Skoltech
Skolkovo Institute of Science and Technology

Causality based event sequence modelling

Nikita Paplavskii, Polina Pilyugina

Models of Sequential Data 2021

Problem statement

Skoltech
Skolkovo Institute of Science and Technology

Event sequence is defined as:

$$E = ((k_1, t_1), (k_2, t_2), \ldots, (k_N, t_N))$$

- K — number of event types
- N — total number of events in sequence
- $k_i \in \{1, 2, \ldots, K\}$ is a particular event type occurred
- $t_1 < t_2 < \ldots < t_N$ is continuous time of event occurrence.

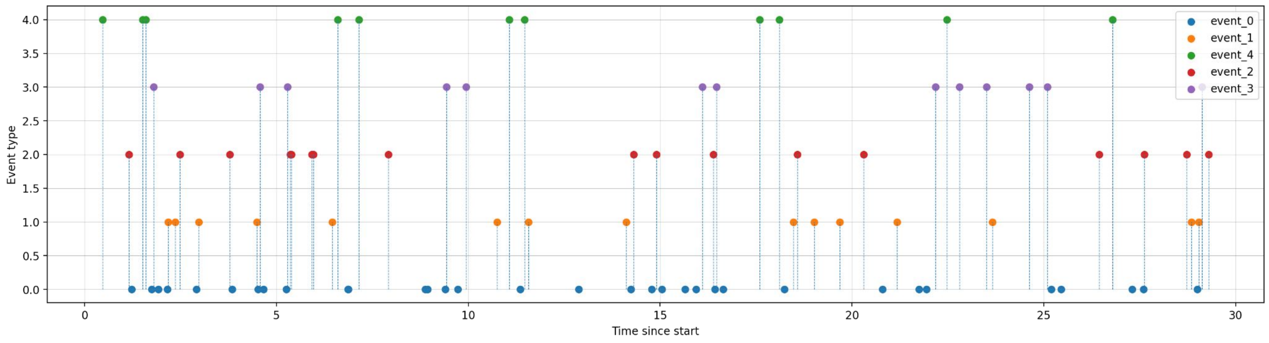


Figure 1. Example of an event sequence from hawkesinhib dataset

Neural Hawkes

Skoltech
Skolkovo Institute of Science and Technology

Hawkes process:

$$\lambda_k(t) = \mu_k + \sum_{h:t_h<t} \alpha_{k_h,k} \exp(-\delta_{k_h,k}(t - t_h))$$

where $\mu_k \geq 0$ is the base intensity of event type k, $\alpha_{j,k} \geq 0$ is the degree to which an event of type j initially excites type k, and $\delta_{j,k} \geq 0$ is the decay rate of that excitation

Neural Hawkes modified process:

$$\tilde{\lambda}_k(t) = \mu_k + \sum_{h:t_h<t} \alpha_{k_h,k} \exp(-\delta_{k_h,k}(t - t_h))$$

which allows inhibition ($\alpha_{j,k} < 0$) and inertia ($\mu_k < 0$)

$$\lambda_k(t) = f_k(\tilde{\lambda}_k(t))$$

and $f : R \rightarrow R^+$ is transfer function to obtain a positive intensity

Neural Hawkes

Neural Hawkes modified process reformulation for CTLSTM:

Given a time $t > 0$, the intensity of type k event $\lambda_k(t)$ is given by the following equations:

$$\lambda_k(t) = f_k(w_k^T h(t))$$
$$h(t) = o_i \odot (2\sigma(2c(t)) - 1) \text{ for } t_\in(t_i, t_{i+1}] \tag{4}$$

where $h(t)$ is hidden state and $c(t)$ is memory cell.

Neural Hawkes

Skoltech
Skolkovo Institute of Science and Technology

Neural Hawkes modified process reformulation for CTLSTM:

Given a time $t > 0$, the intensity of type k event $\lambda_k(t)$ is given by the following equations:

$$\lambda_k(t) = f_k(w_k^T h(t))$$
$$h(t) = o_i \odot (2\sigma(2c(t)) - 1) \text{ for } t_{\in}(t_i, t_{i+1}] \qquad (4)$$

where $h(t)$ is hidden state and $c(t)$ is memory cell.

It **does not have an explicit formulation of the $\alpha_{j,k}$,** which are present in the original formulation
Therefore, it **by construction cannot take advantage of the known-in-advance** matrix of such coefficients

Problem statement

Skoltech
Skolkovo Institute of Science and Technology

Problem:

- Neural Hawkes model for event sequences prediction **does not allow to take causality** into account

Aim:

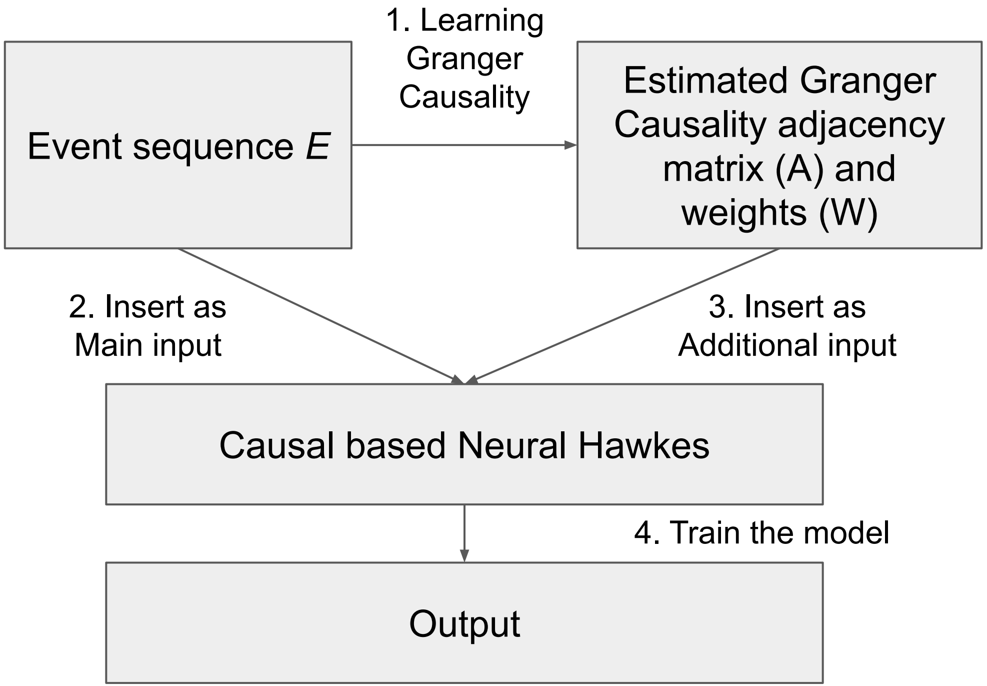
- Implement **causality estimation** for event sequences
- Implement **Causal based Neural Hawkes**
- **Test** it on **synthetic** and **real-world datasets**
- **Compare** it to the original **Neural Hawkes**

Hypothesis:

- Causal based model will **outperform original Neural Hawkes**

Our approach

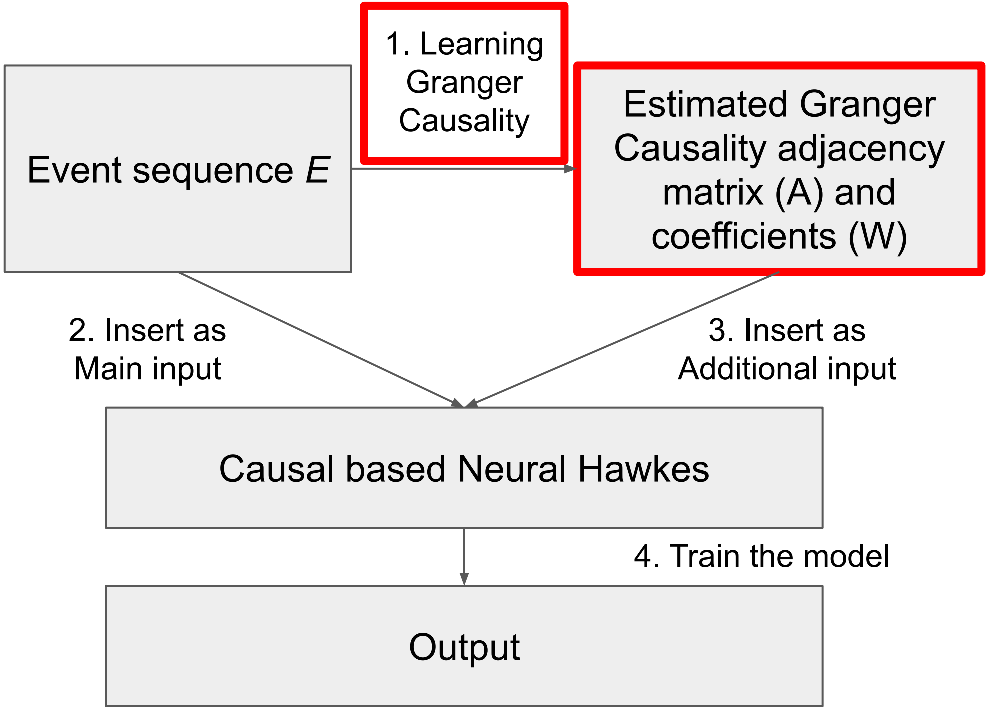
Our generalized approach looks as follows:



Our approach

Skoltech
Skolkovo Institute of Science and Technology

Our generalized approach looks as follows:



Methods: Causality of event sequences

Skoltech
Skolkovo Institute of Science and Technology

Hawkes process:

$$\lambda_k(t) = \mu_k + \sum_{h:t_h<t} \alpha_{k_h,k} \exp(-\delta_{k_h,k}(t - t_h))$$

where $\mu_k \geq 0$ is the base intensity of event type k, $\alpha_{j,k} \geq 0$ is the degree to which an event of type j initially excites type k, and $\delta_{j,k} \geq 0$ is the decay rate of that excitation

According to (1), the binarized *infectivity matrix* A = [sign $\alpha_{k_h,k}$ is the **adjacency matrix** of the corresponding **Granger Causality Graph**

Therefore, using Granger Causality estimation we can adjust the model to learn only those interdependencies which are non-zero in the estimated Granger Causality Graph

1. Eichler, M., Dahlhaus, R., & Dueck, J. (2017). Graphical Modeling for Multivariate Hawkes Processes with Nonparametric Link Functions. *Journal of Time Series Analysis*, *38*(2), 225–242. https://doi.org/10.1111/jtsa.12213

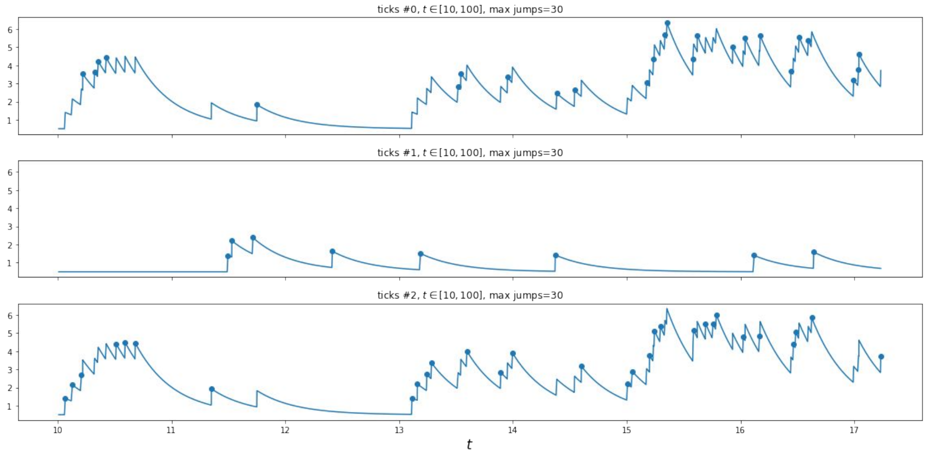
Methods: Causality of event sequences

Skoltech
Skolkovo Institute of Science and Technology

Granger Causality is a statistical concept of **causality based on prediction**

With event sequences, we want to learn whether **type-i event Granger-causes type-j event** and estimate **corresponding adjacency matrix**

In case of three event types, such adjacency matrix can look as follows:



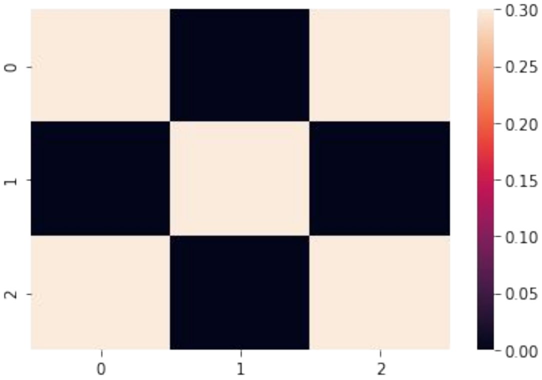


Figure 3. Simulated Hawkes process: its intensities (on the left) and its underlying adjacency matrix (on the right)

Methods: Causality of event sequences

Skoltech
Skolkovo Institute of Science and Technology

There are different methods for Granger Causality Estimation for Hawkes processes:

- **HawkesSumGaussians**[1] — combines the MLE with the sparse-group-lasso to learn the Granger causality graph of the target process.
- **Hawkes ADM4**[2] — estimates the infectivity matrix that is both low-rank and sparse by optimizing nuclear norm and L1 norm simultaneously

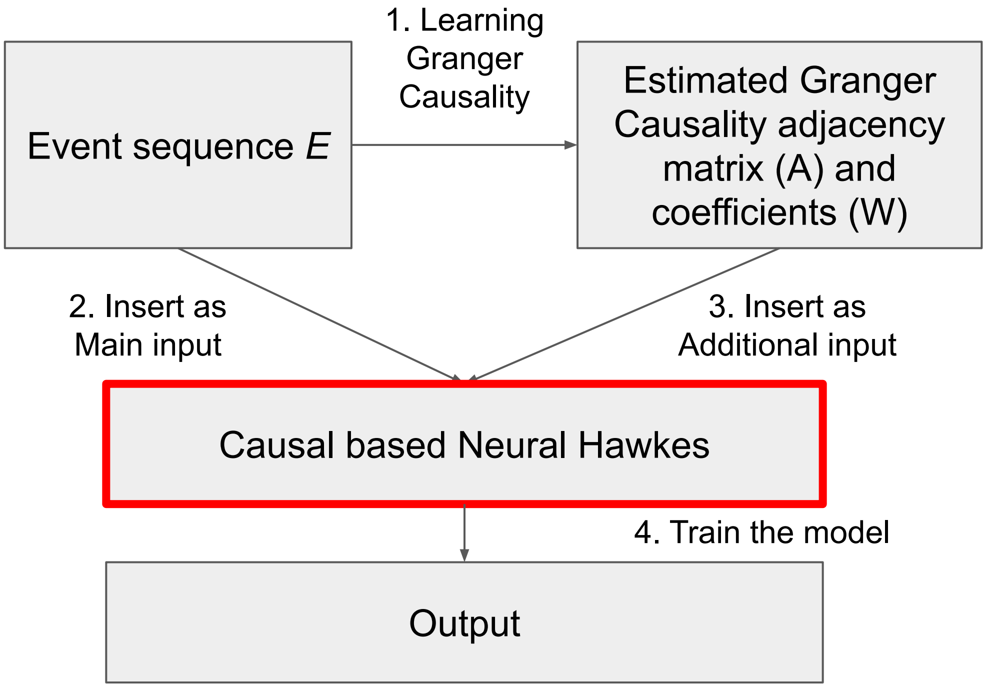
Aims:

1. Evaluate the performance of these methods on **synthetic data** with **known adjacency matrix**
2. Evaluate these methods on **real world datasets** by including them into **Causal based Neural Hawkes model**

1. Xu, Farajtabar, and Zha (2016, June) in ICML, Learning Granger Causality for Hawkes Processes.
2. Zhou, K., Zha, H., & Song, L. (2013, May). Learning Social Infectivity in Sparse Low-rank Networks Using Multi-dimensional Hawkes Processes. In AISTATS (Vol. 31, pp. 641-649).

Our approach

Our generalized approach looks as follows:



Methods: Causal-based NeuralHawkes

Neural Hawkes modified process reformulation for CTLSTM:

Given a time $t > 0$, the intensity of type k event $\lambda_k(t)$ is given by the following equations:

$$\lambda_k(t) = f_k(w_k^T h(t))$$
$$h(t) = o_i \odot (2\sigma(2c(t)) - 1) \text{ for } t_{\in}(t_i, t_{i+1}] \quad (4)$$

where $h(t)$ is hidden state and $c(t)$ is memory cell.

We modify it by multiplying $w_k^T h(t)$ with $A \odot W$ matrix before passing it to the SoftPlus function

$w_k^T h(t)$ — matrix of unnormalized intensities

$A \odot W$ — matrix A is adjacency matrix, W is matrix with estimated interaction coefficients

Methods: Causal-based NeuralHawkes

Skoltech
Skolkovo Institute of Science and Technology

For experiments we have created three versions of the model:

- **CausalNeuralHawkesMasked**:

  we initialize matrix W as all-ones with no gradient

- **CausalNeuralHawkesMaskedWeighted**:

  we initialize matrix W using the estimated interaction coefficients with no gradient

- **CausalNeuralHawkesTrainableWeighted**:

  we initialize matrix W using the estimated interaction coefficients with gradient

Thus obtaining three models we experiment with

1. Xu, Farajtabar, and Zha (2016, June) in ICML, Learning Granger Causality for Hawkes Processes.
2. Zhou, K., Zha, H., & Song, L. (2013, May). Learning Social Infectivity in Sparse Low-rank Networks Using Multi-dimensional Hawkes Processes. In AISTATS (Vol. 31, pp. 641-649).

Results: Data

Skoltech
Skolkovo Institute of Science and Technology

For experiments with model setup we created 4 synthetic datasets with predefined matrices A and W using ***tick*** library:

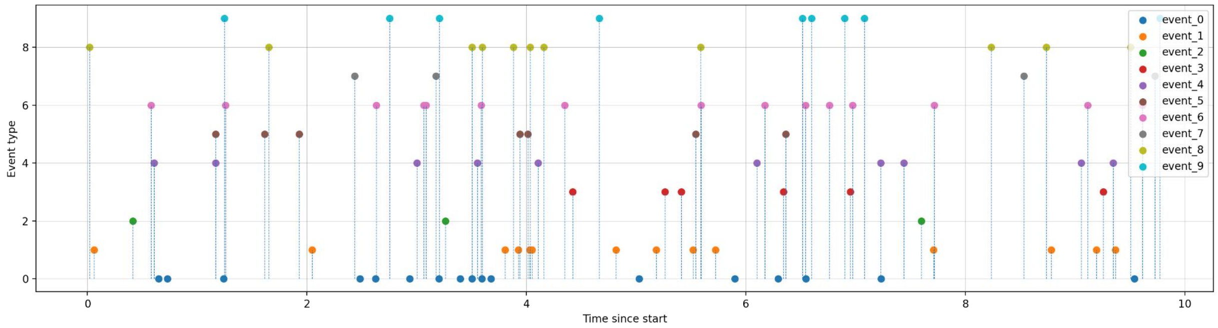
K — number of event types

T — maximal length of sequence

N — number of sequences

dataset_name	K	T	N
data_synth_5_events	5	22	300
data_synth_2_events	2	31	100
data_synth_3_events	3	23	300
data_synth_10_events	10	17	300

Table 1. Description of synthetic datasets used

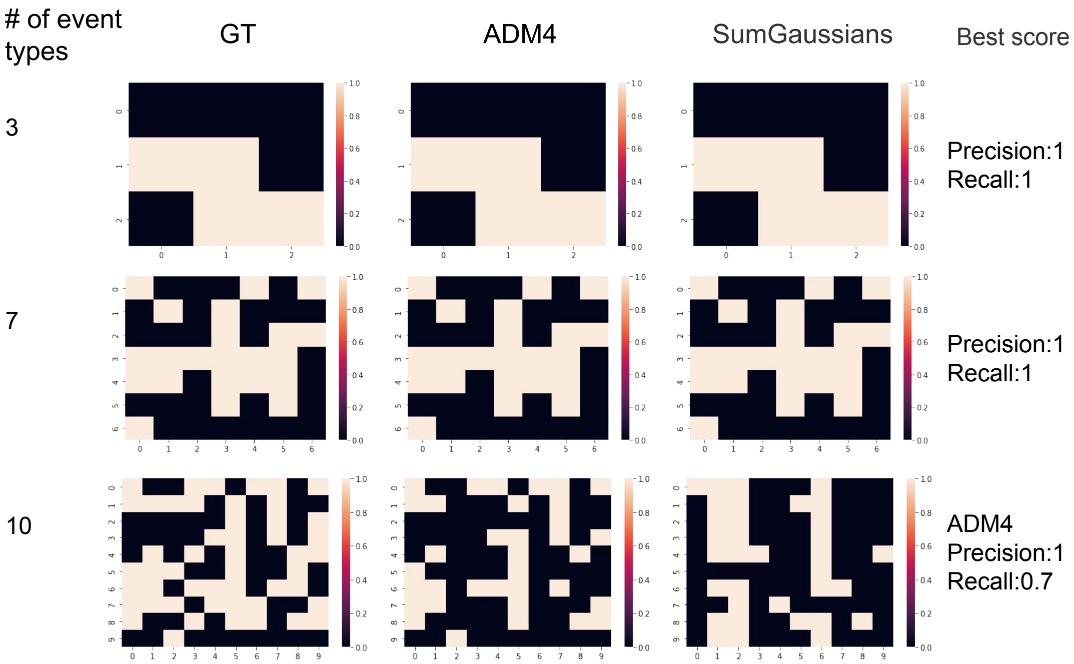


Sample event sequence from data_synth_3_events

Results: Causality estimation

To estimate adjacency matrix we use ***tick***[1] library.

1. Simulated Hawkes processes using **SimuHawkesExpKernels** function with exponential kernels simulation
2. Fitted **ADM4** and **SumGaussians** methods
3. Obtained adjacency matrices via binarization
4. Calculated **Precision and Recall** between Ground Truth and Estimated adjacency matrices



1. Emmanuel Bacry, Martin Bompaire, Stéphane Gaïffas, Soren Poulsen, Tick: a Python library for statistical learning, with a particular emphasis on time-dependent modelling

Results: Evaluation on synthetic data

Skoltech
Skolkovo Institute of Science and Technology

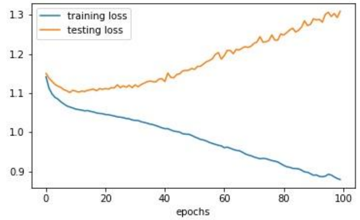
	NH			CNHM			CNHW			CNHTW		
Dataset	L	rmse	acc	L	rmse	acc	L	rmse	acc	L	rmse	acc
data_synth_ 10_events	0,88	0,131	15,79%	1,05	0,201	13,16%	1,06	0,324	14,47%	1,01	0,36	13,16%
data_synth_ 2_events	1,02	0,728	47,62%	0,93	0,876	47,62%	0,98	1,029	**71,43%**	0,94	0,47	61,90%
data_synth_ 3_events	0,86	0,357	30,56%	0,87	0,345	27,78%	0,84	0,367	**47,22%**	0,88	0,35	25,00%
data_synth_ 5_events	0,87	0,42	18,18%	0,89	0,417	33,33%	0,89	0,411	30,30%	0,89	0,38	**33,33%**

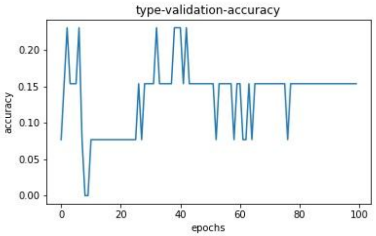
Models performance comparison on synthetic datasets

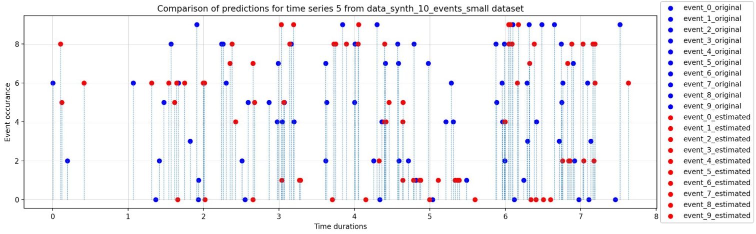
Results: Evaluation on synthetic data

Synth_10_events_small

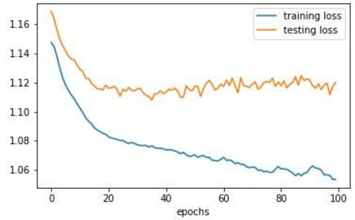
NeuralHawkes

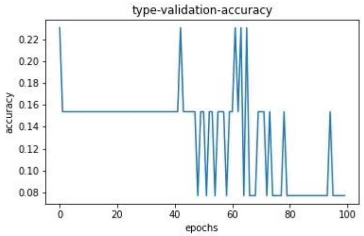


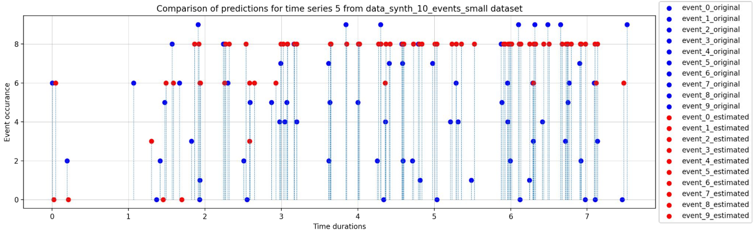




NeuralHawkesMasked





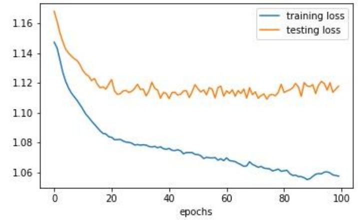


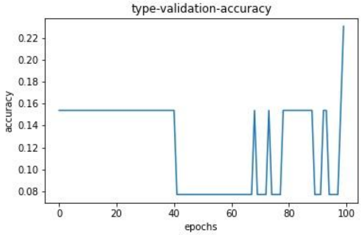
Results: Evaluation on synthetic data

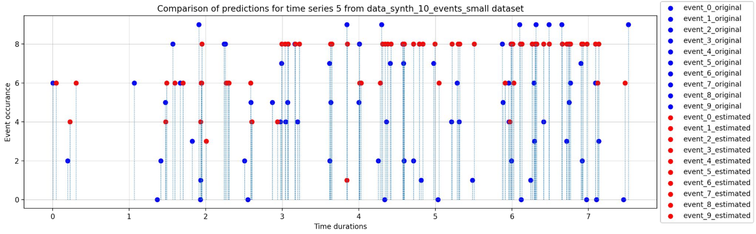
Skoltech
Skolkovo Institute of Science and Technology

Synth_10_events_small

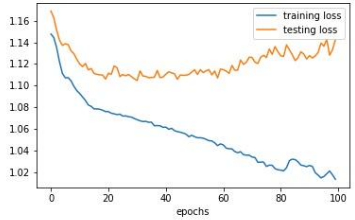
CausalNeuralHawkesMaskedWeighted

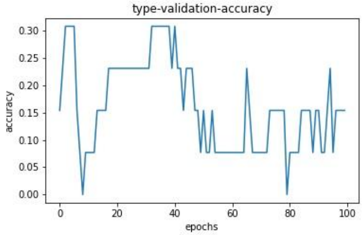


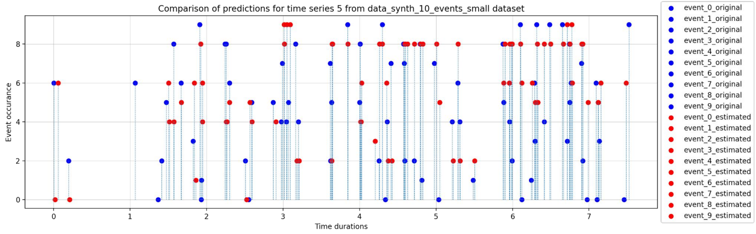




CausalNeuralHawkesTrainableWeighted





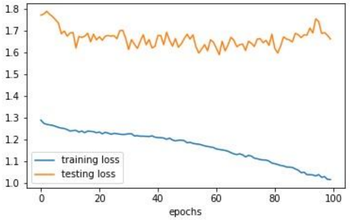


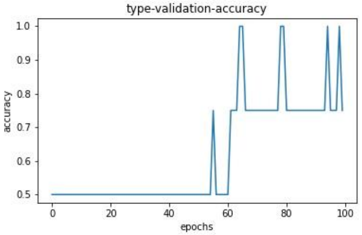
Skoltech
Skolkovo Institute of Science and Technology

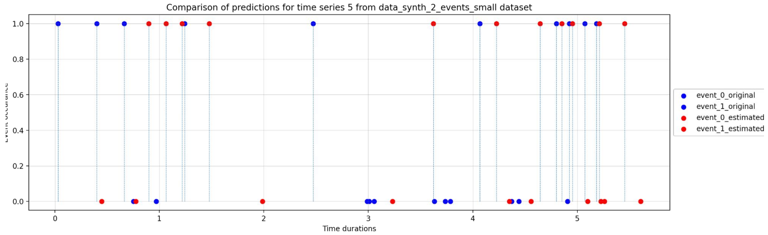
Results: Evaluation on synthetic data

Synth_2_events_small

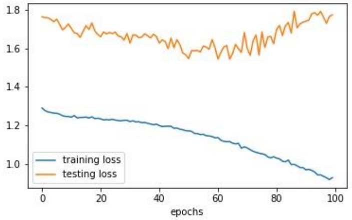
NeuralHawkes

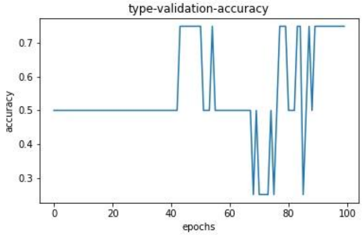


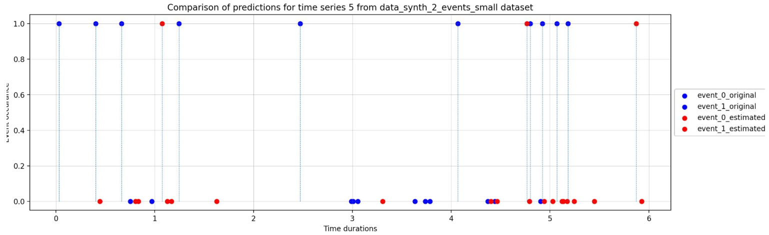




NeuralHawkesMasked



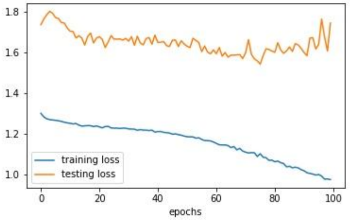


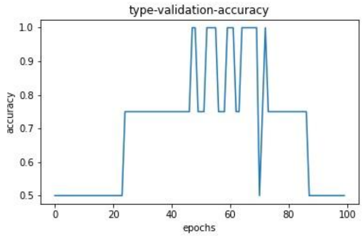


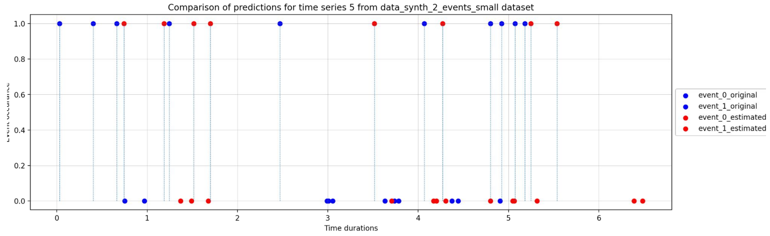
Results: Evaluation on synthetic data

Synth_2_events_small

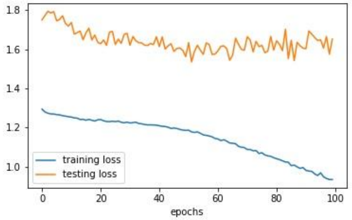
CausalNeuralHawkesMaskedWeighted

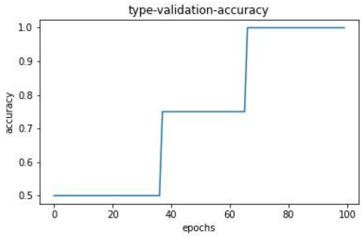


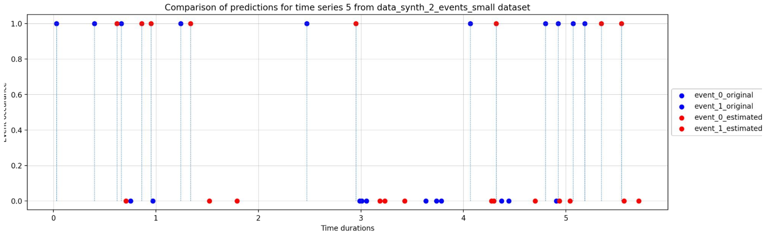




CausalNeuralHawkesTrainableWeighted



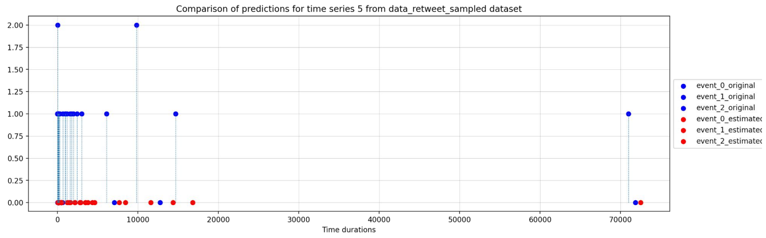




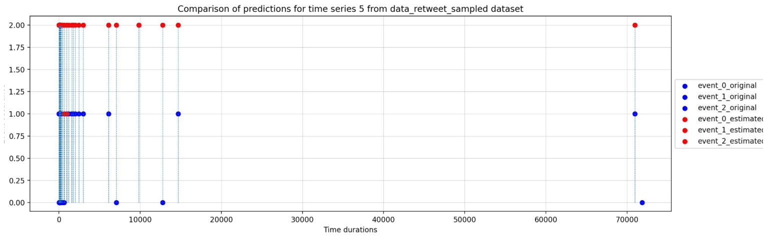
Results: Evaluation on retweet data

Skoltech
Skolkovo Institute of Science and Technology

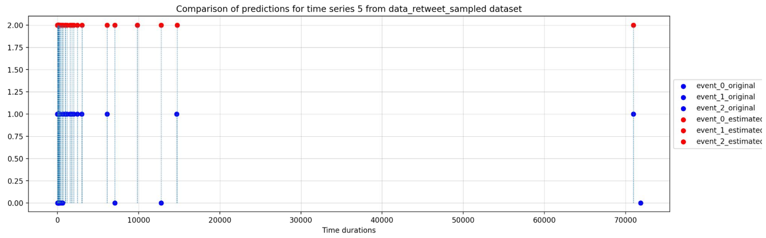
NeuralHawkes



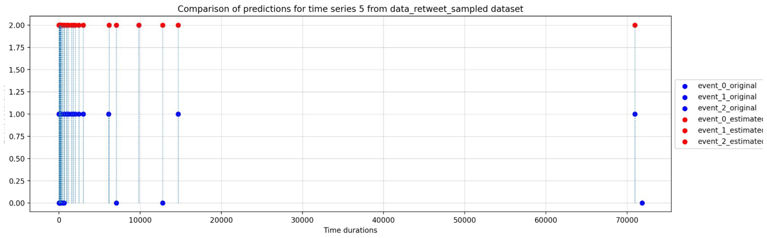
NeuralHawkesMasked



CausalNeuralHawkesMaskedWeighted



CausalNeuralHawkesTrainableWeighted



What we have done: summary

1. Successfully **implemented Causal-based Neural Hawkes** model
2. Experimented with **three versions** of Causal Neural Hawkes
3. Evaluated **models on synthetic dataset** to choose the best one
4. Created a **python package**
5. **Rewritten** the model to be **CUDA compatible**
6. Created **evaluation pipeline** and **plotting functional**

Conclusions

1. Our new model is outperforming the original neural hawkes data on synthetic datasets in terms of accuracy of prediction

2. Among new versions, the CausalNeuralHawkesTrainedWeighted is the best performing version

3. Our modified version is more stable

4. Computational intensity is still a problem of both model training and causality estimation

Skolkovo Institute of Science and Technology

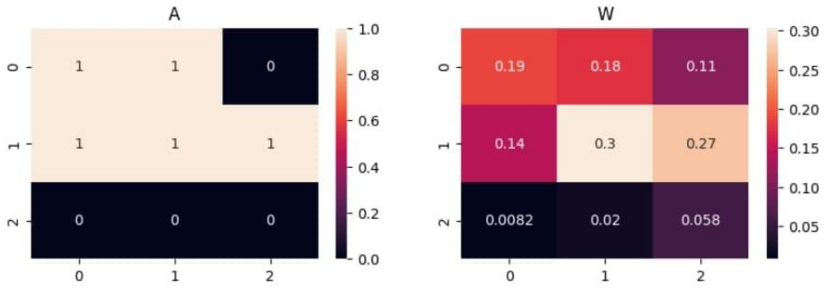
Nikita Paplavskii, Polina Pilyugina

Models of Sequential Data 2021

Results: Evaluation on retweet data

model	L	rmse	acc
NH	9,72	8 116,30	40,00%
CNHM	1 900,14	8 432,72	8,89%
CNHW	1 905,10	8 426,13	4,44%
CNHTW	1 901,95	8 432,60	4,44%

Models performance comparison on retweet datasets



Estimated granger causality graph and intensities

Neural Hawkes

Neural Hawkes model is Continious Time LSTM Model (CTLSTM).

It is based on classical LSTM model, but adjusted to work with continuous times.

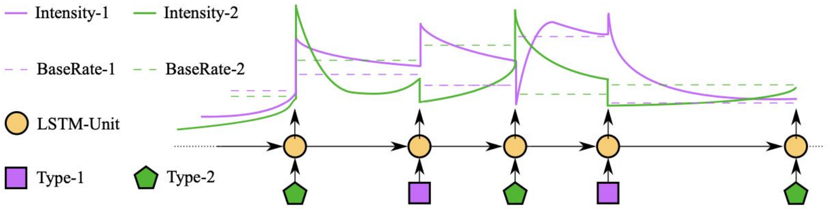


Figure 1: Drawing an event stream from a neural Hawkes process. An LSTM reads the sequence of past events (polygons) to arrive at a hidden state (orange). That state determines the future "intensities" of the two types of events—that is, their time-varying instantaneous probabilities. The intensity functions are continuous parametric curves (solid lines) determined by the most recent LSTM state, with dashed lines showing the steady-state asymptotes that they would eventually approach. In this example, events of type 1 excite type 1 but inhibit type 2. Type 2 excites itself, and excites or inhibits type 1 according to whether the count of type 2 events so far is odd or even. Those are immediate effects, shown by the sudden jumps in intensity. The events also have longer-timescale effects, shown by the shifts in the asymptotic dashed lines.

For the proposed models, the log-likelihood (1) of the parameters turns out to be given by a simple formula—the sum of the log-intensities of the events that happened, at the times they happened, minus an integral of the total intensities over the observation interval $[0, T]$:

$$\ell = \sum_{i:t_i \leq T} \log \lambda_{k_i}(t_i) - \underbrace{\int_{t=0}^{T} \lambda(t)dt}_{\text{call this } \Lambda} \tag{8}$$