University of California

Los Angeles

Non Parametric Estimation of Inhibition for Point Process Data

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Statistics

by

Alexa Beyor

2015

Abstract of the Thesis

Non Parametric Estimation of Inhibition for Point Process Data

by

Alexa Beyor

Master of Science in Statistics

University of California, Los Angeles, 2015

Professor Frederic R. Paik Schoenberg, Chair

For a single geyser one eruption may inhibit another eruption. The objective is to estimate the inhibition function of geyser eruptions using a non parametric algorithm by extending the non parametric estimation method of Marsan and Lengliné(2008) for clustered Hawkes processes to the case where there may be inhibition. The proposed method is tested using simulated geyser eruption data from known densities: Exponential, Pareto, Normal, and Uniform. The method is then applied ot the actual data from the Lone Pine Geyser in Yellowstone National Park. The data consists of 163 eruptions from 2011. The results indicate that geyser eruptions do inhibit other eruptions to some degree.

The thesis of Alexa Beyor is approved.

Nicolas Christou

Ying Nian Wu

Frederic R. Paik Schoenberg, Committee Chair

University of California, Los Angeles

2015

To my mother and father

Table of Contents

List of Figures

CHAPTER 1

Introduction

The goal is to estimate the inhibition function of geyser eruptions using a non parametric algorithm. The idea for estimating inhibition stems from the Marsan and Lengliné(2008) algorithm for estimating the triggering cascade of earthquakes. The Marsan and Lengliné(2008) algorithm estimates the triggering cascade probabilistically for a purely clustered Hawkes point process (Hawkes 1971), whereas here we allow the process to have both clustering and inhibition at different scales. A triggering cascade refers to one earthquake triggering another earthquake which in turn can trigger subsequent earthquakes.

Earthquakes and geyser eruptions are a naturally occurring point processes. A point process is a collection of random points in space such as location or time. Both earthquakes and geyser eruptions alter their respective systems when they occur. Tectonic plates shift during an earthquake, and water is ejected from a chamber which empties and cools during an eruption. Geysers are a fairly rare phenomena that only exist under a specific set of environmental conditions. There must be a source of water, a chamber to hold the water, intense heat, and pressure. According to Rinehard(1980), a geyser occurs when the surface of a hot spring is constricted preventing the circulation of water and heat loss. Pressure in the hot spring increases with depth which can prevent the deepest water from boiling even when the temperature exceeds the surface boiling point. Steam forms and bubbles upwards as the water in the chamber rises. As the bubbles rise they expand. At a critical point, the bubbles will lift the water causing overflow or splashes. The overflow and splashing releases the pressure within the system resulting in intense and violent boiling forcing out the water in the chamber. This eruption of water is a geyser. The eruption ceases when the system cools or the water is depleted.

The process of geyser eruptions appears to imply one eruption would inhibit subsequent eruptions of the same geyser. A period of time must pass to allow for the water to refill the chamber, for pressure to build, and steam to form.

The algorithm proposed in this paper is an extension of the idea from Marsan and Lengliné(2008) to estimate the triggering function of a Hawkes process non parametrically, only here the triggering function is allowed to be negative and thus potentially finding inhibition. Hawkes models are used in the study of earthquakes (Ogata 1988, Ogata 1998), crime (Mohler et al. 2011, Zipkin et al. 2016), and invasivie species (Balderama et al. 2012), among other applications. Historically processes like geysers and other point processes with inhibition have typically been modeled as a renewal process, Daley and Vere-Jones(2003). The difference here is that, with a renewal process, the time until the next eruption only depends on the single previous eruption time, whereas in this paper this assumption is relaxed. For example, suppose a geyser typically erupts every ten days. Also suppose the geyser has recently erupted at intervals of two days, three days, and then ten days. According to the renewal model, it is no more likely than average to take a long time now, and the two and three day intervals are irrelevant. However, by common sense one might expect a longer interval. The method proposed in this paper allows for two and three day intervals to have a lasting effect.

The data set used in this paper is from the Lone Pine Geyser in Yellowstone National Park. In addition to estimating the inhibition function for this particular geyser, the non parametric algorithm with also be evaluated. To assess the performance several simulated data set were created with known densities.

CHAPTER 2

Procedures & Methods

Estimating the inhibition function of a point process data set requires estimating the rate, the number expected due to background, the number observed within a bin, and choosing the interval length, and bin size for the chosen interval. Suppose $g_{u_i u_j}$ is the inhibition function for the interval u between u_i and u_j, then $g(u) = g_{u_i u_j}$ for all u in $[u_i, u_j]$. The non parametric equation to estimate inhibition is:

$$\hat{g}_{u_i u_j} = \frac{\sum_{k=1}^{n} \sum_{l=1}^{n} 1_{t_l - t_k \in [u_i, u_j]}}{n(u_j - u_i)} - \frac{n}{T} \tag{2.1}$$

and the standard error equation is:

$$SE_{g_{u_i u_j}} = \frac{\sqrt{\sum_{k=1}^{n} \sum_{l=1}^{n} 1_{t_l - t_k \in [u_i, u_j]}}}{T(u_j - u_i)} \tag{2.2}$$

Where n is the total number of observed points, T is the duration of the time series, $\frac{n}{T}$ is the rate, t is an event, u is the interval, and u_i is the start of the bin and u_j is the end of the bin within the interval. The number expected due to background is:

$$k = \frac{n(u_j - u_i)}{T} \tag{2.3}$$

The number observed within a bin is:

$$\sum_{k=1}^{n}\sum_{l=1}^{n} 1_{t_l - t_k \in [u_i, u_j]} \tag{2.4}$$

To computer the inhibition estimates and standard errors, iterate through each bin until the end of the chosen interval is reached. Once all the estimates for each bin are computed, they are transformed into density estimates using the following equation:

$$f(u) = \frac{\hat{g}_{u_i u_j}}{(u)(m)\sum \hat{g}_{u_i u_j}} \tag{2.5}$$

Where m is the number of bins. The density estimates of the inhibition and 95% confidence bounds are then plotted for each bin in the interval. The resulting plot is then analyzed.

It should be noted that the choice of interval length and bin size play an intimate role in how well the inhibition estimates and standard errors will recovery the true (underlying) density. If a bin is too small, there may only be a few observed points, if any. Not having enough events in a bin results in poor estimates and poor standard errors. Conversely, if a bin is too large, there may be too many event is a given bin which may result in missing potential inhibition between events. This may happen if the two events are captured within the same bin. The length of the interval is also important as there may be boundary effects if the interval is excessively large.

Choosing the most appropriate interval length and bin size begins with an educated guess. With the initial choice, compute the estimates and standard errors then plot them. Viewing the plotted estimates and standard errors will reveal if a bin is the incorrect size and if the interval is too large. If the confidence bounds become larger with each bin, the bin size is typically too small and should be increased. The bin size should also be increased if the confidence bound of an estimate is non-existent as this implies there were no observed points in the bin. The bin size should be decreased when the plotted estimates do not appear to converge near zero. It is also important to look at the number of observed values within a bin. If a bin only has one or two observed events, the bin is too small. Conversely, if the number of observed events in a bin is significantly larger the number expected due to background, the bin size is too large. To find the optimal interval length and bin size, we adjusted the interval length and bin size and plotted the recomputed estimates and standard errors. Depending on the resulting plots further adjustments may have been made to interval length or bin size.

CHAPTER 3

Simulated Data

In order to assess the performance of the inhibition estimates several sets of simulated data with known densities were created. The simulations are designed to be similar to a single geyser with multiple eruptions. To do this start with a set of random deviates from a Poisson process and arrange them in ascending order. The Poisson process is ideal for creating a point process set of times since it is a random process that will count the number of events in a given interval. Points are removed from the initial set of random deviates using a probabilistic method with a specific probability density function. Models of the form 3.1 are considered, where h is non negative and thus there is inhibition, but not so much inhibition that the conditional intensity dips below zero, which would violate the definition of a point process intensity, Daley and Vere-Jones(2003). This model is simulated using a Poisson process and then thinning it, according to the routine of Lewis and Schedler(1978).

$$\lambda(t) = [10 - \sum_{i=1} h(t - t_i)]^+ \tag{3.1}$$

Where $t_i < t$, t is a point in the sorted set of initial times, $h(t)$ is the probability density function, $\lambda(t)$ is the estimated rate of events, and $\lambda = 10$ is the expected rate of events for all created simulations. Using $\lambda(t)$ and λ a probability, p, is computed using the following equation.

$$p = \frac{\lambda(t)}{\lambda} \tag{3.2}$$

The probability is compared to a randomly generated uniform value between 0 and 1. The point t will be removed if p greater than or equal to the generated random uniform. Repeat this process for every point in the sorted set of initial times. This process will remove approximately half of the points from the initial set. The thinned data will be referred to as a "simulation" from here on out.

Twelve simulations were created using the probability density functions of the Exponential, Pareto, Normal, and Uniform Distributions. The probability density function, pdf, of each distribution is used for $h(t)$. There are three simulations for each distribution with approximately 500, 5,000, and 50,000 events. Each distribution was chosen for a particular reason that will be discussed in further detail in the respective subsections. The various simulations sizes and thinning probability density functions allow for a better evaluation of the inhibition estimation algorithm. The aim is to estimate the known density for each simulation. The main way this will be done is by plotting the known density with the estimates for each simulation.

3.1 Exponential

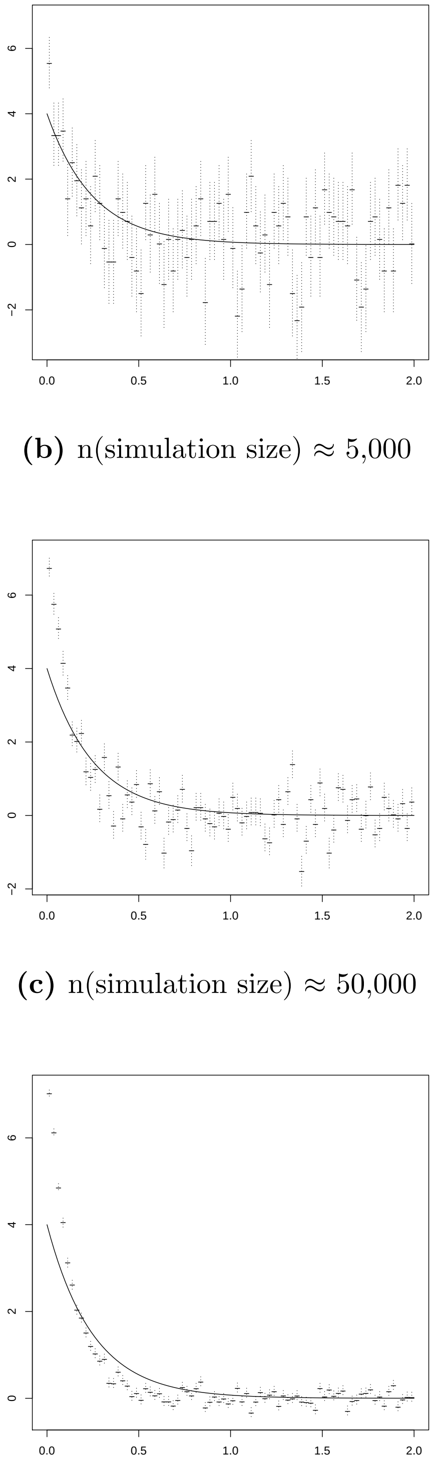
The Exponential Distribution was chosen for its monotonically decreasing probability density function. It also describes the time intervals between events in a Poisson process. This is useful since a decreasing density is expected if there is inhibition, and the initial set of random deviates, times, is created with a Poisson process.

A random rate of $\lambda = 4$ was chosen for the pdf. The following is the pdf of the Exponential Distribution.

$$h_{exponential}(t) = \lambda e^{-\lambda t} \tag{3.3}$$

Figure 3.1 are the plotted estimates and 95% confidence bounds with the exponential pdf. The interval is from 0-2 with 80 bins. It is evident from looking at Figure 1 that as the size of the simulation increases the estimates become less sporadic and the 95% confidence bounds become smaller. Most importantly as the simulation size increases the estimates follow the pdf better. Even though the smallest simulation, Figure 3.1a, has the least accurate estimates it still has the general shape of the exponential pdf. The inhibition estimation algorithm is preforming alright.

Figure 3.1: The Exponential pdf with $\lambda = 4$, estimates, and 95% confidence bounds for the simulations created using the Exponential pdf are shown in the plots The interval is from 0 to 2 with a bin size of .025.



(a) n(simulation size) $\approx$ 500

(b) n(simulation size) $\approx$ 5,000

(c) n(simulation size) $\approx$ 50,000

3.2 Pareto

The Pareto Distribution, similarly to the Exponential Distribution, has a decreasing probability density function. The main difference between the two distributions is the Pareto Distribution has a heavier tail. Using a slightly different pdf allows for a more thorough assesment of the inhibition estimation algorithm.

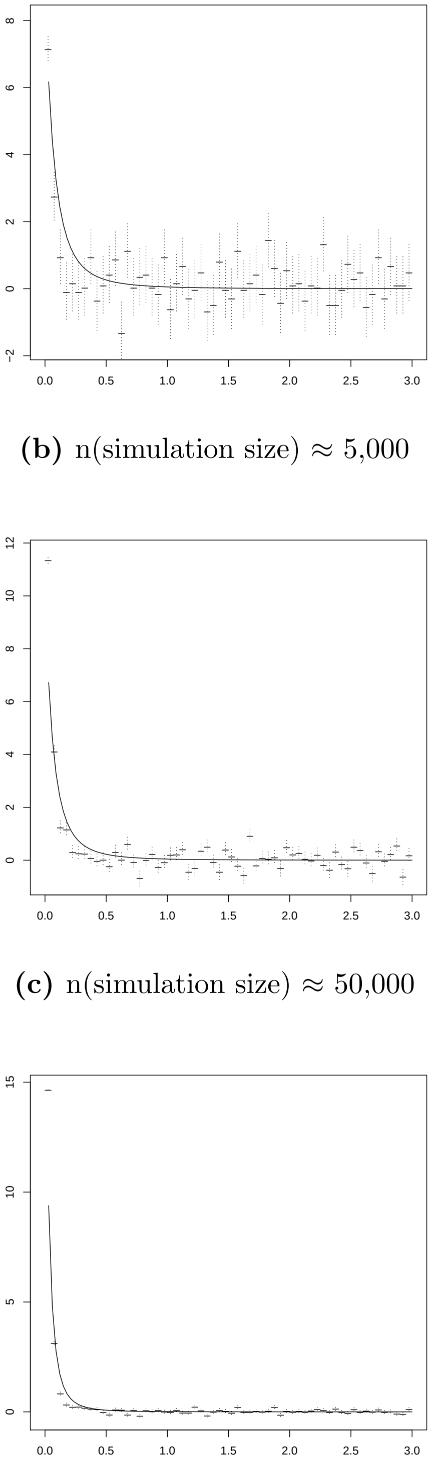
The shape, $\alpha = 2$, for the pdf was chosen at random, and the scale, t_m, is the minimum value of the simulation. The following is the pdf of the Pareto Distribution for all $t > t_m$.

$$h_{pareto}(t) = \frac{\alpha t_m^{\alpha}}{t^{\alpha+1}} \tag{3.4}$$

Figure 3.2 are the plotted estimates and 95% confidence bounds with the Pareto pdf. The interval is from 0 to 3 with 60 bins. Once again as the simulations size increases the estimates become less irregular and the 95% confidence bounds become smaller. The estimates for all three simulations follow the Pareto pdf fairly well. It is clear from looking at Figure 3.2a, 3.2b, or 3.2c that the underlying density is a decreasing one that is similar to a Pareto or exponential. The inhibition estimation algorithm preforms exceptionally well.

Figure 3.2: The Pareto pdf with $\alpha = 2$, estimates, and 95% confidence bounds for the simulations created using the Pareto pdf are shown in the plots below. The interval is from 0 to 3 with a bin size of .05.

(a) n(simulation size) ≈ 500



(b) n(simulation size) ≈ 5,000

(c) n(simulation size) ≈ 50,000

3.3 Normal

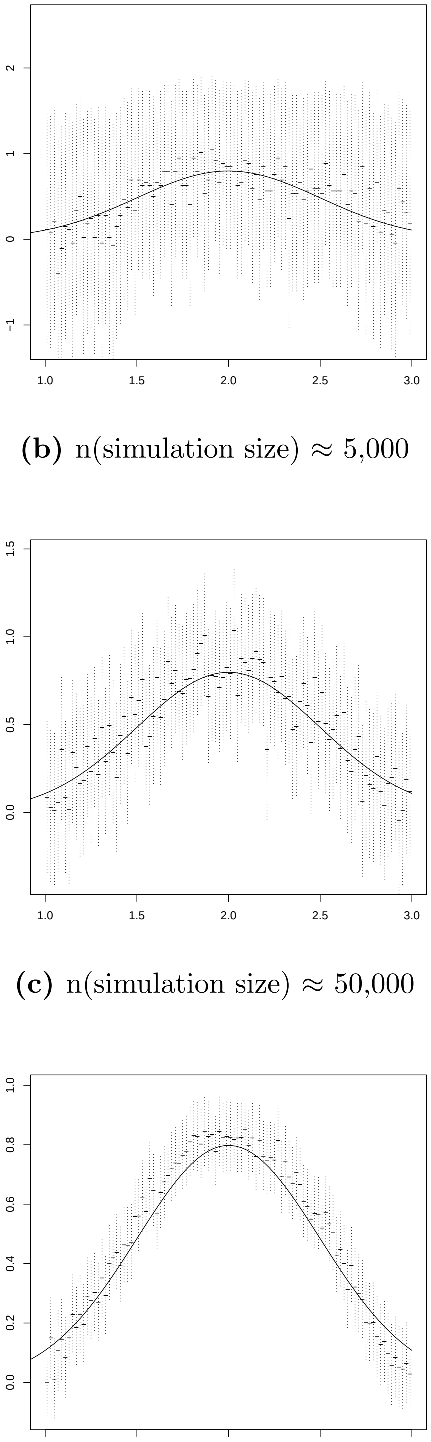
The Normal Distribution was chosen to help gauge how well the inhibition estimation algorithm would preform with a nonzero peak. An nonzero peak implies that part of the density will be increasing and the other part will be decreasing.

For the pdf of the Normal Distribution, a nonstandard mean and standard deviation are used, $\mu = 2$ and $\sigma = .5$. The following is the pdf of the Normal Distribution.

$$h_{normal}(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(t-\mu)^2}{2\sigma^2}} \tag{3.5}$$

Figure 3.3 are the plotted estimates and 95% confidence bounds with the Normal pdf. The interval is from 1 to 3 with 90 bins. Similar to the other simulations, as the simulations size increases the estimates and 95% confidence bounds become less varied and smaller. In Figure 3.3a, there is a slight peak of the estimates and confidence bounds in the middle. This slight hump makes it easier to detect the underlying normal pdf, but it would be difficult to see without the plotted normal pdf. On the other hand, Figure 3.3c, the estimates for the largest simulation follow the normal pdf extremely well. It appears as though even when the peak of the density is not at the beginning or at zero the inhibition estimate algorithm preforms relatively well.

Figure 3.3: The Normal pdf with $\mu = 2$ and $\sigma = .5$, estimates, and 95% confidence bounds for the simulations created using the Normal pdf are shown in the plots below. The interval is from 1 to 3 with a bin size of .02.



(a) n(simulation size) ≈ 500

(b) n(simulation size) ≈ 5,000

(c) n(simulation size) ≈ 50,000

3.4 Uniform

Similarly to the Normal Distribution, the Uniform Distribution was chosen due to the immediate drop from the maximum to minimum value. This will help with determining how well the inhibition estimation algorithm preforms when the difference between the maximum (peak) and minimum is not gradual.

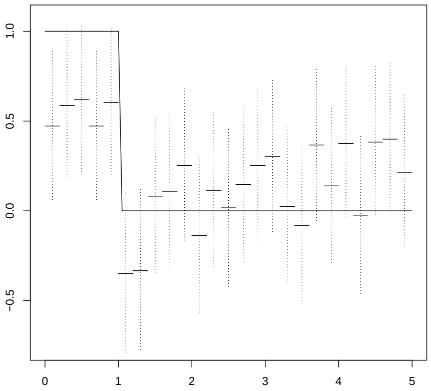
The standard values of the minimum and maximum, $a = 0$ and $b = 1$, were used for the Uniform Distribution. The following is the pdf of the Uniform Distribution for all $t \in [a, b]$.

$$h_{uniform}(t) = \frac{1}{b - a} \tag{3.6}$$

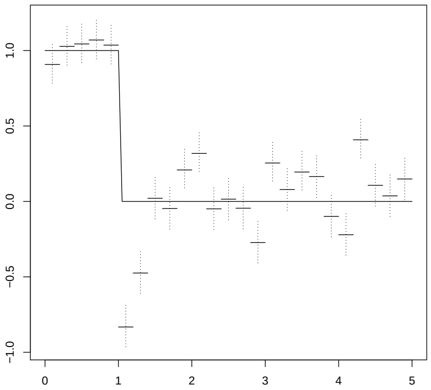
Figure 3.4 are the plotted estimates and 95% confidence bounds with the Uniform pdf. The interval is from 0 to 5 with 25 bins. In figure 3.4a, the estimates are under and then over estimated. In Figure 3.4b, the estimates of the maximum of the Uniform pdf are great, but the estimates for the minimum value are more sporadic and the confidence bounds do not fully capture the plotted Uniform pdf. In figure 3.4c, the maximum value of uniform pdf is overestimated a bit,and the estimates for the minimum value are under estimated but converge fairly quickly to the zero, the minimum value of the pdf. It is interesting to see how the amount of information in each bin affects how well the estimate preforms.

Figure 3.4: The Uniform pdf with $a = 0$ and $b = 1$, estimates, and 95% confidence bounds for the simulations created using the Uniform pdf are shown in the plots below. The interval is from 0 to 5 with a bin size of .2.

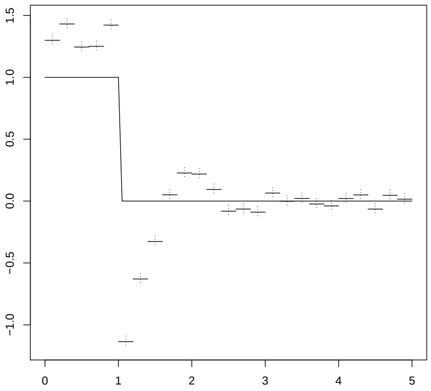
(a) n(simulation size) $\approx$ 500



(b) n(simulation size) $\approx$ 5,000



(c) n(simulation size) $\approx$ 50,000



CHAPTER 4

Lone Pine Geyser

The data set used is from the Lone Pine Geyser in Yellowstone National Park. The most recently recorded information about eruptions of this geyser are from 2011. This is the eruption information used in this paper.

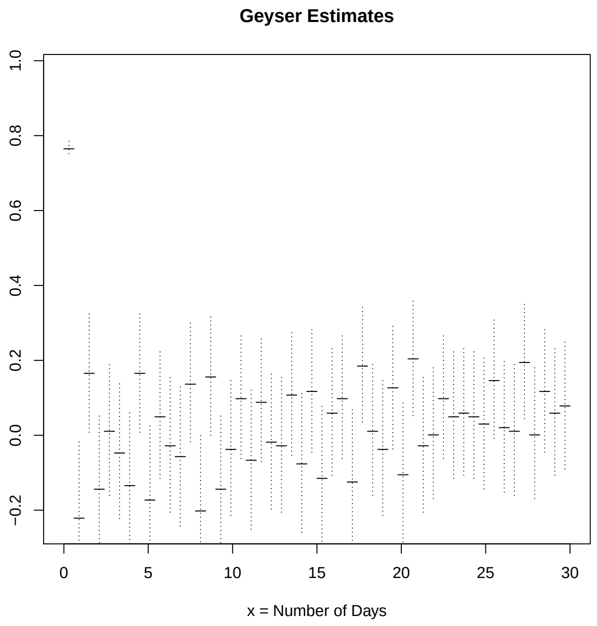
The following information about the Lone Pine Geyser is from the Geyser Observation and Study Association. The eruptions for the Lone Pine Geyser are relatively predictable with an eruption occurring every 24 to 27 hours. The duration for each eruptions is about 20 minutes and can reach around 75 feet high. This would be considered a major eruption. There can also be minor eruptions an hour or two after a major eruption. These minor eruptions do not appear to significantly affect the refill rate. The data logger place downstream records information every minute during the summer months, April to November, and every 3 minutes during the winter months, December to March. The data logger detects changes in temperature, a rise in temperature indicates an eruption. Due to ice formation during the winter months, some eruption information may be lost due to the increased difficulty of detecting temperature changes.

According to the Geyser Observation and Study Association the eruption data for 2011 begins during the winter months, and due to an error with the data logger is missing a month of recordings from June 14 to July 14. To compensate for the month of unrecorded eruptions, the recorded time stamps after July 14 were moved forward month by subtracting the time difference between the last recording in June and the first recording in July. The 2011 data set for the Lone Pine Geyser has 163 recorded eruptions.

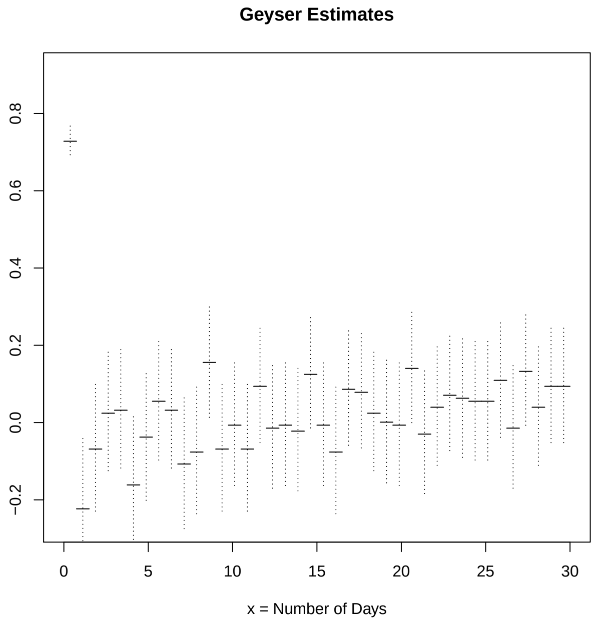
Figure 4.1 shows the estimates and 95% confidence bounds for the geyser data with an interval from 0 to 30 days. The estimates appear to oscillate a bit before converging near zero at the end of the interval. There is one estimate, the first one, which is noticeably larger than the rest. In Figure 4.1a, there is only one observed eruption in the first bin. For each bin in the interval , the expected number of eruptions due to background is approximately 80. In Figure 4.1b, there are six observed eruptions in the first bin while the expected due to background is approximately 100. By having only a few observed eruptions within a bin the resulting confidence bounds are relatively small compared to other bins. This initially large estimate is followed by a convergence to zero which seems to show some inhibition similar to the Exponential or Pareto density.

Figure 4.1: The estimates and 95% confidence bounds for the Lone Pine Geyser data over a 30 day interval.

(a) Estimates for a bin size of .6 (14 hours and 24 minutes)



(b) Estimates for a bin size of .75 (18 hours)



CHAPTER 5

Conclusion

Based on how well the estimates recovered the various probability density functions used to create several simulations with different peaks, it appears the estimates with their confidence bounds are fairly robust. It is clear that number of events within a bin determines how well the estimates will capture the underlying density. Knowing this, the choice of bin size is crucial to the performance of the estimates. As long as the bin size is large enough to contain enough events and small enough to not lump too much information together the estimates computed from the non parametric algorithm should recover the inherent density of the data set. In general, the performance of the algorithm will improve as the size of the data set being worked with increases.

In the future, a way to determine the best interval length and bin size would prove incredibly useful. At the moment it begins as an educated guess and is refined from there. It would also be beneficial to test the algorithm with other simulated data sets and geyser eruption data set of various sizes.

References

[1] Marsan, D., and Lengliné, O. (2008). *Extending Earthquakes's Reach Through Cascading*, Science 319, 1076-1079.

[2] Marsan, D., and Lengliné, O. (2010). *A New Estimation of the Decay of Aftershock Density with Distance to the Mainshock*, Journal of Geophysical Research 115.

[3] Hawkes, A. G., *Spectra of Some Self-Exciting and Mutually Exciting Point Processes*, Biometricka 58, 83-90, 1971

[4] Daley, D.J., and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes: Volume 1: Elementary Theory and Methods, Second Edition*, Springer-Verlag New York, New York, 19-40 & 66-106.

[5] Lewis, P.A.W., and Shedler, G.S. (1978). *Simulation of Nonhomogeneous Poisson Processes by Thinning*, Naval Postgraduate School.

[6] Bryan, T. Scott (1995). *The Geysers of Yellowstone.* University Press of Colorado, Niwot.

[7] Rinehart, John S. (1980). *Geysers and Geothermal Energy.* Springer-Verlag New York, New York.

[8] Taylor, R., and Yellowstone National Park. *Lone Pine*, http://www.geyserstudy.org. The Geyser Observation and Study Association.

[9] Ogata, Y. (1988), *Statistical Models for Earthquake Occurances and Residual Analysis for Point Processes.* Journal of the Americal Statistical Association, Vol. 83, No.401, 9-27.

[10] Ogata Y. (1998), *Space-time Point-process Models for Earthquake Occurrences*, Annals of the Institute of Statistical Mathematics, Vol. 50, No.2, 379-402.

[11] Lewis, E., and Mohler, G. (2011), *A Nonparametric EM Algorithm for Multiscale Hawkes Processes*, Journal of Nonparametric Statistics, Vol. 00, No.00, 1-16.

[12] 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 Math, to appear.

[13] Balderama, E. (2012), *Spatial-Temporal Branching Point Process Models in the Study of Invasive Species* University of California, Los Angeles.