Cardinality-Regularized Hawkes-Granger Model

Tsuyoshi ("Ide-san") Idé, Georgios Kollias, Dzung T. Phan, Naoki Abe, T.J. Watson Research Center, IBM Research

Advances in Neural Information Processing Systems 34 (NeurIPS 2021), to appear. Presentation slides: <u>https://ide-research.net/papers/2021_NeurIPS_Ide_presentation.pdf</u> @article{Ide20NeurIPS, title={Cardinality-Regularized Hawkes-Granger Model}, author={Tsuyoshi Id\'{e} and Georgios Kollias and Dzung T. Phan and Naoki Abe}, journal={Advances in Neural Information Processing Systems}, volume={34}, year={2021}, pages={2682--2694},

- Motivation and problem setting
- Hawkes process and Granger causality
- Minorization-maximization (MM) framework
- LOHawkes
- Experimental evaluation

Motivation: Event causal analysis to answer the question "who caused this?"

- Data: Marked (=multivariate) event sequence
 - Collection of (timestamp, event type)

 $\mathcal{D} = \{(t_0, d_0), (t_1, d_1), \dots, (t_N, d_N)\}$

- ✓ t_n : time stamp of the n-th event
- ✓ d_n : event type of the n-th (one of {1, ..., D})

Typical application: AlOps

- "Artificial Intelligence for IT Operations"
- Many (sub) modules of the IT system generates many error/warning events
- They are massive and myopic: making sense of what caused what is very challenging even to experienced engineers





Problem setting: For event causal analysis, we wish to find instanceand type-level causal relationship

- Given $\mathcal{D} = \{(t_0, d_0), (t_1, d_1), \dots, (t_N, d_N)\}$
 - d = 1 || || || || || |||| || || || 1 1 1 11 1 1 1 11 11 1111 1 11 1 Ш
- Find
 - Instance-level triggering probabilities (for each instance)
 - Type-level causal relationship





time

- Motivation and problem setting
- Hawkes process and Granger causality
- Minorization-maximization (MM) framework
- LOHawkes
- Experimental evaluation

Self-exciting point process (aka Hawkes process) is a good fit for our problem

- Hawkes process has been used in seismology to associate aftershocks with major earthquakes
 - Y. Ogata, "Seismicity analysis through point-process modeling: A review." Seismicity patterns, their statistical significance and physical meaning (1999): 471-507.
- Key quantity: event intensity function $\lambda_d(t \mid H_t)$
 - Probability density of event occurrence in the future, given event history H_t and an event type d



^{*} Picture source: Swiss Seismological Service, http://www.seismo.ethz.ch/en/home/



For event causal analysis, we employ a point-process model called the Hawkes process



- Maximum likelihood fitting of the impact matrix is the same as uncovering type-level Granger causality [Zhou+13][Eichler17] etc.
 - \circ Example: If A_{2,3}=0, the type-2 event is not caused by the type-3 event

Sparse learning is critically important for practical event causal analysis

 Granger causality: If Y shifted backward in time has a significantly high correlation with X, then X is a cause to Y.

○ Tricky part: "significantly high"



- Sparse learning provides a way of systematically ruling out unlikely options from a huge # of possibilities
 - $\circ~$ My PC in NY froze because of a flip of a butterfly in Tibet?
 - Did the sunshine cause Meursault to commit the murder? (Camus, "The Stranger").



- Motivation and problem setting
- Hawkes process and Granger causality
- Minorization-maximization (MM) framework
- LOHawkes
- Experimental evaluation

Vanilla minorization-maximization (MM) framework

Log likelihood

$$L = \sum_{n=1}^{N} \left\{ \boxed{\ln \lambda_{d_n}(t_n \mid \mathcal{H}_{n-1})} - \int_{t_{n-1}}^{t_n} du \lambda_{d_n}(u \mid \mathcal{H}_{n-1}) \right\}$$

$$= \ln \left\{ \mu_{d_n} + \sum_{i=0}^{n-1} A_{d_n,d_i} \phi_{d_n}(t_n - t_i) \right\}$$

$$\geq q_{n,n} \ln \frac{\mu_d}{q_{n,n}} + \sum_{i=0}^{n-1} q_{n,i} \ln \frac{A_{d_n,d_i} \phi_d(t_n - t_i)}{q_{n,i}}$$

4 types of parameters, 4 optimization problems



Finding the tightest bound of Jensen's inequality

- The optimization problem to solve for each n = 1, ..., N $\circ \max_{q_n} \left\{ q_{n,n} \ln \frac{\mu_d}{q_{n,n}} + \sum_{i=0}^{n-1} q_{n,i} \ln \frac{A_{d_n,d_i}\phi_d(t_n - t_i)}{q_{n,i}} \right\}$ subject to $\sum_{i=0}^n q_{n,i} = 1; \forall i, q_{n,i} > 0$ $\checkmark q_n = [q_{n,0}, ..., q_{n,n}]^T$ \circ Lagrangian $\mathcal{L} = q_{n,n} \ln \frac{\mu_d}{q_{n,n}} + \sum_{i=0}^{n-1} q_{n,i} \ln \frac{A_{d_n,d_i}\phi_d(t_n - t_i)}{q_{n,i}} - \lambda(\sum_{i=0}^n q_{n,i} - 1)$
- The objective is convex upward and has a maximum \circ (proof) Differentiate w.r.t. $q_{n,i}$ twice to get $-1/q_{n,i}$, which is always negative.
- The optimality condition is obtained by equating the first derivative of \mathcal{L} to 0

• **Resulting in:**
$$\ln \frac{\mu_d}{q_{n,n}} = \text{constant}, \quad \ln \frac{A_{d_n,d_i}\phi_d(t_n - t_i)}{q_{n,i}} = \text{constant}$$

 $\circ~$ The sum-to-one constraint leads to the solution shown previously

Leveraging Jensen bound for instance-level event causal analysis

The tightest Jensen bound is achieved if

$$q_{n,i} = \begin{cases} \frac{\mu_d}{\mu_d + \sum_{i=1}^{n-1} A_{d,d_i} \phi_{d_n}(t-t_i)} \\ \frac{A_{d,d_i} \phi_{d_n}(t-t_i)}{\mu_d + \sum_{i=1}^{n-1} A_{d,d_i} \phi_{d_n}(t-t_i)} \end{cases}$$

- This can be interpreted as the probability that the *i*-th event caused the *n*-th event, which we call the causal probability
- We use this for instance-level causal analysis



"I just got an event (t_n, d_n) . Tell me which event caused the particular event?"

 $\{ q_{n,i} \}$

 μ_1,\ldots,μ_D

MM solution for the baseline intensity $\mu = [\mu_1, ..., \mu_D]^+$

• Log-likelihood lower bound (collecting terms related to μ) with L2 regularizer

$$^{\circ} \quad L_{1} = \sum_{n=1}^{N} \left\{ q_{n,n} \ln \frac{\mu_{d_{n}}}{q_{n,n}} - \mu_{d_{n}} \Delta_{n,n-1} \right\} - \frac{1}{2} \nu_{\mu} \|\boldsymbol{\mu}\|_{2}^{2}, \quad \text{where} \quad \Delta_{n,n-1} \triangleq t_{n} - t_{n-1}$$

 \circ Second derivative is always negative \rightarrow convex upwards, a maximum exists

Getting maximizer by equating the first derivative to zero

$$\circ \quad 0 = \frac{\partial L}{\partial \mu_k} = \sum_{n=1}^N \delta_{d_n,k} \left\{ \frac{q_{n,n}}{\mu_k} - \Delta_{n,n-1} \right\} - \nu_\mu \mu_k, \quad \text{where } \delta_{d_n,k} \text{ is Kronecker's delta}$$

 $\circ~$ This is a quadratic equation and can be easily solved:

$$D_k^{\mu} = \sum_{n=1}^N \delta_{d_n,k} \Delta_{n,n-1}, \quad N_k^{\mu} = \sum_{n=1}^N \delta_{d_n,k} q_{n,n}, \quad \mu_k = \frac{1}{2\nu_{\mu}} \left(-D_k^{\mu} + \sqrt{(D_k^{\mu})^2 + 4\nu_{\mu} N_k^{\mu}} \right).$$

 $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_D$

MM solution for the decay parameter $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_D]^\top$: General solution

• Log-likelihood lower bound (collecting terms related to $m{eta}$) with L2 regularizer

$$C_{1} = \sum_{n=1}^{N} \sum_{i=0}^{n-1} \left\{ q_{n,i} \ln \frac{\phi_{d_{n}}(\Delta_{n,i})}{q_{n,i}} - A_{d_{n},d_{i}} h_{n,i} \right\} - \frac{1}{2} \nu_{\beta} \|\beta\|_{2}^{2}$$

$$h_{n,i} \triangleq \int_{t_{n-1}}^{t_{n}} \mathrm{d}u \,\phi_{d_{n}}(t-t_{i}) \cdot \frac{1}{2} \sum_{i=0}^{n-1} \frac{1}{2} \left\{ q_{n,i} \ln \frac{\phi_{d_{n}}(\Delta_{n,i})}{q_{n,i}} - A_{d_{n},d_{i}} h_{n,i} \right\} - \frac{1}{2} \left\{ p_{\beta} \|\beta\|_{2}^{2} \right\}$$

First derivative and optimality condition

$$^{\circ} \quad 0 = \frac{\partial L_1}{\partial \beta_k} = \sum_{(n,i)} \left\{ q_{n,i} \frac{\partial \ln \phi_{d_n}(\Delta_{n,i})}{\partial \beta_k} - A_{d_n,d_i} \frac{\partial h_{n,i}}{\partial \beta_k} \right\} - \nu_\beta \beta_k,$$

- Define nondimensional decay function $\varphi(\cdot)$ via $\phi_d(u) = \beta_d \varphi(\beta_d u)$
- General solution:

$$\beta_{k} = \frac{1}{2\nu_{\beta}} \left(-D_{k}^{\beta} + \sqrt{(D_{k}^{\beta})^{2} + 4\nu_{\beta}N_{k}^{\beta}} \right), \qquad N_{k}^{\beta} = \sum_{(n,i)} \delta_{d_{n},k} q_{n,i} = \sum_{n=1}^{\infty} \delta_{d_{n},k} (1 - q_{n,n}) \\ D_{k}^{\beta} = \sum_{(n,i)} \delta_{d_{n},k} \left\{ A_{k,d_{i}} \frac{\partial h_{n,i}}{\partial \beta_{k}} - q_{n,i} \frac{\varphi'(\beta_{k}\Delta_{n,i})}{\varphi(\beta_{k}\Delta_{n,i})} \right\}.$$

N

 β_1, \ldots, β_D

MM solution for the decay parameter $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_D]^\top$: Specific solution for the exponential and power distributions

- Exponential distribution
 - $\circ \varphi(u) = \exp(-u)$

$$D_{k}^{\beta} = \sum_{n=1}^{N} \delta_{d_{n},k} \sum_{i=0}^{n-1} \left[q_{n,i} \Delta_{n,i} + A_{k,d_{i}} \frac{\partial h_{n,i}}{\partial \beta_{k}} \right],$$
$$\frac{\partial h_{n,i}}{\partial \beta_{k}} = \delta_{d_{n},k} \left[\Delta_{n,i} e^{-\beta_{k} \Delta_{n,i}} - \Delta_{n-1,i} e^{-\beta_{k} \Delta_{n-1,i}} \right]$$

• Power distribution $\circ \varphi(u) = \eta (1+u)^{-\eta-1}$

$$D_k^{\beta} = \sum_{n=1}^N \delta_{d_n,k} \sum_{i=0}^{n-1} \left[\frac{(\eta+1)q_{n,i}\Delta_{n,i}}{1+\beta_k\Delta_{n,i}} + A_{k,d_i} \frac{\partial h_{n,i}}{\partial\beta_k} \right],$$
$$\frac{\partial h_{n,i}}{\partial\beta_k} = \delta_{k,d_n} \left\{ \frac{\eta\Delta_{n,i}}{(1+\beta_k\Delta_{n,i})^{\eta+1}} - \frac{\eta\Delta_{n-1,i}}{(1+\beta_k\Delta_{n-1,i})^{\eta+1}} \right\}.$$

Two contributions of this work

- First mathematically consistent approach to *sparse* causal learning through Hawkes process
- Simultaneous instance- and type-level event causal analysis for causal event diagnosis

- Motivation and problem setting
- Hawkes process and Granger causality
- Minorization-maximization (MM) framework

LOHawkes

Experimental evaluation

Existing "sparse" learning algorithms for A in fact cannot produce any sparse solutions

$$\begin{aligned} & \Psi(\mathsf{A}) \triangleq \sum_{n=1}^{N} \sum_{i=0}^{n-1} q_{n,i} \ln A_{d_n,d_i} - \sum_{n=1}^{N} \sum_{i=0}^{n-1} A_{d_n,d_i} \int_{t_{n-1}}^{t_n} \mathrm{d}t \; \phi(t-t_i) - \frac{1}{2} \nu_A \|\mathsf{A}\|_2^2 \\ & = \sum_{k=1}^{D} \sum_{l=1}^{D} (Q_{k,l} \ln A_{k,l} - H_{k,l} A_{k,l}) - \frac{1}{2} \nu_A \|\mathsf{A}\|_2^2 \quad \begin{array}{l} \text{added L2} \\ \text{regularizer} \end{array} \quad & Q_{k,l} \triangleq \sum_{(n,i)} \delta_{d_n,k} \delta_{d_i,l} q_{n,i} \\ & H_{k,l} \triangleq \sum_{(n,i)} \delta_{d_n,k} \delta_{d_i,l} h_{n,i} \\ & h_{n,i} \triangleq \int_{t_{n-1}}^{t_n} \mathrm{d}u \; \phi_{d_n}(t-t_i) \\ & H_{k,l} \triangleq \sum_{(n,i)} \delta_{d_n,k} \delta_{d_i,l} h_{n,i} \\ & H_{k,l} \triangleq \sum_{(n,i)} \delta_{d_n,k} \delta_{d_i,l} h_{n,i} \\ & H_{k,l} \triangleq \int_{t_{n-1}}^{t_n} \mathrm{d}u \; \phi_{d_n}(t-t_i) \\ & H_{k,l} \triangleq \int_{t_{n-1}}^{t_{n-1}} \mathrm{d}u \; \phi_{d$$

Existing sparse causal learning approach use L₁ or L_{2,1} regularizer:

$$\sum_{k=1}^{D} \sum_{l=1}^{D} (Q_{k,l} \ln A_{k,l} - H_{k,l} A_{k,l}) - \frac{1}{2} \nu_A \|\mathsf{A}\|_2^2 - \tau \|\mathsf{A}\|_p$$

Proof: Simply because $\ln 0 = -\infty$ and thus 0 is not allowed (easy!)

○ <u>Theorem 1</u>: For $p \ge 1$, this problem is convex and has a unique solution. The solution cannot be sparse, i.e., $A_{k,l} \ne 0$, if $Q_{k,l} \ne 0$ and $v_A \ne 0$.

How do we get a sparse solution in a legit way? Introducing L_0 -regularized problem with " ϵ -sparsity"

Proposed problem of our Hawkes-Granger framework:

$$\sum_{k=1}^{D} \sum_{l=1}^{D} \left(Q_{k,l} \ln A_{k,l} - H_{k,l} A_{k,l} - \frac{1}{2} \nu_A A_{k,l}^2 \right) - \tau \|\mathbf{A}\|_0$$
vectorized
vectorized
version
$$\max_{\boldsymbol{x}} \left\{ \sum_m \Psi_m(x_m) - \tau \|\boldsymbol{x}\|_0 \right\}, \quad \Psi_m(x_m) \triangleq \left(g_m \ln x_m - h_m x_m - \frac{\nu_A}{2} x_m^2 \right)$$

- Singularity remains at zero
- We introduce a "zero-ness" parameter ϵ and solve:

$$\max_{\boldsymbol{x}} \sum_{m} \left\{ \Psi_m(x_m) - \tau I(x_m > \epsilon) \right\}$$

Semi-analytic solution exists: Rare example of "solvable" L₀-regularized problem.

 $\circ\,$ c.f. [Phan&Ide SDM19] for the first proposal of ϵ -sparsity.

Solving L₀-regularized impact matrix estimation problem

• Objective function has a jump at $x_m = \epsilon$

$$\max_{\boldsymbol{x}} \sum_{m} \left\{ \Psi_m(x_m) - \tau I(x_m > \epsilon) \right\}$$

- The solution covers the three cases below
 - \circ Analytic solution using KKT conditions \rightarrow paper (easy)



- Motivation and problem setting
- Hawkes process and Granger causality
- Minorization-maximization (MM) framework
- LOHawkes
- Experimental evaluation

Comparison with "sparse" Hawkes algorithms: Do they produce a sparse impact matrix?

- Generated 10-dimensional synthetic event data
- Trained L₁- and L_{2,1}-regularized Hawkes models with many different regularization strength values



Comparison with neural Granger approaches:

Can they reproduce a true type-level causal graph?

State-of-the-art neural Granger models [Tank+21]

- cMLP: component-wise multi-layer perceptron
- cLSTM: component-wise long short-term memory
- Generated synthetic 5-dimensional event data with a very simple causal graph
 - For neural methods, the event data were converted into regular time series of counts
- Evaluated as a binary classification problem for each edge
 - True positive and true negative accuracies
- Why neural methods failed?
 - mainly due to equi-time-interval assumption





Table 1: Break-even accuracies in Fig. 3.		
L_0 Hawkes	cLSTM	cMLP
1.00 ± 0.11	0.43 ± 0.09	0.31 ± 0.10

Application of instance-level causal analysis: Event grouping for IT system management

- Real data center warning/errors over about two months
 - N=718, D=14
- (top) Instance triggering probabilities $\{q_{n,i}\}$
 - Sparse due to the sparsity of impact matrix A and time decay effect
- (bottom) The 150-th instance (type ETH_INIT)
 - $\circ~$ ETH_INIT: event type related to network initialization
 - Network-related events are reasonably associated
 - Noise event type "UPDOWN" is successfully suppressed
 - ✓ Informational event type that accounts for more than a half of instances



0.2

0.0

Summary

- Proposed a new LO-regularized Hawkes process for guaranteed sparsity
- Showed that existing sparse Hawkes models do not yield sparse solution
- Developed a new approach to event causal diagnosis, which leverages simultaneous type- and instance-level causal analysis