Journal of Machine Learning Research 25 (2024) 1-26 Submitted 8/23; Revised 4/24; Published 4/24

Granger Causal Inference in Multivariate Hawkes Processes by Minimum Message Length

Kateřina Hlaváčková-Schindler KATERINA.SCHINDLEROVA@UNIVIE.AC.AT
Faculty of Computer Science, University of Vienna
Vienna, Austria, and Institute of Computer Science
Czech Academy of Sciences, Prague, Czechia

Anna Melnykova ANNA.MELNYKOVA@UNIV-AVIGNON.FR
Laboratory of Mathematics
University of Avignon, Avignon, France

Irene Tubikanec IRENE.TUBIKANEC@AAU.AT
Department of Statistics
University of Klagenfurt, Klagenfurt, Austria

Editor: Jin Tian

Abstract

Multivariate Hawkes processes (MHPs) are versatile probabilistic tools used to model various real-life phenomena: earthquakes, operations on stock markets, neuronal activity, virus propagation and many others. In this paper, we focus on MHPs with exponential decay kernels and estimate connectivity graphs, which represent the Granger causal relations between their components. We approach this inference problem by proposing an optimization criterion and model selection algorithm based on the minimum message length (MML) principle. MML compares Granger causal models using the Occam's razor principle in the following way: even when models have a comparable goodness-of-fit to the observed data, the one generating the most concise explanation of the data is preferred. While most of the state-of-art methods using lasso-type penalization tend to overfitting in scenarios with short time horizons, the proposed MML-based method achieves high F1 scores in these settings. We conduct a numerical study comparing the proposed algorithm to other related classical and state-of-art methods, where we achieve the highest F1 scores in specific sparse graph settings. We illustrate the proposed method also on G7 sovereign bond data and obtain causal connections, which are in agreement with the expert knowledge available in the literature.

Keywords: Granger causal inference, multivariate Hawkes processes, minimum message length, model selection

1. Introduction

Many practical applications deal with a large amount of irregular and asynchronous sequential data observed within a fixed time horizon. One can interpret such data as event sequences containing stereotypic events, which can be modeled via multidimensional point processes. These events can be, e.g. user viewing records, patient records in hospitals (in which times, diagnoses or treatments are provided), various levels of earthquakes, high-frequency financial transactions or neuronal activity.

In this paper, we focus on a special type of point processes, known as Hawkes processes (Hawkes, 1971). Their main advantage over other point processes (such as the classical Poisson processes) is that they permit to model the influence of past events, thanks to their "memory" property, as well as possible interactions between different components of the process. In particular, a multivariate Hawkes process (MHP) $\boldsymbol{X} = (\boldsymbol{X}_t)_{t\in[0,T]} = (X_t^1, \dots, X_t^p)_{t\in[0,T]}$ is a p-dimensional temporal point process representing a system of $p \geq 2$ interacting units. One can interpret each component of a MHP as a "particle" or "node" in some given system, e.g. a neuron in a brain, an account in a social network, or a certain type of financial transaction. Here, we consider MHPs with conditional intensity function, at each dimension $i \in \{1, \dots, p\}$ following

$$\lambda_i(t) = \mu_i + \sum_{j=1}^{p} \int_0^t \alpha_{ij} \exp(-\beta_{ij}(t-\tau)) dX_\tau^j, \tag{1}$$

where the $\mu_i > 0$ are positive parameters also known as background intensities, $\beta_{ij} > 0$ are positive decay constants, and $\alpha_{ij} \geq 0$ are non-negative influence parameters, which model the interaction between different components of the process $\boldsymbol{X}$. The conditional intensity function gives an expected number of events on each infinitely small interval of time, i.e.

$$\lambda_i(t) = \mathbb{E}[dX_t^i | \mathcal{F}_t] = \lim_{\Delta t \to 0} \mathbb{E}[X_{t+\Delta t}^i - X_t^i | \mathcal{F}_t], \tag{2}$$

where $\mathcal{F}_t$ is a filtration which contains all the information of the process prior to time t. A realization of the process corresponds to a list of event occurrence times within the time interval $[0, T]$ at which the counts are carried out. In particular, for each $i \in \{1, \dots, p\}$, the vector of observed event times of the i-th particle $(X_t^i)_{t\in[0,T]}$ is given by $\boldsymbol{x}_i := (t_1^i, \dots, t_{n_i}^i)^\top$, where $n_i \in \mathbb{N}$ and $0 < t_1^i < \cdots < t_{n_i}^i \leq T$, and a realization of the entire process $\boldsymbol{X}$ is given by $\boldsymbol{x} = \{\boldsymbol{x}_i\}_{i=1}^p$. The value T is called the time horizon of a MHP. The case when T is of order at most a hundred times the dimension p is referred to as a "short" time horizon. If T is of order at least a thousands times the dimension p, we talk about a "long" time horizon.

Since we focus on interaction functions given by an exponential kernel, in the following we will refer to $\boldsymbol{X}$ as exp-MHP. The main objective of this paper is to infer the connectivity graph, which describes the Granger-causal relationships between the components of exp-MHPs. We use the notion of Granger causality among Hawkes processes based on the definition from Eichler et al. (2017) and say that the component X^j does not *Granger-cause* X^i if and only if the corresponding interaction function is equal to 0 for all $t \in [0, T]$. Since our interaction function is given by $\alpha_{ij} \exp(-\beta_{ij}(t-\tau))$ this holds if and only if the influence parameter $\alpha_{ij} = 0$. In this case, there is no edge leading from j to i in the corresponding graph, otherwise an edge $j \to i$ is present in the graph. In other words, we study the problem of estimating the connections in a directed graph when the underlying model is an exp-MHP.

We approach this inference problem by proposing a model selection algorithm called MMLH for exp-MHPs based on the so called minimum message length (MML) principle. Methods using MML learn through a data compression perspective and are sometimes described as mathematical applications of Occam's razor, see e.g. Grünwald and Roos (2019). The minimum message length principle for statistical and inductive inference as well as machine learning was originally introduced by Wallace and Boulton (1968). It is

a formal information-theoretic restatement of Occam's razor: This means that even when models have a comparable goodness-of-fit accuracy to the observed data, the one generating the shortest overall message is more likely to be correct. In this context, a message consists of a statement of the model, followed by a statement of the data encoded concisely using that model. The MML method considers the model which compresses the data most (i.e., the one with the "shortest message length") as most descriptive for the data.

As the proposed MMLH algorithm to recover causal connections in exp-MHPs is a model selection method, it allows to incorporate possible expert knowledge about the underlying structure. For example, this may be knowledge about the maximum number m of possible causal connections to each node, such as the maximum number of debtors for every trustee in a financial connectivity graph. If such knowledge is available, the algorithm searches over a reduced set of possible structures (those indicating at most m connections), decreasing the number of parameters which have to be simultaneously estimated under a given structure. Parametric inference methods on the contrary, e.g. maximum-likelihood estimation (MLE), require all parameters to be estimated simultaneously, which can result in a poor performance. In contrast to other model selection methods, such as the classical one obtained via the Bayesian information criterion (BIC) or the recent method proposed by Jalaldoust et al. (2022) based on the data compression technique "minimum description length" (MDL), MMLH incorporates prior distributions of relevant model parameters, making the method more flexible in terms of structure-related penalty.

We compare the proposed MMLH algorithm to two state-of-the-art methods (the related MDL-based method from Jalaldoust et al. (2022) and the method ADM4 from Zhou et al. (2013)), as well as to three standard reference methods, namely BIC, AIC (Akaike information criterion), and MLE. We focus on data with short time horizons and consider graphs of dimension seven, ten, and twenty, respectively. MMLH shows the highest F1 accuracy with respect to all considered methods for specific sparse graph settings. We complete the numerical study by applying our approach on a real-world data base, which describes the return volatility of sovereign bonds of seven large economies (see e.g. Demirer et al. (2018); Jalaldoust et al. (2022)). Most discovered causal connections are in accordance with the expert knowledge from the literature.

Notations. Regarding terminology, we use both terms "causal structure" and "connectivity graph", depending on which is more appropriate in the respective context. Regarding notation, scalar variables are denoted by regular letters and vectors and matrices by bold letters. Stochastic processes are denoted by capital letters (e.g. $\boldsymbol{X}$) and any realization or point by a lower-case letter (e.g. $\boldsymbol{x}$). Matrices are denoted by Greek or capital regular letters. Let $\boldsymbol{\alpha}$ be a generic matrix. Then $\boldsymbol{\alpha}_i$ denotes the i-th row of the matrix $\boldsymbol{\alpha}$ and α_{ij} the j-th entry in the i-th row. Moreover, $\boldsymbol{\alpha}^\top$ and $|\boldsymbol{\alpha}|$ denote the transpose and determinant of the matrix $\boldsymbol{\alpha}$, respectively.

This paper is organized as follows. Section 2 discusses related work. Section 3 recalls the general idea of MML as a criterion for model selection and parameter estimation. In Section 4, we apply the MML approach to exp-MHPs. The proposed algorithm MMLH is described in Section 5. In Section 6, we illustrate the performance of the proposed algorithm in comparison to benchmark methods on both synthetic data as well as real-world data. Section 7 concludes the study and outlines perspectives for possible future work. MMLH is

coded in `R`, using the package `Rcpp` (Eddelbuettel and François (2011)). A sample code is available at: `https://github.com/IreneTubikanec/MMLH`

2. Related Work

Related work can be categorized into the work on discovery of Granger causal networks in MHPs and on applying compression based methods (such as MML and MDL) to Granger causal inference.

The problem of inferring the Granger causal structure is relatively new in the context of MHPs, however, it has attracted a lot of attention in recent years, see e.g. Hansen et al. (2015), Xu et al. (2016) and Sulem et al. (2021). Didelez (2008) studied causal connectivity graphs for discrete time events and extended them to marked point processes. Most of the related work deals with variable selection in sparse causal graphs. The recovery of the Granger causal structure is directly linked to the problem of (parametric or nonparametric) estimation of the interaction function, which is studied in the literature, e.g. by Eichler et al. (2017). The most common approach to reconstruct the network is to apply maximum likelihood estimation, see e.g. Ogata (1988), Veen and Schoenberg (2008), Juditsky et al. (2020). Maximum likelihood estimation reveals favorable theoretical properties and is not computationally expensive. However, it does not lead to good scores in practice, especially on small datasets. Some improvement can be achieved when the estimation is done via confidence intervals (as in Wang et al. (2020)), however, they are difficult to compute for a general class of models. Xu et al. (2016) applied an expectation maximization (EM) algorithm based on a penalized likelihood objective leading to temporal and group sparsity to infer a Granger graph in MHPs.

The method ADM4 in Zhou et al. (2013) performs variable selection by using lasso and nuclear norm regularization simultaneously on the parameters to cluster variables as well as to obtain a sparse connectivity graph. The method NPHC (Achab et al. 2017) takes a non-parametric approach in learning the norm of the kernel functions to find the causal connectivity graph. The method uses a moment-matching approach to fit the second-order and third-order integrated cumulants of the process.

To infer a causal connectivity graph, Bacry et al. (2020) optimize a least-square based objective function with lasso and trace norm of the interaction tensor for the intensity process. Trouleau et al. (2021) investigated stability of cumulant-based estimators for causal inference in MHPs with respect to noise. Wei et al. (2023b) recover a Granger causal graph for Hawkes processes coupled with the so-called ReLU link function; It was tested on long time horizons T and, in comparison to other mentioned methods, it considers both exciting and inhibiting effects. Idé et al. (2021) introduced a causal learning framework based on a cardinality-regularized Hawkes process. Hansen et al. (2015) use lasso penalization to infer sparse connectivity graphs in MHPs. Most of the above mentioned methods using lasso-type penalization demonstrated good performance in scenarios with long time horizons T. It is however known that lasso-type penalization methods often suffer from overfitting in the opposite case of short time horizons, see e.g. Reid et al. (2016). To overcome the drawbacks of these methods, we approach penalization based on the MML principle.

MML-based model selection as an inductive inference method based on data compression was first introduced in Wallace and Boulton (1968). Intuitively, the recovery of a connectivity graph using the MML principle is equivalent to selecting an optimal model

for the observed data, where "optimal" means "the one which permits to encode the data in a binary string of the shortest length" in terms of coding theory. There exist papers on the recovery of Granger connectivity graphs by MML for processes having distributions from exponential families, see Hlaváčková-Schindler and Plant (2020a,b), but to the best of our knowledge, not for MHPs. Another compression scheme using Occam's razor in terms of coding representations is the minimum description length (Rissanen (1998)), which was more recently developed in Grünwald (2007) and Grünwald and Roos (2019). In comparison to MML, the MDL principle does not use any knowledge of priors. MDL-based Granger-causal inference has been recently applied to exp-MHPs in Jalaldoust et al. (2022) and to Gaussian processes in Hlaváčková-Schindler and Plant (2020b).

3. Minimum Message Length Criterion and Its Approximation

Methods based on the MML principle consider the model which compresses the data the most (i.e., the one with the "shortest message length"). To be able to decompress this representation of the data, the details of the statistical model used to encode the data must also be a part of the compressed data string. The calculation of the exact message is an NP-hard problem, since it corresponds to the Kolmogorov complexity (see Wallace and Dowe (1999)), which is in general not computable due to the halting problem (see Li and Vitányi (2008)). However, there exist computable approximations of MML, the most used one is the Wallace–Freeman approximation (Wallace and Freeman, 1987), which we will use in this paper (see Section 3.2).

Before we recall the general idea behind MML and outline the aforementioned approximation approach, we define statistical models. Statistical models are families of probability distributions of the form

$$M = \{p(\cdot|\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}, \tag{3}$$

parametrized by a set $\boldsymbol{\Theta}$ (usually a subset of a Euclidean space). They are represented by families of probability distributions

$$\{M_{\boldsymbol{\gamma}} : \boldsymbol{\gamma} \in \boldsymbol{\Gamma}\}, \tag{4}$$

where $\boldsymbol{\Gamma}$ is a countable set of so-called "structures" and, for each structure $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}$,

$$M_{\boldsymbol{\gamma}} = \{p_{\boldsymbol{\gamma}}(\cdot|\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}_{\boldsymbol{\gamma}}\} \tag{5}$$

is a statistical model, parameterized by the space $\boldsymbol{\Theta}_{\boldsymbol{\gamma}}$.

In our setting, the set of structures $\boldsymbol{\Gamma}$ can be interpreted as a countable set of binary vectors, i.e. $\boldsymbol{\Gamma} = \{0,1\}^q$ with $q > 0$. Each element $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}$ is then a q-dimensional vector of zeros and ones, where the number of ones is given by k, and where $\gamma_j = 1$ denotes the presence of the j-th variable in the subset of k variables, and $\gamma_j = 0$ means that the j-th variable is not present. The parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{\boldsymbol{\gamma}} \subset \mathbb{R}^k$ then define the "weights" (which can be interpreted as importance measures) with respect to the variables in $\boldsymbol{\gamma}$. In the graph context, a structure set $\boldsymbol{\Gamma} = \{0,1\}^q$ corresponds to a given node, which can have at most q causes. An element γ_j, $j \in \{1,\dots,q\}$, of a structure $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}$ indicates the presence ($\gamma_j = 1$) or absence ($\gamma_j = 0$) of an incoming edge from node j into the given node. For example, for a graph with $q = 3$ nodes, the corresponding structure set for each node is given by

$\mathbf{\Gamma} = \{0,1\}^3 = \{(0,0,0),(1,0,0),(0,1,0),(0,0,1),(1,1,0),(1,0,1),(0,1,1),(1,1,1)\}$. If $\mathbf{\Gamma}$ corresponds to node 1, then the element $\boldsymbol{\gamma} = (0,1,0) \in \mathbf{\Gamma}$ indicates that node 1 is not self-excitatory (since $\gamma_1 = 0$), has an incoming edge from node 2 (since $\gamma_2 = 1$), and has no incoming edge from node 3 (since $\gamma_3 = 0$).

3.1 Idea Behind the Minimum Message Length Method

The MML principle is a formal information theory restatement of Occam's razor: even when models have a comparable goodness-of-fit to the observed data, the one generating the shortest overall message is more likely to be correct (where the message consists of a statement of the model, followed by a statement of data encoded concisely using that model). Let us describe the idea of the MML method more formally.

Consider some data $\boldsymbol{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ that we would like to send to a receiver by encoding it into a message (e.g. a binary string). The key idea in MML inference is to interpret this message as consisting of the following parts: an encoding (called *assertion*) of the model structure $\boldsymbol{\gamma} \in \mathbf{\Gamma}$ and associated parameters $\boldsymbol{\theta} \in \mathbf{\Theta}_{\boldsymbol{\gamma}}$, a description (called *detail*) of the data $\boldsymbol{y}$ using the model $p_{\boldsymbol{\gamma}}(\boldsymbol{y}|\boldsymbol{\theta})$ specified in the assertion, and a preamble code describing which structure is used. The total message length of the data $\boldsymbol{y}$, model structure $\boldsymbol{\gamma} \in \mathbf{\Gamma}$, and parameterization $\boldsymbol{\theta} \in \mathbf{\Theta}_{\boldsymbol{\gamma}}$ is then given by

$$I(\boldsymbol{y}; \boldsymbol{\theta}; \boldsymbol{\gamma}) = I(\boldsymbol{\theta}; \boldsymbol{\gamma}) + I(\boldsymbol{y}|\boldsymbol{\theta}; \boldsymbol{\gamma}) + I(\boldsymbol{\gamma}), \tag{6}$$

where $I(\boldsymbol{\theta}; \boldsymbol{\gamma})$, $I(\boldsymbol{y}|\boldsymbol{\theta}; \boldsymbol{\gamma})$, and $I(\boldsymbol{\gamma})$ denote the length of the assertion, detail, and structure preamble code, respectively. Equation (6) is also called refined total message length in the literature.

The length of the assertion $I(\boldsymbol{\theta}; \boldsymbol{\gamma})$ is a measure of the model complexity, while the length of the detail $I(\boldsymbol{y}|\boldsymbol{\theta}; \boldsymbol{\gamma})$ is a measure of the goodness-of-fit of the model to the data (model capability). Moreover, for $\gamma \in \mathbf{\Gamma} = \{0,1\}^q$, the set of all possible structures, we set

$$I(\boldsymbol{\gamma}) = \log \binom{q}{k} + \log(q+1), \tag{7}$$

as recommended by Roos et al. (2009). MML seeks the model structure and corresponding parameters that minimize this trade off between model complexity and model capability, i.e.

$$\{\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\theta}}\} = \operatorname{argmin}_{\boldsymbol{\gamma} \in \mathbf{\Gamma}, \boldsymbol{\theta} \in \mathbf{\Theta}_{\boldsymbol{\gamma}}} I(\boldsymbol{y}; \boldsymbol{\theta}; \boldsymbol{\gamma}). \tag{8}$$

3.2 Wallace-Freeman Approximation

In the following, we recall a well-known approximation of the total message length introduced in (6), originally proposed by Wallace and Freeman (1987). A detailed presentation of this approximation can be also found in Chapter 5 of Wallace (2005), including a list of required assumptions given in Section 5.1.1. Similar to Bayesian selection methods, this procedure utilizes prior probability distributions for parameters.

According to the Wallace-Freeman approximation, the codelength of data $\boldsymbol{y}$ for a given model parametrization $\boldsymbol{\theta} \in \mathbf{\Theta}_{\boldsymbol{\gamma}}$ under a fixed structure $\gamma \in \mathbf{\Gamma} = \{0,1\}^q$ (i.e., the detail w.r.t. γ) is given by

$$I(\boldsymbol{y}|\boldsymbol{\theta}; \boldsymbol{\gamma}) \approx -\log p_{\boldsymbol{\gamma}}(\boldsymbol{y}|\boldsymbol{\theta}) + \frac{k}{2}. \tag{9}$$

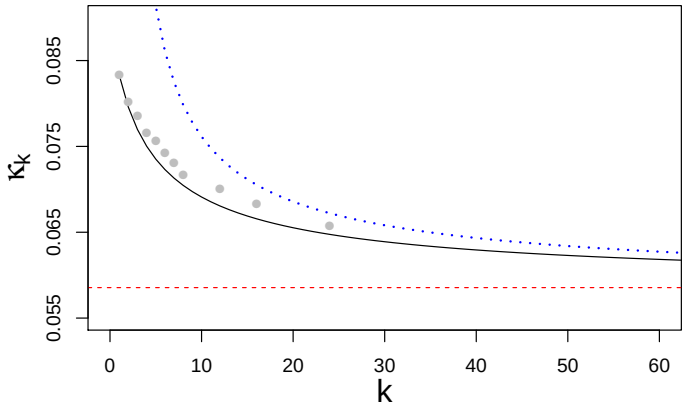

Figure 1: Blue dotted line: Upper bound for κ_k. Black solid line: Lower bound for κ_k. Red dashed line: Limit of the bounds for $k \to \infty$. Grey dots: Known values of κ_k.

Moreover, the codelength of the assertion w.r.t. γ is given by

$$I(\boldsymbol{\theta};\gamma) \approx -\log \pi_{\boldsymbol{\gamma}}(\boldsymbol{\theta}) + \frac{1}{2}\log|J_{\boldsymbol{\gamma}}(\boldsymbol{\theta})| + \frac{k}{2}\log \kappa_k, \tag{10}$$

where $\pi_{\boldsymbol{\gamma}}(\boldsymbol{\theta})$ is a prior probability distribution over $\boldsymbol{\Theta}_\gamma$, $J_\gamma(\boldsymbol{\theta})$ is the expected Fisher information matrix and κ_k is a quantizing lattice constant, which depends on the number of parameters k that is determined by the structure γ.

While an optimal value for κ_k is not available in general, in Wallace and Freeman (1987) the following upper and lower bounds were proposed for $k > 1$:

$$\frac{\Gamma(k/2+1)^{2/k}}{\pi(k+2)} < \kappa_k < \frac{\Gamma(k/2+1)^{2/k}\Gamma(2/k+1)}{\pi k}, \tag{11}$$

where in this case Γ denotes the gamma function. These bounds are reported as function of k (black solid and blue dotted lines) in Figure 1. They both converge to $1/(2\pi e)$ (red dashed line) for $k \to \infty$. Some values of κ_k (for small k) are known explicitly, see e.g. Conway and Sloane (1984), Makalic and Schmidt (2021). Those reported in Table 1 of Conway and Sloane (1984) are added as gray dots to Figure 1. For $k = 1$, it is known that $\kappa_k = 1/12$, and thus the lower bound in (11) is achieved, see Makalic and Schmidt (2021). The choice of approximation for κ_k influences the penalty term with respect to the number of parameters k determined by the structure γ. Following again Wallace and Freeman (1987) and Wallace (2005), pp. 257–258, we focus on the approximation

$$\frac{k}{2}(\log \kappa_k + 1) \approx -\frac{k}{2}\log(2\pi) + \frac{1}{2}\log(k\pi) + \psi(1), \tag{12}$$

where ψ denotes the digamma function and $\psi(1) \approx -0.5772$.

Using (7) and the approximations (9), (10) and (12), the total message length (6) can be approximated as follows:

$$\begin{aligned} I(\boldsymbol{y};\boldsymbol{\theta};\boldsymbol{\gamma}) \approx & -\log p_{\boldsymbol{\gamma}}(\boldsymbol{y}|\boldsymbol{\theta}) - \log \pi_{\boldsymbol{\gamma}}(\boldsymbol{\theta}) + \frac{1}{2}\log|J_{\boldsymbol{\gamma}}(\boldsymbol{\theta})| \\ & - \frac{k}{2}\log(2\pi) + \frac{1}{2}\log(k\pi) + \psi(1) \\ & + \log\binom{q}{k} + \log(q+1). \end{aligned} \tag{13}$$

4. Granger Causal Structure Recovery in Multivariate Hawkes Processes by Minimum Message Length

In this section, we present the proposed MML-based procedure for causal inference in exp-MHPs defined via intensity (1). First, we define the corresponding parameter space $\boldsymbol{\Theta}$. Second, we introduce the components required to define the total message length (13) for exp-MHPs over the parameter space $\boldsymbol{\Theta}$. In particular, we report an explicit expression of the the log-likelihood, derive an approximation for the Fisher information matrix, and choose appropriate prior distributions. Finally, we introduce suitable structures, include them in the aforementioned expressions, and propose a criterion for the total message length in exp-MHPs.

4.1 Parameter Space for Exp-MHPs

Consider an exp-MHP $\boldsymbol{X}$, i.e. a MHP defined via intensity (1). Throughout, we assume for all $i, j \in \{1, \dots, p\}$ that the decay constants β_{ij} are known. This is a common practice in the literature (see, e.g. Juditsky et al. (2020), Wang et al. (2020), Jalaldoust et al. (2022)), since these constants are considered to be part of the model itself (such as the memory kernel which is here exponential). Moreover, we assume that the background intensities μ_i of the i-th particle X^i and the influence vector $\boldsymbol{\alpha}_i = (\alpha_{i1}, \dots, \alpha_{ip})^\top$ on X^i are not known. Considering the entire process $\boldsymbol{X}$, we also introduce the unknown *baseline vector* $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)^\top$ and *influence matrix* $\boldsymbol{\alpha}$, whose i-th row corresponds to the influence vector $\boldsymbol{\alpha}_i$. Recall that by the definition of Granger causality an entry α_{ij} is non-zero if and only if there is an incoming edge from node j to node i. The parameter vector of $\boldsymbol{X}$ is then defined as

$$\boldsymbol{\theta} = [\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top, \dots, \boldsymbol{\theta}_p^\top]^\top \in \boldsymbol{\Theta} = (\mathbb{R}_0^+)^{p+p^2}, \tag{14}$$

where

$$\boldsymbol{\theta}_i = (\mu_i, \boldsymbol{\alpha}_i^\top)^\top \in \boldsymbol{\Theta}_i = (\mathbb{R}_0^+)^{p+1} \tag{15}$$

is the parameter vector of the i-th component X^i, for $i \in \{1, \dots, p\}$.

4.2 Log-Likelihood for Exp-MHPs

In the following, we recall the log-likelihood of an exp-MHP, see, e.g. Ozaki (1979) (univariate case) and Shlomovich et al. (2022) (multivariate case). Consider an observation $\boldsymbol{x}$ of an exp-MHP $\boldsymbol{X}$ and the parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ (14). Then, the log-likelihood can be

decomposed as

$$\log p(\boldsymbol{x}|\boldsymbol{\theta}) = -\sum_{i=1}^{p}\left(\int_0^T \lambda_i(s)ds - \sum_{j=0}^{n_i}\log\lambda_i(t_j^i)\right). \tag{16}$$

Since each summand of this function depends only on the i-th dimension $\boldsymbol{\theta}_i$ (15) of the parameter vector $\boldsymbol{\theta}$ (14), the negative log-likelihood function can be written as the sum:

$$-\log p(\boldsymbol{x}|\boldsymbol{\theta}) = -\sum_{i=1}^{p}\log p_i(\boldsymbol{x}|\boldsymbol{\theta}_i), \tag{17}$$

where each summand represents the marginal negative log-likelihood of the corresponding node. To ease the notations, define $l(\boldsymbol{x}|\boldsymbol{\theta}) := -\log p(\boldsymbol{x}|\boldsymbol{\theta})$ and $l_i(\boldsymbol{x}|\boldsymbol{\theta}_i) := -\log p_i(\boldsymbol{x}|\boldsymbol{\theta}_i)$. The explicit expression for each $l_i(\boldsymbol{x}|\boldsymbol{\theta}_i)$ can be derived using (1) and is given by

$$\begin{aligned} l_i(\boldsymbol{x}|\boldsymbol{\theta}_i) = \mu_i t^{\max} + \sum_{j=1}^{p}\frac{\alpha_{ij}}{\beta_{ij}}\sum_{k=1}^{n_j}\left[1-\exp(-\beta_{ij}(t^{\max}-t_k^j))\right] \\ -\sum_{l=1}^{n_i}\log\left[\mu_i + \sum_{j=1}^{p}\alpha_{ij}\sum_{k:t_k^j<t_l^i}\exp(-\beta_{ij}(t_l^i-t_k^j))\right], \end{aligned} \tag{18}$$

where $t^{\max} \leq T$ is the largest jump time recorded over all nodes.

Remark 1 *The function $l_i(\boldsymbol{x}|\boldsymbol{\theta}_i)$ from* (18) *is convex in $\boldsymbol{\alpha}_i$ and μ_i. Thus, the maximum likelihood estimate (MLE) can be computed using convex optimization. The proof of convexity can be found, e.g. in Ogata (1981).*

4.3 Hessian Matrix for Exp-MHPs

In this section, we derive an explicit expression for the Hessian matrix $H(\boldsymbol{\theta})$ of the negative log-likelihood $l(\boldsymbol{x}|\boldsymbol{\theta})$ based on formula (18), see Shlomovich et al. (2022). Note that the Fisher information matrix J, required in the total message length criterion (13), is defined as the expected Hessian. In general, this expectation is difficult to compute and it is often replaced by the observed Fisher information. The observed Fisher information, in turn, is given by the Hessian, evaluated at an estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$.

To obtain the Hessian, we first need to compute the gradient of the negative log-likelihood $l(\boldsymbol{x}|\boldsymbol{\theta})$. The required first-order derivatives can be computed explicitly and are given by

$$\begin{aligned} \frac{\partial l(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \mu_i} &= t^{\max} - \sum_{l=1}^{n_i}\frac{1}{\mu_i + \sum_{k=1}^{p}\alpha_{ik}A_{ik}(t_l^i)}, \\ \frac{\partial l(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \alpha_{ij}} &= \frac{1}{\beta_{ij}}\sum_{k=1}^{n_j}\left[1-\exp\left(-\beta_{ij}(t^{\max}-t_k^j)\right)\right] - \sum_{l=1}^{n_i}\frac{A_{ij}(t_l^i)}{\mu_i + \sum_{k=1}^{p}\alpha_{ik}A_{ik}(t_l^i)}, \end{aligned} \tag{19}$$

where

$$A_{ij}(t) = \sum_{k:t_k^j<t} \exp\left(-\beta_{ij}(t - t_k^j)\right), \quad \text{for } t \in [0, T]. \tag{20}$$

Intuitively, the term $A_{ij}(t)$ summarizes the "weighted" history of events on node j up to time t: since the β_{ij} are positive, "new" events will always have more importance than the "old" ones.

Furthermore, the second order partial derivatives w.r.t. the parameters μ_i and α_{ij} are given by

$$\begin{aligned}
\frac{\partial^2 l(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \mu_i^2} &= \sum_{r=1}^{n_i} \frac{1}{\left(\mu_i + \sum_{k=1}^p \alpha_{ik} A_{ik}(t_r^i)\right)^2}, \\
\frac{\partial^2 l(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \mu_i \partial \alpha_{ij}} &= \sum_{r=1}^{n_i} \frac{A_{ij}(t_r^i)}{\left(\mu_i + \sum_{k=1}^p \alpha_{ik} A_{ik}(t_r^i)\right)^2}, \\
\frac{\partial^2 l(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \alpha_{ij} \partial \alpha_{ij'}} &= \sum_{r=1}^{n_i} \frac{A_{ij}(t_r^i) A_{ij'}(t_r^i)}{\left(\mu_i + \sum_{k=1}^p \alpha_{ik} A_{ik}(t_r^i)\right)^2}.
\end{aligned} \tag{21}$$

Note that all derivatives $\frac{\partial^2 l(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \mu_i \partial \alpha_{i'j}}, \frac{\partial^2 l(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \alpha_{ij} \partial \alpha_{i'j'}}$ for $i' \neq i$ are equal to 0. Therefore, the Hessian of $l(\boldsymbol{x}|\boldsymbol{\theta})$ can be written as a block-diagonal matrix of the form

$$H(\boldsymbol{\theta}) = \begin{pmatrix} H_1(\boldsymbol{\theta}_1) & & \\ & \ddots & \\ & & H_p(\boldsymbol{\theta}_p) \end{pmatrix}, \tag{22}$$

where, for each $i \in \{1, \dots, p\}$, $H_i(\boldsymbol{\theta}_i)$ is given by the $(p+1) \times (p+1)$-dimensional matrix

$$H_i(\boldsymbol{\theta}_i) = \begin{pmatrix} \frac{\partial^2 l(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \mu_i^2} & \frac{\partial^2 l(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \mu_i \partial \alpha_{i1}} & \cdots & \frac{\partial^2 l(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \mu_i \partial \alpha_{ip}} \\ \frac{\partial^2 l(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \alpha_{i1} \partial \mu_i} & \frac{\partial^2 l(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \alpha_{i1}^2} & \cdots & \frac{\partial^2 l(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \alpha_{i1} \partial \alpha_{ip}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 l(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \alpha_{ip} \partial \mu_i} & \frac{\partial^2 l(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \alpha_{ip} \partial \alpha_{i1}} & \cdots & \frac{\partial^2 l(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \alpha_{ip}^2} \end{pmatrix}, \tag{23}$$

with entries as in (21). Since the determinant of a block-diagonal matrix is equal to the product of the determinants of the diagonal blocks, we have that

$$\log |H(\boldsymbol{\theta})| = \sum_{i=1}^p \log |H_i(\boldsymbol{\theta}_i)|. \tag{24}$$

Note that, in the case when the vector of intensities $\boldsymbol{\mu}$ is known, it is possible to use a more computationally efficient approximation of the Hessian, see Wang et al. (2020). We consider the intensities $\boldsymbol{\mu}$ to be unknown, thus we will rely on the analytical expression for the Hessian, with entries defined via (21).

4.4 Choice of Priors for Exp-MHPs

In this section, we define two possible prior distributions $\pi(\boldsymbol{\theta})$ for the parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ (14), which is now considered to be a random quantity.

First, we assume that two parameter vectors $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ as in (15), corresponding to different nodes $i \neq j$, are independent. Therefore, the negative log-prior function can be expressed as

$$-\log \pi(\boldsymbol{\theta}) = -\sum_{i=1}^{p} \log \pi_i(\boldsymbol{\theta}_i), \tag{25}$$

where $\pi_i(\boldsymbol{\theta}_i)$ is a prior distribution for the parameter vector $\boldsymbol{\theta}_i \in \boldsymbol{\Theta}_i$ (15), corresponding to the i-th node.

Throughout, we further assume that μ_i and all entries α_{ij} of a parameter vector $\boldsymbol{\theta}_i$ (15) are independent and identically distributed (iid), yielding

$$\pi_i(\boldsymbol{\theta}_i) = \pi(\mu_i) \prod_{j=1}^{p} \pi(\alpha_{ij}). \tag{26}$$

In the following, we consider two different prior distributions for the entries μ_i and α_{ij} of a parameter vector $\boldsymbol{\theta}_i \in (\mathbb{R}_0^+)^{p+1}$ as in (15), which both allow to incorporate the prior knowledge that the parameters to be estimated are all non-negative: the uniform distribution $\mathrm{U}[0, b]$, with $b > 0$, and the exponential distribution $\mathrm{Exp}(c)$, with $c > 0$, having support $[0, \infty)$.

Uniform prior. Assuming that μ_i and the α_{ij} are iid as $U[0, b]$, $b > 0$, the prior (26) for the i-th node becomes

$$\pi_i(\boldsymbol{\theta}_i) = \prod_{j=1}^{p+1} \frac{1}{b} = \frac{1}{b^{p+1}}. \tag{27}$$

Thus, the negative log-prior for the i-th node is given by

$$-\log \pi_i(\boldsymbol{\theta}_i) = (p+1)\log(b). \tag{28}$$

Note that, to obtain a flat prior with a non-restrictive domain, one may consider a large value for the hyperparameter b (see e.g. Oliver et al. (1996) for the use of uniform priors in the context of MML).

Exponential prior. Assuming that μ_i and the α_{ij} are iid as $\mathrm{Exp}(c)$, $c > 0$, the prior (26) for the i-th node becomes

$$\pi_i(\boldsymbol{\theta}_i) = c\exp(-c\mu_i) \prod_{j=1}^{p} c\exp(-c\alpha_{ij}) = c^{p+1} \exp\left(-c\mu_i - c\sum_{j=1}^{p} \alpha_{ij}\right). \tag{29}$$

Thus, the negative log-prior for the i-th node is given by

$$-\log \pi_i(\boldsymbol{\theta}_i) = c\mu_i + c\sum_{j=1}^{p} \alpha_{ij} - (p+1)\log(c). \tag{30}$$

In this case, to obtain a flat prior, one may choose a small value for the hyperparameter c.

Remark 2 *The negative log-priors in* (28) *are constant in* $\boldsymbol{\theta}_i$ *and those in* (30) *are linear in* $\boldsymbol{\theta}_i$*. Thus, they are convex in both cases.*

4.5 MML Criterion for Granger Causal Inference in Exp-MHPs

From formulas (17) and (25), it becomes evident that the optimization of the function $\log p(\boldsymbol{x}|\boldsymbol{\theta}) + \log \pi(\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ can be done independently for each node i. Therefore, to perform causal inference in exp-MHPs, for each $i \in \{1, \dots, p\}$, we introduce a structure set $\boldsymbol{\Gamma}_i = \{0,1\}^p$, whose elements $\boldsymbol{\gamma}_i = (\gamma_{i1}, \dots, \gamma_{ip})^\top$ are p-dimensional vectors of zeros and ones with $k_i > 0$ corresponding to the number of ones. It holds then that $\gamma_{ij} = 1$ if and only if events in the j-th node Granger-cause events in the i-th node, and $\gamma_{ij} = 0$ if and only if $\alpha_{ij} = 0$, i.e. there is no impact of node j on node i. This means that causal discovery in exp-MHPs is equivalent to identifying the sparsity pattern in the influence vector $\boldsymbol{\alpha}_i$, for each $i \in \{1, \dots, p\}$.

According to the definition of Granger-causality in exp-MHPs, for a given structure $\boldsymbol{\gamma}_i \in \boldsymbol{\Gamma}_i = \{0,1\}^p$, the corresponding restricted parameter space $\boldsymbol{\Theta}_{\boldsymbol{\gamma}_i}$ contains parameter vectors representing μ_i and $\boldsymbol{\alpha}_i$ such that α_{ij} is present in $\boldsymbol{\alpha}_i$ if and only if $\gamma_{ij} = 1$. Thus, for any $\boldsymbol{\gamma}_i$ containing k_i non-zero entries, the vector $\boldsymbol{\alpha}_i$ has k_i non-zero entries to be estimated. Moreover, the baseline intensity μ_i, which has no influence on causal discovery, has to be estimated as well. Hence, under a given structure $\boldsymbol{\gamma}_i$, a total of $k_i + 1$ non-negative parameters are to be estimated and the restricted parameter space $\boldsymbol{\Theta}_{\boldsymbol{\gamma}_i} = (\mathbb{R}_0^+)^{k_i+1}$.

Recall that the formulas (18), (23), (28) and (30) are formulated for $\boldsymbol{\theta}_i \in \boldsymbol{\Theta}_i = (\mathbb{R}_0^+)^{p+1}$ from (15). For a given structure $\boldsymbol{\gamma}_i \in \boldsymbol{\Gamma}_i = \{0,1\}^p$ and parameter vector $\boldsymbol{\theta}_i \in \boldsymbol{\Theta}_{\boldsymbol{\gamma}_i} = (\mathbb{R}_0^+)^{k_i+1}$ with $\boldsymbol{\alpha}_i$ having $k_i \leq p$ entries, the log-likelihood $\log p_{\boldsymbol{\gamma}_i}(\boldsymbol{x}|\boldsymbol{\theta}_i)$, the Hessian $H_{\boldsymbol{\gamma}_i}(\boldsymbol{\theta}_i)$ and log-priors $\log \pi_{\boldsymbol{\gamma}_i}(\boldsymbol{\theta}_i)$ are obtained from the aforementioned formulas, replacing p by k_i and adjusting all indices properly.

Finally, for each node $i \in \{1, \dots, p\}$, we can now propose the following refined message length criterion for causal inference in exp-MHPs:

$$\begin{aligned} I(\boldsymbol{x}; \boldsymbol{\theta}_i; \boldsymbol{\gamma}_i) = &- \log p_{\boldsymbol{\gamma}_i}(\boldsymbol{x}|\boldsymbol{\theta}_i) - \log \pi_{\boldsymbol{\gamma}_i}(\boldsymbol{\theta}_i) + \frac{1}{2} \log |H_{\boldsymbol{\gamma}_i}(\hat{\boldsymbol{\theta}}_i)| \\ &- \frac{k_i}{2} \log(2\pi) + \frac{1}{2} \log(k_i \pi) + \psi(1) \\ &+ \log \binom{p}{k_i} + \log(p+1), \end{aligned} \tag{31}$$

where, for a given structure $\boldsymbol{\gamma}_i \in \boldsymbol{\Gamma}_i = \{0,1\}^p$, the Hessian matrix is evaluated at the estimate $\hat{\boldsymbol{\theta}}_i$ given by

$$\hat{\boldsymbol{\theta}}_i = \operatorname{argmin}_{\boldsymbol{\theta}_i \in \Theta_{\gamma_i}} \left(- \log p_{\boldsymbol{\gamma}_i}(\boldsymbol{x}|\boldsymbol{\theta}_i) - \log \pi_{\boldsymbol{\gamma}_i}(\boldsymbol{\theta}_i) \right).$$

Remark 3 *i) Note that under the uniform prior* (28) *the estimate* $\hat{\boldsymbol{\theta}}_i$ *coincides with the MLE. ii) One may either remove the* p*-dimensional zero vector from the structure set* $\boldsymbol{\Gamma}_i$ *or allow for* $k_i = 0$ *(node* i *does not receive any input connections) by setting the term* $k_i(\log \kappa_{k_i} + 1)/2$ *(cf. formula* (12)*) to zero in that case. Criterion* (31) *for* $k_i = 0$ *then results into the form, where the second line is not present.*

5. Algorithm MMLH for Granger Causal Inference in Exp-MHPs and its Complexity

The form of the MML criterion (31) leads to the following algorithm, denoted as MMLH, which we propose for causal inference in exp-MHPs.

Algorithm 1: MMLH: Causal inference in exp-MHPs by MML

Input : Dimension p, data $\mathbf{x}$
Output: Estimate $\hat{\boldsymbol{\gamma}} = [\hat{\boldsymbol{\gamma}}_1^\top, \ldots, \hat{\boldsymbol{\gamma}}_p^\top]^\top \in \boldsymbol{\Gamma} := \boldsymbol{\Gamma}_1 \times \ldots \times \boldsymbol{\Gamma}_p$

1 **for** *each* $i \in \{1, \ldots, p\}$ **do**
2 **for** *each* $\boldsymbol{\gamma}_i \in \boldsymbol{\Gamma}_i = \{0,1\}^p$ **do**
3 $\hat{\boldsymbol{\theta}}_i \longleftarrow \operatorname{argmin}_{\boldsymbol{\theta}_i \in \boldsymbol{\Theta}_{\boldsymbol{\gamma}_i}} \big(-\log p_{\boldsymbol{\gamma}_i}(\mathbf{x}|\boldsymbol{\theta}_i) - \log \pi_{\boldsymbol{\gamma}_i}(\boldsymbol{\theta}_i)\big)$;
4 $\hat{c}_{\boldsymbol{\gamma}_i} \longleftarrow I(\mathbf{x}; \hat{\boldsymbol{\theta}}_i; \boldsymbol{\gamma}_i)$ (eq. 31);
5 **end**
6 $\hat{\boldsymbol{\gamma}}_i \longleftarrow \operatorname{argmin}_{\boldsymbol{\gamma}_i \in \boldsymbol{\Gamma}_i} \hat{c}_{\boldsymbol{\gamma}_i}$;
7 **end**
8 **return** $\hat{\boldsymbol{\gamma}} = [\hat{\boldsymbol{\gamma}}_1^\top, \ldots, \hat{\boldsymbol{\gamma}}_p^\top]^\top$;

Remark 4 *Recall that the decay constants β_{ij} are assumed to be known (see Section 4.1). However, adding more unknown parameters would be possible and would require small modifications in Algorithm 8. The estimation procedure in lines 3 and 4 would have to be adjusted accordingly, by extending the parameter space and setting priors for the β_{ij}. Moreover, the Hessian matrix would include double derivatives with respect to β_{ij}, which are available as closed-form expressions in the literature (see Ozaki (1979)).*

We now address the computational complexity of the proposed method. Algorithm 1 consists of p optimizations, each requiring 2^p evaluations of the MML criterion (31). Each such evaluation relies on a parameter learning procedure for the observed data $\boldsymbol{x}$ (line 3 of Algorithm 8) and a function evaluation (line 4 of Algorithm 8), which includes a computation of the corresponding Hessian matrix and its determinant.

The computational complexity of a parameter learning procedure (line 3) depends on the number of parameters to be estimated (for a given structure $\boldsymbol{\gamma}_i \in \boldsymbol{\Gamma}_i = \{0,1\}^p$ with k_i non-zero entries, $k_i + 1$ parameters have to be estimated and $k_i \leq p$) and the size of $\boldsymbol{x}$, i.e. the number of observed events (which in turn depends on T). In our R-implementation, we apply the Nelder-Mead search method (NM), which relies on a user-defined termination. There is no convergence theory providing an estimate for the number of iterations required to satisfy a reasonable accuracy constraint given in the termination test, see Singer and Singer (1999). Thus, to the best of our knowledge, an upper bound on the complexity of the NM method with a fixed termination rule is not available in the literature.

The complexity of computing the determinant of the symmetric Hessian matrix (line 4) also depends on the "size" of $\boldsymbol{\gamma}_i$, the dimension of $H_{\boldsymbol{\gamma}_i}$ being $(k_i + 1) \times (k_i + 1)$. In our R-implementation, we use a LU decomposition procedure, typically having a complexity of order $(k_i + 1)^3$ (note that Aho and Hopcroft (1974) showed, however, that the exponent 3 may be reduced to 2.373 via a fast matrix multiplication method).

In scenarios where the expert knowledge suggests an upper bound on the number of causes for each node $i \in \{1, \ldots, p\}$, the computational complexity of MMLH can be reduced.

Concretely, assume that for each node i the structure space $\mathbf{\Gamma}_i = \{0,1\}^p$ only contains p-dimensional binary vectors with at most $m < p$ non-zero entries. In this case, for the cardinality of $\mathbf{\Gamma}_i$ it holds that

$$|\mathbf{\Gamma}_i| = \sum_{k=0}^{m} \binom{p}{k} < m\binom{p}{m} = O(p^m).$$

Therefore, the number of required evaluations of the MML criterion (31) reduces from $p2^p$ to less than $pp^m = p^{m+1}$, which may be beneficial especially for large dimensions p and upper bounds $m \ll p$. Moreover, since now $k_i \leq m < p$, also the computational cost of each parameter learning procedure (line 3) and determinant computation (line 4) reduces. Note that the upper bound m may be given as input parameter to Algorithm 8 and the corresponding reduction of the structure set (from p-dim. binary vectors with at most p non-zero entries to those with at most $m < p$ non-zero entries) is straight-forward to implement.

Note also that, in the general framework of a non-reduced structure space, the related state-of-the-art MDLH method (Jalaldoust et al. (2022)) also requires to solve p optimization problems, each relying on 2^p evaluations of their MDL function. Such an evaluation also contains a parameter estimation procedure for the observed data $\boldsymbol{x}$. In addition, while this function does not include Hessian matrix and determinant computations, it requires a large number N of Monte Carlo simulations for additional parameter learning (integral estimation). This results in a total of $(N+1)p2^p$ parameter learning procedures required by MDLH.

Remark 5 *In order to optimize the performance of the MMLH method, especially in high-dimensional setups, it is recommended to carry out the p independent optimizations (lines 2–6 of Algorithm 1) in parallel. Additionally, note that when the number of nodes is large, it is possible to replace the exhaustive search algorithm by a genetic algorithm, which can be parallelized as well, see e.g. Hlaváčková-Schindler and Plant (2020a).*

6. Numerical Experiments

In this section, we illustrate the performance of the proposed MMLH algorithm on both simulated synthetic data (with known ground-truth connectivity matrices) and real G7 sovereign bond data.

In all our experiments, we consider both MMLH with uniform prior (28) (denoted by MMLH-u) and with exponential prior (30) (denoted by MMLH-e). These proposed procedures are compared with the related state-of-the-art MDL-based method for causal inference in exp-MHPs (denoted by MDLH) introduced in Jalaldoust et al. (2022). As a representative of the lasso-based procedures we consider the method ADM4 from Zhou et al. (2013). Note that this method has been also considered in the experiments reported in Jalaldoust et al. (2022) and shown to outperform e.g. the method NPHC from Achab et al. (2017). Moreover, we consider two classical related model selection methods, namely the one obtained via the BIC and the one obtained via the AIC (i.e. Algorithm 8 where the criterion (31) is replaced by the BIC and AIC, respectively). Further, we investigate Algorithm 8 where the criterion (31) is reduced to the first term only (the negative log-likelihood). This method is denoted by MLE-ms, where "ms" stands for model selection. We also consider standard maximum likelihood estimation, where node j is assumed to cause node i if and only if the estimated α_{ij}-value is larger than a pre-set threshold (here 0.1). This method is denoted by MLE-thr. The following results for MDLH and ADM4 are based on the code

provided by Jalaldoust et al. (2022), the respective algorithms of the other methods are newly implemented.

6.1 Experiments with Synthetic Data

In our synthetic experiments, we choose different setups inspired by those reported in Jalaldoust et al. (2022). In particular, we investigate exp-MHPs of dimension $p = 7$, 10 and 20, respectively, and focus on short time horizons T, i.e. $T \leq 100p$. Moreover, we use the F1 score to evaluate the accuracy of our inferred connectivity matrices (in comparison to the respective ground truth). The $F1$ score is defined as the harmonic mean of the precision and recall measures:

$$\text{F1 score} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}},$$

where

$$\text{precision} = \frac{\text{number of correctly predicted edges (ones)}}{\text{total number of predicted edges (ones)}},$$

$$\text{recall} = \frac{\text{number of correctly predicted edges (ones)}}{\text{number of edges (ones) present in the ground truth}}.$$

In all considered settings, the experiments are repeated N (here, $N = 100$) times, yielding N estimates $\hat{\boldsymbol{\gamma}}^1, \ldots, \hat{\boldsymbol{\gamma}}^N$ from Algorithm 8. An average F1 score over the N trials, along with the corresponding standard deviation (put in parenthesis), is reported. The reported computing times are also averaged over trials and correspond to an implementation of Algorithm 8, which is parallelized at the level of nodes (see Remark 5). The code was run on p parallel cores of a HPC architecture located at the University of Klagenfurt (AMD EPYC 7532, 2.4 GHz, 32-core processor).

We focus on sparse connectivity graphs (two different settings) and investigate also the mid-dense setting considered in Jalaldoust et al. (2022).

Sparse settings We consider two different sparse settings. In the first one, the causal structure corresponds to a unidirectional (cascade) coupling structure with self-excitation in the first component. This means that all entries α_{ij} of the influence matrix $\boldsymbol{\alpha}$ are zero, except for α_{11} and those in the lower diagonal (i.e., the entries $\alpha_{(i+1),i}$, $i = 1, \ldots, p-1$). In the second setting, each node is influenced either by itself or by one of the other components (single input structure). This means that the influence matrix $\boldsymbol{\alpha}$ has exactly one non-zero entry per row, which (in contrast to the cascade setting) is randomly placed.

Both settings have p connections (out of a total of p^2 possible connections), corresponding to 14.3%, 10% and 5% of edges for $p = 7$, 10, and 20, respectively. For $p = 10$ and $p = 20$, we assume to have some prior expert knowledge on the maximum number of causes per node and reduce the model search to structure sets $\boldsymbol{\Gamma}_i = \{0, 1\}^p$, $i \in \{1, \ldots, p\}$, which contain binary vectors with at most $m = 5$ and $m = 3$ non-zero entries, respectively. Moreover, in both settings we set all non-zero α_{ij}-parameters to 0.55, and consider $\mu_i = 0.5$ and $\beta_{ij} = 1$. Furthermore, we set the parameter of the uniform prior (28) and exponential prior (30) to $b = 10^5$ and $c = 10^{-5}$, respectively, choosing thus very flat and little-informative prior distributions.

The results for the two sparse settings are reported in Table 1 and Table 2, respectively, and also compared to those obtained by randomly assigning one connection per row in the

desired connectivity matrix, a procedure denoted by RAND. We observe that both variants of the proposed algorithm MMLH-u and MMLH-e yield F1 scores higher than those of the other methods and that there is no tangible difference between these two variants. Moreover, the results obtained with MMLH are comparable to those obtained with BIC. This may be explained by the fact that a large value of b (resp. small value of c) leads to a stronger penalty of structures $\boldsymbol{\gamma}_i \in \boldsymbol{\Gamma}_i = \{0,1\}^p$ with large number of non-zero entries k_i, as it happens in BIC. Another observation is the poor performance of the classical maximum likelihood approach MLE-thr, especially for large dimensions p. This may result from the fact that, when p is large, a lot of parameters have to be estimated simultaneously under this method, while the model selection approach reduces the amount of parameters to be estimated, taking different structures into account. Moreover, especially in sparse settings, the limited time horizon may not allow for enough observations to ensure the convergence of the maximum likelihood estimator to the true parameter vector in practice.

The impact of the choice of b (resp. c) for MMLH-u (resp. MMLH-e) on the F1 score is illustrated in Figure 2 (blue lines), where we focus on the cascade scenario with $p = 7$ and $T = 200$. Decreasing b (resp. increasing c) leads to a decrease in the F1 accuracy. However, we observe that the "true positive" (TP) score (red lines) is not strongly influenced by the choice of b (resp. c). This means that when decreasing b (resp. c), MMLH still identifies the correct connections with a high precision, but also proposes connections which are not present in the underlying ground truth graph. Similar observations can be made for the second sparse setting (figures not shown).

	F1 score						
$p =$	7			10 (red. $\boldsymbol{\Gamma}_i$, m=5)			20 (red. $\boldsymbol{\Gamma}_i$, m=3)
$T =$	200	400	700	200	400	700	200
Runtime	12.7 s	40.4 s	116.4 s	78.0 s	277.4 s	821.3 s	521.5 s
MMLH-u	**0.948** (0.089)	**0.979** (0.053)	**0.985** (0.046)	**0.953** (0.072)	**0.968** (0.062)	**0.982** (0.039)	**0.933** (0.067)
MMLH-e	**0.948** (0.089)	**0.979** (0.053)	**0.985** (0.046)	**0.953** (0.072)	**0.968** (0.062)	**0.982** (0.039)	**0.933** (0.067)
MLE-ms	0.572 (0.048)	0.603 (0.038)	0.623 (0.035)	0.540 (0.034)	0.577 (0.038)	0.602 (0.032)	0.505 (0.022)
MLE-thr	0.410 (0.082)	0.416 (0.069)	0.413 (0.078)	0.231 (0.052)	0.240 (0.052)	0.236 (0.051)	0.099 (0.017)
BIC	0.943 (0.090)	0.977 (0.055)	0.983 (0.046)	0.943 (0.080)	0.964 (0.063)	0.977 (0.042)	0.927 (0.068)
AIC	0.850 (0.094)	0.862 (0.079)	0.896 (0.076)	0.804 (0.086)	0.833 (0.074)	0.860 (0.064)	0.719 (0.047)
RAND	0.130 (0.117)	0.159 (0.150)	0.159 (0.131)	0.102 (0.096)	0.095 (0.097)	0.084 (0.093)	0.045 (0.045)
ADM4	0.773 (0.063)	0.782 (0.055)	0.807 (0.049)	0.733 (0.039)	0.759 (0.034)	0.784 (0.050)	0.694 (0.051)
MDLH	0.566 (0.062)	0.533 (0.062)	0.556 (0.061)	0.431 (0.056)	0.421 (0.041)	0.429 (0.060)	0.398 (0.046)

Table 1: Sparse setting 1: Cascade structure. The values for the uniform and exponential priors are $b = 10^5$ and $c = 10^{-5}$, respectively. Moreover, red. $\boldsymbol{\Gamma}_i$ denotes a reduced structure set.

Mid-dense setting Now, we investigate a default scenario considered in Jalaldoust et al. (2022). In this setting, all diagonal entries of the influence matrix $\boldsymbol{\alpha}$ are non-zero, i.e. all nodes are self-excitatory. All non-diagonal entries of the adjacency matrix of the underlying connectivity graph are randomly drawn from a Bernoulli distribution with success probability 0.3. In the case of a success (one), the corresponding α_{ij} is drawn from $U([0.1, 0.2])$ (and so are the α_{ii}). Moreover, each entry μ_i of the baseline vector $\boldsymbol{\mu}$ is drawn from $U([0.5, 1.0])$ and all β_{ij} are again set to 1. Here, we further set the prior parameters b and c to 4 and 0.3,

	F1 score						
$p =$	7			10 (red. $\mathbf{\Gamma}_i$, m=5)			20 (red. $\mathbf{\Gamma}_i$, m=3)
$T =$	200	400	700	200	400	700	200
Runtime	13.0 s	41.2 s	115.1 s	79.5 s	278.9 s	836.1 s	522.2 s
MMLH-u	**0.956** (0.074)	**0.967** (0.068)	**0.978** (0.051)	**0.944** (0.084)	**0.960** (0.070)	**0.958** (0.066)	**0.929** (0.072)
MMLH-e	**0.956** (0.074)	**0.967** (0.068)	**0.978** (0.051)	**0.944** (0.084)	**0.960** (0.070)	**0.958** (0.066)	**0.929** (0.072)
MLE-ms	0.571 (0.043)	0.599 (0.039)	0.617 (0.032)	0.536 (0.037)	0.569 (0.040)	0.592 (0.035)	0.502 (0.025)
MLE-thr	0.414 (0.061)	0.404 (0.081)	0.401 (0.077)	0.243 (0.052)	0.237 (0.054)	0.234 (0.059)	0.096 (0.021)
BIC	0.954 (0.075)	0.961 (0.072)	0.975 (0.056)	0.936 (0.085)	0.954 (0.073)	0.955 (0.068)	0.922 (0.073)
AIC	0.849 (0.082)	0.863 (0.089)	0.875 (0.080)	0.794 (0.075)	0.820 (0.081)	0.837 (0.070))	0.710 (0.053)
RAND	0.127 (0.128)	0.154 (0.149)	0.144 (0.131)	0.093 (0.082)	0.091 (0.090)	0.104 (0.096)	0.052 (0.047)
ADM4	0.471 (0.068)	0.528 (0.043)	0.555 (0.059)	0.519 (0.073)	0.516 (0.063)	0.529 (0.067)	0.353 (0.042)
MDLH	0.768 (0.066)	0.828 (0.076)	0.872 (0.060)	0.742 (0.070)	0.780 (0.071)	0.835 (0.047)	0.686 (0.041)

Table 2: Sparse setting 2: Single input structure. The values for the uniform and exponential priors are $b = 10^5$ and $c = 10^{-5}$, respectively. Moreover, red. $\mathbf{\Gamma}_i$ denotes a reduced structure set.

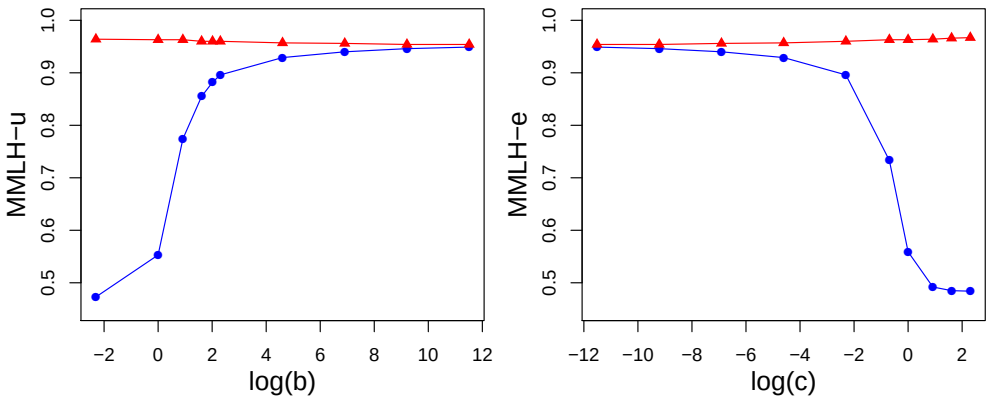

Figure 2: Cascade structure for $p = 7$ and $T = 200$. F1 score (blue lines) and TP-score (red lines) as functions of the uniform prior parameter b (left panel) and exponential prior parameter c (right panel). The x-axes are reported in log-scale.

respectively, reducing the penalty strength on structures $\boldsymbol{\gamma}_i$ with a larger number of non-zero entries. This is motivated by the fact that this scenario may be considered as a mid-dense setting, since the ground truth connectivity matrices contain on average $p + 0.3p(p-1)$ connections. In the case of $p = 7$, this corresponds to an average of 40% of edges.

The results are reported in Table 3 for different values of T. We observe that the method with the highest F1-accuracy is MDLH. Moreover, on shorter time horizons T, MMLH is also outperformed by ADM4 and MLE-thr. For $T \geq 1000$, the two rivals are MDLH and AIC, which give a score close to 0.95 and 0.92, respectively (MMLH, in comparison, is approaching 0.9). For the chosen prior parameters, we observe a slightly better performance for MMLH-u than for MMLH-e, except for the cases $T = 200$ and $T = 400$.

	F1 score					
$p =$	7					
$T =$	200	400	700	1000	1200	1400
Runtime	54.3 s	149.4 s	472.4 s	882.2 s	1818.2 s	2354.5 s
MMLH-u	0.567 (0.106)	0.692 (0.084)	0.787 (0.070)	0.833 (0.059)	0.857 (0.054)	0.872 (0.049)
MMLH-e	0.579 (0.096)	0.692 (0.085)	0.772 (0.068)	0.812 (0.060)	0.841 (0.059)	0.847 (0.050)
MLE-ms	0.597 (0.080)	0.670 (0.087)	0.797 (0.081)	0.767 (0.065)	0.795 (0.070)	0.804 (0.052)
MLE-thr	0.623 (0.096)	0.729 (0.097)	0.786 (0.096)	0.824 (0.075)	0.836 (0.079)	0.863 (0.065)
BIC	0.399 (0.080)	0.471 (0.086)	0.537 (0.075)	0.613 (0.071)	0.645 (0.074)	0.707 (0.070)
AIC	0.490 (0.099)	0.660 (0.103)	0.802 (0.081)	0.877 (0.065)	0.901 (0.059)	0.920 (0.040)
ADM4	0.695 (0.072)	0.748 (0.066)	0.761 (0.067)	0.786 (0.059)	0.784 (0.056)	0.786 (0.063)
MDLH	**0.767** (0.062)	**0.841** (0.058)	**0.900** (0.052)	**0.927** (0.041)	**0.936** (0.046)	**0.942** (0.035)

Table 3: Mid-dense setting 3: Bernoulli random structure. The values for the uniform and exponential priors are $b = 4$ and $c = 0.3$, respectively.

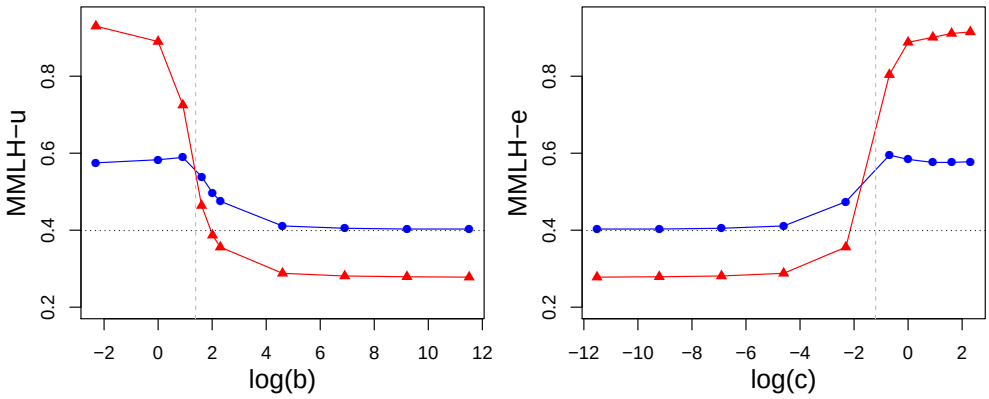

Figure 3: Bernoulli random structure for $p = 7$ and $T = 200$. F1 score (blue lines) and TP score (red lines) as functions of the uniform prior parameter b (left panel) and exponential prior parameter c (right panel). The vertical grey dashed lines indicate the investigated values $b = 4$ and $c = 0.3$, and the horizontal black dotted lines correspond to the F1 score for BIC. The x-axes are reported in log-scale.

In Figure 3, we report again the impact of b (left panel) and c (right panel) on the F1 score (blue lines) and TP score (red lines) for the case $T = 200$. The previously considered values $b = 4$ and $c = 0.3$ are marked as vertical grey dashed lines. Remarkably, while in the sparse scenarios BIC almost reached the performance of MMLH, both MMLH-u and MMLH-e outperform BIC in this mid-dense setting for all considered values of b and c (though only slightly for large b and small c).

Impact of the choice of the decay constants Now we study how the choice of the decay constants β_{ij} influences the performance of the considered methods. We focus on the first sparse setting (cascade structure) and note that similar observations are made also for

the other previously investigated scenarios. In particular, we consider the left column of Table 1 (where $T = 200$ and $\beta_{ij} = 1$) and analyse how the results reported there change for other choices of T and β_{ij}, see Table 4. Note that, the larger the decay constants β_{ij}, the fewer event occurrence times are observed on a fixed time horizon T. Thus, in the first and third columns of Table 4 fewer data points are observed on average, while in the second and fourth column of Table 4 the average number of observed data points is approximately the same as in the left column of Table 1.

It is observed that all methods perform worse for larger values of the decay constants β_{ij}, except for MDLH whose performance seems not to be strongly influenced by the choice of the decay constants. However, there are no changes in the ranking of the considered methods (except for MDLH slightly outperforming ADM4 when $\beta_{ij} = 1.5$). Moreover, we find that larger values of the time horizon T, to adjust for the decrease in the average number of observed data points caused by larger values of the decay constants β_{ij}, do not fully compensate for the loss in performance. Note that these relationships can be also observed for other values of the β_{ij}.

	F1 score			
$p = 7$, T	200	280	200	324
β_{ij}	1.5	1.5	2	2
MMLH-u	**0.837** (0.155)	**0.857** (0.132)	**0.715** (0.158)	**0.798** (0.148)
MMLH-e	**0.837** (0.155)	**0.857** (0.132)	**0.715** (0.158)	**0.798** (0.148)
MLE-ms	0.542 (0.068)	0.552 (0.063)	0.484 (0.071)	0.529 (0.064)
MLE-thr	0.342 (0.048)	0.347 (0.052)	0.352 (0.051)	0.359 (0.044)
BIC	0.834 (0.149)	0.852 (0.133)	0.709 (0.155)	0.790 (0.145)
AIC	0.762 (0.134)	0.769 (0.112)	0.664 (0.137)	0.723 (0.132)
ADM4	0.597 (0.060)	0.612 (0.050)	0.620 (0.051)	0.642 (0.049)
MDLH	0.617 (0.073)	0.615 (0.080)	0.608 (0.082)	0.624 (0.074)

Table 4: Sparse setting 1: Cascade structure. Different values for the decay constants β_{ij} and time horizon T are considered. The values for the uniform and exponential priors are $b = 10^5$ and $c = 10^{-5}$, respectively.

6.2 Experiments with Real-World Data

The goal of this subsection is to illustrate how the proposed MMLH approach performs on real-world data. In particular, we consider 10-year (2003-2014) sovereign bond yield volatilities of seven large economies called the Group of Seven (G7), being composed of the US, Canada, Germany, France, Japan, UK and Italy. This dataset has been investigated in Demirer et al. (2018) and also analysed in Jalaldoust et al. (2022). It is publicly available at: `http://qed.econ.queensu.ca/jae/2018-v33.1/demirer-et-al/`

As this dataset corresponds to time series data, we pre-process it to identify shocks (i.e. events able to be reproduced by a point process) in the return volatilities. In particular, following Jalaldoust et al. (2022), for each country we roll a one-year window over the respective data and register an event if the latest value of the window is among the top 20%

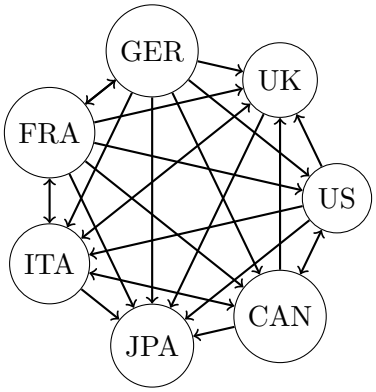

Figure 4: Connectivity graph derived from Umar et al. (2022) (an edge is drawn if it appears in at least one of the three graphs shown in their Figure 1).

of values in the rolling window. The resulting number of events registered for each country is roughly 500. Following Jalaldoust et al. (2022), we assume that this data is observed within a time horizon of $T = 400$ of the investigated exp-MHP, i.e. an instance of a short time horizon.

Our study is conducted as follows: We compare MMLH with comparable methods designed for MHPs (in particular, MDLH, ADM4, BIC, and AIC) as well as with available expert knowledge. For MMLH, ADM4, BIC, and AIC, we prepare the data as described above and launch the respective causal discovery algorithm. For MDLH, we rely on the results reported in Jalaldoust et al. (2022). Moreover, as "expert knowledge" we consider the conclusions reported in Umar et al. (2022), where the network connectedness between sovereign bond yield curve components of the G7 countries is discussed, without using MHPs as a basis. The connectivity graph derived from Umar et al. (2022) is shown in Figure 4. The results of the investigated methods are presented in Figure 5, with the graphs in the top panels corresponding to ADM4 (left) and MDLH (right) and the graphs in the middle panels corresponding to BIC (left) and AIC (right). In the bottom panels, we report the results for MMLH-e with $c = 10^{-5}$ (left), $c = 0.3$ (central), and $c = 2.5$ (right). As expected, similar results are obtained for MMLH-u. For example, using $b = 10^5$ we obtain the same graph as reported in the bottom left panel, and for $b = 4$ the same as shown in the bottom central panel, except for the connection "ITA to FRA" not being present. Note that the red connections in the graphs visualized in Figure 5 are those that are in agreement with the expert knowledge graph of Figure 4.

We observe that ADM4 yields the most connections in agreement with Umar et al. (2022), however it also suggests 6 connections that are not present in Figure 4. As expected from the synthetic experiments, for small c (resp. large b) (i.e., when we have a strong penalty on structures $\boldsymbol{\gamma}_i \in \boldsymbol{\Gamma}_i = \{0,1\}^p$ with many non-zero entries), MMLH performs similar to BIC, which only suggests one connection not present in Figure 4 ("US to GER"). Increasing c to 0.3 (resp. decreasing b to 4) adds further connections to the graph (bottom central panel), of which all are reported in Umar et al. (2022). In particular, for $c = 0.3$ MMLH-e yields 9 connections present in Figure 4 (and a single one that is not reported

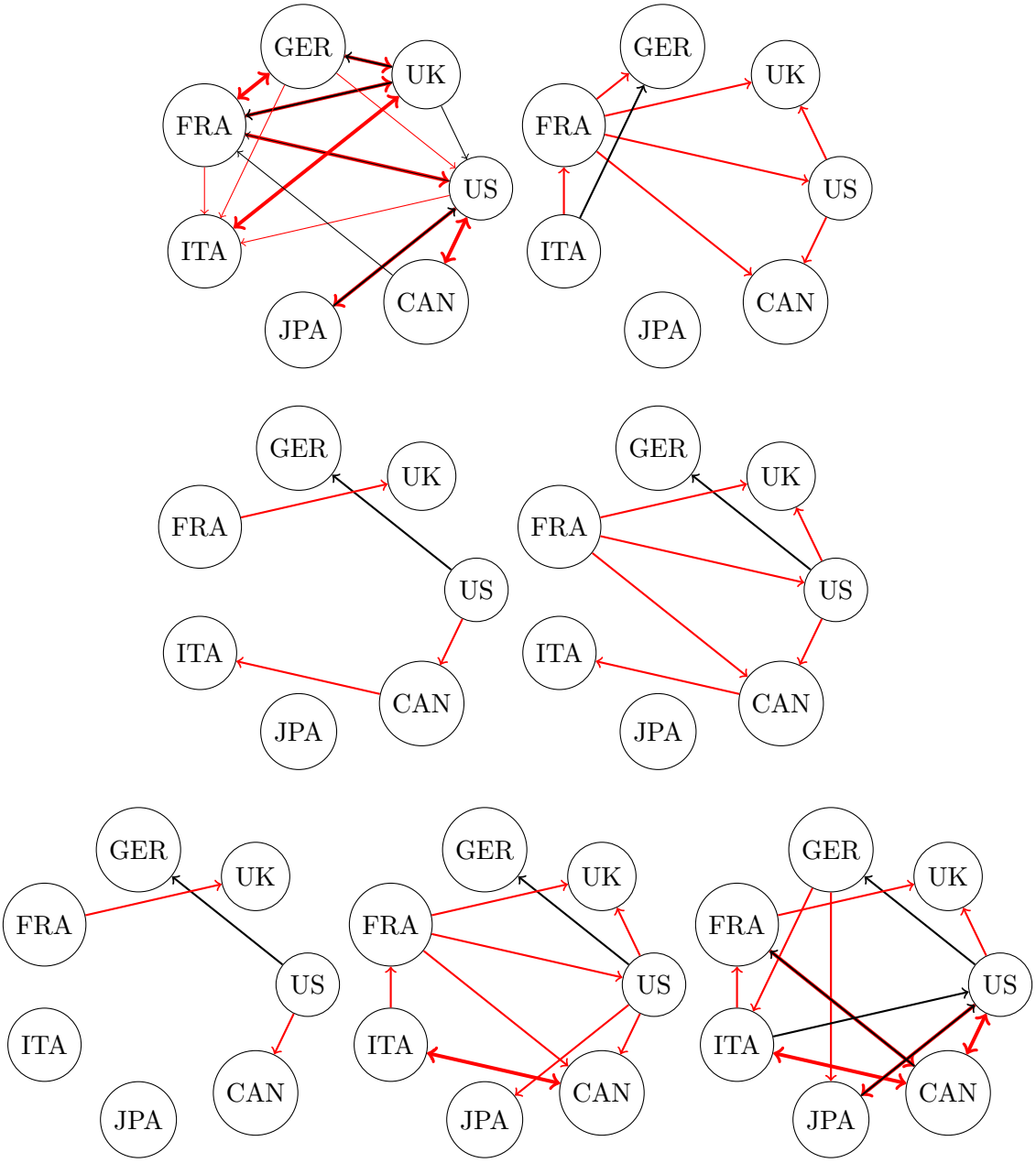

Figure 5: Connectivity graph for ADM4 (top left panel), MDLH (top right panel), BIC (middle left panel), AIC (middle right panel), MMLH-e with $c = 10^{-5}$ (bottom left panel), $c = 0.3$ (bottom central panel), and $c = 2.5$ (bottom right panel). The red connections are those that are in agreement with Umar et al. (2022).

there). In comparison, MDLH and AIC only capture 7 (resp. 6) connections of Figure 4 (and also report one connection that is not reported there). Thus, the proposed MMLH algorithm not only outperforms the classical BIC and AIC methods, but also reports more connections

that are in agreement with the literature than the related state-of-the-art MDLH method. Note also that 6 out of 8 connections detected by MDLH and all connections obtained by AIC are in agreement with MMLH-e for $c = 0.3$.

Moreover, we observe that many outgoing connections from France are captured well by ADM4, MDLH, AIC and MMLH. Note, however, that for MMLH-e the connection "FRA to US" disappears when c is increased from 0.3 (bottom central panel) to 2.5 (bottom right panel). This may be explained by the newly discovered edges now dominating the connection "FRA to US", in the sense that the chosen structure shown in the bottom right panel yields a slightly smaller value of criterion (31) than the same structure including the connection "FRA to US". A similar effect is observed for MMLH-u. Further, the outgoing connections from Germany are only (partially) captured by ADM4. However, this improves for MMLH under large values of c (resp. small values of b). In particular, using $c = 2.5$ and $b = 0.25$, we obtain the connections "GER to ITA" and "GER to JPA". We also observe that Japan is isolated from the other countries, except under ADM4 and MMLH, which report an ingoing connection from the US, in agreement with Umar et al. (2022). Only MMLH-e (for $c = 2.5$) reports a second ingoing connection (from Germany), which is also present in Figure 4. These observations also suggest a solid performance of the proposed MMLH procedure.

7. Conclusion and Discussion

In this paper, we estimated Granger causal relations between components of multivariate Hawkes processes with exponential decay kernels. These relations are described by a connectivity graph, which can be estimated from observations of the process. We approached this problem by proposing an optimization criterion and a model selection algorithm MMLH based on the minimum message length principle.

In contrast to other model selection algorithms, MMLH incorporates prior distributions of the underlying model parameters. While classical model selection criteria (e.g. BIC) are designed to penalize models with a lot of parameters (i.e. structures with many non-zero entries) and thus do not take into account other descriptions of the model, MMLH offers more flexibility in terms of structure-related penalty. This may be particularly beneficial, if some a-priori expert knowledge on the structure of the underlying graph exists. Given the fact that all model parameters to be estimated are non-negative, we investigated both a uniform prior and an exponential prior and observed a similar performance for both of them in all our experiments.

We conducted synthetic experiments in which we compared the proposed algorithm to other related classical and state-of-art methods, focusing on short time horizons. In the considered sparse graph scenarios, MMLH achieved the highest F1 scores among all comparison methods. Concerning the investigated mid-dense scenario, MMLH showed an F1 score comparable to MLE-ms, MLE-thr, AIC and ADM4. However, the performance of MMLH improved with increasing time horizons, the only rivals being the state-of-art MDLH and the classical AIC. The superior F1 score of MMLH on sparse connectivity graphs may be explained by the fact that the minimum message length principle prefers short encodings of the model (i.e. sparse graphs) over longer encodings (i.e. non-sparse graphs) together with a short description of the data using the model.

Finally, we illustrated the proposed method on G7 sovereign bond data and compared the inferred causal connections to those of MDLH, ADM4, BIC, and AIC. We demonstrated that the connectivity graphs obtained via MMLH (with three different parameterizations of the prior function) are in agreement with the expert knowledge extracted from the literature.

As a possible future work one may investigate connectivity graphs in multivariate Hawkes processes with other kernels or intensities given by non-linear functional relationships (e.g., ReLU or sigmoid functions). Another research direction is a modification of the algorithm which allows to increase the performance on non-sparse structures. Moreover, one may study if/how the presented method benefits from considering directed acyclic graph structures. For recent approaches in this regard, we refer to Zheng et al. (2018), Zhang et al. (2022), and Wei et al. (2023a).

Acknowledgments

The authors would like to thank the reviewers for their constructive comments which helped to improve the manuscript. Their support is highly appreciated. K. H.-S. acknowledges a partial support by the Austrian Science Foundation (FWF) project I 5113-N, and by the Czech Academy of Sciences, Praemium Academiae awarded to M. Paluš. A part of this paper was written while I.T. was member of the Institute of Stochastics, Johannes Kepler University Linz, 4040 Linz, Austria. During this time, I.T. was supported by the Austrian Science Fund (FWF): W1214-N15, project DK14.

References

Massil Achab, Emmanuel Bacry, Stephane Gaiffas, Iacopo Mastromatteo, and Jean-Francois Muzy. Uncovering causality from multivariate Hawkes integrated cumulants. In *International Conference on Machine Learning*, pages 1–10. PMLR, 2017.

Alfred V. Aho and John E. Hopcroft. *The Design and Analysis of Computer Algorithms*. Pearson Education India, 1974.

Emmanuel Bacry, Martin Bompaire, Stephane Gaïffas, and Jean-Francois Muzy. Sparse and low-rank multivariate Hawkes processes. *Journal of Machine Learning Research*, 21 (50):1–32, 2020.

John Horton Conway and Neil James Alexander Sloane. On the Voronoi regions of certain lattices. *SIAM Journal on Algebraic Discrete Methods*, 5(3):294–305, 1984.

Mert Demirer, Francis X. Diebold, Laura Liu, and Kamil Yilmaz. Estimating global bank network connectedness. *Journal of Applied Econometrics*, 33(1):1–15, 2018.

Vanessa Didelez. Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):245–264, 2008.

D. Eddelbuettel and R. François. Rcpp: Seamless R and C++ integration. *J. Stat. Soft.*, 40 (8):1–18, 2011. doi: 10.18637/jss.v040.i08. URL `http://www.jstatsoft.org/v40/i08/`.

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.

Peter Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.

Peter Grünwald and Teemu Roos. Minimum description length revisited. *International Journal of Mathematics for Industry*, 11(01):1930001, 2019.

Niels Richard Hansen, Patricia Reynaud-Bouret, and Vincent Rivoirard. Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, 21(1):83–143, 2015.

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

Kateřina Hlaváčková-Schindler and Claudia Plant. Heterogeneous graphical Granger causality by minimum message length. *Entropy*, 22(12):1400, 2020a.

Kateřina Hlaváčková-Schindler and Claudia Plant. Graphical Granger causality by information-theoretic criteria. In *Proceedings of the 24th European Conference on Artificial Intellingence, 2020*, pages 1459–1466. IOS Press, 2020b.

Tsuyoshi Idé, Georgios Kollias, Dzung T. Phan, and Naoki Abe. Cardinality-regularized Hawkes-Granger model. *Advances in Neural Information Processing Systems*, 34:2682–2694, 2021.

Amirkasra Jalaldoust, Kateřina Hlaváčková-Schindler, and Claudia Plant. Causal discovery in Hawkes processes by minimum description length. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, pages 6978–6987. PKP Publishing Services Network, 2022.

Anatoli Juditsky, Arkadi Nemirovski, Liyan Xie, and Yao Xie. Convex parameter recovery for interacting marked processes. *IEEE Journal on Selected Areas in Information Theory*, 1(3):799–813, 2020.

Ming Li and Paul Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*, volume 3. Springer, 2008.

Enes Makalic and Daniel Francis Schmidt. Minimum message length inference of the exponential distribution with type I censoring. *Entropy*, 23(11):1439, 2021.

Yosihiko Ogata. On Lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, 27(1):23–31, 1981.

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

Jonathan J. Oliver, Rohan A. Baxter, and Chris S. Wallace. Unsupervised learning using MML. In *International Conference on Machine Learning*, pages 364–372, 1996.

Taisuke Ozaki. Maximum likelihood estimation of Hawkes' self-exciting point processes. *Annals of the Institute of Statistical Mathematics*, 31(1):145–155, 1979.

Stephen Reid, Robert Tibshirani, and Jerome Friedman. A study of error variance estimation in lasso regression. *Statistica Sinica*, pages 35–67, 2016.

Jorma Rissanen. *Stochastic Complexity in Statistical Inquiry*, volume 15. World Scientific, 1998.

Teemu Roos, Petri Myllymaki, and Jorma Rissanen. MDL denoising revisited. *IEEE Transactions on Signal Processing*, 57(9):3347–3360, 2009.

Leigh Shlomovich, Edward A. K. Cohen, and Niall Adams. A parameter estimation method for multivariate binned Hawkes processes. *Statistics and Computing*, 32:98, 2022. doi: 10.1007/s11222-022-10121-2.

Sanja Singer and Saša Singer. Complexity analysis of Nelder-Mead search iterations. In *Proceedings of the first Conference on Applied Mathematics and Computation*, pages 185–196. PMF–Matematicki Odjel Zagreb, Croatia, 1999.

Deborah Sulem, Vincent Rivoirard, and Judith Rousseau. Bayesian estimation of nonlinear Hawkes process. *arXiv preprint arXiv:2103.17164*, 2021.

William Trouleau, Jalal Etesami, Matthias Grossglauser, Negar Kiyavash, and Patrick Thiran. Cumulants of Hawkes processes are robust to observation noise. In *International Conference on Machine Learning*, pages 10444–10454. PMLR, 2021.

Zaghum Umar, Yasir Riaz, and David Y Aharon. Network connectedness dynamics of the yield curve of G7 countries. *International Review of Economics & Finance*, 79:275–288, 2022.

Alejandro Veen and Frederic P. Schoenberg. Estimation of space–time branching process models in seismology using an EM–type algorithm. *Journal of the American Statistical Association*, 103(482):614–624, 2008.

Chris S. Wallace. *Statistical and Inductive Inference by Minimum Message Length*. Springer, 2005.

Chris S. Wallace and D.M. Boulton. An information measure for classification. *The Computer Journal*, 11(2):185–194, 1968.

Chris S. Wallace and David L. Dowe. Minimum message length and Kolmogorov complexity. *The Computer Journal*, 42(4):270–283, 1999.

Chris S. Wallace and P.R. Freeman. Estimation and inference by compact coding. *Journal of the Royal Statistical Society: Series B (Methodological)*, 49(3):240–252, 1987.

Haoyun Wang, Liyan Xie, Alex Cuozzo, Simon Mak, and Yao Xie. Uncertainty quantification for inferring Hawkes networks. *Advances in Neural Information Processing Systems*, 33:7125–7134, 2020.

Song Wei, Yao Xie, Christopher S Josef, and Rishikesan Kamaleswaran. Causal graph discovery from self and mutually exciting time series. *arXiv preprint arXiv:2106.02600*, 2023a.

Song Wei, Yao Xie, Christopher S. Josef, and Rishikesan Kamaleswaran. Granger causal chain discovery for sepsis-associated derangements via continuous-time Hawkes processes. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2536–2546, 2023b.

Hongteng Xu, Mehrdad Farajtabar, and Hongyuan Zha. Learning Granger causality for Hawkes processes. In *International Conference on Machine Learning*, pages 1717–1726. PMLR, 2016.

Zhen Zhang, Ignavier Ng, Dong Gong, Yuhang Liu, Ehsan Abbasnejad, Mingming Gong, Kun Zhang, and Javen Qinfeng Shi. Truncated matrix power iteration for differentiable dag learning. *Advances in Neural Information Processing Systems*, 35:18390–18402, 2022.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

Ke Zhou, Hongyuan Zha, and Le Song. Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes. In *Artificial Intelligence and Statistics*, pages 641–649. PMLR, 2013.