International Journal of Data Science and Analytics (2020) 10:321–330
https://doi.org/10.1007/s41060-020-00223-3

REGULAR PAPER

Fast multi-resolution segmentation for nonstationary Hawkes process using cumulants

Feng Zhou[1] · Zhidong Li[1] · Xuhui Fan[2] · Yang Wang[1] · Arcot Sowmya[2] · Fang Chen[1]

Received: 30 April 2020 / Accepted: 8 May 2020 / Published online: 8 June 2020

Abstract
The stationarity is assumed in the vanilla Hawkes process, which reduces the model complexity but introduces a strong assumption. In this paper, we propose a fast multi-resolution segmentation algorithm to capture the time-varying characteristics of the nonstationary Hawkes process. The proposed algorithm is based on the first- and second-order cumulants. Except for the computation efficiency, the algorithm can provide a hierarchical view of the segmentation at different resolutions. We extensively investigate the impact of hyperparameters on the performance of this algorithm. To ease the choice of hyperparameter, we propose a refined Gaussian process-based segmentation algorithm, which is proved to be a robust method. The proposed algorithm is applied to a real vehicle collision dataset, and the outcome shows some interesting hierarchical dynamic time-varying characteristics.

Keywords Hawkes process · Nonstationary · Segmentation · Cumulants

1 Introduction

The point process data is a common data type in real applications. To model this kind of point process data, various statistical models have been proposed to disclose its underlying temporal dynamics, such as homogeneous Poisson process [28], inhomogeneous Poisson process [30] and Hawkes process [11]. In this paper, we focus on the Hawkes process.

Hawkes process is widely used to model the self-exciting phenomenon which can be observed in many fields, like crime [16], ecosystem [10], transportation [7] and TV programs [17]. An important way to characterize a temporal point process is through the definition of a conditional intensity. The specific Hawkes process conditional intensity is:

$$\lambda(t) = \mu + \int_0^t \phi(t-s)\mathrm{d}\mathbb{N}(s) = \mu + \sum_{t_i<t} \phi(t-t_i), \qquad (1)$$

where $\mu > 0$ is the baseline intensity which is constant, $\{t_i\}$ are the timestamps of events before time t, $\mathbb{N}(t)$ is the corresponding counting process and $\phi(\cdot) > 0$ is the triggering kernel. The summation of triggering kernels explains the nature of self-excitation, which is the occurrence of events in the past will intensify the intensity of events occurring in the future.

It is straightforward to see that the conditional intensity of Hawkes process is unchanged over timeshifting because μ is a constant and $\phi(\cdot)$ only depends on $\tau = t - t_i$, not on t, which means the stationarity [11,25]. The assumption of stationarity leads to reduced model complexity and easy inference. However, the point process data generated in many real applications has nonstationary properties, which means its first-, second- and higher-order cumulants (moments) are changing over time. Applying the vanilla Hawkes process directly to the nonstationary data is apparently inappropriate. On the other hand, the nonstationarity itself can be an important feature in some applications. For example, the pattern of human heart rate can change from healthy to ill conditions and under different physiological states; in transportation, the influence of a car accident to the road condition is changing between day and night, busy and non-busy hours (see Fig. 1).

One of the common methods of analyzing nonstationary time series is to use segmentation. This kind of problem is also called a change-point problem in mathematics [6,15]. Given a nonstationary point process data, the segmentation algorithm will divide the whole observation period into sev-

✉ Feng Zhou
feng.zhou@uts.edu.au

[1] University of Technology Sydney, Ultimo, Australia

[2] University of New South Wales, Sydney, Australia

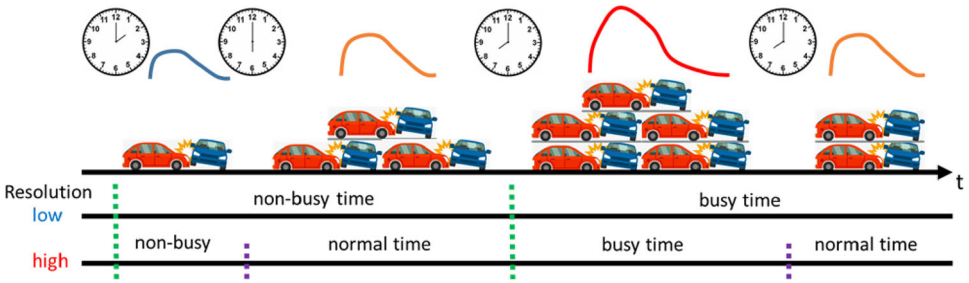

Fig. 1 In transportation, car accidents will have a triggering effect (the arch function above) on the future and the triggering effect is varying between day and night, busy and non-busy hours. An example of the multi-resolution segmentation of the time axis is shown above: a low-resolution (2 segments) and a high-resolution (4 segments) partitions provide a hierarchical insight into the dynamic time-varying characteristics of triggering effect of vehicle collision

eral non-overlapping contiguous segments in such a way that each segment is more approximately stationary than the original data and can be assumed to be stationary.

To the best of our knowledge, no segmentation algorithm has been proposed for the nonstationary Hawkes process. In this paper, we propose the first multi-resolution segmentation (MRS) algorithm for the nonstationary Hawkes process, which can reveal the optimal partition structure in a hierarchical manner. The multi-resolution segmentation is meaningful in real data applications. For example, when the traffic data is analyzed (see Fig. 1), the low-resolution partition (e.g., two segments) corresponds to a "coarser" distinction (e.g., day and night), while the high-resolution partition (e.g., four or six segments) corresponds to a "finer" distinction (e.g., the alternating busy and non-busy hours). The multi-resolution segmentation will help us obtain a hierarchical insight into the nonstationary structure of point process data.

As shown later, the performance of the MRS algorithm depends on the choice of hyperparameters. To ease the choice of one hyperparameter, we propose a revised Gaussian process-based version which is more robust. Overall, our work makes the following contributions:

- We propose the first multi-resolution segmentation algorithm which provides a hierarchical analysis of the dynamic evolution of the nonstationary Hawkes process.
- The MRS depends on the cumulants of the Hawkes process which is fast to compute. Consequently, the MRS (linear computation complexity) is faster than the case of direct estimation of baseline intensity and triggering kernel.
- A more robust revised version of MRS is also proposed to ease the choice of one hyperparameter, which is slower but still acceptable in real applications.

The rest of the paper is organized as follows: In Sect. 2, we summarize some related work. In Sect. 3, we describe the cumulants of the Hawkes process. In Sects. 4 and 5, we propose our MRS algorithm and perform the synthetic data experiment. In Sect. 6, we discuss the Gaussian process MRS (GP-MRS) to ease the choice of one hyperparameter and perform the synthetic data experiment. In Sect. 7, a real vehicle collision dataset is analyzed. Finally, we conclude our work.

2 Related works

2.1 Nonstationary Hawkes process

The generalization of vanilla Hawkes process to nonstationary Hawkes process [25] mainly consists of two cases: the first case is the extension of baseline intensity μ to time-changing $\mu(t)$ and the second case is the extension of triggering kernel $\phi(\tau)$ to time-changing $\phi(\tau, t)$. Plenty of state of the arts have performed inference for a time-changing baseline intensity with a stationary triggering kernel [13,14,27]. For both baseline intensity and triggering kernel being nonstationary, the authors of [24] and [23] provided a general nonparametric estimation theory for the first- and second-order cumulants of a locally stationary Hawkes process. However, this method is inefficient in computation complexity because every point on the two-dimensional covariance function $Cov(\tau, t)$ has to be estimated and it is not applicable to real applications. In this sense, our MRS algorithm can be considered as a "coarser" version of the work in [24]: it combines adjacent small sectors with similar statistical properties into a larger segment and only outputs more heterogeneous segments. Although it is "coarser," the computation complexity is drastically reduced to make it practical.

2.2 Segmentation of time series

The nonstationarity is a common property in time-series data [1,18]. Segmentation is a standard method of data analysis to divide a nonstationary sequential data into a certain

number of non-overlapping contiguous homogeneous segments. A heuristic segmentation algorithm is designed to study the distribution of periods with constant heart rate in [4]. The same method is also applied to analyze changes of the climate [9]. A generalized version is proposed in [5] to overcome the over-segmentation problem caused by heterogeneities induced by correlations. Similarly, the authors of [29] generalizes this existing algorithm for segmenting regime switching processes. All segmentation methods mentioned above cannot be applied directly to the Hawkes process, because they only consider the case of (marked) Poisson process.

3 Cumulants of Hawkes process

The cumulants of Hawkes process [3,12] are used in the MRS algorithm because the utilization of cumulants can accelerate the inference as shown in the experiment of Sec. 5.1. We consider a 1-variate Hawkes process $\mathbb{N}_t$ whose jumps are all of size 1 and whose intensity at time t is $\lambda(t)$. If $\{t_i\}$ denotes the jump times of $\mathbb{N}_t$, the $\lambda(t)$ can be expressed as (1). If $\mathbb{N}_t$ is stationary, the first-order cumulant (mean event rate) is

$$\Lambda \mathrm{d}t = \mathbb{E}(\mathrm{d}\mathbb{N}_t) = \frac{\mu}{1 - \int \phi(\tau)\mathrm{d}\tau}\mathrm{d}t. \tag{2}$$

The second-order cumulant is

$$Cov(\mathrm{d}\mathbb{N}_{t_1}, \mathrm{d}\mathbb{N}_{t_2}) = \mathbb{E}(\mathrm{d}\mathbb{N}_{t_1}\mathrm{d}\mathbb{N}_{t_2}) - \mathbb{E}(\mathrm{d}\mathbb{N}_{t_1})\mathbb{E}(\mathrm{d}\mathbb{N}_{t_2}). \tag{3}$$

Because $\mathbb{N}_t$ is stationary, $Cov(\mathrm{d}\mathbb{N}_{t_1}, \mathrm{d}\mathbb{N}_{t_2})$ only depends on $\tau = t_2 - t_1$ and can be expressed as:

$$v(\tau)\mathrm{d}\tau = \mathbb{E}(\mathrm{d}\mathbb{N}_0\mathrm{d}\mathbb{N}_\tau) - \mathbb{E}(\mathrm{d}\mathbb{N}_0)\mathbb{E}(\mathrm{d}\mathbb{N}_\tau). \tag{4}$$

Or, it can be rewritten in terms of conditional expectations

$$g(\tau)\mathrm{d}\tau = v(\tau)\mathrm{d}\tau/\Lambda = \mathbb{E}(\mathrm{d}\mathbb{N}_\tau|\mathrm{d}\mathbb{N}_0 = 1) - \Lambda\mathrm{d}\tau. \tag{5}$$

The $g(\tau)$ will be used throughout this paper.

A stationary Hawkes process is uniquely defined by its first- and second-order cumulants, and there is a bijection between its second-order statistics $g(\tau)$ and the triggering kernel $\phi(\tau)$ [3].

4 Multi-resolution segmentation

We assume there is a set of observation $\{t_i\}_{i=1}^N$ on $[0, T]$ from a nonstationary Hawkes process where the baseline intensity μ is piecewise constant and the triggering kernel $\phi(\tau)$ is changing over time t. The fundamental idea of MRS is to uniformly divide the observation period $[0, T]$ into M sectors (the highest resolution), i.e., $s_1, \ldots, s_M$, where $\{s_j\}_{j=1}^M$ are sectors and $|s_j|$ is the width of the sector. In each s_j, the point process is assumed to be stationary.

Intuitively, we can estimate the triggering kernel $\phi(\tau)$ in each sector, compare them by adjacent pairs, and find out the possible partition positions. However, the estimation of $\phi(\tau)$ is time consuming no matter in parametric way (maximum likelihood estimation) or nonparametric way (EM algorithm [14], Wiener–Hopf equation [3]), let alone running on all sectors. In order to increase the computation efficiency, we do not estimate $\phi(\tau)$ in each sector directly but use the second-order statistics $g_j(\tau)$ instead which can be estimated faster. The second-order statistics $g_j(\tau)$ in each sector can be empirically estimated using (5).

The reason we can replace $\phi(\tau)$ in each sector with $g_j(\tau)$ is that there is a bijection between them, so the difference between two adjacent $g_j(\tau)$ stands for the nonstationarity of $\phi(\tau)$. The difference of two adjacent $g_j(\tau)$ is written as a normalized mean squared error (NMSE)

$$\mathrm{NMSE} = \mathbb{E}_\tau\left(\left(\frac{g_j(\tau)}{\int g_j(\tau)\mathrm{d}\tau} - \frac{g_{j+1}(\tau)}{\int g_{j+1}(\tau)\mathrm{d}\tau}\right)^2\right). \tag{6}$$

In most cases, $g_j(\tau)$ is an even function for 1-variate Hawkes process when $\tau \to \pm\infty$, $g_j(\tau) \to 0$. If $g_j(\tau)$ is expressed as a histogram function $g_j(\tau) = \sum_{k=1}^K (g_j^k \delta_{kh})$ where $\delta_{kh}(\tau) = 1$ if $(k-1)h \le \tau < kh$ and 0 otherwise, h is the bin-width and $g_j(\tau)$ is 0 beyond the support of Kh, we can write $g_j(\tau)$ as a vector $\mathbf{g}_j = [g_j^k]_{k=1}^K$. Eq. (6) can be rewritten into a discrete version

$$NMSE = \frac{\sum_{k=1}^K \left(\left(\frac{g_j^k}{2h\sum_{k=1}^K g_j^k} - \frac{g_{j+1}^k}{2h\sum_{k=1}^K g_{j+1}^k}\right)^2\right)}{K}. \tag{7}$$

Given the NMSE on all candidate cutting positions, if a desired number of segments (the desired output resolution) R is set, we can pick out the largest $R-1$ cutting positions which is the segmentation.

The scheme of MRS is shown in Fig. 2. By multi-resolution, we mean by increasing (decreasing) the desired output resolution R the segmentation algorithm will output segments at different resolutions in a hierarchical manner. For example, when $R = M$, the partitioner will output the highest resolution (cutting at all candidate positions), as R becomes smaller, the output resolution will be lower (fewer segments will be given out) until there is no cutting at all.

After segmentation, we can piecewisely learn the baseline intensity and triggering kernel on each segment. Specifically, a nonparametric estimation method: Wiener–Hopf equation method [3] is used. As proved in that work, $\phi(\tau)$ and $g(\tau)$ satisfy the Wiener–Hopf equation

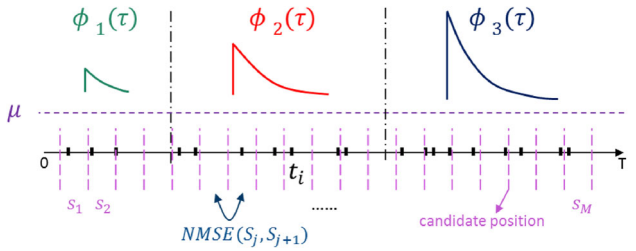

Fig. 2 The scheme of multi-resolution segmentation, for simplicity μ is assumed to be constant and there are 3 different $\phi(\tau)$'s distributed on $[0, T]$

$$g(\tau) = \phi(\tau) + \phi(\tau) * g(\tau), \forall \tau > 0, \tag{8}$$

where $*$ stands for the convolution. In most cases, the Winer–Hopf equation cannot be solved analytically, but there is a lot of literature on how to solve it numerically [2,19]. A common method is the Nystrom method [20]. After the solution of $\phi(\tau)$, μ can be estimated by the first-order cumulant (2). The overall pseudocode of MRS and estimation of μ and $\phi(\tau)$ is formally presented in Alg. 1.

Algorithm 1 Algorithm for MRS and estimation of μ's and $\phi(\tau)$'s

Input: $\{t_i\}_{i=1}^N$, T, R, M, K .
Output: partition positions, μ and $\phi(\tau)$ on each segment .
1: Uniformly divide $[0, T]$ into sectors $\{s_j\}_{j=1}^M$.
2: Estimate the second-order statistics $\mathbf{g}_j = [g_j^k]_{k=1}^K$ on each s_j using (5).
3: Compute the NMSE between two adjacent $\mathbf{g}_j$ using (7).
4: Set a desired output resolution R to obtain the partition positions.
5: After segmentation, estimate the second-order statistics $g(\tau)$ on each segment using (5).
6: Estimate μ and $\phi(\tau)$ on each segment using (8) and (2).
7: **return** partition positions, μ and $\phi(\tau)$ on each segment.

5 Synthetic data experiment of MRS

We use the thinning algorithm [21] to independently generate 40 sets of observations $\{\{t_i\}_{i=1}^{N_l}\}_{l=1}^{40}$ (N_l is the number of points on l-th observation) on $[0, 1000]$ from a nonstationary Hawkes process where μ's are 2, 1.5, 1 and triggering kernels are $\phi_1(\tau) = 1 \cdot \exp(-2\tau)$, $\phi_2(\tau) = 2 \cdot \exp(-4\tau)$ and $\phi_3(\tau) = 3 \cdot \exp(-4\tau)$ distributed on $[0, 200]$, $[200, 600]$ and $[600, 1000]$, respectively (see Fig. 2). The goal is to find the underlying partition structure and estimate μ's and $\phi(\tau)$'s in a nonparametric way. The highest resolution is set to be $M = 10$ ($|s_j| = 100$), $g_j(\tau)$ is expressed as a histogram function $g_j(\tau) = \sum_{k=1}^K (g_j^k \delta_{kh})$ where $h = 0.75$ and $K = 8$. The hyperparameters M and K are fine tuned here and will be discussed in Sect. 5.2.

Table 1 Multi-resolution segmentation results. "New Position" is the newly added partition position

R	1	2	3	4	5
New position	∅	**600**	**200**	500	900
$\frac{\text{Min(Threshold)}}{\text{Max(NMSE)}}$	100%	**88.45%**	**11.41%**	8.05%	7.15%
R	6	7	8	9	10
New position	400	300	700	800	100
$\frac{\text{Min(Threshold)}}{\text{Max(NMSE)}}$	6.88%	5.17%	2.24%	1.25%	0%

We average the estimated $g_j(\tau)$ over 40 sets of independent observations and Table 1 shows the multi-resolution segmentation results in a hierarchical manner as R increases from 1 to the highest resolution 10. We can see when $R = 1$, there is no cutting at all; when $R = 3$, the partition positions match the ground truth; when $R = 10$ the algorithm cuts at every candidate position (the highest resolution). To quantify the NMSE caused by estimation variance, we show the proportion of the minimum threshold corresponding to R over the maximum NMSE in Table 1. We can see the NMSE induced by estimation variance is below 11.41% (the last correct cutting is "200" which corresponds to 11.41%), which means the MRS is robust to obtain the correct segmentation.

Setting $R = 3$, the correct segmentation $[0, 200]$, $[200, 600]$, $[600, 1000]$ is obtained. The next step is to infer μ and $\phi(\tau)$ on each segment. We empirically estimate the second-order statistics $g(\tau)$ on each segment and solve the Winer–Hopf equation (8). The estimated $\hat{\mu}_1 = 2.05$, $\hat{\mu}_2 = 1.64$, $\hat{\mu}_3 = 1.01$ and $\hat{\phi}(\tau)$'s are shown in Fig. 3. We can see the estimation matches the ground truth.

5.1 Complexity of MRS

In this section, we analyze the computation complexity of MRS algorithm. The MRS algorithm has a linear computation complexity which means it is practical. The complexity of MRS mainly depends on two parameters: the highest resolution M and the size of the observation multiplying the number of bins on $g_j(\tau)$: $\mathcal{N}K$ where $\mathcal{N} = \sum_l N_l$.

The complexity of estimation of $g_j(\tau)$ on each sector is $\mathcal{O}(n_j K)$ where n_j is the number of points in s_j, consequently, the complexity of all $g_j(\tau)$ on all independent observations is $\mathcal{O}(\mathcal{N}K)$. The complexity of NMSE between two adjacent $g_j(\tau)$ over M sectors is $\mathcal{O}(M)$. Therefore, the final complexity of MRS is $\mathcal{O}(\mathcal{N}K + M)$. The running time experiments over $\mathcal{N}K$ (M) given M ($\mathcal{N}K$) are shown in Fig. 4 which proves the linear complexity.

We also compare the consuming time of estimation of $\phi(\tau)$ with $g(\tau)$ to prove the necessity of utilizing cumulants as features for acceleration. Given 1,896 observation points, the consuming time of $g(\tau)$ is 0.5 s but 38.4 s for $\phi(\tau)$, which

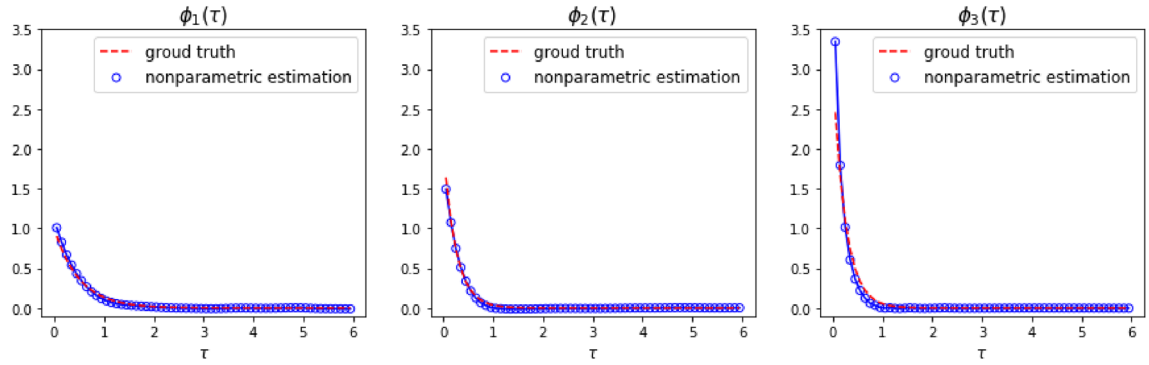

Fig. 3 The estimated $\hat{\phi}_1(\tau)$, $\hat{\phi}_2(\tau)$ and $\hat{\phi}_3(\tau)$. The ground truths are $1 \cdot \exp(-2\tau)$, $2 \cdot \exp(-4\tau)$ and $3 \cdot \exp(-4\tau)$, respectively

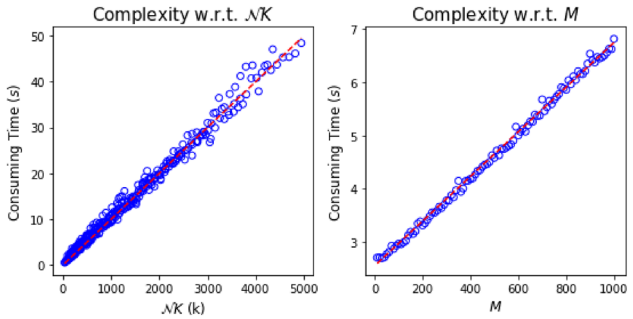

Fig. 4 The consuming time of MRS (left: w.r.t. $\mathcal{N}K$ given $M = 10$; right: w.r.t. M given $\mathcal{N}K = 31,547 \times 8$)

proves replacing $\phi(\tau)$ with $g(\tau)$ is an efficient way to speed up the inference.

5.2 Hyperparameters

The difference between two adjacent estimated $g_j(\tau)$ and $g_{j+1}(\tau)$ is from two sources: the first source is the difference between $\mathbb{E}(g_j(\tau))$ and $\mathbb{E}(g_{j+1}(\tau))$ which is the nonstationarity, the second source is the estimation variance of $g_j(\tau)$ induced by the choice of hyperparameters. There are two hyperparameters affecting the performance of MRS: M and K.

Intuitively, the highest resolution M should not be too small or too large. If too small, there are few sectors as the candidate partition positions; consequently, the segmentation result from MRS degrades. If too large, there will be fewer points in each sector s_j, which means the estimation variance of $g_j(\tau)$ is large, consequently, the segmentation result also degrades.

Given $K = 8$, the experiment is performed with M from 3 to 20. The segmentation and NMSE results with $R = 3$ are shown in Table 2 and Fig. 5. We can see when M is in $[10, 16]$, the segmentation from MRS is close to the ground truth; when $M > 20$, the estimation variance is overwhelming, as a result, the partition positions are misidentified.

Table 2 Segmentation results of MRS w.r.t. M and K

M	3	**10**	16	20
Partition positions	333.3, 666.6	**200**, **600**	187.5,687.5	150,350
K	**10**	**20**	30	40
Partition positions	**200**, **600**	**200**, **600**	100,200	100, 200

Given an appropriate highest resolution M, the performance of MRS is also affected by the hyperparameter K. The reason behind this phenomenon is that as K becomes larger, there are more bins on $g_j(\tau)$ and the estimated $\mathbf{g}_j = [g_j^k]_{k=1}^K$ will be overfitting. To show this problem, we perform experiments given the highest resolution $M = 10$ but with $K = 10, 40$ and 100. The estimated $\mathbf{g}_1$ when $K = 10, 40$ and 100 is shown in Fig. 7 (only the positive half is shown because of even function). It is clear that the $\mathbf{g}_1$ with $K = 100$ is overfitting since there are many spikes up and down. The more bins we have, the larger the estimation variance of $g_j(\tau)$ will be, which will lead to a misidentified segmentation. To prove it, the segmentation and NMSE results when $K = 10, 20, 30$ and 40 with $R = 3$ are shown in Table 2 and Fig. 6. We can see when $K \geq 30$, the segmentation obtained from MRS does not match the ground truth anymore.

6 A refined MRS algorithm: GP-MRS

Intuitively, a model selection experiment can be performed to obtain the optimal hyperparameters M and K. Nevertheless, for a more robust model, we propose a refined MRS algorithm: GP-MRS in this section, by using which we do not need to choose the optimal value of K. We can arbitrarily set a large K as GP-MRS can prevent it from overfitting.

6.1 Description of GP-MRS

The key idea of GP-MRS is to use a standard GP regression to smooth the vector $\mathbf{g}_j = [g_j^k]_{k=1}^K$ in each sector. Given

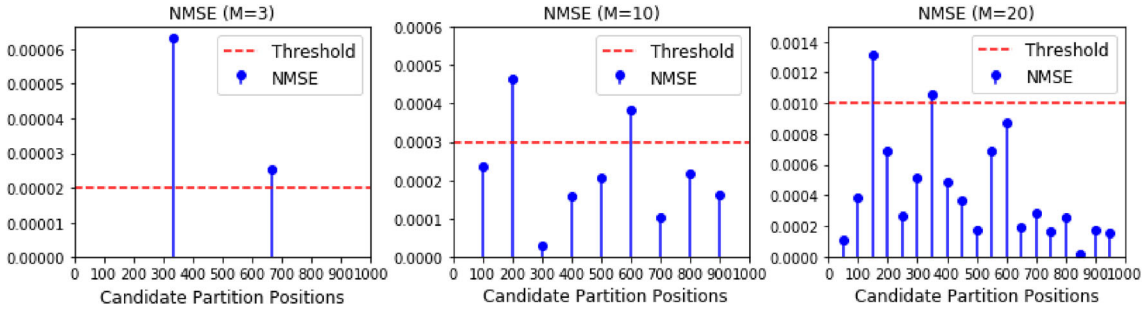

Fig. 5 Given $K = 8$, the NMSE of MRS w.r.t. M. The threshold corresponds to $R = 3$ (Only $M = 3, 10, 20$ are shown)

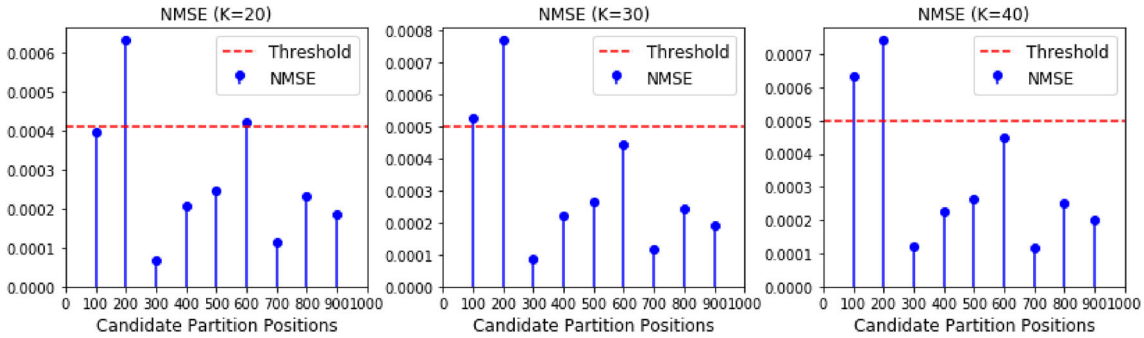

Fig. 6 Given $M = 10$, the NMSE of MRS w.r.t. K. The threshold corresponds to $R = 3$ (Only $K = 20, 30, 40$ are shown)

$\mathbf{g}_j = [g_j^1, g_j^2, \ldots, g_j^K]$ in s_j, the GP regression is performed to evaluate the posterior mean function $\overline{g}_j(\tau|g_j^1, g_j^2, \ldots, g_j^K)$

$$\overline{g}_j(\tau) = \mathbf{d}^{\mathrm{T}}\mathbf{C}_K^{-1}\mathbf{g}_j^{\mathrm{T}}, \tag{9}$$

where $\mathbf{C}_K$ is the matrix of $C(\tau_k, \tau_{k'}) = ker(\tau_k, \tau_{k'}) + \sigma^2\delta_{kk'}$, $\{\tau_{k(k')}\}_{k(k')=1}^K$ are x-values of K training points and σ^2 is the noise variance of training points, the vector $\mathbf{d} = [ker(\tau_1, \tau), \ldots, ker(\tau_K, \tau)]^{\mathrm{T}}$, $\mathbf{g}_j$ are y-values of K training points. Here, the covariance kernel is squared exponential kernel $ker(x, x') = \theta_0 \exp\left(-\frac{\theta_1}{2}\|x - x'\|^2\right)$ where θ_0 and θ_1 are hyperparameters of GP. We use $\overline{g}_j(\tau)$ to replace the directly estimated $\mathbf{g}_j$ in NMSE (7). By using GP-MRS, the NMSE induced by overfitting of $g_j(\tau)$ can be effectively eliminated when K is large (comparison between Figs. 6 and 8).

It is worth noting that GP-MRS cannot be applied to address the problem of M because a too large M will lead to a sparse sector where the GP regression cannot provide a true posterior mean function. To obtain the optimal hyperparameter M, a rule of thumb formula is provided: $M \approx \mathcal{N}/250L$ where L is the number of independent observations.

6.2 Synthetic data experiment of GP-MRS

We apply the GP-MRS algorithm to the same experiment as in the Sect. 5.2. The GP hyperparameters are fine-tuned to $\theta_0 = 1, \theta_1 = 1, \sigma^2 = 0.01$. It is out of the scope of this paper to discuss how to choose the GP hyperparameters. The estimated $\overline{g}_1(\tau)$ when $K = 10, 40$ and 100 is shown in Fig. 7. It is clear that the $\overline{g}_j(\tau)$ from GP-MRS is stable whatever K is. We also analyze the segmentation and NMSE results with $R = 3$ which are shown in Table 3 and Fig. 8; we can see the segmentation and NMSE are both stable whatever K is. Conclusively, the GP-MRS is more robust than the MRS and can provide the correct segmentation in cases where the MRS does not work.

Table 3 Segmentation results of GP-MRS w.r.t. K

K	20	30	40	200
Partition positions	200, 600	200, 600	200, 600	200, 600

6.3 Complexity of GP-MRS

For a standard GP regression, it costs $\mathcal{O}(K^3)$ for matrix inversion when calculating K training points ($\mathbf{g}_j = [g_j^k]_{k=1}^K$). The final complexity of GP-MRS is $\mathcal{O}(\mathcal{N}K + MK^3)$. Unavoidably, the introduce of GP regression into MRS will make it slower. Given $\mathcal{N} = 120,000$ and $M = 10$, the consuming time of GP-MRS when $K = 40$ is 48.64 s on a normal desktop. We can see it is still acceptable when K is not too large.

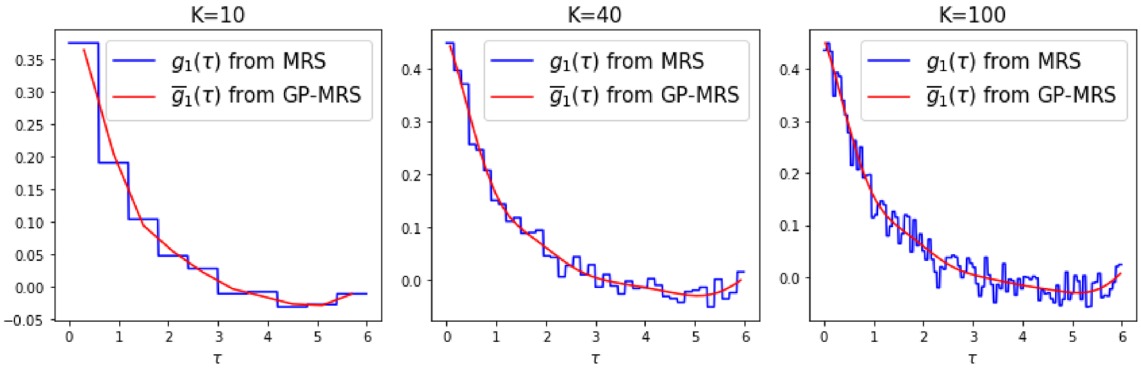

Fig. 7 Given $M = 10$, the estimated $g_1(\tau)$ from MRS and GP-MRS when $K = 10, 40$ and 100

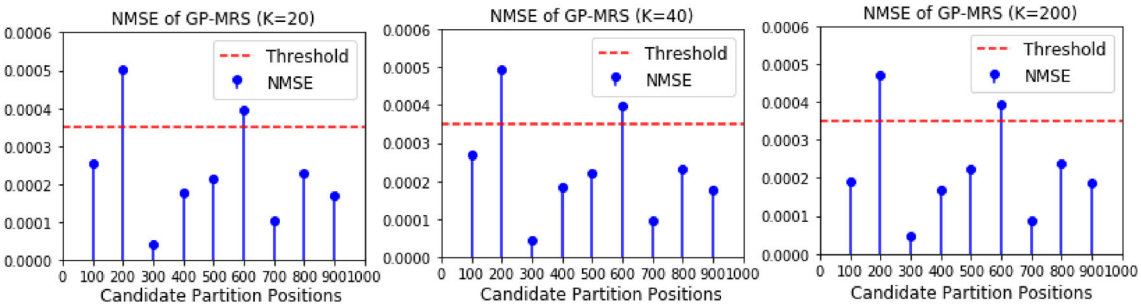

Fig. 8 Given $M = 10$, the NMSE of GP-MRS w.r.t. K. The threshold corresponds to $R = 3$ (Only $K = 20, 40, 200$ are shown).The NMSE is nearly unchanged whatever K is

6.4 Comparison with baseline models

As shown, the GP-MRS algorithm is superior to the MRS algorithm on hyperparameter selection. In this section, we compare our GP-MRS algorithm with several baseline models including

- Inhomogeneous Poisson process (IPP): a Poisson process with a smooth intensity function. The intensity function is estimated by the algorithm in [26].
- Stationary parametric Hawkes process (SPHP): the vanilla Hawkes process with constant μ and exponential decay $\phi(\tau)$. The parameters are estimated by maximum likelihood estimation in [22].
- Stationary nonparametric Hawkes process (SNHP): the nonparametric Hawkes process with constant μ and nonparametric triggering kernel $\phi(\tau)$. The inference is performed by the Wiener–Hopf method in [3].
- Semi-nonstationary nonparametric Hawkes process (SNNHP): the Hawkes process with nonstationary $\mu(t)$ and stationary nonparametric triggering kernel $\phi(\tau)$. The inference is performed by the maximum penalized likelihood estimation in [14].
- Nonstationary nonparametric Hawkes process (NNHP): the Hawkes process with nonstationary $\mu(t)$ and nonstationary nonparametric triggering kernel $\phi(t, \tau)$. The inference is performed by our GP-MRS algorithm.

Table 4 Training and test log-likelihood of all models for synthetic data

	IPP	SPHP	SNHP	SNNHP	NNHP
training log-likeihood	50.3	103.2	153.6	169.4	**192.6**
Test log-likeihood	52.5	80.6	120.1	148.3	**160.9**

We utilize the same 3-segment experimental setup as in Sec. 5 to generate two sets of data. One is used as the training data and the other one as the test data. We measure the training and test log-likelihood to characterize the fitting and prediction ability. The result is shown in Table 4 where we can see the NNHP model with our GP-MRS inference algorithm is better than the alternatives w.r.t. both training and test log-likelihood due to its superior model expression.

7 Real data experiment

The GP-MRS is applied to a real vehicle collision dataset to discover the hierarchical time-varying characteristics.

7.1 Vehicle collisions in New York City

The vehicle collision dataset[1] is provided by the New York City Police Department. It contains about 1.05 million vehicle collision records in New York City from July 2012 to September 2017. The dataset includes the collision date, time, borough, location, contributing factor and so on.

In daily transportation, the vehicle collision occurring in the past will increase the intensity of vehicle collision occurring in the future because of the traffic jam caused by the initial collision, so there exists a triggering effect from the past collision to the future one. There are already some works trying to model the triggering effect using classic Hawkes process (parametric or nonparametric), but they all assume the stationarity is satisfied. However, this is not the case in real life. As shown later, we reveal the hierarchical time-varying characteristics of triggering kernel and baseline intensity of vehicle collision over 24 h by using the GP-MRS algorithm.

7.1.1 Weekdays

We filter out the collision records on all weekdays from May 1st, 2017 to June 30th, 2017. Some collisions are occurring at the same time as the data resolution is at a minute level, which violates the definition of the temporal point process. To avoid this, we add a small time interval to all the simultaneous records to separate them.

The observation every day is assumed to be independent, so there are 45 sets of independent observations. Totally, 137,578 points are observed. We use the GP-MRS for segmentation which is still fast enough in this case. The whole observation period T is set to 1440 min (24 h a day). The support of $\phi(\tau)$ is set to 8 min. The hyperparameters of GP-MRS θ_0, θ_1, σ^2 are set to 1, 1, 0.01 by cross-validation; K can be arbitrarily set to a large number (20 is used here) and M is chosen to be 12 by the rule of thumb, which means the sector size is 120 min (2 h).

When the desired output resolution $R = 2$, the consuming time of GP-MRS is about 10 s and the cutting positions are 2:00 and 8:00. The segmentation is shown in Fig. 9 left and can be understood as the busy time and non-busy time. After segmentation, we estimate μ and $\phi(\tau)$ on each segment. The estimated μ's are $\mu_1 = 0.317$ and $\mu_2 = 0.127$, the estimated $\phi(\tau)$'s are shown in Fig. 10 left. We can see both μ_1 and $\phi_1(\tau)$ are larger than μ_2 and $\phi_2(\tau)$ which is consistent with our common sense because the traffic is more crowd in a busy time. Additionally, the estimated nonparametric triggering kernel is not strictly monotonic decreasing: there is a small bump around 5 min after the initial collision, which proves the superior flexibility of nonparametric estimation.

[1] https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95

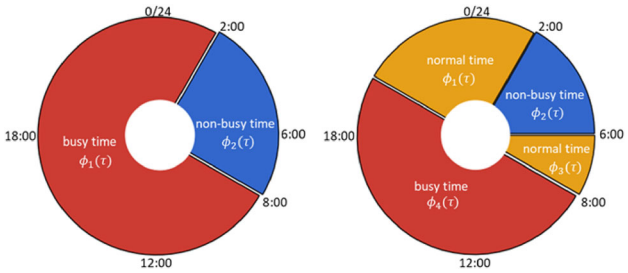

Fig. 9 Weekdays: The 24-h segmentation result of vehicle collisions, 2 segments (left) and 4 segments (right)

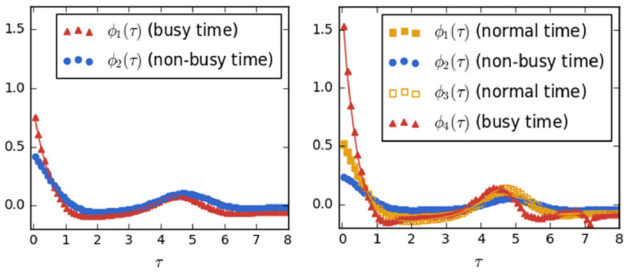

Fig. 10 Weekdays: The estimated $\phi(\tau)$ of vehicle collisions, 2 segments (left) and 4 segments (right)

To show the hierarchical multi-resolution property of GP-MRS, the desired output resolution R is increased to 4 and we obtain a finer segmentation. The consuming time in this case is also about 10 s and the cutting positions are 2:00, 6:00, 8:00 and 20:00. The segmentation is shown in Fig. 9 right. The segmentation can be understood as the normal time, busy time and non-busy time. The late night is between 2:00 and 6:00 which are non-busy hours; the after-work entertainment hours (from 20:00 to 2:00) together with morning commute hours (from 6:00 to 8:00) are the normal time; the daytime (from 8:00 to 20:00) is the busy time. The estimated μ's are $\mu_1 = 0.32$, $\mu_2 = 0.12$, $\mu_3 = 0.29$ and $\mu_4 = 0.59$. The estimated $\phi(\tau)$'s are shown in Fig. 10 right. Two normal-time $\phi(\tau)$'s are almost overlapping; both the baseline intensity and triggering kernel of busy time are larger than normal time, larger than the non-busy time at the initial stage.

7.1.2 Weekends

We also filter out collision records on all weekends from February 1st, 2017 to August 31st, 2017. With $R = 2$, the cutting positions are 2:00 and 8:00 which are same as weekdays. With $R = 3$, we can get a finer segmentation: 2:00, 8:00 and 12:00. The segmentation is shown in Fig. 11. The estimated $\phi(\tau)$'s are shown in Fig. 12.

An interesting phenomenon is that the low-resolution time-varying characteristics of weekdays are similar with that of weekends, but the high-resolution characteristics are very different, e.g., 6:00–8:00 becomes the non-busy time

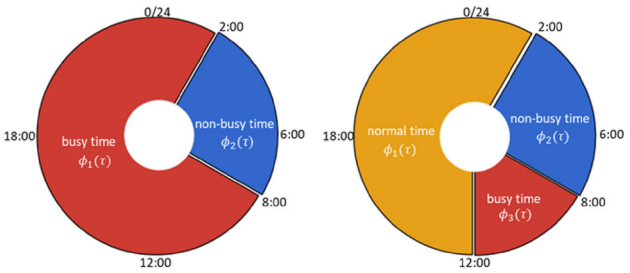

Fig. 11 Weekends: The 24-h segmentation result of vehicle collisions, 2 segments (left) and 3 segments (right)

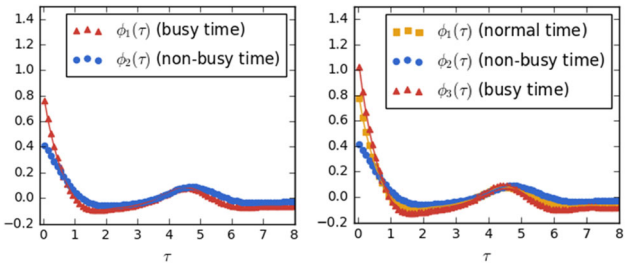

Fig. 12 Weekends: The estimated $\phi(\tau)$ of vehicle collisions, 2 segments (left) and 3 segments (right)

Table 5 Training and test log-likelihood of all models for vehicle collision data

	IPP	SPHP	SNHP	SNNHP	NNHP
Training log-likeihood	−594.3	−581.6	−567.5	−540.9	**−536.9**
Test log-likeihood	−783.5	−708.3	−687.3	−651.6	**−637.2**

on weekends maybe because of late waking up; 12:00–20:00 becomes the normal time maybe because of less heavy traffic. The multi-resolution segmentation provides a hierarchical insight into the dynamic evolution of vehicle collision.

7.2 Comparison with baseline models

As in the synthetic data, we compare the performance of IPP, SPHP, SNHP, SNNHP and NNHP on the real data. The training and test log-likelihood of all models on the vehicle collision training and test data (weekdays) is shown in Table 5 where the NNHP model with our GP-MRS inference algorithm fits the data best w.r.t. both training and test log-likelihood.

8 Conclusions

There has been lots of research work made on modeling the change-point in nonstationary and heterogeneous data. An interesting and promising approach is the online detection of change-point in stochastic processes [8,18]. At the current stage, our proposed MRS algorithm can address the heterogeneous sequence data with a batch method. In the future work, the extension to online learning can be considered.

In this paper, we propose an MRS algorithm to partition the nonstationary Hawkes process, which provides a hierarchical view of the nonstationary structure. In this way, the hierarchical dynamic time-varying characteristics of nonstationary Hawkes process can be discovered. Besides, the algorithm is fast because of the utilization of cumulants as features. After segmentation, the baseline intensity and triggering kernel are estimated in a nonparametric way. Overall, this is a nonstationary and nonparametric Hawkes process. To ease the choice of hyperparameter, a refined GP-MRS algorithm is also proposed at the cost of lower efficiency but still acceptable. Both synthetic and real data experiments show the superiority of our proposed model over the state of the arts.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Allesiardo, R., Féraud, R., Maillard, O.A.: The non-stationary stochastic multi-armed bandit problem. Int. J. Data Sci. Anal. **3**(4), 267–283 (2017)
2. Atkinson, K.: A survey of numerical methods for the solution of Fredholm integral equations of the second kind (1976)
3. Bacry, E., Muzy, J.F.: First-and second-order statistics characterization of Hawkes processes and non-parametric estimation. IEEE Trans. Inf. Theory **62**(4), 2184–2202 (2016)
4. Bernaola-Galván, P., Ivanov, P.C., Amaral, L.A.N., Stanley, H.E.: Scale invariance in the nonstationarity of human heart rate. Phys. Rev. Lett. **87**(16), 168105 (2001)
5. Bernaola-Galván, P., Oliver, J., Hackenberg, M., Coronado, A., Ivanov, P.C., Carpena, P.: Segmentation of time series with long-range fractal correlations. Eur. Phys. J. B **85**(6), 211 (2012)
6. Carlstein, E.G., Müller, H.G., Siegmund, D.: Change-point problems. IMS (1994)
7. Du, N., Dai, H., Trivedi, R., Upadhyay, U., Gomez-Rodriguez, M., Song, L.: Recurrent marked temporal point processes: Embedding event history to vector. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 1555–1564 (2016)
8. Fan, Y., Lu, X.: An online Bayesian approach to change-point detection for categorical data. Knowl. Based Syst. 105792 (2020)
9. Feng, G.L., Gong, Z.Q., Dong, W.J., Li, J.P.: Abrupt climate change detection based on heuristic segmentation algorithm (2005)
10. Gupta, A., Farajtabar, M., Dilkina, B., Zha, H.: Discrete interventions in Hawkes processes with applications in invasive species management. In: IJCAI, pp. 3385–3392 (2018)
11. Hawkes, A.G.: Spectra of some self-exciting and mutually exciting point processes. Biometrika **58**(1), 83–90 (1971)
12. Jovanović, S., Hertz, J., Rotter, S.: Cumulants of Hawkes point processes. Phys. Rev. E **91**(4), 042802 (2015)

13. Lemonnier, R., Vayatis, N.: Nonparametric markovian learning of triggering kernels for mutually exciting and mutually inhibiting multivariate Hawkes processes. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, pp. 161–176 (2014)
14. Lewis, E., Mohler, G.: A nonparametric EM algorithm for multiscale Hawkes processes. J. Nonparametr. Stat. **1**(1), 1–20 (2011)
15. Liu, S., Yamada, M., Collier, N., Sugiyama, M.: Change-point detection in time-series data by relative density-ratio estimation. Neural Netw. **43**, 72–83 (2013)
16. Liu, Y., Yan, T., Chen, H.: Exploiting graph regularized multi-dimensional Hawkes processes for modeling events with spatio-temporal characteristics. In: IJCAI, pp. 2475–2482 (2018)
17. Luo, D., Xu, H., Zhen, Y., Ning, X., Zha, H., Yang, X., Zhang, W.: Multi-task multi-dimensional Hawkes processes for modeling event sequences. In: Twenty-Fourth International Joint Conference on Artificial Intelligence (2015)
18. Miyaguchi, K., Yamanishi, K.: Online detection of continuous changes in stochastic processes. Int. J. Data Sci. Anal. **3**(3), 213–229 (2017)
19. Noble, B., Weiss, G.: Methods based on the Wiener–Hopf technique for the solution of partial differential equations. Phys. Today **12**, 50 (1959)
20. Nyström, E.J.: Über die praktische Auflösung von Integralgleichungen mit Anwendungen auf Randwertaufgaben. Acta Math. **54**(1), 185–204 (1930)
21. Ogata, Y.: Space-time point-process models for earthquake occurrences. Ann. Inst. Stat. Math. **50**(2), 379–402 (1998)
22. Ozaki, T.: Maximum likelihood estimation of hawkes' self-exciting point processes. Ann. Inst. Stat. Math. **31**(1), 145–155 (1979)
23. Roueff, F., Von Sachs, R.: Time-frequency analysis of locally stationary Hawkes processes. arXiv preprint arXiv:1704.01437 (2017)
24. Roueff, F., Von Sachs, R., Sansonnet, L.: Locally stationary Hawkes processes. Stoch. Process. Their Appl. **126**(6), 1710–1743 (2016)
25. Roueff, F., Von Sachs, R., et al.: Time-frequency analysis of locally stationary Hawkes processes. Bernoulli **25**(2), 1355–1385 (2019)
26. Samo, Y.L.K., Roberts, S.: Scalable nonparametric Bayesian inference on point processes with gaussian processes. In: International Conference on Machine Learning, pp. 2227–2236 (2015)
27. Tannenbaum, N.R., Burak, Y.: Theory of nonstationary Hawkes processes. Phys. Rev. E **96**(6), 062314 (2017)
28. Thompson, W.: Point Process Models with Applications to Safety and Reliability. Springer, Berlin (2012)
29. Toth, B., Lillo, F., Farmer, J.D.: Segmentation algorithm for non-stationary compound Poisson processes. Eur. Phys. J. B **78**(2), 235–243 (2010)
30. Weinberg, J., Brown, L.D., Stroud, J.R.: Bayesian forecasting of an inhomogeneous Poisson process with applications to call center data. J. Am. Stat. Assoc. **102**(480), 1185–1198 (2007)