Accurate estimation of cross-excitation in multivariate Hawkes process models of infectious diseases

Youness Diouane
Computer Science Department
Boston College
Boston, USA
diouane@bc.edu

Frederic Schoenberg
Department of Statistics
UCLA
Los Angeles CA, USA
frederic@stat.ucla.edu

George Mohler
Computer Science Department
Boston College
Boston, USA
mohlerg@bc.edu

***Abstract*—Multivariate Hawkes processes are a popular model for estimating Granger causality from event sequences on networks. In this work we show that under certain parameter regimes, such as those that arise when modeling infectious disease transmission, false discovery of cross-excitation becomes a major problem. We first provide evidence through simulation that substantial spurious cross-excitation is present when the largest eigenvalue of the productivity matrix approaches the critical value of 1, which leads to multicollinearity. We then propose and compare several methods for mitigating false cross-excitation, through different types of regularization and staged estimation. Our experimental results include both synthetic data as well as transmission data from the Covid-19 pandemic.**

***Index Terms*—Multivariate Hawkes process, Granger causality, multicollinearity, Covid-19**

I. Introduction

Hawkes processes [1] are point process models of event clustering with application to a variety of phenomena, including seismology [2], [3], finance [4]–[7], and crime modeling [8], [9]. The main advantage of Hawkes processes over other point processes, such as stationary Poisson processes, is their ability to model how past events influence current events of the process due to their memory property. Furthermore, they exhibit a natural notion of Granger causality due to the fact that occurrences of events in the past may increase the probability of events in the future [10].

In this paper, we consider the multivariate Hawkes process (MHP), which has recently been proposed as a method for describing the transmission of infectious diseases across different locations and over time [11]–[13]. The main objective will be to estimate Granger causality across spatial locations through the productivity matrix of the Hawkes process. Previous work [14] has shown that the Granger causality structure of MHPs is fully encoded in the kernel function of the model. Suppose we have a system of M interacting point processes, $\mathbf{H} = (\mathbf{H}_t)_{0\leq t\leq T} = \left(\mathbf{H}_t^1, \mathbf{H}_t^2, \ldots, \mathbf{H}_t^M\right)_{0\leq t\leq T}$, that occur in the time interval $[0, T]$, with $M > 2$ the number of processes, each $\mathbf{H}^i$ consisting of points $\{s_j^i\}$, and with conditional intensity

$$\lambda_i(t) = \mu_i + \sum_{j=1}^{M} \sum_{k:s_k^j<t} \phi_{ij}(t - s_k^j) \tag{1}$$

where $\mu_i > 0$ corresponds to the background rate, and ϕ describes the kernel function. Then we can say that at a given time t, the component $\mathbf{H}_s^j$ Granger-causes $\mathbf{H}_t^i$ if and only if the kernel function $\phi_{ij}(t - s)$ is not vanishing. We adopt this approach to study the dynamics of epidemic diseases between different locations.

The Covid-19 pandemic has sparked a great deal of interest in point process modeling of infectious disease [15]–[20]. For instance, [21] proposed three basic models that can be fit to Covid-19 data emerging from local and national governments. The authors in [22] introduced a novel epidemic model using a latent Hawkes process with temporal covariates for modelling the infections. In this paper, we model the dynamics of epidemic diseases between different locations using the aforementioned MHPs. The aim is to infer the Granger-causality that is implicitly encoded in the kernel function, and more specifically in the productivity matrix. The data are spatial-temporal, where the total number of daily infections in every specific location is recorded. Although a variety of estimation methods have been proposed, the task is nevertheless extremely challenging. Indeed, for the case of data where the precise occurrence times of the individual points are observed, the parameters can be fit by maximum likelihood estimation (MLE) [23] and the resulting estimates are known to have desirable asymptotic properties [24]. Furthermore, even if the model is missing some relevant covariates, given that their effect is small, the MLE will still be consistent under general conditions [25]. The triggering function can also be estimated non-parametrically [26]–[28]. Bayesian methods can also be used to estimate parameters and quantify uncertainty in the Hawkes processes models [29], [30]. However, for epidemic data in which data consist of daily counts rather than precise individual occurrences times, the Hawkes parameters can be

This work was supported in part by NSF grant ATD-2124313, ATD-2317397, SCC-2125319 and AFOSR MURI grant FA9550-22-1-0380.

efficiently fit by least squares, applying methods from time series theory to the bin-count sequences of mutivariate point process data [31], [32].

However, for the case considered here, i.e., spatio-temporal data where in each location, the total number of cases in that location is recorded, and where individuals in one location may infect individuals in another location, both MLE and least squares perform poorly. Indeed, both methods appear to grossly overestimate the cross-productivity, i.e., the rate at which those from one location infect others at a different location. We provide evidence that one issue that arises is multicollinearity: if, for instance, cases are rising in California and New York at similar rates, then the model tends to overestimate the rate at which those in California are infecting those in New York and vice versa, since the model, a priori, has no reason to suspect that those in New York are any less likely to be infected by Californians than by other New Yorkers. Our goal is to mitigate this problem, i.e., false cross-excitations in multivariate Hawkes processes for epidemic diseases using data involving spatial components (different locations) and temporal ones (total daily infections at given individual times). We propose seven different estimation methods. Furthermore, we run three experiments. The first two are based on synthetic data that mimic epidemic disease dynamics between different locations, while the third one relies on real data of Covid-19 in which the total number of daily infections is recorded between four counties in California, USA, namely San Diego, San Francisco, Sacramento and San Jose. The paper is organized as follows: Section II briefly describes the Hawkes model we will be using to fit the data. Section III explains in full detail the seven methods used for our estimation. A short analysis of multicollinearity is provided in Section IV. The issues of over-fitting and parameter tuning are examined in Section V. The various methods are applied to both simulated and real data in Section VI [1] , followed by a discussion of the results in Section VII.

II. Model

The model considered throughout is the MHP, whose conditional intensity (1) in the discrete case can be rewritten as

$$\lambda_i(t) = \mu_i + \sum_{j=1}^{M} \sum_{s=0}^{t} \phi_{ij}(t-s) N(s). \tag{2}$$

Here μ_i is the background rate (considered time independent) for points of type i, where i in this context refers to a location. ϕ is a density function i.e., non negative and integrating to 1, called the triggering density, and N is the number of new infections which varies from day to day. For infectious disease transmission, which can have an incubation period, one possibility for the triggering density governing the serial interval between cases is the Weibull distribution:

$$\phi_{ij}(t) = R_{ij} \left[(1-p)^{t^\beta} - (1-p)^{(t+1)^\beta} \right], \tag{3}$$

where p is a real parameter satisfying $0 < p < 1$ and $\beta > 0$ is the shape parameter. The reproduction matrix, R_{ij}, represents the expected number of secondary cases at location i triggered directly by an infection at location j.

The serial interval distribution of a disease is typically assumed to be constant across locations, and therefore the parameters β and p in the Weibull distribution can be estimated by superimposing the data into a univariate process [33] and estimating the Weibull parameters with Stan [34]. Subsequently, with β and p already estimated, the reproduction matrix can be estimated using a variety of regression based approaches, which is the scope of the next Section.

III. Methods

As discussed in the previous Section, we first estimate the parameters, β and p, of a superimposed univariate Hawkes process. This can be achieved by maximizing the likelihood [23] through a Bayesian approach by sampling from the posterior distribution given priors on the model parameters. In the present case, we first maximize the likelihood function of the Weibull parameters assuming the number of events per day follows a Poisson distribution. For unknown parameters θ, the log-likelihood function is given by,

$$l(\theta) = \log \prod_{t=0}^{T} \frac{\lambda(t)^{N(t)} e^{-\lambda(t)}}{N(t)!}, \tag{4}$$

where T is the length of the time window of observation. Let $L(\theta) = \max_\theta l(\theta)$. Maximizing the log-likelihood (4) is equivalent to maximizing its nominator. Thus

$$\begin{aligned} L(\theta) &= \max_\theta \log \prod_{t=0}^{T} \lambda(t)^{N(t)} e^{-\lambda(t)} \\ &= \max_\theta \sum_{t=0}^{T} \log \left(\lambda(t)^{N(t)} e^{-\lambda(t)} \right). \end{aligned} \tag{5}$$

Therefore, the maximum log-likelihood function can be written as

$$L(\theta) = \max_\theta \sum_{t=0}^{T} \left(N(t) \log \lambda(t) - \lambda(t) \right). \tag{6}$$

We solve the optimization problem in Equation (6) using Stan. We perform Hamiltonian Monte Carlo using 1000 samples. The prior densities we use are

$$R \sim beta(1,1)\ ,\ p \sim beta(1,1) \tag{7}$$

$$\mu \sim cauchy(0,5)\ ,\ \beta \sim cauchy(0,5). \tag{8}$$

The distribution $Beta(1,1)$ is a uniform distribution that constraints the parameters p and R to be bounded between 0 and 1. The prior of $Cauchy(0,5)$ was chosen for μ and β because we initially do not have much information about them. Note that, in the aforementioned estimation, we are superposing [33] the MHP into a single univariate Hawkes

[1] The code can be found at https://github.com/younesszs/False-excitation-mitigation

process, which allows us readily to estimate the serial interval distribution ϕ.

Next, with p and β fixed, we return to estimating the full MHP. In particular, we use a regression framework to estimate the reproduction matrix R. We compare several approaches that we refer to as "Linear estimate", "Ridge estimate", "Ridge Diag estimate", "Elastic Net estimate", "Elastic Net Diag estimate", "Staged1 estimate" and "Staged2 estimate".

A. Linear estimate

The first method we consider is linear regression for estimating the reproduction matrix R_{ij} and the baseline rates μ_i. Note that, with p and β fixed, the MHP intensity can be seen as obeying the linear model:

$$E(N_i(t)) = b_{0,i} + \sum_{j=1}^{M}\sum_{s=0}^{t}\Big[b_{ij}\Big((1-p)^{(t-s-1)^\beta} - (1-p)^{(t-s)^\beta}\Big)N_j(s)\Big]. \tag{9}$$

Plugging in the observed event counts in place of their expected value, this amounts to finding the best solution to the linear equation

$$\vec{y} \sim \vec{b}X. \tag{10}$$

That is, The length of the vector $\vec{y}$ of observed values is MT, where M is the number of nodes in the network and T is the number of days. In the above linear model, the linear regression coefficients $\vec{b}$ determine the baseline rate $\mu_i = b_{0,i}$, and reproduction matrix $R_{ij} = b_{ij}$. The predictor matrix X, which contains the Weibull kernel convolved with past event counts, has shape $MT \times M(M+1)$. Note that the intercept is implicitly absorbed in the predictor matrix by adding a column of 1's to X. We thus estimate the coefficients $\vec{b}$ using least-squares by minimizing

$$MSE = \sum_{i=1}^{MT}\left(y_i - \sum_{j=1}^{M(M+1)} x_{ij}b_{ij}\right)^2, \tag{11}$$

where x_{ij}'s represent the elements of the predictor matrix X.

B. Penalized regression estimates

We next consider four penalized regression methods: Ridge, Ridge Diag, Elastic Net and Elastic Net Diag.

Ridge regression consists of a L2 penalty on the coefficients b_{ij}. L2 penalty is a technique used in machine learning to mitigate the problem of over-fitting by adding a penalty term to the loss function. The penalty term is the sum of the squares of the model coefficients. The primary effect of L2 regularization is to effectively shrink the coefficients towards zero, but unlike L1 regularization, it does not set them to exactly zero. We also consider a modified Ridge regression, which we call Ridge Diag, where we only penalize the off-diagonal coefficients. For motivation, we return to the New York and California scenario from the Introduction. In the case of infectious diseases, it makes sense to have an informative prior that anticipates more transmission within a particular region than across distinct regions. Thus we penalize non-zero off-diagonal coefficients, but not the diagonal coefficients. The cost function that characterizes Ridge and Ridge Diag estimates can be written as

$$\text{MSE} + \lambda\sum_{i=1}^{MT}\sum_{j=1}^{M(M+1)} b_{ij}^2\delta_{ij}, \tag{12}$$

where λ is a regularization parameter. Here δ_{ij} is one in the case of Ridge regression, and $\delta_{ij} = 1_{i\neq j}$ in the case of Ridge Diag.

Elastic Net and Elastic Net Diag have the same structure as Ridge and Ridge Diag, respectively, illustrated in the expression (12), except for the fact that now we use elastic net regularization, that is, a combination of L1 and L2 penalties. Thus the loss function is given by

$$\begin{aligned}\text{MSE} + (1-\alpha)\lambda\sum_{i=1}^{MT}\sum_{j=1}^{M(M+1)} b_{ij}^2\delta_{ij} \\ + \alpha\lambda\sum_{i=1}^{MT}\sum_{j=1}^{M(M+1)} |b_j|\delta_{ij},\end{aligned} \tag{13}$$

where α is a mixing parameter that controls the balance between the ridge and lasso penalties and should be between 0 and 1. In the case where $\alpha = 0$ we retrieve the loss function (12) i.e., ridge regression, whereas $\alpha = 1$ corresponds to the lasso loss function. Note that, for all our case studies, the optimal value of the parameter α was determined using cross validation.

C. Staged estimates

The last two estimation methods we use are denoted by "Staged1" and "Staged2". They are named as such because they utilize two stages to estimate the Hawkes parameters. The first stage consists of independent linear regressions for each node to estimate the diagonal elements of R:

$$y_i(t) = b_0 + b_{ii}\sum_{s=0}^{t}\Big[(1-p)^{(t-s-1)^\beta} - (1-p)^{(t-s)^\beta}\Big]N_i(s). \tag{14}$$

Then in the second stage the diagonal element estimates $\hat{b}_{ii}$ are fixed and the off-diagonal elements are estimated for the entire MHP:

$$\begin{aligned}y_i(t) =& b_{0i} + \sum_{j=1}^{M}\sum_{s=0}^{t}\Big[b_{ij}\Big((1-p)^{(t-s-1)^\beta} - (1-p)^{(t-s)^\beta}\Big)N_j(s)1_{i\neq j}\Big] \\ &+ \hat{b}_{ii}\sum_{s=0}^{t}\Big[\Big((1-p)^{(t-s-1)^\beta} - (1-p)^{(t-s)^\beta}\Big)N_i(s)\Big].\end{aligned} \tag{15}$$

We refer to this method as "Staged1". Because "Staged1" tends to overestimate diagonals in the reproduction matrix, we also consider a modified version, called "Staged2", where we estimate an additional parameter γ which multiplies the diagonal coefficients and allows for a global correction to the diagonals:

$$\begin{aligned} y_i(t) = b_{0i} + \sum_{j=1}^{M}\sum_{s=0}^{t}\Bigg[b_{ij}\Big((1-p)^{(t-s-1)^\beta} \\ - (1-p)^{(t-s)^\beta}\Big) N_j(s) 1_{i\neq j}\Bigg] \\ + \gamma \hat{b}_{ii} \sum_{s=0}^{t}\Bigg[\Big((1-p)^{(t-s-1)^\beta} \\ - (1-p)^{(t-s)^\beta}\Big) N_i(s)\Bigg]. \end{aligned} \tag{16}$$

Like the penalized regression methods with diagonal modifications, the idea behind the staged estimates is that most transmission should happen within a geographical region. Then in the second stage, off-diagonal transmission is fit to the residual error of the diagonal model, the idea being that the estimated transmission between two different geographical areas is only permitted to account for additional excitation after local transmission has already been accounted for.

IV. The challenge of multicollinearity

In this section we briefly discuss the challenge of multicollinearity, which is a major problem when dealing with epidemic diseases. In the case of spatio-temporal data, both MLE and least squares tend to result in vast overestimates of the rate at which the disease spreads between different locations. Multicollinearity is a phenomenon that occurs when two or more predictor variables in a regression model are highly correlated. This is problematic because it makes it difficult to determine the exact effect of each independent variable due to the shared variance among the independent variables. Multicollinearity is an issue with MHPs, as features in Equation 9 are determined by lagged event counts convolved with the triggering kernel. In Fig. 1 on the left, we plot the correlation matrix of MHP features for one simulation. Note that there are blocks of feature pairs whose correlation coefficient is near 1.

To better understand multicollinearity in MHPs, we run 100 simulations with different reproduction matrices R. We find that collinearity gets worse as the largest eigenvalue of the reproduction matrix approaches or exceeds 1. In Fig. 1 on the right, we plot the largest correlation across feature pairs for each simulation vs the largest eigenvalue of the reproduction matrix. We observe that the maximum correlation across feature pairs approaches 1 as the maximum eigenvalue approaches 1.

V. The challenge of over-fitting

A closely related challenge in estimating cross-excitation is over-fitting. Particularly when comparing models with different numbers of parameters, it is important to watch out for

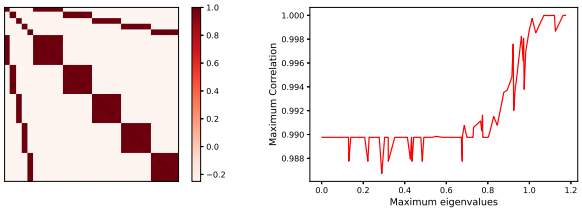

Fig. 1. Collinearity increases as the largest eigenvalue of the reproduction matrix increases. To show this, we simulate 100 MHPs with different reproduction matrices and compare feature correlation with the largest eigenvalue. On the left, we plot the correlation matrix among features for one of the simulations. MHP features are correlated as they determined by lagged event counts convolved with the triggering density. On the right, we plot the largest correlation coefficient between pairs of features against the largest eigenvalue of the reproduction matrix. Note that as the largest eigenvalue approaches 1, then there are a pair of features with correlation that also approaches 1.

the possibility that one model may offer superior fit to the training data to which it was fitted, but significantly worse on external, testing data. Our approach to mitigating this problem is to introduce different regularizations on the diagonal and off diagonal estimates of the reproduction matrix, along with staged estimates that first fit the diagonals and then fit the off-diagonals to residual error.

For the linear estimate without regularization, we observe that, like multicollinearity, over-fitting gets worse as the largest eigenvalue of the reproduction matrix approaches 1. As in the previous section, we run 100 simulations with different reproduction matrices R. We split the data in time where the first half of the time interval is used for training the linear estimate in Equation 9. We then evaluate training error, as well as test error on the held out second half of the data.

In Fig. 2, we plot the training error and the test error vs the largest eigenvalue of the reproduction matrix. While the train error does increase with the largest eigenvalue, the testing error explodes as the largest eigenvalue approaches and exceeds 1. We also plot the ratio of the training vs testing MSE, which approaches zero as the largest eigenvalue approaches 1.

VI. Experiments

In this section we compare the methods outlined in Section III on two synthetic datasets, along with daily COVID-19 case counts from several California counties in 2020.

A. Cycle experiment

In our first experiment, we generate synthetic MHP events on a network with $(M = 5)$ nodes. We use $T = 300$ days for the time window of observation and for the Weibull parameters we set $p = 0.05$ and $\beta = 2.0$. The reproduction matrix R follows a cyclic pattern with diagonal $R_{ii} = 0.8$ and off-diagonal elements $R_{ij} = 0.05$ when $mod(|i-j|/3) = 1$ (and

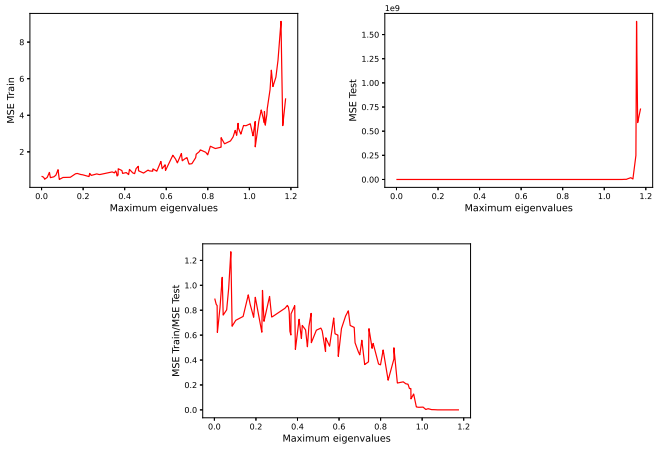

Fig. 2. Train and test error of linear regression applied to 100 simulations of MHPs with different reproduction matrices R. Upper left: Train MSE vs the largest eigenvalue of R for each of the 100 simulations. Upper right: Test MSE vs the largest eigenvalue of R for each of the 100 simulations. Bottom: Ratio of the the train-test MSE vs the largest eigenvalue for 100 MHP simulations.

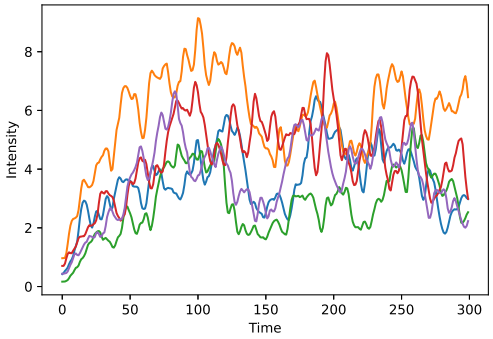

Fig. 3. Conditional intensity of simulated 5-dimesional Hawkes process with a Weibull distribution and a cyclic reproduction matrix R given in equation (17).

zero otherwise). The reproduction matrix therefore is given by:

$$R_{\text{Cycle}}^{\text{Ground truth}} = \begin{pmatrix} 0.8 & 0.05 & 0 & 0 & 0.05 \\ 0.05 & 0.8 & 0.05 & 0 & 0 \\ 0 & 0.05 & 0.8 & 0.05 & 0 \\ 0 & 0 & 0.05 & 0.8 & 0.05 \\ 0.05 & 0 & 0 & 0.05 & 0.8 \end{pmatrix}. \quad (17)$$

The $M = 5$ intensities of one simulation of this process is shown in Fig. 3. We note that the intensities are highly correlated due to having similar parameters and having cross-excitation through the reproduction matrix.

We next run 1000 simulations to estimate the reproduction matrix R and compare the methods discussed in Section III, i.e., the linear estimate, the penalized estimates and the staged estimates. As described earlier, we first fit the Weibull parameters p and β using superposition and Stan. The estimated parameters are given in Table I, and we note that the 95% confidence intervals contain the true values.

TABLE I

ESTIMATED VALUES OF THE WEIBULL DISTRIBUTION PARAMETERS p AND β AND THEIR RESPECTIVE 95% CONFIDENCE INTERVALS FOR THE CYCLE EXPERIMENT.

p	β
$0.079\,(0.026, 0.157)$	$1.807\,(1.316, 2.377)$

TABLE II

MEAN DIAGONAL AND OFF-DIAGONAL ESTIMATIONS FOR THE REPRODUCTION MATRIX R USING ESTIMATION MODELS AS WELL AS THEIR RESPECTIVE MEAN SQUARED ERROR (MSE) FOR THE TRAINING AND TESTING SETS FOR THE CYCLE EXPERIMENT OVER 1000 SIMULATIONS. REMIND THAT THE GROUND TRUTH VALUES FOR THE DIAGONAL AND OFF-DIAGONAL FOLLOW $\left(R_{\text{CYCLE}}^{\text{GROUND TRUTH}}\right)_{diag} = 0.8$ AND $\left(R_{\text{CYCLE}}^{\text{GROUND TRUTH}}\right)_{\text{OFF-DIAG}} = 0.05$, RESPECTIVELY. THE BOLDED VALUES ARE THE BEST ONES IN EACH COLUMN.

Method	***Train MSE***	***Test MSE***	***R Diagonal***	***R Off diagonal***
Linear	**2.305**	14.656	0.133	0.208
Ridge	2.727	6.228	0.471	0.100
Ridge Diag	2.687	6.099	0.746	0.029
Elastic Net	2.664	**6.073**	0.661	0.059
Elastic Net Diag	2.702	6.101	0.747	0.027
Staged1	2.554	7.523	**0.815**	**0.051**
Staged2	2.337	15.293	0.000	0.219

In Fig. 4, we plot the graph representations of the adjacency matrix for each method, using solid black edges when the estimate $R_{ij} > .1$ and dashed blue edges when $R_{ij} \in (.02, .1]$. Here we observe that linear regression results in a fully connected graph, and thus exhibits a high amount of false cross-excitation. Elastic net provides a graph estimate closest to the ground truth graph, with only one false edge.

In Table II, we report the mean value of the diagonal and off diagonal elements of the reproduction matrix R for the 1000 simulations, in addition to the mean squared error characterizing the different models used. As to be expected, Linear regression has the lowest training error due to the lack of parameter penalization. However, Elastic Net has the lowest test MSE. Staged1 is the best method at recovering the true model parameters for the reproduction matrix, thus there seems to be a tradeoff between Elastic Net and Staged1 regression.

B. Karate Club experiment

In the next experiment, we generate synthetic data from a MHP simulated on the karate club graph introduced by Zachary [35]. The network is formed from a study carried out from 1970-1972, where social interactions between 34 karate club members were recorded. As in Section VI-A, we generate MHP data on the karate club graph with $M = 34$ nodes and with the Weibull distribution described in Section II. The simulated events are generated within a time window of 300 days. For the reproduction matrix, we use the parameters:

$$\begin{aligned} \left(R_{\text{Karate club}}^{\text{Ground truth}}\right)_{diag} &= 0.15 \\ \left(R_{\text{Karate club}}^{\text{Ground truth}}\right)_{\text{off-diag}} &= 0.012. \end{aligned} \quad (18)$$

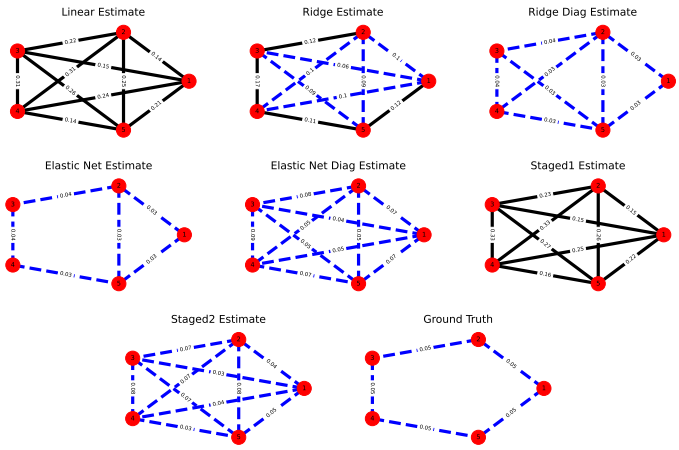

Fig. 4. Graphs corresponding to each estimation of R for the cycle experiment. The dashed lines characterize interactions of weight that belong the interval $(0.02, 0.1]$, while the straight black lines correspond to interactions with weight strictly larger than 0.1.

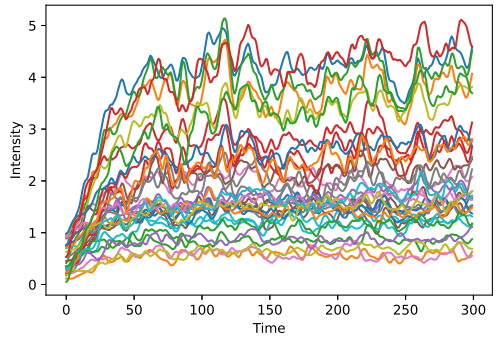

Fig. 5. Conditional intensity of simulated 34-dimesional Hawkes process with a Weibull distribution and reproduction matrix R yielded from Zachary's karate club graph.

The conditional intensity given by Equation (2) for one simulation is plotted in Fig. 5.

We first estimate the Weibull distribution parameters p and β using Stan, and the results are presented in Table III. We again note that the 95% confidence intervals contain the true values.

In Fig. 6, we plot the graph representations of the adjacency matrix for each method, using solid black edges when the estimate $R_{ij} > .1$ and dashed blue edges when $R_{ij} \in (.02, .1]$. We again observe that linear regression results in a near fully connected graph. As before, Elastic net provides a graph estimate closest to the ground truth graph, however still with a high degree of false cross-excitation. Thus even for moderate sized graphs, uncovering causal connections in MHPs may not be possible in certain parameter regimes, even with a high amount of regularization.

In Table IV, we report the mean value of the diagonal and off diagonal elements of the reproduction matrix R for the 100 simulations, in addition to the mean squared error

TABLE III
ESTIMATED VALUES OF THE WEIBULL DISTRIBUTION PARAMETERS p AND β AND THEIR RESPECTIVE 95% CONFIDENCE INTERVALS FOR THE ZACHARY'S KARATE CLUB EXPERIMENT.

p	β
$0.041\,(0.003, 0.142)$	$2.317\,(1.191, 3.571)$

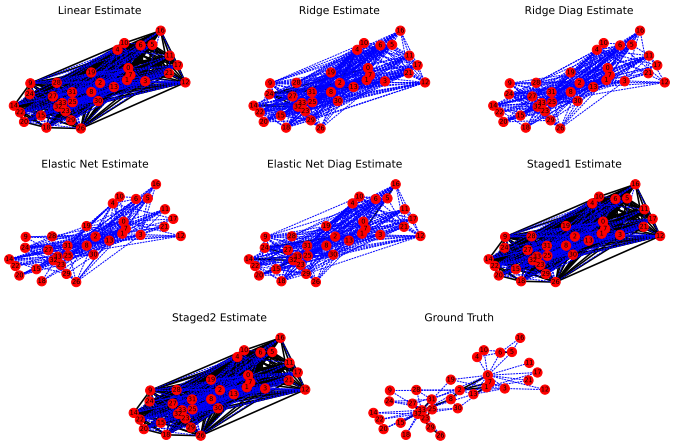

Fig. 6. Estimated reproduction matrix R for each method applied to the karate network MHP. The reproduction adjacency matrix R is visualized as a network using solid black edges when the estimate $R_{ij} > .1$ and dashed blue edges when $R_{ij} \in (.02, .1]$.

characterizing the different models used. Here we again observe that Elastic Net has the lowest test MSE. Again we see that a method that treats the diagonals and off-diagonals differently best recovers the true reproduction matrix, this time with diagonal Elastic Net having the best performance.

C. COVID-19 California County Data

In our last experiment, we fit MHPs to Coronavirus 2019 (Covid-19) daily infection data from four different counties in the state of California, USA during 2020. Data was obtained from the NY Times Covid-19 Github repository[2] . The counties include San Diego, San Francisco, Sacramento and San Jose and the daily infections were recorded for a period of 100 days after March 10, 2020. The evolution of the case count time series is shown in Fig. 7.

The CA COVID-19 data provides a good example of time series of contagion processes that are highly correlated, yet most transmission is realistically within each county rather than across counties during this time period.

We estimate the off-diagonal and diagonal elements of the reproduction matrix R using the different estimation models described above. In this case, R is a 4×4 matrix in which the off-diagonal elements $R_{ij},\ i \neq j$ represent the expected number of secondary infections in county i directly caused primary infections in county j. On the other hand, the diagonal R_{ii} represents the number of secondary infections in county i

[2]https://github.com/nytimes/covid-19-data

TABLE IV

MEAN DIAGONAL AND OFF-DIAGONAL VALUES FOR THE REPRODUCTION MATRIX R USING DIFFERENT ESTIMATION MODELS AND THEIR RESPECTIVE MEAN SQUARED ERROR (MSE) FOR THE TRAINING AND TESTING SETS OVER 100 SIMULATIONS FOR THE KARATE CLUB EXPERIMENT. THE GROUND TRUTH VALUES FOR THE DIAGONAL AND OFF-DIAGONAL FOLLOW $\left(R^{\text{GROUND TRUTH}}_{\text{KARATE CLUB}}\right)_{diag} = 0.15$ AND $\left(R^{\text{GROUND TRUTH}}_{\text{KARATE CLUB}}\right)_{\text{OFF-DIAG}} = 0.012$, RESPECTIVELY. THE BOLD VALUES REPRESENT THE BEST VALUE IN EACH COLUMN.

Method	***Train MSE***	***Test MSE***	***R Diagonal***	***R Off diagonal***
Linear	**0.337**	369.910	0.000	0.186
Ridge	0.737	1.630	0.044	0.034
Ridge Diag	0.599	1.623	0.123	0.022
Elastic Net	0.659	**1.597**	0.040	0.029
Elastic Net Diag	0.651	1.624	**0.157**	**0.015**
Staged1	0.364	347.433	0.295	0.173
Staged2	**0.337**	370.870	0.000	0.186

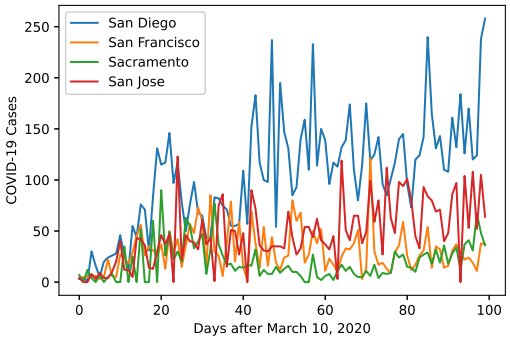

Fig. 7. Daily recorded new infected cases in the four California counties i.e., San Diego, San Francisco, Sacramento and San Jose over the period of 100 days. The recording started after March 10, 2020.

caused by a primary infection within the same county i, i.e., how the epidemic disease spreads from person to person within San Diego for instance. Table V shows the estimated Weibull parameters p and β using Stan, while Table VI summarizes the resulting mean values of R_{ij} $i \neq j$, as well as the corresponding MSE for each of the estimated models. As before, Elastic Net has the lowest test MSE on the data. While we do not know ground truth for R_{ij}, the 0.138 cross-transmission number for Elastic Net seems higher than what might be expected during the lock down. The 0.065 estimate given by diagonal Elastic Net is perhaps more plausible during this time period.

VII. DISCUSSION

In this paper we investigated the task of fitting mutlivariate Hawkes processes to discrete infectious disease count data on networks. In the case of infectious diseases, a shared triggering kernel is plausible and allows for regression based estimation through superimposing the MHPs. However, these shared triggering kernels are convolved with lagged event counts to generate features for the regression, and by construction these features are highly correlated. Thus multicollinearity and over-fitting become problems when fitting discrete MHPs to infectious disease data, especially as the largest eigenvalue

TABLE V

ESTIMATED VALUES OF THE WEIBULL DISTRIBUTION PARAMETERS p AND β AND THEIR RESPECTIVE 95% CONFIDENCE INTERVALS FOR THE CASE OF REAL COVID-19 DATA.

p	β
$0.0004\,(0.0001, 0.001)$	$4.477\,(3.843, 5.161)$

TABLE VI

MEAN VALUE OF DIAGONAL AND OFF-DIAGONAL VALUES FOR THE REPRODUCTION MATRIX R USING DIFFERENT ESTIMATION MODELS AND THEIR RESPECTIVE MEAN SQUARED ERROR (MSE) FOR THE TRAINING AND THE TESTING SETS FOR THE COVID-19 DATA.

Method	***Train MSE***	***Test MSE***	***R Diagonal***	***R Off diagonal***
Linear	**117.567**	4323.397	0.448	0.337
Ridge	183.470	809.936	0.411	0.175
Ridge Diag	182.832	800.446	0.647	0.071
Elastic Net	175.869	**797.403**	0.506	0.138
Elastic Net Diag	181.496	798.640	0.669	0.065
Staged1	121.448	3993.257	0.676	0.295
Staged2	120.373	4461.879	0.139	0.437

of the reproduction matrix exceeds 1 (which is the case during a pandemic). False cross-excitation is estimated by this regression approach, where transmission is predicted across geographies that might not actually exist.

We explored several methods for mitigating false cross-excitation. We find that standard regularization methods, in particular Elastic Net, produce the lowest MSE on held out data. Thus if the goal is short-term forecasting of infectious disease counts, this approach will yield the most accurate estimates. We note that these methods still have false cross-excitation, however it is possible that pooling case counts across nodes reduces variance (even though adding some bias) and may yield improved forecasts even when the nodes are independent.

However, if the goal is to recover true parameters of the infectious disease, then methods that treat diagonal and off-diagonal components of the reproduction matrix separately may yield more accurate estimates. Here we find that diagonal Elastic Net and Staged1 regression yielded the best results, depending on the dataset.

REFERENCES

[1] A. G. Hawkes, "Spectra of some self-exciting and mutually exciting point processes," *Biometrika*, vol. 58, no. 1, pp. 83–90, 1971.

[2] L. Adamopoulos, "Cluster models for earthquakes: Regional comparisons," *Journal of the International Association for Mathematical Geology*, vol. 8, pp. 463–475, 1976.

[3] Y. Ogata, "Space-time point-process models for earthquake occurrences," *Annals of the Institute of Statistical Mathematics*, vol. 50, pp. 379–402, 1998.

[4] Y. Aït-Sahalia, J. Cacho-Diaz, and R. J. Laeven, "Modeling financial contagion using mutually exciting jump processes," *Journal of Financial Economics*, vol. 117, no. 3, pp. 585–606, 2015.

[5] S. Azizpour, K. Giesecke, and G. Schwenkler, "Exploring the sources of default clustering," *Journal of Financial Economics*, vol. 129, no. 1, pp. 154–183, 2018.

[6] V. Chavez-Demoulin*, A. C. Davison, and A. J. McNeil, "Estimating value-at-risk: a point process approach," *Quantitative Finance*, vol. 5, no. 2, pp. 227–234, 2005.

[7] E. Errais, K. Giesecke, and L. R. Goldberg, "Affine point processes and portfolio credit risk," *SIAM Journal on Financial Mathematics*, vol. 1, no. 1, pp. 642–665, 2010.

[8] 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*, vol. 106, no. 493, pp. 100–108, 2011.

[9] J. Olinde and M. B. Short, "A self-limiting hawkes process: interpretation, estimation, and use in crime modeling," in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020, pp. 3212–3219.

[10] A. Jalaldoust, K. Hlaváčková-Schindler, and C. Plant, "Causal discovery in hawkes processes by minimum description length," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 6, 2022, pp. 6978–6987.

[11] M. Kim, D. Paini, and R. Jurdak, "Modeling stochastic processes in disease spread across a heterogeneous social system," *Proceedings of the National Academy of Sciences*, vol. 116, no. 2, pp. 401–406, 2019.

[12] E. Choi, N. Du, R. Chen, L. Song, and J. Sun, "Constructing disease network and temporal progression model via context-sensitive hawkes process," in *2015 IEEE International Conference on Data Mining*. IEEE, 2015, pp. 721–726.

[13] H. J. T. Unwin, I. Routledge, S. Flaxman, M.-A. Rizoiu, S. Lai, J. Cohen, D. J. Weiss, S. Mishra, and S. Bhatt, "Using hawkes processes to model imported and local malaria cases in near-elimination settings," *PLoS computational biology*, vol. 17, no. 4, p. e1008830, 2021.

[14] M. Eichler, R. Dahlhaus, and J. Dueck, "Graphical modeling for multivariate hawkes processes with nonparametric link functions," *Journal of Time Series Analysis*, vol. 38, no. 2, pp. 225–242, 2017.

[15] Á. Gajardo and H.-G. Müller, "Point process models for covid-19 cases and deaths," *Journal of Applied Statistics*, vol. 50, no. 11-12, pp. 2294–2309, 2023.

[16] B. Chen, P. Shrestha, A. L. Bertozzi, G. Mohler, and F. Schoenberg, "A novel point process model for covid-19: Multivariate recursive hawkes process," in *Predicting Pandemics in a Globally Connected World, Volume 1: Toward a Multiscale, Multidisciplinary Framework through Modeling and Simulation*. Springer, 2022, pp. 141–182.

[17] P. Giudici, P. Pagnottoni, and A. Spelta, "Network self-exciting point processes to measure health impacts of covid-19," *Journal of the Royal Statistical Society Series A: Statistics in Society*, vol. 186, no. 3, pp. 401–421, 2023.

[18] F. Schoenberg, "Some statistical problems involved in forecasting and estimating the spread of sars-cov-2 using hawkes point processes and seir models," *Environmental and Ecological Statistics*, vol. 30, no. 4, pp. 851–862, 2023.

[19] Z. Dong, S. Zhu, Y. Xie, J. Mateu, and F. J. Rodríguez-Cortés, "Non-stationary spatio-temporal point process modeling for high-resolution covid-19 data," *Journal of the Royal Statistical Society Series C: Applied Statistics*, vol. 72, no. 2, pp. 368–386, 2023.

[20] Á. Briz-Redón, A. Iftimi, J. Mateu, and C. Romero-García, "A mechanistic spatio-temporal modeling of covid-19 data," *Biometrical Journal*, vol. 65, no. 1, p. 2100318, 2023.

[21] A. L. Bertozzi, E. Franco, G. Mohler, M. B. Short, and D. Sledge, "The challenges of modeling and forecasting the spread of covid-19," *Proceedings of the National Academy of Sciences*, vol. 117, no. 29, pp. 16 732–16 738, 2020.

[22] S. Lamprinakou, A. Gandy, and E. McCoy, "Using a latent hawkes process for epidemiological modelling," *Plos one*, vol. 18, no. 3, p. e0281370, 2023.

[23] A. Bonnet, M. M. Herrera, and M. Sangnier, "Maximum likelihood estimation for hawkes processes with self-excitation or inhibition," *Statistics & Probability Letters*, vol. 179, p. 109214, 2021.

[24] Y. Ogata, "Estimators for stationary point processes," *Ann. Inst. Statist. Math*, vol. 30, no. Part A, pp. 243–261, 1978.

[25] F. P. Schoenberg, "A note on the consistent estimation of spatial-temporal point process parameters," *Statistica Sinica*, pp. 861–879, 2016.

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

[27] E. Bacry and J.-F. Muzy, "Second order statistics characterization of hawkes processes and non-parametric estimation," *arXiv preprint arXiv:1401.0903*, 2014.

[28] D. Marsan and O. Lengline, "Extending earthquakes' reach through cascading," *Science*, vol. 319, no. 5866, pp. 1076–1079, 2008.

[29] J. G. Rasmussen, "Bayesian inference for hawkes processes," *Methodology and Computing in Applied Probability*, vol. 15, pp. 623–642, 2013.

[30] G. Mohler, "Modeling and estimation of multi-source clustering in crime and security data," *The Annals of Applied Statistics*, pp. 1525–1539, 2013.

[31] M. Kirchner, "Hawkes and inar (∞) processes," *Stochastic Processes and their Applications*, vol. 126, no. 8, pp. 2494–2525, 2016.

[32] ——, "An estimation procedure for the hawkes process," *Quantitative Finance*, vol. 17, no. 4, pp. 571–595, 2017.

[33] H. Xu, D. Luo, X. Chen, and L. Carin, "Benefits from superposed hawkes processes," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2018, pp. 623–631.

[34] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell, "Stan: A probabilistic programming language," *Journal of statistical software*, vol. 76, 2017.

[35] W. W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of anthropological research*, vol. 33, no. 4, pp. 452–473, 1977.