Appendix 1. Likelihood and evidence lower bound

Before deriving the likelihood for the mixture of MHPs model, we start with the following Lemma about the relationship between the PDF and the conditional intensity function.

Lemma 1. Given the conditional intensity function of a Hawkes process, $\lambda^*(t)$, the probabilistic density function f^*, the survival function S^* and the cumulative distribution function F^* are given by:

$$f^*(t) = \lambda^*(t) \exp(-\int_{t_n}^{t} \lambda^*(s)ds)$$
$$S^*(t) = \exp(-\int_{t_n}^{t} \lambda^*(s)ds)$$
$$F^*(t) = 1 - \exp(-\int_{t_n}^{t} \lambda^*(s)ds).$$

Proof. By definition, we have:

$$\lambda^*(t) = \frac{f^*(t)}{1 - F^*(t)} = \frac{\frac{d}{dt}F^*(t)}{1 - F^*(t)} = -\frac{d}{dt}\log(1 - F^*(t))$$

The equation for F^* follows from the fundamental theorem of calculus, and the other two relationships can be derived using $f^* = \frac{d}{dt}F^*$ and $S^* = 1 - F^*$. □

The likelihood for the complete data $\{(t_n, Z_n, W_n)\}$ is given by:

$$L(t, Z, W) = \prod_{n=1}^{N} p(Z_n)p(W_n|Z_n)p^*(t_n|Z_n).$$

The first two terms are straightforward, i.e., the smoothing prior and the language model. Let $t_0 = 0$ and $t_{N+1} = T$, the last term can be derived as follows:

$$\begin{aligned}
L_{t|Z} &= \prod_{n=1}^{N} p^*(t_n|Z_n) \times S^*(T) \\
&= \prod_{n=1}^{N}\prod_{m=1}^{M} \lambda^*_{i_n,m}(t_n)^{Z_{nm}} \times \prod_{n=1}^{N}\prod_{m=1}^{M}\prod_{i=1}^{I} S^*_{i,m}(t_n) \times \prod_{m=1}^{M}\prod_{i=1}^{I} S^*_{i,m}(T) \\
&= \prod_{n=1}^{N}\prod_{m=1}^{M} \lambda^*_{i_n,m}(t_n)^{Z_{nm}} \times \prod_{n=1}^{N}\prod_{m=1}^{M}\prod_{i=1}^{I} \exp(-\int_{t_{n-1}}^{t_n} \lambda^*_{i,m}(s)ds) \times \prod_{m=1}^{M}\prod_{i=1}^{I} \exp(-\int_{t_N}^{T} \lambda^*_{i,m}(s)ds) \\
&= \prod_{n=1}^{N}\prod_{m=1}^{M} \lambda_{i_n,m}(t_n)^{Z_{mn}} \times \exp(-\sum_{i=1}^{I}\sum_{m=1}^{M}\int_0^T \lambda_{i,m}(s)ds).
\end{aligned}$$

which gives the Eq(11) in the paper.

The normalization term in the evidence lower bound (ELBO) is derived as follows:

$$
\begin{aligned}
&\sum_{i=1}^{I}\sum_{m=1}^{M}\int_0^T \mathbb{E}_q[\lambda_{i,m}(s)]ds \\
&= \sum_{i=1}^{I}\sum_{m=1}^{M}\int_0^T (\gamma_m\mu_i + \sum_{t_l<t}\phi_{tm}\alpha_{i_l i}\kappa(t-t_l))dt \\
&= T\sum_{i=1}^{I}\sum_{m=1}^{M}\gamma_m\mu_i + \sum_{i=1}^{I}\int_0^T\sum_{t_l<t}\alpha_{i_l i}\kappa(t-t_l)dt\sum_{m=1}^{M}\phi_{tm} \\
&= T\sum_{i=1}^{I}\sum_{m=1}^{M}\gamma_m\mu_i + \sum_{i=1}^{I}\int_0^T\sum_{t_l<t}\alpha_{i_l i}\kappa(t-t_l)dt \\
&= T\sum_{i=1}^{I}\sum_{m=1}^{M}\gamma_m\mu_i + \sum_{i=1}^{I}\sum_{n=1}^{N+1}\sum_{l=1}^{n-1}\alpha_{i_l i}\int_{t_{n-1}}^{t_n}\kappa(t-t_l)dt \\
&= T\sum_{i=1}^{I}\sum_{m=1}^{M}\gamma_m\mu_i + \sum_{i=1}^{I}\sum_{l=1}^{N}\alpha_{i_l i}\sum_{n=l+1}^{N+1}(K(t_n-t_l)-K(t_{n-1}-t_l)) \\
&= T\sum_{i=1}^{I}\sum_{m=1}^{M}\gamma_m\mu_i + \sum_{i=1}^{I}\sum_{l=1}^{N}\alpha_{i_l i}K(T-t_l),
\end{aligned}
$$

which proves Eq(15) in the paper.

Appendix 2. Inference and learning algorithm

In this section, we derive the inference and learning algorithms.

Inference for ϕ. The variational ELBO objective is decomposable for ϕs. Isolating the terms containing ϕ_n we have:

$$
\begin{aligned}
\max \mathfrak{L}_{\phi_n} = &\sum_{m=1}^{M}\phi_{nm}\left(\log\pi_m + \sum_{v=1}^{V}w_{nv}\log\beta_{mv} - \sum_{l=1}^{n-1}\phi_{lm}\eta_{ln}^m\log(\eta_{ln}^m) - \log\phi_{nm}\right) \\
&+ \sum_{m=1}^{M}\phi_{nm}\left(\eta_{nn}^m\log(\gamma_m\mu_{i_n}) + \sum_{\ell=1}^{n-1}\phi_{lm}\eta_{ln}^m\log(\alpha_{i_l i_n}k(t_n-t_l)) + \sum_{\ell=n+1}^{N}\phi_{lm}\eta_{nl}^m\log(\alpha_{i_n i_l}k(t_n-t_l))\right) \\
s.t.\ &\sum_{m=1}^{M}\phi_{nm} = 1.
\end{aligned}
$$

Eq(19) follows from the first-order optimality of the above optimization, i.e., by forming the Lagrangian and setting the derivative to zero.

Inference for η. The branching variables are optimal when the lower-bound in Eq(16) is tight, i.e., when the equality satisfies. This leads to:

$$\begin{aligned}\eta^m_{nn} &\propto \gamma_m \mu_{i_n},\\ \eta^m_{ln} &\propto \alpha_{i_l i_n} k(t_n - t_l),\\ \eta^m_{nn} + \sum_{l=1}^{n-1} \phi_{lm} \eta^m_{ln} &= 1,\end{aligned}$$

which leads to Eq(20–21).

Update α, μ and γ. The update formulas for π and β are trivial. Here we derive the update equations for μ, γ and α, given that the variational parameters are optimal. We have, for α, the ELBO:

$$\mathfrak{L}_\alpha = \sum_{n=1}^{N}\sum_{m=1}^{M} \phi_{nm} \sum_{l=1}^{n-1} \eta^m_{ln} \phi_{lm} \log(\alpha_{i_l i_n} k(t_n - t_l)) - \sum_{i=1}^{I}\sum_{l=1}^{N} \alpha_{i_l i} K(T - t_l).$$

Eq(25) follows by setting the derivative of $\mathfrak{L}_\alpha$ to zero. Eq(24) and Eq(26) can be derived similarly.

Appendix 3. An alternative inference algorithm

In Eq(16), we lower-bound the expectation of the log-intensity by applying Jensen's inequality with a set of branchings that satisfy Eq(17–18). Here, we present an alternative way to define these branchings and lower-bound the expected log-intensity, which leads to another version of the inference algorithm. This algorithm is slightly flawed in its mathematic form but in our experiments, we found that it converges faster than the one presented in the paper.

Again, we define a set of branching variables $\{\eta^m\}$, each of which fill a lower-triangular matrix, i.e., $\eta^m_n = [\eta^m_{1,n}, \ldots, \eta^m_{n,n}]^T$. But different from Eq(17–18), the branchings here define a multinomial distribution such that each $\eta^m_{l,n} \geq 0$ and $\sum_{l=1}^n \eta^m_{ln} = 1$.Let $\mathcal{E}[\eta^m_n] = -\sum_{l=1}^n \eta^m_{ln} \log(\eta^m_{ln})$ be the entropy of η^m_n, we have:

$$\begin{aligned}&\mathbb{E}_q[\log(\lambda_{i_n,m}(t_n))]\\ =&\mathbb{E}_q[\log(\eta^m_{nn} \frac{\gamma_m \mu_{i_n}}{\eta^m_{nn}} + \sum_{\ell=1}^{n-1} \eta^m_{ln} \frac{Z_{lm} \alpha_{i_l,i_n} k(t_n - t_l)}{\eta^m_{ln}})]\\ \geq&\eta^m_{nn} \log(\gamma_m \mu_{i_n}) + \sum_{\ell=1}^{n-1} \eta^m_{ln} (\mathbb{E}_q[\log(Z_{lm})] + \log(\alpha_{i_l,i_n} k(t_n - t_l))) + \mathcal{E}[\eta^m_n]\end{aligned}$$

Unfortunately, in the Equation above, the term $\mathbb{E}_q[\log(Z_{lm})]$, i.e., the expected logarithm of a Bernoulli variable, is *not well-defined*. As a proxy, we substitute it with $\log(\mathbb{E}_q[Z_{lm}]) = \log \phi_{lm}$. Note that this is not mathematically sound.

This new lower bound leads to the following inference and learning algorithms.

Variational inference:

$$\begin{aligned}\phi_{nm}^{\text{new}} \propto\ & \pi_m : \text{prior} \\ & \times \left(\prod_{v=1}^{V} \beta_{mv}^{w_{nv}}\right) : \text{content} \\ & \times (\gamma_m \mu_{i_n})^{\eta_{nn}^m} : \text{self triggering} \\ & \times \prod_{l=1}^{n-1} \left(\phi_{lm}^{\text{old}} \alpha_{i_l i_n} \kappa(t_n - t_l)\right)^{\eta_{ln}^m} : \text{influences from the past} \\ & \times \prod_{l=n+1}^{N} \exp(\eta_{nl}^m \frac{\phi_{lm}^{\text{old}}}{\phi_{nm}^{\text{old}}}) : \text{influences to the future}^{1}\end{aligned}$$

where the branchings are updated via:

$$\eta_{nn}^m = \frac{\gamma_m \mu_{i_n}}{\gamma_m \mu_{i_n} + \sum_{l=1}^{n-1} \phi_{lm} \alpha_{i_l i_n} k(t_n - t_l)}$$
$$\eta_{ln}^m = \frac{\phi_{lm} \alpha_{i_l i_n} k(t_n - t_l)}{\gamma_m \mu_{i_n} + \sum_{l=1}^{n-1} \phi_{lm} \alpha_{i_l i_n} k(t_n - t_l)}$$

Learning:

$$\pi_m \propto \sum_{n=1}^{N} \phi_{nm}$$

$$\beta_{mv} \propto \sum_{n=1}^{N} w_{nv} \phi_{nm}$$

$$\mu_i = \frac{\sum_{n=1}^{N} \delta(i_n = i) \sum_{m=1}^{M} \phi_{nm} \eta_{nn}^m}{\sum_{m=1}^{M} \gamma_m T}$$

$$\alpha_{ij} = \frac{\sum_{m=1}^{M} \sum_{n=1}^{N} \sum_{l=1}^{n-1} \delta(i_l = i) \delta(i_n = j) \phi_{nm} \eta_{ln}^m}{\sum_{n=1}^{N} \delta(i_n = i) K(T - t_n)}$$

$$\gamma_m = \frac{\sum_{n=1}^{N} \phi_{nm} \eta_{nn}^m}{\sum_{i=1}^{I} \mu_i T}$$

Note that the formulas for π and β are the same.

Appendix 4. Sparsity in MHPs and MMHP

In this appendix, we show that simply adding Lasso (i.e., ℓ_1) or ElasticNet (i.e., mixture of ℓ_1 and ℓ_2) penalties to the ELBO objective cannot lead to topological sparsity. To illustrate this, let us consider the plain MHP model (i.e., without latent mixture variables). Suppose $\lambda^*(t) = [\lambda_1^*(t), \ldots, \lambda_I^*(t)]^\top$ be the conditional intensity function of an I-dimensional Hawkes process, where

$$\lambda_i^*(t) = \mu_i + \sum_{t_l < t} \alpha_{i_l i} \kappa(t - t_l)$$

Given a sequence of events $\{(t_n, i_n)|n = 1, \ldots, N\}$. The log-likelihood is given by:

$$\mathcal{L} = \sum_{n=1}^{N} \log(\mu_{i_n} + \sum_{\ell=1}^{n-1} \alpha_{i_l,i_n} \kappa(t_n - t_\ell)) - T \sum_{i=1}^{I} \mu_i - \sum_{n=1}^{N} \sum_{i=1}^{I} \alpha_{i_n,i} K(T - t_n)$$

Introducing the branchings, we have the following lower-bound:

$$\mathcal{L} \geq \mathcal{J} = \sum_{n=1}^{N} (\eta_{nn} \log(\mu_{i_n}) + \sum_{\ell=1}^{n-1} \eta_{ln} \log(\alpha_{i_l,i_n} \kappa(t_n - t_\ell))) - T \sum_{i=1}^{I} \mu_i - \sum_{n=1}^{N} \sum_{i=1}^{I} \alpha_{i_n,i} K(T - t_n)$$

Optimizing this lower-bound given the branchings yields the following MLE for α:

$$\alpha_{ij} = \frac{\sum_{n=1}^{N} \sum_{l=1}^{n-1} \delta_{i_l,i} \delta_{i_n,j} \eta_{ln}}{\sum_{n=1}^{N} \delta_{i_n,i} K(T - t_n)},$$

the inferred infectivity is not sparse.

We want to enforce sparsity of α. Let $\mathcal{J}(\alpha)$ be the objective involving α, we have:

$$\mathcal{J}(\alpha) = \sum_{n=1}^{N} \sum_{\ell=1}^{n-1} \eta_{ln} \log(\alpha_{i_l,i_n} \kappa(t_n - t_\ell)) - \sum_{n=1}^{N} \sum_{i=1}^{I} \alpha_{i_n,i} K(T - t_n)$$

Note that this objective already includes a ℓ_1 regularization on α. Simply adding another ℓ_1 term cannot give us a sparse solution, i.e.:

$$\min \mathcal{J}(\alpha) - C \sum_{i,j=1}^{I} |\alpha_{ij}| = \sum_{n=1}^{N} \sum_{\ell=1}^{n-1} \eta_{ln} \log(\alpha_{i_l,i_n} \kappa(t_n - t_\ell)) - (\sum_{n=1}^{N} K(T - t_n) - C) \sum_{i=1}^{I} \alpha_{i_n,i}$$

yields

$$\alpha_{ij} = \left(\frac{\sum_{n=1}^{N} \sum_{l=1}^{n-1} \delta_{i_l,i} \delta_{i_n,j} \eta_{ln}}{\sum_{n=1}^{N} \delta_{i_n,i} K(T - t_n) - C} \right)^+$$

Note that in the above, although increasing the regularization strength C could lead to zero αs, but the regularization path is not well-behaved (see Figure 1).

Alternatively adding an ℓ_2 term (to make a elastic net type regularization) also doesn't work:

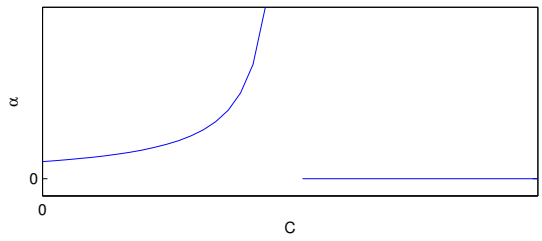

Figure 1: Regularization path.

$$\min \mathcal{J}(\alpha) - C \sum_{i,j} |\alpha_{ij}^2|$$

yields

$$\alpha_{ij} = \frac{-B + \sqrt{B^2 + 4C \sum_{n=1}^{N} \sum_{l=1}^{n-1} \delta_{i_l,i} \delta_{i_n,j} \eta_{ln}}}{2C}$$

where $B = \sum_{n=1}^{N} \delta_{i_n,i} K(T - t_n)$. Unfortunately, this solution is not sparse either (note that $\eta_{ln} = 0$ *iff* the corresponding $\alpha_{ij} = 0$). Also, this solution is not well-posed when the regularization strength is small (e.g., $C \to 0$).

Infeasibility of sparseness in MHPs: the reason. What happened? Why the sparsity encouraging penalties cannot yield solutions that are supposed to be sparse? The key reason has to do with the way we lower-bounding the log-likelihood. Particular, by introducing the branchings and breaking down the log-sum in the log-likelihood $\mathcal{L}$, the infectivities αs become singular points in the lower-bound $\mathcal{J}$, hence, sparsity is infeasible if $\mathcal{J}$ is optimized.

As such, in order to obtain sparsity, we need to optimize the log-likelihood $\mathcal{L}$ directly instead of its lower-bound $\mathcal{J}$. For a plain MHP model, this is not a big deal as the optimization can be solved with any convex optimization algorith. However, for the mixture of MHPs model, optimizing $\mathcal{L}$ directly is intractable. For example, if we didn't introduce the branchings to break down the log-sum, each meme-identity variable Z would be coupled with all Zs in the past and all Zs in the future in a very complicated nonlinear manner, inference of which is unimaginably troublesome.