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,

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

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.,

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

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

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

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

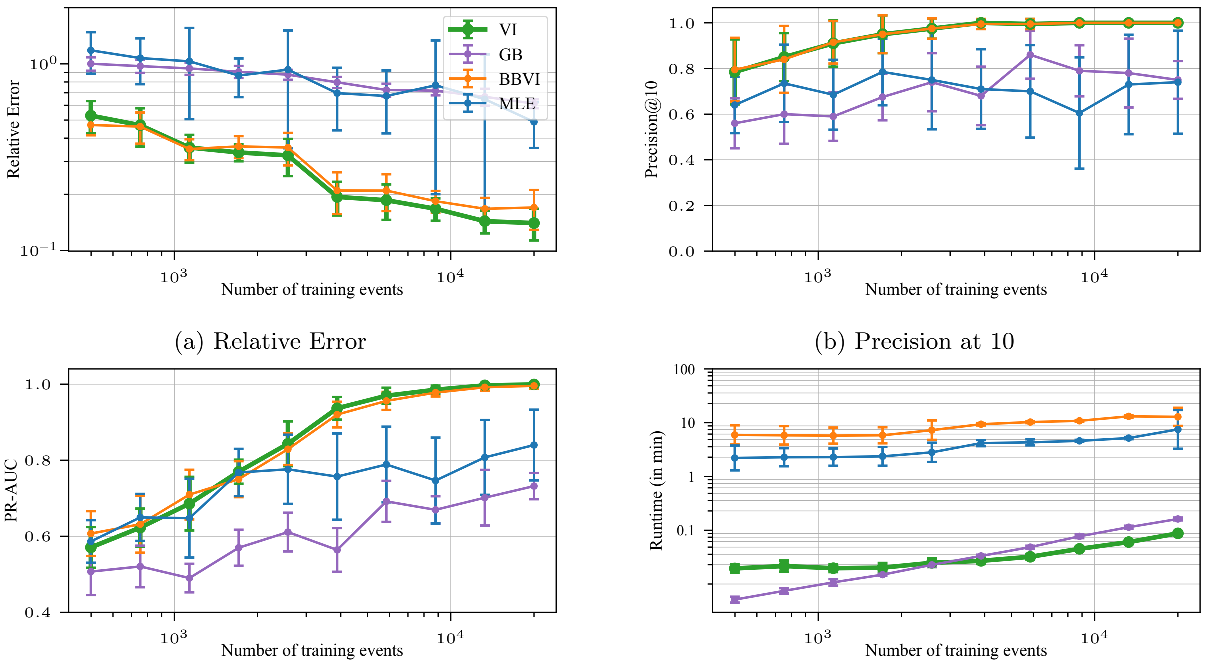
C Additional Experimental Results

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.



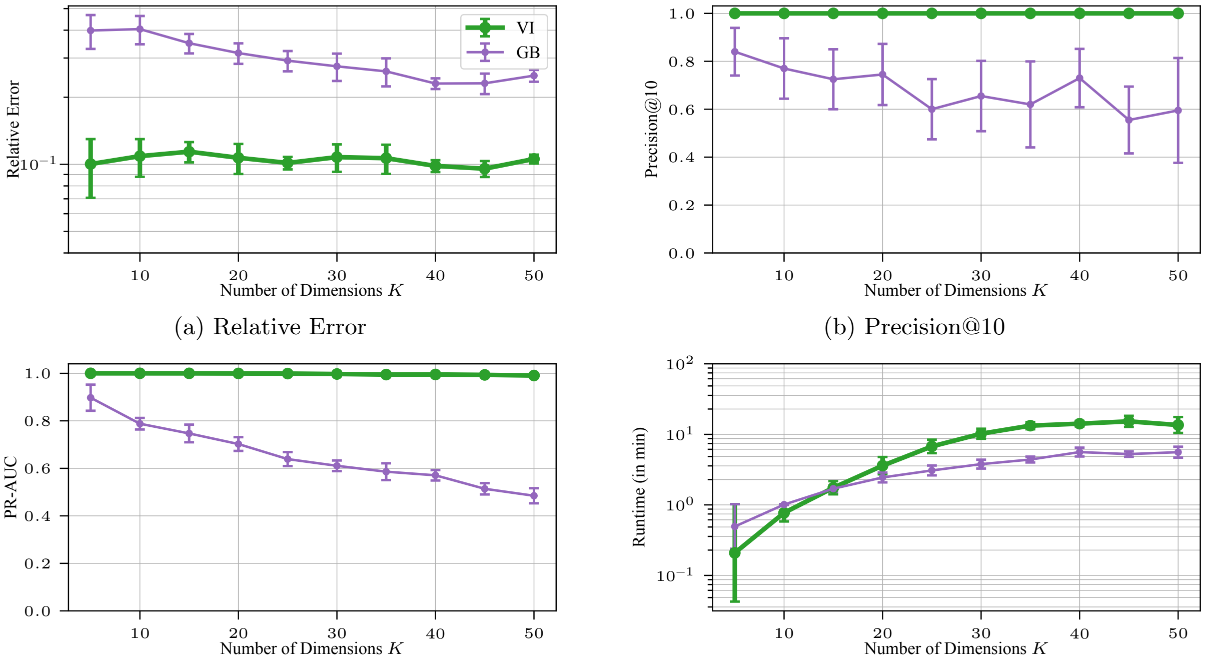
(a) Relative Error

(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.



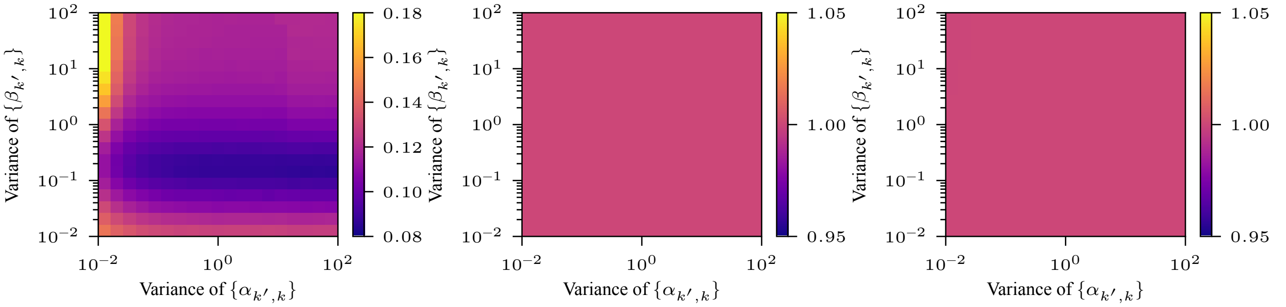
(a) Relative Error

(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.

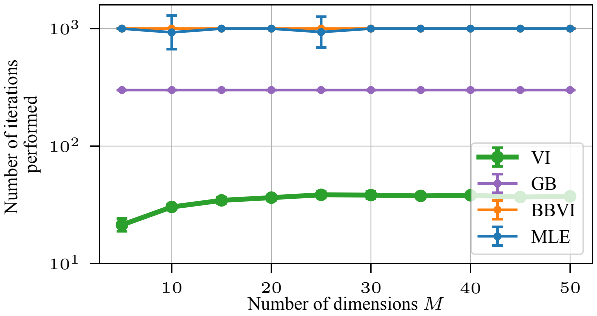


Figure 8: Number of iterations performed in the experiments on synthetic data.

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 $p(\beta_{k',k}) = \text{InverseGamma}(10^4, 10^4)$.