Fast Estimation of Causal Interactions using Wold Processes

Flavio Figueiredo, Guilherme Borges, Pedro O.S., Vaz de Melo, Renato Assunção
Universidade Federal de Minas Gerais (UFMG) (NeurIPS 2018)

A Variational Inference Approach to Learning Multivariate Wold Processes

Jalal Etesami *, William Trouleau*, Negar Kiyavash, Matthias Grossglauser, Patrick Thiran
École Polytechnique Fédérale de Lausanne (EPFL) (AISTATS 2021)

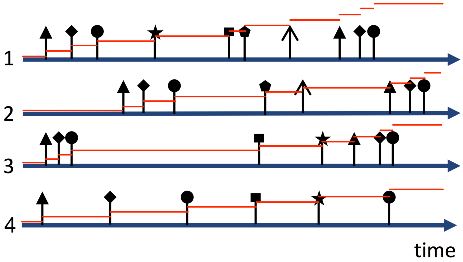
Presenter: Qingmei Wang

March.3rd, 2022

- Background
- GRANGER-BUSCA
- Variational Inference Approach
- Experiments
- Summary

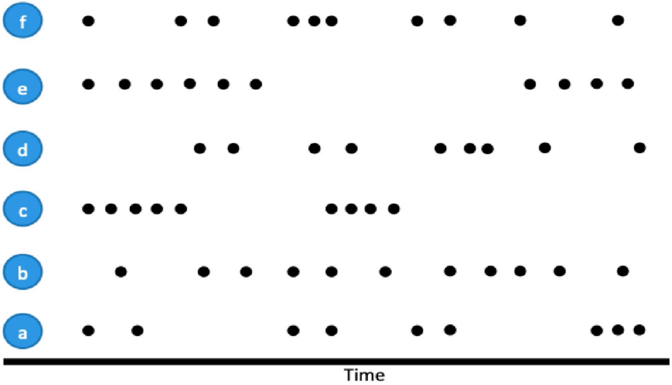
Event Sequences and Temporal Point Processes

Event sequence: $\{(t_i, c_i)\}_{i=1}^{I}$, $c_i \in \mathcal{C}$, or **Counting process:** $N(t) = \{N_c(t)\}_{c \in \mathcal{C}}$.



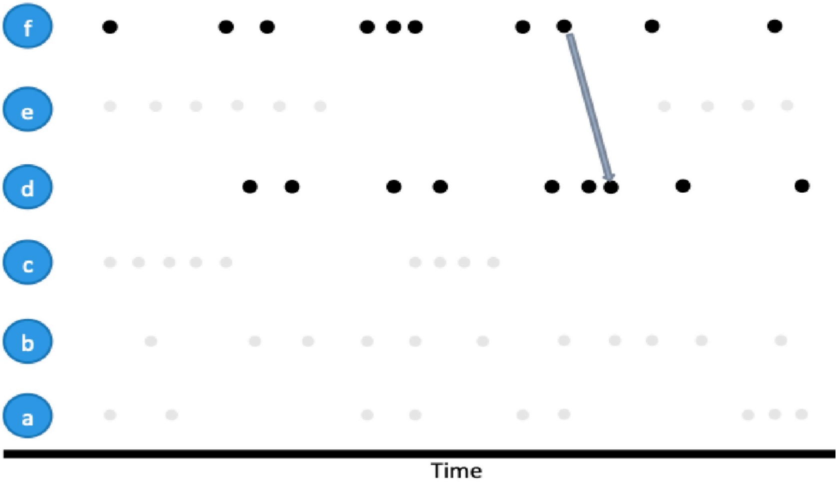
Where $t_i \in [0, T]$ mean timestamps and $c_i \in \mathcal{C} = \{1, \ldots, C\}$ mean event types.

Each entity is viewed as a point process



- Each observation is a timestamp

Notice how f and d are related



- Eventst tends to precede the other

How do we capture this relation?

- Hawkes Processes

Hawkes Process $\mathrm{HP}_{\mathcal{C}}(\boldsymbol{\mu}, \boldsymbol{\Phi})$ models the triggering pattern between different events:

$$\lambda_c(t) = \underbrace{\mu_c}_{\text{base intensity}} + \sum_{(t_i,c_i)\in\mathcal{H}_t} \underbrace{\phi_{cc_i}(t - t_i)}_{\text{impact function}} \tag{1}$$

- $\boldsymbol{\mu} = [\mu_c]$: exogenous fluctuation of the system.
- $\boldsymbol{\Phi} = [\phi_{cc'}(t)]$: endogenous triggering pattern of type-c' on type-c.

Hawkes Processes

- $\phi_{cc}(\cdot)$: the **self**-triggering pattern.
- $\phi_{cc'}(\cdot), c \neq c'$: the **mutually**-triggering pattern.

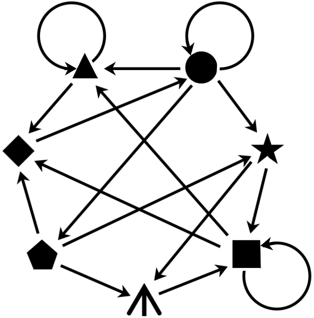
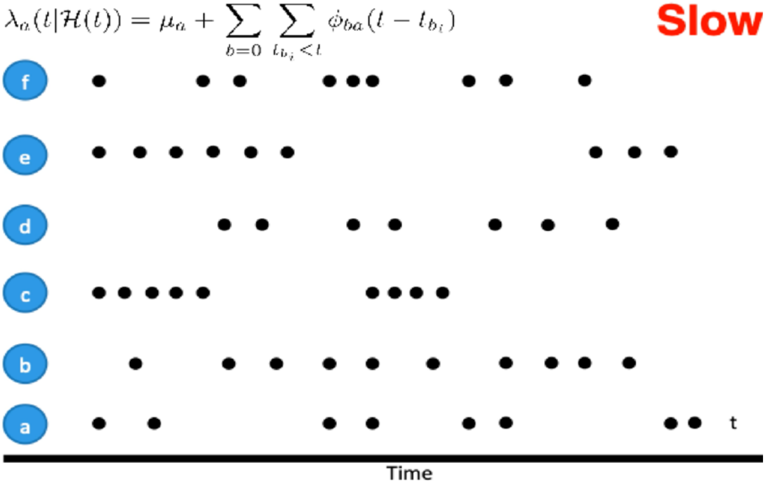


Figure 1: Granger causality

Long memory

$$\lambda_a(t|\mathcal{H}(t)) = \mu_a + \sum_{b=0}^{K-1} \sum_{t_{b_i} < t} \phi_{ba}(t - t_{b_i})$$



Wold Processes

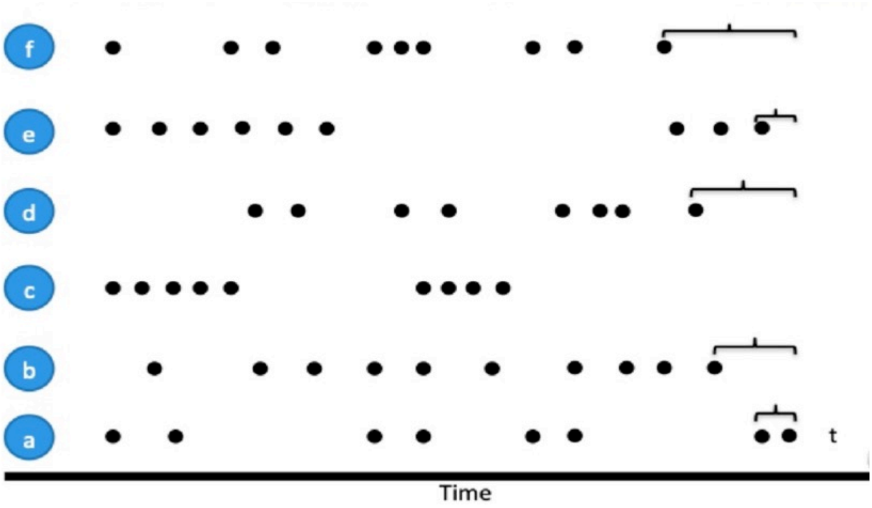
- Different from Hawkes processes, whose intensity function depends on the whole history of previous events, the probability distribution of the *i*-th inter-event time δ_i depends only on the previous inter-event time δ_{i-1}

$$\lambda_a(t \mid \mathcal{H}(t)) = \mu_a + \sum_{b=0}^{K-1} \alpha_{ba}\omega_{ba}(t)$$

Depends on the last increment only

$$\omega_{ba}(t) = \frac{1}{\beta_b + \Delta_{ba}(t)}$$

Fast!



- Background
- GRANGER-BUSCA
- Variational Inference Approach
- Experiments
- Summary

What does Granger Busca look like?

- Busca is another point process model based on Wold processes and it is GRANGER-BUSCA's starting point

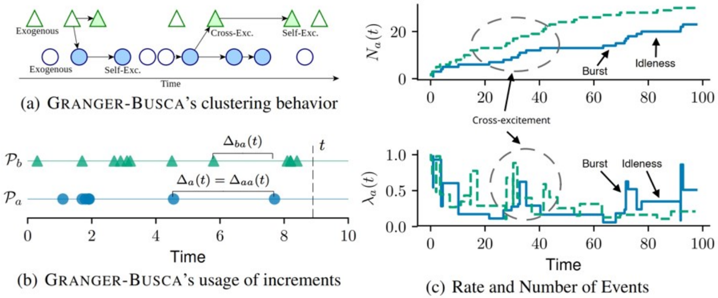


Figure 1: GRANGER-BUSCA at work. Plot (a) shows the events of process $\mathcal{P}_a$ (circles) and process $\mathcal{P}_b$ (triangles). The arrows show the excitement component of the model. Plot (b) illustrates how $\Delta_{aa}(t)$ and $\Delta_{ba}(t)$ are calculated. Plot (c) shows the cumulative random processes $N_a(t)$ and $N_b(t)$ in the top, while the bottom plot shows the random conditional intensity functions $\lambda_a(t)$ and $\lambda_b(t)$.

Goal in this work

- To extract Granger causality from multivariate point process data only
- To develop learning algorithms that are asymptotically fast

Formalize GRANGER-BUSCA

- GRANGER-BUSCA's multivariate conditional intensity function

$$\lambda_a(t) = \underbrace{\mu_a}_{\text{Exogenous Poisson Rate}} + \underbrace{\sum_{b=0}^{K-1} \frac{\alpha_{ba}}{\beta_b + \Delta_{ba}(t)}}_{\text{Endogenous Wold Rate}} \tag{2}$$

Learning GRANGER-BUSCA

- Developed Markov Chain Monte Carlo (MCMC) sampling algorithm to learn GRANGER-BUSCA from data
- Fixed $\beta = 1$ to simplify the learning strategy
- Sample a latent variable, z_{a_i}, which takes a value of $b \in [0, K-1]$when process $\mathcal{P}_a$ influences t_{a_i}. When the stamp is exogenous, set this value to a constant K.
- Learned GRANGER-BUSCA with an Expectation Maximization approach. Hidden labels and the matrix $\mathbf{G}$are estimated in the Expectation step. With the labels, $\boldsymbol{\mu}$ estimated in the maximization step

How to update the z_{a_i} lables

- Given any event at

$$\Pr\left[t_{a_i} \in \text{EXOG.}\right] = \frac{\mu_a}{\mu_a + \sum_{b'=0}^{K-1} \lambda_{b'a}\left(t_{a_i}\right)} \tag{3}$$

$$\Pr\left[t_{a_i} \leftarrow \mathcal{P}_b\right] = \frac{\lambda_{ba}\left(t_{a_i}\right)}{\mu_a + \sum_{b'=0}^{K-1} \lambda_{b'a}\left(t_{a_i}\right)} \tag{4}$$

  - Selected the inducing process based on the conditional probability

$$\Pr\left[t_{a_i} \leftarrow \mathcal{P}_b \mid t_{a_i} \notin \text{ EXOG. }\right] = \frac{\lambda_{ba}\left(t_{a_i}\right)}{\sum_{b'=0}^{K-1} \lambda_{b'a}\left(t_{a_i}\right)} \tag{5}$$

Learning GRANGER-BUSCA

- Sample the hidden labels z_{a_i} as follows

1. For each process $\mathcal{P}_a$
   (a) Sample row a from $\boldsymbol{G}$ as $\sim Dirichlet(\alpha_p)$
2. For each process $\mathcal{P}_a$
   (a) For each observation $t_{a_i} \in \mathcal{P}_a$
      i. Sample $p \sim Uniform(0, 1)$
         A. When $p < e^{-\mu_a(t_{a_i} - t_{\mu_a})}$
            $z_{a_i} \leftarrow$ exogeneous
         B. Otherwise
            Sample $z_{a_i} \sim Multinomial(Eq\ 5)$

- Background
- GRANGER-BUSCA
- Variational Inference Approach
- Experiments
- Summary

Relax all restrictive assumptions

- GRANGER-BUSCA assumes that $\sum_{k=1}^{K} \alpha_{k',k} = 1$ and $\beta_{k',k} = \beta_k$ for all $k' \in [K]$.
- Variational inference approach targeted to learn the set of parameters

$$\begin{aligned} \boldsymbol{\mu} &:= \{\mu_k : k \in [K]\}, \\ \boldsymbol{\alpha} &:= \{\alpha_{k',k} : k', k \in [K]\}, \\ \text{and} \quad \boldsymbol{\beta} &:= \{\beta_{k',k} : k', k \in [K]\}. \end{aligned}$$

Variational Inference Approach

- A method for approximating the posterior distribution over the model parameters given the observations

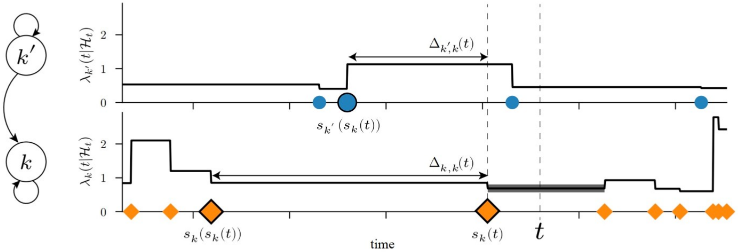


Figure 1: Illustration of the Wold process dynamics on a simple toy example with 2 processes, where process k is influenced by process k' and by itself, *i.e.*, $\alpha_{k',k} > 0$ and $\alpha_{k,k} > 0$, and process k' also influences itself. At the highlighted time t, the intensity in process k depends on the two highlighted inter-event times $\Delta_{k',k}(t)$ and $\Delta_{k,k}(t)$, which remain constant until the next event in process k.

Maximum Likelihood Estimation

$$\log p(\mathcal{P} \mid \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\mu}) = \sum_k \sum_{t_{k,i} \in \mathcal{P}_k} \log \lambda_k \left(t_{k,i} \mid \mathcal{H}_t\right) - \sum_k \int_0^T \lambda_k \left(t \mid \mathcal{H}_t\right) dt \tag{6}$$

- The specific form of Wold process defined makes the log-likelihood function non-convex with respect to β

$$\lambda_k \left(t \mid \mathcal{H}_t\right) = \mu_k + \sum_{k'=1}^{K} \frac{\alpha_{k',k}}{\beta_{k',k} + \Delta_{k',k}(t)} \tag{7}$$

- Moreover, maximum-likelihood estimation of point processes typically scales poorly to high dimensional settings
- Used a variational inference approach to circumvent both issues of non-convexity and scalability.

Variational Inference

- Variational inference (VI) is a method for approximating the posterior distribution over the model parameters given the observations
- Defined an auxiliary variable $\mathbf{z}_{k,i}$ for each event $t_{k,i}$ to be a one-hot vector that indicates the cause of that event
- $\mathbf{z}_{k,i} = \left[z_{k,i}^{(0)}, z_{k,i}^{(1)}, \cdots, z_{k,i}^{(K)} \right]$
- Approximate the posterior distribution $p(\boldsymbol{\mu}, \mathbf{z}, \boldsymbol{\alpha}, \boldsymbol{\beta} \mid \mathcal{P})$ with a variational distribution $q(\boldsymbol{\mu}, \mathbf{z}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ that minimizes the KL-divergence between p and q.

Variational Inference

- VI solves for the optimal variational distribution that minimizes the KL-divergence, or equivalently it maximizes the evidence lower bound (ELBO), given by

$$\text{ELBO}(q) = \mathbb{E}_q[\log p(\boldsymbol{\mu}, \mathbf{z}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathcal{P})] \quad - \mathbb{E}_q[\log q(\boldsymbol{\mu}, \mathbf{z}, \boldsymbol{\alpha}, \boldsymbol{\beta})]. \tag{8}$$

Variational Inference

- Considered a mean-field approximation for the variational distribution. In such an approximation, the variational parameters are assumed to be independent. Therefore

$$
\begin{aligned}
q(\boldsymbol{\mu}, \mathbf{z}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = & \prod_{k=1}^{K} q\left(\mu_k\right) \times \prod_{k=1}^{K} \prod_{i=1}^{|\mathcal{P}_k|} q\left(\mathbf{z}_{k,i}\right) \\
& \times \prod_{k=1}^{K} \prod_{k'=1}^{K} q\left(\alpha_{k',k}\right) q\left(\beta_{k',k}\right)
\end{aligned} \tag{9}
$$

- Using this approximation and coordinate ascent for maximizing (3), we obtain the variational distributions $\left\{q\left(\mu_k\right), q\left(\mathbf{z}_{k,i}\right), q\left(\alpha_{k',k}\right), q\left(\beta_{k',k}\right)\right\}$ by selecting appropriate prior distributions over the parameters

Variational Inference

Variational update of the auxiliary parent variable $\mathbf{z}_{k,i}$. The definition of the auxiliary variable $\mathbf{z}_{k,i}$ implies that $\sum_{k'=0}^{K} z_{k,i}^{(k')} = 1$. As shown in Appendix B, this results in

$$q(\mathbf{z}_{k,i}) = \text{Categorical}\left(K+1; p_{k,i}^{(0)}, ..., p_{k,i}^{(K)}\right), \qquad (5)$$

where the probabilities

$$\begin{aligned} p_{k,i}^{(0)} &\propto \exp\left(\mathbb{E}_{q(\mu_k)}[\log \mu_k]\right) \\ \text{and } p_{k,i}^{(k')} &\propto \exp\big(\mathbb{E}_{q(\alpha_{k',k})}[\log(\alpha_{k',k})] \\ &\quad - \mathbb{E}_{q(\beta_{k',k})}[\log\left(\beta_{k',k} + \Delta_{k',k}(t_{k,i})\right)]\big), \\ &\quad \forall k' \in [K] \end{aligned}$$

are normalized such that $\sum_{k'=0}^{K} p_{k,i}^{(k')} = 1$. In the above equations, the expectations are over the variational distributions.

Variational Inference

Variational update of $\beta_{k',k}$. For this parameter, we select the prior distribution to be Inverse-Gamma with shape $\phi_{k',k}$ and scale $\psi_{k',k}$. This choice of prior results in a variational distribution of $\beta_{k',k}$ proportional to

$$(\beta_{k',k})^{-\phi_{k',k}-1} e^{\left(-\frac{\psi_{k',k}}{\beta_{k',k}}\right)} \prod_{i=1}^{|\mathcal{P}_k|} \Bigg[(\beta_{k',k} + \Delta_{k',k}(t_{k,i}))^{-\mathbb{E}\left[z_{k,i}^{(k')}\right]} \exp\left(- \frac{\mathbb{E}[\alpha_{k',k}](t_{k,i} - t_{k,i-1})}{\beta_{k',k} + \Delta_{k',k}(t_{k,i})} \right) \Bigg]. \tag{8}$$

Variational update of μ_k. Similar to α, we use the Gamma distribution as the prior of μ_k with shape c_k and rate d_k resulting in the posterior

$$q(\mu_k) = \mathrm{Gamma}\left(C_k; D_k\right), \tag{7}$$

where

$$C_k := c_k + \sum_{i=1}^{|\mathcal{P}_k|} \mathbb{E}_{q(z_{k,i}^{(0)})}[z_{k,i}^{(0)}],$$

$$D_k := d_k + \sum_{i=1}^{|\mathcal{P}_k|} (t_{k,i} - t_{k,i-1}).$$

- Background
- GRANGER-BUSCA
- Variational Inference Approach
- Experiments
- Summary

Accuracy

- Precision @ n score
  - Retrieve top neighbors per node

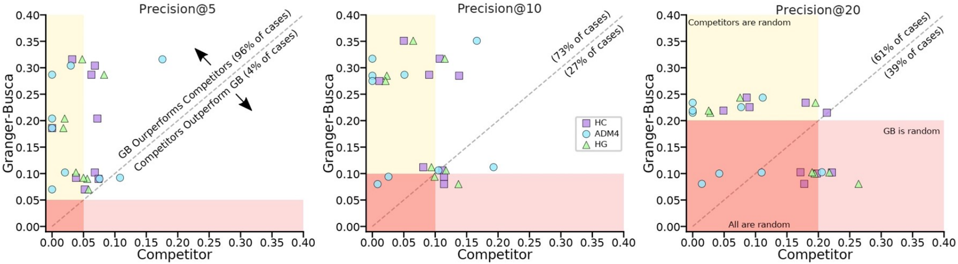


Figure 2: Precision Scores for the Top-100 datasets.

Full Datasets

Table 1: Datasets used for Experiments and Precision Scores for Full Datasets. Due to their sizes, only GRANGER-BUSCA is able to execute in all datasets. To allow comparisons, we execute baselines methods with only the Top-100 destination nodes. Other results are presented in Table 2 and Figure 2.

	# Proc (K)	# Obs. (N)	N (Top-100)	Span	%NZ	P@5	P@10	P@20	TT(s)
bitcoinalpha [28]	3,257	23,399	2,279	5Y	0.2%	0.26	0.14	0.07	3
bitcoinotc [28]	4,791	33,766	2,328	5Y	0.1%	0.25	0.14	0.07	7
college-msg [39]	1,313	58,486	10,869	193D	1.1%	0.36	0.30	0.19	1
email [31, 50]	803	327,677	92,924	803D	3.74%	0.23	0.28	0.32	4
sx-askubuntu [40]	88,549	879,121	58,142	7Y	0.006%	0.25	0.13	0.06	2774
sx-mathoverflow [40]	16,936	488,984	59,602	7Y	0.07%	0.28	0.16	0.09	98
sx-superuser [40]	114,623	1,360,974	64,866	7Y	0.006%	0.26	0.14	0.07	4614
wikitalk [30, 40]	251,154	7,833,140	211,344	6Y	0.003%	0.25	0.14	0.07	27540
memetracker-100 [29]	100	24,665,418	-	9M	9.85%	0.30	0.29	0.22	114
memetracker-500 [29]	500	39,318,989	-	9M	4.44%	0.30	0.30	0.23	274

Memetracker

Table 2: Comparing GRANGER-BUSCA (GB) with Hawkes-Cumulants (HC) Memetracker.

	Precision@5		**Precision@10**		**Precision@20**		**Kendall**		**Rel. Error**		**TT(s)**	
	HC	GB	HC	GB	HC	GB	HC	GB	HC	GB	HC	GB
top-100	0.06	**0.30**	0.09	**0.29**	0.01	**0.22**	0.05	**0.26**	1.0	**0.44**	**87**	114
top-500	0.01	**0.30**	0.01	**0.30**	0.02	**0.23**	0.08	**0.20**	1.8	**0.06**	715	**274**

- Background
- GRANGER-BUSCA
- Variational Inference Approach
- Experiments
- Summary

Summary

Fast Estimation

- By using a Wold Process
  - We can evaluate the intensity in linear time
  - Fast data structures for estimation
- Hawkes Processes
  - Usually quadratic both on time and processes