Causal Discovery in Hawkes Processes by Minimum Description Length (preprint)

Amirkasra Jalaldoust,[1,2] **Kateřina Hlaváčková-Schindler,** [3,4] **Claudia Plant,** [3,5]
[1] Department of Computer Science, Columbia University, New York, USA
[2] Department of Mathematical Science, Sharif University of Technology, Tehran, Iran
[3] Faculty of Computer Science, University of Vienna, Vienna, Austria
[4] Institute of Computer Science, Czech Academy of Sciences, Prague, Czech Republic
[5] ds:UniVie, University of Vienna, Vienna, Austria
jalaldoust@cs.columbia.edu, katerina.schindlerova@univie.ac.at, claudia.plant@univie.ac.at

Abstract

Hawkes processes are a special class of temporal point processes which exhibit a natural notion of causality, as occurrence of events in the past may increase the probability of events in the future. Discovery of the underlying influence network among the dimensions of multi-dimensional temporal processes is of high importance in disciplines where a high-frequency data is to model, e.g. in financial data or in seismological data. This paper approaches the problem of learning Granger-causal network in multi-dimensional Hawkes processes. We formulate this problem as a model selection task in which we follow the minimum description length (MDL) principle. Moreover, we propose a general algorithm for MDL-based inference using a Monte-Carlo method and we use it for our causal discovery problem. We compare our algorithm with the state-of-the-art baseline methods on synthetic and real-world financial data. The synthetic experiments demonstrate superiority of our method in causal graph discovery compared to the baseline methods with respect to the size of the data. The results of experiments with the G-7 bonds price data are consistent with the experts' knowledge.

1 Introduction

In many applications, one needs to deal with multi-dimensional sequential data of irregular or asynchronous nature occurring in a continuous time. The examples can be preferences of users in in social networks, interactions of areas during earthquakes and aftershocks in geophysics (Veen and Schoenberg 2008) or high-frequency signals in financial data (Bacry, Mastromatteo, and Muzy 2015). These data can be seen as event sequences containing multiple event types and modeled by a multi-dimensional Hawkes process (MHP)(Hawkes 1971), also known as a self-exciting temporal point process. The main advantage of using Hawkes processes over e.g. Poisson processes is that they permit to model the influence of the past events on the current behavior of the process, due to their memory property. Eichler (Eichler, Dahlhaus, and Dueck 2017) introduced the graphical Granger causal model for MHP in the form of an autoregressive structure for the intensities of each marginal process.

Methods to discover Granger-causal graphs in MHP search for a solution of a variable selection problem, mostly in terms of an optimum of a corresponding objective function constructed for the MHP at hand (Zhou, Zha, and Song 2013a; Xu, Farajtabar, and Zha 2016; Zhou, Zha, and Song 2013b). These methods usually show good performance in scenarios with "long" horizon T, i.e. when T is greater by several orders of magnitude than the dimension of MHP p, however in the opposite case of "short" horizon, they often suffer from over-fitting. We define short event sequence when the time horizon is at most of order $100 \cdot p$ for a p-dimensional event sequence.

In our paper we approach causal discovery in MHP by the minimum description (MDL) principle which was as a model selection method introduced in (Rissanen 1998) and developed in (Grünwald and Roos 2019). Although MDL has been already applied to graphical Granger-causality in (Hlaváčková-Schindler and Plant 2020a) for Gaussian time series with a significant precision superiority over baseline methods, application of MDL to Hawkes processes is more challenging due to the nature of the processes. The contributions of our paper are summarized as follows.

- We present a general procedure for practical estimation of a MDL objective function based on Monte-Carlo (MC) methods for estimation of integrals.
- Using this procedure we construct an MDL objective function for inference of the Granger-causal graph for multi-dimensional Hawkes processes.
- We evaluate the performance of our causal discovery algorithm on both synthetic and real-world data, and compare our results with the baseline methods. Our method demonstrates a significant superiority in causal discovery for short event sequences.

The paper is organized as follows: The multi-dimensional Hawkes processes and the minimum description length principle are defined in Section 2. Section 3 presents our method to estimate a general MDL function. Causal discovery in Hawkes processes is proposed in Section 4. Related work is discussed in Section 5. Section 6 presents experiments and discussion and our conclusion is in Section 7. The Appendix contains a proof and experimental setup.

2 Preliminaries

2.1 Notation

Scalar variables are denoted by regular letters, multi-dimensional variables bold, random variables by capital letters (e.g. $\boldsymbol{X}$), the support of each random variable by the same calligraphic letter (e.g. $\mathcal{X}$) and any realization or point in the support is denoted by lower-case letter (e.g. $\boldsymbol{x}$). Matrices are denoted by Greek letters, and for matrices such as $\boldsymbol{\omega}$, $\boldsymbol{\omega}_i$ denotes the i-th row. All vectors are column vectors.

2.2 Temporal Point Processes

A temporal point process is a random process which is used to model occurrence of events in time. Each realization is a list of events $\{t_i\}$ with $t_i \in [0, T]$ and T is called horizon. The interval $[0, T]$ denotes the time window in which the process was observed. A temporal point process can be equivalently represented by a counting process U where $U(t)$ for $t \in [0, T]$ is the number of events happened prior to time t. Figure 1a visualizes a realization of a three-dimensional temporal point process. For a temporal point process one may define the conditional intensity function

$$\lambda(t) = \mathbb{E}[dU(t)|\mathcal{H}_t] = \lim_{\Delta t \to 0} \mathbb{E}[U(t+\Delta t) - U(t)|\mathcal{H}_t], \quad (1)$$

where $\mathcal{H}_t$ is called the filtration at time t consisting of all events prior to time t. A multi-dimensional temporal point process is a set of coupled temporal point processes and it can represented by a set of counting processes $\{U_i\}_{i=1}^{p}$, where $U_i(t)$ denotes the number of events in the i-th process prior to time t. Similarly, the conditional intensity function for the i-th dimension is

$$\lambda_i(t) = \mathbb{E}[dU_i(t)|\mathcal{H}_t]. \quad (2)$$

Granger-Causality in Temporal Point Processes In multi-dimensional time-series variable $\boldsymbol{x}_i$ Granger-causes variable $\boldsymbol{x}_j$ when the future of $\boldsymbol{x}_j$ is better predicted when taking into account the past of variable $\boldsymbol{x}_i$ (Granger 1969). In multi-dimensional temporal point processes the events in the j-th dimension Granger-cause the events in the i-th dimension if (Kim et al. 2011)

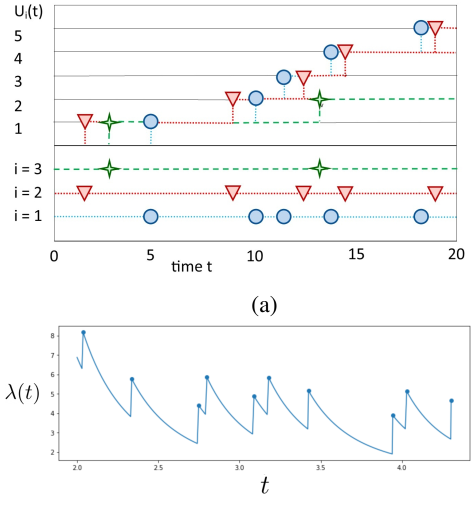
$$\lambda_i(t) = \mathbb{E}[dU_i(t)|\mathcal{H}_t] \neq \mathbb{E}[dU_i(t)|\mathcal{H}_t^{-j}], \quad (3)$$

where $\mathcal{H}_t^{-j}$ is the history prior to time t excluding events in the j-th dimension.

Hawkes Processes Introduced in (Hawkes 1971), a Hawkes process is a temporal point process with the intensity function following the equation

$$\lambda(t) = \mu + \int_{-\infty}^{t} \phi(t-\tau)dU(\tau), \quad (4)$$

where $\mu > 0$ is the exogenous baseline intensity, and $\phi : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ is the kernel function, see e.g. Fig 1b. Similarly, a multi-dimensional Hawkes process (MHP) is defined to be a multi-dimensional temporal point process with conditional intensity function at each dimension $i = 1, \dots, p$ following



(a)

(b)

Figure 1: (a) Realization of a 3-dimensional temporal point process, and its corresponding counting process; (b) The intensity function of a 1-dim Hawkes process with exponential decay kernel.

$$\lambda_i(t) = \boldsymbol{\mu}_i + \sum_{j=1}^{p} \int_{-\infty}^{t} \phi_{ij}(t-\tau)dU_j(\tau). \quad (5)$$

Granger-Causality in MHP Eichler et al. in (Eichler, Dahlhaus, and Dueck 2017) have shown that the Granger-causality structure of MHP is fully encoded in the corresponding kernel functions of the model. The result is stated in the following theorem.

Theorem 1 (Eichler et al., 2017)**.** *Let $U = \{U_i\}_{i=1}^{p}$ be a p-dimensional Hawkes process with conditional intensity functions defined as in Eq. 5. Then the events in the j-th dimension do not Granger-cause events in the i-th dimension if only if $\phi_{ij} \equiv 0$.*

2.3 Minimum Description Length

The minimum description length (MDL), introduced into statistical modeling by (Rissanen 1998; Barron, Rissanen, and Yu 1998) is a principle based on compression of information. The most common version of the method as a model selection principle makes use of two-part codes: the first part represents the information that one is trying to learn, such as the index of a model in a family of models (model selection) or parameter values (parameter estimation); the second part is an encoding of the data, given the model in the first part.

Statistical MDL learning is strongly connected to probability theory and statistics through the correspondence between codes and probability distributions. While Bayesian approach is often useful in constructing efficient MDL codes, the MDL framework also accommodates other codes that have no assumption about the data-generating process.

An example is the Shtarkov normalized maximum likelihood code (Shtarkov 1978), which we will use. Based on (Grünwald and Roos 2019), statistical models (families of probability distributions) are of the form $M = \{p(.|\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ parametrized by a parameter space $\boldsymbol{\Theta}$ (usually a subset of a Euclidean space). Further, we partition the parameter space $\boldsymbol{\Theta}$ into a family of disjoint restricted parameter subspaces $\{\boldsymbol{\Theta}_{\boldsymbol{\gamma}} : \boldsymbol{\gamma} \in \boldsymbol{\Gamma}\}$ so that we can define families of models $\{M_{\boldsymbol{\gamma}} : \boldsymbol{\gamma} \in \boldsymbol{\Gamma}\}$ where each $M_{\boldsymbol{\gamma}} = \{p(.|\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}_{\boldsymbol{\gamma}}\}$ is a statistical model used to model data $\boldsymbol{x} \in \mathcal{X}$, and each $p(.|\boldsymbol{\theta})$ represents a probability density function on $\mathcal{X}$.

Normalized Maximum Likelihood Distribution The normalized maximum likelihood (NML) distribution provides a general technique to apply the minimum description length (MDL) principle to statistical model selection. In its general form, the NML distribution and the MDL estimators depend on a function $v : \boldsymbol{\Theta} \rightarrow \mathbb{R}_0^+$ named luckiness function. The NML distribution for each model $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}$ is given by

$$p_{v|\boldsymbol{\gamma}}^{NML}(\boldsymbol{x}) = \frac{\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_{\boldsymbol{\gamma}}} p(\boldsymbol{x}|\boldsymbol{\theta})v(\boldsymbol{\theta})}{\int_{\mathcal{X}} \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_{\boldsymbol{\gamma}}} p(\boldsymbol{s}|\boldsymbol{\theta})v(\boldsymbol{\theta})d\boldsymbol{s}}. \tag{6}$$

(It is well-defined whenever the normalizing integral in (6) is finite.) The logarithm of this integral, called the model complexity is

$$COMP(M_{\boldsymbol{\gamma}}; v) = \log \int_{\mathcal{X}} \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_{\boldsymbol{\gamma}}} p(\boldsymbol{s}|\boldsymbol{\theta})v(\boldsymbol{\theta})d\boldsymbol{s}. \tag{7}$$

To simplify the notation, for each $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ let $R_v(\boldsymbol{\theta}; \boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{\theta})v(\boldsymbol{\theta})$, and let

$$r_v(\boldsymbol{\theta}; \boldsymbol{x}) = \log R_v(\boldsymbol{\theta}; \boldsymbol{x}) = \log p(\boldsymbol{x}|\boldsymbol{\theta}) + \log v(\boldsymbol{\theta}). \tag{8}$$

For any luckiness function v, we define the MDL estimator based on v for a specific model $M_{\boldsymbol{\gamma}} \subset M$ as

$$\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{x}) = \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_{\boldsymbol{\gamma}}} -r_v(\boldsymbol{\theta}; \boldsymbol{x}). \tag{9}$$

Suppose now a collection of models $M_{\boldsymbol{\gamma}}$ indexed by a finite $\boldsymbol{\Gamma}$ and specified luckiness functions $v_{\boldsymbol{\gamma}}$ on $\boldsymbol{\Theta}_{\boldsymbol{\gamma}}$, for each $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}$ and we pick a uniform distribution π on $\boldsymbol{\Gamma}$. If we base the model selection on NML, we pick over $\boldsymbol{\Gamma}$ the model minimizing function

$$\begin{aligned} L_v(\boldsymbol{\gamma}; \boldsymbol{x}) &= -\log \pi(\boldsymbol{\gamma}) - \log p_{v|\boldsymbol{\gamma}}^{NML}(\boldsymbol{x}) \\ &= -\log \pi(\boldsymbol{\gamma}) - r_v(\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{x}); \boldsymbol{x}) \\ &\quad + COMP(M_{\boldsymbol{\gamma}}; v). \end{aligned} \tag{10}$$

The MDL function incorporates a trade-off between the goodness-of-fit measured by $r_v(\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{x}); \boldsymbol{x})$ and the model complexity measured by $COMP(M_{\boldsymbol{\gamma}}; v)$. Finally, the model selection based on MDL picks

$$\hat{\boldsymbol{\gamma}}^{MDL} = \operatorname*{arg\,min}_{\boldsymbol{\gamma} \in \boldsymbol{\Gamma}} L_v(\boldsymbol{\gamma}; \boldsymbol{x}). \tag{11}$$

3 Our Method to Estimate the MDL function

We propose a new method to estimate the MDL objective function $L_v(\boldsymbol{\gamma}; \boldsymbol{x})$ for a given data $\boldsymbol{x} \in \mathcal{X}$, a fixed model $M_{\boldsymbol{\gamma}} \subset M$, and an appropriate luckiness function v. We further apply this general method to the problem of causal discovery in Hawkes processes.

According to Eq. 10, in order to estimate $L_v(\boldsymbol{\gamma}; \boldsymbol{x})$, the terms $\log \pi(\boldsymbol{\gamma})$, $r_v(\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{x}); \boldsymbol{x})$, and $COMP(M_{\boldsymbol{\gamma}}; v)$ should be computed. As $\pi(\boldsymbol{\gamma})$ is known to us a priori, the first term $\log \pi(\boldsymbol{\gamma})$ is computable. We propose how to compute the other two terms - goodness-of-fit $r_v(\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{x}); \boldsymbol{x})$ and model complexity $COMP(M_{\boldsymbol{\gamma}}; v)$.

3.1 Computing the Goodness-of-Fit

Finding the MDL estimator $\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{x})$ is the result of a minimization process over the restricted parameter space $\boldsymbol{\Theta}_{\boldsymbol{\gamma}}$. Convex optimization problems have a unique solution and there are efficient methods to find it. To this end, we require two conditions:

1. The restricted parameter space $\boldsymbol{\Theta}_{\boldsymbol{\gamma}}$ is a convex set.
2. The objective function measuring the negative goodness-of-fit

$$-r_v(\boldsymbol{\theta}; \boldsymbol{x}) = -\log(p(\boldsymbol{x}|\boldsymbol{\theta})) - \log(v(\boldsymbol{\theta})) \tag{12}$$

   is a convex function.

Condition 1 depends on an appropriate partitioning of the parameter space and can be fulfilled in many scenarios (e.g. when $\boldsymbol{\Theta}_{\boldsymbol{\gamma}}$ is a Euclidean space or a positive cone of it). Condition 2 can be fulfilled by an appropriate choice of luckiness function v; the objective function in Eq. (12) involves two terms:

i. The negative log-likelihood term $-\log p(\boldsymbol{x}|\boldsymbol{\theta})$.
ii. The negative log-luckiness term $-\log v(\boldsymbol{\theta})$.

The term (i) depends on the problem in hand, however, in many statistical scenarios it is convex in the model parameter $\boldsymbol{\theta}$ (e.g. in linear regression models and exponential families). The term (ii) is convex in $\boldsymbol{\theta}$ for any log-concave choice of v. Therefore, the summation of the two convex terms results in a convex objective function. In summary, to have a convex optimization problem for finding the MDL estimator $\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{x})$, we require:

1. $\boldsymbol{\Theta}$ to be partitioned into convex subsets $\{\boldsymbol{\Theta}_{\boldsymbol{\gamma}}\}_{\boldsymbol{\gamma} \in \boldsymbol{\Gamma}}$.
2. Log-likelihood to be a concave function.
3. Luckiness v to be selected as a log-concave function.

3.2 Estimating Model Complexity

In this subsection we propose estimation of the model complexity by using Monte Carlo (MC) simulation for estimation of integrals. According to definition of R_v and Eq. 9 model complexity can be rewritten as

$$COMP(M_{\boldsymbol{\gamma}}; v) = \log \int_{\mathcal{X}} R_v(\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{s}); \boldsymbol{s})d\boldsymbol{s}. \tag{13}$$

As discussed in subsection 3.1, we can efficiently compute a unique MDL estimator $\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{s})$ for any data $\boldsymbol{s} \in \mathcal{X}$. For any parameter $\boldsymbol{z} \in \boldsymbol{\Theta}$, we have

$$\begin{aligned} COMP(M_{\boldsymbol{\gamma}}; v) &= \log \int_{\mathcal{X}} \frac{R_v(\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{s}); \boldsymbol{s})}{p(\boldsymbol{s}|\boldsymbol{\theta} = \boldsymbol{z})} p(\boldsymbol{s}|\boldsymbol{\theta} = \boldsymbol{z}) d\boldsymbol{s} \\ &= \log \mathbb{E}_{\boldsymbol{X} \sim p(.|\boldsymbol{\theta}=z)} [\frac{R_v(\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{X}); \boldsymbol{X})}{p(\boldsymbol{X}|\boldsymbol{\theta} = \boldsymbol{z})}] \\ &= \log \mathbb{E}_{\boldsymbol{X} \sim p} [\frac{R_v(\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{X}); \boldsymbol{X})}{p(\boldsymbol{X}|\boldsymbol{\theta})} |\boldsymbol{\theta} = \boldsymbol{z}]. \end{aligned} \tag{14}$$

Note that in order to rewrite the integral as above, the term $p(\boldsymbol{s}|\boldsymbol{\theta} = \boldsymbol{z})$ should be positive for all data $\boldsymbol{s} \in \mathcal{X}$. The term inside the expectation is a function of the parameter $\boldsymbol{z} \in \boldsymbol{\Theta}$ and the random variable $\boldsymbol{X}$, so for simplicity we define

$$\begin{aligned} Q_{v|\boldsymbol{\gamma}}(\boldsymbol{s}, \boldsymbol{z}) &= \frac{R_v(\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{s}); \boldsymbol{s})}{p(\boldsymbol{s}|\boldsymbol{z})} \\ &= \frac{p(\boldsymbol{s}|\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{s})) v(\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{s}))}{p(\boldsymbol{s}|\boldsymbol{z})} \end{aligned} \tag{15}$$

so that we have

$$COMP(M_{\boldsymbol{\gamma}}; v) = \log \mathbb{E}_{\boldsymbol{X} \sim p} [Q_{v|\boldsymbol{\gamma}}(\boldsymbol{X}, \boldsymbol{\theta}) | \boldsymbol{\theta} = \boldsymbol{z}]. \tag{16}$$

We can also randomize $\boldsymbol{z}$ and then take the expected value. As the general MDL method does not impose any distribution for $\boldsymbol{\theta}$, we can assume any arbitrary full support distribution on $\boldsymbol{\Theta}$ and take random samples of $\boldsymbol{\theta}$ from it. Therefore,

$$\begin{aligned} COMP(M_{\boldsymbol{\gamma}}; v) &= \log \mathbb{E}_{\boldsymbol{\theta} \sim p} [\mathbb{E}_{\boldsymbol{X} \sim p} [Q_{v|\boldsymbol{\gamma}}(\boldsymbol{X}, \boldsymbol{\theta}) | \boldsymbol{\theta} = \boldsymbol{z}]] \\ &= \log \mathbb{E}_{\boldsymbol{X}, \boldsymbol{\theta} \sim p} [Q_{v|\boldsymbol{\gamma}}(\boldsymbol{X}, \boldsymbol{\theta})]. \end{aligned} \tag{17}$$

We estimate the last expectation by taking multiple joint samples of $\boldsymbol{X}, \boldsymbol{\theta}$ (Line 1 and Line 3 of Algorithm 1), and by computing $Q_{v|\boldsymbol{\gamma}}$ at these points (Line 5 of Algorithm 1) we report the average value as an unbiased estimation of $COMP(M_{\boldsymbol{\gamma}}; v)$. One can also compute confidence intervals for our estimation using sample variance of the random draws of $Q_{v|\boldsymbol{\gamma}}(\boldsymbol{X}, \boldsymbol{\theta})$. Our steps to estimate model complexity are summarized in Algorithm 1. First, according to distribution $p(\boldsymbol{\theta})$, we draw random parameters $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_N$ where N is the number of MC simulations. Next, for each parameter $\boldsymbol{z}_i$, we simulate the data $\boldsymbol{s}_i$ according to the distribution $p(\boldsymbol{X}|\boldsymbol{\theta} = \boldsymbol{z}_i)$. Finally, computing the MDL estimator $\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{s}_i)$ enables us to compute

$$Q_i = \frac{p(\boldsymbol{s}_i|\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{s}_i)) v(\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{s}_i))}{p(\boldsymbol{s}_i|\boldsymbol{z}_i)}, \tag{18}$$

which is an i.i.d. sample of $Q_{v|\boldsymbol{\gamma}}(\boldsymbol{X}, \boldsymbol{\theta})$. According to Eq. 17, the mean value of random samples $Q_1, Q_2, \ldots, Q_N$ is an unbiased estimator for model complexity. Furthermore, the accuracy of our estimation increases by increasing number of MC simulations N; In fact, by the central limit theorem, the variance of our estimator for model complexity linearly decreases with increasing N.

Algorithm 1: Estimate $COMP(M_{\boldsymbol{\gamma}}; v)$

Input: Model $\boldsymbol{\gamma}$.
Given: Luckiness v, number of MC simulations N.
Output: Estimation of model complexity $\hat{C}$.

```
1: Draw i.i.d. parameters $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_N$ w.r.t. $p(\boldsymbol{\theta})$
2: for $1 \leq i \leq N$ do
3:     Draw data $\boldsymbol{s}_i$ w.r.t. $p(\boldsymbol{X}|\boldsymbol{\theta} = \boldsymbol{z}_i)$
4:     $\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{s}_i) \leftarrow \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_{\boldsymbol{\gamma}}} -\log p(\boldsymbol{s}_i|\boldsymbol{\theta}) - \log v(\boldsymbol{\theta})$
5:     $Q_i \leftarrow p(\boldsymbol{s}_i|\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{s}_i)) v(\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{s}_i)) / p(\boldsymbol{s}_i|\boldsymbol{z}_i)$
6: end for
7: $\hat{C} \leftarrow \log(\text{mean}(\{Q_i\}_{i=1}^{p}))$
8: return $\hat{C}$
```

4 Causal Discovery in Hawkes Processes

In this section we propose an MDL-based algorithm for causal discovery in Hawkes processes. We restrict ourselves to MHP with exponential decay kernel functions (exp-MHP). This is a class of MHP with kernels defined as

$$\phi_{ij}(t) = \alpha_{ij} \exp(-\beta_{ij} t), \tag{19}$$

where $\boldsymbol{\alpha}$ is a $p \times p$ matrix called the influence coefficient matrix with non-negative entries, and $\boldsymbol{\beta}$ is a known constant $p \times p$ matrix called decay matrix with positive entries. Hence, the intensity function for the i-th dimension is

$$\lambda_i(t) = \mu_i + \sum_{j=1}^{p} \int_{-\infty}^{t} \alpha_{ij} \exp(-\beta_{ij}(t-s)) dU_j(s), \tag{20}$$

where $\boldsymbol{\mu}$ is a p-dimensional vector of non-negative baseline intensities. Exp-MHP with known decay matrix $\boldsymbol{\beta}$ can be characterized by the pair $(\boldsymbol{\mu}, \boldsymbol{\alpha})$, and we denote this pair as the parameters.

We focus on exp-MHP in this paper, however, the methodology can be utilized in other parametric settings as well; e.g. in MHP with power-law kernels (Bacry and Muzy 2016) and in MHP with kernels defined by a set of basis functions. As mentioned above, causal discovery task is to identify the causal influence network among variables. In case of Granger-causality in Hawkes processes, this network is a directed graph where each node corresponds to a Hawkes process. Any directed graph of p nodes can be expressed by its adjacency matrix $\boldsymbol{\eta} \in \{0, 1\}^{p \times p}$ where $\eta_{ij} = 1$ iff there exists an edge from j-th node to the i-th node. For the causal graph of a MHP we have $\eta_{ij} = 1$ iff events in the j-th dimension Granger-cause the events in the i-th dimension. Applying Theorem 1 for exp-MHP it holds $\eta_{ij} = 0$ iff $\alpha_{ij} = 0$, and therefore, causal discovery in exp-MHP corresponds to identifying the sparsity pattern of the influence coefficient matrix $\boldsymbol{\alpha}$.

4.1 Parameter Learning in exp-MHP

Before we construct our MDL function for causal discovery, we need to define parameters $\boldsymbol{\theta}$ and its space as well as to know the log-likelihood function w.r.t. $\boldsymbol{\theta}$. The parameters of the exp-MHP model are the influence coefficient matrix $\boldsymbol{\alpha}$ and the baseline vector $\boldsymbol{\mu}$. In the i-th dimension it is

$$\boldsymbol{\theta}_i = (\mu_i, \boldsymbol{\alpha}_i^\top)^\top \in (\mathbb{R}_0^+)^{p+1}, \tag{21}$$

and we define the parameter vector of exp-MHP as

$$\boldsymbol{\theta} = [\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top, \ldots, \boldsymbol{\theta}_p^\top]^\top \in (\mathbb{R}_0^+)^{p+p^2}. \tag{22}$$

Any realization of a MHP can be seen as the collection of event sequences $\boldsymbol{x} = \{\boldsymbol{x}_i\}_{i=1}^p$ where each event sequence $\boldsymbol{x}_i = (t_1^i, t_2^i, \ldots, t_{n_i}^i)^\top$ denotes the times when the events occurred in the i-th dimension. The multi-dimensional conditional intensity can be computed in interval $[0, T]$ based on realization $\boldsymbol{x}$ and the parameter vector $\boldsymbol{\theta}$ (Daley and Vere-Jones 2003). Hence, the negative log-likelihood for $\boldsymbol{\theta}$ is

$$\begin{aligned} -\log p(\boldsymbol{x}|\boldsymbol{\theta}) &= \sum_{i=1}^{p} -\log p(\boldsymbol{x}|\boldsymbol{\theta}_i) \qquad (23) \\ &= \sum_{i=1}^{p} \left(\int_0^T \lambda_i(s) ds - \sum_{j=0}^{n_i} \log \lambda_i(t_j^i) \right) \end{aligned}$$

and in exp-MHP (Ozaki 1979)

$$\begin{aligned} -\log p(\boldsymbol{x}|\boldsymbol{\theta}_i) &= \mu_i T + \sum_{j=1}^{p} \frac{\alpha_{ij}}{\beta_{ij}} \sum_{k=1}^{n_j} [1 - \exp(-\beta_{ij}(T - t_k^j))] \\ &- \sum_{l=1}^{n_i} \log[\mu_i + \sum_{j=1}^{p} \alpha_{ij} \sum_{k: t_k^j < t_l^i} \exp(-\beta_{ij}(t_l^i - t_k^j))]. \end{aligned} \tag{24}$$

The search space for finding $\boldsymbol{\theta}_i$ which minimizes Eq. 24 is the positive cone $(\mathbb{R}_0^+)^{p+1}$, and the objective function is convex in $\boldsymbol{\theta}_i$ (Ogata 1981). Therefore a unique solution exists and efficient algorithms are available, e.g. (Bacry, Mastromatteo, and Muzy 2015).

4.2 Causal Discovery As Model Selection

Let $\boldsymbol{\Gamma}$ be the set of all binary $p \times p$ matrices. For each binary matrix $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}$ the model $M_{\boldsymbol{\gamma}}$ is the set of all p-dimensional Hawkes process models with their causal graph having adjacency matrix $\boldsymbol{\gamma}$. According to the definition of Granger-causality in exp-MHP, the restricted parameter space $\boldsymbol{\Theta}_{\boldsymbol{\gamma}}$ contains parameter vectors representing $\boldsymbol{\mu}, \boldsymbol{\alpha}$ such that $\alpha_{ij} = 0$ iff $\gamma_{ij} = 0$. The baseline vector $\boldsymbol{\mu}$ has no influence on causal discovery and our proposed model selection formulation.

In this way the parameter space $\boldsymbol{\Theta}$ is partitioned into disjoint subsets $\{\boldsymbol{\Theta}_{\boldsymbol{\gamma}}\}_{\boldsymbol{\gamma} \in \boldsymbol{\Gamma}}$ as $\boldsymbol{\Theta}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\Theta}_{\boldsymbol{\gamma}'}$ do not intersect for any different $\boldsymbol{\gamma}, \boldsymbol{\gamma}' \in \boldsymbol{\Gamma}$, because for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ the sparsity pattern of the influence matrix $\boldsymbol{\alpha}$ uniquely determines the causal graph $\boldsymbol{\gamma}$. Thus finding the true model $M_{\boldsymbol{\gamma}^*} \subset M$ is equivalent to finding the true causal graph with adjacency matrix $\boldsymbol{\gamma}^*$.

MDL Objective Function for exp-MHP We show that the MDL formulation for causal discovery discussed in subsection 4.2 satisfies the three conditions listed in the end of subsection 3.1. For any $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}$ containing $w > 0$ non-zero entries, matrix $\boldsymbol{\alpha}$ has w non-zero entries to be estimated, as well as p dimensions of the baseline vector $\boldsymbol{\mu}$ has to be estimated. Hence, a total of $p + w$ non-negative parameters are to be estimated and therefore, the restricted parameter space $\boldsymbol{\Theta}_{\boldsymbol{\gamma}}$ is the positive cone $(\mathbb{R}_0^+)^{p+w}$ which is a convex set (Condition 1). Further, according to Eq. 24, negative log-likelihood is a convex function (Condition 2). Thus, by choosing luckiness function v to be log-concave (Condition 3) we get a convex optimization problem for finding the MDL estimator

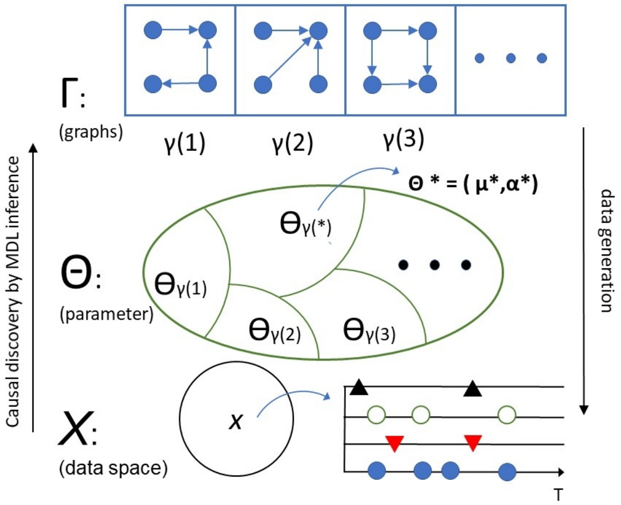


Figure 2: The relationship between the true causal graph $\boldsymbol{\gamma}^{(*)}$, the true parameter $\boldsymbol{\theta}^*$, and the data $\boldsymbol{x}$. The data generation mechanism (e.g., nature, right arrow) first decides the true causal graph $\boldsymbol{\gamma}^{(*)}$, then decides the parameter $\boldsymbol{\theta}^*$ (the excitation coefficients $\boldsymbol{\alpha}^*$ and the exogenous intensities $\boldsymbol{\mu}$) in the restricted parameter space $\boldsymbol{\Theta}_{\boldsymbol{\gamma}^*}$, and finally draws a random realization (i.e., data) $\boldsymbol{x}$ from this fully specified Hawkes process. Conversely, in the causal discovery task (left arrow) we are given the data $\boldsymbol{x}$ and we use this hierarchy to infer the true causal graph.

$$\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}} = \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_{\boldsymbol{\gamma}}} -\log p(\boldsymbol{x}|\boldsymbol{\theta}) - \log v(\boldsymbol{\theta}).$$

For example, an appropriate uninformative choices for the luckiness function v can be $v \equiv 1$, which we used in our experiments. Another example of appropriate choice which penalizes large baseline intensities and influence coefficients can be

$$v(\boldsymbol{\theta}) = \prod_{i=1}^{p} \exp(-\mu_i) \prod_{i,j=1}^{p} \exp(-\alpha_{ij}). \tag{25}$$

Efficient computation of the MDL estimator $\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}$ enables us to estimate model complexity $COMP(M_{\boldsymbol{\gamma}}; v)$ according to Algorithm 1, and thus we may estimate the MDL objective $L_v(\boldsymbol{\gamma}; \boldsymbol{x})$ as in Eq. 10 for any specific causal graph $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}$.

4.3 Causal Discovery Algorithm

According to MDL criterion for model selection, to find the best model for describing the data $\boldsymbol{x}$ and equivalently learning the underlying causal graph, we should find $\hat{\boldsymbol{\gamma}}^{MDL}$ from Eq. 11. This can by found by the exhaustive search over $\boldsymbol{\Gamma}$, or by heuristics based on genetic algorithms, as e.g. in (Hlaváčková-Schindler and Plant 2020b). Figure 2 illustrates the connection of the mechanisms of causal discovery by MDL and of data generation.

By imposing mild conditions on the luckiness function v and distribution π of the models indexed in set $\boldsymbol{\Gamma}$ as in Theorem 2, we reduce the complexity of the above search problem for finding $\hat{\boldsymbol{\gamma}}^{MDL}$.

Theorem 2. *If in MDL-based model selection for exp-MHP,*

$$\pi(\boldsymbol{\gamma}) = \prod_{i=1}^{p} \pi_i(\boldsymbol{\gamma}_i), \quad v(\boldsymbol{\theta}) = \prod_{i=1}^{p} v_i(\boldsymbol{\theta}_i), \tag{26}$$

then the MDL function can be rewritten as p independent terms

$$L_v(\boldsymbol{\gamma}; \boldsymbol{x}) = \sum_{i=1}^{p} L_v^i(\boldsymbol{\gamma}_i; \boldsymbol{x}), \tag{27}$$

such that each $L_v^i(\boldsymbol{\gamma}_i, \boldsymbol{x})$ can be computed by Algorithm 2.

According to Theorem 2, if the distributions over the causal graphs and the luckiness function are independent for each dimension of the Hawkes process, then we may optimize the MDL function for each of the dimensions separately, which enables us to perform parallel computation. The proof follows from the functional form of the MDL function, and it is provided in the Appendix. Algorithm 2 computes estimate of $L_v^i(\gamma_i; \boldsymbol{x})$ based on $COMP$ from Algorithm 1 and on input values $\boldsymbol{x}$, $\boldsymbol{\gamma}$, π_i, v_i and number of MC simulations.

By solving the independent optimization problems

$$\hat{\boldsymbol{\gamma}}_i^{MDL} = \underset{\boldsymbol{\gamma}_i \in \{0,1\}^p}{\arg\min}\ L_v^i(\boldsymbol{\gamma}_i; \boldsymbol{x}), \tag{28}$$

we get

$$\hat{\boldsymbol{\gamma}}^{MDL} = [\hat{\boldsymbol{\gamma}}_1^{MDL} | \hat{\boldsymbol{\gamma}}_2^{MDL} | \dots | \hat{\boldsymbol{\gamma}}_p^{MDL}]^T \tag{29}$$

as the solution to the optimization problem in Eq. 11.

Our causal discovery method, summarised in Algorithm 3, executes the following steps: First, we capture p as the number of coordinates of the data. Next, for every index $1 \leq i \leq p$, we search exhaustively over the model space $\boldsymbol{\Gamma}_i$ to find $\hat{\boldsymbol{\gamma}}_i$ which minimizes our estimation of the MDL objective function $L_v^i(\boldsymbol{\gamma}_i; \boldsymbol{x})$ computed by Algorithm 2. By default, the model space $\boldsymbol{\Gamma}_i$ is the set of all p-dimensional binary vectors. We call Algorithm 3 "MDLH" for MDL-based causal discovery in Hawkes processes.

Computational Complexity Our causal discovery method concluded in Algorithm 3 comprises of p optimizations, each solved by 2^p calls of estimating MDL function by Algorithm 2. The latter consists of i) parameter learning for the data $\boldsymbol{x}$ and ii) a call of estimating model complexity by Algorithm 1 during which we run N MC simulation by generating N random data and learning the parameter for each of them. Therefore, each call of estimating MDL function by Algorithm 2 performs a total of $N+1$ parameter learning procedures. The computational complexity of these parameter learning procedures depends on the number of parameters to be learned (for $\boldsymbol{\gamma}_i$ with w non-zero entries, the number of parameters is $w+1$) and the total number of observed events, however, for simplicity reasons we can assume they have the same computational complexity. In conclusion, to search over all causal graphs, we must perform $(N+1) \cdot p \cdot 2^p$ parameter learning operations under our causal discovery method.

Algorithm 2: Estimate $L_v^i(\boldsymbol{\gamma}_i; \boldsymbol{x})$

Input: The data $\boldsymbol{x}$, model $\boldsymbol{\gamma}$, index i.
Given: Distribution π_i, luckiness v_i, number of MC simulations N.
Output: Estimation of the MDL function $\hat{L}_i$

```
$\hat{\boldsymbol{\theta}}_i \leftarrow \arg\min_{\boldsymbol{\theta}_i \in \boldsymbol{\Theta}_{\boldsymbol{\gamma}_i}} -\log p(\boldsymbol{x}|\boldsymbol{\theta}_i) - \log v_i(\boldsymbol{\theta}_i)$
$\hat{C}_i \leftarrow$ estimate $COMP(M_{\boldsymbol{\gamma}_i}; v)$ by Alg. 1
$\hat{L}_i \leftarrow -\log \pi_i(\boldsymbol{\gamma}_i) - \log p(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_i) - \log v_i(\hat{\boldsymbol{\theta}}_i) + \hat{C}_i$
**return** $\hat{L}_i$
```

Algorithm 3: Causal Discovery by MDL

Input: The data $\boldsymbol{x}$
Given: Distribution π, luckiness v, number of MC simulations N, model spaces $\{\boldsymbol{\Gamma}_i\}_{i=1}^p$ - default: $\boldsymbol{\Gamma}_i = \{0,1\}^p$
Output: Inferred causal graph $\hat{\boldsymbol{\gamma}}$

```
$p \leftarrow$ dimension($\boldsymbol{x}$)
**for** $i \in \{1, 2, \dots, p\}$ **do**
    $\hat{\boldsymbol{\gamma}}_i \leftarrow \arg\min_{\boldsymbol{\gamma}_i \in \boldsymbol{\Gamma}_i}$ estimate $L_v^i(\boldsymbol{\gamma}_i; \boldsymbol{x})$ by Alg. 2
**end for**
$\hat{\boldsymbol{\gamma}} \leftarrow [\hat{\boldsymbol{\gamma}}_1 | \hat{\boldsymbol{\gamma}}_2 | \dots | \hat{\boldsymbol{\gamma}}_p]$        //adjacency matrix
**return** $\hat{\boldsymbol{\gamma}}$
```

Sparse Causal Graphs In scenarios when the expert knowledge suggests a small upper-bound on the number of causes for each variable, we may reduce the computational complexity of our algorithm significantly. If the degree of each of the nodes in the causal graph is bounded by a constant $m \ll p$, we can achieve polynomial computational complexity. I.e., the model space $\boldsymbol{\Gamma}_i$ for the i-th dimension would contain only p-dimensional binary vectors with at most m non-zero entries. Thus, we have

$$|\boldsymbol{\Gamma}_i| = \sum_{k=0}^{m} \binom{p}{k} < m \binom{p}{m} = \mathcal{O}(p^m). \tag{30}$$

Therefore, we would have to perform $\mathcal{O}((N+1) \cdot p^{m+1})$ parameter learning procedures, which is polynomial in p for constant $m \ll p$. We stress here however that the contribution of our work is not in computational complexity but in the methodology and precision of causal discovery.

Amortization In situations when the problem specification (dimension p, decay matrix $\boldsymbol{\beta}$, horizon T, model index distribution π, luckiness v, number of MC simulations N) is fixed, we can amortize most of the computational cost of our causal discovery method as follows: We compute and store the model complexities as in Algorithm 1 for all possible inputs which costs $N \cdot p \cdot 2^p$ times of performing parameter learning procedure in the default scenario ($N \cdot p^{m+1}$ times in sparse graphs). Next, for each query data we may run the causal discovery method in a total of $p \cdot 2^p$ times of performing parameter learning procedures in Line 1 of Algorithm 2 (p^{m+1} times in sparse graphs), while we use the pre-stored values for model complexities.

Parallelization p optimizations in Line 3 of Algorithm 3 can be performed in parallel, as they are independent. Each optimization consists of 2^p calls of Algorithm 2 in default scenario (p^m calls in sparse graphs) which comprises $N + 1$ independent parameter learning procedures. Therefore, all $(N + 1) \cdot p \cdot 2^p$ steps ($(N + 1) \cdot p^{m+1}$ in sparse graphs) are independent and can be performed in parallel.

5 Related Work

We address here the related work on discovery of Granger-causal networks in MHP and on compression schemes related to causal discovery in general. The method ADM4 in (Zhou, Zha, and Song 2013a) performs variable selection by using lasso and nuclear norm regularization simultaneously on the parameters to cluster variables as well as to obtain sparsity of the network. To detect a Granger-causal graph in MHP, (Xu, Farajtabar, and Zha 2016) applied an EM algorithm based on a penalized likelihood objective leading to temporal and group sparsity. The method NPHC (Achab et al. 2017) takes a non-parametric approach in learning the norm of the kernel functions to address the causal discovery problem. A moment-matching method is used fitting the second-order and third-order integrated cumulants of the process. NPHC is the most recent development in the literature and outperforms the state-of-the-art methods in many aspects. These methods show good performance in scenarios with “long” horizon T, however, in the opposite case of “short” horizon, they often suffer from over-fitting.

The MDL principle was applied to bivariate causal inference problem in (Marx and Vreeken 2017, 2018; Budhathoki and Vreeken 2018). The considered causal inference is however not the Granger one, as they discuss i.i.d. setting in contrast to sequentially ordered data. More recently, (Mian, Marx, and Vreeken 2021) extended these works to causal inference into multi-dimensional case using a greedy algorithm based on forward and backward search. The conditions on the convexity of the constructed criterion functions is hard to impose and the proposed greedy algorithm does not guarantee to find a unique solution. The MDL principle in the graphical Granger-causal models for Gaussian time-series was applied in (Hlaváčková-Schindler and Plant 2020a). A similar principle, connecting the idea of compression of information with Bayesian inference, called the minimum message length, has been applied to graphical Granger models for Poisson time series in (Hlaváčková-Schindler and Plant 2020c) and for time series from exponential family in (Hlaváčková-Schindler and Plant 2020b). To our best knowledge, no work on causal discovery in MHP based on MDL has been published yet.

6 Experiments and Discussion

As mentioned before, we call the instance of our general method applied to MHP by MDLH. We evaluate its performance on both synthetic data with known ground truth, and real-world data. MDLH is implemented in python, and the experiments are performed on a one Core Intel Xeon machine with 16 GB RAM. Our implementation and all experimental data are available at https://dm.cs.univie.ac.at/research/downloads/ and https://github.com/Amirkasraj/HawkesMDL

6.1 Synthetic Data

Baseline Methods We consider i) NPHC (Achab et al. 2017) as a state-of-the-art method for causal discovery in MHP since it outperformed many of previous rival methods; ii) ADM4 (Zhou, Zha, and Song 2013b) has the same assumptions as ours and does model selection by a mix of lasso and nuclear regularization; iii) Information-theory based criteria for model selection including AIC, BIC, and HQ (Chen 2016); We refer the method with the best score out of them as IC; iv) Regularized maximum likelihood (ML) and v) regularized least-squares (LS) with lasso, ridge, and elastic net regularizations. Implementation of all these methods is provided in Python by using Tick package (Bacry et al. 2017). For each of the comparison methods we report its best score among all possible regularizations and over a grid of hyperparameters. Each estimation method reports a matrix of kernel norms and we set threshold 0.01 to determine the causal graph from the real valued kernel norms.

Data Generation Process As in Figure 2, we generate synthetic data. In all experiments, we allow for self-excitation in all dimensions, i.e., the diagonal entries of all adjacency matrices are non-zero. In default scenario (low dimensions), each non-diagonal entry of the adjacency matrix of the causal graph is randomly drawn from Bernoulli(r). In sparse graphs (high dimensions), for each dimension, we draw the number of causes (other dimensions which affect this dimension) from $unif(\{0, 1, \ldots, m\})$, and then we uniformly chose a candidate from all possible subsets of other variables with that size. For $p = 7$ we go with default scenario with $r = 0.3$, and for $p = 20$ we go with sparse graphs scenario with $m = 1$. Next, each entry of matrix $\boldsymbol{\alpha}$ is drawn from unif$([0.1, 0.2])$ and each entry of vector $\boldsymbol{\mu}$ is drawn from unif$([0.5, 1.0])$. In our experiments, the decay matrix $\boldsymbol{\beta}$ is the matrix of ones.

Evaluation and Results We use F1 measure to evaluate the methods. F1 is suitable to evaluate the accuracy of estimated directed graphs represented by adjacency matrices, since gives the same importance to causal and non-causal connections. Due to the limitation of our computation resources, the number of MC simulations N was 1000 in our algorithm in all cases; We only consider luckiness function $v \equiv 1$ and uniform distribution π. In each experimental setting we randomly generate 100 data generation processes and draw one random sample from each of them. As Table 1 demonstrates, our method significantly outperforms all baseline methods on short horizon. The total failure of

IC methods (an empty graph in all cases) is presumably due their weak performance for short data.

Table 1: Performance of MDL and baselines in F1.

p	7			20		
T	200	400	700	500	1300	2000
MDLH	**77.4**	**84.7**	**89.3**	**79.4**	**82.8**	**84.4**
ADM4	68.4	72.6	78.5	26.8	29.9	31.5
NPHC	49.3	58.8	61.3	27.3	34.5	40.0
ML	68.7	74.6	80.4	25.8	28.2	29.4
LS	68.3	74.4	76.9	26.4	29.8	31.3
IC	NA	NA	NA	NA	NA	NA
Random	30.0	30.0	30.0	7.5	7.5	7.5

6.2 Real-World Data

G-7 Bonds We use daily return volatility of sovereign bonds of 7 large and developed economies called G-7 including USA, Germany, France, Japan, UK, Canada, Italy from 2003-2014 as in (Demirer et al. 2018). The goal is to discover the underlying influence network among the sovereign bonds.

Shock Identification As the data is a time-series and not a point process, to identify shocks (events in point process) in the daily return volatility, we roll a one year window over the data, and in each dimension if the latest value of the window is among the top 20 percent in the rolling window we register an event in that dimension for that day. The number of events registered in each dimension is around 500, and applying the knowledge gained in the synthetic experiments with a similar scenario, we assumed this data is an instance of our synthetic data with $T = 400$.

Results Only MDLH and ML were tested on the G-7 data, since ML outperfomed other baseline methods for $p = 7$ in the experiments with synthetic data. MDLH discovers the graph depicted in Figure 3. (The self-loops of all nodes were omitted). The structure is plausible when considering the network discovered by (Demirer et al. 2018). Also, the graph corresponds to the expert knowledge from the domain, e.g., that Japanese bond neither influences nor gets influenced by other G-7 countries. Moreover, the influence of French bonds on the US, UK and other big economies can be affected by the fact that France accused U.S of 'economic war' in 2003 which spread in most of the world media (CNN-International.com 2003). On the other hand, ML discovers bi-directed edge between US and Japan which contradicts to the expert knowledge given in (Demirer et al. 2018). To conclude, the MDLH gives a more plausible graph than ML.

7 Conclusion

We presented a general procedure for practical estimation of an MDL-based objective function using Monte-Carlo integral estimation. In this procedure we constructed an MDL objective function for inference of the Granger-causal graph for multi-dimensional Hawkes processes. We demonstrated significant superiority of our method in causal discovery for short event sequences in synthetic experiments. In a real experiment with G-7 bonds, our method gives a more plausible causal graph than the baseline method.

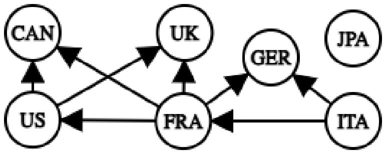


Figure 3: G-7 causal graph inferred by MDLH

Acknowledgements

This work was funded in part by the Austrian Science Fund (FWF) I5113-N, in part within the Austrian Climate and Research Programme ACRP project MEDEA KR19AC0K17614, and in part by the Czech Science Foundation project GA19-16066S.

Appendix

7.1 Proof of Theorem 2

Proof. Define

$$\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}_i}(\boldsymbol{x}) = \operatorname*{arg\,min}_{\boldsymbol{\theta}_i \in \boldsymbol{\Theta}_{\boldsymbol{\gamma}_i}} -\log p(\boldsymbol{x}|\boldsymbol{\theta}_i) - \log v_i(\boldsymbol{\theta}_i),$$

where $\boldsymbol{\Theta}_{\boldsymbol{\gamma}_i}$ is the space of all possible values for $\boldsymbol{\theta}_i$. We have

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{x}) &= \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_{\boldsymbol{\gamma}}} -\log p(\boldsymbol{x}|\boldsymbol{\theta}) - \log v(\boldsymbol{\theta}) \\ &= [\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}_1}(\boldsymbol{x})^T, \hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}_2}(\boldsymbol{x})^T, \dots, \hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}_p}(\boldsymbol{x})^T]^T.\end{aligned}$$

Further we define

$$COMP(M_{\boldsymbol{\gamma}_i}; v) = \log \int_{\mathcal{X}} p(\boldsymbol{s}|\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}_i}(\boldsymbol{s}))v_i(\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}_i}(\boldsymbol{s}))d\boldsymbol{s}.$$

We have

$$\begin{aligned}COMP(M_{\boldsymbol{\gamma}}; v) &= \log \int_{\mathcal{X}} p(\boldsymbol{s}|\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{s}))v(\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{s}))d\boldsymbol{s} \\ &= \log \int_{\mathcal{X}} [\prod_{i=1}^{p} p(\boldsymbol{s}|\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}_i}(\boldsymbol{s}))v_i(\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}_i}(\boldsymbol{s}))]d\boldsymbol{s} \\ &= \log \prod_{i=1}^{p} \int_{\mathcal{X}} p(\boldsymbol{s}|\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}_i}(\boldsymbol{s}))v_i(\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}_i}(\boldsymbol{s}))d\boldsymbol{s} \\ &= \sum_{i=1}^{p} \log \int_{\mathcal{X}} p(\boldsymbol{s}|\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}_i}(\boldsymbol{s}))v_i(\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}_i}(\boldsymbol{s}))d\boldsymbol{s} \\ &= \sum_{i=1}^{p} COMP(M_{\boldsymbol{\gamma}_i}; v).\end{aligned}$$

As in Eq. 22 in the paper, the negative log-likelihood can be written in independent terms for each dimension. Therefore, for each dimension $1 \leq i \leq p$ we can write

$$L_v^i(\boldsymbol{\gamma}_i; \boldsymbol{x}) := -\log \pi_i(\boldsymbol{\gamma}_i) - \log p(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}_i}(\boldsymbol{x})) \qquad (31)$$
$$- \log v_i(\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}_i}(\boldsymbol{x})) + COMP(M_{\boldsymbol{\gamma}_i}; v).$$

Hence, we have

$$\begin{aligned}\sum_{i=1}^{p} L_v^i(\boldsymbol{\gamma}_i; \boldsymbol{x}) =& -\sum_{i=1}^{p} \log \pi_i(\boldsymbol{\gamma}_i) - \sum_{i=1}^{p} \log p(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}_i}(\boldsymbol{x})) \\ & - \sum_{i=1}^{p} \log v_i(\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}_i}(\boldsymbol{x})) \\ & + \sum_{i=1}^{p} COMP(M_{\boldsymbol{\gamma}_i}; v) && (32) \\ =& -\log \pi(\boldsymbol{\gamma}) - \log p(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{x})) \\ & - v(\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{x})) + COMP(M_{\boldsymbol{\gamma}}; v) && (33) \\ =& L_v(\boldsymbol{\gamma}; \boldsymbol{x}). && (34)\end{aligned}$$

Thus, L_v^i as defined above satisfies Eq. 27 as required. Algorithm 2 summarizes the procedure for computing $L_v^i(\boldsymbol{\gamma}_i; \boldsymbol{x})$. First, we compute the MDL estimator $\hat{\boldsymbol{\theta}}_i$ for the i-th dimension by optimizing the goodness-of-fit, which is a convex optimization problem for an appropriate choice of luckiness function v, as discussed in subsection 4.3. Next, we estimate the model complexity by using Algorithm 1. Finally, we compute MDL objective as in Eq. 27. □

7.2 Experimental Setup

The comparison methods are estimation methods which search for the MHP kernel functions and baseline vector based on the data. To extract a causal graph from the achieved output, based on Theorem 1, we put a threshold on the kernel norm to distinguish zero and non-zero kernels. This threshold is set to 0.01.

Each of the comparison methods has a penalty and a level of regularization given. For ML, LS, and NPHC we have penalties: L1 (lasso), L2, elastic net, and none. For ADM4 we have lasso-nuclear penalization. We evaluated each of the baseline methods with all possible penalties and for a set of possible values (levels) for regularization

$$C \in \{1, 2, 5, 10, 20, 50, 100, 200, 500, 1000, 2000, 5000, ...\}.$$

For ADM4, we also used the nuclear lasso ratio; For each of the above settings we also considered the lasso-nuclear-ratio taking value in $\{0, 0.1, 05, 0.9, 1\}$.

The reported numbers for "Random" in Table 1 are the result of a random adjacency matrix with the same number of non-zero entries as the average test case. We report the highest F1 score achieved by each method based on different hyper-parameters. As we do not perform train/test validation and instead we take the highest F1, the validated F1 scores for baseline methods would be presumably lower.

Information criteria (AIC, BIC, and HQ) generally do not perform well for small data and it was the case also in our experiments. These methods rely on reducing model loss (i.e., negative log-likelihood) at the price of adding new parameters to the model. The least price that these models suggest for increasing the size of parameter set is about 1 unit of log-likelihood, hence, these model selection methods allow for adding any edge to the graph only if the log-likelihood could be increased by at least 1 unit compared to the empty graph model. This is not the case for a small data set (”short” data), as the value of log-likelihood that we have for the naive (empty) model and the maximum-likelihood model are both very small (of order 0.01 or 0.1 in all of our experimental settings). Therefore, the 1 unit improvement is not possible and this prevents the IC methods from discovering any edge in the causal graph.

Our method is an MDL-based model selection with no hyper-parameters. However, we can choose the number of MC simulations N for integral estimation, and the higher N the better estimate. Limited by our computational resources, we used 1000 iterations for the default case (for dimension $p = 7$), and 500 iterations for the sparse graph scenario (for dimension $p = 20$). As discussed in Section 4 in Subsection Amortization, we first do the MC simulations and compute model complexity values, which takes about one hour in each experimental setting (i.e., with fixed p and T), and then we perform 100 test runs, each taking a about ten seconds.

Real-world Data We observed n our synthetic experiments that method ML outperformed the other baseline methods for $p = 7$ and $T = 400$. Elastic net regularization was the best regularization for ML also in our synthetic experiments. So we ran this method on the data for the set of levels for regularization as listed above, and in all cases the bi-directed edge between US and Japan as an edge of the causal graph was returned. This is not plausible since it contradicts the expert knowledge from the domain.

References

Achab, M.; Bacry, E.; Gaıffas, S.; Mastromatteo, I.; and Muzy, J.-F. 2017. Uncovering causality from multivariate Hawkes integrated cumulants. In *Proceedings of the International Conference on Machine Learning*, 70:1–10. PMLR.

Bacry, E.; Bompaire, M.; Deegan, P.; Gaïffas, S.; and Poulsen, S. 2017. Tick: a Python library for statistical learning, with an emphasis on Hawkes processes and time-dependent models. *Journal of Machine Learning Research*, 18(1): 7937–7941.

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

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

Barron, A.; Rissanen, J.; and Yu, B. 1998. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44(6): 2743–2760.

Budhathoki, K.; and Vreeken, J. 2018. Origo: Causal inference by compression. *Knowledge and Information Systems*, 56(2): 285–307.

Chen, Y. 2016. Likelihood function for multivariate Hawkes processes. *Preprint (5 pages) available at https://www. math. fsu. edu/ychen/research/HawkesLikelihood. pdf.*

CNN-International.com. 2003. France accuses U.S of 'economic war'. www.edition.cnn.com/2003/WORLD/europe/06/15/paris.airshow.french.us.row/index.html.

Daley, D. J.; and Vere-Jones, D. 2003. Basic properties of the Poisson process. *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*, 19–40.

Demirer, M.; Diebold, F. X.; Liu, L.; and Yilmaz, K. 2018. Estimating global bank network connectedness. *Journal of Applied Econometrics*, 33(1): 1–15.

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

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

Grünwald, P.; and Roos, T. 2019. Minimum description length revisited. *International Journal of Mathematics for Industry*, 11(01): 1930001.

Hawkes, A. G. 1971. Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 33(3): 438–443.

Hlaváčková-Schindler, K.; and Plant, C. 2020a. Graphical Granger causality by information-theoretic criteria. In *Proceedings of the European Conference on Artificial Intelligence*, 1459–1466. IOS Press.

Hlaváčková-Schindler, K.; and Plant, C. 2020b. Heterogeneous graphical Granger causality by minimum message length. *Entropy*, 22(12): 1400.

Hlaváčková-Schindler, K.; and Plant, C. 2020c. Poisson graphical Granger causality by minimum message length. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*, 526–541. Springer.

Kim, S.; Putrino, D.; Ghosh, S.; and Brown, E. N. 2011. A Granger causality measure for point process models of ensemble neural spiking activity. *PLOS Computational Biology*, 7(3): e1001110.

Marx, A.; and Vreeken, J. 2017. Telling cause from effect using MDL-based local and global regression. In *Proceedings of the IEEE International Conference on Data Mining*, 307–316. IEEE.

Marx, A.; and Vreeken, J. 2018. Causal inference on multivariate and mixed-type data. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*, 655–671. Springer.

Mian, O.; Marx, A.; and Vreeken, J. 2021. Discovering fully oriented causal networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1–14.

Ogata, Y. 1981. On Lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, 27(1): 23–31.

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

Rissanen, J. 1998. *Stochastic complexity in statistical inquiry*, volume 15. World Scientific.

Shtarkov, I. M. 1978. Compression of multispectral video data. *Aerial and Spaceborne Observations of the Earth: Computer Processing of Video Data*, 36–55.

Veen, A.; and Schoenberg, F. P. 2008. Estimation of space–time branching process models in seismology using an EM–type algorithm. *Journal of the American Statistical Association*, 103(482): 614–624.

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

Zhou, K.; Zha, H.; and Song, L. 2013a. Learning social infectivity in sparse low-rank networks using multidimensional Hawkes processes. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 641–649. PMLR.

Zhou, K.; Zha, H.; and Song, L. 2013b. Learning triggering kernels for multi-dimensional Hawkes processes. In *Proceedings of the International Conference on Machine Learning*, 1301–1309. PMLR.