Causal Discovery in Hawkes Processes by Minimum Description Length

Kasra Jalaldoust
Columbia University

Katerina Hlaváčková-Schindler, Claudia Plant
University of Vienna

universität wien

COLUMBIA UNIVERSITY

AAAI 2022

Summary

- We investigate the problem of causal discovery in a special class of temporal point processes named Hawkes processes.
- Our approach is based on Minimum Description Length (MDL) principle for statistical inference.
- Causal discovery task is to infer the underlying influence network in a system of random variables.
- 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.

Outline

Preliminaries

- Granger causality in temporal point processes
- MDL-based statistical inference

Our Method to Estimate the MDL Objective Function

Causal Discovery in Hawkes Processes

- Causal Discovery ~ Model selection
- Algorithm

Experiments and Discussion

Granger Causality In Temporal Point Processes

- A multi-dimensional temporal point process is a set of random process which is used to model occurrence of multi-type events in time.
- Each realization of the process is a collection of event lists. The i-th list is denoted as $\{t_1^i, t_2^i, \ldots, t_{n_i}^i\}$ where $t_j^i \in [0,T]$, and T is called the *horizon*.
- A temporal point process can be equivalently 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

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

Where $\mathscr{H}_t$ is the filtration, i.e., the history of the process before time t.

Granger Causality In Temporal Point Processes

- In multi-dimensional time-series variable $\mathbf{x}_j$ Granger-causes variable $\mathbf{x}_i$ when the future of $\mathbf{x}_i$ is better predicted when taking into account the past of variable $\mathbf{x}_j$.

- According to the above definition, the i-th dimension of the process Granger-causes the j-th dimension if

$$\lambda_i(t) = \mathbb{E}[dU_i(t) \mid \mathscr{H}_t] \neq \mathbb{E}[dU_i(t) \mid \mathscr{H}_t^{-j}],$$

  where $\mathscr{H}_t^{-j}$ denotes the history of the process excluding events in the j-th dimension.

Granger Causality In Temporal Point Processes

- Hawkes 1971: A Hawkes process is a temporal point process with the intensity function defined:

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

1. p is the number of dimensions.
2. μ is a vector in positive cone representing the exogenous baseline intensities for each of the dimensions.
3. The functions $\phi_{ij} : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ are the kernel functions specifying how the previous events in the j-th dimension increases the intensity of the i-th dimension at time t.

- **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 $\phi_{ij} \equiv 0$.

MDL-based Statistical Inference

- We denote a parametric statistical model by a family of probability distributions

$$M = \{p(\,.\,|\theta) : \theta \in \mathbf{\Theta}\},$$

  where $\mathbf{\Theta}$ denotes the parameter space.

- Further we partition the parameter space into disjoint subsets $\{\mathbf{\Theta}_\gamma : \gamma \in \mathbf{\Gamma}\}$ to define family of models $\{M_\gamma : \gamma \in \mathbf{\Gamma}\}$.

- We denote data by $\mathbf{x} \in \mathcal{X}$.

MDL-based Statistical Inference

- The Normalized Maximum Likelihood (NML) distribution (Shtarkov 1978) provides a general technique to apply the Minimum Description Length (MDL) principle for statistical model selection.
- The NML distribution depend on the *luckiness function* $v : \boldsymbol{\Theta} \rightarrow \mathbb{R}_0^+$
- The NML distribution for each sub-model M_γ is defined as

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

- The logarithm of the normalizing integral is called the *model complexity*, denoted by

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

MDL-based Statistical Inference

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

$$r_v(\theta; \mathbf{x}) = \log R_v(\theta; \mathbf{x}) = \log p(\mathbf{x} \,|\, \theta) + \log v(\theta)\,.$$

- We define the *MDL estimator* based on luckiness function v for a specific sub-model $M_\gamma \subset M$ as

$$\hat{\theta}_{v|\gamma}(\mathbf{x}) = \arg \min_{\theta \in \mathbf{\Theta}_\gamma} - r_v(\theta; \mathbf{x}).$$

MDL-based Statistical Inference

- Given a distribution π over the model space $\mathbf{\Gamma}$, for data $\mathbf{x} \in \mathcal{X}$ and sub-model $M_\gamma \subset M$, the MDL objective function is defined as

$$
\begin{aligned}
L_v(\gamma; \mathbf{x}) &= -\log \pi(\gamma) - \log p^{NML}_{v|\gamma}(\mathbf{x}) \\
&= -\log \pi(\gamma) - r_v(\hat{\theta}_{v|\gamma}(\mathbf{x}); \mathbf{x}) + COMP(M_\gamma; v)
\end{aligned}
$$

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

- MDL-based model selection criterion suggests the sub-model

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

Our Method to Estimate MDL Objective

Computing goodness-of-fit

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

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

        A. $-\log p(\mathbf{x} \mid \theta)$ is *usually* convex, e.g., in linear regression and exp. families.

        B. $-\log v(\theta)$ is convex for any log-concave choice of luckiness function.

Our Method to Estimate MDL Objective

Estimating model complexity

- **Remark:** $COMP(M_\gamma; v) = \log \int_{\mathcal{X}} R_v(\hat{\theta}_{v|\gamma}(\mathbf{s}); \mathbf{s}) d\mathbf{s}$, and for any $\mathbf{s} \in \mathcal{X}$ we efficiently compute $\hat{\theta}_{v|\gamma}(\mathbf{s})$.

- For any $\mathbf{z} \in \boldsymbol{\Theta}$, we have

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

Our Method to Estimate MDL Objective

Estimating model complexity

- For simplicity, define

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

  and rewrite model complexity as

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

- By randomizing the parameter θ according to any full-support distribution, we might rewrite model complexity

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

Our Method to Estimate MDL Objective

Estimating model complexity

- For simplicity, define

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

  and rewrite model complexity as

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

- By randomizing $\mathbf{z}$ according to any full-support distribution, we might rewrite model complexity

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

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

Causal Discovery in Hawkes Processes

Parameter learning in a parametric class

- We restrict ourselves to multi-dimensional Hawkes processes with exponential-decay kernels (exp-MHP). In this class, the kernels functions are

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

  where the decay matrix β is constant and known, while the influence coefficient matrix α determines the magnitude of the effect, and it is the parameter of the model.

- The baseline vector μ is also parameter of the model.

Causal Discovery in Hawkes Processes

Parameter learning in a parametric class

- Let $\theta_i = (\mu_i, \alpha_{i\cdot}^T)^T \in (\mathbb{R}_0^+)^{p+1}$ denote the parameters associated with the i-th dimension.
- Let $\theta = (\theta_1^T, \theta_2^T, \ldots, \theta_p^T)^T \in \in (\mathbb{R}_0^+)^{p+p^2}$ be the parameter vector of exp-MHP.
- Likelihood for exp-MHP can be computed efficiently (Ozaki 1979).
- The objective is convex in θ (Ogata 1981).
- There are efficient implementations for solving MLE for exp-MHP (Bacry, Mastromatteo, and Muzy 2015).

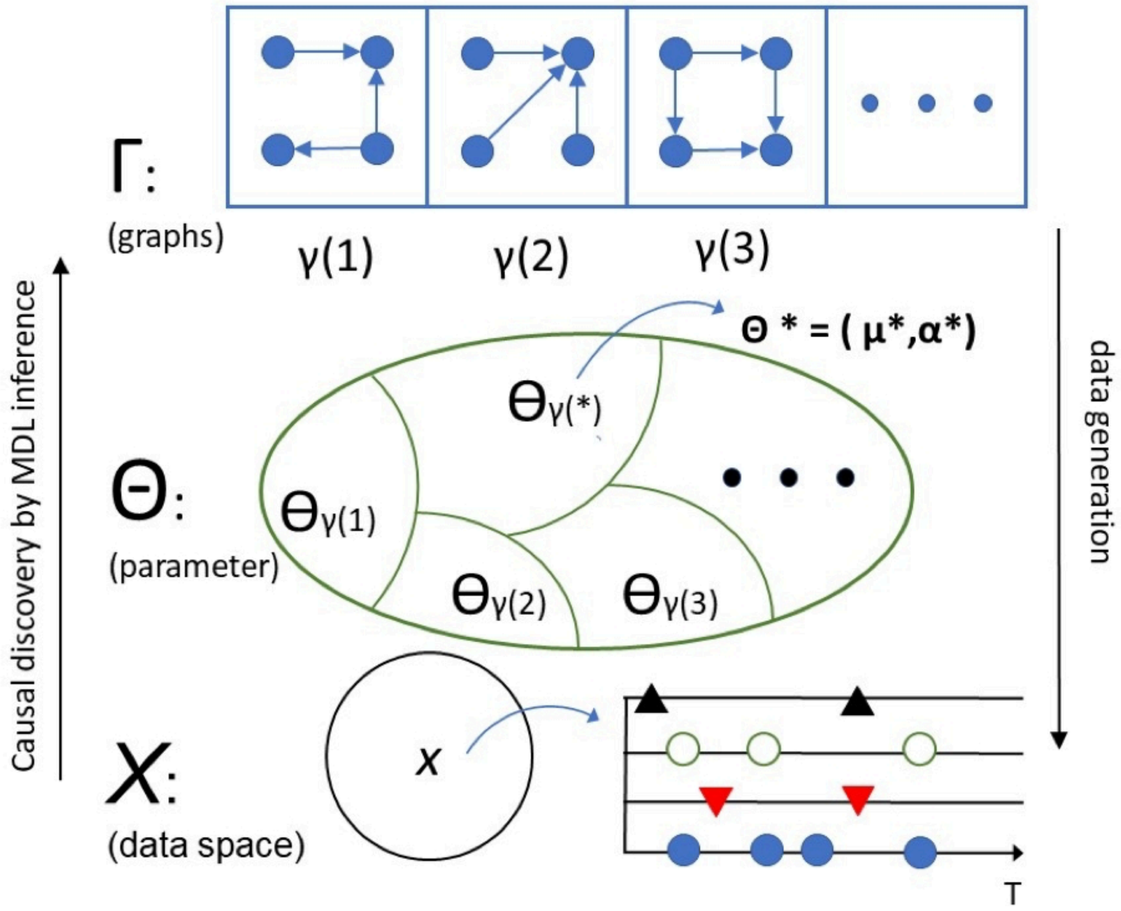
Causal Discovery in Hawkes Processes

Causal discovery as a model selection problem

- Let $\mathbf{\Gamma}$ be the set of all binary $p \times p$ matrices.
- Let M denote all p-dimensional exp-MHP models.
- For each $\gamma \in \mathbf{\Gamma}$, let $M_\gamma \subset M$ denote all p-dimensional exp-MHP models with Granger causal graph γ (adjacency matrix).
- According to Theorem 1, $\alpha_{ij}^* = 0$ if and only if $\gamma_{ij}^* = 0$. 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-model M_{γ^*} which contains the true parameter θ^*.

Causal Discovery in Hawkes Processes

Causal discovery as a model selection problem



Causal Discovery in Hawkes Processes

Algorithm

- **Fact:** conditions for MDL objective being well-behaved are satisfied in this formulation. Details are available in the paper.
- Exhaustive search over $\mathbf{\Gamma}$ is inefficient even for small p; $|\mathbf{\Gamma}| \geq 2^{p^2}$
- **Theorem 2.** If in MDL-based model selection for exp-MHP,

$$\pi(\gamma) = \prod_{i=1}^{p} \pi_i(\gamma_\mathbf{i}), \quad \nu(\theta) = \prod_{i=1}^{p} \nu_i(\theta_i),$$

then the MDL function can be rewritten as p independent terms

$$L_\nu(\gamma; \mathbf{x}) = \sum_{i=1}^{p} L_\nu^i(\gamma_i; \mathbf{x}),$$

such that each term is computable by Algorithm 2 in the paper.

Causal Discovery in Hawkes Processes

Algorithm

- By solving the independent optimization problems

$$\hat{\gamma}_i^{MDL} = \arg \min_{\gamma_i \in \{0,1\}^p} L_v^i(\gamma_i; \mathbf{x}),$$

we achieve the MDL estimator (causal graph)

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

in computational complexity of $O(p2^p)$.

Causal Discovery in Hawkes Processes

Algorithm: Discussion

- **Complexity.** N Monte-Carlo simulations $\Longrightarrow (N+1) \cdot p \cdot 2^p$ exp-MHP parameter estimation procedures.
- **Sparse graphs.** Max degree $\delta \ll p \Longrightarrow$ polynomial complexity $O(N \cdot p^\delta)$.
- **Amortization.** Multiple discovery tasks of the same specification (e.g., horizon, dimension, hyperparameters, etc.) $\Longrightarrow$ run N Monte-Carlo simulations and use the results for all of them.
- **Parallelization.** The p independent optimization problems (Theorem 2) can be performed in parallel. All N Monte-Carlo simulations are independent and can be performed in parallel.

Experiments and Discussion

Synthetic datasets

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

Experiments and Discussion

Real dataset

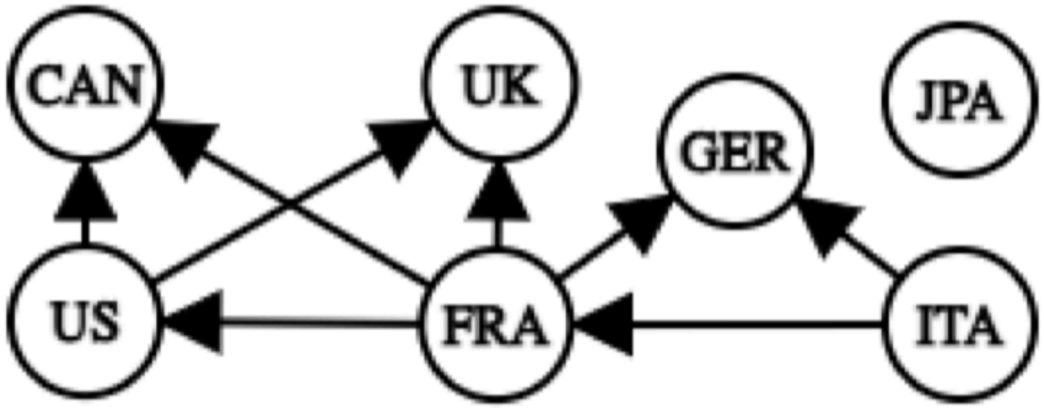


Figure 3: G-7 causal graph inferred by MDLH

Thank you for your attention!