Journal of Machine Learning Research 26 (2025) 1-102 Submitted 8/23; Revised 4/25; Published 8/25

# Scalable and Adaptive Variational Bayes Methods for Hawkes Processes

**Déborah Sulem** DEBORAH.SULEM@BSE.EU
*Barcelona School of Economics*
*Universitat Pompeu Fabra*

**Vincent Rivoirard** VINCENT.RIVOIRARD@CEREMADE.DAUPHINE.FR
*Ceremade, CMRS, UMR 7534*
*Université Paris-Dauphine, PSL University*

**Judith Rousseau** JUDITH.ROUSSEAU@STATS.OX.AC.UK
*Department of Statistics*
*University of Oxford*

**Editor:** Ali Shojaie

## Abstract

Hawkes processes are often applied to model dependence and interaction phenomena in multivariate event data sets, such as neuronal spike trains, social interactions, and financial transactions. In the nonparametric setting, learning the temporal dependence structure of Hawkes processes is generally a computationally expensive task, all the more with Bayesian estimation methods. In particular, for multivariate nonlinear Hawkes processes, Monte-Carlo Markov Chain (MCMC) methods used to sample from the posterior distribution do not scale well to the dimension of the process. Recently, efficient algorithms targeting a mean-field variational approximation of the posterior distribution have been proposed, however, these methods do not allow to perform model selection on the graph of interactions of the Hawkes model. In this work, we propose a novel adaptive Bayesian variational method that performs model selection and can estimate a sparse graphical parameter. For the popular sigmoid Hawkes processes, we design a parallel algorithm which is scalable to high-dimensional point processes and large sequences of events. Furthermore, we unify existing variational Bayes approaches under a general nonparametric inference framework, and analyse the asymptotic properties of these methods under easily verifiable conditions on the prior, the variational class, and the nonlinear model. Finally, through an extensive set of numerical simulations, we demonstrate that our method is able to adapt to the dimensionality of the parameter of the Hawkes process, and is partially robust to certain types of model misspecification.

**Keywords:** temporal point processes, bayesian nonparametrics, connectivity graph, variational approximation

## 1. Introduction

To analyse multiple streams of temporal events with inter-dependence, one is often interested in inferring the local dependence structure between events and in estimating potential interactions between the different streams. In this context, the multivariate Hawkes process is a widely used temporal point process (TPP) model, for instance, in seismology (Ogata, 1999), criminology (Mohler

et al., 2011), finance (Bacry and Muzy, 2015), and social network analysis (Lemonnier and Vayatis, 2014). In particular, the generalised nonlinear multivariate Hawkes model is an extension of the classical self-exciting process (Hawkes, 1971) and can account for different types of temporal interactions often found in event data (Hawkes, 2018; Bonnet et al., 2021), including *excitation* and *inhibition* effects (Hawkes, 2018; Bonnet et al., 2021; Olinde and Short, 2020). There are now several review papers on Hawkes processes to which a reader can refer to: Worrall et al. (2022), Lima (2023), Hawkes (2018).

$$\lambda_t^k = \phi_k\left(\nu_k + \sum_{l=1}^{K}\int_{-\infty}^{t^-} h_{lk}(t-s)dN_s^l\right), \quad k = 1,\dots,K, \tag{1}$$

where $\phi_k$ is the link function, $\nu_k$ the background rate and $h_{lk}$ the interaction function of $N^l$ onto $N^k$. While the link functions $\phi_k$ are typically fixed and known, $\nu_k; h_{lk}, l, k \leqslant K$ are unknown and so might be the graph of interactions, represented by the adjacency matrix with components $\delta_{lk} = \mathbb{1}_{h_{lk}\neq 0}$. The graph of interaction encodes the Granger-causal relationships between the dimensions of the multivariate point process, see Eichler et al. (2017), and is therefore of particular interest; notably in high dimensions as it leads to more interpretable models, see Section 2.2 for more details.

Bayesian estimation of parameters $\nu_k$ and $h_{lk}$ and of the connectivity graph $\delta$ has been studied theoretically in linear models by Donnet et al. (2020) and generalised in nonlinear models by Sulem et al. (2024) where standard nonparametric priors on the $h_{lk}$ such as random histograms, splines and nonparametric mixtures combined with spike and slab layer on $\delta$, enjoy optimal asymptotic properties under mild assumptions on the true parameters. However, the MCMC algorithms used to sample from the posterior distribution do not scale well with the dimension, even in the linear model. The computational issue is even more crucial in nonlinear models, where the likelihood is doubly intractable.

Recently, data augmentation strategies have been used to facilitate the computation of the posterior distribution in the sigmoid Hawkes model. Moreover, mean-field variational Bayes algorithms have been proposed to efficiently obtain an approximation of the posterior distribution derived from certain families of Gaussian priors (Malem-Shinitski et al., 2021; Zhou et al., 2022). However, these methods do not integrate sparse priors and the estimation of the graph of interactions. Moreover, variational Bayes estimation methods have not yet been theoretically analysed in the context of Hawkes processes.

From a frequentist perspective, penalised contrast estimators have been studied in the linear model by Hansen et al. (2015); Bacry et al. (2020) and by Cai et al. (2024) in the nonlinear model. These methods do not provide measures of uncertainty but scale well with the dimension. In this work, we propose to address a methodological challenge and a theoretical gap in the estimation of nonlinear Hawkes processes. In summary, we design a novel sparsity-inducing variational algorithm that performs Bayesian model selection and is applicable to high-dimensional data, and we derive asymptotic guarantees for variational Bayes methods. Our contributions are further detailed below.

- We propose a variational Bayes model selection procedure to infer the connectivity graph parameter of Hawkes processes, and also to construct adaptive estimators.

- We design a sparsity-inducing variational Bayes algorithm that estimates a sparse connectivity graph based on a thresholding heuristic.

- We construct a general nonparametric variational Bayes estimation framework for multivariate Hawkes processes and analyse the concentration rates of variational posterior distributions. We also apply our general results to variational classes of interest, including mean-field and model-selection variational families.

- In addition to being theoretically valid in the asymptotic regime, we show that our algorithm performs very well in practice. In an extensive set of simulations, we observe that it is scalable to large data sets. In particular, on a desktop with only 8 cores, our algorithm takes less than 8 hours to run for processes with 64 dimensions and more than 88000 events. Moreover, it is able to uncover the true graph parameter, and is even robust to certain types of model misspecification.

In Section 2, we describe our general model and inference setup and present our novel adaptive and sparsity-inducing variational algorithm in Section 3 with a special focus on the sigmoid model in Section 4. Section 5 contains our general results, and their applications to prior and variational families of interest in the Hawkes model. Finally, we report in Section 6 the results of an in-depth simulation study. Besides, the proofs of our main results are reported in Appendix D.

## 2. Multivariate Hawkes Processes and Bayesian Nonparametric Inference

In this section, we first introduce some useful notation (Section 2.1) and state some background notions on Hawkes processes (Section 2.2). Then we describe the Bayesian methodology for this model (Section 2.3) and introduce variational methods with model selection (Section 2.4).

### 2.1 Notation

For a function $h$, we denote $\|h\|_1 = \int_{\mathbb{R}} |h(x)|dx$ the $L_1$-norm, $\|h\|_2 = \sqrt{\int_{\mathbb{R}} h^2(x)dx}$ the $L_2$-norm, $\|h\|_\infty = \sup_{x\in\mathbb{R}} |h(x)|$ the supremum norm, and $h^+ = max(h, 0)$, $h^- = max(-h, 0)$ its positive and negative parts. For a $K \times K$ matrix $M$, we denote $r(M)$ its spectral radius, $\|M\|$ its spectral norm, and $tr(M)$ its trace. For a vector $u \in \mathbb{R}^K, \|u\|_1 = \sum_{k=1}^K |u_k|$. The notation $k \in [K]$ is used for $k \in \{1, \dots, K\}$. For a set $B$ and $k \in [K]$, we denote $N^k(B)$ the number of events of $N^k$ in $B$ and $N^k|_B$ the point process measure restricted to the set $B$. For random processes, the notation $\overset{\mathcal{L}}{=}$ corresponds to equality in distribution. We also denote by $C(u, \mathcal{H}_0, d)$ the covering number of a set $\mathcal{H}_0$ by balls of radius $u$ w.r.t. a metric $d$. For any $k \in [K]$, let $\mu_k^0 = \mathbb{E}_0[\lambda_t^k(f_0)]$ be the mean of $\lambda_t^k(f_0)$ under the stationary distribution $\mathbb{P}_0$. For a set $\Omega$, its complement is denoted $\Omega^c$. We also use the notations $u_T \lesssim v_T$ if $|u_T/v_T|$ is bounded when $T \to \infty$, $u_T \gtrsim v_T$ if $|v_T/u_T|$ is bounded and $u_T \asymp v_T$ if $|u_T/v_T|$ and $|v_T/u_T|$ are bounded. We recall that a function $\phi$ is $L$-Lipschitz, if for any $(x, x') \in \mathbb{R}^2$, $|\phi(x) - \phi(x')| \leqslant L|x - x'|$. We denote $\mathbb{1}_n$ and $\mathbf{0}_n$ the all-ones and all-zeros vectors of size $n$. Finally, we denote $\mathcal{H}(\beta, L_0)$ the Hölder class of $\beta$-smooth functions with radius $L_0$. For a set $\mathcal{V}$, we denote $\mathcal{V}^{\otimes K} = \underbrace{\mathcal{V} \times \cdots \times \mathcal{V}}_{K \text{ times}}$.

### 2.2 Multivariate Hawkes Processes

Recall that Hawkes processes are temporal point processes (TPP), i.e. counting processes on $\mathbb{R}^K$, where $K \in \mathbb{N}\backslash\{0\}$ is the number of dimensions (components) of the process, and defined on a

probability space $(\mathcal{X}, \mathcal{G}, \mathbb{P})$. Denoting the observed process by $N = (N_t)_{t\in[0,T]} = (N_t^1, \dots, N_t^K)_{t\in[0,T]}$ over the period $[0, T]$, for each $k = 1, \dots, K$ and time $t \in [0, T]$, $N_t^k \in \mathbb{N}$ counts the number of events that have occurred until $t$ at component $k$, therefore, $(N_t^k)_{t\in[0,T]}$ is an integer-valued, non-decreasing, process. The law of a TPP is characterised by its conditional intensity function (or, in short, intensity), denoted $(\lambda_t)_t = (\lambda_t^1, \dots, \lambda_t^K)_{t\in\mathbb{R}}$, that intuitively gives the infinitesimal probability rate of events, conditionally on the history of the process, i.e,

$$\lambda_t^k dt = \mathbb{P}\left[\text{event at dimension } k \text{ in } [t, t+dt] \middle| \mathcal{G}_t\right], \quad k = 1, \dots, K, \quad t \in [0, T],$$

where $\mathcal{G}_t = \sigma(N_s, 0 \leqslant s < t)$ denotes the history of the process until time $t$. In the nonlinear Hawkes model, the intensity function has the form (1), where the *background* or *spontaneous* rate of events $\nu_k \geqslant 0$ and the *interaction functions* (or *triggering kernels*) $h_{lk} : \mathbb{R}^+ \to \mathbb{R}$ model the causal relation of past events of $N^l$ onto $N^k$. In this regard the graph of interactions (or connectivity graph) represented by the adjacency matrix $\delta = (\delta_{lk})_{l,k\leqslant K}$ with $\delta_{lk} := \mathbb{1}_{h_{lk}\neq 0}$

In the Hawkes model, the graphical model is given by the non-zero interaction functions. Defining for each $(l, k)$, $\delta_{lk} := \mathbb{1}_{h_{lk}\neq 0}$, the graph parameter $\delta := (\delta_{lk})_{l,k} \in \{0, 1\}^{K\times K}$ is a Granger-causal graph, called the connectivity graph or graph of interactions. In particular, it implies that for any $(l, k)$, $N^k$ is *locally-dependent* on $N^l$, if and only if $h_{lk} \neq 0$ (Eichler et al., 2017). Note that in the complete graph, i.e. $\delta = \mathbb{1}\mathbb{1}^T$, the number of functions parametrising $\lambda_t^k$ is $K^2$.

Another essential part of the nonlinear Hawkes model are the link functions $\phi = (\phi_k)_k$, which are here considered to be known and chosen by the practitioner. We will assume that these functions are monotone non-decreasing, so that a value $h_{lk}(x) > 0$ can be interpreted as an excitation effect, and $h_{lk}(x) < 0$ as an inhibition effect, for some $x \in \mathbb{R}^+$. Common choices of link functions are ReLU functions $\phi_k(x) = \max(x, 0) = (x)_+$ (Hansen et al., 2015; Chen et al., 2017; Costa et al., 2020; Lu and Abergel, 2018; Bonnet et al., 2021; Deutsch and Ross, 2022), sigmoid-type functions, e.g., $\phi_k(x) = \theta_k(1 + e^x)^{-1}$ with a scale parameter $\theta_k > 0$ (Zhou et al., 2021, 2022; Malem-Shinitski et al., 2021), softplus functions $\phi_k(x) = \log(1 + e^x)$ (Mei and Eisner, 2017), or clipped exponential functions, i.e., $\phi_k(x) = \min(e^x, \Lambda_k)$ with $\Lambda_k > 0$ (Gerhard et al., 2017; Carstensen et al., 2010). Under some conditions on $\phi$ and the interaction functions, there exist a stationary version of the process (see Brémaud and Massoulié (1996); Sulem et al. (2024)).

We note that when all the interaction functions are non-negative and $\phi_k(x) = x$ for every $k$, the intensity (1) corresponds to the linear Hawkes model. As it will be useful to describe our algorithm in Section 4.1, we define the *linear* part of the intensity as

$$\tilde{\lambda}_t^k = \nu_k + \sum_{l=1}^{K} \int_{-\infty}^{t^-} h_{lk}(t - s) dN_s^l, \quad k = 1, \dots, K, \quad t \in \mathbb{R}. \tag{2}$$

With this notation, we can re-write the nonlinear intensity (1) as $\lambda_t^k = \phi_k(\tilde{\lambda}_t^k)$.

### 2.3 Bayesian Framework

From now on, we assume that the data we observe, $N$, is a realisation from a stationary Hawkes process with known link functions $(\phi_k)_k$ and unknown parameter $f_0 = (\nu_0, h_0)$. We make the common assumption that the interaction functions have bounded support (see for instance Hansen et al. (2015); Donnet et al. (2020); Sulem et al. (2024); Cai et al. (2024)) and let $A \geqslant \sup\{x \in$

$\mathbb{R}^+; \max_{l,k} |h^0_{lk}(x)| > 0\}$ be an upper bound of the support of the $h_{lk}$'s. The length $A$ is sometimes called the memory parameter of the process and we assume that it is fixed.

We assume that we observe $N$ over a window $[-A, T]$, with $T > 0$, which allows us to define the likelihood function for the events in $[0, T]$ - and effectively to base our inference procedure on the observation of $N$ over $[0, T]$. With a slight abuse of notation, we will use $N$ to denote both the process and an observation from it. Then for a parameter $f = (\nu, h)$, the log-likelihood function can be written as a sum of partial log-likelihoods for each dimension:

$$L_T(f) := \sum_{k=1}^{K} L^k_T(f), \quad L^k_T(f) = \left[ \int_0^T \log(\lambda^k_t(f)) dN^k_t - \int_0^T \lambda^k_t(f) dt \right]. \tag{3}$$

In the following, we will denote by $\mathbb{P}_0(.|\mathcal{G}_0)$ the true conditional distribution of $N$, given the initial condition $\mathcal{G}_0$ (that notably includes the observations in $[-A, 0)$). For a parameter $f$, we also define the distribution $\mathbb{P}_f(.|\mathcal{G}_0)$ as $d\mathbb{P}_f(.|\mathcal{G}_0) = e^{L_T(f) - L_T(f_0)} d\mathbb{P}_0(.|\mathcal{G}_0)$. Moreover, we denote $\mathbb{E}_0$ and $\mathbb{E}_f$ the expectations associated to $\mathbb{P}_0(.|\mathcal{G}_0)$ and $\mathbb{P}_f(.|\mathcal{G}_0)$. With a slight abuse of notation, we drop the notation $\mathcal{G}_0$ in the subsequent expressions.

Our goal is to estimate the parameter $f$ from $N$, and we consider a nonparametric setting to do so. We denote by $\mathcal{F}$ the space of possible values for $f$. Given a prior distribution $\Pi$ on $\mathcal{F}$, the posterior distribution is defined, for any subset $B \subset \mathcal{F}$, as

$$\Pi(B|N) = \frac{\int_B \exp(L_T(f)) d\Pi(f)}{\int_{\mathcal{F}} \exp(L_T(f)) d\Pi(f)} =: \frac{N_T(B)}{D_T}, \quad D_T := \int_{\mathcal{F}} \exp(L_T(f)) d\Pi(f). \tag{4}$$

Before studying the problem of computing the posterior distribution, we explicit a hierarchical construction of the prior distribution $\Pi$ that allows to estimate a sparse graphical model.

Firstly, the prior is built so that it puts mass 1 to finite-memory processes, i.e., to parameter $f$ such that the interaction functions $(h_{lk})_{l,k}$ have a bounded support included in $[0, A]$. Secondly, we use a hierarchical spike-and-slab prior based on the connectivity graph parameter $\delta$ similar to Donnet et al. (2020); Sulem et al. (2024). For each pair of dimensions $(l, k) \in [K]^2$, we define $\delta_{lk} \in \{0, 1\}$ with $\delta_{lk} = 0$ if and only if $h_{lk} = 0$. Hence, the $\delta_{lk}$'s are analogous to inclusion variables and $\delta = (\delta_{lk})_{lk} \in \{0, 1\}^{K^2}$ is the connectivity graph associated to $f$. It defines the sparsity structure of $h = (h_{lk})_{l,k}$ and hence the dependency structure in the process, as explained in Section 2.2.

From the previous parametrisation, we thus construct the prior on $h$ as the product of a prior distribution on $\delta$ (the graphical model) and a prior distribution on $h|\delta$. We consider $\Pi_\delta$, a prior distribution on the space $\{0, 1\}^{K^2}$, and, for each $(l, k)$ such that $\delta_{lk} = 1$, $h_{lk} \sim \Pi_{h|\delta}$ where $\Pi_{h|\delta}$ is a prior distribution on functions with support included in $[0, A]$ and is such that $\Pi_{h|\delta}(h_{lk} = 0) = 0$. For clarity of exposition, we consider the case where the prior $\Pi_{h|\delta}$ is built by developing the functions $h_{lk}$, when non null, on a dictionary of functions $(e_j)_{j \geqslant 1}$, such that $e_j : [0, A] \to \mathbb{R}$, $\forall j$. Hence, the non-null interaction functions are parametrised in the following way

$$h_{lk} = \sum_{j=1}^{J_k} h^j_{lk} e_j, \quad h^j_{lk} \in \mathbb{R}, \quad \forall j \in [J_k], \quad J_k \geqslant 1, \quad (l, k) \in [K]^2, \tag{5}$$

where $J_k$ is the number of functions in the dictionary needed to decompose the functions $(h_{lk})_{l \in [K]}$ which are non-null. Note that we assume, for simplicity, that this number is constant for all "incoming" functions $h_{lk}$ at each dimension $k$. Therefore, by specifying a prior distribution $\Pi_J$ on

$J = (J_k)_{k\in[K]}$, a prior distribution $\Pi_{h|\delta,J}$ on the weights $(h^j_{lk})_{j,l,k}$, and a prior distribution $\Pi_\nu$ on $\nu$, we can write the prior distribution on $f$ as

$$d\Pi(f) = d\Pi_\nu(\nu)d\Pi_\delta(\delta)d\Pi_J(J)d\Pi_{h|\delta,J}(h) \tag{6}$$

$$d\Pi_{h|\delta,J}(h) = \prod_{l,k}\Big[(1-\delta_{lk})\delta_{(0)}(h_{lk}) + \delta_{lk}d\tilde{\Pi}_{h|\delta,J}(h_{lk})\Big], \tag{7}$$

where $\delta_{(0)}$ denotes the Dirac measure at 0. This prior is analogous to the spike-and-slab introduced in high-dimensional regression problems (see for instance Castillo and van der Vaart (2012); Castillo et al. (2015) or Hoffmann et al. (2015) to name but a few). Now, we keep this construction general and do not yet further specify the distributions $\Pi_\nu, \Pi_\delta, \Pi_J$ and $\Pi_{h|\delta,J}$.

The posterior distribution (4) resulting from this prior does not have an analytical expression and is generally expensive to compute. However, we note that there is one case where the posterior distribution can be factorised over each dimension, which allows us to perform the computations of each factor independently and in parallel. This happens when $\Pi$ can be written as a product of probability distributions on the dimension-restricted parameters $f_k = (\nu_k, (h_{lk})_{l=1,\dots,K})$, for $k \in [K]$, so that $f = (f_k)_k$. By defining $\mathcal{F}'$ the space of each parameter $f_k$ (assuming for simplicity that it is the same for each dimension), we then have that $\mathcal{F} = \mathcal{F}'^{\otimes K}$. A factorised prior distribution takes the form $d\Pi(f) = \prod_k d\Pi_k(f_k)$ and leads to a similar factorisation for the posterior distribution. This can be seen by looking at the expressions of the log-likelihood function and the intensity function. From (3) and (1), we can see that each partial likelihood term $L^k_T(f)$ depends on $f_k$ and effectively we can write it as a function of $f_k$ only, i.e., $L^k_T(f) = L^k_T(f_k)$. The factorisation of the posterior distribution can be then deduced, since,

$$d\Pi(f|N) = \prod_k d\Pi_k(f_k|N), \quad d\Pi_k(f_k|N) = \frac{\exp\{L^k_T(f_k)\}d\Pi_k(f_k)}{\int_{\mathcal{F}_k}\exp\{L^k_T(f_k)\}d\Pi_k(f_k)}. \tag{8}$$

Note that although the computation of each factor $\Pi_k(.|N)$ can be performed independently and in parallel, it nonetheless requires the whole data $N$.

Despite this possible parallelisation, implementation of Monte-Carlo Markov Chains methods for computing the posterior distribution remains very challenging even in moderate dimensional contexts (Donnet et al., 2020; Zhou et al., 2021; Malem-Shinitski et al., 2021) because (i) for each dimension one has to search among $2^K$ models (corresponding to different graphs of interactions), (ii) the likelihood is doubly intractable due to the nonlinearity in $\phi_k$ and to the integrals $\int_0^T \lambda^k_t(f)dt$. To alleviate this computational bottleneck, we consider in the next section a variational Bayes methodology, which aims to compute an approximation of the posterior distribution. We note that in the rest of this paper, we use the term "high-dimensional" from a practical and computational perspective, to denote data dimensions for which computing the full posterior distribution is not feasible in the nonparametric setting (typically for $K > 3$).

### 2.4 Variational Bayes Inference

For high-dimensional processes, we consider using the variational Bayes methodology to compute an approximation of the posterior distribution. Variational methods in the Hawkes model have been introduced by Zhang et al. (2020) for the linear model and used by Zhou et al. (2022); Malem-Shinitski et al. (2021) for the sigmoid model. In this section, we define a general variational Bayes inference framework and introduce a model selection methodology.

The main idea of variational Bayes inference is to consider a variational class of distributions on the parameter space $\mathcal{F}$, say $\mathcal{V}$, and to approximate the posterior $\Pi(\cdot|N)$ and by a distribution $\hat{Q} : \mathcal{F} \to \mathbb{R}$ defined as

$$\hat{Q} := \arg\min_{Q \in \mathcal{V}} KL\left(Q \| \Pi(.|N)\right), \tag{9}$$

where $KL\left(Q\|\Pi(.|N)\right)$ is the Kullabck-Leibler divergence between $Q$ and $Q'$ defined as

$$KL(Q\|Q') := \begin{cases} \int \log\left(\frac{dQ}{dQ'}\right) dQ, & \text{if } Q \ll Q' \\ +\infty, & \text{otherwise} \end{cases}.$$

The approximating distribution $\hat{Q}$ is called the variational Bayes (VB) posterior distribution and corresponds to the best approximation within $\mathcal{V}$, with respect to the Kullback-Leibler divergence. We first note that under a product posterior (8), the VB posterior also factorises in $K$ factors, $\hat{Q}(f) = \prod_k \hat{Q}_k(f_k)$ where each factor $\hat{Q}_k$ approximates $\Pi_k(.|N)$. Therefore, one can reformulate the objective (9) as

$$\hat{Q}_k := \arg\min_{Q_k \in \mathcal{V}'} KL\left(Q_k \| \Pi_k(.|N)\right), \tag{10}$$

where $\mathcal{V}'$ is a variational class of distributions on $\mathcal{F}'$, and which implies that $\mathcal{V} = \mathcal{V}'^{\otimes K}$.

In the case of the sigmoid Hawkes model, for which $\phi_k(x) \propto (1 + e^{-x})^{-1}$, data augmentation strategies have been proposed to design mean-field variational classes and optimisation algorithms, see Zhou et al. (2022); Malem-Shinitski et al. (2021). We present below a general introduction to these strategies and more details about data augmentation schemes are reported in Appendix B.

Augmenting the likelihood consists in finding a random variable $Z$ and an augmented log-likelihood $L_T^A(f, z)$ such that $z$ has distribution $\mathbb{P}_A$ (conditionally on $N$ and $f$) and

$$\mathbb{E}_{\mathbb{P}_A}(\exp(L_T^A(f, z))|N, f) = \exp(L_T(f)).$$

In this perspective, $z$ can be viewed as a latent variable and the *augmented* posterior distribution is a distribution on $\mathcal{F} \times \mathcal{Z}$ defined as

$$\Pi_A(B|N) = \frac{\int_B \exp(L_T^A(f, z)) d(\Pi(f) \times \mathbb{P}_A(z))}{\int_{\mathcal{F}\times\mathcal{Z}} \exp(L_T^A(f, z)) d(\Pi(f) \times \mathbb{P}_A(z))}, \quad B \subset \mathcal{F} \times \mathcal{Z}.$$

In other words it is the posterior distribution on the augmented space. The aim of such augmentation scheme is to obtain augmented likelihoods $L_T^A(f, z)$ which are tractable. This is the case when the link function is the sigmoid Hawkes function and the set of latent variables corresponds to marks for the point process $N$ and to a realisation of a marked Poisson point process (see Appendix B).

A variational class of interest to approximate $\Pi_A(\cdot|N)$ is the family of distributions on $\mathcal{F} \times \mathcal{Z}$ that factorise over the parameter and latent variable spaces, i.e.,

$$\mathcal{V}_{AMF} = \left\{Q : \mathcal{F} \times \mathcal{Z} \to [0, 1];\ Q(f, z) = Q_1(f)Q_2(z)\right\}. \tag{11}$$

This variational class is called the mean-field family and the variational posterior is then called the *mean-field variational posterior distribution*, defined as

$$\hat{Q}_{AMF} = \arg\min_{Q \in \mathcal{V}_{AMF}} KL(Q\|\Pi_A(.|N)). \tag{12}$$

As will be explained in Section 4.2.2, the mean-field variational class facilitates posterior inference. Zhou et al. (2022) use it together with shrinkage priors on the function weights $(h^j_{lk})$ and design an efficient iterative algorithm to compute the VB posterior. Using instead Gaussian processes priors, Malem-Shinitski et al. (2021) proposes a nonetheless similar algorithm. We build on this previous work and consider the high-dimensional nonparametric setting where selection of the connectivity graph $\delta \in \{0,1\}^{K^2}$ and of the number of functions basis $J = (J_k)_k \in \mathbb{N}^K$ is of interest. We recall that if $\delta$ is a sparse graph, then the effective dimension of $f$ is smaller and the computation of the posterior distribution is facilitated.

## 3. Adaptive Variational Bayes

We now introduce our general model selection procedure within our variational Bayes framework and our more scalable two-step strategy.

### 3.1 Variational Model Selection

We describe our model selection procedure to select the effective dimension of $f$, the parameter of the Hawkes model estimated via a variational Bayes method. We define $m := (\delta, J)$ and call it a *model* for the Hawkes process. We can then re-write our parameter space as

$$\mathcal{F} = \bigcup_{m \in \mathcal{M}} \mathcal{F}_m, \quad \mathcal{F}_m = \{f' \in \mathcal{F};\ \delta' = \delta,\ J' = J\}, \quad m = (\delta, J), \tag{13}$$

where $\mathcal{M}$ is the set of models

$$\mathcal{M} = \left\{m = (\delta, J);\ \delta \in \{0,1\}^{K \times K},\ J \in \mathbb{N}^K\right\}.$$

**Remark 1** *We note that under a factorisable prior, we can define a model for each dimension k as* $m_k = (\delta_{\cdot k}, J_k) \in \mathcal{M}_k$*, with* $\mathcal{M}_k$ *defined similarly to* $\mathcal{M}$*, and further decompose* $\mathcal{F}' = \bigcup_{m_k \in \mathcal{M}_k} \mathcal{F}'_{m_k}$.

We can now construct an *adaptive* variational posterior distribution by considering a mean-field family within each subspace $\mathcal{F}_m$, denoted by $\mathcal{V}^m$. An adaptive VB posterior can be defined in two ways, either by selecting a distribution on a single model (Zhang and Gao, 2020), or by averaging the distributions over multiple models (Ohn and Lin, 2024). We first define the model-restricted variational posterior as

$$\hat{Q}^m = \arg\min_{Q \in \mathcal{V}^m} KL(Q \| \Pi(.|N)), \tag{14}$$

for each model $m$. We then consider two types of adaptive VB posteriors: the model selection $\hat{Q}_{MS}$ and the model averaging $\hat{Q}_{MA}$, as

$$\hat{Q}_{MS} := \hat{Q}^{\hat{m}}, \quad \hat{m} := \arg\max_{m \in \mathcal{M}} ELBO(\hat{Q}^m), \tag{15}$$

$$\hat{Q}_{MA} := \sum_{m \in \mathcal{M}} \hat{\gamma}_m \hat{Q}^m, \tag{16}$$

where $ELBO(\cdot)$ is the *evidence lower bound (ELBO)* defined as

$$ELBO(Q^m) := \mathbb{E}_{Q^m}\left[\log \frac{p(f^m, z, N)}{Q^m(f^m, z)}\right], \quad Q^m \in \mathcal{V}^m, \tag{17}$$

where $p(f^m, z, N)$ is the joint density of a parameter $f^m \in \mathcal{F}_m$, the latent variables $z$, and the data $N$, and $\{\hat{\gamma}_m\}_{m\in\mathcal{M}}$ are the model marginal (variational) probabilities defined as

$$\hat{\gamma}_m = \frac{\Pi_m(m)\exp\left\{ELBO(\hat{Q}_m)\right\}}{\sum_{m\in\mathcal{M}}\Pi_m(m)\exp\left\{ELBO(\hat{Q}_m)\right\}}, \quad \Pi_m(m) := \Pi_\delta(\delta)\Pi_J(J), \quad m \in \mathcal{M}.$$

**Remark 2** *We note that in practice, the VB posterior $\hat{Q}_{MS}$ defined in* (15) *can be simpler to sample from than the model averaging version $\hat{Q}_{MA}$ in* (16)*. However we believe that it is a good idea to track the VB posterior distribution on models $\hat{\gamma}_m$, to choose between* (15) *or* (16)*. In our simulation study, in many cases $\hat{\gamma}_m$ was very concentrated in one model in which case choosing* (15) *is sensible but in some case the posterior weights $\hat{\gamma}_m$ were spread over a couple of models, in which case it was better to use* (16)*.*

In Section 5, we prove that the previous Bayesian model selection approach is theoretically valid for Hawkes processes and we also demonstrate in Section 6 that it performs well in practice for moderately large values of $K$. However, this approach is not feasible for very large dimensions since the set of all possible models $\mathcal{M}$ has cardinality greater than $2^{K^2}$, the cardinality of the graph space $\{0,1\}^{K^2}$. In the next section, we propose an efficient two-step procedure that selects a sparse graph in high-dimensional settings.

### 3.2 The Two-step Procedure

For data with a large number of dimensions $K$, we propose an adaptive and sparsity-inducing variational Bayes procedure for selecting the model and estimating the parameter of Hawkes processes in sparse settings. Before describing our algorithm, we give some intuition and theoretical arguments justifying our methodology.

In Section 5, we prove that, under easy to verify assumptions on the prior and on the parameters, the VB posterior concentrates around the true parameter $f_0$ (in $L_1$-norm defined in (28)) at some rate $\epsilon_T$, which typically depends on the smoothness of the true interaction functions. A consequence of this result is that for each $(l,k) \in [K]^2$, the marginal distribution on $S_{lk} := \|h_{lk}\|_1$ concentrates around the true value $S^0_{lk} := \|h^0_{lk}\|_1$ at the same rate $\epsilon_T$. This also implies that the mean value $\hat{S}_{lk} = \mathbb{E}_{\hat{Q}}[\|h^0_{lk}\|_1]$, with $\hat{Q}$ the adaptive VB posterior $\hat{Q}_{MS}$ or $\hat{Q}_{MA}$ is a consistent estimator and converges towards $S^0_{lk}$ at the rate $\epsilon_T$.

Hence, if for all $(l,k)$ such that $\delta^0_{lk} = 1$, $S^0_{lk}$ is large compared to $\epsilon_T$, then for any threshold $\eta_0$ such that $\epsilon_T << \eta_0 < \min_{l,k} S^0_{lk}$, the mean estimate $\hat{S}_{lk}$ should verify $\hat{S}_{lk} > \eta_0$. This remains true for the adaptive variational posterior within models with the complete graph $\delta_C = \mathbb{1}\mathbb{1}^T$, denoted by $\hat{Q}^C$, since the complete graph necessarily overfits the true graph $\delta_0$. Therefore, we can define the following thresholding estimator of $\delta$

$$\hat{\delta} = (\hat{\delta}_{lk})_{l,k}, \quad \hat{\delta}_{lk} = 1 \quad \Leftrightarrow \quad \hat{S}_{lk} \geqslant \eta_0, \quad \forall (l,k) \in [K]^2, \tag{18}$$

where $\hat{S}_{lk} = \mathbb{E}_{\hat{Q}^C}[\|h^0_{lk}\|_1]$. In Section 5.3, we make the previous argument formal and prove that $\hat{\delta}$ is a consistent estimator of $\delta_0$. To choose the threshold $\eta_0$ in a data-driven way, we proposed a heuristic based on finding a gap in the estimated $L_1$-norms $(\hat{S}_{lk})_{l,k}$ using the variational posterior $\hat{Q}^C$.

More precisely, in our algorithm, we order the estimates $\hat{S}_{lk}$, $(l,k) \in [K]^2$, say $\hat{S}_{(1)} \leqslant \hat{S}_{(2)} \leqslant \cdots \leqslant \hat{S}_{(K^2)}$ and construct 95% credible sets around each $\hat{S}_{lk}$ (using the 2.5% and 97.5% quantiles of the variational posterior on $S_{lk}$). We then choose $\eta_0 \in (\hat{S}_{(i_0)}, \hat{S}_{(i_0+1)})$ by finding $i_0$, the index of the first gap between the credible sets, i.e., the first index for which the credible sets associated to $\hat{S}_{(i_0)}$ do not intersect the ones associated to $\hat{S}_{(i_0+1)}$. In Figure 1, we plot the estimates $(\hat{S}_{(i)})_{i=1,\ldots,16^2}$ (blue dots) and the 95% credible sets, in one of the simulation settings of Section 6 for which $K = 16$. In this case, the true graph $\delta_0$ is sparse and many $S^0_{lk}$ (orange dots) are equal to 0. From this figure, we can see that by choosing $\eta_0$ anywhere between 0.1 and 0.2, we can correctly estimate the true graph $\delta_0$. More details on these results and their interpretation are provided in Section 6.

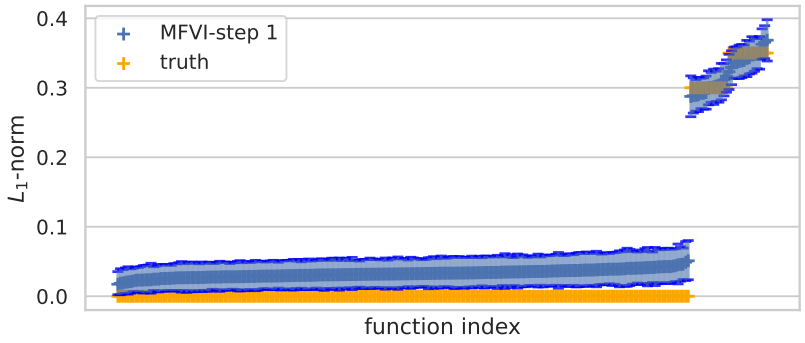



Figure 1: Estimated $L_1$-norms $(\hat{S}_{(i)})_{i\in[K^2]}$ and 95% credible sets (in blue ), based on the mean-field adaptive variational posterior mean and the set of models $\mathcal{M}_C$ containing models with complete graph $\delta_C = \mathbb{1}\mathbb{1}^T$, plotted in increasing order. The orange dots correspond to the true values $S^0_{lk} = \|h^0_{lk}\|_1$. These results correspond to one realisation of the *Excitation* scenario of Simulation 4, for the Hawkes processes with $K = 16$ dimensions.

After estimating the connectivity graph $\hat{\delta}$, we refine the estimation of the non-zero parameters in $f$ by computing the adaptive VB posterior on models with graph parameter $\hat{\delta}$. Intuitively, if the estimate $\hat{\delta}$ corresponds to the true graph, we should obtain a better estimate of $f$. In our numerical simulations, we empirically verify that the second step improves the estimates, in particular, in the higher-dimensional settings (see Appendix G.3).

Our adaptive two-step algorithm is summarised below:

1. *Complete graph VB*:

   (a) compute the adaptive VB posterior $\hat{Q}^C$ ((15) or (16)) associated to the set of models $\mathcal{M}_C$ with the complete graph $\delta_C = \mathbb{1}\mathbb{1}^T$

   $$\mathcal{M}_C := \{m_C = (\delta_C = \mathbb{1}\mathbb{1}^T, J = (J_k)_k);\ J_k \geqslant 1,\ \forall k\}, \tag{19}$$

   and compute the posterior mean estimates $\hat{S}_{lk} = \mathbb{E}^{\hat{Q}^C}[\|h_{lk}\|]$, $\forall (l,k) \in [K]^2$.

   (b) order the values $\hat{S}_{lk}$ in increasing order, say $\hat{S}_{(1)} \leqslant \hat{S}_{(2)} \leqslant \cdots \leqslant \tilde{S}_{(K^2)}$, and define $\hat{\delta}_{lk} = 1$ if and only if $\hat{S}_{lk} > \eta_0$, where $\eta_0$ is a threshold defined by the first gap between the credible sets of two consecutive estimates $\hat{S}_{(i)}, \hat{S}_{(i+1)}$, $i \in [K^2]$.

2. *Graph-restricted VB*: compute the adaptive VB posterior $\hat{Q}^{\hat{\delta}}$ ((15) or (16)) associated to the set of models $\mathcal{M}_E$ with $\delta = \hat{\delta}$:

$$\mathcal{M}_E := \{m_E = (\hat{\delta}, J = (J_k)_k);\ J_k \geqslant 1,\ \forall k\}. \tag{20}$$

We also note that different variants of our two-step strategy are possible. In particular one can choose a different threshold for each dimension $k \in [K]$, since convergence rates could differ across dimensions. Moreover, one can potentially remove the model selection procedure that chooses the number of dictionary functions $J_k$ in the first step 1(a), and compute a variational posterior in only one model $m \in \mathcal{M}_C$, e.g., one with the $J_k$'s sufficiently large. In recent work, Bonnet et al. (2021) also propose a thresholding approach for estimating the connectivity graph $\delta$ in the context of parametric maximum likelihood estimation. In fact, a variant of our procedure inspired by their work would consist in defining the graph estimator as $\hat{\delta}_{lk} = 1 \iff \hat{S}_{lk} > \varepsilon \sum_{l,k} \hat{S}_{lk}$, where $\varepsilon \in (0, 1)$ is a pre-defined or data-driven threshold.

In the next section, we present the algorithm to compute the adaptive VB posterior within a set of models $\mathcal{M}$ for the sigmoid Hawkes model.

## 4. Adaptive Algorithms in the Sigmoid Model

In this section, we focus on the *sigmoid* Hawkes model, for which the link functions in (1) are sigmoid-type functions. We consider the following parametrisation of this model: for each $k \in [K]$, let

$$\phi_k(x) = \theta_k \tilde{\sigma}(x), \quad \tilde{\sigma}(x) = \sigma\left(\alpha(x - \eta)\right), \quad \sigma(x) := (1 + e^{-x})^{-1}, \quad \alpha > 0,\ \eta > 0,\ \theta_k > 0. \tag{21}$$

Here, we assume that the hyperparameters $\alpha, \eta$ and $\theta = (\theta_k)_k$ are known; however, our methodology can be directly extended to estimate an unknown $\theta$, similarly to Zhou et al. (2022) and Malem-Shinitski et al. (2021). We first note that for $\alpha = 0.1, \eta = 10$ and $\theta_k = 20$, the nonlinearity $\phi_k$ is similar to the ReLU and softplus functions on $[-\infty, 20]$ (see Figure 2 in Section 6). This is helpful to compare the impact of the link functions on the inference in our numerical experiments in Section 6.

### 4.1 Inference in a Fixed Model

For a given model $m = (\delta, J)$, using a Gaussian-type of spike-and-slab prior distribution, the model-restricted mean-field variational posterior (14) can be computed via an iterative algorithm, in an almost identical fashion as Zhou et al. (2022), that we describe below.

We recall from this previous work that in the sigmoid model, a data augmentation strategy (described in more details in Appendix B), allows to recover some conjugacy between the prior and the mean-field variational posterior. In this case, the latent variables are $(\omega, \bar{Z})$ where $\omega = ((\omega_i^k)_{i\in[N_k]}; 1 \leqslant k \leqslant K)$ are random marks at each point of the point process $N$ with Polya-Gamma prior density $p_{PG}(\cdot; 1, 0)$ (defined in (46)), and $\bar{Z} = ((\bar{\omega}_j^k, \bar{T}_j^k)_{j\in[\bar{Z}_k]}); 1 \leqslant k \leqslant K)$ is $K$-dimensional marked Poisson point process on $[0, T] \times \mathbb{R}^+$ with prior intensity measure $\Lambda^k(t, \omega) = \theta_k p_{PG}(\omega; 1, 0)$. For a parameter $f_m$ and latent variable $(\omega, \bar{Z})$, the augmented data log-likelihood in the sigmoid

model is hence expressed as

$$L_T(f_m, \omega, \bar{Z}; N) = \sum_{k\in[K]} \left\{ \sum_{i\in[N_k]} \left[ \log\theta_k + g(\omega_i^k, \tilde{\lambda}^k_{T_i^k}(f_m)) + \log p_{PG}(\omega_i^k; 1, 0) \right] \right.$$
$$\left. + \sum_{j\in[\bar{Z}_k]} \left[ \log\theta_k + g(\bar{\omega}_j^k, -\tilde{\lambda}^k_{\bar{T}_j^k}(f_m)) + \log p_{PG}(\bar{\omega}_j^k; 1, 0) \right] - \theta_k T \right\}. \tag{22}$$

with $g(\omega, x) = -\frac{\omega x^2}{2} + \frac{x}{2} - \log 2$. We note that since the $\tilde{\lambda}^k_t(f_m)$ is linear in the parameter $f_m$, each coordinate of $f$ appears in a quadratic term in the previous expression. We recall that, denoting by $\mathbb{P}_A$ the prior distribution on the latent variable space $\mathcal{O}\times\mathcal{Z}$, the augmented posterior distribution is then proportional to $\Pi_A(f_m, \omega, \bar{Z}|N) \propto L_T(f_m, \omega, \bar{Z}; N)\Pi(f_m)\mathbb{P}_A(\omega, \bar{Z})$, and the (augmented) mean-field variational posterior distribution is defined as

$$\hat{Q}^m_{AMF}(f_m, \omega, \bar{Z}) = \arg\min_{Q^m\in\mathcal{V}^m_{AMF}} KL(Q^m \| \Pi_A(\cdot|N)) = \hat{Q}^m_1(f_m)\hat{Q}^m_2(\omega, \bar{Z}),$$

where $\mathcal{V}^m_{AMF} = \left\{ Q : \mathcal{F}_m \times \mathcal{O}\times\mathcal{Z} \to \mathbb{R}^+;\ dQ(f_m, \omega, \bar{Z}) = dQ_1(f_m)dQ_2(\omega, \bar{Z}) \right\}$ is the model-restricted mean-field variational family. It is well known that the mean-field variational posterior then verifies

$$\hat{Q}^m_1(f_m) \propto \exp\left\{ \mathbb{E}_{\hat{Q}^m_2}[\log p(f_m, z, N)] \right\}, \quad \hat{Q}_2(z) \propto \exp\left\{ \mathbb{E}_{\hat{Q}^m_1}[\log p(f_m, z, N)] \right\}, \tag{23}$$

see for instance Blei et al. (2017).

Next, recalling our notation from Section 2.3, we introduce a family of Gaussian prior distributions $\Pi_{h|\delta,J}(h)$ on $\mathcal{F}_m$ which results in a Gaussian form for $\hat{Q}^m_1$ and which allows an iterative variational inference algorithms with closed-forms updates, using (23). With $|J| = \sum_k J_k$, we define

$$\mathcal{H}^J_e = \left\{ h = (h_{lk})_{l,k} \in \mathcal{H};\ h_{lk}(x) = \sum_{j=1}^{J_k} h^j_{lk} e_j(x),\ x\in[0, A],\ \underline{h}^{J_k}_{lk} = (h^1_{lk}, \dots, h^{J_k}_{lk}) \in \mathbb{R}^{J_k},\ \forall (l,k) \in [K]^2 \right\},$$

the subset of $\mathcal{F}$ of functions decomposed with $J$ dictionary functions $(e_j)_j$. For each $(l, k)$,

- if $\delta_{lk} = 1$, we consider a normal prior distribution on the vector $\underline{h}^{J_k}_{lk}$, with mean $\mu_{J_k} \in \mathbb{R}^{J_k}$ and covariance matrix $\Sigma_{J_k} \in \mathbb{R}^{J_k\times J_k}$, i.e., $\underline{h}^{J_k}_{lk} \sim \mathcal{N}(\mu_{J_k}, \Sigma_{J_k})$;
- otherwise if $\delta_{lk} = 0$, we set $\underline{h}^{J_k}_{lk} = \mathbf{0}_{J_k}$.

We then denote by $I_k(\delta) = \{l \in [K]; \delta_{lk} = 1\}$ for each $k$ and by $\mu_m = (\mu^m_k)_k$ with $\mu^m_k = (\mu_{J_k})_{l\in I_k(\delta)} \in \mathbb{R}^{|I_k(\delta)|J_k}$ and $\Sigma_m = Diag((\Sigma^m_k)_k)$ with $\Sigma^m_k = Diag((\Sigma_{J_k})_{l\in I_k(\delta)}) \in \mathbb{R}^{|I_k(\delta)|J_k\times|I_k(\delta)|J_k}$. We also set $\Pi_\nu$, the prior distribution on the background rates, to be a Gaussian distribution and for each $k$, $\nu_k \overset{i.i.d}{\sim} \mathcal{N}(\mu_\nu, \sigma^2_\nu)$ with $\mu_\nu, \sigma_\nu > 0$. With a slight abuse of notation, we denote by $f^m_k = (\nu_k, (\underline{h}^{J_k}_{lk})_{l\in I_k(\delta)}) \in \mathbb{R}^{|I_k(\delta)|J_k+1}$ the vector of non-null parameter for dimension $k$ in model $m$. In the next proposition, we provide analytical expressions of the augmented variational posterior with the above prior construction.

**Proposition 3** *Given the previous Gaussian prior, each factor of the variational distribution* $\hat{Q}_1^m(f_m) = \prod_k \hat{Q}_1^{m,k}(f_k^m)$*, is a Gaussian distribution with mean vector* $\tilde{\mu}_k^m \in \mathbb{R}^{|I_k(\delta)|J_k+1}$ *and covariance matrix* $\tilde{\Sigma}_k^m \in \mathbb{R}^{(|I_k(\delta)|J_k+1)\times(|I_k(\delta)|J_k+1)}$*, i.e.,* $\hat{Q}_1^{m,k}(f_k^m) = \mathcal{N}(\tilde{\mu}_k^m, \tilde{\Sigma}_k^m)$ *with*

$$\tilde{\Sigma}_k^m = \left[\alpha^2 \sum_{i\in[N_k]} \mathbb{E}_{\hat{Q}_2^{m,k}}[\omega_i^k] H(T_i^k) H(T_i^k)^T + \alpha^2 \int_0^T \int_0^{+\infty} \bar{\omega}^k H(t) H(t)^T \Lambda^k(t, \bar{\omega}^k) d\bar{\omega}^k dt + (\Sigma_k^m)^{-1}\right]^{-1}, \tag{24}$$

$$\tilde{\mu}_k^m = \frac{1}{2}\tilde{\Sigma}_k^m \left[\alpha \sum_{i\in[N_k]} (2\mathbb{E}_{\hat{Q}_2^{m,k}}[\omega_i^k]\alpha\eta + 1) H(T_i^k) + \alpha \int_0^T \int_0^{+\infty} \left(2\bar{\omega}^k \alpha\eta - 1\right) H(t)\Lambda^k(t, \bar{\omega}^k) d\bar{\omega}^k dt + 2(\Sigma_k^m)^{-1}\mu_k^m\right], \tag{25}$$

*where* $N_k := N^k[0, T]$ *and*

$$\Lambda^k(t, \bar{\omega}) := \theta_k \frac{\exp\left\{-\frac{1}{2}\mathbb{E}_{\hat{Q}_1^{m,k}}[\tilde{\lambda}_t^k(f_k^m)]\right\}}{2\cosh\frac{c_t^k}{2}} p_{PG}(\bar{\omega}; 1, c_t^k), \quad c_t^k := \sqrt{\mathbb{E}_{\hat{Q}_1^{m,k}}[\tilde{\lambda}_t^k(f_k^m)^2]}$$
$$H(t) := (H^0(t), H^1(t), \ldots, H^K(t)) \in \mathbb{R}^{|J|+1}$$
$$H^0(t) = 1, H^k(t) := (H_j^k(t))_{j=1,\ldots,J_k},\ k \in [K]$$
$$H_j^k(t) := \int_{t-A}^t e_j(t-s) dN_s^k, \quad j \in [J_k], k \in [K].$$

*Moreover, we also have that* $\hat{Q}_2^m(\omega, \bar{Z}) = \hat{Q}_{21}^m(\omega)\hat{Q}_{22}^m(\bar{Z})$ *with* $\hat{Q}_{21}^m(\omega) = \prod_k \prod_{i\in[N_k]} p_{PG}(\omega_i^k; 1, c_{T_i^k}^k)$ *and* $\hat{Q}_{22}^m(\bar{Z}) = \prod_k \hat{Q}_{22}^{m,k}(\bar{Z}^k)$ *where for each k,* $\hat{Q}_{22}^{m,k}$ *is the probability distribution of a marked Poisson point process on* $[0, T] \times \mathbb{R}^+$ *with intensity measure* $\Lambda^k(t, \bar{\omega})$*.*

The proof of Proposition 3 is provided in Appendix C.1 and relies on standard computation (see for instance Donner and Opper (2018) and Zhou et al. (2021)). Note that intuitively, $H_j^k(t)$ is the influence at time $t$ from a point at the $k$-th dimension due to the $j$-th basis function, and that both in (24) and (25), the first term essentially comes from the original data likelihood, the second term mainly comes from the distribution of the augmented variables (which depends on the parameter) and the last term corresponds to the contribution of the prior on the parameters.

From Proposition 3, one can see that $\hat{Q}_2^m$ can be computed if $\hat{Q}_1^m$ is known and vice-versa. Therefore, to compute $\hat{Q}^m$, we use an iterative algorithm that initialises $\hat{Q}_1^m$ and $\hat{Q}_2^m$ at the prior distribution ($\Pi$ and $\mathbb{P}_A$), and then updates each factor $\hat{Q}_1^m$ and $\hat{Q}_2^m$ alternatively. This procedure is summarised in following Algorithm 1.

We note that the updates of the mean vectors and covariance matrices require to compute an integral, which we perform using the Gaussian quadrature method (Golub and Welsch, 1969), where the number of points, denoted $n_{GQ}$, is a hyperparameter of our method. We also recall that under the previous factorisable Gaussian prior, each variational factor $\hat{Q}_k^m$ of $Q^m$ only depends on a subset of the parameter $f_k$, and hence, of the sub-model, $m_k := (\delta_k, J_k)$ and can hence be computed independently and in parallel.

**Remark 4** *The number of iterations $n_{iter}$ in Algorithm 1 is another hyperparameter of our method. In practice, we implement an early-stopping procedure, where we set a maximum number of iterations, such as 100, and stop the algorithm whenever the increase of the ELBO is small, e.g., lower than $10^{-3}$, indicating that the algorithm has converged.*

**Remark 5** *Similarly to Zhou et al. (2021); Malem-Shinitski et al. (2021), we can also derive analytic forms of the conditional distributions of the augmented posterior* (49)*. Therefore, the latter could be computed via a Gibbs sampler, which is provided in Algorithm 4 in Appendix F. However, this Gibbs sampler requires to sample a K-dimensional inhomogeneous Poisson point process (the latent variable) and is therefore computationally much slower than the variational algorithm, which only requires to compute* expectations *wrt to the latent variables distribution.*

**Algorithm 1:** Mean-field variational inference algorithm in a fixed model

**Input:** $N = (N^1, \dots, N^K)$, $m = (\delta, J)$, $J = (J_1, \dots, J_K)$, $\mu_m = (\mu_k^m)_k, \Sigma_m = (\Sigma_k^m)_k$, $n_{iter}$, $n_{GQ}$.
**Output:** $\tilde{\mu}_m = (\tilde{\mu}_k^m)_k, \tilde{\Sigma}_m = (\tilde{\Sigma}_k^m)_k$.

Precompute $(H(T_i^k))_{i,k}$.
Precompute $(p_q, v_q)_{q\in[n_{GQ}]}$ (points and weights for Gaussian quadrature) and $(H(p_q))_{q\in[n_{GQ}]}$ .
**do in parallel for each** $k = 1, \dots, K$
  Initialisation: $\tilde{\mu}_k^m \leftarrow \mu_k^m, \tilde{\Sigma}_k^m \leftarrow \Sigma_k^m$.
  **for** $t \leftarrow 1$ *to* $n_{iter}$ **do**
    **for** $i \leftarrow 1$ *to* $N_k$ **do**
      $\mathbb{E}_{\hat{Q}_1^{m,k}}[\tilde{\lambda}^k_{T_i^k}(f_k^m)^2] = \alpha\left(H(T_i^k)^T\tilde{\Sigma}_k^s H(T_i^k) + (H(T_i^k)^T\tilde{\mu}_k^s)^2 - 2\eta H(T_i^k)^T\tilde{\mu}_k^m + \eta^2\right)$
      $\mathbb{E}_{\hat{Q}_2^{m,k}}[\omega_i^k] = \tanh\left(\sqrt{\mathbb{E}_{\hat{Q}_1^{m,k}}[\tilde{\lambda}^k_{T_i^k}(f_k^s)^2]}\right) / \left(2\sqrt{\mathbb{E}_{\hat{Q}_1^{m,k}}[\tilde{\lambda}^k_{T_i^k}(f_k^m)^2]}\right)$
    **for** $q \leftarrow 1$ *to* $n_{GQ}$ **do**
      $\mathbb{E}_{\hat{Q}_1^{m,k}}[\tilde{\lambda}^k_{p_q}(f_k^m)^2] = \alpha\left(H(p_q)^T\tilde{\Sigma}_k^m H(p_q) + (H(p_q)^T\tilde{\mu}_k^m)^2 - 2\eta H(p_q)^T\tilde{\mu}_k^m + \eta^2\right)$
      $\mathbb{E}_{\hat{Q}_2^{m,k}}[\omega_q^k] = \tanh\left(\sqrt{\mathbb{E}_{\hat{Q}_1^{m,k}}[\tilde{\lambda}^k_{p_q}(f_k^s)^2]}\right) / \left(2\sqrt{\mathbb{E}_{\hat{Q}_1^{m,k}}[\tilde{\lambda}^k_{p_q}(f_k^m)^2]}\right)$
      $\mathbb{E}_{\hat{Q}_1^{m,k}}[\tilde{\lambda}^k_{p_q}(f_k^m)] = \alpha\left((\tilde{\mu}_k^m)^T H(p_q) - \eta\right)$
    Compute $\tilde{\Sigma}_k^m$ and $\tilde{\mu}_k^m$ using (24) and (25)

### 4.2 Adaptive Algorithms

In this section, we present our two adaptive variational algorithms which implements the model-selection and two-step approach from Section 3 for the sigmoid Hawkes model leveraging Algorithm 1. Our first algorithm, denoted *fully-adaptive*, computes the VB posterior (15) or (16), and is suitable for settings where the number of dimensions $K$ is small or moderately large (and all $2^K$ graphical models can be explored). The second algorithm, denoted *two-step adaptive*, relies on a partial model-selection strategy and the two-step approach from Section 3.2, and is more efficient in settings where $K$ is large.

#### 4.2.1 Fully-Adaptive Variational Algorithm

To implement our model selection procedure, we consider that the maximum number of functions $(e_j)$ in the dictionary used to estimate the functions $h_{lk}$ is $J_T \in \mathbb{N}$. We then consider the set of models

$$\mathcal{M}_T = \Big\{ m = (\delta, J = (J_k)_k);\ \delta \in \{0,1\}^{K\times K},\ 1 \leqslant J_k \leqslant J_T,\ k \in [K] \Big\}. \tag{26}$$

We can easily see that in this case $|\mathcal{M}_T| \sim 2^{K^2} J_T$, and that for any $m = (\delta, J) \in \mathcal{M}_T$, the number of parameters in $m$ is equal to $\sum_{l,k} \delta_{lk}(J_k + 1) + 1$. Therefore, exploring all models in $\mathcal{M}_T$ is only computationally feasible for low-dimensional settings. We also further specify $\Pi_m(m)$, the prior distribution on $\mathcal{M}_T$. We consider

$$\Pi_m(m) = \prod_k \Pi_m(m_k) = \prod_k \Pi_{k,\delta}(\delta_{\cdot k})\Pi_{k,J}(J_k), \quad m_k = (\delta_{\cdot k}, J_k),\ \forall k.$$

Without prior information on the dependence structure of the process, one can choose $\Pi_{k,\delta}$ for instance as a product of Bernoulli distribution with parameter $p \in (0,1)$ and set $\Pi_{k,J}$ as the uniform distribution over $[J_T]$.

Using Algorithm 1, for each $m = (m_k)_k$, we compute $\hat{Q}_k^m$ together with the corresponding $ELBO(\hat{Q}_k^m)$ for each $k$. Then we either select the optimal model $\hat{m}_k = \arg\max_{m_k} ELBO(\hat{Q}_k^m)$ and compute the model-selection VB posterior $\hat{Q}_{MS} = \otimes_{k=1}^K \hat{Q}_k^{\hat{m}_k}$, or we compute the model-averaging VB posterior $\hat{Q}_{MA} = \otimes_{k=1}^K \hat{Q}_{k;MA}$ with

$$\hat{Q}_{k;MA} = \sum_{m_k} \hat{\gamma}_k^m \hat{Q}_k^m, \quad \hat{\gamma}_k^m = \frac{\tilde{\gamma}_k^m}{\sum_m \tilde{\gamma}_k^m} \quad \tilde{\gamma}_k^m = \Pi_{k,\delta}(\delta_{\cdot k})\Pi_{k,J}(J_k) \exp\left\{ ELBO(\hat{Q}_k^m) \right\}. \tag{27}$$

This procedure is summarised in Algorithm 2 and we call it the *fully-adaptive mean-field variational inference* algorithm.

**Algorithm 2:** Fully-adaptive mean-field variational inference

**Input:** $N = (N^1, \ldots, N^K)$, $\mathcal{M}_T$, $\mu = (\mu_m)_{m\in\mathcal{M}_T}$, $\Sigma = (\Sigma_m)_{m\in\mathcal{M}_T}$, $n_{iter}$, $n_{GQ}$.
**Output:** $\hat{Q}_{MA}$ or $\hat{Q}_{MS}$.

**do in parallel for each** $m = (\delta, D) \in \mathcal{M}_T$
  Compute the variational posterior $\hat{Q}_m$ using Algorithm 1 with $\mu_m, \Sigma_m$, $n_{iter}$ and $n_{GQ}$ as hyperparameters.
  Compute $(ELBO(\hat{Q}_k^m))_k$ and $(\tilde{\gamma}_k^m)_k$ using (27).
Compute $\{\hat{\gamma}^m = (\hat{\gamma}_k^m)_k\}_{m\in\mathcal{M}_T}$ and $\hat{Q}_{MA}$ or $\hat{Q}_{MS}$.

#### 4.2.2 Two-step Adaptive Mean-Field Algorithm

For settings with moderately large to large values of $K$, we instead use an algorithm based on the two-step approach introduced in Section 3.

We recall that in the latter strategy, we start with a maximal graph $\delta_C$, typically the complete graph $\delta_C = \mathbb{1}\mathbb{1}^T$, and considering the set of models

$$\mathcal{M}_C = \{ m = (\delta_C, J = (J_k)_k);\ 1 \leqslant J_k \leqslant J_T,\ k \in [K] \},$$

where here as well we assume that the number of functions in the dictionary is bounded by $J_T$. In the first step of our fast algorithm, we compute the model-selection adaptive VB posterior $\hat{Q}^C_{MS}$ using Algorithm 2, replacing $\mathcal{M}_T$ by $\mathcal{M}_C$. We note that with $\mathcal{M}_C$, the set of models explored per dimension is $J_T$, hence the optimisation procedure over this set is feasible, even for large values of $K$ (as soon as the computation for each model is fast and $J_T$ is not too large).

Then, in a second step, we use $\hat{Q}^C_{MS}$ to estimate the $L_1$-norms $(\|h_{lk}\|_1)_{l,k}$ and the graph parameter $\hat{\delta}$, with the thresholding method described in Section 3. Next, we consider the second set of models

$$\mathcal{M}_E = \left\{ m = (\hat{\delta}, J = (J_k)_k);\ 1 \leqslant J_k \leqslant J_T,\ k \in [K] \right\},$$

which has the same cardinality as $\mathcal{M}_C$, and compute the adaptive model-selection VB posterior $\hat{Q}_{MS}$ or model-averaging VB posterior $\hat{Q}_{MA}$ using Algorithm 2, replacing $\mathcal{M}_T$ by $\mathcal{M}_E$. This procedure is summarised in Algorithm 3.

In the next section, we provide theoretical guarantees for general variational Bayes approaches, and apply them to our adaptive and mean-field algorithms.

**Algorithm 3:** Two-step adaptive mean-field variational inference

**Input:** $N = (N^1, \dots, N^K)$, $\mathcal{M}_T$, $\mu = (\mu_m)_m, \Sigma = (\Sigma_m)_m$, $n_{iter}$, $n_{GQ}$.
**Output:** $\hat{Q}_{MS}$ or $\hat{Q}_{MA}$

1 Compute $\hat{Q}_{MS}$ using Algorithm 2 with input set $\mathcal{M}_C$ and hyperparameters $\mu = (\mu_m)_m, \Sigma = (\Sigma_m)_m$, $n_{iter}$, $n_{GQ}$.
2 Compute $\hat{\delta}$ using the thresholding of the estimate $\hat{S}$.
3 Compute $\hat{Q}^C_{MS}$ or $\hat{Q}_{MA}$ using Algorithm 2 with input set $\mathcal{M}_E$ and hyperparameters $\mu = (\mu_m)_m, \Sigma = (\Sigma_m)_m$, $n_{iter}$, $n_{GQ}$.

## 5. Theoretical Properties of the Variational Posteriors

In this section, we establish general results on variational Bayes methods for estimating the parameter of Hawkes processes, as well as guarantees for our adaptive and mean-field approaches proposed in Sections 2 and 4. In particular, we derive the concentration rates of variational Bayes posterior distributions, under general conditions on the model, the prior distribution, and the variational family.

We recall that in our problem setting, we assume that the observations $N$ are distributed according to a nonlinear Hawkes process with true parameter $f_0$ (unknown) but the link functions $\phi := (\phi_k)_k$ in the nonlinear intensity (1) are fixed by the statistician and therefore known *a-priori*. Throughout the section we assume that these functions are monotone non-decreasing, $L$-Lipschitz, $L > 0$, and that one of the two following conditions is satisfied:

**(C1)** For a parameter $f = (\nu, h) \in \mathcal{F}$, the matrix defined by $\rho^+ = (\rho^+_{lk})_{l,k} \in \mathbb{R}_+^{K \times K}$ with $\rho^+_{lk} = L \left\| h^+_{lk} \right\|_1$, $\forall l, k$, satisfies $\left\| \rho^+ \right\| < 1$;

**(C2)** For any $k \in [K]$, the link function $\phi_k$ is bounded, i.e., $\exists \Lambda_k > 0, \forall x \in \mathbb{R}, 0 \leqslant \phi_k(x) \leqslant \Lambda_k$.

These assumptions are sufficient conditions to have existence of a stationary version of the Hawkes process (see for instance Brémaud and Massoulié (1996), Deutsch and Ross (2022), or Sulem et al. (2024)).

### 5.1 Variational Posterior Concentration Rates

In this section, we state our general concentration result on the VB posterior distribution. This result relies on the concentration properties of the (exact) posterior distribution for the nonlinear Hawkes model. In particular, the proof of our main theorem, namely subsequent Theorem 7, uses an intermediate result from Sulem et al. (2024), and also relies on theoretical properties of variational posteriors established by Ray and Szabó (2021) and Nieman et al. (2022).

Before stating our result, we introduce an assumption, also used to prove the concentration of the posterior distribution (4) in the nonlinear Hawkes model in Sulem et al. (2024).

**Assumption 6** *We assume that there exist $\varepsilon > 0$ and $L' > 0$ such that for all $f \in \mathcal{F}$, for each $k \in [K]$, the link function $\phi_k$ restricted to the interval $I_k = (\nu_k - \max_{l\in[K]} \left\|h_{lk}^-\right\|_\infty - \varepsilon, \nu_k + \max_{l\in[K]} \left\|h_{lk}^+\right\|_\infty + \varepsilon)$ is bijective from $I_k$ to $J_k = \phi_k(I_k)$ and its inverse is $L'$- Lipschitz on $J_k$. We also assume that at least one of the two following conditions is satisfied.*

*(i) For any $k \in [K]$, $\inf_{x\in\mathbb{R}} \phi_k(x) > 0$.*

*(ii) For any $k \in [K]$, $\phi_k > 0$, and $\sqrt{\phi_k}$ and $\log \phi_k$ are $L_1$-Lipschitz with $L_1 > 0$ .*

This assumption is verified for commonly used link functions and practical parameter spaces $\mathcal{F}$ (see Example 1 in Sulem et al. (2024)). In particular, it holds for sigmoid-type link functions, such as the ones considered in Section 4, when the parameter space is included into a ball with finite radius (see below). We generally define our parameter space $\mathcal{F}$ as follows

$$\mathcal{H}' = \{h : [0, A] \to \mathbb{R};\ \|h\|_\infty < \infty\}, \quad \mathcal{H} = \left\{h = (h_{lk})_{l,k=1}^K \in \mathcal{H}'^{K^2};\ (h, \phi) \text{ satisfy } \textbf{(C1)} \text{ or } \textbf{(C2)}\right\},$$
$$\mathcal{F} = \left\{f = (\nu, h) \in (\mathbb{R}_+\backslash\{0\})^K \times \mathcal{H};\ (f, \phi) \text{ satisfies Assumption } 6\right\}.$$

While $\mathcal{H}'$ is a space of bounded functions (in supremum norm) with support in $[0, A]$, the space $\mathcal{H}$ is a composite parameter space where the parameter is made of $K^2$ functions (each of them belonging to $\mathcal{H}'$). We also define the $L_1$-distance for any $f, f' \in \mathcal{F}$ as

$$\|f - f'\|_1 := \left\|\nu - \nu'\right\|_1 + \left\|h - h'\right\|_1, \quad \left\|h - h'\right\|_1 := \sum_{l,k=1}^K \left\|h_{lk} - h'_{lk}\right\|_1, \quad \left\|\nu - \nu'\right\|_1 := \sum_k |\nu_k - \nu'_k|. \tag{28}$$

For the sigmoid-type link function $\phi_k(x) = \theta_k\sigma(\alpha(x-\eta))$, we define $\mathcal{F}_C = \left\{f = (\nu, h) \in (-C, C)^K \times \mathcal{H}_C\right\}$, with $\mathcal{H}_C = \{h \in \mathcal{H};\ \|h_{lk}\|_\infty < C, \forall l, k\}$ and $C > 0$, and in this case $\phi_k^{-1}$ is $L'_C$-Lipschitz on $J_k = [\phi_k(-2C), \phi_k(2C)]$ with $L'_C = (\alpha\theta_k\sigma(\alpha(-2C - \eta))(1 - \sigma(\alpha(2C - \eta))))^{-1}$. Finally we introduce

$$B_\infty(\epsilon) = \left\{f \in \mathcal{F};\ \nu_k^0 \leqslant \nu_k \leqslant \nu_k^0 + \epsilon,\ h_{lk}^0 \leqslant h_{lk} \leqslant h_{lk}^0 + \epsilon,\ (l, k) \in [K]^2\right\}, \quad \epsilon > 0,$$

a neighbourhood around $f_0$ in supremum norm, and a sequence $(\kappa_T)_T$ defined as

$$\kappa_T := 10(\log T)^r, \tag{29}$$

with $r = 0$ if $(\phi_k)_k$ satisfies Assumption 6 (i), and $r = 1$ if $(\phi_k)_k$ satisfies Assumption 6 (ii).

**Theorem 7** *Let $N$ be a Hawkes process with link functions $\phi = (\phi_k)_k$ and parameter $f_0 = (\nu_0, h_0)$ such that $(\phi, f_0)$ satisfy Assumption 6 and **(C1)** or **(C2)**. Let $\epsilon_T = o(1/\sqrt{\kappa_T})$ be a positive sequence verifying $\log^3 T = O(T\epsilon_T^2)$, $\Pi$ be a prior distribution on $\mathcal{F}$, and $\mathcal{V}$ a variational family of distributions on $\mathcal{F}$. We assume that the following conditions are satisfied for $T$ large enough.*

***(A0)** There exists $c_1 > 0$ such that $\Pi(B_\infty(\epsilon_T)) \geqslant e^{-c_1 T\epsilon_T^2}$.*

***(A1)** There exist $M_0, c_2 > 0$, $\epsilon_0 \in (0, 1]$ such that for any $\epsilon \in [M_0\sqrt{\kappa_T}\epsilon_T, \epsilon_0]$ there exist $\mathcal{H}(\epsilon) \subset \mathcal{H}$, $\zeta_0 > 0$, and $x_0 > 0$ such that*

$$\Pi(\mathcal{H}^c(\epsilon)) = o(e^{-c_2 T\epsilon^2}) \quad \text{and} \quad \log C\left(\zeta_0\epsilon, \mathcal{H}(\epsilon), \|\cdot\|_1\right) \leqslant x_0 T\epsilon^2.$$

*where $\mathcal{H}^c(\epsilon)$ denotes the complement of $\mathcal{H}(\epsilon)$ in $\mathcal{H}$.*

***(A2)** There exists $Q \in \mathcal{V}$ such that $supp(Q) \subset B_\infty(\epsilon_T)$ and $KL(Q\|\Pi) = O(\kappa_T T\epsilon_T^2)$.*

*Then, for any $M_T \to \infty$ and $\hat{Q}$ defined in (9), we have that*

$$\mathbb{E}_0\left[\hat{Q}\left(\|f - f_0\|_1 > M_T\sqrt{\kappa_T}\epsilon_T\right)\right] \xrightarrow[T\to\infty]{} 0$$

$$\mathbb{P}_0\left(\sum_{l,k=1}^{K}\left|\mathbb{E}_{\hat{Q}}[\|h_{lk}\|_1] - \|h_{lk}^0\|_1\right| > M_T\sqrt{\kappa_T}\epsilon_T\right) \xrightarrow[T\to\infty]{} 0$$

The proof of Theorem 7 is reported in Appendix D.2 and leverages the proof on the posterior concentration rates for the nonlinear Hawkes model from Sulem et al. (2024). It also uses an adaptation of Theorem 5 of Ray and Szabó (2021) and Lemma 13 in Nieman et al. (2022) on the theory of variational posteriors for the first result, together with the approach of Zhang and Gao (2020) for the second. We also note that the second statement also holds if one replaces $\mathbb{E}_{\hat{Q}}[\|h_{lk}\|_1]$ by the variational posterior median estimator of $\|h_{lk}\|_1$, and more generally any quantile. The latter can therefore be an alternative estimator of the norms for our two-step procedure.

We make a few remarks related to the previous theorem. Firstly, the role of the subsets $\mathcal{H}_T$ in Assumptions **(A0)** and **(A1)** is a technicality that arises in the construction of tests for establishing the concentration rates (see also for instance Donnet et al. (2020)).

Secondly, similarly to Donnet et al. (2020) and Sulem et al. (2024), Theorem 7 also holds when the assumptions on the prior **(A0)** and on the variational family **(A2)** are verified for neighborhoods around $f_0$ in $L_2$-norm. More precisely, one can replace $B_\infty(\epsilon_T)$ in Assumptions **(A0)** and **(A2)** by

$$B_2(\epsilon_T, B) = \left\{f \in \mathcal{F};\ \max_k |\nu_k - \nu_k^0| \leqslant \epsilon_T,\ \max_{l,k}\|h_{lk} - h_{lk}^0\|_2 \leqslant \epsilon_T,\ \max_l \nu_l + \max_k \|h_{kl}\|_\infty < B\right\},$$

with $B > 0$, but in this case $\kappa_T$ is also replaced by $\kappa_T' = 10(\log\log T)(\log T)^r$.

Thirdly, Theorem 7 also holds under the following more general condition on the variational family:

**(A2')** The variational family $\mathcal{V}$ verifies $\min_{Q\in\mathcal{V}} KL(Q\|\Pi(.|N)) = O(\kappa_T T\epsilon_T^2)$.

However, in practice, one often verifies **(A2)** and deduces **(A2')** using the following steps from Zhang and Gao (2020): for any $Q \in \mathcal{V}$, we have that

$$KL(Q\|\Pi(.|N)) \leqslant KL(Q\|\Pi) + Q(KL(\mathbb{P}_{T,f_0}, \mathbb{P}_{T,f})),$$

where we denote $\mathbb{P}_{T,f_0}$ (resp. $\mathbb{P}_{T,f}$) the distribution with density proportional to $e^{L_T(f_0)}$ (resp. $e^{L_T(f)}$) with respect to a homogeneous Poisson Process over $[0, T]$. Moreover, using Lemma S6.1 from

Sulem et al. (2024), for any $f \in B_\infty(\epsilon_T)$, we also have that

$$\mathbb{E}_0\left[L_T(f_0) - L_T(f)\right] \leqslant \kappa_T T \epsilon_T^2.$$

Therefore, under **(A2)**, there exists $Q \in \mathcal{V}$ such that $KL(Q\|\Pi(.|N)) = O(\kappa_T T\epsilon_T^2)$, which implies **(A2')**. We also note that **(A2)** (or **(A2')**), is the only condition on the variational class, and informally states that this family of distributions can approximate the true posterior conveniently.

Additionally, Assumptions **(A0)** and **(A1)** are sufficient conditions for proving that the posterior concentration rate is at least as fast as $\sqrt{\kappa_T}\epsilon_T$. They are closely related to the ones of Theorem 3.2 in Sulem et al. (2024). Nonetheless, here **(A1)** is a slightly stronger condition than the one needed for posterior concentration rates, and is used to prove the second statement of Theorem 7. This assumption is also related to the assumptions of Theorem 2.1 of Zhang and Gao (2020). In fact, for commonly used nonparametric priors, our assumption **(A1)** is verified (see Section 5.2).

Finally, for sigmoid-type link functions, we propose the analog of Theorem 7 when $\mathcal{F}$ is not included into a ball of finite radius. For this purpose, we replace the $L_1$-norm by the truncated $L_1$-norm

$$\|f - f_0\|_{1,C} = \|f^C - f_0\|_1, \quad f^C = (\nu, (h_{lk}^C)_{l,k}), \quad h_{lk}^C = (h_{lk} \vee -2C) \wedge 2C,$$

where $C > 0$ is any constant such that $C \geqslant \max_{l,k} \|h_{lk}^0\|_\infty$. Then the posterior and variational posterior concentration rate results hold for the truncated $L_1$-norm under the same assumptions. This is stated in the following proposition.

**Proposition 8** *Let $N$ be a Hawkes process with sigmoid link functions $\phi_k(x) = \theta_k\sigma(\alpha(x-\eta))$ and parameter $f_0 = (\nu_0, h_0)$. Let $\epsilon_T = o(1/\sqrt{\kappa_T})$ be a positive sequence verifying $\log^3 T = O(T\epsilon_T^2)$, $\Pi$ be a prior distribution on $\mathcal{F}$, and $\mathcal{V}$ a variational family of distributions on $\mathcal{F}$. We assume that Assumptions (A0)-(A2) from Theorem 7 are satisfied for $T$ large enough with $\kappa_T = 10\log T$. Then, for any $M_T \to \infty$ and $C > 0$ such that $C \geqslant \max_{l,k} \|h_{lk}^0\|_\infty$ we have*

$$\mathbb{E}_0\left[\Pi\left(\|f - f_0\|_{1,C} > M_T\sqrt{\kappa_T}\epsilon_T \middle| N\right)\right] \xrightarrow[T\to\infty]{} 0,$$

*and, with $\hat{Q}$ defined in* (9)*, we also have*

$$\mathbb{E}_0\left[\hat{Q}\left(\|f - f_0\|_{1,C} > M_T\sqrt{\kappa_T}\epsilon_T\right)\right] \xrightarrow[T\to\infty]{} 0$$
$$\mathbb{P}_0\left(\sum_{l,k=1}^K \left|\mathbb{E}_{\hat{Q}}[\|h_{lk}^C\|_1] - \|h_{lk}^0\|_1\right| > M_T\sqrt{\kappa_T}\epsilon_T\right) \xrightarrow[T\to\infty]{} 0$$

We note that in the above result, the constant $C$ can be set arbitrarily large, hence our result is only slightly weaker than Theorem 7. In particular in our numerical experiments in Section 6, we checked that all 95% credible sets of $h_{lk}$ are contained in $[-C_1, C_1]$, for $C_1$ not too large, so we could choose $C = C_1/2$ and we approximate the variational posterior mean on the truncated function norms $\mathbb{E}_{\hat{Q}}[\|h_{lk}^C\|_1]$ by the untruncated ones $\mathbb{E}_{\hat{Q}}[\|h_{lk}\|_1]$ to select the graph estimator. The proof of Proposition 8 is provided in Appendix E. We also note that in the context of Proposition 8 the second statement also holds if one replaces $\mathbb{E}_{\hat{Q}}[\|h_{lk}^C\|_1]$ by the variational posterior median estimator of the unclipped parameter $\|h_{lk}\|_1$, which therefore can be used as an alternative estimator to the clipped estimator. Finally Proposition 8 remains valid for other link functions that are similar to the sigmoid, i.e., bijective from $\mathbb{R}^+$ to an open and bounded interval.

### 5.2 Applications to Variational Classes and Prior Families of Interest

We now apply our general theorem to prove the validity of our proposed variational inference methods based on mean-field variational families and model-selection strategies, introduced in Section 3 and used in our algorithms for the sigmoid model in Section 4. We also verify our general conditions on the prior distribution on two examples of nonparametric prior families, namely random histograms and Gaussian processes, used for instance respectively by Donnet et al. (2020) and Malem-Shinitski et al. (2021). Similarly to Donnet et al. (2020) and Sulem et al. (2024), we also derive explicit concentration rates for the variational posterior distribution and for Hölder classes of functions.

Firstly, we recall from Section 2.3 that we consider a "spike-and-slab" prior distribution on $f$ which decomposes as

$$d\Pi(f) = d\Pi_\nu(\nu) d\Pi_\delta(\delta) d\Pi_{h|\delta}(h), \quad d\Pi_{h|\delta}(h) = \prod_{l,k} d\tilde{\Pi}_{h|\delta}(h_{lk}). \tag{30}$$

We also recall from Sulem et al. (2024) that for this prior, the prior mass condition **(A0)** of Theorem 7 can be replaced by the following assumption on $\Pi_\delta$ and $\Pi_{h|\delta}$:

**(A0')** There exists $c_1 > 0$ such that $\Pi(B_\infty(\epsilon_T)|\delta = \delta_0) \geqslant e^{-c_1 T\epsilon_T^2/2}$ and $\Pi_\delta(\delta = \delta_0) \geqslant e^{-c_1 T\epsilon_T^2/2}$.

For instance, one can choose that under the prior the $\delta_{lk}$'s are i.i.d. Bernoulli random variables with common mean $p \in (0,1)$, and $\Pi_\delta$ would then automatically verify the second part of Assumption **(A0')**. For the first part, $\Pi_{h|\delta}(B_\infty(\epsilon_T)|\delta = \delta_0) \geqslant e^{-c_1 T\epsilon_T^2/2}$, we show in the following that this can also be easily verified for the random histogram and Gaussian processes priors.

#### 5.2.1 Mean-Field Variational Family

In this section, we consider again the mean-field variational inference framework with latent variable augmentation introduced in Section 2.4. We recall that in this approach the mean-field family $\mathcal{V}_{AMF}$ is defined as

$$\mathcal{V}_{AMF} = \{Q : \mathcal{F} \times \mathcal{Z} \to [0,1];\ Q(f,z) = Q_1(f)Q_2(z)\},$$

where $z \in \mathcal{Z}$ is some latent variable. We also recall our notation $\mathbb{P}_A$, for the prior distribution on $z$, which is independent of the prior on $f$ and therefore leads to the augmented prior distribution $\Pi \times \mathbb{P}_A \in \mathcal{V}_{AMF}$. Hence, assumption **(A2)** is equivalent to the prior mass condition (see for instance Zhang and Gao (2020)).

Moreover, assumptions **(A0')** and **(A1)** are the same as in Sulem et al. (2024) and therefore can be applied to any prior family discussed there. In particular, priors on the $h_{lk}$'s based on decompositions on dictionaries like in (5) have been studied in Arbel et al. (2013) or Shen and Ghosal (2015) and their results can be applied to prove assumptions **(A0')** and **(A1)**. Below, we apply Theorem 7 for the mean-field family and random histogram priors, and a similar result for hierarchical Gaussian process priors is also reported in Appendix D.3.

Here, we consider the random histogram prior, i.e., specify $\Pi_{h|\delta}(h)$ as a prior distribution on piecewise-constant functions $h_{lk}$, using the decomposition (5) from Section 4.1. We note that this prior distribution is also similar to the basis decomposition prior in Zhou et al. (2022, 2021). For simplicity, we assume here that $J := J_1 = \cdots = J_k$ and consider a regular partition of $(0, A]$ based

on $(t_j)_{j=0,\dots,J}$ with $t_j = jA/J,\ j = 0,\dots,J$. We now re-write (5) as

$$h^w_{lk}(x) = \sum_{j=1}^{J} w^j_{lk} e_j(x), \quad e_j(x) = \frac{J}{A}\mathbb{1}_{(t_{j-1},t_j]}(x), \quad w^j_{lk} \in \mathbb{R} \quad \forall j \in [J], \forall l,k \in [K].$$

We note that $\|e_j\|_2 = \sqrt{J/A}$ but $\|e_j\|_1 = 1,\ \forall j \in [J]$, therefore, the functions of the dictionary, $(e_j)_j$ are orthonormal in terms of the $L_1$-norm. In this general construction, we can also consider a prior on the number of pieces $J$ with exponential tails, for instance, $J \sim \mathcal{P}(\lambda)$ with $\lambda > 0$, or $J = 2^D$ where $2^D \leqslant J_D < 2^{D+1}$ and $J_D \sim \mathcal{P}(\lambda)$. However, in our algorithms in Section 4, we consider for $\Pi_J$ a probability mass function on $[J_{max}]$ with $J_{max} \geqslant 1$. Finally, given $J$, we fully specify $\Pi_{h|\delta}(h)$ by considering a normal prior distribution on each weight $w^j_{lk}$, i.e.,

$$w^j_{lk}|J \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma_0^2), \quad \sigma_0 > 0 \quad \forall l,k \leqslant K, \quad j \leqslant J. \tag{31}$$

With this prior construction, assumptions **(A0')** and **(A1)** are easily checked. For instance, this Gaussian random histogram prior is a particular case of the spline prior family considered in Sulem et al. (2024), with a spline basis of order $q = 0$. We note that these conditions are also verified easily for other prior distributions on the weights, for instance, the shrinkage prior of Zhou et al. (2021) based on the Laplace distribution $p_{Lap}(w^j_{lk}; 0, b) = (2b)^{-1}\exp\{-|w^j_{lk}|/b\}$ with $b > 0$, and a "local" spike-and-slab prior inspired by the construction in Donnet et al. (2020):

$$w^j_{lk}|J \overset{\text{i.i.d.}}{\sim} p\delta_{(0)} + (1-p)p_{Lap}(.; 0, b), \quad p \in (0,1), \quad b > 0,$$

where $\delta_{(0)}$ is the Dirac measure at 0.

In the following proposition, we further assume that the true functions in $h_0$ belong to a Holder-smooth class of functions $\mathcal{H}(\beta, L_0)$ with $\beta \in (0,1)$, so that explicit variational posterior concentration rates $\epsilon_T$ for the mean-field family and the random histogram prior can be derived.

**Proposition 9** *Let $N$ be a Hawkes process with link functions $\phi = (\phi_k)_k$ and parameter $f_0 = (\nu_0, h_0)$ such that $(\phi, f_0)$ verify Assumption 6. Assume that for any $l,k \in [K]$, $h^0_{lk} \in \mathcal{H}(\beta, L_0)$ with $\beta \in (0,1]$ and $L_0 > 0$. Then, under the above Gaussian random histogram prior* (31) *with $J \asymp (T/\log T)^{1/(2\beta+1)}$, the mean-field variational distribution $\hat{Q}_1$ defined in* (11) *satisfies, for any $M_T \to +\infty$,*

$$\mathbb{E}_0\left[\hat{Q}_1\left(\|f - f_0\|_1 > M_T(\log T)^q (T/\log T)^{-\beta/(2\beta+1)}\right)\right] \xrightarrow[T\to\infty]{} 0,$$

*with $q = 0$ if $\phi$ verifies Assumption 6(i) and $q = 1/2$ if $\phi$ verifies Assumption 6(ii).*

The proof of Proposition 9 is omitted since it is a direct application of Theorem 7 to mean-field variational families in the context of a latent variable augmentation scheme. We note that the variational concentration rates also match the true posterior concentration rates (Sulem et al., 2024).

### 5.2.2 Adaptive Variational Family

In this section, we consider the model-selection adaptive variational posterior distributions (15) and (16), and similarly obtain their concentration rates.

We first note that these $\hat{Q}_{MS}$ and $\hat{Q}_{MA}$ correspond to the following variational families (see also Appendix A.2)

$$\mathcal{V}_{A1} = \cup_{m\in\mathcal{M}}\{\{m\}\times\mathcal{V}^m\}, \quad \mathcal{V}_{A2} = \left\{\sum_{m\in\mathcal{M}}\alpha_m Q_m; \sum_m \alpha_m = 1,\ \alpha_m \geqslant 0,\ Q_m \in \mathcal{V}^m,\ \forall m \in \mathcal{M}\right\},$$

where here, $\mathcal{M}$ is the set of all possible models, i.e.,

$$\mathcal{M} = \left\{m = (\delta, J = (J_1, \ldots, J_K));\ \delta \in \{0,1\}^{K\times K},\ J_k \in \mathbb{N},\ \forall k \in [K]\right\},$$

and for a model $m \in \mathcal{M}$, the variational family $\mathcal{V}^m$ corresponds to a set of distributions on the subspace $\mathcal{F}_m \times \mathcal{Z} \subset \mathcal{F} \times \mathcal{Z}$ and $\bigcup_{m\in\mathcal{M}}\mathcal{F}_m = \mathcal{F}$. We also recall that with $\mathcal{V}_{MS}$, the VB posterior is a distribution on the "optimal" model while with $\mathcal{V}_{MA}$, the variational posterior is a mixture. In this setting, the general results from Zhang and Gao (2020) can be applied, and, here, it is enough to replace the prior assumption **(A0)** by

$$\begin{aligned}\textbf{(A0'')} \quad &\exists c_1 > 0, \Pi\left(B_\infty(\epsilon_T)\,\middle|\delta = \delta_0, J = (J_T J_k^0)_k\right) \geqslant e^{-c_1 T\epsilon_T^2/3},\\ &\Pi_\delta(\delta = \delta_0) \geqslant e^{-c_1 T\epsilon_T^2/3}, \quad \Pi_J\left(J = \left(J_T J_k^0\right)_k\right) \geqslant e^{-c_1 T\epsilon_T^2/3},\end{aligned} \tag{32}$$

where $J_T = \left(\frac{T}{\log T}\right)^{1/(2\beta+1)}$ and the $J_k^0$'s are prior hyperparameters, assuming that, for any $l, k \in [K]$, $h_{lk}^0 \in \mathcal{H}(\beta, L_0)$. To see that **(A0'')** is enough, we note that it implies that

$$-\log\Pi(m = m_0) - \log\Pi\left(B_\infty(\epsilon_T)|m = m_0\right) \leqslant c_1 T\epsilon_T^2, \quad m_0 = \left(\delta_0, \left(J_T J_k^0\right)_k\right),$$

which also implies **(A0)**. For example, under the random histogram prior of Section 5.2.1, it is enough to choose $\Pi_J$ such that, for some sequence $(x_n)_{n\geqslant 1}$ such that $x_n \xrightarrow[n\to\infty]{} \infty$,

$$\Pi_J(J_l > x_n) \lesssim e^{-cx_n}, \quad \Pi_J(J_l = x_n) \gtrsim e^{-cx_n}, \quad \forall n \geqslant 1, \quad c > 0,$$

which is the case for instance when $\Pi_J$ is a Geometric distribution. In the next proposition, we state our result on the model-selection variational family, when using the random histogram prior distribution; however, this result also holds for other prior distributions based on decomposition over dictionaries such as the ones in Arbel et al. (2013); Shen and Ghosal (2015).

**Proposition 10** *Let $N$ be a Hawkes process with link functions $\phi = (\phi_k)_k$, parameter $f_0 = (\nu_0, h_0)$ such that $(\phi, f_0)$ verify Assumption 6. Assume that for any $l, k \in [K]$, $h_{lk}^0 \in \mathcal{H}(\beta, L_0)$ with $\beta \in (0,1]$ and $L_0 > 0$. Then, under the random histogram prior distribution, for the model selection variational posterior* (15) *, we have that, for any $M_T \to +\infty$, for $\hat{Q} = \hat{Q}_{MS}$ or $\hat{Q} = \hat{Q}_{MA}$,*

$$\begin{aligned}&\mathbb{E}_0\left[\hat{Q}\left(\|f - f_0\|_1 > M_T(\log T)^q (T/\log T)^{-\beta/(2\beta+1)}\right)\right] \xrightarrow[T\to\infty]{} 0,\\ &\mathbb{P}_0\left(\sum_{l,k=1}^K \left|\mathbb{E}_{\hat{Q}}[\|h_{lk}\|_1] - \|h_{lk}^0\|_1\right| > M_T(\log T)^q (T/\log T)^{-\beta/(2\beta+1)}\right) \xrightarrow[T\to\infty]{} 0\end{aligned} \tag{33}$$

*with $q = 0$ if $\phi$ verifies Assumption 6(i) and $q = 1/2$ if $\phi$ verifies Assumption 6(ii).*

Since Proposition 10 is a direct consequence of Theorem 7 and Theorem 4.1 in Zhang and Gao (2020) for $\hat{Q} = \hat{Q}_{MS}$ and of adapting Theorem 3.6 of Ohn and Lin (2024) for $\hat{Q} = \hat{Q}_{MA}$, its proof is omitted.

### 5.3 Convergence Rate Associated to the Two-step Algorithm

As previously discussed in Section 2.4, when the number of dimensions $K$ is large, both the distribution $\hat{Q}_{MA}$ and $\hat{Q}_{MS}$ are intractable in practice, due to the necessity of exploring all models in $\mathcal{M}_T$. For this setting, the two-step procedure proposed in Section 3.2 first constructs the estimator of the graph with (18) associated to a given threshold $\eta_0$, then constructs a restricted set of models $\mathcal{M}_E$ and computes the corresponding variational distribution $\hat{Q}^{\hat{\delta}}$. We now show that this two-step procedure is theoretically justified. We first prove that our thresholding graph estimator is consistent. We recall our notation $S^0_{lk} = \left\|h^0_{lk}\right\|_1$, $\hat{S}_{lk} = \mathbb{E}_{Q^c_{MS}}[\|h_{lk}\|_1]$, $\forall l,k \in [K]$ and we order them as $\hat{S}_{(1)} \leqslant \hat{S}_{(2)} \leqslant \cdots \leqslant \hat{S}_{(K^2)}$. We denote by $K_0$ the cardinal of $\mathcal{I}(\delta_0)$.

**Proposition 11** *Assume that the VB posterior $\hat{Q}$ based on the complete graph of interactions concentrates at the rate $\epsilon_T = o(1)$, i.e., there exists $c > 0$ such that*

$$\mathbb{P}_0\left[\sum_{l,k=1}^{K} |\hat{S}_{lk} - S^0_{lk}| > \epsilon_T\right] \xrightarrow[T\to\infty]{} 0. \tag{34}$$

*Assume also that $f_0$ is such that*

$$s_0 = \min_{(l,k)\in\mathcal{I}(\delta_0)} S^0_{lk} \geqslant 8\epsilon_T. \tag{35}$$

*Then, for any $\eta_0$ such that $2\epsilon_T \leqslant \eta_0 \leqslant s_0 - 2\epsilon_T$, we have that*

$$\mathbb{P}_0(\hat{\delta} \neq \delta_0) = o(1). \tag{36}$$

*Moreover, with probability going to 1,*

$$\forall i \leqslant K^2 - K_0 - 1, \hat{S}_{(i+1)} - \hat{S}_{(i)} \leqslant 2\epsilon_T, \quad \hat{S}_{(K^2-K_0+1)} - \hat{S}_{(K^2-K_0)} > s_0 - 4\epsilon_T \geqslant s_0/2,$$

*and as soon as $\epsilon_T = o(s_0)$,*

$$\max_{i\leqslant K^2-K_0-1} \left\{\hat{S}_{(i+1)} - \hat{S}_{(i)}\right\} = o_{P_0}(\hat{S}_{(K^2-K_0+1)} - \hat{S}_{(K^2-K_0)}).$$

**Remark 12** *The assumption* (34) *holds in particular under the assumptions of Theorem 7. Thus, if we further assume $s_0 \geqslant 8\epsilon_T$, the consistency result* (36) *holds.*

The first part of the result (36) shows that $\hat{\delta}$ is consistent under a wide range of threshold $\eta_0$ and the second part that as soon as $s_0$ is much larger than $\epsilon_T$, the first significant jump between $S_{(i)}$ and $S_{(i+1)}$ will take place at $i = K^2 - K_0$, and thus the data driven procedure leads to a consistent $\hat{\delta}$. The proof of Proposition 11 is reported in Appendix D.4.

We note that Proposition 11 applies in fact to any data dependent distribution $\hat{Q}$ and is not restricted to the variational posterior. It applies also to the full posterior or other types of approximations of the posterior as long as the $L_1$-concentration assumption (34) is verified. Note also that in Proposition 11, (35) is a mild requirement on $f_0$ since $\epsilon_T = o(1)$. The previous proposition demonstrates that in practice, the following two thresholding strategies for estimating the graph lead to consistent estimators:

(i) if a lower bound $0 < u \leqslant \min_{(l,k)\in I(\delta_0)} S^0_{lk}$ is known *a-priori*, we can fix the threshold $u > \eta_0 > u/2$ and compute $\hat{\delta} = (\hat{\delta}_{lk})_{l,k}$, $\hat{\delta}_{lk} = \mathbb{1}_{\{\hat{S}_{lk}>\eta_0\}}$, $\forall l,k$.

(ii) otherwise, we can choose a data-dependent threshold $\eta_0 \in (\hat{S}_{(i_0)}, \hat{S}_{(i_0+1)})$, where $(\hat{S}_{(i)})_{i\in[K^2]}$ corresponds to the values $(\hat{S}_{lk})_{l,k}$ in increasing order and $i_0$ is the first index such that $\hat{S}_{(i+1)} - \hat{S}_{(i)}$ is *large*, and we then compute $\hat{\delta}$ as in (i). In our numerical experiments, we consider that $\hat{S}_{(i+1)} - \hat{S}_{(i)}$ is large when the 95% credible sets constructed from $Q^C_{MS}$ around $\hat{S}_{(i+1)}$ and $\hat{S}_{(i)}$ do not intersect.

**Corollary 13** *Under the assumptions of either Proposition 10 or Proposition 9 or Proposition 20 and if condition* (35) *holds with* $\epsilon_T = (\log T)^q T^{-\beta/(2\beta+1)}$ *with* $q$ *given in the Propositions 10, 9 or 20, the adaptive variational posterior* $\hat{Q}_E$ *corresponding to the variational family* $\mathcal{M}_E := \{m_E = (\hat{\delta}, J = (J_k)_k);\ J_k \geqslant 1,\ \forall k\}$ *where* $\delta$ *is the thresholding estimator also concentrates at the rate* $\epsilon_T$ *in* $L_1$ *norm.*

The previous corollary is a direct consequence of Proposition 10 and Proposition 11 since under the event $\{\hat{\delta} = \delta_0\}$, the subspace $\cup_{m\in\mathcal{M}_E}\mathcal{F}_m$ contains the true parameter $f_0$.

## 6. Numerical Results

In this section, we perform a simulation study to evaluate our variational Bayesian method in the context of nonlinear Hawkes processes, and demonstrate its efficiency, scalability, and robustness in various estimation setups.

In low-dimensional settings ($K = 1$ and $K = 2$), we can compare our variational posterior to the true posterior distribution, obtained via an MCMC method. As a preliminary experiment, we analyse the performance of a Metropolis-Hastings sampler in commonly used nonlinear Hawkes processes, namely with ReLU, sigmoid and softplus link functions (Simulation 1). In the subsequent simulations, we focus on the sigmoid model and test our adaptive variational algorithms, in well-specified (Simulations 2-5) and mis-specified settings (Simulation 6), high-dimensional data sets, and for different connectivity graphs (Simulation 4).

In each setting, unless specified otherwise, we sample one observation of a Hawkes process with dimension $K$, link functions $(\phi_k)_k$ and parameter $f_0 = (\nu_0, h_0)$ on $[0, T]$, using the thinning algorithm of Adams et al. (2009). In most simulated settings, the true interaction functions $(h^0_{lk})_{l,k}$ will be piecewise-constant, and we use the random histogram prior described in Section 4.1 in our variational Bayes method. For $D \geqslant 1$, we introduce the notation

$$\mathcal{H}^D_{histo} = \left\{ h_k = (h_{lk})_l;\ h_{lk}(x) = \sum_{j=1}^{2^D} w^j_{lk} e_j(x),\ x \in [0, A],\ l \in [K],\ e_j(x) = \frac{2^D}{A}\mathbb{1}_{[\frac{jA}{2^D}, \frac{(j+1)A}{2^D})}(x) \right\},$$

and for the remaining of this section, we index functions $h_{lk}$ by the histogram depth $D$.

In the next sections, we report the results of the following set of simulations.

- **Simulation 1: Posterior distribution in parametric, univariate, nonlinear Hawkes models.** We analyse the posterior distribution computed from a Metropolis-Hasting sampler (MH) in several nonlinear univariate Hawkes processes ($K = 1$), with ReLU, sigmoid, and softplus link functions. For this sampler, we consider that the dimensionality $D_0$ such that $h_0 \in \mathcal{H}^{D_0}_{histo}$ is known, and therefore, the posterior inference is non-adaptive.

- **Simulation 2: Variational and true posterior distribution in parametric, univariate sigmoid Hawkes models.** In a univariate setting with $h_0 \in \mathcal{H}^{D_0}_{histo}$ and the dimensionality $D_0$ is

known (non-adaptive), we compare the variational posterior obtained from Algorithm 1 to the posterior distribution obtained from two MCMC samplers, i.e., the MH sampler of Simulation 1, and a Gibbs sampler available in the sigmoid model (Algorithm 4).

- **Simulation 3: Fully-adaptive variational algorithm in univariate and bivariate sigmoid models.** This experiment evaluates our first adaptive variational algorithm (Algorithm 2) in sigmoid Hawkes processes with $K = 1$ and $K = 2$, in nonparametric settings where the true interaction functions are either piecewise-constant functions with unknown dimensionality or continuous.

- **Simulation 4: Two-step adaptive variational algorithm in high-dimensional sigmoid models.** This experiment evaluates the performance and scalability of our fast adaptive variational algorithm (Algorithm 3), for sigmoid Hawkes processes with $K \in \{2, 4, 8, 10, 16, 32, 64\}$, in sparse and less sparse settings of the true parameter $h_0 \in \mathcal{H}_{histo}^{D_0}$ with unknown dimensionality $D_0$.

- **Simulation 5: Convergence of the two-step adaptive variational posterior for varying data set sizes.** In this experiment, we evaluate the asymptotic performance of our two-step variational procedure (Algorithm 3), with respect to the number of observations, i.e., the length of the observation horizon $T$, for sigmoid Hawkes processes with $K = 10$.

- **Simulation 6: Robustness of the variational posterior to some types of mis-specification of the Hawkes model.** This experiment aims at evaluating the performance of our variational algorithm for the sigmoid Hawkes model (Algorithm 3) on data sets generated from Hawkes processes with mis-specified nonlinear link functions and memory parameter of the interaction functions.

In all simulations, we set the memory parameter as $A = 0.1$. In the low-dimensional settings in Sections 6.1, 6.2 and 6.3, we evaluate the performance visually. In the moderately large to large-dimensional settings in Sections 6.4, 6.5 and 6.6, we compute $L_1$-risk measures on the continuous parameter and $\ell_0$-error on the graph parameter (defined below). We note that the only existing Bayesian methods for multivariate Hawkes processes that are not parametric are the variational methods by Zhou et al. (2022) and Malem-Shinitski et al. (2021), which do not perform Bayesian model selection for choosing "smoothness" parameters of the interaction functions (e.g., the number of basis functions in Zhou et al. (2022)) and the graph of interaction. Our variational algorithms, Algorithms 2 and 3, build on these methods and propose a principled way of performing graph and model size inference in the context of high-dimensional Hawkes processes, where this methodology is particularly relevant.

**Remark 14** *One important quantity in these synthetic experiments is the number of* excursions *in the generated data, formally defined in Costa et al. (2020) and Lemma 16 in Appendix D.1. Intuitively, the observation window of the data* $[0, T]$ *can be partitioned into contiguous intervals* $\{[\tau_{i-1}, \tau_i)\}_{i=1,\dots,I}$, $\tau_0 = 0, \tau_I = T$, $I \in \mathbb{N}$, *called* excursions*, where the point process measures are i.i.d. The main properties of these intervals are that* $N[\tau_{i-1}, \tau_i) \geqslant 1$ *and* $N[\tau_i - A, \tau_i) = 0$*. Our theoretical results show that the number of excursions grows linearly with* $T$*. However the constant of proportionality depends on* $K$ *and for large* $K$ *and moderate* $T$ *we sometimes observe no excursion although the posterior still concentrates. To capture better the notion of local (on each dimension)*

*effective sample size we introduce the concept of* local *excursions, defined for each dimension k as a partition of* $[0,T] = \bigcup_{i=1}^{I_k} [\tau_{i-1}^{k,loc}, \tau_i^{k,loc})$ *such that* $N^k[\tau_{i-1}^{k,loc}, \tau_i^{k,loc}) \geqslant 1$ *and* $N^k[\tau_i^{k,loc} - A, \tau_i^{k,loc}) = 0$. *It is not clear that they are a good proxy for effective sample size in general, but in our simulation study below they behave as such.*

### 6.1 Simulation 1: Posterior Distribution in Univariate Nonlinear Hawkes Models

In this simulation, we consider univariate Hawkes processes ($K = 1$) with link function $\phi = \phi_1$ of the form

$$\phi(x) = \theta + \Lambda\psi(\alpha(x - \eta)), \tag{37}$$

where $\xi = (\theta, \Lambda, \alpha, \eta)$ and $\psi : \mathbb{R} \to \mathbb{R}^+$ are known and chosen as:

- Sigmoid: $\psi(x) = (1 + e^{-x})^{-1}$ and $\xi = (0.0, 20.0, 0.2, 10.0)$;
- ReLU: $\psi(x) = \max(x, 0)$ and $\xi = (0.001, 1.0, 1.0, 0.0)$;
- Softplus: $\psi(x) = \log(1 + e^x)$ and $\xi = (0.0, 40.0, 0.1, 20.0)$.

Note that the corresponding link functions $\phi$ have similar shapes on a range of values between -20 and 20 (see Figure 2). In all models, we consider a Hawkes process with $h_0 = h_{11}^0 \in \mathcal{H}_{histo}^{D_0}$ with $D_0 = 2$, and three scenarios, called *Excitation only*, *Mixed effect*, and *Inhibition only*, where $h_0$ is respectively non-negative, signed, and non-positive (see Figure 3 for instance). In each of the nine settings, we set $T = 500$ and in Table 1, we report the corresponding number of events and excursions observed in each scenario and model. Note that, as we may expect, more events and less excursions are observed in the data generated in *Excitation only* scenario than in the *Mixed effect* and *Inhibition only* scenarios.

Here, we assume that $D_0$ is known and we consider a normal prior on $\mathcal{H}_{histo}^{D_0}$ such that $w_{11} \sim \mathcal{N}(0, \sigma^2 I)$, and for $\nu_1$, $\nu_1 \sim \mathcal{N}(0, \sigma^2)$, with $\sigma = 5.0$. To compute the (true) posterior distribution, we run a Metropolis-Hasting (MH) sampler implemented via the Python package PyMC4[1] with 4 chains, 40 000 iterations, and a burn-in time of 4000 iterations. We also use the Gaussian quadrature method (Golub and Welsch, 1969) for evaluating the log-likelihood function, except in the ReLU model and *Excitation only* scenario, where the integral term is computed exactly. We note that we also tested a Hamiltonian Monte-Carlo sampler in this simulation, and obtained similar posterior distributions, but within a much larger computational time, therefore these results are excluded from this experiment.

The posterior distribution on $f = (\nu_1, h_{11})$ in the ReLU model and our three scenarios are plotted in Figure 3. For conciseness purpose in this section, our results for the sigmoid and softplus models are reported in Appendix G.1. We note that in almost all settings, the ground-truth parameter $f_0$ is included in the 95% credible sets of the posterior distribution. Nonetheless, the posterior mean is sometimes biased, possibly due to the numerical integration errors in the log-likelihood computation. Moreover, we conjecture that the estimation quality depends on the number of events and the number of excursions, which could explain the differences between the *Excitation only*, *Mixed effect*, and *Inhibition only* scenarios. In particular, the credible sets seem consistently smaller for the second scenario, which realisations have more excursions than the other ones.

1. `https://www.pymc.io/welcome.html`

This simulation therefore shows that the posterior distribution in commonly used nonlinear univariate Hawkes models behaves well and can be sampled from using a simple MH sampler. Nonetheless, we note that the MH iterations are computationally expensive, which prevents from scaling this algorithm to large dimensions. Therefore, we will only use the MH sampler to compute the posterior distribution in the low-dimensional settings, i.e., Simulations 2 and 3, with respectively $K = 1$ and $K = 2$.

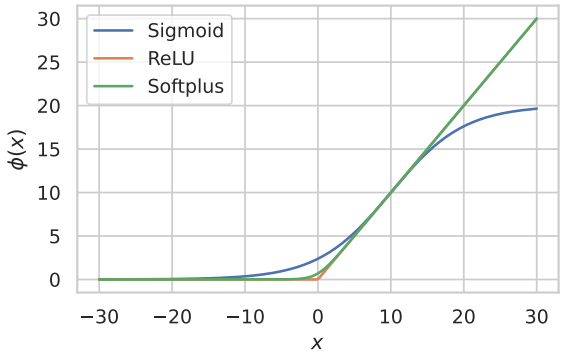



Figure 2: Link functions $\phi$ of the Hawkes model considered in Simulation 1, namely the sigmoid (blue), ReLU (red), and softplus (green) functions.

| Scenario | | Sigmoid | ReLU | Softplus |
|---|---|---|---|---|
| *Excitation* only | # events | 5250 | 5352 | 4953 |
| | # excursions | 1558 | 1436 | 1373 |
| Mixed effect | # events | 3876 | 3684 | 3418 |
| | # excursions | 1775 | 1795 | 1650 |
| Inhibition only | # events | 3047 | 2724 | 2596 |
| | # excursions | 1817 | 1693 | 1588 |

Table 1: Number of events and excursions in the simulated data of Simulation 1 with $T = 500$. We refer to Remark 14 and Lemma 16 in Appendix D.1 for the definition of an excursion in Hawkes processes.

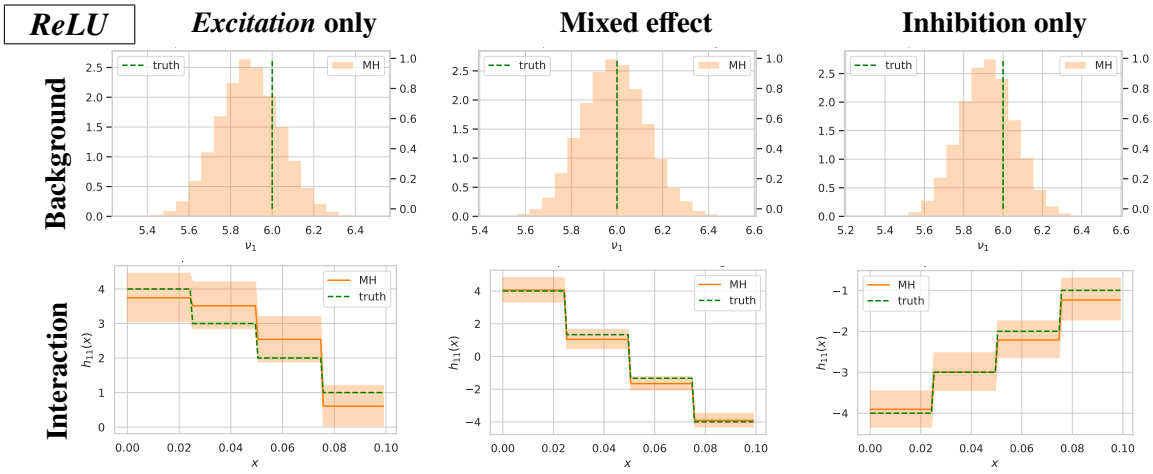



Figure 3: Posterior distribution on $f = (\nu_1, h_{11})$ obtained with the Metropolis-Hastings sampler (MH), in the univariate ReLU models of Simulation 1. The three columns correspond to the *Excitation only* (left), *Mixed effect* (center), and *Inhibition only* (right) scenarios. On the first row, we plot the marginal posterior distribution on the background rate $\nu_1$, and on the second row, the posterior mean (solid orange line) and 95% credible sets (orange areas) on the interaction function $h_{11}$, here piecewise-constant with dimensionality $2^{D_0} = 4$. The true parameter $f_0 = (\nu_1^0, h_{11}^0)$ is plotted in dotted green line.

### 6.2 Simulation 2: Parametric Variational Posterior and Posterior Distribution in the Univariate Sigmoid Model.

In this simulation, we consider the same univariate scenarios as Simulation 1, but only for the sigmoid Hawkes model and compare the variational and true posterior distributions. Here, the dimensionality $D_0$ of the true function $h_0$ is assumed to be known, therefore, the samplers are non-adaptive. Specifically, we compare the performance of the previous MH sampler, the Gibbs sampler (introduced in Remark 5 and described in Algorithm 4 in Appendix F), and our mean-field variational algorithm in a fixed model (Algorithm 1) - here, we fix the dimensionality of $h_{11}$ to $J = 2^{D_0} = 4$. We run 4 chains for 40 000 iterations for the MH sampler, 3000 iterations of the Gibbs sampler, and use our early-stopping procedure for the mean-field variational algorithm.

In Figure 4, we can compare the variational posterior on $f = (\nu_1, h_{11})$ to the posterior distributions, computed either with the Gibbs or MH samplers, in the three estimation scenarios. We note that variational posterior mean is always close to the posterior mean, in particular when computed with the Gibbs sampler. Nonetheless, its credible sets are generally smaller, which is a common empirical observation of mean-field variational approximations.

Besides, the variational posterior seems to be similarly biased as the posterior distribution, as can be seen for the background rate $\nu_1$ in the *Inhibition* scenario. One could therefore test if this bias decreases with more data observations, i.e., larger $T$; however, the Gibbs sampler has a large computational time (between 3 and 5 hours), which is about 6 (resp. 40) times longer than the MH sampler (resp. our mean-field algorithm), due to the expensive latent variable sampling scheme (see Table 2). Finally, we also compare the estimated intensity function using the (variational) posterior means, on a sub-window of the observations in Figure 5. The latter plot shows that all three methods provide fairly equivalent estimates on the nonlinear intensity function.

From this simulation, we conclude that, in the univariate and parametric sigmoid Hawkes model, the mean-field variational algorithm in a fixed model provides a good approximation of the posterior distribution. Moreover, we note that although the Gibbs sampler is slightly better than MH, it is much slower than the latter and therefore cannot be applied to multivariate Hawkes processes in practice. Therefore, in the bivariate simulation in the next section, we only compare to the posterior distribution computed with the MH sampler, which can still be computed within reasonable time for $K = 2$ .

| Scenario | MH | Gibbs | MF-VI |
|---|---|---|---|
| *Excitation* only | 2169 | 16 092 | 416 |
| Mixed effect | 2181 | 13 097 | 338 |
| *Inhibition* only | 2222 | 9 318 | 400 |

Table 2: Computational times (in seconds) of the Gibbs sampler (Algorithm 4), our mean-field variational (MF-VI) algorithm (Algorithm 1), and the Metropolis-Hastings (MH) sampler in each parametric univariate scenario of Simulation 2 with $T = 500$. We note that the Gibbs sampler is much slower than the MH sampler, which is also slower than the mean-field variational algorithm.

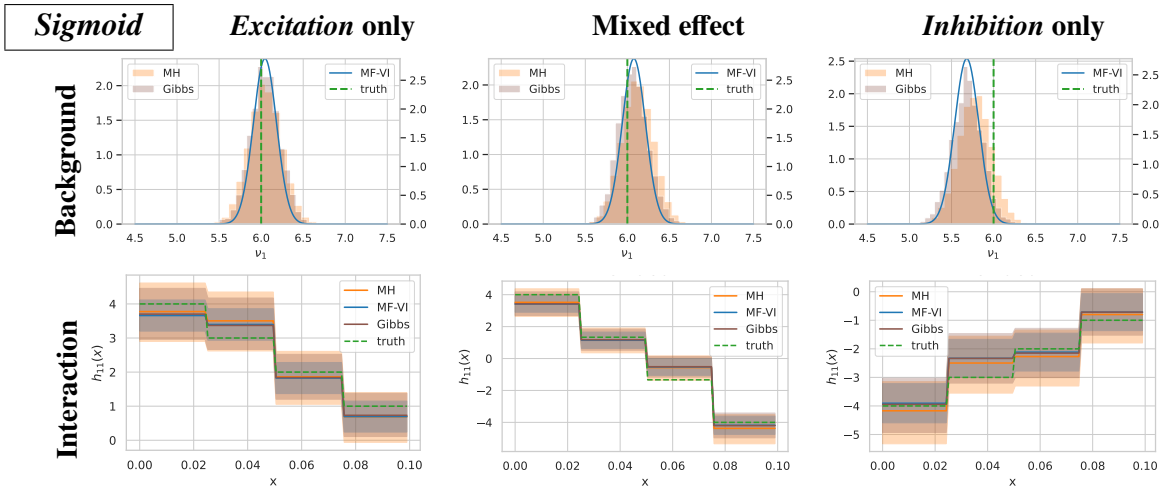
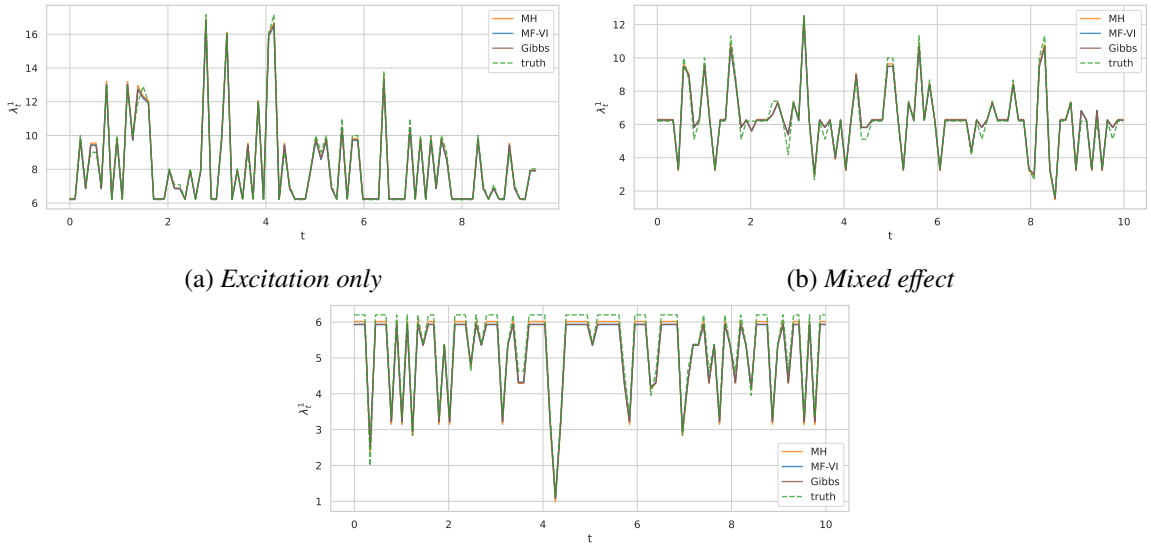



Figure 4: Posterior and variational posterior distributions on $f = (\nu_1, h_{11})$ in the univariate sigmoid model of Simulation 2, evaluated by the MH sampler, the mean-field variational (MF-VI) algorithm in a fixed model (Algorithm 1) and the Gibbs sampler (Algorithm 4). The three columns correspond to the *Excitation only* (left), *Mixed effect* (center), and *Inhibition only* (right) scenarios. The true parameter $f_0$ is plotted in dotted green line. The first row contains the marginal distributions (VB, MH and Gibbs) on the background rate $\nu_1$, and the second row represents the posterior means (solid lines) and 95% credible sets (colored areas) on the (self) interaction function $h_{11}$. We note that the variational posterior is close to the Gibbs posterior distribution, nonetheless, has smaller credible bands.




(a) *Excitation only*

(b) *Mixed effect*

(c) *Inhibition only*

Figure 5: Intensity function on a sub-window of the observation window estimated via the variational posterior mean (blue) or via the posterior mean, computed with the MH sampler (orange) or the Gibbs sampler (purple), in each scenario of Simulation 2. The true intensity $\lambda_t^1(f_0)$ is plotted in dotted green line. We note that all estimates are close in this simulation.

### 6.3 Simulation 3: Fully-Adaptive Variational Method in the Univariate and Bivariate Sigmoid Models.

In this simulation, we test our fully-adaptive variational inference algorithm (Algorithm 2), in the one-dimensional ($K = 1$) and two-dimensional ($K = 2$) sigmoid models, and in two estimation settings:

1. *Well-specified:* $h_0 \in \mathcal{H}_{hist}^{D_0}$ (with $D_0 = 2$);
2. *Mis-specified:* $h_0 \notin \mathcal{H}_{hist}^{D_0}$, and $h_{lk}^0$ is a continuous function, for all $(l, k) \in [K^2]$.

Note that in the well-specified case, $m_0 := (\delta_0, 2^{D_0})$ is unknown for the variational method, nonetheless, we also compute the posterior distribution with the non-adaptive MH sampler using the true $m_0$. In the bivariate model, we choose a true graph parameter $\delta_0$ with one zero entry (see Figure 8a). We also consider an *Excitation* scenario where all the true interaction functions $(h_{lk}^0)_{l,k}$ are non-negative and with $T = 2000$, and an *Inhibition* scenario where the self-interaction functions $(h_{kk}^0)_{k=1,2}$ are non-positive with $T = 3000$. The latter setting aims at imitating the so-called self-inhibition phenomenon in neuronal spiking data, due to the refractory period of neurons (Bonnet et al., 2021). In our adaptive variational algorithm, we set a maximum histogram depth $D_1 = 5$ for $K = 1$, and $D_1 = 4$ for $K = 2$, so that the number of models per dimension is respectively 7 and 76.

In the well-specified setting, we first analyse the ability of Algorithm 2 to recover the true connectivity graph and dimensionality of $h_0$. In Figure 6, we plot the model marginal probabilities $(\hat{\gamma}_m)_m$ in our adaptive variational posterior and in the univariate setting. In the *Excitation* scenario, the marginal probability mass $\hat{\gamma}_{\hat{m}}$ is maximized at the true model, i.e., $\hat{m} = m_0 = (\delta_0 = 1, 2^{D_0} = 2)$, and all the other marginal probabilities $\hat{\gamma}_m$ are negligible. Therefore, in this case, the model-averaging VB posterior (16) is essentially equivalent to the model-selection VB posterior (15). In the *Inhibition* scenario, as seen in Figure 6, $\hat{\gamma}_m$ takes almost the same value at $\hat{m} = (\hat{\delta} = \delta_0 = 1, \hat{D} = 1)$ (mode of $\hat{\gamma}_m$) and at the true model $m_0$, roughly 0.5 in each case. The model-selection VB posterior thus is based on the wrong model for $D$ but not for $\delta$. Interestingly the estimation of $\nu$ remains good, while that of $h$ is obviously biased. The model averaging VB (16) is thus to be preferred in this situation, since it is essentially a mixture of two components, one corresponding to $\hat{D} = 1$, and the second one corresponding to the true model $D_0 = 2$ .

Nonetheless, comparing the estimated nonlinear intensity based on the model-selection variational posterior mean and the posterior mean in Figure 27 in Appendix G, we note that the model selection variational estimate is very close to the true intensity and the non-adaptive MH estimate, despite the error of dimensionality in the *Inhibition* scenario.

We then compare the model selection adaptive variational posterior distribution on the parameter with the true posterior distribution computed with the non-adaptive MH sampler in Figure 7. We note that in the *Excitation* scenario, the variational posterior mean is very close to the posterior mean, however, its 95% credible bands are significantly smaller. Note also that, in the *Inhibition* scenario, in spite of the wrongly selected histogram depth, the estimated interaction function is still not too far from the truth.

In the mis-specified setting, all the marginal probabilities are negligible but one, in both the *Excitation* and *Inhibition* scenarios (see Figure 6), although there is no true $m_0$ in this case. In Figure 29 in Appendix G, we note that the model selection adaptive variational posterior mean approximates quite well the true parameter. Moreover, its 95% credible bands often cover the truth but are once again slightly too narrow.

The previous observations in the well-specified and mis-specified settings can also be made in the two-dimensional setting. In Figure 8, we plot the marginal probabilities in the adaptive variational posterior in the 2 scenarios of our well-specified setting. We note that in this bi-dimensional example, we observe again that the highest probability model is correct, i.e., $\hat{m} = m_0$, in the Excitation scenario but not in the Inhibition scenario. For the latter, the posterior distributions on the intensity and parameter are plotted in Figures 28 and 30 in Appendix G, and we note again that the number of basis function in $\hat{Q}_{MS}$ is 2 instead of 4 - however the intensity is still well estimated. For the Excitation scenario and mis-specified setting, Figure 9 reveals that the parameter is also well estimated, however, the under-coverage phenomenon of the credible regions also occurs in this mis-specified example.

Finally, we note that our fully-adaptive variational algorithm is more than 10 times faster to compute than the non-adaptive MH sampler, as can be seen from the computing times reported in Table 3. This simulation study therefore shows that our fully-adaptive variational algorithm enjoys several advantages in Bayesian estimation for Hawkes processes: it can infer the dimensionality of the interaction functions $D$, the dependence structure through the graph parameter $\delta$, provides a good approximation of the posterior mean, and is computationally efficient.

| # dimensions | Scenario | T | FA-MF-VI | MH |
|---|---|---|---|---|
| $K = 1$ | Excitation | 2000 | 32 | 417 |
| | Inhibition | 3000 | 33 | 445 |
| $K = 2$ | Excitation | 2000 | 189 | 2605 |
| | Inhibition | 3000 | 197 | 2791 |

Table 3: Computing times (in seconds) of our fully-adaptive mean-field variational method (FA-MF-VI) (Algorithm 2) and the Metropolis-Hastings (MH) sampler in the univariate and bivariate sigmoid models and the scenarios of Simulation 3.

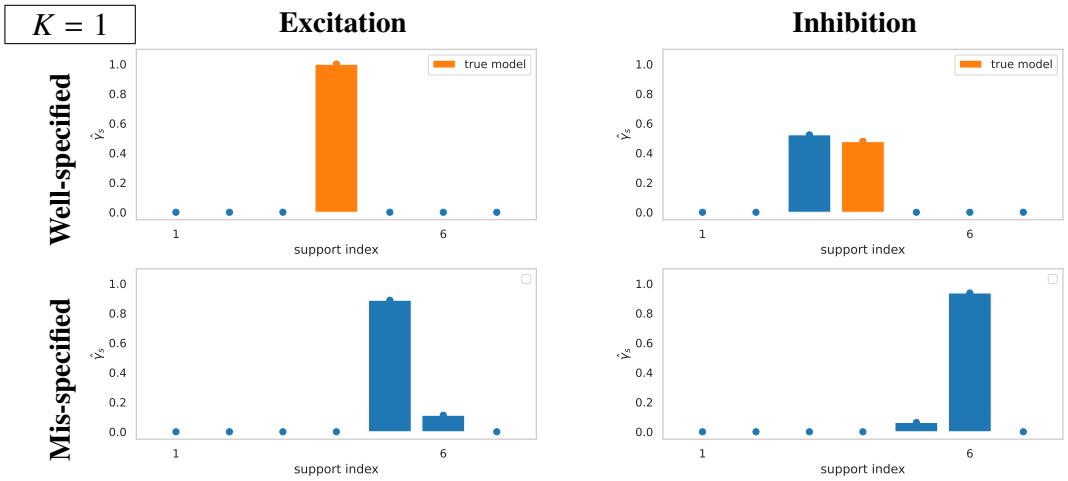



Figure 6: Model marginal probabilities $(\hat{\gamma}_m)_m$ in the adaptive mean-field variational posterior, in the well-specified and mis-specified settings of Simulation 3 with $K = 1$. The left and right panels correspond to the *Excitation* (resp. *Inhibition*) setting. The elements in $\mathcal{S}_1$ are indexed from 1 to 7, and correspond respectively to $m = (\delta = 0, 2^D = 1)$, and $m = (\delta = 1, 2^D)$ with $D = 0, \ldots, 5$. All probabilities are plotted in blue, except for the one corresponding to the true model which is colored in orange.

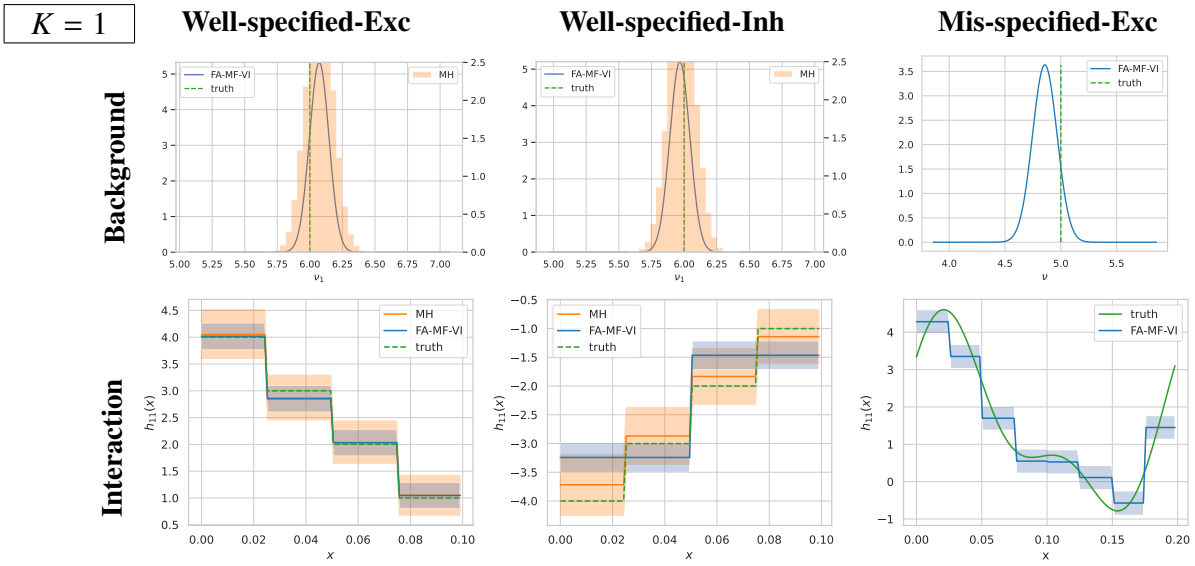



Figure 7: Posterior and model-selection variational posterior distributions on $f = (\nu_1, h_{11})$ in the univariate sigmoid model and settings of Simulation 3, evaluated by the MH sampler and the fully-adaptive mean-field variational (FA-MF-VI) algorithm (Algorithm 2). The three columns correspond respectively to the two well-specified settings, i.e., the *Excitation* (Well-specified-Exc) and *Inhibition* (Well-specified-Exc) scenarios, and one mis-specified setting (Mis-specified-Exc). The first row contains the marginal distribution on the background rate $\nu_1$, and the second row represents the (variational) posterior mean (solid line) and 95% credible sets (colored areas) on the (self) interaction function $h_{11}$. The true parameter $f_0$ is plotted in dotted green line.

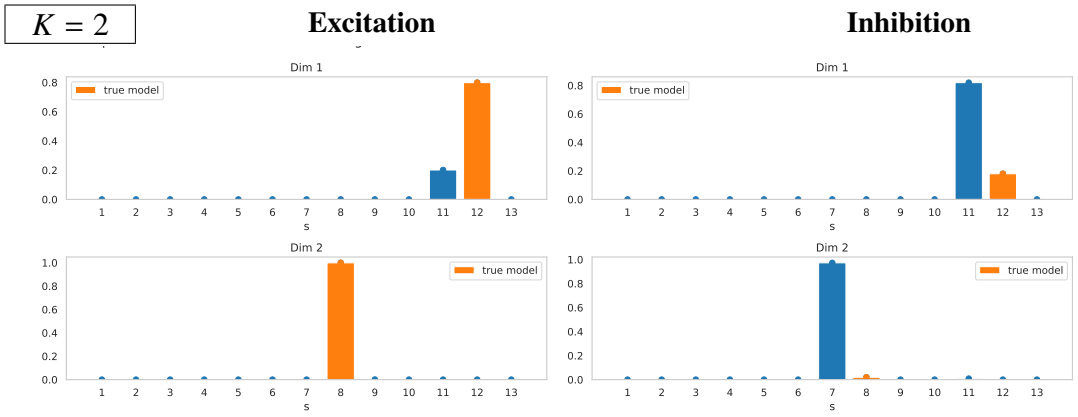



Figure 8: Marginal probabilities on the graph and dimensionality parameter $s_k = (\delta_{\cdot,k}, D_k)$ at the first dimension (first row) and second dimension (second row), i.e., $(\hat{\gamma}^k_{s_k})_{s_k \in \mathcal{S}_2}$, $k = 1, 2$ in the fully-adaptive averaged mean-field variational posterior, in the well-specified setting of Simulation 3 with $K = 2$. The *Excitation* scenario (left column) corresponds to $h_0 \geqslant 0$, while in the *Inhibition* scenario (right column) , $h^0_{11}, h^0_{22} \leqslant 0$. The elements in $\mathcal{S}_2$ are indexed from 1 to 13. All probabilities are plotted in blue, except for the one corresponding to the true model which is colored in orange.

$K = 2$

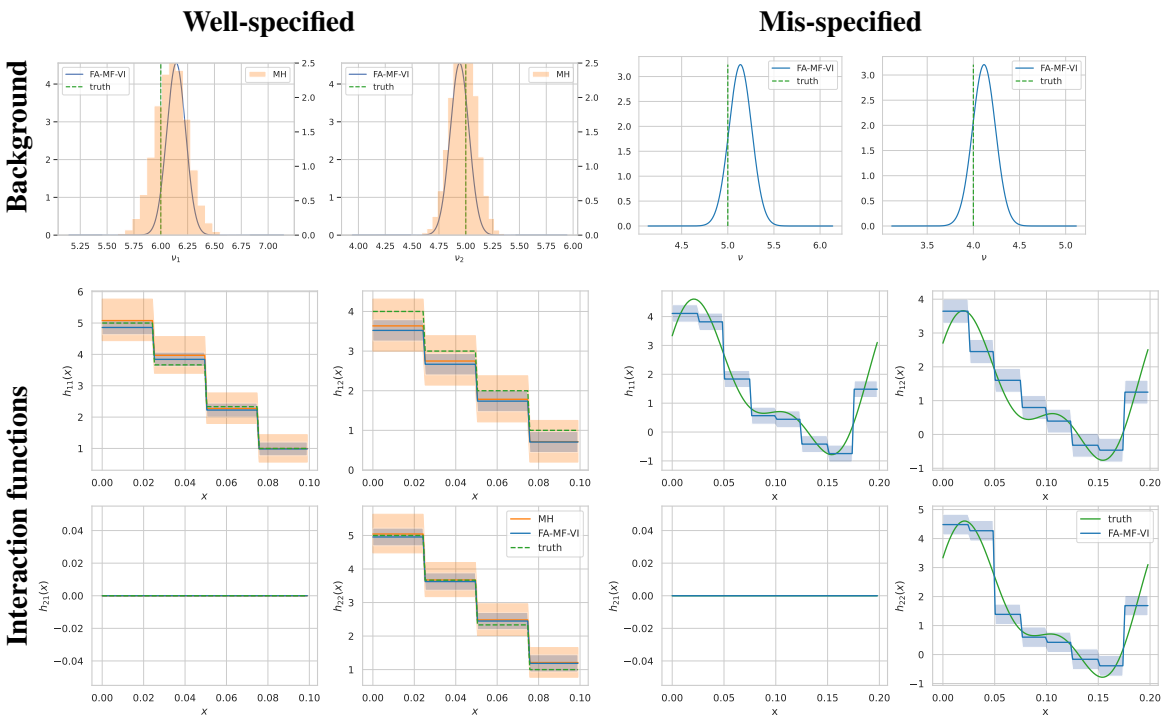



Figure 9: Model-selection variational posterior distributions on $f = (\nu, h)$ in the bivariate sigmoid model, and well-specified and mis-specified settings, and *Excitation* scenario of Simulation 3, computed with the fully-adaptive mean-field variational (FA-MF-VI) algorithm (Algorithm 2). The first row correspond two columns correspond to the *Excitation* (left) and *Inhibition* (right) settings. The first row contains the marginal distribution on the background rates $(\nu_1, \nu_2)$, and the second and third rows represent the (variational) posterior mean (solid line) and 95% credible sets (colored areas) on the four interaction function $h_{11}, h_{12}, h_{21}, h_{22}$. The true parameter $f_0$ is plotted in dotted green line.

### 6.4 Simulation 4: Two-step Variational Posterior in High-dimensional Sigmoid Models.

In this section, we test the performance of our two-step variational procedure (Algorithm 3), first, in sparse settings of the true parameter $h_0$, then, in relatively denser regimes.

#### 6.4.1 Sparse Settings

In this experiment, we consider sparse multivariate sigmoid models with $K \in \{2, 4, 8, 16, 32, 64\}$ dimensions. We recall that to the best of our knowledge, the only Bayesian method that has currently been tested in high-dimensional Hawkes processes is the semi-parametric version of Zhou et al. (2022) where the interaction functions are also decomposed over a dictionary of functions, but the choice of the number of functions is not driven by a model selection procedure and the graph of interaction is not inferred. Here, we construct a well-specified setting with $h_0 \in \mathcal{H}^{D_0}_{hist}$ and $D_0 = 1$, and an *Excitation* scenario and an *Inhibition* scenario, similar to Simulation 3, and a *sparse* connectivity graph parameter $\delta_0$ with $\sum_{l,k} \delta^0_{lk} = 2K - 1$, as shown in Figure 12. In Table 4, we report our chosen value of $T$ in each setting and the corresponding number of events, global and local excursions, and the computing times of our algorithm.

In Table 5 we report multiple performance measures of our method. First, we report the $L_1$-risk of the model-selection variational posterior defined as

$$r_{L_1}(\hat{Q}) := \mathbb{E}_{\hat{Q}}[\|\nu - \nu_0\|_1] + \sum_{l,k} \mathbb{E}_{\hat{Q}}\left[\left\|h_{lk} - h^0_{lk}\right\|_1\right]. \tag{38}$$

We note that in general, the number of terms in the risk grows with $K$ and the number of non-null interaction functions in $h$ and $h_0$ - which thus can be of order $O(K^2)$ in a *dense* setting. We also report the $L_1$-error of the variational posterior mean defined as

$$Err_{L_1}(\hat{Q}) := \left\|\mathbb{E}_{\hat{Q}}[\nu] - \nu_0\right\|_1 + \sum_{l,k} \left\|\mathbb{E}_{\hat{Q}}[h_{lk}] - h^0_{lk}\right\|_1,$$

as well as the posterior mean error on the norms of the interaction functions:

$$Err_h(\hat{Q}) := \sum_{l,k} \left|\mathbb{E}_{\hat{Q}}[\|h_{lk}\|_1] - \left\|h^0_{lk}\right\|_1\right|.$$

We note that for the model selection variational posterior $\hat{Q}_{MS}$ with model $\hat{m} = (\hat{\delta}, \hat{J} = (J_k))$ the posterior mean on the norms of the interaction functions can be exactly computed, since

$$\mathbb{E}_{\hat{Q}}[\|h_{lk}\|_1] = \sum_{j=1}^{J_k} \sqrt{\frac{2}{\pi}[\Sigma^{\hat{J}_k}_{lk}]_{jj}} \exp\left\{-\frac{[\tilde{\mu}^{\hat{J}_k}_{lk}]^2_j}{[\Sigma^{\hat{J}_k}_{lk}]_{jj}}\right\} - [\tilde{\mu}^{\hat{J}_k}_{lk}]_j \left[1 - 2\Phi\left(-\frac{[\tilde{\mu}^{\hat{J}_k}_{lk}]_j}{\sqrt{[\Sigma^{\hat{J}_k}_{lk}]_{jj}}}\right)\right].$$

We estimate the posterior risk and the posterior mean error via a Monte-Carlo estimate with 500 samples. Additionally, we evaluate the accuracy of our algorithm when estimating the graph of interaction and the size $D_k$ at each dimension $k$, defined as

$$Acc_{graph}(\hat{\delta}) = \frac{1}{K^2} \sum_{l,k} \mathbb{1}_{\delta^0_{lk} = \hat{\delta}_{lk}}, \qquad Acc_{dim}(\hat{D}) = \frac{1}{K} \sum_{k} \mathbb{1}_{D^0_k = \hat{D}_k},$$

where $\hat{\delta} = (\hat{\delta}_{lk})_{l,k}$ and $\hat{D} = (\hat{D}_k)_k$ are respectively the estimated graph and the inferred dimensionality of $(h_{.k})_k$ in Algorithm 3.

From Table 5, we note that, in almost all settings, the accuracy of our algorithm for estimating the graph of interaction is equal or very close to 1. Therefore, our method is able to recover almost perfectly the true graph $\delta_0$ (the estimated graphs in the *Excitation* and *Inhibition* scenarios are plotted in Figures 31 and 32 in Appendix G.3). Hence, our gap heuristics for choosing the threshold $\eta_0$ (see Section 4.2.2) after the first step of Algorithm 3 works well in this setting. In Figure 15 (and Figure 34 in Appendix G.3 in the *Inhibition* scenario), we note that the $L_1$-norms of the interaction functions are well estimated in the first step, leading to a gap between the 95 % credible sets of the estimated norms close and far from 0. This gap includes 0.12 for all $K$'s, which is the value we choose for $\eta_0$. We therefore observe that here the variational estimates allow to discriminate between the true signals and the noise and to recover the true graph parameter.

Additionally, from Table 5, we note that our measures of risk and error seem to approximately grow linearly with $K$, which indicates that the estimation does not deteriorate with larger $K$. In Figure 14 (and Figure 33 in Appendix G.3), we plot the error on the $L_1$-norms in the form of a heatmap compared to the true norms. We note that for all $K$'s, these errors are relatively small. Moreover, our variational algorithm estimates well the parameter, as can be visually checked in Figure 18, where we plot the model-selection variational posterior distribution on a subset of the parameter for each value of $K$, in the *Excitation* scenario (see Figure 35 in Appendix G.3 for our results in the *Inhibition* scenario). Besides, the computing times of our algorithm seem to scale relatively well with $K$ and the number of events in these sparse settings, as can be seen from Table 5 and Figure 10. For $K = 64$, our algorithm runs in less than 13 hours, in spite of the large number of events (about 77000) in the largest data set.

Finally, to assess the utility of the graph selection step of Algorithm 3, where the adaptive variational posterior is computed by optimising over the set of models $\mathcal{M}_E$, we have also compared the adaptive variational posteriors obtained after the first optimisation (over the set of models $\mathcal{M}_C$) and after the second one. More precisely, we evaluate the risk and error of the final VB posterior ("step 2") and of the VB posterior obtained after thresholding the small norms and setting the corresponding interaction functions to 0 ("step 1"). For this comparison, we consider the same settings with dimensions $K = 2, 4, 8, 16, 32$ and repeat our experiment 10 times. In Figure 11, we report the averaged risks (with standard deviations). We observe that the risks of the two VB posteriors are very close for low dimensional settings, but for the larger dimensions ($K \geqslant 16$), the risk after the second step is smaller, in particular in the Inhibition scenario. Hence, in addition to learning an interpretable dependency structure for the multivariate process, our graph selection step can significantly improve the estimation of the non-zero parameters. We recall that this is in contrast to existing variational methods such as Zhou et al. (2022) which do not perform any model selection step.

**Remark 15** *We note that while the computing times of our algorithm in these experiments have been evaluated using only two processing units, it could be greatly decreased if more cores are available since several parts of our algorithm can be parallelised. In fact, our algorithm computing the model selection variational posterior for our hierarchical prior can leverage parallel computing in 2 ways:*

*(a) When the prior distribution on the full parameter $f$ factorises into $K$ distributions for each sub-part of the parameter $f_k, k = 1, \dots, K$, the variational posterior is also factorisable, i.e., $\hat{Q}(f) = \prod_{k=1}^{K} \hat{Q}^k(f_k)$, and one can compute independently the $K$ factors $\hat{Q}_k(f_k), k \in [K]$.*

*(b) For each model $m \in \mathcal{M}$, one can compute independently each model-restricted variational posterior $\hat{Q}^{k,m}(f_k^m)$ that enters the expression of $\hat{Q}^k(f_k)$.*

*The expected decrease in computing time obtained by parallelising these computations depends on the number of available machines. If there are $K \times M$ machines, with $M = |\mathcal{N}|$ the number of models, then (a) allows to decrease the computing time by a factor $K$ and (b) allows to decrease it almost by a factor $M$ (since after computing each $\hat{Q}^{k,m}(f_k^m)$ one only needs to compute the variational posterior probabilities $(\hat{\gamma}_k^m)_m$ to compute $\hat{Q}(f)$).*

### 6.4.2 Testing Different Graphs and Sparsity Levels.

In this experiment, we evaluate Algorithm 3 on different settings of the graph parameter $\delta_0$, namely a sparse, a random, and a dense settings, illustrate in Figure 13. The sparse setting is similar to the previous section, while the random setting corresponds to a slightly less sparse regime where additional edges are present in $\delta_0$. Note that these three settings have different numbers of edges in $\delta_0$, therefore, different numbers of non-null interaction functions to estimate. From Table 6, we also note that there are more events and less global excursions in the dense setting that in the two other ones, in particular, in the *Excitation* scenario where this number drops to 2.

Our numerical results in Table 7 show that in the dense setting, the graph accuracy of our estimator is slightly worse, and the risk of the variational posterior is much higher than in the other settings. We conjecture that this loss of performance is related to the smaller number of global excursions, which leads to a more difficult estimation problem. We can also see from Figure 16 that in this particular setting, the estimation of the norms of the interaction functions is deteriorated, and the gap that allows to discriminate between the null and non-null functions is not present anymore. Nonetheless, in the *Inhibition* scenario, for which the number of global excursions is not too small, this phenomenon does not happen and the estimation is almost equivalent in all graph settings.

To further explore the applicability of our thresholding approach in the dense setting, we test the following three-step approach in the *Excitation* scenario, with $K = 10$ and a dense graph $\delta_0$:

- The first step is similar to the one of our two-step procedure, i.e., we estimate an adaptive variational posterior distribution within models that contain the complete graph $\delta_C$.

  Then, if there is no significant gap in the variational posterior mean estimates of the $L_1$-norms, we look for a (conservative) threshold $\eta_1$ corresponding to the first "slope change", and estimate a (dense) graph $\hat{\delta}$.

- In a second step, we compute the adaptive variational posterior distribution within models that contain $\hat{\delta}$ and re-estimate the $L_1$-norms of the functions.

  If we now see a significant gap in the norms estimates, we choose a second threshold within that gap; otherwise, we look again for a slope change and pick a conservative threshold $\eta_2$ to compute a second graph estimate $\hat{\delta}_2$.

- In the third and last step, we repeat the second step with now our second graph estimate, $\hat{\delta}_2$.

In Figure 17, we plot our estimates of the norms after each step of the previous procedure. In this case, we have chosen visually the threshold $\eta_1 = 0.09$ and $\eta_2 = 0.18$ after respectively the first and second step, using the slope change heuristics. We note that the previous method indeed provides a conservative graph estimate in the first step, but in the second step, allows to refine our estimate of the graph and approach the true graph. Besides, we note that the large norms are inflated along the three steps of our procedure. Therefore, our method performs better in sparse settings where a significant gap allows to correctly infer the true graph $\delta_0$.

In conclusion, our simulations in low and high-dimensional settings, with different levels of sparsity in the graph, show that our two-step procedure is able to correctly select the graph parameter and dimensionality of the process in sparse settings, and hence allows to scale up variational Bayes approaches to larger number of dimensions. Nonetheless, from the moderately high-dimensional settings, the estimation of the parameter $f$ becomes sensitive to the difficulty of the problem. In particular, the performance is sensitive to the graph sparsity, tuning the number of non-null functions to estimate, and, as we conjecture, the number of global excursions in the data. Finally, we note that heuristic approaches for the choice of the threshold - needed to estimate the graph parameter - need to further explored in noisier and denser settings.

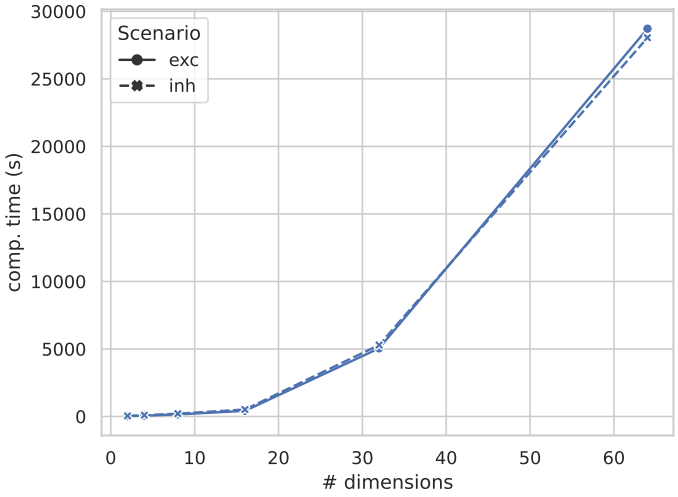
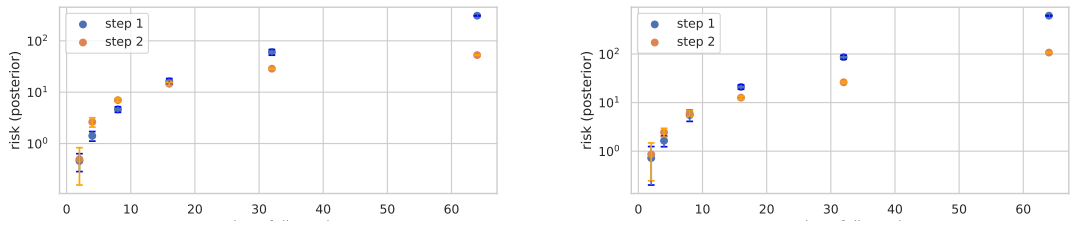


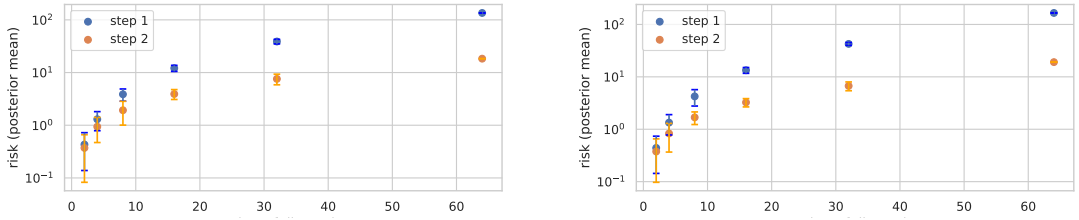

Figure 10: Computational times of our two-step mean-field variational algorithm (Algorithm 3) in the *Excitation* (exc) and *Inhibition* (inh) scenarios and well-specified setting of Simulation 4, for $K = 2, 4, 8, 16, 32, 64$.

| K | Scenario | T | # events | # global excursions | # local excursions | CT (s) |
|---|---|---|---|---|---|---|
| 2 | Exc | 500 | 4,176 | 1,477 | 1,054 | 27 |
| | Inh | 700 | 3,703 | 2,177 | 1,516 | 45 |
| 4 | Exc | 500 | 8,043 | 1,533 | 1,046 | 59 |
| | Inh | 700 | 6,271 | 2,536 | 1,315 | 96 |
| 8 | Exc | 500 | 19,833 | 704 | 1,146 | 153 |
| | Inh | 700 | 15,077 | 1,734 | 1,526 | 208 |
| 16 | Exc | 500 | 38,342 | 122 | 1,104 | 416 |
| | Inh | 700 | 26,541 | 632 | 1,373 | 517 |
| 32 | Exc | 500 | 61,978 | 6 | 1,007 | 5049 |
| | Inh | 700 | 60,638 | 14 | 1,531 | 5293 |
| 64 | Exc | 300 | 88,227 | 0 | 656 | 28714 |
| | Inh | 450 | 79,865 | 0 | 994 | 28046 |

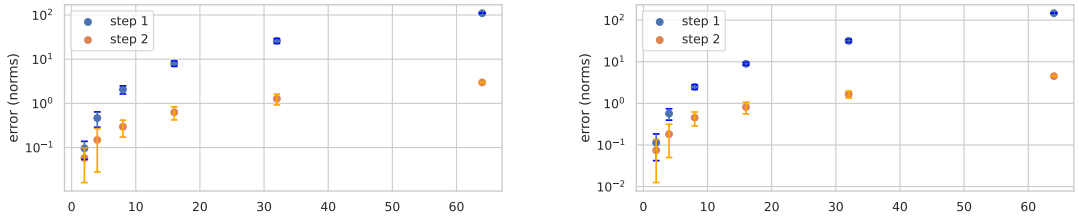
Table 4: Description of the simulated data in Simulation 4 (averaged over 10 repetitions and rounded at the closest integer): number of observed events, excursions (global and local) and computing times (CT) of Algorithm 3.




(a) Variational posterior risk $r_{L_1}(\hat{Q})$.




(b) Error of variational posterior mean $Err_{L_1}(\hat{Q})$.




(c) Error of variational posterior mean on norms $Err_h(\hat{Q})$.

Figure 11: Performance of the variational posteriors obtained after the first step (in blue) and the second step (in orange) of Algorithm 3 in the *Excitation* scenario (left column) and *Inhibition* scenario (right column) of Simulation 4, versus the number of dimensions $K = 2, 4, 8, 16, 32, 64$. From top to bottom row, we report the posterior risk, the posterior mean risk, and the posterior mean error on functions' norms. For each setting (except for $K = 64$), we repeat the experiment 10 times and we plot the mean risk and the intervals at $\pm 2$ standard deviations.

| K | Scenario | Risk | $L_1$-error | Error on norms | Graph accuracy | Dimension Accuracy |
|---|---|---|---|---|---|---|
| 2 | Exc | 0.47 | 0.37 | 0.06 | 1.00 | 1.00 |
| | Inh | 0.46 | 0.38 | 0.07 | 1.00 | 1.00 |
| 4 | Exc | 1.10 | 0.94 | 0.15 | 1.00 | 1.00 |
| | Inh | 1.02 | 0.83 | 0.18 | 1.00 | 0.95 |
| 8 | Exc | 2.28 | 1.93 | 0.30 | 1.00 | 0.99 |
| | Inh | 2.08 | 1.69 | 0.45 | 1.00 | 0.99 |
| 16 | Exc | 4.63 | 3.92 | 0.63 | 1.00 | 0.99 |
| | Inh | 4.03 | 3.25 | 0.81 | 1.00 | 0.98 |
| 32 | Exc | 8.99 | 7.58 | 1.27 | 1.00 | 0.98 |
| | Inh | 8.61 | 6.71 | 1.66 | 1.00 | 0.95 |
| 64 | Exc | 22.16 | 18.34 | 2.98 | 1.00 | 0.98 |
| | Inh | 24.99 | 19.10 | 4.50 | 1.00 | 0.84 |

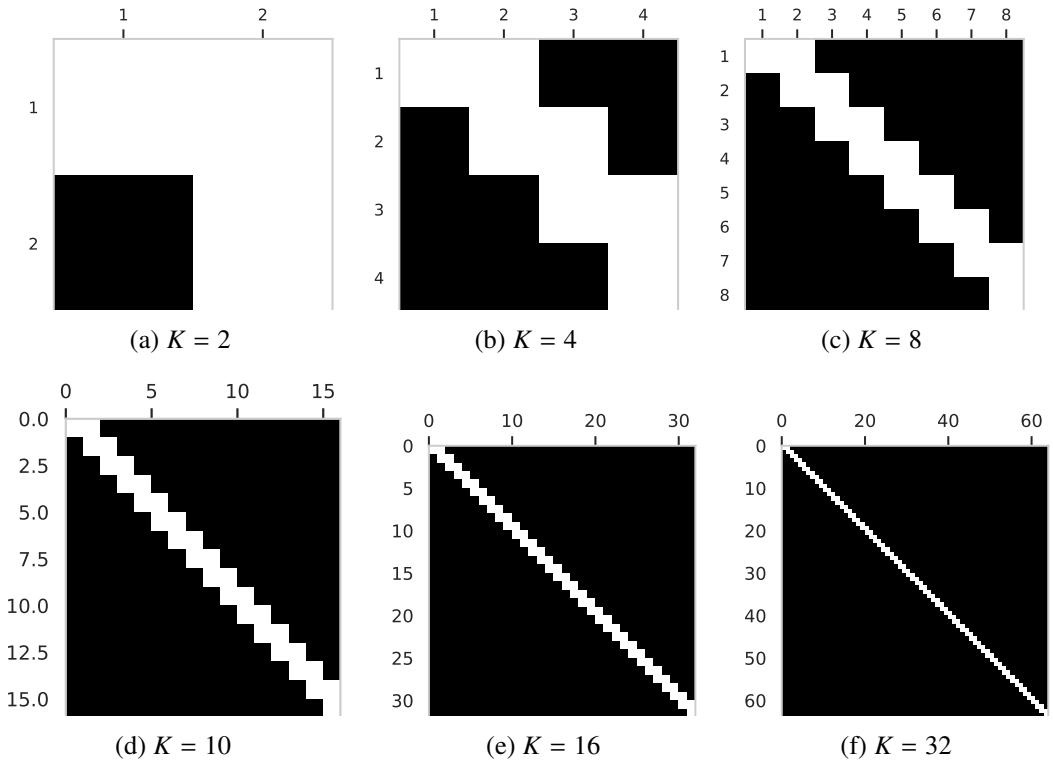
Table 5: Performance of Algorithm 3 (averaged over 10 repetitions, except for $K = 64$) in the multivariate settings of Simulation 4, measured by the variational posterior risk $r_{L_1}(\hat{Q})$ ("Risk"), the $L_1$-error of the variational posterior mean $Err_{L_1}(\hat{Q})$, the error on estimating the norms of interaction functions $Err_h(\hat{Q})$ ("Error on norms"), and the accuracy of our graph estimate $Acc_{graph}(\hat{\delta})$ and of the selected dimensionality of the interaction functions in the model-selection variational posterior, $Acc_{dim}(\hat{D})$ .

| Scenario | Graph | # Edges | # Events | # Excursions | # Local excursions |
|---|---|---|---|---|---|
| Excitation | Sparse | $2K-1$ | 24638 | 431 | 1212 |
| | Random | $3K-1$ | 27475 | 398 | 1262 |
| | Dense | $5K-6$ | 90788 | 2 | 1432 |
| Inhibition | Sparse | $2K-1$ | 22683 | 911 | 1778 |
| | Random | $3K-1$ | 24031 | 884 | 1834 |
| | Dense | $5K-6$ | 35291 | 547 | 2170 |

Table 6: Number of edges, observed events, and excursions in the different graph settings of Simulation 4 ($K = 10$).

| Scenario | Graph | Graph accuracy | Dimension accuracy | $L_1$-error |
|---|---|---|---|---|
| Excitation | Sparse | 1.00 | 1.00 | 2.91 |
| | Random | 1.00 | 1.00 | 4.00 |
| | Dense | 0.5 | 1.00 | 17.67 |
| Inhibition | Sparse | 1.00 | 1.00 | 2.62 |
| | Random | 0.99 | 1.00 | 3.44 |
| | Dense | 1.00 | 1.00 | 2.67 |

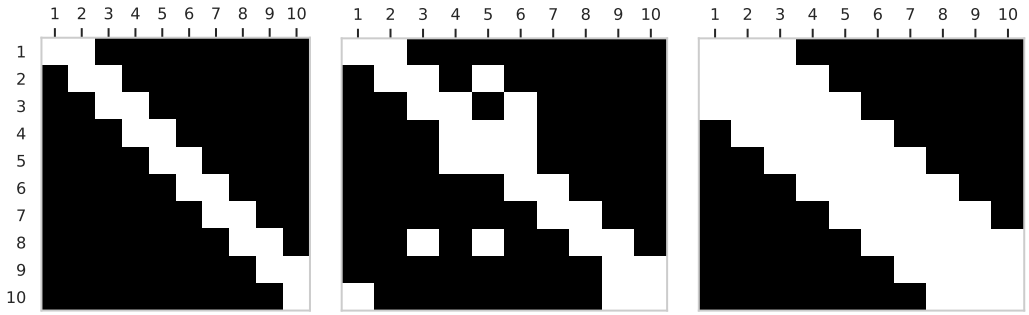
Table 7: Performance of Algorithm 3 in the different graph settings of Simulation 4 ($K = 10$). We note in that the dense graph setting, there are more parameters to estimate, and therefore non-null terms in the risk metric.



(a) $K = 2$ (b) $K = 4$ (c) $K = 8$

(d) $K = 10$ (e) $K = 16$ (f) $K = 32$

Figure 12: True graph parameter $\delta_0$ (black=0, white=1) in the sparse multivariate settings of Simulations 4 with the number of dimensions $K = 2, 4, 8, 10, 16, 32$.



(a) sparse (b) random (c) dense

Figure 13: True graph parameter $\delta_0$ (black=0, white=1) in the sparse, random, and dense settings of Simulations 4 with $K = 10$ dimensions.

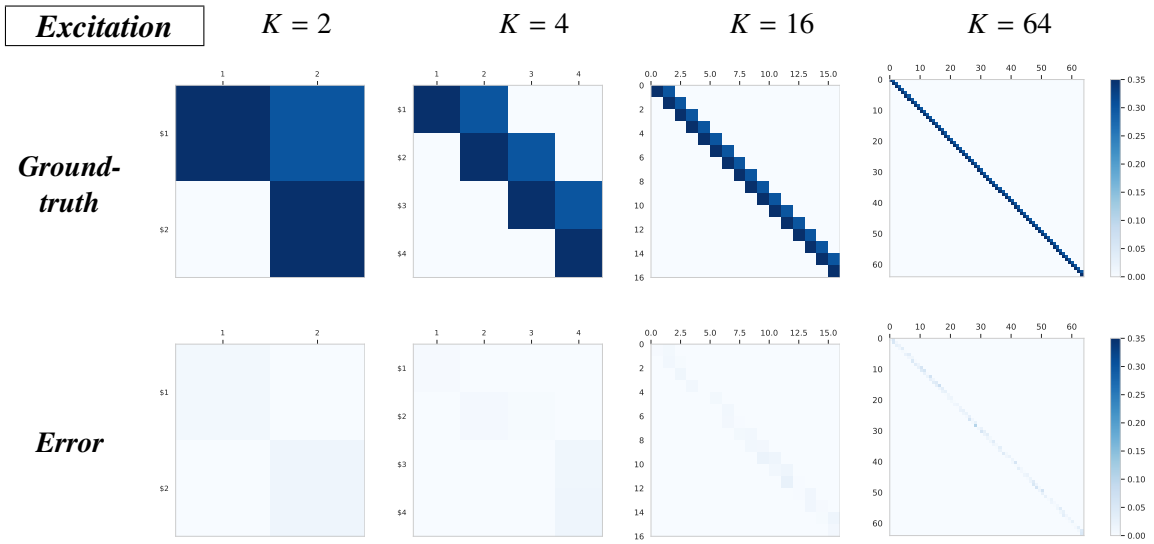
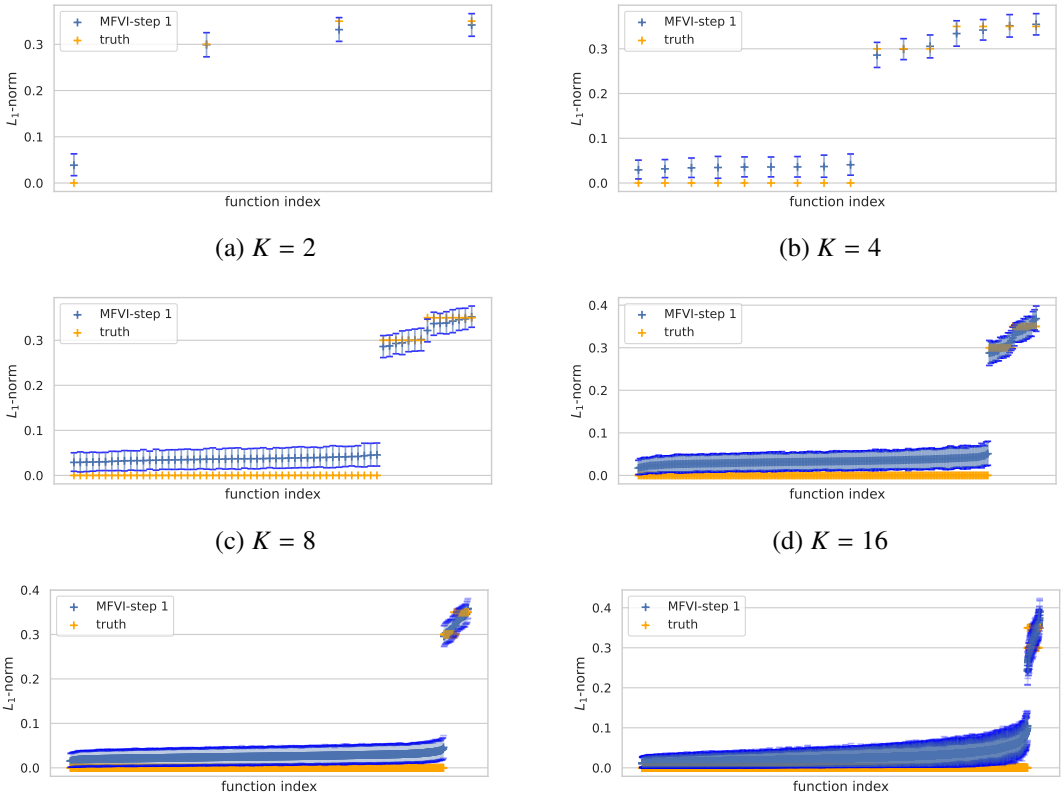



Figure 14: Heatmaps of the $L_1$-norms of the interaction functions, for the true parameter $h_0$, i.e., the entries of the matrix $S_0 = (S^0_{lk})_{l,k} = (\|h^0_{lk}\|_1)_{l,k}$ (first row) and the $L_1$-error of the model-selection variational posterior mean obtained with Algorithm 3 $Err_h(\hat{Q}_{MS})$ (second row), in the *Excitation* scenario of Simulation 4 and for $K = 2, 4, 16, 64$.

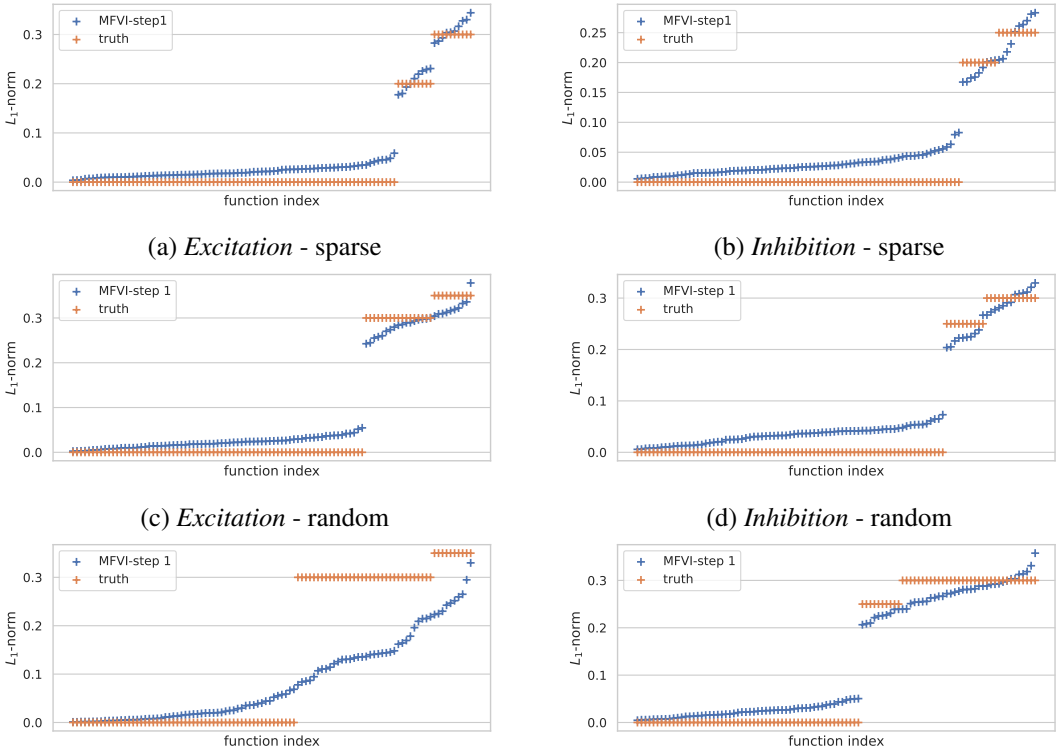



(a) $K = 2$ (b) $K = 4$

(c) $K = 8$ (d) $K = 16$

(e) $K = 32$ (f) $K = 64$

Figure 15: Estimated $L_1$-norms using the model-selection variational posterior obtained after the first step of Algorithm 3, plotted in increasing order, in the *Excitation* scenario of Simulation 4, for the models with $K = 2, 4, 8, 16, 32, 64$. In these settings, our threshold $\eta_0 = 0.15$ is included in the gap between the estimated norms close to 0 and far from 0, therefore, our gap heuristics allows to recover the true graph parameter (see Section 4.2.2).




(a) *Excitation* - sparse

(b) *Inhibition* - sparse

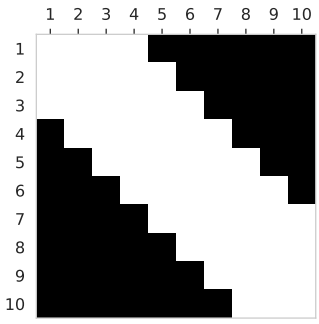
(c) *Excitation* - random

(d) *Inhibition* - random

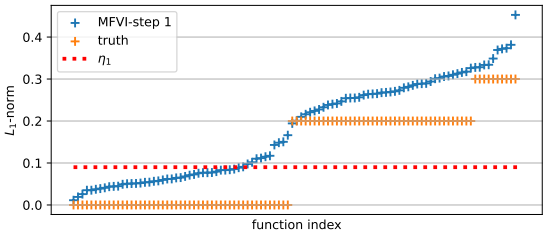
(e) *Excitation* - dense

(f) *Inhibition* - dense

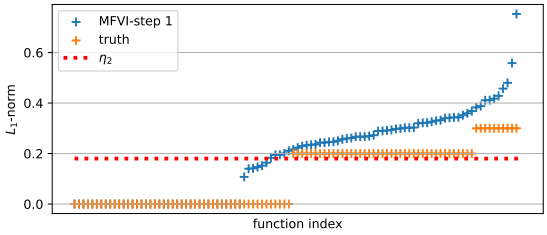
Figure 16: Estimated $L_1$-norms using the model-selection variational posterior obtained after the first step of Algorithm 3, plotted in increasing order, in the different graph settings (sparse, random, and dense $\delta_0$, see Figure 13) and scenarios of Simulation 4 with $K = 10$. We note that in the dense graph setting, although the norms are not very well estimated after the first step, the gap heuristics still allows to recover the true graph parameter.



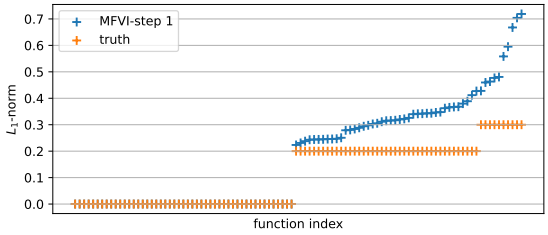

(a) True graph $\delta_0$




(b) Step 1




(c) Step 2




(d) Step 3

Figure 17: True graph $\delta_0$ and estimated norms $\hat{S}_{lk}$ using the model selection adaptive variational posterior obtained after each step of our three-step procedure, proposed for the dense graph setting of Simulation 4. In Step 1 and Step 2, we plot the data-driven thresholds $\eta_1$ and $\eta_2$, chosen with a "slope change" heuristics.

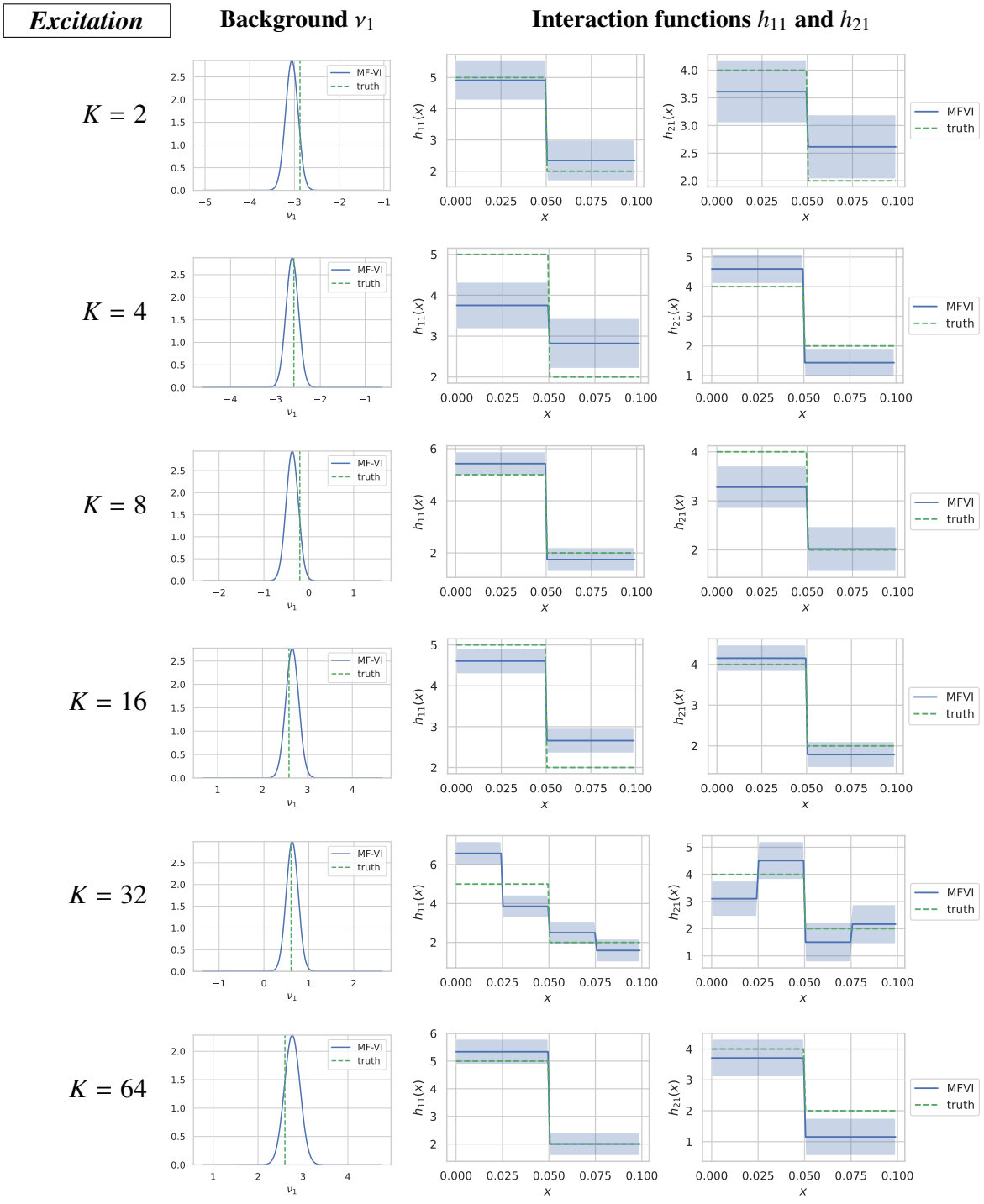



Figure 18: Model-selection variational posterior distributions on $\nu_1$ (left column) and interaction functions $h_{11}$ and $h_{21}$ (second and third columns) in the *Excitation* scenario and multivariate sigmoid models of Simulation 4, computed with our two-step mean-field variational (MF-VI) algorithm (Algorithm 3). The different rows correspond to different multivariate settings $K = 2, 4, 8, 16, 32, 64$.

### 6.5 Simulation 5: Convergence of the Two-step Variational Posterior for Varying Data Set Sizes.

In this experiment, we study the variations of performances of Algorithm 3 with increasing lengths of the observation window, i.e., increasing number of data points. We consider multidimensional data sets with $K = 10$, $T \in \{50, 200, 400, 800\}$, the same connectivity graph as in Simulation 4, and an *Excitation* and an *Inhibition* scenarios. The number of events and excursions in each data sets are reported in Table 9 in Appendix G.4.

We estimate the parameters using the model-selection variational posterior in Algorithm 3 for each data set. From Figure 19, we note that the risk decreases quite rapidly with the number of observations for $T \leqslant 400$, then the decrease between $T = 400$ and $T = 800$ is much weaker. This may be due to the fact that in this intricate model, the asymptotic regime may only be reached for larger values of $T$. We can also see from Figure 20 that the estimation of the $L_1$-norms after the first step of the algorithm improves for larger $T$, leading to a bigger gap between the small and large norms. Finally, in Figure 21 (and Figure 37 in Appendix), we plot the model-selection variational posterior and note that its mean gets closer to the ground-truth parameter and its credible set shrinks for larger $T$.

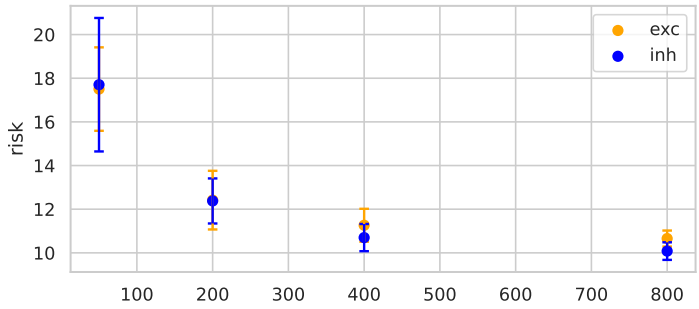
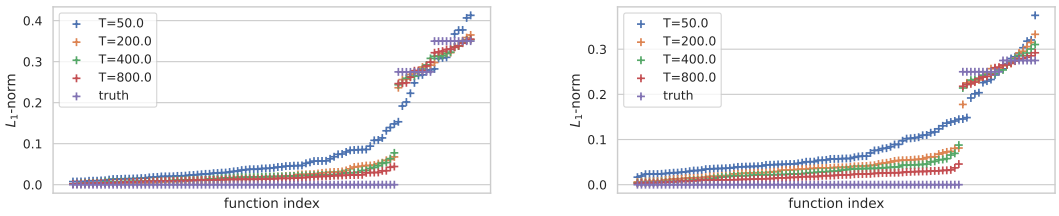


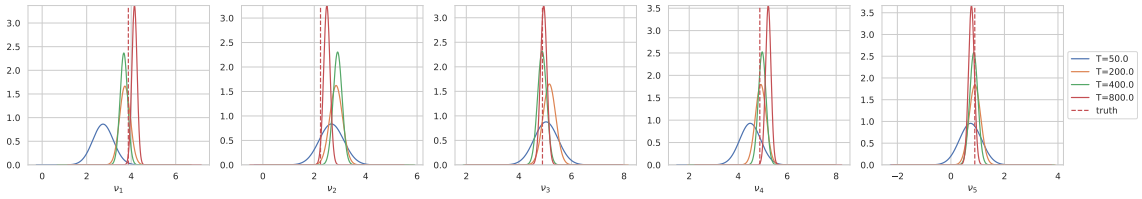

Figure 19: Performance of Algorithm 3 in terms of $L_1$-risk of the variational posterior mean versus the different data set sizes $T \in \{50, 200, 400, 800\}$ in the Excitation (exc) and Inhibition (inh) scenarios of Simulation 5 with $K = 10$. The results are averaged over 10 repetitions and we report the mean $\pm 2$ standard deviations.



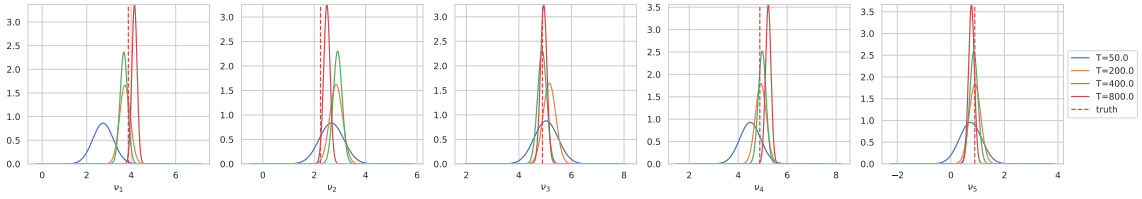

(a) *Excitation* scenario

(b) *Inhibition* scenario

Figure 20: Estimated $L_1$-norms after the first step of Algorithm 3, for different observation lengths $T$, in the *Excitation* and *Inhibition* scenarios of Simulation 5 with $K = 10$. We note that the norms are better estimated, after the first step of our algorithm, for larger $T$, leading to a larger gap between the small and large estimated norms, in both scenarios.




(a) *Excitation* scenario




(b) *Inhibition* scenario

Figure 21: Model-selection adaptive variational posterior on a subset of background rates, $(\nu_1, \ldots, \nu_5)$, for different observation lengths $T \in \{50, 200, 400, 800\}$, in the *Excitation* and *Inhibition* scenarios in Simulation 5 with $K = 10$. The variational posterior behaves as expected in this simulation: as $T$ increases, its mean gets closer to the ground-truth parameter and its variance decreases.

### 6.6 Simulation 6: Robustness to Mis-specification of the Link Function and the Memory Parameter

In this experiment, we first test the robustness of our variational method based on the sigmoid model parametrised by (37) with $\xi = (0.0, 20.0, 0.2, 10.0)$ to mis-specification of the nonlinear link functions $(\phi_k)_k$. Specifically, we set $K = 10$ and construct synthetic mis-specified data by simulating a Hawkes process where for each $k$, the link $\phi_k$ is chosen as:

- ReLU: $\phi_k(x) = (x)_+$;
- Softplus: link $\phi_k(x) = \log(1 + e^x)$;
- Mis-specified sigmoid, with unknown $\theta_k \overset{i.i.d.}{\sim} U([15, 25])$.

We also consider *Excitation* and *Inhibition* scenarios. Here, $T = 300$ in all settings.

In Figure 22, we plot the estimated $L_1$-norms after the first step of Algorithm 3 and note that there is still a gap in all settings and scenarios, although the norms are not well estimated in the case of the ReLU and softplus nonlinearities. The gaps allow to estimate well the connectivity graph parameter, but the other parameters cannot be well estimated for these two links, as can be seen from the risks in Table 8. Nonetheless, the sign of the interaction functions is well recovered in all settings.

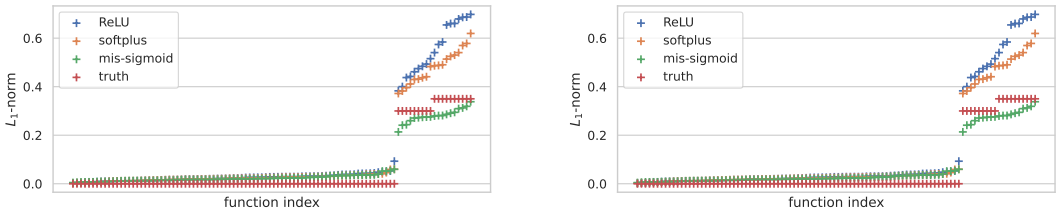
Then, we test the robustness of our variational method to mis-specification of the memory parameter $A$, assumed to be known in our framework. We recall that $A$ corresponds to the upper bound of the support of the interaction functions. For this experiment, we generate data from the sigmoid Hawkes process with $K = 10$ and with ground-truth parameter $A_0 = 0.1$, in two sets of parameters corresponding to an *Excitation* and an *Inhibition* scenarios. Here, we set $T = 500$ and apply our variational method (Algorithm 3) with $A \in \{0.5, 0.1, 0.2, 0.4\}$.

In Figure 23, we plot the estimated $L_1$-norms of the interaction functions, after the first step of Algorithm 3, when using the different values of $A$. We note that when $A$ is smaller than $A_0$, the norms of the non-null functions are underestimated, while if $A$ is larger than $A_0$, the norms are slightly overestimated. We note that, in all settings, the graph can be well estimated with the gap heuristics (see Figure 40 in Appendix). The model-selection variational posterior on a subset of the interaction functions is plotted in Figure 24. We note that for $A = 0.05 = A_0/2$, only the first part of the functions can be estimated, while for $A > A_0$, the mean estimate is close to 0 on the upper part of the support. Nonetheless, in the latter case, the dimensionality of the true functions is not well-recovered.

In conclusion, this experiment shows that our algorithm is robust to the mis-specification of the nonlinear link functions and the memory parameter, for estimating the connectivity graph and the sign of the interaction functions when the latter are either non-negative or non-positive. Nonetheless, the other parameters of the Hawkes model cannot be well recovered.

## 7. Discussion

In this paper, we proposed novel adaptive variational Bayes methods for sparse and high-dimensional Hawkes processes, and provided a general theoretical analysis of variational methods when the dimension $K$ of the process is fixed but possibly large. We notably obtained variational posterior concentration rates, under easily verifiable conditions on the prior and approximating family that



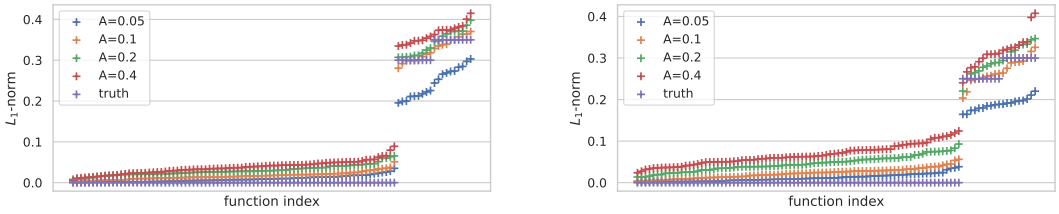

(a) *Excitation* scenario

(b) *Inhibition* scenario

Figure 22: Estimated $L_1$-norms after the first step of Algorithm 3, in the mis-specified settings of Simulation 6. In this simulation, the link functions are set to $\phi_k(x) = 20\sigma(0.2(x - 10))$, $\forall k$, in our algorithm, while the data sets are generated from a Hawkes process with ReLU, softplus, or a mis-specified sigmoid (mis-sigmoid) link functions, in *Excitation* and *Inhibition* scenarios. We note that for the ReLU and softplus link, the norms are not well estimated after the first step, nonetheless, our gap heuristic can still recover the true graph parameter.

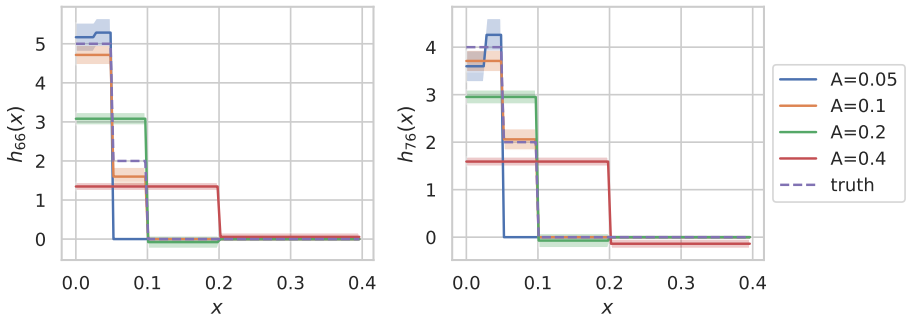
| Scenario | Link | Graph accuracy | Dimension accuracy | $L_1$-risk |
|---|---|---|---|---|
| Excitation | ReLU | 1.00 | 1.00 | 49.58 |
| | Softplus | 1.00 | 1.00 | 34.27 |
| | Mis-specified sigmoid | 1.00 | 1.00 | 19.69 |
| Inhibition | ReLU | 1.00 | 1.00 | 59.95 |
| | Softplus | 1.00 | 1.00 | 33.94 |
| | Mis-specified sigmoid | 0.99 | 1.00 | 15.78 |

Table 8: Performance of Algorithm 3 for the different mis-specified settings and scenarios of Simulation 6 ($K = 10$). We note that the graph parameter and the dimensionality are still recovered in these cases, although, the other parameters cannot be well estimated, as can be seen from the large risk.

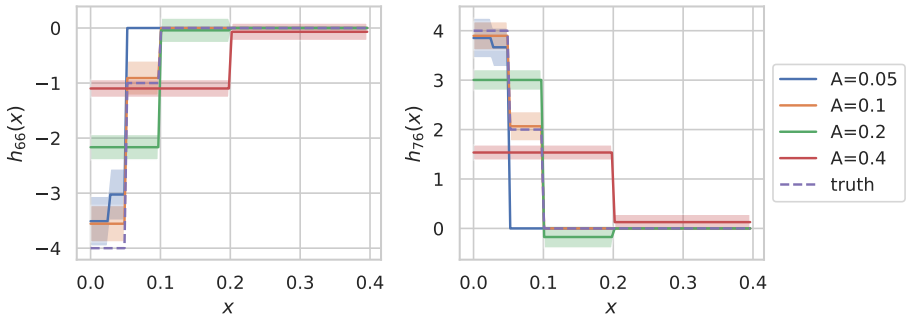



(a) *Excitation* scenario (b) *Inhibition* scenario

Figure 23: Estimated $L_1$-norms of the interaction functions after the first step of Algorithm 3 specified with different values of the memory parameter $A = 0.05, 0.1, 0.2, 0.4$ containing the true memory parameter $A_0 = 0.1$, in the scenarios of Simulation 6. In all cases, we still observe a gap, although the norms are under-estimated (resp. over-estimated) for $A = 0.05$ (resp. $A = 0.4$).




(a) *Excitation* scenario




(b) *Inhibition* scenario

Figure 24: Model-selection variational posterior on the interaction functions $h_{66}$ and $h_{76}$ obtained with Algorithm 3, specified with different values of the memory parameter $A = 0.05, 0.1, 0.2, 0.4$, in the scenarios of Simulation 6 with $K = 10$ and true memory parameter $A_0 = 0.1$. We note that the estimation of the interaction functions is deteriorated when $A$ is mis-specified, however the signs of the functions are still recovered.

we validated commonly used inference set-ups. Our general theory holds in particular in the sigmoid Hawkes model, for which we developed adaptive variational mean-field algorithms, which improve existing ones by their ability to infer the graph parameter and the dimensionality of the interaction functions. Moreover, we demonstrated on simulated data that our most computationally efficient algorithm is able to scale up to high-dimensional processes.

Nonetheless, our theory does not yet cover the setting where $K$ can grow to infinity with the sample size, which is of interest in applications of Hawkes processes to social network analysis and neuroscience. In this limit, previous works have considered sparse models (Cai et al., 2024; Bacry et al., 2020; Chen et al., 2017) and mean-field settings (Pfaffelhuber et al., 2022). We would therefore be interested in extending our results to these models. Moreover, our empirical study shows that the credible sets of variational distributions do not always have good coverage, an observation that sometimes also holds for the posterior distribution. Therefore, it is left for future work to study the property of (variational) posterior credible regions, and potentially design post-processing methods of the latter to improve coverage in practice. Additionally, the thresholding approach for estimating the graph in our two-step adaptive variational procedure could be further explored, in particular, in dense settings.

Finally, it would be of practical interest to develop variational algorithms beyond the sigmoid model, e.g., for the ReLU and softplus Hawkes models. While in the sigmoid model, the conjugacy of the mean-field variational posterior using data augmentation leads to particularly efficient algorithms, it is unlikely that such convenient forms could be obtained for more general models. A potential approach for other models could be to parametrise variational families with normalising flows, as it is for instance done for cut posteriors in Carmona and Nicholls (2022).

## Acknowledgments

The project leading to this work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 834175). The project is also partially funded by the EPSRC via the CDT OxWaSP. The authors would like to thank the anonymous referees and the Associate Editor for valuable comments and suggestions.

## Appendix A. Mean-Field and Model-Selection Variational Inference

In this section, we first recall some general notions on mean field variational Bayes and model selection variational Bayes, then present additional details on the construction of variational families in the case of multivariate Hawkes processes.

### A.1 Mean-Field Approximations

In a general inference context, when the parameter of interest, say $\vartheta$, is decomposed into $D$ blocks, $\vartheta = (\vartheta_1, \dots, \vartheta_D)$ with $D > 1$, a common choice of variational class is a mean-field family that can be defined as $\mathcal{V}_{MF} = \left\{Q;\ dQ(\vartheta) = \prod_{d=1}^{D} dQ_d(\vartheta_d)\right\}$. In this case, the mean-field variational posterior distribution corresponds to $\hat{Q} = \arg\min_{Q \in \mathcal{V}_{MF}} KL(Q||\Pi(.|N)) = \prod_{d=1}^{D} \hat{Q}_d$. Note that the mean-field family removes some dependencies between blocks of coordinates of the parameter in the approximated posterior distribution.

Assuming that the mean-field variational posterior distribution has a density with respect to a dominating measure $\mu = \prod_d \mu_d$, with a slight abuse of notation, we denote $\hat{Q}$ both the distribution and density with respect to $\mu$. An interesting result from Bishop (2006) is that the mean-field variational posterior distribution verifies, for each $d \in [D]$,

$$\hat{Q}_d(\vartheta_d) \propto \exp\left\{\mathbb{E}_{\hat{Q}_{-d}}[\log p(\vartheta, N)]\right\}, \tag{39}$$

where $p(\vartheta, N)$ is the joint density of the observations and the parameter with respect to $\prod_d \mu_d \times \mu_N$ with $\mu_N$ the data density, and $\hat{Q}_{-d} := \prod_{d' \neq d} \hat{Q}_{d'}$. This property (39) can be used to design efficient algorithms for computing the variational posterior, such as the coordinate-ascent variational inference algorithm.

In a general setting where the log-likelihood function of the nonlinear Hawkes model can be augmented with some latent variable $z \in \mathcal{Z}$ (see for instance Zhou et al. (2021, 2022); Malem-Shinitski et al. (2021)), with $\mathcal{Z}$ the latent parameter space, the augmented log-likelihood $L_T^A(f, z)$ leads to an *augmented* posterior distribution, defined as

$$\Pi_A(B|N) = \frac{\int_B \exp(L_T^A(f,z)) d(\Pi(f) \times \mathbb{P}_A(z))}{\int_{\mathcal{F}\times\mathcal{Z}} \exp(L_T^A(f,z)) d(\Pi(f) \times \mathbb{P}_A)(z)}, \quad B \subset \mathcal{F} \times \mathcal{Z},$$

where $\mathbb{P}_A$ is a prior distribution on $z$ which has a density with respect to a dominating measure $\mu_z$. Recalling the mean-field variational from Section 4.1 defined as

$$\mathcal{V}_{AMF} = \{Q : \mathcal{F} \times \mathcal{Z} \to [0,1];\ Q(f,z) = Q_1(f)Q_2(z)\},$$

the augmented mean-field variational posterior corresponds to

$$\hat{Q}_{AMF}(f,z) := \arg\min_{Q \in \mathcal{V}_{AMF}} KL(Q(f,z)||\Pi_A(f,z|N)) =: \hat{Q}_1(f)\hat{Q}_2(z), \tag{40}$$

and, using property (39), verifies

$$\hat{Q}_1(f) \propto \exp\left\{\mathbb{E}_{\hat{Q}_2}[\log p(f,z,N)]\right\}, \quad \hat{Q}_2(z) \propto \exp\left\{\mathbb{E}_{\hat{Q}_1}[\log p(f,z,N)]\right\}, \tag{41}$$

where $p(f, z, N)$ is the joint density of the parameter, the latent variable, and the observations with respect to the measure $\prod_d \mu_d \times \mu_z \times \mu_N$.

### A.2 Model-Selection Variational Posterior

In this section, we present two model-selection variational approaches to approximate the posterior by an adaptive variational posterior distribution. We recall from our construction in Section 4.2 that our parameter $f$ of the Hawkes processes is indexed by a model $m$ of hyperparameters in the form $m = (\delta, J_{lk}, (l,k) \in \mathcal{I}(\delta))$, where $\mathcal{I}(\delta) = \{(l,k);\ \delta_{lk} = 1\}$ is the set of non null functions.

In a model-selection variational approach, one can consider a set of candidate models $\mathcal{M}$ and for any $m \in \mathcal{M}$, a class of variational distributions on $f$ with model $m$, denoted $\mathcal{V}^m$. Then, one can define the total variational class as $\mathcal{V} = \cup_{m\in\mathcal{M}}\{\{m\} \times \mathcal{V}^m\}$, which contains distributions on $f$ localised on one model. Then, given $\mathcal{V}$ and as shown for instance in Zhang and Gao (2020), the variational posterior distribution has the form

$$\hat{Q} := \hat{Q}_{\hat{m}}, \quad \hat{m} := \arg\max_{m\in\mathcal{M}} ELBO(\hat{Q}^m),$$

where $\hat{Q}^m = \arg\min_{Q\in\mathcal{V}^m} KL(Q|||\Pi(.|N))$ and $ELBO(\cdot)$ is called the *evidence lower bound (ELBO)*, defined as

$$ELBO(Q) := \mathbb{E}_Q\left[\log \frac{p(f,z,N)}{Q(f,z)}\right], \quad Q \in \mathcal{V}. \tag{42}$$

The ELBO is a lower bound of the marginal log-likelihood $p(N)$.

An alternative model-selection variational approach consists in constructing a model-averaging variational posterior, also called *adaptive* in Ohn and Lin (2024), as a mixture of distributions over the different models, i.e.,

$$\hat{Q} = \sum_{m\in\mathcal{M}} \hat{\gamma}_m \hat{Q}_m, \tag{43}$$

where $\{\hat{\gamma}_m\}_{m\in\mathcal{M}}$ are marginal probabilities defined as

$$\hat{\gamma}_m = \frac{\Pi_m(m)\exp\left\{ELBO(\hat{Q}_m)\right\}}{\sum_{m\in\mathcal{M}}\Pi_m(m)\exp\left\{ELBO(\hat{Q}_m)\right\}}, \quad \forall m \in \mathcal{M}. \tag{44}$$

In this strategy, the approximating family of distributions corresponds to

$$\mathcal{V} = \left\{\sum_{m\in\mathcal{M}} \alpha_m Q_m; \sum_m \alpha_m = 1,\ \alpha_m \geqslant 0,\ Q_m \in \mathcal{V}^m,\ \forall m\right\}.$$

## Appendix B. Data Augmentation in the Sigmoid Hawkes Model

In this section, we recall the latent variable augmentation strategy and the definition of the augmented mean-field variational distribution in sigmoid-type Hawkes processes, proposed in previous work (Zhou et al., 2022; Malem-Shinitski et al., 2021). In our method in Section 4.2, we use this construction to efficiently compute an approximated posterior distribution on $\mathcal{F}_m \subset \mathcal{F}$, on parameters $f$ within a model $m = (\delta, J_{lk}; (l,k) \in \mathcal{I}(\delta))$.

The first data augmentation step consists in re-writing the sigmoid function as a mixture of Polya-Gamma random variables (Polson et al., 2013), i.e.,

$$\sigma(x) = \mathbb{E}_{\omega\sim p_{PG}(.;1,0)}\left[e^{g(\omega,x)}\right] = \int_0^{+\infty} e^{g(\omega,x)} p_{PG}(\omega;1,0)d\omega, \quad g(\omega,x) = -\frac{\omega x^2}{2} + \frac{x}{2} - \log 2, \tag{45}$$

with $p_{PG}(.;1,0)$ the Polya-Gamma density. We recall that $p_{PG}(.;1,0)$ is the density of the random variable

$$\frac{1}{2\pi^2}\sum_{k=1}^{\infty}\frac{g_k}{(k-1/2)^2}, \quad g_k \overset{\text{i.i.d.}}{\sim} Gamma(1,1),$$

and that the *tilted* Polya-Gamma distribution is defined as

$$p_{PG}(\omega;1,c) = \cosh\left(\frac{c}{2}\right)\exp\left\{-\frac{c^2\omega}{2}\right\}p_{PG}(\omega;1,0), \quad c \geqslant 0, \tag{46}$$

where cosh denotes the hyperbolic cosine function. With a slight abuse of notation, we re-define the linear intensity (2) as $\tilde{\lambda}_t^k(f) = \alpha\Big(\nu_k + \sum_{l=1}^K \int_{-\infty}^{t^-} h_{lk}(t-s)dN_s^l - \eta\Big)$, so that we have $\lambda_t^k(f) = \theta_k\sigma(\tilde{\lambda}_t^k(f)), t \in \mathbb{R}$. For any $k \in [K]$, let $N_k := N^k[0,T]$ and $T_1^k, \ldots, T_{N_k}^k \in [0,T]$ be the times of events at component $N^k$. Now, let $\omega = (\omega_i^k)_{k\in[K], i\in[N_k]}$ be a set of latent variables such that

$$\omega_i^k \overset{\text{i.i.d.}}{\sim} p_{PG}(\cdot;1,0), \quad i \in [N_k], \quad k \in [K].$$

Then, using (45), an *augmented* log-likelihood function can be defined as

$$L_T(f,\omega;N) = \sum_{k\in[K]}\left\{\sum_{i\in[N_k]}\Big(\log\theta_k + g(\omega_i^k, \tilde{\lambda}_{T_i^k}(f)) + \log p_{PG}(\omega_i^k;1,0)\Big) - \int_0^T \theta_k\sigma(\tilde{\lambda}_t^k(f))dt\right\}, \tag{47}$$

and, using that $\sigma(x) = 1-\sigma(-x)$, the integral term on the RHS in (47) can be re-written as

$$\int_0^T \theta_k\sigma(\tilde{\lambda}_t^k(f))dt = \int_0^T\int_0^\infty \theta_k\left[1 - e^{g(\bar{\omega},-\tilde{\lambda}_t^k(f))}\right]p_{PG}(\bar{\omega};1,0)d\bar{\omega}dt.$$

Secondly, Campbell's theorem (Daley and Vere-Jones, 2007; Kingman, 1993) is applied. We first recall here its general formulation. For a Poisson point process $\bar{Z}$ on a space $\mathcal{X}$ with intensity measure $\Lambda : \mathcal{X} \to \mathbb{R}^+$, and for any function $\zeta : \mathcal{X} \to \mathbb{R}$, it holds true that

$$\mathbb{E}\left[\prod_{x\in\bar{Z}} e^{\zeta(x)}\right] = \exp\left\{\int (e^{\zeta(x)} - 1)\Lambda(dx)\right\}. \tag{48}$$

Therefore, for each $k$, we denote by $\bar{Z}^k$ the realisation of a marked Poisson point process on $\mathcal{X} = ([0,T], \mathbb{R}^+)$ with intensity measure $\Lambda^k(t,\omega) = \theta_k p_{PG}(\omega;1,0)$ and we also denote by $\mathbb{P}_{\bar{Z}^k}$ its law. Using Campbell's theorem with $\zeta(t,\omega) := g(\omega, -\tilde{\lambda}_t^k(f))$, conditionally on the realisation of the process $N$ and using that $\sigma(x) = 1-\sigma(-x)$, it holds that

$$\mathbb{E}_{\bar{Z}^k}\Big[\prod_{(\bar{T}_j^k,\bar{\omega}_j^k)\in\bar{Z}^k} e^{g(\bar{\omega}_j^k, -\tilde{\lambda}_{\bar{T}_j^k}^k(f))} \mid N\Big] = \exp\left\{\int_0^T\int_0^\infty \theta_k\left(e^{g(\bar{\omega},-\tilde{\lambda}_t^k(f))} - 1\right)p_{PG}(\bar{\omega};1,0)d\bar{\omega}dt\right\},$$

where we denote by $\bar{Z}^k = (\bar{T}_1^k, \bar{\omega}_1^k), \ldots, (\bar{T}_1^k, \bar{\omega}_{\bar{Z}_k}^k) \in [0,T]\times\mathbb{R}_+$ the times and marks in $\bar{Z}^k$ and by $\bar{Z}_k := \bar{Z}^k[0,T]$ the number of marked points. Moreover, we denote by $\bar{Z} = (\bar{T}_i^k, \bar{\omega}_i^k, 1 \leqslant i \leqslant \bar{Z}_k, 1 \leqslant k \leqslant K)$ the set containing the $K$ realisations of marked Poisson point processes. Then, replacing

the integral term in (47) by a product over the observation $\bar{Z}$, the *doubly augmented* log-likelihood function corresponds to

$$L_T(f,\omega,\bar{Z};N) = \sum_{k\in[K]}\left\{\sum_{i\in[N_k]}\left[\log\theta_k + g(\omega_i^k,\tilde{\lambda}_{T_i^k}(f)) + \log p_{PG}(\omega_i^k;1,0)\right] + \sum_{j\in[\bar{Z}_k]}\left[\log\theta_k + g(\bar{\omega}_j^k,-\tilde{\lambda}_{\bar{T}_j}(f)) + \log p_{PG}(\bar{\omega}_j^k;1,0)\right] - \theta_k T\right\}.$$

The previous augmented log-likelihood function, and the prior distribution $\Pi$ on the parameter and the latent variables distribution $\mathbb{P}_A = p_{PG}(.|1,0)\times\mathbb{P}_{\bar{Z}}$, allow to construct an *augmented* posterior distribution proportional to

$$\Pi(f,\omega,\bar{Z}|N) \propto \prod_k\left\{\prod_{i\in[N_k]}\theta_k e^{g(\omega_i^k,\tilde{\lambda}_{T_i^k}(f))}p_{PG}(\omega_i^k;1,0)\times\prod_{j\in[\bar{Z}^k]}\theta_k e^{g(\bar{\omega}_j^k,-\tilde{\lambda}_{\bar{T}_j}(f))}p_{PG}(\bar{\omega}_j^k;1,0)\right\}\times\Pi(f). \tag{49}$$

## Appendix C. Analytical Derivation in the Sigmoid Hawkes Model

In this section we provide the proof of Proposition 3 and the analytical derivation of the ELBO in the context of sigmoid Hawkes processes.

### C.1 Proof of Proposition 3

We recall that Proposition 3 states the analytic forms of the conditional updates in Algorithm 1, the mean-field variational algorithm with fixed dimensionality described in Section 4.1.

For ease of exposition, in this section we consider a model $m$ and a dimension $k$ and we drop the indices $k$ and $m$, e.g., we use the notation $Q_1, Q_2$ for the variational factors. In the following computation, we use the notation $c$ to denote a generic constant which value can vary from one line to the other. For simplicity, we also assume that $J := J_1 = \cdots = J_K$ and we recall that $\phi_k(x) = \theta_k\sigma(\alpha(x-\eta))$.

From the definition of the augmented posterior (49), we first note that

$$\begin{aligned}\log p(f,N,\omega,\bar{Z}) &= \log\Pi(f,\omega,\bar{Z}|N) + \log p(N) = L_T(f,\omega,\bar{Z};N) + \log\Pi(f) + \log p(N) + c\\ &= \log p(\omega|f,N) + \log p(\bar{Z}|f,N) + \log\Pi(f) + \log p(N) + c.\end{aligned} \tag{50}$$

In the previous equality we have used the facts that $p(\omega|f,N,\bar{Z}) = p(\omega|f,N)$ and $p(\bar{Z}|f,N,\omega) = p(\bar{Z}|f,N)$. We recall our notation $H(t) = (H^0(t),H^1(t),\ldots,H^K(t)) \in \mathbb{R}^{KJ+1}$, $t\in\mathbb{R}$, where for

$k \in [K]$, $H^k(t) = (H_j^k(t))_{j=1,\dots,J}$. In the following, $H(t)$ denotes $H^k(t)$ for the chosen $k$. We have that

$$\begin{aligned}
\mathbb{E}_{Q_2}[\log p(\omega|f,N)] &= \mathbb{E}_{Q_2}\left[\sum_{i\in[N]} g(\omega_i, \tilde{\lambda}_{T_i}(f))\right] + c = \mathbb{E}_{Q_2}\left[\sum_{i\in[N]} -\frac{\omega_i \tilde{\lambda}_{T_i}(f)^2}{2} + \frac{\tilde{\lambda}_{T_i}(f)}{2}\right] + c \\
&= \mathbb{E}_{Q_2}\left[\sum_{i\in[N]} -\frac{\omega_i \alpha^2 (f^T H(T_i) H(T_i)^T f - 2\eta H(T_i)^T f + \eta^2)}{2} + \frac{\alpha H(T_i)^T f}{2}\right] + c \\
&= \mathbb{E}_{Q_2}\left[-\frac{1}{2}\sum_{i\in[N]} \left\{\omega_i \alpha^2 f^T H(T_i) H(T_i)^T f - \alpha(2\omega_i \alpha\eta + 1) H(T_i)^T f + \omega_i \alpha^2 \eta^2\right\}\right] + c \\
&= -\frac{1}{2}\sum_{i\in[N]} \left\{\mathbb{E}_{Q_2}[\omega_i]\alpha^2 f^T H(T_i) H(T_i)^T f - \alpha(2\mathbb{E}_{Q_2}[\omega_i]\alpha\eta + 1) H(T_i)^T f + \mathbb{E}_{Q_2}[\omega_i]\alpha^2\eta^2\right\} + c.
\end{aligned}$$

Moreover, we also have that

$$\begin{aligned}
\mathbb{E}_{Q_2}[\log p(\bar{Z}|f,N)] &= \mathbb{E}_{Q_2}\left[-\frac{1}{2}\sum_{j\in[\bar{Z}]} \left\{\bar{\omega}_j \alpha^2 f^T H(\bar{T}_j) H(\bar{T}_j)^T f - \alpha(2\bar{\omega}_j \alpha\eta - 1) H(\bar{T}_j)^T f + \bar{\omega}_j \alpha^2 \eta^2\right\}\right] + c \\
&= \int_0^T \int_0^\infty \left[-\frac{1}{2}\left(\bar{\omega}\alpha^2 f^T H(t) H(t)^T f - \alpha(2\bar{\omega}\alpha\eta - 1) H(t)^T f + \bar{\omega}\alpha^2\eta^2\right)\right] \Lambda(t,\bar{\omega}) d\bar{\omega} dt + c \\
&= -\frac{1}{2}\left[f^T\left(\alpha^2 \int_0^T \int_0^\infty \bar{\omega} H(t) H(t)^T \Lambda(t,\bar{\omega}) d\bar{\omega} dt\right) f \right. \\
&\qquad \left. + f^T\left(\alpha \int_0^T \int_0^\infty (2\bar{\omega}\alpha\eta - 1) H(t)^T \Lambda(t,\bar{\omega}) d\bar{\omega} dt\right)\right] + c.
\end{aligned}$$

Besides, we have $\mathbb{E}_{Q_2}[\log \Pi(f)] = -\frac{1}{2} f^T \Sigma^{-1} f + f^T \Sigma^{-1}\mu + c$. Therefore, using (23), we obtain that

$$\begin{aligned}
\log Q_1(f) &= -\frac{1}{2}\left[f^T\left(\alpha^2 \sum_{i\in[N]} \mathbb{E}_{Q_2}[\omega_i] H(T_i) H(T_i)^T + \alpha^2 \int_0^T \int_0^\infty \bar{\omega} H(t) H(t)^T \Lambda(t,\bar{\omega}) d\bar{\omega} dt + \Sigma^{-1}\right) f \right. \\
&\qquad \left. - f^T\left(\alpha \sum_{i\in[N]} (2\mathbb{E}_{Q_2}[\omega_i]\alpha\eta + 1) H(T_i)^T + \alpha \int_0^T \int_0^\infty (2\bar{\omega}\alpha\eta - 1) H(t)^T \Lambda(t,\bar{\omega}) d\bar{\omega} dt + 2\Sigma^{-1}\mu\right)\right] + c \\
&=: -\frac{1}{2}(f - \tilde{\mu})^T \tilde{\Sigma}^{-1} (f - \tilde{\mu}) + c,
\end{aligned}$$

therefore $Q_1(f)$ is a normal distribution with mean vector $\tilde{\mu}$ and covariance matrix $\tilde{\Sigma}$ given by

$$\tilde{\Sigma}^{-1} = \alpha^2 \sum_{i\in[N]} \mathbb{E}_{Q_2}[\omega_i] H(T_i) H(T_i)^T + \alpha^2 \int_0^T \int_0^\infty \bar{\omega} H(t) H(t)^T \Lambda(t,\bar{\omega}) d\bar{\omega} dt + \Sigma^{-1}, \tag{51}$$

$$\tilde{\mu} = \frac{1}{2}\tilde{\Sigma}\left[\alpha \sum_{i\in[N]} (2\mathbb{E}_{Q_2}[\omega_i]\alpha\eta + 1) H(T_i)^T + \alpha \int_0^T \int_0^\infty (2\bar{\omega}\alpha\eta - 1) H(t)^T \Lambda(t,\bar{\omega}) d\bar{\omega} dt + 2\Sigma^{-1}\mu\right]. \tag{52}$$

For $Q_2(\omega, \bar{Z})$, we first note that using (23) and (50), we have $Q_2(\omega, \bar{Z}) = Q_{21}(\omega)Q_{22}(\bar{Z})$. Using the same computation as Donner and Opper (2018)) Appendices B and D, one can then show that

$$Q_{21}(\omega) = \prod_{i\in[N]} p_{PG}(\omega_i|1, \underline{\lambda}_{T_i}),$$
$$\underline{\lambda}_t = \sqrt{\mathbb{E}_{Q_1}[\tilde{\lambda}_t(f)^2]} = \alpha^2 \sqrt{H(t)^T\tilde{\Sigma}H(t) + (H(t)^T\tilde{\mu})^2 - 2\eta H(t)^T\tilde{\mu} + \eta^2}, \quad \forall t \in [0, T],$$

and that $Q_{22}$ is a marked Poisson point process measure on $[0, T] \times \mathbb{R}^+$ with intensity

$$\Lambda(t, \bar{\omega}) = \theta e^{\mathbb{E}_{Q_1}[g(\bar{\omega}, -\tilde{\lambda}_t(f)]} p_{PG}(\bar{\omega}; 1, 0) = \theta \frac{\exp(-\frac{1}{2}\mathbb{E}_{Q_1}[\tilde{\lambda}_t(f)])}{2\cosh\frac{\underline{\lambda}_t(f)}{2}} p_{PG}(\bar{\omega}|1, \underline{\lambda}_t(f))$$
$$= \theta\sigma(-\underline{\lambda}_t)\exp\left\{\frac{1}{2}(\underline{\lambda}_t(f) - \mathbb{E}_{Q_1}[\tilde{\lambda}_t(f)])\right\} p_{PG}(\bar{\omega}|1, \underline{\lambda}_t)$$
$$\mathbb{E}_{Q_1}[\tilde{\lambda}_t(f)] = \alpha(H(t)^T\tilde{\mu} - \eta).$$

Therefore, we have that

$$\mathbb{E}_{Q_1}[\omega_i] = \frac{1}{2\underline{\lambda}_{T_i}} \tanh\left(\frac{\underline{\lambda}_{T_i}}{2}\right), \quad \forall i \in [N].$$

### C.2 Analytic Formulas of the ELBO

In this section, we provide the derivation of the evidence lower bound $(ELBO(\hat{Q}_k^m))_k$ for a mean-field variational distribution $\hat{Q}_m(f, \bar{Z}) = \hat{Q}_1^m(f)\hat{Q}_2^m(\bar{Z})$ in a fixed model $m = (\delta, D)$. For ease of exposition, we drop the subscript $m$ and $k$. From (42), we have

$$ELBO(\hat{Q}) = \mathbb{E}_{\hat{Q}}\left[\log \frac{p(f, \omega, \bar{Z}, N)}{\hat{Q}_1(f)\hat{Q}_2(\omega, \bar{Z})}\right]$$
$$= \mathbb{E}_{\hat{Q}_2}\left[-\log \hat{Q}_2(\omega, \bar{Z})\right] + \mathbb{E}_{\hat{Q}_2}\left[\mathbb{E}_{\hat{Q}_1}\left[\log p(f, \omega, \bar{Z}, N)\right]\right] + \mathbb{E}_{\hat{Q}_1}[-\log \hat{Q}_1(f)].$$

Now using the notation of Section 4.1, we first note that defining $K(t) := H(t)H(t)^T$, we have that

$$\mathbb{E}_{\hat{Q}_1}[\tilde{\lambda}_{T_i}(f)^2] = tr(K(t)\tilde{\Sigma}) + \tilde{\mu}^T K(t)\tilde{\mu}$$
$$\mathbb{E}_{\hat{Q}_1}\left[\log \mathcal{N}(f; \mu, \Sigma)\right] = -\frac{1}{2}tr(\Sigma^{-1}\tilde{\Sigma}) - \frac{1}{2}\tilde{\mu}^T\Sigma^{-1}\tilde{\mu} + \tilde{\mu}^T\Sigma^{-1}\mu - \frac{1}{2}\mu^T\Sigma^{-1}\mu - \frac{1}{2}\log|2\pi\Sigma|.$$

Moreover, we have

$$\mathbb{E}_{\hat{Q}_1}[\log \hat{Q}_1(f)] = -\frac{|m|}{2} - \frac{1}{2}\log|2\pi\tilde{\Sigma}|.$$

Using that for any $c > 0$, $p_{PG}(\omega; 1, c) = e^{-c^2\omega/2}\cosh(c/2)p_{PG}(\omega; 1, 0)$, we also have

$$\begin{aligned}
\mathbb{E}_{\hat{Q}_2}\left[-\log \hat{Q}_2(\omega, \bar{Z})\right] &= \sum_{i\in[N]} -\mathbb{E}_{\hat{Q}_2}[\log p_{PG}(\omega_i, 1, 0)] + \frac{1}{2}\mathbb{E}_{\hat{Q}_2}[\omega_i]\mathbb{E}_{\hat{Q}_1}[\tilde{\lambda}_{T_i}(f)^2] - \log\cosh\left(\frac{\underline{\lambda}_{T_i}(f)}{2}\right) \\
&\quad - \int_{t=0}^{T}\int_{0}^{+\infty}[\log\Lambda(t,\bar{\omega})]\Lambda(t,\bar{\omega})d\bar{\omega}dt + \int_{t=0}^{T}\int_{0}^{+\infty}\Lambda(t,\bar{\omega})d\bar{\omega}dt \\
&= \sum_{i\in[N]} -\mathbb{E}_{\hat{Q}_2}[\log p_{PG}(\omega_i, 1, 0)] + \frac{1}{2}\mathbb{E}_{\hat{Q}_2}[\omega_i]\mathbb{E}_{\hat{Q}_1}[\tilde{\lambda}_{T_i}(f)^2] - \log\cosh\left(\frac{\underline{\lambda}_{T_i}(f)}{2}\right) \\
&\quad - \int_{t=0}^{T}\int_{0}^{+\infty}\left[\log\theta - \frac{1}{2}\mathbb{E}_{\hat{Q}_1}[\tilde{\lambda}_{T_i}(f)] - \log 2 - \log\cosh\left(\frac{\underline{\lambda}_{T_i}(f)}{2}\right) - \frac{1}{2}\mathbb{E}_{\hat{Q}_1}[\tilde{\lambda}_{T_i}(f)^2]\bar{\omega}\right. \\
&\quad \left. + \log\cosh\left(\frac{1}{2}\underline{\lambda}_{T_i}(f)\right) + \log p_{PG}(\bar{\omega}; 1, 0) - 1\right]\Lambda(t)p_{PG}(\bar{\omega}; 1, \underline{\lambda}_{T_i}(f))dtd\bar{\omega} \\
&= \sum_{i\in[N]} -\mathbb{E}_{\hat{Q}_2}[\log p_{PG}(\omega_i, 1, 0)] + \frac{1}{2}\mathbb{E}_{\hat{Q}_2}[\omega_i^k]\mathbb{E}_{\hat{Q}_1}[\tilde{\lambda}_{T_i}(f)^2] - \log\cosh\left(\frac{\underline{\lambda}_{T_i}(f)}{2}\right) \\
&\quad - \int_{t=0}^{T}\left[\log\theta - \frac{1}{2}\mathbb{E}_{\hat{Q}_1}[\tilde{\lambda}_{T_i}(f)] - \log 2 - \frac{1}{2}\mathbb{E}_{\hat{Q}_1}[\tilde{\lambda}_{T_i}(f)^2]\mathbb{E}_{\hat{Q}_2}[\bar{\omega}] - 1\right]\Gamma(t)dt \\
&\quad - \int_{t=0}^{T}\int_{0}^{+\infty}\log p_{PG}(\omega; 1, 0)\Gamma(t)p_{PG}(\omega; 1, \underline{\lambda}_{T_i}(f))d\omega dt.
\end{aligned}$$

with $\Gamma(t) = \theta\int_0^\infty \Lambda(t,\bar{\omega})d\bar{\omega} = \frac{e^{-\frac{1}{2}\mathbb{E}_{\hat{Q}_1}[\tilde{\lambda}_{T_i}(f)]}}{2\cosh\frac{\underline{\lambda}_{T_i}(f)}{2}}$. Moreover, we have

$$\begin{aligned}
&\mathbb{E}_{\hat{Q}_2}\left[\mathbb{E}_{\hat{Q}_1}\left[\log p(f, \omega, \bar{Z}, N)\right]\right] = \sum_{i\in[N]}\left\{\log\theta + \mathbb{E}_{\hat{Q}_2}\left[\mathbb{E}_{\hat{Q}_1}\left[g(\omega_i, \tilde{\lambda}_{T_i}(f))\right] + \log p_{PG}(\omega_i; 1, 0)\right]\right\} \\
&+ \log\theta + \mathbb{E}_{\hat{Q}_2}\left[\mathbb{E}_{\hat{Q}_1}\left[g(\bar{\omega}_t, -\tilde{\lambda}_{T_i}(f))\right] + \log p_{PG}(\bar{\omega}_t; 1, 0)\right] + \mathbb{E}_{\hat{Q}_1}\left[\log\mathcal{N}(f; \mu, \Sigma)\right] \\
&= \sum_{i\in[N]} \log\theta - \log 2 - \frac{1}{2}\mathbb{E}_{\hat{Q}_1}\left[\tilde{\lambda}_{T_i}(f)^2\right]\mathbb{E}_{\hat{Q}_2}\left[\omega_i\right] + \frac{1}{2}\mathbb{E}_{\hat{Q}_1}\left[\tilde{\lambda}_{T_i}(f)\right] + \mathbb{E}_{\hat{Q}_2}\left[\log p_{PG}(\omega_i; 1, 0)\right] \\
&+ \int_0^T\int_0^{+\infty}\left[\log\theta - \log 2 - \frac{1}{2}\mathbb{E}_{\hat{Q}_1}\left[\tilde{\lambda}_{T_i}(f)^2\right]\bar{\omega} - \frac{1}{2}\mathbb{E}_{\hat{Q}_1}\left[\tilde{\lambda}_{T_i}(f)\right] + \log p_{PG}(\bar{\omega}; 1, 0)\right]\Gamma(t)p_{PG}(\omega; 1, \underline{\lambda}_{T_i}(f))d\omega dt \\
&+ \mathbb{E}_{\hat{Q}_1}\left[\log\mathcal{N}(f; \mu, \Sigma)\right] - \theta T \\
&= \sum_{i\in[N]} \log\theta - \log 2 - \frac{1}{2}\mathbb{E}_{\hat{Q}_1}\left[\tilde{\lambda}_{T_i}(f)^2\right]\mathbb{E}_{\hat{Q}_2}\left[\omega_i\right] + \frac{1}{2}\mathbb{E}_{\hat{Q}_1}\left[\tilde{\lambda}_{T_i}(f)\right] + \mathbb{E}_{\hat{Q}_2}\left[\log p_{PG}(\omega_i; 1, 0)\right] \\
&+ \int_0^T\left[\log\theta - \log 2 - \frac{1}{2}\mathbb{E}_{\hat{Q}_1}\left[\tilde{\lambda}_{T_i}(f)^2\right]\mathbb{E}_{\hat{Q}_2}\left[\bar{\omega}\right] - \frac{1}{2}\mathbb{E}_{\hat{Q}_1}\left[\tilde{\lambda}_{T_i}(f)\right]\right]\Gamma(t)dt \\
&+ \int_0^T\int_0^{+\infty}\log p_{PG}(\bar{\omega}; 1, 0)\Gamma(t)p_{PG}(\bar{\omega}; 1, \underline{\lambda}_{T_i}(f))d\bar{\omega}dt + \mathbb{E}_{\hat{Q}_1}\left[\log\mathcal{N}(f; \mu, \Sigma)\right] - \theta T.
\end{aligned}$$

Therefore, with $c > 0$ a constant that does not depend on the size of the model, with zero mean prior $\mu = 0$,

$$ELBO(\hat{Q}) = \frac{|m|}{2} + \frac{1}{2}\log|2\pi\tilde{\Sigma}| - \frac{1}{2}tr(\Sigma^{-1}\tilde{\Sigma}) - \frac{1}{2}\tilde{\mu}^T\Sigma^{-1}\tilde{\mu} - \frac{1}{2}\log|2\pi\Sigma| + \sum_{i\in[N]} \log\theta - \log 2 + \frac{\mathbb{E}_{\hat{Q}_1}\left[\tilde{\lambda}_{T_i}(f)\right]}{2} - \log\cosh\left(\frac{\tilde{\lambda}_{T_i}(f)}{2}\right) + \int_{t=0}^{T}\int_{0}^{+\infty} \Lambda(t,\bar{\omega})d\bar{\omega}dt - \theta T.$$

## Appendix D. Proofs

In this section, we provide the proof of our main theoretical result, namely Theorem 7. We first recall a set of useful lemmas from Sulem et al. (2024).

### D.1 Technical Lemmas

In the first lemma, we recall the definition of excursions from Sulem et al. (2024), for stationary nonlinear Hawkes processes verifying conditions (C1) or (C2). Then, Lemma 17, corresponding to Lemma A.1 in Sulem et al. (2024), provides a control on the main event $\tilde{\Omega}_T$ considered in the proof of Theorem 7. Finally, Lemma 18 (Lemma A.4 in Sulem et al. (2024)) is a technical lemma for proving posterior concentration in Hawkes processes.

We also introduce the following notation. For any excursion index $j \in [J_T - 1]$, we denote $(U_j^{(1)}, U_j^{(2)})$ the times of the first two events after the $j$-th renewal time $\tau_j$, and $\xi_j := U_j^{(2)}$ if $U_j^{(2)} \in [\tau_j, \tau_{j+1})$ and $\xi_j := \tau_{j+1}$ otherwise.

**Lemma 16 (Lemma 5.1 in Sulem et al. (2024))** *Let $N$ be a Hawkes process with monotone non-decreasing and Lipschitz link functions $\phi = (\phi_k)_k$ and parameter $f = (\nu, h)$ such that $(\phi, f)$ verify **(C1)** or **(C2)**. Then the point process measure $X_t(.)$ defined as*

$$X_t(.) = N|_{(t-A,t]}, \tag{53}$$

*is a strong Markov process with positive recurrent state $\emptyset$. Let $\{\tau_j\}_{j\geqslant 0}$ be the sequence of random times defined as*

$$\tau_j = \begin{cases} 0 & \text{if } j = 0; \\ \inf\left\{t > \tau_{j-1};\ X_{t^-} \neq \emptyset,\ X_t = \emptyset\right\} = \inf\left\{t > \tau_{j-1};\ N|_{[t-A,t)} \neq \emptyset,\ N|_{(t-A,t]} = \emptyset\right\} & \text{if } j \geqslant 1. \end{cases}$$

*Then, $\{\tau_j\}_{j\geqslant 0}$ are stopping times for the process $N$. For $T > 0$, we also define*

$$J_T = \max\{j \geqslant 0;\ \tau_j \leqslant T\}. \tag{54}$$

*The intervals $\{[\tau_j, \tau_{j+1})\}_{j=0}^{J_T-1} \cup [\tau_{J_T}, T]$ form a partition of $[0, T]$. The point process measures $(N|_{[\tau_j,\tau_{j+1})})_{1\leqslant j\leqslant J_T-1}$ are i.i.d. and independent of $N|_{[0,\tau_1)}$ and $N|_{[\tau_{J_T},T]}$; they are called* excursions *and the stopping times $\{\tau_j\}_{j\geqslant 1}$ are called* regenerative *or* renewal *times.*

**Lemma 17 (Lemma A.1 in Sulem et al. (2024))** *Let $Q > 0$. We consider $\tilde{\Omega}_T$ defined in Section D.2. For any $\beta > 0$, we can choose $C_\beta$ and $c_\beta$ in the definition of $\tilde{\Omega}_T$ such that $\mathbb{P}_0[\tilde{\Omega}_T^c] \leqslant T^{-\beta}$. Moreover, for any $1 \leqslant q \leqslant Q$,*

$$\mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T^c} \max_l \sup_{t\in[0,T]} \left(N^l[t-A,t)\right)^q\right] \leqslant 2T^{-\beta/2}.$$

**Lemma 18 (Lemma A.4 in Sulem et al. (2024))** *For any $f \in \mathcal{F}_T$ and $l \in [K]$, let*

$$Z_{1l} = \int_{\tau_1}^{\xi_1} |\lambda_t^l(f) - \lambda_t^l(f_0)|dt.$$

*Under the assumptions of Theorem 7, for $M_T \to \infty$ such that $M_T > M\sqrt{\kappa_T}$ with $M > 0$ and for any $f \in \mathcal{F}_T$ such that $\|r - r_0\|_1 \leqslant \max(\|r_0\|_1, \tilde{C})$ with $\tilde{C} > 0$, there exists $l \in [K]$ such that on $\tilde{\Omega}_T$,*

$$\mathbb{E}_f[Z_{1l}] \geqslant C(f_0)\Big(\left\|r_f - r_0\right\|_1 + \|h - h_0\|_1\Big),$$

*with $C(f_0) > 0$ a constant that depends only on $f_0$ and $(\phi_k)_k$.*

### D.2 Proof of Theorem 7

We start by defining the stochastic distance $\tilde{d}_{1T}$, stochastic and $L_1$-neighborhoods around $f_0$ as

$$\tilde{d}_{1T}(f,f') = \frac{1}{T}\sum_{k=1}^K \int_0^T \mathbb{1}_{I_2}(t)|\lambda_t^k(f) - \lambda_t^k(f')|dt, \quad I_2 = \bigcup_{j=1}^{J_T-1}[\tau_j, \xi_j] \tag{55}$$

$$B_{d_1}(\varepsilon) = \left\{f \in \mathcal{F};\ \tilde{d}_{1T}(f,f_0) \leqslant \varepsilon\right\},$$

$$B_{L_1}(\varepsilon) = \{f \in \mathcal{F};\ \|f - f_0\|_1 \leqslant \varepsilon\}, \quad \varepsilon > 0,$$

where for each $j \in [J_T]$, $U_j^{(2)}$ is the first event after $U_j^{(1)}$, and $\xi_j := U_j^{(2)}$ if $U_j^{(2)} \in [\tau_j, \tau_{j+1})$ and $\xi_j := \tau_{j+1}$ otherwise.

Let $(\eta_T)_T$ be a positive sequence and $\hat{Q}$ be the variational posterior as defined in (9). For any event $\mathcal{E}_T$, we have

$$\mathbb{E}_0\left[\hat{Q}(B_{d_1}(\eta_T)^c)\right] \leqslant \mathbb{P}_0\left[\mathcal{E}_T^c\right] + \mathbb{E}_0\left[\hat{Q}(B_{d_1}(\eta_T)^c)\mathbb{1}_{\mathcal{E}_T}\right]. \tag{56}$$

We first construct an event $\mathcal{E}_T$ such that $\mathbb{P}_0\left[\mathcal{E}_T^c\right] = o(1)$.

Recall that we consider a general Hawkes model with known link functions $(\phi_k)_k$. Let $r_0 = (r_1^0, \dots, r_K^0)$ with $r_k^0 = \phi_k(\nu_k^0)$. With $C_\beta, c_\beta > 0$, we first define a composite $\tilde{\Omega}_T \in \mathcal{G}_T$ as

$$\tilde{\Omega}_T = \Omega_N \cap \Omega_J \cap \Omega_U,$$

$$\Omega_N = \left\{\max_{k\in[K]} \sup_{t\in[0,T]} N^k[t-A,t) \leqslant C_\beta \log T\right\} \cap \left\{\sum_{k=1}^K \left|\frac{N^k[-A,T]}{T} - \mu_k^0\right| \leqslant \chi_T\right\},$$

$$\Omega_J = \{J_T \in \mathcal{J}_T\}, \quad \Omega_U = \left\{\sum_{j=1}^{J_T-1}(U_j^{(1)} - \tau_j) \geqslant \frac{T}{\mathbb{E}_0[\Delta\tau_1]\|r_0\|_1}\left(1 - 2c_\beta\sqrt{\frac{\log T}{T}}\right)\right\},$$

$$\mathcal{J}_T = \left\{J \in \mathbb{N};\ \left|\frac{J-1}{T} - \frac{1}{\mathbb{E}_0[\Delta\tau_1]}\right| \leqslant c_\beta\sqrt{\frac{\log T}{T}}\right\},$$

with $J_T$ the number of excursions as defined in (54), $\mu_k^0 := \mathbb{E}_0\left[\lambda_t^k(f_0)\right], \forall k, \chi_T = \chi_0\sqrt{\frac{\log T}{T}}$, $\chi_0 > 0$ and $\{U_j^{(1)}\}_{j=1,\dots,J_T-1}$ denoting the first events of each excursion (see Lemma 16 for a precise definition). Intuitively, $\tilde{\Omega}_T$ is an event where the numbers of events in $[0,T]$ and in any interval of length $A$, the number of excursions, and the times of the first event in each excursion are controlled and from Lemma A.1 in Sulem et al. (2024), see also Lemma 17, $\mathbb{P}_0(\tilde{\Omega}_T) = 1 + o(1)$. Then, we define $\Omega'_T \in \mathcal{G}_T$ as

$$\Omega'_T = \left\{\int e^{L_T(f)-L_T(f_0)}d\widetilde{\Pi}(f) > e^{-C_1 T\varepsilon_T^2}\right\}, \quad \widetilde{\Pi}(B) = \frac{\Pi(B\cap B_T)}{\Pi(B_T)}, \quad B, B_T \subset \mathcal{F},$$

with $C_1 > 0$ and $\varepsilon_T, M_T$ positive sequences such that $T\varepsilon_T^2 \to \infty$ and $M_T \to \infty$. From Lemma 17, we have that $\mathbb{P}_0\left[\tilde{\Omega}_T^c\right] = o(1)$. Thus, with $D_T$ defined in (4), $\Omega_T = \tilde{\Omega}_T \cap \Omega'_T$, $B_T = B_\infty(\epsilon_T)$, and $\varepsilon_T = \sqrt{\kappa_T}\epsilon_T$, we obtain that

$$\begin{aligned}
\mathbb{P}_0\left[\mathcal{E}_T^c\right] &\leqslant \mathbb{P}_0\left[\tilde{\Omega}_T^c\right] + \mathbb{P}_0\left[\mathcal{E}_T'^c \cap \tilde{\Omega}_T\right] \\
&= o(1) + \mathbb{P}_0\left[\left\{\int_{B_T} e^{L_T(f)-L_T(f_0)}d\Pi(f) \leqslant \Pi(B_T)e^{-C_1 T\varepsilon_T^2}\right\} \cap \tilde{\Omega}_T\right] \\
&\leqslant o(1) + \mathbb{P}_0\left[\left\{D_T \leqslant \Pi(B_T)e^{-C_1 T\varepsilon_T^2}\right\} \cap \tilde{\Omega}_T\right] = o(1),
\end{aligned}$$

with $C_1 > 1$, using (A0), i.e., $\Pi(B_T) \geqslant e^{-c_1 T\varepsilon_T^2}$, and the following intermediate result from the proof of Theorem 3.2 in Sulem et al. (2024)

$$\mathbb{P}_0\left[\left\{D_T \leqslant \Pi(B_\infty(\epsilon_T))e^{-\kappa_T T\varepsilon_T^2}\right\} \cap \tilde{\Omega}_T\right] = o(1).$$

We now focus on the second term on the RHS of (56) and bound it using the following technical lemma, which is an adaptation of Theorem 5 of Ray and Szabó (2021) and Lemma 13 in Nieman et al. (2022).

**Lemma 19** *Let $B \subset \mathcal{F}$, $\Omega \in \mathcal{G}_T$, and $Q$ be a distribution on $\mathcal{F}$. If there exist $C, u_T > 0$ such that*

$$\mathbb{E}_0\left[\Pi(B|N)\mathbb{1}_\Omega\right] \leqslant Ce^{-u_T}, \tag{57}$$

*then, we have that*

$$\mathbb{E}_0\left[Q(B)\mathbb{1}_\Omega\right] \leqslant \frac{2}{u_T}\left(\mathbb{E}_0\left[KL(Q||\Pi(.|N))\mathbb{1}_\Omega\right] + Ce^{-u_T/2}\right).$$

**Proof** We follow the proof of Ray and Szabó (2021) and use the fact that, for any $g : \mathcal{F} \to \mathbb{R}$ such that $\int_{\mathcal{F}} e^{g(f)}d\Pi(f|N) < +\infty$, it holds true that

$$\int_{\mathcal{F}} g(f)dQ(f) \leqslant KL(Q||\Pi(.|N)) + \log\int_{\mathcal{F}} e^{g(f)}d\Pi(f|N). \tag{58}$$

Applying the latter inequality with $g = \frac{1}{2}u_T\mathbb{1}_B$, we obtain

$$\begin{aligned}
\frac{1}{2}u_T Q(B) &\leqslant KL(Q||\Pi(.|N)) + \log(1 + e^{\frac{1}{2}u_T}\Pi(B|N)) \\
&\leqslant KL(Q||\Pi(.|N)) + e^{\frac{1}{2}u_T}\Pi(B|N).
\end{aligned}$$

Then, multiplying both sides of the previous inequality by $\mathbb{1}_\Omega$ and taking expectation w.r.t. to $\mathbb{P}_0$, using (57), we finally obtain

$$\frac{1}{2}u_T\mathbb{E}_0\left[Q(B)\mathbb{1}_{\Omega_T}\right] \leqslant \mathbb{E}_0\left[KL(Q||\Pi(.|N))\mathbb{1}_\Omega\right] + Ce^{-\frac{1}{2}u_T}.$$

■

We thus apply Lemma 19 with $B = B_{d_1}(\eta_T)^c$, $\Omega = \Omega_T$, $\eta_T = M'_T\varepsilon_T$, $Q = \hat{Q}$, and $u_T = M_T T\varepsilon_T^2$ with $M'_T \to \infty$. We first check that (57) holds, i.e., we show that there exist $C, M_T, M'_T > 0$ such that

$$\mathbb{E}_0\left[\mathbb{1}_{\Omega_T}\Pi[\tilde{d}_{1T}(f, f_0) > M'_T\varepsilon_T|N]\right] \leqslant C\exp(-M_T T\varepsilon_T^2). \tag{59}$$

For any test $\phi$, we have the following decomposition

$$\mathbb{E}_0\left[\mathbb{1}_{\Omega_T}\Pi[\tilde{d}_{1T}(f, f_0) > M'_T\varepsilon_T|N]\right] \leqslant \underbrace{\mathbb{E}_0\left[\phi\mathbb{1}_{\Omega_T}\right]}_{(I)} + \underbrace{\mathbb{E}_0\left[(1-\phi)\mathbb{1}_{\Omega_T}\Pi[B_{d_1}(M'_T\varepsilon_T)^c|N]\right]}_{(II)}.$$

Note that we have

$$\begin{aligned}(II) = \mathbb{E}_0\left[(1-\phi)\mathbb{1}_{\Omega_T}\Pi[B_{d_1}(M'_T\varepsilon_T)^c|N]\right] &= \mathbb{E}_0\left[\int_{B_{d_1}(M'_T\varepsilon_T)^c}\mathbb{1}_{\mathcal{E}_T}(1-\phi)\frac{e^{L_T(f)-L_T(f_0)}}{D_T}d\Pi(f)\right] \\ &\leqslant \frac{e^{C_1 T\varepsilon_T^2}}{\Pi(B_T)}\mathbb{E}_0\left[\sup_{f\in\mathcal{F}_T}\mathbb{E}_f\left[\mathbb{1}_{B_{d_1}(M'_T\varepsilon_T)^c}\mathbb{1}_{\Omega_T}(1-\phi)|\mathcal{G}_0\right]\right] \\ &\quad + \frac{e^{C_1 T\varepsilon_T^2}}{\Pi(B_T)}\Pi(\mathcal{F}_T^c)\end{aligned} \tag{60}$$

since on $\Omega_T, D_T \geqslant \Pi(B_T)e^{-C_1 T\varepsilon_T^2}$. Using the proof of Theorem 5.5 in Sulem et al. (2024), we can directly obtain that for $T$ large enough, there exist $x_1, M, M' > 0$ such that

$$\begin{aligned}(I) &\leqslant 2(2K+1)e^{-x_1 {M'_T}^2 T\varepsilon_T^2} \\ (II) &\leqslant 2(2K+1)e^{-x_1 {M'_T}^2 T\varepsilon_T^2/2},\end{aligned}$$

which implies that

$$\mathbb{E}_0\left[\mathbb{1}_{\Omega_T}\Pi[\tilde{d}_{1T}(f, f_0) > M'_T\varepsilon_T|N]\right] \leqslant 4(2K+1)e^{-x_1 M_T'^2 T\varepsilon_T^2/2},$$

and (59) with $M_T = x_1 M_T'^2/2$ and $C = 4(2K+1)$. Applying Lemma 19 thus leads to

$$\mathbb{E}_0\left[\hat{Q}(B_{d_1}(\eta_T)^c)\mathbb{1}_{\Omega_T}\right] \leqslant 2\mathbb{E}_0\left[\frac{KL(\hat{Q}||\Pi(.|N))] + Ce^{-M_T T\varepsilon_T^2/2}}{M_T T\varepsilon_T^2}\right] \leqslant 2Ce^{-M_T T\varepsilon_T^2/2} + 2\frac{\mathbb{E}_0[KL(\hat{Q}||\Pi(.|N))]}{M_T T\varepsilon_T^2}.$$

Moreover, from (A2) and the remark following Theorem 7, it holds that $\mathbb{E}_0[KL(\hat{Q}||\Pi(.|N))] = O(T\varepsilon_T^2)$, therefore we obtain the following intermediate result

$$\mathbb{E}_0\left[\hat{Q}(B_{d_1}(\eta_T)^c)\right] = o(1).$$

Now, with $M_T > M'_T$, we note that

$$\mathbb{E}_0\left[\hat{Q}(\|f-f_0\|_1 > M_T\varepsilon_T)\right] = \mathbb{E}_0\left[\hat{Q}(\tilde{d}_{1T}(f,f_0) > M'_T\varepsilon_T)\right] + \mathbb{E}_0\left[\hat{Q}(\|f-f_0\|_1 > M_T\varepsilon_T, \tilde{d}_{1T}(f,f_0) < M'_T\varepsilon_T)\mathbb{1}_{\Omega_T}\right] + \mathbb{P}_0[\Omega_T^c].$$

Therefore, it remains to show that

$$\mathbb{E}_0\left[\hat{Q}(\|f-f_0\|_1 > M_T\varepsilon_T, \tilde{d}_{1T}(f,f_0) < M'_T\varepsilon_T)\mathbb{1}_{\Omega_T}\right] = \mathbb{E}_0\left[\hat{Q}(B_{L_1}(M_T\epsilon_T)^c \cap B_{d_1}(M'_T\varepsilon_T))\mathbb{1}_{\Omega_T}\right] = o(1).$$

For this, we apply again Lemma 19 with $B = B_{L_1}(M_T\varepsilon_T)^c \cap B_{d_1}(M'_T\varepsilon_T)$ and $u_T = TM_T^2\varepsilon_T^2$. We have

$$\mathbb{E}_0\left[\mathbb{1}_{\Omega_T}\Pi(B_{L_1}(M_T\varepsilon_T)^c \cap B_{d_1}(M'_T\varepsilon_T)|N)\right] \leqslant \frac{e^{C_1T\varepsilon_T^2}}{\Pi(B_T)}\mathbb{E}_0\left[\int_{B_{L_1}(M_T\varepsilon_T)^c}\mathbb{E}_f\left[\mathbb{1}_{\Omega_T}\mathbb{1}_{B_{d_1}(M'_T\varepsilon_T)}|\mathcal{G}_0\right]d\pi(f)\right].$$

Let $f \in B_{L_1}(M_T\varepsilon_T)^c \cap B_{d_1}(M'_T\varepsilon_T)$. For any $j \in [J_T-1]$ and $l \in [K]$, let

$$Z_{jl} = \int_{\tau_j}^{\xi_j} |\lambda_t^l(f) - \lambda_t^l(f_0)|dt, \quad j \in [J_T-1], \quad l \in [K]. \tag{61}$$

Using Lemma 18 and the integer $l$ introduced in this lemma, for any $f \in B_{L_1}(M_T\epsilon_T)^c$, we have

$$\begin{aligned}
\mathbb{E}_f\left[\mathbb{1}_{\Omega_T}\mathbb{1}_{B_{d_1}(M'_T\varepsilon_T)}|\mathcal{G}_0\right] &\leqslant \mathbb{P}_f\left[\sum_{j=1}^{J_T-1} Z_{jl} \leqslant TM'_T\varepsilon_T|\mathcal{G}_0\right] \\
&\leqslant \sum_{J\in\mathcal{J}_T}\mathbb{P}_f\left[\sum_{j=1}^{J-1} Z_{jl} - \mathbb{E}_f\left[Z_{jl}\right] \leqslant TM'_T\epsilon_T - \frac{T}{2\mathbb{E}_0\left[\Delta\tau_1\right]}C(f_0)M_T\epsilon_T|\mathcal{G}_0\right] \\
&\leqslant \sum_{J\in\mathcal{J}_T}\mathbb{P}_f\left[\sum_{j=1}^{J-1} Z_{jl} - \mathbb{E}_f\left[Z_{jl}\right] \leqslant -\frac{T}{4\mathbb{E}_0\left[\Delta\tau_1\right]}C(f_0)M_T\varepsilon_T|\mathcal{G}_0\right],
\end{aligned}$$

for any $M_T \geqslant 4\mathbb{E}_0\left[\Delta\tau_1\right]M'_T$. Similarly to the proof of Theorem 3.2 in Sulem et al. (2024)), we apply Bernstein's inequality for each $J \in \mathcal{J}_T$ and obtain that

$$\mathbb{E}_f\left[\mathbb{1}_{\Omega_T}\mathbb{1}_{B_{d_1}(M'_T\varepsilon_T)}|\mathcal{G}_0\right] \leqslant \exp\{-c(f_0)'T\}, \quad \forall f \in B_{L_1}(M_T\varepsilon_T)^c,$$

for $c(f_0)'$ a positive constant. Therefore, we can conclude that

$$\mathbb{E}_0\left[\hat{Q}\left(B_{L_1}(M_T\varepsilon_T)^c \cap B_{d_1}(M'_T\varepsilon_T)\right)\mathbb{1}_{\Omega_T}\right] \leqslant \frac{2}{M_TT\varepsilon_T^2}\mathbb{E}_0\left[KL(\hat{Q}\|\Pi(.|N))\right] + o(1) = o(1),$$

since $\mathbb{E}_0\left[KL(\hat{Q}\|\Pi(.|N))\right] = O(T\varepsilon_T^2)$ by assumption (A2). This leads to our final conclusion

$$\mathbb{E}_0\left[\hat{Q}\left(\|f-f_0\|_1 > M_T\varepsilon_T\right)\right] = o(1).$$

We now prove the second statement of Theorem 7. We recall our notation $S = (S_{lk})_{l,k}, S^0 = (S^0_{lk})_{l,k}$ with $S_{lk} = \|h_{lk}\|_1$ and $S^0_{lk} = \|h^0_{lk}\|_1$ and define

$$\|S - S^0\|_1 := \sum_{l,k}|S_{lk} - S^0_{lk}|.$$

Note that $S = S(f)$ is a function of the parameter $f$ but we omit this dependence for the sake of simplicity of our notation. We now assume that there exists $\bar{S}_0 > 0$ such that either (i) $\|S - S^0\|_1 \leqslant \bar{S}_0$ because $\mathcal{F}$ is a bounded space, i.e., $\|h_{lk}\|_\infty \leqslant C$, which implies $S_{lk} = \|h_{lk}\|_1 \leqslant A\|h_{lk}\|_\infty = AC$, or either (ii) $\|S - S^0\|_1 \leqslant \bar{S}_0 + K\|\nu - \nu^0\|_1 A$ because $(\phi_k)$ satisfy Assumption **(C1)** on $\mathcal{F}$ which implies that

$$\|h_{lk}\|_1 = \|h^+_{lk}\|_1 + \|h^-_{lk}\|_1 < 1 + \|h^-_{lk}\|_1 \leqslant 1 + \|h^-_{lk}\|_\infty A \leqslant 1 + |\nu_k|A \leqslant 1 + |\nu^0_k|A + |\nu_k - \nu^0_k|A,$$

for any $f \in B_{d_1}(M'_T \varepsilon_T)$. We also note that in the case where $\mathcal{F}$ is unbounded, the rest of this proof also holds if $S$ is replaced by a clipped version $S^C = (S^C_{lk})_{l,k}$ where $S^C_{lk} = \|h^C_{lk}\|_1$ with $C > \max_{l,k} \|h^0_{lk}\|_\infty$ and $h^C_{lk} = ((-2C) \vee h_{lk}) \wedge 2C$.

Then for case (i), using inequality (58) with $g(f) = z_T T \frac{\|S - S_0\|_1}{\bar{S}_0}$ with $z_T > 0$ defined later, we have, for any distribution $Q$,

$$\mathbb{E}_0[Q(\|S - S^0\|_1)\mathbb{1}_{\Omega_T}] \leqslant \frac{\bar{S}_0}{z_T T}\left(\mathbb{E}_0[\mathbb{1}_{\Omega_T} KL(Q\|\Pi(\cdot|N)] + \mathbb{E}_0[\mathbb{1}_{\Omega_T} \log \int e^{z_T T \frac{\|S - S^0\|_1}{\bar{S}_0}} d\Pi(S|N)]\right). \tag{62}$$

Now since

$$\begin{aligned}
\|S - S^0\|_1 &= \sum_{l,k} \Big| \int_0^A (|h_{lk}(x)| - |h^0_{lk}(x)|)dx \Big| \\
&\leqslant \sum_{l,k} \int_0^A \| |h_{lk}(x)| - |h^0_{lk}(x)| \| dx \\
&\leqslant \sum_{l,k} \int_0^A |h_{lk}(x) - h^0_{lk}(x)|dx = \sum_{l,k} \|h_{lk} - h^0_{lk}\|_1 \leqslant \|f - f_0\|_1,
\end{aligned}$$

this implies that for any $\epsilon > 0$,

$$\Pi(\|S - S^0\|_1 > \epsilon|N) \leqslant \Pi(\|f - f_0\|_1 > \epsilon|N). \tag{63}$$

We now show that if the following holds true: there exists $M_0, c, C, \epsilon_0 > 0$ such that for any $M_0\varepsilon_T \leqslant \epsilon \leqslant \epsilon_0$, it holds that

$$\mathbb{E}_0\left[\mathbb{1}_{\Omega_T} \Pi\left(\|f - f_0\|_1 > \epsilon | N\right)\right] \leqslant C e^{-cT\epsilon^2}, \tag{64}$$

then we obtain our result.

Let $\epsilon \in [M_0\varepsilon_T, \epsilon_0]$ and $M > 1$. We first decompose the RHS of (62) as

$$\begin{aligned}
\int e^{z_T T \frac{\|S - S^0\|_1}{\bar{S}_0}} d\Pi(S|N) &= \int e^{z_T T \frac{\|S - S^0\|_1}{\bar{S}_0}} \mathbb{1}_{\bar{S}_0 > \|S - S^0\|_1 \geqslant \epsilon_0} d\Pi(S|N) + \int e^{z_T T \frac{\|S - S^0\|_1}{\bar{S}_0}} \mathbb{1}_{\epsilon_0 > \|S - S^0\|_1 > M\epsilon} d\Pi(S|N) \\
&\quad + \int e^{z_T T \frac{\|S - S^0\|_1}{\bar{S}_0}} \mathbb{1}_{\|S - S^0\|_1 < M\epsilon} d\Pi(S|N) \\
&=: \mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3.
\end{aligned}$$

For the first term $\mathcal{T}_1$, from (64), we have:

$$\mathbb{E}_0\left[\mathbb{1}_{\Omega_T} \mathcal{T}_1\right] \leqslant e^{z_T T} \mathbb{E}_0\left[\mathbb{1}_{\Omega_T} \Pi(\|S - S^0\|_1 \geqslant \epsilon_0|N)\right] \leqslant C e^{z_T T - cT\epsilon_0^2} \leqslant C e^{-cT\epsilon_0^2/2},$$

if $z_T \leqslant \frac{cT\epsilon_0^2}{2}$. For the second term $\mathcal{T}_2$ we further decompose the integral as

$$\mathbb{E}_0\left[\mathbb{1}_{\Omega_T}\mathcal{T}_2\right] \leqslant \sum_{j=M}^{\frac{\epsilon_0}{\epsilon}} \mathbb{E}_0\left[\mathbb{1}_{\Omega_T}\int e^{z_T T\frac{\|S-S^0\|_1}{\bar{S}_0}}\mathbb{1}_{(j+1)\epsilon>\|S-S^0\|_1\geqslant j\epsilon}d\Pi(S|N)\right]$$
$$\leqslant \sum_{j=M}^{\frac{\epsilon_0}{\epsilon}} e^{z_T T\frac{(j+1)\epsilon}{\bar{S}_0}-cTj^2\epsilon^2} \leqslant \sum_{j=M}^{\frac{\epsilon_0}{\epsilon}} e^{-cTj^2\epsilon^2/2} \leqslant 2e^{-cTM^2\epsilon^2/2},$$

where the third inequality holds if $z_T\frac{(j+1)\epsilon}{\bar{S}_0} \leqslant cTj^2\epsilon^2/2$. For the latter to hold it is sufficient that

$$z_T \leqslant \frac{c\bar{S}_0 j\epsilon}{4}, \qquad \forall j \geqslant M.$$

For the third term $\mathcal{T}_3$, we have

$$\mathbb{E}_0\left[\mathbb{1}_{\Omega_T}\mathcal{T}_3\right] \leqslant e^{z_T T\frac{M\epsilon}{\bar{S}_0}}.$$

Therefore, from Jensen's inequality, we obtain that

$$\mathbb{E}_0\left[\mathbb{1}_{\Omega_T}\log\int e^{z_T T\frac{\|S-S^0\|_1}{\bar{S}_0}}d\Pi(S|N)\right] \leqslant \log\left(Ce^{-cT\epsilon_0^2/2}+2e^{-cTM^2\epsilon^2/2}+e^{z_T T\frac{M\epsilon}{\bar{S}_0}}\right)$$
$$\leqslant \log\left(2e^{z_T T\frac{M\epsilon}{\bar{S}_0}}\right) = \log(2)+z_T T\frac{M\epsilon}{\bar{S}_0},$$

for $T$ sufficiently large, and this holds in particular for $\epsilon = M_0\varepsilon_T$ and $z_T = \frac{c\bar{S}_0 MM_0\varepsilon_T}{4}$. Reporting into (62) and applying it to the variational posterior $\hat{Q}$, we obtain

$$\mathbb{E}_0[\hat{Q}(\|S-S^0\|_1)\mathbb{1}_{\Omega_T}] \leqslant \frac{\bar{S}_0}{z_T T}KL(\hat{Q}\|\Pi(\cdot|N)+2MM_0\varepsilon_T,$$

and since by definition $KL(\hat{Q}\|\Pi(\cdot|N) = \inf_Q KL(Q\|\Pi(\cdot|N) \leqslant \tilde{C}T\varepsilon_T^2$ where $\tilde{C}>0$ using **(A2')**, we finally have

$$\mathbb{E}_0[\hat{Q}(\|S-S^0\|_1)\mathbb{1}_{\Omega_T}] \leqslant \frac{4\tilde{C}}{MM_0}\bar{S}_0\varepsilon_T+2MM_0\varepsilon_T \leqslant M'\varepsilon_T,$$

for some $M'>0$ large enough.

Now, with $\hat{S} = \int S d\hat{Q}(S) = \mathbb{E}_{\hat{Q}}[S]$, we have

$$\mathbb{E}_0[\|\hat{S}-S^0\|_1\mathbb{1}_{\Omega_T}] = \mathbb{E}_0[\|\int(S-S^0)d\hat{Q}(S)\|_1\mathbb{1}_{\Omega_T}] \leqslant \mathbb{E}_0[\int\|S-S^0\|_1 d\hat{Q}(S)\mathbb{1}_{\Omega_T}] = \mathbb{E}_0[\hat{Q}(\|S-S^0\|_1)\mathbb{1}_{\Omega_T}]$$
$$\leqslant M'\varepsilon_T.$$

Thus, for any $M_T\to\infty$, we have

$$\mathbb{P}_0[\Omega_T\cap\{\|\hat{S}-S^0\|_1 > M_T\epsilon_T\}] \leqslant \frac{\mathbb{E}_0[\|\hat{S}-S^0\|_1\mathbb{1}_{\Omega_T}]}{M_T\epsilon_T} \leqslant \frac{M'\epsilon_T}{M_T\epsilon_T} = o(1), \tag{65}$$

and noting that $\|\hat{S} - S^0\|_1 = \sum_{l,k} |\mathbb{E}_{\hat{Q}}(S_{lk}) - S^0_{lk}|$ this implies that

$$\mathbb{P}_0[\sum_{l,k} |\mathbb{E}_{\hat{Q}}(S_{lk}) - S^0_{lk}| > M_T \epsilon_T] \leqslant \mathbb{P}_0[\Omega_T] + o(1) = o(1).$$

Finally, to prove (64), we use the same technique as the proof of Theorem 5.5 in Sulem et al. (2024) and Assumption **(A1)**: $\exists M_0, c_2 > 0$, $\epsilon_0 \in (0,1]$, $\forall \epsilon \in [M_0\varepsilon_T, \epsilon_0]$, $\exists \mathcal{H}_T(\epsilon) \subset \mathcal{H}$, $\zeta_0 > 0$, and $x_0 > 0$ such that

$$\Pi(\mathcal{H}_T^c(\epsilon)) = o(e^{-c_2 T\epsilon^2}) \quad \text{and} \quad \log C\left(\zeta_0\epsilon, \mathcal{H}_T(\epsilon), \|\cdot\|_1\right) \leqslant x_0 T\epsilon^2.$$

We have

$$\begin{aligned}\mathbb{E}_0\left[\mathbb{1}_{\Omega_T}\Pi\left(\|f - f_0\|_1 > \epsilon | N\right)\right] &= \mathbb{E}_0\left[\mathbb{1}_{\Omega_T}\Pi\left(B_{L_1}(\epsilon)^c | N\right)\right] \\ &\leqslant \mathbb{E}_0\left[\mathbb{1}_{\Omega_T}\Pi\left(B_{L_1}(\epsilon)^c \cap B_{d_1}(M_0'\epsilon) | N\right)\right] + \mathbb{E}_0\left[\mathbb{1}_{\Omega_T}\Pi\left(B_{d_1}(M_0'\epsilon)^c | N\right)\right]\end{aligned} \tag{66}$$

For the second term on the RHS of the previous equation, the same computations as the first part of this proof and **(A1)** leads to

$$\mathbb{E}_0\left[\mathbb{1}_{\Omega_T}\Pi\left(B_{d_1}(M_0'\epsilon)^c | N\right)\right] \leqslant C' e^{-c' T\epsilon^2},$$

for some $C', c' > 0$. Then, for the first term, from the previous computation, we also have

$$\mathbb{E}_0\left[\mathbb{1}_{\Omega_T}\Pi\left(B_{L_1}(\epsilon)^c \cap B_{d_1}(M_0'\epsilon) | N\right)\right] \leqslant \frac{e^{C_1 T\epsilon^2/M_0^2}}{\Pi(B_\infty(\epsilon))} \times e^{-c(f_0)' T} + \frac{e^{C_1 T\epsilon^2/M_0^2}}{\Pi(B_\infty(\epsilon))}\Pi(\mathcal{H}_T^c(\epsilon)),$$

using that $\epsilon \geqslant M_0\varepsilon_T$. Moreover we note that $\Pi(B_\infty(\epsilon)) \geqslant \Pi(B_\infty(\varepsilon_T/M_0)) \geqslant e^{-c_1 T\epsilon_T^2} \geqslant e^{-c_1 T\epsilon^2}$, for $T$ large enough. Therefore, using **(A1)** again we can conclude that

$$\mathbb{E}_0\left[\mathbb{1}_{\Omega_T}\Pi\left(B_{L_1}(\epsilon)^c \cap B_{d_1}(M_0'\epsilon) | N\right)\right] \leqslant C'' e^{-c'' T\epsilon^2},$$

for some $C'', c'' > 0$, which from (66) allows to conclude that there exists $C, c > 0$ such that

$$\mathbb{E}_0\left[\mathbb{1}_{\Omega_T}\Pi\left(\|f - f_0\|_1 > \epsilon | N\right)\right] \leqslant C e^{-cT\epsilon^2},$$

for any $\epsilon \in [M_0\varepsilon_T, \epsilon_0]$.

Finally for case (ii) we note that since on $\Omega_T$, there exists $c_0 > 0$ such that $d_{1T}(f, f_0) \geqslant c_0\|\nu - \nu_0\|$, then for any $\epsilon > 0$

$$\mathbb{E}_0[\mathbb{1}_{\Omega_T}\Pi(\|\nu - \nu_0\|_1 > M'\epsilon)] \leqslant \mathbb{E}_0[\mathbb{1}_{\Omega_T}\Pi(B_{d_1}(M'\epsilon)^c)] \leqslant C e^{-cT\epsilon^2 \wedge \epsilon}.$$

Then we can apply (62) with $\bar{S}_0 = 1$ and using that for any $\epsilon > 0$,

$$\Pi(\|S - S^0\|_1 > \epsilon) \leqslant \Pi(KA\|\nu - \nu^0\|_1 > \epsilon),$$

we can write

$$\begin{aligned}\mathbb{E}_0[\mathbb{1}_{\Omega_T} \log \int e^{T\|S - S^0\|_1} d\Pi(f|N)] &\leqslant MT\varepsilon_T + \mathbb{E}_0[\mathbb{1}_{\Omega_T} \log \int_{e^{MT\varepsilon_T}}^\infty \Pi(\|\nu - \nu^0\|_1 > \log s/(TKA))ds] \\ &\leqslant 2MT\varepsilon_T,\end{aligned}$$

and the conclusion follows from the same arguments as case (i).

### D.3 Gaussian Process Prior

In this section, we propose an alternative prior family based on Gaussian processes, a family commonly used for nonparametric estimation of Hawkes processes (see for instance Zhang et al. (2020); Zhou et al. (2020); Malem-Shinitski et al. (2021)). We define a centered Gaussian process distribution with covariance function $k_{GP}$ as the prior distribution $\tilde{\Pi}_{h|\delta}$ on each $h_{lk}$ such that $\delta_{lk} = 1$, $l, k \in [K]$, i.e., for any $n \geqslant 1$ and $x_1, \dots, x_n \in [0, A]$, we have

$$(h_{lk}(x_i))_{i=1,\dots,n} \sim \mathcal{N}\left(0_n, (k_{GP}(x_i, x_j))_{i,j=1,\dots,n}\right).$$

We then verify assumptions **(A0')** and **(A1)** based on the $L_2$-neighborhoods (see comment after Theorem 7), i.e., we check that there exist $\mathcal{H}_T \subset \mathcal{H}$ and $c_1, x_0, \zeta_0 > 0$, such that

$$\Pi(\mathcal{H}_T^c) \leqslant e^{-(\kappa_T + c_1)T\epsilon_T^2}, \quad \log C(\zeta_0 \epsilon_T, \mathcal{H}_T, \|.\|_1) \leqslant x_0 T\epsilon_T^2, \quad \Pi(B_2(\epsilon_T, B)) \geqslant e^{-c_1 T\epsilon_T^2}.$$

We define $\mathcal{H}_T = \mathcal{B}_T^{\otimes K^2}$ with $\mathcal{B}_T \subset L_2([0, A])$. We note that if $\mathcal{B}_T$ verifies

$$\tilde{\Pi}_h(\mathcal{B}_T^c) \leqslant K^{-2} e^{-(\kappa_T + c_1)T\varepsilon_T^2}, \quad \log C(\zeta_0 \epsilon_T, \mathcal{B}_T, \|.\|_1) \leqslant \frac{x_0 T\epsilon_T^2}{K^2}, \quad \tilde{\Pi}_h(\left\|h_{lk} - h_{lk}^0\right\|_2 < \epsilon_T) \geqslant e^{-c_2 T\varepsilon_T^2}, \tag{67}$$

then, for all $\zeta > 0$, there exists $\zeta_2 > 0$ (independent of $T$) such that $\Pi(\mathcal{H}_T^c) \leqslant K^2 \tilde{\Pi}(\mathcal{B}_T^c)$, and

$$\log C(\zeta \epsilon_T, \mathcal{H}_T, \|.\|_1) \leqslant K^2 \log C(\zeta_2 \epsilon_T, \mathcal{B}_T, \|.\|_1), \quad \Pi(B_2(\epsilon_T, B)) \geqslant \prod_{l,k} \tilde{\Pi}_h\left(\left\|h_{lk} - h_{lk}^0\right\|_2 < \epsilon_T\right).$$

Finding $B_T$ that verifies (67) can in fact be deduced from Theorem 2.1 in van der Vaart and van Zanten (2009) that we recall here. Let $\mathbb{H}$ be the Reproducing Kernel Hilbert Space of $k_{GP}$ and $\phi_{h_0}(\varepsilon)$ be the concentration function associated to $\tilde{\Pi}_{h|\delta}$ defined as

$$\phi_{h_0}(\varepsilon) = \inf_{h \in \mathbb{H}, \left\|h_{lk} - h_{lk}^0\right\|_2 \leqslant \varepsilon} \left\|h_{lk} - h_{lk}^0\right\|_{\mathbb{H}} - \log \tilde{\Pi}(\|h_{lk}\|_2 \leqslant \varepsilon), \quad \varepsilon > 0.$$

For any $\epsilon_T > 0$ such that $\phi_{h_0}(\epsilon_T) \leqslant T\epsilon_T^2$, there exists $\mathcal{B}_T \subset L_2([0, A])$ satisfying

$$\tilde{\Pi}_h(\mathcal{B}_T^c) \leqslant e^{-CT\epsilon_T^2}, \quad \log C(3\epsilon_T, \mathcal{B}_T, \|.\|_2) \leqslant 6CT\varepsilon_T^2, \quad \tilde{\Pi}_h(\left\|h_{lk} - h_{lk}^0\right\|_\infty < 2\epsilon_T) \geqslant e^{-T\epsilon_T^2},$$

for any $C > 1$ such that $e^{-CT\epsilon_T^2} < 1/2$. Since $\|h_{lk}\|_1 \leqslant \sqrt{A}\,\|h_{lk}\|_2$, we then obtain that

$$\log C(3\sqrt{A}\epsilon_T, \mathcal{B}_T, \|.\|_1) \leqslant \log C(3\epsilon_T, \mathcal{B}_T, \|.\|_2) \leqslant 6CT\epsilon_T^2,$$

and finally, that $\log \rfloor(\zeta_0 \epsilon_T, \mathcal{H}_T, \|.\|_1) \leqslant 6CK^2 T\epsilon_T^2 \leqslant x_0 T\epsilon_T^2$ with $\zeta_0 = 3\sqrt{A}$, $x_0 = 12CK^2$.

Although more general kernel functions $k_{GP}$ could be considered, we focus on the hierarchical squared exponential kernels for which

$$\forall x, y \in \mathbb{R}, \quad k_{GP}(x, y; \ell) = \exp\left\{-(x - y)^2/\ell^2\right\}, \quad \ell \sim IG(\ell; a_0, a_1), \quad a_0, a_1 > 0,$$

where $IG(.; a_0, a_1)$ with $a_0, a_1 > 0$ is the Inverse Gamma distribution. The hierarchical squared exponential kernel is notably chosen in the variational method of Malem-Shinitski et al. (2021), and its adaptivity and near-optimality has been proved by van der Vaart and van Zanten (2009).

**Proposition 20** *Let $N$ be a Hawkes process with link functions $\phi = (\phi_k)_k$ and parameter $f_0 = (\nu_0, h_0)$ such that $(\phi, f_0)$ verify Assumption 6. Assume that for any $l, k \in [K]$, $h^0_{lk} \in \mathcal{H}(\beta, L_0)$ with $\beta > 0$ and $L_0 > 0$. Let $\tilde{\Pi}_{h|\delta}$ be the above Gaussian Process prior with hierarchical squared exponential kernel $k_{GP}$. Then, under our hierarchical prior, the mean-field variational distribution $\hat{Q}_1$ defined in* (40) *satisfies, for any $M_T \to +\infty$,*

$$\mathbb{E}_0\left[\hat{Q}_1\left(\|f - f_0\|_1 > M_T(\log\log T)^{1/2}(\log T)^q (T/\log T)^{-\beta/(2\beta+1)}\right)\right] \xrightarrow[T\to\infty]{} 0,$$

*with $q = 1$ if $\phi$ verifies Assumption 6(i) and $q = 3/2$ if $\phi$ verifies Assumption 6(ii).*

Given Theorem 7, Proposition 20 is then a direct consequence of Theorem 7 and van der Vaart and van Zanten (2009), therefore its proof is omitted.

**Remark 21** *The Gaussian process prior has been used in variational methods for Hawkes processes when there exists a conjugate form of the mean-field variational posterior distribution, i.e., $\hat{Q}_1$ is itself a Gaussian process with mean function $m_{VP}$ and kernel function $k_{VP}$. This is notably the case in the sigmoid Hawkes model under the latent variable augmentation scheme described in Section 4.1 and used for instance byMalem-Shinitski et al. (2021). Since the computation of the Gaussian process variational distribution is often expensive for large data set, the latter is often further approximated using the sparse Gaussian process approximation via inducing variables (Titsias and Lázaro-Gredilla, 2011). Using results of Nieman et al. (2022), we conjecture that our result in Proposition 20 would also hold for the mean-field variational posterior with inducing variables.*

### D.4 Proof of Proposition 11

We recall that in this proof, we assume that $K$ is fixed (but can be large). Since the complete graph $\delta_C = \mathbb{1}\mathbb{1}^T$ is larger than the true graph $\delta_0$, the subspace $\bigcup_{m\in\mathcal{M}_C} \mathcal{F}_m$ contains the true parameter $f_0$. Hence, Proposition 10 remains valid with $\mathcal{V}_C = \cup_{m\in\mathcal{M}_C}\{\{m\} \times \mathcal{V}^m\}$ and the corresponding adaptive variational posterior $\hat{Q}^C_{MS}$ concentrates on $f_0$ at the rate $\epsilon_T = (\log T)^q T^{-\beta/(2\beta+1)}$.

By assumption, with probability going to 1 as $T \to \infty$, it holds that

$$\sum_{l,k\leqslant K} |\hat{S}_{lk} - S^0_{lk}| \leqslant \epsilon_T,$$

which in particular implies that

$$S^0_{lk} - \epsilon_T \leqslant \hat{S}_{lk} \leqslant S^0_{lk} + \epsilon_T \tag{68}$$

Moreover, we have that

$$\mathbb{P}_0(\hat{\delta} \neq \delta_0) \leqslant \sum_{l,k} \mathbb{P}_0(\hat{\delta}_{lk} \neq \delta^0_{lk}). \tag{69}$$

We now consider 2 cases:

- if $(l, k) \in I(\delta_0)$, then

$$\begin{aligned}\mathbb{P}_0(\hat{\delta}_{lk} \neq \delta^0_{lk}) = \mathbb{P}_0(\hat{S}_{lk} < \eta_0) &\leqslant \mathbb{P}_0(\hat{S}_{lk} - S^0_{lk} < \eta_0 - s_0) \\ &\leqslant \mathbb{P}_0(\hat{S}_{lk} - S^0_{lk} < -2\epsilon_T) = o(1),\end{aligned}$$

  using (68) and since $S^0_{lk} \geqslant s_0 \geqslant \eta_0 + 2\epsilon_T$.

- if $(l,k) \notin I(\delta_0)$, then $S^0_{lk} = 0$ and

$$\mathbb{P}_0(\hat{\delta}_{lk} \neq \delta^0_{lk}) = \mathbb{P}_0(\hat{S}_{lk} > \eta_0) = P_0(\hat{S}_{lk} - S^0_{lk} > \eta_0) = o(1),$$

since $\eta_0 \geqslant 2\epsilon_T$.

Hence, from (69), we deduce (36).

For the second part of Proposition 11. Recall that $K^2 - K_0$ is the first index $i$ such that $S^0_{(i+1)} > S^0_{(i)}$, where $(S^0_{(i)})_{i\in[K^2]}$ corresponds to the values of $(S^0_{lk})_{l,k}$ in increasing order. Using again (68), we have:

- for any $i < K^2 - K_0 - 1$,

$$\begin{aligned}\mathbb{P}_0(\hat{S}_{(i+1)} - \hat{S}_{(i)} > 2\epsilon_T) \leqslant \mathbb{P}_0(\hat{S}_{(i+1)} > 2\epsilon_T) &= \mathbb{P}_0(\forall j \geqslant i,\, \hat{S}_{(j)} > 2\epsilon_T) \\ &\leqslant P_0(\forall j \geqslant K^2 - K_0,\, \hat{S}_{(j)} > 2\epsilon_T).\end{aligned}$$

Now with probability going to 1,

$$\sum_{lk} |\hat{S}_{lk} - S_{lk}| \leqslant 2\epsilon_T$$

hence with probability to 1 the number of $\hat{S}_{lk} > 2\epsilon_T$ is equal to $K_0$ and

$$\max_{i<K^2-K_0-1} \mathbb{P}_0(\hat{S}_{(i+1)} - \hat{S}_{(i)} > 2\epsilon_T) = o(1).$$

- if $i \geqslant K^2 - K_0$, then using the above reasoning,

$$\mathbb{P}_0(\hat{S}_{(K^2-K_0+1)} - \hat{S}_{(K^2-K_0)} \leqslant s_0 - 4\epsilon_T) \leqslant \mathbb{P}_0(\hat{S}_{(K^2-K_0+1)} \leqslant s_0 - 2\epsilon_T) + \mathbb{P}_0(\hat{S}_{(K^2-K_0)} > 2\epsilon_T) = o(1)$$

where the second term is $o(1)$ since with probability going to 1 $\hat{S}_{(K^2-K_0)} \leqslant 2\epsilon_T$ and the second is due to the fact that for all $j > K^2 - K_0$ $\hat{S}_{(j)} \geqslant s_0 - 2\epsilon_T$.

This terminates the proof of Proposition 11.

## Appendix E. Proof of Proposition 8

In this section we recall and prove the posterior and variational posterior concentration rates for the sigmoid Hakwes model, with link function $\phi_k(x) = \theta_k\sigma(\alpha(x-\eta))$, $\theta_k, \alpha > 0$, $\eta \in \mathbb{R}$, $\forall k$. For any $C > 0$ and $f \in \mathcal{F}$, we recall our definition of the truncated $L_1$-norm and interaction function parameter:

$$\|f - f_0\|_{1,C} = \|f^C - f_0\|_1, \quad f^C = (\nu, (h^C_{lk})_{l,k}), \quad h^C_{lk} = (h_{lk} \vee -2C) \wedge 2C,$$

**Theorem 22** *Let $N$ be a Hawkes process with sigmoid link functions $\phi = (\phi_k)_k$ and parameter $f_0 = (\nu_0, h_0)$. Let $\epsilon_T = o(1/\sqrt{\kappa_T})$ be a positive sequence verifying $\log^3 T = O(T\epsilon_T^2)$, $\Pi$ be a prior distribution on $\mathcal{F}$, and $\mathcal{V}$ a variational family of distributions on $\mathcal{F}$. We assume that Assumptions (A0)-(A2) from Theorem 7 are satisfied for $T$ large enough with $\kappa_T = 10\log T$.*

*Then, for any $M_T \to \infty$, $C > 0$ such that $\max_{l,k} \|h^0_{lk}\|_\infty \leqslant C$, and $\hat{Q}$ defined in* (9)*, we have that*

$$\mathbb{E}_0\left[\Pi\left(\left\|f^C - f_0\right\|_1 > M_T\sqrt{\kappa_T}\epsilon_T\right)\right] \xrightarrow[T\to\infty]{} 0$$
$$\mathbb{E}_0\left[\hat{Q}\left(\left\|f^C - f_0\right\|_1 > M_T\sqrt{\kappa_T}\epsilon_T\right)\right] \xrightarrow[T\to\infty]{} 0$$
$$\mathbb{P}_0\left(\sum_{l,k=1}^{K}\left|\mathbb{E}_{\hat{Q}}[\|h^C_{lk}\|_1] - \|h^0_{lk}\|_1\right| > M_T\sqrt{\kappa_T}\epsilon_T\right) \xrightarrow[T\to\infty]{} 0.$$

**Proof** For the first statement on the posterior distirbution, the argument of this proof is the same as the proof of Theorem 5.5 in Sulem et al. (2024) for the concentration in stochastic distance $\tilde{d}_{1T}$ and the following inequality obtained in the proof of Theorem 7 for the sigmoid link function:

$$\mathbb{E}_0\left[\mathbb{1}_{\Omega_T}\Pi[f \in \mathcal{F} : \tilde{d}_{1T}(f, f_0) > M'_T\varepsilon_T|N]\right] \leqslant 4(2K+1)e^{-x_1 M'^2_T T\varepsilon^2_T/2},$$

where $x_1 > 0$, $M'_T > 0$, $\varepsilon_T = \sqrt{\kappa_T}\epsilon_T$ and $\Omega_T$ is defined in the proof of Theorem 7. We now state and prove a result similar to Lemma 18, which holds if the inverse of the link function is Lipschitz on a large enough interval, for the specific case of the sigmoid link function.

**Lemma 23** *Let $f \in \mathcal{F}_T$ such that $\|\nu - \nu_0\|_1 \leqslant C$. Then*

$$\mathbb{E}_f[Z_{1l}] \geqslant \bar{\zeta}\|f^C - f_0\|_1,$$

*with $\bar{\zeta} > 0$ a constant that may depend on $f_0, C$ and $(\alpha, \eta, (\theta_k)_k)$.*

The proof of this lemma follows the same argument as that of Lemma 4 in Sulem et al. (2024) (Section S9.20). With $x > 0$, for each $k \in [K]$, we define the event $\Omega_k$ as

$$\Omega_k = \left\{\max_{k'\neq k} N^{k'}[\tau_1, \tau_2) = 0,\ N^k[\tau_1, \tau_1 + x] = 0,\ N^k[\tau_1 + x, \tau_1 + x + A] = 1,\ N^k[\tau_1 + x + A, \tau_2) = 0\right\}.$$

On $\Omega_k$, we have

$$\mathbb{E}_f[Z_{1l}] \geqslant \sum_k \mathbb{E}_f\left[\mathbb{1}_{\Omega_k}\int_{\tau_1}^{A+U^{(1)}_1}|\lambda^l_t(f) - \lambda^l_t(f_0)|dt\right],$$

where $U^{(1)}_1$ is the time of the first event on $N^k$. Using $\mathbb{Q}$ the point process measure of a homogeneous Poisson process with unit intensity on $\mathbb{R}^+$ and equal to the null measure on $[-A, 0)$, we can rewrite the previous inequality as

$$\mathbb{E}_f[Z_{1l}] \geqslant \sum_k \mathbb{E}_{\mathbb{Q}}\left[\int_{\tau_1}^{U^{(1)}_1+A}\mathcal{L}_t(f)\mathbb{1}_{\Omega_k}|\lambda^l_t(f) - \lambda^l_t(f_0)|\right]dt,$$

with $\mathcal{L}_t(f)$ the likelihood process given by

$$\mathcal{L}_t(f) = \exp\left(Kt - \sum_k\int_{\tau_1}^t\lambda^k_u(f)du + \sum_k\int_{\tau_1}^t\log(\lambda^k_u(f))dN^k_u\right), \quad t \geqslant 0.$$

On $\Omega_k$, for $t \in [\tau_1, U_1^{(1)} + A)$, we have

$$\begin{aligned}\mathcal{L}_t(f) &\geqslant e^{Kt}\lambda^k_{U_1^{(1)}}(f)\exp\left\{-\sum_{k'}\int_{\tau_1}^t \phi_{k'}(\tilde{\lambda}^{k'}_u(f))du\right\}\\ &\geqslant \phi_k(\nu_k)\exp\left\{-(x+A)\sum_k \theta_k\right\}\\ &\geqslant \phi_k(\nu_k^0 - C)\exp\left\{-(x+A)\sum_k \theta_k\right\} = C_1,\end{aligned}$$

using that $f$ is such that $\|\nu - \nu_0\|_1 \leqslant C$. Hence we have

$$\mathbb{E}_f\left[Z_{1l}\right] \geqslant C_1 \sum_k \mathbb{E}_{\mathbb{Q}}\left[\mathbb{1}_{\Omega_k}\int_{U_1^{(1)}}^{U_1^{(1)}+A}\left|\phi_l(\tilde{\lambda}^l_t(f)) - \phi_l(\tilde{\lambda}^l_t(f_0))\right|dt\right].$$

Moreover, since for any $t \in [U_1^{(1)}, U_1^{(1)} + A]$, $|h_{kl}(t - U_1^{(1)} - h^0_{kl}(t - U_1^{(1)})| \geqslant |h^C_{kl}(t - U_1^{(1)} - h^0_{kl}(t - U_1^{(1)})|$, we have

$$\begin{aligned}|\phi_l(\tilde{\lambda}^l_t(f)) - \phi_l(\tilde{\lambda}^l_t(f_0))| &= |\phi_l(\nu_l + h_{kl}(t - U_1^{(1)})) - \phi_l(\nu_l^0 + h^0_{kl}(t - U_1^{(1)}))|\\ &\geqslant |\phi_l(\nu_l + h^C_{kl}(t - U_1^{(1)})) - \phi_l(\nu_l^0 + h^0_{kl}(t - U_1^{(1)}))| = |\phi_l(\tilde{\lambda}^l_t(f^C)) - \phi_l(\tilde{\lambda}^l_t(f_0))|,\end{aligned}$$

and $\tilde{\lambda}^l_t(f^C) \in [-3C + 3C]$. On $[-3C + 3C]$, $\phi_l^{-1}$ is $L_{l,C}$-Lipschitz with $L_{l,C} > 0$, which leads to

$$\begin{aligned}\mathbb{E}_f\left[Z_{1l}\right] &\geqslant C_1 \sum_k \mathbb{E}_{\mathbb{Q}}\left[\mathbb{1}_{\Omega_k}\int_{U_1^{(1)}}^{U_1^{(1)}+A}\left|\phi_l(\tilde{\lambda}^l_t(f^C)) - \phi_l(\tilde{\lambda}^l_t(f_0))\right|dt\right]\\ &\geqslant \frac{C_1}{L_{l,C}}\sum_k \mathbb{E}_{\mathbb{Q}}\left[\mathbb{1}_{\Omega_k}\int_{U_1^{(1)}}^{U_1^{(1)}+A}\left|\nu_l + h^C_{kl}(t - U_1^{(1)} - \nu_l^0 - h^0_{kl}(t - U_1^{(1)})\right|dt\right].\end{aligned}$$

Then the rest of proof is identical to that of Lemma A.4 in Sulem et al. (2024), replacing $h$ by $h^C$ and the constant $\bar{\zeta}$ in the statement of the lemma can depend on $f_0, C$ and $L_{l,C}$.

We now prove Theorem 22. Let $f \in \mathcal{F}_T$ such that $\|f^C - f_0\|_1 > M_T\varepsilon_T$ with $M_T$ such that $M_T \geqslant 4\mathbb{E}_0\left[\Delta\tau_1\right]M'_T$. Note that $\|f - f_0\|_1 \geqslant \|f^C - f_0\|_1$. Using Lemma 23 we have

$$\begin{aligned}\mathbb{E}_f\left[\mathbb{1}_{\Omega_T}\mathbb{1}_{B_{d_1}(M'_T\varepsilon_T)}|\mathcal{G}_0\right] &\leqslant \mathbb{P}_f\left[\sum_{j=1}^{J_T-1} Z_{jl} \leqslant TM'_T\varepsilon_T|\mathcal{G}_0\right]\\ &\leqslant \sum_{J\in\mathcal{J}_T}\mathbb{P}_f\left[\sum_{j=1}^{J-1} Z_{jl} - \mathbb{E}_f\left[Z_{jl}\right] \leqslant TM'_T\epsilon_T - \frac{T}{2\mathbb{E}_0\left[\Delta\tau_1\right]}\bar{\zeta}M_T\epsilon_T|\mathcal{G}_0\right]\\ &\leqslant \sum_{J\in\mathcal{J}_T}\mathbb{P}_f\left[\sum_{j=1}^{J-1} Z_{jl} - \mathbb{E}_f\left[Z_{jl}\right] \leqslant -\frac{T}{4\mathbb{E}_0\left[\Delta\tau_1\right]}\bar{\zeta}M_T\varepsilon_T|\mathcal{G}_0\right],\end{aligned}$$

since $M_T \geqslant 4\mathbb{E}_0[\Delta\tau_1] M'_T$. Besides, denoting $S_k = \{x : h_{kl}(x) \in [-2C, 2C]^c\}$ and $S_+ = \cup S_k$, we have

$$\begin{aligned}
Z_{1l} &= \int_{\tau_1}^{\xi_1} |\phi(\tilde{\lambda}_t^l(f)) - \phi(\tilde{\lambda}_t^l(f_0))|dt = (U_1 - \tau_1)|\phi(\nu_l) - \phi(\nu_l^0| + \int_{U_1}^{\xi_1} |\phi(\tilde{\lambda}_t^l(f)) - \phi(\tilde{\lambda}_t^l(f_0))|dt \\
&\leqslant (U_1 - \tau_1)|\phi(\nu_l) - \phi(\nu_l^0| + |S_+|\theta_l + \int_{U_1}^{U_1+A} |\phi(\nu_l + \sum_k \mathbb{1}_{U_1 \in N^k} h_{kl}^C(t - U_1)) - \phi(\nu_l^0 + \sum_k \mathbb{1}_{U_1 \in N^k} h_{kl}^0(t - U_1))|dt \\
&\leqslant (U_1 - \tau_1)|\phi(\nu_l) - \phi(\nu_l^0| + |S_+|\theta_l + \int_0^A |\phi(\nu_l + \sum_k \mathbb{1}_{U_1 \in N^k} h_{kl}^C(t)) - \phi(\nu_l^0 + \sum_k \mathbb{1}_{U_1 \in N^k} h_{kl}^0(t))|dt.
\end{aligned}$$

Moreover, since $\phi_l$ is $L_l$-Lipschitz with $L_l = \theta_l \alpha$, on the one hand we have

$$\begin{aligned}
&\int_0^A |\phi(\nu_l + \sum_k \mathbb{1}_{U_1 \in N^k} h_{kl}^C(t) - \phi(\nu_l^0 + \sum_k \mathbb{1}_{U_1 \in N^k} h_{kl}^0(t))|dt \\
&\leqslant L_l \int_0^A |\nu_l + \sum_k \mathbb{1}_{U_1 \in N^k} h_{kl}^C(t) - \nu_l^0 + \sum_k \mathbb{1}_{U_1 \in N^k} h_{kl}^0(t)|dt \\
&\leqslant L_l(|\nu_l - \nu_l^0| + \sum_k \|h_{kl}^C - h_{kl}^0\|_1) \leqslant L_l \|f^C - f_0\|_1.
\end{aligned}$$

On the other hand we have $\|h_{kl}^C - h_{kl}^0\|_1 \geqslant |S_k| C$ which implies

$$\|f^C - f_0\|_1 \geqslant \sum_k \|h_{kl}^C - h_{kl}^0\|_1 \geqslant \sum_k |S_k| C \geqslant C|S^+|.$$

Hence we obtain

$$Z_{1l} \leqslant (U_1 - \tau_1)|\phi(\nu_l) - \phi(\nu_l^0)| + (\theta_l C^{-1} + L_l)\|f^C - f_0\|_1 \leqslant [(U_1 - \tau_1)L_l + (\theta_l C^{-1} + L_l)]\|f^C - f_0\|_1.$$

Therefore, since $\mathbb{E}_f[(\xi_1 - \tau_1)^n] = \frac{n!}{\|\phi(\nu)\|_1^n} \leqslant \frac{n!}{\|\phi(\nu_0 - C)\|_1^n}$, we obtain

$$\begin{aligned}
\mathbb{E}_f\left[Z_{1l}^n\right] &\leqslant 2^{n-1}[L_l^n \mathbb{E}_f[(\xi_1 - \tau_1)^n] + (\theta_l C^{-1} + L_l)^n]\|f^C - f_0\|_1^n) \\
&\leqslant 2^{n-1}[L_l^n \frac{n!}{\|\phi(\nu_0 - C)\|_1^n} + (\theta_l C^{-1} + L_l)^n]\|f^C - f_0\|_1^n \\
&\leqslant 2^{n-1} n! \max\left(\frac{L_1}{\|\phi(\nu_0 - C)\|_1^n}, (\theta_l C^{-1} + L_l)\right)^n \|f^C - f_0\|_1^n \\
&\leqslant \frac{1}{2} n! b^{n-2} v^2,
\end{aligned}$$

with $b := 2\max\left(\frac{L_1}{\|\phi(\nu_0 - C)\|_1^n}, (\theta_l C^{-1} + L_l)\right)\|f^C - f_0\|_1$ and $v := 2\max\left(\frac{L_1}{\|\phi(\nu_0 - C)\|_1^n}, (\theta_l C^{-1} + L_l)\right)\|f^C - f_0\|_1$. Applying Bernstein's inequality as in Sulem et al. (2024), we obtain for $J \geqslant \frac{T}{2\mathbb{E}_0[\Delta\tau_1]}$,

$$\mathbb{P}_f\left[\sum_{j=1}^{J-1} Z_{jl} - \mathbb{E}_f\left[Z_{jl}\right] \leqslant -\frac{T}{4\mathbb{E}_0[\Delta\tau_1]} \bar{\zeta} M_T \varepsilon_T | \mathcal{G}_0\right] \leqslant \exp\left\{-\frac{\bar{\zeta}^2 T}{16\bar{m}'}\right\}.$$

since with $\bar{m} := \max\left(\frac{L_1}{\|\phi(\nu_0 - C)\|_1^n}, (\theta_l C^{-1} + L_l)\right)$ we have

$$v^2 + b\frac{T}{4\mathbb{E}_0[\Delta\tau_1]}\bar{\zeta}M_T\varepsilon_T = 4\bar{m}^2\|f^C - f_0\|_1^2 + \bar{m}\frac{T}{4\mathbb{E}_0[\Delta\tau_1]}\bar{\zeta}M_T\varepsilon_T\|f^C - f_0\|_1 \geqslant \bar{m}' M_T^2\varepsilon_T^2.$$

and with $\bar{m}' = 4(\bar{m}^2 \vee \bar{m}\frac{T}{4\mathbb{E}_0[\Delta\tau_1]}\bar{\zeta})$. Following the arguments of Sulem et al. (2024), this proves the first statement of Theorem 22 and the two other statements can be proved in the same way as Theorem 7 (see Appendix D.2), replacing $f$ by $f^C$ since in this case $\|h_{lk^C}\|_1 \leqslant 2CA$ which implies that $\|f^C - f_0\|_1$ is bounded.

■

## Appendix F. Gibbs Sampler in the Sigmoid Hawkes Model

In this section, we describe a non-adaptive Gibbs sampler that computes the posterior distribution in the sigmoid Hawkes model, using the data augmentation scheme of Section 4 (see also Remark 5).

**Algorithm 4:** Gibbs sampler in the sigmoid Hawkes model with data augmentation

**Input:** $N = (N^1, \dots, N^K)$, $n_{iter}$, $\mu, \Sigma$.
**Output:** Samples $S = (f_i)_{i\in[n_{iter}]}$ from the posterior distribution $\Pi_A(f|N)$.

Precompute $(H_k(T_i^k))_i, k \in [K]$.
Initialise $f \sim \mathcal{N}(f, \mu, \Sigma)$ and $S = []$.
**for** $t \leftarrow 1$ *to* $n_{iter}$ **do**
  **for** $k \leftarrow 1$ *to* $K$ **do**
    **for** $i \leftarrow 1$ *to* $N_k$ **do**
      Sample $\omega_i^k \sim p_{PG}(\omega_i^k; 1, \tilde{\lambda}^k_{T_i^k}(f))$
    Sample $(\bar{T}_j^k)_{j=1,R_k}$ a Poisson temporal point process on $[0, T]$ with intensity $\theta_k\sigma(-\tilde{\lambda}_t^k(f))$
    **for** $j \leftarrow 1$ *to* $R_k$ **do**
      Sample $\bar{\omega}_j^k \sim p_{PG}(\omega; 1, \tilde{\lambda}^k_{\bar{T}_j^k}(f))$
    Update $\tilde{\Sigma}_k = [\beta^2 H_k D_k (H_k)^T + \Sigma^{-1}]^{-1}$
    Update $\tilde{\mu}_k = \tilde{\Sigma}_k\left(H_k\left[\beta v_k + \beta^2\eta u_k\right] + \Sigma^{-1}\mu\right)$
    Sample $f_k \sim \mathcal{N}(f_k; \tilde{\mu}_k, \tilde{\Sigma}_k)$
  Add $f = (f_k)_k$ to $S$.

## Appendix G. Additional Results from our Numerical Experiments

In this section, we report results from our simulation study in Section 6 that were not added to the main text for conciseness purposes. Each of the following sub-sections corresponds to one of the simulation set-up.

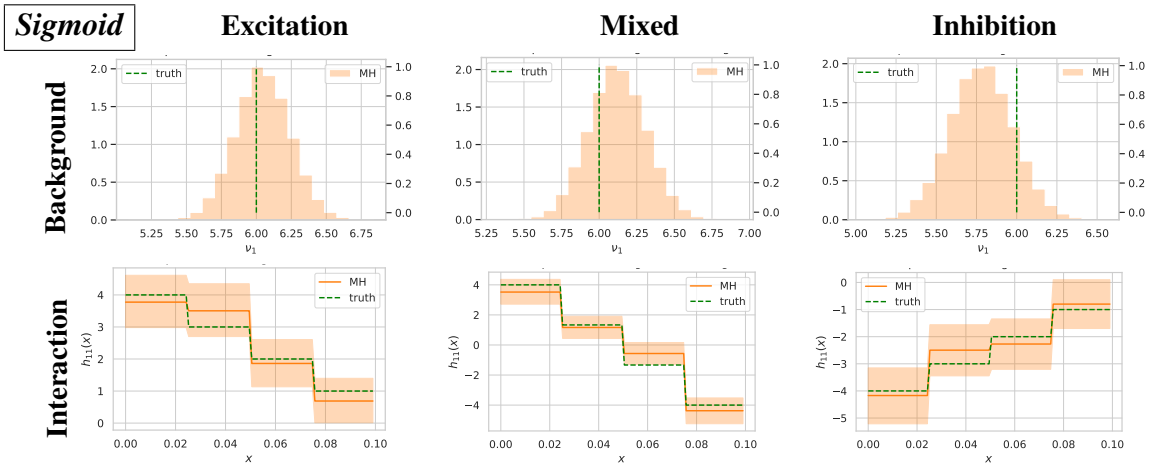



Figure 25: Posterior distribution on $f = (\nu_1, h_{11})$ obtained with the MH sampler in the sigmoid model, in the three scenarios of Simulation 1 ($K = 1$). The three columns correspond to the *Excitation only* (left), *Mixed effect* (center), and *Inhibition only* (right) scenarios. The first row contains the marginal distribution on the background rate $\nu_1$, and the second row represents the posterior mean (solid orange line) and 95% credible sets (orange areas) on the (self) interaction function $h_{11}$. The true parameter $f_0$ is plotted in dotted green line.

### G.1 Simulation 1

This section contains our results for the MH sampler, in the univariate settings of Simulation 1 with sigmoid and softplus link functions (see Figures 25 and 26).

### G.2 Simulation 3

This section contains additional results from Simulation 3: the estimated intensity function in the univariate and bivariate, well-specified settings (Figures 27 and 28), the estimated parameter in the mis-specified settings (Figure 29), and the estimated interaction functions in the bivariate setting and Inhibition scenario (Figure 30).

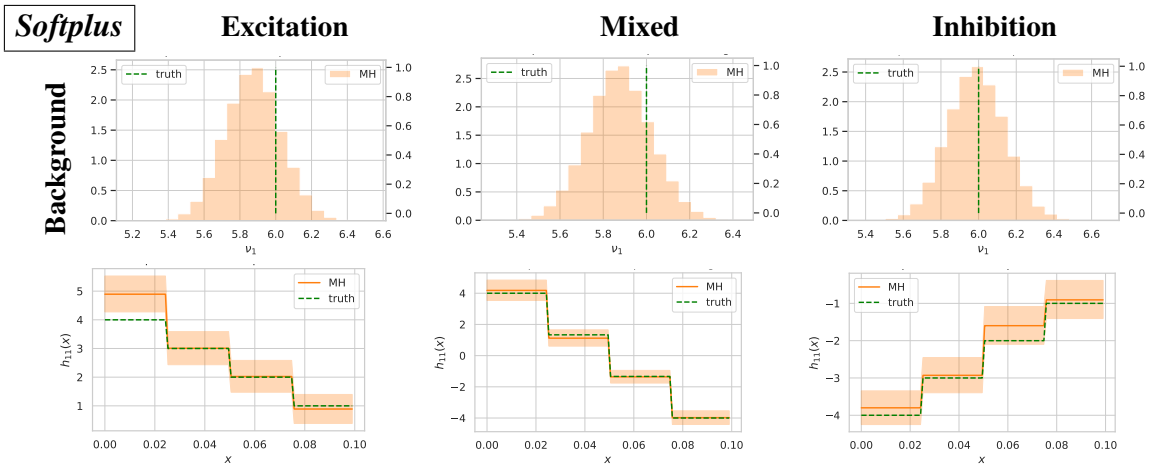
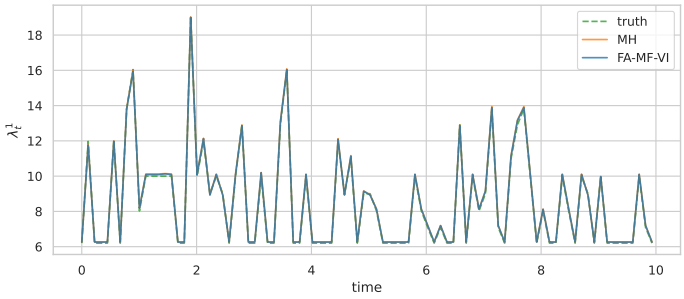
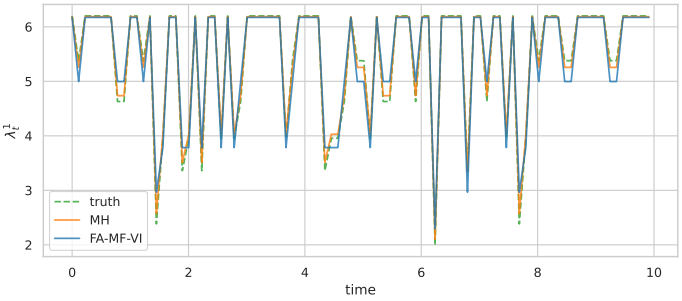



Figure 26: Posterior distribution on $f = (\nu_1, h_{11})$ obtained with the MH sampler in the softplus model, in the three scenarios of Simulation 1 ($K = 1$). The three columns correspond to the *Excitation only* (left), *Mixed effect* (center), and *Inhibition only* (right) scenarios. The first row contains the marginal distribution on the background rate $\nu_1$, and the second row represents the posterior mean (solid orange line) and 95% credible sets (orange areas) on the (self) interaction function $h_{11}$. The true parameter $f_0$ is plotted in dotted green line.

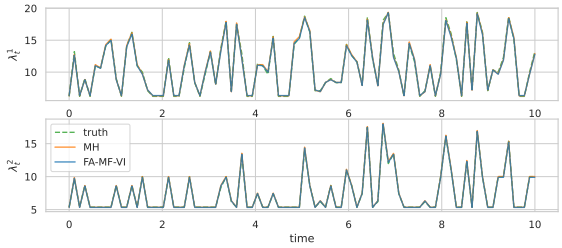



(a) Excitation scenario

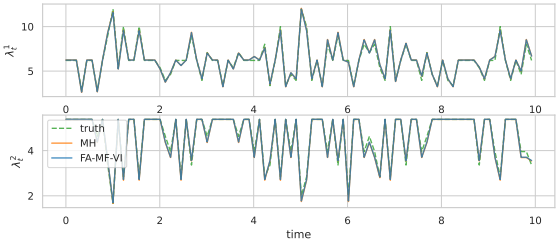



(b) Inhibition scenario

Figure 27: Intensity function on a subwindow of the observation window estimated via the variational posterior mean and via the posterior mean computed with the MH sampler, in the well-specified setting of Simulation 3 on $[0, 10]$, using the fully-adaptive mean-field variational (FA-MF-VI) algorithm (Algorithm 2). The true intensity $\lambda_t^1(f_0)$ is plotted in dotted green line.




(a) *Excitation* scenario




(b) Self-inhibition scenario

Figure 28: Estimated intensity function based on the (variational) posterior mean, in the well-specified and bivariate setting of Simulation 3 on $[0, 10]$, using the fully-adaptive mean-field variational (FA-MF-VI) algorithm (Algorithm 2). The true intensity $\lambda_t(f_0)$ is plotted in dotted green line.

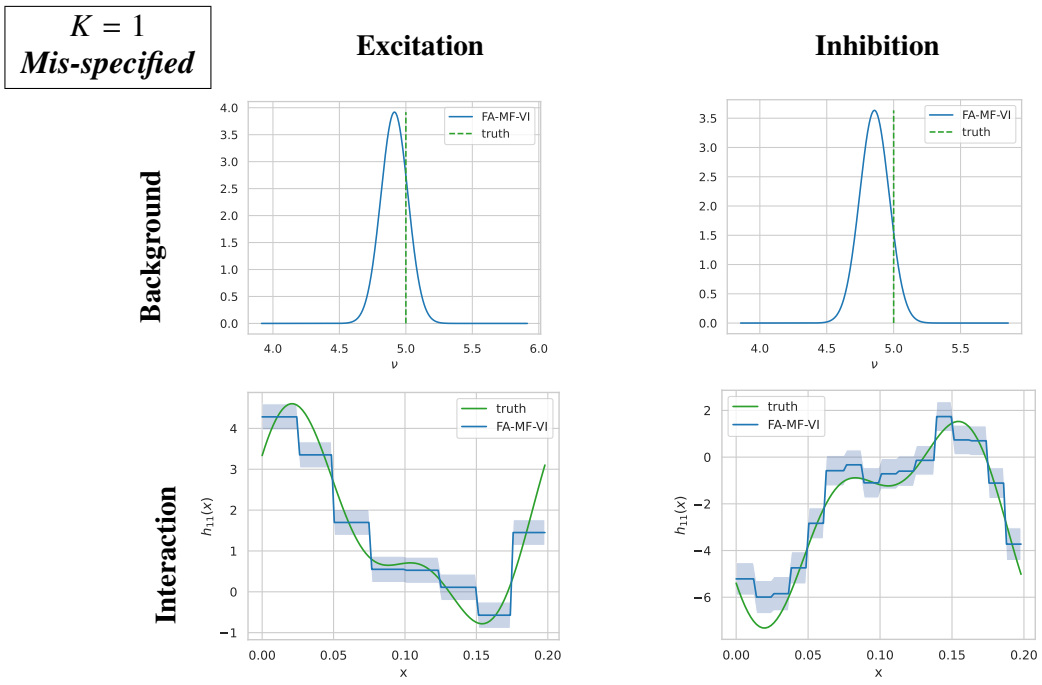



Figure 29: Model-selection variational posterior distributions on $f = (\nu_1, h_{11})$ in the univariate sigmoid model and mis-specified setting of Simulation 3, evaluated by the fully-adaptive mean-field variational (FA-MF-VI) algorithm (Algorithm 2). The two columns correspond to a (mostly) *Excitation* (left) and a (mostly) *Inhibition* (right) settings. The first row contains the marginal distribution on the background rate $\nu_1$, and the second row represents the variational posterior mean (solid line) and 95% credible sets (colored areas) on the (self) interaction function $h_{11}$. The true parameter $f_0$ is plotted in dotted green line.

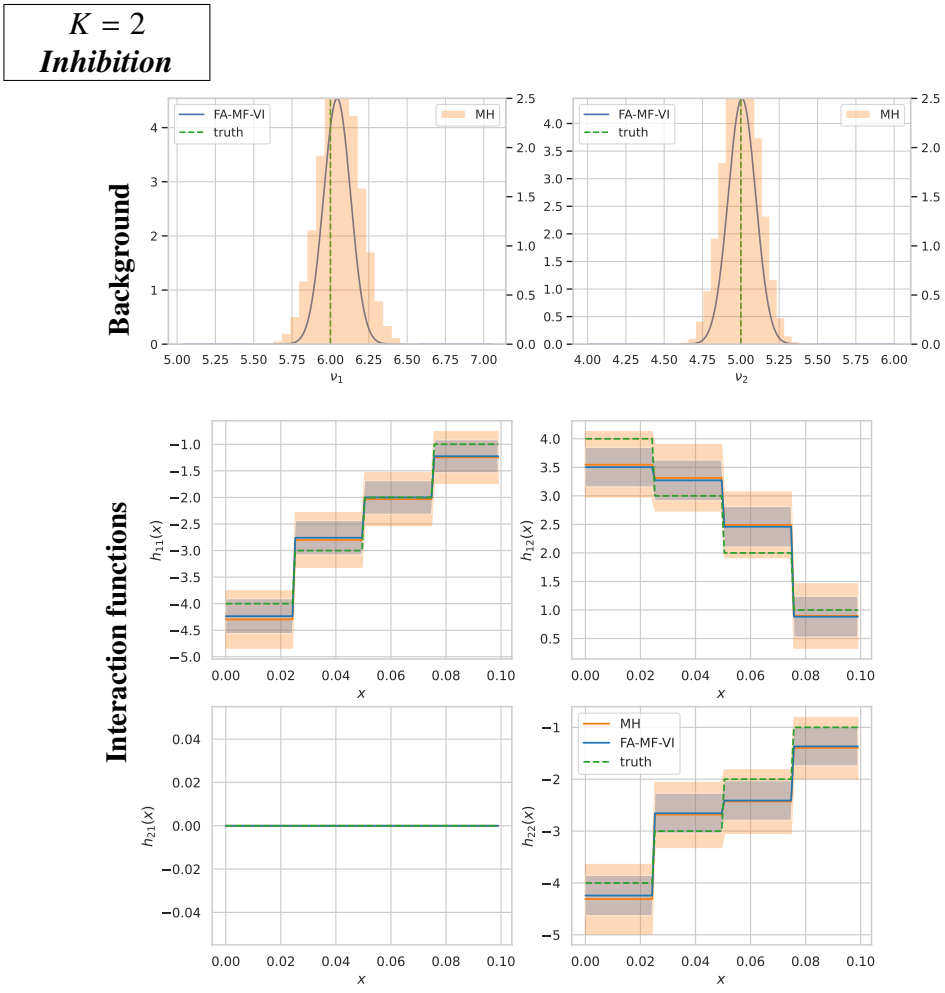
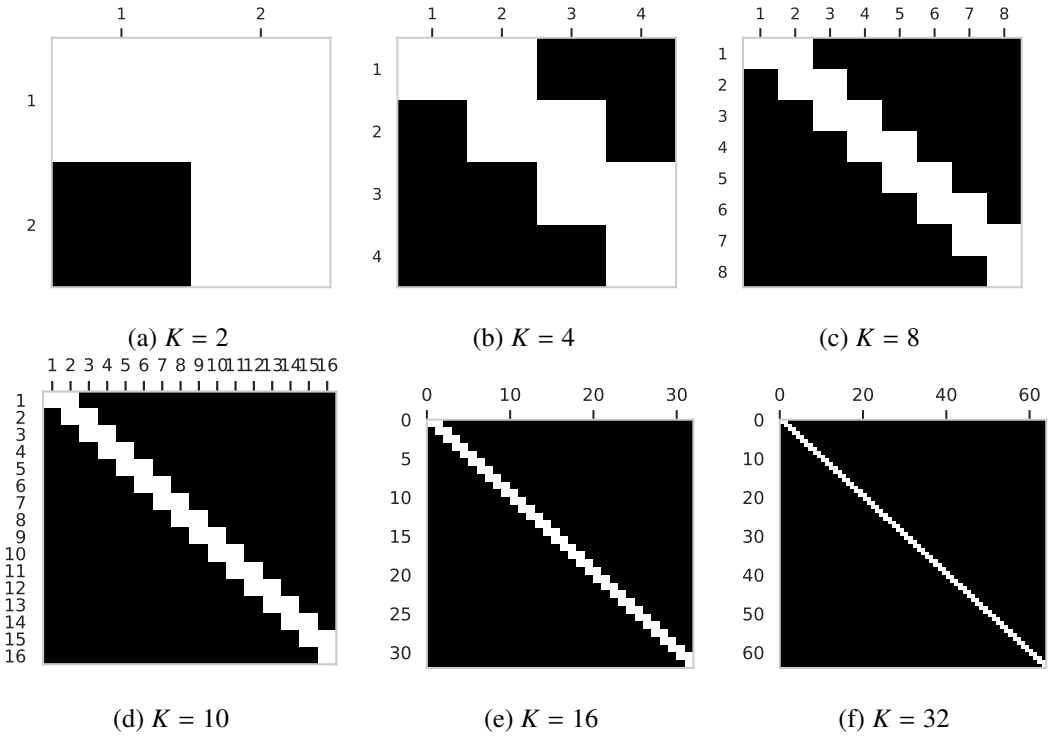



Figure 30: Posterior and model-selection variational posterior distributions on $f = (\nu, h)$ in the bivariate sigmoid model, well-specified setting, and Inhibition setting of Simulation 3, evaluated by the non-adaptive MH sampler and the fully-adaptive mean-field variational (FA-MF-VI) algorithm (Algorithm 2). The first row contains the marginal distribution on the background rates $(\nu_1, \nu_2)$, and the second and third rows represent the (variational) posterior mean (solid line) and 95% credible sets (colored areas) on the four interaction function $h_{11}, h_{12}, h_{21}, h_{22}$. The true parameter $f_0$ is plotted in dotted green line.

### G.3 Simulation 4

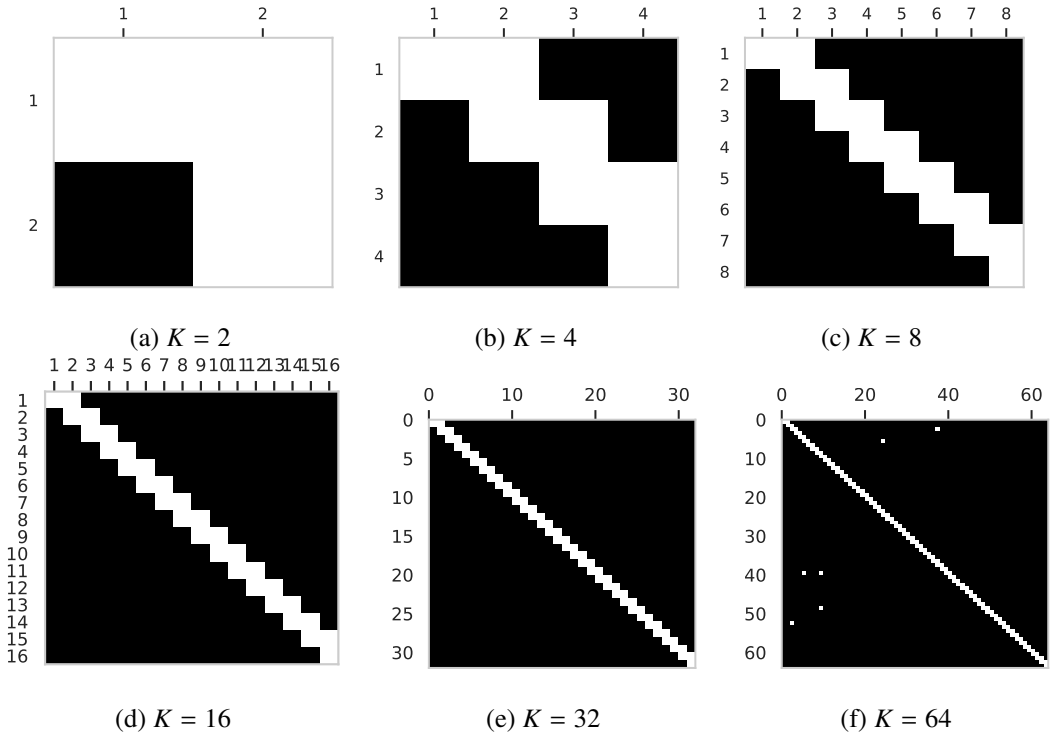
This section contains our results for the Inhibition setting of Simulation 4, i.e., the estimated graphs in (Figures 31 and 32), the heatmaps of the risk on the interaction functions in Figure 33, the estimated $L_1$-norms after the first step of Algorithm 3 in Figure 34, and the variational posterior distribution on the subset of the parameter in Figure 35. We also report a comparison of the risks obtained after the first and second steps of Algorithm 3 in Figure 36.



(a) $K = 2$ (b) $K = 4$ (c) $K = 8$

(d) $K = 10$ (e) $K = 16$ (f) $K = 32$

Figure 31: Estimated graph parameter $\hat{\delta}$ (black=0, white=1) for $K = 2, 4, 8, 16, 32, 64$ in the Excitation scenario of Simulation 4.



(a) $K = 2$ (b) $K = 4$ (c) $K = 8$

(d) $K = 16$ (e) $K = 32$ (f) $K = 64$

Figure 32: Estimated graph parameter $\hat{\delta}$ (black=0, white=1) for $K = 2, 4, 8, 16, 32, 64$ in the Inhibition scenario of Simulation 4.

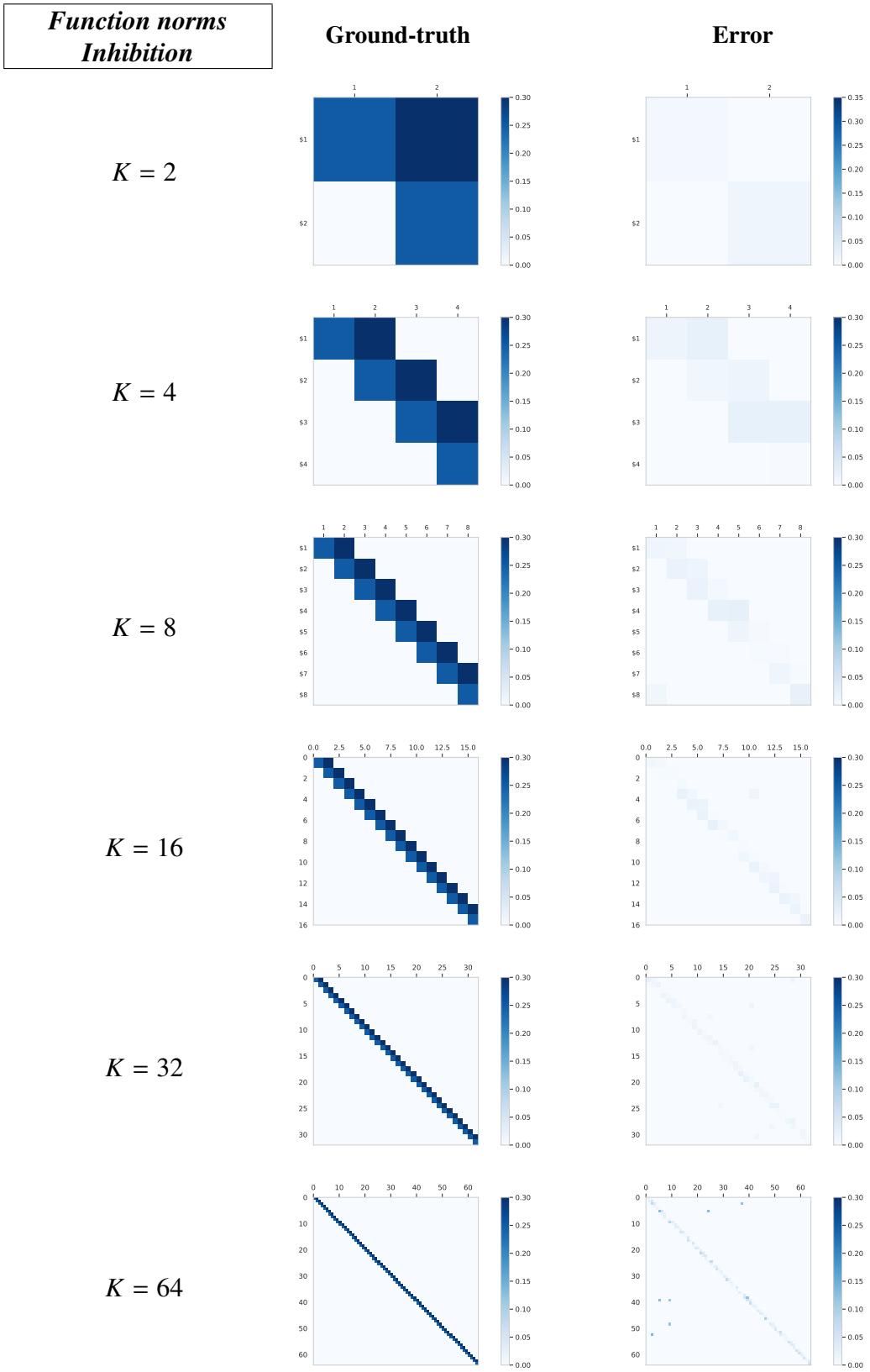
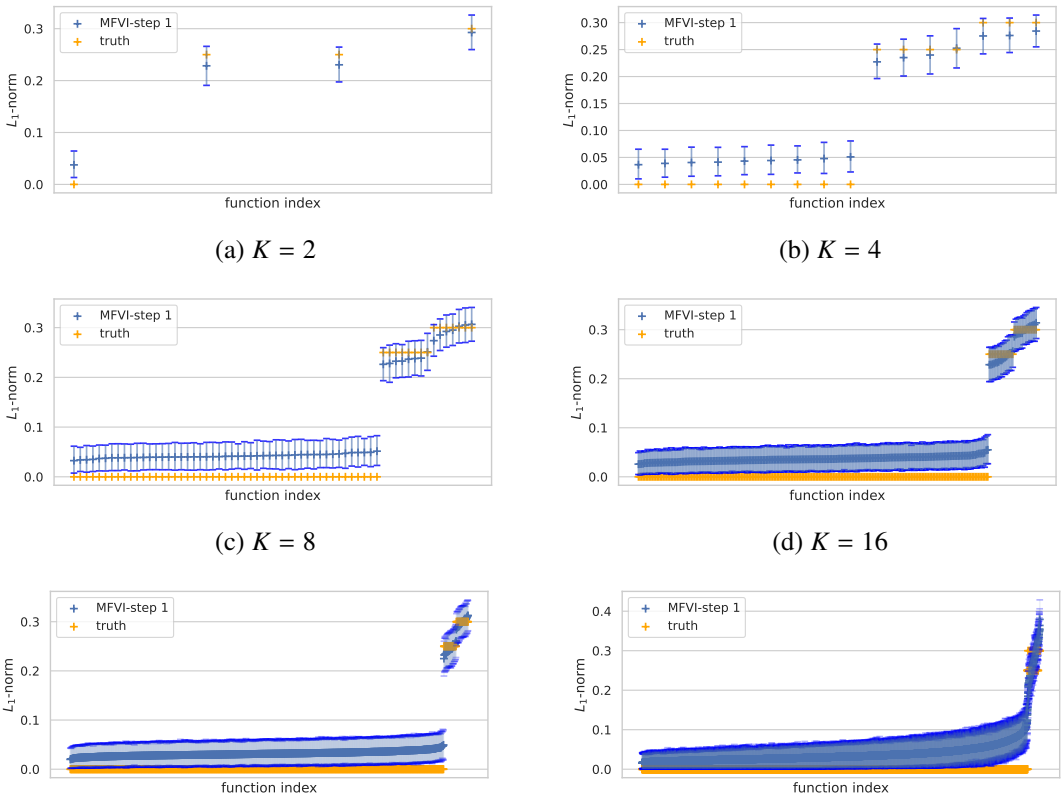



Figure 33: Heatmaps of the $L_1$-norms of the true parameter $h_0$, i.e., the entries of the matrix $S_0 = (S^0_{lk})_{l,k} = (\|h^0_{lk}\|_1)_{l,k}$ (left column) and the $L_1$-error of the model selection variational posterior mean (right column) obtained with Algorithm 3, in the Inhibition scenario of Simulation 4. The rows correspond to $K = 2, 4, 8, 16, 32, 64$.




(a) $K = 2$ (b) $K = 4$

(c) $K = 8$ (d) $K = 16$

(e) $K = 32$ (f) $K = 64$

Figure 34: Estimated $L_1$-norms after the first step of Algorithm 3 (in blue), and ground-truth norms (in orange), plotted in increasing order, in the Inhibition scenario of Simulation 4, for the models with $K \in \{2, 4, 8, 16, 32, 64\}$.

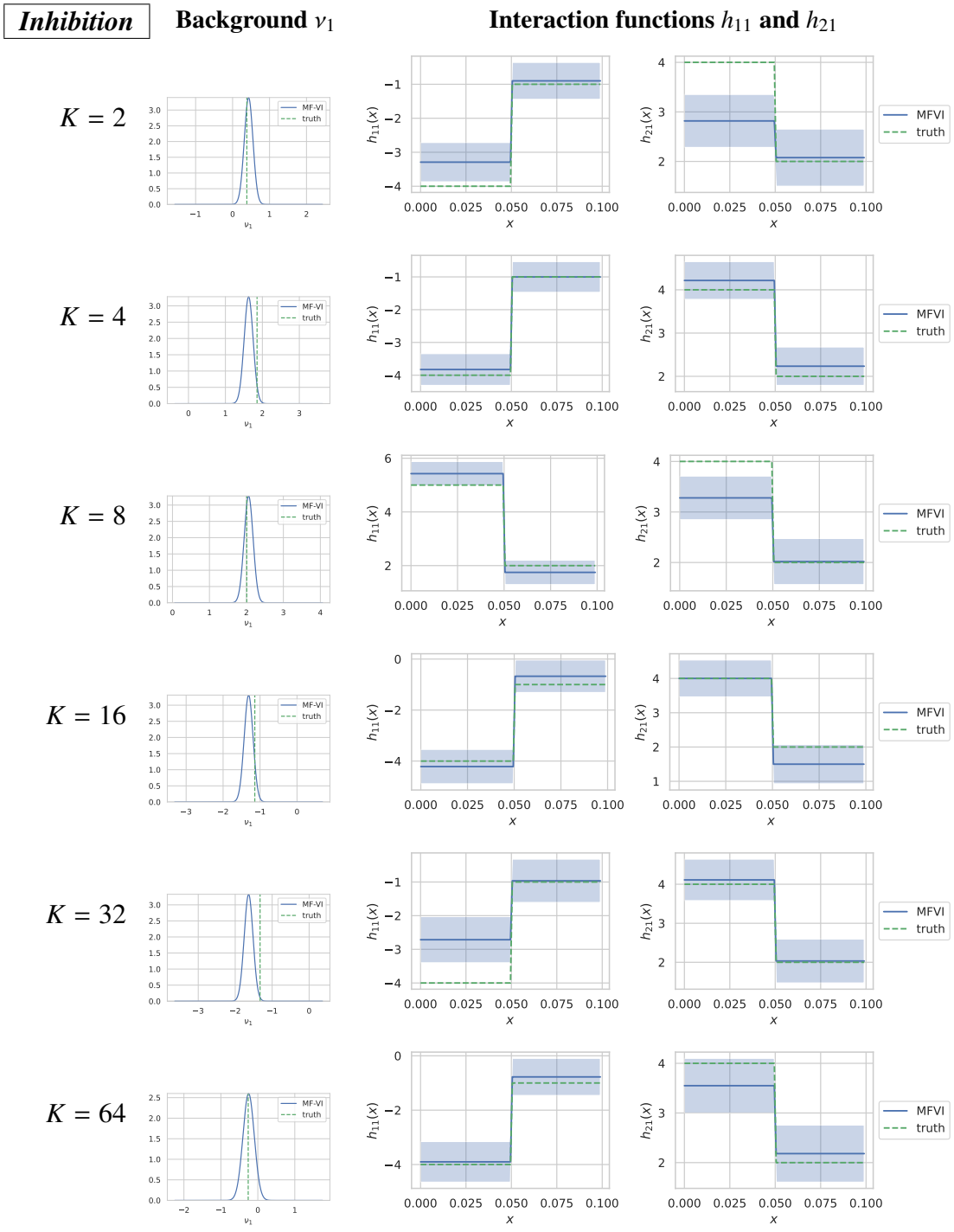
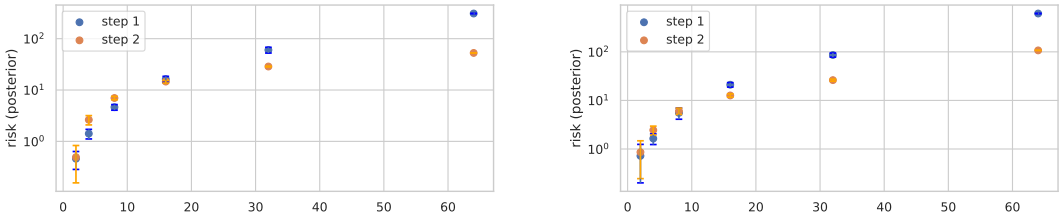



Figure 35: Model-selection variational posterior distributions on $\nu_1$ (left column) and interaction functions $h_{11}$ and $h_{21}$ (second and third columns) in the Inhibition scenario and multivariate sigmoid models of Simulation 4, computed with our two-step mean-field variational (MF-VI) algorithm (Algorithm 3). The different rows correspond to different multivariate settings $K = 2, 4, 8, 16, 32, 64$.




(a) Excitation (b) Inhibition

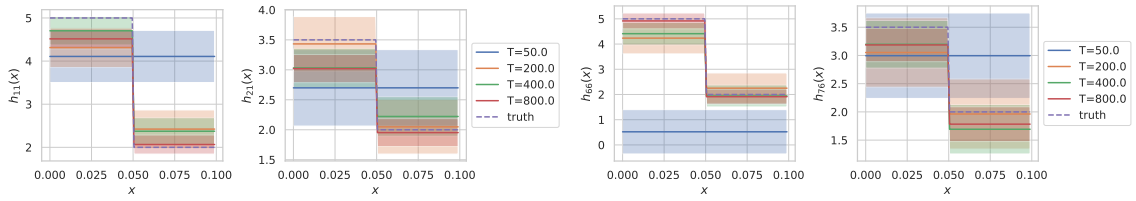
Figure 36: Risks of the variational posteriors obtained after the first step (in blue) and the second step (in orange) of Algorithm 3 in the *Excitation* scenario (left) and *Inhibition* scenario (right) of Simulation 4, for the models with $K = 2, 4, 8, 16, 32$. For each setting, we repeat the experiment 10 times and we plot the mean risk and the intervals at $\pm 2$ standard deviations.

### G.4 Simulation 5

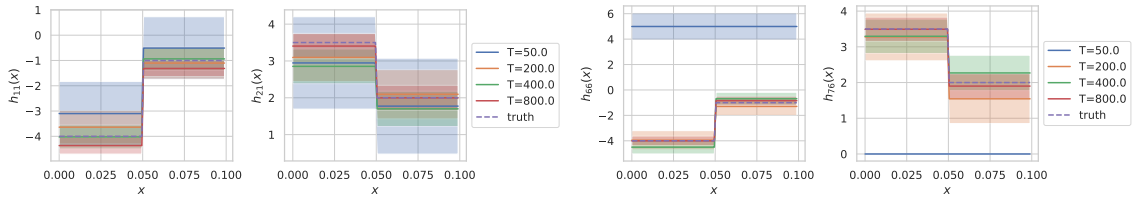
In this section, we report some characteristics of the simulated data in Simulation 5, in particular the number of points and excursions in each setting (see Table 9). Moreover, we report the plots of the posterior distribution in a subset of the parameter in Figure 37.

| Scenario | T | # events | # excursions | # local excursions |
|---|---|---|---|---|
| Excitation | 50 | 2621 | 36 | 114 |
| | 200 | 10,729 | 155 | 473 |
| | 400 | 21,727 | 303 | 957 |
| | 800 | 42,904 | 596 | 1921 |
| Inhibition | 50 | 1747 | 49 | 134 |
| | 200 | 7019 | 222 | 529 |
| | 400 | 13,819 | 466 | 1053 |
| | 800 | 27,723 | 926 | 2118 |

Table 9: Number of points and *global* and average *local* excursions in the multidimensional data sets of Simulation 5 ($K = 10$).




(a) *Excitation* scenario

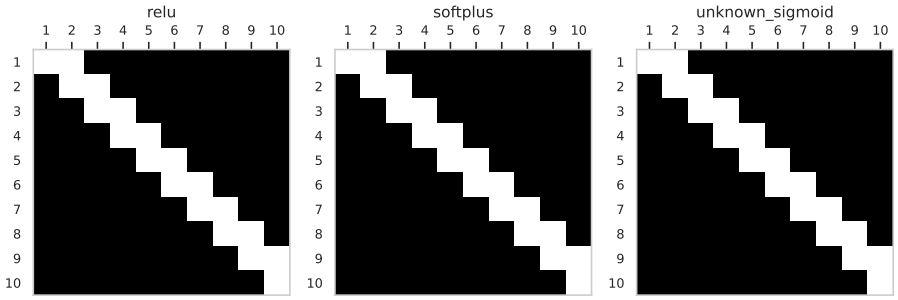



(b) *Inhibition* scenario

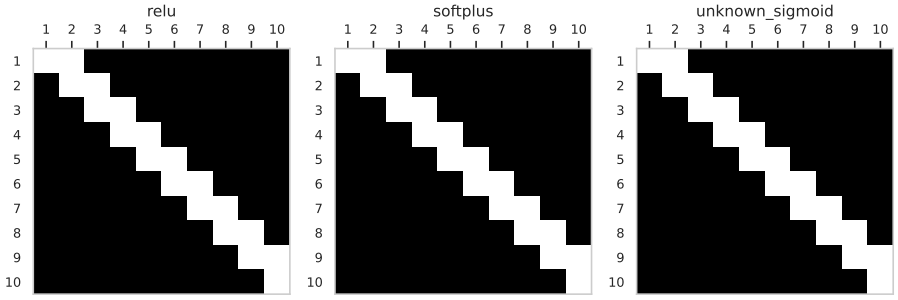
Figure 37: Model-selection variational posterior on two interaction functions $h_{66}$ and $h_{76}$, for different observation lengths $T \in \{50, 200, 400, 800\}$, in the *Excitation* and *Inhibition* scenarios in Simulation 5 with $K = 10$. We note that in this simulation, the true number of basis functions is 2 and is well recovered for all values of $T$. The estimation of these two interaction functions is poor for the smallest $T$, however, it improves when $T$ increases.

### G.5 Simulation 6

This section contains the estimated graphs (Figures 38 and 40), the variational posterior distribution on a subset of the parameter (Figures 39 and 41), in the mis-specified settings of Simulation 6.

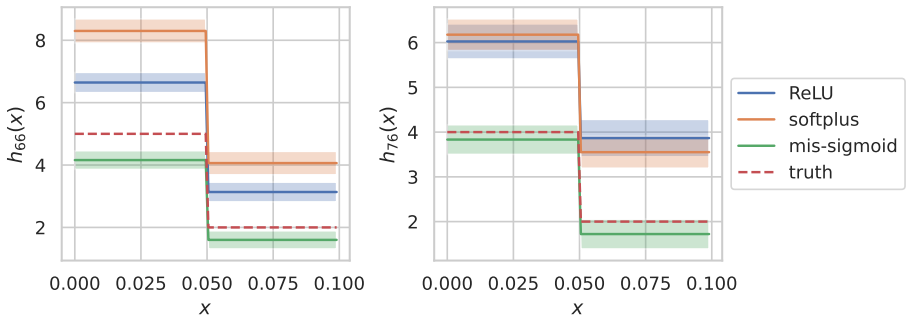



(a) *Excitation* scenario

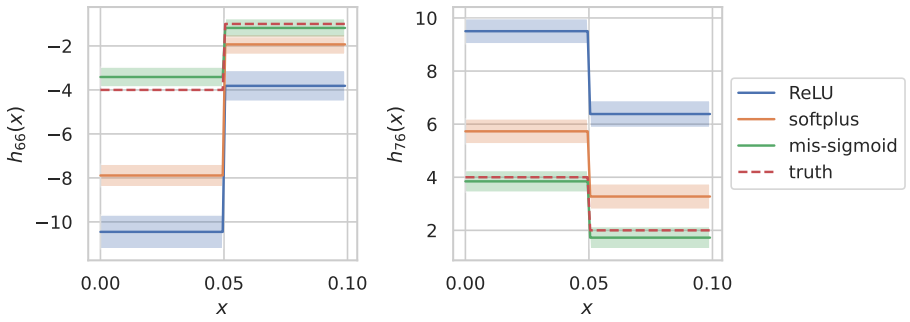



(b) *Inhibition* scenario

Figure 38: Estimated graph after thresholding the $L_1$-norms using the "gap" or "slope change" heuristic, in the different settings of mis-specified link functions of Simulation 6, and in the *Excitation* and *Inhibition* scenarios. We observe that the true graph (with non-null principal and first off-diagonal) is correctly estimated for the ReLU mis-specification setting, while some errors happen in the two other link settings, in particular in the *Inhibition* scenario.

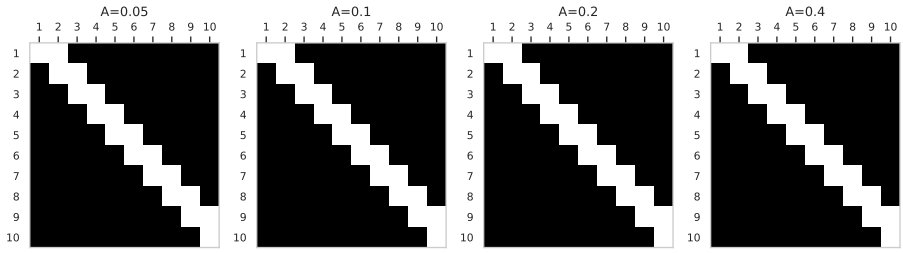



(a) *Excitation* scenario

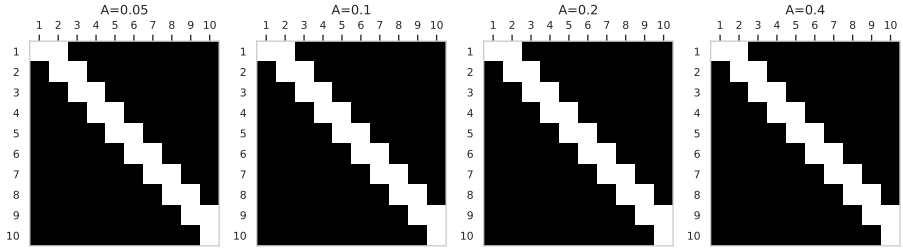



(b) *Inhibition* scenario

Figure 39: Estimated interaction functions $h_{66}$ and $h_{76}$ in the mis-specified settings of Simulation 6, where the data is generated from a Hawkes model with ReLU, softplus, or a mis-specified link function, and in the *Excitation* and *Inhibition* scenarios. We note that the estimation of the interaction functions is deteriorated in these mis-specified cases, however the sign of the functions are still recovered.

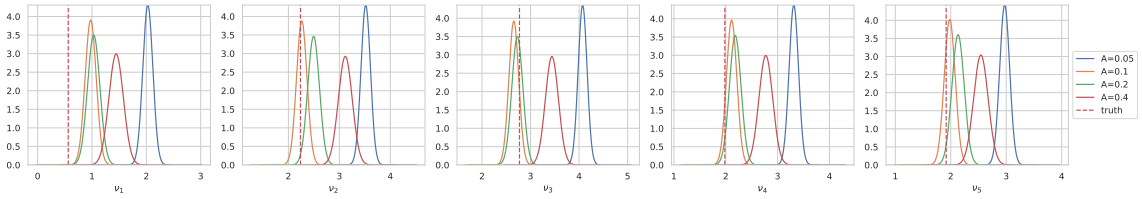



(a) *Excitation* scenario

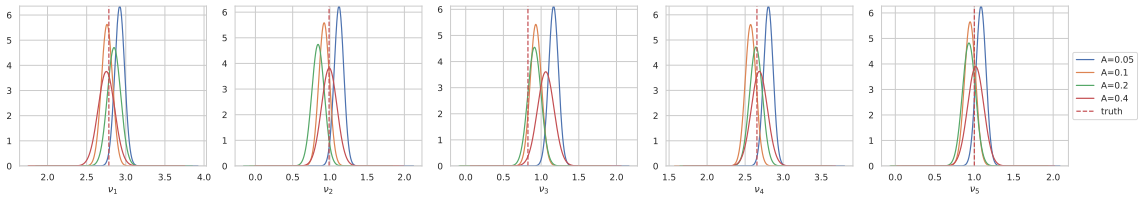



(b) *Inhibition* scenario

Figure 40: Estimated graph after thresholding the $L_1$-norms, when using Algorithm 3 with different support upper bounds $A' \in \{0.5, 0.1, 0.2, 0.4\}$, containing the true memory parameter $A = 0.1$, in the settings of Simulation 7. We note that the true graph (with non-null principal and first off-diagonal) is correctly estimated in all cases, in the *Excitation* scenario (first row) and in the *Inhibition* scenario (second row).




(a) *Excitation* scenario




(b) *Inhibition* scenario

Figure 41: Estimated background rates $\nu_k$ for $k = 1, \ldots, 5$ when using different values of the upper bound parameter $A \in \{0.05, 0.1, 0.2, 0.4\}$, in the two scenarios of Simulation 8. As expected, the background rates are better estimated in the well-specified setting $A = A_0 = 0.1$; nonetheless, when $A$ is not too far above $A_0$, the estimation does not deteriorate too much, in particular in the *Inhibition* scenarios.

## References

Ryan Prescott Adams, Iain Murray, and David J. C. MacKay. Tractable nonparametric bayesian inference in poisson processes with gaussian process intensities. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 9–16, 2009. doi: 10.1145/1553374.1553376.

Julyan Arbel, Ghislaine Gayraud, and Judith Rousseau. Bayesian optimal adaptive estimation using a sieve prior. *Scandinavian journal of statistics*, 40(3):549–570, 2013.

Emmanuel Bacry and Jean-Francois Muzy. Second order statistics characterization of Hawkes processes and non-parametric estimation, 2015.

Emmanuel Bacry, Martin Bompaire, Stéphane Gaïffas, and Jean-Francois Muzy. Sparse and low-rank multivariate Hawkes processes. *Journal of Machine Learning Research*, 21(50):1–32, 2020.

Christopher M. Bishop. *Pattern recognition and machine learning*. Information Science and Statistics. Springer, New York, 2006. doi: 10.1007/978-0-387-45528-0.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, Apr 2017. doi: 10.1080/01621459.2017.1285773.

Anna Bonnet, Miguel Martinez Herrera, and Maxime Sangnier. Maximum likelihood estimation for Hawkes processes with self-excitation or inhibition. *Statistics & Probability Letters*, 179:109214, 2021.

Pierre Brémaud and Laurent Massoulié. Stability of nonlinear Hawkes processes. *The Annals of Probability*, pages 1563–1588, 1996.

Biao Cai, Jingfei Zhang, and Yongtao Guan. Latent network structure learning from high-dimensional multivariate point processes. *Journal of the American Statistical Association*, 119 (545):95–108, 2024.

Chris U Carmona and Geoff K Nicholls. Scalable semi-modular inference with variational meta-posteriors. *arXiv preprint arXiv:2204.00296*, 2022.

Lisbeth Carstensen, Albin Sandelin, Ole Winther, and Niels R Hansen. Multivariate Hawkes process models of the occurrence of regulatory elements. *BMC bioinformatics*, 11(1):1–19, 2010.

I. Castillo and A. van der Vaart. Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics*, 40:2069–2101, 2012.

I. Castillo, J. Schmidt-Hieber, and A. van der Vaart. Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43:1986–2018, 2015.

Shizhe Chen, Ali Shojaie, Eric Shea-Brown, and Daniela Witten. The multivariate Hawkes process in high dimensions: Beyond mutual excitation. *arXiv:1707.04928v2*, 2017.

Manon Costa, Carl Graham, Laurence Marsalle, and Viet Chi Tran. Renewal in Hawkes processes with self-excitation and inhibition. *Advances in Applied Probability*, 52(3):879–915, 2020. doi: 10.1017/apr.2020.19.

Daryl J Daley and David Vere-Jones. *An introduction to the theory of point processes: volume II: general theory and structure*. Springer Science & Business Media, 2007.

Isabella Deutsch and Gordon J. Ross. Bayesian estimation of multivariate hawkes processes with inhibition and sparsity, 2022.

Christian Donner and Manfred Opper. Efficient bayesian inference of sigmoidal Gaussian Cox processes. *Journal of Machine Learning Research*, 19(67):1–34, 2018.

Sophie Donnet, Vincent Rivoirard, and Judith Rousseau. Nonparametric bayesian estimation for multivariate hawkes processes. *Annals of Statistics*, 48(5):2698–2727, 2020.

Michael Eichler, Rainer Dahlhaus, and Johannes Dueck. Graphical modeling for multivariate Hawkes processes with nonparametric link functions. *Journal of Time Series Analysis*, 38(2): 225–242, 2017.

Felipe Gerhard, Moritz Deger, and Wilson Truccolo. On the stability and dynamics of stochastic spiking neuron models: Nonlinear Hawkes process and point process glms. *PLOS Computational Biology*, 13:1–31, 02 2017. doi: 10.1371/journal.pcbi.1005390.

Gene H Golub and John H Welsch. Calculation of Gauss quadrature rules. *Mathematics of computation*, 23(106):221–230, 1969.

Niels Richard Hansen, Patricia Reynaud-Bouret, and Vincent Rivoirard. Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, 21(1):83–143, 2015.

Alan G Hawkes. Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 33(3):438–443, 1971.

Alan G. Hawkes. Hawkes processes and their applications to finance: a review. *Quantitative Finance*, 18(2):193–198, 2018. doi: 10.1080/14697688.2017.1403131.

M. Hoffmann, J. Rousseau, and J. Schmidt-Hieber. On adaptive posterior concentration rates. *The Annals of Statistics*, 43:2259–2295, 2015.

J. F. C. Kingman. *Poisson processes*, volume 3 of *Oxford Studies in Probability*. The Clarendon Press Oxford University Press, New York, 1993.

Remi Lemonnier and Nicolas Vayatis. Nonparametric Markovian learning of triggering kernels for mutually exciting and mutually inhibiting multivariate Hawkes processes. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 161–176. Springer, 2014.

Rafael Lima. Hawkes processes modeling, inference, and control: An overview. *SIAM Review*, 65 (2):331–374, 2023. doi: 10.1137/21M1396927.

Xiaofei Lu and Frédéric Abergel. High-dimensional Hawkes processes for limit order books: modelling, empirical analysis and numerical calibration. *Quantitative Finance*, 18(2):249–264, 2018.

Noa Malem-Shinitski, Cesar Ojeda, and Manfred Opper. Nonlinear Hawkes process with Gaussian process self effects, 2021.

Hongyuan Mei and Jason Eisner. The neural Hawkes process: A neurally self-modulating multivariate point process, 2017.

G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108, 2011. doi: 10.1198/jasa.2011.ap09546.

Dennis Nieman, Botond Szabo, and Harry van Zanten. Contraction rates for sparse variational approximations in Gaussian process regression. *Journal of Machine Learning Research*, 23(205): 1–26, 2022.

Yosihiko Ogata. Seismicity analysis through point-process modeling: A review. *Seismicity patterns, their statistical significance and physical meaning*, pages 471–507, 1999.

Ilsang Ohn and Lizhen Lin. Adaptive variational Bayes: Optimality, computation and applications. *The Annals of Statistics*, 52(1):335–363, 2024.

Jack Olinde and Martin B. Short. A self-limiting Hawkes process: Interpretation, estimation, and use in crime modeling. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 3212–3219, 2020. doi: 10.1109/BigData50022.2020.9378017.

Peter Pfaffelhuber, Stefan Rotter, and Jakob Stiefel. Mean-field limits for non-linear Hawkes processes with excitation and inhibition. *Stochastic Processes and their Applications*, 153:57–78, 2022.

Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349, 2013.

Kolyan Ray and Botond Szabó . Variational Bayes for high-dimensional linear regression with sparse priors. *Journal of the American Statistical Association*, pages 1–12, jan 2021. doi: 10.1080/01621459.2020.1847121.

Weining Shen and Subhashis Ghosal. Adaptive Bayesian procedures using random series priors. *Scandinavian Journal of Statistics*, 42(4):1194–1213, 2015. doi: https://doi.org/10.1111/sjos.12159.

Deborah Sulem, Vincent Rivoirard, and Judith Rousseau. Bayesian estimation of nonlinear Hawkes processes. *Bernoulli*, 30(2):1257–1286, 2024.

Michalis Titsias and Miguel Lázaro-Gredilla. Spike and slab variational inference for multi-task and multiple kernel learning. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24, 2011.

A. W. van der Vaart and J. H. van Zanten. Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. *The Annals of Statistics*, 37(5B), oct 2009.

John Worrall, Raiha Browning, Paul Wu, and Kerrie Mengersen. Fifty years later: new directions in Hawkes processes. *SORT (Statistics and Operations Research Transactions)*, 46(1):3–38, 2022.

Fengshuo Zhang and Chao Gao. Convergence rates of variational posterior distributions. *The Annals of Statistics*, 48(4):2180 – 2207, 2020.

Rui Zhang, Christian Walder, and Marian-Andrei Rizoiu. Variational inference for sparse Gaussian process modulated Hawkes process. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):6803–6810, Apr 2020. doi: 10.1609/aaai.v34i04.6160.

Feng Zhou, Zhidong Li, Xuhui Fan, Yang Wang, Arcot Sowmya, and Fang Chen. Efficient inference for nonparametric Hawkes processes using auxiliary latent variables. *Journal of Machine Learning Research*, 21(241):1–31, 2020.

Feng Zhou, Quyu Kong, Yixuan Zhang, Cheng Feng, and Jun Zhu. Nonlinear Hawkes processes in time-varying system, 2021.

Feng Zhou, Quyu Kong, Zhijie Deng, Jichao Kan, Yixuan Zhang, Cheng Feng, and Jun Zhu. Efficient inference for dynamic flexible interactions of neural populations. *Journal of Machine Learning Research*, 23(211):1–49, 2022.