A Nonparametric EM algorithm for Multiscale Hawkes Processes

Erik Lewis George Mohler

May 9, 2011

Abstract

Estimating the conditional intensity of a self-exciting point process is particularly challenging when both exogenous and endogenous effects play a role in clustering. We propose maximum penalized likelihood estimation as a method for simultaneously estimating the background rate and the triggering density of Hawkes process intensities that vary over multiple time scales. We compare the accuracy of the algorithm with the recently introduced Model Independent Stochastic Declustering (MISD) algorithm and then use the model to examine self-excitation in Iraq IED event patterns.

Keywords: Maximum Penalized Likelihood Estimation, Point Processes, Optimization

1 Introduction

Branching or "self-exciting" point processes are useful in describing chains of events that occur randomly in time and space, where the occurrence of an event increases the likelihood of future events. Self-exciting point processes arise in the modeling of earthquake aftershocks (Ogata, 1988; Ogata, 1998) and more recently in social science applications such as crime (Mohler, Short, Brantingham, Schoenberg, & Tita, 2011), financial contagion (Errais, Giesecke, & Goldberg, 2010), and "viral" processes on the internet (Crane & Sornette, 2008). The Hawkes process (Hawkes, 1971) serves as an archetypal example, where the conditional intensity (rate of events) is modeled by

$$\lambda(t) = \mu(t) + \sum_{t>t_i} g(t - t_i). \tag{1}$$

Here the intensity $\lambda(t)$ depends on past events t_i through the triggering kernel $g(t)$, which determines the elevated risk of events following event t_i. While it is common to assume the background rate,

$\mu(t)$, is constant in time, we will consider the more general case of varying $\mu(t)$.

Self-exciting point processes have been extensively studied in the context of earthquake aftershocks and parametric forms for the triggering kernel are well established (Ogata, 1998). Furthermore, it is widely accepted that on short to intermediate time scales the background rate is approximately constant (Marsan & Lengline, 2008). However, in relatively new fields where self-exciting point processes may explain event patterns, the form of g is not known a priori and it cannot be assumed that the background rate is stationary (Mohler, Short, et al., 2011; Lewis, Mohler, Brantingham, & Bertozzi, 2011). In this situation nonparametric methods are desirable for both estimating the general form of the triggering kernel and determining the extent to which exogenous effects contribute to clustering through the background rate.

In (Marsan & Lengline, 2008), an Expectation-Maximization (EM) algorithm, called "Model Independent Stochastic Declustering" (MISD), is introduced for the nonparametric estimation of self-exciting point processes with time-homogeneous background rate. The method is motivated by stochastic declustering (Zhuang, Ogata, & Vere-Jones, 2002) and makes use of the fact that given the branching structure of a self-exciting point process, estimation of the triggering kernel and background intensity reduce to decoupled density estimation problems (Marsan & Lengliné, 2010; Mohler, Short, et al., 2011). MISD in this context is analogous to the histogram estimator in density estimation.

The efficacy of the MISD algorithm is studied in (Sornette & Utkin, 2009), where the authors find that the ability of MISD to recover key parameters such as the branching ratio (integral of g) depends on the values of the model parameters. In particular, the authors remark that the accuracy of MISD improves as the timescale over which the triggering kernel decays shortens. A similar issue arises in the context of EM algorithms for mixture models, where it has been established that EM algorithms perform better when the mixtures are well separated (Xu & Jordan, 1996). The presence of this issue in both problems follows from Equation (1), which can be viewed as a mixture model with mixtures $\mu(t)$, $g(t-t_1)$,...,$g(t-t_n)$. We revisit this issue in greater detail in Section 2. In particular, we show that parametric EM in the context of Hawkes processes is equivalent to projected gradient, as is the case for EM algorithms for mixture models. Furthermore, the convergence of the EM algorithm, as well as the variance of parameter estimates, depend on how well the time scales of μ and g are separated.

In this paper we introduce a Maximum Penalized Likelihood Estimation (MPLE) approach for the nonparametric estimation of Hawkes processes of the form (1). The method is capable of estimating $\mu(t)$ and $g(t)$ simultaneously, without prior knowledge of their form. Analogous to MPLE in the context of density estimation, the added regularity of the estimates allows for higher accuracy and/or lower sample sizes in comparison to MISD. The layout of this paper is as follows. In Section 2, we investigate why the accuracy of EM algorithms for Hawkes processes depends on the relative

timescales of the background rate and triggering kernel. In Section 3, we review the MISD algorithm and then introduce the MPLE approach to the estimation of Hawkes processes. In Section 4 we compare the accuracy of MPLE to MISD for stationary background rate μ and investigate the ability of MPLE to estimate Hawkes processes of the form (1) when μ is allowed to vary in time. In Section 5, we illustrate the implementation of the MPLE algorithm using a data set of violent civilian deaths occurring in the Iraqi conflict between 2003 and 2007. The algorithm captures both a general escalation of violence occurring over the time scale of 1000 days and self-excitation following violence, occurring over a shorter time scale of a 10 days.

2 Time scale separation and EM convergence

We consider a Hawkes Process of the form (1) with exponential triggering kernel $g(t) = \alpha\omega e^{-\omega t}$ on the time interval $[0, T]$ and the EM algorithm proposed in (Veen & Schoenberg, 2008) for the estimation of the parameters $\Theta = (\mu, \alpha, \omega)$ from a point sample $\{t_i\}_{i=1}^n$ generated according to the intensity given by (1). The log-likelihood function for (1) is given by,

$$l(\Theta) = \sum_{i=1}^{n} \log(\lambda(t_i)) - \int_0^T \lambda(t)dt \tag{2}$$

where the integral approximation, $\mu T + \alpha \sum_{i=1}^{n}(1 - e^{-w(T-t_i)}) \approx \mu T + \alpha n$, can be used when $\omega^{-1} << T$. Let

$$A(i) = \sum_{j=1}^{i-1} e^{-\omega(t_i - t_j)}$$

and

$$B(i) = \sum_{j=1}^{i-1} (t_i - t_j) e^{-\omega(t_i - t_j)}.$$

Then the derivatives of the log-likelihood with respect to the parameters are given by

$$\frac{\partial l}{\partial \mu} = \sum_{i=1}^{n} \frac{1}{\lambda(t_i)} - T \tag{3}$$

$$\frac{\partial l}{\partial \alpha} = \sum_{i=1}^{n} \frac{\omega A(i)}{\lambda(t_i)} - n \tag{4}$$

and

$$\frac{\partial l}{\partial \omega} = \sum_{i=1}^{n} \frac{\alpha}{\lambda(t_i)} (A(i) - \omega B(i)). \tag{5}$$

The EM algorithm for the estimation of the parameters is as follows. Let p_{ij} be the probability that event j triggers event i.. Starting with a guess Θ^0 for the parameters, iterate the following until convergence is reached:

Expectation step:

$$p_{ij}^k = \frac{\alpha^k \omega^k e^{-\omega^k (t_i - t_j)}}{\mu^k + \sum_{j=1}^{i-1} \alpha^k \omega^k e^{-\omega^k (t_i - t_j)}} \tag{6}$$

$$p_{ii}^k = \frac{\mu^k}{\mu^k + \sum_{j=1}^{i-1} \alpha^k \omega^k e^{-\omega^k (t_i - t_j)}} \tag{7}$$

and

Maximization step:

$$\mu^{k+1} = \frac{\sum_{i=1}^{n} p_{ii}^k}{T} \tag{8}$$

$$\alpha^{k+1} = \frac{\sum_{i>j} p_{ij}^k}{n} \tag{9}$$

$$\omega^{k+1} = \frac{\sum_{i>j} p_{ij}^k}{\sum_{i>j} (t_i - t_j) p_{ij}^k}. \tag{10}$$

It is easy to show that each EM iteration satisfies the following:

$$\mu^{k+1} - \mu^k = \frac{\mu^k}{T} \frac{\partial l}{\partial \mu} \tag{11}$$

$$\alpha^{k+1} - \alpha^k = \frac{\alpha^k}{n} \frac{\partial l}{\partial \alpha} \tag{12}$$

and

$$\omega^{k+1} - \omega^k = \frac{\omega^k}{\sum_{i>j} (t_i - t_j) p_{ij}^k} \frac{\partial l}{\partial \omega} \tag{13}$$

where the partial derivatives are evaluated at Θ^k. Thus the EM algorithm is equivalent to a projected gradient ascent algorithm. Such a result has been established for the EM algorithm in the context of Gaussian mixtures (Xu & Jordan, 1996), where the authors show that the projection significantly improves rate of convergence (compared to gradient ascent) when the amount of overlap of the mixtures is small.

For compactness we will write (11)-(13) as

$$\Theta^{k+1} - \Theta^k = P(\Theta^k) \nabla l|_{\Theta^k} \tag{14}$$

where $P(\Theta^k)$ is a diagonal matrix with positive entries. Letting Θ^* denote the MLE estimate of (2)

and Taylor expanding the gradient of the log-likelihood in (14) about Θ^* we get,

$$\|\Theta^{k+1} - \Theta^*\| \leq \|I + P(\Theta^k)H(\Theta^*)\|\|\Theta^k - \Theta^*\| \tag{15}$$

where H is the Hessian of the log-likelihood. Thus the convergence rate of the EM algorithm depends on how well the matrix PH is conditioned.

Boundary correction

When the integral in (2) is evaluated exactly, the partial derivatives of the log-likelihood with respect to α and ω become,

$$\frac{\partial l}{\partial \alpha} = \sum_{i=1}^{n} \frac{\omega A(i)}{\lambda(t_i)} - n + \sum_{i=1}^{n} e^{-w(T-t_i)} \tag{16}$$

and

$$\frac{\partial l}{\partial \omega} = \sum_{i=1}^{N} \frac{\alpha}{\lambda(t_i)}(A(i) - \omega B(i)) - \alpha \sum_{i=1}^{n}(T - t_i)e^{-w(T-t_i)}. \tag{17}$$

While the expectation step stays the same, the maximization step for α and ω become,

$$\alpha^{k+1} = \frac{\sum_{i>j} p_{ij}^k}{n - \sum_{i=1}^{n} e^{-w(T-t_i)}} \tag{18}$$

$$\omega^{k+1} = \frac{\sum_{i>j} p_{ij}^k}{\sum_{i>j}(t_i - t_j)p_{ij}^k + \alpha \sum_{i=1}^{n}(T - t_i)e^{-w^{k+1}(T-t_i)}}. \tag{19}$$

Note that the maximization step for ω^{k+1} is implicit. While this nonlinear equation could be solved at each iteration, we instead replace ω^{k+1} on the right side by ω^k. An additional advantage of this modification is that the algorithm again is projected gradient ascent with the diagonals of $P(\Theta^k)$ given by

$$\frac{\mu^k}{T}, \qquad \frac{\alpha^k}{n - \sum_{i=1}^{n} e^{-w(T-t_i)}}, \qquad \frac{\omega^k}{\sum_{i>j}(t_i - t_j)p_{ij}^k + \alpha \sum_{i=1}^{n}(T - t_i)e^{-w^{k+1}(T-t_i)}}. \tag{20}$$

To investigate the convergence of (14) we simulate realizations of Equation (1) with varying values of ω and fixed paramters $\mu = 1$, $\alpha = .5$, and $T = 2000$, along with initial condition $\lambda(0) = \mu$. For each of the values of $\omega^{-1} = 10^{-2}, 10^{-1}, 10^0, 10^1$ we simulate 10 realizations of the process and run the EM algorithm (with boundary correction) to estimate the parameters (μ, α, ω). A convergence criteria

of 10^{-5} for consecutive iterates in (14) is used to determine when the EM algorithm stops. In Figure 1, we plot the mean and variance of the estimates for α and μ, in addition to the condition number (ratio of the magnitude of the largest to smallest eigenvalue) of $P(\Theta^k)H(\Theta^k)$ at the last iteration. The results follow closely to those observed for the convergence of EM in the context of mixture models (Xu & Jordan, 1996). First, the number of iterations required for the convergence of (14) increases dramatically with ω^{-1}, the time scale over which g decays. For small ω^{-1}, the condition number is close to 1, the condition number of Newton's method. Thus for well separated time scales the EM algorithm behaves like a second order method, which explains the fast convergence observed for certain parameter values. However, as ω^{-1} is increased, the number of iterations required for convergence increases from less than 10 to several thousand iterations in the case of $\omega^{-1} = 10$. Furthermore, the variance of parameter estimates also increases with increasing ω^{-1}. For $\omega^{-1} = .01$, the variance is less than 4% of the true value, increasing to over 12% for $\omega^{-1} = .10$. While these results highlight the need for caution, particularly in interpreting branching ratio estimates, for reasonable sample sizes and large T nonparametric methods seem to be capable of reproducing $g(t)$ and $\mu(t)$ even when a separation of scales of the form $\omega^{-1} << \mu^{-1}$ is not present (see Section 4).

3 MPLE for Hawkes process estimation

For a point process with conditional intensity $\lambda(t)$, the log-likelihood function is given by,

$$L = \sum_{i=1}^{n} \log \lambda(t_i) - \int_0^T \lambda(t)dt \tag{21}$$

In the case of parametric estimation, $\lambda(t)$ is assumed to have a specific functional form. The goal in maximum likelihood estimation, then, is to maximize (21) over all possible parameter choices. The advantages of parametric MLE are that the optimization problem determined by (21) is relatively easy to solve and fewer data points may be required to estimate the parameters compared to nonparametric estimation.

When a parametric form for λ is not known, nonparametric methods are a desirable alternative. For this purpose we will use Maximum Penalized Likelihood Estimation (MPLE), which entails maximizing the likelihood function over functions that possess a certain amount of regularity, but not necessarily taking a particular parametric form. Here we assume that $\lambda(t) = \mu(t) + \sum_{t>t_i} g(t-t_i)$ and then penalize the log-likelihood with regularity terms,

$$L = \sum_{i=1}^{n} \log \lambda(t_i) - \int_0^T \lambda(t)dt - \alpha_1 R(\mu) - \alpha_2 R(g), \tag{22}$$

where R is a roughness penalty (Eggermont & LaRiccia, 2001), for example the L^2 norm of the first derivative of μ or g. While (22) could be maximized directly, the form of λ would lead to an Euler-Lagrange equation comprised of a dense system of nonlinear equations. However, by reformulating the problem in the EM framework, the Euler-Lagrange equations greatly simplify. In order to see how the EM algorithm works in the context of MPLE, it is useful to first consider the Hawkes process when the branching structure is observable.

For an observable branching structure, we define the following random variables

$$\chi_{ij} = \begin{cases} 1 & \text{if event } i \text{ was caused by event } j \\ 0 & \text{otherwise} \end{cases}$$

and

$$\chi_{ii} = \begin{cases} 1 & \text{if event } i \text{ is a background event} \\ 0 & \text{if event } i \text{ is an offspring event} \end{cases}$$

Given the values of these variables, the problem of estimating μ and g reduces to two decoupled density estimation problems. For example, kernel density estimation could be used to estimate $\mu(t)$ and $g(t)$,

$$\mu(t) = \frac{1}{n_b} \sum_{i=1}^{n} \chi_{ii} A_h(t - t_i) \tag{23}$$

$$g(t) = \frac{1}{n_o} \sum_{j=2}^{n} \sum_{i=1}^{j-1} \chi_{ij} A_h(t_i - t_j) \tag{24}$$

where $n_b = \sum \chi_{ii}, n_o = \sum \sum \chi_{ij}$ and $A_h(t) = h^{-1} A(h^{-1} t)$. However, one drawback of KDE is that the number of inter-event times $t_i - t_j$ is order n^2 once the expectation is taken over the χ_{ij}, making KDE computationally expensive for realistic data sets of size $n = O(10^4)$. MPLE is thus a natural choice for smoothing, as the data can be binned in order to make the problem computationally tractable.

In the case of MPLE without penalization, the complete data log-likelihood is given by,

$$L(\mu, g) = \underbrace{\left[\sum_{i=1}^{n} \chi_{ii} \log(\mu(t_i)) \right] - \int_0^T \mu(t) dt}_{L_\mu} + \underbrace{\sum_{i=2}^{n} \left[\sum_{j=1}^{i-1} \left[\chi_{ij} \log(g(t_i - t_j)) \right] - \int_{t_j}^{T} g(t - t_j) dt \right]}_{L_g} \tag{25}$$

This leads to the MISD algorithm developed in (Marsan & Lengline, 2008), which we review here

for the simplified model (1) without space or earthquake magnitude and with μ constant in time.

As the branching structure is unobservable, the MISD algorithm works by maximizing the expectation of the complete data log-likelihood. Thus χ_{ij} is replaced with its expectation, $\mathrm{E}[\chi_{ij}] = p_{ij}$, representing the probability that event i is caused by event j. As a matrix these probabilities take the form,

$$\begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdots & p_{1n} \\ & p_{22} & p_{23} & \cdots & p_{2n} \\ & & p_{33} & \cdots & p_{3n} \\ & & & \ddots & \vdots \\ & & & & p_{nn} \end{bmatrix}$$

where $\sum_{j=1}^{i} p_{ij} = 1$ for all $i = 1, \ldots, n$, as each event i must be caused by one of the previous events ($j = 1, \cdots, i-1$) or must be a background event ($i = j$).

MISD works by iterating between the following two steps:

Expectation step:

The triggering weight is estimated by

$$p_{ij}^{k} = \frac{g^{k}(t_i - t_j)}{\mu^{k} + \sum_{i=1}^{j-1} g^{k}(t_i - t_j)} \tag{26}$$

for $t_i < t_j$. The background weight is estimated as

$$p_{jj}^{k} = \frac{\mu^{k}}{\mu^{k} + \sum_{i=1}^{j-1} g^{k}(t_i - t_j)}. \tag{27}$$

Maximization step:

The updated rates are then computed as

$$g_{m}^{k+1} = \frac{1}{\delta t} \sum_{i,j \in A_m} p_{ij}^{k} \tag{28}$$

where A_m is the set of pairs of events such that $m\delta t \leq |t_i - t_j| \leq (m+1)\delta t$, $g_m = g(m\delta t)$, and δt is a numerical discretization parameter. Similarly, the updated background rate is computed as

$$\mu^{k+1} = \frac{1}{T} \sum_{j=1}^{n} p_{jj}^{k} \tag{29}$$

where T is the duration of the time series (containing n events). Equations (28) (29) follow from setting $\frac{\partial}{\partial \mu} E[L]$ and $\frac{\partial}{\partial g_m} E[L]$ equal to zero. Equation (28) gives the histogram estimator in density estimation.

Next we consider MPLE for the estimation of $\lambda(t)$ when the background rate is allowed to vary. There are many choices for the penalty functional, but we restrict our attention to Good's Penalty (Good & Gaskins, 1971). At each EM-iteration we solve the following minimization problems for μ and g:

$$\text{minimize} \quad -\sum_{i=1}^{n} p_{ii} \log(\mu(t_i)) + \int_0^T \mu(t)dt + \alpha_1 \|(\mu^{1/2})'\|_2^2 \tag{30}$$

$$\text{subject to} \quad \mu \in L^1(\mathbb{R}), \quad \mu \geq 0 \tag{31}$$

and

$$\text{minimize} \quad -\sum_{i=2}^{n} \left[\sum_{j=1}^{i-1} \left[p_{ij} \log(g(t_i - t_j)) \right] - \int_{t_j}^{T} g(t - t_j)dt \right] + \alpha_2 \|(g^{1/2})'\|_2^2 \tag{32}$$

$$\text{subject to} \quad g \in L^1(\mathbb{R}), \quad g \geq 0 \tag{33}$$

As both equations are solved analogously, we illustrate our computational approach for the equation for g. We first make the substitution $u = g^{1/2}$ to get the equivalent problem:

$$\text{minimize} \quad -\sum_{i=2}^{n} \left[\sum_{j=1}^{i-1} \left[p_{ij} \log(u^2(t_i - t_j)) \right] - \int_{t_j}^{T} u^2(t - t_j)dt \right] + \alpha_2 \|u'\|_2^2$$

$$\text{subject to} \quad u \in W^{1,2}(\mathbb{R}), \quad u \geq 0 \tag{34}$$

The solution to the minimization problem in Equation (34) will solve the Euler-Lagrange equation (subject to suitable boundary conditions),

$$-\alpha u''(t) + Cu(t) = \frac{D}{u(t)} \tag{35}$$

where C is the number of events t_j occurring before time t and $D = \sum_{i=2}^{n} \sum_{j=1}^{i-1} p_{ij} \delta(t - (t_i - t_j))$. Solving this ODE is tractable, which is one of the reasons Good's penalty works well (De Montricher, Tapia, & Thompson, 1975).

The ODE is then discretized,

$$-\alpha \frac{u_{m+1} - 2u_m + u_{m-1}}{\delta t^2} + C_m u_m = \frac{D_m}{u_m}$$

where $C_m = \sum_{i:T-t_i<m\delta t} \delta t$ and $D_m = \frac{1}{\delta t} \sum_{i,j \in A_m} p_{ij}$. To solve the discretized system we multiply through by u_m and employ the Seidel type iterative method used in (Mohler, Bertozzi, Goldstein, & Osher, 2011), solving the resulting quadratic equation for u_m implicitly at each sweep.

4 Numerical Examples

Here we compare the accuracy of MISD and MPLE for recovering the triggering kernel and background rate of the Hawkes process (1). The true values for the parameters are $\mu = 1, k_0 = .5$ and $w = 2$, resulting in a realization of 1,907 events total and 975 background events over a time period $T = 1,000$. For each model the EM algorithm is run for 100 iterations and g is assumed to be zero outside the interval $[0, 4]$. For MPLE, 50 bins are used for estimating g and α_2 is selected such that $\|g - \hat{g}_{MPLE}\|_2^2$ is minimized. This yields a smoothing parameter for MPLE of $\alpha_2 = 5.5029$. We then choose the number of bins for MISD such that $\|g - \hat{g}_{MISD}\|_2^2$ is minimized. This leads us to choose 8 bins for estimating g with the MISD algorithm.

In Figure 4, we plot the estimates provided by the MISD and MPLE algorithms. The top row corresponds to MISD, while the bottom row corresponds to MPLE. The true values for μ and g are shown in red while the estimates are shown in blue. The large error in the estimate for μ for early iterations is a result of the initial guess of the p_{ij}'s. We note that the estimate for $\hat{\mu}_{MISD} = 0.862$, while $\hat{\mu}_{MPLE} = 0.9831$, and the convergence of MPLE stabilizes more quickly (around 50 iterations) compared to MISD (around 75 iterations). We also note that both the MISD algorithm and the MPLE algorithm will actually be estimating μ as $975/1000 = .975$ as opposed to the value used to simulate the process, $\mu = 1$. The L_2 error of g for the MISD algorithm is 0.1752 while the L_2 error for the MPLE algorithm is 0.0707. Thus MPLE is an order of magnitude more accurate than MISD in this example.

We also verify convergence of MPLE as the size of the data set increases. In Figure 3, we plot the L_2 error of g with respect to the length of the time window, T. We generate one realization of a point process with a stationary background rate over the time interval [0, 5,000] and take successively larger pieces of it and plot the error. To get an idea of how many events there are, when T= 1,000 there are 445 events and when T=5,000 there are 2,528 events. The grid spacing for g is chosen to be $\delta t = 0.1$. We note that the L_2 error decreases from 0.975 to 0.0446. The grid spacing would need to be decreased further to decrease the error below these levels.

Next we investigate the ability of MPLE to recover Hawkes process intensities when the background rate is changing in time. Here at each iteration of the EM algorithm μ is estimated anal-

ogously to g. We consider the Hawkes process with background rate and triggering kernel given by,

$$\lambda(t) = \underbrace{0.4(\sin(.0004 \cdot 2\pi t - 1000) + 1.1)}_{\mu(t)} + \sum_{t_k < t} \underbrace{e^{-2(t-t_k)}}_{g(t)}$$

Here the background rate varies on a time scale of 1000 days, whereas g decays on a time scale of 10 days. In Figure 4, we plot the true values for $\mu(t)$ and g in red along with the estimated values in blue. There are 2,508 total events total and 1,279 background events where $T = 3,000$. The values for the parameters are $k_0 = 0.5$ and $w = 2$. The EM algorithm is run until consecutive iterates are within a specified tolerance. We use 50 bins for μ over the entire time interval and 50 bins for g over a time interval of $[0, 4]$. The smoothing parameters are $\alpha_1 = 10,000$ and $\alpha_2 = 0.01$.

5 Iraq insurgency violence

The ability of MPLE to handle general background rates and triggering kernels makes it useful for new applications when little is known about the underlying process that is driving activity. For example, we analyze a data set provided by Iraq Body Count (http://www.iraqbodycount.org) comprised of the times of civilian deaths due to insurgent activity in Iraq between 2003 and 2007. We make no a priori assumptions for the functional forms of μ and g except that we assume the timescale over which μ is changing is larger than the timescale over which g decays. Recent analysis of insurgency data (Lewis et al., 2011) makes use of a kernel density estimate of all events as a proxy for the form of the background rate and assumes a parametric form for g.

The three regions that we analyze are Mosul, Fallujah, and Najaf. In Mosul, there are 1,280 events over 1,734 days. In Fallujah, there are 501 events over 1,700 days. And in Najaf, there are 144 events happening over 1,700 days.

In Figures 5 and 6, we plot estimates of the background rate and triggering kernel for the regions Mosul and Fallujah respectively. In both regions a general escalation of violence is observed in the background rate prior to the U. S. troop surge in 2007. This increase in violence occurs over a slow timescale of 1,000 days, where the background rate increases by an order of magnitude. Self-excitation is also present in the data, as the overall rate of violence increases significantly immediately after an event and remains elevated for the several weeks following each event. There is a general decay in time for both triggering densities. For Mosul, there is a large uptick in intensity for g beginning around 20 days. Something similar to this appears in the estimate for Fallujah around 10 days, but the peak is much less pronounced compared to the size of the initial spike at zero. This smaller bump in g for Fallujah could be attributed to variance due to small sample size, though the larger bump for Mosul is harder to attribute to variance. We also note that the general shapes of μ

and g were not sensitive to choices for α_1 and α_2.

Contrast these two examples with the more oscillatory background rate for Najaf in Figure 7. We note that their are fewer events in this data set and that the clustering is more apparent to the naked eye. This could perhaps be a result of the large peaks in the triggering kernel at 1, 7 and 19 days. The heightened rate immediately after an event seems reasonable due to coordinated attacks by insurgents within a group or between groups. As well, the peak at 7 days implies that there may be more attractive days of the week in which to attack a target. This information could aid in the allocation of resources when trying to fight insurgent activity.

6 Conclusion

We extended the density estimation technique of MPLE to the nonparametric estimation of Hawkes process intensities with time-varying background rate. The extension to self-exciting point processes makes use of the EM algorithm, which effectively decouples the problem into two density estimation problems. We evaluated the effectiveness of the algorithm in comparison to MISD and for a Hawkes process with periodic background rate. We then illustrated how the method might be useful in social science applications where prior knowledge of the intensity is not available. In particular, we analyzed data on the Iraq insurgency exhibiting both exogenous and endogenous clustering.

7 Supplemental Materials

EM - Maximum Penalized Likelihood Estimation: There is one function that takes in a sequence of times and a parameter set and it returns a matrix of p_{ij}'s as well as estimates for μ and g. (MPLE.zip which is a zipped matlab file).

References

Crane, R., & Sornette, D. (2008). Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, *105*(41), 15649-15653.

De Montricher, G., Tapia, R., & Thompson, J. (1975). Nonparametric maximum likelihood estimation of probability densities by penalty function methods. *The Annals of Statistics*, *3*(6), 1329–1348.

Eggermont, P. P. B., & LaRiccia, V. (2001). *Maximum penalized likelihood estimation.* Springer Series in Statistics.

Errais, E., Giesecke, K., & Goldberg, L. R. (2010). Affine point processes and portfolio credit risk. *SIAM Journal on Financial Mathematics*, *1*, 642-665.

Good, I., & Gaskins, R. (1971). Nonparametric roughness penalties for probability densities. *Biometrika*, *58*(2), 255.

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, *58*(1), 83-90.

Lewis, E., Mohler, G., Brantingham, P. J., & Bertozzi, A. (2011). Self-exciting point process of insurgency in iraq. *submitted*.

Marsan, D., & Lengline, O. (2008). Extending earthquakes' reach through cascading. *Science*, *319*(5866), 1076.

Marsan, D., & Lengliné, O. (2010). A new estimation of the decay of aftershock density with distance to the mainshock. *Journal of Geophysical Research*, *115*(B9), B09302.

Mohler, G. O., Bertozzi, A. L., Goldstein, T. A., & Osher, S. J. (2011). Fast tv regularization for 2d maximum penalized likelihood estimation. *Journal of Computational and Graphical Statistics*, *0*(0), 1-13.

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., & Tita, G. E. (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Society*.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *J. Amer. Statist. Assoc.*, *83*(401), 9–27.

Ogata, Y. (1998). Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, *50*, 379-402. (10.1023/A:1003403601725)

Sornette, D., & Utkin, S. (2009). Limits of declustering methods for disentangling exogenous from endogenous events in time series with foreshocks, main shocks, and aftershocks. *Physical Review E*, *79*(6), 61110.

Veen, A., & Schoenberg, F. P. (2008). Estimation of spacetime branching process models in seismology using an emtype algorithm. *Journal of the American Statistical Association*, *103*(482), 614-624.

Xu, L., & Jordan, M. I. (1996). On convergence properties of the em algorithm for gaussian mixtures. *Neural Computation*, *8*(1), 129-151.

Zhuang, J., Ogata, Y., & Vere-Jones, D. (2002). Stochastic declustering of space-time earthquake occurrences. *Journal of the American Statistical Association*, *97*(458), 369–380.

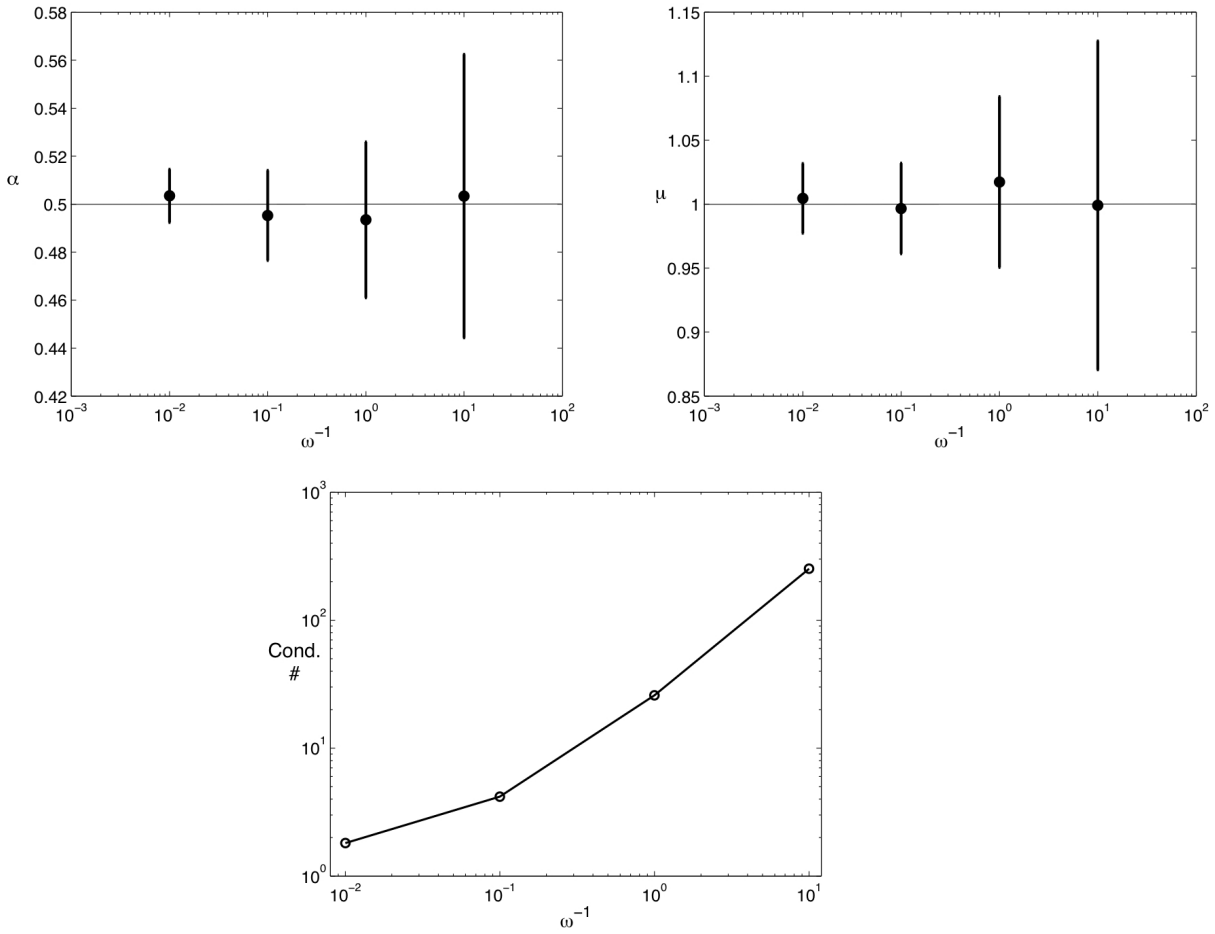

Figure 1: Top: Parameter estimates for branching ratio and background rate (mean and standard deviation) for 10 realizations of the process with parameters $\mu = 1$, $\alpha = .5$, $T = 2000$ and varying ω. Bottom: Condition number for varying ω, computed at the final EM iteration and averaged over the 10 realizations.

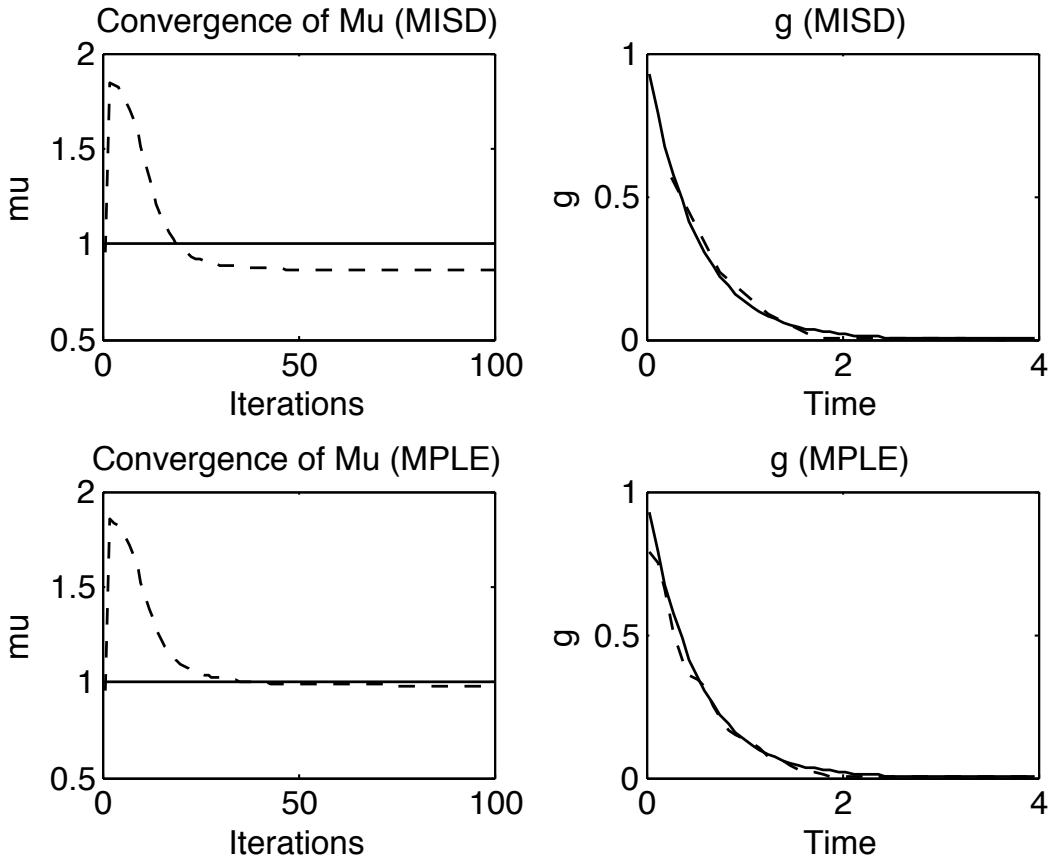

Figure 2: MISD estimate for μ (top left) and estimate for g (top right). MPLE estimate for μ (lower left) and g (lower right). For each figure the dotted line represents the estimate while the solid line represents the true value.

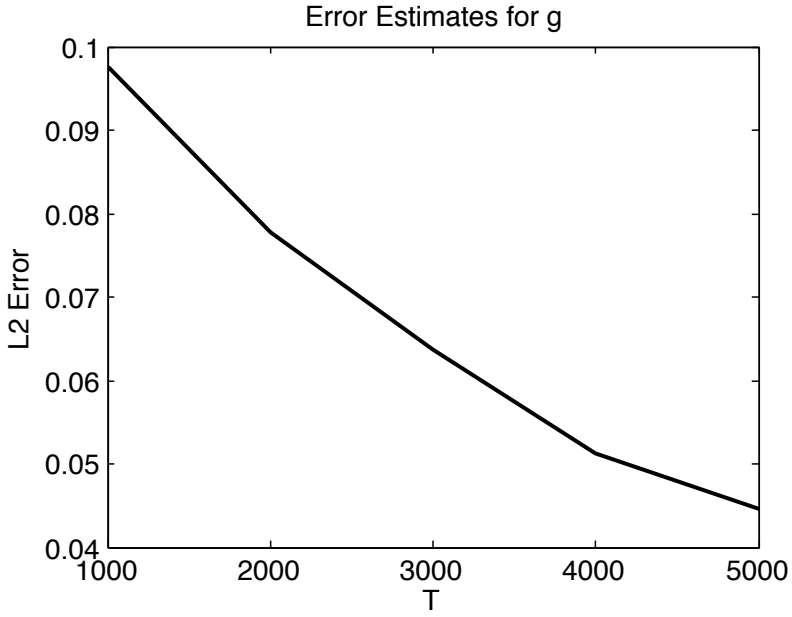

Figure 3: L_2 error of g with respect to the length of the time window, T.

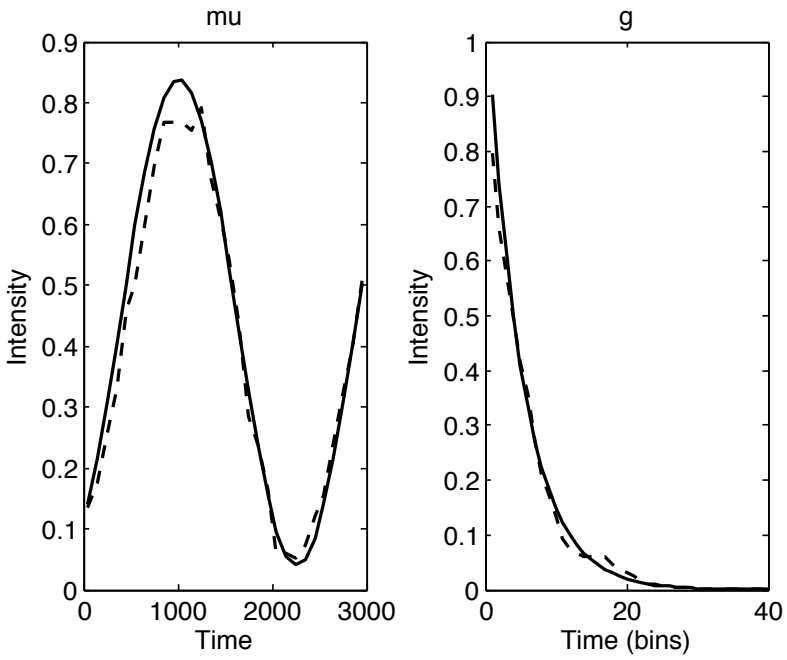

Figure 4: Estimated μ and g (dashed lines) plotted against the true functions (solid lines) using the MPLE algorithm.

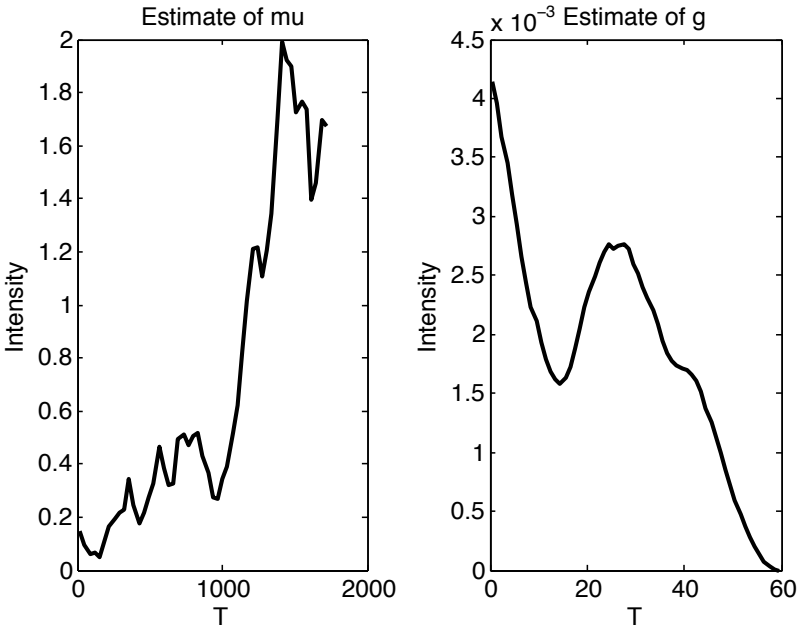

Figure 5: We plot the estimates for the background rate (left) and the triggering density (right) for Mosul. There are 34 bins for μ over T=1,734 days, while g is estimated on [0,60] with 60 bins. Also, $\alpha_1 = 2,000$ while $\alpha_2 = 5,000$.

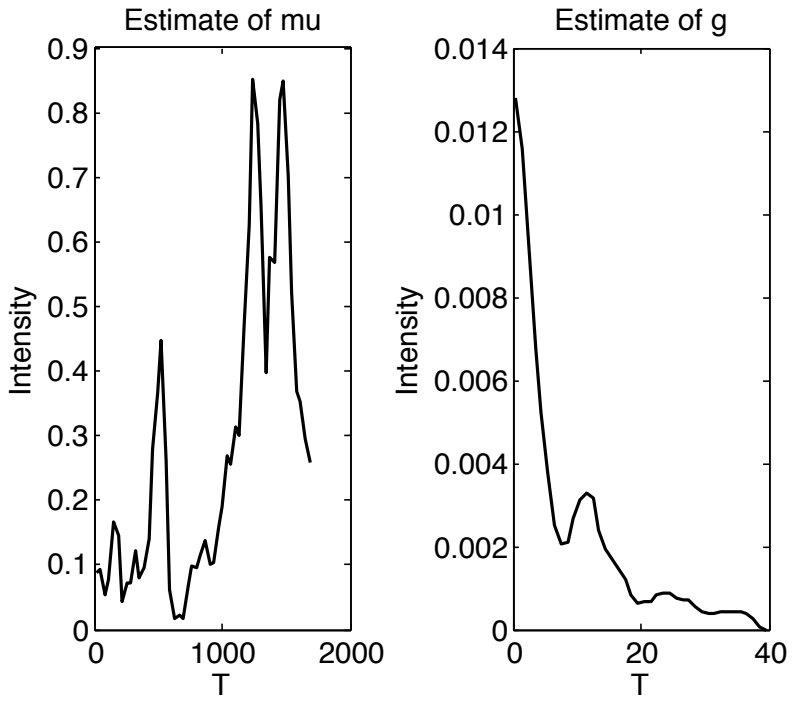

Figure 6: We plot the estimates for the background rate (left) and the triggering density (right) for Fallujah. There are 34 bins for μ over T=1,700 days, while g is estimated on [0,40] with 40 bins. Also, $\alpha_1 = 1,000$ while $\alpha_2 = 500$.

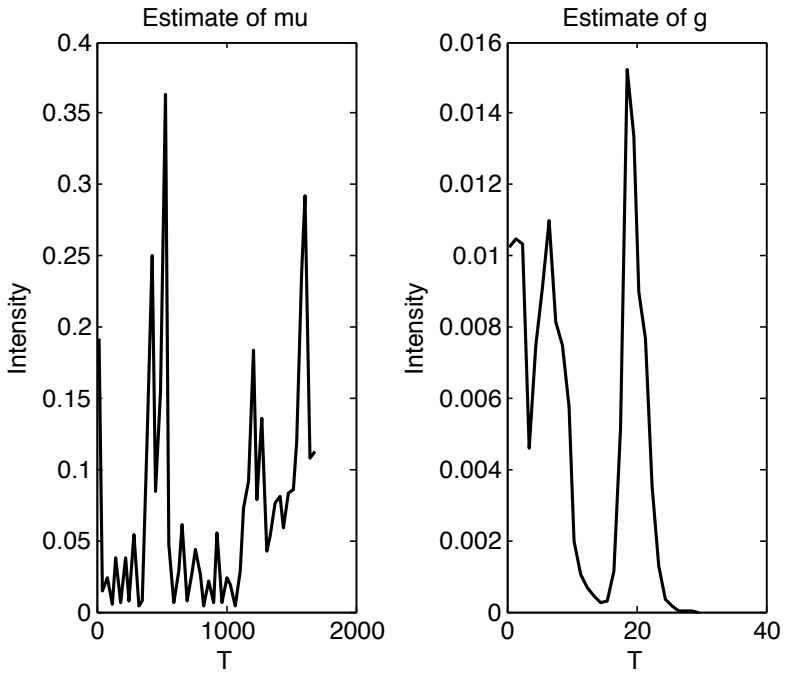

Figure 7: We plot the estimates for the background rate (left) and the triggering density (right) for Najaf. There are 34 bins for μ over T=1,700 days, while g is estimated on [0,30] with 30 bins. Also, $\alpha_1 = 400$ while $\alpha_2 = 100$.