Learning Granger Causality for Hawkes Processes Supplementary File

Hongteng Xu HXU42@GATECH.EDU
School of ECE, Georgia Institute of Technology

Mehrdad Farajtabar MEHRDAD@GATECH.EDU
College of Computing, Georgia Institute of Technology

Hongyuan Zha ZHA@CC.GATECH.EDU
College of Computing, Georgia Institute of Technology

1. Appendix

1.1. Derivation of Surrogate Objective Function

Using the Jensen's inequality, we have following inequality for all c and i:

$$\log\left(\mu_{u_i^c}+\sum\nolimits_{m=1}^{M}\sum\nolimits_{j=1}^{i-1}a^m_{u_i^c u_j^c}\kappa(\tau_{ij}^c)\right)$$
$$\geq p_{ii}\log\left(\frac{\mu_{u_i^c}}{p_{ii}}\right)+\sum_{m=1}^{M}\sum_{j=1}^{i-1}p_{ij}^m\log\left(\frac{a^m_{u_i^c u_j^c}\kappa(\tau_{ij}^c)}{p_{ij}^m}\right).$$

The equation holds if and only if $\mu_u = \mu_u^{(k)}$ and $a^m_{uu'} = a^{m,(k)}_{uu'}$. Therefore, we have $Q_{\Theta|\Theta^{(k)}} \geq \mathcal{L}_\Theta$ and $Q_{\Theta^{(k)}|\Theta^{(k)}} = \mathcal{L}_{\Theta^{(k)}}$.

1.2. Derivation of Learning Algorithm

We have surrogate objective function $F = -Q_{\Theta|\Theta^{(k)}} + \alpha_S\|\boldsymbol{A}\|_1 + \alpha_G\|\boldsymbol{A}\|_{1,2} + \alpha_P E_{\Theta|\Theta^{(k)}}(\boldsymbol{A})$, where $Q = -Q_{\Theta|\Theta^{(k)}} + +\alpha_P E_{\Theta|\Theta^{(k)}}(\boldsymbol{A})$ is the data fidelity term. Similar to (Simon et al., 2013), we choose a group $a_{uu'} = [a^1_{uu'},...,a^M_{uu'}]^\top$ to minimize and fix other parameters. Given current estimate $a^{(k)}_{uu'}$, we majorize Q as

$$Q \leq Q|_{a^{(k)}_{uu'}} + (a_{uu'}-a^{(k)}_{uu'})\nabla_{a_{uu'}}Q|_{a^{(k)}_{uu'}} + \frac{1}{2\eta}\|a_{uu'}-a^{(k)}_{uu'}\|_2^2. \tag{1}$$

Introducing (1) to the surrogate objective function, we rewrite the optimization problem as

$$\min_{a_{uu'}\geq\mathbf{0}} Q|_{a^{(k)}_{uu'}} + (a_{uu'}-a^{(k)}_{uu'})\nabla_{a_{uu'}}Q|_{a^{(k)}_{uu'}} + \frac{1}{2\eta}\|a_{uu'}-a^{(k)}_{uu'}\|_2^2 + +\alpha_S\|a_{uu'}\|_1 + \alpha_G\|a_{uu'}\|_2. \tag{2}$$

Because both $Q|_{a^{(k)}_{uu'}}$ and $\nabla_{a_{uu'}}Q|_{a^{(k)}_{uu'}}$ are known, we add $\frac{\eta}{2}\|\nabla_{a_{uu'}}Q|_{a^{(k)}_{uu'}}\|_2^2$ to the objective function of (2) and reduce $Q|_{a^{(k)}_{uu'}}$ from it, and obtain an equivalent optimization problem

$$\min_{a_{uu'}\geq\mathbf{0}}\frac{1}{2\eta}\|a_{uu'}-(a^{(k)}_{uu'}-\eta\nabla_{a_{uu'}}Q|_{a^{(k)}_{uu'}})\|_2^2 + \alpha_S\|a_{uu'}\|_1+\alpha_G\|a_{uu'}\|_2. \tag{3}$$

The objective function in (3) is convex, so the optimal solution is characterized by the subgradient equations.

$$a^{(k)}_{uu'}-\eta\nabla_{a_{uu'}}Q|_{a^{(k)}_{uu'}} - a_{uu'} = \eta\alpha_S\gamma+\eta\alpha_G\beta. \tag{4}$$

$\gamma = [\gamma_1,...,\gamma_M]^\top$, where $\gamma_m = 1$ if $a^m_{uu'} > 0$, and in $[0,1]$ otherwise. $\beta = \frac{a_{uu'}}{\|a_{uu'}\|_2}$ if $a_{uu'} \neq \mathbf{0}$, and in the set $\{x|\|x\|_2\leq 1\}$ otherwise. Combining the subgradient equations with the basic algebra in (Simon et al., 2013), we get that $a_{uu'} = \mathbf{0}$ if $\|S_{\eta\alpha_S}(a^{(k+1)}_{uu'}-\eta\nabla_{a_{uu'}}Q|_{a^{(k)}_{uu'}})\|_2 \leq \eta\alpha_G$ holds, otherwise $a_{uu'}$ satisfies

$$\left(1+\frac{\eta\alpha_G}{\|a_{uu'}\|_2}\right)a_{uu'} = S_{\eta\alpha_S}(a^{(k)}_{uu'}-\eta\nabla_{a_{uu'}}Q|_{a^{(k)}_{uu'}}), \tag{5}$$

where $S_\alpha(z) = sign(z)(|z|-\alpha)_+$ achieves soft-thresholding for each element of input. Taking the norm on both sides, $\|a_{uu'}\|_2$ can be replaced by

$$(\|S_{\eta\alpha_S}(a^{(k)}_{uu'}-\eta\nabla_{a_{uu'}}Q|_{a^{(k)}_{uu'}})\|_2 - t\eta\alpha_G)_+. \tag{6}$$

Replacing the $\|a_{uu'}\|_2$ in (5) with (6), we obtain the generalized gradient step:

$$a_{uu'}^{(k+1)} = \left(1 - \frac{\eta\alpha_G}{\|S_{\eta\alpha_S}(a_{uu'}^{(k+1)} - \eta\nabla_{a_{uu'}}Q|_{a_{uu'}^{(k)}})\|_2}\right)_+ \times S_{\eta\alpha_S}(a_{uu'}^{(k+1)} - \eta\nabla_{a_{uu'}}Q|_{a_{uu'}^{(k)}}) \quad (7)$$

1.3. Details of Basis Function Selection

In our model, the intensity function of Hawkes process over all dimensions is:

$$\begin{aligned}\lambda(t) &= \sum_{u=1}^{U}\lambda_u(t)\\ &= \sum_{u=1}^{U}\left(\mu_u + \sum\nolimits_{u'=1}^{U}\int_0^t \phi_{uu'}(s)dN_{u'}(t-s)\right)\\ &= \sum_{u=1}^{U}\mu_u + \sum_{u=1}^{U}\sum_{t_i<t}\phi_{uu_i}(t-t_i)\\ &= \sum_{u=1}^{U}\mu_u + \sum_{u=1}^{U}\sum_{t_i<t}\sum_{m=1}^{M}a_{uu_i}^m\kappa_m(t-t_i).\end{aligned} \quad (8)$$

Applying Fourier transform, we have

$$\begin{aligned}\hat{\lambda}(\omega) =& \sum_{u=1}^{U}\mu_u\sqrt{2\pi}\delta(\omega)\\ &+ \sum_{u=1}^{U}\sum_{t_i<t}\sum_{m=1}^{M}a_{uu_i}^m e^{-j\omega t_i}\hat{\kappa}_m(\omega).\end{aligned} \quad (9)$$

In other words, the spectral of $\lambda(t)$ is the weighted sum of those of basis functions. Therefore, the cut-off frequency of basis function is bounded by that of intensity function.

As we show in our paper, given training sequences $\mathcal{S} = \{s_c\}_{c=1}^C$, , where $s_c = \{(t_i^c, u_i^c)\}_{i=1}^{N_c}$, we can estimate $\lambda(t)$ empirically via a Gaussian-based kernel density estimator:

$$\lambda(t) = \sum\nolimits_{c=1}^{C}\sum\nolimits_{i=1}^{N_c}G_h(t-t_i^c). \quad (10)$$

Here t_i^c is the time stamp of the i-th event at the c-th sequence. $G_h(t-t_i^c) = \exp(-\frac{(t-t_i^c)^2}{2h^2})$ is a Gaussian kernel with the bandwidth h.

Because we only care about the selection of basis functions, we just need to estimate the spectral of $\lambda(t)$ rather than compute (10) directly. Specifically, applying Silverman's rule of thumb (Silverman, 1986), we first set optimal $h = (\frac{4\hat{\sigma}^5}{3\sum_c N_c})^{0.2}$, where $\hat{\sigma}$ is the standard deviation of time stamps $\{t_i^c\}$. Applying Fourier transform, we compute an upper bound for the spectral of $\lambda(t)$ as

$$\begin{aligned}|\hat{\lambda}(\omega)| &= \left|\int_{-\infty}^{\infty}\lambda(t)e^{-j\omega t}dt\right|\\ &= \left|\sum_{c=1}^{C}\sum_{i=1}^{N_c}\int_{-\infty}^{\infty}e^{-\frac{(t-t_i^c)^2}{2h^2}}e^{-j\omega t}dt\right|\\ &\le \sum_{c=1}^{C}\sum_{i=1}^{N_c}\left|\int_{-\infty}^{\infty}e^{-\frac{(t-t_i^c)^2}{2h^c}}e^{-j\omega t}dt\right|\\ &= \sum_{c=1}^{C}\sum_{i=1}^{N_c}\left|e^{-j\omega t_i^c}e^{-\frac{\omega^2h^2}{2}}\sqrt{2\pi h^2}\right|\\ &\le \sum_{c=1}^{C}\sum_{i=1}^{N_c}\left|e^{-j\omega t_i^c}\right|\left|e^{-\frac{\omega^2h^2}{2}}\sqrt{2\pi h^2}\right|\\ &= \left(\sum_{c=1}^{C}N_c\sqrt{2\pi h^2}\right)e^{-\frac{\omega^2h^2}{2}}.\end{aligned} \quad (11)$$

Furthermore, we can compute the upper bound of the absolute sum of the spectral higher than ω_0 as

$$\begin{aligned}&\int_{\omega_0}^{\infty}|\hat{\lambda}(\omega)|d\omega\\ \le&\left(\sum_{c=1}^{C}N_c\sqrt{2\pi h^2}\right)\int_{\omega_0}^{\infty}e^{-\frac{\omega^2h^2}{2}}d\omega\\ =&2\pi\left(\sum_{c=1}^{C}N_c\right)\int_{\omega_0}^{\infty}\frac{h}{\sqrt{2\pi}}e^{-\frac{\omega^2h^2}{2}}d\omega\\ =&2\pi\left(\sum_{c=1}^{C}N_c\right)\left(\frac{1}{2}-\int_0^{\omega_0}\frac{h}{\sqrt{2\pi}}e^{-\frac{\omega^2h^2}{2}}d\omega\right)\\ =&2\pi\left(\sum_{c=1}^{C}N_c\right)\left(\frac{1}{2}-\frac{1}{2}\int_{-\omega_0}^{\omega_0}\frac{h}{\sqrt{2\pi}}e^{-\frac{\omega^2h^2}{2}}d\omega\right)\\ =&\pi\left(\sum_{c=1}^{C}N_c\right)\left(1-\frac{1}{\sqrt{2}}\text{erf}(\omega_0h)\right),\end{aligned} \quad (12)$$

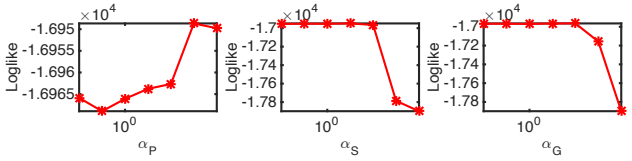
where $\text{erf}(x) = \frac{1}{\sqrt{\pi}}\int_{-x}^{x}e^{-t^2}dt$.

Therefore, give a bound of residual ϵ, we can find an ω_0 guaranteeing $\int_{\omega_0}^{\infty}|\hat{\lambda}(\omega)|d\omega \le \epsilon$, or $\text{erf}(\omega_0h) \ge \sqrt{2} - \frac{\sqrt{2}\epsilon}{\pi\sum_{c=1}^{C}N_c}$. The proposed basis functions $\{\kappa_{\omega_0}(t,t_m)\}_{m=1}^M$ are selected, where ω_0 is the cut-off frequency of basis function and $t_m = \frac{(m-1)T}{M}$, $M = \lceil\frac{T\omega_0}{\pi}\rceil$.

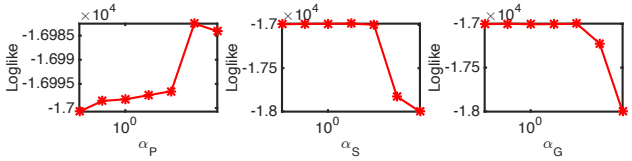
1.4. Configuration of Parameters

With the help of cross validation, we test our algorithm with various parameters in a wide range, where $\alpha_P, \alpha_S, \alpha_G \in [10^{-2}, 10^4]$. According to the measure *Loglike*, we set $\alpha_S = 10$, $\alpha_G = 100$, $\alpha_P = 1000$. The curves of *Loglike* w.r.t. the three parameters are shown in the following

figure. We can find that the learning result is relatively stable when changing the parameters in a wide range.



(a) Sine-like case



(b) Piecewise constant case

Figure 1. The curves of *Loglike* w.r.t. the change of α_P, α_G and α_S are shown. In each subfigure, left: $\alpha_G = 100$, $\alpha_S = 10$, $\alpha_P \in [10^{-2}, 10^4]$; middle: $\alpha_G = 100$, $\alpha_P = 1000$, $\alpha_S \in [10^{-2}, 10^4]$; right: $\alpha_P = 1000$, $\alpha_S = 10$, $\alpha_G \in [10^{-2}, 10^4]$. The number of training sequence is 250.

References

Silverman, Bernard W. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

Simon, Noah, Friedman, Jerome, Hastie, Trevor, and Tibshirani, Robert. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.