Master's Thesis

AUTOMATIC DECOMPOSITION OF SELF-TRIGGERING KERNELS OF HAWKES PROCESSES

Rafael Gonçalves de Lima

Department of Computer Science and Engineering

Graduate School of UNIST

2017

AUTOMATIC DECOMPOSITION OF SELF-TRIGGERING KERNELS OF HAWKES PROCESSES

Rafael Gonçalves de Lima

Department of Computer Science and Engineering

Graduate School of UNIST

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

Automatic Decomposition of Self-Triggering Kernels of Hawkes Processes

A thesis
submitted to the Graduate School of UNIST
in partial fulfillment of the
requirements for the degree of
Master of Science

Rafael Gonçalves de Lima

07. 06. 2017
Approved by

Jaesik Choi

Advisor
Jaesik Choi

Automatic Decomposition of Self-Triggering Kernels of Hawkes Processes

Rafael Gonçalves de Lima

This certifies that the thesis of Rafael Gonçalves de Lima is approved.

07. 06. 2017

Jaesik Choi

Advisor: Jaesik Choi

Committee Member: Se Young Chun

Committee Member: Jun Moon

Abstract

Hawkes Processes (HPs) capture self- and mutual-excitation between events when the arrival of one event makes future ones more likely to happen in time-series data. Identification of the temporal covariance kernel can reveal the underlying structure to better predict future events.

In this work, we present a new framework to represent time-series events with a composition of self-triggering kernels of Hawkes Processes. Our automatic decomposition procedure is composed of three main steps: (1) discretized kernel estimation through frequency domain inversion equation associated with the covariance density, (2) greedy kernel decomposition through four base kernels and their combinations (addition and multiplication), and (3) automated report generation. In addition, we report the first multiplicative kernel compositions along with stationarity conditions for Hawkes Processes. We demonstrate that the new automatic kernel decomposition procedure performs better to predict future events than the existing framework in real-world data.

Contents

List of Figures

List of Tables

List of Abbreviations

CIF Conditional Intensity Function. 8, 9, 11, 14

HP Hawkes Process. 1, 3, 6, 8–11, 13, 14, 17, 48

HPP Homogenous Poisson Process. 6, 8, 9

NHPP Non-homogenous Poisson Process. 6

PP Poisson Process. 8, 9

Chapter I

Introduction

Point Processes [1] have been used for modeling time-events series data. Hawkes Processes (HP) [2] are point processes for modeling self-exciting behavior, i.e., when the arrival of one event makes future events more likely to happen. This type of behavior has been observed in various domains, such as earthquakes, financial markets, web traffic patterns, crime rates [3, 4] and social media [5].

As an example, in high-frequency finance, buyers and sellers of stocks demonstrate herding behavior [6, 7]. After the main earthquake, several aftershocks follow according to a time-clustered pattern [8]. In web data, hyperlink proliferation across pages exhibit self- and mutual-excitation [9]. In criminology, gang-related retaliatory crime patterns are grouped in time [3]. In social media, the ‘infectiousness’ of posts can be shown to be modeled through a self- and mutual-excitement assumption [5].

In Hawkes Processes analysis, some parametric kernels capture intra-domain typical behaviors, e.g., quick time-decaying exponential excitation in the case of finance and web data [10, 9]; slower power-law decay in earthquake-related data [8]; and periodicity-inducing sinusoidal kernel in TV-watching data [11].

Choosing a proper kernel type may significantly contribute to learning the model parameters and better predicting future events. Kernel parameters may be fitted in the data through the gradient descent method over a likelihood function penalized by a regularization criterion (e.g., Akaike Information Criterion) on the number of parameters [12]. Another method of kernel estimation is through the use of the power spectrum of the second order statistics of the process: covariance density and normalized covariance [2]. These are well defined when the self-triggering function induces what is called *stationary behaviour*.

In this work, we present an automatic kernel selection and learning algorithm for Hawkes Processes. Given four types of base kernels (EXP,PWL,SQR and SNS), our algorithm finds the best fitting kernel. For handling composite types of events, we develop a new kernel decomposition method which represents the composition (sum and product) of different kernels. For verifying the stationarity property of each composite kernel, we also derived analytical expressions for the stationarity conditions. To our best knowledge, this is the first multi-class kernel decomposition framework for HPs.

The main steps of the automatic framework, which will be thoroughly explained in the following sections, are then: discretized kernel estimation and greedy kernel decomposition. An automatic framework for extracting relevant typical features of self-excitement from raw data would likely make the Hawkes Processes analysis faster and conveniently expressed through a human-readable report. Thus, the automated generation of such a report is also discussed.

Automatic analysis frameworks for Gaussian Processes (GPs) are proposed in [13] and [14]. However, due the fundamental distinctions between GPs and HPs (such as stationarity conditions and causality assumptions for the latter), the techniques proposed for GPs can not be extended to HPs in a straightforward manner. We include additional references in the Chapter 8.

In the following, we give a general view of Point Processes in Chapter 2. In Chapter 3, we review the Hawkes Processes models. In Chapters 4 and 5, we will introduce the steps of our kernel decomposition algorithm up to two kernels, together with a theoretical analysis of when there may be a composition of more than two kernels. Chapter 7 includes experiments on our decomposition method, followed by conclusions in Chapter 8.

Chapter II

Point Processes

2.1 Introduction

Point processes are probabilistic models, for capturing temporal and spatial characteristics of discrete event arrivals, which arose from the need to deal with integer-valued counting of events such as queues (people arriving in a store) [1], earthquake shocks [8], war death counts [15] and paper citations [16].

This chapter intends to introduce the thesis goal and serve as a general view of the field of Point Processes, with its main qualitative and quantitative aspects being loosely explained along with a good number of real-world examples, for facilitating understanding. It is divided into the following sections:

- Problem Definition and Thesis Goal
- Examples of Point Processes
- Point Processes on the Real Line

2.2 Problem Definition and Thesis Goal

The problem to be studied is the probabilistic prediction of future events of some given type. Given a sequence of said time events, represented through a vector with their corresponding time coordinates, as in:

$$(t_1, t_2, ...t_n),$$

we are interested in predicting the probability of a future event at t' ($t_n < t'$).

In the case of the present work, the types of events studied exhibit self-exciting behaviour, which is when the occurrence of an event makes further events more likely to happen. This type of behaviour, which will be further explained in Chapter 3, can successfully model several types of data. As an example:

- In high-frequency finance, buyers and sellers of stocks demonstrate herding behaviour [6, 7].
- After the main shock, several earthquakes' aftershocks follow according to a time-clusterized pattern [8].

Given our goal of analyzing this type of behaviour, we propose a framework for finding the best predictive model with compositions of heterogeneous types of self-excitation between events, developing a way to:

- Automatically interpret self-exciting behaviour in several domains of data, in a **multi-type** approach
- And make it available to the general public, in the form of an automatically generated report

2.3 Examples of Point Processes

This section aims on giving a brief and intuitive view of the main aspects and concepts related to Point Processes. They are strongly related to the modeling of a variety of real-world situations. As some examples of its importance and pervasiveness on real world phenomena:

- People arriving in a store can be modeled as a sequence of event arrivals, with a time coordinate associated with each person. The rate of arrivals will most likely vary throughout daily, weekly and yearly periods, and properties such as this seasonality of arrivals are well-studied subjects on the field of Queueing Theory.

- Given that fixing machines deals mainly with replacing whole parts, the times of replacements of broken mechanical parts of some machine can be understood as a temporal sequence of discrete events, in which the length of each consecutive interval is independent of the length of the previous ones, being sampled from a probability distribution such as Exponential. This is the main concept behind the subject of Renewal Processes, a subtype of Point Process.

- Spatiotemporal patterns of occurrences of earthquakes have long been studied, mostly in countries prone to high seismic activity, for very practical reasons, with the clustering patterns of both geographical and temporal coordinates of the shocks also being modeled through point processes. In this case, the coordinates of each shock will be associated with the value of its magnitude, usually measured in the Richter scale. This magnitude coordinate can be treated as a so-called *mark* of the event, and the sequence of shocks may then be modelled as a Marked Point Process, a subtype of point process.

- In [15], civilian deaths in Iraq, along with their corresponding times, were modeled as point processes with arrival rates which varied in predictable ways.

- By considering criminal activities with its spatio-temporal coordinates stored as a sequence, and being influenced by criminological research demonstrating a contagion-like dynamics in which crimes can spread through local environments, s.a., burglars repeatedly attacking clusters of nearby targets because the offenders may then be well acquainted with local vulnerabilities, and gang shooting inciting waves of retaliatory violence in the territories of rival gangs, [4] proposes a point process for modeling this formation of crime clusters in space and time, with regularities similar to those of earthquakes' time series. A *Predictive Policing* software, denominated 'Predpol', based on this model, predicted criminal occurrences at twice the rate of the police department's experienced crime analysts, in the city of Los Angeles.

- Determinantal Point Processes (DPP) enjoyed a recent surge in popularity in Machine Learning research,for introducing repulsion and consequently diversity in datasets, thus potentially increasing the generalization power of ML algorithms.[17]. By treating a dataset as a set in which more similar data points are located closer to each other, and considering the instantiation of a specific

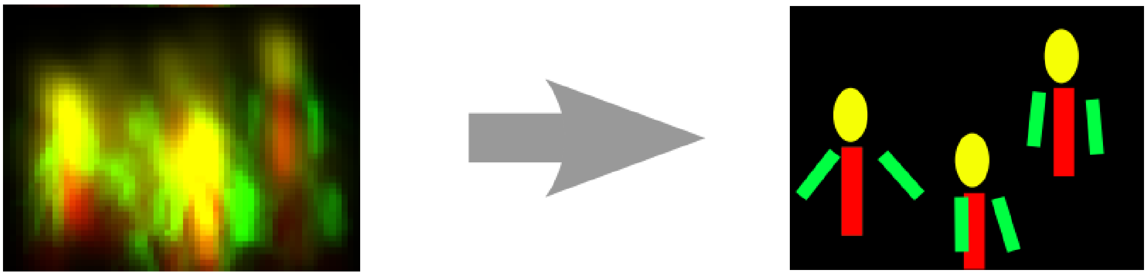


Figure 2.1: Applying diversity to a human detector, previously noisy and uncertain, leads to a separated, cleaner set of predictions. Source: (Kulesza, 2012).

data point as a point process event, a DP works by influencing the instantiation in a way that data points close to the currently selected point are inhibited, forcing the next event to come from a region of less similarity, therefore increasing the diversity of the training data sequence. The usefulness of this repulsion effect is also demonstrated in detection algorithms (see Figure 2.1)

- Academic citation counts may also fall under the scope of Point Process modeling. By considering each time in which a specific paper is cited as being part of a sequence of events, [16] proposes a point process model for predicting and analyzing its future scientific impact.
- Social Media activity can also be understood as a point process. By considering the time coordinate of an specific action, one may obtain sequences of discrete events. [5] then shows how the ‘tweeting’ activity of highly influential people will propagate itself in a cascaded way throughout its ’followers’, then the ’followers of its followers’, and so on.

2.4 Point Processes on the Real Line

Perhaps the type of point process with the most intuitive representation, point processes on the Real Line are those in which the Lebesgue measurable space taken into consideration is a subset of $\mathbb{R}^1$, usually considered the time dimension.

The four possible interpretations given to this kind of Point Process are in terms of:

- Sequences of time intervals
- Counting measures
- Sequences of points
- Non-decreasing integer-valued step functions

When analyzing a point process as a Counting Measure, it is irrelevant whether the process is described on the real line or not. For the three other ways of defining the process, however, the order properties of the reals are considered in an indispensable way. Although the said methods of description may be

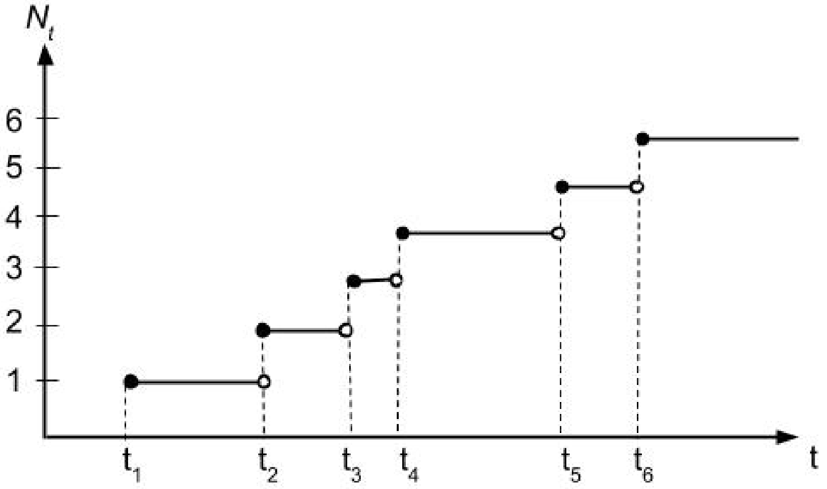


Figure 2.2: An example of a Counting Process N(t).

capable of being extended into dimensions of higher order, they become unavoidably less natural and, regarding the case of the 'sequences of intervals' description, decidedly artificial [1].

A point process representation of a sequence of n time-events is expressed by a vector of the form (t_1,t_2, ... , tn). Treating the real line as a time axis, the vector can be intuitively associated with a so-called Counting Process $N(t)$, s.t.:

- $dN(t) = 1$, if there is an event at time t;
- $dN(t) = 0$, otherwise.

An example of Counting Process is shown in Figure 2.2. We are going to see it later that the dN_t on real line allows for giving a explicit analytical expression for the intensity function of Hawkes Processes (HPs).

This type of point process may be described through its temporal intensity function ($\lambda(t)$), which can be understood as the instantaneous expected rate of arrival of events, or the expectation of derivative of the Counting Process $N(t)$:

$$\lambda(t) = \lim_{h \to 0} \frac{\mathbb{E}[N(t+h) - N(t)]}{h} \tag{2.1}$$

This intensity function, if existing, uniquely characterizes the finite-dimensional distributions of the point process [1]. A simple example of this function would be the constant mean rate of arrival, μ, in the case of a Homogeneous Poisson Processes (HPPs).

In some cases, it is reasonable to assume that this rate of arrival will vary throughout the period of observation, what is the case of the so-called Non-homogeneous Poisson Processes (NHPPs). Consider, for example , the number of people arriving at a restaurant, which is bound to peak strongly at typical mealtimes but drop throughout the remaining hours. This example is illustrated at Figure 2.3, from [18].

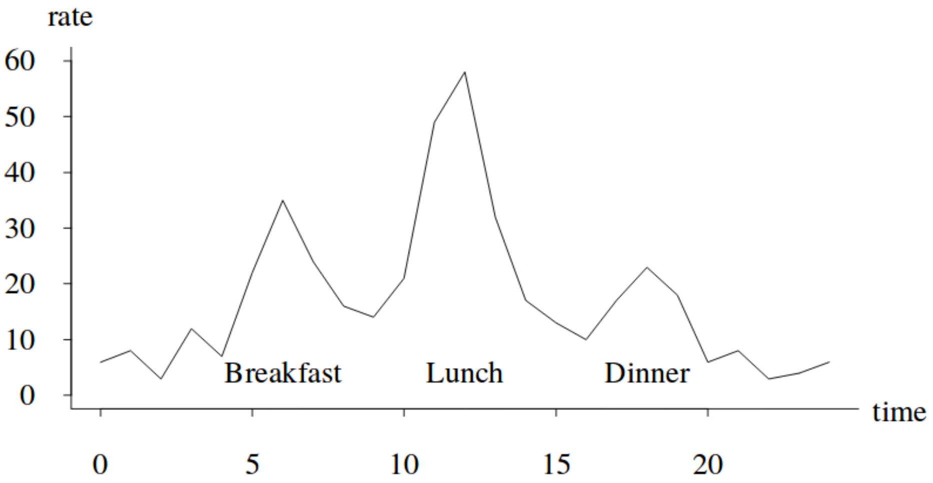


Figure 2.3: An example of a Non-homogeneous Poisson Process. Since the number of people arriving at a restaurant throughout a day peaks during mealtimes and drops during the remaining periods, it is reasonable to assume that the underlying intensity varies with time.

Chapter III

Hawkes Processes

3.1 Introduction

Being arguably the simplest type of Point Process, Poisson Processes (PPs), both in the Homogeneous and Non-homogenous case, have in common the fact that each event is independent of all the others, this being even more evident in the case of processes in the real line.

By glancing back at the examples given in chapter 2, one may notice that many of those point to the possibility of interactions between events: Criminal activities in a certain neighborhood making subsequent ones more likely, Earthquake shocks triggering further aftershocks around the same location, among others.

These may be modeled by a process described with a so-called Conditional Intensity Function (CIF), i.e., a function corresponding to the rate of arrivals in a specific instant in a way that it depends on previously occurred events, the *History* of the process. Moreover, we are mostly interested in those processes in which one event makes future ones more likely, a phenomenon denominated *self-excitation* or *self-triggering*.

HPs, a more general type of point process, which deal with self-exciting phenomena in a rather elegant and analytical way, are the subject of the present chapter.

3.2 Mathematical Definition

Hawkes processes model the intensity function in terms of the self-excitement: when the arrival of an event makes subsequent arrivals more likely to happen [19]; and can be described through the following conditional intensity function $\lambda(t)$:

$$\begin{aligned}\lambda(t) &= \lim_{h\to 0}\frac{\mathbb{E}(N(t+h)-N(t)|\mathcal{H}(t))}{h}\\ &= \mu + \int_{-\infty}^{t}\phi(t-u)dN(u),\end{aligned} \tag{3.1}$$

where:

- $\mathcal{H}(t)$ is the history of the process, the set containing all the events up to time t;
- μ is called *background rate*, or *exogenous intensity*, which is usually fixed as the mean of a HPP;
- $\phi(t)$ is denominated *self-triggering kernel*, or *excitation function*;

From this function, one may notice that the intensity at time t will likely be affected by events which happened before the time t, described by the history of the process. Furthermore, the self-triggering kernel function is assumed to be:

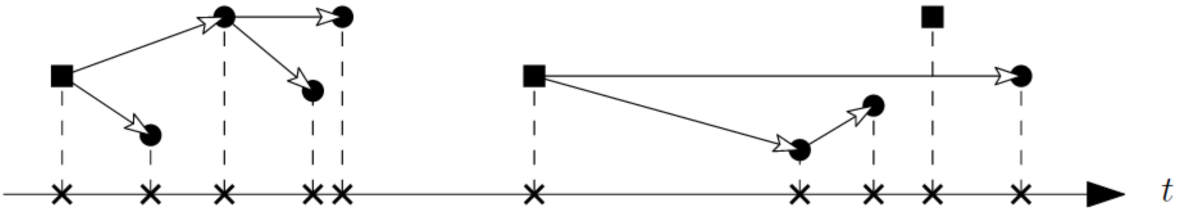


Figure 3.1: "Immigrant-Birth Representation" of a Hawkes Process. A branching process, evidencing the immigrants (squares) and its descendant events (circles). Source: (Laub, 2015).

- Causal,

$$\phi(t) = 0, \text{ for } t < 0$$

- Positive:

$$\phi(t) \geq 0, \text{ for } t \geq 0$$

The first assumption certifies that the CIF of the process is completely defined by its History, while the second one ensures that past events always act in the direction of increasing the intensity, thus exciting, or triggering, the process. Negative-valued kernels are assumed in the case of the so-called self-damping, or self-correcting, processes, which are outside the scope of this work. One may also notice that, if the kernel function is null for all t, the intensity function reduces to that of a HPP.

This expression for the CIF is also written with the convolution operator ('$\star$'):

$$\lambda_t = \mu + \phi_t \star dN_t, \tag{3.2}$$

where the convolution operator corresponds to:

$$A \star B_t = \int_{\mathbb{R}} A_s B_{t-s} ds = \int_{\mathbb{R}} A_{t-s} B_s ds \tag{3.3}$$

From [2], we have that, if :

$$||\phi||= \int_0^\infty \phi(t)dt \leq 1, \tag{3.4}$$

The stationarity criteria can be quickly understood through the point-of-view of what is defined as *Immigrant-Birth Representation* [20]. This accounts for defining a realization of a HP as a combination of events caused by the exogenous intensity, called "Immigrants", and events generated from the excitation effects from these immigrant events, the "Descendants". This derives from the fact that, in essence, a HP realization is a recurrent superposition of PPs: The first PP would be from the baseline intensity considered alone. Then, for each resulting event, a further PP, with the intensity function defined by the kernel function, would result. Then, further realization will be extracted from each of these resulting descendants, indefinitely. Figure 3.1 is a visualization of this concept. The integral expression for the kernel function is the expected number of events in the corresponding process realization which will be caused by a specific event.

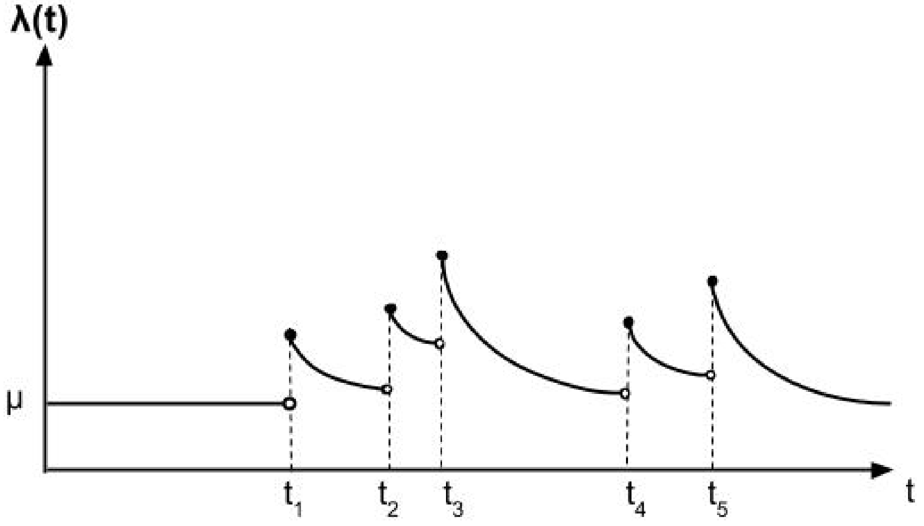


Figure 3.2: An example of an Intensity Function of a self-triggering Point Process $\lambda(t)$.

From this recurrence relation, it is straightforward to see that each event caused by the background rate will result in a tree of descendant events. For the process to no 'blow up', it is necessary that the resulting tree is of finite length, what is only achieved when the stationary criteria, here seen as the branching ratio of the tree, is less than 1.

If the expression in 3.4 is satisfied, then the corresponding process will reach wide-sense stationary behavior, from which the asymptotic steady arrival rate:

$$\Lambda = \frac{\mu}{(1 - ||\phi||)}, \tag{3.5}$$

can be obtained, along with its covariance function, which is independent of t:

$$\nu(\tau) = E(dN(t)dN(t+\tau)). \tag{3.6}$$

Estimating Λ and $\nu(\tau)$, also referred to as first- and second-order statistics, respectively, requires wide-sense stationarity assumptions which, besides being analytically convenient, are also connected to the fact that, in real data, the chain of self-excitedly induced further events will always be of finite type, or without 'blowing up', what corroborates the practicality of the estimated model.

It is worth mentioning, though, that some point processes may have statistics of first- and second-order satisfying stationarity assumptions without consequently being stationary, i.e., non-stationary processes may possess stationary first- and second-order moments [1].

3.3 Methods of Kernel Estimation

Regarding the HPs' field of study, much of the research efforts are directed towards modeling the background rate and kernel function best suited for a given sequence of events.

Accurately estimating the kernel structure may help with understanding the underlying dynamics of the self-excitation and also with predicting future events. These kernel estimation methods can be roughly divided between:

- **Parametric Methods**: in which the estimated kernel function is represented by a continuous function, e.g., an exponentially-shaped function:

$$\phi(t) = \alpha e^{-\beta t} \tag{3.7}$$

, with parameters α and β.

- **Nonparametric Methods**: in which the output kernel estimate is a finite grid, with its values computed at specific coordinates, e.g.:

$$\phi(t) = \begin{bmatrix} \phi(0.1) & \phi(0.34) & \phi(0.42) \end{bmatrix} = \begin{bmatrix} 0.073 & 0.036 & 0.032 \end{bmatrix} \tag{3.8}$$

In the following, we shall give a concise account of the main peculiarities of each type. Then, in Section 3.4, we briefly explain how these different methods are approached regarding some intra-domain problems in the topics of High-frequency Finance, Earthquake Frequency Modeling and Network Analysis of Web Data.

3.3.1 Parametric Methods

Gradient Descent

Originally proposed in [12], for an exponential kernel, this method starts by assuming a continuous parametric function for the kernel.

By this, it is possible to go from Equation 3.2 and a given sequence of events $(t1, ..., t_n)$ to an analytical calculation of the value of the CIF at any point of the time interval for which the sequence is considered. From this CIF expression, it is possible to compute the likelihood value for the sequence, i.e., the probability that the referred sequence was generated by a HP with the assumed parameters [1].

The method then proceeds by attempting to tune the randomly initialized kernel parameters in a way that this likelihood, here in its logarithmic form ('log-likelihood'), is maximized. As an example, for the exponential kernel of Equation 3.7, these derivatives would be calculated w.r.t. μ, α and β.

A pseudocode of the procedure is given in Algorithm 1.

Algorithm 1 Gradient Descent based Kernel Estimation

```
Input learning rate, sequence of events, and initial parameters
Calculate gradient of log-likelihood formula w.r.t. each parameter
Update each parameter with gradient magnitude weighted by the learning rate
Stop when magnitude of gradient is too small or too many iterations were reached
```

[1] The derivation of the likelihood formula in terms of the CIF is given in B

3.3.2 Nonparametric Methods

Expectation-Maximization

For this method, the kernel is defined as a finite grid with a fixed set of m points, equally spaced by time intervals δt. It works by iterating along two steps towards convergence of the values for the background rate and the kernel grid [21, 15].

For a sequence composed of n events, the two steps are:

- Expectation: Calculating probabilities of event j being generated by the background rate (p_{jj}) and being generated by event i (p_{ij}), what results in a n-by-n probability matrix s.a.[2]:

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

  where

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

  and

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

- Update values of mμ and ϕ according to these probabilities:

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

  and

$$\phi(m\delta t)^{k+1} = \frac{1}{\delta t} \sum_{i,j \in A_m} p_{ij}^{k}$$

  where A_m corresponds to all possible pairs of events which satisfy $m\delta t \leq |t_i - t_j| \leq (m+1)\delta t$.

Repeating these steps up to convergence leads to final estimates of μ and the kernel grid.

A variant of this method [15] calculates the matrix of probabilities in the same way, but the Maximization step is carried out by solving a discretized Ordinary Differential Equation with a regularization term for enforcing smoothness of the obtained solution.

Covariance Spectrum-based

In [2], the relations between the Fourier Transforms of both the stationary covariance $\nu(\tau)$ and the excitation function $\phi(t)$ are derived while assuming that the latter belongs to a class of exponential functions,

[2] Note that, by the assumption of causality on the kernel, an event can not have been generated by an event which happened after it, what results in the triangular aspect of the matrix.

i.e., given a $\phi(t)$, assumed to be a linear compositions of exponentials satisfying the stationarity property in 3.4, it is shown to be possible to analytically derive properties of the covariance [3].

The method proposed in [7] works in the inverse direction: Trying to obtain an estimate of $\phi(t)$ through an estimate of $\nu(t)$. In practice, this is carried out by solving a so-called Wiener-Hopf equation, s.a.:

$$m(t) = \phi(t) + \phi(t) \star m(t), \tag{3.9}$$

where

$$\nu(t) = \Lambda m(t). \tag{3.10}$$

In practice, this equation is solved through discretization (Nystrom Method) with the Numerical Integration technique of Gaussian Quadrature. A variant of this method was proposed in [10], and will be throughly explained in Chapter 4.

3.4 Examples in Real-World Data

Effective modeling of interactions between sequential events, such as the one provided by the HPs framework, has been found useful in a variety of domains. In this section, we concisely explain the underlying empirical assumptions made on the models, and more specifically on the self-triggering kernels, when working with Financial, Earthquake and Social Network related data.

3.4.1 Hawkes Processes in Finance

Previous empirical studies have provided evidence that the impact of price itself has, on many aspects, some universal properties and is the main source of price variations, what reinforces the notion of a endogenous nature, in other words, a internal feedback mechanism, of price fluctuations. This is contrasting with the classical notion of an external, exogenously generated, flux of information directing the prices towards a primary value [22].

In this endogeneity scenario, Hawkes Processes have become more and more present in the high-frequency finance domain, due to its structure being naturally adapted to model systems in which the discrete nature of the jumps in N_t is relevant, making the model remarkably suited to modeling high-frequency data [23]. Moreover, Hawkes Processes' parameters, along with their corresponding straightforward interpretations, lead to a notedly simple and flexible interpretation of the complex intraday dynamics of modern electronic markets.

When modeling financial data with HPs, one is mostly concerned with real-time, or 'tick', data, which is more closely related to the aforementioned intra-day dynamics. The goal is to verify how strongly the price variations in a specific stock affects further transactions ('Buys' or 'Sells') on itself or in another stock. Also, since most of the small variations in the price must be due to noise, only large

[3] The Fourier Transform $\hat{f}(\omega)$ of a time function f(t), $\hat{f}(\omega) = \int_{\mathbb{R}} f(t) e^{-i\omega t} dt$, basically consists of decomposing it into its basic frequencies.

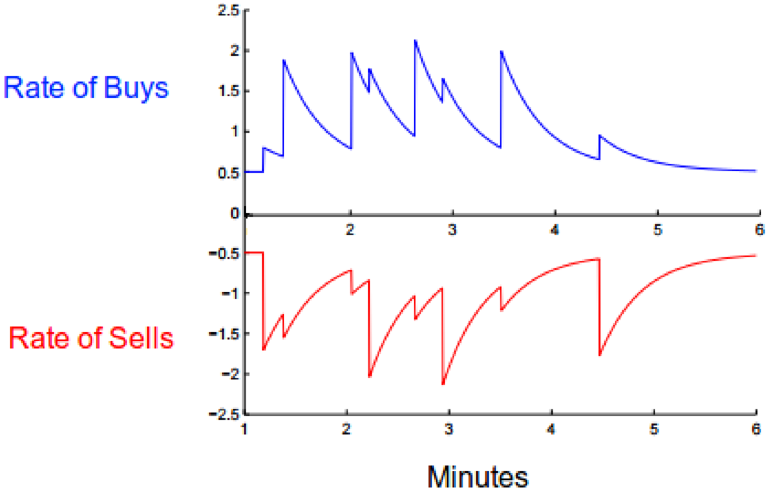


Figure 3.3: Buy (in blue) and Sell (in red) operations in a stock may be seen as two complementary HPs, with price jumps increasing the arrival of Buys while inhibiting the Sell operations

enough jumps are considered, what is done through thresholding over a absolute or percentual value. The Buys and Sells may be considered as the same type of event or separately, as in Figure 3.3[24].

When modeling High-Frequency Financial data, the kernels are usually assumed to be of exponential type. Some recent works, however, point to the better performance of Power-Law shaped kernels in some contexts [25].

3.4.2 Hawkes Processes in Earthquake/Seismicity

Earthquakes' occurrences around some area are also modeled as a point process sequence. Omori's Law (1879) already conjectured that the sequence of aftershocks following a major earthquake shock would distribute itself according to a power-law shaped intensity, s.a:

$$n(t) = \frac{K}{(t+c)^p}, \tag{3.11}$$

where n(t) represents the frequency of aftershocks per unit time interval.

The proposed ETAS (Epidemic-Type Atershock Sequences) model works a slightly different Power-Law shaped self-triggering kernel for the resulting CIF [8]:

$$\lambda(t|H_t) = \mu + \sum_{t_i<t} \frac{K}{(t-t_i+c)^p} \tag{3.12}$$

The parameters of this CIF are fitted through Gradient-Descent based methods, s.a. the ones in subsection 3.3.1. Some additional Power-Law functions may be added to the kernel function, along with some regularization parameter, which will determine if the improvement of the likelihood probability surpasses the computational cost of dealing with a more complex model.

3.4.3 Hawkes Processes for Networked Analysis of Social Networks

Regarding Web Data, the modeling focuses on social networks' actions and hyperlinks insertions on pages. Both of these are treated as hierarchical actions, with 'followers' being influenced by the actions of popular profiles or websites but having little influence in their primary action.

By treating a 'retweet' [5] or a hiperlink insertion [9] as a time event, it is possible to quantify the strength of the influence of some website or celebrity by the number of descendant events it is most likely responsible for.

Chapter IV

Discretized Kernel Estimation

4.1 Introduction

This chapter aims on providing all assumptions, methods and derivations regarding the extraction of a discretized kernel estimation from raw time data sequences. All the steps are originally developed in [10] for the case of symmetrically networked multivariate Hawkes Processes, and we intend here to give a through description of the underlying theory of each step of the method for the univariate case.

Here, the discretized kernel estimation may be optional when a direct optimization of kernel structure is possible. Unfortunately, discontinuous functions (SQR, SNS) do not allow such optimization (e.g., gradient descent). In this work, we use the discretized kernel estimation as an unified method for both continuous (EXP, PWL), discontinuous kernels (SQR, SNS) and, most importantly, their combinations.

Furthermore, another great advantage of this step, compared with traditional sequential methods, is the fact that the value of ν for each value of τ can be calculated independently, while, in gradient descent, the value of the parameters at step t must be obtained before the values for step $t+1$. When combined with parallelization of the loops, in our case, this step greatly improved the speed of obtainment of the most likely parametric representations of the sample processes.

One of the key principles of the method is the one-to-one relation between first- and second-order statistics of stationary HP (normalized stationary temporal covariance) and its corresponding causal self-triggering kernel function, i.e., given a certain pair of first- and second-order statistics, obtained under the stationarity assumption, one can find the only corresponding causal triggering kernel function which induces the aforementioned stationarity. This is explained in the following sections:

- Relation between first- and second-order statistics and kernel function
- Empirical nonparametric estimation of covariance
- Hilbert Transform and Cepstral Analysis
- Phase Indeterminacy and Minimal-Phase condition

4.2 Relation between first- and second-order statistics and kernel function

The aim of this section is to show the relation between the temporal covariance function and the magnitude of the kernel function. It is shown, in [2], that this relation admits closed-form solutions for the case of kernels expressed as summations of exponential terms.

In the case of a more generally shaped kernel, this relation leads to a phase indeterminacy problem which is solved with the help of stationarity assumption and Hilbert Transform-based cepstral analysis techniques for minimal phase filter estimations, i.e, based on the stationarity assumptions, one can show that the true kernel which satisfies the indeterminate phase equation is the minimal phase solution, which is obtained with widely used cepstral analysis techniques.

Regarding Point Processes in general, the covariance function, or second-order statistics, may be described in more than one way. And, in fact, there does not seem to exist a class of point processes with tractability, similarly to Gaussian Processes [26], whose properties of second-order are coextensive with those of general point processes, with HPs getting close to fulfilling this role, but without entirely doing so [1].

The analytical relation between covariance and kernel function developed in [2] for a class of exponential functions was a very remarkable result regarding this topic. In some cases, it may be equivalent simply to the variance on the length of the time intervals among consecutive events. In the present work, however, the concept of covariance corresponds to the expectation of the product of the number of arrivals in different times of the sequences. The nonlinearity of positivity and integer counting of these number of arrivals is usually associated with difficulties surrounding treatment of moment measures in Point Processes.

For expressing the modulus of the kernel function in terms of the stationary covariance, it is convenient to represent the Counting Process function of a stationary process in terms of the asymptotic mean, which would be its expected value, and an uncorrelated variation term of mean zero (i.e., a martingale):

$$dN_t = \lambda_t dt + dM_t \tag{4.1}$$

Using this relation, we may express equation 3.2 as:

$$\lambda_t = \mu + \phi \star dN_t = \mu + \phi \star \lambda_t + \phi \star dM_t, \tag{4.2}$$

from which we have that:

$$(\mathbb{I}\delta - \phi) \star \lambda_t = \mu + \phi \star dM_t. \tag{4.3}$$

Defining:

$$\psi_t = \sum_{i=1}^{\infty} \phi_t^{\star i}, \tag{4.4}$$

with $\phi^{\star n}$ referring to the n-th order auto-convolution operation,e.g., $\phi_t^{\star 2} = \int_{\mathbb{R}} \phi_s \phi_{t-s} ds$, one may observe that the convolution of ϕ_t by the inverse of $(\mathbb{I} - \phi_t)$ is just the term by ψ_t, from which we get to:

$$\lambda_t = (\mathbb{I} - \phi)^{-1} \star \mu + (\mathbb{I} + \psi_t) \star \phi_t dM_t = \Lambda + \psi_t * dM_t, \tag{4.5}$$

given that the term $(\mathbb{I} - \phi_t)^{-1} \star \mu = (\mathbb{I} - |\phi|)^{-1}$, i.e., just the asymptotic mean, Λ. By the very definition of stationarity, one has that the covariance function is independent of t, i.e., it only depends of the relative spacing τ among the countings. In the stationary case, by expanding the Counting Process (dNt) terms in the Equation 4.2 for the covariance function as a summation of the expected number of arrivals

($\lambda_t * dt$) and a uncorrelated brownian motion (martingale) (dM_t), representing the random variations in this expected value, one gets the resulting four terms:

$$\begin{aligned} \nu_\tau &= E(dN_t dN_{t+\tau}) \\ &= E((\lambda_t dt + dM_t)(\lambda_{t+\tau} d(t+\tau) + dM_{t+\tau})) \\ &= E(\lambda_t dt \lambda_{t+\tau}) + E(\lambda_t dM_{t+\tau}) + E(dM_t \lambda_{t+\tau} d(t+\tau)) + E(dM_t dM_{t+\tau}) \end{aligned}$$

The fourth term is the estimation of two uncorrelated brownian motions, and thus the expectation for $\tau \neq 0$ is simply 0. For $\tau = 0$, the expectation will be the asymptotic mean, Λ, given that the jump sizes of the process dN_t are always of size 1:

$$E(dM_t dM_t) = \Lambda dt, \text{ if } \tau = 0 \tag{4.6}$$

A compact way of describing this term would be using the Dirac delta function:

$$E(dM_t dM_{t+\tau}) = \Lambda \delta_\tau dt d(t+\tau). \tag{4.7}$$

Regarding the term $E(\lambda_t dM_{t+\tau})$, we have that, using equation 4.5 for the intensity function:

$$E(\lambda_t dM_{t+\tau})dt \tag{4.8}$$

$$= E(\psi_t \star dM_t dM^{\dagger}_{t+\tau} \tag{4.9}$$

$$= \int_{\mathbb{R}} \psi_{t-s} E(dM_s dM_{t+\tau}) ds \tag{4.10}$$

$$= \int_{\mathbb{R}} \psi_{t-s} \Lambda \delta_{s-t-\tau} = \psi_{-\tau} \Lambda dt d(t+\tau) \tag{4.11}$$

In a similar way, the term $E(dM_t \lambda_{t+\tau})$ results in:

$$E(dM_t \lambda_{t+\tau}) d(t+\tau) = \Lambda \psi_\tau dt d(t+\tau) \tag{4.12}$$

Using the result for the third term, the first term results in:

$$\begin{aligned} E(\lambda_t \lambda_{t+\tau}) dt d(t+\tau) &= (\Lambda E(\lambda_{t+\tau}) E((\psi \star dM_t) \lambda_{t+\tau}))\, dt d(t+\tau) \\ &= \left(\Lambda \Lambda^{\dagger} + \int_{\mathbb{R}} \psi_{t-s} E(dM_s \lambda_{t+\tau}) ds \right) dt d(t+\tau) \\ &= \left(\Lambda \Lambda^{\dagger} + \int_{\mathbb{R}} \psi_{t-s} \Lambda \psi_{s-t-\tau} \right) dt d(t+\tau) \\ &= \left(\Lambda \Lambda^{\dagger} + \tilde{\psi} \star \Lambda \psi_\tau \right) dt d(t+\tau) \end{aligned}$$

, where $\tilde{\psi}_t = \psi_{-t}$. The full expression is then:

$$E(dN_t dN_{t+\tau}) = (\Lambda\Lambda + \Lambda \delta_{-\tau} + \Lambda \psi_\tau + \Lambda \psi_{-\tau} + \tilde{\psi} \star \Lambda \psi_\tau)\, dt d(t+\tau) \tag{4.13}$$

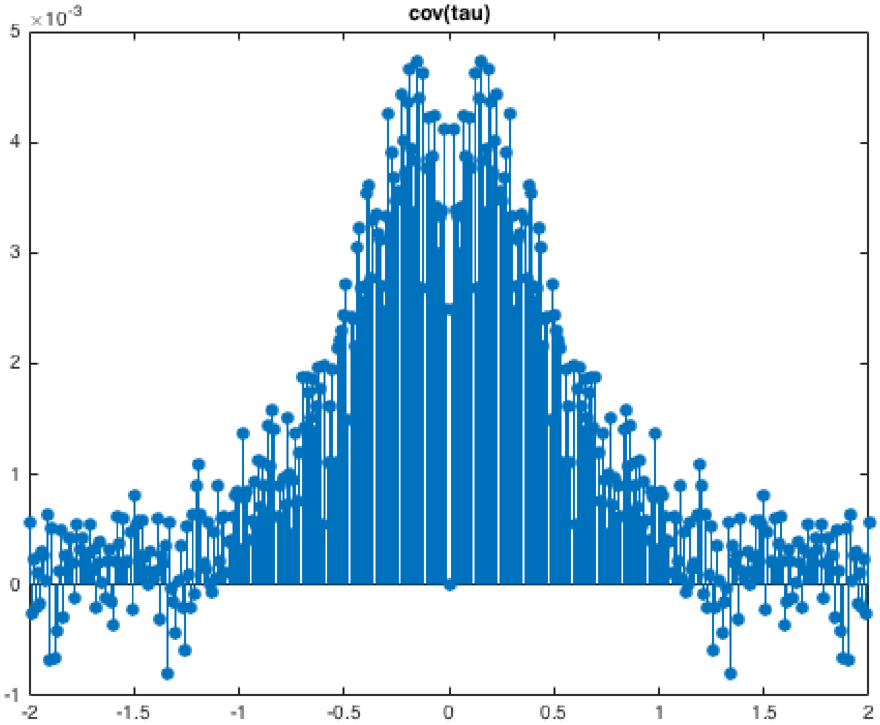


Figure 4.1: Discretized covariance estimate from a sequence generated with exponential kernel.

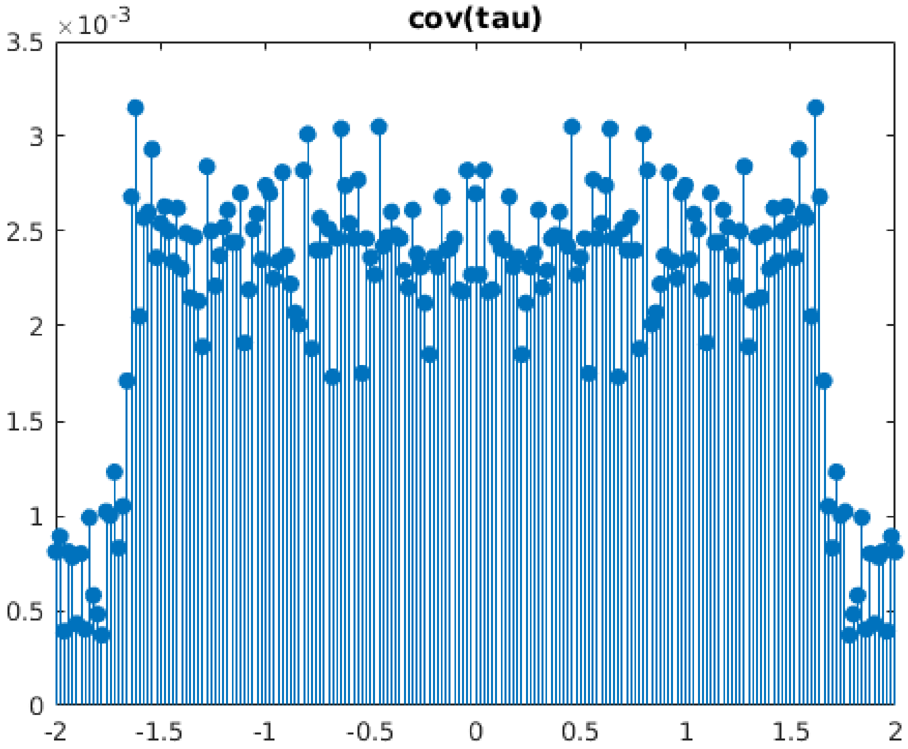


Figure 4.2: Discretized covariance estimate from a sequence generated with square kernel.

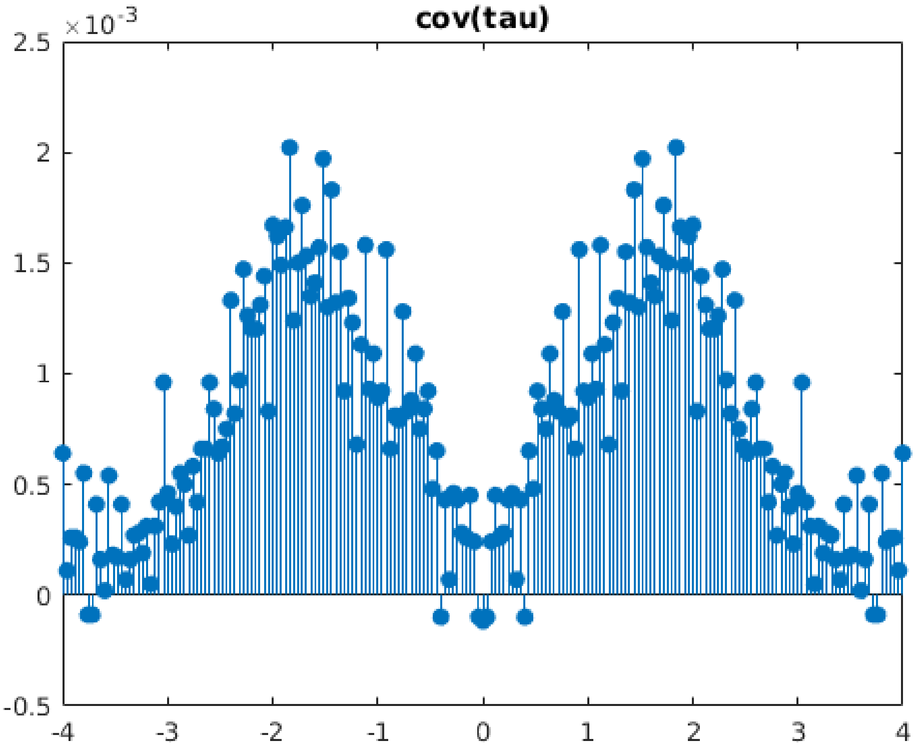


Figure 4.3: Discretized covariance estimate from a sequence generated with sinusoidal kernel.

4.3 Empirical Nonparametric Estimation of Covariance

In practice, the estimation of the covariance from Equation 4.2 is done at discrete time steps δ:

$$v^{(h)}_{\tau,\delta}=\frac{1}{T}\sum_{i=1}^{\lfloor T/\delta\rfloor}(dN^{(h)}_{i\delta}-dN^{(h)}_{(i-1)\delta})(dN^{(h)}_{i\delta+\tau}-dN^{(h)}_{(i-1)\delta+\tau}), \tag{4.14}$$

From [1], we have that, in this case, the extracted function is actually a Kernel Density Estimation of the actual covariance function. The equivalent kernel is a triangular one, with bandwidth h:

$$g^{(h)}_t = \left(1-\frac{t}{|h|}\right)^+, \tag{4.15}$$

as in figure 4.4.

From figure 4.4, one can see that the triangular kernel function can be expressed as the sum of three ramp functions. Working with the Fourier transform restriction, i.e., ($z = i\omega$, with $\omega \in \mathbb{R}$), we have its equivalent expression in the frequency domain:

$$\hat{g}^{(h)}_{i\omega} = \frac{e^{ih}}{z^2h} - \frac{2}{z^2h} + \frac{e^{-zh}}{z^2h} = \frac{2cos(\omega h)-2}{z^2h}$$

$$= \frac{2\left(cos^2\left(\frac{\omega h}{2}\right) - sin^2\left(\frac{\omega h}{2}\right) - cos^2\left(\frac{\omega h}{2}\right) - sin^2\left(\frac{\omega h}{2}\right)\right)}{z^2h}$$

$$= \frac{4}{\omega^2 h}\sin^2(\frac{\omega h}{2}),$$

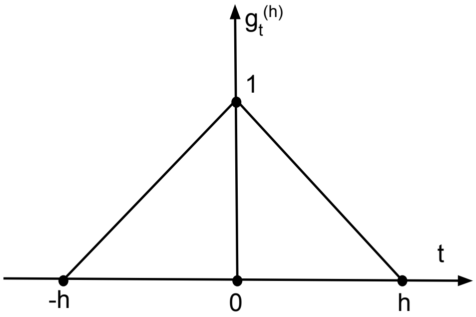


Figure 4.4: Diagram of triangular kernel.

Along with equation 4.13, this result leads to a relation between the estimated covariance and the kernel function. Assuming the window size h for the estimation of the covariance, we have that:

$$\begin{aligned}
\nu_\tau^{(h)} &= \frac{1}{h}E\left(\int_0^h dN_s \int_\tau^{\tau+h} dN_s^\dagger - \int_0^h dN_s\Lambda^\dagger h - \Lambda h\int_\tau^{\tau+h} dN_s^\dagger + \Lambda\Lambda^\dagger h^2\right)\\
&= \frac{1}{h}E\left(\int_0^h \int_\tau^{\tau+h} dN_t dN_{t+\tau} - (\Lambda h)^2\right)\\
&= \frac{1}{h}\int_0^h \int_\tau^{\tau+h} (\Lambda\delta_{-\tau} + \psi_{-\tau}\Lambda + \Lambda\psi_\tau + \tilde{\psi}\star\Lambda\psi_\tau)\, dt d(t+\tau)
\end{aligned}$$

Putting together the assumptions and derivations of sections 4.2 and 4.3 results in:

$$\nu_t^{(h)} = g_t^{(h)}(\mathbb{I}+\psi_t^\dagger)\Lambda(\mathbb{I}+\psi_t^\dagger) \tag{4.16}$$

This can be expressed as:

$$\frac{\nu_t^{(h)}}{\Lambda g_t^{(h)}} = |\mathbb{I}+\psi_t^\dagger|^2, \tag{4.17}$$

which is a phase indeterminacy problem. The solution, to be found in the next sections, will be shown to be, according to the stationarity and causality assumptions for the kernel, the minimal phase solution.

4.4 Hilbert Transform and Cepstral Analysis

This section aims on describing the main concepts related to the Hilbert Transform and the obtainment of the minimal-phase realization of the kernel function. Most of this section, and the next one, draws from the explanations of [27].

4.4.1 Hilbert Transform description

The Hilbert Transform $\hat{f}(t)$ of a function f(t) is defined for all t by:

$$\hat{f}(t) = \frac{1}{\pi}P\int_{-\infty}^{\infty}\frac{f(\tau)}{t-\tau}d\tau, \tag{4.18}$$

when the referred integral exists. This definition in the time domain is a convolution between the Hilbert transformer 1/(πt) and a function f(t).

Usually, it is not possible to analytically calculate the Hilbert Transform description as an ordinary improper integral, because of the pole at $\tau = t$. Generally, what is calculated is the so-called "Cauchy Principal Value" of the integral, the "P" in front of the integral, which is defined as:

$$\lim_{\varepsilon \to 0^+} \left(\int_{\alpha}^{\xi-\varepsilon} f(x)dx + \int_{\xi+\varepsilon}^{\beta} f(x)dx \right) \tag{4.19}$$

It allows the integral in Equation 4.18 to be computed for larger number of functions.

The relation between a real-valued function f(t) and its Hilbert Transform $\hat{f}(t)$ is such that, together, they result in a so-called “strong analytic function”. This “strong analytic signal” is expressed through its amplitude along with its phase, and the derivative function of the phase corresponds to the instantaneous frequency of the signal. When applying the Fourier Transform to this “strong analytic signal”, the resulting spectrum of the signal in the frequency domain is one-sided. We are going to see later that this property is strongly related to the causality of the obtained kernel.

4.4.2 Integrating in the Complex Domain: The Cauchy Integral

It is possible to motivate the Hilbert Transform in a more figurative way through the use of the Cauchy Integral for the calculation of analytical solutions of the definition in Equation 4.18. Visualizing the function in the $\mathbb{C}$ domain will likely make the solution of the integral more concrete and understandable. The intention here is to demonstrate how the singularity of the improper integral can be dealt with, through the use of Complex Variables concepts. Again, most of the material of this section draws from explanations in [27].

First, let be an integral in the complex z-plane defined as:

$$\oint_{\Gamma} \frac{f(z)}{z-a} dz, \tag{4.20}$$

which is denominated as Cauchy Integral. Being f(z) analytic and Γ a piecewise smooth closed contour in a open domain, such as the one in Figure 4.5, we have that the Cauchy Integral theorem can be applied as:

$$\oint_{\Gamma} \frac{f(z)}{z-a} dz = \begin{cases} 2\pi i f(a), & \text{if a is inside } \Gamma \\ 0, & \text{if a is outside } \Gamma \end{cases} \tag{4.21}$$

This is a well known result regarding Functions of one Complex Variable. For understanding what happens when a lies over the contour Γ, it helps to define a new contour such as the one in Figure 4.6. Then as the radius ε goes to zero, the value of the integral along the semi-circle Γ_{ε} tends to $\pi i f(a)$, i.e., half of the value of the residue from when the pole is fully contained in the closed contour.

$$\oint_{\Gamma} \frac{f(z)}{z-a} dz = 2\pi i f(a). \tag{4.22}$$

In practice, the integral over γ_R, in Figure 4.5 vanishes as $R \to 0$ when

$$|f(z)| < \frac{C}{|z|} \tag{4.23}$$

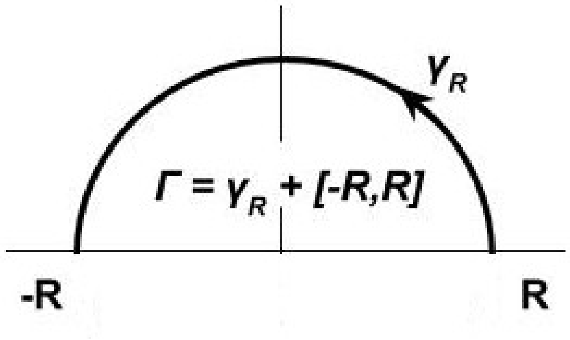


Figure 4.5: A diagram of the integral, in the complex domain, of a piecewise smooth curve.

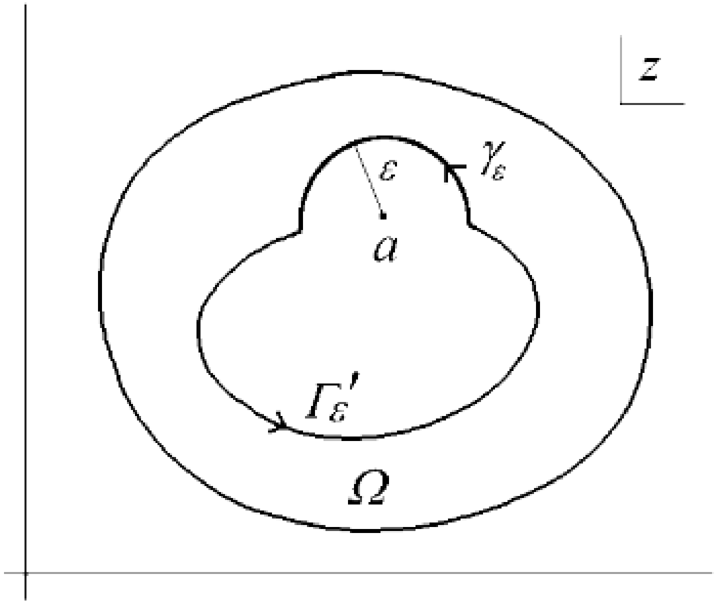


Figure 4.6: New closed contour, defined for when a is located over the old contour. Source: (Johansson, 2006).

for some constant $C > 0$. The same result applies if, for $m > 0$,:

$$|f(z)| < C|e^{imz}|. \tag{4.24}$$

Thus, in the case of singuarities in integrals, such as the one in the definition of Hilbert Transform, it is convenient to visualize the function in the $\mathbb{C}$ domain and, with the help of some results from the theory of Functions of One Complex Variable for the case of integrals over closed contours, obtain analytical solutions for these once unsolvable integrals.

4.4.3 Relation to the Fourier Transform and Cepstral Analysis

In this section, we intend to show the relation of the Hilbert Transform with the Fourier Transform of a signal, and how this relation will be useful for obtaining a causal and positive kernel from the phase indeterminacy relation of Equation 4.17.

The Fourier Transform of a signal f(t) is defined as:

$$F(\omega) = \int_{\mathbb{R}} f(t)e^{-i\omega t}dt, \tag{4.25}$$

with its inverse defined as:

$$f(t) = \frac{1}{2\pi}\int_{\mathbb{R}} F(\omega)e^{i\omega t}d\omega \tag{4.26}$$

In the case of real-valued signals, such as the estimated kernel, the right-sided axis, corresponding to the positive-valued frequencies, contains the whole information of the time-domain waveform.

Considering $F(\omega)$ as the Fourier Transform of a real valued function, it is possible to define a function $Z_f(\omega)$, which is equal to zero for all negative-valued frequencies, i.e.:

$$Z_f(\omega) = F(\omega) + sgn(\omega)F(\omega), \tag{4.27}$$

where:

$$sgn(\omega) = \begin{cases} 1, & \text{for } \omega > 0 \\ 0, & \text{for } \omega = 0 \\ -1, & \text{for } \omega < 0 \end{cases} \tag{4.28}$$

The amplitude of the resulting signal $Z_f(\omega)$ is shown in Figure 4.7. By writing the Inverse Transform of $Z_f(\omega)$ as:

$$z_f(t) = f(t) + ig(t), \tag{4.29}$$

we have that

$$g(t) \xrightarrow{\mathcal{F}} (-isgn(\omega))F(\omega) \tag{4.30}$$

From the fact that $-isgn(\omega)$ is the Inverse Fourier Transform of $1/(\pi t)$, we have that

$$g(t) = f(t) \star \frac{1}{\pi t} = P\int_{\mathbb{R}} \frac{f(\tau)}{t-\tau}d\tau = Hf(t) = \hat{f}(t), \tag{4.31}$$

i.e., g(t) is equivalent to the Hilbert Transform of f(t). Therefore, adding the term $iH(f(t)$ to a signal $f(t)$ will cause it to have a one-sided (positive) Fourier spectrum, what is the main idea behind the minimal phase filter extraction through Cepstral Analysis.

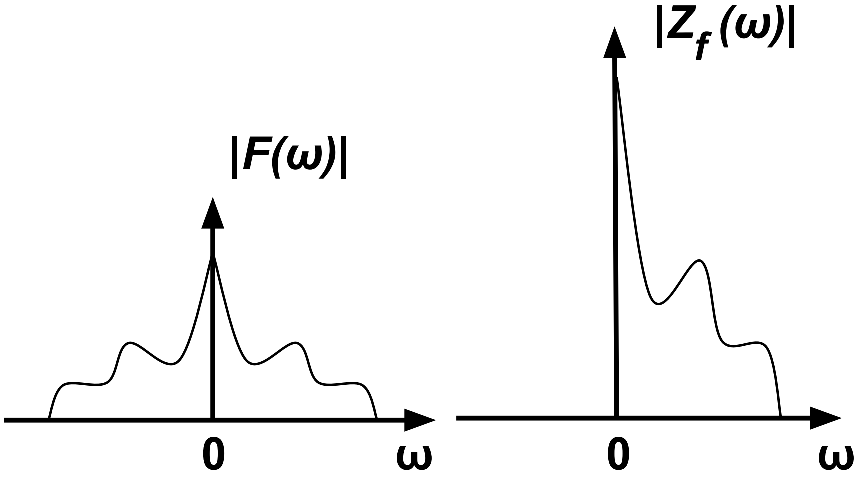


Figure 4.7: Amplitudes of original and resulting $Z_f(\omega)$ function.

4.5 Minimal Phase Solution of the Phase Indeterminacy Problem

The Phase Indeterminacy problem, such as the one in Equation 4.17, is generally a very difficult one. Therefore, it is usually necessary to constrain it to the more tractable case of a minimal phase solution of the system. The key idea for reconstructing a complete Fourier spectrum given only the amplitude is the sufficiency of real and imaginary parts of the Fourier and Hilbert Transforms, in the case of causal sequences [28].

The minimal phase condition of a may be concisely defined as that of being causal, positive and stable. In the continuous case, stability corresponds to having all the poles in the left-side of the complex plane. In the discrete case, it corresponds to not having poles outside the unit circle.

Given a Fourier Transform $F(\omega)$ of a signal, it is possible to summarize three equivalent expressions of the minimum phase condition as the following:

- $\log|F(\omega)|$ and $arg[F(\omega)]$ are Hilbert transforms of each other;
- $F(\omega)$ has no poles or zeros outside the left side of the complex plane;
- There exists a causal and stable inverse system with system function $F^{`1}(\omega)$ such that $F(\omega)F^{`1}(\omega) = 1$.

In Section 4.4.3, it was already shown that if a sequence is causal, then the real and imaginary parts of its Fourier Transform are related by the Hilbert Transform integral.

Based on the stationarity assumptions, it is possible to show that the only causal filter is the obtained minimal phase solution of the Equation 4.17. From the Theorem in [29], we have that, given a real filter with Fourier Transform amplitude $|\hat{f}_{i\omega}|$ satisfying:

$$\int_{\mathbb{R}} \frac{\log(|\hat{f}_{i\omega}|)}{1+\omega^2} d\omega < \infty, \tag{4.32}$$

then the filter defined by

$$g_{i\omega} = e^{-\log\left(|\hat{f}_{i\omega}|\right)+iH\left(\log\left(|\hat{f}_{i\omega}|\right)\right)} \tag{4.33}$$

is the only causal solution of Equation 4.17, therefore, the only one satisfying the assumptions on the self-exciting kernel made in Chapter 3, which is both minimal phase and has amplitude $|\hat{f}_{i\omega}|= |\hat{g}_{i\omega}|$, being simply the minimal phase realization of $|f_{i\omega}|$, .

From

$$(\mathbb{I} + \hat{\psi}_{i\omega}) = (\mathbb{I} - \hat{\phi}_{i\omega})^{-1}, \tag{4.34}$$

let us define

$$E_{i\omega} = |\mathbb{I} - \hat{\phi}_{i\omega}|^{-2} \tag{4.35}$$

It is possible to show that $\hat{g}_{i\omega} = (\mathbb{I} - \hat{\phi}_{i\omega})^{-1}$ is a minimal phase filter by demonstrating that, given the stationarity assumptions on the kernel $\hat{\phi}_{i\omega}$, all the poles and zeros of $\hat{g}_{i\omega}$ have negative real part:

1. If $i\omega$ is a zero of $(\mathbb{I} - \hat{\phi}_{i\omega})^{-1}$, then it is a pole of $\hat{\phi}_{i\omega}$. But we have, from the stationarity condition on $\hat{\phi}_{i\omega}$, that $|\hat{\phi}_{i\omega}| \leq \int_{\mathbb{R}} |e^{-i\omega t}|\phi_t dt$, and thus $\hat{\phi}_{i\omega}$ can only have poles with negative real part.

2. If $i\omega$ is a pole of $(\mathbb{I} - \hat{\phi}_{i\omega})^{-1}$, then $|\hat{\phi}_{i\omega}|$ is equal to 1, what contradicts the stationarity assumption. Therefore, $(1 - \hat{\phi}_{i\omega})^{-1}$ has no pole.

In conclusion, the only solution of the phase indeterminacy problem which satisfies the stationarity assumptions on the self-exciting kernel is the minimal phase solution.

A pseudocode with the main steps of the method is shown in Algorithm 2

Algorithm 2 Discretized kernel Estimation

```
Input: sequence $(t_1, t_2, ..., t_n)$ and resolution of output grid, h.
Calculate $\tau_{max}$ (horizon) and h (window size).
Compute discrete approximation of covariance, $\nu_\tau^{(h)}$.
Undo Kernel Density Estimation of Discretized Covariance in the Frequency Domain ← Divide empirical $\nu$ by $g_{i\omega}^{(h)}$
Extract minimal phase filter realization of $\psi_{i\omega}$
Get $\phi_{i\omega}$ from $\psi_{i\omega}$
Convert $\phi_{i\omega}$ to the time domain, for getting the Discretized Kernel Estimate in the time domain, $\phi(t)$
```

4.6 Summary: Discretized Kernel Estimation

Hereby, a summary of the method is presented. This step is fully described in [10], and basically consists of building an estimator of $\phi(t)$ from empirical measurements of $\nu(\tau)$, the stationary covariance.

Given a finite sequence of ordered time-events in [0,T], we fix a window size of h, and estimate $\nu(\tau)$ as:

$$\nu_\tau^{(h)} = \frac{1}{h} E\left((\int_0^h dN_s - \Lambda h)(\int_\tau^{\tau+h} dN_s - \Lambda h) \right) \tag{4.36}$$

In practice, this estimation is done at discrete time steps δ:

$$\nu^{(h)}_{\tau,\delta}{=}\frac{1}{T}\sum_{i=1}^{\lfloor T/\delta\rfloor}(dN^{(h)}_{i\delta}{-}dN^{(h)}_{(i-1)\delta})(dN^{(h)}_{i\delta+\tau}{-}dN^{(h)}_{(i-1)\delta+\tau}), \tag{4.37}$$

where $dN^{(h)}_{i\delta}$ is the total number of events happening between $t = i\delta$ and $t = i\delta + h$. The speed of this procedure can be greatly improved through parallelization.

From Theorem 1 in [10], we have that, given $g^{(h)}_t = (1 - \frac{|h|}{t})^+$, i.e., a triangular kernel density estimator with bandwidth "h", we have the following relation in Laplace domain:

$$\widehat{\nu_z^{(h)}} = \hat{g_z}^{(h)}(1 + \hat{\psi}^\star_z)\Lambda(1 + \hat{\psi}^\star_z)^\dagger, \tag{4.38}$$

where[1]:

$$\hat{\psi}_z = \sum_{n=1}^{+\infty}\hat{\phi}^n_z = \frac{\hat{\phi}_z}{(1 - \hat{\phi}_z)}. \tag{4.39}$$

Working with the Fourier transform restriction, i.e., ($z = i\omega$, with $\omega \in \mathbb{R}$) and given that:

$$\hat{g}^{(h)}_{i\omega} = \frac{4}{\omega^2 h}\sin^2(\frac{\omega h}{2}), \tag{4.40}$$

we get to:

$$(1 + \hat{\psi}^\star_z)\Lambda(1 + \hat{\psi}^\star_z)^\dagger = \frac{\widehat{\nu_z^{(h)}}}{\hat{g_z}^{(h)}}, \tag{4.41}$$

where we fix $h = \delta$, so we do not bother with cancellations of $\hat{g}^{(h)}_z$. And then, from:

$$|1 + \hat{\psi}_{i\omega}|^2{=}\frac{\hat{\nu}^{(h)}_z}{\Lambda\hat{g_z}^{(h)}}, \tag{4.42}$$

we get to the discretized estimation of ϕ_t by taking the inverse Fourier transform of:

$$\hat{\phi}_{i\omega} = 1 - e^{-\log|1+\hat{\psi}_{i\omega}|+iH(\log|1+\hat{\psi}_{i\omega}|)}, \tag{4.43}$$

in which the operator $H(\cdot)$ refers to the Hilbert transform.

[1] Given a function f_t, $\hat{f}_z$ is its Laplace Transform, and the "$\star$" symbol corresponds to its conjugate

Chapter V

Automatic Kernel Decomposition

5.1 Introduction

- Four base kernels and their typical behaviours
- Stationarity Condition derivation

5.1.1 Self-Exciting Base Kernels

From the definition of the conditional intensity function, the self-excitement of the process is expressed through the kernel function $\phi(t)$. For the kernel decomposition, four base kernels will be used for identifying and estimating typical triggering behaviors:

- **EXP(α,β)**: The decay exponential kernel is parameterized by the amplitude α and decay rate β, and is useful for modeling quick influence decay, such as in finance or web data [9], in which initial transactions/hyperlinks have a lot of impact initially but gradually reduce their influence over time:

$$\mathrm{EXP}(\alpha,\beta) = \alpha e^{-\beta t} \tag{5.1}$$

- **PWL(K,c,p)**: The **po**wer **l**aw kernel is parameterized by the amplitude K, the exponent p, and the constant c, such as in Equation (5.2), and has been prevalent in earthquake [8] and social media-related data [5], modeling a slower decaying trend than the exponential:

$$\mathrm{PWL}(K,c,p) = \frac{K}{(t+c)^p} \tag{5.2}$$

- **SQR(B,L)**: The pulse kernel is described by the amplitude B and the length L. Being a trivial, steady, self-exciting dynamics on its own, it may also work as an offset level for the combined triggering with other kernel types, in the case of addition, and as a horizon truncation, in the case of multiplication[1]:

$$\mathrm{SQR}(B,L) = B(u(t) - u(t-L)) \tag{5.3}$$

- **SNS(A,ω)**: A truncated **sinus**oidal, parameterized by the amplitude A and the angular velocity ω. This type of kernel base function captures well the self-excitement of periodic events, such as TV watching-related data (IPTV) [11], in which watching one episode of a TV program makes the viewer more likely to watch further ones. Since these shows are usually broadcasted weekly,

[1] u(t) is the step function

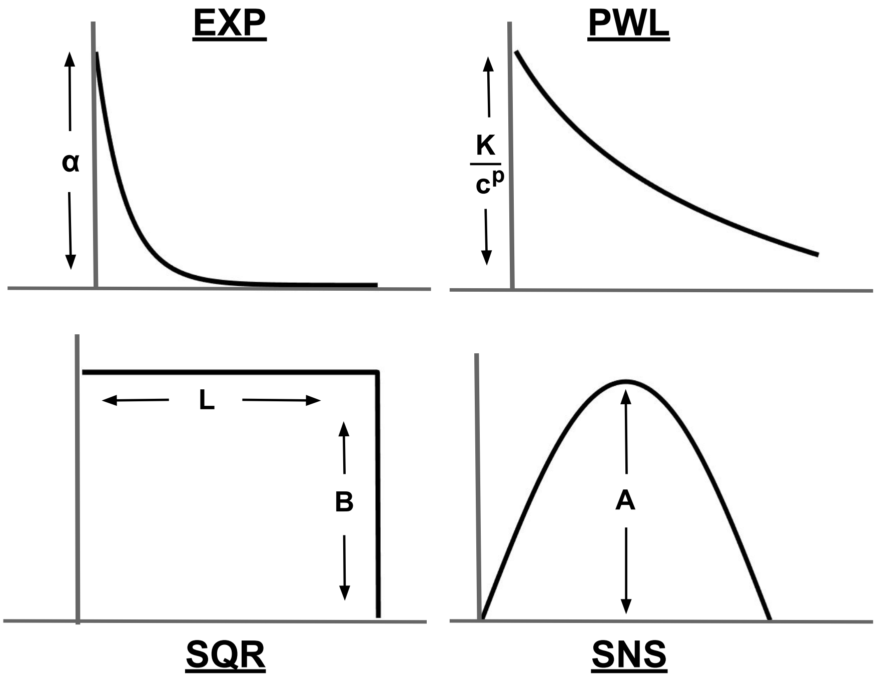


Figure 5.1: The four base kernel types.

the TV-watching behavior will likely demonstrate this weekly self-excitement. Also, according to [3], homicide rates show a pronounced seasonal effect, peaking in the summer and tapering in the winter:

$$\mathrm{SNS}(A,\omega) = Asin(\omega t), \tag{5.4}$$

for $t \in [0, \dfrac{\pi}{\omega}]$, and 0, otherwise.

Type	Equation
Exponential (EXP(α,β))	$\alpha e^{-\beta t}$
Power-Law (PWL(K,c,p))	$\frac{K}{(c+t)^p}$, $(p > 1)$
Pulse (SQR(B,L))	$B(u(t) - u(t-L))$
Sinusoidal (SNS(A))	$Asin(\omega t)$, $t \in \left[0, \frac{\pi}{\omega}\right]$

Table 5.1: Four base kernels

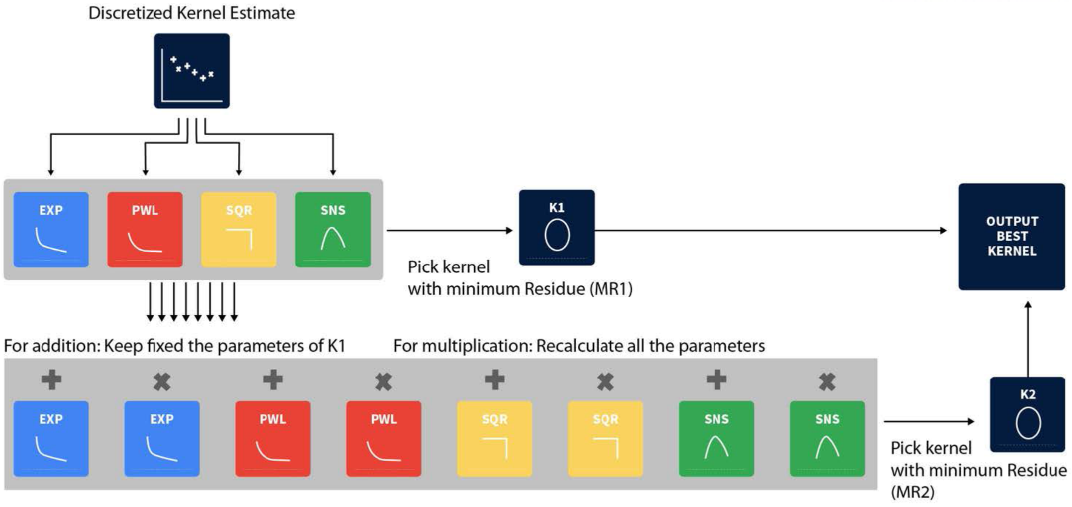

Discretized Kernel Estimate
EXP
PWL
SQR
SNS
K1
Pick kernel
with minimum Residue (MR1)
OUTPUT
BEST
KERNEL
For addition: Keep fixed the parameters of K1
For multiplication: Recalculate all the parameters
EXP
EXP
PWL
PWL
SQR
SQR
SNS
SNS
K2
Pick kernel
with minimum Residue
(MR2)

Figure 5.2: A diagram for the Kernel Decomposition Algorithm.

EXP(α,β)	$\frac{\alpha}{\beta}$	**PWL(K,c,p)**	$\frac{Kc^{1-p}}{p-1}, (p > 1)$	**SQR(B,L)**	BL	**SNS(A,ω)**	$\frac{2A}{\omega}$

Table 5.2: Stationarity Condition of the four Base Kernels.

5.2 Greedy Kernel Decomposition

For expressing the discretized estimation in terms of the four base kernels, the following steps are executed:

1. Calculate residues (L^1-error) w.r.t the four basic kernels {EXP, PWL, SQR, SNS};

2. Check whether the estimated parameters of the kernels satisfy the stationarity condition, by using the closed-form expressions from table 5.2;

3. Among those estimated kernels which satisfy the said condition, select the kernel with the minimum residue MR_1 (denominated K_1);

4. Calculate residues w.r.t. a total of 8 kernel expansions, resulting from 2 operations (addition and multiplication) per base kernel {+EXP, *EXP, +PWL, *PWL, +SQR, *SQR, +SNS, *SNS}, while fixing the optimized parameters for K_1, in the case of Additive Combination, and recalculating all the parameters, in the case of Multiplicative Combination;

5. Among those estimated combinations which satisfy the spectral radius condition (calculated in closed form from Table 5.3), select the kernel with minimum residue MR_2, denominated K_2;

6. If $MR_1 < MR_2/\eta$ (η would act as a regularization parameter), pick K_1 . Else, pick K_2.

A diagram of this algorithm is shown in Algorithm 1.

Algorithm 3 Automatic Decomposition of HP Kernels

```
$input = k_{est}$
$fit_1 = \text{fit}(k_{est}; \emptyset, \{\text{EXP, PWL, SQR, SNS}\})$
$K_1 = index\_of\_kernel(min\_residue(fit_1))$
$MR_1 = min\_residue(fit_1)$
$fit_2 = \text{fit}(k_{est}; K_1, \{\text{+EXP}, \text{* EXP}, \text{+PWL}, \text{*PWL},$
        $\text{+SQR}, \text{*SQR}, \text{+SNS}, \text{*SNS}\})$
$MR_2 = min\_residue(fit_2)$
$K_2 = index(min\_residue(fit_2))$
$output = None$
if $||\phi_{K_1}|| < 1$ then
    $output = K_1$
end if
if $||\phi_{K_2}|| < 1$ then
    if output $\neq$ None then
        if $MR_1 \geq \frac{1}{\eta} MR_2$ then
            $output = K_2$
        end if
    else
        $output = K_2$
    end if
end if
```

Base Kernel	Base Kernel	Condition
EXP(α,β)	EXP(α,β)	$\alpha_1\alpha_2/(\beta_1+\beta_2)$ (closed under multiplication)
EXP(α,β)	PWL(K,c,p)	$\alpha K\beta^{p-1}e^{\beta c}\Gamma(1-p,\beta c)$
EXP(α,β)	SQR(B,L)	$\left(\alpha B(1-e^{-\beta L})\right)/\beta$
EXP(α,β)	SNS(A,ω)	$\left(A\alpha\omega(1+e^{\frac{-\beta\pi}{\omega}})\right)/(\omega^2+\beta^2)$
PWL(K_1,c_1,p_1)	PWL(K_2,c_2,p_2)	$\leq (K_1K_2)/\left((p_1+p_2-1)\min(c_1,c_2)^{(p_1+p_2-1)}\right)$ (upper bound)
PWL(K,c,p)	SQR(B,L)	$\left(KB(c^{-(p-1)}-(c+L)^{-(p-1)})\right)/(p-1)$
PWL(K,c,p)	SNS(A,ω)	$\leq KA\left((c+\frac{\pi}{\omega})^{1-p}-c^{1-p}\right)/(1-p)$ (upper bound)
SQR(B,L)	SQR(B,L)	BL
SQR(B,L)	SNS(A,ω)	$2AB/\omega$
SNS(A,ω)	SNS(A,ω)	$\pi A/(2\omega)$

Table 5.3: Stationarity Condition for Multiplicative Combination of the four Base Kernels.

5.2.1 Stationarity Conditions

As a quick way of evaluating the validity of the estimated models regarding the stationarity criterion, we developed closed-form expressions, either in the form of equality or as an upper bound, listed in Table 5.2, for the case of a single kernel, and Table 5.3, for multiplicative combinations of two kernels [2]. The conditions for additive combinations can be derived from the conditions for single kernels in a straightforward manner.

The kernel is said to induce stationarity if the result of the expression calculated using the estimated parameters belongs to the interval [0,1). This can be justified both from the point-of-view of Hawkes Process as a branching process, also called *immigrant-birth representation* [20], and of the boundedness of the spectral radius (largest absolute value among the eigenvalues) of the excitation matrix. [3]

5.3 Analysis of Higher-Order Kernel Decomposition

A sequential additive decomposition of the discretized estimation vector is rather straightforward, since one may just set the residual vector from the previous stages as the input of the next ones.

In the case of multiplicative decomposition, it is nontrivial to find the result of intraclass decomposition. To the best of our knowledge, no analysis on multiplicative HP kernel decomposition is reported yet.

Here, we provide a new upper bound over an interclass kernel product of unknown degree, as in:

$$[\text{EXP}]^{k_1} \times [\text{PWL}]^{k_2} \times [\text{SQR}]^{k_3} \times [\text{SNS}]^{k_4} \text{ (for } k_i \in \mathbb{Z}^*\text{)},$$

where the operator “$[\cdot]^k$” corresponds to the set of functions which can be decomposed into a k-th order product of kernels, e.g:

[2] $\Gamma(\cdot,\cdot)$ is the well-known Incomplete Gamma Function: $\Gamma(a,y) = \int_y^\infty t^{a-1}e^{-t}dt$

[3] For the Univariate HP case, the excitation matrix has dimension one, being only the excitation function, $\phi(t)$.

$$[EXP]^k = \underbrace{\alpha_1 e^{-\beta_1 x} * \alpha_2 e^{-\beta_2 x} * \ldots * \alpha_k e^{-\beta_k x}}_{\text{k terms}}.$$

By deriving the four possible intraclass kernel products, one may observe that the typical self-exciting behavior features of each kernel type are preserved, as in the following.

5.3.1 EXP

$[\mathrm{EXP}]^{k_1}$ reduces to the case of a single exponential with $\alpha = \prod_{i=1}^{k_1} \alpha_i$ and $\beta = \sum_{i=1}^{k_1} \beta_i$, thus still accounting for its 'quick-decay' behavior: $[\mathrm{EXP}]^{k_1} \subset [\mathrm{EXP}]$;

5.3.2 PWL

$[\mathrm{PWL}]^{k_2}$ is lower bounded by a single PWL kernel with $K = \prod_{i=1}^{k_2} K_i$, $c = max(c_1, ..., c_{k_2})$ and $p = \sum_{i=1}^{k_2} p_i$, thus still accounting for its 'slow-decay' behavior;

5.3.3 SQR

$[\mathrm{SQR}]^{k_3}$ reduces to a single SQR kernel with $B = \prod_{i=1}^{k_4} B_i$ and $L = min(L_1, ..., L_{k_4})$, thus still accounting for its 'steady-triggering' behavior: $[\mathrm{SQR}]^{k_3} \subset [\mathrm{SQR}]$;

5.3.4 SNS

$[\mathrm{SNS}]^{k_4}$ has $A = \prod_{i=1}^{k_4} A_i$ and a 'spikier' aspect (higher bandwidth), thus still accounting for its 'periodicity-inducing' behavior.

Thus, on deepening the decomposition algorithm by overly increasing the number of levels above 2, we may be, in fact, adding little information on the qualitative aspect of the self-exciting behavior analysis of the data while making it more prone to overfitting on the noisiness of the discretized estimation vectors.

5.3.5 Upper Bound

Furthermore, regarding the boundedness of the higher-order decompositions, from the exact results for EXP and SQR intraclass decompositions and the upper bounds for the PWL and SNS ones, we have that:

$$\begin{aligned}
&[\mathrm{EXP}]^{k_1} \times [\mathrm{PWL}]^{k_2} \times [\mathrm{SQR}]^{k_3} \times [\mathrm{SNS}]^{k_4} \\
\leq\ & \alpha e^{-\beta x} \frac{K}{(x + c_{upper})^p} BAsin(\omega x) \\
\leq\ & \frac{\alpha BKAe^{-\beta x}}{(x + c_{upper})^p} \\
=\ & \mathrm{EXP}(\alpha, \beta) \times \mathrm{PWL}(K, c_{upper}, p)^{k_2} \times \mathrm{SQR}(B, L) \times A,
\end{aligned}$$

for $0 \leq x \leq min(L, \dfrac{\pi}{\omega})$, and 0 otherwise.

Chapter VI

Report

6.1 Introduction

Regarding automatic analysis of time series, it is of relevance to make the extracted parameters readily available. One possible way of doing this is through an automatically generated report. The current version of this report is built with HTML and Bootstrap, and read through a web browser. There, the quantitative and qualitative aspects of the analysis are accompanied by visualization plots for the first and second levels of decomposition.

6.2 Automated Report Generation for Hawkes Processes

After estimating the likelihood-maximizing base kernel combination, the report is generated with a description of the corresponding parameters and its most relevant aspects, such as:

- **Decay-Rate**, which would be associated with decaying kernels, EXP and PWL. A exponential decay rate is described as ‘a quick decaying triggering influence’, and the power-law decay may be described as ‘a slowly decaying triggering influence’, with its respective parameters being put in evidence;

- **Steady-Rate Influence**, equivalent to an offset level from the SQR kernel type. It is described as ‘a steadily triggering influence’, with the parameter B put into evidence;

- **Triggering Horizon**, associated with discontinuous kernels, SNS and SQR. It is described as ‘a horizon of triggering influence’, with the parameter L, in the SQR case, or the value of $(\frac{\pi}{\omega})$, in the SNS case, put into evidence;

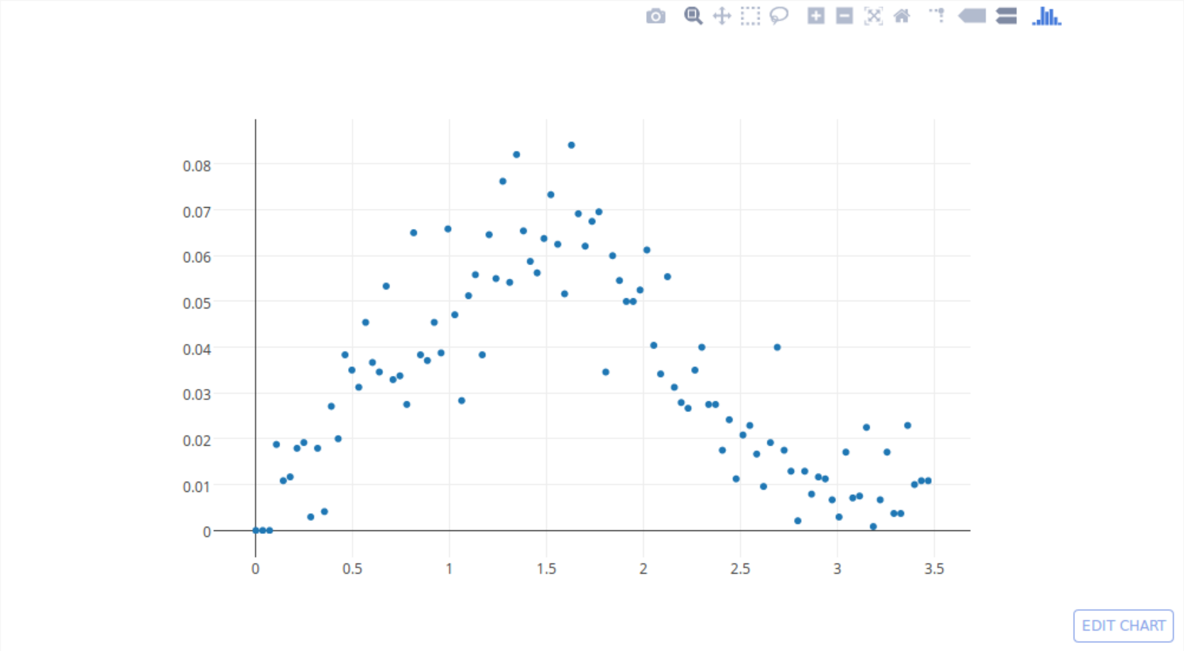
- **Induced Periodicity**, in the case of the SNS kernel type. It is described as ‘a periodicity-inducing triggering influence’, with the value of $(\frac{2\pi}{\omega})$, from the SNS case, put into evidence.

6.3 Examples

In this section, we present some examples of the report:

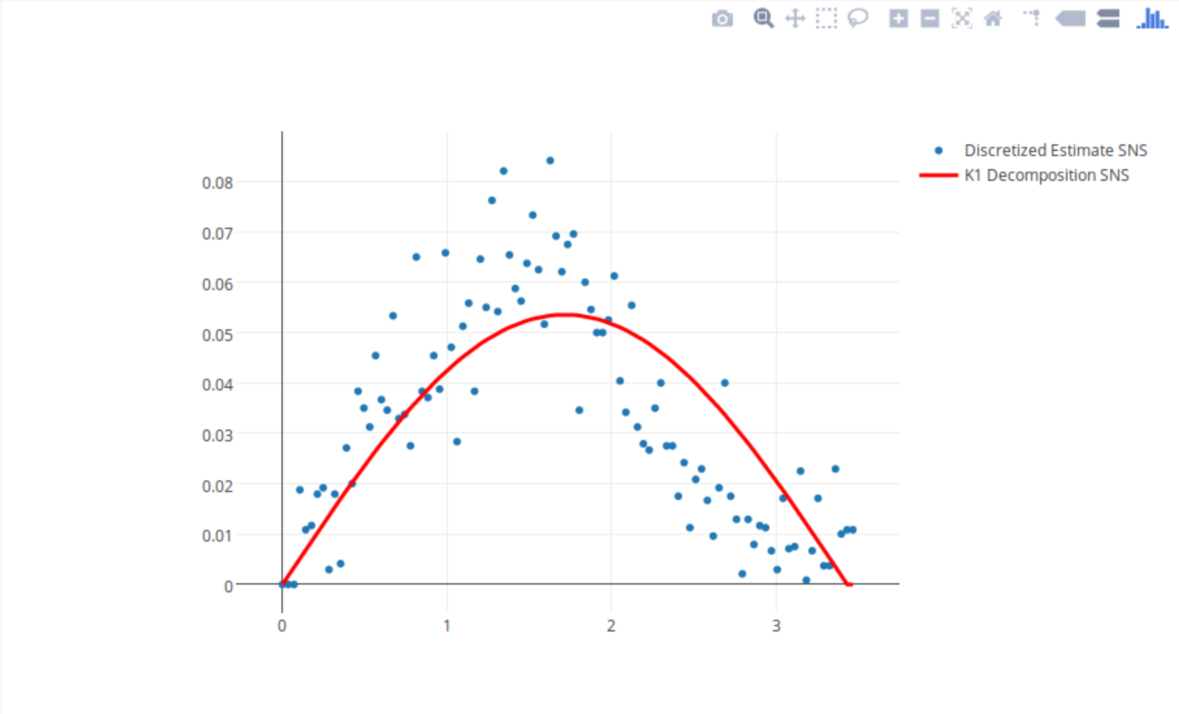
Discretized Kernel Estimation

This part corresponds to the results from the Discretized Kernel Estimation step:



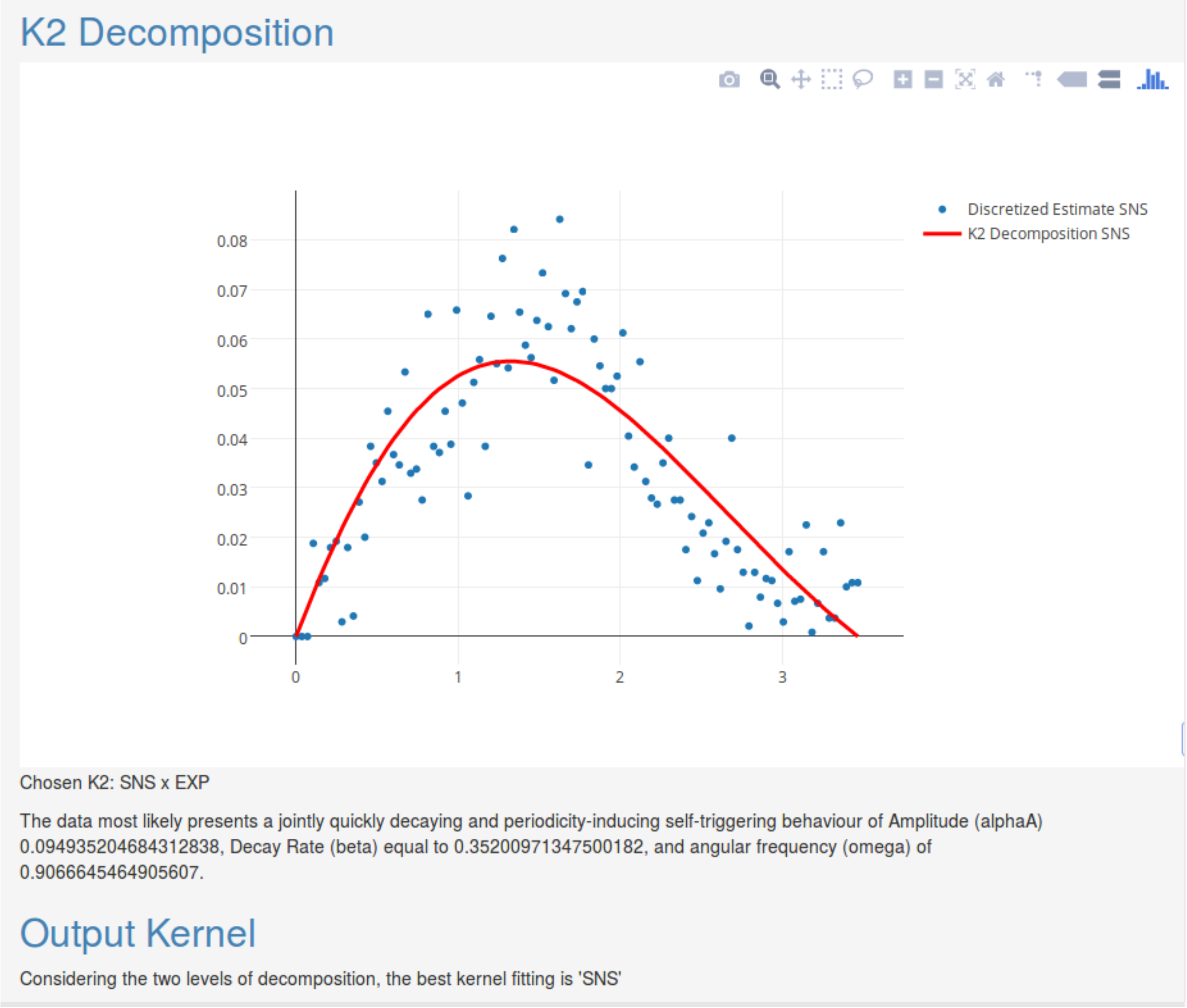
K1 Decomposition

Results from the first level of kernel decomposition (EXP,PWL,SQR,SNS):



Chosen K1: Sinusoidal

The data most likely presents some Periodicity-Inducing self-triggering behaviour with Amplitude (A) 0.053554413238203313 and Angular Frequency (omega) equal to 0.9159162255363827



K2 Decomposition

Chosen K2: SNS x EXP

The data most likely presents a jointly quickly decaying and periodicity-inducing self-triggering behaviour of Amplitude (alphaA) 0.094935204684312838, Decay Rate (beta) equal to 0.35200971347500182, and angular frequency (omega) of 0.9066645464905607.

Output Kernel

Considering the two levels of decomposition, the best kernel fitting is 'SNS'

Chapter VII

Experiments

7.1 Introduction

For testing the validity of the kernel decomposition framework, we conducted experiments with:

- Synthetic Data;
- Financial Data
- Earthquake Data

Synthetic data sequences were used for evaluating the effectiveness of a scale-independence strategy for setting the maximum value of τ in Equation 4.37. In the financial data experiments, quantitative comparisons were made among the first and second levels of the decomposition framework over a set of time sequences, and also among our decomposition framework and an estimation method largely used in financial applications, described in [12]. In the earthquake data experiments, we evaluate the overall triggering behavior of earthquakes through the relative frequency of each estimated kernel type, out of the four base kernels.

7.2 Experimental Results

For real-world data sets, no prior information about the kernel (type and parameters) is available. Thus, we use the log-likelihood of the kernel function over the time sequence as a quality criterion.

Given a realization $(t_1, t_2, ..., t_k)$ of some regular point process on [0,T], its log-likelihood (l) is expressed as:

$$l = \sum_{i=1}^{k} \log(\lambda(t_i)) - \int_0^T \lambda(u)du. \tag{7.1}$$

7.2.1 Synthetic Data

As a means of evaluating the overall precision and efficiency of the kernel decomposition framework, experiments with synthetic data from the corresponding four basic kernel types were performed.

Simulation algorithms related to Hawkes Processes, from which synthetic time sequences data may be obtained, are divided into two main categories: cluster-based [30] and intensity-based [31]. For the proposed experiments, the “Thinning” algorithm, an intensity-based one, was used, mainly because the parametric representations of the kernel types make it very convenient to accurately calculate the value of the intensity function throughout the whole simulation horizon.

		Predicted			
		EXP	PWL	SQR	SNS
Actual	EXP	7	3	0	0
	PWL	4	6	0	0
	SQR	0	3	7	0
	SNS	0	0	0	10

		Predicted			
		EXP	PWL	SQR	SNS
Actual	EXP	8	2	0	0
	PWL	2	8	0	0
	SQR	0	1	9	0
	SNS	0	0	0	10

Table 7.1: Confusion matrices among the four basic kernel types for original horizon length (left) and histogram-based horizon length (right).

The referred algorithm, whose full denomination is “Ogata’s Modified Thinning Algorithm”, basically consists of simulating time sequences using a exceedingly high constant value for the intensity function and then thinning out the generated events using rejection sampling with regards to the locally calculated actual values of the intensity.

For an automatic decomposition framework willing to perform effective analysis of several domains of data, scale-independence is indispensable, as time sequences of disjoint datasets may occur in time scales differing by several orders of magnitude. As an example, earthquake events’ occurrences in a sequence are spaced by intervals of monthly and yearly scales. Thus, setting a horizon of a few months as the maximum value of τ in Equation (4.37) might result in a satisfactory discrete estimation grid, but using the same time length for estimating the triggering behavior of a finance-related sequence would require an impractically large grid resolution for making itself effective.

A histogram of all the time intervals between events in a sequence may be readily generated, and is an indicator of the overall magnitude of the spacing among the events. Thus, as a rule of thumb, the horizon length for τ may be set as the smaller time interval strictly larger than 95% of the sequence’s intervals. In practice, this value of horizon length is obtained with the help of a histogram composed by 100 bins.

First, we generated 10 sequences with the ’Thinning algorithm’ [31], in the time range of 0 to 100000, for each of the four basic kernel types, with predefined kernel parameters and influence horizon of 6.6 [1]. Then, the decomposition framework was used for building confusion matrices over the sets of sequences for both the original horizon length (i.e., 6.6) and the histogram-based horizon length. The confusion matrices are shown in Table 7.1.

7.2.2 Financial Data

For the experimental setting, we picked companies in the NASDAQ and NYSE lists of Yahoo Finance for five categories: Technology, Healthcare, Services, Industrial Goods and Utilities. We extracted tick data from every two minutes of 30 business days (04/07/2017 to 05/18/2017 for Technology and 04/17/2017 to 05/25/2017 for the others). Whenever a stock price changed by some magnitude higher than some

[1] By influence horizon, we mean that the excitation effect of each kernel is considered null for $t > 6.6$.

CATEGORY	STOCKS
Technology	’CSCO’,’GOOGL’,’HPQ’,’INTC’,’IBM’,’MSFT’, ’ORCL’,’TXN’,’XRX’,’AAPL’,’ANGI’,’LITH’,’IAC’, ’FB’,’LOCMQ’,’YELP’,’GRPN’,’AMZN’,’GCI’
Healthcare	’BCRX’,’CUTR’,’AMRI’,’INSY’,’PDCO’,’CERS’, ’ACAD’,’FLDM’,’ACRX’,’ELGX’,’ALXN’,’PFE’,’JNJ’, ’MCK’,’MRK’,’NVS’,’UNH’,’GSK’,’AZN’,’SNY’
Industrial Goods	’GE’,’MMM’,’BA’,’HON’,’UTX’,’LMT’,’CAT’, ’GD’,’DHR’,’ABB’,’ITW’,’RTN’,’NOC’,’DE’, ’EMR’,’ETN’,’WM’,’CRH’,’CMI’,’WY’,’ROP’
Services	’BABA’,’CMCSA’,’HD’,’DIS’,’MCD’,’CHTR’,’PCLN’, ’UPS’,’SBUX’,’UNP’,’WBA’,’TWX’,’COST’,’NFLX’, ’LOW’,’CNI’,’FDX’,’FOXA’,’CSX’,’TJX’,’LVS’
Utilities	’NEE’,’DUK’,’D’,’SO’,’NGG’,’AEP’,’PCG’, ’EXC’,’SRE’,’PPL’,’EIX’,’ED’,’KEP’,’XEL’,’PEG’, ’WEC’,’ES’,’DTE’,’BIP’,’HNP’

Table 7.2: List of stocks selected for each category.

threshold, an event was logged in the corresponding time sequence. Ten different percentual thresholds, increasing at equally spaced intervals from 0.03% to 0.3%, were applied. This procedure resulted in sets of valid sequences specified in Table 7.3, since the remaining ones did not contain enough points for the splitting between training and validation subsequences.

The 80% of the first elements for each sequence were then used as training data, and the remaining 20% were used for validation, i.e., we estimated the parameters of the kernel using the first 24 days and then calculated the log-likelihood on the last 6 days of each sequence. We compared the log-likelihoods of first and second level decompositions, we observed that the second level, with composite kernels, resulted in a higher log-likelihood in a significant number of sequences for all the categories, as shown in Table , what corroborates the idea of a more flexible model of the kernel providing a more accurate

Category	Valid Sequences
Technology	70
Healthcare	117
Industrial Goods	53
Services	61
Utilities	48

Table 7.3: Number of resulting valid sequences for each category.

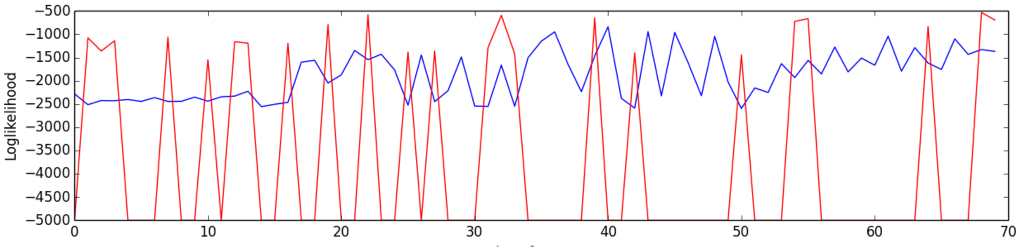


Figure 7.1: Technology: Comparison among loglikelihood of Decomposition Algorithm model (blue) and the usual exponential Hawkes model (red), fitted through Gradient descent.

Dataset	$l(K1,K2) > l(Exp.Hawkes)$	$l(K2) > l(K1)$
Technology	67.14%	98.53%
Healthcare	62.39%	92.31%
Industrial	64.15%	94.34%
Services	54.09%	85.25%
Utilities	77.08%	93.75%

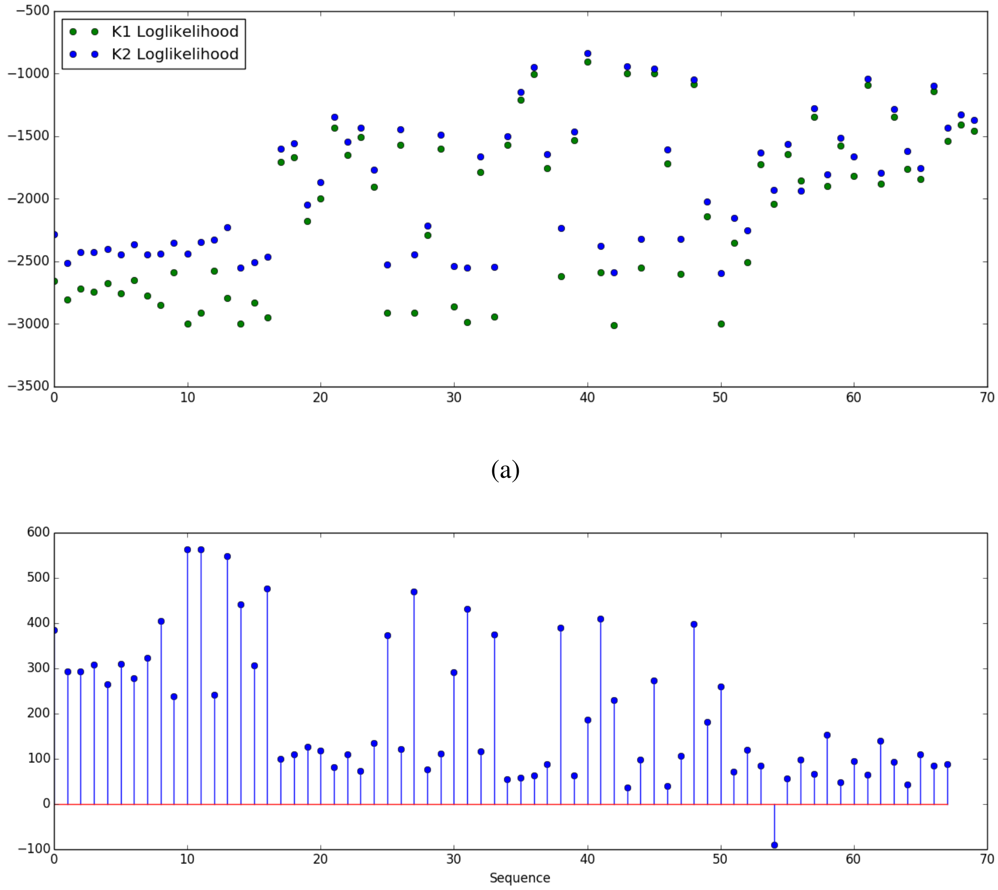
Table 7.4: Aggregate comparison, among the Gradient Descent based HP model and the first- and second-level decompositions of the proposed algorithm, for each of the five datasets.

description of the underlying dynamics of the process. The percentual outperformance of second level decomposition over the first level one, for each category, is shown in Table 7.4. The differences among first and second levels for each sequence are shown in Figures 7.2, 7.4, 7.6, 7.8 and 7.10.

When comparing the performance of the best estimation among the two levels and the usual Exponential HP model used in financial analysis, fitted through the gradient-based method from [12], it is possible to see that the Decomposition Algorithm exhibited a much more robust performance. Although the Exponential HP performed well in some sequences, it tended to get stuck in local maxima with very poor performance, usually leading to unstable or negative combinations of parameters. The percentage of outperformance of the Decomposition Algorithm for each category is shown in Table 7.4. The comparisons for each sequence are shown in Figures 7.1, 7.3, 7.5, 7.7 and 7.9.

7.2.3 Earthquake Data

In lists regularly published by seismological services of most countries habituated to the occurrences of earthquakes, one can obtain relevant information, such as the epicenter of each shock, focal depth, instrumental magnitude and origin time of each earthquake occurrence. Statistical analyses of earthquake catalogues and assessment of earthquake risk in a geophysical area can then be performed through the use of parametric models on the sequences of origin times, obtained by largely ignoring the remaining information. On further analyzing these fitted models, one can then identify and decompose components such as evolutionary trend, periodicity and clustering [8]. Again, the need to insert scale-independency, in the form of the aforementioned histogram-related heuristics, in the decomposition framework makes



(a)

(b)

Figure 7.2: Technology: (a) Comparison among loglikelihood of first- (green) and second-level (blue) of Decomposition Algorithm. (b) Difference among loglikelihood of first- and second-level for each sequence.

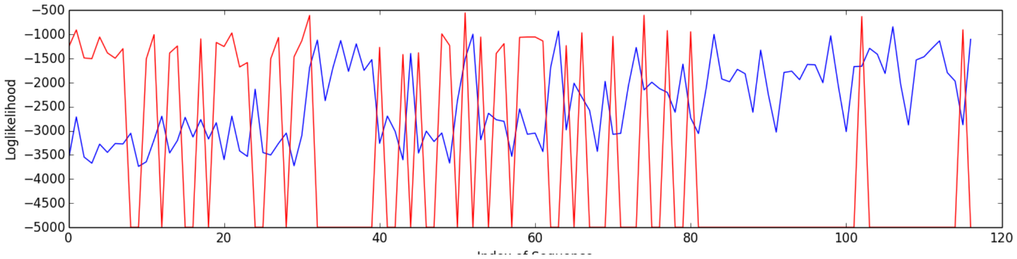
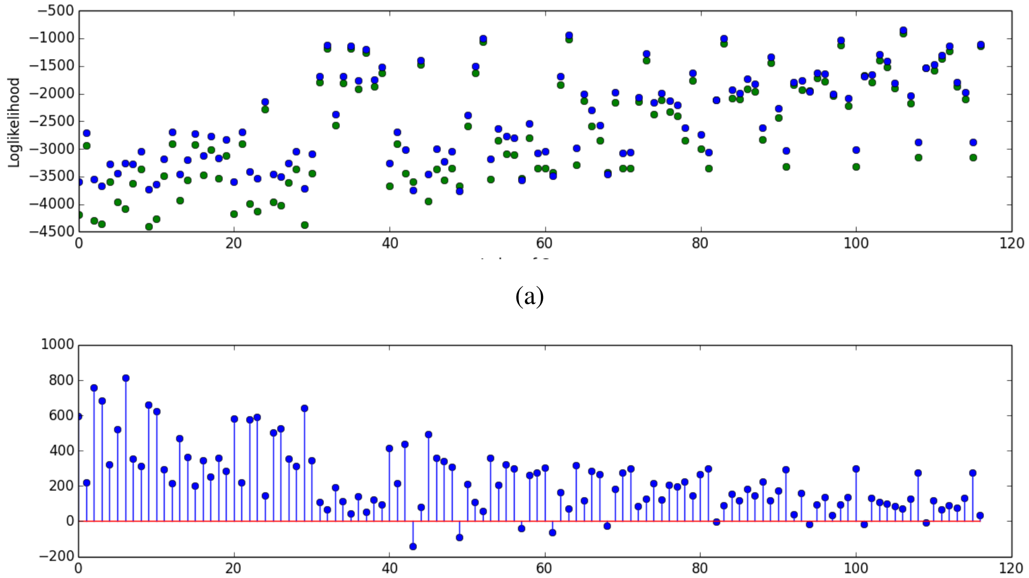


Figure 7.3: Healthcare: Comparison among loglikelihood of Decomposition Algorithm model (blue) and the usual exponential Hawkes model (red), fitted through Gradient descent.



(a)

(b)

Figure 7.4: Healthcare: (a) Comparison among loglikelihood of first- (green) and second-level (blue) of Decomposition Algorithm. (b) Difference among loglikelihood of first- and second-level for each sequence.

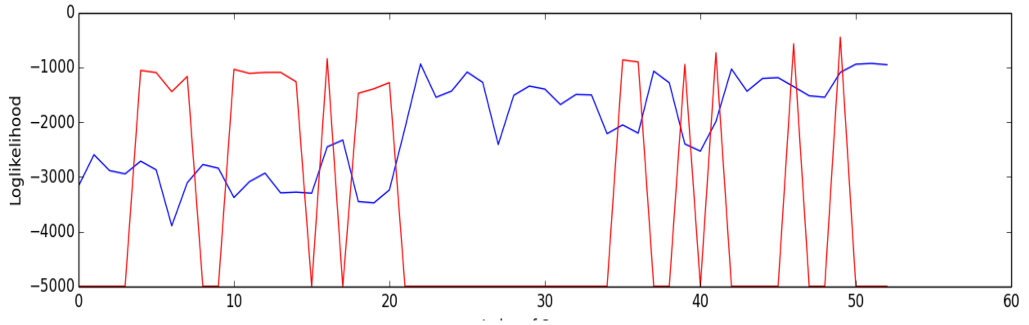
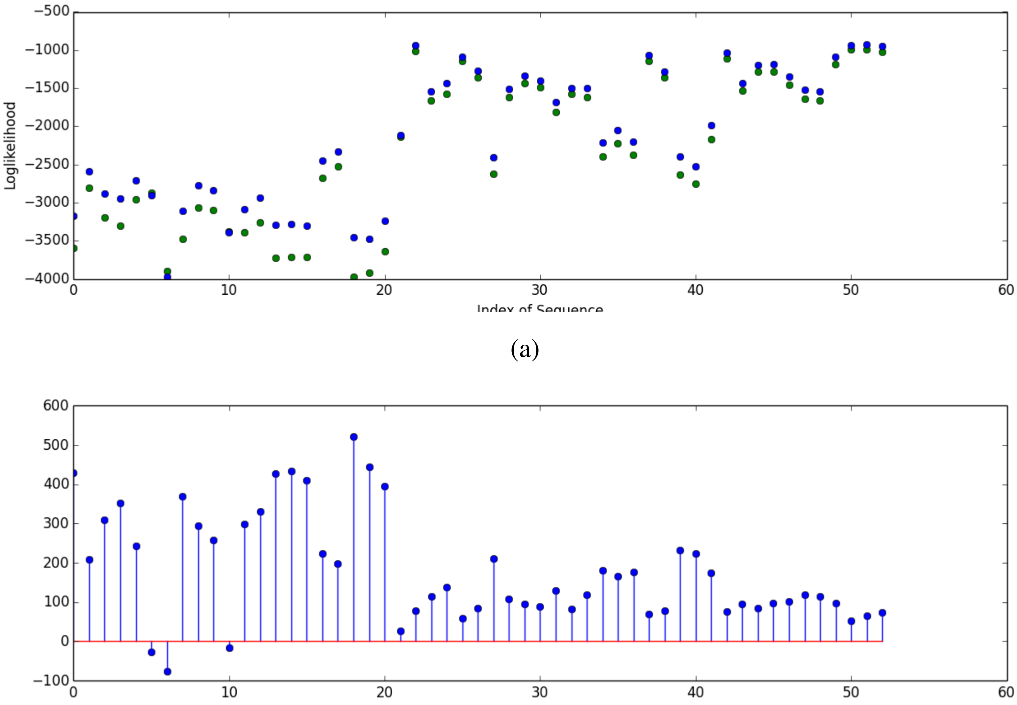


Figure 7.5: Industrial Goods: Comparison among loglikelihood of Decomposition Algorithm model (blue) and the usual exponential Hawkes model (red), fitted through Gradient descent.



(a)

(b)

Figure 7.6: Industrial Goods: (a) Comparison among loglikelihood of first- (green) and second-level (blue) of Decomposition Algorithm. (b) Difference among loglikelihood of first- and second-level for each sequence.

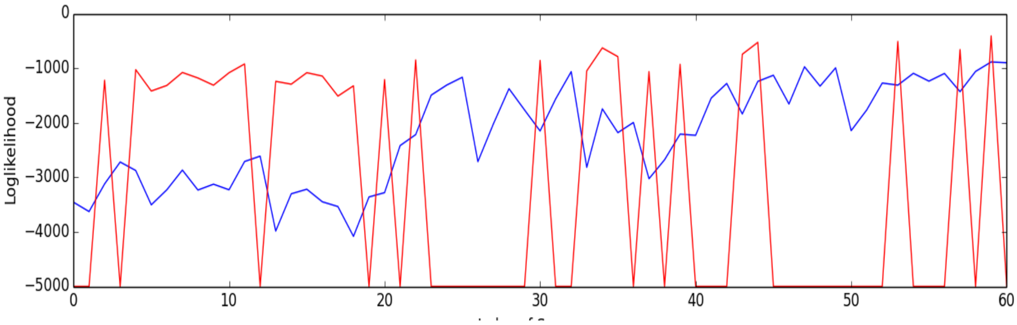
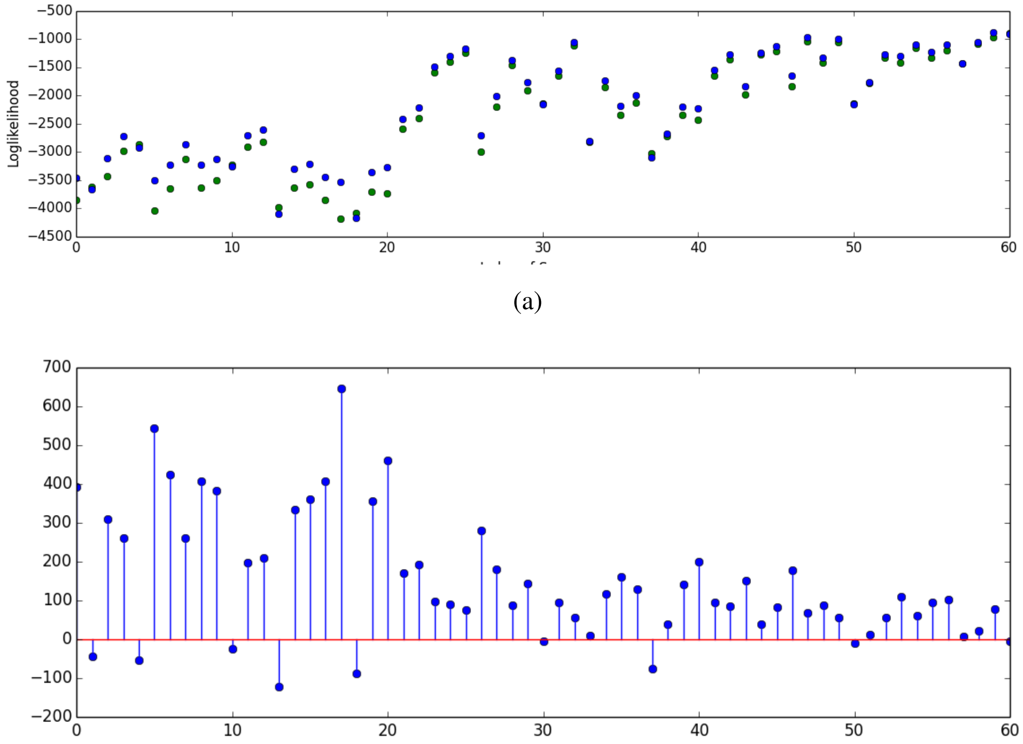


Figure 7.7: Services: Comparison among loglikelihood of Decomposition Algorithm model (blue) and the usual exponential Hawkes model (red), fitted through Gradient descent.



(a)

(b)

Figure 7.8: Services: (a) Comparison among loglikelihood of first- (green) and second-level (blue) of Decomposition Algorithm. (b) Difference among loglikelihood of first- and second-level for each sequence.

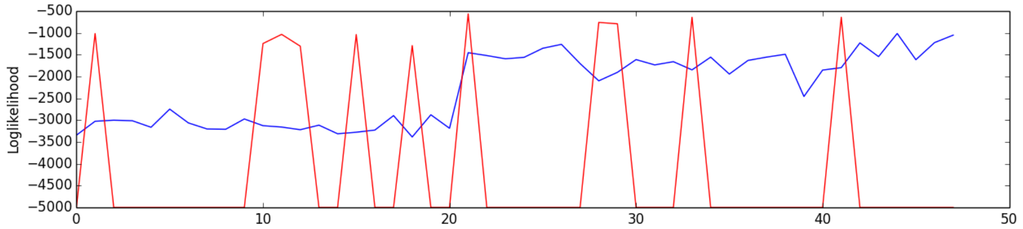
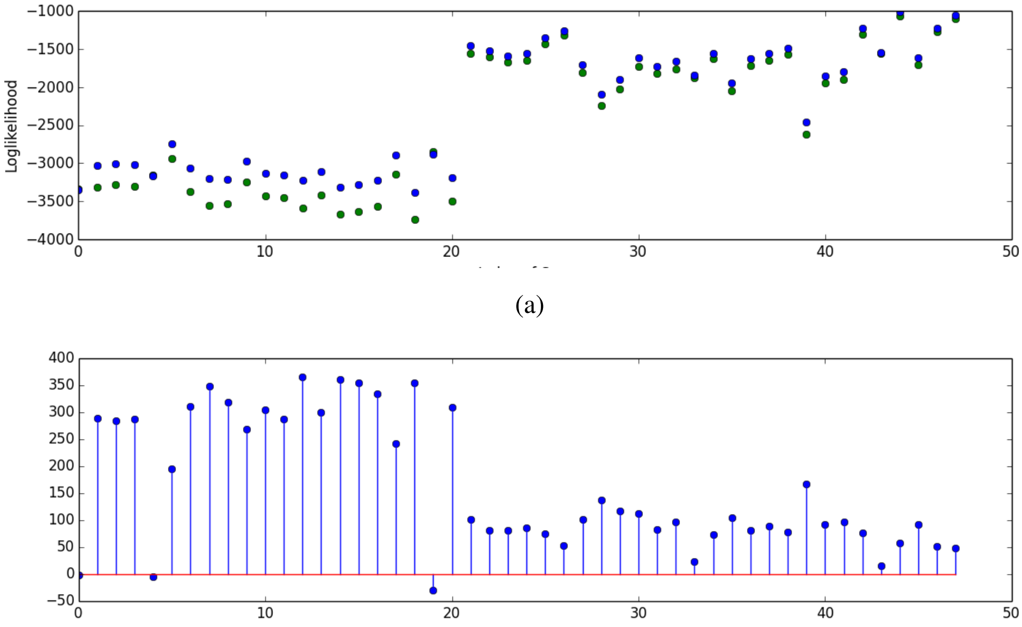


Figure 7.9: Utilities: Comparison among loglikelihood of Decomposition Algorithm model (blue) and the usual exponential Hawkes model (red), fitted through Gradient descent.



(a)

(b)

Figure 7.10: Utilities: (a) Comparison among loglikelihood of first- (green) and second-level (blue) of Decomposition Algorithm. (b) Difference among loglikelihood of first- and second-level for each sequence.

EXP	PWL	SQR	SNS
0	99	1	0

EXP	PWL	SQR	SNS
0	97	2	1

Table 7.5: Frequency of estimated kernel type of first level decomposition for (a) 100-point and (b) 20-point grid resolution.

itself imperative, as earthquakes events are separated by time intervals of monthly or yearly scales. Thus, the estimation horizon for financial data, for example, lasting usually only a few seconds, would hardly capture the overall aspect of the triggering behavior in this case.

The data considered for the experiment was a set of 100 time sequences extracted from the USGS NCSN Catalog (NCEDC database), from the day of 1966/Jan/01 to 2015/Jan/01. The Latitude range was [30,55], and the Longitude range was [-140,-110]. Different length intervals and resulting areas were considered. Whenever the magnitude of an event exceeded some threshold, its time coordinate was added to the corresponding input time sequence. The magnitude thresholds were varied among 2.5, 3.0, 3.5 and 4.0; and the grid resolution was set to 20 and 100 points. For 20-point grid resolution, the relative frequency of each kernel was (EXP,PWL,SQR,SNS) = (0,97,2,1). For the 100-point grid resolution, the relative frequency was (0,99,1,0). The results of first and second level decompositions are shown in Tables 7.5 and 7.6.

The results indicate a strong agreement with the long standing assumption of a power-law shaped kernel for the intensity of aftershocks' ocurrences ('Omori's Law' (1894)). Q-Q plots from the estimated

PWL + EXP	**PWL × SQR**	**SQR + EXP**
58	19	1
PWL × EXP	**PWL + SNS**	
14	8	

PWL + EXP	**PWL × SQR**	**SQR + SQR**
91	4	2
PWL × EXP	**PWL × PWL**	**SNS + EXP**
1	1	1

Table 7.6: Frequency of estimated kernel type of second level decomposition for (a) 100-point and (b) 20-point grid resolution.

models are shown in Figure 7.11.

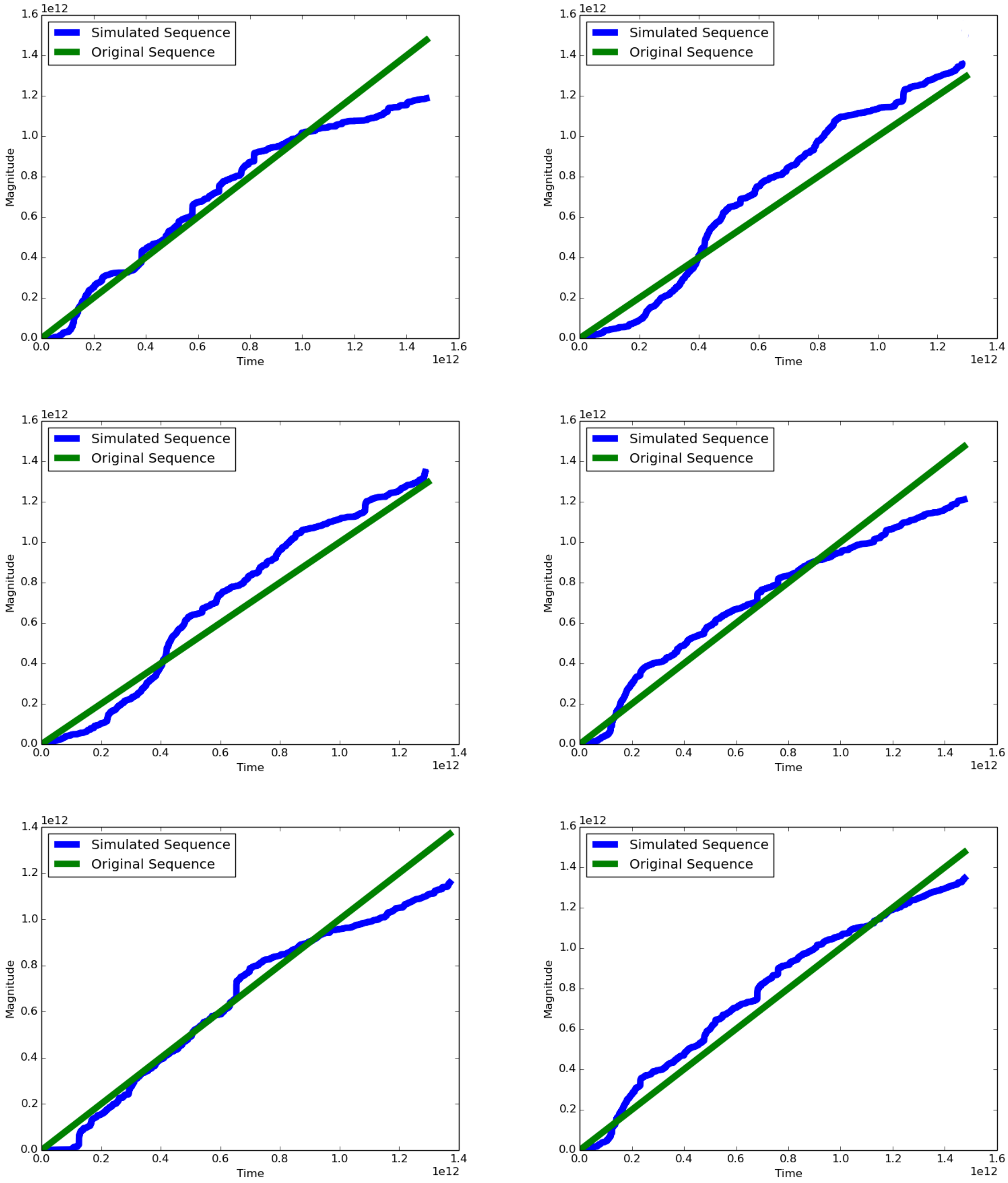


Figure 7.11: Q-Q Plots from Kernel Decomposition estimation over the Earthquake Dataset sequences.

Chapter VIII

Conclusion

Hawkes processes are temporal point processes which capture self-exciting discrete events in time series data. To predict future events with HPs, an appropriate kernel is selected by hands, previously. In this work, we proposed a new temporal covariance-based kernel decomposition method to represent various self-exciting behaviors. We also present a model (structure/parameter) learning algorithm to select the best HP kernel given the temporal discrete events. The stationarity conditions are derived to guarantee the validity of the kernel learning algorithm. In experiments, we demonstrate that the proposed algorithms perform better than existing methods to predict future events by automatically selecting kernels.

8.1 Summary of Thesis Achievements

The contributions of the present work may be summarized as:

- Proposing a framework for automatic estimation, decomposition and analysis for Univariate Hawkes Processes;
- Introducing a Histogram-based heuristics for Horizon-setting in Nonparametric Estimation Methods, along with its advantages;
- Deriving analytical expressions and upper bounds for a variety of combinations of parametric kernel functions;
- Evidencing the advantage of a richer parametric structure for description of HP kernels on real-world data.

8.2 Future Work

Two directions for future work would be on expanding the multi-class kernel framework for more general types of HPs: the multivariate and the spatiotemporal cases.

In the multivariate case, events in one sequence may affect the CIF of another sequence. Thus, not only the overall aspect of the excitation kernel would be of importance, but also the structural properties of the network in which all the concurrent processes are situated. This type of analysis could be applied to the study of Social Networks, in which the attitudes of certain nodes, the celebrities, have a strong influence on the remaining of the users. By identifying and analyzing these hugely influential nodes, companies may optimize the effectiveness of their online marketing strategies.

Regarding the spatiotemporal case, which is useful for the statistical study of criminal occurrences, for example, it would be of interest to analyze possible base kernels for capturing the spatial relations among events. So far, this relation is assumed to depend solely on the magnitude of the distance between the occurrences, but some nonrotationally invariant types of dependency may satisfactorily help modeling some more intricate aspects of the excitation effect.

8.3 Related Work

A spectral analysis approach to a one-dimensional self-exciting point process was introduced in [2]. Recently, the spectral method was extended to non-parametric kernel estimation for symmetrically networked HPs [10].

A Likelihood Maximization method was used on a class of parametric kernels for the excitation matrices in [12], while a series of works for likelihood maximization methods for power-law shaped kernels on seismology and earthquake data modeling were compiled in [8] . However, in [7], it is argued that, being designed for univariate Hawkes Processes, these methods can hardly be used to handle large amounts of data where the kernel function is not well localized compared to the exogenous inter-events time. Subsequently, an extensive analysis of spectral methods for non-parametric kernel estimation on networked HPs is performed.

In [3], Networked HPs are explored for analysis and modeling of stock-trading and crime data. A variant of the spectral method for networked HPs, with excitation matrix composed by linear combinations of decaying-exponentials, there referred to as ‘excitation modes’, is developed in [9]. An iterative log-likelihood maximization for non-parametric kernel estimation is presented in [32].

In [11], a maximum-likelihood estimator with a Sparse-Group-Lasso regularizer is introduced. In [20], an extensive mathematical treatment of the so-called ‘genuine multivariate’ Hawkes Processes, along with some considerations on the case of large-scale and networked data analysis, is provided. In [15], a nonparametric Expectation-Maximization algorithm for Multi-scale Hawkes Processes under the assumption of exponential triggering kernels is proposed, while, in [33], a large-scale inference algorithm for kernels modeled after a summation of exponentials is introduced.

Automatic analysis frameworks for Gaussian Processes (GPs) are proposed in [13] and [14]. However, since the very concept of kernel is distinct between HPs and GPs, the techniques proposed and analyzed in this paper were developed entirely independently.

Acknowledgements

I would like to express my gratitude for:

- My parents, Humberto and Jenice, for the unconditional support throughout this exciting trajectory;
- My supervisor, Jaesik Choi, for the valuable advice towards the completion of this work;
- My siblings, Ricardo and Patricia, for the companionship;
- My aunts, Stella and Jacira, for always taking care of me, even through this very long distance from home;
- My cousins; Bruno, Humbe, Raonni, Julie and Brisa; for the nice laughs, even during those sleepless nights in the lab;
- My colleagues from Statistical Artificial Intelligence Laboratory, for the helpful comments and discussions;
- All the remarkable people I have met in UNIST and all around the country during my wonderful stay in Korea.

Earthquake Data for this work come from the Berkeley Digital Seismic Network (BDSN), doi:10.7932/BDSN, operated by the UC Berkeley Seismological Laboratory, which is archived at the Northern California Earthquake Data Center (NCEDC), doi: 10.7932/NCEDC.

References

[1] D. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*. Springer, 2003.

[2] A. G. Hawkes, "Spectra of some self-exciting and mutually exciting point processes," *Biometrika*, no. 1, pp. 201–213, 1971.

[3] S. W. Linderman and R. P. Adams, "Discovering latent network structure in point process data," in *Proceedings of the International Conference on Machine Learning*, 2014, pp. 1413–1421.

[4] G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita, "Self-exciting point process modeling of crime," *Journal of the American Statistical Association*, vol. 106, no. 493, pp. 100–108, 2012.

[5] Q. Zhao, M. A. Erdogdu, H. Y. He, A. Rajaraman, and J. Leskovec, "Seismic: A self-exciting point process model for predicting tweet popularity," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 1513–1522.

[6] P. Embrechts, T. Liniger, and L. Lin, "Multivariate hawkes processes: an application to financial data," *Applied Probability Trust*, 2011.

[7] E. Bacry and J. Muzy, "First- and second-order statistics characterization of hawkes processes and non-parametric estimation," *IEEE Transactions on Information Theory*, vol. 62, no. 4, pp. 2184–2202, 2016.

[8] Y. Ogata, "Seismicity analysis through point-process modelling: A review," *Pure and Applied Geophysics*, vol. 155, no. 5, pp. 471–507, 1999.

[9] J. Etesami, N. Kiyavash, K. Zhang, and K. Singhal, "Learning network of multivariate hawkes processes: A time series approach," in *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2016.

[10] E. Bacry, K. Dayri, and J. F. Muzy, "Non-parametric kernel estimation for symmetric hawkes processes. application to high frequency financial data," *The European Physical Journal B*, vol. 85, no. 5, pp. 1–12, 2012.

[11] H. Xu, M. Farajtabar, and H. Zha, "Learning granger causality for hawkes processes," in *Proceedings of the International Conference on Machine Learning*, 2016, pp. 1717–1726.

[12] T. Ozaki, "Maximum likelihood estimation of hawkes' self-exciting point processes," *Annals of the Institute of Statistical Mathematics*, no. 31, pp. 145–155, 1979.

[13] D. K. Duvenaud, J. R. Lloyd, R. B. Grosse, J. B. Tenenbaum, and Z. Ghahramani, "Structure discovery in nonparametric regression through compositional kernel search," in *Proceedings of the International Conference on Machine Learning*, 2013, pp. 1166–1174.

[14] Y. Hwang, A. Tong, and J. Choi, "Automatic construction of nonparametric relational regression models for multiple time series," in *Proceedings of the International Conference on Machine Learning*, 2016, pp. 3030–3039.

[15] E. Lewis and G. Mohler, "A nonparametric em algorithm for multiscale hawkes processes," *Journal of Nonparametric Statistics*, vol. 1, no. 1, pp. 1–20, 2011.

[16] S. Xiao, J. Yan, C. Li, B. Jin, X. Wang, X. Yang, S. M. Chu, and H. Zha, "On modeling and predicting individual paper citation count over time," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, 2016, pp. 2676–2682.

[17] A. Kulesza, "Learning with determinantal point processes," Ph.D. dissertation, University of Pennsylvania, 2012.

[18] L. Leemis, "Estimating and simulating nonhomogeneous poisson processes," 2003.

[19] P. Laub, T. Taimre, and P. Pollett, "Hawkes processes," 7 2015.

[20] T. Liniger, "Multivariate hawkes processes," Ph.D. dissertation, ETH Zurich, 2009.

[21] D. Marsan and O. Lengline, "Extending earthquakes' reach through cascading," *Science*, 2008.

[22] J. P. Bouchaud, J. D. Farmer, and F. Lillo, "How markets slowly digest changes in supply and demand," *Handbook of Financial Markets: Dynamics and Evolution*, pp. 57–156, 2008.

[23] E.Bacry, I. Mastromatteo, and J. Muzy, "Hawkes processes in finance," *arXiv*, 2015.

[24] H. H. L. H. H. S. John Gunnar Carlsson, Mao-Ching Foo, "Modelling stock orders using hawkes's self-exciting process," 3 2007.

[25] C. Zhang, "Modeling high frequency data using hawkes processes with power-law kernels," *Procedia Computer Science*, vol. 80, no. 5, pp. 762–771, 2016.

[26] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[27] M. Johansson, "The hilbert transform," Master's thesis, Vaxjo University, Sweden, 2006.

[28] G. Clark, "Phase retrieval from modulus using homeomorphic signal processing and the complex cepstrum: An algorithm for lightning protection systems," Lawrence Livermore National Laboratory, Tech. Rep., 2004.

[29] R. Paley and N. Wiener, *Fourier Transforms in the Complex Domain*. Colloquium Publications - American Mathematical Society, 1934.

[30] J. Moller and J. G. Rasmussen, “Perfect simulation of hawkes processes,” *Advances in Applied Probability*, vol. 37, no. 3, pp. 629–646, 2010.

[31] J. G. Rasmussen, “Temporal point processes: the conditional intensity function,” 2009.

[32] K. Zhou, H. Zha, and L. Song, “Learning triggering kernels for multi-dimensional hawkes processes,” in *Proceedings of the International Conference on Machine Learning*, 2013, pp. 1301–1309.

[33] R. Lemmonier, K. Scaman, and A. Kalogeratos, “Multivariate hawkes processes for large-scale inference,” in *Proceedings of the Conference on Artificial Intelligence*, 2017.

Appendix I

Mathematical Derivation of Stationarity Criteria for Multiplicative Combinations of Kernels

A.1 Introduction

This appendix introduces the full derivations of stationarity criteria for the second order multiplicative compositions of the four base kernels.

A.1.1 EXP x EXP

For the combination “EXPxEXP”, we have that, for stationarity to be achieved:

$$0 \leq \int_0^\infty EXP(\alpha_1, \beta_1) EXP(\alpha_2, \beta_2) dx < 1 \tag{A.1}$$

$$0 \leq \int_0^\infty \alpha_1 e^{\alpha_1 x} \alpha_2 e^{\beta_2 x} dx < 1 \tag{A.2}$$

Thus:

$$\int_0^\infty \alpha_1 e^{-\beta_1 x} \alpha_2 e^{-\beta_2 x} dx = \int_0^\infty (\alpha_1 \alpha_2) e^{-(\beta_1 + \beta_2) x} dx = \int_0^\infty \alpha e^{-\beta x} dx = \frac{\alpha}{\beta} = \frac{\alpha_1 \alpha_2}{\beta_1 + \beta_2} \tag{A.3}$$

So, this case reduces to the case of a single exponential.

A.1.2 EXP x PWL

For the combination “EXPxPWL”, we have that, for stationarity to be achieved:

$$0 \leq \int_0^\infty EXP(\alpha, \beta) PWL(K, c, p) dx < 1 \tag{A.4}$$

$$0 \leq \int_0^\infty \alpha e^{-\beta x} \frac{K}{(x+c)^p} dx < 1 \tag{A.5}$$

Thus:

$$\begin{aligned}
\int_0^\infty \alpha e^{-\beta x} \frac{K}{(x+c)^p} dx &= \alpha K \int_0^\infty (x+c)^{-p} e^{-\beta x} dx \\
&= \alpha K e^{\beta c} \int_{\beta c}^\infty (x+c)^{-p} e^{-\beta (x+c)} dx \\
&= \alpha K e^{\beta c} \beta^p \int_{\beta c}^\infty (\beta (x+c))^{-p} e^{-\beta x} dx \\
&= \alpha K e^{\beta c} \beta^p \Gamma(1-p, \beta c),
\end{aligned} \tag{A.6}$$

where $\Gamma(\cdot, \cdot)$ is the well-known Incomplete Gamma Function: $\Gamma(a, y) = \int_y^\infty t^{a-1} e^{-t} dt$.

A.1.3 EXP x SQR

For the combination “EXPxSQR”, we have that, for stationarity to be achieved:

$$0 \leq \int_0^{\infty} EXP(\alpha,\beta)SQR(B,L)dx < 1 \tag{A.7}$$

$$0 \leq \int_0^{L} \alpha B e^{-\beta x}dx < 1 \tag{A.8}$$

Thus:

$$\int_0^{L} \alpha B e^{-\beta x}dx = \left[\frac{\alpha B e^{-\beta x}}{\beta}\right]_0^L = \frac{\alpha B(1-e^{-\beta L})}{\beta} \tag{A.9}$$

So, in the case of a multiplicative combination, the SQR kernel acts as a truncation horizon.

A.1.4 EXP x SNS

For the combination “EXPxSNS”, we have that, for stationarity to be achieved:

$$0 \leq \int_0^{\infty} EXP(\alpha,\beta)SNS(A,\omega)dx < 1 \tag{A.10}$$

$$0 \leq \int_0^{\frac{\pi}{\omega}} A\alpha e^{-\beta x}sin(\omega x)dx < 1 \tag{A.11}$$

Where:

$$\begin{aligned}
\int_0^{\frac{\pi}{\omega}} A\alpha e^{-\beta x}sin(\omega x)dx &= \int_0^{\frac{\pi}{\omega}} A\alpha e^{-\beta x}\frac{e^{i\omega x}-e^{-i\omega x}}{2i}dx \\
&= \frac{A\alpha}{2i}\left[\frac{e^{(-\beta+i\omega)x}}{-\beta+i\omega} - \frac{e^{(-\beta-i\omega)x}}{-\beta-i\omega}\right]_0^{\frac{\pi}{\omega}} \\
&= \frac{A\alpha}{2i}\left[\frac{(-\beta-i\omega)e^{(-\beta+i\omega)x}-(-\beta+i\omega)e^{(-\beta-i\omega)x}}{\beta^2+\omega^2}\right]_0^{\frac{\pi}{\omega}} \\
&= \left[\frac{A\alpha e^{-\beta x}}{2i}\frac{2i\omega cos(\omega x)-2\beta sin(\omega x)}{\beta^2+\omega^2}\right]_0^{\frac{\pi}{\omega}} \\
&= \frac{A\alpha}{2i}\frac{-2i\omega(e^{\frac{-\beta\pi}{\omega}}-1)}{\beta^2+\omega^2} = \frac{A\alpha\omega(1+e^{\frac{-\beta\pi}{\omega}})}{(\omega^2+\beta^2)}
\end{aligned} \tag{A.12}$$

A.1.5 PWL x PWL

In the case of the combination “PWLxPWL”, an upper bound is derived as follows:

$$0 \leq \int_0^{\infty} PWL(K_1,c_1,p_1)PWL(K_2,c_2,p_2)dx < 1 \tag{A.13}$$

$$0 \leq \int_0^\infty \frac{K_1}{(x+c_1)^{p_1}} \frac{K_2}{(x+c_2)^{p_2}} dx < 1 \tag{A.14}$$

Then:

$$\begin{aligned}\int_0^\infty \frac{K_1}{(x+c_1)^{p_1}} \frac{K_2}{(x+c_2)^{p_2}} dx &\leq \int_0^\infty \frac{K_1 K_2}{(x+min(c_1,c_2))^{p_1+p_2}} dx \\ &= \frac{K_1 K_2}{(p_1+p_2-1)min(c_1,c_2)^{(p_1+p_2-1)}}\end{aligned} \tag{A.15}$$

A.1.6 PWL x SQR

For the combination “PWLxSQR”, we have that, for stationarity to be achieved:

$$0 \leq \int_0^\infty PWL(K,c,p)SQR(B,L)dx < 1 \tag{A.16}$$

$$0 \leq \int_0^L \frac{KB}{(x+c)^p} dx < 1 \tag{A.17}$$

Where:

$$\int_0^L \frac{KB}{(x+c)^p} dx = \left[\frac{KB}{(1-p)(x+c)^{(p-1)}}]\right]_0^L = \frac{KB(c^{-(p-1)} - (c+L)^{-(p-1)})}{p-1} \tag{A.18}$$

So, once again, the SQR kernel acts as a truncation horizon.

A.1.7 PWL x SNS

In the case of the combination “PWLxSNS”, an upper bound is derived as follows:

$$0 \leq \int_0^\infty PWL(K,c,p)SNS(A,\omega)dx < 1 \tag{A.19}$$

$$0 \leq \int_0^{\frac{\pi}{\omega}} \frac{KAsin(\omega x)}{(x+c)^p} dx < 1 \tag{A.20}$$

Where:

$$\begin{aligned}\int_0^{\frac{\pi}{\omega}} \frac{KAsin(\omega x)}{(x+c)^p} dx &\leq \int_0^{\frac{\pi}{\omega}} \frac{KA}{(x+c)^p} dx \\ &= \left[\frac{KA}{(1-p)(x+c)^{(p-1)}}]\right]_0^{\frac{\pi}{\omega}} \\ &= KA\frac{((c+\frac{\pi}{\omega})^{1-p} - c^{1-p})}{1-p}\end{aligned} \tag{A.21}$$

A.1.8 SQR x SQR

For the combination "SQRxSQR", we have that, for stationarity to be achieved:

$$0 \leq \int_0^{\infty} SQR(B_1, L_1)SQR(B_2, L_2)dx < 1 \tag{A.22}$$

$$0 \leq \int_0^{min(L_1,L_2)} B_1B_2dx < 1 \tag{A.23}$$

Where:

$$\int_0^{min(L_1,L_2)} B_1B_2dx = B_1B_2min(L_1, L_2) = BL \tag{A.24}$$

So, the multiplicative combination of two SQR kernels may be reduced to the case of a single SQR kernel.

A.1.9 SQR x SNS

In the case of combinations of discontinuous kernels (SQR and SNS), we assume they have the same starting and ending points, i.e., $L = \frac{\pi}{\omega}$. So, for the combination "SQRxSNS", we have that, for stationarity to be achieved:

$$0 \leq \int_0^{\infty} SQR(B, L)SNS(A, \omega)dx < 1 \tag{A.25}$$

$$0 \leq \int_0^{\frac{\pi}{\omega}} ABsin(\omega x)dx < 1 \tag{A.26}$$

Where:

$$\int_0^{\frac{\pi}{\omega}} ABsin(\omega x)dx = \frac{2AB}{\omega} \tag{A.27}$$

A.1.10 SNS x SNS

In the case of combinations of discontinuous kernels (SQR and SNS), we assume they have the same starting and ending points. So, for the combination "SNSxSNS", we have that, for stationarity to be achieved:

$$0 \leq \int_0^{\infty} SNS(A_1, \omega)SNS(A_2, \omega)dx < 1 \tag{A.28}$$

$$0 \leq \int_0^{\frac{\pi}{\omega}} A_1A_2sin^2(\omega x)dx < 1 \tag{A.29}$$

Where:

$$\int_0^{\frac{\pi}{\omega}} A_1A_2sin^2(\omega x)dx = \int_0^{\frac{\pi}{\omega}} A\frac{(1-cos(2\omega x))}{2}dx = \frac{\pi A}{2\omega} \tag{A.30}$$

Appendix II

Log-likelihood formula for HPs

B.1 Derivation

This derivation follows the steps on [19]. Given a realization $(t_1, t_2, ..., t_k)$ of some regular point process observed over the interval [0,T], the log-likelihood is expressed as:

$$l = \sum_{i=1}^{k} \log(\lambda(t_i)) - \int_0^T \lambda(u)du \tag{B.1}$$

Proof. Let be the joint probability density of the realization:

$$L = f(t_1, t_2, ..., t_k) = \prod_{i=1}^{k} f(t_i) \tag{B.2}$$

It can be written in terms of the Conditional Intensity Function. We can then find f in terms of λ:

$$\lambda(t) = \frac{f(t)}{1 - F(t)} = \frac{\frac{\mathrm{d}F(t)}{\mathrm{d}t}}{1 - F(t)} = -\frac{\mathrm{d}\log(1 - F(t))}{\mathrm{d}t}, \tag{B.3}$$

where, given the history up to last arrival u, $\mathcal{H}(u)$, F(t) is then defined as the conditional cumulative probability distribution of the next arrival time T_{k+1}:

$$F(t) = F(t|\mathcal{H}(u)) = \int_u^t f(s|\mathcal{H}(u))ds \tag{B.4}$$

Integrating both sides of equation (B.3) over (t_k, t):

$$-\int_{t_k}^{t} \lambda(u)du = \log(1 - F(t)) - \log(1 - F(t_k)) \tag{B.5}$$

Given that the realization is assumed to have come from a so-called *simple process*, i.e., a process in which multiple arrivals cannot occur at the same time, we have that $F(t_k) = 0$ as $T_{k+1} > t_k$, which simplifies equation (B.5) to:

$$-\int_{t_k}^{t} \lambda(u)du = \log(1 - F(t)) \tag{B.6}$$

Further rearranging the expression:

$$F(t) = exp\left(-\int_{t_k}^{t} \lambda(u)du\right), \tag{B.7}$$

and

$$f(t) = \lambda(t)exp\left(-\int_{t_k}^{t} \lambda(u)du\right) \tag{B.8}$$

Thus, the likelihood becomes:

$$\begin{aligned} L = \prod_{i=1}^{k} f(t_i) &= \prod_{i=1}^{k} \lambda(t_i) exp\left(-\int_{t_{i-1}}^{t_i} \lambda(u)du\right) \\ &= \left[\prod_{i=1}^{k} \lambda(t_i)\right] exp\left(-\int_{0}^{t_k} \lambda(u)du\right) \end{aligned} \tag{B.9}$$

Given that the process is observed on $[0,T]$, the likelihood must include the probability of seeing no arrivals in $(t_k,T]$:

$$L = \left[\prod_{i=1}^{k} f(t_i)\right](1-F(T)) \tag{B.10}$$

Through using the formulation of F(t), we have that:

$$L = \left[\prod_{i=1}^{k} \lambda(t_i)\right] exp\left(-\int_{0}^{T} \lambda(u)du\right) \tag{B.11}$$

Finally, getting the logarithm of the expression, we have the formula for *l*:

$$l = \sum_{i=1}^{k} \log(\lambda(t_i)) - \int_{0}^{T} \lambda(u)du \tag{B.12}$$

□