Goodness-of-Fit Test for Mismatched Self-Exciting Processes

Song Wei Shixiang Zhu Minghe Zhang Yao Xie
{song.wei,shixiang.zhu,minghe_zhang}@gatech.edu yao.xie@isye.gatech.edu
H. Milton Stewart School of Industrial and Systems Engineering,
Georgia Institute of Technology, Atlanta, Georgia, USA

Abstract

Recently there have been many research efforts in developing generative models for self-exciting point processes, partly due to their broad applicability for real-world applications. However, rarely can we quantify how well the generative model captures the nature or ground-truth since it is usually unknown. The challenge typically lies in the fact that the generative models typically provide, at most, good approximations to the ground-truth (e.g., through the rich representative power of neural networks), but they cannot be precisely the ground-truth. We thus cannot use the classic goodness-of-fit (GOF) test framework to evaluate their performance. In this paper, we develop a GOF test for generative models of self-exciting processes by making a new connection to this problem with the classical statistical theory of Quasi-maximum-likelihood estimator (QMLE). We present a non-parametric self-normalizing statistic for the GOF test: the Generalized Score (GS) statistics, and explicitly capture the model misspecification when establishing the asymptotic distribution of the GS statistic. Numerical simulation and real-data experiments validate our theory and demonstrate the proposed GS test's good performance.

1 Introduction

Self- and mutual- exciting point processes, as known as the Hawkes processes, are introduced by the original papers by Hawkes (1971a,b); Hawkes and Oakes (1974). They become popular in machine learning due to their wide applicability in modeling triggering effect in discrete event data, which is ubiquitous in modern applications ranging from seismology (Ogata, 1988, 1999; Zhuang, 2011), infectious disease modeling (Meyer and Held, 2014; Schoenberg et al., 2019), crime events (Mohler et al., 2011), wildfire occurrence (Peng et al., 2005), civilian deaths in Iraq (Lewis et al., 2012), terrorist activity forecasting (Porter and White, 2012), social network analysis and so on.

Classical Hawkes processes are largely parametric, which focus on modeling the conditional intensity function of the point process (since the conditional intensity function completely specifies the distribution of the process). Hawkes process assumes that the intensity function consists of the sum of a deterministic background intensity (which can be time-varying) and a stochastic term, which captures the influence from the past events. It is common to assume that the influence from past events is additive, and the so-called *triggering function* measures an individual event's influence. One key problem in the Hawkes process is to specify the triggering kernel. Popular parametric triggering functions include exponential kernel, power kernel, and Matérn kernel (Reinhart, 2018).

When facing more complex data with complex temporal triggering patterns, parametric models can become too restrictive and even mis-specified. Thus, recently, there have been many efforts in developing more general generative models for point processes, including probability weighted kernel estimation with adaptive bandwidth (Zhuang et al., 2002), probability weighted histogram estimation (Marsan and Lengline, 2008) and with inhomogeneous spatial background rate (Fox et al., 2016) and neural Hawkes process (Mei and Eisner, 2017).

Since the specified models (including those generative models) are very likely to be incorrect due to the ignorance of the ground-truth, a natural and important question yet to be answered is which model to select in practice. Here, we proposed to use how well those models capture the data, i.e. *goodness-of-fit* of these

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

Hawkes process models, as the metric to rank models in practice. As well-said in Engle (1984): *"At any stage in the specification search, it may be desirable to determine whether an adequate representation of the data has been achieved."* For generative models, since they tend to be further away from the probabilistic framework of Hawkes processes, it is more difficult to evaluate their GOF to the real data. For these generative models, the classic statistical GOF test framework may not apply.

There are two major difficulties in utilizing existing GOF tests for the self-exciting point processes. (1) Generative models typically provide, at most, good approximations to the ground-truth (e.g., through the rich representative power of neural networks), but they cannot be precisely the ground-truth. For instance, it is unlikely that neural networks truly specify the data distribution; rather, the neural networks are being used because of their universal approximation power and can generate a good approximation to the ground-truth (Mei and Eisner, 2017). Typically, it is impractical to access the GOF via testing with unknown ground-truth g^*, as illustrated in the left panel in Figure 1. In both theory and practice, the best we can do is to test how close our fitted model $\widehat{g}$ is to the approximation g_0, as illustrated in the middle panel in Figure 1. Nevertheless, we still consider model misspecification explicitly since it is vital in establishing the asymptotic performance of our proposed GOF test. (2) When we fit conditional intensities from various families, direct comparison between GOF measures from different families is not reasonable; we need to find a unifying space to access comparable GOF measures for all considered families, as illustrated by red lines in the right panel in Figure 1. This space, or rather function family, for GOF, should be carefully chosen such that it is both expressive enough and not too complex to develop a valid, consistent, and tractable test statistic thereon.

GOF test for the whole conditional intensity has been developed by Ogata (1988); Schoenberg (2003), but the theory therein is established under the classic set-up and may fail to generalize to model misspecification setting. Moreover, the triggering effect is the main effect-of-interest in many Hawkes process models since (1) it characterizes the dynamics between events and (2) the background rate can be separately estimated from the well-established declustering procedure (Zhuang et al., 2002; Marsan and Lengline, 2008; Fox et al., 2016). However, the background rate usually dominates the conditional intensity, and the existing tests may not detect subtle triggering function differences. Thus, a principled method to quantify the *goodness-of-fit for triggering effect in Hawkes processes under model misspecification* is essential.

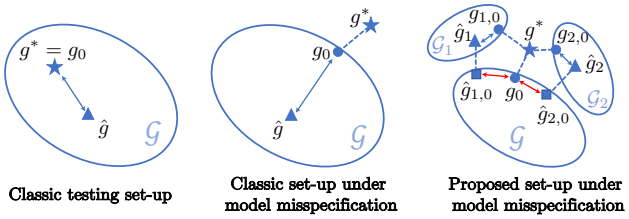


Figure 1: The ground-truth is g^*; the assumed/specified family of candidate models is $\mathcal{G}$ in the left two panels and $\mathcal{G}_1$ and $\mathcal{G}_2$ in the right panel. GOF addresses how close the fitted model $\widehat{g}$ is to the unknown true one g^*. In classic set up, one assumes there exists a $g_0 \in \mathcal{G}$ such that $g^* = g_0$. Under a more general model misspecification case where g^* may not be contained in $\mathcal{G}$, classic GOF measures the distance between $\widehat{g}$ and a good approximate $g_0 \in \mathcal{G}$ to g^* (in K-L divergence sense). When we want to rank models, we need to find the approximation of $\widehat{g}_1$, $\widehat{g}_2$, and g^* in a unifying space $\mathcal{G}$ and compare models $\widehat{g}_1$, $\widehat{g}_2$ therein.

Contribution. In this paper, we present a nonparametric goodness-of-fit (GOF) test statistic, called the *Generalized Score* (GS), which can be broadly applied to evaluating the self-exciting part in Hawkes process generative models. The GS test is constructed by translating the GOF test into a two-sample test: whether the real data and synthetic data from the generative model have the same distribution? Based on this, we derive the likelihood score statistic with estimated piecewise constant kernels, which is flexible and has little model restrictions. We further establish asymptotic properties for MLE of the Quasi-model (QMLE), asymptotic χ^2 null distribution, as well as the power function of GS statistic. The main ingredients of our analysis include (1) making a connection between GS test and the classic theory on MLE under model misspecification (QMLE) (White, 1982) and (2) generalizing the asymptotic properties of MLE of Hawkes process in Ogata (1978) to model misspecification case. Our GS test provides a tool for model diagnosis and comparison of the self-exciting part in Hawkes process generative models. We demonstrate the effectiveness of our proposed test via numerical simulation and real-data examples.

Several features of our GS test include: (1) We develop the test for generative models considering their inherent "model misspecification nature"; (2) we focus on GOF of the triggering effect in Hawkes process models; (3) due to its construction, the GS statistic enjoys simple asymptotic distribution specified by χ^2 distribution and analytical form of the power function, which enables us to calibrate the test without sampling.

Related Work. The one-sample goodness-of-fit prob-

lem is closely related to the two-sample test problem. For independent and identically distributed (i.i.d.) observations, two-sample test is well studied (e.g. energy statistic (Székely and Rizzo, 2004; Baringhaus and Franz, 2004) and maximum mean discrepancy (MMD) (Gretton et al., 2012)) and so is the GOF based on it. Chwialkowski et al. (2016) developed Stein operator based MMD (which they call squared Stein discrepancy) and changed the two-sample test statistic to a one-sample GOF test metric. Bounliphone et al. (2015) reformulated the one-sample GOF problem into a two-sample test problem and developed a model selection tool based on MMD. Extension of those methods to point process is missing until Yang et al. (2019) proposed a kernel goodness-of-fit test by defining a Stein discrepancy for generic point process; However, a common drawback of a kernel-based test is that the null distribution is hard to evaluate (since they depend on infinite series involving the eigenvalues of the kernel). In contrast, our GS statistic follows a simple χ^2 null distribution and is easy to calibrate. Our proposed method allows the distribution under the null to be flexible and estimated from data by comparing the data to the generative model via the test statistic. Other model diagnostics include likelihood of fitted model and the observed data (Schorlemmer et al., 2007) and Information Criterion (IC) (Chen et al., 2018). The likelihood is the most commonly used, but overfitting makes it less convincing and even questionable (as discussed via numerical simulation). Chen et al. (2018) assumed correct model specification, which typically does not hold in the real study, and the consistency result of IC is restricted to exponential triggering function case. For more on the kernel-based two-sample test as well as model diagnosis and selection method of the point process, one can refer to Harchaoui et al. (2013) and Bray and Schoenberg (2013).

2 Problem set-up

We first introduce some necessary mathematical preliminaries, and then formulate the one-sample goodness-of-fit problem into a two-sample test problem.

2.1 Mathematical background

Consider a counting process $\{N(t) : t \geq 0\}$, with associated history $\mathcal{H}_{0,t} = \{t_i : 0 < t_i < t\}$ $(t \geq 0)$ indicating the occurrence time of a sequence of discrete events. For simplicity, we use $\mathcal{H}_t$ instead. A point process is characterized by its conditional intensity, which is defined as:

$$\lambda\left(t|\mathcal{H}_t\right) = \lim_{\Delta t \downarrow 0} \mathbb{E}\left[N\{(t, t+\Delta t)\}|\mathcal{H}_t\right]/\Delta t.$$

Hawkes process is a self-exciting point process with conditional intensity takes the following form:

$$\lambda\left(t|H_t\right) = \mu + \sum_{\{i:t_i<t\}} \phi\left(t - t_i\right), \tag{1}$$

where μ is called the background intensity and $\phi : (0,\infty) \to [0,\infty)$ is called the triggering function.

We assume the separability of triggering function into components for magnitude and time: $\phi\left(t - t_i\right) = \alpha g(t - t_i)$, where temporal triggering function g is a probability density function (p.d.f.) and α represents the magnitude of triggering effect, i.e. how many subsequent events one event can trigger on average. Given the past trajectory $\mathcal{H}_T$ with N events, the log-likelihood over time interval $[0, T]$ can be expressed as:

$$\ell(\theta) = \sum_{i=1}^{N} \log\left(\lambda(t_i|\mathcal{H}_{t_i})\right) - \int_0^T \lambda(u|\mathcal{H}_u)\mathrm{d}u.$$

One can refer to Laub et al. (2015) and Reinhart (2018) for a more comprehensive introduction of Hawkes process and a detailed deviation of its (log-)likelihood function.

2.2 Problem formulation

Suppose we have two data sequences $D_z = (t_1^{(z)}, \ldots t_{N_z}^{(z)})$, $(z = 1, 2)$, which represent the arrival times of a sequence of events. Here, D_1 is from real world and D_2 is generated from the fitted generative model. Assume $D_1 \sim \lambda^*$ and $D_2 \sim \lambda$, where λ^* is the unknown true conditional intensity and $\widehat{\lambda}$ is the fitted one. Further assume both conditional intensities take form in (1). We aim to test

$$H_0' : \lambda^* = \widehat{\lambda}, \quad \text{versus} \quad H_1' : \lambda^* \neq \widehat{\lambda}.$$

Note that λ^* in the above formulation is unknown. As illustrated in Figure 1, we cast the problem above into testing $H_0 : \lambda_0^* = \widehat{\lambda}_0$ by projecting the unknown ground-truth onto a piecewise constant function family $\mathcal{G}$, on which we can develop a tractable goodness-of-fit test statistic. Empirically, this projection is done by mixing D_1 and D_2 and fitting a piecewise constant triggering function to the mixed data. Most importantly, when we have several candidate models, this statistic serves as a quantitative metric to compare models.

We calculate this test statistic in the following three steps: Mix the two data sequences up to get an aggregated sequence; Estimate θ_0, maximizer of the Quasi-likelihood, from a Quasi-parameter space Θ for the aggregated sequence; Compute a test statistic $\widehat{GS}_T$ based on the estimation in the last step.

Remark 1 (Singleton null). In our setting, the triggering function's unknown parameter is infinite-dimensional, so the null hypothesis H_0 is an uncountable set. To make the problem tractable, we cast H_0' to H_0 by representing the unknown triggering function using some basis function (in our case, we use indicator function on mutually disjoint intervals (2)) such that we reduce this into a finite-dimensional problem. Besides, testing with unknown λ^* is impractical, and we can only handle the projected problem H_0 to draw the inference for H_0' anyways.

Remark 2 (Model mismatch). We use the term "Quasi" here since commonly speaking, there will be a mismatch between a machine learning algorithm class we specify and the unknown true intensity, i.e., this class is misspecified as illustrated in Figure 1. We add a prefix "Quasi-" for everything under this class, e.g., Quasi-conditional intensity and Quasi-likelihood function. Since conditional intensity characterizes a point process and we assume the triggering function $\phi^* = \alpha g^*$, we only need to specify the approximate class for g^*. We choose a piecewise constant function class as $\mathcal{G}$. The reason is three-fold: (i) a piecewise constant function can approximate any integrable function arbitrarily well by reducing the size of the discretization bin; (ii) there exists $g_0 \in \mathcal{G}$, which corresponds to our estimand θ_0, serving as a good approximation to g^* and it is identifiable; (iii) most importantly, we can develop an easy-to-calibrate hypothesis test on this family. We will elaborate on these in the next section.

3 Proposed goodness-of-fit test

The idea behind this test comes from a critical observation that under H_0 (or H_0'), mixing two sequences will lead to a Hawkes process with scaled intensity function. Based on this observation, we can derive a Generalized Score (GS) test, which is known to be locally most powerful (Neyman–Pearson lemma).

Step 1: Mix two data sequences and model the aggregated sequence.

In this step, we derive the Quasi-log-likelihood function for the aggregated sequence. The proof is deferred to Appendix A.

Proposition 1 (Log-likelihood of mixing of two Hawkes processes). *Suppose we have two Hawkes processes with conditional intensities*

$$\lambda_z(t|\mathcal{H}_{z,t}) = \mu^{(z)} + \sum_{\{i:t_i^{(z)}<t\}} \phi^{(z)}(t - t_i^{(z)}) \ (z = 1, 2).$$

Define their mixing to be $N(t) = N_1(t) + N_2(t)$. Then it has background intensity $\mu = \mu^{(1)} + \mu^{(2)}$. Denote $T = \max\{t_{N_1}^{(1)}, t_{N_2}^{(2)}\}$ and $\Phi^{(z)}(t) = \int_0^t \phi^{(z)}(u)\mathrm{d}u$. Given the past trajectory: $\mathcal{H}_t = \mathcal{H}_{1,t} \cup \mathcal{H}_{2,t}$, where $\mathcal{H}_{z,t} = \{t_1^{(z)}, \ldots, t_{N_z}^{(z)}\}$, $z = 1, 2$, we have that: (i) Under H_1, let $z' = 2(or\ 1)$ when $z = 1(or\ 2)$, the full model log-likelihood $\ell_1(\mu, \phi^{(1)}, \phi^{(2)}|\mathcal{H}_t)$ is

$$-\mu T + \sum_{z=1}^{2}\sum_{i=1}^{N_z} \log\Big(\mu + \sum_{j<i} \phi^{(z)}(t_i^{(z)} - t_j^{(z)}) + \sum_{j=1}^{N_{z'}} \phi^{(z')}(t_i^{(z)} - t_j^{(z')})\Big) - \Phi^{(z)}(T - t_i^{(z)}).$$

(ii) Under $H_0 : \phi^{(1)} = \phi^{(2)} = \phi$, the sub-model log-likelihood is $\ell_0(\mu, \phi|\mathcal{H}_t) = \ell_1(\mu, \phi, \phi|\mathcal{H}_t)$.

Note that the triggering function takes value zero on $(-\infty, 0]$ and thus we did not consider the triggering effect of events to its own history. By this proposition, we can model the aggregated data via a univariate Hawkes process with the same triggering function under H_0. For each event in process z ($z = 1, 2$), it does not only dependent its original own history, but also depends on the history of another process z' ($z' = 2, 1$). See an illustration of this in Figure 2.

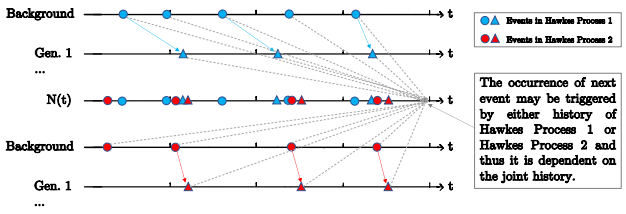


Figure 2: Illustration of mixing of two Hawkes processes $N(t)$. Given the past sample trajectory, the upcoming event of $N(t)$ may (1) come from background poisson process of Hawkes process 1 or 2 OR (2) be a offspring of history $\mathcal{H}_{1,t}$ or $\mathcal{H}_{2,t}$. The grey dashed line in the figure illustrated scenario (2).

Step 2: Discretize triggering function and learn quasi-conditional intensity.

In this step, we choose piecewise constant function as the approximation to the true triggering function for the aggregated sequence. This means we will discretize the time horizon into small intervals (which we call bins) and estimate a "weight" on each interval. In practice, the time horizon we discrete is truncated on $[0, T_0]$ and discretized into finitely many bins, since it is unnecessary to estimate infinite number of weights on infinite time horizon. More specifically, we assume $g_0(t) = \sum_{k=1}^{n_0} g_k \mathbf{1}_{B_k}(t)$ and estimate it from the following class:

$$\mathcal{G} \triangleq \Big\{ g(t) \ \Big| \ 0 \le g_k < \infty \text{ and } \sum_{k=1}^{n_0} g_k \Delta t_k = 1 \Big\}. \quad (2)$$

Here, $0 = \delta t_0 < \delta t_1 < \cdots < \delta t_{n_0} = T_0$ and each bin $B_k = (\delta t_{k-1}, \delta t_k]$ has length $\Delta t_k = \delta t_k - \delta t_{k-1} (k = 1, 2, \ldots, n_0)$.

We apply Probability Weighted Histogram Estimation (Marsan and Lengline, 2008; Fox et al., 2016) to learn the weights g_k on each bin B_k, triggering magnitude α and background intensity μ. Most importantly, our Quasi-conditional intensity defined in (2) satisfies the model assumption in Fox et al. (2016), which guarantees the non-parametric stochastic declustering algorithm as an EM algorithm. It maximizes a lower bound on the Quasi-log-likelihood function, which is in fact the complete-data Quasi-log-likelihood function derived by Veen and Schoenberg (2008). Thus, it outputs the MLE of Quasi-log-likelihood function (QMLE). See Appendix B for further details.

Before moving on, we need to formally define the estimand θ_0 we want to learn. It is the parameter of Quasi-conditional intensity which maximizes the expected Quasi-log-likelihood.

Definition 1 (Estimand). *The estimand θ_0 is*

$$\theta_0 = \arg\max_{\theta \in \Theta} \mathbb{E}[\ell_1(\theta|\mathcal{H}_T)], \tag{3}$$

where the expectation is w.r.t. all trajectories $\mathcal{H}_T$ and the expression of $\ell_1(\theta|\mathcal{H}_T)$ is given in Proposition 1.

Remark (Information theoretic interpretation). Here, $\theta \in \Theta$ in (3) has an information theoretic interpretation (Akaike, 1998). It parameterizes the Quasi-conditional intensity $\lambda = \lambda_\theta$ and θ_0 defined above corresponds to $\lambda_0 = \lambda_{\theta_0}$, which minimizes Kullback-Leibler (K-L) divergence to the unknown ground-truth λ^*:

$$\theta_0 = \arg\min_{\theta \in \Theta} \mathbb{E}[\ell^* - \ell_1(\theta|\mathcal{H}_T)] = \arg\min_{\theta \in \Theta} KL(\lambda^* || \lambda_\theta),$$

where ℓ^* is the true log-likelihood. That's why we call λ_0 "the best approximation to λ^*" or "projection onto the user-specified space" (as illustrated in Figure 1).

Proposition 2 (Global identifiability). *θ_0 defined by (3) is globally identifiable.*

We prove global identifiability by showing (3) is a (strictly) concave program. Most importantly, when the fitted model is actually the same as the unknown true one, θ_0 will lie in Θ_0, i.e. H_0 holds under H_0'. This justifies our projected test H_0, indicating that the difference between mismatched models represents the difference between true models. The detailed proof is deferred to Appendix D. We should make a mild assumption that θ_0 is interior to the convex Quasi-parameter space Θ. This makes sure that we have $\nabla_\theta \mathbb{E}[\ell_1(\theta_0|\mathcal{H}_T)] = 0$, which guarantees that θ_0 is the estimand which our QMLE is consistent for. We will show this in detail later in the Appendix D.

Step 3: Compute GS statistic.

Here, we call the singleton that we want to test a sub-model. We call the Quasi-parameter space under H_0 sub-model Quasi-parameter space and denote it by Θ_0. Similarly, Θ is the full model Quasi-parameter space, or rather, Quasi-parameter space under H_1. Under our proposed approximation class (2), the Quasi-conditional intensity has a parameterization

$$\theta = (\mu, \phi_1^{(1)}, \ldots, \phi_{n_0}^{(1)}, \phi_1^{(2)}, \ldots, \phi_{n_0}^{(2)})^\intercal,$$

where $\mu \triangleq \mu^{(1)} + \mu^{(2)}$ and $\phi_k^{(z)} = \alpha^{(z)} g_k^{(z)}$. The full model Quasi-parameter space is given by

$$\Theta = \Big\{ \theta \;\Big|\; \mu > 0, 0 \le \alpha^{(z)} < 1, \\ g^{(z)} = \sum_{k=1}^{n_0} g_k^{(z)} \mathbf{1}_{B_k} \in \mathcal{G} \;\; (z = 1, 2) \Big\} \subset \mathbb{R}^d,$$

where $d = 1 + 2n_0$. Note that the second constraint guarantees the stationarity and ergodicity. We further denote

$$\phi^{(z)} = (\phi_1^{(z)}, \ldots, \phi_{n_0}^{(z)})^\intercal = (\alpha^{(z)} g_1^{(z)}, \ldots, \alpha^{(z)} g_{n_0}^{(z)})^\intercal$$

to be the Quasi-parameter of the triggering function of Hawkes process z $(z = 1, 2)$. The sub-model Quasi-parameter space is

$$\Theta_0 = \{\theta \in \Theta \mid \phi_k^{(1)} - \phi_k^{(2)} = 0,\ k = 1, \ldots, n_0\} \subset \mathbb{R}^{1+n_0}.$$

Denote the number of constraints (we'll see later it's in fact degree-of-freedom of our test statistic) $r = n_0 = \dim \Theta - \dim \Theta_0$, the null hypothesis $H_0 : \theta_0 \in \Theta_0$ can be re-expressed as $H_0 : h(\theta_0) = \phi^{(1)} - \phi^{(2)} = 0$, where $h : \mathbb{R}^d \to \mathbb{R}^r$. We consider a test:

$$H_0 : h(\theta_0) = 0, \quad \text{versus} \quad H_1 : h(\theta_0) \neq 0,$$

and the following test statistic:

Definition 2 (GS statistic). *Suppose the past sample trajectory is $\mathcal{H}_T$. Denote*

$$S_T(\theta) = \frac{\partial \ell_1(\theta|\mathcal{H}_T)}{\partial \theta} \in \mathbb{R}^d, A_T(\theta) = S_T(\theta) S_T^\intercal(\theta) \in \mathbb{R}^{d \times d},$$

$$H(\theta) = \frac{\partial h(\theta)}{\partial \theta} \in \mathbb{R}^{r \times d}, \; B_T(\theta) = -\frac{\partial^2 \ell_1(\theta|\mathcal{H}_T)}{\partial \theta \partial \theta^\intercal} \in \mathbb{R}^{d \times d},$$

where H exists and has full row rank r, and log-likelihood ℓ_1 is given in Proposition 1. Then, the Generalized Score (GS) test statistic is given by

$$\widehat{GS}_T = S_T^\intercal(\widehat{\theta}_{QMLE}) \widehat{\Sigma}^{-1} S_T(\widehat{\theta}_{QMLE}),$$

where $\widehat{\theta}_{QMLE} \in \Theta_0$ is QMLE under null hypothesis and $\widehat{\Sigma}^{-1}$ is given by:

$$\widehat{\Sigma}^{-1} = B_T^{-1}(\theta) H(\theta)^\intercal \Big(H(\theta) B_T^{-1}(\theta) \\ A_T(\theta) B_T^{-1}(\theta) H(\theta)^\intercal \Big)^{-1} H(\theta) B_T^{-1}(\theta) \Big|_{\theta = \widehat{\theta}_{QMLE}}.$$

Later, we will show $T\widehat{\Sigma}^{-1}$ is a consistent estimator of inverse of covariance matrix of $S_T(\widehat{\theta}_{QMLE})/\sqrt{T}$. Closed-form expression for $\widehat{GS}_T$ is given in Appendix C.

Based on our testing procedure for two single data sequences above (steps $1 \sim 3$), we state a more general version for two sets of data sequences in Algorithm 1.

Algorithm 1 Non-parametric goodness-of-fit test for self-exciting point processes

Input: Two set of i.i.d. data sequences $D_1 = \{D_{1,1}, \ldots, D_{1,L}\}$ and $D_2 = \{D_{2,1}, \ldots, D_{2,L}\}$.
Initialization: n_0 bins on time horizon $[0, T_0]$; repeat times K; number of sequences N to calculate one GS statistic $\widehat{GS}_T$.
Output: K i.i.d. GS statistics .

Step I — Mix $D_{1,i}$ and $D_{2,i}$ to get the aggregated sequence D_i^{agg} $(i = 1, \ldots, L)$.

Step II — Apply Probability Weighted Histogram Estimation to learn QMLE.

Step III — Repeat the procedure for K times: randomly shuffle the order of sequences in the D_1 and repeat step I to get a different set of aggregated sequences, from which we randomly choose N sequences to calculate one $\widehat{GS}_T$.

The stationarity of a stochastic process means the unconditional probability distribution does not change when shifted in time. More specifically, for a stochastic process $N(t)$, for all $t \in \mathbb{R}$, $N(t, t+\delta]$ follows a same probability distribution as long as $\delta > 0$ is fixed. Thus, when $T \to \infty$, we will have

$$\frac{\mathbb{E}[\ell_1(\theta|\mathcal{H}_T)]}{T\ \mathbb{E}[\ell_1(\theta|\mathcal{H}_1)]} \to 1.$$

This shows that the estimand defined by maximum expected log-likelihood principle will not vary with different time horizon T (otherwise, θ_0 is not well-defined). Most importantly, this also shows that learning with L short sequences on time horizon $[0, T_0]$ is equivalent to learning with one long sequence on time horizon $[0, LT_0]$, which justifies our generalization to the testing on two sets of data sequences in Algorithm 1.

4 Theoretical Analysis

Here, we will prove the asymptotic performance of our GS statistics by establishing a novel connection with classic results in statistics for QMLE and the GS test based on it (White, 1982). We provide a generalization of the asymptotic properties of MLE for Hawkes process (Ogata, 1978) to model mismatch case, based on which we get the asymptotic behaviors of testing procedure such as score test and Wald test. The proofs and numerical illustration on why we choose score test over Wald test are deferred to Appendices D and E.

We use θ_0 to denote the projection of ground-truth and test $H_0 : \theta_0 \in \Theta_0$ against $H_1 : \theta_0 \notin \Theta_0$. Apparently, under different hypothesis, θ_0 cannot be the same. To avoid confusion, we say the projection is $\theta_0 = \theta_1 \in \Theta_0$ under H_0 and $\theta_0 = \theta_2 \notin \Theta_0$ under H_1.

Lemma 1 (Asymptotic properties of Quasi-MLE). *Let $\widehat{\theta}_{QMLE}$ and $\widetilde{\theta}_{QMLE}$ be QMLE under H_0 and H_1. For piecewise constant triggering function family (2), QMLE satisfies the following asymptotic properties:*

(i) Convergence to θ_0 almost surely. When $T \to \infty$,

$$\text{under } H_0\text{: } \widehat{\theta}_{QMLE} \overset{a.s.}{\to} \theta_1; \quad \text{under } H_1\text{: } \widetilde{\theta}_{QMLE} \overset{a.s.}{\to} \theta_2;$$

(ii) Asymptotic normality. Define $A(\theta) = \mathbb{E}\left[A_T(\theta)\right]/T$ and $B(\theta) = \mathbb{E}\left[B_T(\theta)\right]/T$, when $T \to \infty$, we will have:

$$\text{Under } H_0\text{: } \sqrt{T}(\widehat{\theta}_{QMLE} - \theta_1) \overset{d}{\to} N(0, \Sigma^{-1}(\theta_1));$$
$$\text{Under } H_1\text{: } \sqrt{T}(\widetilde{\theta}_{QMLE} - \theta_2) \overset{d}{\to} N(0, \Sigma^{-1}(\theta_2)),$$

where $\Sigma^{-1}(\theta) = B^{-1}(\theta)A(\theta)B^{-1}(\theta)$.

(iii) We also have asymptotically normality of the Quasi-score function, no matter under H_0 or H_1:

$$\frac{1}{\sqrt{T}} \frac{\partial \ell_1(\theta|\mathcal{H}_T)}{\partial \theta}\bigg|_{\theta=\theta_0} \overset{d}{\to} N(0, A(\theta_0)) \quad as \quad T \to \infty.$$

Remark. The score function should have the Fisher Information Matrix (FIM) $I(\theta^*)$ as its asymptotic covariance matrix when the model is correct. Using FIM will break the asymptotic χ^2 distribution in the model mismatch case. That's why we need to consider the model mismatch explicitly. Even though we cannot correctly specify the function family for unknown ground-truth, using $A(\theta_0)$ instead of FIM as the covariance matrix will still yield correct asymptotics for our proposed test. Moreover, by Theorem 1 in Ogata (1978), one can verify that Information Matrix Equivalence Theorem in White (1982) still holds for stationary point process, i.e. $\theta_0 = \theta^*$ and $A(\theta_0) = B(\theta_0) = I(\theta_0)$ hold if and only if the model is correctly specified. Thus, our results simplify to the form in Ogata (1978) in the absence of model mismatch. Though the asymptotic covariance matrix of QMLE is no longer inverse of the FIM $I^{-1}(\theta^*)$, we can still estimate it consistently.

Theorem 1 (Asymptotic null distribution of $\widehat{GS}_T$). *Under H_0, the Generalized Score (GS) test statistic has an asymptotic χ^2 distribution. More specifically,*

$$\widehat{GS}_T \overset{d}{\to} \chi^2_r \quad as \quad T \to \infty.$$

Note that here the degree of freedom is $r = n_0$, which is exactly the number of bins we discretize $[0, T_0]$ into.

Theorem 2 (Power function of GS test). *Under H_1, the GS statistic follows a asymptotic noncentral χ^2 distribution with degree of freedom r and noncentrality parameter $T\|\phi^{(1)} - \phi^{(2)}\|_2^2$. For any critical value $c > 0$, when $T \to \infty$, the test power is:*

$$\frac{\mathbb{P}_{H_1}(\widehat{GS}_T > c)}{Q_{r/2}(\sqrt{T}\|\phi^{(1)} - \phi^{(2)}\|_2, \sqrt{c})} \to 1,$$

where $\|\cdot\|_2$ is the vector ℓ_2 norm and $Q_M(a, b)$ is the Marcum-Q-function.

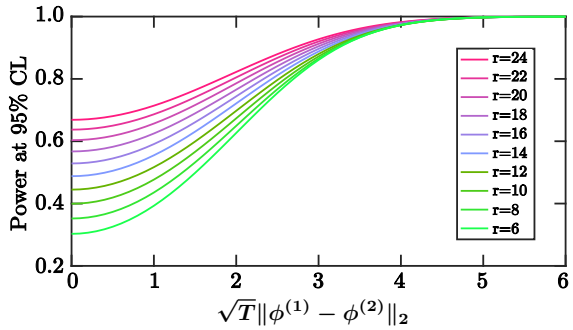


Figure 3: Illustration of asymptotic power of GS test.

The asymptotic power function with the critical value chosen to be the upper 95% quantile of the null distribution is shown in Figure 3. $Q_M(a, b) \to 1$ as $a \to \infty$, indicating our proposed test is consistent. See Appendix F for more on $Q_M(a, b)$.

5 Numerical experiments

In this section, we present numerical simulation to (1) validate the asymptotic property of our method by three simulation experiments; (2) demonstrate the GOF test for synthetic and real data.

5.1 Validation of asymptotic properties

To validate Theorems 1 and 2 presented in Section 4, we conduct three simulation experiments on a synthetic data set. We repeat our experiments on five sub-data sets generated from Hawkes process defined in (1) with 1,000 sequences, where $\mu = 20$ and an exponential triggering function $\phi(t - t_i) = \alpha e^{-10(t-t_i)}, t_i < t$ is adopted; α in each sub-data set is from $\{1.25, 1.5, 1.75, \ldots, 3.75\}$.

The Q-Q plot in Figure 4 (a) shows that the GS statistic follows the χ^2 distribution, which is consisent with Theorem 1; Figure 4 (b) visualizes the mean (red line) and the error bar (green bars) of each testing point for the GS statistics over different sample size N. Clearly, the GS statistics tend to be linear in sample size under

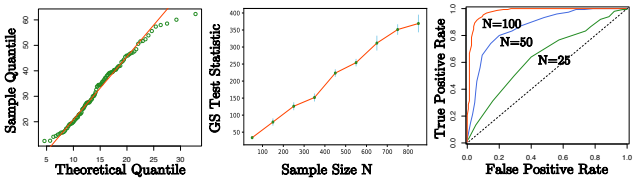


Figure 4: Simulation results: Left (a): quantiles of calculated GS statistics against theoretical quantiles of $\chi^2_{n_0}$ distribution under H_0; Middle (b): mean and variance of GS statistics with increasing N under H_1; Right (c): ROC curve for different N.

H_1, which matches the theoretical results shown in our power study in Theorem 2 and shows that our asymptotic distribution analysis is reasonably accurate. The ROC Curve in Figure 4 (c) shows that the GS statistics has good performance when $N = 100$ (AUC is approximately 1); We choose K to be $20, 5, 150$ for three experiments, respectively. Details on testing procedure can be found in Appendix E.

In short, we have confirmed (a) the χ^2 null distribution; (b) the score is linear in sample size under H_1; (c) the consistency of the proposed test. We also conduct similar experiments for power triggering functions to validate our method is model free. Results are deferred to Figure 8 in Appendix E due to space limitation.

5.2 Effects of number of Bins n_0

We use exponential synthetic data sequence D_1 and D_2 with $\mu_1 = \mu_2 = 20$, $\beta_1 = \beta_2 = 10$, $\alpha_1 = \alpha_2 = 1.5$ under H_0 and $\alpha_1 = 1.5, \alpha_2 = 5$ under H_1. The histogram estimate under H_0 is given in Figure 5. We perform GS test (confidence level 95%) under H_0 and H_1 100 times for each n_0 and report Type I & II errors in Table 1.

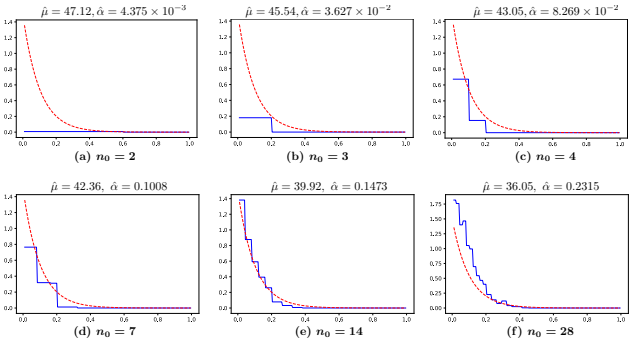


Figure 5: Histogram estimation of exponential kernel with $\mu = \mu_1 + \mu_2 = 40$ and $\alpha = 0.15$ with different n_0. The red dashed line is ground-truth $\alpha e^{-\beta t}$ and the blue solid line is the histogram estimate. The bottom middle panel ($n_0 = 14$) is the most accurate one.

From Figure 5, we can observe that with too many bins,

the histogram will overfit the data (panel (f)), whereas with fewer bins it underfits (panels (a)∼(d)). However, Table 1 shows that $n_0 = 3$ is most powerful in capturing the difference in triggering function. By comparing panel (a) and (b) in Figure 5, even though underfitting still exists, it captures the triggering function, which seems to be sufficient for our setting.

Table 1: Empirical Type I and Type II error (over 100 trials) for different number of Bins.

Number of Bins	2	3	4	7	14	28
Type I error	0.04	0.05	0.06	0.02	0.03	0.02
Type II error	0.59	**0.09**	0.19	0.34	0.79	0.83

5.3 Comparison with existing methods

The basic idea of existing GOF test due to Ogata (1988) is to (i) transform the original process to a residual process by keeping point t_i with probability $\widehat{\mu}/\widehat{\lambda}(t_i|\mathcal{H}_{t_i})$; (ii) test if the residual is a homogeneous Poisson process with rate $\widehat{\mu}$. Commonly used homogeneity test statistic is Ripley's K function (Ripley, 1976) and we use $\widehat{K}(t) = \sum_{i=1}^{N}\sum_{j\neq i}\mathbf{1}_{\{|t_j-t_i|\leq t\}}/\widehat{\mu}N$ as its estimate.

We apply both tests to exponential synthetic data with $\beta = 10$. We still use histogram estimation to estimate the conditional intensity. We calculate the GS statistics with $N = 50, K = 100$ and the average of $\widehat{K}(t)$ over $L = 100$ sequences for time span $t \in \{1, \ldots, 10\}$ but only report $t = 1, 10$ cases since the difference is not large when t doesn't change a lot. The rest is plotted in Figure 9 in Appendix E due to space limitation.

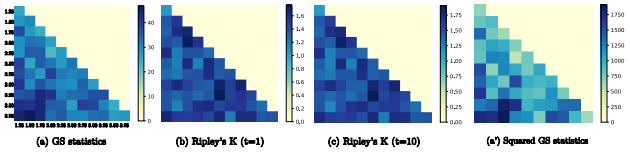


Figure 6: Heat map of (a) GS statistics, (b) $\widehat{K}(1)$, (c) $\widehat{K}(10)$ and (a') Squared GS statistics. For each pixel, the data sequence D_1 and D_2 are exponential synthetic data with α_1 and α_2 specified by the x-axis and the y-axis in (a). Squared GS statistics makes the gradual changing pattern more obvious.

Figure 6 visualizes the GS statistics and $\widehat{K}(t)$ when D_1 and D_2 are generated according to different α's, and show our method has more power in detecting the subtle difference in triggering part over existing methods. This is evident as in (a), the colors of the diagonal pixels are lighter whereas the colors of pixels on the bottom left are darker. This gradual changing pattern shows that GS statistic is larger when two generating distributions (i.e. α's) are further away whereas is smaller when those two distributions are closer, i.e. our proposed test can detect the subtle difference in triggering function accurately. However, in (b) and (c) we do not observe this gradual changing pattern, indicating Ripley's K function values are approximately the same when the true data generation mechanisms of two data sequences vary within a small set. This is because background intensity dominates the conditional intensity and most of the events comes from the background. Thus, testing of whole intensity will fail to detect the subtle triggering function difference.

5.4 Demonstration for model comparison

We perform our proposed test procedure on various synthetic and real data sets to compare four commonly used models. For synthetic experiments, we generate 5,000 sequences for each data sets, which come from the Hawkes process ($\mu = 10$) defined in (1) with different types of triggering functions: (a) exponential (`Exp`): $\phi(t - t_i) = e^{-3(t-t_i)}$; (b) Matern kernel (`Matern`): $\phi(t-t_i) = 0.2\times C_{0.2,2}(t-t_i)$, where $C_{\rho,\nu}(d) = \sigma^2(2^{1-\nu})/\Gamma(\nu)(\sqrt{2\nu}d/\rho)^\nu K_\nu(\sqrt{2\nu}d/\rho)$, where $\Gamma(\cdot)$ is the gamma function, $K_\nu(\cdot)$ is the modified Bessel function of the second kind. For real data experiments, we select a wide range of real data sets including: (c) MIMIC-III (Johnson et al., 2016) (`MIMIC`): 2,246 sequences with average sequence length 4.09; (d) MemeTracker (Leskovec et al., 2009) (`MEME`): randomly-picked 5,000 sequences with average sequence length 24.41. There are 2,500 sequences in (a), (b), (d), and 1,746 sequences in (c) are used for fitting the model and generating new sample sequences. The rest serves as testing data to calculate our GS statistics.

The models we are testing/comparing include (1) exponential triggering function fitted by gradient descent (`Exp GD`); (2) histogram estimation of triggering function fitted by EM algorithm (`Hist EM`) (Marsan and Lengline, 2008; Fox et al., 2016); (3) Long Short Term Memory (`LSTM`) (Hochreiter and Schmidhuber, 1997); (4) Neural Hawkes Process (`NHP`) (Mei and Eisner, 2017); (5) Homogeneous Poisson process with random average intensity (`Random`) as sanity check.

Table 2: GS statistic and Log-Likelihood; lower GS value is better, higher likelihood is better.

	GS statistic					Log-Likelihood		
Data	Exp GD	Hist EM	LSTM	NHP	Random	Exp GD	Hist EM	NHP
Exp	18.25	**11.63**	88.54	14.83	31.78	**21.27**	21.10	20.03
Matern	21.01	**18.40**	81.37	21.86	26.11	19.09	**19.49**	14.91
MIMIC	29.52	27.90	41.34	**25.24**	31.04	**10.46**	8.605	8.973
MEME	36.92	34.29	56.04	**29.98**	39.37	69.51	62.66	**73.15**

We follow the exact testing procedure in Algorithm 1 with $N = 200$, $K = 5$; we choose $n_0 = 15$ for `Exp` and `Matern` data and $n_0 = 13$ for `MIMIC` and `MEME` data. We

report the mean of scores and the likelihood of fitting the model in Table 2. We observe that our proposed GOF test can differentiate models under different settings. In particular, the GS statistics can be used as a ranking criterion. More specifically, the parametric models `Exp GD` and `Hist EM` achieve lower scores (better performance) on synthetic data sets comparing to `NHP` and `LSTM`, since the parametric assumptions of the parametric models (e.g., the additivity in triggering effects) are consistent with the Hawkes process used in generating synthetic data. In the contrast, `NHP` performs better on real data sets, including `MIMIC` and `MEME`, where dynamics between events are more complex and difficult to be captured using parametric models. We also present the corresponding likelihood in Table 2, which is commonly used to measure how well the data are fitted by the model (higher likelihood the better data is fitted). It shows that the likelihood result generally agrees with our GS statistics. Moreover, we also show that as a deterministic time series model, `LSTM` is difficult to compete with other baselines.

We should mention `Exp` data and `Exp GD` method case in particular, where the model is correctly specified. We use GD to maximize the likelihood to obtain MLE of the parameters. We observe that the estimates are further away from ground-truth while the likelihood keeps growing larger (see Figure 10 in Appendix E). This means overfitting occurs and therefore likelihood may be a questionable model comparison metric.

We next show that our proposed test can select the best model. We use the ground-truth to generate the "fitted" sequence, since it is hard to learn the parameters correctly (potentially due to the overly short sequences), and compare it with `Hist EM`. We adopt the same experimental setting with the first row in Table 2 (`Exp` data) and report the result in Table 3.

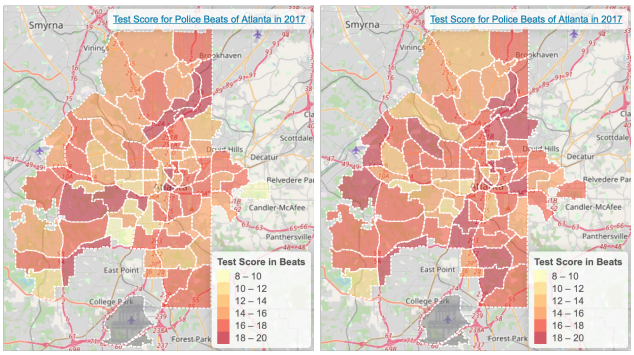
Table 3: Comparison of ground truth and `Hist EM` on `Exp` data. The GS statistic of `Hist EM` is different from that in Table 2 since we use different synthetic data.

Method	GS statistic	log-likelihood
Ground truth	**13.24**	21.64
`Hist EM`	17.38	**21.65**

From this table, we can see that log-likelihood cannot differentiate those two methods and is even misleading, whereas our proposed GS statistic suggests the ground truth is a lot better than the `Hist EM` method. Together with the numerical results in the past experiments, we demonstrate that our proposed GOF test can select the best model in the sense that how well the model captures the self-exciting part in the data.

Goodness-of-fit for 911 call data. To demonstrate the use of our test statistic as a diagnosis tool for the GOF of generative models, we test on 911 call data in 2017 provided by the Atlanta Police. The Atlanta Police Department divides its operation region into 78 beats, so we use this to partition the spatial region and consider a non-homogeneous point process generates sequences in each beat.

We first consider police events data in each beats in one day as a sequence, and for each beat fit generative model using `NHP` and `Exp GD`. Then we calculate the value of the test statistic for each beat. The experiment configurations are as follows: $N = 20$, $K = 1$, $n_0 = 12$. The results are presented in Figure 7.



(a) NHP (b) GD Exp

Figure 7: Goodness-of-fit test for Atlanta 911 call data: (a) for `NHP`; (b) for `Exp GD`. Each polygon in the map represents a police beat in Atlanta. The color depth represents the level of the test score. Lighter color: smaller discrepancy between the generated data and the real data. Overall speaking, we can see `NHP` has better GOF than `Exp GD`, especially in populated area.

Clearly, the generative model has different GOF in each beat. Also, the two generative models have different patterns in their GOF over space. Note that we do not know the ground-truth. This example demonstrates that our tools provide a convenient and flexible diagnosis tool for the GOF for generative models in practice.

6 Acknowledgement

The work is supported by the NSF CAREER Award CCF-1650913, and NSF CMMI-2015787, DMS-1938106, DMS-1830210. The authors would like to thank the Editor and the anonymous referees for the thoughtful comments and suggestions, which led to an improvement of the presentation.

References

Abdel-Aty, S. H. (1954). Approximate formulae for the percentage points and the probability integral of the non-central chi-squared distribution. *Biometrika*, 41(3/4):538–540.

Akaike, H. (1998). *Information Theory and an Extension of the Maximum Likelihood Principle*, pages 199–213. Springer New York, New York, NY.

Baringhaus, L. and Franz, C. (2004). On a new multivariate two-sample test. *Journal of multivariate analysis*, 88(1):190–206.

Bartle, R. G. (1976). *The elements of real analysis*. Wiley.

Boos, D. D. (1992). On generalized score tests. *The American Statistician*, 46(4):327–333.

Bounliphone, W., Belilovsky, E., Blaschko, M. B., Antonoglou, I., and Gretton, A. (2015). A test of relative similarity for model selection in generative models. *arXiv preprint arXiv:1511.04581*.

Bray, A. and Schoenberg, F. P. (2013). Assessment of point process models for earthquake forecasting. *Statistical science*, pages 510–520.

Chen, J., Hawkes, A., Scalas, E., and Trinh, M. (2018). Performance of information criteria for selection of hawkes process models of financial data. *Quantitative Finance*, 18(2):225–235.

Chwialkowski, K., Strathmann, H., and Gretton, A. (2016). A kernel test of goodness of fit. JMLR: Workshop and Conference Proceedings.

Engle, R. F. (1984). Wald, likelihood ratio, and lagrange multiplier tests in econometrics. *Handbook of econometrics*, 2:775–826.

Fox, E. W., Schoenberg, F. P., and Gordon, J. S. (2016). Spatially inhomogeneous background rate estimators and uncertainty quantification for nonparametric hawkes point process models of earthquake occurrences. *The Annals of Applied Statistics*, 10(3):1725–1756.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773.

Harchaoui, Z., Bach, F., Cappe, O., and Moulines, E. (2013). Kernel-based methods for hypothesis testing: A unified view. *IEEE Signal Processing Magazine*, 30(4):87–97.

Hawkes, A. G. (1971a). Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 33(3):438–443.

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

Hawkes, A. G. and Oakes, D. (1974). A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 11(3):493–503.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics*, 40(2):633–643.

Johnson, A. E., Pollard, T. J., Shen, L., Li-wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035.

Laub, P. J., Taimre, T., and Pollett, P. K. (2015). Hawkes processes. *arXiv preprint arXiv:1507.02822*.

Leskovec, J., Backstrom, L., and Kleinberg, J. (2009). Meme-tracking and the dynamics of the news cycle. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09*.

Lewis, E., Mohler, G., Brantingham, P. J., and Bertozzi, A. L. (2012). Self-exciting point process models of civilian deaths in iraq. *Security Journal*, 25(3):244–264.

Marsan, D. and Lengline, O. (2008). Extending earthquakes' reach through cascading. *Science*, 319(5866):1076–1079.

Mei, H. and Eisner, J. M. (2017). The neural hawkes process: A neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems*, pages 6754–6764.

Meyer, S. and Held, L. (2014). Power-law models for infectious disease spread. *The Annals of Applied Statistics*, 8(3):1612–1639.

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., and Tita, G. E. (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108.

Ogata, Y. (1978). The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Annals of the Institute of Statistical Mathematics*, 30(1):243–261.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical association*, 83(401):9–27.

Ogata, Y. (1999). Seismicity analysis through point-process modeling: A review. In *Seismicity patterns, their statistical significance and physical meaning*, pages 471–507. Springer.

Peng, R. D., Schoenberg, F. P., and Woods, J. A. (2005). A space–time conditional intensity model for evaluating a wildfire hazard index. *Journal of the American Statistical Association*, 100(469):26–35.

Porter, M. D. and White, G. (2012). Self-exciting hurdle models for terrorist activity. *The Annals of Applied Statistics*, 6(1):106–124.

Rao, C. R., Rao, C. R., Statistiker, M., Rao, C. R., and Rao, C. R. (1973). *Linear statistical inference and its applications*, volume 2. Wiley New York.

Reinhart, A. (2018). A review of self-exciting spatio-temporal point processes and their applications. *Statistical Science*, 33(3):299–318.

Ripley, B. D. (1976). The second-order analysis of stationary point processes. *Journal of Applied Probability*, 13(2):255–266.

Schoenberg, F. P. (2003). Multidimensional residual analysis of point process models for earthquake occurrences. *Journal of the American Statistical Association*, 98(464):789–795.

Schoenberg, F. P. (2013). Facilitated estimation of etas. *Bulletin of the Seismological Society of America*, 103(1):601–605.

Schoenberg, F. P., Hoffmann, M., and Harrigan, R. J. (2019). A recursive point process model for infectious diseases. *Annals of the Institute of Statistical Mathematics*, 71(5):1271–1287.

Schorlemmer, D., Gerstenberger, M., Wiemer, S., Jackson, D., and Rhoades, D. (2007). Earthquake likelihood model testing. *Seismological Research Letters*, 78(1):17–29.

Sun, Y., Baricz, Á., and Zhou, S. (2010). On the monotonicity, log-concavity, and tight bounds of the generalized marcum and nuttall q-functions. *IEEE Transactions on Information Theory*, 56(3):1166–1186.

Székely, G. J. and Rizzo, M. L. (2004). Testing for equal distributions in high dimension. *InterStat*, 5(16.10):1249–1272.

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.

White, H. (1980). Nonlinear regression on cross-section data. *Econometrica*, 48(3):721–746.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, pages 1–25.

Yang, J., Rao, V., and Neville, J. (2019). A Stein–Papangelou goodness-of-fit test for point processes. In *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 226–235. PMLR.

Zhuang, J. (2011). Next-day earthquake forecasts for the japan region generated by the etas model. *Earth, planets and space*, 63(3):207–216.

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