Uncovering Causality from Multivariate Hawkes Integrated Cumulants

Massil Achab, Emmanuel Bacry, Stéphane Gaïffas, Iacopo Mastromatteo, Jean-François Muzy

Journal of Machine Learning Research 18 (2018) 1-28 Submitted 5/17; Revised 9/17; Published 4/18

Uncovering Causality from Multivariate Hawkes Integrated Cumulants

Massil Achab MASSIL.ACHAB@M4X.ORG
Centre de Mathématiques Appliquées, Ecole polytechnique, Palaiseau, France

Emmanuel Bacry BACRY@CEREMADE.DAUPHINE.FR
Centre de Recherche en Mathématique de la Décision, Université Paris-Dauphine, Paris, France
Centre de Mathématiques Appliquées, Ecole polytechnique, Palaiseau, France

Stéphane Gaïffas GAIFFAS@MATH.UNIV-PARIS-DIDEROT.FR
Laboratoire de Probabilités et Modèles Aléatoires, Université Paris-Diderot, Paris, France

Iacopo Mastromatteo IACOPO.MASTROMATTEO@CFM.FR
Research-Execution, Capital Fund Management, Paris, France

Jean-François Muzy MUZY@UNIV-CORSE.FR
Laboratoire Sciences Pour l'Environnement, Université de Corse, Corte, France

Editor: Edoardo M. Airoldi

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 with respect to the shape of the kernels and gives appealing results on the MemeTracker database and on financial order book data.

Keywords: Hawkes Process, Causality Inference, Cumulants, Generalized Method of Moments

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, see Subrahmanian et al. (2016), the high-frequency variations of signals in finance, see Bacry et al. (2015), the earthquakes and aftershocks in geophysics,

see Ogata (1998), the crime activity, see Mohler et al. (2011) or the position of genes in genomics, see Reynaud-Bouret and Schbath (2010). The succession of the precise timestamps carries a great deal of information about the dynamics of the underlying systems. 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 (i.e., the different components of the counting process), by leveraging on their timestamp patterns, see, for instance, Bacry and Muzy (2016); Lemonnier and Vayatis (2014); Lewis and Mohler (2011); Zhou et al. (2013a); 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 each $\tau \in Z^i$ 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 and Vere-Jones (2003), and patterns in the events time-series can be captured by structuring these intensities.

The Hawkes process introduced in 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, for instance, in social networks, see for instance Crane and 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 *i.e.* with support in $\mathbb{R}_+$ (so that effects don't happen before their cause) 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.

Most of the literature uses a parametric approach for estimating the kernels. Without a doubt, the most popular parametrization form is the exponential kernel $\phi^{ij}(t) = \alpha_{ij}\beta_{ij}e^{-\beta_{ij}t}$ because it definitely simplifies the inference algorithm (e.g., the complexity needed for computing the likelihood is much smaller). When d is large, in order to reduce the number of parameters, some authors choose to arbitrarily share the kernel shapes across the different nodes. Thus, for instance, in Yang and Zha (2013); Zhou et al. (2013b); Farajtabar et al. (2015), they choose $\phi^{ij}(t) = \alpha_{ij}h(t)$ with $\alpha_{ij} \in \mathbb{R}^+$ quantifies 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 $h(t) = \beta t^{-(\beta+1)}$). Both approaches are, most of the time, non-realistic. On the one hand 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). On the other

hand, assuming an exponential shape or a power law shape for a kernel 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 maximal impact.

In order to have more flexibility on the shape of the kernels, nonparametric estimation can be considered. Expectation-Maximization algorithms can be found in Lewis and Mohler (2011) (for $d = 1$) or in Zhou et al. (2013a) ($d > 1$). An alternative method is proposed in Bacry and Muzy (2016) where the nonparametric estimation is formulated as a numerical solving of a Wiener-Hopf equation. Another nonparametric strategy considers a decomposition of kernels on a dictionary of function $h_1, \dots, h_K$, 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 and 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 $k = 1, \dots, K$.

Such methods are computationally-intensive 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 for causality purpose.

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

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

As it can be seen from the cluster representation of Hawkes processes (Hawkes and Oakes (1974)), this integral represents the mean total number of events of type i directly triggered by an event of type j, and then encodes a notion of *causality*. Actually, as detailed below (see Section 2.1), such integral can be related to the Granger causality (Granger (1969)).

The main idea of the method we developed in this paper is to estimate the matrix $\boldsymbol{G}$ directly using a matching cumulants (or moments) method. Apart from the mean, we shall use second and third-order cumulants which correspond respectively to centered second and third-order moments. We first compute an estimation $\widehat{\boldsymbol{M}}$ of these centered moments $M(\boldsymbol{G})$ (they are uniquely defined by $\boldsymbol{G}$). Then, we look for a matrix $\widehat{\boldsymbol{G}}$ that minimizes the L^2 error $\|M(\widehat{\boldsymbol{G}}) - \widehat{\boldsymbol{M}}\|^2$. Thus the integral matrix $\widehat{\boldsymbol{G}}$ is directly estimated without making hardly any assumptions on the shape the involved kernels. As it will be shown, this approach turns out to be particularly robust to the kernel shapes, which is not the case of all previous Hawkes-based approaches that aim causality recovery. We call this method NPHC (Non Parametric Hawkes Cumulant), since our approach is of nonparametric nature. We provide a theoretical analysis that proves the consistency of the NPHC estimator. Our proof is based on ideas from the theory of Generalized Method of Moments (GMM) but requires an original technical trick since our setting strongly departs from the standard parametric statistics with i.i.d observations. Note that moment and cumulant matching techniques proved particularly powerful for latent topic models, in particular Latent Dirichlet Allocation, see Podosinnikova et al. (2015). A small set of previous works, namely Da Fonseca and 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.

The paper is organized as follows: in Section 2, we provide the background on the integrated kernels and the integrated cumulants of the Hawkes process. We then introduce the method, investigate its complexity and explain the consistency result we prove. In Section 3, we estimate the matrix of Hawkes kernels' integrals for various simulated datasets and for real datasets, namely the MemeTracker database and financial order book data. We then provide in Appendix B the technical details skipped in the previous parts and the proof of our consistency result. Section 4 contains concluding remarks.

2. NPHC: The Non Parametric Hawkes Cumulant method

In this Section, we provide the background on integrals of Hawkes kernels and integrals of Hawkes cumulants. We then explain how the NPHC method enables estimating $\boldsymbol{G}$.

2.1 Branching structure and Granger causality

From the definition of Hawkes process as a Poisson cluster process, see Jovanović et al. (2015) or Hawkes and Oakes (1974), 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

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

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 $\boldsymbol{I_d} - \boldsymbol{G}$.

Since the question of a *real causality* is too complex in general, see Imbens and Rubin (2015); Pearl (2009), 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.

Definition 1 (Granger causality for time series) *Given two time series X and Y, we denote $\mathcal{H}(t)$ the set of all information available prior to t, $\mathcal{H}_{-X}(t)$ the previous set in which information coming from X is excluded, and A an arbitrary non-empty set. We say that X Granger-causes Y if*

$$\mathbb{P}[Y(t+1) \in A | \mathcal{H}(t)] \neq \mathbb{P}[Y(t+1) \in A | \mathcal{H}_{-X}(t)].$$

Existing works mainly focus on learning Granger causality for time series, see Arnold et al. (2007); Eichler (2012); Basu et al. (2015), such as vector autoregressive models (VAR), where Granger causality is formulated as a statistical test of the VAR coefficients. In Eichler et al. (2016), the authors extend the definition of Granger (non-)causality to the case of Hawkes processes.

Definition 2 (Granger causality for Hawkes processes) *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 above at Eq. (2).

2.2 Integrated cumulants of the Hawkes process

A general formula for the integral of the cumulants of a multivariate Hawkes process is provided in Jovanović et al. (2015). As explained below, for the purpose of our method, we only need to consider cumulants up to the third order. Given $1 \leq i, j, k \leq d$, 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 martingale representation from Bacry and Muzy (2016) or the Poisson cluster process representation from 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 Appendix B) lead to the following identities:

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

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

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

Equations (8) and (9) are proved in Appendix B. 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 and 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 $M = \{C^{ij}\}_{1 \leq i \leq j \leq d}$ 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 $\boldsymbol{R} \to \boldsymbol{O}\boldsymbol{R}$, 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 $M = \{K^{iij}\}_{1 \leq i,j \leq d}$. Note that the choice of K^{iij} is arbitrary, our method and the theory developed below is unchanged for any choice of d^2 distinct components among the d^3 third-order cumulant. 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, $\boldsymbol{K^c} = \{K^{iij}\}_{1 \leq i,j \leq d}$ 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 us to determine the optimal weighting matrix involved in the loss function. However, this approach is unusable in practice, since the associated complexity is too high. Indeed, since we have d^2 parameters, this matrix has d^4 coefficients and GMM calls for computing its inverse leading to a $O(d^6)$ complexity. In this work, we use the coefficient κ to scale the two terms, as

$$\kappa = \frac{\|\widehat{\boldsymbol{K^c}}\|_2^2}{\|\widehat{\boldsymbol{K^c}}\|_2^2 + \|\widehat{\boldsymbol{C}}\|_2^2},$$

see Section B.4 for an explanation about the link between κ and the weighting matrix. 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 $O(d^3)$ that is cheap compared to the computation of the cumulants when $n = \max_i |Z^i| \gg d$, which is the typical scaling satisfied in applications. 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, see for instance Zhou et al. (2013a); Xu et al. (2016); Zhou et al. (2013b); Bacry and Muzy (2016), where the number of components d in experiments is always around 100 or smaller. Note that, actually, our approach leads to a *much faster* algorithm than the considered state-of-the-art baselines, see Tables 1–4 from Section 3 below.

2.3 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 that is a good approximation if the support of the kernel $\phi^{ij}(t)$ is smaller than H and the spectral norm $\|\boldsymbol{G}\|$ satisfies $\|\boldsymbol{G}\| < 1$.
In this case, given a realization of a stationary Hawkes process $\{\boldsymbol{N}_t : t \in [0, T]\}$, as shown in Appendix B, we can write the estimators of the first three cumulants (3), (4) and (5) as

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

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

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

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 too fast compared to T, then these estimated cumulants are consistent estimators of the theoretical cumulants (see Section 2.6).

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.

2.4 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 6. However, the expectations of cumulants $\boldsymbol{\Lambda}$ and $\boldsymbol{C}$ defined in Eq. (4) and (5) that appear in the definition of $\mathcal{L}(\boldsymbol{R})$ are unknown and should be replaced with $\widehat{\boldsymbol{\Lambda}}$ and $\widehat{\boldsymbol{C}}$. We denote $\widetilde{\mathcal{L}}(\boldsymbol{R})$ the 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):

$$\widetilde{\mathcal{L}}(\boldsymbol{R}) = (1-\kappa)\|\boldsymbol{R}^{\odot 2}\widehat{\boldsymbol{C}}^\top + 2[\boldsymbol{R} \odot (\widehat{\boldsymbol{C}} - \boldsymbol{R}\widehat{\boldsymbol{L}})]\boldsymbol{R}^\top - \widehat{\boldsymbol{K^c}}\|_2^2 + \kappa\|\boldsymbol{R}\widehat{\boldsymbol{L}}\boldsymbol{R}^\top - \widehat{\boldsymbol{C}}\|_2^2 \tag{14}$$

As explained in Choromanska et al. (2015), 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 use AdaGrad from Duchi et al. (2011), a variant of the Stochastic Gradient Descent with adaptive learning rates. AdaGrad scales the learning rates coordinate-wise using the online variance of the previous gradients, in order to incorporate second-order information during training. As detailed in Section 2.5, the optimization step is negligible compared to the computation of the cumulants whenever $n = \max_i |Z^i| \gg d$, which is the typical scaling in applications. The NPHC method is summarized schematically in Algorithm 1.

Algorithm 1 Non Parametric Hawkes Cumulant method

Input: $\boldsymbol{N}_t$
Output: $\widehat{\boldsymbol{G}}$

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

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 Equation (8) write $\boldsymbol{C}^{1/2}\boldsymbol{O}\boldsymbol{L}^{-1/2}$, with $\boldsymbol{L} = \mathrm{diag}(\boldsymbol{\Lambda})$ and $\boldsymbol{O}$ an orthogonal matrix. Our starting point is then simply chosen by setting $\boldsymbol{O} = \boldsymbol{I_d}$ in the previous formula, leading to nice convergence results. 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), which leads to $\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{R}}^{-1}\widehat{\boldsymbol{\Lambda}}$. An efficient implementation of this algorithm with TensorFlow, see Abadi et al. (2016), is available on GitHub: `https://github.com/achab/nphc`.

The optimization problem can be regularized by adding to the function $\mathcal{L}(\boldsymbol{R})$ a regularizing term of the form $\lambda N(\boldsymbol{G})$ that encodes a prior assumption on the structure of $\boldsymbol{G}$. As long as $\boldsymbol{R}$ matches Equation (8) the penalty term can be written as a function of $\boldsymbol{R}$ since $\lambda N(\boldsymbol{G}) = \lambda N(\boldsymbol{I_d} - \boldsymbol{C}^{-1}\boldsymbol{L}\boldsymbol{R}^\top)$. Since the algorithms we compare our method to optimize different objective functions (negative log-likelihood, least-squares, etc.), adding $\lambda N(\boldsymbol{G})$ with the same λ to these functions would trigger different behaviors. Then, in the rest of the document we focus on the unregularized problem, for which we prove the convergence's consistency in the Section 2.6.

2.5 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 and 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. 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. If n denotes the maximum number of events per component (i.e. $n = \max_{1 \leq i \leq d} |Z^i|$) then the complexity of one iteration of the algorithm is

$O(Mn^3d^2 + ML(nd + n^2))$. 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 $\widehat{\boldsymbol{G}}$. In the end, the complexity of the NPHC method is $O(nd^2 + N_{\text{iter}}d^3)$. According to this analysis, summarized in Table 1 below, one can see that in the regime $n \gg d$, the NPHC method outperforms all the other ones.

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

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

2.6 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 condition*: $\mathbb{E}[g(X, \theta)] = 0$ 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 $x_1, \dots, x_n$ 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.

In the following, we use the subscript T to refer to quantities that only depend on the process $(\boldsymbol{N_t})$ in the interval $[0, T]$ (e.g., 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

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

Using these notations, $\widetilde{\mathcal{L}}_T(\boldsymbol{R})$ can be seen as the weighted squared Frobenius norm of $\widehat{g}_T(\boldsymbol{R})$. Moreover, when $T\to+\infty$, one has $\widehat{g}_T(\boldsymbol{R}) \xrightarrow{\mathbb{P}} g_0(\boldsymbol{R})$ under the conditions of the following theorem, where $\xrightarrow{\mathbb{P}}$ stands for convergence in probability.

Theorem 3 (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. *$\boldsymbol{R}\in\Theta$, where Θ 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

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

The proof of the Theorem is given in the subsection B.5 below. 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 allowing to prove consistency for estimators based on moments. Assumption 1 is a standard asymptotic moment condition, that allows to identity parameters from the integrated cumulants.

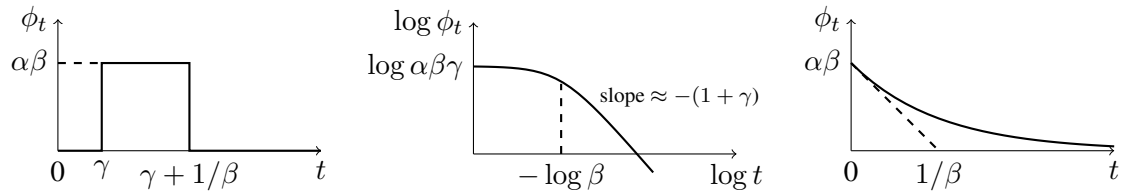
3. Numerical Experiments

In this Section, we provide a comparison of NPHC with the state-of-the art, on simulated datasets with different kernel shapes, the MemeTracker dataset (social networks) and the order book dynamics dataset (finance).

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: rectangular, exponential or power law kernel, see Figure 1 below.

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 3. The matrix $\boldsymbol{G}$ has constant entries α on the three blocks - $\alpha = g^{ij} = 1/6$ for dimension 10 and $\alpha = g^{ij} = 1/10$ for dimension 100 -,

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



(a) Rectangular kernel $\phi_t = \alpha\beta \mathbb{1}_{[0,1/\beta]}(t-\gamma)$

(b) Power law kernel on log-log scale $\phi_t = \alpha\beta\gamma(1+\beta t)^{-(1+\gamma)}$

(c) Exponential kernel $\phi_t = \alpha\beta \exp(-\beta t)$

Figure 1: The three different kernels used to simulate the datasets.

and zero outside. The two other parameters' values are the same for dimensions 10 and 100. The parameter γ is set to $1/2$ on the three blocks as well, 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 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 and the top 200 media sites with the largest number of documents, with about 7 million of events. We name MemeTracker100 the 100-dimensional dataset, and MemeTracker200 the 200-dimensional one. 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.

2. `https://www.memetracker.org/data.html`
3. That is buy orders that are executed and removed from the list.
4. That is buy orders added to the list.
5. That is the number of times a limit order at the ask is canceled: in our dataset, almost 95% of limit orders are canceled before execution.

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 and 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 (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.

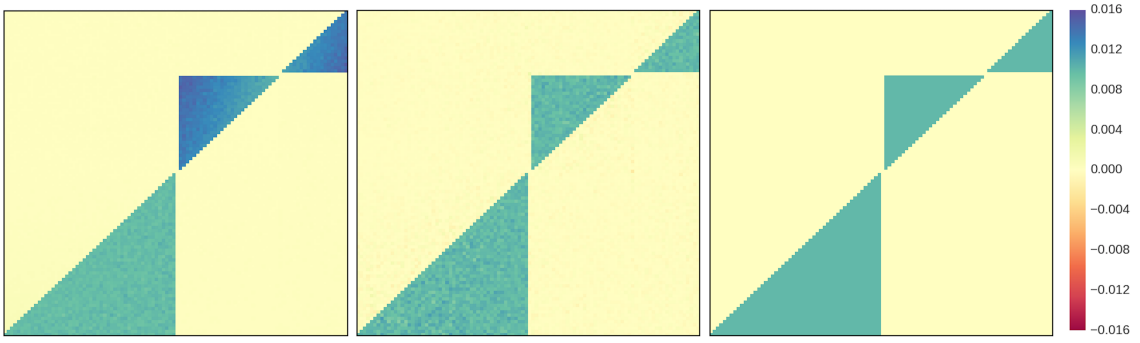


Figure 2: On Exp100 dataset, estimated $\widehat{\boldsymbol{G}}$ with ADM4 (left), with NPHC (middle) and the ground-truth matrix $\boldsymbol{G}$ (right). Both ADM4 and NPHC estimates recover the three blocks. However, ADM4 overestimates the integrals on two of the three blocks, while NPHC gives the same value on each blocks.

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 multi-processed versions of the baseline methods on 56 cores, to decrease the computing time.

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

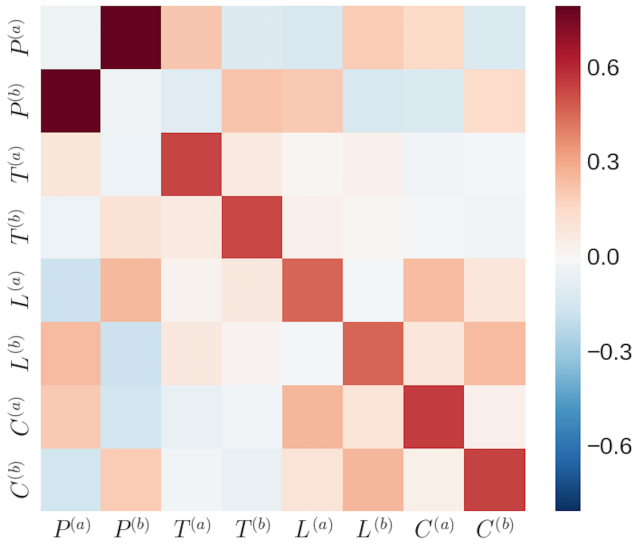


Figure 3: Estimated $\widehat{\boldsymbol{G}}$ via NPHC on DAX order book data.

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 Figure 3, we observe that the matrix $\widehat{\boldsymbol{G}}$ estimated with ADM4 recovers well the block for which $\beta = \beta_0$, i.e. the value we gave to the method, but does not perform well on the two other blocks, while the matrix $\widehat{\boldsymbol{G}}$ estimated with NPHC approximately reaches the true value for each of the three blocks. On these simulated datasets, NPHC obtains a comparable or slightly better Kendall rank correlation, but improves a lot the relative error.

On MemeTracker100, the baseline methods obtain a high relative error between 9% and 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. On MemeTracker200, NPHC outperforms again the baselines, with smaller Kendall rank correlation. The comparison of the computation times for both experiments shows that NPHC scales better than other methods. Plus, 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.

Acknowledgments

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

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**

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**

Appendix A. Additional experiments

We propose below additional experiments and technical details on the theoretical study of the NPHC procedure.

A.1 Convergence curves versus dimension

The NPHC procedure is divided into two parts: the first is the computation of the integrated cumulants' estimators, the second is the minimization of the loss function. As highlighted in the Section 2.5, the bottleneck of the algorithm is the computation of the cumulants' estimators. However, we can still wonder how fast the optimization algorithm converges with respect to the dimensionality of the point process. To answer that question, we simulated ten datasets corresponding to dimensions $d \in \{10, 20, \dots, 100\}$ with $\mu_i = 0.01$ for $i \in [d]$, $g^{ij} = 0.9/d$ for $(i, j) \in [d]^2$ (such that $||\boldsymbol{G}|| = 0.9 < 1$), and $T = 10^7$. We then ran the NPHC method and recorded the loss function's evolution over the iterations, rescaled between 0 and 1. One can see on Figure 4 how the dimension influences the convergence curve: using the same hyperparameters for AdaGrad Duchi et al. (2011), the higher the dimension the more oscillating the convergence curve. Plus, the loss function seems to be flatter in lower dimension ($d = 10$ for instance) since AdaGrad needs more iterations to reach a minimum compared to higher-dimensional cases ($d = 70$ or $d = 100$ for instance).

A.2 Relative error versus number of events

The estimation of the integrated cumulants becomes more accurate when the amount of training data increases. The consistency of the estimators given in Equations (7), (8) and (9) is indeed proved in the theorem's proof in Appendix B. A natural question that arises is then to evaluate the precision of the parameter's estimation when the number of points increases, all other things remaining equal. To quantify this effect, we simulated several datasets similar to Rect10 - described in Section 3 - with

Table 5: Metrics on MemeTracker100: 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**

Table 6: Metrics on MemeTracker200: strong improvement in relative error, rank correlation and computing time.

Method	ODE	GC	ADM4	NPHC
RelErr	0.173	0.212	0.109	**0.084**
MRankCorr	0.062	0.048	0.077	**0.085**
Time (s)	11786	11210	8903	**164**

different simulation's durations. The only difference between those datasets is then the number of points per node. We ran NPHC ten times per dataset to loosely evaluate the variance of the estimate.

We only focus on the relative error metric which gives more interpretable results. The results summarized on the Figure 5 show the decrease of the relative error as the average (over the ten dimensions) number of points per node becomes larger. This decrease comes along with a variance decrease of the estimates.

A.3 Random choice of d^2 third integrated cumulant's entries

The NPHC method arbitrarily computes the d^2 *iij* entries of the third integrated cumulant, among the d^3 entries available, and then minimizes the distance between theoretical and empirical cumulants. We numerically show in this subsection than the computation of d^2 random entries followed by the minimization of the loss function reachs the same performance than NPHC's method. We sampled three sets of random d^2 indices ijk, and ran the methods on the dataset Rect10 introduced in Section 3.

Table 7: Metrics on Rect10: similar relative errors and rank correlations for the different methods.

Set of indices	Random set 1	Random set 2	Random set 3	*iij* (NPHC)
RelErr	0.0013	0.0014	0.0013	0.0013
MRankCorr	0.34	0.34	0.33	0.34

The results summarized on the Table 7 show that, to our knowledge, the procedure is not very sensitive to the selection of the d^2 entries from the third integrated cumulant tensor.

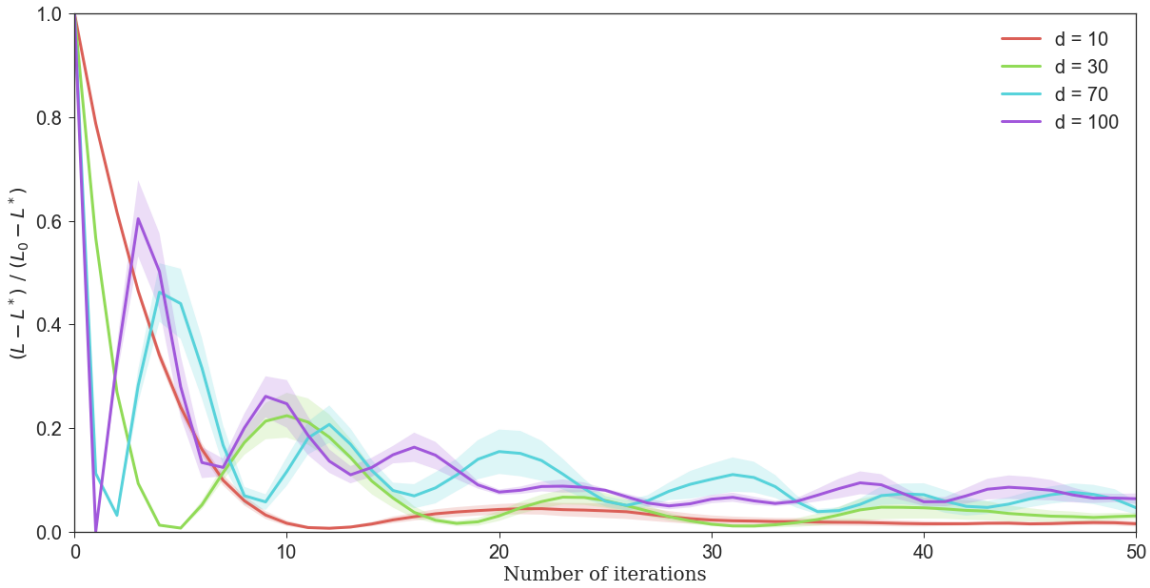


Figure 4: Convergence curves versus dimension.

Appendix B. Technical details

We show in this section how to obtain the equations stated above, the estimators of the integrated cumulants and the scaling coefficient κ that appears in the objective function. We then prove the theorem of the paper.

B.1 Proof of Equation (8)

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

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

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 and 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 $C^{ij} = \sum_m \Lambda^m R_{im} R_{jm}$.

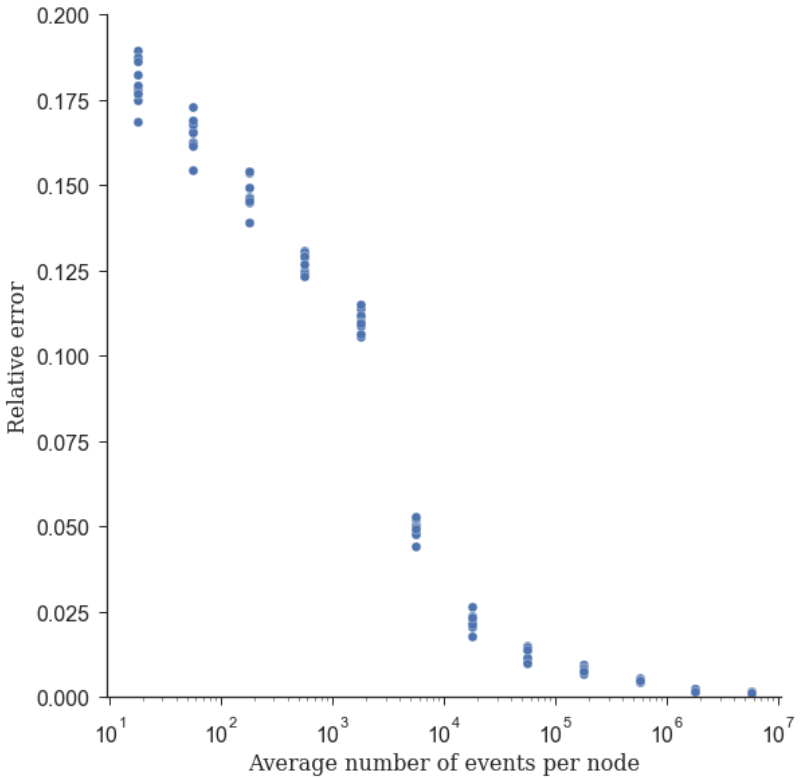


Figure 5: Relative error versus number of events.

B.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).

B.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

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

We estimate the last integral with the remark above.

B.4 Choice of the scaling coefficient κ

Following the theory of GMM, we denote $m(X, \theta)$ a function of the data, where X is distributed with respect to a distribution $\mathbb{P}_{\theta_0}$, which satisfies the *moment conditions* $g(\theta) = \mathbb{E}[m(X, \theta)] = 0$ if and only if $\theta = \theta_0$, the parameter θ_0 being the *ground truth*. For $x_1, \dots, x_N$ observed copies of X, we denote $\widehat{g}_i(\theta) = m(x_i, \theta)$, the usual choice of weighting matrix is $\widehat{W}_N(\theta) = \frac{1}{N}\sum_{i=1}^N \widehat{g}_i(\theta)\widehat{g}_i(\theta)^\top$, and the objective to minimize is then

$$\left(\frac{1}{N}\sum_{i=1}^N \widehat{g}_i(\theta)\right)\left(\widehat{W}_N(\theta_1)\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^N \widehat{g}_i(\theta)\right), \tag{15}$$

where θ_1 is a constant vector. Instead of computing the inverse weighting matrix, we rather use its projection on $\{\alpha \boldsymbol{I_d} : \alpha \in \mathbb{R}\}$. It can be shown that the projection choses α as the mean eigenvalue of $\widehat{W}_N(\theta_1)$. We can easily compute the sum of its eigenvalues:

$$\mathrm{Tr}(\widehat{W}_N(\theta_1)) = \frac{1}{N}\sum_{i=1}^N \mathrm{Tr}(\widehat{g}_i(\theta_1)\widehat{g}_i(\theta_1)^\top) = \frac{1}{N}\sum_{i=1}^N \mathrm{Tr}(\widehat{g}_i(\theta_1)^\top \widehat{g}_i(\theta_1)) = \frac{1}{N}\sum_{i=1}^N ||\widehat{g}_i(\theta_1)||_2^2.$$

In our case, $\widehat{g}(\boldsymbol{R}) = \left[\mathbf{vec}[\widehat{\boldsymbol{K^c}} - \boldsymbol{K^c}(\boldsymbol{R})], \mathbf{vec}[\widehat{\boldsymbol{C}} - \boldsymbol{C}(\boldsymbol{R})]\right]^\top \in \mathbb{R}^{2d^2}$. Considering a block-wise weighting matrix, one block for $\widehat{\boldsymbol{K^c}} - \boldsymbol{K^c}(\boldsymbol{R})$ and the other for $\widehat{\boldsymbol{C}} - \boldsymbol{C}(\boldsymbol{R})$, the sum of the eigenvalues

of the first block becomes $\|\widehat{\boldsymbol{K^c}} - \boldsymbol{K^c}(\boldsymbol{R})\|_2^2$, and $\|\widehat{\boldsymbol{C}} - \boldsymbol{C}(\boldsymbol{R})\|_2^2$ for the second. We compute the previous terms with $\boldsymbol{R}_1 = 0$. All together, the objective function to minimize is

$$\frac{1}{\|\widehat{\boldsymbol{K^c}}\|_2^2}\|\boldsymbol{K^c}(\boldsymbol{R}) - \widehat{\boldsymbol{K^c}}\|_2^2 + \frac{1}{\|\widehat{\boldsymbol{C}}\|_2^2}\|\boldsymbol{C}(\boldsymbol{R}) - \widehat{\boldsymbol{C}}\|_2^2. \tag{16}$$

Dividing this function by $\left(1/\|\widehat{\boldsymbol{K^c}}\|_2^2 + 1/\|\widehat{\boldsymbol{C}}\|_2^2\right)^{-1}$, and setting $\kappa = \|\widehat{\boldsymbol{K^c}}\|_2^2/(\|\widehat{\boldsymbol{K^c}}\|_2^2 + \|\widehat{\boldsymbol{C}}\|_2^2)$, we obtaind the loss function given in Equation (10).

B.5 Proof of the Theorem

The main difference with the usual Generalized Method of Moments, see Hansen (1982), relies in the relaxation of the moment conditions, since we have $\mathbb{E}[\widehat{g}_T(\theta_0)] = m_T \neq 0$. We adapt the proof of consistency given in Newey and McFadden (1994).

We can relate the integral of the Hawkes process's kernels to the integrals of the cumulant densities, from Jovanović et al. (2015). Our cumulant matching method would fall into the usual GMM framework if we could estimate - without bias - the integral of the covariance on $\mathbb{R}$, and the integral of the skewness on $\mathbb{R}^2$. Unfortunately, we can't do that easily. We can however estimate without bias $\int f_t^T C_t^{ij} dt$ and $\int f_t^T K_t^{ijk} dt$ with f^T a compact supported function on $[-H_T, H_T]$ that weakly converges to 1, with $H_T \longrightarrow \infty$. In most cases we will take $f_t^T = \mathbb{1}_{[-H_T,H_T]}(t)$. Denoting $\widehat{C}^{ij,(T)}$ the estimator of $\int f_t^T C_t^{ij} dt$, the term $|\mathbb{E}[\widehat{C}^{ij,(T)}] - C^{ij}| = |\int f_t^T C_t^{ij} dt - C^{ij}|$ can be considered a proxy to the *distance to the classical GMM*. This distance has to go to zero to make the rest of GMM's proof work: the estimator $\widehat{C}^{ij,(T)}$ is then asymptotically unbiased towards C^{ij} when T goes to infinity.

B.5.1 NOTATIONS

We observe the multivariate point process $(\boldsymbol{N}_t)$ on $\mathbb{R}^+$, with Z^i the events of the i^{th} component. We will often write covariance / skewness instead of integrated covariance / skewness. In the rest of the document, we use the following notations.

Hawkes kernels' integrals $\quad \boldsymbol{G}^{\text{true}} = \int \boldsymbol{\Phi}_t dt = (\int \phi_t^{ij} dt)_{ij} = \boldsymbol{I_d} - (\boldsymbol{R}^{\text{true}})^{-1}$

Theoretical mean matrix $\quad \boldsymbol{L} = \text{diag}(\Lambda^1, \dots, \Lambda^d)$

Theoretical covariance $\quad \boldsymbol{C} = \boldsymbol{R}^{\text{true}} \boldsymbol{L} (\boldsymbol{R}^{\text{true}})^\top$

Theoretical skewness $\quad \boldsymbol{K^c} = (K^{iij})_{ij} = (\boldsymbol{R}^{\text{true}})^{\odot 2} \boldsymbol{C}^\top + 2[\boldsymbol{R}^{\text{true}} \odot (\boldsymbol{C} - \boldsymbol{R}^{\text{true}} \boldsymbol{L})](\boldsymbol{R}^{\text{true}})^\top$

Filtering function $\quad f^T \geq 0 \qquad \text{supp}(f^T) \subset [-H_T, H_T] \qquad F^T = \int f_s^T ds \qquad \widetilde{f}_t^T = f_{-t}^T$

Events sets $\quad Z^{i,T,1} = Z^i \cap [H_T, T + H_T] \qquad Z^{j,T,2} = Z^j \cap [0, T + 2H_T]$

Estimators of the mean $\quad \widehat{\Lambda}^i = \frac{N^i_{T+H_T} - N^i_{H_T}}{T} \qquad \widetilde{\Lambda}^j = \frac{N^j_{T+2H_T}}{T+2H_T}$

Estimator of the covariance $\widehat{C}^{ij,(T)} = \frac{1}{T}\sum_{\tau\in Z^{i,T,1}}\left(\sum_{\tau'\in Z^{j,T,2}} f_{\tau'-\tau} - \widetilde{\Lambda}^j F^T\right)$

Estimator of the skewness[6]

$$\widehat{K}^{ijk,(T)} = \frac{1}{T}\sum_{\tau\in Z^{i,T,1}}\left(\sum_{\tau'\in Z^{j,T,2}} f_{\tau'-\tau} - \widetilde{\Lambda}^j F^T\right)\left(\sum_{\tau''\in Z^{k,T,2}} f_{\tau'-\tau} - \widetilde{\Lambda}^k F^T\right)$$
$$- \frac{\widehat{\Lambda}^i}{T+2H_T}\sum_{\tau'\in Z^{j,T,2}}\left(\sum_{\tau''\in Z^{k,T,2}} (f^T\star\widetilde{f}^T)_{\tau'-\tau''} - \widetilde{\Lambda}^k (F^T)^2\right)$$

GMM RELATED NOTATIONS

$$\theta = \boldsymbol{R} \quad \text{and} \quad \theta_0 = \boldsymbol{R}^{\text{true}}$$
$$g_0(\theta) = \textbf{vec}\begin{bmatrix}\boldsymbol{C} - \boldsymbol{R}\boldsymbol{L}\boldsymbol{R}^\top \\ \boldsymbol{K^c} - \boldsymbol{R}^{\odot^2}\boldsymbol{C}^\top - 2[\boldsymbol{R}\odot(\boldsymbol{C}-\boldsymbol{R}\boldsymbol{L})]\boldsymbol{R}^\top\end{bmatrix} \in \mathbb{R}^{2d^2}$$
$$\widehat{g}_T(\theta) = \textbf{vec}\begin{bmatrix}\widehat{\boldsymbol{C}}^{(T)} - \boldsymbol{R}\widehat{\boldsymbol{L}}\boldsymbol{R}^\top \\ \widehat{\boldsymbol{K^c}}^{(T)} - \boldsymbol{R}^{\odot^2}(\widehat{\boldsymbol{C}}^{(T)})^\top - 2[\boldsymbol{R}\odot(\widehat{\boldsymbol{C}}^{(T)} - \boldsymbol{R}\widehat{\boldsymbol{L}})]\boldsymbol{R}^\top\end{bmatrix} \in \mathbb{R}^{2d^2}$$
$$Q_0(\theta) = g_0(\theta)^\top W g_0(\theta)$$
$$\widehat{Q}_T(\theta) = \widehat{g}_T(\theta)^\top \widehat{W}_T \widehat{g}_T(\theta)$$

B.5.2 CONSISTENCY

First, let's remind a useful theorem for consistency in GMM from Newey and McFadden (1994).

Theorem 4 *If there is a function $Q_0(\theta)$ such that (i) $Q_0(\theta)$ is uniquely maximized at θ_0; (ii) Θ is compact; (iii) $Q_0(\theta)$ is continuous; (iv) $\widehat{Q}_T(\theta)$ converges uniformly in probability to $Q_0(\theta)$, then $\widehat{\theta}_T = \arg\max \widehat{Q}_T(\theta) \xrightarrow{\mathbb{P}} \theta_0$.*

We can now prove the consistency of our estimator.

Theorem 5 *Suppose that $(\boldsymbol{N_t})$ is observed on $\mathbb{R}^+$, $\widehat{W}_T \xrightarrow{\mathbb{P}} W$, and*

1. *W is positive semi-definite and $Wg_0(\theta) = 0$ if and only if $\theta = \theta_0$,*
2. *$\theta \in \Theta$, which is compact,*
3. *the spectral radius of the kernel norm matrix satisfies $||\mathbf{\Phi}||_* < 1$,*
4. *$\forall i,j,k \in [d]$, $\int f_u^T C_u^{ij} du \to \int C_u^{ij} du$ and $\int f_u^T f_v^T K_{u,v}^{ijk} du dv \to \int K_{u,v}^{ijk} du dv$,*
5. *$(F^T)^2/T \xrightarrow{\mathbb{P}} 0$ and $||f||_\infty = O(1)$.*

6. When $f_t^T = \mathbb{1}_{[-H_T,H_T]}(t)$, we remind that $(f^T\star\widetilde{f}^T)_t = (2H_T - |t|)^+$. This leads to the estimator we showed in the article.

Then

$$\widehat{\theta}_T \xrightarrow{\mathbb{P}} \theta_0.$$

Remark 6 *In practice, we use a constant sequence of weighting matrices:* $\widehat{W}_T = \boldsymbol{I_d}$.

Proof Proceed by verifying the hypotheses of Theorem 2.1 from Newey and McFadden (1994). Condition 2.1(i) follows by (i) and by $Q_0(\theta) = [W^{1/2} g_0(\theta)]^\top [W^{1/2} g_0(\theta)] > 0 = Q_0(\theta_0)$. Indeed, there exists a neighborhood N of θ_0 such that $\theta \in N \backslash \{\theta_0\}$ and $g_0(\theta) \neq 0$ since $g_0(\theta)$ is a polynom. Condition 2.1(ii) follows by (ii). Condition 2.1(iii) is satisfied since $Q_0(\theta)$ is a polynom. Condition 2.1(iv) is harder to prove. First, since $\widehat{g}_T(\theta)$ is a polynom of θ, we prove easily that $\mathbb{E}[\sup_{\theta\in\Theta} |\widehat{g}_T(\theta)|] < \infty$. Then, by Θ compact, $g_0(\theta)$ is bounded on Θ, and by the triangle and Cauchy-Schwarz inequalities,

$$\begin{aligned}
&\left|\widehat{Q}_T(\theta) - Q_0(\theta)\right| \\
&\quad \leq \left|(\widehat{g}_T(\theta) - g_0(\theta))^\top \widehat{W}_T (\widehat{g}_T(\theta) - g_0(\theta))\right| \\
&\quad + \left|g_0(\theta)^\top (\widehat{W}_T + \widehat{W}_T^\top)(\widehat{g}_T(\theta) - g_0(\theta))\right| + \left|g_0(\theta)^\top (\widehat{W}_T - W) g_0(\theta)\right| \\
&\quad \leq \|\widehat{g}_T(\theta) - g_0(\theta)\|^2 \|\widehat{W}_T\| + 2\|g_0(\theta)\| \|\widehat{g}_T(\theta) - g_0(\theta)\| \|\widehat{W}_T\| + \|g_0(\theta)\|^2 \|\widehat{W}_T - W\|.
\end{aligned}$$

To prove $\sup_{\theta\in\Theta} \left|\widehat{Q}_T(\theta) - Q_0(\theta)\right| \xrightarrow{\mathbb{P}} 0$, we should now prove that $\sup_{\theta\in\Theta} \|\widehat{g}_T(\theta) - g_0(\theta)\| \xrightarrow{\mathbb{P}} 0$. By Θ compact, it is sufficient to prove that $\|\widehat{\boldsymbol{L}} - \boldsymbol{L}\| \xrightarrow{\mathbb{P}} 0$, $\|\widehat{\boldsymbol{C}}^{(T)} - \boldsymbol{C}\| \xrightarrow{\mathbb{P}} 0$, and $\|\widehat{\boldsymbol{K^c}}^{(T)} - \boldsymbol{K^c}\| \xrightarrow{\mathbb{P}} 0$.

PROOF THAT $\|\widehat{\boldsymbol{L}} - \boldsymbol{L}\| \xrightarrow{\mathbb{P}} 0$

The estimator of $\boldsymbol{L}$ is unbiased so let's focus on the variance of $\widehat{\boldsymbol{L}}$.

$$\begin{aligned}
\mathbb{E}[(\widehat{\Lambda}^i - \Lambda^i)^2] &= \mathbb{E}\left[\left(\frac{1}{T}\int_{H_T}^{T+H_T} (dN_t^i - \Lambda^i dt)\right)^2\right] \\
&= \frac{1}{T^2}\int_{H_T}^{T+H_T}\int_{H_T}^{T+H_T} \mathbb{E}[(dN_t^i - \Lambda^i dt)(dN_{t'}^i - \Lambda^i dt')] \\
&= \frac{1}{T^2}\int_{H_T}^{T+H_T}\int_{H_T}^{T+H_T} C^{ii}_{t'-t} dt dt' \\
&\leq \frac{1}{T^2}\int_{H_T}^{T+H_T} C^{ii} dt = \frac{C^{ii}}{T} \longrightarrow 0
\end{aligned}$$

By Markov inequality, we have just proved that $\|\widehat{\boldsymbol{L}} - \boldsymbol{L}\| \xrightarrow{\mathbb{P}} 0$.

PROOF THAT $\|\widehat{\boldsymbol{C}}^{(T)} - \boldsymbol{C}\| \xrightarrow{\mathbb{P}} 0$

First, let's remind that $\mathbb{E}(\widehat{\boldsymbol{C}}^{(T)}) \neq \boldsymbol{C}$. Indeed,

$$\begin{aligned}
\mathbb{E}\left(\widehat{C}^{ij,(T)}\right) &= \mathbb{E}\left(\frac{1}{T}\int_{H_T}^{T+H_T} dN_t^i \int_0^{T+2H_T} dN_{t'}^j f_{t'-t} - \widehat{\Lambda}^i\widetilde{\Lambda}^j F^T\right) \\
&= \mathbb{E}\left(\frac{1}{T}\int_{H_T}^{T+H_T} dN_t^i \int_{-t}^{T+2H_T-t} dN_{t+s}^j f_s - \Lambda^i\Lambda^j F^T\right) + \epsilon^{ij,T,H_T}F^T \\
&= \frac{1}{T}\int_{H_T}^{T+H_T}\int_{-H_T}^{H_T} f_s \mathbb{E}\left(dN_t^i dN_{t+s}^j - \Lambda^i\Lambda^j ds\right) + \epsilon^{ij,T,H_T}F^T \\
&= \int f_s C_s^{ij} ds + \epsilon^{ij,T,H_T}F^T
\end{aligned}$$

Now,

$$\begin{aligned}
\epsilon^{ij,T,H_T} &= \mathbb{E}\left(\Lambda^i\Lambda^j - \widehat{\Lambda}^i\widetilde{\Lambda}^j\right) \\
&= -\frac{1}{T^2}\int_{H_T}^{T+H_T}\int_0^{T+2H_T} \mathbb{E}\left(dN_t^i dN_{t'}^j - \Lambda^i\Lambda^j dt dt'\right) \\
&= -\frac{1}{T^2}\int_{H_T}^{T+H_T}\int_0^{T+2H_T} C_{t-t'}^{ij} dt dt' \\
&= -\frac{1}{T}\int \left(1 + \left(\frac{H_T - |t|}{T}\right)^-\right)^+ C_t^{ij} dt
\end{aligned}$$

Since f satisfies $F^T = o(T)$, we have $\mathbb{E}(\widehat{\boldsymbol{C}}^{(T)}) \longrightarrow \boldsymbol{C}$. It remains now to prove that $\|\widehat{\boldsymbol{C}}^{(T)} - \mathbb{E}(\widehat{\boldsymbol{C}}^{(T)})\| \xrightarrow{\mathbb{P}} 0$.
Let's now focus on the variance of $\widehat{C}^{ij,(T)}$: $\mathbb{V}(\widehat{C}^{ij,(T)}) = \mathbb{E}\left((\widehat{C}^{ij,(T)})^2\right) - \mathbb{E}(\widehat{C}^{ij,(T)})^2$.
Now,

$$\begin{aligned}
&\mathbb{E}\left((\widehat{C}^{ij,(T)})^2\right) \\
&= \mathbb{E}\left(\frac{1}{T^2}\sum_{(\tau,\eta,\tau',\eta')\in(Z^{i,T,1})^2\times(Z^{j,T,2})^2} (f_{\tau'-\tau} - F^T/(T+2H_T))(f_{\eta'-\eta} - F^T/(T+2H_T))\right) \\
&= \mathbb{E}\left(\frac{1}{T^2}\int_{t,s\in[H_T,T+H_T]}\int_{t',s'} dN_t^i dN_{t'}^j dN_s^i dN_{s'}^j (f_{t'-t} - F^T/(T+2H_T))(f_{s'-s} - F^T/(T+2H_T))\right) \\
&= \frac{1}{T^2}\int_{t,s\in[H_T,T+H_T]}\int_{t',s'\in[0,T+2H_T]} \mathbb{E}\left(dN_t^i dN_{t'}^j dN_s^i dN_{s'}^j\right) \\
&\qquad\qquad \cdot (f_{t'-t} - F^T/(T+2H_T))(f_{s'-s} - F^T/(T+2H_T))
\end{aligned}$$

And,

$$\begin{aligned}&\mathbb{E}(\widehat{C}^{ij,(T)})^2\\&= \frac{1}{T^2}\int_{t,s\in[H_T,T+H_T]}\int_{t',s'\in[0,T+2H_T]} \mathbb{E}\left(dN_t^i dN_{t'}^j\right)\mathbb{E}\left(dN_s^i dN_{s'}^j\right)\\&\qquad\qquad\cdot (f_{t'-t}-F^T/(T+2H_T))(f_{s'-s}-F^T/(T+2H_T))\end{aligned}$$

Then, the variance involves the integration towards the difference of moments $\mu^{r,s,t,u}-\mu^{r,s}\mu^{t,u}$.
Let's write it as a sum of cumulants, since cumulants density are integrable.

$$\begin{aligned}\mu^{r,s,t,u}-\mu^{r,s}\mu^{t,u} &= \kappa^{r,s,t,u}+\kappa^{r,s,t}\kappa^u[4]+\kappa^{r,s}\kappa^{t,u}[3]+\kappa^{r,s}\kappa^t\kappa^u[6]+\kappa^r\kappa^s\kappa^t\kappa^u-(\kappa^{r,s}+\kappa^r\kappa^s)(\kappa^{t,u}+\kappa^t\kappa^u)\\&=\kappa^{r,s,t,u}\\&+\kappa^{r,s,t}\kappa^u+\kappa^{u,r,s}\kappa^t+\kappa^{t,u,r}\kappa^s+\kappa^{s,t,u}\kappa^r\\&+\kappa^{r,t}\kappa^{s,u}+\kappa^{r,u}\kappa^{s,t}\\&+\kappa^{r,t}\kappa^s\kappa^u+\kappa^{r,u}\kappa^s\kappa^t+\kappa^{s,t}\kappa^r\kappa^u+\kappa^{s,t}\kappa^r\kappa^u\end{aligned}$$

In the rest of the proof, we denote $a_t = \mathbb{1}_{t\in[H_T,T+H_T]}$, $b_t = \mathbb{1}_{t\in[0,T+2H_T]}$, $c_t = \mathbb{1}_{t\in[-H_T,H_T]}$,
$g_t = f_t - \frac{1}{T+2H_T}F^T$
Before starting the integration of each term, let's remark that:

1. $\mathbf{\Psi}_t = \sum_{n\geq 1}\mathbf{\Phi}_t^{(\star n)} \geq 0$ since $\mathbf{\Phi}_t \geq 0$.

2. The regular parts of C_u^{ij}, $K_{u,v}^{ijk}$ (skewness density) and $M_{u,v,w}^{ijkl}$ (fourth cumulant density) are positive as polynoms of integrals of $\psi_\cdot^{ab}$ with positive coefficients. The integrals of the singular parts are positive as well.

3. (a) $\int a_t b_{t'} f_{t'-t}dtdt' = TF^T$
   (b) $\int a_t b_{t'} g_{t'-t}dtdt' = 0$
   (c) $\int a_t b_{t'} |g_{t'-t}|dtdt' \leq 2TF^T$

4. $\forall t\in\mathbb{R}, a_t(b\star\widetilde{g})_t = 0$, where $\widetilde{g}_s = g_{-s}$.

Fourth cumulant We want here to compute $\int \kappa_{t,t',s,s'}^{i,j,i,j} a_t b_{t'} a_s b_{s'} g_{t'-t} g_{s'-s} dtdt'dsds'$.
We remark that $|g_{t'-t}g_{s'-s}| \leq (||f||_\infty(1+2H_T/T))^2 \leq 4||f||_\infty^2$.

$$\begin{aligned}\left|\frac{1}{T^2}\int \kappa_{t,t',s,s'}^{i,j,i,j} a_t b_{t'} a_s b_{s'} g_{t'-t} g_{s'-s} dtdt'dsds'\right| &\leq \left(\frac{2||f||_\infty}{T}\right)^2 \int dt a_t \int dt' b_{t'} \int ds a_s \int ds' b_{s'} M_{t'-t,s-t,s'-t}^{ijij}\\&\leq \left(\frac{2||f||_\infty}{T}\right)^2 \int dt a_t \int dt' b_{t'} \int ds a_s \int dw M_{t'-t,s-t,w}^{ijij}\\&\leq \left(\frac{2||f||_\infty}{T}\right)^2 \int dt a_t \int M_{u,v,w}^{ijij} dudvdw\\&\leq \frac{4||f||_\infty^2}{T} M^{ijij} \underset{T\to\infty}{\longrightarrow} 0\end{aligned}$$

Third $\times$ First We have four terms, but only two different forms since the roles of (s, s') and (t, t') are symmetric.
First form

$$\begin{aligned}\int \kappa^{i,j,i}_{t,t',s}\Lambda^j G_{\boldsymbol{t}} d\boldsymbol{t} &= \frac{\Lambda^j}{T^2}\int \kappa^{i,j,i}_{t,t',s} a_t b_{t'} a_s b_{s'} g_{t'-t} g_{s'-s} dt dt' ds ds' \\ &= \frac{\Lambda^j}{T^2}\int \kappa^{i,j,i}_{t,t',s} a_t b_{t'} a_s (b \star \widetilde{g})_s g_{t'-t} dt dt' ds \\ &= 0 \qquad \text{since } a_s(b\star\widetilde{g})_s = 0\end{aligned}$$

Second form

$$\begin{aligned}\Big|\int \kappa^{i,j,j}_{t,t',s'}\Lambda^i G_{\boldsymbol{t}} d\boldsymbol{t}\Big| &= \Big|\frac{\Lambda^i}{T^2}\int \kappa^{i,j,j}_{t,t',s'} a_t b_{t'} a_s b_{s'} g_{t'-t} g_{s'-s} dt dt' ds ds'\Big| \\ &= \Big|\frac{\Lambda^i}{T^2}\int \kappa^{i,j,j}_{t,t',s'} a_t b_{t'} g_{t'-t} b_{s'} (a\star g)_{s'} dt dt' ds'\Big| \\ &\leq \frac{\Lambda^i}{T^2} 2||f||_\infty \int ds' b_{s'}(a\star|g|)_{s'}\int dt a_t \int dt' b_{t'} K^{ijj}_{t'-s',t-s'} \\ &\leq 4||f||_\infty K^{ijj}\Lambda^i \frac{F^T}{T} \underset{T\to\infty}{\longrightarrow} 0\end{aligned}$$

Second $\times$ Second
First form

$$\begin{aligned}\Big|\int \kappa^{i,i}_{t,s}\kappa^{j,j}_{t',s'} G_{\boldsymbol{t}} d\boldsymbol{t}\Big| &\leq \frac{2||f||_\infty}{T^2}\int C^{ii}_{t-s} C^{jj}_{t'-s'} a_t b_{t'} |g_{t'-t}| a_s b_{s'} dt dt' ds ds' \\ &\leq \frac{2||f||_\infty}{T^2} C^{ii} C^{jj}\int a_t b_{t'} |g_{t'-t}| dt dt' \\ &\leq 4||f||_\infty C^{ii} C^{jj}\frac{F^T}{T} \underset{T\to\infty}{\longrightarrow} 0\end{aligned}$$

Second form

$$\Big|\int \kappa^{i,j}_{t,s'}\kappa^{i,j}_{t',s} G_{\boldsymbol{t}} d\boldsymbol{t}\Big| \leq 4||f||_\infty (C^{ij})^2 \frac{F^T}{T} \underset{T\to\infty}{\longrightarrow} 0$$

Second $\times$ First $\times$ First
First form

$$\int \kappa^{i,j}_{t,t'}\Lambda^i\Lambda^j G_{\boldsymbol{t}} d\boldsymbol{t} = \frac{\Lambda^i\Lambda^j}{T^2}\int \kappa^{i,j}_{t,t'} a_t b_{t'} g_{t'-t} dt dt' \int a_s b_{s'} g_{s'-s} ds ds' = 0$$

Second form

$$\int \kappa^{i,i}_{t,s}\Lambda^j\Lambda^j G_{\boldsymbol{t}} d\boldsymbol{t} = \left(\frac{\Lambda^j}{T}\right)^2 \int \kappa^{i,i}_{t,s} a_t b_{t'} g_{t'-t} a_s (b\star\widetilde{g})_s dt dt' ds = 0$$

We have just proved that $\mathbb{V}(\widehat{\boldsymbol{C}}^{(T)}) \xrightarrow{\mathbb{P}} 0$. By Markov inequality, it ensures us that $\|\widehat{\boldsymbol{C}}^{(T)} - \mathbb{E}(\widehat{\boldsymbol{C}}^{(T)})\| \xrightarrow{\mathbb{P}} 0$, and finally that $\|\widehat{\boldsymbol{C}}^{(T)} - \boldsymbol{C}\| \xrightarrow{\mathbb{P}} 0$. ■

PROOF THAT $\|\widehat{\boldsymbol{K}^{\boldsymbol{c}}}^{(T)} - \boldsymbol{K}^{\boldsymbol{c}}\| \xrightarrow{\mathbb{P}} 0$

The scheme of the proof is similar to the previous one. The upper bounds of the integrals involve the same kind of terms, plus the new term $(F^T)^2/T$ that goes to zero thanks to the assumption 5 of the theorem.

References

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

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

A. Arnold, Y. Liu, and N. Abe. Temporal causal modeling with graphical granger methods. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 66–75. ACM, 2007.

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

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

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

S. Basu, A. Shojaie, and G. Michailidis. Network granger causality with inherent grouping structure. *The Journal of Machine Learning Research*, 16(1):417–453, 2015.

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

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

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

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

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

M. Eichler. Graphical modelling of multivariate time series. *Probability Theory and Related Fields*, 153(1-2):233–268, 2012.

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

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

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

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

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

L. P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054, 1982.

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

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

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

A. G. Hawkes and D. Oakes. A cluster process representation of a self-exciting process. *Journal of Applied Probability*, pages 493–503, 1974.

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

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

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

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

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

W. K Newey and D. McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245, 1994.

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

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

J. Pearl. *Causality*. Cambridge university press, 2009.

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

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

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

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

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

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

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