The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Proper Loss Functions for Nonlinear Hawkes Processes

Aditya Krishna Menon
Data61 and the Australian National University
aditya.menon@data61.csiro.au

Young Lee*
National University of Singapore
dcsleey@nus.edu.sg

Abstract

Temporal point processes are a statistical framework for modelling the times at which events of interest occur. The Hawkes process is a well-studied instance of this framework that captures *self-exciting* behaviour, wherein the occurrence of one event increases the likelihood of future events. Such processes have been successfully applied to model phenomena ranging from earthquakes to behaviour in a social network.

We propose a framework to design new loss functions to train linear and nonlinear Hawkes processes. This captures standard maximum likelihood as a special case, but allows for other losses that guarantee convex objective functions (for certain types of kernel), and admit simpler optimisation.

We illustrate these points with three concrete examples: for linear Hawkes processes, we provide a least-squares style loss potentially admitting closed-form optimisation; for exponential Hawkes processes, we reduce training to a weighted logistic regression; and for sigmoidal Hawkes processes, we propose an asymmetric form of logistic regression.

Introduction

Temporal point processes are a classical statistical framework for modelling the times at which certain events of interest occur, such as failure times of a hard drive or the impact times of an earthquake (Cox and Isham 1980; Daley and Vere-Jones 2003). The simplest incarnation of these models is the *Poisson process*, which assumes the times between successive events are independent, and the number of events occurring in a time window follows a suitable Poisson distribution (Kingman 1993). Such models are a core tool in queuing theory (Erlang 1909; Kendall 1953).

Despite their versatility, Poisson processes have an important limitation: they are incapable of modelling *self-excitation*, wherein the occurrence of one event increases the likelihood of further events. This characteristic is present in many real-world phenomena, such as the occurrence of an earthquake triggering an aftershock. The *Hawkes process* (Hawkes 1971; Laub, Taimre, and Pollett 2015) is an important extension of the classical Poisson process to allow for such "burstiness". The model has been applied in fields ranging from seismology (Ogata 1988), finance (Bowsher 2007; Hardiman, Bercot, and Bouchaud 2013), and social media (Crane and Sornette 2008; Zhou, Zha, and Song 2013).

Given historical data of event times, the standard way to fit a Hawkes process is to maximise its log-likelihood (Ozaki 1979). This approach is appealing owing to its conceptual simplicity; however, for a generic nonlinear Hawkes process (defined formally in the next section), the resulting objective may be non-convex. Further, even for linear Hawkes processes, optimisation of the likelihood requires an involved iterative optimisation. This raises a natural question: how might we design other losses for training Hawkes processes that have favourable properties compared to the likelihood?

In this paper, we provide a framework to design loss functions for (non-)linear Hawkes processes. Specifically, given a particular choice of nonlinearity, we provide loss function that is suitable for estimating the parameters of the corresponding nonlinear Hawkes process, and which is further convex given a particular structure on the kernel. We study three concrete instantiations of this framework:

(a) for linear Hawkes processes, we propose a loss with a potential closed-form solution; to our knowledge, the only extant closed-form solution for Hawkes processes arises in EM training (Lewis and Mohler 2011).

(b) for exponential Hawkes processes[1], we establish the suitability of the logistic loss, which allows us to reduce training to a logistic regression problem.

(c) for sigmoidal Hawkes processes, we show the viability of a modified logistic regression objective, which provides a convex objective for training.

At a technical level, our proposal rests upon three simple observations: (1) the Hawkes likelihood can be interpreted as a binary classification objective; (2) the asymptotic optimiser of the likelihood is a scaled density estimate; and (3) the broader family of proper losses (Buja, Stuetzle, and Shen 2005) retains this optimal solution, and thus also the fundamental target of interest. While these observations are conceptually simple, the explication of their connections for fitting Hawkes processes is to our knowledge novel, and their implications we believe of interest.

*Work conducted while at Data61.

[1]The term "exponential Hawkes" is sometimes used to mean a linear Hawkes process with exponential kernel. We use the term to mean a Hawkes process with exponential link, but arbitrary kernel.

Background

Our framework requires some background on temporal point processes, as well as loss functions for binary classification. A glossary of important symbols is provided in Table 1.

Temporal point processes

Temporal point processes model the times at which events of interest occur via a stochastic process $(\mathsf{N}_t)_{t\geq 0}$, where $\mathsf{N}_t - \mathsf{N}_s$ measures the number of events that occur in the time interval $(s, t]$. We focus on two such processes.

Inhomogeneous Poisson process Fix some locally integrable $\lambda\colon \mathbb{R}_+ \to \mathbb{R}_+$, and for any $0 \leq s < t$, let $\Lambda(s,t) \doteq \int_s^t \lambda(x)\,\mathrm{d}x$. An *inhomogeneous Poisson process* (*IPP*) *with intensity* $\lambda(\cdot)$ satisfies (Daley and Vere-Jones 2003):

(a) $\mathsf{N}_0 = 0$ almost surely

(b) for any $s < t$, $\mathsf{N}_t - \mathsf{N}_s \sim \mathrm{Poisson}(\Lambda(s,t))$

(c) for any $s < t \leq s' < t'$, $\mathsf{N}_t - \mathsf{N}_s \perp\!\!\!\perp \mathsf{N}_{t'} - \mathsf{N}_{s'}$.

Condition (a) posits that events occur strictly after time 0. Condition (b) posits that the number of events in any interval has a Poisson distribution, with mean given by the integrated intensity in that interval. Condition (c) posits the the number of events in two disjoint intervals is independent.

IPPs may also be understood as the following generative model for event times: given some end time T, the number of events N is drawn from a Poisson with mean $\Lambda(0,T)$, and the N event times are then drawn *i.i.d.* from a distribution P with density (Cox and Isham 1980, pg. 46)

$$p(t) = \lambda(t)/\Lambda(0,T). \tag{1}$$

Given a history $\mathcal{T} \doteq \{t_n\}_{n=1}^N$ of event times, suppose we seek an intensity from a family $\{\lambda(\cdot;\theta) \mid \theta \in \Theta\}$ for suitable parameter space Θ. We may minimise the negative log-likelihood of θ, which for $T \doteq \max_n t_n$ is (upto constants) (Daley and Vere-Jones 2003, Equation 2.1.9)

$$\mathcal{L}_{\mathrm{IPP}}(\theta;\mathcal{T}) \doteq \sum_{n=1}^{N} -\log\lambda(t_n;\theta) + \int_0^T \lambda(u;\theta)\,\mathrm{d}u. \tag{2}$$

When the integral above does not have a closed form, it may be approximated numerically (Davis and Rabinowitz 1984).

Hawkes process The *Hawkes process* extends IPPs so as to model self-excitation. Given a history $\mathcal{T} = \{t_n\}_{n=1}^N$ of event times, a nonlinear Hawkes process with link $F\colon \mathbb{R} \to \mathbb{R}$ posits the intensity (Brémaud and Massoulié 1996)

$$\lambda(t;\mathcal{T}) = F\left(\sum_{n:t>t_n} g(t-t_n)\right) \tag{3}$$

for a *decay function* $g\colon \mathbb{R}_+ \to \mathbb{R}$. The linear Hawkes process (Hawkes 1971) is the case $F(z) = z$ and $g(\cdot)$ nonnegative.

One can parametrise the Hawkes process via a family of decay functions. A popular choice is the *exponential decay*

$$g(z;\theta) = \mu + \alpha \cdot e^{-\delta\cdot z}, \tag{4}$$

with $\theta = \{\mu, \alpha, \delta\}$. For a linear Hawkes process, one requires $\mu, \alpha > 0$, so that there is a background intensity μ, and every time an event occurs, there is a local increase in the probability of further events, *viz*. the phenomena of *self-excitation*. A nonlinear Hawkes process allows $\alpha < 0$, and thus can model *self-inhibition* (Reynaud-Bouret and Schbath 2010).

The negative log-likelihood of the Hawkes process is identical to that of the IPP, with Equation 3 as the intensity. Concretely, given a history $\mathcal{T}$, it is (Ozaki 1979)

$$\mathcal{L}_{\mathrm{HP}}(\theta;\mathcal{T}) \doteq \sum_{n=1}^{N} -\log\lambda(t_n;\mathcal{T},\theta) + \int_0^T \lambda(t;\mathcal{T},\theta)\,\mathrm{d}t. \tag{5}$$

For a general nonlinear Hawkes process, as with IPPs, the integral in Equation 5 must be approximated numerically.

Symbol	Meaning
λ	Intensity function
F	Hawkes link function
g	Decay function
P, p	Event dist. & density
Q, q	Uniform dist. & density

Symbol	Meaning
ℓ	Loss function
Ψ	Proper link function
$\mathcal{T}$	Observed event times
T	Maximal event time
T, T'	Random event times

Table 1: Glossary of important symbols.

Loss functions for binary classification

Given examples of instances paired with labels in $\{\pm 1\}$, the binary classification problem is to predict the labels of unseen instances. Formally, fix an instance space $\mathcal{X}$, distributions P, Q over $\mathcal{X}$, and *loss function* $\ell\colon \{\pm 1\} \times \mathbb{R} \to \mathbb{R}_+$. Then, we seek a *scorer* $s\colon \mathcal{X} \to \mathbb{R}$ with low expected loss or *risk* for positive (negative) instances drawn from P (Q), i.e.[2]

$$\mathcal{L}_{\mathrm{BC}}(s;P,Q) \doteq \mathop{\mathbb{E}}_{\mathsf{X}\sim P}\left[\ell(+1, s(\mathsf{X}))\right] + \mathop{\mathbb{E}}_{\mathsf{X}'\sim Q}\left[\ell(-1, s(\mathsf{X}'))\right]. \tag{6}$$

Given samples $\mathcal{S} \doteq \{(x_i, +1)\}_{i=1}^N \cup \{(x_j, -1)\}_{j=1}^M$, suppose we seek a scorer from a family $\{s(x;\theta) \mid \theta \in \Theta\}$, e.g. linear models $s(x;\theta) = \langle\theta, x\rangle$. We may minimise the weighted *empirical* risk, which for optional weight $w > 0$ is

$$\mathcal{L}_{\mathrm{BC}}(\theta;\mathcal{S}) \doteq \frac{1}{N}\sum_{i=1}^{N} \ell(+1, s(x_i;\theta)) + \frac{w}{M}\sum_{j=1}^{M} \ell(-1, s(x_j;\theta)), \tag{7}$$

Classically, one is interested in the zero-one loss $\ell(y,v) = [\![yv < 0]\!] + 1/2\cdot[\![v = 0]\!]$. Owing to its intractability, it is common to instead use a convex surrogate loss. A useful family of surrogates are that of *strictly proper composite* losses (Buja, Stuetzle, and Shen 2005; Reid and Williamson 2010). These are the fundamental losses of probabilistic classification, where the nonparametric risk minimiser is

$$s^* \doteq \mathop{\mathrm{argmin}}_{s\in\mathbb{R}^{\mathcal{X}}} \mathcal{L}_{\mathrm{BC}}(s;P,Q) = \Psi\circ\eta, \tag{8}$$

i.e. $s^*(x) = \Psi(\eta(x))$, for invertible *link function* Ψ and *label-probability function* $\eta(x) \doteq \Pr(\mathsf{Y} = +1 \mid \mathsf{X} = x)$.

Intuitively, with a powerful class of scorers $\{s(\cdot;\theta) \mid \theta \in \Theta\} = \mathbb{R}^{\mathcal{X}}$, and sufficiently many samples, minimising a

[2] Generally, the expectations over P, Q are weighted by the corresponding marginal label probability, $\Pr(\mathsf{Y} = \pm 1)$. We assume without loss of generality that $\Pr(\mathsf{Y} = +1) = 1/2$, since the weighting may be absorbed into the loss.

strictly proper composite loss allows one to accurately recover the underlying probability of an instance being positive. A canonical example is the logistic loss $\ell(y, v) = \log(1 + e^{-yv})$, for which $\Psi^{-1}(v) = (1 + e^{-v})^{-1}$.

Beyond the likelihood: proper losses and IPPs

Since the Hawkes likelihood is equivalent to that of an IPP, we first generalise the IPP objective via three simple ideas:

(a) the IPP objective can be viewed as a special kind of binary classification (Equations 9, 11),

(b) the non-parametric minimiser of the IPP objective is a scaling of the event times' density (Lemma 1),

(c) any strictly proper composite loss preserves this optimal solution (Lemma 3), and thus provides a risk (Equation 17) which is a candidate alternative to the MLE.

IPPs as binary classification

For fixed $\mathcal{T} = \{t_n\}_{n=1}^N$ with $T = \max_n t_n$, the IPP objective was motivated as the log-likelihood of a particular probabilistic model. Ostensibly, this is an unsupervised or "one-class" learning problem, where the only observations are $\mathcal{T}$.

A simple trick lets us move to a more familiar "two-class" supervised learning problem. Let us approximate[3] the integral in Equation 2 by a Riemann sum, computed on a partition of $[0, T]$ into M uniformly spaced "background events" $\{t_m\}_{m=1}^M$. Since these events are T/M apart, we have

$$\hat{\mathcal{L}}_{\text{IPP}}(\theta; \mathcal{T}) \doteq \sum_{n=1}^{N} -\log \lambda(t_n; \theta) + \frac{T}{M} \cdot \sum_{m=1}^{M} \lambda(t_m; \theta)$$
$$\propto \frac{1}{N} \sum_{n=1}^{N} -\log \lambda(t_n; \theta) + \frac{T}{N} \cdot \frac{1}{M} \sum_{m=1}^{M} \lambda(t_m; \theta), \quad (9)$$

where the second equation simply scales the first by $1/N$. Equation 9 is a weighted empirical binary classification risk (Equation 7), with weight $w = T/N$ and asymmetric loss

$$(\forall v > 0)\, \ell(+1, v) = -\log v \qquad \ell(-1, v) = v. \quad (10)$$

Plainly, this is a classification problem with instances being times in $[0, T]$, the observed event times treated as "positive", and uniformly distributed background event times treated as "negative". Intuitively, IPPs seek to distinguish whether a candidate time comes from the underlying process, or from a uniform background process.

More generally, tackling the integral directly, we have

$$\mathcal{L}_{\text{IPP}}(\theta; \mathcal{T}) \propto \mathop{\mathbb{E}}_{\mathsf{T} \sim \widehat{P}} [-\log \lambda(\mathsf{T}; \theta)] + \frac{T}{N} \cdot \mathop{\mathbb{E}}_{\mathsf{T}' \sim Q} \left[\lambda(\mathsf{T}'; \theta)\right], \quad (11)$$

where $\widehat{P}$ is a discrete distribution that is uniform over $\mathcal{T}$, Q is the uniform distribution over $[0, T]$, and T, T' are random variables distributed according to the respective distributions. This is clearly a binary classification risk as per Equation 6, and we thus may view IPP fitting as solving a particular classification problem.

[3] We do not discretise time entirely, as done in e.g. (Hall and Willett 2016), since we use the exact times for the observed events.

Given this interpretation, a natural question is whether other choices of loss beyond Equation 10 are possible. There are at least two reasons to embark on such a quest. First, certain losses may admit simpler optimisation compared to the standard likelihood. Second, one may wish to model the intensity as $\lambda = F \circ s$, where either $-\log F$ or F is non-convex; alternate convex losses are thus of interest.

We now present a means of exploring other losses, by studying the fundamental target of interest in the IPP risk.

The minimiser of the IPP classification risk

The original (and compelling) justification for the IPP objective is that it arises naturally from the log-likelihood. An alternate justification is that it has a sensible minimiser. In more detail, recall that conditioned on the number of points N, the event times in an IPP are *i.i.d.* from a distribution P with density given by Equation 1. For fixed N, in the regime of infinitely many sample paths, Equation 11 approaches

$$\mathop{\mathbb{E}}_{\mathsf{T} \sim P} [\ell(+1, \lambda(\mathsf{T}; \theta))] + \frac{T}{N} \cdot \mathop{\mathbb{E}}_{\mathsf{T}' \sim Q} [\ell(-1, \lambda(\mathsf{T}'; \theta))]. \quad (12)$$

To explore other losses for the IPP, one strategy is to determine, akin to Equation 8, the choice of λ that minimises Equation 12 in a non-parametric setting where the intensity family $\{\lambda(\cdot; \theta) \mid \theta \in \Theta\} = \mathbb{R}_+^{\mathcal{X}}$. This minimiser is the object we converge to given infinitely many samples, and an arbitrarily flexible class of intensities; it thus represents the fundamental target of interest. One can then choose alternate losses that retain this target, i.e., have the same minimiser.

We now show the optimal λ is simply a scaled version of the underlying density for the event times. This implies IPPs are fundamentally entwined with density estimation.

Lemma 1. *For ℓ per Equation 10, uniform distribution Q over $\mathcal{X} \doteq [0, T]$, and distribution P over $\mathcal{X}$ with density p, let*

$$\lambda^* \doteq \mathop{\text{argmin}}_{\lambda \in \mathbb{R}_+^{\mathcal{X}}} \mathop{\mathbb{E}}_{\mathsf{T} \sim P} [\ell(+1, \lambda(\mathsf{T}))] + \frac{T}{N} \cdot \mathop{\mathbb{E}}_{\mathsf{T}' \sim Q} [\ell(-1, \lambda(\mathsf{T}'))] \quad (13)$$

Then, $\lambda^* = N \cdot p$.

The proof of Lemma 1 is a simple consequence of the fact that ℓ is strictly proper composite with a specific link.

Lemma 2. *The loss ℓ of Equation 10 is strictly proper composite with inverse link $\Psi^{-1}(v) = v/(1 + v)$ for $v > 0$.*

Proof. By (Reid and Williamson 2010, Corollary 12), a differentiable loss is strictly proper composite iff $\ell'(-1, v)/(\ell'(-1, v) - \ell'(+1, v))$ is invertible, in which case this is the inverse link Ψ^{-1}. Since the given ℓ has $\ell'(-1, v) = 1$ and $\ell'(+1, v) = -1/v$ for $v > 0$, the result follows. □

Let us return to Lemma 1 in light of this. Recall that for a strictly proper composite loss, the optimal scorer is a transform of $\eta(t) = \Pr(\mathsf{Y} = +1 \mid \mathsf{T} = t)$. If Q has density q, the optimal scorer is thus also a transform of the density ratio p/q, since by Bayes' rule (and the assumption $\Pr(\mathsf{Y} = +1) = 1/2$),

$$\frac{\eta(t)}{1 - \eta(t)} = \frac{\Pr(\mathsf{Y} = +1 \mid \mathsf{T} = t)}{\Pr(\mathsf{Y} = -1 \mid \mathsf{T} = t)} = \frac{\Pr(\mathsf{T} = t \mid \mathsf{Y} = +1)}{\Pr(\mathsf{T} = t \mid \mathsf{Y} = -1)} = \frac{p(t)}{q(t)}. \quad (14)$$

But since Q is uniform over $[0, T]$ by construction, p/q is merely $T \cdot p$; thus, estimating η implicitly estimates the density p. We now use this to sketch the proof of Lemma 1.

Proof of Lemma 1. Equation 13 concerns the *weighted* loss

$$\ell_{\text{wt}}(+1, v) \doteq \ell(+1, v) \qquad \ell_{\text{wt}}(-1, v) \doteq (T/N) \cdot \ell(-1, v). \quad (15)$$

Lemma 2 implies ℓ has link $\Psi(u) = u/(1-u)$. Thus, by (Menon and Ong 2016, Lemma 5), ℓ_{wt} is also strictly proper composite with link $\Psi_{\text{wt}}(u) = (N/T) \cdot u/(1-u)$. Consequently, by definition, the optimal scorer in Equation 13 is $\Psi_{\text{wt}} \circ \eta$. Now, by Equation 14, $\eta(t)/(1-\eta(t)) = p(t)/q(t) = T \cdot p(t)$, since Q is uniform over $[0, T]$. Thus, the optimal scorer is $(N/T) \cdot T \cdot p(t) = N \cdot p(t)$. □

The generalised IPP objective

Lemma 1 shows the fundamental target of interest in IPPs is the density of P. A reasonable alternative loss to Equation 10 should retain this target, i.e. be optimised by predicting the density. We thus seek to construct losses with this property.

In fact, *any* strictly proper composite loss ℓ is a viable candidate. By definition, any such loss recovers a transform of the class probability η; hence, following the same reasoning as in Lemma 1, it will recover a transform of p. Formally, we have the following analogue of Lemma 1.

Lemma 3. *For strictly proper composite loss ℓ with link Ψ, uniform distribution Q over $\mathcal{X} \doteq [0, T]$, and distribution P over $\mathcal{X}$ with density p, let*

$$s^* \doteq \operatorname*{argmin}_{s \in \mathbb{R}^{\mathcal{X}}} \mathop{\mathbb{E}}_{\mathsf{T} \sim P} [\ell(+1, s(\mathsf{T}))] + \frac{T}{N} \cdot \mathop{\mathbb{E}}_{\mathsf{T}' \sim Q} [\ell(-1, s(\mathsf{T}'))] \quad (16)$$

Then, $F \circ s^* = N \cdot p$ for $F(v) \doteq \Psi^{-1}(v)/(1 - \Psi^{-1}(v))$.

Proof. Equation 16 concerns the weighted loss ℓ_{wt} as in Equation 15. By (Menon and Ong 2016, Lemma 5), ℓ_{wt} is strictly proper composite with link $\Psi_{\text{wt}}(u) = \Psi(u/(w + (1 - w) \cdot u))$, where $w \doteq T/N$. Since $\eta(t) = Tp(t)/(1 + Tp(t))$, we have optimal scorer $s^*(t) = \Psi\left(p(t)/(p(t) + 1/N)\right)$. Further algebra then reveals that $F \circ s^* = N \cdot p$. □

Thus, for scorers $\{s(\cdot; \theta) \mid \theta \in \Theta\}$ and strictly proper composite ℓ, the *generalised IPP procedure* involves fitting

$$\mathcal{L}_{\text{GPP}}(\theta; \mathcal{T}) \doteq \sum_{n=1}^{N} \ell(+1, s(t_n; \theta)) + \int_0^T \ell(-1, s(t; \theta))\, \mathrm{d}t \quad (17)$$

$$\lambda(t; \theta) \doteq F(s(t; \theta)) \qquad F(v) \doteq \frac{\Psi^{-1}(v)}{1 - \Psi^{-1}(v)}, \quad (18)$$

where Ψ is the link function of ℓ. By Lemma 3, this procedure is optimised by an intensity that is a scaled version of the underlying density, as with the standard IPP objective.

The standard IPP procedure is recovered for ℓ as per Equation 10: here, $\Psi^{-1}(v) = v/(1+v)$ (by Lemma 2), and thus $F(v) = v$; consequently, the scorer s is equivalent to the intensity. For general Ψ, however, the intensity is a nonlinear transform of the scorer. We study concrete examples, and discuss suitable choices of loss ℓ, in the next section.

Discussion and related work

The relation between IPPs and density estimation has prior precedent. At a practical level, (Diggle 1985) proposed a variant of kernel density estimation to fit IPPs. At a theoretical level, our Lemma 1 on the non-parametric minimiser of the IPP objective complements a result of (Fithian and Hastie 2013) on a *parametric* minimiser: they showed that fitting *log-linear* IPPs, where $\lambda(t) = \exp(a + \langle b, \Phi(t) \rangle)$ for feature mapping $\Phi \colon \mathbb{R}_+ \to \mathbb{R}^D$, is equivalent to fitting a density estimate with $\widehat{p}(t) \propto \exp(\langle b, \Phi(t) \rangle)$. They did not however propose to generalise the IPP objective (Equation 17).

Viewed as a sibling of density estimation, the connection of IPPs to binary classification (Equations 9, 11) is not surprising: binary classification is an established viewpoint for anomaly detection (Steinwart, Hush, and Scovel 2006), (Hastie, Tibshirani, and Friedman 2009, Section 14.2.4), *viz.* the problem of estimating a level set of the density. Further, the risk of proper losses with a uniform background may be understood in terms of more general proper scoring rules for density estimation (Gneiting and Raftery 2007). However, in the context of IPPs, the explication of this fact and the derivation of its consequences are to our knowledge novel.

Our justification of Equation 17 is that it preserves an asymptotic, nonparametric minimiser; while necessary for any sensible alternative, it is not sufficient, as it ignores finite-sample effects. We shall revisit this point shortly.

New losses for (non-)linear Hawkes processes

The generalised IPP objective (Equation 17) provides an alternative to the log-likelihood. Recall that two motivations for exploring alternate losses is in obtaining objectives that are convex, as well as less involved to optimise. We make these points concrete by applying this framework to nonlinear Hawkes processes (Equation 3), which correspond to a particular form of intensity λ. In particular, we consider nonlinear Hawkes processes satisfying two assumptions:

A1 the link function $F(\cdot)$ is invertible

A2 the decay function (Equation 3) has the form

$$g(z; \theta) = \mu + \sum_{i=1}^{L} \alpha_i \cdot k_i(z) \quad (19)$$

for $\theta = (\mu, \alpha_1, \ldots, \alpha_L)$ and *triggering kernels* $k_i \colon \mathbb{R}_+ \to \mathbb{R}_+$.

Assumption A1 holds for the standard identity link $F(z) = z$, but also for a range of nonlinear links that shall be subsequently discussed. Assumption A2 is equivalent to requiring that, given event history $\mathcal{T}$, the intensity is

$$\lambda(t; \mathcal{T}) = F\left(\mu + \sum_{i=1}^{L} \alpha_i \cdot \sum_{n: t > t_n} k_i(t - t_n)\right).$$

With $L = 1$ and kernel $k(z) = e^{-\delta \cdot z}$, this captures the exponential decay $g(z) = \mu + \alpha \cdot e^{-\delta \cdot z}$ of Equation 4. Note that we assume δ is *fixed*, and not learned. This is a strong assumption, but can be removed at the expense of convexity (see discussion); further, one can mimic learning δ via a (sample size dependent) number of kernels with different δ's (Xu, Farajtabar, and Zha 2016).

We now provide a convex objective for fitting processes satisfying these assumptions. Figure 1 summarises.

INPUT: Invertible nonlinearity $F(\cdot)$; kernels $\{k_i\}_{i=1}^L$

PROCEDURE:

(1) Construct canonical proper loss per Equation 24, *or* alternate loss with link per Equation 23

(2) Find the linear scorer that minimises 22

(3) Estimate intensity using Equation 18

Figure 1: Framework to fit nonlinear Hawkes processes.

A linear model view of Hawkes processes

Before proceeding, we note that a useful way to view the Hawkes process is as follows. Suppose our decay function has the form of Equation 19. Given event times $\mathcal{T} = \{t_n\}_{n=1}^N$, define the feature mapping $\Phi \colon \mathbb{R}_+ \to \mathbb{R}^{L+1}$ by

$$\Phi(t; \mathcal{T}) \doteq \left(1, \Phi_k(t; \mathcal{T})\right) \quad \Phi_k(t; \mathcal{T}) \doteq \left(\textstyle\sum_{t > t_n} k_i(t - t_n)\right)_{i=1}^L . \quad (20)$$

Then, under Assumption A2, the intensity for the nonlinear Hawkes process (Equation 3) may be written compactly as

$$\lambda(t; \mathcal{T}, \theta) = F\left(s(t; \mathcal{T}, \theta)\right) \quad s(t; \mathcal{T}, \theta) = \langle \theta, \Phi(t; \mathcal{T}) \rangle, \quad (21)$$

where $\theta = \left(\mu, \alpha_1, \ldots, \alpha_L\right) \in \mathbb{R}^{L+1}$. That is, the intensity is a nonlinear transform of a linear scoring model. For a linear Hawkes process, the parameters θ must be nonnegative.

Canonical losses for nonlinear Hawkes processes

Equation 21 lets us interpret the negative log-likelihood for the nonlinear Hawkes process (Equation 5) as an instance of the generalised IPP objective (Equation 17). For the linear scorer class $\{s(\cdot; \theta) = \langle \theta, \Phi(\cdot; \mathcal{T}) \rangle \mid \theta \in \Theta\}$, we have

$$\mathcal{L}_{\text{HP}}(\theta; \mathcal{T}) = \sum_{n=1}^{N} \ell(+1, \langle \theta, \Phi(t_n) \rangle) + \int_0^T \ell(-1, \langle \theta, \Phi(t) \rangle) \, \mathrm{d}t \quad (22)$$

for the loss $\ell(+1, v) = -\log F(v)$, $\ell(-1, v) = F(v)$, which is evidently strictly proper composite with link

$$\Psi^{-1}(v) = F(v)/(1 + F(v)), \quad (23)$$

recalling that F is invertible (A1). Thus, by Equation 18, we model $\lambda = F \circ s$ as per the Hawkes intensity (Equation 21).

Inspired by the previous section, suppose we wish to fit a nonlinear Hawkes process with link $F(\cdot)$. Then, a viable alternative is to replace ℓ in Equation 22 with any strictly proper composite loss having inverse link as per Equation 23, since this retains the fundamental target of intensity $\lambda = F \circ s$. One simple choice is the *canonical* proper loss for this choice of Ψ^{-1} (Buja, Stuetzle, and Shen 2005), *viz.*

$$\ell(-1, v) = \int_{c_0}^{v} \frac{F(x)}{1 + F(x)} \, \mathrm{d}x \qquad \ell(+1, v) = \ell(-1, v) - v \quad (24)$$

for a suitable c_0 guaranteeing finiteness of the integral. Importantly, unlike the standard likelihood, this loss is guaranteed to be convex, regardless of the choice of F.

To make this idea concrete, we study its application for three distinct $F(\cdot)$. Our results are summarised in Table 2.

$F(z)$	$\ell(+1, v)$	$\ell(-1, v)$
z	$-v$	$1/2 \cdot v^2$
$\exp(z)$	$\log(1 + \exp(v)) - v$	$\log(1 + \exp(v))$
$(1 + \exp(-z))^{-1}$	$\log(1 + 2 \cdot \exp(v)) - 2 \cdot v$	$\log(1 + 2 \cdot \exp(v))$

Table 2: Proposed losses for various nonlinearities $F(\cdot)$.

Fitting linear Hawkes via the LSIF loss

Consider first the case of a linear Hawkes process, with $F(x) = x$. Despite the convexity of the log-likelihood in θ, a viable alternative has a salient feature: consider

$$(\forall v > 0) \, \ell(+1, v) = -v \qquad \ell(-1, v) = 1/2 \cdot v^2, \quad (25)$$

which is strictly proper composite with $\Psi^{-1}(v) = v/(1 + v)$ (Menon and Ong 2016, Lemma 1), as required by Equation 23. This was studied as the *least squares importance filtering* (*LSIF*) loss by (Kanamori, Hido, and Sugiyama 2009), who showed that for a linear model, as in Equation 22, the loss *potentially* admits a *closed-form solution*

$$\theta^* = (N/T) \cdot \left(\mathop{\mathbb{E}}_{\mathsf{T}' \sim Q} \left[\Phi(\mathsf{T}') \, \Phi(\mathsf{T}')^{\mathrm{T}} \right] \right)^{-1} \cdot \mathop{\mathbb{E}}_{\mathsf{T} \sim \widehat{P}} \left[\Phi(\mathsf{T}) \right]. \quad (26)$$

This is only "potentially" the minimiser as for the linear Hawkes process, the parameters θ are required to be non-negative. There is no guarantee that this holds for Equation 26; as a heuristic, one may threshold the weights at zero, and use this as an initialisation for explicit risk optimisation.

The above closed-form solution has clear conceptual appeal. It is also fast to compute for small L, requiring $\mathcal{O}(N + L^3)$ complexity, where the second term arises from the matrix inverse. We caution however that in fitting a Hawkes process, even computing Φ in Equation 20 naïvely require $\mathcal{O}(N^2)$ time, while the optimisation itself will be $\mathcal{O}(N)$ for standard gradient-following procedures. Nonetheless, the ease of computing Equation 26 suggests it may be minimally useful as an *initialisation* to the standard MLE optimisation.

Such least-squares style loss functions have in fact previously appeared in the Hawkes literature (Reynaud-Bouret and Schbath 2010; Bacry, Gaïffas, and Muzy 2015), albeit derived from very different means. Interestingly, (Reynaud-Bouret and Schbath 2010) establish that with suitable regularisation, the finite-sample minimiser of the objective is *consistent*, as with the standard MLE. We emphasise that such a loss is only one special case of our framework.

Fitting exponential Hawkes via the logistic loss

Consider now the case of $F(x) = \exp(x)$, which we term an *exponential Hawkes process*. By a simple calculation,

$$\int_{-\infty}^{v} \frac{F(x)}{1 + F(x)} \, \mathrm{d}x = \log(1 + \exp(v)).$$

Thus, the corresponding canonical loss is

$$\ell(+1, v) = \log(1 + e^v) - v \qquad \ell(-1, v) = \log(1 + e^v), \quad (27)$$

viz. the logistic loss underpinning logistic regression. Since our underlying scorer is linear (Equation 21), this suggests

that an exponential Hawkes process can be fit via standard logistic regression. In fact, a stronger statement is possible. Given observed and background events $\{t_n\}_{n=1}^N$, $\{t_m\}_{m=1}^M$, let us fit $\theta = (\mu, \alpha_1, \ldots, \alpha_L)$ via the *weighted* logistic risk,

$$\sum_{n=1}^{N} \log(1+\exp(-s(t_n;\theta))+w\cdot \sum_{m=1}^{M} \log(1+\exp(s(t_m;\theta))) \quad (28)$$

for $w > 0$. Then, as $w \to +\infty$, $(\alpha_1, \ldots, \alpha_L)$ will converge *exactly* to those that optimise the log-likelihood of the exponential Hawkes process. The reason is simple: (Fithian and Hastie 2013) showed that fitting of a log-linear IPP model, $\lambda(t) = \exp(a + \langle b, \Phi(t)\rangle)$, is *equivalent* to a weighted logistic regression objective, in the limit of an infinite weight on the negative class. As the Hawkes likelihood is equivalent to that of an IPP with a particular intensity, for sufficiently large w, *fitting an exponential Hawkes process is equivalent to weighted logistic regression, even on a finite sample*.

Fitting sigmoidal Hawkes via modified logistic loss

Our last example is a sigmoid nonlinearity $F(z) = a \cdot (1 + \exp(-z))^{-1}$ for any $a \notin \{-1, 0\}$. For this link, the likelihood objective is non-convex, since $F(\cdot)$ is. The canonical proper composite loss is however convex. By a simple calculation,

$$\int_{-\infty}^{v} \frac{F(x)}{1+F(x)}\,\mathrm{d}x = \frac{a}{1+a} \cdot \log(1 + (1+a)\cdot\exp(v)).$$

Thus, the corresponding canonical loss is (by rescaling)

$$\begin{aligned} \ell(+1, v) &= \log(1 + (1+a)\cdot\exp(v)) - (1+a)/a \cdot v \\ \ell(-1, v) &= \log(1 + (1+a)\cdot\exp(v)). \end{aligned} \quad (29)$$

When $a = 1$, we get the form as shown in Table 2. This can be seen as a modified version of the logistic loss, and its convexity makes it appealing compared to the MLE.

Discussion and related work

A subtlety in the Hawkes process is that replacing the empirical $\widehat{P}$ with P is delicate, as the event times are no longer *i.i.d*. Our framework nonetheless produces meaningful results for the LSIF and logistic loss (confer existing consistency analysis and finite-sample equivalence).

Assumption A1 ensures the function $F(v)/(1 + F(v))$ is invertible; without this one cannot have a valid link for a proper composite loss. Assumption A2 does not obviate the use of proper losses; however, if the kernel parameters are used in a nonlinear manner, then neither the closed-form LSIF solution nor the reduction to vanilla logistic regression is viable. The objectives are nonetheless viable to optimise with generic nonlinear solvers; alternately, one could iterate between optimising the kernel parameters and θ.

To our knowledge, the only existing closed-form solution for the linear Hawkes process is as part of the EM algorithm with an exponential decay function (Lewis and Mohler 2011; Zipkin et al. 2016). By contrast, the closed-form LSIF solution holds for *any* decay of the form in Equation 19.

The reduction of exponential Hawkes fitting to logistic regression is in the spirit of prior connections of IPP optimisation to established statistical techniques, such as Poisson regression (Berman and Turner 1992) and the MAXENT procedure (Renner and Warton 2013).

Literature on fitting of nonlinear Hawkes processes has been relatively sparse. An interesting recent exception is the work of (Wang et al. 2016), who proposed an algorithm that jointly estimates the parameters θ and the link $F(\cdot)$. At its core is the elegant Isotron algorithm for fitting single index models (Kalai and Sastry 2009). Despite its generality, the algorithm suffers from worse sample complexity compared to standard model fitting where $F(\cdot)$ is assumed known.

Empirical illustration

We validate our theoretical analyses by illustrating the viability of using losses other than the standard maximum likelihood to fit various (non-)linear Hawkes processes.

Parameter recovery on synthetic data

We first assess the new loss functions in a controlled setting.

Basic setup For fixed N, we use thinning (Ogata 1981) to generate N samples from a nonlinear Hawkes process with known link $F(\cdot)$, and exponential decay $g(\cdot)$ with $\delta = 1$ and known parameters $\theta^* = (\mu, \alpha)$ (Equation 4). From these samples, we compute a parameter estimate $\hat{\theta}$ via maximum likelihood estimation (**MLE**) – i.e. optimising Equation 2 – and an alternate loss to be specified. Given this $\hat{\theta}$, we compute the mean absolute error (MAE) $1/2 \cdot \|\hat{\theta} - \theta^*\|_1$. We repeat this for 1000 independent samples from the process.

Link & loss We consider the three Hawkes links $F(\cdot)$ of the prequel, and for each compare the MLE to their proper composite alternatives; Table 3 summarises. For the logistic loss, we apply a weighting $w = 10^8$ on the background class, following Equation 28; for other losses, we set $w = 1$.

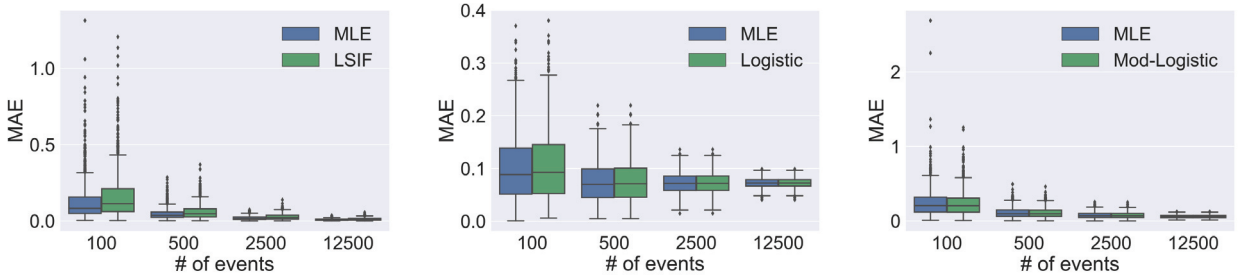
The parameters θ^* are varied for each link, with precise settings derived from previous studies (Ozaki 1979; Wang et al. 2016); Since $\alpha < 0$ for the exponential and sigmoidal link, the corresponding processes exhibit self-inhibition.

Link $F(z)$	θ^*	**Comparison loss**	**Reference**
z	$(0.5, 0.8)$	LSIF	Eqn. 25
$\exp(z)$	$(0.5, -0.1)$	Weighted logistic	Eqn. 27
$(1 + \exp(-z))^{-1}$	$(0.5, -0.1)$	Modified logistic	Eqn. 29

Table 3: Parameter settings for synthetic data.

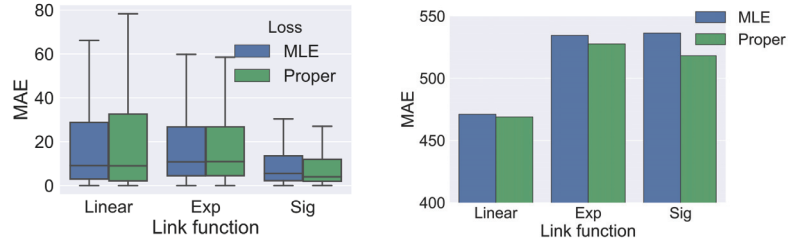
We optimised the MLE and (modified) logistic loss with L-BFGS, enforcing a stationarity constraint that $\alpha < \delta$ and a tolerance criterion of 10^{-8}. For the LSIF loss, we use the closed-form solution of Equation 26.

Results Figure 2 confirms that the proper loss solutions have commensurate accuracy to the MLE for various N. This reassures that when the Hawkes process is well specified, these losses behave sensibly, and indeed recover the optimal parameters asymptotically. Of note is that for the exponential link, the MLE and weighted logistic solution are indistinguishable, as predicted by the theory.



(a) Linear. (b) Exponential. (c) Sigmoid.

Figure 2: MAE for Hawkes parameter estimation with various links. Shown are boxplots over 1000 independent samples.



(a) `LastFM`. Shown are boxplots over 500 (user, artist) pairs.

(b) `bitcoin`. Shown is a barplot for a single trading day.

Figure 3: Absolute error for various nonlinear Hawkes processes trained with MLE and proper loss on real-world datasets.

Event prediction on real data

We compare the various losses on two real-world datasets: `LastFM` (Celma 2010), comprising the times that users listen to songs by artists, where Hawkes processes have previously been applied for temporal recommendation (Du et al. 2015); and `bitcoin`, comprising times of trades on the MtGox Bitcoin exchange for a single day (Heusser 2013). For each dataset, we aim to predict the number of events happening in some specified future time window.

Basic setup For each dataset, we split the recorded event times into a train and test set. We fit various Hawkes processes with an exponential decay ($\delta = 1$ on `LastFM`, $\delta = 0.1$ on `bitcoin`) on the training times. By drawing 100 independent samples from the learned process, we estimate the number of events in the testing period. We compute the absolute error between this prediction and the ground truth.

The precise split methodology varies for each dataset. For `LastFM`, we select 500 random (user, artist) pairs for which there are at least 100 listening events over the span of at least two months. For each pair, we define the testing period to be the last month of the recorded history.

For `bitcoin`, we use all trades occurring in the window 1PM - 3PM for training, and make predictions in the window 3PM - 4PM. This dataset only has a single day's worth of trading, so we simply report the absolute error for this day.

Results Figures 3a and 3b confirm that even on real-world data, the performance of the MLE and the corresponding proper loss are largely indistinguishable. Of interest is that for the sigmoid link, the proper loss solution offers consistent (albeit statistically insignificant) improvement over the MLE, possibly owing to the non-convexity of the latter. We further confirm that for the exponential link, the MLE and logistic regression solution are practically identical.

Overall, we find our proper loss objectives produce sensible results on synthetic and real-world datasets.

Conclusion

We presented a new family of losses to train (non-)linear Hawkes processes, giving three concrete examples: for linear Hawkes processes, we provided a least-squares style loss with a closed-form solution; for exponential Hawkes processes, we showed how training can be reduced to weighted logistic regression; and for sigmoidal Hawkes processes, we proposed a modified form of logistic regression.

There are several directions for future work. On the theoretical end, translating conditions on $F(\cdot)$ that guarantee *stationarity* (Brémaud and Massoulié 1996; Karabash 2012) to conditions on the corresponding canonical proper loss would be of interest. On the practical end, a more detailed empirical study, extensions to stochastic excitations (Lee, Lim, and Ong 2016), multivariate Hawkes processes (Bacry, Mastromatteo, and Muzy 2015), and exploring possible uses of the closed form LSIF solution in conjunction with EM algorithm (Veen and Schoenberg 2008), would be of interest.

References

Bacry, E.; Gaïffas, S.; and Muzy, J.-F. 2015. A generalization error bound for sparse and low-rank multivariate Hawkes processes. *ArXiv e-prints*.

Bacry, E.; Mastromatteo, I.; and Muzy, J.-F. 2015. Hawkes processes in finance. *Market Microstructure and Liquidity* 01(01).

Berman, M., and Turner, T. R. 1992. Approximating point process likelihoods with glim. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 41(1):31–38.

Bowsher, C. 2007. Modelling security market events in continuous time: Intensity based, multivariate point process models. *Journal of Econometrics* 141(2):876 – 912.

Brémaud, P., and Massoulié, L. 1996. Stability of nonlinear Hawkes processes. *Annals of Probability* 24(3):1563–1588.

Buja, A.; Stuetzle, W.; and Shen, Y. 2005. Loss functions for binary class probability estimation and classification: Structure and applications. Technical report, UPenn.

Celma, O. 2010. *Music Recommendation and Discovery in the Long Tail*. Springer.

Cox, D. R., and Isham, V. 1980. *Point Processes*. Chapman & Hall.

Crane, R., and 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.

Daley, D., and Vere-Jones, D. 2003. *An introduction to the theory of point processes. Vol. I*. Springer-Verlag, second edition.

Davis, P., and Rabinowitz, P. 1984. *Methods of Numerical Integration*. Academic Press.

Diggle, P. 1985. A kernel method for smoothing point process data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 34(2):138–147.

Du, N.; Wang, Y.; He, N.; and Song, L. 2015. Time-sensitive recommendation from recurrent user activities. In *Advances in Neural Information Processing Systems*.

Erlang, A. K. 1909. The theory of probabilities and telephone conversations. *Nyt Tidsskrift for Matematik B* 20(B):33–39.

Fithian, W., and Hastie, T. 2013. Finite-sample equivalence in statistical models for presence-only data. *Annals of Applied Statistics* 7(4):19171939.

Gneiting, T., and Raftery, A. E. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 102(477):359–378.

Hall, E. C., and Willett, R. M. 2016. Tracking dynamic point processes on networks. *IEEE Transactions on Information Theory* 62(7):4327–4346.

Hardiman, S. J.; Bercot, N.; and Bouchaud, J.-P. 2013. Critical reflexivity in financial markets: a Hawkes process analysis. *European Physics Journal B* 86(10):442.

Hastie, T.; Tibshirani, R.; and Friedman, J. 2009. *The Elements of Statistical Learning*. Springer, 2nd edition.

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

Heusser, J. 2013. Hawkes process estimation. https://github.com/jheusser/hawkes.

Kalai, A., and Sastry, R. 2009. The Isotron algorithm: High-dimensional isotonic regression. In *Conference on Learning Theory (COLT)*.

Kanamori, T.; Hido, S.; and Sugiyama, M. 2009. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research* 10:1391–1445.

Karabash, D. 2012. On Stability of Hawkes Process. *ArXiv e-prints*.

Kendall, D. 1953. Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded markov chain. *Annals of Math. Stat.* 24(3):338–354.

Kingman, J. 1993. *Poisson Process*. Oxford University Press.

Laub, P. J.; Taimre, T.; and Pollett, P. K. 2015. Hawkes Processes. *ArXiv e-prints*.

Lee, Y.; Lim, K. W.; and Ong, C. S. 2016. Hawkes processes with stochastic excitations. In *International Conference on Machine Learning, (ICML)*.

Lewis, E., and Mohler, G. 2011. A nonparametric EM algorithm for multiscale Hawkes processes. http://math.scu.edu/~gmohler/EM_paper.pdf.

Menon, A. K., and Ong, C. S. 2016. Linking losses for density ratio and class-probability estimation. In *International Conference on Machine Learning*.

Ogata, Y. 1981. On Lewis' simulation method for point processes. *IEEE Trans. Inf. Theor.* 27(1):23–31.

Ogata, Y. 1988. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association* 83:9–27.

Ozaki, T. 1979. Maximum likelihood estimation of Hawkes' self-exciting point processes. *Annals of the Institute of Statistical Mathematics* 31(1):145–155.

Reid, M. D., and Williamson, R. C. 2010. Composite binary losses. *Journal of Machine Learning Research* 11:2387–2422.

Renner, I. W., and Warton, D. I. 2013. Equivalence of maxent and poisson point process models for species distribution modeling in ecology. *Biometrics* 69(1):274–281.

Reynaud-Bouret, P., and Schbath, S. 2010. Adaptive estimation for Hawkes processes; application to genome analysis. *Annals of Statistics* 38(5):2781–2822.

Steinwart, I.; Hush, D.; and Scovel, C. 2006. A classification framework for anomaly detection. *Journal of Machine Learning Research* 6:211232.

Veen, A., and Schoenberg, F. P. 2008. Estimation of space–time branching process models in seismology using an EM–type algorithm. *Journal of the American Statistical Association* 103(482):614–624.

Wang, Y.; Xie, B.; Du, N.; and Song, L. 2016. Isotonic Hawkes processes. In *International Conference on Machine Learning*.

Xu, H.; Farajtabar, M.; and Zha, H. 2016. Learning granger causality for hawkes processes. In *International Conference on Machine Learning*.

Zhou, K.; Zha, H.; and Song, L. 2013. Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes. In *Conference on Artificial Intelligence and Statistics (AISTATS)*, 641–649.

Zipkin, J. R.; Schoenberg, F. P.; Coronges, K.; and Bertozzi, A. L. 2016. Point-process models of social network interactions: Parameter estimation and missing data recovery. *European Journal of Applied Mathematics* 27(3):502529.