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)

Abstract

Temporal point-processes are often used for mathematical modeling of sequences of discrete events with asynchronous timestamps. We focus on a class of temporal point-process models called multivariate Wold processes (MWP). These processes are well suited to model real-world communication dynamics. Statistical inference on such processes often requires learning their corresponding parameters using a set of observed timestamps. In this work, we relax some of the restrictive modeling assumptions made in the state-of-the-art and introduce a Bayesian approach for inferring the parameters of MWP. We develop a computationally efficient variational inference algorithm that allows scaling up the approach to high-dimensional processes and long sequences of observations. Our experimental results on both synthetic and real-world datasets show that our proposed algorithm outperforms existing methods.

1 INTRODUCTION

Multivariate point-processes provide powerful tools to gain insight on the behavior of complex systems such as social networks (Blundell et al., 2012), networks of neurons (Monti et al., 2014), financial markets (Namaki et al., 2011), and television records (Xu et al., 2016).

Wold processes (Daley and Jones, 2003; Wold, 1948), akin to Hawkes processes (Hawkes, 1971), are a type of multivariate point-process that are well suited for modeling discrete events. They are defined in terms of a Markovian joint distribution of *inter-event times*. Specifically, the times between consecutive events t_{i-1} and t_i, also called inter-event times $\delta_i := t_i - t_{i-1}$, form a Markov chain such that the distribution

$$p(\delta_i|\delta_{i-1}, \delta_{i-2}, ..., \delta_1) = p(\delta_i|\delta_{i-1}).$$

Wold processes are suitable for modeling the dynamics of complex systems, and their inherent Markovian property facilitates the learning task (Vaz de Melo et al., 2013, 2015). The interactions among the processes of a multivariate Wold process (MWP) can be visualized using a directed graph in which nodes and edges represent processes and direct influences, respectively.

Recently, Vaz de Melo et al. (2015) showed that Wold processes can model the dynamics of real-world communications more faithfully than the widely used Hawkes processes. Figueiredo et al. (2018) then developed a Markov Chain Monte Carlo (MCMC) sampling-based algorithm, called Granger-Busca, to infer the parameters of a MWP. The choice of prior of the MCMC algorithm in (Figueiredo et al., 2018) required certain restrictive assumptions on the network. For instance, it required that every node in the underlying network of MWP has at least one out-going edge, *i.e.*, at least one child. Clearly, many practical systems violate this assumption. Beside relaxing the limiting assumptions in (Figueiredo et al., 2018), we propose an efficient Bayesian algorithm for learning a general class of MWPs. To achieve scalability, we propose a variational inference (VI) approach to approximate the high-dimensional posterior of the model parameters given the data.

2 RELATED WORKS

The inference problem for multivariate point-process has been mostly studied for Hawkes processes. The main approaches for estimating the parameters of Hawkes processes are of two flavors. Maximum likelihood-based approaches estimate the parameters from the likelihood of observations (Zhou et al., 2013;

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

*Equal contribution.

Yang et al., 2017; Trouleau et al., 2019), while cumulant-based approaches learn the parameters of interest by solving a set of equations obtained from various order statistics of the Hawkes process (Hawkes, 1971; Bacry and Muzy, 2014; Achab et al., 2017). Some studies address the problem from a Bayesian perspective; *e.g.*, in (Linderman and Adams, 2014), the authors develop an MCMC sampling-based algorithm. Due to the long memory of Hawkes processes, the method does not scale well with the number of observations. To improve the scalibility of such Bayesian methods, Linderman and Adams (2015) proposed approximating continuous-time Hawkes processes with a discrete-time formulation. This led to a computationally efficient stochastic variational inference (VI) algorithm that scales well for longer sequences of observations. More recently, Salehi et al. (2019) proposed a black-box VI approach that can learn the parameters of MHPs without discretizing the process. Analogous to (Linderman and Adams, 2015), the Bayesian approach that we propose here uses mean-field VI to learn the parameters of MWPs.

Recall that Wold processes are defined through a Markovian transition probability distribution on the inter-event times, *i.e.*, $p(\delta_{i+1}|\delta_i)$, which measures the probability of the next inter-event time δ_{i+1}, given the preceding one. It turns out that for general Markovian transition probabilities, this model is analytically intractable (Guttorp and Thorarinsdottir, 2012). However, in the univariate setting, when the transition probabilities have the exponential form $p(\delta_{i+1}|\delta_i) = f(\delta_i)\exp\left(-f(\delta_i)\delta_{i+1}\right)$, the process shows interesting properties (Cox, 1955; Daley, 1982; Daley and Jones, 2003). In particular, the next inter-event time δ_{i+1} is then exponentially distributed with rate $f(\delta_i)$. In the case where $f(\delta_i) = \lambda\delta_i^{-1/2}$, the stationary distribution of inter-event times can also be found via Mellin transforms (Wold, 1948). Similarly, in the case where $f(\delta_i) = \beta + \alpha\delta_i$, the stationary distribution $p(\delta_i)$ has the form $(\beta + \alpha\delta_i)^{-1}\exp\left(-\beta\delta_i\right)$. Analytical properties of a specific type of MWP that is an infinite process defined on the unit circle are discussed in Isham (1977).

Recent efforts consider variations of the exponential Wold process (Vaz de Melo et al., 2013; Alves et al., 2016). Instead of defining the Wold process in terms of its inter-event exponential rate, these works define the process in terms of the conditional mean of an exponentially distributed random variable $\mathbb{E}[\delta_i|\delta_{i-1}] = \beta + \alpha\delta_{i-1}$. This class of point processes is called a *self-feeding process*. This form of Wold process is able to capture both exponential and power-law behavior, which often occur simultaneously in real data. Realizations of this process tend to generate bursts of intense activity, followed by long periods of silence. Vaz de Melo et al. (2015) use self-feeding processes to model the time intervals between communication events for different technologies and means of communications, including short-message services (SMSs), mobile phone-calls, and e-mail transactions. Building on this work, Figueiredo et al. (2018) proposed a multivariate version of the self-feeding process and developed an MCMC sampling-based algorithm to learn the parameters. However, the approach requires restrictive structural assumptions on the network of the process, which limits the applicability of the model.

3 MODEL

In this section, we describe the model and the notation used throughout the paper. We first define the univariate Wold process, and then generalize it to the multivariate case.

A temporal point-process $\mathcal{P}$ is a probability model for a collection of times $\{0 \leq t_0 < t_1 < t_2 < \cdots\}$ that index the occurrences of random asynchronous events. Let $N(a, b]$ denote the random number of events of the process $\mathcal{P}$ in the interval $(a, b]$, and let $\mathcal{H}_t$ denote the history of the process $\mathcal{P}$ up to, but not including, time t. The distribution of a point process is characterized by its *conditional intensity function* and is defined as

$$\begin{aligned}\lambda(t|\mathcal{H}_t)dt &= p(N(t, t+dt] > 0 \mid \mathcal{H}_t) \\ &= \mathbb{E}[N(t, t+dt] \mid \mathcal{H}_t].\end{aligned}$$

Let $\mathcal{D} \triangleq \{\delta_i = t_i - t_{i-1}\}_{i\geq 1}$ denote the sequence of inter-event times. $\mathcal{P}$ is called a Wold process if the distribution over the inter-events is Markovian, *i.e.*,

$$p(\delta_{i+1}|\delta_i, \delta_{i-1}, \cdots, \delta_1) = p(\delta_{i+1}|\delta_i).$$

The form of the transition probability specifies the class of Wold process. For instance, in this work, we consider the self-feeding process formulation where transition probabilities have the exponential form given by $p(\delta_{i+1}|\delta_i) = f(\delta_i)\exp(-f(\delta_i)\delta_{i+1})$. In addition, we consider $f(\delta_i)$ to be $1/(\beta + \delta_i)$ so that the conditional mean is linear (Vaz de Melo et al., 2015; Alves et al., 2016).

Now, to define the MWP, consider a set of Wold processes $\mathcal{P} = \bigcup_{k=1}^{K} \mathcal{P}_k$ that are observed simultaneously, where $\mathcal{P}_k = \{t_{k,0} < t_{k,1} < \ldots\}$ and $t_{k,i}$ denotes the i-th event in the k-th process (also called dimension). At a given time t, the conditional intensity of the k-th process depends on the last inter-event times $\{\Delta_{k',k}(t) : k' \in [K]\}$, where $[K] \coloneqq \{1, ..., K\}$ and

$$\Delta_{k',k}(t) \coloneqq s_k(t) - s_{k'}(s_k(t)).$$

In this definition, $s_k(t)$ is the last event time of process k before time t, *i.e.*, $s_k(t) \coloneqq \max\{t_{k,i} : t_{k,i} < t\}$,

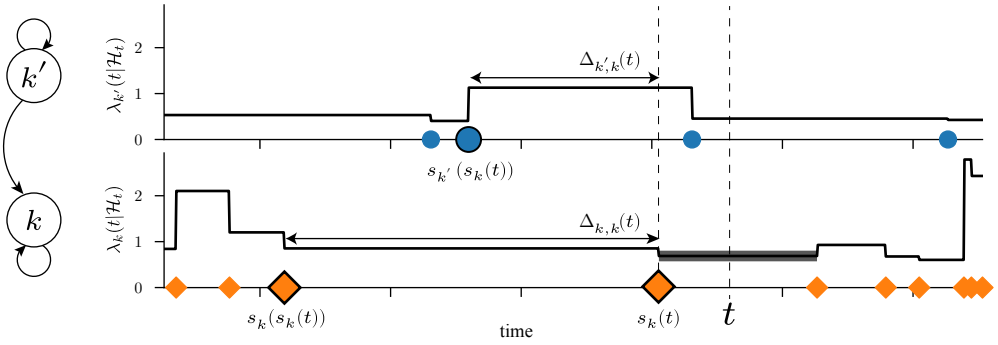


Figure 1: Illustration of the Wold process dynamics on a simple toy example with 2 processes, where process k is influenced by process k' and by itself, *i.e.*, $\alpha_{k',k} > 0$ and $\alpha_{k,k} > 0$, and process k' also influences itself. At the highlighted time t, the intensity in process k depends 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.

and $s_{k'}(s_k(t))$ is the last event of process k' preceding the event $s_k(t)$, *i.e.*, $s_{k'}(s_k(t)) \coloneqq \max\{t_{k',j} : t_{k',j} < s_k(t) < t\}$. An illustration of the process is shown in Figure 1. The conditional intensity function of the k-th process is then

$$\lambda_k(t|\mathcal{H}_t) = \mu_k + \sum_{k'=1}^{K} \frac{\alpha_{k',k}}{\beta_{k',k} + \Delta_{k',k}(t)}, \qquad (1)$$

where $\mu_k \geq 0$ is its background rate, and the influence of process k' on process k at time t is captured by $\alpha_{k',k}/(\beta_{k',k} + \Delta_{k',k}(t))$. The parameter $\alpha_{k',k} \geq 0$ is the weight of the influence and $\beta_{k',k} > 0$ ensures the stability of the process, *i.e.*, that the expected number of events stays finite in a finite time horizon (Daley and Jones, 2003).

Unlike the Hawkes process, the Wold process has finite memory because of its Markov property. In addition, because $\Delta_{k',k}(t)$ changes only when there is an event in dimension k, a given process k in a MWP is influenced by other processes (including itself) only when an event occurs in process k, as illustrated in Figure 1.

In Hawkes processes, the structure of the causal network is encoded in the support of the *excitation matrix* (Eichler et al., 2017; Etesami et al., 2016). Therefore, learning the support of the excitation matrix is sufficient for recovering the network structure. Analogously, one can gather the influences among dimensions of a MWP in a matrix called the *influence matrix* $\mathbf{G}(t) = [\alpha_{k',k}/(\beta_{k',k} + \Delta_{k',k}(t))]_{k',k=1}^{K}$. The main reason for this name is that the influence matrix captures the Granger causality in the network of the MWP. Specifically, the support of $\mathbf{G}(t)$ is the adjacency matrix of the corresponding Granger-causal network. Granger (1969) introduced a notion of causal relationships in a network of time series, which states that process X Granger-causes another process Y, *i.e.*, $X \to Y$, if the past information of X can provide statistically significant information about the future of Y. Based on this definition, a point process k' influences another point process k if $\lambda_k(t|\mathcal{H}_t) \neq \lambda_k(t|\mathcal{H}_t \setminus \mathcal{H}_t^{k'})$, where $\mathcal{H}_t^{k'}$ denotes the history of process k' up to time t. This condition in the network of Wold processes is equivalent to $[\mathbf{G}(t)]_{k',k} \neq 0$. See Appendix A for more details.

In this work, we learn the set of parameters

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

It is worth emphasizing that the algorithm proposed in (Figueiredo et al., 2018) assumes that $\sum_{k=1}^{K} \alpha_{k',k} = 1$ and $\beta_{k',k} = \beta_k$ for all $k' \in [K]$. Herein, we relax all these restrictive assumptions.

4 METHOD

4.1 Maximum Likelihood Estimation

Suppose that we observe a sequence of discrete events $\mathcal{P} = \bigcup_{k=1}^{K} \mathcal{P}_k$ over an observation period $[0, T)$ generated by a MWP. The generic approach to infer the parameters of the model is to use regularized maximum-likelihood estimation. The design of regularization depends on the problem at hand, as well as the necessary conditions we are imposing, *e.g.*, positivity or sparsity of the parameters. The log-likelihood function of a

multivariate point-process can be written as

$$\log p(\mathcal{P}|\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\mu}) = \sum_k \sum_{t_{k,i}\in\mathcal{P}_k} \log \lambda_k(t_{k,i}|\mathcal{H}_t) - \sum_k \int_0^T \lambda_k(t|\mathcal{H}_t)dt. \tag{2}$$

The specific form of Wold process defined in (1), makes the log-likelihood function non-convex with respect to $\boldsymbol{\beta}$. Moreover, maximum-likelihood estimation of point processes typically scales poorly to high dimensional settings. Therefore, we use a variational inference approach to circumvent both issues of non-convexity and scalability.

4.2 Variational Inference

Variational inference (VI) is a method for approximating the posterior distribution over the model parameters given the observations. In order to represent the posterior in a tractable form, it is common to define an auxiliary variable that relates the parameters and the observations (Simma and Jordan, 2010; Linderman and Adams, 2015; Figueiredo et al., 2018). Observing that the conditional intensity in (1) is a summation of $K+1$ terms, we can use the superposition theorem of point processes to define the parent of each event (Daley and Jones, 2003; Linderman and Adams, 2014). More precisely, we define 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. This cause is either the background rate μ_k or one of the processes in $[K]$. Specifically,

$$\mathbf{z}_{k,i} = [z_{k,i}^{(0)}, z_{k,i}^{(1)}, \cdots, z_{k,i}^{(K)}].$$

where $z_{k,i}^{(0)}$ is 1 if and only if $t_{k,i}$ was caused by the background rate μ_k or $z_{k,i}^{(k')}$ is 1 if and only if $t_{k,i}$ was caused by process k'. As an event has only one cause (or parent), $\mathbf{z}_{k,i}$ is a one-hot vector, which means that $\sum_{k'=0}^{K} z_{k,i}^{(k')} = 1$ for all k and i.

Our approach is conceptually similar to the VI algorithm proposed in (Linderman and Adams, 2015) for learning Hawkes processes. However, because Hawkes processes suffer from long memory, each preceding event is a potential parent, so the number of auxiliary variables increases exponentially with the number of events. To overcome this issue, Linderman and Adams (2015) approximate the Hawkes process by discretizing time, which has the drawback of introducing an approximation error. In contrast, as a result of the Markovian nature of MWPs, only the preceding events of each dimension are the potential parents. Thus, the number of potential parents of an event remains constant.

Having defined the auxiliary variable $\mathbf{z}$, we approximate the posterior distribution $p(\boldsymbol{\mu},\mathbf{z},\boldsymbol{\alpha},\boldsymbol{\beta}|\mathcal{P})$ with a variational distribution $q(\boldsymbol{\mu},\mathbf{z},\boldsymbol{\alpha},\boldsymbol{\beta})$ that minimizes the KL-divergence between p and q. In particular, VI solves for the optimal variational distribution that minimizes the KL-divergence, or equivalently it maximizes the evidence lower bound (ELBO), given by

$$\begin{aligned}\text{ELBO}(q) = \; & \mathbb{E}_q\left[\log p(\boldsymbol{\mu},\mathbf{z},\boldsymbol{\alpha},\boldsymbol{\beta},\mathcal{P})\right] \\ & - \mathbb{E}_q\left[\log q(\boldsymbol{\mu},\mathbf{z},\boldsymbol{\alpha},\boldsymbol{\beta})\right].\end{aligned} \tag{3}$$

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

$$\begin{aligned}q(\boldsymbol{\mu},\mathbf{z},\boldsymbol{\alpha},\boldsymbol{\beta}) = & \prod_{k=1}^{K} q(\mu_k) \times \prod_{k=1}^{K}\prod_{i=1}^{|\mathcal{P}_k|} q(\mathbf{z}_{k,i}) \\ & \times \prod_{k=1}^{K}\prod_{k'=1}^{K} q(\alpha_{k',k})\, q(\beta_{k',k}).\end{aligned} \tag{4}$$

Using this approximation and coordinate ascent for maximizing (3), we obtain the variational distributions $\{q(\mu_k), q(\mathbf{z}_{k,i}), q(\alpha_{k',k}), q(\beta_{k',k})\}$ by selecting appropriate prior distributions over the parameters. Coordinate ascent is a commonly used optimization method in VI. It iteratively updates each factor of the mean-field variational density while holding the others unchanged (Wang and Blei, 2013). Next, we give the variational updates. Derivation of these updates can be found in Appendix B.

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

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

where the probabilities

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

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 update of $\alpha_{k',k}$. Selecting a Gamma distribution with shape $a_{k',k}$ and rate $b_{k',k}$ for prior of $\alpha_{k',k}$ results in a Gamma mean-field approximation of the posterior, given by

$$q(\alpha_{k',k}) = \text{Gamma}\,(A_{k',k}; B_{k',k}), \tag{6}$$

where

$$A_{k',k} \coloneqq a_{k',k} + \sum_{i=1}^{|\mathcal{P}_k|} \mathbb{E}_{q(z_{k,i}^{(k')})}[z_{k,i}^{(k')}],$$
$$B_{k',k} \coloneqq b_{k',k} + \sum_{i=1}^{|\mathcal{P}_k|} \mathbb{E}_{q(\beta_{k',k})}\Big[\frac{t_{k,i} - t_{k,i-1}}{\beta_{k',k} + \Delta_{k',k}(t_{k,i})}\Big].$$

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

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

where

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

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

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

This form of the density function is not straightforward to work with as we need to compute $\mathbb{E}[\log(\beta_{k',k} + \Delta_{k',k}(t))]$ and $\mathbb{E}[1/(\beta_{k',k} + \Delta_{k',k}(t))]$. On the other hand, the form of this distribution suggests that it can be well-approximated by an inverse-Gamma distribution. Hence, we approximate this distribution with an Inverse-Gamma and use the following variational update

$$q(\beta_{k',k}) = \text{Inverse-Gamma}(\Phi_{k',k}; \Psi_{k',k}), \tag{9}$$

where $\Phi_{k',k}$ and $\Psi_{k',k}$ are selected so that its moments coincide with the moments of the distribution in (8). This leads to the following form of the parameters

$$\Phi_{k',k} \coloneqq \frac{w x_w - v x_v}{x_w - x_v} - 1,$$
$$\Psi_{k',k} \coloneqq \frac{(w-v) x_v x_w}{x_w - x_v}.$$

In these equations, $w \geq 1$, $w > v \in \mathbb{R}_+$ and x_w denotes the smallest positive real root of equation $g_w(x) = 0$, where

$$g_w(x) \coloneqq \frac{\phi_{k',k} + 1 - w}{x} + \sum_{i=1}^{|\mathcal{P}_k|} \frac{\mathbb{E}_{q(z_{k,i}^{(k')})}[z_{k,i}^{(k')}]}{x + \Delta_{k',k}(t_{k,i})} - \frac{\psi_{k',k}}{x^2} - \sum_{i=1}^{|\mathcal{P}_k|} \frac{\mathbb{E}_{q(\alpha_{k',k})}[\alpha_{k',k}](t_{k,i} - t_{k,i-1})}{(x + \Delta_{k',k}(t_{k,i}))^2}. \tag{10}$$

The following lemma guarantees the existence of such an inverse-Gamma distribution.

Lemma 1. *If $0 < \phi_{k',k} + 1 - w < \phi_{k',k} + 1 - v$ and $w \geq 1$, then $\Phi_{k',k}$ and $\Psi_{k',k}$ exist and are positive.*

Proof. Let $g_u(x)$ denote the function in (10). We have $\lim_{x\to 0_+} g_u(x) = -\infty$ and $\lim_{x\to\infty} g_u(x) = 0_+$ for $u = v, w$. Thus, $g_u(x)$ has at least one positive real root. Without loss of generality, let x_v and x_w be the smallest positive real roots of $g_v(x)$ and $g_w(x)$, respectively. Given the assumption in the lemma, it is clear that $g_v(x) > g_w(x)$ for $x > 0$. Hence, $0 = g_v(x_v) > g_w(x_v)$ and $g_v(x_w) > g_w(x_w) = 0$. Since x_w is the smallest positive root of $g_w(x)$, $\lim_{x\to 0_+} g_u(x) = -\infty$, and $g_w(x_v) < 0$, then $x_w > x_v$. Now, using the facts that $w > v$ and $w \geq 1$, and the equations of $\Phi_{k',k}$ and $\Psi_{k',k}$, we conclude the proof. □

In Section 5.3, we provide an example of realizations of the distribution in (8) to illustrate the goodness of this approximation.

5 EXPERIMENTAL RESULTS

We now provide a comparison of our VI approach with state-of-the-art approaches in two sets of experiments. We first simulate synthetic realizations of MWPs, where the ground-truth parameters are known, to measure the performance and efficiency of each approach to recover the influence matrix. Subsequently, we evaluate our approach on two real-world datasets of multivariate asynchronous time series. For reproducibility, we provide a detailed description of the setup of each experiment in Appendix E and make the code publicly available online at `https://github.com/trouleau/var-wold`.

5.1 Experiments on Synthetic Data

To simulate MWPs, we generated Erdös–Rényi random graphs with K nodes. We sampled background rates $\{\mu_k^*\}$ from Uniform$[0, 0.05]$, edge weights $\{\alpha_{k',k}^*\}$ from Uniform$[0.1, 0.2]$ for all edges, and parameters $\{\beta_{k',k}^*\}$ from Uniform$[1, 2]$, all independently. To evaluate the scalability of an approach with respect to the number

of dimensions, we varied the number of dimensions K between 5 and 50 nodes. The results are averaged over 5 graphs with 4 realizations of MWPs for each graph, with an average of 10 000 training events per dimension. We compared the performance of our approach, denoted as VI, with three other methods:

GB. The MCMC sampling-based approach Granger-Busca from (Figueiredo et al., 2018) is the only other approach designed for MWPs. Note that GB does not estimate a posterior for $\{\beta_{k',k}\}$, but instead uses the data-driven heuristic $\beta_{k',k} = \text{median}(\{t_{k,i+1} - t_{k,i} | t_{k,i} \in \mathcal{P}_k\})/\exp(1)$, referred to as *Busca*, as advised by the authors.

BBVI. To compare with another method based on VI, we adapted the approach in (Salehi et al., 2019), originally designed for Hawkes processes, for learning MWPs. The approach is based on black-box VI and the variational EM algorithm. Details of the adaptation are provided in Appendix E.1.

MLE. For a simple baseline, we also compared with maximum-likelihood estimation with a Tikhonov regularizer.

Note that the three Bayesian approaches VI, BBVI, and GB estimate a posterior over the parameters rather than a point-estimate as done in MLE. Therefore, we use the mean of the posteriors to evaluate the performance of the estimated influence weights $\{\hat{\alpha}_{k',k}\}$.

To evaluate the performance of each approach in learning the influence matrix of the processes, we compared the estimated $\{\hat{\alpha}_{k',k}\}$ with the ground-truth $\{\alpha^*_{k',k}\}$ using three metrics common in the literature (Xu et al., 2016; Figueiredo et al., 2018; Salehi et al., 2019):

- **Relative error.** To evaluate the distance of the estimated weights to the ground-truth ones, we computed the averaged relative error defined as $|\hat{\alpha}_{k',k} - \alpha^*_{k',k}|/\alpha^*_{k',k}$ when $\alpha^*_{k',k} \neq 0$, and $\hat{\alpha}_{k',k}/(\min_{\alpha^*_{n,m}>0} \alpha^*_{n,m})$ when $\alpha^*_{k',k} = 0$ (Xu et al., 2016; Salehi et al., 2019).
- **Precision@n.** To assess the performance of the approaches at recovering the top edges, we used precision@n, which is defined as the average fraction of correctly identified edges in the top n largest estimated weights (Figueiredo et al., 2018).
- **PR-AUC.** Considering that an edge exists in the influence matrix if the learned value $\hat{\alpha}_{k',k} > \eta$, we evaluate the performance of the resulting binary edge classification problem using the area under the precision-recall curve over all thresholds $\eta > 0$.

Our results are depicted in Figure 2. As shown in Figure 2a-2c, both VI and BBVI outperform GB and MLE on all metrics. Despite the non-convexity of the problem, both methods achieve an almost perfect precision@10 and PR-AUC for all numbers of dimensions. On the other hand, MLE showed a large variance in the estimated parameters[1].

The computational complexity per iteration is of order $\mathcal{O}(|\mathcal{P}|K)$ for our VI algorithm and $\mathcal{O}(|\mathcal{P}| \log K)$ for GB, where $|\mathcal{P}|$ is the total number of events and K is the number of dimensions. However, note that each update of our VI approach is easily parallelizable over each of the K^2 edges, while GB can only be parallelized over each of the K nodes. We also observed that VI empirically requires fewer iterations to converge compared to GB. To show this, we compared the runtime of each method in Figure 2d. Note that all methods were implemented in Python, GB was compiled in Cython, and VI used just-in-time compilation with the library Numba. To make runtime comparison fair, all methods were run on a single core on the same machine. Although both VI and BBVI perform well, the runtime of VI is about one order of magnitude faster than BBVI and is similar to GB. More details are available in Appendix C.

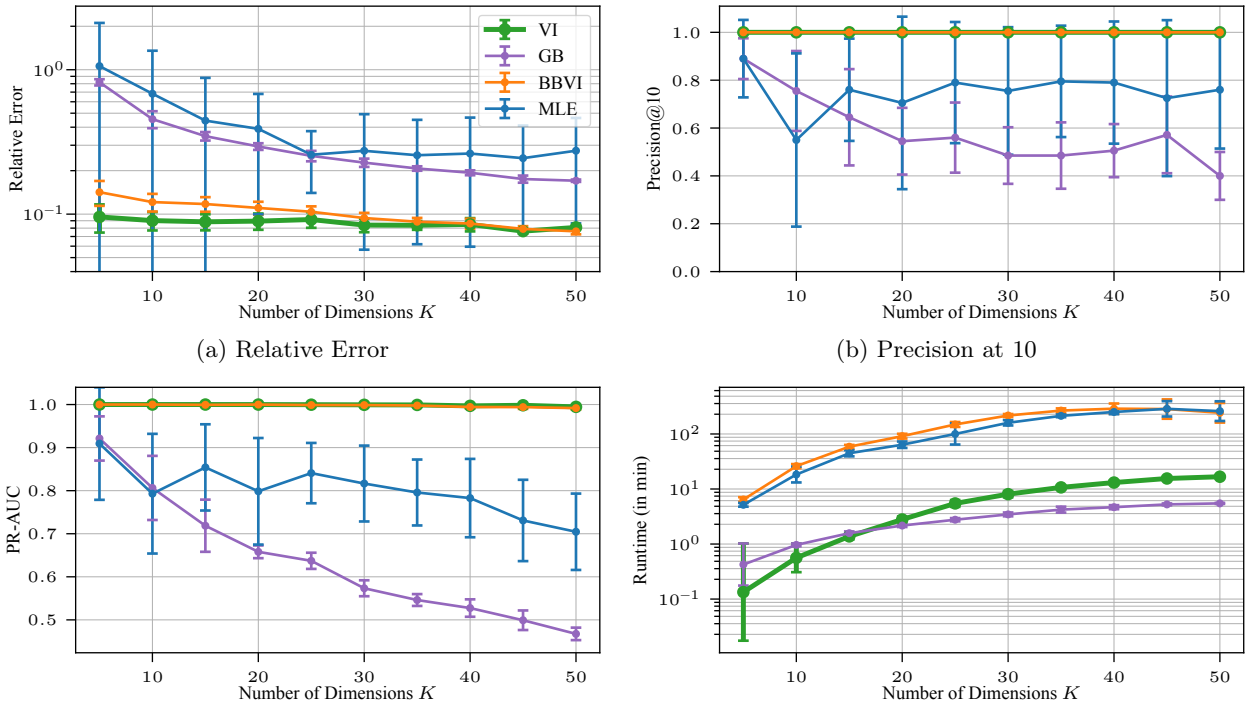
5.2 Experiments on Real Datasets

We evaluated the approaches on two datasets from the Snap Network Repository[2]: (1) the email-Eu-core dataset that contains emails sent between collaborators from a large European research institution (Paranjape et al., 2017; Figueiredo et al., 2018), and (2) the Memetracker dataset containing online blog posts (Leskovec et al., 2009; Achab et al., 2017; Figueiredo et al., 2018). We compare our VI approach on these datasets with GB, which currently is the most scalable approach. In (Figueiredo et al., 2018), the authors showed that MWPs are better suited than Hawkes processes for these two datasets.

Email-EU-core. The dataset consists of source nodes (senders) that send events to destination nodes (receivers) at some time. Each event is represented as a triplet (source, destination, timestamp). Following the same preprocessing as (Figueiredo et al., 2018), we aggregated the events by receiver and considered the top 100 receivers, *i.e.*, those with the most events, resulting in a total of 92 924 events. We hypothesize that the ground-truth influence matrix is determined

[1] To highlight that the discrepancy of performance does not come from the particular experimental setup, we present additional results in Appendix C for an alternative experimental setup matching the structural assumption of GB, *i.e.*, where $\sum_k \alpha_{k',k} = 1$.

[2] https://snap.stanford.edu/data/



(a) Relative Error

(b) Precision at 10

(c) PR-AUC

(d) Runtime

Figure 2: Results on synthetic data for varying number of dimensions K. Panels (a) (log-scale) relative error, (b) precision@10, (c) PR-AUC and panel (d) (log-scale) empirical runtime of each approach in minutes.

by the number of emails sent by a sender to a receiver. More precisely, the ground-truth was defined as a graph whose nodes are both senders and receivers and whose directed edges captured the flow of communication from sender to receiver, weighted by fraction of received emails. We used the first 75% of the dataset for training and the remaining 25% for testing. We evaluated the results for two tasks: (1) An edge-estimation task where we evaluated the performance of each approach to recover the ground-truth influence matrix of the training set, and (2) an event-prediction task where we measured the predictive log-likelihood of the two approaches on the held-out test set. Because both approaches estimate the posterior distribution over the parameters (in contrast with single point-estimators), we characterized the uncertainty using Monte Carlo samples of the parameters from the learned posterior distributions, and we reported the mean and standard deviations among these samples for all the reported metrics.

The results on the Email-EU-core dataset are shown in Table 1. VI outperfoms GB on all metrics. The improvement can be explained by the fact that VI relaxes the restrictive assumptions in GB, *i.e.*, that $\sum_{k=1}^{K} \alpha_{k',k} = 1$ and $\beta_{k',k} = \beta_k \ \forall k \in [K]$, which we discussed in Section 3.

MemeTracker. The dataset consists of the times of publication of online blog posts along with the hyperlinks within. The dataset was originally collected to analyze the propagation of short phrases, called *memes*, and is often modeled as a multivariate point-process (Rodriguez et al., 2014; Achab et al., 2017; Figueiredo et al., 2018). To evaluate the performance of their algorithms, (Achab et al., 2017) and (Figueiredo et al., 2018) extracted a ground-truth influence matrix based on hyperlink references among the websites and reported the precision of their methods, which were low. This could be explained by the presence of noise in the dataset[3], as well as by non-stationarity, (*i.e.*, varying dynamics of the data over time), which were reported in (Rodriguez et al., 2014). Therefore, for the MemeTracker dataset, we focused on the predictive capability of our algorithm compared to GB by evaluating the predictive log-likelihood on held-out data. More precisely, we split the data into observation windows of about 12 days, trained on each window and tested on the following one. The log-likelihood values were

[3]The assumption that a hyperlink (source) appearing in another blog (destination) implies a causal influence might not be accurate. For example, a hyperlink can appear in comments of a blog, unrelated from its main content.

Table 1: Results on the EU-email-core dataset.

	PR-AUC	Precision@10	Precision@50	Precision@200	Pred. log-likelihood
VI	**0.33** (±0.00)	**0.40** (±0.00)	**0.40** (±0.00)	**0.47** (±0.00)	**−5.64** (±1.16e−2)
GB	0.32 (±0.00)	0.20 (±0.00)	0.35 (±0.07)	0.43 (±0.03)	−11.56 (±1.78e−2)

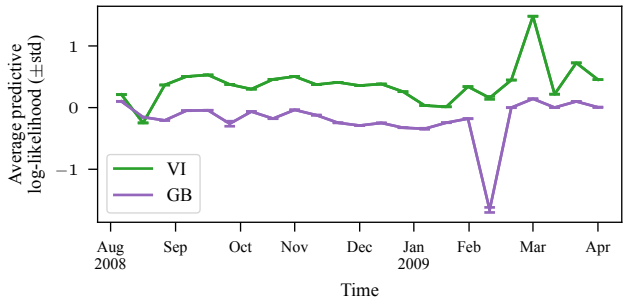


Figure 3: Held-out predictive log-likelihood on the MemeTracker dataset.

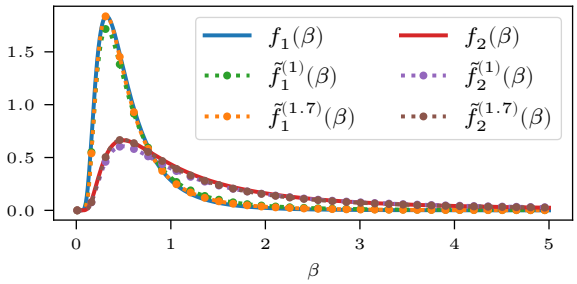


Figure 4: Two examples of the distribution in (8) and their corresponding Inverse-Gamma approximation in (9). The Inverse-Gamma are denoted by tilde and they are obtained by selecting $v = 0$ and $w \in \{1, 1.7\}$.

normalized by number of events[4].

Figure 3 depicts the results on the MemeTracker dataset. Again, to account for uncertainty in the estimation, we also reported the mean and standard deviations of the predictive log-likelihood among Monte Carlo samples of the parameters. We see that VI outperforms GB for all observation windows. Moreover, the values are not stable over time, confirming the findings of Rodriguez et al. (2014) that the dynamics of the data are indeed non-stationary.

5.3 Example of the $q(\beta_{k',k})$ Approximation

Next, we evaluate the goodness of the approximation proposed in (9) for the update of $\beta_{k',k}$. To do so, we considered two realizations of the distribution in (8), given by

$$f_1(\beta) \propto \beta^{-1-1} \cdot e^{-1/\beta} \cdot (\beta + 2.9)^{-0.2} \cdot e^{-\frac{0.6}{\beta+2.9}} \cdot (\beta + 1.7)^{-0.8} \cdot e^{\frac{-1.6}{\beta+1.7}}, \quad (11)$$

$$f_2(\beta) \propto \beta^{-3-1} \cdot e^{-1/\beta} \cdot (\beta + 0.3)^{-0.3} \cdot e^{\frac{-1.6}{\beta+0.3}} \cdot (\beta + 1.1)^{-1.8} \cdot e^{\frac{-0.4}{\beta+1.1}} \cdot (\beta + 2)^{-0.1} \cdot e^{\frac{-.8}{\beta+2}}, \quad (12)$$

and we computed their approximated Inverse-Gamma distribution using (9). We display the resulting distributions in Figure 4. The approximated Inverse-Gamma distributions are denoted by $\widetilde{f}$ and are obtained by selecting $v = 0$ and $w \in \{1, 1.7\}$.

[4]For reproducibility, we provide the detailed preprocessing steps in Appendix E.3.

In order to measure the goodness of the approximation, we also present in Table 2 the KL-divergence between (11) and (12) and their approximated Inverse-Gamma distributions for several choices of w.

Table 2: KL-divergences between the distributions in (11) and (12) and their approximations.

	$\widetilde{f}^{(1)}$	$\widetilde{f}^{(1.3)}$	$\widetilde{f}^{(1.7)}$	$\widetilde{f}^{(1.9)}$	$\widetilde{f}^{(2.5)}$
f_1	0.0370	0.0272	0.0126	0.0070	–
f_2	0.0296	0.0222	0.0151	0.0119	0.0062

6 CONCLUSION

We have addressed the problem of learning the parameters of multivariate temporal point-processes. This problem has been widely studied for the multivariate Hawkes process, but the long memory of such processes makes Bayesian inference difficult. Due to its Markovian intensity function, the Multivariate Wold process does not suffer from the same shortcomings and has therefore recently gained popularity in the literature. We relaxed the limiting structural assumptions of the only available state-of-the-art method and proposed an efficient Bayesian algorithm based on variational inference for multivariate Wold processes with exponential transition probabilities. Our experiments on both synthetic and real-world datasets show that our approach outperforms the state-of-the art and is able to accurately and efficiently recover the influence matrix of the process.

Acknowledgments

The work presented in this paper was supported in part by the Swiss National Science Foundation under grant number 200021-182407.

References

Achab, M., Bacry, E., Gaïffas, S., Mastromatteo, I., and Muzy, J.-F. (2017). Uncovering causality from multivariate hawkes integrated cumulants. *The Journal of Machine Learning Research*, 18(1):6998–7025.

Alves, R. A. d. S., Assuncao, R. M., and Vaz de Melo, P. O. S. (2016). Burstiness scale: A parsimonious model for characterizing random series of events. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1405–1414.

Bacry, E. and Muzy, J.-F. (2014). Second order statistics characterization of hawkes processes and non-parametric estimation. *arXiv preprint arXiv:1401.0903*.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Blundell, C., Beck, J., and Heller, K. A. (2012). Modelling reciprocating relationships with hawkes processes. In *Advances in Neural Information Processing Systems*, pages 2600–2608.

Cox, D. R. (1955). Some statistical methods connected with series of events. *Journal of the Royal Statistical Society: Series B (Methodological)*, 17(2):129–157.

Daley, D. (1982). Stationary point processes by markov-dependent intervals and infinite intensity. *Journal of Applied Probability*, 19(A):313–320.

Daley, D. J. and Jones, D. V. (2003). *An Introduction to the Theory of Point Processes: Elementary Theory of Point Processes*. Springer.

Eichler, M., Dahlhaus, R., and Dueck, J. (2017). Graphical modeling for multivariate hawkes processes with nonparametric link functions. *Journal of Time Series Analysis*, 38(2):225–242.

Etesami, J., Kiyavash, N., Zhang, K., and Singhal, K. (2016). Learning network of multivariate hawkes processes: A time series approach. *UAI*.

Figueiredo, F., Borges, G. R., de Melo, P. O. V., and Assunção, R. (2018). Fast estimation of causal interactions using wold processes. In *Advances in Neural Information Processing Systems*, pages 2971–2982.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438.

Guttorp, P. and Thorarinsdottir, T. L. (2012). What happened to discrete chaos, the quenouille process, and the sharp markov property? some history of stochastic point processes. *International Statistical Review*, 80(2):253–268.

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90.

Isham, V. (1977). A markov construction for a multi-dimensional point process. *Journal of Applied Probability*, 14(3):507–515.

Leskovec, J., Backstrom, L., and Kleinberg, J. (2009). Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, page 497–506, New York, NY, USA. Association for Computing Machinery.

Linderman, S. and Adams, R. (2014). Discovering latent network structure in point process data. In *International Conference on Machine Learning*, pages 1413–1421.

Linderman, S. W. and Adams, R. P. (2015). Scalable bayesian inference for excitatory point process networks. *arXiv preprint arXiv:1507.03228*.

Monti, R. P., Hellyer, P., Sharp, D., Leech, R., Anagnostopoulos, C., and Montana, G. (2014). Estimating time-varying brain connectivity networks from functional mri time series. *NeuroImage*, 103:427–443.

Namaki, A., Shirazi, A., Raei, R., and Jafari, G. (2011). Network analysis of a financial market based on genuine correlation and threshold method. *Physica A: Statistical Mechanics and its Applications*, 390(21-22):3835–3841.

Paranjape, A., Benson, A. R., and Leskovec, J. (2017). Motifs in temporal networks. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, page 601–610, New York, NY, USA. Association for Computing Machinery.

Quinn, C. J., Kiyavash, N., and Coleman, T. P. (2015). Directed information graphs. *IEEE Transactions on information theory*, 61(12):6887–6909.

Rodriguez, M. G., Leskovec, J., Balduzzi, D., and Schölkopf, B. (2014). Uncovering the structure and temporal dynamics of information propagation. *Network Science*, 2(1):26–65.

Salehi, F., Trouleau, W., Grossglauser, M., and Thiran, P. (2019). Learning hawkes processes from a handful of events. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors,

Advances in Neural Information Processing Systems 32, pages 12715–12725. Curran Associates, Inc.

Simma, A. and Jordan, M. I. (2010). Modeling events with cascades of poisson processes. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, UAI'10, page 546–555, Arlington, Virginia, USA. AUAI Press.

Trouleau, W., Etesami, J., Grossglauser, M., Kiyavash, N., and Thiran, P. (2019). Learning hawkes processes under synchronization noise. In *International Conference on Machine Learning*, pages 6325–6334.

Vaz de Melo, P. O. S., Faloutsos, C., Assunção, R., Alves, R., and Loureiro, A. A. (2015). Universal and distinct properties of communication dynamics: How to generate realistic inter-event times. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 9(3):1–31.

Vaz de Melo, P. O. S., Faloutsos, C., Assunção, R., and Loureiro, A. (2013). The self-feeding process: a unifying model for communication dynamics in the web. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1319–1330.

Wang, C. and Blei, D. M. (2013). Variational inference in nonconjugate models. *Journal of Machine Learning Research*, 14(Apr):1005–1031.

Wold, H. (1948). On stationary point processes and markov chains. *Scandinavian Actuarial Journal*, 1948(1-2):229–240.

Xu, H., Farajtabar, M., and Zha, H. (2016). Learning granger causality for hawkes processes. In *International Conference on Machine Learning*, pages 1717–1726.

Yang, Y., Etesami, J., He, N., and Kiyavash, N. (2017). Online learning for multivariate hawkes processes. In *Advances in Neural Information Processing Systems*, pages 4937–4946.

Zhou, K., Zha, H., and Song, L. (2013). Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *Artificial Intelligence and Statistics*, pages 641–649.

A Granger Causality in Multivariate Wold Processes

As we discussed, X Granger-causes Y if knowing the past of X improves our prediction of the future of Y given the past of the remainder of the processes in the network. (Quinn et al., 2015) showed that directed information (DI) (or transfer entropy) captures Granger causality in a network of stochastic processes. More precisely, X Granger-causes Y iff $H(Y|X,\mathbf{Z}) \neq H(Y|\mathbf{Z})$, where H denotes the Shanon entropy and $\mathbf{Z}$ represents all the variables in the network apart from X and Y. Using this notation and following the steps in (Etesami et al., 2016) and (Eichler et al., 2017) that relate Granger causality and the intensity function of multivariate Hawkes processes. It can be shown that in MWPs, dimension k' causes dimension k at time t, if $H(\lambda_k(t|\mathcal{H}_t)|\mathcal{H}_t) \neq H(\lambda_k(t|\mathcal{H}_t)|\mathcal{H}_t \setminus \mathcal{H}_t^{k'})$.

By the definition of the conditional intensity function of MWP, if $\alpha_{k',k}/(\beta_{k',k} + \Delta_{k',k}(t)) = 0$, then $H(\lambda_k(t|\mathcal{H}_t)|\mathcal{H}_t) = H(\lambda_k(t|\mathcal{H}_t)|\mathcal{H}_t \setminus \mathcal{H}_t^{k'})$. In other words, dimension k' does not Granger cause dimension k. As a result, the support of the influence matrix encodes the Granger-causal networkstructure of a MWP.

B Derivations of the Variational Inference Updates

In this section, we present the derivations of variational updates for the MWP parameters. From (Blei et al., 2017), we know that maximizing the ELBO with the mean-field assumption implies that the variational update of a parameter x_j from the parameter set $\mathbf{x}$ given the observation set $\mathbf{d}$ has the following form

$$q(x_j) = \exp\left(\mathbb{E}_{-x_j}\left[\log p(\mathbf{x},\mathbf{d})\right]\right) + \text{const.} \tag{13}$$

In the above expression, $p(\mathbf{x},\mathbf{d})$ denotes the joint distribution of the parameters and the observations. The expectation is taken with respect to the variational density of all the parameters except x_j. Using this update rule, we can explicitly derive all the variational updates of interest. For notational simplicity, we use the following definitions throughout the appendix.

$$\begin{aligned}
&\boldsymbol{\alpha}_k := \{\alpha_{k',k}\}_{k'=1}^K, \quad \boldsymbol{\alpha} := \{\boldsymbol{\alpha}_k\}_{k=1}^K,\\
&\boldsymbol{\beta}_k := \{\beta_{k',k}\}_{k'=1}^K, \quad \boldsymbol{\beta} := \{\boldsymbol{\beta}_k\}_{k=1}^K,\\
&\mathbf{z} := \{\mathbf{z}_{k,i} : i \in [|\mathcal{P}_k|]\}_{k=1}^K, \quad \boldsymbol{\mu} := \{\mu_k\}_{k=1}^K.
\end{aligned}$$

B.1 Variational update for the auxiliary parent variables $\mathbf{z}_{k,i}$

Let $-\mathbf{z}_{k,i}$ denote the set of all parameters except $\mathbf{z}_{k,i}$. From (13), we obtain

$$\log\left(q\left(\mathbf{z}_{k,i}\right)\right) = \mathbb{E}_{-\mathbf{z}_{k,i}}\left[\log p(\boldsymbol{\mu},\mathbf{z},\boldsymbol{\alpha},\boldsymbol{\beta}|\mathcal{P})\right] + \text{const.} = \mathbb{E}_{-\mathbf{z}_{k,i}}\left[\log p\left(\mathbf{z}_{k,i}\middle|\mu_k,\boldsymbol{\alpha}_k,\boldsymbol{\beta}_k,\mathcal{P}\right)\right] + \text{const.}$$

The last equality holds because of the mean-field assumption. In order to obtain the conditional distribution of the parent variable given the rest of the parameters, we use the fact that the number of events in a given interval is distributed according to Poisson distribution. Hence,

$$\begin{aligned}
p\left(\mathbf{z}_{k,i}\middle|\mu_k,\boldsymbol{\alpha}_k,\boldsymbol{\beta}_k,\mathcal{P}\right) &= \text{Poisson}\left(z_{k,i}^{(0)};\mu_k(t_{k,i}-t_{k,i-1})\right)\\
&\times \prod_{k'=1}^K \text{Poisson}\left(z_{k,i}^{(k')};\frac{\alpha_{k',k}(t_{k,i}-t_{k,i-1})}{\beta_{k',k}+\Delta_{k',k}(t_{k,i})}\right)\mathbf{I}_{\{\sum_{k'} z_{k,i}^{(k')}=1\}},
\end{aligned} \tag{14}$$

where $\mathbf{I}$ denotes the indicator function. The product form in (14) results again the mean-field assumption, and the indicator enforces that $\sum_{k'=0}^K z_{k,i}^{(k')} = 1$. Substituting the above conditional distribution into the variational update equation, we obtain

$$\begin{aligned}
\log\left(q\left(\mathbf{z}_{k,i}\right)\right) =& \mathbb{E}_{\mu_k}\left[\log\left(\mu_k(t_{k,i}-t_{k,i-1})\right)^{z_{k,i}^{(0)}}\right] + \mathbb{E}_{\boldsymbol{\alpha}_k,\boldsymbol{\beta}_k}\left[\log\prod_{k'=1}^K\left(\frac{\alpha_{k',k}(t_{k,i}-t_{k,i-1})}{\beta_{k',k}+\Delta_{k',k}(t_{k,i})}\right)^{z_{k,i}^{(k')}}\right]\\
&+ \log\mathbf{I}_{\{\sum_{k'} z_{k,i}^{(k')}=1\}} + \text{const.}\\
=& z_{k,i}^{(0)}\mathbb{E}_{\mu_k}\left[\log\left(\mu_k(t_{k,i}-t_{k,i-1})\right)\right]\\
&+ \sum_{k'=0}^K z_{k,i}^{(k')}\mathbb{E}_{\boldsymbol{\alpha}_k,\boldsymbol{\beta}_k}\left[\log\left(\frac{\alpha_{k',k}(t_{k,i}-t_{k,i-1})}{\beta_{k',k}+\Delta_{k',k}(t_{k,i})}\right)\right] + \log\mathbf{I}_{\{\sum_{k'} z_{k,i}^{(k')}=1\}} + \text{const.}
\end{aligned}$$

Therefore, $q(\mathbf{z}_{k,i})$ is Categorical, i.e.,

$$q(\mathbf{z}_{k,i}) = \text{Categorical}\,(K+1; p_{k,i}^{(0)}, ..., p_{k,i}^{(K)}), \tag{15}$$

where $p_{k,i}^{(k')}$ is the probability that $z_{k,i}^{(k')}$ is one and the others are zero. Therefore, $\{p_{k,i}^{(k')}\}$ is a valid probability distribution, *i.e.*, $\sum_{k'} p_{k,i}^{(k')} = 1$.

B.2 Variational update for $\alpha_{k',k}$

From (13), we have

$$\begin{aligned}
&\log\left(q\left(\alpha_{k',k}\right)\right) = \mathbb{E}_{-\alpha_{k',k}}\left[\log p(\boldsymbol{\mu}, \mathbf{z}, \boldsymbol{\alpha}, \boldsymbol{\beta}|\mathcal{P})\right] + \text{const.} \\
&= \mathbb{E}_{-\alpha_{k',k}}\left[\log p(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathcal{P}) + \log p(\boldsymbol{\alpha}|\mathcal{P})\right] + \text{const.} \\
&= \sum_{i=1}^{|\mathcal{P}_k|} \mathbb{E}_{-\alpha_{k',k}}\left[\log p\left(\mathbf{z}_{k,i}\middle|\mu_k, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_k, \mathcal{P}\right)\right] + \log p(\alpha_{k',k}) + \text{const.} \\
&= \sum_{i=1}^{|\mathcal{P}_k|} \mathbb{E}_{-\alpha_{k',k}}\left[z_{k,i}^{(k')}\ \log\left(\frac{\alpha_{k',k}(t_{k,i} - t_{k,i-1})}{\beta_{k',k} + \Delta_{k',k}(t_{k,i})}\right) - \left(\frac{\alpha_{k',k}(t_{k,i} - t_{k,i-1})}{\beta_{k',k} + \Delta_{k',k}(t_{k,i})}\right)\right] + \log p(\alpha_{k',k}) + \text{const.} \\
&= \sum_{i=1}^{|\mathcal{P}_k|} \mathbb{E}_{z_{k,i}^{(k')}}[z_{k,i}^{(k')}] \log(\alpha_{k',k}) - \alpha_{k',k} \sum_{i=1}^{|\mathcal{P}_k|} \mathbb{E}_{\beta_{k',k}}\left[\frac{t_{k,i} - t_{k,i-1}}{\beta_{k',k} + \Delta_{k',k}(t_{k,i})}\right] + \log p(\alpha_{k',k}) + \text{const.}
\end{aligned}$$

If we select the prior distribution of $\alpha_{k',k}$ to be Gamma with with shape $a_{k',k}$ and rate $b_{k',k}$, the variational posterior remains Gamma, *i.e.*,

$$q(\alpha_{k',k}) = \text{Gamma}\left(A_{k',k}; B_{k',k}\right), \tag{16}$$

where the shape and rate parameters are respectively given by

$$A_{k',k} := a_{k',k} + \sum_{i=1}^{|\mathcal{P}_k|} \mathbb{E}_{z_{k,i}^{(k')}}[z_{k,i}^{(k')}], \qquad B_{k',k} := b_{k',k} + \sum_{i=1}^{|\mathcal{P}_k|} \mathbb{E}_{\beta_{k',k}}\Big[\frac{t_{k,i} - t_{k,i-1}}{\beta_{k',k} + \Delta_{k',k}(t_{k,i})}\Big].$$

B.3 Variational update for μ_k

The update rule for μ_k is similar to the one of $\alpha_{k',k}$.

$$\begin{aligned}
&\log\left(q\left(\mu_k\right)\right) = \mathbb{E}_{-\mu_k}\left[\log p(\boldsymbol{\mu}, \mathbf{z}, \boldsymbol{\alpha}, \boldsymbol{\beta}|\mathcal{P})\right] + \text{const.} = \mathbb{E}_{-\mu_k}\left[\log p(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathcal{P}) + \log p(\boldsymbol{\mu}|\mathcal{P})\right] + \text{const.} \\
&= \sum_{i=1}^{|\mathcal{P}_k|} \mathbb{E}_{-\mu_k}\left[\log p\left(\mathbf{z}_{k,i}\middle|\mu_k, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_k, \mathcal{P}\right)\right] + \log p(\mu_k) + \text{const.} \\
&= \sum_{i=1}^{|\mathcal{P}_k|} \mathbb{E}_{-\mu_k}\left[z_{k,i}^{(0)}\ \log\left(\mu_k(t_{k,i} - t_{k,i-1})\right) - \mu_k(t_{k,i} - t_{k,i-1})\right] + \log p(\mu_k) + \text{const.} \\
&= \sum_{i=1}^{|\mathcal{P}_k|} \mathbb{E}_{z_{k,i}^{(0)}}[z_{k,i}^{(0)}] \log(\mu_k) - \mu_k \sum_{i=1}^{|\mathcal{P}_k|} (t_{k,i} - t_{k,i-1}) + \log p(\mu_k) + \text{const.}
\end{aligned}$$

Selecting a Gamma prior with shape c_k and rate d_k implies the result.

B.4 Variational update for $\beta_{k',k}$

Note that $\beta_{k',k}$ is defined for k', k in $[K] := \{1, ..., K\}$. Similar to the update rule for $\alpha_{k',k}$, we have

$$\begin{aligned}
\log\left(q\left(\beta_{k',k}\right)\right) &= \mathbb{E}_{-\beta_{k',k}}\left[\log p(\boldsymbol{\mu}, \mathbf{z}, \boldsymbol{\alpha}, \boldsymbol{\beta}|\mathcal{P})\right] + \text{const.} \\
&= \mathbb{E}_{-\beta_{k',k}}\left[\log p(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathcal{P}) + \log p(\boldsymbol{\beta}|\mathcal{P})\right] + \text{const.} \\
&= \sum_{i=1}^{|\mathcal{P}_k|} \mathbb{E}_{-\beta_{k',k}}\left[z_{k,i}^{(k')}\ \log\left(\frac{\alpha_{k',k}(t_{k,i} - t_{k,i-1})}{\beta_{k',k} + \Delta_{k',k}(t_{k,i})}\right) - \left(\frac{\alpha_{k',k}(t_{k,i} - t_{k,i-1})}{\beta_{k',k} + \Delta_{k',k}(t_{k,i})}\right)\right] + \log p(\beta_{k',k}) + \text{const.} \\
&= -\sum_{i=1}^{|\mathcal{P}_k|} \mathbb{E}_{z_{k,i}^{(k')}}[z_{k,i}^{(k')}] \log(\beta_{k',k} + \Delta_{k',k}(t_{k,i})) - \mathbb{E}_{\alpha_{k',k}}[\alpha_{k',k}] \sum_{i=1}^{|\mathcal{P}_k|} \frac{t_{k,i} - t_{k,i-1}}{\beta_{k',k} + \Delta_{k',k}(t_{k,i})} + \log p(\beta_{k',k}) + \text{const.}
\end{aligned}$$

If we select an Inverse-Gamma prior for $\beta_{k',k}$ with shape $\phi_{k',k}$ and scale $\psi_{k',k}$, $q(\beta_{k',k})$ will be proportional to

$$(\beta_{k',k})^{-\phi_{k',k}-1} e^{\left(-\frac{\psi_{k',k}}{\beta_{k',k}}\right)} \prod_{i=1}^{|\mathcal{P}_k|} (\beta_{k',k} + \Delta_{k',k}(t_{k,i}))^{-\mathbb{E}_{z_{k,i}^{(k')}}[z_{k,i}^{(k')}]} e^{\left(-\frac{\mathbb{E}\alpha_{k',k}[\alpha_{k',k}](t_{k,i} - t_{k,i-1})}{\beta_{k',k} + \Delta_{k',k}(t_{k,i})}\right)}, \text{ for } k', k \in [K]. \tag{17}$$

This distribution is not analytically tractable, but it can be well-approximated by an inverse-Gamma distribution. Therefore, we approximate the variational update for $\beta_{k',k}$ as an Inverse-Gamma$(\Phi_{k',k}, \Psi_{k',k})$. We choose its parameters $\Phi_{k',k}$ and $\Psi_{k',k}$ such that its resulting moments coincide with the moments of the distribution in (17). Finding the moments of the distribution in (17) tends to be quite challenging. Instead, we use the following observation to obtain our approximation.

Remark 1. *Let $f(x; a, b)$ be the p.d.f. of the Inverse-Gamma distribution with shape a and rate b. The Function $x^u f(x; a, b)$ has a global maximum that occurs at $b/(a + 1 - u)$ for $u \in \mathbb{R}_+$.*

We argue that if the u-th moment of a Inverse-Gamma variable, with shape $\Phi_{k',k}$ and rate $\Psi_{k',k}$, coincides with the u-th moment of the distribution in (17), denoted by $h(x)$, then we should have

$$\int_{\mathbb{R}_+} x^u f(x; \Phi_{k',k}, \Psi_{k',k}) dx = \int_{\mathbb{R}_+} x^u h(x) dx.$$

A sufficient condition for the above equality is that the points that maximize $x^u f(x; \Phi_{k',k}, \Psi_{k',k})$ and $x^u h(x)$ should coincide. This happens if

$$\frac{\Psi_{k',k}}{\Phi_{k',k} + 1 - u} = x_u, \tag{18}$$

where x_u is the point that maximizes $x^u h(x)$. By equating the derivative of $\log(x^u h(x))$ to zero, it is easy to see that x_u is the real root of the following equation

$$\frac{\phi_{k',k} + 1 - u}{x} + \sum_{i=1}^{|\mathcal{P}_k|} \frac{\mathbb{E}_{q(z_{k,i}^{(k')})}[z_{k,i}^{(k')}]}{x + \Delta_{k',k}(t_{k,i})} - \frac{\psi_{k',k}}{x^2} - \sum_{i=1}^{|\mathcal{P}_k|} \frac{\mathbb{E}_{q(\alpha_{k',k})}[\alpha_{k',k}](t_{k,i} - t_{k,i-1})}{(x + \Delta_{k',k}(t_{k,i}))^2} = 0.$$

Since the above function has continuous derivatives, we can use, for example, Halley's method to find its root. Equation (18) alone cannot specify both $\Psi_{k',k}$ and $\Phi_{k',k}$. Thus, by selecting two different u, say $u = v$ and $u = w$, we obtain

$$\frac{\Psi_{k',k}}{\Phi_{k',k} + 1 - v} = x_v, \quad \frac{\Psi_{k',k}}{\Phi_{k',k} + 1 - w} = x_w.$$

Solving for $\Psi_{k',k}$ and $\Phi_{k',k}$, we obtain

$$\Phi_{k',k} = \frac{w x_w - v x_v}{x_w - x_v} - 1, \quad \Psi_{k',k} = \frac{(w - v) x_w x_v}{x_w - x_v}.$$

Lemma 1 implies that such x_v and x_w exist and the above shape and scale are positive for appropriate choices of v, w, and $\phi_{k',k}$.

B.5 Computing the required statistics

Note that the variational updates introduced in Section B depend on each others through some common statistics. For instance, the variational update for the auxiliary variable $\mathbf{z}_{k,i}$ in (15) requires computing $\mathbb{E}_{\alpha_{k',k}}[\log \alpha_{k',k}]$. In this section, we provide analytical expressions of such statistics.

Since $q(\mathbf{z}_{k,i})$ is Categorical, for $k \in [K], i \in \mathcal{P}_k, k' \in [K] \cup \{0\}$, we have

$$\mathbb{E}_{z_{k,i}^{(k')}}[z_{k,i}^{(k')}] = p_{k,i}^{(k')}, \tag{19}$$

where $p_{k,i}^{(k')}$ is the probability that $z_{k,i}^{(k')} = 1$ and $z_{k,i}^{(l)} = 0$ for $l \neq k'$.

Given that $\alpha_{k',k}$ has a Gamma $(A_{k',k}; B_{k',k})$ distribution, we have for $k \in [K], k' \in [K]$,

$$\mathbb{E}_{\alpha_{k',k}}[\alpha_{k',k}] = \frac{A_{k',k}}{B_{k',k}}, \tag{20}$$

$$\mathbb{E}_{\alpha_{k',k}}[\log(\alpha_{k',k})] = \Upsilon(A_{k',k}) - \log(B_{k',k}), \tag{21}$$

where $\Upsilon(\cdot)$ denotes the digamma function. Similarly, we can obtain the required statistics of μ_k.

Because we use an inverse-Gamma distribution for the variational update of $\beta_{k',k}$,

$$\mathbb{E}_{\beta_{k',k}}\Big[\frac{1}{\beta_{k',k} + \Delta_{k',k}(t_{k,j})}\Big] = \int_{\mathbb{R}_+} \frac{1}{y + \Delta_{k',k}(t_{k,j})} y^{-\Phi_{k',k}-1} \exp\left(-\Psi_{k',k}/y\right) \frac{dy}{Z},$$

$$\mathbb{E}_{\beta_{k',k}}\Big[\log(\beta_{k',k} + \Delta_{k',k}(t_{k,j}))\Big] = \int_{\mathbb{R}_+} \log(y + \Delta_{k',k}(t_{k,j})) y^{-\Phi_{k',k}-1} \exp\left(-\Psi_{k',k}/y\right) \frac{dy}{Z},$$

where Z denotes the normalization factor of the inverse-Gamma$(\Phi_{k',k}, \Psi_{k',k})$. The above expressions can be approximated as follows

$$\mathbb{E}\Big[\frac{1}{\beta_{k',k} + \Delta_{k',k}(t_{k,j})}\Big] \approx \frac{1}{\frac{\Psi_{k',k}}{\Phi_{k',k}-1} + \Delta_{k',k}(t_{k,j})}, \tag{22}$$

$$\mathbb{E}\Big[\log\left(\beta_{k',k} + \Delta_{k',k}(t_{k,j})\right)\Big] \approx \log\Big(\frac{\Psi_{k',k}}{\Phi_{k',k}-1} + \Delta_{k',k}(t_{k,j})\Big). \tag{23}$$

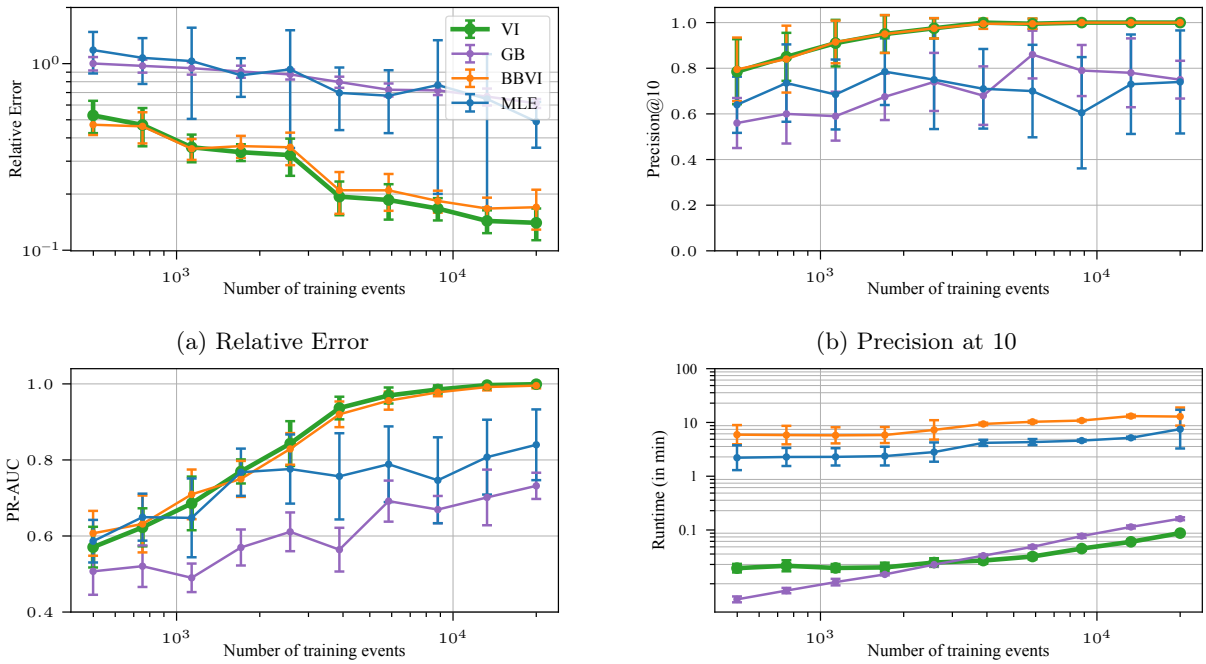
C Additional Experimental Results

In this section, we present some additional experimental results.

Analysis of performance w.r.t. number of training events. To evaluate the number of training samples required to achieve a good performance for each approach, we ran the experiments with the same synthetic simulation setup, fixed the number of dimensions $K = 10$, and varied the number of training events. We present these results in Figure 5. Although BBVI was originally designed to train on small observations sequences, our VI approach does as well or outperforms BBVI.

Alternative simulation setup. To further investigate the effect of the structural constraint required by GB, *i.e.*, $\sum_k \alpha_{k',k} = 1$ for all $k', k \in [K]$, we ran additional experiments on synthetic data where we normalized the ground-truth $\{\alpha_{k',k}\}$ such that $\sum_k \alpha_{k',k} = 1$. The results are shown in Figure 6. We see that, even if GB performs better than in Figure 2, our VI approach still outperforms GB on all metrics.

Robustness to the choice of prior. To investigate the sensitivity of VI to choice of the prior, we ran additional experiments on synthetic data. For $K = 10$ dimensions, we fixed the mean as in the experiments of Section 5.1, and evaluated the performance for variance of the priors of $\{\alpha_{k',k}\}$ and $\{\beta_{k',k}\}$ ranging between 10^{-2} and 10^2. As seen in Figure 7, for a large range of priors, VI remains stable. For all values tested, both the PR-AUC and Precision@10 remained at 1.0.



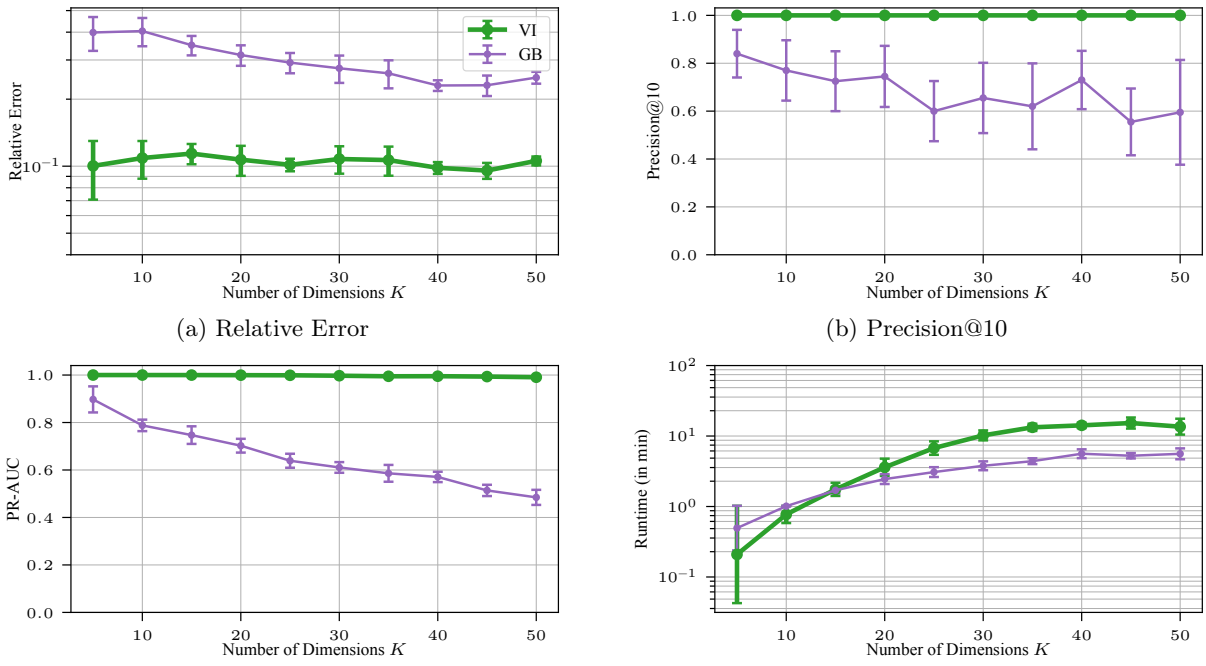
(a) Relative Error

(b) Precision at 10

(c) PR-AUC

(d) Runtime

Figure 5: Results on synthetic data for varying numbers of training events. Panel (a) (log-scale) relative error, (b) precision@10, (c) PR-AUC, and panel (d) (log-scale) empirical runtime of each approach in minutes.



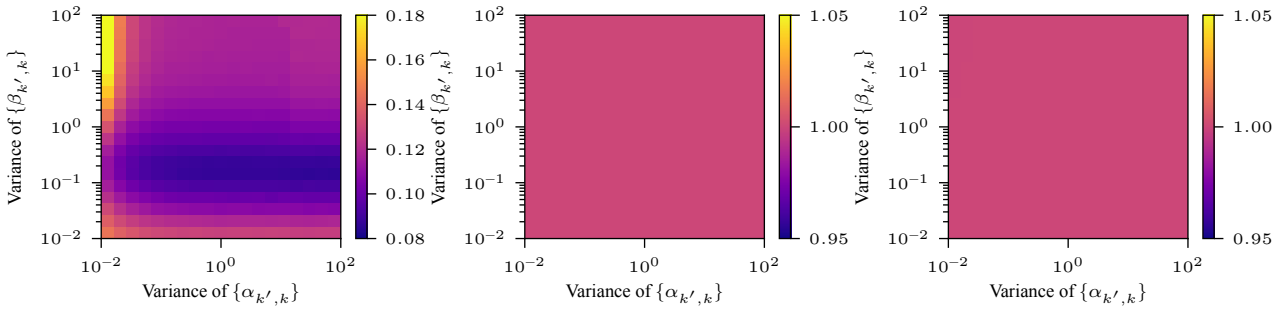
(a) Relative Error

(b) Precision@10

(c) PR-AUC

(d) Runtime

Figure 6: Results on synthetic data for the alternative synthetic simulation setup where we normalize the $\{\alpha_{k',k}\}$ such that $\sum_k \alpha_{k',k} = 1$. Panel (a) (log-scale) relative error, (b) precision@10, (c) PR-AUC, and panel (d) (log-scale) empirical runtime of each approach in minutes.



(a) Relative Error (b) Precision@10 (c) PR-AUC

Figure 7: Analysis of the robustness of VI to the choice of prior. We report the relative error for a wide range of variances for both $\{\alpha k', k\}$, $\{\beta_{k',k}\}$, keeping their mean fixed to the same value used in the experiments.

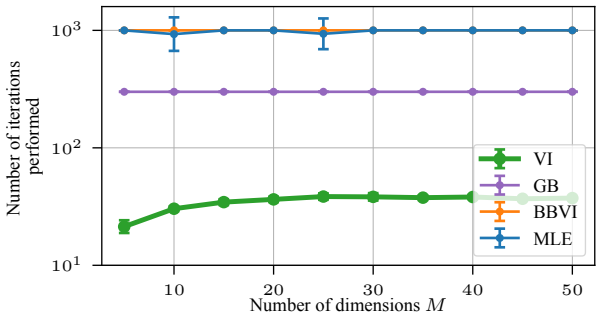


Figure 8: Number of iterations performed in the experiments on synthetic data.

Analysis of the number of iterations. In Figure 2, we discussed the runtime of each algorithm on synthetic data. To make the comparison fair, we also report the number of iterations performed in Figure 8. As stated in Appendix E, we ran VI, BBVI and MLE until convergence or up to maximum 10 000 iterations. As the number of dimensions increases, the number of iterations needed for VI to converge becomes sub-linear. BBVI almost always ran to the cap on the maximum number of iterations because it uses Monte Carlos samples of the posterior at each iteration and hence exhibit a larger variance between iterations. We ran GB for 3000 iterations, which was found to be enough to reach convergence[5].

D Computational Complexity

We report the computational complexity of GB to be $\mathcal{O}(|\mathcal{P}| \log K)$, while the authors of the method originally report $\mathcal{O}(|\mathcal{P}|(\log |\mathcal{P}| + \log K))$ in (Figueiredo et al., 2018). The difference lies in the computation of the inter-event times $\{\Delta_{k',k}(t_{k,i})\}$, where the authors consider the computation of each inter-event time as $\mathcal{O}(\log |\mathcal{P}|)$ at each iteration. However, it suffices to compute these values once and cache them. Therefore, this step is $\mathcal{O}(1)$, which reduces the computational complexity of GB to $\mathcal{O}(|\mathcal{P}| \log K)$.

E Reproducibility

E.1 Simulation setup for synthetic data

We generated Erdös–Rényi random graphs with K nodes. We sampled background rates $\{\mu_k^*\}$ from Uniform$[0, 0.05]$, edge weights $\{\alpha_{k',k}^*\}$ from Uniform$[0.1, 0.2]$ for all edges, and parameters $\{\beta_{k',k}^*\}$ from Uniform$[1, 2]$, all independently. Each algorithm was then run as follows.

VI. We ran the algorithm for a maximum of 10 000 iterations or until convergence. We defined convergence

[5]Note that (Figueiredo et al., 2018) used 300 iterations without further justification.

when the maximum absolut difference of any parameter between two consecutive iterations is less than 10^{-4}. We used priors $p(\mu_k) = \text{Gamma}(0.1, 1)$, $p(\alpha_{k',k}) = \text{Gamma}(0.1, 1)$ and $p(\beta_{k',k}) = \text{InverseGamma}(100, 100)$.

GB. We used the implementation released in (Figueiredo et al., 2018). We used 3000 iterations in all experiments. As advised by the authors, we used the same Dirichlet prior with uniform parameters $1/K$, and set the parameters $\{\beta_{k',k}\}$ to the data-driven heuristic $\beta_{k',k} = \text{median}(\{t_{k,i+1} - t_{k,i} | t_{k,i} \in \mathcal{P}_k\})/\exp(1)$.

BBVI. We adapted the implementation released in (Salehi et al., 2019). More details are provided in E.2. Analogous to VI, we ran the method for a maximum of 10 000 iterations or until convergence. As in (Salehi et al., 2019), we used Log-Normal posterior distributions, Laplacian priors $\{\alpha_{k',k}\}$, and Gaussian priors for $\{\mu_k\}$ and $\{\beta_{k',k}\}$ with the same parameters released in the code of (Salehi et al., 2019).

MLE. Analogous to VI, we ran the method for a maximum of 10 000 iterations or until convergence.

All experiments were run on a single-core, on the same machine with a processor *Intel(R) Xeon(R) CPU E5-2680 v3 @ 2.50GHz* and 256GB of RAM.

E.2 Adaptation of the BBVI approach for Wold processes

BBVI was introduced in (Salehi et al., 2019) to learn the parameters of a Hawkes process. They maximized the ELBO

$$\text{ELBO}(q) = \mathbb{E}_q \left[\log p\left(\mathcal{P}|\theta\right)\right] + \mathbb{E}_q \left[\log p\left(\theta\right)\right] - \mathbb{E}_q \left[\log q\left(\theta\right)\right] \tag{24}$$

over the parameters θ of Hawkes process, using gradient descent with black-box VI. Specifically, a posterior $q(\theta)$ was first postulated (chosen to be Log-Normal), and Monte Carlo samples was used to evalue the expecations in (24). In addition, the variational EM algorithm was used to update the parameters of the prior $p(\theta)$ based on the current estimate of the posterior.

To adapt the approach for MWPs, we only needed to replace the log-likelihood term $\log p\left(\mathcal{P}|\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta}\right)$ in (24) by the likelihood defined in (2).

E.3 Experiments on Real Datasets

E.3.1 Email-EU-core dataset

As explained in Section 5, the Email-EU-core dataset is composed of emails between researchers from a European research institution. Each email in the dataset is a tuple (sender, receiver, timestamp). To build each process from the dataset, we used the same preprocessing steps as Figueiredo et al. (2018). More precisely, we excluded users with no sent email and defined the set of processes as the top-100 users with the most received emails. We then aggregated the timestamps by receivers. The entries in the ground-truth influence matrix are defined by counting the number of emails sent from each sender to each receiver (a weight zero indicates the absence of an edge). The preprocessing code is made available publicly.

For the hyper-parameters, we ran a sweep over the Dirichlet prior of GB over $[0.01, 0.1, 1.0, 10.0, 100.0]$ and reported the best results obtained with 10.0. For VI, we ran a sweep over the of parameters of the priors over $[0.01, 0.1, 1.0, 10.0, 100.0]$ and used $p(\mu_k) = \text{Gamma}(1.0, 1.0)$, $p(\alpha_{k',k}) = \text{Gamma}(1.0, 1.0)$ and $p(\beta_{k',k}) = \text{InverseGamma}(100.0, 100.0)$.

E.3.2 MemeTracker dataset.

The MemeTracker dataset is composed of online blog posts. We used the top-100 blogs with the highest number of published posts and built the processes by aggregating the sequences of published timestamps, resulting in 15 168 774 events in 100 dimensions. The preprocessing code is made available publicly. We ran a sweep over the Dirichlet prior of GB over $[0.01, 0.1, 1.0, 10.0]$, and did not observe a significant difference between the different values and reported the results obtained for 0.01. For VI, we used priors $p(\mu_k) = \text{Gamma}(0.1, 1)$, $p(\alpha_{k',k}) = \text{Gamma}(0.1, 1)$ and $p(\beta_{k',k}) = \text{InverseGamma}(10^4, 10^4)$.