2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Hawkes Network Identification with Log-Sparsity Penalty

Xinhui Rong†, Victor Solo†* and Akila J. Seneviratne‡

Abstract— **Point process networks are emerging in wide range of applications, such as finance and social media. Such networks are often modeled by the multivariate Hawkes processes. In this paper, we propose a log-sparsity penalized least squares (log-LS) estimation method for the Hawkes intensity to capture the network dynamics, while eliminating the inactive nodes. We develop a new continuous-time log-LS formulation correcting an error in previous work by finding an underlying true global minimum. We use a cyclic descent + BIC algorithm for efficient optimization. We finally compare various penalties in simulations demonstrating the advantages of log-sparsity.**

I. Introduction

Point process applications are found in neural science [1], high frequency finance [2], seismology [3], etc, where random event times of one or more sources are collected and modeled to characterize the underlying dynamics.

In the multi-source case, the *event times* of one source are considered to be dependent on the history of its own and also other sources' past events. The sources thus form a *point process network* where each source is a *node* of the network, and a source's impact on each node, including itself, forms a *directed link* between nodes. For example, in the network of neurons in an animal's visual cortex, each neuron is considered a node, the neural spike times are the event time signals, and the existence of the excitation influence from one neuron to another, or itself, forms a link.

The multivariate Hawkes process is the most widely used model for a dynamic point process network. It is a linear history dependent point process model. It assumes that each node has a constant *background rate* and is also statistically and positively excited by all the previous events of nodes having directed links towards it. Hawkes model assumes an exponentially decaying excitement impact. Examples of Hawkes modeling are e.g. neural networks [1], [4], social activity networks [5], [6] and currency exchange networks [7].

The identification of the Hawkes model is generally hard. The problems lie in, e.g. (a) the difficulty of identifying multiple nonlinear exponential time constants, and (b) producing sparsity to eliminate unconnected nodes, i.e. to have exact 0 impact weight parameters for unconnected nodes. We will discuss each of them later in this paper. But for (b), it can be achieved by the *sparse signal reconstruction* methods.

Sparse signal reconstruction is to add a sparsity inducing penalty to the original minimization problem for parameter identification, usually least squares or negative log-likelihoods. Common examples for such non quadratic penalties include the ℓ_0 [8], [9], ℓ_q [10], [11] and ℓ_1 [12], [13], [14] norms, Gaussian entropy [15], log penalty [16], [17], [18], [19], [20], etc.

The log penalty is known to produce much greater sparsity than ℓ_1 [16]. Existing solutions to log sparsity problems include weighed-ℓ_1 algorithms [16], [19], [20], iterative thresholding [17], [9], etc. But the result in [17] has a fundamental flaw, which will be corrected in Section IV.

In point process applications, most works use ℓ_1 penalty [6], [21], [22], [23]. Authors of [24] adopted an ℓ_0 penalized least squares + BIC based criterion on binned data. However, all above but the last, assume the nonlinear terms known; they avoid the difficulty (a) mentioned above.

In this paper, we use Hawkes-Laguerre model, a basis expansion of Hawkes, as the authors of [24] did to reduce to a single nonlinear parameter. We formulate and solve the log sparsity problem for Hawkes-Laguerre, and make the following new contributions:

(a) We develop a new continuous-time log-penalized least squares criterion (log-LS) for Hawkes-Laguerre intensity.
(b) We identify the correct threshold for the log-LS problem.*
(c) We develop a cyclic descent algorithm (log-CD) for log-LS.*
(d) We develop a log-CD + BIC method for Hawkes-Laguerre intensity estimation, including selection of the optimal nonlinear time constants and log-penalty parameters.

* (b) and (c) are results for the general log-LS problems and can be applied to non point process applications.

The remainder of the paper is organized as follows. In Section II, we review point process properties. Sections III to IV address contributions (a) and (b), respectively. Section V introduces contributions (c) and (d). In Section VI, we compare various penalties in simulations. Section VII is conclusions.

II. Point process preliminaries

A point process network is a network each of whose nodes carries a point process. We suppose there are M nodes and denote the counting process at the $m^{(th)}$ node as $N_{m,t} = \#$ events at node m up to and including time t. We also denote the network counting process $N_t = \left[N_{1,t}, \cdots, N_{M,t}\right]^\top$.

• The *conditional intensity function* uniquely characterizes point processes and, for $m^{(th)}$ node, is defined as

$$\lambda_m(t) = \lambda_m(t|N_0^{t^-}) = \lim_{\delta \to 0} \frac{1}{\delta} \Pr[N_{m,t+\delta} - N_{m,t} = 1 | N_0^{t^-}],$$

† are with School of Electrical Engineering and Telecom., University of New South Wales, Sydney, Australia and ‡ is with Woolcock Institute of Medical Research, University of Sydney, Sydney, Australia. * is the corresponding author. (email: v.solo@unsw.edu.au).

where $N_0^{t-} = \{N_{l,u} | l = 1, \cdots, M, 0 < u < t\}$ is the memory/history up to time t. $\lambda_m(t)$ is the probability of an event occurring at node m in a tiny time interval after t, given the history up to t. We denote $\lambda(t) = \left[\lambda_1(t), \cdots, \lambda_M(t)\right]^\top$.

We assume *no-simultaneity*. i.e. only zero or one event can occur in an infinitesimally small time interval.

Note that the history consists of all past event times of all M nodes and also note that the intensity is a stochastic process itself.

• The stochastic intensity obeys [25] the *Martingale increment* property: $\mathrm{E}[\lambda(t)dt - (N_{t+dt} - N_t)|N_0^{t-}] = 0$, This will be used to derive the least squares criterion.

• We model the stochastic intensities with a vector Hawkes model [26] as follows

$$\lambda(t) = c + \int_0^t H(t-u) dN_u, \tag{1}$$

where $c = \left[c_1, \cdots, c_M\right]^\top, c_m > 0$ is the vector of background rates and $H_{M\times M}(u) = [h_{ml}(u)]$ is a matrix of Hawkes impulse responses. Specifying $H(u)$ is a major problem which we overcome by employing a causal basis expansion in terms of Laguerre polynomials

$$h_{ml}(u) = \textstyle\sum_{j=1}^P \alpha_{ml,j} \phi_{m,j}(u) \tag{2}$$

$$\phi_{m,j}(u) = \frac{(\beta_m u)^{j-1}}{(j-1)!} \beta_m e^{-\beta_m u} \tag{3}$$

$$\Rightarrow \int_0^\infty \phi_{m,j}(u) du = 1 \tag{4}$$

$$\Rightarrow \lambda_m(t) = c_m + \textstyle\sum_{l=1}^M \sum_{j=1}^P \alpha_{ml,j} \chi_{ml,j}(t) \tag{5}$$

$$\chi_{ml,j}(t) = \int_0^t \phi_{m,j}(t-u) dN_{l,u} = \textstyle\sum_{T_{l,k}<t} \phi_{m,j}(t - T_{l,k}), \tag{6}$$

where P is the model order and $\chi_{ml,j}$'s are *memory regressors*, $\alpha_{ml,j} > 0$ characterizes the scale of the impact of node l on node m, and $\frac{1}{\beta_m} > 0$ is the time constant at which such impact decays, and $T_{l,k}$'s are the event times at node l. The second last line is a Lebesgue-Stieltjes integral.

Note that if β_m is known, $\lambda_m(t)$ is linear in c_m and $\alpha_{ml,j}$. This makes model fitting for those parameters much simpler.

III. Sparsity Penalized Continuous Time Least Squares

In this section, we develop a new continuous time (CT) log-sparsity penalized least squares criterion (log-LS) for the sparse Hawkes-Laguerre intensity estimation. This will eliminate ('zero out') weak or absent nodal connections helping to capture more accurate dynamics.

We first derive a discrete time (DT) least squares criterion. Then, we get a CT version by taking limits. Then we can add the sparsity penalty.

We collect the α parameters and corresponding regressors into vectors as follows.

$$\begin{aligned} \alpha_m &= \left[\, \alpha_{m1,1} \cdots \alpha_{m1,P} \; \alpha_{m2,1} \cdots \alpha_{mM,P} \,\right]^\top \\ \chi_m(t) &= \left[\, \chi_{m1,1}(t) \cdots \chi_{m1,P}(t) \; \chi_{m2,1}(t) \cdots \chi_{mM,P}(t) \,\right]^\top, \end{aligned}$$

and rewrite (5) in vector form $\lambda_m(t) = \left[\, 1 \; \chi_m^\top(t) \,\right] \left[\begin{smallmatrix} c_m \\ \alpha_m \end{smallmatrix}\right]$. Now discretize the counts in observation period $(0, T]$ into K small bins each of width $\delta = \frac{T}{K}$. Then let $N_{m,i}^\delta$ be the number of events at node m observed in the bin $((i-1)\delta, i\delta]$. This leads to the following DT least squares criterion [24]:

The linear parameters $c_m, \alpha_{ml,j}$ can be estimated by minimizing $J(c, \alpha) = \sum_m J_m(c_m, \alpha_m)$

$$J_m(c_m, \alpha_m) = \textstyle\sum_{i=1}^K \delta \lambda_m(i\delta) \left(\frac{1}{2} \lambda_m(i\delta) \delta - N_{m,i}^\delta \right), \tag{7}$$

for $m = 1, \cdots, M$.

A derivation can be found in [24] using the martingale increment property. It follows a standard least square minimization of the discretized martingale $\lambda_m(i\delta)\delta - N_{m,i}^\delta$ with a term unrelated to parameters dropped.

However, we would like to avoid binning the observed data because information will be lost. Therefore, we obtain a CT criterion by taking limits of the DT criterion. We define the following integrals and sums.

$$\begin{aligned} b_{ml,j}(t) &= \textstyle\int_0^t \chi_{ml,j}(u) du \\ B_{m,(l,j)}^{(l',j')}(t) &= \textstyle\int_0^t \chi_{ml,j}(u) \chi_{ml',j'}(u) du \\ \tilde{B}_{m,(l,j)}^{(l',j')}(t) &= B_{m,(l,j)}^{(l',j')}(t) - \frac{1}{t} b_{ml,j}(t) b_{ml',j'}(t) \\ G_{m,(l,j)}^{(l',j')} &= \frac{1}{\sqrt{\tilde{B}_{m,(l,j)}^{(l,j)}(T) \tilde{B}_{m,(l',j')}^{(l',j')}(T)}} \tilde{B}_{m,(l,j)}^{(l',j')}(T) \\ S_{ml,j} &= \textstyle\sum_{r=1}^{N_{m,T}} \chi_{ml,j}(T_{m,r}) \\ v_{ml,j} &= \frac{1}{\sqrt{B_{m,(l,j)}^{(l,j)}(T)}} \left[S_{ml,j} - \frac{N_{m,T}}{T} b_{ml,j}(T) \right], \end{aligned}$$

where $b_{ml,j}$ is the regressor integral, $B_{m,(l,j)}^{(l',j')}$ is the cross integral, $G_{m,(l,j)}^{(l',j')}$ is the centered and normalized cross integral, $S_{ml,j}$ is the regressor sum and $v_{ml,j}$ is the centered and normalized regressor sum.

We now state the optimization for α_m and c_m in the following result.

Result I. (a) *log-LS*. The continuous-time log-penalized least squares criterion (log-LS) is given

$$J_m(\alpha_m) = \frac{1}{2} \alpha_m^\top G_m \alpha_m - v_m^\top \alpha_m + h_m \textstyle\sum_{l,j} \pi(\alpha_{ml,j}), \tag{8}$$

where $\pi(a) = \ln[(|a| + \gamma)/\gamma]$ is the standard log-sparsity penalty term [16], $0 < \gamma \ll h_m$ are two penalty parameters to be chosen, and

$$\begin{aligned} G_m = G_m^\top &= \left[\, G_{m1,1} \cdots G_{m1,P} \; G_{m2,1} \cdots G_{mM,P} \,\right] \\ G_{ml,j} &= \left[\, G_{m,(l,j)}^{(1,1)} \cdots G_{m,(l,j)}^{(1,P)} \; G_{m,(l,j)}^{(2,1)} \cdots G_{m,(l,j)}^{(M,P)} \,\right]^\top \\ v_m &= \left[v_{m1,1} \quad \cdots \quad v_{m1,P} \quad v_{m2,1} \quad \cdots v_{mM,P} \right]^\top. \end{aligned}$$

(b) Denote by $\alpha_m^* = [\alpha_{m1,1}^* \cdots \alpha_{mM,P}^*]^\top$ the minimizer of (8). Weights $\hat{\alpha}_{ml,j}$ can be estimated as

$$\hat{\alpha}_{ml,j} = \alpha_{ml,j}^* / \sqrt{\tilde{B}_{m,(l,j)}^{(l,j)}(T)}, \tag{9}$$

(c) Background rate $\hat{c}_m$ can be estimated as

$$\hat{c}_m = \frac{1}{T} \left(N_{m,T} - \textstyle\sum_{l=1}^M \sum_{j=1}^P \hat{\alpha}_{ml,j} b_{ml,j}(T) \right). \tag{10}$$

Proof: Due to space limits, we only sketch the derivation. The CT criterion (8) is derived from the DT criterion (7). Dividing (7) by δ, it can be rewritten as

$$\bar{J}_m(c_m, \alpha_m) = \sum_{i=1}^K \frac{1}{2}\left(x_i^\top \begin{bmatrix} c_m \\ \alpha_m \end{bmatrix}\right)^2 - y_i x_i^\top \begin{bmatrix} c_m \\ \alpha_m \end{bmatrix}, \tag{11}$$

where $x_i = \begin{bmatrix}\sqrt{\delta} & \sqrt{\delta}\chi_m^\top(i\delta)\end{bmatrix}^\top$ and $y_i = \frac{N_{m,i}^\delta}{\sqrt{\delta}}$.

Now center and normalize x_i to get $\tilde{x}_i = \frac{x_i - \frac{1}{K}\sum_i x_i}{\|x_i - \frac{1}{K}\sum_i x_i\|_2}$, and center $N_{m,i}^\delta/\sqrt{\delta}$ to get $\tilde{y}_i = y_i - \frac{1}{K\sqrt{\delta}}\sum_i N_{m,i}^\delta = \frac{1}{\sqrt{\delta}}\left(N_{m,i}^\delta - \frac{N_{m,T}}{T}\delta\right)$. Note that the first entry in $\tilde{x}_i$ becomes 0 so c_m disappears.

Denote by $\tilde{J}_m(\alpha_m)$ the new criterion after replacing x_i, y_i by $\tilde{x}_i, \tilde{y}_i$ in $\bar{J}_m$. Now expand $\tilde{J}_m(\alpha_m)$, add the log-sparsity penalty, and use the limits such as

$$\frac{1}{K}\sum_i \chi_{ml,j}(i\delta) = \frac{1}{T}\sum_i \chi_{ml,j}(i\delta)\delta \to \frac{1}{T}\int_0^T \chi_{ml,j}(t)dt$$

$$\sum_i \chi_{ml,j}(i\delta)\chi_{ml',j'}(i\delta)\delta \to \int_0^T \chi_{ml,j}(t)\chi_{ml',j'}(t)dt$$

$$\sum_i \chi_{ml,j}(i\delta)N_{m,i}^\delta \to \sum_{r=1}^{N_{m,T}} \chi_{ml,j}(T_r), \quad \text{as } \delta \to 0,$$

to get the quoted log-LS in (8).

Finally recovering the centering and normalization gives $\hat{\alpha}_m$ in (9) and $\hat{c}_m$ in (10).

■

Remarks.

(a) The criteria $J_m, m = 1, \cdots, M$ can be optimized separately.

(b) G_m can be shown to be positive definite. G_m has unit diagonal entries and off-diagonal entries of magnitude < 1 due to normalization.

(c) Authors of [24] used DT log-penalized least squares by adding the penalty term directly to (7). We use the CT criterion to preserve all information.

(d) We have centered and normalized the regressors and centered the observed counting increment, so that the background rate c_m disappears from the criterion and avoids being penalized - since they must be $\neq 0$ for the Hawkes model to make sense. Another reason for normalization is for use in the cyclic descent algorithm to be developed later. c_m is estimated by undoing the centering and normalization.

(e) Criterion (8) is different from the CT criterion used in [21] because they do not center or normalize and thus the background rate c_m gets penalized causing problems if it is zeroed.

We now turn to solving log-LS in (8).

IV. Log-Sparsity

In order to solve the log-LS in (8), we first need to discuss the solution of the following scalar log-LS criterion

$$\min_f J_s(f; y) = \frac{1}{2}(y - f)^2 + h\pi(f). \tag{12}$$

We show below that the global minimizer f_* is found by the following thresholding

$$f_* = \Psi(y) = \begin{cases} 0, & \text{if } |y| \le \tau_* \\ f_e, & \text{if } |y| > \tau_* \end{cases} \tag{13}$$
$$f_e = \text{sgn}(y)(\sqrt{h}\psi(w) - \gamma),$$

where $\psi(w) = w + \sqrt{w^2 - 1}, w = \frac{|y|+\gamma}{2\sqrt{h}}$, f_e is the local minimizer when $f \neq 0$, optimized by calculus, and τ^* is the threshold derived below.

For the two most common penalties of ℓ_1, ℓ_0, it is well known that scalar penalized least squares problem is solved by thresholding with a single threshold. However for the ℓ_q penalty it turns out the minimization involves two thresholds [27], [10] - roughly corresponding to a local minimum and a global minimum. It turns out that this is also the case for the log-sparsity penalty. We call these thresholds:

i) Euler threshold - the local threshold

ii) Sparsity threshold - the global threshold.

A. Euler Threshold

In [17], the authors find that the log-LS problem, for $f \neq 0$ exists, when $|y| > \tau_e = 2\sqrt{h} - \gamma$ (their formula is equivalent but differs in scale because we have a factor of $\frac{1}{2}$ in J_s). We call τ_e the *Euler threshold*.

The authors then mistakenly claim, without proof, that the Euler threshold also provides the global minimum. We now proceed to find the global minimizer.

B. Sparsity Threshold

The global minimizer is as follows.

Result II. Sparsity Threshold. The scalar log-LS (12) is globally minimized by (13) where the sparsity threshold τ_* is given by

$$\tau_* = \min_{f>0} U(f), \quad U(f) = \frac{1}{2}f + \frac{h\ln[(f+\gamma)/\gamma]}{f}, \tag{14}$$

and $\tau_e < \tau_*$.

Proof: We minimize $J_s(f)$ by comparing $J_s(0)$ with $J_s(f), f \neq 0$.

We have $J_s(0) = \frac{1}{2}y^2$. So 0 is preferred if

$$\frac{1}{2}y^2 < \frac{1}{2}y^2 - yf + \frac{1}{2}f^2 + h\ln[(|f| + \gamma)/\gamma]$$
$$\Rightarrow yf < \frac{1}{2}f^2 + h\ln[(|f| + \gamma)/\gamma].$$

This inequality is non-trivial iff $yf > 0$, i.e. iff $\text{sgn}(y) = \text{sgn}(f)$ whereupon it is equivalent to

$$|y| < \frac{1}{2}|f| + \frac{h\ln[(|f| + \gamma)/\gamma]}{|f|}.$$

This leads us to introduce the sparsity threshold τ_* as in (14). Then, 0 is preferred if $|y| < \tau_*$.

We denote the minimizer of $U(f)$ by f_0. We now show $\tau_e < \tau_*$.

$U(f)$ is convex on $(0,\infty)$, so differentiating $U(f)$ for a minimum, we find

$$\begin{aligned} 0 &= \tfrac{1}{2} + \tfrac{h}{(f_0+\gamma)f_0} - \tfrac{h\ln[(f_0+\gamma)/\gamma]}{f_0^2} \\ \Rightarrow\quad \tfrac{1}{2}f_0 + \tfrac{h}{f_0+\gamma} &= \tfrac{h\ln[(f_0+\gamma)/\gamma]}{f_0^2} \\ \Rightarrow\quad f_0 + \tfrac{h}{f_0+\gamma} &= U(f_0) = \tau_*. \end{aligned}$$

Now consider for $f > 0$, the following function

$$\begin{aligned} f + \tfrac{h}{f+\gamma} &= [(f+\gamma) + \tfrac{h}{f+\gamma}] - \gamma \\ &= \sqrt{h}\left[\tfrac{f+\gamma}{\sqrt{h}} + \tfrac{\sqrt{h}}{f+\gamma}\right] - \gamma \\ &\geq 2\sqrt{h} - \gamma = \tau_e, \end{aligned}$$

since the term in square brackets has the form $x + \frac{1}{x}$ and is minimized at $x = 1$ with minimum $= 2$. Then we get equality iff $f + \gamma = \sqrt{h}$ and the result follows. ∎

Remark. $U(f)$ is convex on $f > 0$, so we can effectively find $\tau_* = U(f_0)$ by numerically finding the solution f_0 to

$$U'(f_0) = \frac{1}{2} + \frac{h}{\gamma^2}\frac{\frac{f_0/\gamma}{f_0/\gamma+1} - \ln(f_0/\gamma+1)}{(f_0/\gamma)^2} = 0.$$

V. Log-cyclic descent (log-CD)

Using the results derived in Section IV, we can now develop a cyclic descent algorithm, which we call log-CD for solving log-LS problem in (8). We state it in Algorithm I.

We can thus state the complete algorithm for sparse Hawkes intensity estimation by the log-CD method, taking into account the optimization for the nonlinear parameter β_m. We select β_m, together with the log-penalty parameter h_m, by a BIC criterion. This is summarized in Algorithm II.

Algorithm I. Cyclic descent for log-LS (log-CD).

To optimize $J_m(\alpha_m)$ in (8), with current iterate $\alpha_m^k = [\alpha_{ml,j}^k]_{l=1,\cdots,M,j=1,\cdots,P}$, we update α_m^{k+1} as followings:

1) Start with an element $\alpha_{ml,j}^k$ in α_m^k.
2) Calculate $z_{ml,j}^k = v_{ml,j} - \sum_{j'\neq j,l'\neq l}\alpha_{ml',j'}^k G_{m,(l,j)}^{(l',j')}$.
3) Solve the scalar log-LS $J_s(\alpha_{ml,j}^{k+1}; z_{ml,j}^k)$ to get $\alpha_{ml,j}^{k+1} = \Psi(z_{ml,j}^k)$ defined in (13), with the threshold τ_* defined in (14).
4) Replace $\alpha_{ml,j}^k$ with $\alpha_{ml,j}^{k+1}$ in α_m^k.
5) If all elements in α_m^k are not exhausted, go to 2).
6) If the terminating condition is met, take the estimate $\hat{\alpha}_m = \alpha_m^{k+1}$ and terminate. If not, let $k \leftarrow k+1$ and go to 1).

Remarks.

(a) Step 4) is essential to the convergence of log-CD.
(b) The terminating condition can be the relative decrement in J_m falling below a tiny number: $\frac{J_m(\alpha_m^k) - J(\alpha_m^{k+1})}{|J(\alpha_m^k)|} < \epsilon$.
(c) Algorithm I can be easily amended to solve the general log-LS $\min_\alpha \frac{1}{2}\|y - X\alpha\|_2^2 + h\sum_i \pi(\alpha_i)$.
(d) Convergence analysis is a challenging issue which we hope to treat elsewhere.

Note that for now we have assumed β_m known. We now treat the nonlinear parameter β_m and also find the optimal tuning parameter h_m by combining the above log-CD with BIC method. We summarize the full algorithm as follows.

Algorithm II. log-CD + BIC algorithm.

The sparse Hawkes-Laguerre intensity can be estimated by separately optimizing each J_m. For each m, do the following:

1) Choose a proper grid of possible values of $\beta_m^1, \cdots, \beta_m^{K_b}$, and $h_m^1, \cdots, h_m^{K_h}$.
2) For each β_m^p, calculate the corresponding $G_{m,(l,j)}^{(l',j')}$ and $v_{ml,j}$.
3) Repeat on each pair of (β_m^p, h_m^q), the log-LS optimization and solve log-LS in (8) by Algorithm I
4) Calculate $\hat{\alpha}_m$ by (9) and $\hat{c}_m$ by (10).
5) Calculate the corresponding BIC: [28] $BIC_m^{(p,q)} = -2\mathcal{L}_m^{(p,q)} + d\ln N_{m,T}$, where $d =$ no. of active parameters $= 2+$ no. of non-zero $\alpha_{ml,j}$'s and the likelihood [29], [25]

$$\begin{aligned} &\mathcal{L}_m^{(p,q)} \triangleq \mathcal{L}_m(\hat{\alpha}_m, \hat{c}_m | \beta_m^p, h_m^q) \\ &= \textstyle\sum_{r=1}^{N_{m,T}} \ln(\hat{c}_m + \hat{\alpha}_m^\top \chi_m(T_{m,r})) - \hat{c}_m T - \hat{\alpha}_m^\top b_m(T), \end{aligned}$$

$b_m(T) = \left[\, b_{m1,1}(T) \;\cdots\; b_{m1,P}(T)\; b_{m2,1}(T) \;\cdots\; b_{mM,P}(T) \,\right]$.

6) Find the pair (β_m^p, h_m^q) that has the smallest BIC and find the corresponding estimates $\hat{\alpha}_m, \hat{c}_m$ and $\hat{\beta}_m = \beta_m^p$.

Remarks.

(a) The rule-of-thumb value for h_m is $h_0 = \sqrt{2\ln(MP)}$.
(b) $G_{m,(l,j)}^{(l',j')}$ and $v_{ml,j}$ are pre-calculated for each β_m^p. Some recursive calculation details can be found in [28].
(c) Only with Hawkes-Laguerre model is it possible to reduce the nonlinear parameter estimation (of β_m) to a simple grid search. Other models may require estimating up to MP nonlinear parameters.

VI. Simulations

Here we run simulations to compare log sparsity penalization with ℓ_0 and ℓ_1 penalties on Hawkes-Laguerre intensity estimation and reveal a number of interesting features.

We simulate $P =$ 3-rd order multivariate Hawkes-Laguerre point processes with $M = 3$ nodes. We use the following node background rates and time constants

$$\begin{aligned} \begin{bmatrix} c_1 & c_2 & c_3 \end{bmatrix} &= \begin{bmatrix} 0.2 & 0.5 & 1 \end{bmatrix} \\ \begin{bmatrix} \frac{1}{\beta_1} & \frac{1}{\beta_2} & \frac{1}{\beta_3} \end{bmatrix} &= \begin{bmatrix} 0.2 & 0.33 & 0.1 \end{bmatrix}. \end{aligned}$$

We set sparse memory regressor weights at

$$\tilde{\alpha}_{23} = \tilde{\alpha}_{31} = \tilde{\alpha}_{33} = 0,$$

where $\tilde{\alpha}_{ml} \triangleq \sum_{j=1}^P \alpha_{ml,j}$. The active weights are chosen such that

$$\begin{aligned} &\begin{bmatrix} \tilde{\alpha}_{11} & \tilde{\alpha}_{12} & \tilde{\alpha}_{13} & \tilde{\alpha}_{21} & \tilde{\alpha}_{22} & \tilde{\alpha}_{32} \end{bmatrix} \\ &= \begin{bmatrix} 0.5 & 0.7 & 0.2 & 0.4 & 0.35 & 0.2 \end{bmatrix}. \end{aligned}$$

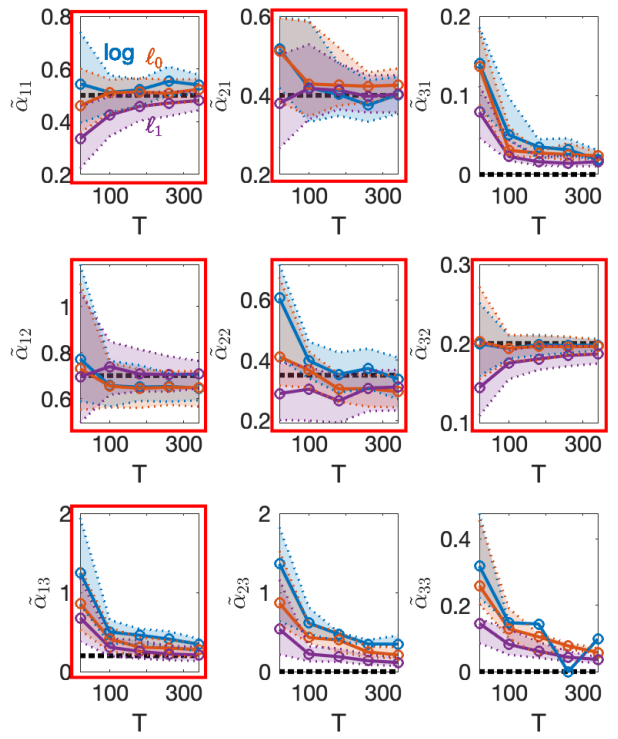

Fig. 1. Quantiles of estimated weights against T. Blue: log, orange: ℓ_0, purple: ℓ_1, black: true. Thick solid lines: median, thin dotted lines: upper and lower 25% quantiles. Red boxes: active weights.

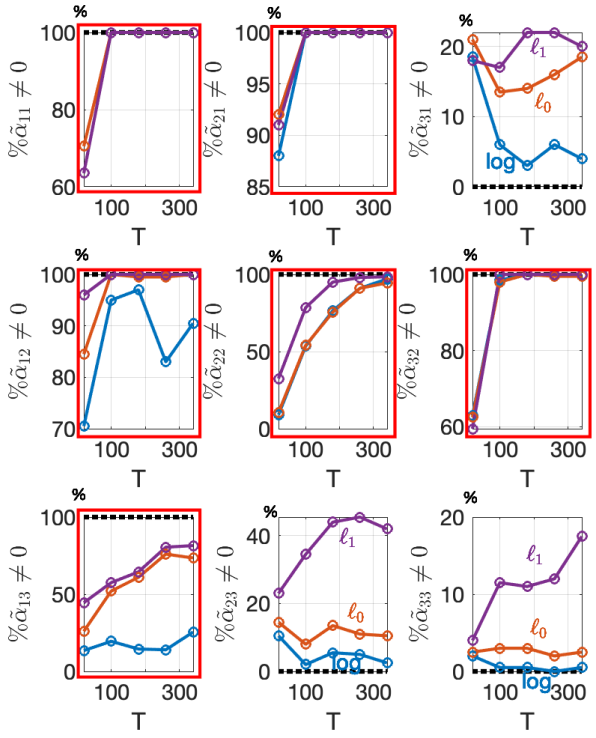

Fig. 2. Percentage of nonzero $\tilde{\alpha}_{ml}$'s against T. Blue: log, orange: ℓ_0, purple: ℓ_1, black: true. Red boxes: active weights.

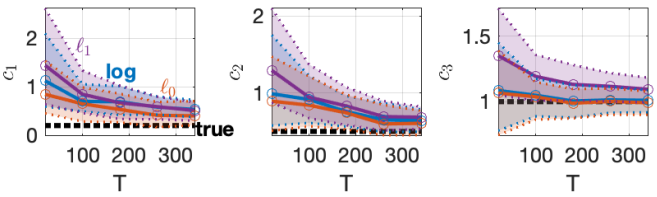

Fig. 3. Quantiles of estimated $\hat{c}_m$ against T. Blue: log, orange: ℓ_0, purple: ℓ_1, black: true. Median: thick solid lines, upper and lower 25% quantiles: thin dotted lines.

Five observation times $T := 20, 100, 180, 260, 340$ are considered. We use the thinning algorithm [30], [31] to simulate $n = 200$ point processes for each T.

To find the optimal $\frac{1}{\beta_m}$ and h_m, we vary $K_b = 15$ time constants $\frac{1}{\beta_m}$ in $[0.067, 1]$ and $K_h = 15$ h_m's in $[0.1, 1]h_0, h_0 = \sqrt{2\ln(MP)}$. We fix $\gamma = 5 \times 10^{-4}$.

We compare (1) the log-sparsity penalty with two other penalties: (2) ℓ_0, (3) ℓ_1. In each case, the same penalty parameter grid is used and optimization is by cyclic descent. All terminating conditions are set to $\frac{J_m(\alpha_m^k) - J(\alpha_m^{k+1})}{|J(\alpha_m^k)|} < \epsilon, \epsilon = 10^{-5}$.

One would also like to compare with results based on using the Euler threshold. But it fails catastrophically. 23.3% of the runs using Euler thresholding terminated because the criterion increased!

Now we discuss the results from the above three penalties.

A. Memory regressor weights

We first show the estimated memory regressor weights $\hat{\alpha}_{ml,j}$. It is impossible to plot all PM^2 $\alpha_{ml,j}$'s, so we plot their sums $\tilde{\alpha}_{ml}$. We plot separately in two figures.

In Fig. 1, we plot the quantiles of weight estimates that are estimated active, i.e. for $\tilde{\alpha}_{ml} \neq 0$. The thick solid lines are the medians of each method, taken from the $n = 200$ repeats at each T. The colored areas between the thin dotted lines indicate the $25\% - 75\%$ quantiles. The back dashed horizontal lines are the true values. The red box frames indicate truly active weights and the others are inactive ones.

In Fig. 2, we plot the percentage of weight estimates that are eliminated by sparsity penalty. i.e. $\%\{\tilde{\alpha}_{ml} = 0\}$. The true value for active weights is 100% and that of inactive ones is 0%, indicated by black dashed line.

We have the following mixed performance observations:

(a) It seems that the log-sparsity is the most severe sparsity penalty. In Fig. 2, for inactive weights (plots without the red box), log-sparsity gives the best sparsity followed by ℓ_0 and then ℓ_1, confirming that log penalty provides better sparsity than ℓ_1. In Fig. 1, log-sparsity estimates are larger, (e.g. $\tilde{\alpha}_{31}, \tilde{\alpha}_{13}, \tilde{\alpha}_{23}, \tilde{\alpha}_{33}$). It means that the log-sparsity tends to eliminate small estimates better than other penalties.

(b) The log-sparsity penalty tends to eliminate some small valued active weights, e.g. for $\tilde{\alpha}_{13}$ in Fig. 2.

(c) For the statistics of nonzero estimates, all three methods perform similarly. It seems that the statistics of nonzero estimates are not heavily affected by each penalty term.

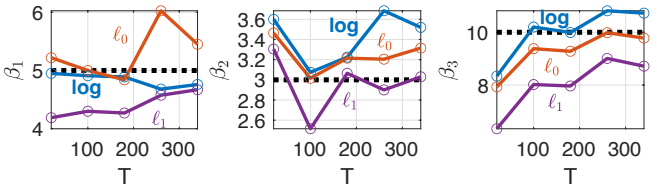

Fig. 4. Means of estimated $\hat{\beta}_m$ against T. Blue: log, orange: ℓ_0, purple: ℓ_1, black: true.

B. Background rates and time constants

The quantiles of estimated $\hat{c}_m$ and the mean values of estimated $\hat{\beta}_m$ are plotted in Fig. 3 and Fig. 4, respectively. We plot mean values of $\hat{\beta}_m$ instead of medians because they are estimated by grid search.

All 3 methods perform similarly in $\hat{c}_1$ and $\hat{c}_2$. But for $\hat{c}_3$, ℓ_1 has larger estimate of $\hat{c}_3$ because it fails to eliminate inactive memory regressors as shown in Fig. 2.

Log-sparsity estimates $\hat{\beta}_1$ best. ℓ_1 estimates $\hat{\beta}_2$ best. ℓ_0 and log-sparsity equally estimate $\hat{\beta}_3$ best.

C. Computational times and negative estimates

We now compare some other performance aspects. First, the average computational times per 100 repeats are: Log-CD: 0.92s + 3.13s, ℓ_0: 1.71s, ℓ_1: 1.99s. Additional 3.13s per 100 repeats for log-CD is used to solve the threshold τ_* in (14), which does not increase with model order P or observation time T.

Second, we compare the percentage of negative estimates as they need to be positive. For log-CD and ℓ_1, all BIC selected estimates are positive. However, for ℓ_0, increasing number of negative estimates are found as T grows. We detect 0.39% negative $\hat{\alpha}_{ml,j}$'s when $T = 20$ but the number increased to 5.7% when $T = 340$.

VII. Conclusions

In this paper, we develop a new continuous-time log-sparsity penalized least squares criterion (log-LS) for Hawkes-Laguerre intensity estimation.

We discover, for the first time, the correct globally minimizing threshold for the general log-LS problem, correcting a major flaw in previous work.

We developed a cyclic descent algorithm for the log-LS problem, which we call log-CD. We selected simultaneously the optimal nonlinear time constant parameters and the penalty weights with BIC. This tuning parameter selection problem is poorly treated or ignored in other literature.

We found interesting features in the comparative simulations. Firstly the erroneous Euler threshold led to complete failure to converge in about 25% of cases. Secondly, the ℓ_0 sparsity penalty produced an increasing number of negative coefficients as the observation period increased. Log-sparsity did not have this problem. As expected, the ℓ_1 penalty led to less sparsity than the log-penalty. Together these results suggest the log-sparsity penalty should enjoy wider use.

In the future, we will tackle likelihood approaches.

References

[1] W. Bialek et. al., *Spikes: Exploring the Neural Code*, MIT press, 1997.

[2] N. Hautsch, *Modelling irregularly spaced financial data: theory and practice of dynamic duration models*, vol. 539, Springer Science & Business Media, 2011.

[3] F. Schoenberg and K. Tranbarger, "Description of earthquake after-shock sequences using prototype point patterns," *Environmetrics*, vol. 19, pp. 271–286, 2008.

[4] P. Dayan and L. F. Abbott, *Theoretical neuroscience: computational and mathematical modeling of neural systems*, MIT press, 2005.

[5] M. Lukasik et. al., "Hawkes processes for continuous time sequence classification: an application to rumour stance classification in twitter," in *Proc. Assoc. Comp. Linguistics, Vol.2*, 2016, pp. 393–398.

[6] K. Zhou, H. Zha, and L. Song, "Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes," in *Artif. Intell. and Stat.* 2013, pp. 641–649, PMLR.

[7] S. A. Pasha and V. Solo, "Distributed topology identification for point process dynamic networks," in *IEEE ICASSP*, 2015, pp. 3681–3685.

[8] A. J Seneviratne and V. Solo, "On vector l0 penalized multivariate regression," in *IEEE ICASSP*, 2012, pp. 3613–3616.

[9] T. Blumensath and M. E. Davies, "Iterative thresholding for sparse approximations," *Journal of Fourier analysis and Applications*, vol. 14, no. 5, pp. 629–654, 2008.

[10] G. Marjanovic and V. Solo, "l_q sparsity penalized linear regression with cyclic descent," *IEEE Trans. Signal Process.*, vol. 62, no. 6, pp. 1464–1475, 2014.

[11] B. D. Rao et. al., "Subset selection in noise based on diversity measure minimization," *IEEE Trans. Signal. Proc.*, vol. 51, pp. 760–770, 2003.

[12] S. Alliney and SA Ruzinsky, "An algorithm for the minimization of mixed l1 and l2 norms with application to bayesian estimation," *IEEE Trans. Signal Process.*, vol. 42, pp. 618–627, 1994.

[13] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. R. Stat. Soc. Ser. A. Stat. Soc.: Series B*, vol. 58, pp. 267–288, 1996.

[14] B. Efron et. al., "Least angle regression," *Ann. Stat.*, vol. 32, no. 2, pp. 407–499, 2004.

[15] B. D Rao and K. Kreutz-Delgado, "An affine scaling methodology for best basis selection," *IEEE Trans. Signal Process.*, vol. 47, no. 1, pp. 187–200, 1999.

[16] E. J. Candes et. al., "Enhancing sparsity by reweighted l1 minimization," *J. Fourier Anal. Appl.*, vol. 14, no. 5, pp. 877–905, 2008.

[17] D. Malioutov and A. Aravkin, "Iterative log thresholding," in *IEEE ICASSP*, 2014, pp. 7198–7202.

[18] J. Weston et. al., "Use of the zero norm with linear models and kernel methods," *J. Mach. Learn. Res.*, vol. 3, pp. 1439–1461, 2003.

[19] D. Wipf and S. Nagarajan, "Solving sparse linear inverse problems: Analysis of reweighted ℓ_1 and ℓ_2 methods," in *APARS*, 2009.

[20] M. A. Khajehnejad et. al., "Analyzing weighted ℓ_1 minimization for sparse recovery with nonuniform sparse models," *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 1985–2001, 2011.

[21] E. Bacry et. al., "Sparse and low-rank multivariate hawkes processes," *Journal of Machine Learning Research*, vol. 21, pp. 1–32, 2020.

[22] H. Xu, F. Mehrdad, and H. Zha, "Learning granger causality for hawkes processes," 2016.

[23] X. Tang and L. Li, "Multivariate temporal point process regression," *J. Am. Stat. Assoc.*, pp. 1–16, 2021.

[24] S. A. Pasha and V. Solo, "Sparse topology identification for point process networks," in *IEEE ICASSP*, 2018, pp. 2196–2200.

[25] D. Daley and David Vere-Jones, *An introduction to the theory of point processes. 2nd ed*, vol. 1, Springer, 2003.

[26] A. G. Hawkes, "Spectra of some self-exciting and mutually exciting point processes," *Biometrika*, vol. 58, no. 1, pp. 83–90, 1971.

[27] G. Marjanovic and V. Solo, "On l_q optimization and matrix completion," *IEEE Trans. Signal Process.*, vol. 60, no. 11, pp. 5714–5724, 2012.

[28] B. I. Godoy et. al., "Truncated hawkes point process modeling: System theory and system identification," *Automatica*, vol. 113, pp. 108733, 2020.

[29] D. Snyder and M. Miller, *Random Point Processes in Time and Space*, Springer-Varlag, New York, 1991.

[30] P. A. W Lewis and G. S. Shedler, "Simulation of nonhomogeneous poisson processes by thinning," *Nav. Res. Logist. Q.*, vol. 26, no. 3, pp. 403–413, 1979.

[31] Y. Ogata, "On lewis' simulation method for point processes," *IEEE Trans. Inf. Theory*, vol. 27, no. 1, pp. 23–31, 1981.