Self-exciting spatio-temporal statistical models for count data with applications to modeling the spread of violence

by

Nicholas John Clark

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Program of Study Committee:
Philip M. Dixon, Major Professor
Ulrike Genschel
Mark Kaiser
Jarad Niemi
Zhengyuan Zhu

Iowa State University
Ames, Iowa
2018

DEDICATION

For Sarah, Piper, and Raegan.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I am extremely grateful for the hard work and dedication of my advisor, Dr. Philip Dixon, in making this dissertation possible. Our twice-weekly meetings, over the last two and a half years, were not only instrumental in shaping my vision for this work, but also quickly became the highlight of my week. I particularly appreciate the many conversations we had about 'telling a story' through our writing and translating ideas into words that others, potentially outside the field of statistics, would actually want to read.

I am also grateful to Profs Genschel, Kaiser, Niemi, and Zhu. Your questions, guidance, and encouragement helped grow a germ of an idea into a scholarly work. Individual conversations from each of you are interwoven into this work.

Going back to graduate school has long been a goal of mine. The camaraderie and professionalism of the faculty and the other graduate students has exceeded any of my expectations. I have genuinely looked forward to every day that I've spent in my cubicle and it has rarely felt like work and for this I am appreciative beyond what I could possibly convey.

ABSTRACT

In this dissertation we provide statistical models and inferential techniques for analyzing the number of violent or criminal events as they evolve over space and time. Our research focuses on a class of models we refer to as self-exciting spatio-temporal models. These are a class of parametric models that allow for dependence in a latent structure as well as dependence in the data model combining what is sometimes referred to as observation driven and parameter driven models. This class of models arise from straight-forward assumptions on how violence or crime evolves over space and time and has use in the statistical modeling of situations where there is an expected repeat or near-repeat victimization. In Chapter 2 we present the spatially correlated self-exciting model and the reaction-diffusion self-exciting model to analyze the number of violent events in different regions in Iraq. We also demonstrate how Laplace approximations can be used to conduct efficient Bayesian inference. We further show how the choice of the latent structure matters in this problem. In Chapter 3 we generalize the spatially correlated self-exciting model and show how it extends the classic integer generalized auto-regressive conditionally heteroskedastic, or INGARCH, model to account for spatial correlation and improves the second order properties of the INGARCH model. We refer to this new class of models as the spatially correlated INGARCH, or SP-INGARCH, model. We show how the spatially correlated self-exciting model is similar to the SPINGARCH(0,1) model. Finally in Chapter 4 we present a fast extended Laplace approximation algorithm for fitting the SPINGARCH(0,1) model demonstrating empirically how the extended Laplace approximation method reduces a bias that exists in the Laplace approximation method while performing much quicker than Markov Chain Monte Carlo approaches.

CHAPTER 1. Introduction

The statistical modeling of violence has generally neglected to properly account for spatial and temporal correlation. In Ratcliffe (2010), the argument is made that potential temporal correlation is often ignored in the statistical modeling of crimes whereas in examples such as Mohler (2013) spatial correlation is ignored. One of the principal challenges is that there are few available statistical models for count data that both offer a mechanism for capturing spatial-temporal correlation as well as result in meaningful inference. The commonly used Poisson-Log Normal model, for example, only allows limited data correlation as described in Aitchison and Ho (1989). Furthermore, this method of modeling assumes that, in the terminology of Cox et al. (1981), large scale structure in the model is parameter-driven. On the other hand, criminologists have shown the presence of repeat victimization in Johnson et al. (1997) and Johnson et al. (2007). To properly account for this phenomena, a statistical model should consider observation-driven structure, commonly referred to as self-excitement.

In this dissertation, we propose a new class of statistical models that accounts for self-excitement in the modeling of the spread of violence. Motivating our work is the assumption that violence at a given space-time region can arise both due to repeat victimization as well as exogenous factors. This assumption allows us to formulate a class of parametric model for the counts of violence that has flexible second order properties and can be fit using standard Bayesian software.

1.1 Spatio-temporal counts of violent events

The number of crimes or other violent events is usually aggregated over a fixed space-time lattice and presented as count data. For example, in this dissertaiton we use two primary datasets that are both aggregated due to privacy issues and convenience. The first dataset is from the Global Terrorism Database (LaFree and Dugan (2007)). We specifically look at the violent incidents in Iraq from 2003-2010 aggregated over province and month. Figure 1.1 depicts the monthly counts of US government reported violent incidents in Iraq aggregated over province during this time period. Here it appears that the violence over this time period spreads out from the center of the country.

We also use crime in Chicago freely available from the city of Chicago available at https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present. The data is aggregated over city block, we further aggregate it over census block group and month. The number of burglaries per census block per month in the south side of Chicago is shown in figure 1.2. Here we see that the spatial correlation is much less obvious than in Iraq, however as we will show it is still present in the data.

Though the spatio-temporal spread of violence in Iraq and Chicago appear to be very different, they both have been associated with the repeat-victimization or self-excitement. In Lewis et al. (2012) and Mohler (2013), civilian deaths were found to be partially attributed to self-excitement and in Mohler et al. (2011) and Mohler (2013) burglaries were also shown to be attributed to self-excitement. A further mathematical model for the spatio-temporal diffusion of burglaries presented in Short et al. (2008) also heavily relied on the notion of self-excitment.

1.2 Self-excitement in a statistical model

The models we present and use also account for self-excitement. The general idea of self-excitement in a statistical model is not new and have been extensively used in modeling phenomena from earthquakes in Ogata (1988) to finance (see e.g. Bacry et al. (2015)).

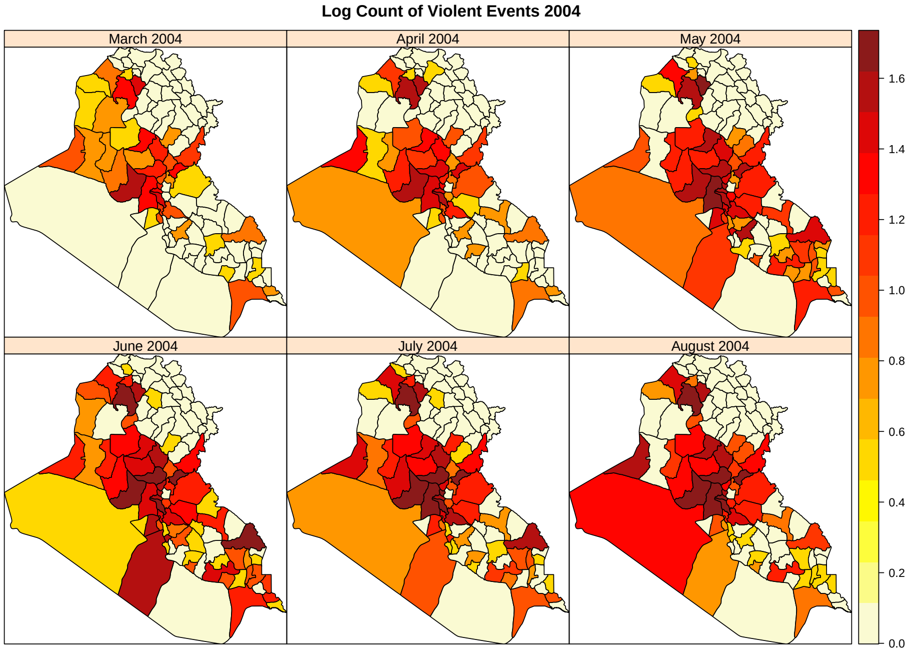

Figure 1.1 The spread of the log-count of violence in Iraq over discrete regions and discrete times.

Self-excitement, often referred to as Hawkes processes, originates from Hawkes (1971) work in point process data where the intensity of a Poisson process, $\lambda(t)$ is assumed to depend both on t as well as past realizations of the process

$$\lambda(t) = \nu + \int_0^t g(t-s)N(ds), \quad t > 0. \tag{1.1}$$

Here $g(t)$ can be thought of as a triggering function that generally decays over time. In this manner, an observed event today increases the probability of a subsequent event occurring the following day. In the modeling of violence this is the notion of repeat, or near repeat victimization shown to exist in burglaries in Johnson et al. (1997) and Johnson et al. (2007).

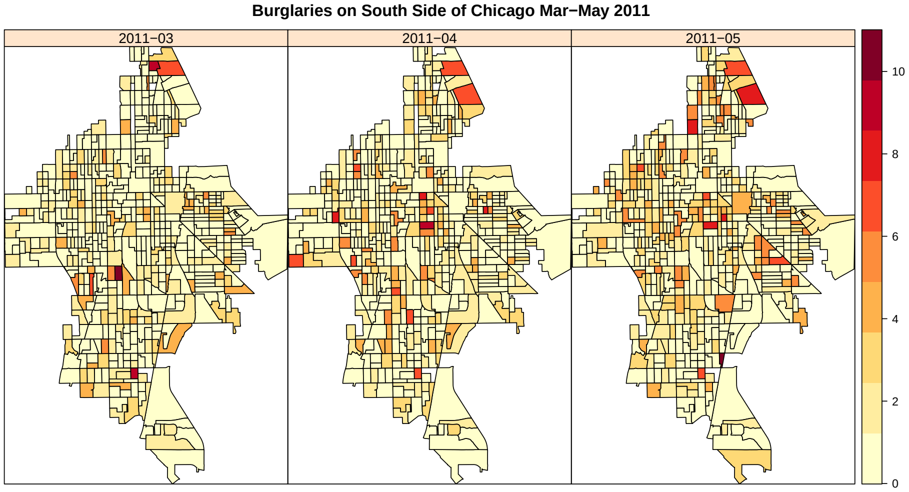

Figure 1.2 Spread of burglaries in the south side of Chicago in 2011.

The resulting statistical model using self-excitation is doubly stochastic as $\int_0^t g(t-s)N(ds)$ depends on a random variable, the number of events between time t and $t-s$.

However, as demonstrated in Mohler (2013), there may be other forms of latent variation in the intensity function. In that manuscript the authors demonstrated how the modeling of crimes in Chicago as well as violence in Iraq is improved through the addition of a latent auto-regressive term. In particular, they assumed that the number of violent events, y_t, between t and some fixed $t - \Delta t$ follows a Poisson distribution with expectation given by

$$\lambda_t = \exp(x_t) + \sum_{t>j} \eta \kappa^{t-j} y_j \tag{1.2}$$

$$\boldsymbol{x} \sim \text{Gau}\left(\mathbf{0}, \Sigma\right) \tag{1.3}$$

$$\Sigma_{t,j} = \sigma^2 a^{|t-j|}. \tag{1.4}$$

That is, the expected number of violent events is a linear combination of a discrete self-excitement term derived from an exponential kernel of a Hawkes process and a log-Gaussian Auto-Regressive (1) model. While this improved the fit of the data they considered, clearly spatial correlation was ignored.

1.3 Self-exciting spatio-temporal parametric models

Historically, the modeling of spatial-temporal count data has relied on the assumption that the log expected counts can be modeled as a latent Gaussian random variable (e.g. in Python et al. (2016)). In our belief this is overly restrictive and makes an realistic assumption that the counts are conditionally independent given a latent process and does not account for the possibility of self-excitement. Attempts at accounting for self-excitement generally fail to address the spatial correlation as in Mohler (2013) or estimate it non-parametrically as in Mohler et al. (2011).

In this dissertation we consider the impacts of combining self-excitement with spatial and spatio-temporal latent structures. In Chapter 2 we propose two different latent structures resulting in a spatially correlated self-exciting statistical model and a reaction diffusion self-exciting statistical model. We further demonstrate an efficient methodology for conducting Bayesian inference based on Laplace approximations and apply the methodology to the counts of violent events in Iraq.

In Chapter 3 we extend the spatially correlated self-exciting model and demonstrate similarities between this model and a relatively recent discrete valued time series model called the integer generalized autoregressive conditional heteroskedasticity, or INGARCH, model. In that spirit, we refer to the resulting model as the spatially correlated INGARCH, or SPINGARCH model. We demonstrate how the model can be fit using off the shelf software for Bayesian inference. We further show how the model out performs the INGARCH model in capturing the second order properties of the number of burglaries in the south side of Chicago.

Finally, in Chapter 4, we present an efficient methodology for Bayesian inference of the spatially correlated self-exciting model (also referred to as the SPINGARCH(0,1) model) based on an extended Laplace approximation. We show how for a large range of the parameter space the extended Laplace approximation matches Markov chain Monte Carlo (MCMC) based techniques while providing a drastic computational speedup.

CHAPTER 2. Modeling and Estimation for Self-Exciting Spatio-temporal Models of Terrorist Activity

Preprint of a manuscript accepted in *The Annals of Applied Statistics.*

Nicholas J. Clark, Philip M. Dixon

Please cite Clark and Dixon (2018)

Abstract

Spatio-temporal hierarchical modeling is an extremely attractive way to model the spread of crime or terrorism data over a given region, especially when the observations are counts and must be modeled discretely. The spatio-temporal diffusion is placed, as a matter of convenience, in the process model allowing for straightforward estimation of the diffusion parameters through Bayesian techniques. However, this method of modeling does not allow for the existence of self-excitation, or a temporal data model dependency, that has been shown to exist in criminal and terrorism data. In this manuscript we will use existing theories on how violence spreads to create models that allow for both spatio-temporal diffusion in the process model as well as temporal diffusion, or self-excitation, in the data model. We will further demonstrate how Laplace approximations similar to their use in Integrated Nested Laplace Approximation can be used to quickly and accurately conduct inference of self-exciting spatio-temporal models allowing practitioners a new way of fitting and comparing multiple process models. We will illustrate this approach by fitting a self-exciting spatio-temporal model to terrorism data in Iraq and demonstrate how choice of process model leads to differing conclusions on the existence of self-excitation in the data

and differing conclusions on how violence spread spatially-temporally in that country from 2003-2010.

2.1 Introduction

A typical spatio-temporal model consists of three levels, a data model, a process model, and a parameter model. A common way to model data then is to assume the data model, $Y(\cdot)$, is conditionally independent given the process model $X(\cdot)$. For example, if observations take place at areal regions, $\boldsymbol{s_i}$, at discrete time periods, t, and $Y(\boldsymbol{s_i}, t)$ are counts, a common model is $Y(\boldsymbol{s_i}, t)|X(\boldsymbol{s_i}, t) \sim \text{Pois}(\exp(X(\boldsymbol{s_i}, t)))$. The spatio-temporal diffusion structure is commonly then placed on the process model which commonly is assumed to have a Gaussian joint distribution of $\boldsymbol{X} \sim \text{Gaus}(\boldsymbol{0}, Q^{-1}(\theta))$. The majority of analysis of these models is done using Bayesian techniques requiring a further parameter model for θ. The challenge in these models is, then, determining an appropriate structure for $\boldsymbol{Q}^{-1}(\theta)$ or $\boldsymbol{Q}(\theta)$. If both the covariance and the precision is chosen to be too dense inference quickly becomes impossible due to the size of $\boldsymbol{Q}^{-1}(\theta)$. In spatio-temporal models it is quite common for the dimension of $\boldsymbol{Q}$ to be larger than, say, $10^4 \times 10^4$. A thorough overview giving many examples of this method of modeling is given in Cressie and Wikle (2015).

In modeling terrorism or crime data one possibility is to use an extremely general spatio-temporal process model to capture variance not explained through the use of covariates. For example Python et al. (2016) use a Matern class covariance function over space and an AR(1) process over time. They then use covariates to test the impact of infrastructure, population, and governance. The general spatio-temporal process models used, in this case, has an extremely sparse precision structure greatly aiding in computations.

While diffusion in spatio-temporal models is often modeled through a latent process, more recent models describing the spread of violence have incorporated self-excitation, or spatio-temporal diffusion that exists linearly in the data model itself. Self-excitation is the theory that in terrorism, or crime, the probability of an event occurring is a function of previous successful events. For instance Mohler et al. (2011) demonstrate that burglars are more likely to visit locations that have previously, successfully, been burgled. Mohler (2013) derived a class of models that allowed for temporal diffusion in both the process model as well as the data model and demonstrated how the two processes were identifiable.

In the modeling of terrorism data Lewis et al. (2012), Porter et al. (2012), and Mohler (2013) have all successfully used the self-excitation approach to model. Most recently, Tench et al. (2016) used a likelihood approach for temporal modeling of IEDs in Northern Ireland using self-excitation. However, in these papers, the existence and analysis of self-excitation was the primary objective and any process model dependency was either ignored or treated as a nuisance. The one exception is in Mohler (2013) where a temporal only model was assumed for the process model and inference was conducted allowing both process model dependence and data model dependence.

In this manuscript, we will consider a spatial and a spatial-temporal process model that allows for self-excitation. We will present two self-exciting models for terrorist activity that have different process models corresponding to different notions of how terrorism evolves in time and space as well as temporal dependency in the data model to account for self-excitation. These two models are specific cases of more general spatio-temporal models that allow dependency in both the process model as well as the data model.

We will further show how Laplace approximations similar to their use in Integrated Nested Laplace Approximation, or INLA, an approximate Bayesian method due to Rue et al. (2009) can be used to conduct inference for these types of models. We will show, via simulation, how INLA, when appropriately modified, can accurately be used to make inference on process level parameters for self-exciting models and aid analysts in determining the appropriate process model when scientific knowledge cannot be directly applied as in Cressie and Wikle (2015). Finally, we will apply this technique to terrorism data in Iraq. We will show that choice of process model, in this case, results in differing conclusions on the impact of self-excitation in the model.

2.2 Self-Exciting Spatio-Temporal Models

The use of self-exciting models in both criminal and terrorism modeling has become increasingly popular over the last decade after being originally introduced in Short et al. (2008). Self-excitement, in a statistical model, directly models copy-cat behavior by letting an observed event increase the intensity (or excites a model) over a specified time or location. Self-exciting models are closely related to Hawkes processes, which are counting processes where the probability of an event occurring is directly related to the number of events that previously occurred. In a self-exciting model, the criminal intensity at a given spatio-temporal location, (x, y, t) is a mixture of a background rate, ν and self excitement function, $f(\mathcal{H}_{x,y,t})$ that is dependent on the observed history at that location, $\mathcal{H}_{x,y,t}$.

A common temporal version of a discretized Hawkes process is

$$\begin{aligned} &Y_t \sim \text{Pois}(\lambda_t) && (2.1)\\ &\lambda_t = \nu(t) + \sum_{j<t} \kappa(t-j)y_j \\ &t \in \{1, 2, ...T\} && (2.2) \end{aligned}$$

In this example, in order for the process to have finite expectation in the limit, $\kappa(t-j)$ must be positive and $\sum_{i=1}^{\infty} \kappa(i) < 1$. $\kappa(t-j)$ can be thought of as the probability that an event at time j triggers an event at time i.

Laub et al. (2015) provides an excellent overview of the mathematical properties of the continuous Hawkes process and the discrete process when $\kappa(t-j)$ is taken to be an exponential decay function.

In Mohler et al. (2011), ν was treated as separable in space and time and was non-parametrically estimated using stochastic declustering, while in Mohler (2013), the spatial correlations were ignored and an AR(1) process was used for ν and an exponential decay was assumed for the self-excitation. In terrorism modeling Lewis et al. (2012) used a piece-wise linear function for ν.

Here we will first define a general model that allows spatial or spatial-temporal correlation to exist in the process model and positive temporal correlation to exist in the data model to allow for self-excitation. First define $\boldsymbol{s_i} \in \mathbb{R}^2$, $i \in \{1, 2, ..., n_d\}$ as locations in a fixed, areal, region. We further define $t \in \{1, 2, ..., T\}$ as discrete time. The general form of a spatial-temporal self-exciting model is then given in (2.3)

$$\begin{aligned} &Y(\boldsymbol{s_i}, t)|\mu(\boldsymbol{s_i}, t) \sim \text{Pois}(\mu(\boldsymbol{s_i}, t)) \\ &\mu(\boldsymbol{s_i}, t) = \exp(X(\boldsymbol{s_i}, t)) + \eta Y(\boldsymbol{s_i}, t-1) \\ &\boldsymbol{X} \sim \text{Gau}(\boldsymbol{0}, Q^{-1}(\theta)) \end{aligned} \tag{2.3}$$

Comparing the above to the Hawkes process, we now have ν as a function of space-time and denote it as $X(\boldsymbol{s_i}, t)$. We use the simplest form of self-excitation letting $\kappa(t-j)$ be a point-mass function such that $\kappa(k) = \eta$ for $k = 1$ and $\kappa(k) = 0$ for $k \neq 1$. In all cases $Y(\boldsymbol{s_i}, t)$ will be discrete, observable, count data.

To contrast (2.3) with a typical spatial model, figure 2.1 depicts the expectation for one areal location $(\boldsymbol{s_i}, t)$ without self-excitation and with self-excitation as shown in Figure 2.1. In this figure, the lower line shows $\mu(\boldsymbol{s_i}, t)$ with $\eta = 0$, and the upper line has $\eta = .4$. The impact of self-excitation is clearly present in time 10-13.

2.2.1 Spatially Correlated Self-Exciting Model

In the first example of (2.3) we assume the background intensity rate, $X(\boldsymbol{s_i}, t)$ has only spatial correlation. This model is motivated through the assumption that the latent dependency, $X(\boldsymbol{s_i}, t)$, is as a continuous measure of violent tendency at region $\boldsymbol{s_i}$ at time period t and regions that are closer together in space are assumed to share common characteristics.

Next, define $N(\boldsymbol{s_i})$ as the neighborhood of location $\boldsymbol{s_i}$ where two regions are assumed to be neighbors if they share a common border. $|N(\boldsymbol{s_i})|$ is the number of neighbors of location $\boldsymbol{s_i}$. The model for $Y(\boldsymbol{s_i}, t)$, or the number of observed violent events at a given space-time location is then given by:

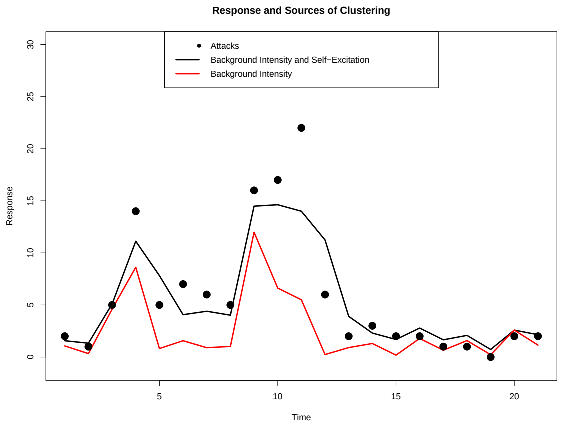

Figure 2.1 This figure shows an example of the expectation of two processes, one with self-excitation and one without. The bottom line is the expectation of a process with no self-excitation, the top has self-excitation of $\eta = .4$. The data realizations are from the process with self-excitation.

$$\begin{aligned}
&Y(\boldsymbol{s_i},t)|\mu(\boldsymbol{s_i},t) \sim \text{Pois}(\mu(\boldsymbol{s_i},t)) \qquad (2.4)\\
&\mu(\boldsymbol{s_i},t) = \exp(X(\boldsymbol{s_i},t)) + \eta Y(\boldsymbol{s_i},t-1)\\
&X(\boldsymbol{s_i},t) = \theta_1 \sum_{\boldsymbol{s_j}\in N(\boldsymbol{s_i})} X(\boldsymbol{s_j},t) + \epsilon(\boldsymbol{s_i},t)\\
&\epsilon(\boldsymbol{s_i},t) \sim \text{Gau}(0,\sigma^2)
\end{aligned}$$

Letting $\boldsymbol{H}$ denote the spatial neighborhood matrix such that $H_{i,j} = H_{j,i} = 1$ if $\boldsymbol{s_i}$ and $\boldsymbol{s_j}$ are neighbors, the full distribution of the joint distribution of the latent state is $\boldsymbol{X} \sim \text{Gau}(\boldsymbol{0},(\boldsymbol{I}_{ns,ns} - \boldsymbol{I}_{n,n} \otimes \theta_1\boldsymbol{H})^{-1}\boldsymbol{L}(\boldsymbol{I}_{ns,ns} - \boldsymbol{I}_{n,n} \otimes \theta_1\boldsymbol{H})^{-1})$ where $\boldsymbol{L} = \text{diag}(\sigma^2,...,\sigma^2)$. The evolution in the latent field is equivalent to the spatial evolution in what is commonly referred to as a Simultaneous or Spatial Auto-regressive model (SAR). Alternatively, the

Conditional Auto-regressive model (CAR) of Besag (1974) could be used to model the latent state modifying the joint density above.

The difference between the SAR and (2.4) is in the self excitement parameter, η. In (2.4), temporal correlation is present, but is present through the data model specification rather than through a temporal evolution in the latent state. Therefore, temporal correlation is a function solely of the self-excitation in the process. η gives the probability of an event at time $t-1$ creating an event at time t. In order for the system to be well-behaved, η is constrained to (0,1). In order for the joint distribution of $\boldsymbol{X}$ to be valid, $\theta_1 \in (\psi_{(1)}^{-1}, \psi_{(n)}^{-1})$ where ψ_i is the ith smallest eigenvalues of $\boldsymbol{H}$.

The critical assumption in this model is that the propensity of a given location to be violent is spatially correlated with its adjacent spatial neighbors and only evolves over time as a function of excitation. If terrorism is diffusing according to this model, regions that are geographically adjacent are behaving in a similar manner. The existence of self-excitation would indicate that individuals within a region are being inspired through the actions of others. While combating terrorism is complex, if terrorism is diffusing in this manner, one suggestion would be to identify the root causes within a geographic area as well as quick action against any malicious actor to discourage copy-cat behavior.

2.2.2 Reaction Diffusion Self-Exciting Model

Alternatively, a model similar to Short et al. (2008) can be used to motivate the process model for the latent state resulting in a non-separable spatio-temporal, $\boldsymbol{X}$. Here we let $X(\boldsymbol{s_i}, t)$ corresponds to a continuous measure of violence due to terrorists or criminals at location $\boldsymbol{s_i}$ at time t. This is still a latent variable as we are not directly measuring $X(\boldsymbol{s_i}, t)$. However, now in order for an area to increase in violent tendency, a neighboring area must decrease as the actors causing the violence move from location to location. Furthermore, if terrorists are removed from the battlefield at a rate proportional to the total number of terrorists present and if terrorists move to fill power vacuums, the process model is similar to the reaction-diffusion partial differential equation (see Cressie and Wikle (2015) for more on the reaction-diffusion model)

$$\frac{\partial X(\boldsymbol{s_i},t)}{\partial t} = \frac{\kappa}{|N(\boldsymbol{s_i})|}\triangle X(\boldsymbol{s_i},t) - \alpha X(\boldsymbol{s_i},t) \tag{2.5}$$

In order to generalize this partial differential equation (PDE) to an irregular lattice, we make use of the graphical Laplacian, Γ, in place of $\triangle$ in (2.5). Γ is a matrix that extends the notion of second derivatives to irregular graphs and can be defined as a matrix of the same dimension as the number of geographical regions with entries given by

$$\Gamma(s_i,s_j)\begin{cases} -|N(s_i)| & j=i \\ 1 & j \in N(s_i) \\ 0 & \text{Otherwise} \end{cases}$$

With the addition of a random noise term assumed to be Gaussian, the full model can be seen as an example of (2.3).

$$\begin{aligned} &Y(\boldsymbol{s_i},t)|\mu(\boldsymbol{s_i},t) \sim \text{Pois}(\mu(\boldsymbol{s_i},t)) \\ &\mu(\boldsymbol{s_i},t) = \exp(X(\boldsymbol{s_i},t)) + \eta Y(\boldsymbol{s_i},t-1) \\ &X(\boldsymbol{s_i},t) = \frac{\kappa}{|N(s_i)|}\sum_{\boldsymbol{s_j}\in N(\boldsymbol{s_i})} X(\boldsymbol{s_j},t-1) + (1-\kappa-\alpha)X(\boldsymbol{s_i},t-1) + \epsilon(\boldsymbol{s},t) \\ &\epsilon(\boldsymbol{s},t) \sim \text{Gau}(0,\sigma^2) \end{aligned} \tag{2.6}$$

In contrast to the Spatially Correlated Self-Exciting (SCSE) Model, the process model dependency exists in both space and time. In order to derive properties of this model we first let $\boldsymbol{M} = \kappa \text{ diag}\left(\frac{1}{|N_{s_i}|}\right)\Gamma + (1-\alpha)\boldsymbol{I}_{s,s}$ and now note that this is equivalent to a Vector Auto-Regressive, VAR, model $\boldsymbol{X}_t = \boldsymbol{M}\boldsymbol{X}_{t-1} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \text{Gau}(\boldsymbol{0},\sigma^2\boldsymbol{I})$.

The VAR(1) model requires all the eigenvalues of $\boldsymbol{M}$ to be between -1 and 1. This can be satisfied by first noting that 0 is always an eigenvalue of diag $\left(\frac{1}{|N_{s_i}|}\right)$ trivially corresponding to the eigenvector of all 1s. The largest eigenvalue is at most 2 as shown in Chung (1997). Due to the structure of $(1-\alpha)\boldsymbol{I}_{s,s}$ this implies maximum eigenvalue of $\boldsymbol{M}$ is $(1-\alpha)$ and minimum is $-2\beta + (1-\alpha)$. Therefore, the parameter spaces for α and κ are $\alpha \in (0,1)$ and $\kappa \in (\frac{-\alpha}{2}, \frac{2-\alpha}{2})$.

Just as in the SCSE Model, if ϵ has a Gaussian distribution, the Reaction Diffusion Self-Exciting (RDSE) Model has an exact solution for the latent Gaussian field, $\boldsymbol{X}$.

Letting Σ_s be the spatial covariance at a fixed period of time which is assumed to be invariant to time , then we can solve for Σ_s by using the relationship $\Sigma_s = \boldsymbol{M}\Sigma_s\boldsymbol{M^T} + \sigma^2\boldsymbol{I}$. As demonstrated by Cressie and Wikle (2015), this leads to

$$\mathrm{vec}(\Sigma_s) = \left(\boldsymbol{I}_{s^2,s^2} - \boldsymbol{M}\otimes\boldsymbol{M}\right)^{-1}\mathrm{vec}\left(\sigma^2\boldsymbol{I}_{s,s}\right), \tag{2.7}$$

where vec () is the matrix operator that stacks each column of the matrix on top of one or another. Recall that s is the size of the lattice that is observed at each time period. The joint distribution for all $\boldsymbol{X}$ is then $\boldsymbol{X} \sim \mathrm{Gau}(\boldsymbol{0}, Q_{rd}^{-1}(\theta))$ where

$$Q_{rd}^{-1}(\theta) = \left[\begin{array}{c|c|c|c} \Sigma_s & M\Sigma_s & ... & M^n\Sigma_s \\ \hline \Sigma_s M^T & \Sigma_s & ... & M^{n-1}\Sigma_s \\ \hline ... & ... & ... & ... \\ \hline \Sigma_s(M^T)^n & \Sigma_s(M^T)^{n-1} & ... & \Sigma_s \end{array}\right] \tag{2.8}$$

.

However, practically, this involves inverting a potentially large matrix $\boldsymbol{I}_{s^2,s^2} - \boldsymbol{M}\otimes\boldsymbol{M}$. Therefore, it is easier to deal with the inverse of (2.8) given in (2.9).

$$Q_{rd}(\theta) = \left[\begin{array}{c|c|c|c|c} \boldsymbol{I}_{n,n} & -M & \boldsymbol{0} & ... & ... \\ \hline -M^T & M^TM + \boldsymbol{I}_{n,n} & -M & \boldsymbol{0} & ... \\ \hline \boldsymbol{0} & -M^T & M^TM + \boldsymbol{I}_{n,n} & -M & ... \\ \hline ... & ... & ... & ... & ... \\ \hline \boldsymbol{0} & ... & -M^T & M^TM + \boldsymbol{I}_{n,n} & -M \\ \hline \boldsymbol{0} & ... & ... & -M^T & \boldsymbol{I}_{n,n} \end{array}\right]\frac{1}{\sigma^2} \tag{2.9}$$

The primary difference between the SCSE model and the RDSE model is that the process model correlation in the SCSE is only spatial while in the RDSE it is spatio-temporal. In the below toy examples, we show the expectation for $X(s_i, t)$ for both the SCSE and the RDSE model on a 4 x 4 lattice structure. We fixed both models with a value of $X(s_1, 1) = 10$ as the upper left hand observation at time 1. As seen in the RDSE model, the high count at time 1 spreads to neighboring regions in time 2 and time 3 whereas the process model has no temporal spread in the SCSE but has a high level of spatial spread.

Spatially Correlated Latent Process Conditional on $(s_1, 1) = 10$

Time 1				Time 2			
10	5	2	1	0	0	0	0
5	4	3	2	0	0	0	0
2	3	2	1	0	0	0	0
1	1	1	1	0	0	0	0

Reaction Diffusion Latent Process Conditional on $(s_1, 1) = 10$

Time 1				Time 2				Time 3			
10	1	1	0	3	3	0	0	2	1	1	0
1	1	0	0	3	1	0	0	1	1	0	0
1	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0

Practically, if data follows the RDSE model, it implies a high terrorism count in one region will manifest into a high terrorism count in a neighboring region at a later time period. In combating terrorism, the RDSE might suggest isolating geographical regions to mitigate the risk of spread while addressing self-excitation through direct action against malicious actors who are inspiring others.

2.3 Model Fitting

In both the RDSE and the SCSE, spatio-temporal diffusion exists in both the process model and the data model. If the diffusion was solely in the process model, a technique for inference would be Integrated Nested Laplace Approximation, or INLA.

INLA was first proposed in Rue et al. (2009) to specifically address the issue of Bayesian Inference of high dimensional Latent Gaussian Random Fields, LGRFs. An example of this for count data is:

$$\begin{aligned} Y(s_i) &\sim \text{Pois}(\mu(s_i, t)) \qquad (2.10)\\ \mu(\boldsymbol{s_i}, \boldsymbol{t}) &= \exp(\lambda(\boldsymbol{s_i}, t))\\ \lambda(\boldsymbol{s_i}, t) &= \beta_0 + \boldsymbol{Z}^t\boldsymbol{\beta} + X(\boldsymbol{s_i}, \boldsymbol{t})\\ \boldsymbol{X} &\sim \text{Gau}(\boldsymbol{0}, Q^{-1}(\theta)) \end{aligned}$$

INLA is often preferable over MCMC for these types of models. An issue with traditional Markov Chain Monte Carlo (MCMC) techniques for these models is that the dimension of X is often very large. Therefore, while MCMC has $O_p(N^{-1/2})$ errors, the N in the errors is the simulated sample size for the posterior. Just getting $N = 1$ may be extremely difficult due to the vast number of elements of X that need to be estimated. In general, MCMC will take hours or days in order to successfully simulate from the posterior making the computational cost of fitting multiple process models extremely high. In Python et al. (2016), terrorism data was fit using a grid over the entire planet using INLA, though without self-excitation in the model.

To address the issues with MCMC use in LGRFs, Rue et al. (2009) developed a deterministic approach based on multiple Laplacian approximations. A LGRF is any density that can be expressed as

$$\pi(\boldsymbol{\theta}, \boldsymbol{X}|\boldsymbol{Y}) \propto \pi(\theta)|Q(\boldsymbol{\theta})|^{1/2} \exp\left[\frac{-1}{2}\boldsymbol{X}^t Q(\boldsymbol{\theta})\boldsymbol{X} + \sum_{\boldsymbol{s}} \log\left(\pi(Y(\boldsymbol{s_i})|X(\boldsymbol{s_i}), \boldsymbol{\theta})\right)\right] \qquad (2.11)$$

In order to conduct inference on this model, we need to estimate $\pi(\boldsymbol{\theta}|\boldsymbol{y})$, $\pi(\theta_i|\boldsymbol{y})$ and $\pi(x_i|\boldsymbol{y})$. The main tool Rue et al. (2009) employ is given in their equation (3) as

$$\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{Y}) \propto \frac{\pi(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{Y})}{\tilde{\pi}_G(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{Y})}|_{X=x^*(\boldsymbol{\theta})} \qquad (2.12)$$

In Rue et al. (2009) they note that the denominator of (2.12) almost always appears to be unimodal and approximately Gaussian. The authors then propose to use a Gaussian approximation to $\pi(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{Y})$ which is denoted above as $\tilde{\pi}_G$. Moreover, (2.12) should hold no matter what choice of $\boldsymbol{X}$ is used, so a convenient choice for $\boldsymbol{X}$ is the mode for a given θ, which Rue et al. (2009) denote as $x^*(\boldsymbol{\theta})$.

Now, $\pi(\boldsymbol{\theta}|\boldsymbol{Y})$ can be explored by calculating the marginal for choices of θ, which if chosen carefully can greatly decrease the computational time. These explored values can then be numerically integrated out to get credible intervals for $\pi(\theta_i|\boldsymbol{Y})$.

Following the exploration of $\theta|Y$, and computation of $\theta_i|Y$, INLA next proceeds to approximate $\pi(X(s_i)|\boldsymbol{\theta}, \boldsymbol{Y})$. The easiest way to accomplish this is to use the marginals that can be derived straightforwardly from $\tilde{\pi}_G(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{Y})$ from (2.12). In this manuscript we will use this technique for simplicity of computation, however, if the latent states are of interest in the problem (and they often are), this can be problematic as it fails to capture any skewness of the posterior of $\boldsymbol{X}$. One way to correct this is to re-apply (2.12) in the following manner:

$$\tilde{\pi}_{LA}(X(\boldsymbol{s_i})|\theta, y) \propto \frac{\pi(\boldsymbol{X}, \theta, \boldsymbol{Y})}{\tilde{\pi}_G(\boldsymbol{X}_{-s_i}|X(s_i), \boldsymbol{\theta}, \boldsymbol{Y})}|_{x_{-i}=\boldsymbol{x_{-i}}^*(x_i,\theta)} \tag{2.13}$$

In (2.13) $\boldsymbol{X}_{-s_i}$ is used to represent $\boldsymbol{X}$ with latent variable $X(s_i)$ removed. This is a reapplication of Tierney and Kadane's marginal posterior density and gives rise to the nested term in INLA.

2.3.1 Laplace Approximation for Spatio-Temporal Self-Exciting Models

While INLA is an attractive technique due to computational speed and implementation, it is not immediately usable for the SCSE and the RDSE as the structure in (2.3) is

$$\begin{aligned} \mu(\boldsymbol{s_i}, t) &= \exp(X(\boldsymbol{s_i}, t)) + \eta Y(\boldsymbol{s_i}, t-1) \\ \eta &\in (0,1) \end{aligned} \tag{2.14}$$

In this structure, $X(.)$ and $Y(.)$ are not linearly related and a Gaussian prior for η is clearly not appropriate due to the parameter space constraints.

However, Laplace approximations can still be used by conducting inference on η at the same time inference is conducted on the the set of latent model parameters. In both the Spatially Correlated Self-Exciting Model and the Reaction Diffusion Self-Exciting model, the full conditional for the latent state is

$$\pi(\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}\boldsymbol{X}^T\boldsymbol{Q(\theta)X} + \sum_{s_i,t}\log\pi\left(Y(\boldsymbol{s_i},t)|X(\boldsymbol{s_i},t),\eta,Y(\boldsymbol{s_i},t-1)\right)\right) \tag{2.15}$$

Here we will let $\boldsymbol{\theta} = (\theta_1, \sigma^2, \eta)^T$ and $\boldsymbol{Q_{sc}(\theta)} = (\boldsymbol{I}_{sn,sn} - \theta_1\boldsymbol{I}_{t,t} \otimes \boldsymbol{H})$ for the Spatially Correlated Self-Exciting Model and use $\boldsymbol{Q_{rd}(\theta)}$ for the RDSEM defined in (2.9).

While $\boldsymbol{\theta}$ in (2.15) does not contain η we next do a Taylor series expansion of

$$\log\pi\left(Y(\boldsymbol{s_i},t)|X(\boldsymbol{s_i},t),\eta,Y(\boldsymbol{s_i},t-1)\right),$$

as a function of $X(\boldsymbol{s_i},t)$ and, for each $\boldsymbol{s_i},t$, expand the term about a guess for the mode, say $\mu_0(\boldsymbol{s_i},t)$. First we write $\boldsymbol{B^*}(\boldsymbol{\theta}|\mu_0)$ as a vector of the same length as $X(\boldsymbol{s_i},t)$ where each element is given by

$$B(\boldsymbol{s_i},t|\mu_0) = \left(\frac{\partial\log\pi\left(Y(\boldsymbol{s_i},t)\right)}{\partial X(\boldsymbol{s_i},t)}\Big|_{X(\boldsymbol{s_i},t)=\mu(\boldsymbol{s_i},t)} - \mu(\boldsymbol{s_i},t)\frac{\partial^2\log\pi\left(Y(\boldsymbol{s_i},t)\right)}{\partial X(\boldsymbol{s_i},t)^2}\Big|_{X(\boldsymbol{s_i},t)=\mu(\boldsymbol{s_i},t)}\right) \tag{2.16}$$

Next, we further define $\boldsymbol{Q^*(\theta)}|\mu_0$ as the updated precision matrix.

$$\boldsymbol{Q^*(\theta)|\mu_0} = \boldsymbol{Q(\theta)} + \text{diag}\left(-\frac{\partial^2\log\pi\left(Y(\boldsymbol{s_i},t)\right)}{\partial X(\boldsymbol{s_i},t)^2}\right)\Big|_{X(\boldsymbol{s_i},t)=\mu(\boldsymbol{s_i},t)} \tag{2.17}$$

Where $\boldsymbol{Q(\theta)}$ is either $\boldsymbol{Q_{sc}(\theta)}$ or $\boldsymbol{Q_{rd}(\theta)}$ depending on the context. Then we can write

$$\pi(\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}\boldsymbol{X}^T\left(\boldsymbol{Q^*(\theta)|\mu_0}\right)\boldsymbol{X} + \boldsymbol{X}^T\left(\boldsymbol{B^*(\theta)|\mu_0}\right)\right) \tag{2.18}$$

While in (2.17), $\boldsymbol{Q(\theta)}$, the original precision matrix, does not contain η, $\boldsymbol{Q^*(\theta)}$, the updated precision matrix, does depend on the self-excitation parameter.

Next we find the values of $\mu(\boldsymbol{s_i})$ that maximize (2.18). This is done through the use of an iterative maximization algorithm by solving for $\boldsymbol{\mu_1}$ in $(Q^*(\boldsymbol{\theta})|\boldsymbol{\mu_0})\,\boldsymbol{\mu_1} = \boldsymbol{B^*}(\boldsymbol{\theta}|\mu_0)$. For a fixed $\boldsymbol{\theta}$, this converges rapidly, due to the sparsity of both $\boldsymbol{Q_{sc}}$ and $\boldsymbol{Q_{rd}}$. .

In (2.12), for a fixed $\boldsymbol{\theta}$, we can then find $x^*(\boldsymbol{\theta})$. When the denominator of (2.12) is evaluated at $x^*(\boldsymbol{\theta})$ it becomes $|\boldsymbol{Q^*(\theta)}\frac{1}{2\pi}|^{1/2}$ which is equivalent to the hyperparameter inference recommended by Lee and Nelder (1996) as pointed out by R. A. Rigby in Rue et al. (2009).

In order to best explore $\pi(\boldsymbol{\theta}|\boldsymbol{Y})$ the posterior mode is first found through a Newton-Raphson based method. In order to do this we approximate the Hessian matrix based off of finite difference approximation to the second derivatives.

After locating the posterior mode of $\pi(\boldsymbol{\theta}|\boldsymbol{Y})$, the parameter space can be explored using the exploration strategy laid out in section 3.1 of Rue et al. (2009).

Now, for the set of diffusion parameters, $\boldsymbol{\theta}$ which contain η, we have a method of estimating $\pi(\boldsymbol{\theta}|\boldsymbol{Y})$. Inference for any further data model covariates can now be conducted in the same manner as done in Rue et al. (2009).

2.3.2 Model Comparison and Goodness of Fit

In order to conduct model comparison, we will use the deviance information criterion (DIC) originally proposed by Spiegelhalter et al. (2002). Goodness of fit will be conducted through the use of posterior predictive p-values, outlined by Gelman et al. (1996).

To approximate the DIC, we first find the effective number of parameters for a given $\boldsymbol{\theta}$. As noted in Rue et al. (2009), we can estimate this by using $n - \text{tr}\left(\boldsymbol{Q}(\boldsymbol{\theta})\boldsymbol{Q}^*(\boldsymbol{\theta})^{-1}\right)$ for both the SCSEM and the RDSEM. This gives the effective number parameters for a given $\boldsymbol{\theta}$, which can then be averaged over $\pi(\boldsymbol{\theta}|\boldsymbol{Y})$ to get the effective number of parameters for the model.

Secondly, we calculate the deviance of the mean

$$-2\sum_{s_i,t}\log\pi\left(Y(s_i,t)|\hat{X}(s_i,t),\boldsymbol{\theta^*}\right) \tag{2.19}$$

where $\boldsymbol{\theta^*}$ is the posterior mode and $\hat{X}(s_i,t)$ is the expectation of the latent state fixing $\theta = \theta^*$. DIC can then be found through deviance of the mean plus two times the effective number of parameters as in chapter 7 of Gelman et al. (2014) and initially recommended by Spiegelhalter et al. (2002).

In order to assess goodness of fit in analyzing the terrorism data in Section 5, we will use posterior predictive P-values as described by Gelman et al. (1996). Here, we pick critical components of the original dataset that we wish to see if the fitted model can accurately replicate, for instance the number of zeros in the dataset which we can designate as $T(\boldsymbol{Y})$. Next, for an index $m = 1...M$, We then draw a value of $\boldsymbol{\theta_m}$ according to $\pi(\boldsymbol{\theta}|\boldsymbol{Y})$ and simulate a set of observations $Y^*(\boldsymbol{s_i, t})_m$ of the same dimension as $\boldsymbol{Y}$ and compute $T(\boldsymbol{Y}^*_m)$. This process is repeated M times and a posterior predictive p-value is computed as $\frac{1}{M}\sum_{m=1}^{M} I\left[T(\boldsymbol{Y}^*_m) > T(\boldsymbol{Y})\right]$ where $I\left[.\right]$ is the indicator function. While not a true P-value, both high and low values of the posterior predictive p-value should cause concern over the fitted models ability to replicate features of the original dataset.

2.4 Simulation

In order to validate the Laplace based methodology for spatially correlated self-exciting models we conducted simulation studies using data on a 8 by 8 Spatial grid assuming a rook neighborhood structure. In order to decrease the edge effect, we wrapped the grid on a torus so each node had four neighbors. For each grid location we simulated 100 observations, creating a spatio-temporal model that had 6400 observations, meaning in (2.12), $\boldsymbol{Q}(\boldsymbol{\theta})$ had a dimension of 6400×6400.

In the first simulation we used (2.4) fixing the parameters at values that generated data that appeared to resemble the data from Iraq used in Section 5. The generating model we used was:

$$\begin{aligned}
&Y(\boldsymbol{s_i}, t)|\mu(\boldsymbol{s_i}, t) \sim \text{Pois}\ (\mu(\boldsymbol{s_i}, t)) \qquad (2.20)\\
&\mu(\boldsymbol{s_i}, t) = \exp(-1 + X(\boldsymbol{s_i}, t)) + .2Y(\boldsymbol{s_i}, t-1)\\
&X(\boldsymbol{s_i}, t) = .22 \sum_{\boldsymbol{s_j} \in N(\boldsymbol{s_i})} X(\boldsymbol{s_j}, t) + \epsilon(\boldsymbol{s_i}, t)\\
&\epsilon(\boldsymbol{s_i}, t) \sim \text{Gau}(0, .4)
\end{aligned}$$

The spatial parameter for model was $\theta_1 = .22$ which suggests a positive correlation between spatially adjacent locations. An η value of 0.2 would suggest that each event that occurs at one time period increases the expected number of events at the next time period by .2. Here we fix σ^2 was fixed at 0.4 and use a value of $\beta_0 = -1$ to reflect that in real world applications the latent process likely is not zero mean.

Once the data were generated, we found $\pi(\boldsymbol{\theta}|\boldsymbol{Y})$ by applying (2.12). Here we note that the numerator of (2.12) is $\pi(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{Y}) = \pi(\boldsymbol{Y}|\boldsymbol{X}, \eta, \boldsymbol{\theta})\pi(\eta)\pi(\boldsymbol{X}|\theta_1, \sigma^2)\pi(\theta_1)\pi(\sigma^2)$ which requires a prior specification for η,θ_1, and σ. In order to reflect an a-priori lack of knowledge we choose vague priors for all parameters. In this model, we use a Half-Cauchy with scale parameter of 25 for σ and a Uniform $(\psi_{(1)}^{-1}, \psi_{(n)}^{-1})$ where $\psi_{(i)}$ is the ith largest eigen value of the spatial neighborhood. As we used a shared-border, or rook, neighbor structure wrapped on a torus, the parameter space is (-0.25,0.25) as each spatial location has four neighbors. The choice of the Half-Cauchy is in line with the recommendations for vague priors for variance components of hierarchical models as outlined in Gelman et al. (2006) and rigorously defended in Polson et al. (2012). We let the prior for η be Uniform(0,1).

Using a gradient descent method with step-halving we found the posterior mode of $\pi(\boldsymbol{\theta}|\boldsymbol{Y})$ to be $\sigma^2 = 0.32$, $\theta_1 = 0.22$, and $\eta = 0.20$. Using the z based parameterization described in Section 3.1 we next explored the parameterization $\log \pi(\boldsymbol{\theta}|\boldsymbol{Y})$ and found credible intervals of $\pi(\sigma^2|\boldsymbol{Y}) = (0.29, 0.36), \pi(\theta_1|\boldsymbol{Y}) = (0.22, 0.23)$ and $\pi(\eta|\boldsymbol{Y}) = (.18, .21)$. Fixing $\boldsymbol{\theta}$ at the posterior mode, we then found an approximate 95% credible interval for β_0 to be (-1.67,.07).

The posterior maximum and credible interval for σ^2 appear to be slightly lower than expected, but the remaining parameter credible intervals covered the generating parameter.

Next we simulated from the reaction-diffusion self-excitation model letting $\beta_0 = 0$, $\alpha = 0.1$, $\kappa = 0.2$, $\sigma^2 = 0.25$, and $\eta = 0.4$

In fitting the model, we again use vague priors for all the parameters. Again, we place a Half-Cauchy prior on σ^2 as described above. In order to conform to the parameter space of α and κ, we let $\pi(\alpha) \sim \text{Unif}\ (0, 1)$ and $\pi(\kappa|\alpha) \sim \text{Unif}\ (-\frac{\alpha}{2}, 1 - \frac{\alpha}{2})$.

Again using the Laplace approximation technique of section 3, we found the posterior mode of $\pi(\boldsymbol{\theta}|\boldsymbol{Y})$ to be at $\alpha = 0.085$, $\kappa = 0.19$, $\sigma^2 = 0.21$, $\eta = 0.35$. 95% credible intervals for the posterior marginals were $\alpha \in (0.07, 0.10)$, $\kappa \in (0.14, 0.24)$, $\sigma^2 \in (0.18, 0.24)$, and $\eta \in (0.32, 0.40)$. At the posterior mode of $\boldsymbol{\theta}$, the posterior marginal for β_0 was approximately (-0.03,0.01). Critically, if there is self-excitement in the data, in all simulations it was differentiable from the latent diffusion. This is a spatial-temporal analogue to the finding in Mohler (2013) where a temporal AR(1) process was differentiable from self-excitement.

In our simulations, the approximations described in this manuscript performed reasonably well for inference on the spatio-temporal diffusion parameters in most cases. However, when σ^2 is large, or when η is large, we have found that the approximations create bias in one or more of the parameters likely due to the high effective number of parameters. However, all approximate likelihood based methods will likely struggle in these cases as well. As noted in Rue et al. (2009), the approximation error in Laplace based methods is related to the number of effective latent variables over the total sample size.

2.5 Spatio-Temporal Diffusion of Violence in Iraq (2003-2010)

2.5.1 Statistical Models and Data

One region where the reasons for the diffusion of terror and crime still remains unclear is in Iraq during 2003 to 2010. While violence undoubtedly spread throughout the country, it remains unclear how or why, spatio-temporally, the spread occurred. Part of the uncertainty is that there still is not agreement over whether violence was due to insurgency, civil war, or organized crime. For example, Hoffman (2006) refers to the violence in Iraq as an insurgency, Fearon (2007) argues that the spread of violence was due to a civil war, and Williams (2009) argues that there was a large presence of organized crime in the country.

A few previous studies have examined the presence or absence of self-excitation. In Lewis et al. (2012), the authors concluded that self-excitation was present in select cities in Iraq during this time period. The presence of the self-excitation finding was echoed in Braithwaite and Johnson (2015) where the authors also noted a correlation between locations that shared microscale infrastructure similarities. This would suggest repeat or near-repeat actions were causing the increase in violence in a region.

However, in both of these cases, the structural form of the latent spatio-temporal diffusion was, a-priori, assumed to be known. In fact, this is likely not the case. In a classic work on the subject, Midlarsky et al. (1980) discuss how heterogeneity between locations can cause correlation in violence or individuals who cause violence can actually physically move from one location to another. In particular, if violence is strictly due to crime we would expect self-excitement and limited diffusion between geographical regions. Whereas if violence is due to insurgencies we would expect more movement of actors as they seek to create widespread disruption in the country. The former theory is reflected in the Spatial Correlation in the Spatially Correlated Self-Exciting model and the later theory would correspond to the Reaction Diffusion component of the second model.

The overarching goal of this analysis, thus, is to determine whether in Iraq the growth of violence in fixed locations was due to the presence or absence of self-excitation. Furthermore, we want to determine whether the latent diffusion of violence is due to the movement of population such as in the Reaction Diffusion model, or whether there is static spatial correlation. We will answer this while controlling for exogenous factors that may also explain the rise in violence in a region.

In order to address this question as well as demonstrate how Laplace based approximations can be used to fit real world data to models of the class of (2.10) we used data from the Global Terrorism Database (GTD) introduced in LaFree and Dugan (2007) to examine the competing theories. The GTD defines terrorist events as events that are intentional, entail violence, and are perpetrated by sub-national actors. Additionally, the event must be aimed at obtaining a political, social, religions, or economic goal and must be conducted in order to coerce or intimidate a larger actor outside of the victim. The majority of lethal events in Iraq from 2003-2010 fit the above category.

The GTD uses a variety of open media sources to capture both spatial and temporal data on terroristic events. The database contains information on what the event was, where it took place, when it took place, and what terrorist group was responsible for the event. From 2003 to 2010 in Iraq, the database contains 6263 terrorist events, the spatial structure is shown in figure 2.2.

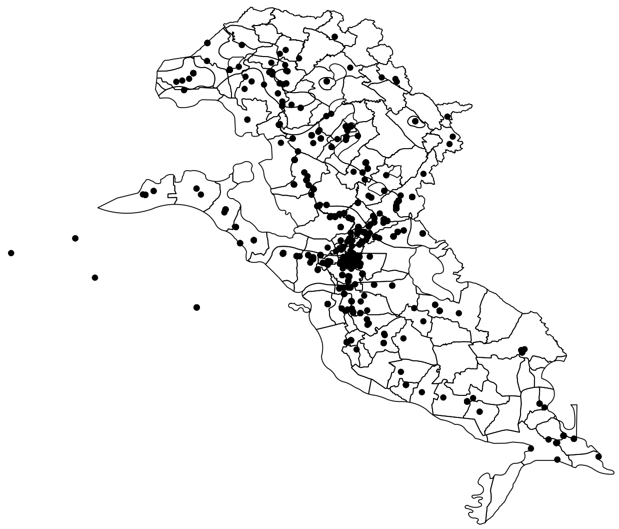

Figure 2.2 Spatial depiction of 6263 events in Iraq.

As seen in this map, the majority of the violence is in heavily populated areas such as Baghdad and in the regions north up the Tigris river to Mosul and west through the Euphrates river. In order to model this data, we aggregated the point data to 155 political districts intersected by ethnicities and aggregated the data monthly for 96 months meaning that $\Sigma(\theta)$ in (2.3) is a 14880 x 14880 matrix. Population for each political district was taken from the Empirical Studies of Conflict Project website, available from https://esoc.princeton.edu/files/ethnicity-study-ethnic-composition-district-level.

We considered covariates controlling for the population density within a fixed region as well as for the underlying ethnicity. We will make the simplifying assumption that both of these are static over time. Previous statistical analysis on terrorism considered macro level covariates, such as democracy in Python et al. (2016) that differ country to country but would not change within a single country as analyzed here. Other studies considered more micro level covariates such as road networks that were found to be statistically related to terrorism in Braithwaite and Johnson (2015). Here we take the view point that the vast majority of the incidents in Iraq were directed against individuals rather than terrorist events directed at fixed locations. Therefore, we would expect a higher population density to provide more targets for a potential terrorist to attack. Furthermore, covariates such as road networks, number of police, or number of US soldiers, would all be highly collinear with population density. We do, though, consider a covariate for ethnicity in a region. Specifically, we add an indicator if the region is predominately Sunni. The disenfranchisement of the Sunnis and high level of violence in Sunni dominated areas has been well established, see for e.g. Baker III et al. (2006). Previous research in Linke et al. (2012) focusing on Granger Causality also suggested indicators for majority Sunni/Shia may be appropriate in any analysis of violence in Iraq.

$$\begin{aligned} &Y(\boldsymbol{s_i},t)|\mu(\boldsymbol{s_i},t) \sim \text{Pois }(\mu(\boldsymbol{s_i},t)) \\ &\mu(\boldsymbol{s_i},t) = \exp[\beta_0 + \beta_1 \log \text{Pop}(\boldsymbol{s_i}) + \beta_2 \text{Sunni }(\boldsymbol{s_i}) + X(\boldsymbol{s_i},t)] + \eta Y(\boldsymbol{s_i},t-1) \\ &\boldsymbol{X} \sim \text{Gau}(\boldsymbol{0}, \boldsymbol{Q}^{-1}(\theta)) \end{aligned} \tag{2.21}$$

The complete statistical model is given in (2.21). We next fit this model letting $\boldsymbol{Q} = \boldsymbol{Q_{sc}}$ and $\boldsymbol{Q} = \boldsymbol{Q_{rd}}$. We further consider fixing $\eta = 0$ to test the presence or absence of self-excitement in the data for both process models.

2.5.2 Results

We fit all four models using the Laplace approximation method described in Section 3.1. For each of the parameters we used vague proper priors to ensure posterior validity. In the SCSEM and the Spatially Correlated models we used a Half-Cauchy with scale parameter of 5 for σ and a Uniform $(\psi_1^{-1}, \psi_n^{-1})$ where ψ are the eigenvalues of $\boldsymbol{H}$. Using the neighborhood structure corresponding to the geographical regions described above, this corresponded to a Uniform (-.3,.13). For each of the exogenous covariates, we used independent Gaussian (0,1000) priors. For the SCSEM model we further assumed a Uniform (0,1) prior for η.

In fitting the RDSEM, we again used a Half-Cauchy with scale parameter of 5 for σ. For the decay parameter, α, we used a Uniform (0,1) prior and for the diffusion parameter, κ, we chose a Uniform $\left(\frac{-\alpha}{2}, 1 - \frac{\alpha}{2}\right)$ in order to ensure we were in the allowable parameter space.

All four models took approximately 30 min to an hour depending on starting values to converge using a Newton-Raphson based algorithm to find the maximum. Gaussian approximations to the 95% credible intervals for the parameters are given for all four models are shown in table 2.1. As can be seen in comparing the SCSEM to the RDSEM, the presence or absence of self-excitation appears to be dependent on the choice of structure of $\boldsymbol{Q}$. Furthermore, the impact of majority Sunni is also dependent on whether the Reaction Diffusion or Spatially Correlated model was used.

Table 2.1 95% Credible Intervals for Model Parameters

	Spatial Correlation Only	SCSEM	Reation Diffusion Only	RDSEM
β_0	(-20.2,-18.4)	(-16, -15.5)	(-22.4,-18.0)	(-22,-18.6)
β_1	(1.2,1.4)	(0.9, 1.1)	(1.1, 1.4)	(1.1, 1.5)
β_2	(1.6,1.9)	(1.1,1.3)	(0.10, 0.33)	(0.17, 0.53)
η	-	(0.35,0.37)	-	(0, 0.04)
σ^2	(1.9,2.7)	(2.1, 2.5)	(0.20, 0.30)	(0.18,0.26)
θ_1	(.08,.10)	(0.09,0.1)	-	-
α	-	-	(0.001, 0.007)	(0.001, 0.007)
κ	-	-	(0.03, 0.07)	(0.03, 0.06)

Table 2.2 Model Assessment and Selection Statistics For Iraq Data

Model	DIC	P-Values Maximum Value	P-Value Zeros
Spatially Correlated Model without Self-Excitation	9370	.02	1
Spatially Correlated Self-Exciting Model	8722	.97	0
Reaction Diffusion Only Model	8664	.53	.81
Reaction Diffusion Self-Exciting Model	8699	.45	.89

Using the methodology described in Section 3.2, we next calculated DIC as well as posterior predictive P-values based off of the maximum observed value and the number of zeros in the dataset. In the original dataset, the maximum number of events observed for all regions and months was 26 and the dataset had 13445 month/district observations that were zero. The model assessment and selection results are shown in Table 2.2.

Clearly from table 2.2, the models with an underlying reaction diffusion process model outperform those with spatial correlation only. Furthermore, the addition of self-excitation in the model appears to have minimal impact. In particular, without self-excitation, the spatial correlation model tends to under count the number of violent activities while the

SCSEM tends to over count. There really is not much difference between the RDSEM and the reaction diffusion model so we prefer the simper reaction diffusion only model. While β_2 is only minimally significant in the reaction diffusion model, the models perform better including the covariate than disregarding it entirely.

2.5.3 Significance

Under all measures of performance, the reaction diffusion model, (2.6), without self-excitation outperforms the other models under consideration. This process model as well as values of the covariates and the lack of self-excitement in the data offer several insights into the causes of the spread of violence in Iraq.

Not surprisingly, the reaction diffusion model has a positive relationship between log population and violence. As the majority of attacks are directed at individuals it would clearly follow that regions that have higher population will offer more targets as well as more potential combatants. Further, higher populated areas also would have had higher number of Iraqi government officials as well as US military presence. The positive, though small, relationship between Sunni and violence is also not surprising as, in general, predominately Sunni areas were generally more disenfranchised following the transition to a new Iraqi government after the downfall of Saddam Hussein.

More significantly, though, was that the reaction diffusion model fit the data better than the SCSEM or the RDSEM. This suggests that increases in violence can be attributed, at least in part, to movement between high violence and low violence areas rather than repeat or near-repeat actors in a fixed location.

While the κ parameter may appear small in 2.1 it still has an impact on the process. For the sake of simplicity, we can demonstrate this on a three node system. For this system we consider a central node that has a high level of violence surrounded by two nodes that have a low level of violence. In this set up we will fix $\beta_0 = -19$, $\beta_1 = 1.3$ and consider each node as having a population of 1000 and let $\kappa \in \{.03, .07\}$. The resultant system over time is shown in Figure 2.3. Even in this simple system, there is a noticeable increase in violence as the center node diffuses throughout the entire system.

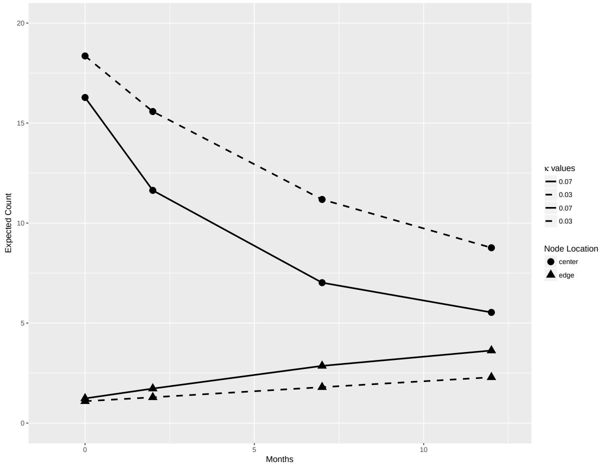

Figure 2.3 This plot shows the expected changes in violence in a simple three node system where the center node starts with a high level of violence and the other two nodes start with a low level of violence for $\kappa = .03$, depicted as dashed lines in above figure, and $\kappa = .07$, depicted as straight lines. As seen here after 12 months for $\kappa = .07$ the nodes are essentially at equilibrium.

The implications of the reaction-diffusion model being preferable over the SCSEM or the spatial correlation only model can be seen by going back to the original PDE that inspired the model, (2.5). The underlying assumption in that model is that violence is that the rate of violence spreading to a region is determined through the levels of violence in neighboring region. From a military planning perspective, this would suggest that if there is a peaceful region surrounded by areas of high violence, the peaceful region should be isolated to prevent the movement of malicious actors. This strategy would be consistent with published military strategy as outlined in Army and Corps (2006).

Finally, this offers insight into the nature of the conflict that was fought in Iraq. For instance, Zhukov (2012) discuss how insurgencies diffuse throughout a population by either physical movement of actors or through movement of ideas, whereas Short et al. (2008) suggest that criminal violence would be expected to have an element of self-excitation.

When accounting for the possibility of spatial diffusion, this self-excitation does not appear to be present in the Iraqi dataset. Therefore, as a diffusion based model fits the data better, this would suggest the spread of violence was due to the physical movement of an insurgent population or ideology rather than a criminal element that would be expected to stay more static at a location.

2.6 Discussion

In this manuscript we develop statistical models that allow for spatio-temporal diffusion in the process model and temporal diffusion in the data model. We relate the models to existing theory in how violence diffuses in space and time. We further developed a Laplace approximation for spatio-temporal models that contain self-excitement. This modification allows for a quick and accurate fit to commonly used models in both the analysis of terrorism and criminology. A critical difference between classical INLA and our proposal is that in our proposal, inference is not only performed on the hyperparameters during the exploration of $\pi(\theta|Y)$ in (2.12) but also on the self-excitation parameter η. While η is not generally thought of as a hyperparameter, when the linear expansion of the log-likelihood is done in (2.18), η enters into $\boldsymbol{Q^*}$ and $\boldsymbol{B^*}$ in a similar manner as the hyperparameters.

While we only considered two process models and self-excitation that only exists for one time period, the methodology outlined above can easily be extended to allow self-excitation to have an exponential decay similar to the modeling technique of Mohler et al. (2011). As shown above, the absence of testing multiple process models may result in premature conclusions about how violence is spreading over regions. While self-excitation may be present in one model, its significance may be dulled through the use of alternative process models resulting in differing conclusions.

Although self-excitation has become increasingly popular, alternate approaches based on Besag's auto-logistic model, as used in Weidmann and Ward (2010) are possible, though care must be taken if count data is used as Besag's auto-Poisson does not permit positive dependency. As shown in Kaiser and Cressie (1997), a Winsorized Poisson must be used if positive dependency is desired, as it most certainly is in terrorism modeling. In this case, the data model dependency would linearly be associated with the log of $\mu(\boldsymbol{s_i}, t)$.

Though the motivation for the models in this manuscript was the spatio-temporal spread of violence, the novel concept of combining latent process dependency and data model dependency has the potential to be used in other fields. For example in the modeling of thunderstorms, self-excitation may be present temporally, while process model dependency may also be appropriate due to small-scale, unobservable, spatial or spatio-temporal dependency. Laplace approximations, as demonstrated in this manuscript, allow for quick and relatively accurate methods to fit multiple types of self-exciting spatio-temporal models for initial inference.

CHAPTER 3. A Spatially Correlated Auto-Regressive Model For Count Data

A paper to be submitted to *The Journal of the American Statistical Association*

Abstract

The statistical modeling of multivariate count data observed on a space-time lattice has generally focused on using a hierarchical modeling approach where space-time correlation structure is placed on a continuous, unobservable, process. The data distribution is then assumed to be conditionally independent given the latent process. However, in many real-world applications, especially in the modeling of criminal or terrorism data, this conditional independence between the observed counts may be inappropriate. In the absence of spatial correlation, the Integer Auto-Regressive Conditionally Heteroskedastic (INGARCH) process could be used to capture this data model dependence however this model does not allow for any unexplained spatial correlation in the data. In this manuscript we propose a class of models that extends the INGARCH process to account for small scale spatial variation, which we refer to as a SPINGARCH process. The resulting model allows both data model dependence as well as dependence in a latent structure. We demonstrate how second-order properties can be used to differentiate between models in this class. Finally, we apply Bayesian inference for the SPINGARCH process demonstrating its use in modeling the spatio-temporal structure of burglaries in Chicago from 2010-2015 and demonstrate how accounting for spatial correlation changes the conclusion on the existence of repeat victimization.

3.1 Introduction

The modeling of count data where each observation takes place on a space-time lattice arises in multiple disciplines. In the disease literature, the number of infected patients is often aggregated over geographic areas and discrete times to protect the confidentiality of patients. In the modeling of terrorism or criminal acts, that we consider in this manuscript, data is often presented aggregated over time and space for security reasons. Even for spatial continuous and temporally continuous data, the analysis is often performed aggregated over fixed spatial and temporal domains as a matter of convenience. The challenge, then, is how to appropriately model the relationship between observations. Assumptions on either the spatial relationship or the temporal relationship between observations are necessary if any statistical analysis is to be performed. In this paper we present a novel approach for structuring space-time dependency for count data through a combination of spatial dependence in a latent, process model, and temporal dependence in the data model.

In the spatial statistics literature, an early attempt at structuring spatial relationships for count data was made in Besag (1974) where the data model distribution was conditionally specified given a fixed spatial neighborhood. However, as shown in Besag (1974), this results in a statistical model that only allows negative correlation. More recently, Kaiser and Cressie (1997) demonstrated how modifications could be made to the statistical model that allowed both negative and positive correlation. A similar methodology was employed in Augustin et al. (2006) to address the spatial dynamics of seed count data in agricultural models. The critical assumption in these classes of models is the distribution of the observed counts can be conditionally specified from the observed counts at spatial neighbors, a Markov assumption in space.

Advances in computation and Bayesian inference have also allowed for modeling through spatial hierarchical models similar to the Poisson log-Normal approach of Aitchison and Ho (1989). Letting $\boldsymbol{s_i}$ be a discrete spatial location and t be discrete time, spatial-temporal dependence can be introduced by assuming the existence of a latent process,

$Y(\boldsymbol{s_i}, t) \sim \text{Gau}(\boldsymbol{\alpha}, \Sigma(\theta))$ that has a spatial-temporal structure characterized by $\Sigma(\theta)$. The data model is then assumed to be independent given the latent state. The idea was extended to incorporate spatial dynamics in Besag et al. (1991). Here a spatial Markov assumption was still made, but it was made in a latent, unobserved, continuous process. The spatial observations were then assumed to be independent given the latent process. This idea was also used in Wolpert and Ickstadt (1998), who used a Poisson-gamma model with spatial dependence in the latent, gamma, structure.

The concept of allowing the spatial dependence to exist only in a latent field overcomes the difficulty of only negative spatial correlation that arises in the auto-Poisson model of Besag (1974). Although this form of modeling only allows for limited dependence in the data as demonstrated in Aitchison and Ho (1989), it has become commonplace in literature. For example in Goicoa et al. (2016) mortality rates are studied using latent conditional effects for space, time, and age. This approach also has given rise to specialized analytical techniques and software to conduct efficient inference, for example Integrated Nested Laplace Approximations of Rue et al. (2009) or spBayes of Finley et al. (2007). However, this approach assumes that the dependence in the data is due to a latent, unobservable process which does not capture repeat victimization that is believed to exist in violence as explained for example in Polvi et al. (1990). Repeat victimization is the belief that an observed crime or violent act increases the likelihood of a future crime occurring at that exact same spot or against the exact same person and can be modeled assuming a data model dependence or as an observation driven process.

While count data in the spatial statistics literature has predominately been addressed through structure in a latent process, in the time series literature it has evolved quite differently. For example, the INGARCH model of Ferland et al. (2006) and Heinen (2003) is a time series model for counts where the data model is Poisson with expectation that is a function of both previous counts and previous expectations. Specifically, if we let Z_t be a time series of counts and $\mathcal{F}_t$ be the σ-field generated by $Z_0, ..., Z_t, \lambda_0$, the INGARCH(p, q)

model is

$$Z_t|\mathcal{F}_{t-1} \sim \text{Pois}(\lambda_t), \quad \lambda_t = d + \sum_{i=1}^{p} a_i \lambda_{t-i} + \sum_{j=1}^{q} b_j Z_{t-j} \tag{3.1}$$

This results in a time series model that is a function of both the data model and a deterministic process model. Ferland et al. (2006) demonstrated how the INGARCH(1,1), given as $\lambda(s_i, t) = d + \kappa\lambda(s_i, t-1) + \eta Z(s_i, t-1)$, is analogous to an ARMA(1,1) for counts. In Fokianos et al. (2009) it was shown that a perturbed INGARCH(1,1) model was geometrically ergodic giving a unique stationary distribution and asymptotic normality of the roots of the likelihood equations. The stationary distribution of the INGARCH(1,1) process is also equivalent to a stochastic process given in Hawkes (1971), often called a self-exciting point process, where

$$\begin{aligned} Z_t|\lambda_t &\sim \text{Pois}(\lambda_t) \\ \lambda(t) &= \nu(t) + \int_0^t g(t-u)N(ds), \end{aligned} \tag{3.2}$$

when the process is sampled at discrete times and $g(t-u) = \eta \exp(-\alpha(t-t_i))$.

While the INGARCH model was motivated to model univariate time series data, there has been some recent effort to apply it to multivariate count data. Heinen and Rengifo (2007) used copulas to model the contemporaneous correlation. However there are issues with using copulas for count data and it is generally less reliable and identifiable than the use of copulas for continuous data, as explained in Genest and Nešlehová (2007). Liu (2012) allows for a spatial lag dependency through treating λ_t as a vector and replacing a_i and b_j with a series of matrices. The author then allows for contemporaneous correlation through the bivariate Poisson. These models, though, do not naturally extend previous spatial models to the INGARCH class nor do they capture how criminologists and others believe crime actually evolves over space and time. Furthermore, as we will show, the variance to mean ratio for the INGARCH process dictates the range for the allowable autocorrelation limiting its practical use.

In this manuscript, we introduce a class of Self-Exciting Spatio-Temporal models for count data we refer to as Spatially Correlated Integer Auto-Regressive Conditionally Heteroskedastic, or SPINGARCH, models. These models maintain many of the stationarity properties of the INGARCH process while allowing for spatial correlation through the addition of a latent log-Gaussian spatially correlated process. We will demonstrate how the models arise from assumptions on how crime and violence evolves over space and time, how they retain the same stationarity properties as the INGARCH model and how they can be differentiated through assessment of second order characteristics. The SPINGARCH model also allows a much wider range of second order properties affording the modeler more flexibility in describing the autocorrelation and variance to mean ratio of the data. We will further show how to conduct inference and model assessment to differentiate between models within this class using burglaries in Chicago as an exemplar.

3.2 General Model

A self-exciting spatio-temporal model is characterized through the existence of a latent process that is allowed to have spatial correlation as well as a data process that has self-excitation, or a positive feedback mechanism. Denote $Z(\boldsymbol{s_i}, t)$ as the data model with $\boldsymbol{s_i} \in \{\boldsymbol{s_1}, \cdots, \boldsymbol{s_{n_d}}\}$ as fixed spatial locations and $t \in \{1, \cdots, T\}$ as discrete points in time. We next introduce $Y(\boldsymbol{s_i}, t)$ as a latent random variable defined on $(\boldsymbol{s_i}, t)$. Finally we define a spatial set: $N_i = \{\boldsymbol{s_j} : \boldsymbol{s_j} \text{ is a spatial neighbor of } \boldsymbol{s_i}\}$. Note that $\boldsymbol{s_i}$ is normally a vector in $\mathbb{R}^2$, often times representing a fixed geographic region. For example a county in the state of Iowa may be given a single representative $\boldsymbol{s_i}$.

Now, letting $\mathcal{H}_{Z(\boldsymbol{s_i},t)}$ denote the history of the data process at location $\boldsymbol{s_i}$ up until time period t, the SPINGARCH process, $\left[Z(\boldsymbol{s_i}, t) | Y(\boldsymbol{s_i}, t), \mathcal{H}_{Z(\boldsymbol{s_i},t)}\right]$ is conditionally Poisson and has a mass function of

$$f(Z(\boldsymbol{s_i},t)|Y(\boldsymbol{s_i},t),\mathcal{H}_{Z(\boldsymbol{s_i},t)}) = \exp\left[A_i(Y(\boldsymbol{s_i},t),\mathcal{H}_{Z(\boldsymbol{s_i},t)})Z(\boldsymbol{s_i},t) - B_i(Y(\boldsymbol{s_i},t),\mathcal{H}_{Z(\boldsymbol{s_i})}) + C(Z(\boldsymbol{s_i},t))\right] \tag{3.3}$$

$$\begin{aligned}
&\exp\left[A_i(Y(\boldsymbol{s_i},t),\mathcal{H}_{Z(\boldsymbol{s_i},t)})\right] = \exp\left[Y(\boldsymbol{s_i},t)\right] + \eta Z(\boldsymbol{s_i},t-1) + \kappa E\left[Z(\boldsymbol{s_i},t-1)|Y(\boldsymbol{s_i},t-1),\mathcal{H}_{Z(\boldsymbol{s_i},t-1)}\right]\\
&B_i(Y(\boldsymbol{s_i},t),\mathcal{H}_{Z(\boldsymbol{s_i},t)}) = \exp\left[A_i(Y(\boldsymbol{s_i},t),\mathcal{H}_{Z(\boldsymbol{s_i},t)})\right]\\
&C(Z(\boldsymbol{s_i},t)) = -\log\left[Z(\boldsymbol{s_i},t)!\right],
\end{aligned}$$

where

$$\begin{aligned}
&Y(\boldsymbol{s_i},t)|\boldsymbol{y_t}(N_i) \sim N(\mu(\boldsymbol{s_i},t),\sigma^2) \qquad (3.4)\\
&\mu(\boldsymbol{s_i},t) = \alpha(\boldsymbol{s_i}) + \zeta \sum_{s_j \in N_i}\{y(s_j,t) - \alpha(s_j)\}.
\end{aligned}$$

The SPINGARCH structure consists of combining (3.3), which is Markovian in time and (3.4) which is Markovian in space. The observation $Z(s_i,t)$ is conditioned on the entire past and current latent process $Y(s_i,t)$ is modeled through the use of full conditional distribution in space at each point in time. Markov assumptions reduce the conditioning to one time step and spatial neighbors.

If we let $\boldsymbol{A}$ be the neighborhood matrix such that entry $(i,j) = 1$ if s_i and s_j are neighbors and restrict $\zeta \in (\psi_{(1)}^{-1},\psi_{(n)}^{-1})$ where $\psi_{(k)}$ is the kth largest eigenvalue of $\boldsymbol{A}$, the resulting latent process model, at each time t, is a Conditional Auto-Regressive or CAR model used in Cressie and Wikle (2015).

Letting $\boldsymbol{\alpha} = (\alpha(\boldsymbol{s_1}),\cdots\alpha(\boldsymbol{s_{n_d}})$, the CAR model has joint distribution $\boldsymbol{Y_t} \sim \text{Gau}(\boldsymbol{\alpha},(I-\boldsymbol{C})^{-1}\boldsymbol{M})$ where C is an $n \times n$ matrix with entries ζ if $s_j,t \in N(\boldsymbol{s_i})$ or 0 otherwise. $\boldsymbol{M}$ is a diagonal matrix with entries σ^2. For notational convenience we define $\lambda(\boldsymbol{s_i},t) \equiv \exp\left[A_i(.)\right]$. We further take $\alpha(\boldsymbol{s_i}) = \alpha$ for $i = 1,\cdots n_d$, but these terms could be further modeled as, for example, functions of spatially varying covariates. To simplify notation, we let $\boldsymbol{\theta}$ represent the vector of parameters, σ^2,ζ and $\Sigma(\boldsymbol{\theta}) \equiv (I-\boldsymbol{C})^{-1}\boldsymbol{M}$ as the $n_d \times n_d$ latent covariance matrix. We further write the covariance between location $\boldsymbol{s_i}$ and $\boldsymbol{s_j}$ as $\Sigma_{i,j}$ with the understanding that this value will depend on whether s_j is in the neighborhood of $\boldsymbol{s_i}$ or not. Similarly, we let all the diagonal elements of Σ (generally assumed to be equal) be expressed as $\Sigma_{i,i}$.

This model extends the INGARCH(1,1) model given in (3.1) by taking the leading term d in that model to itself be a log Gaussian spatial process. We refer to model (3.3) as the Spatially Correlated Integer Auto-Regressive Conditionally Heteroskedastic, or SPINGARCH, model.

3.2.1 Model Motivation and Relationship to Mathematical Models of Crime

Though the original motivation for the INGARCH(1,1) model was as a variance property for time series, e.g. Ferland et al. (2006), it also arises out of an exponentially decaying difference equation with self-excitation and is similar to models used in mathematical criminology. Recall that the INGARCH(1,1) model assumes that $\lambda(\boldsymbol{s_i}, t) = d + \kappa\lambda(\boldsymbol{s_i}, t-1) + \eta Z(\boldsymbol{s_i}, t-1)$. Letting $\kappa = 1 - \chi$ this can be re-written as

$$\frac{\lambda(\boldsymbol{s_i}, t) - \lambda(\boldsymbol{s_i}, t-1)}{\Delta t} = d - \chi\lambda(\boldsymbol{s_i}, t-1) + \eta Z(\boldsymbol{s_i}, t-1), \tag{3.5}$$

where Δt is 1. The difference equation given in (3.5) assumes that there is a natural exponential decay in the λ process as well as self-excitation, or data model dependence, η. The η term captures the expected change that is due to repeat or near-repeat actions, a characteristic of violence that has been shown to exist in the social science literature, see e.g. Polvi et al. (1990) and Pease et al. (1998). Furthermore, at each time period the process is increased by d, some exogenous, potentially spatially varying structure. To account for covariates associated with large scale spatial structure, d could then be parameterized as $d = \exp(\boldsymbol{X}^T\boldsymbol{\beta})$.

The INGARCH(1,1) model is equivalent to equation (2.4) in Short et al. (2008) where the authors formulated mathematical models for burglaries occurring on a lattice. Here, the assumption was made that the the expected number of burglaries were a function of geographic specific measure of attractiveness, d, a natural exponential decay, χ, and repeat victimization, η. While this and other works in the mathematical criminology literature were concerned with the limiting differential equation as $\Delta t \to 0$, we are concerned with fitting aggregated data to the model which would result in fitting (3.5).

Applying the INGARCH model to data on a spatial lattice, the underlying assumption is that all spatial variation is captured through $\boldsymbol{X^T\beta}$. If we view the difference equation, (3.5), as the process of scientific interest, the model assumes that the INGARCH process manifests itself the exact same way at every spatial location. In particular, the INGARCH process assumes that any exogenous increase to the expected count at time t is the exact same at every single point in space and time that share the same large scale spatial structure. More realistically, though, there exist small scale variation between locations that cannot be captured through covariates. If the INGARCH model was used to model the spatio-temporal evolution of violence or crime, this assumption would suggest that any geographic characteristics impacting the expected number of events are fully captured in d.

The SPINGARCH (1,1) model, on the other hand assumes a similar difference equation as (3.5) however it is now assumed that d is a separate, latent, spatial process. In particular, d is assumed to follow a Conditional Auto-regressive, or CAR, process. That is, we now assume there is a separate latent spatial process that describes the increase in the difference equation due to exogenous factors. The use of a CAR process here assumes that given its geographic neighbors, the impact of the exogenous factors at the spatial location is conditionally independent of the impact at all other locations. The assumption in the INGARCH(1,1) described above is therefore relaxed to allow for slight variations in how geographic characteristics impact expected violence. A natural question is what are the impacts of the model properties through relaxing this assumption and what differing roles do η and κ practically play in this model. To answer this we first note that the SPINGARCH model can conveniently be written as

$$
\begin{aligned}
&Z(\boldsymbol{s_i},t)|\lambda(\boldsymbol{s_i},t) \sim \text{Pois}(\lambda(\boldsymbol{s_i},t)) \qquad (3.6)\\
&E[Z(\boldsymbol{s_i},t)|\lambda(\boldsymbol{s_i},t)] = \lambda(\boldsymbol{s_i},t)\\
&\boldsymbol{\lambda_t} = \exp(\boldsymbol{Y_t}) + \eta \boldsymbol{Z_{t-1}} + \kappa \boldsymbol{\lambda_{t-1}}\\
&\boldsymbol{Y_t} \sim \text{Gau}(\boldsymbol{\alpha}, (I_{n_d,n_d} - \boldsymbol{C})^{-1}\boldsymbol{M}),
\end{aligned}
$$

where $\boldsymbol{\lambda_t} = (\lambda(s_1,t), \lambda(s_2,t), \cdots, \lambda(s_{n_d},t))^T$. Here it is clear that $\boldsymbol{\lambda_t}$ is a Markov chain on state space $(\mathbb{R}^+)^{n_d}$. Viewing it this way will allow us to note the existence of a unique stationary distribution with finite moments. The moments, in particular the second moments, will differentiate between the INGARCH(1,1) and the SPINGARCH(1,1). Furthermore, as we will show, changing the variance to mean ratio in the INGARCH(1,1) process impacts the implied autocorrelation whereas the SPINGARCH(1,1) offers more flexibility in controlling the variance to mean ratio and the temporal correlation in the model. The second moments will further differentiate between the SPINGARCH(1,0), that is $\eta = 0$ and the SPINGARCH(0,1), $\kappa = 0$, models.

This means that it matters if we assume that the previous observed violence impacts the current expected violence or we assume that the previous expected level of violence impacts the current expected violence. Though the difference may seem nuanced, the choice of assumptions will result in different second order properties and as we will demonstrate in the simulation data simulated from the SPINGARCH(0,1) cannot be accurately fit to the SPINGARCH(1,0) model.

3.3 Model Properties

The INGARCH(1,1) model, when perturbed with a random sequence that has a density on the positive real line, is geometrically ergodic and subsequently converges to a unique stationary distribution as $T \to \infty$, as proved in Fokianos et al. (2009). A similar argument can be made for the SPINGARCH (1,1) process as $\boldsymbol{\lambda_t} = (\lambda(s_1,t), \cdots, \lambda(s_{n_d},t))^T$ is a Markov chain on the state space $(\mathbb{R}^{+})^{n_d}$ and is perturbed by the multivariate log-Gaussian density. Unlike the proof in Fokianos et al. (2009) further perturbation is not needed as the log Gaussian spatially correlated errors allow the chain to always have a positive probability for visiting any set of positive Lebesgue measure.

Proposition 1. *Under the parameter space restriction, $\eta, \kappa \geq 0$ and $\eta + \kappa < 1$, the SP-INGARCH (1,1) is geometrically ergodic and admits a unique stationary distributions that has finite first two moments.*

A complete proof of Proposition 1 follows closely the development in Fokianos et al. (2009) relying on Markov chain theory and is given in Appendix A. As a result of proposition 1, we can use the stationary distribution to derive first and second order properties for the SPINGARCH (1,1) model. The second order properties will provide methods for differentiating between the INGARCH (1,1) and the SPINGARCH(1,1) as well as between the SPINGARCH(1,0) and the SPINGARCH(0,1) models.

3.3.1 First Order Properties

To derive the expectation for data from the Self-Exciting Poisson CAR model, we first note that $\exp(\boldsymbol{Y})$ has a multivariate log-normal distribution. Therefore, as the natural parameter is linked exponentially with the linear predictor, using properties of the Poisson distribution, we have

$$\begin{aligned}
E\left[Z(\boldsymbol{s_i},t)\right] &= E\left[E\left[Z(\boldsymbol{s_i},t)|\lambda(\boldsymbol{s_i},t)\right]\right] \\
&= E\left[\lambda(\boldsymbol{s_i},t)\right] = E\left[\exp(Y(s_i,t))\right] + \eta E\left[Z(\boldsymbol{s_i},t-1)\right] + \kappa E\left[\lambda(\boldsymbol{s_i},t-1)\right] \\
&= \exp\left(\alpha + \frac{\Sigma_{1,1}}{2}\right) + \eta E\left[\lambda(\boldsymbol{s_i},t-1)\right] + \kappa E\left[\lambda(\boldsymbol{s_i},t-1)\right], \qquad (3.7)
\end{aligned}$$

which, at stationarity, yields, $E\left[Z(\boldsymbol{s_i},t)\right] = \frac{1}{1-\eta-\kappa}\exp(\alpha + \frac{\Sigma_{1,1}}{2})$. The existence of self-excitation within the data model, or $\eta > 0$, increases the marginal expectation for the data model in a manner similar to the INGARCH(1,1) model.

3.3.2 Second Order Properties

The SPINGARCH(1,1) model allows for flexible modeling of variances. In particular, the variance to mean ratio can be manipulated without impacting the autocorrelation in time, which distinguishes it from an INGARCH(1,1) model. In addition, spatial structure may be modeled through a combination of large-scale and small-scale spatial processes.

3.3.2.1 Variance

To see how the variance to mean ratio can be adjusted under the SPINGARCH(1,1) model we can first compute the marginal variance of $Z(\boldsymbol{s_i}, t)$. To find this value we exploit the independence of $Y(\boldsymbol{s_i}, t)$ and $Z(\boldsymbol{s_i}, t-1)$ yielding

$$\begin{aligned}
\mathrm{Var}(Z(\boldsymbol{s_i}, t+1)) =& \mathrm{Var}(E(Z(\boldsymbol{s_i}, t+1)|\lambda(\boldsymbol{s_i}, t+1)) + E(\mathrm{Var}(Z(\boldsymbol{s_i}, t+1)|\lambda(\boldsymbol{s_i}, t+1)) \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (3.8) \\
=& \mathrm{Var}(\lambda(\boldsymbol{s_i}, t+1)) + E(\lambda(\boldsymbol{s_i}, t+1)) \\
=& \kappa^2 \mathrm{Var}(\lambda(\boldsymbol{s_i}, t)) + \eta^2 \mathrm{Var}(Z(\boldsymbol{s_i}, t)) + 2\kappa\eta \mathrm{Var}\ (\lambda(\boldsymbol{s_i}, t)) + \\
& \mathrm{Var}(\exp(Y(\boldsymbol{s_i}, t))) + E(Z(\boldsymbol{s_i}, t)) \qquad (3.9) \\
=& \kappa^2 \mathrm{Var}(Z(\boldsymbol{s_i}, t)) + \eta^2 \mathrm{Var}(Z(\boldsymbol{s_i}, t)) + 2\kappa\eta \mathrm{Var}\ (Z(\boldsymbol{s_i}, t)) + \\
& - \kappa^2 E(Z(\boldsymbol{s_i}, t)) - 2\kappa\eta E(Z(\boldsymbol{s_i}, t)) + \mathrm{Var}(\exp(Y(\boldsymbol{s_i}, t))). \qquad (3.10)
\end{aligned}$$

Under the conditions in Proposition 1 we have second order temporal stationarity and subsequently,

$$\mathrm{Var}\ (Z(\boldsymbol{s_i}, t)) = \frac{1}{1-(\kappa+\eta)^2} \mathrm{Var}\ (\exp(Y(\boldsymbol{s_i}, t))) + \frac{1-\kappa^2-2\kappa\eta}{1-(\kappa+\eta)^2} E(Z(\boldsymbol{s_i}, t)). \qquad (3.11)$$

Therefore, similar to the INGARCH(1,1) process, the SPINGARCH(1,1) process allows for the modeling of overdispersion. Furthermore, as we will show below, overdispersion can be accounted for without impacting the range of possible autocorrelation.

3.3.2.2 Temporal Covariance

To see the impact of adjusting the mean to variance ratio on the temporal covariance we find the lag-one autocorrelation by relying on the second order stationarity implicit in Proposition 1. As derived in Appendix B the autocovariance under the SPINGARCH(1,1) model is

$$\text{Cov}(Z(\boldsymbol{s_i},t),Z(\boldsymbol{s_i},t+1)) = (\eta+\kappa)\,\text{Var}Z(\boldsymbol{s_i},t) - \kappa E[Z(\boldsymbol{s_i},t)]. \tag{3.12}$$

In particular, if $\kappa = 0$, i.e. SPINGARCH(0,1), the lag-one autocorrelation for the process is η and in general, the lag-h autocorrelation is η^h.

The significance of this is that it allows a great deal of flexibility in capturing second order properties of the data. The SPINGARCH (0,1) process, for example, has a lag-one auto correlation of η, and a variance to mean ratio of $\frac{\text{Var}(\exp(Y(\boldsymbol{s_i},t)))}{(1-\eta)\text{E}[\exp(Y(\boldsymbol{s_i},t))]} + \frac{1}{1-\eta^2}$. Therefore, through manipulating the $\frac{\text{Var}(\exp(Y(\boldsymbol{s_i},t)))}{\text{E}[\exp(Y(\boldsymbol{s_i},t))]}$ we can manipulate the variance to mean ratio without impacting the autocorrelation.

3.3.2.3 Spatial Covariance and Correlation

The SPINGARCH(1,1) model also allows for limited spatial correlation Recalling that $\Sigma_{i,j}$ is the marginal covariance between $Y(\boldsymbol{s_i},t)$ and $Y(\boldsymbol{s_j},t)$, the spatial covariance between $Z(\boldsymbol{s_i},t)$ and $Z(\boldsymbol{s_j},t)$ is

$$\text{Cov}(Z(\boldsymbol{s_i},t),Z(\boldsymbol{s_j},t) : \forall i \neq j) = \frac{\exp(2\alpha)}{1-(\eta+\kappa)^2}\left[\exp(\Sigma_{i,i}+\Sigma_{i,j}) - \exp(\Sigma_{i,i})\right]. \tag{3.13}$$

A proof of this is is given in Appendix B. From (3.13) it is clear that the spatial covariance is zero if the marginal covariance between $Y(\boldsymbol{s_i},t)$ and $Y(\boldsymbol{s_j},t)$ is zero. However, if there is a non-zero marginal covariance between the spatial locations, α, η and κ influence the spatial correlation in the data. As $\Sigma_{i,j}$ can be either positive or negative, the spatial covariance, unlike the temporal covariance, can be either positive or negative.

The spatial correlation is therefore

$$\begin{aligned}\operatorname{Corr}(Z(\boldsymbol{s_i},t),Z(\boldsymbol{s_j},t)) &= \frac{\exp(2\alpha)\left(\exp(\Sigma_{i,i}+\Sigma_{i,j})-\exp(\Sigma_{i,i})\right)}{\operatorname{Var}(\exp(Y(\boldsymbol{s_i},t)))+E[Z(\boldsymbol{s_i},t)]} \\ = \operatorname{Corr}(Z(\boldsymbol{s_i},t),Z(\boldsymbol{s_j},t)) &= \frac{(\exp(\Sigma_{i,i}+\Sigma_{i,j})-\exp(\Sigma_{i,i}))}{\exp(2\Sigma_{i,i})-\exp(\Sigma_{i,i})+\exp(-\alpha+\frac{\Sigma_{i,i}}{2})\frac{1}{1-(\kappa+\eta)}}. \qquad (3.14)\end{aligned}$$

The spatial correlation, as seen in (3.14), only depends on η and κ through the expectation of $Z(\boldsymbol{s_i},t)$, however this is a potential limitation as this implies that the range of correlations depends on values of parameters other than the parameters of the CAR process.

3.4 Bayesian Inference

Bayesian analysis of the SPINGARCH model can be accomplished using off the shelf Markov Chain Monte Carlo software such as rStan introduced in Carpenter et al. (2016) through application of the techniques suggested by Joseph (2016). Letting the prior distribution of $\pi(\boldsymbol{\theta}) = \pi(\eta|\kappa)\pi(\kappa)\pi(\boldsymbol{\alpha})\pi(\sigma)\pi(\zeta).$, the full condition of $\boldsymbol{\theta} = \left(\eta,\kappa,\alpha,\zeta,\sigma^2\right)^T$ is

$$\pi(\boldsymbol{\theta}|\boldsymbol{Z},\boldsymbol{Y}) \propto \prod_{t=1}^{T}\pi(\boldsymbol{Z}_t|\boldsymbol{\lambda}_t)\pi(\boldsymbol{\lambda}_t|\boldsymbol{\lambda}_{t-1},\boldsymbol{Z}_{t-1},\boldsymbol{\theta},\boldsymbol{Y}_t)\pi(\boldsymbol{Y}_t|\boldsymbol{\theta})\pi(\boldsymbol{\lambda_0}|\boldsymbol{\theta})\pi(\boldsymbol{Z_0}|\boldsymbol{\lambda_0})\pi(\boldsymbol{\theta}), \qquad (3.15)$$

and the full conditional of $\boldsymbol{Y}$ is

$$\pi(\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{\theta}) \propto \prod_{t}\pi(\boldsymbol{Z}_t|\boldsymbol{\lambda}_t)\pi(\boldsymbol{\lambda}_t|\boldsymbol{\lambda}_{t-1},\boldsymbol{Z}_{t-1},\boldsymbol{\theta},\boldsymbol{Y}_t)\pi(\boldsymbol{Y}_t|\boldsymbol{\theta})\pi(\boldsymbol{\lambda_0}|\boldsymbol{\theta})\pi(\boldsymbol{Z_0}|\boldsymbol{\lambda_0}) \qquad (3.16)$$

In general we assume independence in our priors except for η and κ due to the restriction that $\eta+\kappa<1$. Now note that in order to do any form of Markov Chain Monte Carlo inference we must sample from the density of the full latent state, $\boldsymbol{Y}$ which requires evaluations of

$$\begin{aligned}\log(\boldsymbol{Y}|\boldsymbol{\alpha},\sigma,\zeta)) \propto \frac{-t\times n_d}{2}\log(2\pi)+\frac{1}{2}\log|\Sigma_f^{-1}(\boldsymbol{\theta})| \\ -\frac{1}{2}(Y-\alpha)^T\Sigma_f^{-1}(\boldsymbol{\theta})(Y-\alpha), \qquad (3.17)\end{aligned}$$

where $\Sigma_f(\boldsymbol{\theta})$ is the full space-time covariance matrix $\left(I_{n_d\times T,n_d\times T}-I_{t,t}\otimes\boldsymbol{C}\right)^{-1}I_{t,t}\otimes\boldsymbol{M}$. The sparsity of the covariance structure means that the only computations of $\frac{1}{2}(Y-$

$\alpha)^T\Sigma_f^{-1}(\boldsymbol{\theta})(Y-\alpha)$ that need to occur are for spatial neighbors. Therefore, the most difficult part of the computation of the log-density is the computation of the determinant, $\log|\Sigma_f^{-1}(\boldsymbol{\theta})|$. However, the complicated notation of the covariance structure belies the fact that the precision matrix is both block diagonal and extremely sparse that greatly simplify computations of (3.17). The specific structure for $\Sigma^{-1}(\boldsymbol{\theta})$ allows us to follow Jin et al. (2005). In particular, we have $\log|\Sigma^{-1}(\boldsymbol{\theta})| = \frac{n_d}{2\log\sigma^2} + \log|I_{n_d,n_d} - \zeta N|$ where N is the neighborhood or adjacency matrix. Letting $V\Lambda V^T$ be the spectral decomposition of N we have $|I_{n_d,n_d} - \zeta N| = |V||I_{n_d,n_d} - \zeta\Lambda||V^T| = \prod_{j=1}^{n_d}(1-\zeta\chi_j)$ where χ_j are the eigenvalues of the neighborhood matrix. Also, as $\Sigma_f^{-1}(\boldsymbol{\theta})$ is block diagonal with each block being size $T\times T$ and having structure Σ^{-1}, it follows that $\log|\Sigma_f^{-1}(\boldsymbol{\theta})| = \frac{n_d\times T}{\log\sigma^2} + T\log|\Sigma^{-1}(\boldsymbol{\theta})|$

$$\log|\Sigma_f^{-1}(\boldsymbol{\theta})| = T\log|\Sigma^{-1}(\boldsymbol{\theta})| \tag{3.18}$$

$$\propto \frac{n_d\times T}{\log\sigma^2} + T\sum_{j=1}^{n_d}(1-\zeta\chi_j) \tag{3.19}$$

The greatest advantage of this approach is that the eigenvalues depend only on the neighborhood structure and do not depend on any parameters, therefore they can be computed ahead of time. This means that we never need to deal with matrices of the size of $\Sigma_f(\boldsymbol{\theta})$, rather we just need to find the eigenvalues for the neighborhood matrix. This allows relatively quick fit for the fully Bayesian model using software such as Stan (Carpenter et al. (2016)), which allows user defined log-densities and only requires proportional computations to the log-density.

To conduct model assessment under the above framework we rely on posterior predictive P values, see e.g. Gelman et al. (1996). This technique samples new data sets after sampling parameters from the posterior distribution. Statistics are calculated on the new data and compared to the statistics from the original dataset. For each data set we calculate $T_1(\boldsymbol{Z})$, the spatial Moran's I statistic, $T_2(\boldsymbol{Z})$, the log of the variance to mean ratio and $T_3(\boldsymbol{Z})$, the sample lag-1 auto-correlation. We denote the percentage of times the new test statistic is greater than the statistic from the original data as p_1, p_2, and p_3. Values that deviate significantly from .5 would suggest the fitted model can not accurately replicate the original data characteristic.

3.5 Simulation

Here we demonstrate that the parameters from the SPINGARCH model can be recovered using the Bayesian inference methodology and that different models within the SPINGARCH class can practically be differentiated through their second order properties. In order to demonstrate this we simulate from a SPINGARCH (0,1) and subsequently fit the simulated data to the SPINGARCH (0,1), the SPINGARCH (1,0) model, and the INGARCH(1,1) model. We then simulate from the posterior distributions and calculate posterior predictive P-values for the mean to variance ratio, a spatial Moran's I statistic, and a lag-1 autocorrelation as described above.

The spatial domain we use in the simulations is a 1-D spatial domain wrapped on a cylinder, meaning that each spatial location had two neighbors. This domain restricts $\zeta \in (-0.5, 0.5)$ in order to ensure a joint density exists for $\boldsymbol{Y}$. We use the following data generating process

$$\begin{aligned}
&Z(\boldsymbol{s_i}, t) \sim \text{Pois}(\lambda(\boldsymbol{s_i}, t)) \qquad (3.20)\\
&\lambda(\boldsymbol{s_i}, t) = \exp\left[Y(\boldsymbol{s_i}, t)\right] + 0.66 Z(\boldsymbol{s_i}, t-1)\\
&Y(\boldsymbol{s_i}, t) | \boldsymbol{y_t}(N_i) \sim \text{Gau}(\mu(s_i, t), 0.5)\\
&\mu(\boldsymbol{s_i}, t) = 0 + 0.49 \sum_{\boldsymbol{s_j} \in N_i} \{y(\boldsymbol{s_j}, t)\}.
\end{aligned}$$

A depiction of one simulation from (3.20) is shown in figure 3.1. The counts generated from (3.20) are overdispersed, temporally correlated, and spatially correlated with a log variance to mean ratio of approximately 2.8, a lag-one correlation of 0.66, and a Moran's I statistic of 0.56.

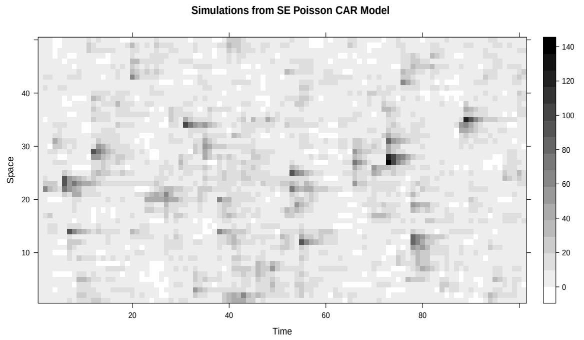

Figure 3.1 Simulated Data from (3.20). The X axis is Time from 1-100, the Y axis is Space. Note that the spatial domain here is 2-D wrapped on a cylinder. In the image horizontal streaking is indicative of high temporal autocorrelation while vertical streaking is indicative of spatial correlation.

We fit each model using Hamiltonian Monte Carlo through the R package, Rstan of Gelman et al. (2015) using vague proper priors for all parameters. For each model 3 Markov Chains of 3000 iterations are used after which the chains exhibited no signs of non-convergence. When using the inferential procedure outlined in Section 3.4, the 95% marginal credible intervals obtained from the data depicted in Figure 3.1 are $\alpha \in (-0.24, 0.1)$, $\sigma^2 \in (0.46, 0.59)$, $\zeta \in (0.486, 0.492)$, and $\eta \in (0.64, 0.67)$ all clearly cover the generating parameters. Posterior predictive P values, using 1000 random draws from each of the posterior densities, are given in Table 3.1.

Table 3.1 Posterior predictive P values from fitting the given models to data generated by a SPINGARCH(0,1) process.

	INGARCH(1,1)	SPINGARCH(1,0)	SPINGARCH(0,1)
p_1 - Moran's I Statistic	0	.05	.46
p_2 - Var to Mean Ratio	0	.99	.65
p_3 - Lag 1 Auto Correlation	0	.45	.60

The resultant posterior predictive P values, as shown in Table 3.1, demonstrate that clearly the SPINGARCH(0,1) model does the most adequate job of replicating the three statistics. The reason that the SPINGARCH(1,0) model is able to accurately replicate the Lag 1 auto-correlation in the data is because of the high variance to mean ratio. In particular, using the parameters in (3.20) the log variance to mean ratio is approximately 2.5. Therefore, values of κ that are near η will be able to generate the same autocorrelation as shown in (3.12). However, the SPINGARCH(1,0) cannot capture the mean to variance ratio, nor the spatial correlation in the data.

In general, if we fit the SPINGARCH(1,0) to data that was generated from the SPINGARCH(0,1) model it is unable to replicate the second order properties. The fit consistently overestimates the variance to mean ratio, potentially by a considerable amount and underestimates, on average, the Moran's I statistic. As is expected the INGARCH(1,1) performs poorly on all three measures though we might have expected it to replicate one of the three measures, it fails to do so. Practically, this means that violence or crime that arises from self-excitation only can be differentiated from violence that arises due to a natural decay in time by examining how well the model is able to replicate the second order statistics from the original dataset.

3.6 Burglaries in Chicago

As a case study we consider a statistical model for burglaries in the south side of Chicago during 2010-2015 using crime data from the city of Chicago. As Chicago is one of the most racially and socio-economically segregated cities in America, we consider only the southside, a relatively racial and socio-economic homogeneous region depicted in figure 3.2. We aggregated the number of burglaries by Census block group and by month. Within the south side of Chicago there are 552 census block groups resulting in a spatial domain of $\boldsymbol{s_i} \in \{\boldsymbol{s_1}, \cdots, \boldsymbol{s_{552}}\}$ and temporal domain of $t \in \{1, 2, \cdots, 72\}$.

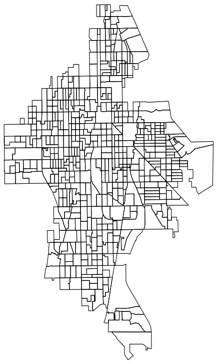

Figure 3.2 552 Census block groups in South Chicago. This area of Chicago is relatively racially and economically homogeneous

While the racially homogeneity eliminates some sources of socio-economic variability in the data, it does not eliminate all of it. To account for this, we further consider covariates that address unique socio-economic and population characteristics for each region. Unemployment, for example, has long been shown to have a relationship with crime, see e.g. Britt (1994) and Raphael and Winter-Ebmer (2001), the later showing property crime in particular has a strong relationship with unemployment. All potential covariates were obtained from the U.S. Census Tiger data available at https://www.census.gov/geo/maps-data/data/tiger-data.html. The maximum number of burglaries in a month in a census block for this subset of the city is 17. The variance to mean ratio in the data is 1.8, suggesting there is some overdispersion in the data. There is both temporal and spatial clustering as evident by the average lag-one autocorrelation, .32, and the Moran's I statistic of .20 using a four nearest neighbor structure for the weight matrix. There is a clear seasonality trend in the data as well as a general downward trend from 2010-2015. In order to account for this we preprocessed the data to remove the seasonality effect and the trend prior to estimating the impact of the spatial covariates and the process covariates.

One possible model that describes the spread of burglaries is similar to the model of Short et al. (2008). We might assume that the change in the rate of burglaries at a location is a function of a base attractiveness due to unique geographical features at that location, $\alpha_{\boldsymbol{s_i}}$, a natural decay over time, ι, and repeat victimization, η. These assumptions lead to the (stochastic) difference equation

$$\frac{\lambda(\boldsymbol{s_i}, t+1) - \lambda(\boldsymbol{s_i}, t)}{\Delta t} = \alpha_{\boldsymbol{s_i}} - \iota\lambda(\boldsymbol{s_i}, t) + \eta Z(\boldsymbol{s_i}, t). \tag{3.21}$$

Upon assuming $\Delta t = 1$, (3.21) is clearly an INGARCH(1,1) model with $\kappa = (1 - \iota)$. The unique geographical features that can be viewed as exogenous factors that increase the expectation. We assume they are structured as

$$\alpha_{\boldsymbol{s_i}} = \exp\left(\beta_0 + \beta_{pop}\log(\text{Pop}_{\boldsymbol{s_i}}) + \beta_{ym}\text{Young Men}_{\boldsymbol{s_i}} + \beta_{wealth}\text{Wealth}_{\boldsymbol{s_i}} + \beta_{unemp}\text{Unemp}_{\boldsymbol{s_i}}\right). \tag{3.22}$$

In comparison to this process, we also consider a SPINGARCH(1,1) process with the following structure,

$$\begin{aligned}
&Z(\boldsymbol{s_i}, t) \sim \text{Pois}(\lambda(\boldsymbol{s_i}, t)) && (3.23)\\
&E[Z(\boldsymbol{s_i}, t)] = \lambda(\boldsymbol{s_i}, t)\\
&\boldsymbol{\lambda_t} = \exp(\boldsymbol{Y_t} + \boldsymbol{U}) + \eta\boldsymbol{Z_{t-1}} + \kappa\boldsymbol{\lambda_{t-1}}\\
&\boldsymbol{Y_t} \sim \text{Gau}(\boldsymbol{\alpha}, \sigma^2_{ind}\boldsymbol{I}_{n_d,n_d})\\
&\boldsymbol{U} \sim \text{Gau}(\boldsymbol{0}, (\boldsymbol{N} - \boldsymbol{C})^{-1}\sigma^2_{sp}). && (3.24)
\end{aligned}$$

The structure of (3.23) differs from (3.6) in the following manner. Instead of a spatially varying effect that manifests itself differently at each point in time, in (3.23) we use a single spatial random variable $\boldsymbol{U}$. To account for the differing number of spatial neighbors, we use the weighted CAR model of Besag et al. (1991). This spatial process assumes that the latent conditional variance for each location in the CAR model is $\frac{\sigma^2_{sp}}{|N(s_i)|}$. This process has a joint density given in (3.24) letting $\boldsymbol{N}$ be a diagonal matrix with entry $(1, 1)$ equal to the number of neighbors of location s_1. Recalling that $\boldsymbol{C}$ is the matrix that has entries ζ

in position $(1, 2)$ if s_1 and s_2 are geographically adjacent, the parameter space of ζ under this formulation is $\zeta \in (-1, 1)$. The $\boldsymbol{U}$ term, then, captures the location specific variability common at all times.

We also add a second term, $\boldsymbol{Y}_t$ which captures unique characteristics at each spatial-temporal location. This term is independent across space and time and captures small spatio-temporal variability. This formulation is similar to what is commonly referred to as the Besag-York-Mollie, or BYM model, given in Besag et al. (1991). Finally, as in the (3.21), we allow $\boldsymbol{\alpha} = (\alpha_{s_1}, \cdots, \alpha_{s_{n_d}})^T$ to describe the dependence of the mean count on socio economic characteristics using the same covariates as in (3.22).

To complete the Bayesian inference we further need to place priors on all parameters in the model. In order to minimize the impact of the prior selection on the posterior densities we select diffuse proper priors for $\beta_0, \beta_1, \beta_2, \eta, \kappa, \sigma_{sp}$ and σ^2 and conducted sensitivity analysis to determine that the choice of prior had minimal impact on the results. To account for the fact that along much of the parameter space of ζ the model is nearly unidentifiable we fix ζ near the edge of the parameter space similar to an intrinsic auto-regressive model (see Wall (2004) for more issues on identifiability of the spatial parameter in a CAR model).

95% credible intervals found from fitting (3.23) using the procedure outlined in Section 3.5 and a similar Bayesian inference for (3.21) are given in Table 3.2. Note that the SPINGARCH(1,1) took approximately 13 hours to fit using 3 Markov chains of 7000 iterations each using the rStan package of Carpenter et al. (2016). $\hat{R}$ and visual examination of the chains indicated no evidence that they had not converged. Divergence transitions for the Markov chains were also checked and eliminated.

The SPINGARCH(1,1) parameters suggest that the $\alpha_{\boldsymbol{s_i}}$ process considered manifests itself differently at each unique spatial-temporal location even when residual spatial variation is accounted for in the CAR model. In other words, there may exist small-scale spatial effects that are captured in the CAR model as well as unique characteristics of each location that the CAR model does not fully explain.

Parameter	SPINGARCH(1,1)	INGARCH(1,1)
β_0	(-3.3,-1.0)	(-4.2,-3.4)
β_{pop}	(0.11,0.34)	(0.33,0.46)
β_{ym}	(-0.75,0.17)	(0.06,0.09)
β_{wealth}	(0.05,0.16)	(-0.04,0.01)
β_{unemp}	(0.006,0.07)	(0.002,0.03)
η	(0.04,0.07)	(0.22,0.24)
κ	(0.31,0.39)	(0.44,0.48)
σ^2_{sp}	(0.40,0.54)	-
σ^2_{ind}	(0.40,0.47)	-

Table 3.2 95% Credible Intervals for parameters of the SPINGARCH(1,1) and INGARCH(1,1) models applied to the Chicago burglary data.

Table 3.3 Posterior Predictive P Values

	SPINGARCH(1,1)	INGARCH(1,1)
p_1 - Moran's I Statistic	0.43	0
p_2 - Variance to Mean Ratio	0.62	0
p_3 - Lag 1 Auto Correlation	0.67	0.74

In order to examine the impact of including the spatial structure we again calculate posterior predictive P-values for the (log) variance to mean ratio, the lag-one correlation, and Moran's I statistic given in 3.3. As seen in the posterior predictive P values, the INGARCH process given in (3.21) is not able to capture the spatial structure nor the variance to mean ratio in the data. Therefore, the large scale effects considered inadequately capture the total spatial structure in the data. The model given in (3.23) does a much better job of capturing the spatial and temporal correlation as well as the variance to mean ratio.

The largest implication of the above analysis is that if the INGARCH(1,1) model was fit, one would be tempted to conclude that there exists significant self-excitement due to the high η value which is not present in the SPINGARCH(1,1) formulation. This may lead to the potentially erroneous conclusion that there is repeat victimization present in the data. However, since the SPINGARCH model appears to fit the data better, the apparent self-excitement may actually be misspecification of the error structure in the model. As the self-excitement parameter captures either repeat burglaries or burglaries motivated by a previous successful action, concluding the existence of self-excitement may have policy implications if the model was used in practice. Contrary to this, previous research in Polvi et al. (1991) suggested that a burglarized home had an elevated risk of another burglary within six weeks, however the elevated risk, may in fact, be explained through unexplained spatial correlation as demonstrated in our analysis. Policy implications of concluding repeat victimization, or self-excitement, is present in an area are discussed in Pease et al. (1998).

While not intending to be a complete treatment of Burglary in Chicago, the above demonstration does show that the SPINGARCH process has practical use in extending the INGARCH process to sociological phenomena such as the modeling of crime and failure to account for spatial correlation may result in potentially misleading conclusions regarding the existence of self-excitement. R code for fitting both the INGARCH model and the SPINGARCH model given in (3.23) is available at `https://github.com/nick3703/Chicago-Data`.

3.7 Discussion

In this manuscript we formulated a statistical model that contains both latent structure spatial dependence and data model temporal dependence extendeding earlier work of Ferland et al. (2006), Fokianos et al. (2009), and Davis and Liu (2016). We also demonstrated how such models can arise from stochastic difference equations where the number of new arrivals into a process are no longer static but rather are themselves stochastic and how these assumptions are consistent with beliefs on how violence and crimes evolve over space and time.

The resulting SPINGARCH(1,1) is novel in its combination of both data model and latent model dependency and greatly extends the uses of the INGARCH(1,1) process. Though the SPINGARCH process unique combines common models from the spatial literature an the time series literature, there are a few other notable models that are similar. In Martínez-Beneito et al. (2008) the count of diseases was modeled on a space time lattice. This model assumed that the number of infected individuals was conditionally Poisson where the natural parameter was structured to be a Log INGARCH (1,0) combined with a latent process model. The latent process model was then conditionally specified similar to a spatial conditional auto-regressive CAR model and a temporal auto-regressive CAR model,

$$Z(\boldsymbol{s_i}, t) \sim \text{Pois}(\exp(r(\boldsymbol{s_i}, t))) \tag{3.25}$$

$$r(\boldsymbol{s_i}, t) = \mu + \alpha_t + \rho\,(r(\boldsymbol{s_i}, t-1) - \alpha_{t-1} - \mu) + \theta(\boldsymbol{s_i}, t) + \epsilon(\boldsymbol{s_i}, t) \tag{3.26}$$

$$\alpha \sim \text{AR}(1) \quad \theta \sim \text{CAR} \quad \epsilon \sim \text{Gau}(0, \sigma^2). \tag{3.27}$$

Here the log-relative risk at location $\boldsymbol{s_i}$ and time t, $r(\boldsymbol{s_i}, t)$ is a linear function of a latent Gaussian conditionally autoregressive term (CAR) in space, a latent Auto-regressive (AR) term in time, as well as a function of the previous log relative risk, $r(\boldsymbol{s_i}, t-1)$.

In Mohler (2013), a discretized Cox-Hawkes model was presented that is an INGARCH(0,q) combined with a latent log-Gaussian process where the Gaussian process follows a classic

AR(1) from time-series literature, for example given in Shumway and Stoffer (2010).

$$Z_t \sim \text{Pois}(\lambda_t), \quad \lambda_t = \exp(Y_t) + \sum_{j<t} \eta\kappa^{t-j} Z_{t-j} \tag{3.28}$$

$$Y_t \sim \text{AR}(1) \tag{3.29}$$

Where AR(1) is the Auto-Regressive (1) model. In Mohler (2013), further structure is placed on κ^{i-j} to give real-world meaning to the parameters.

The SPINGARCH model, the Martínez-Beneito et al. (2008) model, and Mohler (2013) the model are each justified through the assumed existence of two separate processes that impact the expectation, or the log-expectation. Mohler (2013) used a temporal AR(1) latent process that impacts the expectation as well as a 'self-exciting' proces. Similarly, Martínez-Beneito et al. (2008) used a latent Spatio-temporal CAR process combined with a data driven process that impact the log-expectation of the Poisson.

Lastly, as evident in the example, the restriction to a spatial CAR process for the latent variable $Y(s_i, t)$ is not necessary in practice. Oftentimes a CAR specification is preferable as it is easier to model real-world phenomena conditionally. On the other hand, in Clark and Dixon (2018), both a Simultaneous Auto-Regressive (SAR) and a Vector Auto-Regressive (VAR) specification were used to model the latent variable. In the former case, the model properties derived above will still hold, however in the later case the latent state also contained temporal covariance. It is not immediately obvious that the ergodic properties of the model still exist if the latent process is allowed to propagate over time in this manner making derivations of the second order process, then, more difficult.

Appendix A - Proof of Proposition 1

We will first make use of the result given in Athreya and Pantula (1986) that states for an AR(1) process, $\lambda = (\lambda_n : n \geq 0)$, $|\eta| < 1$ given as

$$\lambda_{n+1} = \eta\lambda_n + Z_{n+1}, \tag{3.30}$$

where Z_n are a sequence of random, i.i.d., variables, a sufficient condition for the existence of a unique stationary distribution is $E[\log(1+|Z_1|)] < \infty$. This is extended to Vector AR(1) models in Zeevi and Glynn (2004).

Due to the temporal independence of $\exp(Y(s_i,t))$, we note that the latent process, $\lambda(s_i,t)$, for the SPINGARCH(1,0) process can be written as the VAR(1) process

$$\boldsymbol{\lambda}_t = \exp(\boldsymbol{Y}_t) + \kappa \boldsymbol{\lambda_{t-1}}, \tag{3.31}$$

where $\boldsymbol{\lambda}_t = (\lambda(s_1,t), \lambda(s_2,t), \cdots, \lambda(s_{n_d},t)^T$. Here, $\boldsymbol{Y}_t \sim$ iid Gau $\big(\boldsymbol{\alpha}_t, (I-C)^{-1}M\big)$. Thus, as $E[\log(1+||\boldsymbol{Y}_1||)] < \infty$ we can appeal to Proposition 2 of Zeevi and Glynn (2004) and conclude that λ admits a unique stationary distribution, π and λ converges in distribution to π as $T \to \infty$.

We next prove geometric ergodicity, and hence stationarity,for $\lambda(s_i,t)$ under a more general formulation for $n_d = 1$ which can be shown to hold in general for any spatial domain. We therefore begin with

$$\lambda(s_i,t) = \exp(Y(s_i,t)) + \kappa\lambda(s_i,t-1) + \eta Z(s_i,t-1). \tag{3.32}$$

First we note that by recursion we can write

$$\begin{aligned}
&[\lambda(s_i,t)|\lambda(s_i,0)=B] = \exp(Y(s_i,t)) + \kappa\lambda(s_i,t-1) + \eta Z(s_i,t-1)\\
&= \exp(Y(s_i,t)) + \kappa\left[\exp(Y(s_i,t-1)) + \kappa\lambda(s_i,t-2) + \eta Z(s_i,t-2)\right] + \eta Z(s_i,t-1)\\
&\cdots\\
&= \sum_{k=0}^{t-1} \kappa^k \exp(Y(s_i,t-k)) + \sum_{k=0}^{t-1} \kappa^k \eta Z(s_i,t-k-1) + \kappa^t B.
\end{aligned} \tag{3.33}$$

Intuitively this suggests that the impacts of the initial condition decay at an exponential rate and both the log-Gaussian and the Poisson errors further decay at a geometric rate. The general proofs of geometric ergodicity for INGARCH properties either follow Davis and Liu (2016) and rely on showing a geometric moment contraction condition, or follow Fokianos et al. (2009) and show a drift condition and associated small set condition. The model given in (3.32) cannot easily be shown to satisfy the geometric moment contraction condition as $E[|\exp(\boldsymbol{Y}_{t_i}) - \exp(\boldsymbol{Y}_{t_j})|] > 0$ for $i \neq j$, so therefore we will closely follow Fokianos et al. (2009) and show that a drift condition holds off of a compact set, C, then show that C is a small set, see e.g. Meyn and Tweedie (2009). A general outline we will follow is we will first show the Markov chain is ϕ-irreducible and aperiodic. We will then show the existence of a test function meeting the conditions of Theorem 15.0.1 (iii) of Meyn and Tweedie (2009). Note that here we assume that our Markov chain is initialized at $\lambda(s_i, 0) = \lambda_0$ and $Z(s_i, 0) \sim Po(\lambda_0)$ is a random variable.

First we will show that the Markov chain is ϕ-irreducible. ϕ-irreducibility, as defined in Meyn and Tweedie (2009), formally is that there exists a measure, ϕ, such that for all Borel sets, A, such that $\phi(A) > 0$, $\mathrm{P}_{\lambda_0}(\tau_A < \infty) > 0$, where P_{λ_0} is the Markov chain beginning at λ_0 and τ_A is the hitting time of the Markov chain. In other words, ϕ-irreducibility means that for any set that has positive measure, the Markov chain has a positive probability of eventually entering the set. ϕ-irreducible further implies and is implied by the condition that for every set A with $\phi(A) > 0$, $P(\lambda(s_i, t) \in A | \lambda(s_i, 0) = B) > 0$ for some t.

To show that this holds in the case we consider, consider the sequence, $Z(s_i, 0) = Z(s_i, 1) = \cdots = Z(s_i, t-1) = 0$ which occurs with positive probability due to the conditional Poisson density of $Z(s_i, t)$ and assumption that $\lambda(s_i, 0) < \infty$. If we choose t large enough we can always have $\kappa\lambda(s_i, 0) < \inf A$, hence along this sequence $P(\lambda(s_i, t) \in A | \lambda(s_i, 0) = B) = P[\sum_{k=0}^{t-1} \kappa \exp(Y(s_i, t-k)) \in (A - \kappa^t B)]$. Though there is no close formed density for sums of log-normals, $\sum_{k=0}^{t-1} \kappa^k \exp(Y(s_i, t-k))$ clearly has positive measure on $\mathbb{R}^+$. Therefore for any A with $\phi(A) > 0$, $P[\lambda(s_i, t) \in A | \lambda(s_i, 0) = B] > P[\lambda(s_i, t) \in A | \lambda(s_i, 0) = B, Z(s_i, 0) = Z(s_i, 1) = \cdots = Z(s_i, t-1) = 0] P[Z(s_i, 0) = Z(s_i, 1) = \cdots = Z(s_i, t-1) = 0] | \lambda(s_i, 0) = B > 0$ which implies ϕ irreducibility.

Note that a similar argument gives aperiodicity as to show that the chain is aperiodic it suffices to show that there exists a small set, A with $\phi(A) > 0$ such that for any $\lambda_0 \in A$, $P(\lambda(s_i, 1) \in A | \lambda(s_i, 0) = \lambda_0) > 0$ and $P(\lambda(s_i, 2) \in A | \lambda(s_i, 0) = \lambda_0) > 0$. Therefore, we just need to pick $A = [0, K]$ for K large enough and note that $P(\lambda(s_i, 2) \in A|\lambda(s_i, 0) = \lambda_0) > P(\lambda(s_i, 2) \in A|Z(s_i, 1) = 0, Z(s_i, 0) = 0, \lambda(s_i, 0) = \lambda_0)P(Z(s_i, 1) = 0|\lambda(s_i, 0) = \lambda_0, Z(s_i, 0) = 0)P(Z(s_i, 0)|\lambda(s_i, 0) = \lambda_0))$ and we note that each term on the right hand side of the inequality occurs with positive probability. Similarly $P(\lambda(s_i, 1) \in A|\lambda(s_i, 0) = \lambda_0) > P(\lambda(s_i, 1) \in A|\lambda(s_i, 0) = \lambda_0, Z(s_i, 0) = 0)P(Z(s_i, 0) = 0|\lambda(s_i, 0) = \lambda_0) > 0$ as $P(\lambda(s_i, 1) \in A|\lambda(s_i, 0) = \lambda_0, Z(s_i, 0) = 0) > 0$ and $P(Z(s_i, 0) = 0|\lambda(s_i, 0) = \lambda_0) > 0$

Next, we will appeal to Theorem 15.0.1 (iii) and Lemma 15.2.8 of Meyn and Tweedie (2009) in a manner similar to Fokianos et al. (2009). We will first show that there exists a test function, $V(\lambda)$ where the inequality $E[V(\boldsymbol{\lambda}_{t+1})|\boldsymbol{\lambda}_t = \boldsymbol{\lambda}_*] \leq \psi V(\boldsymbol{\lambda}_*) + L\ \mathrm{I}(\boldsymbol{\lambda}_* \in C)$ holds where $\psi \in (0, 1)$, $L \in (0, \infty)$ and $I(.)$ is the indicator function. Next we will show that C is a small set and hence a petite set. The compact set we will consider is $C \in [0, G]$ where $G \in (0, \infty)$. Note that this requires expanding the parameter space of λ from $(0, \infty)$ to $[0, \infty)$. To do this we define $Z \sim (\text{Po }(0))$ to be degenerately 0.

Akin to Fokianos et al. (2009) we can consider $V(\lambda) = 1 + \lambda^2$. Suppressing the dependency on $\alpha, \zeta,$ and $, \sigma^2$ we write γ as the second moment of the log-Gaussian density, $\exp(Y(s_i, t))$ and have

$$E[V(\lambda_t)|\lambda_{t-1} = \lambda_*] = 1 + E[(\exp(Y(s_i, t)) + \kappa\lambda_{t-1} + \eta Z_{t-1})\,|\lambda_{t-1} = \lambda_*] \tag{3.34}$$

$$= 1 + \gamma + (\kappa + \eta)^2\lambda_*^2 + 2(\eta + \kappa)\exp(\alpha + \frac{\Sigma_{1,1}}{2})\lambda_* \tag{3.35}$$

First consider $\lambda_* \in C^c$, on this set we have

$$1 + \gamma + (\kappa + \eta)^2\lambda_*^2 + 2(\eta + \kappa)\exp(\alpha + \frac{\Sigma_{1,1}}{2})\lambda_* = \\ \left[\left(1 - \frac{\lambda_*^2}{1 + \lambda_*^2} + \frac{(\eta + \kappa)^2\lambda_*^2}{1 + \lambda_*^2}\right) + \frac{\gamma}{1 + \lambda_*^2} + \frac{2(\eta + \kappa)\exp(\alpha + \frac{\Sigma_{1,1}}{2})\lambda_*}{1 + \lambda_*^2}\right](1 + \lambda_*^2). \tag{3.36}$$

Here, as G increases, the supremum of the term inside [.] goes to $(\eta + \kappa)^2$ which is less than 1 by assumptions on the parameter space for η and κ.

Next, for $\lambda_* \in C$, we can still write (3.35) which is bounded by $1 + \gamma + (\kappa + \eta)^2G^2 + 2(\eta + \kappa)\exp(\alpha + \frac{\Sigma_{1,1}}{2})G$. Thus, there exists $C = [0, G]$ such that $E[V(\lambda_{t+1})|\lambda_t = \lambda_*] \leq \psi V(\lambda_*) + L\ \mathrm{I}(\lambda_* \in C)$.

Next, we will show that the set $C = [0, G]$ is a small set. That is, $\exists\ n$ such that

$$\inf_{\lambda \in C} \mathrm{P}^n(\lambda, A) > 0 \tag{3.37}$$

for a set A having Lebesgue measure greater than zero.

To show this, let $\lambda_0 \in C$ and $Z_0 = 0$. Then, there exists a path, $Z_1 = \cdots = Z_m = 0$ that exists with probability greater than zero. Using the recursive formulation, (3.33), it follows that along that path, $\lambda_m = \sum_{k=0}^{m-1} \kappa^k \exp(Y_{t-k}) + (\kappa)^m \lambda_0$. While the geometric sum of uncorrelated log Gaussian terms, $\sum_{k=0}^{m-1} \kappa^k \exp(Y_{t-k})$ has no closed form solution, it has density with regard to the positive Lebesgue measure. Therefore, if we consider an interval with positive Lebesgue measure, $A = (a, b)$, then there exists $m = N$ such that $\kappa^N \lambda_0 < a$. Therefore, for A, $\inf_{\lambda \in C} \mathrm{P}^N(\lambda, A) > \mathrm{P}\ (\kappa \exp(Y_{t-k}) \in (a - \kappa)^N \lambda_0, b - (\kappa)^N \lambda_0)\mathrm{P}\ (Z_1 = \cdots = Z_N = 0) > 0$. Thus, the interval (a, b) is uniformly reachable from $\lambda_0 \in C$. Therefore it follows in a manner similar to Fokianos et al. (2009), that $C = [0, K]$ is a small set. This demonstrates that (3.32) is geometrically ergodic and therefore admits a unique stationary distribution. Furthermore, the specific choice of $V(.)$ used in the drift condition ensures that second moments exist for the stationary distribution.

Appendix B - Derivation of Temporal and Spatial Covariance

In order to derive the temporal covariance, without loss of generality we assume $\boldsymbol{\alpha} = \mathbf{0}$ and we first find

$$E\left[Z(s_i, t) Z(s_i, t+1)\right] = E\left[Z(s_i, t)\left(E[Z(s_i, t+1) | \lambda(s_i, t), Z(s_i, t)]\right)\right] \tag{3.38}$$

$$= E\left[Z(s_i, t)\left(\eta Z(s_i, t) + \kappa \lambda(s_i, t) + \exp(\frac{\Sigma_{1,1}}{2})\right)\right] \tag{3.39}$$

$$= E\left[Z(s_i, t)^2\right] + \kappa E\left[Z(s_i, t)\lambda(s_i, t)\right] + \frac{1}{1 - (\eta + \kappa)} \exp(\Sigma_{1,1}) \tag{3.40}$$

$$= \eta\left(\mathrm{Var}\ (Z(s_i, t)) + E[Z(s_i, t)]^2\right) + \kappa E[\lambda(s_i, t)^2] + \frac{1}{1 - (\eta + \kappa)} \exp(\Sigma_{1,1}) \tag{3.41}$$

$$= \eta \mathrm{Var}\ (Z(s_i, t)) + \frac{\eta}{(1 - (\eta + \kappa))^2} \exp(\Sigma_{1,1}) + \kappa E[\lambda(s_i, t)^2] + \frac{1 - (\eta + \kappa)}{(1 - (\eta + \kappa))^2} \exp(\Sigma_{1,1}). \tag{3.42}$$

Therefore, as $E[Z(s_i,t)]^2 = \frac{1}{(1-(\eta+\kappa))^2}\exp(\Sigma_{1,1})$, the covariance is

$$\text{Cov }(Z(s_i,t)Z(s_i,t+1)) = \eta\text{Var}Z(s_i,t) + \kappa E[\lambda(s_i,t)^2] - \frac{\kappa}{(1-(\eta+\kappa))^2}\exp(\Sigma_{1,1}). \quad (3.43)$$

Next, we note that $E[\lambda(s_i,t)^2] = \text{Var }(\lambda(s_i,t)) + (E[\lambda(s_i,t)])^2 = \text{Var}(Z(s_i,t)) - E[\lambda(s_i,t)] + (E[\lambda(s_i,t)])^2$. Thus we have

$$\text{Cov }(Z(s_i,t)Z(s_i,t+1)) = (\eta+\kappa)\,\text{Var}(Z(s_i,t)) - \kappa E[\lambda(s_i,t)] +$$

$$\frac{\kappa}{(1-(\eta+\kappa))^2}\exp(\Sigma_{1,1}) - \frac{\kappa}{(1-(\eta+\kappa))^2}\exp(\Sigma_{1,1}) \quad (3.44)$$

$$= (\eta+\kappa)\,\text{Var}(Z(s_i,t)) - \kappa E[Z(s_i,t)], \quad (3.45)$$

as desired.

Next we find $E[Z(s_i,t)Z(s_j,t)]$ for arbitrary $i \neq j$. Recall that we let $\Sigma_{i,j}$ be the entry in the covariance matrix at location (i,j).

First note that $E[Z(s_i,t)Z(s_j,t)] = E[E[Z(s_i,t)Z(s_j,t)|\lambda(s_i,t),\lambda(s_j,t)]] = E[\lambda(s_i,t)\lambda(s_j,t)]$. Using this we have

$$E[Z(s_i,t)Z(s_j,t)] = E[\lambda(s_i,t)\lambda(s_j,t)] \quad (3.46)$$

$$= \eta^2 E[Z(s_i,t-1)Z(s_j,t-1)] + \eta\kappa E[Z(s_i,t-1)\lambda(s_j,t-1)] +$$

$$\eta\kappa E[\lambda(s_i,t-1)Z(s_j,t-1)] + 2\eta\frac{1}{1-(\eta+\kappa)}\exp(2\alpha\Sigma_{i,i}) +$$

$$\kappa^2 E[\lambda(s_i,t-1)\lambda(s_j,t-1)] + 2\kappa\frac{1}{1-(\eta+\kappa)}\exp(2\alpha\Sigma_{i,i}) + \exp(2\alpha)\exp(\Sigma_{i,i}+\Sigma_{i,j}) \quad (3.47)$$

$$= (\eta+\kappa)^2 E[Z(s_i,t-1)Z(s_j,t-1)] + 2\frac{\eta+\kappa}{1-(\eta+\kappa)}\exp(\Sigma_{i,i}) + \exp(2\alpha)\exp(\Sigma_{i,i}+\Sigma_{i,j}) \quad (3.48)$$

Relying on second order stationarity in time, this yields

$$E[Z(s_i,t)Z(s_j,t)] = \frac{1}{1-(\eta+\kappa)^2}\left(2\frac{(\eta+\kappa)}{1-(\eta+\kappa)}\exp(2\alpha+\Sigma_{i,i}) + \exp(\Sigma_{i,i}+\Sigma_{i,j})\right). \quad (3.49)$$

Therefore, we have

$$\begin{aligned}
&\mathrm{Cov}(Z(s_i,t)Z(s_j,t) = \frac{1}{1-(\eta+\kappa)^2}\left(2\frac{(\eta+\kappa)}{1-(\eta+\kappa)}\exp(2\alpha+\Sigma_{i,i})+\exp(\Sigma_{i,i}+\Sigma_{i,j})\right)- \\
&\frac{1}{(1-(\eta+\kappa))^2}\exp(2\alpha+\Sigma_{i,i}) \qquad (3.50) \\
&= \frac{\exp(2\alpha)}{1-(\eta+\kappa)^2}\left(\exp(\Sigma_{i,i}+\Sigma_{i,j}-\exp(\Sigma_{i,i}))\right), \qquad (3.51)
\end{aligned}$$

as given in (3.13).

CHAPTER 4. An Extended Laplace Approximation Method for Bayesian Inference of Self-Exciting Spatial-Temporal Models of Count Data

A paper to be submitted to *Computational Statistics and Data Analysis*

Abstract

Self-Exciting models are statistical models of count data where the probability of an event occurring is influenced by the history of the process. In particular, self-exciting spatio-temporal models allow for spatial dependence as well as temporal self-excitation by extending the integer-valued generalized autoregressive conditionally heteroscedastic (INGARCH) model to account for spatial correlation. For large spatial or temporal regions, however, the model leads to an intractable likelihood. An increasingly common method for dealing with large spatio-temporal models is by using Laplace approximations (LA). This method is convenient as it can easily be applied and is quickly implemented. However, as we will demonstrate in this manuscript, when applied to self-exciting Poisson spatial-temporal models, Laplace Approximations result in a significant bias in estimating some parameters. Using this class of models as an exemplar, we will explain why and when this bias exists and offer recommendations for when to use higher order Laplace approximations. We will demonstrate how to do this in a Bayesian setting for Self-Exciting Spatio-Temporal models. We will further show there is a limited parameter space where the extended LA method still has bias. In these uncommon instances we will demonstrate how a more computationally intensive fully Bayesian approach using the Stan software program is possible in those rare

instances. The performance of the extended LA method is illustrated with both simulation and real-world data.

4.1 Introduction

Intractable likelihood functions arise in a multitude of settings in statistics, especially in modeling spatio-temporal data. For spatial or spatio-temporal models it is oftentimes easier to specify the probability of an event occurring at a given location conditional on the occurrence or non-occurrence at neighboring events. In this instance, it is easy to write down the conditional density, but the joint density may not have a closed form expression, or, if it does, the likelihood cannot be evaluated.

For example, in a spatial process observed on a fixed lattice we may have, writing $\boldsymbol{s_i} \in \{\boldsymbol{s_1}, \cdots, \boldsymbol{s_{n_d}}\}$ as fixed locations in $\mathbb{R}^2$, $Z(\boldsymbol{s_i}) \sim \text{Pois}(\lambda(\boldsymbol{s_i}))$ as observed counts at a given location. We may further have $\boldsymbol{\lambda} \sim \text{Log Gau}(\boldsymbol{\alpha}, \Sigma(\theta))$ where $\boldsymbol{\lambda}$ is the vector of all Poisson expectations at each location and Log Gau is the standard multivariate log-Gaussian distribution. Spatial structure may be placed on $\Sigma(\theta)$ by, for example, letting $\Sigma(\theta) = (I_{n_d,n_d} - C)^{-1}M$ where I is the identity matrix, C is a matrix with entries ζ at location i, j if spatial locations $\boldsymbol{s_i}$ and $\boldsymbol{s_j}$ are spatial neighbors, and M is a diagonal matrix with diagonal entries, σ^2. This model is oftentimes called the Poisson-CAR model and is described in detail in Section 4.2 of Cressie and Wikle (2015). The log-likelihood for the spatial parameters is proportional to the intractable integral

$$l_{n_d}(\theta) \propto -\frac{1}{2}\log\det\left(\Sigma(\theta)\right) + \log\int_{\mathbb{R}^{n_d}} \exp\left(\sum_{s_{i=1}}^{s_{i=n_d}} Z(\boldsymbol{s_i})Y(\boldsymbol{s_i}) - \exp(Y(\boldsymbol{s_i})) - \frac{1}{2}\boldsymbol{Y}^T\Sigma^{-1}(\theta)\boldsymbol{Y}\right) \mathrm{d}\boldsymbol{Y}, \tag{4.1}$$

where θ is the set of all spatial parameters.

However, while the integral in (4.1) is intractable, it is of the form $I_n = \int_{\mathbb{R}^n} \exp(-h_d\,(Y))\mathrm{d}y$ allowing for Laplace approximations to be used to conduct inference. In both spatial and spatio-temporal modeling, using Laplace approximations to conduct inference on the spatial or spatio-temporal diffusion parameters has dramatically increased since the advent of the

Integrated Nested Laplace Approximation, or INLA, package from Rue et al. (2009). Rue et al. (2017) provides many examples of INLA being used in literature.

Though the Laplace approximation technique is extremely fast compared to Markov Chain Monte Carlo (MCMC) techniques and it provides consistent estimates for parameters, it only does so asymptotically where the asymptotic error rate decreases as a function of pseudo independent observations. By pseudo independent we mean observations that are separated sufficiently far in either spatial or temporal distance as to have minimal influence on one another. For example, in Cressie (1992), p.15, it is shown how a simple spatial-only model with 10 spatially dependent observations is equivalent to 6 pseudo-independent observations. The growth of the equivalent independent observations is what justifies, asymptotically, the consistency of the Laplace approximations. Meaning, if the correlation structure of $\Sigma(\theta)$ is strong, then increasing the number of observations may only have minimal impact on the validity of the Laplace approximations.

In this manuscript we will re-examine some of the shortfalls of using Laplace approximations for inference of spatial or spatio-temporal diffusion parameters. For a class of models which we will refer to as the self-exciting Poisson CAR models we will show how the assumptions for the first order Laplace approximations of techniques such as INLA may not hold over the entire parameter space. We will demonstrate how, in this case, higher order approximations of Shun and McCullagh (1995) and Evangelou et al. (2011) offer more accurate inference and offer greater consistency in parameter estimation and show how the results are comparable to a fully Bayesian inference using rStan of Gelman et al. (2015).

4.2 Model

In this manuscript we write $Z(\boldsymbol{s_i}, t)$ for observed count data on a spatial temporal lattice where $\boldsymbol{s_i} \in \{\boldsymbol{s_1}, \cdots, \boldsymbol{s_{n_d}}\}$ indexes space and $t \in \{1, 2, ...T\}$ indexes time. Defining $\boldsymbol{Z_t} = (Z(s_1, t), Z(s_2, t), ..., Z(s_{n_d,t}))^T$, consider the SPINGARCH(0,1) model given in Chapter 3

$$
\begin{aligned}
&Z(\boldsymbol{s_i}, t) \sim \text{Pois}(\lambda(\boldsymbol{s_i}, t)) && (4.2)\\
&E[Z(\boldsymbol{s_i}, t)] = \lambda(\boldsymbol{s_i}, t)\\
&\boldsymbol{\lambda_t} = \exp(\boldsymbol{Y_t}) + \eta \boldsymbol{Z_{t-1}}\\
&\boldsymbol{Y_t} \sim \text{Gau}(\boldsymbol{\alpha_t}, (I_{n_d,n_d} - \boldsymbol{C})^{-1}\boldsymbol{M}).
\end{aligned}
$$

This model is also closely related to the Self-Exciting spatially correlated model in Chapter 2.

We again let $\boldsymbol{C}$ be the spatial proximity matrix with entry $(i, j) = \zeta$ if the spatial locations, $\boldsymbol{s_i}, \boldsymbol{s_j}$ are neighbors and 0 otherwise. And let $\boldsymbol{M}$ be a diagonal matrix of dimension $n_d \times n_d$ with diagonal entries σ^2. In order to ensure positive definiteness of the Gaussian covariance matrix we must have $\zeta \in (\psi_{(1)}^{-1}, \psi_{(n)}^{-1})$ where $\psi_{(k)}$ is the kth largest eigenvalue of the neighborhood matrix.

Data level dependence, or what is commonly referred to as self-excitation, is present in the model through the addition of the $\eta \boldsymbol{Z_{t-1}}$ term to the linear predictor of $\boldsymbol{\lambda}$. The expected number of events at space-time location $(\boldsymbol{s_i}, t)$ then is a summation of the expected events due to an underlying, latent CAR process, as well as events due to repeat or copy-cat actors. A sufficient condition to ensure a unique stationary distribution as shown in Chapter 3 is $\eta \in (0, 1)$.

The data model for $Z(\boldsymbol{s_i}, t)$, when conditioned on $Z(\boldsymbol{s_i}, t-1)$ and $Y(\boldsymbol{s_i}, t)$, is then Poisson. In other words, the density of $Z(\boldsymbol{s_i}, t)$ depends on the previously observed $Z(\boldsymbol{s_i}, t-1)$ and a latent, unobserved $Y(\boldsymbol{s_i}, t)$.

This model is closely related to the INGARCH(1,1) model. A typical INGARCH model, for example in Davis et al. (2016), is a univariate model for discrete time series where,

$$
\begin{aligned}
&Z_t \sim \text{Pois}(\lambda_t) && (4.3)\\
&\lambda_t = d + \kappa\lambda_{t-1} + \eta Z_{t-1}. && (4.4)
\end{aligned}
$$

Therefore, the self-exciting Poisson CAR model in (3.6) extends this to incorporate spatial variation through the addition of a spatially structured log-Gaussian error term. The model given in (3.6) is also a spatial version of the discrete Hawkes-Cox model of Mohler (2013) only allowing a time lag of 1.

The latent process model, $Y(\boldsymbol{s_i}, t)$, is a Conditional Auto-Regressive or CAR model given in Cressie and Wikle (2015) and has joint distribution $\boldsymbol{Y_t} \sim \text{Gau}(\boldsymbol{\alpha_t}, (I_{n_d,n_d} - \boldsymbol{C})^{-1}\boldsymbol{M})$. Statistically the self-exciting Poisson CAR model is interesting as it is both hierarchical and conditionally specified at the data level, not at the process level.

As well as being statistically interesting model (4.2) also arises naturally when the expected count at space-time location $(\boldsymbol{s_i}, t)$ is equal to the expected count due to a spatial latent process, $\exp(Y(\boldsymbol{s_i}, t))$ and the expected count due to self-excitation, $\eta Z(\boldsymbol{s_i}, t-1)$. This can occur, for example in the modeling of violence in a region. The latent (unobserved) tension in the region may be solely due to geography or demographics observed at a given space and time. This may be expressed as a function of large-scale variation, $\boldsymbol{\alpha}$ and small scale variation which is captured in the CAR component of the model. The critical assumption is that the small scale variation only exists in space. The second cause of violence in a space-time region may be attributed to the ”repeat-victimization” effect, or the propensity of violent action to be repeated in, or near, the same geographical region. That is, once a violent action occurs, there is some probability that that action will generate copy cats. As a consequence of the model, if we know $\exp(Y(\boldsymbol{s_i}, t))$ and η, then the expected number of violent events that arise from model(3.6) can be seen as the sum of the expected number of events due to the latent process and the expected number of events due to copy cat actors.

The likelihood associated with this model is given in (4.5).

$$\begin{aligned} L(\eta, \alpha, \zeta, \sigma^2 | \boldsymbol{Z}) \propto \int_{\boldsymbol{\Omega}_y} \prod_{i=1}^{n} \prod_{t=1}^{T} & \exp(-\eta Z(\boldsymbol{s_i}, t-1) - \exp(Y(\boldsymbol{s_i}, t))) \\ & \times (\eta Z(\boldsymbol{s_i}, t-1) + \exp(Y(\boldsymbol{s_i}, t)))^{Z(\boldsymbol{s_i}, t)} \, d\mu_{\boldsymbol{Y}} \end{aligned} \tag{4.5}$$

and due to the temporal independence of $\boldsymbol{Y}|\boldsymbol{Z}$, we can simplify this to

$$L(\eta, \alpha, \zeta, \sigma^2 | \boldsymbol{Z}) \propto \prod_{t=1}^{T} \int_{\boldsymbol{\Omega}_{y_t}} \prod_{i=1}^{n} \exp(-\eta Z(\boldsymbol{s_i}, t-1) - \exp(Y(\boldsymbol{s_i}, t))) \\ \times (\eta Z(\boldsymbol{s_i}, t-1) + \exp(Y(\boldsymbol{s_i}, t)))^{Z(\boldsymbol{s_i}, t)} \, d\mu_{\boldsymbol{Y_t}}. \tag{4.6}$$

However, practically, this likelihood cannot be directly maximized due to the intractable integration that is taken with respect to the multivariate Gaussian density associated with $\boldsymbol{Y}$. Bayesian Monte Carlo Markov Chain (MCMC) methods also are extremely challenging in this set-up as MCMC techniques will generally either involve integrating (4.5) or sampling from the latent states. A similar model was analyzed in Mohler (2013) where inference was conducted using Metropolis Adjusted Langevin Algorithm (MALA). The challenge in using MCMC techniques including MALA is that the dimension of $\Sigma(\theta)$ is potentially quite large. Any sampling of Y will require thousands of evaluations of the determinant of this matrix as well evaluations of the log-likelihood. As we will describe in Section 5 this can be sped up through precomputing eigenvalues of the neighborhood matrix but even with this, it remains potentially painfully slow and unfeasible in the model building phase of analysis.

4.3 Laplace Approximation

An approximation method similar to Integrated Nested Laplace Approximation (INLA) was used to fit a Self-Exciting Poisson SAR model in Clark and Dixon (2018). This inferential technique is based on the work done in Tierney and Kadane (1986). Using this, we can approximate $\pi(\theta|Z)$ where Z is the observed data, Y is a latent random variable, and θ is the set of parameters that inference by using the relationship

$$\pi(\theta|Z) \propto \left. \frac{\pi(Z, Y, \theta)}{\pi_G(Y|Z, \theta)} \right|_{Y=Y^*(\theta)}, \tag{4.7}$$

where $\pi(.)$ represent a density function and $\pi(.|.)$ represent a conditional density function and $\pi_G(Y|Z, \theta)$ is the Gaussian approximation to the density $\pi(Y|Z, \theta)$. Both the numerator and the denominator are then evaluated at the mode of Y for a given θ, denoted as $Y^*(\theta)$. The benefit of this, when applied to (4.5) is that it is essentially an integration free method of marginalizing over Y. For (3.6), this becomes

$$\tilde{\pi}(\eta, \zeta, \sigma^2, \alpha|\boldsymbol{Z}) \propto \frac{\pi(\boldsymbol{Z}|\eta, \boldsymbol{Y})\pi(\boldsymbol{Y}|\alpha, \zeta, \sigma^2)\pi(\zeta)\pi(\alpha)\pi(\sigma^2)\pi(\zeta)}{\pi_G(\boldsymbol{Y}|\alpha, \eta, \zeta, \sigma^2, \boldsymbol{Z})}, \tag{4.8}$$

where $\tilde{\pi}(\eta, \zeta, \sigma^2, \alpha|\boldsymbol{Z})$ is an approximation to the marginal posterior density of $\eta, \zeta, \sigma^2, \alpha$, and $\pi_G(\boldsymbol{Y}|\alpha, \eta, \zeta, \sigma^2, \boldsymbol{Z})$ is a Gaussian approximation to the joint density of the latent state $\boldsymbol{Y}$.

The Gaussian approximation given in the denominator of (4.8) is based off of a Taylor series approximation to the log-density of $\pi(\boldsymbol{Z}|\boldsymbol{Y}, \eta)$. That is, as $\pi_G(\boldsymbol{Y}|\alpha, \eta, \zeta, \sigma^2, \boldsymbol{Z}) \propto \pi(\boldsymbol{Z}|\boldsymbol{Y}, \eta)\pi(\boldsymbol{Y}|\alpha, \zeta, \sigma^2)$ we can use the fact that $\pi(\boldsymbol{Y}|\alpha, \zeta, \sigma^2)$ is Gaussian to create a Gaussian approximation to the full conditional of $\boldsymbol{Y}$ by calculating a truncated Taylor series expansion of $\log \pi(\boldsymbol{Z}|\boldsymbol{Y}, \eta)$ about the mode of $\boldsymbol{Y} = \boldsymbol{\mu}$, yielding

$$\begin{aligned}\pi_G(\boldsymbol{Y}|\eta, \zeta, \sigma^2, \boldsymbol{Z}) \propto &(2\pi)^{n/2} \det(\Sigma(\theta))^{1/2} \exp(-\frac{1}{2}(\boldsymbol{Y} - \boldsymbol{\alpha})^t \Sigma^{-1}(\theta)(\boldsymbol{Y} - \boldsymbol{\alpha}) + \\ &\sum_{s_i, t} f(\mu(s_i, t))(Y(s_i, t) - \alpha(s_i, t) + 1/2k(\mu(s_i, t))(Y(s_i, t) - \alpha(s_i, t)^2).\end{aligned} \tag{4.9}$$

Where in (4.9)

$$\begin{aligned}f(\mu(s_i, t)) = &\frac{Z(s_i, t)\exp(\mu(s_i, t))}{\exp(\mu(s_i, t)) + \eta Z(s_i, t-1)} - \exp(\mu(s_i, t)) - \\ &\mu(s_i, t)\left(\frac{Z(s_i, t)\exp(\mu(s_i, t))}{\exp(\mu(s_i, t)) + \eta Z(s_i, t} - \frac{\exp(2\mu(s_i, t))Z(s_i, t)}{(\exp(\mu(s_i, t)) + \eta Z(s_i, t-1))^2} - \exp(\mu(s_i, t))\right),\end{aligned} \tag{4.10}$$

and

$$k(\mu(s_i, t)) = -\frac{Z(s_i, t)\exp(\mu(s_i, t))}{\exp(\mu(s_i, t)) + \eta Z(s_i, t)} + \frac{\exp(2\mu(s_i, t))Z(s_i, t)}{(\exp(\mu(s_i, t)) + \eta Z(s_i, t-1))^2} + \exp(\mu(s_i, t)). \tag{4.11}$$

The expressions $f(.)$ and $k(.)$ given are derived from expanding the log-density of Z as a function of Y about an initial guess for the mode. (4.9) is then maximized as a function of $\boldsymbol{Y}$ and then evaluated at that value. This is equivalent to the Laplace Approximation to the marginal density given in Tierney and Kadane (1986). The computational burden comes in finding the mode of $\pi_G(\boldsymbol{Y}|\alpha, \eta, \zeta, \sigma^2, \boldsymbol{Z})$, however the sparsity of $\Sigma^{-1}(\theta)$ makes this easier as it requires the repeated solution, for each time point, t, for the vector $\boldsymbol{\mu_{n+1}} = (\mu_{n+1}(s_1, t), \cdots \mu_{n+1}(s_{n_d}, t))^T$ in the linear equation $\left(\Sigma^{-1}(\theta) + \text{Diag } k(\mu_n(s_i, t))\right) \boldsymbol{\mu_{n+1}} = f(\boldsymbol{\mu_n})$. The sparsity of $\Sigma^{-1}(\theta)$ ensures the sparsity of $\left(\Sigma^{-1}(\theta) + \text{Diag } k(\mu_n(s_i, t))\right)$ and convergence occurs rapidly.

When (4.9) is evaluated at the posterior mode, it becomes $2\pi^{n/2} \det(W + \Sigma^{-1}(\theta))^{\frac{1}{2}}$ where W is a diagonal matrix of the same dimension as $\Sigma(\theta)$ where each diagonal entry is $k(\mu(s_i, t))$. The numerator of (4.8) is then evaluated at $\mu(s_i, t)$. Therefore, the problem is simply a computation once the posterior mode of the denominator is found.

Inference is then carried out by fixing values of $\eta, \zeta, \sigma^2, \alpha$, then finding the values of $\boldsymbol{Y}$ that maximize the Gaussian approximation. Then, for those fixed parameter values, we obtain an estimate of the posterior probability. The parameter space for $\eta, \zeta, \sigma^2, \alpha$ can be efficiently explored to map out the marginal likelihood surface for that set of parameters. Rue et al. (2009) discuss efficient methods for exploring the parameter space.

From $\tilde{\pi}(\eta, \zeta, \sigma^2, \alpha|\boldsymbol{Z})$ and $\pi_G(\boldsymbol{Y}|\alpha, \eta, \zeta, \sigma^2, \boldsymbol{Z})$ we can then estimate the marginal posterior density $\pi(Y|Z)$ by calculating $\pi(Y|Z) \approx \sum \tilde{\pi}(\eta, \zeta, \sigma^2, \alpha|\boldsymbol{Z})\pi_G(\boldsymbol{Y}|\alpha, \eta, \zeta, \sigma^2, \boldsymbol{Z})$ where the summation is over all values of θ with sufficiently high posterior probability. If inferential concern is on the density of the latent state, we can subsequently improve $\pi_G(\boldsymbol{Y}|\alpha, \eta, \zeta, \sigma^2, \boldsymbol{Z})$ by using a skew-Normal approximation based off of a higher order Taylor series expansion as given in Rue et al. (2009).

Clearly the Gaussian approximation, and hence the Laplace approximation, is asymptotically valid if the Taylor series of Z has a vanishing third and higher derivatives. Otherwise, the practitioner must rely on the assumption that the higher order terms are negligible, which, as we will show, is not the case in this model.

4.3.1 Issues with Laplace Approximation for Spatio-Temporal Data

There are two primary concerns with using this technique. The concerns are somewhat addressed in Rue et al. (2009), but we will make them clear here. The first concern is unavoidable in any parametric modeling of spatio-temporal data. To see this issue, it is instructive to consider spatial sampling with temporal replication where there is no temporal dependence. If we only consider $Z(\boldsymbol{s_i})$ with $\boldsymbol{s_i} \in \{\boldsymbol{s_1}, \cdots, \boldsymbol{s_{n_d}}\}$ and say we sample this T times, then we have complete replication of any spatial patterns to conduct inference from. Without replication, we have to hope that our spatial domain is large enough to create internal replication, that is, that the dependency in the data decays at a sufficient rate. This same issue exists in spatio-temporal data. Now, we have data that has dependence in both space and time and we inevitably only have a single realization of the data. Therefore, our space-time observation must be large enough to break both the space dependence and the time dependence. Essentially, this means that our, unobservable, space-time clusters must be small.

This is an issue with using Laplace approximations as the inferential results are asymptotically justified through the growth of independent samples. The approximation error of Tierney and Kadane (1986) is $\mathcal{O}(n^{-3/2})$, however the meaning of 'n' for spatio-temporal models is not well-defined. The asymptotics are clearly justifiable if both the size of the grid and the number of observations per node increases, but the n that needs to grow is the number of independent space-time observations.

One method of examining whether this has occurred is to look at the effective number of parameters as defined in Spiegelhalter et al. (1998). If the data is completely independent, then n is indeed the number of samples. In this case, the effective number of parameters is the number of large scale parameters in the model. If we examine the ratio of observations to the effective number of parameters we will get an estimate of the number of observations available to estimate each of the effective number of parameters. If, for example, the effective number of parameters is close to n, then the ratio of observations to effective number of parameters will be extremely small indicating that we lack sufficient observations to conduct meaningful analysis.

The above concern really applies for any analysis of space-time data when we directly work with the full log-likelihood. In order to conduct meaningful inference we need to have replication or pseudo-replication of our data. The second issue is more specific to Laplace approximations and appears to be more prevalent in count data. That is, there is a bias in the approximation due to the truncation of the Taylor series that underlies the Gaussian approximation in the denominator of (4.8). This appears to first have been demonstrated in Joe (2008) where clustered (temporal) count data was analyzed assuming a Poisson-log Gaussian mixture where the log Gaussian was assumed to have an AR(1) structure. In Joe (2008), the AR(1) parameter was consistently shown to be biased low and, assuming zero intercept, the variance was biased high. Carroll et al. (2015) also demonstrated bias in the estimation of the Intrinsic Conditional Auto-Regressive (ICAR) parameter when using the INLA software and Clark and Dixon (2018) noted that as η and σ^2 increase in (4.2), the estimation bias in σ^2 becomes pronounced. Rue et al. (2009) recognize the bias in Laplace approximations, but state that it tends to be negligible in practice and only appear in pathological cases. However, as we will demonstrate, issues with truncation of the Taylor series approximation underlying the Laplace approximation are a concern for self-exciting Poisson models like (3.6) for parameter values that potentially arise in practice.

4.4 Extended Laplace Approximation

The primary issue in (4.8) when applied to (3.6) is that we are essentially conducting a Laplace approximation to an integral of the form

$$M = \int_{\mathbb{R}^{n_d \times T}} \exp\left(-g(\boldsymbol{Y}|\boldsymbol{Z}, \eta, \zeta, \sigma^2, \alpha)\right) dY, \tag{4.12}$$

where

$$g(Y|.) = \tfrac{1}{2}\boldsymbol{Y}^T \Sigma^{-1}(\theta)\boldsymbol{Y} - \left(\sum_{i=1}^{n_d} \sum_{t=1}^{T} -\eta Z(\boldsymbol{s_i}, t-1) - \exp(Y(\boldsymbol{s_i}, t)) + Z(\boldsymbol{s_i}, t)\log[\eta Z(\boldsymbol{s_i}, t-1) + \exp(Y(\boldsymbol{s_i}, t))]\right). \tag{4.13}$$

Clearly the size of $g(.)$ matches the dimension of the integration. As demonstrated in Shun and McCullagh (1995), typical Laplace approximations result in a necessarily biased approximation to the integral. In order to improve on this Shun and McCullagh (1995) and Evangelou et al. (2011) conduct an expansion of $\log(M)$ that is correct even when the dimension of the integral in (4.12) is equal to the sample size. Similar to a standard Laplace approximation, they conduct a Taylor's series expansion of $g(Y|.)$ as a function of $\boldsymbol{Y}$ yielding

$$\log M = \log\left[\exp(a_0) E \exp\left(a_i Y(\boldsymbol{s_i}, t) + a_{ij} Y(\boldsymbol{s_i}, t) Y(s_j, t)/2! + a_{ijk} Y(\boldsymbol{s_i}, t) Y(\boldsymbol{s_j}, t) Y(\boldsymbol{s_i}, t) + \cdots\right)\right], \tag{4.14}$$

where in (4.14) we are summing over all combinations of i, j, k and the a terms are functions of the partial derivatives of $g(\boldsymbol{Y}|.)$. The right hand side of (4.14) is the joint cumulant-generating function of $Y(s_i, t)$, $Y(s_i, t)Y(s_j, t)$, $Y(s_i, t)Y(s_j, t)Y(s_k, t), \cdots$ etc. Therefore, by definition the joint cumulant-generating function can also be written as the expansion given in equation (2.40) of Hall (1992) as

$$\sum_{k \geq 1} \frac{1}{k!} \sum_{i_1} \cdots \sum_{i_k} t^{(i_1)} \cdots t^{(i_k)} \kappa^{(i_1, \cdots, i_k)}(\boldsymbol{Y}) = \log E\left(\exp(\boldsymbol{t}^T \boldsymbol{Y})\right), \tag{4.15}$$

where $\kappa^{(i_1, \cdots, i_k)}$ is the multivariate cumulant taken over all partitions of $(i_1, \cdots, i_k)$.

In order to apply this to the self-exciting spatio-temporal model, we use the notation of Evangelou et al. (2011) letting $g_i(Y) = \frac{\partial g(Y)}{\partial Y(\boldsymbol{s_i}, t)}$ and $g_{i,j}(Y) = \frac{\partial^2 g(Y)}{\partial Y(\boldsymbol{s_i}, t) \partial(Y(s_j, t))}$. We will also let $g_{\boldsymbol{Y}}$ be the gradient of g and $g_{\boldsymbol{YY}}$ be the Hessian and $g^{i,j}$ be the (s_i, s_j) element of the inverse of the Hessian matrix. Aiding in our derivation is the fact that $g_{ijk} = 0$ unless $i = j = k$ for all partial derivatives of order 3 or higher and that all cumulants greater than 2 are zero as expectation is taken with respect to the Gaussian density.

Using this notation the first three terms of (4.15) yield

$$\log M \propto -\hat{g} - \frac{1}{2}|\hat{g}_{\boldsymbol{YY}}| - \sum_t \sum_i \frac{1}{8} \hat{g}_{iiii} (\hat{g}^{ii})^2 - \\ \sum_t \sum_i \frac{1}{48} \hat{g}_{iiiiii} (\hat{g}^{ii})^4 + \frac{1}{72} \sum_t \sum_{i,j \leq i} \hat{g}_{iii} \hat{g}_{jjjj} \left(6\left(\hat{g}^{ij}\right)^3 + 9 \hat{g}^{ii} \hat{g}^{jj} \hat{g}^{ij}\right), \tag{4.16}$$

where in (4.16) we denote $\hat{g}$ as the evaluation of the g function at $Y(s_i,t) = \mu(s_i,t)$ where $\mu(s_i,t)$ is the point that maximizes the Gaussian approximation to the full conditional density for Y in the denominator of (4.8). This expansion requires the derivation of the third, fourth and sixth derivatives of g which are

$$\begin{aligned} g_{iii} = -\Bigg[&\exp(Y(\boldsymbol{s_i},t))\left(\frac{Z(\boldsymbol{s_i},t)}{\lambda(Y(\boldsymbol{s_i},t))}-1\right) - 3\exp(2Y(\boldsymbol{s_i},t))\left(\frac{Z(\boldsymbol{s_i},t)}{\lambda(Y(\boldsymbol{s_i},t))^2}\right)+ \\ &2\exp(3Y(\boldsymbol{s_i},t)))\left(\frac{Z(\boldsymbol{s_i},t)}{\lambda(Y(\boldsymbol{s_i},t))^3}\right)\Bigg] \end{aligned} \tag{4.17}$$

$$\begin{aligned} g_{iiii} = -\Bigg[&\exp(Y(\boldsymbol{s_i},t))\left(\frac{Z(\boldsymbol{s_i},t)}{\lambda(Y(\boldsymbol{s_i},t))}-1\right) - 7\exp(2Y(\boldsymbol{s_i},t))\left(\frac{Z(\boldsymbol{s_i},t)}{\lambda(Y(\boldsymbol{s_i},t))^2}\right)+ \\ &12\exp(3Y(\boldsymbol{s_i},t)))\left(\frac{Z(\boldsymbol{s_i},t)}{\lambda(Y(\boldsymbol{s_i},t))^3)}\right) - 6\exp(4Y(\boldsymbol{s_i},t))\left(\frac{Z(\boldsymbol{s_i},t)}{\lambda(Y(\boldsymbol{s_i},t))^4}\right)\Bigg] \end{aligned} \tag{4.18}$$

$$\begin{aligned} g_{iiiiii} = -\Bigg[&\exp(Y(\boldsymbol{s_i},t))\left(\frac{Z(\boldsymbol{s_i},t)}{\lambda(Y(\boldsymbol{s_i},t))}-1\right) - 31\exp(2Y(\boldsymbol{s_i},t))\left(\frac{Z(\boldsymbol{s_i},t)}{\lambda(Y(\boldsymbol{s_i},t))^2}\right)+ \\ &180\exp(3Y(\boldsymbol{s_i},t)))\left(\frac{Z(\boldsymbol{s_i},t)}{\lambda(Y(\boldsymbol{s_i},t))^3)}\right) - 438\exp(4Y(\boldsymbol{s_i},t))\left(\frac{Z(\boldsymbol{s_i},t)}{\lambda(Y(\boldsymbol{s_i},t))^4}\right)]+ \\ &408\exp(5Y(\boldsymbol{s_i},t))\left(\frac{Z(\boldsymbol{s_i},t)}{\lambda(Y(\boldsymbol{s_i},t))^5}\right) - 120\exp(6Y(\boldsymbol{s_i},t))\left(\frac{Z(\boldsymbol{s_i},t)}{\lambda(Y(\boldsymbol{s_i},t))^6}\right)\Bigg] \end{aligned} \tag{4.19}$$

where $\lambda(Y(\boldsymbol{s_i},t)) = \exp(Y(\boldsymbol{s_i},t)) + \eta Z(\boldsymbol{s_i},t-1)$ in (4.18) and (4.17). The final pieces needed are $g^{i,i}$ and $g^{i,j}$ both of which can be found in the appropriate entry upon inverting $\Sigma^{-1}(\theta) + W$ where W is the same as defined in (4.9), which is the equivalent of $g_{i,i}$. Note that as $\Sigma^{-1}(\theta)$ is sparse by construction, therefore $\Sigma^{-1}(\theta) + W$ is also sparse and inversion can be performed using Cholesky factorization for sparse matrices at a low computational cost. Practically, it is only this inversion that makes the extended Laplace approximation computationally slower than the first order Laplace approximation.

The evaluation of (4.16) at this point brings the error from $\mathcal{O}(n^{-1})$ in the Laplace approximations to the marginals, to approximately $\mathcal{O}(n^{-3})$ when the higher order terms are included. While again this n is ill-defined, critically it is the same for both the original and the extended Laplacian, meaning if there is insufficient data to accurately estimate the marginals under (4.8), the further expansion may be an improvement. The inclusion of up to the sixth order terms is consistent with what was done in Raudenbush et al. (2000) though, as we will empirically demonstrate, in most cases g_{iiiiii} can be neglected.

Two points become clear through examining the expansion given in (4.16) and (4.15). The first is that additional terms added to (4.16) will contain $(\hat{g}^{ii})^k$ for $k > 4$. As can be empirically demonstrated, raising σ^2 near 1 and increasing η increase the number of elements of $\hat{g}^{ii}$ that are greater than 1 and also increase the maximum $\hat{g}^{ii}$ value. The second point that can be empirically demonstrated is that increasing η and σ^2 increase the summation terms in (4.16) meaning that there likely will be a bias in estimation of σ^2 and potentially η. Therefore, if these terms are not accounted for, as in the first order Laplace approximation, then the parameter estimates will be biased and credible intervals formed around the posterior mode may not cover the generating values.

4.4.1 General Algorithm For Conducting Bayesian Inference Using Higher Order Laplace Approximation

Here we will outline the general algorithm for using (4.16) to conduct an approximate Bayesian inference for the set of parameters, $\theta = \left(\alpha, \eta, \zeta, \sigma^2\right)$. The first task is finding the mode of $\pi(\theta|\boldsymbol{Z})$. First we fix a value of θ and for that value of θ find the value of $\boldsymbol{Y}^* = \boldsymbol{\mu}^*$ that maximizes (4.9). This is accomplished through repeatedly solving $\left(\Sigma^{-1}(\theta) + \text{diag } k(\mu^*(s_i,t))\right)\boldsymbol{\mu}^* = f(\boldsymbol{\mu}^*)$ where $f(\boldsymbol{\mu}^*)$ is the vector of evaluations of f given in (4.9). The sparsity of $\Sigma^{-1}(\theta) + \text{diag } (k(\mu^*(s_i,t)))$ makes this task extremely fast.

This value, $\boldsymbol{Y}^*$, is then used to evaluate (4.17), (4.18), and (4.19), giving an approximation to the Log-likelihood given in (4.16). As a point of comparison, on a 10 × 10 lattice wrapped on a Torus with 100 observations, finding $\boldsymbol{Y}^*$ and computing (4.16) take approximately 1-1.5 seconds. Using finite differences, the Hessian at that point can then be approximated. This takes an additional 32 evaluations if one covariate is in the model. A Newton-Raphson algorithm can then be used to find the mode of $\tilde{\pi}(\theta|\boldsymbol{Z})$. In the majority of problems considered, this took us approximately 4-5 steps. Finding the mode, again for the 10000 size data set described above this, generally, takes less than 10 minutes.

To initialize the Newton-Raphson algorithm, we recommend starting parameter values near the moment-based estimates derived using the second order calculations of the SP-INGARCH(0,1) process found in Chapter 3. Letting $\widehat{\cdot}_m$ be the method of moments based estimate we can find first $\widehat{\eta}_m = \rho(\bar{1})$ where $\rho(1)$ is the sample lag(1) autocorrelation at a given location. While this can be averaged directly over all locations, to reduce bias it is recommended to do a Fisher's Z-transformation, average the transformed autocorrelations, then do an inverse of the Z-transformation (see for example Silver and Dunlap (1987)). Letting $\Sigma_{i,i}$ be the marginal variance at location i of the unobserved process the moment based estimate of α is

$$\widehat{\alpha}_m = \frac{-1}{2}\left(\log\left(\widehat{\mathrm{Var}(Z)}(1-\widehat{\eta}_m^2)\right) - \bar{Z} + ((1-\widehat{\eta}_m)\bar{Z})^2\right) + 2\log\left((1-\widehat{\eta}_m)\bar{Z}\right) \tag{4.20}$$

where $\widehat{\mathrm{Var}(Z)}$ is the sample variance of the data.

While the starting values for α and η are estimated straight from the data, we cannot directly estimate starting values for σ^2 and ζ. We recommend starting ζ close to the edge of the parameter space as the parameter must be high in order to generate significant spatial correlation in the data. For example, if a 4 nearest neighbor structure is being used and there is spatial correlation in the data, we initialize $\zeta_{est} = .248$. In order to initialize σ^2 we use moment based estimates of the marginal variance, $\Sigma_{i,i}$ and the marginal covariance between nearest neighbors, say $\Sigma_{i,j}$. Moment-based estimates are

$$\widehat{\Sigma_{i,i}}_m = 2\log(\bar{Z}(1-\widehat{\eta}_m)) - 2\widehat{\alpha}_m \tag{4.21}$$

$$\begin{aligned}\widehat{\Sigma_{i,j}}_m &= \widehat{\rho(i,j)}(\exp(2\widehat{\Sigma_{i,i}}_m) - \exp(\widehat{\Sigma_{i,i}}_m) \\ &+ (\exp(-\widehat{\alpha}_m)/(1-\widehat{\eta}_m))\exp(\widehat{\Sigma_{i,i}}_m/2)) + \exp(\widehat{\Sigma_{i,i}}_m)) - \widehat{\Sigma_{i,i}}_m\end{aligned} \tag{4.22}$$

From here, a starting value of σ^2 can be found solving for $\boldsymbol{v}$ in $\left[(I_{n_d,n_d} - \zeta_{est}\boldsymbol{N})^{-1}\right]\boldsymbol{v} = \boldsymbol{s}$ where $\boldsymbol{s}$ is a vector with first entry $\widehat{\Sigma_{i,i}}_m$ and entries of $\widehat{\Sigma_{i,j}}_m$ at the position corresponding to the nearest neighbors of the first entry. For example, if the neighborhood structure is 10×10 wrapped on a lattice, $\boldsymbol{s}$ would have $\widehat{\Sigma_{i,i}}_m$ in the first position and $\widehat{\Sigma_{i,j}}_m$ at positions 2, 10, 11, and 91. The first entry of $\boldsymbol{v}$ can then serve as a starting value for σ^2. This value will be too low by construction but in general will allow the Newton-Raphson algorithm to converge to the posterior mode. If the algorithm fails to take a first step, the estimate of η can be slightly perturbed for example by subtracting .1 which usually eliminates most issues.

At the posterior mode, the posterior parameter space can then be efficiently explored using methods outlined in Rue et al. (2009). Credible intervals for individual elements of θ can be found either through assuming posterior normality and using the Hessian at the posterior mode or through the method outlined in Ferkingstad et al. (2015).

In summary, the primary advantage of using Laplace based techniques is computational speed. A single computation of the log-likelihood for a 10×10 neighborhood structure with $T = 100$ with $\boldsymbol{\alpha}$ as intercept only takes approximately 1 second with the primary computational cost being incurred in finding the mode of the Gaussian approximation to the denominator of (4.8). In using the extended Laplace approximation method in (4.16) there is an additional cost of about .5 of a second per evaluation. As a full exploration of the parameter space may take 600 to 1000 evaluations, the total cost incurred through using the expansion is about 5 to 6 minutes.

4.5 Fully Bayesian Approach

While the size of $\Sigma(\theta)$ makes MCMC techniques challenging, some properties of the model make it feasible to use a flexible modeling language such as Stan to perform inference. To do this, we follow closely the development given in Joseph (2016). First, note that we are trying to find

$$\pi(\theta|\boldsymbol{Z}) \propto \prod_{s_i,t} \pi(Z(s_i,t)|Y(s_i,t), Z(s_i,t-1), \eta)\pi(Y(s_i,t)|\boldsymbol{\alpha}, \sigma, \zeta)\pi(\eta)\pi(\boldsymbol{\alpha})\pi(\sigma)\pi(\zeta) \quad (4.23)$$

In the above, we are required to both sample from and calculate the density of the latent state, $\boldsymbol{Y}$ which requires evaluations of

$$\log(\pi(Y(\boldsymbol{s_i},t)|\boldsymbol{\alpha},\sigma,\zeta)) \propto \frac{1}{2}\log|\Sigma_f^{-1}(\theta)| - \frac{1}{2}(Y-\alpha)^T\Sigma_f^{-1}(\theta)(Y-\alpha) \quad (4.24)$$

To speed up computations, we note that the greatest computational cost in the sampling is the calculation of the determinant of the potentially very large matrix, $\Sigma_f^{-1}(\theta)$. However, the specific structure for $\Sigma_f^{-1}(\theta)$ allows us to follow Jin et al. (2005). First we note that $\log|\Sigma_f^{-1}(\theta)| = T\log|\Sigma^{-1}(\theta)|$ and $\log|\Sigma^{-1}(\theta)| = \frac{n_d}{2\log\sigma} + \log|I_{n_d,n_d} - \zeta N|$ where N is the neighborhood or adjacency matrix. Therefore, we can let $V\Lambda V^T$ be the spectral decomposition of N and then $|I_{n_d,n_d} - \zeta N| = |V||I_{n_d,n_d} - \zeta\Lambda||V^T| = \prod_{j=1}^{n_d}(1-\zeta\lambda_j)$ where λ_j are the eigenvalues of the neighborhood matrix which can be precomputed.

However, even using state of the art MCMC software such as Stan and precomputing all eigenvalues, MCMC still remains slow. For example, if $n_d = 100$ and $T = 100$, a single MCMC chain of length 5000 could take up to 3.5 hours to converge. In this example, the chain hadn't converged after 1000 iterations but exhibited no signs of non-convergence after 5000. In comparison, the Laplace approximation method of section 3, under the same set up, takes less than 10 minutes to find the find the parameters that maximize (4.8) and then another 15-20 minutes to evaluate the posterior parameter space. The extended Laplace approximation incurs an additional cost of about .5 of a second per evaluation and under the above conditions would add about 5 to 6 minutes of computations.

4.6 Simulation Study

In order to compare the Laplace approximation with the higher order Laplace approximation and the MCMC inferential methodology, we simulated data from model (3.6) on a

10×10 grid wrapped on a torus to reduce edge effects using a rook neighborhood structure. We further set $t \in \{1, 2, ..., 100\}$. The choice of these values was made to replicate potential real world situations. For example, counts aggravated over counties in a state or aggregated over neighborhoods in a major metropolitan area often have approximately 100 locations. For instance, there are 99 counties in Iowa, there are 96 named neighborhoods in Chicago, and there are 120 districts in Iraq. $T = 100$ would correspond to approximately two years of data observed weekly.

Next, we simulated from all 80 combinations of $\eta \in \{.1, .2, .3, .4, .5, .6, .7, .8\}$ and $\sigma^2 \in \{.1, .2, .3, .4, .5, .6, .7, .8, .9, 1\}$. For each choice of η and σ^2 we next set $\zeta = .245$ in order to generate significant spatial correlation as the spatial correlation. While we could have considered other choices of ζ note that the spatial correlation between two observations at the same point in time is

$$\text{Corr}(Z(\boldsymbol{s_i}, t)Z(\boldsymbol{s_j}, t)) = \frac{(\exp(\Sigma_{i,i} + \Sigma_{i,j}) - \exp(\Sigma_{i,i}))}{\left(\exp(2\Sigma_{i,i}) - \exp(\Sigma_{i,i}) + \frac{\exp(-\alpha)}{1-\eta}\exp(\frac{\Sigma_{i,i}}{2}\right)}, \tag{4.25}$$

where $\Sigma_{i,j}$ is the (i, j)th entry in the covariance matrix, $\Sigma(\theta)$. In order to have significant correlation in (4.25) ζ needs to be near the edge of the parameter space. The spatial correlation reflects a well known problem for CAR models and is presented in depth in Wall (2004). We further fixed $\alpha(s_i, t) = 0, \forall s_i, t$.

For each of the 100 combinations of parameters we found the values of $\hat{\sigma^2}$, $\hat{\eta}$ and $\hat{\zeta}$ that maximized (4.8) and (4.16) without the g_{iiiiii} term. In all cases, vague proper priors were used for η, σ^2, α and ζ. In order to form 95% credible intervals for the extended LA and LA(1) we inverted the negative Hessian at the posterior mode and used the centered credible intervals from the MCMC output. In all cases, estimates of η, α, and ζ using Laplace approximation and extended LA were generally unbiased and credible intervals in almost all cases covered these parameters. The difficulty lies in estimating the conditional variance, σ^2.

Table 4.1 gives the proportion of 95% credible intervals for the extended Laplace approximation that covered the generating parameter separated by the maximum $\hat{g}^{ii}$ from the extended Laplace approximation. As is clear, when $\hat{g}^{ii} > 2$ the credible intervals for σ^2 from the Extended Laplace rarely cover the generating parameters and when $\hat{g}^{ii} > .5$ the first order Laplace approximation generally fails to cover σ^2

	$0 < \hat{g}^{ii} < .5$	$.5 < \hat{g}^{ii} < 2$	$2 < \hat{g}^{ii}$
LA(1)	12/14	0/39	0/37
Extended LA	13/14	36/39	5/37

Table 4.1 Proportion of 95% credible intervals found by inverting the Hessian at the posterior mode that covered the generating parameter for ranges of maximum $\hat{g}^{ii}$ value.

Making the assumption that in practice (4.8) is preferable over (4.16) due to the simplicity of calculating (4.8) and both of these techniques are preferable over MCMC techniques as they are considerably quicker to fit, the results from Table 4.1 offer practical guidance for which technique to use in various situations. In Figure 4.1, we display, for all parameter combinations, the preferred method for inference.

As a point of reference, table 4.2 gives the results from three of the 80 parameter combinations compared when fit using the first order Laplace approximation, the extended Laplace approximation with and without the sixth order term and full MCMC. The benefits of adding the sixth order correction are not significant here, nor were they obviously significant in any of the other cases we considered. For the MCMC technique, the full parameter space was explored and then the posterior median was used as a point estimate for the parameter to compute relative bias. As shown in table 4.2, the LA(1) and extended LA both were considerably quicker and only for large values of η and σ^2 was the bias of the extended LA significantly larger than the MCMC.

As a general algorithm for fitting, we would generally attempt to fit the extended Laplace approximation first as computationally there is very little advantage to fitting the first order Laplace approximation. Upon examining the $\hat{g}^{ii}$ values, if there are a considerable number of them that are greater than 1 and the maximum is larger than 2 we would proceed to fit

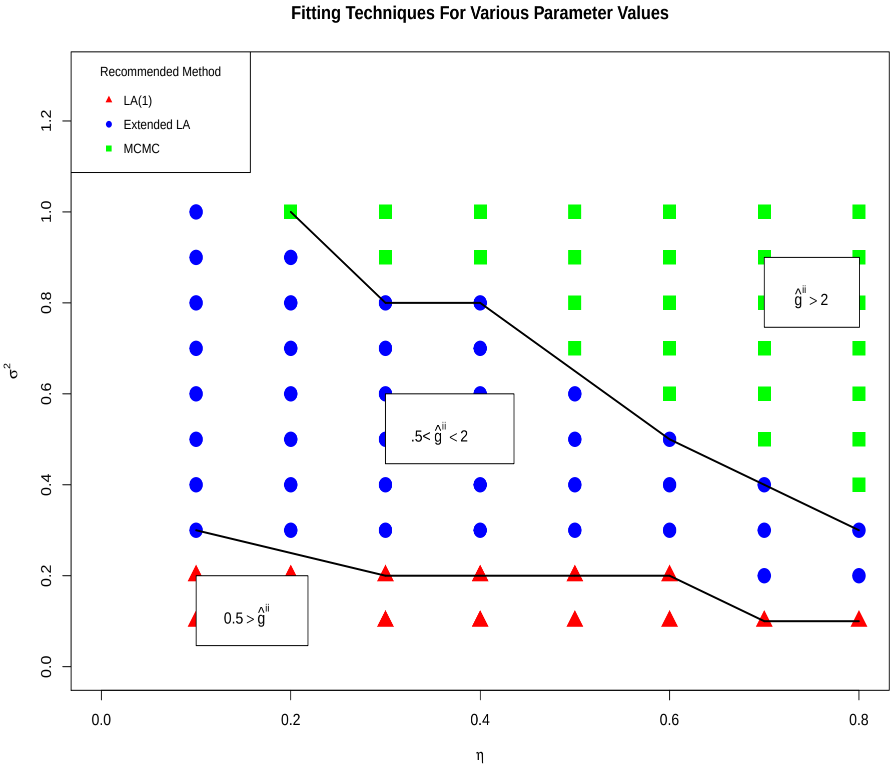

Figure 4.1 Recommended methodology for inference for subspace of parameters. The lines subset the parameter space by maximum $\hat{g}^{ii}$ values when fit using the extended Laplace approximation without the sixth order term. The recommendations are consistent with the results shown in table 4.1

	$\eta = .1$, $\sigma^2 = .4$	$\eta = .4$, $\sigma^2 = .6$	$\eta = .7$, $\sigma^2 = 1$
Relative Bias in LA(1)	.09	.25	.32
Time to Fit LA(1) (min.)	2	3	5
Extended LA Without 6th Order	.03	.02	.14
Extended LA With 6th Order	.03	.07	.25
Time to Fit Extended LA	3	5	5
MCMC	.02	.01	.04
Time to Fit MCMC	50-65	50-65	75-244

Table 4.2 Relative Bias and approximate times to find point estimates. Times are based on a single run and can be expected to vary in practice. Note that the MCMC time is for a full exploration of the parameter space. In general the fit times between LA(1) and the Extended LA are comparable while MCMC took 1 to 3 hours if three chains were run in parallel.

the full MCMC and not, generally, trust the output from the extended Laplace procedure. For example, Figures 4.2 and 4.3 show histograms of the $\hat{g}^{ii}$ values for $\eta = .4$, $\sigma^2 = .6$ and $\eta = .7$, $\sigma^2 = 1$. As evident in the Figure, the majority of the $\hat{g}^{ii}$ terms are less than one when $\eta = .4$ and $\sigma^2 = .6$ and the maximum is less than 2. However, when $\eta = .7$, $\sigma^2 = 1$ many of the $\hat{g}^{ii}$ terms are larger than 1 and the maximum is around 4. The impact to the likelihood is also drastically different as in the former case $\frac{1}{48}\sum_i \hat{g}_{iiiiii}(\hat{g}^{ii})^4 = -41$ and in the later $\frac{1}{48}\sum_i \hat{g}_{iiiiii}(\hat{g}^{ii})^4 = -440$, suggesting that further terms in the Laplace expansion would be necessary in the later case.

As depicted in figure 4.1 there is only a limited parameter space that the extended Laplace approximation method outlined in Section 1 should be used. While this may seem like a strong restrictions, simulated values from this region often result in extremely peaked and variable data, of which is rarely seen in the cases we envision the self-exciting Poisson CAR model being used. For example, if we simulate with $\sigma^2 = 1$ and $\eta = .7$, a situation where the extended LA fails to cover the generating parameters and the maximum $\hat{g}^{ii} > 2$, we would see data realizations such as shown in Figure 4.4. Practically, as depicted here, these parameter settings would correspond to a situation where there where very low counts

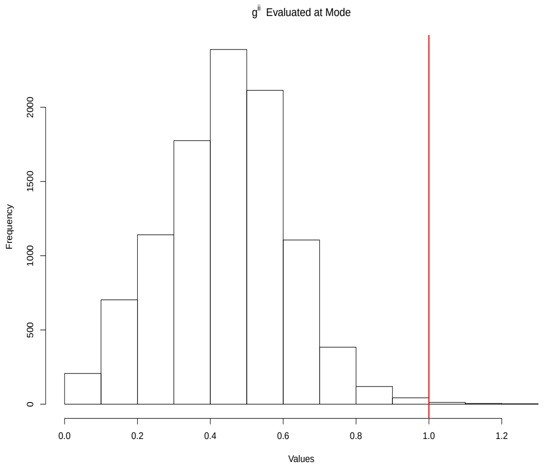

Figure 4.2 $\hat{g}^{ii}$ terms when extended Laplace approximation is used to fit data generated from $\eta = .4$, $\sigma^2 = .6$. The vertical red line is at 1. Here the majority of the $\hat{g}^{ii}$ terms are less than 1 and the maximum is less than 2 suggesting that higher, ignored, terms from Taylor series will only negligibly contribute to likelihood.

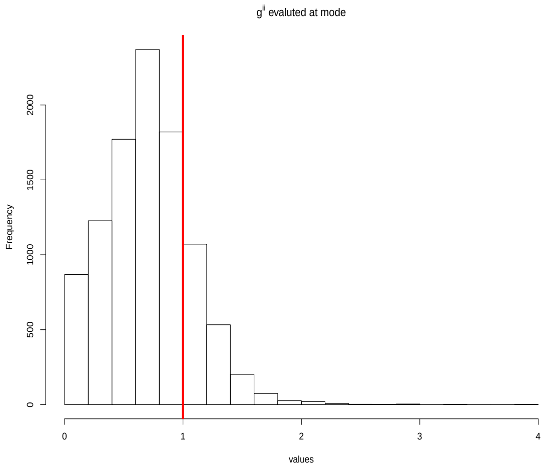

Figure 4.3 $\hat{g}^{ii}$ terms when extended Laplace approximation is used to fit data generated from $\eta = .7$, $\sigma^2 = 1$. Clearly here there are a significant number of terms that are greater than 1 and the maximum is nearly 4 suggesting that higher order, ignored, terms would significantly contribute to the likelihood.

followed by a massive spike and slow decay back to low counts. If the model were to be used to model something like the number of violent crimes in a neighborhood, it would be extremely unlikely that the data would follow this pattern. Further, the computational gains from using the extended Laplace approximation make it ideal for at least a first attempt at inference.

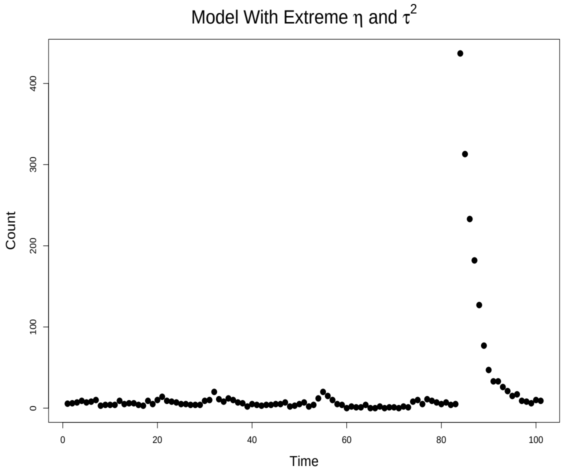

Figure 4.4 Counts from a simulated location with $\eta = .7$ and $\sigma^2 = 1$

4.7 Illustrative Example

In the following section we consider modeling violent crime in the city of Chicago in 2015 using the Self-Exciting Poisson CAR model. The Self-Exciting Poisson CAR model may be appropriate here as there are potentially multiple processes that are giving rise to the violence. Specifically, some crime may be due to a latent tension at a given location and there may be further violence that is due to copy-cat or retaliatory attacks. Previous work including Mohler (2013) analyzed this data in the absence of spatial correlation and concluded that self-excitement was present. Our purpose here is not to fully explore the complex nature of how and why violence occurred in Chicago, but rather to demonstrate how the extended LA could be used by social scientists to quickly explore competing theories within the Self-Exciting Poisson CAR framework allowing the practitioner to capture latent spatial correlation while allowing for the possibility of self-excitation.

The data used for the Chicago crimes is provided via https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2. We then aggregated all violent crimes both weekly and within specific predefined neighborhoods. We considered aggravated assault, aggravated battery, and homicides involving weapons as violent crimes. While there are certainly other violent crimes that could be considered, these crimes in particular seem likely to exhibit self-excitation within a given neighborhood as they potentially spur some form of retaliation. Similar data was used in both Mohler (2013) and Mohler (2014).

While there are no official neighborhoods in Chicago and counts can vary between 77 and 200 named areas, the city of Chicago publishes boundaries at https://data.cityofchicago.org/browse?q=neighborhoods&sortBy=relevance of 77 distinct neighborhoods. These are the neighborhoods we used in the analysis and appear to be consistent with historical norms for both locations and naming conventions within the city. We are not aware of previous statistical studies analyzing crime aggregated to neighborhood levels within the Chicago to compare the choice of neighborhood structure to. Mohler (2013) used data within a specific police beat, which corresponds, approximately, to athird of the size of one of the neighborhoods.

The resulting dataset consists of 9237 violent crimes that occurred in the city over 53 weeks (December 28 2014 - January 2, 2016). A spatial depiction of the crimes aggregated over neighborhoods is given in Figure 4.5. As evident in Figure 4.5, there appears to be spatial clustering in both the south and the western regions of the city. Spatial tests such as Moran's I applied to the aggregated data further give evidence to clustering in both space and time.

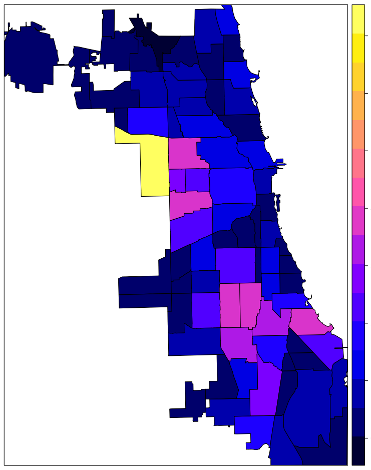

Figure 4.5 Total count of violent crimes for 2015 aggregated over neighborhood.

We fit the data using the model given in (4.26):

$$Z(\boldsymbol{s_i}, t) \sim \text{Pois}(\lambda(\boldsymbol{s_i}, t)) \tag{4.26}$$

$$E[Z(\boldsymbol{s_i}, t)] = \lambda(\boldsymbol{s_i}, t) \tag{4.27}$$

$$\boldsymbol{\lambda_t} = \exp(\boldsymbol{Y_t}) + \eta \boldsymbol{Z_{t-1}} \tag{4.28}$$

$$\boldsymbol{Y_t} \sim \text{Gau}(\boldsymbol{\alpha_t}, (I_{n_d,n_d} - C)^{-1} M). \tag{4.29}$$

A well-known phenomenon in criminology, as shown in Anderson (1987), is that higher temperatures are related to higher levels of both violent and non-violent crimes. To control for this, structure was placed on $\boldsymbol{\alpha_t}$. Specifically, for location (s_i, t), $\alpha(s_i, t) = \beta_0 + \beta_1 x_1(s_i, t) + \beta_2 x_2(s_i, t)$ where $x_1(s_i, t)$ corresponds to the observed average temperature in neighborhood s_i and time t and $x_2(s_i, t)$ corresponds to the log-population of location s_i at time t. Due to data limitations, we assume that temperature is constant across neighborhoods at time t and population is constant across time at neighborhood s_i. To aid in estimation of covariates, we centered and scaled the temperatures. We used census data for each neighborhood from the United States Census Bureau in 2010. For temperature, we used historic temperatures available from the Weather Underground website at www.wunderground.com.

Using the higher order Laplace approximation given in (4.16) we used a numerical estimate of the Hessian matrix allowing us to perform approximate Newton-Raphson maximization for the parameter space. In this and the subsequent inferences we used diffuse proper priors for $\boldsymbol{\theta}$. Specifically, $\pi(\sigma) \sim \text{Ca}^{+}(5)$, $\pi(\zeta) \sim \text{Unif}(0, .185)$, $\pi(\eta) \sim \text{Unif}(0, 1)$, and $\pi(\beta_0), \pi(\beta_1), \pi(\beta_2) \sim \text{Gau}(0, 1000)$. Where Ca^{+} is a half-Cauchy. The parameter space of ζ is dictated by the largest eigenvalue of the spatial adjacency neighborhood, in this case the largest eigenvalue is approximately 5.4 constraining $\zeta \leq .185$. On a Surface Pro 3, the posterior mode was found using the statistical software R in under 10 minutes. The observed maximum was found at $\hat{\theta} = (.52, .179, .50, -5.6, .18, .49)$.

The positive value of β_1 observed here echoes the findings of Anderson (1987) that increasing temperatures increase the probability of violence occurring. Specifically, because of the structure of model (4.26), if, for a given neighborhood, the temperature changes from 50 degrees Fahrenheit to 90 degrees Fahrenheit, the model would suggest that the expected number of violent crimes, due to temperature alone, would increase by a factor of 2 when controlling for self-excitement in the model.

The interpretation of η differs slightly than the large scale parameters in $\boldsymbol{\alpha}$. A value of .49 that each violent events at time period t raises the expected number of events at time period $t + 1$ by .49. In other words, if there were 10 violent events in week 1 at a given location we would expect there to be 5 events in week 2 that were ’copy-cat’ or inspired by the violence in week 1.

In order to generate credible intervals, we assumed the posterior was Gaussian and invert the negative Hessian at the posterior mode to find credible intervals for the marginal density of each element of $\boldsymbol{\theta}$. Alternatively, the posterior space could be explored as in Rue et al. (2009), however, in this, and other cases we considered, $\pi(\boldsymbol{\theta}|\boldsymbol{Z})$ was generally quadratic. As the Hessian was already found in the Newton-Raphson based optimization used earlier, calculating credible intervals comes at no additional computational cost. 95% credible intervals of $\pi(\boldsymbol{\theta}|\boldsymbol{Z})$ were $\sigma^2 \in (.43, .61)$, $\zeta \in (.176, .182)$, $\eta \in (.47, .53)$, $\beta_0 \in (-6.3, -4.9)$,

$\beta_1 \in (.09, .27)$, and $\beta_2 \in (.42, .55)$. Credible intervals for each parameter are given in 4.4. Figure 4.6 shows the $\hat{g}^{ii}$ terms from the posterior mode. As seen in this figure, the majority of the terms are less than one and the maximum is less than two suggesting that the extended LA will be more appropriate than the first order Laplace approximation.

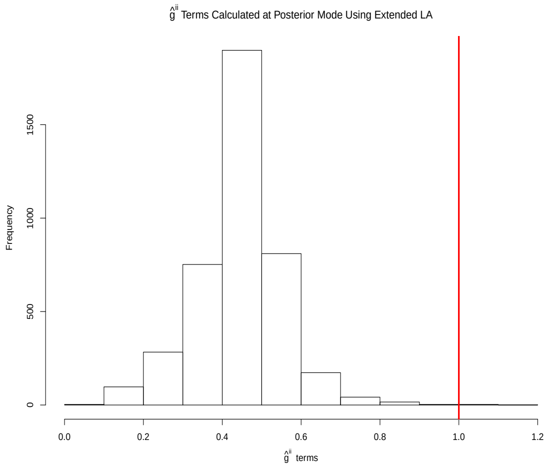

Figure 4.6 Histogram of $\hat{g}^{ii}$ terms from posterior mode of Chicago Data. While the maximum is greater than 1 it is not greater than 2 and the preponderance of the terms are less than one.

To compare this to the standard Laplace approximation we next fit to (4.26) using the first-order Laplace approximation method. We again used a numerical approximation of the Hessian and used a Newton-Raphson method to maximize the posterior. Using this inferential technique the parameters were maximized at $\hat{\theta} = (.38, .180, .50, -5.6, .17, .50)$. Gaussian approximations to the posterior marginals were again found through inverting the negative Hessian as $\sigma^2 \in (.33, .43)$, $\zeta \in (.178, .183)$, $\eta \in (.47, .53)$, $\beta_0 \in (-5.7, -5.4)$, $\beta_1 \in (.11, .23)$, and $\beta_2 \in (.48, .50)$. As is seen in Table 4.3 clearly the largest difference in the point estimation is in σ^2 as the point estimate using LA(1) is over two standard deviations from the estimate using the extended LA method. Furthermore, 95% credible intervals for σ^2 do not overlap, as seen in 4.4.

Finally, to compare the extended Laplace approximation to an MCMC technique we fit the model approach using the rStan software of Gelman et al. (2015) using the technique outlined in section 5. Three chains were run, starting at different locations in the parameter space. The chains were run for 10000 iterations each. Stan uses the first half of the iterations for warm-up, resulting in 15000 posterior samples for each parameter. Convergences was determined through examining the $\hat{R}$ values as well as through visual examination of the trace plots and eliminating all divergent transitions, see e.g. Betancourt (2017). After 10000 iterations there was no evidence that the chains had not converged. The entire process, running the three chains in parallel, took 2 hours.

Point Estimates	σ^2	ζ	η	β_0	β_1	β_2
LA(1)	.38	.180	.50	-5.6	.17	.50
Extended LA	.52	.179	.50	-5.6	.18	.49
MCMC	.50	.179	.50	-5.6	.18	.49

Table 4.3 Point estimates of the parameters from fitting model (3.6) to the Chicago crime data. As evident, the Expanded LA and MCMC techniques are extremely similar, while LA(1) has a bias for σ^2.

95% Credible Intervals	σ^2	ζ	η	β_0	β_1	β_2
LA(1)	(.33,.43)	(.178,.183)	(.47,.53)	(-5.7,-5.4)	(.11,.23)	(.48,.50)
Extended LA	(.43,.61)	(.176,.182)	(.47,.53)	(-6.3,-4.9)	(.09,.27)	(.42,.55)
MCMC	(.42,.59)	(.176,.182)	(.47,.53)	(-6.3,-5.0)	(.09,.27)	(.42,.56)

Table 4.4 Comparison between 95 % credible intervals formed using LA(1), extended LA, and MCMC. Note that the 95 % credible intervals for LA(1) and the extended LA were formed through using finite differences to approximate the Hessian and then using a Gaussian approximation to the posterior.

Using MCMC, 95 % credible intervals were $\sigma^2 \in (.42, .59)$, $\zeta \in (.176, .182)$, $\eta \in (.47, .53)$, $\beta_0 \in (-6.3, -5.0)$, $\beta_1 \in (.09, .27)$, and $\beta_2 \in (.42, .56)$. A comparison of point estimates is given in Table 4.3 and a comparison of credible intervals found through MCMC and extended LA is given in Table 4.4. As is clearly evident, there is not a significant difference between the extended Laplace technique and MCMC, however the time to fit the model was drastically higher using MCMC. While LA(1) and the extended LA were fit in similar time, LA(1) appears to underestimate σ^2, which is consistent with what was expected examining the $\hat{g}^{ii}$ terms in Figure 4.6. R code and data for fitting both the extended Laplace as well as the MCMC using rStan to 4.26 is available at https://github.com/nick3703/ExtendedLA.

4.8 Discussion

In this manuscript we demonstrated how extending Laplace approximations significantly reduces and over a large parameter space eliminates the bias in self-exciting spatio-temporal models. Contrary to the statement in Rue et al. (2009) that primarily only pathological examples resulted in bias from using the first order Laplace approximation, in this manuscript we showed that inferential techniques based on first order Laplace approximations would result in incorrect credible intervals and biased point estimates for a model that arises naturally from assumptions in the social sciences. We note that Ferkingstad et al. (2015) also offers a copula based method for potentially correcting the bias, however, this takes the analysis out of the Laplace framework and it is unclear what proceeding along this line does to the asymptotics. Furthermore, in order to implement the methodology outlined in Ferkingstad et al. (2015) we would need to calculate the skew-normal approximation to $\pi(Y|\theta, Z)$ which would add to the computational burden.

We further showed how a fully Bayesian approach could be considered through exploiting the sparsity of the precision matrix of the spatio-temporal process model. Even with a fully Bayesian approach being possible, the main benefit of using an extended LA methodology for this model is in computational speed. While MCMC takes several hours, the entire process for the extended LA usually takes minutes. The datasets we considered here were moderately sized for spatio-temporal data, if, however, we used larger datasets we would expect there to be an even larger disparity in fitting time.

The obvious cost of using the extended LA methodology is it requires deriving potentially up to sixth order partial derivatives to compute (4.16). Also, under the methodology outlined in this manuscript, exploration of the parameter space would not be efficient for a higher number of covariates in the model. However, as demonstrated above, if a Gaussian approximation to the marginals were to be used the parameter space would not have to be fully explored and second order finite differences could be used to fairly quickly approximate the Hessian.

Finally, we demonstrated how this methodology can be applied to analyze crime in Chicago showing how both spatial and temporal covariates can be considered through placing structure on $\boldsymbol{\alpha}$ and in this instance matches the inference using MCMC techniques. Interestingly, the self-excitement value found in this analysis, $\hat{\eta} = .50$, is similar to what was found in Mohler (2013) where in one police beat, 55% of observed crime was found to be due to repeated actions, or self-excitement. While that manuscript did not consider exogenous covariates, our analysis would suggest that the self-excitement was present even when weather and population size were considered.

CHAPTER 5. Conclusion and Future Direction

5.1 Summary

We have proposed a general class of models that accounts for spatial correlation as well as self-excitement. These models account for multiple sources of variation that are often present in the modeling of violent events. In particular, they arise from the assumption that the underlying intensity at a particular location can be generated from a stochastic difference equation resulting in the statistical model

$$
\begin{aligned}
&Z(\boldsymbol{s_i}, t)|\lambda(\boldsymbol{s_i}, t) \sim \text{Pois}(\lambda(\boldsymbol{s_i}, t)) \qquad (5.1)\\
&E[Z(\boldsymbol{s_i}, t)] = \lambda(\boldsymbol{s_i}, t)\\
&\boldsymbol{\lambda_t} = \exp(\boldsymbol{Y_t}) + \eta \boldsymbol{Z_{t-1}} + \kappa \boldsymbol{\lambda_{t-1}}\\
&\boldsymbol{Y_t} \sim \text{Gau}(\boldsymbol{\alpha_t}, (I_{n_d,n_d} - \boldsymbol{C})^{-1}\boldsymbol{M}).
\end{aligned}
$$

In Chapter 2 we assume that $\kappa = 0$ and call the resulting model a Spatially Correlated Self-Exciting model. In this chapter we further consider relaxing the temporal independence in the latent field, $\boldsymbol{Y_t}$ and consider a latent process that evolves according to a reaction-diffusion partial difference equation. The resulting model is referred to as the Reaction Diffusion Self-Exciting model. We demonstrate that both models give rise to sparse precision matrices and allow us to use an efficient Laplace approximation method to conduct inference. We show how these models can be used to analyze to the number of violent events aggregated monthly and by socio-economic region in Iraq. These novel models offer a new method for the statistical modeling of crime and violence. Most importantly, as we show, they result from preexisting theories on how violence evolves in space and time and do not rely on unrealistic assumptions such as conditional independence given a latent process.

In Chapter 3 we present the SPINGARCH model, a generalization of the Spatially Correlated Self-Exciting model. In particular we look at the SPINGARCH(1,1) given in (5.1). We prove the existence of a unique stationary distribution with finite second order moments under suitable parameter constraints. This allows us to derive the temporal and spatial covariance and demonstrate how the SPINGARCH(1,1) model has much more flexible second order properties than the INGARCH(1,1) model while retaining interpretability. We demonstrate how precomputing the eigenvalues of the neighborhood matrix allows us to conduct Bayesian inference using off the shelf technology such as Stan. We then apply this to analyzing the number of burglaries per month in individual census blocks in Chicago. We conclude that the INGARCH(1,1) model is not able to capture any of the second order properties in the data while a modified version of the SPINGARCH(1,1) model is. Significantly, we believe this extends the practical use of the INGARCH(1,1) model to spatial regions where there is strong temporal (positive) correlation and smaller, though significant, spatial correlation.

Finally, in Chapter 4 we demonstrate how the Laplace approximation method given in Chapter 2 results in bias in the estimation of σ and the bias increases as $\eta \to 1$. We use an extended Laplace approximation method to correct the bias and give a range of parameters where we recommend the extended Laplace approximation method be used. We show how, while a fully Bayesian method is possible along the lines of the work done in Chapter 3, we empirically demonstrate how the extended Laplace method is orders of magnitude faster. This methodology extends the usability of these models to remote field locations, such as military deployments, where robust computational tools are not available and answers are needed quickly.

Overall, this dissertation extends modeling choices for analyzing how violence spreads over space and time. While the resulting models are novel in their combination of both latent and data model dependency, they only require minimal assumptions and have a relationship to preexisting models for time series modeling of count data. Most importantly, the inferential methodology presented for the SPINGARCH(0,1) or Spatially Correlated Self-Exciting model, allows for ease of use.

5.2 Future Work

We believe that there is more work that can be done in extending applications, theory, and computations associated with the SPINGARCH(1,1) model. While the primary application we have considered is the spatio-temporal modeling of violence, there are other examples of data that can be expected to have both latent, spatial, dependence as well as self-excitement. One application that is of particular concern is the spatio-temporal modeling of the number of suicides. Phillips (1974) discusses the notion of 'copy-cat' suicides where it is the actual observed suicide that increases the probability of other individuals following suit. When this is combined with unique spatial factors, both large and small scale, the resulting appropriate model is the SPINGARCH(0,1) or SPINGARCH(1,1) model.

Outside of the social sciences, the modeling of certain weather phenomena may also benefit from the use of the SPINGARCH(1,1) model. For example, the statistical modeling of earthquakes has relied on capturing the notion of self-excitement as in Ogata (1988). These models may, as well, benefit by capturing small scale spatial variability.

Theoretically, while we have demonstrated the existence of a unique stationary distribution for the SPINGARCH(1,1) model and subsequently the Spatially Correlated Self-Excitement model, this proof relies on the temporal independence of the latent term, $\boldsymbol{Y_t}$. It remains to be shown what, if any, additional conditions are needed to ensure stationary exists for the Reaction Diffusion Self-Excitement model where the latent state is temporally correlated. In light of this, we further have not derived the second order properties of the Reaction Diffusion model as we cannot guarantee stationarity.

A further theoretical consideration is the impact of aggregation on inference. Throughout this dissertation we used data that was aggregated over space and time, however the actual violent events occur on continuous space and continuous time. While self-excitement becomes much more difficult to capture as we move to continuous time, Lindgren et al. (2011) offers a technique for approximating continuous space with Gaussian Markov random fields allowing the computational advantage of Laplace approximations to be used in the case of continuous time. A straightforward extension, then, would be to use the techniques of Lindgren et al. (2011) with the extended Laplace approximation methodology to conduct efficient inference for a continuous space, discrete time SPINGARCH(0,1) model. However, this still relies on temporal discretization.

Finally, computationally, we have demonstrated how an extended Laplace approximation technique is able to quickly and accurately conduct Bayesian inference for the SPINGARCH(0,1) model, we have no methodology for the general SPINGARCH(1,1) model. As the SPINGARCH(1,1) model can be viewed as a state-space model, methods such as the bootstrap filter given in Doucet et al. (2010) may offer a way to improve on MCMC techniques. Variational Bayes methods, such as those used in the modeling of violence in Afghanistan through the use of Log-Gaussian Cox Processes by Zammit-Mangion et al. (2013) may offer a method as well.

BIBLIOGRAPHY

Aitchison, J. and Ho, C. (1989). The multivariate poisson-log normal distribution. *Biometrika*, 76(4):643–653.

Anderson, C. A. (1987). Temperature and aggression: Effects on quarterly, yearly, and city rates of violent and nonviolent crime. *Journal of personality and social psychology*, 52(6):1161.

Army, U. and Corps, U. M. (2006). Counterinsurgency, fm 3–24, mcwp 3–33.5.

Athreya, K. B. and Pantula, S. G. (1986). Mixing properties of harris chains and autoregressive processes. *Journal of applied probability*, 23(4):880–892.

Augustin, N. H., McNicol, J., and Marriott, C. A. (2006). Using the truncated auto-poisson model for spatially correlated counts of vegetation. *Journal of agricultural, biological, and environmental statistics*, 11(1):1–23.

Bacry, E., Mastromatteo, I., and Muzy, J.-F. (2015). Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(01):1550005.

Baker III, J. A., Hamilton, L. H., Group, I. S., et al. (2006). *The Iraq study group report*. Vintage.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236.

Besag, J. (1981). On a system of two-dimensional recurrence equations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 302–309.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1):1–20.

Betancourt, M. (2017). Diagnosing biased inference with divergences. http://mc-stan.org/users/documentation/case-studies/divergences_and_bias.html. Accessed: 2017-07-17.

Braithwaite, A. and Johnson, S. D. (2015). The battle for baghdad: Testing hypotheses about insurgency from risk heterogeneity, repeat victimization, and denial policing approaches. *Terrorism and Political Violence*, 27(1):112–132.

Britt, C. L. (1994). Crime and unemployment among youths in the united states, 1958-1990. *American Journal of Economics and Sociology*, 53(1):99–109.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., and Riddell, A. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software*, 20:1–37.

Carroll, R., Lawson, A., Faes, C., Kirby, R., Aregay, M., and Watjou, K. (2015). Comparing inla and openbugs for hierarchical poisson modeling in disease mapping. *Spatial and spatio-temporal epidemiology*, 14:45–54.

Chung, F. R. (1997). *Spectral graph theory*, volume 92. American Mathematical Soc.

Clark, N. J. and Dixon, P. M. (2018). Modeling and estimation for self-exciting spatio-temporal models of terrorist activity. *Annals of Applied Statistics*. (in press).

Cox, D. R., Gudmundsson, G., Lindgren, G., Bondesson, L., Harsaae, E., Laake, P., Juselius, K., and Lauritzen, S. L. (1981). Statistical analysis of time series: Some recent developments [with discussion and reply]. *Scandinavian Journal of Statistics*, pages 93–115.

Cressie, N. (1992). Statistics for spatial data. *Terra Nova*, 4(5):613–617.

Cressie, N. and Wikle, C. K. (2015). *Statistics for spatio-temporal data*. John Wiley & Sons.

Davis, R. A., Holan, S. H., Lund, R., and Ravishanker, N. (2016). *Handbook of discrete-valued time series*. CRC Press.

Davis, R. A. and Liu, H. (2016). Theory and inference for a class of nonlinear models with application to time series of counts. *Statistica Sinica*, 26:1673–1707.

Doucet, A., de Freitas, N., and Gordon, N. (2010). *Sequential Monte Carlo methods in practice*. Springer.

Evangelou, E., Zhu, Z., and Smith, R. L. (2011). Estimation and prediction for spatial generalized linear mixed models using high order laplace approximation. *Journal of Statistical Planning and Inference*, 141(11):3564–3577.

Fearon, J. D. (2007). Iraq's civil war. *Foreign Aff.*, 86:2.

Ferkingstad, E., Rue, H., et al. (2015). Improving the inla approach for approximate bayesian inference for latent gaussian models. *Electronic Journal of Statistics*, 9(2):2706–2731.

Ferland, R., Latour, A., and Oraichi, D. (2006). Integer-valued garch process. *Journal of Time Series Analysis*, 27(6):923–942.

Finley, A. O., Banerjee, S., and Carlin, B. P. (2007). spbayes: an r package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of Statistical Software*, 19(4):1.

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

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). *Bayesian data analysis*, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA.

Gelman, A. et al. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, 1(3):515–534.

Gelman, A., Lee, D., and Guo, J. (2015). Stan: A probabilistic programming language for bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*, 40(5):530–543.

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, pages 733–760.

Genest, C. and Nešlehová, J. (2007). A primer on copulas for count data. *ASTIN Bulletin: The Journal of the IAA*, 37(2):475–515.

Goicoa, T., Ugarte, M., Etxeberria, J., and Militino, A. (2016). Age–space–time car models in bayesian disease mapping. *Statistics in medicine*.

Hall, P. (1992). The bootstrap and edgeworth expansion.

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

Heinen, A. (2003). Modelling time series count data: an autoregressive conditional poisson model.

Heinen, A. and Rengifo, E. (2007). Multivariate autoregressive modeling of time series count data using copulas. *Journal of Empirical Finance*, 14(4):564–583.

Hoffman, B. (2006). Insurgency and counterinsurgency in iraq. *Studies in Conflict & Terrorism*, 29(2):103–121.

Jin, X., Carlin, B. P., and Banerjee, S. (2005). Generalized hierarchical multivariate car models for areal data. *Biometrics*, 61(4):950–961.

Joe, H. (2008). Accuracy of laplace approximation for discrete response mixed models. *Computational Statistics & Data Analysis*, 52(12):5066–5074.

Johnson, S. D., Bernasco, W., Bowers, K. J., Elffers, H., Ratcliffe, J., Rengert, G., and Townsley, M. (2007). Space–time patterns of risk: a cross national assessment of residential burglary victimization. *Journal of Quantitative Criminology*, 23(3):201–219.

Johnson, S. D., Bowers, K., and Hirschfield, A. (1997). New insights into the spatial and temporal distribution of repeat victimization. *The British Journal of Criminology*, 37(2):224–241.

Joseph, M. (2016). Exact sparse car models in stan. `http://mc-stan.org/users/documentation/case-studies/mbjoseph-CARStan.html`. Accessed: 2017-07-17.

Kaiser, M. S. and Cressie, N. (1997). Modeling poisson variables with positive spatial dependence. *Statistics & Probability Letters*, 35(4):423–432.

LaFree, G. and Dugan, L. (2007). Introducing the global terrorism database. *Terrorism and Political Violence*, 19(2):181–204.

Laub, P. J., Taimre, T., and Pollett, P. K. (2015). Hawkes processes. *arXiv preprint arXiv:1507.02822*.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 619–678.

Lewis, E., Mohler, G., Brantingham, P. J., and Bertozzi, A. L. (2012). Self-exciting point process models of civilian deaths in iraq. *Security Journal*, 25(3):244–264.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498.

Linke, A. M., Witmer, F. D., and O'Loughlin, J. (2012). Space-time granger analysis of the war in iraq: A study of coalition and insurgent action-reaction. *International Interactions*, 38(4):402–425.

Liu, H. (2012). *Some models for time series of counts*. Columbia University.

Martínez-Beneito, M. A., López-Quilez, A., and Botella-Rocamora, P. (2008). An autoregressive approach to spatio-temporal disease mapping. *Statistics in medicine*, 27(15):2874–2889.

Meyn, S. P. and Tweedie, R. L. (2009). *Markov chains and stochastic stability*. Cambridge University Press.

Midlarsky, M. I., Crenshaw, M., and Yoshida, F. (1980). Why violence spreads. *International Studies Quarterly*, 24(2):262–298.

Mohler, G. (2013). Modeling and estimation of multi-source clustering in crime and security data. *The Annals of Applied Statistics*, 7(3):1525–1539.

Mohler, G. (2014). Marked point process hotspot maps for homicide and gun crime prediction in chicago. *International Journal of Forecasting*, 30(3):491–497.

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

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

Pease, K. et al. (1998). *Repeat victimisation: Taking stock*, volume 90. Home Office Police Research Group London.

Phillips, D. P. (1974). The influence of suggestion on suicide: Substantive and theoretical implications of the werther effect. *American Sociological Review*, pages 340–354.

Polson, N. G., Scott, J. G., et al. (2012). On the half-cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4):887–902.

Polvi, N. et al. (1990). Repeat victimization. *Journal of Police Science and Administration*, 17:8–11.

Polvi, N., Looman, T., Humphries, C., and Pease, K. (1991). The time course of repeat burglary victimization. *The British Journal of Criminology*, 31(4):411–414.

Porter, M. D., White, G., et al. (2012). Self-exciting hurdle models for terrorist activity. *The Annals of Applied Statistics*, 6(1):106–124.

Python, A., Illian, J., Jones-Todd, C., and Blangiardo, M. (2016). A bayesian approach to modelling fine-scale spatial dynamics of non-state terrorism: World study, 2002-2013. *arXiv preprint arXiv:1610.01215*.

Raphael, S. and Winter-Ebmer, R. (2001). Identifying the effect of unemployment on crime. *The Journal of Law and Economics*, 44(1):259–283.

Ratcliffe, J. (2010). Crime mapping: spatial and temporal challenges. In *Handbook of quantitative criminology*, pages 5–24. Springer.

Raudenbush, S. W., Yang, M.-L., and Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate laplace approximation. *Journal of computational and Graphical Statistics*, 9(1):141–157.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392.

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., and Lindgren, F. K. (2017). Bayesian computing with inla: a review. *Annual Review of Statistics and Its Application*, 4:395–421.

Short, M. B., D'ORSOGNA, M. R., Pasour, V. B., Tita, G. E., Brantingham, P. J., Bertozzi, A. L., and Chayes, L. B. (2008). A statistical model of criminal behavior. *Mathematical Models and Methods in Applied Sciences*, 18(supp01):1249–1267.

Shumway, R. H. and Stoffer, D. S. (2010). *Time series analysis and its applications: with R examples*. Springer Science & Business Media.

Shun, Z. and McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 749–760.

Silver, N. C. and Dunlap, W. P. (1987). Averaging correlation coefficients: should fisher's z transformation be used? *Journal of Applied Psychology*, 72(1):146.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van der Linde, A. (1998). Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. Technical report, Research Report, 98-009.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Tench, S., Fry, H., and Gill, P. (2016). Spatio-temporal patterns of ied usage by the provisional irish republican army. *European Journal of Applied Mathematics*, 27(03):377–402.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the american statistical association*, 81(393):82–86.

Wall, M. M. (2004). A close look at the spatial structure implied by the car and sar models. *Journal of statistical planning and inference*, 121(2):311–324.

Weidmann, N. B. and Ward, M. D. (2010). Predicting conflict in space and time. *Journal of Conflict Resolution*, 54(6):883–901.

Williams, P. (2009). *Criminals, militias, and insurgents: organized crime in Iraq*. Strategic Studies Institute.

Wolpert, R. L. and Ickstadt, K. (1998). Poisson/gamma random field models for spatial statistics. *Biometrika*, 85(2):251–267.

Zammit-Mangion, A., Dewar, M., Kadirkamanathan, V., Flesken, A., and Sanguinetti, G. (2013). Modeling and prediction in conflict: Afghanistan. In *Modeling conflict dynamics with spatio-temporal data*, pages 47–66. Springer.

Zeevi, A. and Glynn, P. W. (2004). Recurrence properties of autoregressive processes with super-heavy-tailed innovations. *Journal of applied probability*, 41(3):639–653.

Zhukov, Y. M. (2012). Roads and the diffusion of insurgent violence: The logistics of conflict in russia's north caucasus. *Political Geography*, 31(3):144–156.