Flexible estimation of temporal point processes and graphs

Déborah Sulem

St.Peter's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2022

Declaration

I hereby declare that except where specific reference is made to the work of others, the intellectual contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university.

My personal contributions are as detailed in the authorship forms at the end of each chapter. This dissertation is my own work except as specified in the text and authorship forms.

Déborah Sulem

Hillary 2023

To my dear, and greatly missed, Prilia

A ma très chère et regrettée Prilia

Acknowledgements

I would first like to deeply thank my supervisors, Judith, Mihai, Xiaowen and Vincent, for their exceptional guidance and support during this PhD. I have been greatly inspired by their research interests and scientific ethics, and I feel extremely privileged to have learnt (so much) from them. I am very grateful to them for their encouragement, advice, and patience, in particular for bearing with my organisation of Dropbox folders.

I would also like to thanks all the members of my two research groups, the *Bayesian nonparametrics* and the Networks group at the Oxford Man Institute, for providing such a friendly and exciting working environment, for their motivation to share numerous social moments, including lunches, coffee breaks, pub nights, dinners, garden parties, and online games. My PhD years definitely feel anything but solitary thanks to them. Special thanks go to Henry, an amazing collaborator and friend, Alice, for being an informal mentor and a concert mate, Matteo, for beeing such a cheerful climbing companion.

I am very grateful to the Department of Statistics which has made this PhD experience outstanding: the OxWaSP directors, Francois and Arnaud for their precious advice, my college advisors, Geoff and Cora for their personal attention, Joanna and Beverley for their incredible work as administrators, the IT team for their priceless help. I am very grateful to the EPSRC UKRI for funding my PhD, and to the ISBA and j-ISBA societies for their additional financial support and for providing such a wonderful scientific community. I also thank the European Research Council (ERC), who provided funding for an academic visit at the University of Chicago, under the European Union's Horizon 2020 research and innovation programme (grant agreement No 834175).

A huge thank to my OxWaSP and StatML family for making the CDT a very special PhD experience. The sense of togetherness has really thrived and helped me in stressful times. A big thank to the EquiStats group for sharing this incredible journey to make Statistics more diverse & fair. A special thank also to Lorenzo, Natalia, Anna, Sahra, Ruth, Romain, Kamelia, Pierre and Badr, for all the wonderful lunches and coffee breaks together.

I would like to personally thank Francesca for being first my college "mother", then my PhD sister, and eventually an unforgettable friend! An enormous thank to my college friend Mariana

for making moments together always memorable, for your generosity and originality. An other enormous thank to Juba for sharing so many Oxford & London experiences, uncountable formal dinners, concerts, garden parties, BoPs, etc. And a third enormous thanks to my friend Ariane, for going to London for the first time with me, and for bounding together during all these different lives in Paris, Oxford & London.

A huge thanks also to my housemates, Jennifer and Surrhabi, for making home the place where I longed for after work.

Finally, to my family, my mother and father, my brother Gabriel, my sister Myriam, and my grandfather, for their love, confidence, and support, and to which I owe everything. I also thank a lot my aunt Catherine and Eduard, for their encouragement and very warm welcome during my academic trips in North America.

Table of Contents

Abstract

Handling complex data types with spatial structures, temporal dependencies, or discrete values, is generally a challenge in statistics and machine learning. In the recent years, there has been an increasing need of methodological and theoretical work to analyse non-standard data types, for instance, data collected on protein structures, genes interactions, social networks or physical sensors. In this thesis, I will propose a methodology and provide theoretical guarantees for analysing two general types of discrete data emerging from interactive phenomena, namely *temporal point processes* and *graphs*.

On the one hand, temporal point processes are stochastic processes used to model event data, i.e., data that comes as discrete points in time or space where some phenomenon occurs. Some of the most successful applications of these discrete processes include online messages, financial transactions, earthquake strikes, and neuronal spikes. The popularity of these processes notably comes from their ability to model unobserved interactions and dependencies between temporally and spatially distant events. However, statistical methods for point processes generally rely on estimating a latent, unobserved, stochastic intensity process. In this context, designing flexible models and consistent estimation methods is often a challenging task.

On the other hand, graphs are structures made of nodes (or agents) and edges (or links), where an edge represents an interaction or relationship between two nodes. Graphs are ubiquitous to model real-world social, transport, and mobility networks, where edges can correspond to virtual exchanges, physical connections between places, or migrations across geographical areas. Besides, graphs are used to represent correlations and lead-lag relationships between time series, and local dependence between random objects. Graphs are typical examples of non-Euclidean data, where adequate distance measures, similarity functions, and generative models need to be formalised. In the deep learning community, graphs have become particularly popular within the field of *geometric deep learning*.

Structure and dependence can both be modelled by temporal point processes and graphs, although predominantly, the former act on the temporal domain while the latter conceptualise spatial interactions. Nonetheless, some statistical models combine graphs and point processes in order to account for both spatial and temporal dependencies. For instance, temporal point processes have been used to model the birth times of edges and nodes in temporal graphs. Moreover, some multivariate point processes models have a latent graph parameter governing the pairwise causal relationships between the components of

the process. In this thesis, I will notably study such a model, called the Hawkes model, as well as graphs evolving in time.

This thesis aims at designing inference methods that provide flexibility in the contexts of temporal point processes and graphs. This manuscript is presented in an integrated format, with four main chapters and two appendices. Chapters 2 and 3 are dedicated to the study of Bayesian nonparametric inference methods in the generalised Hawkes point process model. While Chapter 2 provides theoretical guarantees for existing methods, Chapter 3 also proposes, analyses, and evaluates a novel variational Bayes methodology. The other main chapters introduce and study model-free inference approaches for two estimation problems on graphs, namely spectral methods for the signed graph clustering problem in Chapter 4, and a deep learning algorithm for the *network change point detection* task on temporal graphs in Chapter 5.

Additionally, Chapter 1 provides an introduction and background preliminaries on point processes and graphs. Chapter 6 concludes this thesis with a summary and critical thinking on the works in this manuscript, and proposals for future research. Finally, the appendices contain two supplementary papers. The first one, in Appendix A, initiated after the COVID-19 outbreak in March 2020, is an application of a discrete-time Hawkes model to COVID-related deaths counts during the first wave of the pandemic. The second work, in Appendix B, was conducted during an internship at Amazon Research in 2021, and proposes an explainability method for anomaly detection models acting on multivariate time series.

1 | Introduction

In the first section of this chapter, I introduce the general context unifying the works comprised in this thesis. Then, in Section 1.2, I present a selective review of existing works that motivate the statistical questions listed in Section 1.3.1. Next, I describe the contributions in each of the main chapters in Section 1.3. Finally, Section 1.4 provides the background material for the subsequent chapters.

1.1 Models for discrete data: point processes and graphs

"Of quantities some are discrete, others continuous"

Aristotle, *Categories*, I.6, trans. J. L. Ackrill.

"Matter therefore is discrete being, not continuous."

G.W. Leibniz, *Sämtliche Schriften und Briefe*, 6.3, trans. A. Harmer.

Discrete data refers to a type of data that takes a countable number of values, or contains a countable number of sub-parts. Point processes (PPs) and graphs are two mathematical concepts that are used to model discrete data. A point process is a stochastic process whose realisations are random points in space, e.g., the Euclidean space $\mathbb{R}^d$, $d \in \mathbb{Z}_{\geqslant 0}$, called *spatial* PP, or in time, e.g., on the real line $\mathbb{R}$, called *temporal* PP (TPP). Point processes can therefore be applied to model *event data*, i.e., lists of events with covariates such as locations, dates, and characteristics of their occurrences. Graphs are geometric structures comprising a finite number of nodes and edges, connecting pairs or more generally, subsets of nodes. A typical example of graphs is spatial grids (or lattices), where nodes are regularly spaced and edges connect adjacent nodes.

Many statistical problems on event data are studied with temporal point processes. A TPP can be formally defined as an arrival process of events at random times $(T_1, T_2, \dots)$, $T_1 < T_2 < \dots$, $T_i \in \mathbb{R}$, $i = 1, 2, \dots$. However, it is equivalently, and more frequently, defined as a counting process of the occurrences of events $N = (N_t)_{t \in \mathbb{R}}$, where N_t represents the number of event occurrences up to time $t \in \mathbb{R}$. In practice, data sets of events often contain additional information on the events, such as their location or magnitude on a associated scale, alongside their times of occurrence. These event covariates can be integrated as *marks* associated to each event in the TPP. In particular, when the marks correspond to a finite number $K \geqslant 1$ of locations or agents, a multivariate temporal point process can be defined as $N = (N_t)_{t \in \mathbb{R}}$, where for each t,

NO	YEAR	MO	DY	HR	MN	MAG	C
1	1885	2	9	2	0	6.0	0
2	1885	6	11	9	20	6.9	0
3	1885	7	29	5	30	6.0	0
4	1885	10	30	20	30	6.2	0
5	1885	12	7	13	2	6.3	0
6	1885	12	19	18	26	6.0	2
7	1886	4	13	5	44	6.3	0
8	1886	7	2	12	33	6.3	2
9	1887	5	29	0	50	6.4	0
10	1887	5	29	1	10	6.2	2
11	1888	2	5	0	50	7.1	0
12	1888	11	24	2	3	6.5	0
13	1889	3	31	6	42	6.6	0
14	1890	11	17	9	31	6.3	0
15	1891	4	7	9	49	6.7	0
16	1891	5	5	8	16	6.2	0
17	1891	7	21	20	19	7.0	0
18	1892	10	22	19	9	6.0	0
19	1894	2	25	4	18	6.8	0

Figure 1.1: Extract of the earthquake strikes data analysed by Ogata (1988). The columns NO, YEAR, MO, DY, HR, MN MAG indicate respectively the number, year, month, hour, minute, and magnitude of the earthquakes. In column C, a 0,1, and 2 represent respectively a main shock, a foreshock and an aftershock.

$N_t = (N_t^1, \dots, N_t^K)$ and each component N_t^k counts the number of events that have occurred until time at location $k \in \{1, \dots, K\}$.

Temporal point processes have historically been applied by Hawkes (1971) and Ogata (1978) to model the occurrences of earthquakes and aftershocks. In a seminal paper, Ogata (1988) analyses the temporal and spatial patterns of seismic events in Japan during the years 1885-1980 and estimates the rate of events using a *self-exciting* or *epidemic-type* point process model. An extract from this data and the estimated temporal rate with Ogata's method are reported respectively in Figure 1.1 and Figure 1.2. In the last decade, temporal point processes have successfully been applied in financial contexts, for modelling buy and sell transactions, and on online social media, for analysing and predicting clicks and messages, two main applications that are reviewed by Hawkes (2018) and Rizoiu et al. (2017).

A natural inference problem on event data is to predict *when, where* or *how many* events will happen in the future. These questions are central in the recent biological and social applications of TPPs. For instance, TPPs are leveraged in neuroscience, for predicting neuronal spike patterns (Gerhard et al., 2017), in genomics, for inferring the locations of DNA motifs (Gusto and Schbath, 2005),

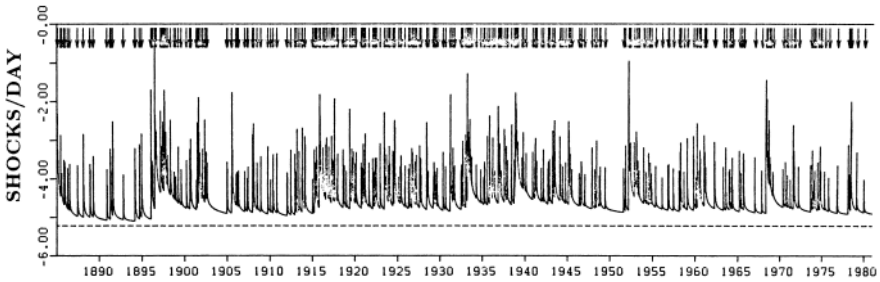


Figure 1.2: Estimated conditional intensity rate of earthquake strikes in the Tohoku area from 1885 to 1980, estimated by Ogata (1988). The rate is plotted in logarithmic scale, along time. The downwards arrows at the top of the plot indicate the occurrence times of the shocks.

in epidemiology, for anticipating the spread of diseases (Meyer et al., 2011), in psychology, for analysing human interactions (Halpin and De Boeck, 2013), and in criminology, for preventing terrorist attacks (Mohler et al., 2011). In each of these applications, experts generally believe that some dependence exists between event occurrences, and that this dependence structure is determined by the underlying phenomenon. For instance, sociologists study the mutual influence of users of online social platforms, and may want to infer if there is a causal relationship between two users' activities.

Temporal point processes models and statistical estimation methods can provide partial answers to those questions. The core object of inference in a TPP model is the probability rate of events over time, called the *conditional intensity function*. In general, the conditional intensity function is a latent, stochastic process, expressed as a function of the time *conditionally* on the history of the point process, i.e., past event occurrences. It can notably be used to determine the expected number of events happening in a future horizon, and the expected time of the next future event.

While TPPs are better suited to model temporal dependencies, relationships between agents are more directly modelled by graphs, or using the equivalent denomination, networks. Graphs are ubiquitous representations for spatially structured data and interactive phenomena, as nodes and edges can conceptualise a diversity of real-world entities and concepts. For instance, nodes are used to represent atoms, individuals, computers, cities, genes, proteins, and edges can model chemical bounds, phone calls, data exchanges, migration flows, gene co-expressions or regulatory relationships. Modelling biological, geographical, or social systems using graphs and extracting the relevant information from the structure using a statistical method allows to answer questions such as: What are the most important agents? Which dis-functioning gene should a drug target? What

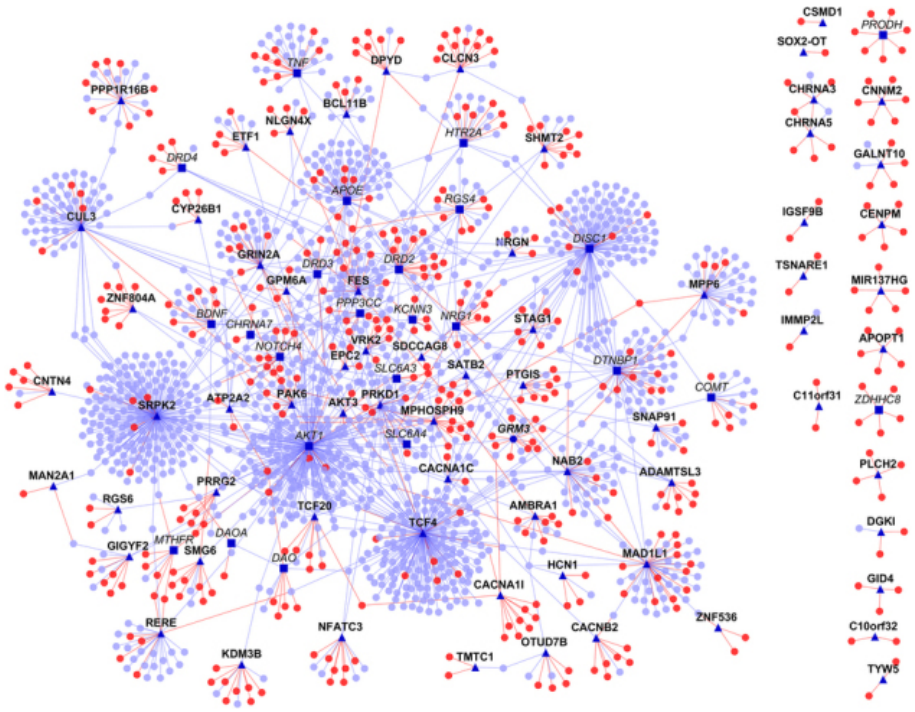


Figure 1.3: Schizophrenia interactome from Ganapathiraju et al. (2016). The interactome is a graph in which nodes are genes, and edges symbolise protein-to-protein interactions, for the proteins associated to each gene. Schizophrenia-associated genes are shown as dark blue nodes, while the newly interacting genes are red nodes. The shape of nodes indicates the source of the schizophrenia genes: triangles relates to genome-wide association studies, and squares to historic associations. The blue and edges correspond respectively to known interactions and to interactions inferred in Ganapathiraju et al. (2016).

physical properties does this molecule have?

One of the first graph-related problems pertains to road networks. It is known as Konigsberg's bridges problem and was solved by Euler (1741). Interestingly, graphs gained a major role in sociology in the 1930's for analysing social interactions, and more recently, for mining online social platforms such as Facebook, Reddit, or Youtube. Network analysis has also been successfully applied to study neurological disorders (Ganapathiraju et al., 2016), protein-to-protein interactions (Koh et al., 2012), financial networks (Ha et al., 2015), political co-voting patterns (Arinik et al., 2019), transportation systems (Sugishita and Masuda, 2021), and migration flows (Fagiolo and Mastrorillo, 2013). For instance, Figure 1.3 is a visualisation of a protein-to-protein interaction graph between Schizophrenia-associated genes, where some of the unseen or missing interactions have been inferred - a semi-supervised learning task called *link prediction*.

In fact, there is a diversity of graph models and tasks that can be performed on these structures. The

simplest graph model is the static, unweighted, and undirected graph, where the node set is $V = \{1, 2, \dots, n\}$, $n \in \mathbb{Z}_{\geqslant 0}$, and the edge set contains pairs of nodes, i.e., $E = \{e = \{u, v\};\, u, v \in V\}$. In this restrictive representation, an edge generally encodes a similarity property or a positive interaction between the nodes. This is also called the *homophily* assumption on the edges, defined for instance in Özgür Şimşek and Jensen (2008). Relaxing this assumption can be done with the *signed* graph model, originating from social psychology, notably the work of Harary (1953). In signed graphs, edges are given a sign, $+1$ or -1, which can represent either a similarity or dissimilarity measure, a friendship or enmity relationship, or a positive or negative correlation. Another important generalisation of simple graphs are *dynamic* or *temporal* networks. The latter incorporate a temporal dimension in the graph model and is therefore more suitable to time-varying systems, as noted by Skarding et al. (2021a).

In summary, temporal point processes and graphs *a-priori* model different types of discrete data with interactions, e.g., temporal dependencies between events for TPPs, and relationships between individuals in a social space. Nonetheless, in the subsequent sections, I will show how these concepts are related, and how the works in the main chapters provide complementary perspectives on the high-level problem of interaction modelling.

1.2 Literature overview

In this section, I aim at providing a selected overview of problems, methods, and theoretical results related to temporal point processes, signed graphs, and finally dynamic networks. These existing works have fostered the statistical questions and contributions of this thesis, listed in Section 1.3.1.

1.2.1 Modelling event data with temporal point processes

> "The event is always that which has just happened and that which is about to happen, but never that which is happening."
>
> Gilles Deleuze, *The Logic of Sense*.

After a presentation of the main goals of event data modelling, I will review existing works on Poisson point processes, which lay the foundations of statistical methods for the Hawkes point process model, studied in Chapters 2 and 3, and other related temporal point process models.

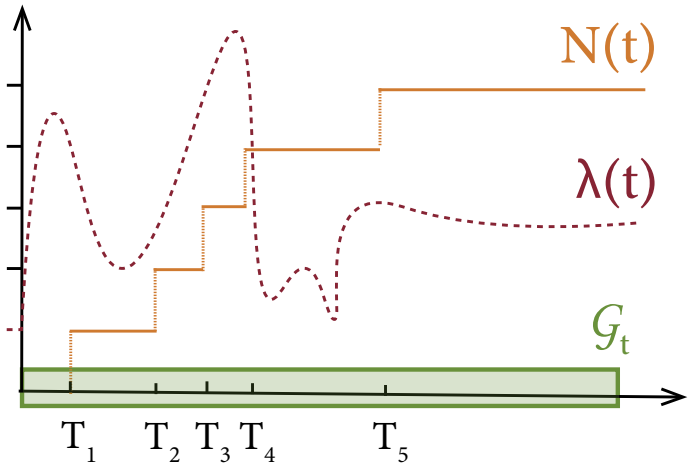


Figure 1.4: Illustration of a univariate temporal point process $N = (N(t))_t$ with a sequence of event times $(T_1, T_2, T_3, \dots)$, conditional intensity function $\lambda(t) = \lambda(t|\mathcal{G}_t)$, and history $\mathcal{G}_t = \sigma(N_s; s < t)$ for $t \in \mathbb{R}$.

Event rate prediction and interaction modelling. Temporal point processes are used to model and predict the rate of events of a phenomenon observed over time, called the conditional intensity function. The latter is in general a function of the time t and of the history of the point process $\mathcal{G}_t$, containing the information of all events before t, possibly infinitely extending backward in time. Informally, the conditional intensity function, denoted by $\lambda(t|\mathcal{G}_t)$, is the infinitesimal probability rate of event, i.e., $\lambda(t|\mathcal{G}_t)dt = \mathbb{P}\left[\text{event in } [t, t+dt]|\mathcal{G}_t\right]$. These concepts are schematically represented in Figure 1.4 for a univariate temporal point process. Additionally, Figure 1.5 illustrates how temporal point processes can be used to model neuronal spike train data.

In a TPP model, one can specify a form of the intensity function that depends on a parameter f, possibly infinite-dimensional. This is then the parameter of interest for estimating $\lambda(t|\mathcal{G}_t)$ and making predictions about future events. In addition to intensity estimation, common statistical questions on event data are related to the causality structure of the temporal process such as: Which events have been caused by an ancestor event? Which components have a causal effect on other components? (Eichler et al., 2017) Which processes are conditionally independent of each other? (Christgau et al., 2022) What are the types of interactions between a set of processes, e.g, inhibitory or excitatory? (Bonnet et al., 2022) Which events contribute to the prediction of target quantities? (Zhang et al., 2021).

In the past few decades, most TPP models have aimed at explaining the *bursting* or *clustering* behaviour of events. The latter corresponds to empirical observation that events often appear at

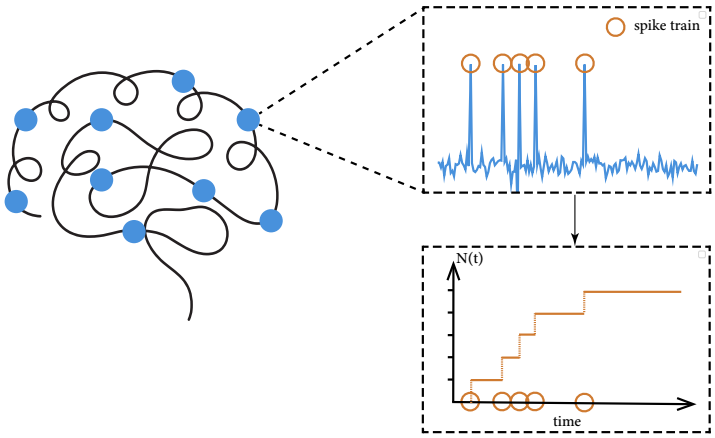


Figure 1.5: Schematic representation of spike train data modelling based with a temporal point process. Each spike train in the neuron's electrical potential time series corresponds to a biological activation, and is modelled as an event of the counting process $(N(t))_t$.

points close in time, separated by periods of inactivity where no event happens. This behaviour is frequently observed in natural phenomena such as earthquake strikes with subsequent aftershocks (Ogata, 1988), and in human communication patterns such as email exchanges (Miscouridou et al., 2018). The clustering behaviour is also sometimes called *contagion* or *excitation* effect, and is notably replicated by TPP models such as the log-Gaussian Cox process, studied for instance by Møller et al. (1998); Teh and Rao (2011), the self-exciting Hawkes process from Hawkes (1971), and the Poisson cluster process, reviewed in the book of Daley and Vere-Jones (2007).

More recently, some TPP models have focused on accounting for the *inhibition* phenomenon, which explains the fact that the probability of occurrence of new events can be decreased by previous events, or by another process. This phenomenon is of particular importance for modelling spiking neurons data, where self- and mutual- regulation mechanisms between neurons are central, as noted by Gerhard et al. (2017) and Duval et al. (2022). The nonlinear Hawkes model, considered by Chen et al. (2017); Cai et al. (2022); Deutsch and Ross (2022), the mutually-regressive point processes proposed by Apostolopoulou et al. (2019); Liu and Hauskrecht (2019), and the neural temporal point processes, introduced by Mei and Eisner (2017) and reviewed by Shchur et al. (2021), are now popular models that can account for both *excitatory* and *inhibitory* dependencies between events.

Poisson process intensity estimation. The Poisson point process is probably the most studied TPP model from the point of view of intensity estimation. In this model, the conditional intensity

is deterministic and only depends on the time variable, or equivalently, is independent of the history of the process. To estimate Poisson intensity functions, penalised maximum likelihood estimation (MLE) methods have been applied, using kernel smoothing in Bartoszynski et al. (1981); Ramlau-Hansen (1983), and reproducing kernel Hilbert spaces in Flaxman et al. (2017). In the Bayesian framework, Adams et al. (2009) proposes an approach based on Gaussian processes priors called Log-Gaussian processes. An efficient block Gibbs sampler for these processes is then introduced by Teh and Rao (2011). Several approximating schemes have also been developed, for instance a Laplace approximation in Cunningham et al. (2008), and a variational inference method in Lloyd et al. (2015).

On the theoretical side, Kirichenko and Van Zanten (2015) notably prove the optimality of Log-Gaussian processes. Additionally, Belitser et al. (2015) establish optimal contraction rates for Bayesian nonparametric smoothing methods based on spline priors. Besides, Reynaud-Bouret (2003) proposes *projection* estimators optimising a criterion called *contrast*, with optimality and adaptivity proved in Reynaud-Bouret and Rivoirard (2008). Moreover, an extension of Poisson processes, that can account for covariates dependence, are Aalen counting processes. Inference in the latter model is also performed via penalised projection estimators in Reynaud-Bouret (2006); Hansen et al. (2015), and Bayesian nonparametric methods based on Dirichlet process mixtures and log-splines priors in Donnet et al. (2017).

Inference methods for Hawkes process In contrast to the Poisson point process, the conditional intensity function of Hawkes processes is a stochastic process. For a general multivariate Hawkes process $(N_t)_t = ((N_t^k)_{k=1,\dots,K})_t$, the intensity of the k-th component is of the form

$$\lambda_k(t|\mathcal{G}_t) = \phi_k\left(\nu_k + \sum_{l=1}^{K}\int_{-\infty}^{t} h_{lk}(t-s)dN_s^l\right), \tag{1.1}$$

where ϕ_k is a *link* or *activation* function and h_{lk} is the interaction from N^l to N^k (see Section 1.4.1 for more details). *Linear* Hawkes processes correspond to the case where $\phi_k(x) = x$, $\forall x$ and $h_{lk} \geqslant 0$, $\forall l, k$. In the univariate setting, where $K = 1$ and $h_{11} = h$, the linear Hawkes model is called the *self-exciting* process (Hawkes, 1971), and is closely related to the *epidemic-type aftershock sequence* model of Ogata (1999). For the latter model, the functional parameter h, also called *triggering kernel*, is often given a parametric form, such as an exponential $h(x) = \alpha e^{-\beta t}$ or a power law $h(x) = \alpha/(t+\beta)^{\gamma}$. Early inference methods for this parametric model are based on

the MLE, for instance in the works of Hawkes (1971); Hawkes and Oakes (1974); Ogata (1978,9). The triggering kernel is also estimated via a nonparametric maximum likelihood estimator using B-spline decomposition by Gusto and Schbath (2005), and implemented by Zhou et al. (2013) in a majorisation-minimisation algorithm.

However, Veen and Schoenberg (2008) show that the MLE can only be efficiently computed for small data sets and propose a more efficient expectation maximisation (EM) algorithm where the latent variables represent the *branching*, or causality, structure of the observed events. Also for linear Hawkes processes, Da Fonseca and Zaatour (2014) leverage the analytical tractability of the process moments to design a moment-matching method, which is computationally more efficient than the MLE, although statistically not necessarily so. Penalised projection estimators are also considered by Reynaud-Bouret and Schbath (2010) and Hansen et al. (2015), who derive oracle inequalities, respectively for univariate and multivariate linear Hawkes process. Yet another estimation method proposed by Bacry and Muzy (2014) consists in finding the causal solution of a system of Wiener-Hopf equations relating the process' covariance structure and the interaction functions. Besides, Bacry and Muzy (2014) show that this method is equivalent to minimising the contrast criterion. Finally, Reynaud-Bouret and Schbath (2010) establish the minimax rate of estimation in the self-exciting Hawkes model.

Amongst Bayesian methods for Hawkes processes, Rasmussen (2013) develops a Metropolis-inside-Gibbs algorithm in a model with exponential interaction functions. Several other Monte-Carlo Markov Chain (MCMC) methods have been designed for parametric and nonparametric linear Hawkes models, for instance a method incorporating slice sampling by Blundell et al. (2012), a block Gibbs sampler in Zhang et al. (2018b), a reversible-jump MCMC in Donnet et al. (2020), and a Sequential Monte-Carlo algorithm in Linderman et al. (2017). Besides, Zhang et al. (2018b) proposed an efficient EM algorithm for computing the maximum-a-posteriori estimator. To achieve better computational efficiency, approximate Bayesian methods have been introduced, such as variational inference algorithms (Zhou et al., 2021b) and Integrated Nested Laplace Approximation techniques (Serafini et al., 2022). Finally, in the context of missing data, Deutsch and Ross (2020) propose an ABC algorithm. While the aforementioned Bayesian works are methodological, a general theoretical perspective is provided by Donnet et al. (2020), who establish general conditions for finding posterior concentration rates and study various families of prior distributions.

Nonlinear Hawkes processes correspond to the case where the links ϕ_k's are nonlinear functions.

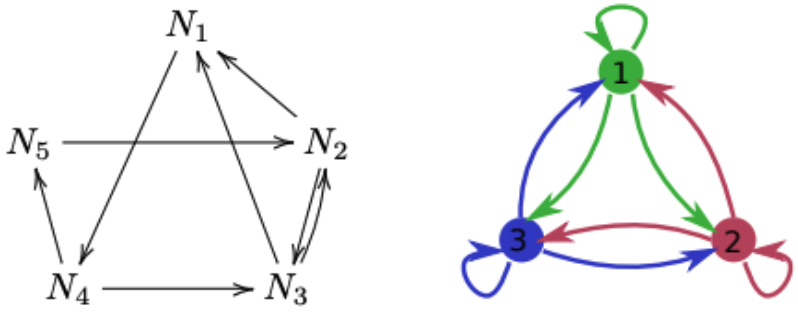
They are often used when the interaction functions h_{lk} can be negative, which is of interest in applied contexts with inhibitory interactions (Reynaud-Bouret et al., 2014). However, relatively much less work has been dedicated to these processes. The setting where the links are the ReLU functions, i.e., $\phi_k(x) = (x)_+$, is often chosen. For instance, Bonnet et al. (2022) propose an algorithm for computing the MLE, for a parametric model with exponential interaction functions. In the same model, Deutsch and Ross (2022) applies a sparsity-inducing prior in an MCMC method. Additionally, the projection estimator (Hansen et al., 2015) and a Reproducing Kernel Hilbert Space estimator (Lemonnier and Vayatis, 2014) have been empirically tested in the ReLU Hawkes model. On the theoretical side, Chen et al. (2017) obtain guarantees for smoothing kernel estimators of the cross-covariances, i.e., second-order statistics of the process that fully characterise the latter in the *linear* Hawkes model (Bacry and Muzy, 2014). Recently, Cai et al. (2022) have proved concentration inequalities for estimating the intensity via the contrast criterion. [1]

A practical difficulty of likelihood-based methods for nonlinear Hawkes processes lies in the complexity of the likelihood function (see Section 1.4.1). However, in the case of sigmoid link functions, for which $\phi_k(x) = \theta_k(1 + e^{-x})$, $\theta_k > 0$, an elegant data augmentation scheme is proposed by Adams et al. (2009). This technique allows to derive Gibbs samplers and mean-field variational inference algorithms, when using certain families of Gaussian priors, for instance in the models of Zhou et al. (2021a); Malem-Shinitski et al. (2022); Zhou et al. (2022). Finally, Wang et al. (2016) consider the problem of estimating the link function of a parametric nonlinear Hawkes model, via a piecewise-constant estimator and a moment-matching method.

Related temporal point process models. In the recent years, several TPP models related to Hawkes processes have been proposed to account for more complex interaction patterns. For instance, the *mutually-regressive* point process of Apostolopoulou et al. (2019) and the *self-limiting* Hawkes process by Olinde and Short (2020) model the inhibition phenomenon via a multiplicative term of the linear Hawkes intensity. Moreover, Liu and Hauskrecht (2019) design a *Gaussian process regressive* point process, where the dependence on the last event at each component is modelled via Gaussian processes.

Additionally, Mei and Eisner (2017) have introduced a deep learning framework for temporal point processes called *neural point process*. This type of models aims at increasing the expressive power

[1] However, the results in this paper rely on strong assumptions on the predictability properties of the involved quantities.



(a) From Etesami et al. (2016) (b) From Reynaud-Bouret et al. (2013)

Figure 1.6: Examples of connectivity graph parameter δ in two Hawkes models. The nodes of the graphs are the components of the point process N^k, $k = 1, \dots, K$, with (a) $K = 5$, and (b) $K = 3$. The arrows symbolise the directed Granger-causal links between two components. A causal link from N^l to N^k is equivalent to a non-null *interaction function* h_{lk} in the conditional intensity function (1.1). We note that the model (b) has a complete graph parameter, i.e., $\delta_{lk} = 1, \ \forall l, k$.

of TPPs by estimating the intensity function without specifying a functional form. In fact, neural point processes achieve state-of-the-art performance on tasks such as customer recommendation (Kumar et al., 2019) and clinical event prediction (Enguehard et al., 2020). Different architectures for these deep learning models have been designed, including recurrent neural networks in Du et al. (2016); Omi and Aihara (2019), generative adversarial networks in Xiao et al. (2017), and reinforcement learning techniques in Li et al. (2018). For better modelling long and short-term dependencies, Zuo et al. (2020) propose a *Transformer Hawkes* model with self-attention mechanisms. Besides, Dubey et al. (2021a) construct a *Bayesian Neural Hawkes* process to add uncertainty quantification on the predictions.

Estimating the causality structure of events. In practice, event data is also analysed from the perspective of its causality structure, a notion formally defined and studied by Didelez (2008). For instance, Gunawardana et al. (2011) capture general event dependencies in a decision tree and piecewise-constant conditional intensity model, estimated using a conjugate posterior. However, one specificity of the multivariate Hawkes model is to directly encode the causality structure in an associated parameter, denoted by $\delta = (\delta_{lk})_{l,k} \in \{0, 1\}^{K \times K}$ and called the *connectivity graph* - or *Granger-causal* graph - which characterises the local dependence structure. In the Hawkes model, the graph δ is a redundant parameter, defined for each ordered pair (l, k) as $\delta_{lk} = 1$ if $h_{lk} \neq 0$ and $\delta_{lk} = 0$ otherwise, with h_{lk} an interaction function defined in the conditional intensity function (1.1). Therefore, nodes in this graph represent the components of the process, and a directed edge

from a “source" component to a “target" component is equivalent to the “target" being locally dependent on the “source". Figure 1.6 show two examples of connectivity graph in the Hawkes model.

In high-dimensional Hawkes processes, i.e., when the number of components K is large, a standard approach consists in estimating a *sparse* connectivity graph. This task is notably applied to estimate the functional connectivity of neurons in Lambert et al. (2018), and disease spread networks in Xu et al. (2016). Several sparse estimation methods for the Hawkes connectivity graph have been proposed, including a likelihood ratio testing procedure (Kim et al., 2011), penalised projection estimators (Eichler et al., 2017; Hansen et al., 2015; Lambert et al., 2018; Cai et al., 2022), ℓ_1-regularised MLE (Xu et al., 2016; Zhou et al., 2013), an ℓ_0-penalised minorisation-maximisation algorithm (Idé et al., 2021), and a thresholding procedure on the cross-covariance estimator (Chen et al., 2017).

Interestingly, Achab et al. (2017) show that in the linear Hawkes model, the first and second integrated cumulants of the process are not sufficient to characterise the L_1-norms of the interaction functions, and therefore the Granger-causality structure. Consequently, Achab et al. (2017) propose a consistent least-square estimator based on the first three integrated cumulants, and related to the Generalised Method of Moments by Hall (2005). In the nonlinear Hawkes model, Cai et al. (2022) prove that the penalised projection estimator is consistent on the connectivity graph, however under strong predictability assumptions on the process. Besides, Bayesian estimation of the connectivity graph parameter has only been empirically tested in Donnet et al. (2020).

1.2.2 Modelling positive and negative relationships using signed graphs

“Social networks are an inevitable part of modern life."

Shahriari and Jalili (2014)

In this section, I first introduce the origins of signed graphs in statistical analysis, then describe a set of problems that have been studied on this extended graph model. Next, I focus on the signed graph clustering task and provide an overview of the state-of-the-art literature.

Signed graphs for modelling human relationships, and more. In social network analysis, an interaction or relationship between two individuals can often have a positive or a negative connotation, e.g., friendly or antagonist messages, agreement or disagreement, collaboration or

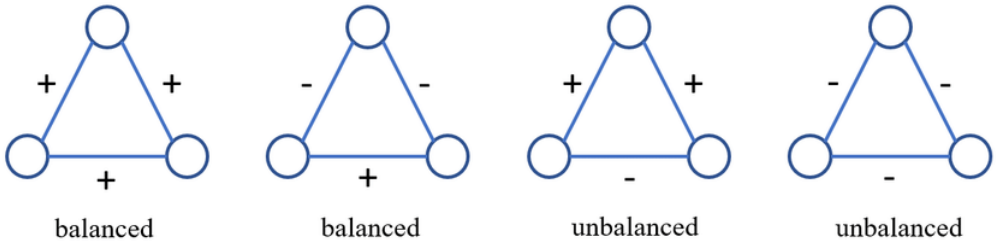


Figure 1.7: Balanced and unbalanced triads, or *triangles*, in the social balance theory of Cartwright and Harary (1956).

conflict, similarity or dissimilarity, and trust or distrust (Easley and Kleinberg, 2010). This binary alternative appears in human sentiments (Harary, 1953), international relations (Moore, 1978), online ratings (Kunegis et al., 2009), Parliament co-voting patterns (Cucuringu, 2015), and can be modelled using a sign graph, where every edge is given a sign, i.e., a label $+1$ or -1.

Seminal work using signed graphs emerges in social psychology, on the basis of social balance theory (Heider, 1958; Cartwright and Harary, 1956). In this paradigm, human relationship patterns follow the rules “a friend of my friend is my friend" and “an enemy of my enemy is my friend", which lead to the notion of *structural balance*. In signed graph theory, a cycle of edges is balanced if the product of the signs of its edges is positive, or equivalently, if it contains an even number of negative edges. For triangles (or triads), the four possible combinations of balanced and unbalanced triangles are listed in Figure 1.7. The balance theorem by Harary (1953) states that if a signed graph is balanced, the node set can be divided into two subsets such as the positive edges only link nodes within each subset, and the negative edges link nodes in different subsets.

To model real-world social networks, Davis (1967) introduced a notion of *weak balance*, which allows triads with all negative edges. Additionally, different measures of partial balance have been proposed to analyse signed networks and reviewed by Aref and Wilson (2017). One such measure is the frustration index, defined by Harary (1953) as the minimum number of edges that need to be deleted to make a signed graph balanced. Computing this index is an NP-hard problem, but several approximate algorithms have been proposed, for instance by Hüffner et al. (2007). Alternatively, the walk-based spectral measure of balance suggested by Estrada and Benzi (2014) can be computed via the eigendecomposition of the signed adjacency matrix.

Other applications of signed graphs are gene regulatory networks with activatory and inhibitory relationships (Karaaslanli et al., 2022), brain networks (Rubinov and Sporns, 2011), and more

broadly, time series correlation networks (Costantini and Perugini, 2014). Correlation networks are built from pairwise correlation coefficients between time series, and are often analysed in the form of a complete, weighted, and signed graph. They are ubiquitous in time series analysis, e.g., in finance (Pavlidis et al., 2006), genomics (Fujita et al., 2012), neuroscience (Smith et al., 2011; Saberi et al., 2021), and climate science (Hlinka et al., 2017).

Most statistical methods on graphs are not designed to handle edges with negative signs and straightforwardly adaptable. In fact, a signed graph $G = (V, E)$ can be defined as the union of two unsigned graphs with the same node set V. The *positive* graph, denoted G^+, contains the positive edges $E^+ = \{e_+ = \{u, v\};\ u, v \in V, \{u, v, +1\} \in E\}$ and the *negative* graph, denoted G^-, has the edge set $E^- = \{e_- = \{u, v\};\ u, v \in V, \{u, v, -1\} \in E\}$. However, there is no *a-priori* systematic rationale on how these two unsigned graphs should be treated in an inference method for G. For instance, Wang et al. (2022) propose a regularised likelihood-based method for signed graph clustering giving different weights to the positive and negative graphs.

In the last decade, inference methods on signed graphs have been introduced to address the problems of signed graph clustering (Kunegis et al., 2010), link (sign) prediction (Leskovec et al., 2010; Kumar et al., 2016; Chiang et al., 2011), node ranking (Shahriari and Jalili, 2014), synchronisation over the group $\mathbb{Z}_2$ (Cucuringu, 2015), graph anomaly detection (Kumar et al., 2014), and data mining (Tang et al., 2016). Another application of signed graphs is proposed by Goldberg et al. (2007) in a semi-supervised node classification task, which consists in labelling every node of a graph using a few pre-labelled nodes. In this context, adding negative edges in the graph between the labelled nodes that belong to different classes leads to better classification performance.

Signed graph clustering Graph clustering or *community detection* is probably the most studied unsupervised learning problem on signed graphs, for Tomasso et al. (2022) also provides a recent and extensive review. This task corresponds to finding a partition of the node set such that there are as many as possible positive edges between nodes in the same subset, called *cluster* or *community*, and negative edges between nodes in different subsets. A related problem is *polarisation discovery*, studied for instance by Bonchi et al. (2019) and Tzeng et al. (2020), where the goal is to uncover two or multiple pairs of polarised communities, linked by mostly negative edges (Bonchi et al., 2019; Tzeng et al., 2020). Signed graph clustering techniques are applied for instance by Karataş and Şahin (2018) to identify criminal or terrorist groups on online social networks and detect fraud

events on telecommunication, and by Bansal et al. (2004) and Aghabozorgi et al. (2015) to cluster time series via correlation networks

For solving the signed graph clustering problem, Yang et al. (2007) and Gómez et al. (2009) propose modularity-based methods, a type of algorithms that maximise a measure of cluster quality in order to find the nodes' partition. Spectral methods have also been designed, for instance by Chiang et al. (2012), Gallier (2016) and Mercado et al. (2019). These algorithms are derived from an eigenvalue problem, and apply a clustering algorithm such as k-means (Hartigan and Wong, 1979) on the spectral node embeddings. In Gallier (2016), the node embeddings correspond to the eigenvectors of the signed Laplacian, while in Mercado et al. (2019), the latter is replaced by the Power Mean Laplacian. Besides, He et al. (2018) and Huang et al. (2019) develop deep learning methods based on graph neural networks to cluster signed graphs.

Similarly to the unsigned graph setting, theoretical guarantees for signed graph clustering algorithms can be obtained in suitable *signed stochastic block models*, as is performed for instance in Cucuringu et al. (2019); Mercado et al. (2019); Wang et al. (2022). These random graph models for signed graphs with an underlying community structure are related to extensions of the stochastic block model (SBM) where the edges are given labels, such as the censored and labeled stochastic block models (see for instance the review of Abbe et al. (2014) for an overview of the SBM). However, existing results on the performance of spectral methods often hold in restricted settings. For instance, the properties of the signed Laplacian are studied with $k = 2$ equal-size communities by Cucuringu et al. (2019), and for general k by Mercado et al. (2019), but in a particular model with possibly edges with both positive and negative signs. Recently, Wang et al. (2022) also obtains an information-theoretic limit of the signed clustering problem.

Finally, another difficulty when solving a task on graphs is the *network sparsity*, a property of most social networks (Tomasso et al., 2022). Informally, a sparse graph is a graph that has very few edges, compared to the number of possible edges. The graph sparsity is often quantified by the average density of edges, defined as $p = \frac{2|E|}{|V|(|V|-1)} \in [0, 1]$. In the asymptotic limit of large graphs (i.e., $n = |V| \gg 1$), the number of edges grows as a function of the number of nodes. The *dense* graph regime corresponds to $p \gtrsim \frac{\log n}{n}$, while in the *sparse* regime, $p = O(\frac{1}{n})$. The intermediate regime $\frac{1}{n} \lesssim p \lesssim \frac{\log n}{n}$ is often called the *relatively dense* regime. In both the unsigned and signed graph setting, the edge sparsity is known to decrease the performance of graph clustering algorithms, in particular spectral methods (Tomasso et al., 2022). In unsigned graphs, this phenomenon has been

explained by Dall'Amico et al. (2021), who study the effect of the power-law degree distribution, and addressed in spectral methods by adding a regularisation step, for instance in Amini et al. (2013); Joseph and Yu (2016); Le et al. (2015). Nonetheless, there is not yet a methodology for signed graph clustering algorithms.

1.2.3 Modelling time-varying structures with dynamic networks

> "To exist is to change, to change is to mature, to mature is to go on creating oneself endlessly."
>
> Henri Bergson, *Creative Evolution*, 1907.

In this section, I aim at providing a summary of dynamic network models and inference approaches. Then, I introduce the change point detection task for dynamic networks and review existing methods.

Models of graphs with a temporal dimension. Temporal, dynamic, time-varying, evolutionary, or evolving networks refer to network models that incorporate a time covariate in the nodes, edges, or their attributes. Real-world data modelled by graphs often comes from evolving systems, for instance online communication platforms (Kumar et al., 2019), co-voting networks (Wilson et al., 2019), fMRI data (Cribben and Yu, 2017), or results from interactions of short duration such as phone calls (Holme and Saramaki, 2012). In consequence, the ability to analyse and control a real-world network can be improved by dynamic network models (Li et al., 2017).

In a dynamic network, nodes and edges appear, and disappear, over time. This temporal process is often described as a sequence of graph *snapshots* at discrete timestamps, i.e., a temporally ordered sequence of static graphs. However, it can also be described as a continuous process, where each node and edge is given a birth and death times (see Section 1.4.4 for more details). Many learning tasks on static graphs have their equivalent dynamic formulation, for instance, dynamic node classification (Pei et al., 2016), link prediction (Rossi et al., 2020), and graph clustering (Rossetti and Cazabet, 2018). Moreover, dynamic network models are used by Holme and Saramaki (2012) to describe the formation of a static graph and a disease contagion process in an interaction structure. Besides, temporal graph models can be used to detect network events such as distribution change points in Yu et al. (2021).

First approaches for dynamic network models aimed at explaining the change in global networks properties, such as the degree distribution and the network diameter. For instance, Leskovec et al. (2005) propose an epidemic-type attachment model accounting for the densification power law and

diameter shrinking phenomena. Then, a variety of models have been introduced to account for more diverse edge and time dependencies, such as temporal exponential random graphs (Robins et al., 2007), dynamic reference models (Holme et al., 2004), dynamic compartmental models (Newman, 2002) and TPP based models such as multivariate Cox processes (Perry and Wolfe, 2013; Vu et al., 2011), Aalen processes (Hunter et al., 2011), and Hawkes processes (Fox et al., 2016; Sanna Passino and Heard, 2022).

Yet one of the most popular model is the dynamic latent space model. The latter is related to the dynamic stochastic block model, which accounts for community structures in temporal graphs, for instance in Kim et al. (2018). In a dynamic latent space model, each node is associated to a trajectory in a latent space, and the edges are conditionally independent given these latent variables. This type of model is applied to study conference papers co-authorship networks (Sarkar and Moore, 2006), bill co-sponsorship networks in the US Congress (Sewell and Chen, 2015), and bilateral trade data (Ward et al., 2013). For these models, maximum likelihood estimation is applied by Sarkar and Moore (2006), while Sewell and Chen (2015) and Yang et al. (2011) propose Monte-Carlo Markov Chain methods. A nonparametric estimation method via Gaussian processes is also designed in Durante and Dunson (2014). Additionally, several variational EM algorithms have been developed, for instance by Yang et al. (2011), Matias and Miele (2017), and Ho et al. (2011).

Relatively few work has been done on designing model-free algorithms, in spite of their interest in some real-world applications. For the dynamic network clustering task, spectral methods have been developed by Pensky and Zhang (2019) and Keriven and Vaiter (2022). In addition, deep learning algorithms have been proposed, such as neural spatio-temporal point processes (Trivedi et al., 2019; Jin et al., 2019), dynamic and spatio-temporal graph neural networks (Skarding et al., 2021b), and dynamic network representation learning (Kazemi et al., 2020). Most of these neural networks architectures combine static graph neural networks architectures with a temporal mechanism, such as recurrent layers (Seo et al., 2018; Pareja et al., 2020) and attention mechanisms (Sankar et al., 2020). Moreover, these data-driven methods often achieve state-of-the-art performance in the dynamic network clustering and link prediction tasks, as is shown for instance by Rossi et al. (2020).

Network change point detection Dynamic networks are sometimes used to model non-stationary real-world processes which dynamics undergo abrupt switches or breaks. For instance, in social

network analysis, interaction patterns can be modified after a "shock" (Rossi et al., 2020). One task of interest in several applications consists in detecting such structural breaks (or *change points*), for instance to segment brain connectivity fMRI data (Ondrus et al., 2021), and to discover *phases* in financial correlation networks (Barnett and Onnela, 2016) or transport networks (Yu et al., 2021). For multivariate time series, detecting change points is a task that has been widely studied for several decades, while for dynamic networks, the equivalent task, often termed *network change point detection*, has recently become more popular. In this task, a statistical method is defined as *offline* if the detection and localisation of change points is performed retrospectively, and *online*, if it is concomitant with the observation time of the network. Figure 1.8 illustrates the change point detection problem in a discrete-time dynamic network, comprising a sequence of *graph snapshots* with a single change point.

Most existing methods for this task rely on some model assumptions. For instance, Peel and Clauset (2015) first estimate the time-varying parameters of a generalised hierarchical random graph, then apply a standard time series change point detection method on the inferred sequence. Penalised maximum likelihood approaches have also been considered in a non-homogeneous Poisson point process model by Corneli et al. (2018), and in dynamic stochastic block models by Wilson et al. (2019) and Bhattacharjee et al. (2020). A model-free method is applied by Miller and Mokryn (2020), who monitor the temporal sequence of the degree distribution of the graph. A similar approach is conducted by Wang et al. (2017) on the edge distribution. Other model-free algorithms leverage pairwise graph comparison to detect change points, for instance a graph kernel in Gretton et al. (2006), a graph similarity function in Koutra et al. (2016), and a graph distance in Hewapathirana et al. (2020).

Yet another popular type of methods is based on the network cumulative sums (CUSUM) statistic, an extensively applied statistic in change point detection problems for sequential data. In the context of temporal sequences of graphs, CUSUM statistics are often based on weighted averages of the adjacency matrices. For instance, Wang et al. (2021) design an offline CUSUM-based algorithm, with consistency and optimality guarantees under a dynamic inhomogeneous Bernoulli graph model. This method and its analysis have been extended to the online inference setting by Yu et al. (2021). Besides, Padilla et al. (2019) propose a refined CUSUM algorithm for sequences of dependent graphs. Moreover, Dubey et al. (2021b) and Enikeeva and Klopp (2021) concomitantly propose CUSUM methods for dynamic networks with missing data. Nonetheless, since existing CUSUM

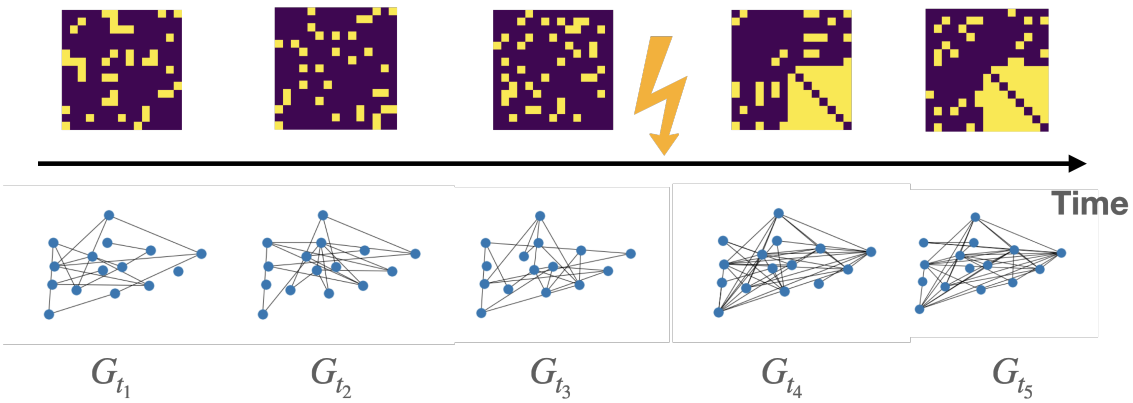


Figure 1.8: Discrete-time dynamic network with five graph snapshots $(G_{t_i})_{i=1,\dots,5}$ and one distribution *change point*. The binary matrices above the time axis correspond to the adjacency matrices of each graph (a yellow and purple square represent respectively a 1 and 0 entry), and the orange flash symbolises an event, causing an abrupt change. The latter affects the structure of the last two snapshots, which visibly contain a denser subset of nodes.

methods only use the adjacency matrices of dynamic networks, they are limited to unattributed networks, i.e., networks without additional attributes such as node or edge features.

1.3 Objective and contributions of this thesis

This thesis is motivated by the expressive power and complementarity of temporal point processes and graphs for modelling interactive phenomena. In this section, I first summarise some motivating questions, then present a comprehensive overview of results, methods, and outputs, provided in each of the main chapters.

1.3.1 Motivating questions and outline

As described in Section 1.2.1, the nonlinear multivariate Hawkes model is a flexible temporal point process model that extends the original linear, or *self-exciting*, model of Hawkes (1971). This generalised model is popular amongst practitioners, thanks to its ability to account for *excitating* and *inhibiting* interactions between entities, in particular biological neurons (Gerhard et al., 2017). Moreover, this model enjoys a nice causality interpretation through its associated *connectivity graph* parameter. However, in comparison to the linear Hawkes model, the nonlinear model has been much less studied.

Firstly, it is not yet known how Bayesian nonparametric methods theoretically and empirically

perform in this model. One preliminary question is to determine which assumptions on the link functions and the parameter are required to define an identifiable model. Then, one would wonder what general conditions on the prior distribution and the model are needed to guarantee consistency and concentration of the posterior distribution. Additionally, if the nonlinear link functions also include a parameter, a natural problem is to jointly infer the latter with the original parameter of the Hawkes model.

Secondly, the consistency of Bayesian methods on the connectivity graph has not yet been studied, even in the linear model. Nonetheless, there is empirical evidence in Donnet et al. (2020) that the posterior distribution is consistent. Thirdly, the computational complexity of Bayesian nonparametric methods in the Hawkes model is a main challenge, limiting its extensive use in practice. Therefore, a widely open problem is the design of efficient algorithms in the nonlinear setting, such as approximate Bayesian methods. It would therefore be interesting to first, analyse the variational Bayes algorithms of Zhou et al. (2021a) and Malem-Shinitski et al. (2022), which apply in restricted settings, then, to extend the theory to general approximate Bayes methods.

In Section 1.2.2, I presented the signed graph clustering problem, its numerous applications, and the practical advantages of spectral algorithms. These model-free approaches essentially perform an eigenvector computation, which can scale up to large and moderately sparse graphs. The preliminary theoretical analysis by Cucuringu et al. (2019) of the signed Laplacian and the SPONGE spectral algorithms, in a simple *signed stochastic block model*, suggests that these results could be extended in several ways.

First, one natural problem is to analyse the *normalised* versions of these algorithms, namely the symmetric signed Laplacian and SPONGEsym algorithms, which empirically perform better than the un-normalised versions. Secondly, it would be of interest to study these methods in a general signed stochastic block model, with arbitrary number of clusters, and general cluster sizes. Another question is to adapt these algorithms for the sparse graph setting, for which their empirical performance is likely to decrease without an appropriate modification. Some regularisation strategies for sparse unsigned graphs could then be tested, and similarly analysed.

Finally, previous work cited in Section 1.2.3 shows that the temporal graph setting can better model real-world evolving networks, and the latter sometimes undergo abrupt distribution changes. The detection and localisation of these changes are important in several applications, such as brain

connectivity state segmentation. However, there are relatively few approaches to solve this task when there is no prior knowledge on the network generative distribution, and the type of change points that may occur. Moreover, most existing approaches cannot accommodate for network covariates, such as node and edge attributes.

Therefore, one problem is to design more flexible methods for detecting change points in dynamic networks. For other graph estimation problems, deep learning algorithms can generally learn from covariates, and achieve state-of-the-art performance, without relying on any model assumption. It would thus be interesting to leverage a learning procedure, such as a graph neural network, for solving the network change point detection task. In this line of research, some central questions would lie around the choice of architecture, training and validation procedures, and the computational efficiency, relatively to standard model-based methods. Moreover, it would be interesting to compare the performance of learning-based algorithms and model-based methods in settings where the model assumptions are verified, and where the networks are attributed.

The subsequent main chapters aim at tackling these relevant questions, in four independent works. Chapters 2 and 3 address open questions in the nonlinear Hawkes model and Bayesian nonparametric framework. Chapter 4 provides insights on spectral methods for the signed clustering problem. Finally, Chapter 5 proposes a deep-learning method for solving the network change point detection task. Chapter 4 has been published in the Journal of Machine Learning Research, Chapters 2 and 5 have been submitted respectively to the Bernoulli journal and to the Machine Learning journal. Chapter 3 has not yet been submitted. The general context of this manuscript is illustrated in Figure 1.9. In the next sections, I summarise the contributions in each chapter.

1.3.2 On Bayesian nonparametric estimation for nonlinear Hawkes processes

Chapter 2 is joint work with J. Rousseau and V. Rivoirard. In this paper, we analyse the asymptotic properties of Bayesian nonparametric methods in the nonlinear multivariate Hawkes model. We consider K-dimensional Hawkes processes $N = (N_t)_t$ with conditional intensity function (1.1), and general nonlinear link functions $(\phi_k)_{k=1,\dots,K}$. We aim at estimating the parameter $f = (\nu, h, \delta)$ of the process, containing the background rates $\nu = (\nu_k)_{k=1,\dots,K}$, the K^2 interaction functions $h = (h_{lk})_{l,k=1,\dots,K}$, and the connectivity graph $\delta = (\delta_{lk})_{l,k=1,\dots,K}$ where for each l, k, $\delta_{lk} = 0 \iff h_{lk} = 0$.

We first assume that the link functions $(\phi_k)_{k=1,\dots,K}$ are known. Then, we also consider a shifted

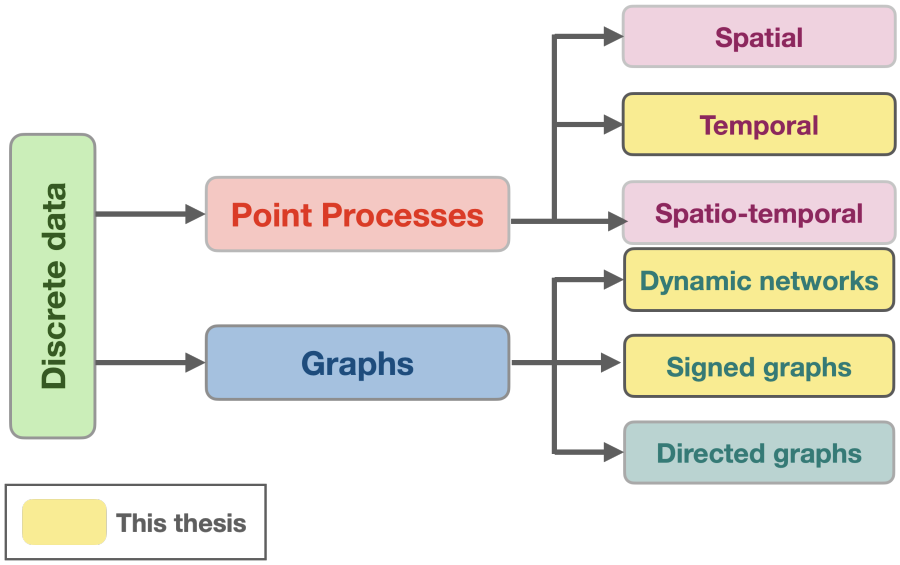


Figure 1.9: Description of the context and content of this thesis.

ReLU model where $\phi_k(x) = \theta_k + (x)_+, \theta_k \geqslant 0, k \in [K]$, with an unknown baseline rate parameter $\theta = (\theta_k)_{k=1,\dots,K}$, which we also aim to estimate. This model includes the ReLU function, brings more flexibility to the latter, and leads to a positive intensity whenever $\theta > 0$. The shifted ReLU model is also motivated by the unboundedness of the likelihood ratio when using the ReLU function, and it can be seen as an alternative to the exponential model by Gerhard et al. (2017).

In the Bayesian framework, we consider a generic nonparametric prior distribution, denoted Π, on the parameter space, and, given an observation N of the process on a window $[0, T]$, study the concentration rates of the posterior distribution $\Pi(.|N)$, when $T \to \infty$. More precisely, we look for the smallest possible sequence $\epsilon_T = o(1)$, and general but easy-to-verify conditions on the prior and the model, such that we can prove that

$$\mathbb{E}_{f_0}[\Pi(d(f, f_0) < \epsilon_T | N)] \xrightarrow[T\to\infty]{} 1, \tag{1.2}$$

where f_0 denotes the true parameter of the observed process, $d(.,.)$ is a loss function, and $\mathbb{E}_{f_0}$ is the expectation under the true distribution of the process. The latter result is formally described in Theorem 3.2. In the shifted ReLU model, the posterior concentration rate on f and the baseline parameter θ (Proposition 3.5) can be similarly written as

$$\mathbb{E}_{f_0}[\Pi(d(f, f_0) + \tilde{d}(\theta, \theta_0) < \epsilon_T | N)] \xrightarrow[T\to\infty]{} 1,$$

where $\tilde{d}(.,.)$ is another loss function. From our results, we also deduce the convergence rate of the posterior mean estimator $\hat{f} = \mathbb{E}^{\Pi}[f|N]$. We verify our main assumptions on several commonly used nonlinear models, and three nonparametric prior families, namely splines, random histograms and mixtures of Beta distributions, for which we also provide explicit concentration rates for Hölder classes of functions,

Additionally, we establish posterior consistency results on the connectivity graph parameter δ (Theorem 3.9), i.e., we prove that under certain assumption on the prior, it holds that

$$\mathbb{E}_{f_0}[\Pi(\delta = \delta_0|N)] \xrightarrow[T\to\infty]{} 1,$$

where δ_0 denotes the true connectivity graph parameter. We note that the previous consistency result is not directly implied by the concentration (1.2) when there exist (l, k) such that $h^0_{lk} = 0$. Moreover, we propose a risk-minimising estimator of the graph based on adequate loss functions. The use of such estimator decouples the design of the prior from the inference on the graph parameter, and therefore leads to a consistent method under weaker conditions on the prior distribution (Theorem 3.11).

Preliminary to our main results, we provide identifiability conditions for nonlinear Hawkes models (Proposition 2.3 and 2.5). For the posterior concentration rates, the argument relies on the general theory of Ghosal and van der Vaart (2007), similarly to the work of Donnet et al. (2020) in the linear Hawkes model. This theory consists in proving a prior condition, a testing condition, and a Kullback-Leibler condition, and the core of our work lies in obtaining each piece.

We prove those general conditions with a novel technique, based on the regenerative properties of Hawkes processes and the concept of *excursion*. Extending and leveraging the probabilistic results of Costa et al. (2020), we design specific tests and control the log-likelihood ratio using a decomposition of the observations into independent and identically distributed sub-parts. This elegant tool allows us to derive the posterior concentration rates, under mild assumptions on the prior distribution and the model.

1.3.3 On scalable variational Bayes methods for Hawkes processes

Chapter 3 is joint work with J. Rousseau and V. Rivoirard. In this work, we unify, extend, and analyse variational Bayes inference methods in the general Hawkes model. This type of methods

aim at approximating the posterior distribution $\Pi(.|N)$ by a *variational* posterior distribution, denoted $\hat{Q}$, belonging to a class of "convenient" distributions, called the *variational* class and denoted $\mathcal{V}$. More precisely, the approximate posterior distribution is defined as

$$\hat{Q} := \arg\min_{Q \in \mathcal{V}} KL\left(Q || \Pi(.|N)\right),$$

where $KL(.||.)$ denotes the Kullback-Leibler divergence.

We first provide general theoretical guarantees on the variational posterior distribution. We notably study the concentration rates in Theorem 3.2, equivalently to (1.2) for the posterior distribution, under general conditions on the prior distribution, the model, and the variational class. For this, we leverage the results of Chapter 2 and the general theory of Zhang and Gao (2020). We apply our main result to the mean-field variational class and a newly introduced spike-and-slab class. The latter is of interest to infer a sparse graph parameter in high-dimensional Hawkes processes. We also apply our results to the random histogram and Gaussian process prior families.

We also propose a novel adaptive and sparsity-inducing method, related to the general model-selection frameworks of Zhang and Gao (2020) and Ohn and Lin (2021), to select the dimensionality of the estimation problem. This approach enjoys provable guarantees and allows us to design efficient algorithms in the sigmoid Hawkes model, with link functions $\phi_k(x) = \theta_k(1 + e^{-x})^{-1}$ with $\theta_k > 0$, $k \in [K]$. Building on existing data augmentation strategies and using Gaussian priors, we construct two *adaptive* mean-field variational algorithms. Our most efficient algorithm (Algorithm 3) first selects the graph parameter, which allows to reduce the computational cost for high-dimensional processes. We empirically demonstrate the effectivess of our approach in an extensive set of simulations. In particular, we show that our variational algorithms are faster than MCMC methods and can correctly infer the graph parameter. Moreover, Algorithm 3 can scale up to large K.

1.3.4 On spectral clustering algorithms and regularisation in signed graphs

Chapter 4 is joint work with M. Cucuringu, A. Singh, and H. Tyagi. This paper contains an analysis of four spectral clustering algorithms for signed graphs. Two of these algorithms come from previous works; the first one is the SPONGE$_{sym}$ method introduced by Cucuringu et al. (2019), and the second one is the symmetric signed Laplacian algorithm, partly analysed in the works of

Kunegis et al. (2010); Gallier (2016); Mercado et al. (2019). The two other ones are regularised algorithms newly designed for the sparse graph regime, and derive from the two first methods and the regularisation scheme of (Amini et al., 2013) for unsigned graphs.

For these four algorithms, we bound their misclustering rates (Theorems 7 and 16). More precisely, we prove that with probability going to 1, and in the large graph limit $n \to \infty$, for a controlled approximate solution of the spectral clustering algorithm, the fraction of misclustered nodes is small. To obtain these guarantees, we study the spectral properties of the SPONGE$_{sym}$ operator and the symmetric signed Laplacian matrix, denoted $\overline{L_{sym}}$, in a suitably defined signed stochastic block model (SSBM), with a general number of clusters and non-equal cluster sizes.

For instance, we first prove that, with high probability, the eigenspace of $\overline{L_{sym}}$, derived from a randomly sampled graph, is close to the corresponding Laplacian matrix in the *expected* graph, denoted $\mathcal{L}_{sym}$, under the SSBM. For this, we apply a general technique that upper bounds the error $\left\|\overline{L_{sym}} - \mathcal{L}_{sym}\right\|$, then lower bounds the *eigengap* of $\mathcal{L}_{sym}$, and applies the Davis-Kahan theorem (Davis and Kahan, 1970). Moreover, we add a perturbation argument to analyse clusters with non-equal-sizes. Finally, the control on the misclustering error is derived using the now standard tool of Lei and Rinaldo (2015).

For the SPONGE$_{sym}$ and symmetric signed Laplacian algorithms, we obtain our results in the relatively dense regime, i.e., when the edge density of the SSBM is such that $p \gtrsim \frac{\log n}{n}$. In the sparse regime $p = O\left(\frac{1}{n}\right)$, we propose to regularise these methods using the technique of Amini et al. (2013). In the unsigned graph setting, the latter method consists in adding a positive weight τ/n, $\tau > 0$ to every entry of the adjacency matrix A. We therefore introduce two regularisation parameters $\gamma^+, \gamma^- > 0$, and define a regularised signed adjacency matrix as $A_\gamma = A + (\gamma^+ - \gamma^-)/n \mathbb{1}\mathbb{1}^T$. We then apply our two first spectral algorithms in this regularised signed graph.

One advantage of such regularisation technique is to solve the problem of non concentration of the adjacency matrix in the sparse regime, noted for instance by Le et al. (2015) under the (unsigned) stochastic block model. Using their techniques, which rely on a decomposition of the graph into a *core* component and a remaining subset of nodes, we prove that our regularised signed Laplacian and SPONGEsym operators concentrate, when the regularisation parameters are adequately chosen, i.e., when $\gamma_+ + \gamma_-$ scale respectively as $(np)^{7/8}$ and $(np)^{6/7}$ (Theorems 6 and 11).

Finally, we provide an extensive numerical analysis of these four algorithms, on simulated data from the SSBM with a small and large numbers of clusters, and three real-world social network data sets. In comparison to existing spectral clustering methods, we demonstrate the superior performance of the SPONGE$_{sym}$ algorithm, followed by the symmetric signed Laplacian, in the dense regime. In the sparse regime, we study the variation of performance with respect to the choice of regularisation parameters, and show that the regularised methods outperform non-regularised ones on real data.

1.3.5 On the change point detection task in dynamic networks

Chapter 5 is joint work with H. Kenlay, M. Cucuringu, and X. Dong. In this paper, we propose a novel learning-based method for detecting change points in discrete-time dynamic networks. Our method is based on a data-driven graph similarity function learnt by a graph deep learning algorithm. The similarity function allows to quantify the discrepancy between the current graph snapshot and the past ones, in an online setting of a streaming network. Our approach is change point agnostic, does not require any tuning of the detection threshold, and is suitable for different types of dynamic networks, in particular those with node attributes.

Our graph similarity learning method is based on a novel siamese graph neural network (GNN) architecture. The latter is quite shallow and includes a generic siamese graph encoder followed by a parsimonious similarity module. In particular, we leverage Sort-k pooling, introduced by Zhang et al. (2018a), to detect both *local* and *global* displacements in the graph structure or attributes. Moreover, for unattributed dynamic networks with a constant node set, we propose to use *identity* positional encodings to increase the expressive power of the graph encoder.

One difficulty in our deep learning approach to the network change point detection problem is to design the training procedure. To the best of our knowledge, this problem has not yet been addressed in previous work. We propose to sample pairs of graphs in the dynamic network training sequence in order to train our graph similarity learning algorithm like a binary classifier task. Nonetheless, after the training procedure, our method is very fast at test time and avoids any detection delay.

We evaluate and compare our method on simulated data from dynamic stochastic block models (DSBM), similarly to the works of Wilson et al. (2019); Wang et al. (2013); Padilla et al. (2019); Yu et al. (2021), and two real-world dynamic correlation networks. In DSBMs, we show the superiority of our method over non-deep learning approaches in several change point scenarios. In real-world correlation networks from physical sensor data, our method shows better performance at detecting

unseen types of change points and extrapolating to unseen networks, in comparison to baseline methods. On the correlation network of S&P stock returns, we qualitatively show that our detected change points better correlate with major financial events.

Remark 1.3.1. *The two works in the appendices were not included as main chapters of this manuscript, although they are connected to the aforementioned works, for the following reasons. I contributed as a second author to the first one, in Appendix A, which contains a case-study of discrete-time Hawkes processes in COVID-19 data. It is now part of the thesis of Dr. Raiha Browning. The second work in Appendix B, is more loosely connected to discrete data and interaction modelling, since it focuses on the anomaly detection problem in time series.*

1.4 Background

In this section, I formally define Hawkes processes, signed graphs and dynamic networks. I also introduce spectral clustering and graph convolutional networks.

Notations For a function $h : \mathbb{R} \to \mathbb{R}$, I denote $\|h\|_1 = \int_{\mathbb{R}} |h(x)|dx$ and $\|h\|_2 = \sqrt{\int_{\mathbb{R}} h^2(x)dx}$ its L_1 and L_2 norms, and $h^+(x) = \max(x, 0) = (x)_+$ and $h^-(x) = \max(-x, 0)$ its positive and negative parts. For an integer $K \geqslant 1$, I denote the set $[K] = \{1, \dots, K\}$. For a matrix $M \in \mathbb{R}^{m\times n}$, $\|M\|$ denoted its spectral norm, i.e., its largest singular value, and $\|M\|_F$ denotes its Frobenius norm. I also denote $tr(M) = \sum_i M_{ii}$ the trace of M, M_i its i-th row, and M^j its j-th column. I use the notation $\mathbb{1} = (1, \dots, 1)$ for the all ones column vector and I for the identity matrix, which sizes depend on the context. For a temporal point process N and for any $a, b \in \mathbb{R}, a < b$, $N[a, b]$ denotes the number of points in $[a, b]$.

1.4.1 Hawkes processes

In this section, I define linear and nonlinear Hawkes processes, first, in the univariate setting, then, in the multivariate setting. I consider a probability space $(\Omega, \mathcal{G}, \mathbb{P})$, and recall that for a K-dimensional temporal point process $N = (N_t)_{t\in\mathbb{R}}$, $N_t = (N_t^1, \dots, N_t^K)$ counts the number of events at each component until time t, for each $t \in \mathbb{R}$. Let $\{\mathcal{G}_t\}_{t\in\mathbb{R}}$ with $\mathcal{G}_t = \sigma(N_s, s \leqslant t) \subset \mathcal{G}$, be the filtration or *history* of the point process. The conditional intensity function $(\lambda_t)_{t\in\mathbb{R}}$ associated

to N is a $\mathcal{G}_t$-predictable and K-dimensional process verifying

$$\mathbb{E}[N[s,t]|\mathcal{G}_s] = \mathbb{E}\left[\int_s^t \lambda_u du \Big| \mathcal{G}_s\right], \quad \forall s,t \in \mathbb{R},\ s < t. \tag{1.3}$$

Moreover, if the intensity function depends on a parameter f and is denoted $(\lambda_t(f))_{t\in\mathbb{R}}$, the log-likelihood function of the point process N observed over a window $[0,T], T > 0$ is defined as

$$L_T(f) := \sum_{k=1}^{K} \left[\int_0^T \log(\lambda_t^k(f)) dN_t^k - \int_0^T \lambda_t^k(f) dt\right].$$

I now define the self-exciting temporal point process, which corresponds to the linear univariate Hawkes model introduced by Hawkes (1971). In the following definition, $K = 1$.

Definition 1.4.1 (Self-exciting process (Hawkes, 1971))**.** *A simple point process $N = (N_t)_{t\in\mathbb{R}}$ with history $\{\mathcal{G}_t\}_{t\in\mathbb{R}}$ is a self-exciting linear Hawkes process if its conditional intensity function $(\lambda_t)_{t\in\mathbb{R}}$ can be written as*

$$\lambda_t = \nu + \int_{-\infty}^{t-} h(t-s) dN_s, \quad \forall t \in \mathbb{R}, \tag{1.4}$$

where $\nu \in R^+\backslash\{0\}$ and $h : \mathbb{R}^+ \to \mathbb{R}^+$ is a non-negative function.

Note that integrals with respect to the point process measure N such as in (1.4) can also be written as sums over the times of events. Let $(T_1, T_2, \dots)$ be the times of events in N. Then, for any function $g : \mathbb{R}^+ \to \mathbb{R}$, it holds that

$$\int_{-\infty}^{t-} g(t-s) dN_s = \sum_{T_i<t} g(t-T_i).$$

Definition 1.4.1 introduces the two parameters of the self-exciting Hawkes process: the *spontaneous* or *background* rate $\nu > 0$, and the *self-exciting* function $h : \mathbb{R}^+ \to \mathbb{R}^+$, also called *triggering kernel*. The background rate models exogenous effects while the self-exciting function accounts for the endogenous dependence on past events. Due to the latter temporal dependence, the self-exciting process is in general non-Markovian - unless h is of exponential form, i.e., $h(x) = \alpha e^{-\beta t}$, $\alpha, \beta > 0$.

Since the self-exciting function h is non-negative, each event in the history $\mathcal{G}_t$ contributes as a non-negative term in the intensity λ_t; this reproduces the so-called *self-exciting* property or *self-*

excitation phenomenon. Therefore, the inter-event times of a Hawkes process are dependent, and the sequences of events display a *bursting* or *clustering* behaviour. Moreover, this process is linear in the sense that its conditional intensity function λ_t is linear in the parameter $f = (\nu, h)$.

For TPPs, the stationarity property corresponds to the intensity function being a stationary process, i.e., $\mathbb{E}[\lambda_t] = \lambda_0,\ \lambda_0 \in \mathbb{R}^+$, or equivalently, for any $s < t$, $\mathbb{E}[N[s,t]]$ only depends on $(t-s)$. For the self-exciting process, and also more general Hawkes processes, the existence of a stationary version is equivalent to the non-explosiveness property, i.e., $\mathbb{E}[N[s,t]] < +\infty$, for any $s < t$. Moreover, the dynamics of a Hawkes process are said to be stable with respect to an initial condition $\mathcal{G}_0 \subset \mathcal{G}$ if the process with history $\mathcal{G}_0$ at $t = 0$ converges in distribution towards the stationary version when $t \to \infty$ (Brémaud and Massoulié, 1996).

In the self-exciting model, there exists a stationary version of the point process if $\|h\|_1 < 1$. If $\|h\|_1 > 1$, the process is explosive, and, in the critical case $\|h\|_1 = 1$, stationarity conditions have been studied by Brémaud and Massoulié (2001). When the dynamics of the process are stable, an alternative definition of Hawkes processes consists in constructing the process as the solution of a system of stochastic equations.

Definition 1.4.2. *Let Q be a $\mathcal{G}_t$-Poisson point process on $(0,+\infty)\times(0,+\infty)$ with unit intensity and let N^0 be an initial condition, i.e., a point process measure on $(-\infty, 0]$. Assume that $h : \mathbb{R}^+ \to \mathbb{R}^+$ is such that $\|h\|_1 < 1$ and let*

$$\begin{cases} N = N_0 + \int_{(0,+\infty)\times(0,+\infty)} \delta(u) \mathbb{1}_{\theta \leqslant \lambda_u} Q(du, d\theta) \\ \lambda_u = \nu + \int_{-\infty}^{u-} h(u-s) dN_s,\ u > 0 \end{cases}, \tag{1.5}$$

where $\delta(.)$ is the Dirac function and $\nu > 0$. Then the unique strong solution of (1.5) is a self-exciting Hawkes process with background rate ν and self-exciting function h.

A straightforward extension of the self-exciting Hawkes process is a multivariate process, called the *mutually-exciting* process, or multivariate linear Hawkes model.

Definition 1.4.3 (Mutually-exciting Hawkes process)**.** *Let $K \in \mathbb{N}\backslash\{0\}$. A multivariate point process $N = (N_t)_t = (N_t^1, \dots, N_t^K)_{t\in\mathbb{R}}$ is a linear Hawkes process if for any $l, k \in [K]$, $(N_t^k)_t$ and $(N_t^l)_t$ cannot have common points and the conditional intensity function $(\lambda_t)_{t\in\mathbb{R}} =$*

$(\lambda_t^1, \dots, \lambda_t^K)_t$ *can be written as*

$$\lambda_t^k = \nu_k + \sum_{l=1}^{K} \int_{-\infty}^{t-} h_{lk}(t-s) dN_s^l, \quad t \in \mathbb{R}, \quad k \in [K],$$

where for each $l, k \in [K]$, $\nu_k > 0$, *and* $h_{lk} : \mathbb{R}^+ \to \mathbb{R}^+$.

In Definition 1.4.3, the functions $(h_{lk})_{l,k}$ are generally called *interaction functions*. These functions model the dependence on past events at each component, and also define a Granger-causality structure between the components of the process. The definition of this notion, or rather its negative, is formulated by Eichler et al. (2017).

Definition 1.4.4 (Granger non-causality)**.** *Let* N *be a stationary multivariate point process with filtration* $\{\mathcal{G}_t\}_t$ *and for a component* i, *let* $\{\mathcal{G}_t^{-i}\}_t$, $\mathcal{G}_t^{-i} := \sigma\left(N_s^j, s < t, j \in [K] \backslash \{i\}\right)$ *be the sub-filtration excluding the* i*-th component. Then the* i*-th component does not Granger-cause the* j*-th component with respect to* $\{\mathcal{G}_t\}_t$ *if the intensity function* λ_t^j *is* $\mathcal{G}_t^{-i}$*-measurable.*

In multivariate Hawkes processes, the previous definition translates into a nullity condition on the interaction functions. A component $(N_t^i)_t$ does not Granger-cause $(N_t^j)_t$ with respect to $\{\mathcal{G}_t\}_t$ if and only if $h_{ij} = 0$ (Proposition 3.2 in Eichler et al. (2017)). Therefore, a Hawkes process is associated to a Granger-causality or *connectivity* graph parameter $\delta = (\delta_{lk})_{l,k=1,\dots,K} \in \{0,1\}^{K \times K}$ where for each $l, k \in [K]$, $\delta_{lk} = 0 \iff h_{lk} = 0$. The intensity function can also be re-written as

$$\lambda_t^k = \nu_k + \sum_{l=1}^{K} \delta_{lk} \int_{-\infty}^{t-} h_{lk}(t-s) dN_s^l, \quad t \in \mathbb{R}, \quad k \in [K],$$

and the parameter f of the process is defined either as $f = (\nu, h)$ where $\nu = (\nu_k)_{k \in [K]}$ and $h = (h_{lk})_{l,k \in [K]}$ or as $f = (\nu, h, \delta)$.

The linear Hawkes model can be generalised to a nonlinear model, where in particular, the non-negativity condition on the interaction functions can be relaxed.

Definition 1.4.5. *(Nonlinear Hawkes process) Let* $K \in \mathbb{N} \backslash \{0\}$. *A multivariate point process* $N = (N_t)_t = (N_t^1, \dots, N_t^K)_{t \in \mathbb{R}}$ *is a Hawkes process if its conditional intensity function* $(\lambda_t)_{t \in \mathbb{R}} =$

$(\lambda_t^1, \dots, \lambda_t^K)_t$ *can be written as*

$$\lambda_t^k = \phi_k \left(\nu_k + \sum_{l=1}^{K} \int_{-\infty}^{t-} h_{lk}(t-s) dN_s^l \right), \quad t \in \mathbb{R}, \quad k \in [K],$$

where for each $k \in [K]$, $\phi_k : \mathbb{R} \to \mathbb{R}^+$, $\nu_k > 0$, *and for each* $l \in [K]$, $h_{lk} : \mathbb{R}^+ \to \mathbb{R}$.

In Definition 1.4.5, the functions $(\phi_k)_{k\in[K]}$ are called the *link* or *activation* functions. Although they do not have to, the link functions are often chosen to be monotone non-decreasing and continuous. In this definition, the functions $(h_{lk})_{l.k}$ can have a non-null negative part $(h_{lk}^-)_{l,k}$, therefore past events can contribute as *negative* terms in the intensity function at t. This characterises the so-called *mutual-inhibition* phenomenon.

In practice, the nonlinear functions $(\phi_k)_k$ are often parametrised in the forms

$$\begin{cases} \phi_k(x) = \theta_k + \psi(x) \\ \text{or} \quad \phi_k(x) = \theta_k \psi(x) \end{cases}, \quad \theta_k \geqslant 0, \quad k \in [K],$$

where $\psi : \mathbb{R} \to \mathbb{R}^+$ is a nonlinear function such as the ReLU function $\psi(x) = \max(x, 0) = (x)_+$ (Hansen et al., 2015; Costa et al., 2020; Deutsch and Ross, 2022) , the sigmoid function $\psi(x) = (1 + e^{-x})^{-1}$ (Zhou et al., 2020,0; Malem-Shinitski et al., 2022), the softplus function $\psi(x) = \log(1+e^x)$ (Mei and Eisner, 2017), or a clipped exponential function $\psi(x) = \min(e^x, \Lambda), \Lambda > 0$ (Gerhard et al., 2017; Carstensen et al., 2010). The additional parameter $\theta = (\theta_k)_{k\in[K]}$ corresponds to a baseline event rate in the additive model, for instance in the shifted ReLU studied in Chapter 2 where $\phi_k(x) = \theta_k + (x)_+, k \in [K]$, and is typically small. In the sigmoid model, where $\phi_k(x) = \theta_k(1 + e^x)^{-1}, k \in [K]$, θ corresponds to the upper limit of the intensity and can potentially be large.

The stability properties of nonlinear Hawkes processes have been studied by Brémaud and Massoulié (1996) under different sets of assumptions on the link functions and the parameter. Moreover, central limit theorems in nonlinear and linear Hawkes processes are derived in Zhu (2013); Bacry et al. (2013). Interestingly, *renewal* or *regenerative* properties in Hawkes processes have been studied by Costa et al. (2020); Graham (2021); Raad (2019). Gaussian and Poisson approximations of functionals of Hawkes processes are also shown by Torrisi (2016,0), while Berry-Essen type bounds have recently been obtained by Hillairet et al. (2021). Mean-field limits in the high-dimensional

setting $K \gg 1$ of Hawkes processes and age-dependent extensions have been studied by Delattre et al. (2014); Chevallier (2017); Raad et al. (2020); Pfaffelhuber et al. (2022); Erny et al. (2022).

1.4.2 Signed graphs, Laplacians and cut functions

In this section, I define signed graphs in the context of simple, unweighted, undirected and unattributed graphs, i.e., graphs with simple edges and without nodes and edges covariates. Then I present common Laplacian operators for signed graphs and their relations to cut functions.

I define a signed graph as $G = (V, E)$, where $V = \{1, \dots, n\}$ is the node set and E is the edge set where each edge $e \in E$ is a triplet $\{u, v, s\}$ with $s \in \{-1, 1\}$ and $u, v \in V$. A signed graph can be decomposed into two unsigned subgraphs, the positive subgraph $G^+ = (V, E^+)$ and the negative subgraph $G^- = (V, E^-)$), where $E^+ = \{e = \{u, v\};\ e \cap \{+1\} \in E\}$ (resp. E^-) is the set of positive (resp. negative) edges. Note that for a simple signed graph, the two subgraphs G^+ and G^- have disjoint edge support, i.e., $E^+ \cap E^- = \emptyset$, and $E^+ \cup E^- = E$. Let m be the size of E, i.e., the number of edges in the graph.

The signed adjacency matrix of G is the matrix $A \in \{0, 1, -1\}$ such that for any $i, j \in [n]$,

$$A_{ij} = \begin{cases} 1 & \text{if } \{i, j, +1\} \in E \\ -1 & \text{if } \{i, j, -1\} \in E \\ 0 & \text{otherwise .} \end{cases}$$

It can also be defined as $A := A^+ - A^-$, where $A^\pm \in \{0, \pm 1\}^{n \times n}$ are the adjacency matrices of the positive and negative subgraphs. The signed adjacency is a symmetric matrix, therefore all its eigenvalues are real numbers. Some open problems and conjectures on the spectral property of the signed adjacency matrix have been listed by Belardo et al. (2019). Moreover, the signed degree matrix is the diagonal matrix $\overline{D} \in \mathbb{N}^{n \times n}$ such that $D_{ii} = \sum_{j=1}^{n} |A_{ij}|$ and $D_{ij} = 0$ if $j \neq i$ for any $i, j \in [n]$. It is also equal to $\overline{D} = D^+ + D^-$, where $D^\pm = Diag(A^\pm \mathbb{1}) \in \mathbb{N}^{n \times n}$.

I also define the notions of connectedness and balance in signed graphs. An example of balanced signed graphs is drawn in Figure 1.10.

Definition 1.4.6. *Let G be a signed graph with positive and negative subgraphs $G^+ = (V, E^+)$ and $G^- = (V, E^-)$.*

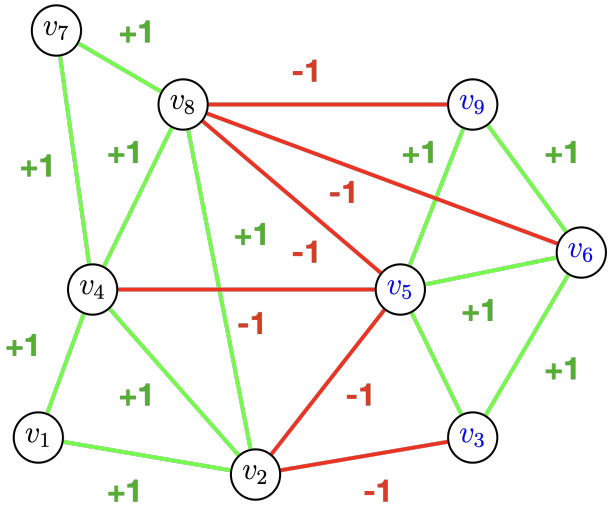


Figure 1.10: Example of balanced signed graph from Gallier (2016).

- *G is said to be connected if for any pair of nodes $\{u, v\} \in V^2$, there exists a path from u to v, i.e., there exists a sequence of nodes $(v_0, v_1, \ldots, v_p)$, $p \in \mathbb{N}$, such that $v_0 = u$, $v_p = v$ and $\{v_i, v_{i+1}\} \in E$, $i = 0, \ldots, p-1$.*
- *G is said to be balanced if there exists is a partition of V into two blocks V_1 and V_2 such that all the positive edges connect pairs of nodes $\{u, v\}$ with $u, v \in V_1$ or $u, v \in V_2$, and all the negative edges connect pair of nodes $\{u, v\}$ with $u \in V_1$ and $v \in V_2$. If G is connected, it is balanced if and only if every cycle (i.e., a path from u to u for $u \in V$) contains an even number of negative edges.*

Graph operators (or matrices) for signed graphs include different variants of graph Laplacians. The unnormalised (or combinatorial) signed Laplacian is defined by Kunegis et al. (2010) as

$$\overline{L} = \overline{D} - A.$$

Two normalized versions of the signed Laplacian exist, the random-walk signed Laplacian $\overline{L_{rw}} = I - \overline{D}^{-1}A$, and the symmetric signed Laplacian $\overline{L_{sym}} = I - \overline{D}^{-1/2}A\overline{D}^{-1/2}$. Note that these normalised Laplacians are well-defined for signed graphs without isolated vertices, i.e., for which $\bar{D}_{ii} > 0, \forall i \in [n]$. The signed Laplacians are symmetric positive semi-definite matrices, as shown in Proposition 5.1 and Proposition 5.2 of Gallier (2016), therefore their eigenvalues are real and non-negative. From Theorem 5.6 of the same reference, for a connected signed graph, the signed

Laplacian is positive definite (therefore its eigenvalues are positive) if and only if it is not balanced.

Graph Laplacians are related to cut functions. Before defining the latter, I define the volume of a subset $C \subset V$ in a signed graph G as

$$\text{Vol}_G(C) := \sum_{i \in C, j \in V} |A_{ij}| = \sum_{i \in C} \bar{D}_{ii}.$$

The volume of C in the positive and negative subgraphs are similarly defined as $\text{Vol}_{G^\pm}(C) := \sum_{i \in C, j \in V} A^\pm_{ij} = \sum_{i \in U} D^\pm_{ii}$. The (signed) cut function between C and its complement $\overline{C} = V \setminus C$ is then defined as

$$\text{Cut}_G(C, \overline{C}) := \sum_{i \in C, j \in \bar{C}} |A_{ij}|.$$

Similarly, in the positive and negative subgraphs, the cut function is defined as $\text{Cut}_{G^\pm}(C, \overline{C}) := \sum_{i \in C, j \in \bar{C}} A^\pm_{ij}$, and therefore,

$$\text{Cut}_G(C, \overline{C}) = \text{Cut}_{G^+}(C, \overline{C}) + \text{Cut}_{G^-}(C, \overline{C}).$$

Let also $links^\pm(C, B) := \sum_{i \in C, j \in B} A^\pm_{ij}$, for any subsets $B, C \subset V$. The connection between quadratic forms of the unnormalised signed Laplacian and the signed cut function is formalised in the following proposition from Gallier (2016).

Proposition 1.4.7 (Proposition 5.3 in Gallier (2016))**.** *Let $C_1, \ldots, C_K$ be a partition of V into $K \geqslant 1$ clusters, i.e., a set of disjoint subsets of V such that $\cup_{j=1}^K C_j = V$. For any $j \in [K]$, let $X^j \in \mathbb{R}^n$ be a vector representing C_j such that, with $x_j > 0$,*

$$X^j_i = \begin{cases} x_j & \text{if } i \in C_j, \\ 0 & \text{if } i \notin C_j, \end{cases}, \quad \forall i \in [n].$$

Then,

$$(X^j)^T \bar{L} X^j = x_j^2 (Cut_G(C_j, \overline{C}_j) + 2links^-(C_j, C_j)). \tag{1.6}$$

Moreover, there is a connection between quadratic forms of the signed degree matrix and the

volume function. For a vector X^j as defined in Proposition 1.4.7, it holds that

$$(X^j)^T \bar{D} X^j = x_j^2 \text{Vol}_G(C_j).$$

Therefore, one can link the signed Laplacian and degree matrix to a normalised signed cut function.

Definition 1.4.8 (Definition 5.2 in Gallier (2016))**.** *Let $C_1, \ldots, C_K$ be a partition of V into $K \geqslant 1$ clusters. The signed normalised cut is defined as*

$$sNcut(C_1, \ldots, C_K) = \sum_{j=1}^{K} \frac{Cut_G(C_j, \overline{C}_j)}{Vol_G(C_j)} + 2 \sum_{j=1}^{K} \frac{links^-(C_j, C_j)}{Vol_G(C_j)}.$$

Therefore, for a matrix $X \in \mathbb{R}\backslash\{0\}^{n \times K}$ where each column X^j represents a cluster C_j, it also holds that

$$sNcut(C_1, \ldots, C_K) = \sum_{j=1}^{K} \frac{(X^j)^T \bar{L} X^j}{(X^j)^T \bar{D} X^j}.$$

Signed cut functions are also related to the level of unbalancedness of the graph and the spectrum of graph Laplacians. The following result provides bounds on the smallest eigenvalue of the unnormalised signed Laplacian.

Proposition 1.4.9. *Proposition 3.3 and Theorem 3.4 in Hou (2005) Let G with a signed graph and λ_1 the smallest eigenvalue of its signed Laplacian. Then it holds that*

$$\Delta - \sqrt{\Delta - \omega(G)^2} \leqslant \lambda_1 \leqslant 4\omega(G),$$

where $\Delta = \max_{i \in [n]} \bar{D}_{ii}$ is the largest degree in G, and

$$\omega(G) = \min_{C \subset V, C \neq \emptyset} \frac{e_{min}(C) + Cut_G(C, \overline{C})}{|C|},$$

with $e_{min}(C)$ the minimum number of edges that need to be removed from the subgraph induced by C to make it balanced.

From Proposition 1.4.9, if a graph G is balanced, then $\omega(G) = \lambda_1 = 0$ and its signed Laplacian is singular. In the context of clustering and link prediction tasks, additional signed Laplacian-type

operators have been defined. The *unsigned* Laplacian $L_u = \bar{D} - |A|$ and the *physics* Laplacian $L_p = D - A$ with $D = Diag(A\mathbb{1})$ have been considered by Knyazev (2018). Chiang et al. (2012) define the Balance Ratio Cut and the Balanced Normalized Cut operators as respectively $\overline{L_{BRC}} = D^+ - A$ and $\overline{L_{BNC}} = \bar{D}^{-1/2}(D^+ - A)\bar{D}^{-1/2}$. The former can be related to cut functions: for any partition $C_1, \dots, C_K$ of V with representation matrix X, and any $j \in [K]$, it holds that

$$(X^j)^T L_{BRC} X^j = x_j^2(Cut_{G^+}(C_j, \overline{C}_j) + links^-(C_j, C_j)). \tag{1.7}$$

Note that in comparison to (1.6), the first term on the RHS of (1.7) is the cut function over the positive subgraph. Mercado et al. (2019) introduce a family of signed graph operators called *Matrix Power Mean of Laplacians* defined as $L_p = \frac{1}{2^{1/p}}((L^+_{sym})^p + (L^-_{sym})^p)^{\frac{1}{p}}, p \in \mathbb{R}$, where $L^\pm_{sym} = I - (D^\pm)^{-1/2} A^\pm (D^\pm)^{-1/2}$ are the symmetric Laplacians of the positive and negative subgraphs. Concomitantly, Cucuringu et al. (2019) propose the SPONGE and SPONGE$_{sym}$ operators, defined as

$$T = (L^- + \tau^+ I)^{-1/2}(L^+ + \tau^- I)(L^- + \tau^+ I)^{-1/2},$$
$$T_{sym} = (L^-_{sym} + \tau^+ I)^{-1/2}(L^+_{sym} + \tau^- I)(L^-_{sym} + \tau^+ I)^{-1/2},$$

where $\tau^\pm > 0$ are *trade-off* parameters and $L^\pm = D^\pm - A^\pm$ are the graph Laplacians of the positive and negative subgraphs. The definition of these operators also originate from an appropriately defined signed cut objective in Cucuringu et al. (2019).

1.4.3 Spectral graph clustering

Spectral graph clustering methods are algorithms for partitioning the node set into a predefined number of clusters $K \geqslant 1$. They have a common pipeline where the main step is to solve an eigenvalue, or generalised eigenvalue, problem. The latter is generally the main computational bottleneck, however there are now efficient solvers such as the Locally Optimal Block Preconditioned Conjugate Gradient method (Knyazev, 2001). The graph operator in the eigenvalue problem comes from the relaxation of an optimisation problem based on cut functions (see Section 1.4.2). Before I describe this derivation, I recall the steps of spectral clustering algorithms. Given a graph G and a choice of operator $O(G)$, e.g., the graph Laplacian, a spectral algorithm has the three following steps:

1. Compute the operator $O(G)$.
2. Compute its K extremal eigenvectors $U_1, \dots, U_K$, i.e., the eigenvectors associated to the K smallest (or largest) eigenvalues of $O(G)$.
3. Stack the eigenvectors colomn-wise into a matrix $U = [U_1, \dots, U_K] \in \mathbb{R}^{n \times K}$ and cluster its rows using the k-means or k-means++ algorithm (Vassilvitskii and Arthur, 2006). The latter gives the final partition of V.

I now illustrate the derivation of the graph operator for a signed graph G from an objective function, using one possible formulation of the signed clustering problem. To find clusters in G, one possibility is to search a partition of the node set such that most of the positive edges connect nodes in the same cluster, or equivalently, few positive edges connect nodes in different clusters. Intuitively, this formulation of the signed clustering problem implies that the clusters should contain nodes that are similar to each other, or have positive relationships, e.g., friendly interactions in a social network. In this case, one aims at minimising the cut function in the positive subgraph, i.e., solving

$$\min_{C_1,\dots,C_K} \text{Cut}_{G^+}(C_i, \overline{C_i}),$$

where the minimum is computed over the partitions $C_1, \dots, C_K$ of V. However, the solution of the latter objective often contains one “giant" cluster spanning most of the nodes, while the other clusters are “small". In practice, this is often not a useful partition. Therefore, it is common to modify the objective function to take into account the size of the clusters, i.e., to solve

$$\min_{C_1,\dots,C_k} \sum_{i=1}^{K} \frac{\text{Cut}_{G^+}(C_i, \overline{C_i})}{|C_i|},$$

which corresponds to a *ratio cut objective*. Moreover, noting that for a vector X^i representing the cluster $i \in [K]$, as defined in Section 1.4.2, it holds that

$$x_i^2 \text{Cut}_{G^+}(C_i, \overline{C_i}) = (X^i)^T L^+ X^i,$$

where $L^+ = D^+ - A^+$, therefore, the previous objective can be also written as

$$\min_X \sum_{i=1}^{K} \frac{(X^i)^T L^+ X^i}{(X^i)^T X^i}, \tag{1.8}$$

where X is the matrix representing the partition $C_1, \ldots, C_K$. Note that the minimum in (1.8) is computed over the following set of matrices:

$$\mathcal{X} = \left\{ X = [X^1, \ldots, X^K];\ X^i \in \{0,1\}^n,\ X^i \neq 0,\ (X^i)^T (X^j) = 0,\ \forall i, j \in [K] \right\}.$$

Moreover, (1.8) is equivalent to the following optimisation problem

$$\min_{X \in \mathcal{X}, X^T X = I} tr(X^T L^+ X).$$

Given the form of $\mathcal{X}$, (1.8) is a NP-hard problem, therefore a common relaxation consists in replacing the constraint $X \in \mathcal{X}$ by $X \in \mathbb{R}^{n \times K}$. Then, the previous trace minimisation problem is equivalent to an eigenvalue problem and the solution is given by $X^* = [U_1, \ldots, U_K]$, where $U_1, \ldots, U_K$ are the eigenvectors associated to the K smallest eigenvalues of L^+.

A variant of this approach consists in normalise the cut function by the volume of the subsets C_i's instead of their size, and thus to solve the alternative problem

$$\min_{C_1,\ldots,C_k} \sum_{i=1}^{K} \frac{\text{Cut}_{G^+}(C_i, \overline{C_i})}{\text{Vol}_{G^+}(C_i)} = \min_{C_1,\ldots,C_K} \sum_{i=1}^{K} \frac{(X^i)^\top L^+ X^i}{(X^i)^\top D^+ X^i},$$

which corresponds to a *normalised cut objective*. If for each $i \in [K]$, $X^i = \frac{1}{\sqrt{\text{Vol}_{G^+}(C_i)}} (D^+)^{1/2} \mathbb{1}_{C_i}$, then the latter problem is equivalent to

$$\min_{C_1,\ldots,C_K} \sum_{i=1}^{K} (X^i)^\top L^+_{sym} X^i. \tag{1.9}$$

Relaxing the discreteness constraint on X in (1.9) also leads to an eigenvalue problem, now formulated as

$$\min_{X^\top X = I} tr\left(X^\top L^+_{sym} X \right).$$

Normalised signed Laplacians spectral clustering algorithms generally have better prediction accuracy, as empirically shown for instance by Chiang et al. (2012); Cucuringu et al. (2019);

Mercado et al. (2019). In unsigned graphs, Sarkar and Bickel (2015) theoretically prove that this type of normalisation decreases the “spread" of clusters in stochastic block models.

Finally, because of their connection with eigenvalue problems, spectral signed clustering algorithms can be theoretically analysed under suitably defined random graph models such as signed stochastic block models in Mercado et al. (2019); Cucuringu et al. (2019). In unsigned graphs, upper bounds on the mis-clustering rates are obtained after an analysis of the spectral properties, i.e., the eigenvalues and eigenvectors, of the Laplacian matrix in Rohe et al. (2011).

1.4.4 Dynamic networks

In this section, I define dynamic networks in the continuous and discrete-time frameworks. In the latter setting, which is further studied in Chapter 5, I also introduce the notion of *node attributes* and *directed* edges. I then review events or change points related to the *community life cycle*. In the continuous framework, a dynamic network is defined as a continuous process where each node and edge is given a continuous birth and death timestamps. The following definition is adapted from Definition 1 in Rossetti and Cazabet (2018).

Definition 1.4.10. *(Continuous dynamic network) A continuous dynamic network, denoted $\mathcal{G} = (V, E)$, is defined by a set of temporal nodes V and a set of temporal edges E. Each element of V is a triplet (v, t_s, t_e) where $t_e, t_s \in \mathbb{R}, t_e \leqslant t_s$ are the birth and death timestamps of the node v. Each element of E is a quadruplet (u, v, t_s, t_e) where $u, v \in V$ nodes and $t_e, t_s \in \mathbb{R}, t_e \leqslant t_s$ are the birth and death timestamps of the edge between u and v.*

In the discrete framework, often called *snapshot network* or *graph snapshots*, a dynamic network is defined as a temporally ordered sequence of graphs at discrete timestamps. In the next definition, adapted from Definition 2 in (Rossetti and Cazabet, 2018), I consider attributed networks, where each node in each snapshot is associated to a vector of attributes or covariates.

Definition 1.4.11. *(Snapshot network) A discrete-time dynamic network or snapshot network is a sequence of graphs $\mathcal{G} = (G_t)_{t\in\mathbb{N}}$. For each $t \in \mathbb{N}$, $G_t = (V_t, E_t, X_t)$ is a static graph with node set V_t, edge set E_t, and node attributes matrix $X \in \mathbb{R}^{n_t \times d_t}$, where $n_t = |V_t|$ and $d_t \in \mathbb{Z}_{\geqslant 0}$ is the dimension of the node attributes at time t.*

In practice, the dimensions of nodes attributes is often constant, i.e., $d_t = d$. A snapshot network is related to a multi-layer graph, however, in the former object, the set of graphs is temporally ordered

and edges only connect nodes within the same snapshot. While the continuous-time framework is more closely related to temporal processes, the discrete-time framework is an easier representation to handle for its similarity to multi-layer graphs. Therefore, it is more frequently used in practice, and can well represent aggregated network data, e.g., number of phone calls or emails over a day or week. It is also the framework adopted in most dynamic random graph models such as dynamic latent space models (Durante and Dunson, 2014), dynamic stochastic block models (Matias and Miele, 2017; Padilla et al., 2019; Bhattacharjee et al., 2020; Yu et al., 2021), low-rank models (Bao and Michailidis, 2018) and the dynamic graphon model (Zhao et al., 2019).

In the dynamic setting, a network can exhibit a dynamic or evolving community structure, where the number of clusters and the memberships can change over time. The transformations that a community can undergo are called *events of the community life cycle* (Rossetti and Cazabet, 2018). Examples of such transformations include “birth" and “death", corresponding to the time stamps at which a community appears and disappears from the network, “growth" and “contraction", i.e., an increase and decrease of the community size, and the “split", the phenomenon by which a community is divided into two components. The reverse of the “split" is the “merge" transformation, where two communities unite into a single one. While the birth, death, split, and merge transformations generally happen at a single time stamp, the growth and contraction events can span several time stamps.

Events modifying the community structure and occurring within a single timestamp can correspond to change points for a dynamic network. More broadly, a change point, or a phase transition, is a timestamp at which an abrupt change occurs, in the topological structure of the network, or in the covariates distribution. In the discrete-time framework, a dynamic network observed over a period of time $T > 0$, denoted $\mathcal{G} = (G_1, \ldots, G_T)$, is said to have a single change point if there exists $\tau \in [T]$ such that

$$G_t \sim \mathcal{P}_1, \quad 1 \leqslant t < \tau,$$
$$G_t \sim \mathcal{P}_2, \quad \tau \leqslant t \leqslant T,$$

where $\mathcal{P}_1, \mathcal{P}_2$ are two graph distributions, or graph generative mechanisms. This definition can be generalised to multiple change points $(\tau_1, \ldots, \tau_K)$, which partition the set of timestamps $[T]$ into $(K + 1)$ segments called epochs (Bao and Michailidis, 2018).

1.4.5 Graph convolutional networks

Graph convolutional networks (GCN) are deep-learning algorithms operating on graphs and belong to the broader of graph-based neural networks. Many variants of GCNs have been defined, notably the spatial GCNs (Niepert et al., 2016) and spectral GCNs (Bruna et al., 2013). In this section, I will describe the simplified spectral version by Welling and Kipf (2016), introduced in the context of semi-supervised classification.

Let G be a (static) graph with adjacency matrix A and possibly some node attributes $X \in \mathbb{R}^{n \times d_0}$. A GCN is a model $f(G)$ defined as a composition of maps (or *layers*), which propagate a hidden representation, or *embedding*, of the graph. Let $L \geqslant 1$ be the number of layers, i.e., the depth of the GCN. For each layer $l \in [L]$, let $d_l \geqslant 1$ be the number of units (i.e., the width). The embedding of G at layer l, denoted $H^{(l)}$, is defined as

$$H^{(l)} = \sigma\left(\tilde{A} H^{(l-1)} W^{(l)} + B^{(l)}\right), \tag{1.10}$$

where σ is a pointwise activation function such as ReLU, $W^{(l)} \in \mathbb{R}^{d_{l-1} \times d_l}$ is a weight matrix, $B^{(l)} \in \mathbb{R}^{d_l}$ is a bias vector, and $\tilde{A} = \tilde{D}^{-1/2}(A + I)\tilde{D}^{-1/2}$ is the *normalised augmented adjacency* matrix with degree matrix $\tilde{D} = \mathrm{diag}((A + I)\mathbb{1})$. The propagation rule (1.10) proposed by Welling and Kipf (2016) comes from a first-order approximation of a graph convolution operation.

The initial representation $H^{(0)}$, i.e., the input of the first layer, is either the node attributes matrix X or a positional encoding matrix, such as $\tilde{D}$, if G is unattributed. The top layer of a GCN depends on the final task and objective function. For a node classification task, the activation function at layer L is generally the softmax function $s(x)_i = \frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_j}}$, $i \in [K]$, $x \in \mathbb{R}^K$, with K the number of classes. For a graph-level task, the final embedding matrix $H^{(L)} \in \mathbb{R}^{n \times d_L}$ is first *pooled* into a graph embedding vector $\tilde{H}^{(L)} \in \mathbb{R}^{d_L}$, before applying the activation function. A pooling operation can be a sum $\tilde{H}^{(L)}_j = \sum_{i=1}^{n} H^{(L)}_{ij}$ or a max operation $\tilde{H}^{(L)}_j = \max_{i \in [n]} H^{(L)}_{ij}$, $\forall j \in [d_L]$.

The weight and bias parameters $W^{(l)}$ and $B^{(l)}$ at each layer $l \in [L]$ are trainable, i.e., they are optimised to minimise a loss function, using a back-propagation algorithm. For instance, in a binary graph classification task, the goal is to learn a model f such that for any test graph G^*, $f(G^*) \in [0, 1]$ is the probability of G^* to belong to one of the two classes. In a supervised learning context, the loss function is often the binary cross entropy (BCE) loss. Given a training data set

$\mathcal{D} = \{(G_i, y_i)\}_{i \in [N]}$, with $y_i \in \{0, 1\}$ the class label of the graph G_i, for each i, the BCE loss function is defined as

$$\mathcal{L}_{BCE}(\mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} \left[y_i \log f(G_i) + (1 - y_i) \log(1 - f(G_i)) \right].$$

In practice, graph convolutional networks are often shallow, i.e., $2 \leqslant L \leqslant 5$, to limit the *over-smoothing* phenomenon in deep GCNs (Cai and Wang, 2020).

2 | Bayesian nonparametric estimation of nonlinear Hawkes processes

Submitted to the Bernoulli journal.

Submitted to Bernoulli

Bayesian estimation of nonlinear Hawkes processes

DÉBORAH SULEM [1], VINCENT RIVOIRARD [2,†] and JUDITH ROUSSEAU [1,*]

[1] *University of Oxford, E-mail:* deborah.sulem@stats.ox.ac.uk; [*] judith.rousseau@stats.ox.ac.uk
[2] *Ceremade, CNRS, UMR 7534, Université Paris-Dauphine, PSL University, 75016 Paris, France.*
E-mail: [†] Vincent.Rivoirard@dauphine.fr

Multivariate point processes (MPPs) are widely applied to model the occurrences of events, e.g., natural disasters, online message exchanges, financial transactions or neuronal spike trains. In the Hawkes process model, the probability of occurrences of future events depend on the past of the process. This model is particularly popular for modelling interactive phenomena such as disease expansion. In this work we consider the nonlinear multivariate Hawkes model, which allows to account for *excitation* and *inhibition* between interacting entities. We provide theoretical guarantees for applying nonparametric Bayesian estimation methods in this context. In particular, we obtain concentration rates of the posterior distribution on the parameters, under mild assumptions on the prior distribution and the model. These results also lead to convergence rates of Bayesian estimators. Another object of interest in event-data modelling is to infer the *graph of interaction* - or Granger causal graph. In this case, we provide consistency guarantees; in particular, we prove that the posterior distribution is consistent on the graph adjacency matrix of the process, as well as a Bayesian estimator based on an adequate loss function.

MSC2020 subject classifications: Primary 62G20, 62G05; secondary 60G65
Keywords: Nonlinear Hawkes processes; Nonparametric Bayesian inference; Causal Graph

1. Introduction

1.1. Nonlinear Hawkes processes

The Hawkes model is a popular temporal point process (PP) for modelling the occurrences of event-type phenomena. Extending the Poisson cluster process [45], this model allows the probability of occurrence of a new event to depend on the history of the process. The first construction by Hawkes [33] aimed at modelling the *self-excitatory* behaviour of earthquakes' strikes with aftershocks, and is called the *linear Hawkes process*. Since then, it has been extensively used, partly due to its interpretable parameters and branching structure representation [51]. This notably leads to tractable inference and simulation methods [3, 9, 32].

Hawkes processes have been largely and successfully applied in various contexts of correlated event-data, including online social popularity [23], stock prices moves [21], topic modelling [19], DNA motifs occurrences [7, 31, 52], and neuronal activity modelling [10, 38, 50]. They are used to infer both diffusion phenomena on networks and the structure of time-dependent networks [44]. Related and extended models include the mutually-regressive PP [1], the age-dependent [48] and marked [37] Hawkes processes, the dynamic contagion process [12], the reactive PP [22], the self-correcting PP [35] and the Dirichlet-Hawkes process [19]. More recently, neural point processes inspired by the Hawkes model *have also been proposed* [18, 42].

In a multivariate temporal PP, each dimension represents an entity, a location or a type of event - it is equivalent to a *marked* point process with finite mark space. For $K \in \mathbb{N}\backslash\{0\}$, the PP can be described

as a counting process $N = (N_t)_t = (N_t^1, \dots, N_t^K)_{t \geqslant 0}$, where N_t^k denotes the number of events that have occurred until time t at location k. Its dynamics are characterised by a conditional intensity function $(\lambda_t)_t = (\lambda_t^1, \dots, \lambda_t^K)_{t \geqslant 0}$, which is informally the infinitesimal rate of event conditionally on the past of the process, i.e, for $k = 1, \dots, K$, $\lambda_t^k dt = \mathbb{P}\left[N_t^k \text{ has a jump in } [t, t+dt] | \mathcal{G}_t\right]$, where $\mathcal{G}_t$ is the history of the process up to time t. In the nonlinear Hawkes model, only one dimension N^k of the process can jump at each time t and the intensity process has the following form

$$\lambda_t^k = \phi_k \left(\nu_k + \sum_{l=1}^{K} \int_{-\infty}^{t^-} h_{lk}(t-s) dN_s^l \right), \quad k = 1, \dots, K. \tag{1}$$

In (1), the parameter $\nu_k > 0$ denotes the *background* - or *spontaneous* - rate of events, and models exogeneous influences. The endogenous effects on the process are parametrised by *interaction functions* $(h_{lk})_{l,k=1}^K$ - or *triggering kernels*. More precisely, for $(l,k) \in [K]^2$, the function $h_{lk} : \mathbb{R} \to \mathbb{R}$ models the influence of component N^l onto component N^k. It can be decomposed into an *excitating* contribution ($h_{lk}^+ = \max(h_{lk}, 0)$) and an *inhibiting* contribution ($h_{lk}^- = \max(-h_{lk}, 0)$). Finally, the *link* or *activation function* $\phi_k : \mathbb{R} \to \mathbb{R}^+$ ensures that the intensity is a non-negative process, and is generally chosen to be monotone non-decreasing. If all the interaction functions h_{lk} are non-negative and all the link functions equal the identity functions, (1) corresponds to the linear Hawkes model.

The dependence on past events in the intensity (1) leads to a notion of *causality*. For Hawkes processes, a Granger-causal relationship between two components of the process corresponds to a non-null interaction function [20]. We can define the *connectivity graph* parameter $\delta \in \{0,1\}^{K^2}$ such that for each (l,k), $\delta_{lk} = 1$ if the function h_{lk} in (1) is non null and $\delta_{lk} = 0$ otherwise. We note that this parameter is redundant with $(h_{lk})_{l,k=1}^K$.

To the best of our knowledge, the estimation of the parameters of nonlinear Hawkes processes $\nu = (\nu_k)_k$, $h = (h_{lk})_{l,k=1}^K$, $\delta = (\delta_{lk})_{l,k=1}^K$ - as well as additional parameters of the link functions $(\phi_k)_k$ has not been theoretically analysed, neither in the frequentist nor in the Bayesian frameworks. In the nonparametric setting, the existing results apply to linear Hawkes processes for the estimation of (ν, h) [17] and for the estimation of the connectivity graph δ [32, 9]. In the nonlinear model, [8] study the estimation of the cross-covariances of the process, and [62] estimate a piecewise-constant link function assuming a parametric form on the interaction functions.

In this work, we analyse the theoretical properties of Bayesian methods for estimating ν, h, δ and additional parameters of the nonlinear functions $(\phi_k)_k$. We consider a prior distribution on the parameters, say Π, and our aim is to study posterior concentration rates in such models. More precisely, we wish to determine $\epsilon_T = o(1)$ and conditions on the model and on Π such that

$$\mathbb{E}_{f_0}[\Pi(d(f, f_0) > \epsilon_T | N)] \xrightarrow[T \to \infty]{} 1,$$

where $f = (\nu, h)$, $d(.,.)$ is some loss function on the parameter space, and $\Pi(.|N)$ denotes the posterior distribution given an observation of the process on $[0, T]$. In the last equation, we assume that the data N is generated by a Hawkes process with *true parameter* f_0, and we denote $\mathbb{P}_{f_0}$ its generating distribution and $\mathbb{E}_{f_0}$ the associated expectation. In particular, a consequence of such result is the construction of estimators on ν, h which converge in the frequentist sense at the rate ϵ_T. We also obtain posterior consistency results on the graph parameter δ, and construct a consistent risk-minimising estimator.

1.2. Related works

There is a rich literature on Hawkes processes in probability, statistics, and more recently in machine learning and deep learning. The stability properties of the nonlinear Hawkes model have been studied

under several assumptions [5, 36], together with the rate of convergence to the stationary solution [6] and the Bartlett spectrum [41]. Regenerative properties of Hawkes processes were investigated for the models with finite [11] and infinite [29, 47] memory. Recently [4, 24, 25] derived functional central limit theorems and large deviations principles for ergodic processes. Malliavin-Stein calculus was applied by [56, 57] to establish Gaussian and Poisson approximations of functionals of the linear Hawkes process, and later by [34] to obtain Berry-Esséen bounds. Stationary distributions of high dimensional Hawkes processes were also studied, notably in the mean-field limit [13, 14, 48].

Many statistical works have been dedicated to designing robust and efficient estimation procedures in the linear Hawkes model. In the seminal work of [46], the interaction functions are given in a parametric form and estimated by maximising the likelihood function. In parametric models, an Expectation-Maximisation algorithm was proposed in [61] to compute the maximum likelihood estimator while MCMC methods were designed for sampling from the posterior distribution [49]. The EM algorithm was extended by [39] to nonparametric Hawkes models using a penalised likelihood objective. Another nonparametric approach was introduced by [52] for the linear univariate model by using a model selection strategy. In the multivariate Hawkes model, Lasso-type estimates were designed by [32]. Still for linear models, Bayesian approaches have also been implemented for nonparametric Hawkes models, see for instance [19]. In [17] the authors study asymptotic properties of the posterior distribution in the linear model.

Causality graphs for discrete-time events were introduced by [30] and extended to marked point processes by [16], with an explicit definition in the case of multivariate Hawkes processes by [20]. In linear parametric models, some approaches optimise a least-square objective based on the intensity process [3, 4]. For nonparametric Hawkes processes, [63] apply an EM algorithm based on a penalized likelihood objective leading to temporal and group sparsity. Still in the linear model, Lasso-type estimates proposed by [32] for nonparametric Hawkes processes naturally lead to sparse connectivity graphs. This procedure has been generalised to high-dimensional processes by [9] by adding an edge screening step.

1.3. Our contributions

This paper considers a general multivariate Hawkes model with a nonlinear and nonparametric form of the intensity function, and provides theoretical guarantees on Bayesian estimation methods. We cover a large range of link functions ϕ_k, which covers most of the nonlinear Hawkes models considered in the literature [11, 32, 26, 7, 8, 43, 42, 15, 58], such as the ReLU $\phi_k(x) = (x)_+ = \max(x, 0)$, clipped exponentials $\phi_k(x) = \min(e^x, \Lambda)$, the sigmoid $\phi_k(x) = (1 + e^{-x})^{-1}$, and the softplus $\phi_k(x) = \log(1 + e^x)$. These models have been notably introduced for neuronal spike-train data modelling, where intense-activity periods alternate with resting states called *refractory periods*[1]. The ReLU function directly extends the original linear Hawkes model to handle negative interaction functions. In [32, 11] it is called the *standard* nonlinear Hawkes model, as it is the closest to the linear Hawkes process. Exponential and sigmoidal functions appear in several applied works [26, 7], where smoothness, saturation and thresholding effects are desirable properties. The softplus function is often preferred in machine learning algorithms as a soft approximation of ReLU [42].

The first question to answer is the identifiability of $f = (\nu, h)$, which is treated in Section 2.2. Building on these results, we study posterior concentration rates in terms of the L_1-norm on f in Section 3.1. Our aim is to describe the posterior concentration rates in terms of conditions on the prior Π and on the true parameter $f_0 = (\nu_0 = (\nu_k)_k, h_0 = (h^0_{lk})_{l,k})$ which are simple to verify and under rather weak

[1] A refractory period is a time interval during which a neuron is unlikely to emit a spike train.

assumptions on the link functions. Interestingly, we eventually reduce the problem to conditions on the prior and the f_0 similar to those found in the literature on density and nonlinear regression estimation (see Theorem 3.2), which makes them easy to verify in a wide range of prior models. From this we derive convergence rates of Bayesian estimators of f_0 (Corollary 3.8) and posterior consistency on δ_0 (Theorem 3.9), with δ_0 the true graph parameter associated to h_0.

We also extend our results to the case where the link functions are partially unknown, in the special case of shifted ReLU link functions. More precisely we consider models in the form $\phi_k(x) = \theta_k + (x)_+$, with $\theta_k > 0$ unknown. For such models we show identifiability of the parameters (f, θ), $\theta = (\theta_k)_{1 \leqslant k \leqslant K}$ and derive a general posterior concentration rate result similar to Theorem 3.2 on both f and θ.

To the best of our knowledge, these results are the first theoretical properties on the nonparametric estimation of both f_0 and δ_0 in the frequentist and Bayesian literature of nonlinear Hawkes processes. Besides, for partially known link functions, in the particular setting of the shifted ReLU model, we also provide the first result on the estimation of the additional parameter θ. We note that recently, computational methods for a related setting have been developed in [65, 66, 40, 64]. In the latter works, a sigmoidal nonlinear Hawkes model is defined with $\phi_k(x) = \theta_k(1 + e^{-x})^{-1}$ and unknown parameter $\theta = (\theta_k)_k$. However, although the theoretical analysis of the latter model is beyond the scope of this paper, it is similar in spirit to our models. In fact, our techniques could potentially be applied to this multiplicative parametrisation, which we leave for future work.

Our results are related to those of [17], obtained in the case of linear Hawkes processes. However, the analysis of the process and our proofs for estimating the parameter rely on *renewal* properties, newly introduced by [11] in the univariate ReLU nonlinear Hawkes model. One key novelty of our work is to leverage the concept of *excursions* in the context of statistical analysis. This concept allows to decompose the trajectory of the process into independent, observable subintervals, and also to analyse the process on specific events where the parameter estimation is simplified. Developing these tools for nonlinear processes is fundamental since classical technical arguments used for linear Hawkes processes and based on Poisson branching structures cannot be applied in this case. We believe that these new proof techniques have an interest in themselves, in addition to weakening some of the assumptions on the prior distribution considered in [17].

The rest of the paper is organised as follows. In Section 2, we define the multivariate stationary nonlinear Hawkes process, present the identifiability results and describe the Bayesian framework. Section 3 presents the posterior concentration results on f and θ and consistency on δ results. Section 4 is dedicated to the construction of prior distributions that satisfy the assumptions of the theorems. The most novel aspects of the proofs are reported in Section 5. Appendix A contains some technical lemmas. Finally, supplementary proofs and results can be found in the supplementary material [55].

Notations. For a function h, we denote $\|h\|_1 = \int_{\mathbb{R}} |h(x)|dx$ the L_1-norm, $\|h\|_2 = \sqrt{\int_{\mathbb{R}} h^2(x)dx}$ the L_2-norm, $\|h\|_\infty = \sup_{x \in \mathbb{R}} |h(x)|$ the supremum norm, and $h^+ = max(h, 0)$, $h^- = max(-h, 0)$ its positive and negative parts. For a $K \times K$ matrix A, we denote $r(A)$ its spectral radius and $\|A\|$ its spectral norm. For a vector $u \in \mathbb{R}^K$, $\|u\|_1 = \sum_{k=1}^K |u_k|$. The notation $k \in [K]$ is used for $k \in \{1, \cdots, K\}$. For a set B and $k \in [K]$, we denote $N^k(B)$ the number of events of N^k in B and $N^k|_B$ the point process measure restricted to the set B. For random processes, the notation $\overset{\mathcal{L}}{=}$ corresponds to equality in distribution. We also denote $\mathcal{N}(u, \mathcal{H}_0, d)$ the covering number of a set $\mathcal{H}_0$ by balls of radius u w.r.t. a metric d. For any $k \in [K]$, let $\mu_k^0 = \mathbb{E}_0[\lambda_t^k(f_0)]$ be the mean of $\lambda_t^k(f_0)$ under the stationary distribution $\mathbb{P}_0$. For a set Ω, its complement is denoted Ω^c. We also use the notations $u_T \lesssim v_T$ if $|u_T/v_T|$ is bounded when $T \to \infty$, $u_T \gtrsim v_T$ if $|v_T/u_T|$ is bounded and $u_T \asymp v_T$ if $|u_T/v_T|$ and $|v_T/u_T|$ are bounded.

2. Problem setup

2.1. Definition and stationary distribution

In this section, we first recall the formal definition of a multivariate Hawkes process. We consider a probability space $(\mathcal{X}, \mathcal{G}, \mathbb{P})$ and a MPP $N = (N_t)_{t\in\mathbb{R}} = (N_t^1, \dots, N_t^K)_{t\in\mathbb{R}}$. Let $\{\mathcal{G}_t\}_{t\in\mathbb{R}}$ be the filtration such that $\mathcal{G}_t = \sigma(N_s, s \leqslant t)$ and for $T > 0$, we assume that $\mathcal{G}_T \subset \mathcal{G}$. We say that $(N_t)_t$ is a multivariate Hawkes process with parameter $f = ((\nu_k)_{k=1}^K, (h_{lk})_{l,k=1}^K, (\theta_k)_{k=1}^K)$ adapted to $\mathcal{G}$ if

i) almost surely, $\forall k, l \in [K]$, $(N_t^k)_t$ and $(N_t^l)_t$ never jump simultaneously;
ii) for all $k \in [K]$, the $\mathcal{G}_t$-predictable intensity process of N^k at $t \in \mathbb{R}$ is given by

$$\lambda_t^k(f) = \phi_k\left(\nu_k + \sum_{l=1}^K \int_{-\infty}^{t^-} h_{lk}(t-s)dN_s^l\right), \quad k = 1, \dots, K.$$

We consider finite-memory Hawkes processes for which interaction functions have a bounded support included in $[0, A]$ with $A > 0$ known - chosen arbitrarily large in practice. We recall that in (1), if for all k, ϕ_k is the identity function and for all l, h_{lk} is non-negative, this PP model corresponds to the classical linear Hawkes process with parameter $\nu = (\nu_k)_{k=1}^K$ and $h = (h_{lk})_{k,l=1}^K$ and intensity process:

$$\tilde{\lambda}_t^k(\nu, h) := \nu_k + \sum_{l=1}^K \int_{t-A}^{t^-} h_{lk}(t-s)dN_s^l. \tag{2}$$

With this notation, the nonlinear intensity can be written as $\lambda_t^k(f) = \phi_k(\tilde{\lambda}_t^k(\nu, h))$. For a nonlinear Hawkes process, the existence and uniqueness of a stationary distribution is proved under some assumptions on the parameters f and the link functions $\phi = (\phi_k)_k$. In the following lemma, we provide two sufficient conditions, which are variants of existing work. We recall that a function ϕ is L-Lipschitz, if for any $(x, x') \in \mathbb{R}^2$, $|\phi(x) - \phi(x')| \leqslant L|x - x'|$.

Lemma 2.1. *Let N be a Hawkes process with parameter f and link functions $(\phi_k)_k$ such that for any $k \in [K]$, $\phi_k : \mathbb{R} \to \mathbb{R}^+$ is monotone non-decreasing and L-Lipschitz, with $L > 0$. If one of the following conditions is satisfied:*

(C1) *The matrix S^+ with entries $S_{lk}^+ = L\left\|h_{lk}^+\right\|_1$ satisfies $r(S^+) < 1$;*
(C2) *For any $k \in [K]$, ϕ_k is bounded, i.e., $\exists \Lambda_k > 0, \forall x \in \mathbb{R}$, $\phi_k(x) \leqslant \Lambda_k$.*

then there exists a unique stationary version of the process N with finite average.

In the previous lemma, the second stationarity condition **(C2)** directly comes from Theorem 7 by [5] and is applied to our (less general) context of Lipschitz and non-decreasing link functions. The first condition **(C1)** is obtained in Theorem 1 of [15], in a more restricted Hawkes model where $\phi_k(x) = (x)_+$ and the interaction functions are of the form $h_{lk} = K_{lk}g(t)$ with $g \geqslant 0$ and $K_{lk} \in \mathbb{R}$, but the same arguments can be applied to prove the stationarity of the process in our more general nonlinear model. However, in the context of inference, we will consider a slightly stronger condition:

(C1bis) The matrix S^+ with entries $S_{lk}^+ = L\left\|h_{lk}^+\right\|_1$ satisfies $\left\|S^+\right\| < 1$.

From now on, we will assume that we observe on a window $[-A, T]$ a stationary Hawkes process with link functions $(\phi_k)_k$ and true parameters $f_0 = ((\nu_k^0)_{k=1}^K, (h_{lk}^0)_{l,k=1}^K)$. We denote $\mathbb{P}_0$ the stationary distribution of N and $\mathbb{P}_0(.|\mathcal{G}_0)$ its conditional distribution given $\mathcal{G}_0$. We note that $\mathbb{P}_0$ is a well-defined

transformation of the probability distribution $\mathbb{P}$ (through its alternative definition in Lemma S9.2 [55]). For $f = ((\nu_k)_{k=1}^K, (h_{lk})_{l,k=1}^K)$ satisfying the assumptions of Lemma 2.1, the log-likelihood of the processs on $[0,T]$ conditionally on $\mathcal{G}_0$ (i.e., conditionally on $N|_{[-A,0)}$) is given by

$$L_T(f) := \sum_{k=1}^K \left[\int_0^T \log(\lambda_t^k(f)) dN_t^k - \int_0^T \lambda_t^k(f) dt \right].$$

Then, for any parameter f, we define the conditional distribution $\mathbb{P}_f$ from the likelihood function

$$d\mathbb{P}_f(.|\mathcal{G}_0) = e^{L_T(f) - L_T(f_0)} d\mathbb{P}_0(.|\mathcal{G}_0). \tag{3}$$

We denote $\mathbb{E}_0$ (resp. $\mathbb{E}_f$) the expectation associated with $\mathbb{P}_0$ (resp. $\mathbb{P}_f$).

2.2. Identifiability of the parameters

In this section, we provide some conditions on the model which ensure that the parameters of the nonlinear Hawkes model defined in (1) are identifiable. To do so we consider the following weak assumption.

***Assumption* 2.2.** For $f = (\nu, h)$, there exists $\varepsilon > 0$ such that for any $k \in [K]$, ϕ_k restricted to $I_k = (\nu_k - \max_{l \in [K]} \|h_{lk}^-\|_\infty - \varepsilon, \nu_k + \max_{l \in [K]} \|h_{lk}^+\|_\infty + \varepsilon)$ is injective.

Proposition 2.3. *Let N be a nonlinear Hawkes process as defined in (1) with link functions $(\phi_k)_k$ and parameter $f = (\nu, h)$ satisfying the conditions of Lemma 2.1 and Assumption 2.2. If N' is a Hawkes processes with the same link functions $(\phi_k)_k$ and parameter $f' = (\nu', h')$, then if N and N' have the same distribution, i.e., $N \overset{\mathcal{L}}{=} N'$, then $\nu = \nu'$ and $h = h'$.*

Note that if the ϕ_k's are injective on $\mathbb{R}$, which holds in particular for the sigmoid and the softplus functions, then Assumption 2.2 is verified for all f. However our result is more general and also covers link functions which are only injective on a sub-interval of $\mathbb{R}$ such as ReLU or shifted ReLU ($\phi_k(x) = \theta_k + \max(x, 0)$) and clipped exponentials ($\phi_k(x) = \min(e^x, \Lambda_k)$). In this case, Assumption 2.2 holds over a restricted parameter space for f. More precisely, ϕ_k needs to be injective over an interval which includes all the possible values of $\nu_k + h_{lk}(s)$, for any $l \in [K]$ and $s \in [0, A]$.

***Remark* 2.4.** One consequence of Assumption 2.2 is that for any $t > 0$ such that $N[t - A, t) \leqslant 1$, then $\lambda_t^k(f) > 0$ (since ϕ_k is non-negative and monotone non-decreasing) for all $k \in [K]$. However, Assumption 2.2 still allows to model the *refractory periods* of biological neurons, i.e., when the neurons cannot or are very unlikely to fire again during a period after firing. Indeed, one can have $\lambda_t^k(f)$ very small for t such that $N^k[t - A, t) = 1$, depending on f and ϕ_k.

Proposition 2.3 supports the feasibility of the parameter estimation when the nonlinear functions ϕ_k's are fully known. It can however be extended to the setup where the link functions are partially known. In the next proposition, we consider the case of $\phi_k(x) = \theta_k + \psi_k(x)$ where ψ_k is a function such that $\lim_{x \to -\infty} \psi_k(x) = 0$ and $\theta_k \geqslant 0$ is an unknown parameter, for each $k \in [K]$. In this case, we denote $\lambda_t(f, \theta)$ the intensity process.

Proposition 2.5. *Let N be a Hawkes process with parameter $f = (\nu, h)$ and link function $\phi_k(x;\theta_k) = \theta_k + \psi_k(x)$ with $\theta_k \geqslant 0$ for any $k \in [K]$ satisfying the conditions of Lemma 2.1 and Assumption 2.2. We also assume that for all $k \in [K]$, $\lim_{x\to-\infty} \psi_k(x) = 0$ and*

$$\exists l \in [K],\ x_1 < x_2, \quad \text{such that } h^-_{lk}(x) > 0, \quad \forall x \in [x_1, x_2]. \tag{4}$$

Then if N' is a Hawkes processes with link functions $\phi_k(x;\theta'_k) = \theta'_k + \psi_k(x)$, $\theta'_k \geqslant 0$ and parameter $f' = (\nu', h')$,

$$N \overset{\mathcal{L}}{=} N' \implies \nu = \nu', \quad h = h', \quad \text{and} \quad \theta = \theta', \quad \theta = (\theta_k)_{k=1}^K, \quad \theta' = (\theta'_k)_{k=1}^K.$$

Besides, in this case we have $\mathbb{P}_f[\inf_{t\geqslant 0} \lambda^k_t(f,\theta) = \theta_k] = 1$.

The proofs of Propositions 2.3 and 2.5 are reported in Section S7.1 in the supplementary material [55]. In Proposition 2.5, the condition (4) implies that each component k receives some inhibition (i.e., $\exists l,\ h^-_{lk} \neq 0$). In particular, we will use this condition in the shifted ReLU model where $\psi_k(x) = (x)_+$. We note that θ_k is not identifiable when no inhibition is received by N^k (i.e., when $\forall l,\ h^-_{lk} = 0$). More precisely, the following lemma - proved in Section S7.1 in the supplementary material [55] - states that in a mutually-exciting ReLU model, the parametrisation of the process is not unique. Informally, our models present a singularity at the parameter "$h^- = 0$".

Lemma 2.6. *Let N be a Hawkes process with parameter $f = (\nu, h)$ and link functions $\phi_k(x;\theta_k) = \theta_k + (x)_+$, $\theta_k \geqslant 0$, $k \in [K]$ satisfying Assumption 2.2, and let $k \in [K]$. If $\forall l \in [K], h_{lk} \geqslant 0$, then for any $\theta'_k \geqslant 0$ such that $\theta_k + \nu_k - \theta'_k > 0$, let N' be the Hawkes process driven by the same underlying Poisson process Q as N (see Lemma S9.2 [55]) with parameter $f' = (\nu', h')$ and link functions $\phi_k(x;\theta'_k) = \theta'_k + (x)_+, k \in [K]$ with $\nu' = (\nu_1, \dots, \nu_k + \theta_k - \theta'_k, \dots, \nu_K) \neq \nu$, $h' = h$, and $\theta' = (\theta_1, \dots, \theta'_k, \dots, \theta_K) \neq \theta$. Then for any $t \geqslant 0, \lambda^k_t(f,\theta) = \lambda^k_t(f',\theta')$, and therefore $N \overset{\mathcal{L}}{=} N'$.*

2.3. Bayesian inference

We can now present our Bayesian estimation framework. We assume that the observed Hawkes process N satisfies the conditions of Lemma 2.1, i.e., the link functions ϕ_k's are monotone non-decreasing, L-Lipschitz with $L > 0$ and either we consider a bounded model $\phi_k(x) \leqslant \Lambda, \forall k, \Lambda > 0$ (condition **(C2)**) or we assume $\left\|S^+_0\right\| < 1$ (condition **(C1bis)**) with $S^+_0 = (L\left\|h^{0+}_{lk}\right\|_1)_{l,k\in[K]^2}$. We define the parameter space for $f = ((\nu_k)_{k=1}^K, (h_{lk})_{l,k=1}^K)$ and the functional space as follows. Let

$$\mathcal{H}' = \{h : [0,A] \to \mathbb{R};\ \|h\|_\infty < \infty\}, \quad \mathcal{H} = \left\{h = (h_{lk})_{l,k=1}^K \in {\mathcal{H}'}^{K^2};\ (h,\phi) \text{ satisfy } \textbf{(C1bis)} \text{ or } \textbf{(C2)}\right\},$$

$$\mathcal{F} = \left\{f = (\nu, h) \in (\mathbb{R}_+ \setminus \{0\})^K \times \mathcal{H};\ f \text{ satisfies Assumption } 2.2\right\}.$$

We recall that for an unbounded link function, condition **(C1bis)** corresponds to $\left\|S^+\right\| < 1$ with $S^+ = (L\left\|h^+_{lk}\right\|_1)_{l,k\in[K]^2}$. We also recall that $A > 0$ is fixed. In the graph estimation problem (see Section 3.2), the parameter of interest is $\delta_0 \in \{0,1\}^{K^2}$ where $h^0_{lk} = 0 \iff \delta^0_{lk} = 0$. With a slight abuse of notations, we sometimes write $f = ((\nu_k)_k, (h_{lk})_{l,k})_k, (\delta_{lk})_{l,k})$ with $\delta \in \{0,1\}^{K^2}$.

***Remark* 2.7.** With ReLU-type link functions, we have $\mathcal{H} = \{h = (h_{l,k}) \in (\mathcal{H}')^{K^2};\ \|S^+\| < 1\}$ and $\mathcal{F} = \{f \in (\mathbb{R}_+\backslash\{0\})^K \times \mathcal{H};\ \|h_{lk}^-\|_\infty < \nu_k,\ (l,k) \in [K]^2\}$. With clipped exponential links $\phi_k(x) = \min(e^x, \Lambda_k)$, we have $\mathcal{H} = \mathcal{H}'^{K^2}$ and $\mathcal{F} = \left\{f \in \mathbb{R}_+\backslash\{0\}^K \times \mathcal{H}'^{K^2};\ \nu_k + \|h_{lk}^+\|_\infty < \log\Lambda_k,\ (l,k) \in [K]^2\right\}$.

We now define our metric on the parameter space $\mathcal{F}$. For any $f = (\nu, h),\ f' = (\nu', h') \in \mathcal{F}$, we define the following L_1-distances:

$$\|\nu - \nu'\|_1 = \sum_{k=1}^K |\nu_k - \nu_k'|, \quad \|h - h'\|_1 = \sum_{l=1}^K \sum_{k=1}^K \|h_{lk} - h_{lk}'\|_1, \quad \|f - f'\|_1 = \|\nu - \nu'\|_1 + \|h - h'\|_1 .$$

Finally, we consider a prior distribution Π on $\mathcal{F}$ and define the posterior distribution on $B \subset \mathcal{F}$ as

$$\Pi(B|N) = \frac{\int_B \exp(L_T(f))d\Pi(f)}{\int_{\mathcal{F}} \exp(L_T(f))d\Pi(f)} = \frac{\int_B \exp(L_T(f) - L_T(f_0))d\Pi(f)}{\int_{\mathcal{F}} \exp(L_T(f) - L_T(f_0))d\Pi(f)} =: \frac{N_T(B)}{D_T}, \tag{5}$$

denoting $N_T(B)$ and D_T our numerator and denominator of the posterior with the form above.

3. Main results

In this section, we state our most important results on the posterior distribution on the parameter f and the restriction on the connectivity graph δ, leading respectively to convergence rates and consistency of some Bayesian nonparametric estimators.

3.1. Posterior concentration rates

We first prove that under mild assumptions on the link functions and the true parameter, we can describe the posterior concentration rate ϵ_T with respect to the L_1-distance on $\mathcal{F}$ defined in Section 2.3, in terms of standard conditions on the prior. We then consider the case where the link functions ϕ_k depend on an unknown parameter, in the special case of shifted ReLU: $\phi_k(x; \theta_k^0) = \theta_k^0 + (x)_+$, for which we prove posterior concentration on both f_0 and θ_0. To do so, we use the following assumption on the true parameter, which is a strengthening of the identifiability condition in Assumption 2.2.

***Assumption* 3.1.** For $f_0 = (\nu_0, h_0)$, we assume that there exists $\varepsilon > 0$ such that for any $k \in [K]$, ϕ_k restricted to $I_k = (\nu_k^0 - \max_{l\in[K]} \|h_{lk}^{0-}\|_\infty - \varepsilon, \nu_k^0 + \max_{l\in[K]} \|h_{lk}^{0+}\|_\infty + \varepsilon)$ is bijective from I_k to $J_k = \phi_k(I_k)$ and its inverse is L'- Lipschitz on J_k, with $L' > 0$. We also assume that one of the two following conditions is satisfied:

i) For any $k \in [K]$, $\inf_{x\in\mathbb{R}} \phi_k(x) > 0$.

ii) For any $k \in [K]$, $\phi_k > 0$ and $\sqrt{\phi_k}$ and $\log\phi_k$ are L_1-Lipschitz with $L_1 > 0$.

The first part of Assumption 3.1, which is a slight strengthening of Assumption 2.2, holds in all cases described previously. The second part considers two cases: (i) the ϕ_k's are lower bounded by a positive constant, which holds for instance when $\phi_k(x; \theta_k) = \theta_k + \psi_k(x)$ with $\theta_k > 0$ and $\psi_k \geqslant 0$ and (ii) the ϕ_k's can approach 0 but satisfy an additional smoothness condition which holds in particular if the

derivatives ϕ_k' are bounded and the functions $\log \phi_k$'s are Lipschitz. It is notably the case for the commonly used Hawkes models [11, 32, 26, 7, 8, 43, 42], see Example 1 below. Note that this assumption excludes the standard ReLU function $\phi_k(x) = (x)_+$, which we will treat separately in Proposition 3.5.

***Example* 1.** The following nonlinear models verify Assumption 3.1. Let $s, t, \Lambda > 0$.

- **Positive or shifted ReLU**-type functions: $\phi_1(x) = \max(sx, t) \geqslant t > 0$, which is injective on $[t/s, +\infty)$, s-Lipschitz and its inverse on $[t, +\infty)$, $\phi_1^{-1}(x) = s^{-1}x$ is s^{-1}-Lipschitz.
- **Clipped exponential** functions: $\phi_2(x) = \min(e^{sx}, \Lambda)$, which is injective on $(-\infty, s^{-1}\log\Lambda]$ and $s\Lambda$-Lispchitz. Its inverse on $(0, \Lambda]$, $\phi_2^{-1}(x) = s^{-1}\log x$ is Lipschitz on any compact of $(0, \Lambda]$ and $\sqrt{\phi_2}(x) = \sqrt{\min(e^{sx}, \Lambda)} = \min(e^{sx/2}, \sqrt{\Lambda})$ and $\log\phi_2 = \min(sx, \log\Lambda)$ are respectively $s\Lambda$–Lispchitz and s-Lipchitz;
- **Sigmoid** functions: $\phi_3(x) = (1 + e^{-s(x-t)})^{-1}$, which is injective on $\mathbb{R}$ and s-Lipschitz. Its inverse $\phi_3^{-1}(x) = t + \frac{1}{s}\log\frac{x}{1-x}$ is Lipschitz on any compact of $(0, 1)$, $\sqrt{\phi_3}$ is s-Lipschitz and $\frac{\phi_3'(x)}{\phi_3(x)} \leqslant s$ thus $\log\phi_3$ is s-Lipschitz;
- **Softplus** functions: $\phi_4(x) = \log(1 + e^{s(x-t)})$, which is injective on $\mathbb{R}$, s-Lipschitz and its inverse $\phi_4^{-1}(x) = \frac{1}{s}\log(e^x - 1) + t$ is Lipschitz on any compact of $\mathbb{R}_+^*$. Finally $\sqrt{\phi_4}$ and $\log\phi_4$ are s-Lipschitz.

To state our first result, we also define the following neighbourhoods in f_0 in supremum and L_2-norms respectively, for $B > 0$:

$$B_\infty(\epsilon_T) = \{f \in \mathcal{F};\ \nu_k^0 \leqslant \nu_k \leqslant \nu_k^0 + \epsilon_T,\ h_{lk}^0 \leqslant h_{lk} \leqslant h_{lk}^0 + \epsilon_T,\ (l,k) \in [K]^2\}.$$

$$B_2(\epsilon_T, B) = \{f \in \mathcal{F};\ \max_k |\nu_k - \nu_k^0| \leqslant \epsilon_T,\ \max_{l,k} \|h_{lk} - h_{lk}^0\|_2 \leqslant \epsilon_T,\ \max_{l,k} \|h_{lk}\|_\infty < B\}.$$

In particular, $B_\infty(\epsilon_T)$ is chosen so that for any $f \in B_\infty(\epsilon_T)$, $k \in [K]$ and $t \geqslant 0$, the intensities verify $\lambda_t^k(\nu, h) \geqslant \lambda_t^k(\nu_0, h_0)$. Finally we define

$$\kappa_T = 10(\log T)^r \tag{6}$$

with $r = 0$ if $(\phi_k)_k$ satisfies Assumption 3.1 (i), $r = 1$ if $(\phi_k)_k$ satisfies Assumption 3.1 (ii).

Theorem 3.2. *Let N be a Hawkes process with known link functions $\phi = (\phi_k)_k$ and parameter $f_0 = (\nu_0, h_0)$ such that (ϕ, f_0) satisfy Assumption 3.1. Let $\epsilon_T = o(1/\sqrt{\kappa_T})$ be a positive sequence verifying $\log^3 T = O(T\epsilon_T^2)$ and Π be a prior distribution on $\mathcal{F}$. We assume that the following conditions are satisfied for T large enough.*

***(A0)** There exists $c_1 > 0$ such that $\Pi(B_\infty(\epsilon_T)) \geqslant e^{-c_1 T\epsilon_T^2}$.*

***(A1)** There exist subsets $\mathcal{H}_T \subset \mathcal{H}$ and $c_2 > 0$ such that, with $\Upsilon_T = \{\nu = (\nu_k)_k,\ 0 < \nu_k \leqslant e^{c_2 T\epsilon_T^2},\ \forall k\}$, $\Pi(\mathcal{H}_T^c) + \Pi(\Upsilon_T^c) = o(e^{-(\kappa_T + c_1)T\epsilon_T^2})$.*

***(A2)** There exist $\zeta_0 > 0$ and $x_0 > 0$ such that $\log \mathcal{N}(\zeta_0\epsilon_T, \mathcal{H}_T, \|.\|_1) \leqslant x_0 T\epsilon_T^2$.*

Then, for $M > 0$ large enough, we have

$$\mathbb{E}_0\left[\Pi(\|f - f_0\|_1 > M\sqrt{\kappa_T}\epsilon_T \big| N)\right] = o(1). \tag{7}$$

The proof of Theorem 3.2 is provided in Section 5.2.

***Remark* 3.3.** In Theorem 3.2, if we replace $B_\infty(\epsilon_T)$ by $B_2(\epsilon_T, B)$ for some $B > 0$ in **(A0)**, then the concentration rate in (7) is $\sqrt{\log\log T\kappa_T}\epsilon_T$ instead of $\sqrt{\kappa_T}\epsilon_T$. Replacing $B_\infty(\epsilon_T)$ by $B_2(\epsilon_T, B)$ can be

useful for some families of priors, as seen in the case of mixtures of Beta distributions in Section S4.1 in the supplementary material [55].

***Remark* 3.4.** Our concentration rate in (7) holds under the stationary distribution $\mathbb{P}_0$, implying in this case that the "initial condition" $N|_{[-A,0]} \subset \mathcal{G}_0$ also comes from the stationary law. However, in practice, one might observe a process on $[-A, T]$ with an arbitrary distribution on $[-A, 0]$. Under the conditions of Lemma 2.1, the dynamics of the resulting process are *stable* (in the sense of Definition 1 of [5]), using the results in [5]. In particular, its distribution $\mathbb{P}_0(.|\mathcal{G}_0)$ converges exponentially fast to the stationary distribution $\mathbb{P}_0$. Therefore, we expect that (7) would still hold under $\mathbb{P}_0(.|\mathcal{G}_0)$, i.e., under a more general initial distribution on $[-A, 0]$.

An interesting aspect of Theorem 3.2 is that the assumptions on the prior **(A0)**, **(A1)** and **(A2)**, are similar to those of simpler estimation problems like density estimation, regression or linear Hawkes processes. This allows to directly derive explicit forms of the posterior concentration rates in the non-linear Hawkes model under common families of priors, such as Gaussian processes, hierarchical Gaussian processes, basis expansions or mixture models (see [59, 60, 2, 53] or Section 2.3.2 of [17]). In Section 4, we illustrate this using splines and mixture models. Additionally, our Theorem 3.2 avoids the unpleasant assumption on the prior in Theorem 3 of [17] which requires that for some $u_0 > 0$, $\Pi(\|S\| > 1 - u_0(\log T)^{1/6}\epsilon_T^{1/3}) \leqslant e^{-2c_1 T\epsilon_T^2}$. This is thanks to our novel proof techniques using regeneration times under the true model $\mathbb{P}_0$ (see Section 5.1).

Theorem 3.2 provides posterior concentration rates for a large class of link functions, as discussed earlier. In particular, it covers the case of shifted ReLU link functions, i.e, $\phi_k(x) = \theta_k + (x)_+$ where $\theta_k > 0$ is a *baseline* rate, which can be arbitrarily small. This link function can be seen as an alternative to the exponential function with positive baseline rate in [26] In [26], neurons firing rates are modelled using nonlinear Hawkes processes with a positive link function, which still allows to account for the refractory periods of neurons (for which the firing rate is small). Moreover, while in the case of the shifted ReLU model, Theorem 3.2 assumes that the baseline rates $\theta = (\theta_k)_k$ are known, we show in the next proposition that we can also estimate θ. Besides, we additionally provide a posterior concentration result when using the standard ReLU function $\phi_k(x) = (x)_+$, under a stronger assumption on the model. We note that for this latter choice of link function, the intensity function is only *non-negative* and the likelihood function is equal to 0 in parts of neighbourhoods of h_0, which causes several issues in the control of the Kullback-Leibler divergence.

Before stating our results, we define neighbourhoods in θ_0, also in supremum and L_2-norms, respectively $B_\infty^\Theta(\epsilon_T) = \{\theta \in \Theta;\ \|\theta - \theta_0\|_\infty \leqslant \epsilon_T\}$ and $B_2^\Theta(\epsilon_T, B) = \{\theta \in \Theta;\ \|\theta - \theta_0\|_2 \leqslant \epsilon_T\}$, and in this case we define $\kappa_T = 10(\log T)^r$ with $r = 0$ in the shifted ReLU model (Case 2 of the following proposition) and $r = 2$ in the standard ReLU model (Case 1).

Proposition 3.5. *Let N be a nonlinear Hawkes process with link functions $(\phi_k)_k$ and parameter $f_0 = (\nu_0, h_0)$ satisfying Assumption 2.2. Let $\epsilon_T = o(1/\sqrt{\kappa_T})$ be a positive sequence verifying $\log^3 T = O(T\epsilon_T^2)$ and Π be a prior distribution on $\mathcal{F}$.*

- ***Case 1 (Standard ReLU):* $\phi_k(x) = (x)_+$*, for all* $k \in [K]$*.** *Under the Assumptions **(A0)**, **(A1)** and **(A2)** of Theorem 3.2, if f_0 verifies the following additional assumption*

$$\limsup_{T\to\infty} \frac{1}{T}\mathbb{E}_0\left(\int_0^T \frac{\mathbb{1}_{\lambda_t^k(f_0)>0}}{\lambda_t^k(f_0)}dt\right) < +\infty, \quad k \in [K], \tag{8}$$

then for $M > 0$ large enough, (7) holds.

- ***Case 2 (Shifted ReLU with θ_0 unknown):*** *$\phi_k(x;\theta_k^0) = \theta_k^0 + (x)_+$, $\theta_k^0 > 0$,* ***for all*** *$k \in [K]$. Let Π_θ be a prior distribution on $\Theta = \{\theta = (\theta_k)_k;\ \theta_k > 0\}$. If the Assumptions **(A0)**, **(A1)** and **(A2)** of Theorem 3.2 are satisfied when replacing $B_\infty(\epsilon_T)$ by $B_\infty(\epsilon_T) \cap B_\infty^\Theta(\epsilon_T)$ for T large enough, and if (4) is verified, then for $M > 0$ large enough,*

$$\mathbb{E}_0\left[\Pi(\|f - f_0\|_1 + \|\theta - \theta_0\|_1 > M\sqrt{\kappa'_T}\epsilon_T \big| N)\right] = o(1).$$

***Remark* 3.6.** In Case 1 of Proposition 3.5 only, $B_\infty(\epsilon_T)$ cannot be replaced by $B_2(\epsilon_T)$ in assumption **(A0)**. This is due to the fact that we need to consider parameters f such that the likelihood at f is positive (i.e., $\exp(L_T(f)) > 0$) in order to control the Kullback-Leibler divergence (see Lemmas S6.1, S6.3 and A.2). In this argument, we also need the additional assumption (8). The latter is a non trivial condition on the intensity of the true model, which we do not expect to hold in many situations. For instance it does not hold if $\tilde{\lambda}_t(f_0)$ is Lipschitz in a neighbourhood of t such that $\tilde{\lambda}_t(f_0) = 0$. We expect that this can happen with significant probability as soon as one interaction functions h_{lk}^0 is Lipschitz and h_{lk}^{0-} is non-null. It is however not clear if this condition is sharp, i.e., if Bayesian or other likelihood-based methods would be suboptimal without this assumption (from our construction of tests, it is easy to construct frequentist estimates of f which converge at the rate $\sqrt{\epsilon_T}$ defined by the testing condition in [28]). This also motivates the study of the shifted ReLU model, as an alternative of interest for modelling positive intensity functions. Nonetheless, in Lemma 4.3, we provide sufficient conditions in a finite-histogram model so that (8) holds. Finally, we note that using Theorem 1.2 of [11] and notation τ_1, τ_2 for the regeneration times defined in Lemma 5.1, (8) is equivalent to $\mathbb{E}_0\left(\int_{\tau_1}^{\tau_2} \frac{\mathbb{1}_{\lambda_t^k(f_0)>0}}{\lambda_t^k(f_0)} dt\right) < +\infty$.

***Remark* 3.7.** In Theorem 3.2, we in fact obtain the posterior concentration rate on $((\phi_k(\nu_k))_k, h)$, i.e.,

$$\mathbb{E}_0\left[\Pi\left(\sum_k |\phi_k(\nu_k) - \phi_k(\nu_k^0)| + \|h - h_0\|_1 > M\sqrt{\kappa_T}\epsilon_T \big| N\right)\right] = o(1),$$

for M a large enough constant. Moreover, if the ϕ_k's are partially known of the form $\phi_k(x;\theta_k) = \theta_k + \psi(x)$ where $\theta_k \geqslant 0$ and ψ is given, then we obtain

$$\mathbb{E}_0\left[\Pi(\|h - h_0\|_1 + \sum_k |\theta_k + \psi(\nu_k) - \theta_k^0 - \psi(\nu_k^0)| > M\sqrt{\kappa_T}\epsilon_T \big| N)\right] = o(1).$$

In the next corollary, we deduce from the previous results the convergence rate of the posterior means

$$(\hat{\nu}, \hat{h}) = \mathbb{E}^\Pi[f|N] = \int_{\mathcal{F}} f d\Pi(f|N), \quad \text{and} \quad \hat{\theta} = \mathbb{E}^\Pi[(\theta)|N] \quad \text{when } \theta_0 \text{ is unknown (in the shifted ReLU model).}$$

Corollary 3.8. *Under the assumptions of Theorem 3.2 or Case 1 of Proposition 3.5, if $\int_{\mathcal{F}} \|f\|_1 d\Pi(f) < +\infty$, then for $M > 0$ large enough, it holds that*

$$\mathbb{P}_0\left[\|\hat{\nu} - \nu_0\|_1 + \|\hat{h} - h_0\|_1 > M\sqrt{\kappa_T}\epsilon_T\right] = o(1).$$

Under the assumptions of Case 2 of Proposition 3.5, we have

$$\mathbb{P}_0\left[\|\hat{\nu} - \nu_0\|_1 + \|\hat{h} - h_0\|_1 + \|\hat{\theta} - \theta_0\|_1 > M\sqrt{\kappa_T}\epsilon_T\right] = o(1).$$

The proofs of Theorem 3.2 and Proposition 3.5 are given in Sections 5.2 and 5.3, and the proof of Corollary 3.8 is reported in Section S3 in the supplementary material [55].

3.2. Consistency on the connectivity graph

In this section, we state our consistency results on the connectivity or Granger causality graph $\delta \in \{0,1\}^{K^2}$, which characterises the fact that interaction functions between pairs of dimensions are null or not, i.e., $\delta_{lk} = 0 \iff h_{lk} = 0$, $(l,k) \in [K]^2$. We note that the definition of Granger causality graph for the linear Hawkes model (see for instance Definition 3.3 in [20]) also holds for the nonlinear model. This leads us to consider the following hierarchical spike-and-slab prior structure. Writing $h_{lk} = \delta_{lk} h_{lk} = \delta_{lk} S_{lk} \bar{h}_{lk}$, with $S_{lk} = \|h_{lk}\|_1$ and $\bar{h}_{lk}$ such that $\left\|\bar{h}_{lk}\right\|_1 = 1$, we define a family of priors:

$$\begin{aligned} \delta \sim \pi_\delta, \quad \mathcal{I}(\delta) = \{(l,k) \in [K]^2;\, \delta_{lk} = 1\}, \\ (h_{lk},\, (l,k) \in \mathcal{I}(\delta)) | \delta \sim \Pi_{h|\delta}(\cdot|\delta) \quad \text{and} \quad \forall (l,k) \notin \mathcal{I}(\delta),\, h_{lk} = 0, \end{aligned} \tag{9}$$

with π_δ a probability distribution on $\{0,1\}^{K^2}$. We can either determine $\Pi_{h|\delta}$ as a distribution on the set of $(h_{lk},\, (l,k) \in \mathcal{I}(\delta))$ and obtain the marginal distribution of $S = (S_{lk})_{lk}$, or construct it as in [17] - see also the prior construction in Section 4. Adapting **(A0)** to the above structure, we recall that δ_0 corresponds to the true connectivity parameter and we consider the following assumption

$$\textbf{(A0')} \quad \Pi(B_\infty(\epsilon_T)|\delta = \delta_0) \geqslant e^{-c_1 T \epsilon_T^2/2}, \quad \pi_\delta(\delta = \delta_0) \geqslant e^{-c_1 T \epsilon_T^2/2}.$$

For instance, one can choose $\pi_\delta = \mathcal{B}(p)^{K^2}$ with $0 < p < 1$, implying that the δ_{lk}'s are i.i.d. Bernoulli random variables. Then for any fixed p, **(A0')** is verified as soon as $\Pi_{h|\delta}(B_\infty(\epsilon_T)|\delta = \delta_0) \geqslant e^{-c_1 T \epsilon_T^2/2}$ holds. This formalism allows us to consider the posterior distribution of δ which is a key object to infer the connectivity graph. The next theorem is our posterior consistency result, which is a consequence of Theorem 3.2 and Proposition 3.5 and holds for all previously considered link functions ϕ.

Theorem 3.9. *Let N be a Hawkes process with function $\phi = (\phi_k)_k$ and parameter $f_0 = (\nu_0, h_0)$, $\epsilon_T = o(1/\sqrt{\kappa_T})$ be a positive sequence and Π be a prior distribution on $\mathcal{F}$ satisfying the conditions of Theorem 3.2 or Proposition 3.5 (replacing **(A0)** by **(A0')**). Then,*

$$\mathbb{E}_0\left[\Pi(\delta_{lk} \neq \delta^0_{lk},\, \forall (l,k) \in \mathcal{I}(\delta_0)|N)\right] = o(1), \quad I(\delta_0) = \{(l,k) \in [K]^2;\, \delta^0_{lk} = 1\}. \tag{10}$$

If in addition the following holds

$$\forall \delta \in \{0,1\}^{K^2},\, \forall C > 0,\, \forall (l,k) \in \mathcal{I}(\delta) \cap \mathcal{I}(\delta_0)^c,\, \Pi_{h|\delta}\left(S_{lk} \leqslant C\epsilon_T\,|\delta\right) = o\left(e^{-(\kappa_T + c_1)T\epsilon_T^2}\right), \tag{11}$$

*with $c_1 > 0$ defined in **(A0')**, then $\mathbb{E}_0\left[\Pi(\delta \neq \delta_0|N)\right] = o(1)$.*

The first part of Theorem 3.9 in (10) is directly obtained from Theorem 3.2 or Proposition 3.5 (Cases 1 and 2) and says that the posterior probability of $\delta_{lk} = 1$ converges to 1, if the edge $l \to k$ is in $I(\delta_0)$, i.e., $\delta^0_{lk} = 1$. The second and more difficult part of Theorem 3.9 is to infer a non-edge $\delta^0_{lk} = 0$. The condition (11) forces the conditional prior distribution $\Pi_{h|\delta}$ to be exponentially small around 0 for all h_{lk} such that $\delta_{lk} = 1$. We note that it also implies that if $h^0_{lk} \neq 0$ and is small, then it may not be detected nor estimated properly. In Section 4, we present two common families of priors on the S_{lk}'s that verify (11).

Interestingly, if the model is more constrained, a much weaker condition on the prior distribution on S_{lk} is required which avoids this issue on the estimation of small "signals" h^0_{lk}. We now consider two restricted Hawkes models, where the interaction functions are either all the same, or only depend on

the "receiver" node. For simplicity of exposition, we consider the case of fully known link functions satisfying the assumptions of Theorem 3.2, however our next proposition remains valid for the ReLU and shifted ReLU models under the assumptions of Proposition 3.5.

- **All-equal model:** we assume that $\forall (l,k) \in [K]^2$, $h_{lk} = \delta_{lk}\tilde{h}$, with $\tilde{h} \in \mathcal{H}'$ so that $\mathcal{F} = \{f = (\nu, \delta, \tilde{h}) \in \mathbb{R}_+ \backslash \{0\}^K \times \{0,1\}^{K^2} \times \mathcal{H}'$; (f,ϕ) satisfy **(C1bis)** or **(C2)** and Assumption 3.1$\}$. When $\delta \neq 0$, then $\tilde{h} \sim \Pi_{\tilde{h}}$ is a probability distribution on $\mathcal{H}' \cap \{\tilde{h} \neq 0\}$.
- **Receiver node dependent model:** we assume that $\forall (l,k) \in [K]^2, h_{lk} = \delta_{lk} h_k$ with $h_k \in \mathcal{H}'$, so that $\mathcal{F} = \{f = (\nu, \delta, (h_k)_k)$; $h_k \in \mathcal{H}'$, $\forall k$, (f,ϕ) satisfy **(C1bis)** or **(C2)** and Assumption 3.1$\}$. We also assume that the prior distribution Π can be written as a product of priors $(\Pi_k)_k$ where for each k, Π_k is a distribution on $(\nu_k, h_k, \delta_{lk}, l \in [K])$, restricted to $\mathcal{F}$. We denote $\delta_{\cdot k} = (\delta_{lk}, l \in [K])$.

Proposition 3.10. *We consider a restricted Hawkes model either defined above as the **All-equal model** or as the **Receiver node dependent model**. Let N be a Hawkes process with function $\phi = (\phi_k)_k$ and parameter $f_0 = (\nu_0, h_0)$ and let Π be a prior distribution on $\mathcal{F}$ such that the prior on ν has positive and continuous density wrt the Lebesgue measure. We also assume that there exists $0 < p_1 < 1/2$ such that for any $(l,k) \in [K]^2$, $p_1 \leqslant \Pi(\delta_{lk} = 1) \leqslant 1 - p_1$.*

- *In the **All-equal model**:*
  1. *If there exists $(l,k) \in [K]^2$ such that $\delta^0_{lk} \neq 0$, then if $\Pi_{\tilde{h}}(h_0 \leqslant \tilde{h} \leqslant h_0 + \epsilon_T) \geqslant e^{-c_1 T \epsilon_T^2/2}$ and if **(A1)**, **(A2)** hold, then $\mathbb{E}_0[\Pi(\delta \neq \delta_0 | N)] = o(1)$.*
  2. *If $\delta_0 = 0$, then if there exists $\mathcal{H}_T \subset \mathcal{H}$ such that for all $\delta \neq 0$, $\Pi_{h|\delta}(\mathcal{H}_T^c | \delta) = o(T^{-K/2})$, if **(A2)** holds with $\epsilon_T = \sqrt{\log T/T}$, and if*

$$\forall C > 0,\ \Pi_{\tilde{h}}\left(0 < \|\tilde{h}\|_1 \leqslant C\sqrt{\log T/T}\right) = o((\log T)^{-K/2}), \tag{12}$$

  then $\mathbb{E}_0[\Pi(\delta \neq 0 | N)] = o(1)$.
- *In the **Receiver node dependent model**: under **(A0')**, **(A1)**, **(A2)**, for any $k \in [K]$,*
  1. *If there exists $l \in [K]$ such that $\delta^0_{lk} \neq 0$, then $\mathbb{E}_0\left[\Pi(\delta_{k_1 k} \neq \delta^0_{k_1 k} | N)\right] = o(1)$, $\forall k_1 \in [K]$.*
  2. *If $\delta^0_{\cdot k} = 0$, if there exists $\tilde{\mathcal{H}}_T \subset \mathcal{H}_1$ such that $\Pi_k(\tilde{\mathcal{H}}_T^{\ c}) = o(T^{-K/2})$, and if for $M > 0$ large enough and $x_0 > 0$, $\zeta_0 > 0$,*

$$\mathcal{N}\left(\zeta_0 M\sqrt{\log T/T}, \tilde{\mathcal{H}}_T, \|.\|_1\right) \leqslant T^{x_0 M},$$

  and if (12) holds with h_k instead of $\tilde{h}$, then $\mathbb{E}_0\left[\Pi(\delta_{\cdot k} \neq \delta^0_{\cdot k} | N)\right] = o(1)$.

Consequently, in those restricted Hawkes models, the above proposition states that the posterior distribution is consistent at δ_0 under the much weaker assumption(12) on the prior compared to (11) of Theorem 3.9. In fact, in the **All-equal model** (resp. the **Receiver node dependent model**), if the true graph has no edge (resp. no edge arriving on node k), then the posterior distribution on h (resp. h_k) concentrates at the paranetric rate $\sqrt{\log T/T}$. This gives a sharp lower bound on the marginal density of N, i.e., on the denominator D_T in (5). We note that (12) is a mild condition which is verified in particular when the prior distribution on $\tilde{S} = \|\tilde{h}\|_1$ (resp. $S_k = \|h_k\|_1$) conditionally on $\tilde{S} \neq 0$ (resp. $S_k \neq 0$) has a density wrt the Lebesgue measure bounded by $\tilde{S}^{-a}$ (resp. S_k^{-a}) with $a > 0$ near 0.

We now study the consistency of Bayesian estimators of the connectivity graph. From Theorem 3.9 or Proposition 3.10, we can directly obtain that the graph estimator based on the $0-1$ loss function defined as $\hat{\delta}^{\Pi}_{lk}(N) = 1 \iff \Pi(\delta_{lk} = 1|N) > \Pi(\delta_{lk} = 0|N)$, is consistent, i.e., $\mathbb{P}_0\left[\hat{\delta}^{\Pi}(N) \neq \delta_0\right] = o(1)$.

This result is obtained with the prior condition (11) in the non-restricted model, which as previously explained can deteriorate the inference of small and non-null interaction functions. We thus propose an alternative graph estimator based on a loss function penalising small signals, which therefore allows us to use prior distributions which do not verify (11). For any graph estimator $\hat{\delta} = (\hat{\delta}_{lk})_{l,k} \in \{0,1\}^{K^2}$ and parameter $f = (\nu, h, \delta) \in \mathcal{F}$, we define

$$L(\hat{\delta}, f) = \sum_{l,k=1}^{K} \mathbb{1}_{\hat{\delta}_{lk}=0}\mathbb{1}_{\delta_{lk}=1} + \mathbb{1}_{\hat{\delta}_{lk}=1}(\mathbb{1}_{\delta_{lk}=0} + \mathbb{1}_{\delta_{lk}=1}F(\|h_{lk}\|_1)),$$

with $F : \mathbb{R}^+ \to [0,1]$ a monotone non-increasing function, with $F(0) = 1$. For a prior Π, the risk of the estimator $\hat{\delta}$ is defined as

$$r(\hat{\delta}, \Pi|N) = \int_{\mathcal{F}} L(\hat{\delta}, f)d\Pi(f|N) = \sum_{l,k} \mathbb{1}_{\hat{\delta}_{lk}=0}\Pi(\delta_{lk} = 1|N) + \mathbb{1}_{\hat{\delta}_{lk}=1}\left[\Pi(\delta_{lk} = 0|N) + \mathbb{E}^{\Pi}(\mathbb{1}_{\delta_{lk}=1}F(\|h_{lk}\|_1)|N)\right].$$

Then the associated risk-minimising estimator, $\hat{\delta}^{\Pi,L}(N) = \arg\min_{\delta\in\{0,1\}^{K^2}} r(\delta, \Pi|N)$, verifies

$$\hat{\delta}^{\Pi,L}_{lk}(N) = 1 \iff \mathbb{E}^{\Pi}[(1 - F(\|h_{lk}\|_1))\mathbb{1}_{\delta_{lk}=1}|N] \geqslant \Pi(\delta_{lk} = 0|N). \tag{13}$$

In the next theorem, we prove that our estimator $\hat{\delta}^{\Pi,L}(N)$ is consistent under the true model $\mathbb{P}_0$ if the penalisation function F satisfies an exponential condition.

Theorem 3.11. *Let N be a Hawkes process with function $\phi = (\phi_k)_k$ and parameter $f_0 = (\nu_0, h_0)$, $\epsilon_T = o(1/\sqrt{\kappa_T})$ be a positive sequence and Π be a prior distribution on $\mathcal{F}$ satisfying the conditions of Theorem 3.2 or Proposition 3.5 (replacing **(A0)** by **(A0')**). Then, if there exists $a > 0$ such that*

$$0 \leqslant 1 - F(M\sqrt{\kappa_T}\epsilon_T) \leqslant e^{-(c_1+a+\kappa_T)T\epsilon_T^2}, \tag{14}$$

for T large enough and with $M > 0$ defined in Theorem 3.2, then for any $(l,k) \in \mathcal{I}(\delta_0)$ such that $1 - F(\|h^0_{lk}\|_1) \geqslant 2e^{-(\kappa_T+c_1)T\epsilon_T^2}$, we have $\mathbb{P}_0\left[\hat{\delta}^{\Pi,L}(N) \neq \delta_0\right] = o(1)$.

***Remark* 3.12.** The assumption on the penalisation function (14) is verified in particular if (i) F is truncated, i.e., $F(x) = \mathbb{1}_{[0,\epsilon]}(x)$ for some (arbitrarily small) $\epsilon > 0$, or if (ii) F is exponentially decreasing around 0, i.e., $F(x) = 1 - \exp\{-\frac{1}{x^p}\}$ with $p > 1/\beta$ if $\epsilon_T = T^{-\beta/2\beta+1}(\log T)^q$ for some $q \geqslant 0$ (see Corollary 4.2 for instance). We note that the choice of penalisation function F determines the detection level of our risk-minimising graph estimator for "small signals". With (i), we will detect "signals" $\|h^0_{lk}\|_1 > \epsilon$ and with (ii), we can detect $\|h^0_{lk}\|_1 > T^{-(p(2\beta+1))^{-1}}$. We also note that this assumption is related to (11), however, since it applies on the penalisation function F and not on the prior distribution, it does not alter the posterior distribution, thus the estimation of ν_0 and h_0.

The proofs of Theorem 3.9, Proposition 3.10 and Theorem 3.11 can be found respectively in Section 5.4, Section S2.2 in the supplementary material [55] and Section 5.5.

4. Prior models

In this section, we construct prior distributions Π that satisfy the assumptions of our main results stated in Section 3 and obtain explicit posterior concentration rates for Hölder-smooth classes of interaction

functions. For ease of exposition, we consider link functions ϕ_k's injective on (m_k, M_k), with $m_k, M_k \in \mathbb{R} \cup \{-\infty, +\infty\}$.

First, we consider a prior on $\nu = (\nu_k)_k$ of the form: $\nu_k \overset{i.i.d.}{\sim} \pi_\nu(\nu_k|(h_{lk})_{l\in[K]})) \propto \pi_\nu(\nu_k)\mathbb{1}_{(m_k,M_k)}(\nu_k)$ with π_ν a positive and continuous probability density on $(0, +\infty)$. To verify **(A1)**, we can for instance choose π_ν such that $\pi_\nu(\nu_k > x) \leqslant x^{-a}$ with $a > 1$. Then it is enough to choose c_2 such that $c_2 > (\kappa_T + c_1)/a$. Moreover in Case 2 of Proposition 3.5 (i.e., shifted ReLU with unknown shift θ_0), we consider a prior on θ such that $\theta_k \overset{i.i.d.}{\sim} \pi_\theta$ with π_θ a density wrt the Lebesgue measure on $(0, +\infty)$.

For the prior on h, we consider the hierarchical structure (9) introduced in Section 3.2 and for the sake of simplicity we assume that $\delta_{lk} \overset{i.i.d.}{\sim} \mathcal{B}(p)$, $\forall (l,k) \in [K]^2$, $p \in (0,1)$, although as previously mentioned, more general priors on δ could be considered. We recall that $I(\delta) = \{(l,k) \in [K]^2;\ \delta_{lk} = 1\}$. We then consider two parametrisation setups. In the first one, $h = (h_{lk},\ (l,k) \in I(\delta))$ is drawn from a truncated distribution of the form

$$d\Pi_h(h|\delta) \propto d\Pi_h^{\otimes|I(\delta)|}(h)\mathbb{1}_{\|S^+\|<1}(h), \tag{15}$$

or simply $d\Pi_h(h|\delta) \propto d\Pi_h^{\otimes|I(\delta)|}(h)$ in the case of a bounded link function (condition **(C2)**), where Π_h is a prior distribution on one function. In the second parametrisation setup,

$$h_{lk} = S_{lk}\bar{h}_{lk}, \quad \left\|\bar{h}_{lk}\right\|_1 = 1, \quad [\bar{h}_{lk}|(l,k) \in I(\delta)] \overset{i.i.d.}{\sim} \Pi_{\bar{h}}, \quad S|\delta \sim \Pi_{S|\delta}, \tag{16}$$

with $\Pi_{\bar{h}}$ is a prior distribution on one L_1-normalised function and $\Pi_{S|\delta}$ is a prior distribution on matrices with non-zero entries δ and, under **(C1bis)**, satisfying $\left\|S^+\right\| < 1$.

Examples of the parametrisation setup (15) are Gaussian processes (or hierarchical Gaussian processes) priors, and prior distributions based on an expansion on some basis, such as Legendre, Fourier, wavelets, splines, etc. As mentioned earlier, the prior assumptions **(A0)-(A2)** are very common in the literature, which allows to directly apply existing results, as we illustrate on spline priors in Section 4.1. In [17], a similar construction is provided using a mixture of Betas distributions in the linear Hawkes model, which leads to the minimax rate of assumption up to a logarithmic factor. We report this construction in the nonlinear model in Section S4.1 in the supplementary material [55] and obtain the same estimation rate up to logarithmic terms. The difficulty in this parametrisation might be to prove condition (11) in Theorem 3.9 for estimating the connectivity graph. In Section 4.2, we illustrate the second parametrisation setup (16) with random histogram priors, which is a setup where condition (11) can be more easily verified. We also consider a prior based on mixtures of Beta distributions in the supplementary material [55]. We denote $\mathcal{H}(\beta, L_0)$ the class of β-smooth functions with radius L_0.

4.1. Spline priors for Π_h

A nonparametric prior Π_h satisfying the assumptions of Theorem 3.2 can be constructed using the family of splines or free knot splines. Without loss of generality, we assume that $A = 1$. For $J \geqslant 1$, let $t_0 = 0 < t_1 < \cdots < t_J = 1$ define a partition of $[0,1]$ and $I_j = (t_{j-1}, t_j)$, $j \in [J]$. We consider splines of order $q \geqslant 0$, i.e., piecewise polynomial functions (on the partition) of degree q and for $q \geqslant 2$, $q-2$ times continuously differentiable. For a given partition, this defines a vector space of dimension $V = q + J - 1$ (see for instance [54, 27]).

For the sake of simplicity, we present the construction of regular partitions, where $t_j = j/J$, however random partitions can be dealt with following the computations of Section 2.3.1 of [17]. Let $B = (B_1, \cdots, B_V)$ be the B-spline basis of order q, as defined in [27]. Recall that for any $j \in [V]$, B_j has

support included in an interval of length q/J, $B_j \geqslant 0$ and $\sum_j B_j(x) = 1$ for all $x \in [0,1]$. We can then define

$$h_{w,J}(x) = w^T B(x), \quad w \in \mathbb{R}^V, \quad J \sim \mathcal{P}(\lambda),$$

where $\mathcal{P}(\lambda)$ is the Poisson distribution with mean λ, and consider the following hierarchical construction of Π_h

$$w_j \overset{i.i.d.}{\sim} \pi_w, \quad 1 \leqslant j \leqslant V = q + J - 1, \tag{17}$$

with π_w a positive and continuous density on $\mathbb{R}$ satisfying $\pi_w(x) \lesssim e^{-a_1|x|^{a_2}}$ for some $a_1, a_2, \lambda > 0$.

Using Lemma 4.1 of [27], if h_0 is $\mathcal{H}(\beta, L_0)$ for some $\beta \leqslant q$ and $L_0 > 0$, then setting $J_T = J_0(T/\log T)^{1/(2\beta+1)}$, $\epsilon_T = (T/\log T)^{-\beta/(2\beta+1)}$, there exist $w_0 \in \mathbb{R}^{V_T}$, $V_T = q + J_T - 1$ and $C > 0$ such that $\|h_0 - h_{w_0,J_T}\|_\infty \leqslant C\epsilon_T$. Moreover using Lemma 4.2 and Lemma 4.3 of [27], we have $\|w_0\|_\infty \leqslant C_0$, for some C_0, and obtain that $\{w \in \mathbb{R}^{V_T}, \|w - w_0\|_\infty \leqslant \epsilon_T\} \subset B_\infty(\epsilon_T)$, which leads to **(A0)**. Similarly, from Lemma 4.2 of [27], $\|h_{w,J} - h_{w',J}\|_1 \lesssim \|w - w'\|_\infty$ and with $\mathcal{H}_T = \{h_{w,J};\ \|w\|_\infty \leqslant T^{B_0},\ J \leqslant J_1 J_T\}$ for some $B_0 > 0$ and $J_1 > 0$, **(A1)** and **(A2)** are also verified. We finally obtain the following result.

Corollary 4.1. *Let N be a Hawkes process with link functions $\phi = (\phi_k)_k$ and parameter $f_0 = (\nu_0, h_0)$ such that (ϕ, f_0) verify the conditions of Lemma 2.1, and Assumption 3.1. Under the above spline prior, if for any $(l,k) \in [K]^2, h^0_{lk} \in \mathcal{H}(\beta, L_0)$ with $\beta \in (0, q+1]$ and $L_0 > 0$, then for $M > 0$ large enough, we have*

$$\mathbb{E}_0\left[\Pi(\|f - f_0\|_1 > M(T/\log T)^{-\beta/(2\beta+1)}(\log T)^{q_0} \big| N)\right] = o(1),$$

where $q_0 = 0$ if ϕ verifies Assumption 3.1(i) and $q_0 = 1/2$ if ϕ verifies Assumption 3.1(ii).

To estimate the connectivity graph δ_0, one can either use the penalised estimator (13), which from the above computations and Corollary 3.11 is consistent, or use the estimator based on the 0-1 loss function if (11) can be verified. In the next section, we consider a prior based on random histograms and illustrate how the latter condition (11) can be satisfied.

4.2. Random histograms prior

Random histograms are a special case of splines with $q = 0$. These piecewise constant functions are of particular interest in the modelling of spike trains emitted by biological neurons, which only interact on certain time periods. We use a similar construction as in Section 2.3.1. of [17], however here the interaction functions are no longer restricted to be non-negative. Using parametrisation (16), the interaction function h_{lk} for $(l,k) \in I(\delta)$ has the form $h_{lk} = S_{lk}\bar{h}_{lk}$ and the $\bar{h}_{lk}$'s are independent and distributed as a random histogram $\bar{h}_{w,\mathbf{t}}$ defined as follows. Given a partition $\mathbf{t} : 0 = t_0 < t_1 < \cdots < t_J = 1$, we define

$$\bar{h}_{w,\mathbf{t}}(x) = \sum_{j=0}^{J-1} \frac{w_j}{t_{j+1} - t_j} \mathbb{1}_{(t_{j-1}, t_j]}, \quad \sum_{j=0}^{J-1} |w_j| = 1, \quad J \sim \mathcal{P}(\lambda),\ \lambda > 0.$$

Similarly to [17], the prior on $(|w_1|, \cdots, |w_J|)$ is constructed by first selecting the non-zero coefficients w_j's, then defining a Dirichlet prior on the vector of non-zero $|w_j|$'s, and finally sampling the sign of

the w_j's. Hence,

$$\forall j \in [J],\ w_j = Z_j u_j, \quad Z_j \in \{-1, 0, 1\}, \quad u_j \geqslant 0, \quad \sum_{j=1}^{J} u_j = 1,$$

and $u_j = 0$ if $Z_j = 0$. We can consider $Z_j \overset{i.i.d.}{\sim}$ Multinomial(p_{-1}, p_0, p_1), with $p_{-1} + p_0 + p_1 = 1$, and given $(Z_1, \cdots, Z_J)$, $(u_{i_1}, \cdots, u_{i_{s_z}}) \sim \mathcal{D}(a_{s_z}, \cdots, a_{s_z})$, $s_z = \sum_j |Z_j|$, where $i_1, \cdots, i_{s_z}$ are the indices of the non zero Z_j's and $\alpha_{-1}, \alpha_0, \alpha_1, a_{s_z} > 0$. Finally if the partition $\mathbf{t}$ is random, we consider a Dirichlet prior $\mathcal{D}(\alpha, \cdots, \alpha)$ on $(t_1, t_2 - t_1, \cdots, 1 - t_{J-1})$. We note that this construction is very similar to Section 2.3.1 of [17], and we therefore obtain the same results as in Corollaries 2 and 3 of [17].

Besides, to estimate the connectivity graph using the 0-1 loss (and to establish our posterior consistency result), we can now verify (11). This condition holds if, with $d\Pi_{S|\delta} = \prod_{(l,k)\in I(\delta)} d\Pi_S(S_{lk}) \mathbb{1}_{\|S^+\|<1}$ (under **(C1bis)**), Π_S has a positive and continuous density π_S on either $[\epsilon, 1]$ if $S^0_{lk} > \epsilon$, or if the density near 0 verifies

$$\pi_S(s^p) \propto s^{-p(\alpha-1)} \exp(-a/s^p) \mathbb{1}_{[0,1]}(s), \quad p > \beta, \quad a > 0.$$

We now present a corollary of Theorem 3.2 in the case of random histograms with random partitions, which is proved as in [17].

Corollary 4.2. *Let N be a Hawkes process with link functions $\phi = (\phi_k)_k$ and parameter $f_0 = (\nu_0, h_0)$ such that (ϕ, f_0) verify Assumption 3.1. Under the above random histogram prior, if for any $(l,k) \in [K]^2, h^0_{lk} \in \mathcal{H}(\beta, L_0)$ with $\beta \in (0,1]$ and $L_0 > 0$, then for M large enough, we have*

$$\mathbb{E}_0\left[\Pi(\|f - f_0\|_1 > M(T/\log T)^{-\beta/(2\beta+1)} (\log T)^q \middle| N)\right] = o(1),$$

where $q = 0$ if ϕ verifies Assumption 3.1(i), and $q = 1/2$ if ϕ verifies Assumption 3.1(ii).

Finally, in the case of the ReLU model (Proposition 3.5), we can also verify (8), in special case of the true parameter $f_0 = (\nu_0, h_0)$ where each h^0_{lk} lie in the space of finite histograms.

Lemma 4.3. *Let N be a nonlinear Hawkes process with parameter $f_0 = (\nu_0, h_0)$ and ReLU link functions $\phi_k(x) = (x)_+, \forall k$, satisfying Assumption 2.2 (and condition **(C1bis)**). If for all $(l,k) \in [K]^2$, there exists $J_0 \in \mathbb{N}^*$ such that $h^0_{lk}(t) = \sum_{j=1}^{J_0} \omega^{lk}_{j0} \mathbb{1}_{I_j}(t)$, with $\{I_j\}_{j=1}^{J_0}$ a partition of $[0,1]$ and $\forall j \in [J_0]$, $\omega^{lk}_{j0} \in \mathbb{Q}$, then (8) holds.*

***Remark* 4.4.** In the previous lemma, the condition that the weights $w^{lk}_{j0}, (l,k) \in [K]^2, j \in [J]$ are rational numbers is a technical argument that allows to find a lower bound on $\tilde{\lambda}^k_t(f_0)$ when $\lambda^k_t(f_0) > 0$. This results from a density argument of the linear combinations of the weights, which, under these conditions, constrains $\lambda^k_t(f_0)$ to take values on a lattice. Besides, we note that our result is in fact more general and applies to any model with Lipschitz link functions such that $\min_{x\in\mathbb{R}} \phi_k(x) = 0$.

Lemma 4.3 is proved in Section S4.2 in the supplementary material [55].

5. Proofs

In this section, we report the proofs of our main theorems on the posterior concentration properties (Theorems 3.2 and Proposition 3.5), and on the estimation of the connectivity graph (Theorems 3.9

and 3.11). Instead of using the clustering structure of linear Hawkes processes like in [17] or a coupling technique like in [8], these proofs leverage the renewal properties of nonlinear Hawkes processes notably studied by Costa et al. in [11]. The novelty of our proofs lies in the selection of parts or special "excursions", that allow us to estimate the parameter at a rate equivalent to the one for a linear Hawkes process. In the following section, we first recall the definitions of the concept of excursions and some properties of the process' renewal times.

5.1. Renewal times and excursions

In the following lemma, we introduce the concept of *excursions* for stationary nonlinear Hawkes processes verifying the conditions of Lemma 2.1. This result extends the ones of Costa et al. in [11] to the multivariate case under condition **(C1bis)** of Lemma 2.1 and to bounded models (condition **(C2)**).

Lemma 5.1. *Let N be a Hawkes process with monotone non-decreasing and Lipschitz link functions $\phi = (\phi_k)_k$ and parameter $f = (\nu, h)$ such that (ϕ, f) verify **(C1bis)** or **(C2)**. Then the point process measure $X_t(.)$ defined as*

$$X_t(.) = N|_{(t-A,t]}, \tag{18}$$

is a strong Markov process with positive recurrent state $\emptyset$. Let $\{\tau_j\}_{j\geqslant 0}$ be the sequence of random times defined as

$$\tau_j = \begin{cases} 0 & \text{if } j = 0; \\ \inf\left\{t > \tau_{j-1};\ X_{t^-} \neq \emptyset,\ X_t = \emptyset\right\} = \inf\left\{t > \tau_{j-1};\ N|_{[t-A,t)} \neq \emptyset,\ N|_{(t-A,t]} = \emptyset\right\} & \text{if } j \geqslant 1. \end{cases}$$

Then, $\{\tau_j\}_{j\geqslant 0}$ are stopping times for the process N. For $T > 0$, we also define

$$J_T = \max\{j \geqslant 0;\ \tau_j \leqslant T\}. \tag{19}$$

The intervals $\{[\tau_j, \tau_{j+1})\}_{j=0}^{J_T-1} \cup [\tau_{J_T}, T]$ form a partition of $[0,T]$. The point process measures $(N|_{[\tau_j,\tau_{j+1})})_{1\leqslant j\leqslant J_T-1}$ are i.i.d. and independent of $N|_{[0,\tau_1)}$ and $N|_{[\tau_{J_T},T]}$; they are called excursions *and the stopping times $\{\tau_j\}_{j\geqslant 1}$ are called* regenerative *or* renewal *times.*

The proof of the previous lemma is omitted since it is a fairly direct multivariate extension of some elements of Proposition 3.1, Proposition 3.4, Theorem 3.5 and Theorem 3.6 in [11], recalled in Section S9 in the supplementary material [55]. For the extension to bounded models, we use a direct consequence of the results in Costa et al. [11] that if N is dominated by a homogeneous Poisson point process, then it also have the regenerative properties of Lemma 5.1. We also note that since A is known, the renewal times τ_j's are observable. In the rest of this article, we denote

$$\Delta\tau_1 = \tau_2 - \tau_1, \tag{20}$$

the length of a generic excursion. For any link functions ϕ_k's and parameter $f = (\nu, h)$, we denote r_f the value of the intensity process at the beginning of each excursion, defined as

$$r_f = (r_1^f, \dots, r_K^f), \quad r_k^f = \phi_k(\nu_k), \quad k \in [K]. \tag{21}$$

In the next two lemmas, we prove some useful results on the distributions of $\Delta\tau_1$, on the number of points in a generic excursion $N[\tau_1, \tau_2)$ and on the number of excursions in the observation window $[-A, T]$, J_T, defined in (19).

Lemma 5.2. *Under the assumptions of Lemma 5.1, the random variables $\Delta\tau_1$ and $N[\tau_1,\tau_2)$ admit exponential moments. More precisely, under condition **(C1bis)**, with $m = \|S^+\| < 1$, we have*

$$\forall s < \min(\|r_f\|_1, \gamma/A), \quad \mathbb{E}_f\left[e^{s\Delta\tau_1}\right] \leqslant \frac{1+m}{2m}, \quad \textit{and} \quad \mathbb{E}_f\left[e^{sN[\tau_1,\tau_2)}\right] < +\infty, \quad \gamma = \frac{1-m}{2\sqrt{K}} \log\left(\frac{1+m}{2m}\right).$$

*Under condition **(C2)**, we have $\forall s < \min_k \Lambda_k$), $\mathbb{E}_f\left[e^{s\Delta\tau_1}\right] \leqslant \frac{\|\Lambda\|_1^2}{(\min_k \Lambda_k - s)^2}$ and $\mathbb{E}_f\left[e^{sN[\tau_1,\tau_2)}\right] < +\infty$. In particular, this implies that $\mathbb{E}_f\left[N[\tau_1,\tau_2) + N[\tau_1,\tau_2)^2\right] < +\infty$.*

***Remark* 5.3.** The previous lemma provides exponential moments of $\Delta\tau_1$ and $N[\tau_1,\tau_2)$, under the assumption that $\|S^+\| < 1$ **(C1bis)**, but we conjecture that results of Lemma 5.2 still holds under the more general conditions $r(S^+) < 1$ **(C1)** of Lemma 2.1.

Lemma 5.4. *Under the assumptions of Lemma 5.1, for any $\beta > 0$, there exists a constant $c_\beta > 0$ such that $\mathbb{P}_f\left[J_T \notin [J_{T,\beta,1}, J_{T,\beta,2}]\right] \leqslant T^{-\beta}$, with J_T defined in (19) and*

$$J_{T,\beta,1} = \left\lfloor \frac{T}{\mathbb{E}_f[\Delta\tau_1]}\left(1 - c_\beta\sqrt{\frac{\log T}{T}}\right)\right\rfloor, \quad J_{T,\beta,2} = \left\lfloor \frac{T}{\mathbb{E}_f[\Delta\tau_1]}\left(1 + c_\beta\sqrt{\frac{\log T}{T}}\right)\right\rfloor.$$

The proofs of Lemmas 5.2 and 5.4 are reported in Section S7.2 in the supplementary material [55].

5.2. Proof of Theorem 3.2 and Case 1 of Proposition 3.5

In this section, we prove our main posterior concentration theorem, Theorem 3.2, as well as Case 1 of Proposition 3.5, which deals with the specific case of the standard ReLU model. The first step of this proof borrows some ideas from the one of Theorem 3 in [17], but also introduces novel elements built from the renewal properties of the process. In particular, the posterior concentration is first proved in terms of a particular distance on the intensity process (see Proposition 5.5 below), which in fact corresponds to a stochastic (pseudo) distance on the parameter space $\mathcal{F}$. This stochastic distance $\tilde{d}_{1T}$ resembles the L_1 stochastic distance used in [17], except that it is restricted to a subset of the observation window $[-A,T]$ which only contains the beginning of each excursion. More precisely for any excursion index $j \in [J_T - 1]$, we denote $(U_j^{(1)}, U_j^{(2)})$ the times of the first two events after the j-th renewal time τ_j (as defined in Lemma 5.1). We note that by definition, $U_j^{(1)} \in [\tau_j, \tau_{j+1})$, $U_j^{(2)} \in [\tau_j, \tau_{j+2}]$ and $\tau_{j+1} \geqslant U_j^{(1)} + A$. We then define our restricted observation window $A_2(T)$ as

$$A_2(T) := \bigcup_{j=1}^{J_T-1} [\tau_j, \xi_j], \tag{22}$$

with $\xi_j := U_j^{(2)}$ if $U_j^{(2)} \in [\tau_j, \tau_{j+1})$ and $\xi_j := \tau_{j+1}$ otherwise. We note that the interval $[\tau_j, \xi_j]$ corresponds either to the beginning of the j-th excursion or to the whole excursion $[\tau_j, \tau_{j+1})$ when the latter contains only one event, implying that $U_j^{(2)} \geqslant \tau_{j+1}$. Moreover, since the renewal times (and J_T) are observable, so is $A_2(T)$.

The construction of $A_2(T)$ is a novel and essential element of our proof. Informally, it corresponds to a set of intervals where the parameters can be inferred in a similar way as in the linear Hawkes

model and which Lebesgue measure is of order T. More precisely, using the renewal properties from Section 5.1, we will prove, using Lemma A.1, that with probability going to 1, $|A_2(T)| \gtrsim T$ under $\mathbb{P}_0$. We can now define our auxiliary stochastic distance as

$$\tilde{d}_{1T}(f, f') = \frac{1}{T} \sum_{k=1}^{K} \int_0^T \mathbb{1}_{A_2(T)}(t) |\lambda_t^k(f) - \lambda_t^k(f')| dt, \tag{23}$$

and state our intermediate posterior concentration rate result, which holds for all models satisfying the conditions of Theorem 3.2 and the ReLU-type models considered in Proposition 3.5.

Proposition 5.5. *Under the assumptions of Theorem 3.2 or Proposition 3.5, for $M'_T = M' \sqrt{\kappa_T}$ with $M' > 0$ a large enough constant,*

$$\mathbb{E}_0 \left[\Pi(\tilde{d}_{1T}(f, f_0) > M'_T \epsilon_T | N) \right] = o(1).$$

The proof of the previous proposition follows the strategy of [17] in Theorem 1, which is based on the now well-known argument by [28]. However, we note that in our setting, this strategy can be applied thanks to the definition of the stochastic distance which restricts the observation window to the set $A_2(T)$. We recall here its main steps. First, we restrict the space of probability events to a subset $\tilde{\Omega}_T$ that has high probability (see below and Lemma A.1). Secondly, we prove a lower bound of the denominator D_T defined in (5), derived from the technical Lemma A.2. Thirdly, we consider a ball centered at the true parameter f_0 of radius $M'_T \epsilon_T$ w.r.t. $\tilde{d}_{1T}$, denoted by $A_{d_1}(M'_T \epsilon_T) \subset \mathcal{F}$. Finally, to find an upper bound of the numerator $N_T(A_{d_1}(M'_T \epsilon_T)^c)$ defined in (5), we partition $A_{d_1}(M'_T \epsilon_T)^c$ into slices $\{S_i\}_i$ on which we can design tests that have exponentially decreasing type I and type II errors (see Lemma S5.1). We then define ϕ as the maximum of the tests on the individual slices S_i. Due to the space constraints, this proof is reported in Section S1 of the supplementary material [55].

From Proposition 5.5, we prove Theorem 3.2 and Case 1 of Proposition 3.5 using the following classical decomposition (see for instance the proof of Theorem 1 in [17]). Let $A, B \in \mathcal{F}_T \subset \mathcal{F}$, with B possibly data dependent, $\phi \in [0, 1]$ be a measurable test, κ_T defined in (6), and $\tilde{\Omega}_T \subset \Omega$. Then,

$$\begin{aligned} \mathbb{E}_0 \left[\Pi(A \cap B | N) \right] \leqslant \mathbb{P}_0 \left[\{ D_T < e^{-(\kappa_T + c_1) T \epsilon_T^2} \} \cap \tilde{\Omega}_T \right] + \mathbb{E}_0 \left[\phi \mathbb{1}_{\tilde{\Omega}_T} \right] + \mathbb{P}_0 [\tilde{\Omega}_T^c] \\ + e^{(\kappa_T + c_1) T \epsilon_T^2} \Pi(\mathcal{F}_T^c) + e^{(\kappa_T + c_1) T \epsilon_T^2} \int_{A \cap \mathcal{F}_T} \mathbb{E}_0 \left[\mathbb{E}_f \left[(1 - \phi) \mathbb{1}_B(f) \mathbb{1}_{\tilde{\Omega}_T}(N) \middle| \mathcal{G}_0 \right] \right] d\Pi(f). \end{aligned} \tag{24}$$

We first introduce the set $\tilde{\Omega}_T$, which from Lemma A.1, has probability $\mathbb{P}_0 \left[\tilde{\Omega}_T^c \right]$ going to 0 at any polynomial rate. For $T > 0$, we denote

$$\mathcal{J}_T := \left\{ J \in \mathbb{N}; \left| \frac{J - 1}{T} - \frac{1}{\mathbb{E}_0[\Delta \tau_1]} \right| \leqslant c_\beta \sqrt{\frac{\log T}{T}} \right\},$$

with $c_\beta > 0$ (and $\beta > 0$) chosen in Lemma A.1, and, with $r_0 := r_{f_0} = (r_1^0, \ldots, r_K^0)$ where $r_k^0 = \phi_k(\nu_k^0)$, and $\mu_k^0 = \mathbb{E}_0 \left[\lambda_t^k(f_0) \right]$, for any k,

$$\Omega_N = \left\{ \max_{k \in [K]} \sup_{t \in [0,T]} N^k[t - A, t) \leqslant C_\beta \log T \right\} \cap \left\{ \sum_{k=1}^{K} \left| \frac{N^k[-A, T]}{T} - \mu_k^0 \right| \leqslant \delta_T \right\},$$

$$\Omega_J = \{ J_T \in \mathcal{J}_T \}, \quad \Omega_U = \left\{ \sum_{j=1}^{J_T - 1} (U_j^{(1)} - \tau_j) \geqslant \frac{T}{\mathbb{E}_0[\Delta \tau_1] \| r_0 \|_1} \left(1 - 2 c_\beta \sqrt{\frac{\log T}{T}} \right) \right\},$$

with $\delta_T = \delta_0 \sqrt{\frac{\log T}{T}}$, $\delta_0 > 0$ and $C_\beta > 0$ chosen in Lemma A.1 and define

$$\tilde{\Omega}_T = \Omega_N \cap \Omega_J \cap \Omega_U. \tag{25}$$

The sets Ω_N, Ω_J and Ω_U control respectively the number of events, the number of excursions and the length of excursions. First, Ω_N corresponds to realisations of N such that the number of events in any interval of length A is upper bounded by $c_\beta \log T$, and the number of events on $[-A, T]$ is close to its expectation under the stationary distribution $\mathbb{P}_0$. Secondly, Ω_J corresponds to the realisations such that the number of excursions in the observation interval $[0, T]$ divided by T, J_T/T, is close to its limit $1/\mathbb{E}_0[\Delta\tau_1]$. Thirdly, on Ω_U, the measure of the subset corresponding to the collections of the beginnings of excursions (from τ_j to the first event $U_j^{(1)}$) is of order T.

Next, we bound the denominator of the posterior D_T from (5). From Lemma A.2, together with the lower bound technique of [28], we have that

$$\mathbb{P}_0\left[D_T < \Pi(B_\infty(\epsilon_T))e^{-\kappa_T T\epsilon_T^2}\right] \leqslant 2\int_{B_\infty(\epsilon_T)} \frac{\mathbb{P}_0[L_T(f) - L_T(f_0) < -\kappa_T T\epsilon_T^2/2]}{\Pi(B_\infty(\epsilon_T))} d\Pi(f) = o(1), \tag{26}$$

which leads to $\mathbb{P}_0\left[D_T < e^{-(\kappa_T + c_1)T\epsilon_T^2}\right] = o(1)$ using asssumption (**A0**).

Then, we find a lower bound on $|A_2(T)|$ on $\tilde{\Omega}_T$. We recall that the point process measures $(N|_{[\tau_j,\tau_{j+1})})_{1\leqslant j\leqslant J_T-1}$ are i.i.d. and *a fortiori* that the random variables $\{U_j^{(1)} - \tau_j\}_j$ are i.i.d. Moreover, for any $j \in [J_T - 1]$, $t \in [\tau_j, U_j^{(1)})$ and $k \in [K]$, the intensity process is by construction equal to $\lambda_t^k(f_0) = r_k^0 = \phi_k(\nu_k^0)$. Therefore, conditionally on τ_j, $U_j^{(1)}$ has the same distribution as an event from a Poisson point process beginning at τ_j, with intensity $\|r_0\|_1$, since the process is the superposition of K univariate Poisson process with intensity r_k^0, $k \in [K]$. Thus, under $\mathbb{P}_0$, each variable $U_j^{(1)} - \tau_j$ follows an exponential distribution with mean $1/\|r_0\|_1$, and on Ω_U, for T large enough, we have that

$$|A_2(T)| = \sum_{j=1}^{J_T-1} (\xi_j - \tau_j) \geqslant \sum_{j=1}^{J_T-1} (U_j^{(1)} - \tau_j) \geqslant c_0 T, \quad c_0 := \frac{1}{2\mathbb{E}_0[\Delta\tau_1]\|r_0\|_1}.$$

Finally, for $R > 0$, we define the balls in L_1 and stochastic distances

$$A_{L_1}(R) := \{f \in \mathcal{F};\ \|f - f_0\|_1 \leqslant R\}, \quad A_{d_1}(R) = \{\tilde{d}_{1T}(f, f_0) \leqslant R\}.$$

We now apply the decomposition (24) with $\phi = 1$, $A := A_{L_1}(M_T\epsilon_T)^c$ and $B := A_{d_1}(M'_T\epsilon_T)$, with $M_T = M\sqrt{\kappa_T}$, $M'_T = M'\sqrt{\kappa_T}$, $M > M'$ and M' defined in Theorem 5.5. As in the proof of Theorem 3 of [17], we are thus left to prove that

$$\sup_{A_{L_1}(M_T\epsilon_T)^c \cap \mathcal{F}_T} \mathbb{P}_f\left[\tilde{\Omega}_T \cap A_{d_1}(M'_T\epsilon_T)|\mathcal{G}_0\right] = o_{\mathbb{P}_0}(e^{-(c_1+\kappa_T)T\epsilon_T^2}), \tag{27}$$

with c_1 defined in assumption (**A0**). We recall that $\mathbb{P}_f$ is the process distribution associated to parameter f defined in (3). To prove (27), we consider $f \in A_{L_1}(M_T\epsilon_T)^c$ such that $\tilde{d}_{1T}(f, f_0) \leqslant M'_T\epsilon_T$ and for $l \in [K]$ and $j \in [J_T - 1]$, we define

$$Z_{jl} := \int_{\tau_j}^{\xi_j} |\lambda_t^l(f) - \lambda_t^l(f_0)|dt. \tag{28}$$

We note that using Lemma 5.1, the random variables $\{Z_{jl}\}_{j\in[J_T-1]}$ are i.i.d., and from (23) we also have that $T\tilde{d}_{1T}(f,f_0) > \max_{l\in[K]} \sum_{j=1}^{J_T-1} Z_{jl}$. In order to derive a Bernstein-type inequality on the sum of the Z_j's, we first find an upper bound of Z_{1l} and its moments. Using that the link functions ϕ_k's are L-Lipschitz, we have

$$\begin{aligned} Z_{jl} &= \int_{\tau_j}^{\xi_j} |\phi_k(\tilde{\lambda}_t^l(\nu,h)) - \phi_k(\tilde{\lambda}_t^l(\nu_0,h_0))|dt \leqslant L\int_{\tau_j}^{\xi_j} |\tilde{\lambda}_t^l(\nu,h) - \tilde{\lambda}_t^l(\nu_0,h_0)|dt \\ &\leqslant L(\xi_j-\tau_j)|\nu_l-\nu_l^0| + L\sum_k \int_{U_j^{(1)}}^{\xi_j} |h_{kl}-h_{kl}^0|(t-U_j^{(1)})dt \\ &\leqslant L(A+U_j^{(1)}-\tau_j)|\nu_l-\nu_l^0| + L\sum_k \|h_{kl}-h_{kl}^0\|_1 \leqslant L(A+1+U_j^{(1)}-\tau_j)\,\|f-f_0\|_1\,. \end{aligned} \tag{29}$$

Moreover, under $\mathbb{P}_f$, for any $j\in[J]$, $U_j^{(1)}-\tau_j$ follows an exponential distribution with mean $1/\left\|r_f\right\|_1$, therefore, for any $n\in\mathbb{N}$, $\mathbb{E}_f\left[(U_j^{(1)}-\tau_j)^n\right] = \frac{n!}{\|r_f\|_1^n}$. Using the standard inequality $(x+y)^n \leqslant 2^{n-1}(x^n+y^n)$, we thus obtain that

$$\begin{aligned} \mathbb{E}_f\left[Z_{1l}^n\right] &\leqslant 2^{n-1}L^n\left((A+1)^n + \mathbb{E}_f\left[(U_j^{(1)}-\tau_j)^n\right]\right)\|f-f_0\|_1^n \\ &\leqslant \frac{1}{2}2n!\left(2L\max\left(A+1,\frac{1}{\left\|r_f\right\|_1}\right)\|f-f_0\|_1\right)^{n-2} \times L^2\max\left(A+1,\frac{1}{\left\|r_f\right\|_1}\right)^2\|f-f_0\|_1^2 \leqslant \frac{1}{2}n!b^{n-2}v^2, \end{aligned} \tag{30}$$

with $b := 2L\max\left(A+1,\frac{2}{\|r_0\|_1}\right)\|f-f_0\|_1$ and $v := L\max\left(A+1,\frac{2}{\|r_0\|_1}\right)\|f-f_0\|_1$. In the last inequality, we have used the fact that $\|r_f-r_0\|_1 \lesssim \tilde{d}_{1T}(f,f_0) \leqslant M_T'\epsilon_T$ on $\tilde{\Omega}_T$. This is because $(U_1^{(1)}-\tau_1)+\cdots+(U_{J_T-1}^{(1)}-\tau_{J_T-1}) \geqslant c_0T/2$, which leads to

$$T\tilde{d}_{1T}(f,f_0) \geqslant \sum_k |r_k^f-r_k^0|\left((U_1^{(1)}-\tau_1)+\cdots+(U_{J_T-1}^{(1)}-\tau_{J_T-1})\right) \geqslant \frac{T\sum_k|r_k^f-r_k^0|}{2\mathbb{E}_0[\Delta\tau_1]\,\|r_0\|_1}. \tag{31}$$

It also implies that $\|r_f\|_1 \geqslant \|r_0\|_1 - \left\|r_f-r_0\right\|_1 \geqslant \|r_0\|_1/2$ for T large enough.

Our final argument consists in using the lower bound on $\mathbb{E}_f[Z_{1l}]$ obtained in Lemma A.4. In this technical lemma, we show that there exists $l\in[K]$ and $C(f_0)>0$ such that $\mathbb{E}_f[Z_{1l}] \geqslant C(f_0)\|f-f_0\|_1$. Therefore, for this l,

$$\begin{aligned} &\mathbb{P}_f\left[\tilde{\Omega}_T\cap\{\tilde{d}_{1T}(f,f_0)\leqslant M_T'\epsilon_T\}\middle|\mathcal{G}_0\right] \leqslant \mathbb{P}_f\left[\tilde{\Omega}_T\cap\left\{\sum_{j=1}^{J_T-1}Z_{jl}\leqslant M_T'T\epsilon_T\right\}\middle|\mathcal{G}_0\right] \\ &\leqslant \mathbb{P}_f\left[\tilde{\Omega}_T\cap\left\{\sum_{j=1}^{J_T-1}(Z_{jl}-\mathbb{E}_f\left[Z_{jl}\right]) \leqslant M_T'T\epsilon_T - (J_T-1)\mathbb{E}_f\left[Z_{jl}\right]\right\}\middle|\mathcal{G}_0\right] \\ &\leqslant \mathbb{P}_f\left[\bigcup_{J\in\mathcal{J}_T}\left\{\sum_{j=1}^{J-1}(Z_{jl}-\mathbb{E}_f\left[Z_{jl}\right]) \leqslant -\frac{C(f_0)T\left\|f-f_0\right\|_1}{4\mathbb{E}_0[\Delta\tau_1]}\right\}\middle|\mathcal{G}_0\right] \leqslant \sum_{J\in\mathcal{J}_T}\mathbb{P}_f\left[\sum_{j=1}^{J-1}(Z_{jl}-\mathbb{E}_f\left[Z_{jl}\right]) \leqslant -\frac{C(f_0)T\left\|f-f_0\right\|_1}{4\mathbb{E}_0[\Delta\tau_1]}\middle|\mathcal{G}_0\right], \end{aligned}$$

where we have used, for the third inequality, that on $\tilde{\Omega}_T$, $J_T - 1 \geqslant \frac{T}{2\mathbb{E}_0[\Delta\tau_1]}$, $\|f - f_0\|_1 \geqslant M_T\epsilon_T$ and $M'_T < M_T$. For each $J \in \mathcal{J}_T$, we can now apply the Bernstein's inequality:

$$\mathbb{P}_f\left[\sum_{j=1}^{J-1}(Z_{jl} - \mathbb{E}_f\left[Z_{jl}\right]) \leqslant x\right] \leqslant \exp\left\{-\frac{x^2}{2(J-1)(v^2+bx)}\right\},$$

with $x = -\frac{C(f_0)T\|f-f_0\|_1}{4\mathbb{E}_0[\Delta\tau_1]}$. We first upper bound the term $v^2 + bx$:

$$\begin{aligned} v^2 + b\frac{C(f_0)\|f-f_0\|_1}{4\mathbb{E}_0[\Delta\tau_1]} &\leqslant L\max\left(A+1, \frac{2}{\|r_0\|_1}\right)\left(L\max\left(A+1, \frac{2}{\|r_0\|_1}\right) + \frac{C(f_0)}{2\|r_0\|_1\mathbb{E}_0[\Delta\tau_1]}\right)\|f-f_0\|_1^2 \\ &= C_1(f_0)\|f-f_0\|_1^2, \end{aligned}$$

with $C_1(f_0) := L\max\left(A+1, \frac{2}{\|r_0\|_1}\right)\left(L\max\left(A+1, \frac{2}{\|r_0\|_1}\right) + \frac{C(f_0)}{2\|r_0\|_1\mathbb{E}_0[\Delta\tau_1]}\right)$. Finally, we obtain that

$$\mathbb{P}_f\left[\sum_{j=1}^{J-1}(Z_{jl} - \mathbb{E}_f\left[Z_{jl}\right]) \leqslant -\frac{C(f_0)T\left\|f-f_0\right\|_1}{4\mathbb{E}_0[\Delta\tau_1]}\middle|\mathcal{G}_0\right] \leqslant \exp\left\{-\frac{C(f_0)^2T^2\left\|f-f_0\right\|_1^2}{8(J-1)C_1(f_0)\left\|f-f_0\right\|_1^2}\right\} \leqslant \exp\left\{-\frac{C(f_0)^2T}{16C_1(f_0)}\right\},$$

and since $\kappa_T\epsilon_T^2 = o(1)$, we can conclude that

$$\mathbb{P}_f\left[\tilde{\Omega}_T \cap \{\tilde{d}_{1T}(f, f_0) \leqslant M'_T\epsilon_T\}\middle|\mathcal{G}_0\right] \leqslant \frac{2T}{\mathbb{E}_0[\Delta\tau_1]}\exp\left\{-\frac{C(f_0)^2T}{16C_1(f_0)}\right\} = o(e^{-(c_1+\kappa_T)T\epsilon_T^2}),$$

which corresponds to (27) and terminates the proof of Theorem 3.2 and Case 1 of Proposition 3.5.

5.3. Proof of Case 2 of Proposition 3.5

We recall that in this case we consider a shifted ReLU model with unknown shift $\theta_0 = (\theta_1^0, \ldots, \theta_K^0)$, corresponding to a particular case of partially known link functions $\phi_k(x; \theta_k) = \theta_k + (x)_+$, and for parameter $f \in \mathcal{F}$ and $\theta \in \Theta$, we denote $\lambda_t(f, \theta)$ the intensity process. We note that in this case, $r_0 = \theta_0 + \nu_0$ and similarly $r_f = \theta + \nu$, with r_f defined in (21). We then prove the posterior concentration rate on both f_0 and θ_0. First, we apply the same steps as in the proof of Theorem 3.2 in Section 5.2, replacing $\|f - f_0\|_1$ by $\left\|r_0 - r_f\right\|_1 + \|h - h_0\|_1 = \|\theta_0 + \nu_0 - \theta - \nu\|_1 + \|h - h_0\|_1$. In particular, we re-define the balls w.r.t. the L_1-distance as (for simplicity we keep the same notation)

We therefore obtain (see also Remark 3.7)

$$\mathbb{E}_0\left[\Pi(\|h - h_0\|_1 + \|\theta_0 + \nu_0 - \theta - \nu\|_1 > M\sqrt{\kappa_T}\epsilon_T\middle|N)\right] = o(1). \tag{32}$$

Secondly, we design a test to separate θ_0 and ν_0. For this, we restrict again the set $\tilde{\Omega}_T$ to a high probability set Ω_A, where θ_0 can be correctly estimated. Let

$$A^k(T) = \{t \in [0, T];\ \tilde{\lambda}_t^k(\nu_0, h_0) < 0\}, \quad \Omega_A = \{|A^k(T)| > z_0T,\ \forall k \in \mathcal{K}\}, \quad 1 \leqslant k \leqslant K,$$

with $z_0 > 0$ defined in the proof of Lemma A.1 (see Section S8.1 in the supplementary material [55]), and define $\tilde{\Omega}'_T = \tilde{\Omega}_T \cap \Omega_A$. Moreover, we define a neighborhood around θ_0, $\bar{A}(R) := \{\theta \in \Theta;\ \|\theta - \theta_0\|_1 \leqslant$

$R\}$ and $\tilde{M}_T = \tilde{M}\sqrt{\kappa_T}$ with $\tilde{M} > M$. Using again the decomposition (24), with $A = \bar{A}(\tilde{M}_T\epsilon_T)^c$, $B = A_{L_1}(M_T\epsilon_T)$, and the subset $\tilde{\Omega}'_T$, we thus only need to construct a test function $\phi \in [0,1]$ verifying:

$$\mathbb{E}_0\left[\phi \mathbb{1}_{\tilde{\Omega}'_T}\right] = o(1), \qquad \sup_{\theta\in\bar{A}(\tilde{M}_T\epsilon_T)^c, f\in A_{L_1}(M_T\epsilon_T)\cap\mathcal{F}_T} \mathbb{E}_0\left[\mathbb{E}_f\left[(1-\phi)\mathbb{1}_{\tilde{\Omega}'_T}\right]\middle|\mathcal{G}_0\right] = o(e^{-(\kappa_T+c_1)T\epsilon_T^2}). \tag{33}$$

To construct this test, we first consider some arbitrary parameter $f_1 = ((\nu_k^1)_k, (h_{lk}^1)_{l,k}) \in A_{L_1}(M_T\epsilon_T)$ and $\theta_1 = (\theta_k^1)_k \in \bar{A}(\tilde{M}_T\epsilon_T)^c$, and for any $k \in [K]$, we define the following subset of the observation window

$$I_k^0(f_1,\theta_1) = \left\{t \in [0,T];\ \lambda_t^k(f_1,\theta_1) = \theta_k^1,\ \lambda_t^k(f_0,\theta_0) = \theta_k^0\right\}. \tag{34}$$

By construction θ_k^0 and θ_k^1 can be identified on the set $I_k^0(f_1,\theta_1)$, hence we need $I_k^0(f_1,\theta_1)$ to be large enough in order to test between θ_k^0 and θ_k^1. We can ensure this by defining a controlled set of excursions $\mathcal{E}$. Let $l \in [K]$ such that $h_{lk}^{0-} \neq 0$, $\delta' = (x_2 - x_1)/3$ with x_1, x_2 defined in condition (4), $c_\star = \min_{x\in[x_1,x_2]} h_{lk}^{0-}(x)$ and $n_1 = \lfloor 2\nu_k^1/(\kappa_1 c_\star)\rfloor + 1$ for some $0 < \kappa_1 < 1$. We consider the following subset of excursions:

$$\mathcal{E} := \{j \in [J_T];\ N[\tau_j, \tau_j+\delta') = N^l[\tau_j,\tau_j+\delta') = n_1, N[\tau_j+\delta',\ \tau_{j+1}) = 0\}, \tag{35}$$

where the τ_j's are the regenerative times defined in Lemma 5.1. Using the intermediate result (S5.13) from the proof of Lemma A.5 in the supplementary material [55], if $|\mathcal{E}|$ is large enough, then we can find a lower bound on $|I_k^0(f_1,\theta_1)|$. We then define our generic test function:

$$\phi(f_1,\theta_1) := \max_{k\in[K]} \min\left(\mathbb{1}_{N^k(I_k^0(f_1,\theta_1))-\Lambda_k^0(I_k^0(f_1,\theta_1))<-v_T} \vee \mathbb{1}_{|\mathcal{E}|<\frac{p_0 T}{2\mathbb{E}_0[\Delta\tau_1]}}, \mathbb{1}_{N^k(I_k^0(f_1,\theta_1))-\Lambda_k^0(I_k^0(f_1,\theta_1))>v_T} \vee \mathbb{1}_{|\mathcal{E}|<\frac{p_0 T}{2\mathbb{E}_0[\Delta\tau_1]}}\right), \tag{36}$$

where $p_0 = \mathbb{P}_0[j \in \mathcal{E}]$, $\Lambda_k^0(I_k^0(f_1,\theta_1)) = \int_0^T \mathbb{1}_{I_k^0(f_1,\theta_1)}\lambda_t^k(f_0,\theta_0)dt$, $v_T = w_T T\epsilon_T$, $w_T = 2\sqrt{\max_k \theta_k^0(\kappa_T + c_1)} + 2x_0$ and x_0 from assumption **(A2)**. From Lemma A.5, there exists $u_1 > 2x_0$ and $\zeta \in (0,1)$ such that

$$\mathbb{E}_0\left[\phi(f_1,\theta_1)\mathbb{1}_{\tilde{\Omega}'_T}\right] \leqslant e^{-u_1 T\epsilon_T^2}, \qquad \sup_{\|f-f_1\|+\|\theta-\theta_1\|\leqslant\zeta\epsilon_T} \mathbb{E}_0\left[\mathbb{E}_f\left[(1-\phi(f_1,\theta_1))\mathbb{1}_{\tilde{\Omega}'_T}\right]\middle|\mathcal{G}_0\right] = o(e^{-(\kappa_T+c_1)T\epsilon_T^2}). \tag{37}$$

To define our global test ϕ, we first cover the space $\bar{A}(\tilde{M}_T\epsilon_T)^c \times A_{L_1}(M_T\epsilon_T) \cap \mathcal{F}_T$ with L_1-balls $\{B_i\}_{1\leqslant i\leqslant\mathcal{N}}$ of radius $\zeta\epsilon_T$, with $\zeta > 0$ and $\mathcal{N} \in \mathbb{N}$ the covering number. For each ball B_i centered at (f_i,θ_i), we define the elementary test $\phi(f_i,\theta_i)$ as in (36). Then we define $\phi := \max_{i\in\mathcal{N}} \phi(f_i,\theta_i)$, and obtain that

$$\mathbb{E}_0\left[\phi\mathbb{1}_{\tilde{\Omega}'_T}\right] \leqslant \mathcal{N}e^{-u_1 T\epsilon_T^2}, \qquad \sup_{\theta\in\bar{A}(\tilde{M}_T\epsilon_T)^c, f\in A_{L_1}(M_T\epsilon_T)\cap\mathcal{F}_T} \mathbb{E}_0\left[\mathbb{E}_f\left[(1-\phi)\mathbb{1}_{\tilde{\Omega}'_T}\right]\middle|\mathcal{G}_0\right] = o(e^{-(\kappa_T+c_1)T\epsilon_T^2}).$$

Next, we find an upper bound of the covering number $\mathcal{N}$ using assumption **(A2)**. We note that if $f \in A_{L_1}(M_T\epsilon_T)$, then for any $(l,k) \in [K]^2$, $\theta_k \leqslant \theta_k + \nu_k = r_k^f \leqslant r_k^0 + \epsilon_T \leqslant 2(\theta_k^0 + \nu_k^0)$. Consequently, using similar computations as in the proof of Proposition 5.5 (see Section S1 of the supplementary material [55]), one can find $x_0' > 0$ such that

$$\mathcal{N} \leqslant \left(\frac{2\max_k(\theta_k^0+\nu_k^0)}{\zeta\epsilon_T}\right)^K \left(\frac{\max_k \nu_k^0 + \epsilon_T}{\zeta\epsilon_T}\right)^K \mathcal{N}(\zeta\epsilon_T, \mathcal{H}_T, \|.\|_1) \lesssim e^{-K\log\epsilon_T} e^{x_0' T\epsilon_T^2} \lesssim e^{K\log T} e^{x_0' T\epsilon_T^2} = o(e^{u_1 T\epsilon_T^2}),$$

since $\log T = o(T\epsilon_T^2)$ by assumption. Hence, reporting into (37), this proves that (33) holds and allows us to conclude that $\mathbb{E}_0\left[\Pi(\bar{A}(\tilde{M}_T\epsilon_T)^c|N)\right] = \mathbb{E}_0\left[\Pi(\|\theta-\theta_0\|_1 > \tilde{M}\sqrt{\kappa_T}\epsilon_T|N)\right] = o(1)$. Finally, since $\tilde{M} > M$, from (32), we also have that $\mathbb{E}_0\left[\Pi(\|\nu+\theta-\nu_0-\theta_0\|_1 + \|h-h_0\|_1 > \tilde{M}\sqrt{\kappa_T}\epsilon_T|N)\right] = o(1)$,. Therefore it only remains to prove that $\mathbb{E}_0\left[\Pi(\|\nu-\nu_0\|_1 > \tilde{M}\sqrt{\kappa_T}\epsilon_T|N)\right] = o(1)$. By the triangle inequality, we have $\|\nu-\nu_0\|_1 \leqslant \|\nu+\theta-\nu_0-\theta_0\|_1 + \|\theta-\theta_0\|_1$, and, up to a modification of the constant $\tilde{M}$,

$$\mathbb{E}_0\left[\Pi(\left\|\nu-\nu_0\right\|_1 > \tilde{M}\sqrt{\kappa_T}\epsilon_T|N)\right] \leqslant \mathbb{E}_0\left[\Pi(\left\|\nu+\theta-\nu_0-\theta_0\right\|_1 > \tilde{M}\sqrt{\kappa_T}\epsilon_T|N)\right] + \mathbb{E}_0\left[\Pi(\left\|\theta-\theta_0\right\|_1 > \tilde{M}\sqrt{\kappa_T}\epsilon_T|N)\right] = o(1),$$

which terminates this proof.

5.4. Proof of Theorem 3.9

In this section, we show that in all the models satisfying the assumptions of Theorem 3.2 or Proposition 3.5, the posterior distribution is consistent on the connectivity graph parameter δ_0. For ease of exposition, we here report the proof for the models considered in Theorem 3.2. We first recall the notation $M_T = M\sqrt{\kappa_T}$, $A_{L_1}(M_T\epsilon_T) = \{f \in \mathcal{F};\ \left\|r_f - r_0\right\|_1 + \|h-h_0\|_1 \leqslant M_T\epsilon_T\}$, and $I(\delta_0) = \{(l,k) \in [K]^2,\ \delta^0_{lk} = 1\}$. We first note that

$$\Pi\left(\delta \neq \delta_0|N\right) = \Pi\left(\exists (l,k) \in [K]^2, \delta^0_{lk} \neq \delta_{lk}\middle|N\right) \leqslant \Pi\left(\exists (l,k) \in I(\delta_0), \delta_{lk} = 0\middle|N\right) + \sum_{(l,k)\notin I(\delta_0)} \Pi\left(\delta_{lk} = 1\middle|N\right). \tag{38}$$

For the first term on the RHS of (38), using Theorem 3.2, we have that

$$\Pi\left(\exists (l,k) \in I(\delta_0), \delta_{lk} = 0\middle|N\right) \leqslant \sum_{(l,k)\in I(\delta_0)} \Pi\left(\{\delta_{lk} = 0\} \cap A_{L_1}(M_T\epsilon_T)\middle|N\right) + o_{\mathbb{P}_0}(1).$$

For large enough T, if $\|h^0_{lk}\|_1 > M_0 M_T\epsilon_T$ with $M_0 > 1$, then

$$\{f \in \mathcal{F}; \delta_{lk} = 0\} \subset \{f \in \mathcal{F}; \left\|h^0_{lk} - h_{lk}\right\|_1 = \left\|h^0_{lk}\right\|_1\} \subset \left\{f \in \mathcal{F}; \left\|h^0_{lk} - h_{lk}\right\|_1 > \frac{\|h^0_{lk}\|_1}{2}\right\} \subset A_{L_1}(M_T\epsilon_T)^c,$$

therefore $\Pi\left(\{\delta_{lk} = 0\} \cap A_{L_1}(M_T\epsilon_T)\middle|N\right) = 0$. For the second term on the RHS of (38), since $(l,k) \notin I(\delta_0)$ implies that $\left\|h^0_{lk}\right\|_1 = 0$ and $\{\delta_{lk} = 1\} \cap A_{L_1}(M_T\epsilon_T) \subset \{f \in \mathcal{F};\ 0 < \|h_{lk}\|_1 \leqslant M_T\epsilon_T\}$, defining $N_T = \int_{\{\delta_{lk}=1\}\cap A_{L_1}(M_T\epsilon_T)} e^{L_T(f)-L_T(f_0)} d\Pi(f)$, and using the decomposition (24) with $A = A_{L_1}(M_T\epsilon_T)$, $B = \{\delta_{lk} = 1\}$ and $\phi = 1$, we obtain that

$$\begin{aligned}\mathbb{E}_0\left[\Pi(\{\delta_{lk} = 1\} \cap A_{L_1}(M_T\epsilon_T)|N)\right] &\leqslant \mathbb{P}_0(D_T < e^{-(\kappa_T+c_1)T\epsilon_T^2} \cap \tilde{\Omega}_T) + \mathbb{P}_0(\tilde{\Omega}_T^c) + e^{(\kappa_T+c_1)T\epsilon_T^2}\Pi(\{\delta_{lk} = 1\} \cap A_{L_1}(M_T\epsilon_T)) \\ &\leqslant o(1) + e^{(\kappa_T+c_1)T\epsilon_T^2} \sum_{\delta\in\{0,1\}^{K^2}} \mathbb{1}_{\delta_{lk}=1}\Pi_{h|\delta}\left(\|h_{lk}\|_1 \leqslant M_T\epsilon_T|\delta\right) = o(1),\end{aligned}$$

where in the last inequality we have used assumptions **(A0)-(A1)**, (11), and the construction of the prior from Section 3.2. Consequently, from (38), we finally arrive at $\mathbb{E}_0\left[\Pi\left(\delta \neq \delta_0|N\right)\right] = o(1)$.

5.5. Proof of Theorem 3.11

We here prove the consistency of the penalised estimator defined in (13). We consider the models satisfying the assumptions of Theorem 3.2, although our proof is also valid for the ReLU-type models

of Proposition 3.5. Besides, for $f \in \mathcal{F}$, we use the shortened notation $d_{1T} := \tilde{d}_{1T}(f, f_0)$ and $\hat{\delta}^{\Pi,L} := \hat{\delta}^{\Pi,L}(N)$. We recall that for $(l,k) \in [K]^2$, $S_{lk} = \|h_{lk}\|_1$ and the notation from previous proofs, $M_T = M\sqrt{\kappa_T}$, $M'_T = M'\sqrt{\kappa_T}$ with $M > M' > 0$. We first note that $\mathbb{P}_0\left[\hat{\delta}^{\Pi,L} \neq \delta_0\right] \leqslant \sum_{l,k} \mathbb{P}_0\left[\hat{\delta}^{\Pi,L}_{lk} \neq \delta^0_{lk}\right]$ and consider two cases for each (l,k).

- **Case 1:** $(l,k) \notin I(\delta_0)$, i.e, $\delta^0_{lk} = 0$. Using (13) and (14), there exists $a > 0$ such that with $c'_1 := a + c_1 + \kappa_T$, for any $\gamma > 0$, we have

$$\begin{aligned}
\mathbb{P}_0\left[\hat{\delta}^{\Pi,L}_{lk} \neq \delta^0_{lk}\right] &= \mathbb{P}_0\left[\hat{\delta}^{\Pi,L}_{lk} = 1\right] \\
&\leqslant \mathbb{P}_0\left[e^{-c'_1 T\epsilon_T^2}\Pi(\delta_{lk} = 1,\, S_{lk} \leqslant M_T\epsilon_T|N) \geqslant \Pi(\delta_{lk} = 0|N) - \Pi(S_{lk} > M_T\epsilon_T|N)\right] \\
&\leqslant \mathbb{P}_0\left[e^{-c'_1 T\epsilon_T^2}\Pi(\delta_{lk} = 1,\, S_{lk} \leqslant M_T\epsilon_T|N) \geqslant \Pi(\delta_{lk} = 0|N)/2\right] \\
&\quad + \mathbb{P}_0\left[\Pi(S_{lk} > M_T\epsilon_T|N) > \Pi(\delta_{lk} = 0|N)/2\right]. \qquad (39)
\end{aligned}$$

To show that the second term in the previous equation is $o(1)$, it is enough to show that

$$\mathbb{P}_0\left[\Pi(d_{1T} > M'_T\epsilon_T|N) > \Pi(\delta_{lk} = 0|N)/4\right] = o(1), \qquad (40)$$

$$\mathbb{P}_0\left[\Pi(d_{1T} \leqslant M'_T\epsilon_T,\, S_{lk} > M_T\epsilon_T|N) > \Pi(\delta_{lk} = 0|N)/4\right] = o(1). \qquad (41)$$

Let $m_T(\delta_{lk} = 0) := \int_{\mathcal{F}_T} e^{L_T(f) - L_T(f_0)} d\Pi(f|\delta_{lk} = 0)$. Similarly to the computations of the lower bound of D_T in Section S1, we have under **(A0')** that $\mathbb{P}_0\left[m_T(\delta_{lk} = 0) \leqslant e^{-\kappa'_T T\epsilon_T^2}\right] = o(1)$ with $\kappa'_T := \kappa_T + c_1$. Using the test function from the proof of Theorem 5.5 in Section S1 in the supplementary material [55] $\phi = \max_i \phi(f_i)$ (with $\phi(f_i)$ defined in Lemma S5.1) we have

$$\begin{aligned}
\mathbb{P}_0\left[\Pi(d_{1T} > M'_T\epsilon_T|N) > \Pi(\delta_{lk} = 0|N)/4\right] &\leqslant \mathbb{E}_0\left[\phi \mathbb{1}_{\tilde{\Omega}_T}\right] + \mathbb{P}_0\left[\tilde{\Omega}_T^c\right] + \Pi(\mathcal{F}_T^c) \\
&\quad + \mathbb{E}_0\left[(1-\phi)\mathbb{1}_{\tilde{\Omega}_T}\mathbb{1}_{\int_{\mathcal{F}_T} \mathbb{1}_{d_{1T} > M'_T\epsilon_T} e^{L_T(f) - L_T(f_0)} d\Pi(f) > \Pi(\delta_{lk}=0) m_T(\delta_{lk}=0)/4}\right] \\
&\leqslant o(1) + \mathbb{E}_0\left[(1-\phi)\mathbb{1}_{\tilde{\Omega}_T}\mathbb{1}_{\int_{\mathcal{F}_T} \mathbb{1}_{d_{1T} > M'_T\epsilon_T} e^{L_T(f) - L_T(f_0)} d\Pi(f) > e^{-\kappa'_T T\epsilon_T^2}/4}\right] \\
&\leqslant o(1) + 4e^{\kappa'_T T\epsilon_T^2} \int_{\mathcal{F}_T} \mathbb{E}_0\left[\mathbb{E}_f\left[\mathbb{1}_{\tilde{\Omega}_T}\mathbb{1}_{d_{1T} > M'_T\epsilon_T}(1-\phi)|\mathcal{G}_0\right] d\Pi(f|\delta_{lk} = 0)\right].
\end{aligned}$$

In the second inequality, we have notably used (S1.4) $\mathbb{E}_0\left[\phi \mathbb{1}_{\tilde{\Omega}_T}\right] = o(1)$ from Section S1. Moreover, from (S1.5), there exists $\gamma_1 > 0$ such that

$$\sum_{i \geqslant M'_T} \int_{\mathcal{F}_T} \mathbb{E}_f\left[\mathbb{1}_{\tilde{\Omega}_T}\mathbb{1}_{f \in S_i}(1-\phi)|\mathcal{G}_0\right] d\Pi(f|\delta_{lk} = 0) \leqslant 4(2K+1)e^{-\gamma_1 M'^2_T T\epsilon_T^2},$$

where the S_i's are the slices defined in (S1.1). Therefore, we obtain (40) using that

$$\mathbb{P}_0\left[\Pi(d_{1T} > M'_T\epsilon_T|N) > \Pi(\delta_{lk} = 0|N)/4\right] \leqslant o(1) + 4e^{\kappa'_T T\epsilon_T^2} 4(2K+1)e^{-(M'_T)^2 T\epsilon_T^2} = o(1).$$

To prove (41), using Markov's inequality and Fubini's theorem, we have, for M' large enough, that

$$\begin{aligned}
&\mathbb{P}_0\left[\Pi(d_{1T} \leqslant M'_T\epsilon_T,\, S_{lk} > M_T\epsilon_T|N) > \Pi(\delta_{lk}=0|N)/4\right] \\
&\quad\leqslant \mathbb{P}_0\left[\{m_T(\delta_{lk}=0) < e^{-\kappa'_T T\epsilon_T^2}\} \cap \tilde{\Omega}_T\right] + \mathbb{P}_0\left[\tilde{\Omega}_T^c\right] \\
&\quad\quad + 4e^{\kappa'_T T\epsilon_T^2}\mathbb{E}_0\left[\int_{\mathcal{F}_T\cap\{S_{lk}>M_T\epsilon_T\}} \mathbb{1}_{\tilde{\Omega}_T}\mathbb{1}_{d_{1T}\leqslant M'_T\epsilon_T} e^{L_T(f)-L_T(f_0)} d\Pi(f|\delta_{lk}=0)\right] \\
&\quad= o(1) + 4e^{\kappa'_T T\epsilon_T^2}\int_{S_{lk}>M_T\epsilon_T} \mathbb{E}_0\left[\mathbb{P}_f\left[\tilde{\Omega}_T \cap \{d_{1T} \leqslant M'_T\epsilon_T\}|\mathcal{G}_0\right]\right] d\Pi(f).
\end{aligned}$$

Moreover, from (27), we have that $\sup_{f\in A_{L_1}(M_T\epsilon_T)^c\cap\mathcal{F}_T} \mathbb{P}_f\left[\tilde{\Omega}_T \cap \{d_{1T} \leqslant M'_T\epsilon_T\}|\mathcal{G}_0\right] = o(e^{-\kappa'_T T\epsilon_T^2})$. Finally, since $\delta^0_{lk} = 0$, $S_{lk} > M_T\epsilon_T$ implies that $f \in A^c_{L_1}(M_T\epsilon_T)$, which thus leads to (41). Reporting into (39), we now have

$$\begin{aligned}
\mathbb{P}_0\left[\hat{\delta}^{\Pi,L}_{lk}=1\right] &\leqslant \mathbb{P}_0\left[e^{-c'_1 T\epsilon_T^2}\Pi(\delta_{lk}=1,\, S_{lk}\leqslant M_T\epsilon_T|N) \geqslant \Pi(\delta_{lk}=0|N)/2\right] + o(1) \\
&\leqslant \mathbb{P}_0\left[e^{-c'_1 T\epsilon_T^2}\Pi(\delta_{lk}=1|N) \geqslant \Pi(\delta_{lk}=0|N)/2\right] + o(1) \\
&= \mathbb{P}_0\left[e^{-c'_1 T\epsilon_T^2} m_T(\delta_{lk}=1) \geqslant \frac{\Pi(\delta_{lk}=0)}{2\Pi(\delta_{lk}=1)} m_T(\delta_{lk}=0)\right] + o(1) \\
&\leqslant \mathbb{P}_0\left[\left\{e^{-c'_1 T\epsilon_T^2} m_T(\delta_{lk}=1) \geqslant \frac{\Pi(\delta_{lk}=0)}{2\Pi(\delta_{lk}=1)} e^{-\kappa'_T T\epsilon_T^2}\right\} \cap \tilde{\Omega}_T\right] + \mathbb{P}_0\left[m_T(\delta_{lk}=0) < e^{-\kappa'_T T\epsilon_T^2}\right] + o(1) \\
&\leqslant \mathbb{P}_0\left[\left\{m_T(\delta_{lk}=1) \geqslant \frac{\Pi(\delta_{lk}=0)}{2\Pi(\delta_{lk}=1)} e^{(c'_1-\kappa'_T)T\epsilon_T^2}\right\} \cap \tilde{\Omega}_T\right] + o(1) \\
&\leqslant \mathbb{E}_0\left[m_T(\delta_{lk}=1)\right] \frac{2\Pi(\delta=1)}{\Pi(\delta_{lk}=0)} e^{-(c'_1-\kappa'_T)T\epsilon_T^2} + o(1) \leqslant \frac{2\Pi(\delta_{lk}=1)^2}{\Pi(\delta_{lk}=0)} e^{-(c'_1-\kappa'_T)T\epsilon_T^2} + o(1) = o(1),
\end{aligned}$$

since $c'_1 > \kappa_T + c_1 = \kappa'_T$ and that $\mathbb{E}_0\left[m_T(\delta=1)\right] = \Pi(\delta_{lk}=1)$ with Fubini's theorem.

- **Case 2:** $(l,k) \in I(\delta_0)$, i.e, $\delta^0_{lk} = 1$. In the case, the computations are slightly simpler since $\{\delta_{lk} = 0\} \implies f \in A_{L_1}(M\sqrt{\kappa_T}\epsilon_T)^c$ and for T large enough, $S^0_{lk} - M_T\epsilon_T > 0$. Thus we can use the fact that $\Pi(\delta_{lk}=0|N) \leqslant \Pi(A_{L_1}(M\sqrt{\kappa_T}\epsilon_T)^c|N)$ (the full computations are reported in Section S2.1 in the supplementary material [55].

6. Conclusion

In this paper we have established concentration and consistency properties of the posterior distribution and of Bayesian estimators of the parameter and connectivity graph, in a general class of nonlinear Hawkes processes. These results validate the common use of these models in different applied contexts. In particular, our results include the commonly used sigmoid and softplus models, as well as the more challenging ReLU model, under some additional restrictions on the parameter space. Moreover, we provide the first theoretical results for estimating an additional parameter of the link functions, in the case of shifted ReLU with unknown shift. To prove those results, we have built a new technique for obtaining model identifiability and concentration inequalities based on the decomposition of the process into excursions, recently introduced by Costa et al. [11]. Finally, our results hold under reasonable

assumptions on the prior distribution and the true model, and we provide practical examples for which those conditions are verified.

Although rather weak assumptions have been used to prove our results, it is likely that the latter hold in more general contexts. In particular, we believe that one could relax the condition on processes with bounded memory ($A < +\infty$) since the regenerative properties of the nonlinear Hawkes processes also hold for processes with unbounded memory. One major improvement of our results would be to consider high dimensional processes ($K \to \infty$), possibly in restricted models such as sparse models [3] or clustering models [48]. Another perspective would be to prove the frequentist minimax rate of estimation, since it would be of great interest to evaluate the optimality of Bayesian procedures in nonlinear Hawkes processes. Some practitioners might also be interested in additional results on the estimation of the link function, through a different parametric or even nonparametric form, like in [62].

Appendix A: Main lemmas

In this section, we state some important lemmas to prove our main results in Section 5. The proofs of Lemmas A.1, A.2, A.4 and A.5 are provided in Sections S1, S5 and S8 in the supplementary material [55]. The first two lemmas are controls respectively of the complement of the main event $\tilde{\Omega}_T$ under the true distribution $\mathbb{P}_0$, and of the deviations of the log likelihood ratio $L_T(f_0) - L_T(f)$.

Lemma A.1. *Let $Q > 0$. We consider $\tilde{\Omega}_T$ defined in (25) in Section 5.2. For any $\beta > 0$, we can choose C_β and c_β in the definition of $\tilde{\Omega}_T$ such that $\mathbb{P}_0[\tilde{\Omega}_T^c] \leqslant T^{-\beta}$. Moreover, for any $1 \leqslant q \leqslant Q$, $\mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T^c} \max_l \sup_{t\in[0,T]} \left(N^l[t-A,t)\right)^q\right] \leqslant 2T^{-\beta/2}$. Finally, the previous results hold when replacing $\tilde{\Omega}_T$ by $\tilde{\Omega}'_T = \tilde{\Omega}_T \cap \Omega_A$ with Ω_A defined in Section 5.3 for the model with shifted ReLU link and unknown shift.*

Lemma A.2. *Under the assumptions of Theorem 3.2 or Proposition 3.5, for any $f \in B_\infty(\epsilon_T)$ and T large enough, we have*

$$\mathbb{P}_0\left[L_T(f_0) - L_T(f) \geqslant \frac{1}{2}\kappa_T T \epsilon_T^2\right] = o(1).$$

with

$$\kappa_T = \begin{cases} 10 & \text{(under Assumption 3.1(i))} \\ 10(\log T) & \text{(under Assumption 3.1(ii))} \\ 10(\log T)^2 & \text{(under Case 1 and condition (8))} \end{cases}$$

***Remark* A.3.** Contrary to the typical approach, the proof of Lemma A.2 is not based on the control of the variance of $L_T(f_0) - L_T(f)$, which is intractable due to the nonlinear form of the log-likelihood function, but on a decomposition of $L_T(f_0) - L_T(f) - KL(f_0, f)$ into a sum of i.i.d. terms T_j defined as:

$$T_j := \sum_k \int_{\tau_j}^{\tau_{j+1}} \log\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right) dN_t^k - \int_{\tau_j}^{\tau_{j+1}} (\lambda_t^k(f_0) - \lambda_t^k(f))dt.$$

The next lemma is a notably used in the proof of Theorem 3.2 in Section 5.2 and bridges the gap between the posterior concentration rate in stochastic distance (see Theorem 5.5) and the rate in L_1-distance (Theorem 3.2).

Lemma A.4. *For $f \in \mathcal{F}_T$ and $l \in [K]$, let*

$$Z_{1l} = \int_{\tau_1}^{\xi_1} |\lambda_t^l(f) - \lambda_t^l(f_0)| dt,$$

where ξ_1 is defined in (22) in Section 5.2. Under the assumptions of Theorem 3.2 and Case 1 of Proposition 3.5, for $M_T \to \infty$ such that $M_T > M\sqrt{\kappa_T}$ with $M > 0$ and for any $f \in \mathcal{F}_T$ such that $\|\nu - \nu_0\|_1 \leqslant \max(\|\nu_0\|_1, \tilde{C})$ with $\tilde{C} > 0$, there exists $l \in [K]$ such that on $\tilde{\Omega}_T$,

$$\mathbb{E}_f[Z_{1l}] \geqslant C(f_0) \|f - f_0\|_1,$$

with $C(f_0) > 0$ a constant that depends only on f_0 and $\phi = (\phi_k)_k$.

Similarly, under the assumptions of Case 2 of Proposition 3.5, for $f \in \mathcal{F}_T$ and $\theta \in \Theta$, let $r_0 = (r_k^0)_k$, $r_f = (r_k^f)_k$ with $r_k^0 = \phi_k(\nu_k^0) = \theta_k^0 + \nu_k^0$, $r_k^f = \phi_k(\nu_k) = \theta_k + \nu_k$, $\forall k$. If $\left\|r_f - r_0\right\|_1 \leqslant \max(\|r_0\|, \tilde{C}')$ with $\tilde{C}' > 0$, then there exists $l \in [K]$ such that on $\tilde{\Omega}_T$,

$$\mathbb{E}_f[Z_{1l}] \geqslant C'(f_0)(\|r_f - r_0\|_1 + \|h - h_0\|_1), \quad C'(f_0) > 0. \tag{42}$$

Finally, this last lemma provides upper bounds on type I and type II errors for the tests used in the proof of Case 2 of Proposition 3.5 in Section 5.3 for estimating the parameter of the link functions θ_0.

Lemma A.5. *Using the notations of Section 5.3, for $\theta_1 \in \bar{A}(\tilde{M}_T \epsilon_T)^c$, $f_1 \in A_{L_1}(M_T \epsilon_T) \cap \mathcal{F}_T$, we define*

$$\phi(f_1, \theta_1) = \max_{k \in [K]} \min\left(\mathbb{1}_{N^k(I_k^0(f_1,\theta_1)) - \Lambda_k^0(I_k^0(f_1,\theta_1)) < -v_T} \vee \mathbb{1}_{|\mathcal{E}| < \frac{p_0 T}{2\mathbb{E}_0[\Delta\tau_1]}}, \mathbb{1}_{N^k(I_k^0(f_1,\theta_1)) - \Lambda_k^0(I_k^0(f_1,\theta_1), f_0) > v_T} \vee \mathbb{1}_{|\mathcal{E}| < \frac{p_0 T}{2\mathbb{E}_0[\Delta\tau_1]}} \right),$$

with $I_k^0(f_1, \theta_1)$ and $\mathcal{E}$ defined in (34) and (35), $p_0 = \mathbb{P}_0[j \in \mathcal{E}]$, $\Lambda_k^0(I_0^k(f_1, \theta_1)) = \int_0^T \mathbb{1}_{I_k^0(f_1,\theta_1)} \lambda_t^k(f_0, \theta_0) dt$ and $v_T = w_T T \epsilon_T$ with $w_T = 2\sqrt{\max_k \theta_k^0(\kappa_T + c_1)} + 2x_0$ and $x_0 > 0$. Then there exists $u_1 > 2x_0$ such that

$$\mathbb{E}_0\left[\phi(f_1, \theta_1) \mathbb{1}'_{\tilde{\Omega}_T}\right] \leqslant e^{-u_1 T \epsilon_T^2}, \quad \sup_{\|\theta - \theta_1\| + \|f - f_1\| \leqslant \zeta \epsilon_T} \mathbb{E}_0\left[\mathbb{E}_f\left[(1 - \phi(f_1, \theta_1)) \mathbb{1}_{\tilde{\Omega}'_T}\right] \middle| \mathcal{G}_0\right] = o(e^{-(\kappa_T + c_1) T \epsilon_T^2}).$$

Supplementary Material

The supplementary material contains nine sections and includes the proofs of our main results, notably Proposition 2.3, Proposition 2.5, Corollary 3.8, Proposition 3.10, Theorem 5.5, Theorem 3.11 (second case). It also includes an alternative construction of the prior distribution and the proofs of our technical lemmas in Section 5 and Appendix A and Lemma 2.6. Finally, the last section contains some useful results, in particular some extensions of the results from [11] related to the regenerative properties of nonlinear Hawkes processes.

Acknowledgements: The project leading to this work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 834175). The project is also partially funded by the EPSRC via the CDT OxWaSP. The authors would like to thank the Editor and two anonymous referees for valuable comments and suggestions.

References

[1] Apostolopoulou, I., Linderman, S., Miller, K. and Dubrawski, A. (2019). Mutually regressive point processes. *Advances in Neural Information Processing Systems* **32**.

[2] Arbel, J., Gayraud, G. and Rousseau, J. (2013). Bayesian optimal adaptive estimation using a sieve prior. *Scandinavian Journal of Statistics* **40** 549–570.

[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** 1–32.

[4] Bacry, E., Delattre, S., Hoffmann, M. and Muzy, J.-F. (2013). Some limit theorems for Hawkes processes and application to financial statistics. *Stochastic Processes and their Applications* **123** 2475–2499.

[5] Brémaud, P. and Massoulié, L. (1996). Stability of nonlinear Hawkes processes. *The Annals of Probability* 1563–1588.

[6] Brémaud, P., Nappo, G. and Torrisi, G. L. (2002). Rate of convergence to equilibrium of marked Hawkes processes. *Journal of Applied Probability* 123–136.

[7] Carstensen, L., Sandelin, A., Winther, O. and Hansen, N. R. (2010). Multivariate Hawkes process models of the occurrence of regulatory elements. *BMC bioinformatics* **11** 456.

[8] Chen, S., Shojaie, A., Shea-Brown, E. and Witten, D. (2017). The multivariate Hawkes process in high dimensions: beyond mutual excitation. *arXiv:1707.04928v2*.

[9] Chen, S., Witten, D. and Shojaie, A. (2017). Nearly assumptionless screening for the mutually-exciting multivariate Hawkes process. *Electronic Journal of Statistics* **11** 1207–1234.

[10] Chornoboy, E., Schramm, L. and Karr, A. (1988). Maximum likelihood identification of neural point process systems. *Biological cybernetics* **59** 265–275.

[11] Costa, M., Graham, C., Marsalle, L. and Tran, V. C. (2020). Renewal in Hawkes processes with self-excitation and inhibition. *Advances in Applied Probability* **52** 879–915.

[12] Dassios, A. and Zhao, H. (2011). A dynamic contagion process. *Advances in Applied Probability* **43** 814–846.

[13] Delattre, S. and Fournier, N. (2016). Statistical inference versus mean field limit for Hawkes processes. *Electronic Journal of Statistics* **10** 1223–1295.

[14] Delattre, S., Fournier, N. and Hoffmann, M. (2016). Hawkes processes on large networks. *Ann. Appl. Probab.* **26** 216–261.

[15] Deutsch, I. and Ross, G. J. (2022). Bayesian estimation of multivariate Hawkes processes with inhibition and sparsity. *arXiv preprint arXiv:2201.05009*.

[16] Didelez, V. (2008). Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70** 245–264.

[17] Donnet, S., Rivoirard, V. and Rousseau, J. (2020). Nonparametric Bayesian estimation for multivariate Hawkes processes. *The Annals of Statistics* **48** 2698 – 2727.

[18] 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* 1555–1564.

[19] Du, N., Farajtabar, M., Ahmed, A., Smola, A. J. and Song, L. (2015). Dirichlet-Hawkes processes with applications to vlustering vontinuous-time document streams. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. *KDD '15* 219–228. Association for Computing Machinery, New York, NY, USA.

[20] Eichler, M., Dahlhaus, R. and Dueck, J. (2017). Graphical modeling for multivariate Hawkes processes with nonparametric link functions. *Journal of Time Series Analysis* **38** 225–242.

[21] Embrechts, P., Liniger, T. and Lin, L. (2011). Multivariate Hawkes processes: an application to financial data. *Journal of Applied Probability* **48** 367–378.

[22] Ertekin, S., Rudin, C. and McCormick, T. H. (2015). Reactive point processes: A new approach to predicting power failures in underground electrical systems. *Ann. Appl. Stat.* **9** 122–144.
[23] Farajtabar, M., Wang, Y., Gomez Rodriguez, M., Li, S., Zha, H. and Song, L. (2015). Coevolve: a joint point process model for information diffusion and network co-evolution. *Advances in Neural Information Processing Systems* **28**.
[24] Gao, F. and Zhu, L. (2018). Some asymptotic results for nonlinear Hawkes processes. *Stochastic Processes and their Applications* **128** 4051–4077.
[25] Gao, X. and Zhu, L. (2018). Functional central limit theorems for stationary Hawkes processes and application to infinite-server queues. *Queueing Systems* **90** 161–206.
[26] Gerhard, F., Deger, M. and Truccolo, W. (2017). On the stability and dynamics of stochastic spiking neuron models: nonlinear Hawkes process and point process GLMs. *PLOS Computational Biology* **13** e1005390.
[27] Ghosal, S., Ghosh, J. K. and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *The Annals of Statistics* **28** 500 – 531.
[28] Ghosal, S. and van der Vaart, A. (2007). Convergence rates of posterior distributions for non iid observations. *The Annals of Statistics* **35** 192-223.
[29] Graham, C. (2021). Regenerative properties of the linear Hawkes process with unbounded memory. *The Annals of Applied Probability* **31** 2844–2863.
[30] Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society* 424–438.
[31] Gusto, G. and Schbath, S. S. (2005). FADO: A statistical method to detect favored or avoided distances between occurrences of motifs using the Hawkes model. *Statistical Applications in Genetics and Molecular Biology* **4** n.p. article n° 24.
[32] Hansen, N. R., Reynaud-Bouret, P. and Rivoirard, V. (2015). Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli* **21** 83–143.
[33] Hawkes, A. G. (1971). Point Spectra of Some Mutually Exciting Point Processes. *Journal of the Royal Statistical Society. Series B (Methodological)* **33** 438–443.
[34] Hillairet, C., Huang, L., Khabou, M. and Réveillac, A. (2021). The Malliavin-Stein method for Hawkes functionals. *arXiv preprint arXiv:2104.01583*.
[35] Isham, V. and Westcott, M. (1979). A self-correcting point process. *Stochastic Processes and their Applications* **8** 335–347.
[36] Karabash, D. (2012). On stability of Hawkes process. *arXiv preprint arXiv:1201.1573*.
[37] Karabash, D. and Zhu, L. (2015). Limit theorems for marked Hawkes processes with application to a risk model. *Stochastic Models* **31** 433–451.
[38] Lambert, R., Tuleau-Malot, C., Bessaih, T., Rivoirard, V., Bouret, Y., Leresche, N. and Reynaud-Bouret, P. (2017). Reconstructing the functional connectivity of multiple spike trains using Hawkes models. *Journal of Neuroscience Methods* **297**.
[39] Lewis, E. and Mohler, G. (2011). A nonparametric EM algorithm for multiscale Hawkes processes. *Journal of Nonparametric Statistics* **1** 1–20.
[40] Malem-Shinitski, N., Ojeda, C. and Opper, M. (2022). Variational Bayesian Inference for Nonlinear Hawkes Process with Gaussian Process Self-Effects. *Entropy* **24**.
[41] Massoulié, L. (1998). Stability results for a general class of interacting point processes dynamics, and applications. *Stochastic Processes and their Applications* **75** 1-30.
[42] Mei, H. and Eisner, J. M. (2017). The neural Hawkes process: A neurally self-modulating multivariate point process. *Advances in neural information processing systems* **30**.
[43] Menon, A. and Lee, Y. (2018). Proper loss functions for nonlinear Hawkes processes. In *Proceedings of the AAAI Conference on Artificial Intelligence* **32**.

[44] Miscouridou, X., Caron, F. and Teh, Y. W. (2018). Modelling sparsity, heterogeneity, reciprocity and community structure in temporal interaction data. *Advances in Neural Information Processing Systems* **31**.

[45] Møller, J. and Rasmussen, J. G. (2005). Perfect simulation of Hawkes processes. *Advances in applied probability* **37** 629–646.

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

[47] Raad, M. B. (2019). Renewal time points for Hawkes processes. *arXiv preprint arXiv:1906.02036*.

[48] Raad, M. B., Ditlevsen, S. and Löcherbach, E. (2020). Stability and mean-field limits of age dependent Hawkes processes. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* **56** 1958–1990. Institut Henri Poincaré.

[49] Rasmussen, J. G. (2013). Bayesian Inference for Hawkes Processes. *Methodology and Computing in Applied Probability* **15** 623–642.

[50] Reynaud-Bouret, P., Rivoirard, V., Grammont, F. and Tuleau-Malot, C. (2014). Goodness-of-fit tests and nonparametric adaptive estimation for spike train analysis. *The Journal of Mathematical Neuroscience* **4** 1–41.

[51] Reynaud-Bouret, P. and Roy, E. (2007). Some non asymptotic tail estimates for Hawkes processes. *Bulletin of the Belgian Mathematical Society-Simon Stevin* **13** 883–896.

[52] Reynaud-Bouret, P. and Schbath, S. (2010). Adaptive estimation for Hawkes processes; application to genome analysis. *The Annals of Statistics* **38** 2781-2822.

[53] Rousseau, J. (2010). Rates of convergence for the posterior distributions of mixtures of Betas and adaptive nonparametric estimation of the density. *Annals of Statistics* **38** 146–180.

[54] Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *The Annals of Statistics* **22** 118 – 171.

[55] Sulem, D., Rivoirard, V. and Rousseau, J. (2022). Supplement to "Bayesian estimation of nonlinear Hawkes processes".

[56] Torrisi, G. L. (2016). Gaussian approximation of nonlinear Hawkes processes. *The Annals of Applied Probability* **26** 2106–2140.

[57] Torrisi, G. L. (2017). Poisson approximation of point processes with stochastic intensity, and application to nonlinear Hawkes processes. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* **53** 679–700. Institut Henri Poincaré.

[58] Truccolo, W., Eden, U. T., Fellows, M. R., Donoghue, J. P. and Brown, E. N. (2005). A Point Process Framework for Relating Neural Spiking Activity to Spiking History, Neural Ensemble, and Extrinsic Covariate Effects. *Journal of Neurophysiology* **93** 1074-1089. PMID: 15356183.

[59] van der Vaart, A. and van Zanten, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. *Annals of Statistics* **37** 2655-2675.

[60] van der Vaart, A. W. and van Zanten, J. H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics* **36** 1435–1463.

[61] 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** 614–624.

[62] Wang, Y., Xie, B., Du, N. and Song, L. (2016). Isotonic Hawkes processes. In *International conference on machine learning* 2226–2234.

[63] Xu, H., Farajtabar, M. and Zha, H. (2016). Learning granger causality for Hawkes processes. *33rd International Conference on Machine Learning, ICML 2016* **4** 2576–2588.

[64] Zhou, F., Kong, Q., Deng, Z., Kan, J., Zhang, Y., Feng, C. and Zhu, J. (2022). Efficient Inference for Dynamic Flexible Interactions of Neural Populations. *Journal of Machine Learning Research* **23** 1–49.

[65] Zhou, F., Kong, Q., Zhang, Y., Feng, C. and Zhu, J. (2021). Nonlinear Hawkes processes in time-varying system. *arXiv preprint arXiv:2106.04844*.

[66] Zhou, F., Luo, S., Li, Z., Fan, X., Wang, Y., Sowmya, A. and Chen, F. (2021). Efficient EM-variational inference for nonparametric Hawkes process. *Statistics and Computing* **31** 1–11.

Submitted to Bernoulli

Supplementary material of Bayesian estimation of nonlinear Hawkes processes

DÉBORAH SULEM [1], VINCENT RIVOIRARD [2,†] and JUDITH ROUSSEAU [3,*]

[1] *University of Oxford, E-mail:* deborah.sulem@stats.ox.ac.uk; *judith.rousseau@stats.ox.ac.uk
[2] *Ceremade, CNRS, UMR 7534, Université Paris-Dauphine, PSL University, 75016 Paris, France. E-mail:* †Vincent.Rivoirard@dauphine.fr
[3] *University of Oxford, E-mail:* deborah.sulem@stats.ox.ac.uk; *judith.rousseau@stats.ox.ac.uk
[4] *CEREMADE, Université Paris Dauphine, E-mail:* †Vincent.Rivoirard@dauphine.fr

This supplementary material contains additional results and proofs that could not be included in the main paper [8] due to space limitations. In Section S1, we report the proofs of Theorem 5.5 and Lemma A.2. Section S2 contains the proofs of two results in the graph estimation problem (second part of Theorem 3.11 and Proposition 3.10). In Section S3 we prove frequentist results of Corollary 3.8. Results regarding the construction of prior distributions can be found in Section S4. In Sections S5, S6 and S7 we report additional technical results and their proofs, notably on the tests used in the main theorems and on the Kullback-Leibler divergence defined for the Hawkes model. Lemmas A.1 and A.4 are proved in Section S8. Finally, we report multivariate extensions of existing results on the regenerative properties of the nonlinear Hawkes model in Section S9.

For the sake of simplicity, all sections, theorems, corollaries, lemmas and equations presented in the supplement are designed with a prefix S. Regarding the others, we refer to the material of the main text [8]. This is not specified at each place.

S1. Proofs of Theorem 5.5 and of Lemma A.2

S1.1. Proof of Theorem 5.5

This section contains the proof of the posterior concentration rate w.r.t. the stochastic distance defined in (23) in [8]. We use the well-known strategy of [4] which has the following steps. First, the space of observations is restricted to a subset $\tilde{\Omega}_T$ defined in (25) which has high probability (see Lemma A.1). Secondly, we use a lower bound of the denominator D_T defined in (5) using Lemma A.2. Thirdly, we consider $A_{d_1}(M'_T\epsilon_T) \subset \mathcal{F}$, the ball centered at f_0 of radius $M'_T\epsilon_T$ w.r.t the auxiliary stochastic distance $\tilde{d}_{1T}$. To find an upper bound of the numerator $N_T(A_{d_1}(M'_T\epsilon_T)^c)$ as defined in (5), $A_{d_1}(M'_T\epsilon_T)^c$ is partitioned into slices S_i on which we can design tests that have exponentially decreasing type I and type II errors (see Lemma S5.1). We then define ϕ as the maximum of the tests on the individual slices S_i. Note that the following proof applies to all estimation scenarios, and for generality here, we consider θ_0 unknown.

We recall the notation $A_{d_1}(\epsilon) = \{f \in \mathcal{F};\ \tilde{d}_{1T}(f, f_0) \leqslant \epsilon\}$. and from (5), $D_T = \int_{\mathcal{F}} e^{L_T(f)-L_T(f_0)} d\Pi(f)$. For a sequence ϵ_T verifying the assumptions of Theorem 3.2 and for $i \geqslant 1$, we denote

$$S_i = \{f \in \mathcal{F}_T;\ Ki\epsilon_T \leqslant \tilde{d}_{1T}(f, f_0) \leqslant K(i+1)\epsilon_T\}, \tag{S1.1}$$

where $\mathcal{F}_T = \{f = (\nu, h) \in \mathcal{F};\ h = (h_{lk})_{l,k} \in \mathcal{H}_T,\ \nu \in \Upsilon_T\}$. Let $M'_T = M'\sqrt{\kappa_T}$ with $M' > 0$ and κ_T defined in (6). Using the decomposition (24) with $A = A_{d_1}(M'_T\epsilon_T)^c$ (and $B = \mathcal{F}$) , for any test function $\phi \in [0,1]$, we have

$$\mathbb{E}_0[\Pi(A_{d_1}(M'_T\epsilon_T)^c|N)] \leqslant \mathbb{P}_0(\tilde{\Omega}_T^c) + \mathbb{P}_0\left(\{D_T < e^{-\kappa_T T\epsilon_T^2}\Pi(B_\infty(\epsilon_T))\} \cap \tilde{\Omega}_T\right) + \mathbb{E}_0[\phi \mathbb{1}_{\tilde{\Omega}_T}]$$
$$+ \frac{e^{\kappa_T T\epsilon_T^2}}{\Pi(B_\infty(\epsilon_T))}\Pi(\mathcal{F}_T^c) + \frac{e^{\kappa_T T\epsilon_T^2}}{\Pi(B_\infty(\epsilon_T))}\left(+\sum_{i=M'_T}^{+\infty}\int_{\mathcal{F}_T}\mathbb{E}_0\left[\mathbb{E}_f\left[\mathbb{1}_{\tilde{\Omega}_T}\mathbb{1}_{f\in S_i}(1-\phi)\right]|\mathcal{G}_0\right]\right]d\Pi(f)\Bigg). \tag{S1.2}$$

For the first term on the RHS of (S1.2), we have $\mathbb{P}_0(\tilde{\Omega}_T^c) = o(1)$ by Lemma A.1 in [8]. For the fourth term of the RHS of (S1.2), under **(A0)** and **(A1)**, we have that

$$\frac{e^{\kappa_T T\epsilon_T^2}}{\Pi(B_\infty(\epsilon_T))}\Pi(\mathcal{F}_T^c) \leqslant e^{(\kappa_T + c_1)T\epsilon_T^2}(\Pi(\mathcal{H}_T^c) + \Pi(\Upsilon_T^c)) = o(1).$$

The second term of (S1.2) is controlled by (26) and goes to 0.

We now deal with the third and fifth terms on the RHS of (S1.2), which require to define a suitable test function ϕ. Let $i \in \mathbb{N}, i \geqslant M'_T$ and $f \in S_i$. On $\tilde{\Omega}_T$, with $A_2(T)$ defined in (22), we have that

$$T\tilde{d}_{1T}(f, f_0) = \sum_{l=1}^{K}\int_{A_2(T)}\left|\lambda_l^k(f) - \lambda_l^k(f_0)\right|dt = \sum_{l=1}^{K}\sum_{j=1}^{J_T-1}\int_{\tau_j}^{\xi_j}\left|\lambda_l^k(f) - \lambda_l^k(f_0)\right|dt$$
$$\geqslant \sum_{l=1}^{K}\sum_{j=1}^{J_T-1}\int_{\tau_j}^{U_j^{(1)}}|r_l^f - r_l^0|dt \geqslant \sum_{j=1}^{J_T-1}(U_j^{(1)} - \tau_j)\sum_l|r_l^f - r_l^0| \geqslant \frac{T}{2\,\|r_0\|_1\,\mathbb{E}_0\,[\Delta\tau_1]}\sum_l|r_l^f - r_l^0|,$$

with $r_f = (\phi_1(\nu_1), \ldots, \phi_K(\nu_K))$, $r_0 = (\phi_1(\nu_1^0), \ldots, \phi_K(\nu_K^0))$ and τ_j, ξ_j, $U_j^{(1)}$, $1 \leqslant j \leqslant J_T - 1$ defined in Sections 5.1 and 5.2 of [8]. Consequently, for any $l \in [K]$, since $\tilde{d}_{1T}(f, f_0) \leqslant K(i+1)\epsilon_T$, we obtain that

$$r_l^f \leqslant r_l^0 + 2K(i+1)\,\|r_0\|_1\,\mathbb{E}_0\,[\Delta\tau_1]\,\epsilon_T \leqslant r_l^0 + 1 + 2K\,\|r_0\|_1\,\mathbb{E}_0\,[\Delta\tau_1]\,i\epsilon_T, \tag{S1.3}$$

for T large enough. Moreover, using Assumption 3.1, ϕ_l^{-1} is L'-Lipschitz on $J_l = \phi_l(I_l)$ and $r_l^0 \in J_l$. With $\varepsilon > 0$ from Assumption 3.1, we now separate the set of indices i in two subsets.

Case 1: i is such that $2L'\,\|r_0\|_1\,\mathbb{E}_0\,[\Delta\tau_1]\,K(i+1)\epsilon_T < \varepsilon$. Then we have that $r_l^f \in J_l$ and $\nu_l \in I_l$ since $|r_l^f - r_l^0| =\leqslant 2\,\|r_0\|_1\,\mathbb{E}_0\,[\Delta\tau_1]\,K(i+1)\epsilon_T$. Consequently, $\frac{1}{L'}|\nu_l - \nu_l^0| \leqslant |r_l^f - r_l^0| \leqslant L|\nu_l - \nu_l^0|$ and in particular,

$$\nu_l \leqslant \nu_l^0 + 2KL'(i+1)\,\|r_0\|_1\,\mathbb{E}_0\,[\Delta\tau_1]\,\epsilon_T.$$

Defining

$$\mathcal{F}_i = \left\{f \in \mathcal{F}_T;\ \nu_l^f \leqslant \nu_l^0 + 1 + 2KL'\,\|r_0\|_1\,\mathbb{E}_0\,[\Delta\tau_1]\,i\epsilon_T, \forall l \in [K]\right\},$$

we therefore have that for any $f \in S_i$ and T large enough, $f \in \mathcal{F}_i$ Let $(f_{i,n})_{n=1}^{\mathcal{N}_i}$ be the centering points of a minimal L_1-covering of $\mathcal{F}_i$ by $\mathcal{N}_i$ balls of radius $\zeta i\epsilon_T$ with $\zeta = 1/(6N_0)$, and N_0 defined in the proof of Lemma S5.1 in Section S5.2. There exists $C_0 > 0$ such that we have

$$\mathcal{N}_i \leqslant \left(\frac{C_0(1 + i\epsilon_T)}{\zeta i\epsilon_T/2}\right)^K \mathcal{N}(\zeta i\epsilon_T/2, \mathcal{H}_T, \|.\|_1).$$

If $i\epsilon_T \leqslant 1$,

$$\mathcal{N}_i \leqslant \left(\frac{4C_0}{\zeta i\epsilon_T}\right)^K \mathcal{N}(\zeta i\epsilon_T/2, \mathcal{H}_T, \|.\|_1) = \left(\frac{4C_0}{\zeta}\right)^K e^{-K\log(i\epsilon_T)} \mathcal{N}(\zeta i\epsilon_T/2, \mathcal{H}_T, \|.\|_1).$$

Otherwise, if $i\epsilon_T \geqslant 1$,

$$\mathcal{N}_i \leqslant \left(\frac{4C_0}{\zeta}\right)^K \mathcal{N}(\zeta i\epsilon_T/2, \mathcal{H}_T, \|.\|_1).$$

Moreover, since $i \mapsto \mathcal{N}(\zeta i\epsilon_T/2, \mathcal{H}_T, \|.\|_1)$ is non-increasing, and if $i \geqslant 2\zeta_0/\zeta$, we have that $\mathcal{N}(\zeta i\epsilon_T/2, \mathcal{H}_T \|.\|_1) \leqslant \mathcal{N}(\zeta_0\epsilon_T, \mathcal{H}_T, \|.\|_1) \leqslant e^{x_0 T\epsilon_T^2}$ using **(A2)**. Consequently, since $\epsilon_T > \epsilon_T^2 > \frac{1}{T}$ when T is large enough, $e^{-\log(i\epsilon_T)} \leqslant e^{\log(\frac{\zeta}{2\zeta_0}T)}$ and we obtain

$$\begin{aligned}\mathcal{N}_i &\leqslant \left(\frac{4C_0}{\zeta}\right)^K \left(\frac{\zeta}{2\zeta_0}\right)^K e^{K\log T} \mathcal{N}(\zeta i\epsilon_T/2, \mathcal{H}_T, \|.\|_1) = \left(\frac{2C_0}{\zeta_0}\right)^K e^{K\log T} \mathcal{N}(\zeta i\epsilon_T/2, \mathcal{H}_T, \|.\|_1) \\ &\leqslant C_K e^{K\log T} e^{x_0 T\epsilon_T^2},\end{aligned}$$

denoting $C_K = \left(\frac{2C_0}{\zeta_0}\right)^K$.

Case 2: $2L' \|r_0\|_1 \mathbb{E}_0[\Delta\tau_1] K(i+1)\epsilon_T > \varepsilon$. Then in this case we define $\mathcal{F}_i = \mathcal{F}_T$ and $\nu_T = e^{c_2 T\epsilon_T^2}$, and the L_1-covering number of $\mathcal{F}_i$ is now upper bounded by

$$\mathcal{N}_i \leqslant \left(\frac{\nu_T}{\zeta i\epsilon_T/2}\right)^K \mathcal{N}(\zeta i\epsilon_T/2, \mathcal{H}_T, \|.\|_1) \leqslant C_0' e^{(x_0+c_2K)T\epsilon_T^2},$$

with $C_0' > 0$ a constant.

In both cases, considering the tests $\phi_i = \max_{n\in[\mathcal{N}_i]} \phi_{f_{i,n}}$ with $\phi_{f_{i,n}}$, $\gamma_1 = \min_l x_{1l}$ defined in Lemma S5.1, and $C_K' = C_K \vee C_0'$, $x_0' = x_0 + c_K$, we have

$$\begin{aligned}&\mathbb{E}_0[\mathbb{1}_{\tilde{\Omega}_T}\phi_i] \leqslant \mathcal{N}_i e^{-\gamma_1 T(i^2\epsilon_T^2 \wedge i\epsilon_T)} \leqslant C_K'(2K+1)e^{K\log T} e^{x_0' T\epsilon_T^2} e^{-\gamma_1 T(i^2\epsilon_T^2\wedge i\epsilon_T)}, \\ &\mathbb{E}_0\left[\mathbb{E}_f\left[\mathbb{1}_{\tilde{\Omega}_T}\mathbb{1}_{f\in S_i}(1-\phi_i)|\mathcal{G}_0\right]\right] \leqslant (2K+1)e^{-\gamma_1 T(i^2\epsilon_T^2\wedge i\epsilon_T)}.\end{aligned}$$

Choosing $\phi = \max_{M_T'\leqslant i\leqslant \mathcal{N}_i} \phi_i$ and since $M_T' \geqslant 2\zeta_0/\zeta$ for T large enough, we obtain

$$\begin{aligned}\mathbb{E}_0[\mathbb{1}_{\tilde{\Omega}_T}\phi] &\leqslant C_K'(2K+1)e^{K\log T} e^{x_0' T\epsilon_T^2}\left[\sum_{i=M_T'}^{\epsilon_T^{-1}} e^{-\gamma_1 i^2 T\epsilon_T^2} + \sum_{i>\epsilon_T^{-1}} e^{-\gamma_1 iT\epsilon_T}\right] \\ &\leqslant C_K'(2K+1)e^{K\log T} e^{x_0' T\epsilon_T^2}\left[\sum_{i=M_T'}^{\epsilon_T^{-1}} e^{-\gamma_1 iM_T' T\epsilon_T^2} + \sum_{i>\epsilon_T^{-1}} e^{-\gamma_1 Ti\epsilon_T}\right] \\ &\leqslant C_K'(2K+1)e^{K\log T} e^{x_0' T\epsilon_T^2}\left[2e^{-\gamma_1 M_T'^2 T\epsilon_T^2} + 2e^{-\gamma_1 T}\right] \\ &\leqslant 4C_K'(2K+1)[e^{-\gamma_1 M_T'^2 T\epsilon_T^2} + e^{-\gamma_1 T}], \qquad \text{(S1.4)}\end{aligned}$$

since $\log^3 T = O(T\epsilon_T^2)$ by assumption. Therefore, we arrive at $\mathbb{E}_0[\mathbb{1}_{\tilde{\Omega}_T}\phi] = o(1)$. Similarly, we can obtain

$$\mathbb{E}_0\left[\sum_{i\geqslant M_T'}\int_{\mathcal{F}_T}\mathbb{E}_f\left[\mathbb{1}_{\tilde{\Omega}_T}\mathbb{1}_{f\in S_i}(1-\phi)|\mathcal{G}_0\right]d\Pi(f)\right] \leqslant (2K+1)\left[\sum_{i=M_T'}^{\epsilon_T^{-1}} e^{-\gamma_1 i^2 T\epsilon_T^2} + \sum_{i>\epsilon_T^{-1}} e^{-\gamma_1 T i\epsilon_T}\right]$$
$$\leqslant 4(2K+1)[e^{-\gamma_1 M_T'^2 T\epsilon_T^2} + e^{-\gamma_1 T}].$$

Therefore, using **(A0)**, we have for the second term in (S1.2),

$$\frac{e^{\kappa_T T\epsilon_T^2}}{\Pi(B_\infty(\epsilon_T))}\left(\sum_{i=M_T'}^{+\infty}\int_{\mathcal{F}_T}\mathbb{E}_0\left[\mathbb{E}_f\left[\mathbb{1}_{\tilde{\Omega}_T}\mathbb{1}_{f\in S_i}(1-\phi)|\mathcal{G}_0\right]\right]d\Pi(f)\right) \leqslant \frac{e^{\kappa_T T\epsilon_T^2}}{e^{-c_1 T\epsilon_T^2}} 4(2K+1)[e^{-\gamma_1 M_T'^2 T\epsilon_T^2} + e^{-\gamma_1 T}]$$
$$\leqslant 4(2K+1)e^{-\gamma_1 M_T'^2 T\epsilon_T^2/2} = o(1), \tag{S1.5}$$

for $M_T' > \sqrt{c_1+\kappa_T}$, which holds true if $M_T' = M'\sqrt{\kappa_T}$ with M' large enough. Aggregating the upper bounds previously obtained, we can finally conclude that

$$\mathbb{E}_0[\Pi(A_{d_1}(M_T'\epsilon_T)^c|N)] \leqslant \mathbb{P}_0(\tilde{\Omega}_T^c) + o(1) = o(1),$$

which terminates the proof of Theorem 5.5.

S1.2. Proof of Lemma A.2

In this section, we prove a control on the log-likelihood ratio of the form $\mathbb{P}_0\left[L_T(f_0) - L_T(f) \geqslant 5z_T\right] = o(1)$, where $z_T = T\epsilon_T^2(\log T)^r$ where $r = 0, 1, 2$ is defined in Lemma S6.3 and depends on the assumptions on the link function. We have

$$L_T(f_0) - L_T(f) = \sum_k \int_0^T \log\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right) dN_t^k - \int_0^T (\lambda_t^k(f_0) - \lambda_t^k(f))dt$$
$$= W_0 + \sum_{j=1}^{J_T-1} T_j + W_T,$$

with

$$W_0 := \sum_k \int_0^{\tau_1} \log\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right) dN_t^k - \int_0^{\tau_1} (\lambda_t^k(f_0) - \lambda_t^k(f))dt,$$
$$W_T := \sum_k \int_{\tau_{J_T}}^T \log\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right) dN_t^k - \int_{\tau_{J_T}}^T (\lambda_t^k(f_0) - \lambda_t^k(f))dt.$$

Let $\mathcal{L}_T = L_T(f_0) - L_T(f) - \mathbb{E}_0[L_T(f_0) - L_T(f)] = L_T(f_0) - L_T(f) - KL(f_0, f)$, with $KL(f_0, f)$ the Kullback-Leibler divergence defined in (S6.19). Then

$$\begin{aligned}
&\mathbb{P}_0[\mathcal{L}_T \geqslant 4z_T] = \mathbb{P}_0\left[\sum_{j=1}^{J_T-1} T_j + W_0 + W_T - KL(f_0, f) \geqslant 4z_T\right] \\
&= \mathbb{P}_0\left[\sum_{j=1}^{J_T-1}(T_j - \mathbb{E}_0[T_j]) + \sum_{j=1}^{J_T-1}\mathbb{E}_0[T_j] - \mathbb{E}_0\left[\sum_{j=1}^{J_T-1} T_j\right] + W_T - \mathbb{E}_0[W_T] + W_0 - \mathbb{E}_0[W_0] \geqslant 4z_T\right] \\
&\leqslant \mathbb{P}_0\left[\sum_{j=1}^{J_T-1} T_j - \mathbb{E}_0[T_j] \geqslant z_T\right] + \mathbb{P}_0\left[(J_T - \mathbb{E}_0[J_T])\mathbb{E}_0[T_1] - \mathbb{E}_0\left[\sum_{j=0}^{J_T-1} T_j - \mathbb{E}_0[T_j]\right] \geqslant z_T\right] + \mathbb{P}_0[W_T - \mathbb{E}_0[W_T] \geqslant z_T] \\
&+ \mathbb{P}_0[W_0 - \mathbb{E}_0[W_0] \geqslant z_T],
\end{aligned} \tag{S1.6}$$

using equation (S6.21) and that

$$\begin{aligned}
KL(f_0, f) = &\underbrace{\sum_k \mathbb{E}_0\left[\int_{\tau_0}^{\tau_1} \log\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right) dN_t^k - \int_0^{\tau_1} (\lambda_t^k(f_0) - \lambda_t^k(f))dt\right]}_{\mathbb{E}_0[W_0]} \\
&+ \underbrace{\sum_k \mathbb{E}_0\left[\int_0^{\tau_{J_T}} \log\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right) dN_t^k - \int_0^{\tau_{J_T}} (\lambda_t^k(f_0) - \lambda_t^k(f))dt\right]}_{=\mathbb{E}_0\left[\sum_{j=1}^{J_T-1} T_j\right]} \\
&+ \underbrace{\sum_k \mathbb{E}_0\left[\int_{\tau_{J_T}}^{T} \log\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right) dN_t^k - \int_{\tau_{J_T}}^{T} (\lambda_t^k(f_0) - \lambda_t^k(f))dt\right]}_{\mathbb{E}_0[W_T]}.
\end{aligned}$$

From Lemma S6.3, we have that $\mathbb{P}_0\left[\sum_{j=1}^{J_T-1} T_j - \mathbb{E}_0[T_j] \geqslant z_T\right] = o(1)$. We now deal with the second term on the RHS of (S1.6). Using Lemma S6.3, we have

$$\begin{aligned}
\mathbb{E}_0\left[\sum_{j=1}^{J_T-1} T_j - \mathbb{E}_0[T_j]\right] &= \mathbb{E}_0\left[\sum_{j=\lfloor T/\mathbb{E}_0[\Delta\tau_1]\rfloor}^{J_T-1} T_j - \mathbb{E}_0[T_j]\right] \\
&\leqslant \mathbb{E}_0\left[\sum_{J\in\mathcal{J}_T} \mathbb{1}_{J_T=J}\left(\sum_{j=\lfloor T/\mathbb{E}_0[\Delta\tau_1]\rfloor}^{J-1} |T_j - \mathbb{E}_0[T_j]|\right)\right] + \sqrt{\mathbb{P}_0[J_T \notin \mathcal{J}_T]}\sqrt{T^2\mathbb{E}_0[T_1^2]} \\
&\leqslant \mathbb{E}_0\left[\sum_{j=\lfloor \frac{T}{\mathbb{E}_0[\Delta\tau_1]}(1-c_\beta\sqrt{\frac{\log T}{T}})\rfloor}^{\lfloor \frac{T}{\mathbb{E}_0[\Delta\tau_1]}(1+c_\beta\sqrt{\frac{\log T}{T}})\rfloor} |T_j - \mathbb{E}_0[T_j]|\right] + T^{1-\beta/2}\sqrt{\mathbb{E}_0[T_1^2]} \\
&\leqslant \frac{2c_\beta}{\mathbb{E}_0[\Delta\tau_1]}\mathbb{E}_0\left[|T_1 - \mathbb{E}_0[T_j]|\right]\sqrt{T\log T} + T^{1-\beta/2}\sqrt{\mathbb{E}_0[T_1^2]}
\end{aligned}$$

$$\lesssim \sqrt{\mathbb{E}_0\left[T_1^2\right]}\sqrt{T\log T} \lesssim \sqrt{T}(\log T)^{3/2}\epsilon_T = o(z_T),$$

since $\log^3 T = O(z_T)$ by assumption. Consequently,

$$\mathbb{P}_0\left[(J_T - \mathbb{E}_0[J_T])\mathbb{E}_0[T_1] - \mathbb{E}_0\left[\sum_{j=0}^{J_T-1} T_j - \mathbb{E}_0\left[T_j\right]\right] \geqslant z_T\right] \leqslant \mathbb{P}_0\left[J_T - \mathbb{E}_0[J_T] \geqslant \frac{z_T}{2\mathbb{E}_0[T_1]}\right]$$
$$\leqslant \mathbb{P}_0\left[J_T - \frac{T}{\mathbb{E}_0[\Delta\tau_1]} \geqslant \frac{z_T}{4\mathbb{E}_0[T_1]}\right],$$

using that $J_T - \mathbb{E}_0[J_T] = J_T - \frac{T}{\mathbb{E}_0[\Delta\tau_1]} + \frac{T}{\mathbb{E}_0[\Delta\tau_1]} - \mathbb{E}_0[J_T]$ and $\frac{T}{\mathbb{E}_0[\Delta\tau_1]} - \mathbb{E}_0[J_T] \leqslant \frac{z_T}{4\mathbb{E}_0[T_1]}$ for T large enough. Consequently, since $\mathbb{E}_0[T_1] \leqslant \sqrt{\frac{z_T}{T}}$, we have with $\eta_T = \sqrt{\frac{z_T}{4\mathbb{E}_0[T_1]}}$ and $B_j = \tau_j - \tau_{j-1} - \mathbb{E}_0[\Delta\tau_1]$, and using the computations as for the proof of Lemma A.1,

$$\mathbb{P}_0\left[J_T - \frac{T}{\mathbb{E}_0[\Delta\tau_1]} \geqslant \eta_T\right] \leqslant \mathbb{P}_0\left[\tau_{\lfloor T/\mathbb{E}_0[\Delta\tau_1]+\eta_T\rfloor} \leqslant T\right]$$
$$= \mathbb{P}_0\left[\sum_{j=1}^{\lfloor T/\mathbb{E}_0[\Delta\tau_1]+\eta_T\rfloor} B_j \leqslant T - \lfloor T/\mathbb{E}_0[\Delta\tau_1] + \eta_T\rfloor\mathbb{E}_0[\Delta\tau_1]\right]$$
$$\leqslant \mathbb{P}_0\left[\sum_{j=1}^{\lfloor T/\mathbb{E}_0[\Delta\tau_1]+\eta_T\rfloor} B_j \leqslant -\mathbb{E}_0[\Delta\tau_1]\eta_T + \mathbb{E}_0[\Delta\tau_1]\right]$$
$$\leqslant \frac{4\lfloor T/\mathbb{E}_0[\Delta\tau_1] + \eta_T\rfloor\mathbb{E}_0\left[\Delta\tau_1^2\right]}{\mathbb{E}_0[\Delta\tau_1]^2\eta_T^2} \lesssim \frac{T}{\eta_T^2} + \frac{1}{\eta_T} \lesssim \frac{1}{z_T} = o(1).$$

For the third term on the RHS of (S1.6), applying Bienayme-Chebyshev's inequality, we have

$$\mathbb{P}_0[W_T - \mathbb{E}_0[W_T] \geqslant z_T] \leqslant \frac{\mathbb{E}_0\left[W_T^2\right]}{z_T^2}. \tag{S1.7}$$

Using similarly computations as in Lemma S6.3, we obtain

$$\mathbb{E}_0\left[W_T^2\right] = \mathbb{E}_0\left[\left(\sum_k \int_{\tau_{J_T}}^T \log\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right)dN_t^k - \int_{\tau_{J_T}}^T (\lambda_t^k(f_0) - \lambda_t^k(f))dt\right)^2\right]$$
$$\lesssim \mathbb{E}_0\left[(T - \tau_{J_T})\int_{\tau_{J_T}}^T \left[\log\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right)\lambda_t^k(f_0) - (\lambda_t^k(f_0) - \lambda_t^k(f))\right]^2 dt\right] + \mathbb{E}_0\left[\int_{\tau_{J_T}}^T \log^2\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right)\lambda_t^k(f_0)dt\right].$$

Then since

$$\mathbb{E}_0\left[(T - \tau_{J_T})\int_{\tau_{J_T}}^T \left[\log\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right)\lambda_t^k(f_0) - (\lambda_t^k(f_0) - \lambda_t^k(f))\right]^2 dt\right] \leqslant \mathbb{E}_0\left[\Delta\tau_1\int_{\tau_1}^{\tau_2} \chi\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right)^2 \lambda_t^k(f_0)^2 dt\right],$$
$$\mathbb{E}_0\left[\int_{\tau_{J_T}}^T \log^2\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right)\lambda_t^k(f_0)dt\right] \leqslant \mathbb{E}_0\left[\int_{\tau_1}^{\tau_2} \log^2\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right)\lambda_t^k(f_0)dt\right],$$

we can use the bounds derived for $\mathbb{E}_0\left[T_j^2\right]$ in Lemma S6.3.

We finally obtain

$$\mathbb{P}_0\left[W_T - \mathbb{E}_0\left[W_T\right] \geqslant z_T\right] \lesssim \frac{(\log^2 T)\epsilon_T^2}{z_T^2} \lesssim \frac{\log^2 T}{T^2\epsilon_T^2} = o(1).$$

With similar computations, we also obtain that $\mathbb{P}_0\left[W_0 - \mathbb{E}_0\left[W_0\right] \geqslant z_T\right] = o(1)$. Consequently, reporting into (S1.6) and using Lemma S6.1, we finally obtain that

$$\mathbb{P}_0\left[L_T(f_0) - L_T(f) > 5z_T\right] \leqslant \mathbb{P}_0\left[\mathcal{L}_T > 5z_T - u_T\right] \leqslant \mathbb{P}_0\left[\mathcal{L}_T > 4z_T\right] = o(1),$$

since $KL(f_0, f) \leqslant u_T \leqslant z_T$ using Lemmas S6.1 and S6.3.

S2. Proof of Theorem 3.11 and Proposition 3.10

S2.1. Proof of Theorem 3.11 (Case 2)

In this section, we prove the second case of the proof of Theorem 3.11 in Section 5.5 of [8]. We recall that in this case we consider $(l,k) \in I(\delta_0)$, i.e., $\delta_{lk}^0 = 1$. We also recall the notation $S_{lk}^0 = \|h_{lk}^0\|_1$ and $M_T = M\sqrt{\kappa_T}$ with $M > 0$.

We first note that if $S_{lk}^0 > M_1\sqrt{\kappa_T}\epsilon_T$ with $M_1 > M$ and $1 - F(S_{lk}^0/2) \geqslant 2e^{-\gamma T\epsilon_T^2}$ for some $\gamma > \kappa_T + c_1 =: \kappa_T'$, then if $\delta_{lk} = 0$, $f \in A_{L_1}(M_1\sqrt{\kappa_T}\epsilon_T)^c$ and

$$\Pi(\delta_{lk} = 0|N) \leqslant \Pi(A_{L_1}(M_T\epsilon_T)^c|N), \quad \text{and} \quad S_{lk}^0 - M_T\epsilon_T \geqslant S_{lk}^0/2.$$

Therefore, since F is non-increasing, $F(S_{lk}^0 - M_T\epsilon_T) \leqslant F(S_{lk}^0/2)$ and

$$\begin{aligned}
&\mathbb{P}_0\left[\hat{\delta}_{lk}^{\Pi,L} = 0\right] \leqslant \mathbb{P}_0\left[\Pi((1 - F(S_{lk}))\mathbb{1}_{\delta=1}(\mathbb{1}_{S_{lk} \geqslant S_{lk}^0 - M_T\epsilon_T} + \mathbb{1}_{S_{lk} < S_{lk}^0 - M_T\epsilon_T})|N) \leqslant \Pi(A_{L_1}(2M_T\epsilon_T)^c|N)\right] \\
&\leqslant \mathbb{P}_0\left[(1 - F(S_{lk}^0/2))\Pi(S_{lk} > S_{lk}^0 - M_T\epsilon_T|N) + \Pi((1 - F(S_{lk}))\mathbb{1}_{S_{lk} < S_{lk}^0 - M_T\epsilon_T})|N) \leqslant \Pi(A_{L_1}(M_1\sqrt{\kappa_T}\epsilon_T)^c|N)\right] \\
&\leqslant \mathbb{P}_0\left[2e^{-\gamma T\epsilon_T^2}\Pi(S_{lk} > S_{lk}^0 - M_T\epsilon_T|N) \leqslant \Pi(A_{L_1}(M_1\sqrt{\kappa_T}\epsilon_T)^c|N)\right] \\
&\leqslant \mathbb{P}_0\left[\Pi(S_{lk} > S_{lk}^0 - M_T\epsilon_T|N) \leqslant 1/2\right] + \mathbb{P}_0\left[\tilde{\Omega}_T \cap \left\{e^{-\gamma T\epsilon_T^2} \leqslant \Pi(A_{L_1}(M_1\sqrt{\kappa_T}\epsilon_T)^c|N)\right\}\right] + \mathbb{P}_0\left[\tilde{\Omega}_T^c\right].
\end{aligned}$$

Similar to the first case where $\delta_{lk}^0 = 0$, we have that $\mathbb{P}_0\left[\tilde{\Omega}_T \cap \left\{e^{-\kappa_T'\epsilon_T^2} \leqslant \Pi(A_{L_1}(M_T\epsilon_T)^c|N)\right\}\right] = o(1)$, and since $\gamma \geqslant \kappa_T'$,

$$\begin{aligned}
\mathbb{P}_0\left[\hat{\delta}_{lk}^{\Pi,L} = 0\right] &\leqslant \mathbb{P}_0\left[\Pi(S_{lk} > S_{lk}^0 - M_T\epsilon_T|N) \leqslant 1/2\right] + o(1) = \mathbb{P}_0\left[\Pi(S_{lk} < S_{lk}^0 - M_T\epsilon_T|N) > 1/2\right] + o(1) \\
&\leqslant \mathbb{P}_0\left[\tilde{\Omega}_T \cap \{\Pi(A_{L_1}(M_1\sqrt{\kappa_T}\epsilon_T)^c|N) > 1/2\}\right] + \mathbb{P}_0\left[\tilde{\Omega}_T^c\right] = o(1),
\end{aligned}$$

which terminates this proof.

S2.2. Proof of Proposition 3.10

In this section, we prove our posterior consistency result on the posterior distribution in the restricted models, the **All equal model** and **Receiver dependent model**, defined in Section 3.2 of [8].

In the **All equal model**, if $I(\delta_0) \neq \emptyset$ then $\exists (l_1, k_1) \in [K]^2, \delta^0_{l_1 k_1} = 1$, and $h_0 \neq 0$. Consequently, for T large enough,

$$\{f \in \mathcal{F};\ \delta_{l_1 k_1} \neq \delta^0_{l_1 k_1}\} = \left\{f \in \mathcal{F};\ \delta_{l_1 k_1} = 0\right\} \subset \left\{f \in \mathcal{F};\ \|h^0_{l_1 k_1} - h_{l_1 k_1}\|_1 = \|h_0\|_1\right\} \subset A_{L_1}(M_T \epsilon_T)^c,$$

leading to $\mathbb{E}_0\left[\Pi(\delta_{l_1 k_1} \neq \delta^0_{l_1 k_1} | N)\right] = o(1)$ using Theorem 3.2. This would hold for the same reasons for any $(l, k) \in I(\delta_0)$. For $(l, k) \notin I(\delta_0)$, we have instead that for T large enough,

$$\begin{aligned}\{f \in \mathcal{F};\ \delta_{lk} \neq \delta^0_{lk}\} = \{f \in \mathcal{F};\ \delta_{lk} = 1\} &\subset \{f \in \mathcal{F};\ \left\|h^0_{lk} - h_{lk}\right\|_1 = \|h\|_1\} \\ &\subset \{f \in \mathcal{F};\ \|h\|_1 + \left\|h^0_{l_1 k_1} - h_{l_1 k_1}\right\|_1 \geqslant \|h_0\|_1\} \subset A_{L_1}(M_T \epsilon_T)^c,\end{aligned}$$

as soon as $\|h_0\|_1 \geqslant 3 M_T \epsilon_T$, since $\|h\|_1 + \left\|h^0_{l_1 k_1} - h_{l_1 k_1}\right\|_1 \geqslant \|h\|_1 + \|h_0\|_1 \wedge \left\|h - h^0\right\|_1 \geqslant (\|h\|_1 + \|h_0\|_1) \wedge (\|h\|_1 + \|h - h_0\|_1) \geqslant \|h_0\|$. Similarly to the proof of Theorem 3.9 in Section 5.4, we then obtain $\mathbb{E}_0\left[\Pi(\delta \neq \delta_0 | N)\right] = o(1)$.

If $I(\delta_0) = \emptyset$, then $\forall (l, k) \in [K]^2, \delta^0_{lk} = 0$, and $h_0 = 0$, and in this case we first show that there exists $C > 0$ such that

$$\mathbb{P}_0\left[\left\{D_T < C T^{-K/2}\right\} \cap \tilde{\Omega}_T\right] = o(1). \tag{S2.8}$$

Since $h_0 = 0$, the log-likelihood function is the one of a K independent homogeneous Poisson PP with parameter r_0, i.e.,

$$L_T(f_0) = L_T(r_0) = \sum_k \log(r^0_k) N^k[0, T) - r^0_k T,$$

with $r^0_k = \phi_k(\nu^0_k)$. Let $\bar{A} = \{f \in \mathcal{F}_T;\ h = 0\}$. For any $f \in \bar{A}$, we also have $L_T(f) = L_T(r_f) = \sum_k \log(r^f_k) N^k[0, T) - r^f_k T$ and the model is also a Poisson PP, which is a regular model, and which parameter is $\phi(\nu)$. Therefore, we have

$$\begin{aligned}L_T(r) - L_T(r_0) &= \sum_k \log(\frac{r_k}{r^0_k}) N^k[0, T) - (r^f_k - r^0_k) T \\ &= \sum_k \left[\frac{r^f_k - r^0_k}{r^0_k} - \frac{1}{2}\left(\frac{r^f_k - r^0_k}{r^0_k}\right)^2 + O_{\mathbb{P}_0}(r^f_k - r^0_k)^3\right] N^k[0, T) - (r^f_k - r^0_k) T \\ &= \sum_k \left(\frac{N^k[0, T)}{r^0_k} - T\right)(r^f_k - r^0_k) - \frac{N^k[0, T)}{2}\left(\frac{r^f_k - r^0_k}{r^0_k}\right)^2 + O_{\mathbb{P}_0}(T(r^f_k - r^0_k)^3).\end{aligned}$$

Also, let $\tilde{\pi}_r$ be the prior density of $r^f_k = \phi_k(\nu_k)$ given by $\tilde{\pi}_r(x) = \phi(\nu)\pi_\nu(\nu)$. Note that in the case of partially known link functions of the form $\phi_k(x) = \theta_k + \psi(x)$, the parameter of the Poisson PP is now (ν, θ) and we can consider a marginal prior density of $r^f_k = \theta_k + \psi(\nu_k)$ given by

$$\tilde{\pi}_r(x) = \int_0^{\psi^{-1}(x)} \pi_\theta(x - \psi(\nu))\pi_\nu(\nu) d\nu.$$

The regularity assumptions on π_ν (and π_θ) and ϕ^{-1} imply that $\tilde{\pi}_r$ is continuous and positive at r_k^0 for all k.

Defining $\bar{A}_T = \bar{A} \cap \{\|r_f - r_0\|_1 \leqslant \epsilon\}$ for $\epsilon > 0$ small enough, we thus have

$$
\begin{aligned}
D_T &= \int_{\mathcal{F}_T} e^{L_T(f)-L_T(f_0)} d\Pi(f) \geqslant \int_{\bar{A}_T} e^{L_T(r)-L_T(r_0)} d\Pi(f) \\
&\geqslant \int_{\bar{A}_T} \prod_{k=1}^{K} \exp\left\{ \left(\frac{N^k[0,T)}{r_k^0} - T \right)(r_k^f - r_k^0) - \frac{N^k[0,T)}{2} \left(\frac{r_k^f - r_k^0}{r_k^0} \right)^2 (1+\epsilon) \right\} \tilde{\pi}(r_k) dr_k \\
&= \prod_{k=1}^{K} \tilde{\pi}_r(r_k^0)(1 + o_{\mathbb{P}_0}(1)) e^{\frac{r_k^0}{2(1+\epsilon)N^k[0,T)} \left(\frac{N^k[0,T)}{r_k^0} - T \right)^2} \times \\
&\quad \int_{|r_k^f - r_k^0| \leqslant \epsilon/K} \exp\left\{ -\frac{N^k[0,T)}{2(r_k^0)^2}(1-\epsilon) \left(r_k^f - r_k^0 - \frac{(r_k^0)^2}{(1+\epsilon)N^k[0,T)} \left(\frac{N^k[0,T)}{r_k^0} - T \right) \right)^2 \right\} dr_k^f \\
&\geqslant \prod_{k=1}^{K} \tilde{\pi}_r(r_k^0) r_k^0 \frac{\sqrt{2\pi}}{[N^k[0,T)(1+\epsilon)]^{1/2}} (1 + o_{\mathbb{P}_0}(1)) \geqslant \prod_{k=1}^{K} \frac{\sqrt{2\pi}\tilde{\pi}_r(r_k^0) r_k^0}{[T(1+\epsilon)]^{1/2}} (1 + o_{\mathbb{P}_0}(1)),
\end{aligned}
$$

since $N^k[0,T)$ is a Poisson random variable with parameter $r_k^0 T$ so that $|N^k[0,T)/T - r_k^0| \leqslant M_T/\sqrt{T}$ with probability going to 1 and $\{|r_k^f - r_k^0| \leqslant \epsilon/K\}$ contains the set

$$
\left| r_k^f - r_k^0 - \frac{(r_k^0)^2}{(1-\epsilon)N^k[0,T)} \left(\frac{N^k[0,T)}{r_k^0} - T \right) \right| \leqslant \frac{\epsilon}{2K},
$$

for T large enough. Therefore we obtain (S2.8) and deduce that $\epsilon_T \lesssim \sqrt{\log T/T}$ using the same arguments as in the proofs of Theorem 3.2. As in Theorem 3.2, it is thus sufficient that

$$
\begin{aligned}
&\Pi(\{0 < \|h\|_1 \leqslant M\sqrt{\log T/T}\} \cap \{\max_k |r_k^f - r_k^0| \leqslant M\sqrt{\log T/T}\}) \\
&\leqslant \Pi(\{0 < \|h\|_1 \leqslant M\sqrt{\log T/T}\} \cap \{\max_k |\nu_k - \nu_k^0| \leqslant \frac{M}{L}\sqrt{\log T/T}\}) = o(T^{-K/2}),
\end{aligned}
$$

for M large enough which boils down to assuming that

$$
\Pi(\{0 < \|h\|_1 \leqslant M\sqrt{\log T/T}\}) = o((\log T)^{-K/2}),
$$

to conclude that $\mathbb{E}_0[\Pi(\delta \neq \delta_0 | N)] = o(1)$.

In the **Receiver node dependent model**, i.e., $\forall (l,k) \in [K]^2, h_{lk} = \delta_{lk} h_k$, we obtain the result similarly to the **All equal model** since the likelihood is also a product of likelihoods per node:

$$
\begin{aligned}
L_T(f) &= \sum_{k=1}^{K} L_T(\nu_k, h_k, \delta(k), \theta_k), \qquad \text{with } \delta(k) := (\delta_{lk}, 1 \leqslant l \leqslant K), \\
L_T(\nu_k, h_k, \delta(k), \theta_k) &:= \sum_{T_i^k} \log \lambda_{T_i^k}^k(f_k) - \int_0^T \lambda_t^k(f_k) dt, \quad f_k = (\nu_k, h_k, \delta(k), \theta_k), \quad k \in [K].
\end{aligned}
$$

If the priors on $(\theta_k, \nu_k, h_k, \delta(k))$ are independent, the posteriors are also independent and we can directly apply the previous result.

S3. Proof of Corollary 3.8

In this section we prove our result on the convergence rate of the posterior mean estimator. In all the considered models with known link functions, the convergence of the posterior mean $\hat{f} = (\hat{\nu}, \hat{h})$ results from the same arguments as in Corollary 1 of [2] (proof in Section 2.3 in the supplementary material). In the case of the shifted ReLU model with unknown shift, we can also use similar computations for $\hat{f} = (\hat{\nu}, \hat{h})$ and $\hat{\theta}$. We first recall some notation from the proofs of Theorem 3.2 and Proposition 3.5: $\bar{A}(\tilde{M}_T\epsilon_T) = \{\theta \in \Theta, \|\theta - \theta_0\|_1 < \tilde{M}_T\epsilon_T\}$, $A_{L_1}(M_T\epsilon_T) = \{(f,\theta) \in \mathcal{F} \times \Theta, \|\theta + \nu - \theta_0 - \nu_0\|_1 + \|h - h_0\|_1 < M_T\epsilon_T\}$ and $\tilde{M}_T = \tilde{M}\sqrt{\kappa_T}$, $M_T = M\sqrt{\kappa_T}$, $\tilde{M} > M > 0$. We note that

$$\left\|\hat{\theta} - \theta_0\right\|_1 \leqslant \tilde{M}_T\epsilon_T + \mathbb{E}^{\Pi}[\|\theta - \theta_0\|_1 \mathbb{1}_{\bar{A}(\tilde{M}_T\epsilon_T)^c}|N].$$

Then, splitting $\bar{A}(\tilde{M}_T\epsilon_T)^c \times \mathcal{F}_T$ into $\bar{A}(\tilde{M}_T\epsilon_T)^c \times \mathcal{F}_T \cap A_{L_1}(M_T\epsilon_T)$ and $\bar{A}(\tilde{M}_T\epsilon_T)^c \times \mathcal{F}_T \cap A_{L_1}(M_T\epsilon_T)^c$, we control $\mathbb{E}^{\Pi}[\|\theta - \theta_0\|_1 \mathbb{1}_{B_T}|N]$ using the following arguments with B_T representing either $\bar{A}(\tilde{M}_T\epsilon_T)^c \times \mathcal{F}_T \cap A_{L_1}(M_T\epsilon_T)$ or $A_{L_1}(M_T\epsilon_T)^c$. Using the decomposition (24), with $\kappa'_T = \kappa_T + c_1$, we have

$$\begin{aligned}
\mathbb{P}_0\left[\mathbb{E}^{\Pi}[\|\theta - \theta_0\|_1 \mathbb{1}_{B_T}|N] > \epsilon_T\right] &\leqslant \mathbb{E}_0\left[\phi \mathbb{1}_{\tilde{\Omega}_T}\right] + \mathbb{P}_0\left[\{D_T < e^{-\kappa'_T T\epsilon_T^2}\} \cap \tilde{\Omega}_T\right] + \mathbb{P}_0\left[\tilde{\Omega}_T^c\right] + \frac{e^{\kappa'_T T\epsilon_T^2}}{\epsilon_T}\Pi(\mathcal{F}_T^c) \\
&\quad + \frac{e^{\kappa'_T T\epsilon_T^2}}{\epsilon_T}\int_{\mathcal{F}_T \cap B_T} \|\theta - \theta_0\|_1 \, \mathbb{E}_0\left[\mathbb{E}_f\left[(1-\phi)\mathbb{1}_{\tilde{\Omega}_T}\right]\middle|\mathcal{G}_0\right] d\Pi(f,\theta) \\
&\leqslant o(1) + o\left(\int_{\mathcal{F}_T \cap B_T} \|\theta - \theta_0\|_1 \, d\Pi(f,\theta)\right) = o(1),
\end{aligned}$$

using the tests defined In Lemma A.5 if $B_T = \bar{A}(\tilde{M}_T\epsilon_T)^c \times \mathcal{F}_T \cap A_{L_1}(M_T\epsilon_T)$ or the tests defined in Lemma S5.1 if $B_T = A_{L_1}(M_T\epsilon_T)^c$, and also that $\log T = o(T\epsilon_T^2)$ to obtain that $\frac{e^{\kappa'_T T\epsilon_T^2}}{\epsilon_T}\Pi(\mathcal{F}_T^c) \leqslant \Pi(\mathcal{H}_T^c)e^{\kappa'_T T\epsilon_T^2 - \log\epsilon_T} = o(1)$, whichs terminates this proof.

S4. Proofs of some results on prior distributions

In this section, we present an alternative construction of the prior distribution using mixtures of Beta distributions and the proof of Lemma 4.3, which gives one example of model where the condition (8) can be verified.

S4.1. Mixtures of Betas priors

This family of prior distributions can be also considered alongside the ones presented in Section 4 of [8]. The following construction is similar to Section 2.3.2 of [2], which is based on [7]. Using the hierarchical structure (15) from Section 4 , we define $\pi_{\bar{h}}$ as follows. For simplicity, we here consider that $A = 1$. Let

$$\tilde{h}_{\alpha,M}(x) = \int_u g_{\alpha,u}(x) dM(u), \quad g_{\alpha,u}(x) = \frac{\Gamma(\alpha/u(1-u))}{\Gamma(\alpha/u)\Gamma(\alpha/(1-u))} x^{-\alpha/(1-u)-1}(1-x)^{-\alpha/u-1},$$

and $\pi_{\bar{h}}$ be the push forward distribution of $\Pi_\alpha \times \Pi_M$ by the transformation $(\alpha, M) \to h_{\alpha,M}$, where Π_α and Π_M are respectively the probability distribution on α and M. Therefore $\pi_{\bar{h}}$ is a bounded signed measure on $[0,1]$. As in [2], we choose $\sqrt{\alpha}$ to follow a Gamma distribution and define Π_M by

$$M(u) = \sum_{j=1}^{J} r_j p_j \delta_{u_j}(u), \quad u_j \overset{i.i.d.}{\sim} G_0, \quad J \sim \mathcal{P}(\lambda),$$

where G_0 is a base measure and the r_j's are independent Rademacher random variables and $(p_1, \cdots, p_J) \sim \mathcal{D}(a_1, \cdots, a_J)$ with $\sum_{j=1}^{J} a_j \leqslant C$ for some fixed $C > 0$. Note that since $\|\bar{h}_{\alpha,M}\|_1 \leqslant 1$, we can define

$$h_{lk} = \tilde{S}_{lk} \tilde{h}_{lk}, \quad \|\tilde{S}^+\| \leqslant 1, \quad \tilde{h}_{lk} \overset{i.i.d.}{\sim} \pi_{\bar{h}},$$

so that the prior distribution on h is the push forward distribution of $\pi_{\bar{h}}^{\otimes |I(\delta)|} \times \pi_{S|\delta}$ by the above transformation, with π_S defined in **(S2)** in Section 4 of [8]. Since $\tilde{S}$ is a (component-wise) upper bound on the matrix S, $\|\tilde{S}^+\| \leqslant 1$ implies $\|S^+\| \leqslant 1$. We then arrive at the following result.

Corollary S4.1. *Let N be a Hawkes process with link functions $\phi = (\phi_k)_k$ and parameter $f_0 = (\nu_0, h_0)$ such that (ϕ, f_0) verify the conditions of Lemma 2.1, and Assumption 3.1. Under the above spline prior, if the prior on S satisfies the conditions defined in **(S1)** (Section 4 of [8]), and also if $\forall (l,k) \in [K]^2$, $h_{lk}^0 \in \mathcal{H}(\beta, L)$ with $\beta > 0$ and $\|S_0^+\| < 1$ then for M large enough,*

$$\mathbb{E}_0\left[\Pi(\|f - f_0\|_1 > MT^{-\beta/(2\beta+1)} \sqrt{\log\log T} (\log T)^q \big| N)\right] = o(1),$$

where $q = 5\beta/(4\beta+2)$ if ϕ verifies Assumption 3.1(i) , and $q = 1/2 + 5\beta/(4\beta+2)$ if ϕ verifies Assumption 3.1(ii) .

S4.2. Proof of Lemma 4.3

Lemma S4.2 (Lemma 4.3)**.** *Let N be a Hawkes process with ReLU link functions $\phi_k(x) = (x)_+, \forall k \in [K]$, and parameter $f_0 = (\nu_0, h_0)$ such that (ϕ, f_0) verify condition **(C1bis)** and for all l, there exists $J_0 \in \mathbb{N}^*$ such that*

$$h_{lk}^0(t) = \sum_{j=1}^{J_0} \omega_{j0}^{lk} \mathbb{1}_{I_j}(t), \quad \omega_{j0}^{lk} \in \mathbb{Q}, \quad \forall j \in [J_0],$$

with $\{I_j\}_{j=1}^{J_0}$ a partition of $[0, A]$. Then, condition (8) of Proposition 3.5 holds, i.e.,

$$\limsup_{T \to \infty} \frac{1}{T} \mathbb{E}_0\left(\int_0^T \frac{\mathbb{1}_{\lambda_t^k(f_0) > 0}}{\lambda_t^k(f_0)} dt \right) < +\infty, \quad k \in [K].$$

Proof. Let f_0 verifying the conditions of the lemma. We first show that there exists $c_0 > 0$ that depends only on the parameters $\{\nu_k^0, \{\omega_{j0}^{kl}\}_{j=1}^{J}\}_{k,l=1}^{K}$ such that $\forall k \in [K], \forall t \geqslant 0$, $\lambda_t^k(f_0) > 0 \implies \lambda_t^k(f_0) \geqslant c_0$. We prove here the result for the unidimensional Hawkes model with $K = 1$, but our proof can be easily generalized to $K > 1$. We therefore use the notation ν_0 and w_{j0} for ν_1^0 and w_{j0}^{11}.

Since $w_{j0} \in \mathbb{Q}$, let $p_j, q_j \in \mathbb{Z}$ such that $w_{j0} = p_j/q_j$ and let $q \in \mathbb{Z}$ be the least common multiple of (p_j, q_j). Thus there exists $a_j \in \mathbb{Z}$ such that $\omega_{j0} = a_j/q$. and for any $t \geqslant 0$, let $n_j(t) = \int_{t-A}^{t} \mathbb{1}_{I_j}(t-s) dN_s$

be the number of events that "activate" the bin j at t. With this notation, we can then write

$$\lambda_t(f_0) = \left(\nu_0 + \sum_{j=1}^{J_0} n_j(t)\frac{a_j}{q}\right)_+ = \left(\nu_0 + \sum_{j=1}^{J_0} n_j(t)\frac{a_j}{q}\right)_+$$
$$= \left(\frac{1}{q}\left[\nu_0 q + \sum_{j=1}^{J_0} n_j(t)a_j\right]\right)_+ .$$

Let $\varepsilon > 0$ such that $\varepsilon = \min_{u\in\mathbb{Z},\nu_0 q+u>0} \nu_0 q + u$. Then $\varepsilon > 0$ and for any $t \geqslant 0$ such that $\tilde{\lambda}_t(f_0) > 0$, since $\sum_{j=1}^{J_0} n_j(t)a_j \in \mathbb{Z}$, then $\nu_0 q + \sum_{j=1}^{J_0} n_j(t)a_j \geqslant \varepsilon > 0$ and $\lambda_t(f_0) \geqslant \varepsilon/q =: c_0 > 0$, which proves that (i) holds. Therefore, in this model, we have

$$\frac{1}{T}\mathbb{E}_0\left(\int_0^T \frac{\mathbb{1}_{\lambda_t(f_0)>0}}{\lambda_t(f_0)}dt\right) \leqslant \frac{1}{T}\mathbb{E}_0\left(\int_0^T \frac{\mathbb{1}_{\lambda_t(f_0)>0}}{c_0}dt\right) \leqslant \frac{1}{T}\mathbb{E}_0\left(\int_0^T \frac{1}{c_0}dt\right) = \frac{1}{c_0} < +\infty,$$

which proves that (8) is satisfied. □

***Remark* S4.3.** We could similarly show that if also $\forall l \in [K], \forall j \in [J]$, $\nu_k^0 \in \mathbb{R}\backslash\mathbb{Q}$, then there exists $d_0 < 0$ depending on $\{\nu_k^0, \{\omega_{j0}^{kl}\}_{j=1}^J\}_{k,l=1}^K$ such that $\forall k \in [K], \forall t \geqslant 0$, $\lambda_t^k(f_0) = 0 \implies \tilde{\lambda}_t^k(\nu_0, h_0) \leqslant d_0$.

S5. Lemmas on tests

In this section we prove two technical lemmas on the test functions used in the proofs of Theorem 5.5 and Proposition 3.5. In Section S5.1, we state and prove our first lemma, Lemma S5.1, which relates to the elementary test functions used in the proof of Theorem 5.5 (Section S1) and in Section S5.2, we prove Lemma A.5, which provides the bound on the error of the tests used in the proof of Case 2 of Proposition 3.5.

S5.1. Lemma S5.1: test used in the proof of Theorem 5.5

Lemma S5.1. *For $i \geqslant 1$, let $\mathcal{F}_i = \{f \in \mathcal{F}_T; \nu_l \leqslant \nu_l^0 + 2K\|r_0\|_1\mathbb{E}_0(\Delta\tau_1)i\epsilon_T,\ \forall l \in [K]\}$ and $f_1 \in \mathcal{F}_i$. We define the test*

$$\phi_{f_1,i} = \max_{l\in[K]} \mathbb{1}_{\{N^l(A_{1l})-\Lambda^l(A_{1l},f_0)\geqslant iT\epsilon_T/8\}} \wedge \mathbb{1}_{\{N^l(A_{1l}^c)-\Lambda^l(A_{1l}^c,f_0)\geqslant iT\epsilon_T/8\}},$$

where for all $l \in [K]$, $A_{1l} = \{t \in [0,T];\ \lambda_t^l(f_1) \geqslant \lambda_t^l(f_0)\}$, $\Lambda^l(A_{1l}, f_0) = \int_0^T \mathbb{1}_{A_{1l}}(t)\lambda_t^l(f_0)dt$ and $\Lambda^l(A_{1l}^c, f_0) = \int_0^T \mathbb{1}_{A_{1l}^c}(t)\lambda_t^l(f_0)dt$. Then

$$\mathbb{E}_0[\mathbb{1}_{\tilde{\Omega}_T}\phi_{f_1,i}] + \sup_{\|f-f_1\|_1\leqslant i\epsilon_T/(12N_0)} \mathbb{E}_0\left[\mathbb{E}_f[\mathbb{1}_{\tilde{\Omega}_T}\mathbb{1}_{f\in S_i}(1-\phi_{f_1,i})|\mathcal{G}_0]\right] \leqslant (2K+1)\max_{l\in[K]} e^{-x_{1l}Ti\epsilon_T(\sqrt{\mu_l^0}\wedge i\epsilon_T)},$$

where for $l \in [K]$, $x_{1l} > 0$ is an absolute constant, $\mu_l^0 = \mathbb{E}_0\left[\lambda_t^l(f_0)\right]$, $N_0 = 1 + \sum_{l=1}^K \mu_l^0$ and S_i is defined in (S1.1).

Proof. For $l \in [K]$, let

$$\phi_{il} = \phi_{il}(f_1) = \mathbb{1}_{\{N^l(A_{1l}) - \Lambda^l(A_{1l}, f_0) \geqslant iT\epsilon_T/8\}}.$$

Mimicking the proof of Lemma 1 of [2], we obtain that

$$\mathbb{E}_0\left[\phi_{il}\mathbb{1}_{\tilde{\Omega}_T}\right] \leqslant e^{-x_1 iT\epsilon_T \min(\sqrt{\mu_l^0}, i\epsilon_T)}. \tag{S5.9}$$

We first consider the event $\{\Lambda^l(A_{1l}, f_1) - \Lambda^l(A_{1l}, f_0) \geqslant \Lambda^l(A_{1l}^c, f_1) - \Lambda^l(A_{1l}^c, f_0)\}$. Let $f \in \mathcal{F}_i$ such that $\|f - f_1\|_1 \leqslant \zeta i\epsilon_T$ with $\zeta = 1/(6N_0)$ and $N_0 = 1 + \sum_l \mu_l^0$. On $\tilde{\Omega}_T$, using that ϕ_l is L-Lipschitz for any l, we have

$$\begin{aligned}
T\tilde{d}_{1T}(f, f_1) &= \sum_{l=1}^K \int_0^T \mathbb{1}_{A_2(T)}(t)|\lambda_t^l(f) - \lambda_t^l(f_1)|dt \leqslant \sum_{l=1}^K \int_0^T |\lambda_t^l(f) - \lambda_t^l(f_1)|dt \\
&\leqslant L\sum_{l=1}^K \int_0^T |\tilde{\lambda}_t^l(\nu, h) - \tilde{\lambda}_t^l(\nu_1, h_1)|dt \\
&\leqslant TL\sum_{l=1}^K |\nu_l - \nu_l^1| + L\sum_{l=1}^K\sum_{k=1}^K \int_0^T \int_{t-A}^t |(h_{kl} - h_{kl}^1)(t-s)|N^k(ds) \\
&\leqslant TL\|\nu - \nu_1\|_1 + \max_l N^l[-A, T]L\sum_{l=1}^K\sum_{k=1}^K \|h_{kl} - h_{kl}^1\|_1 \\
&\leqslant LN_0T\,\|f - f_1\|_1 \leqslant LN_0T\zeta i\epsilon_T.
\end{aligned}$$

Moreover, since $f \in S_i$, on $\tilde{\Omega}_T$, we also have that

$$\int_0^T \mathbb{1}_{A_2(T)}\lambda_t^l(f)dt \leqslant \int_0^T \mathbb{1}_{A_2(T)}\lambda_t^l(f_0)dt + KT(i+1)\epsilon_T \leqslant 2T\mu_l^0 + KT(i+1)\epsilon_T =: \tilde{v}.$$

Applying again inequality (7.7) of [5] with $v = \tilde{v}$ and using the computations of [2], we arrive at

$$\mathbb{E}_f\left[\mathbb{1}_{\tilde{\Omega}_T}\mathbb{1}_{f\in S_i}(1 - \phi_{il})\middle|\mathcal{G}_0\right] \leqslant 2Ke^{-x_{1l}iT\epsilon_T \min(\sqrt{\mu_l^0}, i\epsilon_T)},$$

for some $x_{1l} > 0$. We can obtain similar results for

$$\phi'_{il} = \mathbb{1}_{\{N^l(A_{1l}^c) - \Lambda^l(\bar{A}_{1l}^c, f_0) \geqslant iT\epsilon_T/8\}}.$$

Finally, with $\phi_{f_1,i} = \max_{l\in[K]} \phi_{il} \wedge \phi'_{il}$, we arrive at the final results of this lemma:

$$\mathbb{E}_0\left[\phi_{f_1,i}\mathbb{1}_{\tilde{\Omega}_T}\right] \leqslant \max_l e^{-x_{1l}iT\epsilon_T \min(\sqrt{\mu_l^0}, i\epsilon_T)} \leqslant e^{-(\min_l x_{1l})iT\epsilon_T \min(\sqrt{\mu_l^0}, i\epsilon_T)}$$

$$\mathbb{E}_f[\mathbb{1}_{\tilde{\Omega}_T}\mathbb{1}_{f\in S_i}(1 - \phi_{f_1,i})|\mathcal{G}_0] \leqslant \min_l \mathbb{E}_f[\mathbb{1}_{\tilde{\Omega}_T}\mathbb{1}_{f\in S_i}(1 - \phi_{il})|\mathcal{G}_0] \leqslant 2Ke^{-(\min_l x_{1l})iT\epsilon_T \min(\sqrt{\mu_l^0}, i\epsilon_T)}.$$

□

S5.2. Proof of Lemma A.5

In Lemma A.5, we establish the bound on the type I and type II errors of the tests to estimate the parameter θ in the shifted ReLU link function considered in Case 2 of Proposition 3.5.

We recall that $\Theta = \mathbb{R}_+\backslash\{0\}^K$ and $\bar{A}(R) = \{\theta \in \Theta;\ \|\theta - \theta_0\|_1 \leqslant R\}$. Let $\zeta > 0$ and

$$(f_1, \theta_1) = (\nu_1, h_1, \theta_1) = ((\nu_k^1)_k, (h_{lk}^1)_{l,k}, (\theta_k^1)_k) \in (\bar{A}(\tilde{M}_T\epsilon_T)^c \times \mathcal{F}) \cap A_{L_1}(M_T\epsilon_T),$$

with $\tilde{M}_T = \tilde{M}\sqrt{\kappa_T}$, $M_T = M\sqrt{\kappa_T}$ and $\tilde{M} \geqslant M$. Let $(f, \theta) \in (\bar{A}(\tilde{M}_T\epsilon_T)^c \times \mathcal{F}) \cap A_{L_1}(M_T\epsilon_T)$ such that $\|f - f_1\|_1 \leqslant \zeta\epsilon_T$, i.e,

$$\sum_k |\nu_k - \nu_k^1| + |\theta_k - \theta_k^1| + \sum_{l,k} \left\|h_{lk} - h_{lk}^1\right\|_1 \leqslant \zeta\epsilon_T.$$

Since $\theta \in \bar{A}(\tilde{M}_T\epsilon_T)^c$, there exists $k \in [K]$ such that $|\theta_k^0 - \theta_k| \geqslant \tilde{M}_T\epsilon_T/K$. For this k, from assumption (S7.36), there exists $l \in [K]$ and $x_1, x_2, c_\star > 0$ such that $\forall x \in [x_1, x_2]$, $h_{lk}^0(x) \leqslant -c_\star < 0$.

We first consider the case $\theta_k < \theta_k^0 - \tilde{M}_T\epsilon_T/K$ and recall the notation of Section 5.3: $\delta' = (x_2 - x_1)/3$, $n_1 = \lfloor 2\nu_k^1/(\kappa_1 c_\star)\rfloor + 1$ for some $\kappa_1 \in (0,1)$ and the subset of excursions

$$\mathcal{E} = \{j \in [J_T];\ N[\tau_j, \tau_j + \delta') = N^l[\tau_j, \tau_j + \delta') = n_1,\ N[\tau_j + \delta',\ \tau_{j+1}) = 0\}.$$

We recall that

$$I_k^0(f_1, \theta_1) = \left\{t \in [0,T];\ \lambda_t^k(f_1, \theta_1) = \theta_k^1,\ \lambda_t^k(f_0, \theta_0) = \theta_k^0\right\},$$

and we first state a preliminary lemma on $I_k^0(f_1, \theta_1)$, which is proved at the end of this proof.

Lemma S5.2. *In the Hawkes model with shifted ReLU link function , for any $f_0 \in \mathcal{F}$ such that (S7.36) is satisfied and any $(f_1, \theta_1) \in (\bar{A}(\tilde{M}_T\epsilon_T)^c \times \Theta) \cap A_{L_1}(M_T\epsilon_T)$, on $\tilde{\Omega}_T'$, it holds that*

$$|I_k^0(f_1, \theta_1)| \geqslant \frac{x_2 - x_1}{2} \sum_{j \in [J_T]} \mathbb{1}_{j \in \mathcal{E}},$$

with $\mathcal{E}$ defined in (35).

Let

$$\phi_k(f_1, \theta_1) := \mathbb{1}_{N^k(I_k^0(f_1,\theta_1)) - \Lambda_k(I_k^0(f_1,\theta_1), f_0) < -v_T} \vee \mathbb{1}_{|\mathcal{E}| < \frac{p_0 T}{2\mathbb{E}_0[\Delta\tau_1]}},$$

with $\Lambda_k(I_k^0(f_1, \theta_1), f_0) = \int_0^T \mathbb{1}_{I_k^0(f_1,\theta_1)}(t)\lambda_t^k(f_0)dt$, $p_0 = \mathbb{P}_0[j \in \mathcal{E}]$, $v_T = w_T T\epsilon_T > 0$ with $w_T > 0$ chosen later. We have by definition

$$\mathbb{E}_0\left[\phi_k(f_1, \theta_1)\mathbb{1}_{\tilde{\Omega}_T'}\right] \leqslant \mathbb{P}_0\left[\left\{|\mathcal{E}| < \frac{p_0 T}{2\mathbb{E}_0[\Delta\tau_1]}\right\} \cap \tilde{\Omega}_T'\right] + \mathbb{P}_0\left[\left\{N^k(I_k^0(f_1, \theta_1)) - \Lambda_k(I_k^0(f_1, \theta_1), f_0) < -v_T\right\} \cap \tilde{\Omega}_T'\right]. \tag{S5.10}$$

For the first term on the RHS of (S5.10), we apply Hoeffding's inequality with $X_j = \mathbb{1}_{j\in\mathcal{E}} \overset{i.i.d.}{\sim} \mathcal{B}(p_0)$:

$$\begin{aligned}\mathbb{P}_0\left[\left\{|\mathcal{E}| < \frac{p_0 T}{2\mathbb{E}_0[\Delta\tau_1]}\right\}\cap\tilde{\Omega}'_T\right] &\leqslant \mathbb{P}_0\left[\left\{\sum_{j=1}^{J_T} X_j < \frac{p_0 T}{2\mathbb{E}_0[\Delta\tau_1]}\right\}\cap\tilde{\Omega}'_T\right]\\ &\leqslant \mathbb{P}_0\left[\sum_{j=1}^{T/(2\mathbb{E}_0[\Delta\tau_1])} X_j < \frac{p_0 T}{2\mathbb{E}_0[\Delta\tau_1]}\right] \lesssim e^{-\frac{Tp_0^2}{8\mathbb{E}_0[\Delta\tau_1]}} = o(e^{-u_0 T\epsilon_T^2}),\end{aligned}$$

for $u_0 < p_0^2/(8\mathbb{E}_0[\Delta\tau_1])$ and using that on $\tilde{\Omega}'_T$, $J_T > T/(2\mathbb{E}_0[\Delta\tau_1])$. For the second term of the RHS of (S5.10), we apply inequality (7.7) in [5], with $H_t = \mathbb{1}_{I_k^0(f_1,\theta_1)}(t)$, $H_t^2 \circ \Lambda_t^k(f_0) = \int_0^T \mathbb{1}_{I_k^0(f_1,\theta_1)}(t)\theta_k^0 dt = \theta_k^0|I_k^0(f_1,\theta_1)| \leqslant \theta_k^0 T$, $x = x_3 T\epsilon_T^2$, $x_3 > 0$. If $\sqrt{2\theta_k^0 T x} + x/3 \leqslant w_T T\epsilon_T$ and $x_3 > u_0$, then by (7.7) of [5],

$$\mathbb{P}_0\left[\left\{N^k(I_k^0(f_1,\theta_1)) - \Lambda_k(I_k^0(f_1,\theta_1), f_0) < -v_T\right\}\cap\tilde{\Omega}'_T\right] \leqslant e^{-x_3 T\epsilon_T^2} = o(e^{-u_0 T\epsilon_T^2}).$$

Reporting into (S5.10), we obtain that $\mathbb{E}_0\left[\phi_k(f_1)\mathbb{1}_{\tilde{\Omega}'_T}\right] = o(e^{-u_0 T\epsilon_T^2})$, which proves the first part of Lemma A.5. To prove the second part of Lemma A.5, we first note that

$$\mathbb{E}_f\left[(1-\phi_k(f_1,\theta_1))\mathbb{1}_{\tilde{\Omega}'_T}\right] = \mathbb{P}_f\left[\left\{N^k(I_k^0(f_1,\theta_1)) - \Lambda_k(I_k^0(f_1,\theta_1), f_0) \geqslant -v_T\right\}\cap\left\{|\mathcal{E}| \geqslant \frac{p_0 T}{2\mathbb{E}_0[\Delta\tau_1]}\right\}\cap\tilde{\Omega}'_T\right]. \tag{S5.11}$$

We also have

$$\Lambda_k(I_k^0(f_1,\theta_1), f_0) - \Lambda_k(I_k^0(f_1,\theta_1), f) = \Lambda_k(I_k^0(f_1,\theta_1), f_0) - \Lambda_k(I_k^0(f_1,\theta_1), f_1) \tag{S5.12}$$

$$+ \Lambda_k(I_k^0(f_1,\theta_1), f_1) - \Lambda_k(I_k^0(f_1,\theta_1), f). \tag{S5.13}$$

Firstly, if $|\mathcal{E}| > \frac{p_0}{2\mathbb{E}_0[\Delta\tau_1]}T$, then from Lemma S5.2,

$$|I_k^0(f_1,\theta_1)| \geqslant \frac{(x_2-x_1)p_0}{4\mathbb{E}_0[\Delta\tau_1]}T \tag{S5.14}$$

and

$$\Lambda_k(I_k^0(f_1,\theta_1), f_0) - \Lambda_k(I_k^0(f_1,\theta_1), f_1) = (\theta_k^0-\theta_k^1)|I_k^0(f_1,\theta_1)| \geqslant \frac{(x_2-x_1)p_0}{8K\mathbb{E}_0[\Delta\tau_1]}\tilde{M}_T T\epsilon_T, \tag{S5.15}$$

since $\|\theta-\theta_1\|_1 \leqslant \zeta\epsilon_T$ therefore $\theta_k^0-\theta_k^1 \geqslant |\theta_k^0-\theta_k| - |\theta_k-\theta_k^1| \geqslant \tilde{M}_T\epsilon_T/K - \zeta\epsilon_T \geqslant \frac{\tilde{M}_T}{2K}\epsilon_T$ for T large enough.

Secondly, since $\forall t \in I_k^0(f_1,\theta_1)$, $\tilde{\lambda}_t^k(\nu_1,h_1) \leqslant 0$ and $\tilde{\lambda}_t^k(\nu,h) \leqslant 0$, we have

$$\begin{aligned}\Lambda_k(I_k^0(f_1,\theta_1), f_1) - \Lambda_k(I_k^0(f_1,\theta_1), f) &= (\theta_k^1-\theta_k)|I_k^0(f_1,\theta_1)| - \int_{I_k^0(f_1,\theta_1)}\left((\tilde{\lambda}_t^k(\nu,h))_+ - (\tilde{\lambda}_t^k(\nu_1,h_1))_+\right)dt\\ &\geqslant (\theta_k^1-\theta_k)|I_k^0(f_1,\theta_1)| - \int_{I_k^0(f_1,\theta_1)}|\tilde{\lambda}_t^k(\nu,h) - \tilde{\lambda}_t^k(\nu_1,h_1)|dt\\ &\geqslant -\zeta T\epsilon_T - \int_0^T |\tilde{\lambda}_t^k(\nu,h) - \tilde{\lambda}_t^k(\nu_1,h_1))|dt,\end{aligned} \tag{S5.16}$$

where we have used the fact that by definition $|I_k^0(f_1,\theta_1)| \leqslant T$. Using Fubini's theorem, for any $l \in [K]$, we have

$$\begin{aligned}\int_0^T |\tilde{\lambda}_t^k(\nu,h) - \tilde{\lambda}_t^k(\nu_1,h_1))|dt &= \int_0^T \left|\nu_k - \nu_k^1 + \sum_l \int_{t-A}^{t^-} (h_{lk} - h_{lk}^1)(t-s)dN_s^l\right| dt \\ &\leqslant T|\nu_k - \nu_k^1| + \sum_l \int_{T-A}^T \int_s^{s+A} |h_{lk} - h_{lk}^1|(t)|dtdN_s^l = T|\nu_k - \nu_k^1| + \sum_l \left\|h_{lk} - h_{lk}^1\right\|_1 N^l[-A,T] \\ &\leqslant T\,\|f - f_1\|\left(1 + \sum_l(\mu_l^0 + \delta_T)\right) \leqslant \zeta T\epsilon_T\left(1 + 2\sum_l \mu_l^0\right), \end{aligned} \tag{S5.17}$$

using the definition of $\tilde{\Omega}_T'$ in Section 5.2. Consequently, reporting the previous upper bound into (S5.16), we obtain

$$\Lambda_k(I_k^0(f_1,\theta_1),f_1) - \Lambda_k(I_k^0(f_1,\theta_1),f) \geqslant -\zeta T\epsilon_T(2 + 2\sum_l \mu_l^0).$$

Therefore, using now (S5.15) and (S5.16) in (S5.12), we arrive at

$$\Lambda_k(I_k^0(f_1,\theta_1),f_0) - \Lambda_k(I_k^0(f_1,\theta_1),f) \geqslant \frac{\tilde{M}_T(x_2-x_1)p_0}{8K\mathbb{E}_0[\Delta\tau_1]}T\epsilon_T - \zeta T\epsilon_T(2 + 2\sum_l \mu_l^0) \geqslant \frac{\tilde{M}_T(x_2-x_1)p_0}{16K\mathbb{E}_0[\Delta\tau_1]}T\epsilon_T,$$

since for T large enough, $\tilde{M}_T > \frac{16K\zeta\mathbb{E}_0[\Delta\tau_1](2+2\sum_l \mu_l^0)}{(x_2-x_1)p_0}$. Reporting into (S5.11), we obtain

$$\begin{aligned}&\mathbb{P}_f\left[\left\{N^k(I_k^0(f_1,\theta_1)) - \Lambda_k(I_k^0(f_1,\theta_1),f_0) \geqslant -v_T\right\} \cap \left\{|\mathcal{E}| \geqslant \frac{p_0 T}{2\mathbb{E}_0[\Delta\tau_1]}\right\} \cap \tilde{\Omega}_T'\right] \\ &\leqslant \mathbb{P}_f\left[\{N^k(I_k^0(f_1,\theta_1)) - \Lambda_k(I_k^0(f_1,\theta_1),f) \geqslant -v_T + \frac{\tilde{M}_T(x_2-x_1)p_0}{16\mathbb{E}_0[\Delta\tau_1]}T\epsilon_T\} \cap \tilde{\Omega}_T'\right] \\ &\leqslant \mathbb{P}_f\left[\{N^k(I_k^0(f_1,\theta_1)) - \Lambda_k(I_k^0(f_1,\theta_1),f) \geqslant v_T\} \cap \tilde{\Omega}_T'\right],\end{aligned}$$

if $\tilde{M}_T > \frac{16w_T\mathbb{E}_0[\Delta\tau_1]}{(x_2-x_1)p_0}$, which is true for $\tilde{M}$ large enough (recall that $\tilde{M}_T = \tilde{M}\sqrt{\kappa_T}$) if $w_T \leqslant C\sqrt{\kappa_T}$ with $C > 0$ a constant.

Similarly to the proof of Lemma 1 in [2], we can adapt inequality (7.7) from [5] with $H_t = \mathbb{1}_{I_k^0(f_1,\theta_1)}(t)$ to the conditional probability $\mathbb{E}_f[.|\mathcal{G}_0]$ and the supermartingale $\int_0^T \mathbb{1}_{I_k^0(f_1,\theta_1)}(t)(dN_t - \lambda_t^k(f,\theta)dt)$. With $\tau = T$, $x_T = x_1 T\epsilon_T^2$, we obtain

$$\mathbb{P}_f\left[\{N^k(I_k^0(f_1,\theta_1)) - \Lambda_k(I_k^0(f_1,\theta_1),f) > v_T\} \cap \tilde{\Omega}_T'\right] \leqslant e^{-x_T T\epsilon_T^2} = o(e^{-(\kappa_T+c_1)T\epsilon_T^2}), \quad \text{if } x_T > \kappa_T + c_1. \tag{S5.18}$$

For this to be true, we also need $v_T > \sqrt{2\tilde{v}(\kappa_T+c_1)T\epsilon_T^2} + (\kappa_T+c_1)T\epsilon_T^2/3$ where $\tilde{v}$ is an upper bound of $H_t^2 \circ \Lambda_t^k(f)$. Using the fact that $\forall t \in I_k^0(f_1,\theta_1)$, $\tilde{\lambda}_t^k(\nu_1,h_1) \leqslant 0$, we have

$$H_t^2 \circ \Lambda_t^k(f) = \int_{I_k^0(f_1,\theta_1)} \lambda_t^k(f,\theta)dt = \theta_k|I_k^0(f_1,\theta_1)| + \int_{I_k^0(f_1,\theta_1)\cap\{\tilde{\lambda}_t^k(\nu,h)>0\}} \tilde{\lambda}_t^k(\nu,h)dt$$

$$\leqslant \theta_k |I_k^0(f_1,\theta_1)| + \int_{I_k^0(f_1,\theta_1)\cap\{\tilde{\lambda}_t^k(\nu,h)>0\}} |\tilde{\lambda}_t^k(\nu,h) - \tilde{\lambda}_t^k(\nu_1,h_1)|dt$$

$$\leqslant \theta_k |I_k^0(f_1,\theta_1)| + \zeta T\epsilon_T \left(1 + 2\sum_l \mu_l^0\right) \leqslant T(\theta_k + \tilde{M}_T\epsilon_T/K) \leqslant \theta_k^0 T =: \tilde{v},$$

using (S5.17) and since for T large enough, $\zeta K(1 + 2\sum_l \mu_l^0) < M_T \leqslant \tilde{M}_T$. Consequently, if $w_T > \sqrt{2\theta_k^0(\kappa_T + c_1)} + (\kappa_T + c_1)\epsilon_T/3$ and $w_T \leqslant C\sqrt{\kappa_T}$ (which is possible since $\epsilon_T = o(1/\sqrt{\kappa_T})$ by assumption), then (S5.18) holds and we can finally conclude that $\mathbb{E}_f\left[(1-\phi_k(f_1,\theta_1))\mathbb{1}_{\tilde{\Omega}_T'}\right] = o(e^{-(\kappa_T+c_1)T\epsilon_T^2})$ is verified, which leads to the second part of Lemma A.5.

In the alternative case where $\theta_k > \theta_k^0 + \tilde{M}_T\epsilon_T/K$, similar arguments can be applied with $I_k^0(f_1,\theta_1)$ defined as in (34) and $\mathcal{E}$ defined as in (35) except that $n_1 = \lfloor 2\nu_k^0/(\kappa_1 c_\star)\rfloor + 1$. We then use the following test, with $v_T = w_T T\epsilon_T$

$$\phi_k(f_1,\theta_1) := \mathbb{1}_{N^k(I_k^0(f_1,\theta_1)) - \Lambda_k(I_k^0(f_1,\theta_1),f_0) > v_T} \vee \mathbb{1}_{|\mathcal{E}| < \frac{p_0 T}{2\mathbb{E}_0[\Delta\tau_1]}}.$$

Then Hoeffding's inequality and inequality (7.7) from [5] lead to $\mathbb{E}_0\left[\phi_k(f_1,\theta_1)\mathbb{1}_{\tilde{\Omega}_T}\right] = o(e^{-u_0 T\epsilon_T^2})$. For the second part of Lemma A.5, we first note that in this case, since $\forall t \in I_k^0(f_1,\theta_1)$, $\lambda_t^k(f,\theta) \geqslant \theta_k$ (and also $\lambda_t^k(f_0,\theta_0) = \theta_k^0$, $\lambda_t^k(f_1,\theta_1) = \theta_k^1$), then on the event $|\mathcal{E}| \geqslant \frac{p_0 T}{2\mathbb{E}_0[\Delta\tau_1]}$,

$$\Lambda_k(I_k^0(f_1,\theta_1),f_0) - \Lambda_k(I_k^0(f_1,\theta_1),f) \leqslant (\theta_k^0 - \theta_k^1)|I_k^0(f_1,\theta_1)| + (\theta_k^1 - \theta_k)|I_k^0(f_1,\theta_1)|$$

$$\leqslant (-\tilde{M}_T\epsilon_T/K + \zeta\epsilon_T)|I_k^0(f_1,\theta_1)| \leqslant -\frac{\tilde{M}_T\epsilon_T|I_k^0(f_1,\theta_1)|}{2K} \leqslant -\frac{(x_2-x_1)p_0}{8K\mathbb{E}_0[\Delta\tau_1]}\tilde{M}_T T\epsilon_T,$$

for T large enough and using (S5.14). Consequently,

$$\mathbb{P}_f\left[\{N^k(I_k^0(f_1,\theta_1)) - \Lambda_k(I_k^0(f_1,\theta_1),f_0) \leqslant v_T\} \cap \left\{|\mathcal{E}| \geqslant \frac{p_0 T}{2\mathbb{E}_0[\Delta\tau_1]}\right\} \cap \tilde{\Omega}_T'\right]$$

$$\leqslant \mathbb{P}_f\left[\{N^k(I_k^0(f_1,\theta_1)) - \Lambda_k(I_k^0(f_1,\theta_1),f) \leqslant v_T - \frac{(x_2-x_1)p_0}{8\mathbb{E}_0[\Delta\tau_1]}\tilde{M}_T T\epsilon_T\} \cap \tilde{\Omega}_T'\right]$$

$$\leqslant \mathbb{P}_f\left[\{N^k(I_k^0(f_1,\theta_1)) - \Lambda_k(I_k^0(f_1,\theta_1),f) \leqslant -v_T\} \cap \tilde{\Omega}_T'\right],$$

if $\tilde{M}_T > \frac{16K\mathbb{E}_0[\Delta\tau_1]}{(x_2-x_1)p_0} w_T$. Applying inequality (7.7) from [5], we can finally obtain

$$\mathbb{E}_f\left[(1-\phi_k(f_1,\theta_1))\mathbb{1}_{\tilde{\Omega}_T'}\right] = o(e^{-(\kappa_T+c_1)T\epsilon_T^2}),$$

which ends the proof of Lemma A.5.

Proof of Lemma S5.2

Let $(f_0,\theta_0) \in \mathcal{F}\times\Theta$, $(f_1,\theta_1) \in (\mathcal{F}\times\bar{A}(\tilde{M}_T\epsilon_T)^c) \cap A_{L_1}(M_T\epsilon_T)$ and $k \in [K]$ such that $|\theta_k^1 - \theta_k^0| > \tilde{M}_T\epsilon_T/K$. For this k, from assumption (S7.36), there exists $l \in [K]$ and $x_1, x_2, c_\star > 0$ such that $\forall x \in [x_1,x_2]$, $h_{lk}^0(x) \leqslant -c_\star < 0$. We first consider the case $\theta_k^1 < \theta_k^0 - \tilde{M}_T\epsilon_T/K$. Since $(f_1,\theta_1) \in A_{L_1}(M_T\epsilon_T)$,

we also have that $|\theta_k^1 + \nu_k^1 - \theta_k^0 - \nu_k^0| \leqslant M_T \epsilon_T$, which implies that $\nu_k^1 > \nu_k^0 - (M_T - \tilde{M}_T/K)\epsilon_T > \nu_k^0/2$. For $0 < \kappa_1 < 1$, we define

$$B_1 = \{x \in [0, A];\ h_{lk}^{1-}(x) > \kappa_1 c_\star\}, \quad n_1 = \left\lfloor \frac{2\nu_k^1}{\kappa_1 c_\star} \right\rfloor + 1.$$

Moreover, since $\left\|h_{lk}^0 - h_{lk}^1\right\|_1 \leqslant M_T \epsilon_T$ and $h_{lk}^{0-}(x) \geqslant c_\star$ for $x \in [x_1, x_2]$,

$$|[x_1, x_2] \cap B_1^c| c_\star (1 - \kappa_1) \leqslant \int_{[x_1, x_2] \cap B_1^c} (h_{lk}^1 - h_{lk}^0)(x) dx \leqslant M_T \epsilon_T$$

$$\implies |[x_1, x_2] \cap B_1| \geqslant (x_2 - x_1) - \frac{M_T \epsilon_T}{c_\star (1 - \kappa_1)} \geqslant 3(x_2 - x_1)/4,$$

for T large enough.

Now let $\delta' = (x_2 - x_1)/4$. For $j \in \mathcal{E}$, we denote $T_1, \ldots, T_{n_1}$ the n_1 events occurring on $[\tau_j, \tau_j + \delta']$. For $t \in [\tau_j + x_1 + \delta', \tau_j + x_2]$, we have $t - T_i \in [x_1, x_2]$ for any $i \in [n_1]$ and

$$\tilde{\lambda}_t^k(\nu_0, h_0) = \nu_k^0 + \sum_{i \in [n_1]} h_{lk}^0(t - T_i) < \nu_k^0 - n_1 c_\star < 2\nu_k^1 - n_1 \kappa_1 c_\star < 0,$$

by definition of n_1. Similarly, for $t \in B_1 + [\tau_j, \tau_j + \delta']$, we have $t - T_i \in B_1$ and therefore

$$\tilde{\lambda}_t(\nu_1, h_1) = \nu_k^1 + \sum_{i \in [n_1]} h_{lk}^1(t - T_i) < 2\nu_k^1 - n_1 \kappa_1 c_\star < 0.$$

Consequently, for $t \in ([x_1, x_2] \cap B_1) + [\tau_j, \tau_j + \delta']$, $\lambda_t^k(f_0, \theta_0) = \theta_k^0$ and $\lambda_t^k(f_1, \theta_1) = \theta_k^1$, and thus $([x_1, x_2] \cap B_1) + [\tau_j, \tau_j + \delta'] \subset I_k^0(f_1, \theta_1)$. Moreover, we have

$$\left|([x_1, x_2] \cap B_1) + [\tau_j, \tau_j + \delta']\right| \geqslant 3(x_2 - x_1)/4 - (x_2 - x_1)/4 \geqslant (x_2 - x_1)/2.$$

Consequently,

$$|I_k^0(f_1, \theta_1)| = \sum_{j=0}^{J_T} [\tau_j, \tau_{j+1}] \cap \{t \geqslant 0;\ \lambda_t^k(f_0, \theta_0) = \theta_0,\ \lambda_t^k(f_1, \theta_1) = \theta_1\} \geqslant \sum_{j \in [J_T]} \frac{x_2 - x_1}{2} \mathbb{1}_{j \in \mathcal{E}}.$$

In the alternative case $\theta_k^1 > \theta_k^0 + \tilde{M}_T \epsilon_T / K$, similar computations can be derived by defining n_1 as $n_1 = \min\{n \in \mathbb{N};\ n\kappa_1 c_\star > \nu_k^0\}$.

S6. Lemmas on $L_T(f_0) - L_T(f)$

For $f_0, f \in \mathcal{F}$, we define the Kullback-Leibler (KL) divergence in the Hawkes model as

$$KL(f_0, f) = \mathbb{E}_0[L_T(f_0) - L_T(f)]. \tag{S6.19}$$

With a slight abuse of notation, we still use the same notations $L_T(f_0), L_T(f), KL(f_0, f)$ in the nonlinear model with shifted ReLU link function with the additional shift parameter θ. We also note that with the standard ReLU link function, the KL divergence can be infinite for some $f \in \mathcal{F}$, e.g., if there exists $t \in [0, T]$ such that $dN_t^k = 1$ and $\lambda_t^k(f) = 0$. However, in this model, for any $f \in B_\infty(\epsilon_T)$, $\lambda_t^k(\nu, h) \geqslant \lambda_t^k(\nu_0, h_0)$, which implies that $KL(f_0, f) < +\infty$. The next lemma provides some upper bound on the KL divergence on $B_\infty(\epsilon_T)$ with all the link functions considered in Theorem 3.2 and Proposition 3.5.

S6.1. Lemma to bound the Kullback - Leibler divergence

Lemma S6.1. *Under the assumptions of Theorem 3.2 and of Case 2 of Proposition 3.5, for any $f \in B_\infty(\epsilon_T)$ and T large enough,*

$$0 \leqslant KL(f_0, f) \leqslant \kappa_1 T \epsilon_T^2,$$

and, under the assumptions of Case 1 of Proposition 3.5, we similarly have

$$0 \leqslant KL(f_0, f) \leqslant \kappa_2 (\log T)^2 T,$$

with $\kappa_1, \kappa_2 > 0$ constants that only depends on $(\phi_k)_k$ and f_0.

***Remark* S6.2.** For the models considered in Theorem 3.2 and with the shifted ReLU link function (Case 2 of Proposition 3.5), for $f \in B_2(\epsilon_T, B)$, we instead obtain

$$0 \leqslant KL(f_0, f) \lesssim (\log\log T) T \epsilon_T^2.$$

Moreover, with the standard ReLU link function (Case 1 of Proposition 3.5), without assuming that the additional condition (8) holds, we can also obtain the sub-obtimal bound

$$0 \leqslant KL(f_0, f) \lesssim T \epsilon_T,$$

which would also lead to the sub-optimal posterior concentration rate $\sqrt{\epsilon_T}$.

Proof. For simplicity of exposition, throughout this proof, we use the notation $\lambda_t^k(f), \lambda_t^k(f_0)$ for the intensity in all models, therefore including the case $\lambda_t^k(f,\theta), \lambda_t^k(f_0,\theta_0)$ (Case 2 of Proposition 3.5).

Firstly, similarly to the proof of Lemma 2 of [2], we can easily prove that $KL(f_0, f) \geqslant 0$. Secondly, since intensities are predictable, we have

$$\mathbb{E}_0\left[\int_0^T \log\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right)(dN_t^k - \lambda_t^k(f_0)dt)\right] = 0. \tag{S6.20}$$

Since

$$KL(f_0, f) = \sum_k \mathbb{E}_0\left[\int_0^T \log\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right) dN_t^k + \int_0^T (\lambda_t^k(f) - \lambda_t^k(f_0))dt\right], \tag{S6.21}$$

then, with

$$R_T = \sum_k \mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T^c} \int_0^T \lambda_t^k(f_0) \log\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right) dt\right] + \mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T^c} \int_0^T (\lambda_t^k(f) - \lambda_t^k(f_0))dt\right], \tag{S6.22}$$

$$KL(f_0, f) = \sum_k \mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T}\left(\int_0^T \lambda_t^k(f_0) \log\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right) dt + \int_0^T (\lambda_t^k(f) - \lambda_t^k(f_0))dt\right)\right] + R_T. \tag{S6.23}$$

We first show that $R_T = o(T\epsilon_T^2)$. For the first term on the RHS of (S6.22), if $f \in B_\infty(\epsilon_T)$, we use that $\log x \leqslant x - 1$ for $x \geqslant 1$ and we have

$$\sum_k \mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T^c} \int_0^T \log \frac{\lambda_t^k(f)}{\lambda_t^k(f_0)} \lambda_t^k(f_0) dt\right] \leqslant \sum_k \mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T^c} \int_0^T \mathbb{1}_{\lambda_t^k(f) > \lambda_t^k(f_0)} \log \frac{\lambda_t^k(f)}{\lambda_t^k(f_0)} \lambda_t^k(f_0) dt\right]$$

$$\leqslant \sum_k \mathbb{E}_0\left[\int_0^T \mathbb{1}_{\tilde{\Omega}_T^c}\mathbb{1}_{\lambda_t^k(f_0)>0}\left(\lambda_t^k(f)-\lambda_t^k(f_0)\right)dt\right]$$

$$\leqslant \sum_k TL\left(|\nu_k^0-\nu_k|+\sum_l \left\|h_{lk}-h_{lk}^0\right\|_\infty \mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T^c}\sup_{t\in[0,T]} N^l[t-A,t)\right]\right)$$

$$\leqslant TL\sum_k\left(|\nu_k^0-\nu_k|+\sum_l \left\|h_{lk}-h_{lk}^0\right\|_\infty\right)\mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T^c}\max_l\sup_{t\in[0,T]} N^l[t-A,t)\right]\leqslant LT^{1-\beta}\epsilon_T \tag{S6.24}$$

for T large enough, using Lemma A.1 for $\beta>0$. If the model verifies Assumption 3.1(i), and $f\in B_2(\epsilon_T,B)$, we have

$$\frac{\lambda_t^k(f)}{\lambda_t^k(f_0)}\vee\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\leqslant 2\frac{2\theta_k^0+2L\nu_k^0+L(B+\max_l\left\|h_{lk}^0\right\|_\infty)\sup_t N[t-A,t)}{\inf_x\phi_k(x)},$$

therefore

$$\mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T^c}\int_0^T\left|\log\frac{\lambda_t^k(f)}{\lambda_t^k(f_0)}\right|\lambda_t^k(f_0)dt\right]\lesssim \mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T^c}\max_l\sup_{t\in[0,T]}N^l[t-A,t)\int_0^T\lambda_t^k(f_0)dt\right]$$

$$\lesssim T\mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T^c}\left(\sup_{t\in[0,T]}N[t-A,t)\right)\left(\nu_k^0+\max_l\left\|h_{lk}^0\right\|_\infty\sup_{t\in[0,T]}N^l[t-A,t)\right)\right]$$

$$\lesssim T\mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T^c}\max_l\left(\sup_{t\in[0,T]}N^l[t-A,t)\right)^2\right]\lesssim T^{1-\beta}.$$

If instead the model verifies Assumption 3.1(ii), using that $\log\phi_k$ is L_1-Lipschitz for any k, we can alternatively use that

$$\sum_k\mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T^c}\int_0^T\left|\log\frac{\lambda_t^k(f)}{\lambda_t^k(f_0)}\right|\lambda_t^k(f_0)dt\right]\leqslant L_1\sum_k\mathbb{E}_0\left[\int_0^T\mathbb{1}_{\tilde{\Omega}_T^c}\lambda_t^k(f_0)|\tilde{\lambda}_t^k(f)-\tilde{\lambda}_t^k(f_0)|dt\right]$$

$$\lesssim\sum_k T\left(|\nu_k^0-\nu_k|+\sum_l\left\|h_{lk}-h_{lk}^0\right\|_\infty\mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T^c}\max_l\left(\sup_{t\in[0,T]}N^l[t-A,t)\right)^2\right]\right)\lesssim T^{1-\beta}.$$

We can additionally bound the second term of (S6.22) in a similar fashion and conclude that, in all cases, $R_T=O(T^{1-\beta})=o(T\epsilon_T^2)$ for β large enough.

To bound the first term of the RHS of (S6.23), we consider separately the models satisfying Assumption 3.1(i) and (ii) and Case 1 and Case 2 of Proposition 3.5.

Scenario 1: under Assumption 3.1(i) or Case 2 of Proposition 3.5

Under Assumption 3.1(i), for any $f\in B_\infty(\epsilon_T)$ or $f\in B_2(\epsilon_T,B)$ and $t\geqslant 0$, $\lambda_t^k(f)\geqslant\inf_x\phi_k(x)\geqslant\min_k\inf_x\phi_k(x)$ and $\lambda_t^k(f_0)\leqslant L\nu_k^0+L\sup_{t\in[0,T]}N[t-A,t)\sum_l\|h_{lk}^0\|_\infty$. In Case 2 of Proposition 3.5, for T large enough, $t\in[0,T]$ and $\theta\in B_\infty^\Theta(\epsilon_T)$, $\lambda_t^k(f,\theta)\geqslant\theta_k\geqslant\theta_k^0/2$ and $\lambda_t^k(f_0,\theta_0)\leqslant\theta_k^0+L\nu_k^0+L\sup_{t\in[0,T]}N[t-A,t)\sum_l\|h_{lk}^0\|_\infty$. Therefore, in this scenario, on $\tilde{\Omega}_T$, $\lambda_t^k(f_0)/\lambda_t^k(f)\leqslant\ell_0\log T$ for some $\ell_0>0$. Thus, with $\chi(x)=-\log x+x-1$, we have

$$KL(f_0,f)-R_T=\sum_k\mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T}\left(\int_0^T\lambda_t^k(f_0)\left(\log\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right)+\frac{\lambda_t^k(f)}{\lambda_t^k(f_0)}-1\right)dt\right)\right]$$

$$= \sum_k \mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T}\left(\int_0^T \lambda_t^k(f_0)\chi\left(\frac{\lambda_t^k(f)}{\lambda_t^k(f_0)}\right)dt\right)\right]$$
$$\leqslant \frac{4\log(\ell_0 \log T)}{\min_k \inf_x \phi_k(x)} \sum_k \mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T}\int_0^T (\lambda_t^k(f_0) - \lambda_t^k(f))^2 dt\right],$$

since for any $r_T \in (0, 1/2]$ and $x \geqslant r_T$, we have $\chi(x) \leqslant 4\log r_T^{-1}(x-1)^2$ (see the proof of Lemma 2 of [2]). Note that if $f \in B_\infty(\epsilon_T)$, $\forall t \in [0,T]$, $\lambda_t^k(f) \geqslant \lambda_t^k(f_0)$ and we obtain instead

$$KL(f_0, f) - R_T \leqslant \frac{1}{\min_k \inf_x \phi_k(x)} \sum_k \mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T}\int_0^T (\lambda_t^k(f_0) - \lambda_t^k(f))^2 dt\right].$$

Moreover, since ϕ_k is L-Lipschitz, under Assumption 3.1,

$$|\lambda_t^k(f_0) - \lambda_t^k(f)| = |\phi_k(\tilde{\lambda}_t^k(\nu_0, h_0)) - \phi_k(\tilde{\lambda}_t^k(\nu, h))| \leqslant L|\tilde{\lambda}_t^k(\nu_0, h_0) - \tilde{\lambda}_t^k(\nu, h)|$$
$$\leqslant L|\nu_k - \nu_k^0| + L\sum_l \int_{t-A}^{t^-} |h_{lk} - h_{lk}^0|(t-s)dN_s^l,$$

and in Case 2 of Proposition 3.5, we have

$$|\lambda_t^k(f_0, \theta_0) - \lambda_t^k(f, \theta)| = |\theta_k^0 + \phi_k(\tilde{\lambda}_t^k(\nu_0, h_0)) - \theta_k - \phi_k\tilde{\lambda}_t^k(\nu, h))| \leqslant |\theta_k^0 - \theta_k| + L|\tilde{\lambda}_t^k(\nu_0, h_0) - \tilde{\lambda}_t^k(\nu, h)|$$
$$\leqslant |\theta_k - \theta_k^0| + L|\nu_k - \nu_k^0| + L\sum_l \int_{t-A}^{t^-} |h_{lk} - h_{lk}^0|(t-s)dN_s^l.$$

Using the same computations as in the proof of Lemma 2 of [2], we obtain

$$\sum_k \mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T}\left(\int_0^T (\lambda_t^k(f_0) - \lambda_t^k(f))^2\right)dt\right] \leqslant \gamma_0 T\left(|\nu_k - \nu_k^0|^2 + \sum_l \|h_{lk} - h_{lk}^0\|_2^2\right) \leqslant \gamma_0 T\epsilon_T^2,$$

or

$$\sum_k \mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T}\left(\int_0^T (\lambda_t^k(f_0, \theta_0) - \lambda_t^k(f, \theta))^2\right)dt\right] \leqslant \gamma_0 T\left(\sum_k |\theta_k - \theta_k^0|^2 + |\nu_k - \nu_k^0|^2 + \sum_l \|h_{lk} - h_{lk}^0\|_2^2\right) \leqslant \gamma_0 T\epsilon_T^2,$$

with $\gamma_0 := \max(1, L)\left[3 + 6K\sum_k\left(A\mathbb{E}_0\left[\lambda_0^k(f_0)^2\right] + \mathbb{E}_0\left[\lambda_0^k(f_0)\right]\right)\right]$. Consequently,

$$KL(f_0, f) - R_T \leqslant \begin{cases} \frac{4\log(\ell_0 \log T)}{\min_k \inf_x \phi_k(x)}\gamma_0 T\epsilon_T^2 & \text{if} \quad f \in B_2(\epsilon_T, B) \\ \frac{\gamma_0}{\min_k \inf_x \phi_k(x)} T\epsilon_T^2 & \text{if} \quad f \in B_\infty(\epsilon_T). \end{cases} \tag{S6.25}$$

Therefore, $KL(f_0, f) \leqslant \kappa_1'(\log\log T)T\epsilon_T^2$, with $\kappa_1' = \frac{8\gamma_0}{\min_k \inf_x \phi_k(x)}$ if $f \in B_2(\epsilon_T, B)$ - or $KL(f_0, f) \leqslant \kappa_1 T\epsilon_T^2$ with $\kappa_1 = \frac{2}{\min_k \inf_x \phi_k(x)}$ if $f \in B_\infty(\epsilon_T)$.

Scenario 2: Under Assumption 3.1(ii), i.e., $\phi_k > 0$, and $\log\phi_k$ and $\sqrt{\phi_k}$ are L_1-Lipschitz, $L_1 > 0$.

For $k \in [K]$, let $\Lambda^k(f) := \int_0^T \lambda_t^k(f)dt$. Then for $t \in [0,T]$, we define

$$\alpha_t^k(f) = \frac{\lambda_t^k(f)}{\Lambda^k(f)}.$$

From (S6.23), we have

$$\begin{aligned}
KL(f_0,f)-R_T &= \sum_k \mathbb{E}_0\left[\mathbb{1}_{\tilde\Omega_T}\left(\int_0^T \lambda_t^k(f_0)\log\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right)dt+\int_0^T(\lambda_t^k(f)-\lambda_t^k(f_0))dt\right)\right]\\
&= \sum_k \mathbb{E}_0\left[\mathbb{1}_{\tilde\Omega_T}\left(\Lambda_A^k(f_0)\int_{A^k(T)}\alpha_t^k(f_0)\log\left(\frac{\alpha_t^k(f_0)}{\alpha_t^k(f)}\right)dt+\Lambda^k(f_0)\log\left(\frac{\Lambda^k(f_0)}{\Lambda^k(f)}\right)+(\Lambda^k(f)-\Lambda^k(f_0))\right)\right]\\
&\leqslant \sum_k \mathbb{E}_0\left[\mathbb{1}_{\tilde\Omega_T}\left(\Lambda^k(f_0)\int_0^T\alpha_t^k(f_0)\log\left(\frac{\alpha_t^k(f_0)}{\alpha_t^k(f)}\right)dt+\frac{(\Lambda^k(f_0)-\Lambda^k(f))^2}{\Lambda^k(f_0)}\right)\right],
\end{aligned}$$

where in the last inequality we have used that $\chi(x)\leqslant(x-1)^2$ for $x\geqslant 1/2$, with $x=\frac{\Lambda^k(f)}{\Lambda^k(f_0)}$. In fact, we have

$$|\Lambda^k(f)-\Lambda^k(f_0)|\leqslant TL|\nu_k-\nu_k^0|+L\sum_l\left\|h_{lk}-h_{lk}^0\right\|_1 N^l[-A,T]\leqslant TL\epsilon_T(1+2\max_l\mu_l^0),$$

using that on $\tilde\Omega_T$, $N^l[-A,T]\leqslant T\mu_l^0+T\delta_T\leqslant 2T\mu_l^0$. Moreover, on $\tilde\Omega_T$, using the notations of Section 5.2, we have

$$\Lambda^k(f_0)\geqslant\phi_k(\nu_k^0)\sum_{j=1}^{J_T-1}(U_j^{(1)}-\tau_j)\geqslant\phi_k(\nu_k^0)\frac{T}{2\mathbb{E}_0[\Delta\tau_1]\|r_0\|_1}=:y_0T,$$

for some $y_0>0$. Similarly, for $f\in B_2(\epsilon_T,B)$ or $f\in B_\infty(\epsilon_T)$, we have

$$\Lambda^k(f)\geqslant\phi_k(\nu_k)\sum_{j=1}^{J_T-1}(U_j^{(1)}-\tau_j)\geqslant\phi_k(\nu_k^0/2)\frac{T}{2\mathbb{E}_0[\Delta\tau_1]\|r_0\|_1}.$$

Consequently,

$$\frac12\leqslant 1-\frac{|\Lambda^k(f)-\Lambda^k(f_0)|}{\Lambda_A(f_0)}\leqslant\frac{\Lambda^k(f)}{\Lambda^k(f_0)}\leqslant 1+\frac{|\Lambda^k(f)-\Lambda^k(f_0)|}{\Lambda_A(f_0)}\leqslant 1+\frac{1+2A\max_l\mu_l^0}{y_0}\epsilon_T=1+O(\epsilon_T),$$

for T large enough, and

$$\frac{(\Lambda^k(f)-\Lambda^k(f_0))^2}{\Lambda^k(f_0)}\leqslant\frac{L^2T^2\epsilon_T^2(1+2\max_l\mu_l^0)^2}{\Lambda^k(f_0)}\leqslant\frac{L^2T\epsilon_T^2(1+2A\max_l\mu_l^0)^2}{y_0}.$$

Additionally, on $\tilde\Omega_T$, on the one hand, for $f\in B_2(\epsilon_T,B)$, we also have that for any $t\in[0,T]$, since $\lambda_t^k(f_0)\leqslant\lambda_t^k(f)+\epsilon_T+BC_\beta\log T\implies\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\leqslant M_0\log T$ for some $M_0>0$, then

$$\frac{\alpha_t^k(f_0)}{\alpha_t^k(f)}=\frac{\lambda_t^k(f_0)\Lambda^k(f)}{\lambda_t^k(f)\Lambda^k(f_0)}\leqslant M_0\log T\frac{\Lambda^k(f)}{\Lambda^k(f_0)}\leqslant M\log T+O(M_0\log T\epsilon_T).$$

Applying Lemma 8.7 from [3], we have, for any $M\geqslant M_0$,

$$\int_0^T\alpha_t^k(f_0)\log\left(\frac{\alpha_t^k(f_0)}{\alpha_t^k(f)}\right)dt\leqslant\log(M\log T)\int_0^T\left(\sqrt{\alpha_t^k(f_0)}-\sqrt{\alpha_t^k(f)}\right)^2dt.$$

Moreover,

$$\int_0^T \left(\sqrt{\alpha_t^k(f_0)} - \sqrt{\alpha_t^k(f)}\right)^2 dt \leqslant \int_0^T \frac{1}{\Lambda^k(f_0)}\left(\sqrt{\lambda_t^k(f_0)} - \sqrt{\frac{\Lambda^k(f_0)}{\Lambda^k(f)}\lambda_t^k(f)}\right)^2 dt$$
$$\leqslant \frac{2}{\Lambda^k(f_0)}\int_0^T \left(\sqrt{\lambda_t^k(f_0)} - \sqrt{\lambda_t^k(f)}\right)^2 dt + \frac{1}{\Lambda^k(f_0)}\int_0^T \lambda_t^k(f)\left(1 - \sqrt{\frac{\Lambda^k(f_0)}{\Lambda^k(f)}}\right)^2 dt$$
$$\lesssim \frac{1}{\Lambda^k(f_0)}\int_0^T \left(\sqrt{\lambda_t^k(f_0)} - \sqrt{\lambda_t^k(f)}\right)^2 dt + \frac{(\Lambda^k(f) - \Lambda^k(f_0))^2}{\Lambda^k(f_0)^2}.$$

On the other hand, if $f \in B_\infty(\epsilon_T)$, then $\lambda_t^k(f_0) \leqslant \lambda_t^k(f)$ and we have

$$\int_0^T \alpha_t^k(f_0)\log\left(\frac{\alpha_t^k(f_0)}{\alpha_t^k(f)}\right)dt \leqslant \frac{2}{\Lambda^k(f_0)}\int_0^T \left(\sqrt{\lambda_t^k(f_0)} - \sqrt{\lambda_t^k(f)}\right)^2 dt + \frac{4(\Lambda^k(f) - \Lambda^k(f_0))^2}{\Lambda^k(f_0)^2}.$$

Moreover, in this case,

$$\int_0^T \left(\sqrt{\lambda_t^k(f_0)} - \sqrt{\lambda_t^k(f)}\right)^2 dt = \int_0^T \left(\sqrt{\phi_k(\tilde{\lambda}_t^k(\nu_0, h_0))} - \sqrt{\phi_k(\tilde{\lambda}_t^k(\nu, h))}\right)^2 dt$$
$$\leqslant L_1^2 \int_{A^k(T)} \left(\tilde{\lambda}_t^k(\nu_0, h_0) - \tilde{\lambda}_t^k(\nu, h)\right)^2 dt \lesssim T\epsilon_T^2.$$

Finally, we obtain that

$$KL(f_0, f) \lesssim \begin{cases} (\log\log T)T\epsilon_T^2 & \text{if} \quad f \in B_2(\epsilon_T, B) \\ T\epsilon_T^2 & \text{if} \quad f \in B_\infty(\epsilon_T) \end{cases}.$$

Scenario 3: Case 1 of Proposition 3.5, i.e., $\phi_k(x) = (x)_+, \forall k \in [K]$.

In a Hawkes model with the standard ReLU link function, we can obtain two types of rates, under and without condition (8). We consider $f \in B_\infty(\epsilon_T)$ so that $\forall t \in [0,T], \tilde{\lambda}_t^k(\nu, h) \geqslant \tilde{\lambda}_t^k(\nu_0, h_0)$. Since for any $t \in [0,T]$, $\log(\lambda_t^k(f_0)/\lambda_t^k(f)) \leqslant 0$, we can use that

$$KL(f_0, f) \leqslant \sum_k \mathbb{E}_0\left[\int_0^T (\lambda_t^k(f) - \lambda_t^k(f_0))dt\right] = \sum_k \mathbb{E}_0\left[\Lambda^k(f) - \Lambda^k(f_0)\right],$$

with for any $1 \leqslant k \leqslant K$, $\Lambda^k(f) := \int_0^T \lambda_t^k(f)dt$, and $\Lambda^k(f_0) := \int_0^T \lambda_t^k(f_0)dt$. Since for any t, $\tilde{\lambda}_t^k(\nu, h) \geqslant \tilde{\lambda}_t^k(\nu_0, h_0)$, we have

$$0 \leqslant \Lambda^k(f) - \Lambda^k(f_0) = \int_0^T ((\tilde{\lambda}_t^k(\nu, h))_+ - (\tilde{\lambda}_t^k(\nu_0, h_0))_+ dt \leqslant \int_0^T |\tilde{\lambda}_t^k(\nu, h) - \tilde{\lambda}_t^k(\nu_0, h_0)|dt$$
$$\leqslant T|\nu_k - \nu_k^0| + \sum_l \int_0^T \int_{t-A}^{t^-} |h_{lk} - h_{lk}^0|(t-s)dN_s^l dt \leqslant T(\nu_k - \nu_k^0) + \sum_l \left\|h_{lk} - h_{lk}^0\right\|_1 N^l[-A, T]. \tag{S6.26}$$

Consequently, we arrive at

$$KL(f_0, f) \leqslant KT\epsilon_T(1 + \max_l \mathbb{E}_0\left[N^l[-A, T]\right]) + R_T$$
$$\leqslant T\epsilon_T K(1 + 2\max_l \mu_l^0) + o(T\epsilon_T^2) \lesssim T\epsilon_T.$$

To refine this bound, we will assume that (8) holds. For $k \in [K]$ and $t \in [0, T]$, we define $p_t^k(f) = \lambda_t^k(f)/\Lambda^k(f)$ and similarly for $p_t^k(f_0)$. Using (S6.23), we then have

$$KL(f_0, f) - R_T = \sum_k \mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T}\left(\Lambda^k(f_0)\int_0^T \mathbb{1}_{\lambda_t^k(f_0)>0} p_t^k(f_0)\log\left(\frac{p_t^k(f_0)}{p_t^k(f)}\right)dt + \Lambda^k(f_0)\log\left(\frac{\Lambda^k(f_0)}{\Lambda^k(f)}\right) + (\Lambda^k(f) - \Lambda^k(f_0))\right)\right]$$
$$\leqslant \sum_k \mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T}\left(\Lambda^k(f_0)\int_0^T \mathbb{1}_{\lambda_t^k(f_0)>0} p_t^k(f_0)\log\left(\frac{p_t^k(f_0)}{p_t^k(f)}\right)dt + \frac{(\Lambda^k(f_0) - \Lambda^k(f))^2}{\Lambda^k(f_0)}\right)\right], \tag{S6.27}$$

where in the last inequality, we have used the fact that $-\log x + x - 1 \leqslant (x-1)^2$ for $x \geqslant 1/2$, with $x = \frac{\Lambda^k(f)}{\Lambda^k(f_0)} \geqslant 1$. Moreover, from (S6.26), we have on $\tilde{\Omega}_T$,

$$\Lambda^k(f) - \Lambda^k(f_0) \leqslant T\epsilon_T(1 + 2\max_l \mu_l^0).$$

Besides, on $\tilde{\Omega}_T$, using $A_2(T)$ defined in (22) and noting that in this case, $r_k^0 = \nu_k^0, \forall k$,

$$\Lambda^k(f_0) \geqslant \int_{A_2(T)} \lambda_t^k(f_0)dt \geqslant \sum_{j=1}^{J_T-1}\int_{\tau_j}^{U_j^{(1)}} \lambda_t^k(f_0)dt = \nu_k^0 \sum_{j=1}^{J_T-1}(U_j^{(1)} - \tau_j)$$
$$\geqslant \frac{\nu_k^0 T}{\mathbb{E}_0(\Delta\tau_1)\|\nu_0\|_1}\left(1 - 2c_\beta\sqrt{\frac{\log T}{T}}\right) \geqslant \frac{\nu_k^0 T}{2\mathbb{E}_0(\Delta\tau_1)\|\nu_0\|_1}.$$

Therefore,

$$\Lambda^k(f_0) \leqslant \Lambda^k(f) \leqslant \Lambda^k(f_0) + T\epsilon_T(1 + 2\max_l \mu_l^0)$$
$$\leqslant \Lambda^k(f_0) + \frac{2\Lambda^k(f_0)(1 + 2A\max_l \mu_l^0)\mathbb{E}_0(\Delta\tau_1)\|\nu_0\|_1}{\nu_k^0}\epsilon_T$$
$$\leqslant \Lambda^k(f_0)\left(1 + \frac{2(1 + 2A\max_l \mu_l^0)\mathbb{E}_0(\Delta\tau_1)\|\nu_0\|_1}{\nu_k^0}\epsilon_T\right) \leqslant 2\Lambda^k(f_0), \tag{S6.28}$$

for T large enough. Besides, this implies that $p_t^k(f) = \frac{\lambda_t^k(f)}{\Lambda^k(f)} \geqslant \frac{\lambda_t^k(f_0)}{2\Lambda^k(f_0)} \geqslant p_t^k(f_0)/2$. Using again the inequality $-\log x + x - 1 \leqslant (x-1)^2$ with $x = \frac{p_t^k(f)}{p_t^k(f_0)} \geqslant \frac{1}{2}$ and the fact that $\int_0^T p_t^k(f)dt = \int_0^T p_t^k(f_0)dt = 1$, we have

$$\int_0^T \mathbb{1}_{\lambda_t^k(f_0)>0} p_t^k(f_0)\log\left(\frac{p_t^k(f_0)}{p_t^k(f)}\right)dt = \int_0^T p_t^k(f_0)\log\left(\frac{p_t^k(f_0)}{p_t^k(f)}\right)dt + \int_0^T (p_t^k(f) - p_t^k(f_0))dt$$
$$= \int_0^T p_t^k(f_0)\left(\log\left(\frac{p_t^k(f_0)}{p_t^k(f)}\right) + \frac{p_t^k(f)}{p_t^k(f_0)} - 1\right)dt \leqslant \int_0^T \mathbb{1}_{\lambda_t^k(f_0)>0}\frac{(p_t^k(f_0) - p_t^k(f))^2}{p_t^k(f_0)}dt$$

$$\leqslant \frac{1}{\Lambda^k(f_0)} \int_0^T \mathbb{1}_{\lambda_t^k(f_0)>0} \frac{2\left(\lambda_t^k(f_0) - \lambda_t^k(f)\right)^2 + 2\lambda_t^k(f)^2 \left(1 - \frac{\Lambda^k(f_0)}{\Lambda^k(f)}\right)^2}{\lambda_t^k(f_0)} dt$$

$$\leqslant \frac{2}{\Lambda^k(f_0)} \left[\int_0^T \mathbb{1}_{\lambda_t^k(f_0)>0} \frac{3\left(\lambda_t^k(f_0) - \lambda_t^k(f)\right)^2}{\lambda_t^k(f_0)} + 2\Lambda^k(f_0) \times \frac{(\Lambda^k(f) - \Lambda^k(f_0))^2}{\Lambda^k(f)^2} \right]$$

$$\leqslant \frac{6}{\Lambda^k(f_0)} \int_0^T \mathbb{1}_{\lambda_t^k(f_0)>0} \frac{2\left(\lambda_t^k(f_0) - \lambda_t^k(f)\right)^2}{\lambda_t^k(f_0)} dt + 4\frac{(\Lambda^k(f) - \Lambda^k(f_0))^2}{\Lambda^k(f_0)^2}.$$

In the previous inequalities, we have used $\Lambda^k(f_0) \leqslant \Lambda^k(f)$, and for T large enough, we have the following intermediate result:

$$KL(f_0, f) - R_T \leqslant \sum_k \mathbb{E}_0 \left[\mathbb{1}_{\tilde{\Omega}_T} \left(6 \int_0^T \mathbb{1}_{\lambda_t^k(f_0)>0} \frac{(\lambda_t^k(f_0) - \lambda_t^k(f))^2}{\lambda_t^k(f_0)} dt + 4\frac{(\Lambda^k(f_0) - \Lambda^k(f))^2}{\Lambda^k(f_0)} \right) \right]. \quad \text{(S6.29)}$$

Moreover, on $\tilde{\Omega}_T$, using (S6.28)

$$\Lambda^k(f_0) = \int_0^T \left(\nu_k^0 + \sum_l \int_{t-A}^{t^-} h_{lk}^0(t-s) dN_s^l \right)_+ dt \leqslant T\nu_k^0 + \sum_l \|h_{lk}^{0+}\|_1 N^l[-A, T)$$

$$\leqslant T\nu_k^0 + \frac{3}{2}T \sum_l \|h_{lk}^{0+}\|_1 (\mu_l^0 + \delta_T) \leqslant 2T \left(\nu_k^0 + \sum_l \|h_{lk}^{0+}\|_1 \mu_l^0 \right),$$

for T large enough, since $\delta_T = \delta_0 \sqrt{\frac{\log T}{T}}$. Thus,

$$\frac{(\Lambda^k(f_0) - \Lambda^k(f))^2}{\Lambda^k(f_0)} \leqslant \Lambda^k(f_0) \left(\frac{2(1 + 2A \max_l \mu_l^0)\mathbb{E}_0(\Delta\tau_1)\|\nu_0\|_1}{\nu_k^0} \right)^2 \epsilon_T^2 \leqslant = c_2^0 T \epsilon_T^2,$$

with

$$c_2^0 = 8 \left(\nu_k^0 + \sum_l \|h_{lk}^{0+}\|_1 \mu_l^0 \right) \left(\frac{(1 + 2A \max_l \mu_l^0)\mathbb{E}_0(\Delta\tau_1)\|\nu_0\|_1}{\nu_k^0} \right)^2.$$

Therefore, reporting into (S6.29) we have

$$KL(f_0, f) - R_T \leqslant 6 \sum_k \mathbb{E}_0 \left[\mathbb{1}_{\tilde{\Omega}_T} \int_0^T \mathbb{1}_{\lambda_t^k(f_0)>0} \frac{(\lambda_t^k(f_0) - \lambda_t^k(f))^2}{\lambda_t^k(f_0)} dt \right] + 4Kc_2^0 T \epsilon_T^2.$$

We now bound the first term on the RHS of the previous equation.

$$\sum_k \mathbb{E}_0 \left[\mathbb{1}_{\tilde{\Omega}_T} \int_0^T \mathbb{1}_{\lambda_t^k(f_0)>0} \frac{(\lambda_t^k(f_0) - \lambda_t^k(f))^2}{\lambda_t^k(f_0)} dt \right] \leqslant \sum_k \mathbb{E}_0 \left[\mathbb{1}_{\Omega_{\tilde{\Omega}_T}} \sup_{t\in[0,T]} \mathbb{1}_{\lambda_t^k(f_0)>0} (\lambda_t^k(f) - \lambda_t^k(f_0))^2 \int_0^T \frac{\mathbb{1}_{\lambda_t^k(f_0)>0}}{\lambda_t^k(f_0)} dt \right].$$

Moreover, for any $k \in [K]$ and $t \in [0, T]$, we have on $B_\infty(\epsilon_T)$

$$\mathbb{1}_{\tilde{\Omega}_T} \mathbb{1}_{\lambda_t^k(f_0)>0} (\lambda_t^k(f) - \lambda_t^k(f_0))^2 dt \leqslant 2(\nu_k - \nu_k^0)^2 + 2K \max_l \|h_{lk} - h_{lk}^0\|_\infty^2 \sup_{t\in[0,T]} N^l[t-A, t)^2$$

$$\leqslant 2\epsilon_T^2 + 2KC_\beta^2 \log^2 T\epsilon_T^2 \leqslant 4KC_\beta^2 \log^2 T\epsilon_T^2.$$

Consequently,

$$\sum_k \mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T}\int_0^T \mathbb{1}_{\lambda_t^k(f_0)>0}\frac{(\lambda_t^k(f_0)-\lambda_t^k(f))^2}{\lambda_t^k(f_0)}dt\right] \leqslant 4C_\beta^2 K(\log T)^2 T\epsilon_T^2 \sum_k \mathbb{E}_0\left[\frac{1}{T}\int_0^T \frac{\mathbb{1}_{\lambda_t^k(f_0)>0}}{\lambda_t^k(f_0)}dt\right]$$
$$= 4C_\beta^2 c_1^0 K(\log T)^2 T\epsilon_T^2,$$

using (8), with

$$c_1^0 := \lim\sup_{T\to\infty} \mathbb{E}_0\left[\frac{1}{T}\int_0^T \frac{\mathbb{1}_{\lambda_t^k(f_0)>0}}{\lambda_t^k(f_0)}dt\right] < +\infty.$$

Consequently, reporting into (S6.29), we finally obtain

$$KL(f_0, f) \leqslant 4C_\beta^2 c_1^0 KL(\log T)^2 T\epsilon_T^2 + 4Kc_2^0 T\epsilon_T^2 + o(T\epsilon_T^2)$$
$$\leqslant 8KC_\beta^2 c_1^0 (\log T)^2 T\epsilon_T^2 = \kappa_2 (\log T)^2 T\epsilon_T^2,$$

with $\kappa_2 := 8KC_\beta^2 c_1^0$, which terminates the proof of this lemma.

□

S6.2. Deviations on the log likelihood ratio: Lemma S6.3

The next lemma is a control under $\mathbb{P}_0$ over the centered sum of i.i.d. variables that are used to decompose the log-likelihood ratio in Lemma A.2.

Lemma S6.3. *Under the assumptions of Lemma S6.1, for $f \in B_\infty(\epsilon_T)$ and $j \geqslant 1$, let*

$$T_j := \sum_k \int_{\tau_j}^{\tau_{j+1}} \log\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right) dN_t^k - \int_{\tau_j}^{\tau_{j+1}} (\lambda_t^k(f_0) - \lambda_t^k(f))dt. \tag{S6.30}$$

Then it holds that $\mathbb{E}_0\left[T_j^2\right] \lesssim z_T/T$, with

$$z_T = \begin{cases} T\epsilon_T^2 & \text{(under Assumption 3.1(i))} \\ (\log T)T\epsilon_T^2 & \text{(under Assumption 3.1(ii))} \\ (\log T)^2 T\epsilon_T^2 & \text{(ReLU link)} \end{cases}$$

Moreover, if $\log^3 T = O(z_T)$,

$$\mathbb{P}_0\left[\sum_{j=0}^{J_T-1} T_j - \mathbb{E}_0\left[T_j\right] \geqslant z_T\right] = o(1).$$

***Remark* S6.4.** Under Assumption 3.1, for $f \in B_2(\epsilon_T, B)$, we also obtain similar results with $z_T = (\log\log T)^2 T\epsilon_T^2$.

Proof. Firstly, using the fact that τ_1, τ_2 are stopping times, we have

$$\begin{aligned}
\mathbb{E}_0\left[T_1^2\right] &= \mathbb{E}_0\left[\left(\sum_k \int_{\tau_1}^{\tau_2} \log\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right) dN_t^k - \int_{\tau_1}^{\tau_2} (\lambda_t^k(f_0) - \lambda_t^k(f))dt\right)^2\right] \\
&\lesssim \sum_k \mathbb{E}_0\left[\left(\int_{\tau_1}^{\tau_2} \log\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right)\lambda_t^k(f_0)dt + \int_{\tau_1}^{\tau_2} \log\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right)(dN_t^k - \lambda_t^k(f_0)dt) - \int_{\tau_1}^{\tau_2} (\lambda_t^k(f_0) - \lambda_t^k(f))dt\right)^2\right] \\
&\lesssim \mathbb{E}_0\left[\Delta\tau_1 \int_{\tau_1}^{\tau_2} \chi\left(\frac{\lambda_t^k(f)}{\lambda_t^k(f_0)}\right)^2 \lambda_t^k(f_0)^2 dt\right] + \mathbb{E}_0\left[\int_{\tau_1}^{\tau_2} \log^2\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right)\lambda_t^k(f_0)dt\right], \qquad \text{(S6.31)}
\end{aligned}$$

with $\chi(x) = -\log x + x - 1$. For any $x > 0$, we have $\chi^2(x) \leqslant 2\log^2 x + 2(x-1)^2$. Now, if $f \in B_\infty(\epsilon_T)$, using that $\log^2 x \leqslant (x-1)^2$ for $x = \lambda_t^k(f)/\lambda_t^k(f_0) \geqslant 1$, we have $\chi\left(\frac{\lambda_t^k(f)}{\lambda_t^k(f_0)}\right)^2 \lambda_t^k(f_0)^2 \lesssim (\lambda_t^k(f_0) - \lambda_t^k(f))^2$ and $\log^2\left(\frac{\lambda_t^k(f)}{\lambda_t^k(f_0)}\right)\lambda_t^k(f_0) \lesssim \frac{(\lambda_t^k(f_0)-\lambda_t^k(f))^2}{\lambda_t^k(f_0)}$. Therefore, (S6.31) becomes

$$\begin{aligned}
\mathbb{E}_0\left[T_1^2\right] &\lesssim \mathbb{E}_0\left[\Delta\tau_1 \int_{\tau_1}^{\tau_2} (\lambda_t^k(f_0) - \lambda_t^k(f))^2 dt\right] + \mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T^c} \int_{\tau_1}^{\tau_2} \log^2\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right)\lambda_t^k(f_0)dt\right] \qquad \text{(S6.32)} \\
&\quad + \mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T} \int_{\tau_1}^{\tau_2} \mathbb{1}_{\lambda_t^k(f_0)>0} \frac{(\lambda_t^k(f_0) - \lambda_t^k(f))^2}{\lambda_t^k(f_0)} dt\right].
\end{aligned}$$

With the ReLU link function, we can easily bound the third term on the RHS of (S6.32) using (8):

$$\mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T} \int_{\tau_1}^{\tau_2} \mathbb{1}_{\lambda_t^k(f_0)>0} \frac{(\lambda_t^k(f_0) - \lambda_t^k(f))^2}{\lambda_t^k(f_0)} dt\right] \lesssim \log^2 T \epsilon_T^2 \mathbb{E}_0\left[\int_{\tau_1}^{\tau_2} \frac{\mathbb{1}_{\lambda_t^k(f_0)>0}}{\lambda_t^k(f_0)} dt\right] \lesssim \log^2 T \epsilon_T^2.$$

For the second term on the RHS of (S6.32), using that $\log^2(\lambda_t^k(f))\lambda_t^k(f) \lesssim (\sup_t N[t-A,t))^3$ and similarly for $\lambda_t^k(f_0)$, we have

$$\begin{aligned}
\mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T^c} \int_{\tau_1}^{\tau_2} \log^2\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right)\lambda_t^k(f_0)dt\right] &\lesssim \mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T^c} \int_{\tau_1}^{\tau_2} \log^2(\lambda_t^k(f_0))\lambda_t^k(f_0)dt\right] + \mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T^c} \int_{\tau_1}^{\tau_2} \log^2(\lambda_t^k(f))\lambda_t^k(f)dt\right] \\
&\lesssim \sqrt{\mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T^c} (\sup_t N[t-A,t))^6\right]} \sqrt{\mathbb{E}_0\left[\Delta\tau_1^2\right]} \lesssim T^{-\beta/2} = o(\epsilon_T^2),
\end{aligned}$$

using Lemma A.1. For the first term on the RHS of (S6.32), we have

$$\begin{aligned}
\mathbb{E}_0\left[\Delta\tau_1 \int_{\tau_1}^{\tau_2} (\lambda_t^k(f_0) - \lambda_t^k(f))^2 dt\right] &\lesssim \mathbb{E}_0\left[\Delta\tau_1 \int_{\tau_1}^{\tau_2} (\tilde{\lambda}_t^k(f_0) - \tilde{\lambda}_t^k(f))^2 dt\right] \\
&\leqslant \mathbb{E}_0\left[\Delta\tau_1 \int_{\tau_1}^{\tau_2} (2|\nu_k - \nu_k^0|^2 + 2K \sum_{l=1}^K \left(\int_{t-A}^t (h_{lk} - h_{lk}^0)(t-s)dN_s^l\right)^2 dt\right] \\
&\leqslant 2|\nu_k - \nu_k^0|^2 \mathbb{E}_0\left[\Delta\tau_1^2\right] + 2K \sum_{l=1}^K \mathbb{E}_0\left[\Delta\tau_1 \int_{\tau_1}^{\tau_2} N^l(t-A,t) \int_{t-A}^t (h_{lk} - h_{lk}^0)^2(t-s)dN_s^l dt\right] \\
&= 2|\nu_k - \nu_k^0|^2 \mathbb{E}_0\left[\Delta\tau_1^2\right] + 2K \sum_{l=1}^K \left\|h_{lk} - h_{lk}^0\right\|_2^2 \mathbb{E}_0\left[\Delta\tau_1 N^l[\tau_1,\tau_2)^2\right]
\end{aligned}$$

$$\leqslant 2|\nu_k - \nu_k^0|^2 \mathbb{E}_0\left[\Delta\tau_1\right] + 2K \sum_{l=1}^{K} \left\|h_{lk} - h_{lk}^0\right\|_2^2 \sqrt{\mathbb{E}_0\left[N^l[\tau_1,\tau_2)^4\right]} \sqrt{\mathbb{E}_0\left[\Delta\tau_1^2\right]} \lesssim \epsilon_T^2.$$

Thus, reporting into (S6.32), we can conclude that if (8) holds, $\mathbb{E}_0\left[T_1^2\right] \lesssim \log^2 T \epsilon_T^2$.

Under Assumption 3.1(i), if $f \in B_\infty(\epsilon_T)$, we can use the same computations. If $f \in B_2(\epsilon_T, B)$, for the first term on the RHS of (S6.32) and for the second term, we use instead that $\log^2 x \leqslant 4\log^2(r_T^{-1})(x-1)^2$ for $x \geqslant r_T$ with $x = \frac{\lambda_t^k(f_0)}{\lambda_t^k(f)} \gtrsim r_T := (\log T)^{-1}$ and we obtain,

$$\mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T} \int_{\tau_1}^{\tau_2} \log^2\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right) \lambda_t^k(f_0) dt\right] \lesssim (\log\log T)^2 \mathbb{E}_0\left[\int_{\tau_1}^{\tau_2} (\lambda_t^k(f_0) - \lambda_t^k(f))^2 dt\right]$$
$$\lesssim (\log\log T)^2 \mathbb{E}_0\left[\int_{\tau_1}^{\tau_2} (\tilde{\lambda}_t^k(\nu_0, h_0) - \tilde{\lambda}_t^k(\nu, h))^2 dt\right] \lesssim (\log\log T)^2 \epsilon_T^2,$$

or, in the shifted ReLU model with unknown link (Case 2 of Proposition 3.5),

$$\mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T} \int_{\tau_1}^{\tau_2} \log^2\left(\frac{\lambda_t^k(f_0,\theta_0)}{\lambda_t^k(f,\theta)}\right) \lambda_t^k(f_0,\theta_0) dt\right]$$
$$\lesssim (\log\log T)^2 \left[\mathbb{E}_0\left[\Delta\tau_1\right](\theta_k - \theta_k^0)^2 + \mathbb{E}_0\left[\int_{\tau_1}^{\tau_2} (\tilde{\lambda}_t^k(\nu_0, h_0) - \tilde{\lambda}_t^k(\nu, h))^2 dt\right]\right] \lesssim (\log\log T)^2 \epsilon_T^2,$$

using similar computations to the control of the first term of (S6.32). The remaining term, i.e.,

$$\mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T^c} \int_{\tau_1}^{\tau_2} \log^2\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right) \lambda_t^k(f_0) dt\right],$$

is bounded as the second term of (S6.32).

Finally, under Assumption 3.1(ii), using the fact that $\log \phi_k$ L_1-Lipschitz for any k, we have

$$\mathbb{E}_0\left[\int_{\tau_1}^{\tau_2} \log^2\left(\frac{\lambda_t^k(f_0)}{\lambda_t^k(f)}\right) \lambda_t^k(f_0) dt\right] \lesssim \mathbb{E}_0\left[\int_{\tau_1}^{\tau_2} (\tilde{\lambda}_t^k(\nu_0, h_0) - \tilde{\lambda}_t^k(\nu, h))^2 \lambda_t^k(f_0) dt\right]$$
$$\lesssim \log T \mathbb{E}_0\left[\int_{\tau_1}^{\tau_2} (\tilde{\lambda}_t^k(\nu_0, h_0) - \tilde{\lambda}_t^k(\nu, h))^2 dt\right] + \mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T^c} \int_{\tau_1}^{\tau_2} (\tilde{\lambda}_t^k(f_0) - \tilde{\lambda}_t^k(f))^2 \lambda_t^k(f_0) dt\right]$$
$$\lesssim (\log T)\epsilon_T^2,$$

and the first term of (S6.31) can be bounded similarly.

We now prove the second part of the lemma. We first note that

$$\mathbb{P}_0\left[\sum_{j=0}^{J_T-1} T_j - \mathbb{E}_0\left[T_j\right] \geqslant z_T\right] \leqslant \sum_{J \in \mathcal{J}_T} \mathbb{P}_0\left(\sum_{j=0}^{J-1} T_j - \mathbb{E}_0\left[T_j\right] \geqslant z_T\right) + \mathbb{P}_0\left(\tilde{\Omega}_T^c\right)$$
$$\leqslant T \mathbb{P}_0\left(\sum_{j=0}^{J-1} T_j - \mathbb{E}_0\left[T_j\right] \geqslant z_T\right) + o(1). \tag{S6.33}$$

Let $J \in \mathcal{J}_T$. Since the $\{T_j\}_{1 \leqslant j \leqslant J}$ are i.i.d.. random variables, we apply Fuk-Nagaev inequality (see Proposition S9.3) to the sum of centered variables $T_j - \mathbb{E}[T_j]$ with $\lambda := z_T$ and $x := x_T$ with $x_T \to \infty$ a

sequence determined later. We denote $v := J\mathbb{E}_0\left[T_1^2\right] \leqslant T\mathbb{E}_0\left[T_1^2\right] \lesssim z_T$. Hence, we have $x\lambda/v = x_T z_T/v \gtrsim x_T$. Since $x_T \to \infty$,

$$\left(1+\frac{x\lambda}{v}\right)\log\left(1+\frac{x\lambda}{v}\right) - \frac{x\lambda}{v} \geqslant \frac{x_T\lambda}{v}.$$

From Fuk-Nagaev inequality, we have

$$\mathbb{P}_0\left(\sum_{j=1}^{J}(T_j - \mathbb{E}[T_j]) \geqslant z_T\right) \leqslant J\mathbb{P}_0\left[T_1 - \mathbb{E}[T_1] \geqslant x_T\right] + \exp\left\{-\frac{z_T}{x_T}\right\}. \tag{S6.34}$$

We note that in the second term on the RHS of (S6.34), if $\frac{z_T}{x_T} \geqslant x_0 \log T$ with $x_0 > 0$ large enough, then $\exp\left\{-\frac{z_T}{x_T}\right\} = o(\frac{1}{T})$. Since by assumption, $\log T = o(T\epsilon_T^2)$, then we can choose $x_T = x_0' \frac{z_T}{\log T} \to \infty$ with $x_0' > 0$ a constant small enough. For the first term on the RHS of (S6.34), let us consider $j \in [J]$. From (S6.30), we have

$$T_1 \leqslant \sum_k \left\{\int_{\tau_1}^{\tau_2} |\lambda_t^k(f) - \lambda_t^k(f_0)|dt + \int_{[\tau_1,\tau_2)} |\log\lambda_t^k(f) - \log\lambda_t^k(f_0)|dN_t^k\right\}.$$

Using the first part of the lemma and Cauchy-Schwarz inequality, we have that $\mathbb{E}_0[T_1] \leqslant \sqrt{\frac{z_T}{T}} \leqslant x_T$ since $x_T \gtrsim z_T/\log T$ and $\log^3 T = O(z_T)$. Therefore,

$$\mathbb{P}_0\left[T_1 - \mathbb{E}_0[T_1] \geqslant x_T\right] \leqslant \mathbb{P}_0\left[\tilde{\Omega}_T \cap \left\{\int_{\tau_1}^{\tau_2} |\lambda_t^k(f) - \lambda_t^k(f_0)|dt + \int_{[\tau_1,\tau_2)} |\log\lambda_t^k(f) - \log\lambda_t^k(f_0)| \geqslant x_T\right\}\right] + \mathbb{P}_0\left[\tilde{\Omega}_T^c\right].$$

On the one hand, on $\tilde{\Omega}_T$, under Assumption 3.1(i), using that $|\log x - \log y| \leqslant \frac{|x-y|}{y}$ for $x \geqslant y$,

$$\begin{aligned}
\int_{[\tau_1,\tau_2)} |\log\lambda_t^k(f) - \log\lambda_t^k(f_0)|dN_t^k &\leqslant \frac{2}{\min_k \inf_x \phi_k(x)} \int_{[\tau_1,\tau_2)} |\log\lambda_t^k(f) - \log\lambda_t^k(f_0)|dN_t^k \\
&\leqslant \frac{2LN[\tau_1,\tau_2)}{\min_k \inf_x \phi_k(x)}|\nu_k - \nu_k^0| + \frac{2L}{\min_k \inf_x \phi_k(x)} \int_{[\tau_1,\tau_2)^2} |h_{lk} - h_{lk}^0|(t-s)dN_t^k dN_s^k \\
&\leqslant \frac{4L}{\min_k \inf_x \phi_k(x)}(\epsilon_T N[\tau_1,\tau_2) + N[\tau_1,\tau_2)^2 \left\|h_{lk} - h_{lk}^0\right\|_\infty) \leqslant 3LBN[\tau_1,\tau_2)^2,
\end{aligned}$$

for T large enough. In Case 2 of Proposition 3.5, we similarly have

$$\begin{aligned}
\int_{[\tau_1,\tau_2)} |\log\lambda_t^k(f) - \log\lambda_t^k(f_0)|dN_t^k &\leqslant \frac{2}{\theta_k^0} \int_{[\tau_1,\tau_2)} |\log\lambda_t^k(f) - \log\lambda_t^k(f_0)|dN_t^k \\
&\leqslant \frac{2N[\tau_1,\tau_2)}{\theta_k^0}(|\theta_k - \theta_k^0| + |\nu_k - \nu_k^0|) + \frac{2}{\theta_k^0} \int_{[\tau_1,\tau_2)} \int_{[\tau_1,\tau_2)} |h_{lk} - h_{lk}^0|(t-s)dN_t^k dN_s^k \\
&\leqslant \frac{4}{\theta_k^0}\epsilon_T N[\tau_1,\tau_2) + 2N[\tau_1,\tau_2)^2 \left\|h_{lk} - h_{lk}^0\right\|_\infty \leqslant 3BN[\tau_1,\tau_2)^2,
\end{aligned}$$

Under Assumption 3.1(ii), $\log\phi_k$ is L_1-Lipschitz, therefore,

$$\sum_{t_i \in [\tau_1,\tau_2)} |\log\lambda_{t_i}^k(f) - \log\lambda_{t_i}^k(f_0)| \leqslant L_1 \sum_{t_i \in [\tau_1,\tau_2)} |\tilde{\lambda}_{t_i}^k(\nu,h) - \tilde{\lambda}_{t_i}^k(\nu_0,h_0)| \leqslant L_1 BN[\tau_1,\tau_2)^2.$$

With the ReLU link function, we directly have that $T_1 \leqslant \sum_k \int_{\tau_1}^{\tau_2} (\lambda_t^k(f) - \lambda_t^k(f_0))dt$.
In Case 2 of Proposition 3.5,

$$\int_{\tau_1}^{\tau_2} |\lambda_t^k(f,\theta) - \lambda_t^k(f_0,\theta_0)|dt \leqslant |\theta_k^0 - \theta_k|\Delta\tau_1 + \int_{\tau_1}^{\tau_2} (\tilde{\lambda}_t^k(\nu,h) - \tilde{\lambda}_t^k(\nu_0,h_0))dt$$
$$\leqslant (|\theta_k^0 - \theta_k| + |\nu_k - \nu_k^0|)\Delta\tau_1 + \sum_l \left\|h_{lk} - h_{lk}^0\right\|_1 N^l[\tau_1,\tau_2) \leqslant [2\Delta\tau_1 + N[\tau_1,\tau_2)]\epsilon_T.$$

and in all other cases,

$$\int_{\tau_1}^{\tau_2} |\lambda_t^k(f) - \lambda_t^k(f_0)|dt \leqslant L\int_{\tau_1}^{\tau_2} (\tilde{\lambda}_t^k(\nu,h) - \tilde{\lambda}_t^k(\nu_0,h_0))dt$$
$$\leqslant L|\nu_k - \nu_k^0|)\Delta\tau_1 + L\sum_l \left\|h_{lk} - h_{lk}^0\right\|_1 N^l[\tau_1,\tau_2) \leqslant L[2\Delta\tau_1 + N[\tau_1,\tau_2)]\epsilon_T.$$

Consequently,

$$T_1 \leqslant KC[2\Delta\tau_1 + N[\tau_1,\tau_2)]\epsilon_T + 3KCBN[\tau_1,\tau_2)^2 \leqslant 4KCBN[\tau_1,\tau_2)^2,$$

with $C = \max(1, L, L_1)$ or $C = \max(1, L)$ depending on the assumptions on the link functions, and

$$\mathbb{P}_0\left[T_1 - \mathbb{E}_0[T_1] \geqslant 2x_T\right] \leqslant \mathbb{P}_0\left[N[\tau_1,\tau_2)^2 > \frac{x_T}{2KCB}\right].$$

Using Lemma 5.1, we have for some $s > 0$

$$\mathbb{P}_0\left[N[\tau_1,\tau_2)^2 > \frac{x_T}{2KCB}\right] \leqslant \mathbb{E}_0\left[e^{sN[\tau_1,\tau_2)}\right] e^{-s\sqrt{x_T/(2KCB)}} = o(T^{-2}),$$

if $x_T \geqslant x_0'' \log^2 T$ for some $x_0'' > 0$ large enough, implying that $z_T \geqslant z_0 \log^3 T$ for some $z_0 > 0$. Finally, reporting into (S6.33), we can conclude that

$$\mathbb{P}_0\left(\sum_{j=1}^{J_T} (T_j - \mathbb{E}[T_j]) \geqslant z_T\right) \leqslant T^2\mathbb{P}_0\left[T_1 - \mathbb{E}[T_1] \geqslant x_T\right] + T\mathbb{P}_0\left[\tilde{\Omega}_T^c\right] + T\exp\left\{-\frac{z_T}{x_T}\right\} + o(1) = o(1).$$

□

S7. Proofs of identifiability results and regenerative properties of nonlinear Hawkes models

S7.1. Proofs of Proposition 2.3, Proposition 2.5 and Lemma 2.6

In this section, we prove our two propositions on the model identifiability, i.e., Propositions 2.3 and 2.5, as well as Lemma 2.6 in the mutually-exciting Hawkes model. We recall the results in each case.

Proposition S7.1 (Proposition 2.3). *Let N be a nonlinear Hawkes process as defined in (1) with link functions $(\phi_k)_k$ and parameter $f = (\nu, h)$ satisfying the conditions of Lemma 2.1 and Assumption 2.2. If N' is a Hawkes processes with the same link functions $(\phi_k)_k$ and parameter $f' = (\nu', h')$, then*

$$N \overset{\mathcal{L}}{=} N' \implies \nu = \nu' \quad \text{and} \quad h = h'.$$

Proof. Let $f' = (\nu', h')$ and $N' \sim \mathbb{P}_{f'}$. We recall that $N \sim \mathbb{P}_f$ and $N \overset{\mathcal{L}}{=} N'$ is equivalent to $\lambda_t^l(f) = \lambda_t^l(f')$ for all $t > 0$ and $l \in [K]$. Let τ_1 be the first renewal time of the process N, as defined in Section 5.1. From the proof of Lemma 5.1, with $U_1^{(1)}$ the time of the first event after τ_1 and $V^{(1)} \in [K]$ the index of the component associated with this event, we have that $U_1^{(1)} \sim Exp(\|r_f\|_1) \perp\!\!\!\perp V_1^{(1)}$ with $r_f = (r_1^f, \dots, r_K^f)$ and $r_k^f = \phi_k(\nu_k), \forall k$, and

$$V_1^{(1)} \sim Multi\left(1; \frac{r_1^f}{\|r_f\|_1}, \dots, \frac{r_K^f}{\|r_f\|_1}\right).$$

Therefore we can conclude that

$$N \overset{\mathcal{L}}{=} N' \implies r_f = r_{f'} \iff \phi_k(\nu_k) = \phi_k(\nu_k'),\ \forall k \in [K]. \tag{S7.35}$$

Since for all k, $\nu_k \in I_k$ defined by Assumption 2.2 (ii), then $\nu_k' = \phi_k^{-1}(\phi_k(\nu_k))$ and since ϕ_k is monotone non-decreasing, we obtain $\nu_k = \nu_k', \forall k$.

Moreover, for each $k \in [K]$, we define the event Ω_k as

$$\Omega_k = \left\{ \max_{k' \neq k} N^{k'}[\tau_1, \tau_2) = 0, N^k[\tau_1, \tau_1 + A] = 1, N^k[\tau_1 + A, \tau_2) = 0 \right\}.$$

On Ω_k, for $t \in [\tau_1, \tau_2) \cap [U_1^{(1)}, U_1^{(1)} + A]$ and $l \in [K]$, $\lambda_t^l(f) = \phi_l(\nu_l + h_{kl}(t - U_1^{(1)}))$ and similarly for $\lambda_t^l(f')$. Then, for any $s = t - U_1^{(1)} \in [0, A]$, $\lambda^l_{U_1^{(1)}+s}(f) = \phi_l(\nu_l + h_{kl}(s)) = \phi_l(\nu_l + h'_{kl}(s))$. Consequently, using that ϕ_l is injective on I_l, $h_{kl} = h'_{kl}$ for all $1 \leqslant k, l \leqslant K$ which concludes the proof of this proposition. □

Proposition S7.2. *Proposition 2.5 Let N be a Hawkes process with parameter $f = (\nu, h)$ and link function $\phi_k(x; \theta_k) = \theta_k + \psi_k(x)$ with $\theta_k \geqslant 0$ for any $k \in [K]$ satisfying the conditions of Lemma 2.1 and Assumption 2.2. We also assume that for all $k \in [K]$, $\lim_{x \to -\infty} \psi_k(x) = 0$ and*

$$\exists l \in [K],\ x_1 < x_2, \quad \text{such that } h_{lk}^-(x) > 0, \quad \forall x \in [x_1, x_2]. \tag{S7.36}$$

Then if N' is a Hawkes processes with link functions $\phi_k(x; \theta_k') = \theta_k' + \psi_k(x)$, $\theta_k' \geqslant 0$ and parameter $f' = (\nu', h')$,

$$N \overset{\mathcal{L}}{=} N' \implies \nu = \nu', \quad h = h', \quad \text{and} \quad \theta = \theta', \quad \theta = (\theta_k)_{k=1}^K, \quad \theta' = (\theta_k')_{k=1}^K.$$

Besides, in this case we have $\mathbb{P}_0\left[\inf_{t \geqslant 0} \lambda_t^k(f, \theta) = \theta_k\right] = 1$.

Proof. Using the proof of Proposition 2.3, we first obtain that $\phi_k(\nu_k) = \phi_k(\nu_k')$, therefore

$$\theta_k + \psi_k(\nu_k) = \theta_k' + \psi_k(\nu_k'),\ \forall k \in [K].$$

Secondly, we also have that $\theta_l + \psi_k(\nu_l + h_{kl}(s)) = \theta_l' + \psi_k(\nu_l' + h'_{kl}(s))$ for any $s \in [0, A]$ and all $1 \leqslant k, l \leqslant K$.

We first prove that $\theta = \theta'$ and from the latter we can deduce that $\nu = \nu'$ and finally that $h = h'$ by the injectivity of ψ_k on I_k, for any k. The proof of the identification of θ relies on the construction of

a specific excursion for each $k \in [K]$ in which there exists $t > 0$ such that $\lambda_t^k(f) \in [\theta_k, \theta_k + \epsilon]$ for any $\epsilon > 0$. From that, we will deduce that $N \overset{\mathcal{L}}{=} N' \implies \theta = \theta'$.

Let $k \in [K]$ and consider $l \in [K]$ such that h_{lk} satisfies Assumption S7.36. We first note that

$$\lambda_t^k(f) = \theta_k + \psi_k(\tilde{\lambda}_t^k(\nu, h)) \geqslant \theta_k.$$

Thus, we directly have that $\theta_k \leqslant \inf_{t>0} \lambda_t^k(f)$, a.s. Let $\epsilon > 0$. Using Assumption S7.36 (i), $\exists M > 0, \forall x \leqslant M$, $\psi_k(x) \leqslant \epsilon$. Using now Assumption S7.36 (ii), let $l \in [K]$ and $x_1 < x_2$ such that $[x_1, x_2] \subset B_0 := \{x \in [0, A], h_{lk}(x) \leqslant -c_*\}$. Define $n_1 = \min\{n \in \mathbb{N};\ nc_* > \nu_k^0 - M\}$, $\delta' = (x_2 - x_1)/3$, and we consider an excursion, which we write $[0, \tau]$, and which satisfies

$$\mathcal{E} = \{N[0, \delta'] = N^l[0, \delta'] = n_1,\ N[\delta', \delta' + A] = 0\}.$$

In other words the events only occur on the l-th component of the Hawkes process and only on $[0, \delta']$. Since ψ_k is Lipschitz and injective on $I_k = (\nu_k - \max_l \left\|h_{lk}^-\right\|_\infty - \varepsilon, \nu_k + \max_l \left\|h_{lk}^+\right\|_\infty + \varepsilon)$, it holds that $\mathbb{P}_f[\mathcal{E}] > 0$. For $t \in [x_1 + \delta', x_2]$, $\forall i \in [n_1]$, we have $x_1 \leqslant t - t_i \leqslant x_2$, and therefore,

$$\tilde{\lambda}_t^k(\nu, h) = \nu_k + \sum_{i \in [n_1]} h_{lk}(t - t_i) \leqslant \nu_k - n_1 c_* \leqslant M.$$

Consequently, for $t \in [x_1 + \delta', x_2]$, $\lambda_t^k(f_0) = \theta_k + \psi_k(\tilde{\lambda}_t^k(\nu, h)) \leqslant \theta_k + \epsilon$. We can then conclude that

$$\mathbb{P}_0\left[\exists t \geqslant 0,\ \lambda_t^k(f) \in [\theta_k, \theta_k + \epsilon]\right] > 0,$$

for any $\epsilon > 0$. This is equivalent to

$$\theta_k = \inf_{\omega \in \Omega} \inf_{t \in [0, \tau]} \lambda_t^k(f)(\omega),$$

where $\lambda_t^k(f_0)(\omega)$ denotes the value of the random process $(\lambda_t(f_0))_t$ at time t.

Now, if N' is a Hawkes process with parameter $f' \in \mathcal{F}$ and link functions $\phi_k = \theta_k' + \psi_k$, $k \in [K]$ such that $N \overset{\mathcal{L}}{=} N'$, then for any $t \geqslant 0$ and k such $\lambda_t^k(f) \leqslant \theta_k + \epsilon$, we have $\theta_k' \leqslant \lambda_t^k(f') \leqslant \theta_k + \epsilon$ and thus, $\theta_k \geqslant \theta_k'$. Inversely, if $\lambda_t^k(f') \leqslant \theta_k' + \epsilon$ then $\theta_k \leqslant \theta_k'$ and finally we can conclude that $\theta = \theta'$.

□

Lemma S7.3 (Lemma 2.6)**.** *Let N be a Hawkes process with parameter $f = (\nu, h)$ and link functions $\phi_k(x; \theta_k) = \theta_k + (x)_+$, $\theta_k \geqslant 0$, $k \in [K]$ satisfying Assumption 2.2, and let $k \in [K]$. If $\forall l \in [K], h_{lk} \geqslant 0$, then for any $\theta_k' \geqslant 0$ such that $\theta_k + \nu_k - \theta_k' > 0$, let N' be the Hawkes process driven by the same underlying Poisson process Q as N (see Lemma S9.2) with parameter $f' = (\nu', h')$ and link functions $\phi_k(x; \theta_k') = \theta_k' + (x)_+, k \in [K]$ with $\nu' = (\nu_1, \dots, \nu_k + \theta_k - \theta_k', \dots, \nu_K) \neq \nu$, $h' = h$, and $\theta' = (\theta_1, \dots, \theta_k', \dots, \theta_K) \neq \theta$. Then for any $t \geqslant 0$, $\lambda_t^k(f, \theta) = \lambda_t^k(f', \theta')$, and therefore $N \overset{\mathcal{L}}{=} N'$.*

Proof. We consider $k \in [K]$ such that $\forall l \in [K]$, $h_{lk} \geqslant 0$. For any $t \geqslant 0$, we have

$$\tilde{\lambda}_t^k(\nu, h) = \nu_k + \sum_l \int_{t-A}^{t^-} h_{lk}(t - s) dN_s^l \geqslant \nu_k > 0,$$

and thus $\lambda_t^k(f) = \theta_k + (\tilde{\lambda}_t^k(\nu, h))_+ = \theta_k + \tilde{\lambda}_t^k(\nu, h)$. Moreover, for any $t \geqslant 0$, we have

$$\tilde{\lambda}_t^k(\nu', h') = \nu_k + \theta_k - \theta'_k + \sum_l \int_{t-A}^{t^-} h_{lk}(t-s) dN_s^l \geqslant \nu_k + \theta_k - \theta'_k > 0,$$

and

$$\begin{aligned}\lambda_t^k(f') &= \theta'_k + (\tilde{\lambda}_t^k(\nu', h'))_+ = \theta'_k + \tilde{\lambda}_t^k(\nu', h') \\ &= \theta'_k + \nu_k + \theta_k - \theta'_k + \sum_l \int_{t-A}^{t^-} h_{lk}(t-s) dN_s^l = \theta_k + \tilde{\lambda}_t^k(\nu, h) = \lambda_t^k(f).\end{aligned}$$

Therefore, we obtain that $N =^{\mathcal{L}} N'$.

□

S7.2. Proofs of Lemmas 5.2 and 5.4

In this section, we prove our lemmas related to the renewal properties of the nonlinear Hawles processes, in particular the existence of exponential moments for the generic renewal time $\Delta\tau_1$, and a concentration inequality on J_T. the nunber of excursions in the interval of observation $[0, T]$.

Lemma S7.4 (Lemma 5.2)**.** *Under the assumptions of Lemma 5.1, the random variables $\Delta\tau_1$ and $N[\tau_1, \tau_2)$ admit exponential moments. More precisely, under condition **(C1bis)**, with $m = \|S^+\| < 1$, we have*

$$\forall s < \min(\|r_f\|_1, \gamma/A), \quad \mathbb{E}_f\left[e^{s\Delta\tau_1}\right] \leqslant \frac{1+m}{2m}, \quad \textit{and} \quad \mathbb{E}_f\left[e^{sN[\tau_1,\tau_2)}\right] < +\infty, \quad \gamma = \frac{1-m}{2\sqrt{K}} \log\left(\frac{1+m}{2m}\right).$$

*Under condition **(C2)**, we have $\forall s < \min_k \Lambda_k$, $\mathbb{E}_f\left[e^{s\Delta\tau_1}\right] \leqslant \frac{\|\Lambda\|_1^2}{(\min_k \Lambda_k - s)^2}$ and $\mathbb{E}_f\left[e^{sN[\tau_1,\tau_2)}\right] < +\infty$. In particular, this implies that $\mathbb{E}_f\left[N[\tau_1,\tau_2) + N[\tau_1,\tau_2)^2\right] < +\infty$.*

Proof. Under condition **(C1bis)**, similarly to [1], we use the fact that the multivariate Hawkes model is stochastically dominated by a mutually-exciting process N^+ with parameter $f^+ = (\nu, (h_{lk}^+)_{l,k})$, and driven by the same Poisson process as N (see Lemma S9.2). For N^+, the stopping time $\Delta\tau_1^+$ corresponds to the length of the busy period of a $M^K/G^K/\infty$ queue (see Lemma S9.1, which is a multi-type extension of existing results).

More precisely, since N^+ is mutually-exciting, the cluster representation is available [6], with the ancestor arrival process being a Poisson Point Process equal to the baseline rate r_f, defined in (21). For this process, the duration of the clusters then corresponds to the generic service time H of a queue with an infinite number of servers. In the multidimensional case, this duration may depend on the type of the ancestor (or "customer" in the queuing framework) but the generic service time can be written in a compact form, and is independent of the arrival process

$$H = \sum_{k=1}^K \delta_k H^k,$$

where $\delta_k = 1$ if and only if the ancestor is of type $k \in [K]$. To apply Lemma S9.1, we only need to check that the cluster length H^k, $k \in [K]$ has exponential moments. This can be proved using results from [2].

For the process N^+, let W^k be the number of events in a cluster with an ancestor of type k. By definition of a cluster of events, $H^k \leqslant AW^k$. Moreover, from Lemma 5 in the Supplementary Materials of [2], for a mutually-exciting Hawkes process and for any $t \leqslant \frac{1-\|S^+\|_1}{2\sqrt{K}} \log\left(\frac{1+\|S^+\|}{2\|S^+\|}\right)$ and $k \in [K]$,

$$\mathbb{E}_f\left[e^{tW^k}\right] \leqslant \frac{1+\|S^+\|}{2\|S^+\|}.$$

Therefore, we define $\gamma = (1-\|S^+\|)\left[\log\left(1+\|S^+\|\right) - \log(2\|S^+\|)\right]/(2\sqrt{K})$ and $s_0 = \frac{1+\|S^+\|}{2\|S^+\|}$. For all $0 < t \leqslant \gamma$, we thus have $\mathbb{E}_f\left[e^{tH^k/A}\right] \leqslant s_0$. Consequently, we deduce that the service time H^k has exponential tails, i.e., $\mathbb{P}_f\left[H^k \geqslant t\right] \leqslant s_0 e^{-t\gamma/A}$. We can now use the fact that a.s. $\mathcal{T}_1 = \Delta\tau_1^+$ (cf Lemma S9.2), so that for any $s < \|r_f\|_1 \wedge \gamma/A$, we have $\mathbb{E}_f\left[e^{s\Delta\tau_1^+}\right] < \infty$. Finally using the second part of Lemma S9.2, we have that $\mathbb{P}_f\left[\Delta\tau_1 \leqslant \Delta\tau_1^+\right] = 1$ and, using Lemma S9.1, we arrive at $\forall s < \|r_f\|_1 \wedge \gamma/A$, $\mathbb{E}_f\left[e^{s\Delta\tau_1}\right] < \infty$.

Under condition **(C2)**, we use the fact that the process N is dominated by a K-dimensional homogeneous Poisson point process $N_P = (N_P^1, \dots, N_P^K)$ with rate $\Lambda = (\Lambda_1, \dots, \Lambda_K)$. For the latter process, the generic service time of an ancestor of type k, H_k, is exponentially distributed with mean Λ_k, i.e.,

$$\mathbb{P}_f\left[H_k > t\right] = e^{-\Lambda_k t}, \quad t \geqslant 0.$$

Therefore, denoting $\Delta\tau_1^P$, the corresponding generic stopping time of N^P - with the same definition as in Lemma 5.1 for the Hawkes process (note that the Poisson point process is a renewal process), we have

$$\mathbb{P}_f\left[\Delta\tau_1^P > t\right] \leqslant \mathbb{E}_f\left[N^P[0,t]\right] e^{-\min_k \Lambda_k (t-A)} = \|\Lambda\|_1 \, t e^{-\min_k \Lambda_k (t-A)}.$$

Therefore, for any $s < \min_k \Lambda_k$,

$$\begin{aligned}\mathbb{E}_f\left[e^{s\Delta\tau_1}\right] \leqslant \mathbb{E}_f\left[e^{s\Delta\tau_1^P}\right] &= \int_0^{+\infty} s e^{st} \mathbb{P}_f\left[\Delta\tau_1^P \geqslant t\right] dt \leqslant \|\Lambda\|_1^2 \, e^{\min_k \Lambda_k A} \int_0^{+\infty} t e^{t(s-\min_k \Lambda_k)} dt \\ &\leqslant \|\Lambda\|_1^2 \int_0^{+\infty} \frac{e^{t(s-\min_k \Lambda_k)}}{\min_k \Lambda_k - s} dt = \frac{\|\Lambda\|_1^2}{(\min_k \Lambda_k - s)^2}.\end{aligned}$$

We now consider the number of events in a excursion $N[\tau_1, \tau_2)$. Under condition **(C1bis)**, From Lemma S9.2, we can also deduce that $\mathbb{E}_f\left[N[\tau_1, \tau_2)\right] \leqslant \mathbb{E}_f\left[N^+[\tau_1^+, \tau_2^+)\right]$. We once again use the cluster representation available for N^+. For the latter, let n^τ be the number of ancestors arriving in $[\tau_1^+, \tau_2^+)$ and W_i be the number of points in the cluster with ancestor i for $1 \leqslant i \leqslant n_\tau$. We denote $(NP_t)_t$ the homogeneous Poisson process of intensity $\|r_f\|_1$ corresponding to the arrival times of the ancestors. By definition of τ_1^+, τ_2^+, we have

$$N^+[\tau_1^+, \tau_2^+) = \sum_{i=1}^{n_\tau} W_i. \tag{S7.37}$$

Let $\gamma > s > 0$ and $u < \|r_f\|_1 \wedge \gamma/A$. With $t = \mathbb{E}_f\left[e^{sW_1}\right] \leqslant s_0$, since the W_i's are independent conditionally on n_τ,

$$\begin{aligned}\mathbb{E}_f\left[e^{sN[\tau_1,\tau_2)}\right] &\leqslant \mathbb{E}_f\left[e^{s\sum_{i=1}^{n_\tau} W_i}\right] = \mathbb{E}_f\left[\mathbb{E}_f\left[e^{s\sum_{i=1}^{n_\tau} W_i}|n_\tau\right]\right] = \mathbb{E}_f\left[\mathbb{E}_0\left[e^{sW_1}\right]^{n_\tau}\right] = \mathbb{E}_f\left[\sum_{l=A}^{+\infty} e^{sn_\tau}\mathbb{1}_{\Delta\tau_1\in[l,l+1)}\right] \\ &\leqslant \sum_{l=A}^{+\infty}\mathbb{E}_f\left[e^{sNP[\tau_1,\tau_1+l+1)}\mathbb{1}_{\Delta\tau_1\geqslant l}\right] \leqslant \sum_{l=A}^{+\infty}\sqrt{\mathbb{E}_f\left[e^{2sNP[\tau_1,\tau_1+l+1)}\right]}\sqrt{\mathbb{P}_f\left[\Delta\tau_1 > l\right]} \\ &\leqslant \sqrt{\mathbb{E}_f\left[e^{u\Delta\tau_1}\right]}\sum_{l=A}^{+\infty}\sqrt{\mathbb{E}_f\left[e^{2sNP[\tau_1,\tau_1+l+1)}\right]}e^{-ul/2} = \sqrt{\mathbb{E}_f\left[e^{u\Delta\tau_1}\right]}\sum_{l=A}^{+\infty} e^{\|r_f\|_1(l+1)(e^{2s}-1)/2}e^{-ul/2},\end{aligned}$$

since NP is a homogoneous Poisson process with rate $\|r_f\|_1$. Moreover, since for any $\alpha \in (0,1)$, $\mathbb{E}_f\left[e^{\alpha sW_1}\right] = (\mathbb{E}_f\left[e^{\alpha sW_1}\right]^{1/\alpha})^\alpha \leqslant \mathbb{E}_f\left[e^{sW_1}\right]^\alpha \leqslant s_0^\alpha$, with $t' = \mathbb{E}_f\left[e^{\alpha sW_1}\right]$, we have that $\|r_f\|_1(l+1)(e^{2t'}-1) < u/2$ for α small enough. Consequently,

$$\mathbb{E}_f\left[e^{sN[\tau_1,\tau_2)}\right] \leqslant \sqrt{\mathbb{E}_f\left[e^{u\Delta\tau_1}\right]}\sum_{l=A}^{+\infty} e^{-ul/4} = \frac{\sqrt{\mathbb{E}_0\left[e^{u\Delta\tau_1}\right]}}{1-e^{-u/4}} < \infty.$$

In particular, this implies that $\mathbb{E}_f\left[N[\tau_1,\tau_2)\right] + \mathbb{E}_f\left[N[\tau_1,\tau_2)^2\right] < \infty$. Under condition **(C2)**, the dominating process N^+ is a homogeneous Poisson process with intensity $\Lambda = (\Lambda_1,\dots,\Lambda_K)$ and the previous computations remain valid by replacing r_f by Λ and with $W_i = 1$ for any $i \in [n_\tau]$ (since in this case each cluster only contains the "ancestor" event).

□

Lemma S7.5 (Lemma 5.4)**.** *Under the assumptions of Lemma 5.1, for any $\beta > 0$, there exists a constant $c_\beta > 0$ such that $\mathbb{P}_f\left[J_T \notin [J_{T,\beta,1}, J_{T,\beta,2}]\right] \leqslant T^{-\beta}$, with J_T defined in (19) and*

$$J_{T,\beta,1} = \left\lfloor \frac{T}{\mathbb{E}_f\left[\Delta\tau_1\right]}\left(1 - c_\beta\sqrt{\frac{\log T}{T}}\right)\right\rfloor, \quad J_{T,\beta,2} = \left\lfloor \frac{T}{\mathbb{E}_f\left[\Delta\tau_1\right]}\left(1 + c_\beta\sqrt{\frac{\log T}{T}}\right)\right\rfloor.$$

Proof. Let $c_\beta > 0$ and for $2 \leqslant j \leqslant J_T$, $B_j = \tau_j - \tau_{j-1} - \mathbb{E}_f\left[\Delta\tau_1\right]$. Using Lemma 5.1, the random variables $\{B_j\}_{2\leqslant j\leqslant J_T}$ are i.i.d.. By definition of $J_{T,\beta,2}$, we have

$$\frac{T}{\mathbb{E}_f\left[\Delta\tau_1\right]}\left(1 + c_\beta\sqrt{\frac{\log T}{T}}\right) - 1 < J_{T,\beta,2} \leqslant \frac{T}{\mathbb{E}_f\left[\Delta\tau_1\right]}\left(1 + c_\beta\sqrt{\frac{\log T}{T}}\right).$$

Therefore,

$$\begin{aligned}\mathbb{P}_f\left[J_T \geqslant J_{T,\beta,2}\right] &= \mathbb{P}_0\left[\tau_{J_{T,\beta,2}} \leqslant T\right] = \mathbb{P}_f\left[\tau_0 + \sum_{j=1}^{J_{T,\beta,2}} B_j \leqslant T - J_{T,\beta,2}\mathbb{E}_f\left[\Delta\tau_1\right]\right] \\ &= \mathbb{P}_f\left[\sum_{j=1}^{J_{T,\beta,2}} B_j \leqslant T - J_{T,\beta,2}\mathbb{E}_f\left[\Delta\tau_1\right]\right] \leqslant \mathbb{P}_f\left[\sum_{j=1}^{J_{T,\beta,2}} B_j \leqslant T - T\left(1 + c_\beta\sqrt{\frac{\log T}{T}}\right) + \mathbb{E}_f\left[\Delta\tau_1\right]\right]\end{aligned}$$

$$= \mathbb{P}_f\left[\sum_{j=0}^{J_{T,\beta,2}} B_j \leqslant -c_\beta\sqrt{T\log T} + \mathbb{E}_f[\Delta\tau_1]\right] \leqslant \mathbb{P}_f\left[\sum_{j=1}^{J_{T,\beta,2}} B_j \leqslant -\frac{c_\beta\sqrt{T\log T}}{2}\right].$$

We can now apply the Bernstein's inequality. Using Lemma 5.2, there exists $\alpha > 0$, such that $\mathbb{E}_f\left[e^{\alpha\Delta\tau_1}\right] < +\infty$. Since

$$\mathbb{E}_f\left[e^{\alpha\Delta\tau_1}\right] = \sum_{k=1}^{+\infty}\frac{\alpha^k\mathbb{E}_f\left[(\Delta\tau_1)^k\right]}{k!},$$

we therefore have that

$$\mathbb{E}_f\left[(\Delta\tau_1)^k\right] \leqslant \frac{k!}{\alpha^k}\mathbb{E}_f\left[e^{\alpha\Delta\tau_1}\right] = \frac{1}{2}k!\alpha^{-k+2}\times 2\frac{\mathbb{E}_f\left[e^{\alpha\Delta\tau_1}\right]}{\alpha^2}.$$

In particular, $\mathbb{E}_f\left[(\Delta\tau_1)^2\right] \leqslant 2\frac{\mathbb{E}_0\left[e^{\alpha\Delta\tau_1}\right]}{\alpha^2} =: v$. Consequently, with $b := 1/\alpha$, we obtain $\mathbb{E}_f\left[(\Delta\tau_1)^k\right] \leqslant \frac{1}{2}k!b^{k-2}v$, and therefore,

$$\mathbb{P}_f\left[J_T \geqslant J_{T,\beta,2}\right] \leqslant \exp\left\{\frac{-c_\beta^2 T\log T}{8(\sigma^2 + \frac{c_\beta}{2}\sqrt{T\log T}b)}\right\},$$

with

$$\sigma^2 = \sum_{j=1}^{J_{T,\beta,2}}\mathbb{V}_f(B_j) = J_{T,\beta,2}\mathbb{V}_f(\Delta\tau_1) \leqslant T\left(1 + c_\beta\sqrt{\frac{\log T}{T}}\right)\frac{\mathbb{E}_f\left[\Delta\tau_1^2\right]}{\mathbb{E}_f[\Delta\tau_1]} \leqslant 2T\frac{\mathbb{E}_f\left[\Delta\tau_1^2\right]}{\mathbb{E}_f[\Delta\tau_1]},$$

for T large enough. Therefore, $\sigma^2 + \frac{c_\beta}{2}\sqrt{T\log T}b \leqslant 4T\frac{\mathbb{E}_f\left[\Delta\tau_1^2\right]}{\mathbb{E}_f[\Delta\tau_1]}$ and

$$\mathbb{P}_f\left[J_T \geqslant J_{T,\beta,2}\right] \leqslant \exp\left\{\frac{-c_\beta^2\log T\mathbb{E}_f[\Delta\tau_1]}{32\mathbb{E}_f\left[\Delta\tau_1^2\right]}\right\} = o(T^{-\beta}),$$

for any $\beta > 0$, if $c_\beta > 0$ is chosen large enough. Consequently, with probability greater than $1 - \frac{1}{2}T^{-\beta}$, we have that $J_T \leqslant \frac{T}{\mathbb{E}_f[\Delta\tau_1]}\left(1 + c_\beta\sqrt{\frac{\log T}{T}}\right)$. Similarly, we obtain that

$$\mathbb{P}_f\left[J_T \leqslant J_{T,\beta,1}\right] \leqslant \mathbb{P}_f\left[\sum_{j=1}^{J_{T,\beta,1}} B_j \geqslant c_\beta\sqrt{T\log T}\right] \leqslant \exp\left\{\frac{-c_\beta^2 T\log T}{2(\sigma^2 + c_\beta\sqrt{T\log T}b)}\right\}$$
$$\leqslant \exp\left\{\frac{-c_\beta^2\log T\mathbb{E}_f[\Delta\tau_1]}{4\mathbb{E}_f\left[\Delta\tau_1^2\right]}\right\} = o(T^{-\beta}).$$

Finally, we conclude that with probability greater than $1 - T^{-\beta}$, $J_{T,\beta,1} \leqslant J_T \leqslant J_{T,\beta,2}$.

□

S8. Proof of lemmas A.1 and A.4

S8.1. Proof of Lemma A.1

Lemma S8.1 (Lemma A.1). *Let $Q > 0$. We consider $\tilde{\Omega}_T$ defined in (25) in Section 5.2. For any $\beta > 0$, we can choose C_β and c_β in the definition of $\tilde{\Omega}_T$ such that*

$$\mathbb{P}_0[\tilde{\Omega}_T^c] \leqslant T^{-\beta}.$$

Moreover, for any $1 \leqslant q \leqslant Q$, $\mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T^c} \max_l \sup_{t\in[0,T]} \left(N^l[t-A,t)\right)^q\right] \leqslant 2T^{-\beta/2}$. Finally, the previous results hold when replacing $\tilde{\Omega}_T$ by $\tilde{\Omega}'_T = \tilde{\Omega}_T \cap \Omega_A$ with Ω_A defined in Section 5.3 for the model with shifted ReLU link and unknown shift.

Proof. Let $\beta > 0$. From the definition of $\tilde{\Omega}_T$, we have that

$$\mathbb{P}_0[\tilde{\Omega}_T^c] \leqslant \mathbb{P}_0[\Omega_N^c] + 3\mathbb{P}_0[\Omega_J^c] + \mathbb{P}_0[\Omega_J \cap \Omega_U^c]. \tag{S8.38}$$

For the second term on the RHS of (S8.38), we can directly use Lemma 5.4, and we obtain $\mathbb{P}_0[\Omega_J^c] \leqslant \frac{1}{12}T^{-\beta}$ for c_β large enough. For the first term on the RHS of (S8.38), we use the same strategy as in [2]. Firstly we have

$$\mathbb{P}_0[\Omega_N^c] \leqslant \mathbb{P}_0\left[\max_{k\in[K]} \sup_{t\in[0,T]} N^k[t-A,t) > C_\beta \log T\right] + \sum_{k=1}^K \mathbb{P}_0\left[\left|\frac{N^k[0,T]}{T} - \mu_k^0\right| \geqslant \delta_T\right]. \tag{S8.39}$$

For the first term on the RHS of (S8.39), we use the coupling with the process N^+, i.e., the Hawkes process with parameter $f_0^+ = (\nu_0, h_0^+)$ driven by the same Poisson process. Then for any $l \in [K]$, $\sup_{t\in[0,T]} N^l[t-A,t) \leqslant \sup_{t\in[0,T]} (N^+)^l[t-A,t)$ and consequently,

$$\mathbb{P}_0\left[\max_{k\in[K]} \sup_{t\in[0,T]} N^k[t-A,t) > C_\beta \log T\right] \leqslant \mathbb{P}_0\left[\max_{k\in[K]} \sup_{t\in[0,T]} (N^+)^k[t-A,t) > C_\beta \log T\right].$$

Using Lemma 2 from [2], we obtain that for any $\beta > 0$, there exists $C_\beta > 0$ such that

$$\mathbb{P}_0\left[\max_{k\in[K]} \sup_{t\in[0,T]} (N^+)^k[t-A,t) > C_\beta \log T\right] \leqslant \frac{1}{4}T^{-\beta}.$$

For the second term on the RHS of (S8.39), we use the same arguments as in the proof of Lemma 3 in [2]. For $k \in [K]$, we have

$$\mathbb{P}_0\left[\left|\frac{N^k[0,T]}{T} - \mu_k^0\right| \geqslant \delta_T\right] \leqslant \mathbb{P}_0\left[\left|N^k[0,T] - \int_0^T \lambda_t^k(f_0)\right| \geqslant T\delta_T/2\right] + \mathbb{P}_0\left[\left|\int_0^T \lambda_t^k(f_0) - \mu_k^0 T\right| \geqslant T\delta_T/2\right]. \tag{S8.40}$$

For the second term on the RHS of (S8.40), we can use Corollary 1.1 from [1]. We have that $\lambda_t^k(f_0) = Z(S_t N)$, with

$$Z(N) = \lambda_0^k(f_0) = \phi_k\left(\nu_k^0 + \sum_l \int_{-A}^{0^-} h_{lk}(t-s)dN_s^l\right) \leqslant Lb(1 + N[-A,0)),$$

with $b = \max(\nu_k^0, \max_l \left\|h_{lk}^{0+}\right\|_\infty)$ and for $t \in \mathbb{R}$, $S_t : \mathcal{N}(\mathbb{R}) \to S_t N = N(.+t)$ the shift operator by t units of time. Applying Corollary 1.1 of [1] with $f = Z$, $\pi_A f = \mathbb{E}_0\left[\lambda_0^k(f_0)\right] = \mu_k^0$, $\varepsilon = \delta_T/2$ and $\eta = \frac{1}{4}T^{-\beta}$, we obtain that for δ_0 large enough,

$$\mathbb{P}_0\left[\left|\int_0^T \lambda_t^k(f_0) - \mu_k^0 T\right| \geqslant T\delta_T/2\right] \leqslant \frac{1}{4}T^{-\beta}.$$

For the first term on the RHS of (S8.40), we use the computations of the proof Lemma 3 in the Supplementary Materials of [2] and obtain

$$\mathbb{P}_0\left[\left|N^k[0,T] - \int_0^T \lambda_t^k(f_0)\right| \geqslant T\delta_T/2\right] \leqslant \frac{1}{4}T^{-\beta},$$

for δ_0 large enough.

For the third term on the RHS of (S8.38), we denote $X_j = U_j^{(1)} - \tau_j$ for $1 \leqslant j \geqslant J_T - 1$. We recall that the X_j's are i.i.d. and follow an exponential law with rate $\|r_0\|_1$ under $\mathbb{P}_0$ and $\mathbb{E}_0\left[X_j\right] = \frac{1}{\|r_0\|_1}$. We thus have

$$\begin{aligned}
\mathbb{P}_0[\Omega_J \cap \Omega_U^c] &\leqslant \mathbb{P}_0\left[\Omega_J \cap \left\{\sum_{j=1}^{J_T-1} X_j \leqslant \frac{T}{\mathbb{E}_0[\Delta\tau_1]\|r_0\|_1}\left(1 - 2c_\beta\sqrt{\frac{\log T}{T}}\right)\right\}\right] \\
&\leqslant \mathbb{P}_0\left[\Omega_J \cap \left\{\sum_{j=1}^{J_T-1} X_j - \frac{J_T - 1}{\|r_0\|_1} \leqslant \frac{T}{\mathbb{E}_0[\Delta\tau_1]\|r_0\|_1}\left(1 - 2c_\beta\sqrt{\frac{\log T}{T}} - 1 + c_\beta\sqrt{\frac{\log T}{T}}\right)\right\}\right] \\
&= \mathbb{P}_0\left[\Omega_J \cap \left\{\sum_{j=1}^{J_T-1} X_j - \frac{J_T - 1}{\|r_0\|_1} \leqslant -\frac{c_\beta\sqrt{T\log T}}{\mathbb{E}_0[\Delta\tau_1]\|r_0\|_1}\right\}\right] \leqslant \sum_{J\in\mathcal{J}_\mathcal{T}} \mathbb{P}_0\left[\sum_{j=1}^{J-1} X_j - \frac{J-1}{\|r_0\|_1} \leqslant -\frac{c_\beta\sqrt{T\log T}}{\mathbb{E}_0[\Delta\tau_1]\|r_0\|_1}\right],
\end{aligned}$$

where in the first inequality we have used the fact that on Ω_J,

$$J_T - 1 \geqslant \frac{T}{\mathbb{E}_0[\Delta\tau_1]}\left(1 - c_\beta\sqrt{\frac{\log T}{T}}\right).$$

We apply the Bernstein's inequality using that for any $k \geqslant 1$, $\mathbb{E}_0\left[X_1^k\right] \leqslant k!(\|r_0\|_1)^{-k+2}\mathbb{E}_0\left[X_1^2\right]/2$. Therefore, since $\mathbb{E}_0\left[X_1^2\right] = \|r_0\|_1^{-2}$, we obtain

$$\begin{aligned}
\mathbb{P}_0\left[\sum_{j=1}^{J-1} X_j - \frac{J-1}{\|r_0\|_1} \leqslant -\frac{c_\beta\sqrt{T\log T}}{\mathbb{E}_0[\Delta\tau_1]\|r_0\|_1}\right] &\leqslant \exp -\left\{\frac{c_\beta^2\log T}{\mathbb{E}_0\left[\Delta\tau_1\right]^2\left(1 + \frac{c_\beta\sqrt{\log T}}{\mathbb{E}_0[\Delta\tau_1]\sqrt{T}}\right)}\right\} \\
&\leqslant \exp -\left\{\frac{c_\beta^2\log T}{2\mathbb{E}_0\left[\Delta\tau_1\right]}\right\} \leqslant \frac{1}{4}T^{-\beta},
\end{aligned}$$

for $c_\beta > 0$ large enough. Finally, reporting into (S8.38) we can conclude that for $C_\beta, c_\beta, \delta_0$ large enough,

$$\mathbb{P}_0\left[\tilde{\Omega}_T^c\right] \leqslant T^{-\beta}.$$

For the second part of the lemma, we can use the exact same arguments as in the proof of Lemma 2 in [2] to obtain the result.

For the case of shifted ReLU link function with unknown shift, we similarly have that

$$\mathbb{P}_0[\tilde{\Omega}_T'^c] \leqslant \mathbb{P}_0[\Omega_N^c] + 3\mathbb{P}_0[\Omega_J^c] + \mathbb{P}_0[\Omega_J \cap \Omega_U^c] + \mathbb{P}_0\left[\Omega_J \cap \Omega_A^c\right], \tag{S8.41}$$

and therefore it only remains to bound the last term on the RHS of the previous inequality. Using Assumption S7.36 (ii), let $0 < x_1 < x_2$ and $c_\star$ such that $[x_1, x_2] \subset B_0 = \{x \in [0, A], h_{lk}^0(x) \leqslant -c_*\}$, $n_1 = \min\{n \in \mathbb{N}; nc_* > \nu_k^0\}$, $\delta' = (x_2 - x_1)/3$. We denote $\mathcal{E}_0$ the set of indices satisfying

$$\mathcal{E}_0 = \{j \in [J_T];\ N[\tau_j, \tau_j + \delta'] = N^l[\tau_j, \tau_j + \delta'] = n_1,\ N[\tau_j + \delta', \tau_{j+1}) = 0\}.$$

Since $\forall t \in [\tau_j + x_1 + \delta', \tau_j + x_2]$, $\tilde{\lambda}_t^k(f) < 0$, then $|A^k(f_0)| \geqslant \frac{2(x_2 - x_1)}{3}|\mathcal{E}_0|$ and, with $p_0 = \mathbb{P}_0[j \in \mathcal{E}_0]$,

$$\mathbb{P}_0\left[|A^k(f_0)| < z_0 T\right] \leqslant \mathbb{P}_0\left[|\mathcal{E}_0| < \frac{3z_0}{2(x_2 - x_1)} T\right] \leqslant \mathbb{P}_0\left[|\mathcal{E}_0| < p_0 T/2\right],$$

if $z_0 < 2p_0(x_2 - x_1)/3$. Consequently, applying Hoeffding's inequality with $Y_j = \mathbb{1}_{j \in \mathcal{E}_0} \overset{i.i.d.}{\sim} \mathcal{B}(p_0)$ for $j \in [J_T]$ with $J_T \geqslant 2T/3\mathbb{E}_0[\Delta\tau_1]$, we obtain

$$\mathbb{P}_0\left[|\mathcal{E}_0| < \frac{p_0 T}{2}\right] \leqslant \mathbb{P}_0\left[\sum_{j=1}^{2T/3\mathbb{E}_0[\Delta\tau_1]} Y_j < \frac{p_0 T}{2}\right] \lesssim e^{-\frac{T p_0^2}{6\mathbb{E}_0[\Delta\tau_1]}} \leqslant \frac{1}{4} T^{-\beta}.$$

Consequently, $\mathbb{P}_0\left[\Omega_J \cap \Omega_A^c\right] = o(T^{-\beta})$, which terminates the proof of this lemma.

□

S8.2. Proof of Lemma A.4

Lemma S8.2 (Lemma A.4)**.** *For $f \in \mathcal{F}_T$ and $l \in [K]$, let*

$$Z_{1l} = \int_{\tau_1}^{\xi_1} |\lambda_t^l(f) - \lambda_t^l(f_0)| dt,$$

where ξ_1 is defined in (22) in Section 5.2. Under the assumptions of Theorem 3.2 and Case 1 of Proposition 3.5, for $M_T \to \infty$ such that $M_T > M\sqrt{\kappa_T}$ with $M > 0$ and for any $f \in \mathcal{F}_T$ such that $\|\nu - \nu_0\|_1 \leqslant \max(\|\nu_0\|_1, \tilde{C})$ with $\tilde{C} > 0$, there exists $l \in [K]$ such that on $\tilde{\Omega}_T$,

$$\mathbb{E}_f[Z_{1l}] \geqslant C(f_0)\|f - f_0\|_1,$$

with $C(f_0) > 0$ a constant that depends only on f_0 and $\phi = (\phi_k)_k$.

Similarly, under the assumptions of Case 2 of Proposition 3.5, for $f \in \mathcal{F}_T$ and $\theta \in \Theta$, let $r_0 = (r_k^0)_k$, $r_f = (r_k^f)_k$ with $r_k^0 = \phi_k(\nu_k^0) = \theta_k^0 + \nu_k^0$, $r_k^f = \phi_k(\nu_k) = \theta_k + \nu_k$, $\forall k$. If $\left\|r_f - r_0\right\|_1 \leqslant \max(\|r_0\|, \tilde{C}')$ with $\tilde{C}' > 0$, then there exists $l \in [K]$ such that on $\tilde{\Omega}_T$,

$$\mathbb{E}_f[Z_{1l}] \geqslant C'(f_0)(\|r_f - r_0\|_1 + \|h - h_0\|_1), \quad C'(f_0) > 0. \tag{S8.42}$$

Proof. In this proof, we will show that (S8.42) holds for all the models satisfying the assumptions of Theorem 3.2 and Proposition 3.5, with $r_k^0 = \phi_k(\nu_k^0)$ and $r_k^f = \phi_k(\nu_k)$ for all k. Then, excluding Case 2, we use the fact that for any k, ϕ_k^{-1} is fully known and L'-Lipshitz on $J_k = \phi_k(I_k)$ with I_k defined in Assumption 3.1 (which also holds for the ReLU link function by Assumption 2.2), to show that

$$\begin{aligned}\|r_f - r_0\|_1 + \|h - h_0\|_1 &\geqslant 1/L' \|\nu - \nu_0\|_1 + \|h - h_0\|_1 \\ &\geqslant \min(1, 1/L')(\|\nu - \nu_0\|_1 + \|h - h_0\|_1) = \min(1, 1/L')\,\|f - f_0\|_1\,.\end{aligned}$$

The proof of (S8.42) is inspired by the proof of Lemma 4 in the supplementary material of [2]. The following computations are valid in all our estimation scenarios. We recall that for any k, $r_k^f = \nu_k$ for the ReLU link (Case 1 of Proposition 3.5) and $r_k^f = \theta_k + \nu_k$ for the shifted ReLU link (Case 2 of Proposition 3.5).

Let $A > x > 0$ and $\eta > 0$ such that

$$0 < \frac{(A+x)^2 \eta K^2}{1 - \eta K} < \frac{1}{2} \qquad \text{and} \qquad \eta \leqslant \frac{\min_l r_l^0}{2C_0'}, \tag{S8.43}$$

with C_0' such that $\|r_f - r_0\|_1 + \|h - h_0\|_1 \leqslant C_0'$. Assume that for any $1 \leqslant l' \leqslant K$, $|r_{l'}^f - r_{l'}^0| \leqslant \eta(\|r_f - r_0\|_1 + \|h - h_0\|_1)$ and let $l \in [K]$ such that $\sum_k \|h_{kl} - h_{kl}^0\|_1 = \max_{l'} \sum_k \|h_{kl'} - h_{kl'}^0\|_1$.

Then we have

$$\|r_f - r_0\|_1 + \|h - h_0\|_1 \leqslant \left(\frac{\eta K^2}{1 - \eta K} + K\right) \sum_k \|h_{kl} - h_{kl}^0\|_1. \tag{S8.44}$$

For each $k \in [K]$, we define the event Ω_k as

$$\Omega_k = \left\{ \max_{k' \neq k} N^{k'}[\tau_1, \tau_2) = 0,\ N^k[\tau_1, \tau_1 + x] = 0,\ N^k[\tau_1 + x, \tau_1 + x + A] = 1,\ N^k[\tau_1 + x + A, \tau_2) = 0 \right\}.$$

On Ω_k, we have $\xi_1 = U_1^{(1)} + A$ and thus,

$$\mathbb{E}_f\,[Z_{1l}] \geqslant \sum_k \mathbb{E}_f \left[\mathbb{1}_{\Omega_k} \int_{\tau_1}^{A + U_1^{(1)}} |\lambda_t^l(f) - \lambda_t^l(f_0)| dt \right].$$

Let $\mathbb{Q}$ be the point process measure of a homogeneous Poisson process with unit intensity on $\mathbb{R}^+$ and equal to the null measure on $[-A, 0)$. Then

$$\mathbb{E}_f\,[Z_{1l}] \geqslant \sum_k \mathbb{E}_{\mathbb{Q}} \left[\int_{\tau_1}^{U_1^{(1)} + A} \mathcal{L}_t(f) \mathbb{1}_{\Omega_k} |\lambda_t^l(f) - \lambda_t^l(f_0)| \right] dt,$$

with $\mathcal{L}_t(f)$ the likelihood process given by

$$\mathcal{L}_t(f) = \exp\left(Kt - \sum_k \int_{\tau_1}^t \lambda_u^k(f) du + \sum_k \int_{\tau_1}^t \log(\lambda_u^k(f)) dN_u^k \right).$$

For $t \in [\tau_1, U_1^{(1)} + A)$, since on Ω_k, $\tau_1 + x \leqslant U_1^{(1)} \leqslant \tau_1 + A + x$, we have

$$\mathcal{L}_t(f) \geqslant e^{Kt} \lambda^k_{U_1^{(1)}}(f) \exp\left\{ - \sum_{k'} \int_{\tau_1}^t \phi_{k'}(\tilde{\lambda}^{k'}_u(f)) du \right\}.$$

Under condition **(C2)**, since $\phi_{k'} \leqslant \Lambda_{k'}, \forall k'$, with $\Lambda = (\Lambda_1, \dots, \Lambda_K)$, we directly have that

$$\mathcal{L}_t(f) \geqslant e^{Kt} \lambda^k_{U_1^{(1)}}(f) e^{-\|\Lambda\|_1} \geqslant r_k^f e^{-\|\Lambda\|_1},$$

since at $\lambda^k_{U_1^{(1)}} = r_k^f = \phi_k(\nu_k)$.

Under condition **(C1bis)**, using that ϕ_k is L-Lipschitz, we have

$$\begin{aligned}
\mathcal{L}_t(f) &\geqslant e^{-\sum_{k'} \phi_{k'}(0)(A + U_1^{(1)} - \tau_1)} \lambda^k_{U_1^{(1)}}(f) \exp\left\{ - \sum_{k'} \int_{\tau_1}^{A + U_1^{(1)}} (\phi_{k'}(\tilde{\lambda}^{k'}_u(f)) - \phi_{k'}(0)) du \right\} \\
&\geqslant e^{-\sum_{k'} \phi_{k'}(0)(A + U_1^{(1)} - \tau_1)} \lambda^k_{U_1^{(1)}}(f) \exp\left\{ -L \sum_{k'} \left((A + U_1^{(1)} - \tau_1)\nu_{k'} + \int_{U_1^{(1)}}^{A + U_1^{(1)}} h_{kk'}(u - U_1^{(1)}) du \right) \right\} \\
&\geqslant e^{-\sum_{k'} \phi_{k'}(0)(2A + x)} \lambda^k_{U_1^{(1)}}(f) \exp\left\{ -L \sum_{k'} \left((2A + x)\nu_{k'} + \int_{U_1^{(1)}}^{A + U_1^{(1)}} h^+_{kk'}(u - U_1^{(1)}) du \right) \right\} \\
&\geqslant e^{-\sum_{k'} \phi_{k'}(0)(2A + x)} r_k^f \exp\left\{ -L \sum_{k'} \left((2A + x)\nu_{k'} + \|h^+_{kk'}\|_1 \right) \right\}.
\end{aligned}$$

Moreover, since $\|S^+\|_1 < 1$, then $\forall (k, k') \in [K]^2$, $\|h^+_{kk'}\|_1 < 1$. Thus, we obtain

$$\begin{aligned}
\mathcal{L}_t(f) &\geqslant e^{-\sum_{k'} \phi_{k'}(0)(2A + x)} r_k^f e^{-LK - L(2A + x)\sum_{k'} \nu_{k'}} \\
&\geqslant \frac{e^{-\sum_{k'} \phi_{k'}(0)(2A + x)} r_k^0}{2} e^{-LK - 6AL \max(\tilde{C}, \|\nu_0\|_1)} =: C.
\end{aligned}$$

In the last inequality, we have used our assumption $\|\nu - \nu_0\|_1 \leqslant \max(\|\nu_0\|_1, \tilde{C})$ which implies that

$$\sum_{k'} \nu_{k'} \leqslant 2 \max(\|\nu_0\|_1, \tilde{C}).$$

Moreover, we have that

$$\begin{aligned}
\mathbb{E}_f[Z_{1l}] &\geqslant C \sum_k \mathbb{E}_{\mathbb{Q}} \left[\mathbb{1}_{\Omega_k} \int_{U_1^{(1)}}^{U_1^{(1)} + A} \left| \phi_l(\tilde{\lambda}^l_t(f)) - \phi_l(\tilde{\lambda}^l_t(f_0)) \right| dt \right] \\
&\geqslant \frac{C}{L'} \sum_k \mathbb{E}_{\mathbb{Q}} \left[\mathbb{1}_{\Omega_k} \int_{U_1^{(1)}}^{U_1^{(1)} + A} \left| (\nu_l - \nu_l^0) + (h_{kl} - h^0_{kl})(t - U_1^{(1)}) \right| dt \right],
\end{aligned}$$

in all models except Case 2. In fact, in the latter case, we obtain

$$\mathbb{E}_f[Z_{1l}] \geqslant C \sum_k \mathbb{E}_{\mathbb{Q}} \left[\mathbb{1}_{\Omega_k} \int_{U_1^{(1)}}^{U_1^{(1)} + A} \left| (\theta_l + \nu_l - \theta_l^0 - \nu_l^0) + (h_{kl} - h^0_{kl})(t - U_1^{(1)}) \right| dt \right]$$

$$= C \sum_k \mathbb{E}_{\mathbb{Q}} \left[\mathbb{1}_{\Omega_k} \int_{U_1^{(1)}}^{U_1^{(1)}+A} \left| (r_l^f - r_l^0) + (h_{kl} - h_{kl}^0)(t - U_1^{(1)}) \right| dt \right].$$

On the one hand,

$$\mathbb{E}_{\mathbb{Q}} \left[\mathbb{1}_{\Omega_k} \int_{U_1^{(1)}}^{U_1^{(1)}+A} |\nu_l - \nu_l^0| dt \right] = A|\nu_l - \nu_l^0|\mathbb{Q}(\Omega_k) \leqslant AL'|\phi_l(\nu_l) - \phi_l(\nu_l^0)|\mathbb{Q}(\Omega_k) = AL'|r_l^f - r_l^0|\mathbb{Q}(\Omega_k)$$

$$\leqslant AL' \frac{\eta K^2}{1 - \eta K} \sum_{k'} \|h_{k'l} - h_{k'l}^0\|_1,$$

and in Case 2 we have

$$\mathbb{E}_{\mathbb{Q}} \left[\mathbb{1}_{\Omega_k} \int_{U_1^{(1)}}^{U_1^{(1)}+A} |r_l^f - r_l^0| dt \right] = A|r_l - r_l^0|\mathbb{Q}(\Omega_k) \leqslant A \frac{\eta K^2}{1 - \eta K} \sum_{k'} \|h_{k'l} - h_{k'l}^0\|_1.$$

On the other hand, by definition of $\mathbb{Q}$, $N^k[\tau_1, \tau_1 + x + A] \sim \text{Poisson}(x + A)$. Consequently, with U a random variable with uniform distribution on $[\tau_1 + x, \tau_1 + x + A]$, we obtain

$$\mathbb{E}_{\mathbb{Q}} \left[\mathbb{1}_{\Omega_k} \int_{U_1^{(1)}}^{U_1^{(1)}+A} \left| (h_{kl} - h_{kl}^0)(t - U_1^{(1)}) \right| dt \right] = \mathbb{Q}(\Omega_k) \mathbb{E} \left[\int_U^{U+A} |(h_{kl} - h_{kl}^0)(t - U)| dt \right]$$

$$= \frac{\mathbb{Q}(\Omega_k)}{A} \int_{\tau_1 + x}^{\tau_1 + A + x} \left[\int_s^{A+s} |h_{kl} - h_{kl}^0|(t - s) dt \right] ds \geqslant \mathbb{Q}(\Omega_k) \|h_{kl} - h_{kl}^0\|_1.$$

Moreover, we have

$$\mathbb{Q}(\Omega_k) \geqslant \mathbb{Q}(\max_{k' \neq k} N^{k'}[\tau_1, \tau_1 + x + 2A] = 0, N^k[\tau_1, \tau_1 + x] = 0, N^k[\tau_1 + x, \tau_1 + x + A] = 1)$$

$$= \mathbb{Q}(\max_{k' \neq k} N^{k'}[\tau_1, \tau_1 + x + 2A] = 0) \mathbb{Q}(N^k[\tau_1, \tau_1 + x] = 0) \mathbb{Q}(N^k[\tau_1 + x, \tau_1 + x + A] = 1)$$

$$= e^{-(K-1)(x+2A)} \times e^{-x} \times A e^{-A} := C'.$$

Using (S8.43) together with (S8.44), we obtain

$$\mathbb{E}_f [Z_{1l}] \geqslant \frac{C}{L'} \sum_k \frac{\mathbb{Q}(\Omega_k)}{A} \left(\|h_{kl} - h_{kl}^0\|_1 - A^2 L' \frac{\eta K^2}{1 - \eta K} \|h_{kl} - h_{kl}^0\|_1 \right) \geqslant \frac{C}{L'} \frac{C'}{2} \sum_k \|h_{kl} - h_{kl}^0\|_1$$

$$\geqslant C(f_0)(\|r - r_0\|_1 + \|h - h_0\|_1), \quad C(f_0) = \frac{C}{L'} \frac{C'}{2(K + \eta K^2/(1 - \eta K))}.$$

If there exists $l \in [K]$ such that $|r_l^f - r_l^0| \geqslant \eta(\|r - f - r_0\|_1 + \|h - h_0\|_1)$, we can use similar arguments as in the proof of Lemma 4 of [2]:

$$\mathbb{E}_f [Z_{1l}] \geqslant \mathbb{P}_f \left[\max_k N^k[\tau_1, \tau_1 + A] = 0 \right] \times A|r_l^f - r_l^0|,$$

and

$$\mathbb{P}_f \left[\max_k N^k[\tau_1, \tau_1 + A] = 0 \right] = \mathbb{E}_{\mathbb{Q}} \left[\int_{\tau_1}^{\tau_1 + A} \mathcal{L}_t(f) \mathbb{1}_{\max_k N^k[\tau_1, \tau_1 + A] = 0} dt \right] = \mathbb{E}_{\mathbb{Q}} \left[\int_{\tau_1}^{\tau_1 + A} e^{A\|r\|_1} \mathbb{1}_{\max_k N^k[\tau_1, \tau_1 + A] = 0} dt \right]$$

$$\geqslant A e^{A\|r_f\|_1} e^{-KA},$$

so that

$$\mathbb{E}_f[Z_{1l}] \geqslant C'(f_0)(\|r_f - r_0\|_1 + \|h - h_0\|_1), \quad C'(f_0) = A^2 \eta e^{A\|r_0\|_1/2} e^{-KA}.$$

We can conclude that in all cases,

$$\mathbb{E}_f[Z_{1l}] \geqslant \min(C(f_0), C'(f_0))(\|r_f - r_0\|_1 + \|h - h_0\|_1),$$

and except in Case 2 of Proposition 3.5,

$$\mathbb{E}_f[Z_{1l}] \geqslant \min(C(f_0), C'(f_0), \frac{1}{L'}, 1)\|f - f_0\|_1 .$$

□

S9. Additional results

In this section we recall some useful results on the regenerative properties of the nonlinear Hawkes model, which are mainly straightforward extensions of [1] to our multivariate and general nonlinear setup. Besides, we recall the well-known Fuk-Nagaev's inequality.

The first lemma is an extension of Theorem A.1 [1] for a $M^K/G^K/\infty$ queue when the arrival process is the superposition of K Poisson Point processes, corresponding to K types of customers.

Lemma S9.1. *Consider a $M^K/G^K/\infty$ queue with K types of customers that arrive according to a Poisson process with rate $r = (r_1, \dots, r_K)$. Assume that for each $k \in [K]$, the generic service time H^k for a customer of type k satisfies for some $\gamma > 0$ and for any $t \geqslant 0$:*

$$\mathbb{P}\left[H^k \geqslant t\right] = o(e^{-\gamma t}).$$

Let $\mathcal{T}_1$ the first time of return of the queue to zero.

1. *If $\|r\|_1 < \gamma$, then*

$$\mathbb{P}[\mathcal{T}_1 \geqslant t] \leqslant \left[1 + \frac{\mathbb{E}\left[e^{\gamma B}\right]}{\gamma - \|r\|_1}\right] e^{-\|r\|_1 t},$$

   where B is the length of a busy period of the queue, i.e. $B = \mathcal{T}_1 - V_1$ with V_1 the arrival time of the first customer.
2. *If $\gamma \leqslant \|r\|_1$, then for any $0 < \alpha < \gamma$, $\mathbb{P}[\mathcal{T}_1 \geqslant t] \leqslant c_1(\alpha) e^{-\alpha t}$, with*

$$c_1(\alpha) = \left[1 + \frac{\mathbb{E}\left[e^{\alpha B}\right]}{\|r\|_1 - \alpha}\right].$$

3. $\forall \alpha \leqslant \|r\|_1 \wedge \gamma$, $\mathbb{E}\left[e^{\alpha \mathcal{T}_1}\right] \leqslant \frac{\|r\|_1}{\|r\|_1 + s} \mathbb{E}\left[e^{\alpha B}\right] < +\infty$.

Proof. In this situation, the arrival process of customers, *regardless of their type*, is a superposition of K Poisson processes with individual rate r_k, $k \in [K]$. Consequently, it is equivalent to a Poisson process with rate $\|r\|_1 = \sum_k r_k$. Moreover, the generic service time H of a customer can be written as

$H = \sum_k \delta_k H^k$, with $\delta = (\delta_k)_{k\in[K]}$ a one-hot vector indicating the type of customer. We can easily see that

$$\delta \sim \text{Mult}\left(1, \frac{r_1}{\|r\|_1}, \dots, \frac{r_K}{\|r\|_1}\right), \quad H|\delta \sim \delta\mathcal{P},$$

with $\mathcal{P}$ the vector of service time distributions of the K types of customers. We note that the service time H is independent of the arrival process. Consequently, for $t \geqslant 0$,

$$\mathbb{P}[H \geqslant t] = \sum_k \mathbb{P}\left[H^k \geqslant t,\ \delta_k = 1\right] \leqslant \sum_k \mathbb{P}\left[H^k \geqslant t\right] = o(e^{-\gamma t}).$$

We can therefore conclude that this queue is equivalent to a $M/G/\infty$ queue with rate $\|r\|_1$ and generic service time satisfying $\mathbb{P}[H \geqslant t] = o(e^{-\gamma t})$. We can then apply Theorem A.1 in [1] to obtain the results. □

The next lemma is a direct multivariate extension of the results in Propositions 2.1 and 3.1 and Lemma 3.2 of [1]. It introduces the mutually-exciting process dominating (in the sense of measure) a nonlinear Hawkes process.

Lemma S9.2. *Let Q be a K-dimensional Poisson point process on $(0,+\infty)\times(0,+\infty)^K$ with unit intensity. Let N be the Hawkes process with immigration rate $\nu = (\nu_1, \dots, \nu_K)$, $\nu_k > 0$, $k \in [K]$, interaction functions $h_{lk} : \mathbb{R}_+ \to \mathbb{R}$, $(l,k) \in [K]^2$ and initial measure N_0 on $[-A, 0]$ driven by $(Q_t)_{t\geqslant 0}$ and satisfying one condition of Lemma 2.1. N is the pathwise unique strong solution of the following system of stochastic equations*

$$\begin{cases} N^k = N_0^k + \int_{(0,+\infty)\times(0,+\infty)} \delta(u) \mathbb{1}_{\theta \leqslant \lambda^k(u)} Q^k(du, d\theta), \\ \lambda^k(u) = \phi_k\left(\nu_k + \sum_{l=1}^K \int_{u-A}^u h_{lk}(u-s) dN_s^l\right),\ u > 0, \quad k \in [K] \end{cases}.$$

with $\delta(.)$ the Dirac delta function. Consider the similar equation for a point process N^+ in which h_{lk} is replaced by h_{lk}^+ for any $l, k \in [K]^2$. Then

1. *there exists a pathwise unique strong solution N;*
2. *the same holds for N^+ and $N \leqslant N^+$ a.s. in the sense of measures.*

This also implies that, with $\Delta\tau_1^+$ defined similarly to $\Delta\tau_1$ in (20) for the process N^+,

$$\mathbb{P}\left[\Delta\tau_1 \leqslant \Delta\tau_1^+\right] = 1.$$

Moreover, with $\mathcal{T}_1$ defined as in Lemma S9.1, we also have $\mathbb{P}\left[\Delta\tau_1^+ = \mathcal{T}_1\right] = 1$.

Finally, the last proposition is the Fuk-Nagaev's inequality.

Proposition S9.3. *Let $(X_i)_{i\geqslant 1}$ a sequence of independent and centered random variables with finite variance and $S_n = \sum_{i=1}^n X_i$. With $v = \sum_{i=1}^n \mathbb{V}(X_i)$, for any $x \geqslant 0$ and $\lambda \geqslant 0$, it holds that*

$$\mathbb{P}[S_n \geqslant \lambda] \leqslant \sum_{i=1}^n \mathbb{P}[X_i > x] + \exp\left\{-\frac{v}{x^2} h\left(\frac{x\lambda}{v}\right)\right\},$$

where $h(u) = (1+u)\log(1+u) - u$, $u \geqslant 0$.

References

[1] Costa, M., Graham, C., Marsalle, L. and Tran, V. C. (2020). Renewal in Hawkes processes with self-excitation and inhibition. *Advances in Applied Probability* **52** 879–915.

[2] Donnet, S., Rivoirard, V. and Rousseau, J. (2020). Nonparametric Bayesian estimation for multivariate Hawkes processes. *The Annals of Statistics* **48** 2698 – 2727.

[3] Ghosal, S., Ghosh, J. K. and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *The Annals of Statistics* **28** 500 – 531.

[4] Ghosal, S. and van der Vaart, A. (2007). Convergence rates of posterior distributions for non iid observations. *The Annals of Statistics* **35** 192-223.

[5] Hansen, N. R., Reynaud-Bouret, P. and Rivoirard, V. (2015). Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli* **21** 83–143.

[6] Reynaud-Bouret, P. and Roy, E. (2007). Some non asymptotic tail estimates for Hawkes processes. *Bulletin of the Belgian Mathematical Society-Simon Stevin* **13** 883–896.

[7] Rousseau, J. (2010). Rates of convergence for the posterior distributions of mixtures of Betas and adaptive nonparametric estimation of the density. *Annals of Statistics* **38** 146–180.

[8] Sulem, D., Rivoirard, V. and Rousseau, J. (2022). Bayesian estimation of nonlinear Hawkes processes.

ca

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Bayesian estimation of nonlinear Hawkes processes
Publication Status	□Published □ Accepted for Publication X Submitted for Publication □Unpublished and unsubmitted work written in a manuscript style
Publication Details	Joint work with Professor Judith Rousseau (University of Oxford) and Professor Vincent Rivoirard (Universite Paris-Dauphine). Submitted to the Bernoulli journal.

Student Confirmation

Student Name:	Deborah Sulem		
Contribution to the Paper	I am the first author of this paper. I studied the asymptotic behaviour of the posterior distribution in the nonlinear Hawkes model, proving both concentration and consistency results. I have proposed and analysed a new type of estimators for the connectivity graph parameter		
Signature *Deborah Sulem*		Date	08/11/2022

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Professor Judith Rousseau		
Supervisor comments		
Signature 	Date	

This completed form should be included in the thesis, at the end of the relevant chapter.

3 | Scalable variational Bayes methods for Hawkes processes

Unsubmitted work.

Scalable Variational Bayes methods for Hawkes processes

Déborah Sulem
Department of Statistics
University of Oxford
deborah.sulem@stats.ox.ac.uk

Vincent Rivoirard
Ceremade, CMRS, UMR 7534
Université Paris-Dauphine, PSL University
vincent.rivoirard@dauphine.fr

Judith Rousseau
Department of Statistics
University of Oxford
judith.rousseau@stats.ox.ac.uk

February 11, 2023

Abstract

Multivariate Hawkes processes are temporal point processes extensively applied to model event data with dependence on past occurrences and interaction phenomena, e.g., neuronal spike trains, online messages, and financial transactions. In the nonparametric setting, learning the temporal dependence structure of Hawkes processes is often a computationally expensive task, all the more with Bayesian estimation methods. In the generalised nonlinear Hawkes model, the posterior distribution is non-conjugate and *doubly intractable*, and existing Monte-Carlo Markov Chain methods are often slow and not scalable to high-dimensional processes in practice. Recently, efficient algorithms targeting a mean-field variational approximation of the posterior distribution have been proposed. In this work, we unify existing variational Bayes inference approaches under a general framework, that we theoretically analyse under easily verifiable conditions on the prior, the variational class, and the model. Then, in the context of the popular sigmoid Hawkes model, we design adaptive and sparsity-inducing mean-field variational methods. In particular, we propose a two-step algorithm based on a thresholding heuristic to select the *connectivity graph* parameter of the Hawkes model. Through an extensive set of numerical simulations, we demonstrate that our approach enjoys several benefits: it is computationally efficient, can reduce the dimensionality of the problem by selecting the graph parameter, and is able to adapt to the smoothness of the underlying parameter.

1 Introduction

Modelling point or event data with temporal dependence often implies inferring a local dependence structure between events, or estimating interaction parameters. In this context, the multivariate Hawkes model is a widely used temporal point process (TPP) model, e.g., in seismology [Ogata, 1999], criminology [Mohler et al., 2011], finance [Bacry and Muzy, 2015], and social network analysis [Lemonnier and Vayatis, 2014]. In particular, the generalised nonlinear Hawkes model is able to account for different *types* of temporal interactions, including *excitation* and *inhibition* effects. For event data, the *excitation* phenomenon, sometimes named *contagion* or *bursting behaviour*, corresponds to empirical observation that the occurrence of an event, e.g., a post on a social media, increases the probability of observing similar events in the future, e.g., reaction comments. In contrast, the *inhibition* phenomenon refers to the opposite observation and is prominent in neuronal applications due to biological regulation mechanisms [Bonnet et al., 2021], and in criminology due to the enforcement of policies [Olinde and Short, 2020]. In addition to its expressive power, the multivariate Hawkes model has become popular for the interpretability of its parameter, in particular the *connectivity* or *dependence* graph parameter, which corresponds to a Granger-*causal* graph [Eichler et al., 2017].

In general, a multivariate TPP can be described as a counting process $N = (N_t)_t = (N_t^1, \dots, N_t^K)_{t\in\mathbb{R}}$, where $K \geqslant 1$ is the number of components (or dimensions) of the process. Each component of a TPP can represent a specific type of

event (e.g., a flooding or earthquake, when modelling natural disaster events), or a particular location where events are recorded (e.g., a region or country). For each $k = 1, \dots, K$ and time t, N_t^k denotes the number of events that have occurred until t at component k. We note that a multivariate TPP is also equivalent to a *marked TPP* where the marks belong to the set $\{1, 2, \dots, K\}$ [Daley and Vere-Jones, 2007]. Therefore, multivariate TPPs are of interest for jointly modelling the occurrences of events of distinct types, or recorded at multiple places. In TPPs, the probability distribution of events is characterised by a conditional intensity function (or, more concisely, the intensity), denoted $(\lambda_t)_t = (\lambda_t^1, \dots, \lambda_t^K)_{t \in \mathbb{R}}$. This function is informally the infinitesimal probability rate of event, conditionally on the history of the process, i.e,

$$\lambda_t^k dt = \mathbb{P}\left[N_t^k \text{ has a jump in } [t, t+dt] \middle| \mathcal{G}_t\right], \quad k = 1, \dots, K, \quad t \in \mathbb{R},$$

where $\mathcal{G}_t = \sigma(N_s, s < t)$ denotes the history of the process until time t. In the nonlinear Hawkes model, the intensity is defined as

$$\lambda_t^k = \phi_k\left(\nu_k + \sum_{l=1}^{K} \int_{-\infty}^{t^-} h_{lk}(t-s) dN_s^l\right), \quad k = 1, \dots K, \tag{1}$$

where for each k, $\phi_k : \mathbb{R} \to \mathbb{R}^+$ is a *link* or *activation* function, $\nu_k > 0$ is a *background* or *spontaneous* rate of events, and for each l, $h_{lk} : \mathbb{R}^+ \to \mathbb{R}$ is the *interaction function* or *triggering kernel* from N^l onto N^k. On the one hand, the parameter $\nu = (\nu_k)_k$ characterises the external influence of the environment on the process. Here, we assume that this parameter is constant over time. On the other hand, the functions $h = (h_{lk})_{l,k}$ parametrise the *causal* influence of past events. In particular, for any l, k, there exists a *causal* relationship from N^l to N^k, or in other words, N^k is *locally-dependent* on N^l, if and only if $h_{lk} \neq 0$ [Eichler et al., 2017]. Moreover, defining for each l, k, $\delta_{lk} := \mathbb{1}_{h_{lk} \neq 0}$, the parameter $\delta := (\delta_{lk})_{l,k} \in \{0,1\}^{K \times K}$ defines a Granger-causal graph, called the *connectivity* graph. Finally, the link functions $\phi = (\phi_k)_k$'s are in general nonlinear and monotone non-decreasing. They are an essential part of the model chosen by the practitioner, and frequently set as ReLU functions $\phi_k(x) = \max(x, 0) = (x)_+$ [Hansen et al., 2015, Chen et al., 2017, Costa et al., 2020, Lu and Abergel, 2018, Bonnet et al., 2021, Deutsch and Ross, 2022], sigmoid-type functions, e.g., $\phi_k(x) = \theta_k(1 + e^x)^{-1}$ with a scale parameter $\theta_k > 0$ [Zhou et al., 2021b,a, Malem-Shinitski et al., 2021], softplus functions $\phi_k(x) = \log(1+e^x)$ [Mei and Eisner, 2017], or clipped exponential functions, i.e., $\phi_k(x) = \min(e^x, \Lambda_k)$ with a clip parameter $\Lambda_k > 0$ [Gerhard et al., 2017, Carstensen et al., 2010]. When all the interaction functions are non-negative and $\phi_k(x) = x$ for every k, the intensity (1) corresponds to the linear Hawkes model. Defining the *underlying* or *linear* intensity as

$$\tilde{\lambda}_t^k = \nu_k + \sum_{l=1}^{K} \int_{-\infty}^{t^-} h_{lk}(t-s) dN_s^l, \quad k = 1, \dots K, \tag{2}$$

for any $t \in \mathbb{R}$, the nonlinear intensity (1) can be re-written as $\lambda_t^k = \phi_k(\tilde{\lambda}_t^k)$.

Estimating the parameter of the Hawkes model, denoted $f = (\nu, h)$, and the graph parameter δ, has been theoretically studied in the Bayesian nonparametric framework and the linear model by Donnet et al. [2020] and in general nonlinear models in Sulem et al. [2021]. Moreover, the properties of nonparametric penalised projection estimators have been analysed in the linear model by Hansen et al. [2015] and Bacry et al. [2020] for high-dimensional linear processes, and by Cai et al. [2021] in nonlinear models. Yet, in practice, most methods rely on a parametric framework. In particular, a popular approach in the ReLU Hawkes model consists in estimating a parametric exponential form of the interaction functions, i.e., $h_{lk}(x) = \alpha_{lk} e^{-\beta_{lk} x}$. Then, the estimation of $(\alpha, \beta) = (\alpha_{lk}, \beta_{lk})_{l,k}$ can be performed via the maximum likelihood estimate (MLE) Bonnet et al. [2021], Wang et al. [2016], or a Monte-Carlo Markov Chain (MCMC) method Deutsch and Ross [2022]. Besides, a nonparametric approximated MLE is proposed in Lemonnier and Vayatis [2014] for the ReLU model. In sigmoid models, a data augmentation strategy derived from Donner and Opper [2019] and Adams et al. [2009] for Poisson point processes, allows to design Gibbs sampling algorithms. Such MCMC sampler has been notably proposed in a time-varying Hawkes model and semi-parametric estimation framework in Zhou et al. [2021a], and in a Gaussian process framework in Malem-Shinitski et al. [2021]. However, these algorithms rely on a computationally expensive sampling strategy and are not efficient enough in practice for multivariate processes.

Recently, data augmentation strategies have also been used to derive variational Bayes algorithms in Hawkes models. These novel methods leverage the conjugacy of an augmented mean-field variational posterior distribution with certain families of Gaussian priors. In the linear univariate model (i.e., $K = 1$), the *self-exciting* function $h := h_{11}$ is estimated via a transformation of a Gaussian process, e.g., a quadratic function in Zhang et al. [2020], or a sigmoid function in Zhou et al. [2019, 2020], and an iterative mean-field variational inference (MF-VI) algorithm. In the sigmoid Hawkes model, Zhou et al. [2021b] propose an efficient variational EM algorithm, then Zhou et al. [2021a] develop a related iterative MF-VI algorithm in a time-varying and semi-parametric multivariate model. A similar type of algorithm

is introduced by Malem-Shinitski et al. [2021] in a Gaussian process framework. Nonetheless, these variational approaches have not been yet theoretically analysed. Moreover, the estimation of the graph δ has not been considered in the variational framework, although this parameter can be of interest for scaling up these methods to high-dimensional Hawkes processes. In fact, the connectivity graph also determines the dimensionality and the sparsity of the estimation problem, similarly to the structure parameter in high-dimensional regression [Ray and Szabó , 2021].

In this work, we first provide a general variational Bayes estimation framework for multivariate Hawkes processes that unifies existing approaches, and theoretically analyse variational methods in this context. We notably derive concentration rates for variational posterior distributions, leveraging the general methodology of Zhang and Gao [2017] based on verifying a prior mass, a testing, and a variational class condition. We apply our general results to two variational classes of interest in the Hawkes model, namely the mean-field family and a novel spike-and-slab family, and two families of nonparametric priors. These resuts provide asymptotic guarantees, i.e., in the infinite-data setting, for variational Bayes methods. Then, we propose a novel adaptive and sparsity-inducing variational approach, based on the general methods of Zhang and Gao [2017] and Ohn and Lin [2021] for variational inference with model selection.

Next, building on existing data augmentation strategy, we design two adaptive and sparsity-inducing MF-VI algorithms in the sigmoid Hawkes model. In particular, we propose a two-step procedure based on a thresholding heuristic to select the connectivity graph parameter, which allows to reduce the computational cost for high-dimensional processes. We empirically demonstrate the effectiveness of our algorithms in an extensive set of simulations. We notably show that our adaptive variational algorithms are more computationally efficient than MCMC methods, while enjoying comparable estimation performance. Finally, our algorithms can also correctly infer the connectivity graph parameter, therefore uncovering the causality structure of the true generating process.

Additionally, we note that in the context of sigmoid Hawkes models with link function $\phi_k(x) = \theta_k(1 + e^{-x})^{-1}$ with $\theta_k > 0$, $k \in [K]$, existing algorithms also aim at estimating the scale parameter $\theta = (\theta_k)_k$ [Apostolopoulou et al., 2019, Zhou et al., 2021b, Malem-Shinitski et al., 2021, Zhou et al., 2021a]. However, the latter estimation problem has not been thoroughly analysed yet, neither in the Bayesian nor the frequentist frameworks. Therefore, we also extend the posterior concentration results of Sulem et al. [2021] to the latter model with unknown scale parameter θ, and validate the use of Bayesian methods in this setup.

Outline We first introduce some useful notation. Then, in Section 2, we describe our model and inference setup. Section 3 contains our general results, and their applications to prior and variational families of interest in the Hawkes model. In Section 4, we focus on the sigmoid Hawkes model and present our novel adaptive and sparsity-inducing variational algorithms. Finally, Section 5 contains the results of our numerical experiments. The proofs of our main results are reported in Appendix A.

Notations. For a function h, we denote $\|h\|_1 = \int_{\mathbb{R}} |h(x)|dx$ the L_1-norm, $\|h\|_2 = \sqrt{\int_{\mathbb{R}} h^2(x)dx}$ the L_2-norm, $\|h\|_\infty = \sup_{x\in\mathbb{R}} |h(x)|$ the supremum norm, and $h^+ = max(h, 0)$, $h^- = max(-h, 0)$ its positive and negative parts. For a $K \times K$ matrix A, we denote $r(A)$ its spectral radius, $\|A\|$ its spectral norm, and $tr(A)$ its trace. For a vector $u \in \mathbb{R}^K$, $\|u\|_1 = \sum_{k=1}^K |u_k|$. The notation $k \in [K]$ is used for $k \in \{1, \cdots, K\}$. For a set B and $k \in [K]$, we denote $N^k(B)$ the number of events of N^k in B and $N^k|_B$ the point process measure restricted to the set B. For random processes, the notation $\overset{\mathcal{L}}{=}$ corresponds to equality in distribution. We also denote $\mathcal{N}(u, \mathcal{H}_0, d)$ the covering number of a set $\mathcal{H}_0$ by balls of radius u w.r.t. a metric d. For any $k \in [K]$, let $\mu_k^0 = \mathbb{E}_0[\lambda_t^k(f_0)]$ be the mean of $\lambda_t^k(f_0)$ under the stationary distribution $\mathbb{P}_0$. For a set Ω, its complement is denoted Ω^c. We also use the notations $u_T \lesssim v_T$ if $|u_T/v_T|$ is bounded when $T \to \infty$, $u_T \gtrsim v_T$ if $|v_T/u_T|$ is bounded and $u_T \asymp v_T$ if $|u_T/v_T|$ and $|v_T/u_T|$ are bounded. We recall that a function ϕ is L-Lipschitz, if for any $(x, x') \in \mathbb{R}^2$, $|\phi(x) - \phi(x')| \leqslant L|x - x'|$. We denote $\mathbb{1}_n$ and $\mathbf{0}_n$ the all-ones and all-zeros vectors of size n. Finally, we denote $\mathcal{H}(\beta, L_0)$ the class of β-smooth functions with radius L_0.

2 Model and inference setup

In this section, we first recall the formal definition of multivariate Hawkes processes and our main assumptions. Then, we describe our general variational Bayes inference framework.

2.1 Multivariate Hawkes processes

Let $(\mathcal{X}, \mathcal{G}, \mathbb{P})$ be a probability space, $N = (N_t)_{t\in\mathbb{R}} = (N_t^1, \ldots, N_t^K)_{t\in\mathbb{R}}$ be a K-dimensional temporal point process, and $\{\mathcal{G}_t\}_{t\in\mathbb{R}}$ be the filtration such that $\mathcal{G}_t = \sigma(N_s, s \leqslant t) \subset \mathcal{G}$.

Definition 2.1 (Multivariate nonlinear Hawkes process)**.** *A TPP $(N_t)_t$ is a Hawkes process adapted to $\mathcal{G}$ if*

i) almost surely, $\forall k, l \in [K]$, $(N_t^k)_t$ and $(N_t^l)_t$ never jump simultaneously;

ii) for all $k \in [K]$, the $\mathcal{G}_t$-predictable conditional intensity function of N^k at $t \in \mathbb{R}$ is given by (1).

An alternative definition of Hawkes processes can be formulated via a system of stochastic equations driven by a marked Poisson point process (see for instance Bremaud and Massoulie [1996]).

Definition 2.2. *Let $Q = (Q^1, \dots, Q^k)$ be a K-dimensional point process such that for each k, Q^k is a Poisson point process on $(0, +\infty) \times (0, +\infty)^K$ with unit intensity. Let N_0 a point process measure on $\mathbb{R}_-$. If N is the pathwise unique strong solution of the following system of equations*

$$\begin{cases} N^k = N_0^k + \int_{(0,+\infty)\times(0,+\infty)} \delta(u) \mathbb{1}_{\theta \leqslant \lambda^k(u)} Q^k(du, d\theta), \\ \lambda^k(u) = \phi_k \left(\nu_k + \sum_{l=1}^K \int_{u-A}^u h_{lk}(u-s) dN_s^l \right), \ u > 0, \quad k \in [K] \end{cases}.$$

with $\delta(.)$ the Dirac delta function, then N is a Hawkes process with parameter $\nu = (\nu_k), h = (h_{lk})_{lk}$, link functions $(\phi_k)_k$ and initial measure N_0 on $\mathbb{R}_-$ driven by $(Q_t)_{t \geqslant 0}$

We consider finite-memory and stationary Hawkes processes. More precisely, we assume that the interaction functions $(h_{lk})_{l,k}$ have a bounded support included in $[0, A]$ with $A > 0$ a known constant. We also assume that the activation functions $(\phi_k)_k$ are monotone non-decreasing, L-Lipschitz, $L > 0$, and that one of the two following conditions is satisfied (see for instance Bremaud and Massoulie [1996], Deutsch and Ross [2022], or Sulem et al. [2021]):

(C1) The matrix $S^+ = (S^+_{lk})_{l,k} \in \mathbb{R}_+^{K \times K}$ with $S^+_{lk} = L \left\| h^+_{lk} \right\|_1$, $\forall l, k$, satisfies $\|S^+\| < 1$;

(C2) For any $k \in [K]$, the link function ϕ_k is bounded, i.e., $\exists \Lambda_k > 0, \forall x \in \mathbb{R}, 0 \leqslant \phi_k(x) \leqslant \Lambda_k$.

2.2 Bayesian inference framework

We assume that we observe a stationary K-dimensional Hawkes process N with unknown parameter $f_0 = (\nu_0, h_0)$ and known link functions $(\phi_k)_k$ such that (ϕ, h_0) verifies condition **(C1)**, i.e., $\left\| S_0^+ \right\| < 1$ with $S_0^+ = (L \left\| h_{lk}^{0+} \right\|_1)_{l,k}$. Given an observation of N over a time window $[-A, T]$, with $T > 0$, the log-likelihood function for a parameter $f = (\nu, h)$ is given by

$$L_T(f) := \sum_{k=1}^K L_T^k(f), \quad L_T^k(f) = \left[\int_0^T \log(\lambda_t^k(f)) dN_t^k - \int_0^T \lambda_t^k(f) dt \right]. \tag{3}$$

We then denote $\mathbb{P}_f(.|\mathcal{G}_0)$ the conditional distribution of N defined as

$$d\mathbb{P}_f(.|\mathcal{G}_0) = e^{L_T(f) - L_T(f_0)} \mathbb{P}_0(.|\mathcal{G}_0).$$

We also denote $\mathbb{E}_0$ and $\mathbb{E}_f$ the expectations associated to $\mathbb{P}_0(.|\mathcal{G}_0)$ and $\mathbb{P}_f(.|\mathcal{G}_0)$. With a slight abuse of notation, we will drop the notation $\mathcal{G}_0$. Let $\mathcal{F}$ be the parameter space and Π be a prior distribution on $\mathcal{F}$. The posterior distribution for any subset $B \subset \mathcal{F}$ is defined as

$$\Pi(B|N) = \frac{\int_B \exp(L_T(f)) d\Pi(f)}{\int_{\mathcal{F}} \exp(L_T(f)) d\Pi(f)} =: \frac{N_T(B)}{D_T}, \quad D_T := \int_{\mathcal{F}} \exp(L_T(f)) d\Pi(f). \tag{4}$$

The posterior distribution (4) is often said to be *doubly intractable*, because of the integrals in the log-likelihood function (3) and in the denominator D_T. In general, it is expensive to compute since the parameter f includes K^2 functions.

Remark 2.3. *Let, for each component $k \in [K]$, $f_k = (\nu_k, (h_{lk})_{l=1,\dots,K}) \in \mathcal{F}_k$ so that $f = (f_k)_k$ and $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_K$. If the prior distribution verifies $\Pi(f) = \prod_k \Pi_k(f_k)$, then, given the expressions of the log-likelihood function (3) and the intensity (1), we have that $L_T^k(f) = L_T^k(f_k)$ and the posterior distribution can be written as*

$$\Pi(B|N) = \prod_k \Pi_k(B_k|N), \quad \Pi_k(B_k|N) = \frac{\int_{B_k} \exp(L_T^k(f_k)) d\Pi_k(f_k)}{\int_{\mathcal{F}_k} \exp(L_T^k(f_k)) d\Pi_k(f_k)}, \quad B_k \subset \mathcal{F}_k, \quad \forall k \in [K].$$

The latter factorisation implies that each factor $\Pi_k(.|N)$ of the posterior distribution can be independently computed - nonetheless, given the whole data N.

For the prior distribution Π, we use a construction similar to Donnet et al. [2020] and Sulem et al. [2021]. For ease of exposition, we consider link functions $(\phi_k)_k$ that are injective on $\mathbb{R}$ (see Assumption 3.1 in Section 3.1), however, our construction can be easily adapted if for each k, ϕ_k is injective on a subset of $\mathbb{R}$. We define a prior on f of the form

$$d\Pi(f) = d\Pi_h(h) \prod_k d\Pi_\nu(\nu_k).$$

We use a distribution Π_ν absolutely continuous with respect to the Lebesgue measure, with positive and continuous probability density on $\mathbb{R}^+\backslash\{0\}$, e.g., a gamma distribution. For $h = (h_{lk})_{l,k}$, we use the hierarchical spike-and-slab prior of Donnet et al. [2020] based on the connectivity graph parameter δ. For each l, k, we consider the following reparametrisation

$$h_{lk} = \delta_{lk}\bar{h}_{lk}, \quad \delta_{l,k} \in \{0,1\}, \quad \bar{h}_{lk} \in \mathcal{H}',$$

so that $\delta = (\delta_{lk})_{lk}$ is the connectivity graph. We then consider $\delta \sim \Pi_\delta$, where Π_δ is a prior distribution on $\{0,1\}^{K^2}$. Next, conditionally on δ, we use a truncated distribution on $h|\delta$ of the form

$$d\Pi_h(h|\delta) = \left(\prod_{l,k} d\tilde{\Pi}_{h|\delta}(h_{lk})\right)\mathbb{1}_{\|S^+\|<1}(h),$$

or simply $d\Pi_h(h) = \prod_{l,k} d\tilde{\Pi}_h(h_{lk})$ if $(\phi_k)_k$ satisfies **(C2)**, with

$$\tilde{\Pi}_{h|\delta}(h_{lk}) = \delta_{lk}\tilde{\Pi}_h(\bar{h}_{lk}) + (1-\delta_{lk})\delta_{\{0\}}(\bar{h}_{lk}),$$

where $\delta_{\{0\}}$ is the Dirac measure at $h_{lk} = 0$ and $\tilde{\Pi}_h$ is a nonparametric prior distribution on $\mathcal{H}'$, e.g., a Gaussian process, random histogram, or spline prior (see Sulem et al. [2021]).

Remark 2.4. *From the previous construction, one can see that the graph parameter $\delta \in \{0,1\}^{K^2}$ defines the sparsity structure of h. Besides, one can also use a soft-sparse (or shrinkage) prior distribution for $\tilde{\Pi}_h$.*

2.3 Variational Bayes framework

Previous work on Hawkes processes underlines the difficulty of computing the nonparametric posterior distribution (4) [Donnet et al., 2020, Zhou et al., 2021a, Malem-Shinitski et al., 2021], and scaling up Bayesian methods to high-dimensional processes. Alternatively, variational Bayes methods consist in approximating the posterior distribution within a *variational* class of "convenient" distributions. We first recall the definition of the Kullback-Leibler divergence. For any two distributions Q and Q' on $\mathcal{F}$, it is defined as

$$KL(Q\|Q') := \begin{cases} \int \log \frac{dQ}{dQ'} dQ, & \text{if } Q \ll Q' \\ +\infty, & \text{otherwise} \end{cases}.$$

Let $\mathcal{V}$ be an approximating family of distributions on $\mathcal{F}$. The *variational posterior* distribution, denoted $\hat{Q}$, is defined as the best approximation of the posterior distribution within $\mathcal{V}$, with respect to the Kullback-Leibler divergence, i.e.,

$$\hat{Q} := \arg\min_{Q\in\mathcal{V}} KL\left(Q\|\Pi(.|N)\right). \tag{5}$$

From Remark 2.3, we note that the variational distribution also factorises in K factors, $\hat{Q} = \prod_k \hat{Q}_k$ where each factor $\hat{Q}_k$ approximates $\Pi_k(.|N)$. Therefore, one can choose a variational class $\mathcal{V}'$ of distributions on $\mathcal{F}_1$ and define $\mathcal{V} = \mathcal{V}'^{\otimes K}$.

Mean-field variational inference When the parameter of interest, say ϑ, is multi-dimensional, i.e., $\vartheta = (\vartheta_1, \ldots, \vartheta_D)$ with $D > 1$, a common choice of variational class is a mean-field family [Zhang and Gao, 2017, Ohn and Lin, 2021], that can be defined as

$$\mathcal{V}_{MF} = \left\{Q;\ dQ(\vartheta) = \prod_{d=1}^{D} dQ_d(\vartheta_d)\right\}.$$

Then, the mean-field variational posterior distribution corresponds to $\hat{Q} = \arg\min_{Q\in\mathcal{V}_{MF}} KL\left(Q\|\Pi(.|N)\right) = \prod_{d=1}^{D}\hat{Q}_d$. Note that the mean-field family removes correlation between coordinates of the parameter. From now on, we assume that the mean-field variational posterior distribution has a density with respect to a dominating measure $\mu = \prod_d \mu_d$, and with a slight abuse of notation, we denote $\hat{Q}$ both the distribution and density with respect to μ. An interesting result from Bishop and Nasrabadi [2006], Donnet et al. [2020] is that the mean-field variational posterior distribution verifies, for each $d \in [D]$,

$$\hat{Q}_d(\vartheta_d) \propto \exp\left\{\mathbb{E}_{\hat{Q}_{-d}}[\log p(\vartheta, N)]\right\}, \tag{6}$$

where $p(\vartheta, N)$ is the joint density of the observations and the parameter with respect to $\prod_d \mu_d \times \mu_N$ with μ_N the data density, and $\hat{Q}_{-d} := \prod_{d'\neq d}\hat{Q}_{d'}$. This property (6) can be used to design efficient algorithms for computing the variational posterior, such as the coordinate-ascent variational inference algorithm. In the sigmoid Hawkes model, for which $\phi_k(x) = \theta_k(1+e^x)^{-1}, \forall k$, a mean-field approximating class is used within a latent variable augmentation scheme,

by breaking only the correlation between the original parameter and the latent variable [Malem-Shinitski et al., 2021, Zhou et al., 2021a]. We therefore consider a general setting where the log-likelihood function of the nonlinear Hawkes model can be augmented with some latent variable $z \in \mathcal{Z}$, with $\mathcal{Z}$ the latent parameter space. We denote $L_T^A(f,z)$ the augmented log-likelihood and define the *augmented* posterior distribution as

$$\Pi_A(B|N) = \frac{\int_B \exp(L_T^A(f,z))d(\Pi(f)\times\mathbb{P}_A(z))}{\int_{\mathcal{F}\times\mathcal{Z}} \exp(L_T^A(f,z))d(\Pi(f)\times\mathbb{P}_A)(z)}, \quad B \subset \mathcal{F}\times\mathcal{Z},$$

where $\mathbb{P}_A$ is a prior distribution on z which has a density with respect to a dominating measure μ_z. The approximating mean-field family of $\Pi_A(.|N)$ is then defined as

$$\mathcal{V}_{AMF} = \{Q : \mathcal{F}\times\mathcal{Z} \to [0,1];\ Q(f,z) = Q_1(f)Q_2(z)\}.$$

Thus, using property (6), the (augmented) mean-field variational posterior defined as

$$\hat{Q}_{AMF}(f,z) := \arg\min_{Q\in\mathcal{V}_{AMF}} KL\left(Q(f,z)\|\Pi_A(f,z|N)\right) =: \hat{Q}_1(f)\hat{Q}_2(z), \tag{7}$$

also verifies

$$\hat{Q}_1(f) \propto \exp\left\{\mathbb{E}_{\hat{Q}_2}[\log p(f,z,N)]\right\},$$
$$\hat{Q}_2(z) \propto \exp\left\{\mathbb{E}_{\hat{Q}_1}[\log p(f,z,N)]\right\},$$

where $p(f,z,N)$ is the joint density of the parameter, the latent variable, and the observations with respect to the measure $\prod_d \mu_d \times \mu_z \times \mu_N$.

Spike-and-slab variational inference Another variational class of interest in the context of sparse and high-dimensional models is the spike-and-slab variational family. In the multivariate Hawkes model, we introduce the following spike-and-slab variational family, inspired by the spike-and-slab prior from Section 2.2.

Definition 2.5 (Spike-and-slab variational class)**.** *In the Hawkes model with parameter $f = (\nu, h)$ and connectivity graph δ, the spike-and-slab variational family can be defined as*

$$\mathcal{V}_{SAS} = \left\{Q;\ dQ(f) = dQ_\delta(\delta)dQ_{f|\delta}(f) = dQ_\delta(\delta)dQ_{f|\delta}(\nu,\delta h)\right\}.$$

We note that if Q_δ is deterministic, i.e., Q_δ is the Dirac measure at some $\delta' \in \{0,1\}^{K\times K}$, then $\mathcal{V}_{SAS}$ corresponds to a variational family where the graph parameter is fixed, i.e., the variational posterior has a certain sparsity structure. Moreover, if Q_δ is given a factorised form, i.e., $Q_\delta(\delta) = \prod_{l,k} \bar{Q}_\delta(\delta_{lk})$, then $\mathcal{V}_{SAS}$ corresponds to a mean-field variational family. While standard MCMC methods using spike-and-slab priors are generally untractable, it is sometimes possible to design spike-and-slab variational inference algorithms that enjoy good computational properties (see for instance Titsias and Lázaro-Gredilla [2011], Ray and Szabó [2021] in sparse linear regression).

Variational inference with model selection More generally, variational inference algorithms aim at optimising a lower bound of the marginal log-likelihood, called the evidence lower bound (ELBO), and defined as

$$ELBO(Q) := \mathbb{E}_Q\left[\log\frac{p(f,N)}{Q(f)}\right], \quad Q \in \mathcal{V}, \tag{8}$$

where $p(f,N)$ is the joint distribution of the parameter and the data. The ELBO can also be used within a model-selection variational methodology, and we recall here the two related approaches of Zhang and Gao [2017] and Ohn and Lin [2021]. Let $\mathcal{M}$ be a set of models and for each $m \in \mathcal{M}$, let Π_m be a prior distribution on $\mathcal{M}$ and $\mathcal{V}^m$ be a variational class. The variational posterior in model m is defined $\hat{Q}^m = \arg\min_{Q\in\mathcal{V}^m} KL(Q|||\Pi(.|N))$. Then, a *model-selection* variational posterior, which lies in a selected model, is defined by Zhang and Gao [2017] as

$$\hat{Q} := \hat{Q}_{\hat{m}}, \quad \hat{m} := \arg\max_{m\in\mathcal{M}} ELBO(\hat{Q}^m). \tag{9}$$

We note that the approximating variational family in this case is $\mathcal{V} = \bigcup_{m\in\mathcal{M}} \mathcal{V}^m$. Another possibility is to construct an *adaptive* variational posterior as a mixture of distributions over the different models [Ohn and Lin, 2021], i.e.,

$$\hat{Q}(f) = \sum_{m\in\mathcal{M}} \hat{\gamma}_m \hat{Q}_m, \tag{10}$$

where $\{\hat{\gamma}_m\}_{m\in\mathcal{M}}$ are marginal probabilities defined as

$$\hat{\gamma}_m = \frac{\Pi_m(m)\exp\left\{ELBO(\hat{Q}_m)\right\}}{\sum_{m\in\mathcal{M}}\Pi_m(m)\exp\left\{ELBO(\hat{Q}_m)\right\}}, \quad \forall m \in \mathcal{M}. \tag{11}$$

In this case, the variational family is

$$\mathcal{V} = \left\{\sum_{m\in\mathcal{M}} \alpha_m Q_m; \sum_m \alpha_m = 1,\ \alpha_m \geqslant 0,\ Q_m \in \mathcal{V}^m,\ \forall m\right\}.$$

In Section 3.2.2, we will use this approach to induce sparsity and achieve adaptivity in our variational method. In our context of multivariate Hawkes processes, a "model" m will correspond to the sparsity structure and dimensionality of h, i.e., a graph parameter δ and the truncation level in a basis decomposition for each non-null function h_{lk}.

3 Main results

In this section, we provide theoretical guarantees for using variational Bayes methods to estimate the parameter of nonlinear Hawkes processes. We first derive the concentration rate of the variational posterior distribution (5), under general conditions on the model, the prior distribution, and the variational family. Then, we apply our results to variational methods of practical interest.

3.1 Variational posterior concentration rates

We recall that in our setting, the link functions $\phi = (\phi_k)_k$ in the nonlinear intensity (1) are fixed by the statistician and therefore known *a priori*. To analyse the variational posterior distribution, we first state a general assumption that guarantees the concentration of the posterior distribution (4) in the nonlinear Hawkes model (see Sulem et al. [2021]).

Assumption 3.1. *For a parameter f, we assume that there exists $\varepsilon > 0$ such that for each $k \in [K]$, the link function ϕ_k restricted to $I_k = (\nu_k - \max_{l\in[K]}\left\|h^-_{lk}\right\|_\infty - \varepsilon, \nu_k + \max_{l\in[K]}\left\|h^+_{lk}\right\|_\infty + \varepsilon)$ is bijective from I_k to $J_k = \phi_k(I_k)$ and its inverse is L'-Lipschitz on J_k, with $L' > 0$. We also assume that at least one of the two following conditions is satisfied.*

i) For any $k \in [K]$, $\inf_{x\in\mathbb{R}} \phi_k(x) > 0$.

ii) For any $k \in [K]$, $\phi_k > 0$, and $\sqrt{\phi_k}$ and $\log\phi_k$ are L_1-Lipschitz with $L_1 > 0$.

Assumption 3.1 is needed in Sulem et al. [2021] to prove the posterior concentration rates, and is verified for commonly used link functions (see Example 1 in Sulem et al. [2021]). We also need this assumption to obtain the concentration of the variational posterior distribution since our proofs leverage this existing theory.

We define the parameter space $\mathcal{F}$ as follows

$$\mathcal{H}' = \{h : [0,A] \to \mathbb{R};\ \|h\|_\infty < \infty\}, \quad \mathcal{H} = \left\{h = (h_{lk})_{l,k=1}^K \in {\mathcal{H}'}^{K^2};\ (h,\phi) \text{ satisfy } \mathbf{(C1)} \text{ or } \mathbf{(C2)}\right\},$$
$$\mathcal{F} = \left\{f = (\nu,h) \in (\mathbb{R}_+\backslash\{0\})^K \times \mathcal{H};\ (f,\phi) \text{ satisfies Assumption 3.1}\right\},$$

and the L_1-distance for any $f, f' \in \mathcal{F}$ as

$$\|f - f'\|_1 := \left\|\nu - \nu'\right\|_1 + \left\|h - h'\right\|_1, \quad \left\|h - h'\right\|_1 := \sum_{l,k=1}^K \left\|h_{lk} - h'_{lk}\right\|_1, \quad \left\|\nu - \nu'\right\|_1 := \sum_k |\nu_k - \nu'_k|.$$

In our main assumptions, we will consider the following neighbourhood around f_0 in supremum norm

$$B_\infty(\epsilon) = \left\{f \in \mathcal{F};\ \nu_k^0 \leqslant \nu_k \leqslant \nu_k^0 + \epsilon,\ h_{lk}^0 \leqslant h_{lk} \leqslant h_{lk}^0 + \epsilon,\ (l,k) \in [K]^2\right\}, \quad \epsilon > 0.$$

Finally, we define

$$\kappa_T := 10(\log T)^r, \tag{12}$$

with $r = 0$ if $(\phi_k)_k$ satisfies Assumption 3.1 (i), and $r = 1$ if $(\phi_k)_k$ satisfies Assumption 3.1 (ii).

Theorem 3.2. *Let N be a Hawkes process with link functions $\phi = (\phi_k)_k$ and parameter $f_0 = (\nu_0, h_0)$ such that (ϕ, f_0) satisfy Assumption 3.1. Let $\epsilon_T = o(1/\sqrt{\kappa_T})$ be a positive sequence verifying $\log^3 T = O(T\epsilon_T^2)$, Π be a prior distribution on $\mathcal{F}$ and $\mathcal{V}$ be a variational family of distributions on $\mathcal{F}$. We assume that the following conditions are satisfied for T large enough.*

(A0) *There exists $c_1 > 0$ such that $\Pi(B_\infty(\epsilon_T)) \geqslant e^{-c_1 T\epsilon_T^2}$.*

(A1) *There exist $\mathcal{H}_T \subset \mathcal{H}$, $\zeta_0 > 0$, $c_2 > 0$ and $x_0 > 0$ such that, with $\Theta_T = \{\theta \in \Theta,\ 0 < \theta_k \leqslant e^{c_2 T\epsilon_T^2}\}$, $\Pi(\mathcal{H}_T^c) + \Pi(\Theta_T^c) = o(e^{-(\kappa_T + c_1)T\epsilon_T^2})$ and $\log \mathcal{N}(\zeta_0\epsilon_T, \mathcal{H}_T, \|.\|_1) \leqslant x_0 T\epsilon_T^2$.*

(A2) *There exists $Q \in \mathcal{V}$ such that $supp(Q) \subset B_\infty(\epsilon_T)$ and $KL(Q\|\Pi) = O(\kappa_T T\epsilon_T^2)$.*

Then, for any $M_T \to \infty$ and $\hat{Q}$ defined in (5), we have that

$$\hat{Q}\left(\|f - f_0\|_1 > M_T\sqrt{\kappa_T}\epsilon_T\right) \xrightarrow[T\to\infty]{} 0 \quad \mathbb{P}_0 - a.s.,$$

or equivalently,

$$\mathbb{E}_0\left[\hat{Q}(\|f - f_0\|_1 > M_T\sqrt{\kappa_T}\epsilon_T)\right] \xrightarrow[T\to\infty]{} 0.$$

Remark 3.3. *Similarly to Donnet et al. [2020] Sulem et al. [2021], Theorem 3.2 also holds when the neighborhoods $B_\infty(\epsilon_T)$ around f_0 in supremum norm, considered in assumptions (A0) and (A2), are replaced by the following L_2-balls*

$$B_2(\epsilon_T, B) = \left\{f \in \mathcal{F};\ \max_k |\nu_k - \nu_k^0| \leqslant \epsilon_T,\ \max_{l,k} \|h_{lk} - h_{lk}^0\|_2 \leqslant \epsilon_T,\ \max_l \nu_l + \max_k \|h_{kl}\|_\infty < B\right\},$$

with $B > 0$ and κ_T replaced by $\kappa_T' = 10(\log\log T)(\log T)^r$.

Remark 3.4. *Theorem 3.2 also holds under a more general condition on the variational family*

(A2') *The variational family $\mathcal{V}$ verifies $\min_{Q\in\mathcal{V}} KL(Q\|\Pi(.|N)) = O(\kappa_T T\epsilon_T^2)$.*

*However, in practice, one often verifies **(A2)** and deduces **(A2')** using the following steps from Zhang and Gao [2017]. For any $Q \in \mathcal{V}$, we have that*

$$KL(Q\|\Pi(.|N)) \leqslant KL(Q\|\Pi) + Q(KL(d\mathbb{P}_0, d\mathbb{P}_f)) = KL(Q\|\Pi) + Q(KL(\mathbb{P}_{T,f_0}, \mathbb{P}_{T,f})),$$

where we denote $\mathbb{P}_{T,f_0} = e^{L_T(f_0)}$ and $\mathbb{P}_{T,f} = e^{L_T(f)}$. Using Lemma S6.1 from Sulem et al. [2021], for any $f \in B_\infty(\epsilon_T)$, we also have that

$$\mathbb{E}_0\left[L_T(f_0) - L_T(f)\right] \leqslant \kappa_T T\epsilon_T^2.$$

*Therefore, under **(A2)**, there exists $Q \in \mathcal{V}$ such that $KL(Q\|\Pi(.|N)) = O(\kappa_T T\epsilon_T^2)$, which implies **(A2')** .*

Remark 3.5. *On the one hand, assumptions **(A0)** and **(A1)** are similar to the ones of Theorem 3.2 in Sulem et al. [2021]. They are sufficient conditions for proving that the posterior concentration rate is at least as fast as $\sqrt{\kappa_T}\epsilon_T$. On the other hand, **(A2)** (or **(A2')**), is the only condition on the variational class, and informally states that this family of distributions can approximate well enough the true posterior. Nonetheless, as previously noted by Nieman et al. [2021], we could well have $\min_{Q\in\mathcal{V}} KL(Q\|\Pi(.|N)) \xrightarrow[T\to\infty]{} \infty$ under **(A2)**.*

The proof of Theorem 3.2 is reported in Section A.2.

3.2 Applications to variational classes of interest

In this section, we apply our previous result to variational inference methods of interest in nonlinear Hawkes models. We consider the mean-field and spike-and-slab variational families introduced in Section 2.3, and verify our general conditions on two nonparametric prior families, namely the random histogram prior and the Gaussian process prior. We then obtain explicit concentration rates for the variational posterior distribution and for Hölder-smooth classes of functions.

We recall from Sulem et al. [2021] that in the spike-and-slab prior construction from Section 2.2, we can in fact replace assumption **(A0)** by

(A0') There exists $c_1 > 0$ such that $\Pi(B_\infty(\epsilon_T)|\delta = \delta_0) \geqslant e^{-c_1 T\epsilon_T^2/2}$ and $\Pi_\delta(\delta = \delta_0) \geqslant e^{-c_1 T\epsilon_T^2/2}$,

and that one can choose for instance, $\Pi_\delta = \mathcal{B}(p)^{K^2}$ with $0 < p < 1$, implying that the δ_{lk}'s are i.i.d. Bernoulli random variables. Then, for any fixed p, one only needs to verify $\Pi_{h|\delta}(B_\infty(\epsilon_T)|\delta = \delta_0) \geqslant e^{-c_1 T\epsilon_T^2/2}$.

3.2.1 Mean-field variational family

We consider the mean-field variational inference method within a latent variable augmentation scheme from Section 2.2. We recall our notation $\mathcal{V}_{AMF}$, $\hat{Q}_{AMF}$ and $\mathbb{P}_A$, for respectively the (augmented) mean-field variational family and variational posterior, and the prior distribution on the latent variable. In this section, we apply Theorem 3.2 to $\hat{Q}_{AMF}$ and two nonparametric prior distributions of interest. Practical algorithms in this context will also be designed in Section 4 for the sigmoid Hawkes model. We first note in this scheme, the augmented prior distribution $\Pi \times \mathbb{P}_A \in \mathcal{V}_{AMF}$, therefore assumption **(A2)** is automatically satisfied for $\mathcal{V}_{AMF}$, therefore, we only need to verify conditions **(A0')** and **(A1)** on the prior distribution.

Random histogram prior We consider a random histogram prior for $\tilde{\Pi}_h$, the prior on the slabs from Section 2.2. This prior family is notably used in Donnet et al. [2020], Sulem et al. [2021], and is similar to the basis decomposition prior in Zhou et al. [2021b,a]. We consider a regular partition of $(0, A]$, $(t_j)_{j=0,\dots,J}$ with $t_j = jA/J$ and $J \geqslant 0$, and define piecewise-constant functions as

$$h^w_{lk}(x) = \sum_{j=1}^{J} w^j_{lk} e_j(x), \quad e_j(x) = \frac{J}{A} \mathbb{1}_{(t_{j-1}, t_j]}(x), \quad \forall j \in [J],$$

with $w^j_{lk} \in \mathbb{R}, \forall l, k = 1, \dots, K$ and $j = 1, \dots, J$. We then consider a prior on the number of pieces J, $J \sim \mathcal{P}(\lambda)$ with $\lambda > 0$, then a normal prior distribution on each w_{lk} given J, i.e.,

$$w_{lk}|J \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0_J, K_J), \quad K_J = \sigma_0^2 I_J, \quad \sigma_0 > 0.$$

With this prior construction, assumptions **(A0')** and **(A1)** are directly satisfied. For instance, this Gaussian random histogram prior is a particular case of the spline prior family in Sulem et al. [2021], with a spline basis of order $q = 0$. It would also be straightforward to verify that these conditions also hold for the shrinkage prior of Zhou et al. [2021b] based on the Laplace distribution $p_{Lap}(w^j_{lk}; 0, b) = (2b)^{-1} \exp\{-|w^j_{lk}|/b\}$ with $b > 0$, and for a "locally spike-and-slab" prior inspired by the prior of Donnet et al. [2020], Sulem et al. [2021] such as

$$w^j_{lk}|J \overset{\text{i.i.d.}}{\sim} p\delta_0 + (1-p) p_{Lap}(.; 0, b), \quad p \in (0, 1), \quad b > 0,$$

where δ_0 is the Dirac measure at 0.

Proposition 3.6. *Let N be a Hawkes process with link functions $\phi = (\phi_k)_k$ and parameter $f_0 = (\nu_0, h_0)$ such that (ϕ, f_0) verify Assumption 3.1. Assume that for any $l, k \in [K]$, $h^0_{lk} \in \mathcal{H}(\beta, L_0)$ with $\beta \in (0, 1)$ and $L_0 > 0$. Then, under the above Gaussian random histogram prior, the mean-field variational distribution $\hat{Q}_1$ defined in (7) satisfies, for any $M_T \to +\infty$,*

$$\hat{Q}_1 \left(\|f - f_0\|_1 > M_T (\log T)^q T^{-\beta/(2\beta+1)} \right) \xrightarrow[T\to\infty]{} 0 \quad \mathbb{P}_0 - a.s.,$$

with $q = 0$ if ϕ verifies Assumption 3.1(i) and $q = 1/2$ if ϕ verifies Assumption 3.1(ii).

The proof of Proposition 3.6 is omitted since it is a direct application of Theorem 3.2 to mean-field variational families in the context of a latent variable augmentation scheme.

Gaussian process prior Gaussian process priors are commonly used for nonparametric estimation of Hawkes processes Zhang et al. [2020], Zhou et al. [2020], Malem-Shinitski et al. [2021]. We consider $\tilde{\Pi}_h$ a centered Gaussian process distribution with covariance function k_{GP}, i.e., for each $l, k \in [K]$ and for any $n \geqslant 1$, $x_1, \dots, x_n \in [0, A]$,

$$(h_{lk}(x_i))_{i=1,\dots,n} \sim \mathcal{N}\left(0_n, (k_{GP}(x_i, x_j))_{i,j=1,\dots,n}\right).$$

Here, we verify assumptions **(A0')** and **(A1)** using the L_2-neighborhoods (see Remark 3.3), i.e., we check that there exist $\mathcal{H}_T \subset \mathcal{H}$ and $c_1, x_0, \zeta_0 > 0$ such that

$$\Pi(\mathcal{H}_T^c) \leqslant e^{-(\kappa_T + c_1) T \epsilon_T^2}, \quad \log \mathcal{N}(\zeta_0 \epsilon_T, \mathcal{H}_T, \|.\|_1) \leqslant x_0 T \epsilon_T^2, \quad \Pi(B_2(\epsilon_T, B)) \geqslant e^{-c_1 T \epsilon_T^2}.$$

It is therefore enough to find $\mathcal{B}_T \subset L_2([0, A])$ such that

$$\tilde{\Pi}_h(\mathcal{B}_T^c) \leqslant e^{-(\kappa_T + c_1) T \varepsilon_T^2}, \quad \log \mathcal{N}(\zeta_0 \epsilon_T, \mathcal{B}_T, \|.\|_1) \leqslant x_0 T \epsilon_T^2/2, \quad \tilde{\Pi}_h(\left\|h - h^0_{lk}\right\|_2 < \epsilon_T) \geqslant e^{-c_2 T \varepsilon_T^2}/K^2,$$

and define $\mathcal{H}_T = \mathcal{B}_T^{\otimes K^2}$, since

$$\Pi(\mathcal{H}_T^c) \leqslant \tilde{\Pi}(\mathcal{B}_T^c), \quad \log \mathcal{N}(\zeta \epsilon_T, \mathcal{H}_T, \|.\|_1) \leqslant 2 \log K + \mathcal{N}(\zeta_2 \epsilon_T, \mathcal{B}_T, \|.\|_1), \quad \Pi(B_2(\epsilon_T, B)) \geqslant \prod_{l,k} \tilde{\Pi}_h \left(\left\|h - h^0_{lk}\right\|_2 < \epsilon_T \right).$$

These conditions are easily deduced from Theorem 2.1 in van der Vaart and van Zanten [2009b] that we recall here. Let $\mathbb{H}$ be the Reproducing Kernel Hilbert Space of k_{GP} and $\phi_{h_0}(\varepsilon)$ be the concentration function associated to $\tilde{\Pi}_h$ defined as

$$\phi_{h_0}(\varepsilon) = \inf_{h\in\mathbb{H},\|h-h_0\|_2\leqslant\varepsilon} \|h-h_0\|_{\mathbb{H}} - \log\tilde{\Pi}(\|h\|_2 \leqslant \varepsilon), \quad \varepsilon > 0.$$

For any $\epsilon_T > 0$ such that $\phi_{h_0}(\epsilon_T) \leqslant T\epsilon_T^2$, there exists $\mathcal{B}_T \subset L_2([0,A])$ satisfying

$$\tilde{\Pi}_h(\mathcal{B}_T^c) \leqslant e^{-CT\epsilon_T^2}, \quad \log\mathcal{N}(3\epsilon_T, \mathcal{B}_T, \|.\|_2) \leqslant 6CT\varepsilon_T^2, \quad \tilde{\Pi}_h(\left\|h - h_{lk}^0\right\|_\infty < 2\epsilon_T) \geqslant e^{-T\epsilon_T^2},$$

for any $C > 1$ such that $e^{-CT\epsilon_T^2} < 1/2$. Since $\|h\|_1 \leqslant \sqrt{A}\,\|h\|_2$, we then obtain that

$$\log\mathcal{N}(3\sqrt{A}\epsilon_T, \mathcal{B}_T, \|.\|_1) \leqslant \log\mathcal{N}(3\epsilon_T, \mathcal{B}_T, \|.\|_2) \leqslant 6CT\epsilon_T^2,$$

and finally, that $\log\mathcal{N}(\zeta_0\epsilon_T, \mathcal{H}_T, \|.\|_1) \leqslant 2\log K + 6CT\epsilon_T^2 \leqslant x_0 T\epsilon_T^2$ with $\zeta_0 = 3\sqrt{A}$, $x_0 = 12C$.

Although more general kernel functions k_{GP} could be considered, we focus on squared exponential kernels for which

$$\forall x, y \in \mathbb{R}, \quad k_{GP}(x,y;\ell) = \exp\left\{-(x-y)^2/\ell^2\right\}, \quad \ell \sim IV(\ell; a_0, a_1), \quad a_0, a_1 > 0,$$

where $IV(.; a_0, a_1)$ with $a_0, a_1 > 0$ is the Inverse Gamma distribution. The squared exponential kernel is notably chosen in the variational method of Malem-Shinitski et al. [2021], and its adaptivity and near-optimality has been proved by van der Vaart and van Zanten [2009a].

Proposition 3.7. *Let N be a Hawkes process with link functions $\phi = (\phi_k)_k$ and parameter $f_0 = (\nu_0, h_0)$ such that (ϕ, f_0) verify Assumption 3.1. Assume that for any $l, k \in [K]$, $h_{lk}^0 \in \mathcal{H}(\beta, L_0)$ with $\beta > 0$ and $L_0 > 0$. Let $\tilde{\Pi}$ be a Gaussian Process prior with squared exponential kernel k_{GP}. Then, under the above Gaussian process and inverse Gamma prior, the mean-field variational distribution $\hat{Q}_1$ defined in (7) satisfies, for any $M_T \to +\infty$,*

$$\hat{Q}_1\left(\|f - f_0\|_1 > M_T(\log\log T)^{1/2}(\log T)^q T^{-\beta/(2\beta+1)}\right) \xrightarrow[T\to\infty]{} 0 \quad \mathbb{P}_0 - a.s.,$$

with $q = 1$ if ϕ verifies Assumption 3.1(i) and $q = 3/2$ if ϕ verifies Assumption 3.1(ii).

Proposition 3.7 is also a direct consequence of Theorem 3.2 and van der Vaart and van Zanten [2009a], therefore its proof is omitted. In practice, the Gaussian process prior is used in variational methods for Hawkes processes when there exists a conjugate form of the mean-field variational posterior distribution, i.e., $\hat{Q}_1$ is itself a Gaussian process with mean function m_{VP} and kernel function k_{VP}. For nonlinear Hawkes models, this is notably the case in the sigmoid model, under the latent variable augmentation scheme recalled in Section 4.2 Malem-Shinitski et al. [2021]. Nonethelles, the computation of the Gaussian process variational distribution is often expensive for large data set, therefore Malem-Shinitski et al. [2021] further approximate the posterior distribution using the sparse Gaussian process approximation via inducing variables Titsias and Lázaro-Gredilla [2011]. This leads to an "inducing variable" mean-field variational posterior, which is itself a variational approximation of the original mean-field variational posterior Nieman et al. [2021]. Using the results of Nieman et al. [2021], we conjecture that we could also show that our result in Proposition 3.7 also holds for the "inducing variable" mean-field variational posterior.

3.2.2 Spike-and-slab variational family

In this section, we consider the spike-and-slab variational family $\mathcal{V}_{SAS}$ from Definition 2.5 and first assume that the true connectivity graph parameter δ_0 is known. Then, the variational distribution on δ reduces to a Dirac measure at δ_0. One can then use any family of distributions for $f|\delta_0$, e.g., the augmented mean-field and prior families and our results in Section 3.2.1, directly apply to this context.

Now, if δ_0 is unknown, since the spike-and-slab variational posterior may not be tractable for general distributions $Q_\delta(\delta)$, we propose to infer the graph within a variational Bayes approach with model selection (see Section 2.2). To avoid ambiguity with the Hawkes *model*, we use the term *graph selection*. We then adapt the general results of Zhang and Gao [2017] and Ohn and Lin [2021] to obtain the concentration rates of the *graph selection* variational posterior and the *adaptive* variational posterior, respectively defined in (9) and (10). We recall that the graph $\delta \in \{0,1\}^{K\times K}$ belongs to a set of cardinality 2^K for any fixed $K \geqslant 1$.

From Theorem 4.1 in Zhang and Gao [2017], the prior mass conditions are enough in this case to ensure the concentration of the variational posterior (9). More precisely, we define a *graph-selection* variational posterior as follows. For any $\delta \in \{0,1\}^{K^2}$, we denote

$$\mathcal{V}_{SAS}^{(\delta)} = \left\{Q = Q_{f|\delta}\right\}, \quad \hat{Q}^{(\delta)} = \arg\min_{Q\in\mathcal{V}_{SAS}^{(\delta)}} KL(Q\|\Pi(.|N)), \quad \hat{\delta} = \arg\max_{\delta\in\{0,1\}^{K\times K}} ELBO(\hat{Q}^{(\delta)}),$$

where the ELBO is defined in (8). The graph selection variational posterior distribution is then defined as

$$\hat{Q}_{GS} := \hat{Q}^{(\hat{\delta})}. \tag{13}$$

Proposition 3.8. *Let N be a Hawkes process with link functions $\phi = (\phi_k)_k$, parameter $f_0 = (\nu_0, h_0)$ such that (ϕ, f_0) verify Assumption 3.1. Let $\epsilon_T = o(1/\sqrt{\kappa_T})$ be a positive sequence verifying $\log^3 T = O(T\epsilon_T^2)$ and Π be a prior distribution on $\mathcal{F}$ satisfying **(A0')** and **(A1)**. Then, for the graph selection variational posterior (13), we have that*

$$\hat{Q}_{GS}\left(\|f - f_0\|_1 > M_T\sqrt{\kappa_T}\epsilon_T\right) \xrightarrow[T\to\infty]{} 0 \quad \mathbb{P}_0 - a.s.$$

We note that explicit concentration rates for Hölder-smooth functions can then be derived when using the prior families of Section 3.2.1. Proposition 3.8 is a direct consequence of Theorem 3.2 and Theorem 4.1 in Zhang and Gao [2017], therefore its proof is omitted. Finally, we also obtain a similar result for the adaptive variational posterior, defined in this context as

$$\hat{Q}_{AD}(f) = \sum_{\delta\in\{0,1\}^{K\times K}} \hat{\gamma}_\delta \hat{Q}^{(\delta)}, \tag{14}$$

where $\hat{\gamma}_\delta$ are the marginal probabilities defined as

$$\hat{\gamma}_\delta = \frac{\Pi_\delta(\delta)\exp\left\{ELBO(\hat{Q}^{(\delta)})\right\}}{\sum_{\delta\in\{0,1\}^{K\times K}}\Pi_\delta(\delta)\exp\left\{ELBO(\hat{Q}^{(\delta)})\right\}}, \quad \forall\delta\in\{0,1\}^{K\times K}.$$

We note that from Theorem 2.1 of Ohn and Lin [2021], the adaptive variational posterior can be alternatively defined as $\hat{Q}_{AD} = \arg\min_{Q\in\mathcal{V}_{SAS}} KL(Q\|\Pi(.|N))$ with

$$\mathcal{V}_{SAS} := \left\{Q = \sum_{\delta\in\{0,1\}^{K\times K}} \gamma_\delta Q_{f|\delta},\ \sum_\delta \gamma_\delta = 1,\ \gamma_\delta \geqslant 0,\ \forall\delta\right\}.$$

The following result is adapted from Theorem 3.6 in Ohn and Lin [2021] and directly holds under the same assumptions as Proposition 3.8.

Proposition 3.9. *Let N be a Hawkes process with link functions $\phi = (\phi_k)_k$, parameter $f_0 = (\nu_0, h_0)$ such that (ϕ, f_0) verify Assumption 3.1, and connectivity graph δ_0. Let $\epsilon_T = o(1/\sqrt{\kappa_T})$ be a positive sequence verifying $\log^3 T = O(T\epsilon_T^2)$ and Π be a distribution on $\mathcal{F}$ satisfying **(A0')** and **(A1)**. Then, for the adaptive variational posterior (14), we have that*

$$\hat{Q}_{AD}\left(\|f - f_0\|_1 > M_T\sqrt{\kappa_T}\epsilon_T\right) \xrightarrow[T\to\infty]{} 0 \quad \mathbb{P}_0 - a.s.$$

4 Adaptive mean-field variational algorithms in the sigmoid model

In this section, we consider the Hawkes model with sigmoid link functions, for which an efficient mean-field variational methodology based on data augmentation and Gaussian priors has been previously proposed Malem-Shinitski et al. [2021], Zhou et al. [2021a, 2022]. Here, we consider the following parametrisation of this model with link functions

$$\phi_k(x) = \theta_k\tilde{\sigma}(x), \quad \tilde{\sigma}(x) = \sigma\left(\alpha(x-\eta)\right), \quad \sigma(x) = (1+e^{-x})^{-1}, \quad \alpha,\eta,\theta_k > 0, \quad k\in[K]. \tag{15}$$

We note that for $\alpha = 0.1, \eta = 10$ and $\theta_k = 20$, the nonlinearity ϕ_k is similar to the ReLU and softplus functions on $[-\infty, 20]$ (see Figure 1 in Section 5). For this model, we first prove new concentration results for the posterior distribution, when the true scale parameter, denoted $\theta_0 = (\theta_k^0)_k$, is unknown. Secondly, we describe the latent variable augmentation scheme which allows to obtain a conjugate form of a variational posterior distribution, and, building on prior work, we propose a novel adaptive and sparsity-inducing mean-field variational method. In particular, our approach consists in reducing the dimensionality of the problem by inferring the connectivity graph parameter using the graph selection approach described in Section 3.2.2.

4.1 Posterior concentration rates with unknown scale parameter

First, we state a lemma that ensures the identifiability of the sigmoid Hawkes model with link functions (15) and unknown scale. We will use the following assumption.

Assumption 4.1. *For $f = (\nu, h)$, we assume that*

$$\forall k\in[K],\ \exists l\in[K],\ \exists x_2 > x_1 > 0,\ \exists c_* > 0,\ \forall x\in[x_1, x_2],\ h_{lk}^+(x) > c_*.$$

Remark 4.2. *Assumption 4.1 requires that every component N^k receives some excitation effect from at least one other component. This assumption ensures that the intensity function $\lambda_t^k(f)$ can approach its upper bound θ_k with non-zero probability.*

Lemma 4.3. *Let N be a sigmoid Hawkes process with link functions $(\phi_k)_k$ defined in (15), scale θ, and parameter $f = (\nu, h)$ satisfying Assumption 4.1. If N' is a sigmoid Hawkes process with scale θ', parameter $f' = (\nu', h')$ satisfying Assumption 4.1, then*

$$N \overset{\mathcal{L}}{=} N' \implies \nu = \nu' \quad \text{and} \quad h = h' \quad \text{and} \quad \theta = \theta'.$$

The proof of this lemma is reported in Appendix B. We now consider the problem of estimating both $f = (\nu, h)$ and the scale parameter θ of the link functions (15), and study the concentration properties of the posterior distribution on (f, θ). In this context, we define a bounded parameter space as, for $B > 0$,

$$\mathcal{F}' = \left\{ f = (\nu, h);\ \nu_k \in [-B, B]^K,\ \|h_{lk}\|_\infty < B,\ \forall l, k \right\},$$

and our parameter space is now $\mathcal{F}' \times \Theta$ with $\Theta = (\mathbb{R}_+ \backslash \{0\})^K$. With a slight abuse of notation, we will use $\mathcal{F}$ for $\mathcal{F}'$. We note that for the sigmoid model, we do not need the background rates (ν_k) to be positive. With Π a prior distribution on $\mathcal{F} \times \Theta$, the posterior distribution is defined as

$$\Pi(O|N) = \frac{\int_B \exp(L_T(f)) d\Pi(f, \theta)}{\int_{\mathcal{F} \times \Theta} \exp(L_T(f, \theta)) d\Pi(f, \theta)}, \quad O \subset \mathcal{F} \times \Theta. \tag{16}$$

We assume that $f_0 \in \mathcal{F}$ and define the L_2-neighbourhoods around (f_0, θ_0) as

$$\tilde{B}_2(\epsilon_T, B) = \left\{ (f, \theta) \in \mathcal{F} \times \Theta;\ \max_k |\nu_k - \nu_k^0| \leqslant \epsilon_T, \max_k |\theta_k - \theta_k^0| \leqslant \epsilon_T,\ \max_{l,k} \|h_{lk} - h_{lk}^0\|_2 \leqslant \epsilon_T \right\}.$$

Remark 4.4. *We introduce a bounded parameter space for the sigmoid function to satisfy the Lipschitz condition on the inverse of the link functions in Assumption 3.1. In fact, the sigmoid inverse, i.e., the logit function $\sigma^{-1}(x) = \log \frac{x}{x-1}$, is not Lipschitz on $(0, 1)$. Therefore, this assumption does not hold in the sigmoid Hawkes model, unless the parameter space is bounded in supremum norm so that one can obtain the Lipschitz condition on a bounded domain (since, in this case, the linear intensity $\tilde{\lambda}_t(f)$ defined in (2) is also bounded).*

We now state our concentration result on the posterior distribution (16).

Proposition 4.5. *Let N be a sigmoid Hawkes process with link functions $(\phi_k)_k$ defined in (15), scale parameter θ_0, and parameter $f_0 = (\nu_0, h_0)$ such that f_0 satisfies Assumption 4.1. Let $\kappa_T = 10(\log\log T)\log T$ and $\epsilon_T = o(1/\sqrt{\kappa_T})$ be a positive sequence verifying $\log^3 T = O(T\epsilon_T^2)$. Let $\Pi = \Pi_\nu \times \Pi_\theta \times \Pi_h$ be a prior distribution on $\mathcal{F} \times \Theta$. We assume that for T large enough, the following assumptions hold.*

(A0") *There exists $c_1 > 0$ such that $\Pi(\tilde{B}_2(\epsilon_T, B)) \geqslant e^{-c_1 T\epsilon_T^2}$.*

(A1) *There exist $\mathcal{H}_T \subset \mathcal{H}$, $\zeta_0 > 0$, $x_0 > 0$, such that $\Pi_h(\mathcal{H}_T^c) = o(e^{-(\kappa_T + c_1)T\epsilon_T^2})$ and $\log \mathcal{N}(\zeta_0\epsilon_T, \mathcal{H}_T, \|.\|_1) \leqslant x_0 T\epsilon_T^2$.*

Then, for any $M_T \to \infty$, we have that

$$\mathbb{E}_0 \left[\Pi(\|f - f_0\|_{1,B} + \|\theta - \theta_0\|_1 > M_T \sqrt{\kappa_T}\epsilon_T | N) \right] = o(1). \tag{17}$$

The previous result is an extension of Theorem 3.2 of Sulem et al. [2021], and provides guarantess for Bayesian methods that estimate the scale parameter in the sigmoid model. Moreover, from Proposition 4.5, the concentration rates of a variational posterior on (f, θ) can then be deduced, using similar construction and arguments as for Theorem 3.2. The proof of Proposition 4.5 is reported in Appendix A.3.

4.2 Augmented mean-field variational inference

In this section, we recall existing latent variable augmentation strategy in the sigmoid Hawkes model, and the definition of the augmented mean-field variational distribution in this context Malem-Shinitski et al. [2021], Zhou et al. [2021a]. From now on, we consider that the scale parameter θ is known, however, our methodology can be directly extended to estimate an unknown θ.

The first step consists in re-writing the sigmoid function as a mixture of Polya-Gamma random variables Polson et al. [2012], i.e.,

$$\sigma(x) = \mathbb{E}_{\omega \sim p_{PG}(.;1,0)} \left[e^{g(\omega, x)} \right] = \int_0^{+\infty} e^{g(\omega,x)} p_{PG}(\omega; 1, 0) d\omega, \quad g(\omega, x) = -\frac{\omega x^2}{2} + \frac{x}{2} - \log 2, \tag{18}$$

with $p_{PG}(.;1,0)$ the Polya-Gamma density. We recall that $p_{PG}(.;1,0)$ is the density of the random variable

$$\frac{1}{2\pi^2}\sum_{k=1}^{\infty}\frac{g_k}{(k-1/2)^2}, \quad g_k \overset{\text{i.i.d.}}{\sim} Gamma(1,1),$$

and that the *tilted* Polya-Gamma distribution is defined as

$$p_{PG}(\omega;1,c) = \cosh\left(\frac{c}{2}\right)\exp\left\{-\frac{c^2\omega}{2}\right\}p_{PG}(\omega;1,0), \quad c \geqslant 0,$$

where cosh denotes the hyperbolic cosine function. With a slight abuse of notation, we re-define the linear intensity (2) as

$$\tilde{\lambda}_t^k(f) = 0.1\left(\nu_k + \sum_{l=1}^{K}\int_{-\infty}^{t^-} h_{lk}(t-s)dN_s^l - 10.0\right),$$

so that we have $\lambda_t^k(f) = \theta_k\sigma(\tilde{\lambda}_t^k(f)), t \in \mathbb{R}$. For any $k \in [K]$, let $N_k := N^k[0,T]$ and $T_1^k, \ldots, T_{N_k}^k \in [0,T]$ be the times of events at component N^k. Now, let $\omega = (\omega_i^k)_{k\in[K], i\in[N_k]}$ be a set of latent variables such that

$$\omega_i^k \overset{\text{i.i.d.}}{\sim} p_{PG}(\omega_i^k;1,0), \quad i \in [N_k], \quad k \in [K].$$

Then, using (18), an *augmented* log-likelihood function can be defined as

$$\begin{aligned} L_T(f,\omega;N) &= \sum_{k\in[K]}\left\{\sum_{i\in[N_k]}\left(\log\theta_k + g(\omega_i^k, \tilde{\lambda}_{T_i^k}(f)) + \log p_{PG}(\omega_i^k;1,0)\right) - \int_0^T \theta_k\sigma(\tilde{\lambda}_t^k(f))dt\right\} \\ &= \sum_{k\in[K]}\left\{\sum_{i\in[N_k]}\left(\log\theta_k + g(\omega_i^k, \tilde{\lambda}_{T_i^k}(f)) + \log p_{PG}(\omega_i^k;1,0)\right) - \int_0^T\int_0^\infty \theta_k e^{g(\bar{\omega},\tilde{\lambda}_t^k(f))}p_{PG}(\bar{\omega};1,0)d\bar{\omega}dt\right\}. \end{aligned} \tag{19}$$

Secondly, Campbell's theorem Daley and Vere-Jones [2007], Kingman [1993] is used to re-write the integral term on the RHS in (19). We first recall here its general formulation. For a Poisson point process $\bar{N}$ on a space $\mathcal{X}$ with intensity measure $\Lambda : \mathcal{X} \to \mathbb{R}^+$, and for any function $\zeta : \mathcal{X} \to \mathbb{R}$, it holds true that

$$\mathbb{E}\left[\prod_{x\in\bar{N}} e^{\zeta(x)}\right] = \exp\left\{(e^{\zeta(x)}-1)\Lambda(dx)\right\}. \tag{20}$$

Therefore, using that $\sigma(x) = 1 - \sigma(-x)$, and considering for each k a marked Poisson point process $\bar{N}^k$ on $\mathcal{X} = ([0,T],\mathbb{R}^+)$ with intensity measure $\Lambda^k(t,\omega) = \theta_k p_{PG}(\omega;1,0)$, and distribution $\mathbb{P}_{\bar{N}}$, applying Campbell's theorem with $\zeta(t,\omega) := g(\omega, -\tilde{\lambda}_t^k(f))$, one obtains that

$$\mathbb{E}\left[\prod_{(\bar{T}_j^k,\bar{\omega}_j^k)\in\bar{N}^k} e^{g(\bar{\omega}_k, -\tilde{\lambda}_{\bar{T}_j}^k(f))}\right] = \exp\left\{\int_0^T\int_0^\infty \theta_k\left(e^{g(\bar{\omega},-\tilde{\lambda}_t^k(f))}-1\right)p_{PG}(\bar{\omega};1,0)d\bar{\omega}dt\right\}.$$

Conditionally on N, let $\bar{N} := (\bar{N}^1, \ldots, \bar{N}^K)$ be an observation of the previous Poisson process on $[0,T]$. We denote $\bar{N}_k := \bar{N}^k[0,T]$ and $(\bar{T}_1^k, \bar{\omega}_1^k), \ldots, (\bar{T}_1^k, \bar{\omega}_{\bar{N}_k}^k) \in [0,T]\times\mathbb{R}_+$ the times and marks of $\bar{N}_k$ for each k. Then, a *doubly augmented* log-likelihood function can be defined as

$$L_T(f,\omega,\bar{N};N) = \sum_{k\in[K]}\left\{\sum_{i\in[N_k]}\left[\log\theta_k + g(\omega_i^k, \tilde{\lambda}_{T_i^k}(f)) + \log p_{PG}(\omega_i^k;1,0)\right] + \sum_{j\in[\bar{N}_k]}\left[\log\theta_k + g(\bar{\omega}_j^k, -\tilde{\lambda}_{\bar{T}_j}(f)) + \log p_{PG}(\bar{\omega}_j^k;1,0) - \theta_k T\right]\right\}.$$

This construction allows to define *augmented* posterior distribution as

$$\Pi(f,\omega,\bar{N}|N) \propto \prod_k\left\{\prod_{i\in[N_k]}\theta_k e^{g(\omega_i^k,\tilde{\lambda}_{T_i^k}(f))}p_{PG}(\omega_i^k;1,0)\times\prod_{j\in[\bar{N}^k]}\theta_k e^{g(\bar{\omega}_j^k,-\tilde{\lambda}_{\bar{T}_j}(f))}p_{PG}(\bar{\omega}_j^k;1,0)\right\}\times\Pi(f). \tag{21}$$

Then, in this context, an augmented mean-field variational family can be defined as

$$\mathcal{V}_{AMF} = \left\{Q;\ dQ(f,\omega,\bar{N}) = dQ_1(f)dQ_2(\omega,\bar{N})\right\}, \tag{22}$$

leading to the following variational posterior

$$\hat{Q}_{AMF}(f,\omega,\bar{N}) = \arg\min_{Q\in\mathcal{V}_{AMF}} KL\Big(Q(f,\omega,\bar{N})\|\Pi(f,\omega,\bar{N}|N)\Big) = \hat{Q}_1(f)\hat{Q}_2(\omega,\bar{N}).$$

Using (6), it then holds that

$$\hat{Q}_1(f) \propto \exp\left\{\mathbb{E}_{\hat{Q}_2}[\log p(f,\omega,\bar{N},N)]\right\}, \tag{23}$$

$$\hat{Q}_2(\omega,\bar{N}) \propto \exp\left\{\mathbb{E}_{\hat{Q}_1}[\log p(f,\omega,\bar{N},N)]\right\}. \tag{24}$$

For certain families of Gaussian priors, the variational factors $\hat{Q}_1$ and $\hat{Q}_1$ of the augmented mean-field distribution are conjugate to the prior Π and augmented distribution $\mathbb{P}_A = p_{PG}(.|1,0)\times\mathbb{P}_{\bar{N}}$, which allows to design iterative algorithms with closed-forms updates of (23) and (24). In the next section, we derive a mean-field variational inference algorithm for a fixed dimensionality of the parameter f, a method related to the algorithm of Zhou et al. [2021a].

4.3 Fixed-dimension mean-field variational algorithm

We consider a modification of the random histogram prior family, where the graph parameter $\delta = (\delta_{lk})_{l,k} \in \{0,1\}^{K\times K}$ and the size of the partition $J = 2^D$ are fixed, with $D \geqslant 0$ the partition's depth. We denote $s = (\delta, D)$ and call D the dimensionality of h. We recall our notation from Section 3.2.1 of the basis functions on $(0, A]$,

$$e_j(x) = \frac{J}{A}\mathbb{1}_{I_j}(x), \quad I_j = \left[\frac{J}{A}(j-1), \frac{J}{A}j\right), \quad j \in [J].$$

We also define

$$\mathcal{H}^D_{histo} = \left\{h = (h_{lk})_{l,k} \in \mathcal{H};\ h_{lk}(x) = \sum_{j=1}^{J} h^j_{lk} e_j(x),\ x \in [0,A],\ \underline{h}^D_{lk} = (h^1_{lk},\dots,h^J_{lk}) \in \mathbb{R}^J,\ \forall l,k \in [K]\right\}.$$

For each $l,k \in [K]$ such that $\delta_{lk} = 1$, we consider a normal prior for the distribution on $\underline{h}^D_{lk}$, with mean vector μ_D and covariance matrix Σ_D, i.e.,

$$\underline{h}^D_{lk} \sim \mathcal{N}(\mu_D, \Sigma_D).$$

For each l,k such that $\delta_{lk} = 0$, we set $\underline{h}^D_{lk} = \mathbf{0}_J$, and define $\mu_s = (\delta_{lk}\mu_D)_{l,k} \in \mathbb{R}^{JK^2}$ and $\Sigma_s = Diag((\delta_{lk}\Sigma_D)_{l,k}) \in \mathbb{R}^{JK^2\times JK^2}$. We also consider a normal prior on the background rates, i.e., $\nu_k \sim \mathcal{N}(\mu_\nu, \sigma^2_\nu), k \in [K]$. We denote $f_s := (f^s_k)_k \in \mathcal{F}_s$ where for each k,

$$f^s_k = (\nu_k, \underline{h}^D_{1k}, \dots, \underline{h}^D_{Kk}) \in \mathbb{R}^{KJ+1}.$$

We then consider the data augmentation strategy described in Section 4.2 and an augmented mean-field variational family with fixed $s = (\delta, D)$, i.e.,

$$\mathcal{V}^s_{AMF} = \left\{Q = Q_{f_s|\delta,s} : \mathcal{F}_s \to [0,1];\ dQ(f,\omega,\bar{N}) = dQ_1(f_s)dQ_2(\omega,\bar{N})\right\},$$

and we denote $\hat{Q}_s(f_s,\omega,\bar{N}) = \hat{Q}_{1s}(f_s)\hat{Q}_{2s}(\omega,\bar{N})$ the corresponding variational posterior. Following the same strategy as Donner and Opper [2019], Zhou et al. [2021a], Malem-Shinitski et al. [2021], we can derive analytic forms for $\hat{Q}_{1s}$ and $\hat{Q}_{2s}$.

Introducing the notation $H(t) = (H^0(t), H^1(t), \dots, H^K(t)) \in \mathbb{R}^{KJ+1}$, where $H_0(t) = 1$ and for $k \in [K]$, $H^k(t) = (H^k_j(t))_{j=1,\dots,J}$ with

$$H^k_j(t) := \int_{t-A}^{t} e_j(t-s)dN^k_s, \quad j \in [J], \tag{25}$$

we can prove that with $\hat{Q}_{1s}(f_s) = \prod_k \hat{Q}^k_{1s}(f^s_k)$, for each k, $\hat{Q}^k_{1s}(f^s_k)$ is a normal distribution with mean vector $\tilde{\mu}^s_k \in \mathbb{R}^{KJ+1}$ and covariance matrix $\tilde{\Sigma}^s_k \in \mathbb{R}^{(KJ+1)\times(KJ+1)}$ given by

$$\tilde{\Sigma}^s_k = \left[\alpha^2 \sum_{i\in[N_k]} \mathbb{E}_{\hat{Q}^k_{2s}}[\omega^k_i] H(T^k_i)H(T^k_i)^T + \alpha^2 \int_0^T \int_0^{+\infty} \bar{\omega}^k_t H(t)H(t)^T \Lambda^k(t,\bar{\omega})d\bar{\omega}dt + \Sigma^{-1}_s\right]^{-1},$$

$$\tilde{\mu}^s_k = \frac{1}{2}\tilde{\Sigma}^s_k \left[\alpha \sum_{i\in[N_k]} (2\mathbb{E}_{\hat{Q}^k_{2s}}[\omega^k_i]\alpha\eta + 1)H(T^k_i) + \alpha \int_0^T \int_0^{+\infty} \left(2\bar{\omega}^k\alpha\eta - 1\right) H(t)\Lambda^k(t,\bar{\omega})d\bar{\omega}dt + 2\Sigma^{-1}_s\mu_s\right],$$

where

$$\Lambda^k(t,\bar{\omega}) := \theta_k \frac{\exp\left\{-\frac{1}{2}\mathbb{E}_{Q^k_{1s}}[\tilde{\lambda}^k_t(f^s_k)]\right\}}{2\cosh\frac{(c^k_t)^2}{2}} p_{PG}(\bar{\omega}|1,c^k_t), \quad c^k_t := \sqrt{\mathbb{E}_{Q^k_{1s}}[\tilde{\lambda}^k_t(f)^2]}.$$

Besides, $\hat{Q}_{2s}(\omega,\bar{N}) = \hat{Q}_{21s}(\omega)\hat{Q}_{22s}(\bar{N})$ where

$$\hat{Q}_{21s}(\omega) = \prod_k \prod_{i\in[N_k]} p_{PG}(\omega^k_i|1,c^k_{T^k_i}),$$

and $\hat{Q}_{22s} = \prod_k \hat{Q}^k_{22s}$ where for each k, $\hat{Q}^k_{22s}$ is the probability distribution of a marked Poisson point process on $[0,T]\times\mathbb{R}^+$ with intensity measure $\Lambda^k(t,\bar{\omega})$. The derivation of these formulas is reported in Appendix C.1.

Therefore, given an estimate of $\hat{Q}_{1s}$, one can compute $\hat{Q}_{2s}$, and reciprocally, and the variational posterior distribution can be computed by updating each factor iteratively. This procedure is reported in Algorithm 1. We note that in the updates of $\tilde{\mu}^s_k$ and $\tilde{\Sigma}^s_k$, we need to compute an integral, which we perform using the Gaussian quadrature Golub and Welsch [1969] with n_{GQ} points in our implementation. We also note that in this algorithm, the outer "for" loop that computes each factor $\hat{Q}^s_k$ could be run in parallel.

We additionally note that, similarly to Zhou et al. [2021a], Malem-Shinitski et al. [2021], we can derive analytic forms of the conditional distributions of the augmented posterior distribution (21) and design a Gibbs sampler to compute the latter distribution (see Algorithm 4 in Appendix C.3).

Algorithm 1 Fixed-dimension mean-field variational inference algorithm

Input: $N, \delta, D, \mu_D, \Sigma_D, n_{iter}, n_{GQ}$.
Output: $\tilde{\mu}_D, \tilde{\Sigma}_D$.
  Precompute $(H(T^k_i))_{i,k}$.
  Precompute $(p_q, v_q)_{q\in[n_{GQ}]}$ the points and weights of the Gaussian quadrature method, and $(H(p_q))_{q\in[n_{GQ}]}$.
  for $k \leftarrow 1$ to K **do**
    Initialise $\tilde{\mu}^s_k \leftarrow \mu_s, \tilde{\Sigma}^s_k \leftarrow \Sigma_s$.
    for $t \leftarrow 1$ to n_{iter} **do**
      for $i \leftarrow 1$ to N_k **do**
        $\mathbb{E}_{\hat{Q}_{1s}}[\tilde{\lambda}^k_{T^k_i}(f^s_k)^2] = \alpha\left(H(T^k_i)^T\tilde{\Sigma}^s_k H(T^k_i) + (H(T^k_i)^T\tilde{\mu}^s_k)^2 - 2\eta H(T^k_i)^T\tilde{\mu}^s_k + \eta^2\right)$
        $\mathbb{E}_{\hat{Q}_{2s}}[\omega^k_i] \leftarrow \tanh\left(\sqrt{\mathbb{E}_{\hat{Q}_{1s}}[\tilde{\lambda}^k_{T^k_i}(f^s_k)^2]}\right) / \left(2\sqrt{\mathbb{E}_{\hat{Q}_{1s}}[\tilde{\lambda}^k_{T^k_i}(f^s_k)^2]}\right)$
      end for
      for $q \leftarrow 1$ to n_{GQ} **do**
        $\mathbb{E}_{\hat{Q}_{1s}}[\tilde{\lambda}^k_{p_q}(f^s_k)^2] = \alpha\left(H(p_q)^T\tilde{\Sigma}^s_k H(p_q) + (H(p_q)^T\tilde{\mu}^s_k)^2 - 2\eta H(p_q)^T\tilde{\mu}^s_k + \eta^2\right)$
        $\mathbb{E}_{\hat{Q}_{2s}}[\omega^k_q] \leftarrow \tanh\left(\sqrt{\mathbb{E}_{\hat{Q}_{1s}}[\tilde{\lambda}^k_{p_q}(f^s_k)^2]}\right) / \left(2\sqrt{\mathbb{E}_{\hat{Q}_{1s}}[\tilde{\lambda}^k_{p_q}(f^s_k)^2]}\right)$
        $\mathbb{E}_{\hat{Q}_{1s}}[\tilde{\lambda}^k_{p_q}(f^s_k)] = \alpha\left((\tilde{\mu}^s_k)^T H(p_q) - \eta\right)$
      end for
      $\tilde{\Sigma}^s_k = \left[\alpha^2 \sum_{i\in[N_k]} \mathbb{E}_{\hat{Q}_{2s}}[\omega^k_i] H(T^k_i)H(T^k_i)^T + \alpha^2\theta_k \sum_{q\in[n_{GQ}]} v_q \mathbb{E}_{\hat{Q}_{2s}}[\bar{\omega}^k_q] \frac{\exp(-\frac{1}{2}\mathbb{E}_{\hat{Q}_{1s}}[\tilde{\lambda}^k_{p_q}(f^s_k)])}{2\cosh\frac{1}{2}\mathbb{E}_{\hat{Q}_{1s}}[\tilde{\lambda}^k_{p_q}(f^s_k)^2]} H(p_q)H(p_q)^T + \Sigma_s^{-1}\right]^{-1}$.
      $\tilde{\mu}^s_k = \frac{1}{2}\tilde{\Sigma}^s_k\left[\alpha \sum_{i\in[N_k]} (2\mathbb{E}_{\hat{Q}_{2s}}[\omega^k_i]\alpha\eta + 1)H(T^k_i)^T + \alpha\theta_k \sum_{q\in[n_{GQ}]} v_q(2\mathbb{E}_{\hat{Q}_{2s}}[\bar{\omega}^k_q]\alpha\eta - 1)\frac{\exp(-\frac{1}{2}\mathbb{E}[\tilde{\lambda}^k_{p_q}(f^s_k)])}{2\cosh\frac{1}{2}\mathbb{E}_{\hat{Q}_{1s}}[\tilde{\lambda}^k_{p_q}(f^s_k)^2]} H(p_q)^T + 2\Sigma_s]^{-1}\mu_s\right]$.
    end for
  end for

4.4 Adaptive mean-field variational algorithms

Using the fixed-dimension approach from Section 4.2, we can now design an adaptive and sparsity-inducing variational method, that infers the graph parameter δ and the dimensionality of h. In fact, we propose two algorithms. The first one, Algorithm 2, explores all $s = (\delta, D)$, and outputs an *averaged* or *mode* variational posterior analog to (14) and (9). However, considering the 2^K graphs δ is computationally expensive for moderate K. Therefore, we also propose a more efficient two-step algorithm, Algorithm 3, which first estimates the graph $\hat{\delta}$, then computes an adaptive variational posterior that is a mixture on the restricted set $(\hat{\delta}, D)_D$.

4.4.1 Fully-adaptive mean-field variational algorithm

Given a maximum depth D_T, we first define the set

$$\mathcal{S}_T = \{s = (\delta, D);\ \delta \in \{0,1\}^{K\times K}, 1 \leqslant D \leqslant D_T\}.$$

For simplicity, given a graph δ, we assume that the depth is kept the same for all non-null h_{lk}, therefore, $|\mathcal{S}_T| \sim 2^{K^2} D_T$, and for any $s = (\delta, D) \in \mathcal{S}_T$, $|s| = (D+1)\sum_{l,k} \delta_{lk} + 1$. Let Π_s be a prior distribution on $\mathcal{S}_T$ of the form $\Pi_s(s) = \Pi_\delta(\delta)\Pi_D(D)$. Then, we define the *averaged* variational posterior as

$$\hat{Q}_{AV} = \sum_{s\in\mathcal{S}_T} \hat{\gamma}_s \hat{Q}_s, \tag{26}$$

where for each s, $\hat{Q}_s$ is the variational posterior defined in Section 4.3, and $\hat{\gamma}_s$ is the marginal probability on s defined as

$$\tilde{\gamma}_s = \Pi_\delta(\delta)\Pi_D(D)\exp\left\{ELBO(\hat{Q}_s)\right\}, \quad \hat{\gamma}_s = \tilde{\gamma}_s / \sum_{s\in\mathcal{S}_T} \tilde{\gamma}_s, \tag{27}$$

with

$$ELBO(Q) := \mathbb{E}_Q\left[\log \frac{p(f,\omega,\bar{N},N)}{Q_1(f)Q_2(\omega,\bar{N})}\right]. \tag{28}$$

We also define a *mode* variational posterior as

$$\hat{Q}_{MV} = \hat{Q}_{\hat{s}}, \quad \hat{s} = \arg\max_{s\in\mathcal{S}_T} ELBO(\hat{Q}_s). \tag{29}$$

To obtain $\hat{Q}_{AV}$ or $\hat{Q}_{MV}$, we then compute $\hat{Q}_s$ for every $s \in \mathcal{S}_T$ using Algorithm 1 and the corresponding ELBO (see Appendix C.2 for the latter derivation). We call this procedure the *fully-adaptive mean-field* algorithm, which is summarised in Algorithm 2.

Algorithm 2 Fully-adaptive mean-field variational inference

Input: $N, \mathcal{S}_T, \sigma, n_{iter}, n_{GQ}$.
Output: $\hat{Q}_{AV}$ (or $\hat{Q}_{MV}$)
  for $s = (\delta, D) \in \mathcal{S}_T$ **do**
    Compute the variational posterior $\hat{Q}_s$ using Algorithm 1 with σ, n_{iter} and n_{GQ}.
    Compute $\tilde{\gamma}_s$ using (27).
  end for
  Compute $\{\hat{\gamma}_s\}_{s\in\mathcal{S}_T}$ and $\hat{Q}_{AV}$ (or $\hat{Q}_{MV}$) using (26) and (27) (or (29)).

4.4.2 Two-step adaptive mean-field algorithm

For the multivariate setting $K > 1$, we propose a variant of the previous algorithm that reduces the computational time by avoiding to compute $\hat{Q}_s$ for all possible $s \in \mathcal{S}_T$. We first define the set $\mathcal{S}_T^{\delta_C}$ as

$$\mathcal{S}_T^{\delta_C} = \{s = (\delta_C, D);\ 1 \leqslant D \leqslant D_T\},$$

where $\delta_C = \mathbb{1}\mathbb{1}^T$ is the complete graph. The first step of our method consists in computing the mode variational distribution $\hat{Q}^{\delta_C}$ using Algorithm 2, replacing $\mathcal{S}_T$ by $\mathcal{S}_T^{\delta_C}$. Then, we use $\hat{Q}^{\delta_C}$ to estimate the norms $(\|h_{lk}\|_1)_{l,k}$ and the graph parameter. We denote $\hat{D}_C$ the selected dimensionality in $\hat{Q}^{\delta_C}$ and $J_C = 2^{D_C}$. We then define $\tilde{S} = (\tilde{S}_{lk})_{l,k} \in \mathbb{R}_+^{K\times K}$, where for any l, k,

$$\begin{aligned}\tilde{S}_{lk} = \mathbb{E}_{\hat{Q}_1^{\delta_C}}[\|h_{lk}\|_1] &= \sum_{j=1}^{J_C} \mathbb{E}_{\hat{Q}_1^{\delta_C}}[|h_{lk}^j|] \\ &= \sum_{j=1}^{J_C} \sqrt{\frac{2}{\pi}[\Sigma_{lk}^{D_C}]_{jj}}\exp\left\{-\frac{[\tilde{\mu}_{lk}^{D_C}]_j^2}{[\Sigma_{lk}^{D_C}]_{jj}}\right\} - [\tilde{\mu}_{lk}^{D_C}]_j\left[1 - 2\Phi\left(-\frac{[\tilde{\mu}_{lk}^{D_C}]_j}{\sqrt{[\Sigma_{lk}^{D_C}]_{jj}}}\right)\right],\end{aligned}$$

Given a pre-specified threshold η_0, we then compute our graph estimator $\hat{\delta} = (\hat{\delta}_{lk})_{l,k}$ as

$$\hat{\delta}_{lk} = \mathbb{1}_{\tilde{S}_{lk} > \eta_0}, \quad \forall l, k. \tag{30}$$

Secondly, with

$$\mathcal{S}_T^{\hat{\delta}} = \{s = (\hat{\delta}, D);\ 1 \leqslant D \leqslant D_T\},$$

we compute our *two-step adaptive* variational distribution $\hat{Q}_{TSA}$ using Algorithm 2, replacing $\mathcal{S}_T$ by $\mathcal{S}_T^{\hat{\delta}}$. This procedure is summarised in Algorithm 3.

Moreover, our thresholding procedure to construct an estimator of the graph in (30) can be theoretically justified. From our theory, we know that $\hat{Q}^{\delta_C}$ concentrates at the rate $\varepsilon_T = \sqrt{\kappa_T}\epsilon_T$. Therefore, if there exists $\epsilon_0 > 0$ such that $\forall l, k$, $\left\|h_{lk}^0\right\|_1 \geqslant \epsilon_0$, then for any $M_T \to \infty$ and sequence of thresholds $(\eta_T)_T$ such that $M_T\varepsilon_T \leqslant \eta_T \leqslant \epsilon_0 - M_T\varepsilon_T$, we can show that

$$\mathbb{P}_0\left[\hat{\delta} = \delta_0\right] \xrightarrow[T\to\infty]{} 1.$$

This comes from the fact that with

$$\begin{aligned}\hat{Q}^{\delta_C}\left(\{f; \exists l, k,\ \hat{\delta}_{lk}(\eta_T) = 1 \text{ and } \delta_{lk}^0 = 0\}\right) &= \hat{Q}^{\delta_C}\left(\{f; \exists l, k,\ \|h_{lk}\|_1 > \eta_T \text{ and } \left\|h_{lk}^0\right\|_1 = 0\}\right) \\ &\leqslant \hat{Q}^{\delta_C}\left(\{f; \exists l, k,\ \left\|h_{lk} - h_{lk}^0\right\|_1 > M_T\varepsilon_T \text{ and } \delta_{lk}^0 = 0\}\right) = o_{\mathbb{P}_0}(1).\end{aligned}$$

Similarly, it holds that

$$\hat{Q}^{\delta_C}\left(\{f; \exists l, k,\ \hat{\delta}_{lk}(\eta_T) = 0 \text{ and } \delta_{lk}^0 = 1\}\right) = \hat{Q}^{\delta_C}\left(\{f; \exists l, k,\ \|h_{lk}\|_1 < \eta_T \leqslant \left\|h_{lk}^0\right\|_1 - M_T\varepsilon_T\}\right) = o_{\mathbb{P}_0}(1).$$

An alternative to setting a threshold η_0 is to select it in a data-driven way after the first step of Algorithm 3. One heuristic is to sort the entries of the matrix $\tilde{S}$ in increasing order of magnitude, to spot a "gap" in this list, and to choose somewhere mid-way in this gap. The latter is visible in our simulation study, as can be seen in the plots in Figure 20.

Algorithm 3 Two-step adaptive mean-field variational inference

Input: $N, \{(\mu_D, \Sigma_D)\}_D, n_{iter}, n_{GQ}, \eta_0$.
Output: $\hat{Q}_{TSA}$
Compute $\hat{Q}^{\delta_C}$ using Algorithm 2 with input set $\mathcal{S}_T^{\delta_C}$.
Compute $\hat{\delta}$ using (30).
Compute $\hat{Q}_{TSA}$ using Algorithm 2 with input set $\mathcal{S}_T^{\hat{\delta}}$.

5 Numerical results

In this section, we perform a simulation study and evaluate (variational) Bayesian methods in the context of nonlinear Hawkes processes. We first test a MCMC method in commonly used nonlinear models (Simulation 1), and then, our adaptive variational algorithms derived for the sigmoid model (Simulations 2, 3 and 4). In each setting, we sample one observation of a Hawkes process with dimension K, link functions $(\phi_k)_k$ and parameter $f_0 = (\nu_0, h_0)$ on $[0, T]$ using the thinning algorithm Adams et al. [2009]. In most settings, the true interaction functions $(h_{lk}^0)_{l,k}$ will be piecewise-constant and we will use the random histogram prior described in Section 4.3. For $D \geqslant 1$, we define

$$\mathcal{H}_{histo}^D = \left\{h = (h_{lk})_{l,k};\ h_{lk}(x) = \sum_{j=1}^{2^D} w_{lk}^j e_j(x),\ x \in [0, A],\ l, k \in [K]\right\}.$$

We report the following set of simulations:

- **Simulation 1: MCMC method in univariate nonlinear Hawkes models.** This experiment aims at evaluating a Metropolis-Hasting sampler (MH) in several nonlinear Hawkes models, with ReLU, sigmoid, and softplus link functions, and in a setting where $h_0 \in \mathcal{H}_{histo}^{D_0}$ and the dimensionality D_0 is known. Since this MCMC sampler is quite computationally expensive to run, we only test the univariate setting $K = 1$ in this simulation.

- **Simulation 2: Comparison of MH sampler, Gibbs sampler, and fixed-dimension variational algorithm in the univariate sigmoid model.** In this simulation, we also consider a univariate setting where $h_0 \in \mathcal{H}_{histo}^{D_0}$ and the dimensionality D_0 is known. We compare two MCMC methods, namely the MH sampler and a Gibbs sampler (Algorithm 4), and the fixed-dimension mean-field variational inference algorithm (Algorithm 1).
- **Simulation 3: Adaptive mean-field variational algorithm in the univariate and bivariate sigmoid models.** In this experiment, we test our fully-adaptive variational algorithm (Algorithm 2) for sigmoid Hawkes processes with $K = 1$ and $K = 2$, in nonparametric settings where the true interaction functions are piecewise-constant functions or continuous.
- **Simulation 4: Two-step variational algorithm for multivariate sigmoid models** In this simulation, we test our two-step adaptive mean-foeld algorithm (Algorithm 3) for sigmoid Hawkes processes with $K = 2, 4, 8, 32$, in sparse settings of the true parameter h_0.

In all simulations, we set $A = 0.1$. Additional details on these experiments are reported in Appendix D.

5.1 Simulation 1: MCMC method in univariate nonlinear Hawkes models

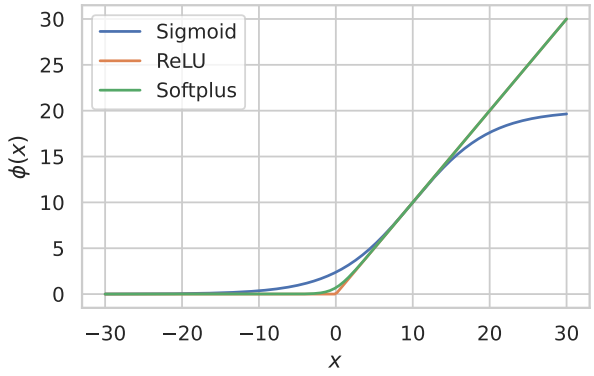


Figure 1: Link functions ϕ considered in Simulation 1, corresponding to our sigmoid, ReLU, and softplus Hawkes models.

In this simulation, we set $K = 1$ and consider a link function ϕ of the form

$$\phi(x) = \theta + \Lambda\psi(\alpha(x - \eta)), \tag{31}$$

where $\xi = (\theta, \Lambda, \alpha, \eta)$ and $\psi : \mathbb{R} \to \mathbb{R}^+$ are known. We consider the following models:

- Sigmoid: $\psi(x) = (1 + e^{-x})^{-1}$ and $\xi = (0.0, 20.0, 0.2, 10.0)$;
- ReLU: $\psi(x) = \max(x, 0)$ and $\xi = (0.001, 1.0, 1.0, 0.0)$;
- Softplus: $\psi(x) = \log(1 + e^x)$ and $\xi = (0.0, 40.0, 0.1, 20.0)$.

The corresponding link functions ϕ are plotted in Figure 1. In all models, we set $\nu_0 = 6$ and $h_0 = h_{11}^0 \in \mathcal{H}_{histo}^{D_0}$ with $D_0 = 2$, and consider three scenarios, namely *Excitation only*, *Mixed effect*, and *Inhibition only*, where h_0 is respectively non-negative, signed and non-positive (see Figure **??** for instance). In this simulation, we assume that D_0 is known and consider a normal prior on $\mathcal{H}_{histo}^{D_0}$ on w_{11}, $w_{11} \sim \mathcal{N}(0, \sigma^2 I)$, and on ν_1, $\nu_1 \sim \mathcal{N}(0, \sigma^2)$, with $\sigma = 5.0$. We set $T = 500$ and report the number of events and excursions (see Lemma A.1 in Appendix A.1 for the definition of this concept), observed in each simulation and model in Table 1. As expected, more events and less excursions are observed in the *Excitation only* scenario than in the *Mixed effect* and *Inhibition only* scenarios.

We run a Metropolis-Hasting sampler implemented via the Python package PyMC4[1] with 4 chains, 40 000 iterations and with a burn-in time of 4000 iterations. The log-likelihood is evaluated using the Gaussian quadrature method Golub and Welsch [1969] for numerical integration, except in the ReLU model and *Excitation only* scenario where the integral is computed exactly. The posterior distribution on $f = (\nu_1, h_{11})$ in the three models and scenarios are plotted in Figure 2, Figure 3, and Figure 4.

[1] https://www.pymc.io/welcome.html

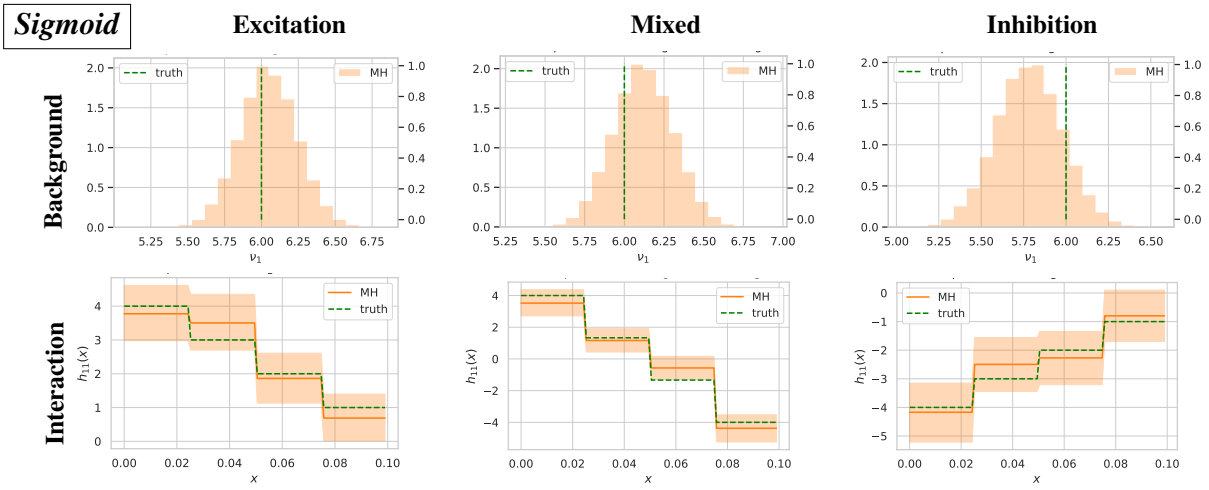


Figure 2: Posterior distribution on $f = (\nu_1, h_{11})$ obtained with the MH sampler in the sigmoid model, in the three scenarios of Simulation 1 ($K = 1$). The three columns correspond to the *Excitation only* (left), *Mixed effect* (center), and *Inhibition only* (right) scenarios. The first row contains the marginal distribution on the background rate ν_1, and the second row represents the posterior mean (solid line) and 95% credible sets (colored areas) on the (self) interaction function h_{11}. The true parameter f_0 is plotted in dotted green line.

We note that in almost all settings, the ground-truth parameter f_0 is included in the 95% credible sets of the posterior distribution, except in the Excitation scenario in the softplus model. Nonetheless, the posterior mean is sometimes biased, in particular in the Excitation scenario, possibly due to the numerical integration errors. One conjecture is that the estimation quality depends on the number of events and the number of excursions, which could explain the differences between the Excitation, Mixed, and Inhibition scenarios. In particular, the credible sets seem consistently smaller for the Mixed scenario, which realisations have more excursions than the Excitation scenario and more events that the Inhibition scenario.

This simulation can be seen as an illustration of the theoretical results of Sulem et al. [2021] for general nonlinear Hawkes models. Moreover, the MH sampler provides a baseline method to compare our variational algorithms in low-dimensional settings, i.e., for $K = 1$ (Simulations 2 and 3) and $K = 2$ (Simulations 3). We note that we also tested a Hamiltonian Monte-Carlo sampler in this simulation, and obtained similar posterior distributions, but in a much larger computational time.

Scenario		Sigmoid	ReLU	Softplus
Excitation only	# events	5250	5352	4953
	# excursions	1558	1436	1373
Mixed effect	# events	3876	3684	3418
	# excursions	1775	1795	1650
Inhibition only	# events	3047	2724	2596
	# excursions	1817	1693	1588

Table 1: Number of events and excursions in the simulated data of Simulation 1. The definition of the concept of excursion in the Hawkes model is recalled in Lemma A.1 in Appendix A.1.

5.2 Simulation 2: Comparison of MH sampler, Gibbs sampler, and fixed-dimension mean-field variational algorithm in the univariate sigmoid model

In this simulation, we consider the sigmoid Hawkes model with $K = 1$ and the three estimation scenarios of Simulation 1, where the dimensionality D_0 is known. We compare the performance of the previous MH sampler, the Gibbs sampler of the latent variable augmentation scheme (Algorithm 4 in Appendix C.3), and our fixed-dimension mean-

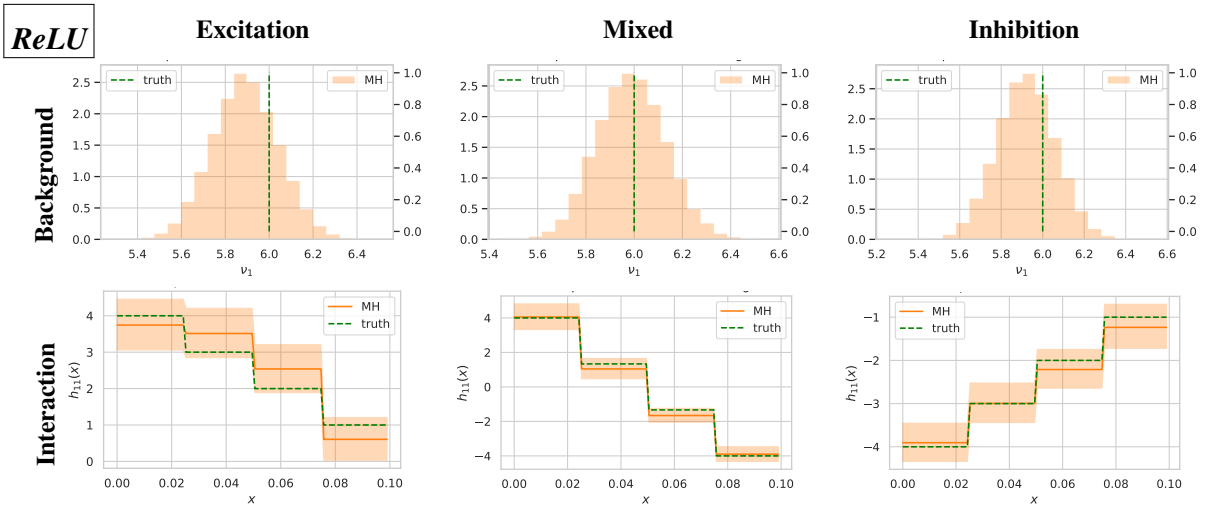


Figure 3: Posterior distribution on $f = (\nu_1, h_{11})$ obtained with the MH sampler in the ReLU model, in the three scenarios of Simulation 1 ($K = 1$). The three columns correspond to the *Excitation only* (left), *Mixed effect* (center), and *Inhibition only* (right) scenarios. The first row contains the marginal distribution on the background rate ν_1, and the second row represents the posterior mean (solid line) and 95% credible sets (colored areas) on the (self) interaction function h_{11}. The true parameter f_0 is plotted in dotted green line.

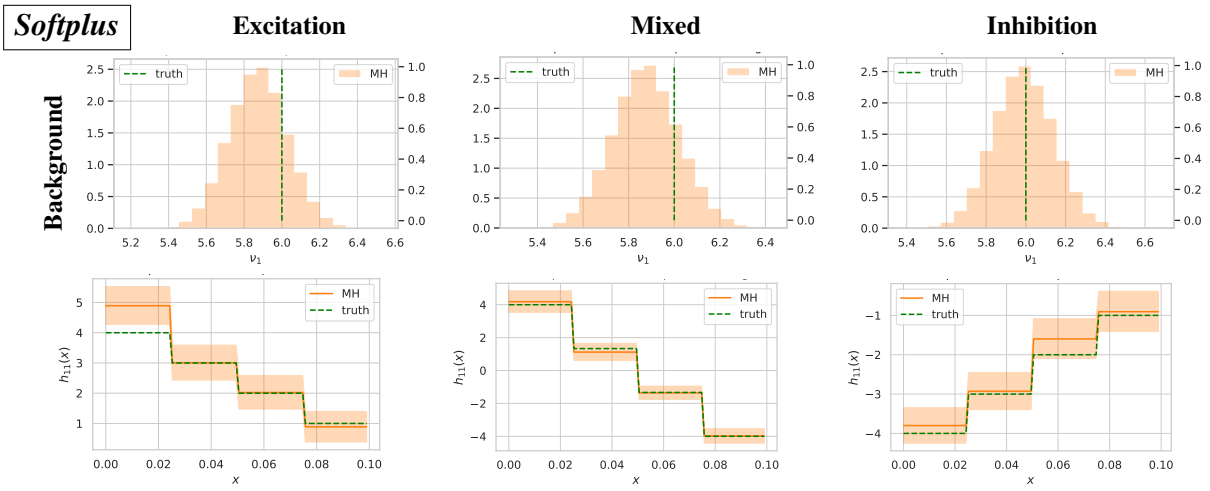


Figure 4: Posterior distribution on $f = (\nu_1, h_{11})$ obtained with the MH sampler in the softplus model, in the three scenarios of Simulation 1 ($K = 1$). The three columns correspond to the *Excitation only* (left), *Mixed effect* (center), and *Inhibition only* (right) scenarios. The first row contains the marginal distribution on the background rate ν_1, and the second row represents the posterior mean (solid line) and 95% credible sets (colored areas) on the (self) interaction function h_{11}. The true parameter f_0 is plotted in dotted green line.

field variational algorithm (Algorithm 1). We run 4 chains for 40 000 iterations for the MH sampler, 3000 iterations of the Gibbs sampler, and 30 iterations of the mean-field variational algorithm.

The (variational) distributions on the parameter $f = (\nu_1, h_{11})$ are plotted in Figure 5. We note that the variational posterior mean is close to the posterior mean, nonetheless, the credible sets of the variational posterior are smaller than the ones of the posterior distribution, which is a common empirical observation in mean-field variational inference. Moreover, in spite of the small number of Gibbs iterations, the Gibbs sampler seems slightly more precise than the other two algorithms; it is about 6 (resp. 40) times longer to run than the MH sampler (resp. our mean-field algorithm), due to the expensive latent variable sampling scheme (see the computational times in Table 2). Besides, the three algorithms seem to be similarly biased, e.g., in the Inhibition scenario. One could therefore test if this bias decreases

with more data observations, i.e., larger T. Finally, we also compare the estimated intensity function on a sub-window in Figure 6 and note that all three methods provide fairly equivalent estimates.

From this simulation, we conclude that, in the univariate and parametric sigmoid Hawkes model, the fixed-dimension mean-field variational algorithm provides a good approximation of the posterior distribution. Moreover, we note that although the Gibbs sampler is slightly better than MH, it is too slow to be applied to multivariate Hawkes processes in practice. Therefore, in the next simulations, we only compare to the posterior distribution computed with the MH sampler.

Scenario	MH	Gibbs	MF-VI
Excitation only	2169	16 092	416
Mixed effect	2181	13 097	338
Inhibition only	2222	9 318	400

Table 2: Computational times (in seconds) of the Gibbs sampler (Algorithm 4), the fixed-dimension mean-field variational (MF-VI) algorithm (Algorithm 1), and the MH sampler in each scenario of Simulation 2 (with $K = 1$). The Gibbs sampler is much slower than the MH sampler, which is also much slower than the MF-VI algorithm.

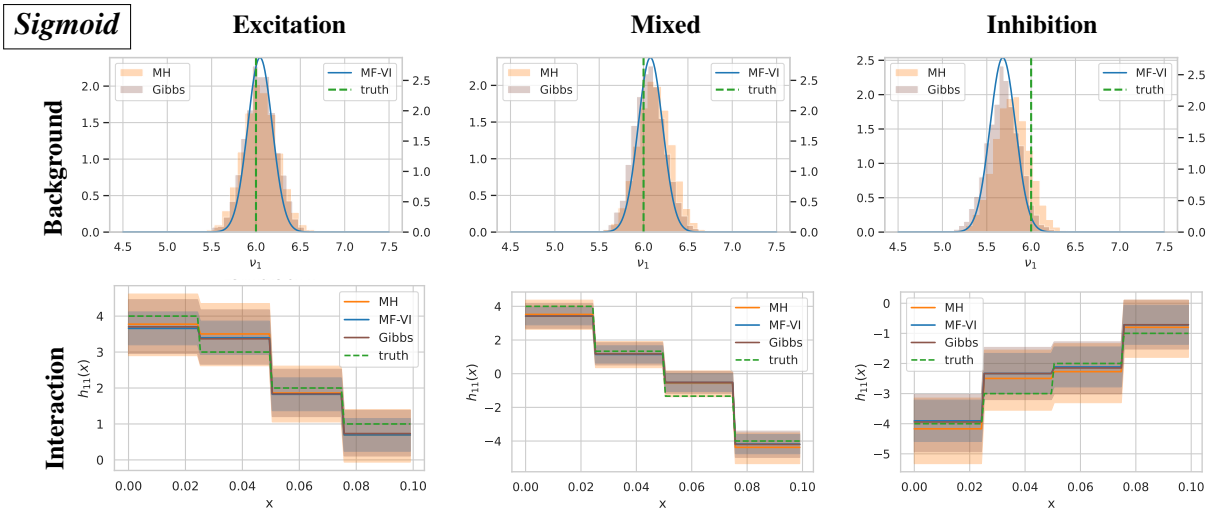
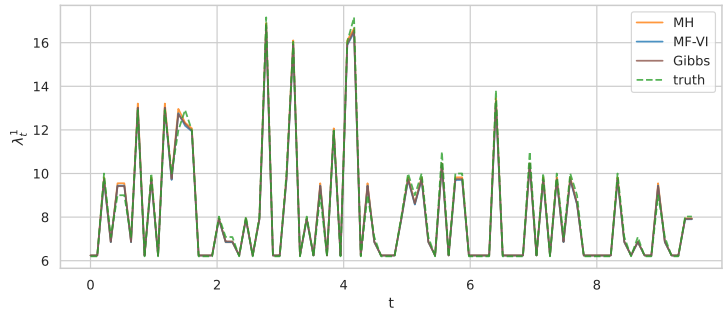


Figure 5: Posterior and variational posterior distributions on $f = (\nu_1, h_{11})$ in the sigmoid model and in the three scenarios of Simulation 2 ($K = 1$), evaluated by the MH sampler, the fixed-dimension mean-field variational (MF-VI) algorithm (Algorithm 1) and the Gibbs sampler (Algorithm 4). The three columns correspond to the *Excitation only* (left), *Mixed effect* (center), and *Inhibition only* (right) scenarios. The first row contains the marginal distribution on the background rate ν_1, and the second row represents the (variational) posterior mean (solid line) and 95% credible sets (colored areas) on the (self) interaction function h_{11}. The true parameter f_0 is plotted in dotted green line.

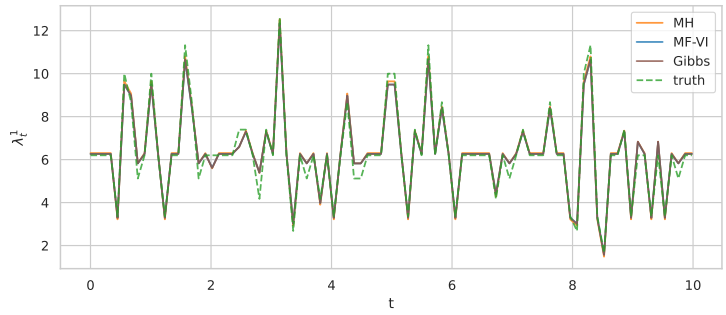
5.3 Simulation 3: Adaptive variational algorithm in the univariate and bivariate sigmoid models

# dimensions	Scenario	FA-MF-VI	MH
$K = 1$	Excitation	2094	3430
	Inhibition	1058	1046
$K = 2$	Excitation	3258	5777
	Inhibition	2616	4679

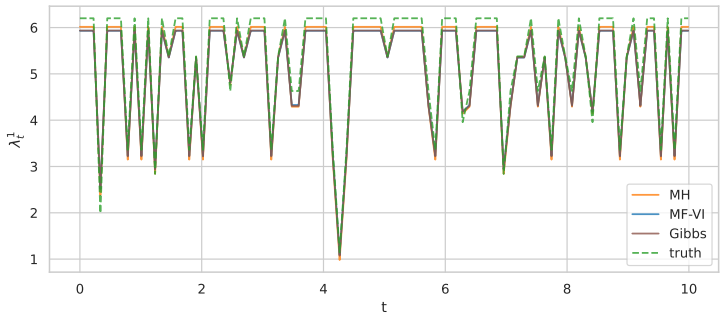
Table 3: Computational times (in seconds) of Algorithm 2 (FA-MF-VI) and MCMC method in the well-specified settings of Simulation 3.



(a) *Excitation only*



(b) *Mixed effect*



(c) *Inhibition only*

Figure 6: Intensity function on a subwindow of the observation window estimated via the variational posterior mean (blue) or via the posterior mean, computed with the MH sampler (orange) or the Gibbs sampler (purple), in each scenario of Simulation 2. The true intensity $\lambda_t^1(f_0)$ is plotted in dotted green line.

In this simulation, we test our fully-adaptive variational Bayes algorithm (Algorithm 2) in the one-dimensional ($K = 1$) and two-dimensional ($K = 2$) sigmoid model. Here, we consider two nonparametric estimation settings:

1. *Well-specified:* $h_0 \in \mathcal{H}_{hist}^{D_0}$ with $D_0 = 2$. In this setting, we compare our variational posterior with the posterior distribution obtained with a non-adaptive MH sampler run with the true $s_0 = (\delta_0, D_0)$;
2. *Mis-specified:* $h_0 \notin \mathcal{H}_{hist}^{D_0}$, and for all l, k, h_{lk}^0 is a continuous function.

Here, $D_0 \geqslant 1$ is unknown and we set $T = 1500$. In the bivariate model, we pick a graph parameter δ_0 with one zero entry (see Figure 12 (a)), i.e., three of the four interaction functions are non-null. We consider an Excitation scenario where h_0 is non-negative and a Self-inhibition scenario where $h_{kk}^0 \leqslant 0$, $k = 1, 2$. We note that the self-inhibition phenomenon is often observed in neuronal spiking data due to their refractory period Bonnet et al. [2021]. In our algorithm, we set a maximum depth $D_1 = 5$ for $K = 1$, so that $|\mathcal{S}| = 7$ and $D_1 = 4$ for $K = 2$, so that $|\mathcal{S}| = 76$.

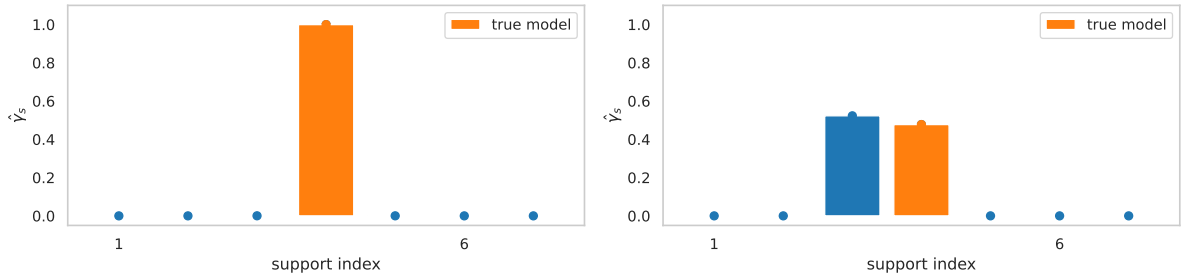
In Figure 7, we plot the marginal probabilities $(\hat{\gamma}_s)_{s \in \mathcal{S}_1}$ in the univariate model and well-specified setting. In the Excitation scenario, the largest marginal probability is on the truth $s_0 = (1, 2)$, i.e., $\hat{\gamma}_{\hat{s}} = \hat{\gamma}_{s_0}$, and all the other marginal

probabilities are negligible. Therefore, in this case, the averaged variational posterior $\hat{Q}_{AV}$ from (26) is essentially equivalent to the mode variational posterior $\hat{Q}_{MV}$ from (29). In the Self-inhibition scenario, we have $\hat{s} = (1, 1)$, but $\hat{\gamma}_{s_0}$ is close to $\hat{\gamma}_{\hat{s}}$, i.e., the marginal probability on s_0 is the second largest. Therefore, in this case the averaged variational posterior is essentially a mixture of the mode variational posterior $\hat{Q}_{MV}$, which is slightly over-regularising in this case, and the variational distribution at the true s_0, $\hat{Q}_{s_0}$. We also plot the estimated intensity based on the mode variational posterior mean in Figure 10 and note that the variational posterior estimates is very close to the true intensity and the non-adaptive MH estimates (see Figure 10).

In Figure 9, we compare the mode variational posterior $\hat{Q}_{MV}$ with the posterior distribution obtained with the non-adaptive MH sampler. We note that in the Excitation scenario, the variational posterior mean is very close to the posterior mean, however its 95% credible bands are significantly smaller. In the Inhibition scenario, in spite of the wrongly selected histogram depth, the estimated parameter is still not too far from the truth. In the mis-specified setting, we note in Figure 8 that the marginal probabilities are also peaked on a single value $\hat{s}$. Moreover, we see on Figure 11 that the true parameter is quite well estimated by $\hat{Q}_{\hat{s}}$. Nonetheless, the 95% credible bands are once again slightly too narrow.

The previous observations can also be made in the two-dimensional model. In the well-specified settings, the largest marginal probabilities are on the true s_0 in both Excitation and Self-inhibition scenarios (see Figure 12 (b) and (c)). Therefore, both the causality structure and the dimensionality are well recovered in this case. We also note from Figures 13 and 14, that the variational posterior mean on the interaction functions approximates well the posterior mean, and thus leads to similar intensity estimates (see Figure 16). In the mis-specified setting, the continuous interaction functions are also quite well estimated, although the under-coverage phenomenon of the credible regions also appears (see Figure 15).

Finally, we note that the computational times of our fully-adaptive variational algorithm is lower than the one of the non-adaptive MH sampler, although here the latter is not adaptive, in particular in the bivariate setting [2], as can be seen in Table 3. This simulation therefore shows that our fully-adaptive variational algorithm enjoys several advantages in Bayesian estimation for Hawkes processes: it can infer the causality structure and provides a good approximation of the posterior mean, and is computationally efficient.



(a) *Excitation* scenario

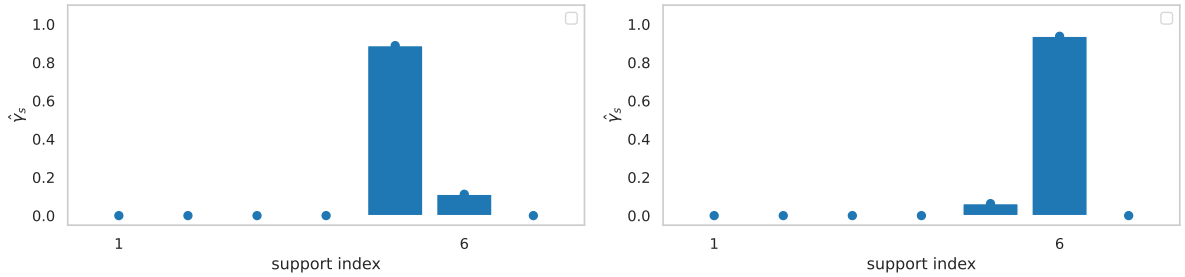
(b) *Inhibition* scenario

Figure 7: Marginal probabilities $(\hat{\gamma}_s)_{s \in \mathcal{S}_1}$ in the adaptive mean-field variational posterior, in the well-specified scenario of Simulation 3 with $K = 1$. The left and right panels correspond to the *Excitation* (resp. *Inhibition*) setting where $h_0 \geqslant 0$ (resp. $h_0 \leqslant 0$). The elements in $\mathcal{S}_1$ are indexed from 1 to 7, and correspond respectively to $s = (0, 0)$, and $s = (1, d)$ with $d = 0, \ldots, 5$. The marginal probability on the true $s_0 = (1, 2)$ is colored in orange.

5.4 Simulation 4: Two-step variational algorithm for multivariate sigmoid models

In this experiment, we test our two-step mean-field variational algorithm (Algorithm 3) in the multivariate sigmoid model with $K = 2, 4, 8, 32$. We note that to the best of our knowledge, the only Bayesian nonparametric method that has currently been tested in high-dimensional Hawkes processes is the Gaussian process model of Malem-Shinitski et al. [2021] in a sigmoid Hawkes model with time-varying background rate. We consider a well-specified setting with $h_0 \in \mathcal{H}^{D_0}_{hist}$ and $D_0 = 1$, and a *sparse* graph parameter δ_0 with $\sum_{l,k} \delta^0_{lk} = 2K - 1$ (see Figure 18). We also design an *Excitation* scenario and a *Self-inhibition* scenario similar to Simulation 3. We only report the results for the former in this section, and the ones for the latter can be found in Appendix D.2. We set $T = 1000$ and report the number of events and excursions in each setting in Table 4. Here, we fix a threshold of $\eta_0 = 0.07$ in Algorithm 3 - this choice will be further discussed below.

[2] We further note that the current implementation of our algorithm has not been yet optimised.



(a) *Excitation* scenario

(b) *Inhibition* scenario

Figure 8: Marginal probabilities $(\hat{\gamma}_s)_{s\in\mathcal{S}_1}$ in the adaptive mean-field variational posterior, in the mis-specified setting of Simulation 3 with $K = 1$. The left and right panels correspond to the (mostly) *Excitation* (resp. *Inhibition*) setting. The elements in $\mathcal{S}_1$ are indexed from 1 to 7, and correspond respectively to $s = (0, 0)$, and $s = (1, d)$ with $D = 0, \dots, 5$.

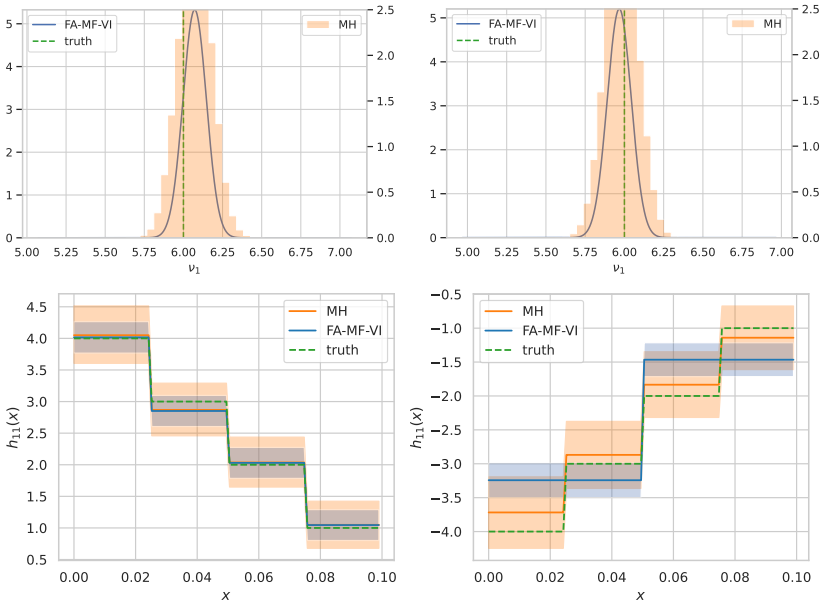
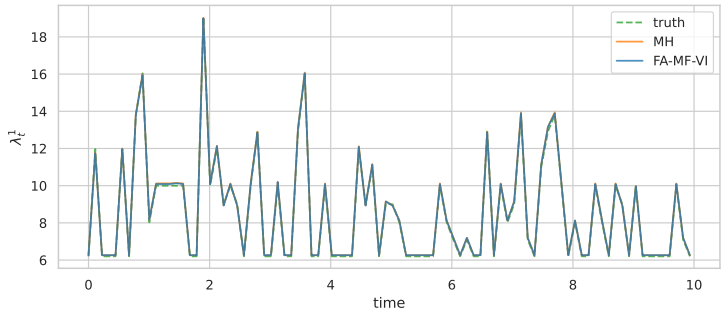


Figure 9: Posterior and mode variational posterior distributions on $f = (\nu_1, h_{11})$ in the univariate sigmoid model and well-specified setting of Simulation 3, evaluated by the MH sampler and the fully-adaptive mean-field variational (FA-MF-VI) algorithm (Algorithm 2). The two columns correspond to the *Excitation* (left) and *Inhibition* (right) settings. The first row contains the marginal distribution on the background rate ν_1, and the second row represents the (variational) posterior mean (solid line) and 95% credible sets (colored areas) on the (self) interaction function h_{11}. The true parameter f_0 is plotted in dotted green line.

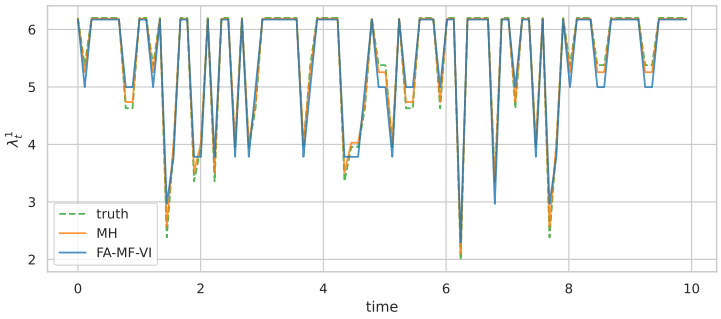
In Table 5, we report the performance of our method, in terms of the L_1-risk of the mode variational posterior on the parameter defined as

$$r_{L_1}(\hat{Q}_{MV}) := \mathbb{E}_{\hat{Q}_{MV}}[\|\nu - \nu_0\|_1] + \sum_{l,k} \mathbb{E}_{\hat{Q}_{MV}}\left[\left\|h_{lk} - h^0_{lk}\right\|_1\right]. \tag{32}$$

We note that the number of terms in the risk grows with K and the number of non-null interaction functions in h and h_0. In every setting, we obtain that $\hat{s} = s_0 = (\delta_0, 1)$, therefore our algorithm is able to recover the true graph δ_0 and dimensionality D_0. Moreover, we note that the risk seems to grow linearly with K, which indicates that the estimation does not deteriorate with larger K. This can be also visually checked in Figure 21, where we plot the variational distribution $\hat{Q}_{\hat{s}}$ on a subset of the parameter f_2 for each K. We note that the 95% credible bands on the background rate ν_2 become larger for larger K, however, this phenomenon does not appear for the interaction functions.



(a) Excitation scenario



(b) Inhibition scenario

Figure 10: Intensity function on a subwindow of the observation window estimated via the variational posterior mean and via the posterior mean computed with the MH sampler, in the well-specified setting of Simulation 3 on $[0, 10]$, using the fully-adaptive mean-field variational (FA-MF-VI) algorithm (Algorithm 2). The true intensity $\lambda_t^1(f_0)$ is plotted in dotted green line.

In Figure 19, we also plot the L_1-errors using $\hat{Q}^{\delta_c}$, i.e., $(\mathbb{E}_{\hat{Q}^{\delta_c}}\left[\left\|h_{lk} - h_{lk}^0\right\|_1\right])_{l,k}$ in the form of a heatmap compared to the true norms. We recall that $\hat{Q}^{\delta_c}$ is mode variational distribution obtained after the first step of Algorithm 3, and is used to estimate the graph for the second step. We note that in all settings, these errors are relatively small, therefore allowing us to select the true graph parameter for the second step. Indeed, in Figure 20, we plot the estimated L_1-norms of the interaction functions using $\hat{Q}^{\delta_c}$, i.e., $(\mathbb{E}_{\hat{Q}^{\delta_c}}\left[\|h_{lk}\|_1\right])_{l,k}$ in increasing order of magnitude and observe a gap between the small and larger "signals". Therefore, with our threshold of $\eta_0 = 0.07$, our algorithm is able to discriminate between the true signals and the noise, but in these settings, the threshold value could have been selected using the "gap" heuristic (see Section 4.4.2). Similar observation can also be made for the Self-inhibition scenario, which results are in Appendix D.2, although the estimation is slightly worse in this scenario. Finally, the computational times of our algorithm seem to scale well with K and the number of events, as can be seen in Figure 17.

This simulation in low and moderately high-dimensional settings therefore shows that our two-step procedure is able to select the causality structure and dimensionality of the process and allows to scale up variational Bayes approaches to larger number of dimensions in sparse settings. Nonetheless, we note that the choice of the threshold and heuristic approaches for this choice need to further explored.

6 Discussion

In this paper, we provided a theoretical study of variational Bayes methods in nonlinear Hawkes processes. We obtained variational concentration rates under easily verifiable conditions on the prior and approximating family, that we validated for estimation set-ups commonly used in practice. Our general theory holds in particular in the sigmoid Hawkes model, for which we also developed adaptive variational mean-field algorithms, that can infer the connectivity graph and the dimensionality of the parameter. Moreover, we demonstrated on simulated data that our most computationally efficient algorithm is able to scale up to high-dimensional processes.

Nonetheless, our theory does not yet cover the high-dimensional setting with $K \to \infty$, which is of practical interest in applications of Hawkes processes in social network analysis and neuroscience. In this limit, previous works have

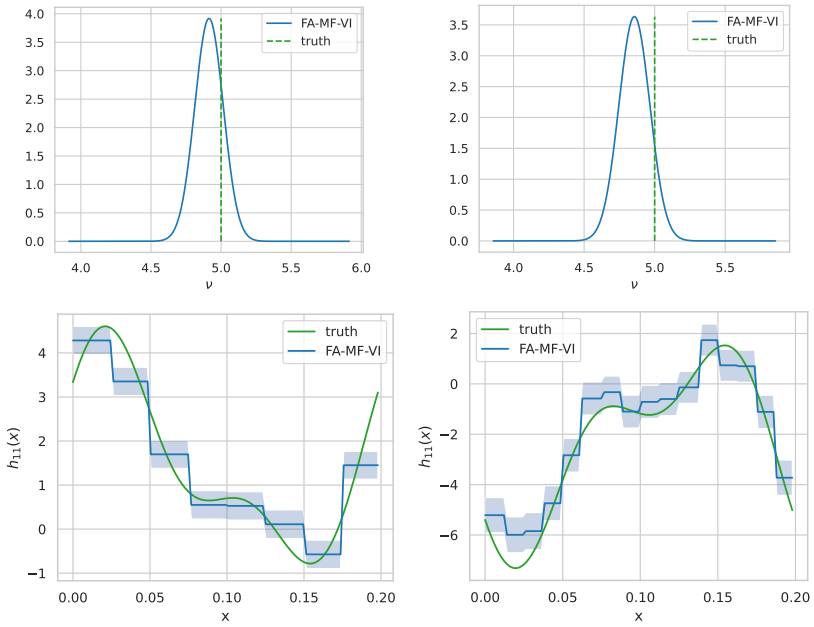
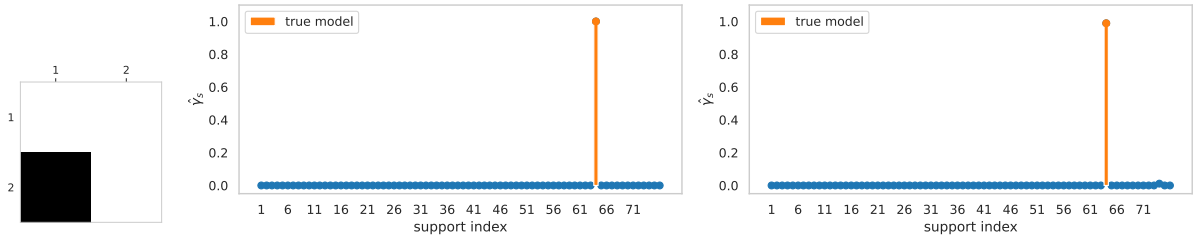


Figure 11: Mode variational posterior distributions on $f = (\nu_1, h_{11})$ in the univariate sigmoid model and mis-specified setting of Simulation 3, evaluated by the fully-adaptive mean-field variational (FA-MF-VI) algorithm (Algorithm 2). The two columns correspond to a (mostly) *Excitation* (left) and a (mostly) *Inhibition* (right) settings. The first row contains the marginal distribution on the background rate ν_1, and the second row represents the variational posterior mean (solid line) and 95% credible sets (colored areas) on the (self) interaction function h_{11}. The true parameter f_0 is plotted in dotted green line.



(a) δ_0 (b) Excitation scenario (c) Self-inhibition scenario

Figure 12: True graph parameter δ_0 (black=0, white=1) (a) and marginal probabilities $(\hat{\gamma}_s)_{s\in\mathcal{S}_2}$ in the adaptive mean-field variational posterior, in the well-specified setting of Simulation 3 with $K = 2$. The *Excitation* scenario (b) corresponds to $h_0 \geqslant 0$, while in the *Self-inhibition* scenario (c) , $h^0_{11}, h^0_{22} \leqslant 0$. The elements in $\mathcal{S}_2$ are indexed from 1 to 76 and the true model corresponds to $s_0 = (\delta_0, 2)$.

considered sparse models Cai et al. [2021], Bacry et al. [2020], Chen et al. [2017] and mean-field settings Pfaffelhuber et al. [2022]. We would therefore be interested in extending our results to these models.

Moreover, our empirical study shows that the credible sets of variational distributions do not always have good coverage, an observation that sometimes also holds for the posterior distribution. Therefore, it is left for future work to study the property of (variational) posterior credible regions, and potentially design post-processing methods of the latter to improve coverage in practice. Additionally, the thresholding approach for estimating the graph in our two-step adaptive variational procedure could be further explored. In particular, designing "good" heuristics to choose the threshold in a data-driven way would be of practical interest.

Finally, it would be of practical interest to develop variational algorithms beyond the sigmoid model, e.g., the ReLU and softplus models. While in the sigmoid model, the conjugacy of the mean-field variational posterior using the data augmentation strategy leads to particularly efficient algorithms, it is unlikely that such convenient forms could be

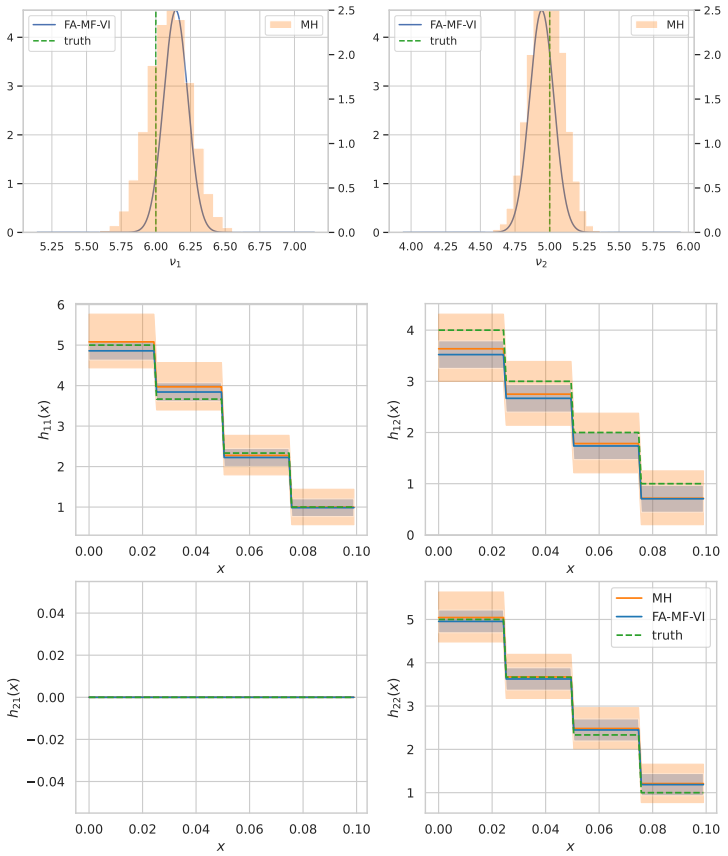


Figure 13: Posterior and mode variational posterior distributions on $f = (\nu, h)$ in the bivariate sigmoid model, well-specified setting, and Excitation setting of Simulation 3, evaluated by the non-adaptive MH sampler and the fully-adaptive mean-field variational (FA-MF-VI) algorithm (Algorithm 2). The first row contains the marginal distribution on the background rates (ν_1, ν_2), and the second and third rows represent the (variational) posterior mean (solid line) and 95% credible sets (colored areas) on the four interaction function $h_{11}, h_{12}, h_{21}, h_{22}$. The true parameter f_0 is plotted in dotted green line.

# dimensions	Scenario	# events	# excursions
2	Excitation	11 719	3295
	Inhibition	8335	3590
4	Excitation	25 509	2362
	Inhibition	17406	2948
8	Excitation	52 390	637
	Inhibition	35530	1043
16	Excitation	108 185	24
	Inhibition	71874	60
32	Excitation	217 320	0
	Inhibition	144741	0

Table 4: Number of observed events and excursions in the multivariate settings of Simulation 4.

obtained for more general models. A potential approach for other models could be to parametrise variational families with normalising flows, as it is for instance done for cut posteriors in Carmona and Nicholls [2022].

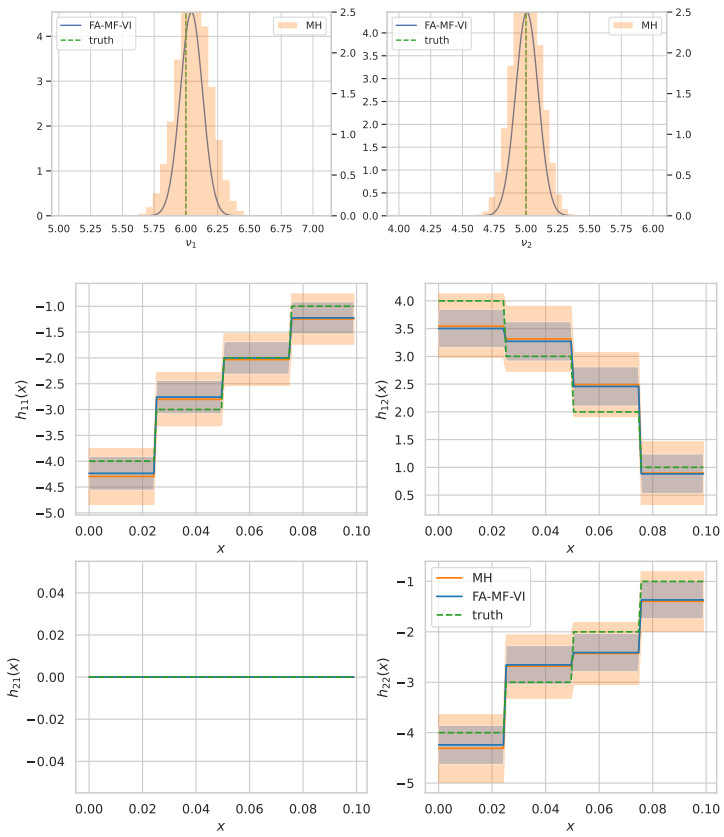


Figure 14: Posterior and mode variational posterior distributions on $f = (\nu, h)$ in the bivariate sigmoid model, well-specified setting, and Self-Inhibition setting of Simulation 3, evaluated by the MH sampler and the fully-adaptive mean-field variational (FA-MF-VI) algorithm (Algorithm 2). The first row correspond two columns correspond to the *Excitation* (left) and *Inhibition* (right) settings. The first row contains the marginal distribution on the background rates (ν_1, ν_2), and the second and third rows represent the (variational) posterior mean (solid line) and 95% credible sets (colored areas) on the four interaction function $h_{11}, h_{12}, h_{21}, h_{22}$. The true parameter f_0 is plotted in dotted green line.

# dimensions	Scenario	$\hat{s} = s_0$	Risk
2	Excitation	Yes	0.408
	Inhibition	Yes	0.277
4	Excitation	Yes	0.697
	Inhibition	Yes	0.767
8	Excitation	Yes	1.672
	Inhibition	Yes	2.312
16	Excitation	Yes	4.692
	Inhibition	Yes	4.688
32	Excitation	Yes	11.066
	Inhibition	Yes	12.074

Table 5: Performance of Algorithm 3 in the multivariate settings of Simulation 4. We report the risk $r_{L_1}(\hat{Q})$ defined in (32) and if the model with largest marginal probability in $\hat{Q}$ corresponds to the true one .

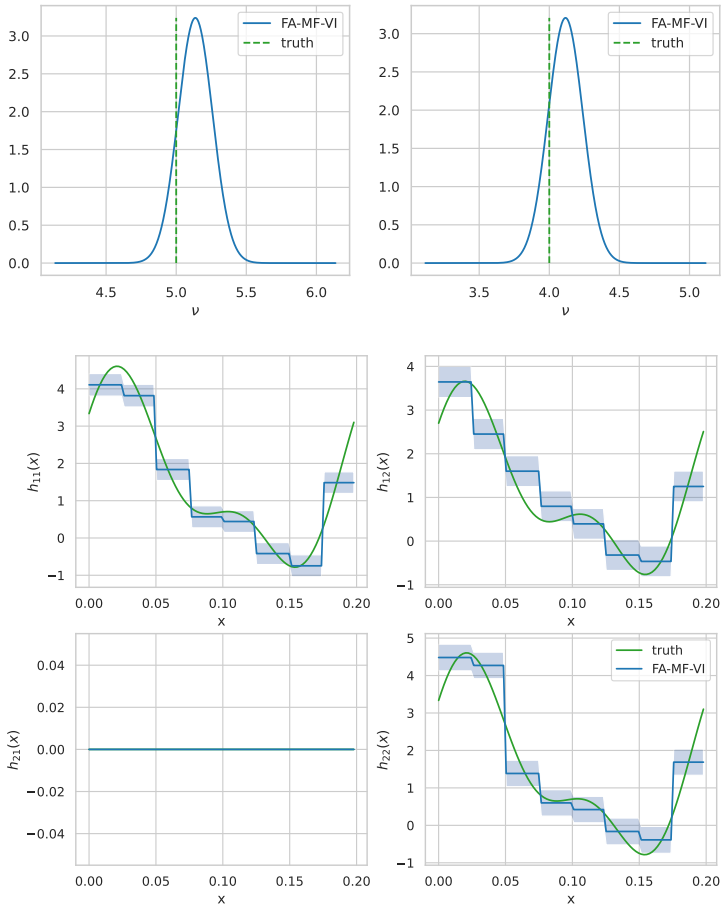
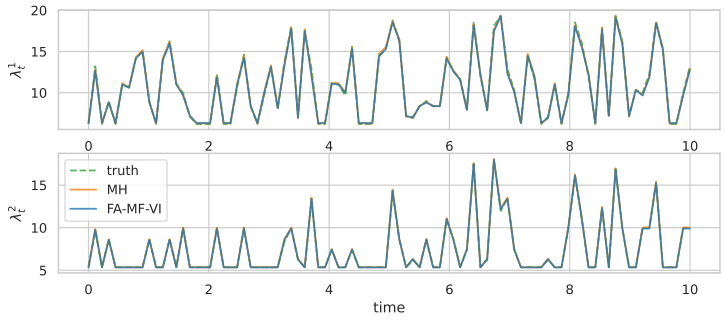
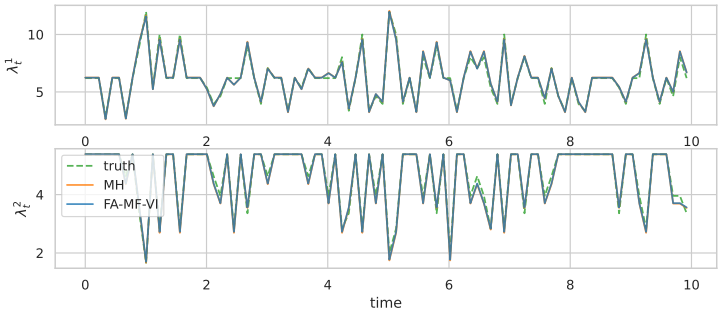


Figure 15: Mode variational posterior distributions on $f = (\nu, h)$ in the bivariate sigmoid model, mis-specified setting, and Excitation setting of Simulation 3, computed with the fully-adaptive mean-field variational (FA-MF-VI) algorithm (Algorithm 2). The first row correspond two columns correspond to the *Excitation* (left) and *Inhibition* (right) settings. The first row contains the marginal distribution on the background rates (ν_1, ν_2), and the second and third rows represent the (variational) posterior mean (solid line) and 95% credible sets (colored areas) on the four interaction function $h_{11}, h_{12}, h_{21}, h_{22}$. The true parameter f_0 is plotted in dotted green line.



(a) Excitation scenario



(b) Self-inhibition scenario

Figure 16: Estimated intensity function based on the (variational) posterior mean, in the well-specified and bivariate setting of Simulation 3 on $[0, 10]$, using the fully-adaptive mean-field variational (FA-MF-VI) algorithm (Algorithm 2). The true intensity $\lambda_t(f_0)$ is plotted in dotted green line.

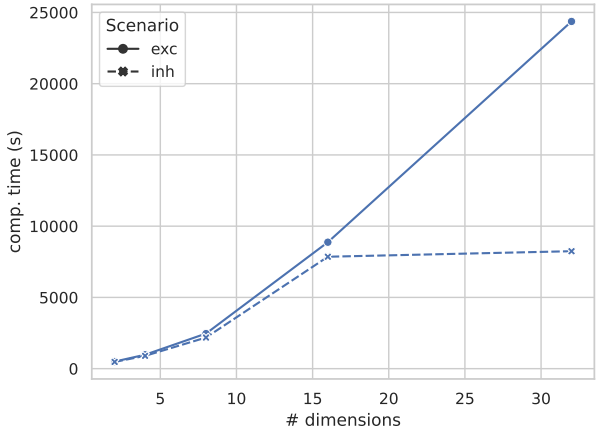
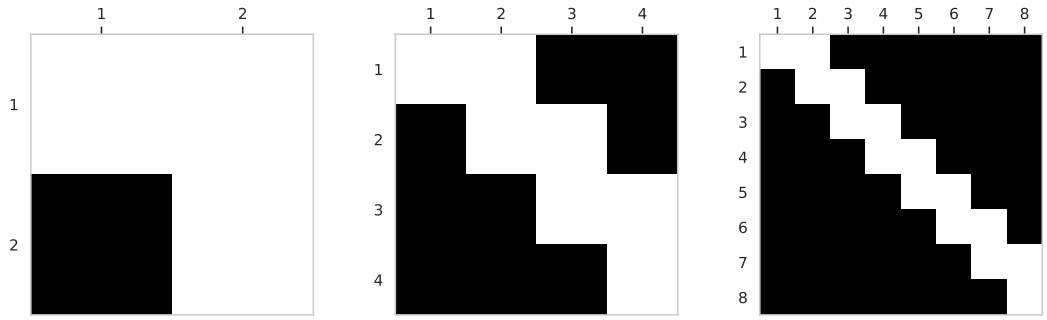
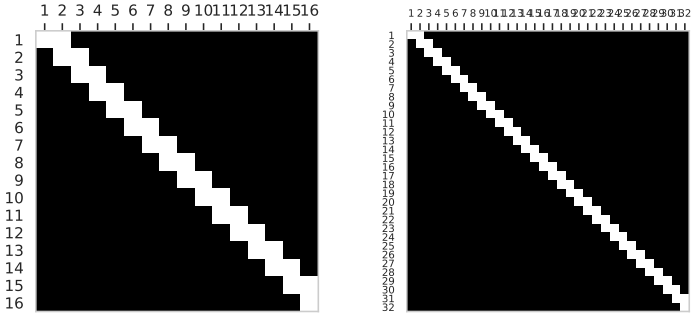


Figure 17: Computational times of our two-step mean-field variational algorithm (Algorithm 3) in the Excitation (exc) and Self-inhibition (inh) scenarios and well-specified setting of Simulation 4, for $K = 2, 4, 8, 16, 32$.



(a) $K = 2$ (b) $K = 4$ (c) $K = 8$



(d) $K = 16$ (e) $K = 32$

Figure 18: True graph parameter δ_0 (black=0, white=1) in the multivariate settings of Simulation 4.

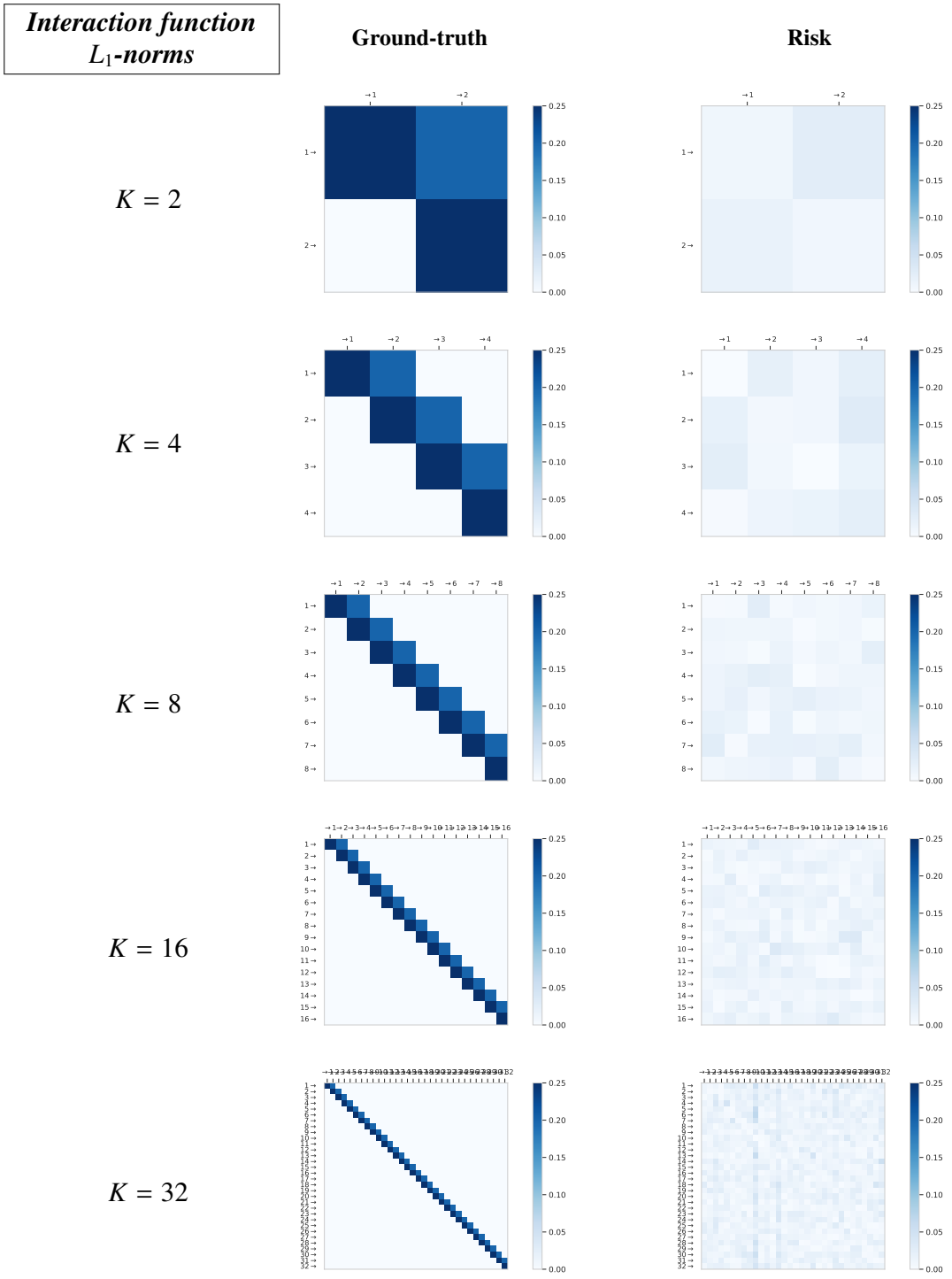
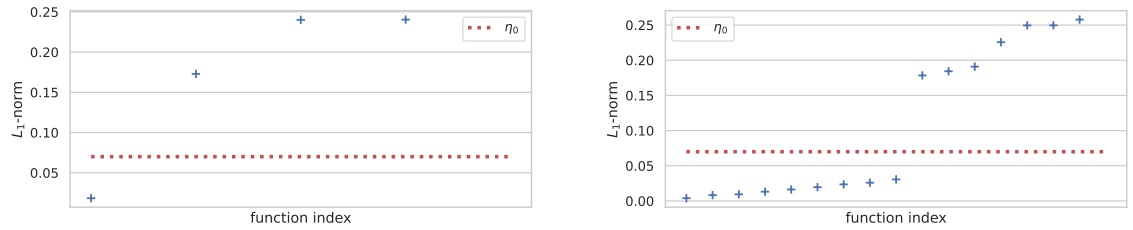
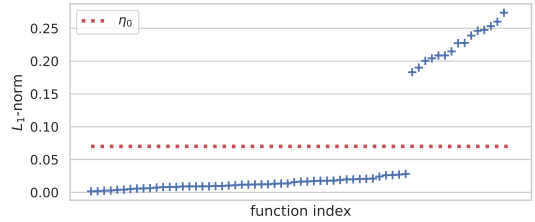


Figure 19: Heatmaps of the L_1-norms of the true parameter h_0, i.e., the entries of the matrix $S_0 = (S^0_{lk})_{l,k} = (\|h^0_{lk}\|_1)_{l,k}$ (left column) and L_1-risk, i.e., $(\mathbb{E}^Q[\|h^0_{lk} - h_{lk}\|_1])_{l,k}$ (right column) after the first step of Algorithm 3, in the Excitation scenario of Simulation 4. The rows correspond to $K = 2, 4, 8, 16, 32$.

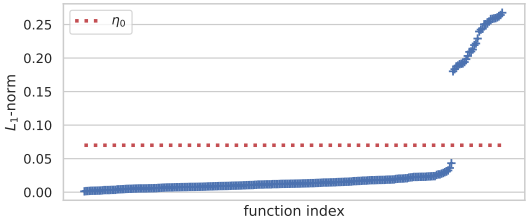


(a) $K = 2$

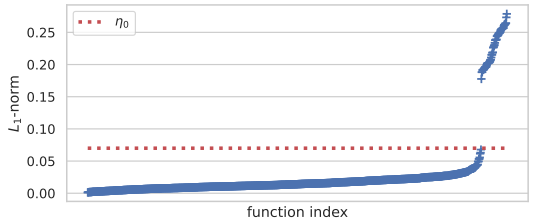
(b) $K = 4$



(c) $K = 8$



(d) $K = 16$



(e) $K = 32$

Figure 20: Estimated L_1-norms after the first step of Algorithm 3, plotted in increasing order, in the Excitation scenario of Simulation 4, for the models with $K = 2, 4, 8, 16, 32$. The threshold in our algorithm $\eta_0 = 0.07$ is plotted in dotted red line.

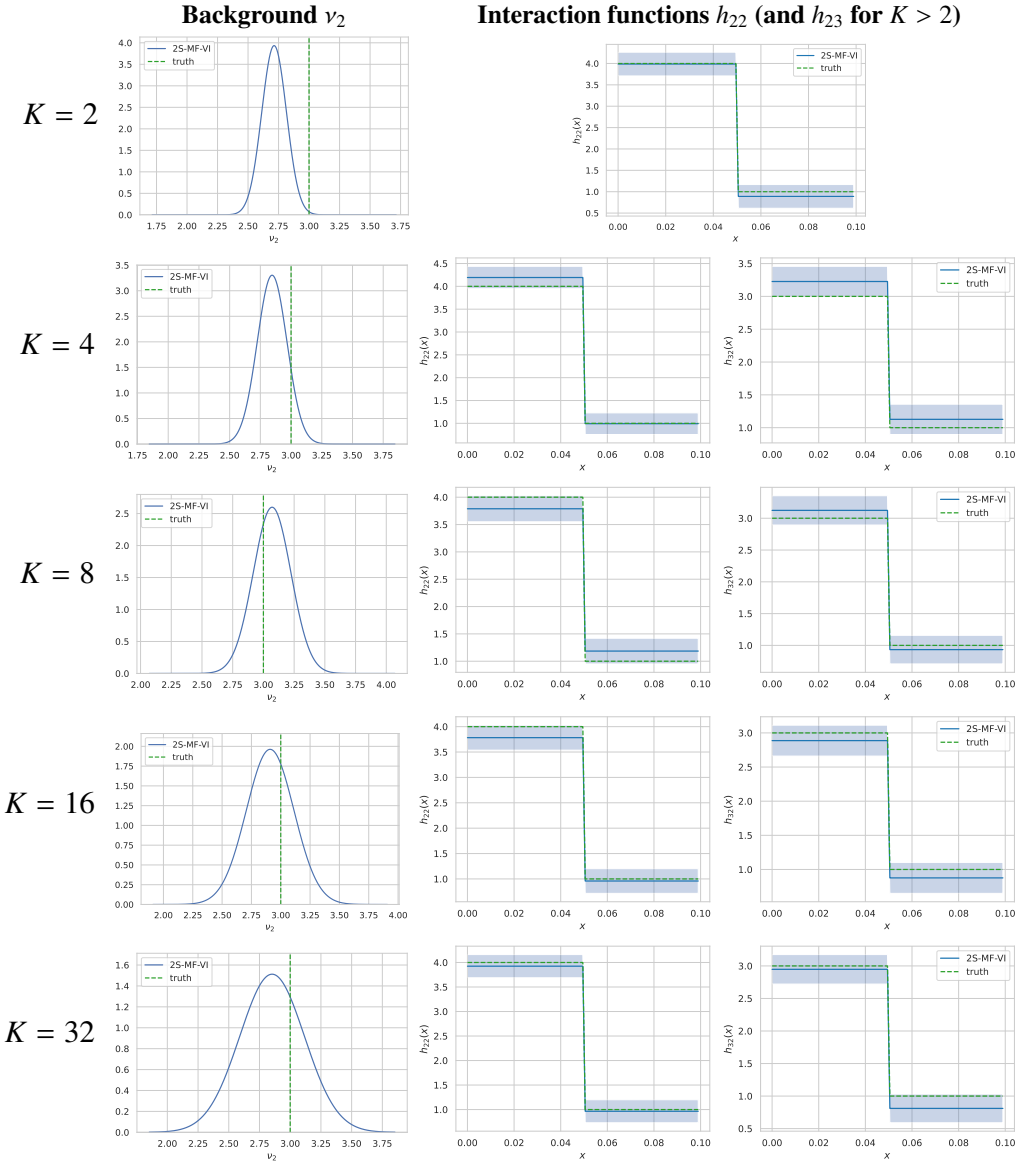


Figure 21: Mode variational posterior distributions on ν_2 (left column) and interaction functions h_{22} and h_{32} (for $K > 2$)(second and third columns) in the Excitation scenario and multivariate sigmoid models of Simulation 4, computed with our two-step mean-field variational (2S-MF-VI) algorithm (Algorithm 3). The different rows correspond to different multivariate settings $K = 2, 4, 8, 16, 32$.

References

Ryan Prescott Adams, Iain Murray, and David J. C. MacKay. Tractable nonparametric bayesian inference in poisson processes with gaussian process intensities. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 9–16, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553376. URL https://doi.org/10.1145/1553374.1553376. 2, 17

Ifigeneia Apostolopoulou, Scott Linderman, Kyle Miller, and Artur Dubrawski. Mutually regressive point processes. *Advances in Neural Information Processing Systems*, 32, 2019. 3

Emmanuel Bacry and Jean-Francois Muzy. Second order statistics characterization of hawkes processes and non-parametric estimation, 2015. 1

Emmanuel Bacry, Martin Bompaire, Stéphane Gaïffas, and Jean-Francois Muzy. Sparse and low-rank multivariate hawkes processes. *Journal of Machine Learning Research*, 21(50):1–32, 2020. 2, 26

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006. 5

Anna Bonnet, Miguel Martinez Herrera, and Maxime Sangnier. Maximum likelihood estimation for hawkes processes with self-excitation or inhibition. *Statistics & Probability Letters*, 179:109214, 2021. 1, 2, 22

Pierre Bremaud and Laurent Massoulie. Stability of nonlinear hawkes processes. *The Annals of Probability*, 1996. 4

Biao Cai, Jingfei Zhang, and Yongtao Guan. Latent network structure learning from high dimensional multivariate point processes, 2021. 2, 26

Chris U. Carmona and Geoff K. Nicholls. Scalable semi-modular inference with variational meta-posteriors, 2022. URL https://arxiv.org/abs/2204.00296. 27

Lisbeth Carstensen, Albin Sandelin, Ole Winther, and Niels R Hansen. Multivariate hawkes process models of the occurrence of regulatory elements. *BMC bioinformatics*, 11(1):1–19, 2010. 2

Shizhe Chen, Ali Shojaie, Eric Shea-Brown, and Daniela Witten. The multivariate hawkes process in high dimensions: Beyond mutual excitation. *arXiv:1707.04928v2*, 2017. 2, 26

Manon Costa, Carl Graham, Laurence Marsalle, and Viet Chi Tran. Renewal in hawkes processes with self-excitation and inhibition. *Advances in Applied Probability*, 52(3):879–915, 2020. doi: 10.1017/apr.2020.19. 2

Daryl J Daley and David Vere-Jones. *An introduction to the theory of point processes: volume II: general theory and structure*. Springer Science & Business Media, 2007. 2, 13

Isabella Deutsch and Gordon J. Ross. Bayesian estimation of multivariate hawkes processes with inhibition and sparsity, 2022. URL https://arxiv.org/abs/2201.05009. 2, 4

Christian Donner and Manfred Opper. Efficient bayesian inference of sigmoidal gaussian cox processes, 2019. 2, 14, 45

Sophie Donnet, Vincent Rivoirard, and Judith Rousseau. Nonparametric Bayesian estimation for multivariate Hawkes processes. volume 48, pages 2698 – 2727. Institute of Mathematical Statistics, 2020. doi: 10.1214/19-AOS1903. URL https://doi.org/10.1214/19-AOS1903. 2, 4, 5, 8, 9

Michael Eichler, Rainer Dahlhaus, and Johannes Dueck. Graphical modeling for multivariate hawkes processes with nonparametric link functions. *Journal of Time Series Analysis*, 38(2):225–242, 2017. 1, 2

Felipe Gerhard, Moritz Deger, and Wilson Truccolo. On the stability and dynamics of stochastic spiking neuron models: Nonlinear hawkes process and point process glms. volume 13, page e1005390, 02 2017. doi: 10.1371/journal.pcbi.1005390. 2

Gene H Golub and John H Welsch. Calculation of gauss quadrature rules. *Mathematics of computation*, 23(106): 221–230, 1969. 15, 18

Niels Richard Hansen, Patricia Reynaud-Bouret, and Vincent Rivoirard. Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, 21(1):83–143, 2015. ISSN 1350-7265. doi: 10.3150/13-BEJ562. URL http://dx.doi.org/10.3150/13-BEJ562. 2, 43

J. F. C. Kingman. *Poisson processes*, volume 3 of *Oxford Studies in Probability*. The Clarendon Press Oxford University Press, New York, 1993. ISBN 0-19-853693-3. Oxford Science Publications. 13

Remi Lemonnier and Nicolas Vayatis. 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, 2014. 1, 2

Xiaofei Lu and Frédéric Abergel. High-dimensional hawkes processes for limit order books: modelling, empirical analysis and numerical calibration. *Quantitative Finance*, 18(2):249–264, 2018. 2

Noa Malem-Shinitski, Cesar Ojeda, and Manfred Opper. Nonlinear hawkes process with gaussian process self effects, 2021. 2, 3, 5, 6, 9, 10, 11, 12, 14, 15, 23

Hongyuan Mei and Jason Eisner. The neural hawkes process: A neurally self-modulating multivariate point process, 2017. 2

G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108, 2011. doi: 10.1198/jasa.2011.ap09546. URL https://doi.org/10.1198/jasa.2011.ap09546. 1

Dennis Nieman, Botond Szabo, and Harry van Zanten. Contraction rates for sparse variational approximations in gaussian process regression, 2021. URL https://arxiv.org/abs/2109.10755. 8, 10, 38

Yosihiko Ogata. Seismicity analysis through point-process modeling: A review. *Seismicity patterns, their statistical significance and physical meaning*, pages 471–507, 1999. 1

Ilsang Ohn and Lizhen Lin. Adaptive variational bayes: Optimality, computation and applications, 2021. 3, 5, 6, 10, 11

Jack Olinde and Martin B. Short. A self-limiting hawkes process: Interpretation, estimation, and use in crime modeling. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 3212–3219, 2020. doi: 10.1109/BigData50022.2020.9378017. 1

Peter Pfaffelhuber, Stefan Rotter, and Jakob Stiefel. Mean-field limits for non-linear hawkes processes with excitation and inhibition. *Stochastic Processes and their Applications*, 2022. 26

Nicholas G. Polson, James G. Scott, and Jesse Windle. Bayesian inference for logistic models using polya-gamma latent variables, 2012. URL https://arxiv.org/abs/1205.0310. 12

Kolyan Ray and Botond Szabó . Variational bayes for high-dimensional linear regression with sparse priors. *Journal of the American Statistical Association*, pages 1–12, jan 2021. doi: 10.1080/01621459.2020.1847121. URL https://doi.org/10.1080. 3, 6, 38, 39

Deborah Sulem, Vincent Rivoirard, and Judith Rousseau. Bayesian estimation of nonlinear hawkes process, 2021. 2, 3, 4, 5, 7, 8, 9, 12, 19, 37, 38, 39, 40, 41, 42, 43, 44

Michalis Titsias and Miguel Lázaro-Gredilla. Spike and slab variational inference for multi-task and multiple kernel learning. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper/2011/file/b495ce63ede0f4efc9eec62cb947c162-Paper.pdf. 6, 10

A. van der Vaart and J. H. van Zanten. Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. volume 37, pages 2655–2675, 2009a. 10

A. W. van der Vaart and J. H. van Zanten. Adaptive bayesian estimation using a gaussian random field with inverse gamma bandwidth. *The Annals of Statistics*, 37(5B), oct 2009b. doi: 10.1214/08-aos678. URL https://doi.org/10.1214. 10

Yichen Wang, Bo Xie, Nan Du, and Le Song. Isotonic hawkes processes. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 2226–2234. JMLR.org, 2016. 2

Fengshuo Zhang and Chao Gao. Convergence rates of variational posterior distributions. *arXiv: Statistics Theory*, 2017. 3, 5, 6, 8, 10, 11

Rui Zhang, Christian Walder, and Marian-Andrei Rizoiu. Variational inference for sparse gaussian process modulated hawkes process. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):6803–6810, Apr 2020. ISSN 2159-5399. doi: 10.1609/aaai.v34i04.6160. URL http://dx.doi.org/10.1609/aaai.v34i04.6160. 2, 9

Feng Zhou, Zhidong Li, Xuhui Fan, Yang Wang, Arcot Sowmya, and Fang Chen. Efficient em-variational inference for hawkes process, 2019. 2

Feng Zhou, Zhidong Li, Xuhui Fan, Yang Wang, Arcot Sowmya, and Fang Chen. Efficient inference for nonparametric hawkes processes using auxiliary latent variables. *Journal of Machine Learning Research*, 21(241):1–31, 2020. URL http://jmlr.org/papers/v21/19-930.html. 2, 9

Feng Zhou, Quyu Kong, Yixuan Zhang, Cheng Feng, and Jun Zhu. Nonlinear hawkes processes in time-varying system, 2021a. 2, 3, 5, 6, 9, 11, 12, 14, 15

Feng Zhou, Yixuan Zhang, and Jun Zhu. Efficient inference of flexible interaction in spiking-neuron networks, 2021b. 2, 3, 9

Feng Zhou, Quyu Kong, Zhijie Deng, Jichao Kan, Yixuan Zhang, Cheng Feng, and Jun Zhu. Efficient inference for dynamic flexible interactions of neural populations. *Journal of Machine Learning Research*, 23(211):1–49, 2022. URL http://jmlr.org/papers/v23/21-1273.html. 11

A Proofs

In this section, we provide the proofs of our main theoretical results, Theorem 3.2 and Proposition 4.5. We first recall a set of useful lemmas from Sulem et al. [2021].

A.1 Technical lemmas

In the first lemma, we recall the definition of excursions from Sulem et al. [2021], for stationary nonlinear Hawkes processes verifying condition (C1) or (C2) . Then, Lemma A.2, corresponding to Lemma A.1 in Sulem et al. [2021], provides a control on the main event $\tilde{\Omega}_T$ considered in the proof of Theorem 3.2. Finally, Lemma A.3 (Lemma A.4 in Sulem et al. [2021]) is a technical lemma for proving posterior concentration in Hawkes processes.

We also introduce the following notation. For any excursion index $j \in [J_T - 1]$, we denote $(U_j^{(1)}, U_j^{(2)})$ the times of the first two events after the j-th renewal time τ_j, and $\xi_j := U_j^{(2)}$ if $U_j^{(2)} \in [\tau_j, \tau_{j+1})$ and $\xi_j := \tau_{j+1}$ otherwise.

Lemma A.1 (Lemma 5.1 in Sulem et al. [2021])**.** *Let N be a Hawkes process with monotone non-decreasing and Lipschitz link functions $\phi = (\phi_k)_k$ and parameter $f = (\nu, h)$ such that (ϕ, f) verify **(C1)** or **(C2)**. Then the point process measure $X_t(.)$ defined as*

$$X_t(.) = N|_{(t-A,t]}, \tag{33}$$

is a strong Markov process with positive recurrent state $\emptyset$. Let $\{\tau_j\}_{j\geqslant 0}$ be the sequence of random times defined as

$$\tau_j = \begin{cases} 0 & \text{if } j = 0; \\ \inf\left\{t > \tau_{j-1};\ X_{t^-} \neq \emptyset,\ X_t = \emptyset\right\} = \inf\left\{t > \tau_{j-1};\ N|_{[t-A,t)} \neq \emptyset,\ N|_{(t-A,t]} = \emptyset\right\} & \text{if } j \geqslant 1. \end{cases}$$

Then, $\{\tau_j\}_{j\geqslant 0}$ are stopping times for the process N. For $T > 0$, we also define

$$J_T = \max\{j \geqslant 0;\ \tau_j \leqslant T\}. \tag{34}$$

The intervals $\{[\tau_j, \tau_{j+1})\}_{j=0}^{J_T-1} \cup [\tau_{J_T}, T]$ form a partition of $[0, T]$. The point process measures $(N|_{[\tau_j,\tau_{j+1})})_{1\leqslant j\leqslant J_T-1}$ are i.i.d. and independent of $N|_{[0,\tau_1)}$ and $N|_{[\tau_{J_T},T]}$; they are called excursions *and the stopping times $\{\tau_j\}_{j\geqslant 1}$ are called* regenerative *or* renewal *times.*

Lemma A.2 (Lemma A.1 in Sulem et al. [2021])**.** *Let $Q > 0$. We consider $\tilde{\Omega}_T$ defined in Section A.2. For any $\beta > 0$, we can choose C_β and c_β in the definition of $\tilde{\Omega}_T$ such that $\mathbb{P}_0[\tilde{\Omega}_T^c] \leqslant T^{-\beta}$. Moreover, for any $1 \leqslant q \leqslant Q$, $\mathbb{E}_0\left[\mathbb{1}_{\tilde{\Omega}_T^c} \max_l \sup_{t\in[0,T]} \left(N^l[t-A,t)\right)^q\right] \leqslant 2T^{-\beta/2}$.*

Lemma A.3 (Lemma 1.4 in Sulem et al. [2021])**.** *For any $f \in \mathcal{F}_T$ and $l \in [K]$, let*

$$Z_{1l} = \int_{\tau_1}^{\xi_1} |\lambda_t^l(f) - \lambda_t^l(f_0)| dt,$$

Under the assumptions of Theorem 3.2, for $M_T \to \infty$ such that $M_T > M\sqrt{\kappa_T}$ with $M > 0$ and for any $f \in \mathcal{F}_T$ such that $\|r - r_0\|_1 \leqslant \max(\|r_0\|_1, \tilde{C})$ with $\tilde{C} > 0$, there exists $l \in [K]$ such that on $\tilde{\Omega}_T$,

$$\mathbb{E}_f[Z_{1l}] \geqslant C(f_0)(\left\|r_f - r_0\right\|_1 + \|h - h_0\|_1),$$

with $C(f_0) > 0$ a constant that depends only on f_0 and $(\phi_k)_k$.

A.2 Proof of Theorem 3.2

We recall that in this result, we consider a general Hawkes model with known link functions $(\phi_k)_k$. Let $r_0 = (r_1^0, \dots, r_K^0)$ with $r_k^0 = \phi_k(\nu_k^0)$. With $C_\beta, c_\beta > 0$, we first define $\tilde{\Omega}_T \in \mathcal{G}_T$ as

$$\begin{aligned}
\tilde{\Omega}_T &= \Omega_N \cap \Omega_J \cap \Omega_U, \\
\Omega_N &= \left\{ \max_{k\in[K]} \sup_{t\in[0,T]} N^k[t-A,t) \leqslant C_\beta \log T \right\} \cap \left\{ \sum_{k=1}^K \left| \frac{N^k[-A,T]}{T} - \mu_k^0 \right| \leqslant \delta_T \right\}, \\
\Omega_J &= \{J_T \in \mathcal{J}_T\}, \quad \Omega_U = \left\{ \sum_{j=1}^{J_T-1} (U_j^{(1)} - \tau_j) \geqslant \frac{T}{\mathbb{E}_0[\Delta\tau_1]\|r_0\|_1} \left(1 - 2c_\beta \sqrt{\frac{\log T}{T}}\right) \right\}, \\
\mathcal{J}_T &= \left\{ J \in \mathbb{Z}_{\geqslant 0};\ \left| \frac{J-1}{T} - \frac{1}{\mathbb{E}_0[\Delta\tau_1]} \right| \leqslant c_\beta \sqrt{\frac{\log T}{T}} \right\},
\end{aligned}$$

with J_T the number of excursions as defined in (34), $\mu_k^0 := \mathbb{E}_0\left[\lambda_t^k(f_0)\right], \forall k$, $\delta_T = \delta_0 \sqrt{\frac{\log T}{T}}$, $\delta_0 > 0$ and $\{U_j^{(1)}\}_{j=1,\dots,J_T-1}$ denoting the first events of each excursion (see Lemma A.1 for a precise definition). Secondly, we define $A'_T \in \mathcal{G}_T$ as

$$A'_T = \left\{ \int e^{L_T(f)-L_T(f_0)} d\widetilde{\Pi}(f) > e^{-C_1 T\varepsilon_T^2} \right\}, \quad \widetilde{\Pi}(B) = \frac{\Pi(B \cap K_T)}{\Pi(K_T)}, \quad K_T \subset \mathcal{F},$$

with $C_1 > 0$ and ε_T, M_T positive sequences such that $T\varepsilon_T^2 \to \infty$ and $M_T \to \infty$. From Lemma A.2, we have that $\mathbb{P}_0\left[\tilde{\Omega}_T^c\right] = o(1)$. Thus, with $A_T = \tilde{\Omega}_T \cap A'_T$, $K_T = B_\infty(\epsilon_T)$, and $\varepsilon_T = \sqrt{\kappa_T}\epsilon_T$, we can obtain that

$$\begin{aligned}
\mathbb{P}_0\left[A_T^c\right] &\leqslant \mathbb{P}_0\left[\tilde{\Omega}_T^c\right] + \mathbb{P}_0\left[A_T'^c \cap \tilde{\Omega}_T\right] = o(1) + \mathbb{P}_0\left[\left\{ \int_{K_T} e^{L_T(f)-L_T(f_0)} d\Pi(f) \leqslant \Pi(K_T) e^{-C_1 T\varepsilon_T^2} \right\} \cap \tilde{\Omega}_T\right] \\
&\leqslant o(1) + \mathbb{P}_0\left[\left\{D_T \leqslant \Pi(K_T) e^{-C_1 T\varepsilon_T^2}\right\} \cap \tilde{\Omega}_T\right] = o(1),
\end{aligned}$$

with $C_1 > 1$, using (A0), i.e., $\Pi(K_T) \geqslant e^{-c_1 T\varepsilon_T^2}$, and the following intermediate result from the proof of Theorem 3.2 in Sulem et al. [2021]

$$\mathbb{P}_0\left[\left\{D_T \leqslant \Pi(B_\infty(\epsilon_T)) e^{-\kappa_T T\varepsilon_T^2}\right\} \cap \tilde{\Omega}_T\right] = o(1).$$

Therefore, we can conclude that

$$\mathbb{P}_0\left[A_T\right] \xrightarrow[T\to\infty]{} 1.$$

We now define the stochastic distance $\tilde{d}_{1T}$ and stochastic neighborhoods around f_0 as

$$\begin{aligned}
\tilde{d}_{1T}(f,f') &= \frac{1}{T} \sum_{k=1}^K \int_0^T \mathbb{1}_{A_2(T)}(t) |\lambda_t^k(f) - \lambda_t^k(f')| dt, \quad A_2(T) = \bigcup_{j=1}^{J_T-1} [\tau_j, \xi_j] \\
A_{d_1}(\varepsilon) &= \left\{ f \in \mathcal{F};\ \tilde{d}_{1T}(f,f_0) \leqslant \varepsilon \right\}, \quad \varepsilon > 0,
\end{aligned} \tag{35}$$

where for each $j \in [J_T]$, $U_j^{(2)}$ is the first event after $U_j^{(1)}$, and $\xi_j := U_j^{(2)}$ if $U_j^{(2)} \in [\tau_j, \tau_{j+1})$ and $\xi_j := \tau_{j+1}$ otherwise. Let η_T be a positive sequence and $\hat{Q}$ be the variational posterior as defined in (5). Using Markov's inequality, we have

$$\mathbb{E}_0\left[\hat{Q}(A_{d_1}(\eta_T)^c)\right] \leqslant \mathbb{P}_0\left[A_T^c\right] + \mathbb{E}_0\left[\hat{Q}(A_{d_1}(\eta_T)^c)\mathbb{1}_{A_T}\right]. \tag{36}$$

We first bound the second term on the RHS of (36) using the following technical lemma, which is an adaptation of Theorem 5 of Ray and Szabó [2021] and Lemma 13 in Nieman et al. [2021].

Lemma A.4. *Let $B_T \subset \mathcal{F}$, $A_T \in \mathcal{G}_T$, and Q be a distribution on $\mathcal{F}$. If there exist $C, u_T > 0$ such that*

$$\mathbb{E}_0\left[\Pi(B_T|N)\mathbb{1}_{A_T}\right] \leqslant Ce^{-u_T}, \tag{37}$$

then, we have that

$$\mathbb{E}_0\left[Q(B_T)\mathbb{1}_{A_T}\right] \leqslant \frac{2}{u_T}\left(\mathbb{E}_0\left[KL(Q\|\Pi(.|N))\mathbb{1}_{A_T}\right] + Ce^{-u_T/2}\right).$$

Proof. We follow the proof of Ray and Szabó [2021] and use the fact that, for any $g : \mathcal{F} \to \mathbb{R}$ such that $\int_{\mathcal{F}} e^{g(f)} d\Pi(f|N) < +\infty$, it holds true that

$$\int_{\mathcal{F}} g(f) dQ \leqslant KL(Q||\Pi(.|N)) + \log \int_{\mathcal{F}} e^{g(f)} \Pi(f|N).$$

Applying the latter inequality with $g = \frac{1}{2} u_T \mathbb{1}_{B_T}$, we obtain

$$\begin{aligned}\frac{1}{2} u_T Q(B_T) &\leqslant KL(Q||\Pi(.|N)) + \log(1 + e^{\frac{1}{2}u_T} \Pi(B_T|N)) \\ &\leqslant KL(Q||\Pi(.|N)) + e^{\frac{1}{2}u_T} \Pi(B_T|N).\end{aligned}$$

Then, multipying both sides of the previous inequality by $\mathbb{1}_A$ and taking expectation wrt to $\mathbb{P}_0$, using (37), we finally obtain

$$\frac{1}{2} u_T \mathbb{E}_0 \left[Q(B_T) \mathbb{1}_{A_T} \right] \leqslant \mathbb{E}_0 \left[KL(Q||\Pi(.|N)) \mathbb{1}_{A_T} \right] + C e^{-\frac{1}{2}u_T}.$$

□

We thus apply Lemma A.4 with $B_T = A_{d_1}(\eta_T)^c$, $\eta_T = M'_T \varepsilon_T$, $Q = \hat{Q}$, and $u_T = M_T T \varepsilon_T^2$ with $M'_T \to \infty$. We first check that (37) holds, i.e., we show that there exists $C, M_T, M'_T > 0$ such that

$$\mathbb{E}_0 \left[\mathbb{1}_{A_T} \Pi[\tilde{d}_{1T}(f, f_0) > M'_T \varepsilon_T | N] \right] \leqslant C \exp(-M_T T \varepsilon_T^2). \tag{38}$$

For any test ϕ, we have the following decomposition

$$\mathbb{E}_0 \left[\mathbb{1}_{A_T} \Pi[\tilde{d}_{1T}(f, f_0) > M'_T \varepsilon_T | N] \right] \leqslant \underbrace{\mathbb{E}_0 \left[\phi \mathbb{1}_{A_T} \right]}_{(I)} + \underbrace{\mathbb{E}_0 \left[(1-\phi) \mathbb{1}_{A_T} \Pi[A_{d_1}(M'_T \varepsilon_T)^c | N] \right]}_{(II)}.$$

Note that we have

$$\begin{aligned}(II) = \mathbb{E}_0 \left[(1-\phi) \mathbb{1}_{A_T} \Pi[A_{d_1}(M'_T \varepsilon_T)^c | N] \right] &= \mathbb{E}_0 \left[\int_{A_{d_1}(M\varepsilon_T)^c} \mathbb{1}_{A_T} (1-\phi) \frac{e^{L_T(f) - L_T(f_0)}}{D_T} d\Pi(f) \right] \\ &\leqslant \Pi(K_T) e^{C_1 T \varepsilon_T^2} \mathbb{E}_0 \left[\sup_{f \in \mathcal{F}_T} \mathbb{E}_f \left[\mathbb{1}_{A_{d_1}(M\varepsilon_T)^c} \mathbb{1}_{A_T} (1-\phi) | \mathcal{G}_0 \right] \right],\end{aligned} \tag{39}$$

since on $A_T, D_T \geqslant \Pi(K_T) e^{C_1 T \varepsilon_T^2}$. Using the proof of Theorem 5.5 in Sulem et al. [2021], we can directly obtain that for T large enough, there exist $x_1, M, M' > 0$ such that

$$\begin{aligned}(I) &\leqslant 2(2K+1) e^{-x_1 M'^2 T \varepsilon_T^2} \\ (II) &\leqslant 2(2K+1) e^{-x_1 M'^2 T \varepsilon_T^2 / 2},\end{aligned}$$

which imply that

$$\mathbb{E}_0 \left[\mathbb{1}_{A_T} \Pi[\tilde{d}_{1T}(f, f_0) > M'_T \varepsilon_T | N] \right] \leqslant 4(2K+1) e^{-x_1 M_T'^2 T \varepsilon_T^2 / 2},$$

and (38) with $M_T = x_1 M_T'^2 / 2$ and $C = 2(2K+1)$. Applying Lemma A.4 thus leads to

$$\mathbb{E}_0 \left[\hat{Q}(A_{d_1}(\eta_T)^c) \mathbb{1}_{A_T} \right] \leqslant 2 \frac{KL(\hat{Q}||\Pi(.|N)) + C e^{-M_T T \varepsilon_T^2 / 2}}{M_T T \varepsilon_T^2} \leqslant 2C e^{-M_T T \varepsilon_T^2 / 2} + 2 \frac{KL(\hat{Q}||\Pi(.|N))}{M_T T \varepsilon_T^2}.$$

Moreover, from (A2) and Remark 3.4, it holds that $KL(\hat{Q}||\Pi(.|N)) = O(T \varepsilon_T^2)$, therefore we obtain the following intermediate result

$$\mathbb{E}_0 \left[\hat{Q}(A_{d_1}(\eta_T)^c) \right] = o(1).$$

Now, with $M_T > M'_T$, we note that

$$\mathbb{E}_0 \left[\hat{Q}(\|f - f_0\|_1 > M_T \varepsilon_T) \right] = \mathbb{E}_0 \left[\hat{Q}(\tilde{d}_{1T}(f, f_0) > M'_T \varepsilon_T) \right] + \mathbb{E}_0 \left[\hat{Q}(\|f - f_0\|_1 > M_T \varepsilon_T, \tilde{d}_{1T}(f, f_0) < M'_T \varepsilon_T) \mathbb{1}_{A_T} \right] + \mathbb{P}_0[A_T^c].$$

Therefore, it remains to show that

$$\mathbb{E}_0\left[\hat{Q}(\|f-f_0\|_1 > M_T\varepsilon_T, \tilde{d}_{1T}(f,f_0) < M'_T\varepsilon_T)\mathbb{1}_{A_T}\right] = \mathbb{E}_0\left[\hat{Q}(A_{L_1}(M_T\epsilon_T)^c \cap A_{d_1}(M'_T\varepsilon_T))\mathbb{1}_{A_T}\right] = o(1).$$

For this, we apply again Lemma A.4 with $B_T = A_{L_1}(M_T\varepsilon_T)^c \cap A_{d_1}(M'_T\varepsilon_T)$ and $u_T = TM_T^2\varepsilon_T^2$. We have

$$\mathbb{E}_0\left[\mathbb{1}_{A_T}\Pi(A_{L_1}(M_T\varepsilon_T)^c \cap A_{d_1}(M'_T\varepsilon_T)|N)\right] \leqslant \Pi(K_T)e^{C_1T\varepsilon_T^2}\mathbb{E}_0\left[\mathbb{E}_f\left[\mathbb{1}_{A_T}\mathbb{1}_{A_{L_1}(M_T\varepsilon_T)^c\cap A_{d_1}(M'_T\epsilon_T)}|\mathcal{G}_0\right]\right].$$

Let $f \in A_{L_1}(M_T\varepsilon_T)^c \cap A_{d_1}(M'_T\varepsilon_T)$. For any $j \in [J_T - 1]$ and $l \in [K]$, let

$$Z_{jl} = \int_{\tau_j}^{\xi_j} |\lambda_t^l(f) - \lambda_t^l(f_0)|dt, \quad j \in [J_T - 1], \quad l \in [K]. \tag{40}$$

Using Lemma A.3, for any $f \in A_{L_1}(M_T\epsilon_T)^c$, we have

$$\begin{aligned}
\mathbb{E}_f\left[\mathbb{1}_{A_T}\mathbb{1}_{A_{d_1}(M'_T\varepsilon_T)}|\mathcal{G}_0\right] &\leqslant \mathbb{P}_f\left[\sum_{j=1}^{J_T-1} Z_{jl} \leqslant TM'_T\varepsilon_T|\mathcal{G}_0\right] \\
&\leqslant \sum_{J\in\mathcal{J}_T} \mathbb{P}_f\left[\sum_{j=1}^{J-1} Z_{jl} - \mathbb{E}_f\left[Z_{jl}\right] \leqslant TM'_T\epsilon_T - \frac{T}{2\mathbb{E}_0\left[\Delta\tau_1\right]}C(f_0)M_T\epsilon_T|\mathcal{G}_0\right] \\
&\leqslant \sum_{J\in\mathcal{J}_T} \mathbb{P}_f\left[\sum_{j=1}^{J-1} Z_{jl} - \mathbb{E}_f\left[Z_{jl}\right] \leqslant -\frac{T}{4\mathbb{E}_0\left[\Delta\tau_1\right]}C(f_0)M_T\varepsilon_T|\mathcal{G}_0\right],
\end{aligned}$$

for any $M_T \geqslant 4\mathbb{E}_0\left[\Delta\tau_1\right]M'_T$. Similarly to the proof of Theorem 3.2 in Sulem et al. [2021]), we apply Bernstein's inequality for each $J \in \mathcal{J}_T$ and obtain that

$$\mathbb{E}_f\left[\mathbb{1}_{A_T}\mathbb{1}_{A_{d_1}(M'_T\varepsilon_T)}|\mathcal{G}_0\right] \leqslant \exp\{-c(f_0)'T\}, \quad \forall f \in A_{L_1}(M_T\varepsilon_T)^c.$$

Therefore, we can conclude that

$$\mathbb{E}_0\left[\hat{Q}\left(A_{L_1}(M_T\varepsilon_T)^c \cap A_{d_1}(M'_T\varepsilon_T)\right)\mathbb{1}_{A_T}\right] \leqslant \frac{2}{M_TT\varepsilon_T^2}\mathbb{E}_0\left[KL(\hat{Q}\|\Pi(.|N))\right] + \exp\{-c(f_0)'T/2\}) = o(1),$$

since $\mathbb{E}_0\left[KL(\hat{Q}\|\Pi(.|N))\right] = O(T\varepsilon_T^2)$ by assumption (A2). This leads to our final conclusion

$$\mathbb{E}_0\left[\hat{Q}\left(\|f-f_0\|_1 > M_T\varepsilon_T\right)\right] = o(1).$$

A.3 Proof of Proposition 4.5

We recall that in this result, we consider the sigmoid Hawkes model with link function $\phi_k(x) = \theta_k(1+e^{-x})^{-1}, x \in \mathbb{R}$ for each $k \in [K]$ with unknown scale parameter $\theta = (\theta_k)_k \in \Theta$. This proposition is an extension of Theorem 3.2 in Sulem et al. [2021], and we prove it using the same strategy, based on the stochastic distance $\widetilde{d}_{1T}(f,f_0)$ (35) and the decomposition into excursions (see Lemma A.1).

We first define

$$\bar{\Upsilon}_T = \mathcal{H}_T \times [-B,B]^K \times \Theta_T = \mathcal{F}_T \times \Theta_T,$$

and note that, since $\Pi(\nu \in [-B,B]^K) = 1$,

$$\Pi(\bar{\Upsilon}_T) = \Pi(\mathcal{H}_T^c) + \Pi(\Theta_T^c).$$

Let $\sigma(x) = (1+e^{-x})^{-1}$, $x \in \mathbb{R}$, $M'_T = M'\sqrt{\kappa_T}$ with $M' > 0$ and $\kappa_T = 10(\log\log T)\log T$, and for $i \geqslant 1$,

$$S_i = \left\{(f,\theta) \in \mathcal{F}\times\Theta;\ Ki\epsilon_T \leqslant \widetilde{d}_{1T}(f,f_0) \leqslant K(i+1)\epsilon_T\right\}.$$

We use the now standard decomposition of the posterior distribution

$$\begin{aligned}
\mathbb{E}_0[\Pi(A_{d_1}(M'_T\epsilon_T)^c|N)] &\leqslant \mathbb{P}_0(\tilde{\Omega}_T^c) + \mathbb{P}_0\left(\{D_T < e^{-\kappa_TT\epsilon_T^2}\Pi(\tilde{B}_2(\epsilon_T,B))\} \cap \tilde{\Omega}_T\right) + \mathbb{E}_0[\phi\mathbb{1}_{\tilde{\Omega}_T}] \\
&\quad + \frac{e^{\kappa_TT\epsilon_T^2}}{\Pi(\tilde{B}_2(\epsilon_T,B))}\left(\Pi(\Upsilon_T^c) + \sum_{i=M'_T}^{+\infty}\int_{\Upsilon_T}\mathbb{E}_0\left[\mathbb{E}_f\left[\mathbb{1}_{\tilde{\Omega}_T}\mathbb{1}_{f\in S_i}(1-\phi)\right]|\mathcal{G}_0\right]\right]d\Pi(f)\right),
\end{aligned} \tag{41}$$

with $\phi \in [0,1]$ a test function, $\tilde{\Omega}_T$ defined in A.2, and D_T from (4). Using previous computation, we know that

$$\mathbb{P}_0(\tilde{\Omega}_T^c) = o(1) \quad \text{and} \quad \mathbb{P}_0\left(\{D_T < e^{-\kappa_T T\epsilon_T^2}\Pi(\tilde{B}_2(\epsilon_T, B))\} \cap \tilde{\Omega}_T\right) = o(1).$$

We also note that using (A1),

$$\frac{e^{\kappa_T T\epsilon_T^2}}{\Pi(\tilde{B}_2(\epsilon_T, B))}\Pi(\Upsilon_T^c) \leqslant e^{(c_1+\kappa_T)T\epsilon_T^2}(\Pi(\Theta_T^c) + \Pi(\mathcal{H}_T^c)) = o(1).$$

For the remaining terms, using the notation of Section A.2, for any $(f,\theta) \in S_i \cap \Upsilon_T$, we have that

$$\begin{aligned} T\widetilde{d}_{1T}(f, f_0) &\geqslant \sum_{k=1}^{K}\sum_{j=1}^{J_T-1}\int_{\tau_j}^{U_j^{(1)}} \left|\lambda_t^k(f) - \lambda_t^k(f_0)\right| dt \\ &\geqslant \sum_{k=1}^{K}\left|\theta_k\sigma(\nu_k) - \theta_k^0\sigma(\nu_k^0)\right| \sum_{j=1}^{J_T-1}(U_j^{(1)} - \tau_j) \\ &\geqslant \sum_{k=1}^{K}\left|r_k - r_k^0\right| \frac{T}{2\mathbb{E}_0[\Delta\tau_1]\|r_0\|_1}, \end{aligned}$$

on $\tilde{\Omega}_T$. Therefore, for any k, with $c(f_0) = \frac{1}{2\mathbb{E}_0[\Delta\tau_1]\|r_0\|_1}$, we have

$$r_k^0 - \frac{K(i+1)\epsilon_T}{c(f_0)} \leqslant r_k \leqslant r_k^0 + \frac{K(i+1).\epsilon_T}{c(f_0)} \tag{42}$$

Since $\nu_k \in [-B, B]$ and $0 \leqslant \sigma(\nu_k) \leqslant 1$, we have from (42) that

$$\frac{\theta_k^0\sigma(\nu_k^0)}{\sigma(B)} - \frac{K(i+1)\epsilon_T}{c(f_0)\sigma(B)} \leqslant \theta_k \leqslant \frac{\theta_k^0\sigma(\nu_k^0)}{\sigma(-B)} + \frac{K(i+1).\epsilon_T}{c(f_0)\sigma(-B)}.$$

Let

$$\mathcal{T}_i = \left\{(f,\theta) \in \Upsilon_T;\ 0 < r_k \leqslant \frac{\theta_k^0\sigma(\nu_k^0)}{\sigma(-B)} + \frac{K(i+1).\epsilon_T}{c(f_0)\sigma(-B)},\ \forall k\right\}. \tag{43}$$

We separate the set of indices i into two cases.

Case 1: $i\epsilon_T \leqslant 1$. Then we have that for any $(f,\theta) \in \mathcal{T}_i$,

$$\theta_k \leqslant \frac{K}{c(f_0)\sigma(-B)},$$

and the covering number denoted $\mathcal{N}_i$ of $\mathcal{T}_i$ by balls of radius $\zeta\epsilon_T$ with $\zeta = 1/(6N_0)$ with $N_0 = 1 + \sum_{k=1}^{K}\mathbb{E}_0\left[\lambda_t^k(f_0)\right]$, verify

$$\mathcal{N}_i \leqslant \left(\frac{C_0BK^2}{(\zeta i\epsilon_T)^2}\right)^K \mathcal{N}(\zeta i\epsilon_T/2, \mathcal{H}_T, \|.\|_1) \leqslant C_0' e^{2K\log T} e^{x_0 T\epsilon_T^2},$$

with $C_0, C_0' > 0$ constants and using **(A1)**.

Case 2: $i\epsilon_T \geqslant 1$. In this case, we have that

$$\mathcal{N}_i \leqslant \left(\frac{C_1K^2 i\epsilon_T}{(\zeta i\epsilon_T)^2}\right)^K \mathcal{N}(\zeta i\epsilon_T/2, \mathcal{H}_T, \|.\|_1) \leqslant C_1' e^{x_0 T\epsilon_T^2},$$

with $C_1, C_1' > 0$ constants.

Then in both cases, using the same tests $(\phi_i)_{i\geqslant M}$ as in the proof of Proposition 5.5 in Sulem et al. [2021], and $\phi = \max_{i\geqslant M}\phi_i$, we have that

$$\begin{aligned} &\mathbb{E}_0\left[\phi \mathbb{1}_{\tilde{\Omega}_T}\right] \lesssim e^{2K\log T} e^{x_0 T\epsilon_T^2}\left[\sum_{i\geqslant M_T'}^{\epsilon_T^{-1}} e^{-x_2 T i^2\epsilon_T^2} + \sum_{i\geqslant\epsilon_T^{-1}} e^{-x_2 T i\epsilon_T}\right] \lesssim e^{-x_2 M_T' T\epsilon_T^2/2} \\ &\sup_{f\in\mathcal{T}_i}\mathbb{E}_0\left[\mathbb{E}_f\left[\mathbb{1}_{\tilde{\Omega}_T}\mathbb{1}_{f\in S_i}(1-\phi_i)|\mathcal{G}_0\right]\right] \leqslant (2K+1)e^{-x_2 T(i^2\epsilon_T^2 \wedge i\epsilon_T)}, \end{aligned}$$

with $x_2 > 0$, which leads to

$$\sum_{i=M'_T}^{+\infty} \int_{\Upsilon_T} \mathbb{E}_0 \left[\mathbb{E}_f \left[\mathbb{1}_{\tilde{\Omega}_T} \mathbb{1}_{f \in S_i} (1-\phi) \right] | \mathcal{G}_0 \right] d\Pi(f) \lesssim e^{-x_2 M'_T T \epsilon_T^2 / 2},$$

and finally to the intermediate result

$$\mathbb{E}_0 \left[\Pi \left(\widetilde{d}_{1T}(f, f_0) > M'_T \epsilon_T | N \right) \right] \xrightarrow[T \to \infty]{} 0,$$

with $M'_T = M' \sqrt{\kappa_T}$ with M' large enough.

Extending Lemma A.4 of Sulem et al. [2021] to the context of sigmoid link with unknown shift, we can easily prove that for $(f, \theta) \in \Upsilon_T$ such that $r_k = \theta_k \sigma(\nu_k) \leqslant \max(r_k^0, c_0)$, $\forall k$, with $c_0 > 0$, there exists $l \in [K]$ and $C(f_0)$ such that on $\tilde{\Omega}_T$,

$$\mathbb{E}_f [Z_{1l}] \geqslant C(f_0) (\left\| r_f - r_0 \right\|_1 + \|h - h_0\|_1).$$

Then, using the same steps as the proof of Theorem 3.2 in Sulem et al. [2021], we can obtain that

$$\mathbb{E}_0 \left[\Pi \left((\left\| r_f - r_0 \right\|_1 + \|h - h_0\|_1) > M_T \epsilon_T | N \right) \right] \xrightarrow[T \to \infty]{} 0,$$

with $M_T = M \sqrt{\kappa_T}$ with $M > M'$. Re-defining the L_1-neighborhood as

$$A_{L_1}(\varepsilon) = \{(f, \theta) \in \mathcal{F} \times \Theta, \|\theta \sigma(\nu) - \theta_0 \sigma(\nu_0)\|_1 + \|h - h_0\|_1 < \varepsilon\}, \quad \varepsilon > 0,$$

the previous result can be re-written as $\mathbb{E}_0 \left[\Pi \left(A_{L_1}(M_T \epsilon_T)^c | N \right) \right] = o(1)$.

We now separate ν and θ using a test similar to the proof of Proposition 3.5 in Sulem et al. [2021] for the shifted ReLU model. For this, for any $\eta > 0$ and with $\theta_T = e^{c_2 T \epsilon_T^2}$, we define

$$A^k(T) = \left\{ t \in [0, T];\ \lambda_t^k(f_0, \theta_0) > \theta_k^0 - \eta \right\}, \quad 1 \leqslant k \leqslant K,$$
$$\Omega_A = \{|A^k(T)| > z_0 T,\ \forall k \in [K]\},$$

with $z_0 > 0$ a constant. We also define $\tilde{\Omega}'_T = \tilde{\Omega}_T \cap \Omega_A$ and a neighborhood around θ_0

$$\bar{A}(R) := \{\theta \in \Theta;\ \|\theta - \theta_0\|_1 \leqslant R\}, \quad R > 0.$$

Let $\tilde{M}_T = \tilde{M} \sqrt{\kappa_T}$ with $\tilde{M} > M$. Using the standard decomposition of the posterior distribution $\Pi(\bar{A}(\tilde{M}_T \epsilon_T)^c | N)$, with $A = \bar{A}(\tilde{M}_T \epsilon_T)^c$, $B = A_{L_1}(M_T \epsilon_T)$, and the subset $\tilde{\Omega}'_T$, we only need to construct a test function $\phi \in [0, 1]$ verifying

$$\mathbb{E}_0 \left[\phi \mathbb{1}_{\tilde{\Omega}'_T} \right] = o(1), \qquad \sup_{(f,\theta) \in A_{L_1}(M_T \epsilon_T) \cap (F_T \times \bar{A}(\tilde{M}_T \epsilon_T)^c \cap \Theta_T)} \mathbb{E}_0 \left[\mathbb{E}_f \left[(1-\phi) \mathbb{1}_{\tilde{\Omega}'_T} \right] \middle| \mathcal{G}_0 \right] = o(e^{-(\kappa_T + c_1) T \epsilon_T^2}). \tag{44}$$

We consider a parameter $(f_1, \theta_1) \in A_{L_1}(M_T \epsilon_T) \cap (F_T \times \bar{A}(\tilde{M}_T \epsilon_T)^c \cap \Theta_T$, and for any $k \in [K]$, we define the following subset of the observation window

$$I_k^0(f_1, \theta_1) = \left\{ t \in [0, T];\ \lambda_t^k(f_1, \theta_1) > \theta_k^1 - \eta,\ \lambda_t^k(f_0, \theta_0) > \theta_k^0 - \eta \right\} \subset A^k(T). \tag{45}$$

To prove that $|I_k^0(f_1, \theta_1)| \gtrsim T$, we construct the following set of excursions $\mathcal{E}$. From Assumption 4.1, let $l \in [K]$ such that $h_{lk}^{0+}(x) \geqslant c_\star$, $\forall x \in [x_1, x_2]$ and $x' = \min(x_1, x_2 - x_1/3)$. For any $n_1^k \in \mathbb{Z}_{\geqslant 0}$, let

$$\mathcal{E}_l(n_1^k) := \{j \in [J_T];\ N[\tau_j, \tau_j + x') = N^l[\tau_j, \tau_j + x') = n_1^k, N[\tau_j + \delta',\ \tau_{j+1}) = 0\}, \tag{46}$$

where the τ_j's are the regenerative times defined in Lemma A.1. For choosing the number of events n_1^k, we separate into two cases.

Case 1: $\theta_k^0 > \theta_k^1$. In this case, we define

$$n_1^k = \lfloor \frac{\log(\theta_k^0 / \eta - 1) + B}{c_\star} \rfloor + 1.$$

Then, we can easily see that for $j \in \mathcal{E}_l(n_1^k)$ and $t \in [\tau_j + x + x_1, \tau_j + x + x_1 + x']$, we have

$$\tilde{\lambda}_t^k(\nu_0, h_0) \geqslant \log \left(\frac{\theta_k^0}{\eta} - 1 \right) \implies \lambda_t^k(f_0, \theta_0) > \theta_k^0 - \eta,$$
$$\tilde{\lambda}_t^k(\nu_1, h_1) \geqslant \log \left(\frac{\theta_k^0}{\eta} - 1 \right) \implies \lambda_t^k(f_1, \theta_1) > \theta_k^1 - \frac{\theta_k^1}{\theta_k^0} \eta > \theta_k^1 - \eta,$$

therefore $t \in I_k^0(f_1, \theta_1)$.

Case 2: $\theta_k^0 \leqslant \theta_k^1$. In this case, we define

$$n_1^k = \lfloor \frac{\log(\theta_k^1/\eta - 1) + B}{c_\star} \rfloor + 1,$$

and we similarly have that $t \in I_k^0(f_1, \theta_1),\ \forall t[\tau_j + x + x_1, \tau_j + x + x_1 + x'],\ j \in \mathcal{E}_l(n_1^k)$.

Moreover, using a proof similar to the one of Lemma A.5 in Sulem et al. [2021], we can show that for $\eta = \eta_T = c\sqrt{\kappa_T}\epsilon_T$ with $c > 0$ large enough, there exist $p_0, u_0 > 0$ such that

$$\mathbb{P}_0\left[|\mathcal{E}_l(n_1^k)| \leqslant p_0 T, \forall k \in [K]\right] = o(e^{-u_0 T\epsilon_T^2}),$$

since $\epsilon_T \gtrsim (\log T)^3 T^{-1}$, and therefore $n_1^k \lesssim \log T$. We can also see that $|I_k^0(f_1, \theta_1)| \geqslant x'|\mathcal{E}_l(n_1^k)|$.

We now define our generic test function

$$\phi(f_1, \theta_1) := \max_{k\in[K]} \mathbb{1}_{|N^k(I_k^0(f_1,\theta_1)) - \Lambda_k^0(I_k^0(f_1,\theta_1)| > v_T} \vee \mathbb{1}_{|\mathcal{E}_l(n_1^k)| < p_0 T}, \tag{47}$$

where $\Lambda_k^0(I_k^0(f_1, \theta_1)) = \int_0^T \mathbb{1}_{I_k^0(f_1,\theta_1)} \lambda_t^k(f_0, \theta_0)dt \geqslant |I_k^0(f_1, \theta_1)|(\theta_k^0 - \eta_T)$, $v_T = w_T T\varepsilon_T$, $w_T = w_0\sqrt{(\kappa_T + c_1)}$, with $w_0 > 0$ a constant chosen later. Using the same steps as in the proof of Lemma A.5 in Sulem et al. [2021], we can show that

$$\mathbb{E}_0\left[\phi(f_1, \theta_1)\mathbb{1}_{\tilde{\Omega}'}\right] = o(e^{-u_0 T\epsilon_T^2}).$$

Moreover, using that

$$\begin{aligned}
\Lambda_k^0(I_k^0(f_1, \theta_1)) &= \int_{I_k^0(f_1,\theta_1)} \lambda_t^k(f_1, \theta_1)dt + \int_{I_k^0(f_1,\theta_1)} (\lambda_t^k(f_0, \theta_0) - \lambda_t^k(f_1, \theta_1))dt \\
&\geqslant \int_{I_k^0(f_1,\theta_1)} \lambda_t^k(f_1, \theta_1)dt + |I_k^0(f_1, \theta_1)|(\theta_k^0 - \eta_T - \theta_k^1) \\
&\geqslant \int_{I_k^0(f_1,\theta_1)} \lambda_t^k(f_1, \theta_1)dt + p_0 T(\bar{M}_T\epsilon_T - \eta_T) \geqslant \int_{I_k^0(f_1,\theta_1)} \lambda_t^k(f_1, \theta_1)dt + p_0 T\bar{M}_T\epsilon_T/2,
\end{aligned}$$

for T large enough, we have that, with $\Lambda_k^1(I_k^0(f_1, \theta_1) := \int_{I_k^0(f_1,\theta_1)} \lambda_t^k(f_1, \theta_1)dt$

$$\begin{aligned}
\left\{|N^k(I_k^0(f_1, \theta_1)) - \Lambda_k^0(I_k^0(f_1, \theta_1)| < v_T\right\} &\subset \left\{N^k(I_k^0(f_1, \theta_1)) - \Lambda_k^1(I_k^1(f_1, \theta_1) + p_0 T\bar{M}_T\epsilon_T/2 < v_T\right\} \\
&\subset \left\{N^k(I_k^0(f_1, \theta_1)) - \Lambda_k^1(I_k^1(f_1, \theta_1) < -p_0 T\bar{M}_T\epsilon_T/4\right\},
\end{aligned}$$

with $\bar{M}_T = M\sqrt{\kappa_T}\epsilon_T$ with $M > 2w_0$ large enough. Therefore, like in Sulem et al. [2021], using inequality (7.7) of Hansen et al. [2015], we can obtain that

$$\mathbb{E}_f\left[(1 - \phi_k(f_1, \theta_1))\mathbb{1}_{\tilde{\Omega}_T'}\right] \leqslant \mathbb{P}_f\left[\left\{|N^k(I_k^0(f_1, \theta_1)) - \Lambda_k(I_k^0(f_1, \theta_1), f_0)| \leqslant v_T\right\} \cap \tilde{\Omega}_T'\right] = o(e^{-(\kappa_T + c_1)T\epsilon_T^2}).$$

Then, to define our global test ϕ, we cover the space $A_{L_1}(M_T\epsilon_T) \cap \mathcal{F}_T \times (\bar{A}(\tilde{M}_T\epsilon_T)^c \cap \Theta_T)$ with L_1-balls $\{B_i\}_{1\leqslant i\leqslant \mathcal{N}}$ of radius $\zeta\epsilon_T$, with $\zeta > 0$ and $\mathcal{N} \in \mathbb{N}$ the covering number, and for each ball B_i centered at (f_i, θ_i), we consider the elementary test $\phi(f_i, \theta_i)$ as in (47). Define $\phi := \max_{i\in\mathcal{N}} \phi(f_i, \theta_i)$, we then obtain that

$$\mathbb{E}_0\left[\phi\mathbb{1}_{\tilde{\Omega}_T'}\right] \leqslant \mathcal{N}e^{-u_0 T\epsilon_T^2}, \qquad \sup_{(f,\theta)\in A_{L_1}(M_T\epsilon_T)\cap\mathcal{F}_T\times\bar{A}(\tilde{M}_T\epsilon_T)^c} \mathbb{E}_0\left[\mathbb{E}_f\left[(1 - \phi)\mathbb{1}_{\tilde{\Omega}_T'}\right]\middle|\mathcal{G}_0\right] = o(e^{-(\kappa_T + c_1)T\epsilon_T^2}).$$

Moreover, we have that

$$\begin{aligned}
\mathcal{N} &\leqslant \left(\frac{2KB\theta_T}{(\zeta\epsilon_T)^2}\right)^K \mathcal{N}(\zeta\epsilon_T, \mathcal{H}_T, \|.\|_1) \lesssim e^{-2K\log\epsilon_T} e^{c_2 T\epsilon_T^2} e^{x_0 T\epsilon_T^2} \\
&\lesssim e^{2K\log T} e^{(c_2 + x_0)T\epsilon_T^2} = o(e^{u_0 T\epsilon_T^2}),
\end{aligned}$$

for u_0 large enough, which implies that $\mathbb{E}_0\left[\phi\mathbb{1}_{\tilde{\Omega}_T'}\right] = o(1)$ and terminates this proof.

B Additional proofs

B.1 Proof of Lemma 4.3

Using the proof of Proposition 2.3 in Sulem et al. [2021], we can easily obtain that $N =^{\mathcal{D}} N'$ implies that

$$\frac{\theta_k}{\theta'_k} = \frac{\tilde{\sigma}(\nu'_k)}{\tilde{\sigma}(\nu_k)} = \frac{\tilde{\sigma}(h'_{lk})}{\tilde{\sigma}(h_{lk})}, \quad \forall l, k.$$

Now, using Assumption 4.1 and the proof of Proposition 2.3 in Sulem et al. [2021], one can show that

$$\mathbb{P}[\sup_{t\geqslant 0} \lambda_t^k(f) = \theta_k] = 1$$
$$\mathbb{P}[\sup_{t\geqslant 0} \lambda_t^k(f') = \theta'_k] = 1.$$

Then, one can conclude that $\theta = \theta'$, implying also that $h = h'$ and $\nu = \nu'$.

C Additional derivation in the sigmoid Hawkes model with data augmentation

C.1 Updates of the fixed-dimension mean-field variational algorithm

In this section, we derive the analytic forms of the conditional updates in Algorithm 1, the mean-field variational algorithm with fixed dimensionality described in Section 4.3. For ease of exposition, we drop the indices k and s and use the notation Q_1, Q_2 for the variational factors. In the following computation, we use the notation c to denote a generic constant which value can vary from one line to the other. We also define $\alpha := 0.1$ and $\eta = 10$.

From the definition of the augmented posterior (21), we first note that

$$\begin{aligned}\log p(f, N, \omega, \bar{N}) &= \log \Pi(f, \omega, \bar{N}|N) + \log p(N) = L_T(f, \omega, \bar{N}; N) + \log \Pi(f) + \log p(N) + c \\ &= \log p(\omega|f, N) + \log p(\bar{N}|f, N) + \log \Pi(f) + \log p(N) + c. \end{aligned} \tag{48}$$

In the previous equality we have used the facts that $p(\omega|f, N, \bar{N}) = p(\omega|f, N)$ and $p(\bar{N}|f, N, \omega) = p(\bar{N}|f, N)$. We recall our notation $H(t) = (H^0(t), H^1(t), \dots, H^K(t)) \in \mathbb{R}^{KJ+1}$, $t \in \mathbb{R}$, where for $k \in [K]$, $H^k(t) = (H_j^k(t))_{j=1,\dots,J}$ and H_j^k defined in (25). We have that

$$\begin{aligned}\mathbb{E}_{Q_2}[\log p(\omega|f, N)] &= \mathbb{E}_{Q_2}\left[\sum_{i\in[N]} g(\omega_i, \tilde{\lambda}_{T_i}(f))\right] + c = \mathbb{E}_{Q_2}\left[\sum_{i\in[N]} -\frac{\omega_i \tilde{\lambda}_{T_i}(f)^2}{2} + \frac{\tilde{\lambda}_{T_i}(f)}{2}\right] + c \\ &= \mathbb{E}_{Q_2}\left[\sum_{i\in[N]} -\frac{\omega_i \alpha^2 (f^T H(T_i) H(T_i)^T f - 2\eta H(T_i)^T f + \eta^2)}{2} + \frac{\alpha H(T_i)^T f}{2}\right] + c \\ &= \mathbb{E}_{Q_2}\left[-\frac{1}{2}\sum_{i\in[N]} \left\{\omega_i \alpha^2 f^T H(T_i) H(T_i)^T f - \alpha(2\omega_i \alpha\eta + 1) H(T_i)^T f + \omega_i \alpha^2 \eta^2\right\}\right] + c \\ &= -\frac{1}{2}\sum_{i\in[N]} \left\{\mathbb{E}_{Q_2}[\omega_i]\alpha^2 f^T H(T_i) H(T_i)^T f - \alpha(2\mathbb{E}_{Q_2}[\omega_i]\alpha\eta + 1) H(T_i)^T f + \mathbb{E}_{Q_2}[\omega_i]\alpha^2\eta^2\right\} + c. \end{aligned}$$

Moreover, we also have that

$$\begin{aligned}\mathbb{E}_{Q_2}[\log p(\bar{N}|f, N)] &= \mathbb{E}_{Q_2}\left[-\frac{1}{2}\sum_{j\in[\bar{N}]} \left\{\bar{\omega}_j \alpha^2 f^T H(\bar{T}_j) H(\bar{T}_j)^T f - \alpha(2\bar{\omega}_j \alpha\eta - 1) H(\bar{T}_j)^T f + \bar{\omega}_j \alpha^2 \eta^2\right\}\right] + c \\ &= \int_0^T \int_0^\infty \left[-\frac{1}{2}\left(\bar{\omega}\alpha^2 f^T H(t) H(t)^T f - \alpha(2\bar{\omega}\alpha\eta - 1) H(t)^T f + \bar{\omega}\alpha^2\eta^2\right)\right] \Lambda(t, \bar{\omega}) d\bar{\omega} dt + c \\ &= -\frac{1}{2}\left[f^T\left(\alpha^2 \int_0^T \int_0^\infty \bar{\omega} H(t) H(t)^T \Lambda(t, \bar{\omega}) d\bar{\omega} dt\right) f + f^T\left(\alpha \int_0^T \int_0^\infty (2\bar{\omega}\alpha\eta - 1) H(t)^T \Lambda(t, \bar{\omega}) d\bar{\omega} dt\right)\right] + c. \end{aligned}$$

Besides, we have $\mathbb{E}_{Q_2}[\log \Pi(f)] = -\frac{1}{2} f^T \Sigma^{-1} f + f^T \Sigma^{-1} \mu + c$. Therefore, using (23), we obtain that

$$\begin{aligned}
\log Q_1(f) = -\frac{1}{2}\Bigg[& f^T \left(\alpha^2 \sum_{i\in[N]} \mathbb{E}_{Q_2}[\omega_i] H(T_i) H(T_i)^T + \alpha^2 \int_0^T \int_0^\infty \bar{\omega} H(t) H(t)^T \Lambda(t,\bar{\omega}) d\bar{\omega} dt + \Sigma^{-1} \right) f \\
& - f^T \left(\alpha \sum_{i\in[N]} (2\mathbb{E}_{Q_2}[\omega_i]\alpha\eta + 1) H(T_i)^T + \alpha \int_0^T \int_0^\infty (2\bar{\omega}\alpha\eta - 1) H(t)^T \Lambda(t,\bar{\omega}) d\bar{\omega} dt + 2\Sigma^{-1}\mu \right) \Bigg] + c \\
=: & -\frac{1}{2}(f-\tilde{\mu})^T \tilde{\Sigma}^{-1} (f-\tilde{\mu}) + c,
\end{aligned}$$

therefore $Q_1(f)$ is a normal distribution with mean vector $\tilde{\mu}$ and covariance matrix $\tilde{\Sigma}$ given by

$$\tilde{\Sigma}^{-1} = \alpha^2 \sum_{i\in[N]} \mathbb{E}_{Q_2}[\omega_i] H(T_i) H(T_i)^T + \alpha^2 \int_0^T \int_0^\infty \bar{\omega} H(t) H(t)^T \Lambda(t,\bar{\omega}) d\bar{\omega} dt + \Sigma^{-1}, \tag{49}$$

$$\tilde{\mu} = \frac{1}{2}\tilde{\Sigma}\left[\alpha \sum_{i\in[N]} (2\mathbb{E}_{Q_2}[\omega_i]\alpha\eta + 1) H(T_i)^T + \alpha \int_0^T \int_0^\infty (2\bar{\omega}\alpha\eta - 1) H(t)^T \Lambda(t,\bar{\omega}) d\bar{\omega} dt + 2\Sigma^{-1}\mu \right]. \tag{50}$$

For $Q_2(\omega, \bar{N})$, we first note that using (24) and (48), we have $Q_2(\omega, \bar{N}) = Q_{21}(\omega) Q_{22}(\bar{N})$. Using the same computation as Donner and Opper [2019]) Appendices B and D, one can then show that

$$\begin{aligned}
Q_{21}(\omega) &= \prod_{i\in[N]} p_{PG}(\omega_i | 1, \underline{\lambda}_{T_i}), \\
\underline{\lambda}_t &= \sqrt{\mathbb{E}_{Q_1}[\tilde{\lambda}_t(f)^2]} = \alpha^2 \sqrt{H(t)^T \tilde{\Sigma} H(t) + (H(t)^T \tilde{\mu})^2 - 2\eta H(t)^T \tilde{\mu} + \eta^2}, \quad \forall t \in [0,T],
\end{aligned}$$

and that Q_{22} is a marked Poisson point process measure on $[0,T] \times \mathbb{R}^+$ with intensity

$$\begin{aligned}
\Lambda(t,\bar{\omega}) &= \theta e^{\mathbb{E}_{Q_1}[g(\bar{\omega}, -\tilde{\lambda}_t(f))]} p_{PG}(\bar{\omega}; 1, 0) = \theta \frac{\exp(-\frac{1}{2}\mathbb{E}_{Q_1}[\tilde{\lambda}_t(f)])}{2\cosh\frac{\underline{\lambda}_t(f)}{2}} p_{PG}(\bar{\omega} | 1, \underline{\lambda}_t(f)) \\
&= \theta\sigma(-\underline{\lambda}_t) \exp\left\{ \frac{1}{2}(\underline{\lambda}_t(f) - \mathbb{E}_{Q_1}[\tilde{\lambda}_t(f)]) \right\} p_{PG}(\bar{\omega} | 1, \underline{\lambda}_t) \\
\mathbb{E}_{Q_1}[\tilde{\lambda}_t(f)] &= \alpha(H(t)^T \tilde{\mu} - \eta).
\end{aligned}$$

Therefore, we have that

$$\mathbb{E}_{Q_1}[\omega_i] = \frac{1}{2\underline{\lambda}_{T_i}} \underline{\lambda}_{T_i}, \forall i \in [N].$$

C.2 Analytic form of the ELBO

In this section, we provide the derivation of the $ELBO(\hat{Q}_s)$ in our adaptive mean-field variational algorithm, Algorithm 2, for each $s = (\delta, D)$. For ease of expositions, we will drop the subscript s. From (28), we have

$$\begin{aligned}
ELBO(\hat{Q}) &= \mathbb{E}_{\hat{Q}}\left[\log \frac{p(f,\omega,\bar{N},N)}{\hat{Q}_1(f)\hat{Q}_2(\omega,\bar{N})} \right] \\
&= \mathbb{E}_{\hat{Q}_2}\left[-\log \hat{Q}_2(\omega,\bar{N}) \right] + \mathbb{E}_{\hat{Q}_2}\left[\mathbb{E}_{\hat{Q}_1}\left[\log p(f,\omega,\bar{N},N) \right] \right] + \mathbb{E}_{\hat{Q}_1}[-\log \hat{Q}_1(f)].
\end{aligned}$$

Now using the notation of Section 4.3, we first note that defining $K(t) := H(t)H(t)^T$, we have that

$$\begin{aligned}
\mathbb{E}_{\hat{Q}_1}[\tilde{\lambda}_{T_i}(f)^2] &= tr(K(t)\tilde{\Sigma}) + \tilde{\mu}^T K(t) \tilde{\mu} \\
\mathbb{E}_{\hat{Q}_1}\left[\log \mathcal{N}(f; \mu, \Sigma) \right] &= -\frac{1}{2} tr(\Sigma^{-1}\tilde{\Sigma}) - \frac{1}{2}\tilde{\mu}^T \Sigma^{-1} \tilde{\mu} + \tilde{\mu}^T \Sigma^{-1} \mu - \frac{1}{2}\mu^T \Sigma^{-1} \mu - \frac{1}{2}\log|2\pi\Sigma|.
\end{aligned}$$

Moreover, we have

$$\mathbb{E}_{\hat{Q}_1}[\log \hat{Q}_1(f)] = -\frac{|m|}{2} - \frac{1}{2}\log|2\pi\tilde{\Sigma}|.$$

Using that for any $c > 0$, $p_{PG}(\omega; 1, c) = e^{-c^2\omega/2}\cosh(c/2)p_{PG}(\omega; 1, 0)$, we also have

$$\begin{aligned}
\mathbb{E}_{\hat{Q}_2}\left[-\log \hat{Q}_{2D}(\omega, \bar{N})\right] &= \sum_k \sum_{i\in[N_k]} -\mathbb{E}_{\hat{Q}_2}[\log p_{PG}(\omega_i^k, 1, 0)] + \frac{1}{2}\mathbb{E}_{\hat{Q}_2}[\omega_i^k]\mathbb{E}_{\hat{Q}_1}[\tilde{\lambda}_{T_i}(f)^2] - \log\cosh\left(\frac{\underline{\lambda}_{T_i}(f)}{2}\right) \\
&\quad - \int_{t=0}^{T}\int_0^{+\infty}[\log\Lambda(t,\bar{\omega})]\Lambda(t,\bar{\omega})d\bar{\omega}dt + \int_{t=0}^{T}\int_0^{+\infty}\Lambda(t,\bar{\omega})d\bar{\omega}dt \\
&= \sum_k \sum_{i\in[N_k]} -\mathbb{E}_{\hat{Q}_2}[\log p_{PG}(\omega_i^k, 1, 0)] + \frac{1}{2}\mathbb{E}_{\hat{Q}_2}[\omega_i^k]\mathbb{E}_{\hat{Q}_{1D}}[\tilde{\lambda}_{T_i}(f)^2] - \log\cosh\left(\frac{\underline{\lambda}_{T_i}(f)}{2}\right) \\
&\quad - \int_{t=0}^{T}\int_0^{+\infty}\left[\log\theta_k - \frac{1}{2}\mathbb{E}_{\hat{Q}_1}[\tilde{\lambda}_{T_i}(f)] - \log 2 - \log\cosh\left(\frac{\underline{\lambda}_{T_i}(f)}{2}\right) - \frac{1}{2}\mathbb{E}_{\hat{Q}_{1D}}[\tilde{\lambda}_{T_i}(f)^2]\bar{\omega}\right. \\
&\quad \left. + \log\cosh\left(\frac{1}{2}\underline{\lambda}_{T_i}(f)\right) + \log p_{PG}(\bar{\omega}; 1, 0) - 1\right]\Lambda^k(t)p_{PG}(\bar{\omega}; 1, \underline{\lambda}_{T_i}(f))dtd\bar{\omega} \\
&= \sum_k \sum_{i\in[N_k]} -\mathbb{E}_{\hat{Q}_2}[\log p_{PG}(\omega_i^k, 1, 0)] + \frac{1}{2}\mathbb{E}_{\hat{Q}_2}[\omega_i^k]\mathbb{E}_{\hat{Q}_{1D}}[\tilde{\lambda}_{T_i}(f)^2] - \log\cosh\left(\frac{\underline{\lambda}_{T_i}(f)}{2}\right) \\
&\quad - \int_{t=0}^{T}\left[\log\theta_k - \frac{1}{2}\mathbb{E}_{\hat{Q}_1}[\tilde{\lambda}_{T_i}(f)] - \log 2 - \frac{1}{2}\mathbb{E}_{\hat{Q}_1}[\tilde{\lambda}_{T_i}(f)^2]\mathbb{E}_{\hat{Q}_2}[\bar{\omega}] - 1\right]\Lambda^k(t)dt \\
&\quad - \int_{t=0}^{T}\int_0^{+\infty}\log p_{PG}(\omega; 1, 0)\Lambda^k(t)p_{PG}(\omega; 1, \underline{\lambda}_{T_i}(f))d\omega dt.
\end{aligned}$$

with $\Lambda^k(t) = \theta_k\int_0^\infty\Lambda^k(t,\bar{\omega})d\bar{\omega} = \frac{e^{-\frac{1}{2}\mathbb{E}_{\hat{Q}_1}[\tilde{\lambda}_{T_i}(f)]}}{2\cosh\frac{\underline{\lambda}_{T_i}(f)}{2}}$.

$$\begin{aligned}
&\mathbb{E}_{\hat{Q}_2}\left[\mathbb{E}_{\hat{Q}_1}\left[\log p(f_k, \omega, \bar{N}, N)\right]\right] = \sum_k\sum_{i\in[N_k]}\left\{\log\theta_k + \mathbb{E}_{\hat{Q}_2}\left[\mathbb{E}_{\hat{Q}_1}\left[g(\omega_i^k, \tilde{\lambda}_{T_i}(f))\right] + \log p_{PG}(\omega_i^k; 1, 0)\right]\right\} \\
&+ \sum_k \log\theta_k + \mathbb{E}_{\hat{Q}_2}\left[\mathbb{E}_{\hat{Q}_1}\left[g(\bar{\omega}_t, -\tilde{\lambda}_{T_i}(f))\right] + \log p_{PG}(\bar{\omega}_t; 1, 0)\right] + \mathbb{E}_{\hat{Q}_1}\left[\log\mathcal{N}(f_k; \mu_k, \Sigma_k)\right] \\
&= \sum_k\sum_{i\in[N_k]}\log\theta_k - \log 2 - \frac{1}{2}\mathbb{E}_{\hat{Q}_1}\left[\tilde{\lambda}_{T_i}(f)^2\right]\mathbb{E}_{\hat{Q}_2}\left[\omega_i^k\right] + \frac{1}{2}\mathbb{E}_{\hat{Q}_1}\left[\tilde{\lambda}_{T_i}(f)\right] + \mathbb{E}_{\hat{Q}_2}\left[\log p_{PG}(\omega_i^k; 1, 0)\right] \\
&+ \int_0^T\int_0^{+\infty}\left[\log\theta_k - \log 2 - \frac{1}{2}\mathbb{E}_{\hat{Q}_1}\left[\tilde{\lambda}_{T_i}(f)^2\right]\bar{\omega} - \frac{1}{2}\mathbb{E}_{\hat{Q}_1}\left[\tilde{\lambda}_{T_i}(f)\right] + \log p_{PG}(\bar{\omega}; 1, 0)\right]\Lambda^k(t)p_{PG}(\omega; 1, \underline{\lambda}_{T_i}(f))d\omega dt \\
&+ \mathbb{E}_{\hat{Q}_1}\left[\log\mathcal{N}(f_k; \mu_k, \Sigma_k)\right] - \theta_k T \\
&= \sum_k\sum_{i\in[N_k]}\log\theta_k - \log 2 - \frac{1}{2}\mathbb{E}_{\hat{Q}_1}\left[\tilde{\lambda}_{T_i}(f)^2\right]\mathbb{E}_{\hat{Q}_2}\left[\omega_i^k\right] + \frac{1}{2}\mathbb{E}_{\hat{Q}_1}\left[\tilde{\lambda}_{T_i}(f)\right] + \mathbb{E}_{\hat{Q}_2}\left[\log p_{PG}(\omega_i^k; 1, 0)\right] \\
&+ \int_0^T\left[\log\theta_k - \log 2 - \frac{1}{2}\mathbb{E}_{\hat{Q}_1}\left[\tilde{\lambda}_{T_i}(f)^2\right]\mathbb{E}_{\hat{Q}_2}[\bar{\omega}] - \frac{1}{2}\mathbb{E}_{\hat{Q}_1}\left[\tilde{\lambda}_{T_i}(f)\right]\right]\Lambda^k(t)dt \\
&+ \int_0^T\int_0^{+\infty}\log p_{PG}(\bar{\omega}; 1, 0)\Lambda^k(t)p_{PG}(\bar{\omega}; 1, \underline{\lambda}_{T_i}(f))d\bar{\omega}dt + \mathbb{E}_{\hat{Q}_1}\left[\log\mathcal{N}(f_k; \mu_k, \Sigma_k)\right] - \theta_k T.
\end{aligned}$$

Therefore, with $c > 0$ a constant that does not depend on the size of the model, with zero mean prior $\mu = 0$,

$$\begin{aligned}
ELBO(\hat{Q}) = \frac{|m|}{2} + \frac{1}{2}\log|2\pi\tilde{\Sigma}| - \frac{1}{2}tr(\Sigma^{-1}\tilde{\Sigma}) - \frac{1}{2}\tilde{\mu}^T\Sigma^{-1}\tilde{\mu} - \frac{1}{2}\log|2\pi\Sigma| \\
+ \sum_k\sum_{i\in[N_k]}\log\theta_k - \log 2 + \frac{\mathbb{E}_{\hat{Q}_1}\left[\tilde{\lambda}_{T_i}^k(f)\right]}{2} - \log\cosh\left(\frac{\tilde{\lambda}_{T_i}^k(f)}{2}\right) \\
+ \int_{t=0}^{T}\int_0^{+\infty}\Lambda(t,\bar{\omega})d\bar{\omega}dt - \theta_k T.
\end{aligned}$$

C.3 Gibbs sampler

From the augmented posterior $\Pi_A(f, \omega, \bar{N})$ defined in (21) and using the Gaussian prior family described in Section 4.2, similar computation as Appendix C.1 can provide analytic forms of the conditional posterior distributions

$\Pi_A(f|\omega, \bar{N}, N), \Pi_A(\omega|N, f)$ and $\Pi_A(\bar{N}|f, N)$. This allows to design a Gibbs sampler algorithm that sequentially samples the parameter f, the latent variables ω and Poisson process $\bar{N}$. With the notation of Appendix C.1, such procedure can be defined as

For every $k \in [K]$,

(Sample latent variables) $\omega_i^k|N, f_k \sim p_{PG}(\omega_i^k; 1, \tilde{\lambda}^k_{T_i^k}(f)), \quad \forall i \in [N_k]$

$\bar{N}^k|f_k$, a Poisson process on $[0, T]$ with intensity $\Lambda^k(t, \bar{\omega}) = \theta_k \sigma(-\tilde{\lambda}_t^k(f)) p_{PG}(\bar{\omega}; 1, \tilde{\lambda}_t^k(f))$

(Update hyperparameters) $R_k = \bar{N}^k[0, T]$

$H_k = [H_{N^k}, H_{\bar{N}^k}], \ [H_{N^k}]_{id} = H_j(T_i^k), \ [H_{\bar{N}^k}]_{jd} = H_b(\bar{T}_j^k), \ d = 0, \dots, KJ, \ i \in [N_k], \ j \in [R_k]$

$D_k = Diag([\omega_i^k]_{i\in[N^k]}, [\bar{\omega}_j^k]_{j\in[R^k]})$

$\tilde{\Sigma}_k = [\beta^2 H_k D_k (H_k)^T + \Sigma^{-1}]^{-1}$

$\tilde{\mu}_k = \tilde{\Sigma}_k \left(H_k \left[\beta v_k + \beta^2 \eta u_k \right] + \Sigma^{-1} \mu \right), \quad v_k = 0.5[\mathbb{1}_{N_k}, -\mathbb{1}_{R_k}], \quad u_k = [[\omega_i^k]_{i\in[N_k]}, [\bar{\omega}_j^k]_{j\in[R_k]}]$

(Sample parameter) $f_k|N, \bar{N}^k, \omega^k \sim \mathcal{N}(f_k; \tilde{m}_k, \tilde{\Sigma}_k)$.

These steps are summarised in Algorithm 4. We note that in this algorithm, one does not need to perform a numerical integration, however, sampling the latent Poisson process is computationally intensive. In our numerical experiments, we use the Python package `polyagamma`[3] to sample the Polya-Gamma variables and a thinning algorithm to sample the inhomogeneous Poisson process.

Algorithm 4 Gibbs sampler in the sigmoid Hawkes model with data augmentation

Input: N, n_{iter}, μ, Σ.
Output: Samples $S = (f_i)_{i\in[n_{iter}]}$ from the posterior $\Pi_A(f|N)$.

```
Precompute (H_k(T_i^k))_i, k ∈ [K].
Initialise f ~ N(f, μ, Σ) and S = [].
for t ← 1 to n_iter do
    for k ← 1 to K do
        for i ← 1 to N_k do
            Sample ω_i^k ~ p_PG(ω_i^k; 1, λ̃^k_{T_i^k}(f))
        end for
        Sample (T̄_j^k)_{j=1,R_k} a Poisson temporal point process on [0, T] with intensity θ_k σ(−λ̃_t^k(f))
        for j ← 1 to R_k do
            Sample ω̄_j^k ~ p_PG(ω; 1, λ̃^k_{T̄_j^k}(f))
        end for
        Update Σ̃_k = [β² H_k D_k (H_k)^T + Σ^{-1}]^{-1}
        Update μ̃_k = Σ̃_k ( H_k [ β v_k + β² η u_k ] + Σ^{-1} μ )
        Sample f_k ~ N(f_k; μ̃_k, Σ̃_k)
    end for
    Add f = (f_k)_k to S.
end for
```

D Additional details and results of the simulation study

D.1 Hyperparameters

We approximate integrals using the Gaussian quadrature method with $n_{GQ} = 2^{D+1} T/A$ points in the univariate settings (Simulation 1,2 and 3). In Simulation 4, we set n_{GQ} to reduce the computational time.

D.2 Self-inhibition scenarios of Simulation 4

[3] https://pypi.org/project/polyagamma/

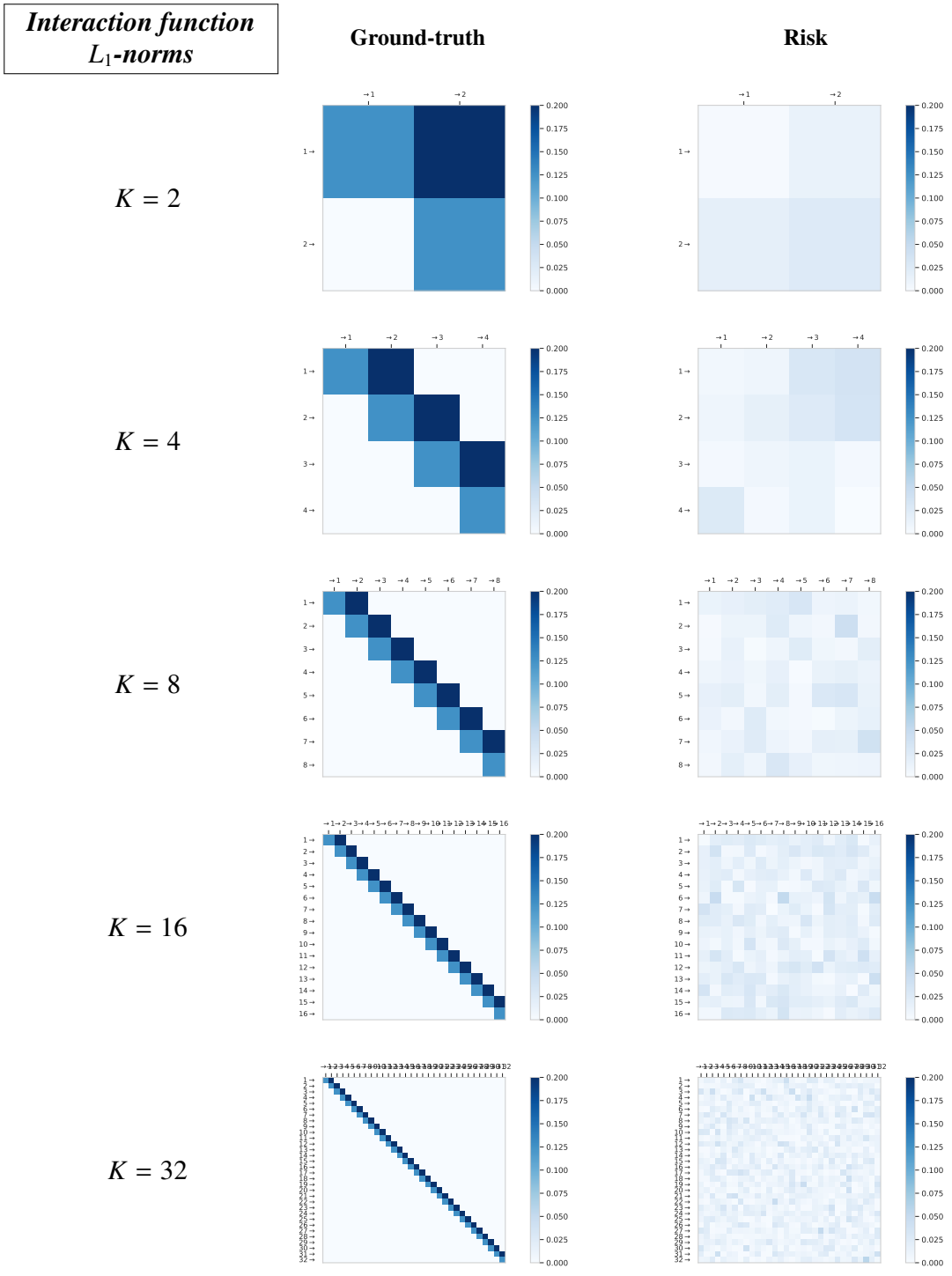
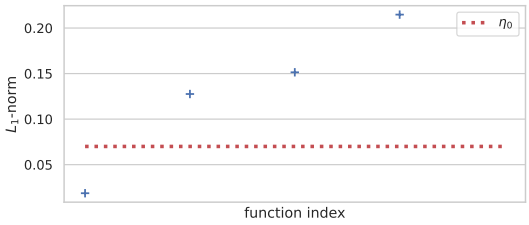
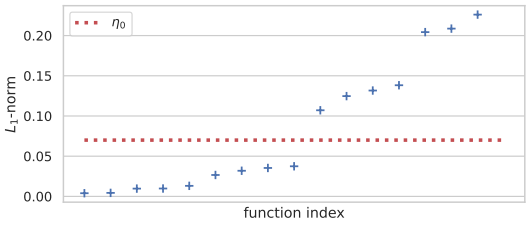


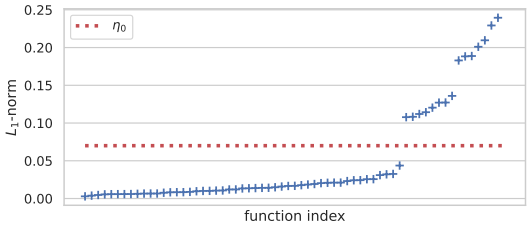
Figure 22: Heatmaps of the L_1-norms of the true parameter h_0, i.e., the entries of the matrix $S_0 = (S^0_{lk})_{l,k} = (\|h^0_{lk}\|_1)_{l,k}$ (left column) and L_1-risk, i.e., $(\mathbb{E}^Q[\|h^0_{lk} - h_{lk}\|_1])_{l,k}$ (right column) after the first step of Algorithm 3, in the Inhibition scenario of Simulation 4. The rows correspond to $K = 2, 4, 8, 16, 32$.



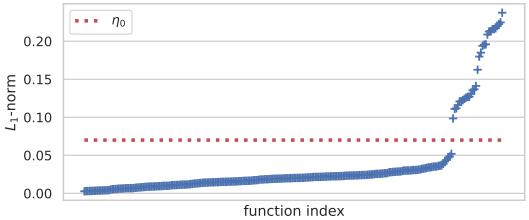
(a) $K = 2$



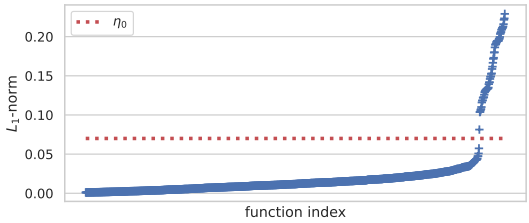
(b) $K = 4$



(c) $K = 8$



(d) $K = 16$



(e) $K = 32$

Figure 23: Estimated L_1-norms after the first step of Algorithm 3, plotted in increasing order, in the Inhibition scenario of Simulation 4, for the models with $K = 2, 4, 8, 16, 32$. The threshold in our algorithm $\eta_0 = 0.07$ is plotted in dotted red line.

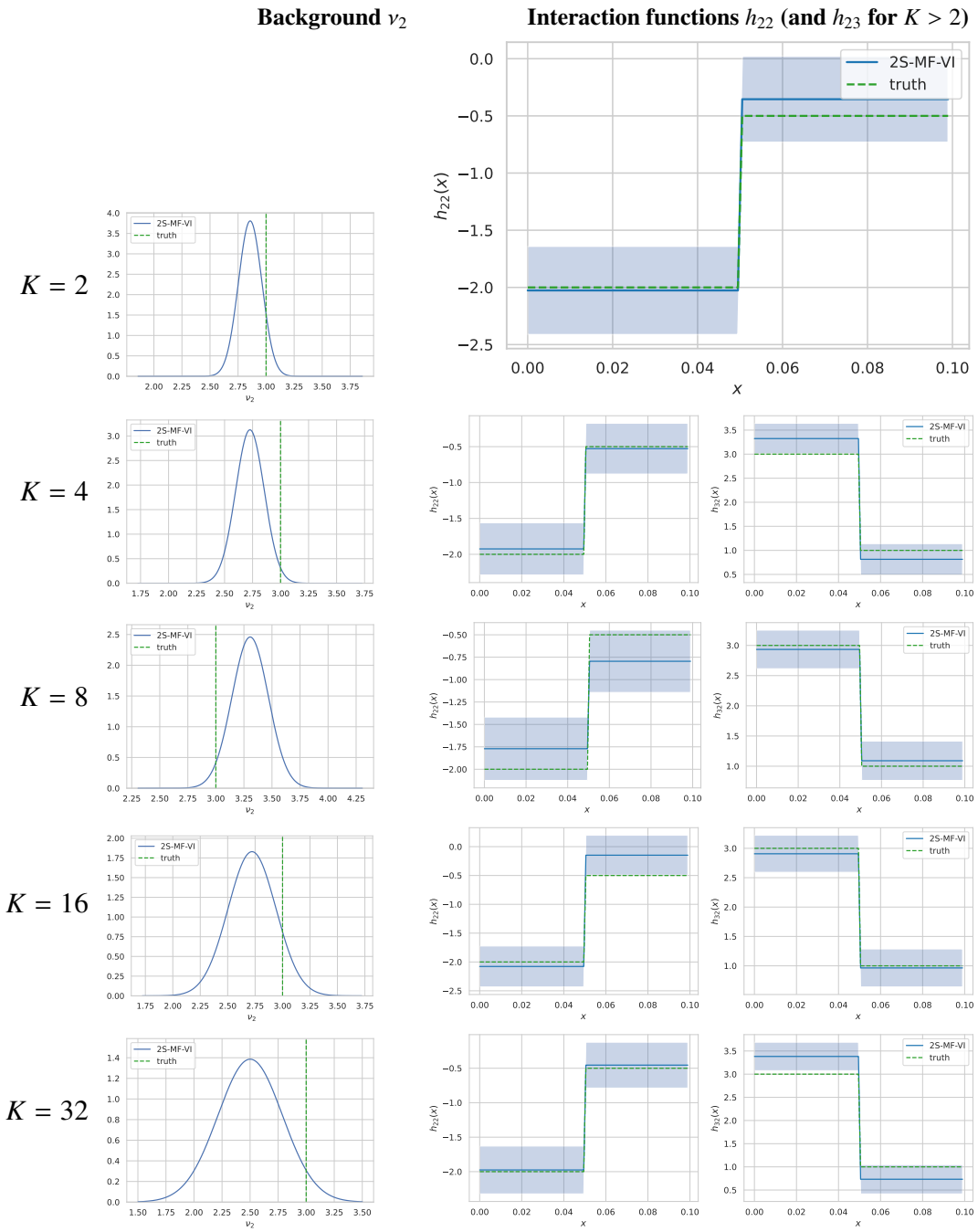


Figure 24: Mode variational posterior distributions on ν_2 (left column) and interaction functions h_{22} and h_{32} (for $K > 2$)(second and third columns) in the *Inhibition* scenario and multivariate sigmoid models of Simulation 4, computed with our two-step mean-field variational (2S-MF-VI) algorithm (Algorithm 3). The different rows correspond to different multivariate settings $K = 2, 4, 8, 16, 32$.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Scalable variational Bayes methods for Hawkes processes
Publication Status	□Published □ Accepted for Publication □Submitted for Publication x Unpublished and unsubmitted work written in a manuscript style
Publication Details	Joint work with Professor Judith Rousseau (University of Oxford) and Professor Vincent Rivoirard (Universite Paris-Dauphine).

Student Confirmation

Student Name:	Deborah Sulem		
Contribution to the Paper	I studied the asymptotic behaviour of the variational Bayes methods in the Hawkes model, proving concentration results on the variational posterior distribution. I proposed an adaptive algorithm with a computationally efficient variant that I implemented for the sigmoid model. I also tested and compared to MCMC methods, that I designed from the PyMC package.		
Signature *Deborah Sulem*		Date	08/11/2022

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Professor Judith Rousseau		
Supervisor comments		
Signature 	Date	

This completed form should be included in the thesis, at the end of the relevant chapter.

4 | Regularized spectral methods for clustering signed networks

This chapter corresponds to the following article:

> Cucuringu, M., Singh, A. V., Sulem, D., & Tyagi, H. (2021). Regularized spectral methods for clustering signed networks. Journal of Machine Learning Research, 22(264), 1-79.

Journal of Machine Learning Research volume (year) pages Submitted 11/20; Revised 05/21; Published date published

Regularized spectral methods for clustering signed networks

Mihai Cucuringu MIHAI.CUCURINGU@STATS.OX.AC.UK
Department of Statistics and Mathematical Institute
University of Oxford
The Alan Turing Institute, London, UK

Apoorv Vikram Singh APOORV.SINGH@NYU.EDU
Department of Computer Science and Engineering
New York University

Déborah Sulem DEBORAH.SULEM@STATS.OX.AC.UK
Department of Statistics
University of Oxford

Hemant Tyagi HEMANT.TYAGI@INRIA.FR
UMR 8524 - Laboratoire Paul Painlevé, F-59000
Inria, Univ. Lille, CNRS *

Editor: Qiaozhu Mei

Abstract

We study the problem of k-way clustering in signed graphs. Considerable attention in recent years has been devoted to analyzing and modeling signed graphs, where the affinity measure between nodes takes either positive or negative values. Recently, Cucuringu et al. (2019) proposed a spectral method, namely SPONGE (*Signed Positive over Negative Generalized Eigenproblem*), which casts the clustering task as a generalized eigenvalue problem optimizing a suitably defined objective function. This approach is motivated by social balance theory, where the clustering task aims to decompose a given network into disjoint groups, such that individuals within the same group are connected by as many positive edges as possible, while individuals from different groups are mainly connected by negative edges. Through extensive numerical experiments, SPONGE was shown to achieve state-of-the-art empirical performance. On the theoretical front, Cucuringu et al. (2019) analyzed SPONGE, as well as the popular Signed Laplacian based spectral method under the setting of a Signed Stochastic Block Model, for $k = 2$ equal-sized clusters, in the regime where the graph is moderately dense.

In this work, we build on the results in Cucuringu et al. (2019) on two fronts for the normalized versions of SPONGE and the Signed Laplacian. Firstly, for both algorithms, we extend the theoretical analysis in Cucuringu et al. (2019) to the general setting of $k \geq 2$ unequal-sized clusters in the moderately dense regime. Secondly, we introduce regularized versions of both methods to handle sparse graphs – a regime where standard spectral methods are known to underperform – and provide theoretical guarantees under the same setting of a Signed Stochastic Block Model. To the best of our knowledge, regularized spectral methods have so far not been considered in the setting of clustering signed graphs. We complement our theoretical results with an extensive set of numerical experiments on synthetic data, and three real world data sets standard in the signed networks literature.

* Authors are listed in alphabetical order.

Keywords: signed clustering, graph Laplacians, stochastic block models, spectral methods, regularization techniques, sparse graphs.

1. Introduction

Signed graphs. The recent years have seen a significant increase in interest for analysis of signed graphs, for tasks such as clustering (Chiang et al., 2014; Cucuringu et al., 2019), link prediction (Leskovec et al., 2010; Kumar et al., 2016) and visualization (Kunegis et al., 2010). Signed graphs are an increasingly popular family of undirected graphs, for which the edge weights may take both positive and negative values, thus encoding a measure of similarity or dissimilarity between the nodes. Signed social graphs have also received considerable attention to model trust relationships between entities, with positive (respectively, negative) edges encoding trust (respectively, distrust) relationships.

Clustering is arguably one of the most popular tasks in unsupervised machine learning, aiming at partitioning the node set such that the average connectivity or similarity between pairs of nodes within the same cluster is larger than that of pairs of nodes spanning different clusters. While the problem of clustering undirected unsigned graphs has been thoroughly studied for the past two decades (and to some extent, also that of clustering directed graphs in recent years), a lot less research has been undertaken on studying signed graphs.

Spectral clustering and regularization. Spectral clustering methods have become a fundamental tool with a broad range of applications in areas including network science, machine learning and data mining (von Luxburg, 2007). The attractivity of spectral clustering methods stems, on one hand, from its computational scalability by leveraging state-of-the-art eigensolvers, and on the other hand, from the fact that such algorithms are amenable to a theoretical analysis under suitably defined stochastic block models that quantify robustness to noise and sparsity of the measurement graph. Furthermore, on the theoretical side, understanding the spectrum of the adjacency matrix and its Laplacians, is crucial for the development of efficient algorithms with performance guarantees, and leads to a very mathematically rich set of problems. One such example from the latter class is that of Cheeger inequalities for general graphs, which relate the dominant eigenvalues of the Laplacian to edge expansion on graphs (Chung, 1996), extended to the setup of directed graphs (Chung, 2005), and more recently, to the graph Connection Laplacian arising in the context of the group synchronization problem (Bandeira et al., 2013), and higher-order Cheeger inequalities for multiway spectral clustering (Lee et al., 2014). There has been significant recent advances in theoretically analyzing spectral clustering methods in the context of stochastic block models; for a detailed survey, we refer the reader to the comprehensive recent survey of Abbe (2017).

In general, spectral clustering algorithms for unsigned and signed graphs typically have a common pipeline, where a suitable graph operator is considered (e.g., the graph Laplacian), its (usually k) extremal eigenvectors are computed, and the resulting point cloud in $\mathbb{R}^k$ is clustered using a variation of the popular k-means algorithm (Rohe et al., 2011). The main motivation for our current work stems from the lack of statistical guarantees in the above literature for the signed clustering problem, in the context of sparse graphs and large number of clusters $k \geq 3$. The problem of k-way clustering in signed graphs aims to find a partition of the node set into k disjoint clusters, such that most edges within clusters are positive,

while most edges across clusters are negative, thus altogether maximizing the number of *satisfied* edges in the graph. Another potential formulation to consider is to minimize the number of (*unsatisfied*) edges violating the partitions, i.e, the number of negative edges within clusters and positive edges across clusters.

A regularization step has been introduced in the recent literature motivated by the observation that properly regularizing the adjacency matrix A of a graph can significantly improve performance of spectral algorithms in the sparse regime. It was well known beforehand that standard spectral clustering often fails to produce meaningful results for sparse networks that exhibit strong degree heterogeneity (Amini et al., 2013; Jin, 2015). To this end, Chaudhuri et al. (2012) proposed the regularized graph Laplacian $L^\tau = D_\tau^{-1/2} A D_\tau^{-1/2}$, where $D_\tau = D + \tau I$, for $\tau \geq 0$. The spectral algorithm introduced and analyzed in Chaudhuri et al. (2012) splits the nodes into two random subsets and only relies on the subgraph induced by only one of the subsets to compute the spectral decomposition. Qin and Rohe (2013) studied the more traditional formulation of a spectral clustering algorithm that uses the spectral decomposition on the entire matrix (Ng et al., 2001), and proposed a regularized spectral clustering which they analyze. Subsequently, Joseph and Yu (2016) provided a theoretical justification for the regularization $A_\tau = A + \tau J$, where J denotes the all ones matrix, partly explaining the empirical findings of Amini et al. (2013) that the performance of regularized spectral clustering becomes insensitive for larger values of regularization parameters, and show that such large values can lead to better results. It is this latter form of regularization that we would be leveraging in our present work, in the context of clustering signed graphs. Additional references and discussion on the regularization literature are provided in Section 1.2.

Motivation & Applications. The recent surge of interest in analyzing signed graphs has been fueled by a very wide range of real-world applications, in the context of clustering, link prediction, and node rankings. Such social signed networks model trust relationships between users with positive (trust) and negative (distrust) edges. A number of online social services such as Epinions and Slashdot that allow users to express their opinions are naturally represented as signed social networks (Leskovec et al., 2010). Banerjee et al. (2012) considered shopping bipartite networks that encode like and dislike preferences between users and products. Other domain specific applications include personalized rankings via signed random walks (Jung et al., 2016), node rankings and centrality measures (Li et al., 2019), node classification (Tang et al., 2016), community detection (Yang et al., 2007; Chu et al., 2016), and anomaly detection, as in Kumar et al. (2014) which classifies users of an online signed social network as malicious or benign. In the very active research area of synthetic data generation, generative models for signed networks inspired by Structural Balance Theory have been proposed in Derr et al. (2018). Learning low-dimensional representations of graphs (network embeddings) have received tremendous attention in the recent machine learning literature, and graph convolutional networks-based methods have also been proposed for the setting of signed graphs, including Derr et al. (2018); Li et al. (2020), which provide network embeddings to facilitate subsequent downstream tasks, including clustering and link prediction.

A key motivation for our line of work stems from time series clustering (Aghabozorgi et al., 2015), an ubiquitous task arising in many applications that consider biological gene

expression data (Fujita et al., 2012), economic time series that capture macroeconomic variables (Focardi, 2005), and financial time series corresponding to large baskets of instruments in the stock market (Ziegler et al., 2010; Pavlidis et al., 2006). Driven by the clustering task, a popular approach in the literature is to consider similarity measures based on the Pearson correlation coefficient that captures linear dependence between variables and takes values in $[-1, 1]$. By construing the correlation matrix as a weighted network whose (signed) edge weights capture the pairwise correlations, we cluster the multivariate time series by clustering the underlying signed network. To increase robustness, tests of statistical significance are often applied to individual pairwise correlations, indicating the probability of observing a correlation at least as large as the measured sample correlation, assuming the null hypothesis is true. Such a thresholding step on the p-value associated to each individual sample correlation (Ha et al., 2015), renders the correlation network as a **sparse** matrix, which is one of the main motivations of our current work which proposes and analyzes algorithms for handling such sparse signed networks. We refer the reader to the popular work of Smith et al. (2011) for a detailed survey and comparison of various methodologies for turning time series data into networks, where the authors explore the interplay between fMRI time series and the network generation process. Importantly, they conclude that, in general, correlation-based approaches can be quite successful at estimating the connectivity of brain networks from fMRI time series.

Paper outline. This paper is structured as follows. The remainder of this Section 1 establishes the notation used throughout the paper, followed by a brief survey of related works in the signed clustering literature and graph regularization techniques for general graphs, along by a brief summary of our main contributions. Section 2 lays out the problem setup leading to our proposed algorithms in the context of the signed stochastic block model we subsequently analyze. Section 3 is a high-level summary of our main results across the two algorithms we consider. Section 4 contains the analysis of the proposed SPONGE$_{sym}$ algorithm, for both the sparse and dense regimes, for general number of clusters. Similarly, Section 5 contains the main theoretical results for the symmetric Signed Laplacian, under both sparsity regimes as well. Section 6 contains detailed numerical experiments on various synthetic and real world data sets, showcasing the performance of our proposed algorithms, as we vary the number of clusters, the relative cluster sizes, the sparsity regimes, and the regularization parameters. Finally, Section 7 is a summary and discussion of our main findings, with an outlook towards potential future directions. We defer to the Appendix additional proof details and a summary of the main technical tools used throughout.

1.1 Notation

We denote by $G = (V, E)$ a signed graph with vertex set V, edge set E, and adjacency matrix $A \in \{0, \pm 1\}^{n \times n}$. We will also refer to the unsigned subgraphs of positive (resp. negative) edges $G^+ = (V, E^+)$ (resp. $G^- = (V, E^-)$) with adjacency matrices A^+ (resp. A^-), such that $A = A^+ - A^-$. More precisely, $A^+_{ij} = \max\{A_{ij}, 0\}$ and $A^-_{ij} = \max\{-A_{ij}, 0\}$, with $E^+ \cap E^- = \emptyset$, and $E^+ \cup E^- = E$. We denote by $\overline{D} = D^+ + D^-$ the signed degree matrix, with the unsigned versions given by $D^+ := A^+\mathbb{1}$ and $D^- := A^-\mathbb{1}$. For a subset of nodes $C \subset V$, we denote its complement by $\overline{C} = V \setminus C$.

For a matrix $M \in \mathbb{R}^{m\times n}$, $\|M\|$ denotes its spectral norm $\|M\|_2$, i.e., its largest singular value, and $\|M\|_F$ denotes its Frobenius norm. When M is a $n \times n$ symmetric matrix, we denote $V_k(M)$ be the $n\times k$ matrix whose columns are given by the eigenvectors corresponding to the k smallest eigenvalues, and let $\mathcal{R}(V_k(M))$ denote the range space of these eigenvectors. We denote the eigenvalues of M by $(\lambda_j(M))_{j=1}^n$, with the ordering

$$\lambda_n(M) \leq \lambda_{n-1}(M) \leq \cdots \leq \lambda_1(M).$$

We also denote M_{i*} to be the i-th row of M. We denote $\mathbb{1} = (1, \ldots, 1)$ (resp. $\mathbb{1}_k$) the all ones column vector of size n (resp. k) and $\chi_1 = \frac{1}{\sqrt{k}}\mathbb{1}_k$. I_m denotes the square identity matrix of size m and is shortened to I when $m = n$. J_{mn} is the $m \times n$ matrix of all ones. Finally, for $a, b \geq 0$, we write $a \lesssim b$ if there exists a universal constant $C > 0$ such that $a \leq b$. If $a \lesssim b$ and $b \lesssim a$, then we write $a \asymp b$.

1.2 Related literature on signed clustering and graph regularization techniques

Signed clustering. There exists a very rich literature on algorithms developed to solve the k-way clustering problem, with spectral methods playing a central role in the developments of the last two decades. Such spectral techniques optimize an objective function via the eigen-decomposition of a suitably chosen graph operator (typically a graph Laplacian) built directly from the data, in order to obtain a low-dimensional embedding (most often of dimension k or $k-1$). A clustering algorithm such as k-means or k-means++ is subsequently applied in order to extract the final partition.

Kunegis et al. (2010) introduced the combinatorial Signed Laplacian $\overline{L} = \overline{D} - A$ for the 2-way clustering problem. For heterogeneous degree distributions, normalized extensions are generally preferred, such as the random-walk Signed Laplacian $\overline{L_{rw}} = I - \overline{D}^{-1}A$, and the symmetric Signed Laplacian $\overline{L_{sym}} = I - \overline{D}^{-1/2}A\overline{D}^{-1/2}$. Chiang et al. (2012) pointed out a weakness in the Signed Laplacian objective for k-way clustering with $k > 2$, and proposed instead a Balanced Normalized Cut (BNC) objective based on the operator $\overline{L_{BNC}} = \overline{D}^{-1/2}(D^+ - A)\overline{D}^{-1/2}$. Mercado et al. (2016) based their clustering algorithm on a new operator called the *Geometric Mean of Laplacians*, and later extended this method in (Mercado et al., 2019) to a family of operators called the *Matrix Power Mean of Laplacians*. Previous work (Cucuringu et al., 2019) by a subset of the authors of the present paper introduced the symmetric SPONGE objective using the matrix operator $T = (L^-_{sym} + \tau^+ I)^{-1/2}(L^+_{sym} + \tau^- I)(L^-_{sym} + \tau^+ I)^{-1/2}$, using the unsigned normalized Laplacians $L^{\pm}_{sym} = I - (D^{\pm})^{-1/2}A^{\pm}(D^{\pm})^{-1/2}$ and regularization parameters $\tau^+, \tau^- > 0$. This work also provides theoretical guarantees for the SPONGE and Signed Laplacian algorithms, in the setting of a Signed Stochastic Block Model.

Mercado et al. (2016) and Mercado et al. (2019) study the eigenspaces - in expectations and in probability - of several graph operators in a certain Signed Stochastic Block Model. However, this generative model differs from the one proposed in Cucuringu et al. (2019) that we analyze in this work. In the former, the positive and negative adjacency matrices do not have disjoint support, contrary to the latter. Moreover, their analysis is performed in the case of equal-size clusters. We will later show in our analysis that their result for the symmetric Signed Laplacian is not applicable in our setting.

Hsieh et al. (2012) proposed to perform low-rank matrix completion as a preprocessing step, before clustering using the top k eigenvectors of the completed matrix. For $k = 2$, Cucuringu (2015) showed that signed clustering can be cast as an instance of the group synchronization (Singer, 2011) problem over $\mathbb{Z}_2$, potentially with constraints given by available side information, for which spectral, semidefinite programming relaxations, and message passing algorithms have been considered. In recent work, Cucuringu et al. (2021) proposed a formulation for the signed clustering problem that relates to graph-based diffuse interface models utilizing the Ginzburg-Landau functionals, based on an adaptation of the classic numerical Merriman-Bence-Osher (MBO) scheme for minimizing such graph-based functionals (Merkurjev et al., 2014). We refer the reader to Gallier (2013) for a recent survey on clustering signed and unsigned graphs.

In a different line of work, known as *correlation clustering*, Bansal et al. (2004) considered the problem of clustering signed complete graphs, proved that it is NP-complete, and proposed two approximation algorithms with theoretical guarantees on their performance. On a related note, Demaine et al. (2006) studied the same problem but for arbitrary weighted graphs, and proposed an $\mathrm{O}(\log n)$ approximation algorithm based on linear programming. For correlation clustering, in contrast to k-way clustering, the number of clusters is not given in advance, and there is no normalization with respect to size or volume.

Regularization in the sparse regime. In many applications, real-world networks are sparse. In this context, regularization methods have increased the performance of traditional spectral clustering techniques, both for synthetic Stochastic Block Models and real data sets (Chaudhuri et al., 2012; Amini et al., 2013; Joseph and Yu, 2016; Le et al., 2015).

Chaudhuri et al. (2012) regularize the Laplacian matrix by adding a (typically small) weight τ to the diagonal entries of the degree matrix $L_\tau = I - D_\tau^{-1/2} A D_\tau^{-1/2}$ with $D_\tau = D + \tau I$. Amini et al. (2013) regularize the graph by adding a weight τ/n to every edge, leading to the Laplacian $\widetilde{L}_\tau = I - D_\tau^{-1/2} A_\tau D_\tau^{-1/2}$ with $A_\tau = A + \tau/n \mathbb{1}\mathbb{1}^T$ and $D_\tau = A_\tau \mathbb{1}$. Le et al. (2017) prove that this technique makes the adjacency and Laplacian matrices concentrate for inhomogeneous Erdős-Rényi graphs. Zhang and Rohe (2018) show that this technique prevents spectral clustering from overfitting through the analysis of dangling sets. In (Le et al., 2017), Le et al. propose a graph trimming method in order to reduce the degree of certain nodes. This is achieved by reducing the entries of the adjacency matrix that lead to high-degree vertices. Zhou and Amini (2018) add a spectral truncation step after this regularization method, and prove consistency results in the bipartite Stochastic Block Model.

Very recently, regularization methods using powers of the adjacency matrix have been introduced. Abbe et al. (2020) transform the adjacency matrix into the operator $A_r = \mathbb{1}\{(I + A)^r \geq 1\}$, where the indicator function is applied entrywise. With this method, spectral clustering achieves the fundamental limit for weak recovery in the sparse setting. Very similarly, Stephan and Massoulié (2019) transform the adjacency matrix into a distance matrix of outreach l, which links pairs of nodes that are l far apart w.r.t the graph distance.

1.3 Summary of our main contributions

This work extends the results obtained in Cucuringu et al. (2019) by a subset of the authors of our present paper. This previous work introduced the SPONGE algorithm, a princi-

pled and scalable spectral method for the signed clustering task that amounts to solving a generalized eigenvalue problem. Cucuringu et al. (2019) provided a theoretical analysis of both the newly introduced SPONGE algorithm and the popular Signed Laplacian-based method (Kunegis et al., 2010), quantifying their robustness against the sampling sparsity and noise level, under the setting of a Signed Stochastic Block Model (SSBM). These were the first such theoretical guarantees for the signed clustering problem under a suitably defined stochastic graph model. However, the analysis in Cucuringu et al. (2019) was restricted to the setting of two equally-sized clusters, which is less realistic in light of most real world applications. Furthermore, the same previous line of work considered the moderately dense regime in terms of the edge sampling probability p, in particular, it operated in the setting where $\mathbb{E}[\overline{D}_{jj}] \gtrsim \ln n$, i.e., $p \gtrsim \frac{\ln n}{n}$. Many real world applications involve large but very sparse graphs, with $p = \Theta\left(\frac{1}{n}\right)$, which provides motivation for our present work.

We summarize below our main contributions, and start with the remark that the theoretical analysis in the present paper pertains to the normalized version of SPONGE (denoted as SPONGE$_{sym}$) and the symmetric Signed Laplacian, while Cucuringu et al. (2019) analyzed only the un-normalized versions of these signed operators. The experiments reported in Cucuringu et al. (2019) also consider such normalized matrix operators, and show their superior performance over their respective un-normalized versions, further providing motivational ground for our current work.

(i) Our first main contribution is to analyze the two above-mentioned signed operators, namely SPONGE$_{sym}$ and the symmetric Signed Laplacian, in the general SSBM model with $k \geq 2$ and unequal-cluster sizes, in the moderately dense regime. In particular, we evaluate the accuracy of both signed clustering algorithms by bounding the mis-clustering rate of the entire pipelines, as achieved by the popular k-means algorithm.

(ii) Our second contribution is to introduce and analyze new regularized versions of both SPONGE$_{sym}$ and the symmetric Signed Laplacian, under the same general SSBM model, but in the sparse graph regime $\mathbb{E}[\overline{D}_{jj}] \gtrsim 1$, a setting where standard spectral methods are known to underperform. To the best of our knowledge, this sparsity regime has not been previously considered in the literature of signed networks; such regularized spectral methods have so far not been considered in the setting of clustering signed networks, or more broadly in the signed networks literature, where such regularization could prove useful for other related downstream tasks. One important aspect of regularization techniques is the choice of the regularization parameters. We show that our proposed algorithms can benefit from careful regularization and attain a higher level of accuracy in the sparse regime, provided that the regularization parameters scale as an adequate power of the average degree in the graph. These findings are supported by our experiments on real-world datasets.

2. Problem setup

This section details the two algorithms for the signed clustering problem that we will analyze subsequently, namely, SPONGE$_{sym}$(Symmetric Signed Positive Over Negative Generalized

Eigenproblem) and the symmetric Signed Laplacian, along with their respective regularized versions.

2.1 Clustering via the SPONGE$_{sym}$ algorithm

The symmetric SPONGE method, denoted as SPONGE$_{sym}$, aims at jointly minimizing two measures of badness in a signed clustering problem. For an unsigned graph G and $X, Y \subset V$, we define the cut function $\mathrm{Cut}_G(X,Y) := \sum_{i\in X, j\in Y} A_{ij}$, and denote the volume of X by $\mathrm{Vol}_G(X) := \sum_{i\in X}\sum_{j=1}^{n} A_{ij}$.

For a given cluster set $C \subset V$, $\mathrm{Cut}_G(C,\overline{C})$ is the total weight of edges crossing from C to $\overline{C}$ and $\mathrm{Vol}_G(C)$ is the sum of (weighted) degrees of nodes in C. With this notation in mind and motivated by the approach of Cucuringu et al. (2016) in the context of constrained clustering, the symmetric SPONGE algorithm for signed clustering aims at minimizing the following two measures of *badness* given by $\frac{\mathrm{Cut}_{G^+}(C,\overline{C})}{\mathrm{Vol}_{G^+}(C)}$ and $\left(\frac{\mathrm{Cut}_{G^-}(C,\overline{C})}{\mathrm{Vol}_{G^-}(C)}\right)^{-1} = \frac{\mathrm{Vol}_{G^-}(C)}{\mathrm{Cut}_{G^-}(C,\overline{C})}$. To this end, we consider "merging" the objectives, and aim to solve

$$\min_{C\subset V} \frac{\frac{\mathrm{Cut}_{G^+}(C,\overline{C})}{\mathrm{Vol}_{G^+}(C)} + \tau^-}{\frac{\mathrm{Cut}_{G^-}(C,\overline{C})}{\mathrm{Vol}_{G^-}(C)} + \tau^+},$$

where $\tau^+ > 0, \tau^- \geq 0$ denote trade-off parameters. For k-way signed clustering into disjoint clusters $C_1, \ldots, C_k$, we arrive at the combinatorial optimization problem

$$\min_{C_1,\ldots,C_k} \sum_{i=1}^{k} \left(\frac{\frac{\mathrm{Cut}_{G^+}(C_i,\overline{C_i})}{\mathrm{Vol}_{G^+}(C_i)} + \tau^-}{\frac{\mathrm{Cut}_{G^-}(C_i,\overline{C_i})}{\mathrm{Vol}_{G^-}(C_i)} + \tau^+} \right). \tag{1}$$

Let D^+, L^+ denote respectively the degree matrix and un-normalized Laplacian associated with G^+, and $L^+_{sym} = (D^+)^{-1/2} L^+ (D^+)^{-1/2}$ denote the symmetric Laplacian matrix for G^+ (similarly for L^-_{sym}, D^-, L^-). For a subset $C_i \subset V$, denote $\mathbb{1}_{C_i}$ to be the indicator vector for C_i so that $(\mathbb{1}_{C_i})_j$ equals 1 if $j \in C_i$, and is 0 otherwise. Now define the normalized indicator vector $x_{C_i} \in \mathbb{R}^n$ where

$$x_{C_i} = \left(\frac{\mathrm{Cut}_{G^-}(C_i,\overline{C_i})}{\mathrm{Vol}_{G^-}(C_i)} + \tau^+ \right)^{-1/2} \frac{1}{\sqrt{\mathrm{Vol}_{G^+}(C_i)}} (D^+)^{1/2} \mathbb{1}_{C_i}.$$

In light on this, one can verify that

$$\begin{aligned}
x_{C_i}^\top x_{C_i} &= \left(\frac{\mathrm{Cut}_{G^-}(C_i,\overline{C_i})}{\mathrm{Vol}_{G^-}(C_i)} + \tau^+ \right)^{-1} \frac{\mathbb{1}_{C_i}^\top D^+ \mathbb{1}_{C_i}}{\mathrm{Vol}_{G^+}(C_i)} = \left(\frac{\mathrm{Cut}_{G^-}(C_i,\overline{C_i})}{\mathrm{Vol}_{G^-}(C_i)} + \tau^+ \right)^{-1}, \\
x_{C_i}^\top L^+_{sym} x_{C_i} &= \left(\frac{\mathrm{Cut}_{G^-}(C_i,\overline{C_i})}{\mathrm{Vol}_{G^-}(C_i)} + \tau^+ \right)^{-1} \frac{\mathbb{1}_{C_i}^\top L^+ \mathbb{1}_{C_i}}{\mathrm{Vol}_{G^+}(C_i)} \\
&= \left(\frac{\mathrm{Cut}_{G^-}(C_i,\overline{C_i})}{\mathrm{Vol}_{G^-}(C_i)} + \tau^+ \right)^{-1} \frac{\mathrm{Cut}_{G^+}(C_i,\overline{C_i})}{\mathrm{Vol}_{G^+}(C_i)}.
\end{aligned}$$

Hence (1) is equivalent to the following discrete optimization problem

$$\min_{C_1,\ldots,C_k} \sum_{i=1}^{k} x_{C_i}^\top (L_{sym}^+ + \tau^- I) x_{C_i} \tag{2}$$

which is NP-Hard. A common approach to solve this problem is to drop the discreteness constraints, and allow x_{C_i} to take values in $\mathbb{R}^n$. To this end, we introduce a new set of vectors $z_1, \ldots, z_k \in \mathbb{R}^n$ such that they are orthonormal with respect to the matrix $L_{sym}^- + \tau^+ I$, i.e., $z_i^\top (L_{sym}^- + \tau^+ I) z_{i'} = \delta_{ii'}$. This leads to the continuous optimization problem

$$\min_{z_i^\top (L_{sym}^- + \tau^+ I) z_{i'} = \delta_{ii'}} \sum_{i=1}^{k} z_i^\top (L_{sym}^+ + \tau^- I) z_i. \tag{3}$$

Note that the above choice of vectors $z_1, ..., z_k$ is not really a relaxation of (2) since $x_{C_1}, \ldots, x_{C_k}$ are not necessarily $(L_{sym}^- + \tau^+ I)$-orthonormal, but (3) can be conveniently formulated as a suitable generalized eigenvalue problem, similar to the approach in Cucuringu et al. (2016). Indeed, denoting $y_i = (L_{sym}^- + \tau^+ I)^{1/2} z_i$, and $Y = [y_1, \ldots, y_k] \in \mathbb{R}^{n \times k}$, (3) can be rewritten as

$$\min_{Y^\top Y = I} \mathrm{Tr}\Big(Y^\top (L_{sym}^- + \tau^+ I)^{-1/2} (L_{sym}^+ + \tau^- I)(L_{sym}^- + \tau^+ I)^{-1/2} Y \Big),$$

the solution to which is well known to be given by the smallest k eigenvectors of

$$T = (L_{sym}^- + \tau^+ I)^{-1/2} (L_{sym}^+ + \tau^- I)(L_{sym}^- + \tau^+ I)^{-1/2},$$

see for e.g. (Sameh and Tong, 2000, Theorem 2.1). However this is not practically viable for large scale problems, since computing T itself is already expensive. To circumvent this issue, one can instead consider the embedding in $\mathbb{R}^k$ corresponding to the smallest k generalized eigenvectors of the symmetric definite pair $(L_{sym}^+ + \tau^- I, L_{sym}^- + \tau^+ I)$. There exist many efficient solvers for solving large scale generalized eigenproblems for symmetric definite matrix pairs. In our experiments, we use the LOBPCG (Locally Optimal Block Preconditioned Conjugate Gradient method) solver introduced in Knyazev (2001).

One can verify that (λ, v) is an eigenpair[1] of T iff $(\lambda, (L_{sym}^- + \tau^+ I)^{-1/2} v)$ is a generalized eigenpair of $(L_{sym}^+ + \tau^- I, L_{sym}^- + \tau^+ I)$. Indeed, for symmetric matrices A, B with $A \succ 0$, it holds true for $w = A^{-1/2} v$ that

$$A^{-1/2} B A^{-1/2} v = \lambda v \iff Bw = \lambda A w.$$

Therefore, denoting $V_k(T) \in \mathbb{R}^{n \times k}$ to be the matrix consisting of the smallest k eigenvectors of T, and $G_k(T) \in \mathbb{R}^{n \times k}$ to be the matrix of the smallest k generalized eigenvectors of $(L_{sym}^+ + \tau^- I, L_{sym}^- + \tau^+ I)$, it follows that

$$G_k(T) = (L_{sym}^- + \tau^+ I)^{-1/2} V_k(T). \tag{4}$$

Hence upon computing $G_k(T)$, we will apply a suitable clustering algorithm on the rows of $G_k(T)$ such as the popular k-means++ (Arthur and Vassilvitskii, 2007), to arrive at the final partition.

[1] With λ denoting its eigenvalue, and v the corresponding eigenvector.

Remark 1 *In Cucuringu et al. (2019), similar arguments as above were shown for the SPONGE algorithm which led to computing the k smallest generalized eigenvectors of the matrix pair* $(L^+ + \tau^- D^-, L^- + \tau^+ D^+)$. SPONGE$_{sym}$ *was proposed in Cucuringu et al. (2019) but no theoretical results were provided.*

Clustering in the sparse regime. We also provide a version of SPONGE$_{sym}$ for the case where G is sparse, i.e., the graph has very few edges and is typically disconnected. In this setting, we consider a regularized version of SPONGE$_{sym}$ wherein a weight is added to each edge (including self-loops) of the positive and negative subgraphs, respectively. Formally, for regularization parameters $\gamma^+, \gamma^- \geq 0$, let us define $A^{\pm}_{\gamma^\pm} := A^{\pm} + \frac{\gamma^\pm}{n}\mathbb{1}\mathbb{1}^\top$ to be the regularized adjacency matrices for the unsigned graphs G^+, G^- respectively. Denoting $D^{\pm}_{\gamma^\pm}$ to be the degree matrix of $A^{\pm}_{\gamma^\pm}$, the normalized Laplacians corresponding to $A^{\pm}_{\gamma^\pm}$ are given by

$$L^{\pm}_{sym,\gamma^\pm} = I - (D^{\pm}_{\gamma^\pm})^{-1/2} A^{\pm}_{\gamma^\pm} (D^{\pm}_{\gamma^\pm})^{-1/2}.$$

Given the above modifications, let $V_k(T_{\gamma^+,\gamma^-}) \in \mathbb{R}^{n\times k}$ denote the matrix consisting of the smallest k eigenvectors of

$$T_{\gamma^+,\gamma^-} = (L^-_{sym,\gamma^-} + \tau^+ I)^{-1/2} (L^+_{sym,\gamma^+} + \tau^- I)(L^-_{sym,\gamma^-} + \tau^+ I)^{-1/2}\,.$$

For the same reasons discussed earlier, we will consider the embedding given by the smallest k generalized eigenvectors of the matrix pencil $(L^+_{sym,\gamma^+} + \tau^- I, L^-_{sym,\gamma^-} + \tau^+ I)$, namely $G_k(T_{\gamma^+,\gamma^-})$ where

$$G_k(T_{\gamma^+,\gamma^-}) = (L^-_{sym,\gamma^-} + \tau^+ I)^{-1/2} V_k(T_{\gamma^+,\gamma^-}),$$

as in (44). The rows of $G_k(T_{\gamma^+,\gamma^-})$ can then be clustered using an appropriate clustering procedure, such as k-means++.

Remark 2 *Regularized spectral clustering for unsigned graphs involves adding* $\frac{\gamma}{n}\mathbb{1}\mathbb{1}^\top$ *to the adjacency matrix, followed by clustering the embedding given by the smallest k eigenvectors of the normalized Laplacian (of the regularized adjacency), see for e.g. Amini et al. (2013); Le et al. (2017). To the best of our knowledge, regularized spectral clustering methods have not been explored thus far in the context of sparse signed graphs.*

2.2 Clustering via the symmetric Signed Laplacian

The rationale behind the use of the (un-normalized) Signed Laplacian $\overline{L}$ for clustering is justified by Kunegis et al. (2010) using the signed ratio cut function. For $C \subset V$,

$$sRCut(C, \overline{C}) = \left(2\mathrm{Cut}_{G^+}(C, \overline{C}) + \mathrm{Cut}_{G^-}(C, C) + \mathrm{Cut}_{G^-}(\overline{C}, \overline{C})\right)\left(\frac{1}{|C|} + \frac{1}{|\overline{C}|}\right). \quad (5)$$

For 2-way clustering, minimizing this objective corresponds to minimizing the number of positive edges between the two classes and the number of negative edges inside each class. Moreover, (5) is equivalent to the following optimization problem

$$\min_{u\in\mathcal{U}} u^\top \overline{L} u,$$

where $\mathcal{U} \in \mathbb{R}^n$ is the set of vectors of the form $\forall i \in [n], u_i = \pm\frac{1}{2}\left(\sqrt{\frac{|C|}{|\overline{C}|}} + \sqrt{\frac{|\overline{C}|}{|C|}}\right)$.

However, Gallier (2016) noted that this equivalence does not generalize to $k > 2$, and defined a new notion of signed cut, called the signed normalized cut function. For a partition $C_1, \dots, C_k$ with membership matrix $X \in \{0,1\}^{n\times k}$,

$$sNCut(C_1, \dots, C_k) = \sum_{i=1}^{k} \frac{\mathrm{Cut}_G(C_i, \overline{C_i})}{\mathrm{Vol}_G(C_i)} + 2\frac{\mathrm{Cut}_{G-}(C_i, C_i)}{\mathrm{Vol}_G(C_i)} = \sum_{i=1}^{k} \frac{(X^i)^\top \overline{L} X^i}{(X^i)^\top \overline{D} X^i},$$

with X^i the i-th column of X. Compared to (5), this objective also penalizes the number of negative edges across two subsets, which may not be a desirable feature for signed clustering. Minimizing this function with a relaxation of the constraint that $X^i \in \{0,1\}^n$ leads to the following problem

$$\min_{Y^\top Y = I} \mathrm{Tr}\Big(Y^\top \overline{L_{sym}} Y\Big).$$

The minimum of this problem is obtained by stacking column-wise the k eigenvectors of $\overline{L_{sym}}$ corresponding to the smallest eigenvalues, *i.e.* $V_k(\overline{L_{sym}})$. Therefore, one can apply a clustering algorithm to the rows of the matrix $V_k(\overline{L_{sym}})$ to find a partition of the set of nodes V.

In fact, we will consider using only the $k-1$ smallest eigenvectors of $\overline{L_{sym}}$ and applying the k-means++ algorithm on the rows of $V_{k-1}(\overline{L_{sym}})$. This will be justified in our analysis via a stochastic generative model, namely the Signed Stochastic Block Model (SSBM), introduced in the next subsection. Under this model assumption, we will see later that the embedding given by the $k-1$ smallest eigenvectors of the symmetric Signed Laplacian of the expected graph has k distinct rows (with two rows being equal if and only if the corresponding nodes belong to the same cluster).

Clustering in the sparse regime. When G is sparse, we propose a spectral clustering method based on a regularization of the signed graph, leading to a regularized Signed Laplacian. To this end, for $\gamma^+, \gamma^- \geq 0$, recall the regularized adjacency matrices $A^{\pm}_{\gamma^\pm}$, with degree matrices $D^{\pm}_{\gamma^\pm}$, for the unsigned graphs G^+, G^- respectively. In light of this, the regularized signed adjacency and degree matrices are defined as follows

$$\begin{aligned} A_\gamma &:= A^+_{\gamma^+} - A^-_{\gamma^-} = A + \frac{\gamma^+ - \gamma^-}{n}\mathbb{1}\mathbb{1}^\top, \\ \overline{D}_\gamma &:= D^+_{\gamma^+} + D^-_{\gamma^-} = D^+ + \gamma^+ I + D^- + \gamma^- I = \overline{D} + (\gamma^+ + \gamma^-)I = \overline{D} + \gamma I, \end{aligned}$$

with $\gamma := \gamma^+ + \gamma^-$. Our regularized Signed Laplacian is the symmetric Signed Laplacian on this regularized signed graph, i.e.

$$L_\gamma := I - (\overline{D}_\gamma)^{-1/2} A_\gamma (\overline{D}_\gamma)^{-1/2}. \tag{6}$$

Similarly to the symmetric Signed Laplacian, our clustering algorithm in the sparse case finds the $k-1$ smallest eigenvectors of L_γ and applies the k-means algorithm on the rows of $V_{k-1}(L_\gamma)$.

Remark 3 *For the choice $\gamma^+ = \gamma^-$, the regularized Laplacian becomes*

$$L_\gamma := I - (\overline{D}_\gamma)^{-1/2} A (\overline{D}_\gamma)^{-1/2},$$

with $\overline{D}_\gamma = \overline{D} + (\gamma^+ + \gamma^-)I$. This regularization scheme is very similar to the degree-corrected normalized Laplacian defined in Chaudhuri et al. (2012).

2.3 Signed Stochastic Block Model (SSBM)

Our work theoretically analyzes the clustering performance of SPONGE$_{sym}$ and the symmetric Signed Laplacian algorithms under a signed random graph model, also considered previously in (Cucuringu et al., 2019; Cucuringu et al., 2021). We recall here its definition and parameters.

- n: the number of nodes in network;
- k: the number of planted communities;
- p: the probability of an edge to be present;
- η: the probability of flipping the sign of an edge;
- $C_1, \ldots, C_k$: an arbitrary partition of the vertices with sizes $n_1, \ldots, n_k$.

We first partition the vertices (arbitrarily) into clusters $C_1, \ldots, C_k$ where $|C_i| = n_i$. Next, we generate a *noiseless* measurement graph from the Erdős-Rényi model $G(n, p)$, wherein each edge takes value $+1$ if both its endpoints are contained in the same cluster, and -1 otherwise. To model noise, we flip the sign of each edge independently with probability $\eta \in [0, 1/2)$. This results in the realization of a signed graph instance G from the SSBM ensemble.

Let $A \in \{0, \pm 1\}^{n \times n}$ denote the adjacency matrix of G, and note that $(A_{jj'})_{j \le j'}$ are independent random variables. Recall that $A = A^+ - A^-$, where $A^+, A^- \in \{0, 1\}^{n \times n}$ are the adjacency matrices of the unsigned graphs G^+, G^- respectively. Then, $(A^+_{jj'})_{j \le j'}$ are independent, and similarly $(A^-_{jj'})_{j \le j'}$ are also independent. But for given $j, j' \in [n]$ with $j \neq j'$, $A^+_{jj'}$ and $A^-_{jj'}$ are dependent. Let $d_i^\pm$ denote the degree of a node in cluster i, for $i \in [k]$ in the graph $\mathbb{E}[A^\pm]$. Moreover, under this model, the expected signed degree matrix is the scaled identity matrix $\mathbb{E}\overline{D} = \overline{d} I$, with $\overline{d} = p(n-1)$.

Remark 4 *Contrary to stochastic block models for unsigned graphs, we do not require (for the purpose of detecting clusters) that the intra-cluster edge probabilities to be different from those of inter-cluster edges, since the sign of the edges already achieves this purpose implicitly. In fact, it is the noise parameter η that is crucial for identifying the underlying latent cluster structure.*

To formulate our theoretical results we will also need the following notations. Let $s_i = n_i/n$ denote the fraction of nodes in cluster i, with l (resp. s) denoting the fraction for the largest (resp. smallest) cluster. Hence, the size of the largest (resp. smallest) cluster is nl (resp. ns). Following the notation in Lei and Rinaldo (2015), we will denote $\mathbb{M}_{n,k}$

to be the class of "membership" matrices of size $n \times k$, and denote $\hat{\Theta} \in \mathbb{M}_{n,k}$ to be the ground-truth membership matrix containing k distinct indicator row-vectors (one for each cluster), i.e., for $i \in [k]$ and $j \in [n]$,

$$\hat{\Theta}_{ji} = \begin{cases} 1 & \text{if node } j \in \text{ cluster } C_i, \\ 0 & \text{otherwise.} \end{cases}$$

We also define the normalized membership matrix Θ corresponding to $\hat{\Theta}$, where for $i \in [k]$ and $j \in [n]$,

$$\Theta_{ji} = \begin{cases} 1/\sqrt{n_i} & \text{if node } j \in \text{ cluster } C_i, \\ 0 & \text{otherwise.} \end{cases}$$

3. Summary of main results

We now summarize our theoretical results for SPONGE$_{sym}$ and the symmetric Signed Laplacian methods, when the graph is generated from the SSBM ensemble.

3.1 Symmetric SPONGE

We begin by describing conditions under which the rows of the matrix $G_k(T)$ approximately preserve the ground truth clustering structure. Before explaining our results, let us denote the matrix $\overline{T}$ to be the analogue of T for the expected graph, i.e.,

$$\overline{T} = (\overline{L^-_{sym}} + \tau^+ I)^{-1/2} (\overline{L^+_{sym}} + \tau^- I)(\overline{L^-_{sym}} + \tau^+ I)^{-1/2},$$

where $\overline{L^{\pm}_{sym}} = I - (\mathbb{E}[D^{\pm}])^{-1/2}\,\mathbb{E}[A^{\pm}](\mathbb{E}[D^{\pm}])^{-1/2}$. We first show that for suitable values of $\tau^+ > 0, \tau^- \geq 0$ (with n large enough), the smallest k eigenvectors of $\overline{T}$, denoted by $V_k(\overline{T})$, are given by $V_k(\overline{T}) = \Theta R$, for some $k \times k$ rotation matrix R. Hence, the rows of $V_k(\overline{T})$ have the same clustering structure as that of Θ. Denoting $G_k(\overline{T}) \in \mathbb{R}^{n \times k}$ to be the matrix consisting of the k smallest generalized eigenvectors of $(\overline{L^+_{sym}} + \tau^- I, \overline{L^-_{sym}} + \tau^+ I)$, and recalling (4), we can relate $G_k(\overline{T})$ and $V_k(\overline{T})$ via

$$G_k(\overline{T}) = (\overline{L^-_{sym}} + \tau^+ I)^{-1/2} V_k(\overline{T}). \tag{7}$$

It turns out that when $V_k(\overline{T}) = \Theta R$, and in light of the expression for $\overline{L^-_{sym}} + \tau^+ I$ from (24), we arrive at $G_k(\overline{T}) = \Theta (C^-)^{-1/2} R$, where $C^- \succ 0$ is as in (18). Since $(C^-)^{-1/2} R$ is invertible, it follows that $G_k(\overline{T})$ has k distinct rows, with the rows that belong to the same cluster being identical. The remaining arguments revolve around deriving concentration bounds on $\left\|T - \overline{T}\right\|$, which imply (for p large enough) that the distance between the column spans of $V_k(T)$ and $V_k(\overline{T})$ is small, i.e., there exists an orthonormal matrix O such that $\left\|V_k(T) - V_k(\overline{T})O\right\|$ is small. Finally, the expressions in (4) and (7) altogether imply that $\left\|G_k(T) - G_k(\overline{T})O\right\|$ is small, which is an indication that the rows of $G_k(T)$ approximately preserve the clustering structure encoded in Θ.

The above discussion is summarized in the following theorem, which is our first main result for SPONGE$_{sym}$ in the moderately dense regime.

Theorem 5 (Restating Theorem 29) (Eigenspace alignment of SPONGE$_{sym}$ in the dense case) *Assuming* $n \geq \max\left\{\frac{2(1-\eta)}{s(1-2\eta)}, \frac{2\eta}{(1-l)(1-\eta)}\right\}$, *suppose that* $\tau^+ > 0, \tau^- \geq 0$ *are chosen to satisfy*

$$\tau^+ > \frac{16\eta}{\beta s(1-2\eta)}, \qquad \tau^- < \frac{\beta}{2}\left(\frac{s(1-2\eta)}{s(1-2\eta)+2\eta}\right)\min\left\{\frac{1}{4(1-\beta)}, \frac{\tau^+}{8}\right\}$$

where β, η *satisfy one of the following conditions*

1. $\beta = \frac{4\eta}{s(1-2\eta)+4\eta}$ *and* $0 < \eta < \frac{1}{2}$, *or*
2. $\beta = \frac{1}{2}$ *and* $\eta \leq \frac{s}{2s+4}$.

Then $V_k(\overline{T}) = \Theta R$ *and* $G_k(\overline{T}) = \Theta(C^-)^{-1/2}R$, *where* R *is a rotation matrix, and* $C^- \succ 0$ *is as defined in* (18). *Moreover, for any* $\varepsilon, \delta \in (0,1)$, *there exists a constant* $\widetilde{c}_\varepsilon > 0$ *such that the following is true. If* p *satisfies*

$$p \geq \max\left\{\widetilde{c}_\varepsilon C_2(s,\eta,l), \frac{256C_1^4(\tau^+,\tau^-)(2+\tau^+)^4}{\delta^4(1+\tau^-)^4(1-\beta)^4}C_2(s,\eta,l), \frac{81}{(1-l)\delta^4}\right\}\frac{\ln(4n/\varepsilon)}{n}$$

with $C_1(\cdot), C_2(\cdot)$ *as in* (45), *then with probability at least* $1-2\varepsilon$, *there exists an orthogonal matrix* $O \in \mathbb{R}^{k\times k}$ *such that*

$$\left\|V_k(T) - V_k(\overline{T})O\right\| \leq \delta, \qquad \textit{and} \qquad \left\|G_k(T) - G_k(\overline{T})O\right\| \leq \frac{\delta}{\sqrt{\tau^+}} + \frac{\delta}{(\tau^+)^2}.$$

Let us now interpret the scaling of the terms n, p, τ^+ and τ^- in Theorem 5, and provide some intuition.

1. In general, when no assumption is made on the noise level η, we have $\beta = \frac{4\eta}{s(1-2\eta)+4\eta}$ and the requirement on n is $n \gtrsim \max\left\{\frac{1}{s(1-2\eta)}, \frac{\eta}{1-l}\right\}$. Then a sufficient set of conditions on $\tau^+ > 0, \tau^- \geq 0$ are

$$\tau^+ \gtrsim 1 + \frac{\eta}{s(1-2\eta)}, \quad \tau^- \lesssim \frac{\eta}{s(1-2\eta)+2\eta}. \tag{8}$$

   Moreover, we see from (45) that $C_1(\tau^+,\tau^-) \lesssim 1/\tau^+$, and thus $\frac{(2+\tau^+)C_1(\tau^+,\tau^-)}{1+\tau^-} \lesssim 1$. Hence, a sufficient condition on p is

$$p \gtrsim \frac{1}{\delta^4}\left(1 + \frac{\eta}{s(1-2\eta)}\right)^4 C_2(s,\eta,l)\frac{\ln n}{n}.$$

2. In the "low-noise" regime where $\eta \leq \frac{s}{2s+4}$, the condition on τ^- in (8) becomes strict, especially as $\eta \to 0$. In this regime, the second condition in Theorem 5 allows for a wider range of values for τ^-; in particular, the following set of conditions suffice

$$\tau^+ \gtrsim 1, \quad \tau^- \lesssim \frac{s(1-2\eta)}{s(1-2\eta)+2\eta}.$$

   Moreover, we then obtain that the condition $p \gtrsim \frac{1}{\delta^4}C_2(s,\eta,l)\frac{\ln n}{n}$ is sufficient.

3. When $\tau^+ \to \infty$, then $\left\|G_k(T) - G_k(\overline{T})O\right\| \to 0$, which might lead one to believe that the clustering performance improves accordingly. This is not the case however, since when τ^+ is large, then $G_k(T) \approx \frac{1}{\sqrt{\tau^+}} V_k(T)$ and $G_k(\overline{T}) \approx \frac{1}{\sqrt{\tau^+}} V_k(\overline{T})$, which means that clustering the rows of $G_k(T)$ (resp. $G_k(\overline{T})$) is roughly equivalent to clustering the rows of $V_k(T)$ (resp. $V_k(\overline{T})$). Moreover, note that for large τ^+, we have $T \approx \frac{1}{\tau^+}(L^+_{sym} + \tau^- I)$ and $\overline{T} \approx \frac{1}{\tau^+}(\overline{L^+_{sym}} + \tau^- I)$ and thus the negative subgraph has no effect on the clustering performance.

SPONGE$_{sym}$ in the sparse regime. Notice that the above theorem required the sparsity parameter $p = \Omega(\ln n/n)$, when n is large enough. This condition on p is essentially required to show concentration bounds on $\left\|L^{\pm}_{sym} - \overline{L^{\pm}_{sym}}\right\|$ in Lemma 27, which in turn implies a concentration bound on $\left\|T - \overline{T}\right\|$ (see Lemma 28). However, in the sparse regime p is of the order $o(\ln n)/n$, and thus Lemma 27 does not apply in this setting. In fact, it is not difficult to see that the matrices $L^{\pm}_{sym}$ will not concentrate[2] around $\overline{L^{\pm}_{sym}}$ in the sparse regime. On the other hand, by relying on a recent result in (Le et al., 2017, Theorem 4.1) on the concentration of the normalized Laplacian of regularized adjacency matrices of inhomogeneous Erdős-Rényi graphs in the sparse regime (see Theorem 31), we show concentration bounds on $\left\|L^+_{sym,\gamma^+} - \overline{L^+_{sym}}\right\|$ and $\left\|L^-_{sym,\gamma^-} - \overline{L^-_{sym}}\right\|$, which hold when $p \gtrsim 1/n$ and $\gamma^+, \gamma^- \asymp (np)^{6/7}$ (see Lemma 32). As before, these concentration bounds can then be shown to imply a concentration bound on $\left\|T_{\gamma^+,\gamma^-} - \overline{T}\right\|$ (see Lemma 33). Other than these technical differences, the remainder of the arguments follow the same structure as in the proof of Theorem 5, thus leading to the following result in the sparse regime.

Theorem 6 (Restating Theorem 34) *Assuming* $n \geq \max\left\{\frac{2(1-\eta)}{s(1-2\eta)}, \frac{2\eta}{(1-\eta)(1-l)}\right\}$, *suppose* $\tau^+ > 0, \tau^- \geq 0$ *are chosen to satisfy*

$$\tau^+ > \frac{16\eta}{\beta s(1-2\eta)}, \qquad \tau^- < \frac{\beta}{2}\left(\frac{s(1-2\eta)}{s(1-2\eta)+2\eta}\right)\min\left\{\frac{1}{4(1-\beta)}, \frac{\tau^+}{8}\right\}$$

where β, η *satisfy one of the following conditions*

1. $\beta = \frac{4\eta}{s(1-2\eta)+4\eta}$ *and* $0 < \eta < \frac{1}{2}$, *or*
2. $\beta = \frac{1}{2}$ *and* $\eta \leq \frac{s}{2s+4}$.

Then $V_k(\overline{T}) = \Theta R$ *and* $G_k(\overline{T}) = \Theta(C^-)^{-1/2}R$, *where* R *is a rotation matrix, and* $C^- \succ 0$ *is as defined in* (18). *Moreover, there exists a constant* $C > 0$ *such that for* $r \geq 1$ *and* $\delta \in (0,1)$, *if* p *satisfies*

$$p \geq \max\left\{1, \left(\frac{4C_1(\tau^+,\tau^-)(2+\tau^+)}{3(\tau^+)^2(1-\beta)(1+\tau^-)}\right)^{28}\right\}\frac{C_4^{14}(r,s,\eta,l)}{\delta^{28}(1-\eta)n},$$

and $\gamma^+, \gamma^- = [np(1-\eta)]^{6/7}$, *then with probability at least* $1 - 2e^{-r}$, *there exists a rotation* $O \in \mathbb{R}^{k\times k}$ *so that*

$$\left\|V_k(T_{\gamma^+,\gamma^-}) - V_k(\overline{T})O\right\| \leq \delta, \qquad \textit{and} \qquad \left\|G_k(T_{\gamma^+,\gamma^-}) - G_k(\overline{T})O\right\| \leq \frac{\delta}{\sqrt{\tau^+}} + \frac{\delta}{(\tau^+)^2}.$$

[2]See for e.g., Le et al. (2017).

Here, $C_4(r,s,\eta,l) := 2^{5/2}Cr^2 + 3\sqrt{2C_2(s,\eta,l)}$*, with* $C_2(s,\eta,l)$ *as defined in* (45)*.*

The following remarks are in order.

1. It is clear that γ^+, γ^- can neither be too small (since this would imply lack of concentration), nor too large (since this would destroy the latent geometries of G^+, G^-). The choice $\gamma^+, \gamma^- \asymp (np)^{6/7}$ provides a trade-off, and leads to the bounds $\left\|L^+_{sym,\gamma^+} - \overline{L^+_{sym}}\right\|$, $\left\|L^-_{sym,\gamma^-} - \overline{L^-_{sym}}\right\| = O((np)^{-1/14})$ when $p \gtrsim 1/n$ (see Lemma 32).

2. In general, for $\eta \in (0,1/2)$, it suffices that τ^+, τ^- satisfy (8) and $n \gtrsim \max\left\{\frac{1}{s(1-2\eta)}, \frac{\eta}{1-l}\right\}$. As discussed earlier, $\frac{(2+\tau^+)C_1(\tau^+,\tau^-)}{1+\tau^-} \lesssim 1$, and hence it suffices that $p \gtrsim \frac{C_4^{14}(r,s,\eta,l)}{\delta^{28}n}$.

Mis-clustering error bounds. Thus far, our analysis has shown that under suitable conditions on n, p, τ^+ and τ^-, the matrix $G_k(T)$ (or $G_k(T_{\gamma^+,\gamma^-})$ in the sparse regime) is close to $G_k(\overline{T})O$ for some rotation O, with the rows of $G_k(\overline{T})$ preserving the ground truth clustering. This suggests that by applying the k-means clustering algorithm on the rows of $G_k(T)$ (or $G_k(T_{\gamma^+,\gamma^-})$) one should be able to approximately recover the underlying communities. However, the k-means problem for clustering points in $\mathbb{R}^d$ is known to be NP-Hard in general, even for $k = 2$ or $d = 2$ (Aloise et al., 2009; Dasgupta, 2008; Mahajan et al., 2012). On the other hand, there exist efficient $(1+\xi)$-approximation algorithms (for $\xi > 0$), such as, for e.g., the algorithm of Kumar et al. (2004) which has a running time of $O(2^{(k/\xi)^{O(1)}}nd)$.

Using standard tools (Lei and Rinaldo, 2015, Lemma 5.1), we can bound the mis-clustering error when a $(1+\xi)$-approximate k-means algorithm is applied on the rows of $G_k(T)$ (or $G_k(T_{\gamma^+,\gamma^-})$), provided the estimation error bound δ is small enough. In the following theorem, the sets S_i, $i = 1, \ldots, k$ contain those vertices in C_i for which we cannot guarantee correct clustering.

Theorem 7 (Re-Stating Theorem 36) *Under the notation and assumptions of Theorem 5, let* $(\widetilde{\Theta}, \widetilde{X}) \in \mathbb{M}_{n\times k} \times \mathbb{R}^{k\times k}$ *be a* $(1+\xi)$*-approximate solution to the* k*-means problem* $\min_{\Theta\in\mathbb{M}_{n\times k}, X\in\mathbb{R}^{k\times k}} \|\Theta X - G_k(T)\|_F^2$*. Denoting*

$$S_i = \left\{ j \in C_i \ : \ \left\|(\widetilde{\Theta}\widetilde{X})_{j*} - (\Theta(C^-)^{-1/2}RO)_{j*}\right\| \geq \frac{1}{2\sqrt{n_i(\tau^+ + \frac{2}{1-l})}} \right\}$$

it holds with probability at least $1 - 2\varepsilon$ *that*

$$\sum_{i=1}^{k} \frac{|S_i|}{n_i} \leq \delta^2(64 + 32\xi)k\left(\tau^+ + \frac{2}{1-l}\right)\left(\frac{(\tau^+)^3 + 1}{(\tau^+)^4}\right). \tag{9}$$

In particular, if δ *satisfies*

$$\delta < \frac{(\tau^+)^2}{\sqrt{(64+32\xi)k(\tau^+ + \frac{2}{1-l})((\tau^+)^3+1)}}$$

then there exists a $k\times k$ permutation matrix π such that $\widetilde{\Theta}_G = \hat{\Theta}_G\pi$, where $G = \cup_{i=1}^k (C_i \backslash S_i)$.

In the sparse regime, the above statement holds under the notation and assumptions of Theorem 6 with $G_k(T)$ replaced with $G_k(T_{\gamma^+,\gamma^-})$, and with probability at least $1-2e^{-r}$.

We remark that when $\tau^+ \to \infty$, the bound on δ becomes independent of τ^+ and is of the form $\delta \lesssim \frac{1}{\sqrt{k}}$. This is also true for the mis-clustering bound in (9), which is of the form $\sum_{i=1}^k \frac{|S_i|}{n_i} \lesssim \delta^2 k$.

3.2 Symmetric Signed Laplacian

We now describe our results for the symmetric Signed Laplacian. We recall that $\mathbb{E}[A] = \mathbb{E}[A^+] - \mathbb{E}[A^-]$ and $\mathbb{E}[\overline{D}]$ denote the adjacency and degree matrices of the expected graph, under the SSBM ensemble. We define

$$\mathcal{L}_{sym} = I_n - (\mathbb{E}[\overline{D}])^{-1/2}\mathbb{E}[A](\mathbb{E}[\overline{D}])^{-1/2}, \tag{10}$$

to be the normalized Signed Laplacian of the expected graph. Moreover, $\rho = \frac{s}{l} \leq 1$ denotes the *aspect ratio*, measuring the discrepancy between the smallest and largest cluster sizes in the SSBM.

We will first show that for ρ large enough, the smallest $k-1$ eigenvectors of $\mathcal{L}_{sym}$, denoted by $V_{k-1}(\mathcal{L}_{sym})$, are given by $V_{k-1}(\mathcal{L}_{sym}) = \Theta R_{k-1}$, with $R_{k-1} \in \mathbb{R}^{k\times(k-1)}$ a matrix whose columns are the $k-1$ smallest eigenvectors of a $k\times k$ matrix $\overline{C}$ defined in Lemma 37. We will then prove that the rows of $V_{k-1}(\mathcal{L}_{sym})$ impart the same clustering structure as that of Θ. The remaining arguments revolve around deriving concentration bounds on $\left\|\overline{L_{sym}} - \mathcal{L}_{sym}\right\|$, which imply, for n, p and ρ large enough, that the distance between the column spans of $V_{k-1}(\overline{L_{sym}})$ and $V_{k-1}(\mathcal{L}_{sym})$ is small, i.e. there exists a unitary matrix O such that $\left\|V_{k-1}(\overline{L_{sym}}) - V_{k-1}(\mathcal{L}_{sym})O\right\|$ is small. Altogether, this allows us to conclude that the rows of $V_{k-1}(\overline{L_{sym}})$ approximately encode the clustering structure of Θ. The above discussion is summarized in the following theorem, which is our first main result for the symmetric Signed Laplacian, in the moderately dense regime.

Theorem 8 (Eigenspace alignment in the dense case) *Assuming $\eta \in [0, 1/2)$, $k \geq 2$, $n \geq 10$, suppose the aspect ratio satisfies*

$$\sqrt{\rho} > 1 - \frac{1}{4k(2+\sqrt{k})}, \tag{11}$$

and suppose that, for $\delta \in (0, \frac{1}{2})$, it holds true that

$$p > C(k,\eta,\delta)\frac{\ln n}{n} \qquad \text{with} \quad C(k,\eta,\delta) = \left(\frac{2Ck}{\delta(1-2\eta)}\right)^2 \quad \text{and}\, C < 43, \tag{12}$$

Then there exists a universal constant $c > 0$, such that with probability at least $1 - \frac{2}{n} - n\exp\left(\frac{-np}{c}\right)$, there exists an orthogonal matrix $O \in \mathbb{R}^{(k-1)\times(k-1)}$ such that

$$\|V_{k-1}(\overline{L_{sym}}) - \Theta R_{k-1}O\| \leq 2\delta,$$

where $R_{k-1} \in \mathbb{R}^{k\times(k-1)}$ is a matrix whose columns are the $(k-1)$ smallest eigenvectors of the matrix $\overline{C}$ defined in Lemma 37.

Remark 9 *(Related work) As previously explained, for the special case where $k = 2$ and with equal-size clusters, a similar result was proved in (Cucuringu et al., 2019, Theorem 3). Under a different SSBM model, the Signed Laplacian clustering algorithm was analyzed by Mercado et al. (2019) for general k. Although their generative model is more general than our SSBM, their results on the symmetric Signed Laplacian do not apply here. More precisely, one assumption of Theorem 3 of Mercado et al. (2019) translates into our model as $p(k-2)(1-2\eta) < 0$, which does not hold for $\eta < \frac{1}{2}$ and $k \geq 2$.*

Remark 10 *(Assumptions) The condition on the aspect ratio* (11) *is essential to apply a perturbation technique, where the reference is the setting with equal-size clusters, i.e. $n_i = \frac{n}{k}, \forall i \in [k]$ (see Lemma 39). In the sparsity condition* (12)*, we note that the constant $C(k, \eta, \delta)$ scales quadratically with the number of classes k and as δ^{-2} with $\delta > 0$ the error on the eigenspace. However, we conjecture that this assumption is only an artefact of the proof technique, and that the result could hold for more general graphs with very unbalanced cluster sizes.*

Regularized Signed Laplacian. We now consider the sparse regime $p = o(\ln n)/n$ and show that we can recover the ground-truth clustering structure up to some small error using the regularized Signed Laplacian L_γ, provided that n, p and ρ are large enough, and that the regularization parameters γ^+, γ^- are well-chosen. We denote $\mathcal{L}_\gamma$ to be the equivalent of the regularized Laplacian for the expected graph in our SSBM, *i.e.*

$$\mathcal{L}_\gamma = I - (\mathbb{E}[\overline{D}_\gamma])^{-1/2}\mathbb{E}[A_\gamma](\mathbb{E}[\overline{D}_\gamma])^{-1/2},$$

with $\mathbb{E}[A_\gamma]$, resp. $\mathbb{E}[\overline{D}_\gamma]$, denoting the adjacency matrix, resp. the degree matrix, of the expected regularized graph. The next theorem is an intermediate result, which provides a high probability bound on $\|L_\gamma - \mathcal{L}_\gamma\|$ and $\|L_\gamma - \mathcal{L}_{sym}\|$.

Theorem 11 *(Error bound for the regularized Signed Laplacian) Assuming $\eta \in [0, 1/2)$, $k \geq 2$, and regularization parameters $\gamma^+, \gamma^- \geq 0$, $\gamma := \gamma^+ + \gamma^-$, it holds true that for any $r \geq 1$, with probability at least $1 - 7e^{-2r}$, we have*

$$\|L_\gamma - \mathcal{L}_\gamma\| \leq \frac{Cr^2}{\sqrt{\gamma}}\left(1 + \frac{\overline{d}}{\gamma}\right)^{5/2} + \frac{32\sqrt{2r}}{\sqrt{\gamma}} + \frac{8}{\sqrt{\overline{d}}}, \tag{13}$$

with $C > 1$ an absolute constant. Moreover, it also holds true that

$$\|L_\gamma - \mathcal{L}_{sym}\| \leq \frac{Cr^2}{\sqrt{\gamma}}\left(1 + \frac{\overline{d}}{\gamma}\right)^{5/2} + \frac{32\sqrt{2r}}{\sqrt{\gamma}} + \frac{8}{\sqrt{\overline{d}}} + \frac{\gamma}{\overline{d} + \gamma}. \tag{14}$$

In particular, for the choice $\gamma = \overline{d}^{7/8}$, if $p \geq 2/n$, we obtain

$$\|L_\gamma - \mathcal{L}_{sym}\| \leq \left(128Cr^2 + 1\right)(\overline{d})^{-\frac{1}{8}}.$$

Remark 12 *The above theorem shows the concentration of our regularized Laplacian L_γ towards the regularized Laplacian* (13) *and the Signed Laplacian* (14) *of the expected graph. More precisely, if for some well-chosen parameters $\gamma^+, \gamma^- \geq 0$, these upper bounds are small, e.g $\|L_\gamma - \mathcal{L}_{sym}\| << 1$, then we have $\|L_\gamma - \mathcal{L}_{sym}\| << \|\mathcal{L}_{sym}\|$ since $\|\mathcal{L}_{sym}\| = 2$ (see Appendix E).*

Using this concentration bound, we can show that the eigenspaces $V_{k-1}(L_\gamma)$ and $V_{k-1}(\mathcal{L}_{sym})$ are "close", provided that $p = \Omega(1/n)$, ρ is close enough to 1, and γ is well-chosen. This is stated in the next theorem.

Theorem 13 (Eigenspace alignment in the sparse case) *Assuming $\eta \in [0, 1/2)$, $k \geq 2$, and $n \geq 10$, suppose that (11) holds true, and for $\delta \in (0, \frac{1}{2})$ and $r \geq 1$, the sparsity p satisfies*

$$p > \left(\frac{2kC_4}{\delta(1-2\eta)}\right)^8 \frac{2}{n} \qquad \text{with} \quad C_4 = 128Cr^2 + 1 \tag{15}$$

and $C > 1$ the constant defined in (13). If the regularization parameters $\gamma^+, \gamma^- \geq 0$ are chosen so that $\gamma = \overline{d}^{7/8}$, then with probability at least $1 - 7e^{-2r} - \frac{2}{n} - ne^{-np/c}$, there exists an orthogonal matrix $O \in \mathbb{R}^{(k-1)\times(k-1)}$ so that

$$\|V_{k-1}(L_\gamma) - \Theta R_{k-1} O\| \leq 2\delta.$$

Remark 14 *In the sparse setting, the constant before the factor $\frac{1}{n}$ in the sparsity condition (15) scales as $\left(\frac{k}{\delta}\right)^8$. However for k fixed, it would hold if $p = \omega(1/n)$ as $n \to \infty$.*

Remark 15 *In practice, one can choose the regularization parameters by first estimating the sparsity parameter p, e.g. from the fraction of connected pairs of nodes*

$$p = \frac{2}{n(n-1)} \sum_{i<j} |A_{ij}|,$$

then choosing $\gamma \geq 0$ so that $\gamma = (\hat{p}(n-1))^{7/8}$. However, from this analysis, it is not clear how one would suitably choose γ^+ and γ^-.

Mis-clustering error bounds. Since $V_{k-1}(\overline{L_{sym}})$ and $V_{k-1}(L_\gamma)$ are "close" to $V_{k-1}(\mathcal{L}_{sym})$, we recover the ground-truth clustering structure up to some error, which we quantify in the following theorem, where we bound the mis-clustering rate when using a $(1+\xi)$-approximate k-means error on the rows of $V_{k-1}(\overline{L_{sym}})$ (resp. $V_{k-1}(L_\gamma)$).

Theorem 16 *(Number of mis-clustered nodes) Let $\xi > 0$ and $\delta \in \left(0, \sqrt{\frac{1}{12(16+8\xi)(k-1)}}\right)$, and suppose that ρ and p satisfy the assumptions of Theorem 8 (resp. Theorem 13 and $r \geq 1$). Let $(\widetilde{\Theta}, \widetilde{R}_{k-1})$ be the $(1+\xi)$-approximation of the k-means problem*

$$\min_{\Theta \in \mathbb{M}_{n,k}, R \in \mathbb{R}^{k\times(k-1)}} \left\|\Theta R - V_{k-1}(\overline{L_{sym}})\right\|_F \qquad \textit{(resp.} \quad \min_{\Theta \in \mathbb{M}_{n,k}, R \in \mathbb{R}^{k\times(k-1)}} \|\Theta R - V_{k-1}(L_\gamma)\|_F \quad \textit{)}.$$

Let $S_i = \left\{ j \in C_i; \left\|(\widetilde{\Theta}\widetilde{R}_{k-1})_{j*} - (\Theta R_{k-1} O)_{j*}\right\|^2 \geq \frac{2}{3n_i} \right\}$ and $\widetilde{V} = \cup_{i=1}^k C_i \backslash S_i$. Then with probability at least $1 - \frac{2}{n} - n\exp(\frac{-np}{c})$ (resp. $1 - 7e^{-2r} - \frac{2}{n} - ne^{-np/c}$), there exists a permutation $\pi \in \mathbb{R}^{k\times k}$ such that $\widetilde{\Theta}_{\widetilde{V}*} = \hat{\Theta}_{\widetilde{V}*}\pi$ and

$$\sum_{i=1}^k \frac{|S_i|}{n_i} \leq 96(2+\xi)(k-1)\delta^2.$$

In particular, the set of mis-clustered nodes is a subset of $\cup_{i=1}^k S_i$.

4. Analysis of SPONGE Symmetric

This section contains the proof of our main results for SPONGE$_{sym}$, divided over the following subsections. Section 4.1 describes the eigen-decomposition of the matrix $\overline{T}$, thus revealing that a subset of its eigenvectors contain relevant information about Θ. Section 4.2 provides conditions on τ^+, τ^- which ensure that $V_k(\Theta) = \Theta R$ (for some rotation matrix R), along with a lower bound on the eigengap $\lambda_{n-k+1}(\overline{T}) - \lambda_{n-k}(\overline{T})$. Section 4.3 then derives concentration bounds on $\left\|T - \overline{T}\right\|$ using standard tools from the random matrix literature. These results are combined in Section 4.4 to derive error bounds for estimating $V_k(\overline{T})$ and $G_k(\overline{T})$ up to a rotation (using the Davis-Kahan theorem). The results summarized thus far pertain to the "dense" regime, where we require $p = \Omega(\ln n/n)$ when n is large. Section 4.5 extends these results to the sparse regime where $p = o(\ln n)/n$, for the regularized version of SPONGE$_{sym}$. Finally, we conclude in Section 4.6 by translating our results from Sections 4.4 and 4.5 to obtain mis-clustering error bounds for a $(1+\xi)$-approximate k-means algorithm, by leveraging previous tools from the literature (Lei and Rinaldo, 2015).

4.1 Eigen-decomposition of $\overline{T}$

The following lemma shows that a subset of the eigenvectors of $\overline{T}$ indeed contain information about Θ, i.e., the ground-truth clustering.

Lemma 17 (Spectrum of $\overline{T}$) *Let*

$$d_i^+ = p\left(n(s_i(1-2\eta)+\eta) - (1-\eta)\right),$$

$$d_i^- = p\left(n(-s_i(1-2\eta)+(1-\eta)) - \eta\right),$$

denote the expected degree of a node in cluster C_i, $i \in [k]$. Let $u^+ = \left(\sqrt{\frac{n_1}{d_1^+}}, \ldots, \sqrt{\frac{n_k}{d_k^+}}\right)^\top$, $u^- = \left(\sqrt{\frac{n_1}{d_1^-}}, \ldots, \sqrt{\frac{n_k}{d_k^-}}\right)^\top$, $\alpha_i^+ = 1 + \tau^- + p(1-\eta)/d_i^+$, and $\alpha_i^- = 1 + \tau^+ + p\eta/d_i^-$, for $i \in [k]$, for some $\tau^+ > 0, \tau^- \geq 0$. Let the columns of $V^\perp$ contain eigenvectors of $\mathbb{E}[D^+]$ which are orthogonal to the column span of Θ. It holds true that

$$\overline{T} = \begin{bmatrix}\Theta R & V^\perp\end{bmatrix} \begin{bmatrix} \mathbf{\Lambda} & & & \\ & \frac{\alpha_1^+}{\alpha_1^-} I_{n_1-1} & & \\ & & \ddots & \\ & & & \frac{\alpha_k^+}{\alpha_k^-} I_{n_k-1} \end{bmatrix} \begin{bmatrix} (\Theta R)^\top \\ {V^\perp}^\top \end{bmatrix}, \tag{16}$$

where R is a $k \times k$ rotation matrix, and Λ is a diagonal matrix, such that $(C^-)^{-1/2}\, C^+\, (C^-)^{-1/2} = R \Lambda R^T$, where

$$C^+ = -p\eta u^+ (u^+)^\top + diag\left(1 + \tau^- + \frac{p}{d_i^+}(1 - \eta - n_i(1-2\eta))\right), \tag{17}$$

$$C^- = -p(1-\eta) u^- (u^-)^\top + diag\left(1 + \tau^+ + \frac{p}{d_i^-}(\eta + n_i(1-2\eta))\right). \tag{18}$$

Proof We first consider the spectrum of D^+, D^-, A^+, A^-, followed by that of $(\overline{L^+_{sym}} + \tau^- I)$ and $(\overline{L^-_{sym}} + \tau^+ I)$, which altogether will reveal the spectral decomposition of $\overline{T}$.

• **Analysis in expectation of the spectra of D^+, D^-, A^+, A^-.** Without loss of generality, we may assume that cluster C_1 contains the first n_1 vertices, cluster C_2 the next n_2 vertices and similarly for the remaining clusters. Note that $\mathbb{E}[D^\pm] = \operatorname{diag}\left(d_1^\pm I_{n_1}, \ldots, d_k^\pm I_{n_k}\right)$, where for $i \in [k]$, straightforward calculations reveal that $d_i^+ = p\left(n(s_i(1-2\eta)+\eta) - (1-\eta)\right)$, and $d_i^- = p\left(n(-s_i(1-2\eta)+(1-\eta)) - \eta\right)$. One can rewrite the matrices $(\mathbb{E}[D^\pm])^{-1}$ in the more convenient form

$$(\mathbb{E}[D^\pm])^{-1} = [\Theta\ V^\perp]\ \operatorname{diag}\left(\frac{1}{d_1^\pm}, ..., \frac{1}{d_k^\pm}, \frac{1}{d_1^\pm} I_{n_1-1}, ..., \frac{1}{d_k^\pm} I_{n_k-1}\right)\ [\Theta\ V^\perp]^\top \tag{19}$$

since the column vectors of Θ are eigenvectors of $(\mathbb{E}[D^\pm])^{-1}$, and the eigenvalues of $(\mathbb{E}[D^\pm])^{-1}$ are apparent because $\mathbb{E}[D^\pm]$ is a diagonal matrix. Note that (19) is true in general, and does not make any assumption on the placement of the vertices into their respective C_i cluster. Furthermore, one can verify that $\mathbb{E}[A^+]$ admits the eigen-decomposition

$$\mathbb{E}[A^+] = \Theta_{n\times k} \begin{bmatrix} n_1 p(1-\eta) & \sqrt{n_1 n_2}p\eta & \ldots & \sqrt{n_1 n_k}p\eta \\ \sqrt{n_2 n_1}p\eta & n_2 p(1-\eta) & \ldots & \sqrt{n_2 n_k}p\eta \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{n_k n_1}p\eta & \sqrt{n_k n_2}p\eta & \ldots & n_k p(1-\eta) \end{bmatrix}_{k\times k} \Theta^\top_{k\times n} - p(1-\eta) I_{n\times n} \tag{20}$$

and similarly, $\mathbb{E}[A^-]$ can be decomposed as

$$\mathbb{E}[A^-] = \Theta_{n\times k} \begin{bmatrix} n_1 p\eta & \sqrt{n_1 n_2}p(1-\eta) & \ldots & \sqrt{n_1 n_k}p(1-\eta) \\ \sqrt{n_2 n_1}p(1-\eta) & n_2 p\eta & \ldots & \sqrt{n_2 n_k}p(1-\eta) \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{n_k n_1}p(1-\eta) & \sqrt{n_k n_2}p(1-\eta) & \ldots & n_k p\eta \end{bmatrix}_{k\times k} \Theta^\top_{k\times n} - p\eta I_{n\times n}\,.$$

• **Analysis of the spectra of $(\overline{L^+_{sym}} + \tau^- I)$ and $(\overline{L^-_{sym}} + \tau^+ I)$.** We start by observing that

$$\begin{aligned} \overline{L^\pm_{sym}} + \tau^\mp I &= I - (\mathbb{E}[D^\pm])^{-1/2}(\mathbb{E}[A^\pm])(\mathbb{E}[D^\pm])^{-1/2} + \tau^\mp I \\ &= (1+\tau^\mp) I - (\mathbb{E}[D^\pm])^{-1/2}(\mathbb{E}[A^\pm])(\mathbb{E}[D^\pm])^{-1/2}\,. \end{aligned} \tag{21}$$

In light of (20), one can write $(\mathbb{E}[D^+])^{-1/2}(\mathbb{E}[A^+])(\mathbb{E}[D^+])^{-1/2}$ as

$$(\mathbb{E}[D^+])^{-1/2}(\mathbb{E}[A^+])(\mathbb{E}[D^+])^{-1/2} = -p(1-\eta)(\mathbb{E}[D^+])^{-1}$$

$$+ \begin{bmatrix}\Theta & V^\perp\end{bmatrix} \begin{bmatrix} \overbrace{\begin{bmatrix} \frac{n_1}{d_1^+}p(1-\eta) & \sqrt{\frac{n_1 n_2}{d_1^+ d_2^+}}p\eta & \ldots & \sqrt{\frac{n_1 n_k}{d_1^+ d_k^+}}p\eta \\ \sqrt{\frac{n_2 n_1}{d_2^+ d_1^+}}p\eta & \frac{n_2}{d_2^+}p(1-\eta) & \ldots & \sqrt{\frac{n_2 n_k}{d_2^+ d_k^+}}p\eta \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{\frac{n_k n_1}{d_k^+ d_1^+}}p\eta & \sqrt{\frac{n_k n_2}{d_k^+ d_2^+}}p\eta & \ldots & \frac{n_k}{d_k^+}p(1-\eta) \end{bmatrix}_{k\times k}}^{\stackrel{\text{def}}{=} B^+} & \mathbf{0}_{k\times(n-k)} \\ \mathbf{0}_{(n-k)\times k} & \mathbf{0}_{(n-k)\times(n-k)} \end{bmatrix} \begin{bmatrix}\Theta^\top \\ V^{\perp\top}\end{bmatrix}. \tag{22}$$

Similarly, using the expression for $\mathbb{E}[A^-]$, the expression for $(\mathbb{E}[D^-])^{-1/2}(\mathbb{E}[A^-])(\mathbb{E}[D^-])^{-1/2}$ can be written as

$$(\mathbb{E}[D^-])^{-1/2}(\mathbb{E}[A^-])(\mathbb{E}[D^-])^{-1/2} = -p\eta(\mathbb{E}[D^-])^{-1} +$$

$$\begin{bmatrix}\Theta & V^\perp\end{bmatrix}\begin{bmatrix}\overbrace{\begin{bmatrix}\frac{n_1}{d_1^-}p\eta & \sqrt{\frac{n_1n_2}{d_1^-d_2^-}}p(1-\eta) & \dots & \sqrt{\frac{n_1n_k}{d_1^-d_k^-}}p(1-\eta)\\ \sqrt{\frac{n_2n_1}{d_2^-d_1^-}}p(1-\eta) & \frac{n_2}{d_2^-}p\eta & \dots & \sqrt{\frac{n_2n_k}{d_2^-d_k^-}}p(1-\eta)\\ \vdots & \vdots & \ddots & \vdots\\ \sqrt{\frac{n_kn_1}{d_k^-d_1^-}}p(1-\eta) & \sqrt{\frac{n_kn_2}{d_k^-d_2^-}}p(1-\eta) & \dots & \frac{n_k}{d_k^-}p\eta\end{bmatrix}_{k\times k}}^{\stackrel{\text{def}}{=}B^-} & \mathbf{0}_{k\times(n-k)}\\ \mathbf{0}_{(n-k)\times k} & \mathbf{0}_{(n-k)\times(n-k)}\end{bmatrix}\begin{bmatrix}\Theta^\top\\ V^{\perp\top}\end{bmatrix}. \tag{23}$$

Combining (19), (22), and (23) into (21), we readily arrive at

$$(\overline{L^\pm_{sym}}+\tau^\mp I) = \begin{bmatrix}\Theta & V^\perp\end{bmatrix}\begin{bmatrix}\underbrace{[\mathrm{diag}(\alpha_i^\pm)-B^\pm]_{k\times k}}_{\stackrel{\text{def}}{=}C^\pm} & \mathbf{0}_{k\times(n-k)}\\ & \alpha_1^\pm I_{n_1-1} & & & \\ & & \alpha_2^\pm I_{n_k-1} & & \\ & & & \ddots & \\ & & & & \alpha_k^\pm I_{n_k-1},\end{bmatrix}\begin{bmatrix}\Theta^\top\\ V^{\perp\top}\end{bmatrix} \tag{24}$$

where $\alpha_i^\pm$ and C^+, C^- are defined as in the statement of the lemma. The spectral decomposition of $\overline{T}$ now follows trivially using (24), along with the spectral decomposition $(C^-)^{-1/2}C^+(C^-)^{-1/2} = R\Lambda R^T$. ■

Lemma 17 reveals that we need to extract the k-informative eigenvectors ΘR from the n-eigenvectors $\begin{bmatrix}\Theta R & V^\perp\end{bmatrix}$ of $\overline{T}$. Clearly, it suffices to recover any orthonormal basis for the column span of Θ, since the rows of any such corresponding matrix (one instance of which is ΘR) will exhibit the same clustering structure as Θ.

4.2 Ensuring $V_k(\overline{T}) = \Theta R$ and bounding the spectral gap

In this section, our aim is to show that, for suitable values of $\tau^+ > 0, \tau^- \geq 0$, the eigenvectors corresponding to the smallest k eigenvalues of $\overline{T}$ are given by ΘR, i.e., $V_k(\overline{T}) = \Theta R$. This is equivalent to ensuring (recall Lemma 17) that

$$\lambda_{n-k+1}(\overline{T}) = \left\|(C^-)^{-1/2}C^+(C^-)^{-1/2}\right\| < \min_{i\in[k]}\frac{\alpha_i^+}{\alpha_i^-} = \lambda_{n-k}(\overline{T}). \tag{25}$$

Moreover, we will need to find a strictly positive lower-bound on the spectral gap $\lambda_{n-k}(\overline{T}) - \lambda_{n-k+1}(\overline{T})$, as it will be used later on, in order to show that the column span of $V_k(T)$ is close to that of $V_k(\overline{T})$. We first consider the equal-sized clusters case, and then proceed to the general-sized clusters case.

4.2.1 Spectral gap for equal-sized clusters

When the cluster sizes are equal, the analysis is considerably cleaner than the general setting. Let us first establish notation specific to the equal-sized clusters case.

Remark 18 (Notation for the equal-sized clusters) *For clusters of equal size, we have that* $n_1 = ... = n_k = n/k$, $d^+ := d_1^+ = ... = d_k^+$, $d^- := d_1^- = ... = d_k^-$, $\alpha^+ := \alpha_1^+ = ... = \alpha_k^+$, *and* $\alpha^- := \alpha_1^- = ... = \alpha_k^-$. *Let* C_e^+, C_e^-, *and* $\overline{T_e}$ *denote the respective counterparts of* C^+, C^-, *and* $\overline{T}$, *for the equal-sized case. In light of* (17) *and* (18), *one can verify that* C_e^+ *and* C_e^- *are simultaneously diagonalizable, which we show in Lemma 60.*

In the following lemma, we show the exact value of $\|\Lambda\| = \left\|(C_e^-)^{-1/2}C_e^+(C_e^-)^{-1/2}\right\|$.

Lemma 19 (Bounding the spectral norm of $(C_e^-)^{-1/2}C_e^+(C_e^-)^{-1/2}$) *For equal-sized clusters, the following holds true*

$$\left\|(C_e^-)^{-1/2}C_e^+(C_e^-)^{-1/2}\right\| = \max\left\{\frac{\tau^-}{\tau^+}, \frac{\tau^- + \frac{pn\eta}{d^+}}{\tau^+ + \frac{pn(1-\eta)}{d^-}}\right\}.$$

Proof The lemma follows directly from Lemma 60. ■

Next, we derive conditions on $\tau^+ > 0, \tau^- \geq 0$ which ensure $V_k(\overline{T}) = \Theta R$.

Lemma 20 (Conditions on τ^- and τ^+) *Suppose* $n \geq \frac{2k(1-\eta)}{1-2\eta}$, *and* $\tau^- \geq 0$, $\tau^+ > 0$. *If* τ^-, τ^+ *satisfy*

1. $$\tau^-\left(1 + \frac{p\eta}{d^-}\right) < \tau^+\left(1 + \frac{p(1-\eta)}{d^+}\right),$$

2. $$\tau^-\left[\frac{(1-2\eta)/k}{(1-\eta) - \frac{1-2\eta}{k}}\right] + \tau^+\left[\frac{(1-2\eta)/k}{\eta + \frac{1-2\eta}{k}}\right] + 1 > \frac{2\eta}{\eta + \frac{1-2\eta}{2k}}.$$

Then it holds true that $V_k(\overline{T}) = \Theta R$, *i.e.,* $\lambda_{n-k+1}(\overline{T}) = \left\|(C_e^-)^{-1/2}C_e^+(C_e^-)^{-1/2}\right\| < \frac{\alpha^+}{\alpha^-} = \lambda_{n-k}(\overline{T})$.

Proof Recalling the expression for $\left\|(C_e^-)^{-1/2}C_e^+(C_e^-)^{-1/2}\right\|$ from Lemma 19, we will ensure that each term inside the max is less than α^+/α^-. To derive the first condition of the lemma, we simply ensure that

$$\frac{\tau^-}{\tau^+} < \frac{1 + \tau^- + p(1-\eta)/d^+}{1 + \tau^+ + p\eta/d^-} \Leftrightarrow \tau^-\left(1 + \frac{p\eta}{d^-}\right) < \tau^+\left(1 + \frac{p(1-\eta)}{d^+}\right).$$

Before deriving the second condition, let us note additional useful bounds on $\frac{np}{d^-}, \frac{np}{d^+}$ which will be needed later.

1. $d^-/np = 1 - \eta - (1-2\eta)/k - \eta/n \leq 1 - \eta$.

2. Since $n \geq k \geq 2$, we obtain that $d^-/np \geq (1-\eta) - (1-3\eta)/k \geq \frac{1-\eta}{2}$. This also implies that $p\eta/d^- \leq 1$.

   Therefore, combining the above two bounds, we arrive at

$$\frac{1}{1-\eta} \leq \frac{np}{d^-} \leq \frac{2}{1-\eta}\,.$$

3. $d^+/np = (1-2\eta)/k + \eta - (1-\eta)/n \leq \eta + (1-2\eta)/k$.

4. Since $n \geq \frac{2k(1-\eta)}{1-2\eta}$, it holds that $d^+/np = (1-2\eta)/k + \eta - (1-\eta)/n \geq \eta + (1-2\eta)/2k$.

5. Therefore, combining the above two conditions yields

$$\frac{1}{\eta + \frac{1-2\eta}{k}} \leq \frac{np}{d^+} \leq \frac{1}{\eta + \frac{1-2\eta}{2k}}\,.$$

To derive the second condition, we need to ensure $\frac{\tau^- + \frac{pn\eta}{d^+}}{\tau^+ + \frac{pn(1-\eta)}{d^-}} < \frac{1+\tau^- + p(1-\eta)/d^+}{1+\tau^+ + p\eta/d^-}$, which is equivalent to

$$\tau^- \left[1 - \frac{np}{d^-}\left((1-\eta) - \frac{\eta}{n}\right)\right] < \tau^+ \left[1 - \frac{np}{d^+}\left(\eta - \frac{1-\eta}{n}\right)\right] + \underbrace{\left[\frac{np(1-\eta)}{d^-}\left(1 + \frac{p(1-\eta)}{d^+}\right) - \frac{np\eta}{d^+}\left(1 + \frac{p\eta}{d^-}\right)\right]}_{\text{term 2}}\,.$$

Now, we can lower bound "term 2" in the above equation as

$$\frac{np(1-\eta)}{d^-}\left(1 + \frac{p(1-\eta)}{d^+}\right) - \frac{np\eta}{d^+}\left(1 + \frac{p\eta}{d^-}\right) \geq 1 - \frac{2\eta}{\eta + \frac{(1-2\eta)}{k}}\,.$$

Hence from the above two equations, we observe that it suffices that τ^+, τ^- satisfy

$$\tau^- \left[\frac{(1-2\eta)/k}{(1-\eta) - \frac{1-2\eta}{k}}\right] + \tau^+ \left[\frac{(1-2\eta)/k}{\eta + \frac{1-2\eta}{k}}\right] + 1 > \frac{2\eta}{\eta + \frac{1-2\eta}{2k}}\,.$$

■

Next, we derive sufficient conditions on τ^+, τ^- which ensure a lower bound on the *spectral gap*

$$\lambda_{n-k}(\overline{T}) - \lambda_{n-k+1}(\overline{T}) = \frac{\alpha^+}{\alpha^-} - \left\|(C_e^-)^{-1/2} C_e^+ (C_e^-)^{-1/2}\right\|.$$

Lemma 21 (Conditions on τ^+, τ^-, and lower-bound on spectral gap) *Suppose $n \geq \frac{2k(1-\eta)}{1-2\eta}$, then the following holds.*

1. *If* $\tau^+ > 0, \tau^- \geq 0$ *satisfy*

$$\tau^+ > \frac{32\eta k}{3(1-2\eta)}, \quad \tau^- < \min\left\{\frac{3}{2}, \frac{3}{16}\tau^+, \frac{3(1-\eta)}{8(\eta + \frac{1-2\eta}{k})}\right\},$$

then $V_k(\overline{T}) = \Theta R$, *and* $\left\|(C_e^-)^{-1/2}C_e^+(C_e^-)^{-1/2}\right\| < \left(1 - \frac{(1-2\eta)}{2k(1-\eta)}\right)\frac{\alpha^+}{\alpha^-}$, *i.e.,* $\lambda_{n-k}(\overline{T}) - \lambda_{n-k+1}(\overline{T}) > \left(\frac{(1-2\eta)}{2k(1-\eta)}\right)\frac{\alpha^+}{\alpha^-}$.

2. *If* $\eta < \frac{1}{3k+2}$ *and* $\tau^+ > 0, \tau^- \geq 0$ *satisfy*

$$\tau^- < \min\left\{\left(\frac{\frac{1-2\eta}{k} - \eta}{\frac{1-2\eta}{k} + \eta}\right), \frac{1}{2}, \frac{\tau^+}{8}\right\},$$

then $V_k(\overline{T}) = \Theta R$, *and* $\left\|(C_e^-)^{-1/2}C_e^+(C_e^-)^{-1/2}\right\| < \frac{\alpha^+}{2\alpha^-}$, *i.e.,* $\lambda_{n-k}(\overline{T}) - \lambda_{n-k+1}(\overline{T}) > \frac{\alpha^+}{2\alpha^-}$.

Proof We need to ensure the following two conditions for a suitably chosen $\beta \in (0, 1]$.

$$\frac{\tau^- + \frac{pn\eta}{d^+}}{\tau^+ + \frac{pn(1-\eta)}{d^-}} < \beta\left(\frac{1 + \tau^- + p(1-\eta)/d^+}{1 + \tau^+ + p\eta/d^-}\right), \tag{26}$$

$$\frac{\tau^-}{\tau^+} < \beta\left(\frac{1 + \tau^- + p(1-\eta)/d^+}{1 + \tau^+ + p\eta/d^-}\right). \tag{27}$$

1. Ensuring (26) We can rewrite (26) as

$$\begin{aligned}&\tau^-\left(1 + \frac{p\eta}{d^-} - \beta\frac{pn(1-\eta)}{d^-}\right) + \tau^+\left(\frac{pn\eta}{d^+} - \beta\left(1 + \frac{p(1-\eta)}{d^+}\right)\right) + \tau^+\tau^-(1-\beta)\\ &< \beta\frac{pn(1-\eta)}{d^-}\left(1 + \frac{p(1-\eta)}{d^+}\right) - \frac{pn\eta}{d^+}\left(1 + \frac{p\eta}{d^-}\right).\end{aligned} \tag{28}$$

Using the expressions for d^+, d^-, we can write the coefficients of the terms τ^+, τ^- as follows.

$$\begin{aligned}1 + \frac{p\eta}{d^-} - \beta\frac{pn(1-\eta)}{d^-} &= \frac{-(\frac{1-2\eta}{k}) + (1-\eta)(1-\beta)}{-(\frac{1-2\eta}{k}) + (1-\eta) - \frac{\eta}{n}},\\ \frac{pn\eta}{d^+} - \beta\left(1 + \frac{p(1-\eta)}{d^+}\right) &= \frac{np}{d^+}(\eta - \beta\frac{1-\eta}{n}) - \beta = \frac{\eta(1-\beta) - \beta(\frac{1-2\eta}{k})}{\frac{1-2\eta}{k} + \eta - \frac{1-\eta}{n}}.\end{aligned}$$

Moreover, using the bounds on $\frac{d^-}{np}, \frac{d^+}{np}$ derived in Lemma 20, we can lower bound the RHS term in (28) as

$$\beta\frac{pn(1-\eta)}{d^-}\left(1 + \frac{p(1-\eta)}{d^+}\right) - \frac{pn\eta}{d^+}\left(1 + \frac{p\eta}{d^-}\right) > \beta - \frac{2\eta}{\eta + \frac{1-2\eta}{k}}.$$

From the above considerations, we see that (28) is ensured provided

$$\tau^-\left[\frac{(\frac{1-2\eta}{k}) - (1-\eta)(1-\beta)}{-(\frac{1-2\eta}{k}) + (1-\eta) - \frac{\eta}{n}}\right] + \tau^+\left[\frac{-\eta(1-\beta) + \beta(\frac{1-2\eta}{k})}{\frac{1-2\eta}{k} + \eta - \frac{1-\eta}{n}}\right] + \beta > \frac{2\eta}{\eta + \frac{1-2\eta}{k}} + \tau^+\tau^-(1-\beta). \tag{29}$$

We outline two possible ways in which (29) is ensured.

- Note that the denominators of the coefficients of τ^+, τ^- in (29) are positive, while the numerators are non-negative provided $1-\beta \leq \frac{(1-2\eta)}{2k(1-\eta)}$. Therefore, choosing

$$\beta = 1 - \frac{(1-2\eta)}{2k(1-\eta)} \quad \left(\geq \frac{3}{4}\right),$$

note that (29) is ensured provided

$$\tau^- \left[\frac{(1-2\eta)}{2k(1-\eta)}\right] + \tau^+ \left[\frac{3(1-2\eta)}{8k\left(\eta + \frac{1-2\eta}{k}\right)}\right] + \frac{3}{4} > \frac{2\eta}{\eta + \frac{1-2\eta}{k}} + \tau^+\tau^- \left[\frac{(1-2\eta)}{2k(1-\eta)}\right]. \quad (30)$$

Finally, we observe that in order for (30) to hold, it suffices that

$$\tau^+\tau^- \left[\frac{(1-2\eta)}{2k(1-\eta)}\right] < \frac{\tau^+}{2} \left[\frac{3(1-2\eta)}{8k\left(\eta + \frac{1-2\eta}{k}\right)}\right] \iff \tau^- < \frac{3(1-\eta)}{8\left(\eta + \frac{1-2\eta}{k}\right)}, \text{ and}$$

$$\frac{2\eta}{\eta + \frac{1-2\eta}{k}} < \frac{\tau^+}{2} \left[\frac{3(1-2\eta)}{8k\left(\eta + \frac{1-2\eta}{k}\right)}\right] \iff \tau^+ > \frac{32\eta k}{3(1-2\eta)}.$$

- Alternatively, by setting $\beta = 1/2$, (29) can be rewritten as

$$\tau^+ \left[\frac{-\frac{\eta}{2} + \frac{1-2\eta}{2k}}{\frac{1-2\eta}{k} + \eta - \frac{1-\eta}{n}}\right] + \frac{1}{2} > \frac{2\eta}{\eta + \frac{1-2\eta}{k}} + \tau^- \left[\frac{-(\frac{1-2\eta}{k}) + \frac{1-\eta}{2}}{-(\frac{1-2\eta}{k}) + (1-\eta) - \frac{\eta}{n}}\right] + \frac{\tau^+\tau^-}{2}. \quad (31)$$

Clearly, it holds true that

$$\frac{1}{2} > \frac{2\eta}{\eta + \frac{1-2\eta}{k}} \iff \eta < \frac{1}{3k+2},$$

which also ensures that the numerator of the coefficient of τ^+ is positive. Therefore, if $\eta < \frac{1}{3k+2}$, then in order for (31) to hold, it suffices that

$$\tau^- < \left[\frac{-\eta + \frac{1-2\eta}{k}}{\frac{1-2\eta}{k} + \eta}\right] \implies \tau^+ \left[\frac{-\frac{\eta}{2} + \frac{1-2\eta}{2k}}{\frac{1-2\eta}{k} + \eta - \frac{1-\eta}{n}}\right] > \frac{\tau^+\tau^-}{2}.$$

2. Ensuring (27) Note that one can rewrite (27) as

$$\tau^-\tau^+(1-\beta) + \tau^- \left(1 + \frac{p\eta}{d^-}\right) < \beta\tau^+ \left(1 + \frac{p(1-\eta)}{d^+}\right). \quad (32)$$

Since $\frac{p\eta}{d^-} \leq 1$, (32) is ensured provided

$$\tau^-\tau^+(1-\beta) + 2\tau^- < \beta\tau^+$$

which in turn holds if each LHS term is respectively less than half of the RHS term. This leads to the condition

$$\tau^- < \min\left\{\frac{\beta}{2(1-\beta)}, \frac{\beta}{4}\tau^+\right\}.$$

Finally, plugging the choices $\beta = 1 - \frac{(1-2\eta)}{2k(1-\eta)} (\geq 3/4)$ and $\beta = \frac{1}{2}$ in the above equation, and combining it with the conditions derived for ensuring (26), we readily arrive (after minor simplifications) at the statements in the Lemma. ■

4.2.2 Spectral gap for the general case

For the general-sized clusters case, it is difficult to find the exact value of $\left\|(C^-)^{-1/2}C^+(C^-)^{-1/2}\right\|$. Therefore, in the following lemma, we show an upper bound on this quantity by bounding the spectral norms of C^+ and $(C^-)^{-1}$.

Lemma 22 (Bounding the spectral norm of $(C^-)^{-1}$ and C^+) *Recall $s := \min_{i\in[k]} n_i/n$. Then it holds true that*

$$\lambda_{\max}(C^+) \leq \tau^- + \frac{n\eta}{n(s(1-2\eta)+\eta) - (1-\eta)}, \tag{33}$$

$$\lambda_{\min}(C^-) \geq \tau^+ . \tag{34}$$

From the above two inequalities, it follows that

$$\left\|(C^-)^{-1/2}C^+(C^-)^{-1/2}\right\| \leq \frac{\lambda_{max}(C^+)}{\lambda_{min}(C^-)} \leq \frac{\tau^- + \frac{n\eta}{n(s(1-2\eta)+\eta)-(1-\eta)}}{\tau^+}.$$

The proof of the above lemma is deferred to Appendix D.

Remark 23 *It is difficult to obtain more precise bounds on $\lambda_{\max}(C^+)$ and $\lambda_{\min}(C^-)$, given the expressions for C^+ in* (17)*, and C^- in* (18)*. Clearly, a tighter bound on $\left\|(C^-)^{-1/2}C^+(C^-)^{-1/2}\right\|$ would yield a tighter analysis in the general case.*

Recall $l := \max_{i\in[k]} n_i/n$; with a slight abuse of notation, let $d_l^{\pm}$ denote the degree of the largest cluster (of size nl). As before, we now derive conditions on $\tau^+ > 0, \tau^- \geq 0$ which ensure $V_k(\overline{T}) = \Theta R$, or equivalently,

$$\lambda_{n-k+1}(\overline{T}) = \left\|(C^-)^{-1/2}C^+(C^-)^{-1/2}\right\| < \min_{i\in[k]} \frac{\alpha_i^+}{\alpha_i^-} = \frac{1+\tau^- + p(1-\eta)/d_l^+}{1+\tau^+ + p\eta/d_l^-} = \frac{\alpha_l^+}{\alpha_l^-} = \lambda_{n-k}(\overline{T}). \tag{35}$$

Additionally, we find sufficient conditions on $\tau^+ > 0, \tau^- \geq 0$ which ensure a lower bound on the *spectral gap* $\lambda_{n-k}(\overline{T}) - \lambda_{n-k+1}(\overline{T}) = \min_{i\in[k]} \frac{\alpha_i^+}{\alpha_i^-} - \left\|(C^-)^{-1/2}C^+(C^-)^{-1/2}\right\|$. These are shown in the following lemma.

Lemma 24 (Conditions on τ^+, τ^-, and Lower-Bound on Spectral Gap) *Suppose $n \geq \max\left\{\frac{2(1-\eta)}{s(1-2\eta)}, \frac{2\eta}{(1-l)(1-\eta)}\right\}$, then the following is true.*

1. *If* $\tau^+ > 0, \tau^- \geq 0$ *satisfy*

$$2\tau^- + \frac{4\eta}{s(1-2\eta)+2\eta} < \frac{s(1-2\eta)}{s(1-2\eta)+2\eta}\tau^+ \tag{36}$$

then $V_k(\overline{T}) = \Theta R$, *i.e.,* $\lambda_{n-k+1}(\overline{T}) = \left\|(C^-)^{-1/2}C^+(C^-)^{-1/2}\right\| < \frac{\alpha_l^+}{\alpha_l^-} = \lambda_{n-k}(\overline{T})$.

2. *For* $\beta = \frac{4\eta}{s(1-2\eta)+4\eta}$ *with* $0 < \eta < \frac{1}{2}$, *if* $\tau^+ > 0, \tau^- \geq 0$ *satisfy*

$$(1-\beta)\tau^-\tau^+ + 2\tau^- + \frac{4\eta}{s(1-2\eta)+2\eta} < \frac{\beta}{2}\left(\frac{s(1-2\eta)}{s(1-2\eta)+2\eta}\right)\tau^+ \tag{37}$$

then $V_k(\overline{T}) = \Theta R$, *and* $\left\|(C^-)^{-1/2}C^+(C^-)^{-1/2}\right\| < \beta\frac{\alpha_l^+}{\alpha_l^-}$, *i.e.,* $\lambda_{n-k}(\overline{T}) - \lambda_{n-k+1}(\overline{T}) > (1-\beta)\frac{\alpha_l^+}{\alpha_l^-}$. *Moreover, for* (37) *to hold, it suffices that*

$$\tau^+ > \frac{16\eta}{\beta s(1-2\eta)}, \quad \tau^- < \frac{\beta}{2}\left(\frac{s(1-2\eta)}{s(1-2\eta)+2\eta}\right)\min\left\{\frac{1}{4(1-\beta)}, \frac{\tau^+}{8}\right\}.$$

3. *The statement in part (*2*) also holds for the choice* $\beta = \frac{1}{2}$, *and provided* $\eta \leq \frac{s}{2s+4}$.

Proof From (35) and Lemma 22, it suffices to show for $\beta \in (0,1]$ that

$$\frac{\tau^- + \frac{\eta}{s(1-2\eta)+\eta-\frac{(1-\eta)}{n}}}{\tau^+} < \beta\left(\frac{1+\tau^-+p(1-\eta)/d_l^+}{1+\tau^++p\eta/d_l^-}\right). \tag{38}$$

For the stated condition on n, it is easy to verify that

$$n \geq \frac{2(1-\eta)}{s(1-2\eta)} \implies s(1-2\eta)+\eta-\frac{(1-\eta)}{n} \geq \frac{s(1-2\eta)}{2}+\eta,$$
$$n \geq \frac{2\eta}{(1-l)(1-\eta)} \implies \frac{p\eta}{d_l^-} \leq \frac{2\eta}{n(1-\eta)(1-l)} \leq 1.$$

Using these bounds in (38), observe that it suffices that τ^+, τ^- satisfy

$$\frac{\tau^- + \frac{2\eta}{s(1-2\eta)+2\eta}}{\tau^+} < \beta\left(\frac{1+\tau^-}{2+\tau^+}\right). \tag{39}$$

Then for $\beta = 1$, we readily see that (39) is equivalent to (36).

To establish the second part of the Lemma, we begin by rewriting (39) as

$$\begin{aligned}(1-\beta)\tau^+\tau^- + 2\tau^- + \frac{4\eta}{s(1-2\eta)+2\eta} &< \left(\beta - \frac{2\eta}{s(1-2\eta)+2\eta}\right)\tau^+ \\ &= \left[\frac{\beta s(1-2\eta)-2\eta(1-\beta)}{s(1-2\eta)+2\eta}\right]\tau^+, \end{aligned} \tag{40}$$

and observe that

$$\beta s(1-2\eta) \geq 4\eta(1-\beta) \iff \beta \geq \frac{4\eta}{s(1-2\eta)+4\eta} \tag{41}$$

This verifies (37) in the statement of the Lemma. The "moreover" part is established by ensuring that each term on the LHS of (37) is a sufficiently small fraction of the RHS term. In particular, it is enough to choose this fraction to be $1/4$ for the first two terms, and $1/2$ for the third term.

Finally, the third part of the Lemma can be shown in the same manner as the second part. The starting point is to ensure (40), and we simply observe that for $\beta = 1/2$, (41) is equivalent to $\eta \leq \frac{s}{2s+4}$. The rest follows identically. ■

4.3 Concentration bound for $\left\|T - \overline{T}\right\|$

In this section, we bound the "distance" between T and $\overline{T}$, i.e., $\left\|T - \overline{T}\right\|$. This is shown via individually bounding the terms $\left\|L^+_{sym} - \overline{L^+_{sym}}\right\|$, and $\left\|L^-_{sym} - \overline{L^-_{sym}}\right\|$. To this end, we first recall the following Theorem from Chung and Radcliffe (2011).

Theorem 25 (Bounding $\left\|L_{sym} - \overline{L_{sym}}\right\|$, (Chung and Radcliffe, 2011)) *Let L_{sym} denote the normalized Laplacian of a random graph, and $\overline{L_{sym}}$ the normalized Laplacian of the expected graph. Let δ be the minimum expected degree of the graph. Choose $\varepsilon > 0$. Then there exists a constant c_ε such that, if $\delta \geq c_\varepsilon \ln n$, then with probability at least $1-\varepsilon$, it holds true that*

$$\left\|L_{sym} - \overline{L_{sym}}\right\| \leq 2\sqrt{\frac{3\ln(4n/\varepsilon)}{\delta}}\,.$$

Remark 26 *A similar result appears in Imbuzeiro Oliveira (2009) for the (unsigned) inhomogeneous Erdős-Rényi model, where $\left\|L_{sym} - \overline{L_{sym}}\right\| = O(\sqrt{\ln n/d_0})$, with d_0 the smallest expected degree of the graph.*

Using Theorem 25, we readily obtain the following concentration bounds for $\left\|L^+_{sym} - \overline{L^+_{sym}}\right\|$ and $\left\|L^-_{sym} - \overline{L^-_{sym}}\right\|$.

Lemma 27 (Bounding $\left\|L^\pm_{sym} - \overline{L^\pm_{sym}}\right\|$) *Assuming $n \geq \max\left\{\frac{2(1-\eta)}{s(1-2\eta)}, \frac{2\eta}{(1-l)(1-\eta)}\right\}$, there exists a constant $c_\varepsilon > 0$ such that if $p \geq \frac{c_\varepsilon \ln n}{n} \max\left\{\frac{1}{s(1-2\eta)+2\eta}, \frac{2}{1-l}\right\}$, then with probability at least $1-2\varepsilon$,*

$$\left\|L^+_{sym} - \overline{L^+_{sym}}\right\| \leq 2\sqrt{\frac{6\ln(4n/\varepsilon)}{np[s(1-2\eta)+2\eta]}}, \qquad \text{and} \qquad \left\|L^-_{sym} - \overline{L^-_{sym}}\right\| \leq 2\sqrt{\frac{12\ln(4n/\varepsilon)}{np(1-l)}}.$$

Proof Note that the minimum expected degrees of the positive and negative subgraphs are given by d^+_s, d^-_l, respectively. For the stated condition on n, it is easily seen that

$$d^+_s \geq \frac{np}{2}\left[s(1-2\eta)+2\eta\right], \quad d^-_l \geq \frac{np}{2}(1-l)(1-\eta) \geq \frac{np(1-l)}{4}. \tag{42}$$

Invoking Theorem 25, and observing that $d_s^+, d_l^- \geq \frac{c_\varepsilon}{2} \ln n$ are ensured for the stated condition on p, the statement follows via the union bound. ■

Next, using the above lemma, we can upper bound $\|T - \overline{T}\|$. This will help us show that $V_k(T)$ and $V_k(\overline{T})$ are "close".

Lemma 28 (Bounding $\|T - \overline{T}\|$) *Let $P = (L_{sym}^- + \tau^+ I)$, $\overline{P} = (\overline{L_{sym}^-} + \tau^+ I)$, $Q = (L_{sym}^+ + \tau^- I)$, and $\overline{Q} = (\overline{L_{sym}^+} + \tau^- I)$. Assume that $\|P - \overline{P}\| \leq \Delta_P$, and $\|Q - \overline{Q}\| \leq \Delta_Q$. Then it holds true that*

$$\|T - \overline{T}\| \leq \frac{(\alpha_s^+ + \Delta_Q)}{\tau^+} \left(\frac{\Delta_P}{\tau^+} + 2\sqrt{\frac{\Delta_P}{\tau^+}} \right) + \frac{\Delta_Q}{\tau^+}$$

where $\alpha_s^+ = 1 + \tau^- + \frac{p(1-\eta)}{d_s^+}$ (see Lemma 17).

Proof Since $P, \overline{P}, Q, \overline{Q}$ are positive definite, therefore using Proposition 59, we obtain the bound

$$\|T - \overline{T}\| \leq \|P^{-1}\| \, \|Q\| \left(\|(\overline{P})^{-1}\| \, \|\overline{P} - P\| + 2 \left\|(\overline{P})^{-1/2}\right\| \|\overline{P} - P\|^{1/2} \right) + \|(\overline{P})^{-1}\| \, \|Q - \overline{Q}\| \,. \tag{43}$$

We know that $\|P^{-1}\| = 1/\tau^+ = \left\|\overline{P}^{-1}\right\|$ and $\|(\overline{P})^{-1/2}\| = 1/\sqrt{\tau^+}$. Moreover, $\|Q\| \leq \|\overline{Q}\| + \Delta_Q$ by Weyl's inequality (Weyl, 1912) (see Appendix B). Hence (43) simplifies to

$$\|T - \overline{T}\| \leq \frac{(\|\overline{Q}\| + \Delta_Q)}{\tau^+} \left(\frac{\Delta_P}{\tau^+} + 2\sqrt{\frac{\Delta_P}{\tau^+}} \right) + \frac{\Delta_Q}{\tau^+} \leq \frac{(\alpha_s^+ + \Delta_Q)}{\tau^+} \left(\frac{\Delta_P}{\tau^+} + 2\sqrt{\frac{\Delta_P}{\tau^+}} \right) + \frac{\Delta_Q}{\tau^+} \,,$$

where the last inequality can be verified by examining the expression of $\overline{Q}$ in (24), and noting from the definition of C^+ that $\|C^+\| < \max\{\alpha_1^+, ..., \alpha_k^+\} = \alpha_s^+$ holds (via Weyl's inequality). ■

4.4 Estimating $V_k(\overline{T})$ and $G_k(\overline{T})$ up to a rotation

We are now ready to combine the results of the previous sections to show that if n, p are large enough, then the distance between the subspaces spanned by $V_k(T)$ and $V_k(\overline{T})$ is small, i.e., there exists an orthonormal matrix O such that $V_k(T)$ is close to $V_k(\overline{T})O$. For τ^+, τ^- chosen suitably, we have seen in Lemma 24 that $V_k(\overline{T}) = \Theta R$ for a rotation R, hence this suggests that the rows of $V_k(T)$ will then also approximately preserve the clustering structure of $V_k(\overline{T})$.

With $P, \overline{P}, Q, \overline{Q}$ as defined in Lemma 28 recall from (4), (7) that $G_k(T), G_k(\overline{T})$ can be written as

$$G_k(\overline{T}) = \overline{P}^{-1/2} V_k(\overline{T}), \quad G_k(T) = P^{-1/2} V_k(T). \tag{44}$$

Therefore if $V_k(\overline{T}) = \Theta R$, then using the expression for $\overline{P}$ from (24) we see that $G_k(\overline{T}) = \Theta(C^-)^{-1/2}R$, and thus the rows of $G_k(\overline{T})$ also preserve the ground truth clustering structure. Moreover, if $\left\|V_k(T) - V_k(\overline{T})O\right\|$ is small, then it can be shown to imply a bound on $\left\|G_k(T) - G_k(\overline{T})O\right\|$. Hence the rows of $G_k(T)$ will approximately preserve the clustering structure of $G_k(\overline{T})$.

Before stating the theorem, let us define the terms

$$C_1(\tau^+, \tau^-) = 3\left(\frac{(3+\tau^-)(2\sqrt{\tau^+}+1)+\tau^+}{(\tau^+)^2}\right), \quad C_2(s,\eta,l) = \max\left\{\frac{1}{s(1-2\eta)+2\eta}, \frac{2}{1-l}\right\}. \tag{45}$$

Theorem 29 *Assuming $n \geq \max\left\{\frac{2(1-\eta)}{s(1-2\eta)}, \frac{2\eta}{(1-l)(1-\eta)}\right\}$, suppose $\tau^+ > 0, \tau^- \geq 0$ are chosen to satisfy*

$$\tau^+ > \frac{16\eta}{\beta s(1-2\eta)}, \quad \tau^- < \frac{\beta}{2}\left(\frac{s(1-2\eta)}{s(1-2\eta)+2\eta}\right)\min\left\{\frac{1}{4(1-\beta)}, \frac{\tau^+}{8}\right\}$$

where β, η satisfy one of the following conditions.

1. *$\beta = \frac{4\eta}{s(1-2\eta)+4\eta}$ and $0 < \eta < \frac{1}{2}$, or*
2. *$\beta = \frac{1}{2}$ and $\eta \leq \frac{s}{2s+4}$.*

Then $V_k(\overline{T}) = \Theta R$ and $G_k(\overline{T}) = \Theta(C^-)^{-1/2}R$ where R is a rotation matrix, and $C^- \succ 0$ is as defined in (18). Moreover, for any $\varepsilon, \delta \in (0,1)$, there exists a constant $\widetilde{c}_\varepsilon > 0$ such that the following is true. If p satisfies

$$p \geq \max\left\{\widetilde{c}_\varepsilon C_2(s,\eta,l), \frac{256 C_1^4(\tau^+,\tau^-)(2+\tau^+)^4}{\delta^4(1+\tau^-)^4(1-\beta)^4}C_2(s,\eta,l), \frac{81}{(1-l)\delta^4}\right\}\frac{\ln(4n/\varepsilon)}{n},$$

with $C_1(\cdot), C_2(\cdot)$ as in (45), then with probability at least $1-2\varepsilon$, there exists an orthogonal matrix $O \in \mathbb{R}^{k\times k}$ such that

$$\left\|V_k(T) - V_k(\overline{T})O\right\| \leq \delta, \qquad \text{and} \qquad \left\|G_k(T) - G_k(\overline{T})O\right\| \leq \frac{\delta}{\sqrt{\tau^+}} + \frac{\delta}{(\tau^+)^2}.$$

Proof We will first simplify the upper bound on $\left\|T - \overline{T}\right\|$ in Lemma 28, starting by bounding α_s^+. If $n \geq \frac{2(1-\eta)}{s(1-2\eta)}$, it is easy to verify that $\frac{(1-\eta)p}{d_s^+} \leq 1$ which implies $\alpha_s^+ \leq 2+\tau^-$. Moreover, we observe from Lemma 27 that $\Delta_P, \Delta_Q \leq 1$ is ensured if $p \geq \widetilde{c}_\varepsilon C_2(s,\eta,l)\frac{\ln(4n/\varepsilon)}{n}$ where $\widetilde{c}_\varepsilon = \max\{24, c_\varepsilon\}$. These considerations altogether imply

$$\begin{aligned}
\left\|T - \overline{T}\right\| &\leq \frac{(3+\tau^-)(2\sqrt{\tau^+}+1)}{(\tau^+)^2}\sqrt{\Delta_P} + \frac{\Delta_Q}{\tau^+} \\
&\leq \frac{(3+\tau^-)(2\sqrt{\tau^+}+1)+\tau^+}{(\tau^+)^2}\max\left\{\sqrt{\Delta_P}, \sqrt{\Delta_Q}\right\} \\
&\leq C_1(\tau^+,\tau^-)C_2^{1/4}(s,\eta,l)\left(\frac{\ln(4n/\varepsilon)}{np}\right)^{1/4}
\end{aligned} \tag{46}$$

where in the penultimate inequality we used $\Delta_Q \leq \sqrt{\Delta_Q}$, and the last inequality uses Lemma 27.

Next, we will use the Davis-Kahan theorem (Davis and Kahan, 1970) (see Appendix B) for bounding the distance $\left\|(I - V_k(\overline{T})V_k(\overline{T})^T)V_k(T)\right\|$. Applied to our setup, it yields

$$\left\|(I - V_k(\overline{T})V_k(\overline{T})^T)V_k(T)\right\| \leq \frac{\left\|T - \overline{T}\right\|}{\lambda_{n-k+1}(T) - \lambda_{n-k}(\overline{T})}, \tag{47}$$

provided $\lambda_{n-k+1}(T) - \lambda_{n-k}(\overline{T}) > 0$. From Weyl's inequality, we know that $\lambda_{n-k+1}(T) \geq \lambda_{n-k+1}(\overline{T}) - \left\|T - \overline{T}\right\|$. Moreover, under the stated conditions on τ^+, τ^-, we obtain from Lemma 24 the bound

$$\lambda_{n-k+1}(\overline{T}) - \lambda_{n-k}(\overline{T}) \geq (1-\beta)\frac{\alpha_l^+}{\alpha_l^-} \geq (1-\beta)\left(\frac{1+\tau^-}{2+\tau^+}\right),$$

where in the last inequality we used the simplifications $p(1-\eta)/d_l^+ \geq 0$ and $p\eta/d_l^- \leq 1$ in the expressions for α_l^+, α_l^-. Hence using (46), we observe that if

$$\begin{aligned} C_1(\tau^+,\tau^-)C_2^{1/4}(s,\eta,l)\left(\frac{\ln(4n/\varepsilon)}{np}\right)^{1/4} &\leq \left(\frac{1-\beta}{2}\right)\left(\frac{1+\tau^-}{2+\tau^+}\right) \\ \iff p &\geq \left(\frac{16C_1^4(\tau^+,\tau^-)C_2(s,\eta,l)(2+\tau^+)^4}{(1+\tau^-)^4(1-\beta)^4}\right)\frac{\ln(4n/\varepsilon)}{n}, \end{aligned}$$

then the RHS of (47) can be bounded as

$$\left\|(I - V_k(\overline{T})V_k(\overline{T})^T)V_k(T)\right\| \leq \frac{2(2+\tau^+)}{(1+\tau^-)(1-\beta)}C_1(\tau^+,\tau^-)C_2^{1/4}(s,\eta,l)\left(\frac{\ln(4n/\varepsilon)}{np}\right)^{1/4}.$$

It follows that there exists an orthogonal matrix $O \in \mathbb{R}^{k\times k}$ so that

$$\begin{aligned} \left\|V_k(T) - V_k(\overline{T})O\right\| &\leq 2\left\|(I - V_k(\overline{T})V_k(\overline{T})^T)V_k(T)\right\| \quad (\text{ using Proposition 57}) \\ &\leq \frac{4(2+\tau^+)}{(1+\tau^-)(1-\beta)}C_1(\tau^+,\tau^-)C_2^{1/4}(s,\eta,l)\left(\frac{\ln(4n/\varepsilon)}{np}\right)^{1/4} \\ &\leq \delta \end{aligned}$$

for the stated bound on p. This establishes the first part of the Theorem.

In order to bound $\left\|G_k(T) - G_k(\overline{T})O\right\|$, we obtain from (44) that

$$\begin{aligned} \left\|G_k(T) - G_k(\overline{T})O\right\| &= \left\|P^{-1/2}(V_k(T) - V_k(\overline{T})O) + (P^{-1/2} - \overline{P}^{-1/2})V_k(\overline{T})O\right\| \\ &\leq \underbrace{\left\|P^{-1/2}\right\|}_{(\tau^+)^{-1/2}}\underbrace{\left\|V_k(T) - V_k(\overline{T})O\right\|}_{\leq\delta} + \left\|P^{-1/2} - \overline{P}^{-1/2}\right\|\underbrace{\left\|V_k(\overline{T})\right\|}_{=1} \\ &\leq \frac{\delta}{\sqrt{\tau^+}} + \left\|P^{-1/2} - \overline{P}^{-1/2}\right\|. \end{aligned} \tag{48}$$

The term $\left\|P^{-1/2} - \overline{P}^{-1/2}\right\|$ can be bounded as

$$\begin{aligned}\left\|P^{-1/2} - \overline{P}^{-1/2}\right\| = \left\|P^{-1}(P^{1/2} - \overline{P}^{1/2})\overline{P}^{-1}\right\| &\leq \frac{\left\|P^{1/2} - \overline{P}^{1/2}\right\|}{(\tau^+)^2} \\ &\leq \frac{\left\|P - \overline{P}\right\|^{1/2}}{(\tau^+)^2} \\ &\leq \frac{3}{(\tau^+)^2}\left[\frac{\ln(4n/\varepsilon)}{np(1-l)}\right]^{1/4}, \end{aligned} \tag{49}$$

where the penultimate inequality uses Proposition 58, and the last inequality follows from Lemma 27 with a minor simplification of the constant. Plugging (49) in (48) leads to the stated bound for $p \geq \frac{81}{(1-l)\delta^4}\frac{\ln(4n/\varepsilon)}{n}$. ■

4.5 Clustering sparse graphs

We now turn our attention to the sparse regime where $p = o(\ln n)/n$. In this regime, Lemma 27 is no longer applicable since it requires $p = \Omega\left(\frac{\ln n}{n}\right)$. In fact, it is not difficult to see that the matrices $L^{\pm}_{sym}$ will not concentrate around $\overline{L^{\pm}_{sym}}$ in this sparsity regime. To circumvent this issue, we will aim to show that the normalized Laplacian $L^{\pm}_{sym,\gamma^{\pm}}$ corresponding to the regularized adjacencies $A^{\pm}_{\gamma^{\pm}} := A^{\pm} + \frac{\gamma^{\pm}}{n}\mathbb{1}\mathbb{1}^\top$ concentrate around $\overline{L^{\pm}_{sym}}$, for carefully chosen values of γ^+, γ^-.

To show this, we rely on the following theorem from Le et al. (2017), which states that the symmetric Laplacian $L_{sym,\gamma}$ of the regularized adjacency matrix $A_\gamma := A + \frac{\gamma}{n}\mathbb{1}\mathbb{1}^\top$ is close to the symmetric Laplacian $\overline{L_{sym,\gamma}}$ of the expected regularized adjacency matrix, for inhomogeneous Erdős-Rényi graphs.

Theorem 30 (Theorem 4.1 of Le et al. (2017)) *Consider a random graph from the inhomogeneous Erdős-Rényi model ($G = (n, p_{ij})$), and let $d = \max_{p_{ij}} np_{ij}$. Choose a number $\gamma > 0$. Then, for any $r \geq 1$, C being an absolute constant, with probability at least $1 - e^{-r}$*

$$\left\|L_{sym,\gamma} - \overline{L_{sym,\gamma}}\right\| \leq \frac{Cr^2}{\sqrt{\gamma}}\left(1 + \frac{d}{\gamma}\right)^{5/2}. \tag{50}$$

The above result leads to a bound on the distance between $L_{sym,\gamma}$ and the normalized Laplacian $\overline{L_{sym}}$ of the expected (un-regularized) adjacency matrix.

Theorem 31 (Concentration of Regularized Laplacians) *Consider a random graph from the inhomogeneous Erdős-Rényi model ($G = (n, p_{ij})$), and let $d = \max_{p_{ij}} np_{ij}$, $d_{\min} = \min_i \sum_j p_{ij}$. Choose a number $\gamma > 0$. Then, for any $r \geq 1$, C being an absolute constant, with probability at least $1 - e^{-r}$*

$$\left\|L_{sym,\gamma} - \overline{L_{sym}}\right\| \leq \frac{Cr^2}{\sqrt{\gamma}}\left(1 + \frac{d}{\gamma}\right)^{5/2} + 3\sqrt{\frac{\gamma}{d_{\min} + \gamma}}. \tag{51}$$

Proof To establish the above lemma we make use of triangle inequality, where we use the fact that $\left\|L_{sym,\gamma} - \overline{L_{sym}}\right\| \leq \left\|L_{sym,\gamma} - \overline{L_{sym,\gamma}}\right\| + \left\|\overline{L_{sym,\gamma}} - \overline{L_{sym}}\right\|$. We know the bound on the first term on the RHS from Lemma 30 (which holds with probability $1 - e^{-r}$). To bound the second term on the RHS, note that

$$\begin{aligned}
\left\|\overline{L_{sym,\gamma}} - \overline{L_{sym}}\right\| &= \left\|\overline{D}^{-1/2}\overline{AD}^{-1/2} - \overline{D}_\gamma^{-1/2}\overline{A}_\gamma\overline{D}_\gamma^{-1/2}\right\| \\
&= \left\|\overline{D}^{-1/2}\overline{AD}^{-1/2} - \overline{D}_\gamma^{-1/2}\overline{AD}_\gamma^{-1/2} + \overline{D}_\gamma^{-1/2}\overline{AD}_\gamma^{-1/2} - \overline{D}_\gamma^{-1/2}\overline{A}_\gamma\overline{D}_\gamma^{-1/2}\right\| \\
&\leq \left\|\overline{D}^{-1/2}\overline{AD}^{-1/2} - \overline{D}_\gamma^{-1/2}\overline{AD}_\gamma^{-1/2}\right\| + \left\|\overline{D}_\gamma^{-1/2}\overline{AD}_\gamma^{-1/2} - \overline{D}_\gamma^{-1/2}\overline{A}_\gamma\overline{D}_\gamma^{-1/2}\right\|.
\end{aligned}$$

The second term of the inequality can be easily bounded as follows.

$$\left\|\overline{D}_\gamma^{-1/2}\overline{AD}_\gamma^{-1/2} - \overline{D}_\gamma^{-1/2}\overline{A}_\gamma\overline{D}_\gamma^{-1/2}\right\| \leq \left\|\overline{D}_\gamma^{-1/2}\right\|^2 \left\|\overline{A} - \overline{A}_\gamma\right\| \leq \frac{\gamma}{d_{\min} + \gamma} \leq \sqrt{\frac{\gamma}{d_{\min} + \gamma}}.$$

To analyse the first term, we observe that

$$\begin{aligned}
&\left\|\overline{D}^{-1/2}\overline{AD}^{-1/2} - \overline{D}_\gamma^{-1/2}\overline{AD}_\gamma^{-1/2}\right\| = \\
&\quad \left\|\overline{D}^{-1/2}\overline{AD}^{-1/2} - \overline{D}_\gamma^{-1/2}\overline{D}^{1/2}\overline{D}^{-1/2}\overline{AD}^{-1/2}\overline{D}^{1/2}\overline{D}_\gamma^{-1/2}\right\| \\
&= \left\|(I - \overline{L_{sym}})(I - \overline{D}^{1/2}\overline{D}_\gamma^{-1/2}) + (I - \overline{D}_\gamma^{-1/2}\overline{D}^{1/2})(I - \overline{L_{sym}})\overline{D}^{1/2}\overline{D}_\gamma^{-1/2}\right\| \\
&\leq \left\|I - \overline{D}^{1/2}\overline{D}_\gamma^{-1/2}\right\| + \left\|I - \overline{D}_\gamma^{-1/2}\overline{D}^{1/2}\right\| \left\|\overline{D}^{1/2}\overline{D}_\gamma^{-1/2}\right\| \\
&\leq \left(1 - \sqrt{\frac{d_{\min}}{d_{\min} + \gamma}}\right) + \left(1 - \sqrt{\frac{d_{\min}}{d_{\min} + \gamma}}\right) \\
&\leq 2\sqrt{\frac{\gamma}{d_{\min} + \gamma}},
\end{aligned}$$

where in the first inequality we use the fact that $\left\|I - \overline{L_{sym}}\right\| \leq 1$, and in the last inequality we use the fact that for two numbers $a, b > 0$ if $a > b$ then $\sqrt{a} - \sqrt{b} \leq \sqrt{a - b}$. We have all the components to plug into the triangle inequality, which yields the desired statement of the theorem. ∎

We now translate Theorem 31 to our setting for G^+, G^- and show that if $p = \Omega(1/n)$ for n large enough, then for the choices $\gamma^+, \gamma^- \asymp (np)^{6/7}$, the bounds $\left\|L^\pm_{sym,\gamma^\pm} - \overline{L^\pm_{sym}}\right\| = O\left(\frac{1}{(np)^{1/14}}\right)$ hold with sufficiently high probability.

Lemma 32 *Let $n \geq \max\left\{\frac{2(1-\eta)}{s(1-2\eta)}, \frac{2\eta}{(1-\eta)(1-l)}\right\}$ and $p \geq \frac{1}{n(1-\eta)}$. Then for the choices $\gamma^+, \gamma^- = [np(1-\eta)]^{6/7}$, and any $r \geq 1$, there exists a constant $C > 0$ such that with probability at least $1 - 2e^r$, it holds true that*

$$\left\|L^+_{sym,\gamma^+} - \overline{L^+_{sym}}\right\| \leq \left(2^{5/2}Cr^2 + \frac{3\sqrt{2}}{\sqrt{s(1-2\eta) + 2\eta}}\right)\frac{1}{[np(1-\eta)]^{1/14}}, \tag{52}$$

$$\left\|L^-_{sym,\gamma^-} - \overline{L^-_{sym}}\right\| \leq \left(2^{5/2}Cr^2 + \frac{6}{\sqrt{1-l}}\right)\frac{1}{[np(1-\eta)]^{1/14}}. \tag{53}$$

Proof We will apply Theorem 31 to the subgraphs G^+, G^-. Let us denote $d^{\pm}$ to be the quantity $\max_{ij} np_{ij}$, and $d^{\pm}_{min}$ to be the minimum expected degree for the positive and negative subgraphs, respectively. From the SSBM model, it can be verified that $d^{\pm} = np(1-\eta)$. We also know that $d^+_{\min} = d^+_s$ and $d^-_{\min} = d^-_l$, where for the stated condition on n, d^+_s, d^-_l satisfy the bounds in (42). The latter can be written as

$$d^+_{\min} \geq \frac{d^+}{2}[s(1-2\eta)+2\eta], \qquad d^-_{\min} \geq \frac{d^-(1-l)}{4}.$$

Let us denote $C_3(s,\eta) = s(1-2\eta)+2\eta$ for convenience. In order to show (52), we obtain from Theorem 31 that, with probability at least $1-e^{-r}$,

$$\begin{aligned}\left\|L^+_{sym,\gamma^+} - \overline{L^+_{sym}}\right\| &\leq \frac{Cr^2}{\sqrt{\gamma^+}}\left(1+\frac{d^+}{\gamma^+}\right)^{5/2} + 3\sqrt{\frac{\gamma^+}{d^+_{\min}+\gamma^+}} \\ &\leq \frac{Cr^2}{\sqrt{\gamma^+}}\left(1+\frac{d^+}{\gamma^+}\right)^{5/2} + 3\sqrt{\frac{\gamma^+}{C_3(s,\eta)d^+}},\end{aligned}$$

where the last inequality uses $d^+_s + \gamma^+ \geq d^+_s$. Now note that if $\gamma^+ \leq d^+$, then the above bound simplifies to

$$\left\|L^+_{sym,\gamma^+} - \overline{L^+_{sym}}\right\| \leq \frac{2^{5/2}Cr^2(d^+)^{5/2}}{(\gamma^+)^3} + \frac{3\sqrt{2}}{\sqrt{C_3(s,\eta)}}\sqrt{\frac{\gamma^+}{d^+}}. \tag{54}$$

Choosing γ^+ such that $\frac{(d^+)^{5/2}}{(\gamma^+)^3} = \sqrt{\frac{\gamma^+}{d^+}}$, or equivalently, $\gamma^+ = (d^+)^{6/7}$, and plugging this in (54), we arrive at (52). Clearly, $\gamma^+ \leq d^+$ is equivalent to the stated condition on p. The bound in (53) follows in an identical manner and is omitted. ■

We are now in a position to write the bound on $\left\|T_{\gamma^+,\gamma^-} - \overline{T}\right\|$ in terms of $\left\|L^{\pm}_{sym,\gamma^{\pm}} - \overline{L^{\pm}_{sym}}\right\|$, in a completely analogous manner to Lemma 28.

Lemma 33 (Adapting Lemma 28 for the sparse regime) *Let* $P_{\gamma^-} = (L^-_{sym,\gamma^-} + \tau^+ I)$, $\overline{P} = (\overline{L^-_{sym}} + \tau^+ I)$, $Q_{\gamma^+} = (L^+_{sym,\gamma^+} + \tau^- I)$, *and* $\overline{Q} = (\overline{L^+_{sym}} + \tau^- I)$. *Assume that* $\left\|P_{\gamma^-} - \overline{P}\right\| \leq \Delta_{P_{\gamma^-}}$, $\left\|Q_{\gamma^+} - \overline{Q}\right\| \leq \Delta_{Q_{\gamma^+}}$. *Then it holds true that*

$$\left\|T_{\gamma^+,\gamma^-} - \overline{T}\right\| \leq \frac{(\alpha^+_s + \Delta_{Q_{\gamma^+}})}{\tau^+}\left(\frac{\Delta_{P_{\gamma^-}}}{\tau^+} + 2\sqrt{\frac{\Delta_{P_{\gamma^-}}}{\tau^+}}\right) + \frac{\Delta_{Q_{\gamma^+}}}{\tau^+},$$

where $\alpha^+_s = 1 + \tau^- + \frac{p(1-\eta)}{d^+_s}$ *(see Lemma 17).*

Next, we derive the main theorem for SPONGE$_{sym}$ in the sparse regime, which is the analogue of Theorem 29. The first part of the Theorem remains unchanged, i.e., for n large enough and τ^+, τ^- chosen suitably, we have $V_k(\overline{T}) = \Theta R$ and $G_k(\overline{T}) = \Theta(C^-)^{-1/2}R$ for a $k \times k$ rotation R, and $C^- \succ 0$. The remaining arguments follow the same outline of Theorem 29, i.e., (a) using Lemma 33 and Lemma 32 to obtain a concentration bound on

$\left\|T_{\gamma^+,\gamma^-} - \overline{T}\right\|$ (when $p = \Omega(1/n)$), and (b) using the Davis-Kahan theorem to show that the column span of $V_k(T_{\gamma^+,\gamma^-})$ is close to $V_k(\overline{T})$. The latter bound then implies that $G_k(T_{\gamma^+,\gamma^-})$ is close (up to a rotation) to $G_k(\overline{T})$, where we recall

$$G_k(\overline{T}) = \overline{P}^{-1/2} V_k(\overline{T}), \quad G_k(T_{\gamma^+,\gamma^-}) = P_{\gamma^-}^{-1/2} V_k(T_{\gamma^+,\gamma^-}) \tag{55}$$

with $P_{\gamma^-}, \overline{P}$ as defined in Lemma 33.

Theorem 34 *Assuming* $n \geq \max\left\{\frac{2(1-\eta)}{s(1-2\eta)}, \frac{2\eta}{(1-l)(1-\eta)}\right\}$, *suppose* $\tau^+ > 0, \tau^- \geq 0$ *are chosen to satisfy*

$$\tau^+ > \frac{16\eta}{\beta s(1-2\eta)}, \quad \tau^- < \frac{\beta}{2}\left(\frac{s(1-2\eta)}{s(1-2\eta)+2\eta}\right) \min\left\{\frac{1}{4(1-\beta)}, \frac{\tau^+}{8}\right\}$$

where β, η *satisfy one of the following conditions.*

1. $\beta = \frac{4\eta}{s(1-2\eta)+4\eta}$ *and* $0 < \eta < \frac{1}{2}$, *or*
2. $\beta = \frac{1}{2}$ *and* $\eta \leq \frac{s}{2s+4}$.

Then $V_k(\overline{T}) = \Theta R$ *and* $G_k(\overline{T}) = \Theta(C^-)^{-1/2} R$ *where* R *is a rotation matrix, and* $C^- \succ 0$ *is as defined in* (18). *Moreover, there exists a constant* $C > 0$ *such that for* $r \geq 1$ *and* $\delta \in (0,1)$, *if* p *satisfies*

$$p \geq \max\left\{1, \left(\frac{4C_1(\tau^+,\tau^-)(2+\tau^+)}{3(\tau^+)^2(1-\beta)(1+\tau^-)}\right)^{28}\right\} \frac{C_4^{14}(r,s,\eta,l)}{\delta^{28}(1-\eta)n},$$

and $\gamma^+, \gamma^- = [np(1-\eta)]^{6/7}$, *then with probability at least* $1 - 2e^{-r}$, *there exists a rotation* $O \in \mathbb{R}^{k\times k}$ *so that*

$$\left\|V_k(T_{\gamma^+,\gamma^-}) - V_k(\overline{T})O\right\| \leq \delta, \qquad \left\|G_k(T_{\gamma^+,\gamma^-}) - G_k(\overline{T})O\right\| \leq \frac{\delta}{\sqrt{\tau^+}} + \frac{\delta}{(\tau^+)^2}.$$

Here, $C_4(r,s,\eta,l) := 2^{5/2}Cr^2 + 3\sqrt{2C_2(s,\eta,l)}$ *with* $C_2(s,\eta,l)$ *as defined in* (45).

Proof We will first simplify the upper bound on $\left\|T_{\gamma^+,\gamma^-} - \overline{T}\right\|$ in Lemma 33. Note that $n \geq \frac{2(1-\eta)}{s(1-2\eta)}$ implies $\alpha_s^+ \leq 2+\tau^-$, and moreover, we can bound $\left\|L^{\pm}_{sym,\gamma^\pm} - \overline{L^{\pm}_{sym}}\right\|$ uniformly (from (52), (53)) as

$$\left\|L^{\pm}_{sym,\gamma^\pm} - \overline{L^{\pm}_{sym}}\right\| \leq \frac{2^{5/2}Cr^2 + 3\sqrt{2C_2(s,\eta,l)}}{[np(1-\eta)]^{1/14}} \leq \frac{C_4(r,s,\eta,l)}{[np(1-\eta)]^{1/14}} \ (= \Delta_{P_{\gamma^-}}, \Delta_{Q_{\gamma^+}}). \tag{56}$$

Note that $\Delta_{P_{\gamma^-}}, \Delta_{Q_{\gamma^+}} \leq 1$ if $p \geq \frac{C_4^{14}(r,s,\eta,l)}{n(1-\eta)}$. Under these considerations, the bound in Lemma 33 simplifies to

$$\begin{aligned}\left\|T_{\gamma^+,\gamma^-} - \overline{T}\right\| &\leq \frac{(3+\tau^-)(2\sqrt{\tau^+}+1)+\tau^+}{(\tau^+)^2} \max\left\{\sqrt{\Delta_{P_{\gamma^-}}}, \sqrt{\Delta_{Q_{\gamma^+}}}\right\} \\ &= \frac{C_1(\tau^+,\tau^-)\sqrt{C_4(r,s,\eta,l)}}{3(\tau^+)^2[np(1-\eta)]^{1/28}}.\end{aligned}$$

Following the steps in the proof of Theorem 29, we observe that

$$\left\|T_{\gamma^+,\gamma^-} - \overline{T}\right\| \leq \frac{1}{2}(\lambda_{n-k+1}(T_{\gamma^+,\gamma^-}) - \lambda_{n-k}(\overline{T})),$$

is guaranteed to hold, provided

$$\begin{aligned}
&\frac{C_1(\tau^+,\tau^-)\sqrt{C_4(r,s,\eta,l)}}{3(\tau^+)^2[np(1-\eta)]^{1/28}} \leq (\frac{1-\beta}{2})\left(\frac{1+\tau^-}{2+\tau^+}\right) \\
\iff\ & p \geq \left(\frac{2C_1(\tau^+,\tau^-)(2+\tau^+)}{3(\tau^+)^2(1-\beta)(1+\tau^-)}\right)^{28} \frac{C_4^{14}(r,s,\eta,l)}{n(1-\eta)}.
\end{aligned}$$

Then, we obtain via the Davis-Kahan theorem that there exists an orthogonal matrix $O \in \mathbb{R}^{k\times k}$ such that

$$\begin{aligned}
\left\|V_k(T_{\gamma^+,\gamma^-}) - V_k(\overline{T})O\right\| &\leq \frac{4\left\|T_{\gamma^+,\gamma^-} - \overline{T}\right\|}{\lambda_{n-k+1}(\overline{T}) - \lambda_{n-k}(\overline{T})} \\
&\leq \frac{4C_1(\tau^+,\tau^-)\sqrt{C_4(r,s,\eta,l)}(2+\tau^+)}{3(\tau^+)^2[np(1-\eta)]^{1/28}(1-\beta)(1+\tau^-)} \leq \delta,
\end{aligned}$$

for the stated bound on p in the theorem. This establishes the first part of the theorem.

In order to bound $\left\|G_k(T_{\gamma^+,\gamma^-}) - G_k(\overline{T})O\right\|$, first observe that

$$\begin{aligned}
\left\|G_k(T_{\gamma^+,\gamma^-}) - G_k(\overline{T})O\right\| &= \left\|P_{\gamma^-}^{-1/2}(V_k(T_{\gamma^+,\gamma^-}) - V_k(\overline{T})O) + (P_{\gamma^-}^{-1/2} - \overline{P}^{-1/2})V_k(\overline{T})O\right\| \\
&\leq \underbrace{\left\|P_{\gamma^-}^{-1/2}\right\|}_{\leq(\tau^+)^{-1/2}} \underbrace{\left\|V_k(T_{\gamma^+,\gamma^-}) - V_k(\overline{T})O\right\|}_{\leq\delta} + \left\|P^{-1/2} - \overline{P}^{-1/2}\right\| \underbrace{\left\|V_k(\overline{T})\right\|}_{=1} \\
&\leq \frac{\delta}{\sqrt{\tau^+}} + \left\|P_{\gamma^-}^{-1/2} - \overline{P}^{-1/2}\right\|. \qquad (57)
\end{aligned}$$

The second term $\left\|P_{\gamma^-}^{-1/2} - \overline{P}^{-1/2}\right\|$ can be bounded as

$$\begin{aligned}
\left\|P_{\gamma^-}^{-1/2} - \overline{P}^{-1/2}\right\| &= \left\|P_{\gamma^-}^{-1}(P_{\gamma^-}^{1/2} - \overline{P}^{1/2})\overline{P}^{-1}\right\| \qquad (58) \\
&\leq \frac{\left\|P_{\gamma^-}^{1/2} - \overline{P}^{1/2}\right\|}{(\tau^+)^2} \leq \frac{\left\|P_{\gamma^-} - \overline{P}\right\|^{1/2}}{(\tau^+)^2} \leq \frac{\sqrt{C_4(r,s,\eta,l)}}{(\tau^+)^2[np(1-\eta)]^{1/28}}, \qquad (59)
\end{aligned}$$

where the penultimate inequality uses Proposition 58, and the last inequality uses (56). Plugging (58) into (57) leads to the stated bound for $p \geq \frac{C_4^{14}(r,s,\eta,l)}{n(1-\eta)\delta^{28}}$. ■

4.6 Mis-clustering rate from k-means

We now analyze the mis-clustering error rate when we apply a $(1+\xi)$-approximate k-means algorithm (e.g., (Kumar et al., 2004)) on the rows of $G_k(T)$ (respectively, $G_k(T_{\gamma^+,\gamma^-})$ in the sparse regime). To this end, we rely on the following result from Lei and Rinaldo (2015),

which when applied to our setting, yields that the mis-clustering error is bounded by the estimation error $\left\|G_k(T) - G_k(\overline{T})O\right\|_F^2$ (or $\left\|G_k(T_{\gamma^+,\gamma^-}) - G_k(\overline{T})O\right\|_F^2$ in the sparse setting). By an $(1+\xi)$-approximate algorithm, we mean an algorithm that is provably within an $(1+\xi)$ factor of the cost of the optimal solution achieved by k-means.

Lemma 35 (Lemma 5.3 of Lei and Rinaldo (2015), Approx. k-means error) *For any $\xi > 0$, and any two matrices $\overline{U}, U$, such that $\overline{U} = \overline{\Theta}\overline{X}$ with $(\overline{\Theta}, \overline{X}) \in \mathbb{M}_{n\times k} \times \mathbb{R}^{k\times k}$, let $(\widetilde{\Theta}, \widetilde{X}) \in \mathbb{M}_{n\times k} \times \mathbb{R}^{k\times k}$ be a $(1+\xi)$-approximate solution to the k-means problem $\min_{\Theta\in\mathbb{M}_{n\times k}, X\in\mathbb{R}^{k\times k}} \|\Theta X - U\|_F^2$ so that*

$$\left\|\widetilde{\Theta}\widetilde{X} - U\right\|_F^2 \leq (1+\xi) \min_{\Theta\in\mathbb{M}_{n\times k}, X\in\mathbb{R}^{k\times k}} \|\Theta X - U\|_F^2$$

and $\widetilde{U} = \widetilde{\Theta}\widetilde{X}$. For any $\delta_i \leq \min_{i'\neq i} \left\|\overline{X}_{i'*} - \overline{X}_{i*}\right\|$, define

$$S_i = \left\{ j \in C_i \ : \ \left\|\widetilde{U}_{j*} - \overline{U}_{j*}\right\| \geq \delta_i/2 \right\}, \qquad \text{then,}$$

$$\sum_{i=1}^{k} |S_i|\, \delta_i^2 \leq 4(4+2\xi) \left\|U - \overline{U}\right\|_F^2 . \tag{60}$$

Moreover, if

$$(16+8\xi) \left\|U - \overline{U}\right\|_F^2 / \delta_i^2 < n_i \qquad \forall i \in [k], \tag{61}$$

then there exists a $k\times k$ permutation matrix π such that $\widetilde{\Theta}_G = \overline{\Theta}_G\pi$, where $G = \cup_{i=1}^k (C_i \backslash S_i)$.

Combining Lemma 35 with the perturbation results of Theorem 29 and Theorem 34, we readily arrive at mis-clustering error bounds for SPONGE$_{sym}$.

Theorem 36 (Mis-clustering error for SPONGE$_{sym}$) *Under the notation and assumptions of Theorem 29, let $(\widetilde{\Theta}, \widetilde{X}) \in \mathbb{M}_{n\times k} \times \mathbb{R}^{k\times k}$ be a $(1+\xi)$-approximate solution to the k-means problem $\min_{\Theta\in\mathbb{M}_{n\times k}, X\in\mathbb{R}^{k\times k}} \|\Theta X - G_k(T)\|_F^2$. Denoting*

$$S_i = \left\{ j \in C_i \ : \ \left\|(\widetilde{\Theta}\widetilde{X})_{j*} - (\Theta(C^-)^{-1/2}RO)_{j*}\right\| \geq \frac{1}{2\sqrt{n_i(\tau^+ + \frac{2}{1-l})}} \right\},$$

it holds with probability at least $1-2\varepsilon$ that

$$\sum_{i=1}^{k} \frac{|S_i|}{n_i} \leq \delta^2(64+32\xi)k \left(\tau^+ + \frac{2}{1-l}\right) \left(\frac{(\tau^+)^3+1}{(\tau^+)^4}\right).$$

In particular, if δ satisfies

$$\delta < \frac{(\tau^+)^2}{\sqrt{(64+32\xi)k(\tau^+ + \frac{2}{1-l})((\tau^+)^3+1)}},$$

then there exists a $k\times k$ permutation matrix π such that $\widetilde{\Theta}_G = \hat{\Theta}_G\pi$, where $G = \cup_{i=1}^k (C_i \backslash S_i)$.

In the sparse regime, the above statement holds under the notation and assumptions of Theorem 34 with $G_k(T)$ replaced with $G_k(T_{\gamma^+,\gamma^-})$, and with probability at least $1-2e^{-r}$.

Proof Since $G_k(T) - G_k(\overline{T})O$ has rank at most $2k$, we obtain from Theorem 29 that

$$\left\|G_k(T) - G_k(\overline{T})O\right\|_F \leq \sqrt{2k}\left\|G_k(T) - G_k(\overline{T})O\right\| \leq \delta\sqrt{2k}\left(\frac{(\tau^+)^{3/2}+1}{(\tau^+)^2}\right). \tag{62}$$

We now use Lemma 35 with $U = G_k(T)$ and $\overline{U} = G_k(\overline{T})O$. It follows from (44) and Lemma 17 that $G_k(\overline{T}) = \Theta(C^-)^{-1/2}R = \Theta\Delta\,\Delta^{-1}(C^-)^{-1/2}R$ where $\Delta = \mathrm{diag}(\sqrt{n_1},\ldots,\sqrt{n_k})$. Denoting $\overline{X} = \Delta^{-1}(C^-)^{-1/2}RO$, we can write $G_k(\overline{T})O = \hat{\Theta}\overline{X}$, where $\hat{\Theta} \in \mathbb{M}_{n\times k}$ is the ground truth membership matrix, and for each $i \neq i' \in [k]$, it holds true that

$$\left\|\overline{X}_{i*} - \overline{X}_{i'*}\right\| \geq \lambda_{\min}((C^-)^{-1/2})\sqrt{1/n_i + 1/n_{i'}} \geq \frac{1}{\sqrt{\lambda_{\max}(C^-)n_i}}\,.$$

From (18), one can verify using Weyl's inequality that

$$\lambda_{\max}(C^-) \leq 1 + \tau^+ + \max_i \frac{p}{d_i^-}(\eta_i + s_i n(1-2\eta)) \leq \tau^+ + \frac{2}{1-l},$$

where the last inequality holds if $n \geq \frac{2\eta}{(1-l)(1-\eta)}$. The above considerations imply that $\delta_i = \frac{1}{\sqrt{n_i(\tau^+ + \frac{2}{1-l})}}$. Now with S_i as defined in the statement, we obtain from (60) and (62) that

$$\sum_{i=1}^k |S_i|\,\delta_i^2 = \frac{1}{\tau^+ + \frac{2}{1-l}}\sum_{i=1}^k \frac{|S_i|}{n_i} \leq \delta^2(32+16\xi)k\frac{((\tau^+)^{3/2}+1)^2}{(\tau^+)^4} \leq \delta^2(64+32\xi)k\left(\frac{(\tau^+)^3+1}{(\tau^+)^4}\right),$$

where the last inequality uses $(a+b)^2 \leq 2(a^2+b^2)$ for $a, b \geq 0$. This yields the first part of the Theorem.

For the second part, we need to ensure (61) holds. Using (62) and the expression for δ_i, it is easy to verify that (61) holds for the stated condition on δ.

Finally, the statement for the sparse regime readily follows in an analogous manner (replacing $G_k(T)$ with $G_k(T_{\gamma^+,\gamma^-})$), by following the same steps as above. ■

5. Concentration results for the symmetric Signed Laplacian

This section contains proofs of the main results for the symmetric Signed Laplacian, in both the dense regime $p \gtrsim \frac{\ln n}{n}$ and the sparse regime $p \gtrsim \frac{1}{n}$. Before proceeding with an overview of the main steps, for ease of reference, we summarize in the Table below the notation specific to this section.

Notation	Description
$\overline{L_{sym}}$	symmetric Signed Laplacian
$\mathcal{L}_{sym}$	population Signed Laplacian
L_γ	regularized Laplacian
$\mathcal{L}_\gamma$	population regularized Laplacian
$\gamma^+, \gamma^- > 0$	regularization parameters
$\gamma = \gamma^+ + \gamma^-$	
$\overline{\alpha} = 1 + \frac{p}{d}(1-2\eta)$	
$\overline{d} = p(n-1)$	expected signed degree
$\rho = \frac{n_{min}}{n_{max}} = \frac{s}{l}$	aspect ratio

The proof of Theorem 8 is built on the following steps. In Section 5.1, we compute the eigen-decomposition of the Signed Laplacian of the expected graph $\mathcal{L}_{sym}$. Then in Section 5.2, we show $\overline{L_{sym}}$ and $\mathcal{L}_{sym}$ are "close", and obtain an upper bound on the error $\left\|\overline{L_{sym}} - \mathcal{L}_{sym}\right\|$. Finally, in Section 5.3, we use the Davis-Kahan theorem (see Theorem 56) to bound the error between the subspaces $V_{k-1}(\overline{L_{sym}})$ and $V_{k-1}(\mathcal{L}_{sym})$. To prove Theorem 11, in Section 5.4, we first use a decomposition of the set of edges $[n] \times [n]$ and characterize the behaviour of the regularized Signed Laplacian on each subset. This leads in Section 5.5 to the error bounds of Theorem 11. Finally, the proof of Theorem 13, that bound the error on the eigenspace, relies on the same arguments as Theorem 8 and can be found in Section 5.6. Similarly to the approach for SPONGE$_{sym}$, the mis-clustering error is obtained using a $(1+\xi)$-approximate solution of the k-means problem applied to the rows of $V_{k-1}(\overline{L_{sym}})$ (resp. $V_{k-1}(L_\gamma)$). This solution contains, in particular, an estimated membership matrix $\widetilde{\Theta}$. The bound on the mis-clustering error of the algorithm given in Theorem 16 is derived using Lemma 35 (Lemma 5.3 of Lei and Rinaldo (2015)), in Section 5.7.

5.1 Analysis of the expected Signed Laplacian

In this section, we compute the eigen-decomposition of the matrix $\mathcal{L}_{sym}$. In particular, we aim at proving a lower bound on the eigengap between the $(k-1)^{th}$ and k^{th} smallest eigenvalues. For equal-size clusters, there is an explicit expression for this eigengap.

5.1.1 Matrix decomposition

Lemma 37 *Let $\Theta \in \mathbb{R}^{n\times k}$ denote the normalized membership matrix in the SSBM. Let $V^\perp \in \mathbb{R}^{n\times(n-k)}$ be a matrix whose columns are any orthonormal base of the subspace orthogonal to $\mathcal{R}(\Theta)$. The Signed Laplacian of the expected graph has the following decomposition*

$$\mathcal{L}_{sym} = [\Theta\, V^\perp] \begin{pmatrix} \overline{C} & 0 \\ 0 & \overline{\alpha} I_{n-k} \end{pmatrix} \begin{bmatrix} \Theta^T \\ (V^\perp)^T \end{bmatrix}, \tag{63}$$

with $\overline{C} = \overline{\alpha} I_k - \overline{B}$, $\overline{\alpha} = 1 + \frac{p}{d}(1-2\eta)$ and $\overline{B}$ is a $k \times k$ matrix such that

$$\overline{B}_{ii'} = \begin{cases} \frac{n_i p}{\overline{d}}(1-2\eta); & \text{if } i = i' \\ -\frac{\sqrt{n_i n_{i'}} p}{\overline{d}}(1-2\eta); & \text{if } i \neq i'. \end{cases} \tag{64}$$

Proof On one hand, we recall from Section 2.3 that the expected degree matrix is a scaled identity matrix $\mathbb{E}[\overline{D}] = \overline{d} I_n$, with $\overline{d} = p(n-1)$. Thus, any vector $v \in \mathbb{R}^n$ is an eigenvector

of $\mathbb{E}[\overline{D}]$ with corresponding eigenvalue $\overline{d}$, and it holds true that

$$\mathbb{E}[\overline{D}]^{-1/2} = \frac{1}{\sqrt{\overline{d}}} I_n = \frac{1}{\sqrt{\overline{d}}} [\Theta \, (V^{\perp})] \, I_n \begin{bmatrix} \Theta^T \\ (V^{\perp})^T \end{bmatrix}. \tag{65}$$

On the other hand, the signed adjacency matrix can be written in the form

$$\mathbb{E}[A] = \mathbb{E}[A^+] - \mathbb{E}[A^-] = M - p(1-2\eta) I_n, \tag{66}$$

where

$$M = \begin{bmatrix} p(1-2\eta) J_{n_1} & -p(1-2\eta) J_{n_1 \times n_2} & \dots & -p(1-2\eta) J_{n_1 \times n_k} \\ -p(1-2\eta) J_{n_2 \times n_1} & p(1-2\eta) J_{n_2} & \dots & -p(1-2\eta) J_{n_2 \times n_k} \\ \vdots & \vdots & \ddots & \vdots \\ -p(1-2\eta) J_{n_k \times n_1} & \dots & \dots & p(1-2\eta) J_{n_k} \end{bmatrix}.$$

The matrix M has the following decomposition

$$M = \overline{d} \Theta \overline{B} \Theta^T = \overline{d} [\Theta \, V^{\perp}] \begin{pmatrix} \overline{B} & 0 \\ 0 & 0 \end{pmatrix} \begin{bmatrix} \Theta^T \\ (V^{\perp})^T \end{bmatrix},$$

with $\overline{B}$ defined in (64). Thus, combining (65) and (66), we arrive at

$$\mathbb{E}[\overline{D}]^{-1/2} \mathbb{E}[A] \mathbb{E}[\overline{D}]^{-1/2} = \frac{1}{\overline{d}} M - p(1-2\eta) \frac{1}{\overline{d}} I_n = [\Theta \, V^{\perp}] \begin{pmatrix} \overline{B} & 0 \\ 0 & 0 \end{pmatrix} \begin{bmatrix} \Theta^T \\ (V^{\perp})^T \end{bmatrix} - (1-2\eta) \frac{p}{\overline{d}} I_n.$$

This finally leads to the decomposition of $\mathcal{L}_{sym}$

$$\mathcal{L}_{sym} = I - \mathbb{E}[\overline{D}]^{-1/2} \mathbb{E}[A] \mathbb{E}[\overline{D}]^{-1/2} = [\Theta \, V^{\perp}] \begin{pmatrix} \overline{C} & 0 \\ 0 & \overline{\alpha} I_{n-k} \end{pmatrix} \begin{bmatrix} \Theta^T \\ (V^{\perp})^T \end{bmatrix},$$

with $\overline{C} = \overline{\alpha} I_k - \overline{B}$ and $\overline{\alpha} = 1 + p(1-2\eta)$. ■

We can infer from Lemma 37 that the spectrum of $\mathcal{L}_{sym}$ is the union of the spectrum of the matrix $\overline{C} \in \mathbb{R}^{k \times k}$ and $\{\overline{\alpha}\}$. Moreover, denoting $u = \frac{1}{\sqrt{\overline{d}}} (\sqrt{n_1}, \dots, \sqrt{n_k})^T$, we have $\overline{C} = p(1-2\eta) u u^T + \operatorname{diag}\left(1 + \frac{p}{\overline{d}}(1-2\eta)(1-2n_i)\right)$. For a SSBM with equal-size clusters, we are able to find explicit expressions for the eigenvalues of $\overline{C}$.

5.1.2 Spectrum of the Signed Laplacian: equal-size clusters

In this section, we assume that the clusters in the SSBM have equal sizes $n_1 = n_2 = \dots = n_k = \frac{n}{k}$. In this case,

$$\frac{1}{\sqrt{\overline{d}}} (\sqrt{n_1}, \dots, \sqrt{n_k})^T = \sqrt{\frac{n}{\overline{d}}} \chi_1,$$

and denoting by $\overline{C}_e$ the matrix $\overline{C}$ in this setting of equal clusters, we may write

$$\overline{C}_e = \frac{np}{\overline{d}} (1-2\eta) \chi_1 \chi_1^T + \left(1 + \frac{p}{\overline{d}} (1-2\eta) \left(1 - 2\frac{n}{k}\right)\right) I_k. \tag{67}$$

Hence, the spectrum of $\overline{C}_e$ contains only two different values. The largest one has multiplicity 1, and χ_1 is the corresponding largest eigenvector. The $k-1$ remaining eigenvalues are all equal. In fact, we have

$$\lambda_i(\overline{C}_e) = \begin{cases} 1 + \frac{p}{\overline{d}}(1-2\eta)(n+1-2\frac{n}{k}); & \text{if } i = 1 \\ 1 + \frac{p}{\overline{d}}(1-2\eta)\left(1-2\frac{n}{k}\right); & \text{if } 2 \le i \le k. \end{cases}$$

One can easily check that these eigenvalues are positive, and that the following inequality holds true

$$\lambda_1(\overline{C}_e) = \overline{\alpha} + \frac{p}{\overline{d}}(1-2\eta)(n-2\frac{n}{k}) \ge \overline{\alpha} > \overline{\alpha} - 2\frac{n}{k}(1-2\eta) = \lambda_2(\overline{C}_e).$$

We finally have

$$\lambda_j(\mathcal{L}_{sym}) = \begin{cases} 1 + \frac{p}{\overline{d}}(1-2\eta)(n+1-2\frac{n}{k}); & \text{if } j = 1 \\ \overline{\alpha}; & \text{if } 2 \le j \le n-k+1 \\ \lambda_2(\overline{C}_e); & \text{if } n-k+2 \le j \le n. \end{cases}$$

Note that for $k = 2$, $\lambda_1(\overline{C}_e) = \overline{\alpha}$ and the spectrum of $\mathcal{L}_{sym}$ contains only two values $\{\overline{\alpha}, \lambda_2(\overline{C}_e)\}$. For $k > 2$, $\lambda_1(\mathcal{L}_{sym}) > \overline{\alpha} > \lambda_2(\overline{C}_e)$. Writing the spectral decomposition

$$\overline{C}_e = R\,\Lambda\,R^T = [R_{k-1}\,\gamma_1]\,\Lambda\,\begin{bmatrix} R_{k-1}^T \\ \gamma_1^T \end{bmatrix},$$

with $\gamma_1 = \chi_1$ and $R_{k-1} \in \mathbb{R}^{k\times(k-1)}$ being the matrix of eigenvectors associated to $\lambda_2(\overline{C}_e)$, we conclude that $V_{k-1}(\mathcal{L}_{sym}) = \Theta R_{k-1}$. In fact, since Θ has k distinct rows and R is a unitary matrix, ΘR also has k distinct rows. As χ_1 is the all one's vector , ΘR_{k-1} has k distinct rows as well. These observations are summarized in the following lemma and lead to the expression of the eigengap.

Lemma 38 (Eigengap for equal-size clusters) *For the SSBM with $k \ge 2$ clusters of equal-size $\frac{n}{k}$, we have that $V_{k-1}(\mathcal{L}_{sym}) = \Theta R_{k-1} \in \mathbb{R}^{n\times(k-1)}$, where R_{k-1} corresponds to the $(k-1)$ smallest eigenvectors of $\overline{C}_e$. Moreover, with the eigengap defined as*

$$\lambda_{gap} := \lambda_{n-k+1}(\mathcal{L}_{sym}) - \lambda_{n-k+2}(\mathcal{L}_{sym}),$$

it holds true that

$$\lambda_{gap} = \overline{\alpha} - \lambda_2(\overline{C}_e) = \frac{2np}{k\overline{d}}(1-2\eta) \ge \frac{2}{k}(1-2\eta). \tag{68}$$

5.1.3 Non-equal-size clusters

In the general setting of non-equal-size clusters, it is difficult to obtain an explicit expression of the spectrum of $\mathcal{L}_{sym}$. Thus, using a perturbation method, we establish a lower bound on the eigengap, provided that the aspect ratio ρ is close to 1. Recall that

$$\begin{aligned} \overline{C} &= p(1-2\eta)uu^T + \operatorname{diag}\left(1 + \frac{p}{\overline{d}}(1-2\eta)(1-2n_i)\right) \\ &= p(1-2\eta)uu^T - 2p(1-2\eta)\operatorname{diag}(u_i^2)_{i=1}^n + \operatorname{diag}\left(1 + \frac{p}{\overline{d}}(1-2\eta)\right). \end{aligned} \tag{69}$$

We note that this matrix is of the form $\Lambda + vv^T$, with Λ being a diagonal matrix and $v \in \mathbb{R}^k$ a vector. Using again the spectral decomposition

$$\overline{C} = R\,\Lambda\,R^T = [R_{k-1}\,\gamma_1]\,\Lambda\,\begin{bmatrix} R_{k-1}^T \\ \gamma_1^T \end{bmatrix}, \tag{70}$$

where γ_1 is the largest eigenvector and $R_{k-1} \in \mathbb{R}^{k\times(k-1)}$ contains the smallest $(k-1)$ eigenvectors of $\overline{C}$, we would like to ensure that the smallest $(k-1)$ eigenvectors of $\mathcal{L}_{sym}$ are related to the $(k-1)$ eigenvectors of $\overline{C}$ in the following way $V_{k-1}(\mathcal{L}_{sym}) = \Theta R_{k-1}$. Note that γ_1 is not necessarily the all one's vector, and ΘR_{k-1} has at least $k-1$ distinct rows. To this end, we will like to ensure that

$$\{\lambda_2(\overline{C}), \ldots, \lambda_{k-1}(\overline{C}), \lambda_k(\overline{C})\} = \{\lambda_{n-k+2}(\mathcal{L}_{sym}), \ldots, \lambda_{n-1}(\mathcal{L}_{sym}), \lambda_n(\mathcal{L}_{sym})\}. \tag{71}$$

From Weyl's inequality (see Theorem 55), we know that

$$|\lambda_i(\overline{C}_e) - \lambda_i(\overline{C})| \le \|\overline{C} - \overline{C}_e\| \quad \forall i = 1, \ldots k,$$

which in particular implies

$$\lambda_2(\overline{C}) \le \lambda_2(\overline{C}_e) + \|\overline{C} - \overline{C}_e\|, \qquad \lambda_1(\overline{C}) \ge \lambda_1(\overline{C}_e) - \|\overline{C} - \overline{C}_e\|.$$

Moreover, $\lambda_1(\overline{C}) = \overline{\alpha}$ when $k = 2$, and $\lambda_1(\overline{C}) > \overline{\alpha}$ when $k > 2$. Thus, for Condition 71 to be true, it suffices to ensure

$$\begin{aligned} \lambda_2(\overline{C}_e) + \|\overline{C} - \overline{C}_e\| < \overline{\alpha} + \|\overline{C} - \overline{C}_e\| &\iff \|\overline{C} - \overline{C}_e\| < \frac{\overline{\alpha} - \lambda_2(\overline{C}_e)}{2} \\ &\iff \|\overline{C} - \overline{C}_e\| < \frac{np}{k\overline{d}}(1-2\eta), \end{aligned}$$

using (68). In this case, we indeed have that $V_{k-1}(\mathcal{L}_{sym}) = \Theta R_{k-1}$. As it will be convenient later, we will ensure a slightly stronger condition, i.e.

$$\|\overline{C} - \overline{C}_e\| < \frac{\overline{\alpha} - \lambda_2(\overline{C}_e)}{4} = \frac{np}{2k\overline{d}}(1-2\eta). \tag{72}$$

Now we compute the error $\|\overline{C} - \overline{C}_e\|$. We recall that $\|u\| = \sqrt{\frac{n}{\overline{d}}}$ and denote $D_u =: \frac{1}{\|u\|^2}\mathrm{diag}(u_i^2)_{i=1}^n$, then (69) becomes

$$\overline{C} = \overline{\alpha} I_k + \frac{np}{\overline{d}}(1-2\eta)\left(\frac{u}{\|u\|}\right)\left(\frac{u}{\|u\|}\right)^T - 2\frac{np}{\overline{d}}(1-2\eta)D_u.$$

Using (67), we obtain

$$\begin{aligned} \|\overline{C} - \overline{C}_e\| &= \left\| \frac{np}{\overline{d}}(1-2\eta)\left(\left(\frac{u}{\|u\|}\right)\left(\frac{u}{\|u\|}\right)^T - \chi_1\chi_1^T\right) - 2\frac{np}{\overline{d}}(1-2\eta)\left(D_u - \frac{1}{k}I_n\right)\right\| \\ &\le \frac{np}{\overline{d}}(1-2\eta)\left\|\left(\frac{u}{\|u\|}\right)\left(\frac{u}{\|u\|}\right)^T - \chi_1\chi_1^T\right\| + 2\frac{np}{\overline{d}}(1-2\eta)\left\|D_u - \frac{1}{k}I_n\right\|. \end{aligned} \tag{73}$$

For the first term on the RHS, we have

$$\left\| \left(\frac{u}{\|u\|}\right)\left(\frac{u}{\|u\|}\right)^T - \chi_1\chi_1^T \right\| \le 2\left\| \frac{u}{\|u\|} - \chi_1 \right\| \le 2\sqrt{k}\max_i \left| \sqrt{\frac{n_i}{n}} - \sqrt{\frac{1}{k}} \right| \le 2\sqrt{k}(\sqrt{l} - \sqrt{s}) \le 2\sqrt{k}(1-\sqrt{\rho}), \tag{74}$$

while for the second term on the RHS, we have

$$\left\| D_u - \frac{1}{k} I_n \right\| = \max_i \left| \sqrt{\frac{n_i}{n}} - \sqrt{\frac{1}{k}} \right| \le 1 - \sqrt{\rho}. \tag{75}$$

By combining (74) and (75) into (73), we arrive at

$$\begin{aligned} \|\overline{C} - \overline{C}_e\| &\le \frac{np}{\overline{d}}(1-2\eta)\sqrt{k}(1-\sqrt{\rho}) + \frac{2np}{\overline{d}}(1-2\eta)(1-\sqrt{\rho}) \\ &\le \frac{np}{\overline{d}}(1-2\eta)(1-\sqrt{\rho})\left(\sqrt{k}+2\right) \\ &\le 2(2+\sqrt{k})(1-2\eta)(1-\sqrt{\rho}), \end{aligned}$$

using that $\frac{np}{\overline{d}} = \frac{n}{n-1} \le 2$. Now since $\frac{np}{2k\overline{d}} \ge \frac{1-2\eta}{2k}$ and from Condition 72, it suffices that ρ satisfies

$$2(2+\sqrt{k})(1-2\eta)(1-\sqrt{\rho}) \le \frac{1-2\eta}{2k} \iff 1-\sqrt{\rho} \le \frac{1}{4k(2+\sqrt{k})}.$$

Finally, we can compute

$$\begin{aligned} \lambda_{gap} &:= \lambda_{n-k+1}(\mathcal{L}_{sym}) - \lambda_{n-k+2}(\mathcal{L}_{sym}) \\ &\ge \overline{\alpha} - \|\overline{C} - \overline{C}_e\| - (\lambda_2(\overline{C}_e) + \|\overline{C} - \overline{C}_e\|) \\ &\ge \overline{\alpha} - \lambda_2(\overline{C}_e) - 2\|\overline{C} - \overline{C}_e\| \\ &\ge \frac{\overline{\alpha} - \lambda_2(\overline{C}_e)}{2} = \frac{np}{k\overline{d}}(1-2\eta) \ge \frac{1-2\eta}{k}. \end{aligned}$$

Hence we arrive at the following lemma.

Lemma 39 (General lower-bound on the eigengap) *For a SSBM with $k \ge 2$ clusters of general sizes $(n_1, \ldots, n_k)$ and aspect ratio ρ satisfying*

$$\sqrt{\rho} > 1 - \frac{1}{4k(2+\sqrt{k})},$$

it holds true that $V_{k-1}(\mathcal{L}_{sym}) = \Theta R_{k-1}$, where $R_{k-1} \in \mathbb{R}^{k\times k-1}$ corresponds to the $(k-1)$ smallest eigenvectors of $\overline{C}$. Furthermore, we can lower-bound the spectral gap λ_{gap} as

$$\lambda_{gap} := \lambda_{n-k+1}(\mathcal{L}_{sym}) - \lambda_{n-k+2}(\mathcal{L}_{sym}) \ge \frac{1-2\eta}{k}.$$

We will now show that $\overline{L_{sym}}$ concentrates around the population Laplacian $\mathcal{L}_{sym}$, provided the graph is dense enough.

5.2 Concentration of the Signed Laplacian in the dense regime

In the moderately dense regime where $p \gtrsim \frac{\ln n}{n}$, the adjacency and the degree matrices concentrate towards their expected counterparts, as n increases. This can be established using standard concentration tools from the literature.

Lemma 40 *We have the following concentration inequalities for A and $\overline{D}$*

1. $\forall 0 < \varepsilon \leq \frac{1}{2}, \exists c_\varepsilon > 0$,

$$\mathbb{P}\bigg(\|A - \mathbb{E}[A]\| \leq ((1+\varepsilon)4\sqrt{2}+2)\sqrt{np} \bigg) \geq 1 - n\exp\bigg(-\frac{np}{c_\varepsilon} \bigg).$$

In particular, there exists a universal constant $c > 0$ such that

$$\mathbb{P}\bigg(\|A - \mathbb{E}[A]\| \leq 12\sqrt{np} \bigg) \geq 1 - n\exp\bigg(-\frac{np}{c} \bigg).$$

2. *If $p > 12\frac{\ln n}{n}$,*

$$\mathbb{P}\bigg(\|\overline{D} - \mathbb{E}[\overline{D}]\| \leq \sqrt{3np\ln n} \bigg) \geq 1 - \frac{2}{n}.$$

Proof For the first statement, we recall that A is a symmetric matrix, with $A_{jj'} = 0$ and with independent entries above the diagonal $(A_{jj'})_{j<j'}$. We denote $Z_{jj'} = A_{jj'} - \mathbb{E}[A_{jj'}]$. If j, j' lie in the same cluster,

$$Z_{jj'} = \begin{cases} 1 - p(1-2\eta) & ; \text{ w. p. } p(1-\eta) \\ -1 - p(1-2\eta) & ; \text{ w. p. } p\eta \\ -p(1-2\eta) & ; \text{ w. p. } 1-p \end{cases}.$$

If j, j' lie in different clusters,

$$Z_{jj'} = \begin{cases} 1 + p(1-2\eta) & ; \text{ w. p. } p\eta \\ -1 + p(1-2\eta) & ; \text{ w. p. } p(1-\eta) \\ p(1-2\eta) & ; \text{ w. p. } 1-p \end{cases}.$$

One can easily check that in both cases, it holds true that

$$\begin{aligned} \mathbb{E}[(Z_{jj'})^2] &= p\big[(1-\eta)(1-p(1-2\eta))^2 + \eta(1+p(1-2\eta))^2 + p(1-2\eta)^2)(1-p)\big] \\ &\leq p(1+\eta(1+p)^2+p) \leq 4p. \end{aligned}$$

Thus we can conclude that for each $j \in [n]$, the following holds

$$\sqrt{\sum_{j'=1}^{n} \mathbb{E}[(Z_{jj'})^2]} \leq \sqrt{4np} = 2\sqrt{np}.$$

Hence, $\widetilde{\sigma} := \max_j \sqrt{\sum_{j'=1}^n \mathbb{E}[(Z_{jj'})^2\,]} \leq 2\sqrt{np}$. Moreover, $\widetilde{\sigma}_* := \max_{j,j'} \left\| Z^+_{jj'} \right\|_\infty = 1 + p(1-2\eta) \leq 2$. Therefore, we can apply the concentration bound for the norm of symmetric

matrices by Bandeira and van Handel (2016, Corollary 3.12, Remark 3.13) (recalled in Appendix 53) with $t = 2\sqrt{np}$, in order to bound $\|Z\| = \|A - \mathbb{E}[A]\|$. For any given $0 < \varepsilon \leq 1/2$, we have that

$$\|A - \mathbb{E}[A]\| \leq \big((1+\varepsilon)4\sqrt{2} + 2\big)\sqrt{np},$$

with probability at least $1 - n\exp\left(\frac{-pn}{c_\varepsilon}\right)$, where c_ε only depends on ε.

For the second statement, we apply Chernoff's bound (see Appendix A.1) to the random variables $\overline{D}_{jj} = \sum_{j'=1}^{n}\left(A^+_{jj'} + A^-_{jj'}\right)$, where we note that $(A^+_{jj'} + A^-_{jj'})_{j'=1}^n$ are independent Bernoulli random variables with mean p. Hence, $\mathbb{E}[D_{jj}] = \overline{d} = p(n-1)$. Let $\delta = \sqrt{\frac{6\ln n}{\overline{d}}}$ and assuming that $p > 12\frac{\ln n}{n}$ (so that $\delta < 1$), we obtain

$$\mathbb{P}\left[\left|\overline{D}_{jj} - \overline{d}\right| \geq \sqrt{6\overline{d}\ln n}\right] \leq \mathbb{P}\left[\left|\overline{D}_{jj} - \overline{d}\right| \geq \sqrt{3np\ln n}\right] \leq 2\exp\big(-2\ln n\big) = \frac{2}{n^2},$$

using that $n - 1 \geq \frac{n}{2}$. Applying the union bound, we finally obtain that

$$\mathbb{P}\left(\|\overline{D} - \mathbb{E}[\overline{D}]\| \geq \sqrt{3np\ln n}\right) \leq \frac{2}{n}.$$

■

Lemma 41 *If $\|A - \mathbb{E}[A]\| \leq \Delta_A$, $\|\overline{D} - \mathbb{E}[\overline{D}]\| \leq \Delta_D$ and $p > 12\frac{\ln n}{n}$, then with probability at least $1 - \frac{2}{n}$, it follows that*

$$\|\overline{L_{sym}} - \mathcal{L}_{sym}\| \leq \frac{\Delta_A}{\overline{d}} + 2\frac{\Delta_D}{\overline{d}} + \frac{\Delta_D^2}{\overline{d}^2}.$$

Proof We first note that using the proof of Lemma 40, with probability at least $1 - \frac{2}{n}$, we have that $\left|\overline{D}_{jj} - \overline{d}\right| \leq \delta\overline{d}, \forall j \in [n]$, with $\delta < 1$. Consequently,

$$\|(\mathbb{E}[\overline{D}])^{-1/2}\overline{D}^{1/2} - I\| = \max_j \left|\sqrt{\frac{\overline{D}_{jj}}{\overline{d}}} - 1\right| \leq \max_j \frac{|\overline{D}_{jj} - \overline{d}|}{\overline{d}} = \frac{\Delta_D}{\overline{d}},$$

since $|\sqrt{x} - 1| \leq |x - 1|$ for $0 < x < 1$. We now apply the first inequality of Proposition 59 with $A^- = \overline{D}, A^+ = A, B^- = \mathbb{E}\left[\overline{D}\right], B^+ = \mathbb{E}\left[A\right]$. We obtain

$$\|\overline{L_{sym}} - \mathcal{L}_{sym}\| \leq \frac{\Delta_A}{\overline{d}} + \left\|\overline{D}^{-1}\right\|\|A\|\left(\frac{\Delta_D^2}{\overline{d}^2} + 2\frac{\Delta_D}{\overline{d}}\right).$$

It remains to prove that $\left\|\overline{D}^{-1}\right\|\|A\| \leq 1$. It holds since $\overline{D}$ is a diagonal matrix, thus $\left\|\overline{D}^{-1}\right\|\|A\| = \left\|\overline{D}^{-1}A\right\|$ and similarly to Lemma 61, it is straightforward to prove that $I - \left\|\overline{D}^{-1}A\right\| \leq 2$, therefore $\left\|\overline{D}^{-1}A\right\| \leq 1$.

■

Combining the results from Lemma 40 and Lemma 41, we arrive at the concentration bound for $\|\overline{L_{sym}} - \mathcal{L}_{sym}\|$.

Lemma 42 *Under the assumptions of Theorem 8, if $n \geq 10$, then with probability at least $1 - n\exp(-\frac{np}{c_\epsilon}) - \frac{2}{n}$ there exists a universal constant $0 < C < 43$ such that*

$$\|\overline{L_{sym}} - \mathcal{L}_{sym}\| \leq C\sqrt{\frac{\ln n}{np}}.$$

Proof If $p \geq \frac{12\ln n}{n}$, the bounds in Lemma 40 hold simultaneously with probability at least $1 - n\exp(-\frac{np}{c}) - \frac{2}{n}$ and we have, with the notations of Lemma 41, $\Delta_A \leq 12\sqrt{np}$ and $\Delta_D \leq \sqrt{3np\ln n}$. Applying Lemma 41, we then obtain

$$\|\overline{L_{sym}} - \mathcal{L}_{sym}\| \leq \frac{12\sqrt{np}}{\overline{d}} + 2\frac{\sqrt{3np\ln n}}{\overline{d}} + \frac{3np\ln n}{\overline{d}^2} \leq \frac{24}{\sqrt{np}} + 4\sqrt{3}\sqrt{\frac{\ln n}{np}} + \frac{12\ln n}{np}.$$

If $n \geq 10$, $\ln n \geq 1$ and $\sqrt{\frac{\ln n}{np}} \geq \frac{1}{\sqrt{np}}$. Moreover, since $p \geq 12\frac{\ln n}{n}$, then $\frac{\ln n}{np} \leq \frac{1}{12} < 1$ and $\sqrt{\frac{\ln n}{np}} \geq \frac{\ln n}{np}$. We finally obtain

$$\|\overline{L_{sym}} - \mathcal{L}_{sym}\| \leq (24 + 4\sqrt{3} + 12)\sqrt{\frac{\ln n}{np}} = C\sqrt{\frac{\ln n}{np}},$$

with $C = 24 + 4\sqrt{3} + 12 \leq 43$. ■

5.3 Proof of Theorem 8

The proof of this theorem relies on the Davis-Kahan theorem. Using Weyl's inequality (see Theorem 55) and Lemma 42, we obtain for all $1 \leq j \leq n$,

$$|\lambda_j(\overline{L_{sym}}) - \lambda_j(\mathcal{L}_{sym})| \leq C\left(\frac{\ln n}{np}\right)^{1/2}.$$

In particular, for the k-th smallest eigenvalue,

$$\begin{aligned}
\lambda_{n-k+1}(\overline{L_{sym}}) &\geq \lambda_{n-k+1}(\mathcal{L}_{sym}) - C\left(\frac{\ln n}{np}\right)^{1/2}, \\
\lambda_{n-k+1}(\overline{L_{sym}}) - \lambda_{n-k+2}(\mathcal{L}_{sym}) &\geq \lambda_{n-k+1}(\mathcal{L}_{sym}) - \lambda_{n-k+2}(\mathcal{L}_{sym}) - C\left(\frac{\ln n}{np}\right)^{1/2} \\
&= \lambda_{gap} - C\left(\frac{\ln n}{np}\right)^{1/2}.
\end{aligned}$$

For $\delta \in (0,1)$, we will like to ensure that

$$\lambda_{gap} - C\left(\frac{\ln n}{np}\right)^{1/2} > \lambda_{gap}\left(1 - \frac{\delta}{2}\right). \tag{76}$$

From Lemma 39, if $\sqrt{\rho} > 1 - \frac{1}{4k(2+\sqrt{k})}$, then $\lambda_{gap} \geq \frac{1}{k}(1-2\eta)$. Then for the previous condition (76) to hold, it is sufficient that

$$C\left(\frac{\ln n}{np}\right)^{1/2} < \frac{\delta}{2k}(1-2\eta) \iff p > \left(\frac{2Ck}{\delta(1-2\eta)}\right)^2 \frac{\ln n}{n} = C(k,\eta,\delta)\frac{\ln n}{n}, \tag{77}$$

with $C(k,\eta,\delta) = \left(\frac{2Ck}{\delta(1-2\eta)}\right)^2$. We note that since $C(k,\eta,\delta) \geq C \geq 12$, hence (77) implies that $p > 12\frac{\ln n}{n}$.

With this condition, we now apply the Davis-Kahan theorem (Theorem 56)

$$\begin{aligned}\|(I - V_{k-1}(\overline{L_{sym}})V_{k-1}(\overline{L_{sym}})^T)V_{k-1}(\mathcal{L}_{sym})\| &\leq \frac{\left\|\overline{L_{sym}} - \mathcal{L}_{sym}\right\|}{\lambda_{gap} - C\left(\frac{\ln n}{np}\right)^{1/2}} \\ &\leq \frac{\delta\lambda_{gap}/2}{\lambda_{gap}(1-\delta/2)} = \frac{\delta/2}{1-\delta/2} \leq \delta.\end{aligned}$$

Using Proposition 57, there then exists an orthogonal matrix $O \in \mathbb{R}^{(k-1)\times(k-1)}$ so that

$$\|V_{k-1}(\overline{L_{sym}}) - \Theta R_{k-1} O\| \leq 2\delta.$$

5.4 Properties of the regularized Laplacian in the sparse regime

The analysis of the signed regularized Laplacian differs from the one of unsigned regularized Laplacian. In particular, Lemma 30 cannot be directly applied, since the trimming approach of the adjacency matrix for unsigned graphs is not available in this case. However, we will also use arguments of Le et al. (2015) and Le et al. (2017) for unsigned directed adjacency matrices in the inhomogeneous Erdős-Rényi model $G(n,(p_{jj'})_{j,j'})$. More precisely, in Section 5.4.1, we will prove that the adjacency matrix concentrates on a large subset of edges called the *core*. On this subset, the unregularized (resp. regularized) Laplacian also concentrates towards the expected matrix $\mathcal{L}_{sym}$ (resp. $\mathcal{L}_\gamma$). In Section 5.4.2, we will show that on the remaining subset of nodes, the norm of the regularized Laplacian is relatively small.

5.4.1 Properties of the signed adjacency and degree matrices

In this section, we adapt the results by Le et al. (2017) for the signed adjacency matrix and the degree matrix in our SSBM. Similarly to Le et al. (2017, Theorem 2.6) (see Theorem 54), the following lemma shows that the set of edges can be decomposed into a large block, and two blocks with respectively few columns and few rows.

Lemma 43 *(Decomposition of the set of edges for the SSBM) Let A be the signed adjacency matrix of a graph sampled from the SSBM. For any $r \geq 1$, with probability at least $1-6n^{-r}$, the set of edges $[n]\times[n]$ can be partitioned into three classes $\mathcal{N},\mathcal{R}$ and $\mathcal{C}$ such that*

1. *the signed adjacency matrix concentrates on $\mathcal{N}$*

$$\|(A - \mathbb{E}A)_{\mathcal{N}}\| \leq Cr^{3/2}\sqrt{\overline{d}(1-\eta)},$$

with $C > 1$ a constant;

2. *$\mathcal{R}$ (resp. $\mathcal{C}$) intersects at most $4n/\overline{d}$ columns (resp. rows) of $[n] \times [n]$;*

3. *each row (resp. column) of $A_{\mathcal{R}}$ (resp. $A_{\mathcal{C}}$) has at most $128r$ non-zero entries.*

Remark 44 *We underline that this lemma is valid because the unsigned adjacency matrices A^+ and A^- have disjoint support. We do not know if similar results could be obtained for the Signed Stochastic Block Model defined by Mercado et al. (2016).*

Proof We denote $A^{\pm}_{sup}$ (resp. $A^{\pm}_{inf}$) the upper (resp. lower) triangular part of the unsigned adjacency matrices. Using this decomposition, we have

$$A = A^+_{inf} + A^+_{sup} - A^-_{inf} - A^-_{sup}.$$

We note that $A^+_{inf}, A^+_{sup}, A^-_{inf}, A^-_{sup}$ have disjoint supports, and each of them has independent entries. We can hence apply Theorem 54 to each of these matrices, where we note that for each matrix

$$d := n \max_{j,j'} \mathbb{E}[A_{jj'}] = np(1-\eta) \leq 2\overline{d}(1-\eta).$$

With probability at least $1 - 2 \times 3n^{-r}$, there exists $\mathcal{N}^{\pm}_{inf}, \mathcal{R}^{\pm}_{inf}, \mathcal{C}^{\pm}_{inf}, \mathcal{N}^{\pm}_{sup}, \mathcal{R}^{\pm}_{sup}, \mathcal{C}^{\pm}_{sup}$ four partitions of $[n] \times [n]$ that have the subsequent properties. For e.g., for A^+_{inf},

- $\|(A^+_{inf} - \mathbb{E}A^+_{inf})_{\mathcal{N}}\| \leq Cr^{3/2}\sqrt{d} \leq Cr^{3/2}\sqrt{2\overline{d}(1-\eta)}$;
- $\mathcal{R}^+_{inf}$ (resp. $\mathcal{C}^+_{inf}$) intersects at most $n/d \leq n/\overline{d}$ columns (resp. rows) of $[n] \times [n]$;
- each row (resp. column) of $(A^+_{inf})_{\mathcal{R}}$ (resp. $(A^+_{inf})_{\mathcal{C}}$) have at most $32r$ ones.

We note that this decomposition holds simultaneously for $A^{\pm}_{inf}$ and $A^{\pm}_{sup}$. Taking the unions of these subsets,

$$\mathcal{N} = \mathcal{N}^+_{inf} \cup \mathcal{N}^+_{sup} \cup \mathcal{N}^-_{inf} \cup \mathcal{N}^-_{sup},$$

and similarly for $\mathcal{R}$ and $\mathcal{C}$, we have, with the triangle inequality

$$\begin{aligned}
&\|(A - \mathbb{E}A)_{\mathcal{N}}\| \\
&\quad = \|(A^+_{inf} - \mathbb{E}A^+_{inf})_{\mathcal{N}^+_{inf}} + (A^+_{sup} - \mathbb{E}A^+_{sup})_{\mathcal{N}^+_{sup}} - (A^-_{inf} - \mathbb{E}A^-_{inf})_{\mathcal{N}^-_{inf}} \\
&\qquad\qquad - (A^-_{sup} - \mathbb{E}A^-_{sup})_{\mathcal{N}^-_{sup}}\| \\
&\quad \leq \|(A^+_{inf} - \mathbb{E}A^+_{inf})_{\mathcal{N}^+_{inf}}\| + \|(A^+_{sup} - \mathbb{E}A^+_{sup})_{\mathcal{N}^+_{sup}}\| + \|(A^-_{inf} - \mathbb{E}A^-_{inf})_{\mathcal{N}^-_{inf}}\| \\
&\qquad\qquad + \|(A^-_{sup} - \mathbb{E}A^-_{sup})_{\mathcal{N}^-_{sup}}\| \\
&\quad \leq 4Cr^{3/2}\sqrt{d} \leq C_1 r^{3/2}\sqrt{\overline{d}(1-\eta)},
\end{aligned}$$

with $C_1 = 4C\sqrt{2}$. Moreover, each row of $\mathcal{R}$ (resp. each column of $\mathcal{C}$) has at most $2 \times 32r$ entries equal to 1 and $2 \times 32r$ entries equal to -1, which means at most $128r$ non-zero entries. Finally $\mathcal{R}$ (resp. $\mathcal{C}$) intersects at most $4n/\overline{d}$ rows (resp. columns) of $[n] \times [n]$. ■

For the degree matrix $\overline{D}$, we use inequality (4.3) from (Le et al., 2017). Recall that the degree of node j is $\overline{D}_{jj} = \sum_{j'=1}^{n}(A^{+}_{jj'} + A^{-}_{jj'})$ which is a sum of n independent Bernoulli variables with bounded variance d/n. We can thus find an upper bound on the error $\|\overline{D} - \mathbb{E}[\overline{D}]\|_F$. This bound is weaker than the one obtained in Lemma 40 with the assumption $p \gtrsim \frac{\ln n}{n}$.

Lemma 45 *There exists a constant $C' > 0$ such that for any $r \geq 1$, with probability at least $1 - e^{-2r}$, it holds true*

$$\sum_{j=1}^{n}(\overline{D}_{jj'} - \overline{d})^2 \leq C' r^2 n d \leq 2C' r^2 n \overline{d}(1-\eta).$$

5.4.2 Properties of the regularized Laplacian outside the core

In this section, we will bound the norm of the Signed Laplacian restricted to the subsets of edges $\mathcal{N}$ and $\mathcal{C}$. The following "restriction lemma" is an extension of Lemma 8.1 in Le et al. (2015) for Signed Laplacian matrices.

Lemma 46 *(Restriction of Signed Laplacian) Let B be a $n \times n$ symmetric matrix, B_γ its regularized form as described in Section 2.2, and $\mathcal{C} \subset [n] \times [n]$. We denote $\overline{D}_\gamma$ the regularized degree matrix , and $\overline{L}_\gamma = \overline{D}_\gamma^{-1/2} B_\gamma \overline{D}_\gamma^{-1/2}$ the modified "Laplacian" and $B_{\mathcal{C}}$ the $n \times n$ matrix such that the entries outside of $\mathcal{C}$ are set to 0. Let $0 < \varepsilon < 1$ such that the degree of each node in $(B_\gamma)_{\mathcal{C}}$ is less that ε times the the corresponding degree in B_γ. Then we have*

$$\|(\overline{L}_\gamma)_{\mathcal{C}}\| \leq \sqrt{\varepsilon}.$$

Proof We denote $\overline{D}_r$ (resp. $\overline{D}_c$) the degree matrix of $(B_\gamma)_{\mathcal{C}}$ (resp. $(B_\gamma)_{\mathcal{C}}^T$) and $\widetilde{L}$ its regularized "Laplacian" (it is not necessarily a symmetric matrix) where

$$\widetilde{L} = (\overline{D}_r^{1/2})^{\dagger} (B_\gamma)_{\mathcal{C}} (\overline{D}_c^{1/2})^{\dagger}.$$

By definition of $\overline{L}_\gamma$, $(\overline{L}_\gamma)_{\mathcal{C}} = \overline{D}_\gamma^{-1/2}(B_\gamma)_{\mathcal{C}}\overline{D}_\gamma^{-1/2}$. Since in $(B_\gamma)_{\mathcal{C}}$, some entries in B are set to 0, we have that for all $1 \leq j \leq n$,

$$(\overline{D}_c)_{jj} \leq [\overline{D}_\gamma]_{jj}.$$

Moreover, by assumption, $(\overline{D}_r)_{jj} \leq \varepsilon [\overline{D}_\gamma]_{jj}$. We denote $X = (\overline{D}_r^{1/2})^{\dagger}$, $Y = (\overline{D}_c^{1/2})^{\dagger}$ and $Z = \overline{D}_\gamma^{-1/2}$, and now we have

$$\overline{L}_{\mathcal{C}} = Z B_{\mathcal{C}} Z = Z X^{\dagger} X B_{\mathcal{C}} Y Y^{\dagger} Z = Z X^{\dagger} \widetilde{L} Y^{\dagger} Z.$$

Because $\|ZX^{\dagger}\| \leq \sqrt{\varepsilon}$ and $\|Y^{\dagger}Z\| \leq 1$, by sub-multiplicativity of the norm, we thus obtain

$$\|\overline{L}_{\mathcal{C}}\| \leq \|ZX^{\dagger}\| \cdot \|\widetilde{L}\| \cdot \|Y^{\dagger}Z\| \leq \sqrt{\varepsilon}\|\widetilde{L}\|.$$

In addition, by considering the $2n \times 2n$ symmetric matrix $\widetilde{L}'$

$$\widetilde{L}' = \begin{pmatrix} 0_n & \widetilde{L} \\ \widetilde{L} & 0_n \end{pmatrix},$$

we have $\|\widetilde{L}'\| = \|\widetilde{L}\| \leq 1$. In fact, $\widetilde{L}'$ is equal to the identity matrix minus the regularized Laplacian of

$$\begin{pmatrix} 0_n & (B_\gamma)_{\mathcal{C}} \\ (B_\gamma)_{\mathcal{C}}^T & 0_n \end{pmatrix}.$$

Using Appendix E, we can conclude that the eigenvalues of $\widetilde{L}'$ are between -1 and 1, leading to $\|\widetilde{L}'\| \leq 1$. Hence, we finally arrive at $\|(\overline{L}_\gamma)_{\mathcal{C}}\| \leq \sqrt{\varepsilon}$. ■

Remark 47 *We note that this lemma is not specific to the rows of the matrix B, and one could also derive the same lemma with the assumptions on the columns of the matrix.*

5.5 Error bounds w.r.t the expected regularized Laplacian and expected Signed Laplacian

In this section, we prove an upper bound on the errors $\|L_\gamma - \mathcal{L}_\gamma\|$ and $\|L_\gamma - \mathcal{L}_{sym}\|$ from Theorem 11. We will use the decomposition of the set of edges $(\mathcal{N}, \mathcal{R}, \mathcal{C})$ from Lemma 43, and sum the errors on each of these subsets of edges. We recall that on the subset $\mathcal{N}$, we have an upper bound on $\|(A - \mathbb{E}A)_{\mathcal{N}}\|$. We will also use the fact that the regularized degrees $[\overline{D}_\gamma]_{jj}$ are lower-bounded by the regularization parameter γ. On the subsets $\mathcal{R}$ and $\mathcal{C}$, we will use Lemma 46 to upper bound the norm of the regularized Laplacian.

Lemma 48 *Under the assumptions of Theorem 11, for any $r \geq 1$, with probability at least $1 - 7e^{-2r}$, we have*

$$\|L_\gamma - \mathcal{L}_\gamma\| \leq \frac{Cr^2}{\sqrt{\gamma}} \left(1 + \frac{\overline{d}}{\gamma}\right)^{5/2} + \frac{32\sqrt{2r}}{\sqrt{\gamma}} + \frac{8}{\sqrt{\overline{d}}}. \tag{78}$$

Proof Let $L_\gamma - \mathcal{L}_\gamma = S + T$ with

$$\begin{aligned} S &= (\overline{D}_\gamma)^{-1/2} A_\gamma (\overline{D}_\gamma)^{-1/2} - (\overline{D}_\gamma)^{-1/2} \mathbb{E}A_\gamma (\overline{D}_\gamma)^{-1/2} = (\overline{D}_\gamma)^{-1/2} (A_\gamma - \mathbb{E}A_\gamma)(\overline{D}_\gamma)^{-1/2}, \\ T &= (\overline{D}_\gamma)^{-1/2} \mathbb{E}A_\gamma (\overline{D}_\gamma)^{-1/2} - (\mathbb{E}\overline{D}_\gamma)^{-1/2} \mathbb{E}A_\gamma (\mathbb{E}\overline{D}_\gamma)^{-1/2}. \end{aligned}$$

We will bound the norm of $S + T$ on $\mathcal{N}$, and the norms of L_γ and $\mathcal{L}_\gamma$ on the residuals $\mathcal{R}, \mathcal{C}$. We first use the triangle inequality to obtain

$$\begin{aligned} &\|L_\gamma - \mathcal{L}_\gamma\| \\ &\quad \leq \| (L_\gamma - \mathcal{L}_\gamma)_{\mathcal{N}} \| + \| ((L_\gamma - I) - (\mathcal{L}_\gamma - I))_{\mathcal{R}} \| + \| (L_\gamma - \mathcal{L}_\gamma)_{\mathcal{C}} \| \\ &\quad \leq \| (L_\gamma - \mathcal{L}_\gamma)_{\mathcal{N}} \| + \| (I - L_\gamma)_{\mathcal{R}} \| + \| (I - \mathcal{L}_\gamma)_{\mathcal{R}} \| + \| (I - L_\gamma)_{\mathcal{C}} \| + \| (I - \mathcal{L}_\gamma)_{\mathcal{C}} \| \\ &\quad = \| (S + T)_{\mathcal{N}} \| + \| (I - L_\gamma)_{\mathcal{R}} \| + \| (I - \mathcal{L}_\gamma)_{\mathcal{R}} \| + \| (I - L_\gamma)_{\mathcal{C}} \| + \| (I - \mathcal{L}_\gamma)_{\mathcal{C}} \| \\ &\quad \leq \|S_{\mathcal{N}}\| + \|T_{\mathcal{N}}\| + \| (I - L_\gamma)_{\mathcal{R}} \| + \| (I - \mathcal{L}_\gamma)_{\mathcal{R}} \| + \| (I - L_\gamma)_{\mathcal{C}} \| + \| (I - \mathcal{L}_\gamma)_{\mathcal{C}} \|. \end{aligned}$$

1. Bounding the norm $\|T_{\mathcal{N}}\|$. Denoting $\gamma = \gamma^+ + \gamma^-$, we have that

$$\begin{aligned}
\|T_{\mathcal{N}}\|^2 &\leq \|T_{\mathcal{N}}\|_F^2 \\
&= \sum_{j,j'=1}^{n} T_{jj'}^2 \\
&= \sum_{j,j'=1}^{n} \left(\mathbb{E}A_{jj'} + (\gamma^+ - \gamma^-)/n\right)^2 \left[\frac{1}{\sqrt{(\overline{D}_{jj}+\gamma)(\overline{D}_{j'j'}+\gamma)}} - \frac{1}{\overline{d}+\gamma}\right]^2 \qquad (79) \\
&\leq \frac{(\overline{d}+\gamma)^2}{2n^2\gamma^6}\left[\sum_{j=1}^{n}(\overline{D}_{jj}+\gamma)^2 \sum_{j'=1}^{n}(\overline{D}_{j'j'}-\overline{d})^2 + n(\overline{d}+\gamma)^2\sum_{i=1}^{n}(\overline{D}_{jj}-\overline{d})^2\right]. \qquad (80)
\end{aligned}$$

To upper bound (79) by (80), we have used the simplification trick in the proof of (Le et al., 2017, Theorem 4.1) which we now recall. Firstly, the second factor of (79) can be upper bounded in the following way. For $1 \leq j, j' \leq n$,

$$\begin{aligned}
&\left|\frac{1}{\sqrt{(\overline{D}_{jj}+\gamma)(\overline{D}_{j'j'}+\gamma)}} - \frac{1}{\overline{d}+\gamma}\right| \\
&= \frac{|(\overline{D}_{jj}+\gamma)(\overline{D}_{j'j'}+\gamma) - (\overline{d}+\gamma)^2|}{(\overline{D}_{jj}+\gamma)(\overline{D}_{j'j'}+\gamma)(\overline{d}+\gamma) + \sqrt{(\overline{D}_{jj}+\gamma)(\overline{D}_{j'j'}+\gamma)}(\overline{d}+\gamma)^2} \\
&\leq \frac{|(\overline{D}_{jj}+\gamma)(\overline{D}_{j'j'}+\gamma) - (\overline{d}+\gamma)^2|}{2\gamma^3} \\
&= \frac{|(\overline{D}_{jj}+\gamma)(\overline{D}_{j'j'}+\gamma) - (\overline{d}+\gamma)(\overline{D}_{jj}+\gamma) + (\overline{d}+\gamma)(\overline{D}_{jj}+\gamma) - (\overline{d}+\gamma)^2|}{2\gamma^3} \\
&= \frac{|(\overline{D}_{jj}-\overline{d})(\overline{D}_{j'j'}+\gamma) + (\overline{d}+\gamma)(\overline{D}_{jj}-\overline{d})|}{2\gamma^3}, \qquad (81)
\end{aligned}$$

where the inequality comes from the fact that $\overline{D}_{jj} + \gamma \geq \gamma$. Secondly, we use the inequality $(a+b)^2 \leq 2(a^2+b^2)$ and we recall that by definition, we can bound the first factor of (79) by $|\mathbb{E}(A_\gamma)_{jj'}| \leq \frac{\overline{d}+\gamma}{n}$. This finally leads to (80).

Now we will bound each term of (80). Using Lemma 45, we have, for any $r \geq 1$, with probability at least $1 - e^{-2r}$,

$$\sum_{j=1}^{n}(\overline{D}_{jj}-\overline{d})^2 \leq 2C'r^2n\overline{d}(1-\eta) \leq 2C'r^2n\overline{d}.$$

If this holds, then the first term of (80) is upper bounded by

$$\begin{aligned}\sum_{i=1}^{n}(\overline{D}_{jj}+\gamma)^2\sum_{j=1}^{n}(\overline{D}_{j'j'}-\overline{d})^2 &\le \left(2\sum_{j=1}^{n}(\overline{D}_{jj}-\overline{d})^2+2n(\overline{d}+\gamma)^2\right)\sum_{j'=1}^{n}(\overline{D}_{j'j'}-\overline{d})^2\\ &\le 2C'r^2n\overline{d}\left(4C'r^2n\overline{d}+2n(\overline{d}+\gamma)^2\right)\\ &\le 2C'r^2n(\overline{d}+\gamma)(1-\eta)\left(4C'r^2nd+2n(\overline{d}+\gamma)^2\right)\\ &\le 2C'r^2n(\overline{d}+\gamma)\left(2(2C'+1)r^2n(d+\gamma)^2\right)\\ &\le C_1r^4n^2(\overline{d}+\gamma)^3,\end{aligned}$$

with $C_1 = 4C'(2C'+1)$. Similarly, we can bound the second term of (80)

$$n(\overline{d}+\gamma)^2\sum_{j=1}^{n}(\overline{D}_{jj}-\overline{d})^2 \le 2C'(\overline{d}+\gamma)^2r^2n^2\overline{d} \le 2C'(\overline{d}+\gamma)^3r^2n^2.$$

Hence, we obtain the following upper bound of (80)

$$\|T_{\mathcal{N}}\|^2 \le \frac{(C_1+2C')r^4}{2\gamma^6}(\overline{d}+\gamma)^5 = \frac{C_2r^4}{\gamma}\left(1+\frac{\overline{d}}{\gamma}\right)^5, \tag{82}$$

with $C_2 = (C_1+2C')/2$.

2. Bounding the norm $\|S_{\mathcal{N}}\|$. We first note that

$$S = (\overline{D}_\gamma)^{-1/2}(A_\gamma - \mathbb{E}A_\gamma)(\overline{D}_\gamma)^{-1/2} = (\overline{D}_\gamma)^{-1/2}(A-\mathbb{E}A)(\overline{D}_\gamma)^{-1/2}.$$

We also recall that $\|\overline{D}_\gamma\| \ge \gamma$. Hence, using Lemma 43, with probability at least $1-6n^{-r}$, we have

$$\begin{aligned}\|S_{\mathcal{N}}\| \le \|\overline{D}_\gamma^{-1/2}\|\,\|(A-\mathbb{E}A)_{\mathcal{N}}\|\,\|\overline{D}_\gamma^{-1/2}\| \le \|(A-\mathbb{E}A)_{\mathcal{N}}\|/\gamma &\le \frac{Cr^{3/2}}{\gamma}\sqrt{\overline{d}(1-\eta)}\\ &\le \frac{Cr^{3/2}}{\gamma}\sqrt{\overline{d}}.\end{aligned} \tag{83}$$

Summing the bounds in (82) and (83), we have the intermediate result

$$\|(L_\gamma-\mathcal{L}_\gamma)_{\mathcal{N}}\| \le \frac{Cr^{3/2}}{\gamma}\sqrt{\overline{d}} + \frac{\sqrt{C_2}r^2}{\sqrt{\gamma}}\left(1+\frac{\overline{d}}{\gamma}\right)^{5/2} \tag{84}$$

$$\le \frac{r^2}{\sqrt{\gamma}}\left(C\sqrt{\frac{\overline{d}}{\gamma}}+\sqrt{C_2}\left(1+\frac{\overline{d}}{\gamma}\right)^{5/2}\right) \tag{85}$$

$$\le \frac{r^2}{\sqrt{\gamma}}(C+\sqrt{C_2})\left(1+\frac{\overline{d}}{\gamma}\right)^{5/2} = \frac{C_3r^2}{\sqrt{\gamma}}\left(1+\frac{\overline{d}}{\gamma}\right)^{5/2}, \tag{86}$$

with $C_3 = C+\sqrt{C_2}$.

3. Bounding $\left\|(L_\gamma)_{\mathcal{R}}\right\|, \left\|(L_\gamma)_{\mathcal{C}}\right\|, \left\|(\mathcal{L}_\gamma)_{\mathcal{R}}\right\|, \left\|(\mathcal{L}_\gamma)_{\mathcal{C}}\right\|$. Using the proof of Lemma 43, each row of $A_{\mathcal{R}}$ has at most $128r$ non-zeros entries and intersects at most $4n/\overline{d}$ columns. Thus, for all $1 \le j \le n$

$$\sum_{j'=1}^{n} \left[(A_\gamma^+ + A_\gamma^-)_{\mathcal{R}}\right]_{jj'} \le 128r + \frac{4\gamma}{\overline{d}} = \gamma\left(\frac{128r}{\gamma} + \frac{4}{\overline{d}}\right) \le \sum_{j'} \left[A_\gamma^+ + A_\gamma^-\right]_{jj'} \left(\frac{128r}{\gamma} + \frac{4}{\overline{d}}\right),$$

as $\sum_{j'}[A_\gamma^+ + A_\gamma^-]_{jj'} \ge n \times \left(\frac{\gamma^+}{n} + \frac{\gamma^-}{n}\right) = \gamma$. We can thus apply Lemma 46 with $\varepsilon = \frac{128r}{\gamma} + \frac{4}{\overline{d}}$, and we arrive at

$$\|(L_\gamma)_{\mathcal{R}}\| \le \sqrt{\frac{128r}{\gamma} + \frac{4}{\overline{d}}}.$$

We also obtain the same bound for $\|(L_\gamma)_{\mathcal{C}}\|$. Similarly, we have $\sum_{j'} \left[\mathbb{E}[A_\gamma^+] + \mathbb{E}[A_\gamma^-]\right]_{jj'} = (n-1)p + \gamma = \overline{d} + \gamma \ge \gamma$ and

$$\begin{aligned}\sum_{j'=1}^{n} \left[(\mathbb{E}[A_\gamma^+] + \mathbb{E}[A_\gamma^-])_{\mathcal{R}}\right]_{jj'} &\le 4\frac{np}{\overline{d}} + \frac{4\gamma}{\overline{d}} \le 8 + \frac{4\gamma}{\overline{d}} = \gamma\left(\frac{8}{\gamma} + \frac{4}{\overline{d}}\right) \\ &\le \sum_{j'} \left[\mathbb{E}[A^+]_\gamma + \mathbb{E}[A^-]_\gamma\right]_{jj'} \left(\frac{8}{\gamma} + \frac{4}{\overline{d}}\right).\end{aligned}$$

We arrive at $\|(\mathcal{L}_\gamma)_{\mathcal{R}}\| \le \sqrt{\frac{8}{\gamma} + \frac{4}{\overline{d}}}$, and finally, we also have $\|(\mathcal{L}_\gamma)_{\mathcal{C}}\| \le \sqrt{\frac{8}{\gamma} + \frac{4}{\overline{d}}}$.

4. Bounding $\|L_\gamma - \mathcal{L}_\gamma\|$. Summing up the bounds obtained in the first three steps, with probability at least $1 - e^{-2r} - 6n^{-r} \ge 1 - 7e^{-2r}$, we finally arrive at the bound

$$\begin{aligned}\|L_\gamma - \mathcal{L}_\gamma\| &\le \frac{C_3 r^2}{\sqrt{\gamma}}\left(1 + \frac{\overline{d}}{\gamma}\right)^{5/2} + 2\sqrt{\frac{128r}{\gamma} + \frac{4}{\overline{d}}} + 2\sqrt{\frac{8}{\gamma} + \frac{4}{\overline{d}}} \\ &\le \frac{C_3 r^2}{\sqrt{\gamma}}\left(1 + \frac{\overline{d}}{\gamma}\right)^{5/2} + 4\sqrt{\frac{128r}{\gamma} + \frac{4}{\overline{d}}} \\ &\le \frac{C_3 r^2}{\sqrt{\gamma}}\left(1 + \frac{\overline{d}}{\gamma}\right)^{5/2} + \frac{32\sqrt{2r}}{\sqrt{\gamma}} + \frac{8}{\sqrt{\overline{d}}}.\end{aligned}$$

■

This bound also provides easily a bound on the norm of $L_\gamma - \mathcal{L}_{sym}$.

Corollary 49 *(Error bound of the regularized Laplacian) With the notations of Theorem 8 and Theorem 11, and $\gamma = \gamma^+ + \gamma^-$, we have*

$$\|L_\gamma - \mathcal{L}_{sym}\| \le \frac{Cr^2}{\sqrt{\gamma}}\left(1 + \frac{\overline{d}}{\gamma}\right)^{5/2} + \frac{32\sqrt{2r}}{\sqrt{\gamma}} + \frac{8}{\sqrt{\overline{d}}} + \frac{\gamma}{\overline{d} + \gamma} =: \Delta_L(\gamma, \overline{d}). \tag{87}$$

In particular, for the choice $\gamma = \overline{d}^{7/8}$, if $p \ge 2/n$, we obtain

$$\|L_\gamma - \mathcal{L}_{sym}\| \le \left(128Cr^2 + 1\right)\overline{d}^{-1/8}.$$

Proof By triangular inequality,

$$\|L_\gamma - \mathcal{L}_{sym}\| \leq \|L_\gamma - \mathcal{L}_\gamma\| + \|\mathcal{L}_\gamma - \mathcal{L}_{sym}\|.$$

For the second term on the RHS, we have

$$\|\mathcal{L}_\gamma - \mathcal{L}_{sym}\| = \left\| \frac{1}{\overline{d}+\gamma}\mathbb{E}A - \frac{1}{\overline{d}}\mathbb{E}A \right\| = \frac{\gamma}{\overline{d}(\overline{d}+\gamma)}\|\mathbb{E}A\| \leq \frac{\gamma}{\overline{d}+\gamma}. \tag{88}$$

The last inequality comes from the fact that $\|\mathbb{E}A\| \leq (n-1)p(1-\eta) \leq \overline{d}$. Thus, by summing the bound obtained in Lemma 48 and (88), we arrive at the expected result in (87). Moreover, if $\gamma \leq \overline{d}$, since $C > 1$, one can readily verify that

$$\|\mathcal{L}_\gamma - \mathcal{L}_{sym}\| \leq 128Cr^2\frac{\overline{d}^{\frac{5}{2}}}{\gamma^3} + \frac{\gamma}{\overline{d}}. \tag{89}$$

If $\gamma = \overline{d}^{7/8}$, then $\gamma \leq \overline{d}$ holds provided $\overline{d} \geq 1$ or equivalently, $p \geq \frac{1}{n-1}$. The latter is ensured if $p \geq 2/n$ (since $n \geq 2$). Plugging this in (89), we then obtain the bound

$$\|\mathcal{L}_\gamma - \mathcal{L}_{sym}\| \leq \left(128Cr^2 + 1\right)\overline{d}^{-1/8}.$$

This concludes the proof of Corollary 49 and Theorem 11. ■

5.6 Error bound on the eigenspaces and mis-clutering rate in the sparse regime

This section provides a bound on the misalignment error of the eigenspaces of L_γ and $\mathcal{L}_{sym}$, which then leads to a bounds on the mis-clustering rate of the k-means clustering step.

5.6.1 Eigenspace alignment

Using the bound from Corollary 49, we can perform the same analysis of the eigenspaces of L_γ and $\mathcal{L}_{sym}$, as in Theorem 8, which will prove Theorem 13. We apply, once again, Weyl's inequality and the Davis-Kahan theorem to bound the distance between the two subspaces $\mathcal{R}(V_{k-1}(L_\gamma))$ and $\mathcal{R}(V_{k-1}(\mathcal{L}_{sym}))$. We have that

$$\lambda_{n-k+1}(L_\gamma) - \lambda_{n-k+2}(\mathcal{L}_{sym}) \geq \lambda_{gap} - \|L_\gamma - \mathcal{L}_{sym}\| \geq \lambda_{gap} - \Delta_L(\gamma, \overline{d}),$$

using Corollary 49. If $\gamma = \gamma_0\overline{d}^{7/8}$, then

$$\Delta_L(\gamma, \overline{d}) \leq \left(128Cr^2 + 1\right)(\overline{d})^{-1/8} := \frac{C_4}{\overline{d}^{1/8}},$$

with $C_4 = 128Cr^2 + 1$. For $0 < \delta < 1/2$, we would like to ensure that

$$\lambda_{gap} - \Delta_L(\gamma, \overline{d}) \geq \lambda_{gap}\left(1 - \frac{\delta}{2}\right).$$

Hence, using the lower bound on the eigengap from Lemma 39, it suffices that

$$\lambda_{gap} - \frac{C_4}{\overline{d}^{1/8}} \geq \lambda_{gap}\left(1 - \frac{\delta}{2}\right) \iff \overline{d}^{1/8} \geq \frac{2kC_4}{\delta(1-2\eta)} \iff p \geq \left(\frac{2kC_4}{\delta(1-2\eta)}\right)^8 \frac{1}{n-1}.$$

Thus, the condition $p \geq \left(\frac{2kC_4}{\delta(1-2\eta)}\right)^8 \frac{2}{n}$ is sufficient. Applying the Davis-Kahan theorem, we arrive at

$$\|(I - V_{k-1}(L_\gamma)V_{k-1}(L_\gamma)^T)V_{k-1}(\mathcal{L}_{sym})\| \leq \frac{\delta\lambda_{gap}/2}{\lambda_{gap}(1-\delta/2)} \leq \frac{\delta/2}{1-\delta/2} \leq \delta,$$

and using once again Proposition 57, there exists an orthogonal matrix $O \in \mathbb{R}^{(k-1)\times(k-1)}$ such that

$$\|V_{k-1}(L_\gamma) - \Theta R_{k-1}O\| \leq 2\delta.$$

5.7 Proof of Theorem 16

In this section, we finally prove our result on the clustering performance of the Signed Laplacian and regularized Laplacian algorithms. The proof essentially relies on the following lemma, which provides a lower bound on the distance between two rows of $\Delta^{-1}R_{k-1}$, with $\Delta = \mathrm{diag}(\sqrt{n_i})$.

Lemma 50 *For all $1 \leq i \neq i' \leq k$, we have $\|(R_{k-1})_{i*} - (R_{k-1})_{i'*}\| \geq 1$. Moreover, for $i \in [k]$, it holds that*

$$\min_{\substack{i,i'\in[k],i\neq i'\\ j\in C_i, j'\in C_{i'}}} \left\|(\Delta^{-1}R_{k-1})_{j*} - (\Delta^{-1}R_{k-1})_{j'*}\right\|^2 \geq \frac{2}{3n_i}.$$

Proof Recall from (69) that $\overline{C} = p(1-2\eta)uu^T + \mathrm{diag}(d_i)$, with $d_i = u_i^2 + \left(1 + \frac{p}{\overline{d}}(1-2\eta)\right)$ and $u_i = \sqrt{\frac{n_i}{\overline{d}}}, 1 \leq i \leq k$. Moreover, from (70), $\overline{C} = R\Lambda R$ with $R = [R_{k-1}\ \gamma_1]$ and γ_1 the largest eigenvector of $\overline{C}$. We first show that the entries of γ_1 are necessarily of the same sign, i.e. $(\gamma_1)_i \geq 0, \forall i$ or $(\gamma_1)_i \leq 0, \forall i$. In fact, by definition, γ_1 is the solution of

$$\max_{\|v\|=1} v^T\overline{C}v = \max_{\|v\|=1} p(1-2\eta)(v^u)^2 + \sum_{i=1}^k d_i v_i^2. \tag{90}$$

Since all the entries of u are positive, it is easy to see that any solution γ_1 of (90) necessarily has entries of the same sign (otherwise you could replace some $(\gamma_1)_i$) by $-(\gamma_1)_i$ and increase the objective function).

Let $i \neq i' \in [k]$. As R has orthonormal rows,

$$\begin{aligned}\langle R_{i*}, R_{i'*}\rangle = 0 &\iff \langle (R_{k-1})_{i*}, (R_{k-1})_{i'*}\rangle + \underbrace{(\gamma_1)_i(\gamma_1)_{i'}}_{\geq 0} = 0\\ &\implies \langle (R_{k-1})_{i*}, (R_{k-1})_{i'*}\rangle \leq 0.\end{aligned}$$

Hence,

$$\begin{aligned}\|(R_{k-1})_{i*} - (R_{k-1})_{i'*}\|^2 &= \|(R_{k-1})i*\|^2 + \left\|(R_{k-1})i'*\right\|^2 - 2\underbrace{\langle (R_{k-1})_{i*}, (R_{k-1})_{i'*}\rangle}_{\leq 0} \\ &\geq \|(R_{k-1})_{i*}\|^2 + \|(R_{k-1})_{i'*}\|^2 \\ &= 2 - \underbrace{[(\gamma_1)_i^2 + (\gamma_1)_{i'}^2]}_{\leq 1} \geq 1.\end{aligned}$$

In particular, this implies that R_{k-1} has k distinct rows. Now let $j, j' \in [n]$ such that $j \in C_i$ and $j' \in C_{i'}$. Recalling that with $\Delta = \mathrm{diag}(\sqrt{n_i})$, $V_{k-1}(\mathcal{L}_{sym}) = \Theta R_{k-1} = \Theta\Delta\Delta^{-1}R_{k-1} = \hat{\Theta}\Delta^{-1}R_{k-1}$, we have

$$\begin{cases}(\Delta^{-1}R_{k-1})_{j*} = \frac{1}{\sqrt{n_i}}(R_{k-1})_{i*}, \\ (\Delta^{-1}R_{k-1})_{j'*} = \frac{1}{\sqrt{n_{i'}}}(R_{k-1})_{i'*}.\end{cases}$$

Hence,

$$\begin{aligned}&\left\|(\Delta^{-1}R_{k-1})_{j*} - (\Delta^{-1}R_{k-1})_{j'*}\right\|^2 \\ &= \frac{1}{n_i}\|(R_{k-1})i*\|^2 + \frac{1}{n_{i'}}\|(R_{k-1})_{i'*}\|^2 - 2\frac{1}{\sqrt{n_i n_{i'}}}\underbrace{\langle (R_{k-1})_{i*}, (R_{k-1})_{i'*}\rangle}_{\leq 0} \\ &\geq \frac{1}{n_i}\|(R_{k-1})_{i*}\|^2 + \frac{1}{n_{i'}}\|(R_{k-1})_{i'*}\|^2 \\ &\geq \frac{1}{n_i} + \frac{1}{n_{i'}} - \frac{(\gamma_1)_i^2}{n_i} - \frac{(\gamma_1)_{i'}^2}{n_i'} \\ &\geq \frac{1}{n_i} + \frac{1}{n_{i'}} - \frac{(\gamma_1)_i^2 + (\gamma_1)_{i'}^2}{ns} \geq \frac{1}{n_i} + \frac{1}{n_{i'}} - \frac{1}{ns} \geq \frac{1}{n_i} + \frac{1}{nl} - \frac{1}{ns}.\end{aligned}$$

Besides, we know that $\frac{1}{nl} \geq \frac{\rho}{n_i}$ and $\frac{1}{ns} \leq \frac{1}{\rho n_i}$. Therefore, we obtain the bound

$$\left\|(\Delta^{-1}R_{k-1})_{j*} - (\Delta^{-1}R_{k-1})_{j'*}\right\|^2 \geq \frac{1}{n_i}\left(1 + \rho - \frac{1}{\rho}\right).$$

We will now prove that with the condition $\sqrt{\rho} > 1 - \frac{1}{4k(2+\sqrt{k})}$, we have $1 + \rho - \frac{1}{\rho} \geq \frac{2}{3}$ and this will lead to the final result. First, we note that $\rho > 1 - \frac{1}{2k(2+\sqrt{k})}$ and $2k(2+\sqrt{k}) \geq 12$, and $\frac{2k(2+\sqrt{k})}{2k(2+\sqrt{k})-1} \leq \frac{5}{4}$ for $k \geq 2$. Thus,

$$1 + \rho - \frac{1}{\rho} \geq 2 - \frac{1}{2k(2+\sqrt{k})} - \frac{2k(2+\sqrt{k})}{2k(2+\sqrt{k})-1} \geq 2 - \frac{1}{12} - \frac{5}{4} = \frac{2}{3}.$$

■

Remark 51 *In the equal-size case $n_i = \frac{n}{k}, \forall 1 \leq i \leq k$, since $\gamma_1 = \chi_1$, R_{k-1} has orthogonal rows and*

$$\|(R_{k-1})_{i*} - (R_{k-1})_{i'*}\|^2 = \|R_{i*} - R_{i'*}\|^2 = 2.$$

This implies that

$$\left\|(\Delta^{-1}R_{k-1})_{j*} - (\Delta^{-1}R_{k-1})_{j'*}\right\|^2 = \frac{2k}{n}.$$

From Lemma 50, we have that $\forall 1 \leq i \leq k$, $\min\limits_{\substack{i,i' \in [k], i \neq i' \\ j \in C_i, j' \in C_{i'}}} \left\|(\Delta^{-1}R_{k-1})_{j*} - (\Delta^{-1}R_{k-1})_{j'*}\right\|^2 \geq \frac{2}{3n_i}$. Hence with $\delta_i^2 := \frac{2}{3n_i}$ and using Lemma 35, we obtain

$$\begin{aligned}
\sum_{i=1}^{k} \delta_i^2 |S_i| = \sum_{i=1}^{k} \frac{2|S_i|}{3n_i} &\leq 4(4+2\xi) \left\|V_{k-1}(\overline{L_{sym}}) - V_{k-1}(\mathcal{L}_{sym})\right\|_F^2 \\
&\leq 4(16+8\xi)(k-1) \left\|V_{k-1}(\overline{L_{sym}}) - V_{k-1}(\mathcal{L}_{sym})O\right\|^2 \\
&\leq 8(16+8\xi)(k-1)\delta^2,
\end{aligned}$$

using Theorem 8 . Moreover, we have

$$\begin{aligned}
\left\|V_{k-1}(\overline{L_{sym}}) - V_{k-1}(\mathcal{L}_{sym})\right\|_F^2 &\leq 2(k-1) \left\|V_{k-1}(\overline{L_{sym}}) - V_{k-1}(\mathcal{L}_{sym})O\right\|^2 \\
&\leq 8(k-1)\delta^2 \\
&< 8(k-1) \cdot \frac{1}{12(16+8\xi)(k-1)} \\
&= \frac{n_i \delta_i^2}{16+8\xi}, \qquad \forall 1 \leq i \leq k.
\end{aligned}$$

Therefore, we can use the second part of Lemma 35 and finally conclude that

$$\sum_{i=1}^{k} \frac{|S_i|}{n_i} \leq 96(2+\xi)\delta^2.$$

For the regularized Laplacian algorithm, the same computations are valid using the result from Theorem 13.

6. Numerical experiments

In this section, we report on the outcomes of several numerical experiments that compare our two proposed algorithms with a suite of state-of-the-art methods from the signed clustering literature. We test the performances of the different algorithms on signed graphs drawn from our Signed Stochastic Block Model, as well as on three real-world data sets that are standard benchmarks in the signed networks literature. We rely on a previous Python implementation of SPONGE and Signed Laplacian (along with their respective normalized

versions), and of other methods from the literature[3], made available in the context of previous work of a subset of the authors of the present paper (Cucuringu et al., 2019). More specifically, we consider algorithms based on the adjacency matrix A, the Signed Laplacian matrix $\overline{L}$, its symmetrically normalized version $\overline{L}_{sym}$ (Kunegis et al., 2010), SPONGE and its normalized version SPONGE$_{sym}$, and the two algorithms introduced in Chiang et al. (2012) that optimize the Balanced Ratio Cut and the Balanced Normalized Cut objectives.

We note that once the low-dimensional embedding has been computed by any of the considered algorithms, the final partition is obtained after running k-means++ (Arthur and Vassilvitskii, 2007), which improves over the popular k-means algorithm by employing a careful seeding initialization procedure and is the typical choice in practice.

6.1 Grid search for choosing the parameters τ^+, τ^-

In the following experiments, the Signed Stochastic Block Model will be sampled with the following set of parameters

- the number of nodes $n = 5000$,
- the number of communities $k \in \{3, 5, 10, 20\}$,
- the relative size of communities (defined in Section 3.2) $\rho = 1$ (equal-size clusters) and $\rho = 1/k$ (non-equal size clusters).

For the edge density parameter p, we choose two sparsity regimes, *Regime I* and *Regime II*, where *Regime II* is strictly harder than *Regime I*, in the sense than for the same value of k, the edge density in *Regime I* is significantly larger compared to *Regime II*. The noise level η is chosen such that the recovery of the clusters is unsatisfactory for a subset of pairs of parameters (τ^+, τ^-). For each set of parameters, we sample 20 graphs from the SSBM and average the resulting ARI.

Our experimental setup is summarized in the following steps.

1. Select a set of parameters (k, ρ, p, η) from the regime of interest;
2. Sample a graph from the SSBM(n, k, ρ, p, η);
3. Extract the largest connected component of the measurement graph (regardless of the sign of the edges);
4. If the size of the latter is too small ($< n/2$), resample a graph until successful;
5. For each pair of parameters (τ^+, τ^-), compute the k-dimensional embeddings using the SPONGE$_{sym}$ algorithm (with the implementation in the *signet* package (Cucuringu et al., 2019));
6. Obtain a partition of the graph into k clusters, and compute the *ARI* between this estimated partition and the ground-truth clusters using the implementation in scikit-learn of the k-means++ algorithm;

[3]Python implementations of a suite of algorithms for signed clustering are available at `https://github.com/alan-turing-institute/signet`

7. Repeat 20 times the steps $2-7$ mentioned above, and record the average performance over the 20 runs.

The results in the dense regimes are reported in Figure 1, while those for the sparse regimes in Figure 2. This set of results indicate that the gradient of the ARI in the space of parameters (τ^+, τ^-) is larger when the cluster sizes are very unbalanced and the edge density is low. We attribute this to the fact that, for suitably chosen values, the parameters (τ^+, τ^-) are performing a form of regularization of the graph that can significantly improve the clustering performance.

6.2 Comparison of a suite of spectral methods

This section performs a comparison of the performance of the following spectral clustering algorithms. We rely on the same notation used in Cucuringu et al. (2019), when mentioning the names of the SPONGE algorithms, namely SPONGE and SPONGE$_{sym}$. The complete list of algorithms compared is as follows.

- the combinatorial (un-normalized) Signed Laplacian $\overline{L} = \overline{D} - A$,
- the symmetric Signed Laplacian $\overline{L}_{sym} = I - \overline{D}^{-1/2} A \overline{D}^{-1/2}$,
- SPONGE and SPONGE$_{sym}$ with a suitably chosen pair of parameters (τ^+, τ^-)
- the Balanced Ratio Cut $L_{BRC} = D^+ - A$
- the Balanced Normalized Cut $L_{BNC} = D^{-1/2}(D^+ - A)D^{-1/2}$.

For the combinatorial and symmetric Signed Laplacians $\overline{L}$ and $\overline{L_{sym}}$, we compute $(k-1)$-dimensional embeddings before applying the k-means++ algorithm. For all other methods, we use the k smallest eigenvectors.

In this experiment, we fix the parameters $n = 5000, k \in \{3, 5, 10, 20\}$ and p, η in a certain set, and for each plot, we vary the aspect ratio $\rho \in [0, 1]$. The relative proportions of the classes $s_i = \frac{n_i}{n}$ are chosen according to the following procedure

1. Fix $s'_1 = 1/k$, pick a value for ρ and compute $s'_k = s'_1/\rho$.
2. For $i \in [2, k-1]$, sample s'_i from the uniform distribution in the interval $[s'_1, s'_k]$.
3. Compute the proportions $s_i = \frac{s'_i}{\sum_{i=1}^k s'_i}$, and then sample the graph from the resulting SSBM.
4. Repeat 20 times the steps 1-3 mentioned above, and record the average performance over the 20 runs.

The results are reported in Figure 3. We note that in almost all settings, the SPONGE$_{sym}$ algorithm outperforms the other clustering methods, in particular for low values of the aspect ratio ρ. With the exception of the symmetric Signed Laplacian, most methods seem to perform worse when the aspect ratio is higher, meaning that the clusters are more unbalanced, which is a more challenging regime.

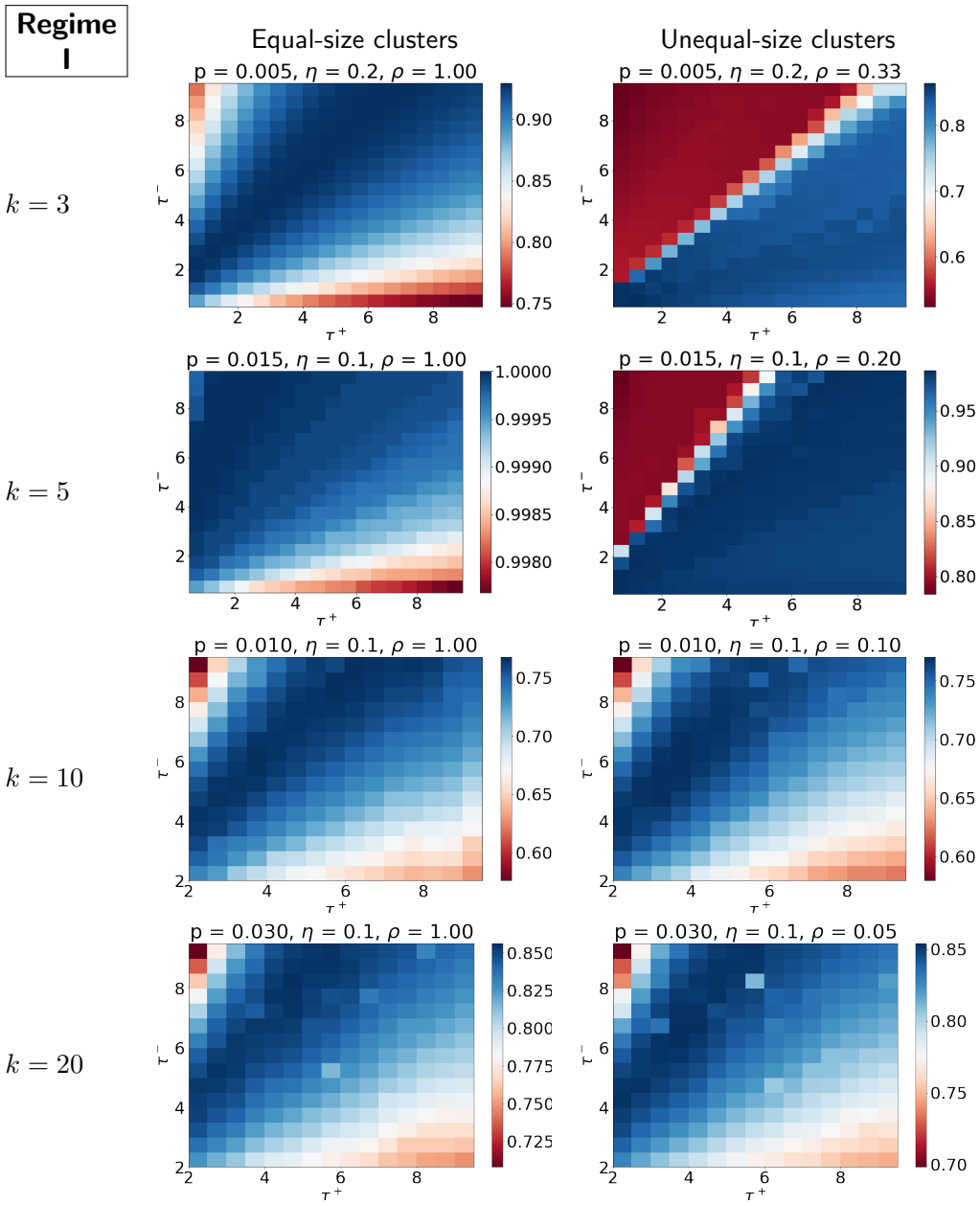


Figure 1: Heatmaps of the Adjusted Rand Index between the ground truth and the partition obtained using the SPONGE$_{sym}$ algorithm with varying regularization parameters (τ^+, τ^-), for a SSBM in *Regime I*, with $n = 5000$ and $k = \{3, 5, 10, 20\}$ clusters of equal sizes (left column) and unequal sizes (right column).

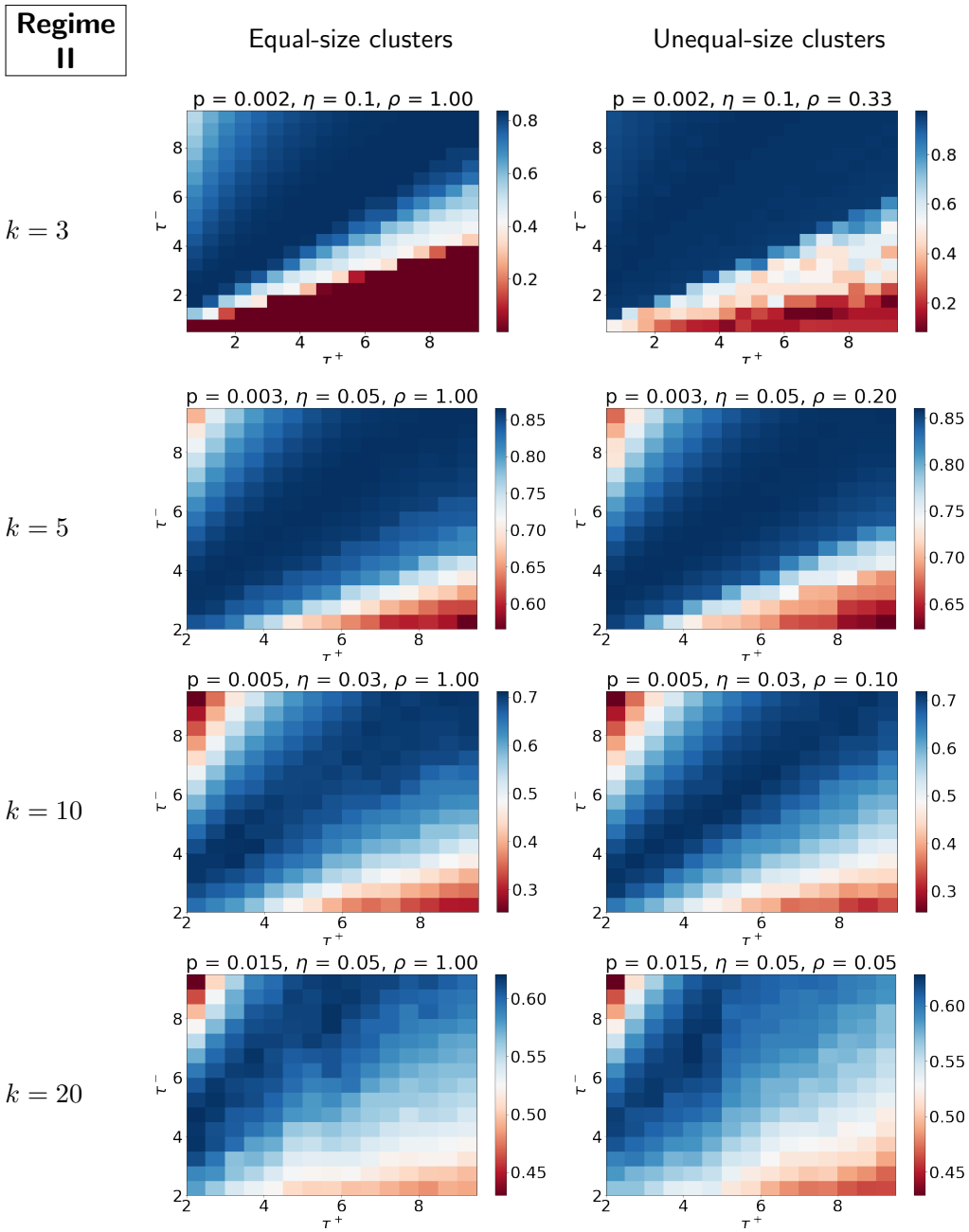


Figure 2: Heatmaps of the Adjusted Rand Index between the ground truth and the partition obtained using the SPONGE$_{sym}$ algorithm with varying regularization parameters (τ^+, τ^-), for a SSBM in *Regime II* with $n = 5000$ and $k = \{3, 5, 10, 20\}$ clusters of equal sizes (left column) and unequal sizes (right column).

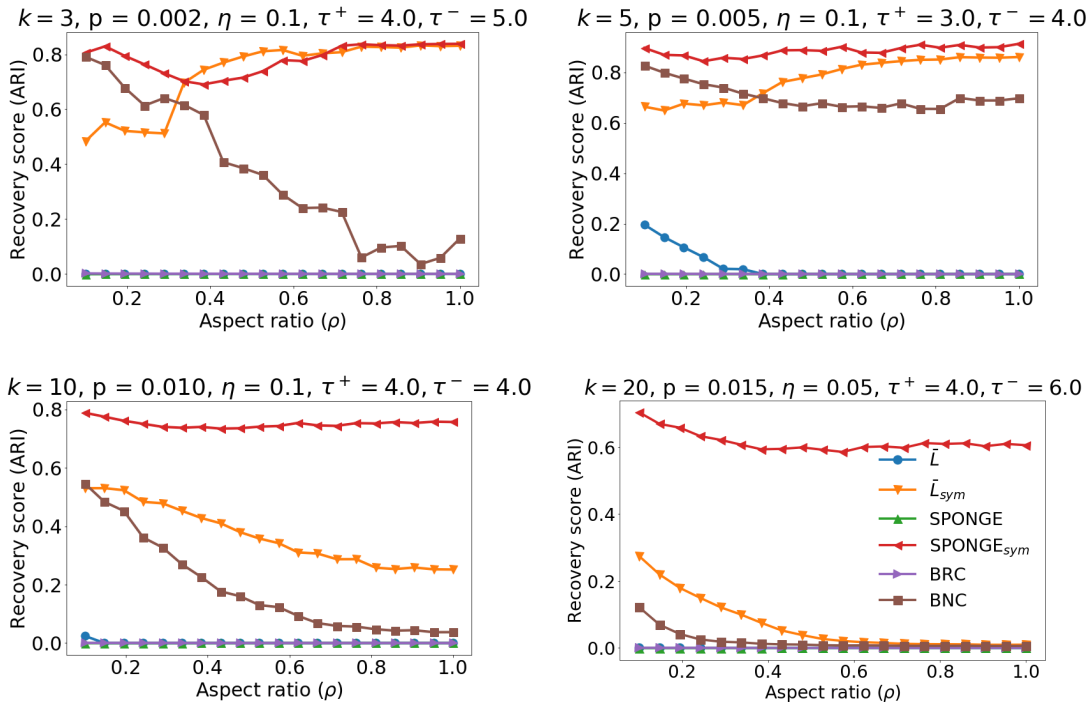


Figure 3: Performance of the various clustering algorithms, as measured by the Adjusted Rand Index, versus the aspect ratio ρ for a SSBM with $k = \{3, 5, 10, 20\}$ for $n = 5000$. For larger number of clusters, $k = 10$ and especially $k = 20$, SPONGE$_{sym}$ is essentially the only algorithm able to produce meaningful results, and clearly outperforms all the other methods. Note that no regularization has been used throughout this set of experiments.

6.3 Performance of the regularized algorithms in the sparse regime

In this batch of experiments, we study how the *regularized* Signed Laplacian and the SPONGE$_{sym}$ sparse algorithms perform. We consider sparse settings of the SSBM ($p \leq 0.003$) with $n = 5000$ nodes. For the SPONGE$_{sym}$ algorithm, we fix the parameters (τ^+, τ^-) in each setting. Our parameter selection procedure is to chose a pair of parameters that leads to a "good" recovery of the clusters for the unregularized algorithm (see Figure 2). We perform a grid search on the parameters (γ^+, γ^-) for each of the two regularized algorithms (see Figure 4 and Figure 5). For the regularized Signed Laplacian algorithm, we observe distinct regions of performance on the space of parameters (γ^+, γ^-). This is not predictable from our theoretical results, where the positive and negative regularization parameters play symmetric roles. We conjecture this to be due to the difference of density of the positive and negative subgraphs in our signed random graph model. For the SPONGE$_{sym}$ sparse algorithm, we note that the gradient of performances in the heatmaps (Figure 4, Figure 5) is similar to what was reported in Figure 2, which could be due to the fact that the parameters (τ^+, τ^-) already have a regularization effect.

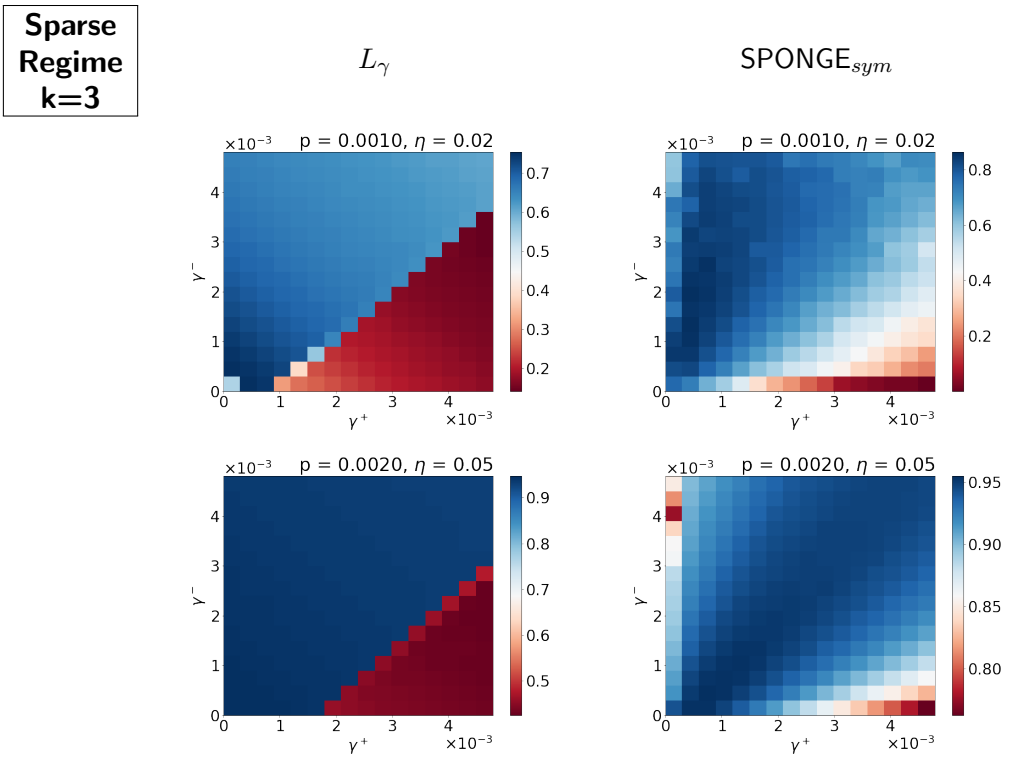


Figure 4: Heatmaps of the Adjusted Rand Index between the ground truth and the partition obtained using the L_γ and SPONGE$_{sym}$ algorithm with fixed parameters (τ^+, τ^-) and varying **regularization** parameters (γ^+, γ^-), for a SSBM in two **sparse** regimes, with $n = 5000$ and $k = 3$ clusters.

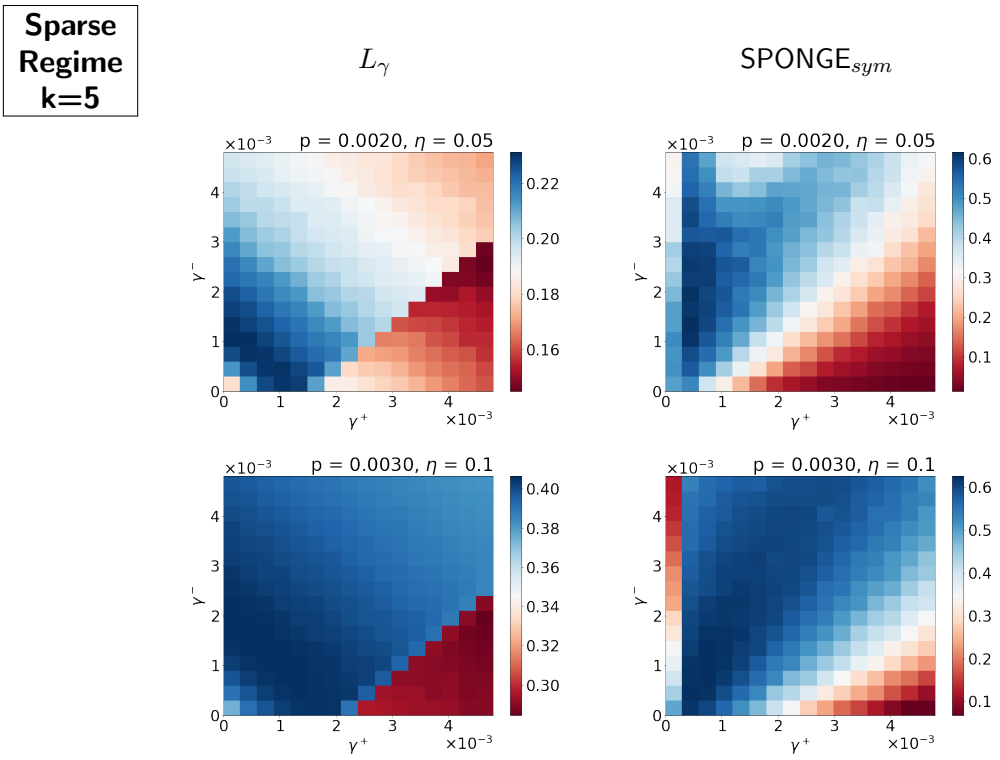


Figure 5: Heatmaps of the Adjusted Rand Index between the ground truth and the partition obtained using the L_γ and SPONGE$_{sym}$ algorithm with fixed parameters (τ^+, τ^-) and varying **regularization** parameters (γ^+, γ^-), for a SSBM in two **sparse** regimes, with $n = 5000$ and $k = 5$ clusters.

Dataset	Number of nodes	Edge density
Wikipedia	$11,259$	2.2×10^{-3}
Slashdot	$82,140$	1.3×10^{-4}
Bitcoin	$5,875$	3.6×10^{-3}

Table 1: Characteristics of the three benchmark data sets.

6.4 Performances on real-world data sets

Finally, we measure the performances of our unregularized and regularized algorithms on three benchmark data sets in the signed clustering problem: the Wikipedia Requests for Adminship, the Slashdot Zoo and Bitcoin data sets from Leskovec and Krevl (2014). These networks are large and sparse (see Table 1 for a summary of the number of nodes and edge densities). Since no ground-truth clusters are available for these networks, we measure the quality of the clustering using an objective score, namely the *normalized adjacency score*. This metric is defined as the sum of ratios of the number of positive edges over the number of negative edges within each cluster. We assume that a higher value of this score indicates a better partition of the node set. Our results are reported in Figure 6. We observe that the regularized versions of our algorithms, namely SPONGE$_{sym}$ and the Symmetric Signed Laplacian, perform much better that their respective unregularized versions, confirming the fact that regularization improves the performance of spectral algorithms in the sparse regime. We also note that the standard Spectral Clustering algorithm based on the signed adjacency matrix - denoted by A in the figure legend - also performs well on these real-world data sets.

7. Concluding remarks and future research directions

In this work, we provided a thorough theoretical analysis of the robustness of the SPONGE$_{sym}$ and symmetric Signed Laplacian algorithms, for graphs generated from a Signed Stochastic Block Model. Under this model, the sign of the edges (rather than the usual discrepancy of the edge densities across clusters versus within clusters) is an essential attribute which induces the underlying cluster structure of the graph. We proved that our signed clustering algorithms, based on suitably defined matrix operators, are able to recover the clusters under certain favorable noise regimes, and under two regimes of edge sparsity. Although the sparse setting is particularly challenging, our algorithms based on regularized graphs perform well, provided that the regularization parameters are suitably chosen. We also expect that the same type of analysis could be adapted to other probabilistic generative models for signed networks. For instance, extensions of the unsigned Stochastic Block Models, such as the Degree-Corrected Stochastic Block Model, that includes degree-heterogeneity could be considered, as well extensions to the setting of polarized communities, in the spirit of those proposed by Bonchi et al. (2019) and Xiao et al. (2020).

One theoretical question that has been not been answered yet relates to the choice of the positive and negative regularization parameters γ_+, γ_-. Having a data-driven approach to tune the regularization parameters would be of great use in many practical applications involving very sparse graphs. An interesting future line of work would be to study the latest

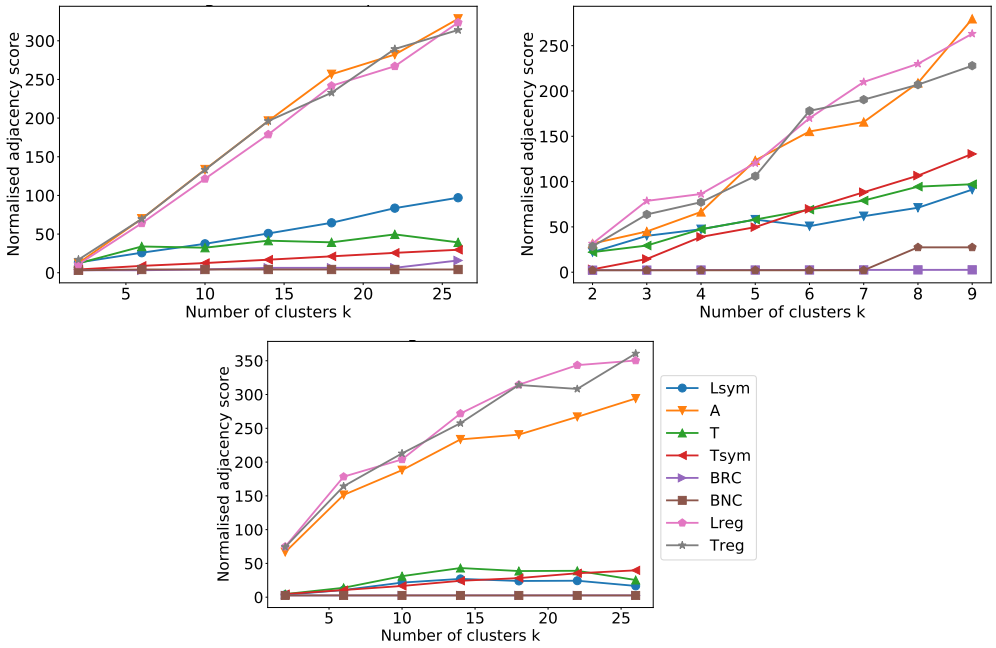


Figure 6: Objective clustering scores attained by the different spectral clustering algorithms, as we vary the number of clusters k on the Wikipedia (top left panel), the Bitcoin (top right panel), and Slashdot (bottom panel) data sets.

regularizing techniques based on powers of adjacency matrices or certain graph distance matrices, in the context of sparse signed graphs.

Yet another approach is to consider a pre-processing stage that performs low-rank matrix completion on the adjacency matrix, whose output could subsequently be used as input for our proposed algorithms. An extension of the Cheeger inequality to the setting of signed graphs, analogue to the generalized Cheeger inequality previously explored in Cucuringu et al. (2016), is another interesting research question. Extensions to the time-dependent setting and online clustering (Liberty et al., 2016; Mansfield et al., 2018), or when covariate information is available (Yan and Sarkar, 2020), are further research directions worth exploring, well motivated by real world applications involving signed networks.

References

Emmanuel Abbe. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

Emmanuel Abbe, Enric Boix-Adserà, Peter Ralli, and Colin Sandon. Graph powering and spectral robustness. *SIAM Journal on Mathematics of Data Science*, 2(1):132–157, 2020.

Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. Time-series clustering–a decade review. *Information Systems*, 53:16–38, 2015.

Daniel Aloise, Amit Deshpande, Pierre Hansen, and Preyas Popat. NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning*, 75(2):245–248, 2009.

Arash A. Amini, Aiyou Chen, Peter J. Bickel, and Elizaveta Levina. Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41 (4):2097–2122, 2013.

David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, page 1027–1035, USA, 2007. Society for Industrial and Applied Mathematics.

Afonso S. Bandeira and Ramon van Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *Ann. Probab.*, 44(4):2479–2506, 07 2016.

Afonso S Bandeira, Amit Singer, and Daniel A Spielman. A Cheeger inequality for the graph Connection Laplacian. *SIAM Journal on Matrix Analysis and Applications*, 34(4): 1611–1630, 2013.

Sujogya Banerjee, Kaushik Sarkar, Sedat Gokalp, Arunabha Sen, and Hasan Davulcu. Partitioning signed bipartite graphs for classification of individuals and organizations. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, pages 196–204. Springer, 2012.

Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation clustering. *Mach. Learn.*, 56 (1-3):89–113, June 2004.

R. Bhatia. *Matrix Analysis*. Springer New York, 1996.

Francesco Bonchi, Edoardo Galimberti, Aristides Gionis, Bruno Ordozgoiti, and Giancarlo Ruffo. Discovering polarized communities in signed networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 961–970, 2019.

Kamalika Chaudhuri, Fan Chung, and Alexander Tsiatas. Spectral clustering of graphs with general degrees in the extended planted partition model. In *25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 35.1–35.23, Edinburgh, Scotland, 2012. JMLR Workshop and Conference Proceedings.

Kai-Yang Chiang, Joyce Jiyoung Whang, and Inderjit S. Dhillon. Scalable clustering of signed networks using balance normalized cut. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, page 615–624, New York, NY, USA, 2012. Association for Computing Machinery.

Kai-Yang Chiang, Cho-Jui Hsieh, Nagarajan Natarajan, Inderjit S. Dhillon, and Ambuj Tewari. Prediction and clustering in signed networks: A local to global perspective. *Journal of Machine Learning Research*, 15:1177–1213, 2014.

Lingyang Chu, Zhefeng Wang, Jian Pei, Jiannan Wang, Zijin Zhao, and Enhong Chen. Finding gangs in war from signed networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1505–1514, 2016.

Fan Chung. Laplacians and the Cheeger inequality for directed graphs. *Annals of Combinatorics*, 9(1):1–19, 2005.

Fan Chung and Mary Radcliffe. On the spectra of general random graphs. *Electronic Journal of Combinatorics*, 18(1):Paper 215, 14, 2011.

Fan RK Chung. Laplacians of graphs and Cheeger's inequalities. *Combinatorics, Paul Erdos is Eighty*, 2(157-172):13–2, 1996.

M. Cucuringu. Synchronization over Z_2 and community detection in multiplex networks with constraints. *Journal of Complex Networks*, 3:469–506, 2015.

Mihai Cucuringu, Ioannis Koutis, Sanjay Chawla, Gary Miller, and Richard Peng. Simple and scalable constrained clustering: a generalized spectral method. *Artificial Intelligence and Statistics Conference (AISTATS) 2016*, 51:445–454, 2016.

Mihai Cucuringu, Peter Davies, Aldo Glielmo, and Hemant Tyagi. SPONGE: A generalized eigenproblem for clustering signed networks. In *Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1088–1098. PMLR, 2019.

Mihai Cucuringu, Andrea Pizzoferrato, and Yves van Gennip. An MBO scheme for clustering and semi-supervised clustering of signed networks. *Communications in Mathematical Sciences*, 19(1):73–109, 2021.

Sanjoy Dasgupta. The hardness of k-means clustering. Technical Report CS2007-0890, University of California, San Diego, 2008.

Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Erik D. Demaine, Dotan Emanuel, Amos Fiat, and Nicole Immorlica. Correlation clustering in general weighted graphs. *Theor. Comput. Sci.*, 361(2):172–187, September 2006.

Tyler Derr, Yao Ma, and Jiliang Tang. Signed graph convolutional networks. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 929–934. IEEE, 2018.

Sergio M Focardi. Clustering economic and financial time series: Exploring the existence of stable correlation conditions. *The Intertek Group*, 2005.

André Fujita, Patricia Severino, Kaname Kojima, João Ricardo Sato, Alexandre Galvão Patriota, and Satoru Miyano. Functional clustering of time series gene expression data by Granger causality. *BMC systems biology*, 6(1):137, 2012.

Jean Gallier. Spectral theory of unsigned and signed graphs. applications to graph clustering: a survey. *CoRR*, abs / 1601.04692:1–122, 2016.

Jean H. Gallier. Notes on elementary spectral graph theory. applications to graph clustering using normalized cuts. *CoRR*, abs/1311.2492, 2013.

Gyeong-Gyun Ha, Jae Woo Lee, and Ashadun Nobi. Threshold network of a financial market using the p-value of correlation coefficients. *Journal of the Korean Physical Society*, 66 (12):1802–1808, 2015.

Cho-Jui Hsieh, Kai-Yang Chiang, and Inderjit S. Dhillon. Low-Rank Modeling of Signed Networks. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2012.

Roberto Imbuzeiro Oliveira. Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges. *arXiv e-prints*, art. arXiv:0911.0600, November 2009.

Jiashun Jin. Fast community detection by SCORE. *The Annals of Statistics*, 43(1):57 – 89, 2015.

Antony Joseph and Bin Yu. Impact of regularization on spectral clustering. *Annals of Statistics*, 44(4):1765–1791, 2016.

Jinhong Jung, Woojeong Jin, Lee Sael, and U Kang. Personalized ranking in signed networks using signed random walk with restart. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 973–978. IEEE, 2016.

A. Knyazev. Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method. *SIAM Journal on Scientific Computing*, 23(2): 517–541, 2001.

A. Kumar, Y. Sabharwal, and S. Sen. A simple linear time $(1 + \varepsilon)$-approximation algorithm for k-means clustering in any dimensions. In *45th Annual IEEE Symposium on Foundations of Computer Science*, pages 454–462, 2004.

Srijan Kumar, Francesca Spezzano, and VS Subrahmanian. Accurately detecting trolls in Slashdot Zoo via decluttering. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, pages 188–195. IEEE, 2014.

Srijan Kumar, Francesca Spezzano, V.S. Subrahmanian, and Christos Faloutsos. Edge weight prediction in weighted signed networks. In *ICDM*, 2016.

Jérôme Kunegis, Stephan Schmidt, Andreas Lommatzsch, Jürgen Lerner, Ernesto W. De Luca, and Sahin Albayrak. *Spectral Analysis of Signed Graphs for Clustering, Prediction and Visualization*, pages 559–570. SIAM, 2010.

Can M. Le, Elizaveta Levina, and Roman Vershynin. Sparse random graphs: regularization and concentration of the laplacian, 2015.

Can M. Le, Elizaveta Levina, and Roman Vershynin. Concentration and regularization of random graphs. *Random Structures & Algorithms*, 51(3):538–561, 2017.

James R Lee, Shayan Oveis Gharan, and Luca Trevisan. Multiway spectral partitioning and higher-order Cheeger inequalities. *Journal of the ACM (JACM)*, 61(6):1–30, 2014.

Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *Ann. Statist.*, 43(1):215–237, 02 2015.

J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. In *WWW*, pages 641–650, 2010.

Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford Large Network Dataset Collection. `http://snap.stanford.edu/data`, June 2014.

Ren-Cang Li. On perturbations of matrix pencils with real spectra. *Mathematics of Computation*, 62(205):231–265, 1994.

Xiaoming Li, Hui Fang, and Jie Zhang. Supervised user ranking in signed social networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 184–191, 2019.

Yu Li, Yuan Tian, Zhang Jiawei, and Yi Chang. Learning signed network embedding via graph attention. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:4772–4779, 04 2020.

Edo Liberty, Ram Sriharsha, and Maxim Sviridenko. An algorithm for online k-means clustering. In *2016 Proceedings of the Eighteenth Workshop on Algorithm Engineering and Experiments (ALENEX)*, pages 81–89. SIAM, 2016.

Meena Mahajan, Prajakta Nimbhorkar, and Kasturi Varadarajan. The planar k-means problem is NP-hard. *Theoretical Computer Science*, 442:13 – 21, 2012.

Philip Andrew Mansfield, Quan Wang, Carlton Downey, Li Wan, and Ignacio Lopez Moreno. Links: A high-dimensional online clustering method, 2018.

Pedro Mercado, Francesco Tudisco, and Matthias Hein. Clustering Signed Networks with the Geometric Mean of Laplacians. In *Advances in Neural Information Processing Systems 29*, pages 4421–4429. Curran Associates, Inc., 2016.

Pedro Mercado, Francesco Tudisco, and Matthias Hein. Spectral clustering of signed graphs via matrix power means. In *36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4526–4536, Long Beach, California, USA, 2019. PMLR.

Ekaterina Merkurjev, Justin Sunu, and Andrea L. Bertozzi. Graph MBO method for multi-class segmentation of hyperspectral stand-off detection video. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 689–693. IEEE, 2014.

Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, New York, NY, USA, 2005.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, page 849–856. MIT Press, 2001.

Nicos G Pavlidis, Vassilis P Plagianakos, Dimitris K Tasoulis, and Michael N Vrahatis. Financial forecasting through unsupervised clustering and neural networks. *Operational Research*, 6(2):103–127, 2006.

Tai Qin and Karl Rohe. Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3120–3128, 2013.

Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878 – 1915, 2011.

Ahmed Sameh and Zhanye Tong. The trace minimization method for the symmetric generalized eigenvalue problem. *Journal of Computational and Applied Mathematics*, 123(1): 155 – 175, 2000. Numerical Analysis 2000. Vol. III: Linear Algebra.

A. Singer. Angular synchronization by eigenvectors and semidefinite programming. *Appl. Comput. Harmon. Anal.*, 30(1):20–36, 2011.

Stephen M. Smith, Karla L. Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F. Beckmann, Thomas E. Nichols, Joseph D. Ramsey, and Mark W. Woolrich. Network modelling methods for FMRI. *NeuroImage*, 54(2):875 – 891, 2011.

Ludovic Stephan and Laurent Massoulié. Robustness of spectral methods for community detection. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2831–2860, Phoenix, USA, 2019. PMLR.

G.W. Stewart and Ji-guang Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

Jiliang Tang, Charu Aggarwal, and Huan Liu. Node classification in signed social networks. In *SDM*, 2016.

U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17:395–416, 2007.

Hermann Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, 71(4):441–479, 1912.

Han Xiao, Bruno Ordozgoiti, and Aristides Gionis. Searching for polarization in signed graphs: a local spectral approach. *Proceedings of The Web Conference 2020*, pages 362–372, 2020. doi: 10.1145/3366423.3380121.

Bowei Yan and Purnamrita Sarkar. Covariate regularized community detection in sparse graphs. *Journal of the American Statistical Association*, 0(0):1–12, 2020.

B. Yang, W. K. Cheung, and J. Liu. Community mining from signed social networks. *IEEE Trans Knowl Data Eng*, 19(10):1333–1348, 2007.

Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

Yilin Zhang and Karl Rohe. Understanding regularized spectral clustering via graph conductance. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Zhixin Zhou and Arash A. Amini. Analysis of spectral clustering algorithms for community detection: the general bipartite setting, 2018.

Hartmut Ziegler, Marco Jenny, Tino Gruse, and Daniel A Keim. Visual market sector analysis for financial time series data. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pages 83–90. IEEE, 2010.

Appendix A. Useful concentration inequalities

A.1 Chernoff bounds

Recall the following Chernoff bound for sums of independent Bernoulli random variables.

Theorem 52 ((Mitzenmacher and Upfal, 2005, Corollary 4.6)) *Let $X_1, \ldots, X_n$ be independent Bernoulli random variables with $\mathbb{P}[X_i = 1] = p_i$. Let $X = \sum_{i=1}^n X_i$ and $\mu = \mathbb{E}[X]$. For $\delta \in (0,1)$, it holds true that*

$$\mathbb{P}[|X - \mu| \geq \delta\mu] \leq 2\exp(-\mu\delta^2/3).$$

A.2 Spectral norm of random matrices

We will make use of the following result for bounding the spectral norm of symmetric matrices with independent, centered and bounded random variables.

Theorem 53 ((Bandeira and van Handel, 2016, Corollary 3.12, Remark 3.13)) *Let X be an $n \times n$ symmetric matrix whose entries X_{ij} $(i \leq j)$ are independent, centered random variables. There there exists for any $0 < \varepsilon \leq 1/2$ a universal constant c_ε such that for every $t \geq 0$,*

$$\mathbb{P}\left[\|X\| \geq (1+\varepsilon)2\sqrt{2}\widetilde{\sigma} + t\right] \leq n\exp\left(-\frac{t^2}{c_\varepsilon \widetilde{\sigma}_*^2}\right) \tag{91}$$

where

$$\widetilde{\sigma} := \max_i \sqrt{\sum_j \mathbb{E}[X_{ij}^2]}, \quad \widetilde{\sigma}_* := \max_{i,j} \|X_{ij}\|_\infty .$$

Note that it suffices to employ upper bound estimates on $\widetilde{\sigma}, \widetilde{\sigma}_*$ in (91). Indeed, if $\widetilde{\sigma} \leq \widetilde{\sigma}^{(u)}$ and $\widetilde{\sigma}_* \leq \widetilde{\sigma}_*^{(u)}$, then

$$\begin{aligned}\mathbb{P}\left[\|X\| \geq (1+\varepsilon)2\sqrt{2}\widetilde{\sigma}^{(u)} + t\right] \leq \mathbb{P}\left[\|X\| \geq (1+\varepsilon)2\sqrt{2}\widetilde{\sigma} + t\right] &\leq n\exp\left(-\frac{t^2}{c_\varepsilon \widetilde{\sigma}_*^2}\right) \\ &\leq n\exp\left(-\frac{t^2}{c_\varepsilon (\widetilde{\sigma}_*^{(u)})^2}\right).\end{aligned}$$

A.3 A graph decomposition result

The following graph decomposition result for inhomogeneous Erdős-Rényi graphs was established in (Le et al., 2017, Theorem 2.6).

Theorem 54 *(Le et al., 2017, Theorem 2.6) Let A be a directed adjacency matrix sampled from an inhomogeneous Erdős-Rényi $G(n, (p_{jj'})_{j,j'})$ model and let $d = n\max_{j,j'} p_{jj'}$. For any $r \geq 1$, with probability at least $1 - 3n^{-r}$, the set of edges $[n] \times [n]$ can be partitioned into three classes $\mathcal{N}, \mathcal{R}$ and $\mathcal{C}$, such that*

1. *the signed adjacency matrix concentrates on $\mathcal{N}$*
$$\|(A - \mathbb{E}A)_{\mathcal{N}}\| \leq Cr^{3/2}\sqrt{d},$$
2. *$\mathcal{R}$ (resp. $\mathcal{C}$) intersects at most n/d columns (resp. rows) of $[n] \times [n]$,*
3. *each row (resp. column) of $A_{\mathcal{R}}$ (resp. $A_{\mathcal{C}}$) have at most $32r$ non-zero entries.*

Appendix B. Matrix perturbation analysis

In this section, we recall several standard tools from matrix perturbation theory for studying the perturbation of the spectra of Hermitian matrices. The reader is referred to Stewart and Sun (1990) for a more comprehensive overview of this topic.

Let $A \in \mathbb{C}^{n\times n}$ be Hermitian with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ and corresponding eigenvectors $v_1, v_2, \ldots, v_n \in \mathbb{C}^n$. Let $\widetilde{A} = A + W$ be a perturbed version of A, with the perturbation matrix $W \in \mathbb{C}^{n\times n}$ being Hermitian. Let us denote the eigenvalues of $\widetilde{A}$ and W by $\widetilde{\lambda}_1 \geq \cdots \geq \widetilde{\lambda}_n$, and $\epsilon_1 \geq \epsilon_2 \geq \cdots \geq \epsilon_n$, respectively.

To begin with, one can quantify the perturbation of the eigenvalues of $\widetilde{A}$ with respect to the eigenvalues of A. Weyl's inequality (Weyl, 1912) is a very useful result in this regard.

Theorem 55 (Weyl's Inequality (Weyl, 1912)) *For each $i = 1, \ldots, n$, it holds that*

$$\lambda_i + \epsilon_n \leq \widetilde{\lambda}_i \leq \lambda_i + \epsilon_1. \tag{92}$$

In particular, this implies that $\widetilde{\lambda}_i \in [\lambda_i - \|W\|, \lambda_i + \|W\|]$.

One can also quantify the perturbation of the subspace spanned by eigenvectors of A, which was established by Davis and Kahan (1970). Before introducing the theorem, we need some definitions. Let $U, \widetilde{U} \in \mathbb{C}^{n\times k}$ (for $k \leq n$) have orthonormal columns respectively, and let $\sigma_1 \geq \cdots \geq \sigma_k$ denote the singular values of $U^*\widetilde{U}$. Also, let us denote $\mathcal{R}(U)$ to be the range space of the columns of U, and similarly for $\mathcal{R}(\widetilde{U})$. Then the k principal angles between $\mathcal{R}(U), \mathcal{R}(\widetilde{U})$ are defined as $\theta_i := \cos^{-1}(\sigma_i)$ for $1 \leq i \leq k$, with each $\theta_i \in [0, \pi/2]$. It is usual to define $k \times k$ diagonal matrices $\Theta(\mathcal{R}(U), \mathcal{R}(\widetilde{U})) := \mathrm{diag}(\theta_1, \ldots, \theta_k)$ and $\sin\Theta(\mathcal{R}(U), \mathcal{R}(\widetilde{U})) := \mathrm{diag}(\sin\theta_1, \ldots, \sin\theta_k)$. Denoting $|||\cdot|||$ to be any unitarily invariant norm (Frobenius, spectral, etc.), the following relation holds (see for eg., (Li, 1994, Lemma 2.1), (Stewart and Sun, 1990, Corollary I.5.4)).

$$|||\sin\Theta(\mathcal{R}(U), \mathcal{R}(\widetilde{U}))||| = |||(I - \widetilde{U}\widetilde{U}^*)U|||.$$

With the above notation in mind, we now introduce a version of the Davis-Kahan theorem taken from (Yu et al., 2015, Theorem 1) (see also (Stewart and Sun, 1990, Theorem V.3.6)).

Theorem 56 (Davis-Kahan) *Fix $1 \leq r \leq s \leq n$, let $d = s - r + 1$, and let $U = (u_r, u_{r+1}, \ldots, u_s) \in \mathbb{C}^{n\times d}$ and $\widetilde{U} = (\widetilde{u}_r, \widetilde{u}_{r+1}, \ldots, \widetilde{u}_s) \in \mathbb{C}^{n\times d}$. Write*

$$\delta = \inf\left\{\left|\hat{\lambda} - \lambda\right| : \lambda \in [\lambda_s, \lambda_r], \hat{\lambda} \in (-\infty, \widetilde{\lambda}_{s+1}] \cup [\widetilde{\lambda}_{r-1}, \infty)\right\}$$

where we define $\widetilde{\lambda}_0 = \infty$ and $\widetilde{\lambda}_{n+1} = -\infty$ and assume that $\delta > 0$. Then

$$|||\sin\Theta(\mathcal{R}(U), \mathcal{R}(\widetilde{U}))||| = |||(I - \widetilde{U}\widetilde{U}^*)U||| \leq \frac{|||W|||}{\delta}.$$

For instance, if $r = s = j$, then by using the spectral norm $\|\cdot\|$, we obtain

$$\sin\Theta(\mathcal{R}(\widetilde{v}_j), \mathcal{R}(v_j)) = \left\|(I - v_j v_j^*)\widetilde{v}_j\right\| \leq \frac{\|W\|}{\min\left\{\left|\widetilde{\lambda}_{j-1} - \lambda_j\right|, \left|\widetilde{\lambda}_{j+1} - \lambda_j\right|\right\}}. \tag{93}$$

Finally, we recall the following standard result which states that given any pair of k-dimensional subspaces with orthonormal basis matrices $U, \widetilde{U} \in \mathbb{R}^{n\times k}$, there exists an alignment of $U, \widetilde{U}$ with the error after alignment bounded by the distance between the subspaces. We provide the proof for completeness.

Proposition 57 *Let $U, \widetilde{U} \in \mathbb{R}^{n\times k}$ respectively consist of orthonormal vectors. Then there exists a $k \times k$ rotation matrix O such that*

$$\left\|\widetilde{U} - UO\right\| \leq 2\left\|(I - UU^T)\widetilde{U}\right\|.$$

Proof Write the SVD as $U^T\widetilde{U} = V\Sigma(V')^T$, where we recall that the ith largest singular value $\sigma_i = \cos\theta_i$ with $\theta_i \in [0, \pi/2]$ denoting the principal angles between $\mathcal{R}(U)$ and $\mathcal{R}(\widetilde{U})$. Choosing $O = V(V')^T$, we then obtain

$$\begin{aligned}
\left\|\widetilde{U} - UV(V')^T\right\| &\leq \left\|\widetilde{U} - UU^T\widetilde{U}\right\| + \left\|UU^T\widetilde{U} - UV(V')^T\right\| \\
&= \left\|(I - UU^T)\widetilde{U}\right\| + \left\|U^T\widetilde{U} - V(V')^T\right\| \\
&= \left\|(I - UU^T)\widetilde{U}\right\| + \|I - \Sigma\| \\
&\leq 2\left\|(I - UU^T)\widetilde{U}\right\|,
\end{aligned}$$

where the last inequality follows from the fact $\|I - \Sigma\| = 1 - \cos\theta_k \leq \sin\theta_k$. ■

Appendix C. Summary of main technical tools

This section collects certain technical results that were used in the course of proving our main results.

Proposition 58 ((Bhatia, 1996, Theorem X.1.1)) *For matrices $A, B \succ 0$,*

$$\left\|A^{1/2} - B^{1/2}\right\| \leq ||A - B||^{1/2}$$

holds as $(\cdot)^{1/2}$ is operator monotone.

Proposition 59 *For symmetric matrices A^+, A^-, B^+ and B^- where $A^-, B^- \succ 0$, the following holds.*

$$\begin{aligned}
&\left\|(A^-)^{-1/2}A^+(A^-)^{-1/2} - (B^-)^{-1/2}B^+(B^-)^{-1/2}\right\| \\
&\leq \left\|(A^-)^{-1}\right\|\left\|A^+\right\|\left(\left\|I - (B^-)^{-1/2}(A^-)^{1/2}\right\|^2 + 2\left\|I - (B^-)^{-1/2}(A^-)^{1/2}\right\|\right) \\
&\qquad + \left\|(B^-)^{-1}\right\|\left\|A^+ - B^+\right\| \\
&\leq \left\|(A^-)^{-1}\right\|\left\|A^+\right\|\left(\left\|(B^-)^{-1}\right\|\left\|(B^-) - (A^-)\right\| + 2\left\|(B^-)^{-1/2}\right\|\left\|(B^-) - (A^-)\right\|^{1/2}\right) \\
&\qquad + \left\|(B^-)^{-1}\right\|\left\|A^+ - B^+\right\|.
\end{aligned}$$

Proof

$$\begin{aligned}
&\left\|(A^-)^{-1/2}A^+(A^-)^{-1/2} - (B^-)^{-1/2}B^+(B^-)^{-1/2}\right\| \\
&= \Big\|(A^-)^{-1/2}A^+(A^-)^{-1/2} - (B^-)^{-1/2}A^+(B^-)^{-1/2} \\
&\qquad\qquad + \ (B^-)^{-1/2}A^+(B^-)^{-1/2} - (B^-)^{-1/2}B^+(B^-)^{-1/2}\Big\| \\
&\leq \left\|(B^-)^{-1/2}(A^+ - B^+)(B^-)^{-1/2}\right\| + \left\|(A^-)^{-1/2}A^+(A^-)^{-1/2} - (B^-)^{-1/2}A^+(B^-)^{-1/2}\right\| .
\end{aligned}$$

Now, we bound the two terms separately. The first term is easy to bound.

$$\begin{aligned}
\left\|(B^-)^{-1/2}(A^+ - B^+)(B^-)^{-1/2}\right\| &\leq \left\|(B^-)^{-1/2}\right\| \left\|A^+ - B^+\right\| \left\|(B^-)^{-1/2}\right\| \\
&= \left\|(B^-)^{-1}\right\| \left\|A^+ - B^+\right\| . \qquad (94)
\end{aligned}$$

To bound the second term, we do the following manipulations,

$$\begin{aligned}
&\left\|(A^-)^{-1/2}A^+(A^-)^{-1/2} - (B^-)^{-1/2}A^+(B^-)^{-1/2}\right\| \\
&= \left\|(A^-)^{-1/2}A^+(A^-)^{-1/2} - (A^-)^{-1/2}(A^-)^{1/2}(B^-)^{-1/2}A^+(B^-)^{-1/2}(A^-)^{1/2}(A^-)^{-1/2}\right\| \\
&= \left\|(A^-)^{-1/2}\left(A^+ - (A^-)^{1/2}(B^-)^{-1/2}A^+(B^-)^{-1/2}(A^-)^{1/2}\right)(A^-)^{-1/2}\right\| \\
&= \left\|(A^-)^{-1/2}\left(A^+ - \left((A^-)^{1/2}(B^-)^{-1/2} - I + I\right)A^+\left((B^-)^{-1/2}(A^-)^{1/2} - I + I\right)\right)(A^-)^{-1/2}\right\| \\
&= \Big\|(A^-)^{\frac{-1}{2}}\big(\ ((A^-)^{\frac{1}{2}}(B^-)^{-\frac{1}{2}} - I)A^+((B^-)^{-\frac{1}{2}}(A^-)^{\frac{1}{2}} - I) + A^+((B^-)^{-\frac{1}{2}}(A^-)^{\frac{1}{2}} - I) \\
&\qquad\qquad + \ ((A^-)^{\frac{1}{2}}(B^-)^{-\frac{1}{2}} - I)A^+\ \big)(A^-)^{-\frac{1}{2}}\Big\| \\
&\leq \left\|(A^-)^{-1}\right\| \left\|A^+\right\| \left(\left\|I - (B^-)^{-1/2}(A^-)^{1/2}\right\|^2 + 2\left\|I - (B^-)^{-1/2}(A^-)^{1/2}\right\|\right) . \qquad (95)
\end{aligned}$$

The first inequality of the lemma follows by adding (95) and (94). To see the second inequality of the lemma, observe that,

$$\begin{aligned}
\left\|I - (B^-)^{-1/2}(A^-)^{1/2}\right\| &= \left\|(B^-)^{-1/2}((B^-)^{1/2} - (A^-)^{1/2})\right\| \\
&\leq \left\|(B^-)^{-1/2}\right\| \left\|(B^-)^{1/2} - (A^-)^{1/2}\right\| \\
&\leq \left\|(B^-)^{-1/2}\right\| \left\|B^- - A^-\right\|^{1/2} \quad \text{(using Proposition 58)} . \qquad (96)
\end{aligned}$$

The second inequality of the lemma follows by substituting (96) in the first inequality of the lemma.

■

Appendix D. Proofs from Section 4

Lemma 60 (Expression for C_e^+ & C_e^-)

$$C_e^+ = -p\eta\frac{n}{d^+}\chi_1\chi_1^\top + \left(1 + \tau^- + \frac{p}{d^+}\left(1 - \eta - \frac{n}{k}(1 - 2\eta)\right)\right) I ,$$

$$C_e^- = -p(1-\eta)\frac{n}{d^-}\chi_1\chi_1^\top + \left(1+\tau^+ + \frac{p}{d^-}\left(\eta + \frac{n}{k}(1-2\eta)\right)\right) I\,.$$

It follows that can be written as $C_e^+ = R\Sigma^+R^\top$ *and* $C_e^- = R\Sigma^-R^\top$*, where* R *is a rotation matrix, and*

$$\Sigma^+ = \begin{bmatrix} \left(1+\tau^- + \frac{p}{d^+}\left(1-\eta-n\left(\eta+\frac{1-2\eta}{k}\right)\right)\right) & \\ & \left(1+\tau^- + \frac{p}{d^+}\left(1-\eta-n\left(\frac{1-2\eta}{k}\right)\right)\right) I_{k-1} \end{bmatrix},$$

$$\Sigma^- = \begin{bmatrix} \left(1+\tau^+ + \frac{p}{d^-}\left(\eta-n\left(1-\eta-\frac{1-2\eta}{k}\right)\right)\right) & \\ & \left(1+\tau^+ + \frac{p}{d^-}\left(\eta+n\left(\frac{1-2\eta}{k}\right)\right)\right) I_{k-1} \end{bmatrix}.$$

The above lemma shows that we know the spectrum of $(C^-)^{-1/2}C^+(C^-)^{-1/2}$ exactly, in the case of equal-sized clusters.

Proof [Proof of Lemma 22] From (17) it follows that,

$$\lambda_{\max}(C^+) \le \max_{i\in[k]}\left(1+\tau^- + \frac{p}{d_i^+}(1-\eta-n_i(1-2\eta))\right).$$

The maximum is achieved for the smallest sized cluster. This shows the proof for (33).

The proof of (34) follows from the fact that in (24) we had decomposed the matrix $\overline{L_{sym}^-} + \tau^+ I$ as a block-diagonal matrix, with block of $C^-, \alpha_1^- I_{n_1-1}, \ldots, \alpha_k^- I_{n_k-1}$. Since $\overline{L_{sym}^-}$ is a symmetric Laplacian, we know that $\lambda_{\min}(\overline{L_{sym}^-} + \tau^+ I) = \tau^+$. Also, $\alpha_i^- > \tau^+$ for $i \in [k]$. Thus the equation follows. ■

Appendix E. Spectrum of Signed Laplacians

This section extends some classical results for the unsigned Laplacian to the symmetric Signed Laplacian and the regularized Laplacian.

Lemma 61 *For all* $x \in \mathbb{R}^n$*,*

$$x^T\overline{L_{sym}}x = \frac{1}{2}\sum_{j,j'}|A_{jj'}|\left(\frac{x_j}{\sqrt{d_j}} - sgn(A_{jj'})\frac{x_{j'}}{\sqrt{d_{j'}}}\right)^2 \tag{97}$$

Moreover, the eigenvalues of $\overline{L_{sym}}$ *and* L_γ *are in the interval* $[0,2]$*.*

Proof Equation (97) is adapted from Proposition 5.2 from Gallier (2016) and is obtained by replacing x by $\overline{D}^{-1/2}x$. The second part of the lemma comes from the fact that $(a\pm b)^2 \le 2(a^2+b^2)$. In fact, for $x \in \mathbb{R}^n$ such that $\|x\| = 1$, we have

$$\begin{aligned} x^T\overline{L_{sym}}x &\le \sum_{j,j'}|A_{jj'}|\left(\frac{x_j^2}{d_j} + \frac{x_{j'}^2}{d_{j'}}\right) \\ &= 2\sum_{j,j'}|A_{jj'}|\frac{x_j^2}{d_j} = 2\sum_j x_j^2 = 2. \end{aligned}$$

Similarly, we have

$$\begin{aligned} x^T L_\gamma x &\leq \sum_{j,j'} |(A_\gamma)jj'| \left(\frac{x_j^2}{\overline{D}_{jj} + \gamma} + \frac{x_{j'}^2}{\overline{D}_{j'j'} + \gamma} \right) \\ &\leq 2 \sum_{j,j'} (|A_{jj'}| + \frac{\gamma}{n}) \frac{x_j^2}{\overline{D}_{jj} + \gamma} \\ &= 2 \sum_j \frac{(\overline{D}_{jj} + \gamma) x_j^2}{\overline{D}_{jj} + \gamma} = 2. \end{aligned}$$

Moreover $\overline{L_{sym}}$ and L_γ are positive semi-definite, thus we can conclude that their eigenvalues are between 0 and 2. ■

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	**Regularized spectral methods for clustering signed networks**
Publication Status	X Published □ Accepted for Publication □Submitted for Publication □Unpublished and unsubmitted work written in a manuscript style
Publication Details	Cucuringu, M., Singh, A. V., Sulem, D., & Tyagi, H. (2021). Regularized spectral methods for clustering signed networks. *J. Mach. Learn. Res.*, *22*, 264-1.

Student Confirmation

Student Name:	Deborah Sulem		
Contribution to the Paper	I have analyzed the theoretic properties of the spectral clustering algorithm based on the symmetric signed Laplacian, as well as its regularized version. I have implemented the regularized algorithms and conducted the numerical experiments.		
Signature *Deborah Sulem*		Date	08/11/2022

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Professor Mihai Cucuringu		
Supervisor comments I confirm that Deborah has made a substantial contribution in line with her description above.		
Signature 	Date	9 Nov 2022

This completed form should be included in the thesis, at the end of the relevant chapter.

5 | Graph similarity learning for detecting change-points in dynamic networks

Submitted to the Machine Learning journal.

Graph similarity learning for change-point detection in dynamic networks

Déborah Sulem[1*], Henry Kenlay[2], Mihai Cucuringu[1,3,4] and Xiaowen Dong[2]

[1]Department of Statistics, University of Oxford.

[2]Department of Engineering Science,University of Oxford.

[3]Mathematical Institute, University of Oxford.

[4]The Alan Turing Institute, London, UK.

*Corresponding author(s). E-mail(s): deborah.sulem@stats.ox.ac.uk;

Contributing authors: kenlay@robots.ox.ac.uk; mihai.cucuringu@stats.ox.ac.uk; xdong@robots.ox.ac.uk;

Abstract

Dynamic networks are ubiquitous for modelling sequential graph-structured data, e.g., brain connectome, population flows and messages exchanges. In this work, we consider dynamic networks that are temporal sequences of graph snapshots, and aim at detecting abrupt changes in their structure. This task is often termed *network change-point detection* and has numerous applications, such as fraud detection and physical motion monitoring. Leveraging a graph neural network model, we design a method to perform online network change-point detection that can adapt to the specific network domain and localise changes with no delay. The main novelty of our method is to use a siamese graph neural network architecture for learning a data-driven graph similarity function, which allows to effectively compare the current graph and its recent history. Importantly, our method does not require prior knowledge on the network generative distribution and is agnostic to the type of change-points; moreover, it can be applied to a large variety of networks, that include for instance edge

weights and node attributes. We show on synthetic and real data that our method enjoys a number of benefits: it is able to learn an adequate graph similarity function for performing online network change-point detection in diverse types of change-point settings, and requires a shorter data history to detect changes than most existing state-of-the-art baselines.

Keywords: dynamic networks, change-point detection, graph similarity learning, siamese graph neural network.

1 Introduction

The study of dynamic - or temporal, evolutionary, time-varying - networks has become very popular in the last decade, with the increasing amount of sequential data collected from structured and evolving systems, e.g. online communication platforms (Kumar et al, 2019), co-voting networks (Wilson et al, 2019), and fMRI data (Cribben and Yu, 2017). In fact, adding a time component to graph-structured data leads to a richer representation and allows more powerful analysis (Skarding et al, 2021). This is particularly important when the network is governed by a non-stationary underlying process, which dynamics undergo abrupt switches or breaks. For instance, social networks appear with different characteristics at different times and can be dependent on global temporal events such as terrorist attacks (Bourqui et al, 2009), thus providing strong motivation for incorporating a temporal dimension in the analysis. Detecting such structural breaks is a common task in diverse applications, from brain connectivity state segmentation (Ondrus et al, 2021) to phase discovery in financial correlation networks (Barnett and Onnela, 2016). Moreover, several real-world dynamic networks are structured around functional groups or densely connected communities (see for instance Rossetti and Cazabet (2018) for a review on *community discovery* in dynamic networks). The evolution of such networks over time has been often measured by the changes in these substructures - sometimes called *community life-cycle* - e.g. growth, decay, merges, splits, etc.

For multivariate time series, change-point detection is a task that has been widely studied in various settings (e.g., nonparametric (Zou et al, 2014), high-dimensional (Wang and Samworth, 2018), and online (Wang et al, 2022)). The equivalent task for dynamic networks is often termed *network change-point detection* (NCPD) and has recently become a popular problem with numerous successful applications in finance (Barnett and Onnela, 2016), neuroscience (Ofori-Boateng et al, 2019), and transport networks (Yu et al, 2021). Depending on the type of problem at hand, dynamic networks have been represented in multiple ways, e.g., with contact sequences, interval graphs, graph snapshots (see Holme and Saramäki (2012) for a precise review of concepts, models, and applications). In this work, we will consider the discrete representation of time-varying networks or *snapshot networks*: we denote a dynamic network $\mathcal{N}_I = \{G_t\}_{t \in I}$ to be a sequence of graph snapshots, where I is an ordered set, chosen as $\mathbb{N}_{>0}$ for simplicity, and each $G_t, t \in I$, is a (static) graph. We note that G_t is a graph that can be directed, have edge weights or node attributes. We define a change-point for the network $\mathcal{N}$ as a timestamp $t \in \mathbb{N}_{>0}$ such that the generative distribution of the graphs before t, $(G_1, \ldots, G_{t-2}, G_{t-1})$ is different from the one of graphs observed from t, $(G_t, G_{t+1}, \ldots)$. More broadly, a change-point for a dynamic network sequence is defined as a timestamp t where a significant shift or deviation can be observed between G_t and the preceding graph snapshots.

In general, a dynamic network may contain multiple change-points and the tasks of detecting and localising the latter therefore correspond to partitioning the observation window $[1, T], T > 0$ into K segments $\mathcal{T}_i = [\tau_{i-1}, \tau_i), 1 \leq i \leq K$ with $\tau_0 = 1$, $\tau_K = T$ and $1 < \tau_1 < \cdots < \tau_{K-1} < T$, such that for each $i \in [K-1]$, the generative distribution of the graph snapshots in $\mathcal{T}_i \cap I$ is the same, while it is different from the distribution generating the graphs in $\mathcal{T}_{i-1} \cap I$ and $\mathcal{T}_{i+1} \cap I$. The set of timestamps $(\tau_i)_{i=1,\ldots,K-1}$ then corresponds to the set of change-points. Intuitively, each temporal segment $[\tau_i, \tau_{i+1})$ can be associated with a state of the underlying process, and each change-point τ_i can be interpreted as a response of the system to an external event. Therefore, NCPD shares some similarity with the task of anomaly detection in temporal graphs Enikeeva and Klopp (2021). In an online setting, one aims to detect such change-points while the graph snapshots are collected, and with minimal detection delay, while in an offline setting, such analysis is conducted *a posteriori* on the whole data sequence. For particular graph generative models, the feasibility of the NCPD task and minimax rates of estimation have been analysed in dynamic random graph models, e.g., Bernoulli networks (Padilla et al, 2019; Enikeeva and Klopp, 2021; Yu et al, 2021; Wang and Samworth, 2018), graphon models (Zhao et al, 2019), stochastic block models (Wilson et al, 2019; Wang et al, 2013) and generalized hierarchical random graphs (Peel and Clauset, 2015). However, most real-world dynamic networks have heterogeneous properties, e.g. sparsity, edge weights, node attributes or nonlinear dynamics (Li et al, 2017) - and neither their generative distribution nor the type of change that can happen are known in advance.

Many existing methods for NCPD measure the discrepancy between two subsets of graphs, and rely on a graph similarity function, kernel or distance for pairwise graph comparisons (Chu and Chen, 2018; Cribben and Yu, 2017; Zhao et al, 2019; Gretton et al, 2008). However, it is often difficult to choose *a priori* an appropriate measure of similarity (or dissimilarity) that can integrate all the network characteristics, while being agnostic to the generating mechanism or type of change-point. Consequently, without any domain knowledge, this choice is often arbitrary, and result in poor performances (Chu and Chen, 2018; Enikeeva and Klopp, 2021; Kriege et al, 2020). Moreover, most online NCPD methods require finely tuning several hyperparameters, such as detection thresholds (Yu et al, 2021) and window sizes (Huang et al, 2020). To address these challenges, we propose a change-point agnostic and end-to-end method for online NCPD that in particular includes learning a data-driven graph similarity function. Our method is therefore adaptive to the network distribution and different types of change-points; in particular, it can easily incorporate general graph features such as node attributes, edge weights or attributes, and can adapt to sparse settings. In summary, our contributions are the following:

- We propose a graph similarity learning model based on a siamese graph neural network able to handle any available node attributes, and demonstrate how it can be leveraged for the online NCPD problem with an

adequate training procedure. In particular, our learnt similarity function is sensitive to both local and global displacements in the graph structure, and can effectively be employed in the context of change-point (and anomaly) detection in temporal networks.

- We use an efficient online NCPD statistic with a short-term history of the graph snapshots that avoids detection delays and requires little additional hyperparameter tuning.
- We empirically demonstrate the advantages of our method on synthetic networks with diverse types of change-points, as well as on two challenging real-world data sets. We notably design a self-supervised training procedure for data without ground-truth labelling of change-points.

Paper outline.

In Section 2, we succinctly review existing work on NCPD and present our general setup and methodology in Section 3. In Section 4, we evaluate our method on synthetic and real-world data sets and compare to several existing NCPD baseline methods. Finally, we conclude in Section 5 with a summary of our results and discuss possible future developments.

2 Related works

The study of dynamic networks, and in particular NCPD, is a relatively recent area of research that has largely incorporated principles from change-point detection in time series, especially in high-dimensional settings. Some NCPD methods estimate the parameters of a network model, e.g., the generalised hierarchical random graph (Peel and Clauset, 2015), a stochastic block model (De Ridder et al, 2016) or the preferential attachment model (Bhamidi et al, 2018), and conduct hypothesis tests to detect changes in the estimated parameters. Other methods maximize a penalized likelihood function, e.g., based on a non-homogeneous Poisson point process model (Corneli et al, 2018) or a dynamic stochastic block model (Wilson et al, 2019; Bhattacharjee et al, 2020). However, for real-world networks, the assumption on a particular model can sometimes be too restrictive.

Several model-agnostic methods for NCPD extract features from the graph snapshots, e.g., the degree distribution (Miller and Mokryn, 2020) or the joint distribution of a set of edges (Wang et al, 2017), and use classical discrepancy measures to quantify the amount of change. Other methods relying on pairwise comparison of graphs use a graph similarity or pseudo-distance, such as the DeltaCon metric (Koutra et al, 2016), the Hamming distance and the Jaccard distances (Donnat and Holmes, 2018), the Frobenius and maximum norms (Barnett and Onnela, 2016), spectral distances based on the Laplacian (Huang et al, 2020; Cribben and Yu, 2017; Hewapathirana et al, 2020), ℓ_2 or ℓ_∞ norms (Zhao et al, 2019) or a graph kernel (Desobry et al, 2005; Gretton et al, 2008; Harchaoui et al, 2009). Nevertheless, these graph metrics suffer from intrinsic

limitations; e.g., the Hamming distance is sensitive to the graph density and the Jaccard distance treats all edges uniformly (Donnat and Holmes, 2018). Furthermore, it has been previously underlined that the choice of graph distance can significantly affect a method's results (Barnett and Onnela, 2016), and therefore requires *a-priori* knowledge or assumption on the network distribution.

One widely popular statistic in change-point detection problems is the cumulative sums (CUSUM) statistic, which has been used in different time series contexts, e.g., in the offline and high-dimensional setting (in combination with the network binary segmentation algorithm) (Wang et al, 2022), and more recently, in the online setting (Wang et al, 2022). Several NCPD methods have adapted this efficient statistic to dynamic networks, e.g., for sparse graphs (Wang and Samworth, 2018), graphs with missing links (Dubey et al, 2021; Enikeeva and Klopp, 2021), in offline (Padilla et al, 2019) and online (Yu et al, 2021) settings, and proved that minimax rates of estimation can be obtained for the overall false alarm probability and the detection delay. However, computing the CUSUM statistic necessitates a "forward" window to detect a change at a given timestamp, and methods based on this statistic often require to tune several hyperparameters (e.g., one or several detection thresholds).

In addition to the aforementioned limitations, most previously cited methods do not provide a principled way to incorporate node attributes or even edge weights. Interestingly, to the best of our knowledge, no prior work has ever considered graph neural networks (GNNs) for the NCPD problem, despite the fact that such architectures can easily handle different types of networks (e.g., signed (Derr et al, 2018) or directed (Huang et al, 2019)), and in particular, can inherently account for any available node attributes (Kipf and Welling, 2016). In dynamic network modelling, graph convolutional recurrent networks (Seo et al, 2018) and dynamic graph convolutional networks (Manessi et al, 2020) were introduced for predicting graph-structured sequences. In the dynamic link prediction task, methods that learn representations of dynamic networks have been proposed, using deep temporal point processes (Trivedi et al, 2019), joint attention mechanisms on nodes neighborhoods and temporal domain (Sankar et al, 2020), memory feature vectors in message-passing architectures (Rossi et al, 2020) or recurrent neural networks (Zhang et al, 2021; Kumar et al, 2019). For anomalous edge detection in dynamic graphs, (Cai et al, 2021) process subgraphs around the target edges through convolution and sort pooling operations, and gated recurrent units. Moreover, one prior work has incorporated GNN layers in a method for change-point detection, but has done so in the context of multivariate time series (Zhang et al, 2020). However, in this method, the GNN encodes the cross-covariances between the time series' dimensions in the spatial layers, and is one part of a complex neural network architecture (the temporal dependencies being encoded by recurrent neural network layers).

Furthermore, while GNNs have proved to effectively learn representations of graphs, they can also be leveraged to learn graph similarity functions in a

data-driven way and for particular tasks in a end-to-end fashion. This now popular problem is called *graph similarity learning (GSL)* (Ma et al, 2021). One common type of model for this task is siamese networks (Koch, 2015), e.g., siamese graph neural networks (Ma et al, 2019)) or graph matching networks (Li et al, 2019; Ling et al, 2021). These architectures allow to learn flexible and adaptive similarity functions and have been successfully applied to several tasks and graph domains, e.g. classification of brain networks (Ma et al, 2019; Liu et al, 2019; Ktena et al, 2017), image classification (Mensink et al, 2012), and detection of vulnerabilities in software systems (Li et al, 2019). In this work, we will leverage such GSL models for the online NCPD task, which avoids the need for choosing *a-priori* a particular graph distance, kernel or embedding.

3 General setup and framework

In this section, we describe our general set-up and NCPD method based on a graph similarity learning model. We will first present our network change-point statistic in Section 3.1, leveraging a similarity function learnt by a GSL model described in Section 3.2, through an adequate training and validation procedures (see Section 3.3). Before presenting our methodology, we introduce some useful notation.

Notation.

We denote $G = (\boldsymbol{A}, \boldsymbol{X}) \in \mathbb{G}$ a graph with $n \geq 1$ nodes denoted by $\{u_1, \ldots, u_n\}$, adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ and node attributes (or features) matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d} \cup \{\emptyset\}$, with $d \geq 1$ attributes. We say that the graph is *attributed* if $\boldsymbol{X} \neq \emptyset$, and *unattributed* otherwise. If $\boldsymbol{A} \in \mathbb{R}_{\geq 0}^{n \times n}$, we also say that the graph is unsigned. We denote $\mathcal{N}_T = \{G_t\}_{1 \leq t \leq T}$ a dynamic network with $T \geq 1$ snapshots, where each graph G_t has the same set of nodes, with the same order in $\boldsymbol{A}$ and $\boldsymbol{E}$. Let $\boldsymbol{I}_n$ and $\mathbb{1}_n$ be respectively the $n \times n$ identity matrix and the all-one vector of size n. For a matrix $\boldsymbol{M}$, we denote $\boldsymbol{M}_{ij}$ an entry, $\boldsymbol{M}_{i:}$ its i-th row and $\boldsymbol{M}_{:j}$ its j-th column. We also denote $\|\boldsymbol{M}\|_F$ and $\|\boldsymbol{M}\|$ respectively the Frobenius norm and operator norm (i.e., the largest singular value). For a vector $\vec{v}$, we denote $\|\vec{v}\|$ its Euclidean norm. For any positive integer J, let $[J]$ denote the set $\{1, 2, \ldots, J\}$.

3.1 Graph similarity function for network change point detection

We consider a single dynamic network $\mathcal{N}_T = \{G_i\}_{1 \leq t \leq T}$ with an unknown number of change-points $1 < \tau_1 < \cdots < \tau_K < T$, $K \geq 1$, such that, for any $k \in [K]$ we have

$$G_i \overset{i.i.d.}{\sim} \mathcal{G}_{k-1}, \quad \tau_{k-1} \leq i < \tau_k, \tag{1}$$

(2)

with $\tau_0 = 1$ and $(\mathcal{G}_0, \dots, \mathcal{G}_K)$ distinct graph generating distributions. We assume that $\forall k \geq 1, \tau_k - \tau_{k-1} \geq L_0$, with $L_0 > 0$ a known lower bound of the minimal spacing between two consecutive change-points. We recall that in our setting, the set of nodes in each graph snapshot G_t is fixed and its ordering is kept unchanged along the sequence. We note that in general, the i.i.d. assumption in (1) is a strong hypothesis on the dynamic network's generative distribution. In practice, this assumption may not be verified since consecutive snapshots of real-world dynamic networks are often correlated. However, this is a standard setting for deriving theoretical results on NCPD methods in dynamic random graph models (Yu et al, 2021; Bhattacharjee et al, 2020; Zhao et al, 2019; Wang and Samworth, 2018). In this work, we consider this set-up for clarity of exposition, nevertheless our method accounts for the possibly existing correlations between the snapshots in the design of the sampling scheme (see Section 3.3).

Assume for now that we have a graph similarity function $s : \mathbb{G} \times \mathbb{G} \to [0, 1]$ that we can use as a binary classifier classifier of graph distribution. In other words, s is such that for any $G_{t_1} \sim \mathcal{G}_{i_1}, G_{t_2} \sim \mathcal{G}_{i_2}$, $s(G_{t_1}, G_{t_2}) > 0.5$ if $\mathcal{G}_{i_1} = \mathcal{G}_{i_2}$ and $s(G_{t_1}, G_{t_2}) \leq 0.5$ otherwise. One can then detect change-points in $\mathcal{N}_T$ by monitoring the following average similarity statistic

$$Z_t(s, L) = \frac{1}{L} \sum_{i=1}^{L} s(G_t, G_{t-i}), \quad t \geq L, \tag{3}$$

where $L < L_0$ is a hyperparameter that controls the length of the past window, and declare a change-point at any timestamp t such that

$$\begin{aligned} &Z_{t'}(s, L) > 0.5, \quad t - L \leq t' < t, \\ &Z_t(s, L) \leq 0.5. \end{aligned} \tag{4}$$

This general method can be applied to recover an arbitrary number of change-points in the dynamic network in an online setting and without any detection delay, i.e., as soon as the data is collected. In practice, one can choose a graph similarity function or kernel $s(\cdot, \cdot)$ and a detection threshold θ, e.g., using a validation criterion (Ranshous et al, 2015) or a significance test procedure using stationary bootstrap (Cribben and Yu, 2017), and declare a change-point (or an anomaly) in the dynamic network whenever $Z_t(s, L) > \theta$. Note that the properties of this method heavily depend on the chosen similarity function and its discriminative power.

Our NCPD method consists of using the statistic $Z_t(s, L)$ and the detection rule (4), together with a data-driven graph similarity function $s(\cdot, \cdot)$ learnt by a s-GNN model, which we describe in the next section.

Remark 1. Our method can also be employed in an offline setting, where one aims at localising changes in a dynamic network after the whole sequence has

been collected, with a slight change of the detection rule. For instance, for a dynamic network with a single change-point, one can localise the latter at $\hat{\tau}$, such that

$$\begin{aligned} \hat{\tau} &= \arg\min_{t\in[L,T]} Z_t(s,L), \\ \text{or} \quad \hat{\tau} &= \arg\max_{t\in[L+1,T]} |Z_t(s,L) - Z_{t-1}(s,L)|. \end{aligned} \tag{5}$$

Additionally, our method could be adapted to a setting where a small detection delay (e.g., of order L) may be tolerated. In this case, we could replace (3) by a more robust change-point statistic that also uses a future (or forward) window, e.g., $(G_t, G_{t+1}, \dots G_{t+L})$. For instance, we could use a two-sample test statistic on the two sets of graphs $(G_{t-1}, \dots, G_{t-L})$ and $(G_t, \dots, G_{t+L})$ such as the maximum kernel Fisher discriminant ratio (Harchaoui et al, 2009) or the maximum mean discrepancy (MMD) (Gretton et al, 2008), for which an unbiased estimate is given by

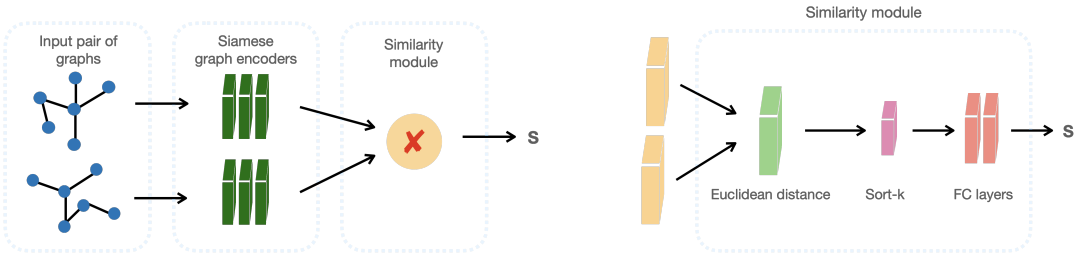
$$Z_t^{MMD} = \sqrt{\frac{1}{L(L+1)} \sum_{i,j=1}^{L+1} \big(s(G_{t-i}, G_{t-j}) + s(G_{t-1+i}, G_{t-1+j}) - s(G_{t-i}, G_{t-1+j})\big)}.$$

Note that this estimate would correspond to the empirical MMD measure between two sets of graphs mapped into a reproducing kernel Hilbert space if the function $s(\cdot,\cdot)$ was a graph kernel function (Gretton et al, 2008).

3.2 Graph similarity learning via siamese graph neural networks

Siamese graph neural networks (s-GNN) are architectures designed to compare pairs of graphs, e.g., for learning a graph similarity function or distance. They can notably be used in graph classification and graph matching tasks Ma et al (2019); Ktena et al (2017) in both supervised and unsupervised settings. More precisely, a general s-GNN takes as input a pair (G_1, G_2), embeds G_1 and G_2 with the same graph encoder (or equivalently, two *siamese* encoders that share the same weights), then combines the embeddings in a symmetric similarity module. The variability of s-GNN architectures mainly lies in the design of these two modules (see for instance Ktena et al (2017); Ma et al (2019); Ling et al (2021)).

In our NCPD method, we propose a s-GNN architecture summarized in Figure 1, for learning a similarity score $s(G_{t_1}, G_{t_2})$ in $[0,1]$ on the space of graph snapshots $(G_1, G_2, \dots, G_t, \dots)$ from the dynamic network. For this purpose, we design a similarity module for comparing the node-level embeddings output by a generic graph encoder (e.g., a graph convolutional network Kipf and Welling (2016), a graph attention network Veličković et al (2018), a GraphSage network Hamilton et al (2017) or a graph isomorphism network (GIN) Xu et al (2019)). Our similarity module consists of a Euclidean distance operation, a pooling layer and two fully-connected layers (see Figure 1b). The pooling operation in



(a) Siamese graph neural network (b) Similarity module

Figure 1: Architecture of our graph similarity learning model. The general pipeline (a) is a siamese GNN where the output module is a similarity module (b). We design the latter with Euclidean distance, Sort-k pooling operations, and fully-connected layers, for measuring the proximity of snapshots in dynamic networks.

this module is *Sort-k pooling* Zhang et al (2018), which consists in selecting and sorting the k largest entries of the input (here, the n-dimensional vector of Euclidean distances). This operation allows to select the subset of nodes having the largest displacement between H_1 and H_2, therefore to measure a local change of the graph. It also limits the number of parameters of the following fully-connected layer.

For the sake of simplicity, we use a simple graph convolutional network (GCN) Kipf and Welling (2016) for undirected and unsigned graphs as the graph encoder in our architecture. However, this block can be replaced by any *ad-hoc* graph encoder. With a GCN, the embedding of a graph $\boldsymbol{H}^{(j)}$ at each layer $j \in [J], J \geq 1$ is computed as follows

$$\boldsymbol{H}^{(j)} = \sigma\left(\tilde{\boldsymbol{A}}\boldsymbol{H}^{(j-1)}\mathbf{W}^{(j)} + \mathbf{B}^{(j)}\right), \tag{6}$$

where $\mathbf{W}^{(j)} \in \mathbb{R}^{h_{j-1} \times h_j}$ is a weight matrix, h_j is the number of hidden units of layer j, $\mathbf{B}^{(j)} \in \mathbb{R}^{h_j}$ is a bias vector, $\tilde{\boldsymbol{A}} = \tilde{\boldsymbol{D}}^{-1/2}(\boldsymbol{A}+\boldsymbol{I}_n)\tilde{\boldsymbol{D}}^{-1/2}$ is the normalized augmented adjacency matrix with degree matrix $\tilde{\boldsymbol{D}} = \mathrm{Diag}((\boldsymbol{A}+\boldsymbol{I}_n)\mathbb{1}_n)$, and σ is the point-wise ReLU activation function, i.e., $\sigma(x) = \max(x, 0)$. The input of the first layer, $\boldsymbol{H}^{(0)}$, is either the node attributes matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ if the input graph is attributed or a positional encoding matrix (see below). Finally, the output of the GCN is the node-level embedding matrix $\boldsymbol{H}^{J} \in \mathbb{R}^{n \times h_J}$ at the last layer. Therefore, for a pair of graphs (G_{t_1}, G_{t_2}), this siamese encoder module computes a pairs of graph embeddings, $(\boldsymbol{H}_1, \boldsymbol{H}_2) := (\boldsymbol{H}^{J}(G_{t_1}), \boldsymbol{H}^{J}(G_{t_2}))$, and the vectors $(\boldsymbol{H}_1)_{i:}$ and $(\boldsymbol{H}_2)_{i:}$ correspond to the representations of the node i respectively in G_{t_1} and G_{t_2}. Intuitively, a large distance between these two embeddings can indicate that node i plays distinct structural roles in G_{t_1} and G_{t_2}.

Then, the pair of embeddings $(\boldsymbol{H}_1, \boldsymbol{H}_2)$ is processed by a similarity module, which first computes a vector of Euclidean distance between the nodes' embeddings, and secondly, applies a Sort-k pooling operation Zhang et al (2018) to select its k largest entries, i.e.,

$$\boldsymbol{P} = (f_{r_1}, \ldots, f_{r_k}), \quad f_i = \|(\boldsymbol{H}_1)_{i:} - (\boldsymbol{H}_2)_{i:}\|_2 \in \mathbb{R}_{\geq 0}, \quad 1 \leq i \leq n,$$

where $r_1, \ldots, r_k$ correspond to the indices of the (sorted) k largest elements of $\{f_i\}_{i\in[n]}$. We note that the Euclidean distance could be also replaced by another distance, similarity, or kernel function such as the cosine similarity or a Gaussian kernel. Next, the pooled vector P is processed by two fully connected layers, each of them containing an affine transformation, batch normalisation and ReLU activation function. Finally, the output of the second fully connected layer is pooled using a sum-pooling layer followed by a sigmoid activation function, so that the final output of the similarity module (and the s-GNN), $s(G_{t_1}, G_{t_2}) \in [0, 1]$, a non-negative similarity score between the two input graphs. This score can be transformed into a similarity label via

$$\hat{y}(G_{t_1}, G_{t_2}) = \begin{cases} 1 & \text{(i.e., } G_{t_1} \text{ and } G_{t_2} \text{ are similar or have the same} \\ & \text{generative distribution)}, \quad \text{if } s(G_{t_1}, G_{t_2}) > 0.5 \\ 0 & \text{(i.e., } G_{t_1} \text{ and } G_{t_2} \text{ are dissimilar or have different} \\ & \text{generative distributions)}, \quad \text{otherwise.} \end{cases} \tag{7}$$

Note that using a Sort-k pooling layer in the design of the similarity module has two main advantages in our context.

• First, since it follows the (node-wise) Euclidean distance operation, it therefore selects the nodes that have the largest discrepancies between their embeddings in the two graphs. Therefore, if a structural change in the dynamic network affects only a few nodes, this change can be picked up by this pooling operation, without being diminished by the absence of change in the rest of the network. This component could be further built upon for identifying which *local* part of the network is mainly driving the change-point, thus enhancing the explainability of the proposed pipeline.

• Second, Sort-k pooling reduces the number of parameters while preserving the most important information for measuring potential and local graph changes. More generally, replacing max or sum pooling by sorted pooling have been proven to increase the accuracy and generalization power of neural networks, in particular in settings with limited data availability, such as one-shot learning Horváth (2020), and can also be used for downsampling graphs Lee et al (2019). In the network change-point-agnostic detection task, we incorporate this pooling layer to mitigate our lack of information on the change-points.

Remark 2. It is often a desirable property of GNN models with graph-level (resp. nodel-level) output to be invariant (resp. equivariant) to nodes' permutations. This is due to the fact that nodes in a graph are generally considered exchangeable, or in other words, the order of node set in the adjacency and node attributes matrix is arbitrary. In our method, the s-GNN model takes as input pairs of graph snapshots from a dynamic network sequence (see Section 3.1), where every snapshot contains the same set of nodes with the same ordering. Therefore, in our context, the invariance property of the learned graph similarity function denotes that the latter is invariant to any permutation of the nodes that is applied on both inputs. More precisely, for any permutation of the node set $\sigma : [n] \to [n]$, denoting $\sigma * G$ the resulting transformation of a graph G under σ (i.e., permutation of the rows and columns of the adjacency and node attributes matrices), the invariance property writes $s(\sigma(G_1), \sigma(G_2)) = s(G_1, G_2)$. This is indeed the case for our method since the node-wise operations, i.e, the graph encoder and the Euclidean distance, are equivariant. Then, since the Sort-k pooling layer is permutation-invariant, i.e. $\boldsymbol{P}(\boldsymbol{H}) = \boldsymbol{P}(\sigma(\boldsymbol{H}))$, so is the final similarity score.

Moreover, when the dynamic network is unattributed, i.e., each graph snapshot contains only structural information $G_i = (\boldsymbol{A}_i, \emptyset)$, one needs to choose an appropriate initialisation of the node features matrix $H^{(0)}$ as input of the s-GNN. Following existing methodology, we propose different variants of our method with different computations of the node encodings, i.e., synthetic node attributes that capture their relative positions in the graph structure or their specific identity. In fact, it has been previously noted that the choice of node encodings is critical for the expressivity of GNN models (Dwivedi et al, 2022). Therefore, in our experiments, we will use and compare four types of encoding, the first three being existing techniques that have been introduced in different graph learning settings, and the last being one that we believe may also be appropriate for certain NCPD tasks.

1. **Degree encoding (s-GNN-D) (Bruna and Li, 2017):** the attribute of a node is a scalar equal to its degree in the graph, i.e. $\boldsymbol{H}^{(0)} = \boldsymbol{A}\mathbb{1}_n \in \mathbb{R}^{n\times 1}$.
2. **Random-Walk encoding (s-GNN-RW) (Li et al, 2020; Dwivedi et al, 2022):** for $k \geq 1$, the vector of attributes of a node $u_i, 1 \leq i \leq n$ is defined as

$$\boldsymbol{H}^{(0)}_{i:} = [\boldsymbol{R}_{ii}, \boldsymbol{R}^2_{ii}, \dots \boldsymbol{R}^k_{ii}] \in \mathbb{R}^k,$$

   where $\boldsymbol{R} = \boldsymbol{A}\boldsymbol{D}^{-1}$ is the random-walk operator and $k \geq 1$ is a hyperparameter.
3. **Laplacian (or positional encoding) (s-GNN-PE) (Dwivedi and Bresson, 2021):** the node attributes are the principal eigenvectors of the symmetric normalised Laplacian matrix $\boldsymbol{L} = \boldsymbol{I}_n - \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$. We note that this is similar to the first steps of spectral clustering algorithms. More

precisely, using the factorisation $\boldsymbol{L} = \boldsymbol{U}^T \boldsymbol{\Lambda} \boldsymbol{U}$ where $\boldsymbol{U}, \boldsymbol{\Lambda}$ respectively contain the ordered set of eigenvectors and eigenvalues of $\boldsymbol{L}$, the Laplacian encodings are defined as

$$\boldsymbol{H}^{(0)} = [\boldsymbol{U}_{:1}^T, \boldsymbol{U}_{:2}^T, \dots, \boldsymbol{U}_{:k}^T] \in \mathbb{R}^{n \times k},$$

where $k \geq 1$ is a hyperparameter.

4. **Identity encoding (s-GNN-I):** we define the initial feature matrix as $\boldsymbol{H}^{(0)} = \boldsymbol{I}_n$, which corresponds to using a *one-hot* encoding of each node. We argue that this is an appropriate choice for the graph siamese encoder in our setting. In fact, these encodings in general break the equivariance property of GNN models; however, this is not the case here since this property has a modified definition when the graphs are snapshots of a dynamic network (see Remark 2). We recall that we have assumed that the set of nodes is constant in the dynamic network, and the global ordering of the nodes, although arbitrary, is common to all graph snapshots. Finally, it has been previously noted in graph learning tasks that taking into account the nodes' identities (Donnat and Holmes, 2018) can be beneficial. We will in particular use these encodings for the real-world dynamic network with a small number of nodes in Section 4.5.

We note that more complex strategies for computing positional encodings include learning them along the training procedure of the s-GNN (Dwivedi et al, 2022). However, we do not consider these latter approaches, which significantly increase the model complexity.

Finally, our s-GNN architecture can classically be trained in a supervised way if a data set of labelled pairs of graphs is available, with $\mathcal{D} = \{(G_1^i, G_2^i, y_i)\}_i$ with $y_i \in \{0, 1\}$ indicating if the graphs come from the same distribution or not. For example, one can optimise its parameters by minimising the cross-entropy loss function

$$\mathcal{L}_{BCE}(G_1, G_2, y) = -y \log s(G_1, G_2) - (1 - y) \log(1 - s(G_1, G_2)),$$

via gradient descent. In the next section, we propose a sampling scheme of the snapshots in the dynamic network and a supervised learning strategy for the s-GNN in the NCPD task.

3.3 Training and validation procedures for NCPD

In the NCPD task, a supervised setting corresponds to the case where a training subsequence of the dynamic network containing ground-truth change-points is available. In this setting, these change-point labels can then be used to design training and validation sets for our GSL model, with these sets containing triplets (G_i^1, G_i^2, y_i), where $y_i \in \{0, 1\}$ is a similarity label ($y_i = 1$ corresponding to "similar"). In this section, we describe our strategy for sampling such triplets from the dynamic network sequence, for both the training and validation steps.

First, we divide the sequence of graph snapshots into training, validation and test subsequences, e.g., using consecutive windows of respectively x%, y% and z% timestamps. In the training and validation sequences, we sample labelled pairs of graphs according to the two following schemes.

1. **Random scheme (training set):** we consider the set of all (non-ordered) pairs of graphs in the training sequence and label each pair (G_{t_1}, G_{t_2}) with $y = 1$ if there is no change-point between t_1 and t_2, and $y_i = 0$ otherwise. Then we uniformly sample a fixed number of pairs with label 1 ("positive" examples) and the same number of pairs with label 0 ("negative" examples), without replacement. The number of pairs is chosen heuristically between T and $10 \times T$ in our experiments.

2. **Windowed scheme (validation set):** we consider the set of all (non-ordered) pairs of graphs in the network sequence that are not distant from each other by more than L timestamps, and label them with the same procedure as in the **Random scheme**.

We note that the different sampling mechanisms for the pairs in the training and validation sets are designed to satisfy a double objective of our learning procedure: we aim to learn an adequate graph similarity function and to detect change-points in a network sequence using the latter. For the first objective, the **Random scheme** allow to subsample pairs of graph snapshots that are further away in the sequence. This design aims at mitigating two possible undesired effects in real-world dynamic networks: on the one hand, the possible temporal correlations between the snapshots in each pair and between the pairs themselves; in the other hand, "transition" phenomenon or gradual changes between generative distributions. Additionally, with the **Random scheme**, we can avoid label imbalance in the training set by sampling the same number of positive and negative pairs, which we assume is favorable for the s-GNN. For the second objective, the **Windowed scheme** builds a validation set of pairs that imitates the test setting of our GSL model. In fact, in our online NCPD method (see Section 3.1), the evaluation of the graph similarity function $s(.,.)$ in the statistic (3) (and detection rule (4)) only applies for pairs of graphs within a sliding window of size L. In particular, in the test setting, the pairs of graphs that are compared by $s(\cdot,\cdot)$ are highly correlated and the number of positive pairs is much larger than the number of negative examples, since there are generally only few change-points in the dynamic network. We finally note that in both the **Random** and **Windowed schemes**, the sampled pairs may have in common (at most) one graph snapshot.

Consequently, in a supervised NCPD setting, we can train our s-GNN model in a supervised way as a binary classifier of pairs (see Section 3.2) using the previous sampling strategies. In an unsupervised NCPD setting, i.e., when the dynamic network does not contain any ground-truth label of change-point, we need to resort to a novel *self-supervised learning* technique (Liu et al, 2021). In this case, we first pre-estimate a set of change-points in the training and validation

sequences using a *spectral* technique, then apply the previous sampling schemes to draw training and validation pairs of graphs (see more details in Section 4.4 where this strategy is applied to the financial network data set).

4 Numerical experiments

In this section, we test and evaluate the performances of our s-GNN method in the online NCPD task, first in a controlled setting of synthetic dynamic networks (Section 4.3), then on real-world correlation networks (Sections 4.4 and 4.5). Moreover, since one of these data sets does not contain ground-truth change-points, we also introduce a self-supervised learning procedure for our method (see Section 4.4).

4.1 Performance metrics

For dynamic network data sets with ground-truth labels of change-points, we evaluate the performance of NCPD methods using the following metrics, in the single or multiple change-point settings:

- **Localisation error (single change-point)** defined as $\text{Error}_{\text{CPD}} = |\hat{\tau} - \tau|$, where $\tau, \hat{\tau}$ are respectively the ground-truth and estimated change-points.
- **Adjusted F1-score (Xu et al, 2018) (multiple change-points)** on the classification of timestamps as change-point (label 1) or not change-point (label 0). We note that the label of timestamps differs from the definition of pair labels in (7) where the label 1 corresponds to "similar" pairs. We tolerate an error of ± 5 timestamps on this task, i.e. all the timestamps within a window of length 11 centered at the ground-truth change-point are given a ground-truth label 1, and a valid detection occurs whenever one of these timestamps is classified as change-point.

For the the data set without ground-truth labels in Section 4.4, we qualitatively discuss our findings, frame them in a financial context, and compare them with previous analysis of similar data. Additionally, in the synthetic data experiments in Section 4.3, we also evaluate the ability of our graph similarity function $\hat{s}$ to discriminate between graphs sampled from the same or different distributions, i.e., to classify pairs of graphs generated from either the same or different random graph models. We measure this property in terms of the accuracy score.

4.2 Baselines

We will compare our data-driven graph similarity function to graph distances, similarity function and kernel previously used in the context of NCPD and graph two-sample-test.

- **Frobenius distance** (Barnett and Onnela, 2016; Nie and Nicolae, 2021; Bao et al, 2018; Dubey et al, 2021), defined as $d_F(\boldsymbol{A}, \boldsymbol{B}) = \|\boldsymbol{A} - \boldsymbol{B}\|_F$,

for two matrices $\boldsymbol{A}, \boldsymbol{B}$ with equal dimensions. Here, we will apply this distance to the adjacency matrices of two graphs. Note that one can also apply it on the graph Laplacian matrices (Bao et al, 2018), and that this distance has also been used in a minimax testing perspective between two graph samples (Ghoshdastidar et al, 2020).

- **Procrustes distance** between Laplacian principal eigenspaces (Hewapathirana et al, 2020). This distance corresponds to the Frobenius distance between the matrices of eigenvectors corresponding to the k largest eigenvalues of the symmetric graph Laplacian $\boldsymbol{L} = \boldsymbol{I}_n - \boldsymbol{D}^{1/2}\boldsymbol{A}\boldsymbol{D}^{1/2}$, after performing an alignment step. The number of eigenvectors k can be prespecified or chosen by finding the optimal low-rank approximation of $\boldsymbol{L}$.
- **DeltaCon similarity (Koutra et al, 2016)**. This graph similarity function is based on the Matusita distance applied to the Fast Belief Propagation graph operators, defined for a graph as $\boldsymbol{S} = [\boldsymbol{I}_n + \epsilon^2\boldsymbol{D} - \epsilon\boldsymbol{A}]^{-1}$ with $\epsilon > 0$. We use the implementation of this similarity function provided in the python package `netrd` [1].
- **Weisfeiler-Lehman (WL) kernel (Shervashidze et al, 2011)**. This graph kernel is notably used in the two-sample-test problem for sets of graphs (Gretton et al, 2008). We use the implementation from the GraKel python package (Siglidis et al, 2020), and fix the number of iterations of the WL kernel algorithm to 5 in our experiments.

We will use the previous baselines in the statistic (3) and detect change-points using a threshold chosen on a validation set. We also compare our NCPD pipeline to methods that do not rely on an explicit graph metric for detecting change-points.

- **Network change-point detection with spectral clustering (SC-NCPD)** (Cribben and Yu, 2017). This method first partitions the node set of each snapshot with a spectral clustering algorithm and compute an inner product between averages of spectral features across a backward (or *past*) and a forward (or *future*) windows. In this method, the number of clusters and the lengths of the windows are pre-specified.
- **Laplacian anomaly detection (LAD)** (Huang et al, 2020). This method applies both to the anomaly detection and change-point detection tasks for dynamic networks, and is based on the anomaly score

$$Z_t = 1 - |\tilde{\sigma}_t \sigma_t|,$$

where σ_t is the vector of top-k singular values of the unormalized Laplacian of the graph G_t and $\tilde{\sigma}_t$ aggregates (e.g. averages) the top-k singular values of each snapshots in a past window of size L, i.e., $(\sigma_{t-L}, \ldots, \sigma_{t-1})$. The

[1] https://netrd.readthedocs.io/

number of singular values k and the length of the window are pre-specified hyperparameters.

- **Network cumulative sums statistic (CUSUM) (Yu et al, 2021).** This method uses a backward and a forward windows of sizes L' to compute a sequence of CUSUM matrices

$$\boldsymbol{C}_t = \frac{1}{\sqrt{2L'}} \left(\sum_{s=t-L'+1}^{t} \boldsymbol{A}_s - \sum_{s=t+1}^{t+L'} \boldsymbol{A}_s \right), \quad L' \leq t \leq T - L'. \tag{8}$$

  Following the methodology in Yu et al (2021), we divide the dynamic network into two samples, $N_A = \{G_{2t}\}_{1 \leq t \leq T/2}$ and $N_B = \{G_{2t-1}\}_{1 \leq t \leq T/2}$, containing the snapshots respectively at even and uneven timestamps. This algorithm monitors two statistics based on the CUSUM matrices (8) of these samples: the Frobenius norm of the Universal Singular Value Threshold (USVT) estimator $\widetilde{B}(t)$ of the CUSUM matrix computed from N_B, and the dot product between $\widetilde{B}(t)/\|\widetilde{B}(t)\|$ and the CUSUM matrix computed from N_A. To avoid tuning the additional threshold parameters, we do not apply the USVT step (or equivalently choose $\tau_1 = 0$ and $\tau_2 = 1$ in USVT). Moreover, we only use the second statistics since the first one is very close to the next baseline.

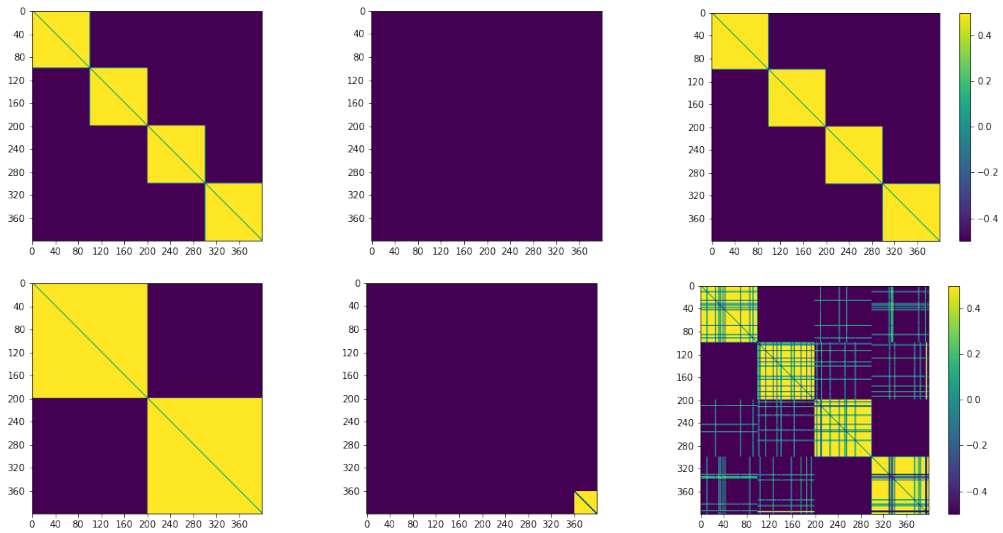
- **Operator norm of network CUSUM (CUSUM 2) (Enikeeva and Klopp, 2021).** We adapt this offline method to the online problem by computing the CUSUM matrix over a past and future windows of size L'. The NCPD statistics is then $z_t = \|\boldsymbol{C}_t\|$.

For these baselines, we fix the number of clusters or singular values to $k = 6$ and the size of windows to $L' = L/2$ when both the past and future are used in the NCPD statistic. We also note that these methods are applied to non-attributed dynamic networks and therefore only use the sequence of adjacency matrices $(\boldsymbol{A}_t)_t$. However, only one of our network data sets is attributed (see Section 4.4) and in this case, the node attributes are ignored by the baseline methods.

4.3 Synthetic data

In this section, we generate dynamic networks from a dynamic stochastic block model (Zhao et al, 2019; Yu et al, 2021; Padilla et al, 2019; Bhattacharjee et al, 2020) with a unique change-point. More precisely, we generate sequences of unattributed graphs $(G_1, \ldots, G_T)$ with $T = 100$ such that for each $t \in [T]$, each graph is independently drawn from a Stochastic Block Model (SBM) with $n = 400$ nodes and

$$G_t \overset{i.i.d}{\sim} \mathcal{G}_1, \quad \text{if } t < \tau,$$
$$G_t \overset{i.i.d}{\sim} \mathcal{G}_2, \quad \text{if } t \geq \tau,$$



(a) "Merge" scenario. (b) "Birth" scenarios. (c) "Swap" scenario.

Figure 2: Expectation of the adjacency matrices of the graphs in the SBMs $\mathcal{G}_1$ (first row) and $\mathcal{G}_2$ (second row) , i.e., before and after the change-point, in our three types of synthetic scenarios, "Merge" (a), "Birth" (b) and "Swaps" (c).

where $\mathcal{G}_1, \mathcal{G}_2$ are two SBM distributions. We recall that an SBM with $K \geq 1$ communities can be defined by a connectivity matrix $\boldsymbol{C} = (p-q)\boldsymbol{I}_K + q\mathbb{1}_K\mathbb{1}_K^T$ with intra- and inter-cluster connectivity parameters $p, q \in [0,1]$, and a membership matrix $\Theta \in \{0,1\}^{n\times K}$. The parameter p (respectively q) corresponds to the probability of existence of an edge between two nodes in the same community (respectively in two different communities), while each row Θ_i of the membership matrix indicates the community a node n_i belongs to.

We consider four different change-point scenarios related to three possible types of events in a community life-cycle (Rossetti and Cazabet, 2018), namely **"Merge"**, **"Birth"** and **"Swaps"**. These community events are illustrated in Figure 2 by heatmaps of the expected adjacency matrices in $\mathcal{G}_1$ and $\mathcal{G}_2$. We note that in all the following settings, the graph snapshots will be relatively sparse.

- **Scenario 1 ("Merge").** In this scenario, the two SBMs $\mathcal{G}_1$ and $\mathcal{G}_2$ have respectively four and two equal-size clusters with inter-cluster connectivity parameter $q = 0.02$ and intra-cluster connectivity parameter $p > q$ which we vary. We design several difficulty levels of this scenario by changing the value of p: the larger p is, the easier the detection problem is.

- **Scenario 2 ("Birth 1").** This scenario mimics the appearance of a community in a dynamic network. In this case, $\mathcal{G}_1$ is the distribution of an Erdos-Renyi model with parameter $q = 0.03$ and $\mathcal{G}_2$ is a SBM with two

communities of size $n - s$ and s, $1 \leq s \leq n/2$, and connectivity matrix

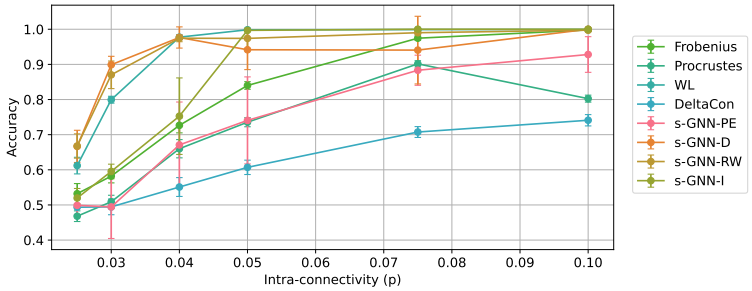
$$C = \begin{pmatrix} q & q \\ q & p \end{pmatrix},$$

with $p = 0.1$. We vary the difficulty of this detection scenario by changing the size of the second cluster s: the bigger s is, the easier the detection problem is.

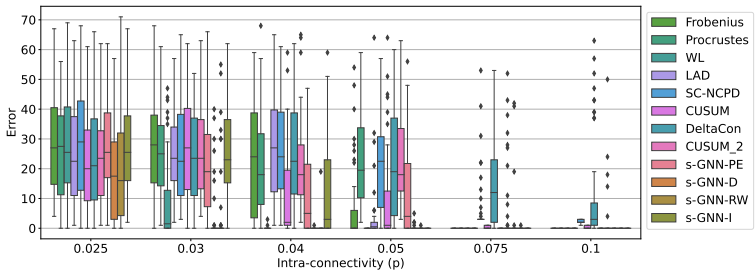
- **Scenario 3 ("Birth 2").** This scenario uses the setting of Scenario 2 but in this case, the size of the dense subgraph is fixed to $s = 100$ and the difficulty level is controlled by the connectivity p. We consider $p > q$ such that the larger p is, the easier the detection problem.
- **Scenario 4 ("Swaps").** In this scenario, the connectivity parameters of the two SBMs are equal but their membership matrices differ. We simulate a recombination of communities where pairs of nodes exchange their community memberships, i.e., two nodes "swap" their community of attachment. The two SBMs have four equal-size clusters with inter-connectivity parameter $q = 0.05$ and intra-connectivity parameter $p = 0.1$. We test different difficulty levels by varying the proportion h of pairs of nodes swapping their memberships; the bigger the h, the easier the detection problem.

Note that Scenario 1 can be considered as a *global* change of the network structure, while the other scenarios correspond to a *local* topological change (i.e., localised on a subset of nodes). For each scenario and each difficulty level, we generate 50 sequences with one change-point uniformly sampled in the interval $[25, 75]$. Moreover, for the pair classification task (see Section 4.1), in each scenario, we also independently generate 1000 labelled pairs of graphs $\{(G_1^i, G_2^i, y_i)\}_i$, where for each i, $G_1^i \sim \mathcal{G}_k, G_2^i \sim \mathcal{G}_l$ with $k, l \in \{1, 2\}$ and $y_i = 1$ if $\mathcal{G}_k = \mathcal{G}_l$ and $y_i = 0$ otherwise. Each of these data sets of pairs is balanced and we use respectively 60%, 20% and 20% of the pairs for training, validation and test. In the NCPD task, we estimate the unique change-point with the detection rule (5), and use a window size $L = 6$. Additional details on the experimental setting can be found in Appendix A.

We test our NCPD method with the four variants of node encodings (**Degree**, **Random Walk**, **Laplacian** and **Identity**) defined in Section 3.2 and report the results of each scenario in Figures 3, 4, 5 and 6. In almost all scenarios and difficulty levels, our method outperforms the other baselines, except for the variant with Laplacian attributes. The drop of performance using the latter type of encodings has been previously attributed to the sign ambiguity in Laplacian eigenvectors (Dwivedi et al, 2022). Moreover, the degree and random walks encodings generally seem to be better than the identity encodings, except for the last scenario. We conjecture that this is due to the fact that in the first three scenarios, nodes in the same cluster are exchangeable in the SBM model, while in the last scenario, this symmetry is broken by the membership exchange



(a) Classification accuracy vs intra-connectivity parameter p.

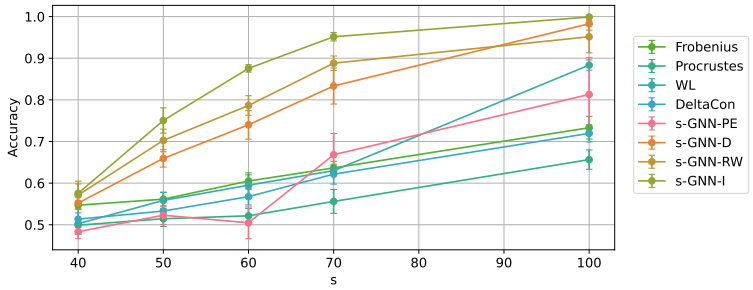


(b) Change-point localisation error for different intra-connectivity parameters p.

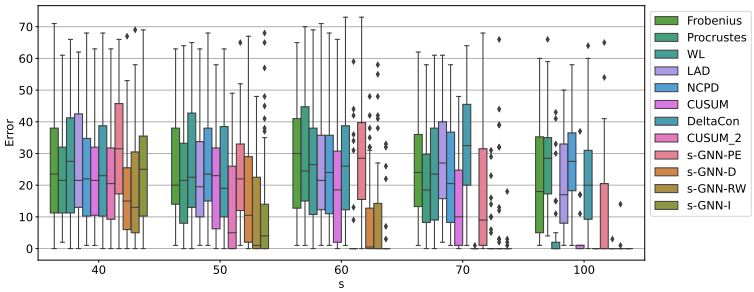
Figure 3: Performances of our s-GNN method and baselines on the classification and detection tasks in the "Merge" scenario. In the first task, pairs of graphs sampled from the same or different SBM distributions are classified using a graph similarity function or a graph distance, therefore, the set of baselines only consists of the latter type of algorithms. In the second task, a single change-point needs to be localised in a dynamic SBM sequencem and the set of baselines include graph distance- (or kernel-) based methods and network change-point detection methods. We remark that for very large values of p, many methods attain zero error and our method achieves a smaller error for all values of p.

mechanism. For networks with a lot of symmetry, the **Identity** encoding might introduce additional noise.

We also observe that the strongest baselines are **CUSUM** and **CUSUM 2**, which have better performances if larger window sizes L are used, while our method is not sensitive to this hyperparameter (see Appendix A.2). In particular, our method performs well even for a short history of data and therefore could also detect change-points that are close to each other in a multiple change-point setting. Consequently, these experiments show that using a data-driven graph similarity function leads to better performances in the NCPD task than existing baselines, in various change-point scenarios.

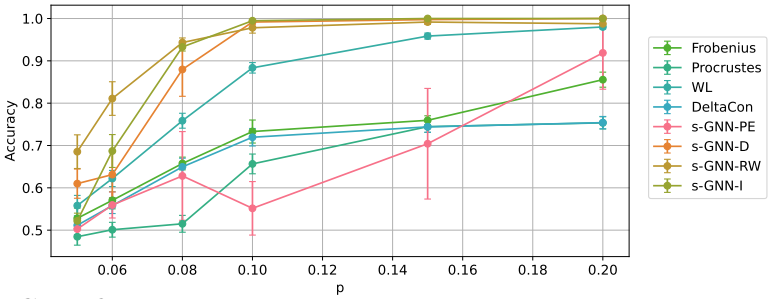


(a) Classification accuracy of pairs vs the community size s.

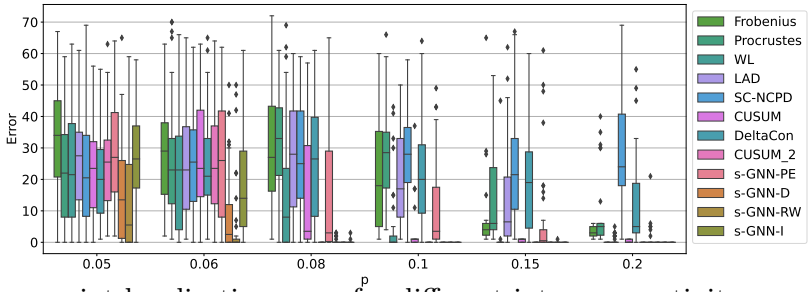


(b) Change-point localisation error for different community sizes.

Figure 4: Performances on the classification (a) and detection (b) tasks in the "Birth 1" scenario.

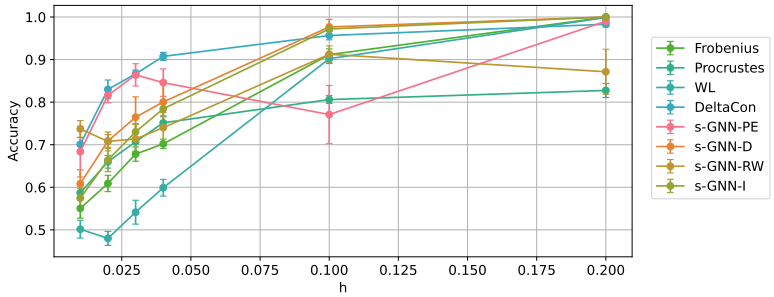


(a) Classification accuracy vs the intra-connectivity parameter p.

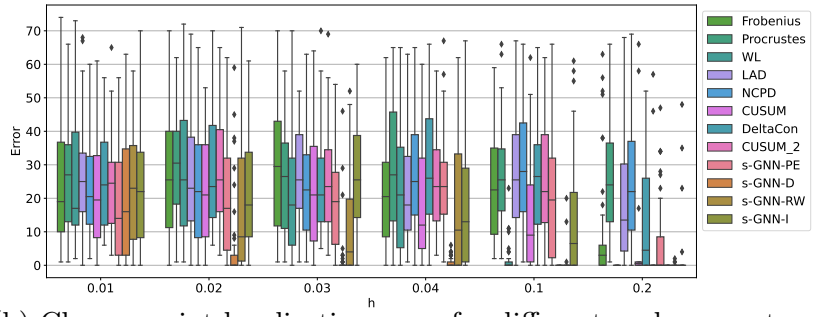


(b) Change point localisation error for different intra-connectivity parameters p.

Figure 5: Performances on the classification (a) and detection (b) tasks in the "Birth 2" scenario.



(a) Classification accuracy vs the exchange rate h.



(b) Change point localisation error for different exchange rates h.

Figure 6: Performances on the classification (a) and detection (b) tasks in the "Swaps" scenario.

4.4 Correlation network from stock returns data

This data set comprising the cross-correlation networks of daily stock returns, computed over a one-month interval, from the S&P 500 index in a period of about 20 years (February 2000 - December 2020). Data sets of stock returns have been previously analysed in different contexts. For online NCPD, Yu et al (2021) and Barnett and Onnela (2016) consider the covariance matrices of S&P 500 weekly log-returns respectively on the period between 1950 and 2000, and between 1982 and 2000, while Dubey et al (2021) analyses the weekly log-returns of 29 stocks from the Dow Jones Industrial Average index, from April 1990 to January 2012. Closely related to our problem, Chakraborti et al (2020) and Samal et al (2021) cluster market behaviours in the USA S&P 500 and Japan Nikkei 225 stock networks during the period from 1985 to 2016.

In this analysis, we consider 685 stocks (therefore nodes in the dynamic network) alongside additional information of their economic activity during each month. The correlation networks are built from the time series of open-to-close (intraday) and close-to-open (overnight) returns. Typically, there are 21 trading days in a calendar month, hence each stock has associated a time series of length 42, since each day of the month contributes with two returns. The resulting stock correlation matrix is the starting point for our network construction. In addition, we employ the following stock properties as node attributes

- volatilities: the standard deviations of the above 42 open-to-close and close-to-open returns, based on which the correlation network was built,

- average daily volume, in shares, over the 21 days of the month,
- average shares outstanding, over the 21 days of the month.

We then construct an attributed and unweighted dynamic network $\mathcal{G}_F = ((\boldsymbol{A}_t, \boldsymbol{X}_t))_{1\leq t\leq 244}$ with 244 snapshots, and for each t, $\boldsymbol{A}_t \in \{0,1\}^{685\times 685}$ and $\boldsymbol{X}_t \in \mathbb{R}^{685\times 4}$, using a truncation procedure of the correlation matrices between stocks. More precisely, we set to 1 the matrix entries that are below the 0.1-quantile and above the 0.9-quantile among the entries of all correlation matrices. We note that after this preprocessing step, each graph snapshot is connected and contains self-loops. A similar procedure has been applied in Yu et al (2021), while other works transform the correlation matrices into complete weighted graphs, e.g., by squaring the correlation coefficients (Chakraborti et al, 2020) or computing the inverse of the ultra-metric distance (Samal et al, 2021). Here we adopt the sparsifying approach to avoid dealing with a large complete graph. In Table 1, we report some properties of the resulting network. Finally, we standardize the node attributes matrices $\{\boldsymbol{X}_t\}_{1\leq t\leq 244}$ across the timestamps: for each column (i.e. each attribute) and each matrix, we center and scale its values by the mean and standard deviation of all the values of this attribute in the graph snapshots.

Although previous work reported changes in the behaviour of the stock market following different economic or global events (Chakraborti et al, 2020), there is no ground-truth knowledge of change-points for this dynamic correlation network. However, there is strong evidence that some major events, such as the ones listed in Table 2, have impacted the dynamics of stock returns and their correlation (Barnett and Onnela, 2016). Therefore, we consider a self-supervised training procedure (Liu et al, 2021) that first pre-estimates a set of change-points in order to train our s-GNN with the procedure described in Section 3.

We first divide our dynamic network into consecutive windows of 50%, 20% and 30% graph snapshots as training, validation and test sequences. Then we pre-estimate change-points in the training and validation sequences using the following methodology. It is common practice to cluster stocks into market sectors (Chakraborti et al, 2020), and we conjecture that this cluster structure is reflected in the correlation network and is a proxy for the underlying state of the financial market at a given time. Therefore, we consider the following three-step procedure:

1. We estimate a cluster structure for each correlation matrix using a spectral clustering algorithm based on the Symmetric Signed Laplacian (Gallier, 2016).
2. We compare the obtained node partitions in each pair of matrices using the Adjusted Rand Index and use the latter as a pairwise measure of similarity between correlation graphs. Then we apply a spectral clustering algorithm based on the Normalised Symmetric Laplacian on the graph

Financial network ($T = 244$, $n = 685$)

	Mean	Median	Standard deviation
Number of edges per graph	46.3×10^3	40.3×10^3	23.6×10^3
Edge density	0.20	0.17	0.10
Average degree	135	96	111
Average shortest path length	1.8	1.8	0.1
Diameter	2.8	3.0	0.5

Table 1: Mean, median and standard deviation of network statistics for the snapshots in the correlation network of S&P index stock returns.

snapshots (i.e., each snapshot is given a label, interpreted as a state or behaviour of the stock market).

3. We estimate change-points by "smoothing" the snapshots' labels: we compute the centroid timestamp of each cluster of snapshots and relabel the latter with the labels of their closest centroid. These new labels now define a partition of the temporal window into consecutive intervals, and therefore pre-estimate change-points in network training sequence.

In the first step, we cluster each correlation matrix into $k = 13$ clusters; this value is chosen by evaluating the silhouette index of the result clustering for different number of clusters $k \in \{10, \dots, 20\}$. In the second step, we cluster the similarity matrix between the graph snapshots based on the ARI (see Figure 11 in Appendix A) into $C = 9$ clusters. We note that in the first step, the clusters correspond to sets of nodes (i.e., stocks) in each graph, while in the second step, the clusters are sets of graph snapshots. The estimated change-points obtained in the third step are plotted in Figure 12a.

Then for the training set, we sample $N = 3000$ pairs of graphs using the "Random scheme" (see Section 3.3) and for the validation set, we use the "Windowed scheme" with a window of size 12, which leads to 684 pairs. For choosing the hyperparameters of the s-GNN, we test every configuration of values with the learning rate in the set $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$, the weight decay in $\{10^{-6}, 10^{-5}\}$, the dropout rate in $\{0.05, 0.2, 0.4\}$, the output size of the Sort-k layer in $\{50, 100, 200\}$, the number of hidden units in $\{32, 64, 128\}$. We select the model with the highest F1-score on the validation set to make predictions on the test set.

Finally, we compute our NCPD statistic (3) with a window size $L = 6$ over the whole network sequence (i.e., training, validation and test), and qualitatively interpret the time series $1 - Z_t(s, L)$ and the detected change-points, plotted in Figure 7, and compared to baselines and the VIX volatility index in Figure 8. We notably compare the peaks of the statistics with a timeline of major market events listed in Table 2. We observe in Figure 7 that some of the detected change-points coincides or happens soon after market events, in particular in the test sequence from November 2014 to December 2020. During this period, a group of peaks are observed around the Chinese Black Monday in August

Major crashes	Period Date
9/11 Financial Crisis	11/09/2001
Stock Market Downturn Of 2002	09/10/2002
US Housing Bubble	2005-2007
Lehman Brothers Crash	16/09/2008
DJ Flash Crash	06/05/2010
Tsunami/Fukushima	11/03/2011
Black Monday / Stock Markets Fall	08/08/2011
Chinese Black Monday	24/08/2015
Dow Jones plunge	02/2018 - 03/2018
WHO public emergency state (COVID-19)	30/01/2020

Table 2: Dates of major crashes and bubbles in the USA market.

2015, and two other large peaks are observed in March 2018 and December 2019, which could be attributed respectively to the Dow Jones index's plunge in February-March 2018 and the emergence of COVID-19 in late 2019. Note that the World Health Organisation declared a Public Health Emergency of International Concern in January 2020. In the training and validation period (from February 2000 to October 2014), the peak in November 2005 could be related to the US Housing Bubble, which spans a period between 2005 and 2007, while the peak in February 2007 could also be linked to the latter event or to the premises of the financial crisis of 2007-2008. Finally, the peaks in August 2009 and January 2012 are difficult to relate to one of the listed events in Table 2 (possibly the Stock Market Fall for the latter one). However, we note that these events are not ground-truth change-points for the correlation network of stocks and other factors unreported in this analysis may also influence its structure.

Nonetheless, in comparison to the baselines, our method is able to detect more events. As can be observed in Figure 8, almost all baselines detect change-points between 2010 and 2012, a period when several financial crashes happened such as the 2010 Dow Jone flash crash, the Fukushima nuclear incident in March 2011 and the Stock Markets Fall of August 2011. However, most baselines fail to detect anything outside this two-year period. One exception holds for the **CUSUM 2**, which indicates network disruptions at roughly six periods: during the financial crisis of 2007-2008, Dow Jones Flash Crash in 2010, the Stock Market Fall in 2011, the Chinese Black Monday in 2015, the Dow Jones plunge in 2018 and the consequences of the COVID-19 pandemic in 2020. However, this method delimits some periods of disruptions rather than clear change-points.

One explanation could be that since our method benefits from using the stock attributes previously listed, which are not taken into account by the baselines. We therefore also tested our method using synthetic attributes (see Section 3.2) and our results are reported in Figure 12d and 12c in Appendix B. We note that our method with synthetic attributes detects much more change-points (Figure 12c) than our method on the attributed network (Figure 12b), and is harder to interpret. We therefore conjecture that the stock attributes are beneficial for our method on this data set.

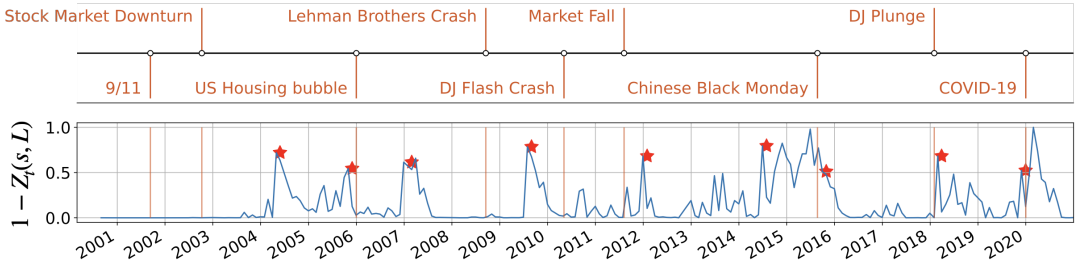


Figure 7: Change-point statistic $1-Z_t(s, L)$ obtained with our graph similarity learning algorithm on the dynamic correlation network of S&P 500 stock returns from February 2000 to December 2020. This period covers a training period from February 2000 to August 2010, a validation period from September 2010 to October 2014 and a test period from November 2014 to December 2020. Main financial events that occured during this period are indicated with vertical red bars. The detected change-points are marked with red stars, and correspond to timestamps verifying (4) and at least 6 months away from the previously detected change-point.

4.5 Correlation networks from physical activity monitoring

This public data set[2] was built for benchmarking time series classifiers on physical activity monitoring (Reiss and Stricker, 2012b,a). This data contains multivariate time series recorded from eight subjects wearing 3D inertial measurement units (IMUs) and performing a protocol of 12 different physical activities such as sitting, walking, descending and ascending stairs and vacuum cleaning. The time series correspond to measurements from 3 IMUs positioned on the subjects' wrist, chest and ankle and containing 3-axis MEMS sensors (an accelerometer, a gyroscope and a magnetometer) with a sampling period of 0.01s. Thus, the dimension of the time series is 27, and there are 8 time series in total (one per subject).

Although this data has also been analysed in the change-point detection task for time series (Zhang et al, 2020), to our knowledge, it has not been used in the context of NCPD. However, previous work noted that the correlations between pairs of axis are particularly useful for differentiating activities based on translations such as walking, running or ascending stairs (Reiss and Stricker, 2012a). Therefore, similarly to Section 4.4, for each subject (i.e., each multivariate time series), we build a dynamic correlation network from the 27 time series, where each node therefore corresponds to an IMU sensor's axis and is associated with a body part.

More precisely, we segment the time series into non-overlapping windows of 100 observations (i.e., a window of length one second) and compute the correlation matrices of these time series over these windows. Then, the correlation matrices

[2] http://www.pamap.org/demo.html

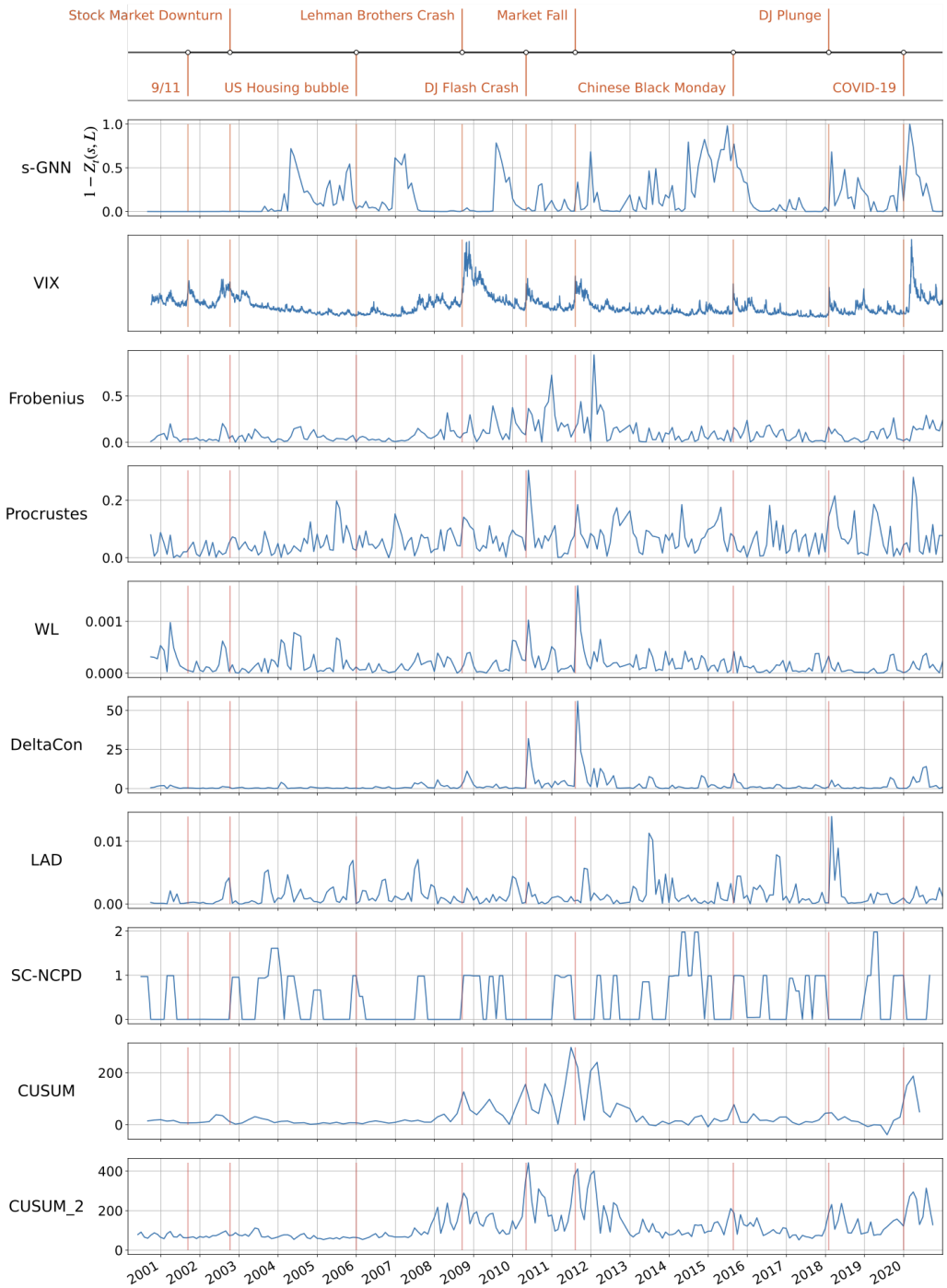


Figure 8: Change-point detection statistics obtained with our method and the baselines on the dynamic correlation network of S&P 500 stock returns from February 2000 to December 2020. This period covers a training period from February 2000 to August 2010, a validation period from September 2010 to October 2014 and a test period from November 2014 to December 2020. Main financial events are indicated with vertical red bars. The different rows correspond, from top to bottom, to a timeline of known market events, our NCPD statistic, the VIX volatility index and the baselines' NCPD statistics.

Subject	Activity performed						Number of change-points	Number of timestamps
	1-4	5	6	7	12,13,16,17	24		
1	✓	✓	✓	✓	✓	✓	13	2490
2	✓	✓	✓	✓	✓	✓	13	2618
3	✓	–	–	–	✓	–	10	1732
4	✓	–	✓	✓	✓	–	11	2302
5	✓	✓	✓	✓	✓	✓	13	2709
6	✓	✓	✓	✓	✓	✓	13	2487
7	✓	✓	✓	✓	✓	–	12	2314
8	✓	✓	✓	✓	✓	✓	13	2606

Table 3: Properties of the dynamic networks obtained from the physical activity monitoring data. Some activities are not listed because they have not been performed by any of the subjects in this experiment.

$\{\boldsymbol{C}_t\}_t$ are transformed into binary adjacency matrices $\boldsymbol{A}_t = \mathbb{1}_{|\boldsymbol{C}_t|>\eta}, t \in [T]$ with a chosen threshold $\eta = 0.2$. We thus obtain an unweighted, unattributed, dynamic network with 27 nodes for each of the 8 subjects. Moreover, each graph snapshot is labelled with the activity performed by the subject during the corresponding temporal window. Therefore, a change of activity between two consecutive snapshots corresponds to a change-point for the network. Table 3 summarises the characteristics of each network.

We then define two NCPD tasks to evaluate our method on, defined as follows.

- **Individual-level NCPD.** Each dynamic network (subject) is used separately and segmented into train, validation and test set. Then we train and test one s-GNN model per network, therefore the learnt graph similarity function is subject-dependent. We note that in this setting, the test sequence contains graphs with unseen activity labels.
- **Cross-individual NCPD.** The eight dynamic networks are combined into a train, validation and test set and we train only one s-GNN model for all sequences. In this case, our method learns only one graph similarity function for all the subjects and its performances evaluated on all of them.

More precisely, in the **Individual-level NCPD** task, we randomly split each dynamic network into 70% training and validation, and 30% test, by isolating a test interval and a validation interval. The latter is a window of size 60 centered around a uniformly sampled change-point, while the lower end of the test interval is uniformly sampled in the whole sequence. We then sample 4000 pairs from the training sequence and 1000 pairs from the validation interval, and train and evaluate our method on each subject separately. In the **Cross-individual NCPD** task, we design the following two evaluation settings.

1. Random split: we concatenate all the training, validation and test sequences from the previous task, as well as the training and validation sets of pairs. Then we subsample respectively 15000 and 3000 pairs with the **Random**

scheme (see Section 3.3). Note that in this case, the s-GNN is trained and tested on sub-sequences from every dynamic network.

2. Leave-one-subject-out (LOSO): in this setting we keep the whole sequence of one subject for testing, and train and validate on the seven remaining sequences. For the latter, we select validation intervals as in the previous task, and sample 3000 training and 1000 validation pairs in each network. Then we subsample 15000 and 3000 pairs from the aggregated training and validation sets.

One can get an insight on the feasibility of these two tasks by looking at (a) the similarity of adjacency matrices within each dynamic network, grouped by activity labels; (b) the similarity of adjacency matrices within the same activity, grouped by network (i.e., subject). We report in Appendix C.1 some insights on the dissimilarity between activities and subjects, measured in terms of the Frobenius distance. In Figure 16, we plot the average adjacency matrices in each activity (the average of all matrices corresponding to the same activity) from the first subject. We note that some of these matrices are quite similar, like the ones for activities $\{1, 2, 3\}$, that correspond respectively to sitting, lying and standing, which are all static activities. In Figure 17, we plot the Frobenius distance between these matrices. We observe that the lowest values of this distance matrix are mostly located on the diagonal, indicating that the graphs that have the same label are more similar to each other than graphs with different labels. In Figure 18, we represent the Frobenius distances between the graphs of different subjects with the same label, for four activities. We observe that for activities 5 and 7, the distances between matrices of different subjects are bigger than the ones from the same subject, however this difference is not always significant and does not seem to appear for activities 1 and 17. Figure 19 confirms this observation: we note that the average Frobenius distance between the graphs with the same label and from different subjects is smaller than the average distance between graphs with different labels.

Since the network is unattributed in this data set and the number of nodes is small, we use the **Identity** encoding for the s-GNN. For our change-point statistic, we use sliding windows of $L = 20$ timestamps. We report the performances of our method and baselines in terms of the adjusted F1-score (Xu et al, 2018) in Table 4, with a tolerance window of ± 5 timestamps. Our method has the best performance in most evaluation settings, and largely outperforms the baselines in the LOSO scheme. This latter result seems to indicate that the s-GNN is able to learn a graph similarity function that is generalisable to unseen subjects, while the good performances in the **Individual-level** task suggests that it can generalise to unseen activities.

5 Conclusion

In this work, we proposed a novel method for detecting change-points in dynamic networks using a data-driven graph similarity function. The latter is

Individual-level NCPD					
Subject	s-GNN-I	Frobenius	SC-NCPD	CUSUM	CUSUM 2
1	0.76 (0.20)	0.62 (0.31)	**0.82 (0.14)**	0.54 (0.30)	**0.81 (0.20)**
2	**0.91 (0.11)**	0.45 (0.22)	0.61 (0.08)	0.45 (0.12)	**0.84 (0.11)**
3	**0.60 (0.18)**	0.37 (0.14)	**0.67 (0.15)**	0.21 (0.27)	0.34 (0.26)
4	**0.73 (0.18)**	0.58 (0.26)	**0.70 (0.08)**	0.60 (0.07)	0.59 (0.22)
5	**0.85 (0.19)**	0.61 (0.22)	0.72 (0.16)	0.36 (0.24)	**0.72 (0.13)**
6	**0.74 (0.19)**	0.73 (0.22)	**0.75 (0.17)**	0.56 (0.30)	0.58 (0.16)
7	**0.90 (0.13)**	**0.79 (0.19)**	0.67 (0.35)	0.57 (0.23)	0.72 (0.37)
8	0.72 (0.24)	**0.88 (0.13)**	0.65 (0.14)	0.57 (0.28)	**0.82 (0.13)**
Cross-individual NCPD					
Setting	s-GNN-I	Frobenius	SC-NCPD	CUSUM	CUSUM 2
Random split	**0.81 (0.07)**	0.75 (0.03)	0.75 (0.03)	0.59 (0.12)	**0.80 (0.04)**
LOSO	**0.89 (0.02)**	0.70 (0.20)	**0.77 (0.06)**	0.62 (0.11)	0.75 (0.12)

Table 4: Adjusted F1-score of our method and baselines in the **Individual-level** and **Cross-individual** NCPD tasks on the physical activity monitoring data. The red, respectively blue, bold values in each row denote the top best, respectively second best, performing methods. The values in the parentheses denote the standard deviation over 10 repetitions of the random splits train/-validation/set, except for the leave-one-subject-out (LOSO) setting for which mean and standard deviation are computed over the 8 folds.

learnt by a siamese GNN model, trained and validated on pairs of graphs from the network sequence. This similarity function allows to effectively compare the current graph to its short-term history for detecting potential displacements, with a simple online statistic. We demonstrated the benefits of our method in synthetic experimental settings of dynamic SBMs, and on two real-world data sets of correlation networks, and concluded that our method is more accurate at distinguishing graphs with different generative distributions and detecting change-points, compared to existing baselines.

As previously noted, one main challenge posed by using a deep-learning based model for NCPD is the training and validation procedures, which necessitate either a data set with change-point labels, or an adequate unsupervised or self-supervised learning procedure. Since the former is quite rare, a future direction for this work could to develop the latter approach, for instance using data

augmentation strategies (Carmona et al, 2021) for introducing artificial change-points in the training set. Another possible extension would be to adapt our framework to more general types of dynamic networks, e.g., snapshots with varying node sets or with missing edges. In certain application domains, it may well be the case that change-points phenomena are localized only in certain parts of the network (as considered in some of our synthetic experiments), and are not affecting the global structure. To this end, yet another interesting addition to the current framework is to be able to pinpoint specifically which part of the network is mainly driving the change-point, to enhance explainability.

Finally, testing the methodology on different types of networks, such as directed networks, is an interesting direction to explore, especially in the context of recent work in the literature that encodes various measures of causality or lead-lag associations in multivariate time series as directed graphs (Bennett et al, 2022; Run, 2019). The structure of such weighted directed graphs may evolve over time, which motivates the need for techniques for change-point detection, a setting where adapting traditional spectral methods for change-point detection would be challenging, due to the asymmetry of the adjacency matrix.

Statements and Declarations

Funding

MC acknowledges support from the EPSRC grant EP/N510129/1 at The Alan Turing Institute. XD gratefully acknowledges support from the Oxford-Man Institute of Quantitative Finance and the EPSRC (EP/T023333/1). DS is supported by the EPSRC and MRC Centre for Doctoral Training in Statistical Science, University of Oxford (grant EP/L016710/1). HK is supported by the EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines and Systems, University of Oxford (EP/L015897/1).

Conflicts of interest

The authors have no conflicts of interest to declare.

Ethics approval

The authors did not need ethics approval for this work.

Authors' contributions

All authors contributed to the writing of the paper. DS performed numerical experiments.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data, material and code

The data and code for running the experiments will be made publicly available.

References

(2019) Detecting causal associations in large nonlinear time series datasets. Science Advances 5(11)

Bao D, You K, Lin L (2018) Network distance based on laplacian flows on graphs. ArXiv abs/1810.02906

Barnett I, Onnela JP (2016) Change Point Detection in Correlation Networks. Scientific Reports 6(1):18,893. https://doi.org/10.1038/srep18893, URL https://doi.org/10.1038/srep18893

Bennett S, Cucuringu M, Reinert G (2022) Lead-lag detection and network clustering for multivariate time series with an application to the us equity market. KDD 2021 MileTS (preliminary workshop version) https://doi.org/10.48550/ARXIV.2201.08283

Bhamidi S, Jin J, Nobel A (2018) Change point detection in network models: Preferential attachment and long range dependence. The Annals of Applied Probability 28(1):35–78. URL https://www.jstor.org/stable/26542304

Bhattacharjee M, Banerjee M, Michailidis G (2020) Change point estimation in a dynamic stochastic block model. 1812.03090

Bourqui R, Gilbert F, Simonetto P, et al (2009) Detecting structural changes and command hierarchies in dynamic social networks. In: 2009 International conference on advances in social network analysis and mining, IEEE, pp 83–88

Bruna J, Li X (2017) Community detection with graph neural networks. stat 1050:27

Cai L, Chen Z, Luo C, et al (2021) Structural Temporal Graph Neural Networks for Anomaly Detection in Dynamic Graphs, Association for Computing Machinery, New York, NY, USA, p 3747–3756. URL https://doi.org/10.1145/3459637.3481955

Carmona CU, Aubet FX, Flunkert V, et al (2021) Neural contextual anomaly detection for time series. ArXiv abs/2107.07702

Chakraborti A, Sharma K, Pharasi HK, et al (2020) Phase separation and scaling in correlation structures of financial markets. Journal of Physics: Complexity 2(1):015,002

Chu L, Chen H (2018) Sequential change-point detection for high-dimensional and non-euclidean data. 1810.05973

Corneli M, Latouche P, Rossi F (2018) Multiple change points detection and clustering in dynamic networks. Statistics and Computing 28:989–1007

Cribben I, Yu Y (2017) Estimating whole-brain dynamics by using spectral clustering. Journal of the Royal Statistical Society Series C: Applied Statistics 66(3):607–627. https://doi.org/10.1111/rssc.12169, https://arxiv.org/abs/1509.03730

De Ridder S, Vandermarliere B, Ryckebusch J (2016) Detection and localization of change points in temporal networks with the aid of stochastic block models. Journal of Statistical Mechanics: Theory and Experiment 2016(11):113,302. https://doi.org/10.1088/1742-5468/2016/11/113302, URL http://dx.doi.org/10.1088/1742-5468/2016/11/113302

Derr T, Ma Y, Tang J (2018) Signed graph convolutional networks. In: 2018 IEEE International Conference on Data Mining (ICDM), IEEE, pp 929–934

Desobry F, Davy M, Doncarli C (2005) An online kernel change detection algorithm. IEEE Transactions on Signal Processing 53(8):2961–2974

Donnat C, Holmes S (2018) Tracking network dynamics: a survey of distances and similarity metrics. 1801.07351

Dubey P, Xu H, Yu Y (2021) Online network change point detection with missing values. 2110.06450

Dwivedi VP, Bresson X (2021) A generalization of transformer networks to graphs. 2012.09699

Dwivedi VP, Luu AT, Laurent T, et al (2022) Graph neural networks with learnable structural and positional representations. 2110.07875

Enikeeva F, Klopp O (2021) Change-point detection in dynamic networks with missing links. 2106.14470

Gallier J (2016) Spectral theory of unsigned and signed graphs. applications to graph clustering: a survey. 1601.04692

Ghoshdastidar D, Gutzeit M, Carpentier A, et al (2020) Two-sample hypothesis testing for inhomogeneous random graphs. The Annals of Statistics 48(4). https://doi.org/10.1214/19-aos1884, URL http://dx.doi.org/10.1214/

19-AOS1884

Gretton A, Borgwardt K, Rasch MJ, et al (2008) A kernel method for the two-sample problem. 0805.2368

Hamilton W, Ying Z, Leskovec J (2017) Inductive representation learning on large graphs. Advances in neural information processing systems 30

Harchaoui Z, Moulines E, Bach FR (2009) Kernel change-point analysis. In: Advances in neural information processing systems, pp 609–616

Hewapathirana IU, Lee D, Moltchanova E, et al (2020) Change detection in noisy dynamic networks: a spectral embedding approach. Social Network Analysis and Mining 10(1):1–22

Holme P, Saramäki J (2012) Temporal networks. Physics Reports 519(3):97–125. https://doi.org/10.1016/j.physrep.2012.03.001, URL http://dx.doi.org/10.1016/j.physrep.2012.03.001

Horváth A (2020) Sorted pooling in convolutional networks for one-shot learning. arXiv preprint arXiv:200710495

Huang J, Shen H, Hou L, et al (2019) Signed graph attention networks. In: International Conference on Artificial Neural Networks, Springer, pp 566–577

Huang S, Hitti Y, Rabusseau G, et al (2020) Laplacian Change Point Detection for Dynamic Graphs https://doi.org/10.1145/3394486.3403077, URL http://arxiv.org/abs/2007.01229%0Ahttp://dx.doi.org/10.1145/3394486.3403077, https://arxiv.org/abs/2007.01229

Kipf TN, Welling M (2016) Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:160902907

Koch GR (2015) Siamese neural networks for one-shot image recognition

Koutra D, Shah N, Vogelstein JT, et al (2016) Deltacon: Principled massive-graph similarity function with attribution. ACM Trans Knowl Discov Data 10:28:1–28:43

Kriege NM, Johansson FD, Morris C (2020) A survey on graph kernels. Applied Network Science 5(1). https://doi.org/10.1007/s41109-019-0195-3, URL http://dx.doi.org/10.1007/s41109-019-0195-3

Ktena SI, Parisot S, Ferrante E, et al (2017) Distance metric learning using graph convolutional networks: Application to functional brain networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, pp 469–477

Kumar S, Zhang X, Leskovec J (2019) Predicting dynamic embedding trajectory in temporal interaction networks. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining https://doi.org/10.1145/3292500.3330895, URL http://dx.doi.org/10.1145/3292500.3330895

Lee J, Lee I, Kang J (2019) Self-attention graph pooling. In: 36th International Conference on Machine Learning, ICML 2019. International Machine Learning Society (IMLS), 36th International Conference on Machine Learning, ICML 2019, pp 6661–6670, funding Information: This work was supported by the National Research Foundation of Korea (NRF-2017R1A2A1A17069645, NRF-2016M3A9A7916996, NRF-2017M3C4A7065887) Publisher Copyright: © 36th International Conference on Machine Learning, ICML 2019. All rights reserved.; 36th International Conference on Machine Learning, ICML 2019 ; Conference date: 09-06-2019 Through 15-06-2019

Li A, Cornelius SP, Liu YY, et al (2017) The fundamental advantages of temporal networks. Science 358(6366):1042–1046. https://doi.org/10.1126/science.aai7488, URL http://dx.doi.org/10.1126/science.aai7488

Li P, Wang Y, Wang H, et al (2020) Distance encoding: Design provably more powerful neural networks for graph representation learning. 2009.00142

Li Y, Gu C, Dullien T, et al (2019) Graph matching networks for learning the similarity of graph structured objects. In: ICML

Ling X, Wu L, Wang S, et al (2021) Multilevel graph matching networks for deep graph similarity learning. IEEE Transactions on Neural Networks and Learning Systems p 1–15. https://doi.org/10.1109/tnnls.2021.3102234, URL http://dx.doi.org/10.1109/TNNLS.2021.3102234

Liu J, Ma G, Jiang F, et al (2019) Community-preserving graph convolutions for structural and functional joint embedding of brain networks. In: 2019 IEEE International Conference on Big Data (Big Data), IEEE, pp 1163–1168

Liu Y, Pan S, Jin M, et al (2021) Graph self-supervised learning: A survey. arXiv preprint arXiv:210300111

Ma G, Ahmed NK, Willke T, et al (2019) Similarity learning with higher-order graph convolutions for brain network analysis. 1811.02662

Ma G, Ahmed NK, Willke TL, et al (2021) Deep graph similarity learning: A survey. Data Mining and Knowledge Discovery 35(3):688–725

Manessi F, Rozza A, Manzo M (2020) Dynamic graph convolutional networks. Pattern Recognition 97:107,000. https://doi.org/10.1016/j.patcog.2019.107000, URL http://dx.doi.org/10.1016/j.patcog.2019.107000

Mensink T, Verbeek J, Perronnin F, et al (2012) Metric learning for large scale image classification: Generalizing to new classes at near-zero cost. In: European Conference on Computer Vision, Springer, pp 488–501

Miller H, Mokryn O (2020) Size agnostic change point detection framework for evolving networks. Plos one 15(4):e0231,035

Nie L, Nicolae DL (2021) Weighted-graph-based change point detection

Ofori-Boateng D, Gel YR, Cribben I (2019) Nonparametric anomaly detection on time series of graphs. bioRxiv

Ondrus M, Olds E, Cribben I (2021) Factorized binary search: change point detection in the network structure of multivariate high-dimensional time series. https://doi.org/10.48550/ARXIV.2103.06347, URL https://arxiv.org/abs/2103.06347

Padilla OHM, Yu Y, Priebe CE (2019) Change point localization in dependent dynamic nonparametric random dot product graphs URL http://arxiv.org/abs/1911.07494, https://arxiv.org/abs/1911.07494

Peel L, Clauset A (2015) Detecting change points in the large-scale structure of evolving networks. In: Twenty-Ninth AAAI Conference on Artificial Intelligence

Ranshous S, Shen S, Koutra D, et al (2015) Anomaly detection in dynamic networks: a survey. Wiley Interdisciplinary Reviews: Computational Statistics 7(3):223–247

Reiss A, Stricker D (2012a) Creating and benchmarking a new dataset for physical activity monitoring. In: Proceedings of the 5th International Conference on PErvasive Technologies Related to Assistive Environments. Association for Computing Machinery, New York, NY, USA, PETRA '12, https://doi.org/10.1145/2413097.2413148, URL https://doi.org/10.1145/2413097.2413148

Reiss A, Stricker D (2012b) Introducing a new benchmarked dataset for activity monitoring. In: 2012 16th international symposium on wearable computers, IEEE, pp 108–109

Rossetti G, Cazabet R (2018) Community discovery in dynamic networks. ACM Computing Surveys 51(2):1–37. https://doi.org/10.1145/3172867, URL http://dx.doi.org/10.1145/3172867

Rossi E, Chamberlain B, Frasca F, et al (2020) Temporal graph networks for deep learning on dynamic graphs. 2006.10637

Samal A, Pharasi HK, Ramaia SJ, et al (2021) Network geometry and market instability. 2009.12335

Sankar A, Wu Y, Gou L, et al (2020) DySAT: Deep Neural Representation Learning on Dynamic Graphs via Self-Attention Networks, Association for Computing Machinery, New York, NY, USA, p 519–527. URL https://doi.org/10.1145/3336191.3371845

Seo Y, Defferrard M, Vandergheynst P, et al (2018) Structured sequence modeling with graph convolutional recurrent networks. In: Cheng L, Leung ACS, Ozawa S (eds) Neural Information Processing. Springer International Publishing, Cham, pp 362–373

Shervashidze N, Schweitzer P, van Leeuwen EJ, et al (2011) Weisfeiler-lehman graph kernels. J Mach Learn Res 12(null):2539–2561

Siglidis G, Nikolentzos G, Limnios S, et al (2020) Grakel: A graph kernel library in python. Journal of Machine Learning Research 21(54):1–5. URL http://jmlr.org/papers/v21/18-370.html

Skarding J, Gabrys B, Musial K (2021) Foundations and modeling of dynamic networks using dynamic graph neural networks: A survey. IEEE Access 9:79,143–79,168. https://doi.org/10.1109/access.2021.3082932, URL http://dx.doi.org/10.1109/ACCESS.2021.3082932

Trivedi R, Farajtabar M, Biswal P, et al (2019) Dyrep: Learning representations over dynamic graphs. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=HyePrhR5KX

Veličković P, Cucurull G, Casanova A, et al (2018) Graph attention networks. 1710.10903

Wang H, Tang M, Park Y, et al (2013) Locality statistics for anomaly detection in time series of graphs. IEEE Transactions on Signal Processing 62(3):703–717

Wang T, Samworth RJ (2018) High dimensional change point estimation via sparse projection. Journal of the Royal Statistical Society Series B: Statistical Methodology 80(1):57–83. https://doi.org/10.1111/rssb.12243, https://arxiv.org/abs/1606.06246

Wang T, Chen Y, Samworth R (2022) High-dimensional, multiscale online changepoint detection. Journal of the Royal Statistical Society Series B: Statistical Methodology

Wang Y, Chakrabarti A, Sivakoff D, et al (2017) Fast change point detection on dynamic social networks. 1705.07325

Wilson JD, Stevens NT, Woodall WH (2019) Modeling and detecting change in temporal networks via the degree corrected stochastic block model. Quality and Reliability Engineering International 35(5):1363–1378. https://doi.org/10.1002/qre.2520, https://arxiv.org/abs/1605.04049

Xu H, Feng Y, Chen J, et al (2018) Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications. Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18 https://doi.org/10.1145/3178876.3185996, URL http://dx.doi.org/10.1145/3178876.3185996

Xu K, Hu W, Leskovec J, et al (2019) How powerful are graph neural networks? 1810.00826

Yu Y, Padilla OHM, Wang D, et al (2021) Optimal network online change point localisation. 2101.05477

Zhang M, Cui Z, Neumann M, et al (2018) An end-to-end deep learning architecture for graph classification. In: Proceedings of the AAAI Conference on Artificial Intelligence

Zhang M, Wu S, Yu X, et al (2021) Dynamic graph neural networks for sequential recommendation. 2104.07368

Zhang R, Hao Y, Yu D, et al (2020) Correlation-aware unsupervised change-point detection via graph neural networks. 2004.11934

Zhao Z, Chen L, Lin L (2019) Change-point detection in dynamic networks via graphon estimation. 1908.01823

Zou C, Yin G, Feng L, et al (2014) Nonparametric maximum likelihood approach to multiple change-point problems. The Annals of Statistics 42(3). https://doi.org/10.1214/14-aos1210, URL http://dx.doi.org/10.1214/14-AOS1210

A Additional details and analysis in the synthetic experiments

In this section, we provide additional details on the hyperparameter selection procedure in the synthetic networks of Section 4.3, and two supplementary experiments, namely a sensitivity analysis of our method to the window size parameter L and to the choice of pooling layer in the similarity module (see Section 3).

A.1 Hyperparameter selection

In each scenario and difficulty level, we train our s-GNN over 100 epochs and and validate using the F1-score. We select one set of hyperparameters of the

s-GNN (i.e., learning rate, number of hidden units, dropout rate and size of Sort-k layer) per scenario by searching over a grid of values in one difficulty level, i.e., $p = 0.03$ in Scenario 1, $s = 60$ in Scenario 2, $p = 0.06$ in Scenario 3, and $h = 0.1$ in Scenario 4. For choosing the hyperparameters we test every configuration of values with the learning rate in the set $\{0.001, 0.01\}$, the dropout rate in $\{0.01, 0.05, 0.1\}$, the size of the Sort-k layer in $\{20, 40, 100\}$, the number of hidden units in $\{16, 32, 64\}$, and select the one with the highest F1-score on the validation set.

For each graph distance baseline d, we use the training and validation sets to choose a classification threshold θ such that the estimated label $\hat{y}_i = 1$ if $d(G_1^i, G_2^i) < \theta$ and $\hat{y}_i = 0$ otherwise (or reversely for the WL kernel).

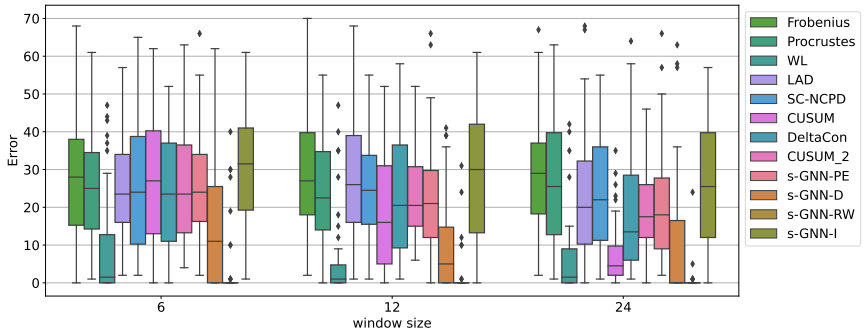
A.2 Sensitivity to the window size

We evaluate the sensitivity of our NCPD method and the baselines to the window size parameter L in the **"Merge"** scenario from Section 4.3. We recall that this hyperparameter corresponds to the amount of past (and future for some baselines) information needed to compute the NCPD statistic. It is therefore also the minimal distance between change-points that a method can detect. In this analysis, we test the performances of the methods using different window sizes in a a synthetic setting from Section 4.3. More precisely, we consider Scenario 1 ("Merge") and three difficulty levels ($p = 0.3, 0.4, 0.5$).

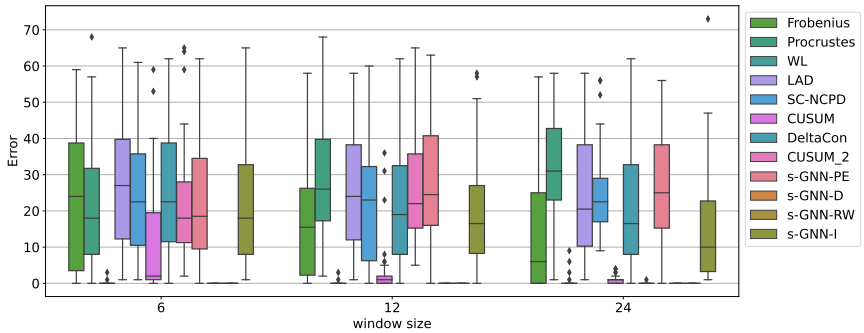
We report our findings in Figure 9. We note that our method is not very sensitive to the window size, in particular our best variant (**s-GNN-RW**) outperforms the baselines for all window sizes. We also remark that the two methods based on the CUSUM statistic (**CUSUM** and **CUSUM 2**) have better performances for larger L, and this effect is larger than for the other baselines. In conclusion, the choice of window size in our NCPD statistic (3) does not have a big impact on the performance of our method, and therefore does not require to be finely tuned.

A.3 Sensitivity to the pooling layer

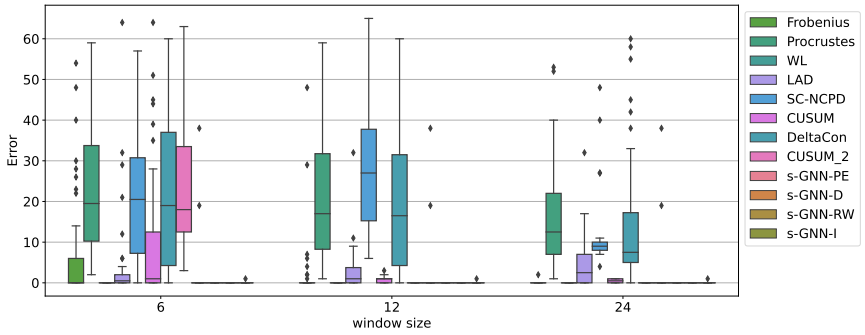
In this section, we test the importance of using a Sort-k pooling layer in our similarity module (Figure 1b). We consider the **"Birth 1"** scenario from Section 4.3 and compare the performance of our method with Sort-k pooling ($k = 100$) with the same method with Max or Average pooling. We report our findings in Figure 10. We note that Max pooling does not have good performances in this experiment and Average pooling has a higher variance than Sort-k, except in the last (and easiest setting). It may be due to the fact that Max pooling is less robust to the sparsity of the network than Sort-k and Average pooling, and that the latter cannot detect local changes in the graphs since it averages the displacement over the whole set of nodes. Therefore, we can conclude that Sort-k pooling is more adapted to detect small distribution changes, while being robust to the sparsity of edges.



(a) Difficult level



(b) Moderate level



(c) Easy level

Figure 9: Performances on the detection task in the **"Merge"** scenario for 3 window sizes $L = 6, 12, 24$ in three difficulty levels: difficult ($p = 0.3$) (a), moderate ($p = 0.4$) (b), easy ($p = 0.5$) (c).

B Additional results on the S&P 500 stock returns data set

In this section, we first report additional figures illustrating the procedure for pre-estimating change-points in the training and validation sequences of the dynamic correlation network; secondly we analyse the network using the eigen-entropy $H(t)$, as previously done in Chakraborti et al (2020) on similar data.

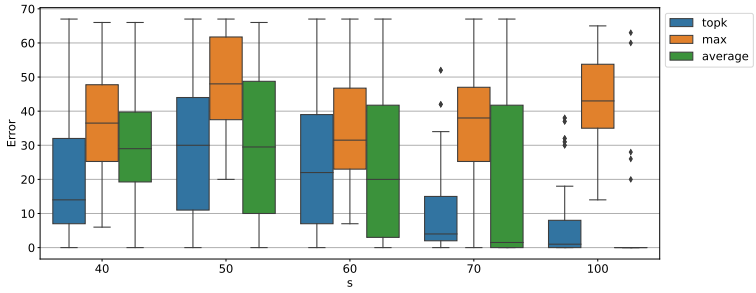


Figure 10: Localisation error of the s-GNN in the **"Birth 1"** scenario with different choices of pooling layers in the similarity module, namely Sort-k, Max and Average pooling.

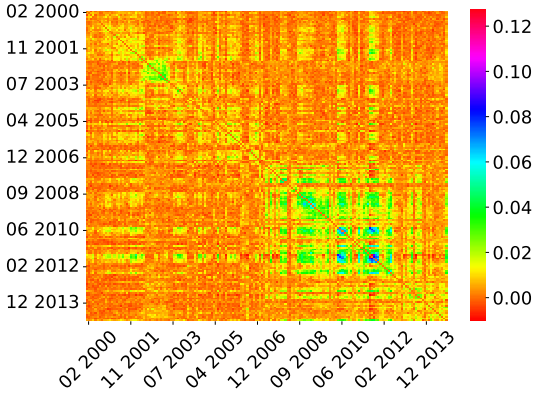
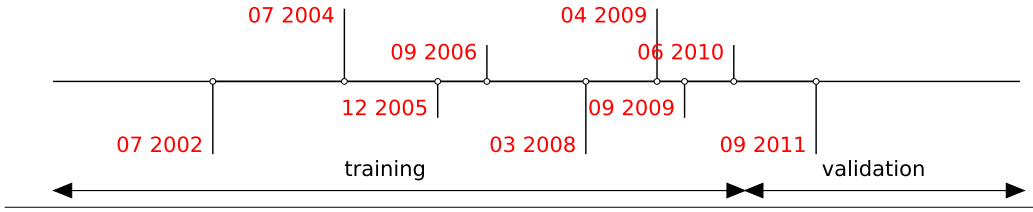


Figure 11: Matrix of Adjusted Rand Index values between the partitions obtained for each pair of graph snapshots in the correlation network of S&P 500 stock returns. The first two digits denote the month, followed by the year.

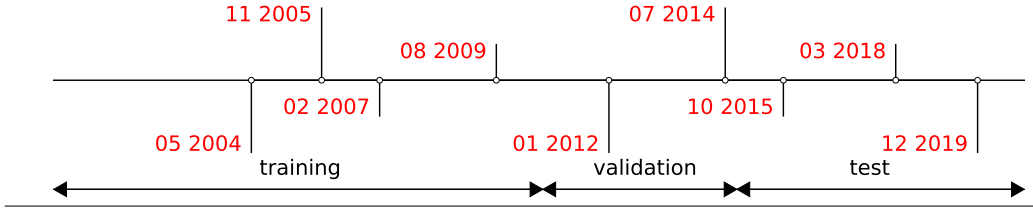
This latter analysis also gives an insight on the possible market phases (i.e., the period in-between our estimated change-points).

In Figure 11, we plot the heatmap of the similarity matrix between the graph snapshots in the training and validation sequences. We recall from Section 4.4 that the similarity score between pairs of snapshots is measured in terms of the Adjusted Rand Index values between the stock partitions obtained for each snapshots. We note that this similarity matrix seems to have a cluster structure; in particular, high similarity scores can be found during the period of the financial crisis from 2007 to 2011 and in 2001-2002. The clustering procedure of this matrix with spectral clustering and a post-processing step (see 4.4) leads to the pre-estimated change-points plotted in Figure 12a.

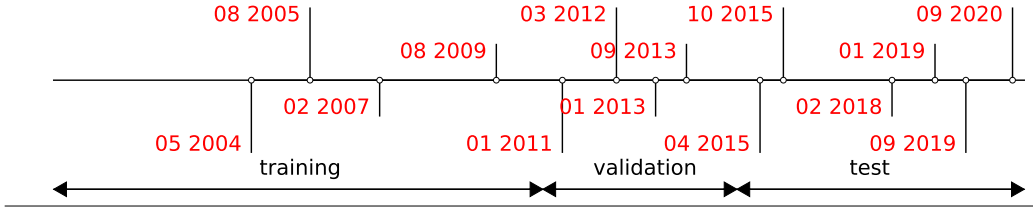
Moreover, we reproduce the analysis of the correlation graphs with the eigen-entropy $H(t)$ from Chakraborti et al (2020). The eigen-entropy of a graph is defined as the entropy of the eigen-centrality vector, which is a L_1-normalised version of the principal eigenvector of the graph adjacency matrix. This principal eigenvector is related to the relative ranks of the different stocks in the market



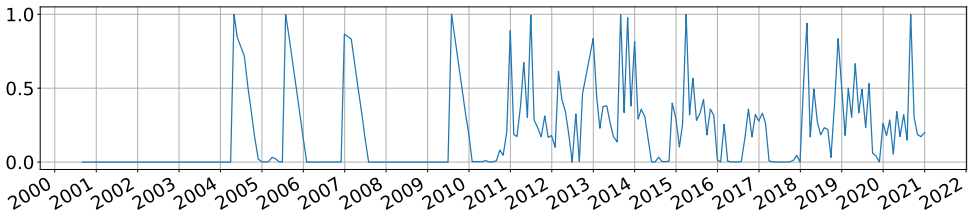
(a) Pre-estimated change-points on the training and validation sequences.



(b) Change-points estimated by our method on the training, validation and test sequences.



(c) Change-points estimated by our method with the **Identity encoding**



(d) Network change-point statistic obtained with our method with the **Identity encoding**.

Figure 12: Change-points estimated on the S&P 500 stock returns correlation network by the pre-estimation procedure described in Section 4.4 (a) and by our method on the attributed data (b) and on the non-attributed data (c). In the latter case, we have used the **Identity encoding** as synthetic attributes and the obtained network change-point statistic is reported in the last panel (d).

and its entropy measures the market "disorder". The correlation matrices $\boldsymbol{C}(t)$'s can be further decomposed into a market mode (principal eigen matrix) $\boldsymbol{C}(t)_M$ and a composite group plus random mode $\boldsymbol{C}(t)_{GR}$ and their corresponding eigen-entropy ($H_M(t), H_{GR}(t)$ can be also computed (see Figure 13). This allows to define a 3D-phase space where the graphs can be separated into types (e.g. market anomalies, crashes, normal behaviour or highest disorder in Chakraborti

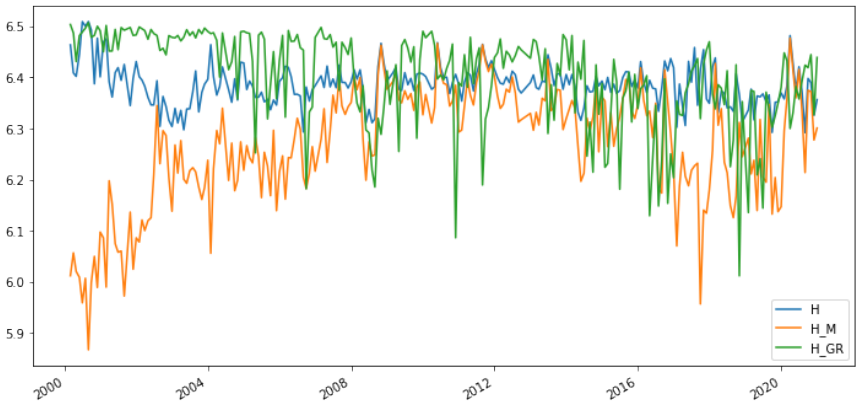


Figure 13: Eigen-entropy of the correlation graph, the market mode and the group plus random mode over time.

et al (2020)). A 2D visualisation of our graphs with their corresponding labels is given in Figure 15a. The distribution of the average eigen-centrality vectors in each class of graphs also indicates that the correlation structure changes in the different phases (see Figure 14). Chakraborti et al (2020) also observe a scaling behaviour by comparing the absolute entropy difference $|H - H_{GR}|$ and the mean market correlation (see Figure 15b).

C Additional experimental results on the physical activity monitoring data

C.1 Similarity between activities and subjects

Additional results on the similarity between activities and subjects are presented in Figure 16, Figure 17, Figure 18, and Figure 19.

C.2 Sensitivity to the tolerance level

In this section, we investigate the sensitivity of our results presented in Table 4 to the tolerance level chosen to compute the adjusted F1-score Xu et al (2018). We consider the **Random split** experiment in the **Cross-Individual** task from Section 4.5, and we reproduce this experiment for different levels of tolerance $\text{tol} = \{1, 3, 5, 7\}$. We report the numerical results in Table 5. We note that our method has the best performance for all considered levels.

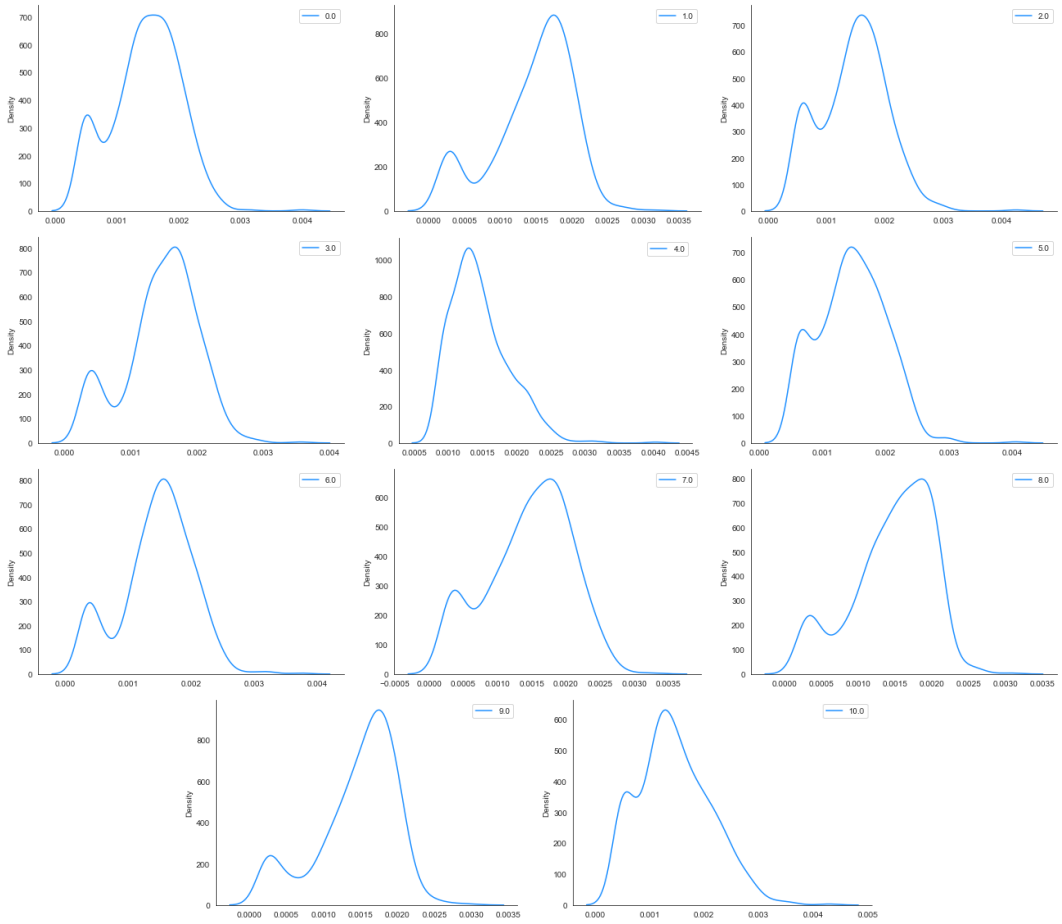
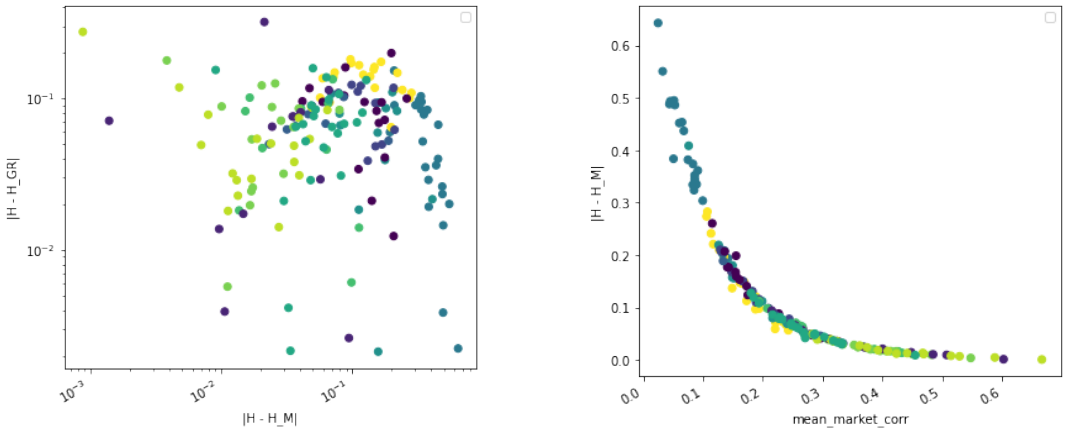


Figure 14: Histogram of the values in the average eigen-centrality vector of each class (phase) of graphs (the subplots correspond to different classes).



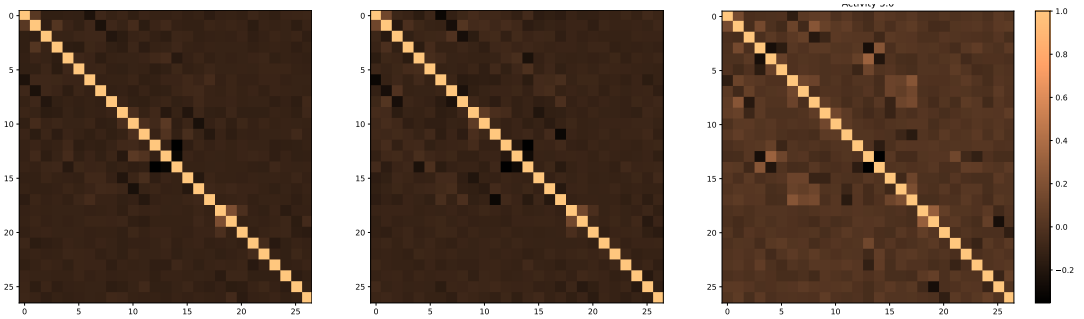
(a) Entropy differences $|H - H_{GR}|$ versus $|H - H_M|$.

(b) Entropy differences $|H - H_M|$ versus mean market correlation.

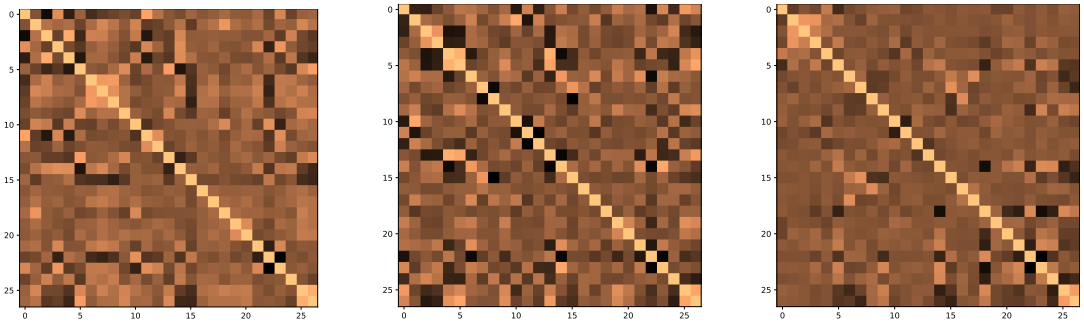
Figure 15: Entropy differences $|H - H_{GR}|$ versus $|H - H_M|$ (in log scale) (a) and entropy differences $|H - H_M|$ versus mean market correlation (b) for the graphs in the financial correlation dynamic network. The colors indicate the class of the graphs in our partition.

Cross-individual NCPD					
Tolerance	s-GNN-I	Frobenius	SC-NCPD	CUSUM	CUSUM 2
1	**0.60 (0.30)**	0.53 (0.22)	0.43 (0.32)	0.33 (0.24)	0.42 (0.30)
3	**0.87 (0.25)**	0.68 (0.20)	0.70 (0.31)	0.53 (0.24)	0.76 (0.29)
5	**0.61 (0.27)**	0.53 (0.20)	0.41 (0.32)	0.27 (0.22)	0.44 (0.31)
7	**0.85 (0.28)**	0.71 (0.21)	0.71 (0.30)	0.56 (0.25)	0.75 (0.29)

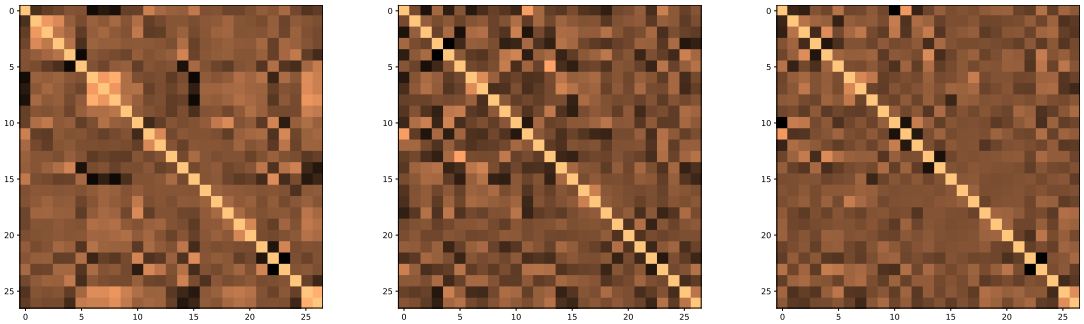
Table 5: Adjusted F1-score of our method (s-GNN) and baselines in the **Cross-individual** NCPD task and the random split setting on the physical activity monitoring data. The bold values in each row denote the top performing method. The values in the parentheses denote the standard deviation over 10 repetitions of the random splits train/validation/set. We remark that different rows corresponding to different tolerance levels essentially amounts to defining a different set of ground truth change-points, and hence we should not necessarily expect a monotonic relationship between tolerance versus the recovery accuracy of all methods; the main take-away message here is that the s-GNN method attains superior performance when compared to other baselines, for the same tolerance level.



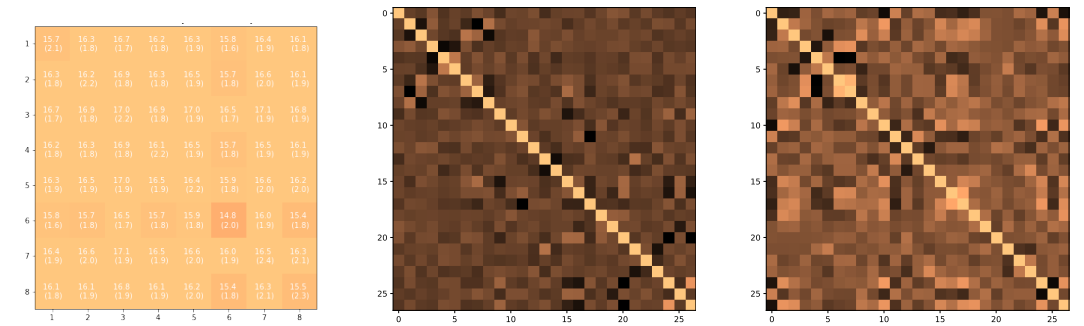
(a) Activity 1 (b) Activity 2 (c) Activity 3



(d) Activity 4 (e) Activity 5 (f) Activity 6

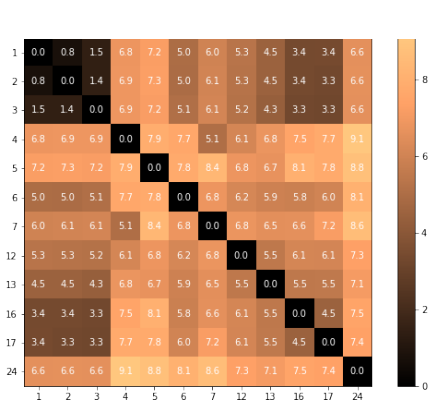


(g) Activity 7 (h) Activity 12 (i) Activity 13

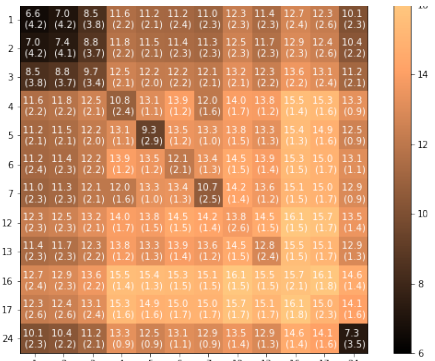


(j) Activity 16 (k) Activity 17 (l) Activity 24

Figure 16: Average adjacency matrices for each of the 12 activities performed by Subject 1.

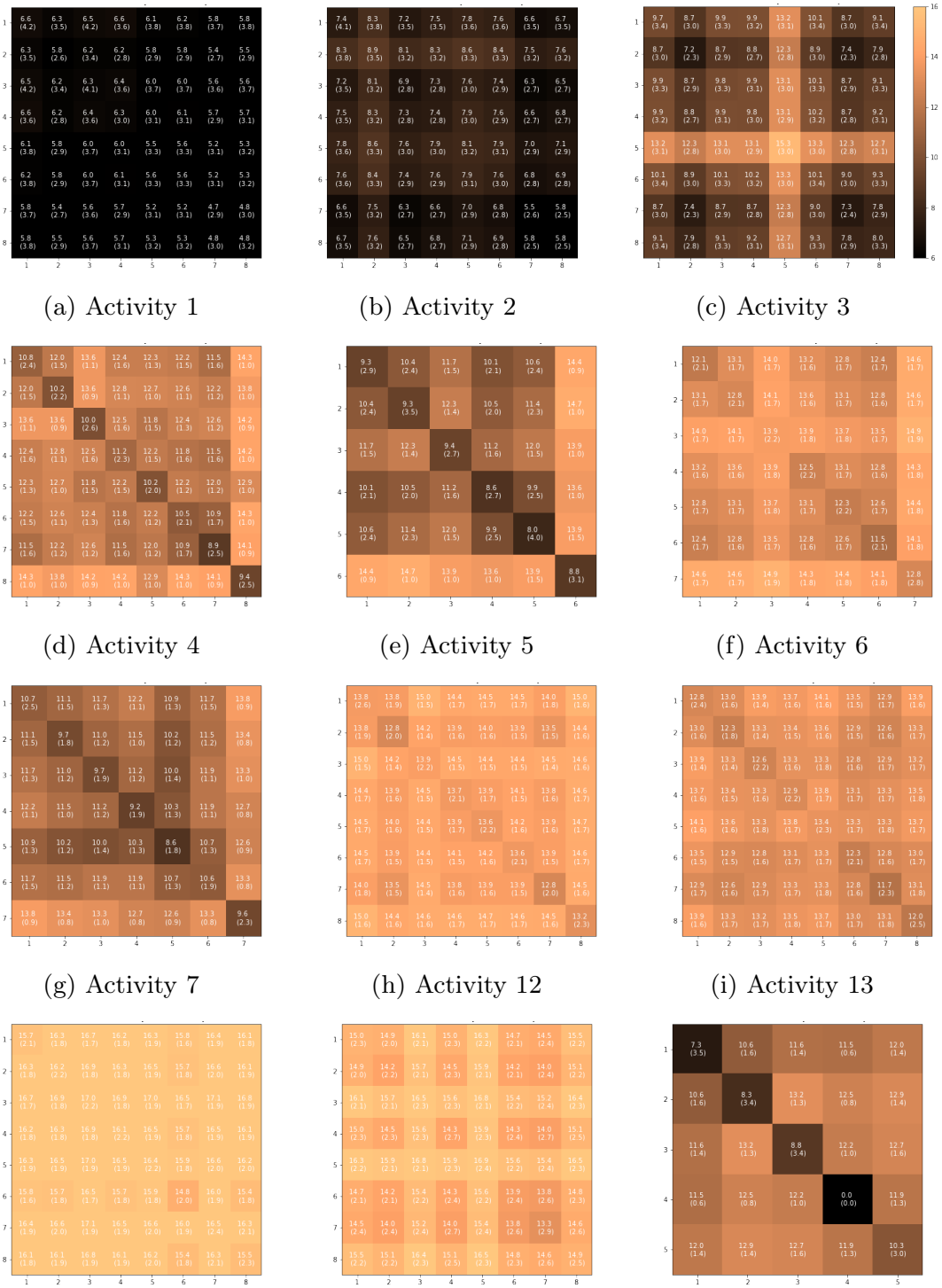


(a) Distance between average adjacency matrices per activity.



(b) Average (and standard deviation) distance between graphs grouped by activities.

Figure 17: Frobenius distance between graphs in the dynamic network corresponding to Subject 1.



(a) Activity 1 (b) Activity 2 (c) Activity 3

(d) Activity 4 (e) Activity 5 (f) Activity 6

(g) Activity 7 (h) Activity 12 (i) Activity 13

(j) Activity 16 (k) Activity 17 (l) Activity 24

Figure 18: Average (and standard deviation) Frobenius distance between the graphs grouped by subjects for each activity label.

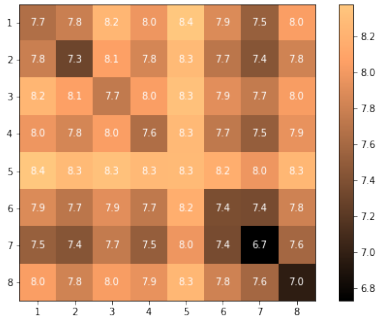


Figure 19: Average Frobenius distance between graphs with the same label grouped by subjects.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Graph similarity learning for change-point detection in dynamic networks
Publication Status	□Published □ Accepted for Publication X Submitted for Publication □Unpublished and unsubmitted work written in a manuscript style
Publication Details	Joint work with Henry Kenlay, Professor Mihai Cucuringu (University of Oxford) and Professor Xiaowen Dong (University of Oxford). Submitted to the Machine Learning Journal.

Student Confirmation

Student Name:	Deborah Sulem		
Contribution to the Paper	I am the first author of this paper. I have proposed the method and the deep-learning architecture, implemented and conducted the numerical experiments.		
Signature *Deborah Sulem*		Date	08/11/2022

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Professor Mihai Cucuringu		
Supervisor comments I confirm that Deborah has made a substantial contribution in line with her description above.		
Signature 	Date	9 Nov 2022

This completed form should be included in the thesis, at the end of the relevant chapter.

6 | Conclusion

This closing chapter first provides a brief summary of the previous chapters, complemented by a critical analysis of their limitations and a review of future work perspectives. Then, I conclude this manuscript with some personal thoughts on the current state-of-the-art research on temporal point processes and graphs.

6.1 Summary of the thesis

This manuscript contains four independent works that tackle the modelling challenges of discrete data and interactive phenomena through the prism of temporal point processes and graphs. Each work comprises novel results, methodology, and/or a numerical study of the statistical method. In Chapters 2 and 3, the inference approach is Bayesian, nonparametric, and based on a temporal point process model. In Chapters 4 and 5, the statistical method is model-free and built respectively from graph spectral and deep learning algorithms. Nonetheless, random graph models are leveraged in the latter works, for deriving theoretical guarantees, and/or validating our method.

In Chapter 2, a general class of nonlinear Hawkes processes has been analysed in the Bayesian nonparametric framework. This flexible model allows to capture *causal*, *excitating* and *inhibiting* interactions between entities, and is commonly applied to neurons spikes data. In this chapter, we have notably established the concentration and consistency properties of the posterior distribution under reasonable and easily verifiable assumptions on the prior distribution and the true model. To prove these results, we have built new technical tools based on the concept of *excursions*, which lead to an elegant analysis of the statistical properties of Hawkes processes, and potentially other regenerative point processes. Moreover, we have exemplified our general results and provided explicit posterior concentration rates for Hölder-smooth classes of functions and three families of nonparametric priors.

In Chapter 3, we have extended some of the aforementioned results to the variational posterior distribution, under mild conditions on the prior distribution, the model, and the variational class. We have applied our general results to existing methods, including mean-field variational inference, and introduced a spike-and-slab variational family to induce sparsity. Moreover, we propose a novel adaptive and sparsity-inducing method via a model-selection approach, for which we have designed practical algorithms in the sigmoid Hawkes model. In particular, we propose an efficient

two-step procedure that can scale up to high-dimensional processes. We provided an extensive set of numerical experiments on simulated data to test our variational Bayes algorithms, and showed the comparative advantages of our approach over Monte-Carlo Markov Chains methods.

In Chapter 4, we have theoretically analysed the signed graph clustering problem and the performance of spectral algorithms. We have first provided a thorough study of the SPONGEsym and symmetric signed Laplacian methods. Then we have proposed and similarly analysed a regularisation strategy for these algorithms. For graphs generated from a signed stochastic block model, we have notably proved that these algorithms are able to recover the latent clusters under certain favorable noise regimes, and under two regimes of edge sparsity, in particular the challenging sparse graph regime. We have also empirically shown on simulated and real-world networks that regularised spectral methods can perform well in sparse graphs, provided that the regularisation parameters are suitably chosen.

In Chapter 5, we have proposed a novel methodology for detecting abrupt distribution changes in discrete-time dynamic networks. Our method relies on learning a graph similarity function from data, to effectively compare pairs of graphs. We have designed a novel, quite parsimonious, and modular siamese graph neural network model that we tested on synthetic experimental settings of dynamic stochastic block models, and on two real-world data sets of correlation networks. We have empirically demonstrated that our method can adapt to different network distributions and types of change points, after an adequate training procedure. In particular, we have empirically shown that learning the similarity function allows to more effectively compare the current graph to its short-term history, and therefore to detect displacements with a simple online statistic.

6.2 Limitations and perspectives for future work

In this section, I discuss certain limitations of the works comprised in this thesis, then propose ways of extending our results and improving our methods.

6.2.1 On nonlinear Hawkes processes

In the results of Chapter 2, we cover some nonlinear Hawkes models commonly used in applications, in particular the ones with the sigmoid and softplus functions as the nonlinearity. However, we have only obtained partial results in the more challenging ReLU Hawkes model, which is a direct

extension of the linear Hawkes model with signed interaction functions. In this model, when some interaction functions are negative, the conditional intensity function $\lambda(t|\mathcal{G}_t)$ can be null for some t. In fact, we have established posterior concentration rates in the shifted ReLU model, and in the standard ReLU model under a strong, and not easily verified, additional assumption - needed in the verification of the Kullback-Leibler condition.

More generally, estimating the link functions or their parameter is an open question in the nonlinear Hawkes model. In our analysis, we have always assumed, except in the shifted ReLU and the sigmoid models, that these nonlinearities were known, i.e., chosen *a-priori* by the statistician. Besides, our asymptotic study of the nonlinear Hawkes model does not provide an insight on how the difficulty of the estimation problem changes depending on these link functions. However, the link functions crucially determine the property of the temporal point process, in particular the positivity and boundedness properties of its intensity function. A practitioner applying the Hawkes model for the first time might not know which one is more suitable to their particular data set. Therefore, they might be wishing for a more flexible inference method, which would also estimate the nonlinearity, or choose the best one amongst a set of predefined types (e.g., ReLU, sigmoid, softplus, exponential,...).

Although the core of our contributions on the Hawkes model is mainly theoretical, Chapter 3 partly tackles the computational challenges of implementing and deploying Bayesian nonparametric inference methods. Several practical questions on the computation of the posterior distribution or an approximate distribution could be further explored. In particular, the computation limits of our two-step mean-field variational inference algorithm need to be empirically tested, both in terms of dimensionality and detection threshold. Moreover, future work could be dedicated to designing efficient variational Bayes algorithms for nonlinear Hawkes models beyond the sigmoid model. In addition, developing faster and better MCMC methods would also be of interest for improving coverage. For instance, the reversible jump approach of Donnet et al. (2020) for the linear model, or more general nonparametric MCMC methods such as the Hamiltonian sampler of Mak et al. (2021), could be adapted to the nonlinear Hawkes model.

Until now, we have considered temporal point processes with a fixed dimension K. Nonetheless, the high-dimensional setting where $K \to \infty$ jointly with $T \to \infty$, is also of interest for applications of the Hawkes model in social network analysis and neuroscience. One perspective for future work is therefore to analyse the properties of Bayesian nonparametric methods in this setting, in a

suitable sparsity regime of the connectivity graph. Another practical task would be to parallelise the computation of the posterior distribution, since each of the K factors can be independently computed on a machine.

Finally, finding the minimax rate of estimation in the nonlinear Hawkes model, in the spirit of Reynaud-Bouret and Schbath (2010) for the linear univariate model, would guarantee the optimality of our estimators, as well as penalised projection estimators. From our analysis, we conjecture that the minimax rates in the nonlinear and linear models are not different. We note that this theory cannot be directly derived from the techniques of Reynaud-Bouret and Schbath (2010), which fundamentally rely on the linearity of the conditional intensity. We are currently studying this problem using our novel technical tools.

6.2.2 On signed and temporal graphs

In Chapter 4, the spectral algorithms for signed graph clustering have been studied in a simple signed stochastic block model. One could therefore extend our results to more general stochastic block models such as the degree-corrected stochastic block model, that includes degree-heterogeneity, or the setting of polarized communities (Bonchi et al., 2019). In addition, extensions to directed, attributed, or temporal graphs are well motivated by real world applications involving signed networks.

Moreover, an open theoretical question on our regularisation strategy relates to the choice of the positive and negative regularisation parameters γ_+, γ_-. Providing a data-driven approach to tune these parameters would be of interesting in practical applications with very sparse networks. Another interesting future line of work would be to adapt and compare the latest graph regularisation techniques based on powers of adjacency matrices by Stephan and Massoulié (2019); Abbe et al. (2020).

In Chapter 5, the proposed methodology for network change point detection relies on training a deep-learning algorithm in a supervised way, and therefore hinges on the availability of labelled data. This can be a limitation of learning-based methods in practice since dynamic network data is rarely annotated with ground-truth change points. Consequently, developing unsupervised or self-supervised learning procedures, for instance based on contrastive learning (Johnson et al., 2022) or on data augmentation strategies (Carmona et al., 2021) could be an interesting improvement of our method.

Moreover, one inconvenient of our graph similarity learning algorithm is the lack of interpretability of the similarity score, and therefore of the detected change points. Therefore, yet another interesting addition to the current framework is to enhance the explainability and the trustworthiness of the method. In particular, practitioners may be interested in understanding which specific part of the network is mainly driving the dissimilarity and/or a detected change point.

Lastly, the simple online statistic used in our network change-point detection methodology may not be optimal in settings where some delay in the detection of change point is acceptable. In these cases, computing a two-sample statistic such as in the maximum mean discrepancy (Gretton et al., 2006) or the Fisher discriminant ratio (Harchaoui et al., 2008), could potentially increase the detection power of our method.

6.3 Concluding remarks

Jointly studying temporal point process models and graphs in this thesis has allowed me to address inference problems with structure and dependence through distinct and complementary perspectives. Besides, it has connected me with different scientific communities.

On the one hand, the framework of the Hawkes model for analysing event data may be considered restrictive, or even out-dated, given the popularity of neural point processes models, and more generally, of model-free inference methods. However, the main asset of the Hawkes model lies in the interpretability of its parameter, and the causality structure between its components. Moreover, it provides a sound probabilistic framework for studying event observations. Undoubtedly, parametric methods in this model are still predominant amongst practitioners, possibly because of the current availability of packages, and the computational complexity of nonparametric methods.

Moreover, it is worth stressing that analysing event data, possibly with some covariates, is a very challenging inference set-up. It is often the case that events are wrongly reported or missing, and that only a subset of components of the phenomenon are observed. For instance, neuroscientists often record spike trains of few tens or hundreds of neurons in brains, which contain several billions of neurons in total. Consequently, for this noisy data, Bayesian methods have the benefits of providing uncertainty bounds, and our theory shows that now standard and well-studied nonparametric priors can be used in this context. Nonetheless, a great advance for Bayesian nonparametric methods on point processes would be the availability of a standard package.

On the other hand, our approaches in the graph-structured data problems start from a task and the design of an algorithm to solve it, instead of a data generative model. However, the utility and applicability of learnt algorithms are mainly driven by empirical proof-of-tests in varied contexts and on diverse data sources. This implies that statisticians developing these methods need access to “good" data and efficient computational resources for testing their methods, which can be the source of disparities amongst the machine learning research community.

Finally, there is a diversity of research approaches on networks and graphs, which interestingly attracts researchers from different backgrounds, e.g., mathematics, physics, computer science, and statistics. Exchanges between with the different communities in this field is therefore a great opportunity to broaden the perspectives in network science.

Bibliography

Abbe, E., Bandeira, A. S., Bracher, A., and Singer, A. (2014). Decoding binary node labels from censored edge measurements: Phase transition and efficient recovery. *IEEE Transactions on Network Science and Engineering*, 1(1):10–22.

Abbe, E., Boix-Adserà, E., Ralli, P., and Sandon, C. (2020). Graph powering and spectral robustness. *SIAM Journal on Mathematics of Data Science*, 2(1):132–157.

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.

Adams, R. P., Murray, I., and MacKay, D. J. (2009). Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. In *Proceedings of the 26th annual international conference on machine learning*, pages 9–16.

Aghabozorgi, S., Shirkhorshidi, A. S., and Wah, T. Y. (2015). Time-series clustering–a decade review. *Information Systems*, 53:16–38.

Amini, A. A., Chen, A., Bickel, P. J., and Levina, E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41(4):2097–2122.

Apostolopoulou, I., Linderman, S., Miller, K., and Dubrawski, A. (2019). Mutually regressive point processes. *Advances in Neural Information Processing Systems*, 32.

Aref, S. and Wilson, M. C. (2017). Measuring partial balance in signed networks. *Journal of Complex Networks*, 6(4):566–595.

Arinik, N., Figueiredo, R., and Labatut, V. (2019). Multiple partitioning of multiplex signed networks: Application to European parliament votes. *Social Networks*, 60:83–102.

Bacry, E., Delattre, S., Hoffmann, M., and Muzy, J.-F. (2013). Some limit theorems for Hawkes processes and application to financial statistics. *Stochastic Processes and their Applications*, 123(7):2475–2499.

Bacry, E. and Muzy, J.-F. (2014). Second order statistics characterization of Hawkes processes and non-parametric estimation. *arXiv preprint arXiv:1401.0903*.

Bansal, N., Blum, A., and Chawla, S. (2004). Correlation clustering. *Machine learning*, 56(1):89–113.

Bao, W. and Michailidis, G. (2018). Core community structure recovery and phase transition detection in temporally evolving networks. *Scientific Reports*, 8(1):1–16.

Barnett, I. and Onnela, J.-P. (2016). Change Point Detection in Correlation Networks. *Scientific Reports*, 6(1):18893.

Bartoszynski, R., Brown, B. W., McBride, C. M., and Thompson, J. R. (1981). Some nonparametric techniques for estimating the intensity function of a cancer related nonstationary Poisson process. *The Annals of Statistics*, pages 1050–1060.

Belardo, F., Cioabă, S. M., Koolen, J. H., and Wang, J. (2019). Open problems in the spectral theory of signed graphs. *arXiv preprint arXiv:1907.04349*.

Belitser, E., Serra, P., and van Zanten, H. (2015). Rate-optimal Bayesian intensity smoothing for inhomogeneous poisson processes. *Journal of Statistical Planning and Inference*, 166:24–35. Special Issue on Bayesian Nonparametrics.

Bhattacharjee, M., Banerjee, M., and Michailidis, G. (2020). Change point estimation in a dynamic stochastic block model. *Journal of Machine Learning Research*, 51.

Blundell, C., Beck, J., and Heller, K. A. (2012). Modelling reciprocating relationships with Hawkes processes. *Advances in neural information processing systems*, 25.

Bonchi, F., Galimberti, E., Gionis, A., Ordozgoiti, B., and Ruffo, G. (2019). Discovering polarized communities in signed networks. In *Proceedings of the 28th acm international conference on information and knowledge management*, pages 961–970.

Bonnet, A., Herrera, M. M., and Sangnier, M. (2022). Inference of multivariate exponential Hawkes processes with inhibition and application toneuronal activity. *arXiv preprint arXiv:2205.04107*.

Brémaud, P. and Massoulié, L. (1996). Stability of nonlinear Hawkes processes. *The Annals of Probability*, pages 1563–1588.

Brémaud, P. and Massoulié, L. (2001). Hawkes branching point processes without ancestors. *Journal of Applied Probability*, 38(1):122–135.

Bruna, J., Zaremba, W., Szlam, A., and LeCun, Y. (2013). Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*.

Cai, B., Zhang, J., and Guan, Y. (2022). Latent network structure learning from high dimensional multivariate point processes. *Journal of the American Statistical Association*, pages 1–40.

Cai, C. and Wang, Y. (2020). A note on over-smoothing for graph neural networks. *arXiv preprint arXiv:2006.13318*.

Carmona, C. U., Aubet, F.-X., Flunkert, V., and Gasthaus, J. (2021). Neural contextual anomaly detection for time series. *arXiv preprint arXiv:2107.07702*.

Carstensen, L., Sandelin, A., Winther, O., and Hansen, N. R. (2010). Multivariate Hawkes process models of the occurrence of regulatory elements. *BMC bioinformatics*, 11(1):456.

Cartwright, D. and Harary, F. (1956). Structural balance: a generalization of heider's theory. *Psychological review*, 63(5):277.

Chen, S., Shojaie, A., Shea-Brown, E., and Witten, D. (2017). The multivariate Hawkes process in high dimensions: Beyond mutual excitation. *arXiv preprint arXiv:1707.04928v2*.

Chevallier, J. (2017). Mean-field limit of generalized Hawkes processes. *Stochastic Processes and their Applications*, 127(12):3870–3912.

Chiang, K.-Y., Natarajan, N., Tewari, A., and Dhillon, I. S. (2011). Exploiting longer cycles for link prediction in signed networks. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1157–1162.

Chiang, K.-Y., Whang, J. J., and Dhillon, I. S. (2012). Scalable clustering of signed networks using balance normalized cut. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, page 615–624, New York, NY, USA. Association for Computing Machinery.

Christgau, A. M., Petersen, L., and Hansen, N. R. (2022). Nonparametric conditional local independence testing. *arXiv preprint arXiv:2203.13559*.

Corneli, M., Latouche, P., and Rossi, F. (2018). Multiple change points detection and clustering in dynamic networks. *Statistics and Computing*, 28:989–1007.

Costa, M., Graham, C., Marsalle, L., and Tran, V. C. (2020). Renewal in Hawkes processes with self-excitation and inhibition. *Advances in Applied Probability*, 52(3):879–915.

Costantini, G. and Perugini, M. (2014). Generalization of clustering coefficients to signed correlation networks. *PloS one*, 9(2):e88669.

Cribben, I. and Yu, Y. (2017). Estimating whole-brain dynamics by using spectral clustering. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(3):607–627.

Cucuringu, M. (2015). Synchronization over z2 and community detection in signed multiplex networks with constraints. *Journal of Complex Networks*, 3(3):469–506.

Cucuringu, M., Davies, P., Glielmo, A., and Tyagi, H. (2019). SPONGE: A generalized eigenproblem for clustering signed networks. In *Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1088–1098. PMLR.

Cunningham, J. P., Shenoy, K. V., and Sahani, M. (2008). Fast Gaussian process methods for point process intensity estimation. In *Proceedings of the 25th international conference on Machine learning*, pages 192–199.

Da Fonseca, J. and Zaatour, R. (2014). Hawkes process: Fast calibration, application to trade clustering, and diffusive limit. *Journal of Futures Markets*, 34(6):548–579.

Daley, D. J. and Vere-Jones, D. (2007). *An introduction to the theory of point processes: volume II: general theory and structure*. Springer Science & Business Media.

Dall'Amico, L., Couillet, R., and Tremblay, N. (2021). A unified framework for spectral clustering in sparse graphs. *Journal of Machine Learning Research*, 22:217–1.

Davis, C. and Kahan, W. M. (1970). The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46.

Davis, J. A. (1967). Clustering and structural balance in graphs. *Human Relations*, 20(2):181–187.

Delattre, S., Fournier, N., and Hoffmann, M. (2014). High dimensional Hawkes processes. *arXiv preprint arXiv:1403.5764*.

Deutsch, I. and Ross, G. J. (2020). Abc learning of Hawkes processes with missing or noisy event times. *arXiv preprint arXiv:2006.09015*.

Deutsch, I. and Ross, G. J. (2022). Bayesian estimation of multivariate Hawkes processes with inhibition and sparsity. *arXiv preprint arXiv:2201.05009*.

Didelez, V. (2008). Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):245–264.

Donnet, S., Rivoirard, V., and Rousseau, J. (2020). Nonparametric Bayesian estimation for multivariate Hawkes processes. *The Annals of Statistics*, 48(5):2698 – 2727.

Donnet, S., Rivoirard, V., Rousseau, J., and Scricciolo, C. (2017). Posterior Concentration Rates for Counting Processes with Aalen Multiplicative Intensities. *Bayesian Analysis*, 12(1):53 – 87.

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.

Dubey, M., Palakkadavath, R., and Srijith, P. (2021a). Bayesian neural Hawkes process for event uncertainty prediction. *arXiv preprint arXiv:2112.14474*.

Dubey, P., Xu, H., and Yu, Y. (2021b). Online network change point detection with missing values. *arXiv preprint arXiv:2110.06450*.

Durante, D. and Dunson, D. B. (2014). Nonparametric bayes dynamic modelling of relational data. *Biometrika*, 101(4):883–898.

Duval, C., Luçon, E., and Pouzat, C. (2022). Interacting Hawkes processes with multiplicative inhibition. *Stochastic Processes and their Applications*, 148:180–226.

Easley, D. and Kleinberg, J. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge university press.

Eichler, M., Dahlhaus, R., and Dueck, J. (2017). Graphical modeling for multivariate Hawkes processes with nonparametric link functions. *Journal of Time Series Analysis*, 38(2):225–242.

Enguehard, J., Busbridge, D., Bozson, A., Woodcock, C., and Hammerla, N. (2020). Neural temporal point processes for modelling electronic health records. In *Machine Learning for Health*, pages 85–113. PMLR.

Enikeeva, F. and Klopp, O. (2021). Change-point detection in dynamic networks with missing links. *arXiv preprint arXiv:2106.14470*.

Erny, X., Löcherbach, E., and Loukianova, D. (2022). Mean field limits for interacting Hawkes processes in a diffusive regime. *Bernoulli*, 28(1):125–149.

Estrada, E. and Benzi, M. (2014). Walk-based measure of balance in signed networks: Detecting lack of balance in social networks. *Physical Review E*, 90(4):042802.

Etesami, J., Kiyavash, N., Zhang, K., and Singhal, K. (2016). Learning network of multivariate Hawkes processes: A time series approach. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, page 162–171.

Euler, L. (1741). Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, pages 128–140.

Fagiolo, G. and Mastrorillo, M. (2013). International migration network: Topology and modeling. *Physical Review E*, 88(1):012812.

Flaxman, S., Teh, Y. W., and Sejdinovic, D. (2017). Poisson intensity estimation with reproducing kernels. *Electronic Journal of Statistics*, 11(2):5081 – 5104.

Fox, E. W., Short, M. B., Schoenberg, F. P., Coronges, K. D., and Bertozzi, A. L. (2016). Modeling e-mail networks and inferring leadership using self-exciting point processes. *Journal of the American Statistical Association*, 111(514):564–584.

Fujita, A., Severino, P., Kojima, K., Sato, J. R., Patriota, A. G., and Miyano, S. (2012). Functional clustering of time series gene expression data by Granger causality. *BMC systems biology*, 6(1):137.

Gallier, J. (2016). Spectral theory of unsigned and signed graphs. applications to graph clustering: a survey. *arXiv preprint arXiv:1601.04692*.

Ganapathiraju, M. K., Thahir, M., Handen, A., Sarkar, S. N., Sweet, R. A., Nimgaonkar, V. L., Loscher, C., Bauer, E. M., and Chaparala, S. (2016). Schizophrenia interactome with 504 novel protein–protein interactions. *NPJ Schizophrenia*, 2.

Gerhard, F., Deger, M., and Truccolo, W. (2017). On the stability and dynamics of stochastic

spiking neuron models: Nonlinear Hawkes process and point process glms. *PLOS Computational Biology*, 13:e1005390.

Ghosal, S. and van der Vaart, A. (2007). Convergence rates of posterior distributions for non iid observations. *The Annals of Statistics*, 35(1):192–223.

Goldberg, A. B., Zhu, X., and Wright, S. (2007). Dissimilarity in graph-based semi-supervised classification. In Meila, M. and Shen, X., editors, *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2 of *Proceedings of Machine Learning Research*, pages 155–162, San Juan, Puerto Rico. PMLR.

Gómez, S., Jensen, P., and Arenas, A. (2009). Analysis of community structure in networks of correlated data. *Physical Review E*, 80(1):016114.

Graham, C. (2021). Regenerative properties of the linear Hawkes process with unbounded memory. *The Annals of Applied Probability*, 31(6):2844–2863.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. (2006). A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19.

Gunawardana, A., Meek, C., and Xu, P. (2011). A model for temporal dependencies in event streams. *Advances in neural information processing systems*, 24.

Gusto, G. and Schbath, S. S. (2005). Fado: A statistical method to detect favored or avoided distances between occurrences of motifs using the Hawkes model. *Statistical Applications in Genetics and Molecular Biology*, 4(1):n.p. article n° 24.

Ha, G.-G., Lee, J. W., and Nobi, A. (2015). Threshold network of a financial market using the p-value of correlation coefficients. *Journal of the Korean Physical Society*, 66(12):1802–1808.

Hall, A. R. (2005). *Generalized method of moments*. Oxford university press.

Halpin, P. and De Boeck, P. (2013). Modelling dyadic interaction with Hawkes processes. *Psychometrika*, 78:793–814.

Hansen, N. R., Reynaud-Bouret, P., and Rivoirard, V. (2015). Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, 21(1):83–143.

Harary, F. (1953). On the notion of balance of a signed graph. *Michigan Math. J.*, 2(2):143–146.

Harchaoui, Z., Moulines, E., and Bach, F. (2008). Kernel change-point analysis. *Advances in neural information processing systems*, 21.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):pp. 100–108.

Hawkes, A. G. (1971). Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 33(3):438–443.

Hawkes, A. G. (2018). Hawkes processes and their applications to finance: a review. *Quantitative Finance*, 18(2):193–198.

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

He, X., Xie, Y., Wu, S.-M., and Lin, F.-C. (2018). Sequential graph scanning statistic for change-point detection. In *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, pages 1317–1321. IEEE.

Heider, F. (1958). *The Psychology of Interpersonal Relations*. Psychology Press.

Hewapathirana, I. U., Lee, D., Moltchanova, E., and McLeod, J. (2020). Change detection in noisy dynamic networks: a spectral embedding approach. *Social Network Analysis and Mining*, 10(1):1–22.

Hillairet, C., Huang, L., Khabou, M., and Réveillac, A. (2021). The malliavin-stein method for Hawkes functionals. *arXiv preprint arXiv:2104.01583*.

Hlinka, J., Hartman, D., Jajcay, N., Tomeček, D., Tintěra, J., and Paluš, M. (2017). Small-world bias of correlation networks: From brain to climate. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27(3):035812.

Ho, Q., Song, L., and Xing, E. (2011). Evolving cluster mixed-membership blockmodel for time-evolving networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 342–350. JMLR Workshop and Conference Proceedings.

Holme, P., Edling, C. R., and Liljeros, F. (2004). Structure and time evolution of an internet dating community. *Social Networks*, 26(2):155–174.

Holme, P. and Saramaki, J. (2012). Temporal networks. *Physics Reports*, 519(3):97–125.

Hou, J. P. (2005). Bounds for the least Laplacian eigenvalue of a signed graph. *Acta Mathematica Sinica*, 21(4):955–960.

Huang, J., Shen, H., Hou, L., and Cheng, X. (2019). Signed graph attention networks. In *International Conference on Artificial Neural Networks*, pages 566–577. Springer.

Hüffner, F., Betzler, N., and Niedermeier, R. (2007). Optimal edge deletions for signed graph balancing. In *International Workshop on Experimental and Efficient Algorithms*, pages 297–310. Springer.

Hunter, D., Smyth, P., Vu, D. Q., and Asuncion, A. U. (2011). Dynamic egocentric models for citation networks. In *Proceedings of the 28th international conference on machine learning*, pages 857–864.

Idé, T., Kollias, G., Phan, D., and Abe, N. (2021). Cardinality-regularized hawkes-granger model. *Advances in Neural Information Processing Systems*, 34:2682–2694.

Jin, W., Qu, M., Jin, X., and Ren, X. (2019). Recurrent event network: Autoregressive structure inference over temporal knowledge graphs. *arXiv preprint arXiv:1904.05530*.

Johnson, D. D., Hanchi, A. E., and Maddison, C. J. (2022). Contrastive learning can find an optimal basis for approximately view-invariant functions. *arXiv preprint arXiv:2210.01883*.

Joseph, A. and Yu, B. (2016). Impact of regularization on spectral clustering. *The Annals of Statistics*, 44(4):1765–1791.

Karaaslanli, A., Saha, S., Aviyente, S., and Maiti, T. (2022). scsgl: kernelized signed graph learning for single-cell gene regulatory network inference. *Bioinformatics*, 38(11):3011–3019.

Karataş, A. and Şahin, S. (2018). Application areas of community detection: A review. In *2018 International congress on big data, deep learning and fighting cyber terrorism (IBIGDELFT)*, pages 65–70. IEEE.

Kazemi, S. M., Goel, R., Jain, K., Kobyzev, I., Sethi, A., Forsyth, P., and Poupart, P. (2020). Representation learning for dynamic graphs: A survey. *J. Mach. Learn. Res.*, 21(70):1–73.

Keriven, N. and Vaiter, S. (2022). Sparse and smooth: Improved guarantees for spectral clustering in the dynamic stochastic block model. *Electronic Journal of Statistics*, 16(1):1330–1366.

Kim, B., Lee, K. H., Xue, L., and Niu, X. (2018). A review of dynamic network models with latent variables. *Statistics Surveys*, 12(none):105 – 135.

Kim, S., Putrino, D., Ghosh, S., and Brown, E. N. (2011). A Granger causality measure for point process models of ensemble neural spiking activity. *PLOS Computational Biology*, 7(3):1–13.

Kirichenko, A. and Van Zanten, H. (2015). Optimality of poisson processes intensity learning with Gaussian processes. *Journal of Machine Learning Research*, 16(1):2909–2919.

Knyazev, A. (2018). *On spectral partitioning of signed graphs*, pages 11–22. SIAM.

Knyazev, A. V. (2001). Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method. *SIAM journal on scientific computing*, 23(2):517–541.

Koh, G. C., Porras, P., Aranda, B., Hermjakob, H., and Orchard, S. E. (2012). Analyzing protein–protein interaction networks. *Journal of proteome research*, 11(4):2014–2031.

Koutra, D., Shah, N., Vogelstein, J. T., Gallagher, B., and Faloutsos, C. (2016). DeltaCon: Principled massive-graph similarity function with attribution. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 10(3):1–43.

Kumar, S., Spezzano, F., and Subrahmanian, V. (2014). Accurately detecting trolls in Slashdot Zoo via decluttering. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, pages 188–195. IEEE.

Kumar, S., Spezzano, F., Subrahmanian, V., and Faloutsos, C. (2016). Edge weight prediction in weighted signed networks. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 221–230. IEEE.

Kumar, S., Zhang, X., and Leskovec, J. (2019). Predicting dynamic embedding trajectory in temporal interaction networks. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Kunegis, J., Lommatzsch, A., and Bauckhage, C. (2009). The slashdot zoo. In *Proceedings of the 18th international conference on World wide web - WWW '09*. ACM Press.

Kunegis, J., Schmidt, S., Lommatzsch, A., Lerner, J., Luca, E. W. D., and Albayrak, S. (2010). *Spectral Analysis of Signed Graphs for Clustering, Prediction and Visualization*, pages 559–570. SIAM.

Lambert, R. C., Tuleau-Malot, C., Bessaih, T., Rivoirard, V., Bouret, Y., Leresche, N., and Reynaud-Bouret, P. (2018). Reconstructing the functional connectivity of multiple spike trains using Hawkes models. *Journal of Neuroscience Methods*, 297:9–21.

Le, C. M., Levina, E., and Vershynin, R. (2015). Sparse random graphs: regularization and concentration of the Laplacian. *arXiv preprint arXiv:1502.03049*.

Lei, J. and Rinaldo, A. (2015). Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237.

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.

Leskovec, J., Huttenlocher, D., and Kleinberg, J. (2010). Predicting positive and negative links in online social networks. In *WWW*, pages 641–650.

Leskovec, J., Kleinberg, J., and Faloutsos, C. (2005). Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187.

Li, A., Cornelius, S. P., Liu, Y.-Y., Wang, L., and Barabási, A.-L. (2017). The fundamental advantages of temporal networks. *Science*, 358(6366):1042–1046.

Li, S., Xiao, S., Zhu, S., Du, N., Xie, Y., and Song, L. (2018). Learning temporal point processes via reinforcement learning. *Advances in neural information processing systems*, 31.

Linderman, S. W., Wang, Y., and Blei, D. M. (2017). Bayesian inference for latent Hawkes processes. *Advances in Neural Information Processing Systems*.

Liu, S. and Hauskrecht, M. (2019). Nonparametric regressive point processes based on conditional Gaussian processes. *Advances in Neural Information Processing Systems*, 32.

Lloyd, C., Gunter, T., Osborne, M., and Roberts, S. (2015). Variational inference for Gaussian process modulated poisson processes. In *International Conference on Machine Learning*, pages 1814–1822. PMLR.

Mak, C., Zaiser, F., and Ong, L. (2021). Nonparametric hamiltonian monte carlo. In Meila, M.

and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7336–7347. PMLR.

Malem-Shinitski, N., Ojeda, C., and Opper, M. (2022). Variational Bayesian inference for nonlinear Hawkes process with Gaussian process self-effects. *Entropy*, 24(3).

Matias, C. and Miele, V. (2017). Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1119–1141.

Mei, H. and Eisner, J. M. (2017). The neural Hawkes process: A neurally self-modulating multivariate point process. *Advances in neural information processing systems*, 30.

Mercado, P., Tudisco, F., and Hein, M. (2019). Spectral clustering of signed graphs via matrix power means. In *36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4526–4536, Long Beach, California, USA. PMLR.

Meyer, S., Elias, J., and Höhle, M. (2011). A space-time conditional intensity model for invasive meningococcal disease occurrence. *Biometrics*, 68(2):607–616.

Miller, H. and Mokryn, O. (2020). Size agnostic change point detection framework for evolving networks. *Plos one*, 15(4):e0231035.

Miscouridou, X., Caron, F., and Teh, Y. W. (2018). Modelling sparsity, heterogeneity, reciprocity and community structure in temporal interaction data. *Advances in Neural Information Processing Systems*, 31.

Mohler, G. O., 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.

Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998). Log Gaussian Cox processes. *Scandinavian journal of statistics*, 25(3):451–482.

Moore, M. (1978). An international application of heider's balance theory. *European Journal of Social Psychology*, 8(3):401–405.

Newman, M. E. J. (2002). Spread of epidemic disease on networks. *Phys. Rev. E*, 66:016128.

Niepert, M., Ahmed, M., and Kutzkov, K. (2016). Learning convolutional neural networks for graphs. In *International conference on machine learning*, pages 2014–2023. PMLR.

Ogata, Y. (1978). The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Annals of the Institute of Statistical Mathematics*, 30(1):243–261.

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

Ogata, Y. (1999). Seismicity analysis through point-process modeling: A review. *Seismicity patterns, their statistical significance and physical meaning*, pages 471–507.

Ohn, I. and Lin, L. (2021). Adaptive variational bayes: Optimality, computation and applications. *arXiv preprint arXiv:2109.03204*.

Olinde, J. and Short, M. B. (2020). A self-limiting Hawkes process: Interpretation, estimation, and use in crime modeling. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 3212–3219.

Omi, T. and Aihara, K. (2019). Fully neural network based model for general temporal point processes. *Advances in neural information processing systems*, 32.

Ondrus, M., Olds, E., and Cribben, I. (2021). Factorized binary search: change point detection in the network structure of multivariate high-dimensional time series. *arXiv preprint arXiv:2103.06347*.

Padilla, O. H. M., Yu, Y., and Priebe, C. E. (2019). Change point localization in dependent dynamic nonparametric random dot product graphs. *arXiv preprint arXiv:1911.07494*.

Pareja, A., Domeniconi, G., Chen, J. J., Ma, T., Suzumura, T., Kanezashi, H., Kaler, T., and Leisersen, C. E. (2020). EvolveGCN: Evolving graph convolutional networks for dynamic graphs. In *AAAI*.

Pavlidis, N. G., Plagianakos, V. P., Tasoulis, D. K., and Vrahatis, M. N. (2006). Financial forecasting through unsupervised clustering and neural networks. *Operational Research*, 6(2):103–127.

Peel, L. and Clauset, A. (2015). Detecting change points in the large-scale structure of evolving networks. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Pei, Y., Zhang, J., Fletcher, G., and Pechenizkiy, M. (2016). Node classification in dynamic social networks. *Proceedings of AALTD*, page 54.

Pensky, M. and Zhang, T. (2019). Spectral clustering in the dynamic stochastic block model. *Electronic Journal of Statistics*, 13(1):678–709.

Perry, P. O. and Wolfe, P. J. (2013). Point process modelling for directed interaction networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(5):821–849.

Pfaffelhuber, P., Rotter, S., and Stiefel, J. (2022). Mean-field limits for non-linear Hawkes processes with excitation and inhibition. *Stochastic Processes and their Applications*.

Raad, M. B. (2019). Renewal time points for Hawkes processes. *arXiv preprint arXiv:1906.02036*.

Raad, M. B., Ditlevsen, S., and Löcherbach, E. (2020). Stability and mean-field limits of age dependent Hawkes processes. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 56(3):1958–1990.

Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions. *The Annals of Statistics*, pages 453–466.

Rasmussen, J. G. (2013). Bayesian inference for Hawkes processes. *Methodology and Computing in Applied Probability*, 15(3):623–642.

Reynaud-Bouret, P. (2003). Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities. *Probability Theory and Related Fields*, 126(1):103–153.

Reynaud-Bouret, P. (2006). Penalized projection estimators of the Aalen multiplicative intensity. *Bernoulli*, 12(4):633 – 661.

Reynaud-Bouret, P. and Rivoirard, V. (2008). Near optimal thresholding estimation of a Poisson intensity on the real line. *Electronic Journal of Statistics*, 4:172–238.

Reynaud-Bouret, P., Rivoirard, V., Grammont, F., and Tuleau-Malot, C. (2014). Goodness-of-fit tests and nonparametric adaptive estimation for spike train analysis. *The Journal of Mathematical Neuroscience*, 4(1):1–41.

Reynaud-Bouret, P., Rivoirard, V., and Tuleau-Malot, C. (2013). Inference of functional connectivity in neurosciences via Hawkes processes. In *2013 IEEE global conference on signal and information processing*, pages 317–320. IEEE.

Reynaud-Bouret, P. and Schbath, S. (2010). Adaptive estimation for Hawkes processes; application to genome analysis. *The Annals of Statistics*, 38(5):2781 – 2822.

Rizoiu, M.-A., Lee, Y., Mishra, S., and Xie, L. (2017). A tutorial on Hawkes processes for events in social media. *arXiv preprint arXiv:1708.06401*.

Robins, G., Pattison, P., Kalish, Y., and Lusher, D. (2007). An introduction to exponential random graph (p*) models for social networks. *Social networks*, 29(2):173–191.

Rohe, K., Chatterjee, S., and Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878 – 1915.

Rossetti, G. and Cazabet, R. (2018). Community discovery in dynamic networks. *ACM Computing Surveys*, 51(2):1–37.

Rossi, E., Chamberlain, B., Frasca, F., Eynard, D., Monti, F., and Bronstein, M. (2020). Temporal graph networks for deep learning on dynamic graphs. *arXiv preprint arXiv:2006.10637*.

Rubinov, M. and Sporns, O. (2011). Weight-conserving characterization of complex functional brain networks. *Neuroimage*, 56(4):2068–2079.

Saberi, M., Khosrowabadi, R., Khatibi, A., Misic, B., and Jafari, G. (2021). Topological impact of negative links on the stability of resting-state brain network. *Scientific reports*, 11(1):1–14.

Sankar, A., Wu, Y., Gou, L., Zhang, W., and Yang, H. (2020). Dysat: Deep neural representation learning on dynamic graphs via self-attention networks. In *Proceedings of the 13th international conference on web search and data mining*, pages 519–527.

Sanna Passino, F. and Heard, N. A. (2022). Mutually exciting point process graphs for modelling dynamic networks. *Journal of Computational and Graphical Statistics*, pages 1–30.

Sarkar, P. and Bickel, P. J. (2015). Role of normalization in spectral clustering for stochastic blockmodels. *The Annals of Statistics*, 43(3):962–990.

Sarkar, P. and Moore, A. W. (2006). Dynamic social network analysis using latent space models. *Advances in neural information processing systems*, 18:1145.

Seo, Y., Defferrard, M., Vandergheynst, P., and Bresson, X. (2018). Structured sequence modeling with graph convolutional recurrent networks. In *International conference on neural information processing*, pages 362–373. Springer.

Serafini, F., Lindgren, F., and Naylor, M. (2022). Approximation of Bayesian Hawkes process models with Inlabru. *arXiv preprint arXiv:2206.13360*.

Sewell, D. K. and Chen, Y. (2015). Latent space models for dynamic networks. *Journal of the American Statistical Association*, 110(512):1646–1657.

Shahriari, M. and Jalili, M. (2014). Ranking nodes in signed social networks. *Social network analysis and mining*, 4(1):1–12.

Shchur, O., Türkmen, A. C., Januschowski, T., and Günnemann, S. (2021). Neural temporal point processes: A review. *arXiv preprint arXiv:2104.03528*.

Skarding, J., Gabrys, B., and Musial, K. (2021a). Foundations and modeling of dynamic networks using dynamic graph neural networks: A survey. *IEEE Access*, 9:79143–79168.

Skarding, J., Gabrys, B., and Musial, K. (2021b). Foundations and modeling of dynamic networks using dynamic graph neural networks: A survey. *IEEE Access*, 9:79143–79168.

Smith, S. M., Miller, K. L., Salimi-Khorshidi, G., Webster, M., Beckmann, C. F., Nichols, T. E., Ramsey, J. D., and Woolrich, M. W. (2011). Network modelling methods for FMRI. *NeuroImage*, 54(2):875 – 891.

Stephan, L. and Massoulié, L. (2019). Robustness of spectral methods for community detection. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2831–2860, Phoenix, USA. PMLR.

Sugishita, K. and Masuda, N. (2021). Recurrence in the evolution of air transport networks. *Scientific reports*, 11(1):1–15.

Tang, J., Chang, Y., Aggarwal, C., and Liu, H. (2016). A survey of signed network mining in social media. *ACM Computing Surveys (CSUR)*, 49(3):1–37.

Teh, Y. and Rao, V. (2011). Gaussian process modulated renewal processes. *Advances in Neural Information Processing Systems*, 24.

Tomasso, M., Rusnak, L. J., and Tešić, J. (2022). Advances in scaling community discovery methods for signed graph networks. *Journal of Complex Networks*, 10(3):cnac013.

Torrisi, G. L. (2016). Gaussian approximation of nonlinear Hawkes processes. *The Annals of Applied Probability*.

Torrisi, G. L. (2017). Poisson approximation of point processes with stochastic intensity, and application to nonlinear Hawkes processes. *Annales de l'Institut Henri Poincare*.

Trivedi, R., Farajtabar, M., Biswal, P., and Zha, H. (2019). Dyrep: Learning representations over dynamic graphs. In *International Conference on Learning Representations*.

Tzeng, R.-C., Ordozgoiti, B., and Gionis, A. (2020). Discovering conflicting groups in signed networks. *Advances in Neural Information Processing Systems*, 33:10974–10985.

Vassilvitskii, S. and Arthur, D. (2006). k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035.

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.

Vu, D., Hunter, D., Smyth, P., and Asuncion, A. (2011). Continuous-time regression models for longitudinal networks. *Advances in neural information processing systems*, 24.

Wang, D., Yu, Y., and Rinaldo, A. (2021). Optimal change point detection and localization in sparse dynamic networks. *The Annals of Statistics*, 49(1):203–232.

Wang, H., Tang, M., Park, Y., and Priebe, C. E. (2013). Locality statistics for anomaly detection in time series of graphs. *IEEE Transactions on Signal Processing*, 62(3):703–717.

Wang, X., Wang, P., and Man-Cho So, A. (2022). Exact community recovery over signed graphs. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I., editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 9686–9710. PMLR.

Wang, Y., Chakrabarti, A., Sivakoff, D., and Parthasarathy, S. (2017). Fast change point detection on dynamic social networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 2992–2998. AAAI Press.

Wang, Y., Xie, B., Du, N., and Song, L. (2016). Isotonic Hawkes processes. In *International conference on machine learning*, pages 2226–2234.

Ward, M. D., Ahlquist, J. S., and Rozenas, A. (2013). Gravity's rainbow: A dynamic latent space model for the world trade network. *Network Science*, 1(1):95–118.

Welling, M. and Kipf, T. N. (2016). Semi-supervised classification with graph convolutional networks. In *J. International Conference on Learning Representations (ICLR 2017)*.

Wilson, J. D., Stevens, N. T., and Woodall, W. H. (2019). Modeling and detecting change in temporal networks via the degree-corrected stochastic block model. *Quality and Reliability Engineering International*, 35(5):1363–1378.

Xiao, S., Farajtabar, M., Ye, X., Yan, J., Song, L., and Zha, H. (2017). Wasserstein learning of deep generative point process models. *Advances in neural information processing systems*, 30.

Xu, H., Farajtabar, M., and Zha, H. (2016). Learning Granger causality for Hawkes processes. In *International conference on machine learning*, pages 1717–1726. PMLR.

Yang, B., Cheung, W., and Liu, J. (2007). Community mining from signed social networks. *IEEE transactions on knowledge and data engineering*, 19(10):1333–1348.

Yang, T., Chi, Y., Zhu, S., Gong, Y., and Jin, R. (2011). Detecting communities and their evolutions in dynamic social networks—a bayesian approach. *Machine learning*, 82(2):157–189.

Yu, Y., Padilla, O. H. M., Wang, D., and Rinaldo, A. (2021). Optimal network online change point localisation. *arXiv preprint arXiv:2101.05477*.

Zhang, F. and Gao, C. (2020). Convergence rates of variational posterior distributions. *The Annals of Statistics*, 48(4):2180–2207.

Zhang, M., Cui, Z., Neumann, M., and Chen, Y. (2018a). An end-to-end deep learning architecture for graph classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Zhang, Q., Lipani, A., and Yilmaz, E. (2021). Learning neural point processes with latent graphs. In *Proceedings of the Web Conference 2021*, WWW '21, page 1495–1505, New York, NY, USA. Association for Computing Machinery.

Zhang, R., Walder, C., Rizoiu, M.-A., and Xie, L. (2018b). Efficient non-parametric bayesian Hawkes processes. *arXiv preprint arXiv:1810.03730*.

Zhao, Z., Chen, L., and Lin, L. (2019). Change-point detection in dynamic networks via graphon estimation. *arXiv preprint arXiv:1908.01823*.

Zhou, F., Kong, Q., Deng, Z., Kan, J., Zhang, Y., Feng, C., and Zhu, J. (2022). Efficient inference for dynamic flexible interactions of neural populations. *Journal of Machine Learning Research*, 23(211):1–49.

Zhou, F., Kong, Q., Zhang, Y., Feng, C., and Zhu, J. (2021a). Nonlinear Hawkes processes in time-varying system. *arXiv preprint arXiv:2106.04844*.

Zhou, F., Luo, S., Li, Z., Fan, X., Wang, Y., Sowmya, A., and Chen, F. (2021b). Efficient em-variational inference for nonparametric Hawkes process. *Statistics and Computing*, 31(4):1–11.

Zhou, F., Zhang, Y., and Zhu, J. (2020). Efficient inference of flexible interaction in spiking-neuron networks. *arXiv preprint arXiv:2006.12845*.

Zhou, K., Zha, H., and Song, L. (2013). Learning triggering kernels for multi-dimensional Hawkes processes. In *International conference on machine learning*, pages 1301–1309. PMLR.

Zhu, L. (2013). Central limit theorem for nonlinear Hawkes processes. *Journal of Applied Probability*, 50(3):760–771.

Zuo, S., Jiang, H., Li, Z., Zhao, T., and Zha, H. (2020). Transformer Hawkes process. In *International conference on machine learning*, pages 11692–11702. PMLR.

Özgür Şimşek and Jensen, D. (2008). Navigating networks by using homophily and degree. *Proceedings of the National Academy of Sciences*, 105(35):12758–12762.

Appendices

A | Discrete-time Hawkes model: a case study of COVID-19

This chapter corresponds to the following article:

> Browning, R., Sulem, D., Mengersen, K., Rivoirard, V., Rousseau, J. (2021) Simple discrete-time self-exciting models can describe complex dynamic processes: A case study of COVID-19. PLoS ONE 16(4): e0250015.

PLOS ONE

RESEARCH ARTICLE

Simple discrete-time self-exciting models can describe complex dynamic processes: A case study of COVID-19

Raiha Browning[1,2]*, **Deborah Sulem**[3], **Kerrie Mengersen**[1,2☯], **Vincent Rivoirard**[4☯], **Judith Rousseau**[3,4☯]

1 School of Mathematical Sciences, Queensland University of Technology, Brisbane, Australia, **2** Australian Research Council, Centre of Excellence for Mathematical and Statistical Frontiers, Brisbane, Australia, **3** Department of Statistics, University of Oxford, Oxford, United Kingdom, **4** Ceremade, Université Paris-Dauphine, Paris, France

☯ These authors contributed equally to this work.
* raihatuitaura.browning@qut.edu.au

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

Editor: Dan Braha, University of Massachusetts, UNITED STATES

Received: November 3, 2020

Accepted: March 29, 2021

Published: April 9, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0250015

Data Availability Statement: The data used in this analysis are available on Github: https://github.com/RaihaTuiTaura/covid-hawkes-paper. This data was obtained from Johns Hopkins University:

Abstract

Hawkes processes are a form of self-exciting process that has been used in numerous applications, including neuroscience, seismology, and terrorism. While these self-exciting processes have a simple formulation, they can model incredibly complex phenomena. Traditionally Hawkes processes are a continuous-time process, however we enable these models to be applied to a wider range of problems by considering a discrete-time variant of Hawkes processes. We illustrate this through the novel coronavirus disease (COVID-19) as a substantive case study. While alternative models, such as compartmental and growth curve models, have been widely applied to the COVID-19 epidemic, the use of discrete-time Hawkes processes allows us to gain alternative insights. This paper evaluates the capability of discrete-time Hawkes processes by modelling daily mortality counts as distinct phases in the COVID-19 outbreak. We first consider the initial stage of exponential growth and the subsequent decline as preventative measures become effective. We then explore subsequent phases with more recent data. Various countries that have been adversely affected by the epidemic are considered, namely, Brazil, China, France, Germany, India, Italy, Spain, Sweden, the United Kingdom and the United States. These countries are all unique concerning the spread of the virus and their corresponding response measures. However, we find that this simple model is useful in accurately capturing the dynamics of the process, despite hidden interactions that are not directly modelled due to their complexity, and differences both within and between countries. The utility of this model is not confined to the current COVID-19 epidemic, rather this model could explain many other complex phenomena. It is of interest to have simple models that adequately describe these complex processes with unknown dynamics. As models become more complex, a simpler representation of the process can be desirable for the sake of parsimony.

https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series.

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Introduction

The outbreak of the novel 2019 coronavirus disease (COVID-19) was declared a Global Health Emergency of International Concern on 30th January 2020, and pronounced a Pandemic on 11th March 2020. It has since spread rapidly with over 116 million confirmed cases and more than 2.5 million deaths as of 7th March 2021 [1]. Since the first reported case in December 2019, countries around the world have fought to contain the virus. In the absence of a vaccine, countries implemented a range of non-pharmaceutical interventions and strategies to reduce the spread of the virus, from measures such as social distancing, mask-wearing and contact tracing, to complete city lockdowns and stay at home orders. These recommendations are guided by mathematical and statistical modelling to quantify the efficacy of these measures [2–9].

There is now an expansive collection of research dedicated to understanding the virus from all perspectives, including its biological, epidemiological, clinical, economic and social impacts. There is also a wealth of knowledge around prevention strategies to control the outbreak. In all of these, statistical and mathematical models are an essential aspect to gaining meaningful insights into how the virus spreads and quantifying its various impacts. A popular choice is compartmental models, with some considering the standard SIR (Susceptible-Infected-Recovered) model [10–12], and further extensions in which additional states are introduced [13–18]. As an alternative to compartmental models, others have used methods such as branching processes to capture the spread of the virus through individual networks [2, 3, 5], log-linear Poisson autoregressive models [19], and other probabilistic models of the infection cycle of the virus [20]. Various models based on growth curves have also been proposed, for example [21–23], who use logistic, exponential and Richards growth curves respectively. More detailed approaches such as agent-based modelling have also been considered by numerous authors [24–27].

A Hawkes process [28] is a stochastic, self-exciting process in which past events influence the short-term probability of future events occurring. They are often used to explain many phenomena that exhibit self-exciting properties, including neuroscience [29–31], crime and terrorism [32–34], seismic activity [35] and social media [36]. Similarly, due to their contagious nature it is also natural to represent infectious diseases, such as the current COVID-19 pandemic, as a Hawkes process.

Hawkes processes have been successfully applied to model epidemics and infectious diseases. For example, for the Ebola outbreaks in West Africa and the Democratic Republic of Congo [37, 38], the Hawkes process is found to outperform the SEIR (Susceptible-Exposed-Infected-Recovered) mechanistic model in terms of short term prediction. Another study employs an extension of the multivariate Hawkes process to understand the transmission routes and regional connectivity for the dengue fever outbreak across regions in Australia [39]. Rocky Mountain Spotty Fever has also been modelled using a recursive Hawkes process, with the expected number of transmissions based on the current conditional intensity of the Hawkes process [40]. Moreover [41], model invasive meningococcal disease using a spatiotemporal extension to the Hawkes process.

The spread of COVID-19 is an extremely complex process, with unknown disease dynamics and huge variations in the preventative measures and responses of different countries. We propose a parsimonious model for COVID-19 deaths, namely discrete-time Hawkes processes (DTHP) [32, 33, 42], to describe the complicated dynamics of the COVID-19 epidemic. In its original form, the Hawkes process is a continuous-time point process; however, the DTHP observes the occurrence of events at a discrete time resolution. Due to this construction, the DTHP can directly model the available data (i.e. daily counts), without artificially imputing the

data onto a continuous timeline, as is generally done in studies using continuous-time Hawkes processes. We also introduce deterministic change points in this study, since the dynamics of the spread vary abruptly as the pandemic progresses and preventative interventions are introduced.

Alternative models, such as the mechanistic and growth curve models discussed previously, primarily focus on estimating the model parameters that govern the system. Hawkes processes, however, are more detailed, as individual events and their respective occurrence times directly influence the likelihood of future events occurring. Hawkes processes also provide additional insights into the infection dynamics of diseases by estimating the level of external cases through the baseline parameter and the triggering kernel, which models the decay in infectivity through time.

Hawkes processes and compartmental models are based on different mathematical principles and rely on different assumptions. However, their connection was explored by [43]. These authors show that, via a modified, finite population variant of the Hawkes model for a particular choice of triggering kernel, the rate of events is equivalent to the SIR model's infection rate. While the SIR family of models is useful if more is known about the system dynamics, a simpler model is often useful for phenomena where there are many unknowns. We show in this study that our model is helpful for this purpose. Additionally, we explore the differences between Hawkes, compartmental models and other approaches further in the discussion.

Related work

An approach to modelling the COVID-19 pandemic using self-exciting branching processes has been suggested by [44]. These authors employ a continuous-time Hawkes model with a nonparametric estimate of the reproduction number, $R(t)$, the average number of secondary cases produced by a single case of the virus. Both death counts and the number of confirmed cases in the early stage of the epidemic, before April 1st, are modelled in three states of the U.S., several European countries and China. Compared to SIR and SEIR models with a fixed reproduction number, their Hawkes model with a dynamic parameter leads to lower estimates of the basic reproduction number, R_0. In the same line of work [45], consider several datasets for the state of Indiana in the early stage of the epidemic. They also compare a nonparametric estimate of the reproduction number, $R(t)$, with an exponentially decreasing function and a step-function, and find that the estimation of R is very sensitive to the type of input data (i.e. deaths or cases), the data source, and the model choice. Similarly [46], adopt a continuous-time Hawkes model with spatial covariates to model both the number of confirmed COVID-19 cases and the number of deaths, for the U.S. at the county level. This study also considers a time-varying reproduction number. Finally [47], also use the continuous-time Hawkes process to illustrate the severity of the virus in France if no preventative action were to be taken.

Two similar approaches to ours are that of [48, 49]. The former proposes a two-phase contagion model based on an extension of the Hawkes process. This study considers a continuous-time Hawkes process, assume the rate of external events varies through time, and estimate the change point in their model. The authors also assume there is no external excitation after the change point. The latter of these is, to the authors' knowledge, the most similar approach to ours. These authors consider a discrete-time Hawkes process to describe the current COVID-19 epidemic. This study focusses on estimating a time-varying reproduction number, ignoring the influence of external activity and considering a fixed excitation kernel.

Several other approaches for modelling COVID-19 that incorporate change points have been proposed to capture the dynamic nature of the pandemic. [50, 51] find that using compartmental models with time-varying infection rates, the estimated change points for Germany

and South Africa, respectively, align with various government interventions in these countries. [52] do not directly estimate the change points; instead, they propose a compartmental model for Italy with piecewise model parameters partitioned into regular time intervals. Alternatively [53], consider a combination of exponential and polynomial regression models to estimate the optimal change points for the COVID-19 outbreak in India. While these studies consider only a single country [54], examine several countries and introduce a single stochastic change point into their compartmental model. [55] present a widespread study across 55 countries using a partially observed Markov process with piecewise transmission rates.

Contributions

In the current literature, the continuous-time Hawkes process requires artificial imputation of the daily count data onto a continuous time resolution, adding a significant computational burden to the implementation and adding additional, potentially unnecessary, noise to the model. We develop a multi-phase approach for the DTHP to directly model the reported daily counts of the number of deaths caused by the virus.

The dynamics of the process before and after the enactment of preventative measures and policy interventions to reduce the spread of the virus are inherently different. The majority of the existing literature on modelling the COVID-19 pandemic using Hawkes processes consider only the early stages of the pandemic. In this work, we develop a variant of the DTHP to model the distinct phases of the COVID-19 epidemic. We modify the traditional Hawkes process to account for this change in dynamics by including deterministic change points in the model.

While [49] also study more recent data, these authors limit parameter estimation to the reproduction number, and fix the remaining parameters of the Hawkes model. In our study, we estimate the excitation kernel for additional flexibility. Regarding external events [48], also assume there is no external excitation in the second phase of their two-phase model. We make no such assumption, and believe considering external excitation throughout the entire course of the pandemic is a valuable consideration. There are still travellers arriving from abroad, and thus exogenous activity is still occurring in later phases at a lower rate. This is particularly relevant as many countries have relatively relaxed quarantine requirements, which means that travellers from abroad are still capable of spreading the virus. Although we study mortality data in this analysis, we are able to make a connection between mortalities and infections. In particular, we show in S1 Appendix that the rate of external events in our model can roughly be interpreted as external infections, times the probability of death given infection. This link is particularly useful in the absence of reliable infection data.

Change point models for Hawkes processes have been considered in other applications [56]. However, these authors assume independence of the observed data between change points, prohibiting events that occur within a time period to influence events in future time periods. This type of model is inappropriate for this application, as the time periods are not independent. While the behaviour of the process varies between time periods, the influence of past events remains active in the memory of the process. Thus, the baseline parameters become artificially inflated if events from different time periods are assumed to be independent. For the current COVID-19 pandemic [49], introduce a method for detecting change points in the reproduction number through augmenting their Hawkes model with state-space methods.

In particular for the COVID-19 epidemic, while other studies directly estimate the change points or partition the timeline into regular intervals to reflect the evolving dynamics of the epidemic, we propose a simple method that incorporates fixed change points. We do not estimate the change points for our model, as it was fairly obvious where a reasonable change point

was in these data, and this avoids complexity arising from different interventions being introduced in each country, with varying levels of restrictions. Furthermore, the delays before tangible results are observed, in addition to the complex and hidden interactions underlying the process, complicate the interpretation of estimated change points. We instead opt for this consistent and simplistic definition of the change point for each country. The change points could however be estimated for more complex trajectories.

We illustrate in this study how a simple model can be used to describe exceedingly complex natural phenomena such as epidemics, and in particular the COVID-19 pandemic. Although it is the same underlying phenomenon, all countries are unique concerning the spread of the virus and the resultant response measures. Our simple model can capture these dynamics. Additionally, while many other studies consider small-scale regions, such as individual counties in the U.S., we are also able to gain insights into the dynamics of the process at a higher-level across entire countries.

Outline

First we define a general form of the DTHP, and contrast this with its continuous-time equivalent. We then introduce the particular model used in the initial stage of this analysis for modelling COVID-19, incorporating a change point into the construction of the DTHP. Next, a brief description of the data and inference methods are provided. Finally, the results for the ten countries of interest are presented, and we also show the results from fitting our model to more recent data. This is followed by a discussion and concluding remarks.

Methods

Discrete-time Hawkes process

The discrete-time Hawkes process is a self-exciting stochastic process whereby events occur at regular intervals on a discrete-time scale. It follows a similar construction to the continuous-time Hawkes process [28]. The conditional intensity function $\lambda(t)$ characterises a Hawkes process, and herein lies the difference between the continuous-time and discrete-time variants. For the DTHP, $\lambda(t)$ represents the expected number of events that occur at time interval t, conditionally on the past. In contrast, for the continuous-time Hawkes process, $\lambda(t)$ is the instantaneous rate of an event occurring at time t. The DTHP model also has an extra layer of flexibility compared to its continuous-time counterpart as the underlying data generating process can be selected as any counting distribution with conditional mean $\lambda(t)$.

Consider a linear univariate discrete-time Hawkes process N, where $N(t)$ represents the number of events up to time interval t. $N(t)$ is dependent on the history of events up to but not including time t, denoted by $H_{t-1} = \{y_s: s \leq t-1\}$, where y_s represents the observed number of events in a given time interval s. Furthermore, $N(t) - N(t-1)$ represents the number of event occurrences at time t, and thus,

$$\begin{aligned} \lambda(t) &= E\{N(t) - N(t-1)|H_{t-1}\} \\ &= \mu + \alpha \sum_{i:t_i<t} y_{t_i} g(t - t_i) \end{aligned} \tag{1}$$

where μ represents the baseline mean of the process and the second term represents the self-exciting component of the Hawkes process, describing the expected number of events during a particular interval t given previous events. The triggering kernel $g(t - t_i)$ describes the influence of past events on the intensity of the process, given the time elapsed since event i, where $t > t_i$. In this study, we specify the triggering kernel to be a proper probability mass function

with strictly positive integer-valued support. Since the sum of the excitation kernel over $\mathbb{Z}_+$ is equal to 1, one can interpret the non-negative magnitude parameter $\alpha \in \mathbb{R}_{\geq 0}$ as the expected number of subsequent events produced by a single event [33].

Model

Daily counts of the reported number of deaths of the novel coronavirus COVID-19 are modelled using the discrete-time Hawkes process, where the number of events observed on day t, namely y_t, are distributed according to the random variable, $Y(t)$, which has conditional mean $E(Y(t)|H_{t-1}) = \lambda(t)$ as defined in Eq (1). In this analysis $Y(t)$ is assumed Poisson distributed, thus $Y(t) \sim \mathcal{P}(\lambda(t))$. The Poisson distribution is selected as it has an intuitive interpretation regarding the generation of daily death counts on a given day, and because it is a natural approximation of a binomial distribution with a large population and low death rate. More detail is given in S1 Appendix. Thus, for the proposed DTHP model, the probability that day t has y events is,

$$P(Y(t) = y|\lambda(t)) = \frac{\lambda(t)^y e^{-\lambda(t)}}{y!}$$

First we consider an initial period up to 25th July 2020, to determine some initial modelling assumptions and study the model performance in the early stages of the pandemic. The conditional intensity function $\lambda(t)$ is altered from Eq (1) to allow for a change point in the process, since the DTHP with fixed parameters is unable to capture the complex dynamics for an epidemic of this scale. The parameters of the DTHP implicitly incorporate environmental and social characteristics that are significant for the spread of the disease, and these characteristics change after preventative measures are introduced. Thus, if the dynamic nature of the epidemic is not taken into account, the model averages the estimated parameters, combining the effects of the initial explosive phase of the pandemic with the downward trend that follows after the implementation of preventative measures.

In the initial period of analysis, to accommodate this shape, we assume in our analysis that two phases can adequately separate the underlying dynamics. Namely, these phases are the initial period where the virus is spreading rapidly and the following period of reduced contagion resulting from the introduction of preventative measures and policies. Many complex interactions are occurring in the deaths process. For example, as medical professionals become more familiar with the virus and treatments are improved, medical facilities are better equipped to deal with COVID-19 patients in critical condition requiring ICU [57, 58]. However, this can be offset by increased demand for hospital beds, resulting in medical facilities becoming overwhelmed and unable to care for all patients that require hospital treatment. Therefore, rather than making explicit assumptions about the underlying processes driving the death dynamics, we link our Hawkes model on the death dynamics to a similar infection model, as we discuss in S1 Appendix.

Thus, we first retrospectively define a single change point at time T_1, where T_1 is the maximum value of deaths, to capture the different dynamics of the epidemic at two distinct stages of the outbreak.

The triggering kernel $g(t - t_i)$ is selected as a geometric excitation kernel, $g(t - t_i;\beta) = \beta(1 - \beta)^{t - t_i - 1}$. The exponential distribution is one of the most commonly used triggering kernels for continuous-time processes. Thus we choose the geometric kernel as it can be shown to be equivalent to the exponential distribution in the context of discrete time. The parameter β represents the success probability in the geometric distribution, and thus the average of the

excitation kernel is $\frac{1}{\beta}$. We also express the expectation of the maximum excitation time in terms of the parameters of the model in S2 Appendix.

The conditional intensity function before T_1 is calculated using one set of model parameters, $(\mu_1, \alpha_1, \beta_1)$. After T_1, the intensity function is calculated using a new set of parameters, $(\mu_2, \alpha_2, \beta_2)$ for the second phase in the epidemic. Thus for one change point at time T_1, $\lambda(t)$ is given by,

$$\lambda(t) = \begin{cases} \mu_1 + \alpha_1 \sum_{i:t_i<t} y_{t_i} g_1(t - t_i), & t \leq T_1 \\ \mu_2 + \alpha_2 \sum_{i:t_i<t} y_{t_i} g_2(t - t_i), & t > T_1 \end{cases} \quad (2)$$

It is straightforward to extend Eq (2) to allow for additional change points. While the majority of this paper considers only the initial stage of the pandemic up to 25th July 2020, we consider subsequent phases after this date as a set of additional analysis. This is to demonstrate how our model can be extended beyond the initial phases of the pandemic, as new data will continue to become available each day for the foreseeable future.

Although we consider the deceased population rather than the infected population, there is a connection between the two under some simplifications. Thus studying deaths is useful for understanding the infection dynamics as well. This is advantageous particularly in the early stages of a pandemic, when no reliable data on infections are available. We do not go into the details here, but the key outcome of this is that α, β and a function of μ are interpreted with respect to infections, not deaths. The full derivation is available in S1 Appendix. As this approximation relies on the assumption of a large population and a low death rate, we would not expect this model to be reasonable for other time series where the rate of occurrence is high, such as COVID-19 recoveries.

For a time series of T days and a given country, the log-likelihood function for this DTHP model with retrospective change point, T_1, up to an additive constant K, is then,

$$\begin{aligned} &\log L(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \\ K \quad &+ \sum_{t=1}^{T_1} \left[y_t \log\left(\mu_1 + \alpha_1 \sum_{i:t_i<t} y_{t_i} \beta_1 (1-\beta_1)^{t-t_i-1}\right) - \left(\mu_1 + \alpha_1 \sum_{i:t_i<t} y_{t_i} \beta_1 (1-\beta_1)^{t-t_i-1}\right) \right] \\ &+ \sum_{t=T_1+1}^{T} \left[y_t \log\left(\mu_2 + \alpha_2 \sum_{i:t_i<t} y_{t_i} \beta_2 (1-\beta_2)^{t-t_i-1}\right) - \left(\mu_2 + \alpha_2 \sum_{i:t_i<t} y_{t_i} \beta_2 (1-\beta_2)^{t-t_i-1}\right) \right] \end{aligned}$$

Data

We use data gathered by the Johns Hopkins University [59] in this work. These data come in the form of daily counts of confirmed cases or deaths by country and region. In this analysis, the number of daily reported deaths for a selection of countries, namely Brazil, China, France, Germany, India, Italy, Spain, Sweden, the United Kingdom and the United States, are considered. We select these countries to represent a global sample of countries that have been adversely affected by the coronavirus outbreak. It is important to note that the definition of deaths due to COVID-19 varies between countries. These differences are ignored in our modelling.

The reported number of deaths was considered a more reliable response variable than the reported number of cases. This is due to data issues that can arise when considering the number of confirmed cases, such as lack of testing or differing testing rates between countries, differences in definitions and differences in the timing for reporting of cases. Additionally, to

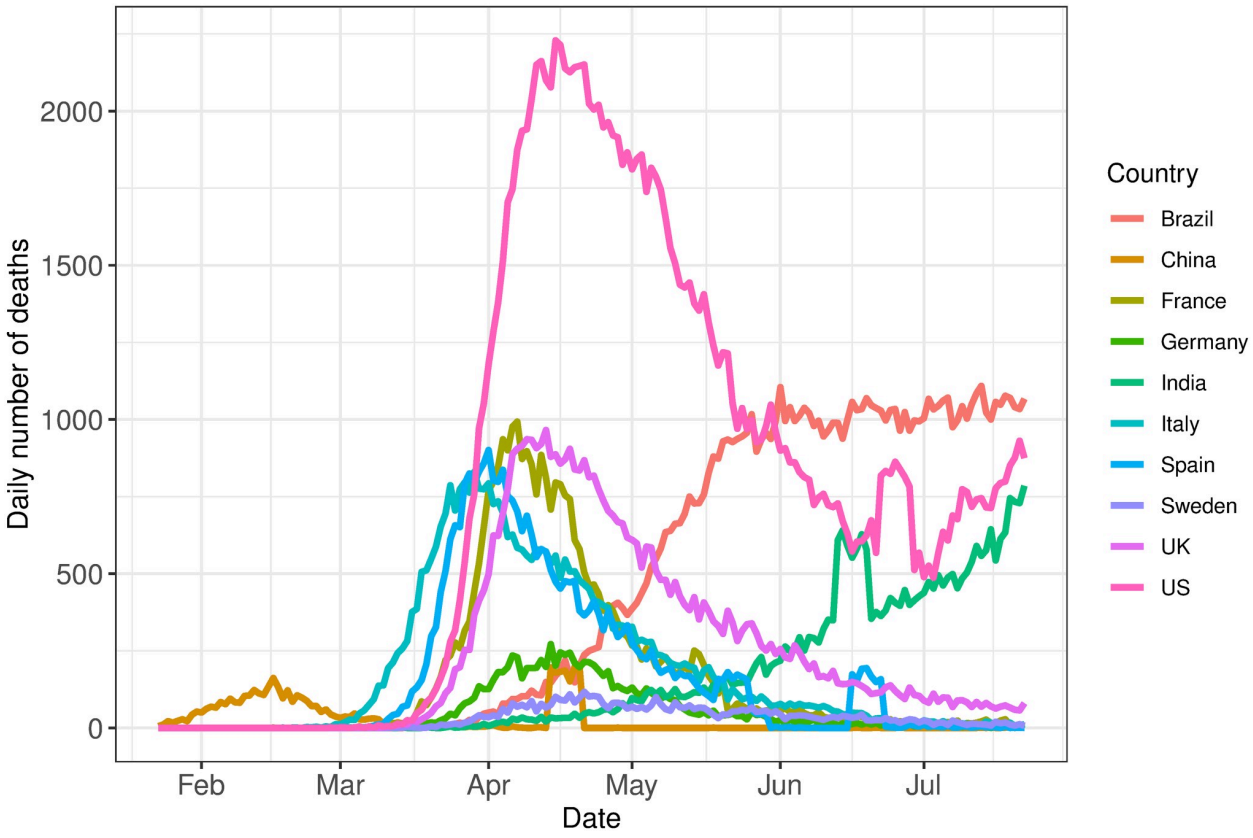


Fig 1. Observed data. Daily volume of deaths due to COVID-19 for the countries selected in this analysis.

https://doi.org/10.1371/journal.pone.0250015.g001

mitigate the effect of systematic influences in reporting, such as lower reporting on weekends [50], the data is smoothed over a rolling window of seven days. The start of the observation window, t_1, for each country is defined as the time the number of deaths exceeds ten. Fig 1 shows the smoothed volume of daily deaths for the countries under consideration up to 25th July 2020.

For the initial stage of this analysis, we consider data up to 25th July 2020. We define a single change point, T_1, as the time where the maximum number of deaths occurs, for the countries with sufficient data in the downward phase of the epidemic by the end of the initial study period. Where there is insufficient evidence for the downward trend, for example, in India and Brazil, no change point was introduced, and only a single phase was modelled. Moreover, the trend for Brazil showed evidence of the curve flattening; however, there was insufficient data for this second phase. Thus the end of the observation window for Brazil is fixed on 1st June 2020. Additionally, as China, India, Spain and the United States experienced large deviations from the current trend towards the end of the observed data, earlier endpoints of 13th April 2020, 12th June 2020, 15th June 2020 and 21st June 2020 were imposed respectively. This avoids the anomalous spikes at the end of these series, since it was not clear whether these aberrations were real or due to reporting definitions or other errors. The endpoint for the remaining countries was set as 25th July 2020. We later extend our analysis to include more recent data, to demonstrate the utility of our model in later phases of the pandemic. A description of the data processing for this is in the relevant Results section.

Parameter inference

Parameter estimation is undertaken using Bayesian methods. We consider a range of prior choices for the baseline parameters μ_1 and μ_2, and perform leave-future-out cross validation with Pareto smoothed importance sampling [60] to assess the performance of each prior choice. The priors considered are,

$$\mu_1, \mu_2 \sim \begin{cases} \log N(1,1) \\ \log N(5,1.5) \\ \text{Gamma}(2,2) \\ \text{Gamma}(5,1) \\ U(0,\infty), \end{cases}$$

where the first term of the log-normal priors represents the mean of the random variable itself, as opposed to the mean of the variable's natural logarithm.

Cross validation with Pareto smoothed importance sampling relies on the expected log predictive density (ELPD), for which a larger value indicates a better model fit. We calculate the ELPD in each country for each of the baseline parameter prior choices, and these results are provided in S1 Table. Based on this analysis, there is no obvious choice of prior that consistently outperforms the rest for each country. On the contrary, the difference in the ELPD is marginal between priors. The remainder of this paper presents the results for $\mu_1, \mu_2 \sim$ Gamma (5, 1), as this is most frequently the highest ELPD, and if not the maximum, is generally very comparable.

Flat priors are selected for α_1, α_2, β_1 and β_2 such that,

- $\pi(\alpha_1, \alpha_2) \propto \mathbb{I}_{(0,\infty)^2}(\alpha_1, \alpha_2)$
- $\beta_1, \beta_2 \sim U(0, 1)$

A Metropolis-adjusted Langevin step [61] is used to jointly update α_1 and β_1, and also to jointly update α_2 and β_2. Denoting the parameters at iteration t by $\alpha^{(t)}, \beta^{(t)}$, the proposals α^*, β^* are simulated from,

$$\begin{bmatrix} \alpha^* \\ \beta^* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \alpha^{(t)} \\ \beta^{(t)} \end{bmatrix} + \frac{\epsilon^2}{2} G \begin{bmatrix} D_\alpha(\alpha^{(t)}, \beta^{(t)}) \\ D_\beta(\alpha^{(t)}, \beta^{(t)}) \end{bmatrix}, \epsilon^2 G \right) \tag{3}$$

where $D_\alpha(.)$ and $D_\beta(.)$ are the gradients of log$L$ with respect to α and β respectively, G is a pre-conditioning matrix accounting for covariance between parameters and ϵ is the step size in the Metropolis-adjusted Langevin algorithm.

The MCMC chain was run for 60,000 iterations discarding the first 20,000. The pre-conditioning matrix G was taken as the covariance matrix from an implementation of the standard Metropolis-Hastings algorithm for each country. The R code and data required to replicate this study are available on Github (https://github.com/RaihaTuiTaura/covid-hawkes-paper).

Results

We first present results from the initial analysis considering data up to 25th July 2020. Fig 2 presents the 95% posterior intervals around the estimated conditional intensity function $\lambda(t)$ against the observed data for each country. The estimated intensity function on day t,

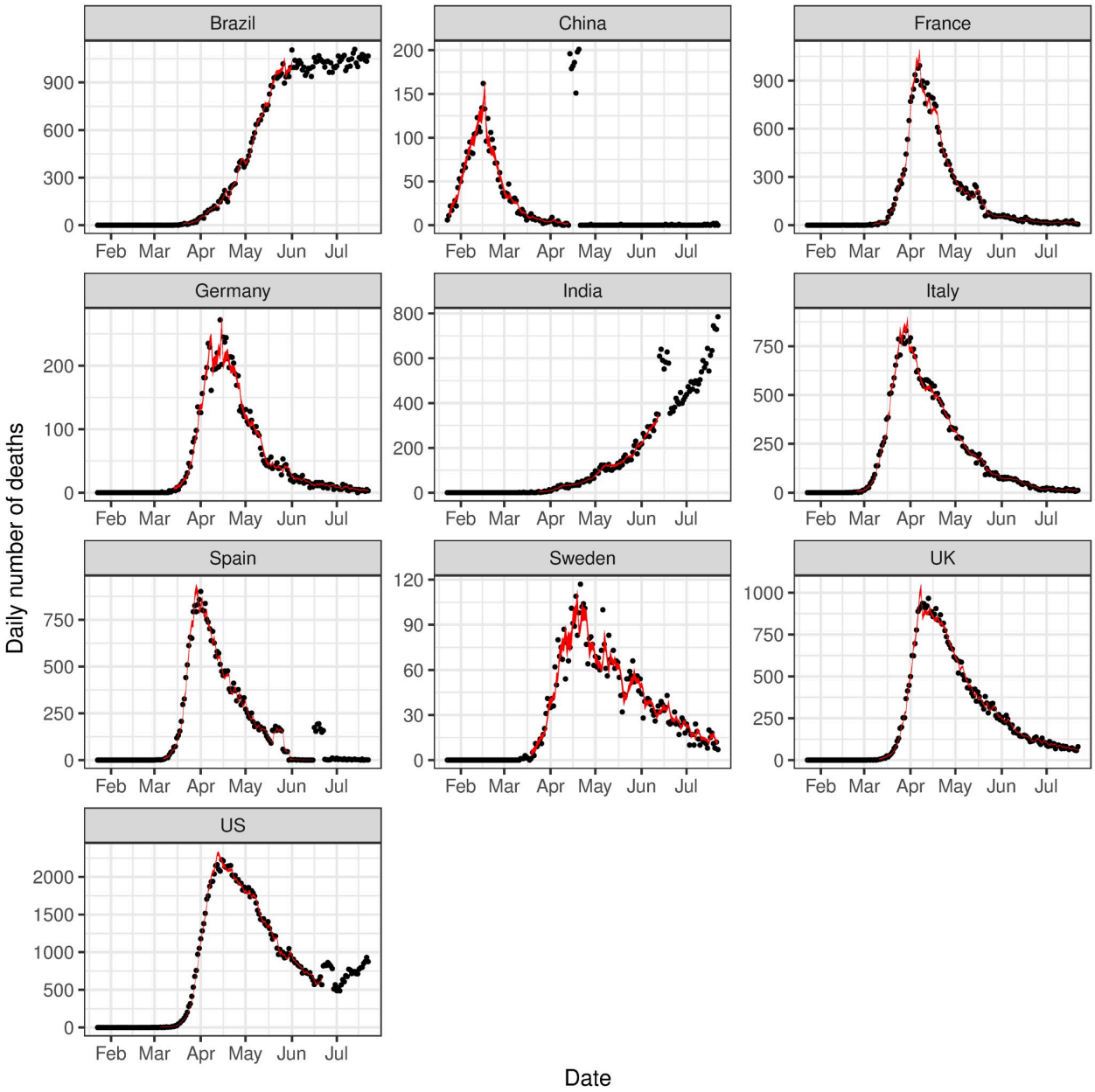


Fig 2. Observed deaths versus estimated deaths. The observed number of deaths (black dots) compared to the 95% posterior interval for the estimated expected number of events, i.e. $\lambda(t)$ (solid red ribbon).

https://doi.org/10.1371/journal.pone.0250015.g002

represents the expected number of events on day t and very closely follows the observed number of deaths. It is also extremely reactive to minor deviations from the observed trend, and more volatile times in the observed data result in wider posterior intervals to account for increased uncertainty in the trend of the data.

Diagnostic plots, including MCMC trace plots, autocorrelation between the MCMC samples and pairwise correlation between parameters were examined and suggest the algorithm has converged. Further details on the posterior distributions of the model parameters, convergence and model diagnostics are provided in S3 Appendix.

Table 1. Phase 1 versus Phase 2 median and 80% intervals for baseline parameters, μ_1 and μ_2.

Country	μ_1	μ_2
Italy	4.39 (3.18,5.71)	1.17 (0.69,1.8)
France	4.57 (3.38,5.91)	1.57 (0.97,2.28)
Spain	5.78 (4.06,7.6)	0.49 (0.28,0.76)
Germany	4.17 (2.89,5.54)	0.95 (0.59,1.39)
Sweden	4.05 (2.88,5.44)	1.79 (1.05,2.68)
U.K.	4.51 (3.08,6)	2.42 (1.32,3.75)
U.S.	4.08 (3.13,5.15)	4.1 (2.16,7.12)
China	8.92 (6.29,11.73)	0.82 (0.48,1.22)
Brazil	4.18 (2.98,5.52)	-
India	2.81 (2.02,3.72)	-

https://doi.org/10.1371/journal.pone.0250015.t001

Table 2. Phase 1 versus Phase 2 median and 80% intervals for magnitude parameters, α_1 and α_2.

Country	α_1	α_2
Italy	1.07 (1.05,1.09)	0.94 (0.93,0.95)
France	1.1 (1.08,1.11)	0.92 (0.91,0.93)
Spain	1.11 (1.09,1.13)	0.96 (0.95,0.97)
Germany	1.06 (1.03,1.09)	0.91 (0.89,0.93)
Sweden	1.07 (1.01,1.13)	0.92 (0.89,0.95)
UK	1.14 (1.11,1.17)	0.95 (0.95,0.96)
US	1.07 (1.06,1.07)	0.97 (0.97,0.98)
China	1.07 (1.01,1.15)	0.8 (0.76,0.84)
Brazil	1.03 (1.02,1.04)	-
India	1.1 (1.07,1.13)	-

https://doi.org/10.1371/journal.pone.0250015.t002

Tables 1–3 present the posterior median and corresponding 80% posterior intervals for the model parameters. Further details for the other baseline parameter priors considered can be found in S4 Appendix. In most countries, the posterior interval for μ_2 is consistently lower than μ_1, indicating a reduction in the baseline rate of events from the beginning to later stages of the epidemic. The exception to this is the U.S. The results for the U.S. are highly sensitive to the prior choice; thus, wider priors return higher posterior estimates than expected when

Table 3. Phase 1 versus Phase 2 median and 80% intervals for triggering kernel parameters, β_1 and β_2 and the means of their respective geometric distributions, β_1^{-1} and β_2^{-1}.

Country	β_1	β_2	β_1^{-1}	β_2^{-1}
Italy	0.88 (0.8,0.95)	0.55 (0.48,0.63)	1.136 (1.053,1.25)	1.818 (1.587,2.083)
France	0.97 (0.92,0.99)	0.64 (0.58,0.7)	1.031 (1.01,1.087)	1.562 (1.429,1.724)
Spain	0.96 (0.9,0.99)	0.91 (0.85,0.95)	1.042 (1.01,1.111)	1.099 (1.053,1.176)
Germany	0.65 (0.57,0.75)	0.51 (0.45,0.59)	1.538 (1.333,1.754)	1.961 (1.695,2.222)
Sweden	0.42 (0.32,0.54)	0.5 (0.39,0.62)	2.381 (1.852,3.125)	2 (1.613,2.564)
UK	0.79 (0.68,0.91)	0.56 (0.5,0.62)	1.266 (1.099,1.471)	1.786 (1.613,2)
US	0.99 (0.98,1)	0.77 (0.66,0.89)	1.01 (1,1.02)	1.299 (1.124,1.515)
China	0.4 (0.28,0.56)	0.43 (0.35,0.54)	2.5 (1.786,3.571)	2.326 (1.852,2.857)
Brazil	0.83 (0.73,0.93)	-	1.205 (1.075,1.37)	-
India	0.33 (0.26,0.41)	-	3.03 (2.439,3.846)	-

https://doi.org/10.1371/journal.pone.0250015.t003

compared to other countries. In an earlier analysis, this behaviour was also prevalent for Sweden and the U.K., although it disappeared when considering a longer time series. This implies that there may be insufficient information in the data for the U.S. to reliably learn the model parameters for the second phase. However, without alternative data, it is not possible to improve modelling for the U.S. by considering a longer time series. This is due to a large anomaly at the end of the series, as discussed in the Data section. Nonetheless, it highlights the importance of having sufficient training data and being cautious when interpreting parameter estimates.

The magnitude parameter in the second phase, α_2, is also consistently lower than the parameter for the first phase, α_1. With a posterior probability (greater than 80%), it can be said for all countries that $\alpha_1 > 1$ and $\alpha_2 < 1$. This implies the process is explosive before the change point and becomes stationary after the change point, likely driven by the introduction of interventions to reduce the rate of infection.

The parameters for the geometric triggering kernel, β_1 and β_2, are similar for Sweden and China. However, for the remaining countries where two phases are considered, the kernel parameter for the first phase, β_1, is larger than β_2, indicating that the self-excitation has a longer memory in the second phase. For reference, $\beta = 0.4$ in the geometric kernel corresponds to an average of 2.5 days for the self-excitation, with the majority of the mass occurring within one week, whereas $\beta = 0.9$ is shorter, corresponding to an average self-excitation of just over 1 day with approximately 2 days of total memory.

Model fit

Several measures are used to assess model fit. First, the model's capability to interpolate missing data is evaluated. Then in-sample and out-of-sample posterior predictive checks are considered. The purpose of prediction in this study is to assess model fit and to discover what can be learned about the process retrospectively.

The first measure of model fit considers how accurately the model can recover missing data. We randomly remove 10% of observations across the entire time series and treat the missing data as parameters in the model to estimate. Table 4 describes the number of interpolated data points for which the observed value lies within both the 95% and 80% credible intervals (CrI) of the posterior distributions for the missing data. Further details can be found in S5 and S6 Appendices. The proportion of data points correctly interpolated is generally high when considering the 95% credible intervals. This reduces when considering the 80% interval, however, is still high for most countries, capturing at least half of the missing data points.

Table 4. Number of missing data points with actual value within 95% and 80% CrIs, out of the total number of missing data points.

Country	95% CrI (average)	80% CrI (average)
France	11/14	7.4/14
Italy	13/15	11/15
Germany	13.4/14	10.2/14
Spain	8/11	6.2/11
Sweden	12.6/13	10.4/13
U.K.	11.8/14	9.2/14
China	8.6/9	7.2/9
U.S.	8.6/11	5.4/11
Brazil	6.6/8	4.6/8
India	7.8/9	6.8/9

https://doi.org/10.1371/journal.pone.0250015.t004

The exception to this is the U.S., with just less than half of the missing data points accurately interpolated.

Prediction is a difficult task, particularly for complex phenomena such as the COVID-19 pandemic. For this particular model, more recent events have a larger impact on the intensity of the process. Thus prediction performed at a time where abnormal behaviour is occurring will be highly uncertain and often unreliable. Moreover, a prediction is only realistic in the short term and generally only at times where there is no evidence of abnormal behaviour. This is consistent with other models in the literature [37, 38, 62–64]. Thus we consider in-sample and out-of-sample posterior predictive checks in this study as a measure of model fit only.

In-sample prediction is performed by generating sample paths of the process for the range of model parameters obtained and comparing these to the observed time series. In particular, a random selection of posterior samples is taken, and the entire time series is simulated from these draws. The posterior predictive intervals from these simulations compared to the observed data are given in Fig 3. In general, the intervals for these simulations encapsulate or are very close to the observed data, however, they can be extremely wide and often underestimate the volume of events in the initial phase of the outbreak. This is likely due to variation in the assumed Poisson data generating distribution, and relatively wide priors on the baseline parameters for the first phase, resulting in a wide range of possible sample paths. Additionally, these sample paths did not adequately capture the observed trend in the U.S. However, we find that including the data from the first phase in the model and predicting the second phase results in improved accuracy of the posterior predictive intervals for all countries. These results are presented in Fig 4.

Out-of-sample (O.O.S.) validation is also performed for each country as a measure of model fit. First, we consider the initial phase of the epidemic before the change point. The model is trained on data from the first 15 days of the sample, followed by a 5-day O.O.S. prediction. We then repeat this process, increasing the length of the training period by 5 days until the change point. As shown in Fig 5, these predictions are reliable only in the short term, and become more unreliable as the end of the first phase approaches. The first phase predictions grow exponentially and quickly surpass the actual growth of the process, as the observed curve flattens due to the effects of preventative measures that have been implemented.

O.O.S. prediction is also considered for the second phase of the model, after the change point. We first train the model on data from the first phase and 15 days of the second phase. We then repeat the same procedure as described above with 10-day O.O.S. predictions. The downward trajectory of the infection cycle is more stable than the upward trajectory, so we consider a longer prediction duration. The posterior predictive intervals are generally very accurate for all countries, as seen in Fig 5. Compared to the O.O.S. validation performed for the first phase, the improvements in accuracy observed in the second phase are likely due to the stationarity of the process in the second phase, resulting in more predictable trends. For both phases, the accuracy of O.O.S. predictions depends on the endpoint of the training period for the model, and the type of behaviour preceding any predictions.

While we do not attempt to predict the course of the epidemic in this study, we do find that O.O.S. predictions may indicate when the peak in the number of events is approaching. This could be useful in countries that have not yet experienced a decline in the number of daily events, for example, Brazil and India in this study. Posterior predictive intervals that surpass the growth rate in the observed data indicate, and could pre-empt, the downward phase of the epidemic. Conversely, where the predictive intervals do encapsulate the observed data, it is unlikely that the peak is being approached. This is evident in Fig 5, where the curve for Brazil is flattening, resulting in unreliable O.O.S. predictions, compared to the more reliable predictions in India due to the strong upward trend.

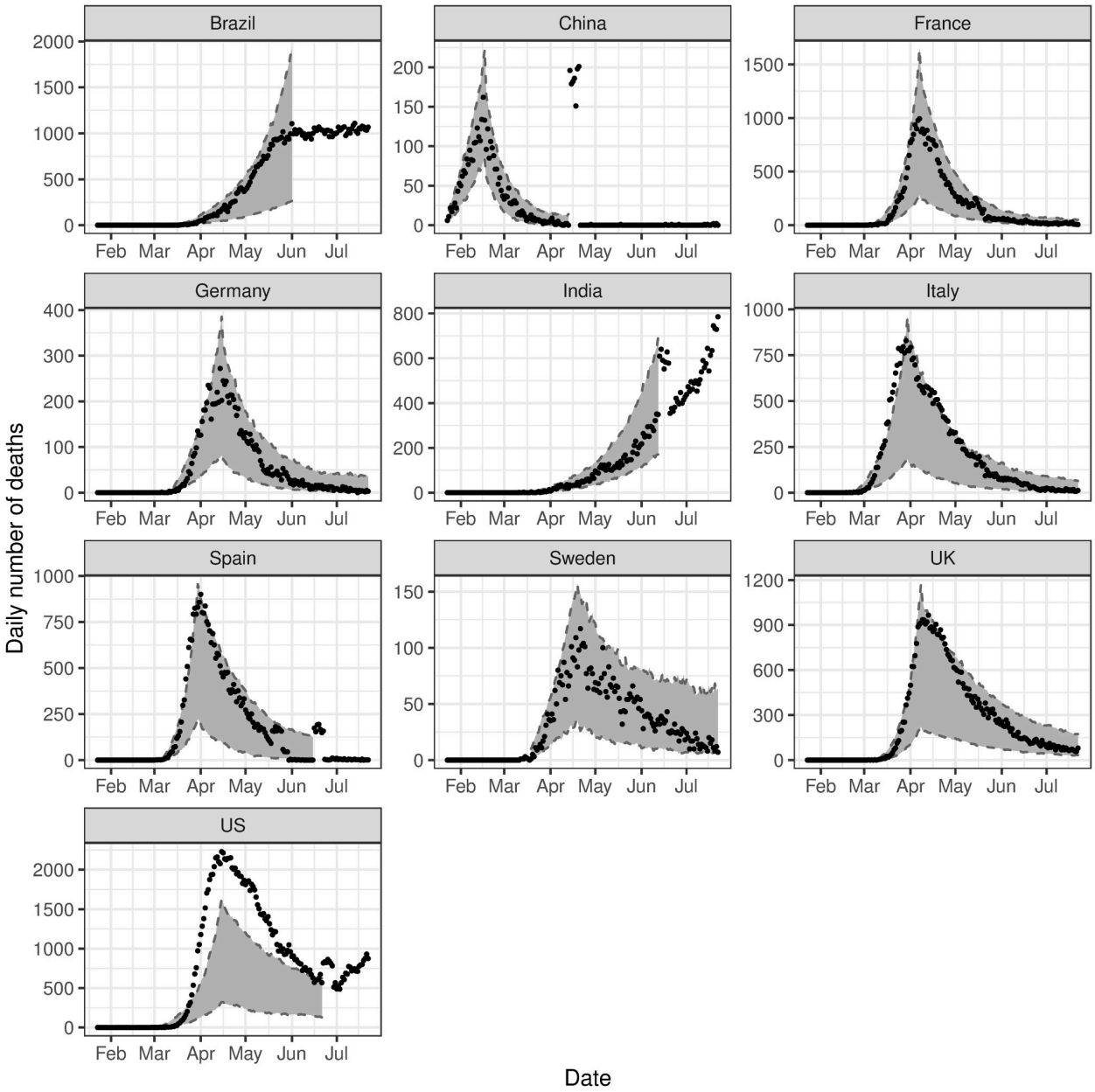


Fig 3. In-sample validation. The observed number of deaths (black dots) compared to the 95% posterior predictive interval for the estimated expected number of events, i.e. $\lambda(t)$ (grey ribbon).

https://doi.org/10.1371/journal.pone.0250015.g003

Fitting subsequent phases

As the pandemic progresses further waves of infection, and thus deaths, are inevitable and will continue to be of interest for the foreseeable future, particularly as a vaccine is rolled out and new variants of the virus are discovered. There is no obvious endpoint to the pandemic, however it is of interest to investigate subsequent waves of infection as well. To address this, we extend our main analysis to determine whether our proposed model is applicable over a longer time period.

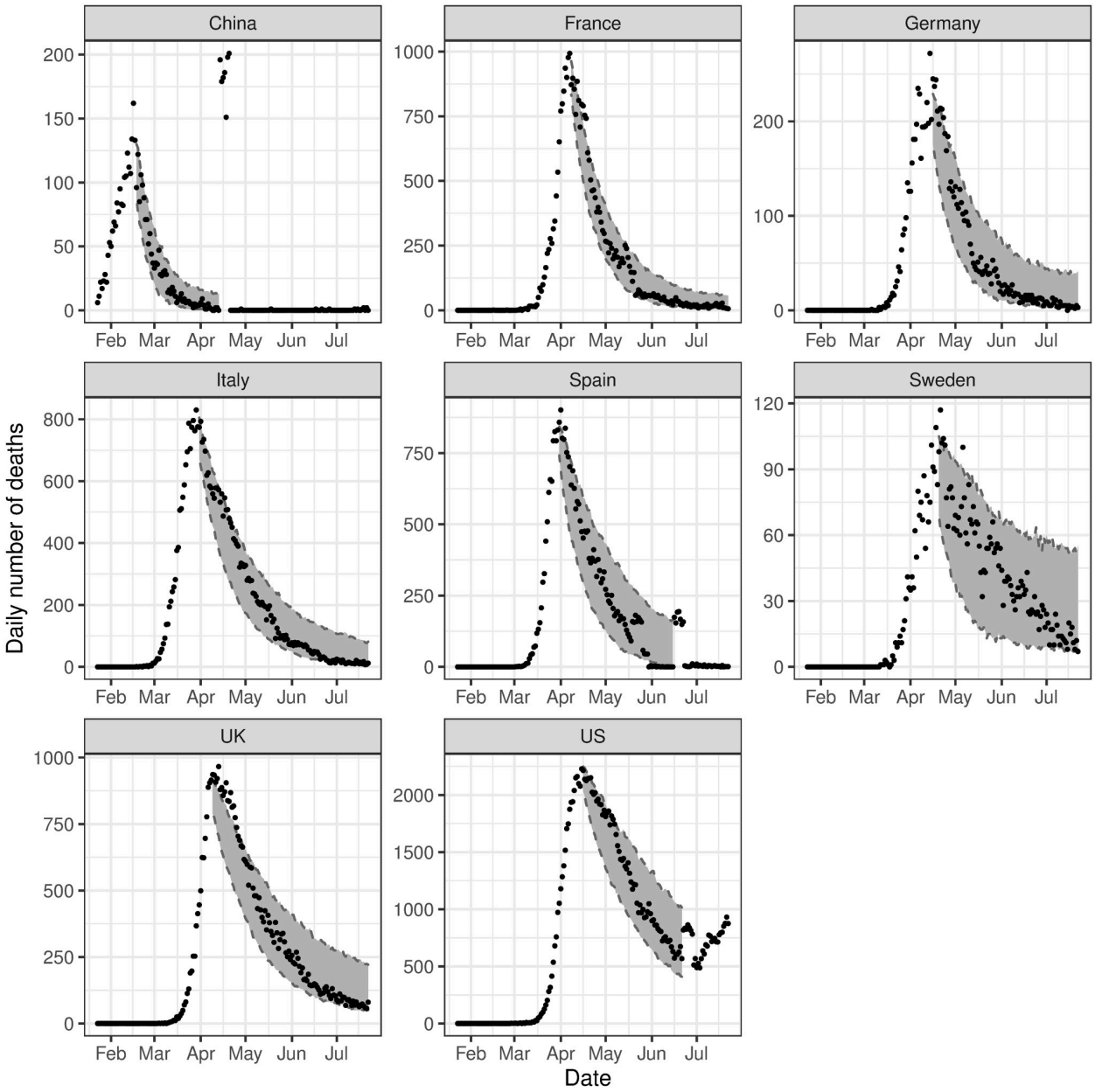


Fig 4. In-sample validation, conditioned on data from the first phase. The observed number of deaths (black dots) compared to the 95% posterior predictive interval for the estimated expected number of events, i.e. $\lambda(t)$ (grey ribbon).

https://doi.org/10.1371/journal.pone.0250015.g004

We consider mortality data from the endpoint of our initial analysis, up to 4th February 2021. Countries with inadequate data to inform another phase were cut short. As such, the observation period for Brazil, U.K and U.S end on 7th January 2021, 24th January 2021 and 12th January 2021 respectively. Furthermore, for many countries there is a period of very low mortality in between the first and second waves of infection, and we do not consider this period. Additionally, China has not experienced a second wave, and thus it is excluded from this subsequent analysis.

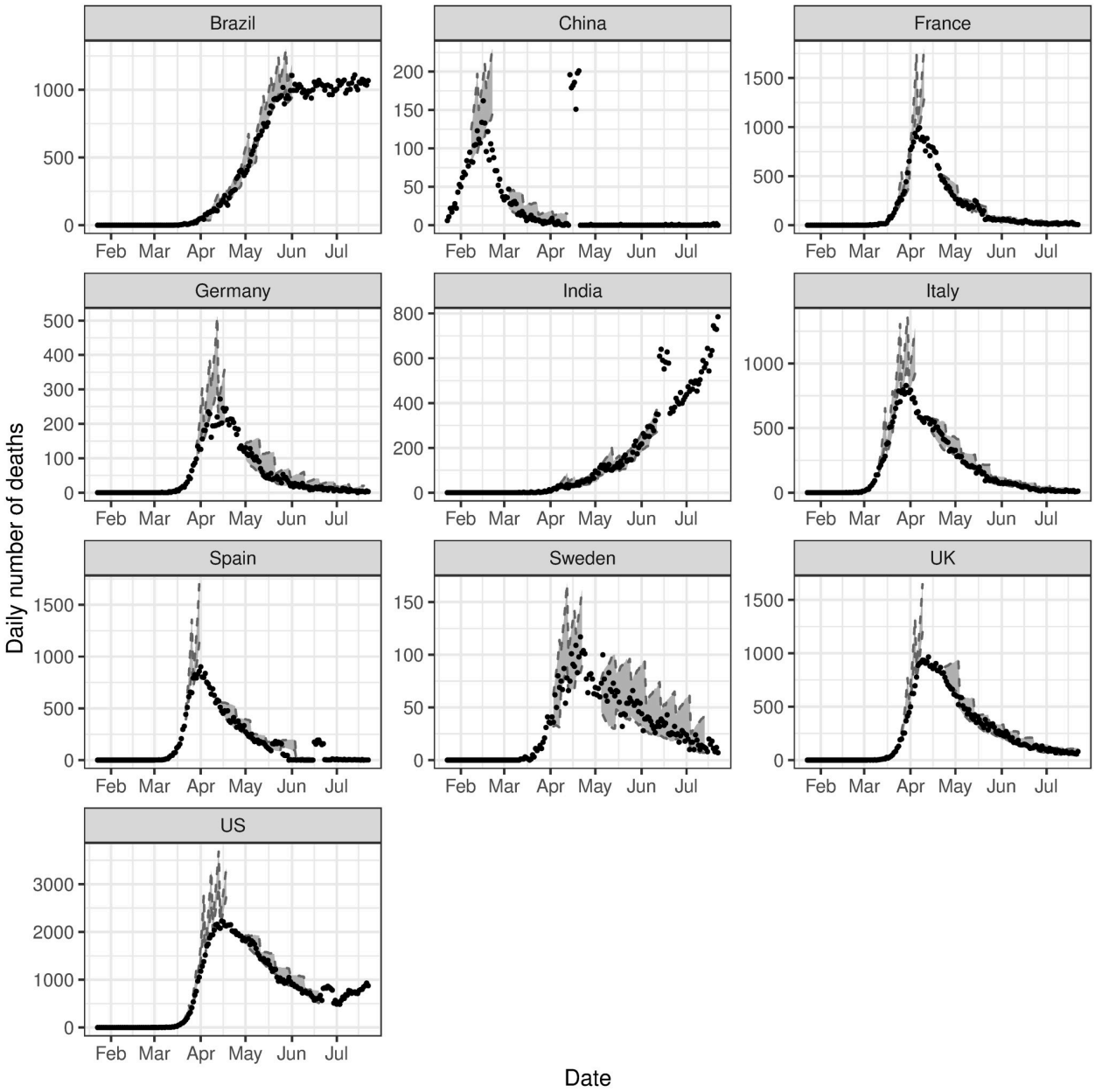


Fig 5. Out-of-sample validation. The observed number of deaths (black dots) compared to the 95% posterior predictive interval for the estimated expected number of events, i.e. $\lambda(t)$ using various training datasets (grey ribbons).

https://doi.org/10.1371/journal.pone.0250015.g005

Change points were selected where there were obvious changes in the trajectory, in a similar fashion as the main analysis. The starting point of the second wave was selected as the time where either the 2 week or 4 week rolling average increases by 50% in a single week. The choice between a 2 or 4 week rolling average is chosen based on which more closely aligns to the start of the second wave upon visual inspection. We note that automatic change point detection algorithms such as the CUSUM algorithm [65] were considered, however, they are not appropriate for our model. These algorithms are generally based on the mean of the time series. Given the self-exciting nature of our model, changes in the intensity of the process do not

necessarily indicate changes in the underlying model parameters. The change points selected can be found in S7 Appendix.

Comparing the parameter estimates between the initial analysis and this subsequent analysis, several observations can be made. The full table of estimates can be found in S2 Table. Generally, while the baseline parameter μ in the initial analysis shows a reduction between the first and second phases, in subsequent phases the baseline mean begins to increase again. This is potentially due to the relaxing of restrictions and the opening of international borders. The magnitude parameter α acts as expected, in other words it is less than 1 for phases with a downward trajectory and greater than 1 for phases with an upward trajectory. In the initial analysis, β is generally close to 1 in the first phase and reduces in subsequent phases.

Fig 6 shows the estimated intensity function against the observed data for the subsequent analysis. We find that the estimated intensity follows very closely to the observed data, as is also seen in the main analysis. We also consider in-sample (Fig 7) and out-of-sample validation (Fig 8), in the same manner as the main analysis. These both show promising results, with both in-sample and out-of-sample predictions aligning very closely to the observed data. The residuals, in this case referring to the difference between the observed data and the estimated intensity, for all phases in both the initial and subsequent analysis are provided in S8 Appendix, and show that the models for both sets of analyses are reasonable.

Discussion

There are many strengths to our work, and some important considerations that needed to be made. We first discuss the main findings of this analysis. This is followed by detailing the limitations and potential extensions. Lastly we compare our model methodology to several popular approaches for modelling this type of phenomena.

DTHP model

Infectious diseases have previously been studied using Hawkes processes. However, the scale, severity and uncertainty of the current COVID-19 pandemic make it a very challenging problem, providing a unique opportunity to evaluate the capacity of Hawkes processes in describing an incredibly complex process. Another source of complexity arises from the definition of what constitutes a COVID-19 death, which differs between countries. This analysis finds that by modifying the DTHP to incorporate change points, our model can adequately capture the overall process as distinct phases, while quickly reacting to and accommodating for some level of abnormal behaviour.

The findings of this work can also quantify the dynamics of these distinct phases in the pandemic. Our results from the initial analysis show that for the baseline parameters, the background rate in the second phase, μ_2, is lower than that for the first phase, μ_1. This is analogous to a reduction in the baseline level of exogenous events, possibly related to reduced travel and general mobility. Another factor could be increased levels of community transmission, affecting the self-exciting component of the intensity function, and thus placing less emphasis on the baseline component. In subsequent phases, μ begins to increase again, which suggests an increase in movement between countries. The exception to this is the U.S., for the reasons stated in previous sections. The baseline parameter could also be affected by the definition of a reported COVID-19 death, as this differs between countries. For example, when the criteria for reporting a death excludes cases where the person suffers from other illnesses in addition to the virus, this could result in an inflated baseline rate, as secondary events from unreported cases could be present in the data.

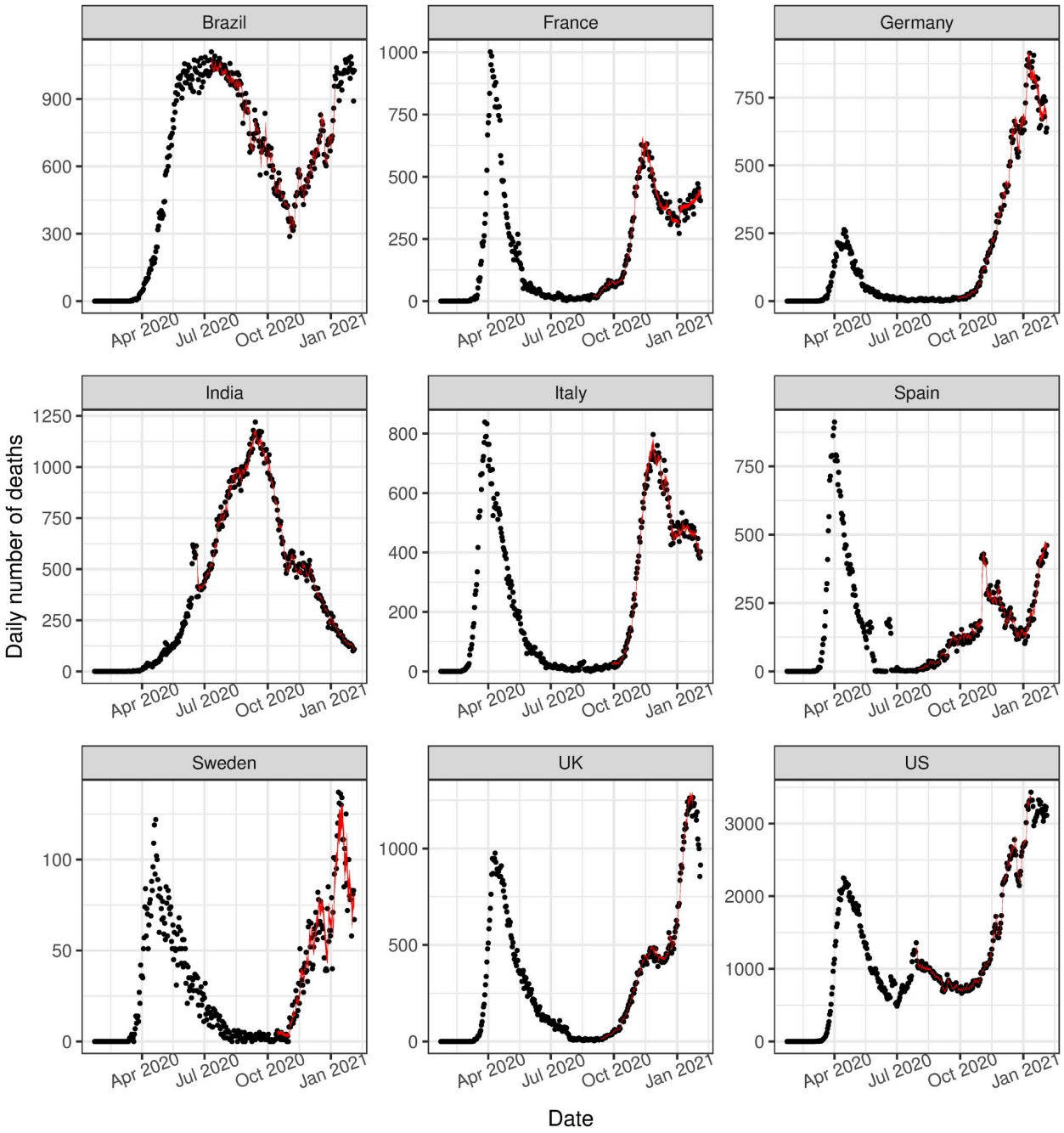


Fig 6. Observed deaths versus estimated deaths (subsequent analysis). The observed number of deaths (black dots) compared to the 95% posterior interval for the estimated expected number of events, i.e. $\lambda(t)$ (solid red ribbon), for the subsequent analysis.

https://doi.org/10.1371/journal.pone.0250015.g006

Our initial results for the magnitude parameters show, with a high degree of certainty, that for the first phase α_1 is greater than 1, and for the second phase α_2 is less than 1. This exhibits the distinct differences between phases, as a magnitude parameter greater than 1 indicates the process itself is non-stationary, and similarly a magnitude parameter less than 1 suggests a stationary process. This pattern is also evident in the analysis of subsequent phases. We discuss

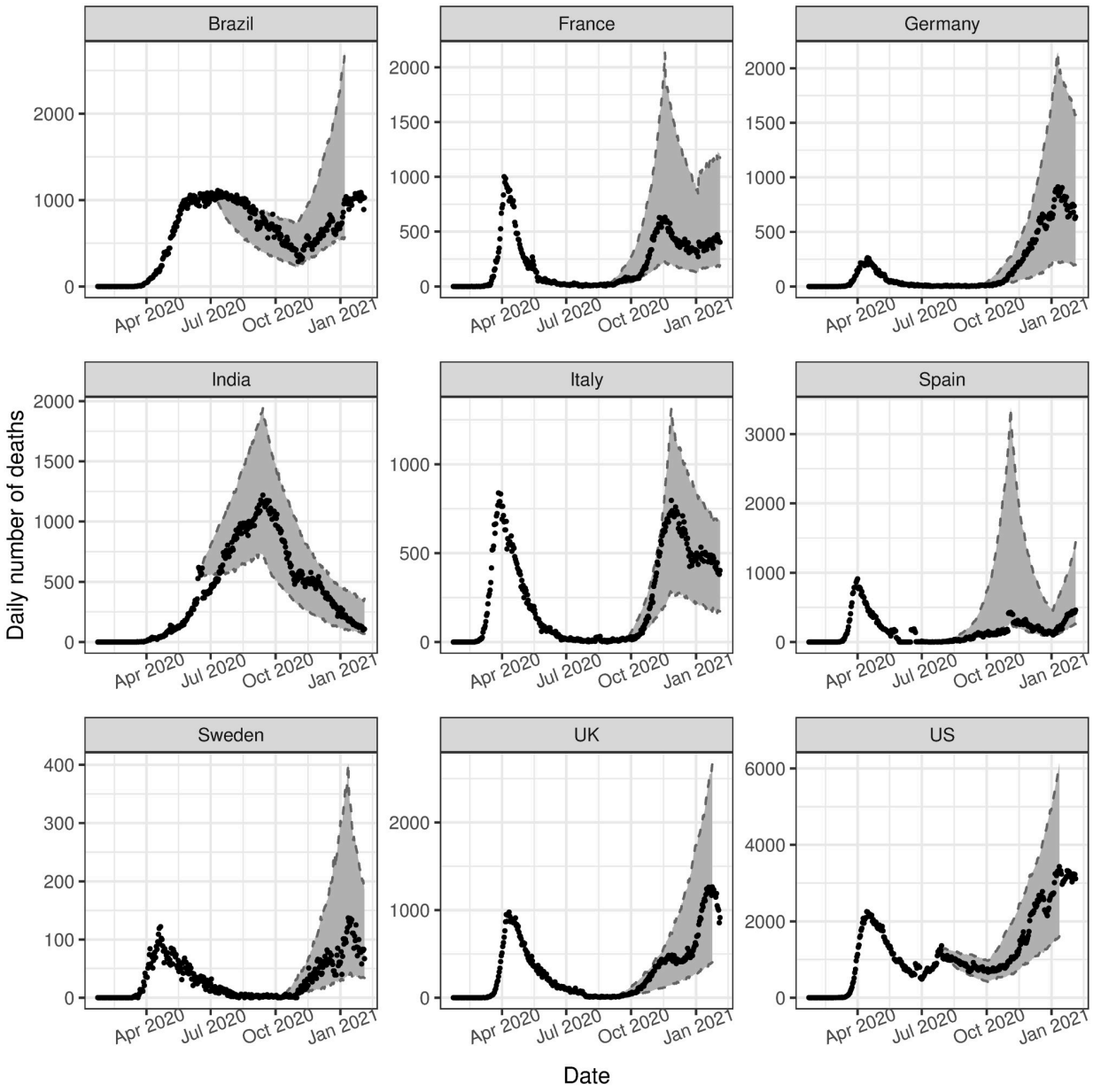


Fig 7. In-sample validation for subsequent analysis, conditioned on data from the initial analysis. The observed number of deaths (black dots) compared to the 95% posterior predictive interval for the estimated expected number of events, i.e. $\lambda(t)$ (grey ribbon), for the subsequent analysis.

https://doi.org/10.1371/journal.pone.0250015.g007

below the similarities between the magnitude parameters in our model and the reproduction number in standard epidemiological models.

The triggering kernel parameter in the first phase, β_1, is higher than that for the second phase, namely β_2, for all countries except Sweden and China. This could suggest that in later stages of the epidemic when preventative measures have been implemented, the time between transmission is longer, as there is less opportunity for transmission. The two exceptions to this, Sweden and China, are on opposite ends of this spectrum. While China enforced very

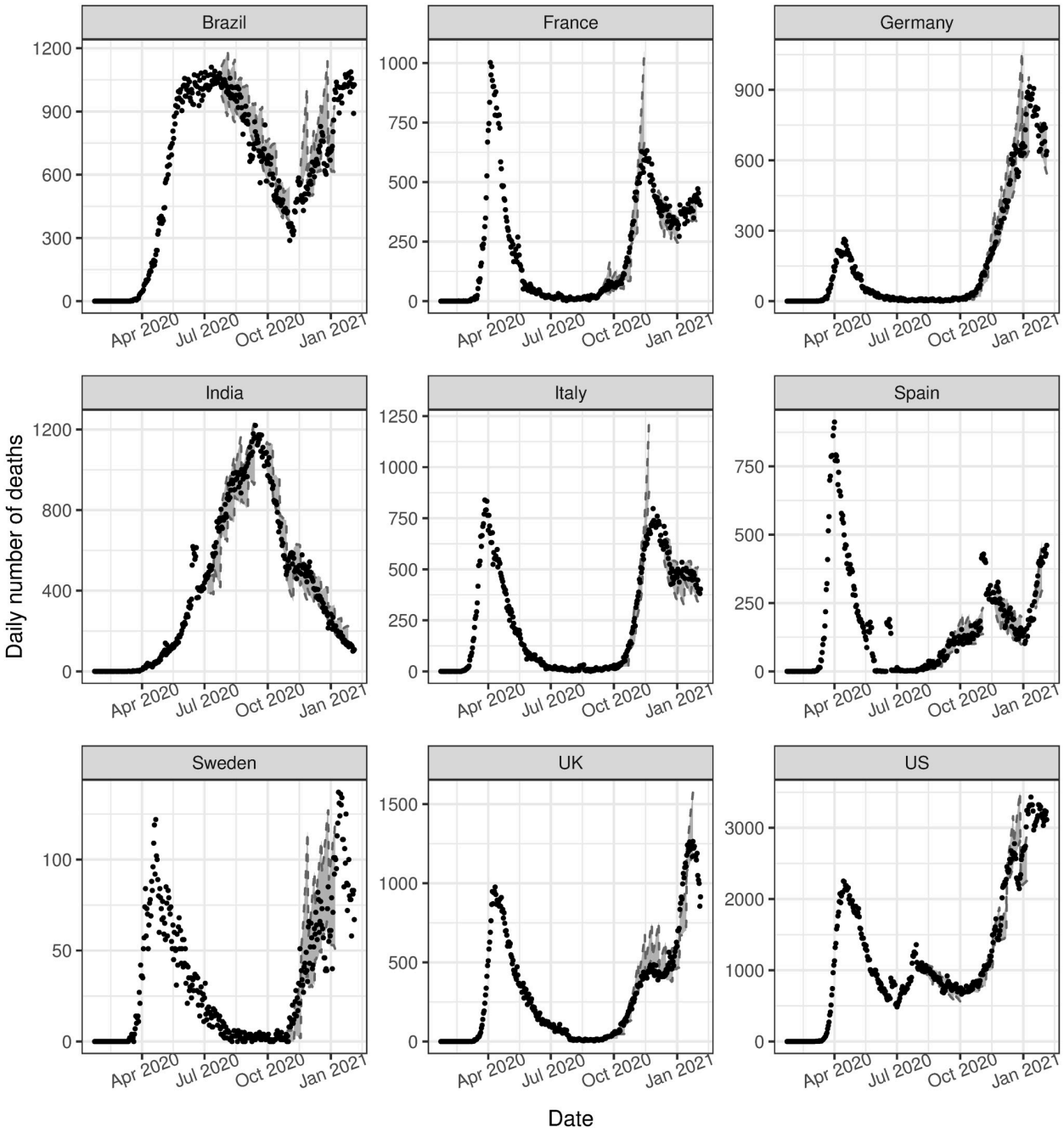


Fig 8. Out-of-sample validation (subsequent analysis). The observed number of deaths (black dots) compared to the 95% posterior predictive interval for the estimated expected number of events, i.e. $\lambda(t)$ using various training datasets (grey ribbons) for the subsequent analysis.

https://doi.org/10.1371/journal.pone.0250015.g008

strict lockdown and quarantine requirements, Sweden adopted a soft approach to lockdown. Large β_1 values could also be an indication of instability in the initial phase of the pandemic, leading to difficulty in predicting and discerning patterns in the data. Additionally, this could be a result of death data being less reliable in early phases, as the process of counting COVID-19 deaths was not yet established.

Throughout the initial stage of this analysis, we have found difficulty in fitting the proposed model for the U.S. In particular, the posterior estimates for the baseline parameter are uncertain as they are heavily influenced by the prior choice. Additionally, in-sample posterior predictive checks found that the sample paths produced by the estimated model parameters do not resemble the observed trend. We consider the U.S. an anomaly, as their response to the virus by the relevant state-level authorities varied widely between states. While this is also true to an extent for other countries, the heterogeneity across the country was arguably more significant for the U.S., implying that the proposed model may need to be applied at a more granular level of regions to obtain more reliable results.

Despite our approach being able to accurately capture the dynamics of this complex process, we now address some limitations and extensions that could be considered. As the epidemic is still ongoing, new data is becoming available each day, and the model must be re-fit and tuned each time the data is updated. While we somewhat manually select change points in this analysis, an algorithm suitable to this model with automatic selection of the number of change points and their respective locations could also be considered. Additional change points need to be determined carefully as there must be sufficient information in each time series to inform parameter estimation. Another consideration is flexible Bayesian nonparametric splines [66] or other methods to provide time-varying parameters. However, the identifiability and existence of this model would need to be established. One could also consider different triggering kernels, including nonparametric kernels in order to improve the flexibility of the model. Another possible extension is considering covariates related to COVID-19 deaths, such as the number of people travelling and number of hospitals per capita.

Comparison with other approaches

Here we discuss several of the many approaches that have been considered to model the ongoing COVID-19 epidemic, and the different perspectives they provide compared to our DTHP model. Compartmental models such as the SIR family of models are among the most popular methods for epidemic modelling. They are more detailed and consider the mechanics of the infection cycle, separating the population into categories such as susceptible, infected and recovered or deceased. Our DTHP model is simplified in the sense that we consider only death events. We chose to model deaths instead of infection numbers as the latter data was very unreliable in the beginning due to lack of testing and different testing policies across countries. However, as we show in S1 Appendix, as a first-order approximation, the death dynamics are helpful to understand the infection dynamics. This approximation is convenient when the infection data are unreliable, as occurred in the early stages of the COVID-19 pandemic. In the presence of data uncertainty such as this, the SIR model requires additional terms to account for this measurement error.

To compare the two frameworks, it is helpful to consider a stochastic variation of the SIR model as a bivariate Poisson process, comprised of infection and recovery events. Infection events are then governed by a Poisson process where the rate is based on the transmission rate and the current size of the susceptible and infected populations, corresponding to the rate of infection in the deterministic SIR model. Our model differs as we consider a discrete time scale, the daily number of events is Poisson-distributed and, conditioned on past events, the rate of events each day is given by Eq (2).

Another significant difference between our model and standard compartmental models is that the latter considers a finite population. In its original form, the Hawkes model assumes that there will be immigrant events arriving at a rate of the baseline mean μ indefinitely, implying an infinite population. However, finite population variants of the Hawkes model do exist

[43]. This differs from the SIR model, which naturally considers a finite population whereby the infection dies out once herd immunity is achieved. The impact of this difference is negligible in our modelling because we predominantly model the pandemic's initial phases, where not enough of the population has been infected or vaccinated to achieve herd immunity. This may not be the case for more prevalent diseases such as the flu, however both models are reasonable. As the flu season ends, there will still be new infections throughout the year, however on a smaller scale.

Hence our approach provides a simple model for unknown and volatile phenomena such as the COVID-19 pandemic, particularly in the early stages of the outbreak. Unlike the common flu, where the dynamics and course of infection are well understood and relatively predictable, COVID-19 is a new and unexplored domain. The various interventions that take place simultaneously result in complex interactions that complicate the dynamics of the process. Our focus is on the early stages of the epidemic where there is a great deal of uncertainty and volatility. The SIR model family is useful for phenomena where the mechanics are well known. However, complicated variants of these models are required to capture the complexity of this pandemic. Our simple model is useful in describing this early stage in the pandemic when there are still many unknowns. Our model also introduces randomness and flexibility that is not afforded in standard compartmental models. This allows our model to adapt to system changes induced by government interventions quickly.

The family of SIR models naturally follow the pattern of infections and deaths rising to a peak and then falling due to a reduction in the susceptible population. However, this is not the cause of the fall observed in the early stage of the pandemic. Instead, the fall is driven by external factors such as social distancing measures, temperature, and improvement in treatments, to name a few. SIR family models have also incorporated change points or time-varying parameters to account for these alternative drivers [51, 52]. Given our analysis's retrospective nature, the change points were quite obvious, and we did not estimate them. However, our Hawkes model can be easily augmented to induce this shape naturally. For example, we could consider a mixture of Hawkes processes for each of these distinct phases, estimate the unknown (or known) change points, or incorporate time-varying parameters.

Another more complex approach is that of agent-based modelling. These are more detailed than compartmental models, and are very useful if you have an understanding of the underlying mechanisms. Recent papers using this approach for the COVID-19 epidemic, referenced in the introduction, reveal the non-random nature of the underlying stochastic processes. Based on fluctuations in social participation and certain biological factors, they lead to the infection spreading, hospitalisation, and eventually to fluctuations of the fatality rate.

Alternatively, one could consider an even more straightforward approach, such as a piecewise exponential model. However, the Hawkes process allows for uncertainty in the model that is not possible with the exponential growth model, which is very strict and captures only the data trend. Allowing fluctuations in the data—particularly for volatile phenomena such as the current pandemic—is an essential aspect of providing a realistic model. The exponential model also becomes less appropriate as the pandemic progresses. In later phases, there are complex interactions that result in trajectories that are inherently not exponential. These are uncertain times, and our model strikes a balance between modelling the dynamics of the whole infection cycle and fitting a generic exponential model. We model some fluctuations motivated by the physical process, but with a simpler model than many others considered in the literature.

While there are many alternative approaches available, the Hawkes model is also a natural model for describing self-exciting phenomena. It provides a flexible and stochastic framework for modelling, and the parameters in our model provide interesting insights into the

pandemic. Namely, α is the average number of secondary infections and is related to the reproduction number, $\frac{\alpha}{\beta}$ is related to the average time an infected individual has infected someone, and μ relates to the occurrence of external excitations, or rather contaminations weighted by the probability of death given contamination. The β parameter on its own also indicates how the time between infections changes throughout time.

The reproduction number, defined as the number of secondary infections from a single case, is a crucial parameter in epidemiological models. Similarly, the magnitude parameters in our model, given by α, also represent the expected number of secondary cases caused by a single parent event. While their respective interpretations are similar at a superficial level, α is not directly comparable to reproduction numbers in epidemiological models. This is due to differences in model assumptions and the underlying mathematical frameworks, as our model's magnitude parameters do not provide the same information as the effective reproduction number. The effective reproduction number informs the level of herd immunity that will bring the virus under control, and the proportion of new infections that must be prevented to change the trend of events from increasing to decreasing [67], whereas our model parameters do not. However, we note that, similarly to reproduction numbers, if $\alpha > 1$ in our model there is exponential growth in the number of events and $\alpha < 1$ leads to a stationary model, which translates into a decrease in the number of deaths if the phase begins at a time with a high event intensity. We also consider a static variable that fundamentally averages over the whole period, rather than varying through time as the effective reproduction number would. We do this as reasonable change points were fairly obvious in the dataset used for this analysis. However, for more complex trajectories, other authors [44, 45] consider a Hawkes model with a time-varying magnitude parameter, which they refer to as a dimensionless reproduction number. This approach could inform the change point's location by observing when the magnitude parameter goes below 1. The change points could also be estimated, for example using the method suggested in [68].

Other key epidemiological parameters are generation times and serial intervals, which describe the time between infection and development of symptoms, respectively, for a pair of individuals. Our model does not capture this type of information, as we do not consider the relationship between specific pairs of individuals. As a result, it is not possible to obtain parameters such as growth rates, which are often of interest in epidemiological models. However, we can gain insight into an alternative temporal aspect of the contagion. The geometric triggering kernel in our model describes how the probability of contagion changes as time elapses. More precisely, we can determine, for a given day, the influence of past events on the expected number of events for that day.

Conclusion

The utility of our model is not restricted to the current coronavirus epidemic, and could be used as a simple model to describe a much broader range of complex phenomena. We have demonstrated through this study that the proposed model is a simple, yet powerful tool for explaining an incredibly complex process. In general, models that attempt to describe complex processes can become increasingly complicated, as more intricate details are embedded and accounted for in the modelling. Thus having a parsimonious model that is flexible enough to competently capture the dynamics of a complex process, without adding too much additional complexity, is very desirable.

In particular for the current pandemic, this study shows that our simple discrete-time Hawkes process can capture the dynamics for different countries, despite the complexities involved with each country's unique response to the virus. The same underlying biological

process is affecting countries in different ways, and there is a significant difference in the impact and severity of the pandemic across different countries. Additionally, the actions that have been taken to stop the spread, and the timing of these also vary widely. These different behaviours between countries mean that the evolution of the pandemic for an individual country is very intricate within itself, and involves many unseen and complex hidden interactions that we cannot model directly. However, the proposed model, while being very simple, can capture these trends surprisingly well.

To adequately model the entire course of the pandemic, we find that we must make provisions as there are multiple distinct phases. Initially, there is exponential growth as the virus spreads, followed by a period of reduced infection rates as actions are taken to slow the spread. These distinct behavioural differences throughout the evolution of the epidemic must be acknowledged, as a single DTHP applied to the entire time series provides uninformative and uninterpretable parameter estimates. Hence a model that accounts for these different phases, such as the model presented in this work, is required.

Fitting a DTHP to the epidemic has led to some other unique insights. Our results show that a discrete-time model is appropriate for this application, avoiding unnecessary computational burden as well as additional noise due to artificial data imputation, as is required for the continuous-time model. This model also provides to an extent, interpretable parameters and an indication of the changing dynamics between distinct phases of the pandemic. We show that despite unique circumstances for individual countries, including the type and timing of non-pharmaceutical interventions, population demographics, and the overall impact of the virus, the model is flexible and can also accomodate some level of volatility in the data. Furthermore, one of the most surprising outcomes of this analysis is that, at the country level, a very simple DTHP model fits remarkably well to the number of deaths, thus capturing the dynamics of the COVID-19 pandemic.

Supporting information

S1 Appendix. Justification for Hawkes model on deaths.
(PDF)

S2 Appendix. About the average excitation duration.
(PDF)

S3 Appendix. Convergence and diagnostic plots for initial and subsequent analysis. Top left hand panel: compares the observed number of deaths (black dots) with the 95% posterior interval for the estimated expected number of events (solid red ribbon). Top right hand panel: shows pairwise correlation between all parameters in the lower triangle, corresponding correlation values in the upper triangle, and the marginal posterior densities for each parameter on the diagonal. Bottom panel: shows trace plots on the top row and the autocorrelation function on the bottom row for each parameter. All figures were generated after thinning the posterior samples.
(PDF)

S4 Appendix. Parameter estimates of baseline parameters for all prior choices. Phase 1 versus Phase 2 median and 80% intervals of baseline parameters for countries with two phases.
(PDF)

S5 Appendix. Missing data interpolation. Tables containing number of missing data points with actual value within 80% and 95% posterior interval, for all prior choices.
(PDF)

S6 Appendix. Figures from missing data interpolation. The histogram represents the estimated posterior distributions for each of the missing data points. The black dashed lines show the 95% credible intervals around the posterior distributions. The solid blue line displays the observed number of deaths.
(PDF)

S7 Appendix. Change point locations.
(PDF)

S8 Appendix. Plot of residuals. For each country and phase, we calculate the estimated expected intensity of the process (i.e. $\lambda(t)$) using the samples of the parameter estimates obtained through the estimation procedure. The histograms then represent the median residual value (median of the difference between the observed number of events and the estimated expected intensity).
(PDF)

S1 Table. Results from leave-future-out cross validation with Pareto smoothed importance sampling. Expected log predictive density (ELPD) for a range of prior choices. Maximum ELPD in bold.
(PDF)

S2 Table. Parameter estimates for original and subsequent analysis. Comparison of median and 80% intervals of parameters for all phases, using the Gamma(5, 1) prior for μ.
(PDF)

Acknowledgments

The authors are grateful to Dr Gentry White, for helpful advice on modelling discrete-time Hawkes processes in the early stages of this project.

Author Contributions

Conceptualization: Raiha Browning, Deborah Sulem, Kerrie Mengersen, Vincent Rivoirard, Judith Rousseau.

Formal analysis: Raiha Browning, Deborah Sulem.

Methodology: Raiha Browning, Deborah Sulem, Kerrie Mengersen, Vincent Rivoirard, Judith Rousseau.

Supervision: Kerrie Mengersen, Vincent Rivoirard, Judith Rousseau.

Validation: Raiha Browning.

Visualization: Raiha Browning.

Writing – original draft: Raiha Browning.

Writing – review & editing: Raiha Browning, Deborah Sulem, Kerrie Mengersen, Vincent Rivoirard, Judith Rousseau.

References

1. World Health Organisation. Weekly Epidemiological Update for Coronavirus disease 2019 (COVID-19) —9 March 2021; 2021. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20210309_weekly_epi_update_30.pdf.

2. Hellewell J, Abbott S, Gimma A, Bosse NI, Jarvis CI, Russell TW, et al. Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. The Lancet Global Health. 2020; 8(4):e488–e496. https://doi.org/10.1016/S2214-109X(20)30074-7 PMID: 32119825
3. Plank MJ, Binny RN, Hendy SC, Lustig A, James A, Steyn N. A stochastic model for COVID-19 spread and the effects of Alert Level 4 in Aotearoa New Zealand. medRxiv. 2020;.
4. Fowler JH, Hill SJ, Obradovich N, Levin R. The effect of stay-at-home orders on COVID-19 cases and fatalities in the United States. medRxiv. 2020;.
5. Peak CM, Kahn R, Grad YH, Childs LM, Li R, Lipsitch M, et al. Individual quarantine versus active monitoring of contacts for the mitigation of COVID-19: a modelling study. The Lancet Infectious Diseases. 2020; 20(9):1025–1033. https://doi.org/10.1016/S1473-3099(20)30361-3 PMID: 32445710
6. Kucharski AJ, Klepac P, Conlan AJK, Kissler SM, Tang ML, Fry H, et al. Effectiveness of isolation, testing, contact tracing, and physical distancing on reducing transmission of SARS-CoV-2 in different settings: a mathematical modelling study. The Lancet Infectious Diseases. 2020; 20(10):1151–1160. https://doi.org/10.1016/S1473-3099(20)30457-6 PMID: 32559451
7. Davies NG, Kucharski AJ, Eggo RM, Gimma A, Edmunds WJ, Jombart T, et al. Effects of non-pharmaceutical interventions on COVID-19 cases, deaths, and demand for hospital services in the UK: a modelling study. The Lancet Public Health. 2020; 5(7):e375–e385. https://doi.org/10.1016/S2468-2667(20)30133-X PMID: 32502389
8. Kretzschmar ME, Rozhnova G, Bootsma MCJ, van Boven M, van de Wijgert JHHM, Bonten MJM. Impact of delays on effectiveness of contact tracing strategies for COVID-19: a modelling study. The Lancet Public Health. 2020; 5(8):e452–e459. https://doi.org/10.1016/S2468-2667(20)30157-2
9. Badr HS, Du H, Marshall M, Dong E, Squire MM, Gardner LM. Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study. The Lancet Infectious Diseases. 2020; 20(11):1247–1254. https://doi.org/10.1016/S1473-3099(20)30553-3
10. Chen Y, Cheng J, Jiang Y, Liu K. A time delay dynamic system with external source for the local outbreak of 2019-nCoV. Applicable Analysis. 2020;. https://doi.org/10.1080/00036811.2020.1732357
11. Wangping J, Ke H, Yang S, Wenzhe C, Shengshu W, Shanshan Y, et al. Extended SIR Prediction of the Epidemics Trend of COVID-19 in Italy and Compared With Hunan, China. Frontiers in Medicine. 2020; 7(169). https://doi.org/10.3389/fmed.2020.00169 PMID: 32435645
12. Roques L, Klein EK, Papaix J, Sar A, Soubeyrand S. Using Early Data to Estimate the Actual Infection Fatality Ratio from COVID-19 in France. Biology. 2020; 9(5). https://doi.org/10.3390/biology9050097 PMID: 32397286
13. Giordano G, Blanchini F, Bruno R, Colaneri P, Di Filippo A, Di Matteo A, et al. Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. Nature Medicine. 2020; 26:855–860. https://doi.org/10.1038/s41591-020-0883-7 PMID: 32322102
14. Warne DJ, Ebert A, Drovandi C, Hu W, Mira A, Mengersen K. Hindsight is 2020 vision: Characterisation of the global response to the COVID-19 pandemic. medRxiv. 2020;. https://doi.org/10.1186/s12889-020-09972-z PMID: 33287789
15. Prem K, Liu Y, Russell TW, Kucharski AJ, Eggo RM, Davies N, et al. The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. The Lancet Public Health. 2020; 5(5):e261–e270. https://doi.org/10.1016/S2468-2667(20)30073-6 PMID: 32220655
16. Zhan C, Tse CK, Lai Z, Hao T, Su J. Prediction of COVID-19 spreading profiles in South Korea, Italy and Iran by data-driven coding. PLOS ONE. 2020; 15(7):e0234763. https://doi.org/10.1371/journal.pone.0234763
17. Li Y, Wang LW, Peng ZH, Shen HB. Basic reproduction number and predicted trends of coronavirus disease 2019 epidemic in the mainland of China. Infectious Diseases of Poverty. 2020; 9(94). https://doi.org/10.1186/s40249-020-00704-4 PMID: 32678056
18. Zu J, Li ML, Li ZF, Shen MW, Xiao YN, Ji FP. Transmission patterns of COVID-19 in the mainland of China and the efficacy of different control strategies: a data- and model-driven study. Infectious Diseases of Poverty. 2020; 9(83). https://doi.org/10.1186/s40249-020-00709-z PMID: 32631426
19. Agosto A, Giudici P. A Poisson Autoregressive Model to Understand COVID-19 Contagion Dynamics. Risks. 2020; 8(3):1–8. https://doi.org/10.3390/risks8030077
20. Flaxman S, Mishra S, Gandy A, Unwin HJT, Mellan TA, Coupland H, et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. Nature. 2020; 584:257–261. https://doi.org/10.1038/s41586-020-2405-7 PMID: 32512579
21. Zou Y, Pan S, Zhao P, Han L, Wang X, Hemerik L, et al. Outbreak analysis with a logistic growth model shows COVID-19 suppression dynamics in China. PLOS ONE. 2020; 15(6):e0235247. https://doi.org/10.1371/journal.pone.0235247 PMID: 32598342

22. Musa SS, Zhao S, Wang MH, Habib AG, Mustapha UT, He D. Estimation of exponential growth rate and basic reproduction number of the coronavirus disease 2019 (COVID-19) in Africa. Infectious Diseases of Poverty. 2020; 9(96). https://doi.org/10.1186/s40249-020-00718-y PMID: 32678037

23. Lee SY, Lei B, Mallick B. Estimation of COVID-19 spread curves integrating global data and borrowing information. PLOS ONE. 2020; 15(7):e0236860. https://doi.org/10.1371/journal.pone.0236860

24. Tadić B, Melnik R. Modeling latent infection transmissions through biosocial stochastic dynamics. PLOS ONE. 2020; 15(10):e0241163. https://doi.org/10.1371/journal.pone.0241163

25. Cuevas E. An agent-based model to evaluate the COVID-19 transmission risks in facilities. Computers in Biology and Medicine. 2020; 121:103827. https://doi.org/10.1016/j.compbiomed.2020.103827

26. Chang SL, Harding N, Zachreson C, Cliff OM, Prokopenko M. Modelling transmission and control of the COVID-19 pandemic in Australia. Nature Communications. 2020; 11(1):1–13. https://doi.org/10.1038/s41467-020-19393-6

27. Burda Z. Modelling Excess Mortality in Covid-19-Like Epidemics. Entropy. 2020; 22:1236. https://doi.org/10.3390/e22111236

28. Hawkes AG. Spectra of some self-exciting and mutually exciting point processes. Biometrika. 1971; 58 (1):83–90. https://doi.org/10.1093/biomet/58.1.83

29. Reynaud-Bouret P, Rivoirard V, Tuleau-Malot C. Inference of functional connectivity in neurosciences via Hawkes processes. In: 2013 IEEE Global Conference on Signal and Information Processing. Austin, TX: IEEE; 2013. p. 317–320.

30. Chornoboy ES, Schramm LP, Karr AF. Maximum likelihood identification of neural point process systems. Biological Cybernetics. 1988; 59(4-5):265–275. https://doi.org/10.1007/BF00332915

31. Apostolopoulou I, Linderman SW, Miller K, Dubrawski A. Multivariate Mutually Regressive Point Processes. In: Advances in Neural Information Processing Systems; 2018. p. 5115–5126.

32. Mohler G. Modeling and estimation of multi-source clustering in crime and security data. Annals of Applied Statistics. 2013; 7(3):1525–1539. https://doi.org/10.1214/13-AOAS647

33. White G, Porter MD, Mazerolle L. Terrorism Risk, Resilience and Volatility: A Comparison of Terrorism Patterns in Three Southeast Asian Countries. Journal of Quantitative Criminology. 2012; 29(2):295–320. https://doi.org/10.1007/s10940-012-9181-y

34. Reinhart A, Greenhouse J. Self-exciting point processes with spatial covariates: modelling the dynamics of crime. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2018; 67(5):1305–1329.

35. Ogata Y. Statistical models for earthquake occurrences and residual analysis for point processes. Journal of Computational and Graphical Statistics. 1988; 83(401):9–27.

36. Chen F, Tan WH. Marked self-exciting point process modelling of information diffusion on Twitter. Annals of Applied Statistics. 2018; 12:2175–2196.

37. Park J, Chaffee AW, Harrigan RJ, Schoenberg FP. A non-parametric hawkes model of the spread of ebola in west africa. Journal of Applied Statistics, Forthcoming. 2018;.

38. Kelly JD, Park J, Harrigan RJ, Hoff NA, Lee SD, Wannier R, et al. Real-time predictions of the 2018–2019 Ebola virus disease outbreak in the Democratic Republic of the Congo using Hawkes point process models. Epidemics. 2019; 28:100354. https://doi.org/10.1016/j.epidem.2019.100354 PMID: 31395373

39. Kim M, Paini D, Jurdak R. Modeling stochastic processes in disease spread across a heterogeneous social system. Proceedings of the National Academy of Sciences. 2019; 116(2):401–406. https://doi.org/10.1073/pnas.1801429116

40. Schoenberg FP, Hoffmann M, Harrigan RJ. A recursive point process model for infectious diseases. Annals of the Institute of Statistical Mathematics. 2019; 71(5):1271–1287. https://doi.org/10.1007/s10463-018-0690-9

41. Meyer S, Elias J, Höhle M. A Space-Time Conditional Intensity Model for Invasive Meningococcal Disease Occurrence. Biometrics. 2011; 68(2):607–616. https://doi.org/10.1111/j.1541-0420.2011.01684.x

42. Linderman SW, Adams RP. Scalable Bayesian Inference for Excitatory Point Process Networks. arXiv. 2015;.

43. Rizoiu MA, Mishra S, Kong Q, Carman M, Xie L. SIR-Hawkes: Linking Epidemic Models and Hawkes Processes to Model Diffusions in Finite Populations. In: Proceedings of the 2018 World Wide Web Conference. Lyon, France: International World Wide Web Conferences Steering Committee; 2018. p. 419–428.

44. Bertozzi AL, Franco E, Mohler G, Short MB, Sledge D. The challenges of modeling and forecasting the spread of COVID-19. Proceedings of the National Academy of Sciences of the United States of America. 2020; 117(29):16732–16738. https://doi.org/10.1073/pnas.2006520117

45. Mohler G, Short MB, Schoenberg F, Sledge D. Analyzing the impacts of public policy on COVID-19 transmission in Indiana: The role of model and dataset selection. 2020;.

46. Chiang WH, Liu X, Mohler G. Hawkes process modeling of COVID-19 with mobility leading indicators and spatial covariates. medRxiv. 2020;.

47. Lesage L. A Hawkes process to make aware people of the severity of COVID-19 outbreak: application to cases in France. Université de Lorraine; University of Luxembourg.; 2020.

48. Chen Z, Dassios A, Kuan V, Lim JW, Qu Y, Surya B, et al. A Two-Phase Dynamic Contagion Model for COVID-19. arXiv. 2020;.

49. Koyama S, Horie T, Shinomoto S. Estimating the time-varying reproduction number of COVID-19 with a state-space method. PLOS Computational Biology. 2021; 17(1):e1008679. https://doi.org/10.1371/journal.pcbi.1008679

50. Dehning J, Zierenberg J, Spitzner FP, Wibral M, Neto JP, Wilczek M, et al. Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions. Science. 2020; 369 (6500). https://doi.org/10.1126/science.abb9789 PMID: 32414780

51. Mbuvha R, Marwala T. Bayesian inference of COVID-19 spreading rates in South Africa. PLOS ONE. 2020; 15(8):e0237126. https://doi.org/10.1371/journal.pone.0237126

52. Piccolomini EL, Zama F. Monitoring Italian COVID-19 spread by a forced SEIRD model. PLOS ONE. 2020; 15(8):e0237417. https://doi.org/10.1371/journal.pone.0237417

53. Sharma VK, Nigam U. Modeling and Forecasting of Covid-19 growth curve in India. medRxiv. 2020;.

54. Paiva HM, Afonso RJM, de Oliveira IL, Garcia GF. A data-driven model to describe and forecast the dynamics of COVID-19 transmission. PLOS ONE. 2020; 15(7):e0236386. https://doi.org/10.1371/journal.pone.0236386

55. Romero-Severson EO, Hengartner N, Meadors G, Ke R. Change in global transmission rates of COVID-19 through May 6 2020. PLOS ONE. 2020; 15(8):e0236776. https://doi.org/10.1371/journal.pone.0236776

56. Detommaso G, Hoitzing H, Cui T, Alamir A. Stein Variational Online Changepoint Detection with Applications to Hawkes Processes and Neural Networks. arXiv. 2019;.

57. Horwitz LI, Jones SA, Cerfolio RJ, Francois F, Greco J, Rudy B, et al. Trends in COVID-19 Risk-Adjusted Mortality Rates. Journal of Hospital Medicine. 2020; 16(2):90–92. https://doi.org/10.12788/jhm.3552

58. Dennis JM, McGovern AP, Vollmer SJ, Mateen BA. Improving Survival of Critical Care Patients With Coronavirus Disease 2019 in England: A National Cohort Study, March to June 2020*. Critical Care Medicine. 2021; 49(2):209–214. https://doi.org/10.1097/CCM.0000000000004747

59. Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. COVID-19 data repository; 2020. https://github.com/CSSEGISandData/COVID-19.

60. Bürkner PC, Gabry J, Vehtari A. Approximate leave-future-out cross-validation for Bayesian time series models. Journal of Statistical Computation and Simulation. 2020; 90(14):2499–2523. https://doi.org/10.1080/00949655.2020.1783262

61. Roberts GO, Tweedie RL. Exponential convergence of Langevin distributions and their discrete approximations. Bernoulli. 1996; 2(4):341–363. https://doi.org/10.2307/3318418

62. Worden L, Wannier R, Hoff NA, Musene K, Selo B, Mossoko M, et al. Projections of epidemic transmission and estimation of vaccination impact during an ongoing Ebola virus disease outbreak in Northeastern Democratic Republic of Congo, as of Feb. 25, 2019. PLOS Neglected Tropical Diseases. 2019; 13 (8):e0007512. https://doi.org/10.1371/journal.pntd.0007512 PMID: 31381606

63. Funk S, Camacho A, Kucharski AJ, Lowe R, Eggo RM, Edmunds WJ. Assessing the performance of real-time epidemic forecasts: A case study of Ebola in the Western Area region of Sierra Leone, 2014-15. PLOS Computational Biology. 2019; 15(2):e1006785. https://doi.org/10.1371/journal.pcbi.1006785

64. Peixoto PS, Marcondes D, Peixoto C, Oliva SM. Modeling future spread of infections via mobile geolocation data and population dynamics. An application to COVID-19 in Brazil. PLOS ONE. 2020; 15(7): e0235732. https://doi.org/10.1371/journal.pone.0235732

65. Killick R, Eckley I. changepoint:an R package for changepoint analysis. Journal of Statistical Software. 2014; 58(3). https://doi.org/10.18637/jss.v058.i03

66. DiMatteo I, Genovese CR, Kass RE. Bayesian curve-fitting with free-knot splines. Biometrika. 2001; 88 (4):1055–1071. https://doi.org/10.1093/biomet/88.4.1055

67. The Royal Society. Reproduction number (R) and growth rate (r) of the COVID-19 epidemic in the UK: methods of estimation, data sources, causes of heterogeneity, and use as a guide in policy formulation; 2020.

68. Li S, Xie Y, Farajtabar M, Verma A, Song L. Detecting Changes in Dynamic Events Over Networks. IEEE Transactions on Signal and Information Processing over Networks. 2017; 3(2):346–359. https://doi.org/10.1109/TSIPN.2017.2696264

B | Counterfactual explanation for anomaly detection in time series

This chapter corresponds to the following article:

> Sulem, D., Donini, M., Zafar, M. B., Aubet, F.-X, Gasthaus, J., Januschowski, T., Das, S., Kenthapandi, K. & Archambeau, C. (2022). Diverse counterfactual explanations for anomaly detection in time series. Submitted to TMLR.

Diverse Counterfactual Explanations for Anomaly Detection in Time Series

Anonymous authors
Paper under double-blind review

Abstract

Data-driven algorithms for detecting anomalies in times series data are ubiquitous, but generally unable to provide helpful explanations for the predictions they make. In this work we propose a post-hoc explainability method that is applicable to any differentiable anomaly detection algorithm for time series. Our method provides explanations in the form of a set of diverse counterfactual examples, i.e., multiple perturbed versions of the original time series that are similar to the latter but not considered anomalous by the detection algorithm. Those examples are informative on the important features of the time series and the magnitude of changes that can be made to render it non-anomalous for the explained algorithm. We call our method *counterfactual ensemble explanation*, and test it on two deep-learning-based anomaly detection models. We apply the latter to univariate and multivariate real-world data sets and assess the quality of our explanations under several explainability criteria such as *Validity, Plausibility, Closeness* and *Diversity*. We show that our algorithm can produce valuable explanations; moreover, we propose a novel visualization of our explanations that can convey a richer interpretation of a detection algorithm's internal mechanism than existing post-hoc explainability methods. Additionally, we design a sparse variant of our method to improve the interpretability of our explanation for high-dimensional time series anomalies. In this setting, our explanation is localized on only a few dimensions and can therefore be communicated more efficiently to the model's user.

1 Introduction

Anomaly detection in time series is a common data analysis task that can be defined as identifying outliers, i.e., observations that do not belong to a reference distribution. For instance, anomaly detection is leveraged to localize a defect in computing systems, disclose a fraud in financial transactions, or diagnose a disease from health records Blázquez-García et al. (2021). Detected outliers often call for further investigation, therefore, the recipient of a detection algorithm outputs generally needs to be able to interpret the algorithm's predictions. Consequently, providing explanations for models that detect anomalies has practical relevance, all the more in the setting of multivariate time series data, where model interpretation is an even more challenging task. This is however a still understudied problem, in particular for machine learning models.

In general, an anomaly detection model classifies each timestamp of a time series as anomalous or not. Several state-of-the-art models involve complex deep learning (DL) classifiers, such as LSTMs Malhotra et al. (2015), RNNs Audibert et al. (2020) or TCNs Bai et al. (2018); Carmona et al. (2021), whose internal mechanisms are opaque. This lack of transparency can prevent these models from being deployed in consequential contexts Brown et al. (2018); Bhatt et al. (2020). Prior work has proposed to include interpretable blocks in machine learning models for anomaly detection (e.g., attention mechanism in RNNs Brown et al. (2018)) or design model-specific explainability methods (e.g., feature-importance scores for Isolation Forests Carletti et al. (2021)). Our work is orthogonal to these methods: we propose a post-hoc and model-agnostic explainability method that can be applied to any existing differentiable anomaly detection model.

The majority of existing post-hoc explainability methods for time series models aims at estimating feature-saliency scores Crabbe & van der Schaar (2021); Pan et al. (2020). The latter ranks the features of the

input data in terms of their relative contribution to the model's prediction. Although these techniques have provided valuable insights in image classification tasks (Fong et al., 2019), it is often a weak form of explanation for anomalies in time series. In fact, they essentially indicate that the time series values at the anomalous time stamps are salient, therefore providing redundant information compared to the anomaly detection model (see for instance Figure 1b, where the salient features are highlighted in green). In practice, a user of an anomaly detection model might be interested in (a) knowing what can be changed in the input data to avoid encountering the anomaly again in the future (preferentially with minimal cost), and (b) understand the model's sensitivity to a particular anomaly. Our proposed method provides explanations satisfying these two requirements.

Counterfactual explanation denotes a type of explainability method that provides insight on the sensitivity of a model's predictions to a change in the input data. They have notably been proposed for interpreting time series classifiers Ates et al. (2021); Delaney et al. (2021); Karlsson et al. (2020). A counterfactual example (or for short, a counterfactual) is an instance-based explanation in the form of a perturbed input on which the model's prediction value is different from the model output on the original data. It thus indicates what modifications of the input must be made to obtain a different prediction. It is generally defined as an instance X' minimizing a cost function such as Wachter et al. (2018):

$$L(X, X', y', \lambda) = \lambda(f(X') - y')^2 + d(X, X'),$$

where X and f are respectively the original input and the prediction model that need to be explained, y' is a desired output value (e.g. a different predicted label in classification contexts), $d(.,.)$ is a distance on the input space and λ is a trade-off parameter. In their basic definition, they are closely related to adversarial examples Verma et al. (2020), however, their properties and their utility are distinct. Adversarial examples are often weakly constrained and used as hard instances to train more robust models, whereas counterfactuals are designed as plausible examples for interpreting an existing model's predictions.

In the context of anomalies detected by a time series model, counterfactual methods aim at generating modified time series (or sub-sequences) that do not contain anomalous observations according to the detection model. With the additional constraint that counterfactuals are somehow similar to the original time series, these time series instances therefore correspond to the closest normal or expected behaviour according to the explained model. For example, when the time series is a temporal record of a patient's blood glucose level with abnormally high values, a counterfactual example can be an alternative record with levels in a non-critical interval. Hence, counterfactual explanations can reveal the boundaries of the normal time series distribution according to the prediction model.

However, a single counterfactual is generally only a partial explanation, satisfying a particular trade-off between predefined criteria Russell (2019). One extension of counterfactual explanation consists in providing an *ensemble* (or set) of *diverse* counterfactual instances Russell (2019); Mothilal et al. (2020); Dandl et al. (2020). Nonetheless, this extension has not been previously considered in the context of time series anomaly detection models. Besides, more broadly, there is no existing strategy to effectively communicate these more complex counterfactual explanations to the model's user. In this work, we propose an approach for generating counterfactual ensemble explanations for anomaly detection models in time series, as well as a visualization method of these explanations.

More precisely, we make the following contributions:

- We introduce a model-agnostic and post-hoc method that explains the predictions of any differentiable anomaly detection model for time series. For any given input and prediction value, our explanation, called *counterfactual ensemble explanation*, is a set of counterfactual examples satisfying different trade-offs between pre-defined criteria. In practice, these examples can be used individually as actionable explanations, or analysed together to investigate the model's sensitivity to perturbations of the input.
- We design a *sparse* variant of our method for high-dimensional time series anomalies, which have been much less studied and generally harder to interpret. In this context, we constraint our counterfactual explanation to make changes only on a few dimensions of the input time series, so that it can be communicated more efficiently to the explanation's recipient.

- We propose an interpretable visualization of our counterfactual ensemble explanation. Our representation shows the range of possible perturbations gaining insight on the model's local decision boundary and sensitivity. Thus, our visualization can increase the actionability of the counterfactual explanation, when the time series features are mutable.
- We investigate the value of our method on two deep-learning anomaly detection models, applied to univariate and multivariate real-world time series data sets. We quantify the quality of our ensemble explanations using metrics previously proposed in other data domains, namely *Validity, Plausibility, Closeness* and *Diversity*. We note that ensemble explanations have never been considered in the context of time series anomalies, therefore there is not yet an equivalent competitive method. However, we also design a *naive* counterfactual ensemble method that we numerically compare to in our experiments.

Figure 1 illustrates our proposed method, and its novelty in contrast to existing explainability methods for time series models. In this univariate example with a spike outlier, a feature-saliency explanation method essentially highlights the time series features near the anomaly (Figure 1b). Besides, a (single) counterfactual explanation proposes a whole new subsequence where the largest feature changes are localized at the anomaly (Figure 1c). In comparison, our ensemble explanation (Figure 1d) is (a) *sparse*, in the sense that it is localized on a few time series features (the anomaly) (b) *optimal* in that it minimally modifies these features and (c) *rich* by diversifying the possible perturbations (see Figure 1e, showing a few examples from our ensemble).

After succinctly reviewing existing work in explainability for time series models and counterfactual explanations in Section 2, we describe the general set-up in Section 3. In Section 4, we present our approach. Then in Section 5, we demonstrate the effectiveness of our method on DL-based models and benchmark anomaly detection data sets. Finally, we discuss our results and propose possible future developments in Section 6.

2 Related work

Explainability methods for users of machine learning models have developed along two paradigms: building models with interpretable blocks or designing model-agnostic methods that can be applied to any model already deployed. For time series data, RETAIN Choi et al. (2016) incorporates an attention-mechanism in an RNN-based model while Dynamic Masks Crabbe & van der Schaar (2021) is a model-agnostic algorithm that produces sparse feature-importance masks on time series using dynamic perturbation operators. In fact, many methods for time series adapt algorithms designed for tabular or image data: for instance, TimeSHAP Bento et al. (2021) extends SHAP, a feature-attribution method that approximates the local behaviour of a model with a linear model using a subset of features. Another interesting line of work interprets CNNs for time series models using Shapelet Learning Ma et al. (2020). Shapelets are subsequences that are learnt from a dataset to build interpretable time series decompositions.

Nonetheless, previously cited work for time series are feature-saliency estimation methods. Although they are notably helpful to localize the important parts of time series (in terms of their contribution to the model's prediction), they can only weakly explain anomaly detection models. Moreover, *instance- or example-based* explanations can be more easily interpreted by a non-expert person Wachter et al. (2018). These methods explain a prediction on a single instance by comparing it to another real or generated example, e.g., the most typical examplar of the observed phenomenon (a *prototype* Hautamaki et al. (2008)) or a contrastive examplar related to a distinct behaviour (a *counterfactual* Ates et al. (2021); Delaney et al. (2021); Karlsson et al. (2020)). For time series classifiers, counterfactuals can be generated by swapping the values of the most discriminative dimensions with those from another training instance Ates et al. (2021). In a causal inference setting, Chernozhukov et al. (2021) construct counterfactual time series as linear combinations of control groups. Unfortunately, these approaches can yield implausible subsequences, that do not belong to the data manifold Carletti et al. (2021), e.g., by breaking correlations between the dimensions of multivariate time series. The Native Guide algorithm Delaney et al. (2021) does not suffer from the previous issue but uses a perturbation mechanism on the Nearest Unlike Neighbor in the training set using the model's internal feature vector. Lastly, for a k-NN and a Random Shapelet Forest classifiers, Karlsson et al. (2020) design a tweaking mechanism to produce counterfactual time series.

However, these methods necessitate knowledge of the model's internal mechanism and/or access to its training dataset, which can be expensive. Additionally, these counterfactual explanations suffer from the so-called *Rashomon effect* Molnar (2019), i.e., the fact that several equally-good perturbed examples might exist and be informative for the model's user. In this case, one might benefit from knowing multiple ones, before choosing the most helpful example in a specific context Mothilal et al. (2020). For linear classifiers of tabular data, a set of diverse counterfactuals can be obtained by sequentially adding constraints along the optimization iterations of the perturbation algorithm Russell (2019), whereas the Multi-Objective Counterfactuals algorithm Dandl et al. (2020) records multiple perturbed examples generated along the iterations of a genetic algorithm. These counterfactual sets therefore contain different trade-offs between conflicting criteria. While in the previous methods, diversity is not explicitly enforced, the DiCE algorithm Mothilal et al. (2020) includes a penalization on counterfactuals' similarity based on Determinantal Point Processes. In a similar fashion, for image classifiers, DiVE Rodriguez et al. (2021) perturbs the latent features in a Variational Auto Encoder and penalises pairwise similarity between perturbations, while Karimi et al. (2020) propose a general framework for generating counterfactual examples with diversity constraints in heterogeneous data. Our paper differs from these works since it considers the problem of generating diverse counterfactual explanations for the time series domain. In particular, we leverage specific time series perturbation mechanisms in order to obtain plausible examples.

To the best of our knowledge, we propose the first method that provides diverse counterfactual explanations for time series. As previously noted Crabbe & van der Schaar (2021), this data domain requires specific treatment of temporal dependencies, therefore existing methods for tabular data cannot be directly applied. Besides, having a diverse set of counterfactual explanations can be particularly helpful for time series where the actionable or mutable features are not known in advance. We introduce our method in the context of anomaly detection, however we believe that our approach could be adapted to other tasks on time series data. Moreover, previous works proposing diverse counterfactual explanations have not discussed the additional challenge of communicating efficiently a set of examples compared to a single one. The visual representation we propose can be related to the "What-If Tool" Wexler et al. (2020), an interactive visual tool designed for general ML model elicitation. Before exposing our method, we describe the general set-up in the next section.

3 General set-up

In this work, we assume that anomalies in a time series are unpredictable and out-of-distribution subsequences. Hence, an anomaly is a significant deviation from a given reference behaviour. In the remainder, we will not make a distinction between anomaly, outlier and anomalous/abnormal/atypical observation. Not-anomalous data points will be considered as belonging to the data distribution, and denoted as the reference/normal/typical/expected behaviour. We will also refer to the latter as the *context*.

For the description of the general set-up, we introduce the following notations: for an integer $k \in \mathbb{N}$, $[k]$ denotes the set $\{i;\ 1 \leq i \leq k\}$ and for $x \in \mathbb{R}$, let $x_+ = \max(0, x)$. For a vector $v \in \mathbb{R}^n$, we denote v_i its i-th coordinate and for $X \in \mathbb{R}^{m \times n}$ a matrix or multivariate time series, X_i denotes respectively the i-th row or the i-th observation.

3.1 Anomaly detection model

We assume that we are given an anomaly detection model which we can use to predict anomalies on a time series of any given length. We consider a general setting where time series are multivariate and the model processes all dimensions (or *channels*) jointly. More precisely, we denote $X \in \mathbb{R}^{T \times D}$ a time series with T time stamps and D dimensions. The prediction function of the model, denoted by f, is used to classify each timestamp $t \in [1, T]$ of X as "anomalous" (i.e., label 1) or "not-anomalous" (i.e., label 0). In fact, the prediction $f(X) \in \mathbb{R}^T$ is a vector of anomaly scores for each timestamp (e.g., probability scores of being anomalous) which transforms into a vector of 0-1 labels using the model's classification rule (e.g. a threshold on these scores). Note that the dimension of the vector $f(X)$ might be smaller than T if the model needs a warm-up interval.

In practice, these models often detect anomalous time stamps by subdividing time series into smaller time windows and classifying the latter (therefore each timestamp or a subset of them in these sub-windows). In other works, to output a prediction on a single timestamp, the "receptive field" of a model is generally a fixed-size (typically small) window. Let's denote $W \in \mathbb{R}^{L \times D}$ a window of size L and consider the following general set-up: the window $W = [W_C, W_S]$ is subdivided by the model into two parts, with $W_C \in \mathbb{R}^{(L-S) \times D}$ a *context* part (that can be empty if the context is implicit once the model is trained) and $W_S \in \mathbb{R}^{S \times D}$ a *suspect* part, for which the model makes a prediction. More precisely, $f(W) \in \mathbb{R}^S$ is the anomaly score of the window W_S and, without loss of generality, we suppose that $f(W) \in [0,1]^S$. We also denote $\theta \in [0,1]$ the anomaly detection rule, i.e., a label 1 is given to W_S if for some $i \in [S]$, $f(W)_i > \theta$.

Examples of anomaly detection models with the previously described mechanism are NCAD Carmona et al. (2021), where the context window has typically thousands of time stamps and the suspect window has 1 to 5 time stamps, and USAD Audibert et al. (2020), where $W = W_S$ and $L = 5$ or 10. In the latter case, the context is implicit and the whole training set is considered as normal data and thus the context of anomalies detected in a test time series.

3.2 Counterfactual explanation

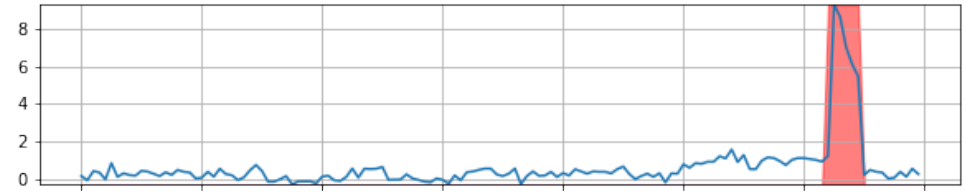
In most cases, a single anomaly is a short subsequence, and can therefore be contained in one or few contiguous subwindows W_S. For ease of exposition, we suppose that an anomaly is contained in one suspect window. An example is shown in Figure 1a where a suspect window W_S (highlighted in red) contains an anomaly. A counterfactual example for model f detecting an anomaly in W_S (i.e., for some $i \in [S]$, $f(W)_i > \theta$), is an alternative window $\widetilde{W} = [W_C, \widetilde{W}_S]$ such that all predicted labels are 0 (i.e., for any $i \in [S]$, $f(\widetilde{W})_i < \theta$). Since the context of the anomaly is also key to its detection by the model, and if W does not contain a context window W_C, we choose to add in the counterfactual example $\widetilde{W}$ a fixed size window W_C, that immediately precedes W in the time series. Note that we implicitly suppose that anomalies are not too close to each other so that the additional context window does not contain any anomaly. With a slight abuse of notations, we still denote $\widetilde{W}$ the obtained counterfactual example.

3.3 Properties of counterfactual explanations

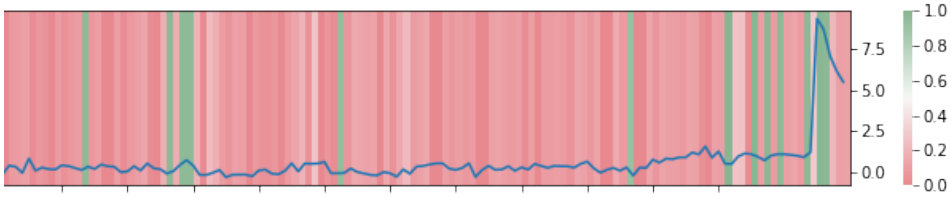
There are four largely consensual properties that convey value and utility to counterfactual explanations in the context of model elicitation Verma et al. (2020):

1. *Validity* or *Correctness*: achieving a desired model output, e.g., changing the predicted class label in classification; this is the key goal of a contrastive explanation.
2. *Parsimony* or *Closeness*: minimally and sparsely changing the original input; this is motivated by practical feasibility of the counterfactual if the input features are actionable, and by readibility of the information communicated to the model's user.
3. *Plausibility*: counterfactual explanations need to contain realistic examples of normal subsequences.
4. *Computational efficiency:* being computable within a reasonable amount of time and with acceptable computing resources.

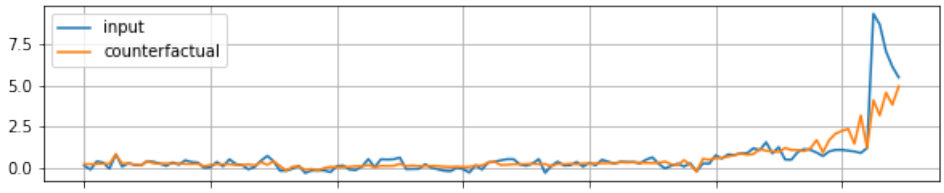
In the context of an anomaly detected in a time series, property (1) is equivalent to flipping the anomaly detection model's prediction label from 1 to 0 (i.e., achieving a anomaly prediction score below the classifier threshold). Property (2) can be enforced by restricting the perturbation of the input on a small window containing the anomaly (i.e., the suspect window W_S) and on few dimensions of the time series (if the anomalous features are only located on some channels). Property (3) requires that the counterfactual belongs to the normal data distribution. If the latter is not known or estimated, this criterion can be complicated to evaluate, but some prior knowledge such as the time series' regularity, seasonality, or bounds can be leveraged. Property (4) potentially depends on the specific setting, in particular the cost of using the model's prediction function or its gradient, and the size of the dataset. However, in our context, we assume that accessing the



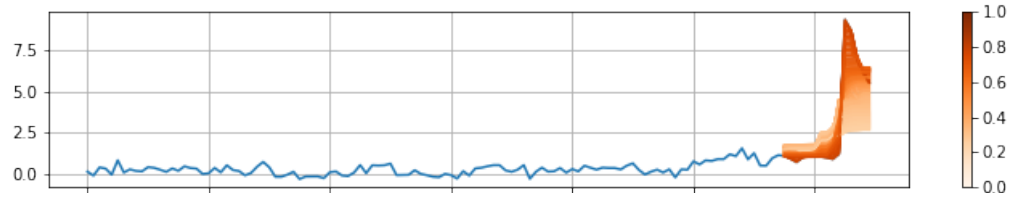
(a) Original time series window



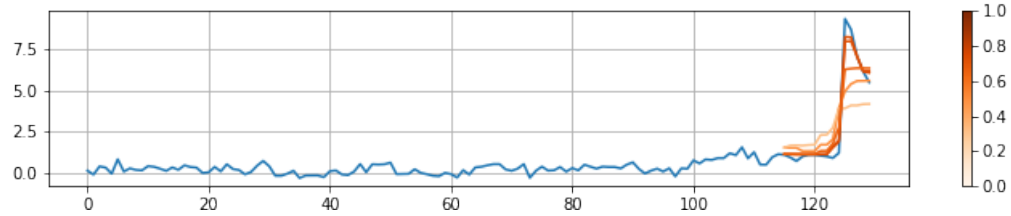
(b) Feature-saliency map



(c) Counterfactual example



(d) Our counterfactual ensemble explanation



(e) Counterfactual examples from our ensemble explanation

Figure 1: Comparison between existing explainability methods for time series and ours, in the context of anomaly detection. The original input (1a) is a univariate time series window containing an anomalous subsequence (a spike outlier, highlighted in red) and the anomaly detection model is NCAD (see Section 5.1.2). The subsequent panels represent the explanations from a feature-importance method (Dynamic Masks Crabbe & van der Schaar (2021)) (1b), an instance-based method (counterfactual example) (1c) and our method (1d). In (1b), the important timestamps have saliency scores closed to one (green color code). In (1d), all the examples from our counterfactual ensemble, which only span the anomalous sub-window, are plotted; the orange color map indicates their anomaly scores (between 0 and 1) given by the explained model. In (1e), we additionally plot five counterfactual examples from this ensemble.

training set of the detection model is particularly expensive, since the latter often decomposes the time series into small windows, leading to a large number of actual training inputs for long time series.

Unfortunately, those properties are often conflicting (e.g., parsimony and plausibility in the context of a spike outlier), therefore a single counterfactual example can only achieve a particular trade-off between them. In the next paragraph, we motivate the use of counterfactual *ensembles* (or sets) as more comprehensive explanations.

3.4 Diversity as an additional property

When the data features are actionable , the counterfactual example informs on the localization and magnitude of change that can be applied to the original time series to obtain non-anomalous data. However, the best or feasible trade-off between the pre-defined criteria might depend on the particular anomaly or user's range of action. In absence of this prior knowledge, previous work Mothilal et al. (2020); Russell (2019) added diversity, or range of perturbation, as one informative criterion. In particular, a set of counterfactual examples can increase the likelihood of finding a helpful explanation Rodriguez et al. (2021).

In this sense, an ensemble of counterfactual explanations for an opaque model is more insightful if the user can discriminate between feasible and non-feasible counterfactual examples when given a set of them. However, this qualitative statement is difficult to quantify in practice since there for most data sets, there is no ground-truth for the notion of actionability of counterfactual explanations.

Moreover, we also argue that this additional complexity in the explanation should be adequately communicated to the explanation's recipient, e.g., with a suitable visualization. Intuitively, the latter should be informative on the different possibilities of features changes and the particular trade-off achieved by a counterfactual example. In Section 5.4, we propose a representation for time series, where all counterfactual examples can be visualized together with their anomaly score under the explained model. One example is shown in Figure (1d) and several case studies are represented in Figure 2.

4 Methodology

In this section, we present our method for generating counterfactual ensemble explanations. Our approach for differentiable anomaly detection models is described in § 4.1 and can be delineated into two variants, whose respective uses depend on prior knowledge of the data distribution. The first one, called *Interpretable Counterfactual Ensembles (ICEs)* (§ 4.1.1), can be applied without any domain knowledge input. The second one, called *Dynamically Perturbed Ensembles (DPEs)* (§ 4.1.2), leverages dynamic perturbation operators (Crabbe & van der Schaar, 2021), which induce a modification of a time series according to a pre-defined mechanism. Next, we design *sparse* variants of this approach, where the perturbations are restricted to a few dimensions of the input (high-dimensional) time series (§ 4.2). Finally, we describe an alternative method for generating our counterfactual ensemble when the model's gradient information is not available (§ 4.3).

4.1 Gradient-based counterfactual ensemble explanations

Most counterfactual algorithms (e.g., Native Guide Delaney et al. (2021), Growing Spheres Laugel et al. (2018), DiCE Mothilal et al. (2020)) rely on adequately perturbing the input W and optimise the perturbation to enforce some properties of the perturbed example. In our method, using the notations of Section 3, we first define an objective function over a single counterfactual example $\widetilde{W} = [W_C, \widetilde{W_S}]$, then use a gradient-descent algorithm starting at the original time series to minimize it. The ensemble of examples is built along the optimization path by collecting adequate perturbations. We define two variants of our method: one, called *Interpretable Counterfactual Ensemble* (ICE), that is a completely unspecified, and another one, *Dynamically Perturbed Ensemble* (DPE), where one can input some domain knowledge and specify a dynamic perturbation mechanism Crabbe & van der Schaar (2021).

4.1.1 Interpretable Counterfactual Ensemble (ICE)

In this variant, the objective function on a counterfactual example is defined as follows:

$$\mathcal{L}_{ICE}(\widetilde{W}) = \mathcal{L}_{pred}(\widetilde{W}) + \mathcal{L}_c(\widetilde{W}) + \mathcal{L}_s(\widetilde{W}), \tag{1}$$

where the first term accounts for the Validity property via a hinge loss on the prediction score on $\widetilde{W}$, i.e.,

$$\mathcal{L}_{pred}(\widetilde{W}) = (f(\widetilde{W}) - c)_+,$$

with $c \in [0, 1]$ is a margin parameter. The second term in equation 1 enforces the Closeness constraint via a penalty similar to the elastic net Zou & Hastie (2005), here using the Frobenius and the L_1 matrix distances:

$$\mathcal{L}_c(\widetilde{W}) = \frac{\lambda_1}{S\sqrt{D}}\|\widetilde{W} - W\|_1 + \frac{\lambda_2}{SD}\|\widetilde{W} - W\|_F,$$

where $\lambda_1, \lambda_2 > 0$ are regularization parameters. Finally, the third term of equation 1 enforces Plausibility through temporal smoothness (see for instance Crabbe & van der Schaar (2021)):

$$\mathcal{L}_s(\widetilde{W}) = \frac{\lambda_T}{(S-1)D}\sum_{i=1}^{D}\sum_{t=1}^{S-1}|[\widetilde{W}_S]_{(t+1)i} - [\widetilde{W}_S]_{ti}|,$$

with $\lambda_T > 0$. The assumption behind this constraint is that normal time series are not too rough and smoother than abnormal windows, therefore realistic perturbations should also be quite smooth.

4.1.2 Dynamically Perturbed Ensemble (DPE)

In this variant, one can specify the perturbation mechanism to obtain the counterfactual ensemble using a dynamic perturbation operator Crabbe & van der Schaar (2021) and a map that spatially and temporally modulates this perturbation. This notably allows to specify the lengthscale of change in the perturbation operator. More precisely, a map is a matrix $M \in [0, 1]^{S\times D}$ that accounts for the amount of change applied to a timestamp and a dimension in the suspect window W_S. A value close to 1 in M indicates a big change while a value close to 0 indicates a small change. Here, the dynamic perturbation operator is a Gaussian blur which takes as input a time series window W, a timestamp $t \in [L - S, L]$, a dimension $i \in [D]$ and a weight $m \in [0, 1]$, and is defined as:

$$\pi_G(W, t, i, m) = \frac{\sum_{t'=1}^{L} W_{t'i}\exp(-(t-t')^2/2(\sigma_{max}(1-m))^2)}{\sum_{t'=1}^{L}\exp(-(t-t')^2/2(\sigma_{max}(1-m))^2)},$$

with $\sigma_{max} \geq 0$, a hyperparameter tuning the blur's temporal bandwidth. We note that the bigger this parameter is, the larger is the smoothing effect of the perturbation. The latter is called *dynamic* in the sense that it modifies a timestamp using its neighbouring times. We also refer to Crabbe & van der Schaar (2021) for more examples of dynamic perturbation operators.

Finally, for a given map M, a perturbed suspect window is given by $[\widetilde{W}_S(M)]_{ti} = \pi(W, L - S + t, i, 1 - M_{ti})$, $t \in [S], i \in [D]$. The objective function is then written in terms of the perturbation map as:

$$\begin{aligned}\mathcal{L}_{DPE}(M) = \mathcal{L}_{pred}(\widetilde{W}(M)) + \frac{\lambda_1}{S\sqrt{D}}\|M\|_1 + \frac{\lambda_2}{SD}\|W - \widetilde{W}(M)\|_F \\ + \frac{\lambda_T}{(S-1)D}\sum_{i=1}^{D}\sum_{t=1}^{S-1}|M_{(t+1)i} - M_{ti}|,\end{aligned} \tag{2}$$

where the first term is the hinge loss, and the second and fourth terms account for the sparsity and smoothness constraints, in this case applied on M rather than $\widetilde{W}$ as in equation 1.

Algorithm 1 Gradient-based counterfactual ensemble explanation algorithm.

Input: The anomalous time series window W, the anomaly detection model f, the anomaly threshold θ, the learning rate η , the number of iterations T , the number of counterfactual examples N.
$\tilde{W}^0 = W$
$I = \{\}$
for $t = 1, \ldots, T$ **do**
  Do one step of Stochastic Gradient Descent $\tilde{W}^t = \tilde{W}^{t-1} - \eta \nabla \mathcal{L}(\tilde{W}^{t-1})$
  if $\forall i \in [S],\ f(\tilde{W}^t)_i < \theta$ **then**
    Add $\tilde{W}^t$ to I
  end if
end for
$J = |I|/N$.
Subsample every J-th elements of I.
Output: The set of N counterfactual examples I.

4.1.3 Optimization and complexity.

Our algorithm for differentiable models has the following steps (see also our pseudo-code in Algorithm 1). We first initialize the counterfactual $\tilde{W}$ at the original anomalous window W. Then, we minimize the objective function equation 1 or equation 2 using T iterations a Stochastic Gradient Descent (SGD) algorithm. At each iteration $t = 1, \ldots, T$, we evaluate the anomaly detection model at the current value $\widetilde{W}^t$ and if $\forall i \in [S], f(\widetilde{W}^t)_i < \theta$, we add $\widetilde{W}^t$ to a set I. After T iterations, we subsample N counterfactuals from the set I to obtain a diverse counterfactual ensemble. In practice, by choosing T around 1000, N around 20-30, and an adequate learning rate, the size of the set I will be much larger than N and for simplicity, we regularly subsample I, ordered by the iteration rank of the examples. Complexity-wise, our method therefore requires to query the anomaly detection model and its gradient at each iteration of the SGD algorithm.

We note that in our method, we do not select only the global optimum of our objective functions, but we collect a set of examples along the optimisation path, as long as these examples are non-anomalous. Our heuristic is that by initializing at the original time series, we hope to collect counterfactual examples that are close to the original time series, for a large range of hyperparameters values. Moreover, since defining an *optimal* counterfactual given the Closeness and Plausibility criteria for each anomaly is not easy to specify, the different examples found along the optimization path achieve different trade-offs between the terms in the objective function. In fact, these examples can be seen as solutions of optimization problems with different sets of weights (hyperparameters) in this objective. Note that similar strategies to ours have been previously used for generating ensemble of counterfactuals in distinct data domains, e.g., in Dandl et al. (2020); Russell (2019); Ley et al. (2022).

Other potential candidates for enforcing diversity. We now discuss other candidates from literature for enforcing diversity in counterfactual explanations. There are three notable alternative strategies: a) define an objective function over a set of counterfactual examples and include a proximity penalty between the examples, as in Mothilal et al. (2020); Ley et al. (2022); b) select the optima of our objective function for N sets of hyperparameters (e.g., chosen over a grid); and c) select the optima of our objective function for N random initialization points of our algorithm. For strategy a), solving such an objective is much more cumbersome for a large number of features and counterfactual examples. In fact, Ley et al. (2022) note that using a Determinantal Point Process penalty like in Mothilal et al. (2020) requires expensive computations of matrix determinants. Besides, using instead a penalty based on pairwise distances like in Bhatt et al. (2021) may be particularly challenging for time series where non-standard distances must be computed. As for strategy b), solving the optimization problem for T sets of hyperparameters would be much less computationally efficient, and in practice, T would need to be much larger than what we use in our method to obtain N valid counterfactuals since most of the hyperparameter configurations would fail. Finally, additionally to being less computationally efficient, we found that strategy c) is not enough to enforce diversity and often leads to redundant solutions in our experiments. This empirical observation has

been previously noted by Ley et al. (2022) in a different data context and may be due to the fact that a fixed set of hyperparameters induces a "strong" minimum of the objective function.

4.2 Sparse counterfactual explanations for high-dimensional time series

A high-dimensional time series would result in a similarly high dimensional explanations. On the other hand, prior work argues that humans prefer simpler explanations Miller (2019). Therefore, one may obtain a simpler explanation by restricting the counterfactual ensemble explanation to span as few dimensions as possible. In this case, the explanation can be more easily visualized and the counterfactual is more actionable, since it then requires to change a minimal number of channels. Besides, anomalies often tend to be concentrated on few dimensions, for instance, when a small subsample of monitoring metrics take abnormal values in a servers network Su et al. (2019b)). Therefore, explanations for these anomalies should also reflect their low-dimensional property. For these reasons, we design a sparse version of our gradient-based method that constraints the counterfactual ensemble explanation to be *spatially* sparse (i.e., sparse or parsimonious in the perturbed dimensions).

4.2.1 Sparse ICE

In the sparse version of ICE, we restrict the number of perturbed dimensions by introducing a vector $w \in [0,1]^D$ and a matrix $Z \in \mathbb{R}^{S\times D}$, and defining $\widetilde{W}_S(w,Z) = (w \otimes \mathbf{1}) \odot Z + ((1-w)\otimes \mathbf{1}) \odot W_S$. The role of w is to select the dimensions in W_S that are perturbed with Z. We then consider an objective function in terms of (Z, w):

$$\begin{aligned}\mathcal{L}_{ICE,SP}(w,Z) = (f(\widetilde{W}(w,Z)) - c)_+ \\ + \frac{\lambda_1}{\sqrt{D}}\|w\|_1 + \frac{\lambda_2}{SD}\|W - \widetilde{W}(w,Z)\|_F \\ + \frac{\lambda_T}{(S-1)D}\sum_{i=1}^{D}\sum_{u=1}^{S-1}|Z_{(u+1)i} - Z_{ui}|. \end{aligned} \tag{3}$$

Contrary to equation 1, where the sparsity penalization is applied globally (i.e., both temporally and spatially), the previous objective enforces spatial sparsity through the L_1-penalisation on w. Another way to see that is to re-interpret objective equation 1 as objective equation 3 with $w = (1, 1, \ldots, 1)$, $Z = \widetilde{W}_S$ and replace the L_1-penalisation on w by $\frac{\lambda_1}{S\sqrt{D}}\|Z - W_S\|_1$.

4.2.2 Sparse DPE

We apply the same idea to the DPE variant by enforcing the perturbation maps to be spatially sparse. More precisely, we define $M(w,t) = t \otimes w$ with $w \in [0,1]^D$ and $t \in [0,1]^T$ and a loss function in terms of (w,t):

$$\begin{aligned}\mathcal{L}_{DPE,SP}(w,t) = (f(\widetilde{W}(w,t)) - c)_+ \\ + \frac{\lambda_1}{\sqrt{D}}\|w\|_1 + \frac{\lambda_2}{SD}\|W - \widetilde{W}(w,t)\|_F \\ + \frac{\lambda_T}{S-1}\sum_{u=1}^{S-1}|t_{u+1} - t_u|. \end{aligned} \tag{4}$$

Here the smoothness constraint is applied on t to guarantee that M is also smooth in the temporal dimension.

4.3 Gradient-free approach: Forecasting Set

If the anomaly detection model is non-differentiable, we propose an alternative algorithm that generates a counterfactual ensemble explanation using an appropriate sampling mechanism. The pseudo-code for this approach is given in Algorithm 2. We describe the steps in detail here. Machine learning models for time series data sometimes rely on sampling in the context of probabilistic forecasting. Here, we will train an

Algorithm 2 Gradient-free counterfactual ensemble explanation algorithm.

Input: The anomalous time series window with its context subwindow $W = [W_C, W_S]$, the anomaly detection model f, the anomaly threshold θ , the training data $\mathcal{D}$, a probabilistic forecasting model g, $W = [W_C, W_S]$, the number of draws T.
Train the model g to predict on $\mathcal{D}$
Obtain the predictive distribution $g(W_C)$
$I_{FS} = \{\}$
for $t = 1, \ldots, T$ **do**
  Sample from $g(W_C)$: $W_F^{(t)} \sim g(W_C)$
  if $\forall i \in [S], f([W_C, W_F^{(t)}])_i < \theta$ **then**
    Add $\tilde{W}^t = [W_C, W_F^{(t)}]$ to I_{FS}
  end if
end for
Output: The set of counterfactual examples $I_{FS} = \{\}$.

auxiliary probabilistic forecasting method and use it as a generative model of counterfactual subsequences. More precisely, given an input window $W_C \in \mathbb{R}^{L-S \times D}$, our auxiliary model g outputs a distribution over a forecast horizon of S time stamps, $g(W_C)$, from which one can sample forecasting paths. We therefore sample T windows $W_F^{(t)} \sim g(W_C)$, $t \in [T]$, then select the ones that are not anomalous according to the anomaly detection model, i.e., our counterfactual ensemble is given by:

$$I_{FS} = \{W_F^{(t)};\ t \in [T] \text{ st } \forall i \in [S], f([W_C, W_F^{(t)}])_i < \theta\}.$$

Note that one could also subsample the set I_{FS} to obtain a fixed number N of examples. Intuitively, since the probabilistic forecasting model is trained to learn the data distribution, it generates realistic forecast samples. However, the sampling model is oblivious to the original input W_S and therefore the forecasting samples are not restricted to be minimally distant from it. Therefore, in this approach, the Closeness and Sparsity properties are not explicitly accounted for. Nonetheless, one could refine this method by selecting the samples which are closer to the original instance. In our experiments, we study the general behaviour of this method without implementing this minor change. In Section 5, we will construct and evaluate this approach with a Feed Forward Neural Network (FFNN) for univariate data and a DeepVAR model Salinas et al. (2019) for multivariate data from the GluonTS package (Alexandrov et al., 2020) [1].

5 Experiments

In this section, we test and compare the performances of our method on two differentiable models, and the relative advantages of its five variants (i.e., ICE, DPE, FS, Sparse ICE, Sparse DPE) in multiple contexts. For this analysis, we have considered two DL anomaly detection models, NCAD Carmona et al. (2021) and USAD Audibert et al. (2020), and four benchmark time series datasets. We report in Section 5.4 a qualitative evaluation of our counterfactual ensemble explanations and their visualization, and in Section 5.5, a quantitative analysis under the previously defined criteria. Note that this study does not include a comparison to existing baselines, since counterfactual ensemble explanations have not been previously considered for time series data. Although some algorithms such as DiCE Mothilal et al. (2020) exist in the context of tabular data, we do not use them in our context since perturbation methods are adapted to each data domain Crabbe & van der Schaar (2021). Nonetheless, for the sake of comparison, we also include a naive baseline, which mechanism is described in Section 5.1. Section 5.2 and Section 5.3 provide additional details on the explainability metrics and the hyperparameters selection procedure.

[1] https://ts.gluon.ai/stable/ (accessed on September 11th 2022)

5.1 Experimental set-up

5.1.1 Datasets

To evaluate our explainability method, we test it on four data sets that are used to benchmark anomaly detection algorithms on time series, see for example Carmona et al. (2021); Su et al. (2019a); Audibert et al. (2020):

- **KPI:** [2] this data set contains 29 univariate time series. It was released in the AIOPS data competition and consists of Key Performance Indicator curves from different internet companies in 1 minute interval.
- **YAHOO:** [3] this data set was published by Yahoo labs and consists of 367 real and synthetic univariate time series.
- **Server Machine Dataset (SMD):** [4] this dataset contains 28 time series with 38 dimensions, collected from a machine in large internet companies Su et al. (2019a).
- **Soil Moisture Active Passive satellite (SMAP):** [5] this NASA data set published by Hundman et al. (2018) contains 55 times series with 25 dimensions.

The main properties of these data sets are summarized in Table 1. These datasets are suitable for evaluating our explainability method since it contains synthetic and real time series anomalies, in diverse time series domains: Key Performance Indicators, server machines, satellite data, etc. We use these datasets since these are commonly used by SOTA anomaly detection methods Carmona et al. (2021); Su et al. (2019a); Audibert et al. (2020). We note that for these data sets, ground-truth labels of anomalies are available. However, this data does not contain additional context or information on the anomalies, consequently there is no ground-truth explanation, *a fortiori* counterfactual example. This is however a common setup in explainability, and, when user studies are not feasible, one needs to resort to proxies for performing a quantitative evaluation Verma et al. (2020). In Section 5.2, we will define our explainability metrics, which have been previously proposed in multiple data domains (see for instance Mothilal et al. (2020) and Verma et al. (2020)).

More precisely, we use the test sets of each dataset, which correspond to the last 50% time stamps of each time series Carmona et al. (2021). When needed, the training and validation sets contain respectively the first 30% and subsequent 20% time stamps. We note that all these datasets have ground-truth anomaly labels on the test set, and in our evaluation, we only compute counterfactual ensemble explanations for the ground-truth anomalies detected by each model (i.e., the *True Positives*).

In practice, our method could be applied on all the detected anomalies, i.e., on both the *True Positives* (TPs) and the *False Positives* (FPs) (i.e., the observations with anomalous predicted labels that are not ground-truth anomalies). However, we consider a practical case where the user is able to analyse only the true anomalies (i.e., the TPs) and wants to know what changes would render this input non-anomalous. However, we also performed a complementary analysis to test our method on FPs (see Appendix B.2). These experiments indicate that the performance of our method on FPs is better in terms of our explainability metrics than on TPs. One explanation for this empirical observation is that a small perturbation of the original FP anomalies is often enough to find good counterfactual explanations using our method. Since explaining TPs is more challenging than FPs, we focus on TPs in the main text and report the FP experiments in Appendix B.2.

5.1.2 Anomaly detection models

In our experimental evaluation, we have selected two differentiable SOTA models with distinct temporal neural networks mechanisms. The first one, Neural Contextual Anomaly Detection (NCAD) Carmona et al.

[2] https://github.com/NetManAIOps/KPI-Anomaly-Detection (accessed on September 11th 2022)
[3] https://webscope.sandbox.yahoo.com/catalog.php?datatype=s&did=70 (accessed on September 11th 2022)
[4] https://github.com/NetManAIOps/OmniAnomaly (accessed on September 11th 2022)
[5] https://github.com/khundman/telemanom (accessed on September 11th 2022)

(2021), uses a temporal convolutional network and subdivides time series into windows that include a context part. The second one, UnSupervised Anomaly Detection (USAD) Audibert et al. (2020), is based on a LSTM Auto-Encoder and predicts anomalies on suspect windows without explicit context windows. Neither of these models are interpretable-by-design, but both have SOTA performances on the benchmark anomaly detection datasets and reasonable training times (around 90 min). Before evaluating our explainability method, we train these models using the procedure described in their respective papers. More details on these models and their detection performance on the benchmark datasets are reported in Appendix A.

5.1.3 Naive counterfactual ensemble explanation

As previously noted, there is no existing method for generating an ensemble of counterfactual examples for time series. We therefore propose a simple interpolation baseline that does not require any training nor optimization procedure. The main idea is similar to the Forecasting Set approach, but here the sampling mechanism is "naive". For each tested window W containing an anomaly in W_S, we draw a sample by interpolating the anomalous window W_S and a constant window with a random weight. The constant window repeats the observation from the timestamp immediately before the anomaly, i.e., $[W_C]_{L-S}$. Thus, for $i \in [N]$, a sample $\widetilde{W}_S^{naive,i}$ is defined as:

$$\widetilde{W}_S^{naive,i} = w_i W_S + (1 - w_i) X_{-1}, \tag{5}$$

where $w_i \overset{i.i.d.}{\sim} U[0,1]$ and $X_{-1} = [[W_C]_{L-S}, \ldots, [W_C]_{L-S}] \in \mathbb{R}^{S \times D}$. As in Section 4.3, we also select the samples that are not anomalous under the model, i.e., the naive counterfactual ensemble is finally:

$$I_N = \{\widetilde{W}^{naive,i} = [W_C, \widetilde{W}_S^{naive,i}];\ i \in [N] \text{ st } \forall t \in [S], f(\widetilde{W}^{naive,i})_t < \theta\}$$

5.2 Explainability metrics

To evaluate the utility of our method, we compute the following metrics as proxies of the criteria defined in Section 3:

- **Failure rate:** This metric accounts for the Validity or algorithm Correctness criteria. For the gradient-based methods (DPE, ICE and their sparse variants), it is defined as the percentage of times our method fails to output an ensemble of N counterfactual examples. For the Forecasting Set approach and naive sampling baseline, the failure rate corresponds to the rejection rate of the sampling scheme.
- **Distance:** The Closeness criterion is measured in terms of the Dynamic Time Warping (DTW) distances between each example of the counterfactual ensemble and the original anomalous window. The DTW distance is generally more adapted to time series data than the Euclidean distance.
- **Implausibility:** since the Plausibility property is not easy to evaluate without expert knowledge of the particular data domain, we decompose it into the three following proxy metrics that cover different notions of deviation from an estimated normal behaviour:
  - DTW distance to a reference time series, here, the median sample from the Forecasting Set approach **(Implausibility 1)**;
  - Temporal Smoothness (**Implausibility 2**), defined as

$$\sum_{i=1}^{D} \sum_{t=1}^{S-1} |[\widetilde{W}_S]_{(t+1)i} - [\widetilde{W}_S]_{ti}|.$$

  - Negative log-likelihood under the probabilistic forecasting distribution g, if available **(Implausibility 3)**.

  We compute the latter metrics for each example of the counterfactual ensemble explanation.

- **Diversity:** the range of values spanned in a counterfactual ensemble is evaluated by the variance of the counterfactual examples at each timestamp.
- **Sparsity correctness:** for multivariate time series, if additional information on the anomalous dimensions in the ground-truth anomalies is available, we compute the precision and recall scores of the sparse variants of DPE and ICE in identifying the dimensions to perturb.

5.3 Hyperparameters selection

The hyperparameters of our counterfactual explanation method with the gradient-based approaches are selected by testing all configurations of $\lambda_1 = \lambda_2, \lambda_T$ in the set $\{0.001, 0.01, 0.1, 1.0\}$, σ_{max} in $\{3, 5, 10\}$ and the learning rate of the SGD algorithm in $\{0.01, 0.1, 1.0, 10.0, 1000.0, 10000.0\}$. As an explainability method can be finely tuned on a particular problem and dataset, the configurations could be evaluated on all the anomalies in the test set. However, for computational time efficiency reasons, we run this evaluation on 100 randomly chosen anomalies, then evaluate the final performance of the chosen configuration on the entire test set. An exception holds for the the SMAP dataset, which contains less than 100 anomalies detected by the models, therefore we run the configurations' evaluation on the whole test set. For each dataset and detection model, we select the set of hyperparameters having the minimal Implausibility 2, given that the failure rate is kept under a pre-defined level, (see Figures 7, 8, 9 and 10 and tables in Appendix D). We note that here focusing on the Implausibility 2 criterion is an arbitrary choice, and one could use instead any other explainability metric. Moreover, we run the SGD algorithm for 1000 iterations and select a maximum of $N = 100$ counterfactual examples along the optimization path. The hyperparameters of the probabilistic models in the Forecasting Set approach are reported in Table 8 in Appendix D. Finally, in order to provide a ready-to-use method, we also suggest a default set of hyperparameters in Table 13 in Appendix D. For all datasets, models and approaches, we use suspect windows of $S = 10$ time stamps and margin parameter $c = 0$.

5.4 Qualitative analysis

Similarly to image classification settings Zeiler & Fergus (2013), visualizations in the time series domain can be human-friendly tools to communicate model explanations, in particular in univariate or low-dimensional settings. In our time series anomaly detection context, we propose to visualize our counterfactual ensemble explanation together with the original time series for which a prediction was made, possibly with an added context window (see Section 3) and on a restricted number of channels. Since the anomaly prediction score given by the explained model is a scalar, we can leverage a color scale to indicate the score of each counterfactual example in the ensemble.

On Figure 2, we present a visualization of our method on two anomalies from the KPI dataset, detected by NCAD and USAD. On each panel, we plot a sub-window of the original time series containing anomalous features in the last 10 time stamps, as well as each counterfactual example given by a variant of our method applied to one of the detection model. Each counterfactual only differs at the anomalous features and the color scheme indicates its anomaly score under the explained model. We argue that this representation allows to deem the range of time series values and prediction scores spanned by the different counterfactuals in our ensemble explanation, and therefore effectively informs on the model's sensitivity and local decision boundary.

We can then visually compare the different variants of our method and the explanations for two detection models. We observe that the counterfactual ensemble explanation from DPE (in red color scale), ICE (in green), and FS (in purple) are quite dissimilar, although they all globally lessen the amplitude of the spike outliers' features. In fact, on the one hand, DPE produces counterfactual ensembles that are less diverse than the other approaches, and relatively close to the original input. This is coherent with the fact that the perturbations are constrained by the dynamic mechanism. On the other hand, ICE's counterfactual sets cover a much larger range of values and therefore allows to visualize more clearly how the anomaly score evolves for different magnitudes of the spikes. This explanation may thus be more informative here since it spans a larger range of time series values. In contrast, the counterfactual ensembles generated by FS do not have the aforementioned interpretation but seem to visually correspond to the expected behaviour given the

shape of the context windows. The previous preliminary observations seem consistent for the two models and confirmed on several other anomalies (see for instance the additional visualizations in Appendix C).

In summary, our counterfactual ensemble explanations effectively contain diverse perturbations of the input time series. These perturbations trigger a change in the detected label of the anomalous sub-sequence, with a small number of altered features. The three approaches, ICE, DPE and FS, bring different insights on the model's prediction, the time series distribution and the possible perturbations to apply to change the former. Their relative advantages may therefore depend on the particular time series context and usage of the counterfactual explanation.

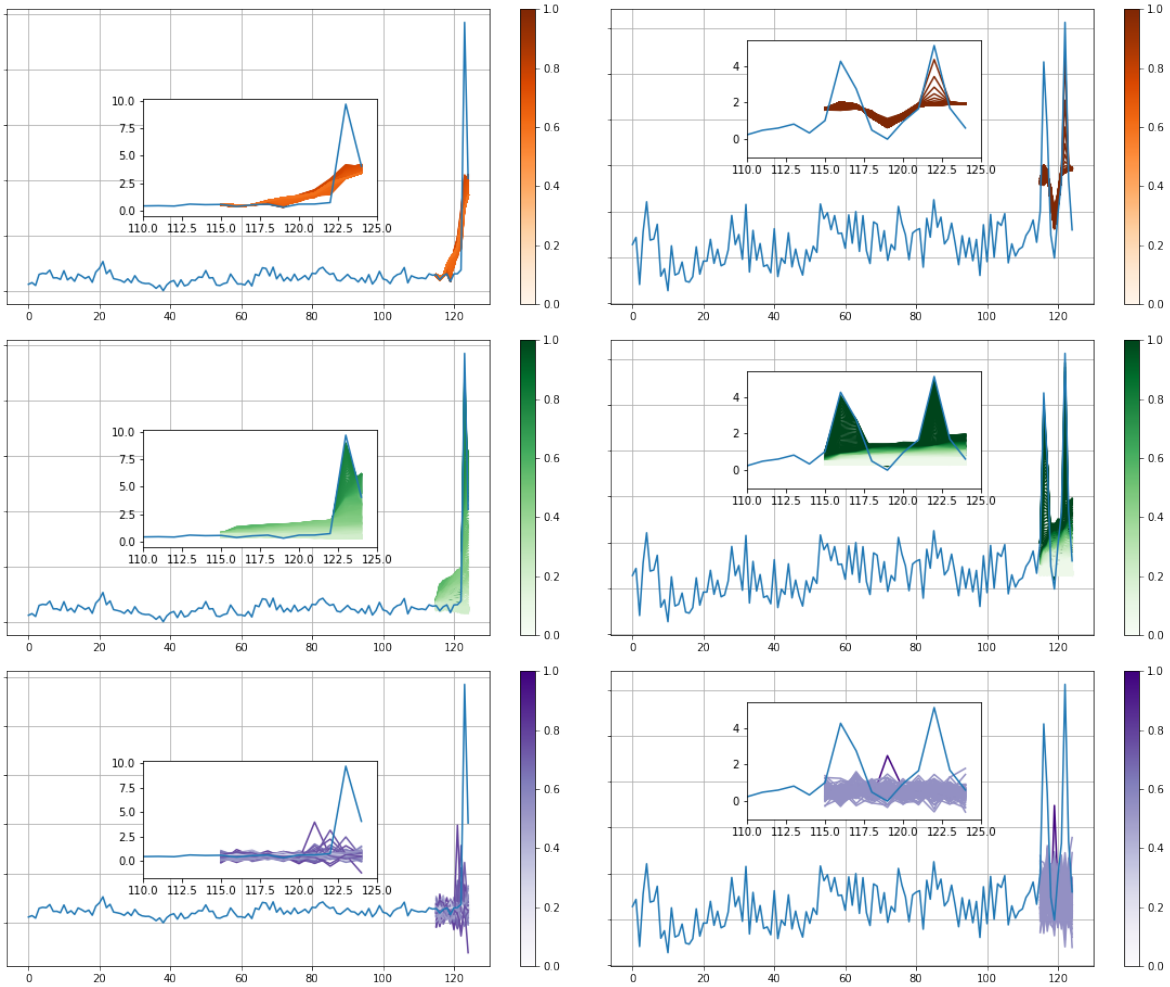


Figure 2: Time series windows containing an anomaly and our counterfactual ensemble explanations, obtained with DPE (first row), ICE (second row) and FS (third row) from the KPI dataset. The first (resp. second) column corresponds to an anomaly that has been detected by NCAD (resp. USAD). Each window includes a context part of 115 time stamps and an abnormal part of 10 time stamps at the end of the window. The original observations are plotted in blue, while the counterfactual examples appear in red, green or purple color scales for respectively DPE, ICE and FS.

Dataset	Dimensions	Number of time series	Total number of time stamps	Total number of anomalies in test set
KPI	1	29	5922913	54560
Yahoo	1	367	609666	2963
SMD	38	28	1416825	29444
SMAP	25	55	584860	57079

Table 1: Succinct description of the four benchmark datasets

5.5 Numerical evaluation

The numerical results discussed in this section are obtained in the set-up described in Section 5.1. However, for parsimony of exposition, our results on the KPI and the SMAP datasets have been moved to Appendix B. We also add a partial sensitivity analysis of our method in Appendix E.

The results on univariate datasets (see Table 2 and Table 5 in Appendix B.1), show that our method has fairly small failure rates (except for the Yahoo dataset and the USAD model). In particular a rate smaller than 10% can be achieved with at least one variant in most pairs (model, dataset), leading to a consequent improvement over the naive procedure. We note that while the DPE variant seems to be valid more often than ICE on the NCAD model, it is the contrary for the USAD model; this difference is possibly due to the distinct internal mechanisms of these models.

Moreover, the analysis of the other explainability metrics supports the qualititative interpretation from Section 5.4. The Distance metric confirms than the gradient-based approaches, DPE and ICE, provides in almost all cases the closest counterfactuals in average, i.e., the least perturbed examples. Note that it sometimes occurs that the naive baseline has a small distance, however it always have a high failure rate. Besides, the Implausibility metrics validate the observation that FS generates the most realistic counterfactual examples in average, in particular in terms of Implausibility 1 (distance to median forecast sample) and Implausibility 3 (NLL under the probabilistic forecasting distribution). This is in fact quite expected since these quantities are directly derived from the forecasting sampling scheme. However, these counterfactuals are less smooth (higher score in Implausibility 2) than for DPE and ICE, which regularize the time series smoothness in the objective functions equation 2 and equation 1.

Finally, DPE and ICE provide a more diverse counterfactual ensemble in most cases in general, but their relative ranking is not clear from these experiments. We conjecture that this metric is particularly sensitive to the learning rate of the SGD algorithm, and the subsampling procedure after the objective minimization (see Section 4.1). In Appendix E, we test our first hypothesis on a small sample of anomalies. We observe in this case that the Diversity criterion is consistently higher for ICE, and greatly increases with the learning rate, at the cost of a higher failure rate.

The numerical results on the multivariate data sets are reported in Table 3 and Table 6 in Appendix B.1. These experiments showcase that our method also generates valid counterfactual ensemble explanations in this setting, with even a failure rate of 0% for the USAD model. Our method fails more frequently on the NCAD model, however, the sparse variants are more often successful. This indicates that imposing a sparsity constraint over the modified dimensions also helps to find valid counterfactuals. Consistent with the univariate datasets, FS produces the most realistic counterfactual examples while the gradient-based approach achieves a better Distance score. We note that in this case the Implausibility 3 metric is not available since the forecast distribution likelihood function in the DeepVAR model is not available [6]. Moreover, the sparse variants seem to correctly identify some of the anomalous channels (precision greater than 0.6 for the USAD model).

Nonetheless, we noted the greater difficulty of tuning the hyperparameters of our method and ranking its variants on these high-dimensional datasets compared to univariate data. In the latter, the default set of

[6] `https://ts.gluon.ai/stable/api/gluonts/gluonts.model.deepvar.html?highlight=deepvar#module-gluonts.model.deepvar` (accessed on September 11th 2022)

NCAD on Yahoo						
Method	Failures (%)	Distance	Implausibility 1	Implausibility 2	Implausibility 3	Diversity
DPE	**9.2**	2.49 (4.91)	1.23 (1.37)	1.42 (2.19)	2.21 (4.76)	0.01
ICE	17.4	**1.54 (1.21)**	0.78 (1.37)	2.26 (1.67)	1.40 (5.17)	0.05
FS	56.6	6.06 (16.38)	**0.27 (0.22)**	3.36 (1.99)	**-0.29 (0.79)**	**0.10**
Naive	72.2	2.69 (5.29)	1.04 (1.26)	**1.32 (1.74)**	1.89 (3.34)	0.05
USAD on Yahoo						
Method	Failures (%)	Distance	Implausibility 1	Implausibility 2	Implausibility 3	Diversity
DPE	29.1	5.20 (18.00)	6.42 (26.81)	0.42 (2.00)	3.74 (6.96)	0.05
ICE	**25.5**	6.66 (25.54)	2.68 (11.46)	**0.40 (1.16)**	2.48 (4.61)	**3.23**
FS	65.1	14.48 (46.03)	**0.48 (0.72)**	0.55 (0.58)	**-0.11 (1.12)**	0.61
Naive	45.8	**4.82 (18.36)**	2.85 (16.31)	0.52 (1.60)	3.25 (6.00)	3.19

Table 2: Performance of our explainability method and the naive baseline in terms of Validity, Closeness, Plausibility and Diversity on the Yahoo dataset and the NCAD (first panel) and USAD (second panel) anomaly detection models. We report the average scores and standard deviations (in brackets) over the counterfactual ensembles. We recall that *Implausibility 1* is the DTW distance to the median forecasting sample, *Implausibility 2* is the temporal smoothness, and *Implausibility 3* is the negative log-likelihood under the probabilistic forecasting output distribution. For all metrics except *Diversity*, we assume that a lower value is better, and the best score is highlighted in bold.

hyperparameters achieves an acceptable performance and allows to quickly compare the relative advantages of an approach for a specific pair (detection model, dataset). We therefore conclude by recalling that example-based explainability methods for multivariate time series are still in their early development, and providing general methods and tuning procedures to generate useful explanations over the instances of a dataset is still an open problem.

6 Discussion & Conclusion

In this work, we have introduced a novel type of post-hoc explainability method called *Counterfactual Ensemble Explanation* for anomaly detection models in time series. Our approach is model-agnostic, can be applied to any differentiable detection model, and is delineated into different variants according to the context. With DPE, one can apply a domain-specific perturbation mechanism to the input time series, while ICE does not require such specification. For high-dimensional time series, our sparse variants, *Sparse DPE* and *Sparse ICE* provide counterfactual examples modifying only a few dimensions of the time series. Additionally, we have proposed a gradient-free approach that uses a probabilistic forecasting technique as a generative scheme and can be applied to any detection model.

Our real-world experiments on four benchmark data sets show that the counterfactual framework, augmented with an ensemble approach, improves the interpretability of two deep-learning models and the anomalies the latter detects. In particular, our visualization tool allows to gauge the change in anomaly scores with respect to a large perturbation range of time series features. In the absence of competitive methods, we quantitatively compare our explanations to a *naive* counterfactual ensemble method using several explainability metrics.

In comparison to existing model-agnostic explainability methods for time series, our approach conveys more quantitative information on the model's sensitivity that a feature-saliency approach such as DynaMask (Crabbe & van der Schaar, 2021) and a richer contrastive explanation than single-counterfactual methods such as Delaney et al. (2021); Ates et al. (2021). Nonetheless, our proposed counterfactual ensemble expla-

NCAD on SMD						
Method	Failures (%)	Precision / Recall	Distance	Implausibility 1	Implausibility 2	Diversity
DPE	**17.1**	-	**8.46 (13.07)**	50.76 (110.40)	12.12 (28.13)	1.21
ICE	42.9	-	79.62 (120.27)	23.69 (28.24)	47.73 (58.32)	**4639.11**
Sparse DPE	20.0	**0.22** / 0.10	36.12 (77.87)	29.03 (54.36)	5.61 (11.30)	**4687.05**
Sparse ICE	20.0	0.20 / **0.33**	26.01 (37.07)	62.65 (107.25)	10.88 (16.72)	174.39
FS	30.0	-	78.46 (157.57)	**1.62 (2.33)**	**1.49 (1.31)**	35.88
Naive	79.8	-	25.06 (53.66)	45.45 (93.03)	9.42 (18.79)	3255.92
USAD on SMD						
Method	Failures (%)	Precision / Recall	Distance	Implausibility 1	Implausibility 2	Diversity
DPE	**0.0**	-	139.02 (261.44)	258.31 (464.18)	41.19 (79.11)	**23339.20**
ICE	**0.0**	-	**31.81 (9.27)**	342.19 (708.88)	22.64 (45.16)	0.52
Sparse DPE	**0.0**	**0.68** / 0.07	115.48 (206.46)	293.70 (679.88)	19.84 (34.98)	105.45
Sparse ICE	**0.0**	0.61 / **0.28**	216.44 (316.05)	172.58 (475.15)	**8.35 (17.52)**	477.43
FS	**0.0**	-	366.57 (672.44)	**18.10 (48.25)**	**8.57 (20.63)**	12175.65
Naive	73.4	-	49.83 (60.13)	475.42 (879.38)	27.45 (45.43)	649.44

Table 3: Performance of our explainability method and the naive baseline in terms of Validity, Closeness, Plausibility and Diversity on the SMD dataset and the NCAD (first panel) and USAD (second panel) anomaly detection models. We report the average scores and standard deviations (in brackets) over the counterfactual ensemble. We recall that *Implausibility 1* is the DTW distance to the median forecasting sample and *Implausibility 2* is the temporal smoothness. For all metrics except *Diversity*, *Precision* and *Recall*, we assume that a lower value is better, and the best score is highlighted in bold.

nation for time series models is an attempt in the interpretation of these models using diverse instance-based methods, in particular in the challenging high-dimensional context.

Although our method offers greater flexibility, better explainability performances and specific interpretation might be achieved if more assumptions are put on the detection model. In particular, similarly to Rodriguez et al. (2021), we could adapt our gradient-based approach to use the internal representations of the model rather than the raw time series. Moreover, aggregating the information contained in diverse explanations is still an open problem. One possible extension of our ensemble method would be to provide a rank over the counterfactual examples according to a utility or feasibility metric.

Broader Impact Statement

We do not see any direct negative impact of our work, however ethical concerns could come from the type of time series data our methodology is applied to. Moreover, our method does not rank the counterfactuals in the set of solutions, and their acceptability needs to be assessed by a domain expert. An extension of our method could be to include fairness constraints in the optimisation objective to obtain “fair" counterfactuals.

Acknowledgments

We thank Chris Russell, Dominik Janzig, Lenon Minorics and Lorenzo Stella for their insightful comments on our work.

References

Alexander Alexandrov, Konstantinos Benidis, Michael Bohlke-Schneider, Valentin Flunkert, Jan Gasthaus, Tim Januschowski, Danielle C. Maddix, Syama Rangapuram, David Salinas, Jasper Schulz, Lorenzo Stella, Ali Caner TÃ¼rkmen, and Yuyang Wang. Gluonts: Probabilistic and neural time series modeling in python. *Journal of Machine Learning Research*, 21(116):1–6, 2020. URL `http://jmlr.org/papers/v21/19-820.html`.

Emre Ates, Burak Aksar, Vitus J. Leung, and Ayse K. Coskun. Counterfactual explanations for multivariate time series. *2021 International Conference on Applied Artificial Intelligence (ICAPAI)*, May 2021. doi: 10.1109/icapai49758.2021.9462056. URL `http://dx.doi.org/10.1109/ICAPAI49758.2021.9462056`.

Julien Audibert, Pietro Michiardi, Frédéric Guyon, Sébastien Marti, and Maria A Zuluaga. Usad : Unsupervised anomaly detection on multivariate time series. In *KDD2020 - The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, San Diego, USA, 2020.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, 2018.

João Bento, Pedro Saleiro, André F. Cruz, Mário A.T. Figueiredo, and Pedro Bizarro. Timeshap: Explaining recurrent models through sequence perturbations. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, Aug 2021. doi: 10.1145/3447548.3467166. URL `http://dx.doi.org/10.1145/3447548.3467166`.

Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 648–657, 2020.

Umang Bhatt, Isabel Chien, Muhammad Bilal Zafar, and Adrian Weller. Divine: Diverse influential training points for data visualization and model refinement. *ArXiv*, abs/2107.05978, 2021.

Ane Blázquez-García, Angel Conde, Usue Mori, and Jose A. Lozano. A review on outlier/anomaly detection in time series data. *ACM Comput. Surv.*, 54(3), apr 2021. ISSN 0360-0300. doi: 10.1145/3444690. URL `https://doi.org/10.1145/3444690`.

Andy Brown, Aaron Tuor, Brian Hutchinson, and Nicole Nichols. Recurrent neural network attention mechanisms for interpretable system log anomaly detection. In *Proceedings of the First Workshop on Machine Learning for Computing Systems*, MLCS'18, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450358651. doi: 10.1145/3217871.3217872. URL https://doi.org/10.1145/3217871.3217872.

Mattia Carletti, Matteo Terzi, and Gian Antonio Susto. Interpretable anomaly detection with diffi: Depth-based isolation forest feature importance, 2021.

Chris U. Carmona, François-Xavier Aubet, Valentin Flunkert, and Jan Gasthaus. Neural contextual anomaly detection for time series, 2021.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. An exact and robust conformal inference method for counterfactual and synthetic controls. *Journal of the American Statistical Association*, 116(536):1849–1864, 2021. doi: 10.1080/01621459.2021.1920957. URL https://doi.org/10.1080/01621459.2021.1920957.

Edward Choi, Mohammad Taha Bahadori, Joshua A. Kulas, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 3512–3520, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Jonathan Crabbe and Mihaela van der Schaar. Explaining time series predictions with dynamic masks. In *ICML*, 2021.

Susanne Dandl, Christoph Molnar, Martin Binder, and Bernd Bischl. Multi-objective counterfactual explanations. *Lecture Notes in Computer Science*, pp. 448–469, 2020. ISSN 1611-3349. doi: 10.1007/978-3-030-58112-1_31. URL http://dx.doi.org/10.1007/978-3-030-58112-1_31.

Eoin Delaney, Derek Greene, and Mark T. Keane. Instance-based counterfactual explanations for time series classification. In Antonio A. Sánchez-Ruiz and Michael W. Floyd (eds.), *Case-Based Reasoning Research and Development*, pp. 32–47, Cham, 2021. Springer International Publishing. ISBN 978-3-030-86957-1.

Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2950–2958, 2019. doi: 10.1109/ICCV.2019.00304.

Ville Hautamaki, Pekka Nykanen, and Pasi Franti. Time-series clustering by approximate prototypes. In *2008 19th International Conference on Pattern Recognition*, pp. 1–4, 2008. doi: 10.1109/ICPR.2008.4761105.

Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, jul 2018. doi: 10.1145/3219819.3219845. URL https://doi.org/10.1145%2F3219819.3219845.

Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. Model-agnostic counterfactual explanations for consequential decisions. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 895–905. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/karimi20a.html.

Isak Karlsson, Jonathan Rebane, Panagiotis Papapetrou, and Aristides Gionis. Locally and globally explainable time series tweaking. *Knowledge and Information Systems*, 62(5):1671–1700, 2020. doi: 10.1007/s10115-019-01389-4.

Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. Comparison-based Inverse Classification for Interpretability in Machine Learning. In Jesús Medina, Manuel Ojeda-Aciego, José Luis Verdegay, David A. Pelta, Inma P. Cabrera, Bernadette Bouchon-Meunier, and Ronald R. Yager (eds.), *17th International Conference on Information Processing and Management of*

Uncertainty in Knowledge-Based Systems (IPMU 2018), Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations, pp. 100–111, Cadix, Spain, June 2018. Springer Verlag. doi: 10.1007/978-3-319-91473-2_9. URL https://hal.sorbonne-universite.fr/hal-01905982.

Dan Ley, Umang Bhatt, and Adrian Weller. Diverse, global and amortised counterfactual explanations for uncertainty estimates. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7):7390–7398, Jun. 2022. doi: 10.1609/aaai.v36i7.20702. URL https://ojs.aaai.org/index.php/AAAI/article/view/20702.

Qianli Ma, Wanqing Zhuang, Sen Li, Desen Huang, and Garrison W. Cottrell. Adversarial dynamic shapelet networks. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 5069–5076. AAAI Press, 2020. URL https://aaai.org/ojs/index.php/AAAI/article/view/5948.

Pankaj Malhotra, Lovekesh Vig, Gautam M. Shroff, and Puneet Agarwal. Long short term memory networks for anomaly detection in time series. In *ESANN*, 2015.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267:1–38, 2019.

Christoph Molnar. *Interpretable Machine Learning*. 2019.

Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Jan 2020. doi: 10.1145/3351095.3372850. URL http://dx.doi.org/10.1145/3351095.3372850.

Qingyi Pan, Wenbo Hu, and Jun Zhu. Series saliency: Temporal interpretation for multivariate time series forecasting. *ArXiv*, abs/2012.09324, 2020.

Pau Rodriguez, Massimo Caccia, Alexandre Lacoste, Lee Zamparo, Issam Laradji, Laurent Charlin, and David Vazquez. Beyond trivial counterfactual explanations with diverse valuable explanations, 2021.

Chris Russell. Efficient search for diverse coherent explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 20–28, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287569. URL https://doi.org/10.1145/3287560.3287569.

David Salinas, Michael Bohlke-Schneider, Laurent Callot, Roberto Medico, and Jan Gasthaus. High-dimensional multivariate forecasting with low-rank gaussian copula processes. In *NeurIPS*, 2019.

Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pp. 2828–2837, New York, NY, USA, 2019a. Association for Computing Machinery. ISBN 9781450362016. doi: 10.1145/3292500.3330672. URL https://doi.org/10.1145/3292500.3330672.

Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2828–2837, 2019b.

Sahil Verma, John Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: A review, 2020.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr, 2018.

James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):56–65, 2020. doi: 10.1109/TVCG.2019.2934619.

Matthew Zeiler and Rob Fergus. Visualizing and understanding convolutional neural networks. volume 8689, 11 2013. ISBN 978-3-319-10589-5. doi: 10.1007/978-3-319-10590-1_53.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320, 2005. ISSN 13697412, 14679868. URL http://www.jstor.org/stable/3647580.

A Technical details and performance of the selected anomaly detection models

In this section, we provide some technical details on the two anomaly detection models selected for the evaluation of our explainability method reported in Section 5. In Table 4, we report their anomaly detection performance on the benchmark datasets, after training with the hyperparameter sets reported in their respective papers when available. Otherwise, we select the models' hyperparameters on a validation set (20% of the time series) using the best adjusted F1-score.

Neural Contextual Anomaly Detection (NCAD) Carmona et al. (2021) : This method splits time series into subwindows $(W^i)_i$ and embeds them using a temporal convolutional network (TCN). Each W^i is subdivided into a context part and a suspect part (typically much smaller than the former), i.e., $W^i = [W^i_C, W^i_S]$. An embedding of the context window W^i_C is also computed by the TCN, then the distance between the embeddings of W_i, denoted z^i, and W^i_C, denoted z^i_C, is evaluated. The algorithm finally labels W^i_S as anomalous if the latter distance is greater than a chosen threshold, i.e, if $d(z^i, z^i_C) > \eta$ with $d(.,.)$ the Euclidean distance for instance and $\eta > 0$. The intuition behind this method is that a large distance between the embeddings of a window and its context part means that the suspect part induces a significant shift of z^i_C in the embedding space. Since the embedding of the context window should reflect the normal behaviour, this deviation thus indicates the presence of an anomaly in W^i_S. For our experiments, we use the open-source implementation. [7]

UnSupervised Anomaly Detection (USAD) Audibert et al. (2020): This reconstruction model splits time series into subwindows that are reconstructed by a LSTM-based AutoEncoder. The latter contains a neural network, called encoder, that embeds each window into a latent representation, and another neural network, called decoder, that maps back the embedding into the original input space. The reconstruction error, i.e., the distance in the time series domain between the original input and the reconstructed output, is used as an anomaly score (a high value of this error leads to the corresponding window to be labelled as anomalous). We use the open source implementation provided by the authors [8] and the hyperparameters provided in the paper for the two multivariate data sets, i.e. SMD and SMAP. For the KPI dataset, the final USAD model is trained for 80 epochs and has windows of size 5, hidden size of 10 and downsampling rate of 0.01. For the Yahoo data, the window size is 10, hidden size of 10 and downsampling rate of 0.05.

Model	KPI	Yahoo	SMD	SMAP
NCAD	0.789	0.772	0.806	0.922
USAD	0.946	0.741	0.643	0.972

Table 4: F1-scores of the two anomaly detection models, i.e., NCAD and USAD, on the four benchmark datasets.

[7] https://github.com/Francois-Aubet/gluon-ts/tree/adding_ncad_to_nursery/src/gluonts/nursery/ncad
[8] https://curiousily.com/posts/time-series-anomaly-detection-using-lstm-autoencoder-with-pytorch-in-python/

NCAD on KPI						
Method	Failures (%)	Distance	Implausibility 1	Implausibility 2	Implausibility 3	Diversity
DPE	**3.9**	5.94 (15.78)	2.16 (4.71)	3.21 (29.62)	2.74 (2.57)	**1.18**
ICE	19.6	**3.08 (1.21)**	15.31 (115.12)	31.67 (206.70)	2.07 (2.06)	0.26
FS	6.0	32.05 (173.76)	**0.21 (0.20)**	**2.42 (1.97)**	**-0.56 (1.14)**	0.12
Naive	53.4	11.82 (74.03)	2.90 (4.49)	4.57 (7.64)	3.48 (2.51)	0.54
USAD on KPI						
Method	Failures (%)	Distance	Implausibility 1	Implausibility 2	Implausibility 3	Diversity
DPE	5.0	25.22 (121.60)	9.40 (65.02)	1.03 (8.16)	3.13 (3.47)	13.10
ICE	**3.5**	**6.52 (7.30)**	4.99 (63.31)	0.50 (4.43)	1.27 (1.98)	0.28
FS	6.8	38.56 (189.88)	**0.33 (0.29)**	**0.38 (0.28)**	**-0.08 (1.12)**	0.26
Naive	45.4	31.93 (154.62)	2.77 (3.93)	1.42 (6.01)	2.81 (2.48)	**69.88**

Table 5: Performance of our explainability method and the naive baseline in terms of Validity, Closeness, Plausibility and Diversity on the KPI dataset and the NCAD (first panel) and USAD (second panel) anomaly detection models. We report the average scores and standard deviations (in brackets) over the counterfactual ensemble. We recall that *Implausibility 1* is the DTW distance to the median forecasting sample, *Implausibility 2* is the temporal smoothness, and *Implausibility 3* is the negative log-likelihood under the probabilistic forecasting output distribution. For all metrics except *Diversity*, we assume that a lower value is better, and the best score is highlighted in bold.

B Additional numerical results

In this section, we report quantitative evaluations of our explainability method that could not be included in the main text due to space limitation. This section notably contains the results on two benchmark datasets using the procedure described in Section 5, and an additional analysis on False Positives.

B.1 Numerical evaluation on the KPI and SMAP datasets

The results on the KPI and SMAP dataset are respectively in Table 5 and Table 6. Note that these results are included in the discussion in Section 5.5.

B.2 Numerical evaluation on False Positives

In the practical use of anomaly detection models, explanations can also be needed when the model wrongly detects an anomaly in a time series. We recall that we call False Positives the anomalies detected by the model that are not ground-truth anomalies. We present here a numerical evaluation on the False Positives detected by NCAD in the KPI benchmarck dataset. The results in Table 7 can be compared to the results obtained on True Positives (i.e., the ground-truth, detected anomalies) reported in the first panel of Table 5 . We observe that in this case ICE achieves 0% failure rate (instead of almost 20 %), and the naive method has also a significantly smaller number of failures. Moreover, all methods seem to perform better in terms of the Distance and Implausibility metrics. This is probably due to the fact that False Positives need less perturbation to become not anomalous for the model, e.g. if they lie close to the model's local decision boundary. Therefore they may inherently be less distant to the normal behaviour than True Positives and

NCAD					
Method	Failures (%)	Diversity	Distance	Implausibility 1	Implausibility 2
DPE	41.7	0.002	0.19 (0.40)	0.21 (0.28)	**0.01 (0.03)**
DPE sparse	27.8	0.004	0.22 (0.42)	0.29 (0.38)	0.03 (0.04)
ICE	**5.6**	**0.067**	0.26 (0.14)	0.39 (0.37)	0.15 (0.09)
ICE sparse	23.6	0.016	0.15 (0.08)	0.22 (0.21)	0.09 (0.06)
FS	87.6	0.012	0.56 (0.77)	**0.05 (0.04)**	0.05 (0.04)
Naive	84.5	0.003	**0.06 (0.08)**	0.09 (0.03)	0.02 (0.03)
USAD					
Method	Failures (%)	Diversity	Distance	Implausibility 1	Implausibility 2
DPE	**0.0**	0.02	0.62 (0.65)	0.96 (0.53)	0.06 (0.05)
DPE sparse	**0.0**	0.02	0.78 (0.85)	0.82 (0.50)	0.05 (0.06)
ICE	**0.0**	0.17	0.74 (0.75)	0.87 (0.43)	**0.04 (0.03)**
ICE sparse	**0.0**	**0.18**	0.72 (0.76)	0.88 (0.42)	0.06 (0.02)
FS	56.8	0.02	2.23 (1.14)	**0.09 (0.01)**	0.10 (0.03)
Naive	46.9	0.01	**0.14 (0.04)**	0.23 (0.02)	0.07 (0.02)

Table 6: Performance of our explainability method and the naive baseline in terms of Validity, Closeness, Plausibility and Diversity on the SMAP dataset and the NCAD (first panel) and USAD (second panel) anomaly detection models. We report the average scores and standard deviations (in brackets) over the counterfactual ensemble. We recall that *Implausibility 1* is the DTW distance to the median forecasting sample and *Implausibility 2* is the temporal smoothness. For all metrics except *Diversity*, *Precision* and *Recall*, we assume that a lower value is better, and the best score is highlighted in bold.

thus easier instances for our counterfactual explanation method. Besides, the Diversity metric is smaller for DPE and ICE, likely as another effect of the smaller amount of perturbation needed.

NCAD						
Method	Failures (%)	Distance	Implausibility 1	Implausibility 2	Implausibility 3	Diversity
DPE	8.8	**2.22 (1.87)**	2.44 (2.50)	2.36 (2.05)	2.15 (2.22)	0.02
ICE	**0.0**	4.36 (3.00)	**0.28 (0.45)**	**0.61 (0.34)**	0.22 (0.93)	0.12
FS	6.6	4.17 (2.91)	0.30 (0.31)	3.54 (3.61)	**-0.22 (0.95)**	**0.43**
Naive	33.3	2.74 (2.22)	2.19 (2.43)	2.97 (2.56)	2.79 (2.29)	0.16

Table 7: Performance of our explainability method and the naive baseline in terms of Validity, Closeness, Plausibility and Diversity on the false positives in the KPI data detected by the NCAD model. We report the average scores and standard deviations (in brackets) over the counterfactual ensemble. We recall that *Implausibility 1* is the DTW distance to the median forecasting sample, *Implausibility 2* is the temporal smoothness, and *Implausibility 3* is the negative log-likelihood under the probabilistic forecasting output distribution. For all metrics except *Diversity*, we assume that a lower value is better, and the best score is highlighted in bold.

C Complementary visualizations of the explanations

In this section, we report additional visualizations of our counterfactual explanations, as well as illustrations of the sparsity induced by the sparse variants of DPE and ICE. Figures 3 and 4 are visualizations applied to the univariate datasets and respectively the NCAD and USAD. The advantage of Sparse ICE compared to the plain version ICE is shown in Figure 5, where only four channels of the multi-dimensional time series window are plotted. For this anomaly, only one of these dimensions contains an anomalous observation

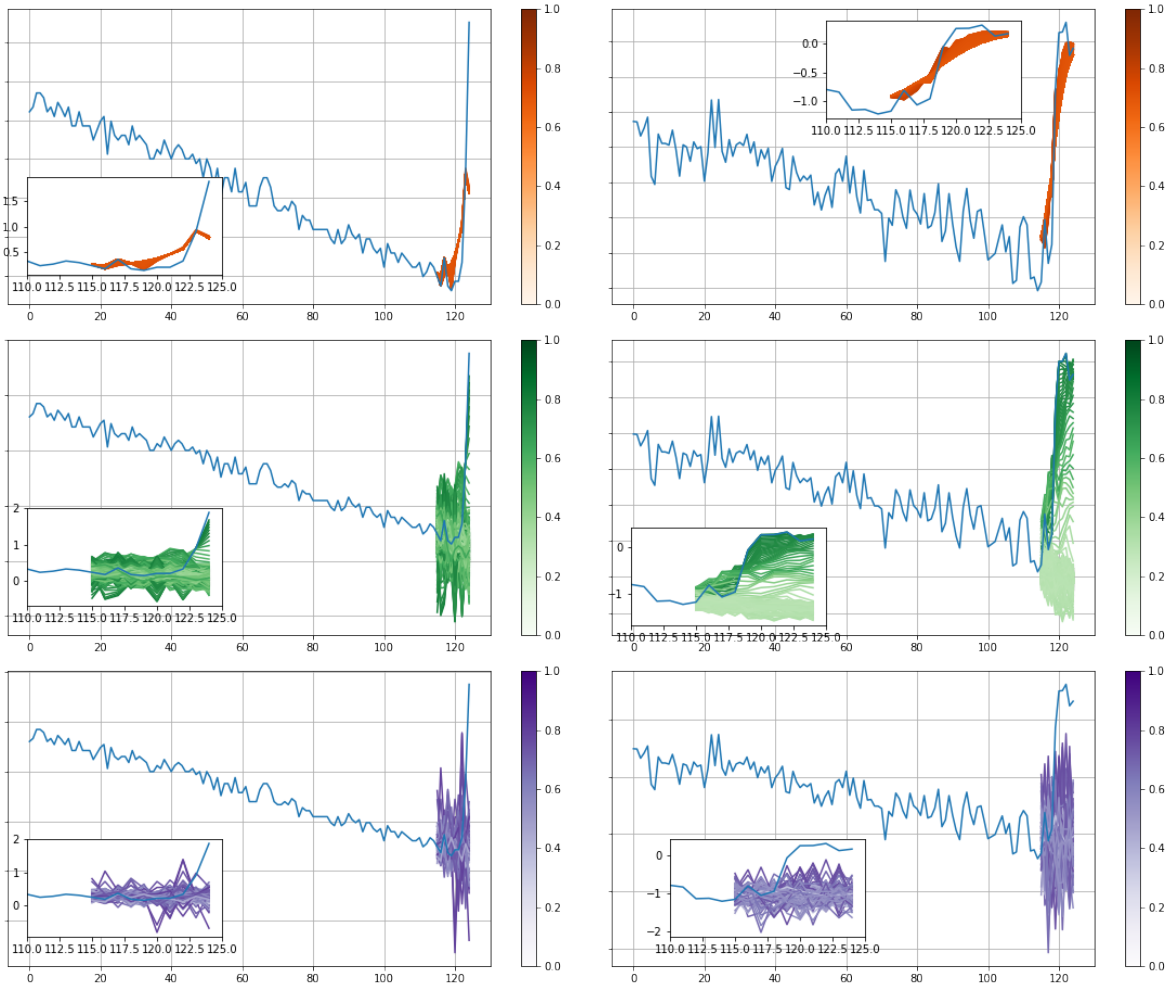


Figure 3: Anomalous windows and counterfactual ensemble explanations obtained with DPE (first row), ICE (second row) and FS (third row) on anomalies in the KPI data set detected by the NCAD model. The columns correspond to two different anomalies. The windows include a context part of 115 time stamps and an abnormal part of 10 time stamps. The original sub-sequence is plotted in blue, while the explanations are in red, green or purple colors for the different variants.

but the counterfactual explanation obtained with the plain ICE perturbs four of them. In contrast, the Sparse ICE variant keeps two dimensions without anomalous features unchanged, leading to a more accurate and readable explanation on this particular anomaly. Similarly, Figure 6 shows two perturbation maps corresponding to examples generated by DPE and its sparse variant. While the plain DPE produces *globally* sparse maps (i.e., in the temporal and dimensional features), Sparse DPE is sparse in dimensions, leading to perturbed examples with few modified channels.

D Illustration of the hyperparameters selection

In this section, we illustrate the hyperparameters selection procedure for our gradient-based method. For each dataset and model, we run our algorithm with several configurations as described in Section 5.3 and select the final one using the failure rate and the Implausibility 1 metric. More precisely, we select a threshold of acceptable failure rate (e.g., 10% or 20%), then amongst the configurations achieving a lower value of the latter, we select the one with the lowest Implausibility 1 value. Figures 7, 8, 10 and 9 show the values of

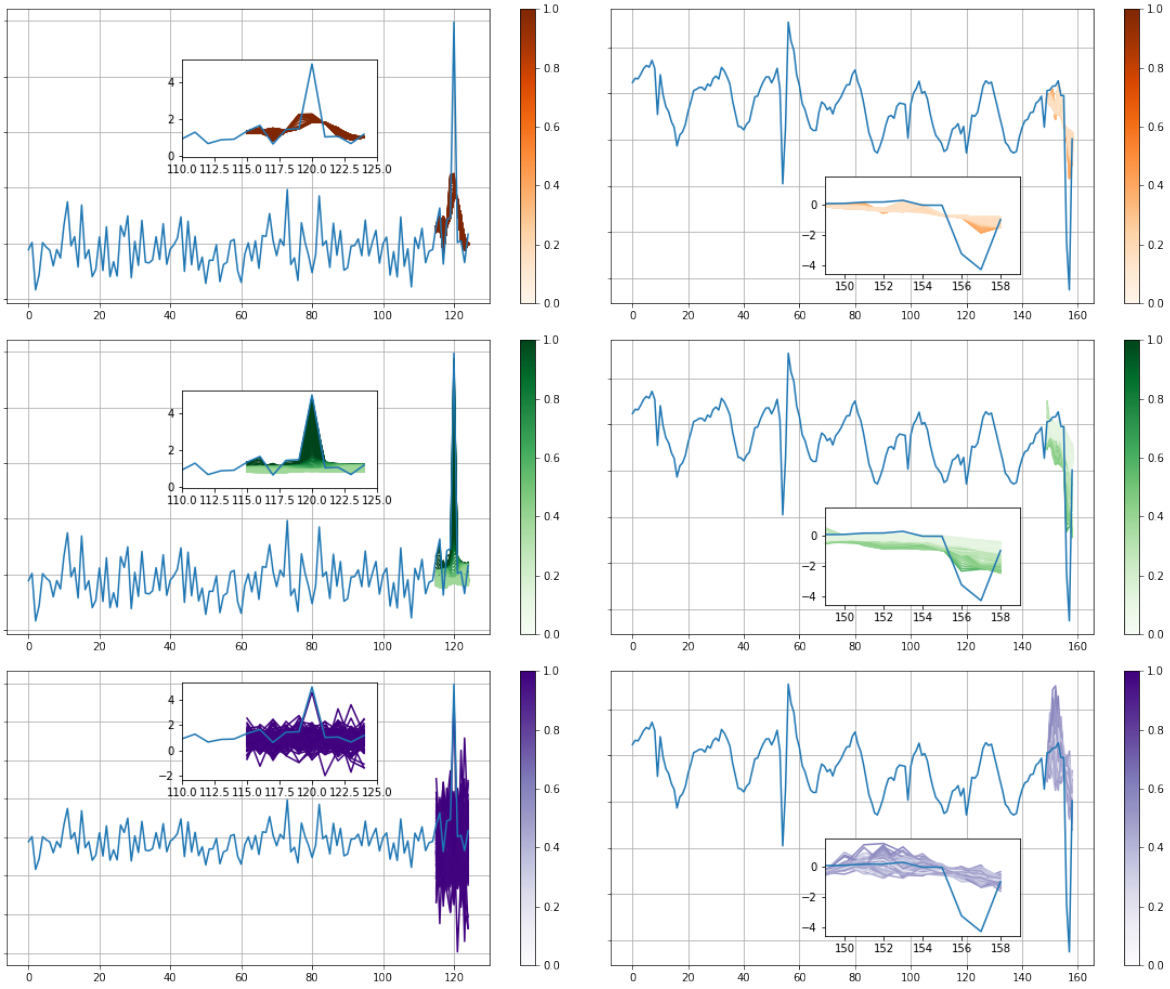
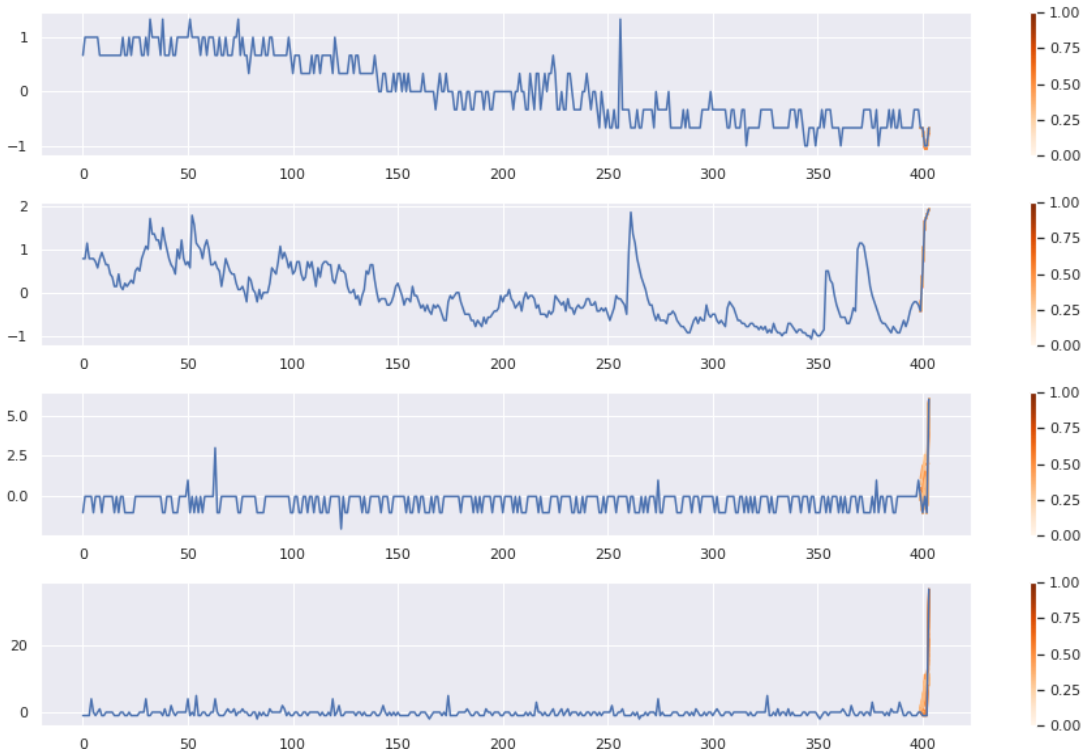


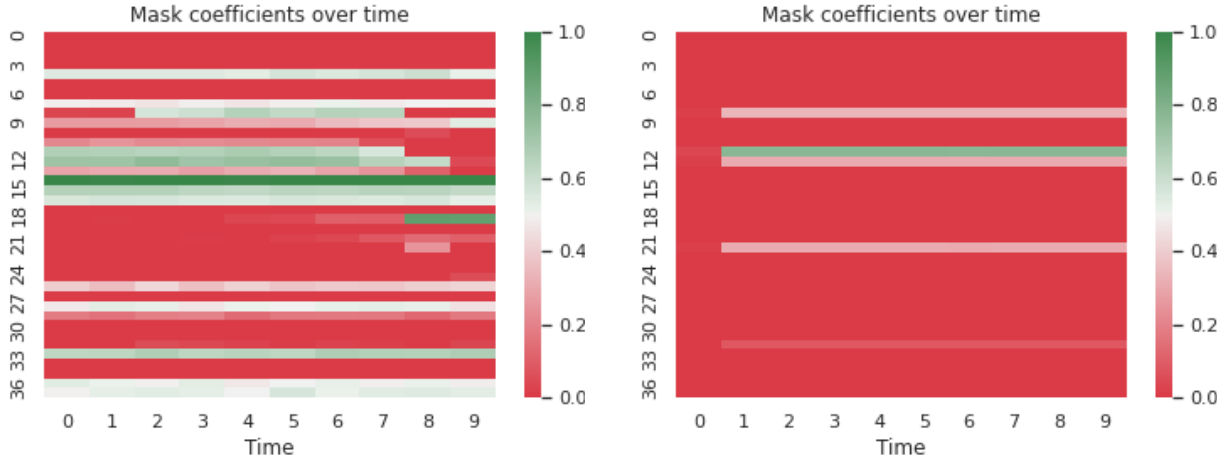
Figure 4: Anomalous windows and counterfactual ensemble explanations obtained with DPE (first row), ICE (second row) and FS (third row) on anomalies in the KPI and Yahoo data sets detected by the USAD model. The rows correspond to different anomalies. The windows include a context part of 115 timestamps and an abnormal part of 10 timestamps. The original subsequence is plotted in blue, while the explanations are in red, green or purple colors for the different variants.

(a) ICE



(b) Sparse ICE

Figure 5: Counterfactual explanation obtained with ICE (a) and the sparse variant (b). The different rows correspond respectively to the first, third, ninth and twelfth dimensions of a subsequence in the SMD dataset. Amongst them, only the fourth two (twelfth dimension) contains an anomalous observation in the last timestamp of the displayed window, detected by the NCAD model. While ICE (a) modifies all the plotted dimensions, Sparse ICE only perturbs the third and fourth (i.e., the ninth and twelfth dimension).



(a) DPE (b) Sparse DPE

Figure 6: Perturbation maps of counterfactual examples in the explanations generated by DPE (a) and its sparse variant (b) on one anomaly in the SMD dataset detected by NCAD. We recall that the rows of each mask correspond to the different dimensions of the time series and the columns to the successive timestamps in the suspect window (see Section 4). The color bars on the right sides of the maps indicate the values (between 0 and 1) of these maps along the time series features.

Dataset	Model type	Number of layers	Hidden size	training epochs	learning rate	prediction length
KPI	FFNN	1	32	100	0.001	10
Yahoo	FFNN	1	32	100	0.001	10
SMD	DeepVAR	4	40	150	0.001	10
SWaT	DeepVAR	4	40	150	0.001	10

Table 8: Hyperparameters of the Probabilistic Forecasting models used in the gradient-free approach on the four benchmark datasets.

these metrics for all explored configurations for each model and dataset. Lastly, in Tables 9, 10, 11 and 12, we report the selected configurations for respectively DPE, ICE, Sparse DPE and Sparse ICE on the benchmark datasets. Besides, the hyperparameters of the gradient-free approach can be found in Table 8.

Dataset	Perturbation	σ_{max}	learning rate	λ_2	λ_T
NCAD-KPI	Gaussian blur	3.0	0.01	0.01	0.1
NCAD-Yahoo	Gaussian blur	10.0	0.01	0.001	0.1
NCAD-SMD	Gaussian blur	20.0	0.01	0.0	1.0
NCAD-SMAP	Gaussian blur	10.0	0.01	1.0	1.0
USAD-KPI	Gaussian blur	3.0	0.01	0.001	1.0
USAD-Yahoo	Gaussian blur	10.0	0.01	0.001	1.0
USAD-SMD	Gaussian blur	20.0	0.1	0.001	0.01
USAD-SMAP	Gaussian Blur	20.0	0.01	0.01	0.1

Table 9: Hyperparameters of the DPE algorithm on the four benchmark datasets.

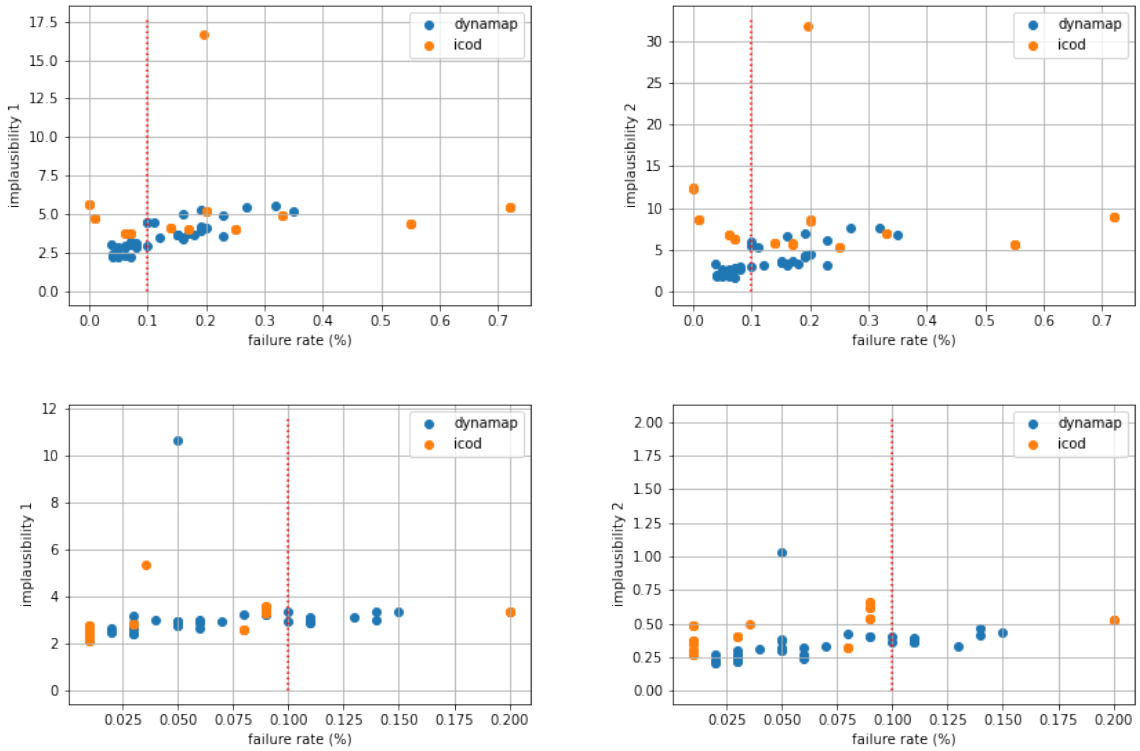


Figure 7: Implausibility measures 1 (left column) and 2 (right column) versus failures rates for different sets of hyperparameters of the ICE and DPE algorithms and their sparse variants applied to the NCAD (first row) and USAD (second row) models on a the KPI dataset. The metrics are computed over a validation set of 5 time series and the failure rate's threshold is 10% (red dotted line).

Dataset	learning rate	λ_1	λ_2	λ_T
NCAD-KPI	0.1	0.01	0.01	1.0
NCAD-Yahoo	0.1	0.01	0.01	1.0
NCAD-SMD	0.1	0.01	0.01	0.1
NCAD-SMAP	0.1	0.1	0.1	1.0
USAD-KPI	0.1	0.001	0.001	1.0
USAD-Yahoo	0.1	0.001	0.001	1.0
USAD-SMD	1000.0	0.01	0.01	1.0
USAD-SMAP	1.0	0.001	0.001	1.0

Table 10: Hyperparameters of the ICE algorithm on the four benchmark datasets.

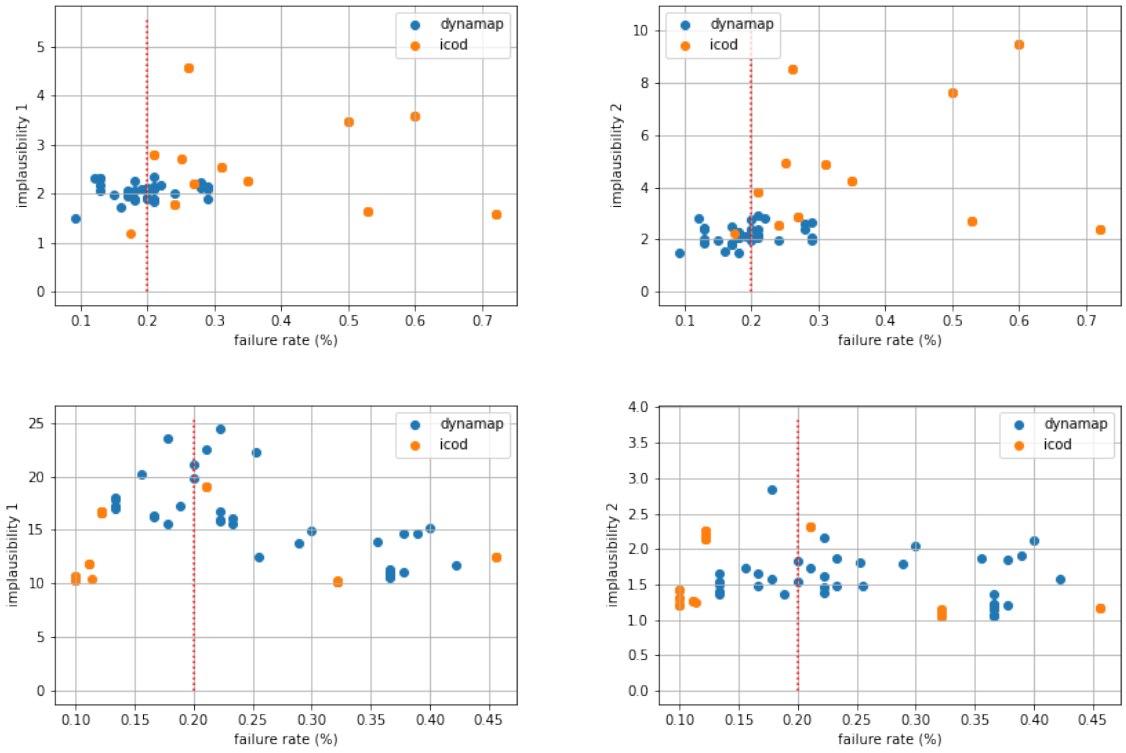


Figure 8: Implausibility measures 1 (left column) and 2 (right column) versus failures rates for different sets of hyperparameters of the ICE and DPE algorithms and their sparse variants applied to the NCAD (first row) and USAD (second row) models on a the Yahoo dataset. The metrics are computed over a validation set of 15 time series and the failure rate's threshold is 25% (red dotted line).

Dataset	Perturbation	σ_{max}	learning rate	λ_1	λ_2	λ_T
NCAD-SMD	Gaussian blur	20.0	0.1	0.01	0.01	0.1
NCAD-SMAP	Gaussian Blur	10.0	0.01	0.1	0.1	0.1
USAD-SMD	Gaussian Blur	20.0	0.01	0.01	0.01	1.0
USAD-SMAP	Gaussian Blur	20.0	0.1	0.01	0.01	0.1

Table 11: Hyperparameters of the Sparse DPE algorithm on the two benchmark multivariate datasets.

Dataset	learning rate	λ_1	λ_2	λ_T
NCAD-SMD	0.1	0.01	0.01	0.1
NCAD-SMAP	0.1	0.1	0.1	1.0
USAD-SMD	10000.0	0.01	0.01	0.1
USAD-SMAP	1.0	0.001	0.001	1.0

Table 12: Hyperparameters of the Sparse ICE algorithm on the two benchmark multivariate datasets.

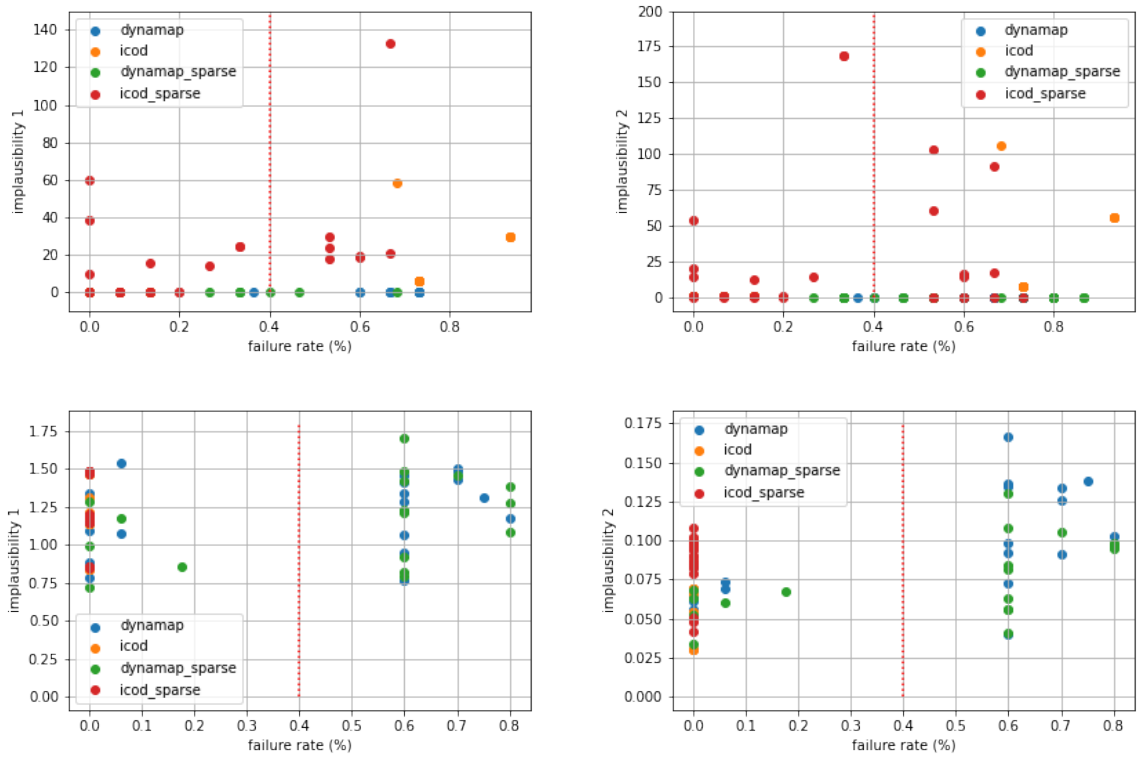


Figure 9: Implausibility measures 1 (left column) and 2 (right column) versus failures rates for different sets of hyperparameters of the ICE and DPE algorithms and their sparse variants applied to the NCAD (first row) and USAD (second row) models on a the SMAP dataset. The metrics are computed over a validation set of 40 time series and the failure rate's threshold is 25% (red dotted line).

Variant	Perturbation	σ_{max}	learning rate	λ_1	λ_2	λ_T	N
ICE	-	-	0.1	0.01	0.01	0.01	100
DPE	Gaussian Blur	3.0	0.01	-	0.1	0.01	100

Table 13: Default set of hyperparameters for our gradient-based counterfactual ensemble method.

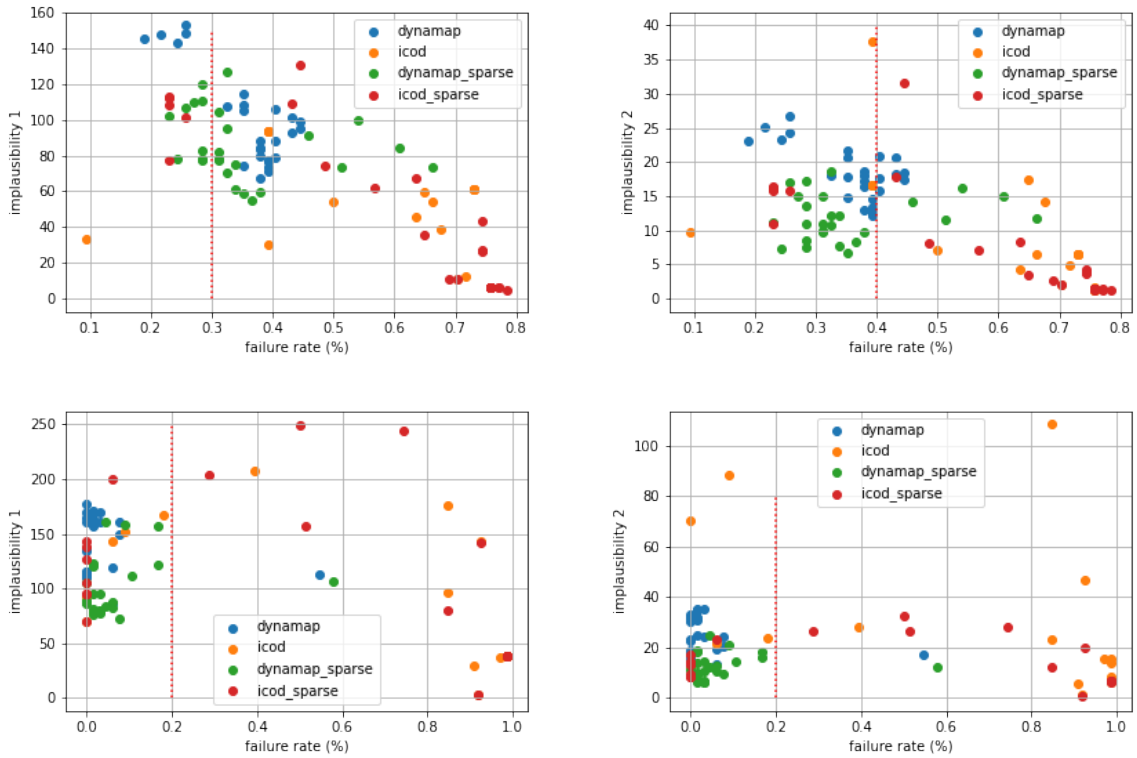


Figure 10: Implausibility measures 1 (left column) and 2 (right column) versus failures rates for different sets of hyperparameters of the ICE and DPE algorithms and their sparse variants applied to the NCAD (first row) and USAD (second row) models on a the SMD dataset. The metrics are computed over a validation set of 6 time series and the failure rate's threshold is 40% for NCAD and 20% for USAD (red dotted lines).

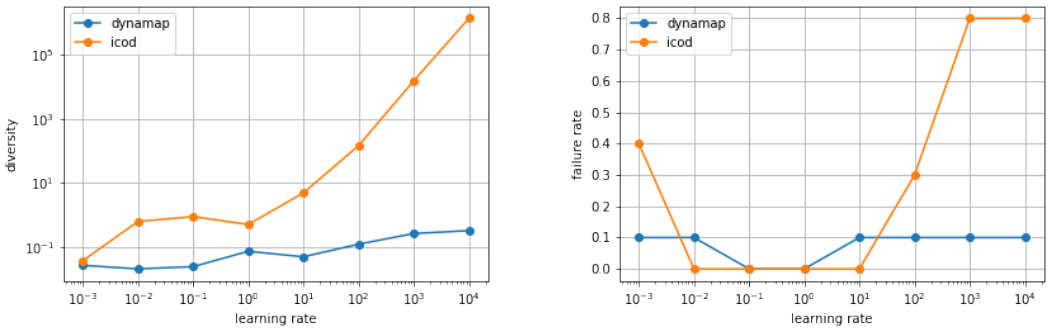


Figure 11: Diversity of the counterfactual ensemble (left) and failure rate of our counterfactual method (right) versus the learning rate of the SGD algorithm for the two variants of our method, ICE and DPE.

E Sensitivity of the Diversity criterion to the learning rate parameter

In this section we report a small-scale study of the influence of the learning rate in the SGD algorithm on the Diversity metric, in our gradient-based approach. We evaluate the latter metric on 10 anomalies detected by the NCAD model in the KPI dataset, obtained with DPE and ICE with learning rates in the set $\{0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000\}$. The other hyperparameters of our method are the same as in Section 5.5. Figure 11 shows the evolution of the Diversity score (left panel) and failure rate (right panel) when the learning rate increases. We observe that the diversity is always higher for ICE than DPE, and dramatically increases when the learning rate is greater than 1 for the former. However, failure rate also skyrockets for high learning rates.