Causal Discovery in Hawkes Processes by Minimum Description Length

Amirkasra Jalaldoust [1] Kateřina Hlaváčková-Schindler [2] Claudia Plant [2]

[1]Department of Computer Science, Columbia University, New York, USA [2]Faculty of Computer Science, University of Vienna, Vienna, Austria

Abstract

We investigate the problem of causal discovery in temporal point processes named Hawkes processes. This is a special class of temporal point processes exhibiting a natural notion of causality, as occurrence of events in the past may increase the probability of events in the future. Our approach to Granger causal inference is based on minimum description length (MDL) principle for statistical inference. Our contribution is two fold: 1. We propose a novel general routine for estimation of MDL objective function based on Monte-Carlo simulations. 2. We employ our method to perform causal discovery in Hawkes processes. Experiments on real and synthetic datasets demonstrate significant superiority of our method in comparison with state-of-the-art algorithms.

Temporal Point Processes

A multivariate-dimensional temporal point process (TPP) is a set of random processes which is used to model occurrence of multi-type events in time. A p-variate TPP can be represented by a collection of counting processes $\{U_i\}$ where $U_i(t)$ for $t \in [0, T]$ is the number of events happened prior to time t. For each dimension i the conditional intensity function is defined as

$$\lambda_i(t) = \mathbb{E}[dU_i(t)|\mathcal{H}_t] = \lim_{\Delta t \to 0} \mathbb{E}[U_i(t + \Delta t) - U_i(t)|\mathcal{H}_t], \tag{1}$$

where $\mathcal{H}_t$ is called the filtration, i.e the history of the process prior to time t.

Granger Causality in Temporal Point Processes

In p-dimensional time-series variable $\boldsymbol{x}_j$ Granger-causes $\boldsymbol{x}_i$ variable when the future of $\boldsymbol{x}_i$ is better predicted when taking into account the past of variable $\boldsymbol{x}_j$.
The i-th dimension of TPP Granger-causes the j-th dimension if

$$\lambda_i(t) = \mathbb{E}[dU_i(t)|\mathcal{H}_t] \neq \mathbb{E}[dU_i(t)|\mathcal{H}_t^{-j}] \tag{2}$$

where $\mathcal{H}_t^{-j}$ denotes the history of the process excluding events in the j-th dimension.
Hawkes 1971: A p-dimensional Hawkes process (MHP) is a p-dim TPP with conditional intensity

$$\lambda_i(t) = \mu_i + \sum_{j=1}^{p} \int_{-\infty}^{t} \phi_{ij}(t - \tau) dU_j(\tau). \tag{3}$$

$\boldsymbol{\mu}$ is a vector of exogenous baseline intensities. Functions $\phi_{ij} : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ are kernel functions specifying how the previous events in the j-th dim influences the intensity of the i-th dim.

Theorem 1. (Eichler et al., 2016). In a Hawkes process, the events in the j-th dimension do not Granger-cause events in the i-th dimension if and only if $\phi_{ij} \equiv 0$.

Parameter learning in a parametric class

We consider p-dim Hawkes processes with exp-decay kernels (exp-MHP) $\phi_{ij} = \alpha_{ij} \exp(-\beta_{ij} t)$ where the decay matrix $\boldsymbol{\beta}$ is constant and known, while the influence coefficient matrix $\boldsymbol{\alpha}$ determines the magnitude of the effect, and it is the parameter of the model. Let $\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \boldsymbol{\alpha}_i^\top)^\top \in (\mathbb{R}_0^+)^{p+1}$ be the parameters associated to i-th dimension. We define the parameter vector of exp-MHP as $\boldsymbol{\theta} = [\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top, \dots, \boldsymbol{\theta}_p^\top]^\top \in (\mathbb{R}_0^+)^{p+p^2}$.
Facts:

1. Likelihood for exp-MHP is convex in θ and has a closed form (Ozaki 1979).
2. The objective is convex in θ (Ogata 1981).
3. Solving MLE for exp-MHP can be done efficiently (Bacry et al. 2015).

Normalized Maximum Likelihood

Denote a parametric statistical model by a family of probability distributions $M = \{p(.|\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ where $\boldsymbol{\Theta}$ is the parameter space. We partition $\boldsymbol{\Theta}$ into disjoint subsets $\{\boldsymbol{\Theta}_{\boldsymbol{\gamma}} : \boldsymbol{\gamma} \in \boldsymbol{\Gamma}\}$ to define family of models $\{M_{\boldsymbol{\gamma}} : \boldsymbol{\gamma} \in \boldsymbol{\Gamma}\}$. For data $\boldsymbol{x} \in \mathcal{X}$, the **normalized maximum likelihood** (NML) distribution (**Shtarkov 1978**) is a general technique applying the MDL principle to statistical model selection. It depends on the luckiness function $v : \boldsymbol{\Theta} \to \mathbb{R}_0^+$ and for each submodel $M_{\boldsymbol{\gamma}}$ is

$$p_{v|\boldsymbol{\gamma}}^{NML}(\mathbf{x}) = \frac{\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_{\boldsymbol{\gamma}}} p(\mathbf{x}|\boldsymbol{\theta}) v(\boldsymbol{\theta})}{\int_{\mathcal{X}} \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_{\boldsymbol{\gamma}}} p(\mathbf{s}|\boldsymbol{\theta}) v(\boldsymbol{\theta}) d\mathbf{s}}. \tag{4}$$

The logarithm of the integral is called the **model complexity** $COMP(M_{\boldsymbol{\gamma}}; v)$.

MDL-Based Statistical Inference

Denote $R_v(\boldsymbol{\theta}, \boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{\theta}) v(\boldsymbol{\theta})$ and let $r_v(\boldsymbol{\theta}; \boldsymbol{x}) := \log R_v(\boldsymbol{\theta}, \boldsymbol{x}) = \log p(\boldsymbol{x}|\boldsymbol{\theta}) + \log v(\boldsymbol{\theta})$ for each $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. We define the MDL estimator based on v for a specific model $M_{\boldsymbol{\gamma}} \subset M$ as $\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{x}) := \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_{\boldsymbol{\gamma}}} -r_v(\boldsymbol{\theta}; \boldsymbol{x})$. The model selection on NML picks over $\boldsymbol{\Gamma}$ the model minimizing

$$L_v(\boldsymbol{\gamma}; \boldsymbol{x}) = -\log \pi(\boldsymbol{\gamma}) - \log p_{v|\boldsymbol{\gamma}}^{NML}(\boldsymbol{x}) = -\log \pi(\boldsymbol{\gamma}) - r_v(\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{x}); \boldsymbol{x}) + COMP(M_{\boldsymbol{\gamma}}; v) \tag{5}$$

where $\pi(.)$ is a given distribution on $\boldsymbol{\Gamma}$.
The **MDL** function incorporates **the trade-off between goodness-of-fit** measured by $r_v(\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{x}); \boldsymbol{x})$ and **the model complexity** $COMP(M_{\boldsymbol{\gamma}}; v)$. The model selection based on MDL selects

$$\hat{\boldsymbol{\gamma}}^{MDL} = \arg\min_{\boldsymbol{\gamma} \in \boldsymbol{\Gamma}} L_v(\boldsymbol{\gamma}; \boldsymbol{x}). \tag{6}$$

Our Method for Estimation of MDL Objective Function

Computation of the goodness-of-fit

To compute the goodness-of-fit term $r_v(\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{x}); \boldsymbol{x})$ we need to efficiently compute the MDL estimator $\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}$ which is the result of a minimization problem over the restricted parameter space $\boldsymbol{\Theta}_{\boldsymbol{\gamma}}$. If the minimization problem is convex, the solution would be unique and efficiently computable. Therefore we require two (mild) conditions:

1. $\boldsymbol{\Theta}_{\boldsymbol{\gamma}}$ is a convex set, e.g., a Euclidean space.
2. $-r_v(\boldsymbol{\theta}; \boldsymbol{x}) = -\log(p(\boldsymbol{x}|\boldsymbol{\theta})) - \log(v(\boldsymbol{\theta}))$ is a convex function.

Estimation of model complexity

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

$$Q_{v|\boldsymbol{\gamma}}(\boldsymbol{s}, \boldsymbol{z}) = \frac{R_v(\hat{\boldsymbol{\theta}}_{v|\boldsymbol{\gamma}}(\boldsymbol{s}); \boldsymbol{s})}{p(\boldsymbol{s}|\boldsymbol{z})} \tag{7}$$

and rewrite model complexity as $COMP(M_{\boldsymbol{\gamma}}; v) = \log \mathbb{E}_{\boldsymbol{X} \sim p}[Q_{v|\boldsymbol{\gamma}}(\boldsymbol{X}, \boldsymbol{\theta})|\boldsymbol{\theta} = \boldsymbol{z}]$.
We randomize $\boldsymbol{z}$ to any full support distribution, we can rewrite model complexity

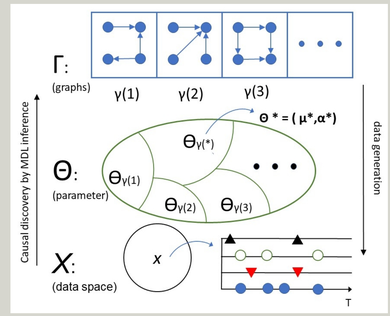
$$COMP(M_{\boldsymbol{\gamma}}; v) = \log \mathbb{E}_{\boldsymbol{X}, \boldsymbol{\theta} \sim p}[Q_{v|\boldsymbol{\gamma}}(\boldsymbol{X}, \boldsymbol{\theta})]. \tag{8}$$

$Q_{v|\boldsymbol{\gamma}}$ is efficiently computable at all points. By taking multiple joint samples of $\boldsymbol{X}$ and using them in $Q_{v|\boldsymbol{\gamma}}$ we may get random draws of $Q_{v|\boldsymbol{\gamma}}(\boldsymbol{X}, \theta)$. Taking the average of these values gives us an unbiased estimator for the above expectation.

Causal Discovery in Hawkes Processes

Causal discovery as a model selection problem

Let $\boldsymbol{\Gamma}$ be the set of all $p \times p$ binary matrices and let M be a set of all p-dim exp-MHP models. For each $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}$, denote $M_{\boldsymbol{\gamma}} \subset M$ a set of all p-dim exp-MHP models with **Granger causal graph** $\boldsymbol{\gamma}$ (adjacency matrix). According to Theorem 1, $\alpha_{ij}^* = 0 \iff \gamma_{ij}^* = 0$. Hence, the true parameter θ^* lies in exactly one sub-model: M_{γ^*}. Therefore, **finding the true causal graph γ^* is equivalent to finding the true sub-bmodel M_{γ^*} which contains the true parameter θ^*.**



Algorithm MDLH:
Solving the independent optimization problems $\hat{\boldsymbol{\gamma}}_i^{MDL} = \arg\min_{\boldsymbol{\gamma}_i \in \{0,1\}^p} L_v^i(\boldsymbol{\gamma}_i; \boldsymbol{x})$, we get the MDL estimator - **causal graph** $\hat{\boldsymbol{\gamma}}^{MDL} = [\hat{\boldsymbol{\gamma}}_1^{MDL} | \hat{\boldsymbol{\gamma}}_2^{MDL} | \dots | \hat{\boldsymbol{\gamma}}_p^{MDL}]^T$ with $\mathcal{O}(p2^p)$ complexity. For sparse graphs with max degree $\delta \ll p$ we get polynomial complexity $\mathcal{O}(Np^\delta)$ for N Monte Carlo simulations. The algorithm can be also amortized and paralellized.

Synthetic and real experiments

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

MDLH and state-of-art methods in F1 measure: ADM4 from Zhou et al. 2013, NPHC from Achab et al. 2017; ML is max. likelihood, LS least squares, IC the best F1 of AIC, BIC and HQ.

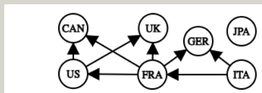


Figure 3: G-7 causal graph inferred by MDLH

Influence network among sovereign bonds of 7 large and developed economies called G-7 from daily return volatility 2003-2014 (data from Demirer et al. 2018). More details in our paper.

Contact: jalaldoust@cs.columbia.edu

36th AAAI Conference on Artificial Intelligence 2022, Vancouver, Canada - Virtual Conference

Contact: katerina.schindlerova@univie.ac.at, claudia.plant@univie.ac.at