Title

Investigating Clustering and Violence Interruption in Gang-Related Violent Crime Data Using Spatial–Temporal Point Processes With Covariates

Journal

Journal of the American Statistical Association, 116(536)

ISSN

0162-1459

Authors

Park, Junhyung
Schoenberg, Frederic Paik
Bertozzi, Andrea L
et al.

Publication Date

2021-10-02

DOI

10.1080/01621459.2021.1898408

Peer reviewed

Annals of the Institute of Statistical Mathematics
https://doi.org/10.1007/s10463-020-00780-1

Fast estimation of multivariate spatiotemporal Hawkes processes and network reconstruction

Baichuan Yuan[1] · **Frederic P. Schoenberg**[2] · **Andrea L. Bertozzi**[1]

Received: 30 March 2020 / Revised: 26 October 2020 / Accepted: 6 November 2020

Abstract
We present a fast, accurate estimation method for multivariate Hawkes self-exciting point processes widely used in seismology, criminology, finance and other areas. There are two major ingredients. The first is an analytic derivation of exact maximum likelihood estimates of the nonparametric triggering density. We develop this for the multivariate case and add regularization to improve stability and robustness. The second is a moment-based method for the background rate and triggering matrix estimation, which is extended here for the spatiotemporal case. Our method combines them together in an efficient way, and we prove the consistency of this new approach. Extensive numerical experiments, with synthetic data and real-world social network data, show that our method improves the accuracy, scalability and computational efficiency of prevailing estimation approaches. Moreover, it greatly boosts the performance of Hawkes process-based models on social network reconstruction and helps to understand the spatiotemporal triggering dynamics over social media.

Keywords Nonparametric estimation · L_2 regularization · Point processes · Social network · Cumulants

Fast estimation of Hawkes processes.

✉ Frederic P. Schoenberg
frederic@stat.ucla.edu

[1] Department of Mathematics, University of California, Los Angeles, 7619D Math-Science Building, Los Angeles, CA 90095-1555, USA

[2] Department of Statistics, University of California, 8142, Math-Science Building, Los Angeles, CA 90095-1554, USA

Published online: 01 January 2021

1 Introduction

The spatiotemporal Hawkes (ST-Hawkes) process has been widely used to model and forecast clustered point process data in the study of earthquakes (Ogata 1998), crimes (Mohler et al. 2011), invasive species (Balderama et al. 2012), terrorist attacks (Porter et al. 2012), infectious disease (Schoenberg et al. 2018a) and finance (Bacry et al. 2015). These models are often characterized by a *triggering density* describing how the occurrence of one event may spark future events nearby. Recently, multivariate Hawkes processes, which can incorporate accompanying information on each event such as event types or magnitude of an earthquake, have been the subject of significant research in the areas of societally harmful events (Mohler 2014; Chiang et al. 2019; Brantingham et al. 2020a), finance (Bacry et al. 2015), neuroscience (Chen et al. 2017) and text analysis (Du et al. 2015; Zhu and Xie 2019). Applications include network reconstruction (Linderman and Adams 2014; Fox et al. 2016; Hall and Willett 2016; Yuan et al. 2019; Mark et al. 2018), causal inference (Achab et al. 2017; Eichler et al. 2017; Brantingham et al. 2020b) and social media cascade modeling (Lai et al. 2016; Farajtabar et al. 2015).

Much of this recent research has been fueled by advances in the nonparametric estimation of Hawkes processes, and in particular by the landmark work of Marsan and Lengliné (2008), who detailed a method for estimating the triggering in a ST-Hawkes process by assuming the triggering density to be a step function and then estimating the step heights via maximum likelihood estimation (MLE). Such nonparametric estimation methods allow the triggering density to be estimated without assuming a particular parametric form which may be subject to misspecification or over-fitting, which can be very serious problems, especially in social science applications (Yuan et al. 2019). Instead, the data drive the estimation of the triggering density, and this is especially attractive for use with the large data sets that are increasingly becoming available in applications. Unfortunately, however, a major limitation of current nonparametric estimation methods is their computational complexity and lack of speed, as existing methods are mainly based on maximum likelihood estimation (MLE) (Reinhart 2018), or variants such as EM-type algorithms (Veen and Schoenberg 2008; Marsan and Lengline 2008), which are typically non-convex problems without closed-form solutions. For applications to crime or social media, for instance, catalogs of millions of ST events are often the subject of study (Wang et al. 2018), and each calculation of the likelihood function with N events requires at least $O(N^2)$ time. In such situations, the estimation of the triggering density using existing methods can be infeasible. As a result, it is important to develop better alternatives to current MLE-based methods (Schoenberg et al. 2018a).

Recent developments in the nonparametric estimation of the Hawkes process provide new insights for this problem, including an analytic method for computing the MLE of the triggering density in the special case where the adjacency matrix is invertible (Schoenberg et al. 2018b), and generalized moments methods (GMM) for the estimation of the triggering matrix (Achab et al. 2017). However,

there are several limitations that prevent us to apply them directly to multivariate ST-Hawkes processes. The analytic MLE method in Schoenberg et al. (2018b) can only be applied to the univariate case, while the GMM method for the temporal process cannot estimate the triggering kernels. In this paper, we propose a new, highly computationally efficient, scalable nonparametric estimator for ST-Hawkes processes, based on a blend of these recent ideas with modern advances in the regularization and inversion of sparse matrices.

The contributions of this paper are threefold. First of all, we extend the analytic formula for the MLE of the step heights in the triggering density (Schoenberg et al. 2018b) to the multivariate ST case and greatly improve the stability of the resulting estimator using regularization. We next extend the cumulant-based estimators of (Achab et al. 2017) to the multivariate ST case and derive GMM estimators of the triggering matrix in this context. Finally, we combine the MLE and GMM estimators to obtain a scalable, consistent and efficient estimator and show that the proposed estimator has a computation complexity linear in the number of events N, allowing one to explore applications to large data sets with millions of events, in which our method outperforms current state-of-the-art methods in terms of both accuracy in network reconstruction and computation time.

The structure of this paper is as follows. We first review background material in Sect. 2. In Sect. 3, we develop the proposed method and show the consistency and computational complexity. The performance of this estimator is inspected using a variety of synthetic and real social network data sets in Sect. 4. Finally, we conclude and discuss important directions for future research in Sect. 5.

2 Multivariate Hawkes processes and nonparametric estimations

In this section, we review the definition of multivariate Hawkes processes and previous researches on inference methods, focusing especially on MLE and GMM.

A point process (Daley and Vere-Jones 2007, 2003) is a σ-finite collection of points $\{\tau_1, \tau_2, \ldots\}$ occurring in some metric space. While the definitions and results below can be extended quite readily to other spaces, we will assume for simplicity throughout that the metric space is a bounded interval $[0, T]$ in time or a bounded interval $B \times [0, T]$ in space-time. A temporal or ST point process is typically modeled via its conditional intensity, $\lambda(t)$ or $\lambda(s, t)$, which represents the infinitesimal rate at which points are accumulating at the particular location in time or space-time, given information on all points occurring prior to time t. Simple point processes are uniquely characterized by their conditional intensity (Daley and Vere-Jones 2007); for models for non-simple point processes, see Schoenberg (2006).

Hawkes processes are typically characterized via their conditional intensities. We refer readers to Daley and Vere-Jones (2007), Brillinger et al. (2002) for details about these concepts. For a simple temporal Hawkes process (Hawkes 1971), the conditional intensity of events at time t can be written

$$\lambda(t) = \mu + K \int_0^t g(t - t') \mathrm{d}N(t'), \tag{1}$$

where $\mu > 0$ is the background rate, $g(v) \geq 0$ is the *triggering density* satisfying $\int_0^\infty g(v)dv = 1$ which describes the conductivity of events, and the constant K is the productivity, which is typically required to satisfy $0 \leq K < 1$ in order to ensure stationarity and subcriticality (Hawkes 1971).

A *multivariate* temporal Hawkes process is conveniently viewed as a sequence of temporal point processes indexed by $u = 1, \ldots, U$, where each subprocess N_u has conditional intensity

$$\lambda_u(t) = \mu_u + \sum_{t_k < t} K_{u_k,u} \, g_{u_k}(t - t_k), \tag{2}$$

and the N points of the entire process may conveniently be labeled (t_k, u_k), for $k = 1, \ldots, N$, where t_k indicates the time of point k, and u_k indicates the index dictating to which subprocess the point belongs. The idea behind Eq. (2) is that the triggering density g_{u_k} and productivity $K_{u_k,u}$ may depend on the index of the point t_k.

In the model (2), μ_u is the background rate, indicating the rate at which points of mark u occur, absent any other prior events. For simplicity, one traditionally assumes a uniform background rate in time. $\boldsymbol{K} \in \mathbb{R}^{U \times U}$ is the triggering matrix, where $K_{u,v}$ is the expected number of events of index v that are triggered by one event of index u. This triggering effect, in this temporal-only case, is closely related to Granger causality (Granger 1969). In fact, subprocess u does not Granger-cause subprocess v if and only if $K_{u,v} = 0$ (Eichler et al. 2017). Similarly, for stationarity and subcriticality, $\boldsymbol{K}$ needs to satisfy $\|\boldsymbol{K}\| < 1$, where $\|\boldsymbol{K}\|$ is the spectral norm of $\boldsymbol{K}$.

In nonparametric estimation of g, one typically assumes that each subprocess has a piecewise-constant or basis-function representation of triggering densities $g_u(t)$ which control how quickly the rate $\lambda_u(t)$ returns to its baseline level μ_u after an event occurs. One can estimate the parameters $\boldsymbol{\mu} = (\mu_u)_u$, $\boldsymbol{K}$, and the triggering densities g via MLE (Ogata 1978) or minimize a regression loss (Chen et al. 2017). Here, we focus on the MLE approach. The log-likelihood function of the intensity function (2) becomes

$$l = \sum_{k=1}^{N} \log(\lambda_{u_k}(t_k)) - \sum_{u=1}^{U} \int_0^{\mathrm{T}} \lambda_u \mathrm{d}t. \tag{3}$$

One can directly maximize this function using off-the-shelf optimization methods or the EM-type algorithm proposed in Veen and Schoenberg (2008). See Yuan et al. (2019) for details about the derivation of the EM-type algorithm for ST-Hawkes processes. Another MLE-based approach, based on their analytic derivation of MLE, is first proposed in Schoenberg et al. (2018b) for the univariate case ($U = 1$). They found that one can solve the MLE problem via solving linear equations in g and two additional linear equations for the background rate μ and productivity K. However, for the multivariate case, the coefficients of these equations depend on the triggering

matrix $\boldsymbol{K}$ and it is no longer a linear system. Also, there is the problem of stability when the matrix of the linear system is singular or nearly singular. The inversion of the matrix is a major problem (Schoenberg et al. 2018b) in its implementation in practice, and in Sect. 3.2, we present the solution to this problem via regularization.

Another kind of estimation method (Achab et al. 2017; Bacry and Muzy 2016) is based on GMM using *cumulants* of Hawkes processes. Define $\mathbf{R} = (\mathbf{I} - \mathbf{K}^{\mathrm{T}})^{-1}$, where $\mathbf{I}$ is the identity matrix. As an alternative to the moments, the first, second and third *cumulant* of Hawkes process $\boldsymbol{\Lambda}$, $\boldsymbol{C}$ and $\boldsymbol{\Gamma}$ can be calculated analytically from $\mathbf{R}$ and $\boldsymbol{\mu}$. The idea of GMM is to match the cumulants from $\mathbf{R}$ and the cumulants approximated numerically from the data. Then, one can obtain the triggering matrix $\boldsymbol{K}^{\mathrm{T}} = \mathbf{I} - \mathbf{R}^{-1}$ by minimizing the non-convex approximation error of the cumulants. This provides a fast estimation procedure for both $\boldsymbol{\mu}$ and $\boldsymbol{K}$. But it does not estimate the triggering density, which plays an important role in the dynamics of the point process. In applications such as stochastic declustering (Zhuang et al. 2002), it is necessary to estimate triggering densities from the data. Some other moment-based methods (Bacry and Muzy 2016) can estimate both of them at the cost of high computation time.

3 Proposed methods for multivariate ST-Hawkes

In this section, we extend the previous discussion to the case of *multivariate* ST-Hawkes processes and derive a fast estimation method via extending and combining the two approaches (MLE and GMM) discussed above. We recommend interested readers to check (Reinhart 2018; Schoenberg et al. 2013) which provide comprehensive reviews of ST point processes. The focus of our method is to reduce the computational burden of the inference and improve the model estimation accuracy. Our motivation comes from the application of network reconstruction. Previous studies have shown the ability of Hawkes process models to uncover the underlying connections between nodes [such as social media users (Yuan et al. 2019), neurons (Chen et al. 2017), email users (Fox et al. 2016) and crime (Linderman and Adams 2014)]. It is essential to develop a scalable method because one often encounters data sets with thousands of nodes (large U) and millions of associated ST events (very large N).

We consider a multivariate ST-Hawkes process with a spatially isotropic triggering density $g(x, y, t)$—i.e., $g(x, y, t) = g(r, t), r = \sqrt{x^2 + y^2}$. ($g$ is only a function of time and distance.) This is a common assumption in real-world applications such as crime and earthquake aftershocks. We assume that $g(t)$ is identical for all subprocesses for simplicity. However, both assumptions can be easily extended to the general case via adding more variables and equations on g like (11). For each subprocess $u = 1, \ldots, U$, the conditional intensity characterizing the multivariate ST-Hawkes process is assumed to have the form

$$\lambda_u(x, y) = \mu_u(x, y) + \sum_{t_k < t} K_{u_k u} g(d_k, t - t_k)\,, \tag{4}$$

where (t_k, x_k, y_k, u_k), for $k = 1, \ldots, N$, denotes the N observed events in $B \times [0, T]$ and $d_k = \sqrt{(x_k - x)^2 + (y_k - y)^2}$. Current MLE-based methods such as the EM-type

algorithm (Veen and Schoenberg 2008; Yuan et al. 2019) are not well suited for large-scale problems due to its $O(N^3)$ computational complexity (Achab et al. 2017). Also, in many applications, it is difficult to ascertain the appropriate triggering density $g(r, t)$. Our proposed method has a linear $O(N)$ complexity and learns triggering densities directly from data. Specifically, we estimate $g(r, t)$ nonparametrically from MLE and $\boldsymbol{K}$, $\boldsymbol{\mu}$ from GMM. This combined method gives a fast and complete estimation of the ST-Hawkes process.

3.1 ST triggering density estimation

We extend the analytic method, first proposed in (Schoenberg et al. 2018b) for the univariate temporal case, to the case of multivariate ST-Hawkes processes.

First, we review the derivation of analytic estimates of the triggering function for the multivariate temporal Hawkes process (2). We assume that $\boldsymbol{\mu}$ and $\mathbf{K}$ are given or well estimated by other means, and the only variables here to be estimated are the heights of the step function comprising the triggering density $g(t) = \sum_{m=1}^{N_t} g_m \mathbb{1}_{t\in(\tau_m,\tau_{m+1})}$ with N_t grids $V_m = \{t \mid t \in (\tau_m, \tau_{m+1})\}, m = 1, \ldots, N_t$ dividing the time window $[0, T]$. The step function is a common assumption for the nonparametric method in Hawkes processes applications (Schoenberg et al. 2018b; Marsan and Lengline 2008). One seeks to obtain the step heights of the triggering density via maximizing the log-likelihood function. The log-likelihood function [from (3)]

$$l = \sum_{k=1}^{N} \log(\lambda_{u_k}(t_k)) - \sum_{u=1}^{U} \left(\mu_u T + \sum_{m=1}^{N_t} g_m \delta_m \sum_{k=1}^{N} K_{u_k u} \right) \tag{5}$$

is concave with respect to $\{g_m\}_m$. We take the derivative with respect to g_m and set it to zero:

$$0 = \frac{\partial l}{\partial g_m} = \sum_{(t_j - t_i)\in V_m} \frac{K_{u_i u_j}}{\lambda_{u_j}(t_j)} - \sum_{u=1}^{U} \sum_{i=1}^{N} K_{u_i u} \delta_m \,, \tag{6}$$

where $\delta_m = \tau_{m+1} - \tau_m$. Using the notation $\lambda = \{\lambda_{u_j}(t_j)\}_j$, $\boldsymbol{A}(k,j) = \sum_{t_j - t_i \in U_k} K_{u_i u_j}$, $\boldsymbol{\beta} = \{g_m\}_m$ and $\boldsymbol{b} = \{\sum_{u=1}^{U} \sum_{i=1}^{N} K_{u_i u} \delta_m\}_m$, we obtain a matrix representation of Eq. (6) as

$$0 = \boldsymbol{A}(1/\lambda) - b \,. \tag{7}$$

Here, $1/\boldsymbol{\lambda}$ is the element-wise reciprocal. The solution of (7) yields an estimate of $\boldsymbol{\lambda}$. Further, Eq. (2) can be rewritten as

$$\boldsymbol{\lambda} = \boldsymbol{\mu} + \boldsymbol{A}^{\mathrm{T}} \boldsymbol{\beta} \,. \tag{8}$$

Solving this equation using the estimate of $\boldsymbol{\lambda}$ from (7) provides the maximum likelihood estimate of $\boldsymbol{\beta}$.

We now focus on the *multivariate* ST-Hawkes process with a piecewise-constant ST triggering density $g(r,\ t)$. We simply assume a uniform background rate $\mu_u(x, y) = \mu_u$. For each subprocess $u = 1, \ldots, U$, the conditional intensity satisfies

$$\lambda_u(x, y, t) = \mu_u + \sum_{t_k<t} K_{u_k u} \sum_{m=1}^{N_t} \sum_{n=1}^{N_r} g_{mn} \mathbb{1}_{t_k - t \in (\tau_m, \tau_{m+1})} \mathbb{1}_{d_k \in (r_n, r_{n+1})} . \tag{9}$$

Here, $d_k = \sqrt{(x_k - x)^2 + (y_k - y)^2}$ and g is defined on a 2-D $N_r \times N_t$ grids with $V_n = \{d_k \mid d_k \in (r_n, r_{n+1})\}, n = 1, \ldots, N_r$ dividing the space and $V_m = \{t_k - t \mid t_k - t \in (\tau_m, \tau_{m+1})\}, m = 1, \ldots, N_t$ dividing the time window. The log-likelihood function of this intensity function is Schoenberg (2013)

$$\begin{aligned} l &= \sum_{k=1}^{N} \log(\lambda_{u_k}(x_k, y_k, t_k)) - \sum_{u=1}^{U} \iint_B \int_0^{\mathrm{T}} \lambda_u(x, y, t) \mathrm{d}t \, \mathrm{d}x \, \mathrm{d}y \,, \\ &= \sum_{k=1}^{N} \log(\lambda_{u_k}(x_k, y_k, t_k)) - \sum_{u=1}^{U} (\mu_u |B| T + \sum_m \sum_n g_{mn} \delta_m \Delta_n \sum_{k=1}^{N} K_{u_k u}) \,, \end{aligned} \tag{10}$$

where $|B|$ is the area of B, $\delta_m = \tau_{m+1} - \tau_m$. and $\Delta_n = \pi(r_{n+1}^2 - r_n^2)$. We calculate the spatial and temporal distance separately as this is the case for most applications. One can consider more complicated spatiotemporal distance via simply replacing $\delta_m \Delta_n$ with the distance measure over the spatiotemporal grid defined by the distance.

Assuming that $\boldsymbol{\mu}$ and $\boldsymbol{K}$ are given, the only variables here are $\{g_{mn}\}_{m,n}$. Maximizing the log-likelihood function will give us the estimation of the triggering density g. Since (10) is concave, we take the derivative of equation w.r.t. g_{mn}

$$0 = \frac{\partial l}{\partial g_{mn}} = \sum_{(t_j - t_i) \in V_m, d_{ij} \in V_n} \frac{K_{u_i u_j}}{\lambda_{u_j}(x_j, y_j, t_j)} - \sum_{u=1}^{U} \sum_{i=1}^{N} K_{u_i u} \delta_m \Delta_n \,, \tag{11}$$

with $d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$. Similar to the temporal case, we define $\boldsymbol{\lambda} = \{\lambda_{u_j}(x_j, y_j, t_j)\}_j$, $\boldsymbol{A}(k(m,n), j) = \sum_{t_j - t_i \in V_m, d_{ij} \in V_n} K_{u_i u_j}$, $\boldsymbol{\beta} = (g_{mn})_{k(m,n)}$ and $\boldsymbol{b} = (\sum_{u=1}^{U} \sum_{i=1}^{N} K_{u_i u} \delta_m \Delta_n)_{k(m,n)}$ with the index $k(m, n) = N_r(m - 1) + n$. Then, we obtain the matrix representation of (11) and (9) as

$$0 = \boldsymbol{A}(1/\boldsymbol{\lambda}) - \boldsymbol{b} \,, \tag{12}$$

$$\boldsymbol{\lambda} = \boldsymbol{\mu} + \boldsymbol{A}^{\mathrm{T}} \boldsymbol{\beta} \,. \tag{13}$$

Finally, we can estimate $\boldsymbol{\beta}$ via solving the above linear equations separately.

3.2 Regularization for linear system

As noted in Schoenberg et al. (2018b), in many applications, the matrix $\boldsymbol{A}$ in (12) and (13) is often ill-conditioned or singular, even with a careful selection of the 2-D grids V_m and V_n. Further, even when it can be obtained, the direct inverse $\boldsymbol{A}^{-1}\boldsymbol{b}$ (or

pseudo inverse $(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{b}$) can give unstable results due to over-fitting. In order to solve the linear equations in a stable and robust fashion, we use regularization procedures to find meaningful approximate solutions. As this is an important topic in inverse problems (Kaipio and Somersalo 2006), both classic regularization methods and statistical inversion theory using the Bayesian framework (Malinverno 2002) can be applied here.

More specifically, we propose the use of the Tikhonov regularization method (Neumaier 1998) with its analytic solution. For example, with the regularization, solving (12) becomes this minimization problem

$$\min_{\boldsymbol{x}} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2 + \|\boldsymbol{\Gamma}\boldsymbol{x}\|^2 , \tag{14}$$

for a Tikhonov matrix $\boldsymbol{\Gamma} = \alpha\boldsymbol{I}$. This is essentially the L_2 regularization, giving preference to solutions with smaller norms. L_1 regularization will typically give a sparse solution with many zero entities. It does not work here due to the fact that each element in $\boldsymbol{x} = 1/\lambda$ is positive and nonzero. Further one could utilize other Tikhonov matrices to guarantee smoothness if the underlying vector is believed to be mostly continuous. Instead of this, for the estimation of the triggering density, we smooth g with the post-processing approach below.

In some applications, one might want to separate the triggering density in space and time (Ogata 1998). As a post-processing step after estimation, we need to disentangle the spatiotemporal density $g(r, t)$ from $\boldsymbol{\beta}$ in order to compare with previous methods (Fox et al. 2016; Yuan et al. 2019), which assuming a separable density. As a result, we can decompose the triggering density $g(r, t)$ into the spatial triggering density $f(r)$ and temporal triggering density $h(t)$ (i.e., $g(r, t) = f(r)h(t)$). If we reshape the $N_r N_t$-by-1 vector $\boldsymbol{\beta}$ as a N_r-by-N_t matrix $\boldsymbol{B}$, then estimating the spatial and temporal triggering density becomes the following unmixing problem

$$\min_{\boldsymbol{f}\geq 0, \boldsymbol{h}\geq 0} \|\boldsymbol{B} - \boldsymbol{f}\boldsymbol{h}\|^2 . \tag{15}$$

Here, $\boldsymbol{B}$ is a nonnegative matrix based on the definition of $g(r, t)$ (triggering density function), $\boldsymbol{f}$ is a nonnegative N_r-by-1 vector and $\boldsymbol{h}$ is a nonnegative 1-by-N_t vector. This is, in fact, a rank-one nonnegative matrix factorization (NMF) (Lee and Seung 1999) $\boldsymbol{B} = \boldsymbol{f}\boldsymbol{h}$ and we solve it using singular value decomposition (SVD). Finally, we use a Gaussian moving average filter to smooth $\boldsymbol{f}$ and $\boldsymbol{h}$ to obtain the estimation of piecewise-constant triggering densities. This is based on our assumption that g is smooth and it can reduce the variance of our estimations. Our numerical experiments show that the regularization procedure described above leads to stable and robust estimations for synthetic and real-world data sets.

3.3 Triggering matrix estimation

In previous sections, we estimate the triggering density with the assumption that both $\boldsymbol{\mu}$ and $\boldsymbol{K}$ are given. In the univariate case, one can remove this assumption by adding two additional linear equations (Schoenberg et al. 2018b). However, in

the multivariate case, because matrix $\boldsymbol{A}$ is depend on matrix $\boldsymbol{K}$, solving $\boldsymbol{\mu}$, $\boldsymbol{K}$ and g simultaneously is no longer a linear problem.

In order to solve this problem, we extend the cumulants (22), (23), (18) to the ST case for a fast estimation of $\boldsymbol{\mu}$ and $\boldsymbol{K}$. For a ST-Hawkes process with U subprocesses, we define its first, second and third cumulant as (Daley and Vere-Jones 2007)

$$\Lambda^i \mathrm{d}t\mathrm{d}x\mathrm{d}y = \mathbb{E}(\mathrm{d}N^i_{t,x,y}), \tag{16}$$

$$C^{ij}\mathrm{d}t\mathrm{d}x\mathrm{d}y = \int_{\tau,a,b\in\mathbb{R}^3} (\mathbb{E}(\mathrm{d}N^i_{t,x,y}\mathrm{d}N^j_{t+\tau,x+a,y+b} - \mathbb{E}(\mathrm{d}N^i_{t,x,y})\mathbb{E}(\mathrm{d}N^j_{t+\tau,x+a,y+b}), \tag{17}$$

$$\begin{aligned}\Gamma^{ijk}\mathrm{d}t\mathrm{d}x\mathrm{d}y = \int_{\tau',a',b'\in\mathbb{R}^3}\int_{\tau,a,b\in\mathbb{R}^3} &(\mathbb{E}(\mathrm{d}N^i_{t,x,y}\mathrm{d}N^j_{t+\tau,x+a,y+b}\mathrm{d}N^k_{t+\tau',x+a',y+b'})\\ &+ 2\mathbb{E}(\mathrm{d}N^i_{t,x,y})\mathbb{E}(\mathrm{d}N^j_{t+\tau,x+a,y+b})\mathbb{E}(\mathrm{d}N^k_{t+\tau',x+a',y+b'})\\ &- \mathbb{E}(\mathrm{d}N^i_{t,x,y}\mathrm{d}N^j_{t+\tau,x+a,y+b})\mathbb{E}(\mathrm{d}N^k_{t+\tau',x+a',y+b'})\\ &- \mathbb{E}(\mathrm{d}N^i_{t,x,y}\mathrm{d}N^k_{t+\tau,x+a,y+b})\mathbb{E}(\mathrm{d}N^j_{t+\tau',x+a',y+b'})\\ &- \mathbb{E}(\mathrm{d}N^j_{t+\tau,x+a,y+b}\mathrm{d}N^k_{t+\tau',x+a',y+b'})\mathbb{E}(\mathrm{d}N^i_{t,x,y})).\end{aligned} \tag{18}$$

Here, $1 \le i,j,k \le U$ and τ, a and b are the variables of integration corresponding to t, x and y.

Cumulants can be numerically estimated from the ST data of events from each subprocess $Z^i = (t_k, x_k, y_k)_k, i = 1,\dots,U$ on the ST bounded area $B\times[0,T]$. Here, we simply assume that B is a rectangular with length X and width Y. We obtain the following estimation formulas for (16), (17) and (18):

$$\hat{\Lambda}^i = \frac{1}{TXY}\sum_{\tau,a,b\in Z^i} = \frac{N^i_{T,X,Y}}{TXY}, \tag{19}$$

$$\hat{C}^{ij} = \frac{1}{TXY}\sum_{\tau,a,b\in Z^i}(N^j_{a+\tilde{X},b+\tilde{Y},\tau+H} - N^j_{a-\tilde{X},b-\tilde{Y},\tau-H} - 8\tilde{X}\tilde{Y}H\hat{\Lambda}^j), \tag{20}$$

$$\begin{aligned}\hat{\Gamma}^{ijk} =& \frac{1}{TXY}\sum_{\tau,a,b\in Z^i}(N^j_{a+\tilde{X},b+\tilde{Y},\tau+H} - N^j_{a-\tilde{X},b-\tilde{Y},\tau-H} - 8\tilde{X}\tilde{Y}H\hat{\Lambda}^j)\times\\ &(N^k_{a+\tilde{X},b+\tilde{Y},\tau+H} - N^k_{a-\tilde{X},b-\tilde{Y},\tau-H} - 8\tilde{X}\tilde{Y}H\hat{\Lambda}^k)\\ &- \frac{\hat{\Lambda}^i}{TXY}\sum_{\tau',a',b'\in Z^k}\sum_{\tau,a,b\in Z^j}(2H - |\tau-\tau'|)^+(2\tilde{X} - |a-a'|)^+(2\tilde{Y} - |b-b'|)^+\\ &+ 64(H\tilde{X}\tilde{Y})^2\hat{\Lambda}^i\hat{\Lambda}^j\hat{\Lambda}^k,\end{aligned} \tag{21}$$

via numerical integration approximations of the cumulants on $[-\tilde{X}, \tilde{X}] \times [-\tilde{Y}, \tilde{Y}] \times [-H, H]$ assuming that the support of the triggering density is within this region [see Appendix B.3 in (Achab et al. 2017) for more details]. One also needs to symmetrize the approximated cumulants via $(\hat{C}^{ij} + \hat{C}^{ji})/2$ and $(2\hat{\Gamma}^{ijj} + \hat{\Gamma}^{jii})/3$ because the actual cumulants satisfy $\Gamma^{ijj} = \Gamma^{ij}$ and $C^{ij} = C^{ji}$.

The first, second and third cumulants of Hawkes process $\boldsymbol{\Lambda}$, $\boldsymbol{C}$ and $\boldsymbol{\Gamma}$ also have the following relationships (Achab et al. 2017) with $\boldsymbol{R}$

$$\boldsymbol{\Lambda}(i) = \Lambda^i = \sum_{m=1}^{U} R^{im} \mu_m \,, \tag{22}$$

$$\boldsymbol{C}(i,j) = C^{ij} = \sum_{m=1}^{d} \Lambda^m R^{im} R^{jm} \,, \tag{23}$$

$$\boldsymbol{\Gamma}(i,j,k) = \Gamma^{ijk} = \sum_{m=1}^{d} (R^{im} R^{jm} C^{km} + R^{im} C^{jm} R^{km} + C^{im} R^{jm} R^{km} - 2\Lambda^m R^{im} R^{jm} R^{km}) \,. \tag{24}$$

Here, $R^{im} = \boldsymbol{R}(i,m)$. Although the definition and numerical estimations of the cumulants are different for the ST case, the above formulas hold for both temporal and spatiotemporal cases because the spatial information can be viewed as "marks" of the temporal point process. According to (23) and (24), we can obtain the triggering matrix $\boldsymbol{K}^{\mathrm{T}} = \mathbf{I} - \mathbf{R}^{-1}$ by minimizing the approximation error of the cumulants with some scaling coefficient κ

$$L(\boldsymbol{R}) = (1-\kappa) \| \boldsymbol{R}^{\odot 2} \hat{\boldsymbol{C}}^{\mathrm{T}} + 2(\boldsymbol{R} \odot (\hat{\boldsymbol{C}} - \boldsymbol{R}\hat{\boldsymbol{L}})) \boldsymbol{R}^{\mathrm{T}} - \hat{\boldsymbol{\Gamma}}^{\boldsymbol{c}} \|_2^2 + \kappa \| \boldsymbol{R}\hat{\boldsymbol{L}}\boldsymbol{R}^{\mathrm{T}} - \hat{\boldsymbol{C}} \|_2^2 \,. \tag{25}$$

Here, $\odot$ is the Hadamard product and $\hat{\boldsymbol{\Gamma}}^{\boldsymbol{c}} = \hat{\boldsymbol{\Gamma}}(i,i,k)$. Given the estimated $\tilde{\boldsymbol{R}}$, we also have $\tilde{\boldsymbol{\mu}} = \tilde{\boldsymbol{R}}^{-1} \tilde{\boldsymbol{\Lambda}}$ from the cumulants equation (22).

Finally, we can plug the approximated cumulants into (25) to estimate $\boldsymbol{\mu}$ and $\boldsymbol{K}$. The error function (25) is a non-convex polynomial and similar to the loss function of a multilayer neural network. As a result, stochastic gradient descend (SGD) with acceleration [e.g., Adam Kingma and Ba (2015) or AdaGrad Duchi et al. (2011)] can be used to minimize the error function. With a good choice of the initial value such as in Achab et al. (2017), SGD often leads to satisfying convergence results and is more accurate than EM-type algorithms in many applications (Achab et al. 2017). The normalization term κ is $\kappa = \frac{\|\hat{\boldsymbol{\Gamma}}^{\boldsymbol{c}}\|_2^2}{\|\hat{\boldsymbol{C}}\|_2^2 + \|\hat{\boldsymbol{\Gamma}}^{\boldsymbol{c}}\|_2^2}$ based on the theory of GMM (Achab et al. 2017). The ratio between the support of the triggering density and the ST bounded area $B \times [0,T]$ matters for the consistency of the GMM (Achab et al. 2017). Usually for specific applications such as social network reconstruction, $B \times [0,T]$ is much larger than the square of the support of the triggering density, which guarantees the consistency of the GMM estimation.

3.4 Consistency guarantee

The consistency of maximum likelihood estimates (Ogata 1978) or GMM estimates (Achab et al. 2017) is guaranteed by general theoretical results. Here, we note that our proposed method, as a combination of GMM and MLE, also yields consistent estimates.

First, as background, note that in Ogata (1978), Ogata showed the MLE of the full vector of parameters is, under quite general conditions, consistent. Also, if only some of the parameters are to be estimated and others, such as in this instance $\boldsymbol{K}$ and $\boldsymbol{\mu}$, are known exactly, then again one may consider the parameter vector to be only those parameters being estimated, and again (Ogata 1978) showed the estimated ones will be consistent. However, we are considering the case where $\boldsymbol{K}$ and $\boldsymbol{\mu}$ are not known but are estimated consistently via GMM, and then, the other parameters are estimated by MLE. To the best of our knowledge, this case has not been studied previously, and the result does not immediately follow from the theorems in Ogata (1978). We show that $\hat{\boldsymbol{\beta}}$ inherits the property of consistency from the MLE and GMM estimators, under the same assumptions as in Achab et al. (2017) and Ogata (1978). For simplicity purposes, we will not list these groups of assumptions, which are mainly related to observations and regularity conditions.

Let Θ denote the full vector of parameters, including $\boldsymbol{K}$ and $\boldsymbol{\mu}$. Let Θ_0 denote the true value of Θ. Let $\boldsymbol{U}$ denote a neighborhood of Θ_0. Let $\boldsymbol{K'}$ and $\boldsymbol{\mu'}$ denote the GMM estimates of $\boldsymbol{K}$ and $\boldsymbol{\mu}$. Let $\Theta = (\boldsymbol{K}, \boldsymbol{\mu}, \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is the vector of other parameters estimated by MLE. Let $\hat{\boldsymbol{K}}$, $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\beta}}$ be the MLEs of these parameters.

Theorem 1 *Assuming the same regularity conditions used in the proofs of consistency of the MLE and GMM estimator in* Achab et al. (2017) *and* Ogata (1978) *(see* "Appendix 3" *for details), the combined estimator* $\hat{\boldsymbol{\beta}} \to \boldsymbol{\beta}$ *in probability as* $T \to \infty$.

Proof Let L denote the log-likelihood divided by T. Thus, L depends on T, but we will suppress this here. Let Θ_1 denote the supremum over $\boldsymbol{U}^c$ of L, which is the MLE outside of $\boldsymbol{U}$.

We are given that $(\boldsymbol{K'}, \mu') \to (\boldsymbol{K}, \mu)$ in probability as $T \to \infty$. Thus, $(\boldsymbol{K'}, \mu')$ are in $\boldsymbol{U}$ with probability going to 1 as $T \to \infty$.

We have $L(\Theta_1) \to \mathbb{E}(L(\Theta_1))$ and $L(\Theta_0) \to \mathbb{E}(L(\Theta_0))$, where this convergence is in probability as $T \to \infty$ and is uniform in Θ. This follows from the same logic as in the proof of Theorem 2 of Ogata (1978).

Similarly, following exactly as in the proof on p253 of Ogata (1978), we have $\mathbb{E}(L(\Theta_0)) > \mathbb{E}(L(\Theta_1))$ and for sufficiently large T, there exists $\epsilon > 0$ such that $|\mathbb{E}(L(\Theta)) - \sup_{\Theta \notin U} \mathbb{E}(L(\Theta))| > \epsilon/2$. This follows from the assumptions in Ogata (1978), particularly the assumption that λ is uniformly bounded away from 0.

Therefore, since $(\boldsymbol{K'}, \mu')$ are in $\boldsymbol{U}$ as $T \to \infty$, for sufficiently large T, we have $\mathbb{E}(L(\Theta_0))$, and therefore, $L(\Theta_0)$ is also maximized within U with probability going to 1. More specifically, for any $\epsilon > 0$, there is $\delta > 0$ and sufficiently large T so that

$$
\begin{aligned}
P(\hat{\boldsymbol{\beta}} \notin \boldsymbol{U}) &= P\{\sup_{\boldsymbol{U}^c} L(\Theta) \geq \sup_{\boldsymbol{U}} L(\Theta)\} \\
&\leq P\{L(\Theta_1) \geq L(\Theta_0)\} \\
&\leq P\{L(\Theta_1) - \mathbb{E}(L(\Theta_1)) \geq \delta\} + P\{\mathbb{E}(L(\Theta_1)) - \mathbb{E}(L(\Theta_0)) > -2\delta\} \\
&\quad + P\{\mathbb{E}(L(\Theta_0)) - L(\Theta_0) \geq \delta\} \\
&\leq \epsilon/2 + 0 + \epsilon/2 \\
&= \epsilon\,.
\end{aligned}
$$

□

3.5 Computational complexity

The state-of-the-art cumulants-based method (NPHC) (Achab et al. 2017) for temporal triggering density estimation has a complexity of $O(NU^2 + N_{\text{iter}}U^3)$, where N_{iter} is the number of iterations for SGD (around 200 for our applications). Our method has a similar complexity $O(NU^2 + N_{\text{iter}}U^3 + (N_r N_t)^3)$ as NPHC since the calculation time of spatiotemporal cumulants is just a constant multiple of temporal cumulants. The additional calculation for triggering density estimation is usually negligible because N_r, N_t are small constants (we use 50 in experiments) and $\boldsymbol{A}$ is usually sparse. For an EM-type algorithm (EM) (Lewis and Mohler 2011), the complexity is $O(N_{\text{iter}}N^3U^2)$ (Achab et al. 2017). With some clever implementation or in some special cases (e.g., temporal Hawkes process with an exponential triggering density), one can reduce this to $O(N^2)$ or better.

Our method outperforms EM when $N \gg U$. Moreover, in many cases, we find that our method is even faster than NPHC. This seems impossible since our method needs to process spatial data in addition to the timestamp. However, for ST data, there are many event pairs that are close in time (within the support of the temporal triggering density) while spatially separated from each other (outside the support of the spatial triggering density). Temporal-only model such as NPHC will calculate these events pairs during the estimation of cumulants. This might cause false positives in causal inference. Our method, on the other hand, uses spatial information to exclude these events. It seems that, for a majority of data sets we examined, this effect is very significant and our method can be much faster than NPHC.

4 Numerical examples

In this section, we compare our method, which is called ST-Hawkes cumulants (STHC) throughout this section, with other popular estimation methods for multivariate Hawkes processes on various data sets. We consider both simulation data and real-world social network data. First, we simulate multiple synthetic data sets with different sizes, triggering matrices and triggering densities. These data sets with ground-truth information allow us to examine different methods in detail. Then, for real-world applications, we further evaluate the performance of these methods on

the task of network reconstruction for multiple location-based social network check-in data sets. Moreover, our method directly estimates spatial and temporal triggering densities, which provides a useful tool for the study of ST dynamics among these check-in events. We conduct all of the experiments on a single machine with a NVIDIA 970 GPU (4 GB memory), 4-core Intel i7-6700K CPU (4.20 GHz) and 16 GB of RAM.

4.1 Synthetic data

Our synthetic data sets are generated using Algorithm 3 in Yuan et al. (2019), which is based on the clustering representation of Hawkes process. We simulate various ST-Hawkes processes and use them to evaluate our method (STHC), the state-of-the-art temporal cumulants method (NPHC) and EM-type algorithm (EM). The details about the simulation, preprocessing and hyperparameters (such as α for the regularization) are described in "Appendix 1." Here, we define some error measurements used in this section.

- *Relative error* between the estimated triggering matrix $\hat{\boldsymbol{K}}$ and the ground-truth matrix $\boldsymbol{K}$:

$$\mathrm{RelErr}\left(\boldsymbol{K}, \hat{\boldsymbol{K}}\right) = \frac{1}{U^2} \sum_{u,v} \left(\frac{|K_{uv} - \hat{K}_{uv}|}{|K_{uv}|} \mathbb{1}_{K_{uv} \neq 0} + |\hat{K}_{uv}| \mathbb{1}_{K_{uv}=0} \right).$$

- *Mean squared error* (MSE) between the estimated triggering densities (temporal $\hat{h}(t)$, spatial $\hat{f}(r)$ and combined $\hat{g}(r,t)$) and the ground-truth triggering densities (temporal $h(t)$, spatial $f(r)$ and combined $g(r, t)$):

$$\mathrm{MSE}_r = \frac{1}{N_r} \sum_{i=1}^{N_r} (f_i - \hat{f}_i)^2, \ \mathrm{MSE}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} (h_i - \hat{h}_i)^2, \ \mathrm{MSE}_\beta = \frac{1}{N_r N_t} \sum_{i=1}^{N_r} \sum_{j=1}^{N_t} (g_{ij} - \hat{g}_{ij})^2 .$$

Here, $\hat{g}_{ij} = \boldsymbol{B}(i,j)$ is the discrete estimation of the triggering density on a 2-D grid of size 50×50 and g_{ij} are the ground-truth values of the triggering density on the grid. $\hat{h}_i = \hat{\boldsymbol{h}}(i)$ and $\hat{f}_i = \hat{\boldsymbol{f}}(i)$ are from the NMF decomposition of $\boldsymbol{B}$, and $h_i = \boldsymbol{h}(i)$ and $f_i = \boldsymbol{f}(i)$ are the ground-truth values of the temporal and spatial triggering densities on the grid accordingly.

4.1.1 Triggering density estimation

We first compare our methods with EM in terms of the triggering density estimation accuracy (NPHC does not estimate triggering densities). Simulation data with 2587 events are from a ST-Hawkes process with $U = 1$, exponential triggering density in time and Gaussian in space. We get a good estimation of the triggering density $f(r)$ ($\mathrm{MSE}_r = 0.001662$), $h(t)$ ($\mathrm{MSE}_t = 0.02876$) in Fig. 1 and the overall estimation for $\boldsymbol{\beta} = (g_{mn})_{k(m,n)}$ ($\mathrm{MSE}_\beta = 0.03400$). This is a relatively small data sets so that we can use EM for ST-Hawkes [ST-EM, see Yuan et al. (2019)] estimation. For ST-EM, we

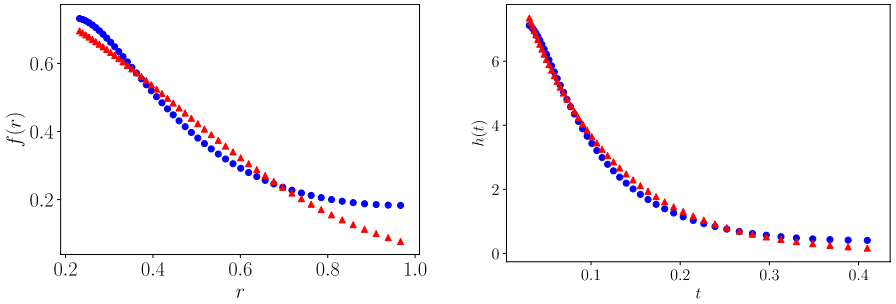


Fig. 1 The estimation results of STHC on $U = 1$ data. Ground-truth spatial triggering density $f(r)$ as red triangles and estimated triggering density as blue circles (left). Temporal triggering density $h(t)$ as red triangles and estimated triggering density as blue circles (right)

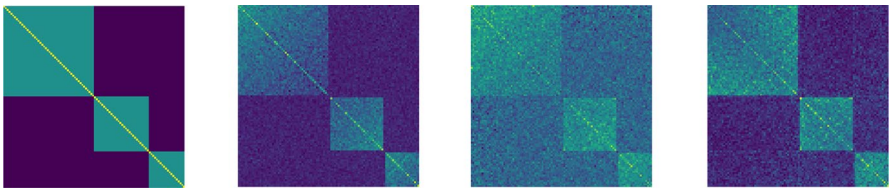


Fig. 2 Ground-truth $\boldsymbol{K}$ matrix, STHC , NPHC result and EM estimation results (from left to right)

get $f(r)$ ($\text{MSE}_r = 0.01485$), $h(t)$ ($\text{MSE}_t = 0.004058$) and $\boldsymbol{\beta}$ ($\text{MSE}_\beta = 0.2533$). Our method is faster (see Table 1) and overall more accurate.

4.1.2 Triggering matrix

Then, we evaluate the ability of our model to recover the triggering matrix $\boldsymbol{K}$. This is important for many applications such as network reconstruction and causal inference. On our existing architecture, the ST-EM method runs out of memory. Instead, we use EM and NPHC implementations in the tick package Bacry et al. (2017) for the following comparisons.

We simulate a ST-Hawkes process with $U = 100$ and a symmetric $\boldsymbol{K}$ matrix (see Fig. 2) because our network reconstruction data sets mainly have undirected social networks. We achieve a relative error of 0.1080. In the same setting, we get a relative error of 0.1626 for NPHC and 0.1459 for EM. The improvement in computation time (see Table 1) is significant.

4.1.3 Combined estimation

Now, we combine the two steps together and give a complete estimation of ST-Hawkes processes. We simulate a ST-Hawkes process with $U = 10$ and 179,176 events in total. From the results in Fig. 3 and Table 1, STHC gives very fast and also

Table 1 The computation time for different methods on synthetic data sets. Here, the time is in second

	STHC	NPHC	EM
$U = 1$	0.165528	–	4.643132
$U = 10$	1.073085	1.093068	4.707377
$U = 100$	2.608996	4.174796	43.781988

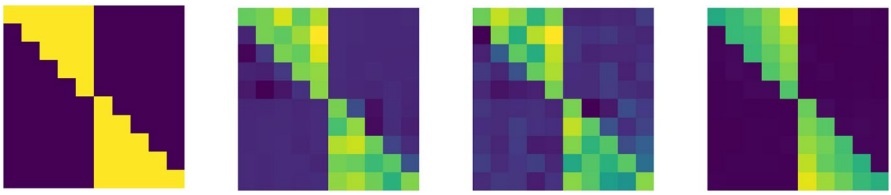


Fig. 3 Ground-truth $\boldsymbol{K}$ matrix, STHC, NPHC and EM estimation results (from left to right)

accurate estimations (RelErr = 0.02901) compared to NPHC (RelErr = 0.04899) and EM (RelErr = 0.03269). We then threshold $\hat{\boldsymbol{K}}$ with $\epsilon = 0.01$ to remove noise. Using $\hat{\boldsymbol{K}}, \hat{\boldsymbol{\mu}}$, we get a good estimation of the triggering density $f(r)$ and $h(t)$ in Fig. 4 with $\text{MSE}_r = 0.002381$, $\text{MSE}_t = 0.06664$ and $\text{MSE}_\beta = 0.1067$, while EM has a much worse MSE ($\text{MSE}_t = 0.9512$) since it does not consider spatial information.

4.1.4 Combined estimation with different triggering densities

We modify the above $U = 10$ data set via replacing the ST triggering density with different functions. We first get accurate estimations of $\tilde{\boldsymbol{K}}$ and $\tilde{\boldsymbol{\mu}}$. Given $\tilde{\boldsymbol{K}}$ and $\tilde{\boldsymbol{\mu}}$, we then estimate the triggering density in space and time. The results are summarized in Table 2. Specifically, we consider Pareto triggering density in time, uniform triggering density in time, power-law triggering density in space and uniform triggering density in space. See "Appendix 1" for more details on generating these synthetic data sets and visualizations of triggering kernels.

In summary, for synthetic data, our method consistently performs better than all baselines in the estimation of triggering matrices and densities in terms of the mean square error and relative error. This is consistent with the finding in Achab et al. (2017) that moment-based methods, which do not require the estimation of triggering densities, are less sensible to model misspecification and more accurate compared with EM. Moreover, our approach is much faster than EM-type algorithms, which are not scalable to large-scale spatiotemporal data.

4.2 Location-based social network reconstruction

In many situations, network data are incomplete and it may not be possible to directly observe the hidden relationships between nodes. Point process models, especially Hawkes processes, are widely used to infer the hidden connections via viewing each

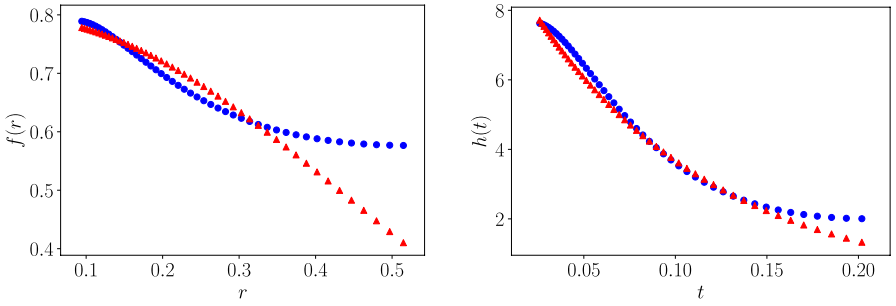


Fig. 4 The estimation results of STHC on $U = 10$ data. Ground-truth spatial triggering density $f(r)$ as red triangles and estimated triggering density as blue circles (left). Temporal triggering density $h(t)$ as red triangles and estimated triggering density as blue circles (right)

Table 2 Error measures for STHC on $U = 10$ data sets with different triggering densities

	MSE_r	MSE_t	$\text{RelErr}(\boldsymbol{K}, \hat{\boldsymbol{K}})$
Pareto in time	0.01244	0.0009966	0.02784
Uniform in time	0.01320	1.296×10^{-5}	0.09306
Power law in space	0.0003904	0.04463	0.0409
Uniform in space	0.0006231	0.1294	0.04552

subprocess as a node in the network. Then, the estimated triggering matrix $\hat{\boldsymbol{K}}$, which uncover the macroscale causality between users (subprocesses), can recover underlying connections between neurons (Linderman and Adams 2014), financial markets (Achab et al. 2017) and social media users (Yuan et al. 2019). The assumption here is that this causality information reflects actual friendship connections. Our task of network reconstruction is to uncover the ground-truth friendship network among social media users using only the information of each user's check-ins.

The Gowalla and Brightkite data sets, collected in Cho et al. (2011), are both from location-based social media websites in which users share their locations by checking in. Authors in Cho et al. (2011) use public APIs to collect user "friendship" networks, location profiles and users' spatiotemporal check-in history. These data sets have already been studied in many areas, such as human mobility (Cho et al. 2011), geometric learning problems (Bronstein et al. 2017) and location-based recommendation systems (Bao et al. 2012; Yuan et al. 2020), and become benchmark data sets of location-based online social networks. Previous studies on these data sets reveal that human movement patterns are often a combination of periodic behaviors and behaviors related to social relationships. Multivariate Hawkes process models could be able to capture both parts via the background rate and mutual triggering between users. Here, we verify this assumption through the network reconstruction task.

In detail, Gowalla has a "friendship" network with 196,591 users, 950,327 edges and a total of 6,442,890 check-ins of these users between February 2009

and October 2010. Brightkite's "friendship" network consists of 58,228 nodes and 214,078 edges, and a total of 4,491,143 check-ins over the period of April 2008–October 2010. Each check-in record includes the latitude and longitude coordinates, a user ID and the time (with a precision of one second). Similar to the Facebook "friendship" network, both the Gowalla and Brightkite friendship networks are undirected and unweighted. We study several subnetworks (Gowalla-SF, Brightkite-LA, Gowalla-CHI, and Brightkite-SD) within these data sets; see "Appendix 2" for details.

To reconstruct these user subnetworks, we model the ST check-ins of each user within a subnetwork as events of one subprocess within a multivariate ST-Hawkes process. Then, we infer relationships between these users (i.e., infer adjacency matrix) from the triggering matrix $\boldsymbol{K}$. We compare our method (STHC) with NPHC and EM in terms of how well the reconstructed networks match the ground-truth friendships. With the prior information that friendship networks are undirected, we first symmetrize the inferred triggering matrix (via $\tilde{\boldsymbol{K}} = \left(\hat{\boldsymbol{K}} + \hat{\boldsymbol{K}}^{\mathrm{T}}\right)/2$) to obtain the estimated weighted adjacency matrix. Then, the network reconstruction becomes a binary classification problem with the probability $\propto \tilde{\boldsymbol{K}}$. Given the ground-truth binary adjacency matrix, we calculate the corresponding receiver operating characteristic (ROC) curves and the area under the curve (AUC) to evaluate the results.

The performances of different methods are examined on various subnetworks with different sizes. Our STHC method consistently outperforms other methods with more than 20% improvement in terms of the AUC in Fig. 5. The improvement is mainly from the ability of our method to exclude false-positive connections. We show an example of network reconstruction results of Brightkite-SD in Fig. 6. For the computation time (see Table 3), STHC scales better than NPHC in all data sets, as explained in Sect. 3.5. EM has the worst scaling due to its superlinear complexity. Finally, we estimate spatiotemporal triggering densities with $N_r = N_t = 50$ for these subnetworks and plot their spatial and temporal marginal triggering densities from (15) in Figs. 7 and 8 separately. The spatial triggering densities for different subnetworks have similar shapes with a cutoff around 10^{-4}. This could come from the fact that the check-in location is usually fixed for a point of interest (POI, such as shop/cafe/gym). The triggering density also implies that the spatial triggering effects between users have a short radius, which mainly occur when they visit the same POI. These temporal triggering densities also share the same trend. The triggering effects only peak a few hours after the event time. This is also observed in other data sets, such as the insurgency activity in Iraq (Lewis and Mohler 2011).

5 Conclusions

We present a novel inference approach of ST-Hawkes processes; it is the most efficient and accurate method in comparison with other popular estimation methods, according to the numerical experiments presented. Moreover, this approach

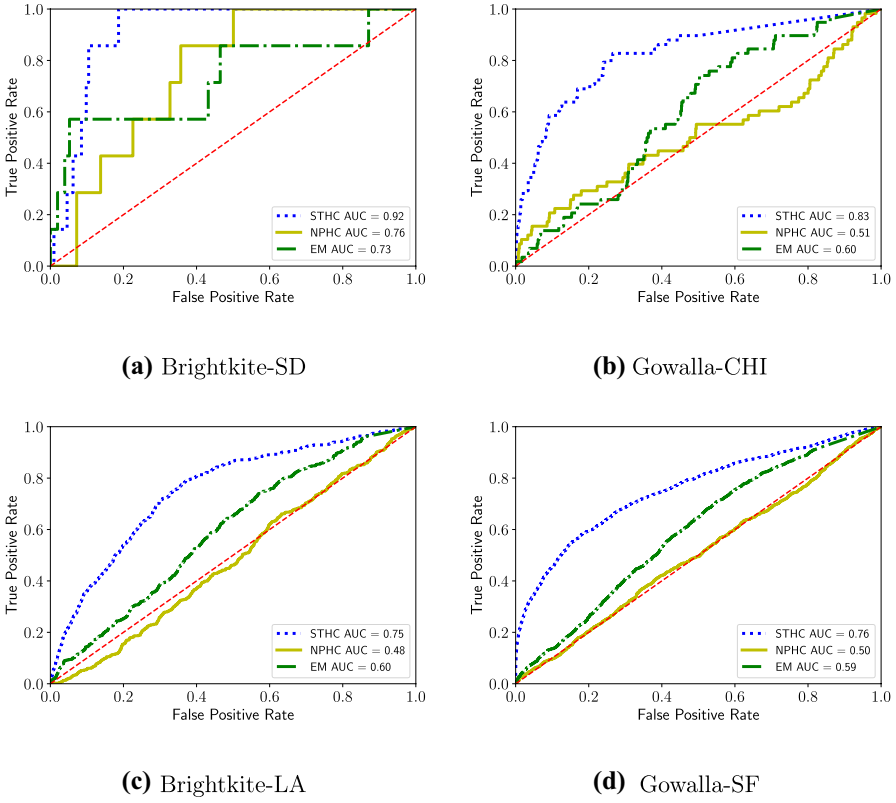


Fig. 5 ROC curves of different methods (STHC, NPHC and EM) on subnetworks in Gowalla and Brightkite data sets. The dashed line (red) is from random guess

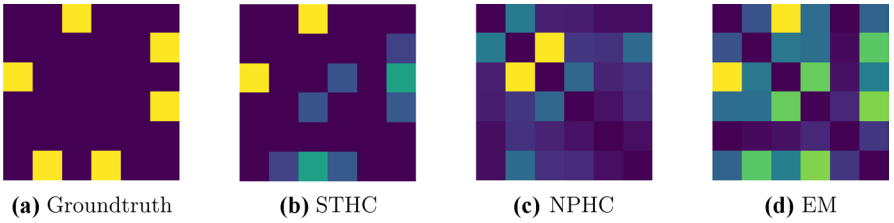


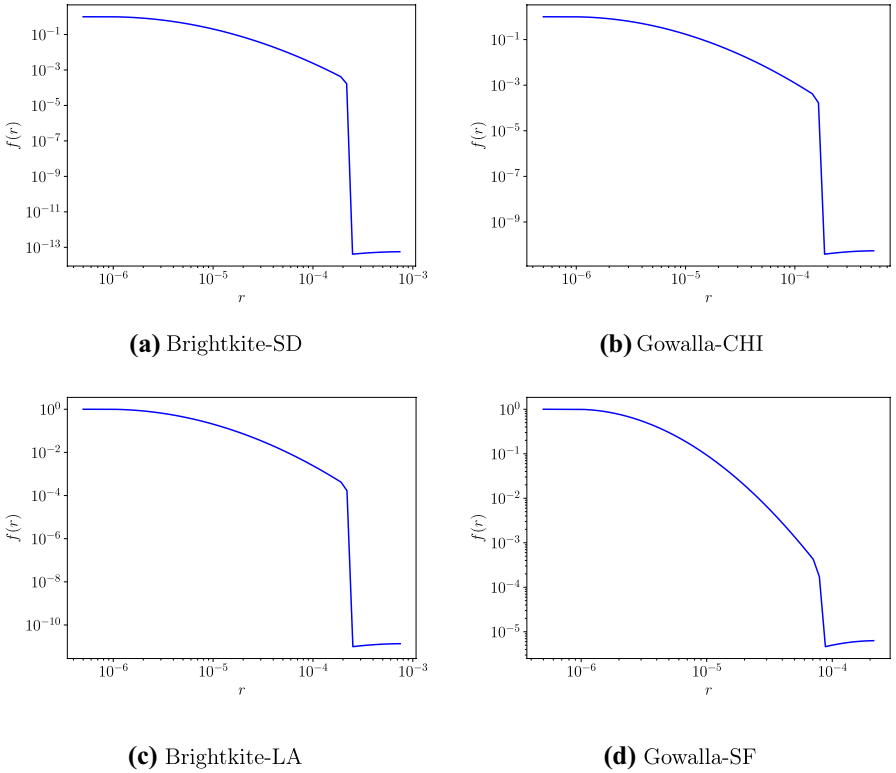
Fig. 6 Friendship network reconstruction using different methods on Brightkite-SD. Here, we zoom in to show a subgraph within the Brightkite-SD network

is successfully applied to network reconstruction problems and leads to promising applications for the inference of causal relationships and social interactions.

A point that should be stressed is that we make a few model assumptions to simplify the estimation procedure. To recapitulate, we assume a constant background

Table 3 The computation time for different methods on Gowalla and Brightkite data sets. Here, the time is in second

	STHC	NPHC	EM
Brightkite-SD	0.271304	2.035561	2.252009
Gowalla-CHI	2.978064	3.869652	15.474624
Brightkite-LA	3.976395	7.001311	36.357789
Gowalla-SF	40.754037	76.514422	180.918273



(a) Brightkite-SD **(b)** Gowalla-CHI

(c) Brightkite-LA **(d)** Gowalla-SF

Fig. 7 Estimated spatial triggering densities of our method on Gowalla and Brightkite data sets. The plot is in log–log scale, and we normalize the triggering density for easy comparison. Hyperparameters include $\alpha = 0.5$ and $N_r = N_t = 50$

rate in space and no boundary effect for events outside the area that we studied. For more general spatial background (inhomogeneous) distributions, one can approximate it using a piecewise-constant function in space by dividing events into spatial grids. Essentially, for each grid, we still have a uniform background for estimation

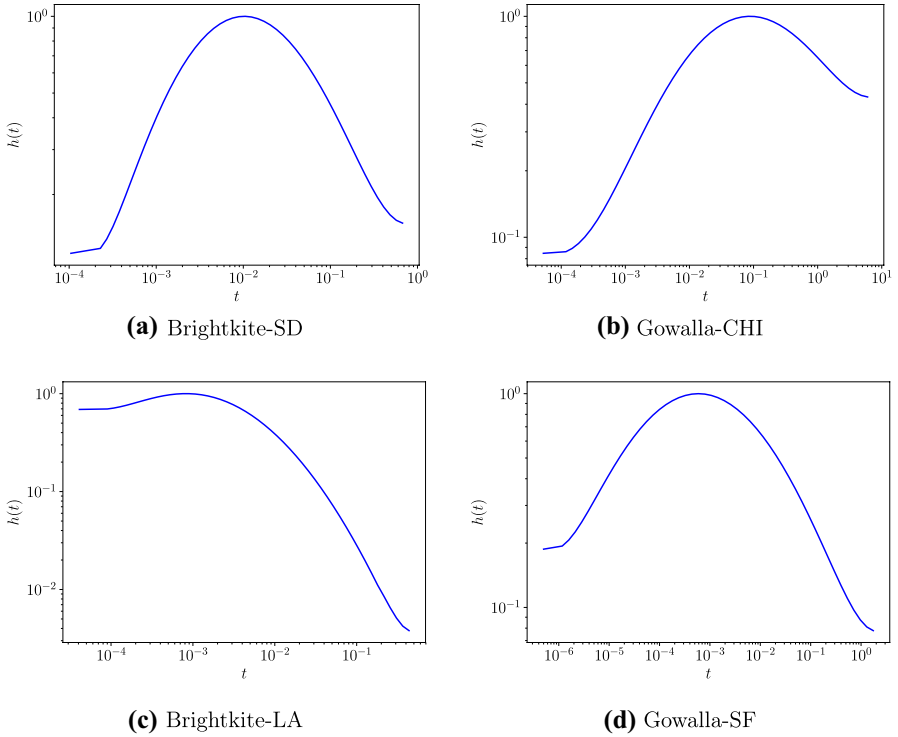


Fig. 8 Estimated temporal triggering densities of our method on Gowalla and Brightkite data sets. The plot is in log–log scale, and we normalize the triggering density for easy comparison. Hyperparameters include $\alpha = 0.5$ and $N_r = N_t = 50$

and then combine them. For applications on large areas with an inhomogeneous background, we expect a piecewise-constant or covariate-based background rate to achieve even better results (Schoenberg et al. 2018b), and incorporating boundary effects helps to remove bias in the estimation of the background rate and triggering densities (Reinhart 2018). Moreover, the current regularization method can be extended to a more general case to utilize the smoothness proprieties of triggering densities.

Finally, while we are focusing on the general case of multivariate ST-Hawkes processes, the current method can be very useful for the estimation of univariate models. The regularization improves the stability and robustness of the analytic method in Schoenberg et al. (2018b). This makes it possible to apply univariate models to the study of large data sets in areas like seismology, epidemiology and criminology.

Acknowledgements This work was supported by the City of Los Angeles Gang Reduction Youth Development Project, by NSF grant DMS-2027277 and by NSF grant DMS-1737770. Baichuan Yuan gratefully acknowledges the fellowship support of the National Institute of Justice (NIJ) under Award Number 2018-R2-CX-0013.

Appendix 1: Simulation data

$U = 1$ data

We simulate a univariate ST-Hawkes process with $K = 1/6$, $\mu = 0.01, T = 2.1 \times 10^5$, $X, Y \in (0, 10)$, $f(r) = \frac{1}{2\pi\sigma^2} \exp(-r^2/2\sigma^2)$ $(\sigma^2 = 0.2)$ and $h(t) = \omega \exp(-\omega t)$ $(\omega = 10)$. The regularization parameter $\alpha = 0.5$.

$U = 100$ data

Using the same triggering densities, this data set has the following parameters: $U = 100$, the background rate $\boldsymbol{\mu} = (0.01, \ldots, 0.01)$. $T = 10^5$, $X, Y \in (0, 10)$, $\sigma^2 = 0.2$ and $\omega = 10$ with 172,943 events. For the triggering matrix in Fig. 2, each yellow pixel is 1/20, cyan pixel is 1/40 and dark pixel is 0.

$U = 10$ data

With the same densities, the parameters are $U = 10$, $\boldsymbol{\mu} = (0.01, \ldots, 0.01)$, $T = 1e6$, $X, Y \in (0, 10)$, $\sigma^2 = 0.2$, $\omega = 10$ and $\boldsymbol{K}$ is shown in Fig. 3. Here, each yellow pixel is 1/6 and dark pixel is 0. The regularization parameter $\alpha = 0.55$.

$U = 10$ data with a Pareto triggering density in time

We keep the same parameters as the $U = 10$ above. The changes on the densities are on the temporal density $h(t) = (p-1)c^{p-1}/(t+c)^p$ with $c = 2$ and $p = 2.5$ and the same spatial triggering density with $\sigma^2 = 0.1$. The regularization parameter $\alpha = 0.38$.

$U = 10$ data with a uniform triggering density in time

Similar to the section above, here we change the temporal densities to be uniform $h(t) = 0.1$ and the spatial triggering density with $\sigma^2 = 0.1$. The regularization parameter $\alpha = 0.4$. We threshold the estimated $\tilde{\boldsymbol{K}}$ with $\epsilon = 0.01$ to remove noise.

$U = 10$ data with a power-law triggering density in space

Similarly, we use the power-law density $f(r) = \frac{1}{(r^2+1)^2}$ in space and the exponential triggering density in time with $\omega = 10$. The regularization parameter $\alpha = 0.28$. We threshold the estimated $\tilde{\boldsymbol{K}}$ with $\epsilon = 0.02$ to remove noise.

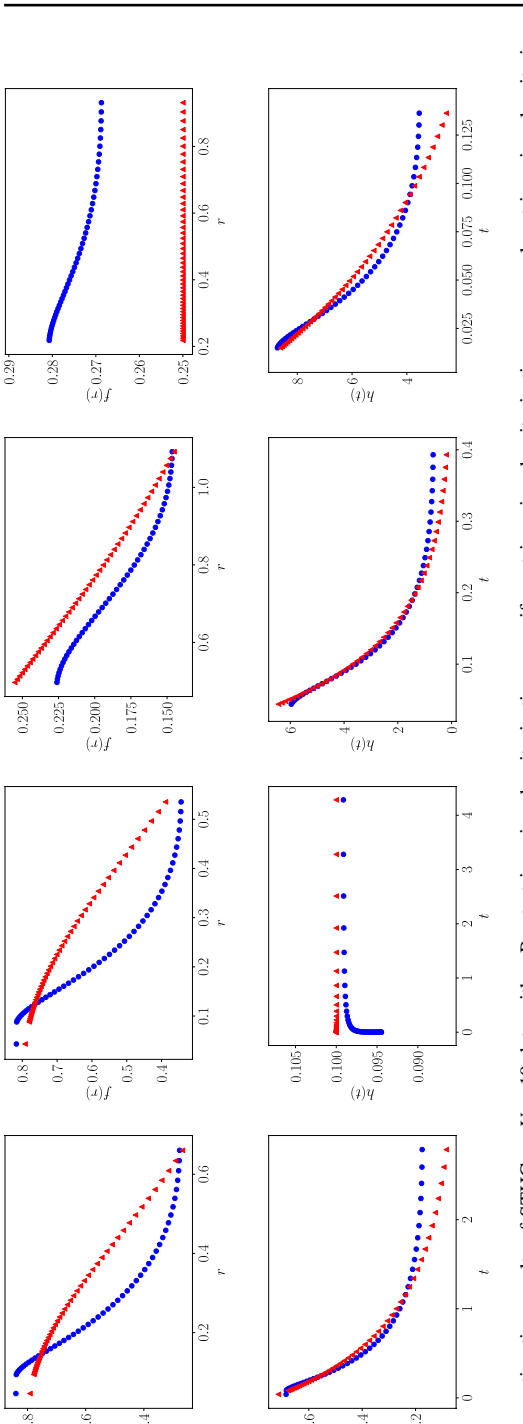


Fig. 9 The estimation results of STHC on $U = 10$ data with a Pareto triggering density in time, a uniform triggering density in time, a power-law triggering density in space and a uniform triggering density in space (from left to right). (Top) Ground-truth spatial triggering density $f(r)$ as red triangles and estimated triggering density as blue circles. (Bottom) Temporal triggering density $h(t)$ as red triangles and estimated triggering density as blue circles

$U = 10$ data with a uniform triggering density in space

Given the same parameters as above, we change the spatial density to $f(r) = 0.25$ and keep the exponential triggering density in time with $\omega = 10$. The regularization parameter $\alpha = 0.36$. We threshold the estimated $\tilde{\boldsymbol{K}}$ with $\epsilon = 0.01$ to remove noise (Fig. 9).

Appendix 2: Gowalla and Brightkite data sets

In this section, we describe the preprocessing procedure for Gowalla and Brightkite data sets. We focus on various local friendship subnetworks within different US cities, including San Diego (SD), Chicago (CHI), Los Angeles (LA) and San Francisco (SF). They have diverse network sizes and ST patterns within the same time period.

Brightkite-SD

We study check-ins in SD for Brightkite data set. We use a bounding box (with a north latitude of 33.1142, a south latitude of 32.5348, an east longitude of -116.9058, and a west longitude of -117.2824)[1] to locate check-ins in SD. We consider "active" users, who have more than 300 check-ins during the period. This gives us a small subnetwork with 25 "active" users and a total of 13,760 check-ins in SD.

Gowalla-CHI

We apply the same procedure as in "Appendix 2" on the Gowalla check-in data for CHI. The bounding box for CHI has a north latitude of 42.0229, a south latitude of 41.6446, an east longitude of -87.5245 and a west longitude of -87.9395. After selecting only active users (with more than 100 check-ins) users, we have a medium-sized subnetwork with 96 users and 27,326 check-ins.

Brightkite-LA

We apply the same procedure as in "Appendix 2" on the Brightkite check-in data in LA. The bounding box for LA has a north latitude of 34.34, a south latitude of 33.70, an east longitude of -118.16 and a west longitude of -118.67. After selecting only active users (with more than 150 check-ins) users, we have a medium-sized subnetwork with 168 users and 89,127 check-ins.

[1] We obtain latitude and longitude coordinates from https://www.flickr.com/places/info.

Gowalla-SF

We apply the same procedure as in "Appendix 2" on the Gowalla check-in data in SF. The bounding box for SF has a north latitude of 37.93, a south latitude of 37.64, an east longitude of − 122.28 and a west longitude of − 123.17. After selecting only active users (with more than 65 check-ins) users, we have a large subnetwork with 515 users and 102,673 check-ins.

Appendix 3: Assumptions for Theorem 1

There are two separate sets of general assumptions for the consistency of GMM and MLE in Hawkes processes. We only list assumptions that are relevant to our proof.

The first set of assumptions is from Ogata (1978) about the point process and intensity functions.

Assumption 1 (Consistency of MLE estimation)

- *Multivariate Hawkes process* $(N_{t,x,y})$ *is stationary, ergodic and absolutely continuous with respect to the standard Poisson process.*
- *The conditional intensity function* λ_Θ *with parameters* Θ *is predictable for all compact metric spaces and continuous in* Θ.
- *When* $t = 0$, λ_Θ*is positive almost surely and* $\lambda_{\Theta_1} = \lambda_{\Theta_2}$*almost surely if and only if* $\Theta_1 = \Theta_2$; *for any* Θ *from a compact metric space, there exists a neighborhood* $U(\Theta)$ *of* Θ *such that for all* $\Theta' \in U(\Theta)$, $|\lambda_{\Theta'}|$ *and* $|\log \lambda_{\Theta'}|$ *are bounded by random variables with finite second moments*.
- *For any* Θ *from a compact metric space, there is a neighborhood* $U(\Theta)$ *of* Θ *such that* $\sup_{\Theta' \in U(\Theta)} |\lambda(\Theta') - \mathbb{E}(\lambda(\Theta'))| \to 0$ *in probability as* $t \to \infty$ *and (for some* $\alpha > 0$) $\sup_{\Theta' \in U(\Theta)} |\log \mathbb{E}(\lambda(\Theta'))|$ *has finite* $(2+\alpha)$th *moment uniform bounded with respect to* t.

On top of Assumption 1, we also need GMM-related assumptions from Achab et al. (2017).

Assumption 2 (Consistency of GMM estimation)

- *For* (25), *the GMM approximation error* $L(\boldsymbol{R}) = 0$ *if and only if* $\boldsymbol{R} = (\boldsymbol{I} - \boldsymbol{K}^{\mathbf{T}})^{-1}$.
- *For* (22–24), *the supports of the triggering density* X, Y, H *satisfy* $\tilde{X}^2/X, \tilde{Y}^2/Y, \tilde{H}^2/T \to 0$ *separately as* $X, Y, H \to \infty$.

References

Achab, M., Bacry, E., Gaïffas, S., Mastromatteo, I., Muzy, J.-F. (2017). Uncovering causality from multivariate Hawkes integrated cumulants. *The Journal of Machine Learning Research*, *18*(1), 6998–7025.

Bacry, E., Bompaire, M., Gaïffas, S., Poulsen, S. (2017). Tick: A python library for statistical learning, with a particular emphasis on time-dependent modelling. arXiv preprint arXiv:1707.03003.

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

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

Balderama, E., Schoenberg, F. P., Murray, E., Rundel, P. W. (2012). Application of branching models in the study of invasive species. *Journal of the American Statistical Association*, *107*(498), 467–476.

Bao, J., Zheng, Y., Mokbel, M. F. (2012). Location-based and preference-aware recommendation using sparse geo-social networking data. In *Proceedings of the 20th international conference on advances in geographic information systems* (pp. 199–208).

Brantingham, P. J., Yuan, B., Herz, D. (2020a). Is gang violent crime more contagious than non-gang violent crime? *Journal of Quantitative Criminology*, https://doi.org/10.1007/s10940-020-09479-1.

Brantingham, P. J., Yuan, B., Sundback, N., Schoenberg, F. P., Bertozzi, A. L., Gordon, J., et al. (2020b). Does violence interruption work? *UCLA preprint*, www.stat.ucla.edu/~frederic/papers/brantingham2.pdf.

Brillinger, D. R., Guttorp, P. M., Schoenberg, F. P., El-Shaarawi, A. H., Piegorsch, W. W. (2002). Point processes, temporal. *Encyclopedia of Environmetrics*, *3*, 1577–1581.

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., Vandergheynst, P. (2017). Geometric deep learning: Going beyond Euclidean data. *IEEE Signal Processing Magazine*, *34*(4), 18–42.

Chen, S., Shojaie, A., Shea-Brown, E., Witten, D. (2017). The multivariate hawkes process in high dimensions: Beyond mutual excitation. arXiv preprint arXiv:1707.04928.

Chiang, W.-H., Yuan, B., Li, H., Wang, B., Bertozzi, A., Carter, J., Ray, B., Mohler, G. (2019). Sos-EW: System for overdose spike early warning using drug mover's distance-based Hawkes processes. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 538–554). Berlin: Springer.

Cho, E., Myers, S. A., Leskovec, J. (2011). Friendship and mobility: User movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1082–1090). ACM.

Daley, D. J., Vere-Jones, D. (2003). *An introduction to the theory of point processes*: *Volume I*: *Probability and its Applications*. New York: Springer.

Daley, D. J., Vere-Jones, D. (2007). *An introduction to the theory of point processes*: *Volume II*: *General theory and structure*. New York: Springer.

Du, N., Farajtabar, M., Ahmed, A., Smola, A. J., Song, L. (2015). Dirichlet–Hawkes processes with applications to clustering continuous-time document streams. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 219–228). ACM.

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

Eichler, M., Dahlhaus, R., Dueck, J. (2017). Graphical modeling for multivariate Hawkes processes with nonparametric link functions. *Journal of Time Series Analysis*, *38*(2), 225–242.

Farajtabar, M., Wang, Y., Rodriguez, M. G., Li, S., Zha, H., Song, L. (2015). Coevolve: A joint point process model for information diffusion and network co-evolution. *Advances in Neural Information Processing Systems*, 1954–1962.

Fox, E. W., Short, M. B., Schoenberg, F. P., Coronges, K. D., Bertozzi, A. L. (2016). Modeling e-mail networks and inferring leadership using self-exciting point processes. *Journal of the American Statistical Association*, *111*(514), 564–584.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*: *Journal of the Econometric Society*, *37*, 424–438.

Hall, E. C., Willett, R. M. (2016). Tracking dynamic point processes on networks. *IEEE Transactions on Information Theory*, *62*(7), 4327–4346.

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

Kaipio, J., Somersalo, E. (2006). *Statistical and computational inverse problems*, Vol. 160. New York: Springer.

Kingma, D. P., Ba, J. (2015). Adam: A method for stochastic optimization. In *International conference on learning representations*.

Lai, E. L., Moyer, D., Yuan, B., Fox, E., Hunter, B., Bertozzi, A. L., Brantingham, P. J. (2016). Topic time series analysis of microblogs. *IMA Journal of Applied Mathematics*, *81*(3), 409–431.

Lee, D. D., Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), p. 788.
Lewis, E., Mohler, G. (2011). A nonparametric EM algorithm for multiscale Hawkes processes. *Journal of Nonparametric Statistics*, *1*(1), 1–20.
Linderman, S., Adams, R. (2014). Discovering latent network structure in point process data. In *International conference on machine learning* (pp. 1413–1421). Beijing, China: JMLR: W&C.
Malinverno, A. (2002). Parsimonious Bayesian Markov chain Monte Carlo inversion in a nonlinear geophysical problem. *Geophysical Journal International*, *151*(3), 675–688.
Mark, B., Raskutti, G., Willett, R. (2018). Network estimation from point process data. *IEEE Transactions on Information Theory*, *65*, 2953–2975.
Marsan, D., Lengline, O. (2008). Extending earthquakes' reach through cascading. *Science*, *319*(5866), 1076–1079.
Mohler, G. O. (2014). Marked point process hotspot maps for homicide and gun crime prediction in Chicago. *International Journal of Forecasting*, *30*(3), 491–497.
Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., Tita, G. E. (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, *106*(493), 100–108.
Neumaier, A. (1998). Solving ill-conditioned and singular linear systems: A tutorial on regularization. *SIAM Review*, *40*(3), 636–666.
Ogata, Y. (1978). The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Annals of the Institute of Statistical Mathematics*, *30*(1), 243–261.
Ogata, Y. (1998). Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, *50*(2), 379–402.
Porter, M. D., White, G., et al. (2012). Self-exciting hurdle models for terrorist activity. *The Annals of Applied Statistics*, *6*(1), 106–124.
Reinhart, A. (2018). A review of self-exciting spatio-temporal point processes and their applications. *Statistical Science*, *33*(3), 299–318.
Schoenberg, F. P. (2006). On non-simple marked point processes. *Annals of the Institute of Statistical Mathematics*, *58*(2), 223–233.
Schoenberg, F. P. (2013). Facilitated estimation of ETAS. *Bulletin of the seismological Society of America*, *103*(1), 601–605.
Schoenberg, F. P., Brillinger, D. R., Guttorp, P. (2013). Point processes, spatial-temporal. *Encyclopedia of Environmetrics*, *4*, 1573–1578.
Schoenberg, F. P., et al. (2018a). Comment on "A review of self-exciting spatio-temporal point processes and their applications" by Alex Reinhart. *Statistical Science*, *33*(3), 325–326.
Schoenberg, F. P., Gordon, J. S., Harrigan, R. J. (2018b). Analytic computation of nonparametric Marsan–Lengliné estimates for Hawkes point processes. *Journal of Nonparametric Statistics*, *30*(3), 742–775.
Veen, A., Schoenberg, F. P. (2008). Estimation of space-time branching process models in seismology using an EM-type algorithm. *Journal of the American Statistical Association*, *103*(482), 614–624.
Wang, B., Luo, X., Zhang, F., Yuan, B., Bertozzi, A. L., Brantingham, P. J. (2018). Graph-based deep modeling and real time forecasting of sparse spatio-temporal data. arXiv preprint arXiv:1804.00684.
Yuan, B., Li, H., Bertozzi, A. L., Brantingham, P. J., Porter, M. A. (2019). Multivariate spatiotemporal Hawkes processes and network reconstruction. *SIAM Journal on Mathematics of Data Science*, *1*(2), 356–382.
Yuan, B., Wang, X., Ma, J., Zhou, C., Bertozzi, A. L., Yang, H. (2020). Variational autoencoders for highly multivariate spatial point processes intensities. In *International conference on learning* (representations).
Zhu, S., Xie, Y. (2019). Spatial–temporal–textual point processes with applications in crime linkage detection. arXiv preprint arXiv:1902.00440.
Zhuang, J., Ogata, Y., Vere-Jones, D. (2002). Stochastic declustering of space-time earthquake occurrences. *Journal of the American Statistical Association*, *97*(458), 369–380.