ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Estimating self-excitation effects for social media using the Hawkes process

Master Thesis

Daniel MacKinlay

Submitted 13th April, 2015;
Corrected 11th May 2015

Advisors: Prof. S. van de Geer / Prof. D. Sornette

Seminar For Statistics / D-MTEC, ETH Zürich

Abstract

Are viral dynamics, the peer-to-peer propagation of ideas and trends, important in social media systems, compared to external influence? Are such dynamics quantifiable? Are they predictable? We suspect they are, in at least some cases, but quantifying how and when such dynamics are significant, and how and when we can detect this, remains an open question.

This thesis investigates how to estimate the parameters of branching dynamics in a large heterogeneous social media time series data set.

The specific model that I use, the class of *Hawkes processes* has been used to model a variety of phenomena characterized by "self-exciting" dynamics - broadly speaking, time-series where "lots of things happening recently" is the best predictor of "more things happening soon", conditional upon the external input to the system.

Variants have been applied as models of seismic events, financial market dynamics, opportunistic crime, epidemic disease spread, and viral marketing. Detecting self-exciting dynamics is of huge importance in these application areas, where it can make a large difference in the certainty and accuracy of prediction, and of the usefulness and practically of interventions to change behavior of the system.

This data I investigate, documenting the time evolution of Youtube views counters, was collected by Crane and Sornette [CS08].

The notoriously viral nature of Youtube popularity suggests that this could supply an opportunity to attempt to quantify these viral dynamics.

The data set has various characteristics which make it a novel source of insight, both into self exciting phenomena, and into the difficulties of estimating them. The time series exhibit a huge variety of different behavioral regimes and different characteristics. While this data contains many observations, it is also incomplete, in the sense that rather than a complete set of occurrence times, there are only sample statistics for that data available.

These qualities present challenges both to the model I attempt to fit, and the estimator that I use to fit the model.

This places some constraints upon how how precisely I can identify branching dynamics, and with how much certainty, and the kind of hypotheses I can support.

This thesis consists of two major phases.

In the *first* phase, I attempt to address the question: What component of the Youtube video views may be ascribed to self-excitation dynamics? In this regard I will attempt to estimate the parameters of generating Hawkes process models for various time series to identify the "branching coefficient" of these models, which is one measure of the significance of viral dynamics.

Based on naive application of the model, I find the evidence is ambiguous. Whilst I cannot reject the hypothesis of branching dynamics, I show that

the model class is unidentifiable within this framework due to several problems.

The first is the unusual form of the dataset; the incompleteness of the time series leads to missing data problems with no immediate and computationally tractable solution.

But even with complete data, I face a second class of problems due to model misspecification. For example, we should be surprised to ever fail to find branching dynamics at work, since branching dynamics is the only explanation permitted for intensity variation in this particular model. The classical Hawkes model assumes away any other source of time-variability, including exogenous influences.

The homogeneity assumption is not essential to modeling self-exciting systems, but merely a convenient assumption in a particular model class.

Therefore, in the *second* phase of the project I consider how to address these difficulties by weakening this assumption. I address the most commonly mentioned source of inhomogeneous behavior, exogenous influence, in what I believe to be a novel fashion.

I use penalized semi-parametric kernel estimators to the data to simultaneously recover exogenous drivers of system behavior and the system parameter. A simple implementation of this idea recovers model parameters under plausible values for the dataset.

The particular combination of estimators and penalties I use here is, to the best of my knowledge, novel, and there are limited statistical guarantees available. I address this deficit with simulations, and discuss how the results might be given more rigorous statistical foundation.

When applied to the data set in hand, the Youtube data, I find that there is support for the significance of branching dynamics; However, the parameters of the inferred process are different to those o the homogeneous estimator. This implies it is crucial to consider the driving process in fitting such models, and the supports the utility of the inhomogeneous methods such as the one I use here to do so.

Contents

Contents **3**

1 Background **3**
1.1 Fluctuation in social systems 3
1.2 Youtube 4

2 The data **5**
2.1 Nomenclature 5
2.2 Outliers and Dragon Kings 10
2.3 Lead Balloons 12
2.4 Hypotheses 14

3 A quick introduction to point process theory **17**
3.1 Univariate temporal point processes 17
3.2 Conditional intensity processes 18
3.3 Kernels 19
3.3.1 Exponential kernel 20
3.3.2 "Basic" power-law kernel families 20
3.4 The Hawkes Process in action 21

4 Estimation of parameters of the homogeneous Hawkes model **23**
4.1 Estimating parameters from occurrence timestamps 24
4.2 Estimation from summary statistics 25
4.3 Model selection 28
4.3.1 The Akaike Information Criterion 28
4.3.2 General difficulties with AIC 29
4.4 Experimental hypotheses 30
4.4.1 Model selection with large data sets 30
4.4.2 Multiple testing considerations 31
4.4.3 Goodness-of-fit tests 31

5 Simulations for the homogenous estimator **33**
5.1 Point estimates . . . 33
5.2 Model selection . . . 39
5.3 Empirical validation of estimators of inhomogenous data . . . 42

6 Results for the homogeneous Hawkes model **47**
6.1 Further work . . . 53
6.1.1 Expanding the candidate set . . . 53
6.1.2 Summary data estimation . . . 54
6.1.3 Goodness of fit . . . 56

7 Estimating branching effects for inhomogeneous processes **57**
7.1 Semiparametric kernel density estimation . . . 58
7.2 The algorithm . . . 59
7.3 Model selection . . . 66

8 Simulations for the inhomogeneous estimator **69**
8.1 Empirical validation on simulated data . . . 69

9 Results for the inhomogeneous Hawkes model **79**
9.1 Single time series detailed analysis . . . 79
9.2 Aggregate analysis . . . 80

10 Conclusions **85**

A Technical Notes **87**
A.1 Data extraction and cleaning . . . 87
A.2 On the complexity of the simplest possible thing . . . 89

Bibliography **91**

Acknowledgements

In addition to Prof. Didier Sornette, who proposed and supported this project, I am indebted to many others.

I am indebted to Dr Vladimir Filimonov for timely feedback on my estimation procedures and the usage of the software, to Prof. Hansruedi Künsch and Spencer Wheatley for useful discussions about Bayesian methods and about kernel estimators, which allowed me to develop my ideas further than I otherwise could. I am thankful to Prof. Sara van de Geer for supporting this chaos on behalf the Seminar for Statistics.

I am also grateful to Miriam Lyons, Markus Hochuli, Katharina Hugentobler and Aline Ulmer for tolerating my absence during the deadlines, and cooking me dinner anyway.

The remaining mistakes, and dirty dishes, are my own responsibility.

Chapter 1

Background

1.1 Fluctuation in social systems

The notorious unpredictability and variability of social systems has achieved new levels of prominence, and possibly new extremes, with the rise of the internet. These unpredictable viral dynamics of social media have variable impact, variable magnitude and impacts, and little connection between these dimensions. Consider:

1. On the 26th of February 2015, a low-quality photograph of a dress of indeterminate color sparking a battle on the internet that garnered 16 million views within 6 hours on Buzzfeed alone. [Sha15]
2. A 61-million person experiment on peer recommendations by Facebook found that strategically applied viral peer-to-peer systems can mobilize citizens politicly on a massive scale. The authors estimate that they were able to garner 280,000 extra votes in the election using this system - enough to strongly influence the outcome of federal elections in the US. [Bon+12]
3. Religious militant organization *Islamic State of Iraq and Syria*, ISIS, exploits viral meme propagation to recruit volunteers and attract funding for its military campaigns by peer recommendation on Youtube and Twitter. [Ber14]

Understanding how, and when, and why this kind of viral propagation takes place is crucial to understanding the function of modern society. Why did that particular dress photograph have such an impact? For that matter, as impressive as the scale of the voter experiment is, it took the backing of a multi-billion dollar corporation to produce this effect, and yet the viral dress photo was simply a thoughtless photograph from a cheap phone camera. And yet, as we see from ISIS, understanding the dynamics of these peer-to-peer systems is implicated in global life-and-death struggles and violent political upheaval.

Learning to understand the dynamics of these systems is economically and politically important. And, thanks to the quantification of communication on the

internet, potentially plausible.

One piece of the puzzle of such systems, which I explore here, is the use of models of self-exciting systems. In such system, activity may be understood to be partly *exogenous*, triggered by influences from the outside world, and partly *endogenous*, triggered by their own past activity. [SMM02; DS05; CSS10] Concretely, this stylized description is the kind of dynamic we observe in, for example, financial markets, where (exogenous) news about a certain company might trigger movement in the price of its stock, but also movement in the price of a company's stock could itself trigger further movements as traders attempt to surf the tide. In social systems, the mysterious popularity of the photograph of a dress viewed 16 million times in a single day is a paradigmatic example of endogenous triggering; there is no plausible news content attached to it.

The particular self-exciting system that I use here is the linear Hawkes process This model has been applied to such diverse systems as earthquakes, [Oga88] product popularity, [DS05; IVV11] financial markets, [HBB13; Fil+14] social media, [CSS10] crime, [Moh+11] neural firing, [Bro+02] and many others. [Sor06]

If we can successfully explain the dynamics of the data using the Hawkes process model, then we are a step closer quantitative predictions of the process behavior, and of future unpredictability by measuring and predicting the importance of the endogenous versus the exogenous component of such systems.

1.2 Youtube

The particular data that I have was collected from Youtube, the social video sharing website. Youtube is owned by Google and headquartered in the USA. It was founded in February 2005 and officially launched in November of the same year.

Distribution of popularity of video on Youtube is often claimed to exhibit classic indicators of the kind of "heavy-tailed" behavior that would indicate certain kinds of self-exciting process behavior. For example, in 2011 a YouTube software engineer was asserted to reveal that 30% of videos accounted for 99% of views on the site. [1] [Whi11]

Shortly before the time that this dataset was collected, YouTube reported that each day it served 100 million videos and accepted more than 65,000 uploads. [REU06]. As at January 2012, they reported approximately 4 billion daily video views. [Ore12] and individual videos with more than 2 billion views. [You14]

They seem, in other words, a perfect test bed to experiment with self exciting models, if we can get the right sort of data about them, and the right methods to analyze it. This brings me to the question of inspecting the data.

[1] This often-cited statistic a published in British newspaper *the Telegraph* without references and I have been unable to find primary sources for its claims. Nonetheless, as I will show later, it is plausible given my dataset.

Chapter 2

The data

I present qualitative description of the cleaned data here. Technical details of the cleaning process are available in the supplement.

As much as possible mathematics and analysis will be reserved for later chapters, with the exception of some essential terms.

2.1 Nomenclature

The data set comprises many separate time series, each comprising certain summary statistics for an underlying view-count process.

The underlying process, the increments of the view counter time series for a given video I will call *occurrences*. The summaries, of how many view counts have occurred at what time, are *observations*. Each time series is made of many observations, and more occurrences. (section 2.1)

I will denote to the underlying view-counter process as $N_v(t)$, where t indexes time and the subscript v indexes over all time series. Normally I will omit the subscript, unless I need to distinguish between two time series.

For a given series time, I have only n observations of the value of the view counter, on an interval $[0, T]$ at times $\tau_{i1<i\leq n}$ where I take $\tau_1 = 0, \tau_n = T$. I write such observation tuples $\{(\tau_i, N(\tau_i))\}_{1<i\leq n}$. It will always be clear from the context *which* time series a given set of timestamps belong to, although it should be understood that there is an implicit index v, i.e. $\{(\tau_{(v,i)}, N_v(\tau_{(v,i)}))\}_{1<i\leq n_{(v)}}$.

The dataset was gathered from 13. October 2006 until 25. May 2007 for use in an article published in 2008, [CS08] Information was *scraped* from Youtube, which is to say, extracted by machine text processing of web page data by an automated web browser; The web pages in question, in this case, are the pages for individual videos displayed on Youtube; To pick one example, the time series encoded as `epUk3T2Kfno` is available at https://www.youtube.com/watch?v=epUk3T2Kfno

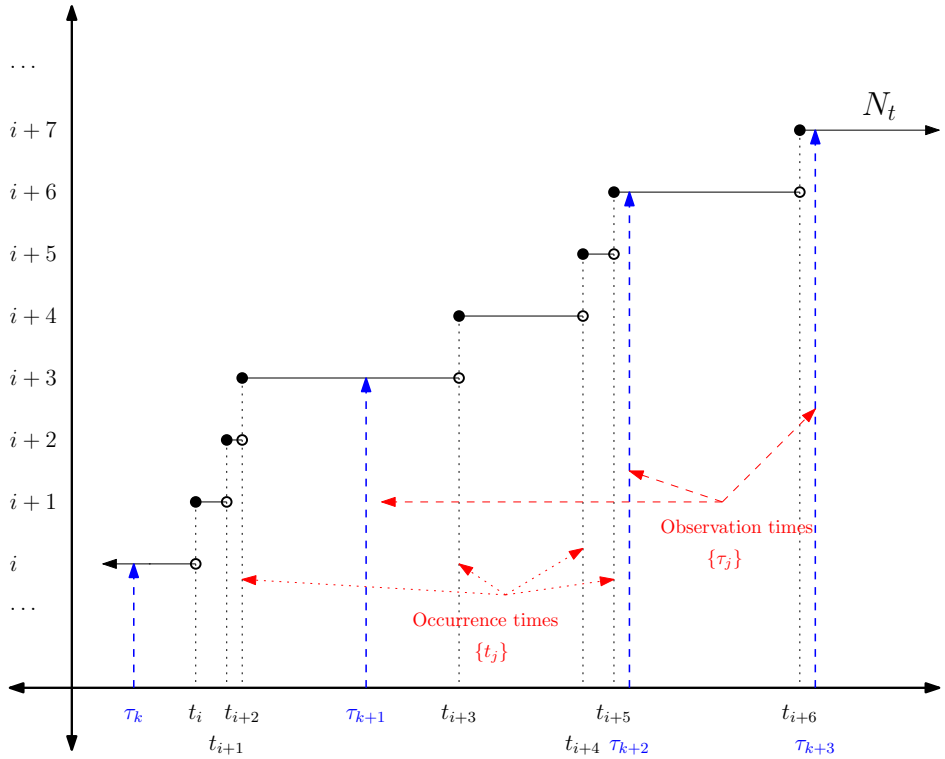

Figure 2.1: Model of the observation procedure of the time series

and corresponds to a video entitled *Otters holding hands*, uploaded on Mar 19, 2007, with summary information

> Vancouver Aquarium: two sea otters float around, napping, holding hands. SO CUTE!

which is an accurate summary of the 88 second cinematic masterpiece. (Figure 2.2)

Source code for the Youtube sampling is no longer available, and limited communication with the author has been possible, so I adopt a conservative approach to interpretation of the available data.

One unusual quality of the data is an administrative one: at the time of data collection, there was no known prohibition against automated data collection from Youtube. At the time of writing, however, the current Youtube Terms Of Service agreement for Switzerland (date 2013/4/3) expressly prohibit the use of automated data collection. Even if I can find a jurisdiction with more permissive Terms of Service, I would have to circumvent complex software defense mechanisms to prevent automated data extraction. I am thus precluded from automated verification of hypotheses developed from this data; I may, however, legally *manually* verify a small number of hypotheses, insofar as that is possible from normal infor-

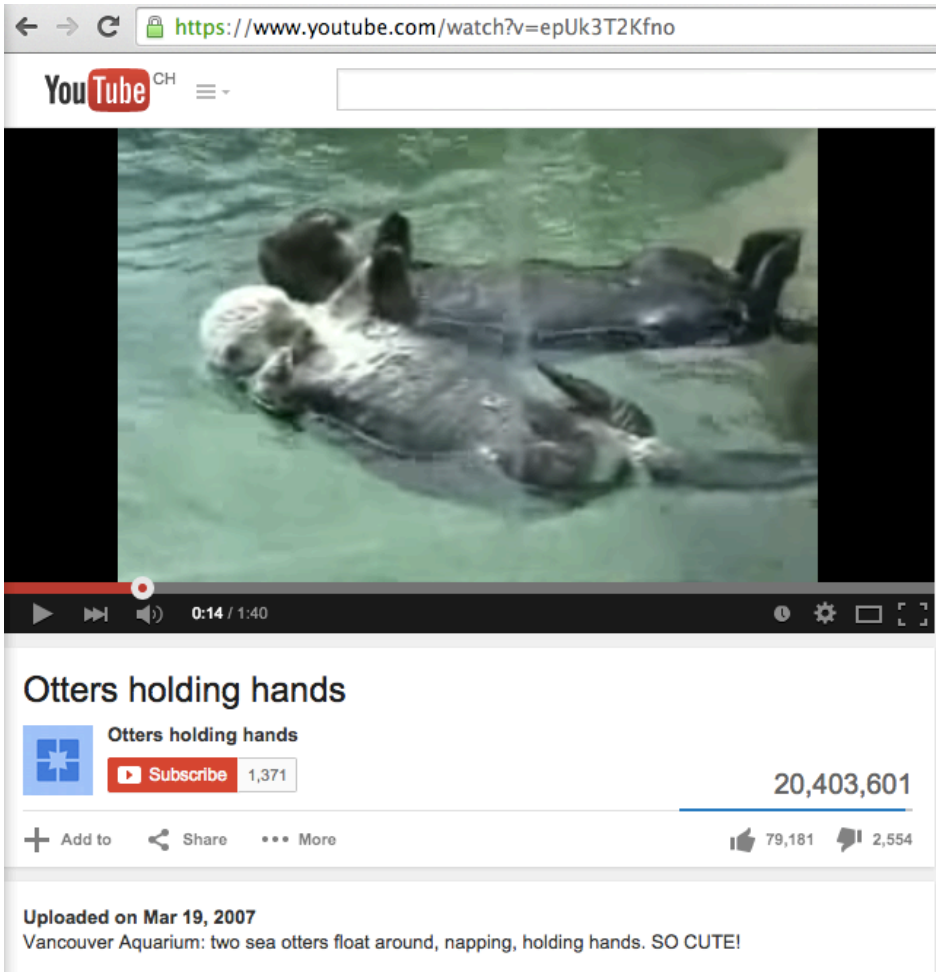

Figure 2.2: Screen capture of "Otters holding hands". Content copyright by the Youtube user "Otters holding hands."

mation available to the user of a browser. This fact will be significant in discussing optimal semiparametric regression strategies later on.

Timespans for individual video series span subsets of the overall interval, and are variously sampled at different rates. The observation interval for a different video can vary from seconds to days - After my data set cleaning, details of which are discussed elsewhere *Data extraction and cleaning*, the rate is approximately 3 observations per calendar day, but varies apparently randomly over time and between videos. There is no obvious correspondence between the observation rates of different videos' time series, or between the observation rate and qualities of the video itself, such as popularity.

The timestamp of the ith such increment I take to be τ_i. One could consider taking this data as a noisy estimate of the true unobserved observation time $\hat{\tau}_i$.

A principled approach to this data decontamination would then be to construct a stochastic process model for the observation time to reflect the stochastic relationship between *recorded* counter value and *true* counter value. One could also attempt to correct for view rates to allow for the time-zone a video is likely to be viewed from and when its viewers would be awake and so forth. The sampling intervals are messy enough that I doubt we could extract such information. An analysis of the robustness of the estimator under perturbation of timestamps to estimate the significance of these assumptions would be wise. I leave that to later work.

As I depend upon asymptotic results in the estimation packages, I cannot learn much from small time series. I discard all series with less than 200 observations. This value is somewhat arbitrary, but is chosen to include a "hump" in the frequency of time series with around 220 observations. This constitutes non-random censoring of the time series due to the data cleaning process, as discussed in the technical supplement. The data is likely already censored, however, as discussed in the technical supplement, and I put this problem aside for future research.

After filtering, 253,326 time series remain. These time series exhibit a range of different behavior, different sampling densities, total number of occurrences, and view rates. (Figure 2.3)

I approximate the instantaneous rate of views for a given time series by a piecewise constant function for visualization.

For compatibility with the notation I use later, I denote this estimate $\hat{\lambda}_{\text{simple}}(t)$, and define it

$$\hat{\lambda}_{\text{simple}}(t) := \sum_{i=2}^{n} \frac{N(\tau_i) - N(\tau_{i-1})}{\tau_i, \tau_{i-1}} \left(\mathbb{I}_{[\tau_{i-1} - \tau_i)}(t) \right) \tag{2.1}$$

$\mathbb{I}_A$ is the indicator function for set A. An example is pictured in Figure 2.4.

Finally we are in a position to actually frame questions about this data.

We might ask if the spikes in this video can be explained by endogenous branching dynamics, or exogenous influence. What could explain the variability in this time series? Is it a video shared for its intrinsic interest, or it is responding to external events?

Sleuthing reveals that the subject of the video, Mexican singer-songwriter Valentin Elizalde, was assassinated at around the upload time of this video. That is a plausible explanation for the initial peak in interest. But the later resurgence?

An biography suggests one hypothesis:

> When he was alive, he never had a best-selling album. But less than four months after his murder and half a year after "To My Enemies"

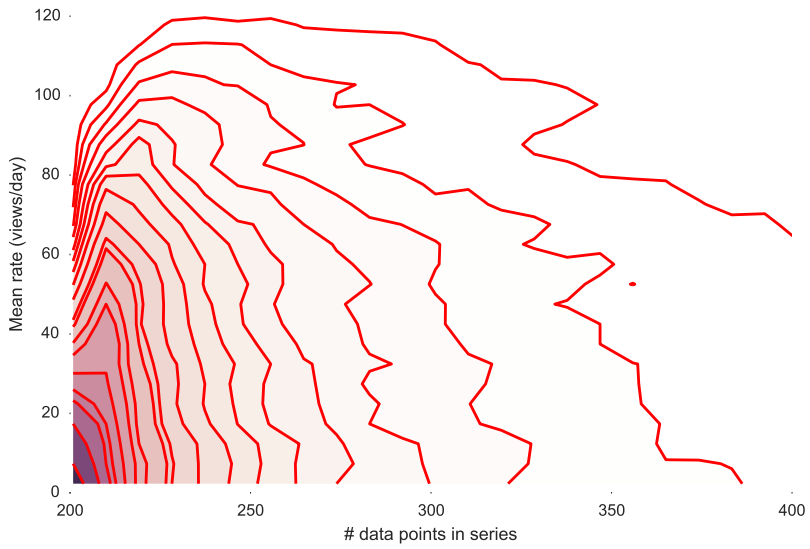

Figure 2.3: Distribution of the 240659 time series with at least 200 observations, in terms of number of data points and mean daily rate. Each successive contour encloses approximately an extra 5% of the total number of time series, totaling 95% of the observations. Some of the final 5% possess mean rate values orders of magnitude greater than the median, and the 0% contour line is therefore excluded for clarity.

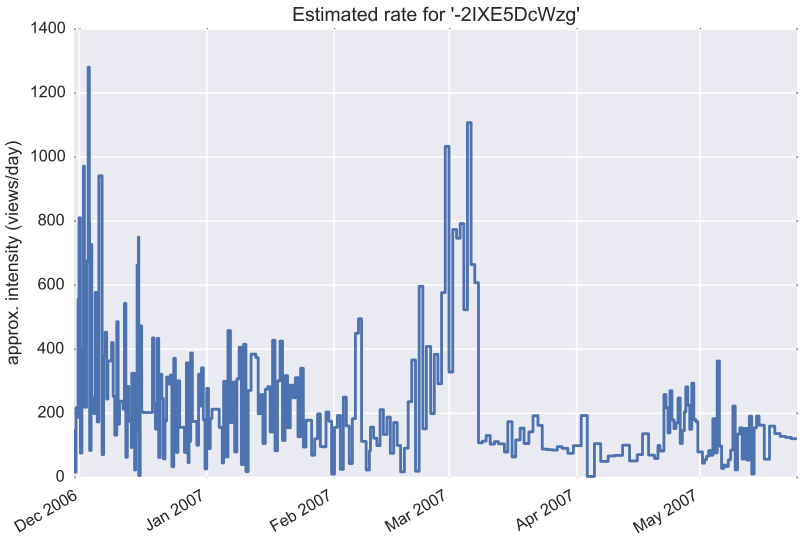

Figure 2.4: View-rate estimate $\hat{\lambda}_{simple}(t)$, for time series *Valentin Elizalde, Volvere a amar*

> became an Internet hit, Elizalde made it big. On March 3, when Billboard came out with its list of best-selling Latin albums in the United States, Elizalde occupied the top two spots. [Roi07]

Was it Elizalde's success in Billboard magazine that lead to the spike in video views? I will return to this question later.

2.2 Outliers and Dragon Kings

We need to consider whether the kind of behavior that we witness amongst large time series, in the sense of having many occurrences recorded, are similar to the results for small time series. For one, this kind of regularity is precisely the kind of thing that we would like to discover. For another thing, if there is no such regularity, that would be nice to know too, as the estimators I use scale very poorly in efficiency with increasing occurrence count.

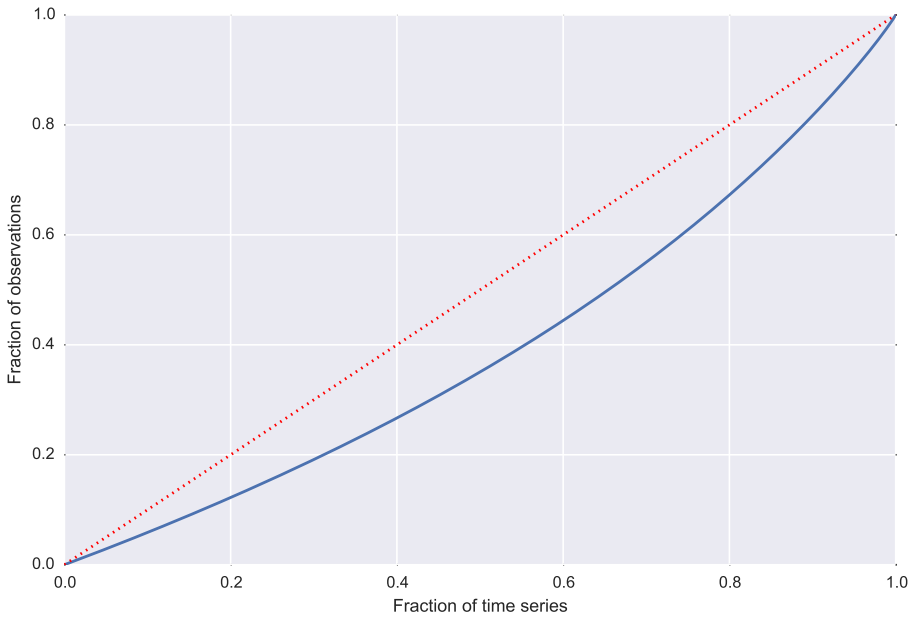

Figure 2.5: Cumulative distribution of **observations** by time series. Dotted red line denotes a the curve of a hypothetical uniform allocation of observations to time series.

I plot here the distribution of sizes *amongst* the time series, in Figure 2.5 and Figure 2.6, and logarithmically in Figure 2.7. We observe an extremely skewed distribution; 25% of the total occurrences recorded by the (filtered) data set are contained in only 671 time series. It is tempting to draw comparison with Sornette's "Dragon King" phenomenon, [Sor09] although given the unknown data censoring process, I will not attempt to draw conclusion about the population

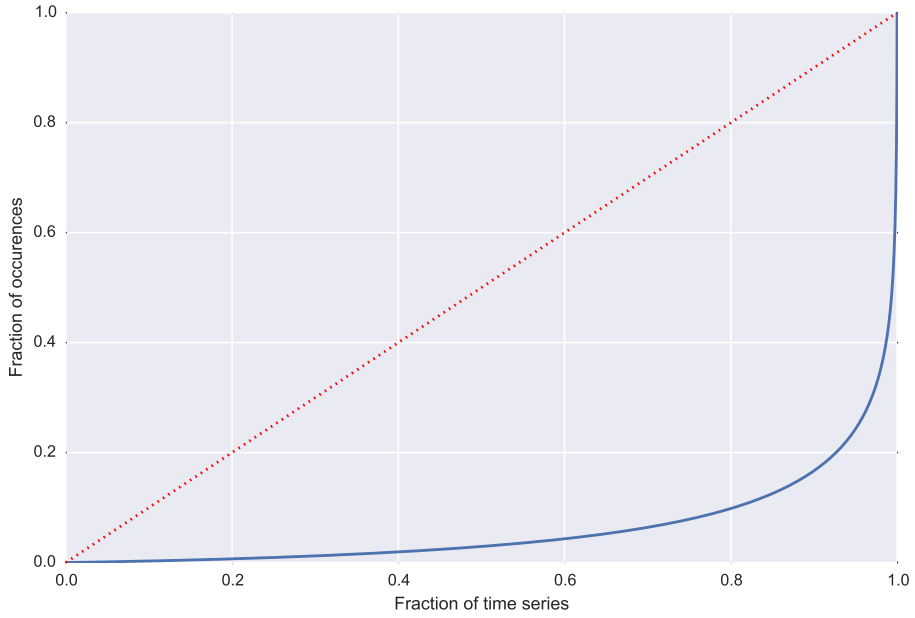

Figure 2.6: Cumulative distribution of **occurrences** by time series. Dotted red line denotes a the curve of a hypothetical uniform allocation of occurrences to time series.

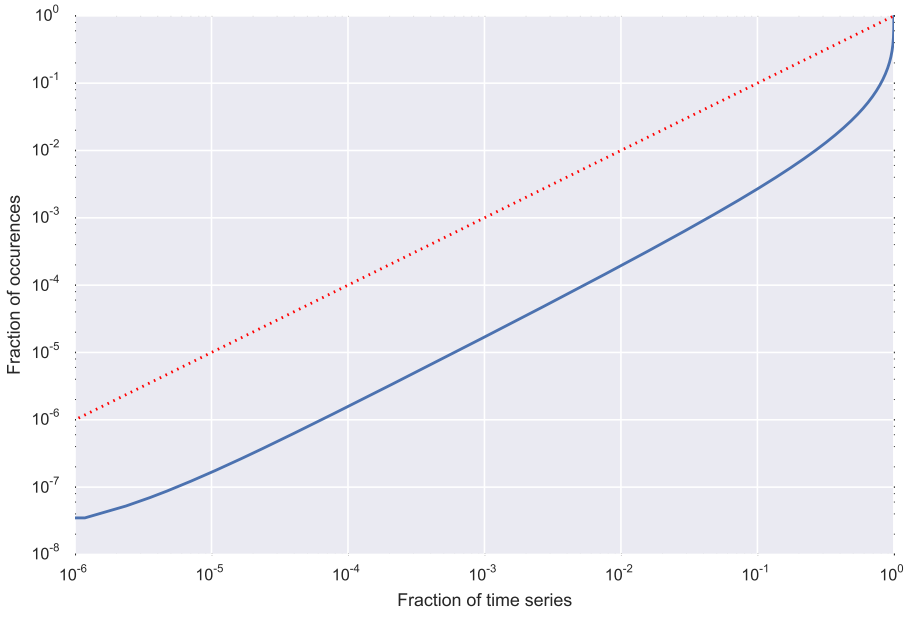

Figure 2.7: Cumulative distribution of **occurrences** by time series, log-log scale. Dotted red line denotes a the curve of a hypothetical uniform allocation of occurrences to time series.

of Youtube videos from sample here. If we wish to ultimately understand this data set, the extremely large number of total views concentrated in a small proportion of total videos will be significant in determining a sampling strategy. If nothing else, the raw number of points in these times series is computationally challenging for our estimators.

2.3 Lead Balloons

The self-exciting model is interesting precisely *because* it can produce variable dynamics. As such, extreme rate variation within a time series or between time series is not necessarily a problem for the model. On the other hand, the Maximum Likelihood estimators that I develop here are sensitive to outliers, so we need to see the kind of problems the data presents, especially where they represent the kind of extreme behavior that will be an outlier with respect to the model.

There are time series where unambiguous external evidence leads us to suspect that the process has undergone an exogenous shock, leading to a sudden increase or decrease in view rate. Sometimes this is due to a clear time limit on a video's relevance. (Figure 2.8)

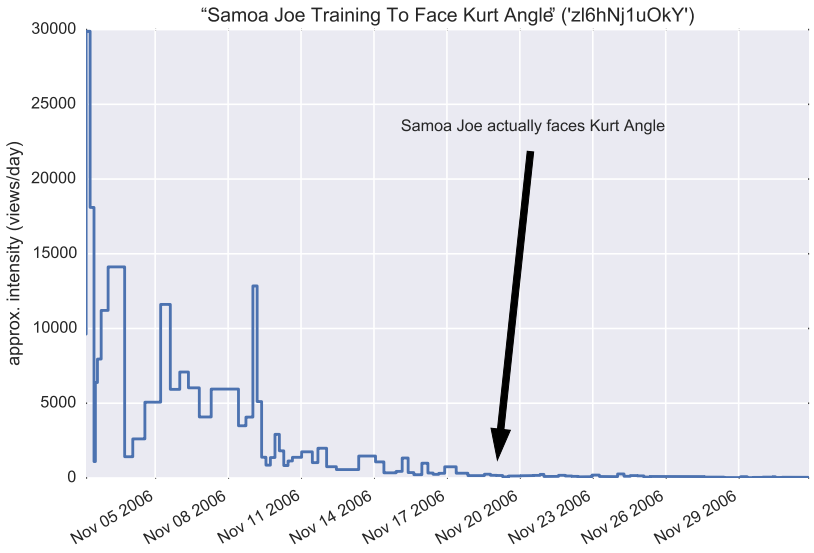

Figure 2.8: A time series with rapid decline

More extreme than sudden loss of interest are the sudden rate "spikes" early in the life of a time series, containing most of the information There is massive

activity at the beginning of the time series, and virtually none thereafter. I call these series *lead balloons*, after their trajectories. (Figure 2.9)

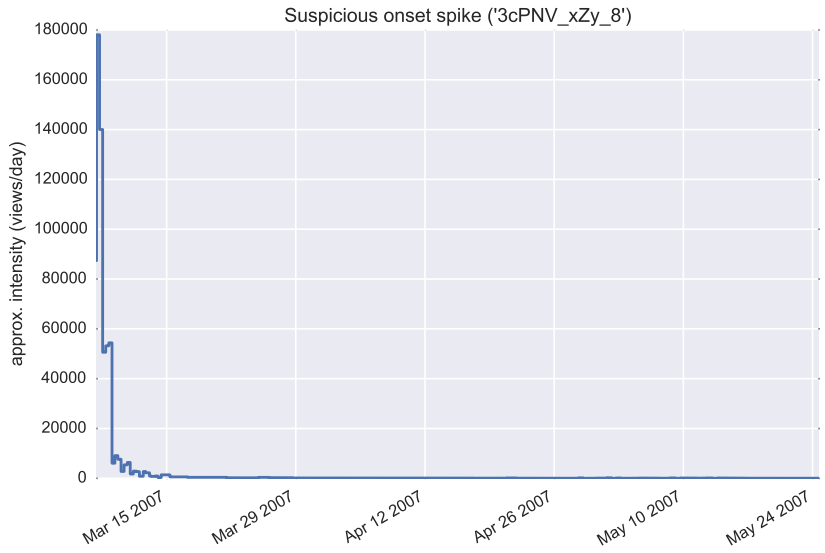

Figure 2.9: A time series with enormous early rate spike. The view rate collapses so suddenly that it is nearly invisible.

It is not immediately clear if these spikes are because of genuine collapses in popularity of a video, or if they are technical artifact. In the case of the last example, the initial spike dwarfs all other activity in the time series, although it never stops entirely. I repeat it on a logarithmic scale, where we can see that the initial rate is orders of magnitude above later activity. (Figure 2.10)

Presuming these spikes a a real phenomenon, one explanation for one would be that something, perhaps a mention on television, has promoted interest, but that the video itself has absolutely no viral potential.

Some sleuthing reveals that this example was video of a notorious brawl at the 2007/3/6 Inter Milan versus Valencia football game leading to a 7 month ban for Valencia player David Navarro. The video was uploaded shortly after the controversial match. It seems plausible that millions of soccer fans who switched off the uneventful game resorted to Youtube to watch the fight they missed at the end; But David Navarro has little viral potential; Once you have seen him brawling once, that is enough.

The majority of these lead balloons have no metadata available in my data set, and one cannot often not acquire any additional metadata even with effort, as videos

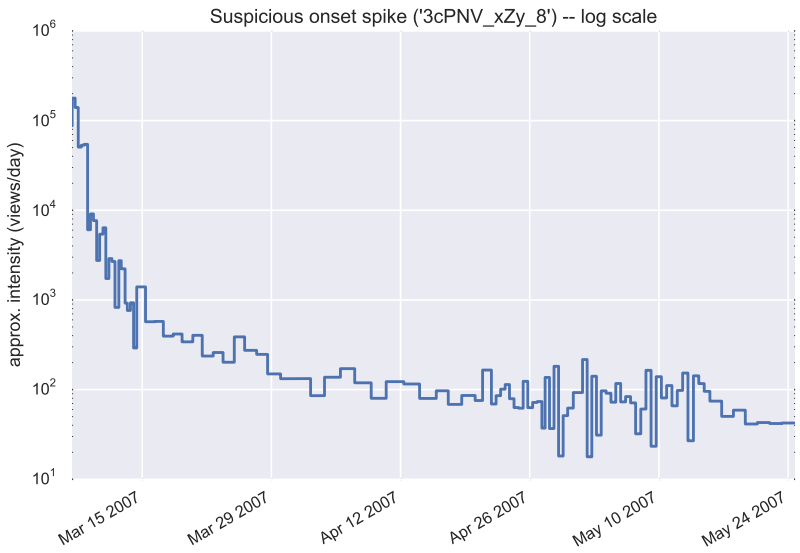

Figure 2.10: The lead balloon time series with enormous early rate spike, log vertical scale to show continued activity.

in this category have often been removed from Youtube. This suggests that perhaps they represent controversial or illegal content which was briefly wildly popular but quickly censored. However, the view counter for deleted videos is, at time of writing, not visible, so we would expect that time series for deleted videos would simply be truncated entirely, not vastly reduced. There is no easy way of deciding this here, but I return to the issue later.

Research on similar systems suggests such sudden spikes are likely to be a common and important part of the dynamics. For example, celebrity mentions affect book sales [DS05; Sor+04] and natural disasters affect charity donations. [CSS10]

There are other classes of stylized dynamics, but the sorts listed here already comprise enough complexity and difficulty for one paper, and accordingly I leave the dataset analysis for the time being.

2.4 Hypotheses

The data has many stylized features of other famous "social contagion" data; It has variable dynamics, a concentration of much activity into a small number of members of the data set and so on.

Whether this fits into the particular framework of the Hawkes process is another question. It seems likely that a branching process fit to such data would be unlikely to support a single background rate or branching ratio for all the data; We

might hypothesis about the distribution of such parameters, e.g. that the generating process is an Omori kernel will a certain exponent. The hypothesis that there might be characteristic timescales or other stylized behavior for such data also seems reasonable. The question is whether the significance of such effects, if any, is quantifiable or identifiable with the tools at we have.

Chapter 3

A quick introduction to point process theory

I construct the point process theory necessary for the model here informally; More formal introductions to the field may be found in the standard textbooks. [DV03; DV08]

I have already introduced basic notation. I return to it here for context.

3.1 Univariate temporal point processes

A temporal point process is a stochastic model for the random placement of points in time. The $N(t)$ function that I discussed in the context of video view counter is the obvious example. If $N(t)$ counts the number of views of a video, and it increments by one every time someone finishes watching a video then we may associate with each video such a counting function. I call each increment of the counting function an occurrences. [1] When I consider the generating process to be a stochastic model I will refer to specific time series as *realizations* generated by that model.

I may equivalently represent the information in that function by the list of occurrence times $0 = t_1, t_2, ..., t_N =: t_{1:N}$, taken to be ordered.

We can see the equivalence by examining $N : \mathbb{R} \mapsto \mathbb{Z}^+$ such that $N(t) \equiv \sum_{i=1}^{N} \mathbb{I}_{\{t_i<t\}}$. Here we will only be considering *simple* processes, which means that for all event indices i, $\Pr(t_i = t_{i+!}) = 0$ - so the increments of the series have size one almost surely.

The Poisson process is the stochastic process whose inter-occurrence times are identically and independently distributed such that $t_i - t_{i-1} \sim \text{Exp}(1/\lambda)$. By

[1] Referred to in the literature also as *events*, which is ambiguous, or *arrivals*, which is an awkward term to describe video views, or *epochs*, which sounds like it has something to do with the extinctions of dinosaurs. Occurrences, as used in seismology, seems the least confusing to me.

convention, we set all $t_i \geq 0$ and thence $N(0) = 0$ a.s. our counting representation. It is easy to show that $N(t) \sim \text{Pois}(\lambda t)$.

It is a standard result, that increments of such process have a Poisson distribution. For $t_j \geq t_i$:

$$N(t_j) - N(t_i) \sim \text{Pois}\left((t_j - t_i)\lambda\right)$$

3.2 Conditional intensity processes

Note also the standard result that

$$\lambda := \lim_{h \to 0} \frac{\text{E}\left(N(t, t+h) - N(t)\right)}{h} \tag{3.1}$$

This suggests that we could generalize the Poisson process to have a variable rate by choosing λ to be a function of time, or even a stochastic process itself. This is indeed possible. In the former case, we have an inhomogeneous Poisson process, and in the latter, a *doubly stochastic* or *Cox* process, where we might condition the event upon some σ algebra, $\mathcal{S}$. We call $\lambda(t|\mathcal{S})$ the *conditional intensity process*.

The Hawkes process is a particular case of the doubly stochastic Poisson process: it is a linear self-exciting process. Its conditional intensity process has a particular form which depends upon the previous values of the process itself. Specifically, given the left-continuous filtration $\mathcal{F}^-_{(-\infty,t)}$ generated by the occurrences of $\{N(s)\}_{s<t}$, its conditional intensity is given, up to an additive constant background rate μ, by the convolution of the path of the process with an interaction kernel ϕ and an "branching ratio" η. The interaction kernel ϕ is taken to be a probability density absolutely dominated by the Lebesgue measure - i.e. it has no atoms. To keep the process causal, we require that the interaction kernel has only positive support. $\phi(t) = 0 \quad \forall t < 0$

$$\begin{aligned}
\lambda(t|\mathcal{F}^-_{(-\infty,t)}) &= \mu + \eta \phi_\theta * N && (3.2)\\
&= \mu + \eta \int_{-\infty}^{\infty} \phi_\theta(t-s) dN(s) && (3.3)\\
&= \mu + \eta \int_{-\infty}^{t} \phi_\theta(t-s) dN(s) && (3.4)\\
&= \mu + \eta \sum_{t_i < t} \phi_\theta(t - t_i) && (3.5)
\end{aligned}$$

(Since we only deal with left-continuous filtrations in temporal point process, I will suppress the "-" superscript henceforth.)

The interpretation here is that each occurrence increases the probability of another occurrence in the near future, or, equivalently, momentarily increases the rate of new occurrences. There are several equivalent ways of thinking about this.

One is the formulation in terms of rate - and this is the basis of the goodness of fit test used for this model. [Oga88]

Another is as a branching process, [HO74] much like the Galton-Watson model of reproduction. In this case the branching ratio η is the expected number of offspring that any given occurrence will have. The offspring may in turn have, on average, η offspring of their own, and so on.

In this branching formulation, μ is the "immigration rate", and reflect the rate of new arrivals to our population of occurrences. The system approaches a stationary distribution if $\mu > 0$ and $1 > \eta \geq 0$. [Haw71]

The importance of this model in the current context is that these models gives us the possibility that observed occurrence in a point process is *exogenously* generated - it is an immigrant, or *endogenously* generated - it was the offspring of a previous occurrence. For Youtube, we could think of Youtube views driven by advertising, or views driven by the recommendations of your peers.

The key parameter in this sense is the branching ratio. Using the usual branching process generating function arguments, one can show that the expected number of occurrences due to a single immigrant is $1/(1-\eta)$. As $\eta \to 1$ the proportion of occurrences attributed to the endogenous dynamics of the system increases rapidly, until, when it passes criticality such that $\eta > 1$ the system diverges to infinity with positive probability.

Where we consider realizations on the half line, meaning with no events before time 0, we usually take by convention $t_0 = 0$ Then we have

$$\lambda(t|\mathcal{F}_{(-\infty,t)}) = \lambda(t|\mathcal{F}_{[0,t)})$$

and we abbreviate the whole thing to $\lambda(t|\mathcal{F}_t)$, or even $\lambda^*(t)$.

This is an *instantaneous* intensity. The greater this intensity at a given moment, the more likely another occurrence in the immediate future.

$$\lambda^*(t) = \lim_{h \to 0} \frac{E\left(N(t, t+h) - N(t)|\mathcal{F}_t\right)}{h}$$

Inspecting the definition of intensity for the process the Hawkes process, this means that, as we had hoped, the more occurrences we've had recently, the more we are likely to have soon.

3.3 Kernels

I have left the kernel unspecified up to now. Apart from demanding "causality", normalization, and continuity, we have in principle the freedom to choose here, and even to non parametrically estimate an interaction kernel. [Moh+11; BDM12; HBB13; BM14a]

There are some classic and widely-supported favorites, however, and I restrict myself to these here as they are the ones that my software supports.

3.3.1 Exponential kernel

The simplest kernel, and the fastest to calculate, is the exponential.

$$\phi(t) := \frac{e^{-t/\kappa}}{\kappa}$$

Such a kernel gives the Hawkes process a number of convenient properties, such as a closed-form linear estimator for the process. [DV03] computationally efficient parameter estimation, [Oza79; OA82] and a Markovian representation. [Oak75]

When comparing the effect of this kernel with other kernel shapes we might wish to ensure that they are operating on "comparable" timescales. We can quantify the "time scale" of this kernel in various ways. One choice is the "mean delay", in the sense that if we take interpret the kernel as a probability density for some random variable $X \sim \phi_{\text{Exp}}(\kappa)$, then its expected value is $EX = \kappa$. We could alternatively choose the median, which is given by $\log(2)\kappa$. I ultimately use both.

3.3.2 "Basic" power-law kernel families

The *Omori law* is a widely used kernel, famous for its long history in earthquake modeling. [Uts70].

In the current context, I use the modified Omori law with the following parameterization, recycling κ as a kernel parameter to economize on limited supplied of greek letters.

$$\phi(t) := \frac{\theta\kappa^{\theta}}{(t+\kappa)^{\kappa+1}}$$

The notable feature of this kernel is that for many parameter values it has a *power law tail* with shape controlled by the θ parameter.

$$\phi(t) \sim \left(t^{-\theta-1}\right), t \gg 0$$

Interpreting the Omori law as an interaction kernel, we can expect long-rage interactions to be comparatively more important than for exponential kernels with the same branching ratio. If we interpret this kernel as a probability density we can see that variables draw from this distribution do not necessarily have finite moments of any order.

The "mean delay" $X \sim \phi_{\text{Omori}}, \theta > 1 \Rightarrow EX = \kappa/(\theta - 1)$. When $\theta \leq 1$ the expectation does not exist. A little calculus reveals that the median point for an Omori-law distributed variable is always defined, and given by $\kappa(2^{1/\theta} - 1)$.

This long range interaction effect, as well as high branching ratio, is implicated in surprisingly variable behavior in various dynamical systems, so we would like to know if our model had this kind of kernel. [DS05; GL08]

3.4 The Hawkes Process in action

Having presented the model I present *why* — I wish to understand how much of the behavior of a time series is generated by endogenous dynamics, and what these dynamics are.

To this end, the branching ratio η of the Hawkes process is a plausible choice to partly quantify this, as it tells me about the criticality and explosive kind of behavior that we can expect, and in particular, the ratio between endogenous and exogenously triggered behavior in the system.

I might also be concerned about the timescale and rate of these dynamics, in which case I will also want to estimate the type and parameters of the influence kernel. ϕ This will tell me, loosely, how *rapidly* these dynamics work, and, to an extent, what kind of evolution we can expect in the system. [DS05].

The background rate, μ seems to be of more instrumental interest. if we truly regard it as an exogenous factor, then it is a "given" whose effects we wish to understand. Nonetheless, we might wish to determine, for example, *why* something went viral, or did not, or identify the effect of a change in background rate. In the next section I consider how we might go about this.

Chapter 4

Estimation of parameters of the homogeneous Hawkes model

Here I discuss estimating parameters of the Hawkes model.

For the current section, this means the techniques for parameter estimation from data where the true generating process is a stationary Hawkes process, as implemented by the `pyhawkes` code which I use for this part. I also cover model selection procedures for this method, i.e. how well we can choose which generating model is the correct one given the data.

Later the vicissitudes of using this estimator for the available data, and the limitations of the methods available.

I begin by discussing the case of estimating the parameters of the homogeneous Hawkes model in the case with *complete* data. The case of "summary" data, where we estimate the model from summary statistics, I will examine shortly.

That is, we are given an interval of length T, taken without loss of generality to be $[0, T]$, and the (monotonic) sequence of occurrence times of the increments of the observed processed on that interval $t_{i1\leq i\leq N} \equiv t_{1:N}$. I say "complete" to denote the common case for time series data; that we have access to the timestamps of every occurrence in the time series realization, $t_{1:N}$.

In this chapter I use θ_i to denote the generic ith component of the model parameter, and $\hat{\theta}_i$ to denote the estimates of it. When I need to be clear, I name components. In general $\theta = (\mu, \eta, \kappa)$ for background rate, branching ratio η and some kernel parameters κ depending upon the kernel under consideration. With the Omori kernel I require an additional kernel parameter, and I occasionally recycle θ when it is unambiguous from context.

4.1 Estimating parameters from occurrence timestamps

Parameter estimation for the Hawkes process models is framed as a Maximum Likelihood (ML) estimation problem. It is not clear that this minimizes prediction error in any norm; for prediction with these models one often uses nonparametric smoothing instead. [Wer+10; HW14] or weighted model averaging [Ger+05]. There exist various methods for estimating parameters via second order statistics. [Bac+12; BM14b; SS11; AS09] There are also Bayesian estimators — see The usual offline methods [Ras13] online sequential Monte Carlo. [MJM13; SB03]

For now, the classic ML method is my starting point. This is the most widely used technique, or at least most widely cited technique, [Oza79; Oga78] I suppose for now that we are interested in estimating the "true" parameters θ of the hypothesized generating model, or as close to that as we can get in some sense, and that this true generating model is a Hawkes process. We assume that we have a realization $t_{1:N}$ of all timestamps from a Hawkes model over an interval $[0, T]$

We consider the hypothesized joint probability density $f_\theta(t_{1:N})$ of that model, here call it the likelihood, and choose the values for the parameter θ which maximize the value of the joint likelihood for the *observed* data $t_{1:N}$. Practically, we maximize the log likelihood given the data $L_\theta(t_{1:N}) := \log f_\theta(t_{1:N})$. I will derive the formula this informally.

$$\hat{\theta}_\pi(t_{1:N}) = \operatorname{argmax}_\theta L_\theta(t_{1:N})$$

Using the regular point process representation of the probability density of the occurrences, we have the following joint log likelihood for all the occurrences, [1] [Oza79]

$$L_\theta(t_{1:N}) := -\int_0^T \lambda_\theta^*(t)dt + \int_0^T \log \lambda_\theta^*(t)dN_t \tag{4.1}$$

$$= -\int_0^T \lambda_\theta^*(t)dt + \sum_{t_j \le t_i} \log \lambda_\theta^*(t_j) \tag{4.2}$$

I recall the intensity for the Hawkes process (Equation 3.4)

$$\lambda^*(t) = \mu + \int_{-\infty}^t \eta\phi(t-s)dNs \tag{4.3}$$

[1] The full log likelihood on $[0, T]$, *pace* Ozaki, includes a final extra term to denote the contribution to likelihood by stipulating that no occurrences were in (t_n, T), i.e. The likelihood of N points observed on $(0, T]$, is the joint density of the occurrences $\{t_1 \dots t_N\}, t_i < T$ and no occurrences on $(t_n, T]$. It is tedious to write this down here. However one can show that it is equivalent to the likelihood function of the extended series N' with an occurrence at time T, such that $N'(t) := (N(t) \wedge N(T)) + \mathbb{I}_{t>T}$. For the duration of these estimation theory chapters, when I refer to $N(\cdot)$ on an interval $(0, T]$, I will really mean $N'(\cdot)$. The difference is in any case small for my data sets, and the increase in clarity is significant.

where $\phi_\kappa(t)$ is the influence kernel with parameter κ, η the branching ratio, and the star $\lambda^*(t)$ is shorthand for $\lambda^*(t|\mathcal{F}_t)$.

Now it only remains to maximize this

$$\hat{\theta}(t_{1:N}) = \text{argmax}_\theta \, L_\theta(t_{1:N})$$

There is no general closed-form representation for the location of the extrema, but they are simple to solve by numerical optimization.

While our observations are by no means independently or identically distributed, this is still recognizably a ML estimator of the sort common in i.i.d. parameter estimation. Indeed, the same sort of large-sample asymptotic theory as $T \to \infty$ for this kind of estimator does apply, given the assumptions of stationarity and certain other technical requirements. [Oga78] Note, however that one does not usually use a large *sample* theory for these estimators, in the sense of collecting many time series and trying to estimate shared parameters for all of them.

There are various disadvantages with this estimator. Naïve maximization can become trapped in local extrema, [OA82] or fail to converge over parameter ranges where the shallow likelihood gradient is dominated by numerical error, [VS08] under mis-specification, [FS13] or timestamp randomization. [HB14] Techniques such as *Expectation Maximization* or logarithmic transformation of the parameters are sometimes used to improve convergence. [VS08]

In addition to the above-identified problems, the normal criticisms of ML estimation can be made - e.g. lack of small sample guarantees, lack of guarantees regarding prediction error under various loss functions and so on.

If we assume, for example, an incorrect shape for the kernel then estimates for the other parameters may become poor. Various researchers have devised non-parametric or semi-parametric estimates of the kernel shape in order to avoid this problem. [BDM12; HBB13; FS13; HB14] Rather than implementing such techniques, I will restrict myself to the "classic" kernel shapes, the exponential and the Omori law which I introduced in the previous chapter, as these two will cause me enough trouble as it is.

Set against these is the practical advantage of being simple to estimate, and the usual benefits of the Maximum Likelihood estimation - specifically, a comprehensive asymptotic theory, and model selection procedures based upon it. The simplicity in particular will turn out to be useful for the current problem, so with those caveats I move on.

4.2 Estimation from summary statistics

Recalling the data, this estimator is useless without some further development. It tells us how to estimate parameter values from a time series realization compris-

ing a sequence of occurrence timestamps $\{t_i\}_{1<i\leq N(T)}$. The data here is, in contrast, a sequence of observation tuples $\{(\tau_j, N(\tau_i))\}_{1<j\leq n}$ for some $n \ll N(T)$ and, under any parameterisation of the Hawkes model, $\forall k$, $\forall j$, $\mathrm{P}(\tau_k = t_j) = 0$.

I take the domain of each realization to be $\tau_1 = 0, \tau_n = T$. (If necessary I translate the observation timestamps to ensure this)

The problem of estimating the process parameters from such summary statistics is an unusual one. There is much work on estimating the model parameters from *censored* data especially in seismology literature [BT05; SCV10] where some proportion of the timestamp data is missing due to some presumed censoring process. However, censored data is a different problem than *summarized* data, with far less work done upon it. [A+08; Vac11]

It is the latter problem that we face here. There are *no* occurrence timestamps available at all, and thus we need to invent some. I will write $\hat{t}_i$ for the $i$the invented occurrence timestamp.

To be a true ML estimator, we would have to choose all estimates $\hat{t}_i$ such that they maximized the likelihood of the model given the data. This would entail choosing them differently depending on, for example, kernel parameters. Conducting such a maximization turns out to be computationally intractable even for tiny numbers of points, however, and some time series have millions, so we need an alternative scheme.

To be plausible, the restrictions are that:

1. by the assumptions of the model, increments of the process occur simultaneously with probability zero, so we cannot place multiple occurrences at the one location;
2. We must place all the occurrences, and only the occurrences, that belong to each interval in that interval, so that $\tau_j \leq \hat{t}_i < \tau_{j+1}$, $\forall N(\tau_j) \leq i < N(\tau_{j+1})$.

Apart from that, there is no *a priori* reason to prefer any particular scheme. We could distribute them uniformly at random, or place them equidistantly, or in a converging sequence etc. I choose uniformly at random. Placing the points uniformly at random upon each interval corresponds to a piecewise constant Poisson rate conditional upon the correct number of occurrences in that time interval. Thus, the approximation that I introduced earlier for plotting purposes becomes the actual estimate of the instantaneous intensity of the process, and I interpolate the occurrences from the observations according to this estimate. See Figure 4.1.

The questions are then: Is this process statistically valid? Can it be improved?

Certainly, applying the Maximum Likelihood estimation software to arbitrarily interpolated data like this trivially does not produce a Maximum Likelihood esti-

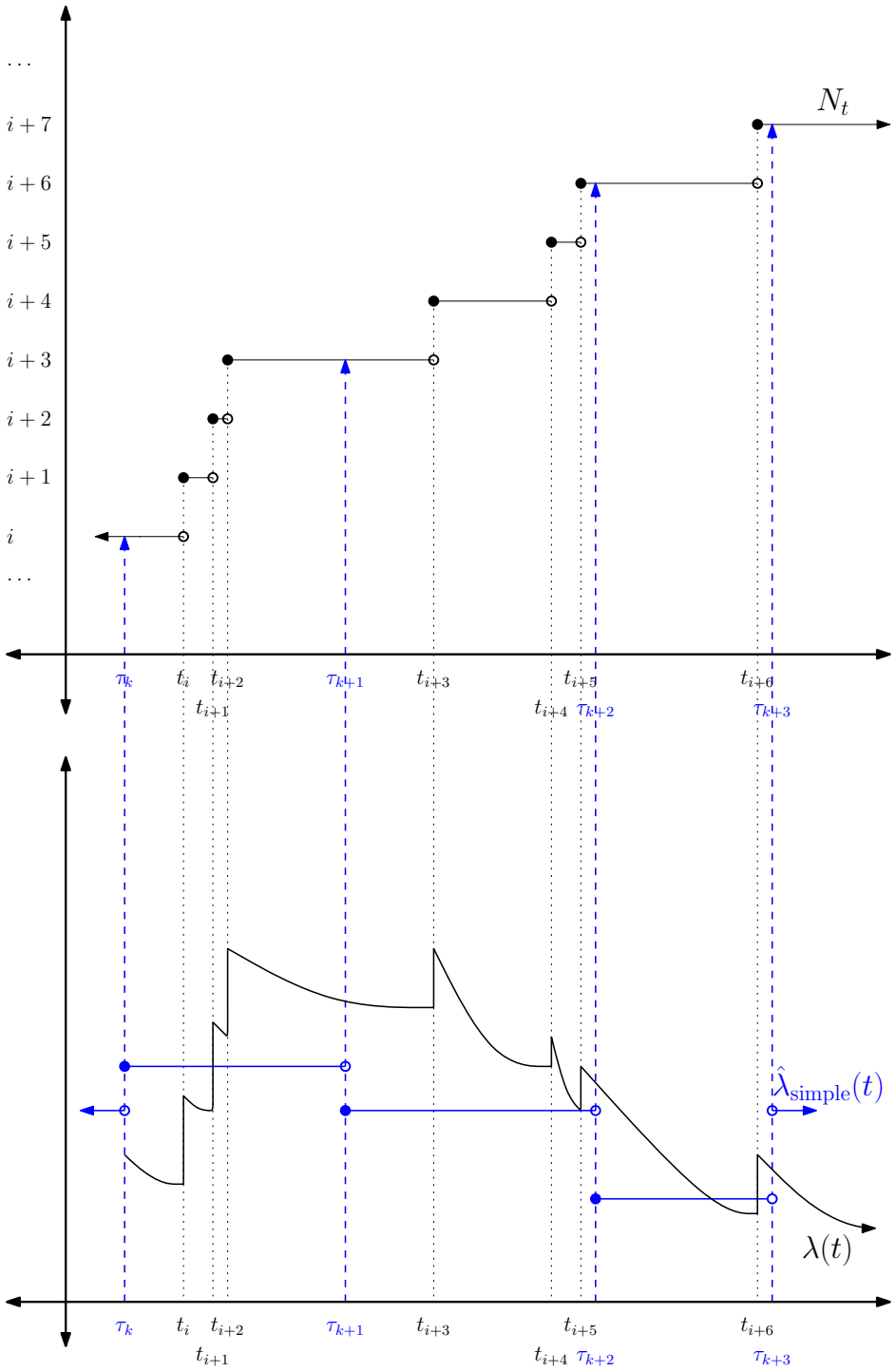

Figure 4.1: Estimating of the hypothetical unobserved true intensity $\lambda(t)$ function by a simple function $\hat{\lambda}_{\text{simple}}(t)$

mator. Maximizing likelihood over all the unknown parameters would additionally include maximizing over the unknown occurrence times. Rather, we guess those other unknown parameters and do not attempt to optimize with respect to them.

Working out how well this procedure approximates an actual ML estimate analytically is not trivial. I will largely ignore this issue here, because that for this particular research group it is a higher priority to see how far we can get with the approximation than to spend much time analyzing the approximation analytically. If we have promising results, then we can attempt to improve the estimator, or to correct for its sampling distribution using a bootstrap procedure. I therefore use simulation to reassure us that the idea is not outright crazy, and that we are plausibly approaching the ML estimator, and move on. I do suggest some ways that the problem might be addressed at the end of the chapter.

4.3 Model selection

4.3.1 The Akaike Information Criterion

Here I discuss the classic model selection procedure for this class of models, the Akaike Information criterion, or AIC. [Aka73; Cla08]

AIC model selection is a well-known procedure for Hawkes-type models. [Oga88] In the literature it is vastly more common than, for example, the Likelihood Ratio identification test of Rubin, [Rub72] although for the special case of nested models they are equivalent. [BA04]

The AIC formula, for a model M fit to a given data set X, for estimated parameter vector $\theta^{\hat{M}}$ with log likelihood L and degrees of freedom d^M

$$\mathrm{AIC}(X, M) = 2d^M - 2L^M(X, \hat{\theta}^M)$$

The degrees of freedom are usually equivalent to the length of the parameter vector θ, although this depends upon the model and fitting procedure. [Efr86]

Given two ML fitted models, $(M_1, \hat{\theta}_1^M)$ and $(M_2\, \hat{\theta}_2^M)$, the difference in their AIC values is an estimate of the relative Kullback-Leibler divergence of the inferred measures $\hat{\mu_1}, \hat{\mu_2}$ from the unknown true distribution, μ i.e.

$$\mathrm{AIC}(X, M_1) - \mathrm{AIC}(X, M_2) \simeq D_{\mathrm{KL}}(\mu \| \hat{\mu_1}) - D_{\mathrm{KL}}(\mu \| \hat{\mu_2})$$

It suffices for the current purposes that it is an information-theoretically-motivated measure of *relative* goodness of model fit to a given dataset. The lower the relative AIC of a model for a given dataset, the more we prefer it. The decision procedure using the AIC is to rank all candidate models fit to a given data set by AIC value, and to select choose the one with the lowest value. Heuristically, we see that a model is "rewarded" for a higher likelihood fit to a given data set, and penalized for the number of parameters it requires to attain this fit.

4.3.2 General difficulties with AIC

There are certain subtleties to the application of this idea.

AIC, as with many ML estimation procedures, is an asymptotic result, relying on the large-sample limiting distribution of the estimators. Just as with the ML procedures in general, it is not robust against outlier observations [Cla08] §2 and we might prefer a robust estimator if the data set has been contaminated by data not easily explained within the model.

Additionally, unlike many test statistics, there is not necessarily a known sampling distribution of AIC difference between two models, even asymptotically. The difference in AIC between two *nested* models approaches a chi^2 distribution under fairly weak assumptions and it becomes an ordinary likelihood test. [Cla08] In the case of non-nested models, we have to estimate the statistics's distribution by simulation or analytic qualities of the particular models.

The derivation of the AIC does not presume that the true generating model is in the candidate set, and so we may use to find the "least worst" in such cases. We could, for example, have several candidate models for a point process, find that they are all rejected by some significance test at the 0.05 level, and the AIC will still give us a "winner" from among this set of rejected models. The "winner" in the Kullback-Leibler metric may of course not give us particularly good performance under other metrics.

More generally Akaike Information Criteria estimates may converge weakly under model misspecification for some models. [KK96] and so our model selection procedure may be faulty. One may introduce model-misspecification guarantees using the more general Takeuchi Information Criterion. [KK96; Cla08] A more commonly preferred solution is simply to expand the candidate set.

Of these difficulties, the problem of model mis-specification will be the more urgent in the current phase. Problems of small-sample corrections I will ignore at this point, but when I add more parameters to the model in the second part of this report, that issue too becomes pressing - see *Model selection*.

Finally, we also need to recall that although I use an ML-*based* estimation procedure, due to the interpolation I am are not really doing true ML estimates from the data, but rather, hoping that my estimates approach the true ML estimates. I know of no results that extend the AIC to this case. Once again I will use simulation to see if this procedure is at least plausible, but we do need to bear this shaky theoretical premise in mind.

4.4 Experimental hypotheses

4.4.1 Model selection with large data sets

The classic setup for use of the AIC is to propose a candidate set of parametric models of the data, with varying numbers of parameters, and then use the AIC to select between them based on the particular tradeoff of goodness-of-fit.

For this Youtube data, for example, we might construct the following set of candidates:

1. Each time series N_v is generated by a Poisson process with rate λ ($d = 1$)
2. Each time series N_v is generated by a renewal process with inter-occurrence times $\{X_i\}_v$ for some common 2-parameter interval distribution, say $X_i \sim \text{Pareto}(\alpha, \beta)$. ($d = 2$)
3. Each time series N_v is generated by a Hawkes process with exponential kernel, background rate μ, branching ratio η, and kernel parameters κ. ($d = 3$)
4. Each time series N_v is generated by a Hawkes process with exponential kernel, background rate μ_v, branching ratio η_v, and kernel parameters κ, where $\mu_v \sim \text{Pareto}(\mu_m in, \alpha_\mu)$ and $\eta_v \sim \text{Beta}(\alpha_\eta, \beta_\eta)$.
5. ...

The more data we have, the more complex a hypothesis we can support. We can include regression here e.g. that the branching ratio is predicted by time of day of upload, [HG10] or that parameters follow a simple trend etc.

We might also ignore some dimensions if consider some of the parameters to be *nuisance* parameters; i.e. we do not care about the distribution of μ_v, but we might suspect that κ has a universal value parameter, [GL08; CSS10] or a limited number of possible values. [CS08].

With the AIC method, the complexity of the hypothesis we can support increases, in general, with the available amount of data. It follows that with this stupendously large data set would support stupendously complex hypotheses; We are faced with, in principle, a combinatorial explosion of possible hypotheses and all of them are computationally expensive to evaluate - and practically, very few of them are supported by the available software.

We can avoid that issue for now since, I argue, we need to infer models that can handle the variation within one series adequately before testing composite hypothesis that couple various parameters together.

4.4.2 Multiple testing considerations

A frequent alternative used, for example, on financial time series, is to give up attempt to finding a small number of universal parameters, and estimate parameters independently on each time series. Then we report the estimated values.

This has issues of its own. If I wish to report, for example, the confidence intervals for 10^6 separate estimates fitted to 10^6 time series, then I am likely to find *something* by sheer force of numbers; This is the multiple testing problem. Moreover, if I am relying on bootstrap estimator variance estimates I face potentially as many bootstrap estimates as I have point estimates. The question of how to construct and report confidence sets or hypothesis tests in this case is a complex and active area of research. [BH95; BY05; Abr+06; MMB09; WR09; GW08; Ben10; MB10; GL11; NG13; Mei14; Gee+14].

While not discounting the importance of these concerns, I argue that there are other methodological questions about the estimator that need to be addressed before I can approach a confidence set for a single times series, let alone multiple ones, and so I set this issue aside.

4.4.3 Goodness-of-fit tests

Traditionally residual analysis is used to diagnose goodness of fit of the Hawkes process parameters using a time change of the process into a uniform unit-rate Poisson process. [Oga88] I do *not* test residuals in this project, since I am aware of no known test that calculates residuals for the summary statistics used here.

Informally, the residual test for point process intensity estimates examine whether the process "looks like" a unit rate Poisson process when scaled, according to estimated intensity, to have unit rate. Since my estimation procedure here involves arbitrary interpolation of the series, we do not have residuals in the per-occurrence sense assumed by Ogata.

Our residual fit must be a last defense against bad model, and therefore if *nothing else*, must be a statistic with some basic guarantees against Type I error. There is no sense going through the motions of applying such model diagnostics, if they can provide, at worst, false confidence, and at best, no information.

In any case, I will provide ample alternative sources of evidence that the fitting procedure is problematic without requiring the goodness of fit test.

Chapter 5

Simulations for the homogenous estimator

5.1 Point estimates

Maximum likelihood estimators for complete times series are well studied. [Cox65; Oga78; Oza79; Sch05] The novel quality of the problem here is the summarization and interpolation step. I estimate the significance of this by simulation.

As mentioned, the observation interval is variable both within and between time series, and there is no known model for the intervals. For want of a better alternative, in my simulations I will use a constant observation interval with each time series. I start with the case of estimating the parameters of a Hawkes model with exponential triggering kernel from data generated by such a process.

I choose 9 different sampling intervals $\{\Delta\tau_i\}_{i=1,2,\ldots,9} = \{2^{-4}, 2^{-3}, \ldots, 2^3, 2^4\}$. I fix the number of observations per realization at $n = 201$, the branching ratio $\eta = 0.9$ and hold the kernel parameters κ fixed. To keep the number of occurrences comparable, I choose $\mu_0 = 5.0$ and $\mu_i = \mu_0/\Delta T_i$. I pick $M = 300$ simulations.

For each $i \in 1, 2, \ldots, 9$, I simulate a realization of a Hawkes process $N_{m,i}(\cdot)$ over the interval $[0, 200\Delta\tau_i]$. I construct maximum likelihood estimate $\hat{\theta}_{\text{complete}}$ for the parameters from the realization. Next, I reduce this realization to summary tuples

$$\{(0, N_{m,i}(0)), (\Delta\tau_i, N_{m,i}(\Delta\tau_i)), (2\Delta\tau_i, N_{m,i}(2\Delta\tau_i)), \ldots, (200\Delta\tau_i, N_{m,i}(200\Delta\tau_i))\}$$

I synthesize "complete data" for these summary tuples with a piecewise constant-rate process to create a new time series $N'_{m,i}(\cdot)$, and estimate the parameters from the new series.

Observe that each realization constructed this way has similar number of occurrences $EN(200\Delta\tau_i)$. Due to edge effects we should expect slightly fewer occurrences when the time step, and hence the whole simulation interval, is shorter.

Each realization has an identical number (201) of observations of those occurrences.

I do not resample a single realization with multiple observation intervals, since that will confound the number of observations with the sample window influence, but rather, I simulate each time with a different seed for the pseudorandom number generator.

Over the 300 simulations I get an estimate of the sampling distribution of the parameter point estimates as the sample step size changes relative to the kernel size. Equivalently I could change the kernel size and leave the step size constant; as there is no other inherent scale in the process, the results are equivalent.

The procedure described thus far corresponds to time series started "from zero", with no history before the start of the data collection, and thus an initially lower rate than the stationary process. An alternative is to sample from the stationary model, Practically, the "stationary" case corresponds to simulating over the interval $[0, C + 200\Delta\tau_i]$ but fitting over the interval $[C, C + 200\Delta\tau_i]$ for some sufficiently large positive C.

Since ultimately the datasets that I use are themselves usually started "from zero", in that the first data point is close to the start of the series the non-stationary case seems more relevant. I display only the "started from zero" data here. That estimation from stationary simulated data generally *worse*, in the sense that the point estimates have larger bias.

I set $\kappa = 1$. This gives a mean delay of 1,

I now plot the inferred sampling distributions against the sampling distribution of the complete data estimator. Across the 300 simulations, I get an average of 8627 occurrences per realization. (See Figure 5.1, Figure 5.2 and Figure 5.3)

The estimates for the branching ratio are somewhat biased, as are those for background rate. It would be interesting to know if this the estimates were *consistent*, with a bias that vanishes in the large sample limit. That question is irrelevant for the current data, which has limited sample sizes.

The estimates for the kernel times scale are the most affected by the resampling process. Compared to the complete data case, almost all estimates have significant bias, and the degree of this bias depends upon the scale of the kernel. In the best case, when the kernel scale is of the same order as the observation interval, it has acquired a mild upward bias by comparison with the complete data estimates. In all other cases the bias is large. *All* re-interpolation, even those whose interval is much smaller than the supposed timescale, introduce additional variance into the estimator. The performance is good enough to use as is for the current experiment. When the sampling window hits 16 the estimate is wrong by a factor of approximately two — but then this estimate is for the parameters of a kernel that is $\frac{1}{16}$ the time scale one might at which one might give up hope of estimating

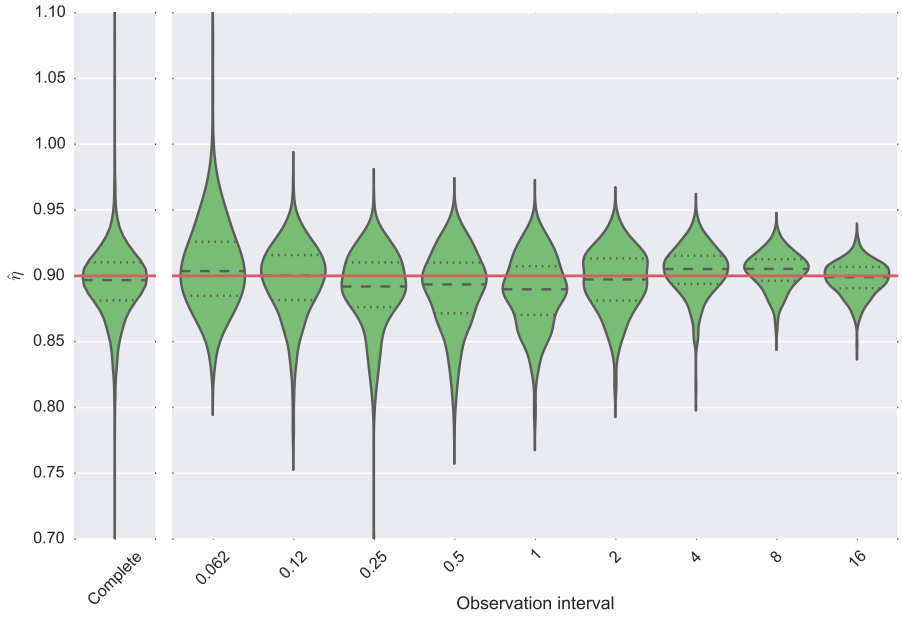

Figure 5.1: Sampling distribution of branching ratio estimates $\hat{\eta}$ under different observation intervals for the Exponential kernel Hawkes process. The true value is marked by the red line.

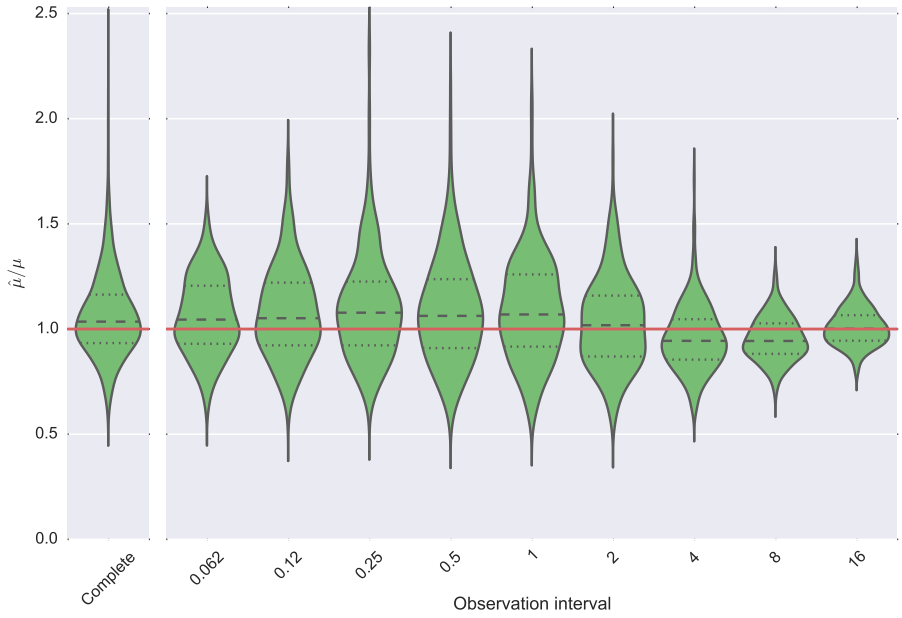

Figure 5.2: Sampling distribution of ratio of background rate estimates to true rate $\hat{\mu}/\mu$ under different observation intervals for the Exponential kernel Hawkes process. The true value is marked by the red line.

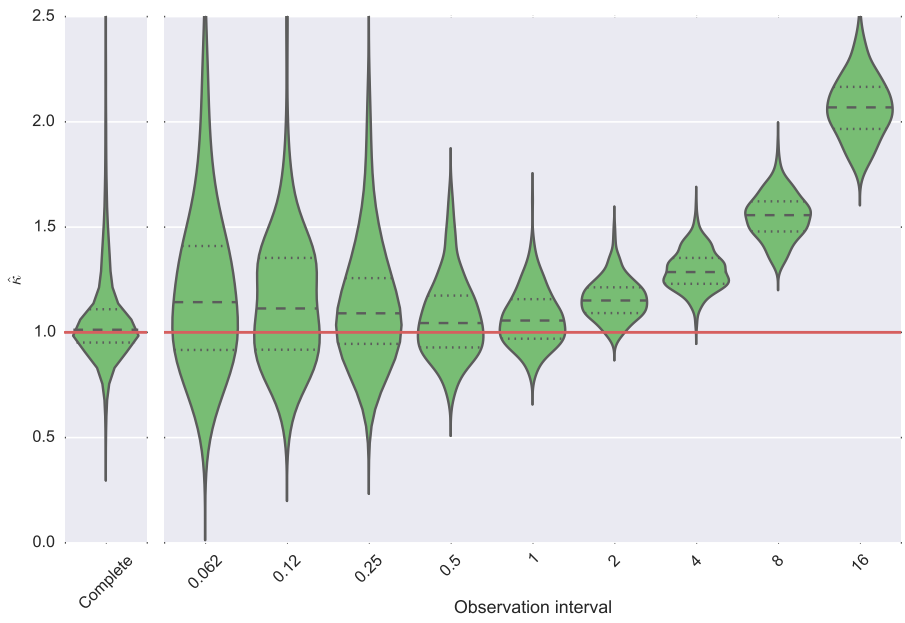

Figure 5.3: Sampling distribution of kernel scale $\hat{\kappa}$ estimates under different observation intervals for the Exponential kernel Hawkes process. The true value is marked by the red line.

the parameters at all. Moreover, the bias is regular enough that one could possibly correct bias for a given parameter estimate by bootstrap simulation: On this range the sample mean point estimate is apparently monotonic, in the sense that, with high frequency, $\hat{\theta}_{\Delta\tau_i} < 2\hat{\theta}_{2\Delta\tau_i}$. I will encounter more pressing problems than correcting this bias however.

These graphs are all of marginal estimator values. Point estimates of components of the parameter vector for any given data set are not independent from the full data estimate in the sense that the Fisher information matrix is not diagonal. [Oga78] Likewise we should not expect the different parameter estimates for fits to the interpolation-based estimator to be independent; For the moment, marginal distributions of the estimates are informative enough.

I now turn to heavy tailed kernel parameter point estimation.

The Omori kernel has two parameters which determine the time scale. Since it is standing in for the whole class of heavy-tailed distributions, it seems wise to test a strongly heavy-tailed parameter set. Accordingly, I choose tail exponent $\theta = 3/2$. Maintaining same mean delay as the exponential kernel requires me to choose the other parameter $\kappa = 1/2$. Note that heavy tailed kernels such as this can lead to estimation uncertainty even with complete data, so problems with this data are likely. [SU09]

Indeed, the variance in the estimates are large. I consider first the branching ratio

estimates and background rates Across the 300 simulations, I get an average of 8823 occurrences per realization. See figures 5.4 and 5.5.

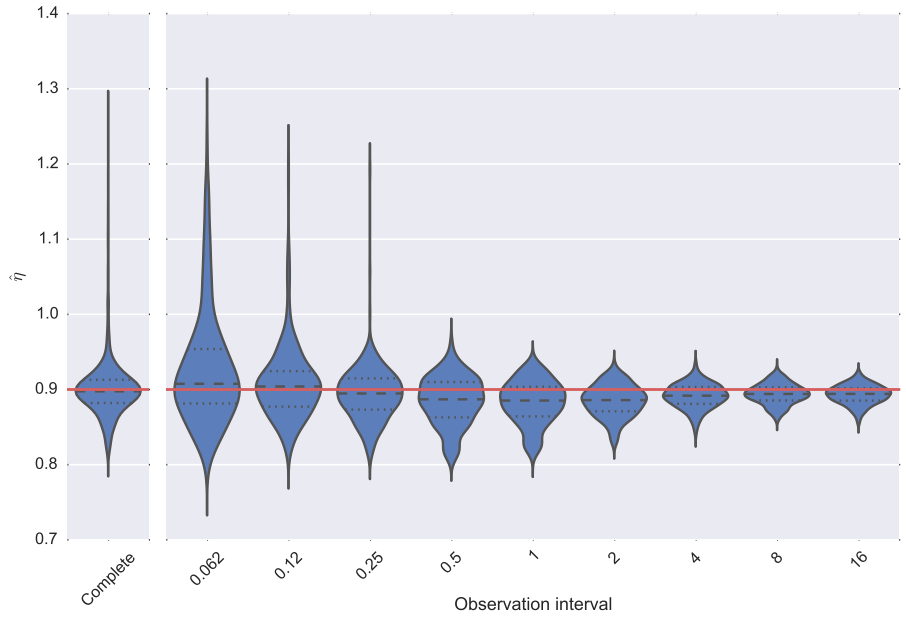

Figure 5.4: Sampling distribution of branching ratio $\hat{\eta}$ estimates under different observation intervals for the Omori kernel Hawkes process. The true value is marked by the red line.

The time scale estimates are bad enough that they present real difficulties in presenting graphically.

Considered individually, the kernel parameter estimates are not consistent, spanning many orders of magnitude. The sampling distribution has more in common with modernist painting than statistical analysis. I show one example in 5.6 although estimates for both parameters are similar.

This problem could be to do with the estimator becoming trapped in local minima, possibly even for pure numerical estimation reasons. Indeed, to save CPU cycles, I restarted the numerical optimizations with only a small number of initial points when estimating parameter values.

I posit that the problem with the estimator is that is getting the *shape* of the kernel wrong, but that the estimates might still be correct in terms of the time scale when the kernel parameters are considered together.

I need some care to plot this, since the Omori law does not necessarily have finite moments of any order, and indeed the estimates often give me a kernel with no finite moments. It seems that mean delay was not a wise choice. I use the classic trick with heavy-tailed distribution and consider the *median* as a measure of cen-

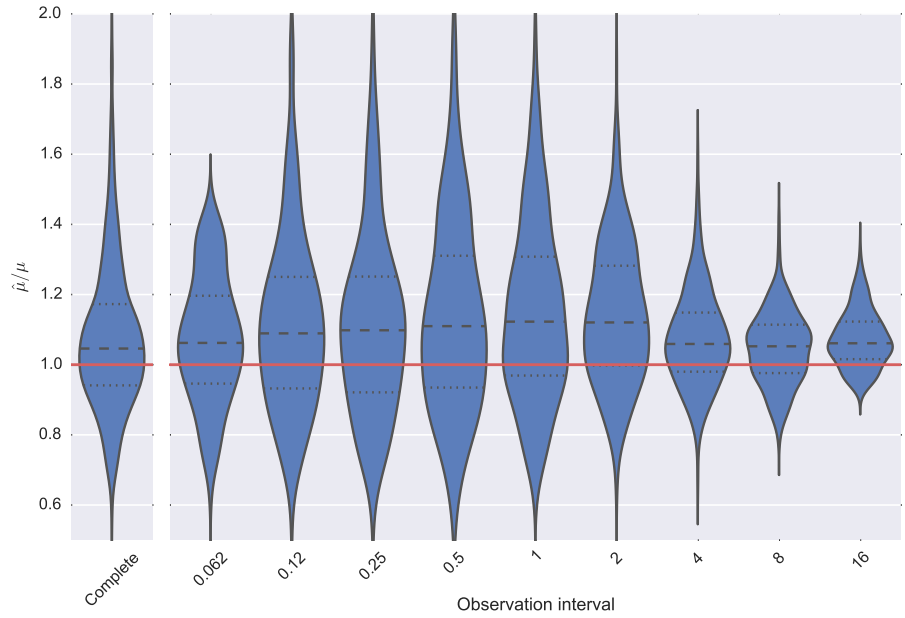

Figure 5.5: Sampling distribution of ratio of background rate estimates to true rate $\hat{\mu}/\mu$ under different observation intervals for the Omori kernel Hawkes process. The true value is marked by the red line.

.

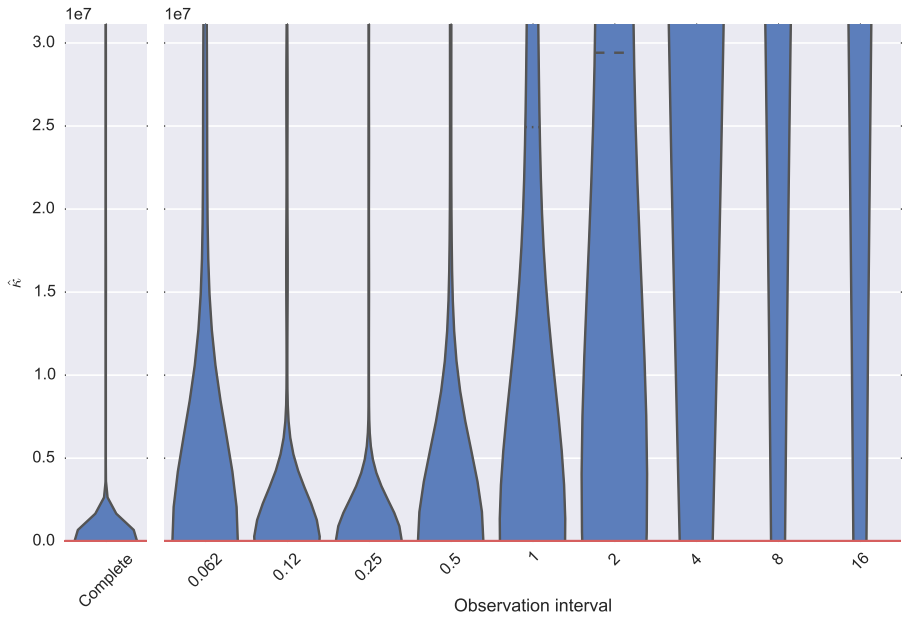

Figure 5.6: Sampling distribution of kernel parameter estimates under different observation intervals for the Omori kernel Hawkes process. The true value is marked by the red line. Note vertical scale.

tral tendency. I use the plug-in estimator of the median given the estimated parameters. Plotting the sampling distribution of this reveals some support for this idea, showing me a somewhat similar picture to the exponential kernel case. (Figure 5.7).

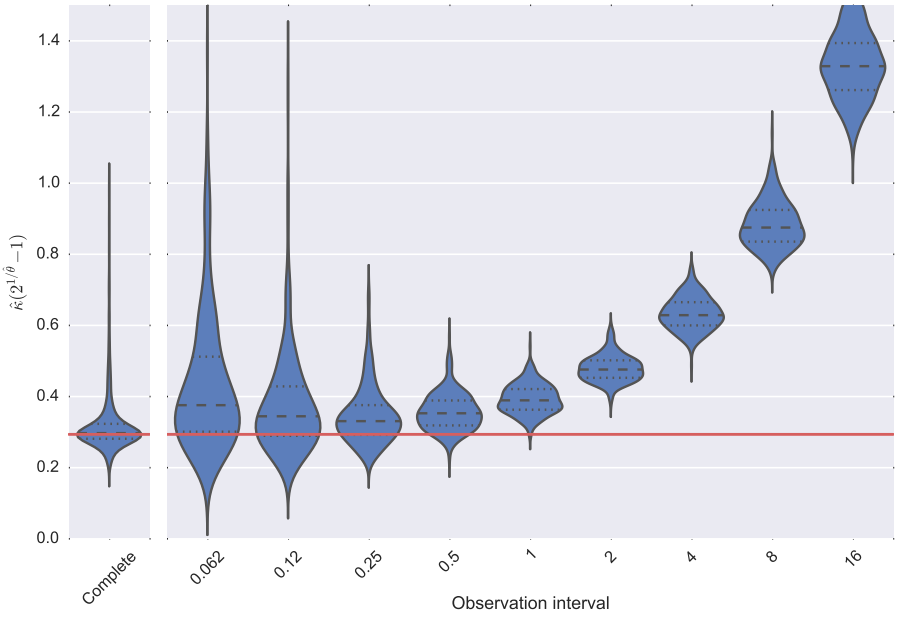

Figure 5.7: Sampling distribution of kernel median delay estimates under different observation intervals for the Omori kernel Hawkes process. The true value is marked by the red line.

Indeed, the Omori kernel has, in an approximate sense informally speaking, "more information"" destroyed by the randomization process than the Exponential kernel. The parameters change not just the average scale of interaction, but the relative weighting of local and long-range interaction.

If we suspect heavy-tailed interaction kernels are important for this kind of data, a different heavy-tailed kernel family might resolve this problem. for now, I will observe that we should be suspicious of the inferred shapes for Omori kernel fits. IN particular, our estimate of heaviness of the tail distribution, which is our motivation for using this kernel, is suspect.

So much for point estimates. I still need to examine the AIC model selection question.

5.2 Model selection

Here I largely follow the methodology and parameter choices of the previous sections.

At this point, I have 3 different candidate models available: The Hawkes model with exponential kernel, the Hawkes model with Omori kernel, and the constant rate Poisson process corresponding to either of the Hawkes models with a zero branching ratio.

Now, instead of merely fitting the Hawkes process with Omori kernel to data generated by a Hawkes process with Omori kernel, I am concerned with whether I can work out *which* model to fit given only the data.

I simulate and interpolate data as before, but now I fit each of the 3 candidate models to the same interpolated data set and compare the AIC statistic for each fit. The goal is to identify the correct model class with high probability.

Even with 3 models there are many permutations here. I will demonstrate only a couple. The primary difference with the previous section is that I will not show the statistic distribution with complete data for technical reasons [1]

First I consider whether I select the Hawkes process with exponential kernel when the true model is in fact a constant rate Poisson process.

Reassuringly, the proposed procedure usually gets this right at all observation intervals, although there is a significant tail of false acceptances of the Hawkes process. Figure 5.8

In the converse case we also select the correct model, although it should be noted that if we were to consider the *magnitude* of the AIC difference as an indicator of certainty, the larger sampling intervals would give us increasingly spurious confidence. (Figure 5.9)

The case is less clear if we try to identify the correct kernel. Trying to select between Omori and Exponential kernels the AIC difference depends strongly on relationship between kernel and observation interval timescales. (Figure 5.10)

Qualitatively, the AIC model selection usually selects a Hawkes model when the true generating process is a Hawkes process and rejects it for a constant rate Poisson process. When we need to select between the two different kernel types, however, the AIC distribution is messy and timescale dependent, and magnitudes of the difference are generally misleading.

This leaves the interpolation-based estimator in a curious position.

Consider a toy world in which all time series are generated by one of the three models I have identified, and in which we must use the interpolation-based estimator, and select between models using the AIC.

In this world, for at least the parameter ranges I have used here, and setting aside the question of the influence of uneven sampling intervals, I can get a good estimate of the presence or absence of self-exciting dynamics. I can get a reasonable

[1] I accidentally deleted it.

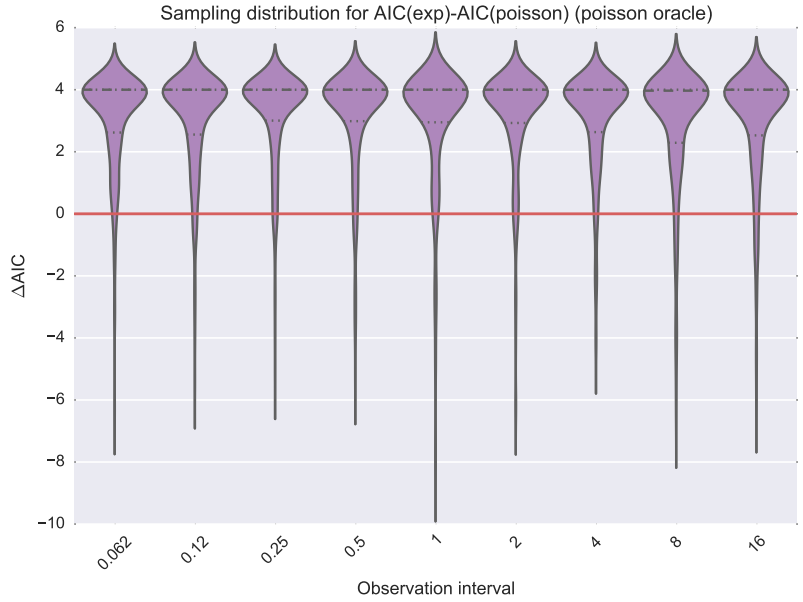

Figure 5.8: Sampling distribution of Δ AIC statistic between estimated Poisson model and Hawkes model (exponential kernel) for data generated by a Poisson process. The line of indifference is marked in red. Positive values select the Poisson model.

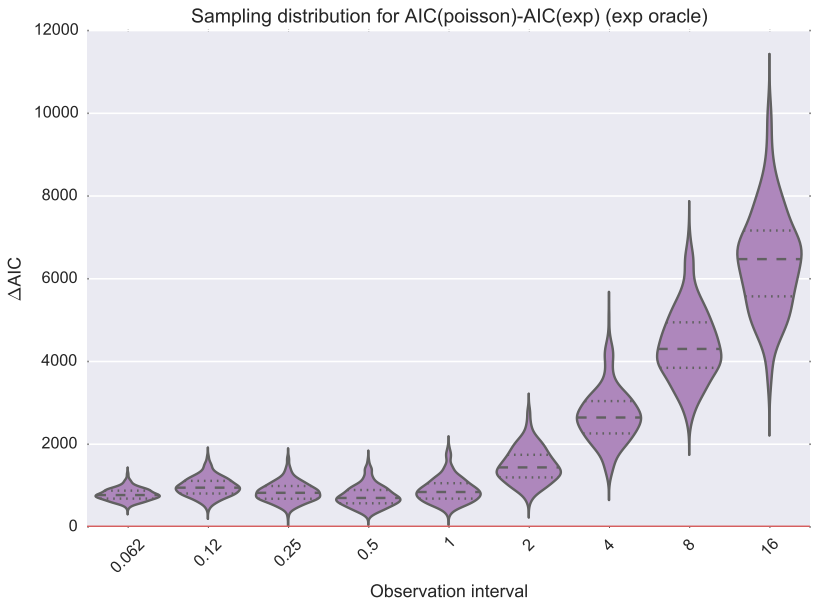

Figure 5.9: Sampling distribution of Δ AIC statistic between estimated Poisson model and Hawkes model (exponential kernel) for data generated by a Hawkes model with exponential kernel. The line of indifference is marked in red. Positive values select the Hawkes process with exponential kernel.

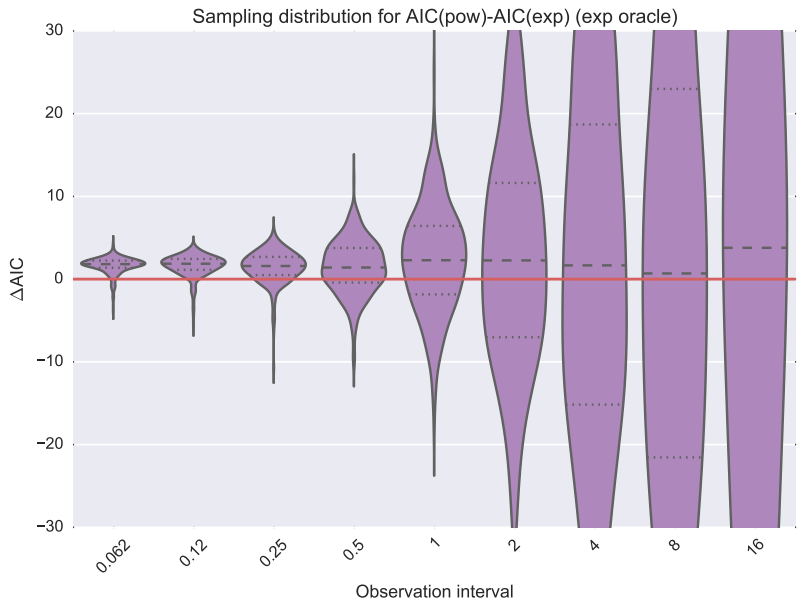

Figure 5.10: Sampling distribution of Δ AIC statistic between estimated Hawkes model with Omori and exponential kernels for data generated by a Hawkes model with exponential kernel. The line of indifference is marked in red. Positive values select the exponential kernel.

estimate of the branching ratio, the background intensity, and even some idea of the characteristic timescale of the process. My ability to estimate specifics of the kernel shape beyond that is virtually non-existent.

This might be acceptable, depending on our purposes. Certainly, in this world we have *some* idea of branching ratio and dynamics for the time series we observe.

I now turn to the question of what happens if we expand the parameters of the toy world to consider the possibility of time series generated by processes from outside this class.

5.3 Empirical validation of estimators of inhomogenous data

I turn to the phenomena of isolated spikes in the data, and consider the behavior that we can expect from the estimators in handling these in articular. We should of course bear in mind that there are many possible inhomogeneities in the data, and many plausible generating mechanisms outside the candidate set. We might nonetheless prefer a restricted candidate model set for easy of interpretation or computational efficiency, so long as the behavior is reasonable despite the mis-specification.

I simulate a stylized "lead balloon" spike. This I define as the inhomogeneous

Poisson process $N(t)$ with the rate function

$$\lambda(t) = \begin{cases} 200 & t \leq 1 \\ 2 & \text{otherwise} \end{cases}$$

I thus have an example of Lead Balloon-type behavior, where the median timestamp should occur somewhere around $t \approx 50$, or 25% of the series total observation window, which is not particularly extreme for this data set. Apart from the single inhomogeneity, the process has zero branching ratio.

Once again I simulate and fit this model 300 times using the estimator. resampling and re-interpolation makes little difference with this piecewise-constant intensity function, so I do not bother variable observation interval sizes and so on, but fit using the complete data estimator.

Using the AIC test, the Poisson model comes last all 300 times. That is, we select a positive branching ratio for some parameters the rest of the time, by a large margin. I picture the evidence, in the form of AIC difference, in favor of the Hawkes models. Since I have been using violin plots so far, I will continue to do that here for consistency, although it should be borne in mind that AIC comparisons are only meaningly with a single dataset, and these comparisons are usually between data sets. Nonetheless we can learn something from the ensemble distribution - for example, that this test never prefers the Poisson model for this kind of data. (Figure 5.11)

The estimation procedure turns out to be reasonably agnostic between the exponential and Omori laws for the kernel shape, much as with the summarized data.

We also get low variance estimates of the series parameter, with large bias. Consider the branching ratio, for example, which is always close to 0.54 for Omori and Exponential kernels. (Figure 5.12) Similarly, the procedure estimates a median timescale with low sampling variance. (Figure 5.13)

These spurious estimators are data-dependent. By choosing, for example, more extreme spikes, I can cause the estimator to pick a higher branching ratio.

However, the real point is not to investigate this particular mis-specified model. Rather, it is to bear in mind that the admirably elegant set set of models that we can fit with `pyhawkes` out of the box is too small to plausibly handle the kind of behavior that the Youtube data suggests, and that all results will be colored by this fact. Nonetheless, I cross my fingers and hope for moment, and turn to the data.

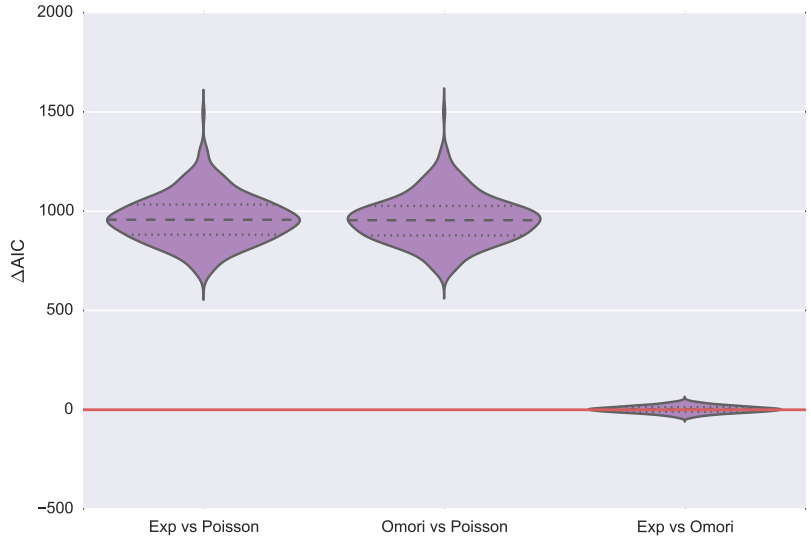

Figure 5.11: Cumulative distribution of Δ AIC values estimated between paris of candidate model for 300 simulated Lead Balloon realizations. Positive values select the *first* named model. The zero line, at which we are indifferent, is marked in red.

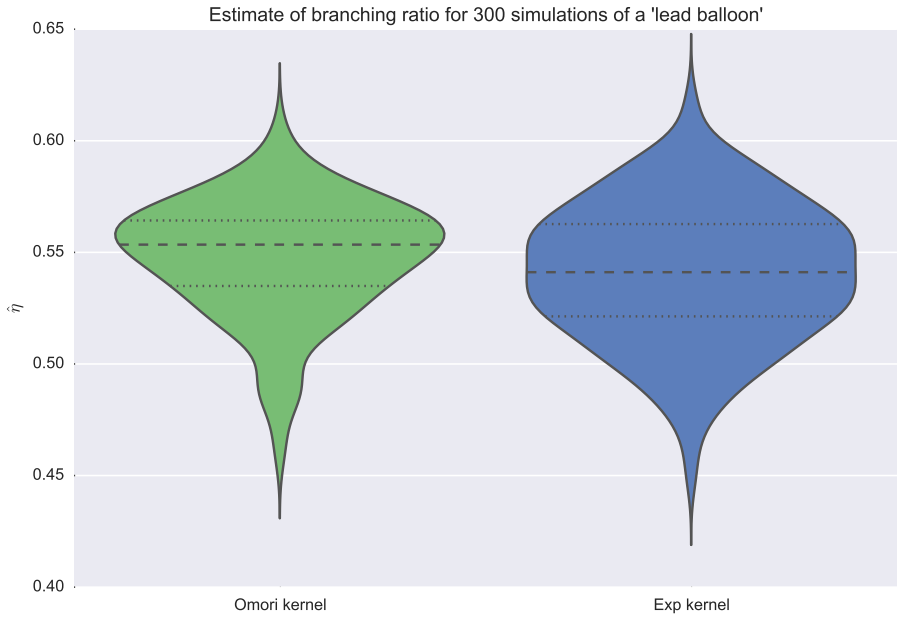

Figure 5.12: Estimated value of branching ratio for 300 simulated Lead Balloon series.

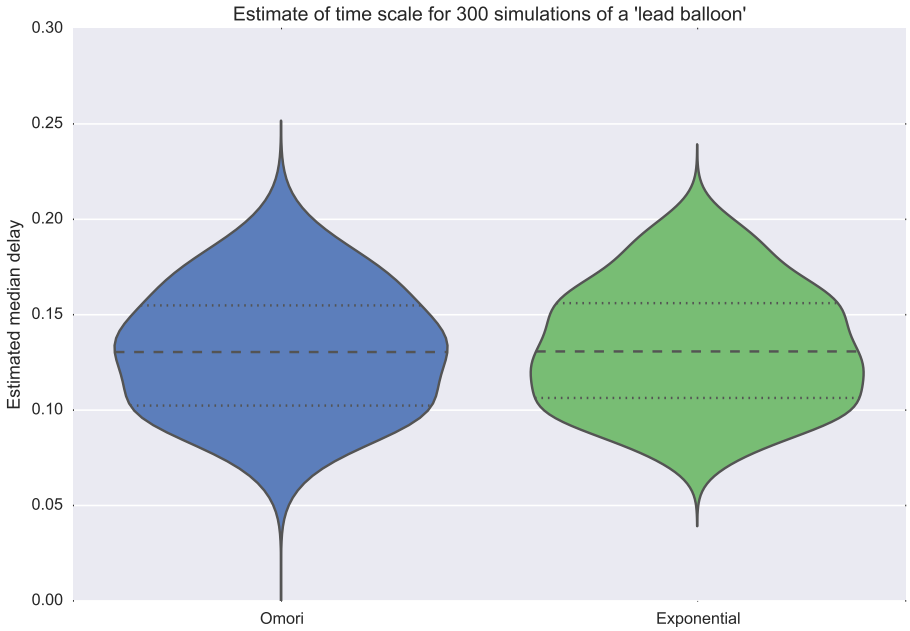

Figure 5.13: Estimated value of median delay for 300 simulated Lead Balloon series.

Chapter 6

Results for the homogeneous Hawkes model

Here I discuss the results applied to various subsets of the time series.

One possible option to deal with the problems identified in the simulation stage would be to manually select some "best" data sets that I believe to be free from inhomogeneity, and fit the estimator to those. This has the unfortunate quality that I have no well-specified notion of what the sampling process of selecting data that "looks like it fits my model" allows me to say about the data more generally. Certainly, finding datasets that "look" endogenous is a trivial procedure, and I have fit some as a diagnostic.

I return to the idea of filtering the data-sets to find ones that are tractable in a principled fashion later, by suggesting that we can simply identify inhomogeneity using the right sort of estimator.

For now, I restrict myself to random sampling. I estimate model parameters from time series chosen uniformly without replacement, from the set of Youtube videos. As discussed earlier, it it not clear if this will give us information about the population of Youtube vides *per se*, but the same criticism could be made of many schemes. I let the computing cluster run through the time series in a random order until the batch jobs are terminated by exceeding their time limits. At the end of the process, there are 92183 time series results.

Using the AIC procedure, I examine the question of which model is selected.

Model selected	Count	%
No self-excitation	7	0.01
Exponential kernel	42259	45.84
Omori kernel	49917	51.15
Total	*92183*	*100.0*

The Poisson distribution is, as expected, rejected apart from a handful of time series of near constant rate. Much as with the misspecified test data, the kernel

choice is reasonably evenly divided between two alternative hypothetical kernel shapes. The data set is too large now for violin plots. However, some histograms convey the idea. I show a raw histogram of estimated results; it is not, for example, weighted by a goodness of fit measure or AIC difference. (Figure 6.1, Figure 6.2) The distribution is qualitatively similar the "lead balloon" fits earlier. We

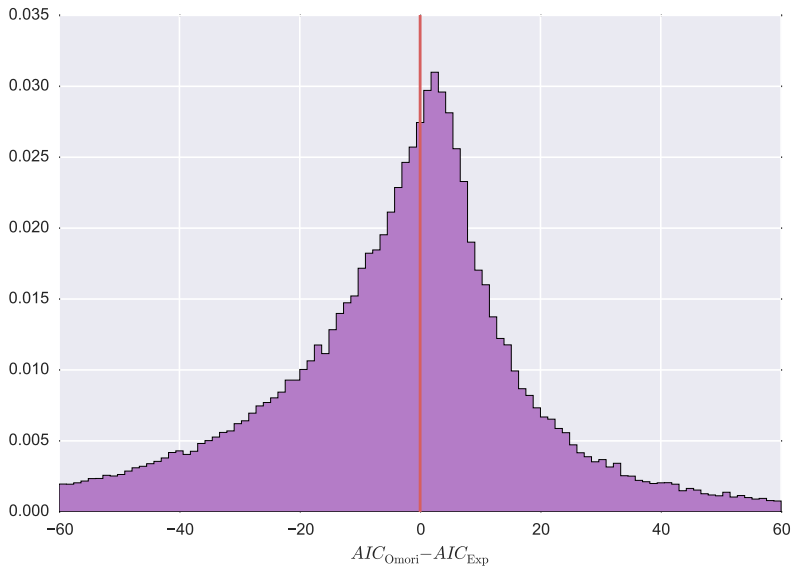

Figure 6.1: Histogram of AIC values for the fit models. Positive numbers select the exponential kernel.

need sharper statistical tools, however, to see if this means anything. Similarly, the estimates would a suggest high branching ratio, although, as discussed, we have excluded the alternative implicitly. (Figure 6.3)

The estimated parameters of the Omori kernels are messy, as with the simulated date. Once again I resort to the plugin kernel median estimate to give us a timescale, and to keep the kernel timescale estimates comparable. The Omori and exponential kernel fits results for the plugin median estimate are given here. The distribution is broad but shows, across both types, a peaks at around 0.1-0.2, corresponding to a median influence decay on the order of 4 hours. This is not an implausible time-scale to estimate from our data. For comparison I plot also the histogram of observation intervals. (Figure 6.4)

Note, however, that *if* we believe these estimates are meaningful, then we need to recall that the interpolation process has introduced upward bias to these values; the "real" time scale is likely even shorter. This effect could be estimated by parametric bootstrap from the estimated parameters.

I might consider how much the estimate might be influenced by the lead bal-

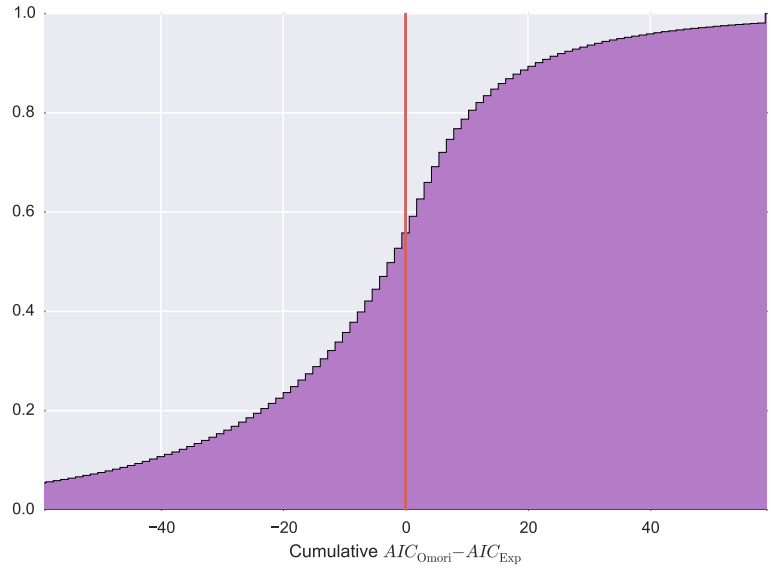

Figure 6.2: Cumulative histogram of AIC values for the fit models. Positive numbers select the exponential kernel.

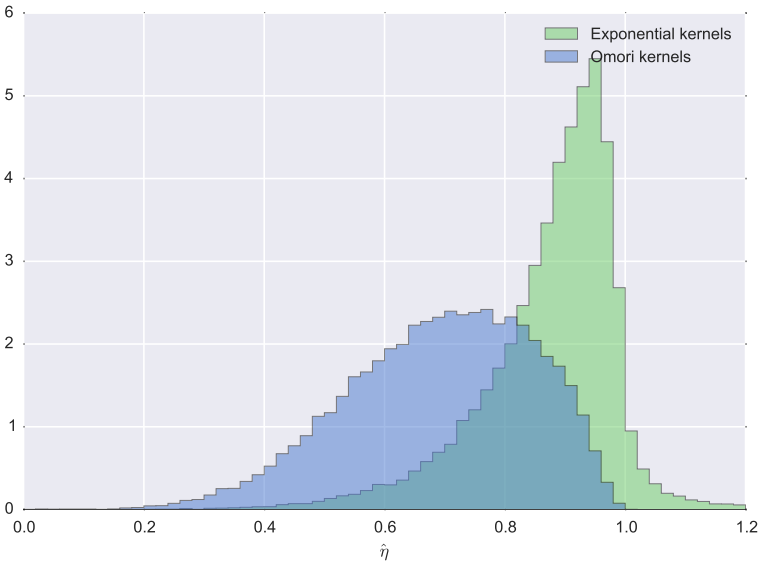

Figure 6.3: Distribution of branching ratio estimates on the Youtube data

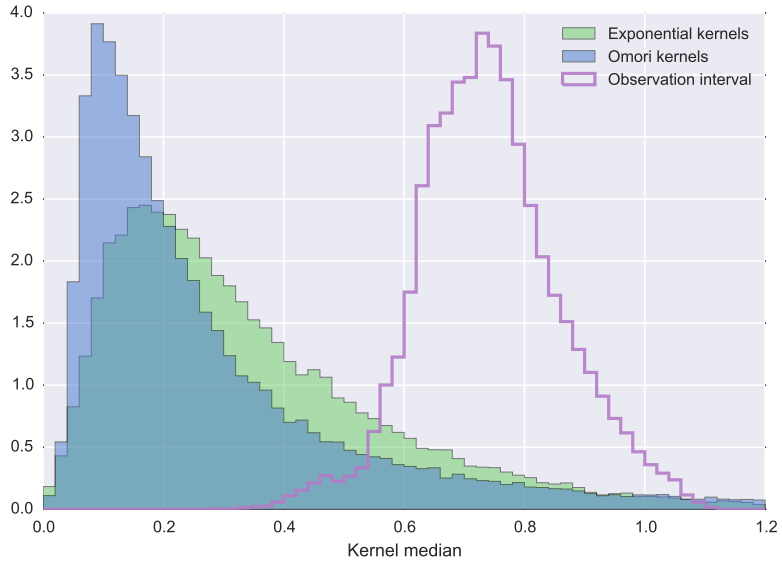

Figure 6.4: Estimated median kernel delay for Omori and Exponential kernel fits to the Youtube data

loons in particular. A simple statistic to quantify this is the estimated median occurrence time in the time series, considered as a fraction of the length. (This is median is the property of a time series as a whole, distinct from the *median interaction time* of the estimated influence *kernel*) If the rate of video views was constant, we would expect this to cluster at the 50% line. If half the views a video ever has were to occur in the first 5% of its sampling window, then it would be at the 5% line. Our videos tend toward the latter type. (Figure 6.5) This prevalence of lead-balloons is itself not uniformly distributed with regard to time series size. Rather, high view-rate time series are disproportionately likely to be lead balloons. (Figure 6.6)

It seems that early success is not necessarily associated with overall success in a simple manner. On one had this shows the interest of the data set -there are clearly non-trivial dynamics in play. On the other hand, these dynamics are ones that we know to be problematic.

We can informally diagnose at least one type of outlier. We see whether the distribution of these estimates is *determined* by the lead-balloon-type outliers, by filtering out all time series whose median sample point occurs *before* the 50% mark. This will restrict the estimates to the 29418 time series that are, by this informal measure, definitely not lead balloons. We are still in exploratory mode here, so I show some histograms to visually inspect the differences in the distributions of estimates. (Figure 6.7, Figure 6.8) The alteration is not spectacular; This particular method of selecting results doesn't get substantially different set

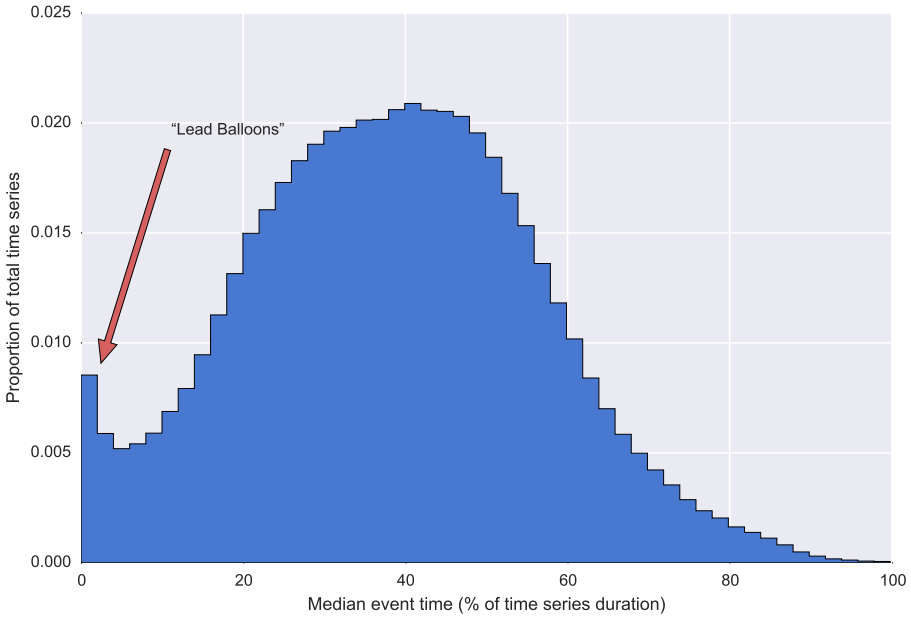

Figure 6.5: Distribution of median occurrence time within each time series, by series

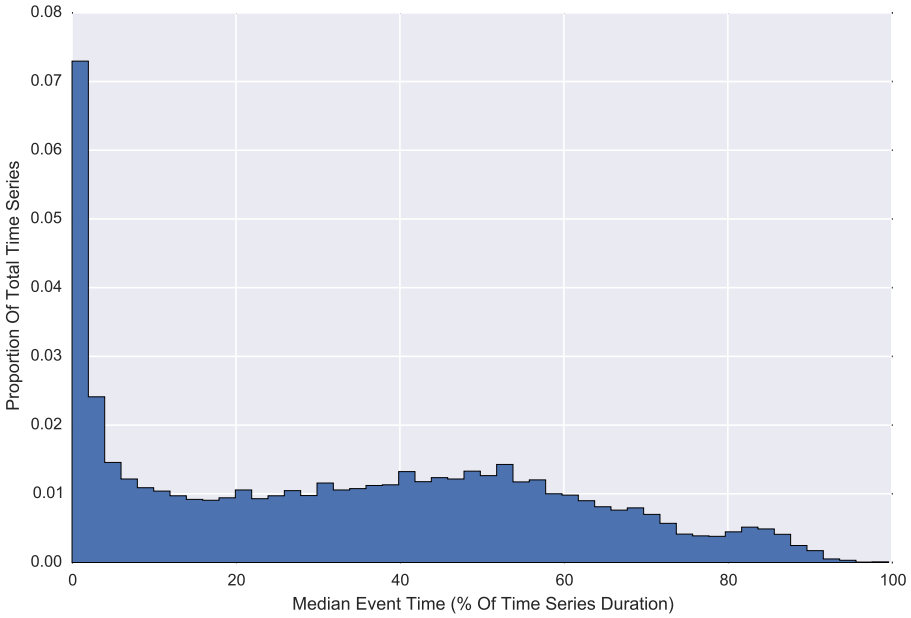

Figure 6.6: Distribution of median occurrence times in the top 5% show an even more extreme skew towards sudden collapse

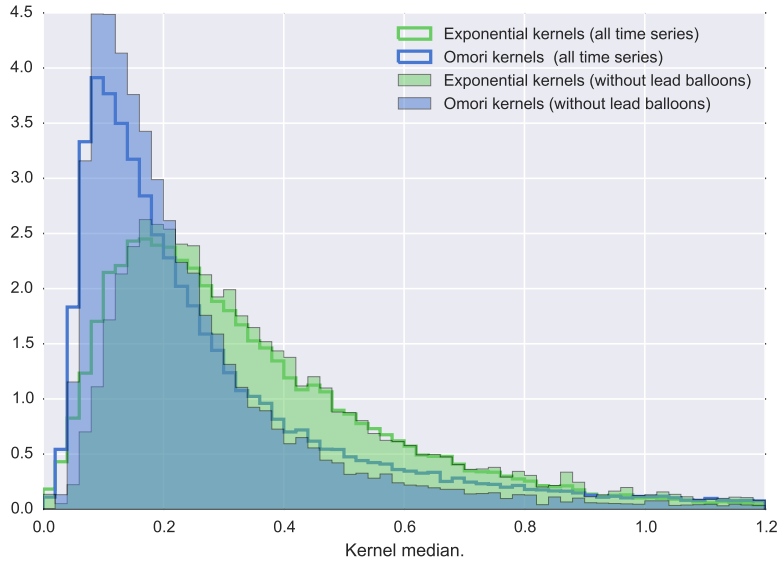

Figure 6.7: Distribution of estimates of kernel median, with and without lead balloons.

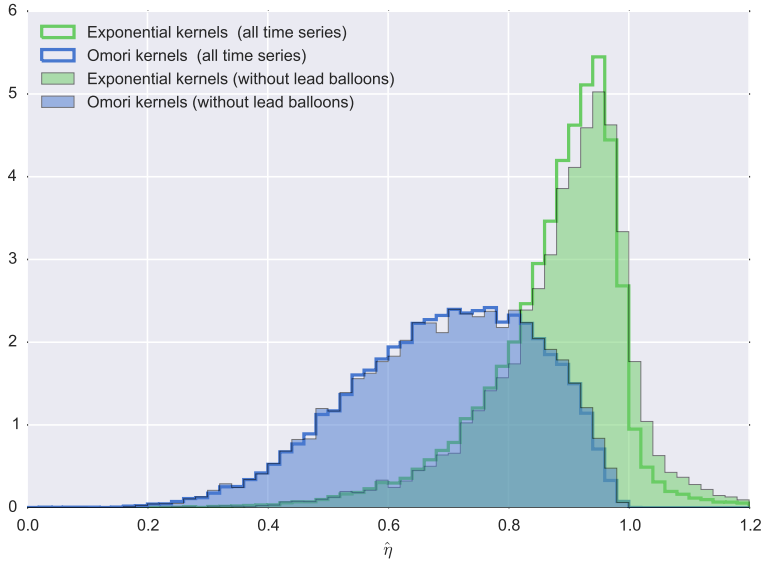

Figure 6.8: Distribution of estimates of kernel median, with and without lead balloons.

of estimates.

If we return to the time series that I showed at the beginning of this report, there are signs that individual estimates are indeed misleading. Consider, for example, the archetypal "lead balloon" that I showed at the beginning, time series of David Navarro's fight, which is interesting for two reasons.

First, I note the estimates for this time series with the exponential kernel, which is selected over the Poisson. We have, to two decimal places $\hat{\mu} = 31.8$, $\hat{\eta} = 0.99$, $\hat{\kappa} = 0.01$. Sure enough, the estimator has determined that this least viral of time series is very nearly critical, and it has detected a time scale of approximately 13 minutes. 13 minutes is so far below the sampling window that it seems implausible to resolve. The extreme criticality estimate, however, shows the estimator is not doing what we'd like. If we believe the model is well specified we can easily bootstrap ourselves some confidence intervals, but even that seems too faith at the moment.

The second reason that it is interesting is that I don't have an Omori kernel estimate for this one. The reason, it turns out, is that the estimation for those kernel parameters, did not terminate in the permitted time. This time series, with 36198 occurrences, is not especially large, but calculation with such time series is slow, and this one was too slow. There is, then a degree of censoring in the data regarding Omori law kernels We can use alternative estimators that approximate the Omori kernel - especially if we think that Omori law kernels are implicate in the extreme behavior of this system On the other hand, since the simulations lead us to believe that we cannot detect precisely the extreme heavy tailed Omori kernels that are of interest here, there does not seem to be immediate benefit to this particular use of CPU cycles.

6.1 Further work

6.1.1 Expanding the candidate set

the case for expanding the class of model we consider in this is clear; It seems likely that even a simple linear trend model for background intensity might be a start, and it has a reasonably simple estimator. [MPW96]

It turns out to be not so easy to do this immediately for basic technical reasons; The software package I use, `pyhawkes`, and indeed, most of the commonly used packages for this estimation procedure, have no support for inference of variable background rate. One can manually simulate a variable background rate by *ad hoc* procedures such as time-transformation of the data.

I did in fact attempt this procedure in an earlier stage of the project, but the problems with goodness of fit and model selection procedures were already severe enough that I decided not to this particular layer of confusion. If I am committed to deriving a new estimator then, I need to choose the one with the highest

practical return on investment, and that would involve a flexible and principled choice of inhomogeneity modeling. I believe the estimators outlines in the next chapter are my best chance in that regard.

6.1.2 Summary data estimation

If we are deeply invested in this data set, or ones with similar missing data problems, we may wish to consider how to reduce the uncertainty due to the data interpolation, since the list of shaky steps in the estimator's construction is clearly too long at the moment in the light of the results of the last chapter.

There are many potential ways to do address this.

Without any change to the current estimator, we could aim to improve confidence intervals by using simulation methods robust agains mis-specification-robust. One can construct confidence intervals, and possibly even bias corrections using the bootstrap; In the case of this model, ideally a nonparametric bootstrap. [Kün89; Lah93; Lah01; Büh02] This is not trivial time-series with long-range dependence, but there is work in the area.

We could try to construct an estimator that addressed the missing data problem analytically. The problem of estimating Hawkes model parameters from summary statistics is not trivial, but the literature suggests potential solutions. I mention some of these here.

1. *The brute force method*:

   Maximize the likelihood with respect to parameters *and* missing time stamps. Then, at least, we are on firmer ground regarding the use of Maximum likelihood estimation theory, since we wil have maximized the likelihood over all the unknowns and not parameters of interest. This idea, while simple in principle, results in an optimization over $N_T + \|\theta\|$ unknowns with a complicated derivative, many constraints and strong coupling between the parameters. This turned out to be computationally prohibitive in my basic implementation.

2. *Stochastic expectation maximization*:

   Expectation Maximization (EM) is a common iterative procedure for estimating "missing-data"-type problems in an ML framework. [DLR77; Wu83] Informally, this estimator alternates between estimating missing data and estimating parameters based on the missing data. While particular form of this estimator does not seem to be any *more* tractable than the brute force method, stochastic variants [CCD95; DLM99; WT90; KL04] allow us to sample from much simpler distributions to approximate missing data, and EM procedures have been used for other problems in Hawkes process inference. [VS08; Hal12] Deriving an estimator for summary Hawkes process

data is address in some recent doctoral research, and the solution ultimately involves an EM-type procedure. [Vac11]

2. *Deconvolution*: Cucala gives a deconvolution kernel estimate of point process intensity, which gives certain robustness guarantees against uncertainties in measurement time, which might possibly be extended to our case. [Cuc08]

3. *Bayesian inference*: There are several attempts to derive Bayesian estimators for the state of self-exciting processes, both offline in Markov Chain Monte Carlo settings [Ras13] and online, as Sequential Monte Carlo [MJM13].

4. *Summary estimator*:

   It seems plausible that an estimator could be constructed that used the observation summary statistics to calculate the full Maximum Likelihood estimate, by calculating likelihood from the summarized observations directly without inferring occurrences. There is no known simple closed form representation for conditional distributions here, but there are certain inequalities. For the exponential kernel higher moments have a simple form [Oak75], and for some other kernels at least a manageable form [BM02; Bac+12]. We could also consider the use of moderate deviation principles and such inequalities to bound estimates from subsampled data. [HSG03; Zhu13]

4. *Indirect Inference*:

   Inference by matching summary statistics between the data and simulations from the hypothesized model is well-established in econometrics. Smith and Gourieroux introduced such methods for continuous-valued time series, [GMR93; Smi93] although there are point process versions. [JT04; CK12] This technique is not purely econometric, having been used in ecology as well. [Ken+05] It is theoretically involved and computationally expensive, but uncontroversial, and comes with its own analogue of the maximum likelihood estimation theory, including various asymptotic results on confidence bounds

Asides from the lack of guarantees, a shortcoming of the constant-rate interpolation estimator is that it has bad computational scaling behavior; while a single Youtube video time series might encompass, say, a hundred thousand views, I still only have a a few hundred sample points statistics to use for inference. And yet evaluating the likelihood function involves simulating a hundred thousand synthetic data points, constructing in turn a kernel density function with a hundred thousand synthetic points, then evaluating that likelihood function at each of the hundred thousand synthetic data points. The computation cost of naïve evaluation the likelihood function scales as $\mathcal{O}(N^2)$ for N occurrences observed at n times, where in general $N \gg n$. Optimizations exist to improve the scaling

properties for for exponential [Oza79] and Omori kernels, [OMK93] although only the first of these is implemented in my software.

Ideally, an estimator whose computational cost scaled with number of observations rather than number of occurrence would be desirable. $\mathcal{O}(n) \lll \mathcal{O}(N)$. I know of no such estimator. Rather, the estimation procedure is slower and has weak guarantees.

Computational efficiency is, like plumbing, something we would rather *work* without paying it any attention. Ss with many large data sets, computational efficiency does become a factor later on in this project when I extend the model.

On potential fix for this problem is exploiting the decomposability of linear simple point processes such as the Hawkes process We may decompose a point process and its intensity function into two simultaneous independent point process. [Rub72] This suggests that we might be able to "downsample" the time series by thinning, construct the estimate on the smaller series, and use those estimates to infer behavior of the larger series. Once again, I leave that for future work.

Regardless, of the method, if we want to handle this data in a principled fashion, and especially if we care about estimating the heavy-tailed kernel types reliably, then pursuing a better estimator for this class of data is essential.

6.1.3 Goodness of fit

Finally, although it was logically clear here that the models fit "badly" in some sense, and thus I didn't need a goodness of fit test, this absence is pressing for future work when we would like a statistical guarantee. Without it, we don't have a general way of diagnosing the shortcoming of the candidate model class, and so expanding the candidate model set is a dangerous proposition.

Chapter 7

Estimating branching effects for inhomogeneous processes

To recap: before us we have a huge data set of variable quality. It seems pregnant with possible conclusions, and yet, the tools we have available to extract them, estimators based upon stationarity assumptions, have limited ability to draw inference from them.

In this section, I extend my toolset in order to solve this problem I relax the assumptions of homogeneity (and also, implicitly, stationarity) that the previous section relied upon.

So far, I have estimated the parameters for an assumed class of models upon lot of data sets, but not managed to reassure myself that I should have confidence in the estimates corresponding to "true" values for a reasonable generating process. The problem is that the estimates' usefulness is restricted by various assumptions whose plausibility is overstretched:

- that I am sampling time series in the large-T limit from a stationary process
- that, therefore, the process has homogeneous parameter values
- that the self excitation dynamics have a certain parametric forms
- that my interpolation from the summary statistics does not unduly bias estimates
- ...and so on.

The question is now if the conclusions of interest can be attained by relaxing some of these assumptions. Does this data set suggest near-critical branching dynamics under more plausible assumptions?

A great deal of the literature on branching-type dynamics, including prior work on this data set, suggests that it is crucial to understand them in terms of isolated "exogenous shocks" external influences upon the system which temporarily

change their behavior. [SH03; Sor+04; DS05; CS08; CSS10; RPL15] Such external shocks can be celebrity endorsements, natural disasters, or news events, etc. The key question is how the model could be extended to account for them. Importance of regime changes in inflating estimates of branching ratio is discussed much in recent work. [FS13; Fil+14; HB14; FWS15] We are invited to consider estimating, for example, the "renormalized" kernel; the mean effect of an imposed exogenous shock upon the system. [HSG03; BM14a]

There are many ways that this can be done. I concern myself specifically with inferring deterministic inhomogeneous time-varying background rates $\mu(t)$. Allowing other parameters, such as kernel parameters to vary is of course possible. [HB14] One can also assume the parameters themselves to be stochastic, then infer the parameters of the underlying distribution. [MSW98; Møl03; OA82; MJM13; GKM11; DZ11] I try one thing at a time, however.

I have not been able to find many examples of explicitly inhomogeneous fits to Hawkes models in the literature, although there is some use of estimates of the parameters on sliding windows of data (e.g [HBB13]). As such, the following work may be novel.

7.1 Semiparametric kernel density estimation

Observe that the Hawkes process is a kind of kernel estimation, in the sense of convolution kernels. It is not qualitatively different in this sense from, for example, kernel density estimators. [Sil82] Using convolution kernels of various types to estimate point process intensity even for non-Hawkes-type processes is well-established area. [Dig85; BD89; Lie11; BD89] Admittedly, the particular case of the Hawkes estimator has unusual features if regarded as a kernel estimation problem.

Firstly, the convolution kernels are causal; that is, the intensity process is predictable with respect to the filtration generated by the occurrence times. Equivalently, the kernel has mass only on the positive half-line. The "classic" kernel-density estimator for example, uses a zero-centered Gaussian density as the kernel.

Secondly, in ML estimation of the Hawkes model parameters, we have an unusual bandwidth selection procedure, based on maximum likelihood estimation of the model's presumed dynamics. Classic kernel density estimation uses different methods, such as rule-of-thumb, or cross-validation procedures. We have, in fact, a parametric estimation problem, whose form happens resemble the non-parametric problems that convolutions kernels are used to solve.

I consider, then, what alternate kernel decompositions are plausible, and in particular the combination of parametric and non parametric estimates. This is precisely the set-up for semi-parametric estimation methods. I argue that there is

a particular semi-parametric method that seems particularly appropriate to this data set.

7.2 The algorithm

Assembling such penalized regression estimators is a job of assembling several different interdependent pieces. Pedagogically it is clearer to introduce each piece as I need it to explain the overall algorithm, and so I do that here.

I consider estimating the parameters of the Hawkes model where the constant background rate μ is replaced with inhomogeneous rate $\mu(t)$. On one hand, this is a sacrifice; in the inhomogeneous case we no longer meet Ogata's sufficient conditions for asymptotic efficiency and normality of the maximum likelihood estimator. [Oga78] On the other hand, there is nothing to lose. The insufficiently general candidate model class was *also* uncertain ground, and even if it were better, the composite interpolated data estimator I am constructing has no such guarantees.. In any case, many point process estimators even with complete data are used without asymptotic efficiency guarantees [Sch05] or unimodal likelihood functions. [FS13]

For technical reasons (see *On the complexity of the simplest possible thing*) I discard the `pyhawkes` library for estimation; I retain it, as a known-good implementation of the ML estimation procedure, for checking the results of my own implementation. (Likelihood estimates agree to within numerical precision for all values, and point estimates usually agree, although my estimator is more prone to numerical instability)

My non-parametric of choice here will be of convolution kernel type. That is, I will introduce an additional convolution kernel ψ, and functions of the form

$$\mu(t) = \mu + \sum_{1 \leq j \leq p} \omega_i \psi_{\nu_j}(t - t_j)$$

for some set of kernel bandwidths $\{\nu_j\}_{1 \leq j \leq p}$, kernel weights $\{\omega_{\nu_j}\}_{1 \leq j \leq p}$, kernel locations $\{t_j\}_{1 \leq j \leq p}$.

There are many kernels available. For reasons of computational efficiency I would like to have one with compact support. For reasons of minimizing errors in my working, I would like to start with the simplest possible option as far as implementation.

By these criteria, the logical choice is the *top hat* kernel, the piecewise-constant function.

$$\psi_\nu(t) := \frac{\mathbb{I}_{0<t\leq\nu}}{\nu}$$

Traditionally the top hat kernel is taken to be centered on 0. I choose it to have positive support because it makes no difference to the working, at this stage but

that, as with the Hawkes influence kernels, it is causal in the right filtration generated by the observation times. In particular, this background rate estimate could be made predictable with respect to the observation process; we could hypothetically do this kind of estimate on an online setting. I stick to the offline setting here, however.

If we would prefer other qualities, such as smoothness, we may of course choose different kernels. Indeed, in an offline setting we can also surrender causality.

Weighted combinations of such functions give me *simple*, i.e. piecewise-constant, functions.

$$\mu(t) = \mu + \sum_{1 \leq j \leq p} \omega_j \frac{\mathbb{I}_{(0,\nu]}(t - t_j)}{\nu}$$

By this point, the parallel should be apparent with the piecewise-constant intensity estimate that I have already used in diagnostic plots;

$$\hat{\lambda}_{\text{simple}}(t) := \sum_{i=2}^{n} \frac{N(\tau_i) - N(\tau_{i-1})}{\tau_i - \tau_{i-1}} \left(\mathbb{I}_{[\tau_{i-1}, \tau_i)}(t) \right)$$

Our interaction kernels are in fact kernel estimates of the conditional intensity function.

This, in turn, suggests a particular form for the nonparametric kernel intensity function, for observations times $\{\tau_j\}_{j \leq n}$

$$\mu_t(t) := \mu + \omega(t)$$

where

$$\omega(t) = \sum_{j=2}^{n} \omega_j \mathbb{I}_{[\tau_{j-1}, \tau_j)}(t)$$

With only summary observations there doesn't seem to be any point in trying to estimate rates on a *finer* scale than this, and so I adopt it as a starting point.

I write $\boldsymbol{\omega}$ for the vector of all ω_j values. Now I take the data vector $\mathbf{t}$ is taken to contain the observation times $\{\tau_j\}$ *and* the occurrence times $\{t_i\}$ although in practice these will be interpolated as before, but we are ignoring that for now. One is free to choose any set of kernel locations - say, one per day or per hour; the $\{\tau_j\}$ values are merely convenient.

I augment the parameter vector to include the kernel weights $\theta' := (\mu, \eta, \kappa, \boldsymbol{\omega})$ The hypothesized generating model now has conditional intensity process

$$\lambda_{\theta'}(t|\mathcal{F}_t) = \mu + \sum_{j=2}^{n} \omega_j \mathbb{I}_{[\tau_{j-1}, \tau_j)}(t) + \eta \sum_{t_i < t} \phi_\kappa(t - t_i)$$

To remain well-defined I require $\forall t,\ \mu(t) > 0 \Rightarrow \forall j,\ \omega_j > -\mu$.

It is not immediately clear what this has gained us, since there are now more parameters to estimate than data points. Additional restrictions are necessary to make the estimation problem well-posed.

One such restriction is the *regularization*, that is, penalization of some parameters. [TG65; HK70] This is often surprisingly effective. Accordingly, I apply an additional penalty to the log likelihood function to penalize particular undesirable sorts of estimates.

For the penalized log likelihood for parameter θ and penalty π, I write

$$L_\pi(\mathbf{t}, \theta') := L_\pi(\mathbf{t}, \theta') - \pi P(\theta')$$

P here is a non-negative functional which penalizes certain values of the parameter vector, and $\pi \geq 0$ is the penalty weight hyperparameter.

As before, the estimate is defined as the maximizer:

$$\hat{\theta}_\pi(\mathbf{t}) = \operatorname{argmax}_\theta L_\pi(\mathbf{t}; \theta)$$

For non-parametric extension to a parametric model, one usually penalizes only the non-parametric extensions to the model, such that they vanish as the penalty term increases. [Gre87]

$$\lim_{\pi \to \infty} \hat{\theta}'_\pi(\mathbf{t}) = \hat{\theta}(\mathbf{t})$$

I follow this practice here, penalizing only the deviations of $\omega(t)$ from the parametric estimate.

Hereafter, I will drop the prime from the augmented parameter vector and simply write θ.

Penalization is more frequently presented in the context of generalized linear regression from i.id. samples, but it also fits within a generalized Maximum Likelihood estimation theory.

Many alternative choices are open at this point We could favor low variation in the background rate by including $\sum |\omega_i - \omega_{i-1}|$ in the penalty. We could penalize values of η, or θ etc. Many other loss functionals are possible. If sparsity were not a priority we could use L_2 norms, or mixtures of norms. We could add additional hyperparameters to weight various penalty functionals differently. This choice will depend on the assumptions on the background rate process. Verifying such assumptions is a whole additional question which I will not address here.

For the kind of "signals" we might expect from the examples that I have shown expect to recover from Youtube data, infrequent spikes, a logical choice penalty would be a sparsifying L_1 penalty on deviations in the background rate. This is the penalty made famous by Lasso regression, compressive sensing and so on. [Tib96; Efr+04; CRT06; Don06] and also applied to point process models [GL05]. Its key feature is favoring estimates of parameter vectors where many entries in that

vector are zero - in other words, it identifies isolated spikes and yet can typically be efficiently numerically solved. Since spikes are what I hope to control for, this sounds like a logical first choice.

$$P((\mu, \eta, \kappa, \boldsymbol{\omega})) := \|\boldsymbol{\omega}_\theta(t)\|_1$$

where $\|\boldsymbol{\omega}_\theta(t)\|_1 = \int_0^T |\omega(t)|dt$ This quality we do not get for "free"; this penalty will introduce bias, and it will still be more computationally expensive than the plain homogenous model. Nonetheless, if the "true" exogenous process is well represented by a sparse $\boldsymbol{\omega}$ vector, this may be what we want. π in this context interpolates between various assumed levels of sparsity.

I am aware of no off-the-shelf penalized estimation software packages for the Hawkes model, and cannot find any suitable published derivations, so I must create my own.

I therefore impose an additional pragmatic restriction: I implement only the exponential kernel for the Hawkes intensity estimation. The exponential kernel is simpler than the Omori, and easier for me to debug, and the indications are anyway that the sparse observation intervals make Omori kernels hard to fit. We may in any case construct general interaction kernels from combinations of exponential kernels, [HBB13] so this is a logical building block to solving the problem of more general kernel types.

With this in mind, I develop the penalized log likelihood function for fixed π. Selection of appropriate penalty π will be left until later.

The derivation of the estimator is an exercise in calculus: Once I have all the derivatives, I can use them to optimize parameters with respect to the log likelihood by gradient ascent (or if you'd prefer, gradient descent for the negative penalized log likelihood loss) Thus, I write down the penalized version of the log likelihood (Equation 4.1)

$$L_\pi(\mathbf{t}; \theta) := -\int_0^T \lambda_\theta(t)dt + \int_0^T \log \lambda_\theta(t) dN_t - \pi\|\boldsymbol{\omega}_\theta(t)\|_1 \tag{7.1}$$

Calculating this likelihood is computationally taxing due to the repeated sums in this likelihood and in the terms of the intensity. (Equation 3.4) Fortunately, we get partial derivatives nearly "free", if we preserve the values of these intermediate sums, so gradient ascent can be done somewhat efficiently.

Following Ozaki, I differentiate with respect to an arbitrary component of the parameter vector θ_x [Oza79]

$$\begin{aligned}
\tfrac{\partial}{\partial\theta_x} L_\pi(\mathbf{t}; \theta) &= -\int_0^T \tfrac{\partial}{\partial\theta_x}\lambda_\theta(t)dt + \int_0^T \tfrac{\partial}{\partial\theta_x} \log \lambda_\theta(t) dN_t - \tfrac{\partial}{\partial\theta_x}\|\boldsymbol{\omega}_\theta(t)\|_1 \\
&= -\int_0^T \tfrac{\partial}{\partial\theta_x}\lambda_\theta(t)dt + \sum_{0\le t_i\le T} \frac{\tfrac{\partial}{\partial\theta_x}\lambda_\theta(t_i)}{\lambda_\theta(t_i)} - \tfrac{\partial}{\partial\theta_x}\|\boldsymbol{\omega}_\theta(t)\|_1
\end{aligned}$$

By construction, $\frac{\partial}{\partial\mu}\|\boldsymbol{\omega}_\theta(t)\|_1 = \frac{\partial}{\partial\eta}\|\boldsymbol{\omega}_\theta(t)\|_1 = \frac{\partial}{\partial\kappa}\|\boldsymbol{\omega}_\theta(t)\|_1 = 0$.

Taking $\theta_x = \mu$,

$$\frac{\partial}{\partial\mu}\lambda_\theta(t|\mathcal{F}_t) = 1$$

I use higher derivatives for μ so that I may optimize this component using a higher order Newton method, since we know that typically the optimal value is particularly slow to converge for this component [VS08] and the values are simple.

$$\begin{aligned}
\frac{\partial}{\partial\mu}L_\pi(\mathbf{t};\theta) &= -T + \sum_{0\le t_i\le T}\frac{1}{\lambda_\theta(t_i)} \\
\frac{\partial^2}{\partial\mu^2}L_\pi(\mathbf{t};\theta) &= \sum_{0\le t_i\le T}\frac{-1}{\lambda_\theta^2(t_i)} \\
\frac{\partial^3}{\partial\mu^3}L_\pi(\mathbf{t};\theta) &= \sum_{0\le t_i\le T}\frac{2}{\lambda_\theta^3(t_i)}
\end{aligned}$$

Now I handle $\theta_x = \eta$,

$$\frac{\partial}{\partial\eta}\lambda_\theta(t|\mathcal{F}_t) = \sum_{t_i<t}\phi_\kappa(t-t_i)$$

so that

$$\frac{\partial}{\partial\eta}L_\pi(\mathbf{t};\theta) = -\int_0^T\sum_{t_i<t}\phi_\kappa(t-t_i)dt + \sum_{0\le t_i\le T}\frac{\sum_{t_k<t_i}\phi_\kappa(t_i-t_k)}{\lambda_\theta(t_i)}$$

Taking $\theta_x = \kappa$, we find

$$\frac{\partial}{\partial\kappa}\lambda_\theta(t|\mathcal{F}_t) = \eta\sum_{t_i<t}\frac{\partial}{\partial\kappa}\phi_\kappa(t-t_i)$$

and so

$$\frac{\partial}{\partial\kappa}L_\pi(\mathbf{t};\theta) = -\eta\int_0^T\sum_{t_i<t}\frac{\partial}{\partial\kappa}\phi_\kappa(t-t_i)dt + \sum_{0\le t_i\le T}\frac{\eta\sum_{t_k<t_i}\frac{\partial}{\partial\kappa}\phi_\kappa(t_i-t_k)}{\lambda_\theta(t_i)}$$

As an implementation detail, I take the *rate* parameterization of the exponential interaction kernel, $\beta = 1/\kappa$, such that, $\phi = \mathbb{I}_{\mathbb{R}^+}(t)\beta e^{-\beta t}$, to make the derivative more compact. It is an easy matter if invert the parameter if you want the more usual parameterization, and I report κ values in the results section, but internally I use β. If we are in fact constructing a well-behaved ML estimator, this kind of smooth invertible transform shouldn't affect the point estimate.

Suppressing the indicator function - we are only evaluating this kernel on its (non-negative) support,

$$\frac{\partial}{\partial\beta}\phi_\beta(t) = e^{-\beta t} - \beta t e^{-\beta t}$$

and

$$\int_0^{t_i} \tfrac{\partial}{\partial\beta}\phi_\beta(t-i)dt = t_i e^{-\beta t_i}$$

giving

$$\tfrac{\partial}{\partial\beta}L_\pi(\mathbf{t};\theta) = -\eta\sum_{t_i<T}\left[(t-t_i)e^{-\beta(t-t_i)}\right]_{t=0\vee t_i}^{T} + \sum_{0\le t_i\le T}\frac{\eta\sum_{t_k<t_i}e^{-\beta(t_i-t_k)}(1-\beta(t_i-t_k)}{\lambda_\theta(t_i)}.$$

One may repeat these steps to produce the Hessian matrix of the parameters of the homogenous model [Oga78; Oza79] but I did not ultimately implement algorithms that made use of those.

Finally, I handle the ω_j values; these are similar to μ part from the un-penalized path.

Taking $\theta_x = \omega_j$, and defining $\Delta\tau_j := \tau_{j-1} - \tau_j$ we find

$$\begin{aligned}
\tfrac{\partial}{\partial\omega_j}\pi\|\boldsymbol{\omega}_\theta(t)\|_1 &= \tfrac{\partial}{\partial\omega_j}\pi\int_{\tau_{j-1}}^{\tau_j}\left|\omega_j\right|dt \\
&= \pi\Delta\tau_j\operatorname{sgn}\omega_j \\
\tfrac{\partial}{\partial\omega_j}\lambda_\theta(t|\mathcal{F}_t) &= \Delta\tau_j\mathbb{I}_{[\tau_{j-1},\tau_j)}(t) \\
\tfrac{\partial}{\partial\omega_j}L_\pi(\mathbf{t};\theta) &= -\Delta\tau_j - \pi\Delta\tau_j\operatorname{sgn}\omega_j + \sum_{\tau_{j-1}\le t_i\le\tau_j}\frac{1}{\lambda_\theta(t_i)}
\end{aligned}$$

Higher partial derivatives are also analogous to the partial derivatives with respect to μ, although of course the penalty introduces a discontinuity at $\omega_j = 0$. This last formula is the key to the gradient ascent algorithm.

Note that the ω_j values are mutually orthogonal, in the sense that $\frac{\partial^2}{\partial\omega_i\omega_j} = 0$ if $i \neq j$. I can treat these components, more or less, as separate univariate components when optimizing, and the Hessian will be sparse off the diagonal.

Note also that although the partial derivative is undefined when $\omega_j = 0$, we can still tell whether to update it in the gradient ascent algorithm by using the *elbow formula*:

$$\left|\sum_{\tau_{j-1}\le t_i\le\tau_j}\frac{1}{\Delta\tau_j\lambda_\theta(t_i)} - 1\right| \le \pi \tag{7.2}$$

that the value of ω_j is "trapped" at 0, in that we know the sign of the partial derivative is different on either side of 0, and we don't need to bother doing any further updates for that parameter.

That completes the set of derivatives that I need to maximize the penalized likelihood for any fixed π, by my choice of numerical optimization method.

This also provides a method for calculating the solution more generally. I now present an approximate forward stage-wise path algorithm.

1. Fit the unpenalized parameters of the model with your choice of numerical optimization method, $\hat{\theta}(\mathbf{t}) = \text{argmax}_\theta L(\mathbf{t};\theta)$ leaving $\boldsymbol{\omega} \equiv 0 : math :.$ Call this estimate $\hat{\theta}_0$. By construction, this is equivalent to taking the penalty weight π large enough that the regularized fit is the same as the non-regularized ML fit.
2. By evaluating the elbow formula (Equation 7.2) for each component ω_j, we can find the smallest value of π such that we would not alter the estimated value for *any* ω if we used it as the penalty parameter. Call this π_0.
3. Now choose a new $\pi_1 = \pi_0 - \Delta$ for Δ small.
4. By construction, $\hat{\theta}_1 := \text{argmax}_\theta L_{\pi_1}(\mathbf{t};\theta) \neq \text{argmax}_\theta L_{\pi_0}(\mathbf{t};\theta)$. Using $\hat{\theta}_0$ as an initial estimate, ascend the penalized log-likelihood gradient until the new maxima is attained. You can use the elbow formula to check which ω_j values are "stuck" at zero without performing updates once all the non-zero parameters have been updated. Any non-stuck parameters are now introduced to the model and their parameters estimated.
5. Choose $\pi_{i+1} = \pi_i - \Delta$. Repeat from step 4.
6. When $\pi_m = 0$ for some m, stop.

The output of this algorithm is the whole regularization path of m different parameters estimates indexed by the hyperparameter π_m Having calculated it, one can choose from the available set by some model selection criteria.

I am deliberately vague about the choice of gradient ascent technique and the choice of step size Δ. These penalized regression problems are typically calculated by using a range of π values and updating the estimates progressively, using the estimate calculated at each state as an approximation to the next stage,as the parameter is varied. For linear regression, for example, there are particularly efficient methods for calculating these regularization paths, and certain guarantees about optimality. [Fri+07; FHT10] Various gradient ascent algorithms are known to perform well for this kind of problem, and there is a significant literature on the details. Frequently simple gradient ascent (resp. descent) algorithms are "good enough". [Sim+11; WL08].

For the Hawkes model I know of no such specific algorithms, and I have not proven that my method will approximate the optimal regularization path. More smaller steps is better, but also more computationally expensive. I could choose number of steps and size adaptively by some procedure. In practice I use a rule-of-thumb logarithmic spacing with a number of steps equal to the number of parameters.

Pragmatically, for performance reasons, my implementation uses a mixed strategy. My algorithm attempts to update marginal estimates for each ω_i parameter more rapidly first via Newton's method [Bat92; Oza79] and uses conjugate gradient descent for the parametric estimates of the Hawkes parameters. If the updates

steps are small this seems to be stable. There are various tuning parameters to such algorithms.

Many variants of pathwise regularization and exist, such as backwards stepwise regularization, versions with random initialization, Least Angle Regression, [Efr+04] selection with integrated cross validation and so on. [FHT10] For this particular model I have found few results giving me analytic guarantees, so I use heuristics and simulation-based verification. I do not declare that this method "best" in any sense, or that it will find the correct global maximum penalized likelihood etc. The method, however, is simple enough to prototype rapidly and, as it turns out, performs well on test data. Simulation results will be presented in the next chapter.

7.3 Model selection

In an exploratory project such as this, I take an informal approach to the complexities of model selection in high dimensional penalized regression. This is a current and active area of research; [Don06; CRT06; Mei07; WR09; MY09; GL11; NG13; ZZ14; Gee+14; Loc+14; BG15] There is, to my knowledge, no published *prêt-à-porter* selection procedure for the particular model family I use here. I therefore proceed heuristically, which is to say, I will adopt rules of thumb from the literature, but observe that if one desires formal quantitative guarantees about confidence intervals, that more work will need to be done. I will assume that we may ignore the question of whether the model is well-specified, and further, I apply use some results derived for the special case of linear regression to the estimation problem here.

Even proceeding informally, we still need to revisit the AIC model selection procedure. I recall the AIC definition, for a model with log likelihood L and degrees of freedom d

$$\mathrm{AIC}(L_{\hat{\theta}}(X)) = 2d - 2\ln(L)$$

This is known to perform well in the large sample limit, where $n \gg d$, but we should be concerned where $d \simeq n$. And that is the case here: We have few data points per time series, and potentially many parameters. The naïve AIC estimate used so far is asymptotically consistent, but negatively biased for small sample sizes. [HT89] Where the number of degrees of freedom in the model is on the same order as the number of parameters, we should expect to see that the use of the AIC formula favors complex models.

To remedy this, I use Sugiura's finite-sample-corrected version, the *AICc*. [Sug78]

$$\mathrm{AICc} := \mathrm{AIC} + \frac{2d(d+1)}{N-d-1}$$

for number of observations $N = |X|$. Whilst Sugiura originally derived this correction for linear regression models, it has been shown in practice to function as

a good model selection statistic for a broader class. [HT89; Cav97] I therefore adopt it as an approximation here.

Other alternatives would include correcting for finite sample bias by deriving an analytic or simulation-based empirical correction e.g. Claeskens and Hjort §6.5 give convenient constructions based on the Hessian of the estimate, which we get "for free" in this estimator. [Cla08] I leave such refinements for a future project.

Taking N as the number of samples in the data set, this leaves the question of the effective degrees of freedom p. For the unpenalized Hawkes process with a single parameter kernel, $p = 3$. [Oga88] For penalized regression the calculation of degrees of freedom can be unintuitive, and can even fail entirely to summarize model complexity. [KR14] However, under certain "niceness" conditions, which I will assume here without proof, there is a remarkably simple solution: the number of non-zero coefficients fit under a particular value of the regularization penalty π in an ℓ_1 regression model provides an unbiased estimate of the effective degrees of freedom of the model and functions as an effective model complexity measure. [Efr04; ZHT07]

Taking these together, we get the following "rule of thumb" model selection statistic:

$$\widehat{\mathrm{AICc}}_\pi(X, \hat{\theta}_\pi) = \frac{2d_\pi N}{N - d_\pi - 1} - 2\ln(L_\pi(X), X))$$

where d_π counts the number of non-zero coefficients estimated under regularization coefficient π.

As with the usual asymptotic AIC, we may choose between models within this regularization path based on this criteria.

$$\pi_{\mathrm{opt}} = \mathrm{argmin}_\pi \, \mathrm{AICc}_\pi$$

Finally, I note that this estimator is both discontinuous and non-monotonic in π by construction, and thus itself could be tricky to optimize. Finally, optimizing $\widehat{\mathrm{AICc}}_\pi$ with respect to a continuum of π values is tedious, so I will evaluate this statistic only at specified finite set of values of $\{\pi_i\}$, and choose models from this subset of all possible values of π. The estimated optimal π value is then,

$$\hat{pi}_{\mathrm{opt}} = \mathrm{argmin}_{\pi \in \{\pi_i\}} \widehat{\mathrm{AICc}}_\pi \tag{7.3}$$

Various alterations to this scheme are possible, such as choosing the regularization parameters over the entire data set, using cross-validation or choosing $\{\pi_i\}$ adaptively during model fit [Ber+11] and so on. A thoroughgoing theoretical treatment of the estimator is, however, inappropriate for this data-driven exploration, and so I move on.

Chapter 8

Simulations for the inhomogeneous estimator

8.1 Empirical validation on simulated data

To summarize the proposed scheme:

My composite estimator combines the interpolated estimator used in the previous chapter with a ℓ_1 penalized estimate of inhomogeneous background rate. The single hyper-parameter, the regularization penalty π is chosen by AIC or AICc, depending which works better. The kernel boundaries are chosen to match the sampling boundaries in the data.

It is time to see if this estimates anything like what we would wish. It turns out that this procedure performs well on simulated data, as I will demonstrate with examples, and with aggregate statistics from simulations.

I begin with a single time series to visualize the kind of output we should expect.

The parameters of the test model are: $\mu_t = 2 + 8\mathbb{I}_{149<t\leq 150}$, $\eta = 0.95$, $\kappa = 1.0$. I simulate this time series over 300 time units, quantize this time series into unit intervals, and fit on randomly interpolated time series based on these intervals.

First, consider the following realization of our generating process, one where it has *nearly* homogeneous parameters. (Figure 8.1)

In this case the homogenous fit is good, returning $\hat{\mu} = 1.945$, $\hat{\eta} = 0.956$, $\hat{\kappa} = 1.010$

In this situation is is not *a priori* clear that we want to bother with inhomogeneous estimators, and indeed that it could be risky to do so, introducing increased estimation variance for no benefit.

Fitting this model using the vanilla AIC indeed results in a significant over-fitting of background rate fluctuations, and also poor matches to the homogeneous pa-

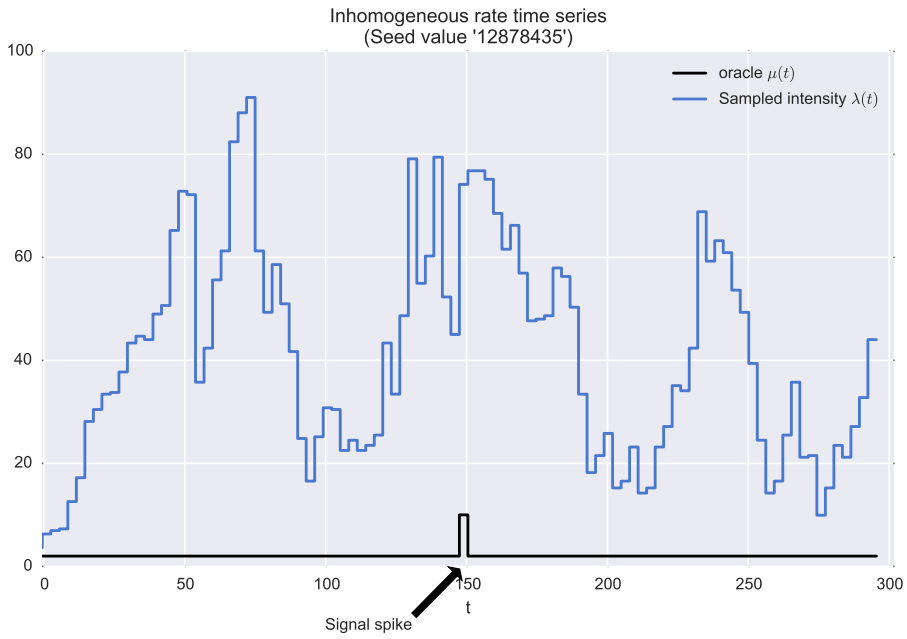

Figure 8.1: μ_t and a realization of the resulting quantized intensity process. $\eta = 0.95$, $\kappa = 1.0$.

rameters, with a concomitant underestimation of the branching ratio. (Figure 8.2)

On the other hand, the finite-sample AICc correction chooses a model which not only recovers the original parameters well, but also estimates the background rate with high accuracy. (Figure 8.3)

Based on this particular graph, and ensemble simulations, it seems that AICc generally performs acceptably in selecting the model, and certainly performs better than the AIC itself. For the remainder of this section I will continue to use it in preference to the AIC.

I reiterate, these point estimates are presented without guarantees or confidence intervals at this stage. It is not immediately clear what the confidence intervals of the complete estimator are, especially in combination with the AIC based model selection.

A single graph is not representative; by changing the seed value I can get results both more and less optimistic than this. In this case, for example, there was a random decline in unconditional intensity in this realization immediately before the spike which intuitively should make the spike "easier" to detect. We will need more comprehensive tests to be persuasive.

I construct a simulation test that looks somewhat like kind of time series I find in the Youtube data set. I choose $\mu_t = 2 + 398\mathbb{I}_{149<t\leq 150}$, $\eta = 0.8$, $\kappa = 1.0$ This

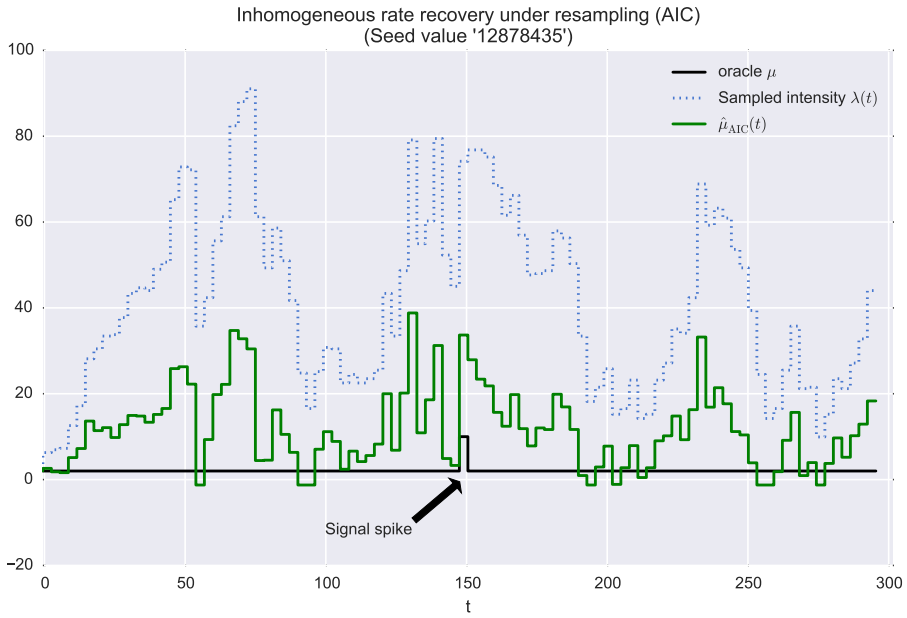

Figure 8.2: μ_t recovery under vanilla AIC. $\eta = 0.95$, $\kappa = 1.0$.$\hat{\eta}_{\mathrm{AIC}} = 0.688$, $\hat{\kappa}_{\mathrm{AIC}} = 3.358$

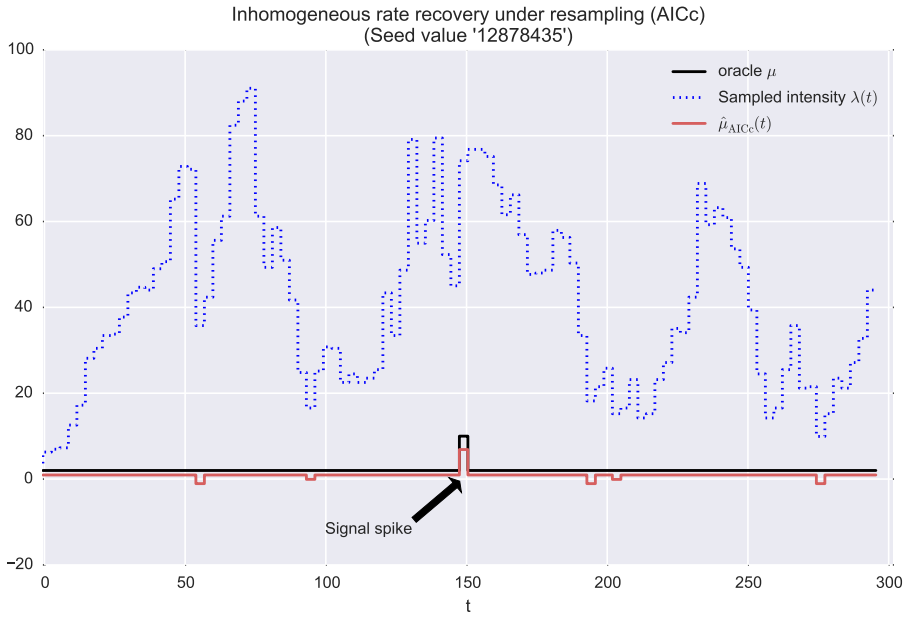

Figure 8.3: μ_t recovery under finite-sample-penalized AICc. $\eta = 0.95$, $\kappa = 1.0$, $\hat{\eta}_{\mathrm{AICc}} = 0.953$, $\hat{\kappa}_{\mathrm{AICc}} = 1.051$

corresponds to $\mu = 2, \omega_{149} = 398$, and $\omega_i = 0 \forall i \neq 149$. I repeat the simulation 100 times, and compare error in estimates. [1]

Performance is variable but generally superior to the inhomogenous test on the same data. (Figure 8.4, Figure 8.5, Figure 8.6)

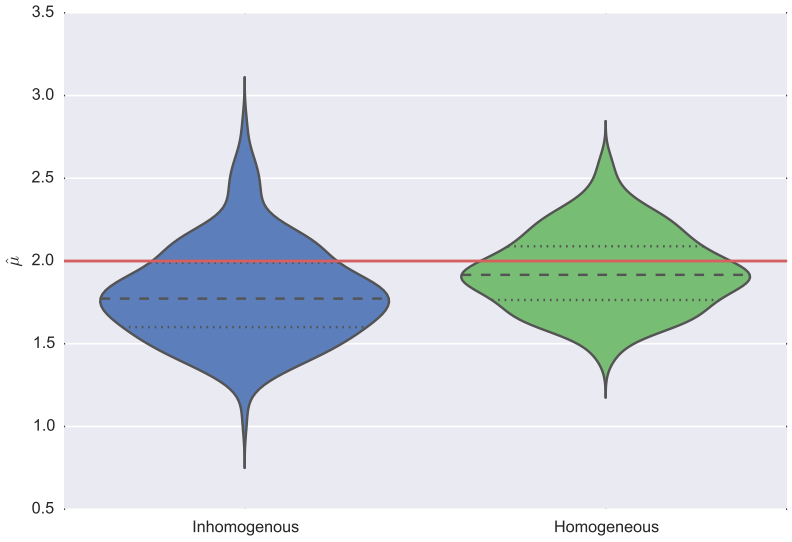

Figure 8.4: μ recovery under finite-sample-penalized AICc.

This inhomogeneous rate estimator still only captures the true value a small fraction of the time, and yet clear it is less biased than the homogenous estimator.

Whether this is what we want is context dependent. We might be especially concerned with "true" values of the parameters of the generating process, or we might be concerned with recovering the "true" background rate. It's hard to plot estimates for whole functions. Instead I will plot the error functional using L_1 norm - specifically $\text{Err} \|\hat{\mu}_t - \mu_t\|_1 / T$. Note that since the homogeneous estimator assumes that $\mu_t \equiv \mu$ but the true generating process is not constant, that the homogeneous estimate cannot attain zero error by this metric. (Figure 8.1)

The last one is most disappointing; it seems that with this estimator the background rate estimates are sometimes worse, when better background rate estimation was a selling point of this estimator.

[1] For this comparison to be fair, I would have used the same parameters as with the "lead balloon" test, i.e. 200 steps, initial spike, and 300 repetitions. The reason for the discrepancy is that the simulations were too CPU intensive to repeat once I had entered the wrong parameters. Fixing this in future work should be trivial, however.

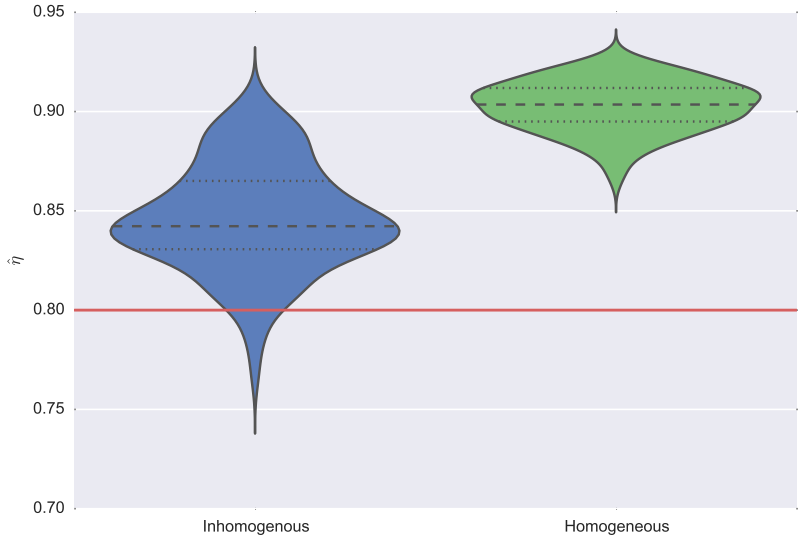

Figure 8.5: η recovery under finite-sample-penalized AICc.

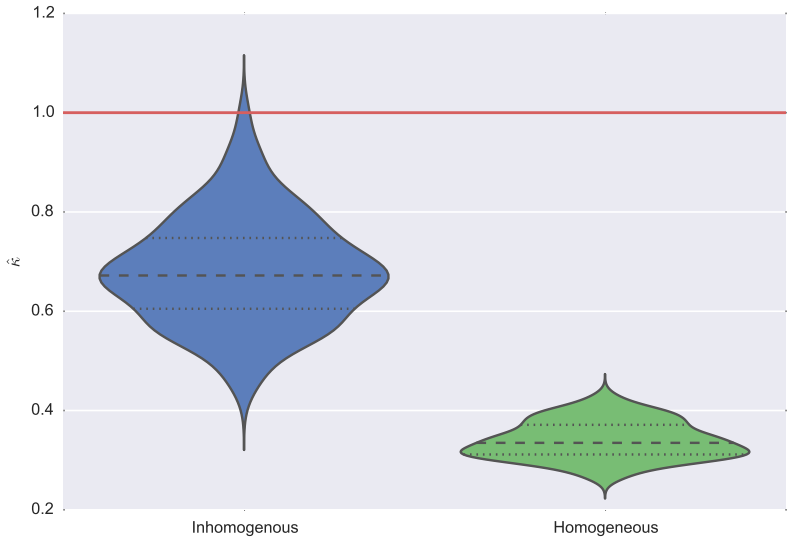

Figure 8.6: κ recovery under finite-sample-penalized AICc.

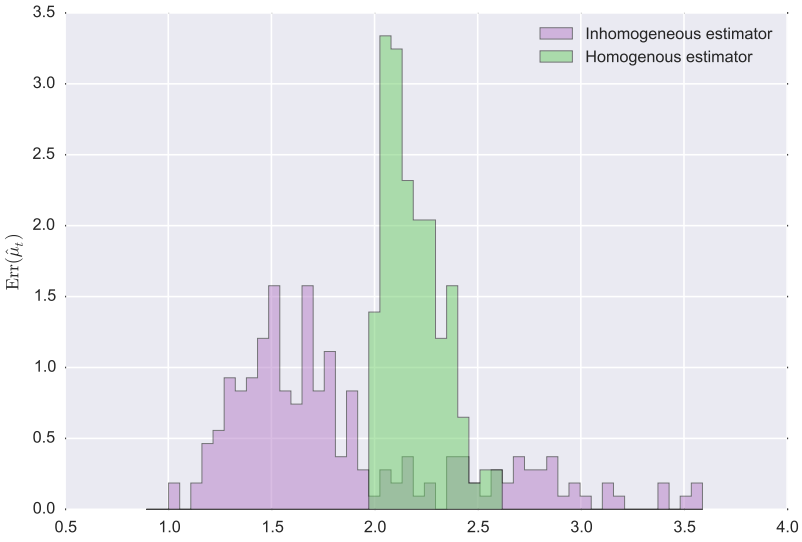

Figure 8.7: Overall error in background rate estimate $\mathrm{Err}\|\hat{\mu}_t - \mu_t\|_1/T$

Alternatively, consider the possibility that the estimator is identifying background rate peaks in terms of size, but it placing them at the incorrect location - say, $\hat{\mu}_t = 2 + 398\mathbb{I}_{148<t\leq 149}$ instead of $\mu_t = 2 + 398\mathbb{I}_{149<t\leq 150}$. In the L_1 penalty, this is more heavily penalized than $\hat{\mu}_t = 2$, which may not be desired behavior.

To diagnose this, I try an alternative error metric for this background rate, based on comparing the background rate and estimate using some smoothing kernel δ.

$$\mathrm{Err}_s(\hat{\mu}_t) := \|\hat{\mu}_t * \delta - \mu_t * \delta\|_1/T$$

I pick the arbitrary value

$$\delta(t) := \frac{1}{5}\mathbb{I}_{[-\frac{5}{2},\frac{5}{2})}$$

Indeed, this smoothed loss function shows the inhomogeneous estimator in a better light, performing nearly always better than the homogeneous competitor. Whether the smoothed version is a more appropriate error metric will depend upon the purpose. (Figure 8.8)

It seems that the inhomogeneous estimator is in fact selecting "good" parameters for this model. To understand what is happening here, I show the estimation paths for all the models in the simulation set for two of the parameters, once again with $\mu_t = 2 + 398\mathbb{I}_{149<t\leq 150}$, $\eta = 0.8$, $\kappa = 1.0$ (Figure 8.9)

As the penalty parameter increases, the estimates of the branching parameters are gradually perturbed over their ranges. Using the AICc selection criteria cor-

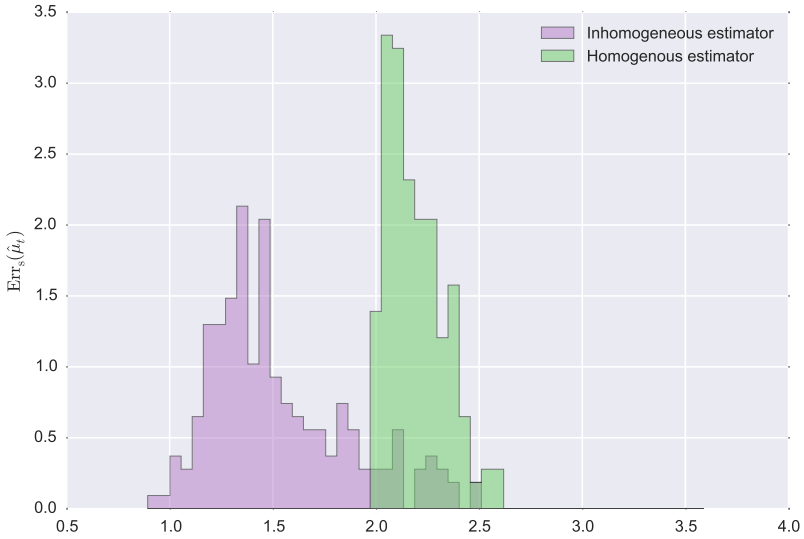

Figure 8.8: Error in smoothed background rate $\mathrm{Err}_s(\hat{\mu}_t)$

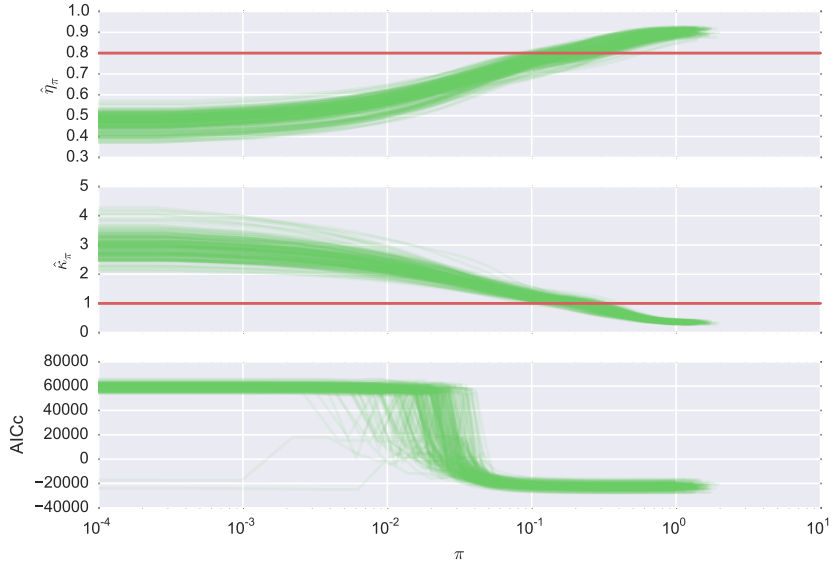

Figure 8.9: Estimated parameter values and AICc for different penalty parameters

responds to presuming that the model's estimates are optimal at the minimum of this AICc estimate. The graph might reassure us of this.

It also shows a potential problem with this procedure, which is that the AICc minimum for any given estimation path is very flat - hardly visible to the naked eye at this scale. It is also not clear whether the minimum necessarily corresponds to the oracle estimates in general. We could possibly do better by choosing some method of choosing the penalty, such as cross validation. For now, the AICc functions well enough for my purposes, and I will consider that question no further.

Now, let us consider the case that was especially poorly handled by the homogeneous estimator - the case of a non-branching, purely heterogeneous process, the "lead balloon" I set $\eta = 0$, and rerun the previous batch of simulations. (Figure 8.10, Figure 8.11, Figure 8.12), Figure 8.13)

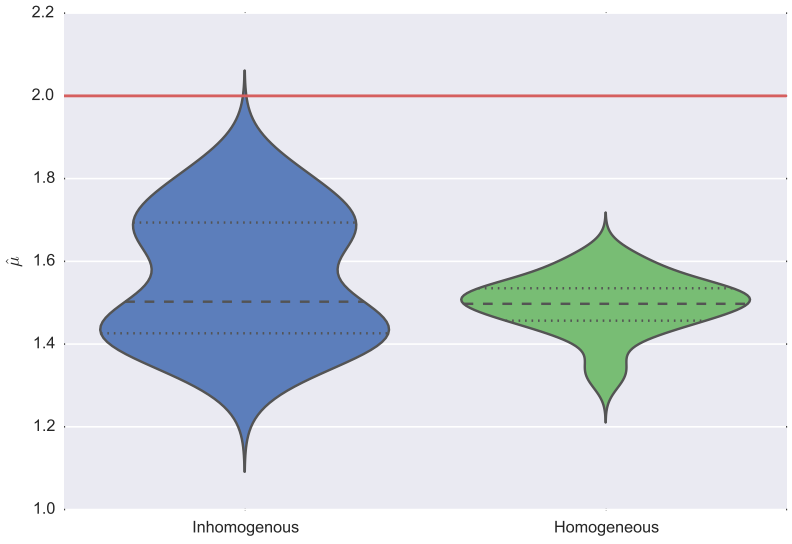

Figure 8.10: μ recovery under finite-sample-penalized AICc.

Once again, the approximation to the oracle values due to the inhomogeneous estimator are imperfect, but superior to the homogenous ones.

The negative values estimated for $\hat{\eta}$ and $\hat{\eta}$ are indications that I should constrain the estimator values to meaningful ranges, which is am omission in my coding.

More, I might improve this result by embedding this estimate in a selection procedure that will reject the Hawkes model entirely when there is no branching,

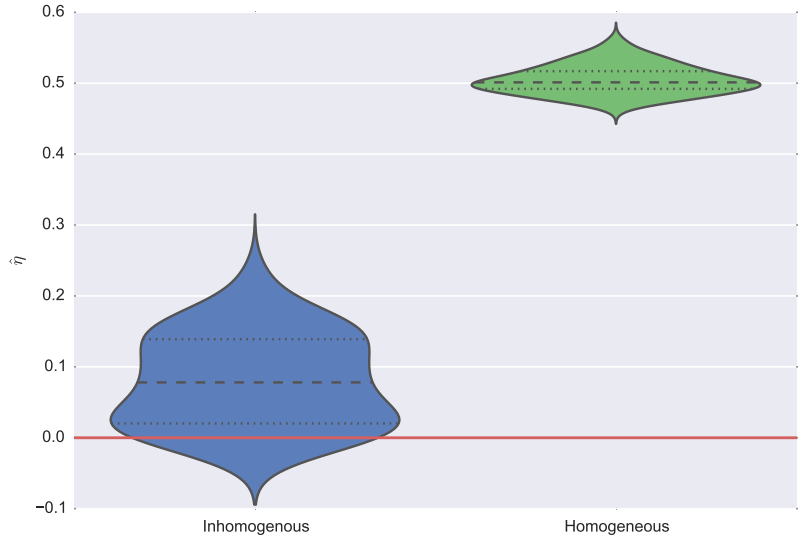

Figure 8.11: η recovery under finite-sample-penalized AICc.

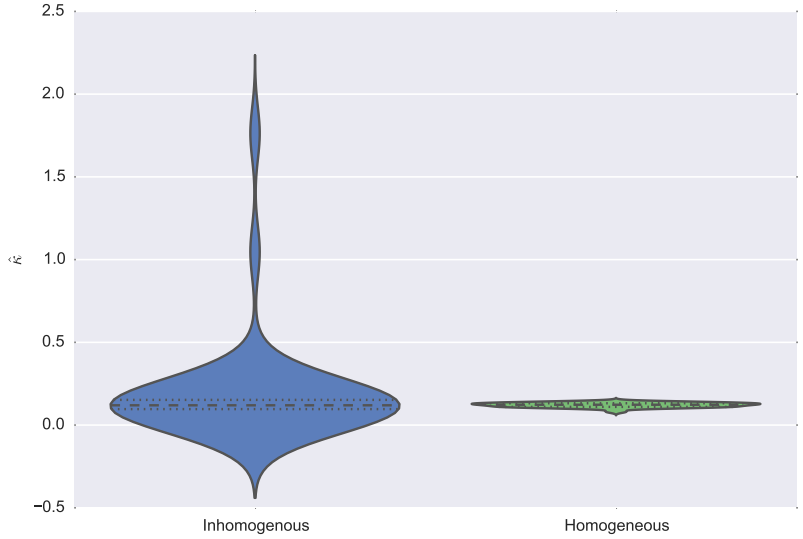

Figure 8.12: κ recovery under finite-sample-penalized AICc.

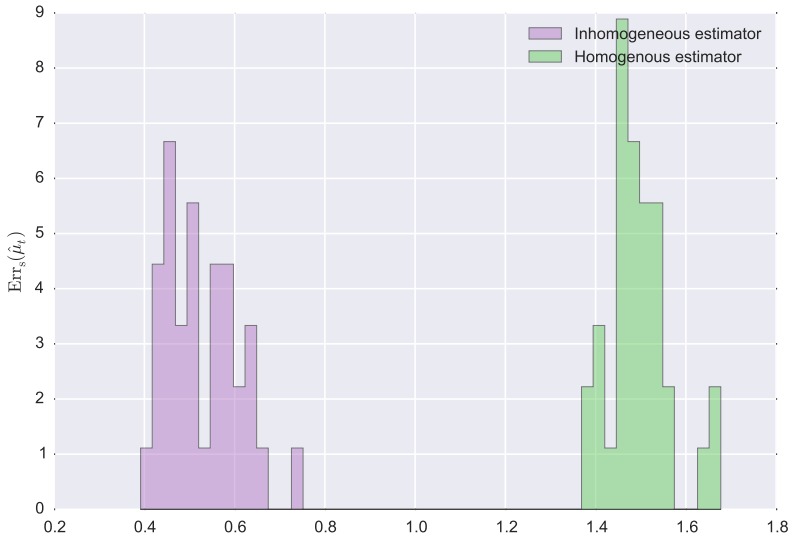

Figure 8.13: Estimation error in smoothed background rate $\mathrm{Err}_s(\hat{\mu}_t)$

much as in the homogeneous case we could reject the Hawkes mode in favor of a constant-rate Poisson process.

However, this is enough to begin.

There are many more steps in evaluating a new estimator than these simulations. I should also test the performance on data generated by non sparse-background rate processes μ_t, and variable bin width, and investigate the influence of observation interval, test alternative kernels and interpolation schemes, and so on.

For now, I have demonstrated that unlike the homogenous estimator, there is at least *some* hope for Youtube-type data, and I move on.

Chapter 9

Results for the inhomogeneous Hawkes model

Once again, I fit the estimator to selected times series from within the Youtube data set, withholding for the moment concrete hypotheses, and report the estimates.

My limits in this section are sharper. My algorithm is far more computationally intensive, and my code far less highly optimized, than `pyhawkes`. I will thus, in an exploratory mode, fit parameters to small subsets to validate that this idea in fact gets us somewhere, and see what further directions are supported for this kind of analysis.

9.1 Single time series detailed analysis

Turning to specific examples, I recall the time series `-2IXE5DcWzg`, *Valentin Elizalde, Volvere a amar*. The question I posed at the start was whether his biographer's hypothesized milestone was the cause of the spike in the data. Did this singer get more views on Youtube because of Billboard magazine listing him, or did Billboard magazine list him because of his surging popularity? Until this moment we've had no tools that could even hint at the answer to this kind of question.

I fit the inhomogeneous estimator to the data. Using AICc, I select the penalty $\pi = 0.466$, corresponding to $\hat{\mu} = 5.59$, $\hat{\eta} = 0.963$, $\hat{\kappa} = 0.186$. (Figure 9.1) Contrast with the homogeneous fit: $\hat{\mu} = 11.3$, $\hat{\eta} = 0.688$, $\hat{\kappa} = 0.992$, the model now selects a far *more* "viral" model for the view-rate growth, with a much shorter timescale and less mean background rate.

Graphing the estimated background intensity, I find that the model does estimate an increased intensity at around the start of that surge in interest. However, the date it suggests is 2007-02-21, substantially before the Billboard on listing 2007-03-3. This model suggests we need to look elsewhere to find an exogenous trigger. At the same time, it suggests that the singer was highly viral,

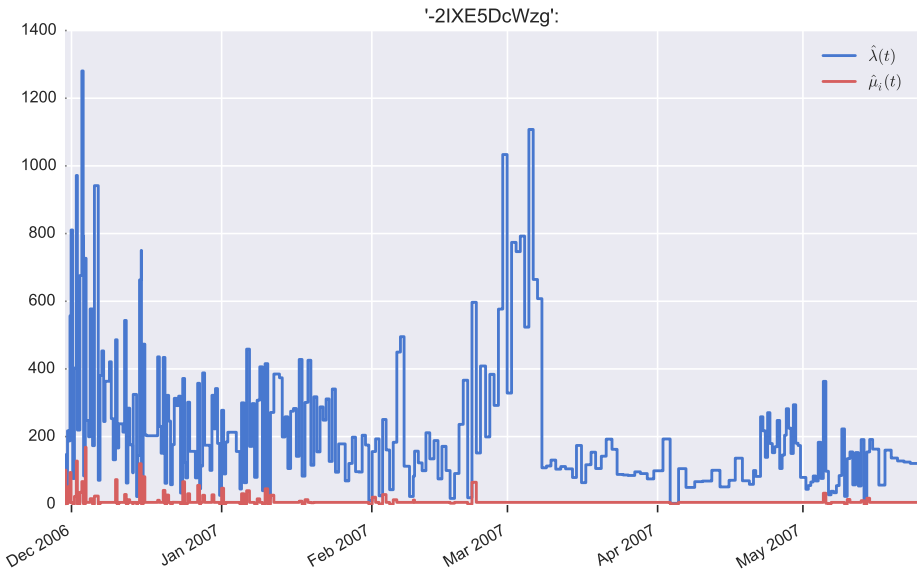

Figure 9.1: Penalized background rate and view-rate estimates $\hat{\lambda}_{simple}(t)$, for time series *Valentin Elizalde, Volvere a amar*

with an branching ratio close to critical. Thanks to simulations, we can argue that this near-criticality is not (necessarily) merely an artifact of inhomogenous background spiking. The estimator is giving plausible results here - It remains to validate them.

9.2 Aggregate analysis

Turning to bulk analysis, I try to fit as many models as possible.

One cost of my improvements in the estimator is a high computational burden. Running the software estimator through the random list of time series, I find that I have only estimated parameters for 913 models before termination of my batch jobs. Note that this also implies a degree of censoring of long time series, since those batch jobs were more likely to be terminated. I give summaries of these here, for illustrative purposes. I do need to know more about the sampling distribution of the estimator estimates in order to draw strong conclusions about population properties, even before I consider how to address the various other difficulties with the data. (Figure 9.2, Figure 9.3, Figure 9.4) As it turns out, the difference in ensemble distribution of estimates is not large, at least to visual inspection.

Despite the significant of the differences considering background rate makes on selected time series, over the ensemble my innovations turn out to make little

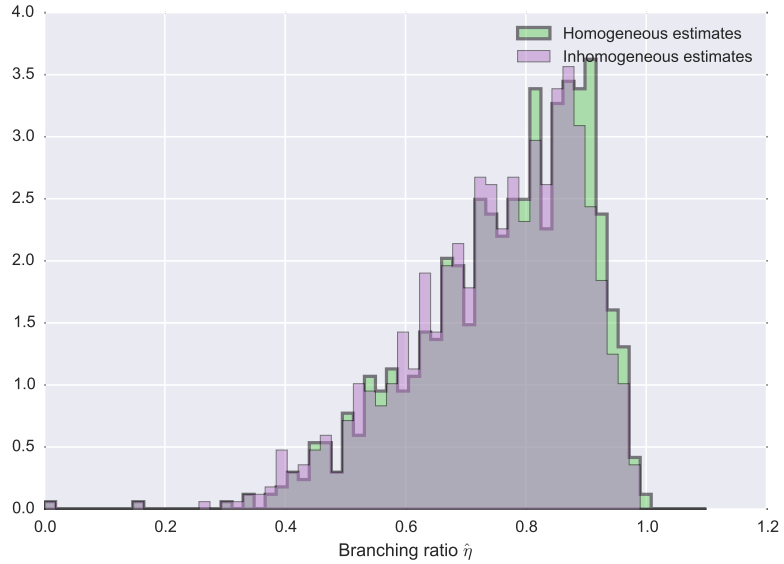

Figure 9.2: Branching ratio estimates for the inhomogeneous estimator

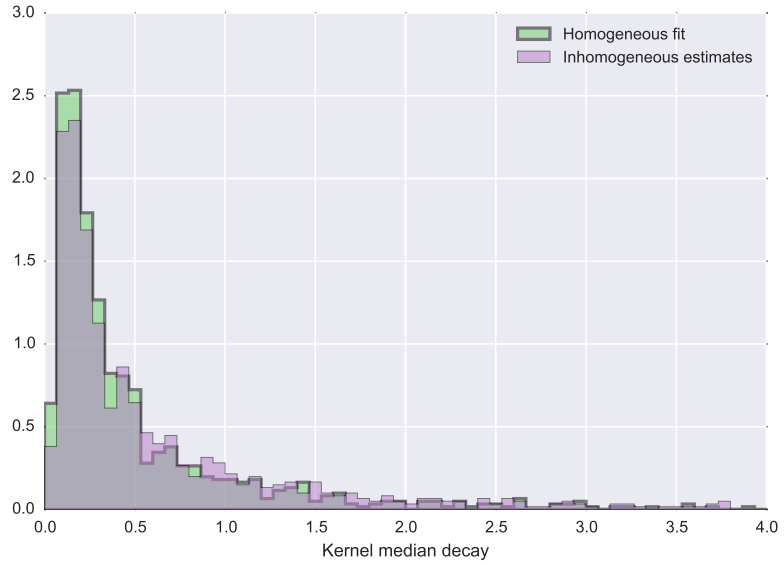

Figure 9.3: Kernel delay estimates for the inhomogeneous estimator.

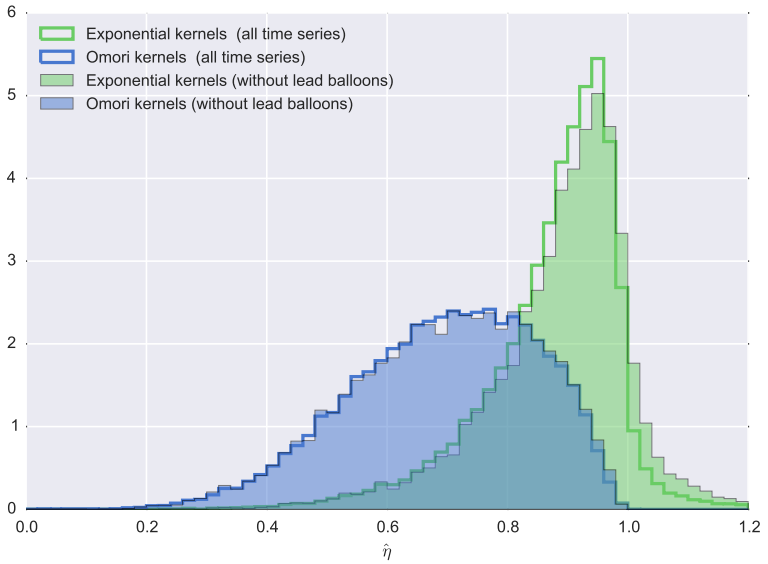

Figure 9.4: Branching ratio estimates in homogenous and inhomogenous cases, showing relationship of estimates.

different to the sampling distribution of randomly chosen series from the data. What is happening here?

We have a couple of possibilities. Firstly, that the problematic "lead balloon", and "spiky" time series I have chosen to test the estimator are *not* significant considered on the scale of the population. As I mentioned at the start, we do need a principled way of making such selections. Or we might be missing other significant types of inhomogeneity, such as slowly varying fluctuations, or that the estimator is often getting trapped in local optima on the messier real data. It could be that the this random sampling is not representative of the population qualities. Or it could be that the censoring of large time series due to their relatively higher computational requirements is discarding interesting results

We might consider improving this constraining our search space, by hypothesizing that the influence kernel of these videos has universal decay time, so that we could exploit the huge amount of data available; Once we are only estimating the background rate and branching ratio but not all the other parameters anew for each individual time series we can exploit the data more effectively. We can also consider optimizing the penalty hyper-parameter over the whole population.

All these require time, of course. The most logical next step, however, would be to set the estimator running on a database of the most problematic sets of time series in the database, and then, while the computing cluster is humming away, get out a pencil and derive new goodness-of-fit test for the model so I can provide

stronger and principled diagnostics diagnostics of these results.

Chapter 10

Conclusions

The Youtube data is both promising and troubling, as far as revealing the secrets of endogenously triggered dynamics. The kinds of problems and promise that it shows are, I believe of general importance, and I encountered many of them in the course of this thesis.

First I presented the seismically inspired Hawkes self exciting process as a potential model for viral dynamics in social media, and mentioned how its parameters might quantify endogenous dynamics in the system. I then presented the methods to estimate such a model, which are uncontroversial.

Ultimately I was not able to recommend tenable estimates for the parameters for this model, however for two reasons.

Firstly, the data is sparsely observed, and the ad hoc interpolation scheme used to approximate the missing information destroys some times of timing information, removing our ability to estimate kernel parameters in the important Omori law case.

Secondly, inhomogeneity in the data lead to extremely poor model identification for the estimator, and the distribution of the estimates so compiled is not a credible predictor of "true" parameters. Using the homogeneous model for this system may give good results for earthquake modeling, where there is no exogenous influence to control for. But where we are concerned with the interaction of endogenous and exogenous factors, these methods are not flexible enough. We cannot meaningfully find the "true" values for the parameter in the model when it is too ill-specified for the estimates of those true values to be informative.

At the same time, the model behind the branching process is much more general than the version we typically fit using off-the-shelf estimators, and I have shown that estimation procedures *can* be extended to estimate these more general models. My attempt to extend the estimator is not the only such attempt, and there are many diverse ways that it might be done. I have demonstrated that this method can solve certain problems, removing the bias due to large spikes,

and potentially identifying exogenous triggers from noisy data. There are clearly other issues to solve. At the same time, the method of penalized regression I have proposed is flexible and could provide the basis for many other such methods by different choice of kernel parameters, penalty functions and so on.

There remains much work to be done; Not only could the estimator be generalized with various penalties and kernel types, but it would also benefit from analysis regarding the sampling distribution, stability and model selection procedures. A practical demonstration, as I give here, is a necessary justification to invest in such work.

In short, whilst I cannot say right now that I have identified the parameters of the generating process of Youtube, I have shown that our evidence before now was indeterminate, and I have given a possible methods that, with a little more work, we could make a claim to identifying these parameters.

Appendix A

Technical Notes

A.1 Data extraction and cleaning

One problem with the data set is the size alone; I begin with an undocumented MySQL database with a disk footprint of approximately 40 gigabytes; Although certain queries run rapidly, most aggregate and summary statistics do not, either terminating due to resource-usage- errors. Based on naming conventions, I identify tables of particular interest; one apparently containing metadata for particular videos, and one containing time series information of video activity. These tables I store as plain Hierarchical Data Format files, divided into 256 "shards" based on the hash value video identifier.

The metadata table is of limited use because of various problems with incomplete or inconsistent data. Metadata about many time series is not available, or contains various types of corrupt or invalid data. Encoding is messy enough that I will not battle LaTeX to try to display it here. In any case, many records have apparently no metadata available, or if it is available, would require more extensive excavation from the database to extract it. Where available I use this metadata, but I do not restrict my investigation only to data points with available metadata.

Leaving metadata aside, I turn to the time series themselves.

I retrieve 676, 638, 684 distinct records from the database, corresponding to 4, 880, 136 distinct videos. Dividing these figures by one another might suggest I have nearly 5 million individual time series, each with more than a thousand observations.

This is not so, for two reasons:

1. Random sampling of time series reveals that time series are not all similarly long. In fact, the data set is dominated by short data-sets, on the order of 10 data points.
2. Even the remaining series can in fact be shorter than expected; The majority of the recorded observations are spurious and must be discarded, as will

	video_id	run_time	view_count
...	...	...	...
10	-2IXE5DcWzg	1165050318	921
11	-2IXE5DcWzg	1165081008	1035
12	-2IXE5DcWzg	1165084724	1035
13	-2IXE5DcWzg	1165115641	1306
14	-2IXE5DcWzg	1165139660	1662
15	-2IXE5DcWzg	1165146641	1726
16	-2IXE5DcWzg	1165177526	1756
17	-2IXE5DcWzg	1165177671	1756
18	-2IXE5DcWzg	1165191787	1876
19	-2IXE5DcWzg	1165209383	1876
20	-2IXE5DcWzg	1165235421	2001
21	-2IXE5DcWzg	1165241236	2001
22	-2IXE5DcWzg	1165243133	2001
23	-2IXE5DcWzg	1165264017	2067
24	-2IXE5DcWzg	1165274487	2067
25	-2IXE5DcWzg	1165306214	2349
...	...	...	...

Table A.1: Filtered time series

be explained below.

The cleaning and analysis that each dataset requires is complex enough that I cannot query these series *per se*. Instead, I download them all and inspect each individually. (Table A.1)

`run_time` I take to correspond to τ_i values. I assume it to be measured in *epoch timestamps* - the number of seconds since new year 1970 UTC. I convert this measure to "days" for convenient in the analysis however. `view_count` I take to denote $N_v(\tau_i)$ and `video_id` is a unique index v of the time series.

Note that many `view_count` values are repeated. Analysis of the data reveals many series like this, with repeated values. This could be evidence that no views occurred in a given time window. However, based on partial notes from the original author, and the sudden extreme increments that are interspersed between these "null increments", there is a more probably explanation: These are "cache hits": stale data presented to the user by the network, for performance reasons, in lieu of current information. I preprocess each time series to remove all non (strictly) monotonic increments, and discard the rest.

With these caveats, I repeat the time series excerpt for video `-2IXE5DcWzg` after preprocessing. (Table A.1)

I have effectively discounted all view increments of size zero. I am effectively

	video_id	run_time	view_count
...	...	...	...
10	-2IXE5DcWzg	1165036079	921
11	-2IXE5DcWzg	1165081008	1035
13	-2IXE5DcWzg	1165115641	1306
14	-2IXE5DcWzg	1165139660	1662
15	-2IXE5DcWzg	1165146641	1726
16	-2IXE5DcWzg	1165177526	1756
18	-2IXE5DcWzg	1165191787	1876
20	-2IXE5DcWzg	1165235421	2001
23	-2IXE5DcWzg	1165264017	2067
25	-2IXE5DcWzg	1165306214	2349
...	...	...	...

Table A.2: Time series with repeated timestamps filtered

also censoring all inactive time series; We cannot "see" any time series with only only zero or one observations - there must least two different view counts to interpolate. There is no clear way to estimate how significant this proportion is given what I know about the data; There is no way of measuring the significance of this choice precisely It could easily be the vast majority of videos which fall into this category. After all, the phrase "long tail" was notoriously popularized by *Wired* in 2004 to describe the preponderance of asymmetric distributions of popularity online [And04] and we should suspect that Youtube is such a system. It would be entirely possible that most of the videos are *never* viewed, and that this data cleaning has censored such videos from the analysis. The simple solution is to exclude this unknown proportion from our analysis. Therefore, throughout this work, it should be understood that the estimates I construct are all *conditional on sustained activity*.

A.2 On the complexity of the simplest possible thing

The first half of the analysis in this report uses the statistical library `pyhawkes`, and the second half hand-built code. The reason for this is technical rather than mathematical.

`pyhawkes` is an amazing project; optimized and featureful, it supports a wide variety of density kernel types, has mathematically and technically sophisticated optimizations and so on. It support multivariate and marked processes, a variety of different kernels etc.

It is also the kind of specialized racing vehicle that requires expert maintenance by qualified service personnel.

I did try to use it for the semi-parameteric regression, but ultimately, when my

needs were simple — optimizing parameters with respect to a simple loss function — I found myself introducing bugs rather than removing them.

When I added features it got messier, in that I encountered different problems. I tried to implement the non-parametric background rate using an off-the-shelf Gaussian Kernel Density Estimator library; The performance of that library was poor, its support for variable width and shape kernels was limited, and to take derivatives with respect to the kernel parameters required me to re-implement large parts of that library.

In the end, rather than modifying and combining and partially reimplementing two high complexity libraries to achieve a mathematically simple end, I judged it safer course to stitch together *simple* components to achieve a simple end.

The upshot is that my code - let us call it `excited` - is not API compatible with `pyhawkes`. Not even close. It uses Python mostly, with numba to dynamically compile the inner loop. It exploits the Scipy library Newton's method and L-BFGS solvers to find optima, which are technical innovations over `pyhawkes`. On the other hand, it does not implement Akaike's recursion relation to optimize calculation of exponential kernels, and is missing the other response kernels available in `pyhawkes`.

This situation is not ideal; In a perfect world, these features would all be combined into one package. In the real world, however, I am enrolled in a *statistics* program rather than *software engineering*, and would be punished accordingly if I sacrificed thoroughness in my statistical analysis in order to observe niceties of software development.

It turned out that the simplest possible bit of code that could solve my statistical problem was in fact complex. Thus, although, access to the code is available upon request, consider yourself warned.

Bibliography

[A+08] Shariar Azizpour, Kay Giesecke, et al. *Self-exciting corporate defaults: contagion vs. frailty*. Stanford University working paper series, 2008. URL: http://web.stanford.edu/dept/MSandE/cgi-bin/people/faculty/giesecke/pdfs/selfexciting.pdf (visited on 03/16/2015).

[Abr+06] Felix Abramovich et al. "Adapting to unknown sparsity by controlling the false discovery rate". In: *The Annals of Statistics* 34.2 (Apr. 2006), pp. 584–653. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/009053606000000074. URL: http://projecteuclid.org/euclid.aos/1151418235 (visited on 04/01/2015).

[Aka73] Hirotogu Akaike. "Information Theory and an Extension of the Maximum Likelihood Principle". In: *Proceeding of the Second International Symposium on Information Theory*. Ed. by Petrovand F Caski. Budapest: Akademiai Kiado, 1973, pp. 199–213. ISBN: 978-1-4612-7248-9, 978-1-4612-1694-0. URL: http://link.springer.com/chapter/10.1007/978-1-4612-1694-0_15 (visited on 04/06/2015).

[And04] Chris Anderson. "The Long Tail". In: *Wired* 12.10 (Oct. 2004). URL: http://archive.wired.com/wired/archive/12.10/tail.html.

[AS09] Giada Adelfio and Frederic Paik Schoenberg. "Point process diagnostics based on weighted second-order statistics and their asymptotic properties". In: *Annals of the Institute of Statistical Mathematics* 61.4 (Dec. 1, 2009), pp. 929–948. ISSN: 0020-3157, 1572-9052. DOI: 10.1007/s10463-008-0177-1. URL: http://link.springer.com/article/10.1007/s10463-008-0177-1 (visited on 01/08/2015).

[BA04] Kenneth P. Burnham and David R. Anderson. "Multimodel Inference Understanding AIC and BIC in Model Selection". In: *Sociological Methods & Research* 33.2 (Nov. 1, 2004), pp. 261–304. ISSN: 0049-1241, 1552-8294. DOI: 10.1177/0049124104268644. URL: http://smr.sagepub.com/content/33/2/261 (visited on 03/27/2015).

[Bac+12] Emmanuel Bacry et al. "Scaling limits for Hawkes processes and application to financial statistics". In: (Feb. 3, 2012). arXiv: 1202.0842. URL: http://arxiv.org/abs/1202.0842 (visited on 06/18/2014).

[Bat92] Roberto Battiti. "First-and second-order methods for learning: between steepest descent and Newton's method". In: *Neural computation* 4.2 (1992), pp. 141–166. ISSN: 0899-7667. DOI: 10.1162/neco.1992.4.2.141. URL: http://rtm.science.unitn.it/~battiti/archive/FirstSecondOrderMethodsForLearning.PDF (visited on 03/20/2015).

[BD89] Mark Berman and Peter Diggle. "Estimating Weighted Integrals of the Second-Order Intensity of a Spatial Point Process". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 51.1 (Jan. 1, 1989), pp. 81–92. ISSN: 0035-9246. URL: https://publications.csiro.au/rpr/pub?list=BRO%5C&pid=procite:d5b7ecd7-435c-4dab-9063-f1cf2fbdf4cb (visited on 03/31/2015).

[BDM12] E. Bacry, K. Dayri, and J. F. Muzy. "Non-parametric kernel estimation for symmetric Hawkes processes. Application to high frequency financial data". In: *The European Physical Journal B* 85.5 (May 1, 2012), pp. 1–12. ISSN: 1434-6028, 1434-6036. DOI: 10.1140/epjb/e2012-21005-8. arXiv: 1112.1838. URL: http://arxiv.org/abs/1112.1838 (visited on 06/18/2014).

[Ben10] Yoav Benjamini. "Simultaneous and selective inference: Current successes and future challenges". In: *Biometrical Journal* 52.6 (Dec. 1, 2010), pp. 708–721. ISSN: 1521-4036. DOI: 10.1002/bimj.200900299. URL: http://onlinelibrary.wiley.com/doi/10.1002/bimj.200900299/abstract (visited on 03/31/2015).

[Ber+11] James S. Bergstra et al. "Algorithms for hyper-parameter optimization". In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2011, pp. 2546–2554. URL: http://papers.nips.cc/paper/4443-algorithms-for-hyper-parameter-optimization (visited on 03/27/2015).

[Ber14] J. M. Berger. *How ISIS Games Twitter*. The Atlantic. June 16, 2014. URL: http://www.theatlantic.com/international/archive/2014/06/isis-iraq-twitter-social-media-strategy/372856/ (visited on 04/10/2015).

[BG15] Peter Bühlmann and Sara van de Geer. "High-dimensional inference in misspecified linear models". In: (Mar. 22, 2015). arXiv: 1503.06426. URL: http://arxiv.org/abs/1503.06426 (visited on 03/27/2015).

[BH95] Yoav Benjamini and Yosef Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1 (Jan. 1, 1995), pp. 289–300. ISSN: 0035-9246. JSTOR: 2346101.

[BM02] P. Brémaud and L. Massoulié. "Power spectra of general shot noises and Hawkes point processes with a random excitation". In: *Advances in Applied Probability* 34.1 (Mar. 2002), pp. 205–222. ISSN: 0001-8678, 1475-6064. DOI: 10.1239/aap/1019160957. URL: http://icwww.epfl.ch/~bremaud/spectra_hawkes.ps (visited on 03/16/2015).

[BM14a] Emmanuel Bacry and Jean-Francois Muzy. "Second order statistics characterization of Hawkes processes and non-parametric estimation". In: (Jan. 5, 2014). arXiv: 1401.0903. URL: http://arxiv.org/abs/1401.0903 (visited on 01/19/2015).

[BM14b] Emmanuel Bacry and Jean-François Muzy. "Hawkes model for price and trades high-frequency dynamics". In: *Quantitative Finance* 14.7 (2014), pp. 1147–1166. ISSN: 1469-7688. DOI: 10.1080/14697688.2014.897000. URL: http://www.cmap.polytechnique.fr/~bacry/ftp/papers/neo13.pdf (visited on 06/18/2014).

[Bon+12] Robert M. Bond et al. "A 61-million-person experiment in social influence and political mobilization". In: *Nature* 489.7415 (Sept. 13, 2012), pp. 295–298. ISSN: 0028-0836. DOI: 10.1038/nature11421. URL: http://www.nature.com/nature/journal/v489/n7415/full/nature11421.html (visited on 04/13/2015).

[Bro+02] ERVRL Brown et al. "The time-rescaling theorem and its application to neural spike train data analysis". In: *Neural computation* 14.2 (Feb. 2002), pp. 325–346. ISSN: 0899-7667. DOI: 10.1162/08997660252741149. URL: http://www.stat.cmu.edu/~kass/papers/rescaling.pdf (visited on 01/08/2015).

[BT05] Adrian Baddeley and Rolf Turner. "Spatstat: an R package for analyzing spatial point patterns". In: *Journal of statistical software* 12.6 (2005), pp. 1–42.

[Büh02] Peter Bühlmann. "Bootstraps for Time Series". In: *Statistical Science* 17.1 (Feb. 1, 2002), pp. 52–72. ISSN: 0883-4237. URL: ftp://stat.ethz.ch/Research-Reports/87.pdf (visited on 02/03/2015).

[BY05] Yoav Benjamini and Daniel Yekutieli. "False Discovery Rate–Adjusted Multiple Confidence Intervals for Selected Parameters". In: *Journal of the American Statistical Association* 100.469 (Mar. 2005), pp. 71–81. ISSN: 0162-1459, 1537-274X. DOI: 10.1198/016214504000001907. URL: http://www.math.tau.ac.il/~yekutiel/papers/JASA%20FCR%20prints.pdf (visited on 03/31/2015).

[Cav97] Joseph E. Cavanaugh. "Unifying the derivations for the Akaike and corrected Akaike information criteria". In: *Statistics & Probability Letters* 33.2 (Apr. 30, 1997), pp. 201–208. ISSN: 0167-7152. DOI: 10.1016/S0167-7152(96)00128-9. URL: http://www.sciencedirect.com/science/article/pii/S0167715296001289 (visited on 03/27/2015).

[CCD95] Gilles Celeux, Didier Chauveau, and Jean Diebolt. *On Stochastic Versions of the EM Algorithm*. report. 1995. URL: https://hal.inria.fr/inria-00074164/document (visited on 03/05/2015).

[CK12] D. R. Cox and Christiana Kartsonaki. "The fitting of complex parametric models". In: *Biometrika* 99.3 (Sept. 1, 2012), pp. 741–747. ISSN: 0006-3444, 1464-3510. DOI: 10.1093/biomet/ass030. URL: http://biomet.oxfordjournals.org/content/99/3/741 (visited on 12/23/2014).

[Cla08] Gerda Claeskens. *Model selection and model averaging*. In collab. with Nils Lid Hjort. Cambridge series in statistical and probabilistic mathematics. Cambridge ; New York: Cambridge University Press, 2008. 312 pp. ISBN: 9780521852258.

[Cox65] D. R. Cox. "On the Estimation of the Intensity Function of a Stationary Point Process". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 27.2 (Jan. 1, 1965), pp. 332–337. ISSN: 0035-9246. JSTOR: 2984202.

[CRT06] Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. "Stable signal recovery from incomplete and inaccurate measurements". In: *Communications on Pure and Applied Mathematics* 59.8 (Aug. 1, 2006), pp. 1207–1223. ISSN: 1097-0312. DOI: 10.1002/cpa.20124. URL: http://arxiv.org/abs/math/0503066 (visited on 08/17/2014).

[CS08] Riley Crane and Didier Sornette. "Robust dynamic classes revealed by measuring the response function of a social system". In: *Proceedings of the National Academy of Sciences* 105.41 (Oct. 14, 2008), pp. 15649–15653. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0803685105. pmid: 18824681. URL: http://www.pnas.org/content/105/41/15649 (visited on 04/06/2014).

[CSS10] Riley Crane, Frank Schweitzer, and Didier Sornette. "Power law signature of media exposure in human response waiting time distributions". In: *Physical Review E* 81.5 (May 3, 2010), p. 056101. DOI: 10.1103/PhysRevE.81.056101. URL: http://link.aps.org/doi/10.1103/PhysRevE.81.056101 (visited on 04/06/2014).

[Cuc08] Lionel Cucala. "Intensity Estimation for Spatial Point Processes Observed with Noise". In: *Scandinavian Journal of Statistics* 35.2 (June 1, 2008), pp. 322–334. ISSN: 1467-9469. DOI: 10.1111/j.1467-9469.

2007.00583.x. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9469.2007.00583.x/full (visited on 03/03/2015).

[Dig85] Peter Diggle. "A Kernel Method for Smoothing Point Process Data". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 34.2 (Jan. 1, 1985), pp. 138–147. ISSN: 0035-9254. DOI: 10.2307/2347366. URL: http://www.maths.tcd.ie/~mnl/store/Diggle1985a.pdf (visited on 02/24/2015).

[DLM99] Bernard Delyon, Marc Lavielle, and Eric Moulines. "Convergence of a stochastic approximation version of the EM algorithm". In: *The Annals of Statistics* 27.1 (Mar. 1999), pp. 94–128. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/aos/1018031103. URL: http://projecteuclid.org/euclid.aos/1018031103 (visited on 03/05/2015).

[DLR77] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 39.1 (Jan. 1, 1977). ArticleType: research-article / Full publication date: 1977 / Copyright © 1977 Royal Statistical Society, pp. 1–38. ISSN: 0035-9246. JSTOR: 2984875.

[Don06] D.L. Donoho. "Compressed sensing". In: *IEEE Transactions on Information Theory* 52.4 (Apr. 2006), pp. 1289–1306. ISSN: 0018-9448. DOI: 10.1109/TIT.2006.871582.

[DS05] Fabrice Deschâtres and Didier Sornette. "Dynamics of book sales: Endogenous versus exogenous shocks in complex networks". In: *Physical Review E* 72.1 (2005), p. 016112. ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.72.016112. URL: http://link.aps.org/doi/10.1103/PhysRevE.72.016112 (visited on 05/21/2014).

[DV03] Daryl J. Daley and D. Vere-Jones. *An introduction to the theory of point processes*. 2nd ed. Vol. 1. Elementary theory and methods. New York: Springer, 2003. ISBN: 0387215646 9780387215648 038795541O 9780387955414. URL: http://ebooks.springerlink.com/UrlApi.aspx?action=summary%5C&v=1%5C&bookid=108085 (visited on 11/11/2014).

[DV08] Daryl J. Daley and David Vere-Jones. *An introduction to the theory of point processes*. 2nd ed. Vol. 2. General theory and structure. Probability and Its Applications. New York: Springer, Jan. 1, 2008. ISBN: 978-0-387-21337-8, 978-0-387-49835-5. URL: http://link.springer.com/chapter/10.1007/978-0-387-49835-5_7 (visited on 11/11/2014).

[DZ11] Angelos Dassios and Hongbiao Zhao. "A dynamic contagion process". In: *Advances in Applied Probability* 43.3 (Sept. 2011). Zentralblatt MATH identifier 05955087 , Mathematical Reviews number (MathSciNet) MR2858222, pp. 814–846. ISSN: 0001-8678, 1475-6064.

DOI: 10.1239/aap/1316792671. URL: http://projecteuclid.org/euclid.aap/1316792671 (visited on 03/05/2014).

[Efr+04] Bradley Efron et al. "Least angle regression". In: *The Annals of Statistics* 32.2 (Apr. 2004), pp. 407–499. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/009053604000000067. URL: http://arxiv.org/abs/math/0406456 (visited on 03/20/2015).

[Efr04] Bradley Efron. "The Estimation of Prediction Error". In: *Journal of the American Statistical Association* 99.467 (Sept. 1, 2004), pp. 619–632. ISSN: 0162-1459. DOI: 10.1198/016214504000000692. URL: http://www.cs.berkeley.edu/~jordan/sail/readings/archive/efron_Cp.pdf (visited on 03/19/2015).

[Efr86] Bradley Efron. "How biased is the apparent error rate of a prediction rule?" In: *Journal of the American Statistical Association* 81.394 (June 1, 1986), pp. 461–470. ISSN: 0162-1459. DOI: 10.1080/01621459.1986.10478291. URL: http://www.stat.washington.edu/courses/stat527/s13/readings/j_am_stat_assoc1986.pdf (visited on 02/08/2015).

[FHT10] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of statistical software* 33.1 (2010), pp. 1–22. ISSN: 1548-7660. pmid: 20808728. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2929880/ (visited on 03/20/2015).

[Fil+14] Vladimir Filimonov et al. "Quantification of the high level of endogeneity and of structural regime shifts in commodity markets". In: *Journal of International Money and Finance*. Understanding International Commodity Price Fluctuations 42 (Apr. 2014), pp. 174–192. ISSN: 0261-5606. DOI: 10.1016/j.jimonfin.2013.08.010. URL: http://www.sciencedirect.com/science/article/pii/S0261560613001125 (visited on 11/20/2014).

[Fri+07] Jerome Friedman et al. "Pathwise coordinate optimization". In: *The Annals of Applied Statistics* 1.2 (Dec. 2007), pp. 302–332. ISSN: 1932-6157, 1941-7330. DOI: 10.1214/07-AOAS131. URL: http://projecteuclid.org/euclid.aoas/1196438020 (visited on 03/20/2015).

[FS13] Vladimir Filimonov and Didier Sornette. *Apparent criticality and calibration issues in the Hawkes self-excited point process model: application to high-frequency financial data*. SSRN Scholarly Paper ID 2371284. Rochester, NY: Social Science Research Network, Aug. 30, 2013. arXiv: 1308.6756. URL: http://arxiv.org/abs/1308.6756 (visited on 06/10/2014).

[FWS15] Vladimir Filimonov, Spencer Wheatley, and Didier Sornette. "Effective measure of endogeneity for the Autoregressive Conditional Duration point processes via mapping to the self-excited Hawkes process". In: *Communications in Nonlinear Science and Numerical Simulation* 22.1–3 (May 2015), pp. 23–37. ISSN: 1007-5704. DOI: 10.1016/j.cnsns.2014.08.042. URL: http://arxiv.org/abs/1306.2245 (visited on 03/30/2015).

[Gee+14] Sara van de Geer et al. "On asymptotically optimal confidence regions and tests for high-dimensional models". In: *The Annals of Statistics* 42.3 (June 2014), pp. 1166–1202. ISSN: 0090-5364. DOI: 10.1214/14-AOS1221. arXiv: 1303.0518. URL: http://arxiv.org/abs/1303.0518 (visited on 12/18/2014).

[Ger+05] Matthew C. Gerstenberger et al. "Real-time forecasts of tomorrow's earthquakes in California". In: *Nature* 435.7040 (May 19, 2005), pp. 328–331. ISSN: 0028-0836, 1476-4679. DOI: 10.1038/nature03622. URL: http://www.nature.com/doifinder/10.1038/nature03622 (visited on 02/24/2015).

[GKM11] K. Giesecke, H. Kakavand, and M. Mousavi. "Exact Simulation of Point Processes with Stochastic Intensities". In: *Operations Research* 59.5 (Oct. 1, 2011), pp. 1233–1245. ISSN: 0030-364X. DOI: 10.1287/opre.1110.0962. URL: http://pubsonline.informs.org/doi/abs/10.1287/opre.1110.0962 (visited on 01/08/2015).

[GL05] Jiang Gui and Hongzhe Li. "Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data". In: *Bioinformatics* 21.13 (July 1, 2005), pp. 3001–3008. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/bti422. pmid: 15814556. URL: http://bioinformatics.oxfordjournals.org/content/21/13/3001 (visited on 03/20/2015).

[GL08] G. Grinstein and R. Linsker. "Power-law and exponential tails in a stochastic priority-based model queue". In: *Physical Review E* 77.1 (Jan. 7, 2008), p. 012101. DOI: 10.1103/PhysRevE.77.012101. URL: http://link.aps.org/doi/10.1103/PhysRevE.77.012101 (visited on 04/08/2015).

[GL11] Sara van de Geer and Johannes Lederer. "The Lasso, correlated design, and improved oracle inequalities". In: (July 1, 2011). arXiv: 1107.0189. URL: http://arxiv.org/abs/1107.0189 (visited on 03/27/2015).

[GMR93] C. Gourieroux, A. Monfort, and E. Renault. "Indirect Inference". In: *Journal of Applied Econometrics* 8 (Dec. 1, 1993), S85–S118. ISSN: 0883-7252. JSTOR: 2285076.

[Gre87] Peter J. Green. "Penalized Likelihood for General Semi-Parametric Regression Models". In: *International Statistical Review / Revue Internationale de Statistique* 55.3 (Dec. 1, 1987), pp. 245–259. ISSN: 0306-7734. DOI: 10.2307/1403404. URL: http://www.maths.bris.ac.uk/~mapjg/papers/green_isr_87.pdf (visited on 03/01/2015).

[GW08] Christopher Genovese and Larry Wasserman. "Adaptive confidence bands". In: *The Annals of Statistics* 36.2 (Apr. 2008), pp. 875–905. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/07-AOS500. URL: http://projecteuclid.org/euclid.aos/1205420522 (visited on 03/27/2015).

[Hal12] Peter F. Halpin. "An EM algorithm for Hawkes process". In: *Psychometrika* 2 (2012). URL: https://www.steinhardt.nyu.edu/scmsAdmin/uploads/007/126/Halpin_Proceedings_Submit.pdf (visited on 11/18/2014).

[Haw71] Alan G. Hawkes. "Point spectra of some mutually exciting point processes". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 33.3 (Jan. 1, 1971), pp. 438–443. ISSN: 0035-9246. JSTOR: 2984686.

[HB14] Stephen J. Hardiman and Jean-Philippe Bouchaud. "Branching-ratio approximation for the self-exciting Hawkes process". In: *Physical Review E* 90.6 (Dec. 11, 2014), p. 062807. DOI: 10.1103/PhysRevE.90.062807. arXiv: 1403.5227. URL: http://arxiv.org/abs/1403.5227 (visited on 03/15/2015).

[HBB13] Stephen J. Hardiman, Nicolas Bercot, and Jean-Philippe Bouchaud. "Critical reflexivity in financial markets: a Hawkes process analysis". In: *The European Physical Journal B* 86.10 (Oct. 1, 2013), pp. 1–9. ISSN: 1434-6028, 1434-6036. DOI: 10.1140/epjb/e2013-40107-3. URL: http://arxiv.org/abs/1302.1405 (visited on 03/05/2014).

[HG10] Asif-ul Haque and Paul Ginsparg. "Last but not least: Additional positional effects on citation and readership in arXiv". In: *Journal of the American Society for Information Science and Technology* 61.12 (Dec. 1, 2010), pp. 2381–2388. ISSN: 1532-2890. DOI: 10.1002/asi.21428. URL: http://arxiv.org/abs/1010.2757 (visited on 04/08/2015).

[HK70] Arthur E. Hoerl and Robert W. Kennard. "Ridge Regression: Biased Estimation for Nonorthogonal Problems". In: *Technometrics* 12.1 (Feb. 1, 1970), pp. 55–67. ISSN: 0040-1706. DOI: 10.1080/00401706.1970.10488634. URL: http://math.arizona.edu/~hzhang/math574m/Read/Ridge.pdf (visited on 04/04/2015).

[HO74] Alan G. Hawkes and David Oakes. "A cluster process representation of a self-exciting process". In: *Journal of Applied Probability* 11.3 (Sept. 1974), p. 493. ISSN: 00219002. DOI: 10.2307/3212693. JSTOR: 3212693.

[HSG03] A. Helmstetter, D. Sornette, and J.-R. Grasso. "Mainshocks are aftershocks of conditional foreshocks: How do foreshock statistical properties emerge from aftershock laws". In: *Journal of Geophysical Research* 108 (B1 2003), p. 2046. ISSN: 0148-0227. DOI: 10.1029/2002JB001991. arXiv: cond-mat/0205499. URL: http://arxiv.org/abs/cond-mat/0205499 (visited on 02/11/2015).

[HT89] Clifford M. Hurvich and Chih-Ling Tsai. "Regression and time series model selection in small samples". In: *Biometrika* 76.2 (June 1, 1989), pp. 297–307. ISSN: 0006-3444, 1464-3510. DOI: 10.1093/biomet/76.2.297. URL: http://biomet.oxfordjournals.org/content/76/2/297 (visited on 03/27/2015).

[HW14] A. Helmstetter and M. J. Werner. "Adaptive Smoothing of Seismicity in Time, Space, and Magnitude for Time-Dependent Earthquake Forecasts for California". In: *Bulletin of the Seismological Society of America* 104.2 (Apr. 1, 2014), pp. 809–822. ISSN: 0037-1106. DOI: 10.1785/0120130105. URL: http://www.bssaonline.org/cgi/doi/10.1785/0120130105 (visited on 02/24/2015).

[IVV11] Raghuram Iyengar, Christophe Van den Bulte, and Thomas W. Valente. "Opinion leadership and social contagion in new product diffusion". In: *Marketing Science* 30.2 (2011), pp. 195–212. ISSN: 0732-2399. DOI: 10.1287/mksc.1100.0566. URL: http://pubsonline.informs.org/doi/abs/10.1287/mksc.1100.0566 (visited on 05/21/2014).

[JT04] Wenxin Jiang and Bruce Turnbull. "The Indirect Method: Inference Based on Intermediate Statistics—A Synthesis and Examples". In: *Statistical Science* 19.2 (May 2004), pp. 239–263. ISSN: 0883-4237, 2168-8745. DOI: 10.1214/088342304000000152. URL: http://www.planchet.net/EXT/ISFA/1226.nsf/769998e0a65ea348c1257052003eb94f/bf9e68719dd7b5f8c1256f560032680f/$FILE/Indirect%20Method%20-%20JIANG%20TURNBULL.pdf (visited on 12/23/2014).

[Ken+05] Bruce E. Kendall et al. "Population cycles in the pine looper moth: Dynamical tests of mechanistic hypotheses". In: *Ecological Monographs* 75.2 (May 1, 2005), pp. 259–276. ISSN: 0012-9615. URL: http://www.sysecol2.ethz.ch/Refs/EntClim/K/Ke169.pdf (visited on 12/23/2014).

[KK96] Sadanori Konishi and Genshiro Kitagawa. "Generalised information criteria in model selection". In: *Biometrika* 83.4 (Dec. 1, 1996), pp. 875–890. ISSN: 0006-3444, 1464-3510. DOI: 10.1093/biomet/83.4.875. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.127.1018%5C&rep=rep1%5C&type=pdf (visited on 02/24/2015).

[KL04] Estelle Kuhn and Marc Lavielle. "Coupling a stochastic approximation version of EM with an MCMC procedure". In: *ESAIM: Probability and Statistics* 8 (Sept. 2004), pp. 115–131. ISSN: 1262-3318. DOI:

10.1051/ps:2004007. URL: http://www.esaim-ps.org/action/article_S1292810004000072 (visited on 03/05/2015).

[KR14] S. Kaufman and S. Rosset. "When does more regularization imply fewer degrees of freedom? Sufficient conditions and counterexamples". In: *Biometrika* 101.4 (Dec. 1, 2014), pp. 771–784. ISSN: 0006-3444, 1464-3510. DOI: 10.1093/biomet/asu034. URL: http://biomet.oxfordjournals.org/content/101/4/771 (visited on 02/08/2015).

[Kün89] Hans Rudolf Künsch. "The Jackknife and the Bootstrap for General Stationary Observations". In: *The Annals of Statistics* 17.3 (Sept. 1, 1989), pp. 1217–1241. ISSN: 0090-5364. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.28.924%5C&rep=rep1%5C&type=pdf (visited on 02/03/2015).

[Lah01] S N Lahiri. "Effects of block lengths on the validity of block resampling methods". In: *Probability Theory and Related Fields* 121 (2001), pp. 73–97. DOI: 10.1007/PL00008798.

[Lah93] S N Lahiri. "On the moving block bootstrap under long range dependence". In: *Statistics & Probability Letters* 18.5 (1993), pp. 405–413. DOI: 10.1016/0167-7152(93)90035-H.

[Lie11] Marie-Colette N. M. van Lieshout. "On Estimation of the Intensity Function of a Point Process". In: *Methodology and Computing in Applied Probability* 14.3 (Aug. 4, 2011), pp. 567–578. ISSN: 1387-5841, 1573-7713. DOI: 10.1007/s11009-011-9244-9. URL: http://link.springer.com/article/10.1007/s11009-011-9244-9 (visited on 02/26/2015).

[Loc+14] Richard Lockhart et al. "A significance test for the lasso". In: *The Annals of Statistics* 42.2 (Apr. 2014), pp. 413–468. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/13-AOS1175. URL: http://arxiv.org/abs/1405.6805 (visited on 01/21/2015).

[MB10] Nicolai Meinshausen and Peter Bühlmann. "Stability selection". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.4 (Sept. 1, 2010), pp. 417–473. ISSN: 1467-9868. DOI: 10.1111/j.1467-9868.2010.00740.x. arXiv: 0809.2932. URL: http://arxiv.org/abs/0809.2932 (visited on 07/18/2014).

[Mei07] Nicolai Meinshausen. "Relaxed Lasso". In: *Computational Statistics & Data Analysis* 52.1 (Sept. 15, 2007), pp. 374–393. ISSN: 0167-9473. DOI: 10.1016/j.csda.2006.12.019. URL: http://stat.ethz.ch/~nicolai/relaxo.pdf (visited on 03/27/2015).

[Mei14] Nicolai Meinshausen. "Group bound: confidence intervals for groups of variables in sparse high dimensional regression without assumptions on the design". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* (Nov. 1, 2014), n/a–n/a. ISSN: 1467-9868. DOI: 10.1111/rssb.12094. arXiv: 1309.3489. URL: http://arxiv.org/abs/1309.3489 (visited on 04/08/2015).

[MJM13] James S. Martin, Ajay Jasra, and Emma McCoy. "Inference for a class of partially observed point process models". In: *Annals of the Institute of Statistical Mathematics* 65.3 (June 1, 2013), pp. 413–437. ISSN: 0020-3157, 1572-9052. DOI: 10.1007/s10463-012-0375-8. arXiv: 1201.4529. URL: http://arxiv.org/abs/1201.4529 (visited on 01/08/2015).

[MMB09] Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann. "p-Values for High-Dimensional Regression". In: *Journal of the American Statistical Association* 104.488 (Dec. 1, 2009), pp. 1671–1681. ISSN: 0162-1459. DOI: 10.1198/jasa.2009.tm08647. URL: http://arxiv.org/abs/0811.2177 (visited on 03/19/2015).

[Moh+11] G. O. Mohler et al. "Self-exciting point process modeling of crime". In: *Journal of the American Statistical Association* 106.493 (Mar. 1, 2011), pp. 100–108. ISSN: 0162-1459. DOI: 10.1198/jasa.2011.ap09546. URL: http://amstat.tandfonline.com/doi/abs/10.1198/jasa.2011.ap09546 (visited on 11/11/2014).

[Møl03] Jesper Møller. "Shot noise Cox processes". In: *Advances in Applied Probability* 35.3 (Sept. 2003), pp. 614–640. ISSN: 0001-8678, 1475-6064. DOI: 10.1239/aap/1059486821. URL: http://www.maphysto.dk/publications/MPS-RR/2002/18.pdf (visited on 12/12/2014).

[MPW96] William A. Massey, Geraldine A. Parker, and Ward Whitt. "Estimating the parameters of a nonhomogeneous Poisson process with linear rate". In: *Telecommunication Systems* 5.2 (Sept. 1, 1996), pp. 361–388. ISSN: 1018-4864, 1572-9451. DOI: 10.1007/BF02112523. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.129.365 (visited on 05/12/2014).

[MSW98] Jesper Møller, Anne Randi Syversveen, and Rasmus Plenge Waagepetersen. "Log Gaussian Cox Processes". In: *Scandinavian Journal of Statistics* 25.3 (Sept. 1, 1998), pp. 451–482. ISSN: 1467-9469. DOI: 10.1111/1467-9469.00115. URL: http://onlinelibrary.wiley.com/doi/10.1111/1467-9469.00115/abstract (visited on 02/26/2015).

[MY09] Nicolai Meinshausen and Bin Yu. "Lasso-type recovery of sparse representations for high-dimensional data". In: *The Annals of Statistics* 37.1 (Feb. 2009), pp. 246–270. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/07-AOS582. URL: http://projecteuclid.org/euclid.aos/1232115934 (visited on 03/27/2015).

[NG13] Richard Nickl and Sara van de Geer. "Confidence sets in sparse regression". In: *The Annals of Statistics* 41.6 (Dec. 2013), pp. 2852–2876. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/13-AOS1170. URL: http://arxiv.org/abs/1209.1508 (visited on 03/27/2015).

[OA82] Yosihiko Ogata and Hirotugu Akaike. "On linear intensity models for mixed doubly stochastic Poisson and self-exciting point processes". In: *Journal of the Royal Statistical Society, Series B* 44 (1982), pp. 269–274. DOI: 10.1007/978-1-4612-1694-0_20. URL: http://bemlar.ism.ac.jp/zhuang/Refs/Refs/ogata1982.pdf (visited on 11/24/2014).

[Oak75] David Oakes. "The Markovian self-exciting process". In: *Journal of Applied Probability* 12.1 (Mar. 1975), p. 69. ISSN: 00219002. DOI: 10.2307/3212408. JSTOR: 3212408.

[Oga78] Yoshiko Ogata. "The asymptotic behaviour of maximum likelihood estimators for stationary point processes". In: *Annals of the Institute of Statistical Mathematics* 30.1 (Dec. 1, 1978), pp. 243–261. ISSN: 0020-3157, 1572-9052. DOI: 10.1007/BF02480216. URL: http://users.iems.northwestern.edu/~armbruster/2007msande444/ogata-78.pdf (visited on 08/19/2014).

[Oga88] Yosihiko Ogata. "Statistical models for earthquake occurrences and residual analysis for point processes". In: *Journal of the American Statistical Association* 83.401 (Mar. 1, 1988), pp. 9–27. ISSN: 0162-1459. DOI: 10.1080/01621459.1988.10478560. URL: http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1988.10478560 (visited on 11/11/2014).

[OMK93] Yosihiko Ogata, Ritsuko S. Matsu'ura, and Koichi Katsura. "Fast likelihood computation of epidemic type aftershock-sequence model". In: *Geophysical Research Letters* 20.19 (Oct. 8, 1993), pp. 2143–2146. ISSN: 1944-8007. DOI: 10.1029/93GL02142. URL: http://onlinelibrary.wiley.com/doi/10.1029/93GL02142/abstract (visited on 12/02/2014).

[Ore12] Alexei Oreskovic. *Exclusive: YouTube hits 4 billion daily video views*. Jan. 23, 2012. URL: http://uk.reuters.com/article/2012/01/23/us-google-youtube-idUSTRE80M0TS20120123 (visited on 03/16/2015).

[Oza79] T. Ozaki. "Maximum likelihood estimation of Hawkes' self-exciting point processes". In: *Annals of the Institute of Statistical Mathematics* 31.1 (Dec. 1, 1979), pp. 145–155. ISSN: 0020-3157, 1572-9052. DOI: 10.1007/BF02480272. URL: http://www.ism.ac.jp/editsec/aism/pdf/031_1_0145.pdf (visited on 04/09/2014).

[Ras13] Jakob Gulddahl Rasmussen. "Bayesian inference for Hawkes processes". In: *Methodology and Computing in Applied Probability* 15.3 (Sept. 1, 2013), pp. 623–642. ISSN: 1387-5841, 1573-7713. DOI: 10.1007/s11009-

011-9272-5. URL: http://vbn.aau.dk/ws/files/45838419/R_2011_03.pdf (visited on 11/18/2014).

[REU06] REUTERS. "YouTube serves up 100 million videos a day online". In: (July 16, 2006). URL: http://usatoday30.usatoday.com/tech/news/2006-07-16-youtube-views_x.htm.

[Roi07] Manuel Roig-Franzia. "Mexican Drug Cartels Leave a Bloody Trail on YouTube". In: *The Washington Post. World* (Apr. 9, 2007). ISSN: 0190-8286. URL: http://www.washingtonpost.com/wp-dyn/content/article/2007/04/08/AR2007040801005_2.html (visited on 04/06/2015).

[RPL15] Marcello Rambaldi, Paris Pennesi, and Fabrizio Lillo. "Modeling FX market activity around macroeconomic news: a Hawkes process approach". In: *Physical Review E* 91.1 (Jan. 26, 2015), p. 012819. DOI: 10.1103/PhysRevE.91.012819. arXiv: 1405.6047. URL: http://arxiv.org/abs/1405.6047 (visited on 01/21/2015).

[Rub72] Izhak Rubin. "Regular point processes and their detection". In: *IEEE Transactions on Information Theory* 18.5 (Sept. 1972), pp. 547–557. ISSN: 0018-9448. DOI: 10.1109/TIT.1972.1054897.

[SB03] A Smith and E Brown. "Estimating a state-space model from point process observations". In: *Neural Computation* 15.5 (May 2003), pp. 965–991. ISSN: 0899-7667. DOI: 10.1162/089976603765202622. URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6789993.

[Sch05] Frederic Paik Schoenberg. "Consistent parametric estimation of the intensity of a spatial–temporal point process". In: *Journal of Statistical Planning and Inference* 128.1 (Jan. 15, 2005), pp. 79–93. ISSN: 0378-3758. DOI: 10.1016/j.jspi.2003.09.027. URL: http://escholarship.org/uc/item/6584c641 (visited on 02/24/2015).

[SCV10] Frederic Paik Schoenberg, Annie Chu, and Alejandro Veen. "On the relationship between lower magnitude thresholds and bias in epidemic-type aftershock sequence parameter estimates". In: *Journal of Geophysical Research: Solid Earth* 115 (B4 Apr. 1, 2010), B04309. ISSN: 2156-2202. DOI: 10.1029/2009JB006387. URL: http://onlinelibrary.wiley.com/doi/10.1029/2009JB006387/abstract (visited on 02/24/2015).

[SH03] D Sornette and A Helmstetter. "Endogenous versus exogenous shocks in systems with memory". In: *Physica A: Statistical Mechanics and its Applications* 318.3–4 (Feb. 15, 2003), pp. 577–591. ISSN: 0378-4371. DOI: 10.1016/S0378-4371(02)01371-7. URL: http://arxiv.org/abs/cond-mat/0206047 (visited on 05/21/2014).

[Sha15] Leslie Shaffer. "The dress that broke the Internet 16 million views in 6 hours." In: (Feb. 27, 2015). URL: http://www.cnbc.com/id/102461771 (visited on 03/13/2015).

[Sil82] B. W. Silverman. "On the Estimation of a Probability Density Function by the Maximum Penalized Likelihood Method". In: *The Annals of Statistics* 10.3 (Sept. 1982), pp. 795–810. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/aos/1176345872. URL: http://oai.dtic.mil/oai/oai?verb=getRecord%5C&metadataPrefix=html%5C&identifier=ADA103875 (visited on 03/06/2015).

[Sim+11] Noah Simon et al. "Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent". In: *Journal of Statistical Software* 39.5 (Mar. 2011). URL: http://www.jstatsoft.org/v39/i05/paper (visited on 03/20/2015).

[Smi93] A. A. Smith. "Estimating nonlinear time-series models using simulated vector autoregressions". In: *Journal of Applied Econometrics* 8 (S1 1993), S63–S84. ISSN: 1099-1255. DOI: 10.1002/jae.3950080506. URL: http://www.econ.yale.edu/smith/2285075.pdf (visited on 01/21/2015).

[SMM02] D. Sornette, Y. Malevergne, and J. F. Muzy. "Volatility fingerprints of large shocks: Endogeneous versus exogeneous". In: (Apr. 30, 2002). What causes crashes? Risk Volume 16 (2), 67-71 (February 2003). arXiv: cond-mat/0204626. URL: http://arxiv.org/abs/cond-mat/0204626 (visited on 04/06/2014).

[Sor+04] Didier Sornette et al. "Endogenous versus exogenous shocks in complex networks: An empirical test using book sale rankings". In: *Physical Review Letters* 93.22 (Nov. 22, 2004), p. 228701. ISSN: 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.93.228701. URL: http://prl.aps.org/abstract/PRL/v93/i22/e228701 (visited on 05/21/2014).

[Sor06] Didier Sornette. "Endogenous versus exogenous origins of crises". In: *Extreme events in nature and society*. The Frontiers Collection. Springer, 2006, pp. 95–119. arXiv: physics/0412026. URL: http://arxiv.org/abs/physics/0412026 (visited on 05/21/2014).

[Sor09] Didier Sornette. "Dragon-Kings, Black Swans and the Prediction of Crises". In: 2.1 (July 24, 2009). arXiv: 0907.4290. URL: http://arxiv.org/abs/0907.4290 (visited on 04/06/2015).

[SS11] A. Saichev and D. Sornette. "Hierarchy of temporal responses of multivariate self-excited epidemic processes". In: (Jan. 8, 2011). arXiv: 1101.1611. URL: http://arxiv.org/abs/1101.1611 (visited on 04/06/2014).

[SU09] D. Sornette and S. Utkin. "Limits of declustering methods for disentangling exogenous from endogenous events in time series with foreshocks, main shocks, and aftershocks". In: *Physical Review E* 79.6 (June 16, 2009), p. 061110. DOI: 10.1103/PhysRevE.79.061110.

arXiv: 0903.3217. URL: http://arxiv.org/abs/0903.3217 (visited on 06/18/2014).

[Sug78] Nariaki Sugiura. "Further analysts of the data by Akaike' s Information Criterion and the finite corrections". In: *Communications in Statistics - Theory and Methods* 7.1 (Jan. 1, 1978), pp. 13–26. ISSN: 0361-0926. DOI: 10.1080/03610927808827599. URL: http://dx.doi.org/10.1080/03610927808827599 (visited on 03/27/2015).

[TG65] A. N. Tikhonov and V. B. Glasko. "Use of the regularization method in non-linear problems". In: *USSR Computational Mathematics and Mathematical Physics* 5.3 (1965), pp. 93–107. ISSN: 0041-5553. DOI: 10.1016/0041-5553(65)90150-3. URL: http://www.sciencedirect.com/science/article/pii/0041555365901503 (visited on 04/05/2015).

[Tib96] Robert Tibshirani. "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1 (Jan. 1, 1996), pp. 267–288. ISSN: 0035-9246. URL: http://statweb.stanford.edu/~tibs/lasso/lasso.pdf (visited on 04/06/2015).

[Uts70] Tokuji Utsu. "Aftershocks and earthquake statistics (1): Some parameters which characterize an aftershock sequence and their interrelations". In: *Journal of the Faculty of Science, Hokkaido University. Series 7, Geophysics* 3.3 (1970), pp. 129–195. URL: http://eprints2008.lib.hokudai.ac.jp/dspace/handle/2115/8683 (visited on 04/12/2015).

[Vac11] Anca Patricia Vacarescu. "Filtering and parameter estimation for partially observed generalized Hawkes processes". PhD thesis. Stanford University, 2011. URL: http://oatd.org/oatd/record?record=oai%5C:purl.stanford.edu%5C:tc922qd0500 (visited on 01/08/2015).

[VS08] Alejandro Veen and Frederic P Schoenberg. "Estimation of Space–Time Branching Process Models in Seismology Using an EM–Type Algorithm". In: *Journal of the American Statistical Association* 103.482 (June 1, 2008), pp. 614–624. ISSN: 0162-1459. DOI: 10.1198/016214508000000148. URL: http://www.stat.ucla.edu/~frederic/papers/em.pdf (visited on 01/19/2015).

[Wer+10] Maximilian J Werner et al. "Adaptively smoothed seismicity earthquake forecasts for Italy". In: *Annals of Geophysics* 3 (Nov. 5, 2010). ISSN: 2037416X. DOI: 10.4401/ag-4839. URL: http://www.annalsofgeophysics.eu/index.php/annals/article/view/4839 (visited on 02/24/2015).

[Whi11] Ben Whitelaw. "Almost all YouTube views come from just 30% of films". In: (Apr. 20, 2011). URL: http://www.telegraph.co.uk/technology/news/8464418/Almost-all-YouTube-views-come-from-just-30-of-films.html (visited on 03/16/2015).

[WL08] Tong Tong Wu and Kenneth Lange. "Coordinate descent algorithms for lasso penalized regression". In: *The Annals of Applied Statistics* 2.1 (Mar. 2008), pp. 224–244. ISSN: 1932-6157, 1941-7330. DOI: 10.1214/07-AOAS147. URL: http://arxiv.org/abs/0803.3876 (visited on 03/20/2015).

[WR09] Larry Wasserman and Kathryn Roeder. "High-dimensional variable selection". In: *Annals of statistics* 37.5A (Jan. 1, 2009), pp. 2178–2201. ISSN: 0090-5364. DOI: 10.1214/08-AOS646. pmid: 19784398. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2752029/ (visited on 03/27/2015).

[WT90] Greg C. G. Wei and Martin A. Tanner. "A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms". In: *Journal of the American Statistical Association* 85.411 (Sept. 1, 1990), pp. 699–704. ISSN: 0162-1459. DOI: 10.1080/01621459.1990.10474930. URL: http://www.biostat.jhsph.edu/~rpeng/biostat778/papers/wei-tanner-1990.pdf (visited on 03/05/2015).

[Wu83] C. F. Jeff Wu. "On the Convergence Properties of the EM Algorithm". In: *The Annals of Statistics* 11.1 (Mar. 1983), pp. 95–103. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/aos/1176346060. URL: http://www.stanford.edu/class/ee378b/papers/wu-em.pdf (visited on 02/11/2015).

[You14] Youtube. *We never thought a video would be watched in numbers greater than a 32-bit integer*. Dec. 1, 2014. URL: https://plus.google.com/+youtube/posts/BUXfdWqu86Q (visited on 04/13/2015).

[ZHT07] Hui Zou, Trevor Hastie, and Robert Tibshirani. "On the "degrees of freedom" of the lasso". In: *The Annals of Statistics* 35.5 (Oct. 2007), pp. 2173–2192. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/009053607000000127. URL: http://projecteuclid.org/euclid.aos/1194461726 (visited on 03/18/2015).

[Zhu13] Lingjiong Zhu. "Moderate deviations for Hawkes processes". In: *Statistics & Probability Letters* 83.3 (Mar. 2013), pp. 885–890. ISSN: 0167-7152. DOI: 10.1016/j.spl.2012.12.011. URL: https://ideas.repec.org/a/eee/stapro/v83y2013i3p885-890.html (visited on 02/16/2015).

[ZZ14] Cun-Hui Zhang and Stephanie S. Zhang. "Confidence intervals for low dimensional parameters in high dimensional linear models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.1 (2014), pp. 217–242. ISSN: 1467-9868. DOI: 10.1111/rssb.12026. URL: http://onlinelibrary.wiley.com/doi/10.1111/rssb.12026/abstract (visited on 12/18/2014).

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

Authored by (in block letters):
For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.