$Z-$Transforms and its Inference on Partially Observable Point Processes

Young Lee, Thanh Vinh Vo, Kar Wai Lim, Harold Soh
National University of Singapore

Abstract

This paper proposes an inference framework based on the $Z-$transform for a specific class of non-homogeneous point processes. This framework gives an alternative method to maximum likelihood estimation which is omnipresent in the field of point processes. The inference strategy is to *couple* or *match* the theoretical $Z-$transform with its empirical counterpart from the observed samples. This procedure fully characterizes the distribution of the point process since there exists a one-to-one mapping with the $Z-$transform. We illustrate how to use the methodology to estimate a point process whose intensity is driven by a general neural network.

1 Introduction

Background and issues. Point processes are statistical models for modelling random times at which events of interest take place. In a variety of applications in machine learning, maximum likelihood estimation can be difficult for point process fitting, depending on the specificities of the point process itself. As the likelihood may not be convex in the parameters, a maximization routine would likely converge to merely a local maximum as opposed to the intended global maximum.

One standard strategy used to identify the global maximum involves using several sets of different initial values for the maximization routine. We remark that this plan of action does not solve the problem in its entirety, and a local maximum may still be identified erroneously as the global maximum. This strategy can be taken further: more than one different maximization routines, with very distinct optimization techniques, may be used collectively with several sets of initial values to infer the parameters of interest. If these different maximization routines identify similar points as the potential global maximum, more confidence can be placed in judgements that the identified points are indeed the actual global maximum.

Another possible approach that is frequently used in the literature is the moment-based estimation wherein parameters are obtained by minimizing a measure of the discrepancy between empirical and theoretical second-moment properties. These *raw* and *uncentered* moments are usually easy to compute and they reveal important aspects of a distribution. For example, the first four moments tell us something about the information regarding the mean, variance and skewness as well as kurtosis, respectively. Using this information, one can immediately place constraints according to our theory on the so-called location, scale or shape and the tail behaviour of the distribution without specifying a full model.

The same line of thought applies to higher order moments. One might ask whether it is possible to match a larger set of moments, or the *entire spectrum* of moments (rather than truncating at the first four moments, say) in order to get an even closer representation of the model. If so, what would be a suitable *quantity* to match? This paper seeks to address these questions and shows that inference is possible through matching a *transform* of the raw and uncentered moments.

Purpose of the present study. We formulate an *alternative* estimation method using the $Z-$transforms for a class of non-homogeneous point processes whose intensity is driven by a general neural network. The $Z-$transform is an integral transform in which this transformation 'compresses' the point process into a single function of an argument, η, which we call the *dummy variable*. This dummy variable η serves as placeholders for the probabilities that determine the distribution of the point process.

$Z-$transforms are in general an alternative specification of stochastic processes that somehow encode the properties of the distributions into a form that is more convenient for certain kinds of probability calculation. We note that the $Z-$transform with dummy η is also known in combinatorics as the probability generating function.

This quantity *fully characterizes* the distribution of the point process in the sense that there exists a one-to-one correspondence between the point process structure and the $Z-$transform. Given the form of our neural network intensity driving the point process, we exploit the structure of the $Z-$transform to estimate the parameters of interest. The inference strategy is to *match* the theoretical $Z-$transform with its empirical counterpart from the observed samples.

In general, we may arrive at an *over-determined* situation: that is, one in which an abundance of $Z-$transform conditions are available. Similarly, there might be too few conditions than parameters to estimate wherein there is insufficient information and the model is *under-determined*. Through our formulation, we match the entire spectrum of the $Z-$transform by regulating η so that there is a surfeit of moment equations outweighing the number of parameters as our ideal situation.

Event censoring. These $Z-$transforms are brought only to illustrate one rather important application in point processes — *censoring*. Censoring is a phenomenon in which the value of an observation is only partially known. In the present context, the times at which events happen are *censored*. Some notable examples include medical trials: the subject may be observed only periodically, rather than assessed in a continuous fashion. The medical practitioners would report the *number* of heartburn per day rather than the *exact times* at which the heartburn occurs within that day for patients.

Contributions. Our noteworthy contributions are as follows:

1. We present a method of inference via the $Z-$transform for a specific class of non-homogeneous point processes whose intensity is driven by a general neural network. The idea is to *match* the theoretical $Z-$transform in its *entirety* with its empirical counterpart from the observed samples. This is done by regulating the number of dummy parameters η so that we have as many equations as there are moment conditions. This would ensure a single solution for the parameters.

2. We use this perspective to solve inference problems motivated by events that are censored. Formally, censored data arises when the event of interest cannot be directly observed but is only known to have occurred during an a period of time. In such cases, the only information we have is the number of events over a given period, but the times at which the event happens are not known. Through $Z-$transforms, we show how to carry out inference in such situations.

3. We explicate a classical connection where the $k-$th moment for our model exhibit a closed-form solution in terms of Stirling numbers and the Bell polynomial. The predictive formula $k = 1$ and expressions for higher moments to calculate variances, skewness and kurtosis are readily available.

4. We present experiments in a controlled setting and show that our inference algorithm are able to indeed recover the optimal parameters.

2 Notational Conventions and Prerequisites

Our framework requires some background on temporal point processes. Let us cover this topic briefly.

2.1 Temporal Point Processes

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

Fix some locally integrable $\lambda : \mathbb{R}_+ \to \mathbb{R}_+$. We call $(N(t))_{t\geq 0}$ an *non-homogeneous Poisson process with intensity* λ if

1. $N_0 = 0$ almost surely
2. for any $s < t$, $N(t) - N(s) \sim \text{Poisson}\left(\int_s^t \lambda(x)\, dx\right)$
3. for any $s < t \leq s' < t'$, $N(t) - N(s) \perp\!\!\!\perp N(t') - N(s')$.

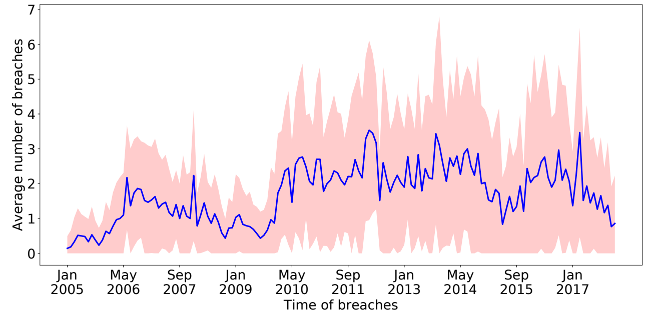


Figure 1: Average number of breaches with standard deviation

One may view non-homogeneous point processes as the following generative model for event times: given some terminal time T, the number of events N is drawn from a Poisson random variable with mean $\int_s^t \lambda(x)\, dx$, and the N event times are then drawn *i.i.d.* from a distribution with probability density function given by

$$\frac{\lambda(t)}{\int_s^t \lambda(x)\, dx}. \tag{1}$$

3 The Model – Neural Network Intensity

The nature of event counts in many applications is likely to be time-varying. A cursory visual inspection of cyber breaches [Privacy Rights Clearinghouse, 2018] over time (see Figure 1) reveals several periodicities. We remark that there is cyclical variability in the sense that there are some cyclical trends or seasonal trends (i.e., some repeated patterns of cyber breaches).

To this end, we formulate an intensity that is general enough to encapsulate these time varying characteristics, while presenting an inference mechanism via matching the $Z-$transform to learn the appropriate activation functions. We present a flexible family of a *neural network*

$$\lambda(t;\theta) = \tau + g\left(\sum_{r=1}^{R} b(t,\delta)\cdot a_r \cdot f(c_r\cdot t + d_r)\right) \tag{2}$$

where $\tau > 0$ being the background intensity driving the point process, $g(\cdot), f(\cdot)$ are non-negative activation functions, $b(t, \delta)$ is a decreasing function in t with δ as the parameter, and R is a fixed number which we call the *hidden units*. This model allows us to control the trend behavior in the intensity function, as can be seen in Figure 2. The parameters to estimate are $\theta = \{\tau, \delta, (a_r, c_r, d_r)_{r=1}^R\}$. Each term $f(c_r\cdot t+d_r)$ may be seen as a learned hidden representation of the input time. By choosing R to be sufficiently large, we can model complex non-linear hidden periodicities within the rate of arrivals governing the number of events. A standing assumption that we make is the follows:

$$\int_0^T \lambda(t)\, dt < \infty. \tag{3}$$

This intensity capture a wide range of other well known intensities. Some examples include

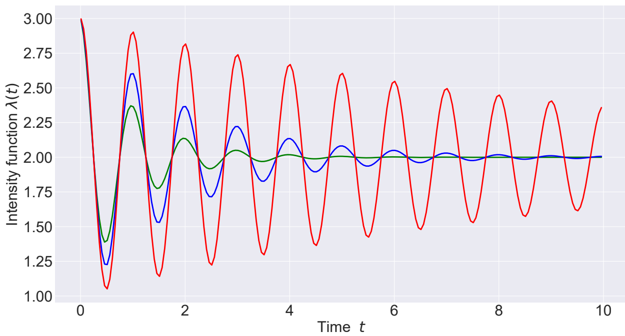


Figure 2: The diminishing effect of the intensity by controlling $b(t, \delta)$

1. **Polynomials.**
$$\lambda(t) = \tau + \sum_{r=0}^{R} a_r \cdot t^r \tag{4}$$
with the output activation function taking the identity function.
2. **The homogeneous Poisson process.**
$$\lambda(t) = \tau \in \mathbb{R}_{++} \tag{5}$$
with the output activation function g being the identity function.
3. **The p-th oscillation intensity.**
$$\lambda(t) = \tau + \sum_{r=1}^{R} e^{-\delta t} \cos^p(c_r \cdot t + d_r) \tag{6}$$
with the decaying function $a_r = 1$, $b(t, \delta) = e^{-\delta t}$. Figure (2) illustrates the dampening of this intensity function by controlling the parameter $\delta > 0$ and the p parameter.
4. **Sigmoid intensity.**
$$\lambda(t) = \tau + \sum_{r=1}^{R} a_r (1 + e^{-t})^{-1} \tag{7}$$
with the output activation function taking the identity function.

Existing work. The idea of using neural network intensities for general point processes is not new; several authors [Kim *et al.*, 2011; Du *et al.*, 2015; Choi *et al.*, 2016; Mei and Eisner, 2016; Du *et al.*, 2016; Xiao *et al.*, 2017; Wang *et al.*, 2017] have recently proposed the use of recurrent neural networks in conjunction with point processes with random intensities. These more complex models require Bayesian inference and likelihood methods, which prevents scalability. Our intensity function is inspired by, and a generalization of, the one proposed in [Menon and Lee, 2017].

3.1 The $Z-$Transform

Here we compute explicitly as possible the $Z-$transform for our model. The $Z-$transform is defined to be
$$E[\eta^{N(t)-N(s)}] \tag{8}$$
for $\eta \in (0, 1)$. As mentioned earlier, this is also known as the probability generating functional [Cox and Isham, 1980] or at times known as exponential tilting. We remark that this quantity uniquely determines the distribution of the entire point process [Kallenberg, 2006].

Proposition 1 *Let N be a point process and $\lambda(\cdot, \theta)$ be its associated intensity process. Let $\eta \in (0, 1)$. Then*
$$E\left[\eta^{N(t)-N(s)}\right] = \exp\left(\int_s^t \lambda(x, \theta)\, dx \cdot (\eta - 1)\right). \tag{9}$$
provided the assumption in equation (3) holds true.

Proof. The proof follows similar lines to that of Watanabe's Charaterization Theorem, confer [Brémaud, 1981]. For a fixed $T < \infty$, the integral in equation (1) is finite due to standing assumption as stated in equation (3). Note however that the original Watanabe's Characterization Theorem deals with expression of the form
$$E[\exp(\eta \cdot (N(t) - N(s)))] \tag{10}$$
but evaluating the quantity of interest in equation (8) requires one to only perform an exponential tilting tweak of a quantity, so the extra details of the proof is elided.

3.2 The First k Moments

The flexibility of theoretical $Z-$transform is borne out in evaluating the first k moments. As pointed out in [Pitman, 1997], one can deduce the following
$$\begin{aligned} E\left[(N(t) - N(s))^k\right] &= \sum_{z=1}^{k} \left\{ \begin{matrix} k \\ z \end{matrix} \right\} \left(\int_s^t \lambda(x, \theta)\, dx\right)^k \\ &= B_k \left(\int_s^t \lambda(x)\, dx\right) \end{aligned} \tag{11}$$
where the $\left\{ \begin{matrix} k \\ z \end{matrix} \right\}$ are known as the Stirling numbers of the second kind that counts the number of ways to partition a set of k elements into z sets and B is the Bell polynomial. This is an interesting connection because, one might be interested in finding the higher order moments of our model, after parameter estimation of $\hat{\theta}$ and this can be evaluated in Python by using the `mpmath.bell()` function.

The predictive formula. The predictive formula for a period of s', t' would take the following form:
$$\int_{s'}^{t'} \lambda(x, \hat{\theta})\, dx \tag{12}$$
This predictive formula gives the expected number of events following the learned parameters $\hat{\theta}$ using the procedures of inference deferred to Section 4.

4 Estimation Procedure

We spell out an optimization technique to learn the parameters θ of our point process driven by a general neural network intensity function. Recall that our intention is to match a *spectrum* of theoretical moments (not just the mean, variance, skewness or kurtosis) through the analytic $Z-$transform of a neural point process as computed in Proposition 1.

It is natural to assume that the following holds: if θ_0 is the true value of the parameter θ, then the theoretical $Z-$transform moments and the empirical moments must match.

Let $\Delta_j(t,s)$ be the observed number of events in sample j during the observation period $(s,t]$. Then the empirical $Z-$transform is given by $\phi_m(\eta,t,s)$ where

$$\phi_m(\eta,s,t) = \frac{1}{m}\sum_{j=1}^{m}\eta^{\Delta_j(s,t)}. \tag{13}$$

From Proposition 1, the theoretical $Z-$transform (repeated here for convenience) for our neural point process model is

$$\phi(\theta;s,t) = \exp\left(\int_s^t \lambda(x;\theta)\cdot(\eta-1)\,dx\right). \tag{14}$$

The parameters are obtained by minimizing the relative entropy between the $Z-$transform and its empirical counterpart, i.e.

$$D_{KL}(\boldsymbol{p}(\eta)\parallel\boldsymbol{q}(\eta)) = \sum_{\eta}\boldsymbol{p}(\eta)\log\frac{\boldsymbol{p}(\eta)}{\boldsymbol{q}(\eta)} \tag{15}$$

where $\boldsymbol{p}$ denotes the normalized empirical probability and $\boldsymbol{q}$ the normalized theoretical probability of the $Z-$transform. We need the normalization since we have to convert the values of $Z-$transforms and its empirical counterpart to probability distributions.

Sigmoid intensity $\lambda_\sigma(\cdot,\theta)$. We illustrate this with a concrete example. Suppose we fix our neural point process model as having a sigmoidal flavored intensity, i.e.

$$\lambda_\sigma(t,\{\tau,c,d\}) := \tau + \frac{1}{e^{(c\cdot t+d)}}. \tag{16}$$

The function has the property that the initial stage of intensity growth is approximately exponential; then, as saturation begins, the value slows, and at maturity, it terminates. Observed that there are three parameters to be estimated, namely $\{\tau,c,d\}$. The theoretical $Z-$transform associated with the sigmoid intensity $\lambda_\sigma(\cdot)$ is calculated to be

$$\phi_\sigma(\eta,s,t) = \exp\left(\tau(t-s) - \frac{1}{c\cdot e^{c(t-s)+d}}\cdot(\eta-1)\right). \tag{17}$$

Note in this particular example, there are three unknowns. The normalized distributions of $\boldsymbol{q}$ and $\boldsymbol{p}$ are calculated as follows:

$$\boldsymbol{q}: \quad \frac{\phi_\sigma(\eta_n,s,t)}{\sum_{n=1}^{\mathcal{N}}\phi_\sigma(\eta_n,s,t)},\quad n=1,2,3,\ldots,\mathcal{N}, \tag{18}$$

$$\boldsymbol{p}: \quad \frac{\phi_m(\eta_n,s,t)}{\sum_{n=1}^{\mathcal{N}}\phi_m(\eta_n,s,t)},\quad n=1,2,3,\ldots,\mathcal{N}. \tag{19}$$

Algorithm 1 Inference of parameters via $Z-$transforms

Input: Fix time points (s,t), m observed number of events for fixed times and $\Delta_j(s,t)$ for $j=1,2,\ldots m$. Define the intensity specification $\lambda(t)$.

1: Given the observed number of events from s to t, $\Delta_j(s,t)$, of m number of paths, compute the empirical moments:

$$\phi_m(\eta,s,t) = \frac{1}{m}\sum_{j=1}^{m}\eta^{\Delta_j(s,t)}$$

2: (a) For the intensity specification that leads to a closed form of $Z-$transforms, compute this quantity via:

$$\phi(\theta;s,t) = \exp\left(\int_s^t \lambda(x;\theta)\cdot(\eta-1)\,dx\right)$$

(b) For the intensity specification that does not inherit a closed form $Z-$transform, compute the theoretical moments approximately with an aid of a using symbolic programming and numerical integration.

3: Compute the normalized distributions $\boldsymbol{q}$ and $\boldsymbol{p}$ via expressions in equations (18) and (19) respectively.

4: Minimize the relative entropy or mean squared error between the theoretical and empirical $Z-$transforms as in equation (15) or equation (21) using standard optimization packages whilst enforcing strict constraints such that $\tau>0,\delta>0$ and $g(\cdot),f(\cdot)$ being non-negative activation functions.

Hence, the estimated parameters associated with λ_σ is then constructed by minimizing the relative entropy between the theorerical $Z-$moment and its empirical version calculated from the observed samples, i.e.

$$\operatorname*{argmin}_{\tau,c,d} D_{KL}(\boldsymbol{p}(\eta)\parallel\boldsymbol{q}(\eta)). \tag{20}$$

Another way to estimate λ_σ is to minimize the mean squared loss between the theoretical $Z-$transform and the empirical version as follows:

$$\operatorname*{argmin}_{\tau,c,d}\sum_{n=1}^{\mathcal{N}}\left(\phi_\sigma(\eta_n,s,t)-\phi_m(\eta_n,s,t)\right)^2. \tag{21}$$

Note that this is the standard loss used in the GMM framework [Hall, 2005].

Overdetermined and underdetermined systems. Let the dimension of θ be p and r be the number of conditions. In general there might be more moments conditions than parameters to estimate $r>p$, i.e., the model is overidentified. If $r<q$, then there is insufficient information and the model is underidentified.

The role of η in the $Z-$transform. Rather than matching the *first* k moments *viz.* mean, variance, skewness and kurtosis, say, we match the entire structure of the $Z-$transform by regulating the number of *different values* for η until there is a surfeit of moment conditions compared to number of unknowns.

	5,000 samples				10,000 samples				50,000 samples			
# of η	τ (KL)	a (KL)	τ (SE)	a (SE)	τ (KL)	a (KL)	τ (SE)	a (SE)	τ (KL)	a (KL)	τ (SE)	a (SE)
2	0.4586	0.8471	0.4647	0.8593	0.4253	0.8798	0.4268	0.8829	0.4510	0.8807	0.4499	0.8785
5	0.4341	0.8597	0.4401	0.8716	0.4558	0.8649	0.4572	0.8677	0.4582	0.8775	0.4569	0.8749
10	0.4400	0.8569	0.4459	0.8687	0.4354	0.8753	0.4367	0.8779	0.5092	0.8522	0.5078	0.8494
20	0.4694	0.8422	0.4753	0.8540	0.4865	0.8497	0.4878	0.8523	0.4460	0.8837	0.4446	0.8810
50	0.4431	0.8554	0.4490	0.8671	0.4669	0.8595	0.4682	0.8621	0.5005	0.8565	0.4991	0.8538
100	0.4325	0.8607	0.4384	0.8724	0.4370	0.8745	0.4382	0.8771	0.4581	0.8776	0.4568	0.8750
500	0.4395	0.8572	0.4454	0.8689	0.4899	0.8480	0.4912	0.8505	0.4977	0.8579	0.4964	0.8552
	τ (MLE): 0.4501 a (MLE): 0.8725				τ (MLE): 0.4612 a (MLE): 0.8736				τ (MLE): 0.4676 a (MLE): 0.8631			

Table 1: Calibrated parameters: linear intensity function

	5,000 samples				10,000 samples				50,000 samples			
# of η	τ (KL)	δ (KL)	τ (SE)	δ (SE)	τ (KL)	δ (KL)	τ (SE)	δ (SE)	τ (KL)	δ (KL)	τ (SE)	δ (SE)
2	2.7696	0.1262	2.8232	0.1249	2.9936	0.1357	2.9854	0.1361	3.0892	0.6859	3.0953	0.6857
5	2.7768	0.3949	2.8316	0.3938	3.0118	0.9035	3.0036	0.9036	3.0889	0.5631	3.0927	0.5632
10	2.7874	0.7920	2.8416	0.7905	3.0045	0.6271	2.9972	0.6274	3.0952	0.8812	3.1007	0.8811
20	2.6493	0.0294	2.8221	0.0332	2.8522	0.1856	2.9868	0.1831	2.9515	0.3221	3.0875	0.3184
50	2.4493	0.7687	2.8413	0.7595	2.6540	0.0037	2.9822	0.0045	2.7229	0.7858	3.0985	0.7765
100	2.4175	0.3755	2.8318	0.3651	2.6445	0.6529	2.9978	0.6442	2.6945	0.2189	3.0849	0.2094
500	2.4802	0.2867	2.8298	0.2783	2.6178	0.4932	2.9940	0.4841	2.6738	0.8038	3.0989	0.7939
	τ (MLE): 3.1160 δ (MLE): 0.6474				τ (MLE): 2.9115 δ (MLE): 0.4596				τ (MLE): 3.0779 δ (MLE): 0.4429			

Table 2: Calibrated parameters: oscillation intensity function

Connections to literature. In most of the point process literature, the moment matching is carried out by matching the first few moments as well as the tilted moments including the autocorrelation functions, [Zhang *et al.*, 2012; Yu, 2004; Gerhardt and Nelson, 2009; Heath *et al.*, 2013; Diggle *et al.*, 2013; Aït-Sahalia *et al.*, 2015; Da Fonseca and Zaatour, 2014]. Similar techniques were extended to inferring non-parametric forms of point processes that relies on second-order and third-order integrated cumulants [Achab *et al.*, 2017].

Depending on the form of the objective and the iterative specificities, these methods can be seen as falling into the framework of the generic method of Generalized Method of Moments for estimating parameters in probabilistic models [Hall, 2005] and are used in other applications of machine learning [Anandkumar *et al.*, 2012; Soufiani *et al.*, 2013].

5 Experiments

5.1 Parameter Recovery on Fictitious Data

Our inference algorithm is tested on synthetic data generated from our model. We validate our theoretical analyses by illustrating the feasibility of using the $Z-$transform matching technique for our neural point process model.

Preliminary setup. For fixed time points (s, t), we use the modified thinning method, confer [Ogata, 1981]; to generate m paths from a neural point process with known parameters θ^*. For each of this sample j, where $j = 1, 2, \ldots, m$, we record the total number of events and we compute the parameter estimate $\hat{\theta}$ via minimizing the procedures set forth in Section 4.

Intensity functions. We repeat this for 5,000 independent samples from the point process for two distinct intensity functions, i.e., **1.** *linear intensity* which takes the form

$$\lambda(t, \theta) = \tau + a \cdot t, \tag{22}$$

and **2.** the *oscillating intensity*, whose intensity diminishes as time goes by taking the following form:

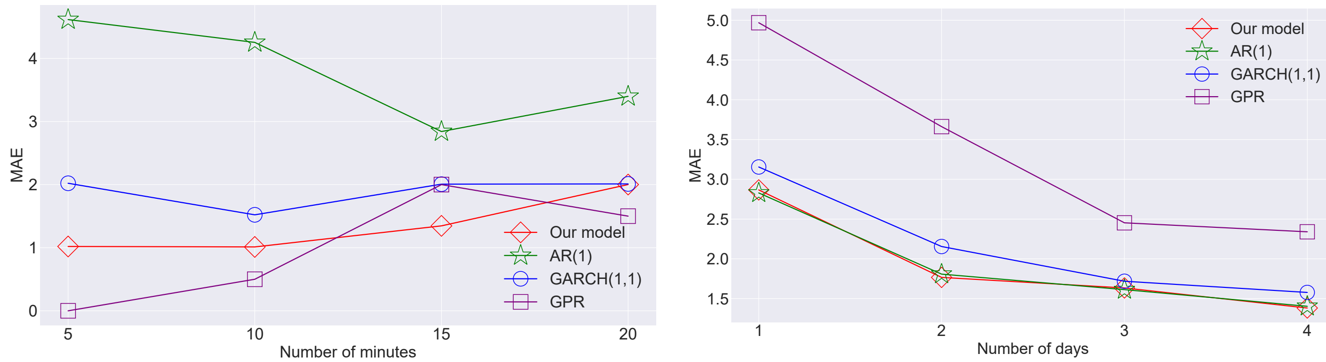
$$\lambda(t) = \tau + e^{-\delta t}(\cos(2\pi \cdot t)). \tag{23}$$

Note that this is a special case of that given in equation (6).

We set the parameters θ^* for the linear intensity as $\theta_{\mathbf{1}}^* = (\tau, a) = (0.5, 0.85)$ and the oscillating intensity as $\theta_{\mathbf{2}}^* = (\tau, \delta) = (3.0, 0.15)$.

Results. Tables 1 and 2 confirm that the $Z-$transform method have commensurate accuracy to the ground truth $\theta_{\mathbf{1}}^*$ and $\theta_{\mathbf{2}}^*$. This reassures that when the $Z-$transform method is used, these inference procedures operate sensibly, and indeed recover the optimal parameters asymptotically.

With regards to the linear intensity (see Table 1), this intensity may seem trivial but this is an example where *both* the MLE as well as the our $Z-$transform method have corresponding accuracy to the ground truth θ^*. However for the oscillation intensity (see Table 2), we see that the MLE method turns out to be problematic, possibly owing to the non-convexity and periodic nature of intensity function. We report that the calibrated parameters by matching the first two moments for linear and oscillation intensities are $(\tau, a) = (0.4000, 0.8997)$ and $(\tau, \delta) = (0.0458, 0.6996)$, respectively. Also in addition, note that the oscillation intensity function has the form constant $+ \exp(-\delta t) \cdot \cos(2\pi\, t)$ where the intensity decays off as time $t \nearrow \infty$. In this particular situation, the effect of the constant dominates the generation of the events. This could be the reason the process of recovering the ground truth for δ is a little challenging, as evident from Table 2.



(a) Calgary

(b) Breaches

Figure 3: Comparison of our model and some baseline models at varying time granularities

Method	Running time (s)
Our model	$\mathbf{0.025 \pm 0.002}$
AR(1)	0.342 ± 0.003
GARCH(1,1)	0.045 ± 0.004
GPR	3.258 ± 0.025

Table 3: Comparison of compute time (in seconds).

As a caveat on the applicability of the our method, we remark that the expression of exponential in equation (9) may at times cause numerical issues since for very large numbers of events, the resulting value for the transform is unrepresentably tiny.

5.2 Event Counts Predictions on Real Data

We compare the various methods on two real-world datasets: breaches data, henchforth known as *Breaches*, comprising times of security violation in a period of 2005 to 2018 [Privacy Rights Clearinghouse, 2018]; and *Calgary*, comprising of access log to the University of Calgary from October 1994 to October 1995 [The Internet Traffic Archive, 2018]. Previous studies have utilized standard time series models to fit and predict potential cyber attacks [Kumar *et al.*, 2006]. For each dataset, we aim to predict the number of events happening in some specified future time window.

Preliminary setup. For each of these datasets, we split the recorded number of events into a train and test set.We then carry out predictions of event counts over periods with different granularities. The precise split methodology varies for each dataset are as follows: For the Breaches dataset, we select the first 3,926 data points as training set to train the models and predict the next 1 to 4 data points (in days). For Calgary, the first 5,000 data points are selected as training set and we predict for the next 20 minutes. The prediction is done via these intervals using the Bell function formula explicated in Section 3.2.

Results. Figure 3 confirms that on real-world data, the performance of our model seemed competitive compared to standard baseline models: AR(1), GARCH$(1,1)$, as well as the Gaussian Process Regression (GPR). In Table 3, we show the training times of our method compared to AR(1), GARCH$(1,1)$, and GPR. The results seemed to indicate that there are benefits over existing models in performing short term predictions. The arguments as to why these might be the case can be found in the recent work of [Menon and Lee, 2017]. Our method is the fastest, while GPR is the slowest. All other methods including our method work well on big datasets; the GPR method seems unable to fit a training set with more than 6,000 points, as it faces a memory usage issue (using the scikit-learn implementation in a Core i7 machine with 8GB RAM); future work may compare against recent sparse Gaussian Process models or online approaches for time-series (e.g., [Soh and Demiris, 2015]).

6 Conclusion

We proposed a method of drawing inferences via the $Z-$transform for a point process whose arrival rate takes a neural network. The idea is to *match* the theoretical $Z-$transforms with its empirical counterpart of the observed samples. We further used this perspective to investigate the inference mechanism in situations wherein events are censored, i.e., the only information we have is the number of events over a given period, but the times at which the event happens are not known. In future work, exploring similar ideas for larger class of point processes whose governing intensity functions are in fact random (confer e.g. [Linderman and Adams, 2014; Flaxman *et al.*, 2017; Ding *et al.*, 2017]) would be of interest. Another area we expect to make progress in is the study that concerns properties of $Z-$transform estimators.

Acknowledgments

Young Lee wishes to thank Aditya K. Menon for his careful and critical reading of our manuscript.

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its National Cybersecurity R&D Programme (NRF2015NCR-NCR002-001).

References

[Achab *et al.*, 2017] Massil Achab, Emmanuel Bacry, Stéphane Gaïffas, Iacopo Mastromatteo, and Jean-François Muzy. Uncovering causality from multivariate Hawkes integrated cumulants. In *ICML*, pages 1–10, 2017.

[Aït-Sahalia *et al.*, 2015] Yacine Aït-Sahalia, Julio Cacho-Diaz, and Roger J. A. Laeven. Modeling financial contagion using mutually exciting jump processes. *Journal of Financial Economics*, 17(3):585–606, 2015.

[Anandkumar *et al.*, 2012] Animashree Anandkumar, Daniel Hsu, and Sham M Kakade. A method of moments for mixture models and hidden Markov models. In *Conference on Learning Theory*, pages 33–1, 2012.

[Brémaud, 1981] Pierre Brémaud. *Point Processes and Queues: Martingale Dynamics*. Springer-Verlag, 1981.

[Choi *et al.*, 2016] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor AI: Predicting clinical events via recurrent neural networks. In *ML for Healthcare Conf.*, pages 301–318, 2016.

[Cox and Isham, 1980] David Roxbee Cox and Valerie Isham. *Point Processes*. Chapman & Hall, 1980.

[Da Fonseca and Zaatour, 2014] José Da Fonseca and Riadh Zaatour. Hawkes process: Fast calibration, application to trade clustering, and diffusive limit. *Journal of Futures Markets*, 34(6):548–579, 2014.

[Diggle *et al.*, 2013] Peter J Diggle, Paula Moraga, Barry Rowlingson, and Benjamin M Taylor. Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm. *Stat. Sci.*, pages 542–563, 2013.

[Ding *et al.*, 2017] Hongyi Ding, Issei Sato, and Masashi Sugiyama. Bayesian nonparametric Poisson process allocation. *arXiv preprint arXiv:1705.07006*, 2017.

[Du *et al.*, 2015] Nan Du, Yichen Wang, Niao He, and Le Song. Time-sensitive recommendation from recurrent user activities. In *NIPS*, pages 3474–3482. 2015.

[Du *et al.*, 2016] Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. Recurrent marked temporal point processes: Embedding event history to vector. In *Proceedings of the 22nd ACM SIGKDD*, KDD '16, pages 1555–1564. ACM, 2016.

[Flaxman *et al.*, 2017] Seth Flaxman, Yee Whye Teh, and Dino Sejdinovic. Poisson intensity estimation with reproducing kernels. In *Proceedings of the 20th Int. Conf. on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 270–279, 2017.

[Gerhardt and Nelson, 2009] Ira Gerhardt and Barry L Nelson. On capturing dependence in point processes: Matching moments and other techniques. Technical report, Northwestern university, 2009.

[Hall, 2005] Alastair R Hall. *Generalized method of moments*. Oxford University Press, 2005.

[Heath *et al.*, 2013] Robert W Heath, Marios Kountouris, and Tianyang Bai. Modeling heterogeneous network interference using Poisson point processes. *IEEE Transactions on Signal Processing*, 61(16):4114–4126, 2013.

[Kallenberg, 2006] Olav Kallenberg. *Foundations of modern probability*. Springer Science & Business Media, 2006.

[Kim *et al.*, 2011] Sanggyun Kim, David Putrino, Soumya Ghosh, and Emery N Brown. A granger causality measure for point process models of ensemble neural spiking activity. *PLoS computational biology*, 7(3):e1001110, 2011.

[Kumar *et al.*, 2006] Vipin Kumar, Jaideep Srivastava, and Aleksandar Lazarevic. *Managing Cyber Threats: Issues, Approaches, and Challenges*, volume 5. Springer, 2006.

[Linderman and Adams, 2014] Scott W. Linderman and Ryan P. Adams. Discovering latent network structure in point process data. In *ICML*, pages 1413–1421, 2014.

[Mei and Eisner, 2016] Hongyuan Mei and Jason Eisner. The neural Hawkes process: A neurally self-modulating multivariate point process. *CoRR*, abs/1612.09328, 2016.

[Menon and Lee, 2017] Aditya Krishna Menon and Young Lee. Predicting short-term public transport demand via inhomogeneous Poisson processes. In *Proceedings of the 2017 ACM on CIKM*, pages 2207–2210. ACM, 2017.

[Ogata, 1981] Yosihiko Ogata. On Lewis' simulation method for point processes. *IEEE TIT*, 27(1):23–31, 1981.

[Pitman, 1997] Jim Pitman. Some probabilistic aspects of set partitions. *The American mathematical monthly*, 104(3):201–209, 1997.

[Privacy Rights Clearinghouse, 2018] Privacy Rights Clearinghouse. Data breaches, 2018.

[Soh and Demiris, 2015] Harold Soh and Yiannis Demiris. Spatio-temporal learning with the online finite and infinite echo-state Gaussian processes. *IEEE Trans. on Neural Networks and Learning Systems*, 26(3):522–536, 2015.

[Soufiani *et al.*, 2013] Hossein Azari Soufiani, William Chen, David C Parkes, and Lirong Xia. Generalized method-of-moments for rank aggregation. In *Advances in Neural Information Processing Systems*, pages 2706–2714, 2013.

[The Internet Traffic Archive, 2018] The Internet Traffic Archive, 2018. http://ita.ee.lbl.gov/html/traces.html.

[Wang *et al.*, 2017] Yongqing Wang, Shenghua Liu, Huawei Shen, Jinhua Gao, and Xueqi Cheng. Marked temporal dynamics modeling based on recurrent neural network. In *Advances in Knowledge Discovery and Data Mining, PAKDD 2017*, pages 786–798. Springer, 2017.

[Xiao *et al.*, 2017] Shuai Xiao, Junchi Yan, Xiaokang Yang, Hongyuan Zha, and Stephen M. Chu. Modeling the intensity function of point process via recurrent neural networks. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 1597–1603, 2017.

[Yu, 2004] Jun Yu. Empirical characteristic function estimation and its applications. *Econometric Reviews*, 23(2):93–123, 2004.

[Zhang *et al.*, 2012] Tonglin Zhang, Zuoyi Zhang, and Ge Lin. Spatial scan statistics with overdispersion. *Statistics in medicine*, 31(8):762–774, 2012.