# Learning Hawkes Processes from a Handful of Events


Farnood Salehi*, William Trouleau*, Matthias Grossglauser, Patrick Thiran

Information and Network Dynamics Group, School of Computer and Communication Sciences, EPFL

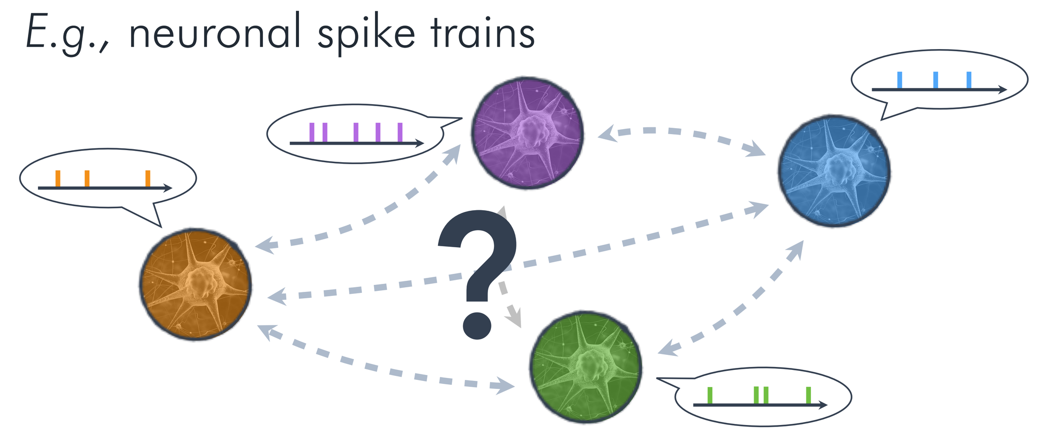
* The first two authors contributed equally to this work


Our code is available on Github!

github.com/trouleau/var-hawkes

## Motivation

- We want to learn the **causal relationships** in a network of discrete events (time series).
  *E.g.*, neuronal spike trains




- **Multivariate Hawkes Processes** (MHPs) are widely used to model mutually exciting patterns in discrete events.
- But experimental **data might be scarce**…

**Question:** How to identify the causal structure of the network **when only small data is available**?

## Model

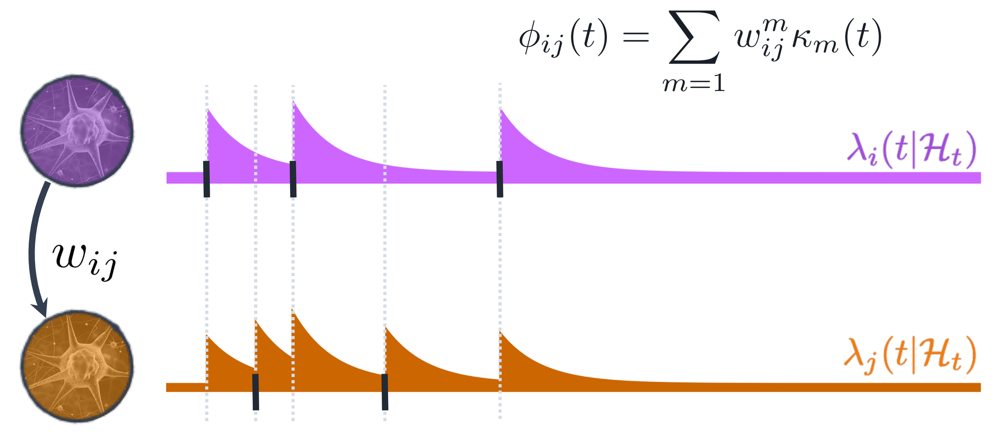
- A Multivariate Hawkes Process is a point process with intensity

$$\lambda_i(t) = \mu_i + \sum_{j=1}^{D} \int_0^t \phi_{ij}(t-\tau)dN_j(\tau).$$

Exogenous intensity: constant, independent of the past

Endogenous intensity: due to excitation from past events, with excitation kernel

$$\phi_{ij}(t) = \sum_{m=1}^{M} w_{ij}^m \kappa_m(t)$$




## Contribution

**1** We model the parameters of an MHP as latent variables, and we develop a probabilistic interpretation of existing maximum likelihood methods.

**2** We assume that each latent variable is sampled from its own prior with some unknown hyper-parameter.

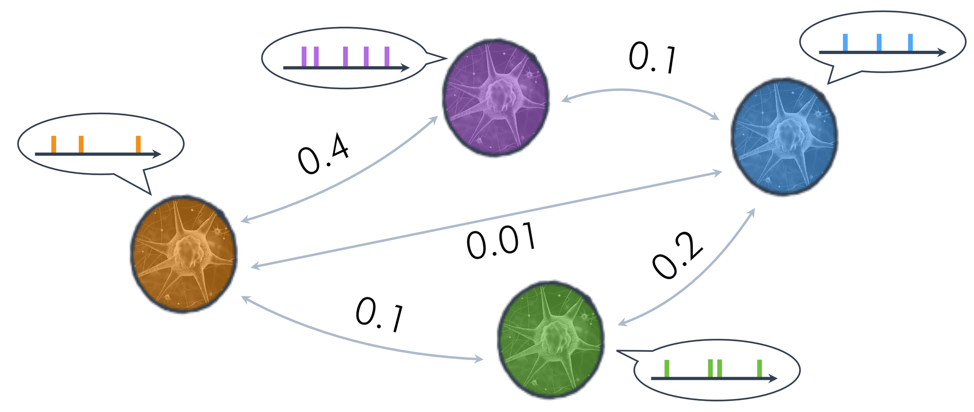
**3** We introduce, VI-MHP, a new Bayesian approach for learning the causal structure of an MHP in the absence of large volume of data. VI-MHP can optimize over thousands of hyper-parameters efficiently.

## Classic Maximum Likelihood Estimation Framework

- **Maximum Likelihood Estimation (MLE)**:

MLE estimates the parameters of the model by maximizing the log-likelihood of a sequence $\mathcal{S}$ w.r.t. the parameters with a regularization term $\mathcal{R}$

$$\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{W}} = \underset{\boldsymbol{\mu}\geq 0, \boldsymbol{W}\geq 0}{\operatorname{argmin}} \; -\log p(\mathcal{S}|\boldsymbol{\mu}, \boldsymbol{W}) + \frac{1}{\alpha}\mathcal{R}(\boldsymbol{\mu}, \boldsymbol{W}).$$



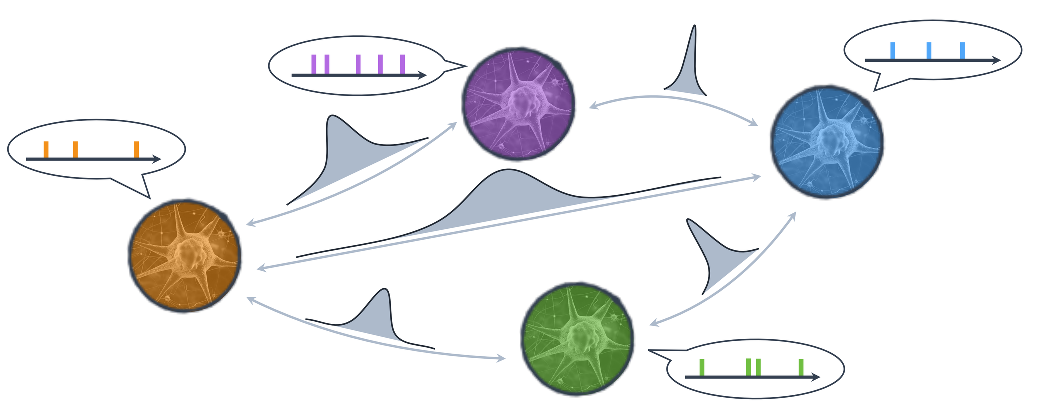

### Challenges

- **Small data** amplifies the **risk of overfitting**.
- It is desirable to control the effect of the penalty with **an independent hyper-parameter** for each of the $MD^2 + D$ **parameters**.
  But finding all of them with grid-search requires exponentially many grid points… ***Impractical!***

## Variational EM for Multivariate Hawkes Processes

- **Bayesian framework**:

Parameters are assumed to be random variables drawn from a prior $p_{\boldsymbol{\alpha}}(\boldsymbol{\mu}, \boldsymbol{W})$. The goal is to find the posterior $p_{\boldsymbol{\alpha}}(\boldsymbol{\mu}, \boldsymbol{W}|\mathcal{S})$ over the parameters.



### Challenges

- **Computing the posterior** is computationally intractable.
- Finding priors, i.e., **tuning hyper-parameters** $\boldsymbol{\alpha}$ with grid search is impossible for $\sim$1000 hyper-parameters.

### Solution

- Use **variational inference** to **estimate the posterior** and to find a lower bound on the marginal likelihood.
- Find the $MD^2 + D$ hyper-parameters by **maximizing the variational lower bound** on the marginal likelihood w.r.t. the hyper-parameters.

- **Variational Inference (VI)** approximates the posterior by a variational distribution $q_{\boldsymbol{\gamma}}(\boldsymbol{\mu}, \boldsymbol{W})$ parameterized by the variational parameters $\boldsymbol{\gamma}$.
- VI finds the best $\boldsymbol{\gamma}$ by minimizing the KL-divergence between the variational distribution $q_{\boldsymbol{\gamma}}(\boldsymbol{\mu}, \boldsymbol{W})$ and the posterior $p_{\boldsymbol{\alpha}}(\boldsymbol{\mu}, \boldsymbol{W}|\mathcal{S})$, which is equivalent to maximizing the ELBO

$$\text{ELBO}(q_{\boldsymbol{\gamma}}, \boldsymbol{\alpha}) = \mathbb{E}_{q_{\boldsymbol{\gamma}}}[\log p_{\boldsymbol{\alpha}}(\boldsymbol{\mu}, \boldsymbol{W}, \mathcal{S})] - \mathbb{E}_{q_{\boldsymbol{\gamma}}}[\log q_{\boldsymbol{\gamma}}(\boldsymbol{\mu}, \boldsymbol{W})].$$

The ELBO is a lower bound on the marginal likelihood

$$\text{ELBO}(q_{\boldsymbol{\gamma}}, \boldsymbol{\alpha}) \leq \log p_{\boldsymbol{\alpha}}(\mathcal{S}).$$

- **We maximize the ELBO over both $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ to increase the marginal likelihood of the data.**
- Since the parameters are non-negative, we use a **log-normal distribution** as the variational distribution to approximate the posterior.
- By interpreting regularization terms as unnormalized priors, we can choose priors that retain the desired properties of common penalties.

### Variational EM Algorithm

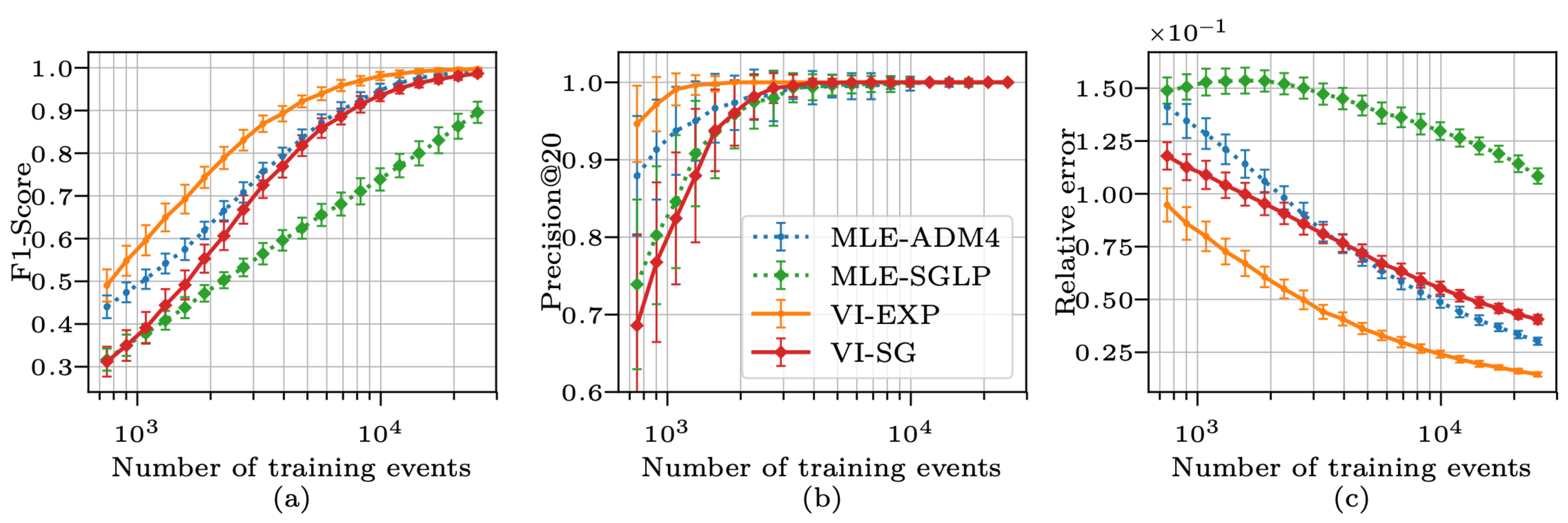
**until** convergence **do**
- **E-step:** Maximize the ELBO w.r.t. $\boldsymbol{\gamma}$.
- **M-step:** Maximize the ELBO w.r.t. $\boldsymbol{\alpha}$ with a closed-form solution (which depends on the choice of prior).

**done**

**Output:** The variational parameters $\boldsymbol{\gamma}$.

## Experimental Results

- Experiments on Synthetic data




Performance measured with respect to the number of training samples (averaged over 10 simulated observations over 30 random graphs, with D=50 fixed).

- Experiments on real dataset

Predictive log-likelihood for the models learned on several datasets with different number of events and dimentions.

| Dataset | Statistics | | Averaged predictive log-likelihood | | | |
|---|---|---|---|---|---|---|
| | #dim ($D$) | #events ($N$) | **VI-SG** | MLE-SGLP | **VI-EXP** | MLE-ADM4 |
| Epidemics | 54 | 5 349 | **−2,06** | −3,03 | −4,31 | −4,61 |
| Stock market | 12 | 7 089 | **−1,00** | −2,45 | −2,82 | −2,81 |
| Enron email | 143 | 74 294 | −0,42 | −1,01 | **−0,23** | −0,40 |


For more information, check out our paper:
F. Salehi, W. Trouleau, M. Grossglauser, P. Thiran. **Learning Hawkes Processes from a Handful of Events**.
Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019).