Estimation of self-exciting point processes from time-censored data

Philipp J. Schneider [*] and Thomas A. Weber [†]
École Polytechnique Fédérale de Lausanne, Station 5, CH-1015 Lausanne, Switzerland

(Received 23 December 2022; accepted 18 May 2023; published 17 July 2023)

Self-exciting point processes, widely used to model arrival phenomena in nature and society, are often difficult to identify. The estimation becomes even more challenging when arrivals are recorded only as bin counts on a finite partition of the observation interval. In this paper, we propose the recursive identification with sample correction (RISC) algorithm for the estimation of process parameters from time-censored data. In every iteration, a synthetic sample path is generated and corrected to match the observed bin counts. Then the process parameters update and a unique iteration is performed to successively approximate the stochastic characteristics of the observed process. In terms of finite-sample approximation error, the proposed RISC framework performs favorably over extant methods, as well as compared to a naïve locally uniform sample redistribution. The results of an extensive numerical experiment indicate that the reconstruction of an intrabin history based on the conditional intensity of the process is crucial for attaining superior performance in terms of estimation error.

DOI: 10.1103/PhysRevE.108.015303

I. INTRODUCTION

Self-exciting point processes have been used in a broad range of areas such as biology [1], credit collections [2], criminology [3], earthquake prediction [4], epidemiology [5,6], information theory [7], marketing [8], neuroscience [9], and social networks [10]. The identification of these processes depends on historical data which are usually assumed to be available in the form of precise time stamps of arrival events. However, there are also numerous practical situations where arrival data are only reported in batches, e.g., as daily aggregates. In the case of COVID-19 disease statistics, for instance, updated data are posted and recorded at the end of any given working day, with weekend arrivals reported only as part of the following Monday statistics. The time censoring appears as a natural consequence of quasiperiodic reporting cycles which are dictated by the absence of live feeds, as well as a need for prerelease verification.[1] At other times, time aggregation into bins may even occur deliberately, as a means of masking the data, either for reasons of privacy or—as often the case in financial markets—to differentiate a low-resolution basic data feed from a high-resolution premium data feed.

This paper considers the identification of self-exciting point processes based on time-censored (or binned) data. Our proposed recursive identification with sample correction (RISC) algorithm takes as input a vector of observed arrival counts on the subintervals (or bins) of a given partition of the observation interval. The key idea is to maintain an estimation based on full-resolution time series which are obtained from simulated (synthetic) sample paths conditional on being consistent with the observed bin counts. An essential element of the algorithm is therefore a sample correction (SC) so as to either subtract samples by thinning or add samples by thickening, until the bin-count vector of the sample-corrected synthetic sequence equals the original bin-count vector. The sequence of parameter estimates is chosen to guarantee an increasing conditional likelihood, which in turn ensures convergence of the algorithm. When tested against other extant estimation methods, the RISC algorithm performs favorably. Our extensive numerical analysis highlights the difficulties of estimating process parameters which result from the generally nonconvex likelihood objective. The latter is to the detriment of other methods which may well exhibit consistent large-sample estimation behavior but might then perform poorly in realistic finite-data use cases.

A. Literature

Self-exciting point processes, also referred to as Hawkes processes, whose intensity for new arrivals depends on the arrival history, were introduced by Hawkes [12,13]. The identification of Hawkes processes is customarily performed using either maximum-likelihood estimation (MLE) [14,15] or expectation-maximization (EM) algorithms [16–18]. While these methods are not without their own challenges (e.g., due to the nonconvexity or local flatness of the likelihood function) [19], our RISC algorithm uses these mainstream estimation techniques for the internal inference of parameters after having removed time censoring; cf. Sec. II A 2.

In the presence of time-censored (binned) data, three main ideas have emerged in the literature to identify the

[*] philipp.schneider@epfl.ch
[†] thomas.weber@epfl.ch

[1] Delayed state information poses challenges for the determination of effective feedback-control strategies. During the COVID-19 pandemic, for instance, delays in the transmission of contact-tracing data proved a salient difficulty for governments in their efforts to effectively control the epidemiological situation [11].

dynamics of Hawkes processes: autoregressive discrete-time approaches, spectral methods using filtering and averaging concepts, and, finally, approaches based on resampling continuous-time histories consistent with observed bin counts. We briefly discuss these three classes of algorithms in turn; the proposed RISC algorithm is most closely related to the last.

The first class of algorithms uses *autoregressive models* to describe the evolution of bin counts when passing along the evenly spaced discrete time periods. For example, Mark *et al.* [20] used a sampled version of the standard log likelihood for Hawkes processes derived by Ozaki [21] to account for the discrete-time evolution of the intensity.[2] In the same spirit, Kirchner [22] introduced an integer-valued autoregression (INAR) as a discrete-time approximation of a Hawkes process, establishing convergence when the order of the autoregression goes to infinity. Finally, the INAR estimation procedure was extended by the same author to also cover multivariate Hawkes processes, with an application to financial limit-order-book data [23]. Kirchner and Bercher [24] showed that for weakly interval-censored data (i.e., for fairly narrow bins), the INAR method achieves similar estimation results as MLE on uncensored data.[3]

Consider now the second class of algorithms, based on *spectral methods*. Cheysson and Lang [25] developed a spectral estimation method based on Whittle's log likelihood [26], building on the work by Dzhaparidze [27]. Consistency and asymptotic normality of the estimator are obtained using the strong mixing properties of Hawkes processes. The authors derived the spectral density via the Bartlett spectrum and account for aliasing.[4] Another interesting development, based on spectral methods, is due to Rizoiu *et al.* [28] who considered a mean behavior Poisson (MBP) process, obtained via Laplace transform, which represents the *expected* intensity of a Hawkes process. The Hawkes-process parameters are then proxied by MBP parameters. In an extension of that work, Calderon *et al.* [29] used a so-called partial mean behavior Poisson process to fit a multivariate Hawkes process when not all components of the event data are time-censored.

The third class of algorithms aims at reconstructing a synthetic event history that most likely caused the observed vector of bin counts. In this vein, Shlomovich *et al.* [30] introduced a binned Hawkes expectation maximization (BH-EM) algorithm for estimating Hawkes process parameters. The method is based on the Monte Carlo implementation of the EM algorithm by Wei and Tanner [31], which uses a random sample of latent data to update the expected value of the log posterior as an arithmetic average (or mixture) before the maximization step. To generate distinct arrival time stamps, the authors iterate over each bin and attempt to determine intermediate event times which maximize (at least locally) the joint probability density conditional on the observed bin count.

[2]Due to the superiority of the method by Shlomovich *et al.* [30], we neglect the approach by Mark *et al.* [20] in our comparison tests.

[3]We test the performance of the RISC algorithm allowing for very significant time censoring; cf. Sec. III.

[4]Aliasing folds high frequencies of the original process into the spectrum of the interval-censored process.

As noted before, our proposed RISC algorithm falls into the third class, combining SC with recursive parameter estimation. Regarding the latter, quite general properties of recursive identification methods were outlined by Söderström *et al.* [32]. Indeed, recursive MLE is a common identification technique in signal processing with nonlinear systems and observational noise [33,34]. Similarly, by recursively identifying the process parameters, successive iterates are associated with a maximizing sequence of likelihood values which leads to convergence, and ultimately to parameters being less affected by the noise introduced through interval censoring of event-arrival times. On the other hand, SC refers to the notion of adjusting a simulated (synthetic) sample path, based on the last parameter estimates, to achieve a high fidelity to the best available estimate of the Hawkes process. Four different SC methods are proposed and tested. Finally, it is important to note that—in contrast to all extant approaches—the RISC algorithm has been constructed for any (nonuniform) partition of the observation interval, thus allowing for different process regimes. This flexibility suggests the formulation of an inverse problem, namely, that of optimizing time censoring in view of achieving certain envelope objectives (e.g., energy efficiency in sensor operations); cf. Sec. IV D.

B. Outline

The remainder of the paper is organized as follows. Section II introduces the preliminaries for the proposed RISC algorithm. We determine various theoretical approaches for reconstructing a valid sample compatible with the observed time-censored data vector by correcting a given continuous-time sample path that reflects the current parameter estimate. Convergence of the RISC algorithm is established formally. In Sec. III, we introduce relative and absolute performance criteria to compare the RISC algorithm against other known solutions. In Sec. IV, we further discuss algorithmic performance and extensions. Section V concludes.

II. RECURSIVE IDENTIFICATION WITH SAMPLE CORRECTION

In line with extant literature, our RISC approach is introduced and tested for the class of self-exciting point processes. We first recall the corresponding law of motion and standard estimation techniques, together with the entropy of time-censored observations (in Sec. II A). Subsequently, the general idea of the RISC method is discussed (in Sec. II B) before we provide a collection of methods to redistribute synthetically generated samples in statistically neutral ways (in Sec. II C) using suitable thinning and thickening algorithms, so the corrected synthetic sample is consistent with a given time-censored sample-path observation that serves as a reference envelope. Finally, we establish the convergence of the RISC algorithm (in Sec. II D).

A. Preliminaries

1. Self-exciting point processes

Self-exciting point processes are a class of spatiotemporal processes introduced by Hawkes [12,13]. The *self-excitation*

characteristic of any such Hawkes process refers to its property that earlier arrivals (at times $t_1, \ldots, t_i > 0$) determine the likelihood of future arrivals (at times $t_j > t_i$). The corresponding law of motion specifies the arrival intensity at any time $t \geqslant 0$ conditional on the event history $\mathcal{H}_t = \{t_i : 0 < t_i < t\}$, so

$$\lambda(t|\mathcal{H}_t) = \mu + \sum_{i:t_i<t} \phi(t - t_i), \quad t \geqslant 0, \tag{1}$$

where $\mu > 0$ is the background rate and $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ is the (continuous and bounded) self-excitation function also referred to as kernel.[5] In the special case where $\phi = 0$, the process simplifies to a homogeneous Poisson process. A popular choice is the exponential kernel,

$$\phi(t) = \alpha\, \beta\, e^{-\beta t}, \quad t \geqslant 0, \tag{2}$$

where $\alpha = \int_0^\infty \phi(t)\, dt > 0$ denotes the branching coefficient and $\beta > 0$ the decay. The exponential kernel satisfies the Markov property.[6]

2. Standard parameter estimation

Common approaches for estimating the process parameters of self-exciting point processes include MLE, EM, and nonparametric algorithms. Both parametric techniques (MLE and EM) determine parameter estimates based on maximizing the (expected) likelihood of an observed event history $\mathcal{H}_T \subset (0, T]$ over an observation horizon $T > 0$.

Maximum-likelihood estimation. Given an observation history $\mathcal{H}_T = \{t_1, \ldots, t_K\}$, an MLE parameter estimate $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ for the Hawkes process in Eq. (1) solves

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} \in \arg\max_{\boldsymbol{\theta}\in\Theta} \left\{ \sum_{i=1}^{K} \ln \lambda\big(t_i|\mathcal{H}_{t_i}\big) - \int_0^T \lambda(s|\mathcal{H}_s)\, ds \right\}; \tag{3}$$

the parameter vector $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^P$ contains the background rate μ, together with $P - 1$ kernel parameters;[7] the parameter space Θ is assumed to be nonempty, convex, and compact. Ogata [14] established that MLE estimators are consistent, asymptotically normal, and efficient.

Expectation maximization. The maximization of the log likelihood in Eq. (3) can be challenging in practice. Since the objective is generally nonconvex (cf. Ogata and Akaike [38]), determining an MLE estimator amounts to finding the solution to a global optimization problem. In addition, the log-likelihood function may exhibit regions with very shallow gradients that can lead to numerical divergence. To cope with these issues, the use of additional structural process information has proved useful. Dempster *et al.* [16] considered parameter estimation in the presence of latent variables. Marsan and Lengline [39], Veen and Schoenberg [18], and Mohler *et al.* [17] adopted this procedure to derive EM-type algorithms for Hawkes processes, maximizing the expected log likelihood of a parameter vector conditional on an observation history $\mathcal{H}_T = \{t_1, \ldots, t_K\}$, as before. Using the immigrant-offspring representation as derived by Hawkes and Oakes [40] as latent information, the stochastic branching-structure matrix $\boldsymbol{P} = [p_{ij}]_{i,j=1}^K$ of the Hawkes process in Eq. (1) is given by

$$p_{ii} = \frac{\mu}{\lambda\big(t_i|\mathcal{H}_{t_i}\big)}, \quad p_{ij} = \frac{\phi(t_i - t_j)}{\lambda\big(t_i|\mathcal{H}_{t_i}\big)}, \quad p_{ji} = 0, \quad 1 \leqslant j < i \leqslant K, \tag{4}$$

where p_{ii} is the probability of an event being an immigrant and p_{ij} is the probability for an offspring event. By offsprings, we refer to events being caused by previous events. Immigrant events, on the other hand, are a product of the background rate and occur independently of the event history. The *conditional* expected (complete) log-likelihood function $Q : \Theta \to \mathbb{R}$ augments the standard log likelihood for the parameter vector $\boldsymbol{\theta}$, with the branching structure $\hat{P} = [\hat{p}_{ij}]_{i,j=1}^K$ conditioned on the parameter vector $\hat{\boldsymbol{\theta}}$ (generally different from $\boldsymbol{\theta}$),

$$Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{K} \hat{p}_{ii} \ln \mu + \sum_{i=2}^{K} \sum_{j=1}^{i-1} \hat{p}_{ij} \ln\big(\phi(t_i - t_j)\big) - \int_0^T \lambda(t|\mathcal{H}_t)\, ds, \quad \boldsymbol{\theta}, \hat{\boldsymbol{\theta}} \in \Theta, \tag{5}$$

where the parameter space $\Theta \subset \mathbb{R}^P$ is as before. An EM estimator $\hat{\boldsymbol{\theta}}_{\text{EM}}$ is obtained iteratively. Given an approximate solution $\hat{\boldsymbol{\theta}}^k$, a solution update $\hat{\boldsymbol{\theta}}^{k+1}$ is found by first calculating the branching structure (E step) and then maximizing $Q(\cdot|\hat{\boldsymbol{\theta}}^k)$ (M step) with respect to $\boldsymbol{\theta} \in \Theta$. The two-step EM process repeats until a convergence criterion is reached, say, at $k = \bar{k}$, so $\hat{\boldsymbol{\theta}}_{\text{EM}} = \hat{\boldsymbol{\theta}}^{\bar{k}}$; see Mark and Weber [19] for details. In recent studies, Nandan *et al.* [41,42] augmented an EM algorithm to also account for spatial variability in modeling earthquake occurrences by including a time- and space-varying exogenous influence.

Nonparametric estimation. In practice, at times there is no ready-to-use parametric kernel, rendering process-identification challenging. This issue is addressed by nonparametric estimation methods that require only limited prior knowledge about the shape or scale of the kernel (such as a Lipschitz constant). Some of the latest nonparametric methods are due to Achab *et al.* [43] and Bacry and Muzy [44].

3. Time-censored observation

In many practical applications, the event history $\mathcal{H}_T = \{t_1, \ldots, t_K\} \subset (0, T]$ cannot be observed at full resolution,

[5]The background rate is assumed constant. This assumption has to be critically assessed based on the application of interest. Recent studies on Hawkes processes analyzing the endoexogeneity in financial markets indicate that this assumption may introduce a certain estimation bias [35,36].

[6]That is, the intensity change $d\lambda(t|\mathcal{H}_t)$ depends only on $\lambda(t|\mathcal{H}_t)$ and the change of the corresponding counting process $N(t|\mathcal{H}_t) = \sum_{t_i \leqslant t} 1$, so $d\lambda(t|\mathcal{H}_t) = -\beta\lambda(t|\mathcal{H}_t)dt + \alpha\beta N(t|\mathcal{H}_t)$; see, e.g., Bacry *et al.* [[37], Prop. 2].

[7]For example, when using the exponential kernel in Eq. (2), it is $\boldsymbol{\theta} = (\mu, \alpha, \beta) \in \Theta \subset \mathbb{R}_+^P$, with $P = 3$.

but only after some time censoring. The latter consists in aggregating event counts X_ℓ, for $\ell \in \{1, \dots, L\}$, relative to a binning partition $\mathscr{P}_T = \{\mathcal{B}_1, \dots, \mathcal{B}_L\}$ of the observation interval $(0, T]$, where L is the number of bins $\mathcal{B}_\ell = (\tau_{\ell-1}, \tau_\ell]$, and $0 = \tau_0 < \tau_1 < \cdots < \tau_L = T$. That is,

$$X_\ell = \sum_{i=1}^{K} \mathbf{1}_{\{t_i \in \mathcal{B}_\ell\}} = N(\tau_\ell|\mathcal{H}_T) - N(\tau_{\ell-1}|\mathcal{H}_T), \quad \ell \in \{1, \dots, L\}. \tag{6}$$

In particular, the sum of the bin counts equals the total number of observations:

$$\sum_{\ell=1}^{L} X_\ell = K = |\mathcal{H}_T|. \tag{7}$$

The norm $\|\mathscr{P}_T\|$ of the binning partition $\mathscr{P}_T$ is equal to the length of its largest bin:

$$\|\mathscr{P}_T\| = \max\{\tau_\ell - \tau_{\ell-1} : \ell \in \{1, \dots, L\}\}.$$

We can conclude that a time-censored observation (relative to the binning partition $\mathscr{P}_T$) yields the bin-count vector $X = (X_1, \dots, X_L)$ instead of the full event history $\mathcal{H}_T$, so Eqs. (6) and (7) hold. To indicate the fact that X is obtained by garbling the information contained in the event history $\mathcal{H}_T$ by means of the binning partition $\mathscr{P}_T$, we write

$$X = \mathcal{H}_T / \mathscr{P}_T. \tag{8}$$

The information in the bin-count vector can be measured by the bin-count entropy:[8]

$$H(X) = -\sum_{\ell=1}^{L} \left(\frac{X_\ell}{K}\right) \ln\left(\frac{X_\ell}{K}\right). \tag{9}$$

As a result, $0 \leqslant H(X) \leqslant \ln L$, where the upper bound is attained for $X_\ell = K/L$ observations in each bin and the lower bound for $L = 1$. Indeed, full time censoring into a single bin (when $L = 1$) cannot convey any information at all, while the maximum amount of information is transmitted by a binning partition that fully separates all observed events (so $X_\ell \in \{0, 1\}$ for all ℓ).

A bin-count partition $\hat{\mathscr{P}}_T = \{\hat{\mathcal{B}}_1, \dots, \hat{\mathcal{B}}_{\hat{L}}\}$ is said to be a refinement of $\mathscr{P}_T$ (written as $\hat{\mathscr{P}}_T \succeq \mathscr{P}_T$) if for any $\hat{\ell} \in \{1, \dots, \hat{L}\}$ there exists an index $\ell \in \{1, \dots, \ell\}$ so $\hat{\mathcal{B}}_{\hat{\ell}} \subset \mathcal{B}_\ell$. It is clear that refining a binning partition can only increase the information contained in a bin-count vector. That is,[9]

$$((X, \hat{X}) = (\mathcal{H}_T/\mathscr{P}_T, \mathcal{H}_T/\hat{\mathscr{P}}_T) \quad \text{and} \quad \hat{\mathscr{P}}_T \succeq \mathscr{P}_T) \quad \Rightarrow \quad H(\hat{X}) \geqslant H(X). \tag{10}$$

Consider now the *relative norm* of the binning partition,

$$\Delta = \frac{\|\mathscr{P}_T\|}{T} \in [0, 1],$$

which is useful for comparison purposes across different observation horizons. Given an event history $\mathcal{H}_T$, Fig. 1 shows the bin-count entropy relative to a random selection of binning partitions (with uniformly distributed breakpoints τ_ℓ, for different $L \in \{1, \dots, |\mathcal{H}_T|\}$), as a function of Δ. The relative spread (coefficient of variation) of entropy values diminishes in the number of bins.

Remark 1 (Uniform time censoring). A *Uniform* binning partition,

$$\mathscr{P}_T = \{((T/L)(\ell - 1), (T/L)\ell]\}_{\ell=1}^{L},$$

has the relative norm $\Delta = 1/L$. For ease of comparison, we restrict attention to uniform binning partitions in the numerical performance evaluation of our method; see Sec. III. This is notwithstanding the fact that in practice nonuniform binning

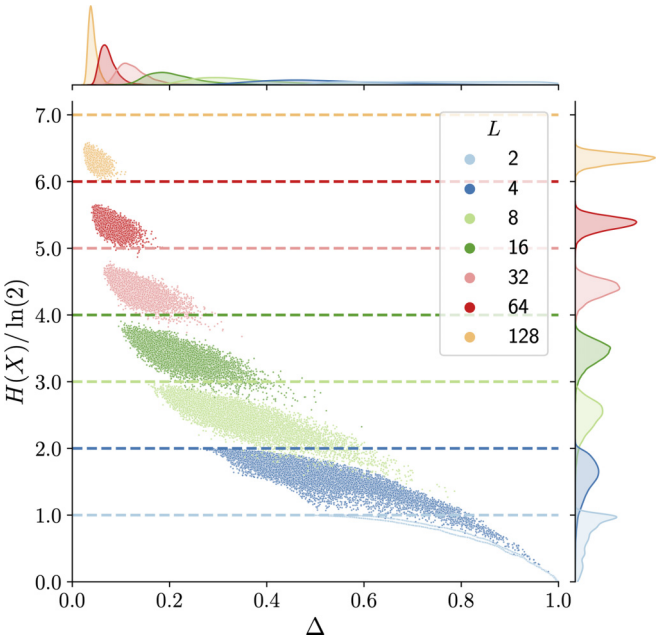


FIG. 1. Bin-count entropy for random partitions with L uniformly distributed breakpoints.

[8] We use the convention (which holds in the limit) that all terms for which X_ℓ is zero do vanish.

[9] To see this, note first that $\mathcal{B}_\ell \cap \hat{\mathcal{B}}_{\hat{\ell}} \neq \emptyset$ implies that $\mathcal{B}_\ell \supset \hat{\mathcal{B}}_{\hat{\ell}}$, for all $(\mathcal{B}_\ell, \hat{\mathcal{B}}_{\hat{\ell}}) \in \mathscr{P}_T \times \hat{\mathscr{P}}_T$, by virtue of the fact that $\hat{\mathscr{P}}_T \succeq \mathscr{P}_T$. Thus, we have that bin counts in $X = \mathcal{H}_T/\mathscr{P}_T$ can be obtained by partially aggregating bin counts in $\hat{X} = \mathcal{H}_T/\hat{\mathscr{P}}_T$, since $X_\ell = |\mathcal{B}_\ell \cap \mathcal{H}_T| = \sum_{\hat{\ell}:\hat{\mathcal{B}}_{\hat{\ell}} \subset \mathcal{B}_\ell} |\hat{\mathcal{B}}_{\hat{\ell}} \cap \mathcal{H}_T| = \sum_{\hat{\ell}:\hat{\mathcal{B}}_{\hat{\ell}} \subset \mathcal{B}_\ell} \hat{X}_{\hat{\ell}}$ for all $\ell \in \{1, \dots, L\}$. But this implies that $H(X) = \sum_{\ell=1}^{L} \sum_{\hat{\ell}:\hat{\mathcal{B}}_{\hat{\ell}} \subset \mathcal{B}_\ell} (\hat{X}_{\hat{\ell}}/K) \ln(X_\ell/K) \leqslant \sum_{\ell=1}^{L} \sum_{\hat{\ell}:\hat{\mathcal{B}}_{\hat{\ell}} \subset \mathcal{B}_\ell} (\hat{X}_{\hat{\ell}}/K) \ln(\hat{X}_{\hat{\ell}}/K) = H(\hat{X})$ as claimed in Eq. (10), due to Jensen's inequality—in combination with the concavity of the logarithm.

may occur naturally. For example, many national COVID-19 statistics use daily bins on weekdays, while weekends are lumped together into larger bins.[10]

B. Recursive identification

Given a binning partition $\mathscr{P}_T$ of the observation interval $(0, T]$, a time-censored observation yields the bin-count vector X in Eq. (8), without the event history $\mathcal{H}_T$ being known. In fact, while many event histories may yield the observed bin-count vector X, the idea of our proposed RISC algorithm is to synthetically construct event histories $\hat{\mathcal{H}}_T^s$, subject to the bin-count constraint

$$X = \hat{\mathcal{H}}_T^s \big/ \mathscr{P}_T, \tag{11}$$

so as to maximize the likelihood of having been generated by a Hawkes process HP($\hat{\boldsymbol{\theta}}$) in Eq. (1), for which an estimate $\hat{\boldsymbol{\theta}}$ of the parameter vector is given.[11] For this purpose, we generate a synthetic event history,

$$\mathcal{H}_T^s = \{t_1^s, \ldots, t_{K_s}^s\},$$

by simulating HP($\hat{\boldsymbol{\theta}}$). Naturally the synthetic bin-count vector $X^s = \mathcal{H}_T^s / \mathscr{P}_T$ might differ from the observed bin-count vector X, which implies the need for thinning (when $X_\ell < X_\ell^s$) or thickening (when $X_\ell > X_\ell^s$), to satisfy the bin-count constraint in Eq. (11). The details of this sample correction within the RISC algorithm are discussed in Sec. II C.

Given a corrected event history $\hat{\mathcal{H}}_T^s = \{\hat{t}_1^s, \ldots, \hat{t}_K^s\}$, which satisfies the bin-count constraint (so $X = \hat{\mathcal{H}}_T^s / \mathscr{P}_T$), the identification portion of the RISC algorithm aims to use one of the standard parametric identification methods in Sec. II A 2 (i.e., MLE or EM) to generate an updated parameter estimate $\hat{\boldsymbol{\theta}}'$. By recursion on the three steps, namely, sample generation, sample correction, and identification (with parameter update), the RISC algorithm gradually identifies the parameters of the time-censored point process and at the same time produces event histories consistent with the time-censored observation. Figure 2 provides an overview. By construction, the corrected synthetic event history has the same cardinality as the (unknown) true history:

$$K = |\mathcal{H}_T| = \sum_{\ell=1}^{L} X_\ell = \big|\hat{\mathcal{H}}_T^s\big|.$$

Hence, the likelihood of the synthetic event histories can be compared across iterations, as well as within a given iteration by using multiple corrected synthetic sample paths instead of just one; selecting the corrected path with the highest likelihood to have been generated by HP($\hat{\boldsymbol{\theta}}$) reduces the simulation noise. The termination criterion,

$$\|\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}'\| \leqslant \epsilon,$$

is then predicated upon the norm of the parameter adjustment after the identification step dropping below a prespecified

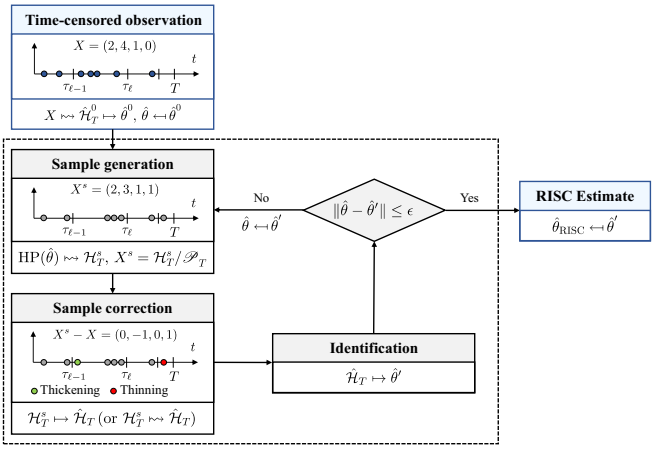


FIG. 2. RISC algorithm: Overview.

positive error ϵ. To initialize the RISC algorithm, one can generate an event history $\hat{\mathcal{H}}_T^0$ by uniformly distributing X_ℓ samples in each interval $(\tau_{\ell-1}, \tau_\ell]$ of the binning partition $\mathscr{P}_T$, so the bin-count constraint is naturally satisfied. This implies that $\hat{\mathcal{H}}_T^0$ is in fact corrected, and the first parameter vector $\hat{\boldsymbol{\theta}}^0$ can be produced using a standard identification step.

C. Sample correction

By means of SC, a synthetically generated event history $\mathcal{H}_T^s$ of a Hawkes process HP($\hat{\boldsymbol{\theta}}$) can be converted to an observation-compatible synthetic event history $\hat{\mathcal{H}}_T^s$ which satisfies the bin-count constraint (11), given the bin-count observation X and the binning partition $\mathscr{P}_T$. The corresponding thinning and thickening procedures described next improve upon the naïve uniform sampling method used to initialize the RISC algorithm. The general goal of more sophisticated SC methods is to produce continuous-time event histories that most closely resemble the arrival characteristics of the Hawkes process under consideration.

1. Thinning

Based on a (possibly already modified) synthetic event history $\mathcal{H}_T^s = \{t_1^s, \ldots, t_{K_s}^s\}$, obtained by simulating the Hawkes process HP($\hat{\boldsymbol{\theta}}$) for a given parameter vector $\hat{\boldsymbol{\theta}}$, we now consider the removal of excess samples from bin $\mathcal{B}_\ell = (\tau_{\ell-1}, \tau_\ell]$. For this, we first compute the intensity $\lambda_T^s(t|\mathcal{H}_T^s)$ by means of Eq. (1), for all $t \in [0, \tau_\ell]$ that is, from the beginning of the observation interval (i.e., $t = 0$) until the end of bin ℓ (i.e., $t = \tau_\ell$). Based on this hazard rate, it is possible to determine the arrival probability of the ith simulated event (conditional on t_{i-1}^s),

$$F_i^s = 1 - \exp\left[-\int_{\max\{\tau_{\ell-1}, t_{i-1}^s\}}^{t_i^s} \lambda_T^s\big(t|\mathcal{H}_T^s\big)\, dt\right], \quad t_i^s \in \big[\mathcal{H}_T^s\big]_\ell, \tag{12}$$

which corresponds to one minus the conditional survival probability on the interval $(\max\{\tau_{\ell-1}, t_{i-1}^s\}, t_i^s]$ for a death-arrival process with hazard rate $\lambda_T^s(\cdot|\mathcal{H}_T^s)$, where $[\mathcal{H}_T^s]_\ell = \mathcal{B}_\ell \cap \mathcal{H}_T^s$ denotes the set of simulated arrivals t_i^s in the interval

[10] Nonuniform (and random) time censoring is discussed in Sec. IV C.

[11] The parameter vector pins down the background rate μ and the self-excitation function ϕ; cf. footnote 7.

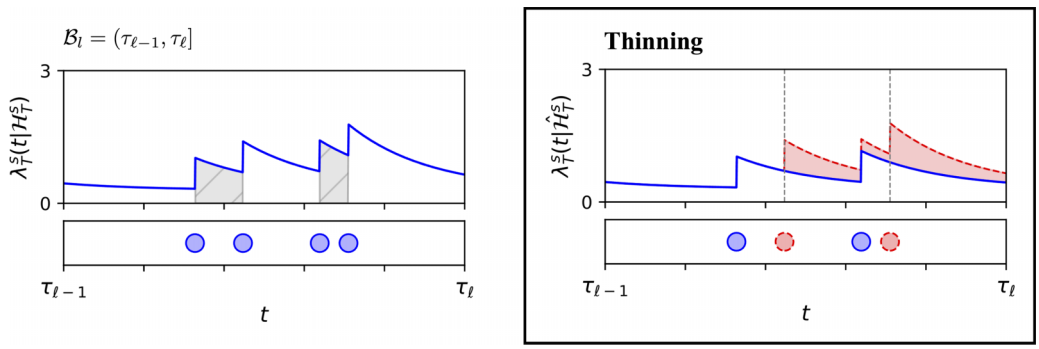


FIG. 3. Thinning process when $X_\ell^s - X_\ell = 2$. Two events with the lowest conditional arrival probabilities (measured as the corresponding areas under the curve) are removed. The dashed grey lines indicate the canceled events.

$(\tau_{\ell-1}, \tau_\ell]$, and where we set $t_0^s = 0$ as initial condition. By assumption, for thinning to be required in bin ℓ, it is

$$X_\ell^s - X_\ell = \left|\left[\mathcal{H}_T^s\right]_\ell\right| - |\mathcal{B}_\ell \cap \mathcal{H}_T| > 0.$$

Thus, a corrected sample history $\hat{\mathcal{H}}_T^s$ is obtained from $\mathcal{H}_T^s$ by removing $X_\ell^s - X_\ell$ samples, namely, those featuring the smallest arrival probabilities, so

$$\left[\hat{\mathcal{H}}_T^s\right]_\ell \in \arg \max_{\{t_i^s\} \subset \left[\mathcal{H}_T^s\right]_\ell} \prod_{t_i^s} F_i^s, \quad \text{s.t. } \left|\{t_i^s\}\right| = X_\ell.$$

The proposed thinning procedure therefore maximizes the joint conditional arrival probability, subject to the ℓth bin-count constraint, whence $\hat{X}_\ell^s = |[\hat{\mathcal{H}}_T^s]_\ell| = X_\ell$. See Fig. 3 for an illustration of the thinning procedure in a case with two excess samples.

Remark 2 (Iterative thinning with intensity update). Instead of removing all excess samples at once from a given bin, a more precise method is to remove only one excess sample at a time (based on the lowest arrival rate), and to then recompute the sample intensity path λ_T^s, repeating the removals and subsequent intensity updates until the bin-count constraint is satisfied.

2. Thickening

Maintaining the same notation as in the preceding thinning procedure, we now consider the statistically neutral adding of samples in bin ℓ when

$$X_\ell^s - X_\ell = \left|\left[\mathcal{H}_T^s\right]_\ell\right| - |\mathcal{B}_\ell \cap \mathcal{H}_T| = -n < 0.$$

Four different thickening procedures (I)–(IV) are discussed in turn. The first method (I), based on order statistics, generates all n missing events at once, whereas the remaining three methods (II)–(IV) add one event at a time recursively, in n iterations.

(I) Order statistics. Given that n events need to be added, the idea is to take the nonhomogeneous arrival intensity $\lambda_T^s(\cdot|\mathcal{H}_T^s)$ as given and add the events at their *expected* time instants, which amount to the expected values of the corresponding n distinct order statistics conditional on n arrivals, pertaining to the first event, the second event, and so forth, until the nth event. By the law of large numbers, these expected arrival times would be approximated by the observed averages of the arrival times taken over sufficiently many n-event sample paths (discarding all sample paths with a different number of events).

To derive closed-form expressions for the thickening arrival times $(\zeta_1, \ldots, \zeta_n)$, recall first that any Poisson process with a constant (positive) intensity, conditional on n arrivals in an interval $\mathcal{B}_\ell = (\tau_{\ell-1}, \tau_\ell]$, features expected arrival times which partition $\mathcal{B}_\ell$ into $n+1$ subintervals of the same length, so

$$\zeta_i = \tau_{\ell-1} + \left(\frac{\tau_\ell - \tau_{\ell-1}}{n+1}\right) i, \quad i \in \{1, \ldots, n\}.$$

When the arrival intensity is nonuniform, the expected insertion times are obtained by means of a standard time change (see, e.g., Daley and Vere-Jones [[45], p. 22]), whence

$$\zeta_i = \Lambda^{-1}\left(\Lambda(\tau_{\ell-1}) + \left(\frac{\Lambda(\tau_\ell) - \Lambda(\tau_{\ell-1})}{n+1}\right) i\right), \quad i \in \{1, \ldots, n\}, \tag{13}$$

using the inverse $\Lambda^{-1}(\cdot)$ of the cumulative rate function,[12]

$$\Lambda(t) = \int_0^t \lambda_T^s\left(\vartheta \,\middle|\, \mathcal{H}_T^s\right) d\vartheta, \quad t \in [0, \tau_\ell], \tag{14}$$

where the intensity $\lambda_T^s(\vartheta|\mathcal{H}_T^s)$ is obtained by means of Eq. (1), for all $\vartheta \in [0, \tau_\ell]$. By construction, the corrected sample history $\hat{\mathcal{H}}_T^s = \mathcal{H}_T^s \cup \{\zeta_1, \ldots, \zeta_n\}$ satisfies the ℓth bin-count constraint, so again $\hat{X}_\ell^s = |[\hat{\mathcal{H}}_T^s]_\ell| = X_\ell$.

(II) Expectation. The idea behind our first *iterative* thickening method is to successively add events at the expected arrival time in the most likely interarrival intervals. For this, consider any interarrival interval $(t_{i-1}^s, t_i^s]$ which has a nonempty intersection with $\mathcal{B}_\ell$, i.e., for all

$$i \in \mathcal{I}_\ell = \left\{i \in \{1, \ldots, K_s\} : \left(t_{i-1}^s, t_i^s\right] \cap \mathcal{B}_\ell \neq \emptyset\right\},$$

where K_s is the number of samples in the current sample history, which is assumed to have been already corrected on all bins to the left of $\mathcal{B}_\ell$ and which may already contain SCs for the current bin under consideration.

[12]For notational convenience, we suppress in Λ the explicit dependence on the sample history $\mathcal{H}_T^s$.

Thus, if

$$i' \in \arg\max_{i\in\mathcal{I}_\ell} \left\{\Lambda\left(\min\left\{\tau_\ell, t_i^s\right\}\right) - \Lambda\left(\max\left\{t_{i-1}^s, \tau_{\ell-1}\right\}\right)\right\} \quad (15)$$

designates an interarrival interval with the highest arrival probability (featuring the largest area under the intensity curve), then one additional sample $\hat{\zeta}$ is added to the interval $(t_{i'-1}^s, t_{i'}^s] \cap \mathcal{B}_\ell$. The new sample is added at the expected arrival time

$$\hat{\zeta} = \mathbb{E}[\zeta | t_{i'-1} < \zeta \leqslant t_{i'}] = \frac{\int_{t_{i'-1}}^{t_{i'}} t\lambda_T^s\left(t\middle|\mathcal{H}_T^s\right)\exp[-(\Lambda(t)-\Lambda(t_{i'-1}))]dt}{1-\exp[-(\Lambda(t_{i'})-\Lambda(t_{i'-1}))]},$$

where $\Lambda(\cdot)$ is the cumulative rate function defined in Eq. (14). Based on an update of the sample history,

$$\mathcal{H}_T^s \leftarrow\!\shortmid \mathcal{H}_T^s \cup \{\hat{\zeta}\},$$

which increments the ℓth bin count X_ℓ^s by 1, we then recompute both the sample-path intensity $\lambda_T^s(\cdot|\mathcal{H}_T^s)$ and the cumulative rate function $\Lambda(\cdot)$ (on $(0, T]$ to be on the safe side) using Eqs. (1) and (14), respectively. The recursion terminates when the bin-count constraint $X_\ell^s = X_\ell$ is satisfied. Then the corrected sample history is obtained by direct assignment: $\hat{\mathcal{H}}_T^s \leftarrow\!\shortmid \mathcal{H}_T^s$, with the corrected bin count $\hat{X}_\ell = X_\ell$, where $\hat{X}_\ell^s \leftarrow\!\shortmid X_\ell^s$.[13]

(III) Exact. Instead of generating a sample at the expected arrival time in the most likely interarrival interval as in method (II), it is also possible to create additional samples $\hat{\zeta}$ via inverse transform sampling. Thus, once the most likely interarrival interval, $(t_{i'-1}, t_{i'}] = (t_{i'-1}^s, t_{i'}^s] \cap \mathcal{B}_\ell$, has been identified according to Eq. (15) (with $t_{i'-1} = \max\{\tau_{\ell-1}, t_{i'-1}^s\}$ and $t_{i'} = \min\{\tau_\ell, t_{i'}^s\}$), an additional arrival is generated based on the cumulative distribution function (CDF) conditional on an observation in the selected interarrival interval within the current sample history $\mathcal{H}_T^s$:

$$G_{i'}^s(t) = \mathbb{P}[\zeta \leqslant t | t_{i'-1} < \zeta \leqslant t_{i'}] = \frac{1-\exp[-(\Lambda(t)-\Lambda(t_{i'-1}))]}{1-\exp[-(\Lambda(t_{i'})-\Lambda(t_{i'-1}))]}, \quad t \in (t_{i'-1}, t_{i'}].$$

For lack of a closed-form inverse, we numerically approximate this sample-based CDF, and subsequently generate $\hat{\zeta}$ via inverse transform sampling. We then proceed as under (II) by updating the sample history, $\mathcal{H}_T^s \leftarrow\!\shortmid \mathcal{H}_T^s \cup \{\hat{\zeta}\}$, recomputing the intensity path $\lambda_T^s(\cdot|\mathcal{H}_T^s)$ in Eq. (1), the arrival probabilities F_i^s in Eq. (12), and the cumulative rate function in Eq. (14). New samples are added in this manner until the ℓth bin-count constraint is satisfied.

(IV) Offspring. A processcentric approach is to base the sample generation on the available knowledge of the law of motion in Eq. (1), and thus to create an additional sample as an offspring using the kernel $\phi(\cdot|\hat{\boldsymbol{\theta}})$ for the current parameter estimate $\hat{\boldsymbol{\theta}}$. The algorithm proceeds as in method (III), using the most likely interarrival interval $(t_{i'-1}, t_{i'}] = (t_{i'-1}^s, t_{i'}^s] \cap \mathcal{B}_\ell$ according to Eq. (15), except that the inverse transform sampling is now based on the CDF for the offspring distribution conditional on $\hat{\boldsymbol{\theta}}$,

$$\hat{G}_{i'}^s(t) = \mathbb{P}[\zeta \leqslant t | t_{i'-1} < \zeta \leqslant t_{i'}] = \frac{1-\exp\left[-\int_{t_{i'-1}}^t \phi(\vartheta|\hat{\boldsymbol{\theta}})\,d\vartheta\right]}{1-\exp\left[-\int_{t_{i'-1}}^{t_{i'}} \phi(\vartheta|\hat{\boldsymbol{\theta}})\,d\vartheta\right]}, \quad t \in (t_{i'-1}, t_{i'}].$$

The latter does not depend directly on the sample history $\mathcal{H}_T^s$, so—depending on the kernel—a computational advantage over method (III) arises when an inverse of the CDF can be obtained explicitly.[14] Samples are generated until the ℓth bin-count constraint holds.

Comparison of the thickening methods. The four thickening methods outlined thus far follow quite different philosophies. Method (I) adds all missing samples at once, based on the time-varying intensity path generated by the sample history $\mathcal{H}_T^s$. Method (II) supplements samples one at a time, each into the most likely interarrival interval, at its expected location, recomputing the intensity path after each such addition. The last two methods generate additional random samples one at a time, based on the intensity path in the most likely interarrival interval [method (III)], or else using the available offspring dynamics based on the current process-parameter estimate [method (IV)]. Figure 4 depicts a situation where two samples need to be added to satisfy the bin-count constraint.

D. Convergence

We now formally establish the convergence of the RISC algorithm, when executed subject to a sample-monotonicity constraint. The added constraint is used to construct a maximizing (joint) sequence of estimators together with corrected-sample histories.[15] In practice, the algorithm converges well without imposing sample monotonicity.

The given reference bin-count vector X distributes the $K = |X|$ binned reference observations into the binning partition $\mathscr{P}_T$. In iteration $k+1$, the RISC algorithm takes the process estimate $\hat{\boldsymbol{\theta}}^k$ from the previous iteration k to simulate $\mathrm{HP}(\hat{\boldsymbol{\theta}}^k)$ so as to obtain a sample history $\mathcal{H}_T^{s,k+1}$, which is then corrected to $\hat{\mathcal{H}}_T^{s,k+1}$ via thinning and thickening on each bin in the given partition $\mathscr{P}_T$. By construction, the bin-count constraint,

$$\hat{\mathcal{H}}_T^{s,k+1}\big/\mathscr{P}_T = X,$$

remains therefore satisfied for all $k \geqslant 0$. The log-likelihood function $\mathcal{L}(\hat{\boldsymbol{\theta}}|\hat{\mathcal{H}}_T^{s,k+1})$ is a continuous function of the $P+K$ real-valued variables consisting of the P-dimensional process-parameter estimate $\hat{\boldsymbol{\theta}}$ and the K-dimensional sample vector $(\hat{t}_1^{s,k+1}, \ldots, \hat{t}_K^{s,k+1})$, containing the time-ordered elements of the corrected sample history $\hat{\mathcal{H}}_T^{s,k+1}$.

[13] In case $X_\ell = 0$, we place a uniform sample in the time interval $(\tau_{\ell-1}, \tau_\ell]$ to avoid potential edge effects biasing the estimation.

[14] Consider the exponential kernel $\phi(\vartheta|\hat{\beta}) = \hat{\beta}\exp(-\hat{\beta}\vartheta)$ for $\vartheta \geqslant 0$. Then $\hat{G}_{i'}^s(t) = y \in (0, 1]$ implies that $t = -(1/\hat{\beta})\ln[-\ln(1-\varkappa y) - \exp(-\hat{\beta}t_{i'-1})] \in (t_{i'-1}, t_{i'}]$, where $\varkappa = 1 - \exp[-\int_{t_{i'-1}}^{t_{i'}} \phi(\vartheta|\hat{\beta})\,d\vartheta]$.

[15] The concept of a maximizing sequence is used here as in the calculus of variations [[46], pp. 193–195].

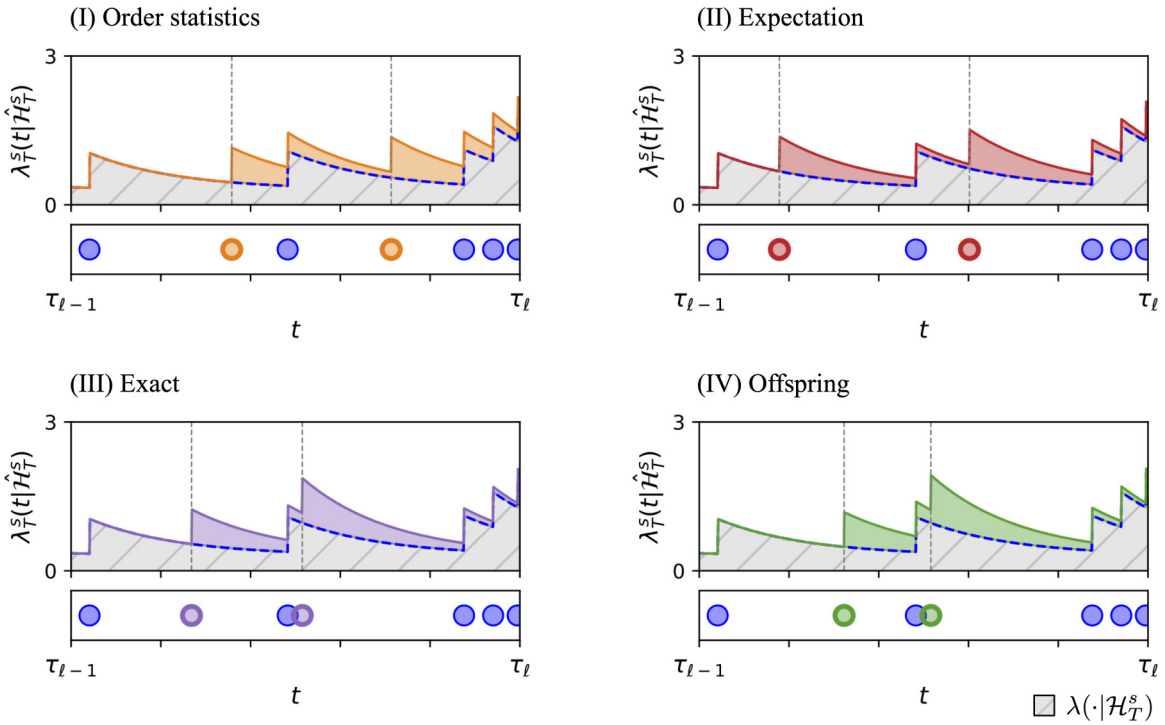


FIG. 4. Thickening process. In bin $\mathcal{B}_\ell = (\tau_{\ell-1}, \tau_\ell]$, two events have to be added (i.e., $X_\ell^s - X_\ell = -2$). For our proposed thickening techniques (I)–(IV), we illustrate the change in the intensity function from $\lambda_T^s(\cdot|\mathcal{H}_T^s)$ to $\lambda_T^s(\cdot|\hat{\mathcal{H}}_T^s)$, depending on the locations of the sample-correction events; the latter are either deterministic (I), (II) or stochastic (III), (IV).

Since by assumption the kernel function is bounded, that is, there exists a finite constant $J > 1/K$ such that $\phi(t) \leqslant J$ for all $t \geqslant 0$, we can conclude by Eq. (1) that the intensity is bounded, as $\lambda(t) \leqslant \mu + KJ$. But this implies that the log likelihood is bounded as well:[16]

$$\mathcal{L}\big(\hat{\boldsymbol{\theta}}\big|\hat{\mathcal{H}}_T^{s,k+1}\big) \leqslant \mathcal{L}^{k+1} \triangleq \mathcal{L}\big(\hat{\boldsymbol{\theta}}^{k+1}\big|\hat{\mathcal{H}}_T^{s,k+1}\big) \leqslant K\ln(\mu + KJ), \quad \hat{\boldsymbol{\theta}} \in \Theta. \tag{16}$$

To obtain monotonicity in the maximized log likelihoods, in the sense that

$$\mathcal{L}^k = \mathcal{L}\big(\hat{\boldsymbol{\theta}}^k|\hat{\mathcal{H}}_T^{s,k}\big) \leqslant \mathcal{L}\big(\hat{\boldsymbol{\theta}}^{k+1}\big|\hat{\mathcal{H}}_T^{s,k+1}\big) = \mathcal{L}^{k+1}, \tag{17}$$

it is enough to subject the corrected history $\hat{\mathcal{H}}_T^{s,k+1}$ [obtained from the sample history $\mathcal{H}_T^{s,k}$ of HP($\boldsymbol{\theta}^k$)] to a sample-monotonicity constraint of the form

$$\mathcal{L}\big(\hat{\boldsymbol{\theta}}^k\big|\hat{\mathcal{H}}_T^{s,k+1}\big) \geqslant \mathcal{L}^k = \mathcal{L}\big(\hat{\boldsymbol{\theta}}^k\big|\hat{\mathcal{H}}_T^{s,k}\big). \tag{18}$$

The sample-monotonicity constraint (18) and the first inequality in Eq. (16) together imply the likelihood monotonicity (17). In addition, Eqs. (16) and (17) yield, by the monotone convergence theorem, that there exists a limit $\mathcal{L}^*$ in the interval $(0, K\ln(\mu + KJ)]$ such that

$$\lim_{k\to\infty} \mathcal{L}^k = \mathcal{L}^*.$$

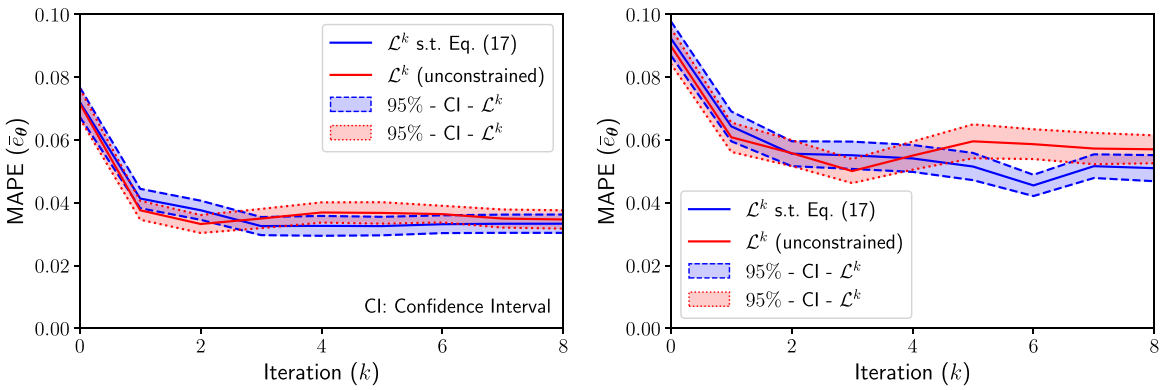
Since $\mathcal{L}$ is an upper semicontinuous function on the compact set $\Theta \times \mathcal{T}_K$, with $\mathcal{T}_K = \{(t_1, \ldots, t_K) \in \mathbb{R}_+^K : t_1 \leqslant \cdots \leqslant t_K\}$, by the Weierstrass theorem there exists a parameter vector $\hat{\boldsymbol{\theta}}^*$ and a (possibly degenerate) history $\hat{\mathcal{H}}_T^{s,*}$ in $\mathcal{T}_K$ so $\mathcal{L}(\hat{\boldsymbol{\theta}}^*|\hat{\mathcal{H}}_T^{s,*}) = \mathcal{L}^*$. In this context, a history is called degenerate if it contains samples that coincide with others. Since simulated histories cannot produce degenerate histories, the maximum of the log likelihood (i.e., $\mathcal{L}^*$) can, in general, only be approximated.

The remaining question is whether the sample-monotonicity constraint (18) can be satisfied by a corrected sample $\hat{\mathcal{H}}_T^{s,k+1}$ of HP($\hat{\boldsymbol{\theta}}^k$) with positive probability. In other words, when repeating the sampling/correction procedure as many times as needed to satisfy the constraint, would this process stop with positive probability after finitely many iterations? That this question has a positive answer can be seen as follows. Indeed, note first that Eq. (18) is satisfied trivially if $\hat{\mathcal{H}}_T^{s,k+1} = \hat{\mathcal{H}}_T^{s,k}$, which in itself may happen with probability zero. However, if we divide an open set $\mathcal{N} \subset \mathcal{T}_K$, which contains $\hat{\mathcal{H}}_T^{s,k}$, into

$$\mathcal{N}_1 = \{\mathcal{H}_T \in \mathcal{N} : \mathcal{L}(\hat{\boldsymbol{\theta}}^k|\mathcal{H}_T) \geqslant \mathcal{L}^k\},$$

and its complement $\mathcal{N}_0 = \mathcal{N} \setminus \mathcal{N}_1$, then we know by assumption that $\mathcal{N}_1$ is nonempty. Moreover, because small variations

[16]By Eq. (1), this bound is tight. For example, the exponential kernel $\phi(t|\beta) = \beta\exp(-\beta t) \leqslant \beta$, for $t \geqslant 0$, can produce intensities arbitrarily close to $\mu + K\beta$ when the sample history $\mathcal{H}_T$ (with $|\mathcal{H}_T| = K$) is concentrated near $t = 0$. Indeed, if for $\varepsilon \in (0, T)$ it is $\mathcal{H}_T \subset [0, \varepsilon]$, then for $\varepsilon \to 0^+$ one would obtain $\lambda(\varepsilon) \to \mu + K\beta$.



(a) $\|\mathscr{P}_T\| = 3$. (b) $\|\mathscr{P}_T\| = 10$.

FIG. 5. Mean absolute percentage error (MAPE) per iteration k.

of $\hat{\mathcal{H}}_T^{s,k}$ produce variations in the log-likelihood function,[17] by the local differentiability of $\mathcal{L}$ in any arrival time t_i (as element of a nondegenerate sample history), the probability measure of $\mathcal{N}_1$ must be greater than zero, so indeed the sample-monotonicity constraint must be satisfied with positive probability.

We can therefore conclude that the RISC algorithm, executed subject to the sample-monotonicity constraint (18), converges in finite time, in the sense that $\hat{\boldsymbol{\theta}}^k$ approximates a maximum-likelihood parameter vector $\hat{\boldsymbol{\theta}}^*$ up to any given error in a finite number of iterations with a finite amount of oversampling.[18]

Figure 5 illustrates the convergence of the Cauchy difference of successive parameter estimates[19] in terms of the mean absolute percentage error (MAPE) introduced in Sec. III A. As a function of the iteration k, the error decreases at first significantly and then levels off, indicating that the simulation noise floor has been reached. Coarser bin partitions suffer from a higher noise floor due to the additional loss of entropy. Lastly, we note that in practice the sample-monotonicity constraint can usually be ignored without a noticeable decrease in the speed of convergence.

III. PERFORMANCE ASSESSMENT

To analyze the usefulness of the RISC method, we introduce two simple evaluation criteria, discuss the setup of our numerical study, and then compare the method performance; first, internally—across our four different SC variants—and then externally with respect to other extant methods for the estimation of time-censored point processes.

A. Evaluation criteria

To quantify the performance of the proposed RISC estimator, we use the (expected) MAPE for tracking deviations of the full parameter estimate $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_P)$ relative to a reference vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_P)$,

$$e(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta}) = \frac{1}{P}\sum_{p=1}^{P} \frac{|\hat{\theta}_p - \theta_p|}{|\theta_p|}, \quad \hat{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \mathbb{R}_{++}^P, \tag{19}$$

as well as the (signed) statistical bias,

$$B(\hat{\theta}_p|\theta_p) = \hat{\theta}_p - \theta_p, \quad \hat{\theta}_p, \theta_p \in \mathbb{R}_{++}, \tag{20}$$

to track systematic deviations of any individual component $\hat{\theta}_p$, for $p \in \{1, \ldots, P\}$, of the estimate from the underlying true value θ_p. To isolate these relative and absolute measures from fluctuations in reference bin counts and randomness stemming from SC, the *expected* performance measures are obtained from Eqs. (19) and (20) as averages over N different instances,

$$\bar{e}_{\boldsymbol{\theta}} = \frac{1}{N}\sum_{n=1}^{N} e(\hat{\boldsymbol{\theta}}^{(n)}|\boldsymbol{\theta}) \quad \text{and} \quad \bar{B}_{\hat{\theta}_p} = \frac{1}{N}\sum_{n=1}^{N} B\big(\hat{\theta}_p^{(n)}|\theta_p\big), \tag{21}$$

for $p \in \{1, \ldots, P\}$, where each estimated parameter vector $\hat{\boldsymbol{\theta}}^{(n)}$ belongs to a bin-count vector $X^{(n)}$, derived from realization $n \in \{1, \ldots, N\}$ of the reference process HP($\boldsymbol{\theta}$).

B. Numerical study: Setup

In our numerical analysis, we concentrate on Hawkes processes with an exponential kernel, characterized by the

[17] Differentiating the log likelihood $\mathcal{L}(\hat{\boldsymbol{\theta}}|\mathcal{H}_T)$ in Eq. (3) with respect to any $t_i \in \mathcal{H}_T \in \mathcal{N} \subset \mathcal{T}_K$ yields

$$\frac{\partial \mathcal{L}(\hat{\boldsymbol{\theta}}|\mathcal{H}_T)}{\partial t_i} = \frac{\sum_{j:t_j<t_i} \phi'(t_i - t_j)}{\lambda(t_i|\mathcal{H}_{t_i})} - \sum_{l=i+1}^{K} \frac{\phi'(t_l - t_i)}{\lambda(t_l|\mathcal{H}_{t_l})} + \phi(T - t_i).$$

Since this derivative does not vanish in any small open neighborhood of $\hat{\mathcal{H}}_T^{s,k}$ (for any nontrivial Hawkes process) conditional on having obtained $\hat{\mathcal{H}}_T^{s,k+1} \in \mathcal{N}$ (which constitutes a positive-probability event), we can conclude that the probability of having attained the global maximum in the previous RISC-iteration k is zero, and correspondingly the probability of having exceeded the preceding log-likelihood value ($\mathcal{L}^k$) is positive.

[18] While we have established the convergence of the monotonic likelihood sequence, and by continuity the sequence of parameter estimates, there is no guarantee with respect to reaching a global optimum due to the inherent nonconvexity of the likelihood function.

[19] Since the Euclidean space $\mathbb{R}^P$ is complete, any Cauchy sequence converges.

TABLE I. Parameter configurations for the time-censored identification of HP($\boldsymbol{\theta}_m$).

$\boldsymbol{\theta}_m$	μ_m	α_m	β_m
$\boldsymbol{\theta}_0$	1.0	0.0	0.0
$\boldsymbol{\theta}_1$	0.4	0.6	0.5
$\boldsymbol{\theta}_2$	0.4	0.6	1.5
$\boldsymbol{\theta}_3$	0.1	0.9	0.5
$\boldsymbol{\theta}_4$	0.1	0.9	1.5

parameter vectors $\boldsymbol{\theta}_m = (\mu_m, \alpha_m, \beta_m)$, for $m \in \{1, \dots, 4\}$. As a reference, we include a pure Poisson process parameter vector for $m = 0$, which can also be viewed as a Hawkes process HP($\boldsymbol{\theta}_0$), for $\boldsymbol{\theta}_0 = (\mu_0, 0, 0)$; see Table I. To match the expected number of arrivals across experiments, the different process scenarios are chosen to feature (approximately) the same expected average arrival rate $\bar{\lambda} \approx \mu/(1-\alpha) = 1$.[20] In other words, the expected number of arrivals over the observation interval $(0, T]$ of length $T = 1000$ is about 1000 for any of the five parameter configurations. For every $\boldsymbol{\theta}_m$, we identify $N = 1000$ synthetically generated Hawkes processes that are time-censored on a uniform binning partition $\mathscr{P}_T$ (cf. Remark 1) of relative norm $\Delta \in [0.1\%, 2\%]$. Each reference sample history $\mathcal{H}_{T,m}^{s,n}$, for $(m, n) \in \{1, \dots, 4\} \times \{1, \dots, N\}$, is obtained via simulation using the Python library TICK [47], which is based on Ogata's thinning algorithm [48]. By iteration of SC, identification, and sample generation, the RISC algorithm produces the corresponding estimators $\hat{\boldsymbol{\theta}}_{\mathrm{RISC}}^{(m,n)}$, using a suitable ϵ-convergence criterion; see Fig. 2 and Appendix B 2 for further details.

Figure 6 illustrates the bin-to-bin variability of HP($\boldsymbol{\theta}_m$) in terms of three summary statistics. First, we consider the bin-to-bin standard deviation,

$$\sigma_m = \left[\frac{1}{L \cdot N} \sum_{\ell,n=1}^{L,N} \left(X_\ell^{(m,n)} - \frac{K^{(m,n)}}{L} \right)^2 \right]^{1/2},$$

where $X^{(m,n)} = (X_1^{(m,n)}, \dots, X_L^{(m,n)})_{\ell=1}^L$ is the bin-count vector associated with simulation run (m, n), and $K^{(m,n)} = \sum_{\ell=1}^L X_\ell^{(m,n)} = |\mathcal{H}_{T,m}^{s,n}|$ records the number of observed samples across the $L = |\mathscr{P}_T| \approx \lfloor T/\Delta \rfloor$ bins. Correspondingly,

$$\mathrm{CV}_m = \frac{(L \cdot N)\, \sigma_m}{\sum_{n=1}^N K^{(m,n)}}$$

is the (measured) bin-to-bin coefficient of variation for HP($\boldsymbol{\theta}_m$). Finally,

$$H_m = \frac{1}{N} \sum_{n=1}^N H(X^{(m,n)})$$

denotes the average of the run-specific bin-count entropy as introduced in Eq. (9). The parameter vectors $\boldsymbol{\theta}_m$ are such that each variability measure is monotonic in $m \in \{0, \dots, 4\}$; a higher m thus indicates a higher degree of differentiation of HP($\boldsymbol{\theta}_m$) from the Poisson reference, HP($\boldsymbol{\theta}_0$). Specifically, the parameter vector $\boldsymbol{\theta}_1$ describes a process similar to a Poisson process with a low bin-to-bin standard deviation and coefficient of variation. By contrast, HP($\boldsymbol{\theta}_4$) is characterized by the highest bin-to-bin standard deviation and coefficient of variation, as a result of the distinctive clusters formed by the near-critical branching ratio α_4 (close to 1), and the relatively fast kernel decay rate β_4. We consider HP($\boldsymbol{\theta}_1$) and HP($\boldsymbol{\theta}_4$) as representative extreme cases in the spectrum of self-exciting point processes, with close-to-Poisson behavior on the one hand and near-critical behavior on the other. The simulation studies below (in Secs. III C–III D) focus on those cases.[21]

Remark 3 (Interrun versus intrarun variability). For a given Hawkes process (with unobserved parameter vector), the RISC estimate $\hat{\boldsymbol{\theta}}_{\mathrm{RISC}}$ fluctuates, since (i) the observed bin-count vector X changes across runs (i.e., across different sample paths) and (ii) even for a fixed X there is randomness built into the RISC algorithm, thus producing slightly varying estimates based on the same input data. As shown in Appendix B 4, the interrun variability, relating to (i), significantly exceeds the intrarun variability, relating to (ii), no matter which particular SC method (I)–(IV) is used.[22] Naturally, the RISC estimates $\hat{\boldsymbol{\theta}}_{\mathrm{RISC}}$ obtained from stochastic sample-correction methods (III) and (IV) are more variable, with a larger inter- and intrarun variability than when using deterministic SC methods (I) and (II); note that the latter also produces stochastic estimates because of the intermittent simulations contained in the RISC algorithm.

C. Comparison of sample-correction methods

SC forms an integral part of the RISC method in Sec. II, developed for the estimation of the parameter vector $\boldsymbol{\theta}$ that characterizes the point process in Eq. (1), subject to time-censored observation. As shown below, the performance of the different SC algorithms in Sec. II C tends to decrease with stronger time censoring. That is, when the norm of the binning partition increases, both the relative error (MAPE) and the expected absolute deviation (measured as the absolute value of the statistical bias) go up.

In addition, the performance of the different SC algorithms depends on the degree of self-excitingness in the underlying process. For example, $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_4$ in Table I illustrate, respectively, both extremes of the spectrum of similarity to a standard Poisson process. The inclusion of a naïve uniform SC method highlights the effect. For HP($\boldsymbol{\theta}_1$), uniform resampling according to the bin-count constraint in Eq. (11) outperforms all of the nontrivial methods, as the implicit bias introduced by the uniformity of the samples happens to work in the right (i.e., Poisson) direction. By contrast, for HP($\boldsymbol{\theta}_4$), uniform resampling comes last due to the significant self-excitation behavior of the process, thus illustrating the benefits of the

[20] See Appendix C 2 for details on how to compute the expected average arrival rate $\bar{\lambda}$ for HP($\boldsymbol{\theta}$) on $(0, T]$.

[21] The corresponding results for the intermediate HP($\boldsymbol{\theta}_2$) and HP($\boldsymbol{\theta}_3$) are reported in Appendix B 6.

[22] By fixing the seed of a random number generator, it is possible to eliminate intrarun variability altogether. Yet, it must be accounted for, at the very least, across different implementations of the same algorithm.

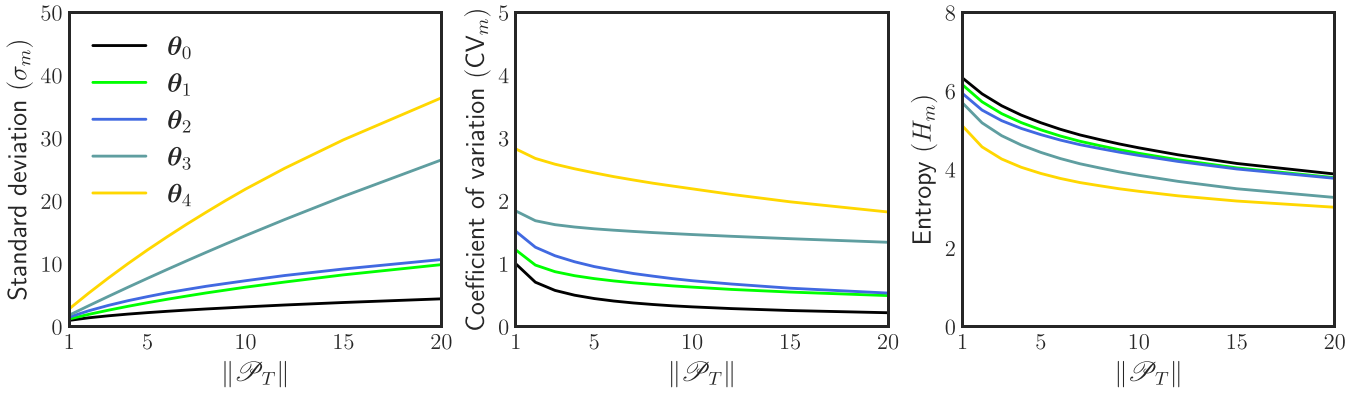


FIG. 6. Bin-to-bin standard deviation σ_m, coefficient of variation CV_m, and entropy H_m for HP($\boldsymbol{\theta}_m$) on a uniform binning partition with norm $\|\mathscr{P}_T\| \in [1, 20]$.[23]

intrabin emulation of nonhomogeneous process characteristics when performing SC.

Estimation performance proves quite sensitive to the coarseness of the binning partition. Specifically, consider first HP($\boldsymbol{\theta}_1$), a Poisson-like process. When time censoring is weak (i.e., for $\|\mathscr{P}_T\| = 1$), all SC methods, including the uniform reference, exhibit very similar performance and identify parameters fairly correctly ($\hat{\mu}_1 \approx \mu_1$, $\hat{\alpha}_1 \approx \alpha_1$), with a slight negative bias for the decay coefficient $\hat{\beta}_1 < \beta_1$; see Fig. 7(a). When time censoring becomes more severe, e.g., through merging blocks of the current (daily) bins into more aggregate (weekly) observations (i.e., for $\|\mathscr{P}_T\| = 7$), differences in the resulting estimation errors begin to emerge.

Remark 4 (Noise floor). Given a finite observation horizon, even an MLE estimator without time censoring (cf. Sec. II A 2) produces a persistent estimation error, which therefore serves as a noise floor, indicated in Figs. 8 and 11 below.[25]

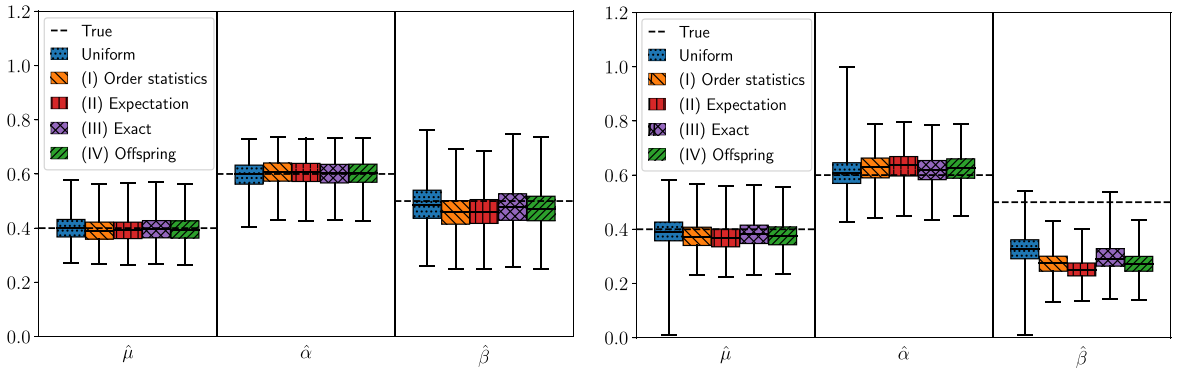
While the estimation quality of the background rate $\hat{\mu}_1$ and the branching coefficient $\hat{\alpha}_1$ is only moderately affected by time censoring, the identification of the decay rate $\hat{\beta}_1$ may be subject to a significant bias; see Fig. 7(b). As bins are merged, the self-excitation behavior in the form of separated clusters is increasingly masked. Therefore, all previously introduced estimation algorithms struggle with the identification, especially of the decay rate; cf. Sec. III D. Stochastic methods (III) and (IV) provide weakly better estimation results, as these methods use, by design, more exploration than the deterministic methods (I) and (II). The naïve uniform SC method displays the best performance, in terms of lowest estimation bias, while at the same time exhibiting the largest spread. The MAPE for $\boldsymbol{\theta}_1$ is dominated by the estimation bias for β_1 and increases monotonically with the degree of time censoring; see Figs. 8 and 9.

For near-critical Hawkes processes, such as HP($\boldsymbol{\theta}_4$), all SC methods provide similar performance results, as long as

[23] Appendix B 1 provides some additional relevant statistics for the different processes.

[24] The boxplots, one for each method in Sec. II, illustrate the parameter estimates. Boxplot whiskers range from minimum to maximum; the box marks the range between the 25th and the 75th percentile (i.e., the second and third quartiles); the median is indicated by a horizontal line within the box. The (dashed) true-parameter line serves as a reference.

[25] The noise floor is also used as a reference in the following sections, e.g., in Figs. 14 and 17.



(a) $\|\mathscr{P}_T\| = 1$. (b) $\|\mathscr{P}_T\| = 7$.

FIG. 7. Estimates $\hat{\mu}, \hat{\alpha}, \hat{\beta}$ of $(\mu_1, \alpha_1, \beta_1) = (0.4, 0.6, 0.5)$.[24]

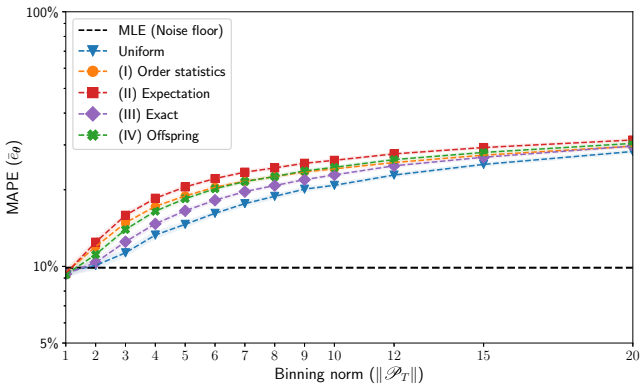


FIG. 8. MAPE $\bar{e}_{\boldsymbol{\theta}}$ for $(\mu_1, \alpha_1, \beta_1) = (0.4, 0.6, 0.5)$.

time censoring remains weak. As the fast decay of $\beta_4 = 1.5$ is difficult to observe, the decay estimates are negatively biased even with a relatively fine binning partition; see Fig. 10(a). Further censoring increases the negative bias; see Fig. 10(b). In comparison with uniform SC, the more sophisticated methods (I)–(IV) provide better performance—in particular for the identification of the decay rate. This is a result of the SC techniques that reconstruct an intrabin arrival history coherent with the self-excitingness of the process. The exact method (III), which corrects the simulation path using the exact conditional intensity function for each interarrival time, exhibits the smallest bias for the decay rate.

The MAPE increases in the severity of time censoring; see Fig. 11. The stochastic methods, in particular (III), show superior performance. The overall estimation error is largely driven by the somewhat unavoidable large bias for the decay rate; see Fig. 12.

Self-exciting behavior is a key feature of Hawkes processes such as HP($\boldsymbol{\theta}_4$). Hence, the reconstruction of an intrabin arrival history consistent with process characteristics is important to mitigate estimation bias—especially for the decay parameter. Focusing on the better-performing stochastic SC methods (III) and (IV), we proceed below with comparing the proposed RISC algorithm against extant estimation algorithms for the identification of time-censored Hawkes processes.

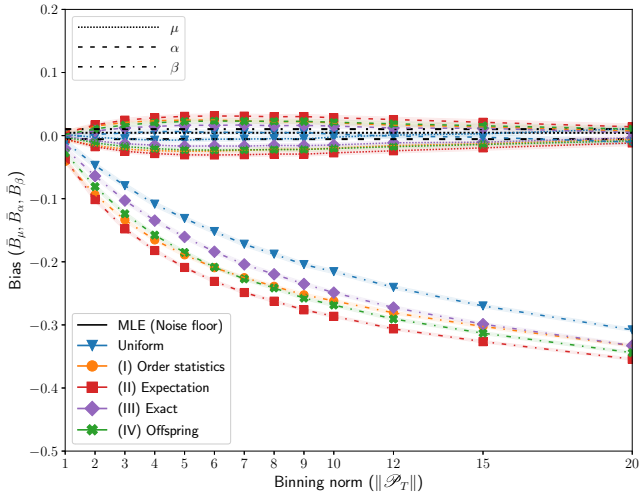


FIG. 9. Bias $\bar{B}_\mu, \bar{B}_\alpha, \bar{B}_\beta$ for $(\mu_1, \alpha_1, \beta_1) = (0.4, 0.6, 0.5)$.

TABLE II. Algorithm features.[26]

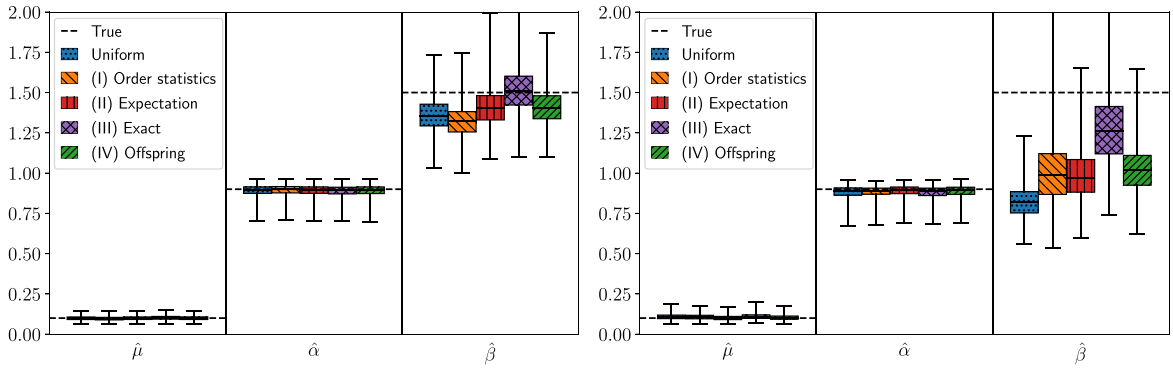
Algorithm	RI	SC
RISC	✓	✓
BH-EM	✓	(✓)
INAR	✗	✗
Whittle	✗	✗

D. Comparison to extant algorithms

There are various algorithms for the identification of self-exciting point processes based on time-censored observations [20,22,25,28–30]; see also our discussion in Sec. I A. Specifically, we compare the RISC algorithm to the BH-EM approach by Shlomovich *et al.* [30], the integer-valued autoregressive time series (INAR) approach by Kirchner [22], and the spectral estimation approach (Whittle) by Cheysson and Lang [25]. As shown in Table II, of the three comparison benchmarks, only the BH-EM algorithm incorporates both recursive estimation and a variation of SC, the latter via successive bin-specific global maximization of a conditional likelihood function. The other two algorithms use a discrete-time (i.e., piecewise constant) approach (INAR), adapting the timescale to the bins, and a spectral-estimation approach (Whittle), respectively. Both of these restrict attention to uniform sampling partitions. For both algorithms (INAR and Whittle), the authors proved that the estimators are consistent and asymptotically normal. However, in the case of INAR, consistency obtains only when the binning norm $\|\mathscr{P}_T\|$ tends towards zero. As we consider coarser time-censored observations, consistency cannot be guaranteed. The strong mixing conditions, a prerequisite for the Whittle algorithm, are satisfied for the exponential kernel in Eq. (2). This implies estimation consistency for $T \to \infty$, assuming that the process remains stationary. Since the BH-EM and the RISC algorithm both use MLE, which is consistent (and asymptotically normal) for uncensored Hawkes processes, their estimates are also asymptotically consistent for $\Delta \to 0^+$.

Comparing the performance of the three alternative approaches to that of the RISC algorithm [with the stochastic sample-correction methods (III) and (IV)], we find that algorithms omitting the sample correction (i.e., those omitting the reconstruction of an event history $\hat{\mathcal{H}}_T$ according to the bin-count observation X) exhibit an elevated estimation error and bias. Interestingly, in our simulation study these estimation issues could be observed even for the smallest considered relative binning norm $\Delta = 0.1\%$ (corresponding to $\|\mathscr{P}_T\| = 1$). In addition, near-critical processes [such as HP($\boldsymbol{\theta}_4$)] demonstrate a further challenge to these algorithms.

[26]The BH-EM algorithm generates samples in $\mathcal{B}_\ell$ by solving an optimization problem, successively for all ℓ. By comparison, the RISC algorithm *corrects* a synthetic sample history in each $\mathcal{B}_\ell$ using thinning or thickening.



(a) $\|\mathscr{P}_T\| = 1$. (b) $\|\mathscr{P}_T\| = 7$.

FIG. 10. Estimates $\hat{\mu}, \hat{\alpha}, \hat{\beta}$ for $(\mu_4, \alpha_4, \beta_4) = (0.1, 0.9, 1.5)$.

While the BH-EM algorithm generates a sample history in a similar spirit as the proposed RISC algorithm, it may exhibit convergence/precision issues, since for every bin a nonconvex optimization problem needs to be solved. Figure 13 depicts significant deviations in the Whittle estimator from the true parameter values, particularly for the decay rate (corresponding to $\hat{\beta}$).

As time censoring becomes more severe, the performance of all algorithms decreases significantly. For the Poisson-like process HP($\boldsymbol{\theta}_1$) the extant solutions tend to produce rather high-variance estimates. Several approaches exhibit convergence issues; see Fig. 13. For $\|\mathscr{P}_T\| \leqslant 8$, the BH-EM algorithm overestimates μ (and thus underestimates α) before converging; see Fig. 14.

The advantage of recreating a statistically consistent sample history can be gleaned from Fig. 14, in terms of MAPE. Additionally, algorithms with SC show lower performance when identifying the background rate μ, and the self-excitation dynamics α.[27] Whittle and INAR estimates for $\hat{\mu}$ deteriorate with increasing time censoring showing signs of parameter overestimation (positive bias). This tendency to overestimate $\hat{\mu}$ results also in an estimation bias for the branching coefficient $\hat{\alpha}$, as both parameters relate to the number of events generated; see Fig. 15 and Appendix C 2.

Consider now the near-critical process HP($\boldsymbol{\theta}_4$), with $\|\mathscr{P}_T\| = 1$. Figure 16 shows that estimates exhibit a rather large spread. The performance of both INAR and Whittle declines fast in the degree of time censoring; see Fig. 17. On the other hand, the RISC algorithm achieves superior results—in terms of MAPE and bias—for most of the binning norms under consideration; see Figs. 17 and 18.

In general, the difficulty of estimating a time-censored Hawkes process depends on both the binning norm and the process (in terms of $\boldsymbol{\theta}$). The average cluster-length ratio $\xi = (\mu/\beta)/(1-\alpha)$ (cf. Appendix C 1) indicates that HP($\boldsymbol{\theta}_1$), with $\xi = 2$, has overlapping clusters. On the other hand, HP($\boldsymbol{\theta}_4$), with $\xi = 2/3$, is more difficult to estimate in the absence of time censoring [cf. the noise floors in Figs. 14 and 17]. To effectively identify Hawkes-process characteristics, taking intrabin offspring behavior into account is essential. For example, when $\|\mathscr{P}_T\| = 7$, the process HP($\boldsymbol{\theta}_1$), conditional on at least one arrival, features a parent and offspring in the

[27] Events caused by self-excitation describe the *endogenous* behavior of the process, while the background rate μ captures the influence of *exogenous* factors. To reliably classify the origin of events is essential for numerous applications, such as for the analysis of trading activity in cryptocurrency markets [49].

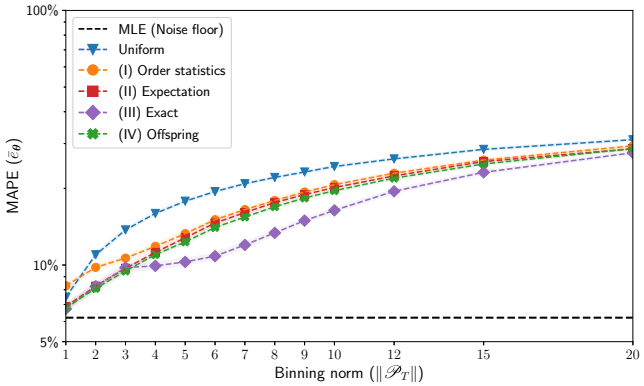


FIG. 11. MAPE $\bar{e}_{\boldsymbol{\theta}}$ for $(\mu_4, \alpha_4, \beta_4) = (0.1, 0.9, 1.5)$.

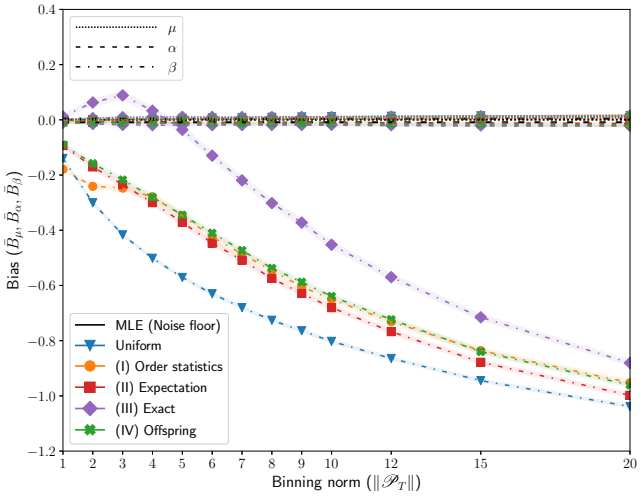
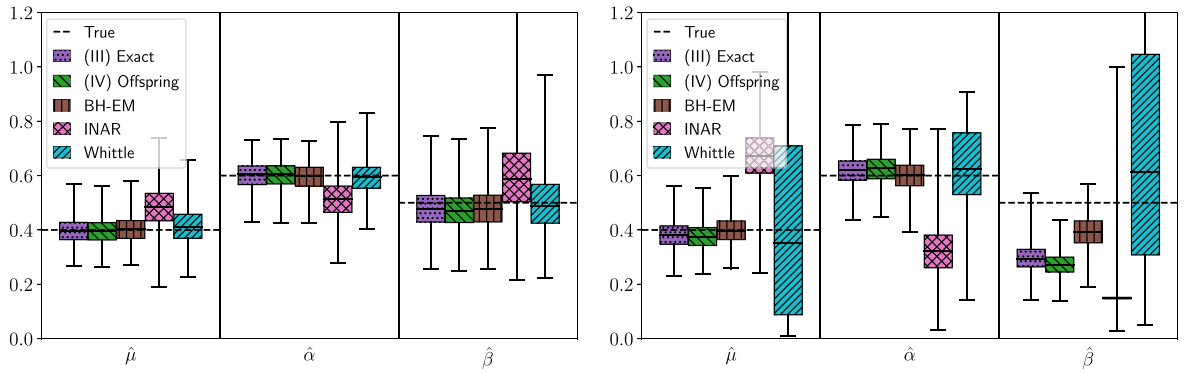


FIG. 12. Bias $\bar{B}_\mu, \bar{B}_\alpha, \bar{B}_\beta$ for $(\mu_4, \alpha_4, \beta_4) = (0.1, 0.9, 1.5)$.



(a) $\|\mathscr{P}_T\| = 1$. (b) $\|\mathscr{P}_T\| = 7$.

FIG. 13. Estimates $\hat{\mu}, \hat{\alpha}, \hat{\beta}$ for $(\mu_1, \alpha_1, \beta_1) = (0.4, 0.6, 0.5)$.

same bin with no less than 72% probability. For HP($\boldsymbol{\theta}_4$), this probability increases to 90%, meaning that censoring tends to mask the self-excitingness of a near-critical process.[28]

IV. DISCUSSION

We are now ready to examine the general usefulness of the RISC algorithm based on additional aspects, such as the sensitivity of performance to data volume, the algorithm's computational complexity, the possibility of nonuniform (or random) time censoring, as well as the relation to interesting practical applications involving a somewhat inverse version of the time-censored estimation problem where the censoring can become a means to achieve an objective such as energy efficiency.

A. Small-sample versus large-sample behavior

The estimation approaches for time-censored Hawkes processes discussed in Sec. III D (with the exception of the Whittle approach) are consistent and asymptotically normal only as the norm of the binning partition tends to zero, i.e., for $\|\mathscr{P}_T\| \to 0^+$. As the time censoring disappears, one recovers the standard problem of estimating process characteristics from an uncensored event history, which can be effectively addressed using standard parameter estimation (cf. Sec. II A 2). For all uniform binning partitions, there remains a bias that monotonically increases with the level of time censoring. The Whittle estimator is consistent and hence converges to the true estimates for a sufficiently large time horizon T. Yet, as shown in our numerical experiments, to benefit from the consistency of the Whittle estimator for Hawkes processes with exponential kernel, the required amount of data can be substantial, to the point that it may be practically infeasible to obtain sufficient observations for the estimation error to even become manageable.[29] The sample size in many applications is limited (and may not even be a choice variable).

[28]The probability that an immigrant at $t_j \in \mathcal{B}_\ell$ causes an offspring in the same bin is $\beta \int_{t_j}^{\tau_\ell} e^{-\beta(t-t_j)}\,dt$. Hence, conditional on an immigrant arrival in $\mathcal{B}_\ell$, the probability of parent and at least one offspring to appear in $\mathcal{B}_\ell$ becomes $[\int_{\tau_{\ell-1}}^{\tau_\ell}(\beta \int_{\vartheta}^{\tau_\ell} e^{-\beta(t-\vartheta)}\,dt)d\vartheta]/(\tau_\ell - \tau_{\ell-1}) = 1 - [1 - e^{-\beta(\tau_\ell - \tau_{\ell-1})}]/[(\tau_\ell - \tau_{\ell-1})\beta]$.

[29]For example, even by increasing the time horizon T from 1000 to 8000, the Whittle estimator—while showing slight improvement—still exhibits substantially larger estimation errors than the RISC algorithm; cf. Appendix B 8.

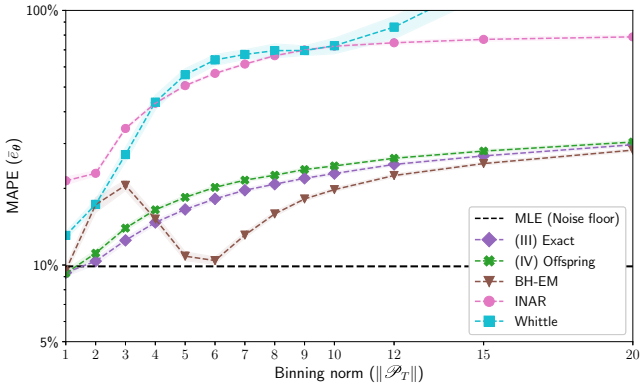


FIG. 14. MAPE $\bar{e}_{\boldsymbol{\theta}}$ for $(\mu_1, \alpha_1, \beta_1) = (0.4, 0.6, 0.5)$.

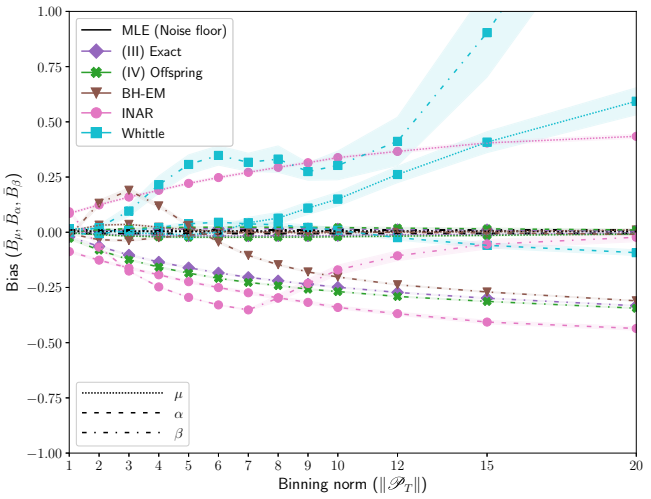
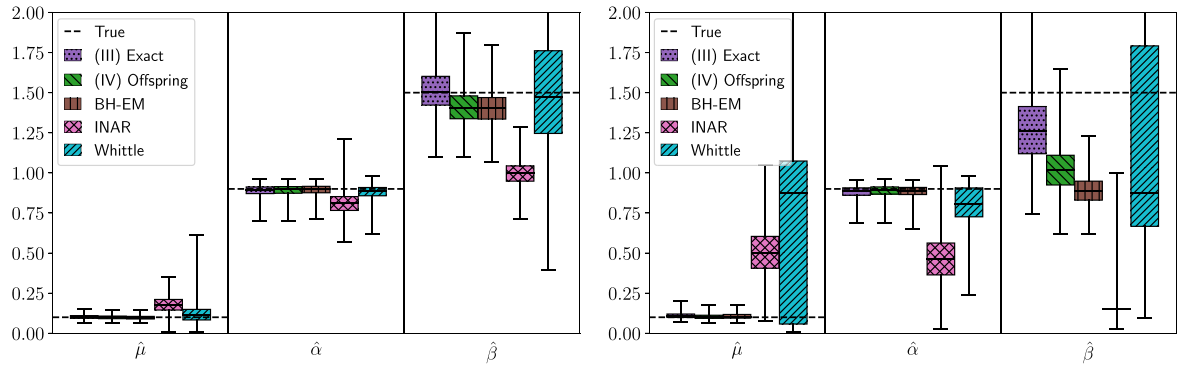


FIG. 15. Bias $\bar{B}_\mu, \bar{B}_\alpha, \bar{B}_\beta$ for $(\mu_1, \alpha_1, \beta_1) = (0.4, 0.6, 0.5)$.



(a) $\|\mathscr{P}_T\| = 1$. (b) $\|\mathscr{P}_T\| = 7$.

FIG. 16. Estimates $\hat{\mu}, \hat{\alpha}, \hat{\beta}$ for $(\mu_4, \alpha_4, \beta_4) = (0.1, 0.9, 1.5)$.

Moreover, large-sample consistency depends on the long-term stationarity of the underlying process, which in practice can be assumed only rarely. For example, the estimation of Hawkes processes in cryptocurrency markets is prone to intraday regime changes [49].[30]

B. Runtime considerations

The numerical study in Sec. III suggests that the proposed RISC algorithm is capable of outperforming other methods in terms of MAPE and bias, at least when the amount of available data is limited. From a computational complexity viewpoint, considering runtime as an evaluation criterion, both INAR and Whittle clearly outperform the other approaches. While INAR features a linear runtime according to $O(L)$, Whittle has an algorithmic time complexity of $O(L\ln(L))$, where L refers to the cardinality of the binning partition. The computational complexity of RISC and BH-EM have to be analyzed by looking at parameter inference and sample generation separately. Parameter inference is performed k times until either the convergence criterion ϵ or the maximum number of iterations M is reached.[31] For Hawkes processes, MLE scales polynomially and results in an $O(K^2)$ complexity, where K denotes the number of events in the (unobserved) sample history $\mathcal{H}_T$. The event count K is equal to the sum of the elements of the bin-count vector X, which is by Eq. (11) the same for any sample-corrected history $\hat{\mathcal{H}}_T$.[32] Therefore, the time complexity required for parameter inference exceeds the runtime of both aforementioned algorithms. The complexity of the sampling procedure, as part of the BH-EM algorithm, can be viewed as solving an optimization problem L times, where the complexity of the optimization problem depends on the

[30] Small-sample statistical inference has been studied extensively for Hawkes processes, and the RISC algorithm can accommodate multiple estimation methods, such as MLE (used throughout in the numerical examples) or EM, which can overcome certain estimation issues by augmenting the parameter space with the branching structure (cf. Sec. II A 2). Other techniques are also available, for example, variational inference [50].

[31] The speed of convergence is sensitive with respect to the choice of $\hat{\boldsymbol{\theta}}_0$. Using uniformly distributed events to derive the initial guess speeds up the convergence of the RISC algorithm when compared to random seeding.

[32] An exception is the Hawkes process with exponential kernel, where the complexity is reduced to $O(K)$ [21].

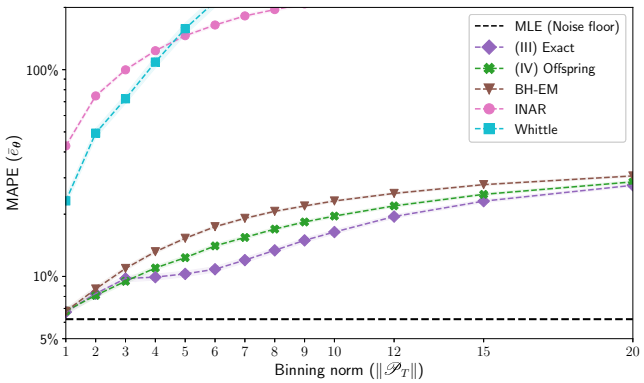


FIG. 17. MAPE $\bar{e}_{\boldsymbol{\theta}}$ for $(\mu_4, \alpha_4, \beta_4) = (0.1, 0.9, 1.5)$.

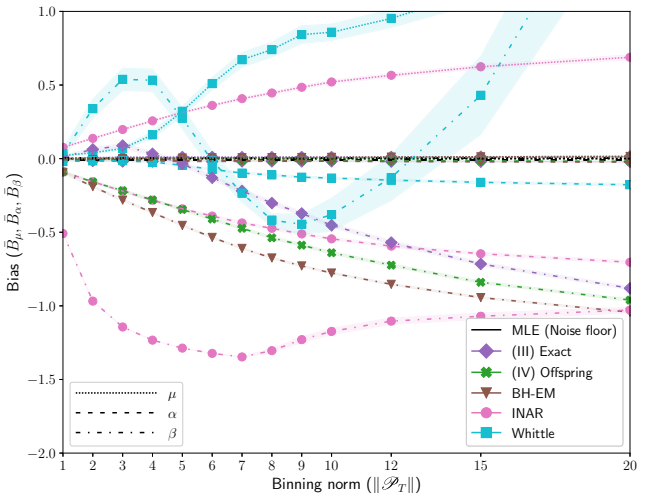


FIG. 18. Bias $\bar{B}_{\mu}, \bar{B}_{\alpha}, \bar{B}_{\beta}$ for $(\mu_4, \alpha_4, \beta_4) = (0.1, 0.9, 1.5)$.

chosen solver.[33] For our approach, the sampling time complexity is dominated by the time required for generating a synthetic sequence. When utilizing Ogata's thinning algorithm [48], the runtime increases linearly in the length of the simulation horizon T. Even under a worst-case scenario of performing K sample adjustments, the RISC algorithm requires significantly less runtime than the BH-EM algorithm, which attempts to find a globally optimal event-time placement within each bin.

C. Nonuniform and random time censoring

To ensure a fair performance comparison of the proposed RISC algorithm to other extant solutions, our numerical experiments relied on uniform binning partitions, despite the fact that the general approach in Sec. II allows for any finite partition of the observation interval. Indeed, in many real-world applications the available observations undergo *nonuniform* time censoring. For instance, self-exciting point processes have been used to model contagious diseases such as COVID-19, where infection statistics would in many countries produce daily infection counts during the week and aggregate counts over the weekend (often resulting in aggregate counts published on Mondays). To remain within the uniform time-censoring framework would require interpolation techniques for deaveraging, which need a separate justification. By contrast, the RISC algorithm in its native form can accommodate any type of time censoring. This includes partial time censoring when censored and uncensored information coexist, in which case one can use a sufficiently fine binning partition on the uncensored portions of the signal.

A *random* partition of time may arise when an interarrival history is checked randomly, for example, when sporadically observing the inventory of a durable-goods monopolist which decreases as a consequence of an intermittent stochastic order-arrival process. Naturally, any given realization of a (nontrivial) random partition is nonuniform with probability 1. Since the relative error curves are usually increasing and concave in the binning norm of a uniform partition, a random partition with the same expected bin width (i.e., a mean-preserving spread of the deterministic uniform partition) tends to have a lower average error, due to Jensen's inequality.[34]

D. Inverse problem and applications

The results in Sec. III indicate how Hawkes-process characteristics influence estimation error and bias. Thus, if a Hawkes process is subject to detectable (or known) regime changes, then a desired estimation error can be achieved via an adapted time-censoring strategy. For example, when monitoring traffic patterns on a social-media platform, during a high-traffic regime a coarser time censoring may be able to achieve the same estimation error as a finer binning partition in a low-traffic regime. This insight suggests a unique stream of research. Assuming the observed bias and process dynamics of an application are known, the matter of time censoring can be reformulated as an inverse problem. The question of interest becomes how much time censoring can be applied while maintaining a certain defined service level (or more generally, while suitably tracking a predefined time-varying service-level requirement). By service level, we refer to the algorithm performance (e.g., in terms of estimation error and runtime) as well as additional user requirements (e.g., aggregate energy use by the time-censored arrival-detection sensors). In some applications, an accurate process monitoring with detailed event history might be very cost intensive. The recurring costs may stem from the computation of updated estimates or from the energy for sensor-based monitoring of an environment, to name just a few. Therefore, the application might operate in different modes when the process dynamics of a system are fairly time invariant, switching from recording high-resolution data to time-censored data.[35] In this context, the design of the binning partition to monitor the process in an economically efficient manner becomes a choice variable.

V. CONCLUSION

In this paper, we introduced a RISC algorithm for the estimation of Hawkes processes from time-censored data. We reconstruct a history of continuous-time samples, which on a given binning partition exhibits the same entropy as an observed reference bin-count sequence (corresponding to the only available actual data). The required thinning and thickening of simulated sample paths is referred to as SC, for which a variety of methods are proposed in Sec. II C and subsequently tested in Sec. III C. In each iteration, a synthetically generated sample path based on the current estimate of the process parameters is corrected by removing (*thinning*) and adding (*thickening*) extra samples consistent with the original bin-count sequence, as well as with the conditional survival probability of each individual sample. The recursive-identification portion of the RISC algorithm refers to a structured update of the process parameters, which ensures convergence; cf. Sec. II D.

The performance of the RISC algorithm is evaluated based on relative and absolute estimation errors in comparison with other extant algorithms. A large-scale simulation study suggests that the proposed algorithm method significantly outperforms a naïve uniform distribution of events matching the presented bin-count sequence. The RISC algorithm is also benchmarked against the INAR approximation proposed by Kirchner [22], the Whittle estimation proposed by Cheysson and Lang [25], and the BH-EM algorithm introduced by Shlomovich *et al.* [30]. For a set of representative

[33] Solvers for nonconvex optimization problems may include the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm (respectively, a limited-memory BFGS algorithm with box constraints: L-BFGS-B) and sequential least-squares quadratic programming (SLSQP).

[34] For HP($\boldsymbol{\theta}_1$) and HP($\boldsymbol{\theta}_4$), with a random partition featuring a constant coefficient of variation of 50% across all expected bin-widths between 1 and 20, we find that the MAPE decreases by about 30% across the board compared to a uniform partition.

[35] Recording might even be switched off completely during stationary phases, substituting missing bin counts by suitably averaged data points *ex post*.

TABLE III. Notation.

Symbol	Description	Range/value
B	Bias	$\mathbb{R}$
$\mathcal{B}_\ell$	Bin (an element of $\mathscr{P}_T$)	$(\tau_{\ell-1}, \tau_\ell]$
CV	Coefficient of variation	$\mathbb{R}_+$
e	Mean absolute percentage error (MAPE)	$\mathbb{R}_+$
H	Bin-count entropy	$\mathbb{R}_+$
$\mathcal{H}_t$	Available information at time t	
k	RISC algorithm iteration index	$\mathbb{N}$
K	Number of arrival events	$\mathbb{N}$
L	Number of bins	$\mathbb{N}$
ℓ	Bin index	$\{1, \ldots, L\}$
$\mathcal{L}$	Log likelihood	$\mathbb{R}$
M	Maximum number of algorithm iterations	$\mathbb{N}$
N	Number of synthetic sample paths	$\mathbb{N}$
$N(t)$	Counting process	$\mathbb{N}$
P	Dimension of parameter space Θ	$\mathbb{N}$
$\mathscr{P}_T$	Binning partition	
t	Current time	$\mathbb{R}_+$
t_j	Arrival time of jth event	$\mathbb{R}_+$
T	Observation horizon	$\mathbb{R}_{++}$
X	Bin-count vector	$\mathbb{N}^K$
α	Branching coefficient	$\mathbb{R}_+$
β	Decay rate	$\mathbb{R}_+$
Δ	Relative norm (of a binning partition)	$\mathbb{R}_+$
ϵ	Tolerance threshold	$\mathbb{R}_{++}$
$\boldsymbol{\theta}$	Vector of process parameters $[\boldsymbol{\theta} = (\mu, \alpha, \beta)]$	$\Theta \subset \mathbb{R}_+^P$
$\bar{\lambda}_T, \bar{\lambda}$	Expected average arrival rate $[\bar{\lambda} = \lim_{T\to\infty} \lambda_T]$	$\mathbb{R}_+$
$\lambda(t\|\mathcal{H}_t)$	Conditional intensity function	$\mathbb{R}_+$
$\Lambda(t\|\mathcal{H}_t)$	Cumulative rate function	$\mathbb{R}_+$
μ	Background rate	$\mathbb{R}_+$
ξ	Average cluster-length ratio $[\xi = (\mu/\beta)/(1-\alpha)]$	$\mathbb{R}_+$
τ_ℓ	Discrete binning time $[\tau_0 = 0; \tau_L = T]$	$(0, T]$
$\phi(\cdot)$	Self-excitation function (kernel)	$\mathbb{R}_+$

parameter vectors $\boldsymbol{\theta}$ and a uniform binning partition, we find that the RISC algorithm achieves robust convergence (despite inherent nonconvexities in the estimation problem) and excellent model performance comparable to and above the level of all extant methods. By varying the level of time censoring extensively (between $\Delta = 0.1\%$ and $\Delta = 2\%$), the sensitivity of the estimation results to the severity of time censoring is examined. In contrast to other algorithms, the proposed approach allows for nonuniform binning partitions, which also invites research into the inverse problem of finding application-optimal partitions; cf. Sec. IV D. The RISC algorithm can be easily adapted to other types of nonhomogeneous Poisson processes (NHPP). In future work, different statistical inference approaches could be explored that ensure convergence to a global optimum and to further improve the runtime of the RISC algorithm.

ACKNOWLEDGMENTS

The authors wish to thank the participants of the Stochastic Networks Conference (SNC) 2022 at Cornell University and three anonymous referees for their helpful comments and suggestions.

APPENDIX A: NOTATION

Table III summarizes the notation.

APPENDIX B: EXTENDED PERFORMANCE ASSESSMENT

1. Distribution of events

Figure 19 illustrates the distribution of events for the parameter vectors $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\theta}_4$, given the binning norm $\|\mathscr{P}_T\| \in \{1, 7, 20\}$. For the Poisson process HP($\boldsymbol{\theta}_0$), and for the noncritical processes HP($\boldsymbol{\theta}_1$) and HP($\boldsymbol{\theta}_2$), the distribution of events per bin approaches a normal distribution with increasing norm $\|\mathscr{P}_T\|$. For the near-critical processes HP($\boldsymbol{\theta}_3$) and HP($\boldsymbol{\theta}_4$), the distribution of events per bin resembles an exponential distribution.

Figure 20 shows representative process realizations for all HP($\boldsymbol{\theta}_m$) in Table I. HP($\boldsymbol{\theta}_1$) and HP($\boldsymbol{\theta}_2$) are Poisson-like processes (with $\alpha \ll 1$), while HP($\boldsymbol{\theta}_3$) and HP($\boldsymbol{\theta}_4$) are near-critical processes (with $\alpha \approx 1$).

2. Algorithmic details

The RISC algorithm, outlined in Sec. II, requires as input the time-censored history in the form of a bin-count

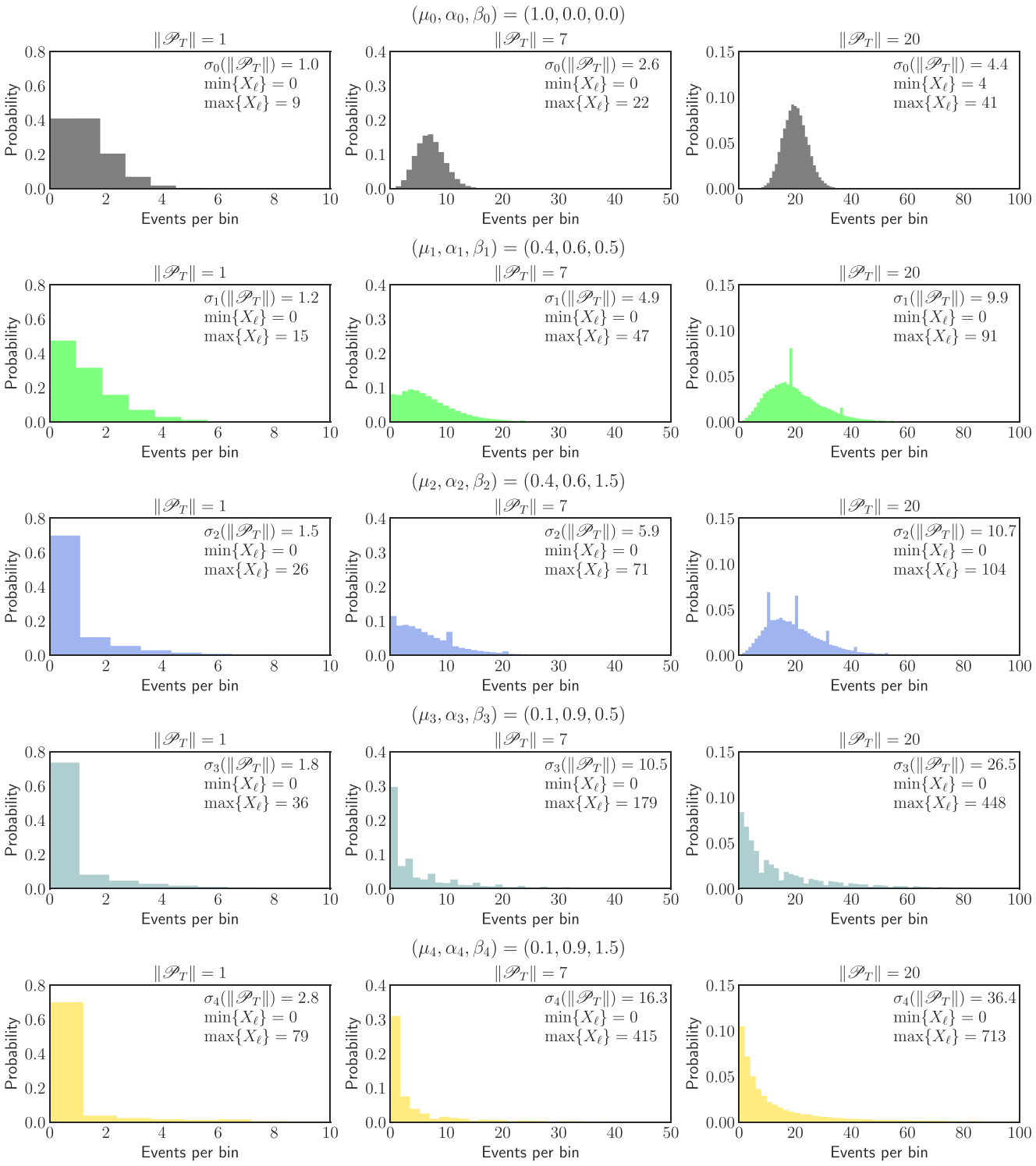


FIG. 19. Distribution of events per bin for HP($\boldsymbol{\theta}_m$) and $\|\mathscr{P}_T\| \in \{1, 7, 20\}$.

vector $X = \mathcal{H}_T/\mathscr{P}_T$ as in Eq. (11), where $\mathcal{H}_T$ is the (unobserved) uncensored sample path on the interval $(0, T]$ (with observation horizon $T > 0$), and $\mathscr{P}_T$ is the binning partition. In our numerical experiments, we use a uniform binning partition $\mathscr{P}_T = \{((T/L)(\ell-1), (T/L)\ell]\}_{\ell=1}^{L}$ (cf. Remark 1) with relative norm $\Delta = \|\mathscr{P}_T\|/T = 1/L$. Specifically, for an observation horizon $T = 1000$ and binning norm $\|\mathscr{P}_T\|$ between 1 and 20, the number of L (which corresponds to the number of bin-count entries X_ℓ in X) varies between 50 and 1000. A tolerance threshold $\epsilon = 0.01$ defines a stopping criterion for the deviation of subsequent parameter iterates $\hat{\boldsymbol{\theta}}^k$ and $\hat{\boldsymbol{\theta}}^{k-1}$ (cf. Fig. 2) for $k \in \{1, \ldots, M\}$, with the maximum number of iterations $M = 20$ acting as a secondary stopping threshold (which is hardly ever reached). To limit the possibility of premature termination of the algorithm due to

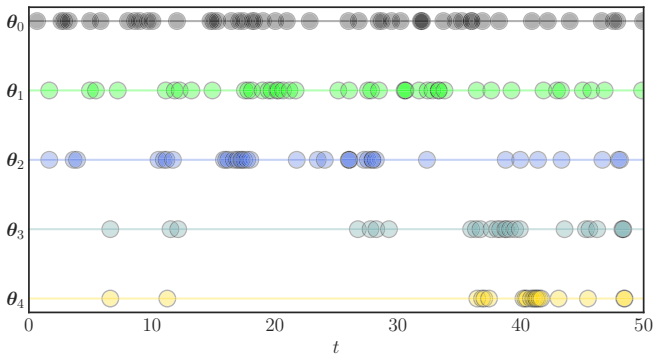


FIG. 20. Representative realizations of HP($\boldsymbol{\theta}_m$), for $m \in \{0, \ldots, 4\}$, on the time interval (0,50].

simulation noise we always perform at least 3 iterations, and test whether

$$\|\hat{\boldsymbol{\theta}}^k - \hat{\boldsymbol{\theta}}^{k-1}\|_2 + \|\hat{\boldsymbol{\theta}}^{k-1} - \hat{\boldsymbol{\theta}}^{k-2}\|_2 + \|\hat{\boldsymbol{\theta}}^{k-2} - \hat{\boldsymbol{\theta}}^{k-3}\|_2 \leqslant 3\epsilon, \quad k \geqslant 3,$$

that is, whether the moving average of subsequent Cauchy deviations satisfies the primary stopping criterion (or not). Naturally, the speed of convergence depends on the chosen SC method (cf. Sec. II C), on the coarseness of the binning partition (quantified by Δ), and the operating point as defined by the initial seed of the parameter estimate, $\hat{\boldsymbol{\theta}}^0$. The latter is obtained by uniformly distributing X_ℓ arrivals on the ℓth bin $((T/L)(\ell-1), (T/L)\ell]$ for all $\ell \in \{1, \ldots, L\}$. To demonstrate the general applicability of the proposed RISC method, the tolerance threshold $\epsilon = 1\%$ remained fixed throughout the study. The fine-tuning of the tolerance threshold as a function of the SC method, partition coarseness, and operating point is left for further research.

3. Solver comparison: SLSQP versus L-BFGS-B

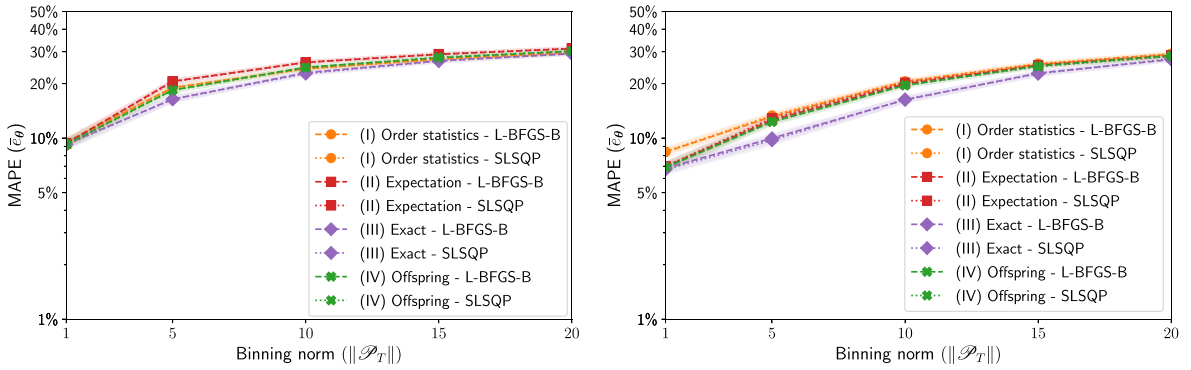
Estimating the process parameters $\hat{\boldsymbol{\theta}}$ requires the solution of a nonconvex optimization problem. Numerical solvers available for this type of problem include the sequential least squares programming (SLSQP) algorithm and the limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm with box constraints (L-BFGS-B). In Figs. 21 and 22, the estimation errors produced by the RISC algorithm (with its different SC methods) and comparison algorithms, respectively, are shown using $N = 250$ sample-path realizations. Note that for RISC both solvers produce almost identical results.

It becomes apparent that both BH-EM and Whittle exhibit convergence issues with the SLSQP solver. By contrast, the INAR estimates feature only a small variation across the two solvers. In our numerical study (cf. Sec. III), we use the L-BFGS-B solver in all comparison algorithms, due to its superior performance, with $\epsilon = 10^{-6}$ as convergence tolerance and $M = 10^5$ as the maximum number of iterations.

Algorithm 1. Recursive identification with sample correction.

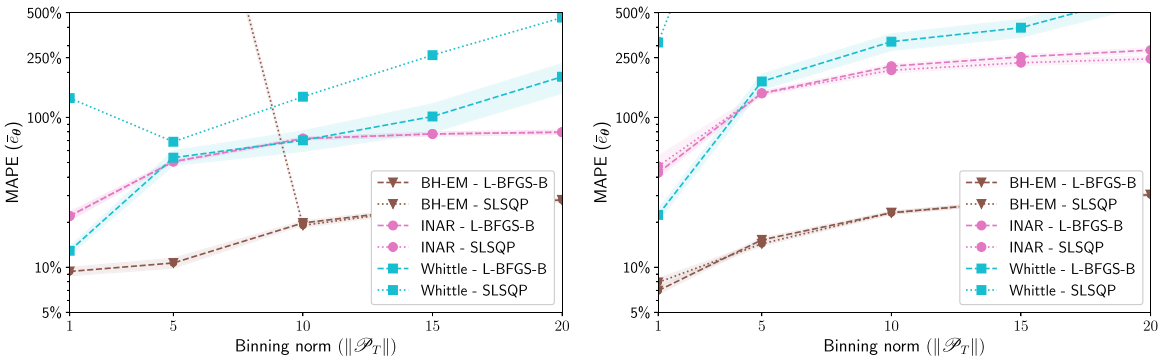
Input: Bin-count vector $X = (X_1, \ldots, X_L)$, observation horizon T, binning partition $\mathscr{P}_T$, tolerance threshold ϵ, maximum number of iterations M, sample-correction methods (I)–(IV)
Output: $\hat{\boldsymbol{\theta}}_{\text{RISC}}$ – RISC estimate of parameter vector

$k \leftarrow 0, \delta \leftarrow \infty$
$\hat{\boldsymbol{\theta}}^k \leftarrow$ Generate initial guess
while $3\epsilon < \delta$ **and** $k < M$ **do**
  $\mathcal{H}_T^{s,k} \leftarrow$ Simulate HP($\hat{\boldsymbol{\theta}}^k$) on $(0, T]$
  $X^{s,k} \leftarrow$ Bin the simulated process history $\mathcal{H}_T^{s,k}$
  for $\ell = 1 \to L$ **do**
    if $X_\ell < X_\ell^{s,k}$ **then** (According to chosen sample-correction method)
      Apply thinning to bin $X_\ell^{s,k}$, and adjust sample history: $[\hat{\mathcal{H}}_T^k]_\ell \leftarrow [\mathcal{H}_T^{s,k}]_\ell$
    else if $X_\ell > X_\ell^{s,k}$ **then**
      Apply thickening to bin $X_\ell^{s,k}$, and adjust sample history: $[\hat{\mathcal{H}}_T^k]_\ell \leftarrow [\mathcal{H}_T^{s,k}]_\ell$
    else
      Continue with next bin
    end if
  end for
  $\hat{\boldsymbol{\theta}}^{k+1} \leftarrow \arg\max_{\boldsymbol{\theta}\in\Theta} \ln \mathcal{L}(\boldsymbol{\theta}|\hat{\mathcal{H}}_T^k)$
  $k \leftarrow k + 1$
  if $k > 3$ **then**
    $\delta \leftarrow \sum_{i=k-2}^{k} \|\hat{\boldsymbol{\theta}}^k - \hat{\boldsymbol{\theta}}^{k-1}\|_2$
  end if
end while



(a) HP($\boldsymbol{\theta}_1$). (b) HP($\boldsymbol{\theta}_4$).

FIG. 21. MAPE $\bar{e}_{\boldsymbol{\theta}}$ for sample-correction methods.



(a) HP($\boldsymbol{\theta}_1$). (b) HP($\boldsymbol{\theta}_4$).

FIG. 22. MAPE $\bar{\epsilon}_{\boldsymbol{\theta}}$ for extant algorithms.[36]

4. Interrun versus intrarun variability

Consider the estimation of the unknown parameter vector $\boldsymbol{\theta}$ of a Hawkes process HP($\boldsymbol{\theta}$). For a fixed seed $\hat{\boldsymbol{\theta}}^0$ of the RISC algorithm (cf. Fig. 2), a given SC method (I)–(IV), a tolerance threshold ϵ, and a maximum number of runs M, the RISC-estimate $\hat{\boldsymbol{\theta}}_{\text{RISC}}$ is generally subject to randomness due to two sources of noise: process uncertainty and simulation uncertainty. The process uncertainty stems from the fact that (as long as $\Delta < 1$) the bin-count vector varies from run to run, resulting in *interrun variability* (measured by $\bar{\sigma}_{\text{RISC}}$). Conditional on a fixed bin-count vector X, the simulation uncertainty produces variations in the estimation output of the RISC algorithm, which arises from the fact that it is necessary to simulate processes based on the various iterates $\hat{\boldsymbol{\theta}}^k$, together with the fact that some SC methods produce a stochastic redistribution of events. The simulation noise produces *intrarun variability* (measured by $\bar{\sigma}_{\text{RISC}}|X$).

Given $N = 1000$ sample-path realizations of HP($\boldsymbol{\theta}$), with the corresponding observations of bin-count vectors $X^{(1)}, \ldots, X^{(N)}$, the pth component of the average RISC-estimate,

$$\bar{\boldsymbol{\theta}}_{\text{RISC}} = (\bar{\theta}_{1,\text{RISC}}, \ldots, \bar{\theta}_{P,\text{RISC}}),$$

is

$$\bar{\theta}_{p,\text{RISC}} = \frac{1}{N}\sum_{n=1}^{N} \hat{\theta}^{(n)}_{p,\text{RISC}}, \quad p \in \{1, \ldots, P\};$$

consequently, the component-specific (unbiased) sample standard deviation becomes

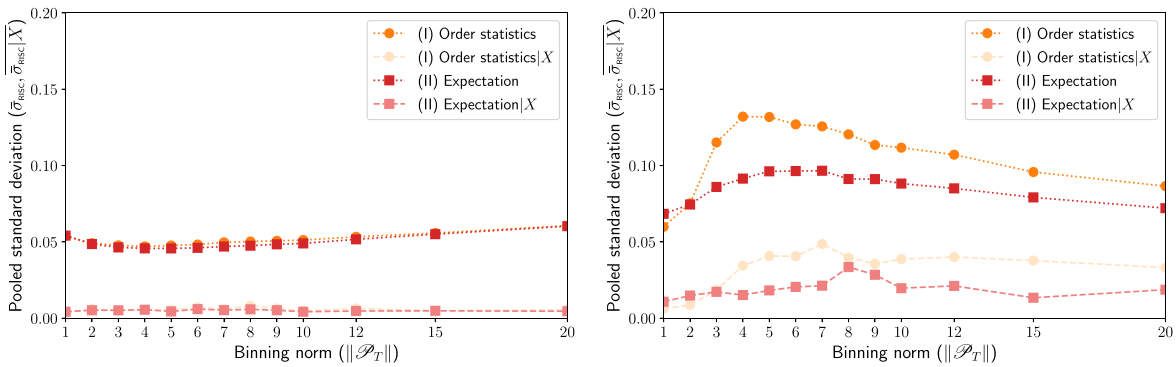
$$\bar{\sigma}_{p,\text{RISC}} = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}\left(\hat{\theta}^{(n)}_{p,\text{RISC}} - \bar{\theta}_{p,\text{RISC}}\right)^2}, \quad p \in \{1, \ldots, P\}.$$

To aggregate across components we use the pooled standard deviation (see, e.g., Cohen [51]),

$$\bar{\sigma}_{\text{RISC}} = \sqrt{\frac{(\bar{\sigma}_{1,\text{RISC}})^2 + \cdots + (\bar{\sigma}_{P,\text{RISC}})^2}{P}},$$

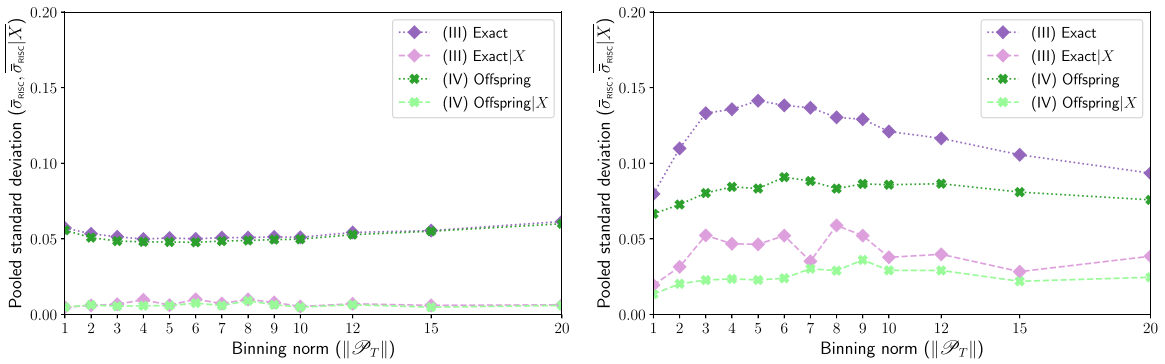
which measures the interrun variability. Given a fixed run with bin-count vector X, it is also possible to rerun the algorithm $\hat{N} = 100$ times and use essentially the same formulas as before (with N replaced by $\hat{N}$) to compute the pooled standard deviation $\bar{\sigma}_{\text{RISC}}|X$, which when averaged over all observed bin-count vectors $X^{(n)}$ measures the (average) intrarun

[36]For the Whittle algorithm with SLSQP solver the confidence intervals are omitted due to salient divergence issues.



(a) HP($\boldsymbol{\theta}_1$). (b) HP($\boldsymbol{\theta}_4$).

FIG. 23. Interrun variability ($\bar{\sigma}_{\text{RISC}}$) versus average intrarun variability ($\overline{\bar{\sigma}_{\text{RISC}}|X}$) for methods (I) and (II).



(a) HP($\boldsymbol{\theta}_1$). (b) HP($\boldsymbol{\theta}_4$).

FIG. 24. Interrun variability ($\bar{\sigma}_{\mathrm{RISC}}$) versus average intrarun variability ($\overline{\bar{\sigma}_{\mathrm{RISC}}|X}$) for methods (III) and (IV).

variability $\overline{\bar{\sigma}_{\mathrm{RISC}}|X}$.[37] As shown in Figs. 23(a) and 23(b) [resp., Figs. 24(a) and 24(b)], for HP($\boldsymbol{\theta}_1$) and HP($\boldsymbol{\theta}_4$), respectively, the interrun variability ($\bar{\sigma}_{\mathrm{RISC}}$) exceeds the average intrarun variability ($\overline{\bar{\sigma}_{\mathrm{RISC}}|X}$) by a significant multiple (between 2 and 5).

5. High-traffic versus low-traffic regime

All process scenarios introduced in Sec. III B possess the same expected arrival rate ($\bar{\lambda} = 1$). To understand how the estimation error deviates in high-traffic regimes ($\bar{\lambda} = 10$), we consider the additional parameter vector $\boldsymbol{\theta}_5 = (1, 0.9, 15)$ featuring a higher background rate while clustering and offspring behaviors remain unchanged (cf. Appendix C 1). Figure 25 illustrates that a high-traffic regime comes with elevated estimation errors. Indeed, even at the finest partition (with $\|\mathscr{P}_T\| = 1$), the intrabin offspring behavior is no longer observable.[38] By contrast, the MLE noise floor (for uncensored sample paths) drops in the high-traffic regime $\boldsymbol{\theta}_5$, in comparison to the low-traffic regime $\boldsymbol{\theta}_4$.

6. Results for intermediate parameter vectors ($\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$)

To complement Sec. III B, we now provide the numerical results for the intermediary parameter vectors $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$; cf. Table I.

a. Performance of sample-correction methods ($\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$)

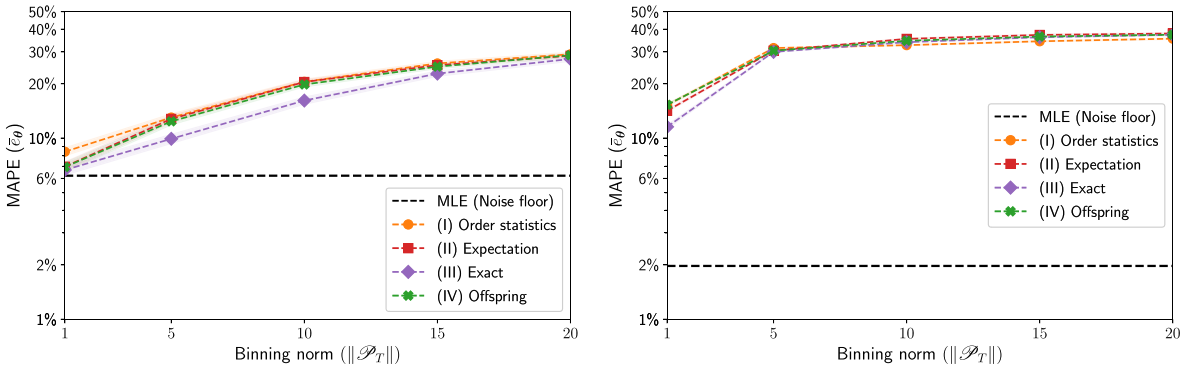
Figures 26 and 28 show the MAPE for $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$ for our proposed RISC algorithm compared to the uniform reference process. Figures 27 and 29, on the other hand, illustrate the bias.

b. Comparison to extant algorithms ($\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$)

We compare the SC methods (III) and (IV) against extant algorithms. Figures 30 and 32 display the MAPE for $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$. Figures 31 and 33 track the bias.

[37] To keep the total run-time manageable, instead of averaging over the pooled standard deviations for *all* runs, we average the results over 100 randomly selected bin-count vectors (from the total of $N = 1000$ bin-count vectors).

[38] The average cluster length $(1/\beta)/(1-\alpha)$ in Eq. (24) is smaller than $\|\mathscr{P}_T\|$ in the simulation study. By comparison, for HP($\boldsymbol{\theta}_4$) the average cluster length is ten times larger: typical Hawkes-process features such as clustered events followed by inactivity, which simplify identification, remain detectable from a bin-count vector on a sufficiently fine binning partition.



(a) HP($\boldsymbol{\theta}_4$). (b) HP($\boldsymbol{\theta}_5$).

FIG. 25. MAPE $\bar{e}_{\boldsymbol{\theta}}$ for low-traffic regime and high-traffic regime.

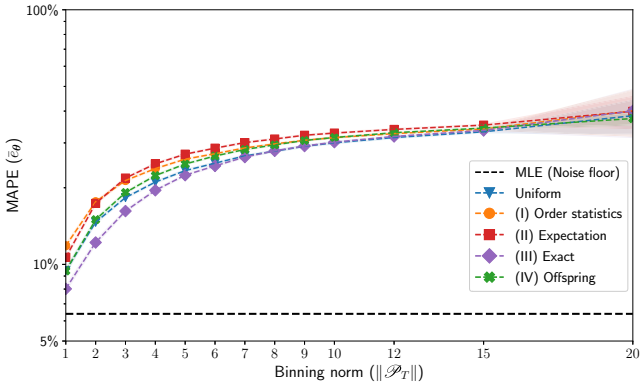


FIG. 26. MAPE $\bar{e}_{\boldsymbol{\theta}}$ for $(\mu_2, \alpha_2, \beta_2) = (0.4, 0.6, 1.5)$.[39]

7. Performance comparison $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_4)$

Section III and Appendix B 6 contained a performance overview as a function of the binning norm for all parameter vectors $\boldsymbol{\theta}$ specified in Table I. We now provide the key results in tabular form for $\|\mathscr{P}_T\| \in \{1, 7, 20\}$. In each table, the naïve uniform SC method serves as baseline. Thereafter, we report the estimation errors for the RISC algorithm with SC methods (I)–(IV) followed by the extant algorithms in Table II.

Tables IV, V, VI, and VII provide the obtained results for $\|\mathscr{P}_T\| = 1$ across all parameter vectors. For $\|\mathscr{P}_T\| = 7$, the corresponding results are presented in Tables VIII, IX, X, and XI. Additionally, Tables XII, XIII, XIV, and XV display the results for $\|\mathscr{P}_T\| = 20$.

[39]For $\|\mathscr{P}_T\| = 20$, the confidence interval widens, which is related to a divergence in the decay-rate parameter β. Mark and Weber [19] observe these types of MLE convergence issues pointing to flat log-likelihood contours as their principal cause.

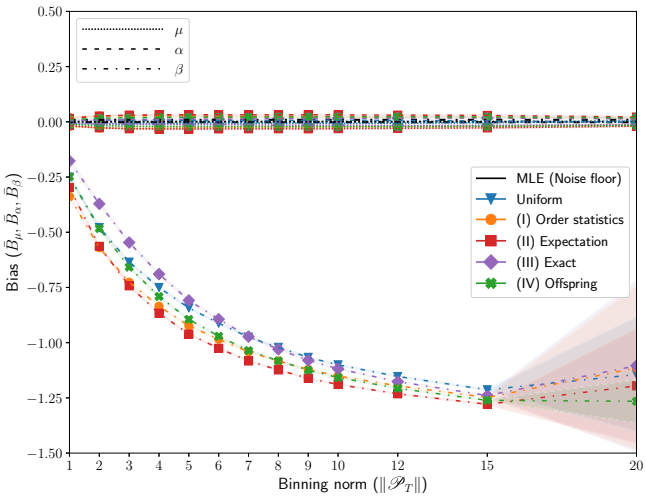


FIG. 27. Bias $\bar{B}_{\mu}, \bar{B}_{\alpha}, \bar{B}_{\beta}$ for $(\mu_2, \alpha_2, \beta_2) = (0.4, 0.6, 1.5)$.

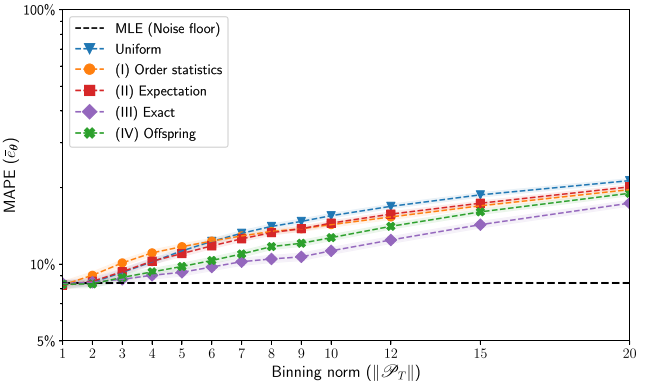


FIG. 28. MAPE $\bar{e}_{\boldsymbol{\theta}}$ for $(\mu_3, \alpha_3, \beta_3) = (0.1, 0.9, 0.5)$.

8. Results for Whittle parameter estimation

Cheysson and Lang [25] proved that their proposed estimator is consistent and asymptotically normal. More specifically, consistency obtains as $T \to \infty$ for a uniform binning partition of a given norm. As the results in Sec. III D indicate, the Whittle algorithm tends to produce an elevated estimator error or, in other words, rather high-variance estimates. To check estimator consistency as a function of sample size, the deviations are evaluated for the observation horizons $T \in \{1000, 2000, 4000, 8000\}$; see Fig. 34. The uniform SC method serves as reference. While the up-to-eightfold increase in the observation horizon tends to improve the estimation error, extending T does not make a major difference. In particular, Whittle does not outperform the naïve uniform SC method when time censoring becomes more severe. These results suggest that the Whittle algorithm requires a sufficiently large sample size and, in particular, also that the observation horizon of $T = 1000$ in our numerical study may simply be insufficient for this inference technique.

APPENDIX C: THEORETICAL BACKGROUND

1. Average cluster size

Statistical inference from time-censored Hawkes processes proves to be challenging. One complicating determinant

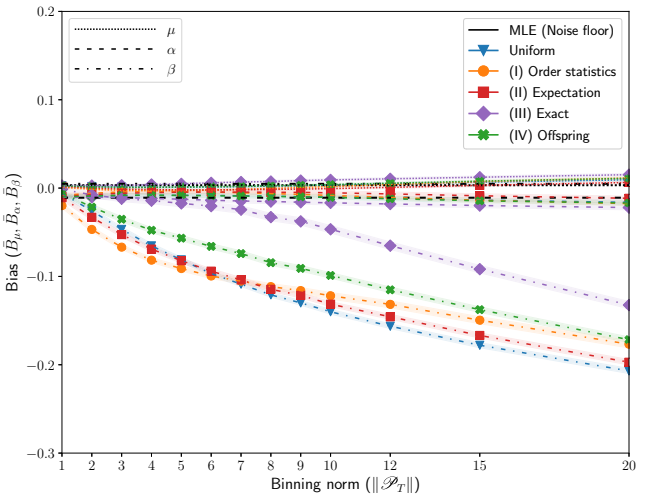


FIG. 29. Bias $\bar{B}_{\mu}, \bar{B}_{\alpha}, \bar{B}_{\beta}$ for $(\mu_3, \alpha_3, \beta_3) = (0.1, 0.9, 0.5)$.

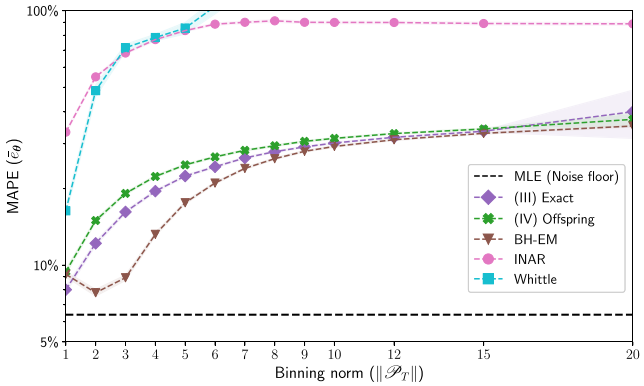


FIG. 30. MAPE $\bar{e}_{\boldsymbol{\theta}}$ for $(\mu_2, \alpha_2, \beta_2) = (0.4, 0.6, 1.5)$.

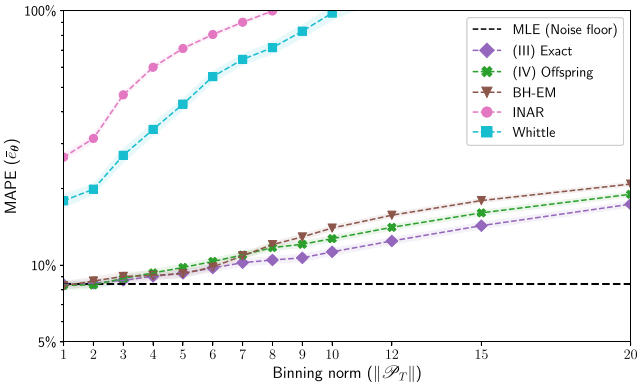


FIG. 32. MAPE $\bar{e}_{\boldsymbol{\theta}}$ for $(\mu_3, \alpha_3, \beta_3) = (0.1, 0.9, 0.5)$.

is whether clusters are overlapping each other or not. To derive the average cluster length,[40] it is useful to introduce the renormalized memory kernel $R(\cdot)$, also known as the response function [52]:

$$R(t) = \frac{\phi(t)}{\alpha} + \int_0^t \phi(t-\vartheta)\, R(\vartheta)\, d\vartheta, \quad t \geqslant 0. \tag{22}$$

Broadly speaking, the memory kernel describes an impulse response, where—in the terminology of branching processes—an impulse refers to an immigrant event, so the considered response is in fact the implied cascade of events.[41] The term $\phi(t)/\alpha$ in Eq. (22) represents the bare memory kernel, describing the probability of an event causing another generation of offsprings. Solving the above Volterra integral equation (of the second kind) in Eq. (22) for the exponential kernel in Eq. (2), one obtains (cf. Rustler [53]):

$$R(t) = \beta \, \exp\left[-(1-\alpha)\beta t\right], \quad t \geqslant 0. \tag{23}$$

Morzywołek [54] derived the average cluster length:

$$\begin{aligned}(1-\alpha) &\int_0^\infty t R(t)\, dt \\ &= (1-\alpha) \int_0^\infty (\beta t) \exp\left[-(1-\alpha)(\beta t)\right] dt = \frac{1/\beta}{1-\alpha}.\end{aligned} \tag{24}$$

To suitably discriminate cluster behavior, we use the ratio ξ of the average cluster length in Eq. (24) to the average distance between immigrants (i.e., $1/\mu$):

$$\xi = \frac{\mu/\beta}{1-\alpha}. \tag{25}$$

For an average cluster-length ratio $\xi \geqslant 1$, clusters are overlapping. By contrast, processes with $\xi < 1$ tend to exhibit well-separated clusters. As an illustration, consider ratios $\xi_m = (\mu_m/\beta_m)/(1-\alpha_m)$ for each $\boldsymbol{\theta}_m = (\mu_m, \alpha_m, \beta_m)$ in Table I with $m \in \{1, \ldots, 4\}$. Since $\xi_2 = \xi_4 = 2/3 < 1 < 2 = \xi_1 = \xi_3$, one expects overlapping clusters for $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_3$ while for $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_4$ clusters are more separated; see Fig. 20.

[40]The average cluster length refers to the time point of the generation of the cluster, an immigrant, to the last offspring that is associated with the generated cascade.

[41]A cascade refers to all generations of offsprings created by a given immigrant.

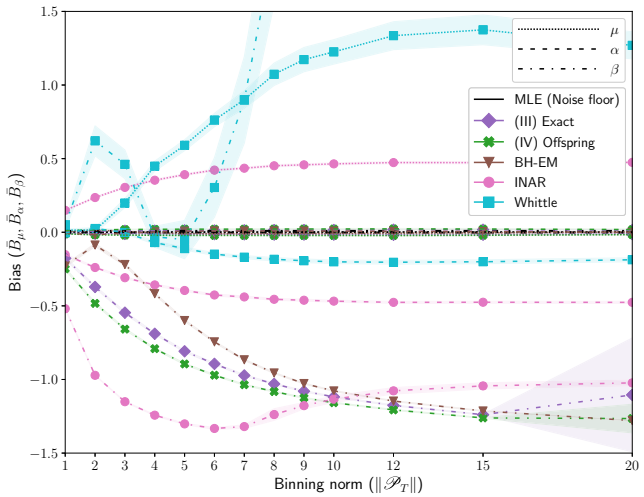


FIG. 31. Bias $\bar{B}_\mu, \bar{B}_\alpha, \bar{B}_\beta$ for $(\mu_2, \alpha_2, \beta_2) = (0.4, 0.6, 1.5)$.

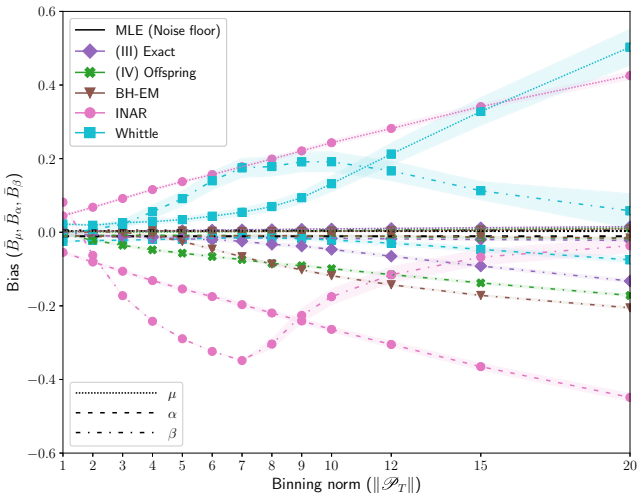


FIG. 33. Bias $\bar{B}_\mu, \bar{B}_\alpha, \bar{B}_\beta$ for $(\mu_3, \alpha_3, \beta_3) = (0.1, 0.9, 0.5)$.

TABLE IV. $(\mu_1, \alpha_1, \beta_1) = (0.4, 0.6, 0.5)$ for $\|\mathscr{P}_T\| = 1$.

	$\bar{e}_{\boldsymbol{\theta}}$		$\bar{e}_{\mu}$		$\bar{e}_{\alpha}$		$\bar{e}_{\beta}$	
Method	mean	stdev	mean	stdev	mean	stdev	mean	stdev
Uniform	0.095	0.059	0.096	0.049	0.068	0.051	0.121	0.075
(I) Order statistics	0.095	0.054	0.094	0.047	0.067	0.049	0.123	0.064
(II) Expectation	0.094	0.054	0.093	0.047	0.066	0.049	0.121	0.065
(III) Exact	0.093	0.057	0.094	0.048	0.067	0.050	0.119	0.071
(IV) Offspring	0.093	0.056	0.094	0.048	0.067	0.050	0.118	0.068
BH-EM	0.095	0.059	0.098	0.050	0.068	0.051	0.120	0.072
INAR	0.214	0.094	0.236	0.075	0.159	0.073	0.247	0.124
Whittle	0.131	0.083	0.139	0.069	0.077	0.058	0.177	0.111

TABLE V. $(\mu_2, \alpha_2, \beta_2) = (0.4, 0.6, 1.5)$ for $\|\mathscr{P}_T\| = 1$.

	$\bar{e}_{\boldsymbol{\theta}}$		$\bar{e}_{\mu}$		$\bar{e}_{\alpha}$		$\bar{e}_{\beta}$	
Method	mean	stdev	mean	stdev	mean	stdev	mean	stdev
Uniform	0.094	0.075	0.067	0.032	0.048	0.035	0.169	0.120
(I) Order statistics	0.118	0.066	0.076	0.031	0.053	0.034	0.226	0.105
(II) Expectation	0.107	0.070	0.071	0.031	0.050	0.034	0.199	0.112
(III) Exact	0.080	0.084	0.065	0.031	0.047	0.035	0.129	0.137
(IV) Offspring	0.094	0.075	0.067	0.031	0.048	0.035	0.168	0.122
BH-EM	0.092	0.083	0.069	0.034	0.050	0.037	0.157	0.135
INAR	0.333	0.076	0.390	0.077	0.263	0.074	0.346	0.077
Whittle	0.164	0.272	0.203	0.104	0.066	0.050	0.222	0.456

TABLE VI. $(\mu_3, \alpha_3, \beta_3) = (0.1, 0.9, 0.5)$ for $\|\mathscr{P}_T\| = 1$.

	$\bar{e}_{\boldsymbol{\theta}}$		$\bar{e}_{\mu}$		$\bar{e}_{\alpha}$		$\bar{e}_{\beta}$	
Method	mean	stdev	mean	stdev	mean	stdev	mean	stdev
Uniform	0.083	0.036	0.142	0.018	0.03	0.034	0.077	0.048
(I) Order statistics	0.083	0.034	0.139	0.018	0.03	0.034	0.080	0.046
(II) Expectation	0.083	0.036	0.140	0.018	0.03	0.034	0.078	0.048
(III) Exact	0.084	0.037	0.143	0.018	0.03	0.035	0.079	0.050
(IV) Offspring	0.083	0.036	0.142	0.018	0.03	0.034	0.078	0.049
BH-EM	0.083	0.036	0.142	0.018	0.029	0.034	0.078	0.049
INAR	0.266	0.078	0.508	0.044	0.069	0.057	0.220	0.115
Whittle	0.179	0.070	0.336	0.051	0.043	0.045	0.159	0.101

TABLE VII. $(\mu_4, \alpha_4, \beta_4) = (0.1, 0.9, 1.5)$ for $\|\mathscr{P}_T\| = 1$.

	$\bar{e}_{\boldsymbol{\theta}}$		$\bar{e}_{\mu}$		$\bar{e}_{\alpha}$		$\bar{e}_{\beta}$	
Method	mean	stdev	mean	stdev	mean	stdev	mean	stdev
Uniform	0.075	0.060	0.098	0.013	0.028	0.032	0.098	0.098
(I) Order statistics	0.083	0.060	0.100	0.012	0.027	0.031	0.122	0.098
(II) Expectation	0.069	0.068	0.098	0.013	0.028	0.032	0.080	0.113
(III) Exact	0.067	0.080	0.102	0.013	0.029	0.033	0.071	0.133
(IV) Offspring	0.068	0.067	0.098	0.013	0.028	0.032	0.078	0.110
BH-EM	0.068	0.063	0.100	0.012	0.027	0.032	0.077	0.104
INAR	0.428	0.066	0.836	0.056	0.108	0.070	0.339	0.071
Whittle	0.232	0.291	0.424	0.060	0.038	0.042	0.233	0.498

TABLE VIII. $(\mu_1, \alpha_1, \beta_1) = (0.4, 0.6, 0.5)$ for $\|\mathscr{P}_T\| = 7$.

	$\bar{e}_{\boldsymbol{\theta}}$		$\bar{e}_{\mu}$		$\bar{e}_{\alpha}$		$\bar{e}_{\beta}$	
Method	mean	stdev	mean	stdev	mean	stdev	mean	stdev
Uniform	0.176	0.056	0.108	0.055	0.076	0.058	0.345	0.054
(I) Order statistics	0.216	0.050	0.116	0.052	0.082	0.055	0.450	0.041
(II) Expectation	0.235	0.047	0.121	0.050	0.085	0.053	0.497	0.036
(III) Exact	0.197	0.051	0.106	0.050	0.075	0.053	0.408	0.050
(IV) Offspring	0.215	0.049	0.112	0.050	0.080	0.053	0.454	0.042
BH-EM	0.131	0.056	0.104	0.053	0.074	0.056	0.215	0.060
INAR	0.616	0.085	0.680	0.102	0.458	0.096	0.708	0.045
Whittle	0.672	0.452	0.804	0.398	0.213	0.153	0.998	0.657

TABLE IX. $(\mu_2, \alpha_2, \beta_2) = (0.4, 0.6, 1.5)$ for $\|\mathscr{P}_T\| = 7$.

	$\bar{e}_{\boldsymbol{\theta}}$		$\bar{e}_{\mu}$		$\bar{e}_{\alpha}$		$\bar{e}_{\beta}$	
Method	mean	stdev	mean	stdev	mean	stdev	mean	stdev
Uniform	0.267	0.057	0.089	0.045	0.061	0.047	0.650	0.074
(I) Order statistics	0.287	0.050	0.099	0.042	0.068	0.044	0.694	0.062
(II) Expectation	0.301	0.047	0.107	0.041	0.074	0.043	0.722	0.056
(III) Exact	0.264	0.057	0.085	0.040	0.059	0.042	0.648	0.080
(IV) Offspring	0.283	0.050	0.093	0.040	0.065	0.043	0.690	0.064
BH-EM	0.240	0.053	0.084	0.041	0.058	0.043	0.578	0.069
INAR	0.900	0.142	1.087	0.120	0.732	0.106	0.880	0.186
Whittle	1.271	2.793	2.345	0.971	0.321	0.166	1.146	4.737

TABLE X. $(\mu_3, \alpha_3, \beta_3) = (0.1, 0.9, 0.5)$ for $\|\mathscr{P}_T\| = 7$.

	$\bar{e}_{\boldsymbol{\theta}}$		$\bar{e}_{\mu}$		$\bar{e}_{\alpha}$		$\bar{e}_{\beta}$	
Method	mean	stdev	mean	stdev	mean	stdev	mean	stdev
Uniform	0.132	0.033	0.148	0.019	0.031	0.035	0.218	0.042
(I) Order statistics	0.129	0.036	0.145	0.018	0.029	0.033	0.214	0.050
(II) Expectation	0.126	0.034	0.139	0.018	0.029	0.034	0.209	0.044
(III) Exact	0.102	0.046	0.157	0.019	0.033	0.037	0.116	0.068
(IV) Offspring	0.110	0.036	0.143	0.018	0.031	0.036	0.156	0.049
BH-EM	0.109	0.035	0.157	0.020	0.031	0.036	0.139	0.043
INAR	0.900	0.072	1.784	0.080	0.220	0.095	0.697	0.005
Whittle	0.643	0.265	1.210	0.177	0.052	0.061	0.668	0.419

TABLE XI. $(\mu_4, \alpha_4, \beta_4) = (0.1, 0.9, 1.5)$ for $\|\mathscr{P}_T\| = 7$.

	$\bar{e}_{\boldsymbol{\theta}}$		$\bar{e}_{\mu}$		$\bar{e}_{\alpha}$		$\bar{e}_{\beta}$	
Method	mean	stdev	mean	stdev	mean	stdev	mean	stdev
Uniform	0.209	0.062	0.141	0.016	0.032	0.035	0.454	0.100
(I) Order statistics	0.165	0.126	0.137	0.017	0.027	0.030	0.331	0.215
(II) Expectation	0.161	0.097	0.114	0.015	0.028	0.033	0.340	0.163
(III) Exact	0.120	0.137	0.145	0.016	0.033	0.035	0.181	0.234
(IV) Offspring	0.154	0.088	0.117	0.015	0.031	0.035	0.316	0.148
BH-EM	0.191	0.057	0.136	0.016	0.031	0.035	0.406	0.091
INAR	1.819	0.123	4.074	0.151	0.486	0.143	0.898	0.049
Whittle	2.616	0.734	7.152	0.800	0.133	0.115	0.563	0.981

TABLE XII. $(\mu_1, \alpha_1, \beta_1) = (0.4, 0.6, 0.5)$ for $\|\mathscr{P}_T\| = 20$.

	$\bar{e}_{\boldsymbol{\theta}}$		$\bar{e}_{\mu}$		$\bar{e}_{\alpha}$		$\bar{e}_{\beta}$	
Method	mean	stdev	mean	stdev	mean	stdev	mean	stdev
Uniform	0.283	0.064	0.136	0.071	0.096	0.074	0.616	0.041
(I) Order statistics	0.297	0.061	0.131	0.068	0.093	0.072	0.666	0.035
(II) Expectation	0.314	0.060	0.136	0.069	0.097	0.073	0.709	0.028
(III) Exact	0.297	0.061	0.131	0.069	0.093	0.072	0.666	0.037
(IV) Offspring	0.304	0.060	0.131	0.068	0.093	0.072	0.688	0.032
BH-EM	0.283	0.060	0.134	0.068	0.094	0.071	0.621	0.033
INAR	0.787	0.213	1.086	0.140	0.726	0.126	0.550	0.318
Whittle	2.110	2.756	1.858	0.985	0.307	0.209	4.164	4.666

TABLE XIII. $(\mu_2, \alpha_2, \beta_2) = (0.4, 0.6, 1.5)$ for $\|\mathscr{P}_T\| = 20$.

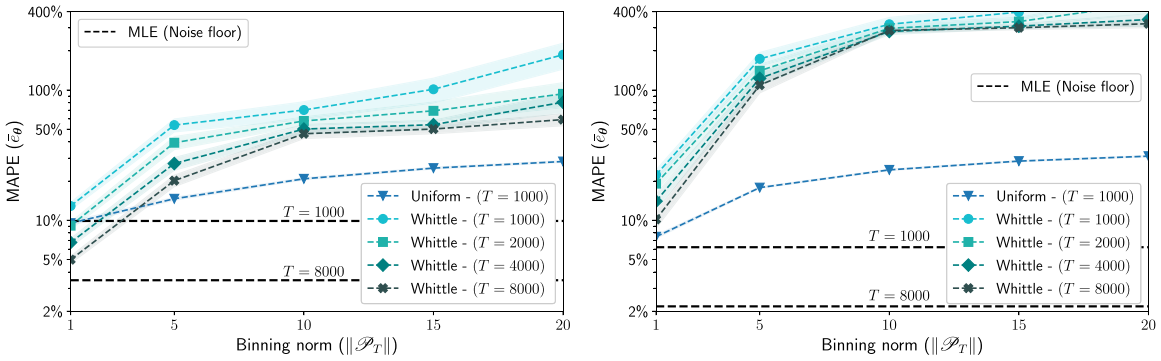
	$\bar{e}_{\boldsymbol{\theta}}$		$\bar{e}_{\mu}$		$\bar{e}_{\alpha}$		$\bar{e}_{\beta}$	
Method	mean	stdev	mean	stdev	mean	stdev	mean	stdev
Uniform	0.384	2.415	0.126	0.066	0.088	0.070	0.937	4.182
(I) Order statistics	0.401	3.445	0.124	0.062	0.087	0.066	0.993	5.966
(II) Expectation	0.399	2.416	0.132	0.065	0.094	0.069	0.972	4.184
(III) Exact	0.401	3.599	0.120	0.061	0.085	0.065	0.998	6.233
(IV) Offspring	0.374	0.909	0.125	0.061	0.089	0.065	0.908	1.572
BH-EM	0.352	0.055	0.120	0.061	0.084	0.065	0.853	0.035
INAR	0.887	0.207	1.186	0.130	0.794	0.112	0.682	0.314
Whittle	2.320	4.418	3.355	1.485	0.390	0.216	3.216	7.503

TABLE XIV. $(\mu_3, \alpha_3, \beta_3) = (0.1, 0.9, 0.5)$ for $\|\mathscr{P}_T\| = 20$.

	$\bar{e}_{\boldsymbol{\theta}}$		$\bar{e}_{\mu}$		$\bar{e}_{\alpha}$		$\bar{e}_{\beta}$	
Method	mean	stdev	mean	stdev	mean	stdev	mean	stdev
Uniform	0.213	0.038	0.190	0.023	0.035	0.040	0.414	0.047
(I) Order statistics	0.196	0.046	0.197	0.024	0.033	0.036	0.358	0.067
(II) Expectation	0.201	0.038	0.176	0.022	0.033	0.038	0.395	0.049
(III) Exact	0.174	0.047	0.213	0.024	0.038	0.041	0.269	0.066
(IV) Offspring	0.190	0.040	0.191	0.023	0.036	0.039	0.344	0.052
BH-EM	0.208	0.035	0.180	0.022	0.034	0.038	0.411	0.041
INAR	1.776	0.232	4.261	0.170	0.501	0.166	0.566	0.324
Whittle	2.114	0.629	5.586	0.783	0.117	0.116	0.641	0.749

TABLE XV. $(\mu_4, \alpha_4, \beta_4) = (0.1, 0.9, 1.5)$ for $\|\mathscr{P}_T\| = 20$.

	$\bar{e}_{\boldsymbol{\theta}}$		$\bar{e}_{\mu}$		$\bar{e}_{\alpha}$		$\bar{e}_{\beta}$	
Method	mean	stdev	mean	stdev	mean	stdev	mean	stdev
Uniform	0.311	0.058	0.205	0.021	0.035	0.036	0.692	0.091
(I) Order statistics	0.294	0.087	0.213	0.022	0.033	0.033	0.635	0.144
(II) Expectation	0.286	0.072	0.161	0.020	0.032	0.035	0.665	0.118
(III) Exact	0.276	0.093	0.205	0.022	0.036	0.037	0.587	0.156
(IV) Offspring	0.287	0.076	0.185	0.021	0.035	0.037	0.640	0.124
BH-EM	0.306	0.053	0.189	0.020	0.034	0.035	0.695	0.082
INAR	2.784	0.256	6.885	0.265	0.782	0.155	0.686	0.319
Whittle	6.019	4.521	15.705	2.417	0.201	0.139	2.152	7.447



(a) HP($\boldsymbol{\theta}_1$). (b) HP($\boldsymbol{\theta}_4$).

FIG. 34. MAPE $\bar{e}_{\boldsymbol{\theta}}$ for an observation horizon $T \in \{1000, 2000, 4000, 8000\}$.

2. Expected average arrival rate

For an even comparison of Hawkes processes generated by different parameter vectors $\boldsymbol{\theta} = (\mu, \alpha, \beta)$, the expected number of events over the given observation interval should be constant, at least approximately. Thus, consider the expected number of arrivals $\mathbb{E}[N(T)]$ for a given horizon $T > 0$, which can be written in the form

$$\mathbb{E}[N(T)] = \mathbb{E}\left[\int_0^T \lambda(t)\,dt\right] = \int_0^T \mathbb{E}[\lambda(t)]\,dt = \bar{\lambda}_T T,$$

where $\bar{\lambda}_T = (1/T)\int_0^T \mathbb{E}[\lambda(t)]\,dt$ is the expected average arrival rate on the interval $(0, T]$. To compute the expected arrival rate we follow Chen *et al.* [55] using the Laplace transform.[42] Indeed, denoting $\varphi(t) = \mathbb{E}[\lambda(t)]$, for $t \geqslant 0$, we retrieve the Laplace transform $\tilde{\varphi}(s)$ from Eq. (1) for $s \in \mathbb{C}$, with the exponential kernel specified in Eq. (2), by first taking the expectation[43] and then the Laplace transform, so

$$\tilde{\varphi}(s) = \int_0^\infty e^{-st}\left[\mu + \alpha\beta\int_0^t e^{-\beta(t-\vartheta)}\varphi(\vartheta)\,d\vartheta\right]dt, \quad s \in \mathbb{C}.$$

Recalling that a convolution in the Laplace domain becomes a multiplication in the time domain, one obtains

$$\tilde{\varphi}(s) = \frac{\mu}{s} + \left(\frac{\alpha\beta}{s+\beta}\right)\tilde{\varphi}(s), \quad s \in \mathbb{C}.$$

This implies, via partial fraction decomposition,

$$\tilde{\varphi}(s) = \frac{\mu}{s}\cdot\frac{s+\beta}{s+(1-\alpha)\beta} = \frac{\mu}{1-\alpha}\left(\frac{1}{s} - \frac{\alpha}{s+(1-\alpha)\beta}\right), \quad s \in \mathbb{C}.$$

Taking the inverse Laplace transform of the preceding expression yields

$$\varphi(t) = \mathbb{E}[\lambda(t)] = \frac{\mu}{1-\alpha}(1 - \alpha\, e^{-(1-\alpha)\beta t}), \quad t \geqslant 0,$$

whence, via integration, we obtain the expected average arrival rate on $(0, T]$:

$$\bar{\lambda}_T = \frac{1}{T}\int_0^T \varphi(t)\,dt = \frac{\mu}{1-\alpha}\left(1 - \frac{1}{T}\frac{\alpha}{(1-\alpha)\beta}(1 - e^{-(1-\alpha)\beta t})\right), \quad T > 0.$$

Thus, for a sufficiently large observation horizon T, the expected average arrival rate on $(0, T]$, approximates the long-run expected average arrival rate $\bar{\lambda}$:

$$\bar{\lambda}_T \approx \bar{\lambda} = \lim_{T\to\infty}\bar{\lambda}_T = \lim_{T\to\infty}\frac{\mathbb{E}[N(T)]}{T} = \frac{\mu}{1-\alpha}.$$

As a result, different processes HP($\boldsymbol{\theta}$), with exponential kernel parameter vectors $\boldsymbol{\theta} = (\mu, \alpha, \beta)$, can be expected to produce a comparable number of events per observation, as long as $\mu/(1-\alpha)$ remains constant, consistent with our choice of the process parameters $\boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_4$ (cf. Table I in Sec. III B).

[42] For details on the derivation of moments for Hawkes processes, see Cui *et al.* [56].

[43] An alternative expression for Eq. (1) is $\lambda(t|\mathcal{H}_t) = \mu + \int_0^t \phi(t-\vartheta)\,dN(\vartheta)$ for $t \geqslant 0$, where the counting process $N(t) = \int_0^t dN(\vartheta)$ is induced by the sample-path realization $\mathcal{H}_t$. Thus, $\mathbb{E}[N(t)] = \int_0^t \varphi(\vartheta)\,d\vartheta$.

[1] P. Reynaud-Bouret and S. Schbath, Adaptive estimation for Hawkes processes; application to genome analysis, Ann. Statist. **38**, 2781 (2010).

[2] N. Chehrazi and T. A. Weber, Dynamic valuation of delinquent credit-card accounts, Manag. Sci. **61**, 3077 (2015).

[3] E. Lewis, G. Mohler, P. J. Brantingham, and A. L. Bertozzi, Self-exciting point process models of civilian deaths in Iraq, Secur. J. **25**, 244 (2012).

[4] Y. Ogata and J. Zhuang, Space–time ETAS models and an improved extension, Tectonophysics **413**, 13 (2006).

[5] A. L. Bertozzi, E. Franco, G. Mohler, M. B. Short, and D. Sledge, The challenges of modeling and forecasting the spread of COVID-19, Proc. Natl. Acad. Sci. USA **117**, 16732 (2020).

[6] M. Kim, D. Paini, and R. Jurdak, Modeling stochastic processes in disease spread across a heterogeneous social system, Proc. Natl. Acad. Sci. USA **116**, 401 (2019).

[7] M. Farajtabar, Y. Wang, M. Gomez-Rodriguez, S. Li, H. Zha, and L. Song, COEVOLVE: A joint point process model for information diffusion and network evolution, J. Mach. Learn. Res. **18**, 1 (2017).

[8] L. Xu, J. A. Duan, and A. B. Whinston, Path to purchase: A mutually exciting point process model for online advertising and conversion, Manag. Sci. **60**, 1392 (2014).

[9] W. Truccolo, U. T. Eden, M. R. Fellows, J. P. Donoghue, and E. N. Brown, A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects, J. Neurophysiol. **93**, 1074 (2005).

[10] K. Zhou, H. Zha, and L. Song, Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes, in *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics, Scottsdale, Arizona, USA*, *Proceedings of Machine Learning Research (PMLR)* (MLR Press, 2013), Vol. 31, pp. 641–649.

[11] M. E. Kretzschmar, G. Rozhnova, M. C. Bootsma, M. van Boven, J. H. van de Wijgert, and M. J. Bonten, Impact of delays on effectiveness of contact tracing strategies for COVID-19: A modelling study, Lancet Public Health **5**, e452 (2020).

[12] A. G. Hawkes, Point spectra of some mutually exciting point processes, J. R. Stat. Soc.: Ser. B (Methodol.) **33**, 438 (1971).

[13] A. G. Hawkes, Spectra of some self-exciting and mutually exciting point processes, Biometrika **58**, 83 (1971).

[14] Y. Ogata, The asymptotic behaviour of maximum likelihood estimators for stationary point processes, Ann. Inst. Stat. Math. **30**, 243 (1978).

[15] I. Rubin, Regular point processes and their detection, IEEE Trans. Inf. Theory, **18**, 547 (1972).

[16] A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc.: Ser. B (Methodol.) **39**, 1 (1977).

[17] G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita, Self-exciting point process modeling of crime, J. Am. Stat. Assoc. **106**, 100 (2011).

[18] A. Veen and F. P. Schoenberg, Estimation of space–time branching process models in seismology using an EM–type algorithm, J. Am. Stat. Assoc. **103**, 614 (2008).

[19] M. Mark and T. A. Weber, Robust identification of controlled Hawkes processes, Phys. Rev. E **101**, 043305 (2020).

[20] B. Mark, G. Raskutti, and R. Willett, Network estimation from point process data, IEEE Trans. Inf. Theory **65**, 2953 (2018).

[21] T. Ozaki, Maximum likelihood estimation of Hawkes' self-exciting point processes, Ann. Inst. Stat. Math. **31**, 145 (1979).

[22] M. Kirchner, Hawkes and INAR(∞) processes, Stoch. Proc. Appl. **126**, 2494 (2016).

[23] M. Kirchner, An estimation procedure for the Hawkes process, Quant. Financ. **17**, 571 (2017).

[24] M. Kirchner and A. Bercher, A nonparametric estimation procedure for the Hawkes process: Comparison with maximum likelihood estimation, J. Stat. Comput. Simul. **88**, 1106 (2018).

[25] F. Cheysson and G. Lang, Spectral estimation of Hawkes processes from count data, Ann. Statist. **50**, 1722 (2022).

[26] P. Whittle, On stationary processes in the plane, Biometrika **41**, 434 (1954).

[27] K. Dzhaparidze, *Parameter Estimation and Hypothesis Testing in Spectral Analysis of Stationary Time Series* (Springer, New York, 1986).

[28] M.-A. Rizoiu, A. Soen, S. Li, P. Calderon, L. Dong, A. K. Menon, and L. Xie, Interval-censored Hawkes processes, J. Mach. Learn. Res. **23**, 1 (2022).

[29] P. Calderon, A. Soen, and M.-A. Rizoiu, Linking across data granularity: Fitting multivariate Hawkes processes to partially interval-censored data, arXiv:2111.02062 (2022).

[30] L. Shlomovich, E. A. Cohen, N. Adams, and L. Patel, Parameter estimation of binned Hawkes processes, J. Comput. Graphical Stat. **31**, 990 (2022).

[31] G. C. Wei and M. A. Tanner, A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms, J. Am. Stat. Assoc. **85**, 699 (1990).

[32] T. Söderström, L. Ljung, and I. Gustavsson, A theoretical analysis of recursive identification methods, Automatica **14**, 231 (1978).

[33] R. Mortensen, Maximum-likelihood recursive nonlinear filtering, J. Optim. Theory Appl. **2**, 386 (1968).

[34] E. Özkan, F. Lindsten, C. Fritsche, and F. Gustafsson, Recursive maximum likelihood identification of jump Markov nonlinear systems, IEEE Trans. Signal Process. **63**, 754 (2014).

[35] A. Wehrli and D. Sornette, The excess volatility puzzle explained by financial noise amplification from endogenous feedbacks, Sci. Rep. **12**, 18895 (2022).

[36] S. Wheatley, A. Wehrli, and D. Sornette, The endo–exo problem in high frequency financial price fluctuations and rejecting criticality, Quant. Financ. **19**, 1165 (2019).

[37] E. Bacry, I. Mastromatteo, and J.-F. Muzy, Hawkes processes in finance, Market Microstruct. Liquidity **01**, 1550005 (2015).

[38] Y. Ogata and H. Akaike, On linear intensity models for mixed doubly stochastic Poisson and self-exciting point processes, J. R. Stat. Soc.: Ser. B (Methodol.) **44**, 102 (1982).

[39] D. Marsan and O. Lengline, Extending earthquakes' reach through cascading, Science **319**, 1076 (2008).

[40] A. G. Hawkes and D. Oakes, A cluster process representation of a self-exciting process, J. Appl. Probab. **11**, 493 (1974).

[41] S. Nandan, G. Ouillon, and D. Sornette, Are large earthquakes preferentially triggered by other large events? J. Geophys. Res.: Solid Earth **127**, e2022JB024380 (2022).

[42] S. Nandan, S. K. Ram, G. Ouillon, and D. Sornette, Is Seismicity Operating at a Critical Point? Phys. Rev. Lett. **126**, 128501 (2021).

[43] M. Achab, E. Bacry, S. Gaïffas, I. Mastromatteo, and J.-F. Muzy, Uncovering causality from multivariate Hawkes integrated cumulants, in *Proceedings of the 34th International Conference on Machine Learning, Sydney, NSW, Australia*, *Proceedings of Machine Learning Research (PMLR)* (MLR Press, 2017), Vol. 70, pp. 1–10.

[44] E. Bacry and J.-F. Muzy, First-and second-order statistics characterization of Hawkes processes and non-parametric estimation, IEEE Trans. Inf. Theory **62**, 2184 (2016).

[45] D. J. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods* (Springer, New York, 2003).

[46] I. M. Gelfand and S. V. Fomin, *Calculus of Variations* (Prentice-Hall, Englewood Cliffs, NJ, 1963).

[47] E. Bacry, M. Bompaire, P. Deegan, S. Gaïffas, and S. V. Poulsen, tick: A Python library for statistical learning, with an emphasis on Hawkes processes and time-dependent models, J. Mach. Learn. Res. **18**, 1 (2018).

[48] Y. Ogata, On Lewis' simulation method for point processes, IEEE Trans. Inf. Theory **27**, 23 (1981).

[49] M. Mark, J. Sila, and T. A. Weber, Quantifying endogeneity of cryptocurrency markets, Eur. J. Finance **28**, 784 (2022).

[50] F. Salehi, W. Trouleau, M. Grossglauser, and P. Thiran, Learning Hawkes processes from a handful of events, in *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada*, Advances in Neural Information Processing Systems (Curran Associates, Red Hook, NY, 2019), Vol. 32, pp. 12715–12725.

[51] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences* (Lawrence Erlbaum Associates, Hillsdale, NJ, 1988).

[52] A. Saichev and D. Sornette, Generation-by-generation dissection of the response function in long memory epidemic processes, Eur. Phys. J. B **75**, 343 (2010).

[53] S. Rustler, Shocks and self-excitation in Twitter: Response function characterization in epidemic processes in social media, Master's thesis, ETH Zurich, Zurich, Switzerland, 2014.

[54] P. Morzywołek, Non-parametric methods for estimation of Hawkes process for high-frequency financial data, Master's thesis, ETH Zurich, Zurich, Switzerland, 2015.

[55] J. Chen, A. Hawkes, E. Scalas, and M. Trinh, Performance of information criteria for selection of Hawkes process models of financial data, Quant. Financ. **18**, 225 (2018).

[56] L. Cui, A. Hawkes, and H. Yi, An elementary derivation of moments of Hawkes processes, Adv. Appl. Probab. **52**, 102 (2020).