Learning Hawkes Processes Under Synchronization Noise

William Trouleau [1] Jalal Etesami [2] Matthias Grossglauser [1] Negar Kiyavash [3 4] Patrick Thiran [1]

Abstract

Multivariate Hawkes processes (MHP) are widely used in a variety of fields to model the occurrence of discrete events. Prior work on learning MHPs has only focused on inference in the presence of perfect traces without noise. We address the problem of learning the causal structure of MHPs when observations are subject to an unknown delay. In particular, we introduce the so-called synchronization noise, where the stream of events generated by each dimension is subject to a random and unknown time shift. We characterize the robustness of the classic maximum likelihood estimator to synchronization noise, and we introduce a new approach for learning the causal structure in the presence of noise. Our experimental results show that our approach accurately recovers the causal structure of MHPs for a wide range of noise levels, and significantly outperforms classic estimation methods.

1. Introduction

Multivariate Hawkes processes (MHPs) are a type of temporal point process where an arrival in one dimension can affect future arrivals in other dimensions. The origin of MHPs dates back to Hawkes (1971), who used them to statistically model earthquakes. Because of their ability to capture mutual excitation between different dimensions of a multivariate counting process, MHPs have become a popular model in a plethora of applications such as finance (Bacry et al., 2012; Hardiman et al., 2013; Linderman & Adams, 2014; Bacry et al., 2015; Etesami et al., 2016), computational biology (Reynaud-Bouret et al., 2010), social network studies (as an alternative for the contagion model) (Yang & Zha, 2013; Farajtabar et al., 2015), and criminology (Mohler et al., 2011; Porter & White, 2012; Linderman & Adams, 2014; Shelton et al., 2018).

Learning the excitation matrix of a MHP, which encodes the causal structure between the processes from a set of observations, has been the focus of recent work (Xu et al., 2016; Etesami et al., 2016). The main approaches for learning MHPs are of two flavors. Maximum likelihood-based approaches estimate the parameters from observations (Ozaki, 1979; Zhou et al., 2013b; Yang et al., 2017); and approaches based on second-order statistics learn the parameters of interest by solving a set of equations obtained from first and second-order statistics of the MHP (Hawkes, 1971; Bacry et al., 2012; Etesami et al., 2016). All the aforementioned work assumes that the observations are noiseless, that is, the arrival times of the events are recorded accurately without any delay. To the best of our knowledge, no work to date has considered learning the causal structure of a noisy MHP. Recent studies tackled the inference of Hawkes processes with missing data (Xu et al., 2017; Shelton et al., 2018), but did not consider noisy (delayed) observations. The inference of temporal point processes in the presence of noisy observations has been studied for non-parametric estimators of spatial Poisson processes (Cucala, 2008; Bar-Hen et al., 2013). However, these studies mostly focus on the special case of independent and known noise and cannot be applied to MHPs.

We study the problem of learning MHPs in the presence of observation noise. More precisely, we consider *synchronization noise*, where the stream of events generated by each source – or dimension – is subject to a random and unknown time shift. This model captures situations where no perfect clock time synchronization is available at different sources, or when the observation process itself introduces source-dependent delays. As an example of the former, consider a network of sensors that record events such as neural spikes or earthquake shocks. It is often the case that the sensors are not perfectly synchronized, because they each rely on a local clock to time-stamp events. As an example of the latter, consider processes where an event can only be observed indirectly after a delay, such as through the symptoms of an infectious disease that manifest themselves some time after the actual infection. We will show that synchronization noise can severely harm the estimation performance of

[1] School of Computer and Communication Sciences, EPFL, Lausanne, Switzerland [2] Bosch Center for Artificial Intelligence [3] Dept. of Electrical and Computer Eng. (ECE), Georgia Institute of Technology [4] Dept. of Industrial and Systems Eng. (ISyE), Georgia Institute of Technology. Correspondence to: William Trouleau <william.trouleau@epfl.ch>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

state-of-the-art learning methods.

1.1. Summary of Results and Organization

Our contribution in this paper is two-fold. First, we show the vulnerability of the state-of-the-art learning algorithms to noisy observations. Second, we provide a novel estimation approach for learning the causal structure of a MHP in the presence of synchronization noise. Unlike previous works on the inference of point processes with noise (Cucala, 2008; Bar-Hen et al., 2013), our approach does not assume that the noise is sampled from a known distribution. Our approach is based on the maximum-likelihood estimation of a novel model called desynchronized multivariate Hawkes process (DESYNC-MHP) in which the parameters of interest consist of the MHP parameters along with the noise. In other words, given a set of observed data, our approach learns the MHP with synchronization noise that maximizes the log-likelihood with respect to both the noise and the MHP parameters. Such log-likelihood function is smooth with respect to the MHP parameters, yet non-convex and non-smooth with respect to the noise parameters. We show that maximizing a smoothed version of this objective function with respect to both the noise and the MHP parameters recovers the excitation matrix and hence the causal structure of the MHP.

The paper is organized as follows. In Section 2, we provide some preliminary definitions and notations. We introduce the synchronization noise in Section 3 and show how it biases the classic maximum likelihood estimation algorithm that assumes the observations to be noiseless. In Section 4, we introduce our methodology to learn Hawkes processes under synchronization noise. Finally, we demonstrate the performance of our approach on synthetic simulations, and we validate it on a dataset of neuronal spike trains in Section 5.

2. Preliminaries

Prior to discussing our results, we introduce the basic notations and definitions used in the paper. Detailed notations will be introduced along the way.

Multivariate Hawkes process (MHP). Formally, a d-dimensional MHP is a collection of d univariate temporal point processes $N_i(t)$, $i = 1, \dots, d$, also called dimension, with a particular form of the conditional intensity function

$$\lambda_i(t|\mathcal{H}_t) = \mu_i + \sum_{j=1}^{d} \sum_{\tau \in \mathcal{H}_t^j} \kappa_{ij}(t - \tau), \tag{1}$$

where $\mathcal{H}_t^j$ is the history of the j-th process up to time t and $\mathcal{H}_t = \bigcup_{i=1}^{d} \mathcal{H}_t^i$. The constant μ_i is the exogenous part of the intensity of the i-th process. The excitation function $\kappa_{ij}(t) \geq 0$ captures the endogenous dynamics of influence of the arrivals in the j-th dimension on the intensity of the i-th dimension.

The matrix $K(t) := [\kappa_{ij}(t)]$ is called the excitation matrix. It has been shown that the support of the excitation matrix encodes the causal structure of the MHP, *i.e.,* process j does not cause process i if and only if $\kappa_{ij}(t) = 0$ (Etesami et al., 2016; Eichler et al., 2017). The causal graph of a d-dimensional MHP is therefore a directed graph on d nodes (each dimension is denoted by a node) and there is a directed edge from node j to node i if and only if $\kappa_{ij}(t) \neq 0$. For more details on MHPs, we refer the interested reader to (Liniger, 2009).

A common choice for the excitation function is the exponential kernel

$$\kappa_{ij}(t) = \alpha_{ij} e^{-\beta t} \mathbb{1}\{t > 0\}, \tag{2}$$

where α_{ij} captures the strength of influence and β captures the time constant (Rasmussen, 2013; Zhou et al., 2013a; Farajtabar et al., 2014; Yan et al., 2015; Shelton et al., 2018). We present our learning approach for exponential kernels, but it is applicable to more general forms of kernels.

Likelihood function of a MHP. Suppose that we observed a sequence of discrete events

$$\boldsymbol{t} := \left\{ \{t_k^i\}_{k=0}^{n_i} \right\}_{i=1}^{d}$$

during a time period $[t_0, T)$, where t_k^i denotes the k-th arrival in the i-th dimension. Let θ denote the parameters of the MHP, which consist of the excitation matrix $\{\alpha_{ij}\}$ and the background intensities $\{\mu_i\}$. Maximum likelihood estimation can be used to learn θ from the observations $\boldsymbol{t}$ (Zhou et al., 2013a; Farajtabar et al., 2014). The log-likelihood of $\boldsymbol{t}$ given θ for a MHP is given by

$$\log \mathbb{P}(\boldsymbol{t}|\theta) = \sum_{i=1}^{d} \left[\sum_{\tau \in \mathcal{H}_T^i} \log \lambda_i(\tau|\mathcal{H}_\tau) - \int_{t_0}^{T} \lambda_i(t|\mathcal{H}_t) dt \right]. \tag{3}$$

It can be shown that the log-likelihood function of Hawkes processes with exponential kernels is convex if the exponential decay β is known (Bacry et al., 2015). It is therefore common practice to define β as a hyper-parameter and to apply maximum likelihood estimation only to

$$\theta := \{\{\mu_i\}_{i=1}^{d}, \{\alpha_{ij}\}_{i,j=1}^{d}\} \in \mathbb{R}_+^{d(d+1)}.$$

As noted before, in the remainder of this paper, we assume that the excitation functions are exponential, as defined in (2).

3. Noisy Observation Framework

In this section, we introduce a particular form of noise, called synchronization noise. We demonstrate its destructive effect on the classic maximum likelihood (ML) estimation methodology, which assumes noiseless observations.

3.1. Synchronization Noise

With synchronization noise, all the arrivals within a dimension are shifted equally by an unknown offset. In other words, for every dimension i, there exists z_i, such that the observed data is

$$\tilde{\boldsymbol{t}} := \{\{\tilde{t}_k^i\}_{k=0}^{\tilde{n}_i}\}_{i=1}^d,$$

where the observed arrival times are not equal to the true arrival times $\boldsymbol{t}$ but instead are related as

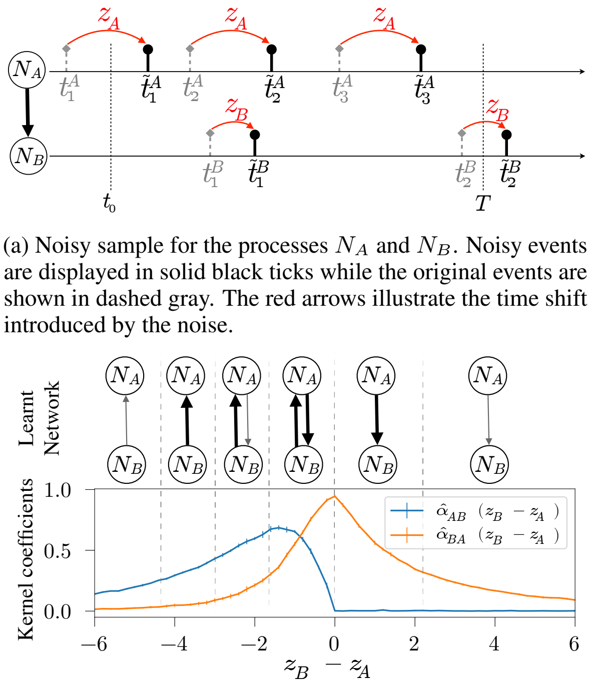
$$\tilde{t}_k^i - t_k^i = z_i \in \mathbb{R}, \forall\, i, k.$$

We denote the collection of noise variables by $\boldsymbol{z} = \{z_i\}_{i=1}^d$. Because of boundary effects due to the finite observation window, the number of noisy observations $\tilde{n}_i$ may differ from n_i as some events can enter or escape the observation window.

To make this more concrete, Figure 1a shows a simple example of the synchronization noise for a 2-dimensional MHP $\{N_A, N_B\}$. The synchronization noise $\{z_A, z_B\}$ do not change the relative orders of the arrivals within a dimension but it affects the relative orders of the arrivals between different dimensions. For instance, in Figure 1a, $t_2^A < t_1^B$ but $t_2^A + z_A = \tilde{t}_2^A > \tilde{t}_1^B = t_1^B + z_B$. Some events can also enter (or escape) the observation window, such as t_1^A (or t_2^B).

3.2. Effect of Noise on Classic Inference Methods

The synchronization noise may swap the relative order of arrivals between different dimensions, which results in estimation errors for classic inference methods, such as ML estimation. Consider once again the simple network of two processes shown in Figure 1a. In this example, the causal graph contains a single edge $N_A \to N_B$, implying that events in process N_A cause future events in process N_B (but not the other way around). Figure 1b displays the result of ML estimation with synchronization noise for these two processes. When $z_A < z_B$, events in N_B tend to occur after their cause (parent) events in N_A, which leads ML estimation to correctly identify the causal direction $N_A \to N_B$. However, as $z_A > z_B$, the causes and effects begin to blur. This forces ML estimation to learn edges in both directions. Finally, as the difference between z_A and z_B gets large, the inferred dependency between N_A and N_B decreases. This is the reason explaining the convergence of the kernel coefficients to zero.



(a) Noisy sample for the processes N_A and N_B. Noisy events are displayed in solid black ticks while the original events are shown in dashed gray. The red arrows illustrate the time shift introduced by the noise.

(b) Maximum likelihood estimate on the toy example of Fig. 1a as a function of noise values. When $z_B - z_A < 0$, ML detects edges in both directions, i.e., $\hat{\alpha}_{AB}$ and $\hat{\alpha}_{BA}$ are both positive.

Figure 1: Illustration of the synchronization noise model on a simple two-dimensional Hawkes process, with process N_A influencing process N_B.

4. Inference under Synchronization Noise

In this section, we introduce a new robust inference approach for learning MHPs in the presence of synchronization noise.

4.1. Model: Desynchronized Multivariate Hawkes Processes (DESYNC-MHP)

We first note that, if the value of the noise $\boldsymbol{z}$ is known, we can simply subtract the value of the noise from each arrival time, and the problem reduces to the inference of a standard (noiseless) MHP. Conditioning on the noise $\boldsymbol{z}$, the log-likelihood (3) can hence be written as the conditional log-likelihood

$$\log \mathbb{P}(\tilde{\mathbf{t}}|\boldsymbol{z}, \theta) = \log \mathbb{P}\left(\{\{\tilde{t}_k^i - z_i\}_{k=0}^{\tilde{n}_i}\}_{i=1}^d \middle| \theta\right)$$
$$= \sum_{i=1}^d \left[\sum_{\tau \in \widetilde{\mathcal{H}}_T^i} \log \lambda_i(\tau - z_i | \widetilde{\mathcal{H}}_{\tau - z_i}) - \int_{t_0 - z_i}^{T - z_i} \lambda_i(t | \widetilde{\mathcal{H}}_t) dt \right], \tag{4}$$

where $\widetilde{\mathcal{H}}_t^i = \{\tilde{t}_k^i \mid \tilde{t}_k^i = t_k^i + z_i < t\}$ is the history of the i-th observed (noisy) process up to time t, and $\widetilde{\mathcal{H}}_t = \bigcup_{i=1}^d \widetilde{\mathcal{H}}_t^i$.

It is important to notice that (4) is a function of the observed history $\widetilde{\mathcal{H}}_t$ due to the conditional intensity function terms. Since the synchronization noise can change the order of the arrivals in different dimensions and consequently the value of the conditional intensity function, it can also change the above conditional log-likelihood. Hence, the noise offset $\boldsymbol{z}$ affects the MHP parameters θ maximizing (4).

We define a new multivariate point process called *desynchronized multivariate Hawkes process* (DESYNC-MHP) that is a MHP with synchronization noise. The parameters of this model are $(\boldsymbol{z}, \theta)$. In other words, a DESYNC-MHP with parameters $(\boldsymbol{z}, \theta)$ is a MHP with parameter θ, where each dimension i is affected by the synchronization noise offset z_i. Therefore, the log-likelihood function of this model, given a set of observed arrivals $\tilde{\boldsymbol{t}}$, can be written as (4). Hence, ML estimation for the DESYNC-MHP amounts to solving the optimization problem

$$\hat{\boldsymbol{z}}, \hat{\theta} = \underset{\boldsymbol{z}\in\mathbb{R},\theta\geq 0}{\operatorname{argmax}} \log \mathbb{P}(\tilde{\mathbf{t}}|\boldsymbol{z}, \theta). \tag{5}$$

An alternative approach to directly maximizing the log-likelihood is to consider the noise as a latent variable and to use the EM algorithm. However, such an approach requires to evalue the posterior distribution, which is intractable because of its coupling with the ordering of the events. It is therefore easier to solve (5) directly. This approach still introduces new challenges that we will address next.

4.2. Challenges

For a *given* noise variable $\boldsymbol{z}$, maximizing (4) with respect to the Hawkes parameters θ results in the ML estimation for the noiseless MHP, which can be often solved efficiently. For instance, in the exponential kernel setting, when $\theta = \{\{\mu_i\}_{i=1}^d, \{\alpha_{ij}\}_{i,j=1}^d\}$, the problem is smooth and convex, and therefore the parameters can be easily estimated using first-order methods.

In contrast, the objective function in (4) is neither smooth nor continuous with respect to the noise $\boldsymbol{z}$. Recall that the intensity function (1) depends on the history $\widetilde{\mathcal{H}}_t$ of the process. However, synchronization noise can invert the order of arrivals in different dimensions, and consequently it can change the past events of some arrivals, which creates discontinuities in the likelihood.

To observe this concretely, consider a 2-dimensional MHP with only two arrival times t_1 and t_2 ($t_1 < t_2$), in dimensions 1 and 2, respectively. Suppose that the observed arrival times $\tilde{t}_1$ and $\tilde{t}_2$, are such that $\tilde{t}_1 < \tilde{t}_2$. The effect of dimension 1 on dimension 2 is captured by

$$\begin{aligned}&\kappa_{21}(\tilde{t}_2 - z_2 - \tilde{t}_1 + z_1)\\ &= \alpha_{21}\, e^{-\beta(\tilde{t}_2 - z_2 - \tilde{t}_1 + z_1)}\, \mathbb{1}\{\tilde{t}_2 - z_2 - \tilde{t}_1 + z_1 > 0\}.\end{aligned}$$

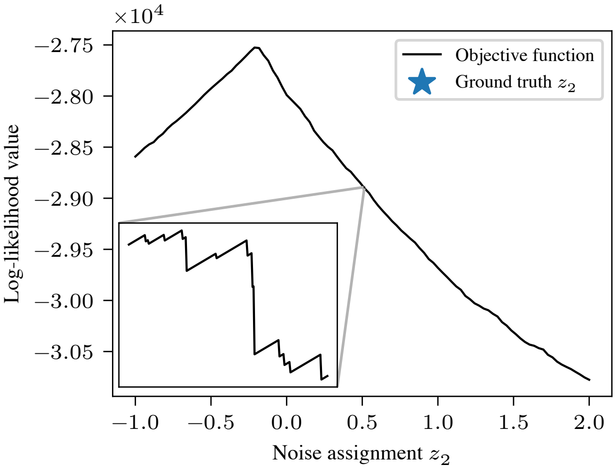


Figure 2: Illustration of the discontinuities of the objective function (4) for a two-dimensional MHP as a function of z_2, when z_1 is fixed to its true value and $\beta = 1$. The inset shows a fine zoom on the objective function in 0.5 ± 0.005.

Hence, for a given z_1, as z_2 increases, the excitation function increases until $z_2 = \tilde{t}_2 - \tilde{t}_1 + z_1$. At this point, the arrival orders are switched and the effect of the arrival at t_1 on the arrival at t_2 disappears. Formally, at $\tau = \tilde{t}_2 - z_2 - \tilde{t}_1 + z_1$, we have

$$\lim_{\tau\to 0^+} \kappa_{21}(\tau) = \alpha_{21} \neq 0 = \lim_{\tau\to 0^-} \kappa_{21}(\tau).$$

This results in a discontinuity in the objective function.

Figure 2 illustrates the objective function as a function of z_2, when z_1 is fixed to its true value, for a two-dimensional process. These discontinuities in the conditional log-likelihood function will prevent gradient-based algorithms from converging. Even worse, the objective function is particularly ill-conditioned: it decreases at the points of discontinuity, but increases everywhere in between. The presence of synchronization noise therefore transforms the computationally efficient estimation of MHP parameters into a particularly ill-conditioned optimization problem.

Below, we discuss our approach to tackle this issue in two steps. We first introduce a novel approach for smoothing the objective function, which allows us to subsequently find an optimum solution by using stochastic gradient descent.

Smoothing the objective function. Recall that the source of the discontinuities (jumps) in the objective function are the swapped arrivals and the discontinuities of the excitation kernels at $t = 0$. If the excitation kernels $\{\kappa_{ij}(t)\}$ were differentiable for all $t \in \mathbb{R}$, such sudden jumps in the intensity function would be avoided and consequently the likelihood function would be smooth. This observation leads us to approximate the excitation kernels with functions that are

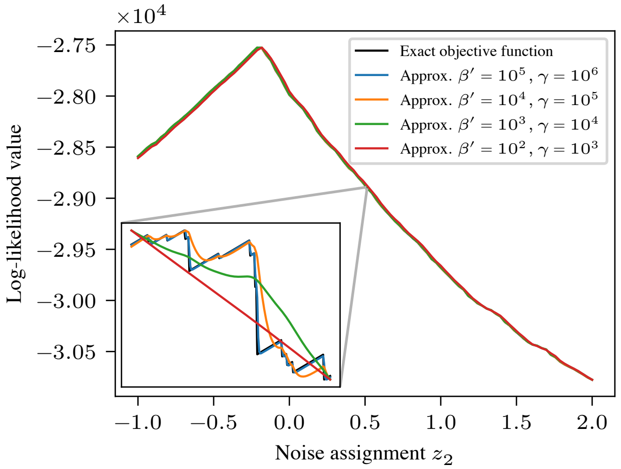


Figure 3: Illustration of the smoothing of the objective function (4) for a two-dimensional MHP as a function of z_2, when $z_1 = z_1^*$ and $\beta = 1$. The inset shows a fine zoom on the objective function in 0.5 ± 0.005.

differentiable everywhere. For instance, one candidate for approximating the exponential kernel is

$$\tilde{\kappa}_{ij}(t) \triangleq \alpha_{ij} \left(\sigma(\gamma t) e^{-\beta t} + (1 - \sigma(\gamma t)) e^{\beta' t} \right), \quad (6)$$

where $\sigma(t) = 1/(1 + e^{-t})$ is the sigmoid function[1]. Since $\lim_{\beta', \gamma \to +\infty} \tilde{\kappa}_{ij}(t) = \kappa_{ij}(t)$, the approximated kernel can be made arbitrarily close to $\kappa_{ij}(t)$. Selecting β' and γ large enough will therefore preserve the causal structure of the MHP. Figure 3 illustrates how $\tilde{\kappa}_{ij}(t)$ affects the objective function for various values of β' and γ.

Stochastic gradient descent. The kernel approximation (6) addresses the non-smoothness of the objective function with respect to the noise $\boldsymbol{z}$. But the issue of convexity remains, as illustrated in the inset of Figure 3 for large values of β'. This means that choosing the right β' is crucial. On the one hand, a small β' makes the objective function smoother and removes some local minima. On the other hand, a small β' degrades the quality of the approximation and hence introduces a larger bias in the optimization problem.

Stochastic gradient descent (SGD) is often used to escape local minima in non-convex optimization. In our case, SGD randomizes the discontinuities, and hence enables us to evade the local minima. We apply a mini-batch version of SGD with a set of C independent observations $\{\tilde{\mathbf{t}}_1, \ldots, \tilde{\mathbf{t}}_C\}$. Due to the ergodicity of stationary MHPs, a set of short independent observations of an MHP is statistically equivalent to a single long observation of that MHP.

Algorithm 1 DESYNC-MHP ML estimation

```
Input: Data $\{\tilde{\mathbf{t}}_1, \ldots, \tilde{\mathbf{t}}_C\}$, hyper-parameters $(\beta, \beta', \gamma)$.
Initialize $\boldsymbol{z}_0$ and $\theta_0$ to random values
$k \leftarrow 0$
repeat
    $\tilde{\mathbf{t}}_k \sim \mathrm{Uniform}\{\tilde{\mathbf{t}}_1, \ldots, \tilde{\mathbf{t}}_C\}$
    $\boldsymbol{z}_{k+1} \leftarrow \boldsymbol{z}_k + \delta_k \nabla_{\boldsymbol{z}} \log \widetilde{\mathbb{P}}(\tilde{\mathbf{t}}_k | \boldsymbol{z}_k, \theta_k)$
    $\theta_{k+1} \leftarrow \max\big(\theta_k + \delta_k \nabla_\theta \log \mathbb{P}(\tilde{\mathbf{t}}_k | \boldsymbol{z}_k, \theta_k), 0\big)$
    $k \leftarrow k + 1$
until convergence
```

Algorithm 1 summarizes the steps of our approach[2]. Since smoothing is only necessary for optimizing $\log \mathbb{P}(\tilde{\mathbf{t}} | \boldsymbol{z}, \theta)$ with respect to $\boldsymbol{z}$, we use the gradient[3] of the smooth approximation of the log-likelihood, denoted by $\nabla_{\boldsymbol{z}} \log \widetilde{\mathbb{P}}(\tilde{\mathbf{t}} | \boldsymbol{z}, \theta)$, to update $\boldsymbol{z}$, and we keep the gradient of the exact log-likelihood to update the MHP parameters θ, denoted by $\nabla_\theta \log \mathbb{P}(\tilde{\mathbf{t}}_k | \boldsymbol{z}_k, \theta_k)$.

5. Experimental Results

We performed two sets of experiments. First, we used synthetic data to show that, despite the non-smoothness and non-convexity of (5), our approach can accurately recover the excitation matrix of the MHP and significantly outperform the classic ML estimator. We further investigated the effects of dimensionality d and the scale of the noise on the performance of our estimator. Second, we validated our approach using a dataset of neuronal spike trains obtained from measurements of the motor cortex of a monkey.

5.1. Experiments on Synthetic Data

We set the exponential decay to $\beta = 1$. For smoothing, we used $\beta' = 50$ and $\gamma = 500$, which were found to work well in practice. For each experiment, we chose small positive background intensities $\{\mu_i\}$ and generated a random[4] excitation matrices with entries $\{\alpha_{ij}\} \in \{0, 1\}$ by sampling edges randomly with probability $2/d$. The average in-degree and out-degree of each nodes was hence close to two. We then rescaled the entries to obtain a spectral radius of 0.95 to ensure that the simulated processes are stable[5]. We generated $C = 5$ realizations of $50,000$ samples from the MHP using Ogata's thinning algorithm[6] (Ogata, 2006).

[1] Note that this choice of kernel is non-causal, in the sense that the kernels are non-zero for $t < 0$.

[2] Source code of the algorithm is available publicly.

[3] The derivation of the gradient with respect to the noise parameters and the parameters of MHP is provided in the Appendix.

[4] Experiments were performed on other random graph models with qualitatively similar results.

[5] Experiments were not found to be sensitive to this choice of value.

[6] We used the Python library *tick* to generate synthetic samples of the processes (Bacry et al., 2017).

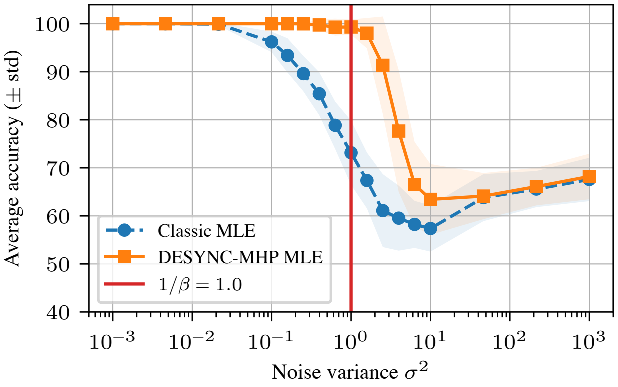


Figure 4: Analysis of the sensitivity to the noise scale with 4 different noise regimes. ($d = 10$ is fixed.)

We repeated each experiment 10 times over 10 different matrices for each set of parameters. We solved the optimization problem (5) using stochastic gradient descent with Lasso regularization on the parameters $\{\alpha_{ij}\}$. We compared our approach against the state-of-the-art maximum likelihood estimation method (Zhou et al., 2013a), which solves the classic maximum likelihood estimation problem with the same regularization (denoted by the label "classic MLE" in the figures below).

The accuracy reported on the y-axes of our figures is the percentage of correctly identified edges (*i.e.,* non-zero kernels in the support of the excitation matrix) which is

$$1 - \frac{\sum_{ij} |\mathbb{1}\{\alpha^*_{ij} > 0\} - \mathbb{1}\{\hat{\alpha}_{ij} > \eta\}|}{d^2},$$

where $\{\alpha^*_{ij}\}$ and $\{\hat{\alpha}_{ij}\}$ denote the ground truth and the estimated coefficients, respectively. A threshold $\eta = 0.05$ was used to zero out the small coefficients.

Sensitivity to the noise level σ^2. We studied the sensitivity of our approach, DESYNC-MHP MLE, to the level of noise and compared it to the classic ML estimator. Figure 4 shows the mean and standard deviation accuracy for difference noise variance σ^2. We observe four different noise regimes:

1. In the low-noise regime, virtually no event order is swapped, meaning that the cause (parent) events always occur before their effect. Both the classic ML estimator and our approach therefore recover the causal structure accurately.
2. When the noise level is increased to $\sigma^2 = 1/\beta = 1$ (indicated by the red vertical line in Figure 4), our approach still recovers the true causal structure with an accuracy close to 100%, contrary to the classic ML estimator which misidentifies more than 25% of edges on average.

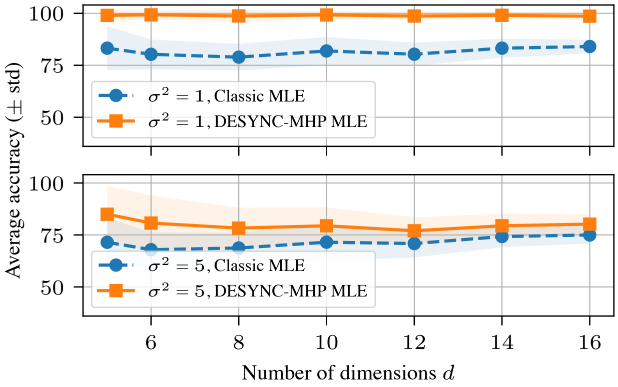


Figure 5: Analysis of the sensitivity to the number of dimensions for two values of noise variance: on the top panel $\sigma^2 = 1$, and on the bottom panel $\sigma^2 = 5$.

3. In the third regime, for noise levels between $\sigma^2 = 1/\beta$ up to one order of magnitude larger than $1/\beta$, our approach gets trapped in local optima more frequently, and hence it loses its accuracy. Yet, it still clearly outperforms the classic ML estimation.

4. In the high-noise regime, the MHP signal gets completely lost in the noise. The log-likelihood function therefore rapidly decreases around the true noise $\boldsymbol{z}_*$ and becomes more and more flat for all $\boldsymbol{z}$ far from $\boldsymbol{z}_*$. Thus, iterative gradient-based algorithms such as Algorithm 1 and the classic ML estimator stay trapped around their initial points $\boldsymbol{z}_0$. Note that our algorithm with fixed $\boldsymbol{z} = \mathbf{0}$ becomes the classic ML algorithm. As the noise variance increases, neither of the two estimators is able to correctly learn the causal structure in the observations, and both algorithms converge toward sparser excitation matrices. More details are given in the Appendix.

   However, between the 3rd and 4th regimes, the noise is not strong enough to completely hide the MHP signal. Consequently, the outputs of the classic ML estimator and DESYNC-MHP ML estimator are driven mostly by the noise and their accuracies are worse than random guesses.

Sensitivity to the number of dimensions d. The number of parameters to estimate grows quadratically with the dimensionality of the process (*i.e.,* $\boldsymbol{z} \in \mathbb{R}^d$, $\theta \in \mathbb{R}^{d^2+d}$). Consequently, the optimization problem becomes harder for larger-sized problems. However, we analyzed the sensitivity of our approach to the number of dimensions d of the MHP in Figure 5. We see that the accuracy of our approach remains fairly constant as we increase the number of dimensions d.

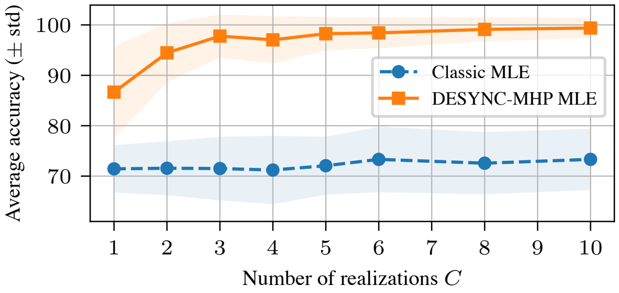


Figure 6: Analysis of the sensitivity to the number of realizations. ($d = 10$, $\sigma^2 = 1$ are fixed.)

Sensitivity to the number of realizations C. Recall that we used SGD in order to evade local minima in the conditional log-likelihood function. Figure 6 shows that with only $C = 3$ independent mini-batches each consisting of $50,000$ samples suffice to obtain an accuracy close to 100%.

5.2. Application to Real Data

In addition to simulations on synthetic data, we also evaluated our approach on an experimental dataset of neuronal spike trains from Wu & Hatsopoulos (2006); Quinn et al. (2011). The dataset consists in measurements of an electrode array located on the motor cortex of a macaque monkey performing a series of tasks involving a specific arm movement. The local field potentials in the motor cortex were recorded and processed to obtain the neuronal spike train data (discrete event times). More details can be found in (Wu & Hatsopoulos, 2006). The dataset contains the spike train data from 115 identified neurons for a duration of an hour, quantized at the resolution of 1 millisecond. Since each spike train was recorded by an independent sensor, some synchronization noise between the dimensions could be expected. For ease of visualization, we kept only a subset of data containing the top $d = 10$ neurons with highest number of spikes, leading to a total of $354\,285$ spikes. We used the first 70% of the dataset for training and kept the last 30% for testing. We set the hyper-parameters (β, β', γ) to $(0.0047, 0.16, 1.6)$ using grid-search.

We compared the predictive log-likelihood on the test set for the models learnt by the baseline classic ML estimator and the DESYNC-MHP ML estimator in Table 1. Since problem (5) is non-convex, the optimization was started from multiple starting points and we report both the average and standard deviation of both estimators.

We see that the DESYNC-MHP ML estimate consistently improves the predictive log-likelihood over the classic ML estimate. Our algorithm identifies a small synchronization noise with an average value of 12.5ms, which is less than the average inter-event time of 88.9ms. The causal graphs learned by the two methods is shown in Figure 7. The two graphs agree on 91% of the edges. In a previous analysis of causality of the dataset, Quinn et al. (2011) identified a dominant direction of influence on both graphs from the lower left to the upper right corner of the array, which might correspond to the direction of propagating local field potential waves discussed in Wu & Hatsopoulos (2006). The causal graphs in Figure 7 are consistent with these findings. A dominant direction is indeed noticeable on both graphs and is particularly striking on the graph learnt by DESYNC-MHP MLE in Figure 7b.

Table 1: Predictive log-likelihood for the models learnt by both approaches. Results are reported averaged over several random initialization points (± standard deviation).

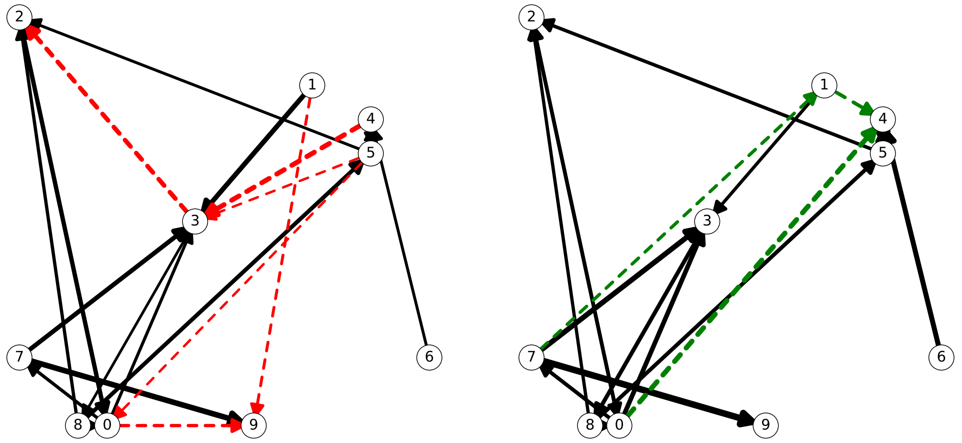
Classic MLE	DESYNC-MHP MLE
$0.4282 \pm 3.5\text{e}{-5}$	$\mathbf{0.4311 \pm 3.0e{-4}}$

To evaluate the robustness of our approach to larger synchronization noise, we added additional shifts the arrivals in different dimensions randomly with various noise variances σ^2 and computed the predictive log-likelihood both for our algorithm and for the classic ML estimator. The results are reported in Figure 8. We identify different noise regimes. For low noise, with a variance smaller than $\sigma^2 = 10\text{ms}$, DESYNC-MHP MLE consistently leads to more likely estimate than the classic MLE. This is consistent with the log-likelihood values computed in Table 1. For higher noise variance, the likelihood of both approaches decreases, but the DESYNC-MHP ML estimate always outperforms the classic one. It is interesting to note that, on this dataset, the shift in noise regime occurs before $1/\beta$. This might come from the noise initially present in the data.

Although our approach shows better results compared to the classic MLE, the gains are not as large as in the case of the synthetic experiments. Since our approach is not limited to the exponential kernel, results could certainly be improved by using a more flexible form of excitation function. For instance, using non-parametric learning approaches for Hawkes processes inspired by Zhou et al. (2013b); Yang et al. (2017) might better fit the true excitation dynamics of the neurons.

6. Conclusion

We addressed the problem of learning the causal structure of multivariate Hawkes processes (MHP) under synchronization noise, which can arise both for technical reasons or as a feature of the observation process. We showed that the classic maximum likelihood (ML) estimator fails when observations are noisy, because delays perturb the order of events across dimensions. In particular, we showed that,



(a) Causal graph learned by the classic MLE.

(b) Causal graph learned by DESYNC-MHP MLE.

Figure 7: Causal graphs of the neuronal spike train dataset. Each node indicates a different neuron. The relative position of the nodes corresponds to the relative position of the electrode on the array. The differences between the two graphs is highlighted with dashed edges. Edges appearing only in the classic ML estimate are highlighted in red in Figure 7a, and edges appearing only in the DESYNC-MHP ML estimate are highlighted in green in Figure 7b. The labels of the nodes correspond to the ordering of the neurons sorted by number of observed events.

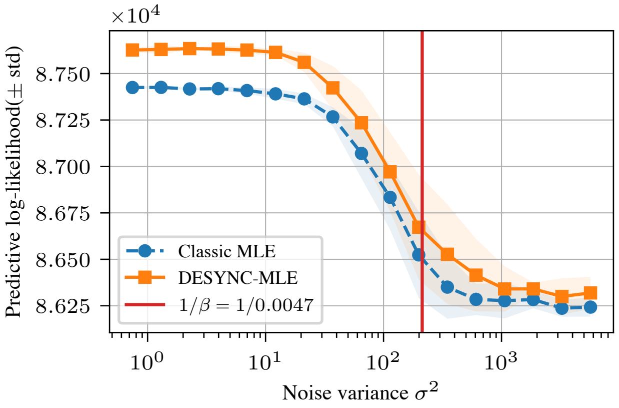


Figure 8: Analysis of the sensitivity to the noise scale on the neuronal spike train dataset.

even with small noise with variance $\sigma^2 \approx 1/\beta$, the classic ML estimator misidentifies on average more than 25% of the edges in the causal structure of a random network.

To address the learning problem in the presence of noisy observations, we introduced a novel multivariate point process, called DESYNC-MHP, which is a MHP with synchronization noise. A DESYNC-MHP with parameters $(\boldsymbol{z}, \theta)$ is a MHP with parameters θ, where each dimension i is affected by the synchronization noise offset z_i. The log-likelihood function of DESYNC-MHP is non-smooth and non-continuous with respect to the noise, making off-the-shelf gradient-based approaches infeasible. We introduced a novel smoothing approach based on a smooth approximation of the excitation kernels, in conjunction with SGD, to tackle the problem. The experimental results show that, despite the non-convexity of the objective, our approach significantly outperforms the classic ML estimator and accurately recovers the causal structure of MHPs for a wide range of noise.

Acknowledgements

The work presented in this paper was supported in part by the Swiss National Science Foundation under grant number 200021-182407, by MURI grant ARMY W911NF-15-1-0479, and by ONR grant W911NF-15-1-0479.

References

Bacry, E., Dayri, K., and Muzy, J.-F. Non-parametric kernel estimation for symmetric Hawkes processes. application to high frequency financial data. *The European Physical Journal B-Condensed Matter and Complex Systems*, 85 (5):1–12, 2012.

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

Bacry, E., Bompaire, M., Gaïffas, S., and Poulsen, S. tick: a Python library for statistical learning, with a particular emphasis on time-dependent modeling. *ArXiv e-prints*, July 2017.

Bar-Hen, A., Chadœuf, J., Dessard, H., and Monestiez, P. Estimating second order characteristics of point processes with known independent noise. *Statistics and Computing*, 23(3):297–309, May 2013. ISSN 1573-1375. doi: 10.1007/s11222-011-9311-7. URL https://doi.org/10.1007/s11222-011-9311-7.

Cucala, L. Intensity estimation for spatial point processes observed with noise. *Scandinavian Journal of Statistics*, 35:322–334, 06 2008. doi: 10.1111/j.1467-9469.2007.00583.x.

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

Etesami, J., Kiyavash, N., Zhang, K., and Singhal, K. Learning network of multivariate Hawkes processes: A time series approach. 2016.

Farajtabar, M., Du, N., Gomez Rodriguez, M., Valera, I., Zha, H., and Song, L. Shaping social activity by incentivizing users. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 27*, pp. 2474–2482. Curran Associates, Inc., 2014.

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

Hardiman, S., Bercot, N., and Bouchaud, J.-P. Critical reflexivity in financial markets: a hawkes process analysis. 2013.

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

Linderman, S. W. and Adams, R. P. Discovering latent network structure in point process data. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pp. II–1413–II–1421. JMLR.org, 2014. URL http://dl.acm.org/citation.cfm?id=3044805.3045050.

Liniger, T. J. *Multivariate Hawkes processes*. PhD thesis, Eidgenössische Technische Hochschule ETH Zürich, 2009.

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., and Tita, G. E. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108, 2011.

Ogata, Y. On lewis' simulation method for point processes. *IEEE Trans. Inf. Theor.*, 27(1):23–31, September 2006. ISSN 0018-9448. doi: 10.1109/TIT.1981.1056305. URL http://dx.doi.org/10.1109/TIT.1981.1056305.

Ozaki, T. Maximum likelihood estimation of Hawkes' self-exciting point processes. *Annals of the Institute of Statistical Mathematics*, 31(1):145–155, 1979.

Porter, M. D. and White, G. Self-exciting hurdle models for terrorist activity. *Ann. Appl. Stat.*, 6(1):106–124, 03 2012. doi: 10.1214/11-AOAS513. URL https://doi.org/10.1214/11-AOAS513.

Quinn, C. J., Coleman, T. P., Kiyavash, N., and Hatsopoulos, N. G. Estimating the directed information to infer causal relationships in ensemble neural spike train recordings. *Journal of computational neuroscience*, 30(1):17–44, 2011.

Rasmussen, J. G. Bayesian inference for hawkes processes. *Methodology and Computing in Applied Probability*, 15 (3):623–642, Sep 2013. ISSN 1573-7713. doi: 10.1007/s11009-011-9272-5.

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

Shelton, C. R., Qin, Z., and Shetty, C. Hawkes process inference with missing data. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Wu, W. and Hatsopoulos, N. G. Evidence against a single coordinate system representation in the motor cortex. *Experimental Brain Research*, 175:197–210, 2006.

Xu, H., Farajtabar, M., and Zha, H. Learning Granger causality for Hawkes processes. *International Conference on Machine Learning*, 48:1717–1726, 2016.

Xu, H., Luo, D., and Zha, H. Learning hawkes processes from short doubly-censored event sequences. *arXiv preprint arXiv:1702.07013*, 2017.

Yan, J., Zhang, C., Zha, H., Gong, M., Sun, C., Huang, J., Chu, S., and Yang, X. On machine learning towards predictive sales pipeline analytics, 2015.

Yang, S.-H. and Zha, H. Mixture of mutually exciting processes for viral diffusion. *International Conference on Machine Learning*, 28:1–9, 2013.

Yang, Y., Etesami, J., He, N., and Kiyavash, N. Online learning for multivariate Hawkes processes. *Neural Information Processing Systems*, 2017.

Zhou, K., Zha, H., and Song, L. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *AISTATS*, volume 31 of *JMLR Workshop and Conference Proceedings*, pp. 641–649. JMLR.org, 2013a.

Zhou, K., Zha, H., and Song, L. Learning triggering kernels for multi-dimensional Hawkes processes. In *International Conference on Machine Learning*, volume 28, pp. 1301–1309, 2013b.