca & Zaatour, 2014; Aït-Sahalia et al., 2010) already used method of moments with Hawkes processes, but only in a parametric setting. Our work is the first to consider such an approach for a nonparametric counting processes framework.

2. NPHC: The Non Parametric Hawkes Cumulant method

The simplest moment-based quantities M that can be explicitly written as a function of $\boldsymbol{G}$ are the integrated cumulants of the Hawkes process.

2.1. Integrated cumulants of the Hawkes process

A general formula for these cumulants is provided in (Jovanović et al., 2015) but, as explained below, for the purpose of our method, we only need to consider cumulants up to the third order. Given $1 \le i, j, k \le d$, the first three integrated cumulants of the Hawkes process can be defined as follows thanks to stationarity:

$$\Lambda^i dt = \mathbb{E}(dN_t^i) \tag{3}$$

$$C^{ij} dt = \int_{\tau \in \mathbb{R}} \Big(\mathbb{E}(dN_t^i dN_{t+\tau}^j) - \mathbb{E}(dN_t^i)\mathbb{E}(dN_{t+\tau}^j) \Big) \tag{4}$$

$$\begin{aligned} K^{ijk} dt = \iint_{\tau, \tau' \in \mathbb{R}^2} \Big(& \mathbb{E}(dN_t^i dN_{t+\tau}^j dN_{t+\tau'}^k) \\ & + 2\mathbb{E}(dN_t^i)\mathbb{E}(dN_{t+\tau}^j)\mathbb{E}(dN_{t+\tau'}^k) \\ & - \mathbb{E}(dN_t^i dN_{t+\tau}^j)\mathbb{E}(dN_{t+\tau'}^k) \\ & - \mathbb{E}(dN_t^i dN_{t+\tau'}^k)\mathbb{E}(dN_{t+\tau}^j) \\ & - \mathbb{E}(dN_{t+\tau}^j dN_{t+\tau'}^k)\mathbb{E}(dN_t^i) \Big), \end{aligned} \tag{5}$$

where Eq. (3) is the mean intensity of the Hawkes process, the second-order cumulant (4) refers to the integrated covariance density matrix and the third-order cumulant (5) measures the skewness of $\boldsymbol{N}_t$. Using the Laplace transform (Bacry & Muzy, 2016) or the Poisson cluster process representation (Jovanović et al., 2015), one can obtain an explicit relationship between these integrated cumulants and the matrix $\boldsymbol{G}$. If one sets

$$\boldsymbol{R} = (\boldsymbol{I_d} - \boldsymbol{G})^{-1}, \tag{6}$$

straightforward computations (see Section 5) lead to the following identities:

$$\Lambda^i = \sum_{m=1}^{d} R^{im} \mu^m \tag{7}$$

$$C^{ij} = \sum_{m=1}^{d} \Lambda^m R^{im} R^{jm} \tag{8}$$

$$\begin{aligned} K^{ijk} = \sum_{m=1}^{d} (& R^{im} R^{jm} C^{km} + R^{im} C^{jm} R^{km} \\ & + C^{im} R^{jm} R^{km} - 2\Lambda^m R^{im} R^{jm} R^{km}). \end{aligned} \tag{9}$$

Our strategy is to use a convenient subset of Eqs. (3), (4) and (5) to define $\boldsymbol{M}$, while we use Eqs. (7), (8) and (9) in order to construct the operator that maps a candidate matrix $\boldsymbol{R}$ to the corresponding cumulants $M(\boldsymbol{R})$. By looking for $\widehat{\boldsymbol{R}}$ that minimizes $\boldsymbol{R} \mapsto \|M(\boldsymbol{R}) - \widehat{\boldsymbol{M}}\|^2$, we obtain, as illustrated below, good recovery of the ground truth matrix $\boldsymbol{G}$ using Equation (6).

The simplest case $d = 1$ has been considered in (Hardiman & Bouchaud, 2014), where it is shown that one can choose $M = \{C^{11}\}$ in order to compute the kernel integral. Eq. (8) then reduces to a simple second-order equation that has a unique solution in $\boldsymbol{R}$ (and consequently a unique $\boldsymbol{G}$) that accounts for the stability condition ($\|\boldsymbol{G}\| < 1$).

Unfortunately, for $d > 1$, the choice $M = \{C^{ij}\}_{1 \le i \le j \le d}$ is not sufficient to uniquely determine the kernels integrals. In fact, the integrated covariance matrix provides $d(d+1)/2$ independent coefficients, while d^2 parameters are needed. It is straightforward to show that the remaining $d(d-1)/2$ conditions can be encoded in an orthogonal matrix $\boldsymbol{O}$, reflecting the fact that Eq. (8) is invariant under the change $\boldsymbol{R} \to \boldsymbol{OR}$, so that the system is under-determined.

Our approach relies on using the third order cumulant tensor $\boldsymbol{K} = [K^{ijk}]$ which contains $(d^3 + 3d^2 + 2d)/6 > d^2$ independent coefficients that are sufficient to uniquely fix the matrix $\boldsymbol{G}$. This can be justified intuitively as follows: while the integrated covariance only contains symmetric information, and is thus unable to provide causal information, *the skewness given by the third order cumulant in the estimation procedure can break the symmetry between past and future so as to uniquely fix* $\boldsymbol{G}$. Thus, our algorithm consists of selecting d^2 third-order cumulant components, namely $M = \{K^{iij}\}_{1 \le i,j \le d}$. In particular, we define the estimator of $\boldsymbol{R}$ as $\widehat{\boldsymbol{R}} \in \operatorname{argmin}_{\boldsymbol{R}} \mathcal{L}(\boldsymbol{R})$, where

$$\mathcal{L}(\boldsymbol{R}) = (1 - \kappa)\|\boldsymbol{K^c}(\boldsymbol{R}) - \widehat{\boldsymbol{K^c}}\|_2^2 + \kappa\|\boldsymbol{C}(\boldsymbol{R}) - \widehat{\boldsymbol{C}}\|_2^2, \tag{10}$$

where $\| \cdot \|_2$ stands for the Frobenius norm, $\boldsymbol{K^c} = \{K^{iij}\}_{1 \le i,j \le d}$ is the matrix obtained by the contraction of the tensor $\boldsymbol{K}$ to d^2 indices, $\boldsymbol{C}$ is the covariance matrix, while $\widehat{\boldsymbol{K^c}}$ and $\widehat{\boldsymbol{C}}$ are their respective estimators, see Equations (12), (13) below. It is noteworthy that the above mean square error approach can be seen as a peculiar Generalized Method of Moments (GMM), see (Hall, 2005). This framework allows to determine the optimal weighting matrix involved in the loss function, which is a question to be addressed in an extended version of the present work. In this work, we use the coefficient κ to scale the two terms, by setting $\kappa = \|\widehat{\boldsymbol{K^c}}\|_2^2 / (\|\widehat{\boldsymbol{K^c}}\|_2^2 + \|\widehat{\boldsymbol{C}}\|_2^2)$. Finally the estimator of $\boldsymbol{G}$ is straightforwardly obtained as

$$\widehat{\boldsymbol{G}} = \boldsymbol{I_d} - \widehat{\boldsymbol{R}}^{-1},$$

from the inversion of Eq. (6). Let us mention an important point: the matrix inversion in the previous formula is not the bottleneck of the algorithm. Indeed, its has a complexity $O(d^3)$ which is cheap compared to the computation of the cumulants when $n = \max_i |Z^i| \gg d$, which is typically satisfied (see next subsection). Solving the considered problem on a larger scale, say $d \gg 10^3$, is an open question, even with state-of-the-art parametric and nonparametric approaches (Zhou et al., 2013a; Xu et al., 2016; Zhou et al., 2013b; Bacry & Muzy, 2016), where the number of components d in experiments is always around 100 or smaller. Actually, our approach leads to a *much faster* algorithm than the considered state-of-the-art baselines (see Tables 1–4 below).

2.2. Estimation of the integrated cumulants

In this section we present explicit formulas to estimate the three moment-based quantities listed in the previous section, namely, $\boldsymbol{\Lambda}$, $\boldsymbol{C}$ and $\boldsymbol{K}$. We first assume there exists $H > 0$ such that the truncation from $(-\infty, +\infty)$ to $[-H, H]$ of the domain of integration of the quantities appearing in Eqs. (4) and (5), introduces only a small error. In practice, this amounts to neglecting border effects in the covariance density and in the skewness density, and it is a good approximation if i) the support of the kernel $\phi^{ij}(t)$ is smaller than H and ii) the spectral norm $\|\boldsymbol{G}\|$ is sufficiently distant from the critical point $\|\boldsymbol{G}\| = 1$.

In this case, given a realization of a stationary Hawkes process $\{\boldsymbol{N}_t : t \in [0, T]\}$, as shown in Section 5, we can write the estimators of the first three cumulants (3), (4) and (5) as

$$\widehat{\Lambda}^i = \frac{1}{T}\sum_{\tau \in Z^i} 1 = \frac{N_T^i}{T} \tag{11}$$

$$\widehat{C}^{ij} = \frac{1}{T}\sum_{\tau \in Z^i}\left(N^j_{\tau+H} - N^j_{\tau-H} - 2H\widehat{\Lambda}^j\right) \tag{12}$$

$$\begin{aligned}\widehat{K}^{ijk} = &\frac{1}{T}\sum_{\tau \in Z^i}\left(N^j_{\tau+H} - N^j_{\tau-H} - 2H\widehat{\Lambda}^j\right)\\ &\cdot\left(N^k_{\tau+H} - N^k_{\tau-H} - 2H\widehat{\Lambda}^k\right)\\ &-\frac{\widehat{\Lambda}^i}{T}\sum_{\tau \in Z^j}\sum_{\tau' \in Z^k}(2H - |\tau' - \tau|)^+\\ &+ 4H^2\widehat{\Lambda}^i\widehat{\Lambda}^j\widehat{\Lambda}^k.\end{aligned} \tag{13}$$

Let us mention the following facts.

Bias. While the first cumulant $\hat{\Lambda}^i$ is an unbiased estimator of Λ^i, the other estimators $\widehat{C}^{ij}$ and $\widehat{K}^{ijk}$ introduce a bias. However, as we will show, in practice this bias is small and hardly affects numerical estimations (see Section 3). This is confirmed by our theoretical analysis, which proves that if H does not grow to fast compared to T, then these estimated cumulants are consistent estimators of the theoretical cumulants (see Subsection 2.5).

Complexity. The computations of all the estimators of the first, second and third-order cumulants have complexity respectively $O(nd)$, $O(nd^2)$ and $O(nd^3)$, where $n = \max_i |Z^i|$. However, our algorithm requires a lot less than that: it computes only d^2 third-order terms, of the form $\widehat{K}^{iij}$, leaving us with only $O(nd^2)$ operations to perform.

Symmetry. While the values of Λ^i, C^{ij} and K^{ijk} are symmetric under permutation of the indices, their estimators are generally not symmetric. We have thus chosen to symmetrize the estimators by averaging their values over permutations of the indices. Worst case is for the estimator of $\boldsymbol{K^c}$, which involves only an extra factor of 2 in the complexity.

2.3. The NPHC algorithm

The objective to minimize in Equation (10) is non-convex. More precisely, the loss function is a polynomial of $\boldsymbol{R}$ of degree 10. However, by replacing $\boldsymbol{\Lambda}$ and $\boldsymbol{C}$ appearing in Eq. (4) and (5) with $\widehat{\boldsymbol{\Lambda}}$ and $\widehat{\boldsymbol{C}}$ helps us to decrease the degree from 10 to 6, which makes the optimization problem less difficult. We denote $\widetilde{\mathcal{L}}(\boldsymbol{R})$ this simpler objective function, where the expectations of cumulants Λ^i and C^{ij} have been replaced with their estimators in the right-hand side of Eqs. (8) and (9). Thanks to (Choromanska et al., 2015), we know that the loss function of a typical multilayer neural network with simple nonlinearities can be expressed as a polynomial function of the weights in the network, whose degree is the number of layers. Since the loss function of NPHC writes as a polynomial of degree 6, we expect good results using optimization methods designed to train deep multilayer neural networks. We used AdaGrad (Duchi et al., 2011), a variant of the Stochastic Gradient Descent algorithm. It scales the learning rate coordinate-wise using the online variance of the previous gradients, in order to captures second-order information. Our problem being non-convex, the choice of the starting point has a major effect on the convergence. Here, the key is to notice that the matrices $\boldsymbol{R}$ that match relation Eq. (8) writes $\boldsymbol{C}^{1/2}\boldsymbol{O}\boldsymbol{L}^{-1/2}$, with $\boldsymbol{L} = \mathrm{diag}(\boldsymbol{\Lambda})$ and $\boldsymbol{O}$ an orthogonal matrix. In our setting, this algorithm gave nice convergence results for $\boldsymbol{O} = \boldsymbol{I_d}$. The NPHC method is described schematically in Algorithm 1.

Even though our main concern is to retrieve the matrix $\boldsymbol{G}$, let us notice we can also obtain an estimation of the baseline intensities' from Eq. (3): $\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{R}}^{-1}\widehat{\boldsymbol{\Lambda}}$. An efficient implementation of this algorithm with TensorFlow, see (Abadi et al., 2016), is available on GitHub: `https://github.com/achab/nphc`.

Algorithm 1 Non Parametric Hawkes Cumulant method

```
Input: $\boldsymbol{N_t}$
Output: $\widehat{\boldsymbol{G}}$
Estimate $\widehat{\Lambda}^i$, $\widehat{C}^{ij}$, $\widehat{K}^{iij}$ from Eqs. (11, 12, 13)
Design $\widetilde{\mathcal{L}}(\boldsymbol{R})$ using the computed estimators.
Minimize numerically $\widetilde{\mathcal{L}}(\boldsymbol{R})$ so as to obtain $\widehat{\boldsymbol{R}}$
Return $\widehat{\boldsymbol{G}} = \boldsymbol{I_d} - \widehat{\boldsymbol{R}}^{-1}$.
```

Table 1. Complexity of state-of-the-art methods. NPHC's complexity is very low since, especially in the regime $n \gg d$.

Method	Total complexity
ODE (Zhou et al., 2013a)	$O(N_{\text{iter}} M(n^3 d^2 + L(nd + n^2)))$
GC (Xu et al., 2016)	$O(N_{\text{iter}} M n^3 d^2)$
ADM4 (Zhou et al., 2013b)	$O(N_{\text{iter}} n^3 d^2)$
WH (Bacry & Muzy, 2016)	$O(nd^2 L + d^4 L^3)$
NPHC	$O(nd^2 + N_{\text{iter}} d^3)$

2.4. Complexity of the algorithm

Compared with existing state-of-the-art methods to estimate the kernel functions, e.g., the ordinary differential equations-based (ODE) algorithm in (Zhou et al., 2013a), the Granger Causality-based algorithm in (Xu et al., 2016), the ADM4 algorithm in (Zhou et al., 2013b), and the Wiener-Hopf-based algorithm in (Bacry & Muzy, 2016), our method has a very competitive complexity. This can be understood by the fact that those methods estimate the kernel functions, while in NPHC we only estimate their integrals. Let us recall that d is the number of components and $n = \max_i |Z^i| \gg d$ the maximum number of events on a single component. Let us recall complexities given in (Xu et al., 2016) together with other ones. The ODE-based algorithm is an EM algorithm that parametrizes the kernel function with M basis functions, each being discretized to L points. The basis functions are updated after solving M Euler-Lagrange equations. The complexity of one iteration of the algorithm is then $O(Mn^3d^2 + ML(nd + n^2))$, with n the maximum number of events and d the dimension. The Granger Causality-based algorithm is similar to the previous one, without the update of the basis functions, that are Gaussian kernels. The complexity per iteration is $O(Mn^3d^2)$. The algorithm ADM4 is similar to the two algorithms above, as EM algorithm as well, with only one exponential kernel as basis function. The complexity per iteration is then $O(n^3d^2)$. The Wiener-Hopf-based algorithm is not iterative, on the contrary to the previous ones. It first computes the empirical conditional laws on many points, and then invert the Wiener-Hopf system, leading to a $O(nd^2L + d^4L^3)$ computation. Similarly, our method first computes the integrated cumulants, then minimize the objective function with N_{iter} iterations, and invert the resulting matrix $\widehat{\boldsymbol{R}}$ to obtain $\widehat{\boldsymbol{G}}$. At the end, the complexity of the NPHC method is $O(nd^2 + N_{\text{iter}}d^3)$. This is summarized in Table 1

2.5. Theoretical guarantee: consistency

The NPHC method can be phrased using the framework of the Generalized Method of Moments (GMM). GMM is a generic method for estimating parameters in statistical models. In order to apply GMM, we have to find a vector-valued function $g(X, \theta)$ of the data, where X is distributed with respect to a distribution $\mathbb{P}_{\theta_0}$, which satisfies the *moment conditions*: $\mathbb{E}[g(X, \theta)] = 0$ if and only if $\theta = \theta_0$, where θ_0 is the "ground truth" value of the parameter. Based on i.i.d. observed copies $x_1, \ldots, x_n$ of X, the GMM method minimizes the norm of the empirical mean over n samples, $\| \frac{1}{n} \sum_{i=1}^n g(x_i, \theta) \|$, as a function of θ, to obtain an estimate of θ_0. In the theoretical analysis of NPHC, we use ideas from the consistency proof of the GMM, but the proof actually relies on very different arguments. Indeed, the integrated cumulants estimators used in NPHC are not unbiased, as the theory of GMM requires, but asymptotically unbiased. Moreover, the setting considered here, where data consists of a single realization $\{\boldsymbol{N}_t\}$ of a Hawkes process strongly departs from the standard i.i.d setting. Our approach is therefore based on the GMM idea but the proof is actually not using the theory of GMM.

Now, we use the subscript T to refer quantities used or computed when we observe the process on $(\boldsymbol{N_t})$ on $[0, T]$, like the truncation term H_T, the estimated integrated covariance $\widehat{\boldsymbol{C}}_T$, or the estimated kernel norm matrix $\widehat{\boldsymbol{G}}_T$. In the next equation, $\odot$ stands for the Hadamard product and $\odot 2$ stands for the entrywise square of a matrix. We denote $\boldsymbol{G}_0 = \boldsymbol{I_d} - \boldsymbol{R}_0^{-1}$ the true value of $\boldsymbol{G}$, and the $\mathbb{R}^{2d \times d}$ valued vector functions

$$g_0(\boldsymbol{R}) = \begin{bmatrix} \boldsymbol{C} - \boldsymbol{R}\boldsymbol{L}\boldsymbol{R}^\top \\ \boldsymbol{K^c} - \boldsymbol{R}^{\odot 2}\boldsymbol{C}^\top - 2[\boldsymbol{R} \odot (\boldsymbol{C} - \boldsymbol{R}\boldsymbol{L})]\boldsymbol{R}^\top \end{bmatrix}$$

$$\widehat{g}_T(\boldsymbol{R}) = \begin{bmatrix} \widehat{\boldsymbol{C}}_T - \boldsymbol{R}\widehat{\boldsymbol{L}}_T\boldsymbol{R}^\top \\ \widehat{\boldsymbol{K}^{\boldsymbol{c}}_{\boldsymbol{T}}} - \boldsymbol{R}^{\odot 2}(\widehat{\boldsymbol{C}}_T)^\top - 2[\boldsymbol{R} \odot (\widehat{\boldsymbol{C}}_T - \boldsymbol{R}\widehat{\boldsymbol{L}}_T)]\boldsymbol{R}^\top \end{bmatrix}$$

so that $\widetilde{\mathcal{L}}_T(\boldsymbol{R})$ is a weighted squared Frobenius norm of $\widehat{g}_T(\boldsymbol{R})$, and $\widehat{g}_T(\boldsymbol{R}) \xrightarrow{\mathbb{P}} g_0(\boldsymbol{R})$ under the conditions of the following theorem, where $\xrightarrow{\mathbb{P}}$ stands for convergence in probability.

Theorem 2.1 (Consistency of NPHC)**.** *Suppose that* $(\boldsymbol{N_t})$ *is observed on* $\mathbb{R}^+$ *and assume that*

1. $g_0(\boldsymbol{R}) = 0$ *if and only if* $\boldsymbol{R} = \boldsymbol{R}_0$;
2. $\boldsymbol{R} \in \Theta$*, which is a compact set*;
3. *the spectral radius of the kernel norm matrix satisfies* $\|\boldsymbol{G}_0\| < 1$;

4. $H_T \to \infty$ *and* $H_T^2/T \to 0$.

Then

$$\widehat{\boldsymbol{G}}_T = \boldsymbol{I_d} - \left(\arg\min_{\boldsymbol{R}\in\Theta} \widetilde{\mathcal{L}}_T(\boldsymbol{R}) \right)^{-1} \overset{\mathbb{P}}{\to} \boldsymbol{G}_0.$$

Remark 1. Assumption 3 is mandatory for stability of the Hawkes process, and Assumptions 3 and 4 are sufficient to prove that the estimators of the integrated cumulants defined in Equations 11, 12 and 13 are asymptotically consistent. Assumption 2 is a very mild standard technical assumption, note Θ is compact so that the minima of the considered functionals of $\boldsymbol{R}$ are reached within Θ. Assumption 1 is a standard asymptotic moment condition, that allows to identity parameters from the integrated cumulants.

The proof of the Theorem is given in the supplementary material.

3. Numerical Experiments

Simulated datasets. We simulated several datasets with Ogata's Thinning algorithm (Ogata, 1981) using the open-source library `tick`[1], each corresponding to a shape of kernel.

$$\text{exponential kernel: } \phi(t) = \alpha\beta\exp(-\beta t) \quad (14)$$

$$\text{power law kernel: } \phi(t) = \alpha\beta\gamma(1+\beta t)^{-(1+\gamma)} \quad (15)$$

$$\text{rectangular kernel: } \phi(t) = \alpha\beta\mathbb{1}_{[0,1/\beta]}(t-\gamma) \quad (16)$$

The integral of each kernel on its support equals α, $1/\beta$ can be regarded as a characteristic time-scale and γ is the scaling exponent for the power law distribution and a delay parameter for the rectangular one. We consider a non-symmetric block-matrix $\boldsymbol{G}$ to show that our method can effectively uncover causality between the nodes, see Figure 1. The parameter α take the same constant value on the three blocks, but we set three very different β_0, β_1 and β_2 from one block to the other, with ratio $\beta_{i+1}/\beta_i = 10$ and $\beta_0 = 0.1$. The matrix $\boldsymbol{G}$ has constant entries on the blocks - $g^{ij} = 1/6$ for dimension 10 and $g^{ij} = 1/10$ for dimension 100 -, and zero outside, and the number of events is roughly equal to 10^5 on average over the nodes. We ran the algorithm on three simulated datasets: a 10-dimensional process with rectangular kernels named Rect10, a 10-dimensional process with power law kernels named PLaw10 and a 100-dimensional process with exponential kernels named Exp100.

MemeTracker dataset. We use events of the most active sites from the MemeTracker dataset[2]. This dataset contains the publication times of articles in many websites/blogs from August 2008 to April 2009, and hyperlinks between posts. We extract the top 100 media sites with the largest number of documents, with about 7 million of events. We use the links to trace the flow of information and establish an estimated ground truth for the matrix $\boldsymbol{G}$. Indeed, when an hyperlink j appears in a post in website i, the link j can be regarded as a direct ancestor of the event. Then, Eq. (2) shows g^{ij} can be estimated by $N_T^{i\leftarrow j}/N_T^j = \#\{\text{links } j \to i\}/N_T^j$.

Order book dynamics. We apply our method to financial data, in order to understand the self and cross-influencing dynamics of all event types in an order book. An order book is a list of buy and sell orders for a specific financial instrument, the list being updated in real-time throughout the day. This model has first been introduced in (Bacry et al., 2016), and models the order book via the following 8-dimensional point process: $N_t = (P_t^{(a)}, P_t^{(b)}, T_t^{(a)}, T_t^{(b)}, L_t^{(a)}, L_t^{(b)}, C_t^{(a)}, C_t^{(b)})$, where $P^{(a)}$ (resp. $P^{(b)}$) counts the number of upward (resp. downward) price moves, $T^{(a)}$ (resp. $T^{(b)}$) counts the number of market orders at the ask[3] (resp. at the bid) that do not move the price, $L^{(a)}$ (resp. $L^{(b)}$) counts the number of limit orders at the ask[4] (resp. at the bid) that do not move the price, and $C^{(a)}$ (resp. $C^{(b)}$) counts the number of cancel orders at the ask[5] (resp. at the bid) that do not move the price. The financial data has been provided by QuantHouse EUROPE/ASIA, and consists of DAX future contracts between 01/01/2014 and 03/01/2014.

Baselines. We compare NPHC to state-of-the art baselines: the ODE-based algorithm (ODE) by (Zhou et al., 2013a), the Granger Causality-based algorithm (GC) by (Xu et al., 2016), the ADM4 algorithm (ADM4) by (Zhou et al., 2013b), and the Wiener-Hopf-based algorithm (WH) by (Bacry & Muzy, 2016).

Metrics. We evaluate the performance of the proposed methods using the computing time, the Relative Error

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

and the Mean Kendall Rank Correlation

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

where $\text{RankCorr}(x,y) = \frac{2}{d(d-1)}(N_{\text{concordant}}(x,y) - N_{\text{discordant}}(x,y))$ with $N_{\text{concordant}}(x,y)$ the number of pairs

[1] `https://github.com/X-DataInitiative/tick`

[2] `https://www.memetracker.org/data.html`

[3] i.e. buy orders that are executed and removed from the list

[4] i.e. buy orders added to the list

[5] i.e. the number of times a limit order at the ask is cancelled: in our dataset, almost 95% of limit orders are cancelled before execution.

(i, j) satisfying $x_i > x_j$ and $y_i > y_j$ or $x_i < x_j$ and $y_i < y_j$ and $N_{\text{discordant}}(x, y)$ the number of pairs (i, j) for which the same condition is not satisfied. Note that RankCorr score is a value between -1 and 1, representing rank matching, but can take smaller values (in absolute value) if the entries of the vectors are not distinct.

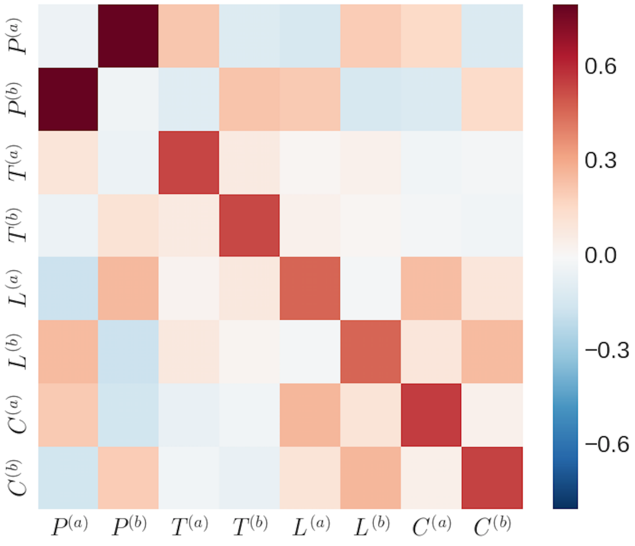

Figure 1. Estimated $\widehat{\boldsymbol{G}}$ via NPHC on DAX order book data.

Table 2. Metrics on Rect10: comparable rank correlation, strong improvement for relative error and computing time.

Method	ODE	GC	ADM4	WH	NPHC
RelErr	0.007	0.15	0.10	0.005	**0.001**
MRankCorr	0.33	0.02	0.21	**0.34**	**0.34**
Time (s)	846	768	709	933	**20**

Table 3. Metrics on PLaw10: comparable rank correlation, strong improvement for relative error and computing time.

Method	ODE	GC	ADM4	WH	NPHC
RelErr	0.011	0.09	0.053	0.009	**0.0048**
MRankCorr	0.31	0.26	0.24	**0.34**	0.33
Time (s)	870	781	717	946	**18**

Discussion. We perform the ADM4 estimation, with exponential kernel, by giving the exact value $\beta = \beta_0$ of one block. Let us stress that this helps a lot this baseline, in comparison to NPHC where nothing is specified on the shape of the kernel functions. We used $M = 10$ basis functions for both ODE and GC algorithms, and $L = 50$ quadrature points for WH. We did not run WH on the 100-dimensional datasets, for computing time reasons, because its complexity scales with d^4. We ran multiprocessed versions of the baseline methods on 56 cores, to decrease the computing time.

Table 4. Metrics on Exp100: comparable rank correlation, strong improvement for relative error and computing time.

Method	ODE	GC	ADM4	NPHC
RelErr	0.092	0.112	0.079	**0.008**
MRankCorr	0.032	0.009	**0.049**	0.041
Time (s)	3215	2950	2411	**47**

Table 5. Metrics on MemeTracker: strong improvement in relative error, rank correlation and computing time.

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

Our method consistently performs better than all baselines, on the three synthetic datasets, on MemeTracker and on the financial dataset, both in terms of Kendall rank correlation and estimation error. Moreover, we observe that our algorithm is roughly 50 times faster than all the considered baselines.

On Rect10, PLaw10 and Exp100 our method gives very impressive results, despite the fact that it does not uses any prior shape on the kernel functions, while for instance the ADM4 baseline do. On these simulated datasets, NPHC obtains a comparable or slightly better Kendall rank correlation, but improves a lot the relative error.

On MemeTracker, the baseline methods obtain a high relative error of from 9% to 19% while our method achieves a relative error of 7% which is a strong improvement. Moreover, NPHC reaches a much better Kendall rank correlation, which proves that it leads to a much better recovery of the relative order of estimated influences than all the baselines. Indeed, it has been shown in (Zhou et al., 2013a) that kernels of MemeTracker data are not exponential, nor power law. This partly explains why our approach behaves better.

On the financial data, the estimated kernel norm matrix obtained via NPHC, see Figure 3, gave some interpretable results (see also (Bacry et al., 2016)):

1. Any 2×2 sub-matrix with same kind of inputs (i.e. Prices changes, Trades, Limits or Cancels) is symmetric. This shows empirically that ask and bid have symmetric roles.

2. The prices are mostly cross-excited, which means that a price increase is very likely to be followed by a price decrease, and conversely. This is consistent with the wavy prices we observe on financial markets.

3. The market, limit and cancel orders are strongly self-excited. This can be explained by the persistence of order flows, and by the splitting of meta-orders into sequences of smaller orders. Moreover, we observe that orders impact the price without changing it. For example, the increase of cancel orders at the bid causes downward price moves.

4. Conclusion

In this paper, we introduce a simple nonparametric method (the NPHC algorithm) that leads to a fast and robust estimation of the matrix $\boldsymbol{G}$ of the kernel integrals of a Multivariate Hawkes process that encodes Granger causality between nodes. This method relies on the matching of the integrated order 2 and order 3 empirical cumulants, which represent the simplest set of global observables containing sufficient information to recover the matrix $\boldsymbol{G}$. Since this matrix fully accounts for the self- and cross- influences of the process nodes (that can represent agents or users in applications), our approach can naturally be used to quantify the degree of endogeneity of a system and to uncover the causality structure of a network.

By performing numerical experiments involving very different kernel shapes, we show that the baselines, involving either parametric or non-parametric approaches are very sensible to model misspecification, do not lead to accurate estimation, and are numerically expensive, while NPHC provides fast, robust and reliable results. This is confirmed on the MemeTracker database, where we show that NPHC outperforms classical approaches based on EM algorithms or the Wiener-Hopf equations. Finally, the NPHC algorithm provided very satisfying results on financial data, that are consistent with well-known stylized facts in finance.

5. Technical details

5.1. Proof of Equation (8)

We denote $\boldsymbol{\nu}(z)$ the matrix

$$\nu^{ij}(z) = \mathcal{L}_z\Big(t \to \frac{\mathbb{E}(dN_u^i dN_{u+t}^j)}{du dt} - \Lambda^i\Lambda^j\Big),$$

where $\mathcal{L}_z(f)$ is the Laplace transform of f, and $\psi_t = \sum_{n\geq 1}\phi_t^{(\star n)}$, where $\phi_t^{(\star n)}$ refers to the n^{th} auto-convolution of ϕ_t. Then we use the characterization of second-order statistics, first formulated in (Hawkes, 1971) and fully generalized in (Bacry & Muzy, 2016),

$$\boldsymbol{\nu}(z) = (\boldsymbol{I_d} + \mathcal{L}_{-z}(\boldsymbol{\Psi}))\boldsymbol{L}(\boldsymbol{I_d} + \mathcal{L}_z(\boldsymbol{\Psi}))^\top,$$

where $\boldsymbol{L}^{ij} = \Lambda^i\delta^{ij}$ with δ^{ij} the Kronecker symbol. Since $\boldsymbol{I_d} + \mathcal{L}_z(\boldsymbol{\Psi}) = (\boldsymbol{I_d} - \mathcal{L}_z(\boldsymbol{\Phi}))^{-1}$, taking $z = 0$ in the previous equation gives

$$\begin{aligned}\boldsymbol{\nu}(0) &= (\boldsymbol{I_d} - \boldsymbol{G})^{-1}\boldsymbol{L}(\boldsymbol{I_d} - \boldsymbol{G}^\top)^{-1},\\ \boldsymbol{C} &= \boldsymbol{R}\boldsymbol{L}\boldsymbol{R}^\top,\end{aligned}$$

which gives us the result since the entry (i, j) of the last equation gives $C^{ij} = \sum_m \Lambda^m R_{im}R_{jm}$.

5.2. Proof of Equation (9)

We start from (Jovanović et al., 2015), cf. Eqs. (48) to (51), and group some terms:

$$\begin{aligned}K^{ijk} &= \sum_m \Lambda^m R_{im}R_{jm}R_{km}\\ &+ \sum_m R_{im}R_{jm}\sum_n \Lambda^n R_{kn}\mathcal{L}_0(\psi^{mn})\\ &+ \sum_m R_{im}R_{km}\sum_n \Lambda^n R_{jn}\mathcal{L}_0(\psi^{mn})\\ &+ \sum_m R_{jm}R_{km}\sum_n \Lambda^n R_{in}\mathcal{L}_0(\psi^{mn}).\end{aligned}$$

Using the relations $\mathcal{L}_0(\psi^{mn}) = R_{mn} - \delta^{mn}$ and $C^{ij} = \sum_m \Lambda^m R_{im}R_{jm}$, proves Equation (9).

5.3. Integrated cumulants estimators

For $H > 0$ let us denote $\Delta_H N_t^i = N_{t+H}^i - N_{t-H}^i$. Let us first remark that, if one restricts the integration domain to $(-H, H)$ in Eqs. (4) and (5), one gets by permuting integrals and expectations:

$$\begin{aligned}\Lambda^i dt &= \mathbb{E}(dN_t^i)\\ C^{ij}dt &= \mathbb{E}\left(dN_t^i(\Delta_H N_t^j - 2H\Lambda^j)\right)\\ K^{ijk}dt &= \mathbb{E}\left(dN_t^i(\Delta_H N_t^j - 2H\Lambda^j)(\Delta_H N_t^k - 2H\Lambda^k)\right)\\ &\quad - dt\Lambda^i\mathbb{E}\left((\Delta_H N_t^j - 2H\Lambda^j)(\Delta_H N_t^k - 2H\Lambda^k)\right).\end{aligned}$$

The estimators (11) and (12) are then naturally obtained by replacing the expectations by their empirical counterparts, notably

$$\frac{\mathbb{E}(dN_t^i f(t))}{dt} \to \frac{1}{T}\sum_{\tau\in Z^i} f(\tau).$$

For the estimator (13), we shall also notice that

$$\begin{aligned}&\mathbb{E}((\Delta_H N_t^j - 2H\Lambda^j)(\Delta_H N_t^k - 2H\Lambda^k))\\ &\quad = \int\int \mathbb{1}_{[-H,H]}(t)\mathbb{1}_{[-H,H]}(t')C_{t-t'}^{jk}dtdt'\\ &\quad = \int (2H - |t|)^+ C_t^{jk}dt.\end{aligned}$$

We estimate the last integral with the remark above.

Acknowledgements

This work benefited from the support of the École Polytechnique fund raising - Data Science Initiative. The authors want to thank Marcello Rambaldi for fruitful discussions on order book data's experiments.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

Aït-Sahalia, Y., Cacho-Diaz, J., and Laeven, R. JA. Modeling financial contagion using mutually exciting jump processes. Technical report, National Bureau of Economic Research, 2010.

Bacry, E. and Muzy, J. F. Hawkes model for price and trades high-frequency dynamics. *Quantitative Finance*, 14(7):1147–1166, 2014.

Bacry, E. and Muzy, J.-F. First- and second-order statistics characterization of hawkes processes and non-parametric estimation. *IEEE Transactions on Information Theory*, 62(4):2184–2202, 2016.

Bacry, E., Mastromatteo, I., and Muzy, J.-F. Hawkes processes in finance. *Market Microstructure and Liquidity*, 1 (01):1550005, 2015.

Bacry, E., Jaisson, T., and Muzy, J.-F. Estimation of slowly decreasing hawkes kernels: application to high-frequency order book dynamics. *Quantitative Finance*, pp. 1–23, 2016.

Choromanska, A., Henaff, M., Mathieu, M., Ben Arous, G., and LeCun, Y. The loss surfaces of multilayer networks. In *AISTATS*, 2015.

Crane, R. and 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.

Da Fonseca, J. and Zaatour, R. Hawkes process: Fast calibration, application to trade clustering, and diffusive limit. *Journal of Futures Markets*, 34(6):548–579, 2014.

Daley, D. J. and Vere-Jones, D. *An Introduction to the Theory of Point Processes Volume I: Elementary Theory and Methods*. Springer Science & Business Media, 2003.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159, 2011.

Eichler, M., Dahlhaus, R., and Dueck, J. Graphical modeling for multivariate hawkes processes with nonparametric link functions. *Journal of Time Series Analysis*, pp. n/a–n/a, 2016. ISSN 1467-9892.

Farajtabar, M., Wang, Y., Rodriguez, M., Li, S., Zha, H., and Song, L. Coevolve: A joint point process model for information diffusion and network co-evolution. In *Advances in Neural Information Processing Systems*, pp. 1945–1953, 2015.

Gomez-Rodriguez, M., Leskovec, J., and Schölkopf, B. Modeling information propagation with survival theory. *Proceedings of the International Conference on Machine Learning*, 2013.

Granger, C. W. J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969. ISSN 00129682, 14680262.

Hall, A. R. *Generalized Method of Moments*. Oxford university press, 2005.

Hansen, N. R., Reynaud-Bouret, P., and Rivoirard, V. Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, 21(1):83–143, 2015.

Hardiman, S. J. and Bouchaud, J.-P. Branching-ratio approximation for the self-exciting Hawkes process. *Phys. Rev. E*, 90(6):062807, December 2014. doi: 10.1103/PhysRevE.90.062807.

Hawkes, A. G. Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 33(3):438–443, 1971. ISSN 00359246.

Jovanović, S., Hertz, J., and Rotter, S. Cumulants of Hawkes point processes. *Phys. Rev. E*, 91(4):042802, April 2015. doi: 10.1103/PhysRevE.91.042802.

Lemonnier, R. and Vayatis, N. Nonparametric markovian learning of triggering kernels for mutually exciting and mutually inhibiting multivariate hawkes processes. In *Machine Learning and Knowledge Discovery in Databases*, pp. 161–176. Springer, 2014.

Lewis, E. and Mohler, G. A nonparametric em algorithm for multiscale hawkes processes. *Journal of Nonparametric Statistics*, 2011.

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

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

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

Podosinnikova, A., Bach, F., and Lacoste-Julien, S. Rethinking lda: moment matching for discrete ica. In *Advances in Neural Information Processing Systems*, pp. 514–522, 2015.

Reynaud-Bouret, P. and Schbath, S. Adaptive estimation for hawkes processes; application to genome analysis. *The Annals of Statistics*, 38(5):2781–2822, 2010.

Subrahmanian, V.S., Azaria, A., Durst, S., Kagan, V., Galstyan, A., Lerman, K., Zhu, L., Ferrara, E., Flammini, A., and Menczer, F. The darpa twitter bot challenge. *Computer*, 49(6):38–46, 2016.

Xu, H., Farajtabar, M., and Zha, H. Learning granger causality for hawkes processes. In *Proceedings of The 33rd International Conference on Machine Learning*, pp. 1717–1726, 2016.

Yang, S.-H. and Zha, H. Mixture of mutually exciting processes for viral diffusion. In *Proceedings of the International Conference on Machine Learning*, 2013.

Zhou, K., Zha, H., and Song, L. Learning triggering kernels for multi-dimensional hawkes processes. In *Proceedings of the International Conference on Machine Learning*, pp. 1301–1309, 2013a.

Zhou, K., Zha, H., and Song, L. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. *AISTATS*, 2013b.