Tensor Kernel Recovery for Discrete Spatio-Temporal Hawkes Processes

Heejune Sheen, Xiaonan Zhu, and Yao Xie, *Member, IEEE*

***Abstract*—We introduce a new discrete spatio-temporal Hawkes process model by formulating the general influence of the Hawkes process as a tensor kernel. Based on the low-rank structure assumption of the tensor kernel, we cast the estimation of the tensor kernel as a convex optimization problem using the Fourier transformed nuclear norm. We provide theoretical performance guarantees for our approach and present an algorithm to solve the optimization problem. In particular, our upper bound of squared estimation error has the convergence rate of $O(lnK/\sqrt{K})$, where K is the number of samples in the time horizon. The efficiency of our estimation is demonstrated with numerical simulations on synthetic data and the analysis of real-world data from Atlanta burglary incidents.**

***Index Terms*—Hawkes process, spatio-temporal data, low-rank tensor, transformed tensor nuclear norm, convex optimization.**

I. Introduction

HAWKES processes, a type of self- (and mutual) exciting point processes, have gained substantial attention in machine learning and statistics due to their wide applicability in capturing complex interactions of discrete events over space, time, and possible networks. Such problem arises from many applications such as seismology [27], criminology [36], finance [17], [28], genomics [29], and social network [24], [35]. One advantage of Hawkes process modeling is that interactions between the history and a current event can be represented in the structure easily, as the Hawkes process, in general, has an intensity function consisting of two parts, a baseline intensity, and a triggering effect.

A central problem in Hawkes process modeling is to estimating the triggering effect through the so-called influence functions, which capture how different locations interact with each other. Estimating the triggering effects with Hawkes process models has been conducted in several prior works [1], [16], [22], [23], [35]. Bacry et al. [1] and Zhou et al. [35] proposed a convex optimization approach with sparse and low-rank assumptions on the interaction adjacency matrix or tensor to estimate an influence in a social network. In particular, they assumed the triggering function as a form of a product between the interaction coefficients and the fixed kernel functions that decay exponentially with continuous time.

Low-dimensional structures are very common in high-dimensional data, such as low-rank matrices and low-rank tensors. A recent motivation for studying low-rank matrices is due to the matrix completion problem [6], [7], [8], [10], [25]. One of the popular approaches is convex relaxations with a matrix nuclear norm to estimate a low-rank matrix. There has been much effort in modeling with low-rank tensors by extending the results on low-rank matrices, including [2], [3], [5], [14], [15]. However, unlike matrices, the rank of a tensor is not uniquely defined, and it can have multiple ranks such as the CP rank [11], [18], Tucker rank [32], tubal rank, and multi-rank. The tubal and multi-rank for the low-rank kernel tensor were proposed by Kilmer & Martin [21] with the algebra for a tensor and the corresponding Fourier-transformed tensor nuclear norm (TNN).

We are interested in a low-rank structure in this work for the tensor kernel capturing the interaction, which can be viewed as a low-rank approximation to capture the dominant mode of the influence functions. In particular, we consider the tubal and multi-rank for the low-rank kernel tensor.

The main purpose of this paper is to propose a discrete Hawkes process model, which is derived from the spatio-temporal Hawkes process approximation. More precisely, spatio-temporal influence functions for the Hawkes process are first parameterized as a low-rank tensor kernel in our model. Then, an approach to estimate the tensor kernel is presented using maximum likelihood with constrained Fourier transformed nuclear norm on the tensor, which leads to a convex optimization problem. We also prove theoretical performance guarantees for the squared recovery error. It is shown that the squared recovery error of our model converges to 0 at the rate of $O(\ln K/\sqrt{K})$ as the number of samples in time horizon K increases to infinity. To solve the optimization problem, a computationally efficient algorithm is designed based on the alternating direction method of multipliers (ADMM). The computational efficiency of our estimation procedure is illustrated with numerical simulations of synthetic and real data. We emphasize that our approach is different from the previous works [1], [35] since the influence

Manuscript received 3 July 2022; revised 13 October 2022 and 22 November 2022; accepted 28 November 2022. Date of current version 28 December 2022. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yuxin Chen. The work of Yao Xie was supported by NSF CAREER under Grants CCF-1650913, CMMI-2015787, DMS-1938106, and DMS-1830210. *(Corresponding author: Heejune Sheen.)*

Heejune Sheen is with the Department of Statistics and Data Science, Yale University, New Haven, CT 06511 USA (e-mail: heejune.sheen@yale.edu).

Xiaonan Zhu is with the Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: xz8451@princeton.edu).

Yao Xie is with the H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: yao.xie@isye.gatech.edu).

Digital Object Identifier 10.1109/TSP.2022.3229642

function is considered as a tensor kernel in the discrete-time and discrete location set-up.

The rest of the paper is organized as follows. Section II presents our model and the problem setup. Section III contains the main theoretical performance upper bound. Section IV proposes an ADMM-based algorithm to solve the optimization problem. Section V contains the numerical study, and finally, Section VI concludes the paper. The proofs are delegated to the Appendix.

II. Model

A. Discrete Hawkes Processes

We first describe continuous spatio-temporal Hawkes processes to motivate our discrete model. Consider a spatio-temporal point process whose events occur at time $t \in [0, T]$, and the location $(x, y) \in \mathcal{A} \subset \mathbb{R}^2$. Define a counting process $N : \mathcal{A} \times [0, \infty) \to \mathbb{Z}_{>0}$, such that $N(B, C)$ is the number of events in the region $B \in \mathcal{B}$ and the time window $C \in \mathcal{C}$, where $\mathcal{B}$ and $\mathcal{C}$ are the Borel σ-algebras of $\mathcal{A}$ and $[0, \infty)$. Let $\mathcal{H}_t$ be the σ-algebra generated by history of the process N up to time t. The conditional intensity function of a point process is defined as

$$\lambda(x, y, t) := \lim_{\Delta_x, \Delta_y, \Delta_t \downarrow 0} \frac{\mathbb{E}(N([x, x+\Delta_x] \times [y, y+\Delta_y] \times [t, t+\Delta_t] | \mathcal{H}_t)}{\Delta_x \Delta_y \Delta_t}. \quad (1)$$

For Hawkes processes, we can define the conditional intensity function with the following form:

$$\lambda(x, y, t) = \mu(x, y) + \int_0^t \iint_B g(x - u_1, y - u_2, t - u_3) N(d(u_1 \times u_2) \times du_3), \quad (2)$$

where $\mu(x, y) \geq 0$ is the base intensity at location (x, y) and $g : \mathbb{R}^2 \times [0.\infty) \to \mathbb{R}_{\geq 0}$ is the kernel function.

Suppose that the event data lie in bounded region $[0, n_1\Delta_x] \times [0, n_2\Delta_y]$ and time $[0, K\Delta_t]$ for some $n_1, n_2, K \in \mathbb{Z}_{>0}$. To discretize the process in both space and time, we define "bin counts" over the discrete space

$$\{i, j : 1 \leq i \leq n_1, 1 \leq j \leq n_2\}$$

and time horizon

$$\{k : -p + 1 \leq k \leq K\} :$$
$$Z_{ijk} = N([(i-1)\Delta_x, i\Delta_x] \times [(j-1)\Delta_y, j\Delta_y] \times [(k-1)\Delta_t, k\Delta_t]). \quad (3)$$

Let $\Delta = \Delta_x \Delta_y \Delta_t$. For a given data preceding discrete time k, the expected bin counts can be approximated as follows:

$$\begin{aligned}
&\mathbb{E}[Z_{ijk} | \mathcal{H}_{k-1}] \approx \Delta \lambda((i-1)\Delta_x, (j-1)\Delta_y, (k-1)\Delta_t) \\
&\approx \Delta \mu((i-1)\Delta_x, (j-1)\Delta_y) \\
&\quad + \Delta \sum_{k'=k-p}^{k-1} \sum_{i'=1}^{n_1} \sum_{j'=1}^{n_2} g\left((i-i')\Delta_x, (j-j')\right. \\
&\quad \left. \times \Delta_y, (k-k')\Delta_t\right) Z_{i'j'k'} \\
&:= \Delta \left(\mu_{ij} + \sum_{k'=k-p}^{k-1} \sum_{i'=1}^{n_1} \sum_{j'=1}^{n_2} \mathcal{G}_{i-i'+n_1, j-j'+n_2, k-k'} Z_{i'j'k'} \right),
\end{aligned}$$

where $\mu \in \mathbb{R}^{n_1 \times n_2}$ and $\mathcal{G} \in \mathbb{R}^{(2n_1-1) \times (2n_2-1) \times p}$ are discretized versions of the base intensity $\mu(x, y)$ and the kernel $g(x, y, t)$, respectively. For the first approximation above, $\mathbb{E}[Z_{ijk} | \mathcal{H}_{k-1}]$ is approximated using (1) with small Δ. For the second approximation, (2) and (3) are used to derive the discrete form.

It is commonly assumed in literature [1], [35] that g has the following form:

$$g(x, y, t) = h(x, y) f(t), \quad (4)$$

where f is a monotonically decreasing non-negative function, and $f(t)$ goes to zero for large t. An example of $f(t)$ is the class of exponential kernels, $f(t) = \alpha e^{-\alpha t}$. For our model, we relax these assumptions so that the kernel function does not need to follow the form (4) and g is not necessarily monotonically decreasing non-negative in time t. The history data with memory depth p is instead exploited to approximate the expected current bin counts. Thus, our model can be applied to more general cases.

Now, we propose a discrete spatio-temporal Hawkes process model with the conditional intensity function defined as follows:

$$\begin{aligned}
\lambda_{ijk}(\mu, \mathcal{G}) &:= \lambda((i-1)\Delta_x, (j-1)\Delta_y, (k-1)\Delta_t) \\
&= \mu_{ij} + \sum_{k'=k-p}^{k-1} \sum_{i'=1}^{n_1} \sum_{j'=1}^{n_2} \mathcal{G}_{i-i'+n_1, j-j'+n_2, k-k'} Z_{i'j'k'},
\end{aligned}$$

for $1 \leq i \leq n_1, 1 \leq j \leq n_2$, and $1 \leq k \leq K$.

Our model is derived from the spatio-temporal Hawkes process. The structure of our model is different from that of Kirchner's approximation [22] of the temporal Hawkes process with the model INAR(p). Thus, the proposed model has various advantages over Kirchner's model when analyzing spatio-temporal data. A more complex setting with higher dimensions (two dimensions in location and one in time) is dealt with in our model, and the location space and time-space are simultaneously discretized with tensor $\mathcal{G}$. Moreover, our interpretation of the discretized version of the kernel function enables the presence of space-time interactions. With constraints imposed on the rank of the tensor and entry-wise bounds on the estimators, a better estimation of the Hawkes process can be obtained when its coefficients are low-rank. Consequently, a convex optimization problem is constructed based on the likelihood function and our corresponding regularization, while [22] employed the projection method on the approximated time series model INAR(p).

To estimate the base intensity matrix μ and the underlying tensor $\mathcal{G}$, the followings are assumed: First, we assume that each entry of μ and $\mathcal{G}$ has the upper and the lower bound, i.e., there exist non-negative constants a_1, b_1, a_2, b_2 such that $a_1 \leq \mu_{ij} \leq$

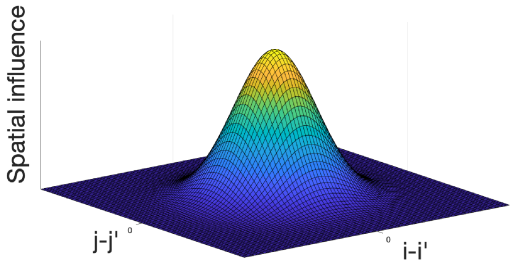


Fig. 1. For any given (i, j) grid, it displays the case where the spatial influence from (i, j) to $(i,' j')$ follows a standard gaussian function.

$b_1, a_2 \leq \mathcal{G}_{ijk} \leq b_2$ and $a_1 + a_2 > 0$. This assumption ensures that our problem is well-posed.

Second, we also assume that the tensor $\mathcal{G}$ has a low multi-rank $(r_1, \ldots, r_p)$, where $r_k = \operatorname{rank}(\widetilde{\mathcal{G}}^{(k)})$ and $\mathcal{G}^{(k)}$ is the kth frontal slice of transformed tensor $\widetilde{\mathcal{G}}$ (i.e., $\mathcal{G}^{(k)} := \mathcal{G}(:, :, k)$ using MATLAB notation). The transformed tensor $\widetilde{\mathcal{G}}$ is obtained by applying the Discrete Fourier Transform (DFT) to the mode-3 fibers of $\mathcal{G}$ (Lemma 1). In other words, a small sum of multi-rank $\gamma := \sum_{i=1}^{p} r_k$ is assumed. This assumption is based on the high correlations that exist within locations and time. For instance, for any given grid (i, j), suppose that the spatial influence from (i, j) to $(i,' j')$ is proportional to a standard Gaussian function (i.e., $\propto e^{-((i-i')^2+(j-j')^2)/2}$) as illustrated in Fig. 1, and the temporal influence follows a decreasing function in time. Then, the tensor $\mathcal{G} \in \mathbb{R}^{(2n_1-1)\times(2n_2-1)\times p}$ in our model has a low multi-rank at most $(1, \ldots, 1)$ and a small sum of the multi-rank less or equal to p.

Our problem considers a transformed multi-rank of a kernel tensor and a TNN over other tensor ranks and norms. The rank of a tensor can be defined in several ways, for instance, the CP rank, the Tucker rank, and the multi-rank. Depending on the tensor rank, the corresponding tensor nuclear norms should be utilized. It is known that the computation of the CP rank is NP-hard [19] and its relaxation is intractable in general. For this reason, the tractable Tucker rank and its relaxation are usually used. One of the popular relaxations is the sum of the matrix nuclear norms of matrices obtained by unfolding a tensor [26]. It is, however, not the tightest convex relaxation of the Tucker rank [30] and the matrix norm may be inefficient when the unfolding matrices have a significant difference in the number of rows and columns. We use TNN since it is the convex envelope of the multi-rank and can successfully interpret the low-rank structure of a kernel tensor.

B. Low-Rank Tensor Regularization

In this section, we review the tensor nuclear norm (TNN), which is used to guarantee the low-rankness of $\mathcal{G}$. We begin by introducing some notation. For matrix X, let $\|X\|$ be the matrix spectral norm, $\|X\|_*$ the matrix nuclear norm, and $\|X\|_F = (\sum_{i,j} X_{ij}^2)^{1/2}$ the Frobenius norm. For 3-way tensor $\mathcal{G} \in \mathbb{R}^{N_1\times N_2\times N_3}$, MATLAB notation is used to denote k-th horizontal, lateral, and frontal slice by $\mathcal{G}(k, :, :)$, $\mathcal{G}(:, k, :)$, and $\mathcal{G}(:, :, k)$ respectively. Specifically, the k-th frontal slice of $\mathcal{G}$ is denoted by $\mathcal{G}^{(k)} := \mathcal{G}(:, :, k)$ for $k = 1, \ldots, N_3$. The k-th mode-3 fiber of a 3-way tensor is defined by holding the first two indices fixed and varying the third, and denoted by $\mathcal{G}(k, k, :)$. The norm of tensor is defined as $\|\mathcal{G}\|_1 = \sum_{i,j,k} |\mathcal{G}_{ijk}|$, $\|\mathcal{G}\|_F = (\sum_{i,j,k} \mathcal{G}_{ijk}^2)^{1/2}$. The tensor spectral norm $\|\mathcal{G}\|_{\text{spec}}$ is defined later in Definition 1.

We introduce the following operators for the tensor algebra: the block circulation, the block diagonalization, and the fold and unfold command of tensor $\mathcal{G}$.

$$\mathsf{bcirc}(\mathcal{G}) = \begin{pmatrix} \mathcal{G}^{(1)} & \mathcal{G}^{(N_3)} & \mathcal{G}^{(N_3-1)} & \cdots & \mathcal{G}^{(2)} \\ \mathcal{G}^{(2)} & \mathcal{G}^{(1)} & \mathcal{G}^{(N_3)} & \cdots & \mathcal{G}^{(3)} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathcal{G}^{(N_3)} & \mathcal{G}^{(N_3-1)} & \mathcal{G}^{(N_3-2)} & \cdots & \mathcal{G}^{(1)} \end{pmatrix},$$

$$\mathsf{blockdiag}(\mathcal{G}) = \begin{pmatrix} \mathcal{G}^{(1)} & & & \\ & \mathcal{G}^{(2)} & & \\ & & \ddots & \\ & & & \mathcal{G}^{(N_3)} \end{pmatrix},$$

$$\mathsf{unfold}(\mathcal{G}) = \begin{pmatrix} \mathcal{G}^{(1)} \\ \mathcal{G}^{(2)} \\ \vdots \\ \mathcal{G}^{(N_3)} \end{pmatrix}, \quad \text{and} \quad \mathsf{fold}(\mathsf{unfold}(\mathcal{G})) = \mathcal{G}.$$

For two tensors $\mathcal{G}_1 \in \mathbb{R}^{N_1\times N_2\times N_3}$ and $\mathcal{G}_2 \in \mathbb{R}^{N_2\times N_4\times N_3}$, the *t-product* is defined as

$$\mathcal{G}_1 * \mathcal{G}_2 = \mathsf{fold}(\mathsf{bcirc}(\mathcal{G}_1)\mathsf{unfold}(\mathcal{G}_2)) \in \mathbb{R}^{N_1\times N_4\times N_3}.$$

Note that Kilmer and Martin [21] proposed a singular value decomposition (SVD) method for three-way tensors, and based on the tensor SVD, TNN is proposed by Semerci et al. [31]. We first review some background materials on the tensor SVD to introduce TNN. See [21] for more information. For a tensor $\mathcal{G} \in \mathbb{R}^{N_1\times N_2\times N_3}$, the block diagonalization property of block circulant matrices is described in the following lemma.

Lemma 1: [21] For $\mathcal{G} \in \mathbb{R}^{N_1\times N_2\times N_3}$, $\mathsf{bcirc}(\mathcal{G}) \in \mathbb{R}^{N_3N_1\times N_3N_2}$, we have

$$(F_{N_3} \otimes I_{N_1})\mathsf{bcirc}(\mathcal{G})(F_{N_3}^* \otimes I_{N_2}) = \mathsf{blockdiag}(\widetilde{\mathcal{G}}),$$

where $\otimes$ is the Kronecker product, I_{N_1} and I_{N_2} are identity matrices in $\mathbb{R}^{N_1\times N_1}$ and $\mathbb{R}^{N_2\times N_2}$, respectively, $F_{N_3} \in \mathbb{R}^{N_3\times N_3}$ is the normalized DFT matrix, which is unitary, and F^* denotes its conjugate transpose. The matrix $\mathsf{blockdiag}(\widetilde{\mathcal{G}})$ is the transformation of $\mathsf{bcirc}(\mathcal{G})$ into the Fourier domain, and the tensor $\widetilde{\mathcal{G}}$ is obtained by performing the DFT to the mode-3 fibers of $\mathcal{G}$ as mentioned earlier.

Based on the matrix SVD, we have

$$\mathsf{blockdiag}(\widetilde{\mathcal{G}}) = \mathsf{blockdiag}(\widetilde{\mathcal{U}})\mathsf{blockdiag}(\widetilde{\mathcal{S}})\mathsf{blockdiag}(\widetilde{\mathcal{V}}), \tag{5}$$

in the Fourier domain, where $\widetilde{\mathcal{G}}^{(k)} = \widetilde{\mathcal{U}}^{(k)}\widetilde{\mathcal{S}}^{(k)}(\widetilde{\mathcal{V}}^{(k)})^\top$ is the SVD. The equivalent decomposition of three-way tensors to (5) is characterized as a tensor-SVD [33]. The tensor-SVD for three-way tensors is described as follows.

Theorem 2 [21]: Any tensor $\mathcal{G} \in \mathbb{R}^{N_1 \times N_2 \times N_3}$ can be factored as

$$\mathcal{G} = \mathcal{U} * \mathcal{S} * \mathcal{V}^\top = \sum_{i=1}^{N_1 \wedge N_2} \mathcal{U}(:,i,:) * \mathcal{S}(i,i,:) * \mathcal{V}(:,i,:)^\top,$$

Combining Theorem 2 and (5), TNN can be defined as follows.

Definition 1 (Theorem 2.4.1 in [34]): The tensor nuclear norm (TNN) of $\mathcal{G}$ is defined as the sum of the singular values of all the frontal slices of $\widetilde{\mathcal{G}}$:

$$\|\mathcal{G}\|_{\mathrm{TNN}} = \sum_{i=1}^{N_1 \wedge N_2} \sum_{j=1}^{N_3} \widetilde{S}_{i,i,j}.$$

Note that the dual norm of the tensor nuclear norm is the tensor spectral norm $\|\mathcal{G}\|_{\mathrm{spec}} := \|\mathsf{bcirc}(\mathcal{G})\|$.

C. Problem Formulation

We use the notation:

$$Z^t := \{Z_{ijk}, 1 \le i \le n_1, 1 \le j \le n_2, -p+1 \le k \le t\},$$

$$Z_q^t := \{Z_{ijk}, 1 \le i \le n_1, 1 \le j \le n_2, q \le k \le t\}.$$

Assume that the discrete data follow the Poisson distribution:

$$Z_{ijk} | \mathcal{H}_{k-1} \sim \mathrm{Poisson}\left(\Delta \lambda_{ijk}\right).$$

Note that for fixed k, Z_{ijk} are conditionally independent for all i, j given $\mathcal{H}_{k-1}$, and λ_{ijk} depends only on the history of data before k, not on the data at time k. We also mention that our analysis can be applied to the alternative, the Bernoulli distribution assumption with slight modifications. Our goal is to estimate the true parameters μ and $\mathcal{G}$ of the discrete Hawkes process model. By our assumptions on the low-rank tensor $\mathcal{G}$, we have

$$\begin{aligned}\|\mathcal{G}\|_{\mathrm{TNN}} &= \|\mathsf{bcric}(\mathcal{G})\|_* = \|\mathsf{blockdiag}(\widetilde{\mathcal{G}})\|_* \\ &\le \sqrt{\gamma}\|\mathsf{blockdiag}(\widetilde{\mathcal{G}})\|_F = \sqrt{\gamma}\|\mathsf{blockdiag}(\mathcal{G})\|_F \\ &\le b_2\sqrt{\gamma(2n_1-1)(2n_2-1)p}.\end{aligned}$$

Accordingly, the candidate set $\mathcal{D}$ for the true value $(\mu, \mathcal{G})$ is defined as:

$$\begin{aligned}\mathcal{D} := \{(\mu, \mathcal{G}) |\ & a_1 \le \mu_{ij} \le b_1,\ a_2 \le \mathcal{G}_{ijk} \le b_2, \\ & \|\mathcal{G}\|_{\mathrm{TNN}} \le b_2\sqrt{\gamma(2n_1-1)(2n_2-1)p}\ \}.\end{aligned}$$

We consider a formulation by maximizing the log-likelihood function of the optimization variable μ and $\mathcal{G}$ given the observations Z^K. The negative log-likelihood function is given by

$$F(\mu, \mathcal{G}) := \sum_{k=1}^{K}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}(\Delta\lambda_{ijk}(\mu,\mathcal{G}) - Z_{ijk}\ln(\Delta\lambda_{ijk}(\mu,\mathcal{G}))).$$

Therefore, the estimators $(\widehat{\mu}, \widehat{\mathcal{G}})$ can be obtained by solving the following convex optimization problem:

$$(\widehat{\mu}, \widehat{\mathcal{G}}) = \underset{(\mu,\mathcal{G})\in\mathcal{D}}{\arg\min}\, F(\mu, \mathcal{G}). \tag{6}$$

Remark 1: Note that the convex optimization problem (6) constrained in a candidate set $\mathcal{D}$ can also be formulated as a regularized maximum likelihood function problem. Indeed, there exists a constant $\tau \in \mathbb{R}$ such that problem (6) equals

$$(\widehat{\mu}, \widehat{\mathcal{G}}) = \arg\min\{F(\mu, \mathcal{G}) + \tau\|\mathcal{G}\|_{\mathrm{TNN}}\}$$

with the natural constraint upon the entries:

$$a_1 \le \mu_{ij} \le b_1,\ a_2 \le \mathcal{G}_{ijk} \le b_2,\ \forall\, i, j, k.$$

It follows from the duality theory in optimization [4]. We use the regularized form to derive the algorithm in Section IV.

III. Theoretical Guarantee

We present an upper bound for the sum of squared errors of the two estimators, which is defined by

$$R[(\mu, \mathcal{G})||(\widehat{\mu}, \widehat{\mathcal{G}})] := \|\mu - \widehat{\mu}\|_F^2 + \|\mathcal{G} - \widehat{\mathcal{G}}\|_F^2,$$

where $\widehat{\mu}$ and $\widehat{\mathcal{G}}$ are the optimal solutions to (6).

To state our theoretical guarantee, we define the condition number as in [20].

Definition 2: Given $X \in \mathbb{R}^{d\times d}, X \succeq 0$ and $p \in [1, \infty]$, the condition number is defined by

$$\delta_p[X] := \max\{\delta \ge 0 : g^\top X g \ge \delta\|g\|_p^2,\ \forall g \in \mathbb{R}^d\},$$

where $X \succeq 0$ ($X \succ 0$) denotes that X is a positive semidefinite matrix (a positive definite matrix, respectively). Note that $\delta_p[X] > 0$ when $X \succ 0$.

Now, we present our main theorem.

Theorem 3 (Estimation error driven by data): Assume that $(\mu, \mathcal{G}) \in \mathcal{D}$ and $(\widehat{\mu}, \widehat{\mathcal{G}})$ are the optimal solution to (6). Let

$$\underline{J} = a_1 + a_2 \min_k\{||Z_{k-p}^{k-1}||_1\},$$

$$\bar{J} = b_1 + b_2\sqrt{\gamma(2n_1-1)(2n_2-1)p}\max_{1\le k\le K}\{\|Z_{k-p}^{k-1}\|_{\mathrm{spec}}\},$$

and let $A[\cdot] : \mathbb{R}^{n_1\times n_2\times(K+p)} \to \mathbb{R}^{d\times d}$ be a mapping defined in Appendix A. Then, for every Z^K, $\alpha_1, \alpha_2 \in (0,1)$, it holds that

$$\begin{aligned}&R[(\mu, \mathcal{G})||(\widehat{\mu}, \widehat{\mathcal{G}})] \\ &\le \frac{16\sqrt{2}n_1n_2\bar{J}^2\ln(\bar{J}/\underline{J})}{\sqrt{K}(1-e^{-2\bar{J}})\Delta\delta_2[A[Z^K]]}\sqrt{\ln\frac{n_1n_2}{\alpha_2}} \\ &\qquad \cdot \max\left\{2\sqrt{\Delta\bar{J}\ln\frac{n_1n_2K}{\alpha_1}},\ 4\ln\frac{n_1n_2K}{\alpha_1}\right\}\end{aligned} \tag{7}$$

with probability at least $1 - 2\alpha_1 - 2\alpha_2$.

Remark 2: If $\delta_2[A[Z^K]] > 0$, for large n_1, n_2, p and K, there exists a constant $C > 0$ such that the following bound holds with high probability.

$$R[(\mu, \mathcal{G})||(\widehat{\mu}, \widehat{\mathcal{G}})] \le C\frac{n_1n_2\bar{J}^2\ln(\bar{J})\sqrt{\ln(n_1n_2)}\cdot\ln(n_1n_2K)}{\sqrt{K}}.$$

Remark 3: From Remark 2, we observe that, for given data Z^K, the upper bound can be regarded as an increasing function of $\bar{J}$. More precisely, the estimation error for the upper bound increases with the upper bound on the tensor nuclear norm $b_2[\gamma(2n_1-1)(2n_2-1)p]^{1/2}$. It implies that the upper bound

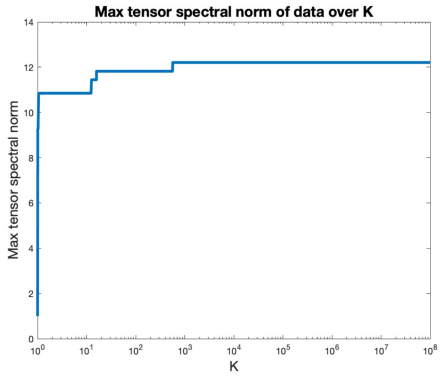


Fig. 2. The simulation of $\max_k \|Z_{k-p}^{k-1}\|_{\text{spec}}$.

of the estimation error will be small if we have a small sum of multi-rank γ. It is a characteristic that we can expect from the low-rank tensor recovery.

Remark 4: We observe that by fixing, n_1, n_2, Δ, a_1, a_2, b_1, b_2, and γ, the upper bound (7) tends to 0 as $K \to \infty$ at the rate of $O(\ln K/\sqrt{K})$. We experimentally show that $\max_{1\le k\le K}\{\|Z_{k-p}^{k-1}\|_{\text{spec}}\}$ is bounded above by O(lnK) in Fig. 2.

The proof for Theorem 3 is presented in Appendix A. In the proof, the Kullback-Leibler (KL) divergence and Hellinger distance are defined between two Poisson distributions. For any two Poisson mean p and q, the KL -divergence is defined as

$$D(p||q) := p\ln(p/q) - (p-q),$$

and the Hellinger distance as

$$H^2(p||q) := 2 - 2\exp\left(-\frac{1}{2}(\sqrt{p}-\sqrt{q})^2\right).$$

Then, a lower bound is derived for $\sum_{k=1}^{K}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} D(\lambda_{ijk}(\mu,\mathcal{G})||\lambda_{ijk}(\widehat{\mu},\widehat{\mathcal{G}}))$ with the Hellinger distance and Lemma 8 in [9]. Furthermore, we establish an upper bound on the sum of the KL divergence using the Azuma Hoeffding's inequality. We then obtain the upper bound for the estimation error by combining the lower and upper bound.

Corollary 1 immediately follows from Theorem 3. In particular, it demonstrates the data-driven upper bound for the sum of KL divergence between the estimated and the true intensity functions.

Corollary 1: Assume that $(\mu,\mathcal{G}) \in \mathcal{D}$ and $(\widehat{\mu},\widehat{\mathcal{G}})$ are the optimal solution to (6). With the notation defined in Theorem 3, for every Z^K, $\alpha_1, \alpha_2 \in (0,1)$, it holds that

$$\sum_{k=1}^{K}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\frac{D(\lambda_{ijk}(\mu,\mathcal{G})||\lambda_{ijk}(\widehat{\mu},\widehat{\mathcal{G}}))}{n_1 n_2 K}$$
$$\le \sqrt{\frac{2}{K}}\ln\frac{\bar{J}}{\underline{J}}\ln\frac{n_1 n_2}{\alpha_2}$$
$$\cdot \max\left\{2\sqrt{\Delta\bar{J}\ln\frac{n_1 n_2 K}{\alpha_1}},\ 4\ln\frac{n_1 n_2 K}{\alpha_1}\right\}$$

with probability at least $1 - 2\alpha_1 - 2\alpha_2$.

IV. Algorithm

For the proposed convex optimization problem (6), we apply ADMM and the majorization-minimization (MM) algorithms. Based on the ADM4 algorithm proposed by [35], we design our algorithm for problem (6). To start with, the constraint sets for μ and $\mathcal{G}$ are separated to the following two closed convex sets:

$$\Gamma_1 := \{\mu \mid a_1 \le \mu_{ij} \le b_1,\ \forall (i,j) \in [\![n_1]\!] \times [\![n_2]\!]\},$$
$$\Gamma_2 := \{\mathcal{G} \mid a_2 \le \mathcal{G}_{ijk} \le b_2,$$
$$\forall (i,j,k) \in [\![2n_1-1]\!] \times [\![2n_2-1]\!] \times [\![p]\!]\},$$

where $[\![n]\!] := \{1,2,\ldots,n\}$. Then, problem (6) can be written as

$$\begin{aligned} \min \quad & F(\mu,\mathcal{G}) + \tau\|\mathcal{G}\|_{\text{TNN}} \\ \text{subject to} \quad & \mu \in \Gamma_1,\ \mathcal{G} \in \Gamma_2. \end{aligned} \tag{8}$$

ADMM is employed to convert the above optimization problem to several sub-problems that are easier to solve. More specifically, the problem is separated into the first term of the objective function, the regularization term, and the constraints. To that end, three auxiliary variables, m, G and R are introduced, and (8) can be equivalently expressed as

$$\begin{aligned} \min \quad & F(\mu,\mathcal{G}) + \tau\|R\|_{\text{TNN}} \\ \text{subject to} \quad & m \in \Gamma_1,\ G \in \Gamma_2, \\ & \mu = m,\ \mathcal{G} = G,\ \mathcal{G} = R. \end{aligned} \tag{9}$$

We then define the following augmented Lagrangian function of (9):

$$\begin{aligned} &\mathcal{L}_\rho(\mathcal{G},\mu,R,G,m,Y_1,Y_2,Y_3) \\ &:= F(\mu,\mathcal{G}) + \tau\|R\|_{\text{TNN}} + \psi_{\Gamma_1}(m) + \psi_{\Gamma_2}(G) \\ &\quad + \rho\langle Y_1, \mathcal{G}-R\rangle + \rho\langle Y_2, \mathcal{G}-G\rangle + \rho\langle Y_3, \mu-m\rangle \\ &\quad + \frac{\rho}{2}\|\mathcal{G}-R\|_F^2 + \frac{\rho}{2}\|\mathcal{G}-G\|_F^2 + \frac{\rho}{2}\|\mu-m\|_F^2, \end{aligned}$$

where Y_1, Y_2, and Y_3 are the dual variables associated with the constraints $\mathcal{G} = R$, $\mathcal{G} = G$, and $\mu = m$, respectively. The constant ρ is a penalty parameter, and the functions $\psi_{\Gamma_1}(m)$ and $\psi_{\Gamma_2}(G)$ are defined as

$$\psi_{\Gamma_1}(m) := \begin{cases} 0 & \text{if } m \in \Gamma_1, \\ +\infty & \text{otherwise,} \end{cases}$$

$$\psi_{\Gamma_2}(G) := \begin{cases} 0 & \text{if } G \in \Gamma_2, \\ +\infty & \text{otherwise.} \end{cases}$$

Notice that two blocks of variables $(\mathcal{G}^{t+1},\mu^{t+1})$ and $(R^{t+1},m^{t+1},G^{t+1})$ are separable in the augmented Lagrangian function. Thus, ADMM can be applied as the following iterations.

$$(\mathcal{G}^{t+1},\mu^{t+1}) = \arg\min_{\mu,\mathcal{G}} \mathcal{L}_\rho(\mathcal{G},\mu,R^t,G^t,m^t,Y_1^t,Y_2^t,Y_3^t), \tag{10}$$

$$\begin{aligned} &(R^{t+1},m^{t+1},G^{t+1}) \\ &\quad = \arg\min_{R,m,G} \mathcal{L}_\rho(\mathcal{G}^{t+1},\mu^{t+1},R,G,m,Y_1^t,Y_2^t,Y_3^t), \end{aligned} \tag{11}$$

$$Y_1^{t+1} = Y_1^t + (\mathcal{G}^{t+1} - R^{t+1}),$$
$$Y_2^{t+1} = Y_2^t + (\mathcal{G}^{t+1} - G^{t+1}),$$
$$Y_3^{t+1} = Y_3^t + (\mu^{t+1} - m^{t+1}).$$

It has been shown in [13] that the above application of ADMM on the two-block convex minimization problem converges.

It remains to solve (10) and (11) respectively. We start with deriving the updating step for $\mathcal{G}$ and μ as follows. Note that for (10), the following optimization problem is considered:

$$\min_{\mu,\mathcal{G}} g(\mu,\mathcal{G}) := \sum_{k'=1}^{K}\sum_{i'=1}^{n_1}\sum_{j'=1}^{n_2}[\Delta\lambda_{i'j'k'}(\mu,\mathcal{G}) - Z_{i'j'k'}\ln(\Delta\lambda_{i'j'k'}(\mu,\mathcal{G}))] + \frac{\rho}{2}\|\mathcal{G} - R^t + Y_1^t\|_F^2 + \frac{\rho}{2}\|\mathcal{G} - G^t + Y_2^t\|_F^2 + \frac{\rho}{2}\|\mu - m^t + Y_3^t\|_F^2. \quad (12)$$

Since no closed-form solutions exist, we apply the MM algorithm as in [35]. For any $\mu, \mathcal{G}$, let $Q(\mu,\mathcal{G};\mu^{(q)},\mathcal{G}^{(q)})$ be a convex function such that

$$g(\mu,\mathcal{G}) \le Q(\mu,\mathcal{G};\mu^{(q)},\mathcal{G}^{(q)}) \quad (13)$$
$$g(\mu^{(q)},\mathcal{G}^{(q)}) = Q(\mu^{(q)},\mathcal{G}^{(q)};\mu^{(q)},\mathcal{G}^{(q)}), \quad (14)$$

where $\mu^{(q)}$ and $\mathcal{G}^{(q)}$ are estimates of μ and $\mathcal{G}$. Then, we can obtain the optimal solutions to convex problem (12) by using the iterative procedure:

$$\left(\mu^{(q+1)},\mathcal{G}^{(q+1)}\right) = \arg\min_{\mu,\mathcal{G}} Q(\mu,\mathcal{G};\mu^{(q)},\mathcal{G}^{(q)}).$$

Let

$$\Omega = \{k \mid Z_{ijk} \neq 0 \text{ for some } i \text{ and } j\}$$

and

$$l(k) = \{(i,j) \mid Z_{ijk} \neq 0\}.$$

Define $Q(\mu,\mathcal{G};\mu^{(q)},\mathcal{G}^{(q)})$ that satisfies (13) and (14) as follows:

$$\begin{aligned}
&Q(\mathcal{G},\mu;\mathcal{G}^{(q)},\mu^{(q)}) \\
&= -\sum_{k'\in\Omega}\sum_{(i',j')\in l(k')}\left[Z_{i'j'k'}\ln\Delta + Z_{i'j'k'}\left(p_{i'j'k'}\ln\frac{\mu_{i'j'}}{p_{i'j'k'}}\right.\right. \\
&\left.\left. + \sum_{k=k'-p}^{k'-1}\sum_{(i,j)\in l(k)} p_{ijk,i'j'k'}\ln\frac{\mathcal{G}_{i'-i+n_1,j'-j+n_2,k'-k}Z_{ijk}}{p_{ijk,i'j'k'}}\right)\right] \\
&+ \sum_{k'=1}^{K}\sum_{i'=1}^{n_1}\sum_{j'=1}^{n_2}\Delta\left(\sum_{k=k'-p}^{k'-1}\sum_{(i,j)\in l(k)}\mathcal{G}_{i'-i+n_1,j'-j+n_2,k'-k}Z_{ijk}\right) \\
&+ \frac{\rho}{2}\|\mathcal{G} - R^t + Y_1^t\|_F^2 + \frac{\rho}{2}\|\mathcal{G} - G^t + Y_2^t\|_F^2 \\
&+ \frac{\rho}{2}\|\mu - m^t + Y_3^t\|_F^2 + n_3\Delta\sum_{i'=1}^{n_1}\sum_{j'=1}^{n_2}\mu_{i'j'},
\end{aligned}$$

where

$$p_{i'j'k'} = \frac{\mu_{i'j'}^{(q)}}{\lambda_{i'j'k'}^{(q)}}, \text{ and } p_{ijk,i'j'k'} = \frac{\mathcal{G}_{i'-i+n_1,j'-j+n_2,k'-k}^{(q)}Z_{ijk}}{\lambda_{i'j'k'}^{(q)}}.$$

Let $a = i' - i + n_1$, $b = j' - j + n_2$, $c = k' - k$, and $l'(k) = l(k) \cap \{i' - i + n_1 = a, j' - j + n_2 = b, k' - k = c\}$. By taking derivative, a closed form solution to $Q(\mu,\mathcal{G};\mu^{(q)},\mathcal{G}^{(q)})$ is

$$\mu_{i'j'}^{(q+1)} = \frac{-B + \sqrt{B^2 - 4\rho C}}{2\rho}, \quad (15)$$

$$\mathcal{G}_{abc}^{(q+1)} = \frac{-U + \sqrt{U^2 - 8\rho V}}{4\rho}, \quad (16)$$

where

$$\begin{aligned}
B &= n_3\Delta + \rho\left[-(\mu^t)_{i'j'} + (Y_3^t)_{i'j'}\right], \\
C &= -\sum_{k'\in\Omega}\sum_{(i',j')\in l(k')} Z_{i'j'k'}p_{i'j'k'}, \\
U &= \sum_{k'=1}^{K}\sum_{i'=1}^{n_1}\sum_{j'=1}^{n_2}\left(\sum_{k=k'-p}^{k'-1}\sum_{(i,j)\in l'(k)}\Delta Z_{ijk}\right) \\
&\quad + \rho\left[-(R^t)_{abc} + (Y_1^t)_{abc} - (G^t)_{abc} + (Y_2^t)_{abc}\right], \\
V &= -\sum_{k'\in\Omega}\sum_{(i',j')\in l(k')}\sum_{k=k'-p}^{k'-1}\sum_{(i,j)\in l'(k)} Z_{i'j'k'}p_{ijk,i'j'k'}.
\end{aligned}$$

Recall that R, m, and G in the objective function of the problem (11) are separable. Hence they can be calculated one by one. We start with updating step R. The optimal solution to (11) regarding R is given by

$$\begin{aligned}
R^{t+1} &= \arg\min_R \tau\|R\|_{\text{TNN}} + \frac{\rho}{2}\|-R_1 + Y_1^t + \mathcal{G}^{t+1}\|_F^2 \\
&= \text{Prox}_{(\tau/\rho)\|\cdot\|_{\text{TNN}}}\left(Y_1^t + \mathcal{G}^{t+1}\right) \\
&= \mathcal{U} * \mathcal{S}_{\rho/\tau} * \mathcal{V}^\top,
\end{aligned}$$

where $\mathcal{U} * \mathcal{S} * \mathcal{V}^\top$ is a tensor singular value decomposition of $Y_1^t + \mathcal{G}^{t+1}$, $\mathcal{S}_{\rho/\tau} = \texttt{IFT}(\widehat{\mathcal{S}}_{\rho/\tau}, [\,], 3)$ for the third frontal slices, and $\widehat{\mathcal{S}}_{\rho/\tau} := \max\{\widehat{S} - \rho/\tau, 0\}$. Here, the operator IFT corresponds to an inverse Fourier transform.

For updating G, an optimal solution to problem (11) for G is obtained as

$$\begin{aligned}
G^{t+1} &= \arg\min_G \psi_{\Gamma_2}(G) + \frac{\rho}{2}\|G - (\mathcal{G}^{t+1} + Y_2^t)\|_F^2 \\
&= \mathcal{P}_{\Gamma_2}(\mathcal{G}^{t+1} + Y_2^t),
\end{aligned}$$

where $\mathcal{P}_{\Gamma_2}$ is a projection onto Γ_2. Similarly, for m, we obtain an optimal solution as follows.

$$\begin{aligned}
m^{t+1} &= \arg\min_m \psi_{\Gamma_1}(m) + \frac{\rho}{2}\|m - (\mu^{t+1} + Y_3^t)\|_F^2 \\
&= \mathcal{P}_{\Gamma_1}(\mu^{t+1} + Y_3^t),
\end{aligned}$$

where $\mathcal{P}_{\Gamma_1}$ is a projection onto Γ_1.

Finally, all the steps are summarized in Algorithm 1.

Algorithm 1: Algorithm for Solving (9).

Input: Given data $Z^K \in \mathbb{R}^{n_1 \times n_2 \times (K+p)}$, $\rho, \tau, a_1, a_2, b_1, b_2$

Output: Matrix $\widehat{\mu}$ and tensor $\widehat{\mathcal{G}}$

Initialize $\mu^{(0)}, \mathcal{G}^{(0)}, R^{(0)}, m^{(0)}, G^{(0)}, Y_1^{(0)}, Y_2^{(0)}, Y_3^{(0)}$, and set $t = 1$.

repeat

Update μ^{t+1} and $\mathcal{G}^{t+1}$ by the following steps:

while not converge **do**

Update $\mu, \mathcal{G}$ by the (15) and (16).

end while

Update R^{t+1}, G^{t+1}, m^{t+1} by solving (11)

Update

$$Y_1^{t+1} = Y_1^t + (\mathcal{G}^{t+1} - R^{t+1})$$
$$Y_2^{t+1} = Y_2^t + (\mathcal{G}^{t+1} - G^{t+1})$$
$$Y_3^{t+1} = Y_3^t + (\mu^{t+1} - m^{t+1})$$

$t = t + 1$.

until Termination criterion is met.

V. Numerical Examples

A. Synthetic Data

We first experiment with Algorithm 1 on synthetic data to see the performance of our method. We generate the true $\mathcal{G} \in \mathbb{R}^{(2n_1-1)\times(2n_2-1)\times p}$ with multi-rank $(r_1, r_2, \ldots, r_p) = (1, 1, \ldots, 1)$ by $\mathcal{G}_{ijk} = u_i^{(1)} u_j^{(2)} u_k^{(3)}$, where $u_i^{(1)}$, $u_j^{(2)}$, and $u_k^{(3)}$ are from uniform distribution $U(0, 1)$. By our discrete approximation to Hawkes processes with memory depth p, we use a non-increasing function of k for the kth frontal slice $\mathcal{G}^{(k)}$ for $k = 1, \ldots, p$. We also generate μ randomly from $U(0, 1)$. The μ and $\mathcal{G}$ are rescaled for a well-defined point process. With the true μ, $\mathcal{G}$, we generate the synthetic data by

$$Z_{ijk} | \mathcal{H}_{(k-1)} \sim \text{Poisson}\,(\Delta \lambda_{ijk})$$

for $i \in [\![n_1]\!], j \in [\![n_2]\!]$ and $k \in [\![K]\!]$ with given initial data Z_{1-p}^0. The initialization $\mu^{(0)}$ and $\mathcal{G}^{(0)}$ are randomly generated with similar scales to their true values. To ensure that the error terms in different cases are at the same scale, we use the relative error

$$\text{Merr} := \frac{\|\mu - \widehat{\mu}\|_F}{\|\mu\|_F} \text{ and } \text{Gerr} := \frac{\|\mathcal{G} - \widehat{\mathcal{G}}\|_F}{\|\mathcal{G}\|_F}$$

to evaluate the estimation of μ and $\mathcal{G}$, respectively.

We test the performance of our method with different n_1, n_2 and p, and compare it with the model without the low-rank constraint on $\mathcal{G}$, and a widely used Hawkes process model with an exponential temporal decay function (e.g., [35]). We denote the proposed method as "TNN," the maximum likelihood estimation method without the low-rank constraint as "MLE," the estimation method with the exponential decay function (i.e., $\alpha e^{-\alpha k}$, where $k = 1, \ldots p$ and $\alpha > 0$ is a decay parameter.) as "EXP," and the estimation method with the matrix nuclear norm and the exponential decay function as "MNN".

TABLE I
Evaluation of TNN, MLE, EXP and MNN With the Atlanta Crime Dataset

	FRQ (1)	FRQ (60)	NLR
TNN	0.3474	0.3677	4126.3
MLE	0.4456	0.3732	4197.3
EXP	0.4509	0.3790	4138.0
MNN	0.3823	0.3681	4129.2

The parameters for each method are chosen based on the dataset that is disjoint from the training dataset. For parameter ρ in Algorithm 1, it is set to 0.002 for TNN, 0.001 for MLE, 0.06 for EXP and MNN. Since ρ serves as a dual step-size in ADMM, the performance of methods is robust to the mild change of ρ. For hyper-parameter τ on the regularized terms, a cross-validation-like method is exploited to tune the parameter and it is set to 0.5 for TNN and 0.1 for MNN. Two cases of experiments are carried out, and the representative results are shown in Fig. 3. The results of each case are averaged over five runs. Table II in Appendix B shows the detailed estimation errors for each case.

In the experiments, we observe that both the relative error of the matrix estimation and the tensor estimation decrease as the sample number grows and becomes close to zero. We notice that the numerical results correspond to our theoretical result on the upper bound of the estimation error mentioned in Remark 4. Moreover, TNN outperforms MLE, EXP, and MNN. It shows the computational efficiency of implementing the Fourier-transformed TNN when the tensor kernel has a low-rank structure, which commonly occurs in the real world due to high correlations between locations and time. It can be observed that TNN has advantages over EXP and MNN for data with unknown or non-exponential temporal decay functions. Moreover, TNN can capture more general spatio-temporal correlations than MNN, where the kernel is decomposed into a matrix and a temporal decay function. We note that our model, TNN, can be easily applied to a wide range of discrete data.

B. Real Data

We next apply our method to real-world data, the crime dataset in Atlanta, USA. The dataset contains 47,245 burglary incidents in Atlanta from January 1, 2015, to February 28, 2017. The events in the region where the latitude is from 33.71 to 33.76 and the longitude are from -84.43 to -84.38 are considered. In the region, 9937 burglary incidents occurred during 789 days. We discretize the area into 5×5 discrete space and the time with a 4-hour interval unit.

We use $p = 5$ as a memory depth for the data, and the parameter τ is set to 3.5 for TNN and 0.4 for MNN. The model is trained and tested with 80% and 20% of the data sequence, respectively.

To evaluate our model, two metrics are employed: First, the metric FRQ, defined as the sum of the absolute difference in frequency of events between the predicted data and the true test data, is used. It is the frequency difference of burglary

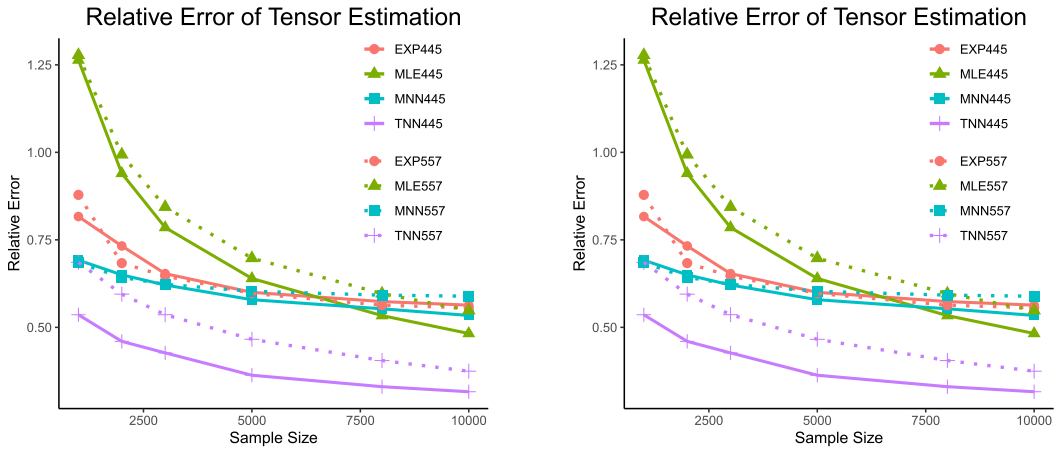


Fig. 3. The relative errors of the estimated tensor kernel (left) and base intensity matrix (right) under different n_1, n_2, p, and sample size K.

TABLE II
THE NUMERICAL ESTIMATION ERROR WITH DIFFERENT n_1, n_2, p AND K

Case	Method	Error	K=1000	K=2000	K=3000	K=5000	K=8000	K=10000
$n_1 = 4,\ n_2 = 4,\ p = 5$	TNN	Gerr	0.536	0.460	0.427	0.363	0.330	0.316
		Merr	0.202	0.131	0.094	0.080	0.064	0.061
	MLE	Gerr	1.264	0.940	0.785	0.640	0.534	0.482
		Merr	0.202	0.144	0.110	0.094	0.066	0.063
	EXP	Gerr	0.817	0.732	0.654	0.600	0.574	0.563
		Merr	0.296	0.174	0.133	0.112	0.103	0.109
	MNN	Gerr	0.692	0.650	0.621	0.597	0.553	0.534
		Merr	0.215	0.163	0.143	0.148	0.144	0.137
$n_1 = 5,\ n_2 = 5,\ p = 7$	TNN	Gerr	0.686	0.595	0.536	0.466	0.405	0.375
		Merr	0.241	0.178	0.192	0.133	0.116	0.105
	MLE	Gerr	1.278	0.994	0.844	0.698	0.596	0.548
		Merr	0.378	0.315	0.297	0.237	0.249	0.241
	EXP	Gerr	0.878	0.984	0.646	0.597	0.563	0.556
		Merr	0.386	0.456	0.385	0.312	0.267	0.239
	MNN	Gerr	0.684	0.643	0.623	0.603	0.592	0.589
		Merr	0.462	0.380	0.323	0.256	0.217	0.200

incidents in 25 subregions. Second, the metric NLR, the sum of the negative log-likelihood function, is compared.

Table I shows numerical results on TNN, MLE, EXP and MNN. The numbers in the columns of FRQ (1) and FRQ (60) represent one instance of FRQ and the average FRQ over 60 runs, respectively. In all metrics, TNN provides better results than MLE, EXP and MNN. We observe again the clear advantage of exploiting a low-rank structure of the tensor kernel. We note that TNN is implemented without any predefined decay parameter, whereas it is necessary for EXP and MNN.

VI. CONCLUSION

We have studied the recovery of the base intensity matrix and the tensor of the discretized version of the kernel function for spatio-temporal Hawkes processes. Using TNN, a formulation of the maximum likelihood estimation with the constraints has been proposed. Specifically, a precise theoretical upper bound for the sum of square errors of the proposed estimators has been presented. We have also applied the ADMM and MM algorithms to solve the proposed convex optimization problem. The numerical experiments demonstrate the efficiency of our method and support the theoretical results. For future work, non-convex optimization techniques will be investigated to estimate the matrix and the tensor kernel in the problem. It will be interesting to study whether the convex relaxation gap can be estimated and reduced by employing non-convex optimization methods.

APPENDIX A
PROOF OF THEOREM 3

We prove Theorem 3. For the simplicity of analysis, we let $\eta := (\mu, \mathcal{G})$. Then, the problem is expressed as:

$$\min F(\eta) := \sum_{k=1}^{K}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}[\Delta\lambda_{ijk}(\eta) - Z_{ijk}\ln(\Delta\lambda_{ijk}(\eta))]$$

$$\text{subject to}\ \ \eta \in \mathcal{D} := \{(\mu,\mathcal{G})|\ a_1 \le \mu_{ij} \le b_1,\ a_2 \le \mathcal{G}_{ijk} \le b_2,$$
$$\|\mathcal{G}\|_{\text{TNN}} \le b_2\sqrt{\gamma(2n_1-1)(2n_2-1)p}\ \Big\}, \qquad (17)$$

where $\lambda, Z \in \mathbb{R}^{n_1\times n_2\times K}$.

We now define the KL-divergence between two Poisson distributions. For any two Poisson mean p and q, the KL divergence

is defined as follows:

$$D(p||q) := p\ln(p/q) - (p-q).$$

Similarly, the Hellinger distance for Poisson distributions is defined:

$$H^2(p||q) := 2 - 2\exp\left\{-\frac{1}{2}(\sqrt{p}-\sqrt{q})^2\right\}.$$

Let θ be a true parameter that we aim to estimate. For any $\eta \in \mathcal{D}$, we have

$$\begin{aligned}
&F(\eta) - F(\theta)\\
&= \sum_{k=1}^{K} F_k(\eta) - F_k(\theta)\\
&= \sum_{k=1}^{K} \{F_k(\eta) + E_{k-1}[F_k(\eta)] - E_{k-1}[F_k(\eta)] - F_k(\theta)\\
&\qquad + E_{k-1}[F_k(\theta)] - E_{k-1}[F_k(\theta)]\}\\
&= \sum_{k=1}^{K} \{E_{k-1}[F_k(\eta) - F_k(\theta)] + F_k(\eta) - E_{k-1}[F_k(\eta)]\\
&\qquad - F_k(\theta) + E_{k-1}[F_k(\theta)]\}, \qquad (18)
\end{aligned}$$

where $F_k(\eta) := \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}(\Delta\lambda_{ijk}(\eta) - Z_{ijk}\ln(\Delta\lambda_{ijk}(\eta)))$ and E_{k-1} denotes the conditional expectation taken with respect to Z^k given $\mathcal{H}_{k-1}$.

Observe that

$$\begin{aligned}
&\sum_{k=1}^{K} E_{k-1}[F_k(\eta) - F_k(\theta)]\\
&= \sum_{k=1}^{K}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \Delta\lambda_{ijk}(\theta)\ln\frac{\lambda_{ijk}(\theta)}{\lambda_{ijk}(\eta)} - \Delta(\lambda_{ijk}(\theta) - \lambda_{ijk}(\eta))\\
&= \sum_{k=1}^{K}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \Delta D(\lambda_{ijk}(\theta)||\lambda_{ijk}(\eta)). \qquad (19)
\end{aligned}$$

Since our estimator $\widehat{\theta}$ is the optimal solution to the problem (17), we obtain that $F(\widehat{\theta}) - F(\theta) \le 0$. From (18) and (19), we have

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \Delta D(\lambda_{ijk}(\theta)||\lambda_{ijk}(\widehat{\theta}))\\
&\le \sum_{k=1}^{K} -F_k(\widehat{\theta}) + E_{k-1}[F_k(\widehat{\theta})] + F_k(\theta) - E_{k-1}[F_k(\theta)]\\
&= \sum_{k=1}^{K}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\{-(\Delta\lambda_{ijk}(\theta) - Z_{ijk})\ln(\Delta\lambda_{ijk}(\widehat{\theta}))\\
&\qquad + (\Delta\lambda_{ijk}(\theta) - Z_{ijk})\ln(\Delta\lambda_{ijk}(\theta))\}\\
&= \sum_{k=1}^{K}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}(\Delta\lambda_{ijk}(\theta) - Z_{ijk})\ln\left(\frac{\lambda_{ijk}(\theta)}{\lambda_{ijk}(\widehat{\theta})}\right). \qquad (20)
\end{aligned}$$

We first derive the lower bound and then the upper bound for the inequality (20).

A. Lower Bound for KL-Divergence

For a fixed Z^K, we describe how to obtain the lower bound for

$$\sum_{k=1}^{K}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\Delta D(\lambda_{ijk}(\theta)||\lambda_{ijk}(\widehat{\theta})).$$

From the information theory, we know that

$$D(\lambda_{ijk}(\theta)||\lambda_{ijk}(\widehat{\theta})) \ge H^2(\lambda_{ijk}(\theta)||\lambda_{ijk}(\widehat{\theta})).$$

To obtain the lower and upper bound for any $\lambda_{ijk}(\theta)$ with Z^K, we define the mapping $W^{ij}(\cdot) : \mathbb{R}^{n_1\times n_2\times p} \to \mathbb{R}^{(2n_1-1)\times(2n_2-1)\times p}$ as follows:

$$(W^{ij}(Z_{k-p}^{k-1}))_{i'j'k'} \triangleq$$

$$\begin{cases} Z_{n_1-(i'-i),n_2-(j'-j),k-k'}, & \text{if } i \le i' \le i+n_1-1,\\ & \quad j \le j' \le j+n_2-1,\\ & \quad 1 \le k' \le p\\ 0, & \text{otherwise.}\end{cases}$$

Then, we can express $\lambda_{ijk}(\theta)$ as

$$\lambda_{ijk}(\theta) = \mu_{ij} + \langle W^{ij}(Z_{k-p}^{k-1}), \mathcal{G}\rangle,$$

where $\langle\cdot,\cdot\rangle$ is an inner product for tensors.

Let $\mathbb{E} \in \mathbb{R}^{(2n_1-1)\times(2n_2-1)\times p}$ be a tensor of all ones. We define

$$l := \min_k\{\langle\mathbb{E}, W^{ij}(Z_{k-p}^{k-1})\rangle\} = \min_k\{\|Z_{k-p}^{k-1}\|_1\}$$

and

$$u := \max_{1\le k\le K}\{\|W^{ij}(Z_{k-p}^{k-1})\|_{\text{spec}}\} = \max_{1\le k\le K}\{\|Z_{k-p}^{k-1}\|_{\text{spec}}\}.$$

For any $\lambda_{ijk}(\theta|\mathcal{H}_{k-1})$, the lower bound is

$$\lambda_{ijk}(\theta) \ge a_1 + a_2\langle\mathbb{E}, W^{ij}(Z_{k-p}^{k-1})\rangle \ge a_1 + a_2\, l,$$

and the upper bound is

$$\begin{aligned}
\lambda_{ijk}(\theta) &\le b_1 + \langle\mathcal{G}, W^{ij}(Z_{k-p}^{k-1})\rangle\\
&\le b_1 + \|W^{ij}(Z_{k-p}^{k-1})\|_{\text{spec}}\|\mathcal{G}\|_{\text{TNN}}\\
&\le b_1 + ub_2\sqrt{\gamma(2n_1-1)(2n_2-1)p}
\end{aligned}$$

by Cauchy-Schwartz inequality and the assumptions. As a result, given Z^K,

$$\begin{aligned}
&\underline{J} := a_1 + a_2\, l \le \lambda_{ijk}(\theta)\\
&\le b_1 + ub_2\sqrt{\gamma(2n_1-1)(2n_2-1)p} := \bar{J}, \forall i,j,k,\forall\theta\in\mathcal{D}.
\end{aligned}$$

By Lemma 8 in [9], for all $T \ge \frac{1}{2}(\sqrt{\lambda_{ijk}(\theta)} - \sqrt{\lambda_{ijk}(\widehat{\theta})})^2$, it holds that

$$H^2(\lambda_{ijk}(\theta)||\lambda_{ijk}(\widehat{\theta})) \ge \frac{1-e^{-T}}{4\bar{J}T}[\lambda_{ijk}(\theta) - \lambda_{ijk}(\widehat{\theta})]^2.$$

Taking $T = 2\bar{J}$, we have

$$H^2(\lambda_{ijk}(\theta)||\lambda_{ijk}(\widehat{\theta})) \ge \frac{1-e^{-2\bar{J}}}{8\bar{J}^2}[\lambda_{ijk}(\theta) - \lambda_{ijk}(\widehat{\theta})]^2.$$

For the subsequent discussion, we need the following notation. Let $E_{ij} \in \mathbb{R}^{n_1 \times n_2}$ be a matrix whose ijth entry is one and all the other entries are zero. We also denote the number of parameters as $d := n_1 n_2 + (2n_1 - 1)(2n_2 - 1)p$. Let

$$\beta := [\mathsf{vec}(\mu); \mathsf{vec}(\mathcal{G})] \in \mathbb{R}^d,$$
$$c_{ij}(Z_{k-p}^{k-1}) := [\mathsf{vec}(E_{ij}); \mathsf{vec}(W^{ij}(Z_{k-p}^{k-1}))] \in \mathbb{R}^d,$$
$$A_{ij}(Z_{k-p}^{k-1}) := c_{ij}(Z_{k-p}^{k-1}) c_{ij}(Z^{k-1})_{k-p}^{\top} \in \mathbb{R}^{d \times d},$$
$$A[Z^K] := \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} A_{ij}(Z_{k-p}^{k-1}) \in \mathbb{R}^{d \times d},$$

where vec is a vectorization operator.

Then, we can represent $\lambda_{ijk}(\theta)$ as follows.

$$\lambda_{ijk}(\theta) = c_{ij}(Z_{k-p}^{k-1})^{\top} \beta.$$

Thus,

$$\begin{aligned}
&H^2(\lambda_{ijk}(\theta) || \lambda_{ijk}(\widehat{\theta})) \\
&\geq \frac{1 - e^{-2\bar{J}}}{8\bar{J}^2} [\lambda_{ijk}(\theta) - \lambda_{ijk}(\widehat{\theta})]^2 \\
&= \frac{1 - e^{-2\bar{J}}}{8\bar{J}^2} [c_{ij}(Z_{k-p}^{k-1})^{\top} \beta - c_{ij}(Z_{k-p}^{k-1})^{\top} \widehat{\beta}]^2 \\
&= \frac{1 - e^{-2\bar{J}}}{8\bar{J}^2} (\beta - \widehat{\beta})^{\top} (c_{ij}(Z_{k-p}^{k-1}) c_{ij}(Z_{k-p}^{k-1})^{\top}) (\beta - \widehat{\beta}) \\
&= \frac{1 - e^{-2\bar{J}}}{8\bar{J}^2} (\beta - \widehat{\beta})^{\top} A_{ij}(Z_{k-p}^{k-1}) (\beta - \widehat{\beta}).
\end{aligned}$$

Note that for given data Z_{k-p}^{k-1}, $A_{ij}(Z_{k-p}^{k-1})$ is positive semidefinite ($A_{ij}(Z_{k-p}^{k-1}) \succeq 0$). We use the condition number in Definition 2 to obtain the lower bound for (20):

$$\begin{aligned}
&\sum_{k=1}^{K} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \Delta D(\lambda_{ijk}(\theta) || \lambda_{ijk}(\widehat{\theta})) \\
&\geq \sum_{k=1}^{K} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \Delta H^2(\lambda_{ijk}(\theta) || \lambda_{ijk}(\widehat{\theta})) \\
&\geq \Delta K \sum_{k=1}^{K} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \frac{1 - e^{-2\bar{J}}}{8\bar{J}^2 K} (\beta - \widehat{\beta})^{\top} A_{ij}(Z_{k-p}^{k-1}) (\beta - \widehat{\beta}) \\
&\geq \Delta K \frac{1 - e^{-2\bar{J}}}{8\bar{J}^2} \delta_2[A[Z^K]] \|\beta - \widehat{\beta}\|_2^2 \\
&= \Delta K \frac{1 - e^{-2\bar{J}}}{8\bar{J}^2} \delta_2[A[Z^K]] (\|\mu - \widehat{\mu}\|_F^2 + \|\mathcal{G} - \widehat{\mathcal{G}}\|_F^2). \quad (21)
\end{aligned}$$

The last inequality follows from Definition 2.

B. Upper Bounds for the Random Term

We next derive the upper bound on (20). The upper bound can be written as

$$\begin{aligned}
&\sum_{k=1}^{K} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} (\Delta \lambda_{ijk}(\theta) - Z_{ijk}) \ln \left(\frac{\lambda_{ijk}(\theta)}{\lambda_{ijk}(\widehat{\theta})} \right) \\
&= \sum_{k=1}^{K} \left\langle E, (\Delta \lambda^{(k)}(\theta) - Z^{(k)}) \circ \ln \left(\frac{\lambda^{(k)}(\theta)}{\lambda^{(k)}(\widehat{\theta})} \right) \right\rangle \\
&\leq \sup_{\eta \in \mathcal{D}} \sum_{k=1}^{K} \left\langle E, (\Delta \lambda^{(k)}(\theta) - Z^{(k)}) \circ \ln \left(\frac{\lambda^{(k)}(\theta)}{\lambda^{(k)}(\eta)} \right) \right\rangle, \quad (22)
\end{aligned}$$

where $E \in \mathbb{R}^{n_1 \times n_2}$ is a matrix of all ones, $\lambda^{(k)}, Z^{(k)}$ are kth frontal slice of tensor λ and Z, respectively, $\circ$ is the Hadamard product, and $\ln \left(\frac{\lambda^{(k)}(\theta)}{\lambda^{(k)}(\widehat{\theta})} \right) \in \mathbb{R}^{n_1 \times n_2}$ is a matrix whose ijth entry is equal to $\ln \left(\frac{\lambda_{ijk}(\theta)}{\lambda_{ijk}(\widehat{\theta})} \right)$. For the analysis of (22), we define

$$\xi_k := \mathsf{vec} \left[(\Delta \lambda^{(k)}(\theta) - Z^{(k)}) \circ \ln \left(\frac{\lambda^{(k)}(\theta)}{\lambda^{(k)}(\eta)} \right) \right] \in \mathbb{R}^{n_1 n_2}.$$

Note that ξ_k is a *martingale difference* vector.

In the subsequent discussion, we will apply the Azuma-Hoeffding inequality and union bound property to derive an upper bound. We need the condition, $|(\xi_k)_s| \leq b_s$ for all $s = 1, \ldots, n_1 n_2$, to apply the Azuma-Hoeffding inequality. The bounds can be obtained by applying the following Poisson concentration inequality.

Lemma 4: For $Y \sim Pois(\lambda)$, for all $t > 0$, it holds that

$$\mathbb{P}\{|Y - \lambda| \geq t\} \leq 2e^{-\frac{t^2}{2(\lambda + t)}}.$$

By Lemma 4, for $\epsilon > 0$,

$$\begin{aligned}
\mathbb{P}\{|\Delta \lambda_{ijk}(\theta) - Z_{ijk})| \geq \epsilon | \mathcal{H}_{k-1}\} \quad &\leq 2e^{-\frac{\epsilon^2}{2(\lambda_{ijk}(\theta) + \epsilon)}} \\
&\leq 2e^{-\frac{\epsilon^2}{2(\Delta \bar{J} + \epsilon)}}.
\end{aligned}$$

By the tower property for conditional expectations,

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{P}\{|\Delta \lambda_{ijk}(\theta) - Z_{ijk})| \geq \epsilon | \mathcal{H}_{k-1}\}\right] \\
&= \mathbb{P}\{|\Delta \lambda_{ijk}(\theta) - Z_{ijk}| \geq \epsilon\} \leq 2e^{-\frac{\epsilon^2}{2(\Delta \bar{J} + \epsilon)}}.
\end{aligned}$$

Therefore, by applying the union bound property,

$$|\Delta \lambda_{ijk}(\theta) - Z_{ijk}| \leq \epsilon, \forall i, j, k,$$

with probability $1 - 2n_1 n_2 K e^{-\frac{\epsilon^2}{2(\Delta \bar{J} + \epsilon)}}$. Since $\theta, \eta \in \mathcal{D}$, we have

$$\left| (\Delta \lambda_{ijk}(\theta) - Z_{ijk}) \ln \left(\frac{\lambda_{ijk}(\theta)}{\lambda_{ijk}(\eta)} \right) \right| \leq \epsilon \ln \frac{\bar{J}}{\underline{J}} := \epsilon,'$$

with probability $1 - 2n_1 n_2 K e^{-\frac{\epsilon^2}{2(\Delta \bar{J} + \epsilon)}}$. Note that this shows each entry in ξ_k is not upper bounded by ϵ' with a small probability.

Now we apply the following Theorem.

Theorem 5 (Theorem 32, 33 in [12]): Consider a random variable X and a filtration $\{\mathcal{F}_0, \ldots \mathcal{F}_n\}$. Suppose $X_0, X_1, \ldots X_n$ is a martingale sequence such that $X_i = \mathbb{E}[X | \mathcal{F}_i]$. For $t > 0$, it holds that

$$\mathbb{P}(|X - \mathbb{E}X| \geq t) \leq 2e^{-\frac{t^2}{2\sum_i c_i^2}} + \sum_i \mathbb{P}(|X_i - X_{i-1}| \geq c_i),$$

where $c_1, \ldots, c_n$ are non-negative values.

For fixed s, we define "bad events" as a set such that $|(\xi_k)_s| > \epsilon'$ for any $k = 1 \dots K$. By Theorem 5, the generalized Azuma-Hoeffding inequality can be applied to the sum of unbounded martingale difference with a probability of the "bad events".

For $t > 0$, we obtain

$$\mathbb{P}\left\{\left|\left(\sum_{k=1}^{K}\xi_k\right)_s\right| \geq t\right\} \leq 2e^{-\frac{t^2}{2K\epsilon'^2}} + \mathbb{P}(\text{"bad events"})$$

and it implies that for $x > 0$,

$$\mathbb{P}\left\{\left|\left(\sum_{k=1}^{K}\xi_k\right)_s\right| \geq \sqrt{2\epsilon'^2 xK}\right\} \leq 2e^{-x} + 2Ke^{-\frac{\epsilon^2}{2(\Delta\bar{J}+\epsilon)}}.$$

By the union bound, we have

$$\left\|\sum_{k=1}^{K}\xi_k\right\|_\infty \leq \sqrt{2\epsilon'^2 xK}$$

with probability $1 - 2n_1n_2Ke^{-\frac{\epsilon^2}{2(\bar{J}+\epsilon)}} - 2n_1n_2e^{-x}$.

Let $\alpha_1 = n_1n_2Ke^{-\frac{\epsilon^2}{2(\Delta\bar{J}+\epsilon)}}$ and $\alpha_2 = n_1n_2e^{-x}$, where $\bar{J} = b_1 + u\sqrt{\gamma(2n_1-1)(2n_2-1)p}$ and $x > 0$. By simple computation, we have

$$\begin{aligned}\epsilon &= \ln\frac{n_1n_2K}{\alpha_1} + \sqrt{\ln^2\frac{n_1n_2K}{\alpha_1} + 2\Delta\bar{J}\ln\frac{n_1n_2K}{\alpha_1}} \\ &\leq \max\left\{2\sqrt{\Delta\bar{J}\ln\frac{n_1n_2K}{\alpha_1}},\ 4\ln\frac{n_1n_2K}{\alpha_1}\right\}.\end{aligned}$$

Hence, it follows that

$$\begin{aligned}\left\|\sum_{k=1}^{K}\xi_k\right\|_\infty &\leq \sqrt{2K}\ln\frac{\bar{J}}{\underline{J}}\sqrt{\ln\frac{n_1n_2}{\alpha_2}} \\ &\quad\cdot\max\left\{2\sqrt{\Delta\bar{J}\ln\frac{n_1n_2K}{\alpha_1}},\ 4\ln\frac{n_1n_2K}{\alpha_1}\right\}\end{aligned}$$

with probability $1 - 2\alpha_1 - 2\alpha_2$. Finally, the upper bound is

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\Delta D(\lambda_{ijk}(\theta)||\lambda_{ijk}(\widehat{\theta})) \\ &\leq \sup_{\eta\in\mathcal{D}}\sum_{k=1}^{K}\left\langle E, (\Delta\lambda^{(k)}(\theta) - Z^{(k)})\circ\ln\left(\frac{\lambda^{(k)}(\theta)}{\lambda^{(k)}(\eta)}\right)\right\rangle \\ &\leq \|\mathsf{vec}(E)\|_1\|\sum_{k=1}^{K}\xi_k\|_\infty \\ &\leq n_1n_2\sqrt{2K}\ln\frac{\bar{J}}{\underline{J}}\sqrt{\ln\frac{n_1n_2}{\alpha_2}} \\ &\quad\cdot\max\left\{2\sqrt{\Delta\bar{J}\ln\frac{n_1n_2K}{\alpha_1}},\ 4\ln\frac{n_1n_2K}{\alpha_1}\right\} \qquad (23)\end{aligned}$$

with probability at least $1 - 2\alpha_1 - 2\alpha_2$. We obtain Theorem 3 by combining (21) and (23).

APPENDIX B
NUMERICAL RESULTS FOR SIMULATION

Table II demonstrates the results for the simulations in Section V. "TNN" denotes our method, which involves low-rank constraints using Fourier transformed nuclear norm, while "MLE" denotes the maximum likelihood method without such constraint, "EXP" denotes the estimation method with fixed exponential temporary decay function, and "MNN" denotes the estimation method with the matrix nuclear norm and the exponential decay function. For each case and each sample size, the experiment was repeated five times for each method. The visualization is presented in Fig. 2 in Section V.

REFERENCES

[1] E. Bacry, M. Bompaire, S. Gax'iffas, and J.-F. Muzy, "Sparse and low-rank multivariate Hawkes processes," *J. Mach. Learn. Res.*, vol. 21, no. 50, pp. 1–32, 2020.

[2] M. T. Bahadori, Q. R. Yu, and Y. Liu, "Fast multivariate spatio-temporal analysis via low rank tensor learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3491–3499.

[3] J. A. Bengua, H. N. Phien, H. D. Tuan, and M. N. Do, "Efficient tensor completion for color image and video recovery: Low-rank tensor train," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2466–2479, May 2017.

[4] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[5] C. Cai, G. Li, H. V. Poor, and Y. Chen, "Nonconvex low-rank tensor completion from noisy data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1863–1874.

[6] E. J. Candes and Y. Plan, "Matrix completion with noise," *Proc. IEEE*, vol. 98, no. 6, pp. 925–936, Jun. 2010.

[7] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, 2009, Art. no. 717.

[8] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2053–2080, May 2010.

[9] Y. Cao and Y. Xie, "Poisson matrix recovery and completion," *IEEE Trans. Signal Process.*, vol. 64, no. 6, pp. 1609–1620, Mar. 2016.

[10] Y. Cao and Y. Xie, "Poisson matrix completion," in *Proc. IEEE Int. Symp. Inf. Theory*, 2015, pp. 1841–1845.

[11] J. D. Carroll and J.-J. Chang, "Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition," *Psychometrika*, vol. 35, no. 3, pp. 283–319, 1970.

[12] F. Chung and L. Lu, "Concentration inequalities and martingale inequalities: A survey," *Internet Math.*, vol. 3, no. 1, pp. 79–127, 2006.

[13] M. Fazel, T. K. Pong, D. Sun, and P. Tseng, "Hankel matrix rank minimization with applications to system identification and realization," *SIAM J. Matrix Anal. Appl.*, vol. 34, no. 3, pp. 946–977, 2013.

[14] S. Gandy, B. Recht, and I. Yamada, "Tensor completion and low-n-rank tensor recovery via convex optimization," *Inverse Problems*, vol. 27, no. 2, 2011, Art. no. 025010.

[15] D. Goldfarb and Z. Qin, "Robust low-rank tensor recovery: Models and algorithms," *SIAM J. Matrix Anal. Appl.*, vol. 35, no. 1, pp. 225–253, 2014.

[16] C. Guo and W. Luk, "Accelerating maximum likelihood estimation for Hawkes point processes," in *Proc. IEEE 23rd Int. Conf. Field Programmable Log. Appl.*, 2013, pp. 1–6.

[17] S. J. Hardiman, N. Bercot, and J.-P. Bouchaud, "Critical reflexivity in financial markets: A Hawkes process analysis," *Eur. Phys. J. B*, vol. 86, no. 10, 2013, Art. no. 442.

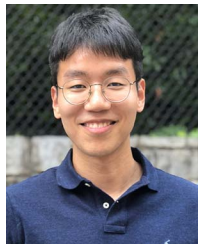
[18] R. A. Harshman et al., "Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multimodal factor analysis," *UCLA Working Papers Phonetics*, vol. 16, pp. 1–84, 1970.

[19] J. C. Hillar and L.-H. Lim, "Most tensor problems are np-hard," *J. ACM*, vol. 60, no. 6, pp. 1–39, 2013.

[20] A. Juditsky, A. Nemirovski, L. Xie, and Y. Xie, "Convex recovery of marked spatio-temporal point processes," *IEEE J Selected Areas in Inf. Theory*, 1, no.3, pp. 799–813. 2020.

[21] M. E. Kilmer and C. D. Martin, "Factorization strategies for third-order tensors," *Linear Algebra Appl.*, vol. 435, no. 3, pp. 641–658, 2011.

[22] M. Kirchner, "An estimation procedure for the Hawkes process," *Quantitative Finance*, vol. 17, no. 4, pp. 571–595, 2017.
[23] M. Kirchner and A. Bercher, "A nonparametric estimation procedure for the Hawkes process: Comparison with maximum likelihood estimation," *J. Stat. Comput. Simul.*, vol. 88, no. 6, pp. 1106–1116, 2018.
[24] R. Kobayashi and R. Lambiotte, "TiDeH: Time-dependent Hawkes process for predicting retweet dynamics," in *Proc. Int. AAAI Conf. Web Social Media*, 2016, pp. 191–200.
[25] V. Koltchinskii et al., "Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion," *Ann. Statist.*, vol. 39, no. 5, pp. 2302–2329, 2011.
[26] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 208–220, Jan. 2013.
[27] Y. Ogata, "Space-time point-process models for earthquake occurrences," *Ann. Inst. Stat. Math.*, vol. 50, no. 2, pp. 379–402, 1998.
[28] T. Omi, Y. Hirata, and K. Aihara, "Hawkes process model with a time-dependent background rate and its application to high-frequency financial data," *Phys. Rev. E*, vol. 96, no. 1, 2017, Art. no. 012303.
[29] P. Reynaud-Bouret et al., "Adaptive estimation for Hawkes processes; Application to genome analysis," *Ann. Statist.*, vol. 38, no. 5, pp. 2781–2822, 2010.
[30] B. Romera-Paredes and M. Pontil, "A new convex relaxation for tensor completion," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2967–2975.
[31] O. Semerci, N. Hao, M. E. Kilmer, and E. L. Miller, "Tensor-based formulation and nuclear norm regularization for multienergy computed tomography," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1678–1693, Apr. 2014.
[32] R. L. Tucker, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, vol. 31, no. 3, pp. 279–311, 1966.
[33] Z. Zhang and S. Aeron, "Exact tensor completion using t-SVD," *IEEE Trans. Signal Process.*, vol. 65, no. 6, pp. 1511–1526, Mar. 2017.
[34] Z. Zhang, G. Ely, S. Aeron, N. Hao, and M. Kilmer, "Novel methods for multilinear data completion and de-noising based on tensor-SVD," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3842–3849.
[35] K. Zhou, H. Zha, and L. Song, "Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes," in *Proc. Artif. Intell. Statist.*, 2013, pp. 641–649.
[36] J. Zhuang and J. Mateu, "A semiparametric spatiotemporal Hawkes-type point process model with periodic background for crime data," *J. Roy. Stat. Soc.: Ser. A. (Statist. Soc.)*, vol. 182, no. 3, pp. 919–942, 2019.



Heejune Sheen received the M.S. degree in statistics from H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA, in 2020. He is currently working toward the Ph.D. degree with the Department of Statistics and Data Science, Yale University, New Haven, CT, USA.

Xiaonan Zhu received the B.S. degree in statistics from the University of Science and Technology of China, Hefei, China, in 2021. She is currently working toward the Ph.D. degree with the Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ, USA.

Yao Xie (Member, IEEE) received the Ph.D. degree in electrical engineering (minor in mathematics) from Stanford University, Stanford, CA, USA, in 2012. She is currently an Associate Professor and Harold R. and Mary Anne Nash Early Career Professor with the H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA, and an Associate Director of the Machine Learning Center. She was a Research Scientist with Duke University, Durham, NC, USA. Her research interests include statistics (particularly change-point detection and spatio-temporal data modeling), machine learning, and signal processing, providing the theoretical foundation and developing computationally efficient and statistically powerful algorithms for engineering problems. She was the recipient of the National Science Foundation CAREER Award in 2017. She is currently an Associate Editor for IEEE TRANSACTIONS ON SIGNAL PROCESSING, *Sequential Analysis: Design Methods and Applications*, *INFORMS Journal on Data Science*, and serves on the Editorial Board of *Journal of Machine Learning Research*.