Statistical learning of Hawkes models and market microstructure

Saad Labyad
St John's College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Hilary 2024

This thesis is dedicated to my parents, Jamila and Abderrahim, and to my sister Salma, to whom I owe more than words can express.

Lorsqu'ils voient un figuier donner des fruits en hiver,
les montagnards de Kabylie disent qu'il a rêvé du printemps.

- Fellag, *Le dernier chameau.*

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Professor Samuel Cohen and Professor Álvaro Cartea, for their help and encouragement. Thank you for many fruitful discussions and useful comments, and for the opportunities you offered me at the Mathematical Institute, New College, the Oxford–Man Institute, and the Alan Turing Institute. I would also like to acknowledge the crucial support of St John's college, under the Ioan and Rosemary James scholarship, and of my college advisor, Professor Jan Obłój, whose advice and help allowed me to start my DPhil in St John's in great conditions. Finally, I would like to thank my friend Pierre Banquy for his practitioner's opinion in several stimulating discussions on algorithmic trading and the hedge fund industry, and for his help with the earlier versions of the `ASLSD` implementation.

Abstract

Multivariate Hawkes processes (MHP) are a fundamental class of point processes with self-excitation. When estimating parameters for these processes, a difficulty is that the two main error functionals, the log-likelihood and the least squares error (LSE), as well as the evaluation of their gradients, have a quadratic complexity in the number of observed events. In practice, this prohibits the use of exact gradient-based algorithms for parameter estimation in many settings. Furthermore, MHP models are not designed for non-stationary training data, and they cannot incorporate event information besides their timestamps: we introduce the marked time-dependent linear Hawkes (MTLH) model to overcome these limitations. We construct an adaptive stratified sampling estimator of the gradient of the LSE of Hawkes models. This results in the `ASLSD` algorithm, a fast parametric estimation method for MHP and MTLH with general kernels, applicable to large datasets, which compares favourably with existing methods. We evaluate our algorithm on synthetic and real-world data.

We use the `ASLSD` algorithm to model high-frequency mid-price movements in the Nasdaq equities market using Hawkes models with multi-modal kernels, time varying baselines and multidimensional continuous marks. This approach allows us to capture the different frequencies of excitation of price movements, and the non-Markovian, non-stationary nature of price changes, while getting a better fit to market data. We leverage the branching representation of fitted models to build a counterfactual price impact model, and to understand exogenous price movements.

Keywords: Hawkes processes; stochastic gradient descent; point processes; Monte Carlo methods; adaptive stratified sampling, market microstructure, price impact.

MSC: 60G55,62M09,90C52,93E10, 91B28.

Contents

Notation

Integers Denote by $\mathbb{N}^*$ the set of strictly positive integers. For positive integers $n < m$, denote by $[\![n, m]\!]$ the set of integers from n to m, and by $[n]$ the set of integers from 1 to n.

Indexing To simplify notation, we use variable names i, j, k to refer to event types, and variable names m, n to refer to event indices such that the n-th event chronologically precedes the m-th event. We use variable names t, s to refer to event timings with $s < t$; and we use variable names τ, σ to refer to time lags.

Matrices and vectors Denote by $\mathcal{M}_d(\mathbb{R})$ (resp. $\mathcal{M}_d(\mathbb{C})$) the set of square matrices of size $d \times d$ on $\mathbb{R}$ (resp. $\mathbb{C}$). Given a matrix $M \in \mathcal{M}_d(\mathbb{C})$, denote by $\mathbf{adj}(M)$ the adjugate of M, by $M^\intercal$ its transpose, and by $\rho(M)$ its spectral radius. Denote by $\mathbb{I}_d$ the identity matrix of rank d, and by $\mathbb{1}_{\boldsymbol{d}}$ the d-dimensional vector with all coefficients equal to 1.

Given a d-dimensional vector $\boldsymbol{X}$, denote by $\mathbf{diag}(\boldsymbol{X}) \in \mathcal{M}_d(\mathbb{R})$ the diagonal matrix where for $i \in [d]$, $\mathbf{diag}(\boldsymbol{X})_{ii} = X_i$. To avoid confusion, denote multi-dimensional stochastic processes in bold $\boldsymbol{X}$, and their components X_k.

Positivity For $x \in \mathbb{R}$, denote by $(x)_+ := \max(x, 0)$ the positive part of x. By abuse of notation, for a d-dimensional vector $\boldsymbol{x}$, we denote by $(\boldsymbol{x})_+$ the element-wise positive part, i.e.the d-dimensional vector where for all $i \in [d]$, $\big((\boldsymbol{x})_+\big)_i = (x_i)_+$. For a d-dimensional vector $\boldsymbol{x}$, we write $\boldsymbol{x} > 0$ if all the components of $\boldsymbol{x}$ are positive, and similarly for $\boldsymbol{x} \geq 0$.

L_p metrics Denote by $||f||_1$ the L_1 norm of integrable functions $f : [0, +\infty) \to \mathbb{R}$. Let $\boldsymbol{X}, \boldsymbol{Y}$ be two d-dimensional stochastic processes and fix two reals $a < b$. For $k \in [d]$, define the component-wise L_2 distance between X_k, Y_k by

$$\|X_k - Y_k\|_{2,[a,b]} := \sqrt{\frac{1}{b-a}\int_a^b \mathbb{E}[(X_k(t) - Y_k(t))^2]\mathrm{d}t}.$$

Define the L_2 distance between the vector processes $\boldsymbol{X}, \boldsymbol{Y}$ by

$$\|\boldsymbol{X} - \boldsymbol{Y}\|^2_{2,[a,b]} := \sum_{k=1}^{d} \|X_k - Y_k\|^2_{2,[a,b]}.$$

By abuse of notation, let $\|\boldsymbol{X} - \boldsymbol{Y}\|^2_{2,b} := \|\boldsymbol{X} - \boldsymbol{Y}\|^2_{2,[0,b]}$.

Chapter 1

Introduction

Temporal point processes are widely applied as models of asynchronous streams of events. One way to specify these models is through their conditional intensity, that is, the expected infinitesimal rate of events per unit of time, conditioned on the history of the process. A parsimonious class of conditional models is Hawkes processes, where the conditional intensity at a given time is given by a linear auto-regression on the previous jumps of the process, parameterised by a kernel matrix and a bias term (background rate); see Hawkes [47]. In this work, we are interested in two types of Hawkes models: the multi-variate Hawkes process (MHP), usually referred to in the literature as the linear Hawkes process; and the marked time-dependent linear Hawkes process (MTLH). The widespread use of Hawkes processes is mainly due to their explainability: their matrix of kernel functions accounts for self-excitation and cross-excitation between different types of events, and their cluster representation can be a proxy for causality between events. Hawkes processes have applications in a variety of domains including *finance*, particularly in market microstructure (see Bacry et al. [9] and Hawkes [48] for an extensive review); *social networks*, with an emphasis on modelling information cascades such as retweets (see Zhao et al. [110], Kobayashi and Lambiotte [55] and Chen et al. [23]); *seismology*, to study the occurrence of earthquakes and their aftershocks (see Veen and Schoenberg [104]); and *criminology*, to examine criminal contagion mechanisms, notably in burglaries and gang violence (see Mohler et al. [70], Lewis et al. [59], and Mohler [69]).

Because of the auto-regressive nature of this conditional intensity model, the estimation of the kernel matrix and background rates of Hawkes models is a difficult problem, usually giving rise to objective functions whose gradient is expensive to compute without strong simplifying assumptions on the kernel matrix. In practice, the absence of a fast parametric estimation method prohibits the use of Hawkes models with significant amounts of data (i.e., of order higher than 10^6–10^7 jumps in an observed sample path), and with arbitrary kernels, in particular non-Markovian kernels. These limitations arise because the evaluation

of the conditional intensity at each time t has linear complexity in the number of jumps up to time t, leading to quadratic complexity overall. Therefore, objective functions based on the conditional intensity of Hawkes processes are expensive to evaluate and to minimize (see Section 4.2 for a detailed analysis). A notable exception is the Hawkes process with exponential kernels (see Section 4.6), where the conditional intensity can be evaluated recursively, which explains the predominance of exponential MHP in the literature.

In the first part of this work, we overcome these limitations by developing a stochastic optimization algorithm for parametric and semi-parametric estimation of Hawkes models that does not directly evaluate the conditional intensity; we call this the `ASLSD` algorithm (Adaptively Stratified Least Squares Descent). This algorithm is computationally efficient, accurate for a wide range of sample sizes, and flexible enough to allow for regularization and sparsity terms to be easily included. In the second part of this work, we use `ASLSD` to model mid-price jumps on the Nasdaq equities market and build a counter-factual impact model.

1.1 Estimating Hawkes processes

We briefly review the state of the art for the estimation of Hawkes processes. The time complexity, assumptions, objective function, and regularization type of the algorithms discussed here are summarized in Table 4.1. To the best of our knowledge, there is no parametric estimation approach for MTLH models in the literature. MHP estimation procedures fall into three main categories:

- *Method of moments.* These procedures are typically based on spectral properties of the MHP, and usually aim to convert the estimation problem into solving a system of equations. Most of these methods are non-parametric, and require mild assumptions beyond stationarity of the MHP as they use second order properties of the process.
- *Maximum likelihood estimation.* As in other statistical problems, this approach benefits from sound theoretical guarantees. However, the evaluation of the log-likelihood of the MHP has a quadratic time complexity in the number of jumps observed. The application of the expectation-maximization (EM) algorithm to MHP estimation usually improves the convergence of optimization algorithms, but the high computational cost of the E step in EM methods does not allow for efficient algorithms.
- *Least squares estimation.* This class of methods is rarely used in the context of MHP. The cost of evaluating the least squares objective function is roughly as expensive as that of evaluating the log-likelihood. Nonetheless, in this work we show that, unlike

the log-likelihood, the least squares error (LSE) has an additive decomposition that is particularly suitable for efficient stochastic approximation.

Method of moments Hawkes [46, 47] applies methods developed for the general analysis of the spectra of point processes by Bartlett [13]. Hawkes shows a link between the Laplace transform of the autocovariance function ν of the increments of stationary MHP and the mean conditional intensity and kernels of the MHP. Bacry et al. [5] use this link to propose a non-parametric estimation method in the specific case of stationary MHP with symmetric kernels and Laplace transforms diagonalizable in the same orthogonal basis. All these assumptions (except stationarity) are relaxed by Bacry and Muzy [10], who show that the MHP parameters solve a system of Wiener–Hopf equations; we use this algorithm as a nonparametric baseline in our numerical examples. Achab et al. [2] use the first three cumulants of the MHP to propose a non-parametric estimation method for the adjacency matrix of the MHP (the matrix of L_1 norms of the kernels). This algorithm is fast, as it depends linearly on the number of jumps of the MHP; however, this method is not meant for the estimation of the kernels themselves. Finally, the work of Gao et al. [38] relies on the spectrum of the cumulative number of jumps of different types instead of the autocovariance property. While algorithms based on the method of moments apply to a wide range of models, they are particularly inefficient when the number of observations is small. These moment-based methods are also particularly prone to the curse of dimensionality (with respect to the number of dimensions d of the MHP), and regularization for the sake of dimensionality reduction seems difficult for these models.

Maximum likelihood estimation A different paradigm consists in maximising the log-likelihood of the sample path, see Daley and Vere-Jones [28]. To the best of our knowledge, the fastest parametric approach, in the case where the kernels are a sum of exponentials with fixed decay rates, is that in Bompaire et al. [15]; we use this algorithm as a parametric baseline in our numerical examples. Lemonnier and Vayatis [56] use Bernstein polynomials to give a density argument to justify the choice of a linear combination of exponential decays. In the case of linear combinations of non-exponential kernels, Bacry et al. [8] propose a mean field approximation of the log-likelihood to speed up standard parametric estimation. Despite the speed of this method, it is difficult to generalize it due to the mean-field and linearity assumptions. Another limitation of log-likelihood methods for MHP estimation is the flatness of the log-likelihood. A classic approach to solve this issue is the EM procedure introduced by Veen and Schoenberg [104] and Lewis and Mohler [58], which is based on Hawkes and Oakes's [49] cluster representation of the MHP. In the general

case, the complexity of an EM iteration remains quadratic, but significantly smoothes the objective. The ADM4 algorithm of Zhou et al. [111] also builds on the EM approach, with the assumption that the kernels of the MHP are of a fixed form with a single scale coefficient. This method uses sparsity and low rank penalties to estimate high-dimensional MHP. Finally, Zhou et al. [112] show that the kernels satisfy an Euler–Lagrange equation and use a Seidel method to solve it numerically. Again, these methods are not applicable to general kernels without a significant computational burden.

Least squares estimation Among M-estimation methods for MHP, the log-likelihood is significantly more popular than the least squares functional. To the best of our knowledge, the work of Reynaud-Bouret and Schbath [90] is the first to introduce this objective for MHP. Their estimation method is meant for piece-wise constant kernels with finite support, with a view towards applications to genomics. Bacry et al. [3] develop an approach for more conventional kernels; namely, linear combinations of exponential kernels with fixed decays. They are interested in dimensionality reduction via sparsity inducing penalties, as the number of kernels in the MHP is quadratic in the number of event types. However, their method is not applicable to general kernels.

1.2 Price modelling with Hawkes processes

The majority of electronic equity markets are order-driven platforms operating with a limit order book (LOB) using a price-display-time priority queue. Empirically, several properties of prices and order flows in this system are identified as stylized facts such as heavy-tailed return distributions, volatility clustering, long memory in order flow, and auto-correlation and long memory of returns, see for example Abergel et al. [1] and Gould et al. [42]. Building accurate high-frequency financial models is a well studied problem that requires to find a trade-off between analytical tractability, interpretability, and good fit of the data. Some of the first jump-diffusion models taking these constraints into account can be found in Cont et al. [27]. In particular, during the last twenty years, there has been a growing interest towards Hawkes models in high-frequency finance. The nature of price and volume changes in a LOB is discrete, which cannot be captured by diffusive models; and asynchronous, which cannot be captured by time series models: point processes can account for these realistic features. The majority of financial applications of Hawkes models focus on MHP with exponential kernels, notably because of the Markovianity of exponential Hawkes kernels that leads to interesting numerical properties and ergodicity properties. But for general kernels, the computation of the conditional intensity of a Hawkes model at all jump times has quadratic complexity.

In this work, we focus on a specific application of Hawkes models in market microstructure: mid-price models. In their pioneering work, Bacry et al. [6] consider an MHP $\boldsymbol{N}$ with two event types: upward and downward moves of the mid-price. They assume $\boldsymbol{N}$ doesn't show any self excitation ($\phi_{11} = \phi_{22} = 0$) and use the same exponential decay function for the cross-excitation kernels. They model the mid-price p as a functional of $\boldsymbol{N}$:

$$p_t = p_0 + N_t^1 - N_t^2. \tag{1.1}$$

Therefore, N^1 (resp. N^2) models the upward (resp. downward) jumps of the mid-price, which are assumed to be jumps of constant size. The assumptions $\phi_{11} = \phi_{22} = 0$ and $\phi_{12} = \phi_{21}$ guarantee the mean reversion of the mid-price in accordance with empirical observations at the high frequency level. They get an analytical formula for the quadratic variation estimator of this mid-price and show its re-scaled version converges to a Brownian motion. The authors also consider the case of a coupling between two assets, with a 4-dimensional exponential Hawkes process with events: upward jumps of asset A, downward jumps of asset A, upward jumps of asset B, downward jumps of asset B. They also add some sparsity assumptions on the kernel matrix and assume it is symmetric. In this framework, they get an analytical formula of the correlation between the two mid-prices which behaviour is qualitatively coherent with the well known Epps effect. Bacry et al. [7] prove a diffusive limit for the model in (1.1) and show it also verifies the lead-lag effect. Another major step in the study of mid-price models is achieved by Jaisson et al. [53]. The authors noted that across several numerical applications of Hawkes processes to estimate models underlying (1.1), the fitted kernels are close to instability. Therefore, the authors proposed to study a sequence of stable uni-dimensional MHPs $(N^T)_T$ with general kernels $(\phi^T)_T$ such that $\phi^T = a_T \phi$, where $\|\phi\|_1 = 1$ and the sequence (a_T) is in $(0, 1)$ with $\lim_{T \to +\infty} a_T = 1$. As the authors note, this is a very specific convergence to an unstable limit, but they manage to prove convergence of this model to a CIR model, and that the model of Bacry et al. [6] converges to a Heston model.

1.3 Work outline

This thesis is separated into two parts. Part I is dedicated to our work on least-squares estimation of Hawkes processes. As an accompaniment to our derivation and analysis, an implementation in python is available at `https://github.com/saadlabyad/aslsd`. Part II is focused on high-frequency mid-price models in the Nasdaq equities market with the `ASLSD` method. We propose a python implementation of our analysis tools in `https://github.com/saadlabyad/lob`. All proofs of our results are at the end of this work, in Appendix A.

Least-squares estimation of Hawkes models Chapter 2 presents the least-squares estimation framework for point processes, our methodology for model evaluation, and some useful classes of Poisson models. Chapter 3 discusses the two classes of Hawkes models we are interested in. MHP models are explainable models that capture first and second order moments of stationary data. However, they have several limitations, notably that the least squares fit of non-stationary data with MHP models can be particularly misleading. Therefore, we introduce the richer class of MTLH models. In Chapter 4, Theorem 4.3.1 provides a decomposition of the LSE as a sum of functions. The rationale behind this decomposition is that, if these functions and their partial derivatives can be evaluated quickly, then a Monte Carlo estimator of the gradient of the LSE is inexpensive to evaluate. We construct this Monte Carlo estimator using adaptive stratified sampling for variance reduction purposes, allowing for general kernels for the Hawkes model. We combine this estimator with numerical schemes from the stochastic gradient descent literature to propose `ASLSD`, a new fast estimation method for MHP and MTLH models with general kernels and large datasets. We discuss different parametric and semi-parametric families of kernel densities for the `ASLSD` algorithm. Finally, in Chapter 5, we evaluate our method on synthetic data and benchmark it against state of the art algorithms; then give two example applications of our method with real-world data: first in epidemiology modelling malaria infections in China, then modelling news cycles with the MemeTracker dataset.

Hawkes models of Nasdaq equities prices Chapter 6 gives some institutional background about the Nasdaq equities market, and introduces our mathematical notation and conventions. Chapter 7 presents the dataset we construct for this work, that is composed of 48 Nasdaq tickers, studied on a period of 64 trading days ranging from Wednesday 1 June 2022 to Friday 2 September 2022. Chapter 8 discusses empirical properties of mid-prices in this dataset, which guide our modelling choices. In Chapter 9, our parametric estimation procedure for MHP and MTLH allows us to consider mid-price models with general kernels, in particular, with delayed and multi-modal kernels. In high frequency financial data, consecutive events might happen too fast to have triggered one another because of latency: there might be several events happening in between the arrival of an order and the arrival of the order it triggers. This feature cannot be captured by any decreasing kernel (notably exponential or power law kernels).

Part I

Least-squares estimation of Hawkes models

Chapter 2

Conditional intensity modelling

2.1 Point processes

In this section, fix $d \in \mathbb{N}^*$ the dimension (number of event types) of the considered processes.

2.1.1 Definition and basic properties

We briefly recall basic definitions of point processes, their conditional intensity, and their compensator based on Daley and Vere-Jones [28].

Definition 2.1.1 (Point process)**.** *A d-dimensional orderly point process is a random sequence of times $\mathcal{T} = \left\{t_m^i : m \in \mathbb{N}^*, i \in [d], t_m^i < t_{m+1}^i\right\}$. The associated counting process $\boldsymbol{N}$ is defined for times $t \geq 0$ by $\mathbf{N_t} := \left(N_t^i\right)_{i\in[d]}$, where $N_t^i := \sum_{i=1}^{+\infty} \mathbb{1}_{\left\{t \geq t_i^j\right\}}$. We denote the total number of jumps up to time t by $N_t := \sum_{i=1}^{d} N_t^i$.*

Consider a d-dimensional point process $\mathcal{T}$ and the associated counting process $\boldsymbol{N}$. Let $\mathcal{F} = (\mathcal{F}_t)_{t\geq 0}$ be the natural filtration of $\boldsymbol{N}$. The counting process $\boldsymbol{N}$ is characterized by its conditional rate of events per unit time given the history of the process.

Definition 2.1.2 (Conditional Intensity)**.** *For $i \in [d]$, the conditional intensity of N^i is defined by $\lambda_i(t) := \lim_{h\downarrow 0} \frac{\mathbb{E}[N_{t+h}^i - N_t^i = 1 | \mathcal{F}_{t-}]}{h}$, and we write $\boldsymbol{\lambda} := (\lambda_1, \ldots, \lambda_d)^\intercal$.*

We define the compensator associated to a counting process, which is simply the integral of the conditional intensity.

Definition 2.1.3 (Compensator)**.** *For $i \in [d]$, the compensator of N^i is defined by $\Lambda_i(t) := \int_0^t \lambda_i(t)\mathrm{d}t$, and we write $\boldsymbol{\Lambda} := (\Lambda_1, \ldots, \Lambda_d)^\intercal$.*

The Doob–Meyer decomposition theorem states that for all event types $i \in [d]$, the process defined for times $t \geq 0$ by $\tilde{M}_t^i := N_t^i - \Lambda_i(t)$, is a martingale. Furthermore, we note that $\langle \tilde{M}^i \rangle_t = \Lambda_i(t)$, hence the stochastic process $\left((\tilde{M}_t^i)^2 - \Lambda_i(t)\right)_{t\geq 0}$ is a martingale.

2.1.2 Moments

For event types $k \in [d]$, and for bounds $a, b \in [0, +\infty)$ with $a < b$, define $\eta^k_{[a,b]} := \frac{N^k([a,b])}{b-a}$. By abuse of notation, for all event types $k \in [d]$, we define the cumulative event rate process η^k for all times $t > 0$ by $\eta^k_t := \eta^k_{[0,t]}$. It is clear that $\boldsymbol{\eta_t}^\intercal := (\eta^1_t, \dots, \eta^d_t)$ is right-continuous, therefore, we also define this process for $t = 0$ by $\boldsymbol{\eta_0} = 0$. The event rate process, which is empirically observed given a data path of $\boldsymbol{N}$, is of particular interest in the study of point processes. First, as we discuss in Section 2.4.1, the vector of rates $\boldsymbol{\eta_T}$ is both the maximum likelihood estimator and the least squares estimator of the conditional intensity under a homogeneous Poisson model. Second, for MHP (see Section 3.1), this quantity is at the center of the law of large numbers for Hawkes processes.

Now fix an event type $k \in [d]$. We get the moments of the cumulative event rate process $\boldsymbol{\eta^k_t}$ using the Doob–Meyer decomposition of N^k.

Proposition 2.1.1 (Moments of the cumulative event rate)**.** *For event types $k \in [d]$, and for times $t > 0$, the moments of the rate process are*

$$\mathbb{E}[\eta^k_t] = \frac{1}{t}\mathbb{E}[\Lambda_k(t)], \qquad \mathrm{Var}[\eta^k_t] = \frac{1}{t^2}\left(\mathbb{E}\left[\Lambda_k(t) + 2\tilde{M}^k_t \Lambda_k(t)\right] + \mathrm{Var}[\Lambda_k(t)]\right). \tag{2.1}$$

We give a proof of this result in Appendix A.1.1.

Definition 2.1.4 (Covariance function)**.** *Fix a sampling period $h > 0$ and a lag $\tau \geq 0$. We call covariance function of the counting process $\boldsymbol{N}$ at time $t \geq 0$ the squared matrix $\nu^{(h)}_\tau(t)$, where for all event types $i, j \in [d]$, the entry $\nu^{(h)}_{ij,\tau}(t)$ is defined by*

$$\nu^{(h)}_{ij,\tau}(t) := \frac{1}{h}\mathrm{Cov}\left[N^j\Big([t, t+h)\Big), N^i\Big([t+\tau, t+\tau+h)\Big)\right]. \tag{2.2}$$

If the counting process $\boldsymbol{N}$ is stationary, then the covariance function $\nu^{(h)}_\tau(t)$ does not depend on the time t and $\nu^{(h)}_\tau(t) = \nu^{(h)}_\tau(0)$ for all $t \geq 0$. In this case, and when there is no ambiguity, we write $\nu^{(h)}_\tau$ instead of $\nu^{(h)}_\tau(t)$. For times $t \geq 0$, and sampling periods $h > 0$, the triangular function $f^{(h)}_{\mathrm{Tr}}$ is

$$f^{(h)}_{\mathrm{Tr}}(t) := \left(1 - \frac{t}{h}\right)_+. \tag{2.3}$$

This function appears in the autocovariance function of the point processes in this work.

2.2 Model estimation framework

We now formalise the estimation problems of conditional intensity models.

Data setup Let $\boldsymbol{N}$ be a d-dimensional counting process. $\boldsymbol{N}$ is our data generating process, denote by $\boldsymbol{\lambda}^\diamond$ its ground truth intensity. For the rest of this subsection, fix a horizon $T > 0$. In this work, we distinguish two setups:

1. the **long path setup**; where we observe a single sample path of the counting process $\boldsymbol{N}$ on $[0, T]$. The underlying assumption is that we observe the process over a sufficiently large horizon T. This work is mostly focused on the long path setup.
2. the **episodic setup**; where we observe multiple sample paths of the counting process $\boldsymbol{N}$ on $[0, T]$. In this case, the time horizon T might be small, but we need a large enough number of episodes (sample paths).

Whether we place ourselves in the long path setup or the episodic setup, we assume that observed paths are non-trivial, that is

- we observe at least one event of each type (*i.e.* for event type $i \in [d]$, $N_T^i > 1$);
- for each event type, the last observed event of that type is preceded by at least one event of every type (*i.e.* using the notation of Chapter 4, for event types $i, j \in [d]$, $\varpi(i, j) < N_T^i$).

Intensity modelling Let Θ be a subset of a Euclidean space denoting the space of parameters. To each given value of parameters $\theta \in \Theta$, we associate a conditional intensity model $\boldsymbol{\lambda}^{(\boldsymbol{\theta})}$. The idea behind least squares estimation of point processes is to find parameter values that minimize the L_2 distance between the model intensity $\boldsymbol{\lambda}^{(\boldsymbol{\theta^*})}$ and the ground truth intensity $\boldsymbol{\lambda}^\diamond$. In Section 2.2.1, we motivate this approach by briefly discussing the identifiability of conditional intensity models. Then in Section 2.2.2, we introduce the usual definition of our loss function, the least squares error (LSE). The separability of the LSE allows to parallelize the minimization of this loss function. We show some results on the empirical evaluation of the LSE.

2.2.1 Model identifiability

Let $\mathbb{M}$ denote a subset of d-dimensional conditional intensity processes (for example, MHP or MTLH conditional intensities). Clearly, $\theta_A = \theta_B$ implies $\boldsymbol{\lambda}^{(\boldsymbol{\theta_A})} = \boldsymbol{\lambda}^{(\boldsymbol{\theta_B})} \quad a.s.$.

Definition 2.2.1 (Identifiable Model)**.** *We say that a class of models Θ is weakly identifiable if the map*

$$V \colon \Theta \to \mathbb{M}; \quad \theta \mapsto \boldsymbol{\lambda}^{(\boldsymbol{\theta})}. \tag{2.4}$$

is injective, that is $\forall \theta_A, \theta_B \in \Theta$

$$\boldsymbol{\lambda}^{(\boldsymbol{\theta_A})} = \boldsymbol{\lambda}^{(\boldsymbol{\theta_B})} \quad a.s. \implies \quad \theta_A = \theta_B. \tag{2.5}$$

We say that a class of models is path-wise identifiable if $\forall \theta_A, \theta_B \in \Theta$

$$\theta_A \neq \theta_B \implies \boldsymbol{\lambda}^{(\boldsymbol{\theta_A})} \neq \boldsymbol{\lambda}^{(\boldsymbol{\theta_B})} \quad a.s.. \tag{2.6}$$

It is clear that if a family of models is path-wise identifiable, then it is weakly identifiable, but the converse is not necessarily true.

2.2.2 Loss function: the least squares error

2.2.2.1 Definition and properties

We can now define the LSE of a model.

Definition 2.2.2 (LSE)**.** *For all event types* $k \in [d]$*, define the* k*-th partial LSE of a conditional intensity model* λ_k *as the random variable*

$$\mathcal{R}_T^{(k)}(\lambda_k) := \frac{1}{T} \int_0^T \lambda_k(t)^2 \, \mathrm{d}t - \frac{2}{T} \sum_{m=1}^{N_T^k} \lambda_k \left(t_m^k \right). \tag{2.7}$$

The LSE of the conditional intensity model $\boldsymbol{\lambda}$ *is the random variable*

$$\mathcal{R}_T(\boldsymbol{\lambda}) := \sum_{k=1}^{d} \mathcal{R}_T^{(k)}(\lambda_k). \tag{2.8}$$

Given a parametric family of models $(\boldsymbol{\lambda}^{\boldsymbol{\theta}})_{\boldsymbol{\theta} \in \Theta}$, for all parameters $\boldsymbol{\theta} \in \Theta$, we write by abuse of notation $\mathcal{R}_T(\boldsymbol{\theta})$, and $\mathcal{R}_T^{(k)}(\boldsymbol{\theta})$ for all event types $k \in [d]$.

Motivation A well known result obtained using the Doob–Meyer decomposition of $\boldsymbol{N}$ is that for all $\boldsymbol{\theta} \in \Theta$

$$\mathbb{E}\left[\mathcal{R}_T(\boldsymbol{\theta})\right] = \|\boldsymbol{\lambda}^{(\boldsymbol{\theta})} - \boldsymbol{\lambda}^{\diamond}\|_T^2 - \|\boldsymbol{\lambda}^{\diamond}\|_T^2. \tag{2.9}$$

Since the norm of the ground truth conditional intensity $\|\boldsymbol{\lambda}^{\diamond}\|_T^2$ is unknown but fixed, minimizing the expected LSE $\mathbb{E}\left[\mathcal{R}_T(\boldsymbol{\theta})\right]$ over $\boldsymbol{\theta} \in \Theta$ is equivalent to minimizing the L_2 distance between models and ground truth intensities $\|\boldsymbol{\lambda}^{(\boldsymbol{\theta})} - \boldsymbol{\lambda}^{\diamond}\|_T^2$ over $\boldsymbol{\theta} \in \Theta$. The parameters estimator we keep is a solution to the minimization program $\mathcal{P}$ defined as

$$\boldsymbol{\theta}^{\star} := \min_{\boldsymbol{\theta}} \quad \mathbb{E}[\mathcal{R}_T(\boldsymbol{\theta})] \qquad \text{s.t.} \quad \boldsymbol{\theta} \in \Theta \tag{2.10}$$

Denote by $\bar{\boldsymbol{N}}$ the counting process with fitted conditional intensity $\boldsymbol{\lambda}^{(\boldsymbol{\theta}^{\star})}$. It is not clear whether this estimator is unbiased or consistent and, if so, at which rate does it converge. Nevertheless, this estimator gives satisfactory results in practice, as seen in Reynaud-Bouret and Schbath [90], Gaïffas and Guilloux [37], Hansen et al. [44], Bacry et al. [3], and in

our numerical experiments (see Chapter 5). Using Equation (2.9), a lower bound for the expected LSE is

$$\mathbb{E}\left[\mathcal{R}_T(\boldsymbol{\theta})\right] \geq -\|\boldsymbol{\lambda}^\diamond\|_T^2, \quad \forall \boldsymbol{\theta} \in \Theta. \tag{2.11}$$

If the ground truth $\boldsymbol{\lambda}^\diamond$ belongs to the space of admissible conditional intensity models Θ, then this lower bound on the expected LSE is attained.

Parallelization We do not impose a specific structure on the optimization program $\mathcal{P}$. To simplify the presentation of our methodology, we work under the following general assumptions:

1. For each event type $k \in [d]$, we assume that the conditional intensity model λ_k has its own sub-vector of parameters $\boldsymbol{\theta_k}$, that is independent from the parameterization of the other conditional intensities. For example, in the case of the Hawkes models discussed in this work, this means that each baseline, each kernel, and each impact function is parameterized independently, and we do not impose symmetry conditions on the kernel matrix model.
2. There is no coupling between the sets of different parameters. For instance, in the case of the Hawkes models in this work, we do not allow a stability constraint on the model adjacency matrix, which would couple the L_1 weights parameters of the kernels.

Note that the `ASLSD` method can be easily adapted to incorporate further constraints such as symmetry conditions and coupling constraints. Under the general assumptions above, for all event types $k \in [d]$, the partial LSE $\mathcal{R}_T^{(k)}(\boldsymbol{\theta})$ only depends on $\boldsymbol{\theta_k}$ and the observed jumps. Therefore, using Equation (2.8), we get

$$\mathcal{R}_T(\boldsymbol{\theta}) := \sum_{k=1}^{d} \mathcal{R}_T^{(k)}(\boldsymbol{\theta_k}). \tag{2.12}$$

For each event type $k \in [d]$, define the minimization program $\mathcal{P}_k$ as

$$\min_{\boldsymbol{\theta_k}} \quad \mathcal{R}_T^{(k)}(\boldsymbol{\theta_k}) \qquad \text{s.t.} \quad \boldsymbol{\theta_k} \in \Theta_k \tag{2.13}$$

This implies that the minimization program $\mathcal{P}$ is equivalent to solving the d independent minimization programs $(\mathcal{P}_k)_{k\in[d]}$. In the remainder of this section, we fix an event type $k \in [d]$ and focus on the program $\mathcal{P}_k$.

2.2.2.2 Long path setup

Empirical evaluation As discussed above, the motivation for the use of the LSE as a loss function is that minimizing its expectation ensures that the model intensity is as close as possible to the ground truth intensity in the L_2 sense. However, in the long path setup, we only observe one sample path of the counting process $\boldsymbol{N}$, and therefore, only one realisation of the LSE $\mathcal{R}_T(\boldsymbol{\theta})$ for any given model $\boldsymbol{\theta}$. Therefore, the practical question we face is the following: for large enough observation windows T, is the LSE $\mathcal{R}_T(\boldsymbol{\theta})$ close enough to its expectation? In order to answer this question, we define the notion of temporal consistency.

Definition 2.2.3 (Temporal consistency)**.** *Let $(Z_t)_{t\geq 0}$ be an integrable stochastic process. We say that Z is temporally consistent if*

$$|Z_T - \mathbb{E}[Z_T]| \xrightarrow[T\to+\infty]{P} 0. \tag{2.14}$$

In general, it is not clear under which conditions on the ground truth conditional intensity $\boldsymbol{\lambda}^\diamond$ and the space of models Θ is the LSE temporally consistent. Using Markov's inequality, a sufficient condition to show that the LSE $\mathcal{R}_T(\boldsymbol{\theta})$ is temporally consistent is if

$$\lim_{T\to+\infty} \mathrm{Var}[\mathcal{R}_T(\boldsymbol{\theta})] = 0. \tag{2.15}$$

If we assume temporal consistency, in the long path setup, the objective of the estimation procedure we consider is to solve the optimization program (P)

$$\min_{\boldsymbol{\theta}} \quad \mathcal{R}_T(\boldsymbol{\theta}) \qquad \text{s.t.} \quad \boldsymbol{\theta} \in \Theta. \tag{2.16}$$

Sign of the LSE Fix an event type $k \in [d]$. A first difference between the LSE (in the context of point processes estimation) and the mean squared error in regression problems (such as the ordinary least squares estimation of linear regression models) is that the partial LSE $\mathcal{R}_T^{(k)}$ is not necessarily positive. In fact, if constant conditional intensity models (*i.e.* homogeneous Poisson, see Section 2.4.1) are allowed in the parametric family of models, which is the case for the Hawkes models we consider in this work, then the minimal value of the partial LSE is almost surely negative. This is because, if we evaluate the LSE at the constant intensity model $\lambda_k = \eta_T^k$, then the partial LSE is

$$\mathcal{R}_T^{(k)}(\lambda_k) = -(\eta_T^k)^2. \tag{2.17}$$

This implies an almost sure upper bound for the minimal value of the partial LSE

$$\min_{\boldsymbol{\theta_k}\in\Theta_k} \mathcal{R}_T^{(k)}(\boldsymbol{\theta_k}) \leq -(\eta_T^k)^2 \quad a.s. \tag{2.18}$$

Over-fitting Depending on the class of admissible conditional intensity models Θ, the LSE is not necessarily almost surely lower bounded. For instance, we build the following over-fitting model by analogy with kernel density estimation. Let $\epsilon > 0$ such that

$$\epsilon < \min_{m \in [N_T^k - 1]} \left(t_{m+1}^k - t_m^k \right), \tag{2.19}$$

and consider the piece-wise constant Poisson [1] conditional intensity model defined by

$$\lambda_k^{(\epsilon)}(t) := \frac{1}{\epsilon} \sum_{k=1}^{N_T^k} \mathbb{1}_{[t_m^k, t_m^k + \epsilon)}(t). \tag{2.20}$$

Figure 2.1 illustrates this. This model achieves a partial LSE

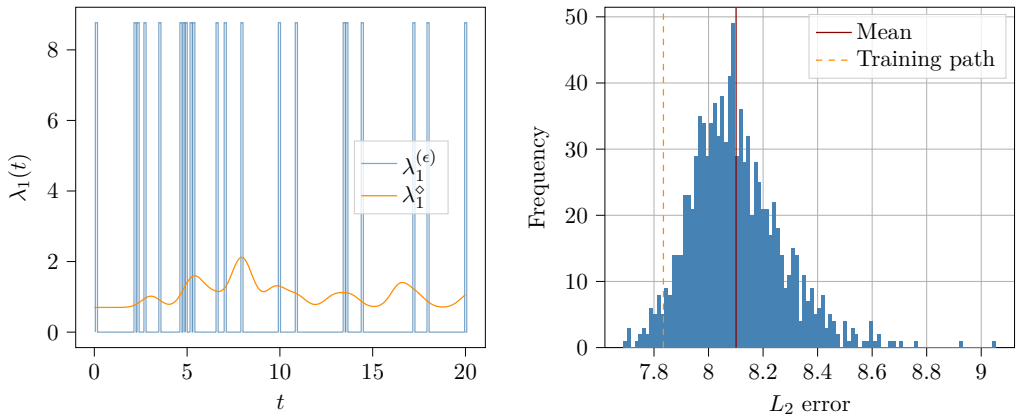


Figure 2.1: A trivial over-fitting model

We simulate a path from a uni-dimensional MHP with Gaussian kernel up to $T = 20$. The corresponding path contains 21 events, and we pick $\epsilon \sim 0.11$ satisfying the constraints above. Left: Orange line plots the ground truth conditional intensity $\lambda_1^\diamond$, blue line plots the over-fitting model $\lambda_1^{(\epsilon)}$. We see that even on the observed data path, the model intensity $\lambda_1^{(\epsilon)}$ is far from the ground truth conditional intensity in the L_2 sense, and localizes around the observed times. Right: we simulate 10^3 paths of the same ground truth MHP, and for each of these paths, we compute the L_2 distance from the deterministic model intensity $\lambda_1^\diamond$. We plot the distribution of these L_2 errors. Red line plots the empirical mean of this distribution, dashed orange line plots the L_2 error with the conditional intensity on the training path.

$$\mathcal{R}_T^{(k)}(\lambda_k) = -\frac{1}{\epsilon} \eta_T^k. \tag{2.21}$$

Therefore, for any given conditional intensity model λ_k, we can always choose $\epsilon > 0$ such that

$$\mathcal{R}_T^{(k)}\left(\lambda_k^{(\epsilon)}\right) < \mathcal{R}_T^{(k)}(\lambda_k). \tag{2.22}$$

This model over-fits the data by not being able to generalize in the two following senses:

[1] To avoid confusion, by this we mean that we observe a path of the point process $(t_m^k)_{m \in [N_T^k]}$, and define the deterministic intensity above based on that single path observed.

1. the model performs poorly on new simulated paths from the ground truth: Equation (2.22) does not hold when taking expectations against the ground truth conditional intensity;
2. the model performs poorly beyond the training horizon: conditional on that training path up to a horizon T, this conditional intensity model is null for times $t > T$.

It is important to note, again, the analogy between this least-squares estimation problem for point processes and least-squares density estimation for i.i.d.(independent and identically distributed) data, as opposed to the least-squares regression problem for i.i.d.data. In the least-squares regression problem for i.i.d.data in the absence of noise, with features x and target y, and a ground truth $f^\diamond$, an over-fitting model f will have the same value as the ground truth on training data

$$f(x_i) = y_i = f^\diamond(x_i). \tag{2.23}$$

As we see in the example of Figure 2.1, this is not the case in point processes estimation: the over-fitting conditional intensity does not match the ground truth intensity, even on the training data, but it does match its empirical version. In fact, if we denote by δ the Dirac delta distribution, then

$$\lambda_k^{(\epsilon)} \xrightarrow[\epsilon \to 0]{P} \sum_{m=1}^{N_T^k} \delta(t - t_m^k). \tag{2.24}$$

This limit process is the process is the constant point process with the same jump times as the training data.

Of course, this over-fitting issue is not specific to the fact that the interval bounds $[t_m^k, t_m^k + \epsilon)$ of this model depend on the training data. For instance, we get a similar problem with piece-wise constant Poisson models in general, with arbitrarily small bounds that do not depend on the training data. In practice, this issue raises the question of out-of-sample testing for point process models (Section 2.3.4), and that of a careful model parameterization according to the scale of the training data.

2.2.2.3 Episodic setup

We now consider extensions of our estimation procedure in the episodic setup where we observe multiple sample trajectories of a given point process. Let $\boldsymbol{N}$ be a d-dimensional counting process. Fix a horizon $T > 0$, and assume we observe n_p sample paths of jump times of the counting process $\boldsymbol{N}$ on the interval $[0, T]$. We know that the conditional intensity model $\boldsymbol{\lambda}$ minimizing the expectation of the least squares error $\mathbb{E}[\mathcal{R}_T]$ also minimizes

the L_2 distance between the intensity and model and the ground truth $\boldsymbol{\lambda}^\diamond$,

$$\frac{1}{T}\sum_{k=1}^{d}\int_0^T \mathbb{E}\left[(\lambda_k(t)-\lambda_k^\diamond(t))^2\right]\mathrm{d}t.$$

For a sample path $\rho \in [n_p]$, we denote by $\mathcal{R}_T^{(k,\rho)}$ the k-th partial LSE evaluated on sample path $\mathcal{T}_T^\rho$. We denote by $\mathcal{G}_T^{(k,\rho)}(\theta_k)$ the estimator of the gradient of the LSE $\mathcal{R}_T^{(k,\rho)}$ on a given sample path. Define

$$\mathcal{E}_T^{(k)} := \frac{1}{n_p}\sum_{p=1}^{n_p}\mathcal{G}_T^{(k,\rho)}. \tag{2.25}$$

Then $\mathcal{E}_T^{(k)}$ is an unbiased estimator of the expected gradient of the LSE $\mathbb{E}\left[\mathcal{R}_T^{(k)}\right]$.

2.3 Model evaluation framework

In this section, we propose different methods to evaluate the quality of a fitted model. Suppose we are in the estimation framework discussed above, with a ground truth intensity $\boldsymbol{\lambda}^\diamond$ and a fitted conditional intensity model $\boldsymbol{\lambda}^\star$ that minimizes the LSE over some class of parametric models. The ground truth intensity is not observable: how can we assess the quality of the model $\boldsymbol{\lambda}^\star$? In Section 2.3.1, we present the standard goodness of fit test for point processes: residual analysis. We propose exact formulas for the residuals of Hawkes models, whose computation has quadratic time complexity. Therefore, we briefly discuss numerical methods for the acceleration of this computation. We discuss the Kolmogorov–Smirnov test for residuals and complementary evaluation frameworks for point process models.

For the rest of this section, suppose we observe a path of the counting process N up to T with ground conditional intensity $\boldsymbol{\lambda}^\diamond$. Let $\boldsymbol{\lambda}$ denote a conditional intensity model, evaluated on the observed path, and let $\hat{\boldsymbol{N}}$ denote the counting process associated to this model. We assume we can simulate paths of $\hat{\boldsymbol{N}}$ exactly.

2.3.1 Residual analysis

Principle Residual analysis is the state of the art goodness-of-fit test for MHP. For event indices $m \in \mathbb{N}^*$, and for event types $k \in [d]$, define the compensator transformed times

$$s_m^{(k)} := \Lambda_k\left(t_m^k\right). \tag{2.26}$$

Fix an event type $k \in [d]$. Define the transformed point process $\mathcal{S}^{(k)} := \left\{s_m^{(k)} : m \in \mathbb{N}^*\right\}$. The inter-arrival times $r^{(k)}$ of the point process $\mathcal{S}^{(k)}$, defined by

$$r_m^{(k)} := s_{m+1}^{(k)} - s_m^{(k)}, \tag{2.27}$$

are usually called residuals in the point processes literature. For convenience, denote by $\mathcal{S} = \left(\mathcal{S}^{(k)}\right)_{k\in[d]}$ the d-dimensional compensator transformed point process. Following Ogata [84], if $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{\diamond}$ *a.s.*, then the compensator transformed processes $\left(\mathcal{S}^{(k)}\right)_{k\in[d]}$ are independent standard Poisson processes. Equivalently, for all event types $k \in [d]$, the residuals $\boldsymbol{r^{(k)}}$ are independent and follow a standard exponential distribution.

Testing identical distribution of residuals To assess if fitted residuals are correctly distributed, we display the Q-Q plots of the residuals against a standard exponential distribution. As visual comparison of the fit of different models can be difficult, we also use the probability plots of residuals, which are defined by

$$z_m^{(k)} := 1 - \exp\left(-s_m^{(k)}\right), \text{ and } \mathcal{J}^{(k)} := \left\{z_m^{(k)} : m \in \mathbb{N}^*\right\}. \tag{2.28}$$

If the residuals $\mathcal{S}^{(k)}$ are independent exponentially distributed, then $\left(\mathcal{J}^{(k)}\right)_{k\in[d]}$ are independent random variables, uniformly distributed on $[0, 1]$. For improved clarity, we subtract the $y = x$ line from the probability plots and rescale $z_m^{(k)}$ with a multiplicative factor $\sqrt{N_T^k - 1}$; for large N_T^k, Donsker's theorem indicates that this results in a process which is approximately a Brownian bridge. Beyond visual inspection, we use the Kolmogorov–Smirnov (KS) test to evaluate the residuals. In each probability plot (for example in Chapter 5), we plot dashed lines corresponding to the 99% critical value for the KS test.

Testing independence of residuals The time transformation property states the residuals are *i.i.d.* standard exponential variables. The KS test, which is the only test used in a significant part of the Hawkes estimation literature, only tests the identical distribution of residuals, but not their independence.

This methodological issue can cause important problems in practice; we illustrate this with the following example. Fix $T > 0$, and consider a path $(t_m^1)_{m\in[N_T^1]}$ of a uni-dimensional standard homogeneous Poisson process $\boldsymbol{N}$ up to time T. Denote by $\bar{\boldsymbol{r}}$ the ordered residuals of $\boldsymbol{N}$, sorted in increasing order. Given this path, construct the uni-dimensional counting process $\bar{\boldsymbol{N}}$ with jump times

$$\tau_0 := \min_{m\in[N_T^1]} r_m^1, \qquad \tau_{m+1} := \tau_m + \bar{r}_m \quad \forall m \in [N_T^1 - 1]. \tag{2.29}$$

The residuals $\bar{\boldsymbol{r}}$ of the counting process $\bar{\boldsymbol{N}}$ are the residuals of the counting process (except sorted in increasing order), therefore will pass the KS test. However, this process $\bar{\boldsymbol{N}}$ is clearly not a standard homogeneous Poisson process since its increments are not independent.

Wald and Wolfowitz [106] propose the the Wald–Wolfowitz test, also known as the runs test: this is a one-sample hypothesis test of independence. In the context of residual analysis,

fix an event type $k \in [d]$, and suppose we want to test for the independence of residuals $\boldsymbol{r^k}$. Denote by $\tilde{r}^k$ the median of this sample. First, construct a sequence $(x_m)_{m \in [N_T^k]}$ where we compare each term of the sample $\boldsymbol{r^k}$ to the median $\tilde{r}^k$; formally for indices $m \in [N_T^k - 1]$

$$x_m := \begin{cases} 1 & \text{if } r_m^k > \tilde{r}^k, \\ 0 & \text{otherwise.} \end{cases} \tag{2.30}$$

Denote by n_1 (resp. n_0) the number of 1 (resp. 0) in the sequence $\boldsymbol{x}$. Wald and Wolfowitz [106] define a run as a sub-sequence $(x_p, x_{p+1}, \ldots, x_{q-1}, x_q)$ with $p \leq q$ such that:

- all elements of the sub-sequence are the same, *i.e.* for all indices $l \in [\![p, q]\!]$, $x_l = x_n$.
- the last element of the sequence $\boldsymbol{x}$ preceding the sub-sequence has a different value; *i.e.* either $p = 0$ or $x_{p-1} \neq x_p$.
- the first element of the sequence $\boldsymbol{x}$ following the sub-sequence has a different value; *i.e.* either $q = N_T^k - 1$ or $x_{q+1} \neq x_q$.

The idea of the Wald–Wolfowitz test is that, under the null hypothesis that the observations in $\boldsymbol{x}$ are independent, the number of runs n_r is normally distributed conditional on n_1 and n_0. The test statistic they define is

$$Z := \frac{n_r - m_r}{\sigma_r}; \tag{2.31}$$

where $m_r := 1 + \frac{2n_0 n_1}{n_0 + n_1}$ and $\sigma_r^2 := \frac{(m_r - 1)(-m_r 2)}{N_T^k - 2}$. Under the null hypothesis, Z must be normally distributed conditional on n_0, n_1, which allows to compute a p-value for this test. Variations of this test, by replacing the median in Equation (2.30) by another value, can also be considered.

2.3.2 One-step-ahead prediction

In this paragraph, we briefly define one-step ahead predictions for the evaluation of point process models, inspired by the work of Du et al. [33] and Mei and Eisner [66]. Consider the following statistical problem. For times $t \in [0, T]$, let $\mathcal{T}_t$ denote the path of the point process $\boldsymbol{N}$ up to t. We are interested in modelling jointly the event type $\iota_t \in [d]$ and the lag s_t until the next jump of $\boldsymbol{N}$ conditional on the history (the path) $\mathcal{T}_t$. Predicting the event type ι_t is a classification problem, and predicting s_t is a regression problem.

For several classes of conditional intensity models, notably for the Hawkes models in this work, the distribution of inter-arrival times of events is not known in closed-form. We simulate one step ahead[2] from the model $\boldsymbol{\lambda}$ conditional on the path $\mathcal{T}_t$. This way, we

[2]For Hawkes models, we use an exact branching simulation algorithm to simulate conditional on a historical path.

get samples of the next event $(\iota_t, t + s_t)$, hence an empirical estimate of the distribution of (ι_t, s_t). To define the prediction $(\hat{\iota}_t, \hat{s}_t)$ for the next event, we first simulate n_{paths} paths one-step-ahead of the model. We then define $\hat{\iota}_t$ to be the mode of the empirical distribution of event types in the simulated samples; and $\hat{s}_t$ to be the empirical mean of lags s_t conditional on the event type of samples being $\hat{\iota}_t$. To assess the quality of this prediction over the observed window, we fix a number $n_{\text{grid}} \in \mathbb{N}^*$, and consider a time grid $G := \Big\{t_i : i \in [n_{\text{grid}}]\Big\}$. To avoid edge effects that may result from the data collection process, we choose $G \subset [0.2 \times T, 0.8 \times T]$. Define the evaluation metrics

$$\texttt{OSA}_R := \sqrt{\frac{\sum_{t \in G}(\hat{s}_t - s_t)^2}{\sum_{t \in G}(s_t)^2}}, \quad \texttt{OSA}_C := \frac{\#\Big\{t \in G : \hat{\iota}_t = \iota_t\Big\}}{\#G}. \tag{2.32}$$

By analogy with the time series literature, for example in Cerqueira et al. [22], we assess these values using a `Naive` benchmark, predicting the next event type and lag to be the same as the last observed value.

2.3.3 Empirical moments

The first and second empirical moments of a point process are summary statistics giving insights on the dynamics of that process. Under the right assumptions, they are unbiased estimators of the moments of the process, computed from observed data. In fact, for some classes of point processes, the conditional intensity is fully characterized by the first and second order moment of the process: this is the case for MHP models (see Section 3.1), the standard class of linear Hawkes models. Therefore, if we compute the empirical moments of data observed from a ground truth point process, we can compare them to empirical moments computed on data simulated from a fitted model for evaluation.

Fix a sampling period $h \in [0, T]$, and divide the time interval $[0, T]$ into $n_{\text{bins}} := \Big\lfloor T/h \Big\rfloor$ intervals of length h. As illustrated in Figure 2.2, for event types $i \in [d]$, and for $m \in [\![0, n_{\text{bins}} - 1]\!]$, we observe the bin count $c_{i,m}^{(h)}$, that is, the number of events of type i in the m-th bin

$$c_{i,m}^{(h)} := N^i\Big([mh, (m+1)h)\Big). \tag{2.33}$$

Analysing the joint distribution of bin counts is a tedious task; instead, the point processes literature focuses on their sample mean and covariance.

First order: empirical event rate Using the Doob–Meyer decomposition of $\boldsymbol{N}$, the conditional intensity of the point process is related to the following empirical intensity.

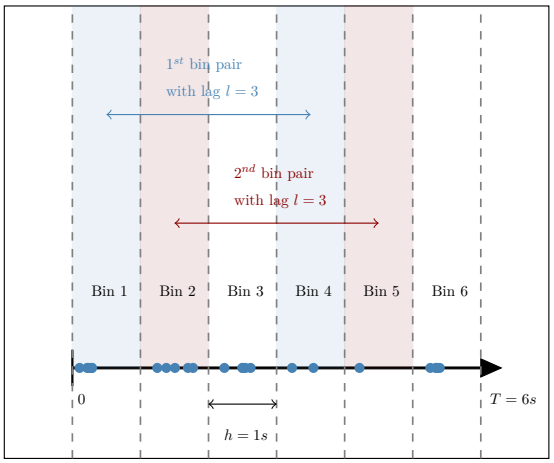


Figure 2.2: Illustration of bin structure

Definition 2.3.1 (Empirical intensity)**.** *For event types $i \in [d]$, and for bin indices $m \in [\![0, n_{bins} - 1]\!]$, the empirical intensity is*

$$L_{i,m}^{(h)} := \frac{c_{i,m}^{(h)}}{h} = \eta_{[mh,(m+1)h)}^{i}. \tag{2.34}$$

Proposition 2.3.1 (Empirical and conditional intensity)**.** *For event types $i \in [d]$, and for bin indices $m \in [\![0, n_{bins} - 1]\!]$, the empirical intensity satisfies*

$$\mathbb{E}\left[L_{i,m}^{(h)}\right] := \frac{1}{h}\int_{mh}^{(m+1)h} \mathbb{E}\left[\lambda_i^{\diamond}(t)\right] \mathrm{d}t. \tag{2.35}$$

Furthermore, for times $t \in [0, T]$,

$$\lim_{h \downarrow 0} \mathbb{E}\left[L_{i,\lfloor \frac{t}{h} \rfloor}^{(h)}\right] := \mathbb{E}\left[\lambda_i(t)\right]. \tag{2.36}$$

To evaluate model fits, we simulate paths from the conditional intensity model $\boldsymbol{\lambda}$, compute $\left(\bar{L}_{i,m}^{(h)}\right)$ the empirical intensity of a simulated path, and define

$$\texttt{M1} := \sqrt{\frac{\sum_{i=1}^{d}\sum_{m=1}^{n_{\text{bins}}}\left(\bar{L}_{i,m}^{(h)} - L_{i,m}^{(h)}\right)^2}{\sum_{i=1}^{d}\sum_{m=1}^{n_{\text{bins}}}\left(L_{i,m}^{(h)}\right)^2}}. \tag{2.37}$$

The denominator of `M1` is positive since we assume the training path is not trivial.

Second order: empirical covariance We are now interested in the dependence between the count of events of type i in a given bin $c_{j,n}^{(h)}$ and the count of events of type j in a bin $c_{i,n+l}^{(h)}$ that is distant from a lag l. Define the empirical averages

$$\bar{c}_{j,l,\text{lead}}^{(h)} := \frac{1}{n_{\text{bins}} - l}\sum_{n=1}^{n_{\text{bins}}-l} c_{j,n}^{(h)}, \quad \bar{c}_{i,l,\text{lag}}^{(h)} := \frac{1}{n_{\text{bins}} - l}\sum_{m=l+1}^{n_{\text{bins}}} c_{i,m}^{(h)}. \tag{2.38}$$

Definition 2.3.2 (Empirical covariance function)**.** *We define the empirical covariance matrix* $\mathcal{V}_l^{(h)}$ *as the squared matrix with entries* $(\mathcal{V}_{ij,l}^{(h)})_{i,j\in[d]}$*. The coefficient* $\mathcal{V}_{ij,l}^{(h)}$ *is the empirical covariance function from events of type* j *to events of type* i *at lag* l *with sampling period* h

$$\mathcal{V}_{ij,l}^{(h)} := \frac{1}{h(n_{bins}-l)} \sum_{n=1}^{n_{bins}-l} \left(c_{j,n}^{(h)} - \bar{c}_{j,l,lead}^{(h)}\right)\left(c_{i,n+l}^{(h)} - \bar{c}_{i,l,lag}^{(h)}\right) \tag{2.39}$$

If the counting process $\boldsymbol{N}$ is stationary, the empirical covariance $\mathcal{V}_l^{(h)}$ at lag $l = \lfloor \frac{\tau}{h} \rfloor$ is an unbiased estimator of the covariance function $\nu_\tau^{(h)}(0)$. This is not necessarily true for a non-stationary data-generating process like a non-homogeneous Poisson process (see Section 2.4) or an MTLH (see Section 3.3). To evaluate model fits, simulate paths from the conditional intensity model $\boldsymbol{\lambda}$, compute $\left(\bar{\mathcal{V}}_{ij,l}^{(h)}\right)$ the empirical covariance of a simulated path, and define

$$\texttt{M2} := \sqrt{\frac{\sum_{i,j\in[d]} \sum_{m=1}^{n_{\text{bins}}} \left(\bar{\mathcal{V}}_{ij,l}^{(h)} - \mathcal{V}_{ij,l}^{(h)}\right)^2}{\sum_{i,j\in[d]} \sum_{m=1}^{n_{\text{bins}}} \left(\mathcal{V}_{ij,l}^{(h)}\right)^2}}. \tag{2.40}$$

2.3.4 Out-of-sample testing

In any statistical estimation problem, out-of-sample (OOS) testing is a fundamental step to avoid over-fitting the training data. However, to the best of our knowledge, there is no standard methodology in the literature for OOS testing in the context of point process estimation. Suppose we are in one of the two estimation setups discussed above. Denote by $T' > 0$ the total observation horizon. Fix a training horizon $T < T'$; we want to fit a conditional intensity model using the training data on $[0, T]$, and test its OOS performance on $[T, T']$.

Metric We propose the following natural extension to the LSE.

Definition 2.3.3 (OOS LSE)**.** *The OOS LSE of a conditional intensity model is*

$$\mathcal{R}_{[T,T']}(\boldsymbol{\theta}) := \frac{1}{T'-T} \sum_{k=1}^{d} \int_T^{T'} \lambda_k(t)^2 \, \mathrm{d}t - \frac{2}{T'-T} \sum_{k=1}^{d} \sum_{m=N_T^k+1}^{N_{T'}^k} \lambda_k\left(t_m^k\right). \tag{2.41}$$

Using the Doob–Meyer decomposition of $\boldsymbol{N}$, this definition satisfies

$$\mathbb{E}\left[\mathcal{R}_{[T,T']}(\boldsymbol{\theta})\right] = \|\boldsymbol{\lambda} - \boldsymbol{\lambda}^\diamond\|_{[T,T']}^2 - \|\boldsymbol{\lambda}^\diamond\|_{[T,T']}^2. \tag{2.42}$$

Therefore, minimizing the expected OOS LSE $\mathbb{E}\left[\mathcal{R}_{[T,T']}(\boldsymbol{\theta})\right]$ is equivalent to minimizing the L_2 distance $\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^\diamond\|_{[T,T']}^2$. On the space of all conditional intensity models, the expected LSE is minimized by a conditional intensity model $\boldsymbol{\lambda}$ that satisfies

$$\lambda_k(t) = \lambda_k^\diamond(t) \quad a.s. \quad \text{for almost all} \quad t \in [T, T'], \quad k \in [d]. \tag{2.43}$$

We conveniently express the OOS LSE $\mathcal{R}_{[T,T']}$ based on the LSE at times T and T'.

Proposition 2.3.2. *The OOS LSE verifies*

$$\mathcal{R}_{[T,T']}(\boldsymbol{\theta}) = \frac{T'}{T'-T}\mathcal{R}_{T'}(\boldsymbol{\theta}) - \frac{T}{T'-T}\mathcal{R}_{T}(\boldsymbol{\theta}). \tag{2.44}$$

Finally, note that the moment metrics `M1` and `M2` above can also apply for OOS testing.

Projectibility of the model In this train-test framework, we illustrate the projectibility of the model in time. Consider a sample path from a uni-dimensional point process on $[0, T']$. We want to fit a piece-wise constant Poisson model to this data (see Section 2.4.2), with conditional intensity

$$\lambda_1(t) := b_0 \mathbb{1}_{[0,\beta_1)}(t) + b_1 \mathbb{1}_{[\beta_1,+\infty)}(t), \tag{2.45}$$

where the interval bounds β_1 is a hyper-parameter. The parameters of the model are the weights $\boldsymbol{b} := (b_0, b_1)$. If we set $\beta_1 = T$, it is clear that $\frac{\partial \mathcal{R}_T^{(k)}}{\partial b_1}(\boldsymbol{b}) = 0$ for all parameter values $\boldsymbol{b} > 0$. Because of the parameterization of the model, the training LSE $\mathcal{R}_T^{(k)}$ does not depend on the parameter b_1, and therefore this parameter cannot be learned in this setup. In a more general context, this type of issue is known as Goodman's paradox, or the new riddle of induction (see Henderson [51]). By construction, the projectibility issue does not appear for stationary models such as the homogeneous Poisson process (Section 2.4.1) or the MHP (Section 3.1). However, this is a problem for non-stationary models, such as non-homogeneous Poisson processes, and the MTLH model (Section 3.3). In the example above, we fix the projectibility issue simply by choosing $\beta_1 < T$. However, this highlights the need to take $T - \beta_1$ sufficiently large to observe enough data in that interval. Despite being a trivial model, piece-wise constant baselines have the disadvantage from a projectibility perspective that their parameters are not coupled at all in the case of a Poisson model, and not directly coupled in the case of an MTLH model.

2.4 A building block: the Poisson process

Deterministic conditional intensity models are called Poisson processes. This section first recalls definitions and general properties of Poisson processes, and introduces some objects of interest. We then discuss some classes of Poisson intensities, their simulation, and properties of their least-squares fit.

Definition 2.4.1 (Poisson process)**.** *Let $\boldsymbol{N}$ be a d-dimensional counting process with conditional intensity $\boldsymbol{\lambda}$. We say that $\boldsymbol{N}$ is a Poisson process if for all event types $i \in [d]$ and for all times $t \geq 0$,*

$$\lambda_i(t) = \mu_i(t), \tag{2.46}$$

where $\forall i \in [d]$, $\mu_i : [0, +\infty) \to (0, +\infty)$ *is differentiable by parts. We write in vector notation* $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^\intercal$. *If the function* μ_i *is constant for all event types* $i \in [d]$, *we say that* $\boldsymbol{N}$ *is a homogeneous Poisson process. Otherwise, we say that* $\boldsymbol{N}$ *is a non-homogeneous Poisson process.*

Of course, for event types $k \in [d]$, the compensator process Λ_k of a Poisson process is deterministic, and defined by $\Lambda_k(t) := \int_0^t \mu_k(s)ds$. We introduce the function M_k, the temporal average of the squared intensity μ_k, as it appears in the LSE of different model classes in this work.

Definition 2.4.2 (Mean squared intensity)**.** *For event types* $k \in [d]$, *and for times* $t > 0$, *the mean squared intensity* M_k *is*

$$M_k(t) := \frac{1}{t} \int_0^t \mu_k^2(t) \mathrm{d}t. \tag{2.47}$$

This function has a simple probabilistic interpretation. Using the Doob–Meyer decomposition of the counting process N^k,

$$M_k(t) = \mathbb{E}\left[\frac{1}{t} \int_0^t \mu_k(t) \mathrm{d}N_t^k \right]. \tag{2.48}$$

Poisson processes have independent increments; that is, for integers $p \in \mathbb{N}^*$, and for sorted collections of times $(\tau_m)_{m \in [p]}$, the $p-1$ random variables $\left(N^k_{\tau_{m+1}} - N^k_{\tau_m}\right)$ are independent. Furthermore, for times $s < t$, the increment $\left(N_t^k - N_s^k\right)$ follows a Poisson distribution with parameter

$$\Lambda_k(t) - \Lambda_k(s) = \int_s^t \mu_k(u) \mathrm{d}u. \tag{2.49}$$

The moments of Poisson processes have simple analytic expressions. The cumulative event rate process η^k verifies the following result.

Proposition 2.4.1 (Cumulative event rate of a Poisson process)**.** *For event types* $k \in [d]$, *and for times* $T > 0$, *the first and second moments of the cumulative event rate process are*

$$\mathbb{E}[\eta_T^k] = \frac{1}{T} \int_0^T \mu_k(t) \mathrm{d}t, \quad \mathrm{Var}[\eta_T^k] = \frac{1}{T^2} \int_0^T \mu_k(t) \mathrm{d}t. \tag{2.50}$$

We give a proof of this result in Appendix A.1.2. Furthermore, the independent second moments property implies a simple covariance matrix for the Poisson process.

Proposition 2.4.2 (Poisson covariance)**.** *Consider event types* $i, j \in [d]$ *with* $i \neq j$. *The covariance matrix of* $\boldsymbol{N}$ *at time* $t > 0$, *lag* $\tau \geq 0$, *and sampling period* $h > 0$ *is*

$$\nu_{ii,\tau}^{(h)}(t) = \left(\frac{1}{h} \int_{t+\tau}^{t+h} \mu_i(s) \mathrm{d}s \right) \mathbb{1}_{\{\tau \leq h\}}, \quad \nu_{ij,\tau}^{(h)}(t) = 0. \tag{2.51}$$

We give a proof of this result in Appendix A.1.2. The two main paradigms for the simulation of Poisson processes up to a horizon $T > 0$ are based on classic simulation methods for random variables.

- **Compensator inversion:** this is the point process counterpart of simulation by CDF inversion. The compensator Λ_k is monotonically increasing; denote by Λ_k^{-1} its generalised inverse. Cinlar [24] (Corollary 7.8) shows that if we simulate a path of a standard Poisson process (see next section) up to time $\Lambda_k(T)$, with event times $(\tau_1, \tau_2, \dots)$, then $\left(\Lambda_k^{-1}(\tau_1), \Lambda_k^{-1}(\tau_2), \dots\right)$ is a path the Poisson process with intensity μ_k.
- **Thinning:** this is the point process counter-part of acceptance-rejection simulation, proposed by Lewis and Shedler [60] (Theorem 1). Consider a Poisson process with conditional intensity $\mu_u \geq \mu_k$; denote by $(\tau_1, \tau_2, \dots)$ its event times. If at each event time τ_m, we reject this event with probability $1 - \mu_k u(\tau_m)/\mu_u(\tau_m)$, the resulting point process has conditional intensity μ_k.

We now discuss four classes of Poisson intensities. Section 2.4.1 discusses the homogeneous Poisson process, which is at the basis of the MHP model (Section 3.1). Non-homogeneous Poisson processes form the basis of the MTLH model (Section 3.3). For each class of Poisson intensity, we characterize its compensator, both for simulation purposes, and because this term appears in the expression of the MTLH LSE. We discuss the moments of the model, and its least-squares fit. In Section 2.4.2, we discuss the basic case of piece-wise constant Poisson processes. This is a simple extension of the homogeneous Poisson case, where several results are obtained in closed-form, and is also an interesting case as an upper bound of other classes of conditional intensities for simulation by thinning. Section 2.4.3 discusses periodic intensities, a useful modelling tool for systems with seasonality. Finally, Section 2.4.4 presents linear intensities as a case of unbounded intensity.

2.4.1 Constant intensities

Definition 2.4.3 (Homogeneous Poisson process)**.** *Let $\boldsymbol{N}$ be a d-dimensional counting process with conditional intensity $\boldsymbol{\lambda}$. We say that $\boldsymbol{N}$ is a homogeneous Poisson process if for all $i \in [d]$ and for all $t \geq 0$,*

$$\lambda_i(t) = \mu_i, \tag{2.52}$$

where $\forall i \in [d]$, $\mu_i > 0$. We write in vector notation $\boldsymbol{\mu} = (\mu_1, \dots, \mu_d)^\intercal$.

In the rest of this subsection, let $\boldsymbol{N}$ be a homogeneous Poisson process

Properties The increments of the homogeneous Poisson process are independent, stationary and follow an exponential distribution. For a horizon $T > 0$, and event types $i \in [d]$, the total event count N_T^i follows a Poisson distribution with distribution $\mu_i T$. The simulation of a homogeneous Poisson process with rate $\boldsymbol{\mu}$ up to a horizon $T > 0$ is well studied and relies on the previous two properties: start by sampling N_T^i following this distribution, then sample N_T^i points following a random uniform distribution on $[0, T]$. Using Proposition 2.4.1 and Proposition 2.4.2, moments of the homogeneous Poisson process have simple expressions.

Proposition 2.4.3 (Moments of the homogeneous Poisson process)**.** *Consider event types $i, j \in [d]$ with $i \neq j$. For times $T > 0$, the first and second moments of the cumulative event rate process are*

$$\mathbb{E}[\eta_T^i] = \mu_i, \quad \mathrm{Var}[\eta_T^i] = \frac{\mu_i}{T}. \tag{2.53}$$

The covariance matrix of the homogeneous Poisson process at time $t > 0$, lag $\tau \geq 0$, and sampling period $h > 0$ is

$$\nu_{ii,\tau}^{(h)}(t) = \mu_i f_{Tr}^{(h)}(\tau), \quad \nu_{ij,\tau}^{(h)}(t) = 0. \tag{2.54}$$

Least-squares fit Least squares estimation of homogeneous Poisson processes can be solved analytically, and leads to similar results than log-likelihood estimation. Consider the long path estimation setup with horizon $T > 0$, and fix an event type $k \in [d]$. For rates $\mu_k \geq 0$, the partial LSE of this model

$$\mathcal{R}_T^{(k)}(\mu_k) = \mu_k^2 - 2\eta_T^k \mu_k. \tag{2.55}$$

Proposition 2.4.4. *The LSE of the homogeneous Poisson model has variance*

$$\mathrm{Var}\big[\mathcal{R}_T^k(\mu_k)\big] = 4\mu_k^2 \mathrm{Var}[\eta_T^k]. \tag{2.56}$$

If the ground truth process is a (possibly non-homogeneous) Poisson model with bounded intensity, then the LSE is temporally consistent.

We give a proof of this result in Appendix A.1.2. It is clear that the minimizer of the LSE $\mu_k^\star$ verifies

$$\mu_k^\star = \eta_T^k, \qquad \mathcal{R}_T^{(k)}(\mu_k^\star) = -(\eta_T^k)^2. \tag{2.57}$$

2.4.2 Piece-wise constant intensities

Fix a number of intervals $J \in \mathbb{N}^*$. Consider a family of strictly positive reals, the interval intercepts $\boldsymbol{b} = (b_j)_{j \in [\![0, J-1]\!]} \in (0, +\infty)^J$. Finally, fix an ordered finite sequence of positive reals, the interval bounds $\boldsymbol{\beta} = (\beta_j)_{j \in [\![0, J]\!]} \in \bar{\mathbb{R}}_+^J$ such that:

$$\beta_0 = 0, \quad \beta_j < \beta_{j+1} \quad \forall j \in [\![0, J-1]\!], \quad \beta_J = +\infty. \tag{2.58}$$

Definition 2.4.4 (Piece-wise constant intensity)**.** *We say that a conditional intensity function μ is piece-wise constant if it satisfies*

$$\mu(t) := \sum_{j=0}^{J-1} b_j \mathbb{1}_{[\beta_j, \beta_{j+1})}(t). \tag{2.59}$$

For times $t > 0$, the mean-squared intensity M is

$$M(t) = \frac{1}{t}\Big(\sum_{j=0}^{g(t)-1} b_j^2(\beta_{j+1} - \beta_j) + b_{g(t)}^2(t - \beta_{g(t)}) \Big), \tag{2.60}$$

and for indices $j \in [\![0, J-1]\!]$, the derivatives of μ and M with respect to b_j are

$$\frac{\partial \mu}{\partial b_j}(t) = \mathbb{1}_{[\beta_j, \beta_{j+1})}(t), \qquad \frac{\partial M}{\partial b_j}(t) = \frac{1}{t}\Big(\mathbb{1}_{j < g(t)} 2 b_j (\beta_{j+1} - \beta_j) + \mathbb{1}_{j = g(t)} 2 b_{g(t)} (t - \beta_{g(t)}) \Big). \tag{2.61}$$

Properties Let $\boldsymbol{N}$ be a uni-dimensional non-homogeneous Poisson process with piece-wise constant conditional intensity μ. Define the ranking function g for times $t \geq 0$

$$g(t) := \min\{ j \in [\![0, J-1]\!], \beta_{j+1} > t \}. \tag{2.62}$$

Define the integral ranking function for $y > 0$

$$G(y) := \max\{ i \in [\![0, J-1]\!], \sum_{j=0}^{i-1} b_j(\beta_{j+1} - \beta_j) \leq y \}; \tag{2.63}$$

with the convention

$$G(y) = 0 \quad \text{if} \quad \{ i \in [\![0, J-1]\!], \sum_{j=0}^{i-1} b_j(\beta_{j+1} - \beta_j) \leq y \} = \emptyset. \tag{2.64}$$

Proposition 2.4.5 (Compensator)**.** *For times $t \geq 0$, the compensator of the counting process $\boldsymbol{N}$ is*

$$\Lambda(t) = \sum_{j=0}^{g(t)-1} b_j(\beta_{j+1} - \beta_j) + b_{g(t)}(t - \beta_{g(t)}). \tag{2.65}$$

For $y > 0$, the inverse compensator of $\boldsymbol{N}$ is

$$\Lambda^{-1}(y) = \beta_{G(y)} + \frac{1}{b_{G(y)}}(y - \sum_{j=0}^{G(y)-1} b_j(\beta_{j+1} - \beta_j)). \tag{2.66}$$

We get the second moment of $\boldsymbol{N}$ is available in closed form.

Proposition 2.4.6 (Covariance function)**.** *The covariance of the Poisson process at time $t > 0$, lag $\tau \geq 0$, and sampling period $h > 0$ is*

$$\nu_{11,\tau}^{(h)}(t) = b_{g(t+\tau)} f_{Tr}^{(h)}(\tau), \quad ifg(t+\tau) = g(t+h). \tag{2.67}$$

If $g(t+\tau) < g(t+h)$, then

$$\begin{aligned} \nu_{11,\tau}^{(h)}(t) = \mathbb{1}_{\{\tau<h\}} \Bigg(& \frac{\beta_{g(t+\tau)+1} - (t+\tau)}{h} b_{g(t+\tau)} + \sum_{p=g(t+\tau)+1}^{g(t+h)-1} \frac{\beta_{p+1} - \beta_p}{h} b_p \\ & + \Big(1 - \frac{\beta_{g(t+h)} - t}{h}\Big) b_{g(t+h)} \Bigg). \end{aligned} \tag{2.68}$$

Least-squares fit Consider the long path estimation setup with horizon $T > 0$. For intercept parameters $\boldsymbol{b} \geq 0$, the partial LSE of the piece-wise constant intensity model is

$$\mathcal{R}_T^{(1)}(\boldsymbol{b}) = \frac{1}{T} \sum_{j=0}^{J-2} b_j^2(\beta_{j+1} - \beta_j) - \frac{2}{T} \sum_{j=0}^{J-2} b_j N^k([\beta_j, \beta_{j+1})). \tag{2.69}$$

The analytic minimum of the LSE is analogous to the homogeneous Poisson case.

Proposition 2.4.7 (Minimal LSE of the piece-wise constant model)**.** *The minimizer $\boldsymbol{b}^\star$ of the LSE is*

$$b_j^\star = \frac{N^k([\beta_j, \beta_{j+1}))}{\beta_{j+1} - \beta_j}, \quad \forall j \in [\![0, J-1]\!]. \tag{2.70}$$

The minimal value of the LSE is

$$\mathcal{R}_T^{(k)}(\boldsymbol{b}^\star) = -\frac{1}{T} \sum_{j=0}^{J-2} \frac{\Big(N^k([\beta_j, \beta_{j+1}))\Big)^2}{\beta_{j+1} - \beta_j}. \tag{2.71}$$

As discussed in the example of over-fitting intensity in Section 2.3.4, Equation (2.71) implies that for all $\alpha > 0$, there exists $J \in \mathbb{N}^*$ and a subdivision $\boldsymbol{\beta}$ of size J such that

$\mathcal{R}_T^{(k)}(\boldsymbol{b}^\star) < -\alpha$. In fact, if $\boldsymbol{\lambda}_{\boldsymbol{k}}^{(\mathrm{nhp})}$ is the least-squares fit of a piece-wise constant non-homogeneous Poisson model to the data, and $\boldsymbol{\lambda}_{\boldsymbol{k}}^{(\mathrm{hp})}$ is the least-squares fit of a homogeneous Poisson model to the data, then

$$\mathcal{R}_T^{(k)}\left(\boldsymbol{\lambda}_{\boldsymbol{k}}^{(\mathrm{nhp})}\right) \leq \mathcal{R}_T^{(k)}\left(\boldsymbol{\lambda}_{\boldsymbol{k}}^{(\mathrm{hp})}\right). \tag{2.72}$$

2.4.3 Periodic intensities

Periodic intensities are particularly interesting for non-stationary systems with seasonality. First, we discuss the general setup of periodic Poisson intensities and computationally efficient exact simulation of these models, second, we present the example of cosine intensities.

2.4.3.1 General case

Fix a period $T_p > 0$. Let $f : \mathbb{R}_+ \to \mathbb{R}$ be a bounded, piece-wise continuous, T_p-periodic function.

Definition 2.4.5 (Periodic intensity)**.** *For times $t \geq 0$, the periodic conditional intensity function is*

$$\mu(t) := c + \alpha f(at + b), \tag{2.73}$$

where the parameters are the frequency $a > 0$, the phase $b > 0$, the re-scaling parameter $\alpha > 0$, and the intercept $c \geq 0$.

For times $t \geq 0$, the primitive of f is $F(t) := \int_0^t f(s)ds$. In the rest of this paragraph, consider a uni-dimensional Poisson process $\boldsymbol{N}$ with periodic conditional intensity μ.

Simulation Simulating paths of $\boldsymbol{N}$ up to a fixed time horizon $T > 0$ is critical for some of the Hawkes models used in this work. In this paragraph, we discuss exact simulation of periodic Poisson models by compensator inversion. Denote by Λ the compensator of $\boldsymbol{N}$. Let $D_h \subset \mathbb{R}_+^*$ be the set such that:

$$D_h = \begin{cases} \mathbb{R}_+^* & \text{if } f \geq 0, \\ \left(0, -\frac{1}{\min f}\right) & \text{otherwise.} \end{cases} \tag{2.74}$$

For all $\epsilon \in D_h$, define h_ϵ by

$$h_\epsilon \colon [0, T) \to [0, T + \epsilon F(T)); \quad \theta \mapsto \theta + \epsilon \int_0^\theta f(s)ds.$$

The function h_ϵ is continuous and strictly increasing, therefore it admits an inverse.

Proposition 2.4.8 (Inverse compensator - with intercept)**.** *For $y \geq 0$, the inverse compensator Λ^{-1} is*

$$\Lambda^{-1}(y) = \frac{kT_p + \theta - b}{a}; \tag{2.75}$$

where

$$\begin{aligned} k &= \left\lfloor \frac{a}{cT_p + \alpha F(T)} \left(y + \frac{\alpha F(b) + cb}{a} \right) \right\rfloor, \\ \theta &= h_{\frac{\alpha}{c}}^{-1} \left(\frac{a}{c} \left(y - \frac{cT + \alpha F(T)}{a} k + \frac{\alpha F(b) + cb}{a} \right) \right). \end{aligned} \tag{2.76}$$

Equation (2.76) requires computing the inverse of the function h_ϵ. While this inverse might be available in closed-form for some choices of periodic intensity μ, even simple cases require the use of numerical inversion methods. In this work, we use Newton's method to compute this inverse. We use a simple heuristic for the initialization of the method: let $y \in [0, T + \epsilon F(T))$ such that we want to compute $h_\epsilon^{-1}(y)$. First, fix a grid $\boldsymbol{\theta}_{\text{grid}} := \left(\theta_{\text{grid}}^{(k)}\right)_{k \in [0, n_{\text{grid}}]}$ on $[0, T]$, with the convention $\theta_{\text{grid}}^{(0)} = 0$, and $\theta_{\text{grid}}^{(n_{\text{grid}})} = T$. Evaluate the function h_ϵ on the grid to get $\boldsymbol{y}_{\text{grid}} = h_\epsilon(\boldsymbol{\theta}_{\text{grid}})$. Since h_ϵ is strictly increasing,

$$y_{\text{grid}}^k \leq y < y_{\text{grid}}^{k+1} \implies \theta_{\text{grid}}^k \leq h_\epsilon^{-1}(y) < \theta_{\text{grid}}^{k+1}. \tag{2.77}$$

Therefore, we sample our initial guess for $h_\epsilon^{-1}(y)$ uniformly at random in the interval $(\theta_{\text{grid}}^k, \theta_{\text{grid}}^{k+1})$. Note that if $f > 0$, we can set $c = 0$ while ensuring positivity of the conditional intensity $\mu > 0$. In this case, the inverse of the compensator is as follows.

Proposition 2.4.9 (Inverse compensator - without intercept)**.** *For $y \geq 0$, the inverse compensator Λ^{-1} is*

$$\Lambda^{-1}(y) = \frac{kT_p + \theta - b}{a}; \tag{2.78}$$

where

$$k = \left\lfloor \frac{a}{\alpha F(T)} \left(y + \frac{\alpha F(b)}{a} \right) \right\rfloor, \quad \theta = F^{-1} \left(\frac{a}{\alpha} \left(y - \frac{\alpha F(T)}{a} k + \frac{\alpha F(b)}{a} \right) \right). \tag{2.79}$$

Mean squared intensity We briefly give the formulas for the corresponding mean squared intensity M, which is essential for the LSE of this model and for the MTLH model in Section 3.3. Let $f_-(t) = \max(0, f(t))$. Define $m_f = |\min(f_-)|$. We parameterize the periodic intensity μ by (α, a, b, δ) such that

$$\mu(t) := \delta + \alpha m_f + \alpha f(at + b). \tag{2.80}$$

Let $c := \delta + \alpha m_f$. For $t \geq 0$, the primitive F_Q of f^2 is $F_Q(t) := \int_0^t f^2(s) ds$.

Proposition 2.4.10 (The mean-squared intensity M)**.** *For times $t > 0$, the mean-squared intensity M is*

$$M(t) = c^2 + \frac{2\alpha c}{at}\Big(kF(T_p) + F(\theta) - F(b)\Big) + \frac{\alpha^2}{at}\Big(kF_Q(T_p) + F_Q(\theta) - F_Q(b)\Big), \tag{2.81}$$

with

$$k := \left\lfloor \frac{at+b}{T_p} \right\rfloor, \quad \theta := at + b - kT_p. \tag{2.82}$$

The derivatives of M with respect to model parameters are

$$\begin{aligned}
\frac{\partial M}{\partial \alpha}(t) &= 2m_f c + \frac{2(\delta + 2\alpha m_f)}{at}\left(kF(T_p) + F(\theta) - F(b)\right) + \frac{2\alpha}{at}\left(kF_Q(T_p) + F_Q(\theta) - F_Q(b)\right); \\
\frac{\partial M}{\partial a}(t) &= \frac{1}{a^2 t}\Bigg(2\alpha c\left(atf(\theta) - kF(T_p) - F(\theta) + F(b)\right) \\
&\qquad + \alpha^2\left(atf^2(\theta) - kF_Q(T_p) - F_Q(\theta) + F_Q(b)\right)\Bigg); \\
\frac{\partial M}{\partial b}(t) &= \frac{1}{at}\left(2\alpha c\left(f(\theta) - f(b)\right) + \alpha^2\left(f^2(\theta) - f^2(b)\right)\right); \\
\frac{\partial M}{\partial \delta}(t) &= 2c + \frac{2\alpha}{at}\left(kF(T_p) + F(\theta) - F(b)\right).
\end{aligned} \tag{2.83}$$

2.4.3.2 Cosine intensity

We consider the case $f = \cos$, for which we get closed-form expressions of the previous results. In this case, the period of f is $T_p = 2\pi$, and $D_h = (0, 1)$. Fix $\epsilon \in (0, 1)$. We get

$$h_\epsilon(\theta) = \theta + \epsilon \sin(\theta) = \theta - \epsilon \sin(\pi + \theta). \tag{2.84}$$

As discussed previously, in order to compute the inverse compensator of the associated periodic Poisson model using our results, we need to invert the function h_ϵ. Here, this inversion is linked to solving Kepler's equation, a well-studied problem. Define the function

$$\mathrm{Kep}_\epsilon \colon \mathbb{R} \to \mathbb{R}; \quad \theta \mapsto \theta - \epsilon \sin\theta. \tag{2.85}$$

Hence for $\theta \in [0, T)$, $h_\epsilon(\theta) = \mathrm{Kep}_\epsilon(\pi + \theta) - \pi$. The function Kep_ϵ is continuous and strictly increasing, therefore it is invertible. Denote by $\mathrm{Kep}_\epsilon^{-1}$ the inverse function. For $y \in [0, 2\pi)$, the inverse of h_ϵ is

$$h_\epsilon^{-1}(y) = \mathrm{Kep}_\epsilon^{-1}(\pi + y) - \pi. \tag{2.86}$$

First, note that the inverse of Kep_ϵ on $[\pi, 3\pi)$ can be written as a function of $\mathrm{Kep}_\epsilon^{-1}|_{[0,\pi)}$ the restriction of the inverse function on $[0, \pi)$, since

$$\mathrm{Kep}_\epsilon^{-1}(y) = \begin{cases} 2\pi - \mathrm{Kep}_\epsilon^{-1}(2\pi - y) & \text{if } y \in [\pi, 2\pi), \\ 2\pi + \mathrm{Kep}_\epsilon^{-1}(y - 2\pi) & \text{if } y \in [2\pi, 3\pi). \end{cases} \tag{2.87}$$

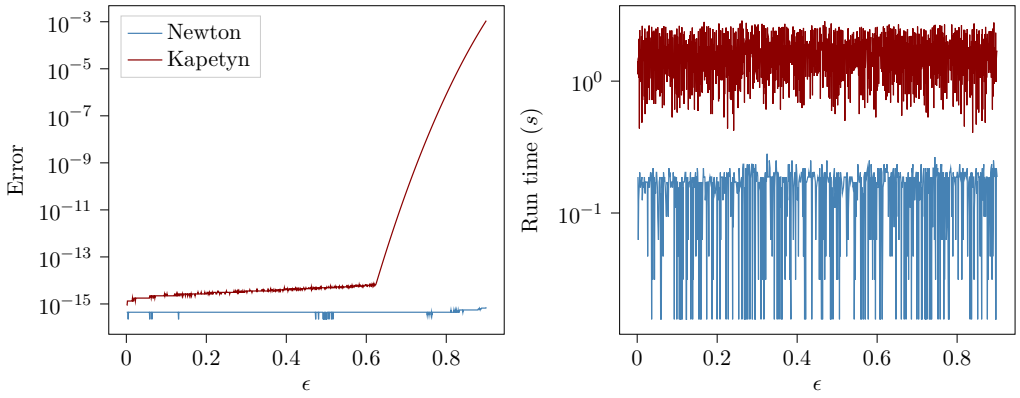
Therefore, we only need to invert Kep_ϵ on $[0, \pi)$, that is, for $y \in [0, \pi)$, we need to solve the equation

$$\theta - \epsilon \sin \theta = y, \tag{2.88}$$

in the unknown $\theta \in [0, \pi)$. This equation is known as the Kepler equation. This equation has a closed-form solution, the Kapetyn series

$$\theta = y + \sum_{n=1}^{\infty} \frac{2}{n} J_n(n\epsilon) \sin(ny), \tag{2.89}$$

where J_n is the $n - th$ Bessel function of the first kind. Figure 2.3 shows that the Newton scheme with adequate initialization is more accurate and faster than the evaluation of the Kapetyn series, validating our method. In particular, our method is robust to large values of ϵ close to 1, which is known to cause numerical instabilities. However, we cannot exclude that more sophisticated series acceleration techniques might improve the performance of the Kapetyn series approach.



(a) Accuracy (b) Speed

Figure 2.3: Comparison of numerical methods

Define a uniform grid G_ϵ for $\epsilon \in [10^{-3}, 0.9]$ with $n_\epsilon = 10^3$ values, and a uniform grid G_θ for $\theta \in [0, \pi - 10^{-2}]$ with $n_\theta = 10^6$ values. For $\epsilon \in G_\epsilon$, and for $\theta \in G_\theta$, compute $f_\epsilon\big(\mathrm{Kep}_\epsilon(\theta)\big)$, where f_ϵ approximates $\mathrm{Kep}_\epsilon^{-1}$ using the Newton method approach, then using the Kapetyn series approach. For each $\epsilon \in G_\epsilon$, measure the total run-time for the inversion of all the values in the grid $\theta \in G_\theta$, and compute the error $\max_{\theta \in G_\theta} |\theta - f_\epsilon\big(\mathrm{Kep}_\epsilon(\theta)|$.

2.4.4 Linear intensities

Definition 2.4.6 (Linear baseline)**.** *For $t \geq 0$, the linear baseline is*

$$\mu(t) = at + b, \tag{2.90}$$

and the parameters are the slope $a \geq 0$, and the intercept $b \geq 0$.

For times $t > 0$, the mean-squared intensity M is $M(t) = \frac{1}{3}a^2t^2 + abt + b^2$; and the derivatives of M with respect to model parameters are $\frac{\partial M}{\partial b}(t) = at + 2b$ and $\frac{\partial M}{\partial a}(t) = \frac{2}{3}at^2 + bt$.

Properties Let $\boldsymbol{N}$ be a uni-dimensional non-homogeneous Poisson process with piecewise constant conditional intensity μ.

Proposition 2.4.11 (Compensator)**.** *For times $t \geq 0$, the compensator of the Poisson process $\boldsymbol{N}$ is*

$$\Lambda(t) = a\frac{t^2}{2} + bt. \tag{2.91}$$

For $y \geq 0$, the inverse compensator Λ^{-1} is

$$\Lambda^{-1}(y) = \begin{cases} \frac{y}{b} & \text{if } a = 0 \text{ and } b > 0, \\ \frac{\sqrt{b^2+8ay}-b}{a} & \text{if } a > 0. \end{cases} \tag{2.92}$$

The second moment of this process has a simple analytic expression.

Proposition 2.4.12 (Covariance)**.** *Consider event types $i, j \in [d]$ with $i \neq j$. The covariance matrix of the Poisson process $\boldsymbol{N}$ at time $t > 0$, lag $\tau \geq 0$, and sampling period $h > 0$ is*

$$\nu^{(h)}_{11,\tau}(t) = \left(a\Big(t + \frac{h+\tau}{2}\Big) + b\right) f^{(h)}_{Tr}(\tau). \tag{2.93}$$

Least-squares fit Consider the long path estimation setup with horizon $T > 0$. Since this problem is in dimension $d = 1$, fix $k = 1$. For parameters $a, b \geq 0$, the partial LSE of the linear intensity model is

$$\mathcal{R}^{(k)}_T(a, b) = \frac{1}{3}a^2T^2 + abT + b^2 - 2a(\frac{1}{T}\sum_{m=1}^{N^k_T} t^k_m) - 2b\eta^k_T. \tag{2.94}$$

We want so solve the minimization program P_k defined by

$$\min_{b,a} \quad \mathcal{R}^{(k)}_T(b, a) \qquad \text{s.t.} \quad b \geq 0, \quad a \geq 0. \tag{2.95}$$

The problem P_k is a quadratic program with positivity constraints. For parameters $a, b \geq 0$, the Hessian of the partial LSE is

$$H_{a,b}\mathcal{R}^{(k)}_T = \begin{bmatrix} 2 & T \\ T & \frac{2}{3}T^2 \end{bmatrix}. \tag{2.96}$$

It is clear that this matrix is definite positive. Therefore, the LSE is strictly convex, and the estimation problem P_k admits a unique solution. Define

$$\tau_k := \frac{1}{TN^k_T}\sum_{m=1}^{N^k_T} t^k_m. \tag{2.97}$$

It is clear that $\tau_k \in (0,1)$. We express the regimes of solutions to problem P_k relatively to τ_k. Recall that the cumulative rate η_T^k is computed from the observed data and does not depend on the model.

Proposition 2.4.13. *The solutions $(b^\star, a^\star)$ to problem P_k depend on the value of τ_k. When normalized by $(\eta_T^k)^2$, the minimal value of the partial LSE $\frac{\mathcal{R}_T^{(k)}(b^\star,a^\star)}{(\eta_T^k)^2}$ only depends on τ_k. We distinguish three regimes of solutions.*

1. *If $\tau_k < \frac{1}{2}$, the solution to problem P_k is the homogeneous Poisson solution*

$$b^\star = \eta_T^k, \quad a^\star = 0. \tag{2.98}$$

In this case, the normalized minimal value of the LSE is

$$\frac{\mathcal{R}_T^{(k)}(b^\star, a^\star)}{(\eta_T^k)^2} = -1. \tag{2.99}$$

2. *If $\tau_k \in (\frac{1}{2}, \frac{2}{3})$, the solution to problem P_k is the affine Poisson solution*

$$b^\star = \frac{6}{T}\Big(\frac{2}{3}N_T^k - \frac{1}{T}\sum_{m=1}^{N_T^k} t_m^k\Big), \quad a^\star = \frac{12}{T^2}\Big(\frac{1}{T}\sum_{m=1}^{N_T^k} t_m^k - \frac{N_T^k}{2}\Big). \tag{2.100}$$

In this case, the normalized minimal value of the LSE is

$$\frac{\mathcal{R}_T^{(k)}(b^\star, a^\star)}{(\eta_T^k)^2} = -9\Big(\frac{2}{3} - \tau_k\Big)^2 - 3\Big(\tau_k\Big)^2. \tag{2.101}$$

3. *If $\tau_k \in (\frac{2}{3}, 1)$, the solution to problem P_k is the linear Poisson solution*

$$b^\star = 0, \quad a^\star = \frac{3}{T^3}\sum_{m=1}^{N_T^k} t_m^k. \tag{2.102}$$

In this case, the normalized minimal value of the LSE is

$$\frac{\mathcal{R}_T^{(k)}(b^\star, a^\star)}{(\eta_T^k)^2} = -3(\tau_k)^2. \tag{2.103}$$

We give a proof of this result in Appendix A.1.2.

Chapter 3

Hawkes models

3.1 Multivariate Hawkes process

This section is focused on the multivariate Hawkes process (MHP), the fundamental class of linear Hawkes models. Section 3.1.1 recalls the standard definition of MHP models in the literature and some of their properties that we use in this work: the branching representation of the MHP, and its first and second order moments. We show that uni-dimensional MHP models are path-wise identifiable, and formalize the parameterization of these models. Section 3.1.2 shows properties of the uni-dimensional MHP ($d = 1$), and Section 3.1.3 is focused on the multi-dimensional MHP ($d \geq 2$).

3.1.1 Properties

We define the MHP model as in Liniger [61].

Definition 3.1.1 (MHP)**.** *Let $\boldsymbol{N}$ be a d-dimensional orderly counting process with conditional intensity $\boldsymbol{\lambda}$. We say that $\boldsymbol{N}$ is a (linear) MHP if for all $i \in [d]$ and for all $t \geq 0$,*

$$\lambda_i(t) = \mu_i + \sum_{j=1}^{d} \sum_{\{m : t_m^j < t\}} \phi_{ij}(t - t_m^j), \tag{3.1}$$

where

- *$\forall i, j \in [d]$, $\phi_{ij} : [0, +\infty) \to [0, +\infty)$ is right-continuous and in L_1. The functions ϕ_{ij} are called the kernels of the MHP, and we write in matrix notation $\Phi(\cdot) = \big(\phi_{ij}(\cdot)\big)_{ij}$.*
- *$\forall i \in [d]$, $\mu_i > 0$. The scalars μ_i are called baseline intensities, and we write in vector notation $\boldsymbol{\mu} = (\mu_1, \dots, \mu_d)^\intercal$.*

We refer to such a process as a $(\boldsymbol{\mu}, \Phi)$-MHP, and to the pair $(\boldsymbol{\mu}, \Phi)$ as model functions.

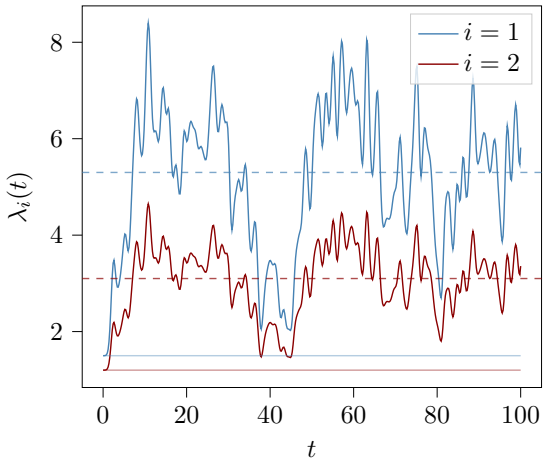


Figure 3.1: Conditional intensity of a bi-dimensional MHP

Consider a bi-dimensional ($d = 2$) MHP model with Gaussian kernels. We simulate a path of this MTLH up to $T = 100$. For times $t \in [0, T]$, and for event types $i \in [d]$, the thick solid lines plots the conditional intensities $\lambda_i(t)$ given the simulated path, the thin solid lines plot the baselines μ_i, and the dashed horizontal lines plot the stationary regime intensities $\eta_\star^{(i)}$.

Figure 3.1 plots an example of MHP conditional intensity. Without loss of generality, we assume in the rest of this work that for all event types $i, j \in [d]$, the kernel ϕ_{ij} is a linear combination of r_{ij} probability density functions $(\tilde{\phi}_{ij,l})_{l \in [r_{ij}]}$, of the form

$$\phi_{ij} = \sum_{l=1}^{r_{ij}} \omega_{ij,l} \tilde{\phi}_{ij,l}. \tag{3.2}$$

For all mixture indices $l \in [r_{ij}]$, the function $\tilde{\phi}_{ij,l} : [0, +\infty) \to [0, +\infty)$ is differentiable by parts and normalized to $\|\tilde{\phi}_{ij,l}\|_1 = 1$. The function $\tilde{\phi}_{ij,l}$ is effectively a PDF; in this work we call $\tilde{\phi}_{ij,l}$ a kernel density, and refer to the random variables with PDF $\tilde{\phi}_{ij,l}$ as the random offsets $\tilde{\tau}_{ij,l}$. As discussed below, these random offsets play a crucial role in the branching representation of MHP, and therefore in the branching simulation of this type of process. Denote by $\tilde{\psi}_{ij,l}$ the associated CDF, defined for lags $t \geq 0$ by

$$\tilde{\psi}_{ij,l}(t) := \int_0^t \tilde{\phi}_{ij,l}(u) \mathrm{d}u. \tag{3.3}$$

The weights $\omega_{ij,l} \geq 0$ control the L_1 norm of the linear combination. Define $\omega_{ij} = \sum_{l=1}^{r_{ij}} \omega_{ij,l}$; it is clear that $||\phi_{ij}||_1 = \omega_{ij}$. Forcing the mixture weights $(\omega_{ij,l})_{l \in [r_{ij}]}$ to be non-negative ensures that the conditional intensity of the MHP is almost surely non-negative at all times. Finally, to simplify notation, for all mixture indices $l \in [r_{ij}]$, we denote $\phi_{ij,l} := \omega_{ij,l} \tilde{\phi}_{ij,l}$, and $\psi_{ij,l} := \omega_{ij,l} \tilde{\psi}_{ij,l}$. It is clear that the primitive of kernel ϕ_{ij} is the function $\psi_{ij} := \sum_{l=1}^{r_{ij}} \psi_{ij,l}$.

3.1.1.1 Branching representation

Below, we see that the matrix $\|\Phi\|_1 := (\|\phi_{ij}\|_1)_{i,j\in[d]}$ plays a special role for the MHP. A heuristic way to see the MHP is to imagine that events are generated following two mechanisms:

- First, events arrive following a homogeneous Poisson process of rate $\boldsymbol{\mu}$. These events are usually referred to in the literature as "immigrants".
- Then, apply the following procedure to each event. Fix an event and let $j \in [d]$ denote the type of this event. For each event type $i \in [d]$, if the kernel ϕ_{ij} is not null, the fixed parent gives rise to "first generation descendants" of type i. The number of descendants of type i follows a Poisson distribution of mean $\|\phi_{ij}\|_1$. The time offsets between a parent event and its descendants are i.i.d.and follow the distribution $\frac{\phi_{ij}}{\|\phi_{ij}\|_1}$. This procedure is then applied iteratively to each descendant.

We illustrate this procedure in Figure 3.2 for a bi-dimensional MHP. The branching repre-

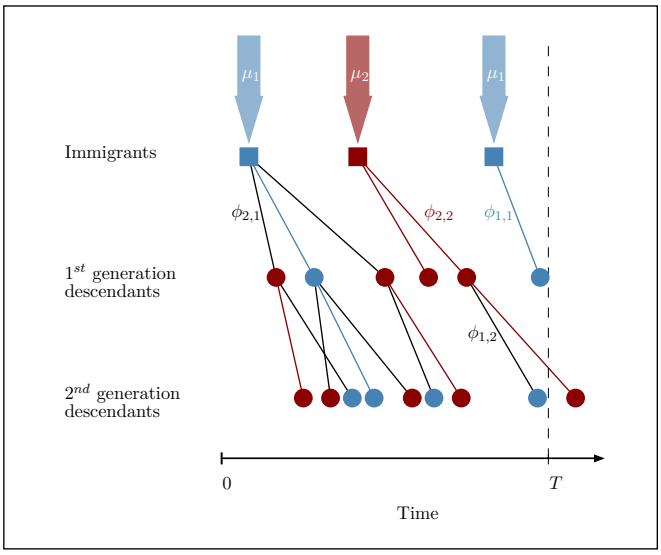


Figure 3.2: Branching representation of an MHP.
Events of type 1 (resp. 2) correspond to the blue (resp. red) nodes. Immigrants of type 1 (resp. type2) arrive with a Poisson rate μ_1 (resp. μ_2) and correspond to the squares. The descendants correspond to the circles.

sentation of Hawkes processes gives rise to an efficient simulation algorithm.

Cluster probabilities The inverse problem, that is, given a path of Hawkes process, identify which events are immigrants, and which events triggered each other, is referred to in the literature as stochastic declustering. It is not possible to infer exactly the branching structure from a path of an MHP, however, its distribution is known. Formally, let $i, j \in [d]$,

$m \in [N_T^i]$ and $n \in [N_T^j]$. We denote by $p_{i,m,j,n}$ the probability that the m-th jump of N^i is triggered by the n-th jump of N^j. Then

$$p_{i,m,j,n} = \begin{cases} \frac{\phi_{ij}(t_m^i - t_n^j)}{\lambda_i(t_m^i)} & \text{if } t_m^i > t_n^j, \\ 0 & \text{otherwise.} \end{cases} \tag{3.4}$$

We denote by $p_{i,m,m}$ the probability that the m-th jump of N^i is a background event of N^i. Then

$$p_{i,m,m} = \frac{\mu_i}{\lambda_i(t_m^i)}. \tag{3.5}$$

Stability We denote by $\mathcal{A} \in \mathcal{M}_d(\mathbb{R})$ the matrix such that for all $i, j \in [d]$, $\mathcal{A}_{ij}$ is the expected number of direct descendants of type i from a parent of type j. For branching processes, this matrix is usually referred to as the adjacency matrix of the process. The spectral radius $\rho(\mathcal{A})$ is referred to in the literature as the branching ratio of the process.

We say that a branching process is stable if any event has an almost surely finite number of descendants. A branching process is stable if and only if $\rho(\mathcal{A}) < 1$. In that case, the expected number of descendants of a given event is $\frac{1}{1-\rho(\mathcal{A})}$. A stable branching process with branching ratio close to 1 is usually called near-critical.

In the case of an MHP, $\mathcal{A} := \|\Phi\|_1$. For certain fields of applications of MHP to real-life systems, the branching ratio of fitted models appears to be consistently near-critical when fitting different MHP to different datasets. In that case, this might be reported in that specific literature as a stylized fact of that system. For instance, this is the case when modelling the mid-price in high frequency financial data using MHP. In the literature on the estimation of MHP, the stability condition is usually not explicitly imposed as an optimization constraint due to its intractability. However, it might be an implicit assumption in some methods, for example the ones that rely on stationarity properties of the MHP such as in Bacry and Muzy [10].

There is no simple analytic characterization of stability in terms of the coefficients of the adjacency matrix, which are usually the parameters of direct interest to the modeller. but we can give some necessary conditions.

Proposition 3.1.1 (Necessary conditions for stability)**.** *A necessary condition for stability is that diagonal coefficients of the adjacency matrix are smaller than* 1*. Formally, we get*

$$\rho(||\Phi||_1) < 1 \implies \omega_{ii} < 1 \quad \forall i \in [d]. \tag{3.6}$$

Furthermore, if $d \geq 2$ *and the matrix* $||\Phi||_1$ *is symmetric with strictly positive coefficients,*

$$\rho(||\Phi||_1) < 1 \implies \omega_{ij} < 1 \quad \forall i, j \in [d]. \tag{3.7}$$

We give a proof of this result in Appendix A.2.3. These necessary conditions have a simple interpretation from a branching representation perspective. The first condition asserts that for an MHP with any given adjacency matrix, the total number of descendants of type $i \in [d]$ with parents exclusively of type i is finite. The second condition asserts that for an MHP with a symmetric adjacency matrix with strictly positive coefficients, alternating cycles of descendants leads to a finite total descendance on expectation: for event types $i, j \in [d]$ such that $i \neq j$, starting from an element of type i, the total number of its descendants (direct and indirect) such that we only count events of type i (resp. j) that have a parent of type j (resp. i) is finite on expectation. However in general, stability does not imply that non-diagonal elements must be smaller than 1.

Remark 3.1.1 (Dirac distributions are not valid MHP kernels)**.** *Note that by definition, Dirac delta functions cannot be used as kernel densities for an MHP as they are not valued in* $[0, +\infty)$*. In fact, if we consider an MHP with Dirac kernels, then we can no longer guarantee that event times are almost surely distinct. For example, consider an MHP model and suppose there exists event types* $i, j \in [d]$ *such that* $\tilde{\phi}_{ij} = \omega_{ij}\tilde{\delta}_a$*, with* $\omega_{ij} > 0$*, and* $a \geq 0$*. Let* $n \in \mathbb{N}^*$*, and consider the direct descendants of the event* t_n^j *of type* i*. If* $a = 0$*, then any type* i *descendant of* t_n^j *will occur at* $t_m^i = t_n^j$*. If* $a > 0$*, then if* t_n^j *has* 2 *or more descendants, they occur at the same time.*

Exogeneity ratio Consider a stable $(\boldsymbol{\mu}, \Phi)$-MHP $\boldsymbol{N}$. Suppose we generate paths of this MHP up to a terminal horizon $T > 0$ using its branching representation. For any given path, and for event types $i \in [d]$, the baseline events of type i are generated using a homogeneous Poisson $N^{(\mu_i)}$ process of rate μ_i. We are interested in two types of ratios representing the asymptotic proportion of baseline events.

Definition 3.1.2 (Exogeneity ratio of MHP)**.** *For event types* $i \in [d]$*, define the type* i *exogeneity ratio as the asymptotic proportion of baseline events of type* i*, that is*

$$\texttt{exo}_i = \lim_{T \to +\infty} \frac{N_T^{(\mu_i)}}{N_T^i}. \tag{3.8}$$

By construction, $\texttt{exo}_i \in [0, 1]$*; define the type* i *endogeneity ratio as* $\texttt{endo}_i = 1 - \texttt{exo}_i$*.*

Define the total exogeneity ratio as the asymptotic proportion of baseline events, that is

$$\texttt{exo} = \lim_{T \to +\infty} \frac{\sum_{i=1}^d N_T^{(\mu_i)}}{\sum_{i=1}^d N_T^i}. \tag{3.9}$$

By construction, $\texttt{exo} \in [0, 1]$*; define the total endogeneity ratio* $\texttt{endo} = 1 - \texttt{exo}$*.*

Using the law of large numbers for MHP (see Section 3.1.1.2), for event types $i \in [d]$, the type i exogeneity ratio verifies

$$\texttt{exo}_i = \frac{\mu_i}{[(\mathbb{I}_d - ||\Phi||_1)^{-1}\boldsymbol{\mu}]_i}; \tag{3.10}$$

and the total exogeneity ratio verifies

$$\texttt{exo} = \frac{||\boldsymbol{\mu}||_1}{||(\mathbb{I}_d - ||\Phi||_1)^{-1}\boldsymbol{\mu}||_1} = \frac{1}{||(\mathbb{I}_d - ||\Phi||_1)^{-1}\frac{\boldsymbol{\mu}}{||\boldsymbol{\mu}||_1}||_1}. \tag{3.11}$$

This ratio depends on the normalised baseline vector $\frac{\boldsymbol{\mu}}{||\boldsymbol{\mu}||_1}$, this implies that in the uni-dimensional case (see below), the total exogeneity ratio does not depend on the baseline parameter.

3.1.1.2 First order moment

Unless mentioned otherwise, we assume in this section that the MHP $\boldsymbol{N}$ is stable. Bordenave and Torrisi [16] give a law of large numbers (LLN) result for MHP

$$\lim_{T \to +\infty} \boldsymbol{\eta_T} = (\mathbb{I}_d - ||\Phi||_1)^{-1}\boldsymbol{\mu} \quad a.s., \tag{3.12}$$

and study large deviations from this limit. We use the notation

$$\boldsymbol{\eta_\star} := \lim_{T \to +\infty} \boldsymbol{\eta_T}, \tag{3.13}$$

and refer to $\boldsymbol{\eta_\star}$ as the stationary regime intensity (SRI) of the MHP $\boldsymbol{N}$. This result implies that the cumulative event rate process $\boldsymbol{\eta}$ is temporally consistent. Of course, this first order property is insufficient to fully characterize the MHP, as the map

$$(\boldsymbol{\mu}, ||\Phi||_1) \mapsto (\mathbb{I}_d - ||\Phi||_1)^{-1}\boldsymbol{\mu}, \tag{3.14}$$

is clearly not injective. In fact, Bacry and Muzy [10] show that the additional use of a second order statistic, the auto-covariance of the process, is necessary to fully characterize an MHP. Furthermore, the stationary regime intensity $\boldsymbol{\eta_\star}$ does not depend on the kernel densities $(\tilde{\phi}_{ij})_{i,j \in [d]}$.

Interpretation To understand Equation (3.12), note that since the MHP $\boldsymbol{N}$ is stationary, the spectral radius of $||\Phi||_1$ is smaller than one. Therefore, the matrix $\mathbb{I}_d - ||\Phi||_1$ is invertible, and its inverse is the power series

$$(\mathbb{I}_d - ||\Phi||_1)^{-1} = \sum_{p=0}^{+\infty} ||\Phi||_1^p. \tag{3.15}$$

In particular, this implies that the matrix $(\mathbb{I}_d - ||\Phi||_1)^{-1}$ has non-negative coefficients. Therefore, the vector

$$\boldsymbol{\eta_\star} = (\mathbb{I}_d - ||\Phi||_1)^{-1}\boldsymbol{\mu}, \tag{3.16}$$

has non-negative coefficients. Furthermore, the stationary regime intensity $\boldsymbol{\eta_\star}$ has another interpretation. Consider the branching representation of the MHP $\boldsymbol{N}$. For all event types $i \in [d]$, denote by $N^{(\mu_i)}$ the associated baseline homogeneous Poisson process. For all times $t > 0$, $N_t^i \leq N_t^{(\mu_i)} \quad a.s.$. Now let $\hat{N}_t^i$ be the process constructed by counting the number of baseline events up to time t, $N_t^{(\mu_i)}$, and adding all their descendants on $[0, +\infty)$. By construction, this process upper-bounds almost surely the original counting process $N_t^i \leq \hat{N}_t^i \quad a.s.$. Denote by $\hat{\boldsymbol{N}}_{\boldsymbol{t}}$ the vector process

$$\hat{\boldsymbol{N}}_{\boldsymbol{t}}^{\intercal} := (\hat{N}_t^1, \dots, \hat{N}_t^d). \tag{3.17}$$

We want to compute the expectation $\mathbb{E}[\hat{\boldsymbol{N}}_{\boldsymbol{t}}]$. The expected number of baseline events on $[0, t]$ is $\mu_i t$. Therefore, it is clear that the expected number of total descendants of these baseline events is

$$\sum_{p=1}^{+\infty} ||\Phi||_1^p \boldsymbol{\mu} t = ||\Phi||_1 \sum_{p=0}^{+\infty} ||\Phi||_1^p \boldsymbol{\mu} t. \tag{3.18}$$

Hence, we get the following formula for the expectation

$$\mathbb{E}[\hat{\boldsymbol{N}}_{\boldsymbol{t}}] = t\boldsymbol{\mu} + t\sum_{p=1}^{+\infty} ||\Phi||_1^p \boldsymbol{\mu} = t\sum_{p=0}^{+\infty} ||\Phi||_1^p \boldsymbol{\mu} = t(\mathbb{I}_d - ||\Phi||_1)^{-1}\boldsymbol{\mu}. \tag{3.19}$$

This implies $\mathbb{E}\left[\frac{1}{t}\hat{\boldsymbol{N}}_{\boldsymbol{t}}\right] = \boldsymbol{\eta_\star}$. Therefore, the stationary regime intensity $\boldsymbol{\eta_\star}$ is also the expected number of baseline events on $[0, t]$ and their expected number of total descendants on $[0, +\infty]$. Since $\mathbb{E}[\boldsymbol{\eta_t}] \leq \mathbb{E}[\frac{1}{t}\hat{\boldsymbol{N}}_{\boldsymbol{t}}]$, this implies an upper bound for the expectation of the cumulative rate process $\boldsymbol{\eta_t}$.

Corollary 3.1.1 (Cumulative rate and SRI)**.** *For times $t \geq 0$, $\mathbb{E}[\boldsymbol{\eta_t}] \leq \boldsymbol{\eta_\star}$.*

First order characterization Fix a value of stationary regime intensity $\boldsymbol{\eta_\star} > 0$. Define the set of stable MHP with stationary regime intensity $\boldsymbol{\eta_\star}$ as

$$\mathbb{F}_{\text{MHP}}(\boldsymbol{\eta_\star}) := \left\{(\boldsymbol{\mu}, ||\Phi||_1) : \quad \rho(||\Phi||_1) < 1, \quad (\mathbb{I}_d - ||\Phi||_1)^{-1}\boldsymbol{\mu} = \boldsymbol{\eta_\star}\right\}. \tag{3.20}$$

This set consists of all stable MHP with the same asymptotic rate of events. However, since this first order statistic is not sufficient to characterize fully the MHP, all the elements of this set present different trade-offs between exogenous and endogenous events.

In general, it is difficult to get an analytic characterization of $\mathbb{F}_{\text{MHP}}(\boldsymbol{\eta_\star})$ in terms of the coefficients of $(\boldsymbol{\mu}, ||\Phi||_1)$. First, it clear that this set is non-empty. For example, consider a

sequence $\alpha_n \in [0,1)^{\mathbb{N}}$, and for all $n \in \mathbb{N}$, define the element $f_n := (\boldsymbol{\mu}^{(n)}, ||\Phi||_1^{(n)})$ where for all event types $i, j \in [d]$ such that $i \neq j$

$$\mu_i^{(n)} = \alpha_n \eta_\star^{(i)}, \quad \omega_{ii}^{(n)} = 1 - \alpha_n, \quad \omega_{ij}^{(n)} = 0. \tag{3.21}$$

It is clear that for all $n \in \mathbb{N}$, $f_n \in \mathbb{F}_{\text{MHP}}(\boldsymbol{\eta_\star})$. We note two particular elements in the boundary (for the Euclidean topology) of $\mathbb{F}_{\text{MHP}}(\boldsymbol{\eta_\star})$:

- the homogeneous Poisson model with $\boldsymbol{\mu} = \boldsymbol{\eta_\star}$ and no endogeneity, *i.e.* $||\Phi||_1 = 0$.
- the degenerate model with $\boldsymbol{\mu} = \mathbf{0}$, and $||\Phi||_1 = \mathbb{I}_d$; which is equivalent to a null counting process: any baseline event would almost surely lead infinitely many descendants in finite time, but since $\boldsymbol{\mu} = \mathbf{0}$ there is almost surely no baseline event happening. Note that this element is the limit of $(f_n)_{n\in\mathbb{N}}$ for $(\alpha_n) = (\frac{1}{n+1})$.

In dimension $d > 1$, there exist stable adjacency matrices $||\Phi||_1$ such that for any baseline $\boldsymbol{\mu} \geq \mathbf{0}$, $(\boldsymbol{\mu}, ||\Phi||_1) \notin \mathbb{F}_{\text{MHP}}(\boldsymbol{\eta_\star})$. Not all stable adjacency matrices lead to non-negative baselines by inverting the LLN using

$$\boldsymbol{\mu} = (\mathbb{I}_d - ||\Phi||_1)\boldsymbol{\eta_\star}. \tag{3.22}$$

For instance, take $d = 2$, and consider an adjacency matrix

$$||\Phi||_1 = \begin{bmatrix} 0.5 & 0.4 \\ 0.25 & 0.7 \end{bmatrix}. \tag{3.23}$$

This adjacency matrix is stable since the associated branching ratio is $\rho(||\Phi||_1) = 0.93$. Consider a positive vector $\boldsymbol{\eta_\star} = (1.5, 2)^\intercal$. Then

$$(\mathbb{I}_d - ||\Phi||_1)\boldsymbol{\eta_\star} = \begin{bmatrix} -0.05 \\ 0.225 \end{bmatrix}. \tag{3.24}$$

The first component of the vector $(\mathbb{I}_d - ||\Phi||_1)\boldsymbol{\eta_\star}$ is negative. Therefore, for a given stable adjacency matrix $||\Phi||_1$ in dimension $d > 1$, not all stationary regime intensities $\boldsymbol{\eta_\star}$ are achievable. Therefore, given a vector $\boldsymbol{\eta} > 0$, and an adjacency matrix $||\Phi||_1$, we refer to the inequality

$$(\mathbb{I}_d - ||\Phi||_1)\boldsymbol{\eta} \geq 0 \tag{3.25}$$

as the LLN condition. If this condition is satisfied, then

$$\Big((\mathbb{I}_d - ||\Phi||_1)\boldsymbol{\eta}, ||\Phi||_1\Big) \in \mathbb{F}_{\text{MHP}}(\boldsymbol{\eta}). \tag{3.26}$$

Note that in the uni-dimensional case $d = 1$, it is clear that

$$\rho(||\Phi||_1) < 1 \implies (\mathbb{I}_d - ||\Phi||_1)\boldsymbol{\eta_\star} > 0, \quad \forall \boldsymbol{\eta_\star} > 0. \tag{3.27}$$

Estimating the stationary regime intensity The value of the stationary regime intensity $\boldsymbol{\eta_\star}$ is often used by moment methods for the estimation of MHP, such as in Bacry et al. [5]. In this work, we use $\boldsymbol{\eta_\star}$ as a heuristic for the initialization of MHP parameters in our iterative method. However, in MHP estimation, under the hypothesis that the ground truth is a stable MHP, the value of $\boldsymbol{\eta_\star}$ is not directly observed. Therefore, in order to use of the LLN as a heuristic, we face the practical question of estimating the stationary regime intensity $\boldsymbol{\eta_\star}$. The data available for this task depends on the estimation setup.

- In the long path setup: we observe one path of the cumulative event rate process $(\boldsymbol{\eta_t})_{t\in[0,T]}$, where the observation window T is usually large. Since the LLN implies that, under the hypothesis that the ground truth is a stable MHP, the cumulative rate $\boldsymbol{\eta}$ is temporally consistent, a natural question arising is whether the window T is large enough for $\boldsymbol{\eta_T}$ to be close enough to the stationary intensity $\boldsymbol{\eta_\star}$.
- In the episodic setup: we observe n_{paths} paths of the event rate $(\boldsymbol{\eta_t})_{t\in[0,T]}$, usually with a small observation window T. For all $t \in [0, T]$, we can approximate the expected event rate $\mathbb{E}[\boldsymbol{\eta_t}]$ with a standard unbiased Monte Carlo estimator over the episodes

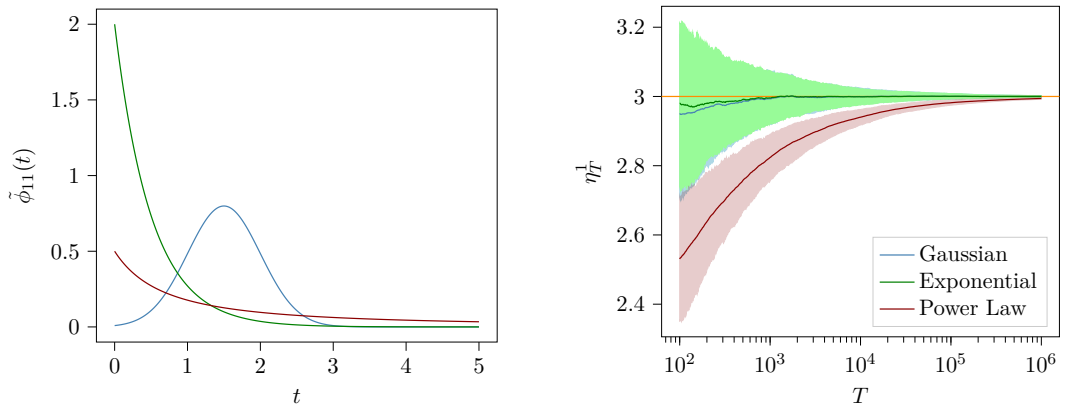
$$\frac{1}{n_{\text{paths}}} \sum_{p=1}^{n_{\text{paths}}} \boldsymbol{\eta_t^{(p)}}. \tag{3.28}$$

  As discussed in Corollary 3.1.1, for all times $t > 0$, the descendants of some events do not fall into the observation window $[0, t]$. Because of this boundary effect, the expected rate $\mathbb{E}[\boldsymbol{\eta_t}]$ under-estimates the stationary intensity $\boldsymbol{\eta_\star}$: $\mathbb{E}[\boldsymbol{\eta_t}] \leq \boldsymbol{\eta_\star}$.

We propose two numerical experiments to illustrate how for different MHPs with the same value of stationary regime intensity $\boldsymbol{\eta_\star}$, the cumulative rate process $\boldsymbol{\eta_t}$ and its expectation have different dynamics and dependencies.

For the first experiment, consider 3 uni-dimensional MHP models: with Gaussian, exponential and power law kernels. We fix the same baseline and the L_1 weight values for all 3 models, that is $\mu_1 = 1.5$, $\omega_{11} = 0.5$. Therefore, all 3 models have the same stationary regime intensity $\eta^1_\star = 3$. For each model, we simulate $n_{\text{paths}} = 10^3$ paths up to $T_{\max} = 10^6$. We compute η^1_T for each path, and for each horizon $T \in [10^2, 10^6]$ (using a logarithmic discretization of this interval with 10^3 points). Figure 3.3 plots the different kernel densities and the mean and percentiles of the distribution of η^1_T for each model, at each time T. First, as expected, we note that for all 3 models, if the observation window T is small, the expected event rate $\mathbb{E}[\boldsymbol{\eta_T}]$ strictly under-estimates the stationary regime intensity $\boldsymbol{\eta_\star}$, and the variance of the distribution of $\boldsymbol{\eta_T}$ seems to decrease with time. Second, we note that although the baseline μ_1 and branching ratio ω_{11} fully characterize the asymptotic regime

intensity $\eta_\star^{(1)}$, they are not sufficient to fully characterize neither the dynamics of the cumulative rate process $\boldsymbol{\eta_T}$ nor the dynamics of its expectation, since the 3 MHP models have different means $\mathbb{E}[\boldsymbol{\eta_T}]$ and variances. This is coherent with the fact that, even if the baseline and branching ratio are fixed, the branching edge effect, that is, the number of events excluded at the edge, depends on the kernel densities $\tilde{\phi}_{ij}$. Intuitively, for the same baseline and branching ratio, an MHP with a heavy-tailed kernel density is more likely to have descendants falling outside of the observation window. Note that this is coherent with the order of magnitude of the difference $\eta_\star^{(1)} - \mathbb{E}[\eta_T^1]$ in this first experiment: for example, if $T \sim 10^2$, then excluding 1 to 5 events on average will under-estimate $\eta_\star$ by 0.01 to 0.05.



(a) Kernel densities

(b) Convergence of η_T^1

Figure 3.3: Empirical illustration of the LLN for differents MHPs

Left: For each one of the three uni-dimensional MHP models, we plot the kernel density $\tilde{\phi}_{11}$ against time. Right: for each model, solid lines plot the expected event rate $\mathbb{E}[\eta_T^1]$ against the window size T. We fill the area between the 25% and 75% empirical percentiles of the values of the rate η_T^1. The horizontal orange line is the theoretical value of the asymptotic intensity $\eta_\star^1$.

Furthermore, the exclusion of descendants at the edge is also driven by the branching ratio of the MHP, since this controls the number of events closer to the edge of the simulation interval. We highlight this with a second experiment. We now only use the previous Gaussian uni-variate MHP with the same kernel density parameters, but we vary the baseline and the branching ratio while maintaining the same stationary regime intensity as before, $\eta_\star^{(1)} = 3$. Figure 3.4 plots the mean and percentiles of the distribution of η_T^1 for each model, at each time T. The results of this experiment seem to imply that for fixed stationary regime intensity and fixed kernel density, When the branching ratio increases, the variance $\mathrm{Var}[\eta_T^1]$ and the error $\eta_\star^{(1)} - \mathbb{E}[\eta_T^1]$ both increase.

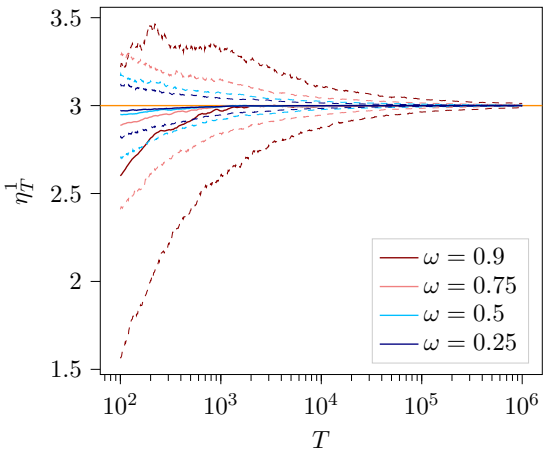


Figure 3.4: LLN.

We compute the total cumulative event rate η_T^1 for each simulated path from the models. Dashed lines plot the 25% and 75% percentile of these values; solid lines plot their empirical mean. In orange, we plot the theoretical value of the asymptotic.

3.1.1.3 Second order moment

Notation and conventions Let $f : [0, +\infty) \to \mathbb{R}$ be an integrable function. We denote by $\mathcal{F}[f]$ the Fourier transform of f, where for all frequencies $x \in \mathbb{R}$

$$\mathcal{F}[f](x) := \int_0^{+\infty} f(t) e^{-i2\pi xt} \mathrm{d}t. \tag{3.29}$$

By abuse of notation, if f is a matrix of functions, we denote by $\mathcal{F}[f]$ the matrix of Fourier transforms where $\mathcal{F}[f]_{ij} = \mathcal{F}[f_{ij}]$. We denote by $\mathcal{F}^{-1}$ the inverse Fourier transform, defined for all functions $g : \mathbb{R} \to \mathbb{R}$ and for all times $t \in \mathbb{R}$ by

$$\mathcal{F}^{-1}[g](x) := \int_{-\infty}^{+\infty} g(x) e^{i2\pi xt} \mathrm{d}x. \tag{3.30}$$

The triangle function $f_{\mathrm{Tr}}^{(h)}$ plays a special role in the expression of the covariance of MHP. Note that for all frequencies $x \geq 0$, the Fourier transform of the triangle function is

$$\mathcal{F}\left[f_{\mathrm{Tr}}^{(h)}\right](x) = h \,\mathrm{sinc}^2(\pi h x). \tag{3.31}$$

Furthermore, the following lemma is particularly useful

Lemma 3.1.1 (Convolution with a triangle)**.** *Consider a function $f : \mathbb{R} \to \mathbb{R}$. Denote by F a primitive of f. For all lags $\tau \geq 0$,*

$$f_{Tr}^{(h)} * f(\tau) = \frac{1}{h}\Big(\int_{\tau}^{\tau+h} F(y)\mathrm{d}y - \int_{\tau-h}^{\tau} F(y)\mathrm{d}y \Big). \tag{3.32}$$

We give a proof of this result in Appendix A.2.4.

For all complex numbers z, denote by z^* the complex conjugate. By abuse of notation, for all matrices $M \in \mathcal{M}_d(\mathbb{C})$, denote by M^* the matrix of element-wise complex conjugates of M. Finally, denote by $M^\dagger$ the conjugate transpose of M.

Second order characterization Fix a sampling period $h > 0$. By abuse of notation, we see $\nu_\tau^{(h)}$ as a matrix of functions of the lag τ, and denote by $\mathcal{F}[\nu^{(h)}]$ the Fourier transform of this matrix. Bacry et al. [5] show[1] that for all frequencies $x \in \mathbb{R}$,

$$\mathcal{F}\left[\nu^{(h)}\right](x) = \mathcal{F}\left[f_{\mathrm{Tr}}^{(h)}\right](x)\Big(\mathbb{I}_d - \mathcal{F}[\Phi]^*(x)\Big)^{-1}\mathbf{diag}(\boldsymbol{\eta}_\star)\Big(\mathbb{I}_d - \mathcal{F}[\Phi]^\intercal(x)\Big)^{-1}. \tag{3.33}$$

We introduce some different terms. Since $\rho(||\Phi||_1) < 1$, $\det(\mathbb{I}_d - \mathcal{F}[\phi]) \in (0, 1)$. Therefore, for times $t \geq 0$, we can define the function

$$Z(t) := \mathcal{F}^{-1}\left[\frac{1}{\det(\mathbb{I}_d - \mathcal{F}[\phi])} - 1\right](t). \tag{3.34}$$

For all periods $x \geq 0$, denote $\mathcal{Z}(x) := |1 + \mathcal{F}[Z](x)|^2 - 1$.

Lemma 3.1.2 (Reformulating the auto-covariance)**.** *For all frequencies $x \in \mathbb{R}$*

$$\mathcal{F}\left[\nu^{(h)}\right](x) = \mathcal{F}\left[f_{Tr}^{(h)}\right](x)(1+\mathcal{Z}(x))\mathbf{adj}\Big(\mathbb{I}_d-\mathcal{F}[\Phi]^*(x)\Big)\mathbf{diag}(\boldsymbol{\eta}_\star)\mathbf{adj}\Big(\mathbb{I}_d-\mathcal{F}[\Phi]^\intercal(x)\Big). \tag{3.35}$$

We give a proof of this result in Appendix A.2.4. For all times $s \geq 0$, define the convolution

$$z^{(2)}(s) := \int_0^{+\infty} Z(u)Z(u+s)\mathrm{d}u. \tag{3.36}$$

For lags $\tau \geq 0$, define the primitives

$$Z^{(1)}(\tau) := \int_0^\tau Z(t)\mathrm{d}t, \quad Z^{(2)}(\tau) := \int_0^\tau z^{(2)}(s)\mathrm{d}s. \tag{3.37}$$

We express the inverse Fourier transform of $\mathcal{Z}$ in terms of the functions Z and $z^{(2)}$.

Lemma 3.1.3 (Determinant of auto-convolutions)**.** *For all times $s \in \mathbb{R}$,*

$$\mathcal{F}^{-1}\Big[\mathcal{Z}\Big](s) = Z(s) + Z^{(-)}(s) + z^{(2)}(|s|). \tag{3.38}$$

Therefore, for all times $\tau \geq 0$,

$$\int_0^\tau \mathcal{F}^{-1}\Big[\mathcal{Z}\Big](s)\mathrm{d}s = Z^{(1)}(\tau) + Z^{(2)}(\tau). \tag{3.39}$$

We give a proof of this result in Appendix A.2.4.

[1] As mentioned by the authors, the relation between the auto-covariance function and the kernels is originally shown by Hawkes [47].

3.1.1.4 Modelling with MHP

Model identifiability Model identifiability can be understood by the following question: considering two MHP models, can we compensate a lower background rate by a larger kernel value to get the same intensity value for both models? As we see below, the answer to this question is negative; 1-dimensional MHP are path-wise identifiable, the parameters of a model are fully characterized by a single path of the conditional intensity process.

Proposition 3.1.2 (1-dimensional MHP)**.** 1-*dimensional MHP are path-wise identifiable.*

We give a proof of this result in Appendix A.2.2. The identifiability of MHP in dimension $d \geq 2$ is beyond the scope of this work.

Parameterization Fix an event type $k \in [d]$. For all event type $i \in [d]$, we consider kernel functions of the form $\phi_{ki} = \sum_{l \in [r_{ki}]} \omega_{ki,l} \tilde{\phi}_{ki,l}$, with previously discussed assumptions. We denote by $\boldsymbol{\omega_k}$ the vector of L_1 weights $\omega_{ki,l}$ sorted in lexicographic order of event type $i \in [d]$, and mixture index $l \in [r_{ki}]$. Therefore, $\boldsymbol{\omega_k}$ is a vector of size $r_k := \sum_{i=1}^{d} r_{ki}$

$$\boldsymbol{\omega_k}^\intercal = (\omega_{k1,1}, \omega_{k1,2}, \ldots, \omega_{k1,r_{k1}}, \omega_{k2,1}, \ldots, \omega_{kd,r_{kd}}). \tag{3.40}$$

For all event types $i \in [d]$, for all mixture indices $l \in [r_{ki}]$, denote by $\tilde{\boldsymbol{\theta}}_{\boldsymbol{ki,l}}$ the vector of parameters of density kernel $\tilde{\phi}_{ki,l}$. Denote by $\tilde{r}_{ki,l}$ the number of these parameters, and let $\tilde{r}_{ki} = \sum_{l=1}^{r_{ki}} \tilde{r}_{ki,l}$. We denote by $\tilde{\boldsymbol{\theta}}_{\boldsymbol{k}}$ the vector of parameters of density kernels, defined similarly to $\boldsymbol{\omega_k}$ by concatenating the vectors $\tilde{\boldsymbol{\theta}}_{\boldsymbol{ki,l}}$ in lexicographic order of event type and mixture index. This vector $\tilde{\boldsymbol{\theta}}_{\boldsymbol{k}}$ has size $\tilde{r}_k = \sum_{i=1}^{d} \tilde{r}_{ki}$. Finally, define $\boldsymbol{\theta_k}$, the vector of parameters of MHP model functionals of dimension k by $\boldsymbol{\theta_k}^\intercal = (\mu_k, \boldsymbol{\omega_k}^\intercal, \tilde{\boldsymbol{\theta}}_{\boldsymbol{k}}^\intercal)$. The total number of parameters of the MHP model across dimensions $k \in [d]$ is $n_{\text{param}} := d + r_k + \tilde{r}_k$. If for all event types $k \in [d]$, the kernel density parameters $\tilde{\boldsymbol{\theta}}_{\boldsymbol{k}}$ are fixed, we refer to the Hawkes model as a sum of basis functions (SBF) MHP; these are widely used in the literature. Our method applies to general MHP, and includes SBF MHP as a special case. In this case, the minimization of $\mathcal{R}_T^{(k)}$ is a quadratic program (QP).

Residuals Unlike for Poisson processes, the compensator of an MHP is not deterministic. We give a formula for the residuals of MHP models. In the rest of this paragraph, fix an event type $k \in [d]$. Using the same approach as in Lemma A.3.1, we get the following formula for MHP residuals.

Proposition 3.1.3 (Residuals of MHP models)**.** *For event indices $m \in [N_T^k]$, the m-th compensator-transformed time of type k is*

$$s_m^{(k)} = t_m^k \mu_k + \sum_{i=1}^{d} \mathbb{1}_{\{m \geq \varpi(k,i)\}} \sum_{n=1}^{\kappa(i,k,m)} \psi_{ki}(t_m^k - t_n^i). \tag{3.41}$$

If for event types $i \in [d]$, the kernel densities $\tilde{\phi}_{ki}$ are compactly supported, that is, there exists $\delta_{ki}^{(L)}, \delta_{ki}^{(R)} \in [0, +\infty)$ with $\delta_{ki}^{(L)} < \delta_{ki}^{(R)}$, such that $\tilde{\phi}_{ki}(t) = 0$ for all times $\forall t \notin \left[\delta^{(L)}, \delta^{(R)}\right)$, then for event times $t_n^i < t_m^k$,

$$\tilde{\psi}_{ki}(t_m^k - t_n^i) = \begin{cases} 0 & \text{if } t_n^i \geq t_m^k - \delta_{ki}^{(L)}, \\ 1 & \text{if } t_n^i \leq t_m^k - \delta_{ki}^{(R)}. \end{cases} \tag{3.42}$$

This accelerates the computation of the residuals, but does not hold for kernels which are not compactly supported. Proposition 3.1.3 implies that the computation of all compensator-transformed times $\boldsymbol{s^k}$ has quadratic time complexity $\mathcal{O}(N_T^2)$, without further assumptions on the kernels. This prohibits residual analysis with large amounts of training data. For general kernels, since the primitives ψ_{ki} are monotonically increasing to ω_{ki}, we propose a simple approximation of the compensator-transformed times $\boldsymbol{s^k}$, with linear time complexity and arbitrary precision, which consists in "trimming" the memory of the observed point process path. For the sake of brevity, we discuss this approximation in dimension $d = 1$, but this approach holds in general dimension. Suppose $d = 1$, and let $k = i = 1$. Fix a cutoff parameter $c \in [N_T^k - 1]$. For all event indices $m \in [N_T^k]$, the cutoff c is the number of events preceding t_m^k on which we evaluate the CDF $\tilde{\psi}_{ki}$. Formally, we approximate the compensator-transformed time $s_m^{(k)}$ by

$$\hat{s}_m^{(k)} := \begin{cases} t_m^k \mu_k + (m - c - 1)\omega_{ki} + \sum_{n=m-c}^{m-1} \psi_{ki}(t_m^k - t_n^i) & \text{if } m > c + 1, \\ s_m^{(k)} & \text{otherwise.} \end{cases} \tag{3.43}$$

This approximation clearly poses a speed-accuracy trade-off. The approximation of all compensator-transformed times $\hat{\boldsymbol{s}}^{\boldsymbol{k}}$ has time complexity $\mathcal{O}(cN_T)$; the speed improvement relies on choosing $c \ll N_T$. However, in order to conduct reliable KS tests or Wald–Wolfowitz tests based on this approximation, it is necessary to control the relative approximation error

$$\max_{m \in [N_T^k]} \frac{|\hat{s}_m^{(k)} - s_m^{(k)}|}{s_m^{(k)}}. \tag{3.44}$$

This approximation error is monotonically decaying with the cutoff c, and the approximation over-estimates the true value of the compensator-transformed time, that is, $\hat{s}_m^{(k)} \geq s_m^{(k)}$ for all $m \in [N_T^k]$. However, we can achieve arbitrary small estimation error, depending on the speed of convergence of the survival function $1 - \tilde{\psi}_{ki}$ to 0, and the observed point process path. Since $\tilde{\psi}_{ki}$ is a CDF, it is monotonically increasing and $\lim_{\tau \to +\infty} \tilde{\psi}_{ki} = 1$. Denote by $\tilde{\psi}_{ki}^{-1}$ the generalized inverse of this CDF.

Proposition 3.1.4 (Residual approximation error)**.** *Fix a cutoff $c \in [N_T^k - 1]$. Let*

$$\tau_c := \min_{m \in [\![c+2, N_T]\!]} t_m^k - t_{m-c-1}^k, \quad \epsilon_c := 1 - \tilde{\psi}_{ki}(\tau_c). \tag{3.45}$$

Then the relative approximation error satisfies the upper-bound

$$\max_{m\in[N_T^k]} \frac{|\hat{s}_m^{(k)} - s_m^{(k)}|}{s_m^{(k)}} \leq \frac{\epsilon_c}{1-\epsilon_c}. \tag{3.46}$$

Conversely, let $\epsilon \in [0,1)$. If there exists a cutoff value $c \in [N_T^k - 1]$ such that

$$\min_{m\in[\![c+2,N_T]\!]} \left(t_m^k - t_{m-c-1}^k\right) \geq \tilde{\psi}_{ki}^{-1}\Big(\frac{1}{1+\epsilon}\Big), \tag{3.47}$$

Then the relative approximation error satisfies the upper-bound

$$\max_{m\in[N_T^k]} \frac{|\hat{s}_m^{(k)} - s_m^{(k)}|}{s_m^{(k)}} \leq \epsilon. \tag{3.48}$$

We give a proof of this result in Appendix A.2.1.

3.1.2 Uni-dimensional MHP

In this paragraph, we show some results for uni-dimensional MHP ($d = 1$). Let $\boldsymbol{N}$ be an MHP with adjacency matrix $||\Phi||_1 = (\omega_{11})$ and baseline $\boldsymbol{\mu} = (\mu_1)$.

3.1.2.1 First order moment

This MHP has branching ratio $\rho(||\Phi||_1) = \omega_{11}$. The stability condition is trivial:

$$\rho(||\Phi||_1) < 1 \iff \omega_{11} < 1. \tag{3.49}$$

In the uni-dimensional case, the LLN condition is equivalent to the stability condition. Suppose the adjacency matrix $||\Phi||_1$ satisfies the stability condition. Let $\boldsymbol{\eta} \in (0, +\infty)$. The LLN condition is

$$(\mathbb{I}_d - ||\Phi||_1)\boldsymbol{\eta} > 0 \iff \omega_{11} < 1. \tag{3.50}$$

The associated stationary regime intensity is $\eta_\star^{(1)} = \frac{\mu_1}{1-\omega_{11}}$, and the associated exogeneity ratio is $\texttt{exo} = 1 - \omega_{11}$. As discussed previously, the total exogeneity ratio `exo` of uni-dimensional MHP does not depend on the baseline μ_{11}. In this case, the parameter ω_{11} is also the endogeneity ratio of events in the system.

3.1.2.2 Second order moment

In this paragraph, we use the results of Bacry et al. [5] to show a slightly different formulation of the covariance of a 1-dimensional MHP, which simplifies closed form expressions for some parametric families of kernels. For times $t \geq 0$,

$$Z(t) = \mathcal{F}^{-1}\left[\frac{\mathcal{F}[\phi_{11}]}{1-\mathcal{F}[\phi_{11}]}\right](t). \tag{3.51}$$

An integral operator For sampling periods $h > 0$, define the linear operator $\mathcal{Q}^{(h)}$ acting on positive causal functions $f : [0, +\infty) \to [0, +\infty)$ such that for all non-negative lags $\tau \geq 0$

$$\mathcal{Q}^{(h)}[f](\tau) := \begin{cases} \frac{1}{h}\Big(\int_{[\tau,\tau+h]} f - \int_{[\tau-h,\tau]} f \Big) & \text{if } \tau \in [h, +\infty), \\ \frac{1}{h}\Big(\int_{[\tau,\tau+h]} f - \int_{[h-\tau,\tau]} f \Big) & \text{if } \tau \in [\frac{h}{2}, h), \\ \frac{1}{h}\Big(2\int_{[\tau,h-\tau]} f + \int_{[h-\tau,h+\tau]} f \Big) & \text{if } \tau \in [0, \frac{h}{2}). \end{cases} \tag{3.52}$$

If the function f is increasing, then $\mathcal{Q}^{(h)}[f](\tau) \geq 0 \quad \tau \geq 0$. Regarding the asymptotic behaviour of $\mathcal{Q}^{(h)}[f]$, note that

$$\lim_{h \to 0^+} \mathcal{Q}^{(h)}[f](\tau) = 0, \quad \forall \tau \geq 0, \qquad \mathcal{Q}^{(h)}[f](0) = \frac{2}{h} \int_{[0,h]} f, \quad \forall h > 0. \tag{3.53}$$

Note that if the function f is constant on the exterior of a segment $[\delta_L, \delta_R]$, then $\mathcal{Q}^{(h)}[f]$ is compactly supported on $[(\delta_L - h)_+, \delta_R + h]$.

A formula for the covariance

Proposition 3.1.5 (Covariance of MHP in $d = 1$)**.** *For a sampling period $h > 0$ and a lag $\tau \geq 0$, the covariance function of $\boldsymbol{N}$ is*

$$\nu_\tau^{(h)} = \eta_\star^1 f_{Tr}^{(h)}(\tau) + \eta_\star^1 \mathcal{Q}^{(h)}[Z^{(1)} + Z^{(2)}](\tau). \tag{3.54}$$

We give a proof of this result in Appendix A.2.4. Note that, by construction, the functions $Z^{(1)}$ and $Z^{(2)}$ are both non-decreasing. Therefore the terms $\mathcal{Q}^{(h)}[Z^{(1)}](\tau)$ and $\mathcal{Q}^{(h)}[Z^{(2)}](\tau)$ are both non-negative. This implies that the covariance is always non-negative. An immediate corollary is that the L_1 norm of the auto-covariance of the MHP is known in closed form.

Corollary 3.1.2 (L_1 norm of the covariance function)**.** *The L_1 norm of the MHP is*

$$||\nu_{11}^{(h)}||_1 = \frac{h\eta_\star^{(1)}}{(1 - \omega_{11})^2}. \tag{3.55}$$

Note that the L_1 norm of the covariance function $||\nu_{11}^{(h)}||_1$ does not depend on the kernel density $\tilde{\phi}_{11}$. Furthermore, the parameters (μ_{11}, ω_{11}) are fully characterized by the moments $(\eta_\star^1, ||\nu_{11}^{(h)}||_1)$. This is the idea behind the cumulants method of Achab et al. [2]. One can potentially use this characterization of (μ_{11}, ω_{11}) for the initialization of the `ASLSD` estimation method, given a computationally cheap but noisy empirical estimator of $||\nu_{11}^{(h)}||_1$.

Example: exponential MHP We now conclude this discussion of the uni-dimensional MHP with a concrete example of kernels: the classic case of an MHP with a unique ($r_{11} = 1$) exponential kernel. Consider an L_1 weight $\omega_{11} \in (0, 1)$, and an exponential decay rate parameter $\beta_{11} > 0$. For all times $t \geq 0$, the exponential kernel density $\tilde{\phi}_{11}$ is

$$\tilde{\phi}_{11}(t) := \beta_{11} \exp(-\beta_{11} t). \tag{3.56}$$

Fix a frequency $x \in \mathbb{R}$. The Fourier transform of the kernel ϕ_{11} is simply

$$\mathcal{F}[\phi_{11}](x) = \frac{\omega_{11}\beta_{11}}{\beta_{11} + i2\pi x}. \tag{3.57}$$

This implies that

$$\frac{\mathcal{F}[\phi_{11}]}{1 - \mathcal{F}[\phi_{11}]}(x) = \frac{\omega_{11}\beta_{11}}{(1 - \omega_{11})\beta_{11} + i2\pi x}. \tag{3.58}$$

Define the modified L_1 weight $a_{11} := \frac{\omega_{11}}{1-\omega_{11}}$, and the coupled decay rate $b_{11} := (1-\omega_{11})\beta_{11}$. It is clear that $a_{11} \geq 0$, and is a strictly increasing function of ω_{11}. Note that a_{11} is also the expected number of descendants of any given event in the branching representation of this MHP. Now for all times $t \geq 0$, we get

$$Z(t) = a_{11} b e^{-bt}. \tag{3.59}$$

Therefore, the function Z still belongs to the parametric class of exponential kernels, with L_1 weight a_{11} and decay rate b_{11}. In this case, it is simple to compute the moments $Z^{(1)}$, $z^{(2)}$, and $Z^{(2)}$ using the formulas of $\tilde{\psi}$ and $\tilde{\Upsilon}$ for an exponential kernel (see Section 4.6). For all times $s > 0$, we get

$$z^{(2)}(s) = \frac{a_{11}^2}{2} b_{11} e^{-b_{11} s}. \tag{3.60}$$

For all times $t \geq 0$,

$$Z^{(1)}(t) = a_{11}(1 - e^{-b_{11}t}), \quad Z^{(2)}(t) = \frac{a_{11}^2}{2}(1 - e^{-b_{11}t}). \tag{3.61}$$

Hence

$$Z^{(1)}(t) + Z^{(2)}(t) = (a_{11} + \frac{a_{11}^2}{2})(1 - e^{-b_{11}t}). \tag{3.62}$$

Note that for all sampling periods $h > 0$, and for all lags $\tau \geq 0$

$$\mathcal{Q}^{(h)}[1 - e^{-b_{11}t}](\tau) := \begin{cases} 2\frac{\cosh(b_{11}h) - 1}{b_{11}h} e^{-b_{11}\tau} & \text{if } \tau \in [h, +\infty), \\ 2 f_{\mathrm{Tr}}^{(h)}(\tau) + \frac{2}{b_{11}h}\Big(e^{-b_{11}h}\cosh(b_{11}\tau) - e^{-b_{11}\tau}\Big) & \text{if } \tau \in [0, h). \end{cases} \tag{3.63}$$

Eventually, we get the following formula for the covariance of this MHP.

Corollary 3.1.3 (Covariance of exponential MHP in $d = 1$)**.** *For sampling periods $h > 0$, and for lags $\tau \geq 0$*

$$\nu^{(h)}_{11,\tau} = \begin{cases} \eta^1_\star(a_{11} + \frac{a_{11}^2}{2}) 2\frac{\cosh(b_{11}h)-1}{b_{11}h} e^{-b_{11}\tau} & \text{if } \tau \geq h, \\ f^{(h)}_{Tr}(\tau)\frac{\eta^1_\star}{(1-\omega_{11})^2} + \eta^1_\star(a_{11} + \frac{a_{11}^2}{2})\frac{2}{b_{11}h}\Big(e^{-b_{11}h}\cosh(b_{11}\tau) - e^{-b_{11}\tau}\Big) & \text{if } \tau \in [0, h). \end{cases} \tag{3.64}$$

We comment on this expression in more detail. First, for all $x \in \mathbb{R}$, define the function

$$f_{\text{hyp}}(x) := \frac{\cosh(x) - 1}{x}. \tag{3.65}$$

Note that if we denote by sinhc the hyperbolic cardinal sine function $\text{sinhc}(x) := \frac{\sinh(x)}{x}$, then we get the formula

$$f_{\text{hyp}}(x) = \frac{x}{2}\,\text{sinhc}^2\Big(\frac{x}{2}\Big). \tag{3.66}$$

It is clear that the function f_{hyp} is strictly increasing. Furthermore,

$$f_{\text{hyp}} \underset{x\downarrow 0}{\sim} \frac{x}{2}, \quad \text{and} \quad \lim_{x\to 0^+} f_{\text{hyp}}(x) = 0. \tag{3.67}$$

Therefore, for $x > 0$, $f_{\text{hyp}}(x) > 0$. We start by analysing the result for lags $\tau \geq h$. We re-write the auto-covariance function in this case as

$$\nu^{(h)}_{11,\tau} = \eta^1_\star\Big(\frac{1}{(1-\omega_{11})^2} - 1\Big) f_{\text{hyp}}(b_{11}h) e^{-b_{11}\tau}. \tag{3.68}$$

Therefore, in this case, the covariance $\nu^{(h)}_{11,\tau}$ only depends on the sampling period h through the term $f_{\text{hyp}}(b_{11}h)$, which increases with h. The covariance only depends on the lag τ through the exponential decay term $e^{-b_{11}\tau}$, where both kernel parameters $(\omega_{11}, \beta_{11})$ are coupled through the decay rate b_{11}.

First, we discuss the dependence in the L_1 weight ω_{11}. In the homogeneous Poisson limiting case $\omega \downarrow 0$, as expected, we get no correlation $\lim_{\omega\downarrow 0} \nu^{(h)}_{11,\tau} = 0$. In order to analyse the other limiting case, that of the critical regime limit $\omega \uparrow 1$, we note that $\nu^{(h)}_{11,\tau} \underset{\omega\uparrow 1}{\sim} \eta^1_\star \frac{\beta_{11}h}{2(1-\omega_{11})}$. Therefore, the auto-covariance of the exponential MHP explodes when near-critical, that is $\lim_{\omega_{11}\uparrow 1} \nu^{(h)}_{11,\tau} = +\infty$. Second, we discuss the dependence in the exponential decay rate β_{11}. The limiting case $\beta_{11} \downarrow 0$ is actually to the homogeneous Poisson case since $\lim_{\beta_{11}\downarrow 0} \tilde{\phi}_{11} = 0$. Unsurprisingly, we get $\lim_{\beta_{11}\downarrow 0} \nu^{(h)}_{11,\tau} = 0$. As we expect, there is no correlation either for infinitely fast decay rates: $\lim_{\beta_{11}\to+\infty} \nu^{(h)}_{11,\tau} = 0$. Note that in the limit $\beta_{11} \to +\infty$, $\tilde{\phi}_{11}$ goes to a Dirac distribution at 0.

3.1.3 Multi-dimensional MHP

In the multi-dimensional MHP case, characterizing the stability condition and the LLN condition in terms of adjacency matrix coefficients is particularly difficult. Furthermore, general estimation methods for Hawkes models suffer a curse of dimensionality with d. Therefore, we focus our discussion on the bi-dimensional case $d = 2$; relevant in many applications of MHP models, notably in finance where some models consider two event types corresponding to upward and downward price movements of a security; and on a bi-variate model in dimension $d \geq 2$, where we assume the same kernels for all self-excitation, and the same kernels for all cross-excitation.

3.1.3.1 The bi-dimensional case $d = 2$

We consider here the case of a general bi-dimensional adjacency matrix

$$||\Phi||_1 = \begin{bmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{bmatrix}. \tag{3.69}$$

We give an analytic characterization of both the stability condition and the LLN condition in terms of the coefficients of the adjacency matrix. Note that

$$\det(\mathbb{I}_d - ||\Phi||_1) = (1 - \omega_{11})(1 - \omega_{22}) - \omega_{12}\omega_{21}. \tag{3.70}$$

If the cross excitation terms are null ($\omega_{12} = \omega_{21} = 0$), and the self-excitation terms have the same value ($\omega_{11} = \omega_{22}$), then the adjacency matrix $||\Phi||_1$ has one eigenvalue, ω_{11}, with multiplicity 2. Otherwise, the matrix $||\Phi||_1$ has 2 distinct eigenvalues:

$$\begin{aligned} e_1 &:= \frac{\omega_{11} + \omega_{22} - \sqrt{(\omega_{11} - \omega_{22})^2 + 4\omega_{12}\omega_{21}}}{2}, \\ e_2 &:= \frac{\omega_{11} + \omega_{22} + \sqrt{(\omega_{11} - \omega_{22})^2 + 4\omega_{12}\omega_{21}}}{2}. \end{aligned} \tag{3.71}$$

It is clear that the spectral radius of the adjacency matrix $||\Phi||_1$ is

$$\rho(||\Phi||_1) = \frac{\omega_{11} + \omega_{22} + \sqrt{(\omega_{11} - \omega_{22})^2 + 4\omega_{12}\omega_{21}}}{2}. \tag{3.72}$$

We note that by construction, $\max(\omega_{11}, \omega_{22}) \leq \rho(||\Phi||_1)$; and

$$\omega_{12}\omega_{21} = (\rho(||\Phi||_1) - \omega_{11})(\rho(||\Phi||_1) - \omega_{22}). \tag{3.73}$$

Proposition 3.1.6 (Stability condition in dimension $d = 2$)**.** *The stability condition in dimension $d = 2$ is*

$$\rho(||\Phi||_1) < 1 \iff \omega_{12}\omega_{21} < (1 - \omega_{11})(1 - \omega_{22}) \quad \textit{and} \quad \omega_{11}, \omega_{22} \in [0, 1). \tag{3.74}$$

It is clear that if the stability condition holds, then the self-excitation terms are such that $\omega_{11}, \omega_{22} \in [0, 1)$ and the cross-excitation terms are such that $\omega_{12}\omega_{21} < 1$.

Proposition 3.1.7 (LLN condition in dimension $d = 2$)**.** *Suppose the adjacency matrix* $||\Phi||_1$ *satisfies the stability condition. Let* $\boldsymbol{\eta} \in (0, +\infty)^2$*. The LLN condition is*

$$(\mathbb{I}_d - ||\Phi||_1)\boldsymbol{\eta} > 0 \iff \begin{cases} \frac{\omega_{21}}{1-\omega_{22}} < \frac{\eta^{(2)}}{\eta^{(1)}} < \frac{1-\omega_{11}}{\omega_{12}} & \text{if } \omega_{12} > 0, \\ \frac{\omega_{21}}{1-\omega_{22}} < \frac{\eta^{(2)}}{\eta^{(1)}} & \text{otherwise.} \end{cases} \tag{3.75}$$

It is clear that in this bi-dimensional case, if there exists $\boldsymbol{\eta} \in (0, +\infty)^2$ satisfying Equation (3.75), then Equation (3.74) is satisfied.

Proposition 3.1.8 (Power series)**.** *Suppose the adjacency matrix* $||\Phi||_1$ *satisfies the stability condition. Therefore, the matrix* $\mathbb{I}_d - ||\Phi||_1$ *is invertible, and*

$$(\mathbb{I}_d - ||\Phi||_1)^{-1} = \frac{1}{(1-\omega_{11})(1-\omega_{22}) - \omega_{12}\omega_{21}} \begin{bmatrix} 1-\omega_{22} & \omega_{12} \\ \omega_{21} & 1-\omega_{11} \end{bmatrix}. \tag{3.76}$$

Proposition 3.1.9 (Stationary regime intensity)**.** *Consider a 2-dimensional MHP with adjacency matrix* $||\Phi||_1$ *and baseline* $\boldsymbol{\mu}$*. The associated stationary regime intensity is*

$$\boldsymbol{\eta_\star} = \frac{1}{(1-\omega_{11})(1-\omega_{22}) - \omega_{12}\omega_{21}} \begin{bmatrix} (1-\omega_{22})\mu_1 + \omega_{12}\mu_2 \\ \omega_{21}\mu_1 + (1-\omega_{11})\mu_2 \end{bmatrix}. \tag{3.77}$$

Proposition 3.1.10 (Exogeneity ratio)**.** *Consider a 2-dimensional MHP with adjacency matrix* $||\Phi||_1$ *and baseline* $\boldsymbol{\mu}$*. The associated stationary regime intensity is*

$$\begin{aligned} \texttt{exo} &= \frac{1 - \omega_{11} - \omega_{22} + \omega_{11}\omega_{22} - \omega_{12}\omega_{21}}{1 - \frac{\mu_1}{\mu_1+\mu_2}(\omega_{22} - \omega_{21}) - \frac{\mu_2}{\mu_1+\mu_2}(\omega_{11} - \omega_{12})}, \\ \texttt{exo} &= \frac{(1 - \rho(||\Phi||_1))(1 + \rho(||\Phi||_1) - \omega_{11} - \omega_{22})}{1 - \frac{\mu_1}{\mu_1+\mu_2}(\omega_{22} - \omega_{21}) - \frac{\mu_2}{\mu_1+\mu_2}(\omega_{11} - \omega_{12})}. \end{aligned} \tag{3.78}$$

3.1.3.2 A bi-variate matrix model for $d \geq 2$

In the general multi-dimensional case $d > 1$, it is difficult to get similar analytic characterizations like the ones discussed previously for $d = 1$ and $d = 2$. In this paragraph, we discuss briefly a particular case of kernel matrices in dimension $d > 1$, which consists in a bi-variate kernel matrix model with one self-excitation kernel and one cross excitation kernel.

The model Define the square matrix $J \in \mathcal{M}_d(\mathbb{R})$ where for all $i, j \in [d]$, $J_{ij} = 1$. This matrix clearly has rank 1, so 0 is an eigenvalue of multiplicity $d-1$. It is clear that d is the only other eigenvalue of J, with associated eigenvector $\mathbb{1}_{\boldsymbol{d}}$. By induction, we see that for all $p \in \mathbb{N}^*$, $J^p = d^{p-1}J$. Finally, $Jx = \|\boldsymbol{x}\|_1 \mathbb{1}_{\boldsymbol{d}}$ for all vectors $\boldsymbol{x} \geq 0$. For all $a, b \in \mathbb{R}$, denote

by $J_{(a,b)} \in \mathcal{M}_d(\mathbb{R})$ the matrix which diagonal elements are $(J_{(a,b)})_{ii} = a$ for all $i \in [d]$, and which non-diagonal elements are $(J_{(a,b)})_{ij} = b$. We notice that this matrix satisfies

$$J_{(a,b)} = bJ + (a-b)\mathbb{I}_d. \tag{3.79}$$

By abuse of notation, for real valued functions f_a, f_b, we denote by $J_{(f_a,f_b)}$ the matrix of functions such that for all $t \geq 0$

$$J_{(f_a,f_b)}(t) := J_{(f_a(t),f_b(t))}. \tag{3.80}$$

Denote by $\omega_S > 0$ (resp. $\omega_C > 0$) the self-excitation (resp. cross-excitation) L_1 weight. Consider a self-excitation (resp. cross-excitation) kernel density $\tilde{\phi}_S$ (resp. $\tilde{\phi}_C$). Finally, define the associated kernels $\tilde{\phi}_S = \omega_S \tilde{\phi}_S$ and $\tilde{\phi}_C = \omega_C \tilde{\phi}_C$. We can now define the bi-variate kernel matrix model

$$\Phi := J_{(\phi_S,\phi_C)}. \tag{3.81}$$

Note that in dimension $d = 2$, all bi-symmetric kernel matrices are of bi-variate models $J_{(\phi_S,\phi_C)}$. In fact, this case is of particular interest in mid-price modelling in finance since empirically, kernel fits appear to be close to bi-variate (see Chapter 9). For instance, this hypothesis is used in Bacry et al. [5]. This bi-variate model can also serve as an L_2 approximation of general kernel matrices in the following sense. Consider a general kernel matrix $\Phi = (\phi_{ij})_{i,j\in[d]}$. The L_2 projection of the kernel matrix Φ on the space of bi-variate matrix models is $J_{(\bar{\phi}_S,\bar{\phi}_C)}$, where $\bar{\phi}_S$ (resp. $\bar{\phi}_C$) is the average of self-excitation (resp. cross-excitation) kernels of $||\Phi||_1$. Formally, $\bar{\phi}_S = \frac{1}{d}\sum_{i=1}^{d} \phi_{ii}$, and $\bar{\phi}_C = \frac{1}{d(d-1)}\sum_{i,j\in[d],i\neq j} \phi_{ij}$. The approximation error depends on the respective sample variances of self excitation kernels and cross-excitation kernels. Formally, the L_2 distance of a d-dimensional kernel matrix Φ to the space of bi-variate matrix models is

$$d\sigma_S^2 + d(d-1)\sigma_C^2, \tag{3.82}$$

where $\sigma_S^2 = \frac{1}{d}\sum_{i=1}^{d}(\phi_{ii} - \bar{\phi}_S)^2$ and $\sigma_C^2 = \frac{1}{d(d-1)}\sum_{i,j\in[d],i\neq j}(\phi_{ij} - \bar{\phi}_C)^2$.

First order moment In this paragraph, we discuss first order properties of MHP with a bi-variate kernel matrix. Even though we do not impose such constraints on the kernel matrix in the `ASLSD` procedure, we allow the use of these results to initialize the L_1 weights $||\Phi||_1$ and the baseline $\boldsymbol{\mu}$ in high dimension. It is clear that the adjacency matrix of the bi-variate model is itself bi-variate of the form $||\Phi||_1 = J_{(\omega_S,\omega_C)}$. We now characterize first order feasibility for the bi-variate model.

Proposition 3.1.11 (Stability condition)**.** *The branching ratio of the adjacency matrix* $||\Phi||_1$ *under the bi-variate model is* $\rho(||\Phi||_1) = (d-1)\omega_C + \omega_S$*. Hence, the stability condition for this matrix is*

$$\rho(||\Phi||_1) < 1 \iff \omega_C < \frac{1-\omega_S}{d-1}. \tag{3.83}$$

We give a proof of this result in Appendix A.2.3. Note that if the adjacency matrix $||\Phi||_1$ is stable, this implies that $\omega_S \in [0, 1)$ and $\omega_C \in [0, \frac{1}{d-1})$.

Proposition 3.1.12 (LLN condition)**.** *Let* $\boldsymbol{\eta} \in (0, +\infty)^d$*. Denote by*

$$m(\boldsymbol{\eta}) := \frac{\min_{k\in[d]} \eta^k}{\sum_{i=1}^d \eta^i}. \tag{3.84}$$

The LLN condition for the bi-variate model is

$$(\mathbb{I}_d - ||\Phi||_1)\boldsymbol{\eta} > 0 \iff \frac{\omega_C}{1+\omega_C-\omega_S} < m(\boldsymbol{\eta}). \tag{3.85}$$

This is equivalent to

$$(\mathbb{I}_d - ||\Phi||_1)\boldsymbol{\eta} > 0 \iff \omega_C < \frac{m(\boldsymbol{\eta})}{1-m(\boldsymbol{\eta})}(1-\omega_S). \tag{3.86}$$

It is clear that if there exists $\boldsymbol{\eta} > 0$ such that the LLN condition holds, then the stability condition must hold too. Indeed, let $\boldsymbol{\eta} > 0$ such that $(\mathbb{I}_d - ||\Phi||_1)\boldsymbol{\eta} > 0$. Since $m(\boldsymbol{\eta}) \leq \frac{1}{d}$, then $\frac{m(\boldsymbol{\eta})}{1-m(\boldsymbol{\eta})} \leq \frac{1}{d-1}$. Therefore, the LLN condition implies the stability condition under the bi-variate model. In the rest of this paragraph, consider a bi-variate adjacency matrix $||\Phi||_1$ satisfying the stability condition $\rho(||\Phi||_1) < 1$. This implies that the matrix $\mathbb{I}_d - ||\Phi||_1$ is invertible. The following lemma computes its inverse, which is still of the bi-variate form. We then use this lemma to get both the stationary regime intensity and the exogeneity rate of this model in closed form.

Lemma 3.1.4 (Power series of the adjacency matrix)**.** *The power series of the adjacency matrix is*

$$(\mathbb{I}_d - ||\Phi||_1)^{-1} = J_{(a_S, a_C)}; \tag{3.87}$$

with

$$\begin{aligned} a_S &:= \frac{1}{1-(\omega_S-\omega_C)}\Big(1 + \frac{\omega_C}{1-(\omega_S+(d-1)\omega_C)}\Big), \\ a_C &:= \frac{\omega_C}{(1-(\omega_S-\omega_C))\,(1-(\omega_S+(d-1)\omega_C))}. \end{aligned} \tag{3.88}$$

We give a proof of this result in Appendix A.2.3. We can now compute a simple expression for the stationary regime intensity $\boldsymbol{\eta}_*$ under the bi-variate model.

Proposition 3.1.13 (Stationary regime intensity)**.** *The stationary regime intensity of the MHP is*

$$\boldsymbol{\eta_*} = \frac{1}{1-(\omega_S-\omega_C)}\left(\frac{\omega_C}{1-(\omega_S+(d-1)\omega_C)}||\boldsymbol{\mu}||_1\mathbb{1}_{\boldsymbol{d}}+\boldsymbol{\mu}\right). \tag{3.89}$$

Proposition 3.1.14 (Exogeneity ratio)**.** *Let $J_{(\omega_S,\omega_C)}$ satisfying the stability condition $\rho(J_{(\omega_S,\omega_C)}) < 1$. Then the exogeneity ratio is*

$$\texttt{exo} = 1 - \left(\omega_S + (d-1)\omega_C\right). \tag{3.90}$$

Balanced classes case We refer to the balanced classes case as the MHP where the kernel matrix is of the bi-variate form, and where the stationary regime intensity is constant across event types, that is for all $i, j \in [d]$, $\eta_\star^i = \eta_\star^j$. Empirically, this case is particularly relevant to mid-price modelling in finance (see Chapter 9). Note that for a bi-variate model, having constant stationary regime intensity across event types is equivalent to having constant baseline across event types. That is, if the adjacency matrix is bi-variate, then

$$\eta_\star^i = \eta_\star^j \quad \forall i,j \in [d] \iff \mu_i = \mu_j \quad \forall i,j \in [d]. \tag{3.91}$$

This follows directly from Proposition 3.1.13. Therefore, by abuse of notation, we write

$$\boldsymbol{\eta_\star} = \eta_\star \mathbb{1}_d, \qquad \boldsymbol{\mu} = \mu \mathbb{1}_d. \tag{3.92}$$

Now it is clear that in the balanced classes case, $m(\boldsymbol{\eta}) = \frac{1}{d}$. Therefore, the stability condition and the LLN condition are the same, and simply read

$$\omega_C < \frac{1-\omega_S}{d-1}. \tag{3.93}$$

3.2 The least-squares problem for MHPs

In this section, we discuss the structure of the least squares estimation problem for MHP. As discussed previously, least-squares estimation of MHP models is closer to a kernel density estimation problem for *i.i.d.* data than a linear regression problem in supervised learning. We emphasize this analogy in Section 3.2.1, we introduce a temporal average operator and a counting average operator and re-formulate the LSE minimization problem using our decomposition. Section 3.2.2 discusses the structure and solutions of the LSE minimization problem in two steps: first, we fix kernel density parameters, and discuss the minimization of the LSE with respect to first order parameters $\boldsymbol{f}_k$ only, which is a non-negative quadratic program. As expected, our analysis shows analogous results to ordinary least squares regression for *i.i.d.* data, and we compute the minimal LSE for fixed kernel density

parameters. This leads to a new objective function to minimize with respect to kernel density parameters. Section 3.2.3 briefly shows some properties of the LSE minimizer. In this section, we consider only the long path estimation setup for the sake of brevity, but one can reach similar conclusions in the episodic setup. Recall that least squares estimation in the long path setup corresponds to solving the d independent minimization programs $(P_k)_{k\in[d]}$

$$\min_{\mu_k,\boldsymbol{\omega_k},\tilde{\boldsymbol{\theta}}_{\boldsymbol{k}}} \quad \mathcal{R}_T^{(k)}(\mu_k,\boldsymbol{\omega_k},\tilde{\boldsymbol{\theta}}_{\boldsymbol{k}}) \qquad \text{s.t.} \quad \mu_k \geq 0, \quad \boldsymbol{\omega_k} \geq 0, \quad \tilde{\boldsymbol{\theta}}_{\boldsymbol{k}} \in \tilde{\Theta}_k. \tag{3.94}$$

Now fix an event type $k \in [d]$, we focus on the analysis of problem P_k.

3.2.1 Problem formulation

3.2.1.1 Feature process

For event types $j \in [d]$, for mixture indices $l \in [r_{kj}]$, and for all times $t \geq 0$, define the feature process $\varphi_j^{(k)}$ as the stochastic convolution

$$\varphi_{jl}^{(k)}(t) := \int_0^t \tilde{\phi}_{kjl}(t-s)\mathrm{d}N_s^j. \tag{3.95}$$

This is also a first-order "kernelized" signature of the counting process $\boldsymbol{N}$. By construction, for given kernel density parameters $\tilde{\boldsymbol{\theta}}_{\boldsymbol{k}}$, we can compute a path of the feature process $\varphi_{jl}^{(k)}$ for the observed data path. Define the baseline feature as the constant process $\varphi_0^{(k)} := 1$. For a baseline parameter $\mu_k \geq 0$ and a vector of L_1 weights parameters $\boldsymbol{\omega_k} \geq 0$, define the vector of first order parameters $\boldsymbol{f_k}$ such that $\boldsymbol{f_k}^\intercal := (\mu_k, \boldsymbol{\omega_k}^\intercal)$. Denote by $\boldsymbol{\varphi^{(k)}}$ the vector of feature processes, defined similarly to $\boldsymbol{f_k}$ by concatenating the terms $\varphi_{jl}^{(k)}$ in lexicographic order of event type j and mixture index l. For event types $k \in [d]$, and for times $t \geq 0$, the intensity λ_k of the MHP model is therefore

$$\lambda_k(t) = \mu_k + \sum_{j=1}^{d}\sum_{l=1}^{r_{kj}} \omega_{kjl}\varphi_{jl}^{(k)}(t) = \boldsymbol{f_k}^\intercal \boldsymbol{\varphi^{(k)}(t)}. \tag{3.96}$$

3.2.1.2 Empirical averages

Temporal averages We introduce the time averaging operator $\langle\!\langle \cdot \rangle\!\rangle$. Let f, g be square-integrable stochastic processes on the real line. Define the (path-wise) temporal average of f on $[0,T]$ by

$$\langle\!\langle f \rangle\!\rangle_T := \frac{1}{T}\int_0^T f(t)\mathrm{d}t, \tag{3.97}$$

and the (path-wise) temporal covariance of f, g on $[0,T]$ by

$$\langle\!\langle f, g \rangle\!\rangle_T := \langle\!\langle fg \rangle\!\rangle_T - \langle\!\langle f \rangle\!\rangle_T \langle\!\langle g \rangle\!\rangle_T. \tag{3.98}$$

We refer to $\langle\!\langle f, f \rangle\!\rangle_T$ as the temporal variance of f. Using the Cauchy–Schwarz inequality, the temporal variance is clearly positive and

$$\langle\!\langle f, f \rangle\!\rangle_T = 0 \iff \exists c \in \mathbb{R}, \quad f(t) = c \quad \forall t \in [0, T]. \tag{3.99}$$

We now introduce notation for first and second order temporal averages of the feature process in the MHP estimation problem. For event types $i, j \in [d]$, and mixture indices $l \in [r_{ki}]$ and $l' \in [r_{kj}]$, define the first order terms

$$m_{jl}^{(X,k)} := \langle\!\langle \varphi_{jl}^{(k)} \rangle\!\rangle_T, \tag{3.100}$$

and the second order terms

$$m_{il,jl'}^{(XX,k)} := \langle\!\langle \varphi_{il}^{(k)} \tilde{\varphi}_{jl'}^{(k)} \rangle\!\rangle_T, \quad q_{il,jl'}^{(XX,k)} := \langle\!\langle \varphi_{il}^{(k)}, \tilde{\varphi}_{jl'}^{(k)} \rangle\!\rangle_T. \tag{3.101}$$

We denote by $\boldsymbol{m^{(X,k)}}$ the vector of first order temporal averages, defined similarly to $\boldsymbol{f_k}$ by concatenating the terms $m_{jl}^{(X,k)}$ in lexicographic order of event type j and mixture index l. Similarly, we define the matrices $m^{(XX,k)}$ and $q^{(XX,k)}$. A process of particular interest is $\tilde{\varphi}_{jl}^{(k)}$, the standardized version of the feature process $\varphi_{jl}^{(k)}$, that we define as

$$\tilde{\varphi}_{jl}^{(k)} := \frac{\varphi_{jl}^{(k)} - \langle\!\langle \varphi_{jl}^{(k)} \rangle\!\rangle_T}{\langle\!\langle \varphi_{jl}^{(k)}, \varphi_{jl}^{(k)} \rangle\!\rangle_T}; \tag{3.102}$$

so that this process satisfies $\langle\!\langle \tilde{\varphi}_{jl}^{(k)} \rangle\!\rangle_T = 0$, and $\tilde{\varphi}_{jl}^{(k)}, \tilde{\varphi}_{jl}^{(k)} \rangle\!\rangle_T = 1$.

Counting averages Define the empirical first order terms

$$m^{(Y,k)} := \frac{1}{T} \int_0^T \mathrm{d}N_t^k, \tag{3.103}$$

and the empirical second order terms

$$m_{jl}^{(XY,k)} := \frac{1}{T} \int_0^T \varphi_{jl}^{(k)}(t) \mathrm{d}N_t^k, \qquad q_{jl}^{(XY,k)} := m_j^{(XY,k)} - m_j^{(X,k)} m^{(Y,k)}. \tag{3.104}$$

We denote by $\boldsymbol{m^{(XY,k)}}$ and $\boldsymbol{q^{(XY,k)}}$ the vectors of counting averages, defined similarly to $\boldsymbol{\omega_k}$ by concatenating the terms $m_{jl}^{(XY,k)}$ and $m_{jl}^{(XY,k)}$ in lexicographic order of event type j and mixture index l. Like in the linear regression problem, the counting averages are unbiased empirical estimators of their analogous temporal averages. Using the Doob–Meyer decomposition, we see that $\mathbb{E}\Big[m^{(Y,k)}\Big] = \mathbb{E}\Big[\langle\!\langle \lambda_k \rangle\!\rangle_T\Big]$, and $\mathbb{E}\Big[m_{jl}^{(XY,k)}\Big] = \mathbb{E}\Big[\langle\!\langle \varphi_{jl}^{(k)} \lambda_k \rangle\!\rangle_T\Big]$. Note that the standardized feature process satisfies

$$\langle\!\langle \tilde{\varphi}_i^{(k)}, N^k \rangle\!\rangle_{N,T} = \frac{\langle\!\langle \varphi_i^{(k)}, N^k \rangle\!\rangle_{N,T}}{\sqrt{\langle\!\langle \varphi_i^{(k)}, \varphi_i^{(k)} \rangle\!\rangle_T}}. \tag{3.105}$$

3.2.1.3 Reformulating the LSE

In order to conduct our analysis, we propose a different formulation of the LSE. Define the vector $\boldsymbol{c_k}$ and the matrix Q_k by

$$\boldsymbol{c_k}^\intercal := \left(m^{(Y,k)}, \boldsymbol{m^{(XY,k)\intercal}}\right), \quad Q_k := \begin{pmatrix} 1 & \boldsymbol{m^{(X,k)\intercal}} \\ \boldsymbol{m^{(X,k)}} & m^{(XX,k)} \end{pmatrix}. \tag{3.106}$$

Note that $\boldsymbol{c_k} = \langle\!\langle \boldsymbol{\varphi^{(k)}} \rangle\!\rangle_{N^k,T}$, and $\quad Q_k = \langle\!\langle \boldsymbol{\varphi^{(k)}\varphi^{(k)\intercal}} \rangle\!\rangle_T$.

Proposition 3.2.1 (LSE)**.** *For parameters $\boldsymbol{\theta_k} \in \Theta_k$, the partial LSE $\mathcal{R}_T^{(k)}$ satisfies the decomposition*

$$\begin{aligned} \mathcal{R}_T^{(k)}(\boldsymbol{\theta_k}) = \mu_k^2 &+ 2\mu_k \sum_{j=1}^{d}\sum_{l=1}^{r_{kj}} \omega_{kjl} m_{jl}^{(X,k)} + \sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{l=1}^{r_{ki}}\sum_{l'=1}^{r_{kj}} \omega_{kil}\omega_{kjl'} m_{il,jl'}^{(XX,k)} \\ &- 2\mu_k m^{(Y,k)} - 2\sum_{j=1}^{d}\sum_{l=1}^{r_{kj}} \omega_{kjl} m_{jl}^{(XY,k)}. \end{aligned} \tag{3.107}$$

In matrix notation, we get

$$\mathcal{R}_T^{(k)}(\boldsymbol{\theta_k}) = \boldsymbol{f_k}^\intercal Q_k \boldsymbol{f_k} - 2\boldsymbol{f_k}^\intercal \boldsymbol{c_k}. \tag{3.108}$$

We give a proof of this result in Appendix A.2.5.

3.2.2 Solutions

As discussed in introduction, we analyse the LSE minimization problem P_k in two steps: first, for fixed kernel density parameters, we minimize the LSE with respect to first order parameters, leading to a new objective function to minimize with respect to kernel density parameters.

3.2.2.1 A quadratic program

Fix the kernel density parameters $\tilde{\boldsymbol{\theta}}_{\boldsymbol{k}}$. We are interested in the minimisation of the partial LSE $\mathcal{R}_T^{(k)}$ with respect to first order parameters $\boldsymbol{f_k} = (\mu_k, \boldsymbol{\omega_k})$ only. Formally, define the program $p_k(\tilde{\boldsymbol{\theta}}_{\boldsymbol{k}})$ by

$$\min_{\boldsymbol{f_k}} \quad \mathcal{R}_T^{(k)}(\boldsymbol{f_k}, \tilde{\boldsymbol{\theta}}_{\boldsymbol{k}}) \qquad \text{s.t.} \quad \boldsymbol{f_k} \geq 0. \tag{3.109}$$

Denote by $\boldsymbol{f_k^\star}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{k}})$ the solutions of this program. This problem is a quadratic program (QP) with positivity constraints. One might be tempted to first compute the matrices in this QP formulation, then use standard solvers to estimate the model parameters (for example, dual-primal methods, see Vandenberghe [102]). The major difficulty is the pre-computation of the QP formulation, which in general, takes quadratic time. This complexity can be reduced

by assuming $r_{ij} = r$, $\tilde{\phi}_{ijl} = \tilde{\phi}_l$, for all i, j, l, and by choosing $\tilde{\phi}_l$ to be an exponential decay. In this case, the pre-computation of the QP formulation will still have linear complexity in time, which is slower than the method we develop in this work. The derivatives of the LSE are simply obtained, and the gradient of the LSE with respect to first order parameters $\boldsymbol{f_k}$ is

$$\nabla_{\boldsymbol{f_k}} \mathcal{R}_T^{(k)} = 2(Q_k \boldsymbol{f_k} - \boldsymbol{c_k}), \tag{3.110}$$

that is, for event types $i \in [d]$ and mixture indices $l \in [r_{ki}]$

$$\begin{aligned} \frac{\partial \mathcal{R}_T^{(k)}}{\partial \mu_k}(\mu_k, \boldsymbol{\omega_k}, \tilde{\boldsymbol{\theta}}_{\boldsymbol{k}}) &= 2\Big(\mu_k + \sum_{j=1}^{d} \sum_{l=1}^{r_{kj}} \omega_{kjl} m_{jl}^{(X,k)} - m^{(Y,k)}\Big), \\ \frac{\partial \mathcal{R}_T^{(k)}}{\partial \omega_{kil}}(\mu_k, \boldsymbol{\omega_k}, \tilde{\boldsymbol{\theta}}_{\boldsymbol{k}}) &= 2\Big(\mu_k m_{il}^{(X,k)} + \sum_{j=1}^{d} \sum_{l'=1}^{r_{kj}} \omega_{kjl'} m_{il,jl'}^{(XX,k)} - m_{jl'}^{(XY,k)}\Big). \end{aligned} \tag{3.111}$$

Similarly, the Hessian of the LSE with respect to first order parameters $\boldsymbol{f_k}$ is simply $H_{\boldsymbol{f_k}} \mathcal{R}_T^{(k)} = 2Q_k$, that is, for event types $i, j \in [d]$, and for mixture indices $l \in [r_{ki}]$ and $l' \in [r_{kj}]$

$$\begin{aligned} &\frac{\partial^2 \mathcal{R}_T^{(k)}}{\partial \mu_k^2}(\mu_k, \boldsymbol{\omega_k}, \tilde{\boldsymbol{\theta}}_{\boldsymbol{k}}) = 2, \qquad \frac{\partial^2 \mathcal{R}_T^{(k)}}{\partial \mu_k \partial \omega_{kil}}(\mu_k, \boldsymbol{\omega_k}, \tilde{\boldsymbol{\theta}}_{\boldsymbol{k}}) = 2m_{il}^{(X,k)}, \\ &\frac{\partial^2 \mathcal{R}_T^{(k)}}{\partial \omega_{kil} \partial \omega_{kjl'}}(\mu_k, \boldsymbol{\omega_k}, \tilde{\boldsymbol{\theta}}_{\boldsymbol{k}}) = 2m_{il,jl'}^{(XX,k)}. \end{aligned} \tag{3.112}$$

The existence and uniqueness of solutions to this problem is characterized by the eigenvalues of the Hessian. Let $u \in \mathbb{R}^{d+1}$, such that $u = (u_0, \ldots, u_d)$. We see that

$$u^\intercal H_{\boldsymbol{f_k}} \mathcal{R}_T^{(k)} u = \langle\!\langle (u_0 + \sum_{j=1}^{d} u_j \tilde{\varphi}_j^{(k)}(t))^2 \rangle\!\rangle_T \geq 0. \tag{3.113}$$

Proposition 3.2.2 (Positivity of the Hessian)**.** *The Hessian of the LSE $H_{\boldsymbol{f_k}} \mathcal{R}_T^{(k)}$ is positive semi-definite. Furthermore, in the uni-dimensional case $d = 1$, the Hessian is definite-positive. If $d > 1$, this is not necessarily the case.*

The previous result implies the existence, but not necessarily the uniqueness, of unconstrained minimizers of the LSE. We are now interested in these unconstrained minimizers using the stationary points of the LSE with respect to first order parameters $\boldsymbol{f_k}$. For a given matrix M, we denote by M^+ is Moore–Penrose inverse. Denote by r the rank of the matrix Q_k, and by $(d_i)_{i \in [r]}$ its strictly positive eigenvalues. Denote by D the diagonal matrix

$$D := \mathbf{diag}\Big(d_1, \ldots, d_r, 0, \ldots, 0\Big) \tag{3.114}$$

Using the spectral theorem, there exists an orthogonal matrix P such that

$$Q_k = PDP^\intercal. \tag{3.115}$$

It is clear that the Moore–Penrose inverses of D and Q_k are

$$D^+ = \mathbf{diag}\Big(\frac{1}{d_1}, \dots, \frac{1}{d_r}, 0, \dots, 0\Big), \quad Q_k^+ = PD^+P^\intercal. \tag{3.116}$$

The stationary points of the LSE for first order parameters are simply obtained using the gradient formula.

Proposition 3.2.3 (Stationary points of the LSE)**.** *The set of stationary points of the LSE $\mathcal{R}_T^{(k)}(\cdot, \tilde{\boldsymbol{\theta}}_{\boldsymbol{k}})$ with respect to first order parameters is*

$$\bar{\boldsymbol{f}}_{\boldsymbol{k}} = Q_k^+ \boldsymbol{c}_{\boldsymbol{k}} + \ker Q_k = PD^+P^\intercal \boldsymbol{c}_{\boldsymbol{k}} + P \ker D. \tag{3.117}$$

Furthermore, if the matrix Q_k is positive definite, then $\bar{\boldsymbol{f}}_{\boldsymbol{k}} = Q_k^{-1}\boldsymbol{c}_{\boldsymbol{k}}$.

We give a proof of this result in Appendix A.2.5. A simple corollary of this result is the value of the LSE at stationary points, and the value of the conditional intensity model at stationary points.

Corollary 3.2.1. *Let $\bar{\boldsymbol{f}}_{\boldsymbol{k}}$ be a stationary point of the LSE for first order parameters, that is, there exists $\boldsymbol{u} \in \ker Q_k$ such that $\bar{\boldsymbol{f}}_{\boldsymbol{k}} := Q_k^+ \boldsymbol{c}_{\boldsymbol{k}} + u$. Then the value of the LSE at $\bar{\boldsymbol{f}}_{\boldsymbol{k}}$ is*

$$\mathcal{R}_T^{(k)}(\bar{\boldsymbol{f}}_{\boldsymbol{k}}, \tilde{\boldsymbol{\theta}}_{\boldsymbol{k}}) = -\bar{\boldsymbol{f}}_{\boldsymbol{k}}{}^\intercal \boldsymbol{c}_{\boldsymbol{k}}, \quad = -\boldsymbol{c}_{\boldsymbol{k}}{}^\intercal Q_k^+ \boldsymbol{c}_{\boldsymbol{k}} - 2\boldsymbol{c}_{\boldsymbol{k}}{}^\intercal \boldsymbol{u}. \tag{3.118}$$

The corresponding conditional intensity is

$$\lambda_k(t) = \boldsymbol{c}_{\boldsymbol{k}}{}^\intercal Q_k^+ \boldsymbol{\varphi}_{\boldsymbol{k}} + \boldsymbol{\varphi}_{\boldsymbol{k}}{}^\intercal \boldsymbol{u}. \tag{3.119}$$

The stationary points of the LSE do not characterize the solutions to problem $p_k(\tilde{\boldsymbol{\theta}}_{\boldsymbol{k}})$ because of the positivity constraints $\boldsymbol{f}_{\boldsymbol{k}} \geq 0$, as discussed below. To deal with these constraints, the Lagrange multiplier associated to this problem is

$$\mathcal{L}_T^{(k)}(\boldsymbol{f}_{\boldsymbol{k}}, \tilde{\boldsymbol{\theta}}_{\boldsymbol{k}}, \boldsymbol{L}) := \mathcal{R}_T^{(k)}(\boldsymbol{f}_{\boldsymbol{k}}, \tilde{\boldsymbol{\theta}}_{\boldsymbol{k}}) - 2\boldsymbol{L}^\intercal \boldsymbol{f}_{\boldsymbol{k}}, \tag{3.120}$$

with a vector of coefficients $\boldsymbol{L}$ of the same size as $\boldsymbol{f}_{\boldsymbol{k}}$. Note that the gradient of the Lagrange multiplier is simply

$$\nabla_{\boldsymbol{f}_{\boldsymbol{k}}} \mathcal{L}_T^{(k)}(\boldsymbol{f}_{\boldsymbol{k}}, \tilde{\boldsymbol{\theta}}_{\boldsymbol{k}}, \boldsymbol{L}) = 2(Q_k \boldsymbol{f}_{\boldsymbol{k}} - \boldsymbol{c}_{\boldsymbol{k}} - \boldsymbol{L}). \tag{3.121}$$

Therefore, the KKT conditions associated to the minimization program $p_k(\tilde{\boldsymbol{\theta}}_{\boldsymbol{k}})$ are

1. Stationarity condition:

$$Q_k \boldsymbol{f}_{\boldsymbol{k}} = \boldsymbol{c}_{\boldsymbol{k}} + \boldsymbol{L}. \tag{3.122}$$

2. Primal feasibility:

$$\boldsymbol{f_k} \geq 0. \tag{3.123}$$

3. Dual feasibility:

$$\boldsymbol{L} \geq 0. \tag{3.124}$$

4. Complementary slackness

$$\boldsymbol{f_k}^\intercal \boldsymbol{L} = 0. \tag{3.125}$$

Note that $\boldsymbol{f_k} = 0$ can be excluded from the analysis of the complementary slackness condition as it violates the stationarity condition. In the rest of this section, we focus on solutions of the problem in the uni-dimensional case $d = 1$ with one feature $r_{11} = 1$, and leave the analysis of the general case for future work. In this case, we solve for $k = d = 1$, and first order parameters are then simply $\boldsymbol{f_1} = (\mu_1, \omega_{11})^\intercal$. As discussed above, the matrix Q_1 is positive definite in this case. Using the results above, the stationary points of the LSE in dimension $d = 1$ are

$$\bar{\mu}_1 = m^{(Y,1)} - m_{11}^{(X,1)} \frac{q_{11}^{(XY,1)}}{q_{11,11}^{(XX,1)}}, \quad \bar{\omega}_{11} = \frac{q_{11}^{(XY,1)}}{q_{11,11}^{(XX,1)}}. \tag{3.126}$$

The sign of the critical L_1 weight $\bar{\omega}_{11}$ is the sign of the temporal covariance $q_{11}^{(XY,1)}$, which depends on the feature process. Therefore it is clear that critical points of the LSE are not necessarily solutions of the constrained minimization problem. The Lagrange multiplier in this case is

$$\mathcal{L}_T^{(1)}(\mu_1, \omega_{11}, L_0, L_1) := \mathcal{R}_T^{(1)}(\mu_1, \omega_{11}, \tilde{\boldsymbol{\theta}}_{\boldsymbol{1}}) - 2L_0\mu_1 - 2L_1\omega_{11}. \tag{3.127}$$

Therefore, there are only three sets of parameters satisfying the complementary slackness condition in Equation (3.125): $(\mu_1, L_1) = (0, 0)$, $(L_0, \omega_1) = (0, 0)$, or $(L_0, L_1) = (0, 0)$. Using these conditions, we classify the solutions of the constrained problem into four regimes depending on the values of the temporal correlation $\frac{q_{11}^{(XY,1)}}{q_{11,11}^{(XX,1)}}$ and the ratio $\frac{m^{(Y,1)}}{m_{11}^{(X,1)}}$. Note that

$$m_{11}^{(X,1)} < m^{(Y,1)}. \tag{3.128}$$

Proposition 3.2.4 (Solutions of problem $p_1(\tilde{\boldsymbol{\theta}}_{\boldsymbol{1}})$)**.** *The solutions $\boldsymbol{f_1^\star} = (\mu_1^\star, \omega_{11}^\star)^\intercal$ to the constrained LSE minimization problem $p_1(\tilde{\boldsymbol{\theta}}_{\boldsymbol{1}})$ are:*

1. *Homogeneous Poisson solution $\Big(\mu_1^\star = \eta_T^1,\ \omega_{11}^\star = 0\Big)$.*
   If we have negative temporal covariance $q_{11}^{(XY,1)} < 0$, then the solution is the homogeneous Poisson fit. Note that in this case, the stationary point of the LSE corresponds to a positive baseline $\bar{\mu}_1 > 0$, and a negative L_1 weight $\bar{\omega}_{11} < 0$.

2. *Stable MHP solution* $\left(\mu_1^\star = \bar{\mu}_1, \omega_{11}^\star = \bar{\omega}_{11}\right)$.

   If we have temporal covariance satisfying $\frac{q_{11}^{(XY,1)}}{q_{11,11}^{(XX,1)}} \in [0,1)$*, then the solution is a stable MHP, and the first order parameters are critical points of the LSE.*

3. *Unstable MHP solution* $\left(\mu_1^\star = \bar{\mu}_1,\ \omega_{11}^\star = \bar{\omega}_{11}\right)$.

   If we have temporal covariance satisfying $\frac{q_{11}^{(XY,1)}}{q_{11,11}^{(XX,1)}} \in \left[1, \frac{m^{(Y,1)}}{m_{11}^{(X,1)}}\right)$*, then the solution is also an MHP, and the first order parameters are also critical points of the LSE. However, note that* $\bar{\omega}_{11} > 1$ *in this case, therefore the fitted MHP is unstable.*

4. *Degenerate MHP solution* $\left(\mu_1^\star = 0,\ \omega_{11}^\star = \frac{m_{11}^{(XY,1)}}{m_{11,11}^{(XX,1)}}\right)$.

   If we have excess temporal correlation $\frac{q_{11}^{(XY,1)}}{q_{11,11}^{(XX,1)}} \in \left[\frac{m^{(Y,1)}}{m_{11}^{(X,1)}}, +\infty\right)$*, then the solution is a degenerate MHP, because the fitted baseline* $\mu_1^\star$ *is null, and the corresponding* L_1 *weight* $\omega_{11}^\star$ *is strictly greater than* 1*. Therefore, when simulating paths from this fitted model, there is a.s. no event happening, but if a baseline event happens, then this event has a.s. infinite descendance in finite time. Note that in this case, the stationary point of the LSE corresponds to a negative baseline* $\bar{\mu}_1 < 0$*, and a positive* L_1 *weight* $\bar{\omega}_{11} > 0$*.*

   We give a proof of this result in Appendix A.2.5.

3.2.2.2 Solving for kernel densities

Recall that $\boldsymbol{f}_{\boldsymbol{k}}^\star(\tilde{\boldsymbol{\theta}}_{\boldsymbol{k}})$ denotes the first order parameters minimizing the LSE, and denote the minimal value of the LSE for kernel density parameters $\tilde{\boldsymbol{\theta}}_{\boldsymbol{k}}$ by

$$\tilde{\mathcal{R}}_T^{(k)}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{k}}) := \mathcal{R}_T^{(k)}\big(\boldsymbol{f}_{\boldsymbol{k}}^\star(\tilde{\boldsymbol{\theta}}_{\boldsymbol{k}}), \tilde{\boldsymbol{\theta}}_{\boldsymbol{k}}\big). \tag{3.129}$$

We are now interested in finding solutions $\tilde{\boldsymbol{\theta}}_{\boldsymbol{k}}$ to the problem $\tilde{P}_k$ defined by

$$\min_{\tilde{\boldsymbol{\theta}}_{\boldsymbol{k}}} \quad \tilde{\mathcal{R}}_T^{(k)}\big(\tilde{\boldsymbol{\theta}}_{\boldsymbol{k}}\big) \qquad \text{s.t.} \quad \tilde{\boldsymbol{\theta}}_{\boldsymbol{k}} \in \tilde{\Theta}_k. \tag{3.130}$$

We get

$$\tilde{\mathcal{R}}_T^{(k)}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{k}}) = \begin{cases} -\left(m^{(Y,1)}\right)^2 & \text{if } \frac{q_{11}^{(XY,1)}}{q_{11,11}^{(XX,1)}} < 0, \\ -\frac{\left(q_{11}^{(XY,1)}\right)^2}{q_{11,11}^{(XX,1)}} - \left(m^{(Y,1)}\right)^2 & \text{if } \frac{q_{11}^{(XY,1)}}{q_{11,11}^{(XX,1)}} \in \left[0, \frac{m^{(Y,1)}}{m_{11}^{(X,1)}}\right), \\ -\frac{\left(m_{11}^{(XY,1)}\right)^2}{m_{11,11}^{(XX,1)}} & \text{if } \frac{q_{11}^{(XY,1)}}{q_{11,11}^{(XX,1)}} \geq \frac{m^{(Y,1)}}{m_{11}^{(X,1)}}. \end{cases} \tag{3.131}$$

Note that the function $\tilde{\mathcal{R}}_T^{(k)}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{k}})$ depends continuously on the temporal and counting averages. The following lemma helps compare the values of the LSE on the different domains.

Lemma 3.2.1 (Bounds on the LSE)**.** *The LSE* $\tilde{\mathcal{R}}_T^{(k)}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{k}})$ *satisfies the following bounds. If* $\frac{q_{11}^{(XY,1)}}{q_{11,11}^{(XX,1)}} \in \left[0, \frac{m^{(Y,1)}}{m_{11}^{(X,1)}}\right)$*, that is, in the well-defined MHP case , then*

$$-\frac{m_{11,11}^{(XX,1)}}{\left(m_{11}^{(X,1)}\right)^2} < \frac{\tilde{\mathcal{R}}_T^{(k)}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{k}})}{\left(m^{(Y,1)}\right)^2} \leq -1. \tag{3.132}$$

If $\frac{q_{11}^{(XY,1)}}{q_{11,11}^{(XX,1)}} \geq \frac{m^{(Y,1)}}{m_{11}^{(X,1)}}$*, that is, in the degenerate MHP case, then*

$$\frac{\tilde{\mathcal{R}}_T^{(k)}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{k}})}{\left(m^{(Y,1)}\right)^2} \leq -\frac{m_{11,11}^{(XX,1)}}{\left(m_{11}^{(X,1)}\right)^2}. \tag{3.133}$$

We give a proof of this result in Appendix A.2.5. Recall that $m^{(Y,1)} = \eta_T^1$ only depends on the observed data path and not on the MHP model. This lemma gives an insight on the solution regimes for $\tilde{P}_k$, but we could not get an analytical expression for the temporal correlation $\frac{q_{11}^{(XY,1)}}{q_{11,11}^{(XX,1)}} \in \left[0, \frac{m^{(Y,1)}}{m_{11}^{(X,1)}}\right)$. Therefore:

- The solution to the full problem P_k is the homogeneous Poisson solution if and only for all kernel density parameters $\tilde{\boldsymbol{\theta}}_{\boldsymbol{k}}$, the temporal correlation is negative.
- If there exists kernel density parameters $\tilde{\boldsymbol{\theta}}_{\boldsymbol{k}}$ such that the temporal correlation $\frac{q_{11}^{(XY,1)}}{q_{11,11}^{(XX,1)}}$ is positive, it is not clear under which condition the solution is non-degenerate. it is not clear whether for a fixed value of the ratio $\frac{m_{11,11}^{(XX,1)}}{\left(m_{11}^{(X,1)}\right)^2}$, we can control the ratio $\frac{m_{11}^{(XY,1)}}{m_{11}^{(X,1)}}$.

3.2.3 Properties of fitted models

In this paragraph, suppose the least-squares fit is a non-degenerate MHP. In this case, for all event types $k \in [d]$, the optimal first order parameters $(\boldsymbol{\mu}_{\boldsymbol{k}}^{\star}, \boldsymbol{\omega}_{\boldsymbol{k}}^{\star})$ are stationary points of the partial LSE $\mathcal{R}_T^{(k)}$.

3.2.3.1 Fitted baseline and first order moment

Recall that for event types $k \in [d]$, stationary points of the LSE verify

$$\mu_k^{\star} = \eta_T^k - \boldsymbol{\omega}_{\boldsymbol{k}}^{\star} \boldsymbol{m}^{(\boldsymbol{X},\boldsymbol{k})}. \tag{3.134}$$

Define the $d \times d$ squared matrix F, where for all event types $i, j \in [d]$

$$F_{ij} := \frac{1}{N_T^j} \sum_{l=1}^{r_{ij}} \omega_{ijl}^{\star} \sum_{n=1}^{N_T^j} \tilde{\psi}_{ij}(T - t_n^j). \tag{3.135}$$

This directly implies the following result.

Proposition 3.2.5 (Fitted baseline)**.** *If the parameters minimizing the LSE are stationary points of the LSE, then*

$$\boldsymbol{\mu}^\star = \big(\mathbb{I}_d - F\big)\boldsymbol{\eta_T}. \tag{3.136}$$

Note that for all $i, j \in [d]$, $F_{ij} \leq \omega^\star_{ijl}$, hence in matrix notation $F \leq ||\Phi^\star||_1$. Therefore, the minimizer vector $\bar{\boldsymbol{\mu}}$ verifies the bounds

$$\big(\mathbb{I}_d - ||\Phi^\star||_1\big)\boldsymbol{\eta_T} \leq \boldsymbol{\mu}^\star \leq \boldsymbol{\eta_T}. \tag{3.137}$$

Now assume that the kernel parameter values are such that the fitted MHP is stationary, and denote by $\bar{\boldsymbol{\eta}}_\star$ the stationary regime intensity of the fitted MHP. Using our result above, we can see how the stationary regime intensity $\bar{\boldsymbol{\eta}}_\star$ of the least squares fitted MHP relates to the observed empirical rate $\boldsymbol{\eta_T}$. In least squares estimation of homogeneous Poisson processes, the stationary regime intensity of the least squares fit is equal to the empirical rate. For the estimation of MHP besides the least squares method, the moment estimation method of Bacry and Muzy [10] matches moments of fitted models and empirical moments by construction. This is not the case for the LSE minimizer. Define the matrix $G := ||\Phi^\star||_1 - F$. For all event types $i, j \in [d]$,

$$G_{ij} = \frac{1}{N_T^j} \sum_{l=1}^{r_{ij}} \omega^\star_{ijl} \sum_{n=1}^{N_T^j} \Big(1 - \tilde{\psi}_{ijl}\big(T - t_n^j\big)\Big). \tag{3.138}$$

Proposition 3.2.6 (Fitted stationary regime intensity)**.** *The stationary regime intensity of the least squares minimizer verifies*

$$\bar{\boldsymbol{\eta}}_\star = \boldsymbol{\eta_T} + (\mathbb{I}_d - ||\Phi^\star||_1)^{-1} G \boldsymbol{\eta_T}. \tag{3.139}$$

We give a proof of this result in Appendix A.2.5. This result implies, in particular, that

$$\bar{\boldsymbol{\eta}}_\star > \boldsymbol{\eta_T}. \tag{3.140}$$

Remark 3.2.1 (Uni-dimensional case)**.** *Note that in dimension $d = 1$, the previous result reads*

$$\bar{\eta}^1_\star = \eta^1_T + \frac{\omega^\star_{11}}{1 - \omega^\star_{11}} \frac{1}{N_T^j} \sum_{n=1}^{N_T^j} \Big(1 - \tilde{\psi}_{ij}\big(T - t_n^j\big)\Big)\eta_T^{(1)}. \tag{3.141}$$

The function $\omega_{11} \mapsto \frac{\omega_{11}}{1-\omega_{11}}$ is a strictly increasing bijection from $[0, 1)$ to $[0, +\infty)$, hence the relative difference between the fitted SRI $\bar{\eta}^1_\star$ and the observed rate η^1_T increases with the fitted L_1 weight $\omega^\star_{11}$ and diverges as the fitted MHP gets to criticality.

3.2.3.2 Fitted intensity conditional on the training path

In the long path setup, we observe path of a counting process $\boldsymbol{N}$, generated by a ground truth conditional intensity $\boldsymbol{\lambda}^{\diamond}$. Denote by $\boldsymbol{\lambda}^{\star}$ the fitted conditional intensity model evaluated on that training path.

Proposition 3.2.7 (Properties of the least squares fitted intensity)**.** *Then for times $t \geq 0$*

$$\lambda_1(t) = m^{(Y,1)} + \frac{q_{11}^{(XY,1)}}{\sqrt{q_{11,11}^{(XX,1)}}} \tilde{\varphi}_{11}^{(1)}(t). \tag{3.142}$$

Therefore, the temporal average of the conditional intensity λ_1 on the training path is

$$\langle\!\langle \lambda_1 \rangle\!\rangle_T = m^{(Y,1)} = \eta_T^1. \tag{3.143}$$

The second order temporal averages of the conditional intensity λ_1 on the training path are

$$\langle\!\langle (\lambda_1)^2 \rangle\!\rangle_T = \left(m^{(Y,1)}\right)^2 + \frac{\left(q_{11}^{(XY,1)}\right)^2}{q_{11,11}^{(XX,1)}}, \qquad \langle\!\langle \lambda_1, \lambda_1 \rangle\!\rangle_T = \frac{\left(q_{11}^{(XY,1)}\right)^2}{q_{11,11}^{(XX,1)}}. \tag{3.144}$$

Finally, the counting average and the second order temporal average are equal

$$\langle\!\langle \lambda_1 \rangle\!\rangle_{N^1,T} = \langle\!\langle (\lambda_1)^2 \rangle\!\rangle_T. \tag{3.145}$$

3.3 Marked time-dependent linear Hawkes process

In this section, we propose a more general family of point processes with linear autoregressive structure: the marked time-dependent linear Hawkes process (MTLH). This class of models aims to overcome two basic limitations of MHP models.

The first such limitation is that MHP models are not designed to use information associated to system events other than their timings. For instance, consider a system where each event has a scale, and where the intensity of events is affected not only by the timing of events, but also by this scale of events: magnitude of an earthquake in seismology, size of a price jump in finance, number of followers of a social media account in sociology.

Below, we emphasize the second fundamental limitation of MHP models. In real-world applications, there are several instances where point process data is clearly non-stationary: systems with seasonality, systems with regime switching, etc. MHP models are not designed to capture stable non-stationary dynamics, since MHP models are non-stationary if and only if their adjacency matrix is critical: non-stationarity and instability are equivalent for an MHP model. In fact, an MHP fit to non-stationary data can often be misleading. For example, consider the estimation of a non-homogeneous Poisson ground truth process and

an MHP model, that we develop below. Of course, the non-homogeneous Poisson ground truth does not belong to the parametric space of MHP models, nor to its closure, and a poor fit is to be expected: but the bigger issue is that even at a qualitative level, the fitted model and the ground truth have very distinct properties. Indeed, we give an example where least-squares fitting Poisson data with an MHP does not lead to a homogeneous Poisson fit. Therefore, the fitted model shows auto-covariance that is not present in the data: the fitted MHP with non-null kernels does not have independent increments, whereas the ground truth does.

3.3.1 Motivation

Consider an observation horizon $T > 0$, and a uni-dimensional data generating process $\lambda^\diamond$ which is a piece-wise constant Poisson process

$$\mu(t) := \mu_L \mathbb{1}_{[0,\frac{T}{2})}(t) + \mu_H \mathbb{1}_{[\frac{T}{2},T)}(t), \tag{3.146}$$

where $\mu_L < \mu_H$. This a simple non-homogeneous Poisson model with a low activity regime μ_L on the first half of the observation window, and a high activity regime μ_H for the second half. We want to fit an MHP model to this data. Note that the least squares minimizing homogeneous Poisson fit to this data is

$$\bar{\mu} = \eta_T^1, \quad \textit{with} \quad \mathbb{E}[\eta_T^1] = \frac{\mu_L + \mu_H}{2}. \tag{3.147}$$

First, consider an exponential MHP model with fixed decay rate $\beta = 0.5$. We simulate 10^3 paths of the ground truth up to T and compute the analytic minimizer of the LSE in (μ, ω). Figure 3.5 plots the distribution of this minimizer, and we see clearly that the result is an MHP with a relatively high self-excitation, on average $\omega \simeq 0.78$. Figure 3.6 plots the

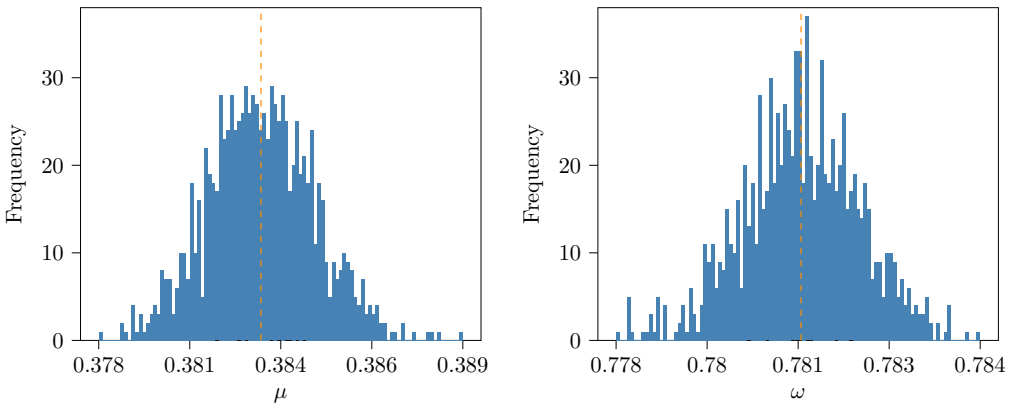


Figure 3.5: Distribution of the LSE analytic minimizer

Empirical distribution of analytic minimizer of the LSE for μ (left) and ω (right). Dashed orange lines plots the empirical averages.

distribution of the difference between the LSE of the fitted MHP model and the LSE of the homogeneous Poisson fit, for each path. We see that this difference is always strictly negative, which emphasizes the fact that the MHP with non-null kernels is indeed a better least squares fit to this data than the homogeneous Poisson. However, Figure 3.7 shows

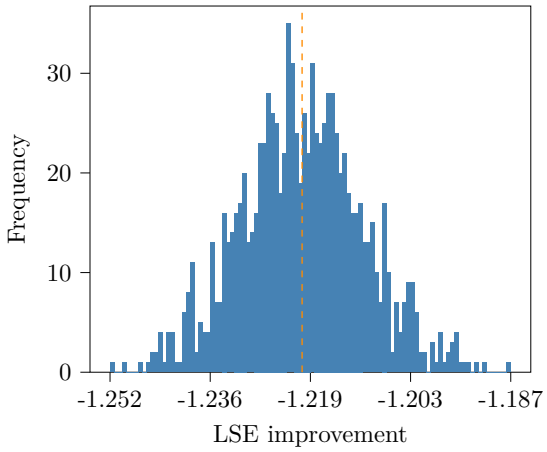


Figure 3.6: Distribution of the LSE improvement over Poisson model

Dashed orange line plots the empirical average.

that, while the MHP fit captures the total cumulative rate of the data, it does not have a satisfactory auto-covariance. Figure 3.8 plots the analytic minimizers $(\mu^\star, \omega^\star)$ of the LSE

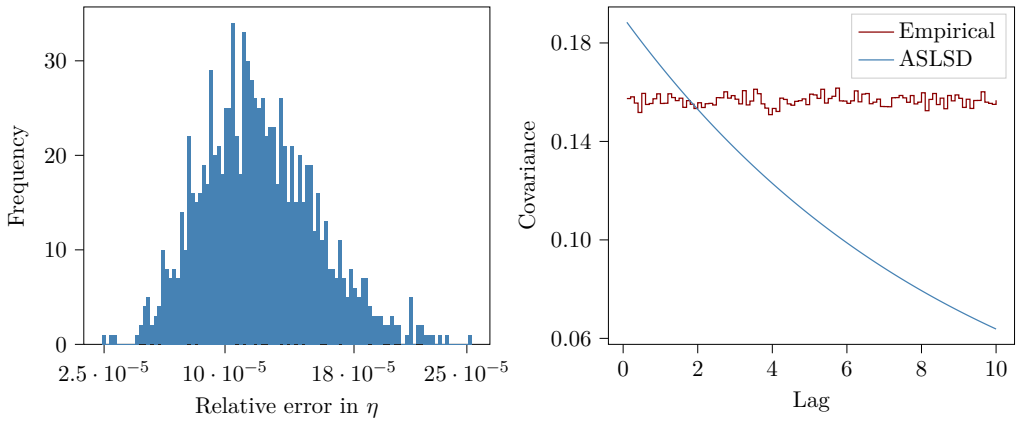


Figure 3.7: Distribution of moments

Left: we simulate 10^3 paths from the model, and for each path compute the cumulative event rate $\bar{\eta}_T$. We plot the empirical distribution of the relative errors $\frac{|\bar{\eta}_T - \eta_T|}{\eta_T}$. Right: Empirical covariance (red) and theoretical covariance of least-squares fit (blue) for sampling period $h = 10^{-1}$.

of the exponential SBF model with fixed decay rate β against β, and the corresponding LSE. The minimal LSE is monotonically increasing with the decay rate β, hence the LSE fit of non-SBF models degenerates to the MHP with null decay rate β, null baselines μ, and critical L_1 weight ω. These results indicate that significant care needs to be taken

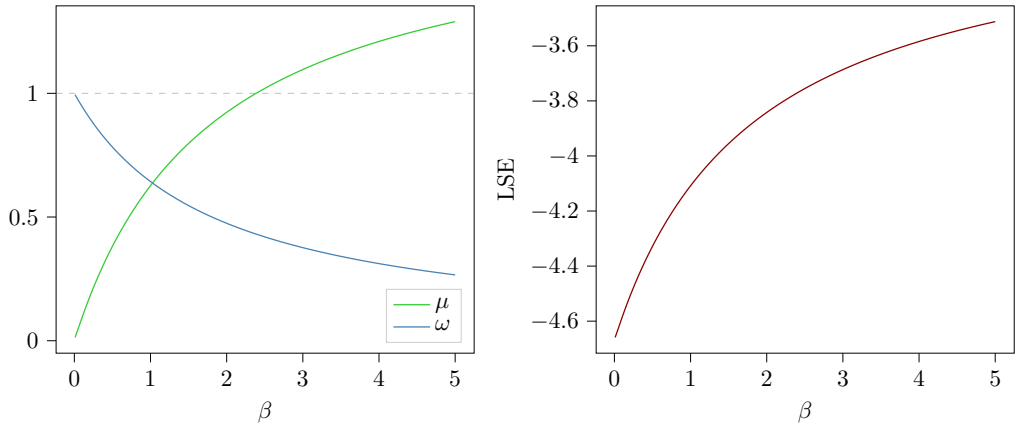


Figure 3.8: Analytic minimizer of the LSE

Left: Blue (resp. green) line plots the analytic minimizer ω (resp. μ) against the decay rate β. Right: minimal LSE against β.

when fitting self-exciting models in non-stationary envrionments, as the difference between self-excitation and other causes of baseline variation may be hard to distinguish.

3.3.2 Properties

Let $\boldsymbol{N}$ be a d-dimensional counting process with conditional intensity $\boldsymbol{\lambda}$. For all event types $i \in [d]$, fix $d^i_{mark} \in \mathbb{N}^*$. Let $\mathcal{Y}^i$ be a compact subset of $\mathbb{R}^{d^i_{mark}}$. For all event index $m \in \mathbb{N}^*$, we associate a mark $\xi^i_m \in \mathcal{Y}^i$ to the event happening at time t^i_m, where $(\xi^i_m)_{m\in\mathbb{N}^*}$ are i.i.d.random variables with distribution $\mathbb{P}_i$. Without loss of generality, we consider spaces of marks $\mathcal{Y}^i$ that are Cartesian products of either of the following sets

- Set of integers of the forms $[\![0, n]\!]$: for categorical marks;
- The closed interval $[0, 1]$: which by the compactness assumption represents all sets of continuous marks up to re-scaling.

Definition 3.3.1 (MTLH)**.** *We say that $\boldsymbol{N}$ is a marked time-dependent linear Hawkes process (MTLH) if, for event types $i \in [d]$ and times $t \geq 0$,*

$$\lambda_i(t) = \mu_i(t) + \sum_{j=1}^{d} \sum_{m:t^j_m<t} \phi_{ij}(t - t^j_m)\mathcal{I}_{ij}(\xi^j_m), \tag{3.148}$$

where

- *$\forall i, j \in [d]$, $\phi_{ij} : [0, +\infty) \to [0, +\infty)$ is in L_1. The functions ϕ_{ij} are called the kernels of the MHP, and we write, in matrix notation, $\Phi(\cdot) = \big(\phi_{ij}(\cdot)\big)_{ij}$.*
- *$\forall i \in [d]$, $\mu_i : [0, +\infty) \to (0, +\infty)$ is differentiable by parts. The functions μ_i are called baseline intensities, and we write, in vector notation, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^\intercal$.*

- $\forall i, j \in [d]$, $\mathcal{I}_{ij} : \mathcal{Y}^j \to [0, +\infty)$ *is differentiable by parts. The functions* $(\mathcal{I}_{ij})_{i,j\in[d]}$ *are called impact functions. For all* $i, j \in [d]$*, we impose the model identifiability condition*

$$\mathcal{I}_{ij}(0) = 1. \tag{3.149}$$

We refer to such a process as a $(\boldsymbol{\mu}, \Phi, \boldsymbol{\mathcal{I}})$*-MTLH, and to the triplet* $(\boldsymbol{\mu}, \Phi, \boldsymbol{\mathcal{I}})$ *as model functions.*

Figure 3.9 plots an example of MTLH conditional intensity. The set of MHP is obviously

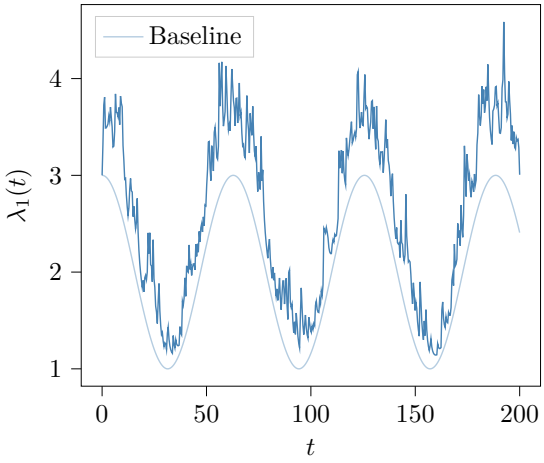


Figure 3.9: Conditional intensity of a uni-dimensional MTLH

Consider a uni-dimensional MTLH model with cosine baseline, a power law kernel with L_1 weight $\omega_{11} = 0.4$ and decay exponent $\alpha_{11} = 1$, random uniform marks on $[0, 1]$, and exponential impact function. We simulate a path of this MTLH up to $T = 200$. For times $t \in [0, T]$, the thick blue line plots the conditional intensity $\lambda_1(t)$ for the simulated path, and the thin blue line plots the baseline $\mu_1(t)$.

included in the set of MTLH. We choose our normalisation condition (on impact functions) so that the only admissible constant impact function is $\mathcal{I}_{ij} = 1$. Hence, a $(\boldsymbol{\mu}, \Phi, \boldsymbol{\mathcal{I}})$-MTLH with constant baselines $\boldsymbol{\mu}$ and constant impact functions $\boldsymbol{\mathcal{I}}$ is of course a $(\boldsymbol{\mu}, \Phi)$-MHP. The set of MTLH models is richer through the addition of two features:

- First, through the addition of time dependence in the baselines. While this results in the MTLH not being a stationary process, several real life systems, in particular financial markets, exhibit multi-scale behaviours which cannot be accounted for solely by the self-excitation terms.
- Second, through the presence of the impact functions. The influence of impact functions on the conditional intensity appears as a multiplicative term in front of the time kernels, which does not affect the distribution of any given time offset in the branching representation of the MTLH, but allow marks to vary the distribution of the number of

descendants. Indeed, the times of the offsets are still distributed according to the density $\tilde{\phi}_{ij}$. Given a mark ξ_m^j, the number of descendants of type i of the event at t_m^j is Poisson distributed with mean $\mathcal{I}_{ij}(\xi_m^j)||\tilde{\phi}_{ij}||_1$, instead of $||\tilde{\phi}_{ij}||_1$ in the MHP case.

Remark 3.3.1 (Marked Hawkes Process)**.** *In the literature, different authors sometimes use different definitions for the notion of "Marked Hawkes Process". In our MTLH model, there are d^2 independent impact functions $\boldsymbol{\mathcal{I}}$. Liniger [61] defines marked Hawkes processes (with i.i.d. marks) using d impact functions: the author assumes that for all event types $i, j, k \in [d]$, $\mathcal{I}_{ik} = \mathcal{I}_{jk}$. We keep a more general definition to preserve the decoupling of estimation programs across event types, as we discuss in the next subsection. This property is not necessary for the* `ASLSD` *method to work, but allows us to parallelize independent optimization programs.*

3.3.2.1 Modelling with MTLH

Parameterization We now introduce notation for MTLH models. Similarly to MHP models, define the vectors of kernel L_1 weights $\boldsymbol{\omega_k}$, and the vector of kernel densities parameters $\tilde{\boldsymbol{\theta}}_{\boldsymbol{k}}$. For event types $i \in [d]$, denote the vector of parameters of the impact function $\mathcal{I}_{ki}$ by $\boldsymbol{\vartheta_{ki}} = (\vartheta_{ki,p})_{p \in r_{ki}^{(\mathcal{I})}}$. We concatenate these vectors to define the vector $\boldsymbol{\vartheta_k}$ by

$$\boldsymbol{\vartheta_k}^\intercal = (\boldsymbol{\vartheta_{k1}}^\intercal, \dots, \boldsymbol{\vartheta_{kd}}^\intercal). \tag{3.150}$$

We denote the total number of impact parameters by $r_k^{\mathcal{I}} = \sum_{i=1}^d r_{ki}^{\mathcal{I}}$. Denote by $\boldsymbol{b_k}$ the parameters of the baseline function μ_k as $\boldsymbol{b_k}^\intercal := (b_{kp})_{p \in [r_k^{(\mu)}]}$. Finally, define $\boldsymbol{\theta_k}$, the vector of parameters of MTLH model functionals of dimension k by

$$\boldsymbol{\theta_k}^\intercal = (\boldsymbol{b_k}^\intercal, \boldsymbol{\omega_k}^\intercal, \tilde{\boldsymbol{\theta}}_{\boldsymbol{k}}^\intercal, \boldsymbol{\vartheta_k}^\intercal).$$

The total number of parameters of the MHP model across dimensions $k \in [d]$ is

$$n_{\text{param}} := \sum_{k=1}^d \left(r_k^{(\mu)} + r_k + \tilde{r}_k + r_k^{(\mathcal{I})} \right).$$

Residuals In the rest of this paragraph, fix an event type $k \in [d]$. For event indices $m \in [N_T^k]$, define the m-th compensator-transformed time of type k

$$s_m^{(k)} := \mathbb{E}_{\xi^j}[\, \Lambda_k \left(t_m^k \right)]\,. \tag{3.151}$$

Using the same approach as in Lemma A.3.1, we get the following formula for MTLH residuals.

Proposition 3.3.1 (Residuals of MTLH models)**.** *For event indices $m \in [N_T^k]$, the m-th compensator-transformed time of type k is*

$$s_m^{(k)} = \int_0^{t_m^k} \mu_k(t)\mathrm{d}t + \sum_{i=1}^{d} \mathbb{E}[\mathcal{I}_{ki}(\xi_m^i)]\, \mathbb{1}_{\{m \geq \varpi(k,i)\}} \sum_{n=1}^{\kappa(i,k,m)} \psi_{ki}(t_m^k - t_n^i), \tag{3.152}$$

3.3.2.2 Marks and impact functions

Following our definition of MTLH, marks can be discrete or continuous, uni-dimensional or multi-dimensional. In this paragraph, we propose some parametric families of impact functions that might be useful for modelling. The case of continuous marks is significantly less studied in the Hawkes literature. In this paragraph, we discuss two examples of simple impact functions for continuous marks: the sigmoid impact for uni-dimensional continuous marks, and the perceptron impact for multi-dimensional continuous marks.

Categorical marks First, we discuss the standard set of uni-dimensional, discrete marks. Suppose the marks set $\mathcal{Y}$ has finite cardinality J. Let $\mathcal{I} : \mathcal{Y} \to [0, +\infty)$ be an impact function. Using the positivity and model identifiability conditions on the impact function , there exists $\boldsymbol{\beta} = (\beta_k)_{k \in [J-1]} \in \mathbb{R}^{J-1}$ such that

$$\mathcal{I}(\xi) = \begin{cases} 1 & \text{if } \xi = 0 \\ \beta_\xi & \text{if } \xi \in [J-1]. \end{cases} \tag{3.153}$$

Therefore, the categorical impact function is simply parameterized by the vector of weights $\boldsymbol{\beta}$. The gradient of this impact function can be efficiently parallelized since for indices $p \in [J-1]$,

$$\frac{\partial \mathcal{I}}{\partial \beta_p}(\xi) = \mathbb{1}_{\{\xi = p\}}. \tag{3.154}$$

Note than an MTLH model with categorical marks and constant baseline corresponds to an MHP of higher dimension.

Perceptron We place ourselves in the case of a continuous, multi-dimensional compact marks set $\mathcal{Y} = [0,1]^{d_{\text{marks}}}$. Denote by f_σ the standard logistic function, such that for $x \in \mathbb{R}$

$$f_\sigma(x) := \frac{1}{1+e^{-x}}, \quad f'_\sigma(x) := f_\sigma(x)(1 - f_\sigma(x)). \tag{3.155}$$

We define the perceptron impact following standard deep learning definitions, with the normalisation condition of impact functions.

Definition 3.3.2 (Perceptron impact)**.** *For marks $\boldsymbol{\xi} \in [0,1]^{d_{marks}}$, the perceptron impact is*

$$\mathcal{I}(\boldsymbol{\xi}) := \frac{f_\sigma(a_0 + \boldsymbol{a}^\intercal \boldsymbol{\xi})}{f_\sigma(a_0)}, \tag{3.156}$$

where the coefficient $a_0 \in \mathbb{R}$ is a bias term, and the parameters $(a_l)_{\{l \in [d_{marks}]\}} \in \mathbb{R}^{d_{marks}}$.

The gradient of the perceptron impact is particularly simple to compute in parallel,

$$\frac{\partial \mathcal{I}}{\partial a_0}(\boldsymbol{\xi}) = f_\sigma(a_0)\mathcal{I}(\boldsymbol{\xi})(1 - \mathcal{I}(\boldsymbol{\xi})), \tag{3.157}$$

and for indices $p \in [d_{\text{marks}}]$

$$\frac{\partial \mathcal{I}}{\partial a_p}(\boldsymbol{\xi}) = \xi_p \mathcal{I}(\boldsymbol{\xi})(1 - f_\sigma(a_0)\mathcal{I}(\boldsymbol{\xi})). \tag{3.158}$$

3.3.3 Cluster modelling

The branching representation of Hawkes models is particularly insightful for several modelling tasks as the parent-child logic of the branching representation can provide practical and interpretable models. Given a Hawkes model $\boldsymbol{\lambda}$, we study the branching representation of events under this model given data paths in two setups:

- **synthetic setup**: the data paths are simulated from the model $\boldsymbol{\lambda}$ through our exact branching simulation algorithm. Therefore, in this setup, the true branching representation of each path is known. This setup serves two objectives. First, to study numerically properties of the branching representation of data paths under correctly specified models, since many statistical properties of event clusters under Hawkes models are not known theoretically (for example, the distribution of relaxation times). Second, this setup is fundamental to tune cluster models before applying them to real data in the next setup.
- **realized setup**: the data paths are observed and the modeller does not know whether they have been generated by the model $\boldsymbol{\lambda}$. For example, these data paths could come from a given real-world dataset, and the model $\boldsymbol{\lambda}$ is the Hawkes fit to this data.

Stochastic declustering is the identification of probable parent–child links in an observed dataset. Discussing the full stochastic declustering of a data path, in any of the 2 setups above, is beyond the scope of this work. Instead, we focus on two aspects of cluster modelling. In Section 3.3.3.1, we discuss the binary classification problem of endogenous and exogenous events in a data path. In Section 3.3.3.2, we discuss data censoring and the impact of events. This discussion applies to both MHP and MTLH models: since these models classes are nested, we use the notation of MTLH models.

3.3.3.1 Classifying exogenous events

We are interested in the binary classification of events in the observed path as exogenous (to which we associate the label 1) or endogenous (label 0) under the model. In accordance with standard terminology of binary classification problems, we may also refer to exogenous (label 1) events as positives, and endogenous (label 0) events as negatives.

Decision Fix a decision threshold $\epsilon \in [0,1]$. We restrict ourselves to a very simple class of models Clf_ϵ.

Definition 3.3.3 (Exogenous event classifier)**.** *The model Clf_ϵ is such that for event types $i \in [d]$, and jump indices $m \in [N_T^i]$, the predicted label of the m-th jump of type i is*

$$Clf_\epsilon(i,m) = \begin{cases} 1 & \text{if } \frac{\mu_i(t_m^i)}{\lambda_i(t_m^i)} \geq \epsilon, \\ 0 & \text{otherwise.} \end{cases} \tag{3.159}$$

Of course, this model is motivated by the cluster probabilities of Hawkes models, which state that the exogeneity probability is $p_{i,m,m} := \frac{\mu_i(t_m^i)}{\lambda_i(t_m^i)}$. It is clear that the set of events classified as exogenous is monotonically decreasing (for the inclusion) with the decision threshold ϵ. For all values of $\epsilon \in [0,1]$, the set of positives is non-null: the first event of the path is always classified exogenous. For a maximal decision threshold $\epsilon = 1$, if all kernels and all impact functions are strictly positive, the first event in the path is the only one classified as exogenous. For a null decision threshold $\epsilon = 0$, all the events in the path are classified as exogenous. A fundamental question now is the fine tuning of the decision threshold ϵ.

Fine-tuning ϵ In order to fine tune the hyper-parameter ϵ, we place ourselves in the synthetic setup, simulate paths from the Hawkes model for which the ground truth classes are known, and use standard evaluation metrics for binary classification problems. Definition for these metrics can be found, for example, in Murphy [71] (page 183, 5.7.2.3): in particular, we focus on the precision $P(\text{Clf}_\epsilon)$ and the recall $R(\text{Clf}_\epsilon)$

$$P(\text{Clf}_\epsilon) := \frac{TP}{TP+FP}, \quad R(\text{Clf}_\epsilon) := \frac{TP}{TP+FN}, \tag{3.160}$$

and the $F1$ score $F_1(\text{Clf}_\epsilon)$, the harmonic mean of precision and recall

$$F_1(\text{Clf}_\epsilon) = 2\frac{P(\text{Clf}_\epsilon) \times R(\text{Clf}_\epsilon)}{P(\text{Clf}_\epsilon) + R(\text{Clf}_\epsilon)}. \tag{3.161}$$

The relative importance given to precision with respect to recall depends on the objective of a study and the type of application. The F_1 score gives the same importance to precision and recall. Therefore, as discussed by Baeza-Yates and Ribeiro-Neto [11] (page 327, 8.6.4), the F_β score can be a more appropriate metric. Let $\beta \geq 0$, the F_β score $F_\beta(\text{Clf}_\epsilon)$ is

$$F_\beta(\text{Clf}_\epsilon) = (\beta^2+1)\frac{P(\text{Clf}_\epsilon) \times R(\text{Clf}_\epsilon)}{\beta^2 P(\text{Clf}_\epsilon) + R(\text{Clf}_\epsilon)}. \tag{3.162}$$

The importance given to recall increases with the parameter β.

3.3.3.2 Modelling descent

Fix a horizon $T > 0$ and suppose we observe a path of the Hawkes model on $[0, T]$. Fix an event type $i \in [d]$, and an event index $m \in [N_T^i]$. We consider the impact of the event t_m^i.

Definition 3.3.4 (Descendants set)**.** *Denote by $\mathcal{D}_{T,i,m}^{(j,p)}$ the set of indices of p-th generation descendants of type $j \in [d]$ of the event t_m^i in the window $[0, T]$ by*

$$\mathcal{D}_{T,i,m,p}^{(j)} := \{n \in [N_T^j] : \quad t_n^j \text{ is a } p\text{-th generation descendant of } t_m^i \text{ and } t_n^j < T\}. \quad (3.163)$$

When $p = 1$, $\mathcal{D}_{T,i,m,p}^{(j)}$ is the set of direct descendants of type j of t_m^i.

Denote by $\mathcal{D}_{T,i,m}^{(j)}$ the set of indices of descendants of type $j \in [d]$ of the event t_m^i in the window $[0, T]$,

$$\mathcal{D}_{T,i,m}^{(j)} := \{n \in [N_T^j] : \quad t_n^j \text{ is a descendant of } t_m^i \text{ and } t_n^j < T\}. \quad (3.164)$$

Of course, $\mathcal{D}_{T,i,m}^{(j)} = \bigcup_{p=1}^{+\infty} \mathcal{D}_{T,i,m,p}^{(j)}$. By abuse of notation, we refer to $\mathcal{D}_{T,i,m}$ as the set of descendants of t_m^i of all types $\mathcal{D}_{T,i,m} = \bigcup_{j=1}^{d} \mathcal{D}_{T,i,m}^{(j)}$.

We censor the event t_m^i and its descendants from the path of the counting process as follows.

Definition 3.3.5 (Censored path)**.** *For all event types $j \in [d]$, we denote by $\tilde{N}^j$ the counting process obtained by censoring the event t_m^i and its descendants from N^j. That is, $N_t^j = \tilde{N}_t^j$ for all $j \in [d]$, and for all $t < t_m^i$. For all $t \geq t_m^i$,*

$$N_t^j = \tilde{N}_t^j + N^j(\{t_m^i\}) + N^j\left((t_m^i, t] \cap \mathcal{D}_{i,m}^{(j)}\right). \quad (3.165)$$

Definition 3.3.6 (Relaxation time)**.** *We define the relaxation time $\tau_{i,m} \geq 0$ of the perturbation induced by t_m^i by*

$$\tau_{i,m} := \max\{t_n^j, j \in [d], n \in \mathcal{D}_{T,i,m}^{(j)}\} - t_m^i, \quad (3.166)$$

with the convention that $\max\{t_n^j, j \in [d], n \in \mathcal{D}_{i,m}^{(j)}\} = t_m^i$ if $\cup_{j\in[d]} \mathcal{D}_{i,m}^{(j)} = \emptyset$, that is: we set the relaxation time $\tau_{i,m}$ to zero if the event t_m^i has no descendants in $[0, T]$.

Chapter 4

The ASLSD method

4.1 Notation and definitions

This chapter presents our procedure for the estimation of Hawkes models. Consider a d-dimensional Hawkes model $\boldsymbol{N}$. For the sake of brevity, assume $\boldsymbol{N}$ is an MTLH model, unless specified otherwise. Since MHP models are a particular case of MTLH models, the ASLSD procedure and all the results in this chapter apply directly to MHP by setting all impact functions to the constants $\mathcal{I}_{ki} = 1$. We present our procedure in the long path setup with horizon $T > 0$, but the extension to the episodic setup is immediate (see the example in Section 5.1). For the rest of this chapter, fix an event type $k \in [d]$.

4.1.1 Book-keeping

In dimension $d > 1$, the ordering of events times across dimensions requires some care. For our method, we first need to track the index of last events of a certain type before a given time. For all event types $i \in [d]$, and for all times $t > 0$, define

$$\kappa(i, t) := \lim_{s \uparrow 0} N^i_{t-s}. \tag{4.1}$$

Equivalently, $\kappa(i, t) = \sup\{m \in \mathbb{N}^*; t^i_m < t\}$, with the convention that if $\{m \in \mathbb{N}^*; t^i_m < t\} = \emptyset$ then $\kappa(i, t) = 0$. By abuse of notation, for all $i, j \in [d]$, for all $n \in \mathbb{N}^*$ define

$$\kappa(i, j, n) := \kappa(i, t^j_n); \tag{4.2}$$

that is, the index of the last event of type i before the n-th event of type j. Note that $\kappa(i, i, n) = n - 1$.

Second, for our method we need to track the index of the first element of a type preceded by a fixed number of events. Hence, for all event types $i, j \in [d]$, and for all indices $h \in \mathbb{N}^*$ define the upper inverse of $\kappa(j, i, \dot{})$ by

$$\varpi(i, j, h) := \inf\{p \in \mathbb{N}^* : \kappa(j, i, p) \geq h\}. \tag{4.3}$$

that is, the index of the earliest jump of type i preceded by at least h jumps of type j. For simplicity, we write

$$\varpi(i,j) := \varpi(i,j,1). \tag{4.4}$$

4.1.2 Correlation functionals

We now introduce a few functions measuring the cross-correlation of kernels and baselines. Fix event types $i, j \in [d]$, and lags $\tau, \sigma \geq 0$.

Kernels For mixture indices $l \in [r_{ki}]$, and $l' \in [r_{kj}]$, the cross-correlation function $\Upsilon_{ijk,ll'}$ between kernel densities $\phi_{ki,l}$ and $\phi_{kj,l'}$ is

$$\tilde{\Upsilon}_{ijk,ll'}(\tau,\sigma) := \int_0^\tau \tilde{\phi}_{ki,l}(u)\tilde{\phi}_{kj,l'}(u+\sigma)\mathrm{d}u, \tag{4.5}$$

and denote $\Upsilon_{ijk,ll'} := \omega_{ki,l}\omega_{kj,l'}\tilde{\Upsilon}_{ijk,ll'}$. The cross-correlation function Υ_{ijk} between the kernels ϕ_{ki} and ϕ_{kj} is

$$\Upsilon_{ijk} := \sum_{l=1}^{r_{ki}} \sum_{l'=1}^{r_{kj}} \Upsilon_{ijk,ll'}. \tag{4.6}$$

Note that the function $\tilde{\Upsilon}_{ijk}$ is bounded; using the Cauchy–-Schwarz inequality, for lags $\tau, \sigma \geq 0$

$$\tilde{\Upsilon}_{ijk}(\tau,\sigma) \leq ||\tilde{\phi}||_2 ||\tilde{\phi}_\star||_2. \tag{4.7}$$

For lags $\sigma \geq 0$, the function $\tau \mapsto \tilde{\Upsilon}_{ijk}(\tau,\sigma)$ is clearly non-decreasing.

Kernel–baseline For a mixture index $l \in [r_{ki}]$, the baseline-kernel cross-correlation $\tilde{K}_{ki}$ between the kernel density $\tilde{\phi}_{ki,l}$ and the baseline μ_k is

$$\tilde{K}_{ki,l}(\tau,\sigma) := \int_0^\tau \phi_{ki,l}(u)\mu_k(u+\sigma)\mathrm{d}u, \tag{4.8}$$

and denote $K_{ki,l} := \omega_{ki,l}\tilde{K}_{ki,l}$. Finally, the cross-correlation function K_{ki} between the kernel ϕ_{ki} and the baseline μ_k is

$$K_{ki} := \sum_{l=1}^{r_{ki}} K_{ki,l}. \tag{4.9}$$

Note that if $\boldsymbol{N}$ is an MHP, then the baseline-kernel cross correlation verifies

$$K_{ki}(\tau,\sigma) = \mu_k \psi_{ki}(\tau). \tag{4.10}$$

4.2 Motivation

Directly minimizing the LSE of Hawkes models, for example with exact first order methods, is particularly inefficient without further assumptions because of the auto-regressive structure of the conditional intensity.

Evaluating the LSE is expensive For times $t \geq 0$, evaluating the conditional intensity model $\lambda_i(t)$ based on Equation (3.148) has linear time complexity in N_t, the number of jumps of $\boldsymbol{N}$ up to time t. Therefore, evaluating the second term

$$\frac{2}{T}\sum_{k=1}^{d}\sum_{m=1}^{N_T^k}\lambda_k(t_m^k) \tag{4.11}$$

in the LSE in Equation (2.7) has time complexity $\mathcal{O}(N_T^2)$, without further assumptions on the kernel matrix Φ. Now, consider the evaluation of the first term

$$\frac{1}{T}\sum_{k=1}^{d}\int_0^T \lambda_k(t)^2 \mathrm{d}t \tag{4.12}$$

in the LSE. If the integral in Equation (4.12) is approximated numerically using a quadrature rule, the cost of such approximation is at least linear since the conditional intensity has a discontinuity at each jump time. If we re-use the evaluations of the conditional intensity at jump times, which we computed for the evaluation of Equation (4.11), we still face the problem that the conditional intensity varies between jump times. For general kernels, it is not clear how to interpolate the conditional intensity accurately between jump times, therefore, we might also have to evaluate $\boldsymbol{\lambda}$ between consecutive jump times. Another drawback of numerical integration is that the evaluation of the gradient would typically require a finite difference method or operator overloading, increasing the overall complexity.

In a nutshell, evaluating the LSE at a given parameter vector $\boldsymbol{\theta}$ has complexity $\mathcal{O}(N_T^2)$ in the general case, and the evaluation of the gradient in closed-form is roughly as expensive (if we ignore for now the curse of dimensionality coming from the number of parameters of the model). This makes exact first order methods impractical for general kernels, unless the kernels considered are sum-of-basis exponential functions.

Towards stochastic optimization A classic strategy to accelerate first order methods is the use of Stochastic Gradient Descent (SGD), which relies on an approximation of the gradient of the objective function; see Robbins and Monro [92]. The application of SGD to the minimization of the LSE in Equation (2.7) faces some difficulties. For example, for $d = 1$, and if $\boldsymbol{N}$ is an MHP model, write

$$\mathcal{R}_T(\theta) = \frac{1}{T}(t_1\mu^2 - 2\mu) + \frac{1}{T}\int_{t_{N_T}}^T \lambda(t)^2\mathrm{d}t + \frac{1}{T}\sum_{m=1}^{N_T-1}\left(\int_{t_m}^{t_{m+1}} \lambda(t)^2\mathrm{d}t - 2\lambda(t_{m+1})\right), \tag{4.13}$$

and define

$$f_m(\theta) := \begin{cases} \int_{t_m}^{t_{m+1}} \lambda(t)^2\mathrm{d}t - 2\lambda(t_{m+1}) & \text{if } m \in [\![1, N_T - 1]\!], \\ \int_{t_{N_T}}^{T} \lambda(t)^2\mathrm{d}t & \text{if } m = N_T, \end{cases} \tag{4.14}$$

to yield the decomposition

$$\mathcal{R}_T(\theta) = \frac{1}{T}(t_1\mu^2 - 2\mu) + \frac{1}{T}\sum_{m=1}^{N_T} f_m(\theta). \tag{4.15}$$

This decomposition preserves the chronological order of the data because, for all m, the function f_m only depends on jump times up to t_{m+1}. However, computing $f_m(\theta)$ and its derivatives has complexity roughly of order $\mathcal{O}(m)$. Due to this linear cost, we propose a new additive decomposition of the LSE and build a fast, yet accurate, Monte Carlo (MC) estimator of large finite sums.

4.3 LSE decomposition

Rewriting the LSE We state our decomposition of the partial LSE $\mathcal{R}_T^{(k)}$ as a sum involving the kernels $(\phi_{ki})_{i\in[d]}$, the baseline-kernel cross-correlations $(K_{ki})_{i\in[d]}$, the kernel cross-correlations $(\Upsilon_{ijk})_{i,j\in[d]}$, and the mean squared background rate M_k.

Theorem 4.3.1 (Least squares error)**.** *For parameters $\boldsymbol{\theta_k} \in \Theta_k$, the partial LSE $\mathcal{R}_T^{(k)}$ satisfies the decomposition*

$$\begin{aligned}\mathcal{R}_T^{(k)}(\boldsymbol{\theta_k}) = &\sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{m=\varpi(i,j)}^{N_T^i}\sum_{n=1}^{\kappa(j,i,m)}\frac{2}{T}\Upsilon_{ijk}(T-t_m^i, t_m^i - t_n^j)\mathcal{I}_{ki}(\xi_m^i)\mathcal{I}_{kj}(\xi_n^j)\\ &- \frac{2}{T}\left(\sum_{j=1}^{d}\sum_{m=\varpi(k,j)}^{N_T^k}\sum_{n=1}^{\kappa(j,k,m)}\phi_{kj}(t_m^k - t_n^j)\mathcal{I}_{kj}(\xi_n^j)\right) + M_k(T) - \frac{2}{T}\sum_{m=1}^{N_T^k}\mu_k(t_m^k)\\ &+ \frac{2}{T}\sum_{i=1}^{d}\sum_{m=1}^{N_T^i}K_{ki}(T-t_m^i, t_m^i)\mathcal{I}_{ki}(\xi_m^i) + \frac{1}{T}\sum_{i=1}^{d}\sum_{m=1}^{N_T^i}\Upsilon_{iik}(T-t_m^i, 0)\mathcal{I}_{ki}(\xi_m^i)^2.\end{aligned} \tag{4.16}$$

We give a proof of this expansion in Appendix A.3.1. Fix event types $i, j \in [d]$, and mixture indices $l \in [r_{ki}]$, $l' \in [r_{kj}]$, and define some notation for the different terms involved in the LSE decomposition of Theorem 4.3.1. First, we have single-indexed sums

$$\begin{aligned}Z_{K_{ki,l}} &:= \sum_{m=1}^{N_T^i} K_{ki,l}(T-t_m^i, t_m^i)\mathcal{I}_{ki}(\xi_m^i), \qquad Z_{\Upsilon_{iik,ll'}} := \sum_{m=1}^{N_T^i}\Upsilon_{iik,ll'}(T-t_m^i, 0)\mathcal{I}_{ki}(\xi_m^i)^2,\\ Z_{\mu_k} &:= \sum_{m=1}^{N_T^k}\mu_k(t_m^k).\end{aligned} \tag{4.17}$$

Second, we have double-indexed sums

$$\begin{aligned} S_{\Upsilon_{ijk,ll'}} &:= \sum_{m=\varpi(i,j)}^{N_T^i} \sum_{n=1}^{\kappa(j,i,m)} \Upsilon_{ijk,ll'}(T - t_m^i, t_m^i - t_n^j)\mathcal{I}_{ki}(\xi_m^i)\mathcal{I}_{kj}(\xi_n^j), \\ S_{\phi_{kj,l'}} &:= \sum_{m=\varpi(k,j)}^{N_T^k} \sum_{n=1}^{\kappa(j,k,m)} \phi_{kj,l'}(t_m^k - t_n^j)\mathcal{I}_{kj}(\xi_n^j). \end{aligned} \tag{4.18}$$

Therefore, the k-th partial LSE satisfies

$$\begin{aligned} \mathcal{R}_T^{(k)}(\theta_k) = \frac{2}{T}\sum_{i=1}^{d}\sum_{l=1}^{r_{ki}} Z_{K_{ki,l}} + \frac{1}{T}\sum_{i=1}^{d}\sum_{l=1}^{r_{ki}}\sum_{l'=1}^{r_{ki}} Z_{\Upsilon_{iik,ll'}} - \frac{2}{T} Z_{\mu_k} + M_k(T) \\ + \frac{2}{T}\sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{l=1}^{r_{ki}}\sum_{l'=1}^{r_{kj}} S_{\Upsilon_{ijk,ll'}} - \frac{2}{T}\sum_{j=1}^{d}\sum_{l'=1}^{r_{kj}} S_{\phi_{kj,l'}}. \end{aligned} \tag{4.19}$$

The only negative terms in the expression of the partial LSE are $-\frac{2}{T}\sum_{j=1}^{d}\sum_{l'=1}^{r_{kj}} S_{\phi_{kj,l'}}$ and $-\frac{2}{T} Z_{\mu_k}$. Each sum term is positive and minimized for null kernels, impact functions and baselines. Fix event types $i, j \in [d]$ with $i \neq j$; the double-indexed sums $\left(S_{\Upsilon_{ijk,ll'}}\right)_{ll'}$ allow the coupling between the kernels ϕ_{ki} and ϕ_{kj}, and the impact functions $\mathcal{I}_{ki}$ and $\mathcal{I}_{kj}$. The single-indexed sums $\left(Z_{K_{ki,l}}\right)_{l\in[r_{ki}]}$ allow the coupling between the baseline μ_k, the kernel ϕ_{ki}, and the impact function $\mathcal{I}_{ki}$.

Complexity The brute force computation of each of the double sums $S_{\Upsilon_{ijk,ll'}}$ and $S_{\phi_{kj,l'}}$ has quadratic time complexity in the number of events. To compute $S_{\Upsilon_{ijk,ll'}}$, at each event index $m \in [\![\varpi(i,j), N_T^i]\!]$, we need to evaluate the functions $\Upsilon_{ijk,ll'}$ and $\mathcal{I}_{kj}$ on the $\kappa(j,i,m)$ events of type j anterior to t_m^i. A crude inequality[1] for $\kappa(j,i,m)$ is $\kappa(j,i,m) \leq N_T^j$. Hence the inequality

$$\sum_{m=\varpi(i,j)}^{N_T^i} \kappa(j,i,m) \leq N_T^i N_T^j. \tag{4.20}$$

Therefore, computing $S_{\Upsilon_{ijk,ll'}}$ has time complexity $\mathcal{O}(N_T^i N_T^j)$. The computation of any of the single sums $Z_{K_{ki,l}}$, $Z_{\Upsilon_{iik,ll'}}$, or Z_{μ_k} has time complexity $\mathcal{O}(N_T^i)$. If the number of basis functions is the same for all kernels, that is, $r_{ki} = r$ for all event types $k, i \in [d]$, and the number of observed events is of the same order of magnitude across dimensions, then the exact brute force evaluation of the LSE has time complexity $\mathcal{O}(d^2 r^2 (N_T^k)^2)$, like for the log-likelihood. The quadratic multiplicative factor $d^2 r^2$ is the total number of basis functions in the model. Overcoming this curse of dimensionality is beyond the scope of this work and remains an open question in the literature of Hawkes processes that prohibits the estimation

[1]This bound is attained in the worst case where all the events of type j precede the first event of type i.

of Hawkes models for large networks without significant simplifying assumptions. Instead, we focus on reducing the quadratic dependence in the number of events $(N_T^k)^2$ by efficiently approximating the double and single sums.

4.4 LSE gradient estimator

Next, for parameters $\boldsymbol{\theta_k} \in \Theta_k$, we construct an unbiased estimator of the gradient $\nabla \mathcal{R}_T^{(k)}(\theta_k)$ of the partial LSE, which we use as an input to our optimization. To do this, we use variance reduction techniques to construct an efficient unbiased estimator using the additive structure of Equation (4.16). In particular, we use stratified MC sampling for the single sums from before; and for the double sums, we propose an adaptive stratified MC sampling scheme with a carefully chosen stratification space: that of lags between event indices. We use an adaptive scheme to allow the algorithm to learn to target lags with the highest impact.

4.4.1 Monte Carlo approximation problem

We first provide some notation to deal with the different types of sums involved in Theorem 4.3.1. For the rest of this section, fix event types $i, j \in [d]$, and mixture indices $l \in [r_{ki}]$ and $l' \in [r_{kj}]$. For some domain $\mathcal{E}$ consider a function $f_{\boldsymbol{\theta_k}} : \mathcal{E} \to [0, +\infty)$ parameterized by $\boldsymbol{\theta_k}$. Let $\mathcal{E}_S \subset \mathcal{E}$ be finite, with $N_S := |\mathcal{E}_S|$, and consider the generic problem of estimating

$$S_T(\theta_k) := \sum_{x \in \mathcal{E}_S} f_{\boldsymbol{\theta_k}}(x).$$

In our problem, we need to address the following configurations:

1. The single sums
   - Kernel correlation $Z_{\Upsilon_{iik,ll'}}$; corresponds to $\mathcal{E}_S = \left\{ \left(T - t_m^i, \xi_m^i\right) : m \in [\, N_T^i] \right\}$, and $f_{\boldsymbol{\theta_k}}(x) = \Upsilon_{iik,ll'}(T - t_m^i, 0)\mathcal{I}_{ki}^2(\xi_m^i)$ for all $x \in \mathcal{E}_S$.
   - Kernel-baseline correlation $Z_{K_{ki,l}}$; where $\mathcal{E}_S = \left\{ \left(T - t_m^i, t_m^i, \xi_m^i\right) : m \in [\, N_T^i] \right\}$, and $f_{\boldsymbol{\theta_k}}(x) = K_{ki,l}(T - t_m^i, t_m^i)\mathcal{I}_{ki}(\xi_m^i)$ for all $x \in \mathcal{E}_S$.
   - Baseline term Z_{μ_k}; corresponds to $\mathcal{E}_S = \left\{ t_m^k : m \in [\, N_T^k] \right\}$, and $f_{\boldsymbol{\theta_k}}(x) = \mu_k(t_m^k)$ for all $x \in \mathcal{E}_S$. In the MHP case, this sum simplifies to $\frac{1}{T} Z_{\mu_k} = \eta_T^k \mu_k$, so we compute it exactly.
2. The double sums
   - Kernel term $S_{\phi_{kj,l'}}$; corresponds to $\mathcal{E}_S = \mathcal{B}_T^j$, where

     $$\mathcal{B}_T^j := \left\{ \left(t_m^k - t_n^j, \xi_n^j\right) : m \in [\![\varpi(k,j), N_T^k]\!], n \in [\, \kappa(j,k,m)] \right\}, \tag{4.21}$$

     and $f_{\boldsymbol{\theta_k}}(x) = \phi_{kj,l'}\left(t_m^k - t_n^j\right)\mathcal{I}_{kj}\left(\xi_n^j\right)$ for all $x \in \mathcal{E}_S$.

- Kernel correlation $S_{\Upsilon_{ijk,ll'}}$; corresponds to $\mathcal{E}_S = \tilde{\mathcal{B}}_T^{ij}$ where

$$\tilde{\mathcal{B}}_T^{ij} := \Big\{ \big(T - t_m^i, t_m^i - t_n^j, \xi_m^i, \xi_n^j\big) : m \in [\![\varpi(i,j), N_T^i]\!], n \in [\,\kappa(j,i,m)] \Big\}, \tag{4.22}$$

and $f_{\boldsymbol{\theta_k}}(x) = \Upsilon_{ijk,ll'}\big(T - t_m^i, t_m^i - t_n^j\big)\mathcal{I}_{ki}\big(\xi_m^i\big)\mathcal{I}_{kj}\big(\xi_n^j\big)$ for all $x \in \mathcal{E}_S$.

The vanilla MC approach to this problem is to uniformly sample $N_{MC} < N_S$ elements of $\mathcal{E}_S$, denoted by $\mathcal{E}_{MC}$, and to consider the unbiased estimator

$$\hat{S}_T(\theta_k) := \frac{N_S}{N_{MC}} \sum_{x \in \mathcal{E}_{MC}} f_{\boldsymbol{\theta_k}}(x).$$

In practice, only mild variations of this approach are needed to achieve satisfactory MC estimation of the single sums, even in the case of a nearly critical Hawkes model (see Section 5.1). However, the estimation of the double sums is significantly more challenging and, in practice, vanilla MC is too imprecise for our problem because it does not capture the variations of $f_{\boldsymbol{\theta_k}}$ on the domain $\mathcal{E}_S$. For this reason, we develop a stratified sampling approach.

4.4.2 Estimating the single sums

First, note that in all the single-indexed sums described above, there exists an event type $i \in [d]$ such that $|\mathcal{E}_S| = N_T^i$. Therefore, for parameter values $\boldsymbol{\theta_k}$, we summarize the single-indexed sum estimation problems above as the estimation of

$$Z_T(\boldsymbol{\theta_k}) = \sum_{m=1}^{N_T^i} f_{\boldsymbol{\theta_k}}(\boldsymbol{x_m}),$$

for a real-valued function $f_{\boldsymbol{\theta_k}}$, and where $\mathcal{E}_S = (\boldsymbol{x_m})_{m \in [N_T^i]}$. Fix $n_{\max} \in \mathbb{N}$ and consider a fixed increasing sequence of integers $b_0 < b_1 < \cdots < b_{n_{\max}}$, with $b_0 := 1$ and $b_{n_{\max}} < N_T^i$. The integer intervals $[\![b_p, b_{p+1} - 1]\!]$ are the strata of a stratified MC estimator of

$$\sum_{m=1}^{b_{n_{\max}}-1} f_{\boldsymbol{\theta_k}}(\boldsymbol{x_m}).$$

Define an unbiased estimator of $Z_T(\theta_k)$ by

$$\hat{Z}_T(\theta_k) := \sum_{p=1}^{n_{\max}-1} \frac{b_{p+1} - b_p}{q_p} Z_p + \sum_{m=b_{n_{\max}}}^{N_T^i} f_{\boldsymbol{\theta_k}} \boldsymbol{x_m}), \tag{4.23}$$

where for each p, we sample uniformly q_p integers $(m_1^{(p)}, \ldots, m_{q_p}^{(p)})$ in the integer interval $[\![b_p, b_{p+1} - 1]\!]$ without replacement and define

$$Z_p := \sum_{l=1}^{q_p} f_{\boldsymbol{\theta_k}}(\boldsymbol{x}_{\boldsymbol{m_l^{(p)}}}). \tag{4.24}$$

We do not set $b_{n_{\max}} = N_T^i$ because, even for a stable Hawkes model, there are values of m such that $t_m^i \sim T$, and these values may contribute significantly to the sum. We summarize this procedure in Algorithm 1.

Algorithm 1: Estimation of a single sum

Result: Estimator $\hat{Z}_T(\boldsymbol{\theta_k})$
Initialize $\hat{Z}_T(\boldsymbol{\theta_k}) = \sum_{m=b_{n_{\max}}}^{N_T^i} f_{\boldsymbol{\theta_k}}(\boldsymbol{x_m})$;
for p *in* $[\, n_{max-1}]$ **do**
  Sample q_p integers $(m_1^{(p)}, \ldots, m_{q_p}^{(p)})$;
  Use (4.24) to compute Z_p ;
  Increment $\hat{Z}_T(\boldsymbol{\theta_k})$ by $\frac{b_{p+1}-b_p}{q_p} Z_p$;
end

We base the construction of this estimator on the following heuristics. First, note that the sequences $m \mapsto \psi_{ki,l}(T - t_m^i)$ and $m \mapsto \Upsilon_{iik,ll}(T - t_m^i, 0)$ are decreasing. Second, qualitatively, we expect $t_m^i \ll T$ for a stable process, except for the largest values of index m. Finally, we expect the variance of $f_{\boldsymbol{\theta_k}}(T - t_m^i)$ to increase with m. The hyper-parameters of this estimator are the bounds of the strata, b_p, and the number of points sampled in the strata, q_p. In our numerical experiments, we choose the index $b_{n_{\max}}$ such that $N_T^i - b_{n_{\max}} \sim 10^3$. One can imagine schemes where $b_{n_{\max}}$ is chosen adaptively, but our experiments suggested this was not useful in practice. Given $b_{n_{\max}}$, we choose $b_{n_{\max}-1} = b_{n_{\max}} - \delta$ where $\delta \in \mathbb{N}$ is a hyper-parameter, and choose the other bounds (b_p), such that $b_{p+1} - b_p \sim (b_{p+2} - b_{p+1})^2$ for the sums $Z_{K_{ki,l}}$ and $Z_{\Upsilon_{iik,ll'}}$. For the sum Z_{μ_k}, we choose a uniform grid, that is, $b_{p+1} - b_p$ is constant.

4.4.3 Estimating the double sums

For model parameters $\boldsymbol{\theta_k} \in \Theta_k$, we now consider the estimation of the double-indexed sums $S_T(\boldsymbol{\theta_k})$ above, corresponding to the terms

$$\sum_{m=\varpi(i,j)}^{N_T^i} \sum_{n=1}^{\kappa(j,i,m)} \Upsilon_{ijk}(T - t_m^i, t_m^i - t_n^j) \mathcal{I}_{ki}(\xi_m^i) \mathcal{I}_{kj}(\xi_n^j), \qquad \sum_{m=\varpi(k,j)}^{N_T^k} \sum_{n=1}^{\kappa(j,k,m)} \phi_{kj}(t_m^k - t_n^j) \mathcal{I}_{kj}(\xi_n^j),$$

which appear in Theorem 4.3.1. Section 4.4.3.1 outlines the structure of our estimator, and Section 4.4.3.2 gives a detailed description of our adaptive stratification.

4.4.3.1 Estimator structure

The key idea is to sample pairs of points, by stratifying over the index differences between events – that is, we group pairs of points by the number of events between them, and sample from these groups. For each training step, we update the number of points sampled from each stratum adaptively, in order to reduce variance. Approaches similar to those used to

estimate the single sums fail because of the larger number of terms in these sums, many of which make very small contributions, leading to a high variance in a simple estimation. Ideally, one could use stratified sampling of the data by time differences $t_m^i - t_n^j$ to reduce the variance of the estimator $\hat{S}(\boldsymbol{\theta_k})$. However, constructing strata based directly on the time differences would require pre-computation of the density of time differences, which is computationally expensive, both in time and memory. Instead, stratified sampling by index differences acts as a proxy for time differences, which is significantly faster and memory efficient; in particular we do not need to store any additional data beyond strata definition.

Sadly, working with this stratification, with multiple types of events, requires somewhat convoluted notation, which we now present. For event indices m, n such that $t_n^j < t_m^i$, the difference $\kappa(j, i, m) - n$ is the number of events of type j in the interval (t_n^j, t_m^i). We refer to the quantity $h = \kappa(j, i, m) - n + 1$ as the lag between indices m, n. For lags $h \in [\,\kappa(j, i, N_T^i)]$, define the sets of event times with a given lag in their indices

$$\mathcal{B}_T^{j,h} := \{t_m^k - t_n^j : m \in [\![\varpi(k, j), N_T^k]\!], n = \kappa(j, k, m) - h + 1\},$$
$$\tilde{\mathcal{B}}_T^{ij,h} := \{(T - t_m^i, t_m^i - t_n^j) : m \in [\![\varpi(i, j), N_T^i]\!], n = \kappa(j, i, m) - h + 1\}.$$

For each element of the set $\mathcal{B}_T^{j,h}$ (resp. $\tilde{\mathcal{B}}_T^{ij,h}$), the pair (m, n) is such that t_n^j is the h-th to last jump of type j before t_m^i (resp. t_m^k). Let $\mathcal{E}_T^{ij,h}$ denote $\mathcal{B}_T^{j,h}$ or $\tilde{\mathcal{B}}_T^{ij,h}$ as appropriate for the sum under consideration. Define

$$S_T^h(\boldsymbol{\theta_k}) := \sum_{x \in \mathcal{E}_T^{ij,h}} f_{\boldsymbol{\theta_k}}(x), \quad \text{hence} \quad S_T(\boldsymbol{\theta_k}) = \sum_{h=1}^{\kappa(j,i,N_T^i)} S_T^h(\boldsymbol{\theta_k}).$$

The sets $(\mathcal{E}_T^{ij,h})_h$ form a partition of $\mathcal{E}_T^{ij}$ of size $\kappa(j, i, N_T^i)$, which is typically still too large to use as a stratification.[2] Our heuristic for this estimator is to first note that

$$\lim_{t \to \infty} \phi_{ki}(t) = 0 \quad \text{and} \quad \lim_{s \to \infty} \Upsilon_{ijk}(t, s) = 0.$$

We expect the contribution of $S_T^h(\boldsymbol{\theta_k})$ to $S_T(\boldsymbol{\theta_k})$ to decrease after a certain lag $h_{\max}$, because the sequence $(\min \mathcal{E}_T^{ij,h})_h$ is strictly decreasing. Hence, we focus on the estimation of

$$S_T^{\max}(\boldsymbol{\theta_k}) := \sum_{h=1}^{h_{\max}} S_T^h(\boldsymbol{\theta_k})$$

separately from the estimation of the remainder

$$S_T^{\text{rest}}(\boldsymbol{\theta_k}) := \sum_{h=h_{\max}+1}^{\kappa(j,i,N_T^i)} S_T^h(\boldsymbol{\theta_k}).$$

[2] For example, one can think of the case $j = i$, where $\kappa(j, i, N_T^i) = N_T^i - 1$.

Our estimation approach is to use an adaptive stratified sampling for $S_T^{\max}(\boldsymbol{\theta_k})$, hence we refer to the set of lags $[h_{\max}]$ as the adaptive domain; and a classic stratified sampling approach for the remainder $S_T^{\text{rest}}(\boldsymbol{\theta_k})$, hence we refer to the set of lags $[\![h_{\max}+1, \kappa(j,i,N_T^i)]\!]$ as the non-adaptive domain. Therefore, we focus our presentation on the estimation of $S_T^{\max}(\boldsymbol{\theta_k})$. First, we group several index lag sets to reduce the number of strata in the adaptive domain $[h_{\max}]$. Formally, let $n_B \in \mathbb{N}^*$. Consider a partition $B = (\boldsymbol{b_1}, \dots, \boldsymbol{b_{n_B}})$ of $[h_{\max}]$. By abuse of notation, for all strata $\boldsymbol{b} \in B$ take the disjoint union and sum

$$\mathcal{E}_T^{ij,\boldsymbol{b}} := \bigcup_{h \in \boldsymbol{b}} \cdot \, \mathcal{E}_T^{ij,h}, \quad S_T^{\boldsymbol{b}}(\boldsymbol{\theta_k}) := \sum_{h \in \boldsymbol{b}} S_T^h(\boldsymbol{\theta_k}).$$

Figure 4.1 illustrates the construction of this stratification in the case $i = j$. Now that we

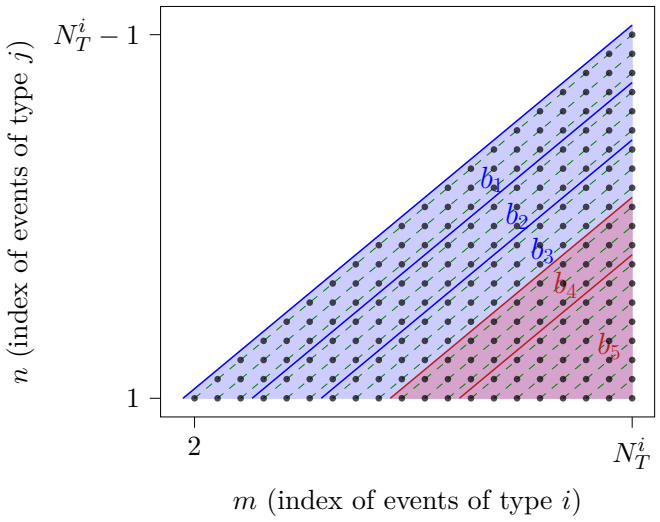


Figure 4.1: Illustration of our stratification.

Here $i = j$ and $N_T^i = 20$. Black points are the points (m, n) such that $1 \leq n < m \leq N_T^i$. Dashed green lines are the lines of equation $m - n = h$ for $h \in [N_T^i - 1]$. With $h_{\max} = 9$, the blue (respectively red) domain contains all the points (m, n) such that $m - n \leq h_{\max}$ (respectively $m - n > h_{\max}$). In this example, we consider 3 (respectively 2) strata in the blue (respectively red) domain, (b_1, b_2, b_3) (respectively (b_4, b_5)), delimited by solid blue (respectively red) lines. The MC estimator on the blue domain is constructed adaptively, whereas it is non-adaptive in the red domain.

have define our strata, we explicit our MC estimator, given a number of points to sample. For each stratum index $p \in [n_B]$, sample $q^{(p)}$ points $(x_1^{\boldsymbol{b_p}}, \dots, x_{q^{(p)}}^{\boldsymbol{b_p}})$ uniformly and without replacement from the stratum $\mathcal{E}_T^{ij,\boldsymbol{b_p}}$, and define an unbiased estimator of $S_T^{\boldsymbol{b_p}}(\boldsymbol{\theta_k})$

$$\hat{S}_T^{\boldsymbol{b_p},q^{(p)}}(\boldsymbol{\theta_k}) = \frac{|\mathcal{E}_T^{ij,\boldsymbol{b}}|}{q^{(p)}} \sum_{m=1}^{q^{(p)}} f_{\boldsymbol{\theta_k}}(x_m^{\boldsymbol{b_p}}). \tag{4.25}$$

Fix in advance the total number Q of points we want to sample across all strata $Q := \sum_{p=1}^{n_B} q^{(p)}$. Let $\boldsymbol{q} := (q^{(1)}, \dots, q^{(n_B)})$ denote the absolute allocation vector. For each stratum

index $p \in [\, n_B]$, define the relative allocation $\tilde{q}^{(p)} := q^{(p)}/Q$, and write $\tilde{\boldsymbol{q}} = (\tilde{q}^{(1)}, \dots, \tilde{q}^{(n_B)})$. An unbiased estimator of $S_T^{\max}(\boldsymbol{\theta_k})$ is

$$\hat{S}_T^{\boldsymbol{q}}(\boldsymbol{\theta_k}) = \sum_{p=1}^{n_B} \hat{S}_T^{\boldsymbol{b_p}, q^{(p)}}(\boldsymbol{\theta_k}). \tag{4.26}$$

In practice, one could fix $\boldsymbol{q}$ a priori, and choose $h_{\max}$ of order 30 to 50, for example. We see in Chapter 5 that this leads to satisfactory results in a variety of cases, particularly for monotonically decaying kernels. However, for general kernels, this approach is not sufficiently robust; we instead adaptively determine the allocation of sampled points per stratum in Section 4.4.3.2.

As in the single sum case, use a standard stratified MC approach with a fixed allocation to estimate the remainder $S_T^{\text{rest}}(\boldsymbol{\theta_k})$ defined above. Denote the estimator of the remainder by $\hat{S}_T^{\text{rest}}(\boldsymbol{\theta_k})$, and the estimator of the sum $S_T(\boldsymbol{\theta_k})$ by

$$\hat{S}_T(\boldsymbol{\theta_k}) = \hat{S}_T^{\boldsymbol{q}}(\boldsymbol{\theta_k}) + \hat{S}_T^{\text{rest}}(\boldsymbol{\theta_k}). \tag{4.27}$$

This procedure is summarized in Algorithm 2, and relies on the discussion of adaptive sampling in Section 4.4.3.2.

Algorithm 2: Estimation of a double sum

Result: Estimator $\hat{S}_T(\boldsymbol{\theta_k})$
Initialize $\tilde{\boldsymbol{q}}_*^{(\mathbf{0})}(\boldsymbol{\theta_k})$;
for s *in* $[\, n_K]$ **do**
  for p *in* $[\, n_B]$ **do**
    Set $\Delta q_s^{\boldsymbol{b_p}} = \tilde{q}_*^{\boldsymbol{b_p},(s-1)}(\boldsymbol{\theta_k})(n_B + \Delta Q_s)$;
    Sample without replacement $\Delta q_s^{\boldsymbol{b_p}}$ points $(x_m^{\boldsymbol{b_p}})_m$ in $\mathcal{E}_T^{ij,\boldsymbol{b_p}}$;
    Use (4.31) to compute $\hat{S}_T^{\boldsymbol{b_p},(\Delta s)}(\boldsymbol{\theta_k})$ and $\hat{\sigma}_T^{\boldsymbol{b_p},(\Delta s)}(\boldsymbol{\theta_k})^2$;
    Use (4.33) to compute $\hat{S}_T^{\boldsymbol{b_p},(s)}(\boldsymbol{\theta_k})$ and $\hat{\sigma}_T^{\boldsymbol{b_p},(s)}(\boldsymbol{\theta_k})^2$;
  end
end
Use (4.35) to compute $\hat{S}_T^{\text{rest}}(\boldsymbol{\theta_k})$;
Use (4.27) to compute $\hat{S}_T(\boldsymbol{\theta_k})$;

4.4.3.2 Adaptive stratification

Allocation criterion For a given number of sample points Q, we build an adaptive strategy to allocate our points between the n_B strata of the adaptive domain $B = (\boldsymbol{b_1}, \dots, \boldsymbol{b_{n_B}})$. Recall that $\boldsymbol{q}$ is the vector of numbers of points allocated per stratum, that we call the absolute allocation; and $\tilde{q} := (q^{(p)}/Q)_{p \in [n_B]}$ is the relative allocation. A naive allocation criterion is to minimize the variance of $\hat{S}_T^{\boldsymbol{q}}(\boldsymbol{\theta_k})$, the estimator of $S_T^{\max}(\boldsymbol{\theta_k})$. For each stratum

index $p \in [n_B]$, define the total standard deviation in a stratum $\boldsymbol{b_p}$ by

$$\sigma_T^{\boldsymbol{b_p}}(\boldsymbol{\theta_k}) := \sqrt{\frac{1}{|\mathcal{E}_T^{ij,\boldsymbol{b_p}}|}\sum_{x\in\mathcal{E}_T^{ij,\boldsymbol{b_p}}} f_{\boldsymbol{\theta_k}}^2(x) - \frac{1}{|\mathcal{E}_T^{ij,\boldsymbol{b_p}}|^2}\Big(\sum_{x\in\mathcal{E}_T^{ij,\boldsymbol{b_p}}} f_{\boldsymbol{\theta_k}}(x)\Big)^2}.$$

The variance of the estimator $\hat{S}_T^{\boldsymbol{q}}(\boldsymbol{\theta_k})$, accounting for sampling without replacement, is

$$\begin{aligned}\mathrm{Var}\Big[\,\hat{S}_T^{\boldsymbol{q}}(\boldsymbol{\theta_k})\Big] &= \sum_{p=1}^{n_B}\frac{|\mathcal{E}_T^{ij,\boldsymbol{b_p}}|^2}{q^p}\Big(1-\frac{q^p-1}{|\mathcal{E}_T^{ij,\boldsymbol{b_p}}|-1}\Big)\sigma_T^{\boldsymbol{b_p}}(\boldsymbol{\theta_k})^2,\\ &= \frac{1}{Q}\sum_{p=1}^{n_B}\frac{|\mathcal{E}_T^{ij,\boldsymbol{b_p}}|^3}{(|\mathcal{E}_T^{ij,\boldsymbol{b_p}}|-1)\tilde{q}^p}\sigma_T^{\boldsymbol{b_p}}(\boldsymbol{\theta_k})^2 - \sum_{p=1}^{n_B}\frac{|\mathcal{E}_T^{ij,\boldsymbol{b_p}}|^2}{|\mathcal{E}_T^{ij,\boldsymbol{b_p}}|-1}\sigma_T^{\boldsymbol{b}_p}(\boldsymbol{\theta_k})^2.\end{aligned} \tag{4.28}$$

Use Jensen's inequality to write

$$\mathrm{Var}\Big[\,\hat{S}_T^{\boldsymbol{q}}(\boldsymbol{\theta_k})\Big] \geq \frac{1}{Q}\Bigg(\sum_{p=1}^{n_B}|\mathcal{E}_T^{ij,\boldsymbol{b_p}}|\sqrt{\frac{|\mathcal{E}_T^{ij,\boldsymbol{b_p}}|}{|\mathcal{E}_T^{ij,\boldsymbol{b_p}}|-1}}\sigma_T^{\boldsymbol{b_p}}(\boldsymbol{\theta_k})\Bigg)^2 - \sum_{p=1}^{n_B}\frac{|\mathcal{E}_T^{ij,\boldsymbol{b_p}}|^2}{|\mathcal{E}_T^{ij,\boldsymbol{b_p}}|-1}\sigma_T^{\boldsymbol{b}_p}(\boldsymbol{\theta_k})^2. \tag{4.29}$$

For each stratum index $p \in [n_B]$, define

$$\tilde{q}_*^p(\boldsymbol{\theta_k}) := \frac{\sqrt{\frac{|\mathcal{E}_T^{ij,\boldsymbol{b_p}}|}{|\mathcal{E}_T^{ij,\boldsymbol{b_p}}|-1}}\sigma_T^{\boldsymbol{b_p}}(\boldsymbol{\theta_k})}{\sum_{p'=1}^{n_B}\sqrt{\frac{|\mathcal{E}_T^{ij,\boldsymbol{b_{p'}}}|}{|\mathcal{E}_T^{ij,\boldsymbol{b_{p'}}}|-1}}\sigma_T^{\boldsymbol{b_{p'}}}(\boldsymbol{\theta_k})}, \tag{4.30}$$

and let $\tilde{\boldsymbol{q}}_*(\boldsymbol{\theta_k}) = \Big(\tilde{q}_*^1(\boldsymbol{\theta_k}),\dots,\tilde{q}_*^{n_B}(\boldsymbol{\theta_k})\Big)$. The inequality in (4.29) is tight, and this lower bound is attained for $\tilde{\boldsymbol{q}} = \tilde{\boldsymbol{q}}_*(\boldsymbol{\theta_k})$. Therefore, given a total number of sample points Q, the relative allocation $\tilde{\boldsymbol{q}} = \tilde{\boldsymbol{q}}_*(\boldsymbol{\theta_k})$ minimizes the variance of the estimator $\hat{S}_T^{\boldsymbol{q}}(\boldsymbol{\theta_k})$, and we refer to it as the optimal allocation. The computation of the optimal allocation $\tilde{\boldsymbol{q}}_*(\boldsymbol{\theta_k})$ is expensive, because the computation of the vector of variances $\big(\sigma_T^{\boldsymbol{b_p}}(\boldsymbol{\theta_k})\big)_p$ has quadratic worst-case complexity. In Section 5.1, we discuss cases where it is not necessary to choose $\tilde{\boldsymbol{q}} = \tilde{\boldsymbol{q}}_*(\boldsymbol{\theta_k})$ for the estimation procedure to converge, notably for some decreasing kernels. But for general kernels, setting an arbitrary allocation $\tilde{\boldsymbol{q}}$ does not lead to satisfactory estimates in practice, this is why we propose an adaptive estimator of the optimal allocation $\tilde{\boldsymbol{q}}_*(\boldsymbol{\theta_k})$ below.

Efficient adaptive estimation of the optimal allocation We slightly modify the work of Etoré and Jourdain [34] on adaptive stratified MC sampling to the case of simple random sampling without replacement. Fix $\boldsymbol{\theta_k}$ and an initial allocation guess $\tilde{\boldsymbol{q}}_*^{(\mathbf{0})}(\boldsymbol{\theta_k})$. Let n_K denote the number of iterations used to estimate $\tilde{\boldsymbol{q}}_*(\boldsymbol{\theta_k})$, and Q the total number of points

that we sample in this procedure. Fix $(\Delta Q_s)_{s\in[\,n_K]}$ such that at each step $s \in [\,n_K]$ we sample $Q_s := n_B + \Delta Q_s$ points. Denote the points sampled in stratum $\boldsymbol{b}$ at step s by

$$\Delta q_s^{\boldsymbol{b}} := 1 + \delta q_s^{\boldsymbol{b}},$$

so that we sample at least one point in each stratum at each step k. Denote by $q_s^{\boldsymbol{b}} := \sum_{s'=1}^{s} \Delta q_{s'}^{\boldsymbol{b}}$ the total number of points sampled in stratum $\mathcal{E}_T^{ij,\boldsymbol{b}}$ up to and including step s. Begin the procedure with an initial guess of the optimal allocation $\tilde{\boldsymbol{q}}_*^{(\mathbf{0})}(\boldsymbol{\theta_k})$. For all $p \in [\,n_B]$, denote by $\hat{S}_T^{\boldsymbol{b_p},(s)}(\boldsymbol{\theta_k})$, $\hat{\sigma}_T^{\boldsymbol{b_p},(s)}(\boldsymbol{\theta_k})$, and $\tilde{\boldsymbol{q}}_*^{(\boldsymbol{s})}(\boldsymbol{\theta_k})$, our estimates constructed inductively, based on all the samples up to step s inclusive. At step s, for all p, compute

$$\Delta q_s^{\boldsymbol{b_p}} = \tilde{q}_*^{\boldsymbol{b_p},(s-1)}(\boldsymbol{\theta_k})(n_B + \Delta Q_s).$$

Next, for all p, sample without replacement $\Delta q_s^{\boldsymbol{b_p}}$ points $(x_m^{\boldsymbol{b_p}})_m$ in $\mathcal{E}_T^{ij,\boldsymbol{b_p}}$, and obtain

$$\begin{aligned}
\hat{S}_T^{\boldsymbol{b_p}}(\boldsymbol{\theta_k}) &= \hat{S}_T^{\boldsymbol{b_p},(n_B)}(\boldsymbol{\theta_k}),\\
\hat{q}_*^{\boldsymbol{b_p}}(\boldsymbol{\theta_k}) &= \hat{q}_*^{\boldsymbol{b_p},(n_B)}(\boldsymbol{\theta_k}),\\
\hat{\sigma}_T^{\boldsymbol{b_p}}(\boldsymbol{\theta_k})^2 &= \hat{\sigma}_T^{\boldsymbol{b_p},(\Delta s)}(\boldsymbol{\theta_k})^2.
\end{aligned}$$

In some cases, using a fixed, preset allocation $\tilde{\boldsymbol{q}}$ is sufficient to get satisfactory numerical results. In particular, for decaying kernels, we note that the sequence $\left(S_T^h(\boldsymbol{\theta_k})\right)_h$ is decreasing with the lag h. For a sufficiently fine stratification, a monotonically decaying allocation leads to good results. This is not the case for more general kernels, hence the importance of constructing an adaptive estimator $\hat{\boldsymbol{q}}_*(\boldsymbol{\theta_k})$ of the optimal allocation $\tilde{\boldsymbol{q}}_*(\boldsymbol{\theta_k})$ without increasing significantly the complexity of the procedure.

At each step s, compute the sample mean and sample variance of the corresponding batches

$$\hat{S}_T^{\boldsymbol{b_p},(\Delta s)}(\boldsymbol{\theta_k}) = \frac{|\mathcal{E}_T^{ij,\boldsymbol{b_p}}|}{\Delta q_s^{\boldsymbol{b_p}}} \sum_{m=1}^{\Delta q_s^{\boldsymbol{b_p}}} f_{\boldsymbol{\theta_k}}(x_m^{\boldsymbol{b_p}}), \tag{4.31}$$

$$\hat{\sigma}_T^{\boldsymbol{b_p},(\Delta s)}(\boldsymbol{\theta_k})^2 = \frac{|\mathcal{E}_T^{ij,\boldsymbol{b_p}}| - 1}{|\mathcal{E}_T^{ij,\boldsymbol{b_p}}|} \frac{1}{\Delta q_s^{\boldsymbol{b_p}} - 1} \sum_{m=1}^{\Delta q_s^{\boldsymbol{b_p}}} \left(f_{\boldsymbol{\theta_k}}(x_m^{\boldsymbol{b_p}}) - \hat{S}_T^{\boldsymbol{b_p},(\Delta s)}(\boldsymbol{\theta_k}) \right)^2. \tag{4.32}$$

To avoid unnecessary computations, update the running estimates in batches with

$$\hat{S}_T^{\boldsymbol{b_p},(s)}(\boldsymbol{\theta_k}) = \frac{q_{s-1}^{\boldsymbol{b_p}}}{q_s^{\boldsymbol{b_p}}} \hat{S}_T^{\boldsymbol{b_p},(s-1)}(\boldsymbol{\theta_k}) + \frac{\Delta q_s^{\boldsymbol{b_p}}}{q_s^{\boldsymbol{b_p}}} \hat{S}_T^{\boldsymbol{b_p},(\Delta s)}(\boldsymbol{\theta_k}), \tag{4.33}$$

$$\begin{aligned}
\hat{\sigma}_T^{\boldsymbol{b_p},(s)}(\boldsymbol{\theta_k})^2 = {} & \frac{q_{s-1}^{\boldsymbol{b_p}} - 1}{q_s^{\boldsymbol{b_p}} - 1} \hat{\sigma}_T^{\boldsymbol{b_p},(s-1)}(\boldsymbol{\theta_k})^2 + \frac{\Delta q_s^{\boldsymbol{b_p}}}{q_s^{\boldsymbol{b_p}} - 1} \hat{\sigma}_T^{\boldsymbol{b_p},(\Delta s)}(\boldsymbol{\theta_k})^2 \\
& + \frac{|\mathcal{E}_T^{ij,\boldsymbol{b_p}}| - 1}{|\mathcal{E}_T^{ij,\boldsymbol{b_p}}|} \frac{q_{s-1}^{\boldsymbol{b_p}} \Delta q_s^{\boldsymbol{b_p}}}{(q_s^{\boldsymbol{b_p}} - 1) q_s^{\boldsymbol{b_p}}} \left(\hat{S}_T^{\boldsymbol{b_p},(s-1)}(\boldsymbol{\theta_k}) - \hat{S}_T^{\boldsymbol{b_p},(\Delta s)}(\boldsymbol{\theta_k}) \right)^2.
\end{aligned} \tag{4.34}$$

Sensitivity of the optimal allocation to parameters Note that for an SBF Hawkes model such that each kernel ϕ_{ij} is represented by only one basis function; *i.e.* only one corresponding parameter $\omega_{ij} > 0$, we have $\nabla \tilde{\boldsymbol{q}}_*(\boldsymbol{\theta_k}) = 0$. In the general case though, the optimal allocation $\tilde{\boldsymbol{q}}_*(\boldsymbol{\theta_k})$ depends on model parameters $\boldsymbol{\theta_k}$, which constitutes an additional estimation challenge. Consider a gradient-based optimization procedure; after $t \in \mathbb{N}^*$ gradient iterations of this procedure we have computed a sequence of parameters

$$\left(\boldsymbol{\theta_k^{(1)}}, \dots, \boldsymbol{\theta_k^{(t)}}, \boldsymbol{\theta_k^{(t+1)}}\right),$$

and estimates for the corresponding optimal allocations

$$\tilde{\boldsymbol{q}}_*\left(\boldsymbol{\theta_k^{(1)}}\right), \dots, \tilde{\boldsymbol{q}}_*\left(\boldsymbol{\theta_k^{(t)}}\right).$$

As discussed above, we need an initial guess to estimate the partial LSE gradient evaluated in $\boldsymbol{\theta_k^{(t+1)}}$. To this effect, we use an exponential moving average to get a heuristic for $\hat{\boldsymbol{q}}_*\left(\boldsymbol{\theta_k^{(t+1)}}\right)$, the optimal allocation at $t+1$; using the previous optimal allocation estimates. Formally, denote by $w > 0$ the EMA weight, fixed at the beginning of the procedure, and write

$$\tilde{\boldsymbol{q}}_*^{(0)}\left(\boldsymbol{\theta_k^{(t+1)}}\right) = w \cdot \tilde{\boldsymbol{q}}_*\left(\boldsymbol{\theta_k^{(t)}}\right) + (1-w) \cdot \tilde{\boldsymbol{q}}_*\left(\boldsymbol{\theta_k^{(t-1)}}\right).$$

Estimation of the remainder $S_T^{\text{rest}}(\boldsymbol{\theta_k})$ As in the single sum case, use a standard stratified MC approach with a fixed allocation to estimate the remainder

$$S_T^{\text{rest}}(\boldsymbol{\theta_k}) := \sum_{h=h_{\max}+1}^{\kappa(j,i,N_T^i)} S_T^h(\boldsymbol{\theta_k}).$$

Let $n_R \in \mathbb{N}^*$ denote the number of strata we use for this estimation, and consider a partition $B' = (\boldsymbol{b'_1}, \dots, \boldsymbol{b'_{n_R}})$ of $[\![h_{\max}+1, \kappa(j,i,N_T^i)]\!]$. Sample q_p points $(x_1^{\boldsymbol{b'_p}}, \dots, x_{q_p}^{\boldsymbol{b'_p}})$ uniformly and without replacement from $\mathcal{E}_T^{ij,\boldsymbol{b'_p}}$ for all $p \in [\![n_R]\!]$. Fix in advance the number of sample points per stratum, which we denote by $\boldsymbol{q} = (q_1, \dots, q_p)$. Next, use the unbiased estimator of $S_T^{\text{rest}}(\boldsymbol{\theta_k})$ defined by

$$\hat{S}_T^{rest}(\boldsymbol{\theta_k}) = \sum_{p=1}^{n_R} \frac{|\boldsymbol{b'_p}|}{q_p} \sum_{n=1}^{q_p} f_{\boldsymbol{\theta_k}}(x_{q_p}^{\boldsymbol{b'_p}}). \tag{4.35}$$

4.4.4 Summary of the gradient estimator

We use the sum estimators constructed above to define the gradient estimate $\boldsymbol{\mathcal{G}_T^{(k)}}(\theta_k)$. For all $i \in [d]$, denote the sum estimates by

- $\hat{S}_{\psi,ki,T}(\theta_k)$ the estimator of $\sum_{m=1}^{N_T^i} \psi_{ki}(T - t_m^i)$,
- $\hat{S}_{\Upsilon,ik,T}(\theta_k)$ the estimator of $\sum_{m=1}^{N_T^i} \Upsilon_{iik}(T - t_m^i)$,

- $\hat{S}_{\phi,kj,T}(\theta_k)$ the estimator of $\sum_{m=\varpi(k,j)}^{N_T^k} \sum_{n=1}^{\kappa(j,k,m)} \phi_{kj}(t_m^k - t_n^j)$,
- $\hat{S}_{\Upsilon,ijk,T}(\theta_k)$ the estimator of $\sum_{m=\varpi(i,j)}^{N_T^i} \sum_{n=1}^{\kappa(j,i,m)} \Upsilon_{ijk}(T - t_m^i, t_m^i - t_n^j)$.

We denote by $\boldsymbol{\mathcal{G}_T^{(k)}}|_{\mu_k}$ the μ_k component of $\boldsymbol{\mathcal{G}_T^{(k)}}$ (the estimator of the partial derivative of the partial LSE $\mathcal{R}_T^{(k)}$ with respect to μ_k), defined by

$$\boldsymbol{\mathcal{G}_T^{(k)}}|_{\mu_k}(\theta_k) = 2\bigg(\mu_k - \eta_T^k + \frac{1}{T}\sum_{i=1}^{d} \hat{S}_{\psi,ki,T}(\theta_k)\bigg). \tag{4.36}$$

For $p \in [d]$ and $l \in [\,\rho_{ij}\,]$, let ϑ_{kpl} be the l-th parameter of ϕ_{kp}. Denote by $\boldsymbol{\mathcal{G}_T^{(k)}}|_{\vartheta_{kpl}}$ the ϑ_{kpl} component of $\boldsymbol{\mathcal{G}_T^{(k)}}$ (the estimator of the partial derivative of the partial LSE $\mathcal{R}_T^{(k)}$ with respect to ϑ_{kpl}) defined by

$$\begin{aligned}\boldsymbol{\mathcal{G}_T^{(k)}}|_{\vartheta_{kpl}}(\theta_k) = \frac{2}{T}\sum_{i=1,i\neq p}^{d} \frac{\partial \hat{S}_{\Upsilon,ipk,T}}{\partial \vartheta_{kpl}} + \frac{\partial \hat{S}_{\Upsilon,pik,T}}{\partial \vartheta_{kpl}} + \frac{2}{T}\frac{\partial \hat{S}_{\Upsilon,ppk,T}}{\partial \vartheta_{kpl}} - \frac{2}{T}\frac{\partial \hat{S}_{\phi,kp,T}}{\partial \vartheta_{kpl}} \\ + \frac{2\mu_k}{T}\frac{\partial \hat{S}_{\psi,kp,T}}{\partial \vartheta_{kpl}} + \frac{1}{T}\frac{\partial \hat{S}_{\Upsilon,pk,T}}{\partial \vartheta_{kpl}}.\end{aligned} \tag{4.37}$$

Our method can easily be extended to include regularization terms, for example to encourage sparsity. To give a concise expression for the complexity in time of the computation of the estimate $\boldsymbol{\mathcal{G}_T^{(k)}}(\theta_k)$, suppose that

- the number of parameters per kernel is constant; *i.e.* there exists $\rho \in \mathbb{N}$ such that $\rho_{ij} = \rho$ for all $i, j \in [d]$,
- the total sample size for the estimation of each single sum is constant, denoted $Q^{(1)}$,
- the total sample size for the estimation of each double sum is constant, denoted $Q^{(2)}$.

Under these assumptions, the complexity in time of the computation of a gradient estimate is roughly of order $\mathcal{O}(\rho d^2 Q^{(2)} + \rho d Q^{(1)})$.

4.4.5 Approximating model functionals

Computing the gradient estimator requires to evaluate the functionals Υ and K. For some specific parametric classes of kernels and baselines, there exists simple closed-form expressions for these functionals. However, for several other parametric classes, the integrals defining these functions are either not solvable analytically, or involve special functions which evaluation can be computationally expensive. For instance, if two kernels are power laws (see Section 4.6), the integral defining their cross-correlation Υ is

$$\Upsilon(t,s) = \int_0^t \frac{\alpha_p \alpha_q \beta_p \beta_q}{(1+\beta_p u)^{1+\alpha_p}(1+\beta_q(u+s))^{1+\alpha_q}} \mathrm{d}u. \tag{4.38}$$

This integral does not admit an analytical solution for general parameter values, except for integer decay rates α_p and α_q. If two kernels are Rayleigh, their cross correlation Υ is

$$\tilde{\Upsilon}(t,s) = \frac{2\pi}{\tilde{\beta}} f_{\mathcal{N}}(\tilde{s}) \left[b_1^2 \tilde{s} f_{\mathcal{N}}(a_1 \tilde{s}) - b_2^2 \tilde{x}_2 f_{\mathcal{N}}(a_1 \tilde{x}_1) + b_1 b_2 (1 - \tilde{s}^2) \Big(F_{\mathcal{N}}(a_1 \tilde{x}_1) - F_{\mathcal{N}}(a_1 \tilde{s}) \Big) \right], \tag{4.39}$$

where $(a_i)_{i\in[2]}$, $(b_i)_{i\in[2]}$, and $\tilde{\beta}$ depend on $\beta, \beta_\star$, and $(\tilde{x}_i)_{i\in[2]}$ and $\tilde{s}$ also depend on the lags t, s. The derivatives of this expression with respect to β and $\beta_\star$ are particularly lengthy and require evaluations of the normal PDF $f_{\mathcal{N}}$ and the normal CDF $F_{\mathcal{N}}$ which in general cannot be cached. Finally, if a kernels is in the Gamma family and a baselines is constant, the integral defining their correlation K is

$$K(t,s) = \mu \int_0^t \beta_l^{\alpha_l} \frac{u^{\alpha_l - 1} \mathrm{e}^{-\beta_l u}}{\Gamma(\alpha_l)} \mathrm{d}u, \tag{4.40}$$

and it does not admit an analytical solution. In this subsection, we propose an alternative solution to approximate the functionals Υ and K, in order to get rid of this requirement. This solution consists in estimating these functionals using a one-shot MC estimator.

CDF Fix event types $i, j \in [\,d]$, a mixture index $l \in [\,r_{ki}]$, and a kernel density parameter index $p \in [\,\tilde{r}_{ij}]$.

Proposition 4.4.1 (MC approximation of ψ)**.** *For lags $t \geq 0$, the primitive $\psi_{ij,l}$ verifies*

$$\psi_{ij,l}(t) = \omega_{ijl} \mathrm{E}_{\tau \sim \tilde{\phi}_{ij,l}} \left[\mathbb{1}_{[0,t]}(\tau) \right] \tag{4.41}$$

and the partial derivatives of $\psi_{ij,l}$ with respect to model parameters verify

$$\begin{aligned} \frac{\partial \psi_{ij,l}}{\partial \omega_{ijl}}(t) &= \mathrm{E}_{\tau \sim \tilde{\phi}_{ij,l}} \left[\mathbb{1}_{[0,t]}(\tau) \right] , \\ \frac{\partial \psi_{ij,l}}{\partial \tilde{\theta}_{ijlp}}(t) &= \omega_{ijl} \mathrm{E}_{\tau \sim \tilde{\phi}_{ij,l}} \left[\mathbb{1}_{[0,t]}(\tau) \frac{\partial \log(\tilde{\phi}_{ij,l})}{\partial \theta_{ijlp}}(\tau) \right] . \end{aligned} \tag{4.42}$$

We give a proof of this result in Appendix A.3.2.

Kernel cross-correlation Fix event types $i, j, k \in [\,d]$, mixture indices $l \in [\,r_{ki}]$ and $l' \in [\,r_{kj}]$, and kernel density parameter indices $p \in [\,\tilde{r}_{ki}]$ and $q \in [\,\tilde{r}_{kj}]$.

Proposition 4.4.2 (MC approximation of Υ)**.** *For lags $t, s \geq 0$, the kernel correlation $\Upsilon_{ijk,ll'}$ verifies*

$$\Upsilon_{ijk,ll'}(t,s) = \omega_{kil} \omega_{kjl'} \mathrm{E}_{\tau \sim \tilde{\phi}_{ki,l}} \left[\mathbb{1}_{[0,t]}(\tau) \tilde{\phi}_{kj,l'}(\tau + s) \right] . \tag{4.43}$$

If $(i,l) = (j,l')$, the partial derivatives of $\Upsilon_{iik,ll}$ with respect to model parameters verify

$$\begin{aligned}
\frac{\partial \Upsilon_{iik,ll}}{\partial \omega_{kil}}(t,s) &= 2\omega_{kil}\mathrm{E}_{\tau\sim\tilde{\phi}_{ki,l}}\left[\mathbb{1}_{[0,t]}(\tau)\tilde{\phi}_{ki,l}(\tau+s)\right], \\
\frac{\partial \Upsilon_{iik,ll}}{\partial \tilde{\theta}_{kilp}}(t,s) &= \omega_{kil}^2\mathrm{E}_{\tau\sim\tilde{\phi}_{ki,l}}\left[\mathbb{1}_{[0,t]}(\tau)\left(\frac{\partial \log\tilde{\phi}_{ki,l}}{\partial\tilde{\theta}_{kilp}}(\tau)\tilde{\phi}_{ki,l}(\tau+s) + \frac{\partial\tilde{\phi}_{ki,l}}{\partial\tilde{\theta}_{kilp}}(\tau+s)\right)\right].
\end{aligned} \tag{4.44}$$

If $(i,l) \neq (j,l')$, the partial derivatives of the $\Upsilon_{ijk,ll'}$ with respect to model parameters verify

$$\begin{aligned}
\frac{\partial \Upsilon_{ijk,ll'}}{\partial \omega_{kil}}(t,s) &= \omega_{kjl'}\mathrm{E}_{\tau\sim\tilde{\phi}_{kil}}\left[\mathbb{1}_{[0,t]}(\tau)\tilde{\phi}_{kjl'}(\tau+s)\right], \\
\frac{\partial \Upsilon_{ijk,ll'}}{\partial \tilde{\theta}_{kilp}}(t,s) &= \omega_{kil}\omega_{kjl'}\mathrm{E}_{\tau\sim\tilde{\phi}_{kil}}\left[\mathbb{1}_{[0,t]}(\tau)\frac{\partial \log\tilde{\phi}_{kil}}{\partial\tilde{\theta}_{kilp}}(\tau)\tilde{\phi}_{kjl'}(\tau+s)\right], \\
\frac{\partial \Upsilon_{ijk,ll'}}{\partial \omega_{kjl'}}(t,s) &= \omega_{kil}\mathrm{E}_{\tau\sim\tilde{\phi}_{kil}}\left[\mathbb{1}_{[0,t]}(\tau)\tilde{\phi}_{kjl'}(\tau+s)\right], \\
\frac{\partial \Upsilon_{ijk,ll'}}{\partial \tilde{\theta}_{kjl'q}}(t,s) &= \omega_{kil}\omega_{kjl'}\mathrm{E}_{\tau\sim\tilde{\phi}_{kil}}\left[\mathbb{1}_{[0,t]}(\tau)\frac{\partial\tilde{\phi}_{kjl'}}{\partial\tilde{\theta}_{kjl'q}}(\tau+s)\right].
\end{aligned} \tag{4.45}$$

We give a proof of this result in Appendix A.3.2.

Kernel–baseline cross-correlation As discussed previously, the separation of time variation scales between kernel densities and baseline functions is an important modelling constraint. For a fixed kernel ϕ, different classes of baselines μ lead to very different kernel–baseline cross-correlations K. There are some classes of baselines for which we can get closed-form expressions of K without further assumptions on ϕ. For instance, if μ is piecewise constant (see notation in Section 2.4.2), then for all lags $t, s \geq 0$, the baseline-kernel correlation K is

$$\begin{aligned}
K(t,s) = {} & \mathbb{1}_{\{g(s)=g(t+s)\}} b_{g(s)}\psi(t) \\
& + \mathbb{1}_{\{g(s)<g(t+s)\}}\Big(b_{g(s)}\psi\big(\beta_{g(s)+1}-s\big) + b_{g(t+s)}\Big(\psi(t) - \psi\big(\beta_{g(t+s)}\big)\Big) \\
& \qquad + \sum_{j=g(s)+1}^{g(t+s)-1} b_j(\psi(\beta_{j+1}-s) - \psi(\beta_j - s))\Big),
\end{aligned} \tag{4.46}$$

which depends on the primitive ψ. If μ is linear (see notation in Section 2.4.4), then

$$K(t,s) = \mu(t+s)\psi(t) - a\int_0^t \psi(u)du. \tag{4.47}$$

In this case, K depends on the primitive of ψ, for which there is not necessarily a closed form expression. Therefore, to simplify the use of the `ASLSD` procedure with general baselines and kernels without further assumptions, we propose an MC approximation of K and its derivatives. Fix event types $i, k \in [\,d]$, a mixture index $l \in [r_{ki}]$, a kernel density parameter index $p \in [\,\tilde{r}_{ki}]$, and a baseline parameter index $q \in [\,r_k^{(\mu)}]$.

Proposition 4.4.3 (MC approximation of the baseline-kernel correlation K)**.** *For lags $t, s \geq 0$, the baseline-kernel correlation K verifies*

$$K_{ki,l}(t,s) = \omega_{kil} \mathrm{E}_{\tau \sim \tilde{\phi}_{ki,l}} \left[\mathbb{1}_{[0,t]}(\tau) \mu_k(\tau + s) \right] , \tag{4.48}$$

the partial derivatives of $K_{ki,l}$ with respect to kernel parameters verify

$$\begin{aligned} \frac{\partial K_{ki,l}}{\partial \omega_{kil}}(t,s) &= \mathrm{E}_{\tau \sim \tilde{\phi}_{kil}} \left[\mathbb{1}_{[0,t]}(\tau) \mu_k(\tau + s) \right] , \\ \frac{\partial K_{ki,l}}{\partial \tilde{\theta}_{kilp}}(t,s) &= \omega_{kil} \mathrm{E}_{\tau \sim \tilde{\phi}_{kil}} \left[\mathbb{1}_{[0,t]}(\tau) \frac{\partial \log(\tilde{\phi}_{kil})}{\partial \tilde{\theta}_{kilp}}(\tau) \mu_k(\tau + s) \right] , \end{aligned} \tag{4.49}$$

and the partial derivative of $K_{ki,l}$ with respect to baseline parameters verify

$$\frac{\partial K_{ki,l}}{\partial b_{kq}}(t,s) = \omega_{kil} \mathrm{E}_{\tau \sim \tilde{\phi}_{kil}} \left[\mathbb{1}_{[0,t]}(\tau) \frac{\partial \mu_k}{\partial b_{kq}}(\tau + s) \right] . \tag{4.50}$$

We give a proof of this result in Appendix A.3.2.

Interpretation This property allows to re-interpret the LSE

Lemma 4.4.1. *For parameters $\boldsymbol{\theta_k} \in \Theta_k$, the partial LSE $\mathcal{R}_T^{(k)}$ verifies*

$$\begin{aligned} \mathcal{R}_T^{(k)}(\boldsymbol{\theta_k}) = \frac{2}{T} \sum_{i=1}^{d} \sum_{m=1}^{N_T^i} \Big(\mathcal{I}_{ki}(\xi_m^i) \omega_{ki} \mathbb{E}_{\tau \sim \tilde{\phi}_{ki}} [\mathbb{1}_{\tau \in [0, T - t_m^i]} \lambda_k(t_m^i + \tau \mid \mathcal{F}_{t_m^i})] - \lambda_k(t_m^i) \Big) \\ + \frac{1}{T} \sum_{i=1}^{d} \sum_{m=1}^{N_T^i} \mathcal{I}_{ki}^2(\xi_m^i) \omega_{ki}^2 \mathbb{E}_{\tau \sim \tilde{\phi}_{ki}} \Big[\mathbb{1}_{\tau \in [0, T - t_m^i]} \tilde{\phi}_{ki}(\tau) \Big] \\ + \mathbb{E}_{N^{(\mu)} \sim \mu_k} \Big[\frac{1}{T} \int_0^T \mu_k(t) \mathrm{d} N_t^{(\mu)} \Big] . \end{aligned} \tag{4.51}$$

We give a proof of this result in Appendix A.3.2. Taking a closer look at this expression, fix an event type $i \in [d]$ and an event index $m \in [N_T^i]$. Recall that the direct descendants of type k of the event at t_m^i, $\mathcal{D}_{T,i,m,1}^{(k)}$, are i.i.d.and there random offsets τ from t_m^i follow the distribution $\tilde{\phi}_{ki}$. The number of direct descendants of type k follows a Poisson distribution with parameter $\mathcal{I}_{ki}(\xi_m^i)\omega_{ki}$. Therefore, we get

$$\mathbb{E} \left[\sum_{n \in \mathcal{D}_{T,i,m,1}^{(k)}} \lambda_k(t_n^k \mid \mathcal{F}_{t_m^i}) \big| \mathcal{F}_{t_m^i} \right] = \mathcal{I}_{ki}(\xi_m^i) \omega_{ki} \mathbb{E}_{\tau \sim \tilde{\phi}_{ki}} [\mathbb{1}_{\tau \in [0, T - t_m^i]} \lambda_k(t_m^i + \tau \mid \mathcal{F}_{t_m^i})] . \tag{4.52}$$

Similarly, we re-write the term

$$\mathbb{E} \left[\sum_{n \in \mathcal{D}_{T,i,m,1}^{(k)}} \mathcal{I}_{ki}(\xi_m^i) \omega_{ki} \tilde{\phi}_{ki}(t_n^k - t_m^i) \big| \mathcal{F}_{t_m^i} \right] = \mathcal{I}_{ki}^2(\xi_m^i) \omega_{ki}^2 \mathbb{E}_{\tau \sim \tilde{\phi}_{ki}} \Big[\mathbb{1}_{\tau \in [0, T - t_m^i]} \tilde{\phi}_{ki}(\tau) \Big] . \tag{4.53}$$

Finally, this implies the following re-formulation of the LSE.

Corollary 4.4.1 (Re-formulating the LSE)**.**

$$\begin{aligned}\mathcal{R}_T^{(k)}(\boldsymbol{\theta_k}) = 2\sum_{i=1}^{d}\frac{1}{T}\int_0^T \left(\mathbb{E}\left[\sum_{n\in\mathcal{D}^{(k)}_{T,i,N_t^i,1}} \lambda_k(t_n^k \mid \mathcal{F}_t)\big|\mathcal{F}_t\right] - \lambda_k(t)\right)\mathrm{d}N_t^i \\ + \mathbb{E}_{N^{(\mu)}\sim\mu_k}\left[\frac{1}{T}\int_0^T \mu_k(t)\mathrm{d}N_t^{(\mu)}\right] \\ + \sum_{i=1}^{d}\frac{1}{T}\int_0^T \mathbb{E}\left[\sum_{n\in\mathcal{D}^{(k)}_{T,i,N_t^i,1}} \mathcal{I}_{ki}(\xi^i_{N_t^i})\omega_{ki}\tilde{\phi}_{ki}(t_n^k - t)\big|\mathcal{F}_t\right]\mathrm{d}N_t^i.\end{aligned} \tag{4.54}$$

The first term in this expression is the counting average of the variation of the conditional intensity when simulating from the model

$$2\sum_{i=1}^{d}\frac{1}{T}\int_0^T \left(\mathbb{E}\left[\sum_{n\in\mathcal{D}^{(k)}_{T,i,N_t^i,1}} \lambda_k(t_n^k \mid \mathcal{F}_t)\big|\mathcal{F}_t\right] - \lambda_k(t)\right)\mathrm{d}N_t^i. \tag{4.55}$$

The other 2 terms are L_2 regularization terms.

4.5 The procedure

Initialize the parameter values at $\boldsymbol{\theta_k^{(0)}} \in \Theta_k$. This starting value can be deterministic or sampled using some heuristics; we discuss this in more detail in Section 4.5.1. We then construct iteratively a sequence of parameter estimates $\left(\boldsymbol{\theta_k^{(t)}}\right)_{t\geq 1}$ with updates of the general form

$$\boldsymbol{\theta_k^{(t+1)}} := \mathrm{proj}_{\Theta_k}\left(\boldsymbol{\theta_k^{(t)}} + \boldsymbol{\Delta\theta_k^{(t)}}\right). \tag{4.56}$$

The function proj_{Θ_k} is the projection operator on the space of parameters Θ_k. It ensures that parameter estimates remain in the set of admissible parameter values. In Section 4.5.2, we construct the update vector $\boldsymbol{\Delta\theta_k^{(t)}}$ using numerical schemes with demonstrated efficiency in highly non-convex problems with large amounts of data, notably in the deep learning literature. In most of these schemes, this update vector at a step t depends on the current parameter values $\theta_k^{(t)}$ and the current gradient estimate, but also on functions of the sequence of preceding parameters and gradient estimates up to step t. To reduce furthermore the variance of the gradient estimates, we introduce a gradient clipping method. The algorithm terminates after a fixed number of iterations n_{iter}, or using a convergence criterion.

4.5.1 Initialization

Iterative gradient methods require an initial guess of parameter values, and aim to initialize parameter values as close as possible to the optimum to speed up convergence of the procedure. Initialization may be deterministic, if the modeller can use domain knowledge or fit

results from similar problems, but in general, we need to initialize parameters at random. In this subsection, we discuss our random initialization heuristics, based on necessary conditions on the LSE minimizer under some assumptions on the data generating process, and the statistical challenges that arise to use these heuristics.

4.5.1.1 Initializing MHP models

First, we discuss random initialization of MHP models. If the training data is generated by a stable $(\boldsymbol{\mu}^\diamond, \Phi^\diamond)$-MHP. The stationarity characterization for MHP implies $\rho(||\Phi^\diamond||_1) < 1$, and the LLN implies that there exists $\boldsymbol{\eta_\star} \in (0, +\infty)^d$ such that

$$\boldsymbol{\eta_\star} = \lim_{t \to +\infty} \boldsymbol{\eta_t} \quad a.s., \qquad \boldsymbol{\eta}_\star = (\mathbb{I}_d - ||\Phi^\diamond||_1)^{-1} \boldsymbol{\mu}^\diamond. \tag{4.57}$$

Of course, even under the hypothesis that the ground truth process is an MHP, the stationary regime intensity $\boldsymbol{\eta_\star}$ is not directly observed, as discussed in Section 3.1.1.2. We use an estimator $\hat{\boldsymbol{\eta}}_\star$ instead. Therefore, we expect a satisfactory adjacency matrix model $||\Phi||_1$ to verify the stability condition

$$\rho(||\Phi||_1) < 1, \tag{4.58}$$

and the LLN condition

$$(\mathbb{I}_d - ||\Phi||_1)\hat{\boldsymbol{\eta}}_\star > 0. \tag{4.59}$$

We then choose the baseline $\boldsymbol{\mu}$ accordingly to the LLN

$$\boldsymbol{\mu} = (\mathbb{I}_d - ||\Phi||_1)\hat{\boldsymbol{\eta}}_\star. \tag{4.60}$$

There exists different approaches to generate stable matrices with positive entries: for example, using Girko's circular law. However, these methods focus on generating eigenvectors of the adjacency matrix. The modeller usually does not have a prior on the eigenvectors of the adjacency matrix of the MHP model, but on the coefficients of this adjacency matrix. For instance, depending on the system being modelled, the modeller may expect a symmetric adjacency matrix. The modeller may also have a prior on the relative scales of different L_1 weights, and potentially the nullity of some coefficients: the modeller might believe that the self-excitation in the system is negligible compared to the cross-excitation. This can be the case in systems which exhibit mean-reverting properties, like in mid-price modelling in finance, or in large systems which exhibit clustered behaviours and where the adjacency matrix is believed to be sparse. However, there is no simple analytic characterization of the stability condition and the LLN condition. Therefore, we rely on our analysis above from Section 3.1.2 and Section 3.1.3.

Uni-dimensional case Consider bounds $a, b \in [0, 1]$ with $a < b$, such that we want the initial guess of the L_1 weight ω_{11} to satisfy $\omega_{11} \in (a, b)$. The bounds a, b are heuristics that may be suggested by domain knowledge, or fit results from similar problems. As discussed in Section 3.1.2, the LLN condition is automatically satisfied for stable uni-dimensional MHP models. Therefore, the random initialization of feasible first order parameters $(\boldsymbol{\mu}, ||\Phi||_1)$ is particularly simple in this case.

Algorithm 3: Random initialization of first order parameters, $d = 1$

Data: An estimator $\hat{\eta}^1_\star > 0$, prior bounds (a, b) on ω_{11}.
Result: First order feasible initial guess $(\mu_1, \omega_{11}) \in \mathbb{F}_{\text{MHP}}(\hat{\eta}^1_\star)$.
Sample ω_{11} following a uniform distribution on $[a, b]$;
$\mu_1 \leftarrow (1 - \omega_{11})\hat{\eta}^1_\star$.

Bi-dimensional case In the bi-dimensional case, it is not always possible to impose bounds on all L_1 weights ω_{ij} freely, that is, independently of the ratio $\frac{\hat{\eta}^2_\star}{\hat{\eta}^1_\star}$. Therefore, we choose to prioritise respecting bounds on the self-excitation L_1 weights ω_{ii}. For event types $i \in [d]$, consider bounds $a_{ii}, b_{ii} \in [0, 1]$ with $a_{ii} < b_{ii}$ such that we want the initial guess of self-excitation L_1 weights ω_{ii} to satisfy $\omega_{ii} \in [a_{ii}, b_{ii}]$.

Algorithm 4: Random initialization of first order parameters, $d = 2$

Data: An estimator $\hat{\boldsymbol{\eta}}_\star > 0$, prior bounds (a_{ij}, b_{ij}) on ω_{ij}.
Result: First order feasible initial guess $(\mu_1, \omega_{11}) \in \mathbb{F}_{\text{MHP}}(\hat{\eta}^1_\star)$.
$r \leftarrow \frac{\hat{\eta}^2_\star}{\hat{\eta}^1_\star}$;
Sample ω_{11} following a uniform distribution on $[a_{11}, b_{11}]$;
Sample ω_{12} following a uniform distribution on $(0, \frac{1-\omega_{11}}{r})$;
Sample ω_{22} following a uniform distribution on $[a_{22}, b_{22}]$;
Sample ω_{21} following a uniform distribution on $(0, r(1 - \omega_{22}))$;
$\boldsymbol{\mu} \leftarrow (\mathbb{I}_d - ||\Phi||_1)\hat{\boldsymbol{\eta}}_\star$.

Bi-variate case If $d \geq 2$, we do not have a general characterization of first order conditions; unless the adjacency matrix follows the bi-variate model in Section 3.1.3.2. In this case, the LLN condition is stronger than the stability condition, leading to a simple algorithm.

Algorithm 5: Random initialization of first order parameters, bi-variate case

Data: An estimator $\hat{\boldsymbol{\eta}}_\star > 0$, prior bounds $(a_S, b_S) \in [0, 1)$ on ω_S.
Result: First order feasible initial guess $(\boldsymbol{\mu}, \omega_S, \omega_C) \in \mathbb{F}_{\text{MHP}}(\hat{\boldsymbol{\eta}}_\star)$
$m(\hat{\boldsymbol{\eta}}_\star) \leftarrow \frac{\min_{k \in [d]} \hat{\eta}^k_\star}{\sum_{i=1}^d \hat{\eta}^i_\star}$;
Sample ω_S following a uniform distribution on $[a_S, b_S]$;
Sample ω_C following a uniform distribution on $[0, \frac{m(\hat{\boldsymbol{\eta}}_\star)}{1-m(\hat{\boldsymbol{\eta}}_\star)}(1 - \omega_S)]$;
$\boldsymbol{\mu} \leftarrow (\mathbb{I}_d - ||\Phi||_1)\hat{\boldsymbol{\eta}}_\star$.

4.5.2 Parameter updates

Gradient clipping Despite the effort in variance reduction to construct of the unbiased LSE gradient estimator, large deviations from the true gradient value can still occur. To avoid these large deviations, we first clip the gradient estimate to an interval of admissible values before proceeding to the parameter updates. Gradient clipping is a classic technique in the deep learning literature, for example in Goodfellow et al. [41] (Section 8.2.4 and 10.11.1), or Zhang et al. [109]. First, define the clipping operator for vectors $\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}$ of the same length by

$$\boldsymbol{y} = \mathrm{clip}\left(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}\right), \tag{4.61}$$

where the m-th component of the clipped vector has components

$$y_m = \begin{cases} u_m & \text{if } x_m \leq u_m, \\ x_m & \text{if } x_m \in [u_m, v_m], \\ v_m & \text{otherwise.} \end{cases} \tag{4.62}$$

To construct the clipped gradient estimate $\boldsymbol{g_k^{(t)}}$ at step t, our heuristic is to force the absolute value of the components of the gradient estimate at step t to be less or equal than the components a positive reference vector $\bar{\boldsymbol{G}}_{\boldsymbol{k}}^{(\boldsymbol{t})}$. The reference $\bar{\boldsymbol{G}}_{\boldsymbol{k}}^{(\boldsymbol{t})}$ is a function of the maximum of absolute gradients over a window of $c_{\text{window}} > 0$ iterations. Formally, define the reference

$$\bar{\boldsymbol{G}}_{\boldsymbol{k}}^{(\boldsymbol{t})} := \boldsymbol{c}_{\text{off}} + c_{\text{tol}} \max_{h \in [c_{\text{window}}]} \left\{ \left| \boldsymbol{g}_{\boldsymbol{k}}^{(\boldsymbol{t-h})} \right| \right\}. \tag{4.63}$$

The hyper-parameter $c_{\text{tol}} \geq 1$ sets how much we can deviate from the maximum. If this coefficient was smaller than 1, the dependence in the previous values would vanish. This coefficient can be taken to be equal to 1, but we face the risk of mitigating too strongly the gradients. Typically, we use $c_{\text{tol}} = 1.1$. The hyper-parameter $\boldsymbol{c}_{\text{off}}$ is a small positive offset. This coefficient is here to avoid getting stuck in a situation where previous gradients are null and not allowing large moves.

$$\boldsymbol{G}_{\boldsymbol{k}}^{(\boldsymbol{t})} := \mathrm{clip}\left(\boldsymbol{\mathcal{G}}_{\boldsymbol{T}}^{(\boldsymbol{k})}, -\bar{\boldsymbol{G}}_{\boldsymbol{k}}^{(\boldsymbol{t})}, \bar{\boldsymbol{G}}_{\boldsymbol{k}}^{(\boldsymbol{t})}\right). \tag{4.64}$$

Schemes As discussed above, we define the parameters update vector $\boldsymbol{\Delta\theta_k^{(t)}}$ using standard schemes from the deep learning literature. Our only modification is a time varying learning rate hyper-parameter $t \mapsto a_{\text{rate}}(t) > 0$, because a constant learning rate does not consistently lead to satisfactory results in our numerical experiments. Denote the element-wise product between vectors or matrices by $\odot$. Denote the two moment hyper-parameters by $a_{\text{M1}} > 0$ and $a_{\text{M2}} > 0$; and a last hyper-parameter $a_{\text{E}} > 0$, which is used to avoid division

by zero. Initialize $\boldsymbol{g_1^{(0)}} = \boldsymbol{g_2^{(0)}} = 0$, and consider a gradient step $t \in \mathbb{N}^*$. We focus on three numerical schemes.

- `Momentum` from Qian [86]; improves the vanilla SGD algorithm by incorporating a momentum term in order to use information from previous gradient updates to smooth parameters trajectory. The associated update is

$$\boldsymbol{\Delta\theta_k^{(t)}} := -a_{\text{rate}}(t) \cdot \boldsymbol{G_k^{(t)}} + a_{\text{M1}} \cdot \boldsymbol{\Delta\theta_k^{(t-1)}}. \tag{4.65}$$

- `RMSprop` from Tieleman and Hinton [99]; is an adaptive learning rate method for each parameters. The associated update is

$$\boldsymbol{g_2^{(t)}} := a_{\text{M2}} \cdot \boldsymbol{g_2^{(t-1)}} + (1 - a_{\text{M2}}) \cdot \boldsymbol{G_k^{(t)}} \odot \boldsymbol{G_k^{(t)}}, \tag{4.66}$$

$$\boldsymbol{\Delta\theta_k^{(t)}} := -a_{\text{rate}}(t) \cdot \frac{\boldsymbol{G_k^{(t)}}}{\sqrt{\boldsymbol{g_2^{(t)}}} + a_{\text{E}}}. \tag{4.67}$$

- `ADAM` from Kingma and Ba [54]; combines the adaptive learning rate idea in `RMSprop` with a more effective use of momentum. One can see `RMSprop` as a special case of `ADAM` with $a_{\text{M1}} = 0$ (no momentum for the gradient term) and no debiasing step. The associated update is

$$g_1^{(t)} := a_{\text{M1}} \cdot \boldsymbol{g_1^{(t-1)}} + (1 - a_{\text{M1}}) \cdot \boldsymbol{G_k^{(t)}}, \qquad \boldsymbol{\bar{g}_1^{(t)}} = \frac{1}{1 - a_{\text{M1}}^t} \cdot \boldsymbol{g_1^{(t)}}, \tag{4.68}$$

$$\boldsymbol{g_2^{(t)}} := a_{\text{M2}} \cdot \boldsymbol{g_2^{(t-1)}} + (1 - a_{\text{M2}}) \cdot \boldsymbol{G_k^{(t)}} \odot \boldsymbol{G_k^{(t)}}, \qquad \boldsymbol{\bar{g}_2^{(t)}} := \frac{1}{1 - a_{\text{M2}}^t} \cdot \boldsymbol{g_2^{(t)}}, \tag{4.69}$$

$$\boldsymbol{\Delta\theta_k^{(t)}} := -a_{\text{rate}}(t) \cdot \frac{\boldsymbol{\bar{g}_1^{(t)}}}{\sqrt{\boldsymbol{\bar{g}_2^{(t)}}} + a_{\text{E}}}. \tag{4.70}$$

In these schemes, the learning rates do not verify Robbins–Monro conditions (see Robbins and Monro [92]). However, Défossez et al. [30] prove the convergence of `ADAM` with specific hyper-parameter values, for smooth objectives with bounded gradients. These schemes do not always perform well in the `ASLSD` method when using a constant learning rate a_{rate}. Empirically, we get our most consistent results using an exponential learning rate: fix an initial learning value $a_{\text{rate}}(0) > 0$, a number of steps before division $a_{\text{steps}} \in \mathbb{N}^*$, a coefficient $a_{\text{div}} \in \mathbb{N}^*$, and a minimal rate $a_{\text{min}} > 0$; and consider

$$a_{\text{rate}}(t) := \max\Big(\frac{a_{\text{rate}}(0)}{a_{\text{div}}^{\lfloor t/a_{\text{steps}} \rfloor}}, a_{\text{min}}\Big). \tag{4.71}$$

The hyper-parameter values that performed well in our numerical experiments are $a_{\text{rate}}(0) = 10^{-1}$, $a_{\text{steps}} = 200$, $a_{\text{div}} = 2$, and $a_{\text{min}} = 10^{-3}$.

4.5.3 Comparison with other estimation methods

Table 4.1 summarizes some features of our algorithm in comparison with existing techniques. The **Algorithm** column contains the reference of a given algorithm, and its name if it has been named in the publication. **Parametric** specifies whether this is a parametric or non-parametric method. **Complexity** is the time complexity of the algorithm to complete a number of iterations n_{iter}, and using the notation of this work. This specific algorithmic complexity does not take into account the fact that different estimation techniques can have different convergence rates. However, for many of these techniques, including ours, there are no theoretical results on convergence rates. Some algorithms consider a discretization of the kernels, we denote by n_{res} the resolution of this discretization. In case the algorithms contain an inner loop, we denote its number of iterations by n'_{iter}. We denote by n_{samples} the number of observed sample paths of the MHP in case the method considers several sample paths. **Assumptions** refers to the additional assumptions made on the MHP (`SBF exp.` for an SBF MHP with exponential kernels, `SBF uni.` for an SBF MHP with $r = 1$ and a unique basis function $\tilde{\phi}_{ij} = \tilde{\phi}$, `exp.` for an MHP with exponential kernels). **Type** is either the type of objective function used (*LSE* for Least Squares Error, *LL* for log-likelihood, *LL-EM* for marginal likelihood in an EM framework) or *MM* in the case of the method of moments. **Regularization** refers to the type of penalty used in the algorithm, if any. In case no penalty is used in the paper, but could be incorporated to the method with mild modifications, we write an asterisk.

4.6 Choosing kernel densities

4.6.1 Motivation

Our algorithm fits within the family of parametric and semi-parametric estimation algorithms for Hawkes models. To the best of our knowledge, robust estimation of Hawkes models and model error have not been extensively explored, and like any other parametric or semi-parametric estimation algorithm, we leave certain modelling decisions to the user based on their knowledge of the system being modeled. Recall that in this work, we consider kernel functions $\phi : [0, +\infty) \to [0, +\infty)$ such that $\phi := \sum_{l=1}^{r} \omega_l \tilde{\phi}_l$. We call the functions $\tilde{\phi}_l : [0, +\infty) \to [0, +\infty)$ kernel densities, with the normalisation condition $\int_{[0,+\infty)} \tilde{\phi}_l = 1$.

As discussed previously, first order properties of an MHP model are fully determined by the constant baseline $\boldsymbol{\mu}$, and the adjacency matrix $||\Phi||_1$. The kernels densities $(\tilde{\phi}_{ij})_{i,j\in[d]}$ control the distribution of offsets in the branching representation of Hawkes models, and appear in second order statistics of the MHP. When using a parametric family of kernel densities, like in any other statistical problem, not all parametric spaces of functions are

Algorithm	Parametric	Complexity	Assumptions	Type	Regularization
ASLSD	param.	$\mathcal{O}(n_{\text{iter}}\rho d^2 Q^{(2)} + n_{\text{iter}}\rho d Q^{(1)})$	-	LSE	*
[3]	param.	$\mathcal{O}(N_T \rho^2 d + n_{\text{iter}}\rho^2 d^3)$	SBF exp.	LSE	Sparsity, Low Rank
MF [8]	param.	$\mathcal{O}(d^2 r \Lambda T \times \max(\Lambda T, d\rho) + d^4\rho^3)$	SBF, Stable, Mean field	LL	*
(SumExp) [15]	param.	$\mathcal{O}(\rho N_T^2)$	SBF exp.	LL	*
[104]	param.	$\mathcal{O}(n_{\text{iter}} N_T^2)$	exp.	LL-EM	-
MPLE [58]	non-param.	$\mathcal{O}(n_{\text{iter}} N_T^2)$	-	LL-EM	Good's penalty
ADM4 [111]	param.	$\mathcal{O}(d^3 n_{\text{iter}} + d^2 N_T n_{\text{samples}} n_{\text{iter}}' + n_{\text{samples}} N_T^2)$	SBF uni.	LL-EM	Sparsity, Low rank
MMEL [112]	non-param.	$\mathcal{O}(\rho N_T^3 d^2 n_{\text{iter}} + \rho n_{\text{res}} n_{\text{iter}}(d N_T + N_T^2))$	-	LL-EM	Kernel smoothing
MLE-SGLP [108]	param.	$\mathcal{O}(\rho N_T^3 d^2 n_{\text{iter}})$	SBF	LL-EM	Sparse, Lasso, Pairwise similarity
(WH) [10]	non-param.	$\mathcal{O}(N_T d^2 n_{\text{res}} + d^4 n_{\text{res}}^3)$	Stable	Autocovariance	-
NPHC [2]	param.	$\mathcal{O}(N_T d^2 + n_{\text{iter}} d^3)$	Stable	Integrated cumulants	-

Table 4.1: Comparison of the computational complexity of our algorithm ASLSD with state of the art estimation of MHP. Our two baseline cases are denoted by SumExp and WH, both here and in subsequent sections.

capable of approximating all models in the space of data generating processes. In this case, the choice of a relevant class of kernels depends, for example, on what level of smoothness and structure is expected, whether kernels should be compactly supported, monotonically decaying, or if kernels should exhibit a specific delayed structure. In practice, exploratory analysis and out-of-sample verification may often be needed.

Notation To simplify notation, fix a kernel density $\tilde{\phi}$. For lags $\tau \geq 0$, the CDF $\tilde{\psi}$ is

$$\tilde{\psi}(\tau) := \int_0^\tau \tilde{\phi}(u)\mathrm{d}u.$$

Fix a kernel density $\tilde{\phi}_\star$, from the same parametric class as $\tilde{\phi}$ or from a different one. For lags $\tau, \sigma \geq 0$, the kernel correlation $\tilde{\Upsilon}$ and the reversed kernel correlation $\tilde{\Upsilon}_\star$ are

$$\tilde{\Upsilon}(\tau, \sigma) := \int_0^\tau \tilde{\phi}(u)\tilde{\phi}_\star(u+\sigma)\mathrm{d}u, \qquad \tilde{\Upsilon}_\star(\tau, \sigma) := \int_0^\tau \tilde{\phi}_\star(u)\tilde{\phi}(u+\sigma)\mathrm{d}u.$$

Fix a time-dependent baseline function μ. For lags $\tau, \sigma \geq 0$, the baseline correlation functional $\tilde{K}$ is

$$\tilde{K}(\tau, \sigma) := \int_0^\tau \tilde{\phi}(u)\mu(u+\sigma)\mathrm{d}u.$$

Numerical requirements Given a parametric family of kernels and baselines, which functionals must be numerically evaluated so that the modeller can use the `ASLSD` algorithm? For the estimation of a Hawkes model using the `ASLSD` method with the MC approximation of functionals from Section 4.4.5, the modeller must be able to evaluate the kernel density function $\tilde{\phi}$ and its derivatives; the logarithm of the kernel function $\log \tilde{\phi}$ and its derivatives; and simulate random variables with probability density $\tilde{\phi}$. Without the MC approximation of functionals, the modeller must be able to evaluate the kernel density function $\tilde{\phi}$ and its derivatives; evaluate the primitive of the kernel function $\tilde{\psi}$ and its derivatives for an MHP model, or evaluate the kernel-baseline correlation function $\tilde{K}$ and its derivatives for an MTLH model; and evaluate the kernel correlation function $\tilde{\Upsilon}$ and its derivatives. We summarize these requirements in Table 4.2. The study of further numerical approximations, such as the approximation of derivatives by finite differences, or the approximation of the logarithm of functions, is out of the scope of this work.

Functional	Estimation (non-MC)	Estimation (MC)	Residuals
$\tilde{\phi}$	✓	✓	✗
$\log \tilde{\phi}$	✗	✓	✗
$\tilde{K}$	✓	✗	✓
$\tilde{\Upsilon}$	✓	✗	✗

Table 4.2: Requirements for Hawkes models

Outline In this technical section, we discuss different parametric classes of kernel densities $\tilde{\phi}$ that we use in our numerical experiments (see Chapter 5) and in our models of Nasdaq equities prices (see Part II). Section 4.6.2 gives results for the exponential kernel density, and introduces `RecExp`, our estimation algorithm for MTLH with exponential kernels based on the recursive exponential trick. Section 4.6.4 discusses power law and Rayleigh kernel densities, two other classic examples. Section 4.6.3 gives results for four families of kernel densities which are dense in the space of Hawkes kernels for the weak topology: uniform, triangular, Gaussian, and Gamma kernels. These families are also suitable for semi-parametric estimation approaches. Finally, Section 4.6.5 discusses delaying kernel densities.

4.6.2 Exponential kernels and Markovian models

Exponential kernels are the fundamental class of Hawkes models in the literature. We give basic results for the standard exponential kernel, and discussion the estimation of MTLH models with exponential kernels.

4.6.2.1 Exponential kernel

Definition 4.6.1 (Exponential kernel density)**.** *For lags $\tau \geq 0$, the exponential kernel density is*

$$\tilde{\phi}(\tau) := \beta e^{-\beta\tau},$$

and the parameter is the decay $\beta \in (0, +\infty)$.

Note that because of our normalisation constraint, our definition differs with some definitions of exponential kernels in the literature. The exponential kernel density degenerates to the null kernel density for $\beta \to 0$; and to the Dirac distribution in 0 for $\beta \to +\infty$.

Proposition 4.6.1 (Kernel density derivatives)**.** *For lags $\tau \geq 0$, the derivatives of $\log \tilde{\phi}$ with respect to model parameters are*

$$\frac{\partial \log \tilde{\phi}}{\partial \beta}(\tau) = \frac{1}{\beta} - \tau.$$

We get the CDF associated to this kernel density below.

Proposition 4.6.2 (CDF)**.** *For lags $\tau \geq 0$, the exponential CDF is*

$$\tilde{\psi}(\tau) = 1 - e^{-\beta\tau}, \tag{4.72}$$

and the derivatives of $\tilde{\psi}$ with respect to model parameters are

$$\frac{\partial \tilde{\psi}}{\partial \beta}(\tau) = \tau e^{-\beta\tau}. \tag{4.73}$$

If U is a uniform random variable on $[\,0,1\,]$, then the random offset $\tilde{\tau} := -\frac{1}{\beta}\log(U)$ has density $\tilde{\phi}$. The moments of $\tilde{\tau}$ are $\mathbb{E}[\tilde{\tau}] = \frac{1}{\beta}$, and $\mathrm{Var}[\tilde{\tau}] = \frac{1}{\beta^2}$. We now give closed-form expressions for the kernel correlation functions of $\tilde{\phi}$ with a kernel $\tilde{\phi}_\star$. First, suppose $\tilde{\phi}_\star = \tilde{\phi}$.

Proposition 4.6.3 (Auto-correlation)**.** *For lags $\tau, \sigma \geq 0$, the auto-correlation of the exponential kernel density is*

$$\tilde{\Upsilon}(\tau,\sigma) = \frac{\beta}{2}e^{-\beta\sigma}(1 - e^{-2\beta\tau}), \tag{4.74}$$

and the derivatives of $\tilde{\Upsilon}$ with respect to model parameters are

$$\frac{\partial\tilde{\Upsilon}}{\partial\beta}(\tau,\sigma) := \frac{1}{2}e^{-\beta\sigma}\bigg(1 - \beta\sigma + e^{-2\beta\tau}(\beta(2\tau+\sigma) - 1)\bigg). \tag{4.75}$$

Note that for lags $\tau, \sigma \geq$, $\tilde{\Upsilon}(\tau,\sigma) \leq \frac{\beta}{2}e^{-\beta\sigma}$. Furthermore, $\frac{\partial\tilde{\Upsilon}}{\partial\beta}(\tau,0) \geq 0$ and

$$\frac{\partial\tilde{\Upsilon}}{\partial\beta}(\tau,0) = 0 \iff \beta = \frac{1}{2t}. \tag{4.76}$$

Suppose $\tilde{\phi}_\star$ is an exponential kernel density with parameter $\beta_\star$ such that $\tilde{\phi}_\star \neq \tilde{\phi}$. Define

$$b := \frac{\beta}{\beta+\beta_\star}, \qquad b_\star := \frac{\beta_\star}{\beta+\beta_\star}. \tag{4.77}$$

Proposition 4.6.4 (Correlation with exponential kernel density)**.** *For lags $\tau, \sigma \geq 0$, the cross-correlation Υ is*

$$\tilde{\Upsilon}(\tau,\sigma) = \frac{\beta\beta_\star}{\beta+\beta_\star}e^{-\beta_\star\sigma}\big(1 - e^{-(\beta+\beta_\star)\tau}\big), \tag{4.78}$$

and the derivatives of $\tilde{\Upsilon}$ with respect to model parameters are

$$\begin{aligned}
\frac{\partial\tilde{\Upsilon}}{\partial\beta}(\tau,\sigma) &= b_\star e^{-\beta_\star\sigma}\bigg(b_\star + (\beta\tau - b_\star)e^{-(\beta+\beta_\star)\tau}\bigg),\\
\frac{\partial\tilde{\Upsilon}}{\partial\beta_\star}(\tau,\sigma) &= be^{-\beta_\star\sigma}\bigg(b - \beta_\star\sigma + \big(\beta_\star(\tau+\sigma) - b\big)e^{-(\beta+\beta_\star)\tau}\bigg).
\end{aligned} \tag{4.79}$$

Finally, suppose $\tilde{\phi}_\star$ is a Gaussian kernel density (see Section 4.6.3.3) with parameters $(\beta_\star, \delta_\star)$. For $\sigma \geq 0$, define the transformed lag $\sigma_\star := \delta_\star - \sigma - \beta\beta_\star^2$. Since $\tilde{\phi}$ is causal, for $x \in \mathbb{R}$, define

$$g_{\mathcal{E}}(x) := \beta\exp(-\beta x), \tag{4.80}$$

Proposition 4.6.5 (Correlation with Gaussian kernel density)**.** *For lags $\tau, \sigma \geq 0$, the cross-correlation Υ is*

$$\tilde{\Upsilon}(\tau,\sigma) = g_{\mathcal{E}}\left(\delta_\star - \sigma - \frac{\beta\beta_\star^2}{2}\right)\left(F_{\mathcal{N}}\left(\frac{\tau - (\delta_\star - \sigma - \beta\beta_\star^2)}{\beta_\star}\right) - F_{\mathcal{N}}\left(-\frac{\delta_\star - \sigma - \beta\beta_\star^2}{\beta_\star}\right)\right), \quad (4.81)$$

and the derivatives of $\tilde{\Upsilon}$ with respect to model parameters are

$$\begin{aligned}\frac{\partial\tilde{\Upsilon}}{\partial\beta}(\tau,\sigma) = g_{\mathcal{E}}\left(\sigma_\star + \frac{\beta\beta_\star^2}{2}\right)\Bigg[&\beta_\star\left(f_{\mathcal{N}}\left(\frac{\tau-\sigma_\star}{\beta_\star}\right) - f_{\mathcal{N}}\left(-\frac{\sigma_\star}{\beta_\star}\right)\right)\\ &+\left(\frac{1}{\beta} - \sigma_\star\right)\left(F_{\mathcal{N}}\left(\frac{\tau-\sigma_\star}{\beta_\star}\right) - F_{\mathcal{N}}\left(-\frac{\sigma_\star}{\beta_\star}\right)\right)\Bigg],\end{aligned} \quad (4.82)$$

$$\begin{aligned}\frac{\partial\tilde{\Upsilon}}{\partial\beta_\star}(\tau,\sigma) = g_{\mathcal{E}}\left(\sigma_\star + \frac{\beta\beta_\star^2}{2}\right)\Bigg[&\beta^2\beta_\star\left(F_{\mathcal{N}}\left(\frac{\tau-\sigma_\star}{\beta_\star}\right) - F_{\mathcal{N}}\left(-\frac{\sigma_\star}{\beta_\star}\right)\right)\\ &+\left(\left(2\beta + \frac{\sigma_\star - \tau}{\beta_\star^2}\right)f_{\mathcal{N}}\left(\frac{\tau-\sigma_\star}{\beta_\star}\right) - \left(2\beta + \frac{\sigma_\star}{\beta_\star^2}\right)f_{\mathcal{N}}\left(-\frac{\sigma_\star}{\beta_\star}\right)\right]\end{aligned}, \quad (4.83)$$

$$\begin{aligned}\frac{\partial\tilde{\Upsilon}}{\partial\delta_\star}(\tau,\sigma) = -g_{\mathcal{E}}\left(\sigma_\star + \frac{\beta\beta_\star^2}{2}\right)\Bigg[&\beta\left(F_{\mathcal{N}}\left(\frac{\tau-\sigma_\star}{\beta_\star}\right) - F_{\mathcal{N}}\left(-\frac{\sigma_\star}{\beta_\star}\right)\right)\\ &+\frac{1}{\beta_\star}\left(f_{\mathcal{N}}\left(\frac{\tau-\sigma_\star}{\beta_\star}\right) - f_{\mathcal{N}}\left(-\frac{\sigma_\star}{\beta_\star}\right)\right)\Bigg].\end{aligned} \quad (4.84)$$

Finally, we give formulas for the kernel–baseline correlation functional $\tilde{K}$ of $\tilde{\phi}$ with different classes of baselines μ. First, suppose μ is a linear baseline with parameters (a, b) (see Section 2.4.4).

Proposition 4.6.6 ($\tilde{K}$ with linear baselines)**.** *For lags $\tau, \sigma \geq 0$, the cross-correlation $\tilde{K}$ is*

$$\tilde{K}(\tau,\sigma) = \left(a\left(\sigma + \frac{1}{\beta}\right) + b\right) - \left(a\left(\tau + \sigma + \frac{1}{\beta}\right) + b\right)e^{-\beta\tau}; \quad (4.85)$$

the derivative of $\tilde{K}$ with respect to the kernel parameter is

$$\frac{\partial\tilde{K}}{\partial\beta}(\tau,\sigma) = -\frac{a}{\beta^2} + \left(at^2 + ast + \left(b + \frac{a}{\beta}\right)\tau + \frac{a}{\beta^2}\right)e^{-\beta\tau}, \quad (4.86)$$

and the derivatives of $\tilde{K}$ with respect to baseline parameters are

$$\frac{\partial\tilde{K}}{\partial a}(\tau,\sigma) = \left(\sigma + \frac{1}{\beta}\right) - \left(\tau + \sigma + \frac{1}{\beta}\right)e^{-\beta\tau}, \qquad \frac{\partial\tilde{K}}{\partial b}(\tau,\sigma) = 1 - e^{-\beta\tau}. \quad (4.87)$$

Now suppose μ is a cosine baseline with parameters (α, δ, a, b) (see Section 2.4.3).

Proposition 4.6.7 ($\tilde{K}$ with cosine baselines)**.** *For lags $\tau, \sigma \geq 0$, the cross-correlation $\tilde{K}$ is*

$$\begin{aligned}\tilde{K}(\tau,\sigma) = (\alpha+\delta)(1-\exp(-\beta\tau)) \\ + \frac{\beta\alpha}{a^2+\beta^2}\Bigg(\beta\cos(a\sigma+b) - a\sin(a\sigma+b) \\ - e^{-\beta\tau}\Big(\beta\cos\big(a(\tau+\sigma)+b\big) - a\sin\big(a(\tau+\sigma)+b\big)\Big)\Bigg).\end{aligned} \tag{4.88}$$

4.6.2.2 Recurrent exponential algorithm

To the best of our knowledge, there is no method for the parametric estimation of MTLH models in the Hawkes literature besides `ASLSD`. In this paragraph, we propose an estimation algorithm with exact evaluation of the LSE gradient, in the special case of an MTLH were all kernels are exponential. We extend the well-known recurrence formula of exponential MHP to this exponential MTLH, and compute exactly the LSE and residuals with linear time complexity. This provides a reference case to verify our algorithms with large amounts of data. We refer to this algorithm as `RecExp`. Formally, assume all kernels are of the form

$$\phi_{ij}(x) := \sum_{l=1}^{r_{ij}} \omega_{ij,l}\tilde{\phi}_{ij,l}(x), \quad \text{with kernel densities} \quad \tilde{\phi}_{ij,l}(x) := \beta_{ij,l}e^{-\beta_{ij,l}x}.$$

We do not make assumptions on the background intensity nor the impact functions.

Exponential trick This method relies on the following property, closely related to the Markovianity of the MHP with exponential kernels. In the rest of this paragraph, fix event types $i, j, k \in [d]$, a mixture index $l' \in [r_{kj}]$, and an event index $m \in [\![\varpi(i,j), N_T^i[\![$. Let $U := (U_n)_{n\in\mathbb{N}}$ be a real valued sequence. Define the exponential sum operator $S^{(m)}_{ijk,l'}$ acting on the sequence U by

$$S^{(m)}_{ijk,l'}[U] := \sum_{n=1}^{\kappa(j,i,m)} e^{-\beta_{kjl'}(t_m^i - t_n^j)}U_n. \tag{4.89}$$

This operator is essential for our `RecExp` method as we show that expressions of this form appear in the LSE and residuals formulas. Without pre-computations, the computation of the m-th exponential sum term $S^{(m)}_{ijk,l'}[U]$ has time complexity $\mathcal{O}(\kappa(j,i,m))$. Therefore, the brute force computation of the full sequence of exponential sum terms for $m \in [\![\varpi(i,j), N_T^i[\![$ has time complexity $\mathcal{O}\left((N_T^j)^2\right)$. However, the following property offers a recurrence formula that simplifies computations to linear time.

Proposition 4.6.8 (Recurrence property)**.** *The $(m+1)$-th exponential sum term is computed from the m-th exponential sum term using*

$$S^{(m+1)}_{ijk,l'}[(U_n)_n] = e^{-\beta_{kjl'}(t^i_{m+1}-t^i_m)} S^{(m)}_{ijk,l'}[(U_n)_n] + \mathbb{1}_{\{\kappa(j,i,m+1)>\kappa(j,i,m)\}} \sum_{n=\kappa(j,i,m)+1}^{\kappa(j,i,m+1)} e^{-\beta_{kjl'}(t^i_{m+1}-t^j_n)} U_n. \tag{4.90}$$

This formula allows to compute all the terms of the exponential sum by recurrence with time complexity $\mathcal{O}(N^i_T)$. Note that we also compute recursively the derivative of these terms with respect to the decay rate $\beta_{kjl'}$, since

$$\frac{\partial S^{(m)}_{ijk,l'}[U]}{\partial \beta_{kjl'}} = S^{(m)}_{ijk,l'}[(t^j_n U_n)_n] - t^i_m S^{(m)}_{ijk,l'}[U]. \tag{4.91}$$

We compute the conditional intensity of the `RecExp` model at all jump times in linear time. Formally, for event types $k \in [d]$, and event indices $m \in [N^k_T]$, the conditional intensity of the `RecExp` model satisfies the recursive formula

$$\lambda_k(t^k_m) = \mu_k(t^k_m) + \sum_{j=1}^{d} \sum_{l'=1}^{r_{kj}} \omega_{kj,l'} \beta_{kj,l'} S^{(m)}_{kjk,l'}[\boldsymbol{\mathcal{I}_{kj}}(\boldsymbol{\xi^j})]\,. \tag{4.92}$$

We also give a recursive expression for the residuals of the `RecExp` model

Proposition 4.6.9 (Linear expression for transformed times)**.** *For all event indices $m \in [N^k_T]$, the m-th transformed time of the model is*

$$s^{(k)}_m = \int_0^{t^k_m} \mu_k(t)\mathrm{d}t + \sum_{j=1}^{d} \mathbb{E}[\mathcal{I}_{kj}(\xi^j_m)]\, \mathbb{1}_{\{m \geq \varpi(k,j)\}} \sum_{l'=1}^{r_{kj}} \omega_{kjl} \left(\kappa(j,k,m) - S^{(m)}_{kjk,l'}[(1)_n] \right). \tag{4.93}$$

Estimation procedure We show how the exponential trick allows to compute the LSE and its gradient in linear time. In the rest of this section, fix event types $i, j, k \in [d]$, and mixture indices $l' \in [r_{ki}]$ and $l' \in [r_{kj}]$. Define the decay rate shares

$$b^{(1)}_{ijk,ll'} := \frac{\beta_{kil}}{\beta_{kjl'} + \beta_{kjl'}}, \quad b^{(2)}_{ijk,ll'} := \frac{\beta_{kjl'}}{\beta_{kjl'} + \beta_{kjl'}},$$

and the half of the harmonic average of decay rates

$$b_{ijk,ll'} := \frac{\beta_{kil}\beta_{kjl'}}{\beta_{kjl'} + \beta_{kjl'}}.$$

For indices $m \in [\![\varpi(i,j), N^i_T]\!]$, define the exponential term

$$\epsilon^{(m)}_{ijk,ll'} := e^{-(\beta_{kil}+\beta_{kjl'})(T-t^i_m)}.$$

Finally, denote the sequences

$$\boldsymbol{\mathcal{I}_{kj}}(\boldsymbol{\xi^j}) := \left(\mathcal{I}_{kj}(\xi^j_n)\right)_{\{n \in [N^j_T]\}}, \qquad \boldsymbol{t^j \mathcal{I}_{kj}}(\boldsymbol{\xi^j}) := \left(t^j_n \mathcal{I}_{kj}(\xi^j_n)\right)_{\{n \in [N^j_T]\}}.$$

First, we get formulas for the single sums $Z_{\Upsilon_{iik,ll'}}$ and their derivatives.

Proposition 4.6.10 (Linear formulas for $Z_{\Upsilon_{iik,ll'}}$)**.** *The sum $Z_{\Upsilon_{iik,ll'}}$ satisfies*

$$Z_{\Upsilon_{iik,ll'}} = \omega_{kil}\omega_{kil'}b_{iik,ll'}\sum_{m=1}^{N_T^i}\mathcal{I}_{ki}(\xi_m^i)^2\big(1-\epsilon_{iik,ll'}^{(m)}\big). \tag{4.94}$$

If the mixture indices verify $l = l'$, the expression of $Z_{\Upsilon_{iik,ll}}$ becomes

$$Z_{\Upsilon_{iik,ll}} = \frac{1}{2}\omega_{kil}^2\beta_{kil}\sum_{m=1}^{N_T^i}\mathcal{I}_{ki}(\xi_m^i)^2\big(1-\epsilon_{iik,ll}^{(m)}\big), \tag{4.95}$$

and the derivatives of $Z_{\Upsilon_{iik,ll}}$ with respect to model parameters are

$$\begin{aligned}
\frac{\partial Z_{\Upsilon_{iik,ll'}}}{\partial\omega_{kil}} &= \omega_{kil}\beta_{kil}\sum_{m=1}^{N_T^i}\mathcal{I}_{ki}(\xi_m^i)^2\big(1-\epsilon_{iik,ll}^{(m)}\big),\\
\frac{\partial Z_{\Upsilon_{iik,ll'}}}{\partial\beta_{kil}} &= \frac{1}{2}\omega_{kil}^2\sum_{m=1}^{N_T^i}\mathcal{I}_{ki}(\xi_m^i)^2\big(1+(2\beta_{kil}(T-t_m^i)-1)\epsilon_{iik,ll}^{(m)}\big).
\end{aligned} \tag{4.96}$$

If the mixture indices verify $l \neq l'$, the derivative of $Z_{\Upsilon_{iik,ll'}}$ with respect to the L_1 weights parameters are

$$\begin{aligned}
\frac{\partial Z_{\Upsilon_{iik,ll'}}}{\partial\omega_{kil}} &= \omega_{kil'}b_{iik,ll'}\sum_{m=1}^{N_T^i}\mathcal{I}_{ki}(\xi_m^i)^2\big(1-\epsilon_{iik,ll'}^{(m)}\big),\\
\frac{\partial Z_{\Upsilon_{iik,ll'}}}{\partial\omega_{kil'}} &= \omega_{kil}b_{iik,ll'}\sum_{m=1}^{N_T^i}\mathcal{I}_{ki}(\xi_m^i)^2\big(1-\epsilon_{iik,ll'}^{(m)}\big),
\end{aligned} \tag{4.97}$$

and the derivatives of $Z_{\Upsilon_{iik,ll'}}$ with respect to the decay rates parameters are

$$\begin{aligned}
\frac{\partial Z_{\Upsilon_{iik,ll'}}}{\partial\beta_{kil}} &= \omega_{kil}\omega_{kil'}b_{iik,ll'}^{(2)}\sum_{m=1}^{N_T^i}\mathcal{I}_{ki}(\xi_m^i)^2\left(b_{iik,ll'}^{(2)}+\big((T-t_m^i)\beta_{kil}-b_{iik,ll'}^{(2)}\big)\epsilon_{iik,ll'}^{(m)}\right),\\
\frac{\partial Z_{\Upsilon_{iik,ll'}}}{\partial\beta_{kil'}} &= \omega_{kil}\omega_{kil'}b_{iik,ll'}^{(1)}\sum_{m=1}^{N_T^i}\mathcal{I}_{ki}(\xi_m^i)^2\left(b_{iik,ll'}^{(1)}+\big((T-t_m^i)\beta_{kil'}-b_{iik,ll'}^{(1)}\big)\epsilon_{iik,ll'}^{(m)}\right).
\end{aligned} \tag{4.98}$$

Second, we get formulas for the double-indexed sums $S_{\Upsilon_{ijk,ll'}}$ and $S_{\phi_{kj,l}}$, and their derivatives.

Proposition 4.6.11 (Linear computation of $S_{\phi_{kj,l'}}$)**.** *The double sum $S_{\phi_{kj,l'}}$ satisfies*

$$S_{\phi_{kj,l'}} = \omega_{kj,l'}\beta_{kj,l'}\sum_{m=\varpi(k,j)}^{N_T^k}S_{kjk,l'}^{(m)}[\boldsymbol{\mathcal{I}_{kj}(\xi^j)}]\,. \tag{4.99}$$

The derivative of $S_{\phi_{kj,l'}}$ with respect to the L_1 weight parameter is

$$\frac{\partial S_{\phi_{kj,l'}}}{\partial\omega_{kj,l'}} = \beta_{kj,l'}\sum_{m=\varpi(k,j)}^{N_T^k}S_{kjk,l'}^{(m)}[\boldsymbol{\mathcal{I}_{kj}(\xi^j)}]\,. \tag{4.100}$$

The derivative of $S_{\phi_{kj,l'}}$ with respect to the decay rate parameter is

$$\frac{\partial S_{\phi_{kj,l'}}}{\partial \beta_{kj,l'}} = \omega_{kj,l'} \sum_{m=\varpi(k,j)}^{N_T^k} (1-\beta_{kj,l'} t_m^k) S_{kjk,l'}^{(m)} [\boldsymbol{\mathcal{I}_{kj}}(\boldsymbol{\xi^j})] + \omega_{kj,l'} \beta_{kj,l'} \sum_{m=\varpi(k,j)}^{N_T^k} S_{kjk,l'}^{(m)} [\boldsymbol{t^j \mathcal{I}_{kj}}(\boldsymbol{\xi^j})] \,. \tag{4.101}$$

Proposition 4.6.12 (Linear computation of $S_{\Upsilon_{ijk,ll'}}$)**.** *The double sum $S_{\Upsilon_{ijk,ll'}}$ verifies*

$$S_{\Upsilon_{ijk,ll'}} = \omega_{kil} \omega_{kjl'} b_{ijk,ll'} \sum_{m=\varpi(i,j)}^{N_T^i} \mathcal{I}_{ki}(\xi_m^i)(1-\epsilon_{ijk,ll'}^{(m)}) S_{ijk,l'}^{(m)} [\boldsymbol{\mathcal{I}_{kj}}(\boldsymbol{\xi^j})] \,. \tag{4.102}$$

If the mixture indices verify $(i,l) = (j,l')$, this expression becomes

$$S_{\Upsilon_{iik,ll}} = \frac{1}{2} \omega_{kil}^2 \beta_{kil} \sum_{m=\varpi(i,i)}^{N_T^i} \mathcal{I}_{ki}(\xi_m^i)(1-\epsilon_{iik,ll}^{(m)}) S_{iik,l}^{(m)} [\boldsymbol{\mathcal{I}_{ki}}(\boldsymbol{\xi^i})] \,. \tag{4.103}$$

The derivative of $S_{\Upsilon_{iik,ll}}$ with respect to the L_1 weight ω_{kil} is

$$\frac{\partial S_{\Upsilon_{iik,ll}}}{\partial \omega_{kil}} = \omega_{kil} \beta_{kil} \sum_{m=\varpi(i,i)}^{N_T^i} \mathcal{I}_{ki}(\xi_m^i)(1-\epsilon_{iik,ll}^{(m)}) S_{iik,l}^{(m)} [\boldsymbol{\mathcal{I}_{ki}}(\boldsymbol{\xi^i})] \,, \tag{4.104}$$

and the derivative of $S_{\Upsilon_{iik,ll}}$ with respect to the decay rate β_{kil} is

$$\begin{aligned} \frac{\partial S_{\Upsilon_{iik,ll}}}{\partial \beta_{kil}} = \frac{1}{2} \omega_{kil}^2 \sum_{m=\varpi(i,i)}^{N_T^i} \mathcal{I}_{ki}(\xi_m^i)(1-\beta_{kil} t_m^i + (\beta_{kil}(2T-t_m^i)-1)\epsilon_{iik,ll}^{(m)}) S_{iik,l}^{(m)} [\boldsymbol{\mathcal{I}_{ki}}(\boldsymbol{\xi^i})] \\ + \frac{1}{2} \omega_{kil}^2 \beta_{kil} \sum_{m=\varpi(i,i)}^{N_T^i} \mathcal{I}_{ki}(\xi_m^i)(1-\epsilon_{iik,ll}^{(m)}) S_{iik,l}^{(m)} [\boldsymbol{t^i \mathcal{I}_{ki}}(\boldsymbol{\xi^i})] \,. \end{aligned} \tag{4.105}$$

If the mixture indices verify $(i,l) \neq (j,l')$, the derivatives of $\Upsilon_{ijk,ll'}$ with respect to the L_1 weight parameters are

$$\begin{aligned} \frac{\partial S_{\Upsilon_{ijk,ll'}}}{\partial \omega_{kil}} &= \omega_{kjl'} b_{ijk,ll'} \sum_{m=\varpi(i,j)}^{N_T^i} \mathcal{I}_{ki}(\xi_m^i) \Big(1-\epsilon_{ijk,ll'}^{(m)}\Big) S_{ijk,l'}^{(m)} [\boldsymbol{\mathcal{I}_{kj}}(\boldsymbol{\xi^j})] \,, \\ \frac{\partial S_{\Upsilon_{ijk,ll'}}}{\partial \omega_{kjl'}} &= \omega_{kil} b_{ijk,ll'} \sum_{m=\varpi(i,j)}^{N_T^i} \mathcal{I}_{ki}(\xi_m^i) \Big(1-\epsilon_{ijk,ll'}^{(m)}\Big) S_{ijk,l'}^{(m)} [\boldsymbol{\mathcal{I}_{kj}}(\boldsymbol{\xi^j})] \,. \end{aligned} \tag{4.106}$$

The derivative of $\Upsilon_{ijk,ll'}$ with respect to the decay rate β_{kil} is

$$\begin{aligned} \frac{\partial S_{\Upsilon_{ijk,ll'}}}{\partial \beta_{kil}} = \omega_{kil} \omega_{kjl'} b_{ijk,ll'}^{(2)} \sum_{m=\varpi(i,j)}^{N_T^i} \mathcal{I}_{ki}(\xi_m^i) \Bigg(\Big((T-t_m^i)\beta_{kil} - b_{ijk,ll'}^{(2)}\Big) \epsilon_{ijk,ll'}^{(m)} \\ + b_{ijk,ll'}^{(2)} \Bigg) S_{ijk,l'}^{(m)} [\boldsymbol{\mathcal{I}_{kj}}(\boldsymbol{\xi^j})] \,. \end{aligned} \tag{4.107}$$

and the derivative of $\Upsilon_{ijk,ll'}$ *with respect to the decay rate* $\beta_{kjl'}$ *is*

$$\begin{aligned}\frac{\partial S_{\Upsilon_{ijk,ll'}}}{\partial \beta_{kjl'}} = \omega_{kil}\omega_{kjl'} b^{(1)}_{ijk,ll'} \sum_{m=\varpi(i,j)}^{N_T^i} \left(b^{(1)}_{ijk,ll'} - \beta_{kjl'} t^i_m + \left(T\beta_{kjl'} - b^{(1)}_{ijk,ll'}\right)\epsilon^{(m)}_{ijk,ll'} \right) \\ \times \mathcal{I}_{ki}(\xi^i_m) S^{(m)}_{ijk,l'}[\boldsymbol{\mathcal{I}_{kj}(\xi^j)}] \\ + \omega_{kil}\omega_{kjl'}\beta_{kjl'} b^{(i)}_{ijk,ll'} \sum_{m=\varpi(i,j)}^{N_T^i} \mathcal{I}_{ki}(\xi^i_m)\left(1-\epsilon^{(m)}_{ijk,ll'}\right) S^{(m)}_{ijk,l'}[\boldsymbol{t^j \mathcal{I}_{kj}(\xi^j)}]\,.\end{aligned} \tag{4.108}$$

Conclusion Based on the previous formulas, we can now compute exactly the gradient of the k-th partial LSE $\nabla_{\boldsymbol{\theta_k}} \mathcal{R}^{(k)}_T(\boldsymbol{\theta_k})$ in linear time complexity. If the exponential decay rate parameters $\boldsymbol{\beta}$ are fixed, we can accelerate this method by pre-computing the exponential sum terms with linear time complexity and storing them; in that case, each gradient iteration has trivial time complexity $\mathcal{O}(1)$. However, for each LSE minimization problem $k \in [d]$, the pre-computation step has memory complexity $\mathcal{O}(r^2 N_T)$: while this is linear in the number of events, the memory complexity is quadratic in the number of basis functions r. Numerically, this requires additional care in the type of data structures used to store the pre-computed exponential sums.

4.6.3 Dense kernel families

In this section, we discuss families of kernels which mixtures are dense in the set of kernels for both the weak topology and the Euclidean topology: uniform, triangular, Gaussian, and Gamma kernel densities. Because of their density property, these families of kernels represent a semi-parametric approach to the kernel estimation problem. The decision of which specific dense family to choose should again be guided by the properties expected of the true kernel. For example, if the true kernel is believed to have bounded support, a mixture of triangular functions could be an adequate choice. If the modeler decides to use a mixture of truncated Gaussian kernels and believes some specific modes are of interest in the system studied, they can be targeted by the choice of Gaussian means in the mixture.

Weak topology Consider the weak topology on probability measures on $[0, +\infty)$ defined by the convergence in distribution. A simple strategy to show that a family of kernels is dense for the weak topology is to show that for all $x \geq 0$, we can construct a sequence of kernel densities $\tilde{\phi}^{(p)}$ such that $\tilde{\phi}^{(p)} \underset{p\to+\infty}{\to} \delta_x$ weakly, where δ_x is the Dirac delta function in x. Since the family of Dirac distributions is dense for the weak topology, the result follows immediately.

Euclidean topology In the deep learning literature, the work on universal approximation theorems focused notably on density of neural network models in L_p spaces. For instance, Theorem 3 of Hornik [52] states that shallow neural networks with arbitrarily large numbers of hidden units, and bounded non-constant activation function of class $\mathcal{C}_p$, are dense in L_p for finite measures with compact support. The L_2 density of the semi-parametric families of kernels below is interesting to us because of the quadratic structure of the LSE. However, although for a well-chosen activation function, each of these families can be expressed as a subset of shallow neural networks with non-negative weights[3], it is not clear whether mixtures of these kernels are dense in the space of kernels for the Euclidean topology.

4.6.3.1 Uniform

Mixtures of uniform kernels are a particularly simple yet fundamental model. In the Hawkes literature, Reynaud-Bouret and Schbath [90] use piece-wise constant kernel functions for instance. With uniform densities, we get the different kernel and baseline functionals appearing in the `ASLSD` method in closed-form for general kernels, and their evaluation is numerically inexpensive.

Definition 4.6.2 (Uniform kernel density)**.** *For lags $\tau \geq 0$, the uniform kernel is*

$$\tilde{\phi}(\tau) := \frac{1}{a} \mathbb{1}_{\{0 \leq \tau - \delta \leq a\}},$$

and the parameters are the lower bound $\delta \in (0, +\infty)$, and the interval size $a \in (0, +\infty)$.

Fix an estimation horizon $b > 0$, and a number of basis functions $r \in \mathbb{N}^*$. Consider a ground truth kernel $\phi^\diamond$, that we want to approximate in L_2 on $[0, b]$ using a SBF uniform kernel with r basis functions. Define the uniform step size $a := b/r$; and the edges $\delta_l := a(l-1)$ for $l \in [r]$. For SBF models $\phi_{\boldsymbol{\omega}} := \sum_{l=1}^r \omega_l \mathbb{1}_{[\delta_l, \delta_l + a)}$, we want to minimize the L_2 projection error

$$\mathcal{L}(\boldsymbol{\omega}) := \frac{1}{2} \|\phi_{\boldsymbol{\omega}} - \phi^\diamond\|_{2,T}^2.$$

The minimizer $\boldsymbol{\omega}^\star$ of the loss function $\mathcal{L}$ is

$$\boldsymbol{\omega}^\star = \Big(\int_{\delta_i}^{\delta_{i+1}} \phi^\diamond \Big)_{l \in [r]}, \qquad \mathcal{L}(\boldsymbol{\omega}^\star) = \frac{1}{2} \|\phi^\diamond\|_{2,T}^2 - \frac{1}{2} \frac{1}{a} \sum_{l=1}^r \Big(\int_{\delta_i}^{\delta_{i+1}} \phi^\diamond \Big)^2. \tag{4.109}$$

We get the CDF associated to this kernel density below.

Proposition 4.6.13 (CDF)**.** *For lags $\tau \geq 0$, the uniform CDF is*

$$\tilde{\psi}(\tau) = \min \left(\frac{(\tau - \delta)_+}{a}, 1 \right), \tag{4.110}$$

[3]The kernel families discussed in this chapter are stable by composition with a linear function.

and the derivatives of $\tilde{\psi}$ with respect to model parameters are

$$\frac{\partial\tilde{\psi}}{\partial\delta}(\tau) = -\frac{1}{a}\mathbb{1}_{\left\{\frac{\tau-\delta}{a}\in(0,1)\right\}}, \qquad \frac{\partial\tilde{\psi}}{\partial a}(\tau) = -\frac{\tau-\delta}{a^2}\mathbb{1}_{\left\{\frac{\tau-\delta}{a}\in(0,1)\right\}}. \tag{4.111}$$

Note that if U is a uniform random variable on $[0,1]$, then the random offset $\tilde{\tau} := aU + \delta$ is distributed following $\tilde{\phi}$. With our parametrization, the moments of $\tilde{\tau}$ are $\mathbb{E}[\tilde{\tau}] = \delta + a/2$ and $\mathrm{Var}[\tilde{\tau}] = \frac{a^2}{12}$. We now get the kernel correlation functionals of the uniform kernel density in closed form. First, suppose $\tilde{\phi}_\star = \tilde{\phi}$.

Proposition 4.6.14 (Auto-correlation)**.** *For lags $\tau, \sigma \geq 0$, the auto-correlation of the uniform kernel density is*

$$\tilde{\Upsilon}(\tau,\sigma) = \frac{1}{a^2}\min\Big((\tau-\delta)_+, (a-\sigma)_+\Big), \tag{4.112}$$

the derivative of $\tilde{\Upsilon}(\tau,\sigma)$ with respect to the lower bound parameter δ is

$$\frac{\partial\tilde{\Upsilon}}{\partial\delta}(\tau,\sigma) = -\frac{1}{a^2}\mathbb{1}_{\{\tau\in[\delta,\delta+a),\quad \sigma\in[0,a),\quad \tau+\sigma\in[0,\delta+a)\}}, \tag{4.113}$$

and the derivative of $\tilde{\Upsilon}(\tau,\sigma)$ with respect to the interval size parameter a is

$$\frac{\partial\tilde{\Upsilon}}{\partial a}(\tau,\sigma) = \mathbb{1}_{\sigma\in[0,a)}\left(\mathbb{1}_{\{\tau+\sigma\geq\delta+a\}}\left(-\frac{2}{a^3}(a-\sigma)+\frac{1}{a^2}\right) - \frac{2}{a^3}(\tau-\delta)\mathbb{1}_{\{\tau\in[\delta,\delta+a),\quad \tau+\sigma\in[0,\delta+a)\}}\right). \tag{4.114}$$

Now suppose $\tilde{\phi}_\star$ is a kernel density such that $\tilde{\phi}_\star \neq \tilde{\phi}$. We do not make assumptions of the parametric family of $\tilde{\phi}_\star$.

Proposition 4.6.15 (Correlation with a general density)**.** *For lags $\tau, \sigma \geq 0$, the cross-correlation Υ is*

$$\tilde{\Upsilon}(\tau,\sigma) = \begin{cases} 0 & \text{if } \tau\in[0,\delta), \\ \frac{1}{a}\Big(\tilde{\psi}_\star(\tau+\sigma) - \tilde{\psi}_\star(\delta+\sigma)\Big) & \text{if } \tau\in[\delta,\delta+a), \\ \frac{1}{a}\Big(\tilde{\psi}_\star(\delta+a+\sigma) - \tilde{\psi}_\star(\delta+\sigma)\Big) & \text{if } \tau\geq\delta+a, \end{cases} \tag{4.115}$$

the derivative of $\tilde{\Upsilon}(\tau,\sigma)$ with respect to the lower bound parameter δ is

$$\frac{\partial\tilde{\Upsilon}}{\partial\delta}(\tau,\sigma) = \begin{cases} 0 & \text{if } \tau\in[0,\delta), \\ -\frac{1}{a}\tilde{\phi}_\star(\delta+\sigma) & \text{if } \tau\in[\delta,\delta+a), \\ \frac{1}{a}\Big(\tilde{\phi}_\star(\delta+a+\sigma) - \tilde{\phi}_\star(\delta+\sigma)\Big) & \text{if } \tau\geq\delta+a, \end{cases} \tag{4.116}$$

and the derivative of $\tilde{\Upsilon}(\tau,\sigma)$ with respect to the interval size parameter a is

$$\frac{\partial\tilde{\Upsilon}}{\partial a}(\tau,\sigma) = \begin{cases} 0 & \text{if } \tau\in[0,\delta), \\ -\frac{1}{a^2}\Big(\tilde{\psi}_\star(\tau+\sigma) - \tilde{\psi}_\star(\delta+\sigma)\Big) & \text{if } \tau\in[\delta,\delta+a), \\ \frac{1}{a}\tilde{\phi}_\star(\delta+a+\sigma) - \frac{1}{a^2}\Big(\tilde{\psi}_\star(\delta+a+\sigma) - \tilde{\psi}_\star(\delta+\sigma)\Big) & \text{if } \tau\geq\delta+a. \end{cases} \tag{4.117}$$

For numerical efficiency, we rewrite this derivative as

$$\frac{\partial \tilde{\Upsilon}}{\partial a}(\tau,\sigma) = \mathbb{1}_{\{\tau \geq \delta + a\}} \frac{1}{a} \tilde{\phi}_\star(\delta + a + \sigma) - \frac{1}{a} \tilde{\Upsilon}(\tau,\sigma) \tag{4.118}$$

Note that for $\tau \geq \delta + a$, $\tilde{\Upsilon}(\tau,\sigma)$ does not depend on τ. In particular, this implies that for lags $\sigma \geq 0$

$$\tilde{\Upsilon}(+\infty,\sigma) = \frac{1}{a}\Big(\tilde{\psi}_\star(\delta + a + \sigma) - \tilde{\psi}_\star(\delta + \sigma)\Big). \tag{4.119}$$

Proposition 4.6.16 (Reverse correlation with another kernel density)**.** *For lags $\tau, \sigma \geq 0$, the reverse cross-correlation $\Upsilon_\star$ is*

$$\tilde{\Upsilon}_\star(\tau,\sigma) = \begin{cases} \frac{1}{a}\mathbb{1}_{\{\tau \geq \delta - \sigma\}} \left(\tilde{\psi}_\star\Big(\min(\tau, \delta + a + \sigma)\Big) - \tilde{\psi}_\star(\delta - \sigma) \right) & \text{if } \sigma \in [0, \delta), \\ \frac{1}{a}\tilde{\psi}_\star\Big(\min(\tau, \delta + a + \sigma)\Big) & \text{if } \sigma \in [\delta, \delta + a), \\ 0 & \text{if } \sigma \geq \delta + a. \end{cases} \tag{4.120}$$

4.6.3.2 Triangular

Triangular densities are more seldom in the Hawkes literature. Like uniform densities, their correlation functionals are inexpensive to evaluate, and available in closed-form for general kernels.

Definition 4.6.3 (Triangular kernel density)**.** *For lags $\tau \geq 0$, the triangular kernel density is*

$$\tilde{\phi}(\tau) := 2\frac{\tau - \alpha}{\beta(\beta + \delta)} \mathbb{1}_{\{\tau \in [\alpha, \alpha + \beta)\}} - 2\frac{\tau - (\alpha + \beta + \delta)}{\delta(\beta + \delta)} \mathbb{1}_{\{\tau \in [\alpha + \beta, \alpha + \beta + \delta)\}}, \tag{4.121}$$

and the parameters are the left corner $\alpha \in (0, +\infty)$, the distance to the altitude foot $\beta \in (0, +\infty)$, and the distance between the altitude foot and the right corner $\delta \in (0, +\infty)$.

Figure 4.2 plots a general triangular density to illustrate our parameterization. Note that for numerical purposes, the following formula might be more convenient to use. For lags $\tau \geq 0$,

$$\tilde{\phi}(\tau) = \frac{2}{\beta + \delta} \left(\frac{\tau - \alpha}{\beta} \mathbb{1}_{\{(\tau - \alpha) \in [0, \beta)\}} + \Big(1 + \frac{\beta}{\delta} - \frac{\tau - \alpha}{\delta}\Big) \mathbb{1}_{\{(\tau - \alpha) \in [\beta, \beta + \delta)\}} \right). \tag{4.122}$$

For more concise equations, denote the corners of the triangle by $c_1 := \alpha$, $c_2 := \alpha + \beta$, and $c_3 := \alpha + \beta + \delta$.

Proposition 4.6.17 (Triangular kernel density derivatives)**.** *For lags $\tau \geq 0$, the derivatives of $\tilde{\phi}$ with respect to model parameters are*

$$\frac{\partial \tilde{\phi}}{\partial \alpha}(\tau) = \frac{-2}{\beta + \delta} \left(\frac{1}{\beta} \mathbb{1}_{\{\tau \in [c_1, c_2)\}} - \frac{1}{\delta} \mathbb{1}_{\{\tau \in [c_2, c_3)\}} \right), \tag{4.123}$$

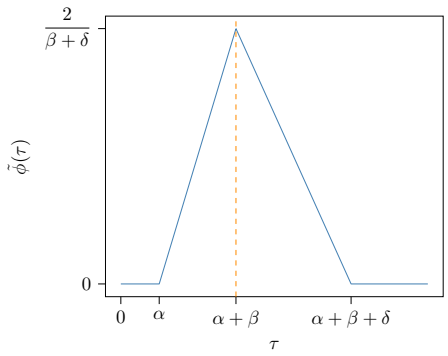


Figure 4.2: Plot of a triangular kernel density against the lag τ.

$$\frac{\partial\tilde{\phi}}{\partial\beta}(\tau) = \frac{-2}{(\beta+\delta)^2}\left(\left(2+\frac{\delta}{\beta}\right)\frac{\tau-\alpha}{\beta}\mathbb{1}_{\{\tau\in[c_1,c_2)\}} - \frac{\tau-\alpha}{\delta}\mathbb{1}_{\{\tau\in[c_2,c_3)\}}\right), \tag{4.124}$$

$$\frac{\partial\tilde{\phi}}{\partial\delta}(\tau) = \frac{-2}{(\beta+\delta)^2}\left(\frac{\tau-\alpha}{\beta}\mathbb{1}_{\{\tau\in[c_1,c_2)\}} + \left(\left(1+\frac{\beta}{\delta}\right)^2 - \left(2+\frac{\beta}{\delta}\right)\frac{\tau-\alpha}{\delta}\right)\mathbb{1}_{\{\tau\in[c_2,c_3)\}}\right), \tag{4.125}$$

We get the CDF associated to this kernel density below.

Proposition 4.6.18 (CDF)**.** *For lags $\tau \geq 0$, the triangular CDF is*

$$\tilde{\psi}(\tau) = \begin{cases} 0 & \text{if } \tau \in [0, c_1), \\ \frac{(\tau-\alpha)^2}{\beta(\beta+\delta)} & \text{if } \tau \in [c_1, c_2), \\ 1 - \frac{(\tau-(\alpha+\beta+\delta))^2}{\delta(\beta+\delta)} & \text{if } \tau \in [c_2, c_3), \\ 1 & \text{if } \tau \geq c_3, \end{cases} \tag{4.126}$$

and the derivatives of $\tilde{\psi}$ with respect to model parameters are

$$\begin{aligned}
\frac{\partial\tilde{\psi}}{\partial\alpha}(\tau) &= -\frac{2}{\beta+\delta}\left(\frac{\tau-\alpha}{\beta}\mathbb{1}_{\{\tau\in[c_1,c_2)\}} + \Big(1+\frac{\beta}{\delta}-\frac{\tau-\alpha}{\delta}\Big)\mathbb{1}_{\{\tau\in[c_2,c_3)\}}\right) = -\tilde{\phi}, \\
\frac{\partial\tilde{\psi}}{\partial\beta}(\tau) &= -\frac{1}{(\beta+\delta)^2}\Bigg((2\beta+\delta)\Big(\frac{\tau-\alpha}{\beta}\Big)^2\mathbb{1}_{\{\tau\in[c_1,c_2)\}} \\
&\qquad + \delta\Big(\Big(1+\frac{\beta}{\delta}\Big)^2 - \Big(\frac{\tau-\alpha}{\delta}\Big)^2\Big)\mathbb{1}_{\{\tau\in[c_2,c_3)\}}\Bigg), \\
\frac{\partial\tilde{\psi}}{\partial\delta}(\tau) &= -\frac{1}{(\beta+\delta)^2}\Bigg(\delta\left(\beta\Big(1+\frac{\beta}{\delta}\Big)\Big(\frac{\tau-\alpha}{\delta} - (1+\frac{\beta}{\delta})\Big) - (\beta+2\delta)\Big(\frac{\tau-\alpha}{\delta}\Big)^2\right)\mathbb{1}_{\{\tau\in[c_2,c_3)\}} \\
&\qquad + \beta\Big(\frac{\tau-\alpha}{\beta}\Big)^2\mathbb{1}_{\{\tau\in[c_1,c_2)\}}\Bigg).
\end{aligned} \tag{4.127}$$

Note that if U is a uniform random variable on $[\,0,1\,]$, then the random offset

$$\tilde{\tau} := \begin{cases} \alpha + (\beta+\delta)\sqrt{\frac{\beta}{\beta+\delta}U} & \text{if } U < \frac{\beta}{\beta+\delta}, \\ \alpha + (\beta+\delta)\Big(1 - \sqrt{\frac{\delta}{\beta+\delta}(1-U)}\Big) & \text{otherwise,} \end{cases} \tag{4.128}$$

is distributed following $\tilde{\phi}$. The moments of $\tilde{\tau}$ are $\mathbb{E}[\tilde{\tau}] = \alpha + \frac{2\beta+\delta}{3}$, $\mathrm{Var}[\tilde{\tau}] = \frac{\beta^2+\delta^2+\beta\delta}{18}$. We now give closed-form expressions for the cross-correlation of triangular kernel densities with general kernels. Suppose $\tilde{\phi}_\star$ is another kernel density such that $\tilde{\phi}_\star \neq \tilde{\phi}$. We do not make assumptions on the parametric family of $\tilde{\phi}_\star$.

Proposition 4.6.19 (Correlation with a general density)**.** *For lags $\tau, \sigma \geq 0$, the cross-correlation Υ is*

$$\tilde{\Upsilon}(\tau,\sigma) = \begin{cases} 0 & \text{if } \tau \in [0, c_1), \\ \frac{2}{\beta+\delta}\Big(\frac{\tau-\alpha}{\beta}\tilde{\psi}_\star(\tau+\sigma) - \frac{1}{\beta}\int_{[c_1+\sigma,\tau+\sigma]}\tilde{\psi}_\star\Big) & \text{if } \tau \in [c_1, c_2), \\ \frac{2}{\beta+\delta}\Big(\frac{c_3-\tau}{\delta}\tilde{\psi}_\star(\tau+\sigma) + \frac{1}{\delta}\int_{[c_2+\sigma,\tau+\sigma]}\tilde{\psi}_\star - \frac{1}{\beta}\int_{[c_1+\sigma,c_2+\sigma]}\tilde{\psi}_\star\Big) & \text{if } \tau \in [c_2, c_3), \\ \frac{2}{\beta+\delta}\Big(\frac{1}{\delta}\int_{[c_2+\sigma,c_3+\sigma]}\tilde{\psi}_\star - \frac{1}{\beta}\int_{[c_1+\sigma,c_2+\sigma]}\tilde{\psi}_\star\Big) & \text{if } \tau \geq c_3, \end{cases} \tag{4.129}$$

and the derivative of $\tilde{\Upsilon}(\tau,\sigma)$ with respect to the left corner parameter α is

$$\frac{\partial\tilde{\Upsilon}}{\partial\alpha}(\tau,\sigma) = \begin{cases} 0 & \text{if } \tau \in [0, c_1), \\ \frac{2}{\beta(\beta+\delta)}\Big(\tilde{\psi}_\star(\alpha+\sigma) - \tilde{\psi}_\star(\tau+\sigma)\Big) & \text{if } \tau \in [c_1, c_2), \\ \frac{2}{\beta(\beta+\delta)}\Big(\frac{\beta}{\delta}\tilde{\psi}_\star(\tau+\sigma) - (1+\frac{\beta}{\delta})\tilde{\psi}_\star(\alpha+\beta+\sigma) + \tilde{\psi}_\star(\alpha+\sigma)\Big) & \text{if } \tau \in [c_2, c_3), \\ \frac{2}{\beta(\beta+\delta)}\Big(\frac{\beta}{\delta}\tilde{\psi}_\star(c_3+\sigma) - (1+\frac{\beta}{\delta})\tilde{\psi}_\star(c_2+\sigma) + \tilde{\psi}_\star(c_1+\sigma)\Big) & \text{if } \tau \geq c_3. \end{cases} \tag{4.130}$$

4.6.3.3 Gaussian

Gaussian mixture models are commonly used in density estimation for their analytical properties, for instance in Zhou et al. [112]. In this section, we discuss kernel densities which are PDF of Gaussian random variables truncated on non-negative values. Denote by $f_{\mathcal{N}}$ (resp. $F_{\mathcal{N}}$) the PDF (resp. CDF) of a standard normal random variable. Denote by $L_{\mathcal{N}}$ the log-derivative of $F_{\mathcal{N}}$, that is $L_{\mathcal{N}} := \frac{f_{\mathcal{N}}}{F_{\mathcal{N}}}$. Note that on $[0,+\infty)$, this function is positive, decreasing, and convex, with $L_{\mathcal{N}}(0) = \sqrt{2/\pi}$, and $\lim_{x\to+\infty} L_{\mathcal{N}}(x) = 0$.

Definition 4.6.4 (Gaussian kernel density)**.** *For lags $\tau \geq 0$, the Gaussian kernel density is*

$$\tilde{\phi}(\tau) := \frac{1}{F_{\mathcal{N}}(\delta/\beta)\sqrt{2\pi\beta^2}}\exp\left(-\frac{1}{2}\left(\frac{\tau-\delta}{\beta}\right)^2\right) = \frac{1}{\beta F_{\mathcal{N}}(\delta/\beta)}f_{\mathcal{N}}\left(\frac{\tau-\delta}{\beta}\right),$$

and the parameters are the location $\delta \in (0,+\infty)$, and the scale $\beta \in (0,+\infty)$.

The Gaussian kernel density degenerates to the null kernel density for $\beta \to +\infty$; and to the Dirac distribution in δ for $\beta \downarrow 0$. This implies that mixtures of Gaussian kernel densities are dense for the weak topology.

Proposition 4.6.20 (Kernel derivatives)**.** *For lags $\tau \geq 0$, the derivatives of $\log \tilde{\phi}$ with respect to model parameters are*

$$\frac{\partial \log \tilde{\phi}}{\partial \beta}(\tau) = \frac{1}{\beta}\left(\left(\frac{\tau - \delta}{\beta}\right)^2 + \frac{\delta}{\beta} L_{\mathcal{N}}(\delta/\beta) - 1\right), \qquad \frac{\partial \log \tilde{\phi}}{\partial \delta}(\tau) = \frac{1}{\beta}\left(\frac{\tau - \delta}{\beta} - L_{\mathcal{N}}(\delta/\beta)\right). \tag{4.131}$$

We now compute the CDF of the Gaussian kernel density and its derivatives with respect to model parameters.

Proposition 4.6.21 (CDF)**.** *For lags $\tau \geq 0$, the CDF of the Gaussian kernel density $\tilde{\phi}$ is*

$$\tilde{\psi}(\tau) = 1 - \frac{1 - F_{\mathcal{N}}\left(\frac{\tau - \delta}{\beta}\right)}{F_{\mathcal{N}}(\delta/\beta)}, \tag{4.132}$$

and the derivatives of $\tilde{\psi}$ with respect to model parameters are

$$\frac{\partial \tilde{\psi}}{\partial \beta}(\tau) = \frac{-1}{\beta^2 F_{\mathcal{N}}(\delta/\beta)}\left((\tau - \delta) f_{\mathcal{N}}\left(\frac{\tau - \delta}{\beta}\right) + \delta \frac{f_{\mathcal{N}}(\delta/\beta)}{F_{\mathcal{N}}(\delta/\beta)}\left(1 - F_{\mathcal{N}}\left(\frac{\tau - \delta}{\beta}\right)\right)\right), \tag{4.133}$$

$$\frac{\partial \tilde{\psi}}{\partial \delta}(\tau) = \frac{-1}{\beta F_{\mathcal{N}}(\delta/\beta)}\left(f_{\mathcal{N}}\left(\frac{\tau - \delta}{\beta}\right) - \frac{f_{\mathcal{N}}(\delta/\beta)}{F_{\mathcal{N}}(\delta/\beta)}\left(1 - F_{\mathcal{N}}\left(\frac{\tau - \delta}{\beta}\right)\right)\right). \tag{4.134}$$

If U is a uniform random variable on $[\,0, 1\,]$, then the random offset $\tilde{\tau} := \delta + \beta F_{\mathcal{N}}^{-1}\Big(1 - F_{\mathcal{N}}(\delta/\beta)U\Big)$ is distributed following $\tilde{\phi}$. In practice, rejection sampling is a more efficient simulation method for truncated Gaussian random variables. While the Gaussian kernel density $\tilde{\phi}$ does indeed have a unique mode in δ, the parameters δ, β do not correspond to the moments of the associated random offset. In fact, a random offset with PDF $\tilde{\phi}$ has expectation

$$\bar{\delta} := \delta + \beta L_{\mathcal{N}}(\delta/\beta), \tag{4.135}$$

which is strictly greater than δ, and variance

$$\bar{\beta}^2 := \beta^2\left(1 - \frac{\delta}{\beta} L_{\mathcal{N}}(\delta/\beta) - \Big(L_{\mathcal{N}}(\delta/\beta)\Big)^2\right), \tag{4.136}$$

which is strictly less than β^2. Therefore, from a modelling standpoint, we need to be careful about how the parameters (δ, β) relate to the statistical features of the random offset. For example, if the initial Gaussian is centered, that is $\delta = 0$,

$$\bar{\delta} = \sqrt{\frac{2}{\pi}}\beta, \qquad \bar{\beta} = \sqrt{1 - \frac{2}{\pi}}\beta. \tag{4.137}$$

In this centered case, the standard deviation to mean ratio $\frac{\bar{\beta}}{\bar{\delta}}$ is constant

$$\frac{\bar{\beta}}{\bar{\delta}} = \sqrt{\frac{\pi}{2} - 1} \simeq 75\%. \tag{4.138}$$

However, both the mean $\bar{\delta}$ and the standard deviation $\bar{\beta}$ of the truncated Gaussian go to $+\infty$ when the standard deviation of the initial Gaussian β goes to $+\infty$. In fact, the location $\delta = 0$ lies approximately 1.3 standard deviations $\bar{\beta}$ away from the truncated Gaussian mean $\bar{\delta}$, and as we see with the formula of the CDF $\tilde{\psi}$ below, we even have $\mathbb{P}(\tilde{\tau} \leq \delta) = 0$. More generally, if $\delta > 0$, denote by α the original standard deviation to mean ratio $\alpha := \frac{\beta}{\delta}$. The moment shift ratios $\frac{\bar{\delta}}{\delta}$ and $\frac{\bar{\beta}}{\beta}$ only depend on α, with

$$\frac{\bar{\delta}}{\delta} = 1 + \alpha L_{\mathcal{N}}\Big(\frac{1}{\alpha}\Big), \qquad \frac{\bar{\beta}}{\beta} = \sqrt{1 - \frac{1}{\alpha} L_{\mathcal{N}}\Big(\frac{1}{\alpha}\Big) - L_{\mathcal{N}}^2\Big(\frac{1}{\alpha}\Big)}. \tag{4.139}$$

The mean shift ratio $\frac{\bar{\delta}}{\delta}$ is an increasing function of the original standard deviation to mean ratio α, with $\frac{\bar{\delta}}{\delta}(\alpha = 0) = 1$, and $\lim_{\alpha \to +\infty} \frac{\bar{\delta}}{\delta} = +\infty$. The standard deviation shift ratio $\frac{\bar{\beta}}{\beta}$ is a decreasing function of the original variance to mean ratio α, with $\frac{\bar{\beta}}{\beta}(\alpha = 0) = 1$, and a non-null limit

$$\lim_{\alpha \to +\infty} \frac{\bar{\beta}}{\beta} = \sqrt{1 - \frac{2}{\pi}} \simeq 60\%. \tag{4.140}$$

We now give closed-form expressions of the kernel correlation function $\tilde{\Upsilon}$ for different

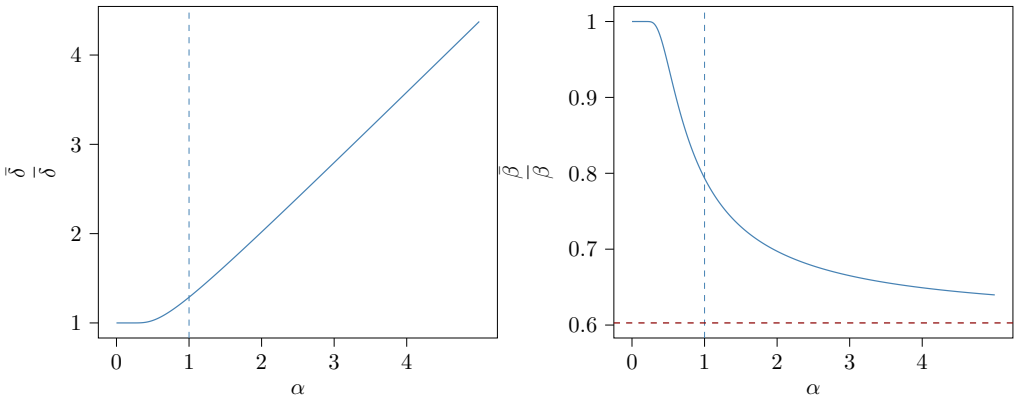


Figure 4.3: Shifted moments of the Gaussian kernel density

Dependence of the shifted moments (mean $\frac{\bar{\delta}}{\delta}$ on the left, standard deviation $\frac{\bar{\beta}}{\beta}$ on the right) on the original standard deviation to mean ratio $\alpha := \frac{\beta}{\delta}$. Red horizontal line plots the asymptote of $\frac{\bar{\beta}}{\beta}$ at $+\infty$, given by $\sqrt{1 - \frac{2}{\pi}}$.

parametric classes of the second kernel density $\tilde{\phi}_\star$. First, suppose $\tilde{\phi}_\star = \tilde{\phi}$, and define the following parameters. Define the re-scaled variance $\tilde{\beta}$ and the dimension-less parameter $\tilde{\delta}$ by

$$\tilde{\beta} := \frac{\beta}{\sqrt{2}}, \qquad \tilde{\delta} := \frac{\delta}{\tilde{\beta}}. \tag{4.141}$$

Define the log-derivative

$$q := L_{\mathcal{N}}(\delta/\beta). \tag{4.142}$$

For lags $\tau, \sigma \geq 0$, define the dimension-less transformations of the time variables

$$\tilde{\sigma} := \frac{\sigma/2}{\tilde{\beta}}, \qquad \tilde{x} := \frac{\tau + \sigma/2}{\tilde{\beta}}. \tag{4.143}$$

Proposition 4.6.22 (Auto-correlation of Gaussian kernel densities)**.** *For lags $\tau, \sigma \geq 0$, the auto-correlation of $\tilde{\phi}$ is*

$$\tilde{\Upsilon}(\tau, \sigma) = \frac{1}{2\tilde{\beta}(F_{\mathcal{N}}(\delta/\beta))^2} f_{\mathcal{N}}(\tilde{\sigma}) \Big(F_{\mathcal{N}}(\tilde{x} - \tilde{\delta}) - F_{\mathcal{N}}(\tilde{\sigma} - \tilde{\delta}) \Big), \tag{4.144}$$

the derivative of $\tilde{\Upsilon}(\tau, \sigma)$ with respect to the location parameter δ is

$$\frac{\partial \tilde{\Upsilon}}{\partial \delta}(\tau, \sigma) = -\frac{f_{\mathcal{N}}(\tilde{\sigma})}{2\tilde{\beta}^2(F_{\mathcal{N}}(\delta/\beta))^2} \left(f_{\mathcal{N}}(\tilde{x} - \tilde{\delta}) - f_{\mathcal{N}}(\tilde{\sigma} - \tilde{\delta}) + \sqrt{2} q \Big(F_{\mathcal{N}}(\tilde{x} - \tilde{\delta}) - F_{\mathcal{N}}(\tilde{\sigma} - \tilde{\delta}) \Big) \right), \tag{4.145}$$

and the derivative of $\tilde{\Upsilon}(\tau, \sigma)$ with respect to the scale parameter β is

$$\begin{aligned} \frac{\partial \tilde{\Upsilon}}{\partial \beta}(\tau, \sigma) = -\frac{f_{\mathcal{N}}(\tilde{\sigma})}{2\sqrt{2}\tilde{\beta}^2(F_{\mathcal{N}}(\delta/\beta))^2} \Bigg(& (\tilde{x} - \tilde{\delta}) f_{\mathcal{N}}(\tilde{x} - \tilde{\delta}) - (\tilde{\sigma} - \tilde{\delta}) f_{\mathcal{N}}(\tilde{\sigma} - \tilde{\delta}) \\ & + (1 - \sqrt{2} q \tilde{\delta} + \tilde{\sigma}^2) \Big(F_{\mathcal{N}}(\tilde{x} - \tilde{\delta}) - F_{\mathcal{N}}(\tilde{\sigma} - \tilde{\delta}) \Big) \Bigg). \end{aligned} \tag{4.146}$$

Suppose $\tilde{\phi}_\star$ is another Gaussian kernel density with parameters $(\delta_\star, \beta_\star)$. Define the log-derivatives

$$q_1 := \frac{f_{\mathcal{N}}(\delta/\beta)}{F_{\mathcal{N}}(\delta/\beta)}, \qquad q_2 := \frac{f_{\mathcal{N}}(\delta_\star/\beta_\star)}{F_{\mathcal{N}}(\delta_\star/\beta_\star)}. \tag{4.147}$$

Define the two following variance parameters: first, the L_2 norm of the vector $(\beta, \beta_\star)$

$$b := \frac{1}{2}\sqrt{\beta^2 + \beta_\star^2}, \tag{4.148}$$

then

$$\tilde{\beta} := \frac{\beta \beta_\star}{\sqrt{\beta^2 + \beta_\star^2}}. \tag{4.149}$$

Define the following dimension-less transformations of kernel parameters

$$b_1 := \frac{\beta}{\sqrt{\beta^2 + \beta_\star^2}}, \qquad b_2 := \frac{\beta_\star}{\sqrt{\beta^2 + \beta_\star^2}}. \tag{4.150}$$

Note that $b_1^2 + b_2^2 = 1$. For lags $\tau, \sigma \geq 0$, define the following dimension-less transformations of the time variables

$$\tilde{D} := \frac{\delta - (\delta_\star - \sigma)}{2b}, \qquad \tilde{\sigma} := -\frac{b_2^2 \delta + b_1^2(\delta_\star - \sigma)}{\tilde{\beta}}, \qquad \tilde{x} := \frac{\tau - \Big(b_2^2 \delta + b_1^2(\delta_\star - \sigma) \Big)}{\tilde{\beta}}. \tag{4.151}$$

Proposition 4.6.23 (Cross-correlation with Gaussian $\tilde{\phi}_\star$)**.** *For lags $\tau, \sigma \geq 0$, the cross-correlation of $\tilde{\phi}$ with $\tilde{\phi}_\star$ is*

$$\tilde{\Upsilon}(\tau,\sigma) = \frac{1}{2bF_{\mathcal{N}}(\delta/\beta)F_{\mathcal{N}}(\delta_\star/\beta_\star)} f_{\mathcal{N}}(\tilde{D})\Big(F_{\mathcal{N}}(\tilde{x}) - F_{\mathcal{N}}(\tilde{\sigma})\Big), \tag{4.152}$$

the derivatives of $\tilde{\Upsilon}(\tau,\sigma)$ with respect to the location parameters are

$$\begin{aligned}
\frac{\partial\tilde{\Upsilon}}{\partial\delta}(\tau,\sigma) &= \frac{-1}{2bF_{\mathcal{N}}(\delta/\beta)F_{\mathcal{N}}(\delta_\star/\beta_\star)} f_{\mathcal{N}}(\tilde{D})\Bigg[\frac{1}{2b}\Big(\frac{q_1}{b_1} + \tilde{D}\Big)\Big(F_{\mathcal{N}}(\tilde{x}) - F_{\mathcal{N}}(\tilde{\sigma})\Big) \\
&\qquad + \frac{b_2^2}{\tilde{\beta}}\Big(f_{\mathcal{N}}(\tilde{x}) - f_{\mathcal{N}}(\tilde{\sigma})\Big)\Bigg], \\
\frac{\partial\tilde{\Upsilon}}{\partial\delta_\star}(\tau,\sigma) &= \frac{-1}{2bF_{\mathcal{N}}(\delta/\beta)F_{\mathcal{N}}(\delta_\star/\beta_\star)} f_{\mathcal{N}}(\tilde{D})\Bigg[\frac{1}{2b}\Big(\frac{q_2}{b_2} - \tilde{D}\Big)\Big(F_{\mathcal{N}}(\tilde{x}) - F_{\mathcal{N}}(\tilde{\sigma})\Big) \\
&\qquad + \frac{b_1^2}{\tilde{\beta}}\Big(f_{\mathcal{N}}(\tilde{x}) - f_{\mathcal{N}}(\tilde{\sigma})\Big)\Bigg],
\end{aligned} \tag{4.153}$$

and the derivatives of $\tilde{\Upsilon}(t,\sigma)$ with respect to the scale parameters are

$$\begin{aligned}
\frac{\partial\tilde{\Upsilon}}{\partial\beta}(\tau,\sigma) &= \frac{-f_{\mathcal{N}}(\tilde{D})}{4b^2F_{\mathcal{N}}(\delta/\beta)F_{\mathcal{N}}(\delta_\star/\beta_\star)}\Bigg[\Bigg(b_1(1-\tilde{D}^2) - \frac{\delta q_1}{2bb_1^2}\Bigg)\Big(F_{\mathcal{N}}(\tilde{x}) - F_{\mathcal{N}}(\tilde{\sigma})\Big) \\
&\qquad + \frac{b_2^2}{b_1}\Bigg(\Big(\tilde{t} - (\tilde{\sigma} + \frac{2\delta}{\beta})\Big)f_{\mathcal{N}}(\tilde{x}) + \Big(\tilde{\sigma} + \frac{2\delta}{\beta}\Big)f_{\mathcal{N}}(\tilde{\sigma})\Bigg)\Bigg],
\end{aligned} \tag{4.154}$$

$$\begin{aligned}
\frac{\partial\tilde{\Upsilon}}{\partial\beta_\star}(\tau,\sigma) &= \frac{-f_{\mathcal{N}}(\tilde{D})}{4b^2F_{\mathcal{N}}(\delta/\beta)F_{\mathcal{N}}(\delta_\star/\beta_\star)}\Bigg[\Big(b_2(1-\tilde{D}^2) - \frac{\delta q_2}{2bb_2^2}\Big)\Big(F_{\mathcal{N}}(\tilde{x}) - F_{\mathcal{N}}(\tilde{\sigma})\Big) \\
&\qquad + \frac{b_1^2}{b_2}\Bigg(\Big(\tilde{x} + \frac{b_2^2}{b_1^2}(\tilde{\sigma} + \frac{\delta}{\beta})\Big)f_{\mathcal{N}}(\tilde{x}) \qquad (4.155) \\
&\qquad - (1-b_2^4)\Big(\tilde{\sigma} + \frac{2b_2^2}{1+b_2^2}\frac{\delta}{\beta}\Big)f_{\mathcal{N}}(\tilde{\sigma})\Bigg)\Bigg].
\end{aligned}$$

4.6.3.4 Gamma

The Gamma kernel is notably used in the seismology literature, originally in the *linlin* model in Hawkes and Oakes [49]. Recall the definition of the Gamma function for all $\alpha > 0$

$$\Gamma(\alpha) := \int_0^{+\infty} t^{\alpha-1}\mathrm{e}^{-t}\mathrm{d}t. \tag{4.156}$$

For $\alpha > 0$ and for $x > 0$, the lower incomplete Gamma function is

$$\gamma(\alpha, x) := \int_0^x t^{\alpha-1}\mathrm{e}^{-t}\mathrm{d}t, \tag{4.157}$$

and the digamma function Ψ_0 is

$$\Psi_0(\alpha) := \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}. \tag{4.158}$$

Definition 4.6.5 (Gamma kernel density)**.** *For lags $\tau \geq 0$, the Gamma kernel density is*

$$\tilde{\phi}(\tau) = \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} \mathrm{e}^{-\beta\tau}, \tag{4.159}$$

and the parameters are the shape $\alpha \in (1, +\infty)$, and the rate $\beta \in (0, +\infty)$.

Proposition 4.6.24 (Kernel derivatives)**.** *For lags $\tau \geq 0$, the derivatives of $\log \tilde{\phi}$ with respect to model parameters are*

$$\frac{\partial \log \tilde{\phi}}{\partial \alpha}(\tau) = \log(\beta\tau) - \Psi_0(\alpha), \qquad \frac{\partial \log \tilde{\phi}}{\partial \beta}(\tau) = \frac{\alpha}{\beta} - \tau. \tag{4.160}$$

We get the CDF associated to this kernel density below.

Proposition 4.6.25 (CDF)**.** *For lags $\tau \geq 0$, the Gamma CDF is*

$$\tilde{\psi}(\tau) = \frac{\gamma(\alpha, \beta\tau)}{\Gamma(\alpha)}, \tag{4.161}$$

and the derivatives of $\tilde{\psi}$ with respect to model parameters are

$$\frac{\partial \tilde{\psi}}{\partial \omega}(\tau) = \frac{\gamma(\alpha, \beta\tau)}{\Gamma(\alpha)}, \qquad \frac{\partial \tilde{\psi}}{\partial \beta}(\tau) = \frac{\omega}{\Gamma(\alpha)} \tau^\alpha \beta^{\alpha-1} e^{-\beta\tau}. \tag{4.162}$$

4.6.4 Heavy tailed densities

4.6.4.1 Power law

Heavy tailed densities are instrumental to model systems with long-memory excitation; in this work, we focus on the classic power law kernel density.

Definition 4.6.6 (Power law kernel density)**.** *For lags $\tau \geq 0$, the power law kernel density is*

$$\tilde{\phi}(\tau) := \frac{\alpha\beta}{(1 + \beta\tau)^{1+\alpha}}, \tag{4.163}$$

and the parameters are the scale $\beta \in (0, +\infty)$, and the decay $\alpha \in (0, +\infty)$.

The power law kernel density degenerates to the null kernel density for $\alpha \to 0$ or $\beta \to 0$; and the Dirac distribution in 0 for $\alpha \to +\infty$ or $\beta \to +\infty$. Note that the function $\tau \mapsto 1/(1 + \beta\tau)^{1+\alpha}$ is not integrable for $\alpha = 0$. The derivatives of this kernel density are simply obtained below. We also display its log derivatives for computational efficiency.

Proposition 4.6.26 (Kernel derivatives)**.** *For lags $\tau \geq 0$, the derivatives of $\log \tilde{\phi}$ with respect to model parameters are*

$$\frac{\partial \log \tilde{\phi}}{\partial \alpha}(\tau) = \frac{1 - \alpha \log(1+\beta\tau)}{\alpha}, \qquad \frac{\partial \log \tilde{\phi}}{\partial \beta}(\tau) = \frac{1 - \alpha\beta\tau}{\beta(1+\beta\tau)}. \tag{4.164}$$

We get the CDF associated to this kernel density below.

Proposition 4.6.27 (CDF)**.** *For lags $\tau \geq 0$, the power law CDF is*

$$\tilde{\psi}(\tau) = 1 - \frac{1}{(1+\beta\tau)^{\alpha}}, \tag{4.165}$$

and the derivatives of $\tilde{\psi}$ with respect to model parameters are

$$\frac{\partial \tilde{\psi}}{\partial \alpha}(\tau) = \frac{\log(1+\beta\tau)}{(1+\beta\tau)^{\alpha}}, \qquad \frac{\partial \tilde{\psi}}{\partial \beta}(\tau) = \frac{\alpha\tau}{(1+\beta\tau)^{1+\alpha}}. \tag{4.166}$$

If U is a uniform random variable on $[0,1]$, then the random offset $\tilde{\tau} := \frac{U^{-1/\alpha}-1}{\beta}$ has density $\tilde{\phi}$. We now discuss the kernel correlation function $\tilde{\Upsilon}$ between $\tilde{\phi}$ and a second power law kernel density $\tilde{\phi}_\star$, with parameters $(\alpha_\star, \beta_\star)$. It is not clear whether we can derive a closed-form expression of $\tilde{\Upsilon}$ for general values of the parameters $(\alpha, \beta, \alpha_\star, \beta_\star)$. While this issue motivates the MC approximation of model functionals we develop in Section 4.4.5, we discuss here a configuration where we can obtain a closed-form expression for $\tilde{\Upsilon}$ and its derivatives in order to evaluate our approximations with synthetic data in Section 5.1. First, assume $\tilde{\phi}_\star = \tilde{\phi}$, that is, $\tilde{\Upsilon}$ is the auto-correlation function of $\tilde{\phi}$. Second, fix the exponent $\alpha = 1$. To avoid confusion, we refer to this kernel density $\tilde{\phi}$ as the squared power law kernel density. For lags $\sigma > 0$, define the dimension-less variables

$$\tilde{\sigma}_1 = \frac{1}{\beta^2\sigma^2} + \frac{2}{\beta^3\sigma^3}, \qquad \tilde{\sigma}_2 = \frac{1}{\beta^2\sigma^2} - \frac{2}{\beta^3\sigma^3}. \tag{4.167}$$

Proposition 4.6.28 (Auto-correlation)**.** *For lags $\tau, \sigma > 0$, the auto-correlation of the squared power law kernel densities is*

$$\begin{aligned}\tilde{\Upsilon}(\tau,\sigma) = {} & \frac{1}{\beta\sigma^2}\Big(1 - \frac{1}{1+\beta(\tau+\sigma)} - \frac{1}{1+\beta\tau} + \frac{1}{1+\beta\sigma}\Big) \\ & + \frac{2}{\beta^2 s^3}\Big(\log\big(1+\beta(\tau+\sigma)\big) - \log\big(1+\beta\tau\big) - \log\big(1+\beta\sigma\big)\Big),\end{aligned} \tag{4.168}$$

and the derivative of $\tilde{\Upsilon}(\tau,\sigma)$ with respect to the scale parameter β is

$$\begin{aligned}\frac{\partial \tilde{\Upsilon}}{\partial \beta}(\tau,\sigma) = {} & -\tilde{\sigma}_1 + \frac{\tilde{\sigma}_1}{1+\beta\tau} - \frac{\tilde{\sigma}_2}{1+\beta\sigma} + \frac{\tilde{\sigma}_2}{1+\beta(\tau+\sigma)} \\ & + \frac{1}{\beta\sigma^2}\Big(\frac{\tau}{(1+\beta\tau)^2} - \frac{\sigma}{(1+\beta\sigma)^2} + \frac{\tau+\sigma}{\big(1+\beta(\tau+\sigma)\big)^2}\Big) \\ & - \frac{4}{\beta^3\sigma^3}\Big(\log\big(1+\beta(\tau+\sigma)\big) - \log\big(1+\beta\tau\big) - \log\big(1+\beta\sigma\big)\Big).\end{aligned} \tag{4.169}$$

For null lags, the auto-correlation of $\tilde{\phi}$ is

$$\tilde{\Upsilon}(\tau, 0) = \frac{\beta}{3}\Big(1 - \frac{1}{(1+\beta\tau^3)}\Big), \tag{4.170}$$

and the derivative of $\tilde{\Upsilon}(\tau, 0)$ with respect to the scale parameter β is

$$\frac{\partial\tilde{\Upsilon}}{\partial\beta}(\tau, 0) = \frac{1}{3}\Big(1 - \frac{1}{(1+\beta\tau)^3}\Big) + \frac{\beta\tau}{(1+\beta\tau)^4}, \tag{4.171}$$

4.6.4.2 Truncated power law

We introduce the truncated power law kernel to avoid the non-integrability of functions$x \mapsto x^{-1-\alpha}$ at 0 for $\alpha > 0$, and to retain the long memory propriety of power laws while constraining the kernel on a finite support.

Definition 4.6.7 (Truncated power law kernel)**.** *For lags $\tau \geq 0$, the truncated power law kernel is*

$$\tilde{\phi}(x) := \frac{\alpha\beta}{x^{1+\alpha}}\mathbb{1}_{x\in[\delta_L,\delta_R)},$$

and the parameters are the exponent $\alpha \in (0, +\infty)$, the delay $\delta_L \in (0, +\infty)$, and the truncation threshold $\delta_R \in (0, +\infty)$. To simplify notation, define

$$\beta := \frac{\delta_L^\alpha \delta_R^\alpha}{\delta_R^\alpha - \delta_L^\alpha}, \qquad b := \beta^{1/\alpha}, \qquad \beta_R := \frac{\delta_R^\alpha}{\delta_R^\alpha - \delta_L^\alpha}. \tag{4.172}$$

We get the derivatives of this kernel simply.

Proposition 4.6.29 (ϕ derivatives)**.** *For lags $\tau \geq 0$, the derivative of $\log\tilde{\phi}$ with respect to the exponent α is*

$$\frac{\partial\log\tilde{\phi}}{\partial\alpha}(\tau) = \left(\frac{1}{\alpha} + \beta_R\log\delta_L - \beta_L\log\delta_R - \log\tau\right)\mathbb{1}_{\tau\in[\delta_L,\delta_R)}. \tag{4.173}$$

We get the CDF associated to this kernel density below.

Proposition 4.6.30 (CDF)**.** *For lags $\tau \geq 0$, the truncated power law CDF is*

$$\tilde{\psi}(\tau) = \beta_R\left(1 - \Big(\frac{\delta_L}{\tau}\Big)^\alpha\right)\mathbb{1}_{\tau\in[\delta_L,\delta_R)} + \mathbb{1}_{\tau\geq\delta_R}, \tag{4.174}$$

and the derivatives of $\tilde{\psi}$ with respect to model parameters are

$$\frac{\partial\tilde{\psi}}{\partial\alpha}(\tau) = \beta_R\left(\Big(\beta_L\log\frac{\delta_R}{\delta_L} + \log\frac{t}{\delta_L}\Big)\Big(\frac{\delta_L}{t}\Big)^\alpha - \beta_L\log\frac{\delta_R}{\delta_L}\right). \tag{4.175}$$

Note that if U is a uniform random variable on $[\,0, 1\,]$, then the random offset $\tilde{\tau} := \frac{b}{(\beta_R - U)^{1/\alpha}}$ is distributed following $\tilde{\phi}$.

4.6.5 Delaying kernels

In nature, there are systems generating streams of data with non-monotonic kernels; for example, when events might trigger each other with some delay. In this section, we discuss families of kernel densities obtained by adding a delay to a given kernel density. Formally, let $\tilde{\phi}$ by a kernel density, and denote by θ a parameter of $\tilde{\phi}$. Consider a delay parameter $\delta \geq 0$. For lags $\tau \geq 0$, define the delayed kernel density $\tilde{\phi}^{(D)}$ by

$$\tilde{\phi}^{(D)}(\tau) := \tilde{\phi}(\tau - \delta)\mathbb{1}_{\{\tau \geq \delta\}}. \tag{4.176}$$

In the rest of the section, we use the superscript (D) to denote functionals related to the delayed kernel density $\tilde{\phi}^{(D)}$. For lags $\tau \geq 0$, the delayed CDF is

$$\tilde{\psi}^{(D)}(\tau) = \tilde{\psi}(\tau - \delta)\mathbb{1}_{\{\tau \geq \delta\}}, \tag{4.177}$$

and the derivatives of the delayed CDF with respect to model parameters are

$$\frac{\partial \tilde{\psi}^{(D)}}{\partial \theta}(\tau) = \frac{\partial \tilde{\psi}}{\partial \theta}(\tau - \delta)\mathbb{1}_{\{\tau \geq \delta\}}, \qquad \frac{\partial \tilde{\psi}^{(D)}}{\partial \delta}(\tau) = -\tilde{\phi}^{(D)}(\tau). \tag{4.178}$$

Second, we express the correlation functionals of the delayed kernel in terms of the correlation functionals of the original kernel.

Proposition 4.6.31 (Auto-correlation functional)**.** *For lags $\tau, \sigma \geq 0$, the auto-correlation function $\tilde{\Upsilon}^{(D)}$ of the delayed kernel density $\tilde{\phi}^{(D)}$ is*

$$\tilde{\Upsilon}^{(D)}(\tau, \sigma) = \tilde{\Upsilon}\big(\tau - \delta, \sigma\big)\mathbb{1}_{\{\tau \geq \delta\}}, \tag{4.179}$$

And the derivatives of $\tilde{\Upsilon}^{(D)}$ with respect to model parameters are

$$\frac{\partial \tilde{\Upsilon}^{(D)}}{\partial \theta}(\tau, \sigma) = \frac{\partial \tilde{\Upsilon}}{\partial \theta}\big(\tau - \delta, \sigma\big)\mathbb{1}_{\{\tau \geq \delta\}}, \qquad \frac{\partial \tilde{\Upsilon}^{(D)}}{\partial \delta}(\tau, \sigma) = -\tilde{\phi}(\tau - \delta)\tilde{\phi}(\tau - \delta + \sigma)\mathbb{1}_{\{\tau \geq \delta\}}. \tag{4.180}$$

As we expect, there is no auto-correlation up to lags $\tau < \delta$. For lags $\tau \geq \delta$, delaying the kernel density shifts the sampling window from $[0, \tau]$ to $[0, \tau - \delta]$, but does not affect the dependence in the lag σ. In least-squares estimation of Hawkes models, the kernel correlation $\tilde{\Upsilon}^{(D)}$ is evaluated on pairs $(T - t_m^i, t_m^i - t_n^j)$, for event types $i, j \in [d]$ and event timings $t_n^j < t_m^i$. We see that

$$\tilde{\Upsilon}^{(D)}(T - t_m^i, t_m^i - t_n^j) = \tilde{\Upsilon}\Big(T - (t_m^i + \delta), (t_m^i + \delta) - (t_n^j + \delta)\Big)\mathbb{1}_{\{t_m^i + \delta \leq T\}}. \tag{4.181}$$

This is consistent with the fact that the LSE of a delayed Hawkes model on a data path is equal to the LSE of the non-delayed Hawkes model on the data path where all jump times are shifted by δ, and where we exclude the shifted jump times larger than the observation

window T. We discuss this property in more detail in the motivational paragraph below. Before that, we link the cross-correlation functional $\tilde{\Upsilon}^{(D)}$ of the delayed kernel with other kernel densities, with that of the original kernel $\tilde{\Upsilon}$. Consider a second kernel density $\tilde{\phi}_\star$, and a delay parameter $\delta_\star \geq 0$. Denote by $\tilde{\phi}_\star^{(D)}$ the associated delayed kernel density. For lags $\sigma \geq 0$, define

$$D(\sigma) := \max(\delta, \delta_\star - \sigma). \tag{4.182}$$

Proposition 4.6.32 (Correlation with other kernel densities)**.** *For $\tau, \sigma \geq 0$, the correlation function $\tilde{\Upsilon}^{(D)}$ between the delayed kernel densities $\tilde{\phi}^{(D)}$ and $\tilde{\phi}_\star^{(D)}$ is*

$$\tilde{\Upsilon}^{(D)}(\tau, \sigma) = \begin{cases} \tilde{\Upsilon}\Big(\tau - D(\sigma), |\delta - (\delta_\star - \sigma)|\Big) \mathbb{1}_{\{\tau \geq D(\sigma)\}} & \text{if } \delta \geq \delta_\star - \sigma, \\ \tilde{\Upsilon}_\star\Big(\tau - D(\sigma), |\delta - (\delta_\star - \sigma)|\Big) \mathbb{1}_{\{\tau \geq D(\sigma)\}} & \text{otherwise.} \end{cases} \tag{4.183}$$

Chapter 5

Numerical experiments

5.1 Synthetic data

In this section, we evaluate `ASLSD` on data simulated with an exact cluster based algorithm, for a variety of data generating processes and model classes including both MHP in Section 5.1.1 and MTLH in Section 5.1.2 [1].

5.1.1 MHP Data

5.1.1.1 Exponential

The estimation of exponential MHP is a well studied problem that we discuss here as a basic verification case, in the long path setup.

Uni-dimensional We start with the classic case of the uni-dimensional MHP with one exponential kernel. Consider parameters

$$\mu_1^\diamond = 1.55, \quad \omega_{11}^\diamond = 0.45, \quad \beta_{11}^\diamond = 1.95. \tag{5.1}$$

We simulate a path of this MHP up to $T' = 10^5$, resulting in 283 099 events. For illustration purposes, first consider a uni-dimensial MHP model with an SBF exponential kernel, *i.e.* fix the model decay rate $\beta_{11} = \beta_{11}^\diamond$. Using the exponential Markovianity trick, we compute the exact LSE in linear time for given parameter values. Figure 5.1 plots the contour map of the model's LSE on the simulated path, in the (μ_1, ω_{11}) plane, and paths of the `ASLSD` procedure for different solvers using the same initial point. Fitting the decay rate of exponential Hawkes models is a difficult problem; Figure 5.2 shows that `ASLSD` yields satisfactory results. We train the model on $[0, T]$ with $T = 0.75T'$, and test it OOS on $[T, T']$. Outside of the uni-dimensional SBF MHP case, it is difficult to visualize the trajectory of

[1] To reproduce the results in this section, see the code in `https://github.com/saadlabyad/aslsd/experiments/`.

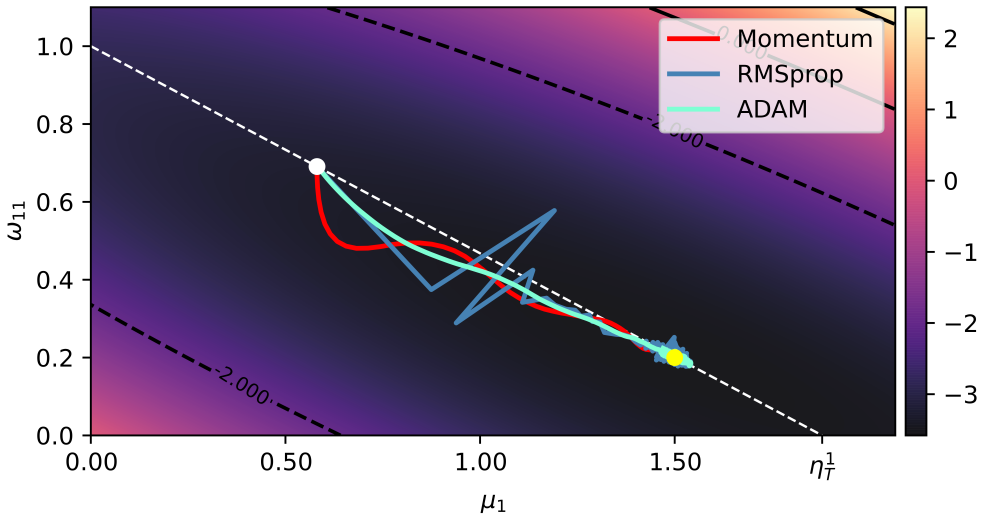


Figure 5.1: Contour plot of the LSE

Contour plot of the exact LSE in the (μ_1, ω_{11}) plane. White segment plots the first order feasible set $\mathbb{F}_{\text{MHP}}(\boldsymbol{\eta}_T)$. White circle is the random initial point, yellow circle plots ground truth parameters, and solid lines are the trajectories of parameter estimates for `ASLSD` with different solvers.

an estimation method without projections. However, Figure 5.3 plots the exact train and test LSE for the parameters at each gradient iteration, which shows a satisfactory decay. To analyze the solver path at the level of each parameter, Figure 5.4 plots the different parameter updates and gradient estimates at each iteration. Again, we see a satisfactory convergence towards the ground truth, despite noisier gradient paths for the decay rate β_{11} Figure 5.5 gives a closer look at the distribution of our gradient estimator evaluated at model parameters ($\mu_1 = 3, \quad \omega_{11} = 0.7, \quad \beta_{11} = 2$). This figure indicates unbiased estimates of the derivatives of the LSE with respect to the L_1 weight ω_{11}, but shows a small bias for derivatives with respect to the exponential decay rate β_{11}.

Bi-dimensional Consider a bi-dimensional MHP with exponential kernels, and true parameters

$$\boldsymbol{\mu}^\diamond = \begin{pmatrix} 1.5 \\ 1 \end{pmatrix}, \quad \boldsymbol{\omega}^\diamond = \begin{pmatrix} 0.25 & 0.5 \\ 0.45 & 0.15 \end{pmatrix}, \quad \boldsymbol{\beta}^\diamond = \begin{pmatrix} 1.05 & 1.5 \\ 2.35 & 1.35 \end{pmatrix}.$$

This ground truth MHP has a moderate branching ratio $\rho \simeq 0.67$. For this problem, consider an estimation model `mhp_exp_2d1r` which is an exponential model as in Section 4.6. We simulate one path of this process up to $T = 10^6$, resulting in 7 768 386 jumps. We fit the `mhp_exp_2d1r` model; Figure 5.6 plots the estimated kernels.

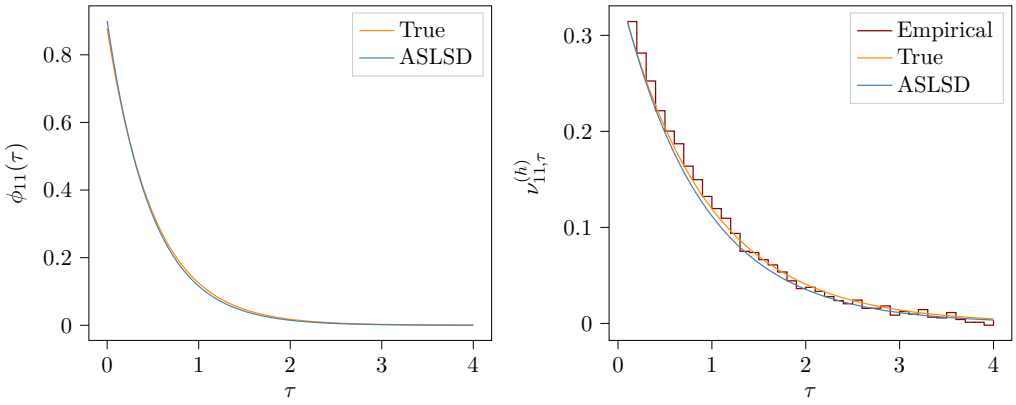


Figure 5.2: Fit results for `mhp_exp_1d1r`
Left: ground truth (orange) and fitted (blue) kernel. Right: Empirical covariance of the training path (red); theoretical covariance of the ground truth and fitted MHP (sampling period $h = 10^{-1}$).

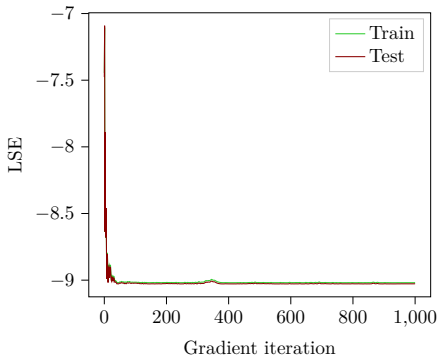


Figure 5.3: LSE at gradient iterations

5.1.1.2 Gaussian

Consider a uni-dimensional MHP with Gaussian kernels (see Section 4.6.3.3), and

$$\mu_1^\diamond = 1.5, \quad \omega_{11}^\diamond = 0.5, \quad \beta_{11}^\diamond = 0.5, \quad \delta_{11}^\diamond = 1.5.$$

We consider a Gaussian MHP model, `mhp_gauss_1d1r`, with the same parameterisation. We discuss both the long path setup and the episodic setup.

Long path setup We simulate one path of the ground truth process up to $T_{\text{long}} = 10^6$, which results in 3 006 805 jumps. Figure 5.7 presents the evolution of model parameter values (left column) and their implied moments (right column) through gradient iterations. There are four parameters fitted (simultaneously) in this model, in order to simply the visualisation of gradient we discuss first order parameters (μ_1, ω_{11}), and then kernel density parameters $(\beta_{11}, \delta_{11})$.

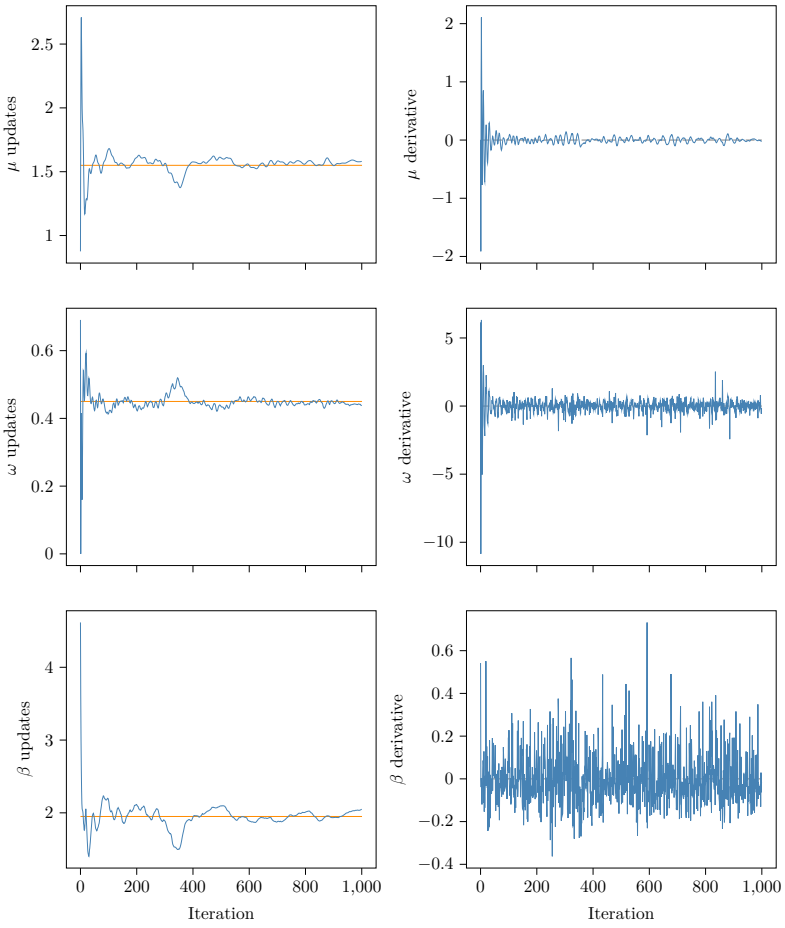


Figure 5.4: Gradient updates for the `mhp_exp_1d1r` model
Orange line corresponds to the true parameter values in the parameter updates column (left).

The baseline μ_1 and the L_1 weight ω_{11} characterize first order properties of the MHP model. Figure 5.7 shows that the stationary regime intensity $\eta_1^\star = \mu_1/(1-\omega_{11})$ implied by parameters updates of `ASLSD` oscillates around the ground truth with decreasing amplitude. The first few updates of (μ_1, ω_{11}) are larger in absolute value, because of the initial learning rate of our ADAM solver being relatively high.

The location parameter δ_{11} and the scale parameter β_{11} further characterize the second order properties of the MHP. Figure 5.7 shows that the dynamics of learned parameters in the (δ, β) plane are very different from the first order behaviour, particularly at the start of training:

- during the first 15 gradient iterations, the location and scale are both decreasing with an affine relation towards the minimal location value allowed in our procedure (10^{-10});
- during the following 28 gradient iterations, the scale parameter β_{11} decreases while the location remains quasi-null.

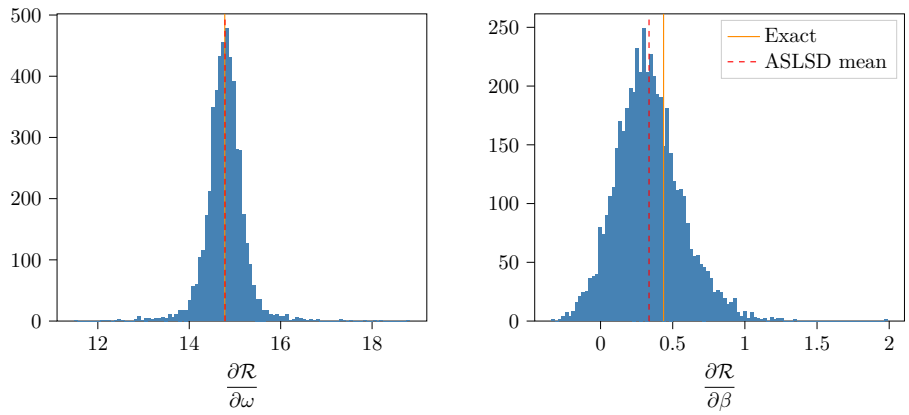


Figure 5.5: Distribution of LSE estimates at a given parameter values
Distribution of LSE gradient estimates for derivatives against the L_1 weight ω_{11} (left) and the exponential decay rate β_{11} (right).

As discussed previously, method of moments approaches to MHP estimation such as Bacry and Muzy [10] exploit the characterization of MHP by their covariance to fit the kernels. However, it is not clear what link LSE minimization has with the MHP. In order to answer this question, we simulate[2] an MHP path up to $T = 10^5$ at each parameters updates, and compute the corresponding empirical covariance up to a lag $\tau = 20$ with sampling period $h = 10^{-1}$. We plot this result in Figure 5.8. Again, we see an oscillatory behaviour of the implied second moment mode around the ground truth throughout gradient iterations.

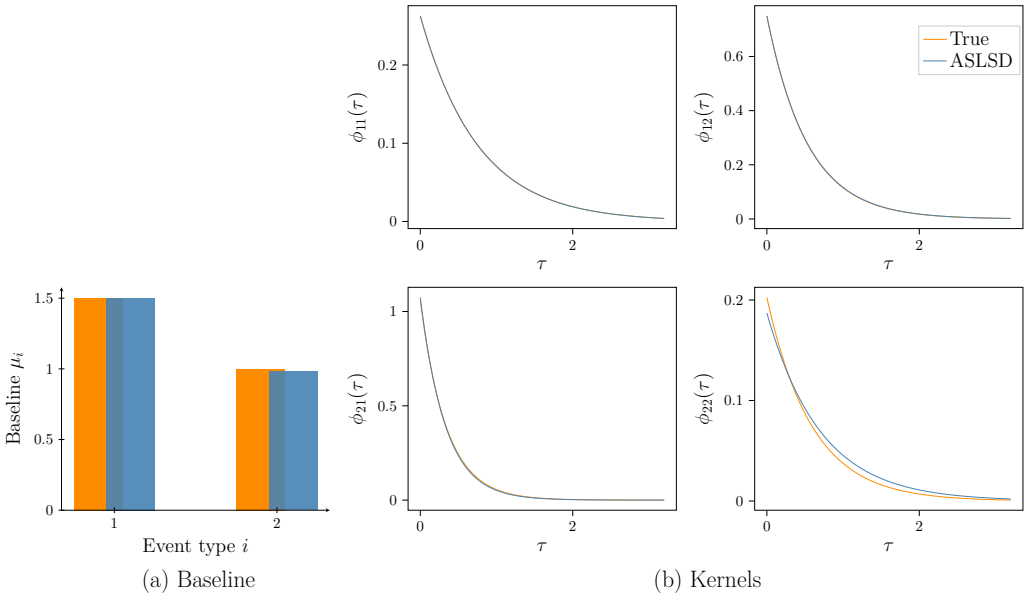
Episodic setup We simulate $n_{\text{ep}} = 1000$ paths of the ground truth Gaussian MHP up to $T_{\text{ep}} = 100$, which results in 296 131 jumps, and use our procedure to estimate the `mhp_gauss_1d1r` model.

Besides the numerical challenges of the `ASLSD` procedure in the episodic case, a first question is whether episodic data is sufficient to characterize a ground truth MHP. Empirically, the order of magnitude of the total number of events is consistent with the long path setup. However, the small observation window $T_{\text{ep}} = 100$ and the preliminary results in Section 4.5.1 invite to caution. Figure 5.8 shows that the empirical moments computed from the episodic data almost match those from the long path data, therefore, it seems reasonable to attempt estimating a Hawkes model in the episodic case. Finally, Figure 5.9 shows fitted kernels for the long path setup and the episodic setup.

5.1.1.3 Power laws

Parametric estimation of power law kernels poses different numerical challenges, coming notably from the long memory of the MHP which implies that large index lags may still

[2]We could not get a closed-form expression for the covariance of a Gaussian MHP.



(a) Baseline (b) Kernels

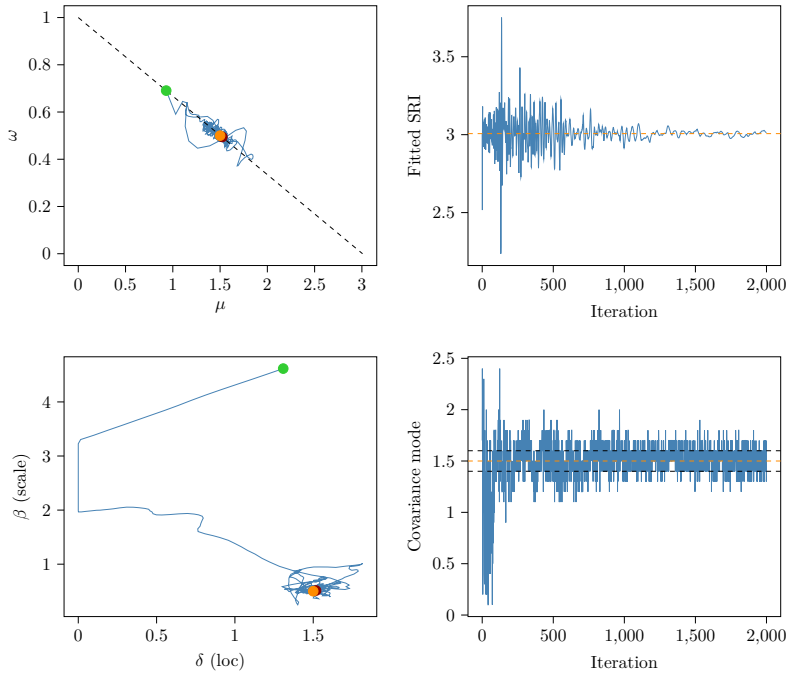
Figure 5.6: Fit results for `mhp_exp_2d1r`

contribute to the stratified MC estimator we build. Consider a ground truth uni-dimensional MHP with power law kernel (see Section 4.6.4.1), and with true parameters

$$\mu_1^\diamond = 1.5, \quad \omega_{11}^\diamond = 0.75, \quad \alpha_{11}^\diamond = 1, \quad \beta_{11}^\diamond = 1.5.$$

To evaluate our method, we set the true exponent value $\alpha_{11}^\diamond = 1$ because we have a closed form expression for the kernel auto-correlation Υ_{111} in this case. We simulate a path of this model up to $T = 10^6$, resulting in 6 024 396 events. We fit a power law MHP model `mhp_power_1d1r` to this data. Figure 5.10 (left) plots the evolution of the relative contribution of each sum term to the LSE (following our notation in Chapter 4) at each gradient iteration: relative contributions oscillate but remain stable after a few gradient iterations, and these contributions are dominate by the double sums S_Υ and S_ϕ. The right side of this figure plots relative allocations of MC samples at each lag stratum for the double sums, at the first gradient iteration (dashed lines) and the last (solid lines). We see that relative allocations at the terminal iteration are monotonically decaying with the stratum index (and therefore with the index lag); and that their contribution is negligible beyond the 25-th stratum, which explains why our estimator works even with a heavy tailed kernel. We evaluate this model by residual analysis. Following the simple approximation in Proposition 3.1.4, fix a cutoff $c = 10^4$. On the training path, compute

$$\tau_c := \min_{m \in [\![c+2, N_T]\!]} t_m^k - t_{m-c-1}^k = 1.384 \times 10^3, \quad \epsilon_c := 1 - \tilde{\psi}_{ki}(\tau_c) = 5.66 \times 10^{-5}. \tag{5.2}$$



(a) Parameters (b) Moments

Figure 5.7: `ASLSD` updates for `mhp_gauss_1d1r`

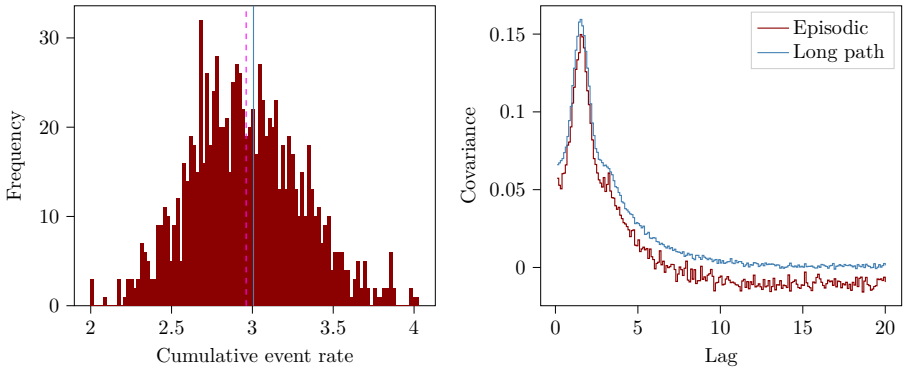
Left: parameter updates of `ASLSD` represented in the (μ, ω) plane (up) and in (δ, β) plane (bottom). Green point is the initial value of the parameters, red point their terminal value, and orange point the true value. The black dashed line in the upper left figure is the line $\mu = 1 - \mu/\eta_T^1$. Right: Moments implied by parameter updates of `ASLSD`. Upper right is the stationary regime intensity $\eta_\star^1 := \mu/(1-\omega)$ at each `ASLSD` update, dashed orange line plots the ground truth value. Bottom right is the mode of the empirical covariance implied by the parameter updates. This empirical covariance is obtained by simulating data from the models up to $T = 10^5$, and computed up to lag $\tau = 20$ with sampling period $h = 10^{-1}$. The mode value is computed as the argument maximizing the empirical covariance. These mode values are discrete with increments of the sampling period $h = 10^{-1}$; the horizontal dashed black lines are respectively at $\eta_T^1 + h$ and $\eta_T^1 - h$.

Figure 5.11 plots the empirical distribution of cut-off timelags and CDF approximation errors. The relative approximation error verifies the upper-bound

$$\max_{m \in [N_T^k]} \frac{|\hat{s}_m^{(k)} - s_m^{(k)}|}{s_m^{(k)}} \leq \frac{\epsilon_c}{1 - \epsilon_c} = 5.66 \times 10^{-5}. \tag{5.3}$$

5.1.1.4 Semi-parametric fit

In this paragraph, we discuss a use case of `ASLSD` in a semi-parametric estimation context. Formally, we define a semi-parametric Hawkes model as an MHP or MTLH with SBF kernels from one of the four dense families of kernels in Section 4.6.3 and in our accompanying code: uniform, triangular, Gaussian, or gamma.



(a) First order (b) Second order

Figure 5.8: Empirical moments of long path and episodic setup

Left: Empirical distribution of the cumulative event rate $\eta_{T_{ep}}$ for each episodes (red), dashed pink line plots the empirical mean of this distribution. Blue line plots the cumulative event rate $\eta_{T_{long}}$ in the long path setup. Right: Red line plots the empirical covariance on the episodes, blue line plots the empirical covariance on the long path, for sampling frequency $h = 0.1s$.

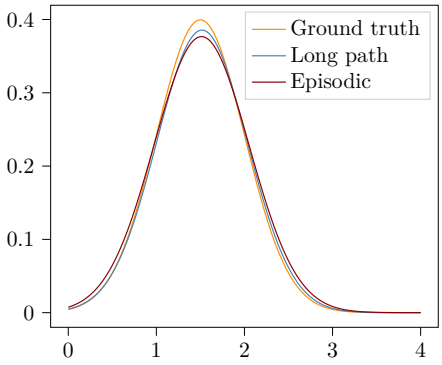


Figure 5.9: Fitted kernels for `mhp_gauss_1d1r`

In the example we discuss here, the data generating process is a uni-dimensional MHP, with true background rate $\mu^\diamond = 0.8$. The only kernel of this MHP, $\phi^\diamond_{11}$, is a mixture of three Rayleigh kernels. For lags $\tau \geq 0$, the Rayleigh kernel density is

$$\tilde{\phi}(\tau) := \frac{\tau}{\beta^2} \exp\left(-\frac{1}{2}\left(\frac{\tau}{\beta}\right)^2\right),$$

and the parameter is the scale $\beta \in (0, +\infty)$. Fix the scale parameters of the mixture

$$\beta^\diamond_{11,1} = 1.25, \quad \beta^\diamond_{11,2} = 3.5, \quad \beta^\diamond_{11,3} = 6.1, \tag{5.4}$$

and L_1 weights

$$\omega^\diamond_{11,1} = 0.2, \quad \omega^\diamond_{11,2} = 0.5, \quad \omega^\diamond_{11,3} = 0.1. \tag{5.5}$$

To reflect typical usage, the model is somewhat misspecified, as the Rayleigh kernel mixture cannot be expressed in terms of any dense families above. We deliberately choose a

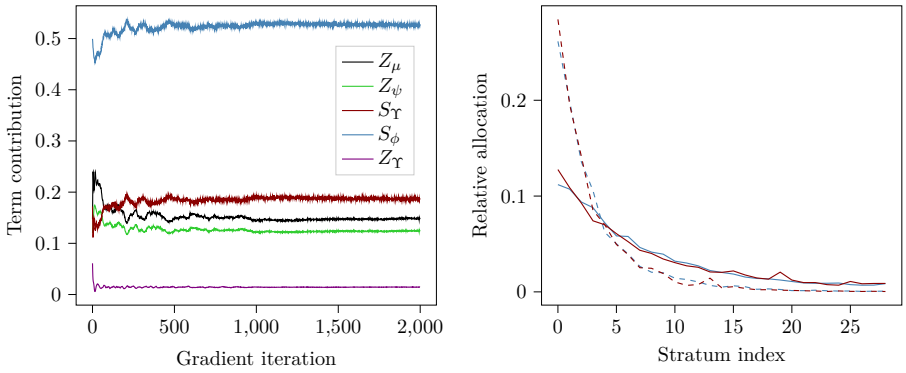


Figure 5.10: Stratified estimator

Left: relative contribution of each term in the LSE estimator. For each sum term, define the relative contribution as the term's absolute value divided by the sum of absolute values of all other sum terms; so that relative allocations sum to 1 at each gradient iteration. Right: Relative allocation of sample points per lag stratum in the double sums; dashed (resp. solid) lines correspond to the allocation at the first (resp. last) gradient iteration. Red lines correspond to S_Υ, blue lines to S_ϕ.

class of models that does not contain the ground truth MHP (and so a perfectly fitting model is impossible to achieve), as a particular challenge where our algorithm is going to struggle due to model error. This type of ground truth process is relatively unusual for the Hawkes literature, however, note that our specification is a favourable setup for kernel estimation: most of the events from this ground truth MHP are asymptotically the result of self-excitation (corresponding to a relatively small exogeneity rate $\text{exo} = 0.2$), this MHP is far from critical (its branching ratio is $\rho = 0.8$), and the Rayleigh distribution does not exhibit long memory. We place ourselves in the long path setup, simulate a path of this process up to $T = 10^6$, which results in $4,005,567$ events.

There are four important modelling choices in the semi-parametric approach:

1. **mixture support:** the modeller must decide of an interval $[\delta_L, \delta_R]$ on which they want to focus the modelling effort, since the uniform and triangular densities are compactly supported, and the Gaussian and gamma densities decay fast. This implies semi-parametric approaches may not be suitable for heavy-tailed ground truth kernels.

2. **kernel family:** families like Gaussian and gamma kernels have more representation power than uniform and triangular kernels, but are more computationally expensive because they require calls to special functions (normal PDF, gamma function). Effectively, the choice of dense kernel family consists in finding a trade-off between accuracy and speed.

3. **number of basis kernel densities:** fixing the hyper-parameter r suffers from a speed-accuracy trade-off. Increasing the number of basis kernel densities increases the time

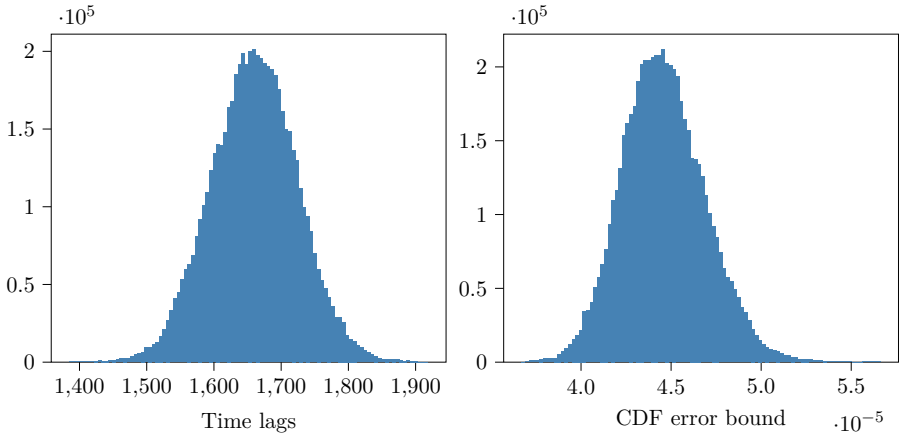


Figure 5.11: Resdiuals approximation error for `mhp_powlaw_1d1r`

Left: empirical distribution of time lags $t^1_m - t^1_{m-c-1}$ between events with index lag $1+c = 1+10^4$.
Right: empirical distribution of the CDF approximation $1 - \tilde{\psi}_{11}(t^1_m - t^1_{m-c-1})$ of these events.

complexity of gradient iterations quadratically (see Section 4.5.3).

4. **location of basis kernel densities:** the last modelling decision is to allocate the r basis kernel densities on the support $[\delta_L, \delta_R]$. For varying numbers of basis kernels r, kernel density models are not necessarily nested, and depending on the grid we specify, increasing the number of basis kernels may reduce the accuracy of the model.

Figure 5.12 plots some empirical properties of the observed data. The empirical covariance[3] shows a distinct mode at $\tau = 1.6s$, and covariance values fall by 60% (compared to that maximum value) for lags greater than $\tau = 20s$. Therefore, we focus our modelling effort on the interval $[\delta_L, \delta_R] = [0, 15]$; only a negligible L_1 mass of the true kernel lies outside of this interval, with $\frac{\int_0^{15} \phi^\diamond_{11}}{\int_0^{+\infty} \phi^\diamond_{11}} \simeq 99\%$.

The distribution of inter-arrival times can be misleading, as the link between this distribution and kernels is unclear. In the present case, the distribution of inter-arrival times is monotonically decaying, unlike the ground truth kernel; and is supported on $[0, 2]$, when $\frac{\int_0^{2} \phi^\diamond_{11}}{\int_0^{+\infty} \phi^\diamond_{11}} \simeq 28\%$.

We now investigate the impact of the choice of families of kernel densities, and of the number of basis kernels densities r_{11}, on the accuracy and speed of the approximation. For each family of kernel densities, we specify our grid of r_{11} kernel densities on $[\delta_L, \delta_R]$ as follows. Define the step size $h := \frac{\delta_R - \delta_L}{r_{11}}$, and for basis indices $l \in [r_{11}]$, set

[3] Thanks to the work of Bacry and Muzy [10], we expect the empirical auto-covariance to be particularly informative as it fully characterizes the kernels, alongside the first order moment. However, as discussed in Section 3.1.2, the link between auto-covariance and kernels is made through the auto-convolutions of kernels (corresponding, heuristically, to descendants of an event over multiple generations): even if a kernel is compactly supported, we will still observe non-zero auto-correlation for large lag values (outside of the kernel support). Despite how informative the empirical covariance is, note that low-resolution covariance estimates might be preferred to limit the computational cost.

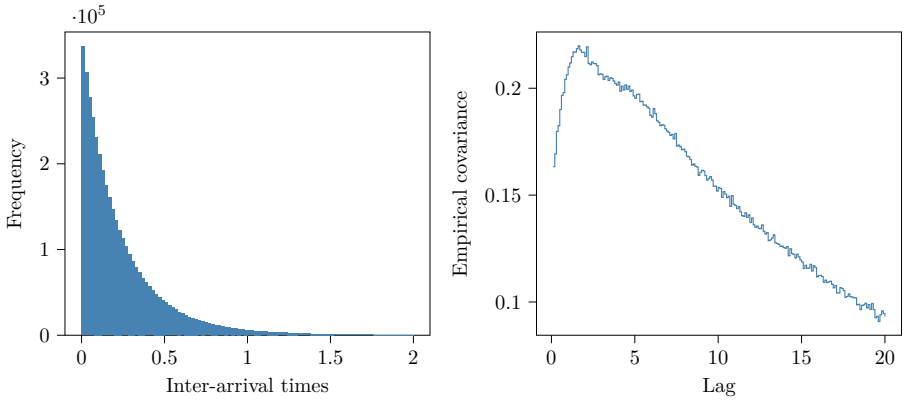


Figure 5.12: Empirical properties of simulated data

Left: empirical distribution of inter-arrival times of events in the simulated path. Right: Empirical covariance of the simulated path, for sampling period $h = 0.1s$.

- for uniform densities: constant interval sizes $a_l := h$, and locations $\delta_l := h(l-1)$;
- for triangular densities: constant $\beta_l = \delta_l = h$, and locations $\alpha_l := h(l-1)$;
- for Gaussian densities: constant scales $h/2$, and locations $\delta_l := h(l-1)$;
- for gamma densities: constant scales h, and locations $\delta_l = 1 + h(l-1)$.

For each kernel family except for uniform densities, our specification ensures overlapping support between basis functions, as this produces signficantly better results in practice. Figure 5.13 plots the run time and L_2 projection error for each kernel family, with varying numbers of basis functions. We fit the Gamma mixture model with 10 basis functions;

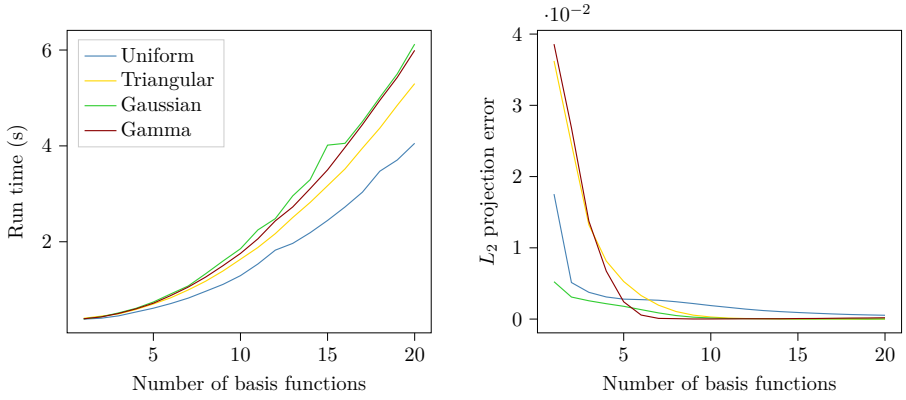


Figure 5.13: Speed-accuracy trade-off

Left: For each class of basis kernel densities, average run time in seconds for one gradient iteration against the number of basis functions in the mixture. Right: For each class of basis kernel densities, L_2 distance between the ground truth and that SBF family against the number of basis functions in the mixture.

Figure 5.14 plots the fitted functionals.

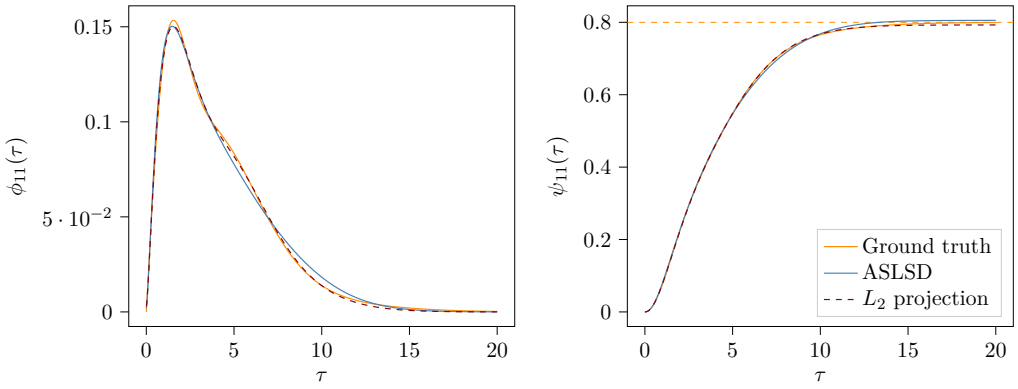


Figure 5.14: Fitted functions for `mhp_rayleigh_1d3r`

5.1.1.5 Benchmarking

We compare the performance of `ASLSD` to two state of the art MHP estimation methods implemented in the python package `tick` [4] with an optimized `C++` backend.

- `SumExp`: an SBF exponential MHP model, fitted using the algorithm in Bompaire et al. [15]. This is an interesting benchmark to evaluate the quality of the fit for the decay parameter β in a non SBF exponential using our method. For general kernels, `SumExp` is our naïve benchmark.
- `WH`: the algorithm proposed in Bacry and Muzy [10], a non-parametric estimation method which solves a Wiener–Hopf system derived from the autocovariance of the MHP. This method applies to any stationary MHP.

Evaluation metrics Consider a $(\boldsymbol{\mu}^\diamond, \Phi^\diamond)-$MHP observed on a window $[0, T]$, and a model $(\boldsymbol{\mu}, \Phi)$. We define metrics to evaluate the performance of our algorithms:

- `L2RelErr` : We define this metric by

$$\texttt{L2RelErr} := \frac{\|\boldsymbol{\mu}^\diamond - \boldsymbol{\mu}\|_2^2}{\|\boldsymbol{\mu}^\diamond\|_2^2} + \frac{\|\boldsymbol{\Phi}^\diamond - \Phi\|_2^2}{\|\boldsymbol{\Phi}^\diamond\|_2^2}, \quad \text{where} \quad \|\boldsymbol{\Phi}\|_2^2 := \sum_{i,j} \int_0^{+\infty} \phi_{ij}^2(t) \mathrm{d}t. \tag{5.6}$$

- `WassErr`: The (first) Wasserstein distance between probability measures f and g is given by

$$\mathbb{W}_1(f, g) := \inf_{\pi \in \Gamma(f,g)} \int_{[0,+\infty)\times[0,+\infty)} |x - y| \mathrm{d}\pi(x, y),$$

where $\Gamma(f, g)$ is the space of measures on $[0, +\infty) \times [0, +\infty)$ with marginals f, g. Define

$$\texttt{WassErr} := \sum_{i=1}^d |\mu_i^\diamond - \mu_i| + \sum_{i=1}^d \sum_{j=1}^d \mathbb{W}_1 \left(\frac{\phi_{ij}^\diamond}{||\phi_{ij}^\diamond||_1}, \frac{\phi_{ij}}{||\phi_{ij}||_1} \right) + \sum_{i=1}^d \sum_{j=1}^d \left| ||\phi_{ij}^\diamond||_1 - ||\phi_{ij}||_1 \right|. \tag{5.7}$$

Procedure Define four ground truth MHP: a uni-dimensional exponential MHP `Exp1D`, a bi-dimensional exponential MHP `Exp2D`, a uni-dimensional Gaussian MHP `Gauss1D`, and a uni-dimensional Gaussian MHP that is a mixture of three Gaussian densities `Semi1D`. For each ground truth MHP, we simulate paths $(\mathcal{T}^{(p)})_{p\in[n_{\text{paths}}]}$ of the process up to a terminal horizon T. We consider $(T_q)_{q\in[n_{\text{times}}]}$, an increasing sequence of times. For each $(\boldsymbol{\mu}^\diamond, \Phi^\diamond)$, for each simulated path $\mathcal{T}^{(p)}$, and for each integer $q \in [n_{\text{times}}]$, we define

$$\mathcal{T}^{(p,q)} := \mathcal{T}^{(p)} \cap [\,0, T_q]^{\,d};$$

i.e., the path of the process truncated at T_q, containing $N_{T_q}^{(p)}$ jumps. Finally, for each ground truth MHP, for each evaluation metric, and for each of the three algorithms considered (our method and the two benchmarks), we fit a given MHP model to the observations $\mathcal{T}^{(p,q)}$ and compute the error $\epsilon^{(p,q)}$. For each time discretization step q, we compute the empirical mean (resp. 25th percentile and 75th percentile) of $(\epsilon^{(p,q)})_{p\in[n_{\text{paths}}]}$ and denote it by $M^{(q)}$ (resp. $Q_{0.25}^{(q)}$ and $Q_{0.75}^{(q)}$). Define the mean number of jumps per path

$$\tilde{N}_q := \frac{1}{n_{\text{paths}}} \sum_{p=1}^{n_{\text{paths}}} N_{T_q}^{(p)}.$$

Figure 5.15 plots the mean errors $M^{(q)}$ against the mean number of jumps $\tilde{N}_q$ at each index $q \in [n_{\text{times}}]$, and the shaded area between the lower quartiles $(Q_{0.25}^{(q)})_{q\in[n_{\text{times}}]}$ and the upper quartiles $(Q_{0.75}^{(q)})_{q\in[n_{\text{times}}]}$. We see that our algorithm outperforms the `WH` benchmark with respect to each evaluation metric in the two exponential cases, `Exp1D` and `Exp2D`. In the unimodal Gaussian case `Gauss1D`, `ASLSD` outperforms `WH` for smaller datasets (under 10^6 jumps) in the Wasserstein metric `WassErr`. For the multi-modal Gaussian example `Semi1D`, `WH` typically outperforms `ASLSD`, however this is due to the (deliberate) misspecification of the SBF Gaussian model which is seen to approach its lower L^2 error bound. Our procedure consistently outperforms the `SumExp` benchmark for all ground truths.

5.1.2 MTLH Data

5.1.2.1 Periodic baseline

Consider a uni-dimensional MTLH model with cosine baseline (frequency $a^\diamond := \pi \times 10^{-5}$, phase $b^\diamond := \pi/2$, scale $\alpha^\diamond := 0.5$, intercept $\delta^\diamond = 1$) and Gaussian kernel (L_1 weight $\omega_{11}^\diamond =$, location $\delta_{11}^\diamond = 3.5$, scale $\beta_{11}^\diamond = 2$).

5.2 Epidemic propagation

In this application, we study the propagation of Malaria in Yunan province, China, for $T = 1000$ days between 1 January 2011 and 24 September 2013. Malaria is not contagious

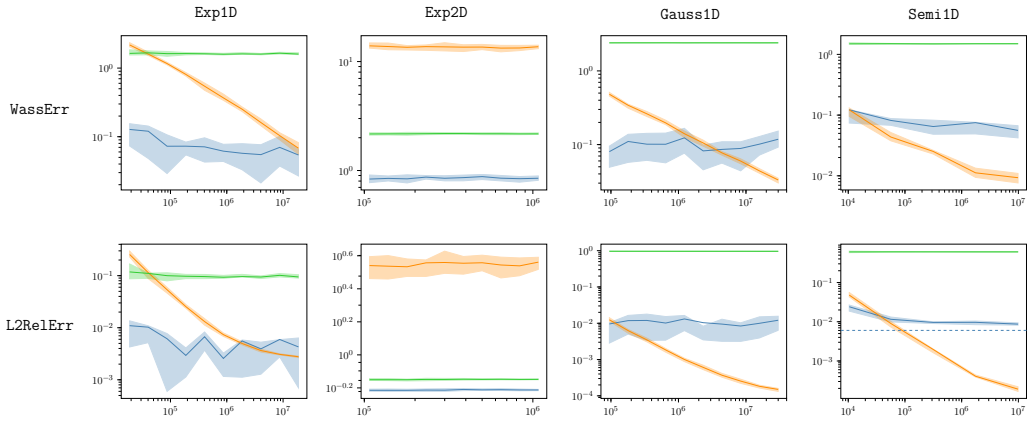


Figure 5.15: Performance of `ASLSD` and benchmarks (MHP)
Error plots for the various ground truth MHP. Blue lines correspond to our algorithm, orange lines to the `WH` benchmark, and green lines to the `SumExp` benchmark. In the lower-rightmost plot, dashed blue line is the lower bound of the `L2RelErr` between the ground truth MHP and a `Semi1D` MHP. This lower bound corresponds to the `L2RelErr` between the ground truth MHP and the L_2 projection of the ground truth MHP on the parametric space of kernels of `Semi1D`.

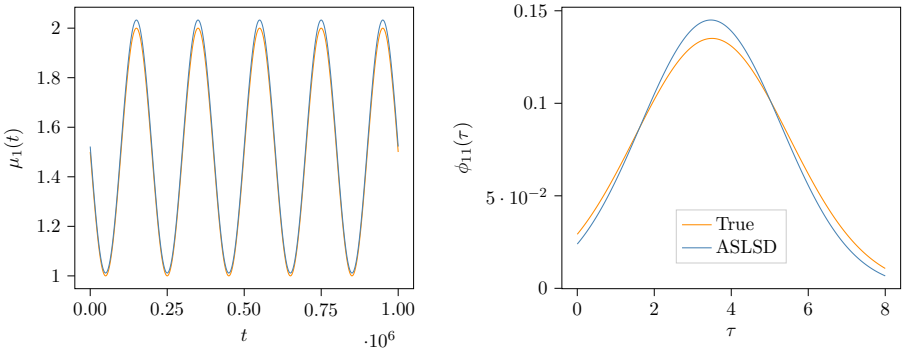


Figure 5.16: Fitted functionals for `mtlh_cos_gauss`

between humans: a human cannot be infected by casual contact or directed exposure to infected blood. Instead, the parasite is transmitted from humans to mosquitoes, and vice versa. In the study of infectious disease within a specific geographic zone in epidemiology, there is a natural exogenous-endogenous event dichotomy. Autochthonous (or locally acquired) cases refer to infections occurring in the studied zone, while imported cases are already infected individuals who arrive in the geographic zone after an infection somewhere else. In this work, we model timings of new infections using Hawkes models, then leverage these models to build an imported cases classifier. To reproduce the results in this section, see the code in the *experiments* folder of `https://github.com/saadlabyad/aslsd/`.

5.2.1 Data analysis

Dataset We use the dataset of Unwin et al. [100] who study the disease transmission; this data is available in Unwin et al. [101]. The dataset contains 2153 recorded infections, with

a decimal time code in days up to an unspecified resolution. The smallest inter-arrival time is about $15s$, therefore the resolution seems to be at least of order a few seconds. This data also records whether the case is imported or local. The first event timestamp is at $t_1^1 = 0$ days, and the last timestamp is at $t_{N_T^1}^1 \simeq 997.05$ days. In Hawkes modelling, if we consider that the ground truth process is observed on $[0, T]$, the underlying assumption is that there is a null probability to observe an event occurring at $t = 0$ or $t = T$. Therefore, we offset the timestamp t_1^1 of the first event by a small random value. It is unclear how sensitive the solutions of the LSE minimization problem are to this random offset. The dataset identifies which infections are imported (1) or autochthonous (0).

Non-stationarity Figure 5.17 shows a clear non-stationarity of the observed point process. In fact, the data seems to exhibit seasonality in cycles of 1 year. Figure 5.17 plots the empirical intensity estimator for this data, with a sampling period $h = 10$ days. The infections present a mode around the summer, and decrease during winter. This is well known in epidemiology, and is explained with the fact that the density of mosquitoes is highly correlated with temperature and humidity levels; the Yunnan region has a tropical monsoon climate, with a rainy season usually occurring from May to October, peaking from June to August. As mentioned above, this dataset records whether registered cases are im-

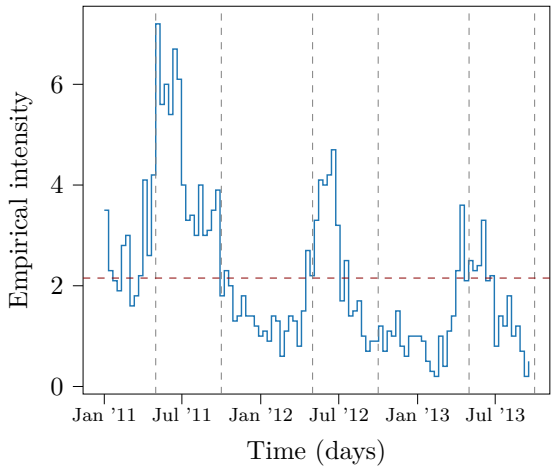


Figure 5.17: Empirical intensity

Empirical intensity of the Malaria data. Dashed vertical lines plot the 1st May and 1st October of each year, corresponding roughly to the start and end of rainy season in Yunnan. The dashed horizontal line corresponds to the average event rate η_T^0.

ported or local. Figure 5.18 plots the empirical intensity of imported and local cases (left), and the cumulative exogeneity ratio (right). We note that 74.8% of cases in this dataset are imported. First, the empirical intensities seem to display a lead–lag relationship between the imported and local cases; note that the difference between the modes of these intensities

is at 50 days. The timings of imported events are clearly non-stationary, and incompatible with homogeneous Poisson dynamics as confirmed by a KS test. This implies MHPs are not appropriate models for this data. Second, on the right picture, the exogeneity ratio seems fluctuate less after the first 3 months of data and stabilize to a value close to 75%, which is compatible with an MTLH with periodic baseline.

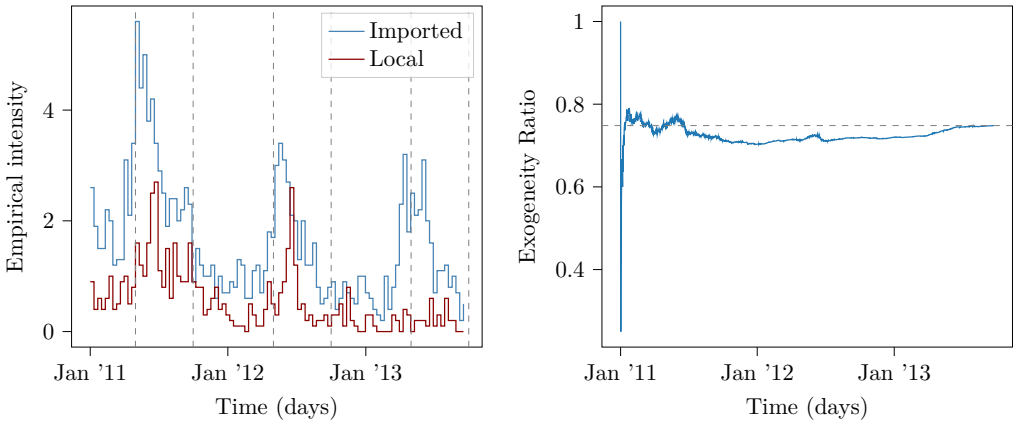


Figure 5.18: Dynamics of imported cases

Covariance Figure 5.19 plots the empirical auto-covariance of the data up to a lag $\tau = 180$ days, with a sampling period $h = 1$ day. The empirical auto-covariance seems to be roughly decreasing with the lag, and we note two important features. First, the empirical auto-covariance gets negative for large lags, above $\tau = 120$ days. This corresponds to the order of magnitude of 4 months or 1 season, and might be due to to the seasonality of the data. Negative auto-covariance cannot be captured by MHP models. Second, the empirical auto-covariance decays relatively slowly compared to the order of magnitude of the incubation period of Malaria. With MHP models, we expect this feature to cause near-critical fits: either with large branching ratios or heavier tails in kernel densities.

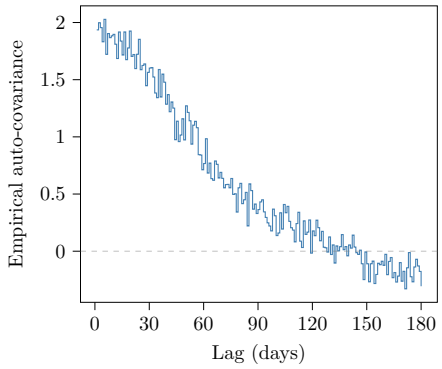


Figure 5.19: Empirical auto-covariance

5.2.2 Modelling

Unwin et al. [100] use a slightly modified uni-dimensional MHP with a delayed Rayleigh kernel for their study. To fit their model, given the typically small number of observations in the applications they consider, they compute exactly the log-likelihood of their observations and input it to a standard optimization solver. In this study, we use Poisson models as benchmarks: `poisson_hom` is a homogeneous Poisson model. `poisson_cos` is a cosine Poisson model (see Section 2.4.3) with fixed frequency $a := 0.02\text{days}^{-1}$ (*i.e.* the intensity has period 314 days) and fixed phase $b := 3.22$ days, and where we fit the parameters α, δ. `poisson_pc` is a piece-wise constant Poisson model (see Section 2.4.2) with 5 intervals of 200 days. We use two MHP models: `mhp_exp` is an MHP with one exponential kernel; and `mhp_gauss` is an MHP with one Gaussian kernel.

Our two main models are MTLH: `mtlh_exp_pc` is an MTLH with no marks, piece-wise constant baseline specified as in `poisson_pc`, and one exponential kernel; and `mtlh_semi_pc` is an MTLH with no mark, piece-wise constant baseline specified as in `poisson_pc`, and semi-parametric kernel based on 10 Gaussian kernels with locations uniformly spaced on $[0, 20]$ days, and scales equal to 50% of the difference between consecutive locations.

We fit `poisson_hom` and `poisson_pc` using the analytic minimizer of their LSE. We fit all other models using `ASLSD` with 2000 gradient iterations: for `mtlh_semi_pc`, we use our gradient approximation, for all other models, we compute the exact gradient given the small size of the data. Given the small number of observations, the standard implementation of the `WH` method would not run on this dataset. Table 5.1 summarises our results; where **LSE** is the exact LSE of the model on the training path, **KS** and **WW** are the p-values of the respective tests, and **M1** and **M2** are the empirical moments metrics defined in Section 2.3.3. `mtlh_semi_pc` performs best, achieving satisfactory results.

Model	**LSE**	**KS** (%)	**WW** (%)	**M1** (%)	**M2** (%)	**OSA** (%)
`poisson_hom`	-4.635	0.0	0.0	58.7	96.3	87.6
`poisson_cos`	-4.903	0.0	0.0	56.7	87.0	88.3
`poisson_pc`	-6.098	0.0	0.0	39.2	23.3	73.5
`mhp_exp`	-6.375	**94.6**	82.9	72.1	64.4	74.7
`mhp_gauss`	-6.359	90.1	21.1	80.3	36.3	73.8
`mtlh_exp_pc`	-6.443	77.8	**96.6**	39.9	38.8	**68.4**
`mtlh_semi_pc`	**-6.457**	65.2	82.9	**37.5**	**21.7**	68.5

Table 5.1: Evaluation of fitted models (Malaria).

Fitted models Figure 5.20 displays fitted baselines and kernels. As expected, MTLH fits show clearly non-stationary baselines which are correlated with the piece-wise constant Poisson fit. The values of MTLH baseline are less than half those of the piece-wise constant Poisson fit, which is consistent with the poor WW p-value of Poisson models. `mtlh_semi_pc` has a branching ratio of 63.4%, and 60.6% for `mtlh_exp_pc`. MHP models are both close to instability and seem to model the data non-stationarity through near-critical fits (see the discussion in Section 3.3): their branching ratios are 91.7% for `mhp_exp`, and 91.2% for `mhp_gauss`. Note that the fitted scale of `mhp_gauss` is 8.15 days, and the fitted decay rate of `mhp_exp` is 0.15 days. As discussed in Section 4.6, the limit of the Gaussian (resp. exponential) kernel when the scale (resp. decay rate) parameter goes to infinity (resp. zero) is the null kernel, which reinforces the observation that MHP fits are unreliable in this case. Unwin et al. [100] propose that the self-excitation of malaria infection is better modelled

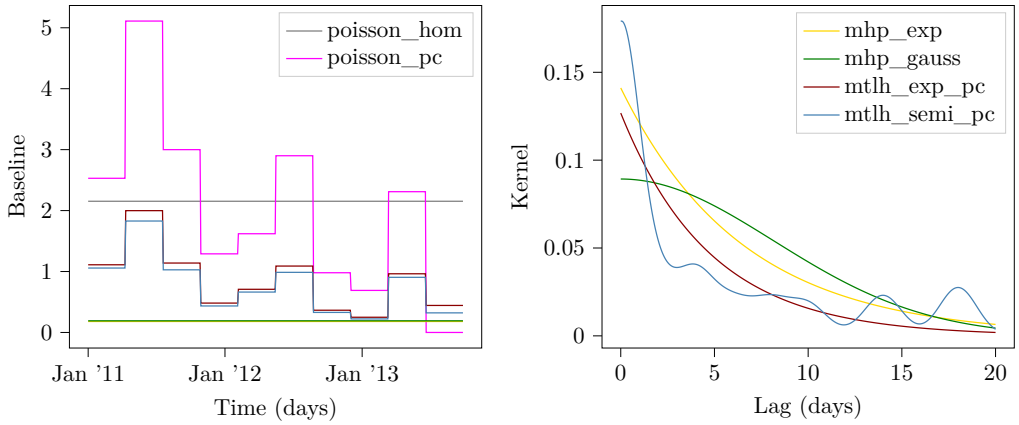


Figure 5.20: Fitted models (Malaria)

by a non-monotonic kernel, because of the delay in contagions due to the mosquito phase of the disease. They force this by using a delayed Rayleigh kernel, with a fixed delay of 15 days. It is therefore not surprising that `mtlh_semi_pc` outperforms the other models as it incorporates multi-modal as well as non-stationary modelling. The fitted location of the Gaussian kernel in `mhp_gauss` is null, but we see clearly two modes in the tail of `mtlh_semi_pc`, located at 14 and 18 days: Figure 5.21 shows that 20% of the L_1 mass of the fitted kernels is located between 14 and 20 days. Given the delays in contagion caused by the malaria life cycle (in particular during the mosquito phase, as discussed, for example, in Stopard et al. [95] or Unwin et al. [100]), we might expect the true kernel to have negligible mass below 10 days. However, we speculate that this dataset may not be based on full observation of malaria cases (for example, due to some cases remaining unreported). This can introduce significant biases in the estimated self-excitation structure, which may explain that 61.8% of the fitted kernel mass in `mtlh_semi_pc` is under 6 days,

with a primary mode at 4 days. It is important to underline that this remains a heuristic interpretation: to the best of our knowledge, in the Hawkes models estimation literature, there is no study of influential observations, censored data, and leverage analogous with regression analysis in the *i.i.d.* case.

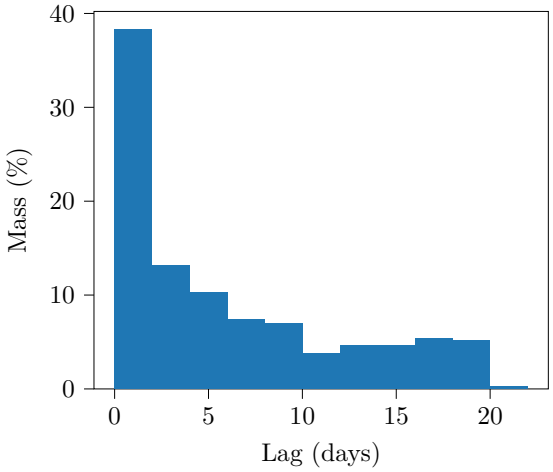


Figure 5.21: Mass distribution (Malaria)
Each bar with edges x_L, x_R corresponds to the probability $\tilde{\psi}_{11}(x_R) - \tilde{\psi}_{11}(x_L)$, where $\tilde{\psi}_{11}$ is the fitted CDF of `mtlh_semi_pc`.

Empirical moments The **M1** and **M2** metrics above show the superior accuracy of the `mtlh_semi_pc` model in fitting the empirical moments of the training data. On the left, we compute the conditional intensity on the training path of each fitted model, and compare it to the empirical intensity of the training data. On the right, we simulate a path from each fitted model, and compute the empirical intensity on that path. This figure raises an interesting paradox for MHP models: despite the clear non-stationarity of training data, the conditional intensity of MHP models on the training data fits the ground truth empirical intensity correctly, however, the empirical intensity of data simulated from these models is far from the ground truth (`M1` $\geq 70\%$). Figure 5.23 plots the empirical covariance of simulated paths from fitted models, comparing it to the empirical covariance of training data. MTLH models achieve a satisfactory `M2` error, as expected from the good performance of the piece-wise constant Poisson model. Note that `mhp_gauss` achieves a surprising low `M2` error. Due to the small size of the dataset, it is difficult to draw conclusions from this covariance analysis.

Residuals Figure 5.24 plots the residuals of fitted models. As shown in Table 5.1, Poisson residuals display poor fits, failing the KS and WW residual test. All Hawkes models in this study pass both tests. Figure 5.25 plots the distribution of inter-arrival times simulated

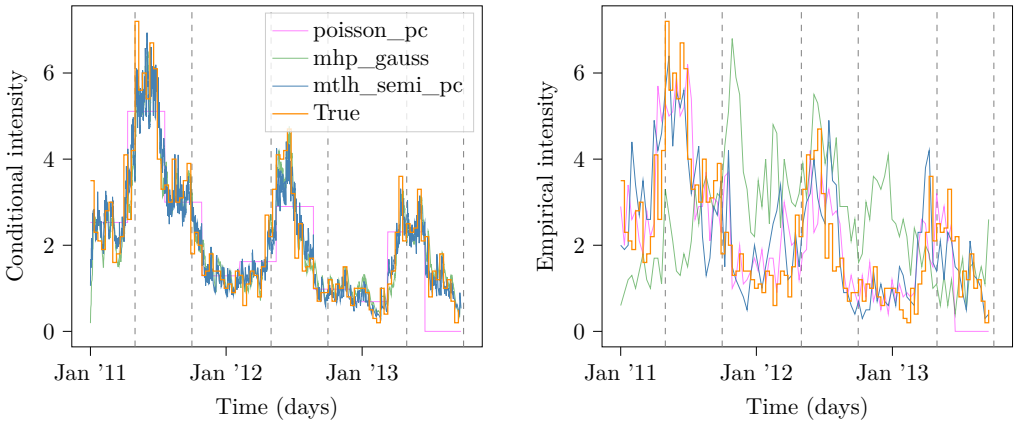


Figure 5.22: Intensity evaluation (Malaria)

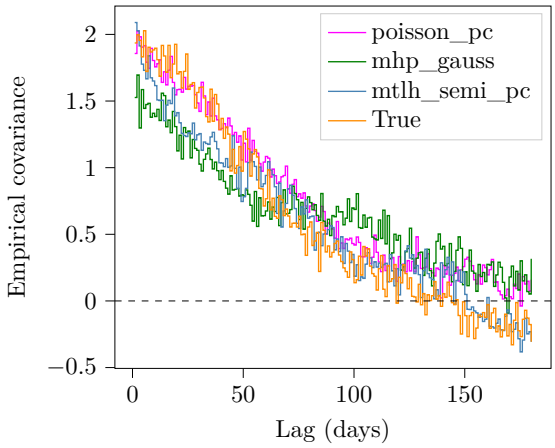


Figure 5.23: Covariance evaluation (Malaria)

from the fitted models, which compare favourably to the ground truth.

OSA For one–step–ahead evaluation of the models (see Section 2.3.2), we consider a uniform time grid on $[200, 800]$ with $n_{\text{grid}} = 50$, and $n_{\text{paths}} = 1000$. Figure 5.26 displays our results for `mtlh_semi_pc`. The predicted lags to next event are positively correlated (60% Pearson correlation), and with prediction error $\texttt{OSA}_R = 68.5\%$. The simulated lags have the same first and second moment as the empirical distribution of observed lags, however, they struggle to recover extreme values of this distribution.

5.2.3 Classifying imported cases

In epidemiology, identifying imported cases is critical for designing appropriate public health response to an epidemic. As discussed in Section 5.2.1, the identification of imported and local cases is a binary classification problem with imbalanced classes. In this paragraph,

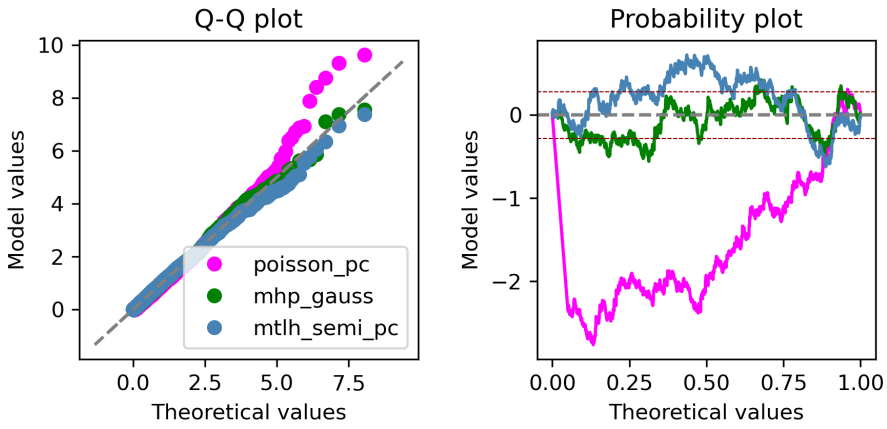


Figure 5.24: Residuals (Malaria)

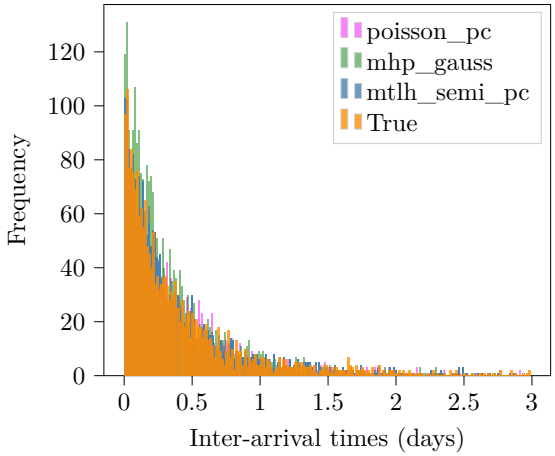


Figure 5.25: Simulated inter-arrival times (Malaria)

we discuss the use of exogeneity probabilities of the `mtlh_semi_pc` fit for this classification task.

Calibration The exogeneity probabilities give target scores for our classifier, we first need to fine tune our decision threshold. We simulate a path from the model up to T, where we know which events are exogenous, and compute the exogeneity probability for all events in this path. Figure 5.27 plots the exogeneity probability at each event time in the simulated path, and the empirical distribution of exogeneity probabilities. This model has a branching ratio of 63.4%: as expected, most events in the simulated path are endogenous and the exogeneity probabilities have relatively small values. Surprisingly, the modes of exogeneity probabilities are not located during the monsoon periods, but rather in November and December. Figure 5.28 plots classic metrics and the receiver operating characteristic (ROC) curve for our classifier. This model achieves a low AUC of 0.59. The

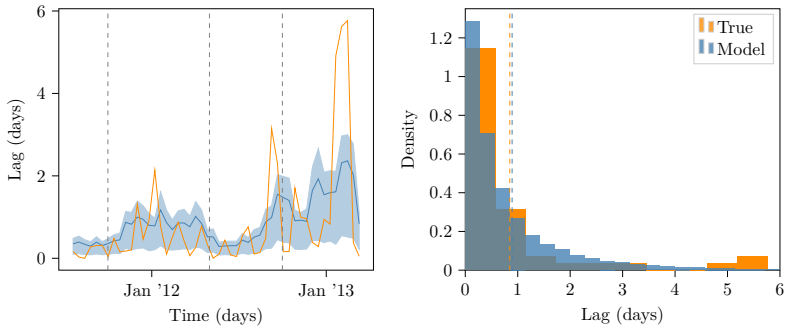


Figure 5.26: OSA evaluation (Malaria)

Left: temporal dynamics of predicted (blue) and observed (orange) lags until next jump, and shaded area between the lower and upper quartiles of simulated values. Dashed vertical lines plot the start and end dates of monsoon in Yunan. Right: empirical distribution of simulated (blue) and observed (orange) lags until next jump. Dashed vertical lines plot empirical means.

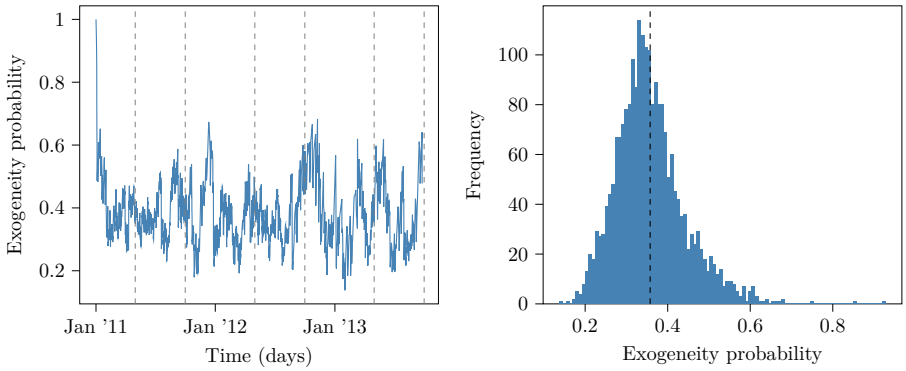


Figure 5.27: Dynamics of exogenous classifications, calibration (Malaria)

Dashed line is the empirical mean of the distribution.

positives and recall plots are particularly close.

Results We now evaluate this classifier on the real data.

Similarly to the calibration results, this classifier achieves an AUC of 0.56 on the real data; all Hawkes models fitted in this study have a similar AUC value. Figure 5.30 plots different metrics and the ROC curve. Figure 5.31 shows indeed that the empirical distribution of exogeneity probabilities of imported cases is very similar to the distribution of the exogeneity probabilities of local cases.

5.3 News cycles

In this application, we study the diffusion of information across media platforms: in particular, we model publication timings of news articles with keywords related to the British

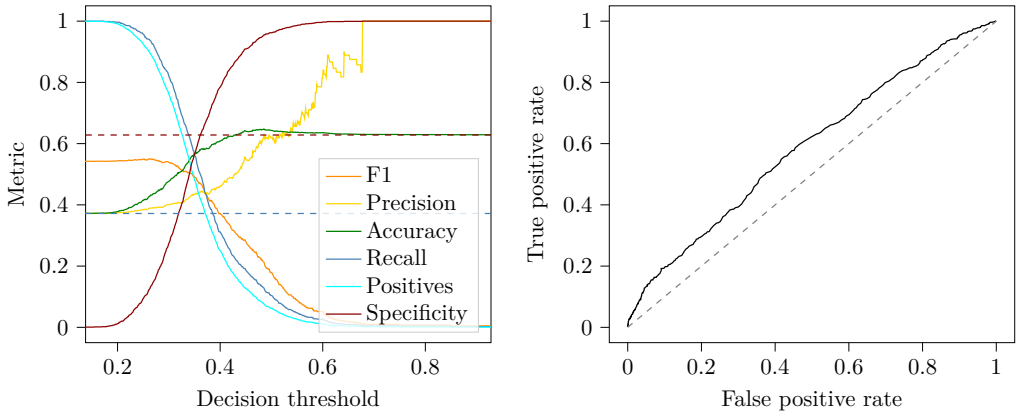


Figure 5.28: Metrics of exogenous classifications, calibration (Malaria)
Blue (resp. red) horizontal dashed line is the true rate of exogenous (resp. endogenous) events in the simulated data.

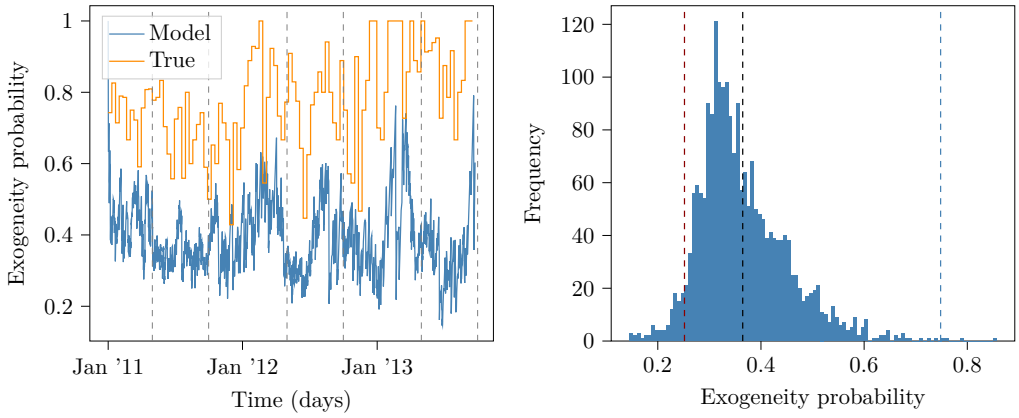


Figure 5.29: Dynamics of exogenous classifications, real data (Malaria)
Left: Blue plots exogeneity probabilities against time. Orange line plots the empirical intensity of true imported cases against time. Right: empirical distribution of exogeneity probabilities. Black dashed line is the empirical mean of the distribution; blue (resp. red) dashed line is the true rate of exogenous (resp. endogenous) events in the data.

Royal family [4] after the wedding of Prince William and Catherine Middleton on Friday 29 April 2011. We train our models on publications occurring on the week from Monday 6 June 2011 midnight UTC to Saturday 11 June 2011 midnight UTC. In particular, we are interested in modelling the interaction between publications in the UK and the USA. During that period, the UK uses British Summer Time (BST) corresponding to UTC+1, and the East coast of the USA uses Eastern Daylight Time (EDT) corresponding to UTC-4. In online news coverage, information can be propagated through different media with high intensity: this is refered to as viral news, which is studied for instance by Lu and Szymanski [64] or Benson [14]. Similar text phrases can occur in news articles from different media;

[4]The exact keyword in the MemeTracker dataset is *prince william-william-kate middleton-kate-middleton-westminster-watch-marriage-queen-king-elizabeth-charles*.

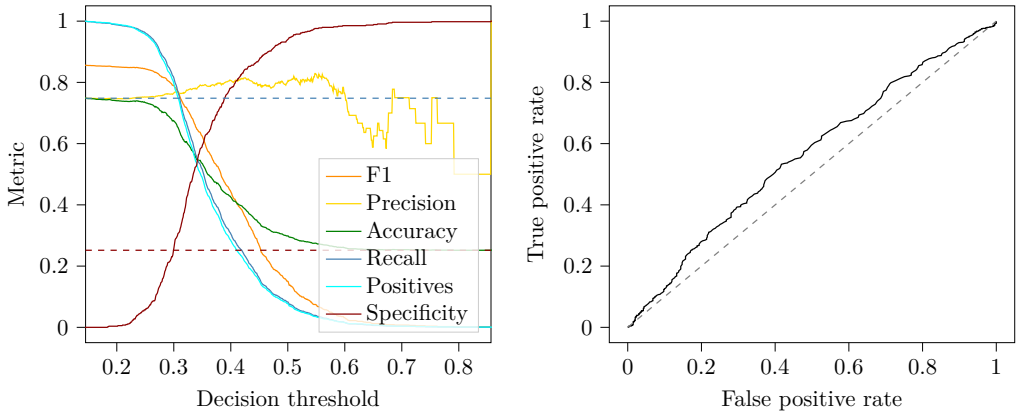


Figure 5.30: Metrics of exogenous classifications, real data (Malaria)
Blue (resp. red) horizontal dashed line is the true rate of exogenous (resp. endogenous) events in the data.

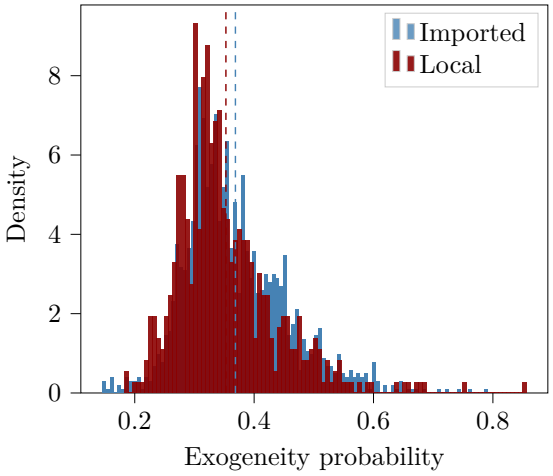


Figure 5.31: Distribution of exogenous probabilities, real data (Malaria)

Leskovec et al. [57] study this phenomenon, which they refer to as memes. Gomez Rodriguez et al. [40] compiled news articles from several websites that mention a selection of keywords into the MemeTracker dataset. In our application, we model the excitation in the occurrences of a given keyword irrespective of the meme to which it belongs to. There is an inherent causality structure and exogenous-endogenous event dichotomy at the level of an event cascade: for a fixed meme, the first news article in the cascade causes the occurrence of that meme in the subsequent articles in the cascade[5]. However, when we shift from meme cascades to news articles, the causal structure is more ambiguous because a given news article (A) can contain more than one meme, in this case:

- there are different other news articles causing the occurrence of these memes in the article (A), and we cannot attribute (A) to a unique parent;

[5]This assumes no anterior mention of the meme has been missed.

- if article (A) is the first recorded in a given cascade, but is not the first recorded in another cascade, then we cannot consider article (A) to be exogenous anymore.

5.3.1 Data analysis

Dataset The MemeTracker dataset was initially developed by Leskovec et al. [57] with the objective of studying propagation of memes across different media; we will use the updated version of Gomez Rodriguez et al. [39]. Each file lists timestamps of occurrences of a given keyword across 5 000 media outlets. The data is timestamped with a resolution in seconds. When two or more posts have the same timestamp, we only keep the event that appears first in the dataset.

Seasonality Figure 5.32 plots the empirical intensity of publication timings between Monday 6 June 2011 at midnight UTC and Monday 28 November 2011 at midnight UTC. We see a clear decrease of publication frequency during weekends. Figure 5.33 plots the empirical

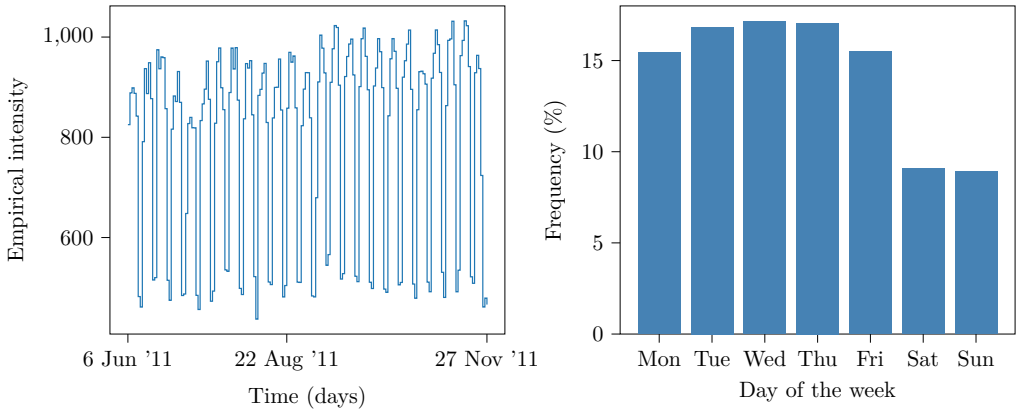


Figure 5.32: Seasonality of news cycles

Left: Empirical intensity of publications. Right: Empirical distribution of publication days.

moments of publication timings during the training week starting Monday 6 June 2011, for media of all nationalities. The vertical dashed lines correspond to 02:00 PM UTC.

Media nationality Diffusion dynamics of news related to the British Royal family differ significantly between UK, USA and Australian media, and those from other nationalities. It is not sufficient to use the top-level domain of the website to deduce its country.[6] We manually verify the nationality of the media sources; the list of media nationalities is available as a *csv* file in the *Applications* folder of our repository. In this study, we focus on

[6]For example, several major news websites had a .com top-level domain at the time of data collection of MemeTracker, such as *The Economist* (British), *El País* (Spanish) or *Globo* (Brazilian).

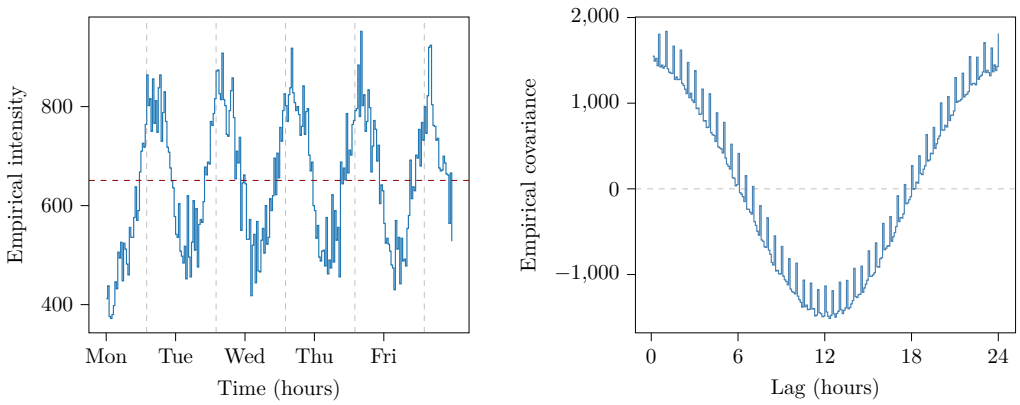


Figure 5.33: Empirical moments (News 1D)

Left: empirical intensity of publication dates with 30 minutes sampling period. Horizontal dashed line is the cumulative event rate on the training period, at 10.8 articles per minute. Right: empirical covariance of publication dates with 5 minutes sampling period.

the interaction between publications in the UK and in the USA. Figure 5.34 plots the correlation matrix of the number of daily publications for a selection of media, computed on the 25 weeks period starting 6 June 2011. Dashed lines delimit UK media from US media. As expected, we note a high correlation cluster corresponding to the websites of generalist British media, differentiated from British tabloids. Note that all US media, with the exception of *CNN*, are highly correlated with the news agency *Reuters*. Among US media, we observe a clear cluster of the websites of generalist news channels, distinct from primarily paper–based or online–based US media. We investigate the lead-lag relationship between

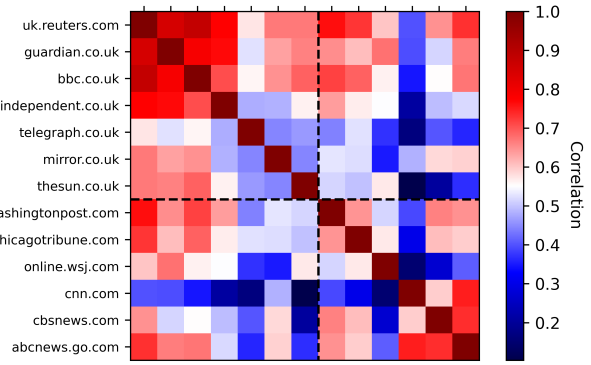


Figure 5.34: Correlation matrix of daily publications

UK and US publications. Figure 5.35 plots the empirical intensity of publications with 30 minutes sampling period. Blue (resp. red) vertical dashed lines are located at 09:00 AM BST (resp. EDT). In media from both countries, publications have a daily periodicity, with

a clear delay due to the different time zones. Note the higher intensity of US publications (1.67 publication per minute) than UK publications (0.91 publication per minute): we attribute this to the predominance of US media in the dataset. Finally, Figure 5.36 plots

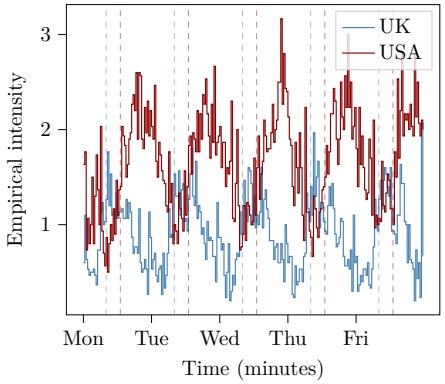


Figure 5.35: Empirical intensity (News 2D)

the empirical covariance of this data with 5 minute sampling period. The auto-covariance of UK publications has similar dynamics to the auto-covariance of US publications, with a different scaling factor. The cross-correlation from UK to US publications displays a clear mode at 6 hours; note that the time difference between the UK and the East coast of the USA is of 5 hours during the data period. Negative correlation values in this empirical estimator might result from data non-stationary rather than inhibitory effects.

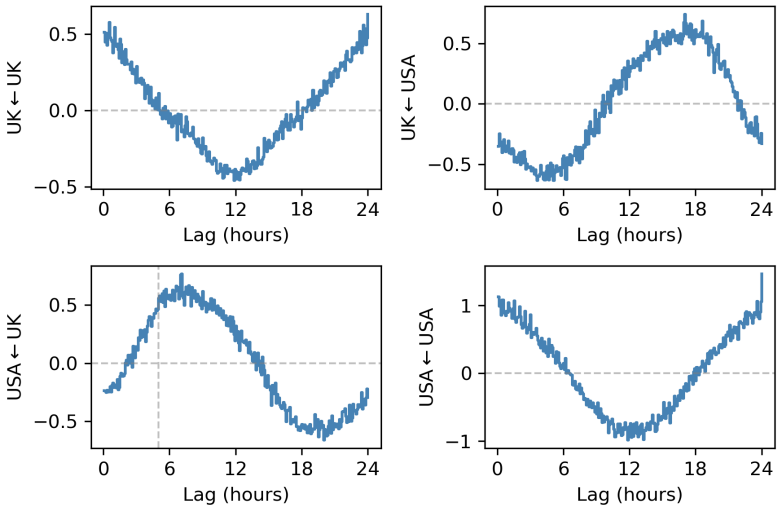


Figure 5.36: Empirical covariance (News 2D)

5.3.2 Modelling

We aggregate all publication timestamps in the training week from UK media (dimension $i = 1$) and US media (dimension $i = 2$) into a path of a bi-dimensional point process. We measure timestamps as minutes elapsed since Monday 6 June 2011 midnight UTC, with one second resolution.

We use Poisson and MHP models as benchmarks: `poisson_hom` is a homogeneous Poisson model. `poisson_cos` is a cosine Poisson model (see Section 2.4.3) with fixed frequency $a := \frac{2\pi}{24\times 60} \simeq 4.3 \times 10^{-3}\text{min}^{-1}$ (*i.e.* a period of 1 day), where we fit the scaling parameters α, δ and the phase parameter b. `poisson_pc` is a piece-wise constant Poisson model (see Section 2.4.2) with 12 intervals of 720 minutes (12 hours) each. `mhp_gauss` is an MHP model with one Gaussian kernel.

Our two main models are MTLH: `mtlh_exp_cos` is an MTLH with no marks, cosine baseline specified as in `poisson_cos`, and one exponential kernel. `mtlh_semi_cos` is an MTLH with no mark, cosine baseline specified as in `poisson_cos`, and semi-parametric kernels based on 7 Gaussian basis kernels with locations uniformly spaced from 0 360 minutes (6 hours), and scales equal to 50% of the difference between consecutive locations.

We fit `poisson_hom` and `poisson_pc` using the analytic minimizer of their LSE. We fit all other models using `ASLSD` with 2000 gradient iterations.

Fitted models Figure 5.37 displays fitted baselines. The `mtlh_semi_pc` fit displays a small baseline for US publications Figure 5.38 plots fitted kernels. As expected, the cross-

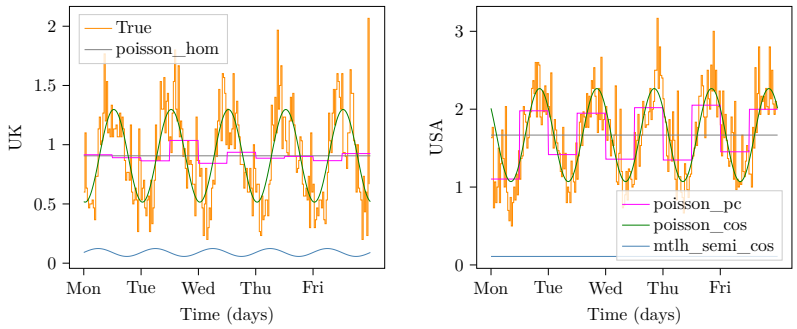


Figure 5.37: Fitted baselines (News 2D)

excitation from UK to UK publications is numerically negligeable. The cross-excitation from UK to US publications displays a clear mode around 5 hours. The self-excitation kernels are both monotonically decreasing and simlar, although the US self-excitation presents more mass around 3 hours, possibly due to the time difference between the East and West coast.

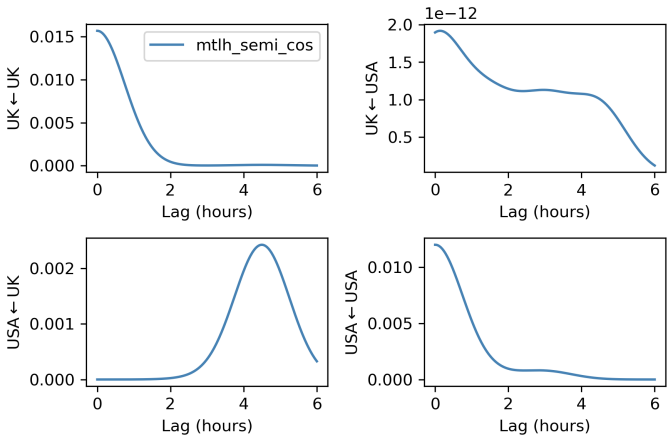


Figure 5.38: Fitted kernels (News 2D)

Part II

Hawkes models of Nasdaq equities prices

Chapter 6

The Nasdaq equities market

6.1 System overview

We now describe the system we are interested in modelling in Part II of this work: mid-prices on the Nasdaq equities market. An electronic equity market is composed of different market participants who can buy or sell shares of a company. In the US, electronic equity markets are managed by exchanges, such as Nasdaq or the New York Stock Exchange (NYSE). The Nasdaq stock market (shortened to Nasdaq), is operated by Nasdaq, Inc., and based in New York City. Therefore, in the rest of this work, all timings are in Eastern Standard Time (EST) unless otherwise specified.

6.1.1 Orders

On public electronic equities markets, each market participant (MP) can place orders on the exchange indicating their willingness to buy or sell shares of a ticker for a given price and quantity. In the market microstructure literature, orders are usually categorized into two simple types:

- Limit Orders (LOs) allow liquidity provision. These are orders to buy or sell with both a size and price specification. As there are not necessarily available counterparties when an LO is posted, it may not be executed immediately. To ensure the liquidity of the market, most exchanges have designated market makers with quoting requirements.
- Market Orders (MOs) allow liquidity taking. These are orders to buy or sell with a size specification only, seeking immediate execution at the best available price.

In practice, for major equities exchanges, all orders are LOs: an MO is simply an LO with a very aggressive limit price. Furthermore, the trading activity can be enriched by several other features of orders, called order attributes on Nasdaq. An LO rests in a Limit Order Book (LOB) until one of the following events occurs.

- The LO gets filled by an MO; *The LO can be filled totally by one MO, or incrementally by partial fills from different MOs.*
- The LO gets cancelled totally or partially by the MP who posted the LO;
- The LO gets cancelled totally by the exchange; *As we discuss below, there are different reasons why an exchange may eventually cancel an LO, based on the order attributes specified by the MP. We give two examples:*

  1. *The MP specified that the LO may not be stored for more than a given duration*[1]*. Once that duration is reached, the exchange deletes the LO totally.*
  2. *The MP specified that they do not accept partial fills. Assume that an incoming MO arrived on the market at some point, and that if the LO had not been totally deleted by Nasdaq, the LO would have been partially but not totally filled by the MO. In this case the LO is totally deleted by the exchange.*

MPs place orders on the exchange through different types of connectivity protocols. In Nasdaq, there are 4 proprietary connectivity protocols in addition to the international standard for exchange connectivity FIX (Financial Information eXchange): FLITE (FIX Lite), OUCH, RASH, and QIX. For a given exchange, each connectivity protocol leads to a different processing of orders, notably in how the order will be automatically repriced or modified during its lifetime.

6.1.2 Market schedule

Most electronic equity markets are usually closed on weekends and holidays. On a normal trading day, these markets are not open for 24 hours: the period for which the system is active and receives and posts orders is referred to as system hours. Market hours refer to a sub-period of time included in system hours, and pre-market (respectively post-market hours) refer to the time period before the start (respectively after the end) of Market hours[2]. This distinction is made because the rules of the system usually vary between market hours on the one hand, and pre-market and post-market hours on the other hand.

Figure 6.1 illustrates the Nasdaq market schedule on a normal trading day. The Nasdaq system opens at 4 AM and closes at 8 PM. Market hours start with an opening cross at 09:30 AM, and end with a closing cross at 4 PM. Despite designated times at which the opening and closing crosses should happen, in practice there exists an offset between the

[1] This duration is called the time-in force in the case of Nasdaq

[2] The term market hours can be ambiguous as one might think that no order executions can happen during pre and post-market hours: this is not true for all market centers. In Nasdaq, executions are possible outside of market hours.

announced times and the times at which the crosses actually happen. This offset depends on the activity on the matching engine, in Nasdaq rule 4701 paragraph (g) (see [73]) it is stated that "Nasdaq Opening Crosses for different System Securities occur sequentially rather than simultaneously". Nasdaq's announced intent behind the opening and closing crosses is to improve price discovery and mitigate the reaction to quarterly reports.

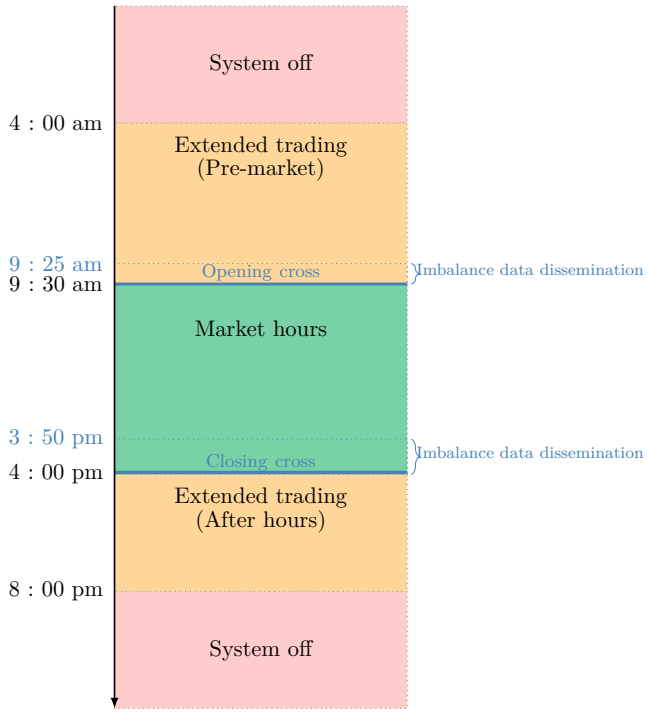


Figure 6.1: Daily market schedule on Nasdaq

Market schedule of the Nasdaq equities market on a normal trading day (without early closing).

6.2 Limit order books

A limit order book (LOB) stores the collection of active LOs. Limit prices can be specified on a discrete grid with constant increment size referred to as a tick. For liquid Nasdaq equities, the tick size is usually 1 cent (USD 0.01). Sizes of limit orders are also specified on a discrete grid with constant increment size referred to as a lot. For liquid Nasdaq equities, the lot size is equal to 1. Limit orders have many more attributes than their price and size, on which we do not focus in this study.

Priority queue In Nasdaq, LOs are posted and stored in the LOB following a price–display–time priority queue. This means that if an incoming buy (resp. sell) order is sent to the market, potentially with a limit price, the routing algorithm executes trades until exhaustion of the size or price of the incoming order using the ranking criteria:

1. LOs with more aggressive prices are executed first;
2. If LOs have the same price, visible orders are executed before hidden orders [3];
3. If limit orders have the same price and display type, the orders with earlier submission times are executed first.

Clock In the rest of this study, denote by $T = 23\,400s$ the total observation window over trading hours, that is, the number of seconds from market opening at $t_i :=$ 09:30 AM (EST) to closing at $t_f :=$ 04:00 PM (EST).

Price levels Fix a time $t \in [0, T]$. The set of limit orders to buy (resp. sell) displayed in the LOB at time t is called the Bid (resp. Ask) side. Consider a price level p, in ticks. We say that the price level p is occupied at the bid (resp. ask) side at time t if there exists at least one displayed buy (resp. sell) limit order posted at level p at time t.

Definition 6.2.1 (Price levels)**.** *Fix a time $t \in [0, T]$ and a positive integer index $i \in \mathbb{N}^*$. We denote by $p_t^{(B,i)}$ (resp. $p_t^{(A,i)}$) the i-th best Bid (resp. Ask) price at time t, that is, the unique price such that this price is occupied on the bid (resp. ask) side at time t, and there exists exactly $i-1$ occupied levels at the bid (resp. ask) side at time t for prices strictly larger (resp. smaller) than $p_t^{B,i}$ (resp. $p_t^{A,i}$).*

The price $p_t^{B,1}$ (resp. $p_t^{A,1}$) is called the bid (resp. ask) price. We denote the total displayed size posted at the i-th best bid (resp. ask) at time t by $s_t^{B,i}$ ($s_t^{B,i}$). The total non-displayed size posted is not a quantity that is accessible in the public Nasdaq data feed. The main object of interest of this work is the mid-price, defined as the arithmetic average between the bid-price and ask price $p_t := \frac{p_t^{A,1}+p_t^{B,1}}{2}$. The mid-price takes discrete values that change in half-tick increments.

6.3 Operating mechanisms

6.3.1 The principle of best price execution

A fundamental LOB design principle A fundamental operating principle for the majority of lit exchanges around the world is the prohibition of locking and crossing. We say that a system is locked (resp. crossed) if the best bid price equals (resp. is strictly greater than) the best ask price. Such a situation is sub-optimal for the Market participants owning all the locked or crossed orders as they would get price improvement by executing against

[3] An important attribute is that a limit order can be displayed (visible) in the order flow, or non-displayed (hidden) in the order flow. A non-displayed order is useful to mitigate one's market impact.

each other. This is why electronic markets are usually designed to avoid locking and crossing: any incoming order can receive price improvement and the matching engine checks whether it can execute against standing orders before posting to the LOB. In Nasdaq, this is explicit in Rule 4757 of the equity market, see [73]. This implies that the bid-ask spread ς_t is always strictly positive on a market that cannot be locked or crossed.

Inter-market locking and crossing In the US, locking or crossing a protected quote is prohibited by the American financial regulator, the SEC, in rule 600 of regulation NMS, see [94]. This also refers to inter-market locking or crossing stocks that are traded on different venues of different exchanges: we say that a quote is locked (resp. crossed) if there exists two markets such that the best bid price on a market is equal to (resp. strictly greater than) the best ask price in the other market. To enforce this rule, it is necessary for the regulator and exchanges to consider the notion of National Best Bid and Offer (NBBO) of a quote which is defined at any point in time[4] as the highest bid price and lowest ask price for the security over all registered exchanges where it is traded. Inter-market locking (resp. crossing) on registered exchanges and alternative trading systems is normally avoided by repricing or cancelling the locking (resp. crossing) orders, and/or routing them to the market that realizes the NBBO.

6.3.2 Life cycle of an order

We summarize the life cycle of an order on Nasdaq as follows:

1. The order is submitted to the system by an MP.
2. The order is processed according to its type and attributes to determine whether it may execute against any contra-side LOs[5] on the Nasdaq LOB.
3. If there are remaining shares, the rest of the processing of the order depends on the time-in-force (TIF) of the order, chosen by the participant. If the order has a TIF of immediate-or-cancel (IOC), then the order is cancelled and returned to the participant. If the order is not IOC, the price of the order might be adjusted depending on its type, attributes, the state of the LOB and the NBBO, and is either routed to a different venue or posted in the Nasdaq LOB.
4. Once an LO has been posted on the Nasdaq LOB, it can be later deleted or modified by the MP. A modification consists in modifying the attributes of the order, such as its price. The modified order will get a new timestamp and will be processed as a new

[4] Up to system latency.

[5] In the Nasdaq terminology in [73], a contra-side order refers to an order in the opposite of the market as the order considered.

order [6]. There is a cap on the number of times an MP can modify the order. The order can also be automatically adjusted by Nasdaq through time.

5. The order remains in the LOB until it gets executed, cancelled, or its TIF expires.

6.3.3 Why does the mid-price move?

A simple yet fundamental question is: what order book event is responsible for an observed mid-price jump? By construction, the mid-price jumps at a given time t as the result of:

- a uni-lateral move: the Ask price jumps but the Bid price doesn't; or the Bid price jumps but the Ask price doesn't.
- a bi-lateral price move: both the Ask price and the Bid price jump, and the sum of their signed jump sizes is non-null.

By construction, uni-lateral moves can occur because of

Scenario name	Trader action	Market conditions
LO in the spread	Submission of a LO at a price strictly between Bid and Ask prices.	Spread strictly greater than 1 tick.
OM in the spread	OM at a price strictly between Bid and Ask prices	Spread strictly greater than 1 tick.
Multi-level execution	Submission of a MO	MO entirely consumes at least one level of orders and no remaining size is posted.
Solitary deletion	Total deletion of LO at the best quote	Deleted LO was the only order at that quote.

A bi-lateral price move is a more peculiar event: it results from the modification of a solitary limit order to a more aggressive price, with execution of all contra-side orders at the first level.

6.3.4 Latency

In order to interpret event timestamps on the Totalview-ITCH data feed and how they may relate to each other, it is necessary to consider the latency experienced by different MPs. In the literature, the term latency may refer to different concepts related to the speed at which given market operations can be realized between MPs and exchanges. In this work, we follow Hasbrouck and Saar [45] and define latency as the total duration for

[6] unless the modification is a short sell specification.

a market participant to: receive information about a market event from Nasdaq; process that information; send an order to Nasdaq in response; and wait for Nasdaq to process this order. This definition of latency is sometimes referred to as the tick-to-trade.

6.3.4.1 Decomposing tick-to-trade

High frequency traders (HFTs) are a sub-class of algorithmic traders who leverage optimized hardware and software infrastructure to place orders with low latency. In order to reduce their latency, market participants can opt for co-location services offered by most lit exchanges, including Nasdaq: co-location consists in storing the market participant's servers and other hardware in a data center that belongs to the exchange [78]. It is well-known that HFTs account for a significant part of trading activity on equities markets: Carrion [20] report that 68% of traded volumes on Nasdaq equities are due to 26 HFTs. This emphasizes the necessity of understanding the orders of magnitude of HFTs tick-to-trade when reading Totalview-ITCH data [7]. Figure 6.2 illustrates connectivity between a co-located MP and Nasdaq.

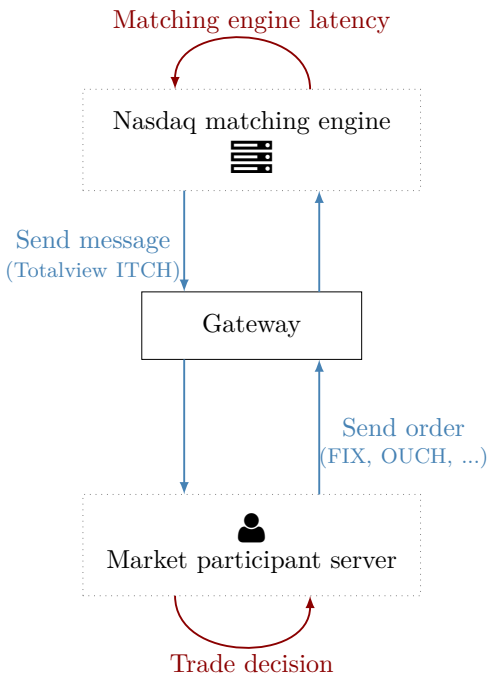


Figure 6.2: Tick-to-trade of HFTs

Connectivity with the exchange of a market participant co-located in the Nasdaq data center. In black, we draw hardware components in this network. Blue arrows draw data transfers, with the associated connectivity protocol between parenthesis. Red arrows draw data processing tasks.

[7] If the MP is not co-located, their latency corresponds mostly to the network transit time between their servers and the Nasdaq data center. Note that a lower bound for network transit time is imposed by the speed of light in vacuum, which gives an order of magnitude of $3.3\mu s/km$.

Matching engine latency To the best of our knowledge, the exact value of the Nasdaq matching engine latency at the time of the present work is not publicly disclosed. As noted by Baldauf and Mollner [12], it is well-known that exchange latency may vary depending on market conditions: the general and ticker-specific trading activity on Nasdaq, and the trading activity on other markets for order-routing purposes. However, in this commercial document [77] used by Nasdaq to promote their exchange solutions to other countries[8], Nasdaq states that their "technology features among the lowest latency in the industry at sub-40μs, with [Nasdaq's] fastest production implementation at 14μs latency door-to-door". It remains unclear how this order of magnitude compares with the latency of the Nasdaq engine in New York. In the literature, Hagströmer and Nordén [43] and Menkveld and Zoican [67] report an average processing time of 250μs for the INET trading system on Nasdaq OMX. While Nasdaq (in New York) is also build on INET, it is difficult to conclude whether values reported in this work for Nordic markets, in 2013, are comparable to the current engine latency on Nasdaq.

Order processing latency At the MP level, the operations that must be conducted are: handling the data feed; compute responses; and make order decisions. There are two paradigms within the HFT community. The first paradigm is software-based HFTs: these are MPs who use general purpose systems, notably processor based systems, for MP processing. For software HFTs, there exist non-compressible times added crossing the PCI bus, and a non-deterministic time added when interacting with the OS layer.

The second paradigm is hardware-based HFTs, sometimes referred to as ultra-high frequency traders: these are co-located MPs using integrated circuits to circumvent physical limitation of CPUs and GPUs. As discussed by De Schryver [29], while Application-specific integrated circuits (ASICs) offer very low latency and have become widely used for cryptocurrency mining, they suffer significant constraint for HFT applications: they are specialized by design, have long production times, and high unit costs. Therefore, Field Programmable Gate Arrays (FPGAs) have been progressively introduced in the finance industry since the early 2010s and, at the time of writing, are widespread on both the buy and sell side. FPGAs are re-configurable integrated circuits which can be programmed using simple logics. FPGAs do not involve the operating system nor need to cross the PCI bus, hence reducing their latency [9]. Nasdaq offers an FPGA version of their market data feed: Totalview-ITCH FPGA [80], which is proposed through the UDP version of ITCH 5.0

[8] Nasdaq, Inc. also offers technological solutions to other market places, notably to operate matching engines, such as for the nuam exchange (Chile, Peru, Colombia), or the Stock Exchange of Thailand.

[9] Therefore, a first large area of applications of FPGAs in finance is for high performance computing, notably MC simulations in derivatives pricing for instance (see Vanderbauwhede and Benkrid [103], Chapters 1 and 2)

(MoldUDP64). Pottathuparambil et al. [85] discuss the use of FPGAs to improve the speed of a Nasdaq Totalview-ITCH 4.0 feed handler, achieving a processing time of $2.7\mu s$, against $38\mu s$ for a CPU benchmark. Furthermore, they note that their FPGA processing time is deterministic and does not vary with time nor with in-bound feed activity, as opposed to the CPU solution. We expect this processing to be significantly lower in current implementations, notably because of the transition to ITCH 5.0 FPGA, and hardware evolution. This seems consistent with the latest research on feed handlers for data encoded in the FAST (FIX Adapted for STreaming) protocol. For instance, Tang et al. [96] achieve a 0.5–1.3μs processing time, and Dou et al. [32] even get to $0.447\mu s$.

Communication times The Nasdaq-MP and MP-Nasdaq communication times are unknown, and executed using different protocols. The Nasdaq-MP communication is done using the Totalview-ITCH protocol, potentially with an FPGA variant. The MP-Nasdaq communication is usually done through OUCH or FIX lite, as FIX is known to be slower.

6.3.4.2 Latency estimation

Latency still constitutes a major technological competitive advantage within the HFT industry, despite co-location solutions ending the arms race for exchange proximity. Therefore, HFTs (and algorithmic trading firms in general) do not disclose their typical latency values. Because of this, during the past 20 years, the study of latency in the market microstructure literature has been focused on theoretical problems rather than empirical studies. The significant theoretical interest in latency has focused on the study of latency arbitrage and front running in stylized LOB models through game theoretical approaches, allowing prescriptive conclusions for regulatory purposes: Cohen and Szpruch [26] derive the optimal latency arbitrage in a two–agents model and study the positive market implications of a Tobin tax; Menkveld and Zoican [67] uses a recursive model to study optimal strategies and equilibrium, with empirical applications on Nasdaq OMX; Baldauf and Mollner [12] studies order-to-trade latency, and suggests frequent bath auctions. Manahov [65] propose a computational approach by simulating a synthetic market model with HFTs, and also suggest auctions to improve market quality.

The empirical study of latency is relatively scarce because of the difficulty of access to very specific proprietary data. Even with exchange proprietary data, it is still not possible to compute the true tick-to-trade of MPs as their algorithmic processing remains unknown: when we see a message on the data feed, it is not possible to know with certainty what triggered the decision to send that message. Thus, researchers often prefer to engage in empirical research on the impact of HFTs on market quality rather than latency estimation.

For instance, Brogaard [18], Carrion [20], and Brogaard et al. [19] use the same Nasdaq proprietary dataset which contains trades and best quotes for 120 stocks traded on Nasdaq, with an indicator of whether the MP is an HFT or not. As noted by Carrion [20], this Nasdaq classification remains subjective, and misclassifications seem to bias their results [10]. Hasbrouck and Saar [45] use Totalview-ITCH data for 351 (resp. 399) Nasdaq tickers in October 2007 (resp. June 2008), and estimate the HFT latency to be 2–3 milliseconds. Using proprietary exchange data for 30 Nasdaq OMX tickers, Hagströmer and Nordén [43] estimate HFT latency around 100–600 (resp. 90–360) microseconds in August 2011 (resp. February 2012) using a proxy measure.

[10] Nasdaq only revealed anonymized MPIDs to researchers in the dataset.

Chapter 7

Dataset

To conduct the present work, we select LOB data from specific Nasdaq tickers over a specific time period. The tickers we select must satisfy certain criteria, and we allocate them in groups of 6 tickers each corresponding to their economic fundamentals. The group division we propose is finer than that of the Nasdaq screener, and allows us to study the impact of these fundamentals as well. In this section, we present our selection methodology, as well as the actual tickers composing the study.

Our market data is available through a third party, Lobster [62]. Lobster is a data provider of all LOBs of US equity markets available on the Nasdaq market data feed (Nasdaq TotalView-ITCH) for academic research. The LOB is reconstructed using the Nasdaq market data feed, Nasdaq TotalView-ITCH. The messages in the data feed are timestamped with nanosecond resolution during all system hours. The equities LOBs on Nasdaq TotalView-ITCH are not only the ones for stocks primarily listed on Nasdaq: it also contains stocks listed on NYSE and AMEX.

7.1 Data selection

In this work, we consider LOB data up to a depth of 5 levels for different tickers on different days. The study period and tickers must be chosen such that our results can be sufficiently generalized, at least qualitatively: specifically, we select liquid stocks which are representative of the different economic sectors, on a time period which is recent enough but not affected by very unusual market conditions.

7.1.1 Selection methodology

7.1.1.1 Time period

We want to use recent data while avoiding periods significantly affected by global crisis situations. We avoid data close to the Covid-19 pandemic starting January 2020, because

of the subsequent market crash and global recession. We also exclude data close to the Russo-Ukrainian war starting 24 February 2022, and the resulting energy crisis starting in the last quarter of 2022.

Therefore, we select data ranging from Wednesday 1 June 2022, to Friday 2 September 2022, inclusive. This represent 3 calendar months (14 weeks) of data, corresponding to a total of $n_{\text{days}} = 64$ full trading days. During this period, there are 2 holidays during which the Nasdaq exchange was closed Nasdaq [82]: Monday 20 June 2022, and Monday 4 July 2022. Over this period, some important market events must be taken into account for data processing

1. **Stock splits**: some firms increase the number of outstanding shares with a given ratio [81]. This means that a the moment of the stock split, the number of outstanding shares is multiplied by the split ratio, and the trading prices are divided by the split ratio. Among the tickers selected in this study (below), two are affected by a stock split in the considered period:

   - `AMZN`, with a $20:1$ stock split going into effect on 6 June 2022 [35].
   - `GOOGL`, with a $20:1$ stock split going into effect on Monday 18 July 2022 [36].

   None of the other tickers in the study underwent a stock split in the year 2022.

2. **Symbol changes**: Nasdaq symbols are not immutable. a firm can decide to change its symbol, for instance because the legal name of the company changed. In this study, it is the case for the company Meta (formerly Facebook), which was listed as `FB`, and changed to `META` on 9 June 2022 [89]. We only use the ticker symbol `META`.

7.1.1.2 Tickers

At the time of writing of this work, there are 3790 Nasdaq listed stocks across 3 market tiers: The Nasdaq Global Select Market; the Nasdaq Global; and The Nasdaq Capital Market. Nasdaq listed and traded symbols are available in a FTP directory refreshed every night [83]. This database is also accessible through the Nasdaq stock screener [79]. In the rest of this work, we refer to stocks using their official symbol.

We want to select $n_{\text{tickers}} = 48$ tickers split into 8 Groups of 6 tickers each. Each Group should be composed of stocks belonging to the same economic sector. This reference is distinct from the "sector" and "industry" fields in the Nasdaq stock screener data, because of the level of granularity of these fields. To select our 48 tickers, we only consider stocks satisfying the following heuristic criteria during the considered period.

1. **Survival:** up to the time of writing of the current work, the stock is not de-listed.

2. **Uninterrupted trading:** the stock's market was not halted.
3. **Large market capitalization:** the stock's total market capitalization exceeds USD 90 billion. *By market capitalization, we refer to the value of this field in the Nasdaq tickers dataset at the time of ticker selection. It it is unclear which definition Nasdaq uses for market capitalization.*
4. **High liquidity:** the stock's total available volume exceeds $800,000$ shares;
5. **Sensible price range:** the stock's average trading price lies between USD 25 and USD 600. *The tick size of all the stocks in this work is the same, and corresponds to one cent, therefore mid-price jumps are always in half a cent increments. This rule excludes, for example,* `BRK.A` *(Berkshire Hathaway Inc.), and* `UNH` *(UnitedHealth Group), which satisfy all the previous criteria.*
6. **Uniqueness of companies:** stocks must be emitted by distinct companies, and US based stocks must be class A Common Stocks. *This last rule aims to avoid capturing correlations between different classes of stocks from the same company. For instance, both* `GOOGL` *and* `GOOG`*, respectively class A common stock and class C capital stock of Alphabet Inc, satisfy all the other criteria.*

In our selection process, we dismissed a few largely traded tickers that were satisfying all the above selection criteria, such as

- Financial services providers such as `V` (Visa) and `MA` (Mastercard): these tickers are among the largest market capitalizations in the financial sector, however, most of the financial institutions represented in this study are banks. Therefore, we preferred to preserve the homogeneity in activity of that group.
- `TSLA` (Tesla): considered on the Nasdaq stock screener as within the Consumer Discretionary economic sector, its industry ("automobile constructor") is significantly different from the other groups of companies in a similar economic sector. This is the only automobile constructor satisfying the selection criteria.

7.1.2 Selected tickers

In the summary tables below, **Country** is the alpha-2 country code of the corresponding company. **Exchange** is the exchange on which the stock is listed: in our study, this will be either Nasdaq or NYSE. **Market cap** is the total market capitalization of the ticker in billion USD, as reported by Nasdaq in March 2023.

G1: Semiconductors From 2020 to 2023, this sector was affected by a global shortage of semiconductors, referred to as the global chip shortage (see for instance Voas et al. [105]). Table 7.1 summarizes the tickers in this Group.

Symbol	Company	Country	Exchange	Market cap (10^9 USD)	Mean price (USD)	Jumps rate (s^{-1})
AMD	Advanced Micro Devices Inc.	US	Nasdaq	128	90.20	11.60
AVGO	Broadcom Inc.	US	Nasdaq	249	522.42	5.36
INTC	Intel Corporation	US	Nasdaq	114	37.32	0.63
NVDA	NVIDIA Corporation	US	Nasdaq	526	169.75	16.94
TSM	Taiwan Semiconductor Manufacturing Company Ltd.	TW	NYSE	467	86.13	3.15
TXN	Texas Instruments Inc.	US	Nasdaq	159	166.37	7.31

Table 7.1: Summary of selected semiconductors companies.

There exists some fundamental differences in activity between these firms: AMD and NVDA do not manufacture their own chips themselves; INTC and TXN design and manufacture their own chips, in the US; and TSM design and manufacture their own chips, but not in the US. One direct implication of these differences is a disparity in how the CHIPS and Science Act [97] applies to each of these companies. This Act offers significant subsidies and investment tax credit to US based chip manufacturers, as part of the China-US trade war. This Act was particularly discussed in summer 2022, which lies in our dataset:

- On 18 July 2022, some US semiconductors manufacturers expressed their need for this Act, and some bills were set to be voted on 19 July 2022 [88];
- the Act was voted by the US senate on 27 July 2022, and by the US House of representatives on 28 July 2022;
- the Act was signed on 9 August 2022.

We have chosen not to include ASML in this economic Group, but rather include it in Computer manufacturing and equipment: ASML is the largest European technology company by market capitalization, but is a provider for semiconductor companies.

G2: Computer manufacturing and equipment Table 7.2 summarizes the tickers in this Group.

Symbol	Name	Country	Exchange	Market cap (10^9 USD)	Mean price (USD)	Jumps rate (s^{-1})
AAPL	Apple Inc.	US	Nasdaq	2,640	153.14	11.73
ASML	ASML Holding N.V.	NL	Nasdaq	263	518.92	2.06
CSCO	Cisco Systems Inc.	US	Nasdaq	209	44.72	0.68
IBM	International Business Machines Corporation	US	NYSE	122	135.24	3.03
QCOM	QUALCOMM Inc.	US	Nasdaq	142	139.28	10.67
TMUS	T-Mobile US Inc.	US	Nasdaq	182	138.57	2.86

Table 7.2: Summary of selected computer manufacturing and equipment companies.

In the Nasdaq stock screener, CSCO and TMUS are considered telecommunications company, but in the "telecommunications" sector the vast majority of firms are classified within the "Cable & Other Pay Television Services" industry. There is only one other "Telecommunications Equipment" company: VOD (Vodafone), which trades at a very low price and that was therefore excluded from the study.

G3: Computer software and internet services Companies developing computer software and/or providing internet services have particularly large market capitalization.

Table 7.3 summarizes the tickers present in this group.

Symbol	Name	Country	Exchange	Market cap (10^9 USD)	Mean price (USD)	Jumps rate (s^{-1})
ADBE	Adobe Inc.	US	Nasdaq	163	399.21	5.91
CRM	Salesforce Inc.	US	NYSE	165	176.31	8.08
GOOGL	Alphabet Inc.	US	Nasdaq	1,210	1084.44	8.20
META	Meta Platforms Inc.	US	Nasdaq	448	169.80	11.10
MSFT	Microsoft Corporation	US	Nasdaq	1,920	267.72	18.87
ORCL	Oracle Corporation	US	Nasdaq	235	73.34	1.78

Table 7.3: Summary of selected companies in computer software and internet services.

G4: Pharmaceuticals Pharmaceutical companies dominate the large tickers from the healthcare sector: among the 7 largest market capitalizations from this sector traded on Nasdaq or NYSE, 6 out of 7 belong to the sector of "Biotechnology: Pharmaceutical Preparations".

With the proximity of the COVID-19 pandemic, we note that 3 of these companies developed COVID-19 vaccines: AZN, JNJ, and PFE. MRK did not develop a COVID-19 vaccine, however, it helped manufacture and supply the vaccine of JNJ. Neither LLY nor NVO produced a COVID-19 vaccine, but developed drug treatments. Table 7.4 summarizes the tickers in this Group.

Symbol	Name	Country	Exchange	Market cap (10^9 USD)	Mean price (USD)	Jumps rate (s^{-1})
AZN	AstraZeneca PLC	UK	Nasdaq	214	65.37	1.13
JNJ	Johnson & Johnson	US	NYSE	419	172.44	3.59
LLY	Eli Lilly and Company	US	NYSE	312	315.07	1.11
MRK	Merck & Company Inc.	US	NYSE	278	90.01	1.62
NVO	Novo Nordisk A/S	DK	NYSE	319	109.77	0.89
PFE	Pfizer Inc.	US	NYSE	243	50.15	0.84

Table 7.4: Summary of selected pharmaceutical companies.

G5: Banks Banks dominate the "Finance" sector in NYSE. All the tickers in this group are listed on NYSE. Table 7.5 summarizes the tickers in this Group.

Symbol	Name	Country	Exchange	Market cap (10^9 USD)	Mean price (USD)	Jumps rate (s^{-1})
BAC	Bank of America Corporation	US	NYSE	283	33.59	0.55
HSBC	HSBC Holdings plc.	UK	NYSE	149	32.02	0.25
JPM	JP Morgan Chase & Co.	US	NYSE	417	116.82	5.78
MS	Morgan Stanley	US	NYSE	167	82.37	4.23
SCHW	The Charles Schwab Corporation	US	NYSE	150	66.95	2.87
WFC	Wells Fargo & Company	US	NYSE	182	42.49	0.96

Table 7.5: Summary of selected banking companies.

G6: Oil and gas The "Energy" sector in the Nasdaq stock screener is dominated by companies which industry is either "Integrated oil Companies" or "Oil & Gas Production". This is again a group of NYSE only tickers. Half of the tickers in this group are not American. Table 7.6 summarizes the tickers in this Group.

Symbol	Name	Country	Exchange	Market cap (10^9 USD)	Mean price (USD)	Jumps rate (s^{-1})
BP	BP p.l.c.	UK	NYSE	120	29.90	0.44
COP	ConocoPhillips	UK	NYSE	127	99.05	5.37
CVX	Chevron Corporation & Co.	US	NYSE	315	155.30	7.77
SHEL	Royal Dutch Shell PLC	NL	NYSE	211	52.69	1.58
TTE	TotalEnergies SE	FR	NYSE	153	52.39	0.82
XOM	Exxon Mobil Corporation	US	NYSE	454	92.05	6.29

Table 7.6: Summary of oil and gas companies.

G7: Retail distribution Table 7.7 summarizes the tickers in this Group.

Symbol	Name	Country	Exchange	Market cap (10^9 USD)	Mean price (USD)	Jumps rate (s^{-1})
AMZN	Amazon.com Inc.	US	Nasdaq	989	232.81	12.81
BABA	Alibaba Group Holding Ltd.	CN	NYSE	265	101.29	5.49
COST	Costco Wholesale Corporation	US	Nasdaq	225	509.49	6.13
HD	The Home Depot Inc.	US	NYSE	324	296.47	4.38
PG	The Procter & Gamble Company	US	NYSE	330	143.64	2.65
WMT	Walmart Inc.	US	NYSE	395	128.06	2.48

Table 7.7: Summary of retail distribution companies.

G8: Food and beverages Table 7.8 summarizes the tickers in this Group.

Symbol	Name	Country	Exchange	Market cap (10^9 USD)	Mean price (USD)	Jumps rate (s^{-1})
BUD	Anheuser-Busch Inbev SA	BE	NYSE	117	53.36	0.93
KO	The Coca-Cola Company	US	NYSE	260	62.82	0.86
MCD	McDonald's Corporation	US	NYSE	198	253.56	3.56
MDLZ	Mondelez International Inc.	US	Nasdaq	91	62.64	0.84
PEP	PepsiCo Inc.	US	Nasdaq	243	170.53	3.98
SBUX	Starbucks Corporation	US	Nasdaq	123	81.27	3.21

Table 7.8: Summary of food and beverages companies.

7.1.3 Context overview

We now give an overview of the features of the different tickers in this dataset. Figure 7.2 plots the empirical correlation between daily returns of the selected tickers over the selected period. For a given ticker, we define the (arithmetic) daily return[1] as the difference between the mid-price in the associated LOB at exactly 04:00 PM, minus the mid-price in the LOB at exactly 09:30 AM. This heat map shows some level of clustering around diagonal blocks, which indicates a higher similarity between stocks of a same economic Group.

7.1.3.1 Exchanges

Dual listings Certain securities are listed on two or more exchanges: this situation is usually referred to as dual listing [76]. In the case of US equities markets, Nasdaq launched its dual listing program in January 2004 [72], [98]. According to Nasdaq, the objective of this program is to attract large corporations listed on NYSE by allowing them to reach more trade opportunities, while receiving lower general fees than if they were primarily listed on Nasdaq. This constitutes an incentive offered to these firms to eventually be primarily listed on Nasdaq (see Hegde et al. [50], and [93]). The arrival of RegNMS in 2005 regulates dual listings, by ensuring markets of dual listed securities do not lock nor cross.

[1] Sometimes, daily returns are defined as the difference between the closing price and opening price, which are themselves sometimes defined as the price of the last and first transaction of the day. It is clear that this definition does not necessarily coincide with ours.

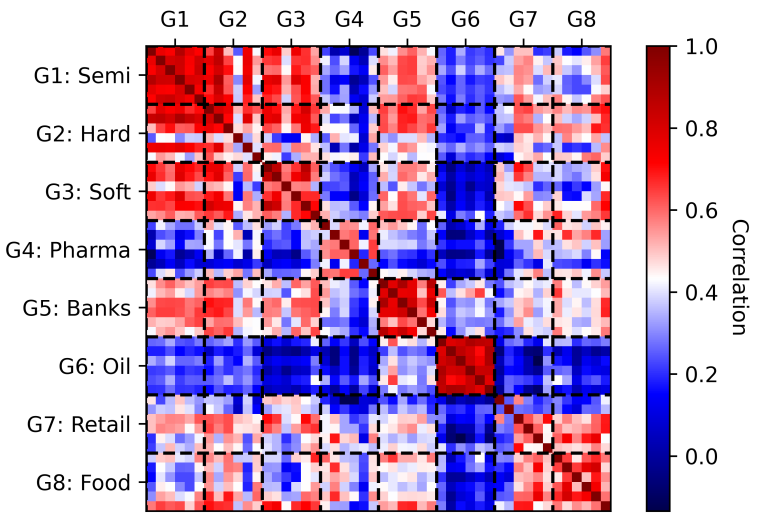


Figure 7.1: Empirical correlation of daily returns

Tickers in the rows and columns are sorted by economic group first, then in alphabetical order of symbol. To clarify visualization, dashed black lines delimit the different Economic Groups. Each unit in the heat map displays the empirical correlation of daily returns between the tickers of the corresponding row and column.

Representation in the dataset Among the 48 tickers we selected, 20 are primarily listed on Nasdaq, and 28 on NYSE. This is because Nasdaq listed firms are predominantly technology companies. Therefore, selected tickers are biased towards NYSE primary listings.

This raises the question of the importance of the exchange in the daily returns correlations we observed previously is due to the primary listing exchange. In Figure 7.2, we permutate the rows and columns of Figure 7.2 by exchange: this shows that the exchange does indeed play a role, but that there is still correlation between Nasdaq and NYSE tickers.

7.1.3.2 Countries

Non-US firms can be listed on Nasdaq and NYSE if they comply with local regulations. Non-US companies can issue securities which are traded exclusively on US exchanges in USD and paying dividends to their holders in USD, through two types of contracts. The first is the American Depositary Receipt (ADR) [74], a negotiable ownership certificate emitted by a US depository bank, representing a multiple of (ordinary) shares of the foreign company. Note that American Depositary Shares (ADS) [75] refer to the actual tradable asset. The second is New York Registry Shares (NYRSs) [25]: USD denominated shares which do not require depository receipts from a US bank. In practice, this is an ad-hoc contract for Dutch

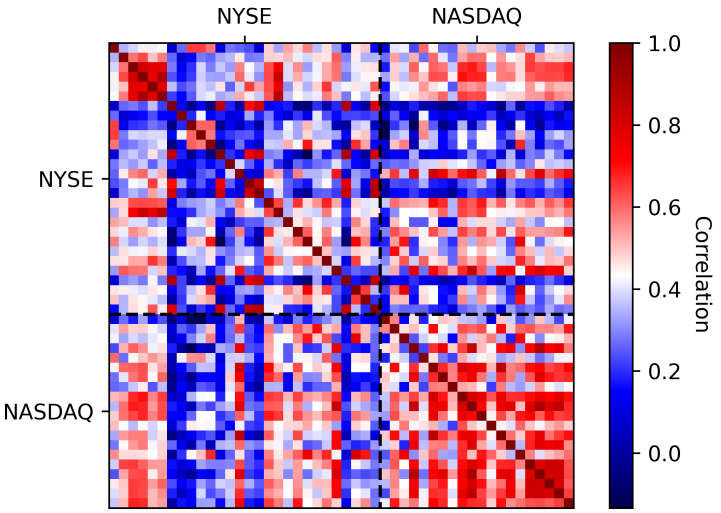


Figure 7.2: Empirical correlation of daily returns, sorted by exchange

Tickers in the rows and columns are sorted by exchange first, then by economic group, then in alphabetical order of symbol. To clarify visualization, dashed black lines delimit the NYSE and Nasdaq tickers. Each unit in the heat map displays the empirical correlation of daily returns between the tickers of the corresponding row and column.

companies. 38 (about 80%) of the tickers in the dataset are US companies. The 10 non-US tickers are from 7 countries: Belgium, Denmark, France, Great Britain (3 tickers), and the Netherlands (2 tickers) in Europe; and China and Taiwan in Asia.

7.1.3.3 Indices

The fact that a security belongs to a major index is often regarded as a significant feature impacting how the security is traded. For instance, a direct consequence of a stock belonging to an index is that the portfolio of ETFs tracking that index will have to be re-balanced to reflect changes in the underlying index.

Index	S&P 500	Nasdaq 100	DJIA
Count	38(7.6%)	28(28%)	15(50%)

Table 7.9: Representation of the main US indices in the dataset.

For each index, we count the number of tickers in the dataset which are within the components of said index. Between parenthesis, we give the ratio of index components in the dataset divided by the total number of components, as a percentage.

S&P 500 Because we do not impose any domiciliation condition, some of the companies emitting these stocks are not primarily listed in the US, or do not derive the majority of

their revenue in the US. In our dataset, the only stocks that are not included in the S&P 500 index are the 10 tickers which are not primarily listed in the US.

Nasdaq 100 In our dataset, the tickers which are not part of the Nasdaq 100 index are exactly those primarily listed on NYSE.

DJIA The Dow Jones Industrial Average (DJIA) is composed of 30 tickers. 15 of these 30 constituents of the DJIA in summer 2022 are present in the dataset.

7.2 Pre-processing messages

Messages on Lobster and Nasdaq TotalView-ITCH are timestamped up to nanosecond precision. In the Lobster data, we observe several sequences of messages with the same timestamp. Because of the hardware and data processing constraints on the matching engine, and the scale of market activity, it seems extremely improbable that messages occurring at the same timestamp correspond to distinct orders arriving at the engine within the same nanosecond. Therefore, our interpretation of simultaneous messages is that they correspond to automatic order adjustments. However, it is not specified in the Nasdaq documentation whether automatic order adjustments are simultaneous, or if there is a processing delay that depends on the activity in the matching engine. In this section, we classify and interpret the different types of sequences of simultaneous messages we observe in Lobster data. We hope a more detailed study of this system will help guide the modelling phase.

The problem of message collisions If a given event happens on the market, this might trigger an automatic sequence of adjustments of orders with specific order types and attributes. This type of mechanism might reduce the quality of order-book models, particularly when order adjustments are made because of events that are unobserved in the Nasdaq market data feed: for instance, if the NBBO moves, because of the Nasdaq inside quotes or because of activity in a market center other than Nasdaq.

We analyse such sequences for the 5 first levels of the LOBs of the ticker AAPL, on 5 days of trading (3 February 2020 to 7 February 2020). The total number of such sequences in this period is $127,961$; their length is small on average with a mean of 2.42 messages and a standard deviation of 3.26. In fact, $106,263$ sequences (83.04%) are of length 2, but 7.06% of these sequences have a length larger than 4 with a maximum length of 660. The non-zero inter-arrival times during these five days are strictly superior to 100 nanoseconds. This, in addition to normal orders of magnitude of market latency for the fastest market

participants, suggests that it would be very unlikely that messages timestamped at the same nanosecond would results from orders sent by different market participants.

Multi-order executions and order modifications Given the format of Lobster messages, we expect the majority of such sequences to come from two phenomena. The first is order modifications: when an order is modified, this would result in two messages in the data feed with the same timestamp. The first one should be a cancellation, and the second one should be a submission of a new LO on the same side of the market. $61,996$ (48.44%) sequences have a structure similar to an order modification: they are sequences of two messages, first a cancellation, then a submission of an LO, both on the same side.

The second is MOs executing several orders. Since the messages are given from a limit order perspective, a market order that executes several LOs will result in several execution messages with the same timestamp on the same side of the market. Of course, if some or all of these executions affect hidden orders, then it is not possible to observe if they are on the same side. $62,251$ (48.66%) sequences have a structure similar to an MO executing several orders.

Crosses We focus on the remaining $3,714$ (2.90%) sequences. In Lobster data, there is a specific message disseminated when a cross happens. This message has its own event type indicator; its price corresponds to the crossing price and its volume to the total traded volume. During the 5 days considered, each opening cross message is included in a sequence of simultaneous messages. We observe two different structures among these messages:

- Three sequences start with the opening cross trade message. The rest of the simultaneous sequence is formed of submissions of LOs. Such a behaviour is coherent with the fact that limit-on-close (LOC) orders, which are orders that do not interact with the continuous trading book prior to the opening cross, can be posted to the continuous trading book if they are not crossed and the participant requested it. It is not specified in Nasdaq documentation how the submission of these orders is handled by the system, but a simultaneous submission of these non-crossed LOCs would be coherent with this behaviour.
- Two sequences start with multiple executions of limit orders, then a cross trade message and finally several submissions of limit orders. We do not have an interpretation of the execution messages yet.

Among the five closing cross trade messages, only two are part of sequences of simultaneous messages. They both have the same structure: they start with multiple executions of LOs,

then a cross trade message. We do not have an interpretation of the execution messages yet.

Submissions Among the 3707 remaining sequences, the majority contain at least one LO. In this paragraph, we focus on the 3254 such sequences. These sequences can be classified as follows:

- There are only 14 sequences with two LO submissions or more. These sequences contain exclusively LO submission messages, happen between 9:30:00 AM and 9:30:10 AM, and these limit orders have the same direction. We do not have an interpretation yet of these sequences.
- There are $3,240$ sequences with exactly one LO submission message. We observe that for all these sequences, the LO submission message is at the end of the sequence. We can classify these sequences into the following categories:
  - $3,139$ of these sequences contain only execution messages and a submission message. For each of the $1,951$ sequences with at least one execution of a visible LO (*i.e.* sequences with executions that do not consist only in executions of hidden orders), all executed visible LOs have the same direction and the limit order is always posted on the opposite direction. We interpret this behaviour as the arrival of an order that executes several LOs and stops walking the book because of its own limit price.
  - There are 25 sequences containing only cancellation messages and exactly one submission message. All cancelled LOs have the same direction. For 7 of these sequences, the submitted LO has the same direction as the cancelled LO, whereas for the other 18 sequences the submission is on the other side of the market.
  - 101 sequences are composed of at least a cancellation message and an execution message, and ended by a submission message.

Remaining sequences The last 453 remaining sequences are exactly the sequences composed of at least one cancellation message, and potentially execution messages, but no submission nor cross trade.

- Only 3 sequences contain only cancellations. These sequences don't seem to correspond to an automatic cancellation of many orders triggered by the expiry of their time-in-force (these sequences are timestamped between 10 AM and 2 PM). Nonetheless, we do notice that all cancelled LOs in a given sequence have the same direction.

- 450 sequences contain at least one execution message and at least one cancellation message and no other message type. Note that in these sequences, if the LOs had not been cancelled, they would have been executed following price-time priority rules. This behaviour is consistent with cancellations to prevent self-execution.

Summary of event types We pre-process Lobster data to define the following event types based on simultaneous Lobster messages. A **trade** is any sequence of simultaneous messages which contains at least one execution of a LO. A **deletion** is a cancellation message that is not simultaneous with any other message. A **modification** is a simultaneous sequence of one cancellation message followed by one LO submission message on the same side of the market. A **liquidity provision** event is a submission of one and only one LO.

Chapter 8

Empirical analysis of mid-prices

In this chapter, we discuss some statistical properties of mid-prices in the dataset, which guide our modelling choices. This empirical analysis is based on our dataset of 48 tickers selected, observed for 64 days, during which a total of $3,344,673,42$ mid-price jumps occurred.

In order to keep this work concise, it is not always possible to plot data for each ticker and each day in this study: for example, with empirical distributions. In such cases, we display data fom 1 week (5 trading days), from Monday 18 July 2022 to Friday 22 July 2022, for 4 tickers: `AAPL`, `AMZN`, `INTC`, and `JPM` (see Table 8.1).

Symbol	**Name**	**Exchange**	**Market cap** (10^9 USD)	**Average sparsity**	**Average value** (10^3 USD)	**Mean price** (USD)	**Event rate** (s^{-1})	**Jumps rate** (s^{-1})	**Min time** (ns)	**Max time** (s)
`AAPL`	Apple Inc.	Nasdaq	2,640	0.42	870	152.23	109.85	12.80	154	15.6
`AMZN`	Amazon.com Inc.	Nasdaq	989	1.91	446	120.15	62.93	11.08	169	12.8
`INTC`	Intel Corporation	Nasdaq	114	0.01	1,462	39.83	40.44	0.70	200	108.5
`JPM`	JP Morgan Chase & Co.	NYSE	417	1.40	219	114.04	27.79	5.98	168	21.7

Table 8.1: Sample data

Event rate is the ratio of total number of LOB events over the considered period, divided by the total observation duration in seconds. **Jumps rate** is the ratio of total number of mid-price jumps over the considered period, divided by the total observation duration in seconds. **Min time** is the average of daily minimum inter-arrival times of mid-price jumps, in nanoseconds. **Max time** is the average of daily maximum inter-arrival times of mid-price jumps, in seconds.

For each ticker, and each trading day, denote by N^1 (resp. N^2) the process counting the number of upward (resp. downward) mid-price jumps up to time $t \in [0, T]$.

8.1 LOB mechanisms

We start by investigating the types of LOB events that make the mid-price change. As discussed above, not all LOB events cause mid-price jumps. In Table 8.1 we see that the rate of LOB events, the rate of mid-price jumps, and the ratio of mid-price jumps to LOB events vary significantly across tickers. However, this ratio seems relatively stable through days for each ticker. In particular, `JPM` has consistently the highest ratio value with the

smallest number of LOB events; and INTC has consistently the lowest ratio with the smallest market capitalization.

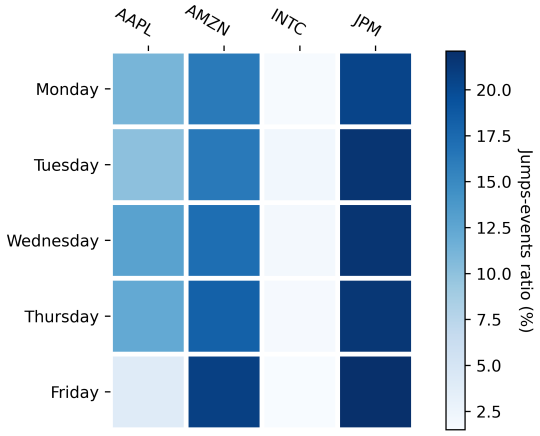


Figure 8.1: Jump-events ratio.

We are now interested in the distribution of event types. Figure 8.2 plots our results. Before diving into their interpretation, we see that when fixing a stock, the distributions considered are stable through time. However, across stocks, we see consistent behaviour between AAPL, AMZN, and JPM, with INTC acting very differently.
As expected, **Limits** are the primary cause of mid-price jumps for all tickers. However, a first counter-intuitive result for AAPL, AMZN, and JPM, is that **Deletions** are the secondary mechanism of jumps, and significantly more predominant than **Trades**. The case of INTC is quite different, when restricting to jump times only, we note that that proportion of posted **Limits** is the same as for the other tickers, but **Trades** are significantly more predominant than **Deletions**. We observe that LO submissions are the primary mechanism of mid-price moves. By construction, a **Limit** event can only make the mid-price jump if the order is posted in the spread. Therefore, this can only happen if the spread is higher than half a tick. Surprisingly, solitary deletions are the secondary mechanism of mid-price jumps for AAPL, AMZN and JPM, with over 30% of jumps.

8.2 Inter-arrival times

We now study the distribution of inter-arrival times of mid-price jumps at the high-frequency level. For each ticker and each day of the dataset, consider the sequence of times at which the mid-price changes. By inter-arrival times, we refer to the difference between consecutive times at which the mid-price changes. Inter-arrival times of mid-price jumps are linked to the latency of market participants. Figure 8.3, we plot systematically two vertical lines:

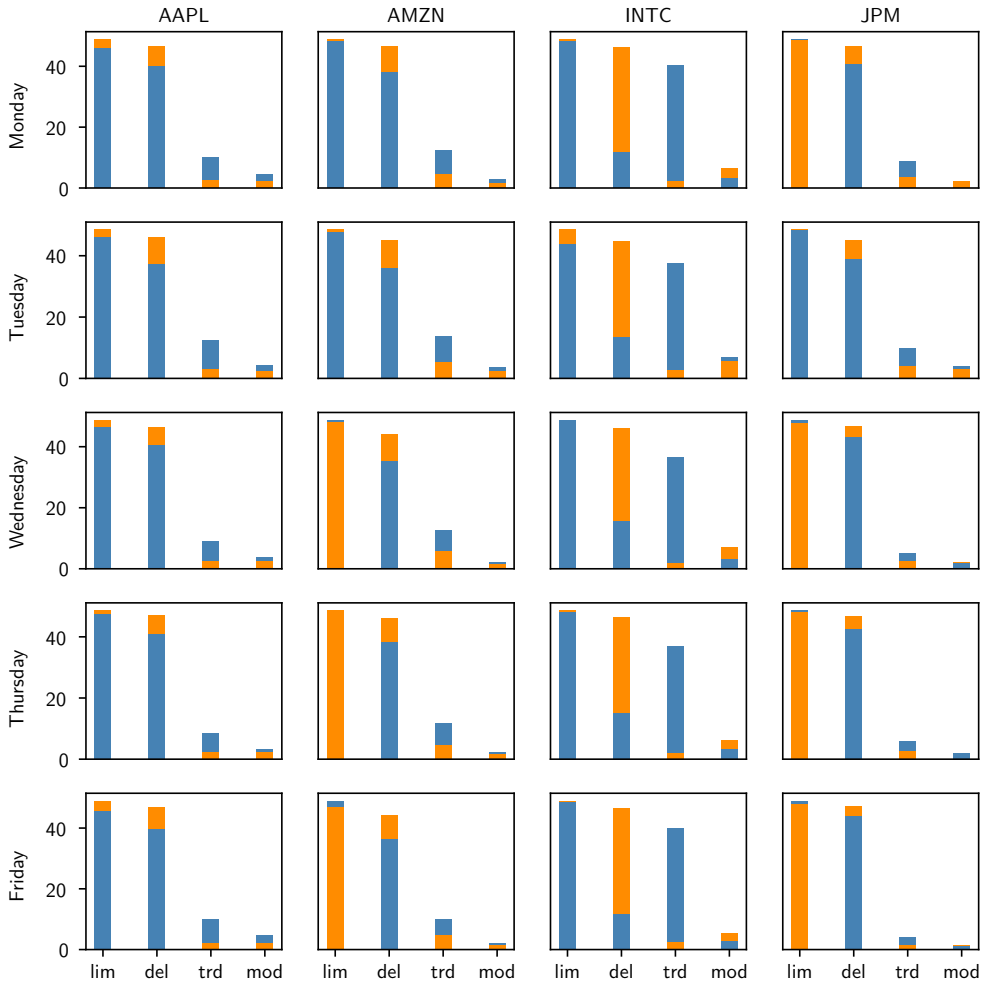


Figure 8.2: Distribution of event types

Distribution of LOB event types (as percentages): for all events (orange), and only to events corresponding to a mid-price jump (blue).

one at $12.5\mu s$, and one at $22.5\mu s$; there are consistently two modes present in the previous distributions which are close to these fixed times.

8.3 Jump size

While the processes N^1 and N^2 count the number of mid-price jumps for a given ticker on a given trading day, mid-price jumps do not all have the same size. In this work, we refer to mid-price jumps of more than half a tick in absolute value as leaps. In Section 8.3.1, we discuss empirical properties of mid-price jump sizes in general: Section 8.3.2 explores the relationship between jump sizes and trading price; and Section 8.3.3 analyzes the bias introduced by the `AMZN` and `GOOGL` data before their stock splits. Section 8.3.4 is focused

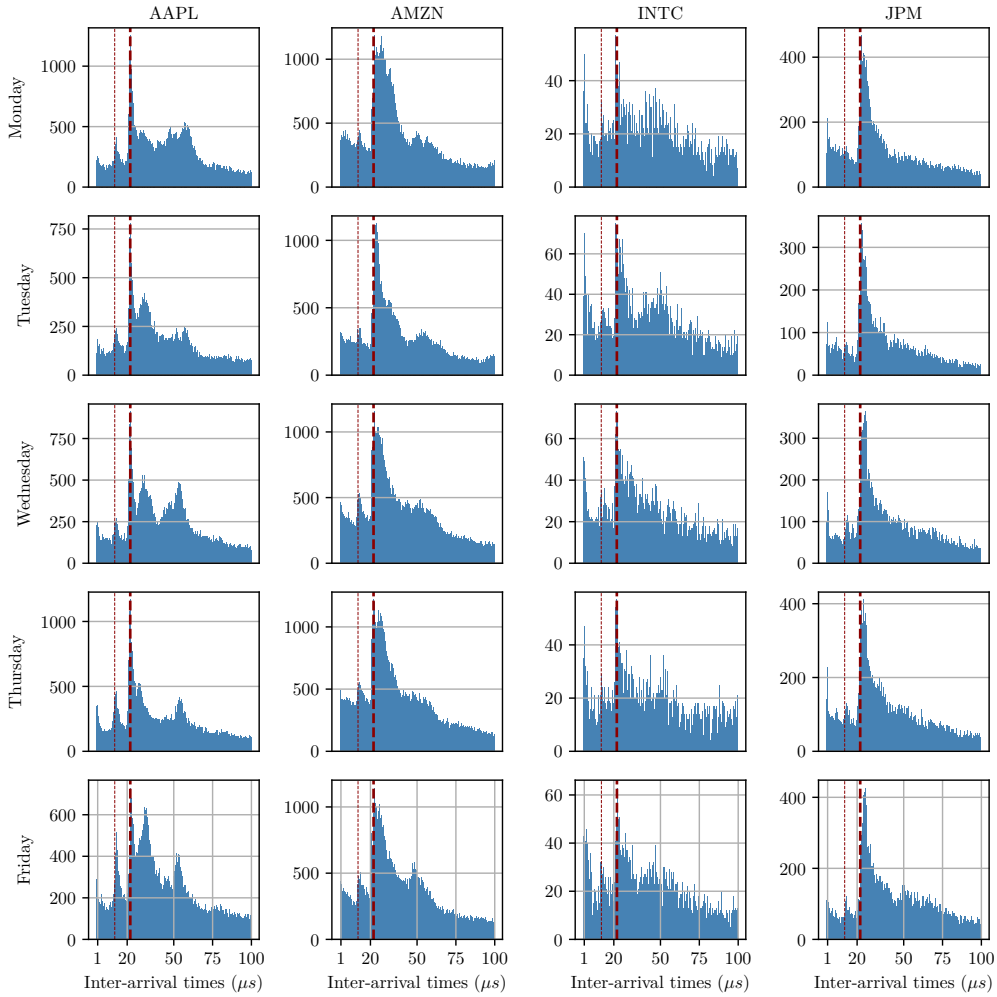


Figure 8.3: Histogram of order inter-arrival times for some Nasdaq stocks.

We plot the histogram of inter-arrival times of mid-price jumps below $100\mu s$ for each ticker on each date with μs resolution.

on empirical properties of leaps.

8.3.1 Overview

Distribution of jump sizes Mid-price jumps are usually small: the empirical average (over all tickers and trading days) of the absolute value of all the jumps observed in this study is 0.819 ticks. As shown in Figure 8.4, 84.79% of observed jumps are of half a tick in absolute value, while only 3.82% of all observed jumps are of 2.5 ticks or more. However, these numbers might be misleading when looking at a specific stock. As we see in Figure 8.5, this is because tickers with larger leap probabilities have smaller range of activities (orders per second).

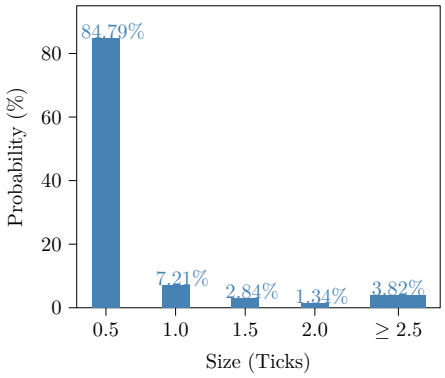


Figure 8.4: Empirical distribution of absolute jump sizes

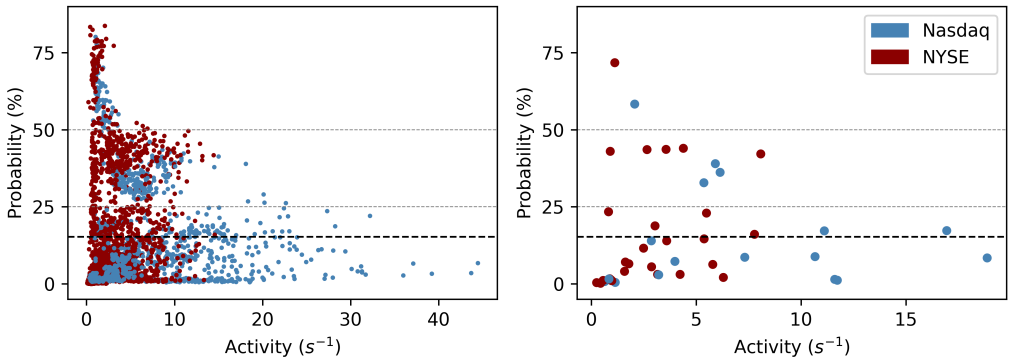


Figure 8.5: Leap probability and activity

Empirical probability of leaps against total activity η for each ticker and each day (left), and for each ticker averaged on all trading days (right). Solid dashed line plots the average leap probability over all tickers.

Time dynamics Figure 8.6 plots the average jump size and leap probability over all tickers against time. As we discuss in Section 8.3.3, the noticeable decreases in average jump size are mainly due to two stock splits.

8.3.2 Relationship between jump sizes and price

We now consider the relationship between the average jump size δ (in ticks) and the time-averaged price (in USD)

$$\langle\!\langle p \rangle\!\rangle_T := \int_0^T p_t \mathrm{d}t. \tag{8.1}$$

A relationship is unsurprising, as our stocks have a common tick size independently of their price. Figure 8.7 shows that average jump size increases with the average price in a convex fashion; with different regimes depending on the primary listing exchange; and that this relationship is better described by an exponential model than by a power law overall (see Table 8.2 and Table 8.3). To fit the relationship between size and price, we consider an

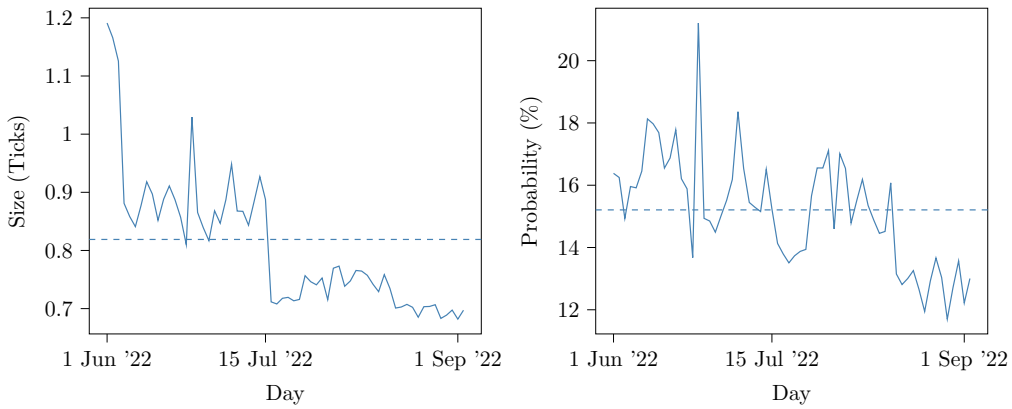


Figure 8.6: Temporal dynamics of jump sizes

Left: Empirical daily mean of jump sizes over all tickers against time, dashed line plots the empirical mean over the time period. Right: Empirical daily mean of leap probabilities over all tickers against time, dashed line plots the empirical mean over the time period.

exponential model and a power law model of the form

$$\delta := c + b \exp\left(a \langle\!\langle p \rangle\!\rangle_T\right), \quad \delta := c + b \langle\!\langle p \rangle\!\rangle_T^a. \tag{8.2}$$

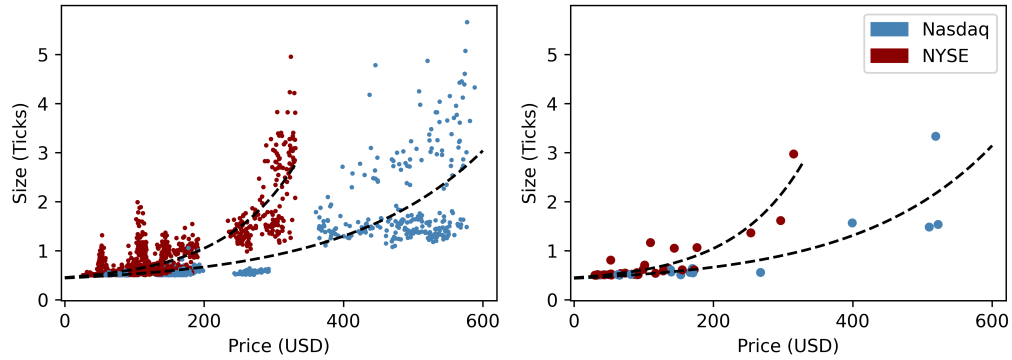


Figure 8.7: Average jump size against time-averaged price

Left: Each point plots the average jump size against the time-averaged price for a given ticker on a given day. Right: Each point plots the average jumps size against the time-averaged price for a given ticker, averaged on all trading days.

Exchange	**Exponent** a ($\times 10^{-3}\text{USD}^{-1}$)	**Scale** b (Ticks)	**Bias** c (Ticks)	L_2 **loss**
Nasdaq	5.097(5.212)	0.128(0.124)	0.314(0.313)	0.177(0.128)
NYSE	9.445(9.804)	0.106(0.0977)	0.347(0.351)	0.0666(0.0419)

Table 8.2: Size–price fit (exponential).

Exchange	**Exponent** a (USD^{-1})	**Scale** b ($\times 10^{-7}$ Ticks)	**Bias** c (Ticks)	L_2 **loss**
Nasdaq	2.321(2.275)	8.124(10.95)	0.5(0.5)	0.177(0.128)
NYSE	2.725(2.441)	2.966(14.87)	0.5(0.5)	0.0691(0.0472)

Table 8.3: Size–price fit (power law).

The increase of average size with time-averaged price results from the increase of leap probabilities with time-averaged price. Figure 8.8 plots this relationship.

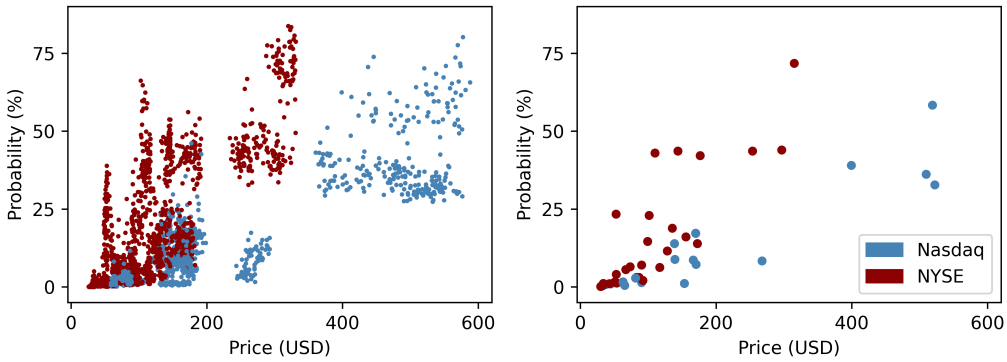


Figure 8.8: Leap probability and price

Left: Each point plots the average leap probability against the time-averaged price for a given ticker on a given day. Right: Each point plots the average jumps leap probability against the time-averaged price for a given ticker, averaged on all trading days.

8.3.3 The effect of stock splits

The previous results are biased by the dynamics of `AMZN` and `GOOGL` before their respective stock splits. Figure 8.9 plots time-average prices against trading day for these two stocks, and benchmarks them against the empirical mean of the time-averaged price over all other tickers. We see clearly that the time-averaged price of `AMZN` and `GOOGL` before split is significantly higher than the average of the other tickers, and has a comparable order of magnitude after split. Figure 8.10 shows that the average absolute jump size increases in the days preceding a stock split.

8.3.4 Leaps

Leaps occur through different order book mechanisms: they can correspond to a trade executing multiple levels of the LOB, a trade or solitary deletion with a sparse LOB, or an LO submission when the spread is more than half a tick. Figure 8.11 plots the temporal dynamics of the proportion of leaps caused by each event type. For `AAPL` and `AMZN`, leaps

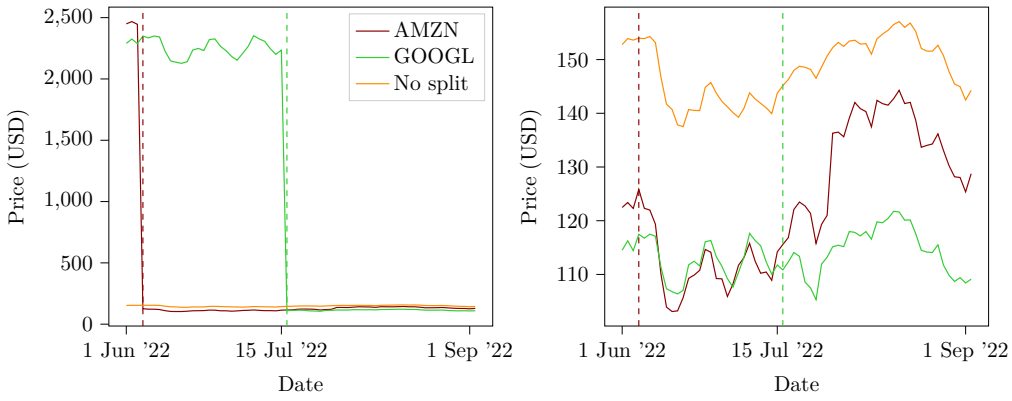


Figure 8.9: Stock splits and price

Time-averaged price against trading day. Red (resp. green) line corresponds to `AMZN` (resp. `GOOGL`), vertical dashed lines mark stock split dates. Orange line corresponds to the empirical mean of time-averaged price on all other stocks. Left figures plots prices as is, right figure adjusts prices by dividing them by their respective split factor.

result predominantly from trades and liquidity provision. For most days, the relative share of solitary deletions among leaps stagnates through the day: this is consistent with the fact that for solitary deletions to cause leaps, the first and second price level should not be contiguous, which is more likely near market opening. For `JPM`, the contributions of each event type are close and have very similar temporal dynamics. On `INTC`, leaps are almost exclusively caused by trades. Finally, the contribution of modifications is negligible for all tickers.

8.4 First order moments

8.4.1 Total cumulative jump rate

We are interested in the total cumulative event rate $\boldsymbol{\eta_T} := (\eta_T^1, \eta_T^2)$. This rate summarises the total price activity at a first order level. For each ticker, Figure 8.12 plots the average over all trading days of (η_T^1, η_T^2). Again, we see that the total event rate is symmetric, and Nasdaq tickers usually have higher activity than NYSE tickers. Figure 8.13 plots the empirical distribution of jump sizes for a smaller subset of data. These distributions have small Fisher-Pearson coefficient of skewness, with a fast decay.

8.4.2 Empirical intensity

Figure 8.14 plots the empirical intensity of mid-price jumps with sampling period $h = 10$ minutes. There is consistently a peak in the jump rate located at the first hour of trading. The first hours of the day constitute a decay phase for the jump rate, followed by a phase where the jump rate fluctuations are small. For some tickers and during some trading

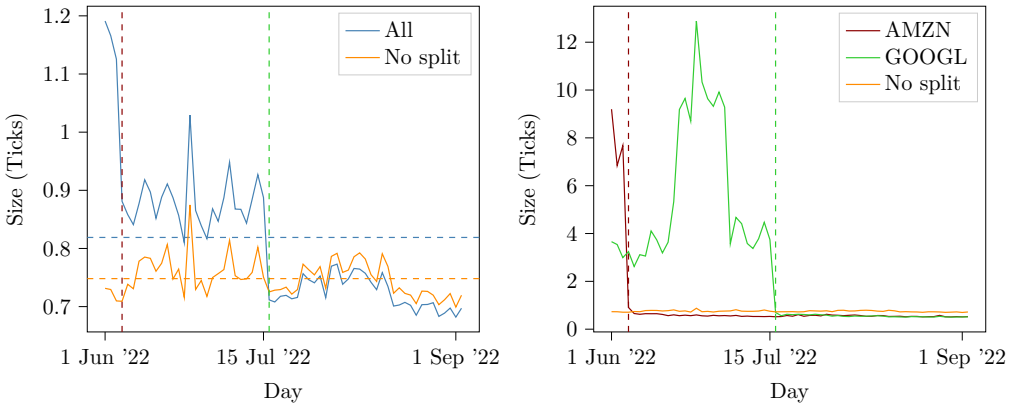


Figure 8.10: Stock splits and jump size

Left: empirical mean of average jump size against trading day over all tickers (blue) and over all tickers excluding `AMZN` and `GOOGL` (orange). Right: Red (resp. green) line plots the average jump size for `AMZN` (resp. `GOOGL`), and vertical dashed lines plot the stock split dates.

days (for example `AAPL` on Monday, `JPM` on Wednesday), there exists a third phase in the afternoon, where the jump rate increases again and decays rapidly to another stable regime before closure. We now look at the contribution of different LOB mechanisms to the number of mid-price jumps through the day. Figure 8.15.

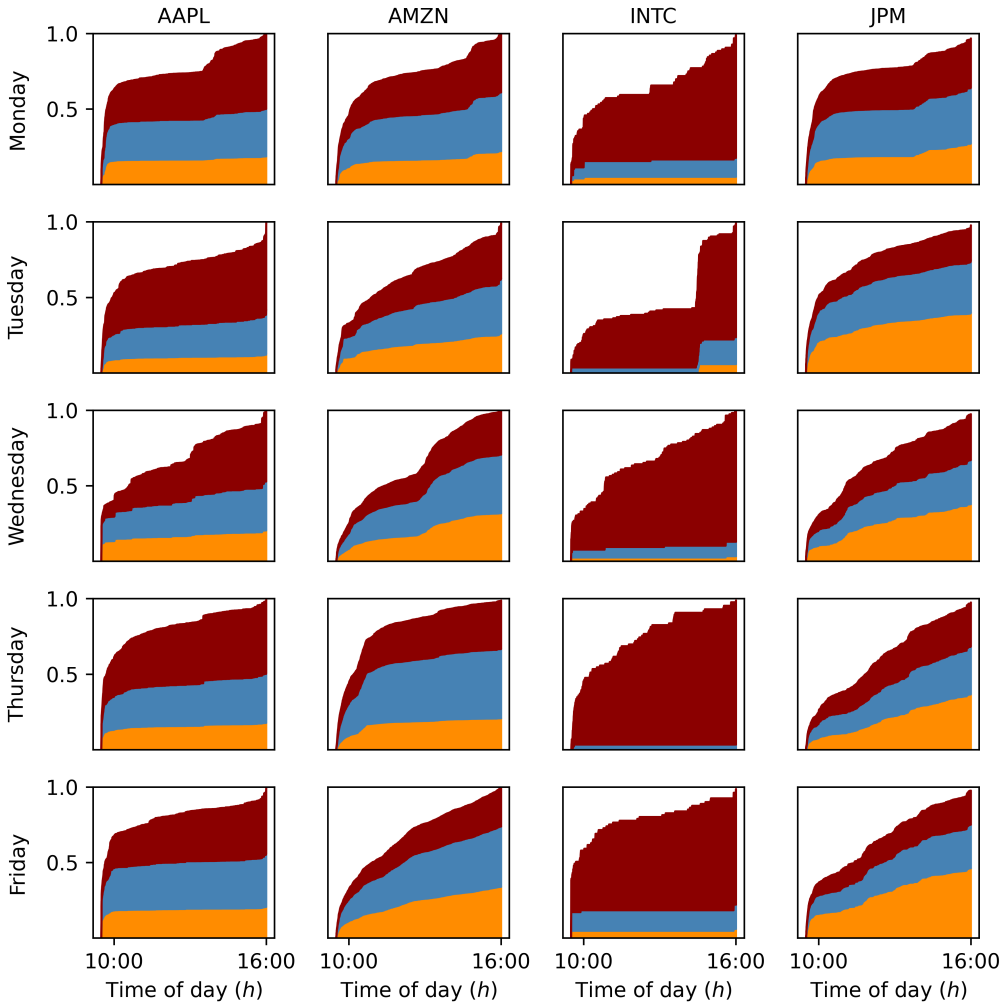


Figure 8.11: Event types of leaps

For each ticker and trading day, ratio of deletions (orange), liquidity provision (blue), and trades (red) divided be the total number leaps, against time of the day.

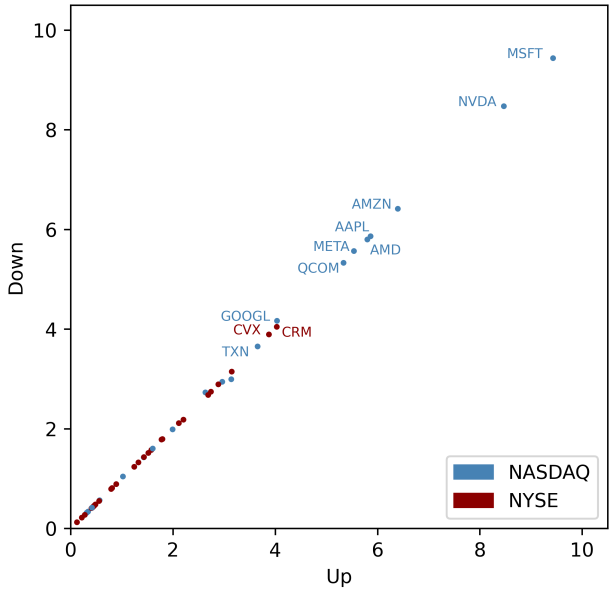


Figure 8.12: Average upward and downward jump rates

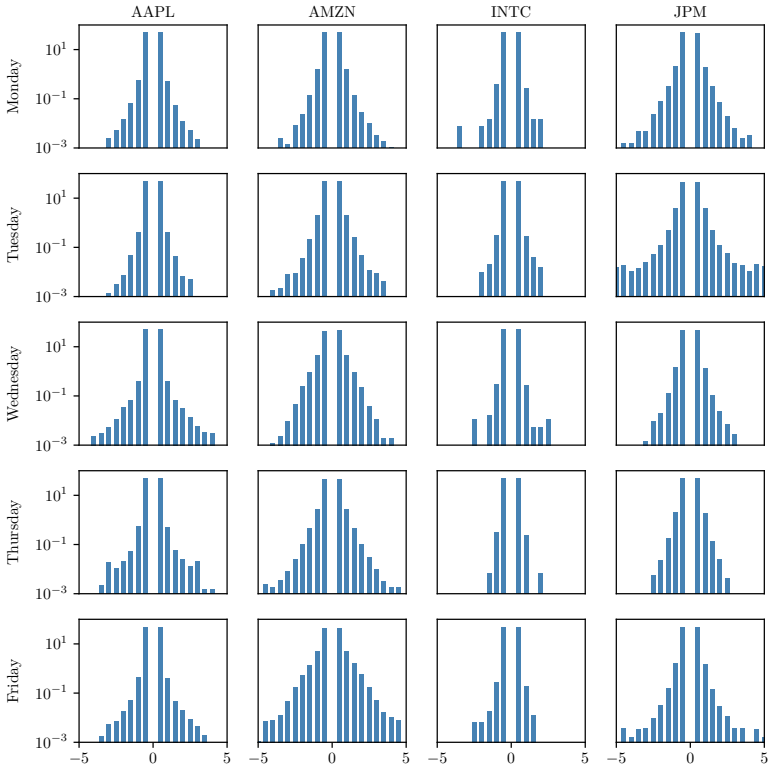


Figure 8.13: Distribution of jump sizes

We plot the histogram of mid-price jump values in ticks for each ticker on each date. The y-axis is set in logarithmic scale and displays frequencies as a percentage.

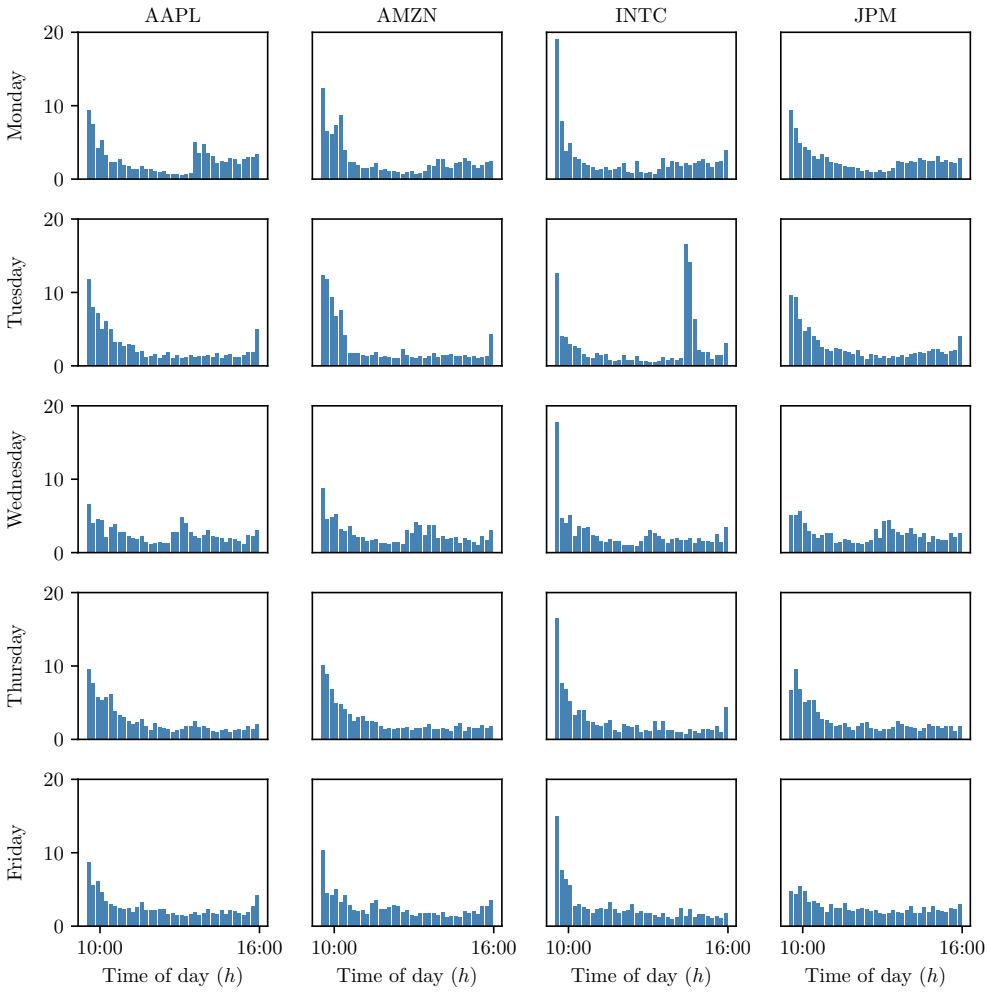


Figure 8.14: Distribution of mid-price jump times

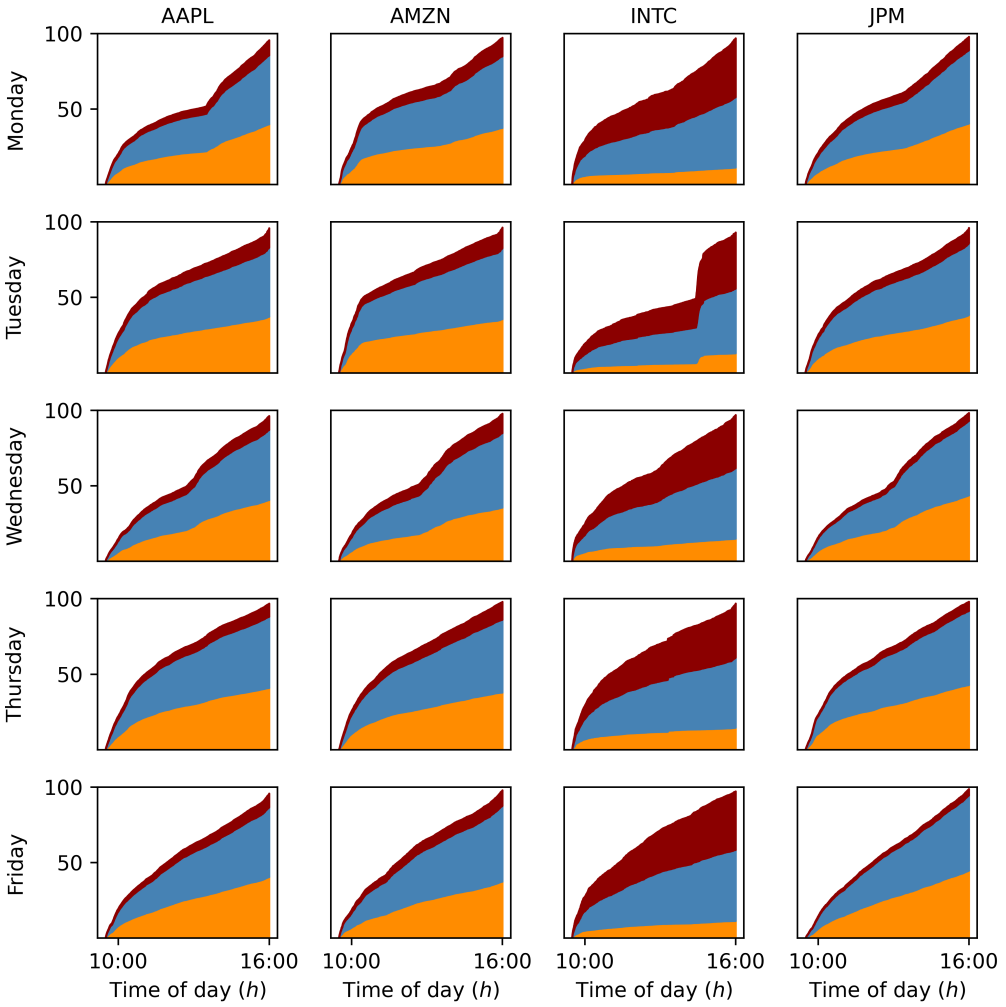


Figure 8.15: Event types of jumps through time.

For each ticker and each trading day, number of deletions (orange), liquidity provisions (blue), and trades (red), divided by the total number of jumps on that day, against time of the day.

Chapter 9

Hawkes price model

9.1 Price model

This chapter aims to demonstrate the use of the `ASLSD` algorithm to enrich the standard class of point process based mid-price models in the literature (used by Bacry et al. [6] for instance). Let $\boldsymbol{N}$ be a bi-dimensional counting process, where N^1 (resp. N^2) counts upward (resp. downward) jumps of the mid-price. Let $p_0 > 0$ be the initial value of the mid-price in ticks. As in the previous chapter, we consider a total observation window of size $T = 23\,400s$, corresponding to the duration of a trading day on Nasdaq. For times $t \in [0, T]$, the standard mid-price model (SPM) is

$$p_t = p_0 + \frac{1}{2}(N_t^1 - N_t^2). \tag{9.1}$$

Under the SPM, the moments of p relate simply to those of $\boldsymbol{N}$; for instance, the auto-covariance of returns and the realized variance are

$$\nu(p)_\tau^h(t) = \frac{1}{4}\Big(\nu_{11,\tau}^h(t) + \nu_{22,\tau}^h(t) - \nu_{12,\tau}^h(t) - \nu_{21,\tau}^h(t)\Big), \quad [p]_t = \frac{N_t^1 + N_t^2}{4}. \tag{9.2}$$

It is clear that the SPM presents several conceptual limitations. For instance, it does not guarantee that for times $t \in [0, T]$, $p_t \geq 0$ *a.s.*, *i.e.* $N_t^2 - N_t^1 \leq 2p_0 \quad a.s.$; nor that $\mathbb{E}[p_t] \geq 0$. The SPM also assumes all price jumps have size 0.5 ticks; we refer to this as the adjusted mid-price.

Definition 9.1.1 (Adjusted price)**.** *For a given observed series of mid price jumps, with associated counting processes N^1 and N^2 (where jumps can be of any size), we define the adjusted price process to be that given by $p^{\text{adj}} = p_0 + (N^1 - N^2)/2$, in units of ticks.*

We shall see that p^{adj} typically overstates the volatility of the price process, as leaps often correspond to mean-reversion. For simplicity, we will often write p for p^{adj}, as this will be our main process of interest.

Again, we use this as a first example application of our procedure in market microstructure, which can be significantly improved. We consider different classes of intensity models for $\boldsymbol{N}$.

Homogeneous Poisson We use a benchmark homogeneous Poisson model `poisson_hom`.

Proposition 9.1.1 (Fitted price moments)**.** *For times $t \in [0, T]$, the moments of the fitted `poisson_hom` price model satisfy*

$$\mathbb{E}[\bar{p}_t] = p_0 + \frac{p_T - p_0}{T} t, \qquad \mathrm{Var}[\bar{p}_t] = \frac{t}{T}[p]_T. \tag{9.3}$$

For a sampling period $h > 0$, the auto-covariance of returns at lag $\tau \geq 0$ is

$$\nu[\bar{p}]^h_\tau = \frac{[p]_T}{T} f^{(h)}_{Tr}(\tau). \tag{9.4}$$

On $[0, T]$, the mean fitted price under `poisson_hom` is a linear function of time, with the same initial and terminal values as the training data. The variance of this model increases linearly with time, from 0 to the quadratic variation of the training data. Figure 9.1 plots these quantities for the ticker `AAPL`, on Monday 18 July 2022. Positivity constraints are

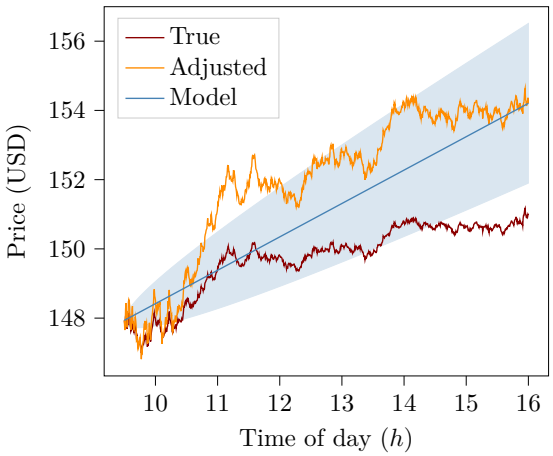


Figure 9.1: The `poisson_hom` price model
For times $t \in [0, T]$, blue line is the mean fitted price $\mathbb{E}[\bar{p}_t]$, orange line is the true adjusted mid-price p_t, and red line is the true mid-price, and We fill in blue the area one standard deviation away from the mean. The axes are re-scaled to display time as hours of the day instead of seconds, and prices in USD instead of ticks.

always satisfied in expectation for fitted models, *i.e.* $\mathbb{E}[\bar{p}_t] > 0$ for times $t \in [0, T]$. We get the distribution of the price process in closed form. Denote by I_q the modified Bessel function of the first kind, such that for all integers $q \in \mathbb{Z}$ and all $x \in \mathbb{R}$

$$I_q(x) := \left(\frac{x}{2}\right)^q \sum_{k=0}^{+\infty} \frac{\left(\frac{x^2}{4}\right)^k}{k!(k+q)!}. \tag{9.5}$$

Proposition 9.1.2 (Price distribution)**.** *For times $t \in [0,T]$, and all signed prices q such that $2q \in \mathbb{Z}$, under the fitted* `poisson_hom` *model,*

$$\mathbb{P}(\bar{p}_t = q) = \Big(\frac{N_T^1}{N_T^2}\Big)^{q-p_0} \exp\Big(- (N_T^1 + N_T^2)\frac{t}{T}\Big) I_{2(q-p_0)}\Big(2\sqrt{N_T^1 N_T^2}\frac{t}{T}\Big). \tag{9.6}$$

We give a proof of this result in Appendix A.4.

Non-homogeneous Poisson We consider `poisson_piececonst`, a non-homogeneous Poisson model with piece-wise constant intensity (see Section 2.4.2). The bounds of the intervals correspond to a uniform grid of the duration of the trading day, in bins of 30 minutes each. This model also acts as a benchmark in our study. The fitted hon-homogeneous Poisson piece-wise constant intensities provide upper-bounds for the baseline parameters of MTLH models with piece-wise constant baselines.

Consider a fitted `poisson_piececonst` model $(\bar{p}_t)_t$.

Proposition 9.1.3 (Fitted price moments)**.** *For all times $t \in [0,T]$, the expectation of the fitted* `poisson_piececonst` *price model is*

$$\mathbb{E}[\bar{p}_t] = p_{\beta_{g(t)}} + \frac{p_{\beta_{g(t)+1}} - p_{\beta_{g(t)}}}{\beta_{g(t)+1} - \beta_{g(t)}}(t - \beta_{g(t)}), \tag{9.7}$$

and the variance is

$$\mathrm{Var}[\bar{p}_t] = [p]_{\beta_{g(t)}} + ([p]_{\beta_{g(t)+1}} - [p]_{\beta_{g(t)}})\frac{t - \beta_{g(t)}}{\beta_{g(t)+1} - \beta_{g(t)}}. \tag{9.8}$$

Under `poisson_piececonst`, the fitted mean price is a piece-wise linear interpolation of the original price path. The variance of the fitted price through time is a piece-wise linear interpolation of the quadratic variation of the training price. For times $t \in [0,T]$, whether the variance of the fitted price $\bar{p}_t$ is greater or lower than the variance of `poisson_hom` depends on the dynamic of the true quadratic variation $[p]_t$ on the training data. Figure 9.2 plots these quantities for the ticker `AAPL`, on Monday 18 July 2022.

MHP models Because of their branching representation, we do not expect MHP mid-price models with monotonically decaying kernels to lead to realistic inter-arrival time distributions for this problem. Fix an event type $i \in [d]$. If, for all event types $j \in [d]$, the kernel ϕ_{ij} is decreasing, then for all j the sequence $(p_{i,m,j,n})_n$ is also decreasing. Therefore, conditional on the jump at time t_m^i not being a background jump, the most likely candidates to trigger this jump are the jumps immediately preceding it in each dimension; that is, the jumps at times $(t_{n_j}^j)$ for $j \in [d]$ such that $t_{n_j}^j < t_m^i, \quad t_{n_j+1}^j \geq t_m^i$. This implication

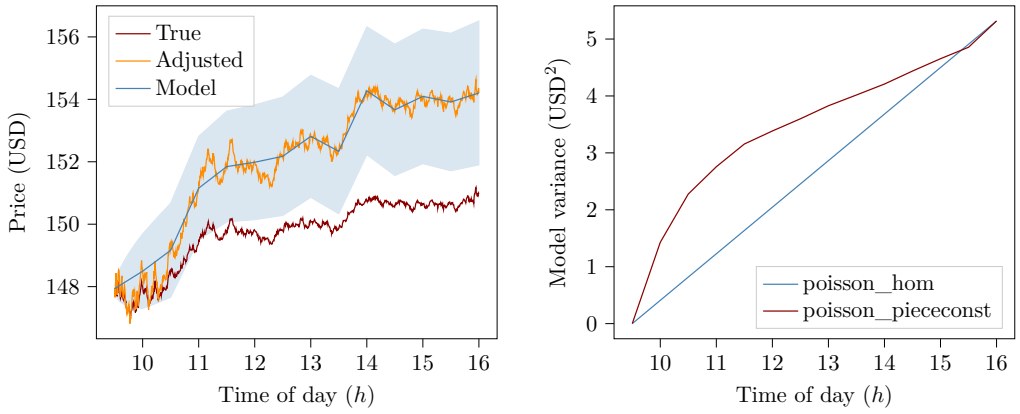


Figure 9.2: The `poisson_piececonst` price model

Left: For times $t \in [0, T]$, blue line is the mean fitted price $\mathbb{E}[\bar{p}_t]$, orange line is the true adjusted mid-price p_t, and red line is the true mid-price, and We fill in blue the area one standard deviation away from the mean. The axes are re-scaled to display time as hours of the day instead of seconds, and prices in USD instead of ticks. Right: we compare model variance as function of time for `poisson_piececonst` and `poisson_hom`.

is not coherent with the typical latency of algorithmic traders, even of ultra-high-frequency traders (see Figure 8.3). This suggests that non-monotonically decaying kernels might be more appropriate. Therefore, we use 3 MHP models

- `mhp_semi`: semi-parametric MHP model with a learnt mixture of Gaussian kernels with 21 logarithmically spaced means in $[25.10^{-6}s, 25s]$, variances between 12% and 50% of means.
- `mhp_exp`: MHP with exponential kernels. This model serves as a Hawkes benchmark because it is very commonly used in Hawkes applications in market-microstructure.

MTLH model `mtlh_pc` is an MTLH model with piece-wise constant baseline as in `poisson_pc`. Each kernel is a mixture of a truncated power law kernel with $\delta_L := 20$ microseconds and $\delta_R := 50$ milliseconds, and 3 SBF Gaussian kernel densities with scale 5 microseconds located at 10, 20, and 30 microseconds.

9.2 Fitted models

A universal delayed power-law decay First, we focus on the results of the fit of `mhp_semi`, as a semi-parametric model designed to guide our modelling choices. Figure 9.3 plots the fitted cross excitation kernels ϕ_{12} in logarithmic scale. There is a clear delay in kernel activity. For times ranging from $20\mu s$ to $100\mu s$, the fitted kernel exhibits a clear Gaussian mode. We fit a linear regression curve (in log-scale) to the kernel values in the interval of times greater than $100\mu s$; for times ranging from $100\mu s$ to $1s$, the fitted Gaussian

kernel resembles a power law of exponent close to -1. Figure 9.4 shows that fitted MHP

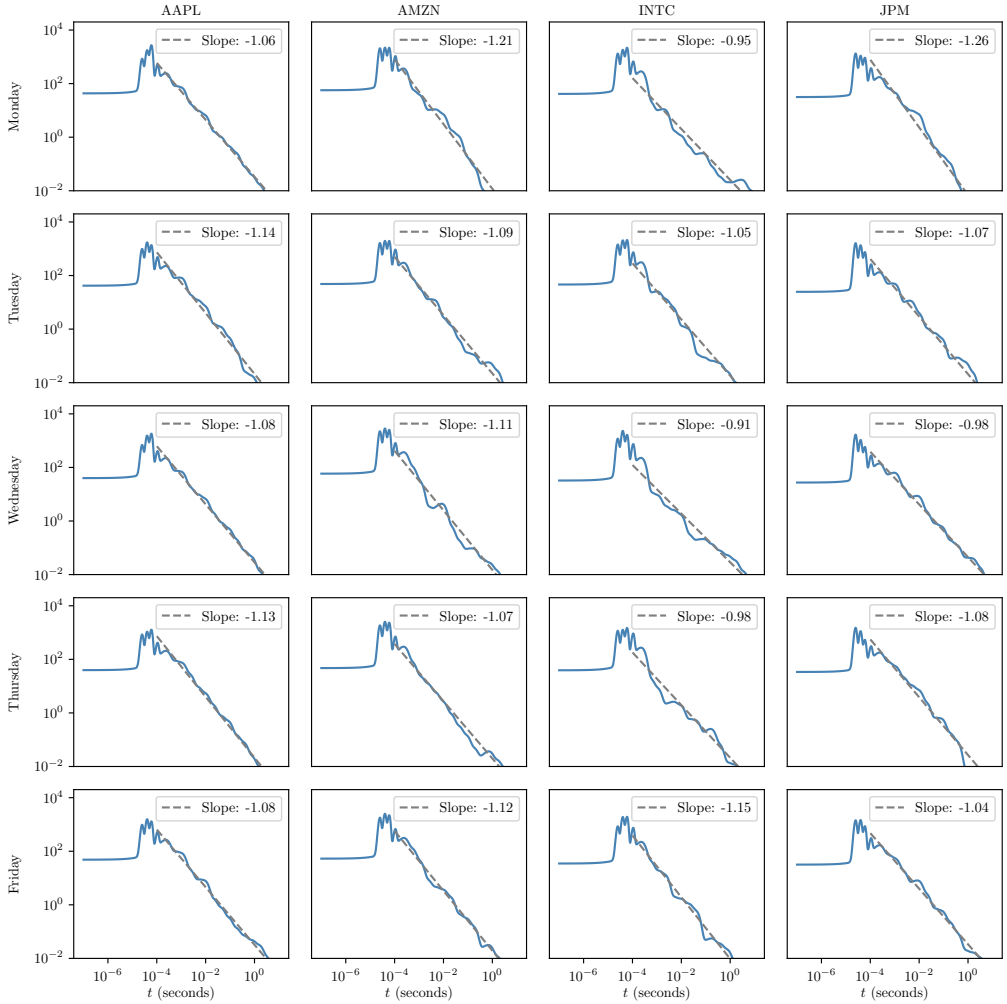


Figure 9.3: Fitted Gaussian kernels

models are stable but near-critical.

MTLH Figure 9.5 plots fitted baselines for the MTLH model. Figure 9.6 plots fitted kernels for the MTLH model. The kernel matrix is clearly bi-symmetric, with a heavy-tailed cross-excitation consistent with the mean reverting dynamics of mid-prices. This model correctly retrieves the empirical moments of the training data. Figure 9.7 plots the empirical intensity of fitted models on simulated paths, and compares it to the empirical intensity of the training path, with sampling period $h = 1$ minute. Figure 9.8 plots the empirical covariance of fitted models with sampling period $h = 1$ second.

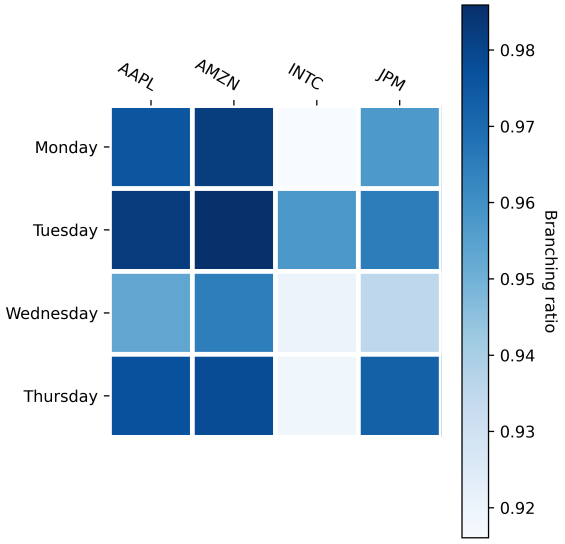


Figure 9.4: Branching ratios of fitted `mhp_gauss` models.

9.3 Exogenous price moves

When applying Hawkes models to real world data, there is no statistical test of stochastic declustering, using only sample paths of the counting process. In the specific case of financial prices modelling, a further difficulty is that we cannot define rigorously what constitutes an exogenous or endogenous price move outside of the Hawkes model, as opposed to our previous application in epidemic propagation, for instance. In finance, neither the definition of an exogenous-endogenous price move dichotomy is clear, nor the existence and uniqueness of causality between different price moves. We use stochastic declustering of Hawkes models in this work as a stylized representation of information mechanisms in an LOB. For a heuristic definition, we see exogenous events as those mid-price moves which do not respond to other mid-price moves of the same ticker on the same trading day: for instance, mid-price moves at trades motivated solely by information on the fundamental, or block trades occurring when re-balancing a portfolio. We see endogenous events as mid-price moves which are directly triggered by another mid-price move: for instance, a trading algorithm reacting to a new value of mid-price that passed a certain threshold, and placing a MO that results in a multi-level execution. It is clear that many mid-price moves do not fall in either of these two cases. While there is a significant literature in market microstructure studying lead-lag relationships between tickers, for example by Cartea et al. [21], these usually consider the impact of returns over a given period rather than individual price moves. In fact, even simple cases of mid-price moves are difficult to classify, for instance: mid-price jumps occurring at liquidation times of a TWAP liquidation strategy.

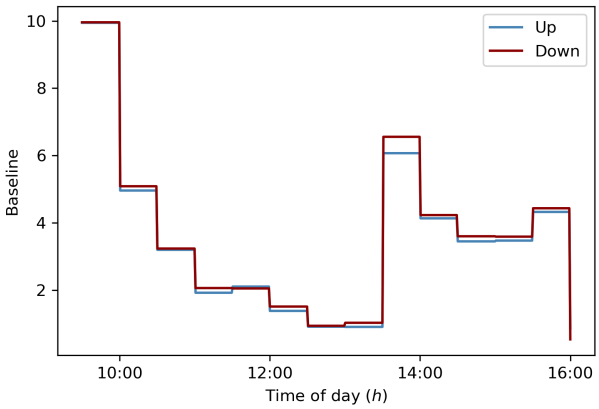


Figure 9.5: Fitted baselines (SPM)

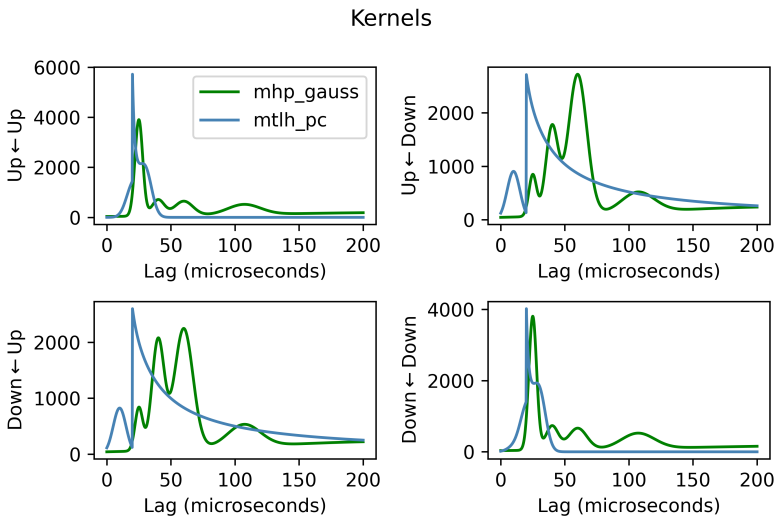


Figure 9.6: Fitted kernels (SPM)

Define an exogenous events classifier as in Section 3.3.3.1 based on the `mtlh_pc` fit. We first calibrate our classifier's decision threshold by simulating data from the Hawkes model with an exact branching algorithm. Figure 9.9 plots different evaluation metrics for this classifier, as well as its ROC curve which displays a AUC of 97.1%. These results show satisfactory performance for this classifier, which we now apply to the training data. Figure 9.10 plots the rate of positives against the decision threshold for the training data, and the empirical distribution of exogeneity probabilities. Following the calibration results, fix a decision threshold $\epsilon = 60\%$. This implies a 43% positivity rate on the training data. This order of magnitude is also consistent with the branching ratio of the model. Figure 9.11 plots the empirical exogeneity probability conditional on the LOB event type, and shows that trades are particularly over-represented.

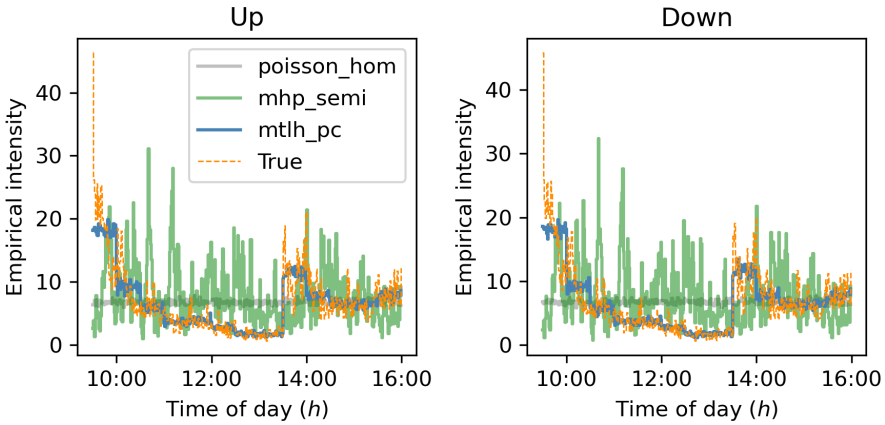


Figure 9.7: Empirical intensity of fitted models on simulated paths (SHM)

9.4 Counter-factual impact model

In market microstructure, market impact describes how a given order or market event changes market dynamics such as price, order-flow, or volatility. The branching representation of Hawkes models gives rise to a natural notion of market impact, by censoring an event and its descendants from the path of the point process as in Section 3.3.3.

9.4.1 Hawkes price impact

Let $\boldsymbol{N}$ be a bi-dimensional counting process and let p be the associated SPM as in Equation (9.1). Fix $i \in [d]$, and $m \in [N_T^i]$. Denote by $\tilde{\boldsymbol{N}}_{\boldsymbol{t}}$ the associated censored path, obtained by removing the event in t_m^i and its descendants. Let $\tau_{i,m}$ be the relaxation time of the system. For times $t \in [0, T]$ the resulting censored price $\tilde{p}_\cdot$ is $\tilde{p}_t := f(\tilde{\boldsymbol{N}}_{\boldsymbol{t}})$.

Definition 9.4.1 (Price impact)**.** *The price impact of the event in t_m^i on the mid-price p is the cad-lag stochastic process $\Delta_{i,m}$, such that for lags $u \in [-t_m^i, T - t_m^i]$*

$$\Delta_{i,m}(u) = p_{t_m^i + u} - \tilde{p}_{t_m^i + u}, \tag{9.9}$$

It is clear that the price impact is causal, that is $\Delta_{i,m}(u) = 0$, $\forall u < 0$, and for times $t \in [0, T]$, $p_t = \tilde{p}_t + \Delta_m^i(t - t_m^i)$. The impact function Δ_m^i is constant for times $u > \tau_{i,m}$. Denote its limit by $\Delta_{i,m}(\infty) := \lim_{u \to (\tau_{i,m})^+} \Delta_{i,m}(u)$. We call $\Delta_{i,m}(\infty)$ the residual price impact. In general, the residual price impact is not necessarily null: in this sense, this notion of relaxation time differs from the one used for modelling dynamic systems in physics; the LOB does not revert back to the original equilibrium, but goes back to a perturbed equilibrium.

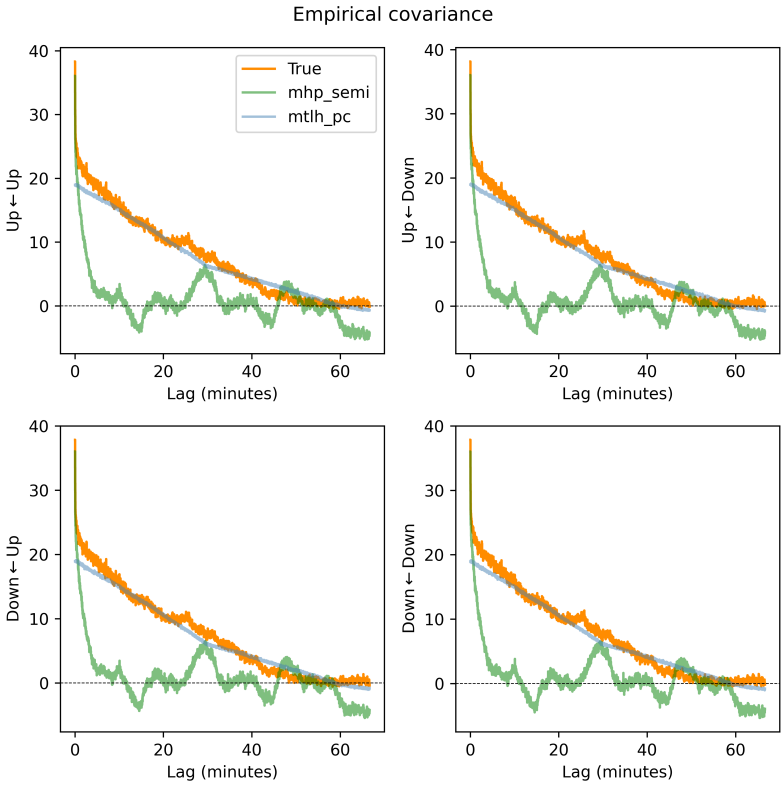


Figure 9.8: Empirical covariance of fitted models (SPM)

Proposition 9.4.1 (Price impact)**.** *The price impact Δ_m^i of the event in t_m^i on the mid-price p is given by the cad-lag stochastic process defined for lags $u \in [-t_m^i, T - t_m^i]$ by*

$$\Delta_{i,m}(u) = \begin{cases} \chi_m^i + \zeta_m^i(u) & \text{if } u \geq 0, \\ 0 & \text{otherwise,} \end{cases} \tag{9.10}$$

where

$$\begin{aligned} \chi_m^i &:= N^1(\{t_m^i\}) - N^2(\{t_m^i\}), \\ \zeta_m^i(u) &:= N^1\left((t_m^i, t_m^i + u] \cap \mathcal{D}_{T,i,m}^{(1)}\right) - N^2\left((t_m^i, t_m^i + u] \cap \mathcal{D}_{T,i,m}^{(2)}\right). \end{aligned} \tag{9.11}$$

We refer to the scalar χ_m^i as the instantaneous impact, and to the function ζ_m^i as the transient impact.

Another interesting property of the SPM is that under this linear model, the price verifies a kernel equation: the price process is the superposition of impacts of exogenous events. Recall that for a Hawkes model with baselines $\boldsymbol{\mu}$, for event types $i \in [d]$, $N^{(\mu_i)}$ denotes the baseline process, that is, the Poisson process with intensity μ_i.

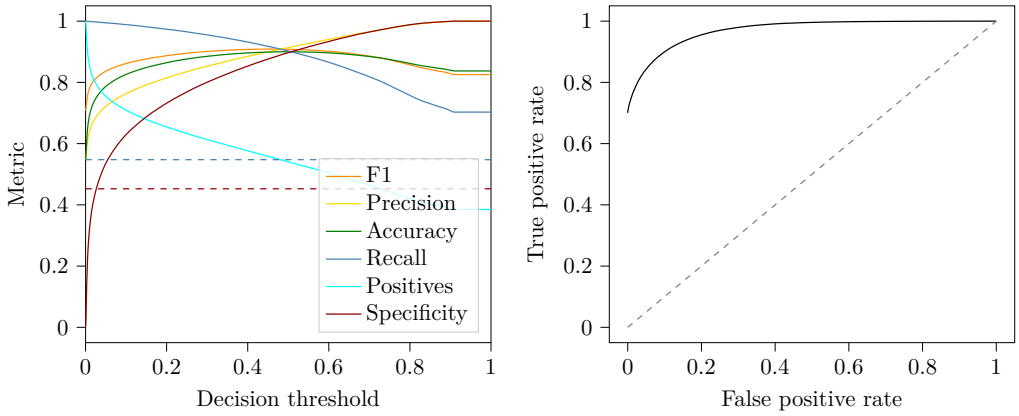


Figure 9.9: Metrics of exogenous classifications, calibration (SPM)
Blue (resp. red) horizontal dashed line is the true rate of exogenous (resp. endogenous) events in the simulated data.

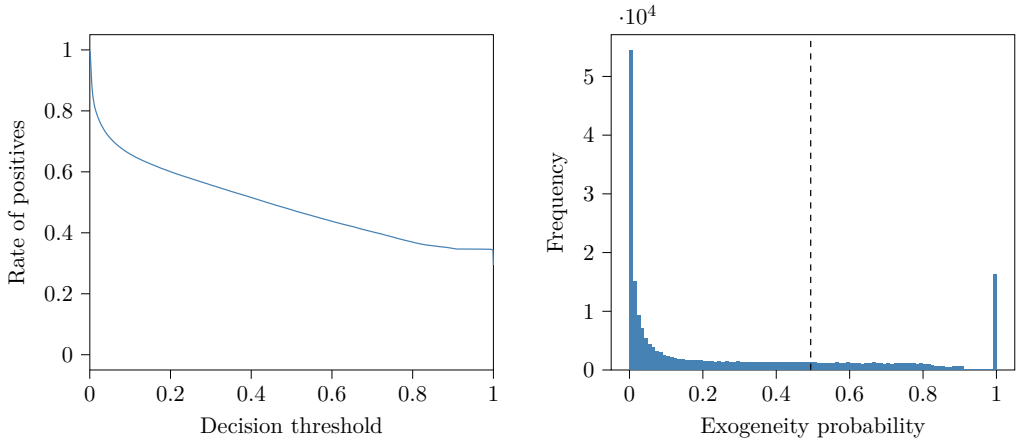


Figure 9.10: Exogeneity probabilities (SPM)

Proposition 9.4.2 (Pricing kernel equation for the SPM)**.** *For times $t \in [0, T]$, the mid-price p_t verifies*

$$p_t = p_0 + \sum_{i=1}^{d} \int_0^t \Delta_{i,N_s^i}(t-s) \mathrm{d}N_s^{(\mu_i)}. \tag{9.12}$$

9.4.2 Results

We apply our results for the `mtlh_pc` fit. Figure 9.12 plots the expected price impact implied by this model, averaged over 10^3 simulated paths. The expected price impact is conditional on the event type of the parent event: upward or downward mid-price jump. For a downward parent jump, we multiply the price impact by a factor -1. In both cases, the price impact relaxes almost monotonically to its asymptotic value in a few milliseconds. Note that 63.1% (resp 61.6%) of paths induced by an upward (resp. downward) parent move do not contain any descendant, which is consistent in order of magnitude with the adjacency

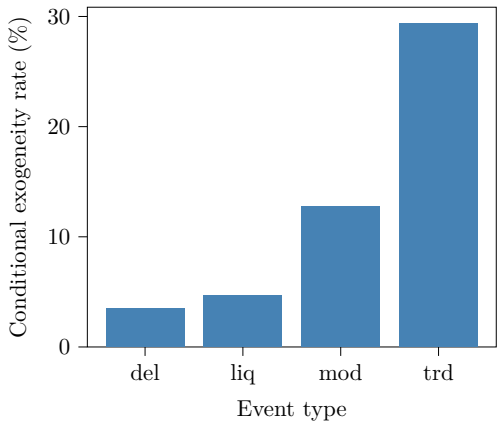


Figure 9.11: Empirical exogeneity probability conditional on event type

matrix of the fitted model. Figure 9.13 plots the empirical distribution of the residual price

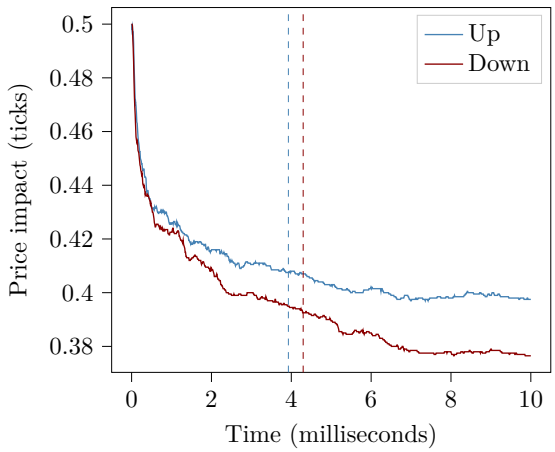


Figure 9.12: Expected price impact

impact. This impact model displays clear adverse selection, as 74.8% (resp. 72.9%) of residual impacts induced by an upward (resp. downward) jump parent are strictly positive (resp. negative). Finally, Figure 9.14 the empirical distribution of relaxation times. The mean relaxation time of an upward (resp. downward) mid-price move is 3.93 milliseconds (resp. 4.30 milliseconds).

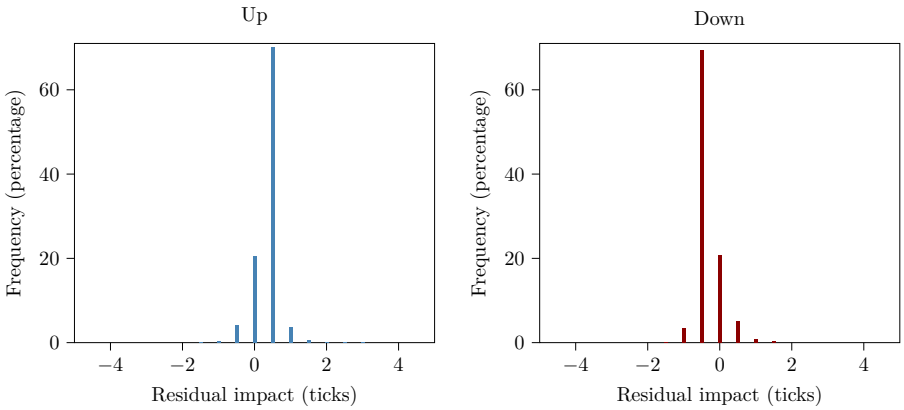


Figure 9.13: Empirical distribution of residual price impact

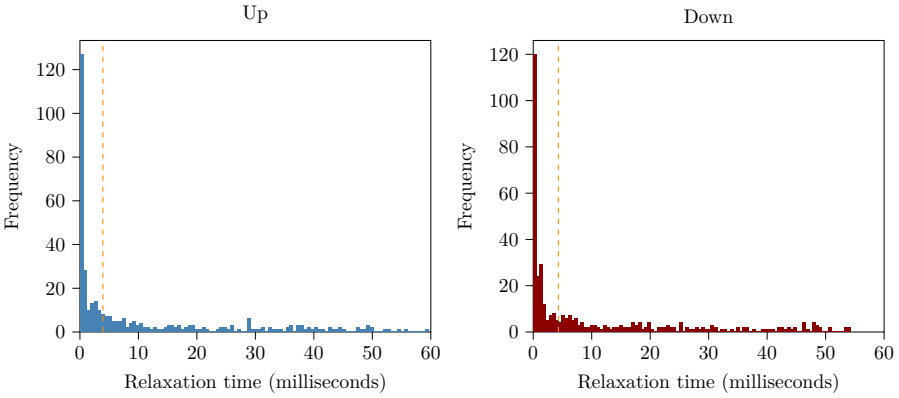


Figure 9.14: Empirical distribution of relaxation times

Chapter 10

Conclusion

10.1 Least-squares estimation of Hawkes processes

In this work, we proposed a new estimation method for MHP and MTLH models, which applies to large datasets and general kernels.

10.1.1 Advantages

Competitivity The numerical experiments we conduct in Section 5.1 show that the precision of this algorithm competes with state-of-the-art methods with significant computational advantages. In Table 4.1, we see that the time complexity of our algorithm is lower than any other method. We also note that `ASLSD` has memory complexity in $\mathcal{O}(1)$ during the gradient iterations as we only store the values of strata allocations. There is a fixed linear cost of $\mathcal{O}(N_T)$ while pre-processing the data (before preforming the gradient iterations), to store the book-keeping lists ϖ and κ.

Explainability An exciting alternative approach to calibrating general models for marked point processes, is to use neural network parameterizations of the conditional intensity function. Examples of this can be found in Du et al. [33], Mei and Eisner [66], Zhu et al. [113], and Dong et al. [31]. In contrast to neural network methods for Hawkes model estimation, our method cannot be transposed simply to a non-linear Hawkes case. However, the semi-parametric nature of our method retains some critical advantages – the semi-parametric formulation allows easy model explainability, as the kernel can be sketched; the connection of the MHP with a branching process is preserved; and (in contrast to models based on neural networks, cf. Rigaki and Garcia [91]), it is straightforward to verify whether the model causes privacy issues when fitted with private data, as one simply checks the fitted kernel is appropriate to publish. These advantages suggest that parametric and

semi-parametric methods retain a significant role in many situations. Compared to non-parametric estimation algorithms for MHP, our semi-parametric families are dense in the space of kernels while maintaining explainability and data privacy. Furthermore, some non-parametric estimation algorithms for MHP (in particular, the deep learning based methods in Du et al. [33] and Mei and Eisner [66]) lose the branching structure of the MHP as they reconstruct the conditional intensity without reconstructing linear Hawkes kernels.

10.1.2 Limitations

Curse of dimensionality in event types Each gradient iteration in `ASLSD` has cubic complexity in the number of event types d. Estimating systems with more than 4 event types on data paths with over 10^5 events for each type turns out to be very impractical without sparsity assumptions like in Zhou et al. [111]. The `ASLSD` method does not seem to be easily enhanced to accommodate large numbers of event types: for instance, it is not clear whether randomly sampling which kernels ϕ_{ij} to update in the stochastic gradient approximation instead of updating all of them would be sufficiently accurate. In the Hawkes literature, this curse of dimensionality is usually circumvented by reducing the number of events to 1 or 2 fundamental types, and encoding the actual type with categorical marks. This clearly limits modelling ability.

Curse of dimensionality in mixture size Because of the cross-correlation terms Υ_{ij}, each gradient iteration in `ASLSD` has quadratic complexity in the number of terms r_{ij} in each kernel mixture model $(\phi_{ijl})_{l \in r_{ij}}$. This limits the use of our method for fine modelling of highly variable kernels. This issue is particularly annoying as unfortunately, model selection is still an unexplored area in the Hawkes literature.

10.1.3 Future work

MHP and MTLH models are insightful, explainable models, which capture first and second order properties of stationary and non-stationary data streams. However, it is clear that they remain stylized models with very small numbers of parameters and several limitations. To extend our estimation method to another class of conditional intensity model as is, it is necessary that the LSE of that model class satisfies an additive decomposition like the one we propose in Theorem 4.3.1, and that the MC estimation of this decomposition is still computationally practical. Several simple extensions to the MTLH models verify this; such as the state-dependent linear Hawkes model, where for a stochastic process $\boldsymbol{X}$, and baseline

functions $\boldsymbol{\mu}$ dependent of time and state $\boldsymbol{X}$, we consider the model

$$\lambda_i(t) := \mu_i(t, \boldsymbol{X_t}) + \sum_{j=1}^{d} \sum_{n=1}^{N_t^j} \mathcal{I}_{ij}(\xi_n^j)\phi_{ij}(t - t_n^j), \quad \forall i \in [d], \quad \forall t \geq 0. \tag{10.1}$$

It is also the case for MTLH with with regime switching, where kernels ϕ_{ij} are no longer translation invariant, that is, they do not only depend on the time elapsed since an event, but also on the timing of the event

$$\lambda_i(t) := \mu_i(t) + \sum_{j=1}^{d} \sum_{n=1}^{N_t^j} \mathcal{I}_{ij}(\xi_n^j)\phi_{ij}(t - t_n^j, t_n^j), \quad \forall i \in [d], \quad \forall t \geq 0. \tag{10.2}$$

Finally, our decomposition also extends to MTLH with external impact, inspired by Rambaldi et al. [87], where for a counting process Y and a uni-variate function f, we consider the models

$$\lambda_i(t) := \mu_i(t) + \sum_{j=1}^{d} \sum_{n=1}^{N_t^j} \mathcal{I}_{ij}(\xi_n^j)\phi_{ij}(t - t_n^j) + \int_0^t f(t-s)dY_s, \quad \forall i \in [d], \quad \forall t \geq 0. \tag{10.3}$$

However, as discussed above, the LSE of the non-linear Hawkes processes (Brémaud and Massoulié [17]) does not verify a practical additive decomposition. The conditional intensity of a point process has to remain positive, and the fact that MHP kernels are positive is sufficient to satisfy this condition. However, because their kernels are positive, MHP and MTLH models are unable to model inhibition between events. In high-frequency LOB data for example, inhibitory behaviours are well known: for example, Lu and Abergel [63] observe empirically that order cancellations that change the spread may inhibit submission of contra-side market orders. Non-linear Hawkes processes can model inhibitory interactions, but the estimation of non-linear Hawkes processes is a hard problem and the literature is scarce, with the notable exceptions of the work of Wang et al. [107] and Menon and Lee [68]. To the best of our knowledge, there exists no fast method to calibrate non-linear Hawkes processes with general kernels. Suppose we define non-linear MTLH model for event types $i \in [d]$ and for times $t \geq 0$ by

$$\lambda_i(t) = A_i\Big(\mu_i(t) + \sum_{j=1}^{d} \sum_{m:t_m^j<t} \phi_{ij}(t - t_m^j)\mathcal{I}_{ij}(\xi_m^j)\Big), \tag{10.4}$$

where $A_i : \mathbb{R} \to [0, +\infty)$ is an activation function. If A_i is a positive polynomial, for instance $A_i(x) := x^2$ for all $x \in \mathbb{R}$, then we still get an additive decomposition of the LSE following the approach in Theorem 4.3.1. However, this decomposition is impractical as it involves cross-correlation terms that depend on 4 event times

$$\int_0^{T-t_{m_1}^{i_1}} \phi_{ki_1}(u)\phi_{ki_{i_2}}(u + t_{m_1}^{i_1} - t_{m_2}^{i_2})\phi_{ki_{i_3}}(u + t_{m_1}^{i_1} - t_{m_3}^{i_3})\phi_{ki_{i_4}}(u + t_{m_1}^{i_1} - t_{m_4}^{i_4})\mathrm{d}u. \tag{10.5}$$

For general classes of activation functions, the additive decomposition in Theorem 4.3.1 cannot be directly transposed to the LSE of this model; in fact it seems improbable to obtain any useful additive decomposition of the LSE in this case. Following the approach of Menon and Lee [68], it would be interesting to investigate the use of other contrast functions than the LSE for the estimation of non-linear Hawkes processes.

10.2 Hawkes models of Nasdaq equities prices

Our procedure allows us to overcome the limitations of Hawkes models in market microstructure by getting a stylized but more realistic price model, that is non-Markovian and non-stationary. However, this is still only a first step into leveraging the `ASLSD` estimation algorithm for Hawkes mid-price models, and therefore, presents several limitations.

Absence of volume data We expect the impact of order-flow and volume data on mid-price movements to be better modeled as state variables rather than marks and impact functions. Furthermore, by omitting order-flow data, our model does not use executions of non-displayed orders.

Learning framework For practical applications, it remains unclear whether price dynamics can be transferred across days or tickers. Even if the Hawkes price model is improved by including order flow and other features in a state-dependent Hawkes model, or allowing inhibitory effects in a non-linear Hawkes model, we still fit these models for each trading day and each ticker.

Post-open and pre-close dynamics Several price models exclude data near the opening and closing of the trading day as those are particularly active periods with increased volatility. The closing cross on Nasdaq is in place since April 2004 and concerns all Nasdaq listed securities. Before market closing, in parallel to the continuous book (the LOB for continuous trading), market participants have the possibility to place On Close orders that will be executed or rejected only at market closing. These orders do not interact with the continuous book at the time they are posted, and constitute a parallel book that is not displayed in the data feed. Nevertheless, aggregated information about this book is disseminated at fixed times from 03:50 PM to 04:00 PM

Appendix A

Proofs

A.1 Conditional intensity modelling

A.1.1 Point processes

Proof of Proposition 2.1.1. Fix $k \in [d]$ and $t > 0$. Since $\tilde{M}_0^k = 0$, the Doob–Meyer decomposition of N^k gives $\mathbb{E}[N_t^k] = \mathbb{E}[\lambda_k(t)]$, and the mean of η_t^k follows. Now note that

$$(N_t^k)^2 = (\tilde{M}_t^k)^2 + 2\tilde{M}_t^k \Lambda_k(t) + \Lambda_k^2(t). \tag{A.1}$$

Since $(\tilde{M}^k)^2 - \Lambda_k$ is a martingale, $\mathbb{E}\left[(\tilde{M}_t^k)^2\right] = \mathbb{E}\left[\Lambda_k(t)\right]$. By taking the expectation in (A.1), we get

$$\mathbb{E}\left[(N_t^k)^2\right] = \mathbb{E}\left[\Lambda_k(t)\right] + \mathbb{E}\left[2\tilde{M}_t^k \Lambda_k(t) + \Lambda_k^2(t)\right]. \tag{A.2}$$

Therefore,

$$\mathrm{Var}\left[N_t^k\right] = \mathbb{E}\left[\Lambda_k(t) + 2\tilde{M}_t^k \Lambda_k(t)\right] + \mathrm{Var}[\Lambda_k^2(t)]. \tag{A.3}$$

□

A.1.2 Poisson processes

Proof of Proposition 2.4.1. For a Poisson process, $\Lambda_k(T)$ is deterministic. Therefore, $\mathrm{Var}[\Lambda_k(T)] = 0$, and using Proposition 2.1.1, $\mathbb{E}[\tilde{M}_T^k \Lambda_k(T)] = \Lambda_k(T)\mathbb{E}[\tilde{M}_T^k] = 0$. □

Proof of Proposition 2.4.2. Fix a time $t > 0$, event types $i, j \in [d]$ with $i \neq j$, a lag $\tau \geq 0$, and a sampling period $h > 0$. By independence of the counting processes N^i and N^j, we get $\nu_{ij,\tau}^{(h)}(t) = 0$. By definition,

$$\nu_{ii,\tau}^{(h)}(t) := \frac{1}{h}\mathrm{Cov}\left[N^i\Big(\big[t, t+h\big)\Big), N^i\Big(\big[t+\tau, t+\tau+h\big)\Big)\right]. \tag{A.4}$$

If $\tau \geq h$, the intervals $\big[t, t+h\big)$ and $\big[t+\tau, t+\tau+h\big)$ are disjoint. By independence of the increments of N^i, we get $\nu^{(h)}_{ii,\tau}(t) = 0$. If $\tau < h$,

$$\begin{aligned}\nu^{(h)}_{ii,\tau}(t) &= \frac{1}{h}\mathrm{Cov}\Big[N^i\Big(\big[t, t+\tau\big)\Big) + N^i\Big(\big[t+\tau, t+h\big)\Big), \\ &\qquad N^i\Big(\big[t+\tau, t+h\big)\Big) + N^i\Big(\big[t+h, t+\tau+h\big)\Big)\Big], \\ &= \frac{1}{h}\mathrm{Cov}\left[N^i\Big(\big[t+\tau, t+h\big)\Big), N^i\Big(\big[t+\tau, t+h\big)\Big)\right], \\ &= \frac{1}{h}\mathrm{Var}\left[N^i\Big(\big[t+\tau, t+h\big)\Big)\right],\end{aligned} \tag{A.5}$$

where we used again the independence of increments of N^i. □

A.1.2.1 Constant intensities

Proof of Proposition 2.4.4. If the ground truth process $\lambda^\diamond_k$ is a (possibly non-homogeneous) Poisson model, then

$$\mathrm{Var}\big[\mathcal{R}^k_T(\mu_k)\big] = 4\mu_k^2 \frac{1}{T^2}\int_0^T \lambda^\diamond_k(t)\mathrm{d}t. \tag{A.6}$$

If $\lambda^\diamond_k$ is bounded, denote its maximum value by $||\lambda^\diamond_k||_\infty$. Since $\mathrm{Var}\big[\mathcal{R}^k_T(\mu_k)\big] \leq 4\mu_k^2 \frac{||\lambda^\diamond_k||_\infty}{T}$, we get $\lim_{T\to+\infty} \mathrm{Var}\big[\mathcal{R}^k_T(\mu_k)\big] = 0$. Using Markov's inequality, the LSE is temporally consistent. □

A.1.2.2 Linear intensities

Proof of Proposition 2.4.13. For parameters $b, a \geq 0$, the derivatives of the partial LSE are

$$\frac{\partial \mathcal{R}^{(k)}_T}{\partial b} = aT + 2b - 2\eta^k_T, \qquad \frac{\partial \mathcal{R}^{(k)}_T}{\partial a} = \frac{2}{3}aT^2 + bT - 2(\frac{1}{T}\sum_{m=1}^{N^k_T} t^k_m). \tag{A.7}$$

Define the Langrange multiplier of problem P_k by

$$\mathcal{L}^{(k)}_T(b, a, L_0, L_1) := \mathcal{R}^{(k)}_T(b, a) - 2bL_0 - 2aL_1. \tag{A.8}$$

The primal and dual optimal variables must satisfy the following KKT conditions.

1. Stationarity:

$$\frac{\partial \mathcal{L}^{(k)}_T}{\partial b} = aT + 2b - 2\eta^k_T - 2L_0, \qquad \frac{\partial \mathcal{L}^{(k)}_T}{\partial a} = \frac{2}{3}aT^2 + bT - 2(\frac{1}{T}\sum_{m=1}^{N^k_T} t^k_m) - 2L_1. \tag{A.9}$$

2. Primal feasibility: $b \geq 0$, $a \geq 0$.

3. Dual feasibility: $L_0 \geq 0$, $L_1 \geq 0$.

4. Complementary slackness: $bL_0 = 0$,$aL_1 = 0$.

First, we discuss the different possibilities for the complementary slackness conditions to be satisfied. The optimum cannot be $(a, b) = (0, 0)$. By contradiction, if this was the case, then the stationarity conditions imply $L_0 = -\eta_T^k$, and $L_1 = -\frac{1}{T}\sum_{m=1}^{N_T^k} t_m^k$. It is clear that dual feasibility is not respected, leading to a contradiction. Therefore, there are only 3 possible cases satisfying the slackness conditions:

1. $(L_0, L_1) = (0, 0)$. In this case, the stationarity conditions imply

$$a = \frac{12}{T^2}\left(\frac{1}{T}\sum_{m=1}^{N_T^k} t_m^k - \frac{N_T^k}{2}\right), \qquad b = \frac{6}{T}\left(\frac{2}{3}N_T^k - \frac{1}{T}\sum_{m=1}^{N_T^k} t_m^k\right). \tag{A.10}$$

2. $(L_0, a) = (0, 0)$. In this case, the stationarity conditions imply

$$b = \eta_T^k, \qquad L_1 = \frac{N_T^k}{2} - \frac{1}{T}\sum_{m=1}^{N_T^k} t_m^k. \tag{A.11}$$

3. $(b, L_1) = (0, 0)$. In this case, the stationarity conditions imply

$$L_0 = -\frac{3}{2T}\left(\frac{2}{3}N_T^k - \frac{1}{T}\sum_{m=1}^{N_T^k} t_m^k\right), \quad a = \frac{3}{T^3}\sum_{m=1}^{N_T^k} t_m^k. \tag{A.12}$$

If $\tau_k < \frac{1}{2}$, the primal feasibility condition is not satisfied in case 1, and dual feasibility is not satisfied in case 3. All KKT conditions are satisfied in case 2. If $\tau_k \in (\frac{1}{2}, \frac{2}{3})$, then dual feasibility is not satisfied in case 2 and case 3, and all KKT conditions are satisfied in case 1. Finally, if $\tau_k \in (\frac{2}{3}, 1)$, then primal feasibility is not satisfied for the variable b in case 1, an dual feasibility is not satisfied for the variable L_1 in case 2, whereas all KKT conditions are satisfied in case 3. □

A.2 Multivariate Hawkes processes

A.2.1 Residuals

Proof of Proposition 3.1.4. Let $m \in [N_T^k]$. Since $\hat{s}_m^{(k)} > s_m^{(k)}$, the absolute estimation error is

$$|\hat{s}_m^{(k)} - s_m^{(k)}| = \sum_{n=1}^{m-c-1} \left(\omega_{ki} - \psi_{ki}(t_m^k - t_n^i)\right) = \omega_{ki} \sum_{n=1}^{m-c-1} \left(1 - \tilde{\psi}_{ki}(t_m^k - t_n^i)\right). \tag{A.13}$$

The estimation error relative to $\hat{s}_m^{(k)}$ verifies the inequality

$$\begin{aligned}\frac{|\hat{s}_m^{(k)} - s_m^{(k)}|}{\hat{s}_m^{(k)}} &\leq \frac{|\hat{s}_m^{(k)} - s_m^{(k)}|}{(m-c-1)\omega_{ki}} \\ &\leq \frac{1}{m-c-1}\sum_{n=1}^{m-c-1}\left(1-\tilde{\psi}_{ki}(t_m^k - t_n^i)\right) \\ &\leq \frac{1}{m-c-1}\sum_{n=1}^{m-c-1}\left(1-\tilde{\psi}_{ki}(t_m^k - t_{m-c-1}^i)\right) \\ &\leq 1-\tilde{\psi}_{ki}(t_m^k - t_{m-c-1}^i) \leq 1-\tilde{\psi}_{ki}(\tau_c).\end{aligned} \tag{A.14}$$

Hence $\frac{|\hat{s}_m^{(k)} - s_m^{(k)}|}{\hat{s}_m^{(k)}} \leq \epsilon_c$. We are interested in the estimation error relative to the ground truth $s_m^{(k)}$, that is

$$\frac{|\hat{s}_m^{(k)} - s_m^{(k)}|}{s_m^{(k)}} = \frac{\frac{|\hat{s}_m^{(k)} - s_m^{(k)}|}{\hat{s}_m^{(k)}}}{1 - \frac{|\hat{s}_m^{(k)} - s_m^{(k)}|}{\hat{s}_m^{(k)}}}. \tag{A.15}$$

The function $x \mapsto \frac{x}{1-x}$ is a monotonically increasing bijection from $[0,1)$ to $[0,+\infty)$, therefore

$$\frac{|\hat{s}_m^{(k)} - s_m^{(k)}|}{s_m^{(k)}} \leq \frac{\epsilon_c}{1-\epsilon_c}. \tag{A.16}$$

Conversely, fix $\epsilon \in [0,1)$. If $\tau_c \geq \tilde{\psi}_{ki}^{-1}(1/(1+\epsilon))$, then $1-\tilde{\psi}_{ki}(\tau_c) \leq \epsilon/(1+\epsilon)$. Therefore,

$$\epsilon_c \leq \epsilon/(1+\epsilon), \quad i.e. \quad \frac{\epsilon_c}{1-\epsilon_c} \leq \epsilon. \tag{A.17}$$

□

A.2.2 Model identifiability

Proof of Proposition 3.1.2. Let $\boldsymbol{\lambda^{(A)}}$ (resp. $\boldsymbol{\lambda^{(B)}}$) be a 1-dimensional conditional intensity vector of MHP with baselines and kernels $(\mu_1^{(A)}, \phi_{11}^{(A)})$ (resp. $(\mu_1^{(B)}, \phi_{11}^{(B)})$), and assume that

$$\lambda_1^{(A)}(t) = \lambda_1^{(B)}(t), \quad \forall t \in [0,T]. \tag{A.18}$$

By evaluating Equation (A.18) for $t < t_1^1$, we get $\mu_1^{(A)} = \mu_1^{(B)}$. Therefore,

$$\int_0^t \phi_{11}^{(A)}(t-s)\mathrm{d}N_s^1 = \int_0^t \phi_{11}^{(B)}(t-s)\mathrm{d}N_s^1, \quad \forall t \in [0,T]. \tag{A.19}$$

For $m \in [\![2, N_T^1]\!]$, define the assertion

$$H^{(1)}(m): \quad \phi_{11}^{(A)}(t) = \phi_{11}^{(B)}(t), \quad \forall t \in [0, t_m^1 - t_1^1]. \tag{A.20}$$

We now prove by induction that for all $m \in [\![2, N_T^1]\!]$, the assertion $H^{(1)}(m)$ is true. For $m = 2$, by evaluating Equation (A.19) for $t \in (t_1^1, t_2^1]$, we get $\phi_{11}^{(A)}(s) = \phi_{11}^{(B)}(s)$, $\forall s \in (0, t_2^1 - t_1^1]$.

Since the kernel functions are right continuous, $\phi_{11}^{(A)}(s) = \phi_{11}^{(B)}(s)$, $\forall s \in [0, t_2^1 - t_1^1]$, and the assertion $H^{(1)}(2)$ is true. Now let $m \in [\![2, N_T^1 - 1]\!]$ such that $H^{(1)}(m)$ is true. We want to prove that $H^{(1)}(m+1)$ is also true. First, by evaluating Equation (A.19) for $t \in (t_m^1, t_{m+1}^1]$, and $s \in (0, t_{m+1}^1 - t_m^1]$, we get

$$\phi_{11}^{(A)}(s + t_m^1 - t_1^1) + \sum_{n=2}^{m} \phi_{11}^{(A)}(s + t_m^1 - t_n^1) = \phi_{11}^{(B)}(s + t_m^1 - t_1^1) + \sum_{n=2}^{m} \phi_{11}^{(B)}(s + t_m^1 - t_n^1). \quad \text{(A.21)}$$

Now, let $s \in (0, \min(t_{m+1}^1 - t_m^1, t_2^1 - t_1^1)]$. We note that $s + t_m^1 - t_n^1 \in [0, t_m^1 - t_1^1]$, $\forall n \in [\![2, m]\!]$. Therefore, $H^{(1)}(m+1)$ implies that

$$\phi_{11}^{(A)}(s + t_m^1 - t_n^1) = \phi_{11}^{(B)}(s + t_m^1 - t_n^1) \quad \forall n \in [\![2, m]\!]. \quad \text{(A.22)}$$

In particular, this implies that

$$\sum_{n=2}^{m} \phi_{11}^{(A)}(s + t_m^1 - t_n^1) = \sum_{n=2}^{m} \phi_{11}^{(B)}(s + t_m^1 - t_n^1). \quad \text{(A.23)}$$

By plugging this in Equation (A.21), we get $\phi_{11}^{(A)}(s + t_m^1 - t_1^1) = \phi_{11}^{(B)}(s + t_m^1 - t_1^1)$. By change of variable, this implies that, for all $x \in (t_m^1 - t_1^1, t_m^1 - t_1^1 + \min(t_{m+1}^1 - t_m^1, t_2^1 - t_1^1)]$,

$$\phi_{11}^{(A)}(x) = \phi_{11}^{(B)}(x).$$

Since the kernel functions are right continuous, this last equation is true for all $x \in [t_m^1 - t_1^1, t_m^1 - t_1^1 + \min(t_{m+1}^1 - t_m^1, t_2^1 - t_1^1)]$. Since $H_m^{(1)}$ is true, for all $x \in [0, t_m^1 - t_1^1 + \min(t_{m+1}^1 - t_m^1, t_2^1 - t_1^1)]$,

$$\phi_{11}^{(A)}(x) = \phi_{11}^{(B)}(x).$$

Now, for all $p \in \mathbb{N}^*$, define the assertion

$$H^{(2)}(p): \quad \left\{ \phi_{11}^{(A)}(x) = \phi_{11}^{(B)}(x), \quad \forall x \in \left[0, t_m^1 - t_1^1 + \min\left(t_{m+1}^1 - t_m^1, p(t_2^1 - t_1^1)\right)\right] \right\}. \quad \text{(A.24)}$$

We want to show by induction that, for all $p \in \mathbb{N}^*$, the assertion $H^{(2)}(p)$ is true. For $p = 1$, it is clear that $H^{(2)}(1)$ is true based on the above. Now let $p \in \mathbb{N}^*$ such that $H^{(2)}(p)$ is true. Let $s \in \left[0, \min\left(t_{m+1}^1 - t_m^1, (p+1)(t_2^1 - t_1^1)\right)\right]$. We note that for all $n \in [\![2, m]\!]$,

$$s + t_m^1 - t_n^1 \in \left[0, t_m^1 - t_1^1 + \min\left(t_{m+1}^1 - t_m^1, p(t_2^1 - t_1^1)\right)\right].$$

Assertion $H^{(2)}(p)$ implies

$$\phi_{11}^{(A)}(s + t_m^1 - t_n^1) = \phi_{11}^{(B)}(s + t_m^1 - t_n^1) \quad \forall n \in [\![2, m]\!].$$

Therefore,

$$\sum_{n=2}^{m} \phi_{11}^{(A)}(s + t_m^1 - t_n^1) = \sum_{n=2}^{m} \phi_{11}^{(B)}(s + t_m^1 - t_n^1). \tag{A.25}$$

By plugging this in Equation (A.21), we get

$$\phi_{11}^{(A)}(s + t_m^1 - t_1^1) = \phi_{11}^{(B)}(s + t_m^1 - t_1^1).$$

By change of variable, for all $x \in \left[0, t_m^1 - t_1^1 + \min\left(t_{m+1}^1 - t_m^1, (p+1)(t_2^1 - t_1^1)\right)\right]$

$$\phi_{11}^{(A)}(s + t_m^1 - t_1^1) = \phi_{11}^{(B)}(s + t_m^1 - t_1^1).$$

Hence $H^{(2)}(p+1)$ is true. This proves that for all $p \in \mathbb{N}^*$,

$$\left\{ \phi_{11}^{(A)}(x) = \phi_{11}^{(B)}(x), \quad \forall x \in \left[0, t_m^1 - t_1^1 + \min\left(t_{m+1}^1 - t_m^1, p(t_2^1 - t_1^1)\right)\right] \right\}.$$

Note that for $p > \lfloor \frac{t_{m+1}^1 - t_m^1}{t_2^1 - t_1^1} \rfloor$, $\min\left(t_{m+1}^1 - t_m^1, p(t_2^1 - t_1^1)\right) = t_{m+1}^1 - t_m^1$. Therefore, we conclude that

$$\left\{ \phi_{11}^{(A)}(x) = \phi_{11}^{(B)}(x), \quad \forall x \in \left[0, t_{m+1}^1 - t_1^1\right] \right\}.$$

Hence $H^{(1)}(m+1)$ is true. This proves that for all $m \in [\![2, N_T^1]\!]$

$$\phi_{11}^{(A)}(t) = \phi_{11}^{(B)}(t), \quad \forall t \in [0, t_m^1 - t_1^1]. \tag{A.26}$$

□

A.2.3 First order results

Proof of Proposition 3.1.1. This stability condition is equivalent to the convergence of the power sequence of the adjacency matrix:

$$\lim_{n \to +\infty} ||\Phi||_1^n = 0. \tag{A.27}$$

Now assume $d \geq 2$ and suppose that the matrix $||\Phi||_1$ is symmetric with strictly positive coefficients. Using the Perron–Frobenius theorem, the spectral radius $\rho(||\Phi||_1)$ is an eigenvalue of $||\Phi||_1$, with strictly positive eigenvector $\boldsymbol{X}$. By contradiction, suppose there exists $p, q \in [d]$ with $p < q$ such that $a_{pq} > 1$. Hence,

$$\begin{aligned} \omega_{pq} X_q + \omega_{pp} X_p + \sum_{i \in [d], i \notin \{p,q\}} \omega_{pi} X_i &= \rho(||\Phi||_1) X_p \\ \omega_{pq} X_p + \omega_{qq} X_q + \sum_{i \in [d], i \notin \{p,q\}} \omega_{qi} X_i &= \rho(||\Phi||_1) X_q. \end{aligned} \tag{A.28}$$

Summing both equations, we get

$$(\omega_{pp}+\omega_{pq})X_p+(\omega_{pq}+\omega_{qq})X_q+\sum_{i\in[d],i\notin\{p,q\}}(\omega_{pi}+\omega_{qi})X_i=\rho(||\Phi||_1)(X_p+X_q). \quad (A.29)$$

Since $\omega_{pq}>1$, $\omega_{pq}(X_p+X_q)>\rho(||\Phi||_1)(X_p+X_q)$. Therefore the LHS is strictly greater than the RHS, and we get a contradiction. □

Proof of Proposition 3.1.11. We discuss the eigenvalues of this bi-variate adjacency matrix model. If $d=2$ and $\omega_S=0$, then the matrix $J_{(\omega_S,\omega_C)}$ has a single eigenvalue, ω_C, of multiplicity 2. Otherwise, the matrix $J_{(\omega_S,\omega_C)}$ has 2 distinct eigenvalues: $\omega_C(d-1)+\omega_S$ is an eigenvalue of multiplicity 1, and $\omega_S-\omega_C$ is an eigenvalue of multiplicity $d-1$. This allows us to express simply the stability condition for the bi-variate matrix model $J_{(\omega_S,\omega_C)}$. □

Proof of Lemma 3.1.4. Let $n\in\mathbb{N}$. Since the matrices $\mathbb{I}_d$ and J commute, the binomial formula gives

$$\begin{aligned}J^n_{(\omega_S,\omega_C)}&=\sum_{k=0}^{n}\binom{n}{k}(\omega_S-\omega_C)^{n-k}\omega_C^k J^k,\\&=(\omega_S-\omega_C)^n\mathbb{I}_d+\sum_{k=1}^{n}\binom{n}{k}(\omega_S-\omega_C)^{n-k}\omega_C^k J^k.\end{aligned} \quad (A.30)$$

Using $J^k=d^{k-1}J$, we get

$$\begin{aligned}J^n_{(\omega_S,\omega_C)}&=(\omega_S-\omega_C)^n\mathbb{I}_d+\frac{1}{d}\sum_{k=1}^{n}\binom{n}{k}(\omega_S-\omega_C)^{n-k}\omega_C^k d^k J,\\&=(\omega_S-\omega_C)^n\mathbb{I}_d+\frac{1}{d}\left(\big(\omega_S+(d-1)\omega_C\big)^n-(\omega_S-\omega_C)^n\right)J.\end{aligned} \quad (A.31)$$

Since $\rho(J_{(\omega_S,\omega_C)})<1$, the power series $\sum_{n\geq 0}J^n_{(\omega_S,\omega_C)}$ converges to $(\mathbb{I}_d-J_{(\omega_S,\omega_C)})^{-1}$. And using Equation (A.31),

$$\begin{aligned}\sum_{n=0}^{+\infty}J^n_{(\omega_S,\omega_C)}&=\sum_{n=0}^{+\infty}(\omega_S-\omega_C)^n\mathbb{I}_d+\frac{1}{d}\left(\sum_{n=0}^{+\infty}\big(\omega_S+(d-1)\omega_C\big)^n-\sum_{n=0}^{+\infty}(\omega_S-\omega_C)^n\right)J,\\&=\frac{1}{1-(\omega_S-\omega_C)}\mathbb{I}_d+\frac{1}{d}\left(\frac{1}{1-(\omega_S+(d-1)\omega_C)}-\frac{1}{1-(\omega_S-\omega_C)}\right)J.\end{aligned} \quad (A.32)$$

Therefore,

$$(\mathbb{I}_d-J_{(\omega_S,\omega_C)})^{-1}=\frac{\omega_C}{(1-(\omega_S-\omega_C))\,(1-(\omega_S+(d-1)\omega_C))}J+\frac{1}{1-(\omega_S-\omega_C)}\mathbb{I}_d. \quad (A.33)$$

□

A.2.4 Second order results

Proof of Lemma 3.1.1. By definition, for all lags $\tau \geq 0$

$$\begin{aligned} f_{\mathrm{Tr}}^{(h)} * f(\tau) &= \int_{-\infty}^{+\infty} f_{\mathrm{Tr}}^{(h)}(z) f(\tau - z) \mathrm{d}z, \\ &= \int_{-h}^{h} \Big(1 - \frac{|z|}{h}\Big) f(\tau - z) \mathrm{d}z, \\ &= \int_{-h}^{h} f(\tau - z) \mathrm{d}z + \frac{1}{h} \int_{-h}^{0} z f(\tau - z) \mathrm{d}z - \frac{1}{h} \int_{0}^{h} z f(\tau - z) \mathrm{d}z, \\ &= \int_{\tau-h}^{\tau+h} f(z) \mathrm{d}z + \frac{1}{h} \int_{-h}^{0} z f(\tau - z) \mathrm{d}z - \frac{1}{h} \int_{0}^{h} z f(\tau - z) \mathrm{d}z. \end{aligned} \tag{A.34}$$

In order to compute these integrals, denote by F the primitive of f such that for all $x \in \mathbb{R}$

$$F(x) := \int_{0}^{x} f(y) \mathrm{d}y. \tag{A.35}$$

The first term is of course

$$\int_{\tau-h}^{\tau+h} f(z) \mathrm{d}z = F(\tau + h) - F(\tau - h). \tag{A.36}$$

We write the second term as

$$\frac{1}{h} \int_{-h}^{0} z f(\tau - z) \mathrm{d}z = \frac{\tau}{h} \int_{\tau}^{\tau+h} f(y) \mathrm{d}y - \frac{1}{h} \int_{\tau}^{\tau+h} y f(y) \mathrm{d}y. \tag{A.37}$$

By integration by parts we get

$$-\frac{1}{h} \int_{\tau}^{\tau+h} y f(y) \mathrm{d}y = -\frac{\tau}{h} \Big(F(\tau + h) - F(\tau) \Big) - F(\tau + h) + \frac{1}{h} \int_{\tau}^{\tau+h} F(y) \mathrm{d}y. \tag{A.38}$$

Hence the second term is

$$\frac{1}{h} \int_{-h}^{0} z f(\tau - z) \mathrm{d}z = -F(\tau + h) + \frac{1}{h} \int_{\tau}^{\tau+h} F(y) \mathrm{d}y. \tag{A.39}$$

Similarly, we get for the third term

$$\frac{1}{h} \int_{0}^{h} z f(\tau - z) \mathrm{d}z = F(\tau - h) - \frac{1}{h} \int_{\tau-h}^{\tau} F(y) \mathrm{d}y. \tag{A.40}$$

Eventually, we get

$$f_{\mathrm{Tr}}^{(h)} * f(\tau) = \frac{1}{h} \Big(\int_{\tau}^{\tau+h} F(y) \mathrm{d}y - \int_{\tau-h}^{\tau} F(y) \mathrm{d}y \Big). \tag{A.41}$$

□

Proof of Lemma 3.1.2. Since

$$\Big(\mathbb{I}_d - \mathcal{F}[\Phi]^*(x)\Big)^{-1} = \frac{1}{\det(\mathbb{I}_d - \mathcal{F}[\Phi](x))^*}\mathbf{adj}(\mathbb{I}_d - \mathcal{F}[\Phi]^*(x)), \tag{A.42}$$

and

$$\Big(\mathbb{I}_d - \mathcal{F}[\Phi]^\intercal(x)\Big)^{-1} = \frac{1}{\det(\mathbb{I}_d - \mathcal{F}[\Phi](x))}\mathbf{adj}(\mathbb{I}_d - \mathcal{F}[\Phi]^\intercal(x)), \tag{A.43}$$

we get the result from Equation (3.33). □

Proof of Lemma 3.1.3. Let $s \in \mathbb{R}$. Note that

$$|1 + \mathcal{F}[Z]|^2 - 1 = 2\Re(\mathcal{F}[Z]) + |\mathcal{F}[Z]|^2. \tag{A.44}$$

Since

$$\mathcal{F}^{-1}\Big[2\Re(\mathcal{F}[Z])\Big](s) = Z(s) + Z^{(-)}, \tag{A.45}$$

and

$$\mathcal{F}^{-1}\Big[|\mathcal{F}[Z]|^2\Big](s) = Z * Z(s), \tag{A.46}$$

we get the desired result. □

Proof of Proposition 3.1.5. Using Lemma 3.1.2, for all frequencies $x \in \mathbb{R}$,

$$\begin{aligned}\mathcal{F}\left[\nu^{(h)}\right](x) &= \eta_\star^1 \mathcal{F}\left[f_{\mathrm{Tr}}^{(h)}\right](x)(1 + \mathcal{Z}(x)),\\ &= \eta_\star^1 \mathcal{F}\left[f_{\mathrm{Tr}}^{(h)}\right](x) + \eta_\star^1 \mathcal{F}\left[f_{\mathrm{Tr}}^{(h)}\right](x)\mathcal{Z}(x).\end{aligned} \tag{A.47}$$

Therefore, in the time domain we get, for all lags $\tau \geq 0$,

$$\nu_\tau^{(h)} = \eta_\star^1 f_{\mathrm{Tr}}^{(h)}(\tau) + \eta_\star^1 f_{\mathrm{Tr}}^{(h)} * \mathcal{F}^{-1}\Big[\mathcal{Z}\Big](\tau). \tag{A.48}$$

Fix a lag $\tau \geq 0$. For a given function $f : \mathbb{R} \to \mathbb{R}$ with primitive F, Lemma 3.1.1 implies

$$f_{\mathrm{Tr}}^{(h)} * f(\tau) = \frac{1}{h}\Big(\int_\tau^{\tau+h} F(y)\mathrm{d}y - \int_{\tau-h}^{\tau} F(y)\mathrm{d}y\Big). \tag{A.49}$$

Now suppose $f = \mathcal{F}^{-1}\Big[\mathcal{Z}\Big]$. We need to compute a primitive of f. Let F be $F(y) := \int_0^y f(t)\mathrm{d}t$. Note that for all times $y \in \mathbb{R}$

$$F(x) = \mathbf{sign}(x)F(|x|). \tag{A.50}$$

Using Lemma 3.1.3, we get that for all times $y \in \mathbb{R}$

$$F(y) = \mathbf{sign}(y)Z^{(1)}(|y|) + \mathbf{sign}(y)Z^{(2)}(|y|). \tag{A.51}$$

It is clear that

$$\frac{1}{h}\int_\tau^{\tau+h} F(x)\mathrm{d}x = \frac{1}{h}\int_\tau^{\tau+h} Z^{(1)}(x)\mathrm{d}x + \frac{1}{h}\int_\tau^{\tau+h} Z^{(2)}(x)\mathrm{d}x. \tag{A.52}$$

If $\tau \geq h$, then

$$\frac{1}{h}\int_{\tau-h}^{\tau} F(x)\mathrm{d}x = \frac{1}{h}\int_{\tau-h}^{\tau} Z^{(1)}(x)\mathrm{d}x + \frac{1}{h}\int_{\tau-h}^{\tau} Z^{(2)}(x)\mathrm{d}x. \tag{A.53}$$

If $\tau < h$, then

$$\frac{1}{h}\int_{\tau-h}^{\tau} F(x)\mathrm{d}x = \frac{1}{h}\int_{-|\tau-h|}^{0} F(x)\mathrm{d}x + \frac{1}{h}\int_{0}^{\tau} F(x)\mathrm{d}x. \tag{A.54}$$

This implies

$$\frac{1}{h}\int_{\tau-h}^{\tau} F(x)\mathrm{d}x = \frac{1}{h}\int_{0}^{\tau} (Z^{(1)}(x) + Z^{(2)}(x))\mathrm{d}x - \frac{1}{h}\int_{0}^{|\tau-h|} (Z^{(1)}(x) + Z^{(2)}(x))\mathrm{d}x. \tag{A.55}$$

Hence, if $\tau < h$

$$\frac{1}{h}\int_{\tau-h}^{\tau} F(x)\mathrm{d}x = \begin{cases} \frac{1}{h}\int_{h-\tau}^{\tau}(Z^{(1)}(x) + Z^{(2)}(x))\mathrm{d}x & \text{if } \tau \in [\frac{h}{2}, h), \\ -\frac{1}{h}\int_{\tau}^{h-\tau}(Z^{(1)}(x) + Z^{(2)}(x))\mathrm{d}x & \text{if } \tau \in [0, \frac{h}{2}). \end{cases} \tag{A.56}$$

Therefore, we get

$$f_{\mathrm{Tr}}^{(h)} * f(\tau) = \begin{cases} \frac{1}{h}(\int_{[\tau,\tau+h]}(Z^{(1)} + Z^{(2)}) - \int_{[\tau-h,\tau]}(Z^{(1)} + Z^{(2)})) & \text{if } \tau \in [h, +\infty), \\ \frac{1}{h}(\int_{[\tau,\tau+h]}(Z^{(1)} + Z^{(2)}) - \int_{[h-\tau,\tau]}(Z^{(1)} + Z^{(2)}) & \text{if } \tau \in [\frac{h}{2}, h), \\ \frac{1}{h}(2\int_{[\tau,h-\tau]}(Z^{(1)} + Z^{(2)}) + \int_{[h-\tau,h+\tau]}(Z^{(1)} + Z^{(2)})) & \text{if } \tau \in [0, \frac{h}{2}). \end{cases} \tag{A.57}$$

The desired result follows immediately. □

A.2.5 Least-squares fit

Proof of Proposition 3.2.1. By definition

$$\mathcal{R}_T^{(k)}(\boldsymbol{\theta_k}) = \frac{1}{T}\int_0^T \lambda_k^2 - \frac{2}{T}\int_0^T \lambda_k(t)\mathrm{d}N_t^k. \tag{A.58}$$

Recall that for times $t \geq 0$, $\lambda_k(t) = \boldsymbol{f_k}^\intercal \boldsymbol{\varphi^{(k)}(t)}$. First, we get

$$\frac{1}{T}\int_0^T \lambda_k(t)\mathrm{d}N_t^k = \boldsymbol{f_k}^\intercal \frac{1}{T}\int_0^T \boldsymbol{\varphi^{(k)}(t)}\mathrm{d}N_t^k = \boldsymbol{f_k}^\intercal \boldsymbol{c_k}. \tag{A.59}$$

Second, for times $t \geq 0$

$$\lambda_k^2(t) = \Big(\boldsymbol{f_k}^\intercal \boldsymbol{\varphi^{(k)}(t)}\Big)^2 = \boldsymbol{f_k}^\intercal \boldsymbol{\varphi^{(k)}(t)}\big(\boldsymbol{\varphi^{(k)}(t)}\big)^\intercal \boldsymbol{f_k}. \tag{A.60}$$

Hence

$$\frac{1}{T}\int_0^T \lambda_k^2 = \Big(\boldsymbol{f_k}^\intercal \Big(\frac{1}{T}\int_0^T \boldsymbol{\varphi^{(k)}\varphi^{(k)\intercal}}\Big)\boldsymbol{f_k} \tag{A.61}$$

Since $Q = \frac{1}{T}\int_0^T \boldsymbol{\varphi^{(k)}\varphi^{(k)\intercal}}$, we get the desired result. □

Proof of Proposition 3.2.3. Using the gradient formula, the stationary condition with respect to first order parameters is characterized by the linear system

$$\nabla_{\boldsymbol{f_k}} \mathcal{R}_T^{(k)} = 0 \iff Q_k \boldsymbol{f_k} = \boldsymbol{c_k}. \tag{A.62}$$

If the matrix Q_k is positive definite, the stationarity equation admits a unique solution $\bar{\boldsymbol{f}}_{\boldsymbol{k}} = Q_k^{-1} \boldsymbol{c_k}$. If the matrix Q_k is positive semi-definite, using the diagonalization of the matrix Q_k in the stationarity equation we get $DP^\intercal \boldsymbol{f_k} = P^\intercal \boldsymbol{c_k}$. The set of solutions of the linear equation $D\boldsymbol{u} = P^\intercal \boldsymbol{c_k}$ of unknown $\boldsymbol{u}$ is

$$\{D^+ P^\intercal \boldsymbol{c_k} + \boldsymbol{v}, \quad \boldsymbol{v} \in \ker D\}. \tag{A.63}$$

Therefore, the solutions of $DP^\intercal \boldsymbol{f_k} = P^\intercal \boldsymbol{c_k}$ are the set $PD^+ P^\intercal \boldsymbol{c_k} + P \ker D$. It is simple to show by double inclusion that $P \ker D = \ker Q$, hence the desired result. □

Proof of Proposition 3.2.4. In order to prove this result, we distinguish three cases depending on the value of the temporal correlation $\frac{q_{11}^{(XY,1)}}{q_{11,11}^{(XX,1)}}$, and for each case, identify which parameter values satisfy the slackness condition and the other KKT conditions.

First, consider the case of negative temporal covariance: $q_{11}^{(XY,1)} < 0$. Note that

$$q_{11}^{(XY,1)} < 0 \iff \frac{m_{11}^{(XY,1)}}{\left(m_{11}^{(X,1)}\right)^2} < \frac{m^{(Y,1)}}{m_{11}^{(X,1)}}. \tag{A.64}$$

Using the Cauchy–Schwarz inequality, the temporal variance satisfies $m_{11,11}^{(XX,1)} > \left(m_{11}^{(X,1)}\right)^2$; therefore

$$q_{11}^{(XY,1)} < 0 \implies \frac{m_{11}^{(XY,1)}}{m_{11,11}^{(XX,1)}} < \frac{m^{(Y,1)}}{m_{11}^{(X,1)}}. \tag{A.65}$$

We discuss the different parameter values that satisfy the slackness condition and confront them to the other KKT conditions. By contradiction, assume $(\mu_1^\star, L_1^\star) = (0, 0)$. The stationarity condition implies

$$L_0^\star = m_{11}^{(X,1)} \left(\frac{m_{11}^{(XY,1)}}{m_{11,11}^{(XX,1)}} - \frac{m^{(Y,1)}}{m_{11}^{(X,1)}} \right). \tag{A.66}$$

Hence $L_0^\star < 0$, contradicting the dual feasibility condition, therefor $(\mu_1^\star, L_1^\star) \neq (0, 0)$. By contradiction, assume $(L_0^\star, L_1^\star) = (0, 0)$. The stationarity condition implies $\omega_{11}^\star = \frac{q_{11}^{(XY,1)}}{q_{11,11}^{(XX,1)}}$. Hence $\omega_{11}^\star < 0$, contradicting the primal feasibility condition. Therefore, $(L_0^\star, L_1^\star) \neq (0, 0)$. The last possibility for the slackness condition to be satisfied is if $(L_0^\star, \omega_{11}^\star) = (0, 0)$. If we set $\mu = m^{(Y,1)}$, and $L_1 = -q_{11}^{(XY,1)}$; then all KKT conditions are satisfied, therefore we get the solution in the negative temporal covariance case.

Second, consider the case of temporal correlation satisfying: $\frac{q_{11}^{(XY,1)}}{q_{11,11}^{(XX,1)}} \leq \frac{m^{(Y,1)}}{m_{11}^{(X,1)}}$. Note that

$$\frac{q_{11}^{(XY,1)}}{q_{11,11}^{(XX,1)}} > \frac{m^{(Y,1)}}{m_{11}^{(X,1)}} \iff \frac{m_{11}^{(XY,1)}}{m_{11,11}^{(XX,1)}} > \frac{m^{(Y,1)}}{m_{11}^{(X,1)}}. \tag{A.67}$$

We discuss the different parameter values that satisfy the slackness condition and confront them to the other KKT conditions. By contradiction, assume $(L_0^\star, \omega_{11}^\star) = (0,0)$. The stationarity conditions imply

$$L_1^\star = -q_{11}^{(XY,1)}. \tag{A.68}$$

Hence $L_1 < 0$, contradicting the dual feasibility condition. By contradiction, assume $(\mu_1^\star, L_1^\star) = (0,0)$. The stationarity conditions imply

$$L_0^\star = m_{11}^{(X,1)} \left(\frac{m_{11}^{(XY,1)}}{m_{11,11}^{(XX,1)}} - \frac{m^{(Y,1)}}{m_{11}^{(X,1)}} \right), \quad \omega^\star = \frac{q_{11}^{(XY,1)}}{q_{11,11}^{(XX,1)}}, \tag{A.69}$$

Using Equation (A.67), we conclude that $\mu_1^\star < 0$, contradicting the primal feasibility condition. The last possibility for the slackness condition to be satisfied is if $(L_0^\star, L_1^\star) = (0,0)$. If we set $\mu_1^\star = m^{(Y,1)} - m_{11}^{(X,1)} \frac{q_{11}^{(XY,1)}}{q_{11,11}^{(XX,1)}}$, then all KKT conditions are satisfied, therefore we get the solution in the excess temporal correlation case.

Finally, consider the case of excess temporal correlation: $\frac{q_{11}^{(XY,1)}}{q_{11,11}^{(XX,1)}} > \frac{m^{(Y,1)}}{m_{11}^{(X,1)}}$. Note that

$$\frac{q_{11}^{(XY,1)}}{q_{11,11}^{(XX,1)}} > \frac{m^{(Y,1)}}{m_{11}^{(X,1)}} \iff \frac{m_{11}^{(XY,1)}}{m_{11,11}^{(XX,1)}} > \frac{m^{(Y,1)}}{m_{11}^{(X,1)}}. \tag{A.70}$$

We discuss the different parameter values that satisfy the slackness condition and confront them to the other KKT conditions. By contradiction, assume $(L_0^\star, \omega_{11}^\star) = (0,0)$. The stationarity conditions imply

$$L_1^\star = -q_{11}^{(XY,1)}. \tag{A.71}$$

Hence $L_1 < 0$, contradicting the dual feasibility condition. By contradiction, assume $(L_0^\star, L_1^\star) = (0,0)$. The stationarity conditions imply $\mu_1^\star = m^{(Y,1)} - m_{11}^{(X,1)} \frac{q_{11}^{(XY,1)}}{q_{11,11}^{(XX,1)}}$. Using Equation (A.70), we conclude that $\mu_1^\star < 0$, contradicting the primal feasibility condition. The last possibility for the slackness condition to be satisfied is if $(\mu_1^\star, L_1^\star) = (0,0)$. If we set

$$L_0^\star = m_{11}^{(X,1)} \left(\frac{m_{11}^{(XY,1)}}{m_{11,11}^{(XX,1)}} - \frac{m^{(Y,1)}}{m_{11}^{(X,1)}} \right), \quad \omega^\star = \frac{q_{11}^{(XY,1)}}{q_{11,11}^{(XX,1)}}, \tag{A.72}$$

then all KKT conditions are satisfied, therefore we get the solution in the excess temporal correlation case. □

Proof of Lemma 3.2.1. If $\frac{q_{11}^{(XY,1)}}{q_{11,11}^{(XX,1)}} \in \left[0, \frac{m^{(Y,1)}}{m_{11}^{(X,1)}}\right)$, then $\frac{\left(q_{11}^{(XY,1)}\right)^2}{q_{11,11}^{(XX,1)}} < \left(\frac{m^{(Y,1)}}{m_{11}^{(X,1)}}\right)^2 q_{11,11}^{(XX,1)}$. Hence

$$-\frac{\left(q_{11}^{(XY,1)}\right)^2}{q_{11,11}^{(XX,1)}} - \left(m^{(Y,1)}\right)^2 < -\left(m^{(Y,1)}\right)^2\left(1 + \frac{q_{11,11}^{(XX,1)}}{\left(m_{11}^{(X,1)}\right)^2}\right). \tag{A.73}$$

The result follows immediately. If $\frac{q_{11}^{(XY,1)}}{q_{11,11}^{(XX,1)}} \geq \frac{m^{(Y,1)}}{m_{11}^{(X,1)}}$, then $\frac{m_{11}^{(XY,1)}}{m_{11,11}^{(XX,1)}} \geq \frac{m^{(Y,1)}}{m_{11}^{(X,1)}}$, and the result follows immediately. □

Proof of Proposition 3.2.6. By definition,

$$\bar{\boldsymbol{\eta}}_{\star} := (\mathbb{I}_d - ||\Phi^{\star}||_1)^{-1}\boldsymbol{\mu}^{\star}. \tag{A.74}$$

Using Proposition 3.2.5, we get

$$\begin{aligned}
\bar{\boldsymbol{\eta}}_{\star} &= (\mathbb{I}_d - ||\Phi^{\star}||_1)^{-1}\big(\mathbb{I}_d - F\big)\boldsymbol{\eta}_{\boldsymbol{T}}, \\
&= (\mathbb{I}_d - ||\Phi^{\star}||_1)^{-1}\big(\mathbb{I}_d - ||\Phi^{\star}||_1 + ||\Phi^{\star}||_1 - F\big)\boldsymbol{\eta}_{\boldsymbol{T}}, \\
&= \boldsymbol{\eta}_{\boldsymbol{T}} + (\mathbb{I}_d - ||\Phi^{\star}||_1)G\boldsymbol{\eta}_{\boldsymbol{T}}.
\end{aligned} \tag{A.75}$$

□

A.3 The ASLSD method

A.3.1 LSE decomposition

We now prove the LSE decomposition of Theorem 4.3.1. Consider a $d-$dimensional MTLH model and fix an event type $k \in [d]$. In Section 3.2.1, we define the feature processes of an MHP; we now extend this definition to the MTLH case. Fix types $i, j \in [d]$. For times $t \geq 0$, the feature process $\varphi_j^{(k)}$ is the stochastic convolution

$$\varphi_j^{(k)}(t) := \int_0^t \tilde{\phi}_{kj}(t-s)\mathcal{I}_{kj}(\xi^j_{N_s^j})\mathrm{d}N_s^j.$$

Note that $\varphi_j^{(k)}(t) = 0$ for times $t \leq t_1^j$. Define

$$C_j^{(k)}(t) := \int_0^t \mu_k(u)\varphi_j^{(k)}(u)\mathrm{d}u, \qquad I_{ij}^{(k)}(t) := \int_0^t \sum_{l=1}^{r_{ki}}\sum_{l'=1}^{r_{kj}} \varphi_i^{(k)}(t)\varphi_j^{(k)}(t)\mathrm{d}t.$$

Lemma A.3.1 (Primitive of the feature process)**.** *For types $j \in [d]$ and times $t \geq 0$,*

$$C_j^{(k)}(T) = \sum_{n=1}^{N_T^j} K_{kj}(T - t_n^j, t_n^j)\mathcal{I}_{kj}(\xi_m^j). \tag{A.76}$$

Proof. Fix $j \in [d]$. By definition

$$C_j^{(k)}(T) = \int_0^T \mu_k(t) \sum_{n=1}^{\kappa(j,t)} \phi_{kj}(t - t_n^j)\mathcal{I}_{kj}(\xi_n^j)\mathrm{d}t = \int_{t_1^j}^T \sum_{n=1}^{\kappa(j,t)} \mu_k(t)\phi_{kj}(t - t_n^j)\mathcal{I}_{kj}(\xi_n^j)\mathrm{d}t.$$

Split the integral

$$C_{j,l}^{(k)}(T) = \int_{t^j_{N_T^j}}^T \sum_{n=1}^{\kappa(j,t)} \phi_{kj}(t - t_n^j)\mu_k(t)\mathcal{I}_{kj}(\xi_n^j)\mathrm{d}t + \sum_{m=1}^{N_T^j-1} \int_{t_m^j}^{t_{m+1}^j} \sum_{n=1}^{\kappa(j,t)} \phi_{kj}(t - t_n^j)\mu_k(t)\mathcal{I}_{kj}(\xi_n^j)\mathrm{d}t.$$

For times $t \in [\, t^j_{N_T^j}, T]$, $\kappa(j,t) = N_T^j$. For indices $m \in [\, N_T^j - 1]$, and for times $t \in [\, t_m^j, t_{m+1}^j]$ $\kappa(j,t) = m$. Hence,

$$C_{j,l}^{(k)}(T) = \int_{t^j_{N_T^j}}^T \sum_{n=1}^{N_T^j} \phi_{kj}(t - t_n^j)\mu_k(t)\mathcal{I}_{kj}(\xi_n^j)\mathrm{d}t + \sum_{m=1}^{N_T^j-1} \int_{t_m^j}^{t_{m+1}^j} \sum_{n=1}^{m} \phi_{kj}(t - t_n^j)\mu_k(t)\mathcal{I}_{kj}(\xi_n^j)\mathrm{d}t.$$

Use Fubini's Theorem to write

$$C_{j,l}^{(k)}(T) = \sum_{n=1}^{N_T^j} \int_{t^j_{N_T^j}}^T \phi_{kj}(t - t_n^j)\mu_k(t)\mathcal{I}_{kj}(\xi_n^j)\mathrm{d}t + \sum_{m=1}^{N_T^j-1} \sum_{n=1}^{m} \int_{t_m^j}^{t_{m+1}^j} \phi_{kj}(t - t_n^j)\mu_k(t)\mathcal{I}_{kj}(\xi_n^j)\mathrm{d}t.$$

Re-index the sums of the second term and use a change of variable to conclude

$$\begin{aligned} C_{j,l}^{(k)}(T) &= \sum_{n=1}^{N_T^j} \int_{t^j_{N_T^j} - t_n^j}^{T - t_n^j} \phi_{kj}(u)\mu_k(u + t_n^j)\mathcal{I}_{kj}(\xi_n^j)\mathrm{d}u \\ &\quad + \sum_{n=1}^{N_T^j-1} \int_0^{t^j_{N_T^j} - t_n^j} \phi_{kj}(u)\mu_k(u + t_n^j)\mathcal{I}_{kj}(\xi_n^j)\mathrm{d}u, \\ &= \sum_{n=1}^{N_T^j} \mathcal{I}_{kj}(\xi_n^j) \int_0^{T - t_n^j} \phi_{kj}(u)\mu_k(u + t_n^j)\mathrm{d}u. \end{aligned}$$

□

Lemma A.3.2 (Second moment of the feature process)**.** *For types $i, j \in [d]$ with $i \neq j$,*

$$\begin{aligned} I_{ii,ll'}^{(k)}(T) &= \sum_{m=1}^{N_T^i} \tilde{\Upsilon}_{iik,ll'}(T - t_m^i, 0) + 2\sum_{m=1}^{N_T^i} \sum_{n=1}^{m-1} \tilde{\Upsilon}_{iik,ll'}(T - t_m^i, t_m^i - t_n^i), \\ I_{ij,ll'}^{(k)}(T) &= \sum_{m=1}^{N_T^i} \sum_{n=1}^{\kappa(j,i,m)} \tilde{\Upsilon}_{ijk,ll'}(T - t_m^i, t_m^i - t_n^j) + \sum_{n=1}^{N_T^j} \sum_{m=1}^{\kappa(i,j,n)} \tilde{\Upsilon}_{jik,l'l}(T - t_n^j, t_n^j - t_m^i). \end{aligned} \tag{A.77}$$

Proof. Fix event types $i, j \in [d]$ with $i \neq j$. First, define the ordered times of events of type i or j

$$(t_q)_{q\in\left[N_T^i+N_T^j\right]} := \left(t_m^i\right)_{m\in\left[N_T^i\right]} \cup \left(t_n^j\right)_{n\in\left[N_T^j\right]},$$

such that for $q \in \left[N_T^i + N_T^j - 1\right]$, $t_q < t_{q+1}$. For all events $q \in [N_T^i + N_T^j]$, denote by $\varsigma(q) \in \{i, j\}$ the type of the event, and by $\iota(q) \in \left[N_T^{\varsigma(q)}\right]$ its index, such that $t_q = t_{\iota(q)}^{\varsigma(q)}$. We prove by induction that for all event indices $q \in \left[N_T^i + N_T^j\right]$,

$$\begin{aligned} I_{ij}^{(k)}(t_q) = &\sum_{m=1}^{\kappa\left(i,\varsigma(q),\iota(q)\right)} \sum_{n=1}^{\kappa(j,i,m)} \int_{t_m^i}^{t_q} \phi_{ki}(t-t_m^i)\phi_{kj}(t-t_n^j)\mathcal{I}_{ki}(\xi_m^i)\mathcal{I}_{kj}(\xi_n^j)\mathrm{d}t \\ &+ \sum_{n=1}^{\kappa\left(j,\varsigma(q),\iota(q)\right)} \sum_{m=1}^{\kappa(i,j,n)} \int_{t_n^j}^{t_q} \phi_{ki}(t-t_m^i)\phi_{kj}(t-t_n^j)\mathcal{I}_{ki}(\xi_m^i)\mathcal{I}_{kj}(\xi_n^j)\mathrm{d}t. \end{aligned}$$

For $q = 1$, the sum is empty so the result is true. Let $q \in [N_T^i + N_T^j - 1]$, assume the property is true for q. It is clearly true for $q + 1$ using

$$\begin{aligned} I_{ij}^{(k)}(t_{q+1}) - I_{ij}^{(k)}(t_q) &= \int_{t_q}^{t_{q+1}} \varphi_{ki}(t)\varphi_{kj}(t)\mathrm{d}t, \\ &= \sum_{m=1}^{\kappa_i}\sum_{n=1}^{\kappa_j} \int_{t_q}^{t_{q+1}} \phi_{ki}(t-t_m^i)\phi_{kj}(t-t_n^j)\mathcal{I}_{ki}(\xi_m^i)\mathcal{I}_{kj}(\xi_n^j), \end{aligned}$$

where $\kappa_i := \kappa\big(i, \varsigma(q+1), \iota(q+1)\big)$ and $\kappa_j := \kappa\big(j, \varsigma(q+1), \iota(q+1)\big)$. Therefore,

$$\begin{aligned} I_{ij}^{(k)}(T) = &\sum_{m=1}^{N_T^i} \sum_{n=1}^{\kappa(j,i,m)} \int_{t_m^i}^{T} \phi_{ki}(t-t_m^i)\phi_{kj}(t-t_n^j)\mathcal{I}_{ki}(\xi_m^i)\mathcal{I}_{kj}(\xi_n^j)\mathrm{d}t \\ &+ \sum_{n=1}^{N_T^j} \sum_{m=1}^{\kappa(i,j,n)} \int_{t_n^j}^{T} \phi_{ki}(t-t_m^i)\phi_{kj}(t-t_n^j)\mathcal{I}_{ki}(\xi_m^i)\mathcal{I}_{kj}(\xi_n^j)\mathrm{d}t. \end{aligned}$$

Use the change of variables $x := t - t_m^i$ and $y := t - t_n^j$, to write

$$\begin{aligned} I_{ij}^{(k)}(T) = &\sum_{m=1}^{N_T^i} \sum_{n=1}^{\kappa(j,i,m)} \int_{0}^{T-t_m^i} \phi_{ki}(x)\phi_{kj}(x+t_m^i-t_n^j)\mathrm{d}x \\ &+ \sum_{n=1}^{N_T^j} \sum_{m=1}^{\kappa(i,j,n)} \int_{0}^{T-t_n^j} \phi_{kj}(y)\phi_{ki}(y+t_n^j-t_m^i)\mathcal{I}_{ki}(\xi_m^i)\mathcal{I}_{kj}(\xi_n^j)\mathrm{d}y. \end{aligned}$$

□

We can now prove the LSE decomposition of Theorem 4.3.1.

Proof of Theorem 4.3.1. Fix $\boldsymbol{\theta_k} \in \Theta_k$ and denote by λ_k the associated MTLH model. It is clear that

$$\sum_{m=1}^{N_T^k} \lambda_k\big(t_m^k\big) = \sum_{m=1}^{N_T^k} \mu_k\big(t_m^k\big) + \sum_{j=1}^{d} \sum_{m=1}^{N_T^k} \sum_{n=1}^{\kappa(j,k,m)} \phi_{kj}\big(t_m^k - t_n^j\big)\mathcal{I}_{kj}\big(\xi_n^j\big).$$

By definition

$$\begin{aligned}\frac{1}{T}\int_0^T \lambda_k(t)^2 \mathrm{d}t &= \frac{1}{T}\int_0^T \left(\mu_k^2(t) + 2\mu_k(t)\sum_{j=1}^{d}\varphi_j^{(k)}(t) + \left(\sum_{j=1}^{d}\varphi_j^{(k)}(t)\right)^2\right)\mathrm{d}t, \\ &= M_k(T) + \frac{2}{T}\sum_{j=1}^{d} C_j^{(k)}(T) + \frac{1}{T}\int_0^T \left(\sum_{j=1}^{d}\varphi_j^{(k)}(t)\right)^2 \mathrm{d}t.\end{aligned}$$

We note that

$$\int_0^T \left(\sum_{j=1}^{d}\varphi_j^{(k)}(t)\right)^2 \mathrm{d}t = \sum_{i=1}^{d}\sum_{j=1}^{d}\int_0^T \varphi_i^{(k)}(t)\varphi_j^{(k)}(t)\mathrm{d}t = \sum_{i=1}^{d}\sum_{j=1}^{d} Q_{ij}^{(k)}.$$

The result follows directly from the application of Lemma A.3.1 and Lemma A.3.2. □

A.3.2 Approximating model functionals

Proof of Proposition 4.4.1. Fix a time $t \geq 0$. By definition,

$$\psi_{ij,l}(t) := \omega_{ij,l}\int_0^t \tilde{\phi}_{ij,l}(u)\mathrm{d}u = \omega_{ij,l}\int_0^{+\infty} \mathbb{1}_{[0,t]}(u)\tilde{\phi}_{ij,l}(u)\mathrm{d}u. \tag{A.78}$$

By definition of the expectation operator,

$$\psi_{ij,l}(t) = \omega_{ijl}\mathrm{E}_{\tau\sim\tilde{\phi}_{ij,l}}\left[\,\mathbb{1}_{[0,t]}(\tau)\right]\,. \tag{A.79}$$

Using Equation (A.78), we get the partial derivative of $\psi_{ij,l}$ with respect to the L_1 weight parameter ω_{ijl}

$$\frac{\partial\psi_{ij,l}}{\partial\omega_{ij,l}}(t,s) = \tilde{\psi}_{ij,l}(t) = \mathbb{E}_{\tau\sim\tilde{\phi}_{ij,l}}\Big[\mathbb{1}_{[0,t]}(\tau)\Big], \tag{A.80}$$

and the partial derivative of $\psi_{ij,l}$ with respect to the kernel density parameter $\tilde{\theta}_{ij,lp}$

$$\begin{aligned}\frac{\partial\psi_{ij,l}}{\partial\tilde{\theta}_{ij,lp}}(t,s) &= \omega_{ij,l}\int_0^{+\infty}\mathbb{1}_{[0,t]}(u)\frac{\partial\tilde{\phi}_{ij,l}}{\partial\tilde{\theta}_{ij,lp}}(u)\mathrm{d}u, \\ &= \omega_{ij,l}\int_0^{+\infty}\mathbb{1}_{\{(u\in[0,t])\cap(\tilde{\phi}_{ij,l}(u)>0)\}}\frac{\partial\log\tilde{\phi}_{ij,l}}{\partial\tilde{\theta}_{ij,lp}}(u)\tilde{\phi}_{ij,l}(u)\mathrm{d}u, \\ &= \omega_{ijl}\mathbb{E}_{\tau\sim\tilde{\phi}_{ij,l}}\left[\mathbb{1}_{[0,t]}(\tau)\frac{\partial\log\tilde{\phi}_{ij,l}}{\partial\tilde{\theta}_{ij,lp}}(\tau)\right].\end{aligned} \tag{A.81}$$

□

Proof of Proposition 4.4.2. Fix lags $t, s \geq 0$. By definition,

$$\Upsilon_{ijk,ll'}(t,s) := \omega_{kil}\omega_{kjl'} \int_0^t \tilde{\phi}_{kil}(u)\tilde{\phi}_{kjl'}(u+s)\mathrm{d}u. \tag{A.82}$$

We re-write this as

$$\Upsilon_{ijk,ll'}(t,s) = \omega_{kil}\omega_{kjl'} \int_0^{+\infty} \mathbb{1}_{u\in[0,t]}\tilde{\phi}_{kjl'}(u+s)\tilde{\phi}_{kil}(u)\mathrm{d}u. \tag{A.83}$$

By definition of the expectation operator,

$$\Upsilon_{ijk,ll'}(t,s) = \omega_{kil}\omega_{kjl'}\mathrm{E}_{\tau\sim\tilde{\phi}_{ki,l}}\left[\mathbb{1}_{\tau[0,t]}\tilde{\phi}_{kj,l'}(\tau+s)\right]. \tag{A.84}$$

If $(i,l) = (j,l')$, then

$$\Upsilon_{iik,ll}(t,s) = \omega_{kil}^2 \int_0^{+\infty} \mathbb{1}_{u\in[0,t]}\tilde{\phi}_{kil}(u+s)\tilde{\phi}_{kil}(u)\mathrm{d}u. \tag{A.85}$$

Using Equation (A.85), the partial derivative of $\Upsilon_{iik,ll}$ with respect to the ω_{kil} is

$$\frac{\partial\Upsilon_{iik,ll}}{\partial\omega_{kil}}(t,s) = 2\omega_{kil}\mathrm{E}_{\tau\sim\tilde{\phi}_{ki,l}}\left[\mathbb{1}_{[0,t]}(\tau)\tilde{\phi}_{ki,l}(\tau+s)\right], \tag{A.86}$$

and the partial derivative of $\Upsilon_{iik,ll}$ with respect to the kernel density parameter $\tilde{\theta}_{kilp}$ is

$$\begin{aligned}\frac{\partial\Upsilon_{iik,ll}}{\partial\tilde{\theta}_{kilp}}(t,s) &= \omega_{kil}^2 \int_0^{+\infty} \mathbb{1}_{u\in[0,t]}\frac{\partial\tilde{\phi}_{kil}}{\partial\tilde{\theta}_{kilp}}(u+s)\tilde{\phi}_{kil}(u)\mathrm{d}u \\ &\quad + \omega_{kil}^2 \int_0^{+\infty} \mathbb{1}_{[0,t]}(u)\tilde{\phi}_{kil}(u+s)\frac{\partial\tilde{\phi}_{kil}}{\partial\tilde{\theta}_{kilp}}(u)\mathrm{d}u, \\ &= \omega_{kil}^2\mathrm{E}_{\tau\sim\tilde{\phi}_{ki,l}}\left[\mathbb{1}_{\tau\in[0,t]}\left(\frac{\partial\log\tilde{\phi}_{ki,l}}{\partial\tilde{\theta}_{kilp}}(\tau)\tilde{\phi}_{ki,l}(\tau+s) + \frac{\partial\tilde{\phi}_{ki,l}}{\partial\tilde{\theta}_{kilp}}(\tau+s)\right)\right].\end{aligned} \tag{A.87}$$

If $(i,l) \neq (j,l')$, then using Equation (A.83), we get the partial derivatives of $\Upsilon_{ijk,ll'}$ with respect to the L_1 weight parameters ω_{kil} and $\omega_{kjl'}$

$$\begin{aligned}\frac{\partial\Upsilon_{ijk,ll'}}{\partial\omega_{kil}}(t,s) &= \omega_{kjl'}\mathrm{E}_{\tau\sim\tilde{\phi}_{ki,l}}\left[\mathbb{1}_{\tau\in[0,t]}\tilde{\phi}_{kj,l'}(\tau+s)\right], \\ \frac{\partial\Upsilon_{ijk,ll'}}{\partial\omega_{kjl'}}(t,s) &= \omega_{kil}\mathrm{E}_{\tau\sim\tilde{\phi}_{ki,l}}\left[\mathbb{1}_{\tau\in[0,t]}\tilde{\phi}_{kj,l'}(\tau+s)\right].\end{aligned} \tag{A.88}$$

Using Equation (A.83), the derivative of $\Upsilon_{ijk,ll'}$ with respect to the kernel density parameter $\tilde{\theta}_{kilp}$ is

$$\begin{aligned}\frac{\partial\Upsilon_{ijk,ll'}}{\partial\tilde{\theta}_{kilp}}(t,s) &= \omega_{kil}\omega_{kjl'} \int_0^{+\infty} \mathbb{1}_{u\in[0,t]}\tilde{\phi}_{kjl'}(u+s)\frac{\partial\tilde{\phi}_{kil}}{\partial\tilde{\theta}_{kilp}}(u)\mathrm{d}u, \\ &= \omega_{kil}\omega_{kjl'} \int_0^{+\infty} \mathbb{1}_{(u\in[0,t])\cap(\tilde{\phi}_{kil}(u)>0)}\tilde{\phi}_{kjl'}(u+s)\frac{\partial\log\tilde{\phi}_{kil}}{\partial\tilde{\theta}_{kilp}}(u)\tilde{\phi}_{kil}(u)\mathrm{d}u, \\ &= \omega_{kil}\omega_{kjl'}\mathrm{E}_{\tau\sim\tilde{\phi}_{kil}}\left[\mathbb{1}_{\tau\in[0,t]}\tilde{\phi}_{kjl'}(\tau+s)\frac{\partial\log\tilde{\phi}_{kil}}{\partial\tilde{\theta}_{kilp}}(\tau)\right].\end{aligned} \tag{A.89}$$

Using Equation (A.83), the derivative of $\Upsilon_{ijk,ll'}$ with respect to the kernel density parameter $\tilde{\theta}_{kjl'q}$ is

$$\begin{aligned}\frac{\partial \Upsilon_{ijk,ll'}}{\partial \tilde{\theta}_{kjl'q}}(t,s) &= \omega_{kil}\omega_{kjl'}\int_0^{+\infty} \mathbb{1}_{u\in[0,t]}\frac{\partial \tilde{\phi}_{kjl'}}{\partial \tilde{\theta}_{kjl'q}}(u+s)\tilde{\phi}_{kil}(u)\mathrm{d}u, \\ &= \omega_{kil}\omega_{kjl'}\mathrm{E}_{\tau\sim\tilde{\phi}_{kil}}\left[\mathbb{1}_{\tau\in[0,t]}\frac{\partial \tilde{\phi}_{kjl'}}{\partial \tilde{\theta}_{kjl'q}}(\tau+s)\right].\end{aligned} \tag{A.90}$$

□

Proof of Proposition 4.4.3. Fix lags $t, s \geq 0$. By definition

$$K_{ki,l}(t,s) := \omega_{kil}\int_0^t \tilde{\phi}_{kil}(u)\mu_k(u+s)\mathrm{d}u. \tag{A.91}$$

We re-write this as

$$K_{ki,l}(t,s) = \omega_{kil}\int_0^{+\infty} \mathbb{1}_{[0,t]}(u)\mu_k(u+s)\tilde{\phi}_{kil}(u)\mathrm{d}u. \tag{A.92}$$

Using (A.92), the partial derivative of $K_{ki,l}$ with respect to ω_{kil} is

$$\frac{\partial K_{ki,l}}{\partial \omega_{kil}}(t,s) = \int_0^{+\infty} \mathbb{1}_{[0,t]}\mu_k(u+s)\tilde{\phi}_{kil}(u)\mathrm{d}u = \mathbb{E}_{\tau\sim\tilde{\phi}_{kil}}\Big[\mathbb{1}_{[0,t]}(\tau)\mu_k(\tau+s)\Big], \tag{A.93}$$

the partial derivative of $K_{ki,l}$ with respect to the kernel density parameter $\tilde{\theta}_{kilp}$ is

$$\begin{aligned}\frac{\partial K_{ki,l}}{\partial \tilde{\theta}_{kilp}}(t,s) &= \omega_{kil}\int_0^{+\infty} \mathbb{1}_{[0,t]}(u)\mu_k(u+s)\frac{\partial \tilde{\phi}_{kil}}{\partial \tilde{\theta}_{kilp}}(u)\mathrm{d}u, \\ &= \omega_{kil}\mathbb{E}_{\tau\sim\tilde{\phi}_{kil}}\Big[\mathbb{1}_{[0,t]}(\tau)\mu_k(\tau+s)\frac{\partial \log \tilde{\phi}_{kil}}{\partial \tilde{\theta}_{kilp}}(\tau)\Big],\end{aligned} \tag{A.94}$$

and the partial derivative of $K_{ki,l}$ with respect to the baseline parameter b_{kq} is

$$\begin{aligned}\frac{\partial K_{ki,l}}{\partial b_{kq}}(t,s) &= \omega_{kil}\int_0^{+\infty} \mathbb{1}_{[0,t]}(u)\frac{\partial \mu_k}{\partial b_{kq}}(u+s)\tilde{\phi}_{kil}(u)\mathrm{d}u, \\ &= \omega_{kil}\mathbb{E}_{\tau\sim\tilde{\phi}_{kil}}\left[\mathbb{1}_{[0,t]}(\tau)\frac{\partial \mu_k}{\partial b_{kilp}}(\tau+s)\right].\end{aligned} \tag{A.95}$$

□

Proof of Lemma 4.4.1. Using Theorem 4.3.1 and the results from this section, we get

$$\begin{aligned}\mathcal{R}_T^{(k)}(\theta_k) = &\sum_{i,j\in[d]}\sum_{m=\varpi(i,j)}^{N_T^i}\sum_{n=1}^{\kappa(j,i,m)}\frac{2}{T}\mathcal{I}_{ki}(\xi_m^i)\mathcal{I}_{kj}(\xi_n^j)\omega_{ki}\omega_{kj}\\ &\qquad\times \mathrm{E}_{\tau\sim\tilde{\phi}_{ki}}\left[\mathbb{1}_{[0,T-t_m^i]}(\tau)\tilde{\phi}_{kj}(\tau+t_m^i-t_n^j)\right]\\ &-\frac{2}{T}\left(\sum_{j=1}^{d}\sum_{m=\varpi(k,j)}^{N_T^k}\sum_{n=1}^{\kappa(j,k,m)}\phi_{kj}(t_m^k-t_n^j)\mathcal{I}_{kj}(\xi_n^j)\right)+M_k(T)-\frac{2}{T}\sum_{m=1}^{N_T^k}\mu_k(t_m^k)\\ &+\frac{2}{T}\sum_{i=1}^{d}\sum_{m=1}^{N_T^i}\mathcal{I}_{ki}(\xi_m^i)\omega_{ki}\mathrm{E}_{\tau\sim\tilde{\phi}_{ki}}\left[\mathbb{1}_{[0,T-t_m^i]}(\tau)\mu_k(t_m^i+\tau)\right]\\ &+\frac{1}{T}\sum_{i=1}^{d}\sum_{m=1}^{N_T^i}\omega_{ki}^2\mathcal{I}_{ki}(\xi_m^i)^2\mathrm{E}_{\tau\sim\tilde{\phi}_{ki}}\left[\mathbb{1}_{[0,T-t_m^i]}(\tau)\tilde{\phi}_{ki}(\tau)\right].\end{aligned} \tag{A.96}$$

The result follows directly by re-arranging the terms. □

A.4 The standard price model

Proof of Proposition 9.1.2. Let $n\in\mathbb{N}$.

$$\mathbb{P}(N_t^2-N_t^1=n)=\sum_{k=0}^{+\infty}\mathbb{P}(N_t^1=k)\mathbb{P}(N_t^2=n+k). \tag{A.97}$$

Developing this, we get

$$\mathbb{P}(N_t^2-N_t^1=n)=(\mu_2 t)^n e^{-(\mu_1+\mu_2)t}\sum_{k=0}^{+\infty}\frac{\left(\mu_1\mu_2 t^2\right)^k}{k!(k+n)!}. \tag{A.98}$$

Finally, by definition of the modified Bessel function of the first kind, we get

$$\mathbb{P}(N_t^2-N_t^1=n)=\left(\frac{\mu_2}{\mu_1}\right)^{\frac{n}{2}}e^{-(\mu_1+\mu_2)t}\sum_{k=0}^{+\infty}I_n\left(2\sqrt{\mu_1\mu_2}t\right). \tag{A.99}$$

By symmetry, it is clear that this holds for all $n\in\mathbb{Z}$. The result follows immediately. □

Bibliography

[1] Abergel, F., Bouchaud, J.-P., Foucault, T., Lehalle, C.-A., and Rosenbaum, M. (2012). *Market microstructure: confronting many viewpoints*. John Wiley & Sons.

[2] Achab, M., Bacry, E., Gaıffas, S., Mastromatteo, I., and Muzy, J.-F. (2017). Uncovering causality from multivariate Hawkes integrated cumulants. In *International Conference on Machine Learning*, pages 1–10. PMLR.

[3] Bacry, E., Bompaire, M., Gaïffas, S., and Muzy, J.-F. (2020). Sparse and low-rank multivariate Hawkes processes. *Journal of Machine Learning Research*, 21(50):1–32.

[4] Bacry, E., Bompaire, M., Gaïffas, S., and Poulsen, S. (2017). Tick: a python library for statistical learning, with a particular emphasis on time-dependent modelling. *arXiv preprint arXiv:1707.03003*.

[5] Bacry, E., Dayri, K., and Muzy, J.-F. (2012). Non-parametric kernel estimation for symmetric Hawkes processes. application to high frequency financial data. *The European Physical Journal B*, 85(5):1–12.

[6] Bacry, E., Delattre, S., Hoffmann, M., and Muzy, J.-F. (2013a). Modelling microstructure noise with mutually exciting point processes. *Quantitative finance*, 13(1):65–77.

[7] Bacry, E., Delattre, S., Hoffmann, M., and Muzy, J.-F. (2013b). Some limit theorems for Hawkes processes and application to financial statistics. *Stochastic Processes and their Applications*, 123(7):2475–2499.

[8] Bacry, E., Gaïffas, S., Mastromatteo, I., and Muzy, J.-F. (2016). Mean-field inference of Hawkes point processes. *Journal of Physics A: Mathematical and Theoretical*, 49(17):174006.

[9] Bacry, E., Mastromatteo, I., and Muzy, J.-F. (2015). Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(01):1550005.

[10] Bacry, E. and Muzy, J.-F. (2016). First-and second-order statistics characterization of Hawkes processes and non-parametric estimation. *IEEE Transactions on Information Theory*, 62(4):2184–2202.

[11] Baeza-Yates, R. and Ribeiro-Neto, B. (2010). *Modern Information Retrieval: The Concepts and Technology Behind Search*. ACM Press Books. Addison Wesley, 2 edition.

[12] Baldauf, M. and Mollner, J. (2020). High-frequency trading and market performance. *The Journal of Finance*, 75(3):1495–1526.

[13] Bartlett, M. S. (1967). The spectral analysis of line processes. In *Proc. Fifth Berkeley Symp. Math. Statist. Probab*, volume 3, pages 135–152.

[14] Benson, M. (2020). Predicting virality of online news articles using textual content.

[15] Bompaire, M., Bacry, E., and Gaïffas, S. (2018). Dual optimization for convex constrained objectives without the gradient-Lipschitz assumption. *arXiv preprint arXiv:1807.03545*.

[16] Bordenave, C. and Torrisi, G. L. (2007). Large deviations of poisson cluster processes. *Stochastic Models*, 23(4):593–625.

[17] Brémaud, P. and Massoulié, L. (1996). Stability of nonlinear Hawkes processes. *The Annals of Probability*, pages 1563–1588.

[18] Brogaard, J. (2010). High frequency trading and its impact on market quality. *Northwestern University Kellogg School of Management Working Paper*, 66.

[19] Brogaard, J., Hendershott, T., and Riordan, R. (2014). High-frequency trading and price discovery. *The Review of Financial Studies*, 27(8):2267–2306.

[20] Carrion, A. (2013). Very fast money: High-frequency trading on the nasdaq. *Journal of Financial Markets*, 16(4):680–711.

[21] Cartea, Á., Cucuringu, M., and Jin, Q. (2023). Detecting lead-lag relationships in stock returns and portfolio strategies. *Available at SSRN*.

[22] Cerqueira, V., Torgo, L., and Mozetič, I. (2020). Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning*, 109(11):1997–2028.

[23] Chen, F., Tan, W. H., et al. (2018). Marked self-exciting point process modelling of information diffusion on Twitter. *The Annals of Applied Statistics*, 12(4):2175–2196.

[24] Cinlar, E. (1975). *Introduction to Stochastic Processes*. Dover Publications.

[25] Citibank (2024 (accessed July 30, 2024)). *Citi's Depositary Receipt Services Glossary*. `https://depositaryreceipts.citi.com/adr/glossary/glossary.aspx?pageId=8&subpageID=101#American%20Depositary%20Receipts%20(ADRs)`.

[26] Cohen, S. N. and Szpruch, L. (2012). A limit order book model for latency arbitrage. *Mathematics and Financial Economics*, 6:211–227.

[27] Cont, R., Stoikov, S., and Talreja, R. (2010). A stochastic model for order book dynamics. *Operations research*, 58(3):549–563.

[28] Daley, D. J. and Vere-Jones, D. (2003). *An introduction to the theory of point processes. Vol. I. Probability and its Applications*. Springer-Verlag, New York.

[29] De Schryver, C. (2015). *FPGA Based Accelerators for Financial Applications*, volume 10. Springer.

[30] Défossez, A., Bottou, L., Bach, F., and Usunier, N. (2022). A simple convergence proof of adam and adagrad. *Transactions on Machine Learning Research*.

[31] Dong, Z., Cheng, X., and Xie, Y. (2022). Spatio-temporal point processes with deep non-stationary kernels. *arXiv preprint arXiv:2211.11179*.

[32] Dou, Y., Zhou, Y., and Xin, B. (2019). An accelerator for decoding market data based on fpga. *Journal of Circuits, Systems and Computers*, 28(03):1950050.

[33] Du, N., Dai, H., Trivedi, R., Upadhyay, U., Gomez-Rodriguez, M., and Song, L. (2016). Recurrent marked temporal point processes: Embedding event history to vector. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1555–1564.

[34] Etoré, P. and Jourdain, B. (2010). Adaptive optimal allocation in stratified sampling methods. *Methodology and Computing in Applied Probability*, 12(3):335–360.

[35] Forbes (2024 (accessed July 30, 2024)a). *Amazon Stock Split: What You Need To Know*. `https://www.forbes.com/advisor/investing/amazon-stock-split/`.

[36] Forbes (2024 (accessed July 30, 2024)b). *Google And GameStop Are Splitting Stocks*. `https://www.forbes.com/sites/qai/2022/08/10/google-and-gamestop-are-splitting-stocks-and-heres-what-it-means-for-investors/?sh=88d45127076c`.

[37] Gaïffas, S. and Guilloux, A. (2012). High-dimensional additive hazards models and the lasso. *Electronic Journal of Statistics*, 6:522–546.

[38] Gao, X., Zhou, X., and Zhu, L. (2018). Transform analysis for Hawkes processes with applications in dark pool trading. *Quantitative Finance*, 18(2):265–282.

[39] Gomez Rodriguez, M., Leskovec, J., and Schölkopf, B. (2012 (accessed July 30, 2024)). *MemeTracker dataset*. `http://snap.stanford.edu/infopath/data.html`.

[40] Gomez Rodriguez, M., Leskovec, J., and Schölkopf, B. (2013). Structure and dynamics of information pathways in online media. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 23–32.

[41] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

[42] Gould, M. D., Porter, M. A., Williams, S., McDonald, M., Fenn, D. J., and Howison, S. D. (2013). Limit order books. *Quantitative Finance*, 13(11):1709–1742.

[43] Hagströmer, B. and Nordén, L. (2013). The diversity of high-frequency traders. *Journal of Financial Markets*, 16(4):741–770.

[44] Hansen, N. R., Reynaud-Bouret, P., and Rivoirard, V. (2015). Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, 21(1):83–143.

[45] Hasbrouck, J. and Saar, G. (2013). Low-latency trading. *Journal of Financial Markets*, 16(4):646–679.

[46] Hawkes, A. G. (1971a). Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 33(3):438–443.

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

[48] Hawkes, A. G. (2018). Hawkes processes and their applications to finance: a review. *Quantitative Finance*, 18(2):193–198.

[49] Hawkes, A. G. and Oakes, D. (1974). A cluster process representation of a self-exciting process. *Journal of Applied Probability*, pages 493–503.

[50] Hegde, S., Lin, H., and Varshney, S. (2010). Competitive stock markets: Evidence from companies' dual listings on the nyse and nasdaq. *Financial Analysts Journal*, 66(1):77–87.

[51] Henderson, L. (2022). The Problem of Induction. In Zalta, E. N. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2022 edition.

[52] Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257.

[53] Jaisson, T., Rosenbaum, M., et al. (2015). Limit theorems for nearly unstable Hawkes processes. *The annals of applied probability*, 25(2):600–631.

[54] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[55] Kobayashi, R. and Lambiotte, R. (2016). Tideh: Time-dependent Hawkes process for predicting retweet dynamics. In *Tenth International AAAI Conference on Web and Social Media*.

[56] Lemonnier, R. and Vayatis, N. (2014). Nonparametric Markovian learning of triggering kernels for mutually exciting and mutually inhibiting multivariate Hawkes processes. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 161–176. Springer.

[57] Leskovec, J., Backstrom, L., and Kleinberg, J. (2009). Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506.

[58] Lewis, E. and Mohler, G. (2011). A nonparametric EM algorithm for multiscale Hawkes processes. *Journal of Nonparametric Statistics*, 1(1):1–20.

[59] Lewis, E., Mohler, G., Brantingham, P. J., and Bertozzi, A. L. (2012). Self-exciting point process models of civilian deaths in iraq. *Security Journal*, 25(3):244–264.

[60] Lewis, P. W. and Shedler, G. S. (1979). Simulation of nonhomogeneous poisson processes by thinning. *Naval research logistics quarterly*, 26(3):403–413.

[61] Liniger, T. J. (2009). *Multivariate Hawkes processes*. PhD thesis, ETH Zurich.

[62] Lobster (2024 (accessed July 30, 2024)). *Lobster*. https://lobsterdata.com/index.php.

[63] Lu, X. and Abergel, F. (2018). High-dimensional Hawkes processes for limit order books: modelling, empirical analysis and numerical calibration. *Quantitative Finance*, 18(2):249–264.

[64] Lu, X. and Szymanski, B. (2017). Predicting viral news events in online media. In *2017 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 1447–1456. IEEE.

[65] Manahov, V. (2016). A note on the relationship between high-frequency trading and latency arbitrage. *International review of financial analysis*, 47:281–296.

[66] Mei, H. and Eisner, J. M. (2017). The neural Hawkes process: A neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems*, pages 6754–6764.

[67] Menkveld, A. J. and Zoican, M. A. (2014). Need for speed? exchange latency and market quality. *Journal of Financial Economics*, 14:71–100.

[68] Menon, A. K. and Lee, Y. (2018). Proper loss functions for nonlinear Hawkes processes. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[69] Mohler, G. (2014). Marked point process hotspot maps for homicide and gun crime prediction in chicago. *International Journal of Forecasting*, 30(3):491–497.

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

[71] Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

[72] Nasdaq (2004 (accessed July 30, 2024)e). *Nasdaq announces dual listing program with six participating companies*. https://ir.nasdaq.com/news-releases/news-release-details/nasdaq-announces-dual-listing-program-six-participating.

[73] Nasdaq (2020 (accessed April 2020)). *Marketplace Rules (4000-7000)*. http://nasdaq.cchwallstreet.com/NASDAQTools/bookmark.asp?id=nasdaq-rule_4000&manual=/nasdaq/main/nasdaq-equityrules/.

[74] Nasdaq (2024 (accessed July 30, 2024)a). *American Depositary Receipt*. https://www.nasdaq.com/glossary/a/american-depositary-receipts.

[75] Nasdaq (2024 (accessed July 30, 2024)b). *American Depositary Share*. https://www.nasdaq.com/glossary/a/american-depositary-share.

[76] Nasdaq (2024 (accessed July 30, 2024)c). *Dual listing*. https://www.nasdaq.com/glossary/d/dual-listing.

[77] Nasdaq ("2024 (accessed July 30, 2024)"d). *Exchange Trading Technology*. https://www.nasdaq.com/solutions/marketplace-technology/nasdaq-trading-technology/exchange-matching#:~:text=The%20technology%20features%20among%20the,latency%20door%2Dto%2Ddoor.

[78] Nasdaq (2024 (accessed July 30, 2024)f). *Nasdaq Co-Location & Wireless Connectivity*. https://www.nasdaq.com/solutions/nasdaq-co-location.

[79] Nasdaq (2024 (accessed July 30, 2024)g). *Nasdaq Stock Screener*. https://www.nasdaq.com/market-activity/stocks/screener.

[80] Nasdaq (2024 (accessed July 30, 2024)h). *Nasdaq TotalView-ITCH FPGA*. https://www.nasdaqtrader.com/content/ProductsServices/DataProducts/TotalView/FPGAITCHFAQ.pdf.

[81] Nasdaq (2024 (accessed July 30, 2024)i). *Stock split*. https://www.nasdaq.com/glossary/s/stock-split.

[82] Nasdaq (2024 (accessed July 30, 2024)j). *U.S. Equity and Options Markets Holiday Schedule*. https://www.nasdaqtrader.com/trader.aspx?id=calendar.

[83] NasdaqTrader (2024 (accessed July 30, 2024)). *Symbol Look-Up/Directory Data Fields and Definitions*. http://www.nasdaqtrader.com/trader.aspx?id=symboldirdefs.

[84] Ogata, Y. (1981). On lewis' simulation method for point processes. *IEEE transactions on information theory*, 27(1):23–31.

[85] Pottathuparambil, R., Coyne, J., Allred, J., Lynch, W., and Natoli, V. (2011). Low-latency fpga based financial data feed handler. In *2011 IEEE 19th Annual International Symposium on Field-Programmable Custom Computing Machines*, pages 93–96. IEEE.

[86] Qian, N. (1999). On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151.

[87] Rambaldi, M., Pennesi, P., and Lillo, F. (2015). Modeling foreign exchange market activity around macroeconomic news: Hawkes-process approach. *Physical Review E*, 91(1):012819.

[88] Reuters (2022 (accessed July 30, 2024)a). *Chip designers warm to U.S. bill despite big benefits to Intel*. https://www.reuters.com/technology/us-chip-industry-split-over-chips-act-benefits-intel-sources-2022-07-18/.

[89] Reuters (2022 (accessed July 30, 2024)b). *Meta Platforms to trade under META ticker from June 9*. https://www.reuters.com/technology/meta-platforms-trade-under-meta-ticker-june-9-2022-05-31/.

[90] Reynaud-Bouret, P. and Schbath, S. (2010). Adaptive estimation for Hawkes processes; application to genome analysis. *The Annals of Statistics*, 38(5):2781–2822.

[91] Rigaki, M. and Garcia, S. (2020). A survey of privacy attacks in machine learning. *arXiv preprint arXiv:2007.07646*.

[92] Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

[93] Securities and Exchange Commision (2004 (accessed July 30, 2024)). *Notice of Filing and Immediate Effectiveness of Proposed Rule Change to Amend the Exchange's Listing Fees*. https://www.nytimes.com/2004/01/12/business/nasdaq-is-expected-to-announce-the-dual-listings-of-some-big-stocks.html.

[94] Securities and Exchange Commision (2005 (accessed April 2020)). *Regulation NMS*. https://www.sec.gov/rules/final/34-51808.pdf.

[95] Stopard, I. J., Churcher, T. S., and Lambert, B. (2021). Estimating the extrinsic incubation period of malaria using a mechanistic model of sporogony. *PLoS computational biology*, 17(2):e1008658.

[96] Tang, Q., Su, M., Jiang, L., Yang, J., and Bai, X. (2016). A scalable architecture for low-latency market-data processing on fpga. In *2016 IEEE Symposium on Computers and Communication (ISCC)*, pages 597–603. IEEE.

[97] The White House (2022 (accessed July 30, 2024)). *FACT SHEET: CHIPS and Science Act will lower costs, create jobs, strengthen supply chains, and counter China*. https://www.whitehouse.gov/briefing-room/statements-releases/2022/08/09/fact-sheet-chips-and-science-act-will-lower-costs-create-jobs-strengthen-supply-chains-and-counter-china/.

[98] Thomas Jr., L. (2004 (accessed July 30, 2024)). *Nasdaq Is Expected to Announce the Dual Listings of Some Big Stocks*. https://www.nytimes.com/2004/01/12/business/nasdaq-is-expected-to-announce-the-dual-listings-of-some-big-stocks.html.

[99] Tieleman, T. and Hinton, G. (2012). Lecture 6.5-rmsprop, coursera: Neural networks for machine learning. *University of Toronto, Technical Report*, 6.

[100] Unwin, H. J. T., Routledge, I., Flaxman, S., Rizoiu, M.-A., Lai, S., Cohen, J., Weiss, D. J., Mishra, S., and Bhatt, S. (2021a). Using Hawkes processes to model imported and local malaria cases in near-elimination settings. *PLoS computational biology*, 17(4):e1008830.

[101] Unwin, J., Routledge, I., Flaxman, S., Rizoiu, M.-A., Lai, S., Cohen, J., Weiss, D. J., Mishra, S., and Bhatt, S. (2021b). Replication data for: Using Hawkes processes to model imported and local malaria cases in near-elimination settings. *Harvard Dataverse*. `https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/YPRLIL`.

[102] Vandenberghe, L. (2010). The CVXOPT linear and quadratic cone program solvers. *Online: http://cvxopt. org/documentation/coneprog. pdf*.

[103] Vanderbauwhede, W. and Benkrid, K. (2013). *High-performance computing using FPGAs*, volume 3. Springer.

[104] Veen, A. and Schoenberg, F. P. (2008). Estimation of space–time branching process models in seismology using an EM–type algorithm. *Journal of the American Statistical Association*, 103(482):614–624.

[105] Voas, J., Kshetri, N., and DeFranco, J. F. (2021). Scarcity and global insecurity: the semiconductor shortage. *IT Professional*, 23(5):78–82.

[106] Wald, A. and Wolfowitz, J. (1940). On a test whether two samples are from the same population. *The Annals of Mathematical Statistics*, 11(2):147–162.

[107] Wang, Y., Xie, B., Du, N., and Song, L. (2016). Isotonic Hawkes processes. In *International conference on machine learning*, pages 2226–2234.

[108] Xu, H., Farajtabar, M., and Zha, H. (2016). Learning Granger causality for Hawkes processes. In *International Conference on Machine Learning*, pages 1717–1726. PMLR.

[109] Zhang, J., He, T., Sra, S., and Jadbabaie, A. (2020). Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*.

[110] Zhao, Q., Erdogdu, M. A., He, H. Y., Rajaraman, A., and Leskovec, J. (2015). Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1513–1522.

[111] Zhou, K., Zha, H., and Song, L. (2013a). Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes. In *Artificial Intelligence and Statistics*, pages 641–649. PMLR.

[112] Zhou, K., Zha, H., and Song, L. (2013b). Learning triggering kernels for multi-dimensional Hawkes processes. In *International Conference on Machine Learning*, pages 1301–1309. PMLR.

[113] Zhu, S., Wang, H., Cheng, X., and Xie, Y. (2022). Neural spectral marked point processes. In *International Conference on Learning Representations*.