THE UNIVERSITY *of* EDINBURGH

Modeling and Applications of Discrete-Time Hawkes Processes: Flexible and Scalable Methods

Trinnhallen Brisley

Doctor of Philosophy

THE UNIVERSITY OF EDINBURGH

2025

Abstract

Many count time series show short-term clustering: after one event, more tend to follow soon after. This thesis develops fast and flexible methods for modeling such self-exciting behavior in data observed on a regular time grid using discrete-time Hawkes models. Our contributions are twofold: first, we derive linear-time algorithms for the log-likelihood and its gradient, which enable efficient estimation for long multivariate sequences. We also build a marked, multivariate model for operational risk in a forensic psychiatric hospital, with an alarm mark for rare severe incidents that prompt a hospital-wide response; in that setting the model improves prediction and yields management-relevant risk signals. Second, we propose the Gaussian Process Discrete Hawkes Process (GP-DHP), a Bayesian nonparametric model that places independent Gaussian process priors on the baseline intensity and the excitation kernel. A collapsed representation over the latent additive intensity supports scalable maximum a posteriori estimation with complexity proportional to the number of observations, and a post hoc decomposition recovers smooth, data-adaptive baseline and excitation functions that separate exogenous background from endogenous feedback. We evaluate the methods on synthetic data and on two applications: weekly Cryptosporidiosis counts and U.S. terrorism incidents. Across these studies GP-DHP outperforms parametric discrete Hawkes baselines in predictive accuracy while revealing a flexible excitation function and seasonal structure. The overall result is a practical toolkit for discrete-time self-excitation that is both scalable and interpretable.

Lay Summary

Everyday life is full of chain reactions: one event can set off another. A violent event in one ward of a psychiatric hospital may lead to further incidents nearby; a news story can spark a flurry of social media posts online; a small outbreak of illness can grow before it fades. This thesis is about recognising, measuring, and forecasting these "after-effects" in data that are recorded as simple counts over time. The work uses a family of models sometimes called *self-exciting* processes.

Why this matters. If we can tell when today's events make tomorrow's events more likely, we can act sooner and better. Hospitals can plan staffing and de-escalation strategies; public-health teams can spot and respond to unusual increases; security analysts can monitor rising risk. In practice, however, existing tools are often too slow for long datasets or too rigid to capture real-world patterns.

What this thesis contributes. The work makes progress on two fronts.

- **Speed.** It introduces step-by-step update rules that let a computer keep track of "what the past implies now" without re-reading the entire history at every time point. In plain terms, calculations that used to slow down as timelines got longer can now be done in near linear time, even for many interacting series. This unlocks routine fitting, testing, and forecasting on long sequences.
- **Flexibility.** It develops a new, more adaptable model that lets the data learn two things at once: (i) the *background level*-how many events we would expect even if nothing had just happened; and (ii) the *after-effect shape*-how much, and for how long, one event tends to raise the chance of another. Instead of hard-wiring a single mathematical formula, the model learns smooth functions for both parts, then cleanly separates "background" from "knock-on effects" for interpretation.

What we studied. The methods are tested on synthetic data and on three real applications that typify decisions made from counted events:

- **Hospital safety.** Violent incidents from a multi-ward forensic psychiatric hospital were analysed at five-minute resolution. The model distinguishes ordinary incidents from those that triggered an alarm heard across the hospital and allows for day–night differences. It finds that non-alarm incidents mostly echo within the same ward, whereas alarm-triggered incidents create wider, hospital-level ripple effects that last longer-insight that can inform staffing and response protocols.

- **Infectious disease.** Weekly counts of Cryptosporidiosis cases are decomposed into a smooth background pattern and short bursts, helping to tell regular seasonal variation from true spikes that may warrant attention.
- **Security events.** Historical U.S. terrorism incidents-mostly long stretches of quiet with sudden clusters-are captured more faithfully by the new model than by standard baselines, improving short-horizon predictive accuracy.

What we found. Across studies, the faster algorithms make previously slow calculations routine, and the flexible model improves out-of-sample prediction while producing transparent summaries: a background trend plus an "impact curve" that shows how long after-effects last. In the hospital case, alarms act as system-wide signals that raise risk across wards for sustained periods, whereas ordinary events mainly have short, local impact. In the public-health and security settings, the method separates steady patterns from bursts, aiding monitoring and communication.

How it can be used. The toolkit supports (i) simple dashboards that display current background levels and recent after-effects; (ii) "what-if" scenarios (for example, how risk changes if alarms become more or less frequent); and (iii) rapid, repeated forecasting for planning. The approach is not a replacement for professional judgement, and it works best with reliable, routinely collected counts.

In short: the thesis delivers fast, flexible, and interpretable methods for counted event data in which one event can trigger another-helping practitioners see when today's events are likely to echo into tomorrow.

Acknowledgements

I am deeply grateful to my advisers, Dr Gordon Ross and Dr Daniel Paulin, for their guidance and support.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Trinnhallen Brisley

Contents

Figures and Tables

Figures

Tables

Chapter 1

Introduction

1.1 Hawkes Processes

The Hawkes process (Hawkes, 1971) is a temporal, self-exciting point process: the occurrence of an event increases the short-term likelihood of further events, leading to clustered behaviour. In the simplest univariate form, the model specifies the conditional intensity (the instantaneous event rate) as

$$\lambda(t \mid H(t)) \;=\; \mu(t) \;+\; \sum_{t_i<t} g(t-t_i)\,,$$

where $H(t) = \{t_i : t_i < t\}$ denotes the history of events before time t. The first term $\mu(t)$ is a baseline rate that operates independently of past events, while the second term is a self-exciting contribution: each past event raises the intensity for a period of time according to the memory kernel $g(\cdot)$. This decomposition makes clear the two driving forces of the process: exogenous arrivals from the baseline and endogenous reinforcement from past events. A formal definition is given later in Subsection 2.6.1.

An equivalent representation was established by (Hawkes & Oakes, 1974), who showed that a Hawkes process can be viewed as a Poisson *cluster* process. Each observed event is either an *immigrant*, generated independently of the history by a Poisson process with rate equal to the baseline, or an *offspring* triggered by earlier events according to the self-exciting mechanism. Immigrants act as cluster centres and stochastically generate offspring, thereby producing the observed clustering. The conditional-intensity and the clustering representations are mathematically equivalent, and together they provide both inferential tractability and interpretability. Since their introduction (Hawkes, 1971), Hawkes processes have been widely applied to real-world phenomena where feedback is intrinsic, including seismicity and aftershock sequences, high-frequency financial activity, contagious social or criminal events, and neural spike trains.

In this thesis we predominantly adopt a discrete-time formulation that is well suited to data recorded on regular time grids or with rounded time stamps. We work with event counts aggregated over successive, equally spaced intervals, together with their cumulative totals. The expected number of events in each interval, given the past, plays the role of the intensity. Conceptually, we separate this intensity into a baseline component, representing exogenous background activity, and a self-exciting component, capturing how earlier events elevate the likelihood of subsequent events within and across intervals. In multivariate settings, this self-exciting mechanism also allows events in one series to influence others (cross-excitation). The full notation is introduced at the start of Chapter 2 and used consistently thereafter.

1.2 Motivation

There exists a substantial literature on Hawkes processes; nevertheless, there remain important methodological and computational gaps that motivate the present work. The first focus of the thesis is to advance flexible, principled estimation for Hawkes processes in a manner that respects how data are often collected in practice, namely in discrete, regular time intervals and occasionally with multiple events recorded in the same bin. Although continuous-time models are classical, discrete-time variants (e.g. Browning, Sulem, Mengersen, Rivoirard, & Rousseau, 2021; White, Porter, & Mazerolle, 2013) remain comparatively understudied, despite their suitability for rounded or batched observations and their natural compatibility with integer-valued event counts.

A second theme is methodological flexibility in the specification of the baseline and self-exciting mechanism. The triggering kernel, the function governing how past events influence present risk, is a critical component of a Hawkes model as it characterises temporal dependence and memory. Much applied work assumes a parametric functional form such as exponential or power-law decay. While such specifications can be effective, the true generative mechanism is rarely known *a priori*, and restrictive parametric assumptions may fail to capture complex dynamics observed in practice. In this thesis we therefore emphasise *flexible* approaches that allow the data to inform the shape and scale of excitation while retaining interpretability and stability.

A third motivation is to develop scalable procedures for *multivariate* and *marked* Hawkes processes. Multivariate specifications enable simultaneous modelling of self-excitation within a dimension and cross-excitation across dimensions, providing substantially richer insight into directional interactions. However, such models are more computationally demanding due to the increased number of excitation pathways and associated parameters. In discrete time, careful algorithmic design can markedly reduce complexity by exploiting event-time recursions and cumulative-mass identities for geometric-type kernels, thereby allowing likelihoods

and gradients to be evaluated in essentially linear time in the number of event bins. This thesis leverages such identities to provide estimation methods that scale to high-dimensional settings while supporting structural regularisation (e.g., sparsity across cross-excitation pathways) to embed prior knowledge into the optimisation problem.

A fourth line of motivation stems from how exogenous structure and context are incorporated. In many domains, exogenous drivers, diurnal and weekly cycles, seasonality, or operational regimes, coexist with endogenous clustering. We therefore adopt baselines $\mu(t)$ that can encode structured variation over time (e.g., periodic components or piecewise-constant factors) and allow marks to modulate the excitation pathway. For example, binary or categorical marks can amplify or attenuate the strength or persistence of excitation, facilitating interpretable decompositions between ordinary and exceptional events. While this thesis focuses primarily on temporal processes, the same modelling principles extend to spatio-temporal settings in which the triggering kernel is decomposed into separable temporal and spatial factors (Mohler, Short, Brantingham, Schoenberg, & Tita, 2011; Reinhart, 2018; Schoenberg, 2016).

Finally, practical uptake is often impeded by the absence of readily available samplers or optimisers for complex discrete-Hawkes models, particularly for multivariate, marked, or flexible-kernel specifications. The computational burden grows quickly with model richness and data size. A further motivation of this thesis is to narrow this gap by developing bespoke inference procedures that exploit problem structure to reduce computational costs without sacrificing statistical fidelity. Taken together, these directions address the dual need for *flexibility* (to capture real-world complexity) and *scalability* (to handle modern data volumes), within a discrete-time Hawkes framework tailored to practical data-collection regimes.

1.3 Research Contributions

This thesis develops flexible and scalable methodology for multivariate, discrete-time Hawkes processes and tackles a novel dataset; a hospital-scale application: violent-incident monitoring across multiple wards in a forensic psychiatric hospital. Here, *flexible* means moving beyond rigid parametric forms for both the baseline and the self-exciting mechanism while preserving interpretability and stability.

1.3.1 Core Methodological Contributions

1. **Event-linear computations in discrete time.** We derive likelihood and gradient recursions that exploit event-time structure together with cumulative-mass identities, so evaluation is linear in the number of events. This makes inference practical on long horizons and in higher dimensions.

2. **Nonparametric baseline and excitation.** We introduce a discrete-time model with independent Gaussian process priors for the baseline and the excitation kernel. A collapsed representation enables scalable MAP estimation, and a post hoc decomposition separates background activity from feedback, yielding smooth, data-adaptive components.
3. **Marked and multivariate structure.** We extend the framework to cross-series excitation and mark-dependent effects (for example, alarm flags or operational regimes), and we encourage sparse, interpretable interactions through designed regularisation.

1.3.2 Application Contributions

1. **Hospital incident analysis.** We analyse violent incidents at five-minute resolution from a multi-ward forensic psychiatric hospital, including an alarm mark that indicates rare severe events. The model quantifies self- and cross-excitation, captures time-of-day and day-of-week structure in the baseline, and produces ward-level and hospital-level one-step-ahead risk signals.
2. **Broader validation.** We evaluate the approach on weekly Cryptosporidiosis counts and on U.S. terrorism incidents, compare predictive log-likelihood with parametric discrete Hawkes baselines, and illustrate baseline and excitation decompositions that reveal bursts, lags, and seasonal structure.

By combining marked multivariate modelling, flexible components, and event-linear evaluation in discrete time, and by demonstrating the approach at hospital scale, the chapter provides a practical toolkit for regularly sampled count data with possible multiple events per interval. The methodology achieves flexibility without sacrificing interpretability, through regularisation and a transparent post hoc decomposition.

1.4 Thesis Outline

This thesis is presented with six chapters and one appendix. Each chapter provides its own focused review and methodological context so it can be read independently. References are consolidated in a single bibliography at the end; technical derivations, algorithms, and additional diagnostics are collected in the appendix.

Chapter 1 introduces Hawkes processes, motivates the discrete-time focus, and states the aims and objectives of the work.

Chapter 2 provides background on frequentist and Bayesian parameter estimation (including MAP as penalised likelihood), and on nonparametric modelling with Gaussian processes. It then reviews temporal point processes, Poisson processes, Hawkes processes (including multivariate and marked forms), and discrete-time Hawkes processes, fixing notation used throughout.

Chapter 3 develops *efficient estimation and forecasting* for multivariate discrete-time Hawkes processes. It derives an event-time log-likelihood identity, $O(1)$ recursions for intensities and gradients across successive event times, and a time–step recursion for simulation/forecasting, together with empirical runtime comparisons.

Chapter 4 presents an *application to incident monitoring* in a multi-ward forensic psychiatric hospital using a marked multivariate discrete Hawkes specification. The model separates baseline structure from self- and cross-excitation, incorporates day/night regimes and alarm-triggered marks, and evaluates predictive performance and interpretability.

Chapter 5 introduces a *Gaussian Process Discrete Hawkes Process (GP-DHP)* that models both baseline and excitation nonparametrically. A collapsed latent-intensity GP prior enables scalable MAP inference and a post hoc decomposition into baseline and excitation components. The chapter reports results on synthetic data and real case studies.

Chapter 6 concludes with a synthesis of contributions, limitations, and directions for future work.

Appendix A gathers additional material for Chapter 3, including a simulation study and a justification of the Poisson observation model, along with further diagnostics.

Chapter 2

Background

2.1 Overview

This chapter provides a general introduction to parameter estimation in probabilistic modelling. The focus is on three pillars that recur throughout the thesis: maximum likelihood estimation, maximum a posteriori estimation viewed as penalised likelihood, and nonparametric modelling with Gaussian processes. Specific Hawkes process details are deferred to later chapters, so the exposition here is model agnostic.

2.2 Frequentist Parameter Estimation

In the frequentist approach, parameters are fixed but unknown and in this thesis we estimate them by maximising the likelihood of the observed data. For independent observations $y = (y_1, \ldots, y_n)$ with model $p_\theta(y_i)$,

$$L(\theta \mid y) = \prod_{i=1}^{n} p_\theta(y_i), \qquad \ell(\theta) = \sum_{i=1}^{n} \log p_\theta(y_i),$$

and an MLE is any maximiser $\hat{\theta} \in \arg\max_{\theta \in \Theta} \ell(\theta)$.

When we want *additional structure*, for example, stability under multicollinearity or sparsity, we maximise a penalised objective

$$Q_\lambda(\theta) = \ell(\theta) - \lambda P(\theta),$$

with tuning parameter $\lambda \geq 0$ and penalty function $P(\theta)$. Two standard choices are:

- ℓ_2 **(ridge) penalty** (Hoerl & Kennard, 1970): $P(\theta) = \frac{1}{2}\|\theta\|_2^2$. This shrinks coefficients smoothly towards zero, improving stability but without setting parameters exactly to zero.
- ℓ_1 **(lasso) penalty** (Tibshirani, 1996): $P(\theta) = \|\theta\|_1$. This encourages sparsity by driving some coefficients exactly to zero, aiding interpretability and automatic variable selection.

Penalised likelihood therefore extends maximum likelihood in a way that lets us encode stability or sparsity while retaining the same optimisation framework.

2.3 Bayesian Parameter Estimation

Let $y = (y_1, \dots, y_n)$ denote the observed data, with each y_i taking values in some observation space, and let $\theta \in \mathbb{R}^p$ be a vector of unknown parameters. A Bayesian analysis combines the likelihood $p(y \mid \theta)$ with prior information $p(\theta)$ to form the posterior

$$p(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{p(y)}, \qquad p(y) = \int p(y \mid \theta)\, p(\theta)\, d\theta,$$

where $p(y)$ is the model evidence (marginal likelihood). The posterior $p(\theta \mid y)$ is the primary object of inference: it quantifies uncertainty about θ after observing y. The likelihood $p(y \mid \theta)$ measures data fit, the prior $p(\theta)$ encodes beliefs before observing data, and the evidence $p(y)$ acts as the normalising constant and supports model comparison. In general $p(y)$ is not available in closed form. When the prior is conjugate to the likelihood (for example, exponential-family models with matching priors), the posterior is available in closed form, which allows direct calculation of posterior summaries. Otherwise, one uses approximate inference. A large class of methods is based on simulation: Markov chain Monte Carlo (MCMC) constructs a Markov chain with stationary distribution $p(\theta \mid y)$ and produces draws from the posterior up to the unknown normalising constant. Prominent examples include the Metropolis–Hastings algorithm (Hastings, 1970) and the Gibbs sampler (Gelfand & Smith, 1990). Gradient-informed proposals can improve mixing; for instance, the Metropolis-adjusted Langevin algorithm (MALA) uses local gradient information to move towards higher posterior mass (Roberts & Tweedie, 1996). Other Monte Carlo approaches include sequential Monte Carlo (Liu & Chen, 1998) and likelihood-free methods such as approximate Bayesian computation (Beaumont, Zhang, & Balding, 2002).

2.4 Nonparametric Modelling and Gaussian Processes

Nonparametric modelling replaces rigid parametric forms with flexible structures whose effective complexity can increase with the amount of data. Common approaches include Gaussian processes, splines, and kernel methods.

We focus on Gaussian processes because they provide flexible priors over unknown functions, with smoothness and scale controlled by a covariance kernel. They also admit tractable finite-dimensional inference: for any collection of inputs $X = \{x_1, \dots, x_n\}$, the corresponding function values $\mathbf{f}_X = (f(x_1), \dots, f(x_n))^\top$ follow a multivariate Gaussian distribution (C. E. Rasmussen & Williams, 2006). In the chapters that follow, GP priors are placed on the baseline intensity

and on the excitation kernel in discrete-time Hawkes models, and estimation is carried out by MAP. In this setting, the GP prior contributes the quadratic regularisation term

$$\mathbf{f}_X^\top K_X^{-1} \mathbf{f}_X,$$

where K_X is the covariance matrix induced by the kernel over the input set X.

2.4.1 Gaussian Process Regression: Definition and Properties

We introduce Gaussian processes in the context of regression, since this is the setting used throughout this thesis. Let $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^n$ denote observed input–output pairs, where $x_i \in \mathscr{X}$ and $y_i \in \mathbb{R}$. We assume that each observation is generated from an underlying latent function f corrupted by additive Gaussian noise:

$$y_i = f(x_i) + \varepsilon_i, \qquad \varepsilon_i \overset{\text{iid}}{\sim} \mathscr{N}(0, \sigma_n^2), \qquad i = 1, \dots, n.$$

Equivalently, writing

$$\mathbf{y} = (y_1, \dots, y_n)^\top, \qquad \mathbf{f}_X = \big(f(x_1), \dots, f(x_n)\big)^\top,$$

we have

$$\mathbf{y} \mid \mathbf{f}_X \sim \mathscr{N}(\mathbf{f}_X, \sigma_n^2 I_n).$$

A Gaussian process (GP) places a prior directly on the unknown latent function f. Specifically, f is said to follow a Gaussian process with mean function m and covariance function k if, for every finite collection of inputs $X = \{x_1, \dots, x_n\} \subset \mathscr{X}$,

$$\mathbf{f}_X \sim \mathscr{N}(\mathbf{m}_X, K_X),$$

where

$$\mathbf{m}_X = \big(m(x_1), \dots, m(x_n)\big)^\top, \qquad [K_X]_{ij} = k(x_i, x_j).$$

We write

$$f \sim \mathscr{GP}\big(m(\cdot), k(\cdot, \cdot)\big).$$

In most applications, and throughout this thesis unless stated otherwise, the mean function is taken to be zero, so that

$$f \sim \mathscr{GP}\big(0, k(\cdot, \cdot)\big).$$

Under this model, the observed responses satisfy

$$\mathbf{y} \sim \mathscr{N}\big(\mathbf{m}_X,\ K_X + \sigma_n^2 I_n\big).$$

Thus, GP regression can be viewed as Bayesian nonparametric regression in which the covariance function k determines how function values at different inputs co-vary. Intuitively, if $k(x,x')$ is large, then the model expects $f(x)$ and $f(x')$ to be similar; if $k(x,x')$ is small, the dependence is weak.

For prediction at a new input x_*, define

$$k_* = \big(k(x_1,x_*),\dots,k(x_n,x_*)\big)^\top, \qquad k_{**} = k(x_*,x_*).$$

Then the joint Gaussianity of $(\mathbf{y}, f(x_*))$ implies the standard GP regression formulas

$$f(x_*) \mid \mathbf{y}, X, x_* \sim \mathcal{N}\big(\mu(x_*),\, \sigma^2(x_*)\big),$$

with

$$\mu(x_*) = m(x_*) + k_*^\top \big(K_X + \sigma_n^2 I_n\big)^{-1}\big(\mathbf{y} - \mathbf{m}_X\big),$$

and

$$\sigma^2(x_*) = k_{**} - k_*^\top \big(K_X + \sigma_n^2 I_n\big)^{-1} k_*.$$

These expressions show the two main outputs of GP regression: a posterior mean function and an associated predictive uncertainty.

Weight-space intuition. Although Gaussian processes are most naturally interpreted as distributions over functions, they can also be motivated from Bayesian linear regression in a feature space. Suppose

$$f(x) = \phi(x)^\top w,$$

where $\phi(x) \in \mathbb{R}^N$ is a feature map and

$$w \sim \mathcal{N}(0, \Sigma_w),$$

with Σ_w symmetric positive semidefinite. Then for any finite set of inputs $X = \{x_1,\dots,x_n\}$, the vector $\mathbf{f}_X$ is Gaussian, and

$$\operatorname{cov}\big(f(x), f(x')\big) = \operatorname{cov}\big(\phi(x)^\top w, \phi(x')^\top w\big) = \phi(x)^\top \Sigma_w \phi(x').$$

Thus the induced covariance function is

$$k(x,x') = \phi(x)^\top \Sigma_w \phi(x').$$

Since Σ_w is symmetric positive semidefinite, it admits a spectral decomposition

$$\Sigma_w = U D U^\top,$$

where U is orthogonal and D is diagonal with nonnegative entries. Hence

$$\Sigma_w^{1/2} = UD^{1/2}U^\top$$

is well defined, and we may write

$$k(x,x') = \phi(x)^\top \Sigma_w^{1/2}\Sigma_w^{1/2}\phi(x') = \left(\Sigma_w^{1/2}\phi(x)\right)^\top \left(\Sigma_w^{1/2}\phi(x')\right).$$

Therefore, defining

$$\psi(x) = \Sigma_w^{1/2}\phi(x),$$

gives

$$k(x,x') = \psi(x)^\top \psi(x'),$$

so the covariance function is an inner product in a transformed feature space. This shows that a GP may be viewed as the function-space representation of Bayesian linear regression with possibly high-dimensional, or even infinite-dimensional, features. In practice, one usually works directly with the kernel function k, without constructing ϕ or ψ explicitly.

Common covariance functions. The covariance function, or kernel, encodes prior assumptions about the latent function such as smoothness, scale, and periodicity.

Squared-exponential (RBF) kernel. A standard choice is

$$k_{\mathrm{rbf}}(x,x') = \sigma_f^2 \exp\left(-\frac{\|x-x'\|^2}{2\ell^2}\right),$$

where σ_f^2 controls the marginal variance and ℓ is a length-scale parameter. This kernel produces smooth sample paths and gives strongly correlated function values for nearby inputs.

Periodic kernel. For periodic structure, one may instead use

$$k_{\mathrm{per}}(x,x') = \sigma_f^2 \exp\left(-\frac{2\sin^2\left(\pi\|x-x'\|/p\right)}{\ell^2}\right),$$

where $p > 0$ is the period. This kernel favours functions that repeat regularly over the input space.

Polynomial kernel. Polynomial structure can be represented using

$$k_{\mathrm{poly}}(x,x') = (c + x^\top x')^r,$$

with degree $r \in \mathbb{N}$ and offset $c \geq 0$. This corresponds to Bayesian linear regression in a finite-dimensional polynomial feature space.

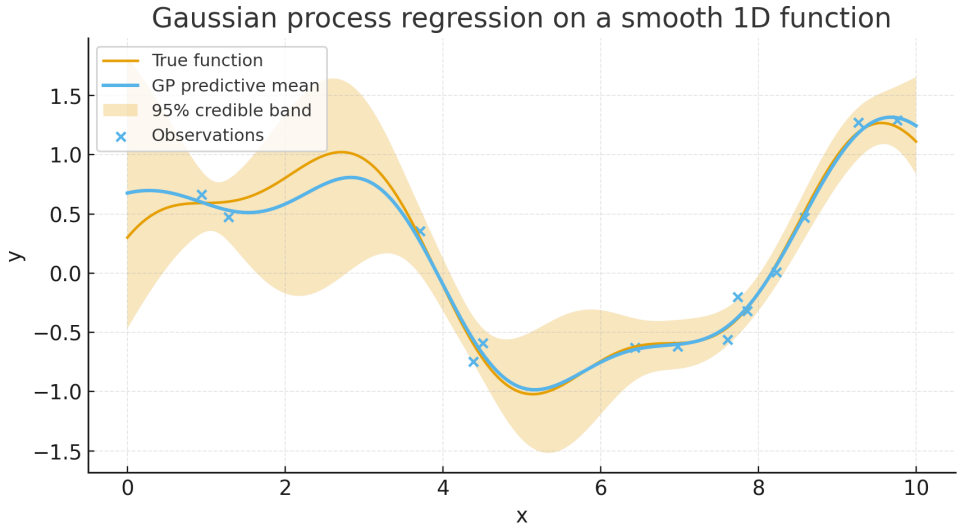


Figure 2.1: Gaussian process regression with an RBF kernel fitted to noisy observations from a smooth one-dimensional signal. The solid curve is the posterior mean and the shaded region is the pointwise 95% credible band.

Illustrative example: GP regression on a smooth 1D signal. To illustrate the model, consider noisy observations from the latent signal

$$f_\star(x) = \sin(0.8x) + 0.3\cos(2x), \qquad y_i = f_\star(x_i) + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma_n^2),$$

with non-uniform inputs $x_i \in [0, 10]$ and noise level $\sigma_n = 0.15$. We place a zero-mean GP prior with RBF kernel

$$k_{\mathrm{rbf}}(x, x') = \sigma_f^2 \exp\left(-\frac{\|x - x'\|^2}{2\ell^2}\right),$$

using $\sigma_f = 1$ and $\ell = 1.2$. The posterior mean and pointwise 95% credible band are shown in Figure 2.1. As expected, uncertainty contracts near observations and increases away from the data.

Gaussian processes are therefore attractive regression models because they provide a flexible prior over functions, admit closed-form posterior inference under Gaussian noise, and quantify uncertainty in a principled way through the predictive distribution.

2.5 Temporal Point Processes

Temporal point processes model the occurrence of events along a continuous time axis. Each time an event occurs, its time of occurrence is recorded, producing an ordered sequence of event times $\{t_1, \ldots, t_N\} \subset [0, T]$ for N observed events, with

$$0 \leq t_1 \leq \cdots \leq t_N \leq T, \qquad t_i \in \mathbb{R}_+, \quad i = 1, \ldots, N.$$

Typical examples include earthquake occurrences and arrivals of calls to a call centre.

Definition 2.5.1 (Temporal point process). A temporal point process on $[0, T]$ is a random, nonnegative, and nondecreasing sequence of event times taking values in $[0, T]$; it specifies the times at which events occur.

Point-process models describe how events are distributed over time and can be used to predict when future events are likely to occur. They can incorporate seasonal patterns and exogenous covariates, and they can capture interactions among events themselves, either increasing the likelihood of subsequent events (self-excitation and clustering) or decreasing it (inhibition). Such behaviour may arise from genuine causal mechanisms or from data-generating processes that promote clustering; appropriately chosen point-process specifications can represent either. As the name suggests, temporal point processes carry an intrinsic temporal structure: events occur one after another in continuous time, and the data are the event *times* themselves. This is fundamentally different from standard time-series settings, where measurements are taken at pre-specified sampling times regardless of whether any event has occurred. When the arrival mechanism is the quantity of interest, a temporal point-process formulation is the natural choice.

2.5.1 Counting Processes

We now examine counting processes, the basic device for recording how many events have occurred by any given time in a temporal point process. Intuitively, a counting process $N(t)$ starts at zero and increases by one each time an event arrives, thereby providing a cumulative total of events observed up to time $t \geq 0$. Between event times the process remains flat, and it never decreases. Formally, we view N as a stochastic process with state space $\mathbb{N}_0$ and time index $t \in [0, \infty)$.

Definition 2.5.2 (Counting process). A stochastic process $N(t)$, $t \geq 0$, is called a counting process if it satisfies the following properties (Daley & Vere-Jones, 2003):

(i) $N(t) \geq 0$ for all t (nonnegative);
(ii) $N(t) \in \mathbb{N}_0$ for all t (integer valued);
(iii) if $s < t$, then $N(s) \leq N(t)$ (nondecreasing in time);
(iv) for any $s < t$, the increment $N(t) - N(s)$ equals the number of events that occurred in the interval $(s, t]$.

We adopt the common convention $N(0) = 0$ when no events occur before time 0.

Consider a point process with observed data $Y = \{y_1, \ldots, y_N\}$, where each observation consists solely of an event time, that is, $y_i = (t_i)$ for $i = 1, \ldots, N$. The event times are recorded in continuous time and arranged in nondecreasing order,

$$0 < t_1 \leq t_2 \leq \cdots \leq t_N \leq T,$$

so that no events occur prior to time 0. The associated counting process can be written explicitly as

$$N(t) := \sum_{i=1}^{N} \mathbf{1}\{t_i \leq t\}, \tag{2.1}$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function. An equivalent set-based formulation counts points falling in any measurable set $A \subseteq [0, T]$:

$$N(A) := \sum_{i=1}^{N} \mathbf{1}\{t_i \in A\}. \tag{2.2}$$

In particular, for an interval $(s, t] \subseteq [0, T]$ we have

$$N(t) - N(s) = N((s, t]),$$

which recovers property (iv) above. As a path in time, $N(t)$ is a right-continuous, piecewise-constant step function that increases by exactly one at each event time and stays constant between events. These unit jumps coincide with the arrival times $\{t_i\}_{i=1}^{N}$, while the flat segments reflect the absence of new events. This staircase shape is the canonical visual signature of a counting process and is routinely used in figures to illustrate the connection between individual event times and the cumulative count over the observation window.

2.5.2 Poisson Processes

We now consider Poisson processes, beginning with the inhomogeneous case, which is a simple but fundamental model for random events on a time axis. An inhomogeneous Poisson process is fully described by a nonnegative intensity function $\lambda(t)$ that controls the local event rate as time evolves. The definition below encodes the two key features of such processes, namely independent increments and Poisson distributed counts over any interval.

Definition 2.5.3 (Poisson process). A counting process $N(t)$ with rate function $\lambda(\cdot) > 0$ is called an inhomogeneous Poisson process if the following conditions hold (Kingman, 1993):

(i) The increments are independent: for any collection of disjoint sets or disjoint intervals $A_1, \ldots, A_K$, the random variables $N(A_1), \ldots, N(A_K)$ are mutually independent.

(ii) For any $s < t$, the increment $N(t) - N(s)$ has a Poisson distribution with mean equal to the integrated intensity over $(s,t]$, that is

$$\mathbb{P}\big(N(t) - N(s) = n\big) = \exp\Big\{ -\int_s^t \lambda(u)\,du \Big\} \frac{\Big(\int_s^t \lambda(u)\,du\Big)^n}{n!}, \qquad n \in \mathbb{N}_0.$$

These properties imply

$$\mathbb{E}[N(t) - N(s)] = \int_s^t \lambda(u)\,du \qquad \text{and} \qquad \mathrm{Var}[N(t) - N(s)] = \int_s^t \lambda(u)\,du,$$

with counts over disjoint intervals being independent.

Consider now a Poisson process with intensity function $\lambda(t)$ and observed events at times $t_1, \ldots, t_N$, where each $t_i \in [0,T]$ and $0 < t_1 \leq \cdots \leq t_N \leq T$. The joint density of the observed event times over $[0,T]$ can be written in the product-exponential form

$$p\big(\{t_i\}_{i=1}^N\big) = \left\{ \prod_{i=1}^N \lambda(t_i) \right\} \exp\big\{ -\Lambda \big\}, \qquad \Lambda := \int_0^T \lambda(z)\,dz, \tag{2.3}$$

where Λ is the cumulative intensity over the observation window $[0,T]$. This likelihood is sometimes referred to as the Poisson likelihood or conditional-intensity likelihood, and it arises by combining the gap probabilities of having no events between consecutive observed times with the hazard contributions at the observed events (Snyder & Miller, 1991). The factor $\exp\{-\Lambda\}$ accounts for the probability of zero events outside the observed times, while the product $\prod_i \lambda(t_i)$ collects the instantaneous rates at the actual event locations.

An important structural property of Poisson processes is superposition. If $N_1(t), \ldots, N_K(t)$ are independent Poisson processes with respective intensities $\lambda_1(t), \ldots, \lambda_K(t)$, then their sum

$$N(t) = \sum_{k=1}^K N_k(t) \tag{2.4}$$

is itself a Poisson process with intensity

$$\lambda(t) = \sum_{k=1}^K \lambda_k(t). \tag{2.5}$$

Conversely, independent thinning of a Poisson process produces independent Poisson subprocesses with scaled intensities, a fact frequently used for simulation and model construction (Cox & Isham, 1980).

A special and widely used case is the homogeneous Poisson process, obtained when the intensity is constant in time, namely

$$\lambda(t) \equiv \lambda, \tag{2.6}$$

for some $\lambda > 0$. In this case the increments are not only independent but also stationary, so $N(t) \sim \text{Poisson}(\lambda t)$ for each $t \geq 0$. Moreover, the interevent or interarrival times $D_i = t_i - t_{i-1}$, with the convention $t_0 = 0$, are independent and identically distributed with $\text{Exponential}(\lambda)$ law, so the event times $\{t_i\}$ form a renewal process. Recall that a renewal process is a point process whose successive interarrival times are independent and identically distributed. In the present case, the interarrival density is

$$f_D(d) = \lambda e^{-\lambda d}, \qquad d \geq 0,$$

while the renewal function is $m(t) = \mathbb{E}[N(t)] = \lambda t$, and hence the renewal density is

$$m'(t) = \lambda.$$

For the inhomogeneous case, a time change by the cumulative intensity

$$\Lambda(t) = \int_0^t \lambda(u)\, du$$

transforms the process into a homogeneous Poisson process, a useful device for inference, goodness-of-fit, and simulation (Brown, Barbieri, Ventura, Kass, & Frank, 2002).

2.6 Hawkes Processes

We now turn to the Hawkes process, a self-exciting point process that captures clustered event arrivals, which will be the core modeling tool in the chapters that follow. The framework was introduced by (Hawkes, 1971) in his seminal paper.

2.6.1 Univariate Hawkes Processes

We focus on a one-dimensional Hawkes process on a finite observation window $[0, T]$. The data consist of N arrivals recorded at times $t_1, \ldots, t_N$ with $0 \leq t_1 \leq \cdots \leq t_N \leq T$. It is convenient to write the data as $Y = \{y_1, \ldots, y_N\}$ with the shorthand $y_i = (t_i)$. In one dimension this notation may look redundant, but it will allow a seamless transition to the multivariate setting where each event carries a component label. Let $H(t) = \{y_i : t_i < t\}$ denote the history strictly prior to time t. The conditional intensity function specifies the instantaneous event rate given the past. In the univariate Hawkes specification we decompose the rate into a background part and an excitation contribution from previous events:

$$\lambda(t \mid \theta, H(t)) \;=\; \mu(t) \;+\; \sum_{i:\, t_i < t} K\, g(t - t_i)\,. \tag{2.7}$$

Here $\mu(t) \geq 0$ is the baseline intensity, K is a nonnegative scalar controlling the overall strength of excitation, and $g(\cdot)$ is a nonnegative kernel that shapes how the influence of an event decays over time. To isolate the total contribution of a single arrival, we adopt the normalization

$$g(z) \geq 0, \qquad \int_0^\infty g(z)\, dz = 1,$$

so that each past event contributes an area K to the future intensity and the function g allocates that mass across lags $z = t - t_i$. We group the unknowns as $\theta = (\theta_\mu, K, \theta_g)$, where θ_μ parameterizes $\mu(\cdot)$ and θ_g parameterizes the kernel. Nonnegativity of the intensity requires $\mu(t) \geq 0$ for all t. In many applications $\mu(t)$ is strictly positive and may vary in time to capture regular patterns that are not driven by event-to-event feedback. Allowing $\mu(t)$ to depend on covariates or to evolve smoothly provides a mechanism for representing seasonality, long-term trends, or exogenous effects. Popular choices include log-Gaussian Cox constructions (Møller, Syversveen, & Waagepetersen, 1998), Gaussian process priors that encourage smooth but adaptive baselines (Adams, Murray, & MacKay, 2009), and fully nonparametric specifications such as Dirichlet process mixtures (Ferguson, 1973). When such flexibility is unnecessary, one can take a constant background $\mu(t) \equiv \mu$, in which case $\theta_\mu = \mu$ is a single nonnegative scalar. The excitation term aggregates the impact of prior events. Because g has been normalized as a density on $[0, \infty)$, the scalar K plays the role of a gain or magnitude parameter: larger K leads to stronger clustering because each arrival injects more mass into the future rate. Taking $K \geq 0$ yields self-excitation, meaning that an event increases the chance of subsequent events for some time after its occurrence. One can also consider inhibitory behavior by allowing negative contributions, corresponding to $K < 0$, in which case arrivals temporarily suppress the rate. In what follows we restrict attention to the excitation-only regime. A further restriction is needed to avoid explosive behavior: with the density normalization on g, imposing $K < 1$ ensures that the process does not produce infinitely many arrivals in a finite time interval. This stability requirement is the standard safeguard that prevents runaway cascades of events (Hawkes & Oakes, 1974). The choice of kernel g determines the temporal profile of the excitation generated by a single event. Decreasing kernels are frequently used so that recent events have greater influence than distant ones, but alternative shapes may be more suitable when domain knowledge suggests delayed impacts or multiple characteristic time scales. The menu of viable options is broad. In some applications a piecewise-constant or histogram form is adopted to reflect discrete lag structure or to encode empirically estimated influence windows (Lewis & Mohler, 2011). Parametric kernels are also common because they provide interpretable time scales and admit closed-form calculations in likelihoods and moments. A widely used example is the exponential kernel

$$g(z) \;=\; \beta\, e^{-\beta z}, \qquad \beta > 0,\, z \geq 0, \tag{2.8}$$

which places most of its mass near zero and decays at rate β. Under (2.8) the parameter set becomes $\theta = (\theta_\mu, K, \beta)$. Other families, such as sums of exponentials or heavy-tailed choices, can be used when the data suggest more persistent memory or multiple decay regimes (Reinhart, 2018). From a modeling perspective, the representation in (2.7) is attractive because it is both interpretable and modular. The baseline $\mu(t)$ can be enriched or simplified depending on what the application demands, without altering the basic mechanics of self-excitation. The magnitude K provides a single knob that controls the overall clustering tendency. The kernel g encapsulates the shape of the memory effect and can be tailored to the problem at hand. The normalization of g ensures that these three roles do not confound one another: the baseline governs the steady level, the kernel governs the time profile, and the scalar K governs the size of the excitation. Throughout, we will omit the explicit conditioning on $H(t)$ when it is clear from context to keep formulas compact, but the dependence on the past is always implicit in the definition of $\lambda(t \mid \theta, H(t))$. In summary, the univariate Hawkes process takes observed event times $t_1, \ldots, t_N$ and posits an intensity that is the sum of a nonnegative background $\mu(t)$ and contributions from prior events, each contribution shaped by a kernel g and scaled in total by a gain K. The parameter vector is $\theta = (\theta_\mu, K, \theta_g)$, or $\theta = (\theta_\mu, K, \beta)$ in the exponential case. Typical modeling choices enforce $\mu(t) \geq 0$, $K \geq 0$, and $K < 1$, and select g to reflect how influence should decay with time. These ingredients together produce clustered or bursty patterns that are characteristic of self-exciting phenomena and that are often observed in practice (Mohler et al., 2011).

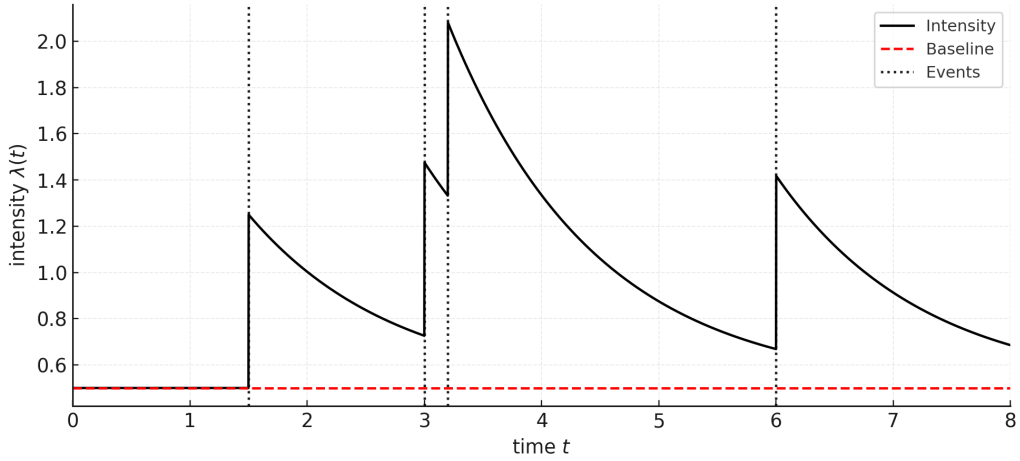


Figure 2.2: Hawkes intensity on $[0, 8]$ showing intensity (black), baseline (red dashed), and event times (black dotted). Parameters: baseline $\mu = 0.5$, jump size $K = 0.75$, decay rate $\beta = 0.8$; events at $t = \{1.5, 3, 3.2, 6\}$.

2.6.2 Branching or Cluster Representation

Hawkes and Oakes established an equivalent Poisson cluster representation (Hawkes & Oakes, 1974). Immigrants arrive according to the baseline process with rate $\mu(t)$. Each event at time t_i independently generates offspring according to a Poisson process with rate $\phi(\cdot)$. The entire process is the superposition of independent clusters rooted at immigrant events. This representation provides intuition for endogeneity and stability, supports simulation by drawing immigrants and then their descendants, and underlies EM-type inference. A convenient computational device is to introduce latent parent indicators $B = \{b_{ij}\}$ that record whether event t_j is the parent of event t_i with $j < i$, and b_{i0} indicates that t_i is an immigrant. Conditional on the past, the probability that t_i is an immigrant is

$$p_{i0} = \frac{\mu(t_i)}{\mu(t_i) + \sum_{t_j < t_i} \phi(t_i - t_j)},$$

and the probability that t_j triggered t_i is

$$p_{ij} = \frac{\phi(t_i - t_j)}{\mu(t_i) + \sum_{t_k < t_i} \phi(t_i - t_k)}, \qquad j < i.$$

These $\{p_{ij}\}$ act as responsibilities in an EM algorithm. In the E-step, expected counts are allocated to immigrant and offspring processes using $\{p_{ij}\}$. In the M-step, parameters in μ and ϕ are updated by maximizing the expected complete-data log-likelihood, which separates into independent Poisson terms for the immigrant and offspring processes (Lewis & Mohler, 2011; J. G. Rasmussen, 2011). Similar augmentations extend to multivariate models and can be embedded in Gibbs or Metropolis-within-Gibbs schemes (Veen & Schoenberg, 2008; R. Zhang, Walder, & Tanaka, 2019).

2.6.3 Multivariate Hawkes Processes

In many applications events come in multiple types or occur across several interacting components. Let $N^{(m)}(t)$ be the counting process for component $m \in \{1, \ldots, M\}$. The multivariate Hawkes model specifies

$$\lambda^{(m)}(t) = \mu^{(m)}(t) + \sum_{l=1}^{M} \int_0^{t-} \phi_{l,m}(t-s)\, dN^{(l)}(s) = \mu^{(m)}(t) + \sum_{l=1}^{M} \sum_{t_i^{(l)} < t} \phi_{l,m}\big(t - t_i^{(l)}\big), \qquad (2.9)$$

where $\phi_{l,m}$ governs the influence of component l on component m. With marks q_i attached to events, one may write $\phi_{l,m}(t - t_i^{(l)}, q_i)$ to allow the triggering effect to depend on event attributes. Define the *branching matrix* G with entries

$$G_{l,m} = \int_0^{\infty} \phi_{l,m}(u)\, du.$$

A sufficient stability condition is that the spectral radius $\rho(G) < 1$, which generalises the univariate condition $\bar{n} < 1$ (Brémaud & Massoulié, 1996). The multivariate log-likelihood sums the univariate contributions over m and includes the corresponding compensators:

$$\ell(\theta) = \sum_{m=1}^{M} \left\{ \sum_{t_i^{(m)} \le T} \log \lambda^{(m)}(t_i^{(m)}) - \int_0^T \lambda^{(m)}(u)\, du \right\}.$$

Estimation can be carried out by maximum likelihood with recursive filters for exponential kernels, or by penalised likelihood to encourage sparse cross-excitation in large networks (Zhou, Zha, & Song, 2013).

To illustrate Eq. (2.9), we simulate a three-component Hawkes process with exponential kernels on the horizon $T = 200$. The intensity for component m is

$$\lambda^{(m)}(t) = \mu^{(m)} + \sum_{l=1}^{3} \sum_{t_i^{(l)} < t} \phi_{l,m}\big(t - t_i^{(l)}\big), \qquad \phi_{l,m}(u) = A_{l,m}\, \beta\, e^{-\beta u} \ \text{ for } u > 0,$$

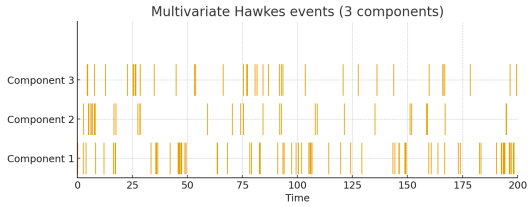
with baseline $\mu = (0.20, 0.10, 0.15)$, decay $\beta = 1.8$, and branching matrix

$$A = \begin{pmatrix} 0.35 & 0.05 & 0.00 \\ 0.10 & 0.25 & 0.05 \\ 0.00 & 0.20 & 0.20 \end{pmatrix}.$$

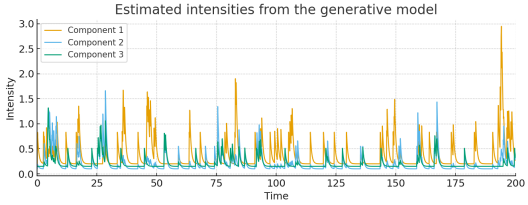
The spectral radius is $\rho(A) = 0.40 < 1$, so the process is subcritical and stable. In one realisation the component counts were 76, 35, and 42 for components 1, 2, and 3. Figure 2.3 The raster shows self-exciting bursts (diagonal A) and spillovers due to cross-excitation, for example $A_{3,2} = 0.20$ and $A_{2,1} = 0.10$. Intensities $\lambda^{(m)}(t)$ jump at events then decay at rate β; binned counts rise above baselines μ during clusters. The heatmap displays the branching matrix A, whose entries equal the expected number of direct children from l to m.

2.6.4 Simulation, Computation, and Applications

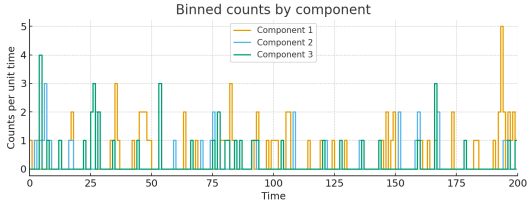
Simulation can proceed by Ogata's thinning applied to the conditional intensity or by the cluster construction, which is often simpler for stationary models (Hawkes & Oakes, 1974; Ogata, 1988). Recursive filters for exponential kernels enable fast likelihood and gradient evaluation, which is essential for large datasets and multivariate settings (Ozaki, 1979). Penalised likelihood or MAP estimation introduces structure, for example sparsity across components, while respecting stability constraints (Zhou et al., 2013).



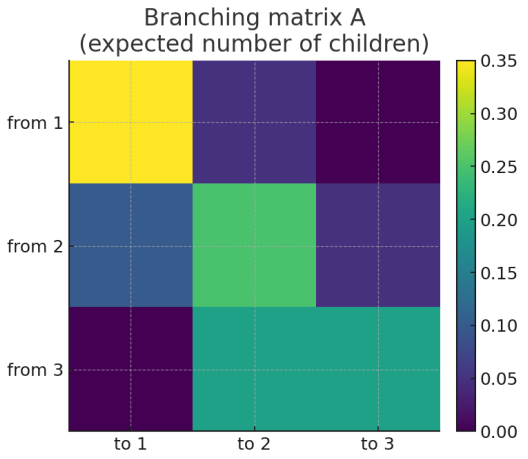
(a) Event raster by component.



(b) Component intensities $\lambda^{(m)}(t)$.



(c) Binned counts per unit time.



(d) Branching matrix A (expected children).

Figure 2.3: Multivariate Hawkes simulation with $\mu = (0.20, 0.10, 0.15)$, $\beta = 1.8$, and branching matrix A. The process is stable since $\rho(A) = 0.40 < 1$.

2.7 Discrete-Time Hawkes Processes

In many applications, events are not observed as exact continuous-time arrival times, but rather as counts aggregated over regular intervals such as hours, days, or weeks. In this setting, a discrete-time Hawkes process provides a natural analogue of the continuous-time Hawkes model by describing self-exciting dependence in a sequence of count observations. Instead of modeling exact event times, the process models the number of events occurring in each time bin.

This formulation can be viewed as a discretization of the usual continuous-time Hawkes process. If time is partitioned into intervals $(t-1, t]$, and Y_t denotes the number of events in the t-th interval, then the conditional mean count in each interval depends on a baseline term together with contributions from previous counts. In this way, past events increase the expected number of future events, but now through lagged bin counts rather than exact event times. Such discrete-time self-exciting models have been studied as practical and interpretable approximations when data are recorded on a regular grid (Browning, Rousseau, & Mengersen, 2022; Browning et al., 2021; Kirchner, 2016).

2.7.1 Univariate Discrete Hawkes Process

We now consider the univariate discrete-time Hawkes process, in which a single sequence of counts is observed over time.

Let N_t denote the cumulative number of events observed up to and including interval $(t-1,t]$, and define the count in interval t by

$$Y_t := \mathrm{N}_t - \mathrm{N}_{t-1}, \qquad t \in \mathbb{N}.$$

The history available strictly before time t is

$$\mathrm{H}_{t-1} = \{Y_s : s \leq t-1\}.$$

The conditional mean intensity at time t is then defined as

$$\lambda(t) = \mathbb{E}[Y_t \mid \mathrm{H}_{t-1}] = \mu(t) + \sum_{d=1}^{t-1} Y_{t-d}\,\Phi(d), \tag{2.10}$$

where $\mu(t) > 0$ is the baseline and $\Phi(d) \geq 0$ is a discrete lag kernel governing how past counts increase the current expected count.

Link with the continuous-time Hawkes model. The usual continuous-time Hawkes process is a point process with conditional intensity

$$\lambda^*(u) = \mu^*(u) + \sum_{t_i<u} g^*(u-t_i), \qquad u \in [0,T].$$

To pass to discrete time, partition the observation window into bins $(t-1,t]$, $t \in \mathbb{N}$, and let

$$Y_t = N((t-1,t])$$

denote the number of events in the t-th interval. Then the conditional mean count in bin t is obtained by integrating the continuous-time intensity over that interval:

$$\mathbb{E}[Y_t \mid \mathrm{H}_{t-1}] = \mathbb{E}\left[\int_{t-1}^{t} dN(u) \,\middle|\, \mathrm{H}_{t-1}\right] = \mathbb{E}\left[\int_{t-1}^{t} \lambda^*(u)\,du \,\middle|\, \mathrm{H}_{t-1}\right].$$

This motivates a discrete conditional-mean model of the form

$$\lambda(t) = \mathbb{E}[Y_t \mid \mathrm{H}_{t-1}] = \mu(t) + \sum_{d=1}^{t-1} Y_{t-d}\,\Phi(d),$$

where $\mu(t)$ represents the baseline contribution aggregated over bin t, and $\Phi(d)$ is a discrete lag kernel summarizing the effect of events that occurred d intervals in the past. Thus, the discrete-time Hawkes model should be viewed as a binned or count-valued analogue of the continuous-time Hawkes process. The additional assumption

$$Y_t \mid \mathrm{H}_{t-1} \sim \mathrm{Poisson}\big(\lambda(t)\big)$$

is a modelling choice for the conditional distribution of the bin counts, rather than a direct consequence of binning itself. A formal connection between continuous-time Hawkes processes and discrete-time count models is developed by Kirchner (2016); see also Browning et al. (2022, 2021) for discrete-time Hawkes formulations in practice.

Observation model. The conditional distribution of Y_t given H_{t-1} may be chosen from a broad class of count models, provided its conditional mean is $\lambda(t)$. Examples include the Poisson distribution, the negative binomial distribution for overdispersion, the binomial distribution when there is a natural upper bound per interval, and zero-inflated variants when excess zeros are present. In this thesis we adopt the Poisson observation model

$$Y_t \mid \mathrm{H}_{t-1} \sim \mathrm{Poisson}\big(\lambda(t)\big).$$

The specification of $\lambda(t)$ in (4.1) is independent of this choice.

Marked extension. If intervals carry additional information, let m_s denote a *mark* attached to interval $(s-1,s]$. A mark is any observed attribute of the events or context in that interval, for example an alarm indicator (0 or 1), an event-type label, a severity score, or a day-versus-night flag. Allowing the excitation to depend on both lag and mark gives

$$\lambda(t) = \mu(t) + \sum_{d=1}^{t-1} Y_{t-d}\,\Phi\big(d, m_{t-d}\big),$$

where $\Phi(d,m) \geq 0$ modulates the strength and persistence of excitation as a function of lag d and mark m. For instance, in the hospital application one may take $m_s = 1$ if at least one incident in $(s-1,s]$ triggered a hospital-wide alarm and $m_s = 0$ otherwise, so that $\Phi(d,1)$ may be larger or longer-lived than $\Phi(d,0)$.

If marked counts $Y_t^{(m)}$ are introduced explicitly, their observation model should likewise be stated conditionally on the past, for example

$$Y_t^{(m)} \mid \mathrm{H}_{t-1} \sim \mathrm{Poisson}\big(\lambda^{(m)}(t)\big),$$

with $\lambda^{(m)}(t)$ defined according to the chosen marked discrete-time Hawkes specification.

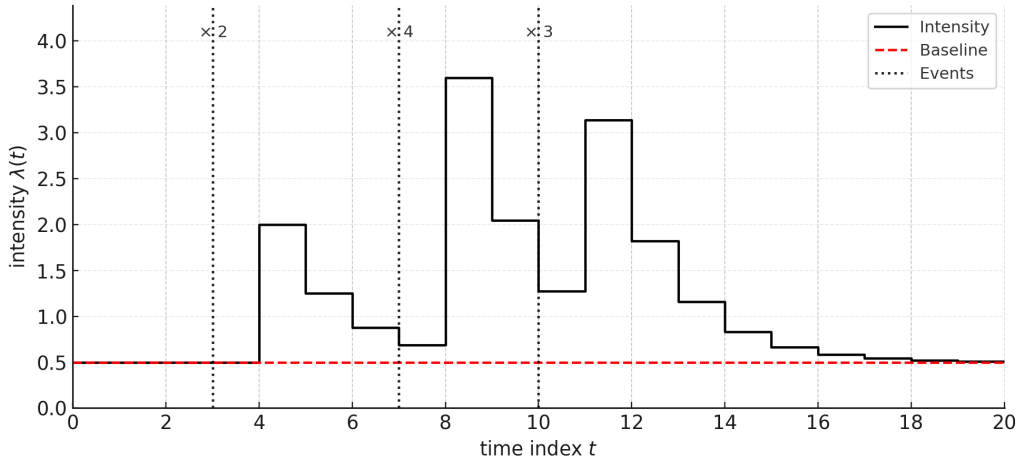


Figure 2.4: Discrete-time Hawkes intensity on $[0,20]$ with geometric excitation. The plot shows intensity (black step), baseline (red dashed), and event times (black dotted; multiplicities annotated at the top). Parameters: baseline $\mu = 0.5$, jump size $K = 0.75$, geometric decay $\beta = 0.5$; events at $t = 3,7,10$ with multiplicities $2,4,3$, respectively.

2.7.2 Branching Representation in Discrete Time

The self-exciting structure admits a branching interpretation analogous to the continuous time cluster view. For each t, decompose Y_t into immigrant and offspring totals via a latent vector $\boldsymbol{y}^{(t)} = (y_0^{(t)}, y_1^{(t)}, \ldots, y_{t-1}^{(t)})$ satisfying

$$Y_t = \sum_{d=0}^{t-1} y_d^{(t)}.$$

Here $y_0^{(t)}$ counts immigrant events at time t, and for $d \geq 1$, $y_d^{(t)}$ counts offspring at t triggered by events that occurred d intervals earlier. Under the Poisson observation model, the conditional allocation of the Y_t events to these sources follows a multinomial law with cell probabilities proportional to their contributions to $\lambda(t)$:

$$\left(y_0^{(t)}, y_1^{(t)}, \ldots, y_{t-1}^{(t)}\right) \sim \text{Multinomial}\left(Y_t;\ \frac{\mu(t)}{\lambda(t)},\ \frac{Y_{t-1}\Phi(1)}{\lambda(t)},\ \ldots,\ \frac{Y_1\Phi(t-1)}{\lambda(t)}\right).$$

This factorisation exploits the splitting property of independent Poisson components. For alternative observation models the same responsibility proportions provide useful weights in EM style or variational updates, although the exact multinomial sampling step may no longer hold. While Gibbs style samplers built on this augmentation can be effective for parametric specifications, they may exhibit slow mixing when highly flexible priors are placed on components of $\lambda(t)$, for example Gaussian process priors. In later chapters we therefore develop MAP based inference that scales efficiently and avoids such pathologies.

2.7.3 Multivariate Discrete Hawkes Process

Let $\mathbf{N}_t = \{\mathsf{N}_t^{(1)}, \ldots, \mathsf{N}_t^{(M)}\}$ be cumulative counts up to $(t-1, t]$ and $\mathbf{Y}_t = \{Y_t^{(1)}, \ldots, Y_t^{(M)}\}$ the interval counts with $Y_t^{(m)} = \mathsf{N}_t^{(m)} - \mathsf{N}_{t-1}^{(m)}$. Write $\mathbf{H}_{t-1} = \{\mathbf{Y}_s : s \leq t-1\}$ for the multivariate history, and let m_s denote any mark attached to interval $(s-1, s]$. For $m \in \{1, \ldots, M\}$ the conditional mean intensity in lag form is

$$\lambda^{(m)}(t) = \mathbb{E}\{Y_t^{(m)} \mid \mathbf{H}_{t-1}\} = \mu^{(m)}(t) + \sum_{l=1}^{M} \sum_{d=1}^{t-1} Y_{t-d}^{(l)} \, \Phi_{l,m}\big(d, m_{t-d}\big), \tag{2.11}$$

where $\mu^{(m)}(t) > 0$ is the baseline for component m, and $\Phi_{l,m}(d, m) \geq 0$ is a discrete excitation kernel that governs how past counts in component l at lag d influence the expected count in component m, possibly as a function of the mark m_{t-d}. Cross-excitation refers to increases in the probability of events across dimensions due to past events in other dimensions. When marks are not used, set $\Phi_{l,m}(d, m) \equiv \Phi_{l,m}(d)$.

Observation model. As in the univariate case, any count distribution with conditional mean $\lambda^{(m)}(t)$ can be used for $Y_t^{(m)} \mid \mathbf{H}_{t-1}$, for example Poisson, negative binomial, binomial, or zero-inflated variants. In this thesis we adopt

$$Y_t^{(m)} \sim \text{Poisson}\big(\lambda^{(m)}(t)\big), \qquad m = 1, \ldots, M.$$

The branching interpretation extends component wise by allocating each $Y_t^{(m)}$ to immigrant or offspring sources within and across dimensions, exactly multinomial under Poisson splitting, and approximately via responsibility weights for alternative observation models.

Chapter 3

Efficient Estimation and Forecasting for Multivariate Discrete Hawkes Processes

3.1 Introduction and Motivation

This chapter develops a practical and scalable toolkit for maximum likelihood estimation, gradient computation, and multi-step forecasting in multivariate *discrete-time* Hawkes models on a fixed grid. The statistical setting and notation for likelihood-based estimation are given in Chapter 2, especially Section 2.2. Formal definitions of temporal point processes and Hawkes processes appear in Section 2.5, Section 2.6.1, Section 2.6.3, and Section 2.7. Here we specialise those definitions to kernel choices that deliver closed-form recursions: exponential in continuous time (for context later) and geometric on the grid (our main focus). These recursions allow efficient likelihood, score, and forecast evaluation.

A naive implementation recomputes past influence at every grid time and for every component. When most grid times have no events, that is wasteful. We replace those repeated resummations by three ingredients that match computational effort to information: an event-time identity for the log-likelihood that collapses the compensator to sums over event times only, constant-time event-time recursions for intensities and for all score components, and a time-step recursion that drives multi-step mean forecasting. The continuous-time analogue with exponential kernels admits a simple event-time state for intensities and a recursive likelihood; we summarise this later as a bridge for context. The Appendix collects complete proofs for the discrete-time results.

A primary reason for developing the discrete-time toolkit here is to support the marked setting analysed later in Chapter 4. The event-time likelihood identity, constant-time intensity recursion, and matched gradient recursions derived in this chapter will be reused verbatim and then extended to handle marks through mark-dependent baselines and excitation kernels. The same optimisation principles from Section 2.2 apply, with the event-time sweeps providing the computational core. The results below will be used as the foundation for the fitting

the application-driven model in Chapter 4. In that chapter, we consider a marked discrete Hawkes process, which is a generalization of the discrete Hawkes process in this chapter. Nevertheless, the techniques for computationally efficient inference and simulation covered here will be used in the application.

3.2 Discrete-Time Hawkes: Model and Notation

We specialise the discrete-time Hawkes process introduced in Section 2.7 to a multivariate, unmarked Poisson observation model, while keeping the notation of Chapter 2. Let $t = 1,\dots,N$ index the discrete time grid, and let $Y_t^{(m)} \in \{0,1,2,\dots\}$ denote the count in interval $(t-1,t]$ for component $m \in \{1,\dots,M\}$. Given the history

$$\mathscr{H}_{t-1} = \{\mathbf{Y}_1,\dots,\mathbf{Y}_{t-1}\},$$

we assume

$$Y_t^{(m)} \mid \mathscr{H}_{t-1} \sim \text{Poisson}\big(\lambda^{(m)}(t)\big).$$

The conditional mean intensity follows the lag formulation introduced in Section 2.7:

$$\lambda^{(m)}(t) = \mu^{(m)} + \sum_{l=1}^{M}\sum_{d=1}^{t-1} Y_{t-d}^{(l)}\,\Phi_{l,m}(d), \qquad \Phi_{l,m}(d) \geq 0. \tag{3.1}$$

Here $\mu^{(m)}$ is the baseline intensity for component m, and $\Phi_{l,m}(d)$ quantifies the contribution at lag d of past events in component l to the current expected count in component m.

It is convenient to decompose the lag kernel as

$$\Phi_{l,m}(d) = K_{l,m}\,g_{l,m}(d), \qquad \sum_{d\geq 1} g_{l,m}(d) = 1,$$

where $K_{l,m} \geq 0$ controls the total excitation transmitted from component l to component m, and $g_{l,m}(d)$ describes how that excitation is distributed across lags. A convenient and widely used discrete-time choice is the geometric kernel (Browning et al., 2022, 2021; Kirchner, 2016),

$$g_{l,m}(d) = \beta_{l,m}(1-\beta_{l,m})^{d-1}, \qquad 0 < \beta_{l,m} < 1, \quad d \in \mathbb{N}^+. \tag{3.2}$$

This kernel is the discrete-time analogue of exponential decay, with $\beta_{l,m}$ controlling the rate at which excitation from past events diminishes across successive lags.

Under this parameterisation, $K_{l,m}$ is the expected number of direct offspring of type m generated by a single event of type l, while $\beta_{l,m}$ determines the memory scale of that excitation. To ensure stability of the multivariate process, we require the spectral radius of the branching matrix $K = (K_{l,m})_{l,m=1}^{M}$ to satisfy

$$\rho(K) < 1.$$

Finally, let

$$\tau := \Big\{ t \in \{1,\dots,N\} : \sum_{m=1}^{M} Y_t^{(m)} > 0 \Big\}, \qquad t_1 < \cdots < t_{N_{\text{events}}}$$

denote the ordered set of event times, that is, the time bins in which at least one event is observed. In what follows, all sums over t that contain a data term are understood to be restricted to τ.

3.3 Event-Time Log-Likelihood

We begin from the Poisson log-likelihood

$$\log L(\theta \mid \tau) = \sum_{t=1}^{N} \sum_{m=1}^{M} \Big\{ Y_t^{(m)} \log \lambda^{(m)}(t) - \lambda^{(m)}(t) \Big\}.$$

This separates into a data term and a compensator:

$$\log L(\theta \mid \tau) = \sum_{t \in \tau} \sum_{m=1}^{M} Y_t^{(m)} \log \lambda^{(m)}(t) - \sum_{t=1}^{N} \sum_{m=1}^{M} \lambda^{(m)}(t),$$

where $\tau = \{ t \in \{1,\dots,N\} : \sum_{m=1}^{M} Y_t^{(m)} > 0 \}$ is the set of event times. The data term contributes only at times $t \in \tau$, since $Y_t^{(m)} = 0$ for all m whenever no events occur in bin t. By contrast, the compensator contributes at every grid time $t = 1,\dots,N$, since it accounts for the full cumulative intensity over the observation window.

A naive implementation evaluates $\lambda^{(m)}(t)$ from scratch for every t and every m:

$$\lambda^{(m)}(t) = \mu^{(m)} + \sum_{l=1}^{M} \sum_{d=1}^{t-1} Y_{t-d}^{(l)} K_{l,m}\, g_{l,m}(d).$$

The inner sum has length $t-1$, so the total work is proportional to $1+2+\cdots+(N-1)$, which is $O(N^2)$ per pair (l,m) and $O(M^2N^2)$ overall. Most of that effort is spent at grid times with no events, purely to evaluate the compensator.

With any nonnegative lag kernel $g_{l,m}(\cdot)$ normalised on the grid so that $\sum_{d=1}^{\infty} g_{l,m}(d) = 1$, this computation can be simplified by collapsing the compensator into a form that depends on event times only. Write the cumulative kernel

$$G_{l,m}(u) := \sum_{s=1}^{u} g_{l,m}(s), \qquad u \in \mathbb{N}^+.$$

Since all sums here are finite, we may reorder them directly. Doing so moves the time summation inside the kernel contribution and replaces the traversal of all $t = 1, \ldots, N$ by a sum over the observed event times $\{t_i\}$. For the geometric special case (3.2),

$$G_{l,m}(u) = 1 - (1 - \beta_{l,m})^u.$$

Proposition 3.3.1 (Event-time likelihood identity)**.** *Under* (3.1) *with* $\Phi_{l,m}(d) = K_{l,m} g_{l,m}(d)$ *and any normalised nonnegative kernel* $g_{l,m}$*,*

$$\log L(\theta \mid \tau) = \sum_{m=1}^{M} \sum_{t \in \tau} Y_t^{(m)} \log \lambda^{(m)}(t) \; - \; \sum_{m=1}^{M} \sum_{l=1}^{M} K_{l,m} \sum_{t \in \tau} Y_t^{(l)} \, G_{l,m}(N - t) \; - \; N \sum_{m=1}^{M} \mu^{(m)}. \quad (3.3)$$

Proof. See Appendix A.

The right-hand side has three components. The last term is $O(MN)$. The middle term is $O(M^2 N_{\text{events}})$ because it sums only over event times. The first term requires $\lambda^{(m)}(t)$ at $t \in \tau$, which we will obtain using the event-time recursion in Section 3.4. The resulting evaluation cost matches $O(M^2 N_{\text{events}})$.

3.4 Constant-Time Intensity Recursion at Event Times

We now specialise the excitation to the geometric form, so that for every pair (l, m)

$$g_{l,m}(d) = \beta_{l,m} (1 - \beta_{l,m})^{d-1}, \qquad d \in \mathbb{N}^+, \;\; 0 < \beta_{l,m} < 1,$$

and hence $\Phi_{l,m}(d) = K_{l,m} \, g_{l,m}(d)$. Under this kernel each past contribution decays by the same factor $(1 - \beta_{l,m})$ per grid step, which allows all historical influence to be compressed into a single event-time state that carries forward between successive event times.

For $j \geq 1$ define

$$R(j, l, m) := \sum_{i : t_i < t_j} Y_{t_i}^{(l)} \, \beta_{l,m} (1 - \beta_{l,m})^{t_j - t_i - 1}, \qquad R(1, l, m) = 0. \quad (3.4)$$

Write $\Delta_j := t_{j+1} - t_j$. The recursion for R and its link to the intensity are given next in Proposition 3.4.1.

Algorithm 1 Multi-step mean forecasting on the grid

Input: forecast origin t, observed count vector $\mathbf{Y}_t$, current state $R^3(t,l,m)$, parameters, horizon h.
for $l = 1$ to M **do**
  for $m = 1$ to M **do**
    $\widehat{R}^3(t+1,l,m) \leftarrow (1-\beta_{l,m})\,R^3(t,l,m) + Y_t^{(l)}\,\beta_{l,m}$
  end for
end for
for $m = 1$ to M **do**
  $\widehat{\lambda}^{(m)}(t+1) \leftarrow \mu^{(m)} + \sum_l K_{l,m}\widehat{R}^3(t+1,l,m)$
end for
for $k = 1$ to $h-1$ **do**
  for $l = 1$ to M **do**
    for $m = 1$ to M **do**
      $\widehat{R}^3(t+k+1,l,m) \leftarrow (1-\beta_{l,m})\,\widehat{R}^3(t+k,l,m) + \widehat{\lambda}^{(l)}(t+k)\,\beta_{l,m}$
    end for
  end for
  for $m = 1$ to M **do**
    $\widehat{\lambda}^{(m)}(t+k+1) \leftarrow \mu^{(m)} + \sum_l K_{l,m}\widehat{R}^3(t+k+1,l,m)$
  end for
end for
Output: $\{\widehat{\lambda}^{(m)}(t+k)\}_{k=1}^{h}$.

Proposition 3.4.1 (Event-time recursion)**.** *For $j = 1,\dots,N_{events}-1$ and all (l,m),*

$$R(j+1,l,m) = (1-\beta_{l,m})^{\Delta_j} R(j,l,m) \;+\; Y_{t_j}^{(l)}\,\beta_{l,m}(1-\beta_{l,m})^{\Delta_j - 1}, \tag{3.5}$$

and

$$\lambda^{(m)}(t_j) = \mu^{(m)} + \sum_{l=1}^{M} K_{l,m} R(j,l,m). \tag{3.6}$$

Proof. See Appendix A.

3.5 Gradients and Fast Recursions

Optimisers in Section 2.2 require the score vector. A naive score recomputes long history sums at every event time, which makes the gradient far more expensive than the likelihood. The primitive score formulas make the algebra transparent. We then show how to evaluate them in constant time per event by augmenting the event-time state with two additional accumulators that track derivatives of the geometric kernel.

Throughout let $g_{p,q}(u) = \beta_{p,q}(1-\beta_{p,q})^{u-1}$ and $G_{p,q}(u) = 1-(1-\beta_{p,q})^{u}$.

Lemma 3.5.1 (Primitive score formulas)**.** *Under* (3.1) *with* $\Phi_{l,m}(d) = K_{l,m} g_{l,m}(d)$*,*

$$\frac{\partial}{\partial \mu^{(p)}} \log L(\theta \mid \tau) = \sum_{t \in \tau} \frac{Y_t^{(p)}}{\lambda^{(p)}(t)} - N, \tag{3.7}$$

$$\frac{\partial}{\partial K_{p,q}} \log L(\theta \mid \tau) = \sum_{t \in \tau} \frac{Y_t^{(q)}}{\lambda^{(q)}(t)} \sum_{i:\, t_i < t} Y_{t_i}^{(p)} g_{p,q}(t - t_i) - \sum_{t \in \tau} Y_t^{(p)} G_{p,q}(N - t), \tag{3.8}$$

$$\frac{\partial}{\partial \beta_{p,q}} \log L(\theta \mid \tau) = \sum_{t \in \tau} \frac{Y_t^{(q)}}{\lambda^{(q)}(t)} K_{p,q} \sum_{i:\, t_i < t} Y_{t_i}^{(p)} \frac{\partial}{\partial \beta_{p,q}} g_{p,q}(t - t_i) - \sum_{t \in \tau} K_{p,q} Y_t^{(p)} \frac{\partial}{\partial \beta_{p,q}} G_{p,q}(N - t), \tag{3.9}$$

with

$$\frac{\partial}{\partial \beta} g(u) = (1-\beta)^{u-2}\big(1 - \beta u\big), \qquad \frac{\partial}{\partial \beta} G(u) = u(1-\beta)^{u-1}.$$

Proof. See Appendix A.

The derivative $\partial g / \partial \beta$ splits into a difference of two geometric terms. This motivates two additional event-time accumulators,

$$R^1(j,l,m) := \sum_{i:\, t_i < t_j} Y_{t_i}^{(l)} (1 - \beta_{l,m})^{t_j - t_i - 2}, \qquad R^2(j,l,m) := \sum_{i:\, t_i < t_j} Y_{t_i}^{(l)} (1 - \beta_{l,m})^{t_j - t_i - 2} \beta_{l,m} (t_j - t_i), \tag{3.10}$$

which update with the same constant-time structure as R.

Proposition 3.5.2 (Fast event-time updates for R^1 and R^2)**.** *With* $\Delta_j = t_{j+1} - t_j$*,*

$$R^1(j{+}1,l,m) = (1 - \beta_{l,m})^{\Delta_j} R^1(j,l,m) + Y_{t_j}^{(l)} (1 - \beta_{l,m})^{\Delta_j - 2}, \tag{3.11}$$

$$R^2(j{+}1,l,m) = (1 - \beta_{l,m})^{\Delta_j} R^2(j,l,m) + \beta_{l,m} \Delta_j R^1(j{+}1,l,m). \tag{3.12}$$

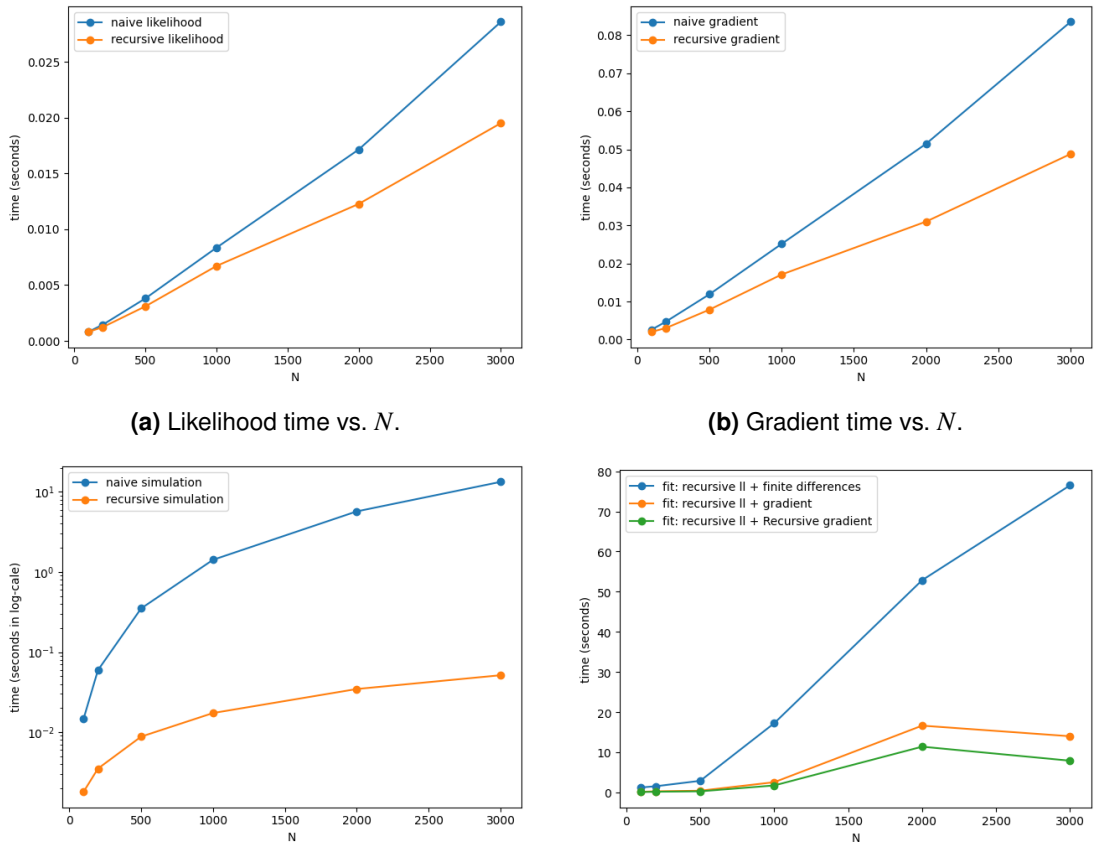
Proof. See Appendix A.

Substituting $\lambda^{(m)}(t_j) = \mu^{(m)} + \sum_l K_{l,m} R(j,l,m)$ and rewriting the inner sums in the primitive scores using R, R^1, R^2 yields the fast event-time scores

$$\frac{\partial}{\partial \mu^{(p)}} \log L(\theta \mid \tau) = \sum_j \frac{Y_{t_j}^{(p)}}{\mu^{(p)} + \sum_{l=1}^M K_{l,p} R(j,l,p)} - N, \tag{3.13}$$

$$\frac{\partial}{\partial K_{p,q}} \log L(\theta \mid \tau) = \sum_j \frac{Y_{t_j}^{(q)} R(j,p,q)}{\mu^{(q)} + \sum_{l=1}^M K_{l,q} R(j,l,q)} - \sum_{t \in \tau} Y_t^{(p)} \big[1 - (1 - \beta_{p,q})^{N-t}\big], \tag{3.14}$$

$$\frac{\partial}{\partial \beta_{p,q}} \log L(\theta \mid \tau) = \sum_j \frac{Y_{t_j}^{(q)}}{\mu^{(q)} + \sum_{l=1}^M K_{l,q} R(j,l,q)} K_{p,q} \Big(R^1(j,p,q) - R^2(j,p,q) \Big) - \sum_{t \in \tau} K_{p,q} Y_t^{(p)} (N-t) (1 - \beta_{p,q})^{N-t-1}. \tag{3.15}$$



(a) Likelihood time vs. N.

(b) Gradient time vs. N.

(c) Simulation time vs. N (log scale).

(d) Fitting time vs. N.

Figure 3.1: Computational benchmarks for the discrete-time Hawkes recursions with geometric kernels. Horizontal axis in all panels: N, the number of grid points. Curves compare naive implementations that re-sum past contributions at every step with the event-time and time-step recursions of Sections 3.3–3.6. Experiments use $M = 3$ components. For each N, dataset, initialisation, and seeds are held fixed.

3.6 Forecasting

We consider h-step-ahead mean forecasting conditional on the observed history up to time t, denoted $\mathscr{H}_t$. The object of interest is

$$\widehat{Y}_{t+k}^{(m)} := \mathbb{E}\left[Y_{t+k}^{(m)} \mid \mathscr{H}_t\right], \qquad k = 1, \dots, h.$$

Since the conditional mean of the count process is the intensity, the one-step-ahead forecast satisfies

$$\widehat{Y}_{t+1}^{(m)} = \mathbb{E}\left[Y_{t+1}^{(m)} \mid \mathscr{H}_t\right] = \lambda^{(m)}(t+1 \mid \mathscr{H}_t).$$

A naive calculation based directly on (3.1) is expensive. If one were to evaluate the model equation at time $t+k$, one would write

$$\lambda^{(m)}(t+k) = \mu^{(m)} + \sum_{l=1}^{M} \sum_{d=1}^{t+k-1} Y_{t+k-d}^{(l)} \Phi_{l,m}(d).$$

However, for forecasting from time t, this expression cannot be used literally when $k \geq 2$, because the counts $Y_{t+1}^{(l)}, \dots, Y_{t+k-1}^{(l)}$ are not yet observed. Thus, a true h-step-ahead forecast must be defined conditionally on $\mathscr{H}_t$, with future unknown counts replaced by their conditional mean forecasts.

Formally, for $k \geq 1$, define

$$\widehat{\lambda}^{(m)}(t+k \mid \mathscr{H}_t) := \mathbb{E}\left[Y_{t+k}^{(m)} \mid \mathscr{H}_t\right].$$

Then the forecasting recursion is obtained by writing

$$\widehat{\lambda}^{(m)}(t+k \mid \mathscr{H}_t) = \mu^{(m)} + \sum_{l=1}^{M} \sum_{d=1}^{t+k-1} \widetilde{Y}_{t+k-d}^{(l)} \Phi_{l,m}(d),$$

where

$$\widetilde{Y}_s^{(l)} = \begin{cases} Y_s^{(l)}, & s \leq t, \\ \widehat{\lambda}^{(l)}(s \mid \mathscr{H}_t), & s > t. \end{cases}$$

Thus, observed counts are used up to time t, while future counts are replaced recursively by their forecast means. For readability, we will often suppress the conditioning on $\mathscr{H}_t$ and write $\widehat{\lambda}^{(m)}(t+k)$, with the understanding that all forecast quantities are conditional on the observed history at forecast origin t. Since $R^3(t,l,m)$ only contains contributions from times strictly before t, the one-step-ahead state $\widehat{R}^3(t+1,l,m)$ is first obtained by updating with the observed count $Y_t^{(l)}$.

Without a state representation, evaluating forecasts in this way is costly. At each forecast step, a naive computation re-sums all past contributions, so at time $t+k$ the work is proportional to $t+k-1$. Over $k=1,\dots,h$, the total work is therefore

$$\sum_{k=1}^{h}(t+k-1) = ht + \frac{h(h-1)}{2},$$

which is $O(ht+h^2)$. If forecasts are produced repeatedly across a full window of length N, the total naive cost is on the order of $\sum_{s=1}^{N} s = O(N^2)$ per pair (l,m), that is $O(M^2N^2)$ overall. The cost explosion arises because the same historical contributions are re-summed at every forecast step.

To avoid this, we carry a single time-step state that propagates all past influence forward with a constant-time update per step:

$$R^3(t,l,m) := \sum_{i:t_i<t} Y_{t_i}^{(l)}\,\beta_{l,m}(1-\beta_{l,m})^{t-t_i-1}, \qquad R^3(1,l,m) = 0. \tag{3.16}$$

Lemma 3.6.1 (Time-step recursion)**.** *For $t = 1,\dots,N-1$,*

$$R^3(t+1,l,m) = (1-\beta_{l,m})\,R^3(t,l,m) + Y_t^{(l)}\,\beta_{l,m}, \qquad \lambda^{(m)}(t) = \mu^{(m)} + \sum_{l=1}^{M} K_{l,m} R^3(t,l,m). \tag{3.17}$$

Proof. See Appendix A.

At forecast origin t, the one-step-ahead mean is

$$\widehat{Y}_{t+1}^{(m)} = \widehat{\lambda}^{(m)}(t+1 \mid \mathscr{H}_t).$$

For $k \geq 1$, true multi-step forecasting is obtained by replacing the unknown future counts in the recursion by their forecast means:

$$\widehat{R}^3(t+k+1,l,m) = (1-\beta_{l,m})\,\widehat{R}^3(t+k,l,m) + \widehat{\lambda}^{(l)}(t+k \mid \mathscr{H}_t)\,\beta_{l,m},$$

and

$$\widehat{\lambda}^{(m)}(t+k \mid \mathscr{H}_t) = \mu^{(m)} + \sum_{l=1}^{M} K_{l,m}\widehat{R}^3(t+k,l,m).$$

This yields a closed linear recursion with $O(M^2)$ work per forecast step.

Algorithm 2 Multi-step mean forecasting on the grid

Input: forecast origin t, current state $R^3(t,l,m)$, parameters, horizon h.
Set $\widehat{R}^3(t,l,m) \leftarrow R^3(t,l,m)$.
for $k = 1$ to h **do**
  for $m = 1$ to M **do**
    $\widehat{\lambda}^{(m)}(t+k) \leftarrow \mu^{(m)} + \sum_l K_{l,m} \widehat{R}^3(t+k,l,m)$
  end for
  if $k < h$ **then**
    for $l = 1$ to M **do**
      for $m = 1$ to M **do**
        $\widehat{R}^3(t+k+1,l,m) \leftarrow (1-\beta_{l,m})\,\widehat{R}^3(t+k,l,m) + \widehat{\lambda}^{(l)}(t+k)\,\beta_{l,m}$
      end for
    end for
  end if
end for
Output: $\{\widehat{\lambda}^{(m)}(t+k)\}_{k=1}^{h}$.

3.7 Conclusion

This section established an event-time likelihood identity, constant-time intensity recursions, fast score evaluations, and a time-step forecasting state for multivariate discrete-time Hawkes processes on a grid. These tools cut computation from $O(M^2N^2)$ to $O(M^2N_{\text{events}})$ for likelihood and gradients, and to $O(M^2h)$ per forecast horizon, h. In Chapter 4 we build on this foundation to fit an application-driven *marked* discrete Hawkes model. The same identities and recursions are reused there with mark-dependent baselines and excitation kernels, providing the computational core for estimation, inference, and forecasting in the application.

Chapter 4

Estimation of Multivariate Marked Discrete Hawkes Processes: An Application to Incident Monitoring

4.1 Introduction

Our goal in this chapter is to examine the timing of violent behavior in a forensic psychiatric hospital, focusing on the interactions between incidents, in order to develop predictive models. These models can inform better protocols, such as optimizing daily routines like meal times and treatment schedules, both of which are linked to increased stress levels (Hinsby & Baker, 2004; Meehan, McIntosh, & Bergen, 2006; Wright, Duxbury, Baker, & Crumpton, 2014). The data, recorded across different hospital wards, display clustering, seasonality, and ward interactions, which complicate the analysis. This chapter builds directly on the discrete-time likelihood and recursion results developed in Chapter 3. Here we extend the framework to a *multivariate marked* discrete Hawkes process tailored to incident monitoring. Marks encode interval-level attributes (for example, alarm status, incident type, or severity) and enter through mark-dependent excitation kernels and baselines. The event-time likelihood identity and constant-time recursions from Chapter 3 remain the computational core, with minor modifications to accommodate mark-specific kernels and cross-excitation. We use the same notation and estimation principles as in Section 2.7 and Section 2.2: likelihood evaluation and gradients are carried out with event-time sweeps, and stability and identifiability are treated as before, now with mark effects. The result is an efficient estimation and forecasting toolkit for high-resolution, marked incident data, which we apply to the hospital monitoring setting.

Previous studies, such as (Barnard, Robbins, Newman, & Carrera, 1984; Beck et al., 2018), primarily used univariate methods which did not account for potential interactions between wards. (Beck et al., 2018) found bursts of aggression that were inconsistent with Poisson models, emphasizing the need for more sophisticated approaches. Seasonality in event timing

– tied to hospital routines – and the clustering of incidents among repeat offenders (Barnard et al., 1984; Hinsby & Baker, 2004; Meehan et al., 2006; Wright et al., 2014) highlight the importance of timely interventions. Incident reporting systems (Stavropoulou, Doherty, & Tosey, 2015) have also demonstrated value in improving patient safety.

The hospital's multivariate data, with incidents recorded across wards, requires modeling cross-ward influences. A further key feature is the hospital alarm system, which can sound during particularly severe incidents. The alarm is audible hospital-wide and prompts staff intervention; operationally, we encode it as a binary mark attached to each event. Because alarms are heard beyond the originating ward, they plausibly increase the chance that activity in one ward coincides with or influences events in others. Capturing this global signal is therefore essential and we include the alarm mark as a covariate in our discrete marked multivariate Hawkes specification.

We make three contributions. First, we rewrite the discrete-time Hawkes likelihood in event time and derive constant-cost, event-indexed recursions for the likelihood and the score, so evaluation scales quadratically with the number of dimensions and linearly with the number of observed events, enabling routine fitting in genuinely multivariate settings. Second, we propose a marked specification that separates ordinary within-ward excitation from an alarm-mediated cross-ward channel, with a simple day/night scaling; parameters are constrained to be nonnegative and estimated with sparsity penalties that encourage few cross-ward effects and a parsimonious alarm channel. Third, we provide a compatible forecasting recursion and simulators for out-of-sample assessment and scenario generation. We emphasize predictive interpretation—alarm effects are treated as associations rather than causal impacts—and verify stability by requiring the spectral radius of the cross-series excitation matrix to be less than one. Together these elements form a reusable recipe for fast estimation and forecasting of multivariate, marked discrete Hawkes processes in applied settings.

The paper is structured as follows: Section 4.2 describes the data, Section 4.3 introduces the Hawkes process, and Section 4.4 defines the discrete marked Hawkes process and a regularization strategy. Section 4.5 presents efficient estimation and simulation methods, while Section 4.7 reports model results. Section 4.8 summarizes contributions and future directions. The supplementary material includes detailed algorithms and additional implementation notes for non-marked discrete Hawkes processes.

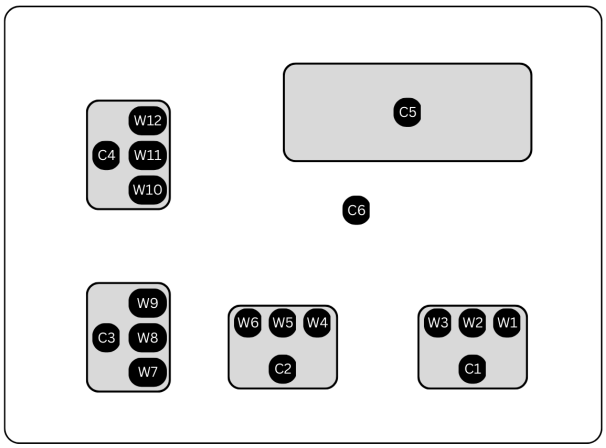


Figure 4.1: Ward Layout: W1–W12 denote wards; C1–C4 are section-specific communal areas; C5 is a hospital-wide communal space; C6 aggregates external/common areas (e.g., corridors and outdoor spaces)

4.2 Hospital Incidents Data

We consider a dataset of event times that represents violent behavior in 18 locations within a forensic psychiatric hospital, covering both wards and communal areas. The data, recorded by hospital employees, were anonymized with patient numbers to prevent identification and access was granted by the National Health Service. It spans a nine-year period from June 26, 2012, to September 30, 2021, although we only used data up to January 1, 2020, to avoid COVID pandemic-related behavioral changes that fall outside typical operations. The hospital consists of four sections, each containing three wards and a communal area, as depicted in Figure 4.1. In addition, the hospital has indoor and outdoor communal areas available to all patients, yielding 18 distinct locations where violent events may occur.

Table 4.1 provides a summary of the number of incidents in different sections of the hospital over study period. It can be seen that the number of incidents in communal areas is significantly lower compared to those within the wards. Nevertheless, communal spaces may facilitate information flow between wards (e.g., through staff movement or patient transfers), so we retain them in our analysis. Regarding C6, we lack detailed information on the locations of incidents within this area, implying inhomogeneous interactions with other sections; including C6 in the multivariate model would substantially increase the parameter count and complicate identifiability, so we treat C6 conservatively in what follows.

A visual inspection of the event times, shown in Figure 4.2, suggests burst-like behavior. We plotted event times as vertical lines for a selection of wards from each hospital section, revealing clear within-ward clustering over short horizons. This pattern is consistent with the self-exciting dynamics that Hawkes processes are designed to capture. The study site

Section	# Events	Section	# Events
C1	22	W7	612
W1	725	W8	283
W2	198	W9	129
W3	82	C4	18
C2	41	W10	1045
W4	520	W11	783
W5	2285	W12	82
W6	337	C5	24
C3	6	C6	191

Table 4.1: Total number of events recorded in each area of the hospital.

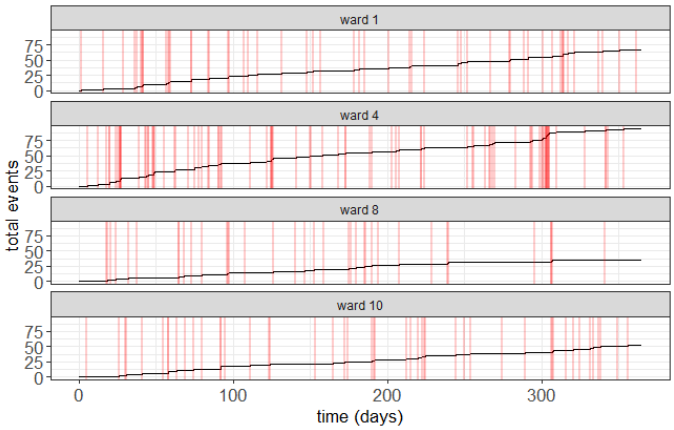


Figure 4.2: Cumulative sum of events for a selection of wards over a one year period. Short vertical tick marks indicate recorded event times in the corresponding ward.

is a high-security forensic hospital serving patients with severe and enduring mental health conditions who cannot be housed in less restrictive settings. Many patients arrive via the criminal justice system, which contributes to a high baseline rate of incidents even in the absence of specific triggers.

The dataset presents several challenges. First, event times were manually recorded by on-duty staff who had to manage incidents before logging them, so timestamps are commonly rounded to the nearest five minutes. This leads to uncertainty in the exact timing of events, and simultaneous events at the recorded resolution, which violates the no-ties assumption often made in continuous-time Hawkes models. A discrete-time formulation naturally accommodates such ties. Table 4.2 shows the proportion of total events for which multiple incidents occurred within the same five-minute interval, underscoring the need for a discrete approach.

A second challenge is computational. In Hawkes models (Hawkes, 1971), including the discrete variant used here, the conditional mean (intensity) aggregates the influence of past events. Evaluating this intensity at all time points scales with the number of timestamps, which is infeasible for long series at five-minute resolution. Following standard recursions

Number of simultaneous events	Proportion of total events
1	90%
2	9%
≥ 3	1%

Table 4.2: Proportion of total events where there were multiple events occurring in the same five-minute interval.

(continuous-time analogues in (Ozaki, 1979) and discrete-time analogues used here), the intensity can be written in terms of cumulative kernel mass, allowing event-indexed updates that scale with the number of events rather than the number of time bins. This substantially reduces runtime when events are sparse relative to the grid, although repeated evaluations can still be nontrivial when the total number of events is large.

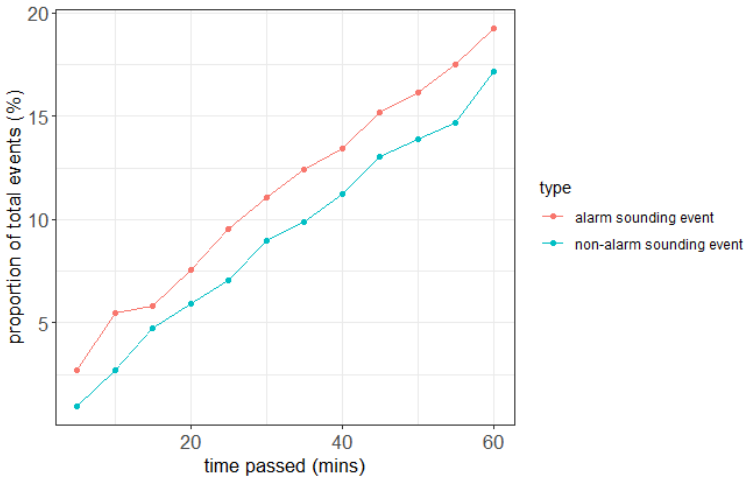


Figure 4.3: Proportion of the total events over the entire hospital that occur at a given amount of time immediately after another event which did/did not caused an alarm to be sounded.

Alarms are triggered during particularly violent incidents, as judged by on-shift staff, and require urgent assistance from hospital employees. Because alarms are audible hospital-wide, they provide a global signal that may raise short-term incident rates outside the originating ward. We therefore treat the alarm as a binary mark attached to each event and allow alarm-marked events to contribute distinct excitation in the multivariate model. To explore this empirically, we compare the subsequent event profiles following incidents with and without alarms. Figure 4.3 summarizes the proportion of events at short lags by alarm status; post-alarm periods show higher near-term activity, consistent with alarms functioning as a global signal. We note that alarms tend to coincide with more severe incidents, so part of the difference may reflect severity rather than the alarm signal per se.

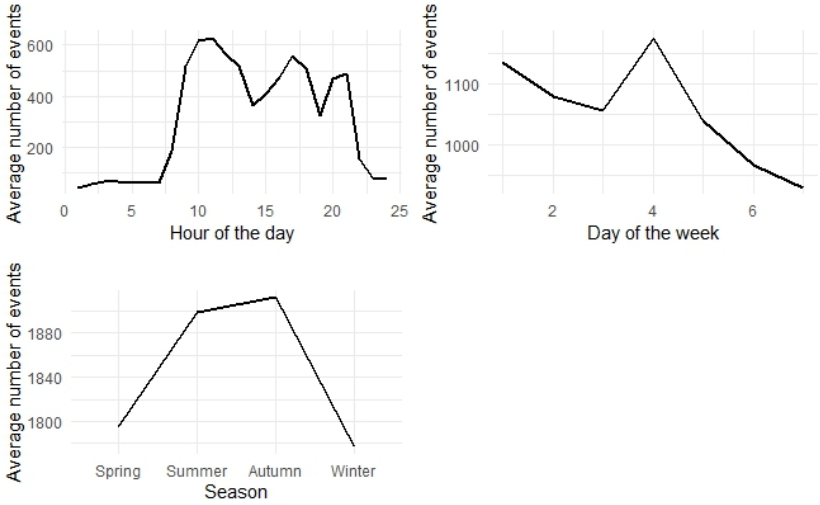


Figure 4.4: Average number of events across the hospital, broken down by hour of the day, day of the week, and season.

Next, we investigate seasonality in our dataset. Previous studies have discussed seasonal characteristics in similar data (Beck et al., 2018; Peluola, Mela, & Adelugba, 2012). For example, (Peluola et al., 2012) analyzed the records of violent incidents over five years in a multilevel secure forensic hospital in Canada, finding that violent events were more frequent during winter months, often linked to unstructured activities. Figure 4.4 presents the average number of events per hour of the day, day of the week, and per season. It can be seen that there is a marked decrease in events between roughly 11 p.m. and 7 a.m., likely reflecting night operations when patients are confined to dorms. We also observe higher average counts in summer and autumn. Studies have linked temperature and weather to aggression and violence, which may contribute to this pattern, although (Peluola et al., 2012) report the opposite in their setting. By day of week, there is a roughly linear decline from Monday to Sunday with a distinct peak on Wednesday; this may reflect operational factors (e.g., visiting or appointment schedules), although we lack the metadata to confirm.

Finally, Figure 4.5 illustrates event frequencies for three high-incident patients, highlighting ward allocation and management policies. In our data, the most violent patients spend the majority of time in a single ward, as seen in the concentrated event distributions for patients 102, 179, and 197. This reflects deliberate containment strategies rather than random allocation and contributes to between-ward variation in incident counts.

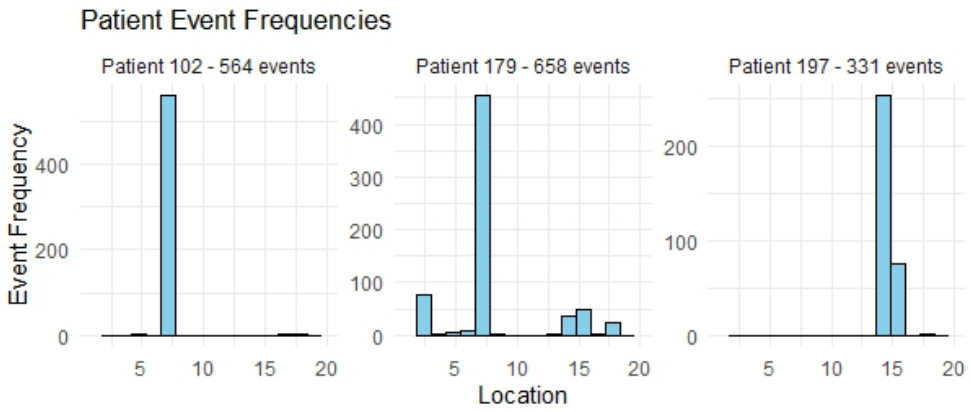


Figure 4.5: Histograms showing the distribution of violent events initiated by the three most violent patients across different areas of the hospital

4.3 Background

Our aim is to develop a model for the number of events in each area of the forensic hospital. In order to deal with the discrete bins, we use the discrete Hawkes framework, which we will introduce in this section. A marked discrete Hawkes process is an extension of the discrete Hawkes process that describes not only the timings and locations of random events but also additional information attached to each event, known as a mark. Although alternative models, such as general clustering processes (Xu et al., 2020; J. Zhang et al., 2023), could be used to represent burst-like behavior, they do not explicitly model the decaying influence of past events over time. The Hawkes process, in contrast, effectively captures these dynamics, making it a more suitable choice for modeling the complex dependencies observed in our data. In later sections, we use an alarm indicator as the mark, allowing events that triggered an alarm to have different excitation than non-alarm events. The remainder of this section reviews the Hawkes process and its discrete and multivariate discrete variants.

4.3.1 Discrete Hawkes Process

Consider a marked univariate discrete Hawkes process (Browning et al., 2021), denoted by N, where N_t is the cumulative number of events up to and including interval $(t-1,t]$. Define

$$Y_t \;=\; \mathrm{N}_t - \mathrm{N}_{t-1}$$

to be the count in interval $(t-1,t]$, and let m_t denote the mark associated with that interval. The mark m_t represents any observed attribute attached to the events or context in interval $(t-1,t]$, for example an event type, a severity category, or a binary alarm indicator. The history available strictly before time t is therefore

$$\mathrm{H}_{t-1} \;=\; \{(Y_s, m_s) : s \le t-1\}.$$

The conditional mean count at time t is

$$\lambda(t) = \mathbb{E}[Y_t \mid \mathrm{H}_{t-1}] = \mu(t) + \sum_{i:t_i \le t-1} Y_{t_i} \Phi(t - t_i, m_{t_i}), \tag{4.1}$$

where $\mu(t) > 0$ is the baseline rate and t_i are the previous time indices such that $Y_{t_i} > 0$. The function $\Phi(\cdot,\cdot) \ge 0$ is a discrete excitation kernel that governs how previous counts increase the current expected count, with the mark m_{t_i} modulating the strength or persistence of that excitation. We assume the conditional observation model

$$Y_t \mid \mathrm{H}_{t-1} \sim \mathrm{Poisson}\big(\lambda(t)\big).$$

4.3.2 Multivariate Discrete Hawkes Process

Consider now a marked multivariate discrete Hawkes process $\mathbf{N}$, where

$$\mathbf{N}_t = \big(\mathrm{N}_t^{(1)}, \mathrm{N}_t^{(2)}, \ldots, \mathrm{N}_t^{(M)}\big)$$

records the cumulative number of events up to and including interval $(t-1,t]$ in each of the M dimensions. For each dimension m, define

$$Y_t^{(m)} = \mathrm{N}_t^{(m)} - \mathrm{N}_{t-1}^{(m)},$$

and write

$$\mathbf{Y}_t = \big(Y_t^{(1)}, Y_t^{(2)}, \ldots, Y_t^{(M)}\big)$$

for the vector of counts in interval $(t-1,t]$. Let m_t denote the mark associated with interval $(t-1,t]$; as above, this may encode any observed attribute relevant to excitation, such as event type, severity, or an external alarm indicator. The history strictly prior to time t is then

$$\mathbf{H}_{t-1} = \{(\mathbf{Y}_s, m_s) : s \le t-1\}.$$

The conditional mean count in dimension m at time t is defined by

$$\lambda^{(m)}(t) = \mathbb{E}\Big[Y_t^{(m)} \mid \mathbf{H}_{t-1}\Big] = \mu^{(m)}(t) + \sum_{l=1}^{M} \sum_{i:t_i \le t-1} Y_{t_i}^{(l)} \Phi_{l,m}(t - t_i, m_{t_i}), \tag{4.2}$$

where $\mu^{(m)}(t) > 0$ is the baseline rate in dimension m, and t_i are times prior to t such that at least one event occurred in interval $(t_i - 1, t_i]$. The function $\Phi_{l,m}(\cdot,\cdot) \ge 0$ is a discrete excitation kernel governing how past events in dimension l affect the expected count in dimension m, with the mark m_{t_i} modulating this effect. When $l \ne m$, this corresponds to cross-excitation between dimensions.

For each $m = 1, \ldots, M$, we assume the conditional observation model

$$Y_t^{(m)} \mid \mathbf{H}_{t-1} \sim \text{Poisson}\big(\lambda^{(m)}(t)\big).$$

4.4 Model

We model the occurrences of violent events in the hospital using a specific form of the discrete multivariate Hawkes model as in Section 4.3.2. Our model introduces a particular form of the discrete excitation function, baseline intensity, and formulation of marks. We begin by discussing the choice of excitation functions.

4.4.1 Self- and Cross-Excitation

Events in different hospital wards often occur in bursts, with instances in one ward triggering others in the same or different wards shortly after. These bursts might result from patients directly witnessing events or hearing about them in the communal areas shared by multiple wards. This necessitates modeling excitation across wards using a multivariate Hawkes process rather than treating wards independently. While wards are grouped into sections of three, communal areas like C1 to C5 (Figure 4.1) have fewer than 40 events, making them unsuitable for modeling. Communal area C6 encompasses all other hospital areas accessible to patients, leading to inhomogeneous relationships between C6 events and others. Thus, we model event occurrences with a twelve-dimensional discrete Hawkes process corresponding to the twelve wards.

The excitation kernels $\Phi_{l,m}(\cdot,\cdot)$ follow a geometric distribution, analogous to the exponential distribution used in continuous-time models. This captures the sharp increase and exponential decay of event probabilities. Recursive algorithms for the likelihood function and gradient enable efficient estimation, as detailed in Section 4.5. Besides self- and cross-excitation from direct observation or word of mouth, alarms triggered by severe events introduce additional excitation. These alarms, determined by on-shift employees, require urgent assistance and are heard hospital-wide. Our model accounts for this by encoding events with a mark indicating whether an alarm was triggered, and whether the event occurred during daytime or nighttime hours. Because alarm triggers correlate with severity and staffing, the effects of alarms in our model should be read as predictive associations; we refrain from making direct causal claims. Our kernel is hence:

$$\Phi_{l,m}(t - t_i, m_{t_i}) = f_{l,m}(m(t_i)) g_{l,m}(m_{t_i}, t - t_i), \quad \text{for } l, m = \{1, 2, ..., 12\}, \tag{4.3}$$

where we define the function $g_{l,m}$ and $f_{l,m}$ as follows,

$$g_{l,m}(m_{t_i}, t-t_i) = \begin{cases} \beta_{cross}(1-\beta_{cross})^{t-t_i-1}, & l \neq m \text{ and } t_i \notin A^{(l)} \\ \beta_{cross}^{alarm}(1-\beta_{cross}^{alarm})^{t-t_i-1}, & l \neq m \text{ and } t_i \in A^{(l)} \\ \beta_{self}(1-\beta_{self})^{t-t_i-1}, & l = m \text{ and } t_i \notin A^{(l)} \\ \beta_{self}^{alarm}(1-\beta_{self}^{alarm})^{t-t_i-1}, & l = m \text{ and } t_i \in A^{(l)} \end{cases} \tag{4.4}$$

with $0 < \beta_\bullet \leq 1$ so that each piece is a geometric probability mass function on positive lags, which determines the shape of the decay function over time from ward l to ward m and $f_{l,m}$ controls the extent to which events cluster together across wards l to m and is a function of the mark. We define $A^{(l)} = \{\{t_i\}$: an alarm was sounded in ward l during the interval.$(t_{i-1}, t_i]\}$. We allow 4 different parameters for the excitation kernels to be fitted, namely β_{cross} for cross-excitations, β_{cross}^{alarm} for cross-excitations when an alarm was sounded, β_{self} for self-excitations, and β_{self}^{alarm} for self-excitations when an alarm was sounded. Finally we set:

$$f_{l,m}(m_{t_i}) = \begin{cases} K_{l,m}, & t_i \text{ during daytime hours and } t_i \notin A^{(l)} \\ K_n \cdot K_{l,m}, & t_i \text{ during nighttime hours and } t_i \notin A^{(l)} \\ K_{l,m} + \alpha_{l,m}, & \text{during daytime hours and } t_i \in A^{(l)} \\ K_n \cdot K_{l,m} + \alpha_n \cdot \alpha_{l,m}, & \text{during nighttime hours and } t_i \in A^{(l)}. \end{cases} \tag{4.5}$$

In this setup, $K_{l,m}$ controls the extent to which events cluster together when an event without an alarm occurs during daytime hours. During nighttime hours, $K_{l,m}$ is scaled by parameter $K_n > 0$. Similarly, $K_{l,m} + \alpha_{l,m}$ controls the extent to which events cluster together when an event with an alarm occurs during daytime hours. Finally, $\alpha_n > 0$ scales $\alpha_{l,m}$ during the nighttime hours. The reasoning here is that we expect that excitation levels will be different when the patients are locked in their dorms during nighttime hours compared to daytime. We can thus interpret $\alpha_{l,m}$ as the extra excitation attributed to alarm-marked event during the daytime and $K_n \cdot \alpha_{l,m}$ during nighttime operations.

4.4.2 Baseline component

From our preliminary analysis in Section 4.2, we found that there is evidence of a diurnal cycle in the data, as well as day-of-week effects and a seasonal component. To include this component in our model, we allow a part of the baseline rate to vary deterministically over time.

$$\mu^{(m)}(t) = \eta^{(m)}\Big(y(t)\Big) \cdot \gamma\Big(h(t), d(t), s(t)\Big), \tag{4.6}$$

where $y : \{1, \ldots, N\} \to \{1, ..., 7\}$, where N is the total number of time intervals in our dataset, gives the year index from the beginning date of our dataset that the time point t falls in. Note that there is a total of seven years in our analysis window. $h : \{1, \ldots, N\} \to \{1, \ldots, 24\}$ gives the hour of the day that time point t falls in, $d : \{1, \ldots, N\} \to \{\text{Monday}, \ldots, \text{Sunday}\}$ gives

the day of the week that t falls on, and $s : \{1,\dots,N\} \to \{\text{Spring, Summer, Fall, Winter}\}$. We compute $\gamma(h,d,s)$ from the training data as the product of three marginal profiles (hour-of-day, day-of-week, and season), each scaled to have mean 1; we then renormalize the product so that $\frac{1}{T}\sum_{t=1}^{T}\gamma(h_t,d_t,s_t) = 1$. During Hawkes fitting, γ is held fixed, so excitation parameters are interpreted conditional on this baseline.

We define the function $\eta^{(m)} : \{1,\dots,N\} \to \mathbb{R}^+$ for $m = 1,\dots,M$, as follows.

$$\eta^{(m)}(n) = \begin{cases} \eta_1^{(m)}, & n = 1 \\ \vdots & \\ \eta_7^{(m)}, & n = 7 \end{cases} \tag{4.7}$$

where $\eta_1^{(m)},\dots,\eta_7^{(m)} > 0$ for $m = 1,..,M$. This allows us to incorporate a time-varying baseline rate which is ward specific. This allows us to consider changing patient populations between wards over each year. We estimate these parameters during the maximum likelihood estimation procedure.

4.4.3 Regularization

Denote the collection of parameters to be fitted in our model during the maximum likelihood estimation procedure by θ. Note that we have a total of 288 matrix excitation parameters to train in our 12-dimensional model for the matrices K and α. To avoid overfitting, it is important to regularize these parameters. To do so, we use a lasso penalty term in the likelihood function. Namely, we define our regularized log-likelihood function $\log L_R(\cdot)$ by

$$\log L_R(\theta) = \log L(\theta) - P(K,\alpha), \tag{4.8}$$

where $\log L(\cdot)$ gives the non-regularized logarithmic likelihood function of our model and

$$P(K,\alpha) \;=\; \lambda_K \sum_{l\neq m} \big|K_{l,m}\big| \;+\; \lambda_\alpha \sum_{l,m=1}^{M} \big|\alpha_{l,m}\big|.$$

That is, the penalization term is given by the sum of the off-diagonal terms in the K-matrix and the sum of the entries of the matrix α. The motivation behind this choice of penalization is to embed our prior belief that the act of witnessing a violent event first-hand should have a greater effect than learning about the event through word of mouth. This regularization method encourages sparsity. We do not enforce a particular structure on α; every entry is penalized equally. Let $\mathbf{T}_{\text{val}}$ denote the set of time indices assigned to the validation period. We select the hyperparameters λ_K and λ_α by cross-validation using this separate validation set $\mathbf{T}_{\text{val}}$. Model assumptions cannot be verified directly; we therefore include simulation-based checks and other diagnostics in Appendix B.1.

4.5 Estimation

Under the multivariate discrete marked Hawkes process, and given $\mathbf{H}_{t-1}$, the count $Y_t^{(m)}$ is Poisson distributed, as described in Section 4.3.2. That is,

$$P\left(Y_t^{(m)} = y \,\middle|\, \mathbf{H}_{t-1}\right) = \frac{\left(\lambda^{(m)}(t)\right)^y e^{-\lambda^{(m)}(t)}}{y!}. \tag{4.9}$$

In addition to the event counts, the model includes binary alarm information. Conditionally on the count $Y_t^{(m)}$, we assume that the number of alarms of type m at time t satisfies

$$A_t^{(m)} \mid Y_t^{(m)} \sim \text{Binomial}\big(Y_t^{(m)}, p_m\big), \tag{4.10}$$

where

$$p_m = \Pr(\text{an event of type } m \text{ triggers an alarm}).$$

Thus p_m is an additional model parameter governing the probability that a type-m event generates an alarm.

Suppose there are N time intervals in total. Let

$$\tau = \{t_1, \ldots, t_{N_{\text{events}}}\}$$

denote the set of times such that at least one event occurred in the interval $(t_i - 1, t_i]$, and let N_{events} be the total number of such intervals. Then τ is the set of arrival times over the entire process.

Under the conditional Poisson model for the counts and the conditional Binomial model for the alarms, the full log-likelihood can be written as

$$\log L_{\text{full}}(\theta, \mathbf{p} \mid \tau, \mathbf{A}) = \log L(\theta \mid \tau) + \log L_{\text{alarm}}(\mathbf{p} \mid \mathbf{A}, \mathbf{Y}), \tag{4.11}$$

where θ is the collection of parameters appearing in the intensity function, $\mathbf{p} = (p_1, \ldots, p_M)$, and

$$\log L(\theta \mid \tau) = \sum_{t=1}^{N} \sum_{m=1}^{M} \left(Y_t^{(m)} \log \lambda^{(m)}(t) - \lambda^{(m)}(t)\right) \tag{4.12}$$

is the Poisson log-likelihood contribution from the event counts. The alarm contribution is

$$\log L_{\text{alarm}}(\mathbf{p} \mid \mathbf{A}, \mathbf{Y}) = \sum_{t=1}^{N} \sum_{m=1}^{M} \left[A_t^{(m)} \log p_m + \left(Y_t^{(m)} - A_t^{(m)}\right) \log(1 - p_m)\right] + \text{const.} \tag{4.13}$$

Here the combinatorial term

$$\log \binom{Y_t^{(m)}}{A_t^{(m)}}$$

does not depend on the parameters and is therefore absorbed into the constant.

This decomposition makes clear how p_m is estimated. The parameters $\mathbf{p}$ appear only in the alarm part of the likelihood, whereas the Hawkes parameters θ appear only in $\log L(\theta \mid \tau)$. Hence estimation corresponds to splitting the full log-likelihood into these two components and maximizing them separately. In particular, for each $m = 1, \ldots, M$, the maximum-likelihood estimator of p_m is

$$\widehat{p}_m = \frac{\sum_{t=1}^{N} A_t^{(m)}}{\sum_{t=1}^{N} Y_t^{(m)}}, \tag{4.14}$$

provided $\sum_{t=1}^{N} Y_t^{(m)} > 0$. Thus $\widehat{p}_m$ is the observed proportion of type-m events that triggered an alarm over the training period.

The remaining parameters θ are estimated by maximizing $\log L(\theta \mid \tau)$. The problem with directly using the form in (4.12) is that, if the data are observed over a large number of time steps, with N large, then the sum becomes expensive to evaluate. Naively computing the excitation term over all timestamps has complexity $O(N^2)$, and is therefore not feasible for long time series. In the next section we rewrite $\log L(\theta \mid \tau)$ so that it can be evaluated using only event times.

Note however that the intensity function we use in this application, as defined in Section 4.4, takes the following form:

$$\lambda^{(m)}(t) = \mu^{(m)}(t) + \sum_{l=1}^{M} \sum_{i:t_i \leq t-1} Y_{t_i}^{(l)} f_{l,m}(m_{t_i}) g_{l,m}(m_{t_i}, t - t_i), \tag{4.15}$$

where $g_{l,m}(m_{t_i}, t - t_i)$ is a geometric distribution for all $l, m = 1 \ldots, M$. Hence, we are able to reduce the computational complexity of the log-likelihood function in Equation (4.12) to order $O(N_{events}^2)$ which substantially reduces computational overhead. The proof of this proposition is analogous to the proof of Proposition 3.3.1 in the previous chapter.

Proposition 4.5.1. *For the intensity function as defined by Equation* (4.15)*, we can rewrite the log-likelihood function in Equation* (4.12) *as follows*

$$\log L(\theta|\tau) = \sum_{m=1}^{M} \sum_{t \in \tau} Y_t^{(m)} \log(\lambda^{(m)}(t)) - \sum_{m=1}^{M} \sum_{l=1}^{M} \sum_{t \in \tau} Y_t^{(l)} f_{l,m}(m_t) G_{l,m}(m_t, N - t) - \sum_{m=1}^{M} \sum_{t=1}^{N} \mu^{(m)}(t) \tag{4.16}$$

where $G_{l,m}(m_t, \cdot)$ is the cumulative mass function of the probability mass function $g_{l,m}(m_t, \cdot)$.

Proof. See the appendix in the supplementary material. □

We have achieved a reduction in the complexity of the calculation of the intensity function by the fact that $N_{events} \leq N$. However, still, when the number of events is high, the calculation of the repeated intensity function will become expensive. In the next section, we develop a faster approach that takes only linear-time complexity to calculate the intensity function. This is achieved by developing a recursive formulation for the intensity function by maintaining two additional data structures that take a constant space per event. This is a marked version of Poposition 3.4.1.

Proposition 4.5.2. *For $j \in \{1,\dots,N_{events}\}$ and $m \in \{1,\dots,M\}$, the conditional intensity function from Section 4.4 can be expressed recursively as:*

$$\begin{aligned}\lambda^{(m)}(t_j) = \mu^{(m)}(t_j) &+ \sum_{l\neq m} K_{l,m}\big(R^{\beta_{cross}}_{day}(j,l) + K_n R^{\beta_{cross}}_{night}(j,l)\big) \\ &+ K_{m,m}\big(R^{\beta_{self}}_{day}(j,l) + K_n R^{\beta_{self}}_{night}(j,l)\big) \\ &+ \sum_{l\neq m} \alpha_{l,m}\big(AR^{\beta^{alarm}_{cross}}_{day}(j,l) + \alpha_n AR^{\beta^{alarm}_{cross}}_{night}(j,l)\big) \\ &+ \alpha_{m,m}\big(AR^{\beta^{alarm}_{self}}_{day}(j,l) + \alpha_n AR^{\beta^{alarm}_{self}}_{night}(j,l)\big),\end{aligned} \tag{4.17}$$

where for $l \in \{1,\dots,M\}$, $j = 1,\dots,N_{events}-1$, $x \in \{day, night\}$, and $\beta \in \{\beta_{cross}, \beta_{self}\}$:

$$R^{\beta}_{x}(j+1,l) := (1-\beta)^{t_{j+1}-t_j} R^{\beta}_{x} + \mathbf{1}_{\{t_j \text{ during } x\}} Y^{(l)}_{t_j} \beta (1-\beta)^{t_{j+1}-t_j-1}, \tag{4.18}$$

and for $\beta^{alarm} \in \{\beta^{alarm}_{cross}, \beta^{alarm}_{self}\}$:

$$AR^{\beta^{alarm}}_{x}(j+1,l) := (1-\beta^{alarm})^{t_{j+1}-t_j} AR^{\beta^{alarm}}_{x} + \mathbf{1}_{\{t_j \text{ during } x\}} A^{(l)}_{t_j} \beta^{alarm} (1-\beta^{alarm})^{t_{j+1}-t_j-1}. \tag{4.19}$$

Initialize $R^{\beta}_{x}(1,l) := AR^{\beta^{alarm}}_{x}(1,l) := 0$ for $l = 1,\dots,M$.

Proof. See Appendix in the supplementary material. □

Note that closed-form score expressions and their $O(1)$ implementations are presented in Appendix B.

Proposition 4.5.1 and Proposition 4.5.2 serve different purposes. Proposition 4.5.1 rewrites the log-likelihood so that the compensator term is expressed as a sum over event times rather than over the full grid $t = 1,\dots,N$, thereby reducing the cost of likelihood evaluation when events are sparse. Proposition 4.5.2, by contrast, concerns the efficient evaluation of the conditional intensities $\lambda^{(m)}(t_j)$ appearing in the event-time likelihood. It introduces recursive state variables that allow these intensities to be updated in constant time from one event time to the next. Thus, the first proposition reduces the likelihood to an event-time form, whereas

the second makes the event-time intensity calculation itself computationally efficient. With the event-time and $O(1)$ recursions, likelihood and gradient evaluation scale as $O(M^2 N_{\text{events}})$ (Appendix B). A sufficient stability condition is $\rho(K) < 1$ because the geometric kernels integrate to 1 in discrete time.

4.6 Simulation and Forecasting for the 12D-HPA Model

This section presents efficient simulation and forecasting algorithms tailored to the marked discrete Hawkes process introduced in this chapter. The constructions build directly on the recursion established in Lemma 3.6.1.

In a frequentist forecasting setup, prediction under a Hawkes-type model is obtained by conditioning on the observed history up to time T_0, computing the conditional intensity thereafter, and simulating many future paths to produce a predictive distribution for event counts in a target window. In continuous time this is commonly done by Ogata's thinning method (Ogata, 1988). In discrete time, while the derivations are simpler, a naive implementation is computationally heavy because each step recomputes a sum over *all* past lags; standard discrete Hawkes simulators (Cinlar, 2013) therefore suffer an $O(t)$ update per time index t.

Our 12D-HPA (twelve-dimensional Hawkes-with-Periodic/Alarm covariates) model admits a convenient *recursive* representation that reduces each update to $O(1)$ per component and covariate. The key idea is to maintain exponentially weighted running sums-one for day, one for night, and their alarm-modulated analogues-so that the intensity at time t depends only on these states rather than on the full event history.

In addition to event counts, the model uses binary/indicator "alarm" variables. Conditionally on $Y_t^{(m)}$ (the number of type-m events at time t), we model the number of alarms as $A_t^{(m)} \mid Y_t^{(m)} \sim \text{Binomial}\big(Y_t^{(m)}, p_m\big)$, where $p_m = \Pr(\text{an event of type } m \text{ triggers an alarm})$.

Let $x \in \{\text{day}, \text{night}\}$ indicate whether t falls in the day or night window, and let $\mathbf{1}_{\{t \in x\}}$ be the corresponding indicator. Multiplicative night-time factors K_n (for events) and α_n (for alarms) upweight the excitation during night hours when appropriate.

Proposition 4.6.1 (Fast recursion for the discrete 12D-HPA intensity)**.** *For components* $m \in \{1,\dots,M\}$ *and times* $t = 1,\dots,N$*, the conditional intensity in Equation* (4.15) *can be written as*

$$\begin{aligned}\lambda^{(m)}(t) = \mu^{(m)}(t) &+ \sum_{l\neq m} K_{l,m} R^{\beta_{\text{cross}}}_{day}(t,l) + K_n \sum_{l\neq m} K_{l,m} R^{\beta_{\text{cross}}}_{night}(t,l) \\ &+ K_{m,m} R^{\beta_{\text{self}}}_{day}(t,m) + K_n K_{m,m} R^{\beta_{\text{self}}}_{night}(t,m) \\ &+ \sum_{l\neq m} \alpha_{l,m} AR^{\beta^{\text{alarm}}_{\text{cross}}}_{day}(t,l) + \alpha_n \sum_{l\neq m} \alpha_{l,m} AR^{\beta^{\text{alarm}}_{\text{cross}}}_{night}(t,l) \\ &+ \alpha_{m,m} AR^{\beta^{\text{alarm}}_{\text{self}}}_{day}(t,m) + \alpha_n \alpha_{m,m} AR^{\beta^{\text{alarm}}_{\text{self}}}_{night}(t,m), \end{aligned} \tag{4.20}$$

where, for $l \in \{1,\dots,M\}$*, the state variables satisfy the first-order recursions*

$$R^{\beta}_x(t+1,l) = (1-\beta) R^{\beta}_x(t,l) + \mathbf{1}_{\{t\in x\}}\, \beta(1-\beta) Y^{(l)}_t, \qquad R^{\beta}_x(1,l) = 0, \tag{4.21}$$

$$AR^{\beta^{\text{alarm}}}_x(t+1,l) = (1-\beta^{\text{alarm}}) AR^{\beta^{\text{alarm}}}_x(t,l) + \mathbf{1}_{\{t\in x\}}\, \beta^{\text{alarm}}(1-\beta^{\text{alarm}}) A^{(l)}_t, \quad AR^{\beta^{\text{alarm}}}_x(1,l) = 0, \tag{4.22}$$

with $x \in \{day, night\}$*,* $\beta \in \{\beta_{\text{cross}}, \beta_{\text{self}}\}$*, and* $\beta^{\text{alarm}} \in \{\beta^{\text{alarm}}_{\text{cross}}, \beta^{\text{alarm}}_{\text{self}}\}$*.*

Proof. For any $x \in \{\text{day}, \text{night}\}$ and $t \geq 1$,

$$\sum_{s\leq t} \mathbf{1}_{\{s\in x\}} Y^{(l)}_s \beta(1-\beta)^{t-s} = (1-\beta) \sum_{s\leq t-1} \mathbf{1}_{\{s\in x\}} Y^{(l)}_s \beta(1-\beta)^{t-1-s} + \mathbf{1}_{\{t\in x\}} \beta(1-\beta) Y^{(l)}_t.$$

Defining $R^{\beta}_x(t,l)$ by the recursion in (4.21) makes the identity exact; the same argument with $A^{(l)}_t$ and β^{alarm} yields (4.22). Substituting these into the original (discrete-time) intensity expansion gives (4.20). □

The naive update at time t sums over $t-1$ lags. With (4.21)-(4.22), each R or AR update costs $O(1)$. Per time step we update $2M$ states for events (day/night, self/cross through the coefficients) and $2M$ for alarms, giving overall $O(M)$ state updates and $O(M^2)$ intensity assembly due to the $\sum_{l\neq m}$ cross-terms-independent of the forecasting horizon length.

We provide two simulation modes because the goals and initial conditions differ:

- *Unconditional simulation* (Alg. 3) generates synthetic trajectories from the model *ab initio*. This is useful for sanity checks, parametric bootstrap, stress testing, and studying qualitative behaviours of the fitted parameters (e.g., stability under different day/night schedules). The state variables R and AR start at zero and are driven entirely by simulated events and alarms.

- *Conditional forecasting* (Alg. 4) produces future paths *given* an observed history up to time T_0. This is the operational setting for prediction: the past is fixed, and uncertainty concerns only future events. The key difference is the *initialisation*: we fold the observed history into the recursion (4.21)–(4.22) to obtain the states at T_0+1, and then simulate forward for H steps. Repeating this B times yields a Monte Carlo approximation to the predictive distribution of any forecast target (e.g., total events by type in $[T_0+1, T_0+H]$).

Both algorithms share the same update mechanics: at each time step we (i) assemble $\lambda^{(m)}(t)$ via the fast representation (4.20), (ii) sample event counts $Y_t^{(m)}$ (Poisson or Bernoulli-thinned, as required), (iii) sample alarm counts $A_t^{(m)}$ (Binomial) unless alarms are exogenous and provided, and (iv) update the state variables with (4.21)–(4.22). Thanks to these recursions, the per-step cost is $O(M)$ to update states and $O(M^2)$ to assemble cross-component intensities—independent of how far ahead we forecast.

In practice:

1. Use Alg. 3 for prior/parameter checks, synthetic experiments, and to visualise qualitative effects of $(K_{l,m}, \alpha_{l,m}, \beta\text{s}, K_n, \alpha_n)$.
2. Use Alg. 4 for plug-in forecasting (frequentist) by fixing parameters at their estimates and drawing B conditional paths; if alarm indicators are observed in the forecast window, plug them in directly (skip the Binomial step).

Algorithm 3 Simulate-12D-HPA (unconditional discrete-time simulation)

Require: Horizon N; components $M{=}12$; day/night map $x(t) \in \{\text{day}, \text{night}\}$; baselines $\mu^{(m)}(t)$; gains $K_{l,m}, K_n$ and $\alpha_{l,m}, \alpha_n$; decays $\beta_{\text{self}}, \beta_{\text{cross}}, \beta_{\text{self}}^{\text{alarm}}, \beta_{\text{cross}}^{\text{alarm}}$; alarm probs p_m.

Ensure: Arrays $\{\lambda^{(m)}(t)\}$, $\{Y_t^{(m)}\}$, $\{A_t^{(m)}\}$ for $t{=}1{:}N$, $m{=}1{:}M$.

Initialise $R_x^{\beta}(1,l) \leftarrow 0$ and $AR_x^{\beta^{\text{alarm}}}(1,l) \leftarrow 0$ for all l, $x \in \{\text{day}, \text{night}\}$, $\beta \in \{\beta_{\text{self}}, \beta_{\text{cross}}\}$, $\beta^{\text{alarm}} \in \{\beta_{\text{self}}^{\text{alarm}}, \beta_{\text{cross}}^{\text{alarm}}\}$.
for $t = 1$ **to** N **do**
  for $m = 1$ **to** M **do**
    Compute $\lambda^{(m)}(t)$ using the recursive form (4.20).
  end for
  for $m = 1$ **to** M **do**
    Draw $Y_t^{(m)} \sim \text{Poisson}\big(\lambda^{(m)}(t)\big)$.
    Draw $A_t^{(m)} \sim \text{Binomial}\big(Y_t^{(m)}, p_m\big)$.
  end for
  for $l = 1$ **to** M **do**
    for $x \in \{\text{day}, \text{night}\}$ **do**
      $R_x^{\beta_{\text{self}}}(t{+}1,l) \leftarrow (1{-}\beta_{\text{self}}) R_x^{\beta_{\text{self}}}(t,l) + \mathbf{1}_{\{t \in x\}} \beta_{\text{self}}(1{-}\beta_{\text{self}}) Y_t^{(l)}$.
      $R_x^{\beta_{\text{cross}}}(t{+}1,l) \leftarrow (1{-}\beta_{\text{cross}}) R_x^{\beta_{\text{cross}}}(t,l) + \mathbf{1}_{\{t \in x\}} \beta_{\text{cross}}(1{-}\beta_{\text{cross}}) Y_t^{(l)}$.
      $AR_x^{\beta_{\text{self}}^{\text{alarm}}}(t{+}1,l) \leftarrow (1{-}\beta_{\text{self}}^{\text{alarm}}) AR_x^{\beta_{\text{self}}^{\text{alarm}}}(t,l) + \mathbf{1}_{\{t \in x\}} \beta_{\text{self}}^{\text{alarm}}(1{-}\beta_{\text{self}}^{\text{alarm}}) A_t^{(l)}$.
      $AR_x^{\beta_{\text{cross}}^{\text{alarm}}}(t{+}1,l) \leftarrow (1{-}\beta_{\text{cross}}^{\text{alarm}}) AR_x^{\beta_{\text{cross}}^{\text{alarm}}}(t,l) + \mathbf{1}_{\{t \in x\}} \beta_{\text{cross}}^{\text{alarm}}(1{-}\beta_{\text{cross}}^{\text{alarm}}) A_t^{(l)}$.
    end for
  end for
end for

4.7 Results

We fit five model variations to the twelve wards of the forensic psychiatric hospital between 26 June 2012 and 1 January 2020. We denote time interval over which we train our models as $\mathbf{T}_{train}$. A validation window of approximately 4.5 months ($\mathbf{T}_{val}$) is used for time-blocked cross-validation of regularization hyperparameters. A holdout test set consisting of the final 4.5 months ($\mathbf{T}_{test}$) is used to evaluate out-of-sample predictive performance. Unless stated otherwise, model selection is performed on $\mathbf{T}_{train} \cup \mathbf{T}_{val}$ and final estimates are refit on this union before evaluation on $\mathbf{T}_{test}$.

We first define the five models which will be compared on our dataset. Our first and simplest model is the inhomogeneous Poisson process. We then consider twelve individual univariate Hawkes processes (one per ward), followed by a multivariate Hawkes process, and finally a multivariate Hawkes process with alarm marks as defined in Section 4.4. For completeness, we also include a univariate-with-alarm variant for each ward. We emphasize that comparisons are based on held-out predictive log-likelihood; the results are read in a predictive/descriptive sense rather than as explicit causal claims. More formally, we define our models as follows.

Algorithm 4 Forecast-12D-HPA (conditional Monte Carlo)

Require: History $\{Y_s^{(l)}, A_s^{(l)} : s \le T_0,\ l{=}1{:}M\}$; horizon H; paths B; model objects as in Alg. 3. (Optionally: future alarms $\{A_t^{(l)} : T_0{<}t \le T_0{+}H\}$.)

Ensure: Simulated paths $\{Y_t^{(m,b)} : t{=}T_0{+}1{:}T_0{+}H,\ m{=}1{:}M,\ b{=}1{:}B\}$ and corresponding intensities.

Initialise $R_x^{\beta}(1,\cdot){=}AR_x^{\beta^{\text{alarm}}}(1,\cdot){=}0$.
for $s = 1$ **to** T_0 **do** ▷ build states from the observed history
    Update $R_x^{\beta}(s{+}1,\cdot)$ and $AR_x^{\beta^{\text{alarm}}}(s{+}1,\cdot)$ using (4.21)-(4.22) with observed (Y_s, A_s).
end for
for $b = 1$ **to** B **do**
    Copy states at $T_0{+}1$ into working buffers.
    for $t = T_0{+}1$ **to** $T_0{+}H$ **do**
        for $m = 1$ **to** M **do**
            Compute $\lambda^{(m)}(t)$ via (4.20).
        end for
        for $m = 1$ **to** M **do**
            Draw $Y_t^{(m,b)} \sim \text{Poisson}\big(\lambda^{(m)}(t)\big)$.
            if future alarms are exogenous **then**
                $A_t^{(m,b)} \leftarrow A_t^{(m)}$
            else
                Draw $A_t^{(m,b)} \sim \text{Binomial}\big(Y_t^{(m,b)}, p_m\big)$
            end if
        end for
        Update the working R, AR states using (4.21)-(4.22) with $(Y_t^{(\cdot,b)}, A_t^{(\cdot,b)})$.
    end for
end for

- **Inhomogeneous Poisson Process (IPP).** Consider an inhomogeneous Poisson process where the number of events in the interval t in the m^{th} ward is Poisson distributed with rate

$$\lambda^{(m)}(t) = \eta^{(m)}\Big(y(t)\Big) \cdot \gamma\Big(h(t), d(t), s(t)\Big),$$

  where the parameters of this intensity are defined as in Section 4.4.2. No excitation terms are included.

- **Univariate Discrete Hawkes Processes (1D-DHP).** Consider a twelve-dimensional discrete Hawkes process as defined in Section 4.3.2, where we model each ward by a dimension in the process. That is,

$$\lambda^{(m)}(t) = \eta^{(m)}\Big(y(t)\Big) \cdot \gamma\Big(h(t), d(t), s(t)\Big) + \sum_{i:t_i \le t-1} Y_{t_i}^{(m)} \Phi_m(t - t_i, m_{t_i}),$$

where for $m = 1, \ldots, M$,

$$\Phi_m((t - t_i, m_{t_i}) = f_m(m_{t_i}) g(m_{t_i}, t - t_i),$$

and we define the functions g and f_m as follows:

$$g(m_{t_i}, t - t_i) = \beta_{self}(1 - \beta_{self})^{t - t_i - 1},$$

$$f_m(m_{t_i}) = \begin{cases} K_m, & t_i \text{ during daytime hours,} \\ K_n \cdot K_m, & t_i \text{ during nighttime hours.} \end{cases}$$

The parameters of the baseline intensity are defined as in Section 4.4.2.

- **Multivariate Discrete Hawkes Process (12D-DHP).** Consider a twelve-dimensional discrete Hawkes process as in Section 4.3.2, where we model each ward by a dimension in the process. That is,

$$\lambda^{(m)}(t) = \eta^{(m)}\Big(y(t)\Big) \cdot \gamma\Big(h(t), d(t), s(t)\Big) + \sum_{l=1}^{M} \sum_{i: t_i \leq t-1} Y_{t_i}^{(l)} \Phi_{l,m}(t - t_i, m_{t_i}),$$

where for $l, m = 1, \ldots, M$,

$$\Phi_{l,m}(t - t_i, m_{t_i}) = f_{l,m}(m_{t_i}) g_{l,m}(m_{t_i}, t - t_i),$$

and we define the functions $g_{l,m}$ and $f_{l,m}$ as follows:

$$g_{l,m}(m_{t_i}, t - t_i) = \begin{cases} \beta_{cross}(1 - \beta_{cross})^{t - t_i - 1}, & l \neq m \text{ and } t_i \notin A^{(l)}, \\ \beta_{self}(1 - \beta_{self})^{t - t_i - 1}, & l = m \text{ and } t_i \notin A^{(l)}. \end{cases}$$

$$f_{l,m}(m_{t_i}) = \begin{cases} K_{l,m}, & t_i \text{ during daytime hours,} \\ K_n \cdot K_{l,m}, & t_i \text{ during nighttime hours.} \end{cases}$$

The parameters of the baseline intensity are defined as in Section 4.4.2. We apply an $L1$-norm penalty to off-diagonal entries of K to encourage sparsity in cross-ward excitation; the hyperparameter λ_K is selected by time-blocked cross-validation on $\mathbf{T}_{val}$ and used in refits on $\mathbf{T}_{train} \cup \mathbf{T}_{val}$.

- **Univariate Discrete Hawkes Process with Alarm Covariates (1D-DHPA).** Consider a twelve-dimensional discrete Hawkes process as defined in Section 4.3.2, where we model each ward by a dimension in the process. That is,

$$\lambda^{(m)}(t) = \eta^{(m)}\Big(y(t)\Big) \cdot \gamma\Big(h(t), d(t), s(t)\Big) + \sum_{i: t_i \leq t-1} Y_{t_i}^{(m)} \Phi_m(t - t_i, m_{t_i}),$$

where for $m = 1, \ldots, M$,

$$\Phi_m(t - t_i, m_{t_i}) = f_m(m_{t_i}) g_m(m_{t_i}, t - t_i),$$

and we define the functions g and f_m as follows:

$$g_m(m_{t_i}, t - t_i) = \begin{cases} \beta_{self}(1-\beta_{self})^{t-t_i-1}, & t_i \notin A^{(m)}, \\ \beta_{self}^{alarm}(1-\beta_{self}^{alarm})^{t-t_i-1}, & t_i \in A^{(m)}. \end{cases}$$

$$f_m(m_{t_i}) = \begin{cases} K_m, & t_i \text{ during daytime hours and } t_i \notin A^{(m)}, \\ K_n \cdot K_m, & t_i \text{ during nighttime hours and } t_i \notin A^{(m)}, \\ K_m + \alpha_m, & t_i \text{ during daytime hours and } t_i \in A^{(m)}, \\ K_n \cdot K_m + \alpha_n \cdot \alpha_m, & t_i \text{ during nighttime hours and } t_i \in A^{(m)}. \end{cases}$$

The parameters of the baseline intensity are defined as in Section 4.4.2. This model captures within-ward excitation with alarm-marked differences in decay/scale.

- **Multivariate Discrete Hawkes Process with Alarm Covariates (12D-DHPA).** Consider a twelve-dimensional discrete Hawkes process as defined in Section 4.3.2, where we model each ward by a dimension in the process. That is,

$$\lambda^{(m)}(t) = \eta^{(m)}\Big(y(t)\Big) \cdot \gamma\Big(h(t), d(t), s(t)\Big) + \sum_{l=1}^{M} \sum_{i: t_i \le t-1} Y_{t_i}^{(l)} \Phi_{l,m}(t - t_i, m_{t_i}),$$

where for $l, m = 1, \ldots, M$,

$$\Phi_{l,m}(t - t_i, m_{t_i}) = f_{l,m}(m_{t_i}) g_{l,m}(m_{t_i}, t - t_i),$$

and we define the functions $g_{l,m}$ and $f_{l,m}$ as follows:

$$g_{l,m}(m_{t_i}, t - t_i) = \begin{cases} \beta_{cross}(1-\beta_{cross})^{t-t_i-1}, & l \neq m \text{ and } t_i \notin A^{(l)}, \\ \beta_{cross}^{alarm}(1-\beta_{cross}^{alarm})^{t-t_i-1}, & l \neq m \text{ and } t_i \in A^{(l)}, \\ \beta_{self}(1-\beta_{self})^{t-t_i-1}, & l = m \text{ and } t_i \notin A^{(l)}, \\ \beta_{self}^{alarm}(1-\beta_{self}^{alarm})^{t-t_i-1}, & l = m \text{ and } t_i \in A^{(l)}. \end{cases}$$

$$f_{l,m}(m_{t_i}) = \begin{cases} K_{l,m}, & t_i \text{ during daytime hours and } t_i \notin A^{(l)}, \\ K_n \cdot K_{l,m}, & t_i \text{ during nighttime hours and } t_i \notin A^{(l)}, \\ K_{l,m} + \alpha_{l,m}, & t_i \text{ during daytime hours and } t_i \in A^{(l)}, \\ K_n \cdot K_{l,m} + \alpha_n \cdot \alpha_{l,m}, & t_i \text{ during nighttime hours and } t_i \in A^{(l)}. \end{cases}$$

Ward	pLL^{IPP}	$pLL^{1D/12D-DHP}$	$pLL^{1D-DHPA}$	$pLL^{12D-DHPA}$
1	-94.1	-94.5	-94.6	**-93.2**
2	-115.5	**-110.5**	-110.6	-110.7
3	-8.3	-4.0e-5	**-4.4e-10**	-1.6e-4
4	-150.9	-148.5	-148.5	**-147.1**
5	-299.2	-283.9	-284.0	**-282.2**
6	-43.4	-43.6	-43.6	**-41.6**
7	**-201.1**	-213.2	-213.5	-204.2
8	-49.0	-44.0	-44.0	**-43.1**
9	-17.2	-13.5	-13.7	**-12.6**
10	-514.9	-375.4	-375.5	**-375.1**
11	-183.8	-173.8	-173.8	**-169.7**
12	-26.4	**-4.0e-5**	-1.9e-7	-1.6e-4
Overall	-1703.8	-1501.2	-1501.7	**-1479.4**
Difference	-224.4 (7.1)	-21.8 (3.7)	-22.3 (3.7)	**0** (0)

Table 4.3: Predictive log-likelihood values for each of the specified models on the test dataset, $\mathbf{T}_{test}$, with approximate standard deviation estimates in brackets. In the Difference row, we report the overall difference in log-likelihood between each model and the best model (12D-DHPA), with corresponding standard deviation estimates in brackets. These estimates were obtained by computing the log-likelihood values, and their differences, on 4 equal sized non-overlapping sections of$\mathbf{T}_{test}$.

The parameters of the baseline intensity are defined as in Section 4.4.2. We also subtract two L1-norm penalties for the off-diagonal elements of the K-matrix and α-matrix to the log-likelihood function to encourage sparsity in the cross-excitations of the event excitations and sparsity over the entire hospital for the alarm excitations. The hyperparameters $\lambda_K = 1$ and $\lambda_\alpha = 0.4$ were selected by time-blocked cross-validation on $\mathbf{T}_{val}$ and then used in refits on $\mathbf{T}_{train} \cup \mathbf{T}_{val}$.

4.7.1 Predictive Model Comparison

We evaluated predictive performance using the predictive log-likelihood metric (Mukhopadhyay & Sathish, 2018), applied to the events in the test data, given the previously observed event counts. Specifically, we calculated the probability of observing the events of the test set $\mathbf{Y}_{test}$, conditioned on all prior observations, including those of the training and validation sets. The predictive log-likelihood (pLL) is then obtained by summing the logarithms of these conditional probabilities.

Model	Baseline	Event excitation (self/cross)	Alarm excitation (self/cross)
IPP	100%	0%	0%
1D-DHP	69%	31% (100% / 0%)	0%
1D-HPA	68%	31% (100% / 0%)	1% (100% / 0%)
12D-DHPA	50%	25% (100% / 0%)	25% (33% / 67%)

Table 4.4: Branching responsibilities (per event). Each entry gives the fraction of an event's intensity attributable to component c, computed as $\lambda_c^{(m)}(t)/\lambda^{(m)}(t)$; rows sum to 1 (baseline and excitation components included). Larger values indicate greater expected parentage from component c.

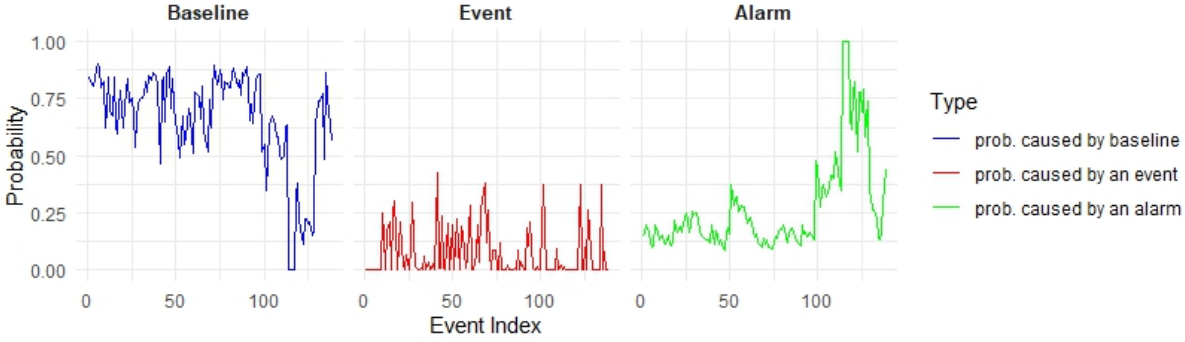


Figure 4.6: Different sources of excitation in Ward 2 over the entire time period. Left: probabilities that events are attributed to the baseline. Middle: probabilities attributed to non-alarm excitation. Right: probabilities attributed to alarm-marked excitation. Attributions are computed under the 12D-DHPA model's branching representation.

We calculate the predictive log-likelihood for each of our models defined at the beginning of this section, using the notation pLL superscripted by the model abbreviations. We report these for each ward individually and for the entire hospital. The best, i.e., the highest predictive log-likelihood, model for each ward is shown in bold. We use this as a metric for each model's predictive performance. Ward indices correspond to the layout in Figure 4.1. All parameters are constrained to be nonnegative during optimization, ensuring nonnegative intensities.

The 1D-DHP and 12D-DHP models perform similarly because cross-validation selected a 12D-DHP with negligible off-diagonal elements in K; effectively, it functions like independent univariate models. Accordingly, we interpret both as univariate Hawkes specifications without alarm marks.

The IPP baseline underperforms markedly, indicating that excitation terms materially improve predictive fit. The 1D-DHP and 1D-DHPA models are very close, suggesting that adding an alarm mark without cross-ward excitation provides little incremental predictive benefit. In contrast, the 12D-DHPA achieves the best overall pLL, indicating that cross-ward excitation associated with alarm-marked events materially improves out-of-sample prediction.

Table 4.4 shows the responsibility decomposition, highlighting how much each event process contributes to the overall intensity. This is computed using the branching representation: for an event at (t, m), the contribution of each source (baseline, self/cross, alarm/non-alarm) is its intensity component divided by $\lambda^{(m)}(t)$; Table 4.4 reports event-wise averages of these weights.

Notably, only around 1% of events are attributed to alarms in the 1D-DHPA model, as shown in Table 4.4, whereas in the 12D-DHPA model about 25% are attributed to alarm-marked excitation, 67% of which are cross-ward. Additionally, Figure 4.6 shows that, in some periods, the attributed probability that an event is linked to an alarm-marked ancestor can approach 1.

It is also worth noting that the 12D-DHPA model retains the same structure for the K-matrix as the univariate models, with zeros on off-diagonal entries, indicating no cross-ward excitation from non-alarm events. In contrast, the excitation matrix α in the 12D-DHPA model is more widely distributed, reflecting the hospital-wide effect of an alarm-triggered event in generating further violent incidents in different wards.

To further assess predictive stability, we split the test set into four equally sized sections and compute pLL for each model (12D-DHPA, 1D-DHPA, 1D-DHP, and IPP) in each section. We then compute the section-wise differences between each model and 12D-DHPA. Approximate standard deviations for the sums of these differences are obtained by treating the four sections as roughly independent: we compute the standard deviation of the section-wise differences and multiply by 2 to obtain the reported standard deviation for the sum over four sections. This provides a quick variability check for the overall pLL differences across $\mathbf{T}_{test}$, and is how the standard errors in Table 4.3 were derived. Additional predictive checks (Appendix A) report a density estimate of pLL from 50 forecasts simulated over the test interval under the fitted model.

4.7.2 Interpretation of Results

We now analyze the parameter fits for the best performing 12D-DHPA model. Figure 4.7 and Figure 4.8 present heat maps representing the values of the excitation matrices in the 12D-DHPA model during daytime and nighttime operational hours, respectively. Specifically, these figures display the matrices K and α in Figure 4.7, and the matrices $K_n \cdot K$ and $\alpha_n \cdot \alpha$ in Figure 4.8. The magnitude of the entry $K_{i,j}$ in the excitation matrix K indicates the model-implied expected number of first-generation events ("direct offspring," under the branching representation) in ward j attributed to a non-alarm event in ward i. Subsequent generations arise recursively through the same mechanism. The magnitude of entry $\alpha_{i,j}$ in the excitation matrix α represents the corresponding expected number of first-generation events in ward j attributed to an alarm-triggered event in ward i.

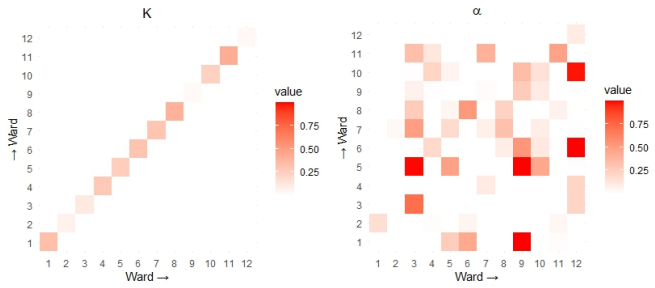


Figure 4.7: Heatmap of excitation matrices K and α under the 12D-DHPA model. These represent the excitation matrices during daytime operations.

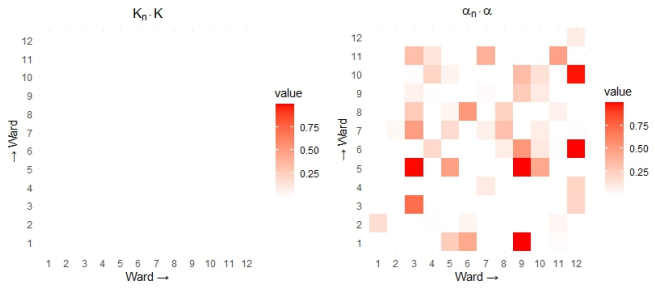


Figure 4.8: Heatmap of excitation matrices $K_n \cdot K$ and $\alpha_n \cdot \alpha$ under the 12D-DHPA model. These represent the excitation matrices during nighttime operations.

Examining the K-matrices for daytime hours, we observe that the excitation is heavily concentrated along the diagonal, indicating primarily within-ward excitation. In practical terms, non-alarm events tend to raise near-term risk in the same ward, with little evidence of spillover to other wards. Furthermore, during the estimation process, we found that $K_n = 0$ indicating that there is effectively no non-alarm excitation during nighttime hours. Thus, an event that occurs at night via the non-alarm channel has negligible impact on future events in the hospital. We note also that our fitted K satisfied $\rho(K) < 1$, showing the process is subcritical.

In contrast, the excitation matrix α reveals a more distributed pattern, with significantly less concentration along the diagonal. This spread is expected, as an alarm-triggered event is audible throughout the hospital, likely influencing event rates across various wards. There is also evidence of increased excitation within the same ward due to an alarm, which is consistent with the expectation that witnessing an event leading to an alarm would affect the likelihood of subsequent events within that ward.

Figure 4.9 presents the excitation functions for three types of events: those that triggered an alarm in different wards, those that triggered an alarm within the same ward, and non-alarm events within the same ward. The vertical dotted lines represent the expected value of the corresponding geometric distribution, calculated as $1/\beta$.

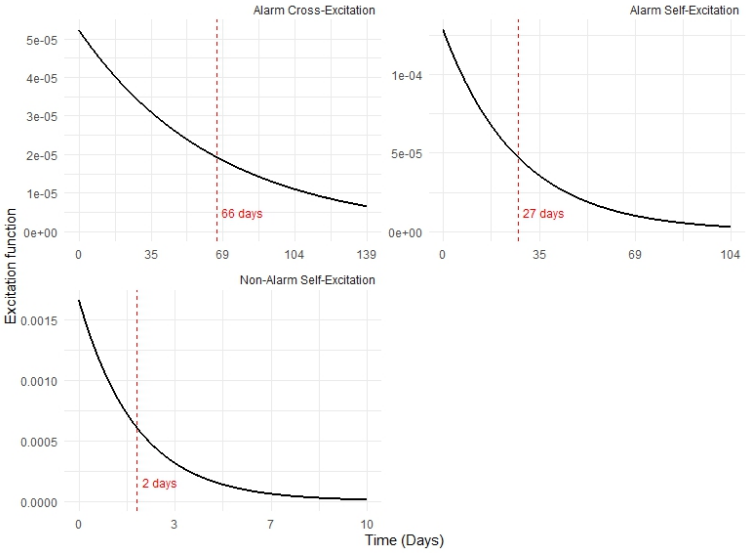


Figure 4.9: Plots of excitation functions for events that triggered an alarm across different wards, events that triggered an alarm within the same ward, and non-alarm events within the same ward. The dotted vertical lines indicate the expected value of the corresponding geometric distribution, given by $1/\beta$.

For non-alarm events occurring within the same ward, the model-implied memory of the event is approximately 2 days, meaning that the impact on future event rates is expected to persist for around 2 days following the initial event. In contrast, the excitation functions for alarm-triggered events exhibit significantly longer memory effects: about 27 days for alarm-triggered events within the same ward and approximately 66 days for those that triggered an alarm in a different ward.

This suggests that patients and staff retain a longer-lasting memory of events when an alarm is sounded, particularly if the alarm originates from a different ward. We view this interpretation as descriptive and model-based. The prolonged impact may stem from the uncertainty of hearing an alarm without context, potentially heightening anxiety or alertness.

We also present an example of the intensity function over a short time frame for wards 1 and 5, as shown in Figure 4.11, focusing on daytime operation hours. The plot highlights three events: one in ward 5 without an alarm (red), one in ward 9 with an alarm (blue), and one in ward 6 with an alarm (green).

The sharp intensity increase in Figure 4.11a reflects the impact of a self-exciting event in ward 5. This rise, followed by a rapid decline, corresponds to the approximately two-day memory effect for same-ward events with alarms, as depicted in Figure 4.9. This illustrates how recent events influence intensity, with local effects dissipating more quickly.

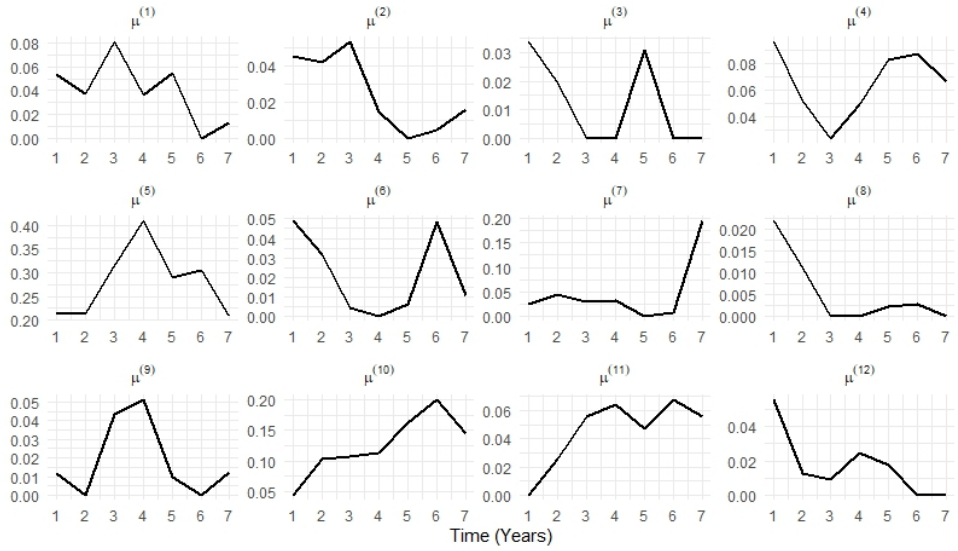
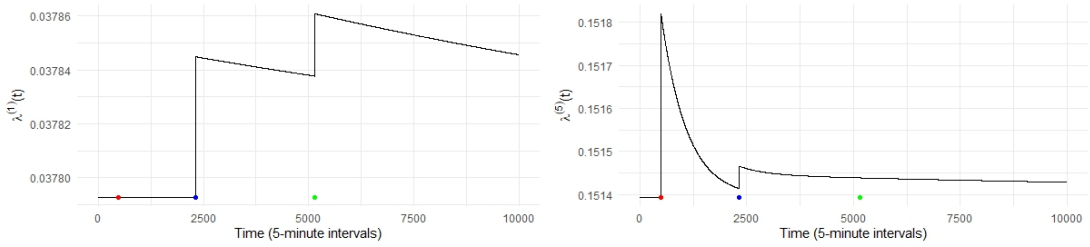


Figure 4.10: Plots of the baseline proportionality constants $\mu^{(m)}$ for each ward $m = 1, \ldots, M$ over the seven-year period of the dataset.



(a) Intensity function for ward 1 **(b)** Intensity function for ward 5

Figure 4.11: Intensity functions for wards 1 and 5 over a short daytime period. Colored dots indicate events: red (ward 5, no alarm), blue (ward 9, alarm), and green (ward 6, alarm).

The event from ward 9, depicted in blue, influences the intensity functions of wards 1 and 5, as shown in the plots. This cross-excitation exhibits a memory effect lasting approximately 66 days, with intensity functions gradually decaying over time.

The event from ward 6, shown in green, affects the intensity function in ward 1 but not ward 5, aligning with the cross-excitation patterns in the heat map of the matrix α in Figure 4.7. Its excitation function resembles that of ward 9's event, differing mainly in scaling due to excitation parameters.

4.7.3 Forecasting

In Table 4.5, we present predictions for the total number of events over the final 4.5 months of our data set, denoted $\mathbf{T}_{test}$. These predictions were generated using the prediction algorithm described in Appendix 4.6, where the probability of an event triggering an alarm was treated as a Bernoulli parameter in the model. This estimate was derived by maximum likelihood on $\mathbf{T}_{train}$. We include all previous historical events when forecasting; i.e., the conditional intensity at each time t incorporates all prior events from the training and validation windows. The forecasts integrate uncertainty by simulating both event counts and the alarm mark via a Bernoulli model estimated on $\mathbf{T}_{train}$. Monte Carlo standard errors come from repeated simulations under the fitted model.

Ward	True	Forecast (σ)	Ward	True	Forecast (σ)
1	10	6 (3)	7	23	58 (9)
2	12	4 (2)	8	4	1 (1)
3	0	0 (0)	9	1	4 (2)
4	17	21 (5)	10	49	44 (8)
5	34	67 (10)	11	20	23 (6)
6	4	5 (3)	12	0	0 (0)

Table 4.5: Comparing the mean predicted number of events over the test interval for each ward with the true number of events. Forecast standard errors (σ) are Monte Carlo estimates from repeated simulations under the fitted model.

Table 4.5 compares the mean predicted number of events for each ward over the interval with the actual observed number of events. The results indicate that the 12D-DHPA model generally performs well in forecasting these unseen data. However, it appears to overestimate the number of events in Wards 5 and 7. This may reflect (i) the long alarm-marked memory estimated for certain cross-ward pairs, and/or (ii) a conservative uncertainty treatment that propagates alarm-triggered cascades across the test window.

4.7.4 Limitations

Our analysis has several limitations. First, event times are manually recorded and rounded to five-minute bins, permitting ties and potential under-reporting; the discrete specification accommodates ties but cannot recover true ordering. Second, excitation kernels are constrained to be nonnegative and thus do not capture short-lag inhibition or refractory behavior that may follow staff intervention; allowing signed or bi-phasic kernels (e.g., difference-of-geometrics) are a natural extension. Third, alarms are modeled as a stationary Bernoulli mark estimated on training data and likely correlate with incident severity and staffing, so alarm-related effects should be interpreted as predictive associations rather than causal impacts. Finally, we hold the seasonal baseline fixed and omit additional covariates (e.g., weather, staffing levels, clinical acuity); future work could relax these assumptions and assess calibration/recalibration of forecasts in operational settings.

4.8 Conclusion

This work addresses two main challenges. First, we developed an efficient discrete Hawkes model for large datasets, using a geometric distribution for the excitation kernel and a constant-space, event-indexed structure for likelihood and gradient computations. Second, we specified a marked discrete Hawkes process to model violent events in a forensic psychiatric hospital, offering insights into self- and cross-excitation between wards and the influence of alarms.

Our model improves predictive performance over standard discrete Hawkes models and could support practical applications, such as data-driven adjustments to staffing, treatment programs, and patient exposure to alarms. We interpret the attribution results under the model's branching representation. The observed strong excitation, both with and without alarms, suggests the need for further investigation into underlying mechanisms.

Risk assessments in forensic psychiatric settings rely heavily on historical data, but incorporating dynamic situational factors could enhance their effectiveness (Desmarais, Nicholls, Wilson, & Brink, 2012). The observed excitation effects may reflect model-based signatures consistent with social contagion—where awareness of violent incidents is associated with subsequent aggression (Beck et al., 2018)—particularly given patients' trauma histories and sensitivity to danger signals like alarms.

Chapter 5

A Nonparametric Discrete Hawkes Model with a Collapsed Gaussian-Process Prior

5.1 Introduction

Most existing DHP models rely on restrictive parametric assumptions. The baseline intensity is often modeled as constant or sinusoidal, while excitation kernels are restricted to geometric or negative-binomial forms. While effective in some settings, such assumptions can limit the ability to capture long memory, nonstationarity, or changes in excitation over time.

To address these limitations, (Browning et al., 2022) introduced a nonparametric DHP using a random histogram prior over the excitation kernel with trans-dimensional MCMC for inference. While this approach allows data-driven excitation structure, the use of a fixed intercept for the baseline may limit its ability to capture smooth or evolving background dynamics. Moreover, the sampling scheme can be costly for long processes.

In this chapter, we introduce the *Gaussian Process Discrete Hawkes Process (GP-DHP)*, a nonparametric model for discrete-time self-exciting count data. GP-DHP places independent Gaussian process priors on the baseline and excitation functions, enabling smooth, data-adaptive estimation without strong parametric constraints. Crucially, we collapse these priors to a single latent GP over the additive intensity, enabling efficient MAP inference and an interpretable decomposition into exogenous (baseline) and endogenous (excitation) components. This yields near-linear-time $\mathcal{O}(T \log T)$ complexity in practice via FFT-based multiplications and structured kernel interpolation. To our knowledge, there is currently no analogous GP-based formulation for discrete-time Hawkes that models both baseline and excitation nonparametrically; our work fills this gap.

Beyond introducing GP-DHP and the collapsed latent-GP inference described above, we also contribute:

1. A closed-form projection from the latent trajectory to interpretable baseline/excitation components;

2. Practical identifiability and stability diagnostics, including the branching-ratio statistic

$$\widehat{\kappa} = \sum_{d=1}^{D_{\max}} \max\{\widehat{\Phi}(d), 0\},$$

where $\widehat{\Phi}(d)$ denotes the estimated discrete excitation kernel at lag d, for $d = 1, \ldots, D_{\max}$, and $D_{\max}$ is the maximum lag considered. The quantity $\widehat{\kappa}$ therefore measures the total estimated positive excitation over the lag window and serves as an empirical analogue of the branching ratio.
3. Empirical evidence on simulations and two case studies (U.S. terrorism; weekly Cryptosporidiosis) showing improved test predictive log-likelihood and faithful decompositions; and
4. an open-source implementation to facilitate replication and reuse.

5.2 Proposed Model

5.2.1 Discrete Hawkes Process

We recall the discrete-time Hawkes process (DHP) as introduced in Section 2.7 and contrasted with its continuous-time counterpart in Section 2.6.3. In continuous time the conditional intensity is an instantaneous rate, whereas in discrete time we work on a fixed grid $t \in \mathbb{N}$ and model the *expected count per interval* via an additive decomposition into an exogenous baseline and an endogenous, history-driven excitation. This discrete formulation is appropriate when observations arrive at regular intervals and aligns with the notation and likelihood framework laid out in Chapter 2 (Section 2.2).

The DHP captures the empirical regularity that recent events raise the chance of future events across nearby intervals, that is, self-excitation, while accommodating cross-excitation in the multivariate setting of Section 2.7. It is therefore well suited to clustered count sequences such as infectious disease case totals, financial transaction volumes, or incident logs recorded hourly, daily, or weekly, where new occurrences can trigger follow-up events in subsequent intervals.

Let $Y_t \in \mathbb{N}$ denote the number of events observed during the interval $(t-1, t]$, and let the history available strictly before time t be denoted by

$$\mathscr{H}_{t-1} = \{Y_s : s \leq t-1\}.$$

The conditional mean intensity at time t is

$$\lambda(t) = \mathbb{E}[Y_t \mid \mathscr{H}_{t-1}] = \mu(t) + \sum_{d=1}^{t-1} Y_{t-d}\, \Phi(d), \tag{5.1}$$

where:

- $\mu(t) > 0$, for $t \in \mathbb{N}$, is the baseline intensity, representing spontaneous, that is, exogenous, events not triggered by past observations. It can capture systematic background variation beyond the influence of past events. In practice, this often reflects smooth trends or seasonal cycles. For example, in epidemiology $\mu(t)$ may encode annual patterns in disease incidence, while in security contexts it may capture differences between weekdays and weekends. This separation allows the model to distinguish predictable rhythms from bursty self-excitation.
- $\Phi(d) \in \mathbb{R}$, non-negative in the standard case, is the excitation kernel, describing the influence that events occurring d steps in the past exert on the present rate. For discrete Hawkes processes, stability requires the excitation kernel to be absolutely summable with total branching ratio

  $$\kappa = \sum_{d=1}^{\infty} \Phi(d) < 1.$$

  This condition prevents explosive growth of events and guarantees a well-defined mean process. If signed kernels are allowed, stability is enforced by bounding the sum of the positive part,

  $$\sum_{d \geq 1} \max\{\Phi(d), 0\} < 1.$$

The observed counts are modeled conditionally on the past as

$$Y_t \mid \mathscr{H}_{t-1} \sim \text{Poisson}\big(\lambda(t)\big).$$

This likelihood is widely used in DHPs because it naturally models integer-valued events and supports the additive decomposition of the intensity function. We emphasize that this is a modeling assumption, not a consequence of discretizing the continuous-time process. That is, a Poisson likelihood is not implied by discretizing a continuous-time Hawkes process; rather, it is a standard choice in the discrete-time literature for tractability and interpretability. Our nonparametric approach refers to the structure of the intensity function rather than the observation model.

Branching Representation of the Discrete Hawkes Process

The discrete Hawkes model above is often interpreted through the Hawkes cluster representation: events arise either as *immigrants* from the baseline $\mu(t)$ or as *offspring* triggered by past events, with integer lags distributed according to $\Phi(d)$ (Hawkes, 1971; Hawkes & Oakes, 1974). In this view, bursts form around immigrant arrivals as offspring generate further descendants. We call this the branching process interpretation.

This branching process interpretation, in some cases, facilitates practical inference and simulation. Each event at time t is assigned a parent source, either an earlier event (endogenous or triggered) or a spontaneous immigrant (exogenous event). The set of all such parent–child relationships is known as the branching structure.

Let $Y_t \in \mathbb{N}$ denote the observed number of events at time t, and define a latent vector $\boldsymbol{y}^{(t)} = (y_0^{(t)}, y_1^{(t)}, \ldots, y_{t-1}^{(t)})$, where: - $y_0^{(t)}$ is the number of immigrant (baseline) events at time t, - $y_d^{(t)}$ is the number of offspring events at time t triggered by events at time $t-d$, for $d \in \{1, \ldots, t-1\}$.

These quantities satisfy the constraint:

$$Y_t = \sum_{d=0}^{t-1} y_d^{(t)}.$$

Given the branching structure $\{\boldsymbol{y}^{(t)}\}_{t=1}^{T}$, the complete-data likelihood factorizes into a product of contributions from baseline and excitation sources. The probability of an event being triggered by a parent $t-d$ is proportional to the weight $\Phi(d)$, and the probability of being an immigrant is proportional to $\mu(t)$. The normalizing constant is the total expected intensity at time t, that is, $\lambda(t) = \mu(t) + \sum_{d=1}^{t-1} Y_{t-d}\Phi(d)$.

For inference, the branching assignments can be interpreted as draws from a multinomial distribution:

$$(y_0^{(t)}, y_1^{(t)}, \ldots, y_{t-1}^{(t)}) \sim \text{Multinomial}\left(Y_t; \frac{\mu(t)}{\lambda(t)}, \frac{Y_{t-1}\Phi(1)}{\lambda(t)}, \ldots, \frac{Y_1\Phi(t-1)}{\lambda(t)}\right).$$

This formulation enables efficient simulation and inference schemes, such as Gibbs sampling, by marginalizing over the latent branching structure or sampling it explicitly.

While Gibbs sampling is effective for parametric DHPs, it can perform poorly for the model in this chapter due to slow mixing under Gaussian-process priors. Thus, we develop an alternative MAP-based inference procedure that scales efficiently and avoids these issues.

5.2.2 Proposed Model

Standard inference in Hawkes processes typically involves estimating the baseline and excitation functions either through maximum likelihood under parametric forms, or via sampling-based Bayesian approaches. In the discrete-time setting, parametric models often assume constant or periodic baselines and simple excitation forms, for example geometric or exponential-type decay (Browning et al., 2021), which can limit flexibility. Fully Bayesian approaches with nonparametric priors exist but tend to be computationally intensive, especially when multiple latent functions must be estimated jointly. In contrast, our approach collapses the GP priors

over the baseline and excitation functions into a single prior over the latent intensity trajectory, enabling efficient MAP inference with scalable optimisation. This improves interpretability while avoiding expensive sampling or marginalisation steps, making the method suitable for larger temporal datasets.

To this end, we propose a nonparametric extension of the discrete-time Hawkes process in which both the baseline and the excitation kernel are modelled using Gaussian process priors. While several works have explored GP-based intensity modelling in the continuous-time Hawkes setting, see for example (Adams et al., 2009; Lloyd, Gunter, Osborne, & Roberts, 2015; Malem-Shinitski, Ojeda, & Opper, 2022), to our knowledge there are currently no analogous GP-based formulations for discrete-time Hawkes processes. Our work addresses this gap by developing a discrete-time nonparametric model that retains the interpretability of its parametric counterparts. We first briefly recall Gaussian processes.

Gaussian processes

A Gaussian process (GP) is a Bayesian nonparametric prior over functions. Formally, it is a collection of random variables such that every finite subset is jointly Gaussian. A GP is specified by a mean function $m(x)$ and a covariance function $k(x,x')$, and is written

$$f(x) \sim \mathcal{GP}\big(m(x), k(x,x')\big).$$

For inputs $\{x_1,\dots,x_n\}$, the vector of evaluations

$$\mathbf{f} = \big(f(x_1),\dots,f(x_n)\big)^\top$$

satisfies

$$\mathbf{f} \sim \mathcal{N}(\mathbf{m},\mathbf{K}), \qquad \mathbf{m}_i = m(x_i), \qquad \mathbf{K}_{ij} = k(x_i,x_j).$$

GPs are well suited to modelling time-varying functions when the functional form is unknown but expected to exhibit structure such as smoothness, periodicity, or long-term trends. Appropriate covariance functions encode these assumptions while preserving tractable finite-dimensional inference; common examples include squared-exponential kernels for smoothness, periodic kernels for seasonality, and linear kernels for trend-like behaviour. For a comprehensive introduction, see (C. E. Rasmussen & Williams, 2006).

Gaussian Process Discrete Hawkes Process (GP-DHP)

We model the latent process $\ell(t)$ through the discrete-time Hawkes decomposition

$$\ell(t) = \mu(t) + \sum_{d=1}^{t-1} Y_{t-d}\,\Phi(d), \tag{5.2}$$

where $\mu(t)$ is the baseline function, $\Phi(d)$ is the discrete excitation kernel, and $d \in \{1,2,\dots\}$ indexes lags. The observed counts satisfy

$$\lambda(t) = \max\{0,\ell(t)\}, \qquad Y_t \mid \mathrm{H}_{t-1} \sim \mathrm{Poisson}\big(\lambda(t)\big),$$

where $\mathrm{H}_{t-1} = \{Y_s : s \le t-1\}$ denotes the history strictly prior to time t. We use the rectifier $\max\{0,\ell(t)\}$ throughout to ensure a nonnegative intensity. Although the map is nondifferentiable at 0, in practice we found that MAP optimisation with damping and line search converges reliably.

As in any additive Hawkes decomposition, $\mu(t)$ and the excitation contribution $\sum_d Y_{t-d}\Phi(d)$ are not uniquely identifiable without additional structure. If $\mu(t)$ were modelled by an unconstrained GP, slowly varying temporal patterns could be attributed either to the baseline or to accumulated excitation, leading to ambiguous decompositions. We address this by imposing a seasonally structured prior on $\mu(t)$ and a nonstationary lag-dependent prior on $\Phi(d)$ that shrinks long-range effects. Together these priors restrict the solution space and yield stable, interpretable decompositions; see Section 5.4.1.

GP prior for the baseline $\mu(t)$. We place a zero-mean GP prior on the baseline,

$$\mu \sim \mathcal{GP}\big(0, k_b\big).$$

On a finite grid $t = 1,\dots,T$, this induces the baseline covariance matrix

$$\mathbf{K}_b \in \mathbb{R}^{T\times T}, \qquad [\mathbf{K}_b]_{tt'} = k_b(t,t').$$

We choose

$$k_b(t,t') = \underbrace{\sigma^2_{\mathrm{per}} \exp\left(-\frac{2\sin^2\big(\pi(t-t')/P\big)}{\ell^2_{\mathrm{per}}}\right)}_{\text{periodic seasonal term}} + \underbrace{\sigma^2_{\mathrm{lin}}\, tt'}_{\text{linear trend term}} + \underbrace{\varepsilon_b^2\,\delta_{tt'}}_{\text{jitter}}, \tag{5.3}$$

where P is the seasonal period, for example $P = 52$ for weekly data or $P = 365$ for daily data, ℓ_{per} controls seasonal smoothness, σ^2_{lin} controls the magnitude of long-term linear variation, and $\delta_{tt'}$ is the Kronecker delta, equal to 1 if $t = t'$ and 0 otherwise.

GP prior for the excitation kernel $\Phi(d)$. We place a zero-mean GP prior on the excitation kernel,

$$\Phi \sim \mathscr{GP}\big(0, k_f\big),$$

defined over discrete lags $d \in \{1, \ldots, D_{\max}\}$. On this lag grid, the corresponding excitation covariance matrix is

$$\mathbf{K}_f \in \mathbb{R}^{D_{\max} \times D_{\max}}, \qquad [\mathbf{K}_f]_{dd'} = k_f(d, d').$$

Rather than multiplying by an explicit exponential envelope, we encode nonstationarity through both an amplitude envelope and an input warping of the lags. Let

$$a(d) = \sigma_f \exp\left(-\frac{\beta d}{2}\right), \qquad g(d) = \frac{1 - e^{-\beta d}}{\beta\, \ell_f},$$

with $\sigma_f > 0$ an amplitude scale, $\ell_f > 0$ a base length-scale, and $\beta \geq 0$ controlling the rate at which the effective metric and amplitude change with lag. Define the stationary squared-exponential kernel on the warped inputs by

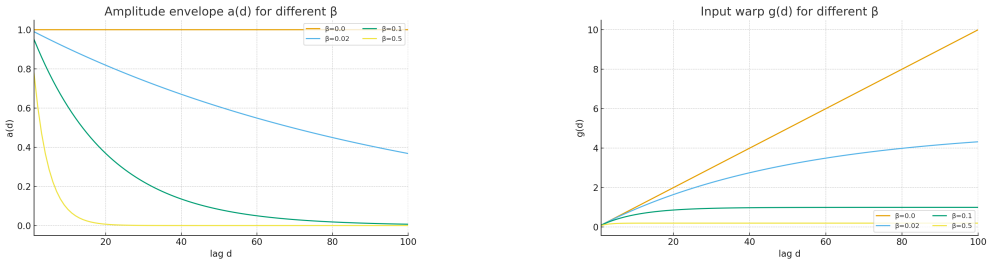
$$k_{\mathrm{RBF}}\big(g(d), g(d')\big) = \exp\left(-\frac{1}{2}\big(g(d) - g(d')\big)^2\right).$$

We then set

$$k_f(d, d') = a(d)a(d')\, k_{\mathrm{RBF}}\big(g(d), g(d')\big) + \varepsilon_f^2\, \delta_{dd'}. \tag{5.4}$$

This construction yields smooth short-lag correlations while simultaneously shrinking long-lag variability via $a(d)$ and compressing large lags through the warp $g(\cdot)$. As $\beta \to 0$, we recover a stationary RBF prior on the original lag scale because $a(d) \to \sigma_f$ and $g(d) \to d/\ell_f$. For $\beta > 0$, the prior increasingly attenuates and decorrelates remote lags, which discourages the excitation kernel $\Phi(d)$ from absorbing slow-moving trends that should instead be attributed to the baseline $\mu(t)$.

Figure 5.2 illustrates how the GP prior on the excitation kernel behaves as the attenuation parameter β increases. The top row shows sample draws of $\Phi(d)$ under different β values, highlighting how larger β suppresses long-lag fluctuations and concentrates mass near the origin. The bottom row displays the corresponding covariance heatmaps for $\mathbf{K}_f$. As β increases, the excitation functions flatten at longer lags and exhibit smoother decay, reflecting a reduced influence of the distant past.



(a) $a(d)$ **(b)** $g(d)$

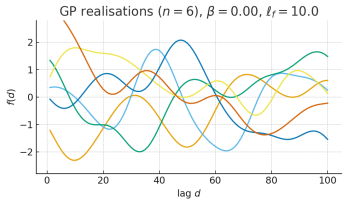
Figure 5.1: Sweep over the attenuation parameter β for the excitation-kernel components defined in (5.4): the envelope $a(d) = \sigma_f e^{-\beta d/2}$, which controls the marginal variance across lags, and the warp $g(d) = (1 - e^{-\beta d})/(\beta \ell_f)$, which controls the effective length-scale through input warping. Larger β shrinks long-lag variability and compresses large lags, discouraging the excitation kernel $\Phi(d)$ from absorbing slow trends that should instead be attributed to the baseline $\mu(t)$. Fixed parameters: $\sigma_f = 1$, $\ell_f = 10$, $D_{\max} = 100$, and $\beta \in \{0, 0.02, 0.1, 0.5\}$.

Discussion. The combination of (5.3) and (5.4) provides the inductive bias needed for identifiability: repeated seasonal structure and trend are captured by the baseline $\mu(t)$, whereas $\Phi(d)$ explains short- to medium-range self-excitation with a principled decay built into the prior (Malem-Shinitski et al., 2022), without imposing a fixed parametric excitation shape.

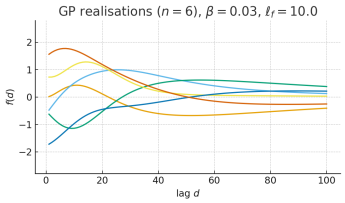
5.3 Inference

Our inference procedure centres on the latent intensity process $\ell(t)$, defined in Equation (5.2), which we treat as the primary object of estimation. Since both the baseline $\mu(t)$ and the excitation kernel $\Phi(d)$ are endowed with Gaussian process priors, the resulting latent trajectory $\ell(t)$ inherits a well-defined Gaussian prior induced by these two components (C. E. Rasmussen & Williams, 2006). Rather than estimating $\mu(t)$ and $\Phi(d)$ jointly as separate latent functions, we integrate them out analytically to obtain a collapsed prior over $\ell(t)$. This yields a single GP prior over the full latent trajectory, improving computational efficiency while preserving the additive Hawkes structure.

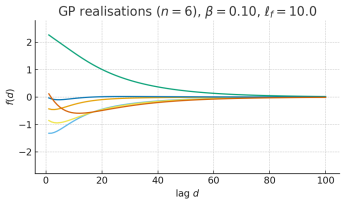
While discrete Hawkes models are often interpreted through a branching representation, we do not use that representation for inference. In our experiments, standard sampling-based approaches based on branching variables mixed poorly when combined with GP priors. This motivates the collapsed latent-intensity representation adopted here, together with direct MAP estimation, which avoids explicit branching variables and enables stable, scalable inference.



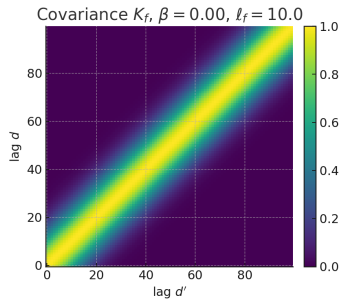
(a) $f(d)$, a GP realisation of the excitation kernel, with $\beta = 0.00$ and $\ell_f = 10$



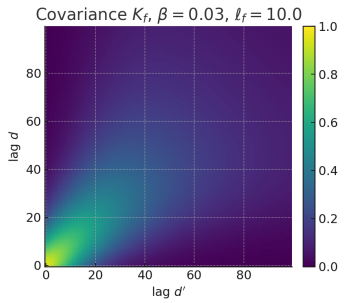
(b) $f(d)$, a GP realisation of the excitation kernel, with $\beta = 0.03$ and $\ell_f = 10$



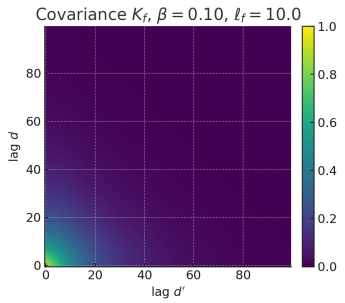
(c) $f(d)$, a GP realisation of the excitation kernel, with $\beta = 0.10$ and $\ell_f = 10$



(d) $\mathbf{K}_f$ heatmap, $\beta = 0.00$, $\ell_f = 10$



(e) $\mathbf{K}_f$ heatmap, $\beta = 0.03$, $\ell_f = 10$



(f) $\mathbf{K}_f$ heatmap, $\beta = 0.10$, $\ell_f = 10$

Figure 5.2: Draws from the GP prior over the excitation kernel $\Phi(d)$ for increasing values of β at fixed ℓ_f. Top: sample draws of $\Phi(d)$. Bottom: corresponding heatmaps of the finite covariance matrix $\mathbf{K}_f$.

5.3.1 Collapsed GP Prior

Inference over two separate functions, the baseline $\mu(t)$ and the excitation kernel $\Phi(d)$, is computationally expensive because Φ enters the model through a sum over all past observations at each time point. We therefore collapse these priors into a single GP prior over the latent intensity vector $\boldsymbol{\ell}$, so that optimisation takes place in a T-dimensional space. This reduces memory and computation, and enables fast matrix–vector operations through the structure of the baseline covariance matrix $\mathbf{K}_b$ and the excitation covariance matrix $\mathbf{K}_f$.

Recall from Equation (5.2) that the latent intensity at time t is

$$\ell(t) = \mu(t) + \sum_{d=1}^{t-1} Y_{t-d}\,\Phi(d).$$

Let

$$\boldsymbol{\ell} = \big(\ell(1), \ldots, \ell(T)\big)^\top$$

denote the vector of latent intensities over a horizon of length T. Because

$$\mu \sim \mathcal{GP}(0, k_b) \qquad \text{and} \qquad \Phi \sim \mathcal{GP}(0, k_f)$$

are assumed independent, it follows that ℓ is multivariate Gaussian with mean zero and covariance matrix $\mathbf{K} \in \mathbb{R}^{T\times T}$, whose entries are

$$[\mathbf{K}]_{ts} = k_b(t,s) + \sum_{d=1}^{t-1}\sum_{d'=1}^{s-1} Y_{t-d}\, Y_{s-d'}\, k_f(d,d'), \qquad t,s = 1,\dots,T. \tag{5.5}$$

Now let $\mathbf{X} \in \mathbb{R}^{T\times(T-1)}$ denote the strictly lower-triangular lagged-count design matrix with entries

$$X_{t,d} = \begin{cases} Y_{t-d}, & 1 \le d \le t-1, \\ 0, & \text{otherwise}, \end{cases} \qquad t = 1,\dots,T, \quad d = 1,\dots,T-1.$$

Let $\mathbf{K}_f \in \mathbb{R}^{(T-1)\times(T-1)}$ be the excitation covariance matrix with entries

$$[\mathbf{K}_f]_{dd'} = k_f(d,d'), \qquad d,d' = 1,\dots,T-1,$$

and let $\mathbf{K}_b \in \mathbb{R}^{T\times T}$ be the baseline covariance matrix with entries

$$[\mathbf{K}_b]_{ts} = k_b(t,s), \qquad t,s = 1,\dots,T.$$

Then the collapsed covariance matrix can be written compactly as

$$\mathbf{K} = \mathbf{K}_b + \mathbf{X}\mathbf{K}_f\mathbf{X}^\top.$$

This formulation avoids the need to represent $\mu(t)$ and $\Phi(d)$ explicitly during inference. It is analogous to collapsed GP constructions used in GP-modulated Poisson processes in continuous time (Malem-Shinitski et al., 2022), but adapted here to the discrete-time Hawkes setting. Unlike Malem-Shinitski et al. (2022), which uses an explicit exponential envelope together with a stationary RBF covariance function for the excitation, we remove the deterministic envelope from the mean structure and instead encode attenuation and smoothness directly within the GP prior for Φ.

5.3.2 Decomposition into Baseline and Excitation Components

While inference is performed on the latent intensity vector ℓ^*, recovering the individual components $\mu(t)$ and $\Phi(d)$ is desirable for interpretability, particularly in epidemiological or social applications where disentangling baseline dynamics from self-excitation can provide mechanistic insight.

To this end, we approximate the MAP estimate of the latent intensity in the additive form

$$\ell^* \approx \boldsymbol{\mu} + \mathbf{X}\boldsymbol{\Phi},$$

where $\boldsymbol{\mu} \in \mathbb{R}^T$ is the baseline vector, $\boldsymbol{\Phi} \in \mathbb{R}^{T-1}$ is the excitation kernel evaluated at discrete lags, and $\mathbf{X} \in \mathbb{R}^{T\times(T-1)}$ is the lagged-count design matrix introduced above. Its entries are

$$X_{t,d} = \begin{cases} Y_{t-d}, & \text{if } 1 \le d \le t-1, \\ 0, & \text{otherwise,} \end{cases} \qquad t = 1,\dots,T,\ \ d = 1,\dots,T-1.$$

This decomposition mirrors the additive structure of Equation (5.2) in matrix form. It therefore provides a natural way to project the fitted latent intensity $\boldsymbol{\ell}^*$ back onto interpretable baseline and excitation components.

Ideally, one would recover $\boldsymbol{\mu}$ and $\boldsymbol{\Phi}$ by solving the constrained optimisation problem

$$\min_{\boldsymbol{\mu},\boldsymbol{\Phi}} \left\{ \frac{1}{2}\boldsymbol{\mu}^\top \mathbf{K}_b^{-1}\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{\Phi}^\top \mathbf{K}_f^{-1}\boldsymbol{\Phi} \right\} \quad \text{subject to } \boldsymbol{\ell}^* = \boldsymbol{\mu} + \mathbf{X}\boldsymbol{\Phi}.$$

This corresponds to the maximum a posteriori decomposition under the constraint that $\boldsymbol{\ell}^*$ is exactly represented as the sum of baseline and excitation terms. We next show that this constrained optimisation problem has a unique closed-form solution.

Proposition 5.3.1 (Hard-constraint decomposition: existence, uniqueness, and closed-form solution)**.** *Let $\mathbf{K}_b \in \mathbb{R}^{T\times T}$ and $\mathbf{K}_f \in \mathbb{R}^{(T-1)\times(T-1)}$ be symmetric positive definite matrices, let $\mathbf{X} \in \mathbb{R}^{T\times(T-1)}$, and let $\boldsymbol{\ell}^* \in \mathbb{R}^T$. Consider*

$$\min_{\boldsymbol{\mu},\boldsymbol{\Phi}} J(\boldsymbol{\mu},\boldsymbol{\Phi}) := \frac{1}{2}\boldsymbol{\mu}^\top \mathbf{K}_b^{-1}\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{\Phi}^\top \mathbf{K}_f^{-1}\boldsymbol{\Phi} \quad \textit{subject to} \quad \boldsymbol{\ell}^* = \boldsymbol{\mu} + \mathbf{X}\boldsymbol{\Phi}. \tag{5.6}$$

Define the collapsed covariance matrix

$$\mathbf{K} := \mathbf{K}_b + \mathbf{X}\mathbf{K}_f\mathbf{X}^\top \in \mathbb{R}^{T\times T}.$$

Then:

1. *The feasible set is nonempty and the objective is strictly convex on it. Hence there is a unique minimiser.*
2. *The unique minimiser $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Phi}})$ is*

$$\widehat{\boldsymbol{\mu}} = \mathbf{K}_b\mathbf{K}^{-1}\boldsymbol{\ell}^*, \qquad \widehat{\boldsymbol{\Phi}} = \mathbf{K}_f\mathbf{X}^\top\mathbf{K}^{-1}\boldsymbol{\ell}^*.$$

3. *The minimum value equals*

$$\frac{1}{2}\boldsymbol{\ell}^{*\top}\mathbf{K}^{-1}\boldsymbol{\ell}^*.$$

Proof. Feasibility and strict convexity. Feasibility holds since $(\boldsymbol{\mu},\boldsymbol{\Phi}) = (\boldsymbol{\ell}^*, 0)$ satisfies the constraint $\boldsymbol{\ell}^* = \boldsymbol{\mu} + \mathbf{X}\boldsymbol{\Phi}$. The objective

$$J(\boldsymbol{\mu},\boldsymbol{\Phi}) = \frac{1}{2}\boldsymbol{\mu}^\top \mathbf{K}_b^{-1}\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{\Phi}^\top \mathbf{K}_f^{-1}\boldsymbol{\Phi}$$

is strictly convex because $\mathbf{K}_b^{-1} \succ 0$ and $\mathbf{K}_f^{-1} \succ 0$. The feasible set is an affine subspace. Therefore a unique minimiser exists.

KKT conditions and solution. Form the Lagrangian

$$\mathscr{L}(\boldsymbol{\mu}, \boldsymbol{\Phi}, \boldsymbol{\lambda}) = \frac{1}{2}\boldsymbol{\mu}^\top \mathbf{K}_b^{-1}\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{\Phi}^\top \mathbf{K}_f^{-1}\boldsymbol{\Phi} + \boldsymbol{\lambda}^\top(\boldsymbol{\ell}^* - \boldsymbol{\mu} - \mathbf{X}\boldsymbol{\Phi}),$$

with multiplier $\boldsymbol{\lambda} \in \mathbb{R}^T$. Stationarity gives

$$\nabla_{\boldsymbol{\mu}}\mathscr{L} = \mathbf{K}_b^{-1}\boldsymbol{\mu} - \boldsymbol{\lambda} = 0 \;\Rightarrow\; \boldsymbol{\mu} = \mathbf{K}_b\boldsymbol{\lambda},$$

and

$$\nabla_{\boldsymbol{\Phi}}\mathscr{L} = \mathbf{K}_f^{-1}\boldsymbol{\Phi} - \mathbf{X}^\top\boldsymbol{\lambda} = 0 \;\Rightarrow\; \boldsymbol{\Phi} = \mathbf{K}_f\mathbf{X}^\top\boldsymbol{\lambda}.$$

Primal feasibility enforces

$$\boldsymbol{\ell}^* - \boldsymbol{\mu} - \mathbf{X}\boldsymbol{\Phi} = \boldsymbol{\ell}^* - \mathbf{K}_b\boldsymbol{\lambda} - \mathbf{X}\mathbf{K}_f\mathbf{X}^\top\boldsymbol{\lambda} = 0,$$

that is,

$$\mathbf{K}\boldsymbol{\lambda} = \boldsymbol{\ell}^*, \qquad \mathbf{K} := \mathbf{K}_b + \mathbf{X}\mathbf{K}_f\mathbf{X}^\top.$$

For any nonzero $v \in \mathbb{R}^T$,

$$v^\top\mathbf{K}v = v^\top\mathbf{K}_b v + (\mathbf{X}^\top v)^\top\mathbf{K}_f(\mathbf{X}^\top v) > 0,$$

since $\mathbf{K}_b \succ 0$ and $\mathbf{K}_f \succ 0$. Hence $\mathbf{K} \succ 0$ and is invertible. Thus

$$\boldsymbol{\lambda} = \mathbf{K}^{-1}\boldsymbol{\ell}^*,$$

and substituting back yields

$$\widehat{\boldsymbol{\mu}} = \mathbf{K}_b\mathbf{K}^{-1}\boldsymbol{\ell}^*, \qquad \widehat{\boldsymbol{\Phi}} = \mathbf{K}_f\mathbf{X}^\top\mathbf{K}^{-1}\boldsymbol{\ell}^*.$$

These satisfy the KKT system and the constraint, hence are optimal. Uniqueness follows from strict convexity.

Minimum value. Using $\widehat{\boldsymbol{\mu}} = \mathbf{K}_b\boldsymbol{\lambda}$ and $\widehat{\boldsymbol{\Phi}} = \mathbf{K}_f\mathbf{X}^\top\boldsymbol{\lambda}$ with $\boldsymbol{\lambda} = \mathbf{K}^{-1}\boldsymbol{\ell}^*$,

$$J(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Phi}}) = \frac{1}{2}\boldsymbol{\lambda}^\top\mathbf{K}_b\boldsymbol{\lambda} + \frac{1}{2}\boldsymbol{\lambda}^\top\mathbf{X}\mathbf{K}_f\mathbf{X}^\top\boldsymbol{\lambda} = \frac{1}{2}\boldsymbol{\lambda}^\top\mathbf{K}\boldsymbol{\lambda} = \frac{1}{2}\boldsymbol{\ell}^{*\top}\mathbf{K}^{-1}\boldsymbol{\ell}^*.$$

This completes the proof. □

Interpretation. The solution $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Phi}})$ is the unique minimum-norm decomposition, in the reproducing kernel Hilbert space (RKHS) norms induced by $\mathbf{K}_b$ and $\mathbf{K}_f$, that exactly reconstructs $\boldsymbol{\ell}^*$. Equivalently, it is the orthogonal projection of $\boldsymbol{\ell}^*$ onto the sum of the baseline and excitation function spaces weighted by their priors. This yields an interpretable and reproducible mapping from the fitted latent trajectory back to baseline and excitation components without introducing any additional tuning parameters.

5.3.3 Computational Complexity and Efficiency

A key advantage of our framework is that it scales to long sequences while maintaining tractable computational cost. Classical implementations of discrete Hawkes processes, as defined in Equation (5.1), often incur $\mathcal{O}(T^2)$ complexity. This is because evaluating the excitation term

$$\sum_{d=1}^{t-1} Y_{t-d}\,\Phi(d)$$

at each time t requires summing over the full event history, and this must be repeated for each time step $t = 1, \ldots, T$. While truncation strategies can reduce this burden in parametric settings, such techniques are less effective in nonparametric models because ignoring long-range dependence can introduce bias.

In contrast, our GPDHP formulation allows scalable inference with efficient evaluation of the log-likelihood and its gradients during optimisation. This is primarily due to two design features:

- **Collapsed latent GP formulation:** Rather than placing priors on the baseline $\mu(t)$ and excitation kernel $\Phi(d)$ separately and sampling or optimising them directly, we collapse these into a single Gaussian process prior over the latent additive intensity $\ell(t)$, as described in Equation (5.5). This enables MAP estimation over the latent trajectory directly, avoiding repeated explicit recomputation of excitation sums at each time step.
- **Efficient log-likelihood evaluation:** With a Poisson observation model and an additive latent intensity, the log-likelihood decomposes into a sum over $t = 1, \ldots, T$, where each term depends only on $\ell(t)$. Temporal dependence is encoded in the GP prior, so likelihood evaluation itself contains no nested history summations.

An additional computational benefit arises from the structure of the collapsed excitation covariance matrix. Using the lagged-count design matrix $\mathbf{X}$ introduced above, the excitation contribution to the covariance can be written as

$$\mathbf{K}_{\text{exc}} = \mathbf{X}\mathbf{K}_f\mathbf{X}^\top,$$

whose entries satisfy

$$[\mathbf{K}_{\text{exc}}]_{ts} = \sum_{d=1}^{t-1}\sum_{d'=1}^{s-1} Y_{t-d}\,Y_{s-d'}\,[\mathbf{K}_f]_{dd'}.$$

Here $\mathbf{X} \in \mathbb{R}^{T\times(T-1)}$ is the strictly lower-triangular lagged-count design matrix, and $\mathbf{K}_f \in \mathbb{R}^{(T-1)\times(T-1)}$ is the finite covariance matrix induced by the GP prior on the excitation kernel $\Phi(d)$. Under our model, $\mathbf{K}_f$ admits the factorisation

$$\mathbf{K}_f = \mathbf{A}\,\mathbf{K}_{\text{stat}}\,\mathbf{A}, \qquad \mathbf{A} := \operatorname{diag}\big(a(1),\dots,a(T-1)\big), \quad a(d) = \sigma_f e^{-\beta d/2},$$

where $\mathbf{K}_{\text{stat}}$ is the stationary covariance matrix obtained by applying the RBF covariance function to the warped lags $u(d) = g(d) = (1-e^{-\beta d})/(\beta \ell_f)$:

$$[\mathbf{K}_{\text{stat}}]_{dd'} = \exp\Big(-\tfrac{1}{2}[u(d)-u(d')]^2\Big).$$

Hence

$$\mathbf{K}_{\text{exc}} = (\mathbf{X}\mathbf{A})\,\mathbf{K}_{\text{stat}}\,(\mathbf{X}\mathbf{A})^\top.$$

Fast matrix–vector multiplications (MVMs). MAP inference with a GP prior requires repeated application of covariance matrices to vectors, so efficiency hinges on fast matrix–vector multiplications:

1. *Multiplication by* $\mathbf{X}$ *and* $\mathbf{X}^\top$ *as convolutions.* For any vector $\boldsymbol{y}$, the products $\mathbf{X}^\top\boldsymbol{y}$ and $\mathbf{X}\boldsymbol{w}$ are cross-correlations between $\boldsymbol{y}$ or $\boldsymbol{w}$ and the count sequence $Y_\cdot$. These can be computed in $\mathcal{O}(T\log T)$ time via FFT-based linear convolution, after standard zero-padding, and the FFT of $Y_\cdot$ can be reused across iterations.
2. *Multiplication by* $\mathbf{A}$ *is diagonal.* Left or right multiplication by $\mathbf{A}$ costs $\mathcal{O}(T)$.
3. *Multiplication by* $\mathbf{K}_{\text{stat}}$. Although $\mathbf{K}_{\text{stat}}$ is stationary in the warped input $u(d)$, it is not Toeplitz with respect to the integer lag index d. We therefore use a structured-kernel interpolation, or inducing-grid, approximation: choose $M \ll T$ inducing points on a uniform grid in the u-domain, form a sparse interpolation matrix $\mathbf{W} \in \mathbb{R}^{(T-1)\times M}$, and approximate

   $$\mathbf{K}_{\text{stat}} \approx \mathbf{W}\mathbf{K}_U\mathbf{W}^\top,$$

   where $\mathbf{K}_U$ is the stationary RBF covariance matrix on the uniform inducing grid. Because this grid is uniform, $\mathbf{K}_U$ is Toeplitz, up to standard circulant embedding, and supports $\mathcal{O}(M\log M)$ matrix-vector multiplications via FFT; see (Xia, 2007). Since $\mathbf{W}$ is very sparse, $\mathbf{W}\boldsymbol{v}$ and $\mathbf{W}^\top\boldsymbol{v}$ cost $\mathcal{O}(T)$.

Combining these steps yields the following complexity for the excitation block:

$$\boldsymbol{y} \mapsto \mathbf{K}_{\text{exc}}\boldsymbol{y} = \mathbf{X}\mathbf{A}\,\underbrace{(\mathbf{W}\mathbf{K}_U\mathbf{W}^\top)}_{\approx\mathbf{K}_{\text{stat}}}\,\mathbf{A}\mathbf{X}^\top\boldsymbol{y} \quad \Rightarrow \quad \mathcal{O}(T\log T) + \mathcal{O}(M\log M) + \mathcal{O}(T).$$

The baseline covariance matrix $\mathbf{K}_b$ is the sum of a periodic component, amenable to FFT or circulant embedding for $\mathcal{O}(T \log T)$ matrix–vector multiplications, and low-rank terms such as linear or constant components. Hence $\mathbf{K}_b \boldsymbol{v}$ also costs $\mathcal{O}(T \log T)$. In practice, the overall per-iteration cost is dominated by FFTs on vectors of length $\mathcal{O}(T)$, and $\mathcal{O}(M)$ for the inducing grid, while memory is linear in T plus the storage for $\mathbf{W}$, proportional to its number of nonzero entries.

Uncertainty via Laplace approximation and projection. To quantify uncertainty, we use a Laplace approximation around the MAP latent trajectory $\boldsymbol{\ell}^*$. Writing the negative log posterior as

$$\mathscr{L}(\boldsymbol{\ell}) = -\sum_{t=1}^{T} \left\{ Y_t \log \lambda(t) - \lambda(t) \right\} + \frac{1}{2} \boldsymbol{\ell}^\top \mathbf{K}^{-1} \boldsymbol{\ell},$$

with

$$\lambda(t) = \max\{0, \ell(t)\},$$

the posterior precision at $\boldsymbol{\ell}^*$ is

$$\mathbf{H} = \mathbf{K}^{-1} + \mathbf{D}, \qquad \mathbf{D} = \operatorname{diag}\Big(\tfrac{Y_t}{\lambda(t)^2} \mathbf{1}\{\ell(t) > 0\} \Big)_{t=1}^{T},$$

and the latent covariance is approximated by

$$\boldsymbol{\Sigma}_\ell \approx \mathbf{H}^{-1}.$$

The baseline and excitation estimates are linear in $\boldsymbol{\ell}^*$, from Proposition 5.3.1,

$$\widehat{\boldsymbol{\mu}} = \mathbf{P}_b \boldsymbol{\ell}^*, \qquad \widehat{\boldsymbol{\Phi}} = \mathbf{P}_f \boldsymbol{\ell}^*,$$

with

$$\mathbf{P}_b = \mathbf{K}_b \mathbf{K}^{-1}, \qquad \mathbf{P}_f = \mathbf{K}_f \mathbf{X}^\top \mathbf{K}^{-1}.$$

Uncertainty therefore propagates according to

$$\operatorname{Cov}\begin{pmatrix} \widehat{\boldsymbol{\mu}} \\ \widehat{\boldsymbol{\Phi}} \end{pmatrix} \approx \begin{pmatrix} \mathbf{P}_b \\ \mathbf{P}_f \end{pmatrix} \boldsymbol{\Sigma}_\ell \begin{pmatrix} \mathbf{P}_b \\ \mathbf{P}_f \end{pmatrix}^\top,$$

yielding pointwise intervals from the corresponding marginals. In practice, we report posterior means and 95% intervals for μ and Φ obtained from these marginals.

Identifiability and regularization. Without additional structure the model is not identifiable: both $\mu(t)$ and the excitation term $\sum_{d=1}^{t-1} Y_{t-d}\,\Phi(d)$ enter additively, and the effective number of latent degrees of freedom grows with T. In our formulation, identifiability is improved by: (i) the periodic component of the baseline covariance matrix $\mathbf{K}_b$, which captures recurrent seasonal patterns; (ii) the linear and constant components, which absorb long-term level and trend; and (iii) the nonstationary prior structure

$$\mathbf{K}_f = \mathbf{A}\mathbf{K}_{\text{stat}}\mathbf{A},$$

where the amplitude envelope $a(d) = \sigma_f e^{-\beta d/2}$ down-weights distant lags and the warp $g(d) = (1 - e^{-\beta d})/(\beta\,\ell_f)$ compresses the effective metric at larger lags. We do not place any explicit exponential envelope in the mean of the excitation; attenuation is instead induced within the covariance through $a(\cdot)$ and $g(\cdot)$. This reduces the tendency of the excitation kernel $\Phi(\cdot)$ to absorb slow trends that should instead be attributed to the baseline $\mu(t)$, and thereby encourages stable decompositions.

Stability and diagnostics. For nonnegative excitation, a sufficient stability condition is the usual branching ratio

$$\kappa = \sum_{d\geq 1} \Phi(d) < 1.$$

For signed kernels, a conservative check is

$$\sum_{d\geq 1} \max\{\Phi(d), 0\} < 1.$$

Our nonstationary prior on $\Phi(d)$, through the amplitude envelope $a(d)$ and the warp $g(d)$, down-weights distant lags and thereby encourages a small branching ratio in practice. A simple post-fit diagnostic is

$$\widehat{\kappa} = \sum_{d=1}^{D_{\max}} \max\{\widehat{\Phi}(d), 0\},$$

computed from the projected excitation estimate $\widehat{\mathbf{\Phi}}$ in Proposition 5.3.1.

5.4 Experiments

5.4.1 Synthetic Data

We evaluate the ability of GPDHP to recover latent self-exciting dynamics by simulating from known discrete Hawkes models and assessing whether the model can reconstruct both the baseline and excitation components from observed count data. Neither the baseline function nor the excitation kernel is assumed known during inference.

Model Specification Each simulated time series is generated from a discrete Hawkes process of length $T = 6000$. The latent intensity is composed of a baseline function $\mu(t)$ and an excitation kernel $\Phi(d)$, as described in Section 5.2. The baseline is assigned a Gaussian process prior with covariance function

$$k_b(t,t') = \sigma_b^2 \exp\left(-\frac{2\sin^2\big(\pi(t-t')/P\big)}{\ell_b^2}\right) + \sigma_{b,c}^2 tt',$$

where $P = 52$ encodes annual periodicity for weekly data, ℓ_b controls the smoothness of the seasonal component, and $\sigma_{b,c}^2$ controls the magnitude of the linear trend.

The excitation kernel $\Phi(d)$ is assigned a zero-mean Gaussian process prior that is nonstationary in lag, built from an amplitude envelope and a warped RBF covariance function:

$$a(d) = \sigma_f \exp\Big(-\tfrac{\beta d}{2}\Big), \qquad g(d) = \frac{1-e^{-\beta d}}{\beta\,\ell_f},$$

$$k_f(d,d') = a(d)a(d')\exp\Big(-\tfrac{1}{2}\big(g(d)-g(d')\big)^2\Big).$$

This construction smoothly correlates nearby lags while attenuating long-range effects through $a(d)$ and compressing large lags via $g(d)$. Both GP priors are centered at zero. Inference proceeds over the latent additive intensity, followed by the closed-form projection to μ and Φ; see Section 5.3.

Hyperparameter Selection via Cross-Validation We select all kernel hyperparameters and the lag-attenuation parameter β via grid-based cross-validation. The first 4,000 time steps are used as a training set, and the remaining 2,000 as a validation set. Hyperparameters are chosen to maximize the Poisson log-likelihood on the validation set.

The following grid is used during cross-validation in both synthetic and real-data experiments:

Hyperparameter	**Search Range**
Lag attenuation β	$\{0.1, 0.2, 0.3, 0.4\}$
Baseline variance σ_b	$\{0.0001, 0.01, 1.0\}$
Baseline periodic length-scale ℓ_b	$\{1, 5, 100\}$
Linear trend variance $\sigma_{b,c}$	$\{0, 10^{-2}, 10^{-4}\}$
Excitation variance σ_f	$\{0.5, 1, 2\}$
Excitation length-scale ℓ_f	$\{5, 10, 20, 30\}$

Table 5.1: Hyperparameter ranges explored via grid search during cross-validation in both synthetic and real-data experiments.

Simulation Design: Excitation Function Recovery

To assess the flexibility of GPDHP, we simulate twelve synthetic datasets from discrete Hawkes models with fixed baseline structure but varying excitation dynamics. In each case, the goal is to evaluate whether the model can accurately recover the latent components from count observations.

Baseline Structure Across all simulations, we use a baseline function with a combination of linear and seasonal terms:

$$\mu(t) = \mu_0 + \mu_1 t + \mu_2 \sin\left(\frac{2\pi t}{P}\right) + \mu_3 \cos\left(\frac{2\pi t}{P}\right),$$

where $\mu_0, \mu_1, \mu_2, \mu_3$, and P are fixed constants. This form mimics typical nonstationary temporal behaviour observed in epidemiological and social event data. The baseline is fitted jointly with the excitation kernel in each experiment.

Excitation Kernel Families

Each dataset varies only in the excitation kernel $\Phi(d)$, which is drawn from one of four parametric families. In total, twelve configurations are used, with three parameterisations per family:

- *Negative Binomial (NB):* A heavy-tailed, overdispersed kernel with long memory:

  $$\Phi(d) = \alpha \binom{d+r-1}{d} (1-p)^d p^r,$$

  where $r > 0$ and $p \in (0,1)$. The parameter r controls peakiness, while p governs tail decay.
- *Geometric:* A memoryless excitation kernel, corresponding to the special case $r = 1$ of the negative binomial:

  $$\Phi(d) = \alpha\, p(1-p)^{d-1}, \qquad d \geq 1, \quad p \in (0,1).$$

  Smaller p induces longer-range dependence.
- *Power law:* A polynomially decaying kernel with tunable amplitude, near-lag width, and tail exponent:

  $$\Phi(d) = \alpha\, (\gamma + d)^{-\beta},$$

  where $\alpha > 0$ sets the overall scale, $\gamma \geq 0$ controls early-lag width or onset, and $\beta > 1$ determines tail heaviness, with larger β implying faster decay.

- *Bimodal Gaussian mixture:* A continuous mixture of two Gaussian modes:

$$\Phi(d) = \alpha \left[\frac{1}{2\sqrt{2\pi}\sigma} \exp\left(-\frac{(d-\mu_1^*)^2}{2\sigma^2} \right) + \frac{1}{2\sqrt{2\pi}\sigma} \exp\left(-\frac{(d-\mu_2^*)^2}{2\sigma^2} \right) \right],$$

  with $\sigma > 0$ and locations μ_1^*, μ_2^* specifying the modes. This form captures multimodal excitation with symmetric weight and width.

In each case, a time series of length $T = 6000$ is simulated, the model is fitted, with the first 4,000 entries used for training and the next 2,000 used for validation, and the projected posterior mean of the excitation kernel $\widehat{\Phi}(d)$ is compared with the ground truth. As shown in Figure 5.3, GPDHP accurately recovers a broad range of excitation structures despite the presence of a shared nonstationary baseline $\mu(t)$.

Figure 5.3 presents the fitted excitation kernels for all twelve simulations, grouped by kernel family. The model consistently isolates excitation dynamics across these distinct functional forms.

Simulation Design: Baseline Function Recovery

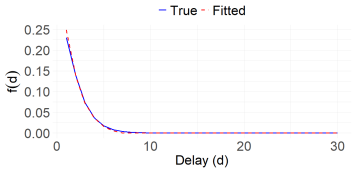
To assess the identifiability and robustness of our decomposition procedure, we conduct a controlled experiment in which the excitation kernel is held fixed, but is still estimated jointly with the baseline function, while the baseline function varies. Specifically, we simulate three synthetic sequences of length $T = 6000$, using the first 4,000 observations for training and the next 2,000 for validation, each with the same latent excitation kernel $\Phi(d)$ drawn from a negative binomial form:

$$\Phi(d) = \alpha \binom{d+r-1}{d} (1-p)^d p^r.$$

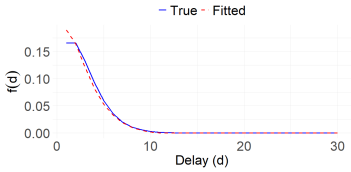
Only the baseline structure differs across the three cases, corresponding to three qualitatively distinct functional forms:

1. Constant,
2. linearly increasing,
3. linearly increasing with an added periodic component.

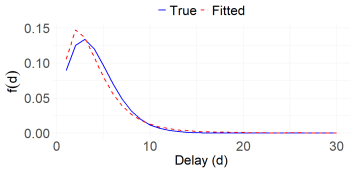
After fitting GPDHP to each dataset, we apply the post hoc decomposition described in Section 5.3 to recover the estimated baseline and excitation components. Figure 5.4 presents the results. In all three cases, the model accurately recovers the shape of the true baseline, and the estimated excitation kernels remain stable across experiments. This demonstrates that GPDHP can effectively disentangle spontaneous activity from triggered self-excitation, even when baseline dynamics differ substantially.



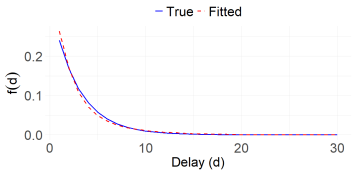
(a) Negative Binomial, $(\alpha, p, r) = (0.6, 0.6, 2)$



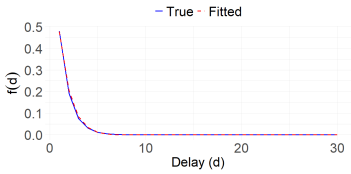
(b) Negative Binomial, $(\alpha, p, r) = (0.6, 0.6, 4)$



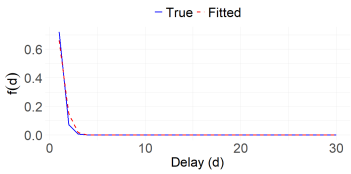
(c) Negative Binomial, $(\alpha, p, r) = (0.6, 0.6, 6)$



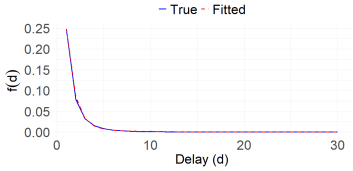
(d) Geometric, $(\alpha, p) = (0.8, 0.3)$



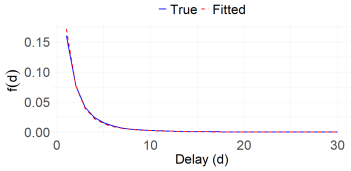
(e) Geometric, $(\alpha, p) = (0.8, 0.6)$



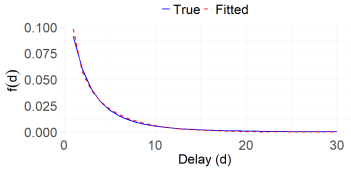
(f) Geometric, $(\alpha, p) = (0.8, 0.9)$



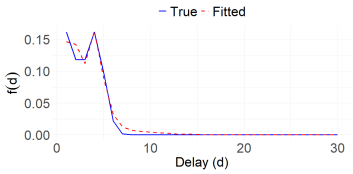
(g) Power law, $(\alpha, \gamma, \beta) = (20, 2, 4)$



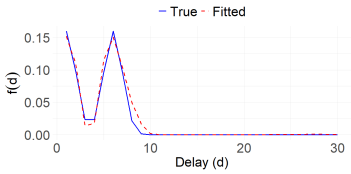
(h) Power law, $(\alpha, \gamma, \beta) = (100, 4, 4)$



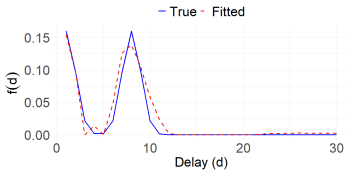
(i) Power law, $(\alpha, \gamma, \beta) = (600, 8, 4)$



(j) Bimodal Gaussian, $(\alpha, \mu_1^*, \mu_2^*, \sigma) = (0.8, 1, 6, 1)$



(k) Bimodal Gaussian, $(\alpha, \mu_1^*, \mu_2^*, \sigma) = (0.8, 1, 8, 1)$



(l) Bimodal Gaussian, $(\alpha, \mu_1^*, \mu_2^*, \sigma) = (0.8, 1, 10, 1)$

Figure 5.3: MAP estimates of the excitation kernels, denoted $\widehat{\Phi}(d)$, for twelve synthetic scenarios grouped by kernel family. In the plots, the label $f(d)$ refers to the ground-truth excitation function used to simulate the data; in the notation of this chapter, this corresponds to the excitation kernel $\Phi(d)$. Rows correspond to: (1) Negative Binomial with increasing shape r at fixed α and p; (2) Geometric with varying decay parameter p at fixed α; (3) power law $\Phi(d) = \alpha(\gamma + d)^{-\beta}$ with fixed $\beta = 4$ and increasing width via γ, with corresponding changes in α; and (4) bimodal Gaussian mixtures with fixed μ_1^* and σ and increasing separation via μ_2^*. All datasets share the same baseline $\mu(t)$.

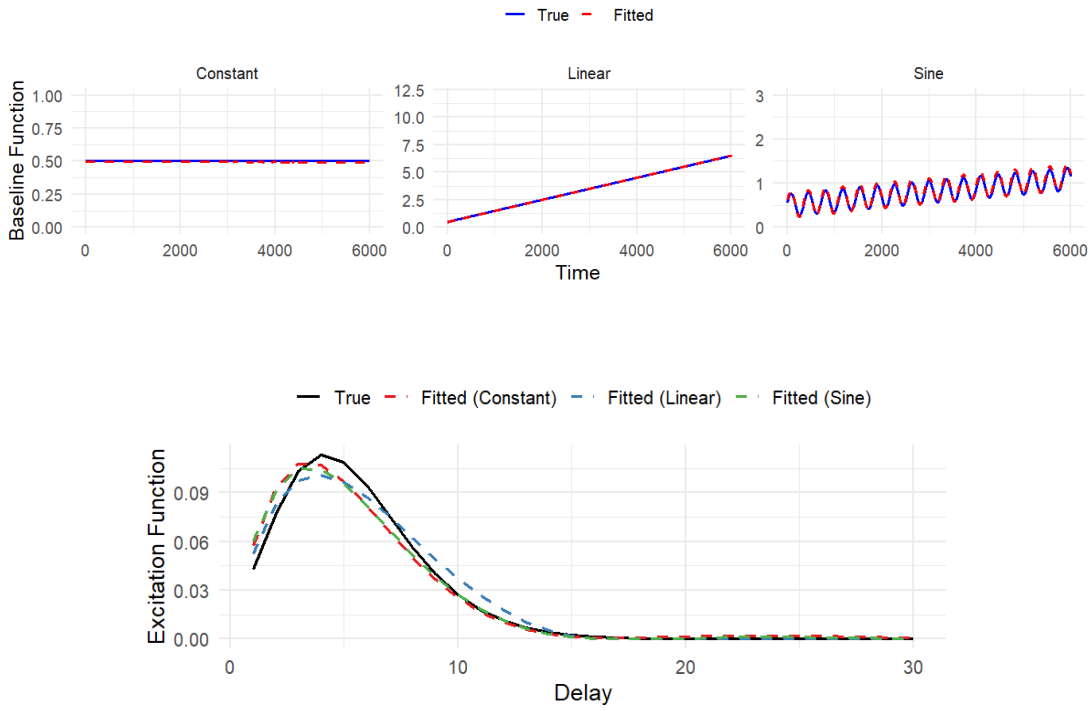


Figure 5.4: Recovery of baseline and excitation components across three distinct baseline settings. **Top panel:** Comparison of true (solid blue) and estimated (dashed red) baseline functions for each experiment. The functional forms correspond to constant, linear, and periodic baselines, respectively. **Bottom panel:** Estimated excitation kernels $\widehat{\Phi}(d)$ (coloured dashed lines) overlaid on the true excitation kernel (black solid). Despite varying baselines, the recovered excitation functions are consistent, confirming decomposition stability.

5.4.2 Real Data

We evaluate GPDHP and four benchmark parametric discrete-time Hawkes models on two real-world count datasets. For the benchmark models, we fix the excitation kernel to a negative-binomial form, a commonly used parametric choice in self-exciting count models (Porter & White, 2012), whereas GPDHP models the excitation kernel nonparametrically through a Gaussian process prior. All benchmark models share the same negative-binomial excitation kernel:

$$\Phi(d) = \binom{d+r-1}{d}(1-p)^d p^r,$$

where $d \in \mathbb{N}$, $r > 0$ is a dispersion parameter, and $p \in (0,1)$ controls temporal decay. This family encompasses both heavy-tailed and short-memory excitation. When $r = 1$, the kernel reduces to the geometric distribution:

$$\Phi(d) = (1-p)^d p,$$

which has been used in recent infectious-disease applications, including modelling COVID-19 mortality (Browning et al., 2021). The more flexible negative-binomial kernel has also been used in terrorism modelling (Porter & White, 2012).

Each model is evaluated using the one-step-ahead predictive log-likelihood (pLL) on a held-out test set. For a time series of length T, with test indices $\mathscr{T}_{\text{test}} \subset \{1,\dots,T\}$, we compute

$$\text{pLL} = \sum_{t \in \mathscr{T}_{\text{test}}} \log p(Y_t \mid \mathrm{H}_{t-1}),$$

where

$$\mathrm{H}_{t-1} = \{Y_s : s \le t-1\}$$

denotes the history strictly prior to time t, and $p(Y_t \mid \mathrm{H}_{t-1})$ is the predictive distribution under the fitted model.

Hyperparameters for GPDHP are selected by cross-validation using the same grid as in Table 5.1. The benchmark models differ only in their specification of the baseline intensity function $\mu(t)$:

- **Discrete DHP:** $\mu(t) = \gamma_0$
- **Linear DHP:** $\mu(t) = \gamma_0 + \gamma_1 t$
- **Sinusoidal DHP:** $\mu(t) = \gamma_0 + \gamma_1 \sin(2\pi t/P)$, with $P = 52$ for Cryptosporidiosis and $P = 365$ for U.S. terrorism events.
- **Linear + Sinusoidal DHP:** $\mu(t) = \gamma_0 + \gamma_1 t + \gamma_2 \sin(2\pi t/P)$, with $P = 52$ for Cryptosporidiosis and $P = 365$ for U.S. terrorism events.

Model	pLL
Discrete DHP	-607.7
Linear DHP	-631.0
Sinusoidal DHP	-607.6
Linear + Sinusoidal DHP	-628.1
GP-DHP	**-573.2**

Table 5.2: Predictive log-likelihood (pLL) on U.S. terrorism data (test set).

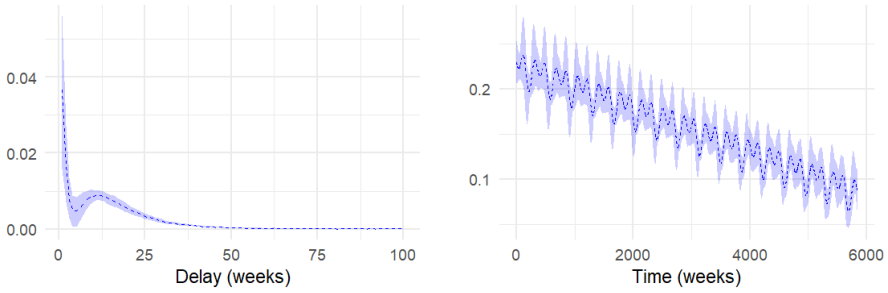
U.S. Terrorism

We analyze a 21-year time series of daily terrorist incidents within the United States, drawn from the Global Terrorism Database (GTD) (National Consortium for the Study of Terrorism and Responses to Terrorism (START), 2021). The data spans from 1972 onward. All events are aggregated to daily resolution and partitioned into training (1972–1981), validation (1982–1987), and test (1988–1992) intervals.

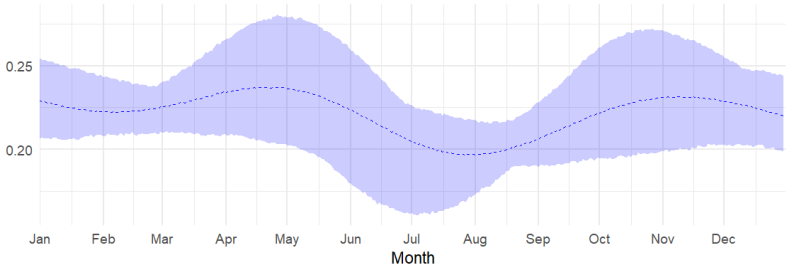
Terrorist activity over this period is characterized by long periods of inactivity interspersed with sudden bursts of incidents, often linked to domestic protest groups or separatist campaigns. This sparse and irregular structure poses a challenge for traditional autoregressive models, which lack the ability to explicitly model causal excitation between events.

Self-exciting models offer a principled alternative. Prior work by (Porter & White, 2012) applied a discrete Hawkes process with a negative binomial excitation kernel to terrorism data in Southeast Asia, demonstrating how such models can recover meaningful structure in burst-prone time series and yield interpretable quantities such as volatility and resilience.

In our application, GP-DHP achieves the highest predictive log-likelihood on the test set, outperforming all parametric baselines (Table 5.2). Notably, the linear DHP performs worse than the constant model, suggesting no benefit from including a long-term trend. The sinusoidal baseline yields a negligible improvement. In contrast, GP-DHP captures the underlying dynamics without imposing rigid structure, adapting flexibly to both long stretches of inactivity and short-term bursts. On top of this, the branching ratio for the excitation function was $\sum_{d\geq 1} \max\{\Phi(d), 0\} = 0.24$. Since the branching ratio is less than one, we get a stable process.



(a) MAP estimates under the GPDHP fit to the U.S. terrorism data. Left: excitation kernel $\Phi(d)$ as a function of lag in weeks. Right: baseline function $\mu(t)$ over time in weeks.



(b) MAP estimate of the seasonal baseline component $\mu(t)$ over one calendar year (January–December).

Figure 5.5: MAP estimates for the GPDHP fit to U.S. terrorism data, including a close-up view of the seasonal baseline component over one year (daily aggregation).

Cryptosporidiosis

We analyse a 365-week series of weekly Cryptosporidiosis case counts, available via the `liboschik2017jss` package in R (Liboschik, Fokianos, & Fried, 2017). Cryptosporidiosis is a gastrointestinal disease caused by the protozoan parasite *Cryptosporidium*, which is typically spread through ingestion of contaminated water. While low-level background transmission persists in many regions, the disease is most notable for its sporadic and localised outbreaks. One of the most severe documented events occurred in Milwaukee in 1993, where over 400,000 residents were affected due to a contaminated municipal water supply (Hunter et al., 2003; Yoder, Harral, & Beach, 2010). These outbreak dynamics are characterised by sharp, short-term increases in reported cases, followed by a rapid return to low endemic levels.

Such temporal patterns are challenging for classical autoregressive count models, such as Poisson or INGARCH-type models (Fokianos, Rahbek, & Tjøstheim, 2009; Fokianos & Tjøstheim, 2011), which assume dependence based on past counts but do not directly model the self-exciting nature of outbreaks. These models often struggle to distinguish between sustained background intensity and transient, event-driven clustering.

Model	pLL
Discrete DHP	-287.1
Linear DHP	-290.1
Sinusoidal DHP	-287.1
Linear + Sinusoidal DHP	-349.8
GPDHP	**-285.1**

Table 5.3: Predictive log-likelihood (pLL) on the Cryptosporidiosis test set.

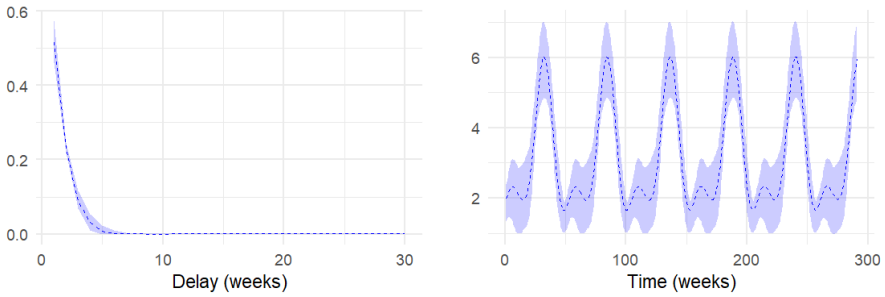


Figure 5.6: MAP estimates under the GPDHP fit to the weekly Cryptosporidiosis data. Left: excitation kernel $\Phi(d)$ as a function of lag in weeks. Right: baseline function $\mu(t)$ over time in weeks.

In contrast, GPDHP explicitly decomposes the latent event rate into two components: a baseline function $\mu(t)$ and a self-exciting kernel $\Phi(d)$. This separation provides epidemiologically meaningful insight: the baseline reflects ongoing exposure or environmental risk, while the excitation term captures outbreak-driven propagation.

Recent work by (Browning et al., 2021) demonstrated the use of discrete Hawkes models with geometric excitation kernels for modelling daily COVID-19 mortality in multiple countries. The geometric kernel is a special case of the negative-binomial excitation used here, namely when $r = 1$, and has proven effective for modelling short-memory self-excitation in epidemic settings. By generalising to a full negative-binomial kernel, GPDHP can adapt more flexibly to varying outbreak shapes and durations. For this dataset, the fitted branching-ratio diagnostic was

$$\widehat{\kappa} = \sum_{d=1}^{D_{\max}} \max\{\widehat{\Phi}(d), 0\} = 0.86.$$

Since this value is less than one, the fitted excitation is consistent with a stable process.

On this dataset, GPDHP achieves a modest improvement in predictive performance, see Table 5.3, compared to the larger gains on terrorism shown in Table 5.2, and produces interpretable decompositions of the observed dynamics in Figure 5.6. The method therefore not only improves predictive accuracy, but also provides a principled way to distinguish irregular spikes from sustained background trends, which is useful for outbreak monitoring and forecasting in public-health surveillance.

GPDHP attains the highest pLL, narrowly outperforming all parametric baselines. While the improvements are modest, this reflects the benefit of data-driven structure when temporal dynamics are weak or irregular.

5.5 Conclusion

We introduced GP-DHP, a scalable nonparametric model for discrete-time Hawkes processes that learns both baseline and excitation directly from binned count data. By collapsing the GP priors into a single latent process, GP-DHP admits efficient MAP optimization with $O(T \log T)$ cost and then uses a closed-form projection to recover interpretable components. Across controlled simulations and two applied datasets, GP-DHP consistently matches or exceeds the predictive performance of parametric baselines while yielding diagnostics and summaries that practitioners expect, including excitation strength, effective memory, and seasonal background behavior.

Practically, the model is attractive when baseline structure is unknown or nonstationary and when excitation is unlikely to follow a fixed parametric shape. The collapsed formulation makes the method usable at the time scales typical of surveillance and monitoring data, and the decomposition clarifies exogenous versus endogenous drivers for domain interpretation.

Limitations include reliance on MAP rather than full Bayesian inference and the current focus on univariate series. Future work includes multivariate and spatial extensions for interacting series, and data-driven kernel selection to adapt automatically to domain-specific seasonality and memory.

Chapter 6

Conclusion

This thesis investigated modelling, inference, and forecasting for discrete time Hawkes processes, with equal emphasis on statistical clarity and computational usability. The central theme was to build models that captured clustered dynamics and contextual effects while remaining simple enough to estimate, simulate, and deploy. On the modelling side, we introduced a multivariate specification that separated a structured baseline from self- and cross-excitation, included a day–night regime, and incorporated alarm indicators as distinct channels of influence. On the computational side, we derived recursive updates that removed repeated sums over the full past—a device used in continuous-time Hawkes processes but, until now, not utilised in discrete time—so that intensities and their components could be evaluated in constant time at each step. Together, these ideas allowed the analyst to move from a description of bursts and spillovers to timely forecasts and decision support.

After reviewing background on Bayesian parameter estimation, MAP optimisation, kernels and Gaussian processes, and point process notation, the thesis first discussed the discrete time representation that rewrote the usual intensity expansions in terms of running states. The result was a set of first-order recursions that carried forward weighted contributions for each component and covariate. In place of a cost that grew with the length of history, each update became a fixed set of arithmetic operations. This change in viewpoint was small in algebra but large in practice: likelihood evaluation, gradient-based optimisation, and forward simulation became straightforward and could be repeated many times for model selection, cross-validation, and uncertainty assessment.

The thesis then built an application-driven model for incident monitoring that we referred to as the 12D HPA specification. Here the baseline accounted for predictable variation in time, including diurnal structure, and the excitation terms were split into self and cross effects. Alarm events were treated as marked triggers that could excite several series at once. The separation between ordinary and alarm-driven excitation was important for interpretation and for interventions. The empirical analysis showed that ordinary events tended to reinforce activity within the same series, while alarms produced wider, hospital-level responses. Forecasts that conditioned on recent history and on the presence or absence of alarms were sharper

and better calibrated than those from simpler competitors such as inhomogeneous Poisson models or unmarked multivariate Hawkes models. The patterns uncovered by the estimated gains were easily visualised and aligned with operational knowledge, which supported the usefulness of the specification beyond purely statistical measures of fit.

To complement this structured model, the thesis proposed a nonparametric framework in which the baseline and the excitation shape were treated as functions to be learned. Gaussian process priors provided smoothness and allowed the data to determine the degree of persistence and the location of important lags. Inference was carried out with a MAP approach that kept computation tractable while retaining the benefits of flexible function learning. On synthetic data with several excitation shapes, and on real count data with seasonality and bursts, this approach recovered interpretable components and produced competitive or better predictive performance. The function-space view also clarified the connection with kernels, since the choice of covariance encoded the beliefs about smoothness, periodicity, and decay appropriate for the application at hand.

A first avenue for future research is to place the nonparametric model on a fully Bayesian footing and develop scalable posterior inference. While the MAP approach proved effective, a full Bayesian treatment would yield calibrated uncertainty for both the baseline and the excitation functions, as well as for forecasts and functionals such as reproduction numbers and branching ratios. Variational approximations and Laplace methods around sparse Gaussian-process representations offer practical starting points, and could be compared against sequential Monte Carlo or MCMC schemes that exploit the recursive state construction. Particular attention should be paid to uncertainty decomposition across baseline, self-excitation, and cross-excitation so that credible intervals remain interpretable to practitioners.

A second line of enquiry is a multivariate extension of the nonparametric framework introduced in Section 5. Multi-output Gaussian processes could endow baselines and excitation shapes with cross-series covariance, allowing shared structure and partial pooling across related streams. Coregionalisation (e.g., linear model of coregionalisation) and low-rank kernels would control complexity while learning which pairs of series share lags or seasonal effects. This extension would directly model cross-excitation functions as latent surfaces over lag and source–target indices, permitting richer, data-driven interaction patterns than fixed parametric forms.

A closely related priority is to deploy the nonparametric framework on hospital data at scale. Hospital operations exhibit strong diurnal, weekly, and holiday patterns, intermittent bursts tied to alarms, and unit-level heterogeneity. A tailored analysis could quantify how alarm types propagate across wards, estimate delay distributions from trigger to response, and

separate routine circadian variation from true cascades. Embedding the model within an operational dashboard would allow forecasters to condition on observed or hypothetical alarm sequences and to translate predictions into staffing and bed-management decisions with explicit uncertainty.

Beyond hospital-specific deployments, spatio-temporal sharing across sites and services is promising. Hierarchical priors could link hospitals within a trust, wards within a hospital, or ambulance stations within a region, enabling information borrowing where data are sparse while preserving local idiosyncrasies. Graph-structured Hawkes components, with edges learned from data or constrained by known referral pathways, would capture how events propagate across the organisational network. Time-varying adjacency matrices could represent changing connectivity during major incidents or seasonal pressures.

Another important strand is online and streaming inference. Warm-started optimisation and incremental state updates already suggested how to refresh forecasts as new counts arrived; future work could formalise this into amortised inference procedures that update parameters and hyperparameters in near real time. Techniques from stochastic variational inference and forgetting factors would help track concept drift, while change-point models layered atop the recursions could detect regime shifts promptly. Rigorous prequential evaluation would ensure that improvements translate to better short-horizon performance under operational constraints.

Handling nonstationarity more broadly would also be valuable. While the day–night regime captured coarse shifts, many applications exhibit gradual changes in baseline load, evolving excitation strength, or seasonally modulated lags. Time-varying kernels—implemented via nonstationary Gaussian-process priors, neural kernel parameterisations, or spline-based state evolution—could allow excitation shapes to adapt over months or years. Coupling such flexibility with stability constraints would prevent explosive behaviour while acknowledging real-world drift.

Robustness is another priority. Count streams often contain outliers, reporting delays, and periods of saturation where additional events inhibit rather than excite further activity. Extensions with heavy-tailed observation models, censoring-aware likelihoods, and inhibitory or saturating kernels would better reflect such realities. Constraints or priors on the effective branching ratio would help maintain subcritical dynamics, and diagnostic tools could flag when the model veered towards unstable regimes.

Model assessment and interpretability can be strengthened further. Discrete-time adaptations of thinning-based residual checks, calibration curves for probabilistic counts, and Probability Integral Transform (PIT) diagnostics would complement likelihood-based comparisons. Structured attribution—decomposing forecasts into baseline, self-, and cross-excitation contributions over lag—could be formalised as part of routine reporting, aiding accountability and trust. For multi-series settings, pairwise influence maps over lag would provide a compact, actionable summary of learned interactions.

Scaling and engineering considerations will also matter for impact. Implementations that exploit sparse linear algebra and automatic differentiation, together with GPU acceleration for batched recursions, would increase throughput for large multivariate systems. Packaging the methods as an open-source library with reproducible workflows, standardised data schemas, and carefully curated benchmarks would facilitate adoption and fair comparison. Synthetic data generators with controllable ground truth could support stress testing across regimes.

Privacy-preserving and federated learning approaches are particularly relevant for sensitive domains such as healthcare. Differentially private training for the recursive likelihood, coupled with secure aggregation across sites, would enable cross-hospital learning without sharing raw events. Aggregation-aware likelihoods that handle varying time bins, delayed reporting, and partial censoring would make the methodology usable under realistic data governance constraints.

Theoretical developments round out the agenda. Identifiability conditions for discrete time Hawkes models with function-valued components, posterior consistency for Gaussian-process-driven intensities, and rates of contraction under misspecification are all open questions. Understanding discretisation error relative to continuous-time Hawkes limits, and characterising when constant-time recursions approximate their continuous counterparts, would clarify the trade-offs that motivate the discrete representation. Such results would underpin principled guidance on bin widths, priors, and regularisation.

Finally, integrating decision models with forecasting remains a fertile area. Coupling the intensity forecasts to queueing approximations for patient flow, or to sequential decision frameworks for staffing and diversion, would close the loop from prediction to action. Optimising over policies using forecast-aware objectives—while accounting for uncertainty and intervention feedback through the excitation structure—could yield measurable performance gains and provide a template for decision support in other bursty count domains.

In conclusion, this thesis showed that discrete time Hawkes models can be both expressive and practical. By combining interpretable structure with constant-time recursions, the models scaled to long horizons and many interacting series without sacrificing the ability to explain why events clustered and how context shaped risk. The 12D HPA specification offered a clear

operational narrative for incident data, while the nonparametric framework learned baseline and excitation functions that adapted to the data and improved forecasts. The resulting toolkit supported simulation, estimation, and forecasting in a way that was transparent to practitioners and faithful to the dynamics of bursty count data.

Bibliography

Adams, R. P., Murray, I., & MacKay, D. J. C. (2009). Tractable nonparametric bayesian inference in poisson processes with gaussian process intensities. In *Proceedings of the 26th international conference on machine learning (icml)* (pp. 9–16).

Barnard, G. W., Robbins, L., Newman, G., & Carrera, F. (1984). A study of violence within a forensic treatment facility. *Bulletin of the American Academy of Psychiatry and the Law*, *12*(4), 339–348.

Beaumont, M. A., Zhang, W., & Balding, D. J. (2002). Approximate bayesian computation in population genetics. *Genetics*, *162*(4), 2025–2035.

Beck, N. C., Tubbesing, T., Lewey, J. H., Ji, P., Menditto, A. A., & Robbins, S. B. (2018). Contagion of violence and self-harm behaviors on a psychiatric ward. *The Journal of Forensic Psychiatry & Psychology*, *29*(6), 989–1006.

Brémaud, P., & Massoulié, L. (1996). Stability of nonlinear hawkes processes. *The Annals of Probability*, *24*(3), 1563–1588.

Brown, E. N., Barbieri, R., Ventura, V., Kass, R. E., & Frank, L. M. (2002). The time-rescaling theorem and its application to neural spike train data. *Neural Computation*, *14*(2), 325–346.

Browning, R., Rousseau, J., & Mengersen, K. (2022). A flexible, random histogram kernel for discrete-time hawkes processes. *arXiv preprint arXiv:2208.02921*.

Browning, R., Sulem, D., Mengersen, K., Rivoirard, V., & Rousseau, J. (2021). Simple discrete-time self-exciting models can describe complex dynamic processes: A case study of covid-19. *PLOS ONE*, *16*(4), e0250015. doi: 10.1371/journal.pone.0250015

Cinlar, E. (2013). *Introduction to stochastic processes*. Courier Corporation.

Cox, D. R., & Isham, V. (1980). *Point processes*. Chapman and Hall.

Daley, D. J., & Vere-Jones, D. (2003). *An introduction to the theory of point processes, volume i: Elementary theory and methods* (2nd ed.). Springer.

Desmarais, S. L., Nicholls, T. L., Wilson, C. M., & Brink, J. (2012). Using dynamic risk and protective factors to predict inpatient aggression: Reliability and validity of start assessments. *Psychological Assessment*, *24*(3), 685.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, *1*(2), 209–230.

Fokianos, K., Rahbek, A., & Tjøstheim, D. (2009). Poisson autoregression. *Journal of the American Statistical Association*, *104*(488), 1430–1439.

Fokianos, K., & Tjøstheim, D. (2011). Log-linear poisson autoregression. *Journal of Multivariate Analysis*, *102*(3), 563–578.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*(410), 398–409.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, *57*(1), 97–109.

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, *58*(1), 83–90.

Hawkes, A. G., & Oakes, D. (1974). A cluster process representation of a self-exciting process. *Journal of Applied Probability*, *11*(3), 493–503.

Hinsby, K., & Baker, M. (2004). Patient and nurse accounts of violent incidents in a medium secure unit. *Journal of Psychiatric and Mental Health Nursing*, *11*(3), 341–347.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Hunter, P. R., Chalmers, R. M., Syed, Q., Hughes, L. S., Woodhouse, S., & Swift, L. (2003). Foot and mouth disease and cryptosporidiosis: Possible interaction between two emerging infectious diseases. *Emerging Infectious Diseases*, *9*(1), 109–112.

Kingman, J. F. C. (1993). *Poisson processes*. Clarendon Press.

Kirchner, M. (2016). Hawkes and inar(∞) processes. *Stochastic Processes and their Applications*, *126*(8), 2494–2525.

Lewis, E., & Mohler, G. (2011). *A nonparametric em algorithm for multiscale hawkes processes.*

Liboschik, T., Fokianos, K., & Fried, R. (2017). tscount: An R package for analysis of count time series following generalized linear models. *Journal of Statistical Software*, *82*(5), 1–51. doi: 10.18637/jss.v082.i05

Liu, J. S., & Chen, R. (1998). Sequential monte carlo methods for dynamic systems. *Journal of the American Statistical Association*, *93*(443), 1032–1044.

Lloyd, C., Gunter, T., Osborne, M., & Roberts, S. (2015). Variational inference for gaussian process modulated poisson processes. In *Proceedings of the 32nd international conference on machine learning (icml)* (pp. 1814–1822).

Malem-Shinitski, N., Ojeda, C., & Opper, M. (2022). Variational bayesian inference for nonlinear hawkes process with gaussian process self-effects. *Entropy*, *24*(3), 356. doi: 10.3390/e24030356

Meehan, T., McIntosh, W., & Bergen, H. (2006). Aggressive behaviour in the high-secure forensic setting: The perceptions of patients. *Journal of Psychiatric and Mental Health Nursing*, *13*(1), 19–25.

Mohler, G., Short, M., Brantingham, P. J., Schoenberg, F., & Tita, G. (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, *106*(493), 100–108.

Møller, J., Syversveen, A. R., & Waagepetersen, R. P. (1998). Log gaussian cox processes. *Scandinavian Journal of Statistics*, *25*(3), 451–482.

Mukhopadhyay, S., & Sathish, V. (2018). Predictive likelihood for coherent forecasting of count time series. *Journal of Forecasting*, *38*(3), 222–235.

National Consortium for the Study of Terrorism and Responses to Terrorism (START). (2021). *Global terrorism database (1970–2020).* `https://www.start.umd.edu/gtd/`.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, *83*(401), 9–27.

Ozaki, T. (1979). Maximum likelihood estimation of hawkes' self-exciting point processes. *Annals of the Institute of Statistical Mathematics*, *31*(1), 145–155.

Peluola, A., Mela, M., & Adelugba, O. (2012). A review of violent incidents in a multilevel secure forensic psychiatric hospital: Is there a seasonal variation? *Medicine, Science and the Law*, *53*(2), 72–79.

Porter, M. D., & White, G. (2012). Self-exciting hurdle models for terrorist activity. *The Annals of Applied Statistics*, *6*(1), 106–124.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press.

Rasmussen, J. G. (2011). Temporal point processes: The conditional intensity function. *Lecture Notes, University of Copenhagen*. (Tutorial reference)

Reinhart, A. (2018). A review of self-exciting spatio-temporal point processes and their applications. *Statistical Science*, *33*(3), 299–318.

Roberts, G. O., & Tweedie, R. L. (1996). Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, *2*(4), 341–363.

Schoenberg, F. P. (2016). A note on the consistent estimation of the etas model. *Journal of Seismology*, *20*(5), 1245–1256.

Snyder, D. L., & Miller, M. I. (1991). *Random point processes in time* (2nd ed.). Springer-Verlag.

Stavropoulou, C., Doherty, C., & Tosey, P. (2015). How effective are incident-reporting systems for improving patient safety? a systematic literature review. *BMJ Quality & Safety*, *24*(4), 826–836.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, *58*(1), 267–288.

Veen, A., & Schoenberg, F. P. (2008). Estimation of space–time branching process models for earthquake occurrences. *The Annals of Applied Statistics*, *2*(1), 124–149.

White, G., Porter, M. D., & Mazerolle, L. (2013). Terrorism risk, resilience and volatility: A comparison of terrorism patterns in three southeast asian countries. *Journal of Quantitative Criminology*, *29*(2), 295–320.

Wright, K. M., Duxbury, J. A., Baker, A., & Crumpton, A. (2014). A qualitative study into the attitudes of patients and staff towards violence and aggression in a high security hospital. *Journal of Psychiatric and Mental Health Nursing*, *21*(2), 184–188.

Xia, J. (2007). A fast algorithm for toeplitz matrix–vector multiplication using the discrete fourier transform. *SIAM Journal on Matrix Analysis and Applications*, *29*(3), 843–860. doi: 10.1137/050642997

Xu, G., Wang, M., Bian, J., Huang, H., Burch, T. R., Andrade, S. C., ... Guan, Y. (2020). Semi-parametric learning of structured temporal point processes. *Journal of Machine Learning Research*, *21*(1), 7851–7889.

Yoder, J. S., Harral, C., & Beach, M. J. (2010). Cryptosporidiosis surveillance — united states, 2006–2008. *MMWR Surveillance Summaries*, *59*(6), 1–14.

Zhang, J., Cai, B., Zhu, X., Wang, H., Xu, G., & Guan, Y. (2023). Learning human activity patterns using clustered point processes with active and inactive states. *Journal of Business & Economic Statistics*, *41*(2), 388–398.

Zhang, R., Walder, C., & Tanaka, M. (2019). Gaussian process hawkes processes. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(1), 6828–6835.

Zhou, K., Zha, H., & Song, L. (2013). Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *Proceedings of aistats* (pp. 641–649).

Appendix A

Proofs for Chapter 3

This appendix collects detailed proofs for the discrete-time results in Chapter 3. Throughout we use the notation introduced in Section 3.2. We write $g_{l,m}(d) = \beta_{l,m}(1-\beta_{l,m})^{d-1}$ and $G_{l,m}(u) = \sum_{s=1}^{u} g_{l,m}(s) = 1-(1-\beta_{l,m})^{u}$. Sums are finite and all terms are nonnegative, so Tonelli's theorem justifies interchanging the order of summation. Differentiation under the summation signs is justified by dominated convergence because $0 < \beta_{l,m} < 1$ implies absolute summability of the geometric tails and $\lambda^{(m)}(t) > 0$.

Proof of Proposition 3.3.1 (event-time likelihood identity)

Proof. Start from the Poisson log-likelihood

$$\log L(\theta \mid \tau) = \sum_{t=1}^{N}\sum_{m=1}^{M}\Big(Y_t^{(m)}\log\lambda^{(m)}(t) - \lambda^{(m)}(t)\Big).$$

The data term equals $\sum_m \sum_{t\in\tau} Y_t^{(m)} \log\lambda^{(m)}(t)$ because $Y_t^{(m)} = 0$ when $t \notin \tau$. For the compensator, substitute $\lambda^{(m)}(t) = \mu^{(m)} + \sum_l K_{l,m}\sum_{i:\,t_i\le t-1} Y_{t_i}^{(l)} g_{l,m}(t-t_i)$ to obtain

$$\sum_{t=1}^{N}\lambda^{(m)}(t) = N\mu^{(m)} + \sum_{t=1}^{N}\sum_{l=1}^{M} K_{l,m}\sum_{i:\,t_i\le t-1} Y_{t_i}^{(l)} g_{l,m}(t-t_i).$$

Interchange the sums over t and i (Tonelli) and change variables $u = t - t_i \in \{1,\dots,N-t_i\}$:

$$\sum_{t=1}^{N}\sum_{i:\,t_i\le t-1} Y_{t_i}^{(l)} g_{l,m}(t-t_i) = \sum_{i=1}^{N_{\text{events}}} Y_{t_i}^{(l)} \sum_{u=1}^{N-t_i} g_{l,m}(u) = \sum_{t\in\tau} Y_t^{(l)} G_{l,m}(N-t).$$

Collecting terms across m yields (3.3). □

Proof of Proposition 3.4.1 (event-time recursion)

Proof. Fix $j \in \{1,\dots,N_{\text{events}}-1\}$ and (l,m). By definition,

$$R(j{+}1,l,m) = \sum_{i:t_i<t_{j+1}} Y_{t_i}^{(l)} \beta_{l,m}(1-\beta_{l,m})^{t_{j+1}-t_i-1}.$$

Split at t_j:

$$R(j{+}1,l,m) = \sum_{i:t_i<t_j} Y_{t_i}^{(l)} \beta_{l,m}(1-\beta_{l,m})^{(t_j-t_i-1)+\Delta_j} \;+\; Y_{t_j}^{(l)} \beta_{l,m}(1-\beta_{l,m})^{\Delta_j-1}.$$

Factor $(1-\beta_{l,m})^{\Delta_j}$ out of the first sum to obtain

$$R(j{+}1,l,m) = (1-\beta_{l,m})^{\Delta_j} \sum_{i:t_i<t_j} Y_{t_i}^{(l)} \beta_{l,m}(1-\beta_{l,m})^{t_j-t_i-1} + Y_{t_j}^{(l)} \beta_{l,m}(1-\beta_{l,m})^{\Delta_j-1}.$$

The bracketed sum is $R(j,l,m)$, which proves (3.5). For (3.6), substitute $\Phi_{l,m}(d) = K_{l,m} g_{l,m}(d)$ into (3.1) at $t = t_j$ and recognise $R(j,l,m)$. □

Proof of Lemma 3.5.1 (primitive score formulas)

Proof. Differentiate (3.3). For $\mu^{(p)}$, $\partial\lambda^{(p)}(t)/\partial\mu^{(p)} = 1$ yields

$$\frac{\partial}{\partial\mu^{(p)}} \log L = \sum_{t\in\tau} Y_t^{(p)} \frac{1}{\lambda^{(p)}(t)} \cdot 1 - N = \sum_{t\in\tau} \frac{Y_t^{(p)}}{\lambda^{(p)}(t)} - N.$$

For $K_{p,q}$, only $\lambda^{(q)}(t)$ depends on $K_{p,q}$, with

$$\frac{\partial\lambda^{(q)}(t)}{\partial K_{p,q}} = \sum_{i:t_i<t} Y_{t_i}^{(p)} g_{p,q}(t-t_i).$$

The compensator term contributes $-\sum_{t\in\tau} Y_t^{(p)} G_{p,q}(N-t)$. Putting these together gives (3.8). For $\beta_{p,q}$, use

$$\frac{\partial}{\partial\beta_{p,q}} g_{p,q}(u) = (1-\beta_{p,q})^{u-2}\big(1-\beta_{p,q}u\big), \qquad \frac{\partial}{\partial\beta_{p,q}} G_{p,q}(u) = u(1-\beta_{p,q})^{u-1},$$

to obtain (3.9). Differentiation under the sums is justified because geometric tails are absolutely summable and $\lambda^{(m)}(t) > 0$. □

Proof of Proposition 3.5.2 (fast updates for R^1 and R^2)

Proof. Write $R^1(j{+}1,l,m) = \sum_{i:t_i<t_{j+1}} Y_{t_i}^{(l)}(1-\beta)^{t_{j+1}-t_i-2}$ with $\beta = \beta_{l,m}$. Split at t_j and note $t_{j+1}-t_i-2 = (t_j-t_i-2)+\Delta_j$ for $t_i < t_j$, then factor $(1-\beta)^{\Delta_j}$. The $t_i = t_j$ term contributes $(1-\beta)^{\Delta_j-2}Y_{t_j}^{(l)}$. This yields (3.11).

For R^2,

$$R^2(j{+}1,l,m) = \sum_{i:t_i<t_{j+1}} Y_{t_i}^{(l)}(1-\beta)^{t_{j+1}-t_i-2}\beta(t_{j+1}-t_i).$$

For $t_i < t_j$, write $t_{j+1}-t_i = (t_j-t_i)+\Delta_j$ and expand

$$\beta(t_{j+1}-t_i)(1-\beta)^{t_{j+1}-t_i-2} = \beta(t_j-t_i)(1-\beta)^{t_j-t_i-2}(1-\beta)^{\Delta_j} + \beta\Delta_j(1-\beta)^{t_j-t_i-2}(1-\beta)^{\Delta_j}.$$

Summing over $t_i < t_j$ gives $(1-\beta)^{\Delta_j}R^2(j,l,m) + \beta\Delta_j(1-\beta)^{\Delta_j}R^1(j,l,m)$. The $t_i = t_j$ term equals $\beta\Delta_j(1-\beta)^{\Delta_j-2}Y_{t_j}^{(l)}$. Recognising $R^1(j{+}1,l,m) = (1-\beta)^{\Delta_j}R^1(j,l,m) + Y_{t_j}^{(l)}(1-\beta)^{\Delta_j-2}$ yields (3.12). □

Proof of fast score formulas (3.13)–(3.15)

Proof. For $\partial \log L/\partial \mu^{(p)}$, rewrite $\lambda^{(p)}(t_j) = \mu^{(p)} + \sum_l K_{l,p}R(j,l,p)$ using (3.6). Substituting into (3.7) gives (3.13).

For $\partial \log L/\partial K_{p,q}$, in (3.8) the data term contains

$$\sum_{i:t_i<t_j} Y_{t_i}^{(p)} g_{p,q}(t_j-t_i) = \sum_{i:t_i<t_j} Y_{t_i}^{(p)}\beta_{p,q}(1-\beta_{p,q})^{t_j-t_i-1} = R(j,p,q),$$

so the numerator at event time t_j is $Y_{t_j}^{(q)}R(j,p,q)$ and the denominator is $\lambda^{(q)}(t_j) = \mu^{(q)} + \sum_l K_{l,q}R(j,l,q)$. The compensator term is already in closed form, which yields (3.14).

For $\partial \log L/\partial \beta_{p,q}$, differentiate g and G. The data term involves $\sum_{i:t_i<t_j} Y_{t_i}^{(p)}\partial g_{p,q}(t_j-t_i)/\partial\beta_{p,q}$. Using

$$\frac{\partial}{\partial\beta}g(u) = (1-\beta)^{u-2}\big(1-\beta u\big) = (1-\beta)^{u-2} - \beta u(1-\beta)^{u-2},$$

the first part sums to $R^1(j,p,q)$ and the second part sums to $R^2(j,p,q)$. Multiplying by $K_{p,q}$ and dividing by $\lambda^{(q)}(t_j)$ yields the first line of (3.15). The compensator derivative gives the second line. This proves (3.15). □

Notes on regularity and complexity. All sums are finite on a fixed window and the geometric tails ensure absolute summability of lag contributions. Baselines $\mu^{(m)}$ are positive, so $\lambda^{(m)}(t) > 0$. The event-time likelihood in Algorithm **??** takes $O(M^2 N_{\text{events}})$ time and $O(M^2)$ memory. The multi-step forecasting in Algorithm 2 takes $O(M^2 h)$ time and $O(M^2)$ memory.

Appendix B

Chapter 3 Appendix

B.1 Simulation Study

We present the results of simulations conducted using the 12D-HPA model. These simulations were performed over the same time period as the training data, utilizing parameters estimated from fitting the model to the training data. Furthermore, the hyperparameter values were set based on those determined during the cross-validation procedure.

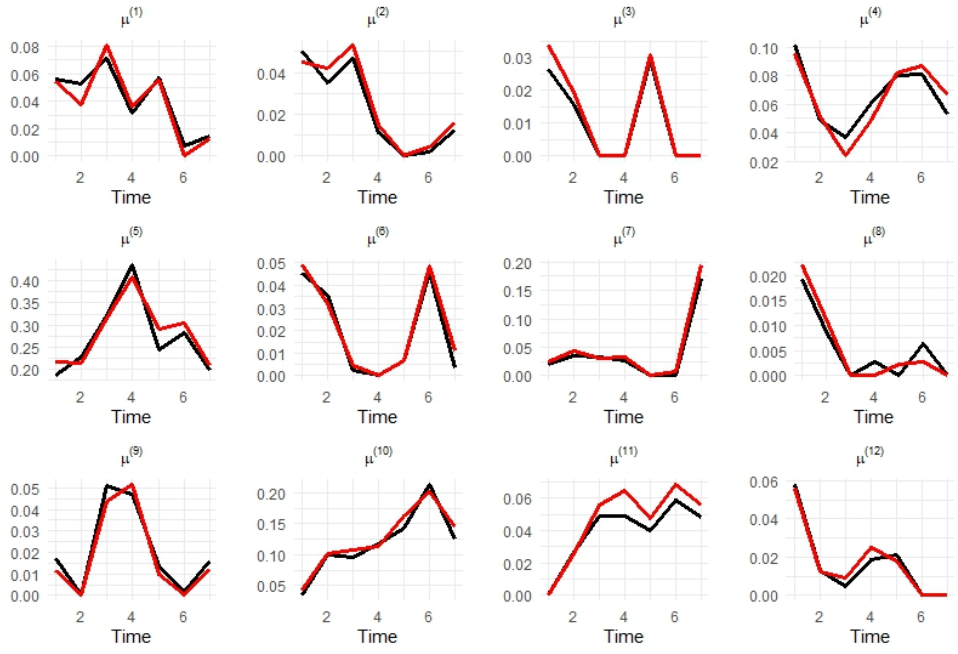


Figure B.1: Plots of the baseline proportionality constants $\mu^{(m)}$ for each ward $m = 1, \ldots, M$ over the seven-year period of the dataset. Red lines represents the true baseline proportionality constants, while the black line represents the estimates background proportionality constants.

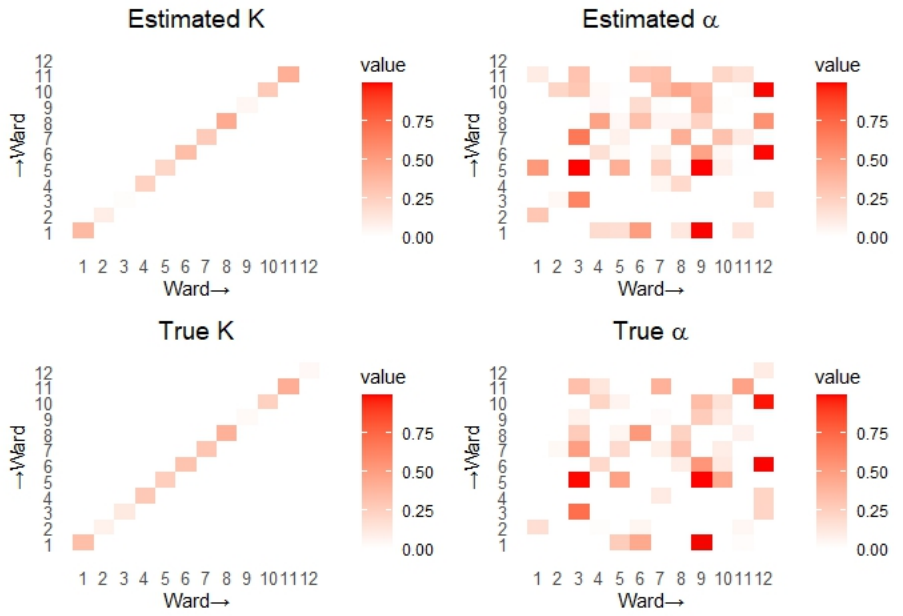


Figure B.2: Heatmaps of the excitation matrices K and α under the 12D-HPA model: Comparison of Estimated and True Values.

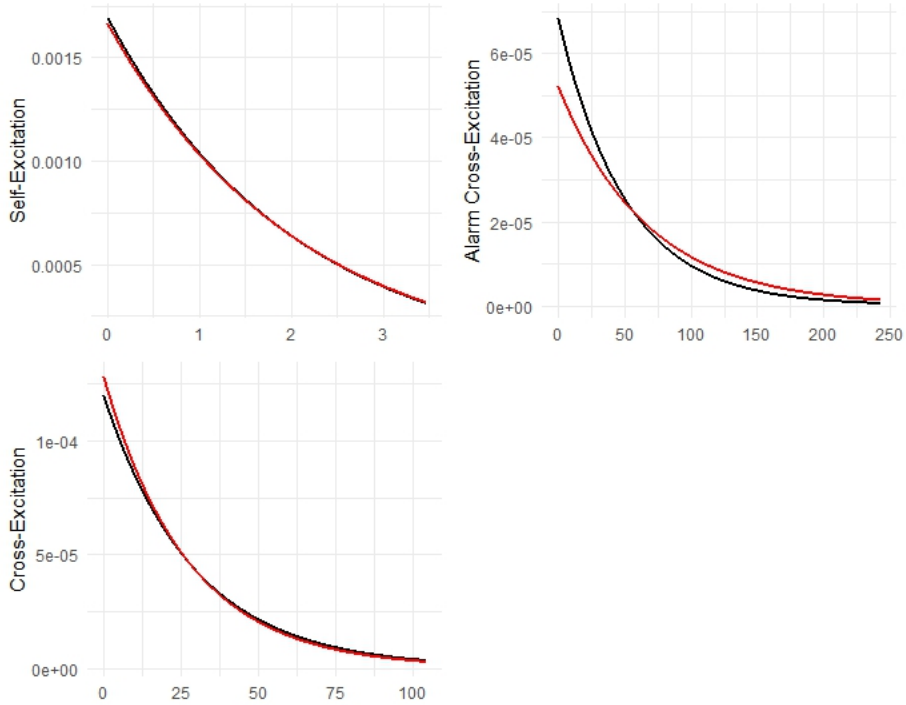


Figure B.3: Plots of excitation functions for events triggering alarms across different wards, events triggering alarms within the same ward, and non-alarm events within the same ward. The red line represents the true excitation function, while the black line represents the estimated excitation function.

True parameter	Estimated parameter
$K_n = 0$	$\hat{K}_n = 0$
$\alpha_n = 1$	$\hat{\alpha}_n = 1$

Table B.1: Comparison of true and estimated parameter values for K_n and α_n.

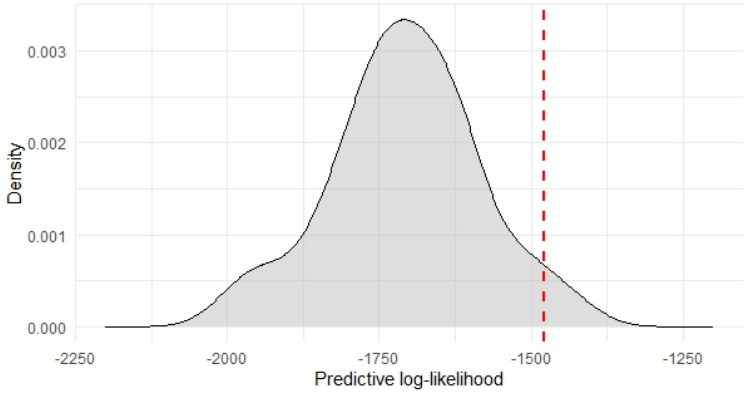


Figure B.4: Density plot of predictive log-likelihoods from 50 forecasts over the test set interval. The red dashed line indicates the log-likelihood calculated on the actual test dataset. We showed that the predictive log-likelihood value calculated on the observed test data falls within the range of values we found on the simulations. That is, the true value is not a significant outlier, so we can not reject the possibility that the model is correct. If the model was incorrect then we would expect to see a low predictive log-likelihood for the observed test set compared to the simulated test data, which is not what we see here.

B.2 Poisson Assumption Motivation

The assumption that event counts follow a Poisson distribution, given the past, can be justified through an analogy to previous work on modeling deaths as a result of infections in Browning et al. (2021). We present this now. In our setting, violent events can be thought of as analogous to infections, where each initial event has the potential to "trigger" subsequent events.

In detail, let us denote the number of new violent events at time t as Y_t. Given the past history, Y_t follows a Poisson distribution with rate parameter λ_t, where λ_t is influenced by a baseline rate of occurrence and the excitation effects from past events. Specifically, we have:

$$Y_t \sim \text{Poisson}(\lambda_t), \quad \lambda_t = \mu + \alpha \sum_{k>0} Y_{t-k}(1-\beta)^{k-1},$$

where:

- μ is the baseline rate of violent events,

- α is the parameter representing the excitation effect of past events,

- β controls the decay of influence over time.

This setup is similar to how the number of deaths was approximated using a Poisson distribution in the context of infections. Consider that each violent event may trigger subsequent violent incidents with a small probability, similar to how infections may lead to deaths. If we assume that there are Y_{t-k} violent events that occur at a previous time $t-k$, and each of these events has a small probability p of triggering a subsequent event, the number of new violent events at time t could be modeled as:

$$Y_t \sim \text{Binomial}(Y_{t-k}, p).$$

When the probability p is small and the number of previous events Y_{t-k} is large, the binomial distribution can be well approximated by a Poisson distribution with parameter pY_{t-k}. Thus, the rate parameter λ_t for the number of new events can be approximated as:

$$\lambda_t = p\left(\mu + \alpha \sum_{k>0} Y_{t-k}(1-\beta)^{k-1}\right).$$

This approximation allows us to use the Poisson assumption to model the counts of violent events while preserving computational efficiency. Moreover, this assumption aligns with the contagion-like dynamics of violent incidents, where each event may influence future events but the individual triggering probability remains small. Furthermore, given that our data are

collected in discrete time intervals, the Poisson assumption provides a reasonable and tractable way to model the number of events occurring within each interval. If the resolution of the data were fine enough to capture the exact ordering of events, we could utilize the more general continuous-time Hawkes process.