

# Bayesian estimation of decay parameters in Hawkes processes

Tiago Santos<sup>a,\*</sup>, Florian Lemmerich<sup>b</sup> and Denis Helic<sup>a</sup>

<sup>a</sup>*Institute of Interactive Systems and Data Science, TU Graz, Graz, Austria*

<sup>b</sup>*Faculty of Computer Science and Mathematics, University of Passau, Passau, Germany*

**Abstract.** Hawkes processes with exponential kernels are a ubiquitous tool for modeling and predicting event times. However, estimating their decay parameter is challenging, and there is a remarkable variability among decay parameter estimates. Moreover, this variability increases substantially in cases of a small number of realizations of the process or due to sudden changes to a system under study, for example, in the presence of exogenous shocks. In this work, we demonstrate that these estimation difficulties relate to the noisy, non-convex shape of the Hawkes process' log-likelihood as a function of the decay. To address uncertainty in the estimates, we propose to use a Bayesian approach to learn more about likely decay values. We show that our approach alleviates the decay estimation problem across a range of experiments with synthetic and real-world data. With our work, we support researchers and practitioners in their applications of Hawkes processes in general and in their interpretation of Hawkes process parameters in particular.

Keywords: Hawkes process, decay rate, Bayesian inference

## 1. Introduction

As a method for modeling and predicting temporal event sequences (henceforth *event streams*), Hawkes processes have seen broad application, ranging from estimating social dynamics in online communities [17], through measuring financial market movements [2] to modeling earthquakes [34]. Researchers and practitioners derive utility from Hawkes processes due to their ability to model irregularly spaced timestamps of events of interest. Therein lies the crucial difference with respect to time series models, which assume observations arrive at time intervals of constant size. Adopting Hawkes processes to model temporal data thus allows for capturing granular temporal signals which time series models disregard or aggregate. As an example, consider the problem of modeling user behavior on Twitter. User activity is bursty [6], i.e., it is characterized by short time spans of increased activity interspersed by longer breaks. A Hawkes process model of user behavior on Twitter explicitly fits the irregularly spaced timestamps of a user's Tweets. Its fitted parameters provide insights into user excitation, i.e., how strong and prolonged her bursts are, and allow for predicting the event times of future Tweets, and thus overall user engagement on Twitter [17]. Similarly, a Hawkes process model fitted to the timestamps of earthquakes facilitates estimations of the timing of future bursts of tectonic activity, such as earthquakes and aftershocks [34]. Hawkes process applications in the realm of finance may leverage the timestamps of

---

\*Corresponding author: Tiago Santos, Institute of Interactive Systems and Data Science, TU Graz, Inffeldgasse 16, 8010 Graz, Austria. E-mail: tsantos.tugraz@gmail.com.

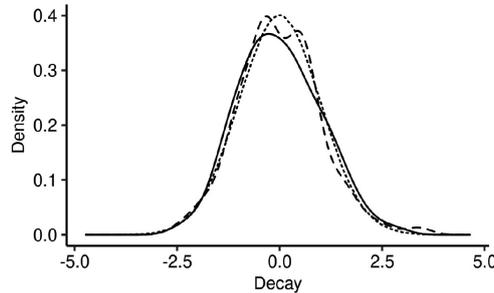


Fig. 1. Remarkable uncertainty in fitted decay parameter values. We illustrate the normalized distribution of decay values estimated with L-BFGS-B across 100 realizations of the same Hawkes process (solid line). The discrepancy between that distribution and the standard Gaussian distribution (dotted line) suggests unique uncertainty in Hawkes process decay estimations. Fitting the decay parameter in Hawkes processes with breaks in stationarity results in other kinds of uncertainty (cf. dashed line).

trades to estimate the growth and decay of clusters of market movements across geographies [2]. Overall, Hawkes processes are well-established methods to model and predict temporal data with irregularly distributed event times.

Technically, Hawkes processes model event streams via the conditional intensity function, the infinitesimal event rate given the event history. Events cause *jumps* in the conditional intensity function, which *decays* to a *baseline* level following a pre-defined functional form, the so-called *kernel*. This kernel is often chosen as an exponential function. The reasons for this choice are manifold, as Hawkes processes with an exponential kernel are (i) efficient to simulate and estimate [18,32], (ii) parsimonious, and (iii) realistic in practical applications [43,48].

### Problem

While baseline and jump parameters of Hawkes processes are typically derived via convex optimization, the estimation of decay parameters in exponential kernels remains an open issue. Previous work simply assumed the decay parameters to be constants [3,12,18], cross-validated decay parameter values [11, 17,39], or estimated them with a range of different optimization approaches [4,13,19,29,35,40,48]. Such estimation approaches result in point estimates that can be considered as sufficient for simulating and predicting event streams. However, researchers frequently interpret Hawkes process parameters directly [5,34,44,48], e.g., to infer the directions of temporal dependency [23,28,46]. Even though such applications rely on estimates of decay parameter values, the current state-of-the-art neglects that, in practice, we derive these values only within a degree of certainty, in particular in a common case of a small number of realizations of a given process. This can lead to qualitatively inaccurate conclusions if researchers do not quantify the uncertainty of decay parameter estimates. Moreover, there are only initial studies [36,40] on how exponential growth, exogenous shocks to a system, or changes in the underlying process mechanics compromise key stationarity assumptions and aggravate estimation errors.

In Fig. 1, we illustrate this problem in relation to a commonly used nonlinear optimization approach for fitting the decay, L-BFGS-B [10]. We estimate the decay value in two cases: (i) a small number of realizations, and (ii) change of process parameters due to a stationarity break. The solid line in the figure represents the normalized distribution of the fitted decay parameter for case (i). We observe that line deviates remarkably from the standard Gaussian depicted as a dotted line, suggesting the need to account for the uncertainty surrounding Hawkes process decay parameter estimations. Repeating the same fitting procedure for case (ii), we observe, in the dashed line, yet other distributional properties of fitted decay values. This suggests a specific need to separately account for uncertainty in the decay parameter estimation in the presence of breaks in stationarity.

### *This work*

The contribution of this paper is two-fold: First, we focus on the potential uncertainty in the decay parameter of Hawkes processes and explore its sources and consequences. In that regard, we uncover that the non-convex and noisy shape of the Hawkes process log-likelihood as a function of the decay is one cause of the variability in decay parameter estimates. Second, we propose to integrate the estimation of the decay parameter in a Bayesian approach. This allows for (i) quantifying uncertainty in the decay parameter, (ii) diagnosing estimation errors, and (iii) addressing breaks of the crucial stationarity assumption. In our approach, we formulate and evaluate closed-form and intractable hypotheses on the value of the decay parameter. Specifically, we encode hypotheses for the decay as parameters of a prior distribution. Then, we consider estimations of the decay for single Hawkes process realizations as samples from a likelihood. These likelihood samples form the data that we combine with the prior to perform Bayesian inference of posterior decay values.

We show with synthetic data as well as in a broad range of real-world domains that our Bayesian inference procedure for fitting the decay parameter allows for quantifying uncertainty, diagnosing estimation errors, and analyzing breaks in stationarity. In our first application, a study of earthquakes in Japanese regions [34], we uncover low uncertainty in certain geographical relationships. Second, we support Settles and Meeder's [41] hypothesis that vocabulary learning effort on the Duolingo mobile app correlates to the estimated difficulty of the learned words [33]. Finally, leveraging a dataset of Tweets before and after the Paris terror attacks of November 2015, we measure the relation between a stationarity-breaking exogenous shock and collective effervescence [20].

Overall, our work sheds light on fitting a widely used class of Hawkes processes, i.e., Hawkes processes with exponential kernels. Better understanding these models and explicitly surfacing uncertainty in their fitted values facilitates their use by practitioners and researchers. We expect the impact of our study to be broad,<sup>1</sup> as our results influence the application of a key analysis approach for studying time-dependent phenomena across most diverse domains such as seismology, finance, or online communities.

## **2. Related work**

Framed in a broad context, our work complements literature on temporal data analysis methods and applications. Temporal data analysis often models phenomena of interest as time series for tasks such as, prominently, prediction [27], outlier detection [21] or similarity search [49]. In particular, the latter is the goal of the broad research stream on time series representation methods (cf. Lin et al.'s [31] citation graph), which derive compact symbolic representations for segments of a time series. Crucially, such representation and segmentation approaches also facilitate the analysis of a more general class of temporal data, where, in contrast to time series, events are not equally spaced in time. To learn from such data, which we term event streams, previous research proposes, for example, (1) deriving binary codes to infer directions of causality between event streams [9], (2) detecting multiple levels of temporal granularity via self-adapting segmenting of event streams [30], (3) or dynamically defining differently sized temporal windows to cluster users of recommender systems [50]. Complementing that research, we demonstrate how to model irregular temporal signatures with Hawkes processes, thus providing a stepping stone for future work to incorporate, rather than discretize, such information. Applications such as Aghababaei and Makrehchi's [1] predictions of crime trends from Twitter data, or Rocha and Rodrigues' [37] forecasts of

---

<sup>1</sup>We make our code available at [https://github.com/tfts/hawkes\\_exp\\_bayes](https://github.com/tfts/hawkes_exp_bayes).

emergency department admissions could benefit from our approach by capturing rich signals embedded in the event times of Tweets or of hospital admissions. Thus, much like Tai and Hsu [45] with their study on commodity price prediction with deep learning, we also provide guidance for practitioners to apply Hawkes processes. Finally, on the methodological side, we also complement existing research highlighting challenges and approaches to learn from non-stationary data [38].

We now position our work within the more specific context of Hawkes process research. One of the first fields to leverage the seminal work by Hawkes [24] includes seismology [14,34]. Since then, Hawkes process theory and practice emerged in the realm of finance [4,5,46], as well as, more recently, in modeling user activity online [23,28,29,36,40,43,44,48,52,53]. More specifically, the latter body of work extended Hawkes processes to predict diffusion and popularity dynamics of online media [36,52,53], model online learning [43,48], capture the spread of misinformation [44], and understand user behavior in online communities [28,40], in online markets [23] and in the context of the offline world [29]. As all of those previous references interpreted the parameter values of Hawkes processes (and variations thereof), they may benefit from our study of the decay parameter, especially as we uncover its properties and assess and mitigate estimation issues with Bayesian inference.

Perhaps closest to our work is Bacry et al.'s [4] study of mean field inference of Hawkes process values. In particular, those authors inspected the effect of varying the decay parameter across a range of values: With increasing decay, fitted self- and cross-excitations decrease while baseline intensity increases. We go beyond their study by deepening our understanding of the (noisy) properties of the Hawkes log-likelihood as a function of the decay. Methodologically, our Bayesian approach relates to Hosseini et al.'s [25]. Those authors infer the decay parameter by assuming a Gamma prior and computing the mean of samples from the posterior (as part of a larger inference problem). In our work, we instead focus on the Bayesian approach as a means to quantify estimation uncertainty. Further, as our Bayesian changepoint model captures breaks in stationarity, we simplify previous work [36,40] which relies on additional assumptions, such as estimating stationarity via the time series of event counts. Our work also complements recent efforts [46] to learn Hawkes processes from small data.

### 3. Decay estimation in Hawkes processes

In this section, we first summarize fundamentals of Hawkes process modelling before describing the problem of decay estimation.

#### 3.1. Hawkes processes

We briefly discuss temporal point processes, a set of mathematical models for discrete events randomly arriving over time. Temporal point processes capture the time of an upcoming event given the times of all previous events via the so-called *conditional intensity function* (or simply *intensity*)  $\lambda^*(t)$ . Mathematically,

$$\lambda^*(t)dt = \text{P}(\text{event in } [t, t + dt] | \mathcal{H}_t), \quad (1)$$

where  $\mathcal{H}_t$  represents the event history up to (but excluding) time  $t$ . Dividing Eq. (1) by  $dt$  (in the Leibniz notational sense), we see  $\lambda^*(t)$  equals the conditional probability of an event in an interval of (infinitesimal) length  $dt$  per such interval  $dt$ . Hence, we interpret  $\lambda^*(t)$  as a history-dependent *event rate*. Such temporal point processes are often termed doubly-stochastic [14], as events occur randomly over time, and the event model  $\lambda^*(t)$  is a random process too. We also note  $\lambda^*(t)$  characterizes temporal point processes as counting processes  $N(t)$  for the number of events up to time  $t$ .

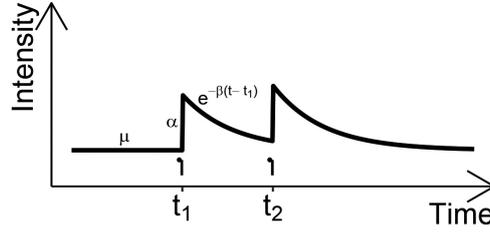


Fig. 2. Illustration of Hawkes process intensity  $\lambda^*$ . This two event sample from a Hawkes process exemplifies how its intensity  $\lambda^*$  changes over time. Starting with a minimal level of intensity  $\mu$ ,  $\lambda^*$  jumps by a constant  $\alpha$  at each event time  $t_i$ , and then decays exponentially at rate  $\beta$  over time.

Hawkes processes [24] assume  $\lambda^*(t)$  follows a certain functional form. Specifically, given a Hawkes process *realization*, i.e., a set of  $n$  events occurring at times  $t_i \in \mathbb{R}^+$ , the conditional intensity of a Hawkes process is

$$\lambda^*(t) = \mu + \alpha \sum_{t_i \in \mathcal{H}(t)} \kappa_\beta(t - t_i), \quad (2)$$

where  $\mu \in \mathbb{R}^+$  is the *baseline* intensity and  $\alpha \in \mathbb{R}^+$  the *self-excitation*, i.e., magnitude of an increase in  $\lambda^*(t)$  at each event time  $t_i$ . Immediately after each  $t_i$ , the intensity decreases according to the kernel  $\kappa_\beta$ . A common choice [17,18,40,46,48,52] for the kernel is an exponential function parametrized by the *decay* rate  $\beta$ , i.e.,  $\kappa_\beta(t) = e^{-\beta t}$ ,  $\beta \in \mathbb{R}^+$ . Plugging this kernel in Eq. (2) we obtain

$$\lambda^*(t) = \mu + \alpha \sum_{t_i \in \mathcal{H}(t)} e^{-\beta(t-t_i)}, \quad (3)$$

which we illustrate in Fig. 2.

Multivariate Hawkes processes with an exponential kernel generalize univariate ones by introducing parameters for self-excitation and for the decay per dimension. Beyond self-excitation, they also capture *cross-excitation*, the intensity jump an event in one dimension causes in another. Formally, the intensity of dimension  $p$  of an  $M$ -variate Hawkes process is

$$\lambda^{*p}(t) = \mu_p + \sum_{q=1}^M \sum_{t_i^q < t} \alpha_{pq} e^{-\beta_{pq}(t-t_i^q)}. \quad (4)$$

Notice that we index each dimension's intensity function with  $\lambda^{*p}$ , and its baseline with  $\mu_p$ . This generalization also includes an excitation matrix with self-excitation and cross-excitation entries  $\alpha_{pp}$  and respectively  $\alpha_{pq}$ , as well as, analogously, a matrix of decay rates  $\beta_{pq}$ . Note that  $\alpha_{pq}$  captures the increase in intensity in dimension  $p$  following an event in dimension  $q$ . In matrix notation, we write  $\boldsymbol{\mu} \in \mathbb{R}^M$ ,  $\boldsymbol{\alpha} \in \mathbb{R}^{M \times M}$  and  $\boldsymbol{\beta} \in \mathbb{R}^{M \times M}$ . We point to the work by Linniger [32] for more on multivariate Hawkes process theory.

We now introduce the notions of *stationarity* and *causality* in the multi-dimensional Hawkes process context. Stationarity implies translation-invariance in the Hawkes process distribution, and, in particular, it also implies that the intensity does not increase indefinitely over time and therefore stays within bounds. More formally, a Hawkes process with an exponential kernel is stationary if the spectral radius  $\rho$ , i.e., the largest eigenvalue of the  $L^1$ -norm of  $\boldsymbol{\alpha}/\boldsymbol{\beta}$ , satisfies  $\rho < 1$ . Note that assessing stationarity of a one-dimensional Hawkes process with an exponential kernel reduces to evaluating  $\alpha/\beta < 1$ .

Recent work [16] established a connection between Granger causality and the excitation matrix: In the particular case of our exponential kernel, dimension  $q$  Granger-causes dimension  $p$  if and only if  $\alpha_{pq} > 0$ .

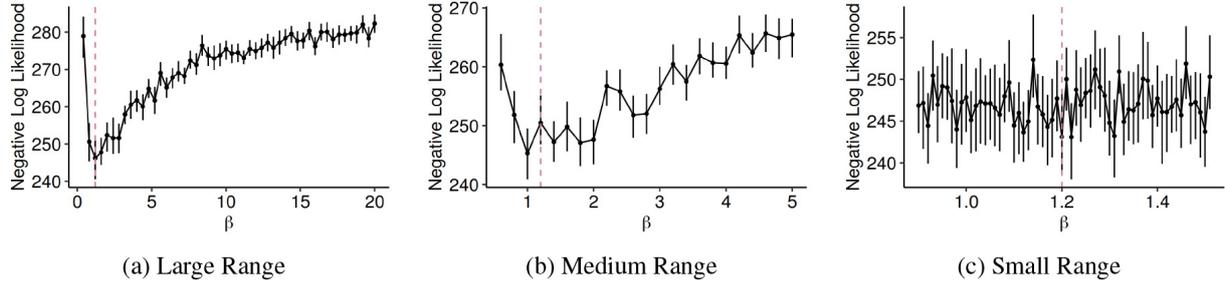


Fig. 3. The negative log-likelihood of Hawkes processes as a function of  $\beta$  is non-convex and noisy. We simulate three separate sets of 100 realizations from a one-dimensional Hawkes process with  $\beta^* = 1.2$  (marked with a vertical dashed line), and we evaluate the process' log-likelihood with one set of realizations per each of three ranges of  $\beta$  around  $\beta^*$ : a large (cf. Fig. 3a), a medium (cf. Fig. 3b) and a small (cf. Fig. 3c) range. Error bars indicate 95% confidence intervals. In Fig. 3a, it is apparent the negative log-likelihood is neither concave nor convex in  $\beta$ . Zooming in around  $\beta^*$  reveals a wide and noisy basin and, in Fig. 3b, the negative log-likelihood is not minimal at  $\beta^*$ . Both observations explain difficulties in optimizing the log-likelihood for the decay parameter. This finding is robust across Hawkes process parameter configurations satisfying the stationarity constraint.

Beyond this result, we interpret the magnitude of excitations  $\alpha_{pq}$  as the strength and direction of temporal dependency between dimensions  $p$  and  $q$ : For example, we say dimension  $q$  influences dimension  $p$  more strongly if and only if  $\alpha_{pq} > \alpha_{qp}$ .

Finally, we define the Hawkes process likelihood function. We work with the log-likelihood function due to its mathematical manipulability and to avoid computational underflows. The equation for the one-dimensional log-likelihood of the Hawkes process with an exponential kernel is as follows [35]:

$$\begin{aligned} \log L(\{t_i\}_{i=1}^n) = & -\mu t_n - \frac{\alpha}{\beta} \sum_{i=1}^n (1 - e^{-\beta(t-t_i)}) \\ & + \sum_{i=1}^n \log(\mu + \alpha A(i)), \end{aligned} \quad (5)$$

with  $A(1) = 0$  and, for  $i > 1$ ,  $A(i) = \sum_{t_j < t_i} e^{-\beta(t_i - t_j)}$ .

Ozaki [35] also proposes a (computationally less intensive) recursive formulation for Eq. (5). We refer the interested reader to Daley and Vere-Jones [14] for a general formulation of the log-likelihood of multivariate temporal point processes. Henceforth, as we focus on Hawkes processes with exponential kernels, we refer to them simply as Hawkes processes.

### 3.2. Decay estimation

To learn from streams of events, practical applications start by fitting Hawkes processes, i.e., optimizing the log-likelihood given in Eq. (5) to a set of event timestamps. Practitioners then inspect fitted parameters to understand inherent temporal dependencies, and perform downstream tasks such as prediction via simulation of the fitted processes [29,40]. As inferring and interpreting (all) fitted Hawkes process parameters is crucial in many real-world applications [5,23,28,34,40,44,48], we turn our attention to the challenges in fitting Hawkes process parameters, and especially, in estimating the decay parameter in the exponential kernel. Previous research has shown that the baseline  $\mu$  and excitation jump  $\alpha$  can be efficiently computed since the log-likelihood is amenable for convex optimization of these parameters [5,18]. However, that does not hold for the decay  $\beta$ , neither in the univariate nor in the multivariate case.

Previous work suggested a wide range of methods to address the decay estimation problem with approaches that provide point estimates. These approaches include setting  $\beta$  to a given constant value [3,12,18], cross-validation over a range of values [11,17,39], or the application of a general optimization method. Those methods comprise non-linear optimization [13,35], Bayesian hyperparameter optimization [19,40], expectation-maximization [29,47] or visual inspection of the log-likelihood function [4,48]. While these methods suffice for obtaining a point estimate with comparable errors, none of them explicitly addresses uncertainty quantification.

We argue that aforementioned problems in estimating decay parameters (cf. Fig. 1) are caused by a noisy, non-convex log-likelihood in  $\beta$ . We illustrate these properties of the log-likelihood with the following exemplary experiment. We consider a univariate Hawkes process with  $\mu = 0.1$ ,  $\alpha = 0.5$  and  $\beta^* = 1.2$  (the same as in Fig. 1), and we then compute the negative log-likelihood for different values of  $\beta$ . We generate three sets of 100 realizations from that Hawkes process. In Fig. 3, we evaluate the negative log-likelihood with one set of realizations per each of three ranges of  $\beta$  around  $\beta^*$ , namely a large (cf. Fig. 3a), a medium (cf. Fig. 3b) and a small (cf. Fig. 3c) range. In the large range, it appears there is a convex basin around  $\beta^*$  (which we annotate with a vertical dashed line), but this function's shape shifts to a concave curve with increasing decay values. The function then converges on the right, as  $\lim_{\beta \rightarrow +\infty} \log L = -\mu t_n + n \log(\mu)$ . Inspecting the seemingly "convex" region more closely uncovers a wide and noisy basin around  $\beta^*$ , where  $\beta^*$  does not always feature minimal negative log-likelihood (cf. Fig. 3b). This explains difficulties in obtaining precise estimations regardless of the optimization strategy. We repeated this experiment with a wide range of parameter values corresponding to stationary Hawkes processes. We confirm that these observations are robust to all such alternative configurations.

#### 4. Bayesian decay estimation

In this section, we present a novel Bayesian approach for the decay parameter estimation in Hawkes process. We begin with the overall goals and the intuition behind our approach before formally introducing the approach and its operationalization. We also provide a short reflection on a frequentist alternative.

##### *Goals*

The aim of our approach is threefold. Firstly, we try to explicitly quantify  $\beta$  estimation uncertainty and the magnitude of potential estimation errors across  $\beta$  fitting approaches (i.e., point estimates). Secondly, considering previous use cases [34] that involve encoding and validating hypotheses on the decay value, we aim to systematically diagnose such hypotheses. Thirdly, current methods return only decay values which fulfill the stationarity constraint  $\rho < 1$ . However, previous work [36,40] studied applications with non-stationary changes such as exponential growth and exogenous shocks. Hence, we see an opportunity for an extension of current decay fitting methods to address potential violations of the stationarity assumption.

##### *Intuition*

To that end, we propose a parsimonious Bayesian inference procedure for encoding and validating hypotheses on likely values for  $\beta$ . In our Bayesian approach, we sequentially consider a series of univariate Hawkes process realizations one by one. With each subsequent realization, we fit  $\beta$  with a given optimization method and obtain an ever-increasing set of estimated decay values that we denote

as  $\{\hat{\beta}\}_{k=1}^K$ .<sup>2</sup> After this collection of  $\hat{\beta}$ , we apply Bayes' theorem to make inferences about the true  $\beta$ . Algorithm 1 outlines the individual steps of our approach.

---

**Algorithm 1:** Bayesian Decay Estimation

---

**Input** : Stream of  $k$  Hawkes process realizations, model for hypothesis on decay parameter  $\mathcal{M}_H$ , Hawkes process decay fitting algorithm `FIT_DECAY`  
**Output**: Posterior statistics on the Hawkes process decay parameter  $\beta$

- 1 Initialization: list of realizations  $R \leftarrow \emptyset$ , list of fitted decay parameters  $D \leftarrow \emptyset$ ;
- 2 **for** incoming  $k$ -th realization with timestamps  $\{t_1, t_2, \dots, t_n\}$  and  $k \leftarrow 1, 2, 3, \dots$ , **do**
- 3      $R \leftarrow \text{APPEND}(R, \{t_1, t_2, \dots, t_n\})$ ;
- 4      $\hat{\beta} \leftarrow \text{FIT\_DECAY}(R)$ ;
- 5      $D \leftarrow \text{APPEND}(D, \hat{\beta})$ ;
- 6      $\text{UPDATE\_POSTERIOR}(D, \mathcal{M}_H)$ ;
- 7 **end**
- 8 Output posterior (or posterior predictive) statistics;                     // e.g. 95% credible interval for true  $\beta$

---

A key difference between our approach and typical applications of Bayesian inference can intuitively be described as follows: The classical Bayesian inference setup typically places a prior distribution for an unknown parameter of interest in a probability distribution which captures the likelihood of given data. Then, applying Bayes' theorem allows for inferring likely values of the unknown parameter given the data. In our approach, we assume the unknown parameter, i.e., the decay parameter, is also the data, which consists of the aforementioned sequence of decay parameter estimations. This setup enables the freedom to choose between computing posterior (i.e., in parameter space) or posterior predictive (i.e., in data space) distributions to learn about the decay. We find that this flexibility is useful and can improve performance in applications.

### Formalization

Given data  $\{\hat{\beta}\}_{k=1}^K$  and a model  $\mathcal{M}_H$  parametrized by hypothesis  $H$  for the parameter of interest  $\beta$ , we propose computing the Bayesian posterior

$$P(\beta | \{\hat{\beta}\}_{k=1}^K, \mathcal{M}_H) \propto P(\{\hat{\beta}\}_{k=1}^K | \beta, \mathcal{M}_H) P(\beta | \mathcal{M}_H), \quad (6)$$

where  $P(\{\hat{\beta}\}_{k=1}^K | \beta, \mathcal{M}_H)$  is the likelihood and  $P(\beta | \mathcal{M}_H)$  the prior encoding a hypothesis. Again, note the accordance between the model parameter and the data in our Bayesian approach:  $\beta$  is the model parameter and, at the same time, our data  $\{\hat{\beta}\}_{k=1}^K$  contains the estimates of that same parameter. Having derived or estimated a posterior distribution density, we have multiple inference possibilities for  $\beta$ . In particular, we can (i) obtain a point estimate for  $\beta$  directly from the posterior (e.g., mean, median, or maximum), or (ii) compute a new estimate  $\tilde{\beta}$  or a statistic (e.g., mean) from the posterior predictive distribution, which is given by (omitting for simplicity  $\mathcal{M}_H$ ):

$$P(\tilde{\beta} | \{\hat{\beta}\}_{k=1}^K) = \int P(\tilde{\beta} | \beta, \{\hat{\beta}\}_{k=1}^K) P(\beta | \{\hat{\beta}\}_{k=1}^K) d\beta. \quad (7)$$

---

<sup>2</sup>Note that the set  $\{\hat{\beta}\}$  does not contain independent observations (as e.g. the realizations used to obtain  $\{\hat{\beta}\}_{k=1}^2$  are also in  $\{\hat{\beta}\}_{k=1}^3$ ). Anticipating that practitioners use as much data as available, we consider  $\{\hat{\beta}\}$  as previously defined. However, repeating all our experiments using only iid  $\hat{\beta}_k$  (i.e., each fitted only on a single realization), we obtain similar though noisier results.

We can also derive credible intervals – roughly Bayesian counterparts of the frequentist confidence interval – for such statistics.

### Operationalization

Thus, this setup leaves us with the task of deciding on appropriate distributions for the prior and the likelihood, as well as mapping hypotheses to specific parametrizations of the prior. While we are, in general, free to choose from the broad spectrum of available techniques for Bayesian inference, we propose to assume the likelihood  $P(\{\hat{\beta}\}_{k=1}^K | \beta) \sim \text{Exponential}(\lambda)$  and a conjugate prior, namely  $P(\beta) \sim \text{Gamma}(a_0, b_0)$ . These assumptions lead to the following practical advantages: (i) there is a closed form solution to the posterior predictive density mean,  $\beta' = E[\tilde{\beta} | \{\hat{\beta}\}_{k=1}^K] = (b_0 + \sum_{k=1}^K \hat{\beta}_k) / (a_0 - 1)$ , and (ii) users of this approach only have to encode hypotheses on  $\beta$  as the prior's  $b_0$  parameter, as they can set  $a_0$  simply to  $K$ . We note that assuming other distributions of the exponential family such as Pareto for the log-likelihood and the conjugate prior is also a valid choice,<sup>3</sup> for which we experimentally obtain equivalent results. Note that our choice of a Bayesian setup also allows for complex inference setups, which require the application of e.g. Markov chain Monte Carlo (MCMC). Also, in our approach, we restrict Bayesian inference to the decay and estimate confidence intervals for other parameters via (frequentist) bootstrap. This enables the estimation of other parameters via more efficient convex optimization routines. We also remark that our approach is agnostic to the choice of decay estimation method, and, in our experiments, we illustrate its application with several of the methods we mentioned in Section 3.2.

So far, we focused the presentation of our approach on univariate Hawkes processes. To generalize to the multivariate case, we set  $\beta_{pq} = \beta \forall p, q$ , a common simplification of the decay estimation problem [17, 39, 40, 44], and we then proceed as previously. Alternatively, optimization of  $\beta_{pq}$  is also possible, but with a quadratic computational complexity in the number of dimensions.

### A frequentist alternative

Finally, we reflect on a bootstrap-based alternative formulation of our Bayesian approach. Repeatedly resampling  $\{\hat{\beta}\}_{k=1}^K$  with replacement yields the bootstrap distribution of decay estimates. Testing hypotheses on the decay value then amounts to comparing a statistic of the bootstrap distribution (such as a mean or a percentile)<sup>4</sup> with the hypothesized value. While it is also possible to extend this bootstrap-based procedure to more complex, intractable inferential setups, the computational cost may surpass that of Bayesian inference (as we discuss in the context of an experiment). In our setting of sequential collection of realizations, Bayesian inference becomes a natural choice to update beliefs about true decay values (cf. Chapter 3.4 of Efron and Hastie [15]), especially in the case of a small number of realizations. Further, as the bootstrap may be interpreted as a nonparametric, noninformative posterior approximation (cf. Chapter 8.4 of Hastie et al. [22]), Bayesian inference with informative priors may, again, lead to lower computational costs in the presence of few observations.

## 5. Experiments

We illustrate experimentally how our approach allows for quantifying the uncertainty in the decay and other Hawkes process parameter values, diagnosing mis-estimation and addressing breaks in stationarity.

<sup>3</sup>We suggest choosing likelihood distributions with positive support as  $\beta > 0$ .

<sup>4</sup>We recommend using empirical bootstrap statistics to correct for sample bias.

Table 1  
Dataset characteristics per goal

Goal	Dataset type	Number of realizations	Mean number of events per realization
Quantify Uncertainty	Synthetic	100	2140
	Real-World	4	19
Diagnose Mis-Estimation	Synthetic	100	100
	Real-World	56	10
Address Stationarity Breaks	Synthetic	100	100
	Real-World	410	22

For each of those three goals, we illustrate how our approach achieves the goal (i) with a synthetic dataset and (ii) within a real-world application. We summarize dataset characteristics for each of those goals in Table 1. Overall, we demonstrate that, besides achieving the previously mentioned goals, our approach is broadly applicable across practical scenarios.

### 5.1. Quantifying the uncertainty

We begin by addressing the problem of quantifying the uncertainty of fitted decay values  $\hat{\beta}$ , as well as potential consequences of mis-estimation on other Hawkes process parameters. One prominent application of multi-dimensional Hawkes processes consists in the estimation of directions of temporal dependency, e.g., when studying influence in online communities [28,40], or in approximating complex geographical [34] or cortical [46] relationships. Recall that inferring such relations between a pair of Hawkes process dimensions may be framed as a problem of estimating which cross-excitation between the two dimensions is higher. We demonstrate how our Bayesian approach helps in quantifying the uncertainty in such inferred relationships and how to estimate the impact of potential errors with synthetic and real-world data.

#### Synthetic data

We consider a two dimensional Hawkes process with parameters  $\boldsymbol{\mu} = \begin{pmatrix} 0.1 \\ 0.5 \end{pmatrix}$ ,  $\boldsymbol{\alpha} = \begin{pmatrix} 0.1 & \alpha_{12} \\ \alpha_{21} & 0.2 \end{pmatrix}$ , and  $\beta_{pq} = \beta \forall_{1 \leq p, q \leq 2}$ . We assume  $\beta = 1.2$ . For the cross-excitation parameters, we set  $\alpha_{21} = 0.7$  and, successively,  $\alpha_{12} = \alpha_{21} * c$  for a range of 10 linearly spaced values of  $c \in [0.75, 1.25]$ . This implies that each configuration encodes a different direction and strength of influence, where dimension 2 more strongly influences dimension 1 for  $c < 1$ , and vice-versa for  $c > 1$ . For each such configuration, we simulate  $K = 100$  realizations with a stopping time of  $T = 1000$  (corresponding to a mean number of events per realization of 2140). As the decay estimation approach we apply a non-linear optimization routine L-BFGS-B [10] after the arrival of each such realization to obtain  $\{\hat{\beta}\}_{k=1}^{100}$  per configuration. Using our closed-form Bayesian inference approach, we hypothesize that  $\beta$  equals 1.5 by setting  $b_0 = 1.5$  in the previously described Gamma prior. We then perform the aforementioned Bayesian inference on each set of  $\{\hat{\beta}\}_{k=1}^{100}$  to derive  $\beta'_{0.025}$  and  $\beta'_{0.975}$ , the lower and upper bounds of the 95% credible interval of the posterior predictive density. For 100 linearly spaced values of the decay in  $[\beta'_{0.025}, \beta'_{0.975}]$ , we fit the remaining Hawkes process parameters (i.e. baseline and excitations) and check the accuracy of the inferred influence direction between dimensions. The accuracy captures how many of the 100 decay values lead to correct recovery of the relation between  $\alpha_{12}$  and  $\alpha_{21}$ . We bootstrap 95% confidence intervals<sup>5</sup> for the accuracy directly from that distribution of 100 decay values. This procedure provides (i)

<sup>5</sup>We report bootstrap statistics here to refrain from unnecessarily extending our Bayesian approach to the other Hawkes process parameters, which, as previously mentioned, should be estimated with convex optimization routines.

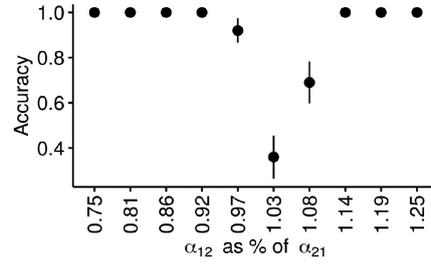


Fig. 4. Quantifying uncertainty when inferring directions of influence. We first fit  $\beta$  on realizations from two-dimensional Hawkes processes with cross-excitation  $\alpha_{12}$  varying from 75% to 125% of  $\alpha_{21}$ . We apply closed-form Bayesian inference to estimate the uncertainty in fitted decay as the posterior predictive 95% credible interval. For a set of decay values in that interval, we estimate the other parameters, and measure the accuracy in recovering the encoded influence direction between dimensions 1 and 2. The accuracy is low and features larger error bars when  $\alpha_{12}$  is close to  $\alpha_{21}$ , where many decay values in the 95% credible intervals lead to wrong estimations of the influence direction.

an estimate of the uncertainty of the fitted decay value via the 95% credible interval, (ii) an estimation of the robustness of the temporal dependency between dimensions 1 and 2, and (iii) empirical evidence for the consequences of fitting Hawkes processes with misaligned  $\beta$ . Beyond the settings presented here, we also experimented with slightly different parametrizations of the Hawkes process and alternative decay fitting approaches such as the Bayesian hyperparameter optimization Hyperopt [8]. In all such settings, we obtained qualitatively the same results.

Figure 4 summarizes the outcome of this experiment. As expected, we observe lower accuracies in inferring the direction of influence between dimension 1 and dimension 2 for  $\alpha_{12}$  close to  $\alpha_{21}$ . This implies that many decay values in the 95% credible interval of such configurations lead to mis-estimations of the direction of influence, and the large error bars reflect this as well. Overall, we believe practitioners may leverage this approach to quantify the uncertainty of the decay estimation and consequently of the inferred directions of influence as well.

#### Earthquakes and aftershocks

We illustrate the outlined uncertainty quantification procedure with a dataset of earthquakes in the Japanese regions of Hida and Kwanto, as initially studied by Ogata et al. [34]. We consider the data listed in Table 6 of that manuscript, i.e., a dataset of 77 earthquakes from 1924 to 1974. We employ the decay value listed in Table 5 of that manuscript as the prior's parameter in our closed-form Bayesian approach.<sup>6</sup> We assume a two-dimensional Hawkes process with a single  $\beta$  value, where the dimensions represent earthquakes in the Japanese regions of Hida and Kwanto. As pre-processing, we split earthquake occurrences into  $K = 4$  equally sized segments which we treat as process realizations, and we convert the event timescale to decades.

We replicate the seismological relationship that earthquakes in the Japanese Hida region precede those in Kwanto, as we obtain  $\alpha_{\text{Kwanto Hida}} > \alpha_{\text{Hida Kwanto}} = 0$ . Our Bayesian estimation procedure yields the posterior predictive density mean of  $\beta' \approx 31.71$ , which corresponds to an intensity half-life of about  $\log(2)/31.71 \approx 0.02$  decades. The inferred relationship between Hida and Kwanto is present for all but one value of the 95% credible interval for  $\beta'$ , the lower extremity  $\beta'_{0.025}$ . These results underscore the low uncertainty of the inferred direction of influence.

<sup>6</sup>We choose that prior for demonstration purposes, as that decay value was estimated with a different Hawkes process from the one we study in this work.

## 5.2. Mis-estimation & misaligned hypotheses

In this section, we demonstrate that our Bayesian approach facilitates diagnosing (inevitable) estimation errors and misaligned hypotheses as over- or under-estimates, as well as the magnitude of that error. Hence, we address a need, which previous work [4,34,48] implies, to encode, validate and diagnose estimations and hypotheses on the decay parameter value. Again, we illustrate how our approach meets that need with synthetic and real-world data.

### *Synthetic data*

We consider a univariate Hawkes process with parameters  $\mu = 1.2$ ,  $\alpha = 0.6$  and  $\beta = 0.8$  (comparable choices of parameters lead to the same qualitative results). Using this parametrization, we generate  $K = 100$  realizations with 100 events each and we estimate decay after the arrival of each individual realization to obtain  $\{\hat{\beta}\}_{k=1}^{100}$  per approach. To illustrate that our Bayesian approach is agnostic to the choice of decay estimation method, we apply the decay estimation approaches L-BFGS-B [10], Grid Search (across 10 evenly distributed values on a log scale in  $[-1, 2]$ , similarly to Salehi et al. [39]) and Hyperopt [8]. Using our closed-form Bayesian inference approach, we leverage a Gamma prior with  $b_0 = 1$ , an estimate slightly larger than true  $\beta$ , to illustrate our approach. We then perform Bayesian inference on each set of  $\{\hat{\beta}\}_{k=1}^{100}$  and compare the RMSE (root mean squared error) between the resulting  $\beta'$  estimates and  $\beta$ . We repeat this whole process 100 times to derive uncertainty per fitting method via the Bayesian bootstrap distribution.

We observe that L-BFGS-B and Hyperopt return the decay estimates with lowest RMSE, while Grid Search performs remarkably worse. Aiming to reduce RMSE across all approaches, we inspect the estimates more closely. Looking into  $\hat{\beta}_k$  reveals that they are consistently below our hypothesized  $b_0$ . We suspect this discrepancy arises due to our prior parameter value  $b_0 = 1$ , which is an over-estimate. Therefore, we compute the difference between that  $b_0$  and the posterior predictive density mean  $\beta'$  per fitting approach. Except for Grid Search ( $b_0 - \beta' = -0.002 \pm 0.013$ ), the approaches boast positive differences (Hyperopt and L-BFGS-B both yield  $b_0 - \beta' = 0.183 \pm 0.006$ ), which imply the prior parameter is larger than the posterior predictive density mean. The direction and magnitude of these shifts of the posterior away from the prior suggest the use of our Bayesian approach as a diagnosis tool, which correctly signals that our hypothesis likely over-estimates true  $\beta$ .

### *Vocabulary learning intensity*

We investigate a scenario proposed as future work by Settles and Meeder's [41] study of user behavior on the Duolingo language learning app: The authors speculate that vocabulary learning intensity in Duolingo correlates with word difficulty as defined by the CEFR language learning framework [33]. We complement the Duolingo data with a dataset of English-language vocabulary and its corresponding CEFR level,<sup>7</sup> and we build two groups of words: those from the easiest CEFR levels, A1 and A2 (A-level group), and those from the hardest ones, C1 and C2 (C-level group). We observe that there are 28 users with 10 vocabulary learning events in the C-level. To control for total learning events per user, we randomly sample a set of 28 users with 10 learning events in the A-level. Increasing the number of events to 11 or 12 leads to qualitatively similar results, but decreased statistical power due to smaller sample size. We repeat this random sampling for a total of 100 times to Bayesian bootstrap 95% credible intervals. We posit that each Duolingo user learning the A-level set of words represents a realization of a univariate

<sup>7</sup><http://www.englishprofile.org/american-english>.

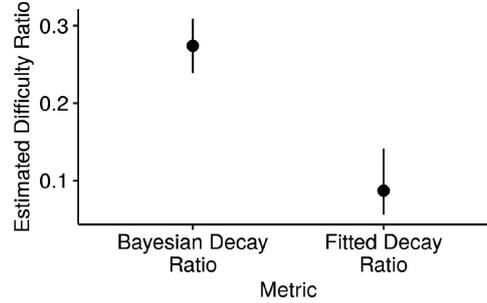


Fig. 5. In the Duolingo app, users learning C-level words have longer learning bursts than those learning A-level words. We fit two Hawkes processes to users with 10 word learning events on Duolingo: one for users studying “hard” (C-level) words, and the other for users learning “easy” (A-level) words. We posit that the decay value of the C-level process is half as large as the A-level one, and we depict the ratio of the former to the latter. The ratio of fitted values (“Fitted Decay Ratio”) is lower than that of posterior predictive density means (“Bayesian Decay Ratio”), due to a conservative prior parametrization. Lower decay values for the same number of events imply longer event bursts, suggesting that C-level words require extended learning effort.

Hawkes process, and users learning the C-level words represent realizations of another univariate Hawkes process. Mapping the six CEFR levels (A1, A2, B1, B2, C1 and C2) to a scale from 1 to 6, we naively assume, for illustration purposes, that the C-levels may be more than twice as hard as A-levels. If the data reflects that hypothesis, then we expect that a short learning burst suffices for grasping A-level words, in contrast to the C-level words, which may require perhaps more than two times as much effort over time. We encode this hypothesis in our closed-form Bayesian approach (with L-BFGS-B) as  $b_{\text{C-level}} = 1$  and  $b_{\text{A-level}} = 2$ , since, after controlling for the total event count, we interpret the former as corresponding to longer periods of higher intensity, when compared with the latter.

In Fig. 5, we depict the ratio of posterior predictive density means  $\beta_{\text{C-level}}^t / \beta_{\text{A-level}}^t$  (“Bayesian Decay Ratio”), as well the analogue ratio of the means of actual L-BFGS-B estimations for both levels (“Fitted Decay Ratio”), i.e.,  $\{\hat{\beta}_{\text{C-level}} / \hat{\beta}_{\text{A-level}}\}_{k=1}^{28}$ . Overall, we confirm Settles and Meeder’s hypothesis that word difficulty correlates positively with the effort required to learn them: The posterior (and fitted) decay values of the C-level words are lower than those of the A-level words, resulting in more prolonged learning bursts in the former vs. the latter. Moreover, we underscore that this practical example illustrates the usefulness of our approach as a diagnosis tool: We observe a moderate shift away from the hypothesized 1:2 difficulty ratio and towards a posterior predictive density mean value slightly below 3:10. This shift is even more severe according to the “Fitted Decay Ratio”, which is not influenced by a prior. This finding indicates that the CEFR language levels may not directly translate to numerical scales for quantifying learning progress. Further, the relatively small size of the user sample highlights the importance of using Bayesian inference to not only diagnose a-priori estimates, but also to explicitly surface estimation uncertainty due to a small dataset. Therefore, although we suggest caution in extrapolating our results, we believe that our observations may contribute to ongoing research on the challenges of quantifying language learning progress [26].

### 5.3. Addressing breaks in stationarity

We turn our attention to the assumption of stationarity and its effect on fitting the decay parameter of Hawkes processes. Recall that stationarity implies that the intensity of a Hawkes process is translation-invariant. In practical applications such as the study of virality of online content [36] or the growth of online communities [40], exogenous shocks or exponential growth break the stationarity assumption.

With synthetic data and a real-world example, we show how our Bayesian approach allows for assessing and capturing breaks in stationarity caused by exogenous shocks.

### *Synthetic data*

We start with the same experimental setup as in Section 5.2, but we introduce two key differences. First, we assume that there was an underlying change in  $\beta$  at some point during the  $K = 100$  realizations. We set the index of that change to  $k^* = 50$ , but our conclusions also hold for other choices of  $k^*$  (such as  $k^* \in [30, 70]$ ). For  $k < k^*$ , we simulate the Hawkes process as previously, but, for  $k \geq k^*$ , we increment  $\beta$  by 1 while keeping the other parameters unchanged, and we simulate from that updated process instead. With  $\beta_1$  we denote  $\beta$  before the change, and with  $\beta_2 = \beta_1 + 1$  the  $\beta$  afterwards. Second, we build a Bayesian inference setup to reflect the hypothesis that the  $\beta$  value changed at some point in the set of realizations. Such models are termed *changepoint models*. We propose the following intractable setup: A prior with  $b_0 = \begin{cases} b_1, & k < \kappa \\ b_2, & k \geq \kappa \end{cases}$ , where we set  $b_1 \sim \text{Exponential}(1)$ ,  $b_2 \sim \text{Exponential}(0.7)$  and  $\kappa \sim U\{1, 100\}$ , and an exponentially distributed likelihood. Note that the hypothesis  $b_1 > b_2$  contradicts the true second part of simulated realizations. As the metrics for this experiment, we first measure the RMSE between the mean (respectively median) of samples from the posterior,  $\bar{\beta}_1$  and  $\bar{\beta}_2$  (resp.  $\kappa/K$ ), and both true  $\beta$  (resp.  $k^*/K$ ). Beyond RMSE, we also assess the estimation accuracy as the relative frequency of a correctly inferred ordering  $\bar{\beta}_1 < \bar{\beta}_2$ . As before, we demonstrate that our Bayesian approach is agnostic to the choice of decay fitting methods by employing L-BFGS-B, Grid Search and Hyperopt.

As a side result, we note that L-BFGS-B and Hyperopt again outperform Grid Search with respect to RMSE. Further, we report mean accuracy values close to 1 for Hyperopt and L-BFGS-B (both  $0.98 \pm 0.014$ ) and remarkably lower values for Grid Search ( $0.65 \pm 0.047$ ). Although we encoded prior parameter values which contradict the data, we still almost always recover the correct relationship  $\bar{\beta}_1 < \bar{\beta}_2$  with the L-BFGS-B and Hyperopt methods. Qualitatively, this suggests identifying the direction of distributional changes in the decay is feasible. Quantitatively, we expect this procedure to yield conservative estimates of the magnitude and timing of the change: Bayesian inference features an “inertia” of a few realizations when updating the posterior after the shock.

### *Strength of collective effervescence*

Our third real-world scenario concerns the manifestation of collective effervescence on Twitter in response to the 13. November 2015 terrorist attacks in Paris, as studied by Garcia and Rimé [20]. They proposed future work could analyze how tweet timings reflect collective emotions surrounding the attacks. We address this suggestion by fitting the changepoint Bayesian model with the L-BFGS-B method. Specifically, we begin by extracting the timestamps of tweets by users in a two week period centered on the day of the attacks. We model each user’s behavior per week as a realization of a univariate Hawkes process, and we also control for tweeting activity per user: We extract all 205 users who tweeted between 20 and 25 times in the week before and in the week after.<sup>8</sup> Lowering the activity bounds yields more users each with less events to fit, while increasing those bounds has the opposite effect. However, by setting the activity bounds to different ranges of 5 tweets (specifically, 10 to 15, 15 to 20, . . . , or 45 to 50 tweets), we qualitatively observe the same outcomes. We hypothesize that Twitter users, who partake in the collective emotion as a reaction to the shock, feature a sustained burst of activity. We expect such a

<sup>8</sup>Note that this extraction process results in a total of 410 realizations, i.e., 205 realizations before the shock and another 205 afterwards. The mean number of events per realization is 22.

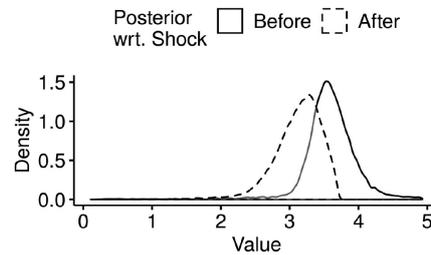


Fig. 6. Tweet timing reflects collective emotions. We depict the result of fitting an MCMC-based changepoint detection model to users' Tweet timings in the two weeks surrounding the November 2015 Paris attacks: The estimated posterior density for the Hawkes process intensity decay assigns more probability mass to higher decay regions before the shock in comparison to afterwards. This suggests that collective effervescence manifests on Tweet timings, as lower decay values after the shock reflect more sustained bursts of activity.

burst of activity to translate into a decrease of the decay value after the shock. To numerically capture this hypothesis, we simply set  $b_1 = 1.5$ ,  $b_2 = 1$ . However, repeating this experiment with an opposing hypothesis (e.g.,  $b_1 = 1$ ,  $b_2 = 1.5$ ), again leads to the same results.

Figure 6 depicts the density of the distribution of the inferred decay posterior before and after the shock. As expected, we confirm the decay value goes down in the week after the attacks, suggesting more sustained bursts of Tweeting activity in response to the attacks. This, in turn, supports the hypothesis that Garcia and Rimé advanced: Reaction timings, in the form of longer bursts of tweets afforded by a 15% lower mean posterior decay after the shock, reflect this collective emotion. We note that this is a conservative estimate of the decrease in the parameter, since activity levels quickly revert back to a baseline within the week after the attacks themselves, as Garcia and Rimé report. Further, this changepoint detection approach also over-estimates the time of the change at realization number 235, i.e., 7.1% later than the first Hawkes process realization after the shock, corresponding to realization number 206.

## 6. Discussion

We reflect on practical aspects of our Bayesian approach vs. a frequentist alternative. Bayesian and frequentist inference are two intensively discussed alternative statistical schools for estimating parameters. In recent years, Bayesian approaches received increasing appreciation by the web and machine learning communities (e.g. [7,42,51]). Beyond that body of work and theoretical considerations (cf. also Sec. 4), we outline practical differences between our Bayesian approach and a bootstrap-based frequentist alternative. Quantifying the uncertainty of decay estimates with bootstrap amounts to deriving confidence (rather than credible) intervals for the decay from the empirical bootstrap distribution. Diagnosing mis-estimates and misaligned hypotheses with the bootstrapped decay distribution is not an integral part of the frequentist inference procedure, on the contrary to our Bayesian approach, which integrates the hypothesis in the inference procedure. However, explicitly formulating one- or two-sided statistical hypotheses to be tested against the bootstrap distribution is just as viable. In the case of addressing breaks of stationarity, the bootstrap-based approach features remarkably higher complexity than our proposed Bayesian approach. While we infer all three parameters capturing the stationarity break simultaneously, the straightforward bootstrap-based alternative would require alternatively computing the distribution of one parameter while fixing the other two (as the inferred distribution is intractable). Hence, we argue that our Bayesian approach may also be the more natural choice for intractable inference. Nevertheless, we stress that the frequentist alternative we outlined is a viable alternative to our Bayesian approach.

## 7. Conclusion

In this work, we studied the problem of fitting the decay parameter of Hawkes processes with exponential kernels. The inherent difficulties we found in accurately estimating the decay value, regardless of the fitting method, relate to the noisy, non-convex shape of the Hawkes log-likelihood as a function of the decay. Further, we identified problems in quantifying uncertainty and diagnosing fitted decay values, as well as in addressing breaks of the stationarity assumption. As a solution, we proposed a parsimonious Bayesian approach. We believe our extensive evaluation of that approach across a range of synthetic and real-world examples demonstrates its broad practical use.

Optimization techniques such as constructing convex envelopes or disciplined convex-concave programming may, in the future, help to optimize the Hawkes process likelihood as a function of the decay. We also believe exploring the potential of the vast Bayesian statistics toolbox for learning more from fitted (decay) parameter values is promising future work.

## Acknowledgments

We thank David Garcia for giving access to the extended Twitter data. We also thank Roman Kern for valuable feedback on the manuscript. Tiago Santos was a recipient of a DOC Fellowship of the Austrian Academy of Sciences at the Institute of Interactive Systems and Data Science of the Graz University of Technology.

## References

- [1] S. Aghababaei and M. Makrehchi, Mining twitter data for crime trend prediction, *Intelligent Data Analysis*, 2018.
- [2] Y. Ait-Sahalia, J. Cacho-Diaz and R. Laeven, Modeling financial contagion using mutually exciting jump processes, *Journal of Financial Economics*, 2015.
- [3] E. Bacry, M. Bompain, P. Deegan, S. Gaïffas and S. Poulsen, Tick: a python library for statistical learning, with an emphasis on hawkes processes and time-dependent models, *The Journal of Machine Learning Research*, 2018.
- [4] E. Bacry, S. Gaïffas, I. Mastromatteo and J.-F. Muzy, Mean-field inference of hawkes point processes, *Journal of Physics A: Mathematical and Theoretical*, 2016.
- [5] E. Bacry, I. Mastromatteo and J.-F. Muzy, Hawkes processes in finance, *Market Microstructure and Liquidity*, 2015.
- [6] A.-L. Barabasi, The origin of bursts and heavy tails in human dynamics, *Nature*, 2005.
- [7] D. Barber, *Bayesian reasoning and machine learning*, Cambridge University Press, 2012.
- [8] J. Bergstra, R. Bardenet, Y. Bengio and B. Kégl, Algorithms for hyper-parameter optimization, in: *NeurIPS*, 2011.
- [9] K. Budhathoki and J. Vreeken, Causal inference on event sequences, in: *SDM*, 2018.
- [10] R. Byrd, P. Lu, J. Nocedal and C. Zhu, A limited memory algorithm for bound constrained optimization, *SIAM Scientific Computing*, 1995.
- [11] E. Choi, N. Du, R. Chen, L. Song and J. Sun, Constructing disease network and temporal progression model via context-sensitive hawkes process, in: *ICDM*, 2015.
- [12] J. Choudhari, A. Dasgupta, I. Bhattacharya and S. Bedathur, Discovering topical interactions in text-based cascades using hidden markov hawkes processes, in: *ICDM*, 2018.
- [13] J. Da Fonseca and R. Zaatour, Hawkes process: Fast calibration, application to trade clustering, and diffusive limit, *Journal of Futures Markets*, 2014.
- [14] D. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes: Elementary Theory of Point Processes*, Springer, 2003.
- [15] B. Efron and T. Hastie, *Computer age statistical inference*, CUP, 2016.
- [16] J. Etesami, N. Kiyavash, K. Zhang and K. Singhal, Learning network of multivariate hawkes processes: A time series approach, in: *UAI*, 2016.
- [17] M. Farajtabar, N. Du, M. Gomez-Rodriguez, I. Valera, H. Zha and L. Song, Shaping social activity by incentivizing users, in: *NIPS*, 2014.

- [18] M. Farajtabar, Y. Wang, M. Gomez-Rodriguez, S. Li, H. Zha and L. Song, Coevolve: A joint point process model for information diffusion and network co-evolution, in: *NeurIPS*, 2015.
- [19] F. Figueiredo, G. Borges, P. de Melo and R. Assunção, Fast estimation of causal interactions using wold processes, in: *NeurIPS*, 2018.
- [20] D. Garcia and B. Rimé, Collective emotions and social resilience in the digital traces after a terrorist attack, *Psychological Science*, 2019.
- [21] M. Gupta, J. Gao, C. Aggarwal and J. Han, Outlier detection for temporal data: A survey, *IEEE Transactions on Knowledge and data Engineering*, 2013.
- [22] T. Hastie, R. Tibshirani and J. Friedman, *The elements of statistical learning: Data mining, inference, and prediction*, Springer Science & Business Media, 2009.
- [23] T. Hatt and S. Feuerriegel, Early detection of user exits from clickstream data: A markov modulated marked point process model, in: *WWW*, 2020.
- [24] A. Hawkes, Spectra of some self-exciting and mutually exciting point processes, *Biometrika*, 1971.
- [25] S. Hosseini, A. Khodadadi, A. Arabzadeh and H. Rabiee, Hnp3: A hierarchical nonparametric point process for modeling content diffusion over social media, in: *ICDM*, 2016.
- [26] J. Hulstijn, The shaky ground beneath the cefr: Quantitative and qualitative dimensions of language proficiency, *The Modern Language Journal*, 2007.
- [27] R. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, OTexts, 2018.
- [28] R. Junuthula, M. Haghdan, K. Xu and V. Devabhaktuni, The block point process model for continuous-time event-based dynamic networks, in: *WWW*, 2019.
- [29] T. Kurashima, T. Althoff and J. Leskovec, Modeling interdependent and periodic real-world action sequences, in: *WWW*, 2018.
- [30] H. Li and Y. Yu, Detecting a multigranularity event in an unequal interval time series based on self-adaptive segmenting, *Intelligent Data Analysis*, 2021.
- [31] J. Lin, E. Keogh, S. Lonardi and B. Chiu, A symbolic representation of time series, with implications for streaming algorithms, in: *SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, 2003.
- [32] T. Liniger, *Multivariate Hawkes Processes*, PhD thesis, ETH, 2009.
- [33] C. of Europe, *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Cambridge University Press, 2001.
- [34] Y. Ogata, H. Akaike and K. Katsura, The application of linear intensity models to the investigation of causal relations between a point process and another stochastic process, *Annals of the Institute of Statistical Mathematics*, 1982.
- [35] T. Ozaki, Maximum likelihood estimation of hawkes' self-exciting point processes, *Annals of the Institute of Statistical Mathematics*, 1979.
- [36] M.-A. Rizoïu, L. Xie, S. Sanner, M. Cebrian, H. Yu and P. Van Hentenryck, Expecting to be hip: Hawkes intensity processes for social media popularity, in: *WWW*, 2017.
- [37] C. Rocha and F. Rodrigues, Forecasting emergency department admissions, *Journal of Intelligent Information Systems*, 2021.
- [38] A. Rossi, B. De Souza, C. Soares, A. de Leon Ferreira de Carvalho and C. Ponce, A guidance of data stream characterization for meta-learning, *Intelligent Data Analysis*, 2017.
- [39] F. Salehi, W. Trouleau, M. Grossglauser and P. Thiran, Learning hawkes processes from a handful of events, in: *NeurIPS*, 2019.
- [40] T. Santos, S. Walk, R. Kern, M. Strohmaier and D. Helic, Self- and cross-excitation in stack exchange question & answer communities, in: *WWW*, 2019.
- [41] B. Settles and B. Meeder, A trainable spaced repetition model for language learning, in: *ACL*, 2016.
- [42] P. Singer, D. Helic, A. Hotho and M. Strohmaier, Hyptrails: A bayesian approach for comparing hypotheses about human trails on the web, in: *Proceedings of the 24th International Conference on World Wide Web*, 2015, pp. 1003–1013.
- [43] B. Tabibian, U. Upadhyay, A. De, A. Zarezade, B. Schölkopf and M. Gomez-Rodriguez, Enhancing human learning via spaced repetition optimization, *PNAS*, 2019.
- [44] B. Tabibian, I. Valera, M. Farajtabar, L. Song, B. Schölkopf and M. Gomez-Rodriguez, Distilling information reliability and source trustworthiness from digital traces, in: *WWW*, 2017.
- [45] L. Tai and C. Hsu, On the profitability and errors of predicted prices from deep learning via program trading, *Intelligent Data Analysis*, 2020.
- [46] W. Trouleau, J. Etesami, M. Grossglauser, N. Kiyavash and P. Thiran, Learning hawkes processes under synchronization noise, in: *ICML*, 2019.
- [47] A.C. Türkmen, hawkeslib, [github.com/canerturkmen/hawkeslib](https://github.com/canerturkmen/hawkeslib), 2019.
- [48] U. Upadhyay, I. Valera and M. Gomez-Rodriguez, Uncovering the dynamics of crowdlearning and the value of knowledge, in: *WSDM*, 2017.

- [49] X. Wang, A. Mueen, H. Ding, G. Trajcevski, P. Scheuermann and E. Keogh, Experimental comparison of representation methods and distance measures for time series data, *Data Mining and Knowledge Discovery*, 2013.
- [50] C. Xiao, K. Xia, X. Liu, Y. Zhang, W. Huo and N. Aslam, Improving session-based temporal recommendation by using dynamic clustering, *Intelligent Data Analysis*, 2017.
- [51] X. Yang, Y. Guo and Y. Liu, Bayesian-inference-based recommendation in online social networks, *IEEE Transactions on Parallel and Distributed Systems* **24**(4) (2012), 642–651.
- [52] K. Zhou, H. Zha and L. Song, Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes, in: *AISTATS*, 2013.
- [53] K. Zhou, H. Zha and L. Song, Learning Triggering Kernels for Multi-Dimensional Hawkes Processes, in: *ICML*, 2013.