Hierarcical Bayesian Modeling of Diffusing Point Process

Sunwoo Lim

December 9, 2022

1. Introduction

The main goal of this project is to build a spatio-temporal model to help us understand the diffusion pattern of a point process across the space and time. My main interest includes 1) how fast the disease diffuses to the population and 2) whether the number of patients tends to increase or decrease. Although the main application I considered is an infectious disease, but the model can be applied in any other diffusion pattern.

There are two or more possible approaches for this research question. I first thought of a direct modeling via a spatio-temporal point process model. I used the term "direct" because I directly model the spatio-temporal point pattern by the spatio-temporal point process model. The idea behind is an analogy between contagion of a disease and latent parents generating offspring points, which is an idea learnt in class. However, the details about the model have to be different from the ones dealt in class, which are focused on attraction / repulsion, not diffusion.

Another approach is binary classification, which is a detour. The main idea is discretizing the domain and modeling whether the virus has spreaded in each grid region using spatial information, so that I regard the grid where the virus has spreaded as the realization of the point process. I applied the model of Hefley et al. (2017), whose original application is on Chronic Wasting Disease (CWD), a diffusive pollutant found in cervids in Wisconsin.

I illustrate the contribution of the project. I found that this model has a serious disadvantage : disability to model time varying growth rate and diffusion rate (introduced later). However, I failed due to my lack of knowledge about the Partial Differential Equation (PDE). Thus instead, I found it is also worthwhile to perform a simulation study that resembles the spreading pattern of an infectious disease and figure out which pattern this model can illustrate and which pattern it cannot.

2. Methodology

2.1 Model structure

I review the model of Hefley et al. (2017). Let $Y_i, i = 1, ..., n$ be the Bernoulli response variable of the i^{th} observation at location $\boldsymbol{s_i} = (s_{1i}, s_{2i}) \in \Omega = [a, b] \times [c, d]$ and time point t_i. Note that index i contains information of both space and time. As in the Generalized Linear Model (GLM), let Y_i rely on the Bernoulli probability $p_i \in [0, 1]$, which relies on $u(\boldsymbol{s_i}, t_i)$, the spatio-temporal effect and $\boldsymbol{x_i}$, the covariate information of the i^{th} observation. Practically, $\boldsymbol{x_i}$ can be understood as average information of people on $\boldsymbol{s_i}$ at t_i. Detail of the model is as follows.

$$\text{Data Model : } Y_i|p_i \sim Ber(p_i) \text{ ; } p_i = g^{-1}\big(u(\boldsymbol{s_i}, t_i)e^{\boldsymbol{x_i'\beta}}\big), \text{ where } u(\cdot,\cdot) > 0, g : [0, \infty) \to [0, 1] \quad (1)$$

$$\text{Diffusion Model : } \frac{\partial}{\partial t}u(\boldsymbol{s}, t) = \Big(\frac{\partial^2}{\partial s_1^2} + \frac{\partial^2}{\partial s_2^2}\Big)[\mu(\boldsymbol{s})u(\boldsymbol{s}, t)] + \lambda(\boldsymbol{s})u(\boldsymbol{s}, t) \quad (2)$$

$$log(\mu(\boldsymbol{s})) = \boldsymbol{z(s)^T\alpha} \text{ ; } \lambda(\boldsymbol{s}) = \boldsymbol{w(s)^T\gamma} \quad (3)$$

$$\text{Prior Model : } \boldsymbol{\beta} \sim MVN(\boldsymbol{0}, 10^6\boldsymbol{I}) \text{ ; } \boldsymbol{\alpha} \sim MVN(\boldsymbol{0}, 10^6\boldsymbol{I}) \text{ ; } \boldsymbol{\gamma} \sim MVN(\boldsymbol{0}, 10^6\boldsymbol{I}) \text{ ; } \phi \sim TN(0, 10^6) \text{ ; } \theta \sim TN(0, 10^6) \quad (4)$$

$$\text{Boundary Condition : } u(\boldsymbol{s}, t) = 0, \forall t > 0 \quad (5)$$

$$\text{Initial Condition : } u(\boldsymbol{s}, 0) = \frac{\theta e^{\frac{-|\boldsymbol{s}-\boldsymbol{d}|}{\phi^2}}}{\int_{\boldsymbol{s}} e^{\frac{-|\boldsymbol{s}-\boldsymbol{d}|}{\phi^2}}} \quad (6)$$

In the data model (1), p_i is connected with the spatio-temporal effect $u(\boldsymbol{s_i}, t_i)$ and the covariate effect $e^{\boldsymbol{x_i'\beta}}$ via the link function $g : [0, \infty) \to [0, 1]$ and Hefley et al. (2017) uses $g^{-1}(x) = \sqrt{\frac{2}{\pi}}\int_0^x e^{x^2/2}dx$. The spatio-temporal effect $u(\boldsymbol{s_i}, t_i)$ is modeled by the diffusion equation in (2). $\mu(\boldsymbol{s}) > 0$ is the diffusion rate of location $\boldsymbol{s}$ and $\lambda(\boldsymbol{s})$ is the growth rate of location $\boldsymbol{s}$. $\boldsymbol{z(s)}$ and $\boldsymbol{w(s)}$ are vectors containing spatial covariates (e.g, forests and population). Note that in (3), nonparametric functions $\mu(\cdot)$ and $\lambda(\cdot)$ that consists $u(\cdot,\cdot)$ are parameterized by $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ respectively.

The model is suitable to illustrate the evolution of an infectious disease and parameters are interpretable. Diffusion $\mu(\boldsymbol{s})$ expresses how fast the disease tends to spread at $\boldsymbol{s}$. If $\mu(\boldsymbol{s})$ is big, the disease does not tend to stay at $\boldsymbol{s}$ for long. The growth rate $\lambda(\boldsymbol{s})$ is more easily interpretable. Big $\lambda(\boldsymbol{s})$ means that the number of patients near $\boldsymbol{s}$ explodes. $\lambda(\cdot)$ can take negative values and it can also have multimodalities. Later in the simulation study, I mainly illuminate the affect of this growth rate brings to the class probability $P(Y_i = 1)$. θ in the initial condition regards the baseline level of $P(Y_i = 1)$.

The model can be viewed as a combination of ecological diffusion model and Bayesian hierarchy. Ecological diffusion describes the random walk movement of individuals across time, where the movement probability is based on the local information. Figure 1 is an easy illustraiton of the ecological diffusion. The point process shows a random walk movement across each neighbor across time. Bayesian hierarchy is helpful for the uncertainty quantification of the parameters and $P(Y_i = 1)$.

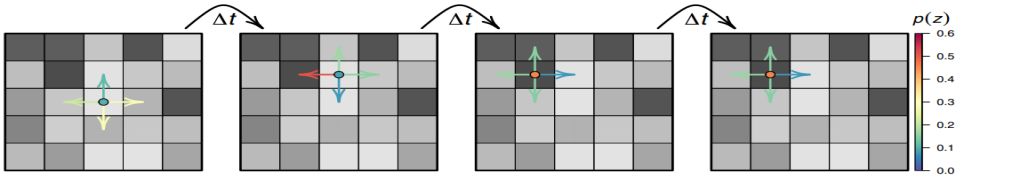

Figure 1 Illustration of an individual random walk based on local habitat information that results in the population-level distribution of individuals described by the ecological diffusion PDE (eqn 1). At each transition time (Δt), an individual (○) can move to a neighbouring habitat patch indicated by the arrows or stay in the current habitat patch. The probability $p(z)$ that an individual moves to a neighbouring patch or resides in the current patch, as indicated by the colour of arrows and dot, depends on the value of the habitat z. Darker habitat patches have a higher value of $p(z)$ and retain individuals longer, resulting in variable residence times. As Δt and the grid cell size become infinitesimally small $p(z)$ is proportional to the diffusion rate $\mu(\mathbf{s}, t)$ in eqn 1.

Figure 1: Easy illustration of ecological diffusion, figure from Hefley et al. (2017)

2.3 Metropolis-Hastings Algorithm

Let $\Theta := (\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \phi, \theta)$ be the parameter of interest. I sample from $p(\Theta|y)$ from the following blocked Gibbs sampler. Since Y is a Bernoulli random variable and priors for each of Θ are MVN or truncated normal, there is no semi-conjugacy. Thus, each parameter update includes a Metropolis-Hastings. Specifically, it is the random walk Metropolis, which uses a symmetric proposal distribution (e.g, normal distribution with mean : current state). Hefley et al. (2017) uses a normal proposal.

Step 1 : Set initial values for Θ

Step 2 : For $i < n_{mcmc}$, iterate the following:

Step 3 : Update $u(\boldsymbol{s}, t)$ and sample (θ, ϕ) jointly from $p(\theta, \phi|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ using random walk Metropolis-Hastings

Step 4 : Update $u(\boldsymbol{s}, t)$ and sample $\boldsymbol{\alpha}$ from $p(\boldsymbol{\alpha}|\theta, \phi, \boldsymbol{\beta}, \boldsymbol{\gamma})$ using random walk Metropolis-Hastings

Step 5 : Update $u(\boldsymbol{s}, t)$ and sample $\boldsymbol{\beta}$ from $p(\boldsymbol{\beta}|\theta, \phi, \boldsymbol{\alpha}, \boldsymbol{\gamma})$ using random walk Metropolis-Hastings

Step 6 : Update $u(\boldsymbol{s}, t)$ and sample $\boldsymbol{\gamma}$ from $p(\boldsymbol{\gamma}|\theta, \phi, \boldsymbol{\alpha}, \boldsymbol{\beta})$ using random walk Metropolis-Hastings

3. Simulation Analysis

I conduct three simulations. Basic setting of the simulation is the same: 30 different times points on a rectangular domain $[0, 1] \times [0, 1]$ with grid size 100×100. Regarding the MCMC, the MCMC sample size is 20000 with burn-in = 10000 and thinning = 10. In the first simulation, the number of points increase and in the second simulation, it decreases. However, they are common in that the points are clustered around one center. The third simulation 3 is more realistic in that there are multiple clusters of points. For instance, COVID-19 in South Korea showed multiple sources of spread (Daegu, Incheon, etc).

For each simulation setting, we obtain

1. Trace plot of the parameters $\Theta := (\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \phi, \theta)$
2. Pointwise posterior mean $\tilde{p}_i := E[Y_i|\boldsymbol{\Theta}]$
3. Uncertainty quantification via pointwise (across the space and time) 95% credible interval
4. Frequentist method of goodness of fit via computing $MAE(\tilde{p}, p)$ or CrossEntropy$(\tilde{p}, p)$.
5. Bayesian method of goodness of fit via posterior predictive check using $p(\boldsymbol{y_{new}}|\boldsymbol{y}) = \int_{\boldsymbol{\Theta}} p(\boldsymbol{y_{new}}|\boldsymbol{y}, \boldsymbol{\Theta}) p(\boldsymbol{\Theta}|\boldsymbol{y}) d\boldsymbol{\Theta}$.

3.1 Simulation 1

In simulation 1, the number of points increase based on the domain center. According to Figure 2, the growth rate $\lambda(\cdot)$ is bimodal, but the overall values do not vary much. The true probabilities $P(\boldsymbol{s}, t)$ well illustrates the diffusion pattern. At $t = 1$, there is a high probability at the center at the domain, but the probabilities diffuse fastly to the periphery. The bottom-right figure shows the true point process generated from the true probabilities $P(\boldsymbol{s}, t)$. The title of each panel shows how many points exist at time t. The number of points tends to increase and the points spread wider throughout the domain across time.

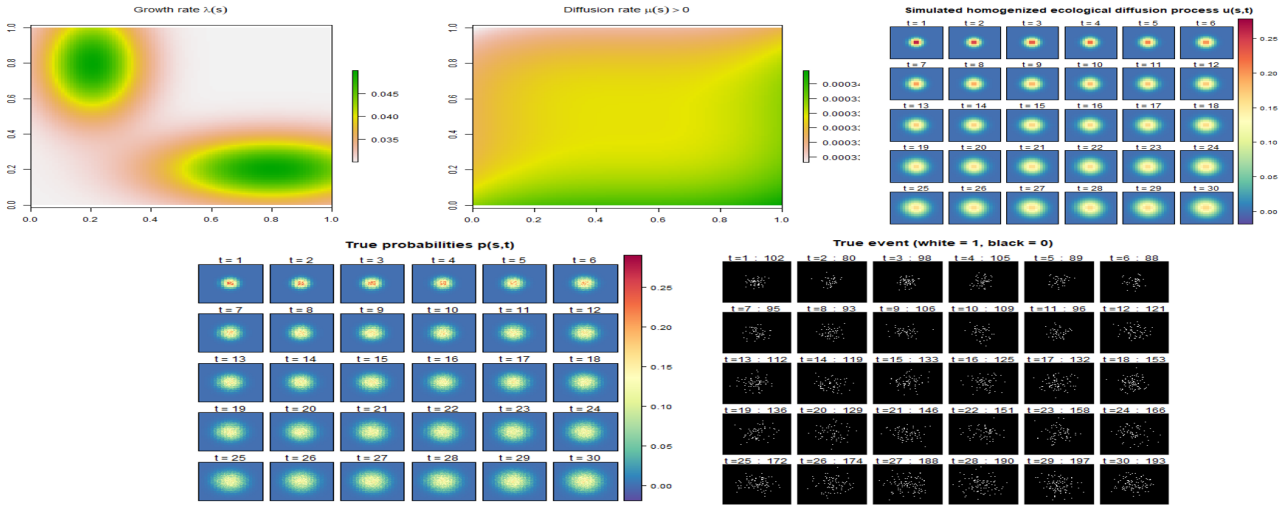

Figure 2: Growth rate, Diffusion rate, $u(\boldsymbol{s}, t)$, true probabilities and simulated binary events of simulation 1

Figure 3 contains the trace plot of the paramters in MCMC, pointwise posterior mean and quantiles of $P(\boldsymbol{s}, t)$. Burn in samples are got rid of. Due to the page limit, in later simulations, I do not post the pointwise quantiles of $P(\boldsymbol{s}, t)$ in the report.

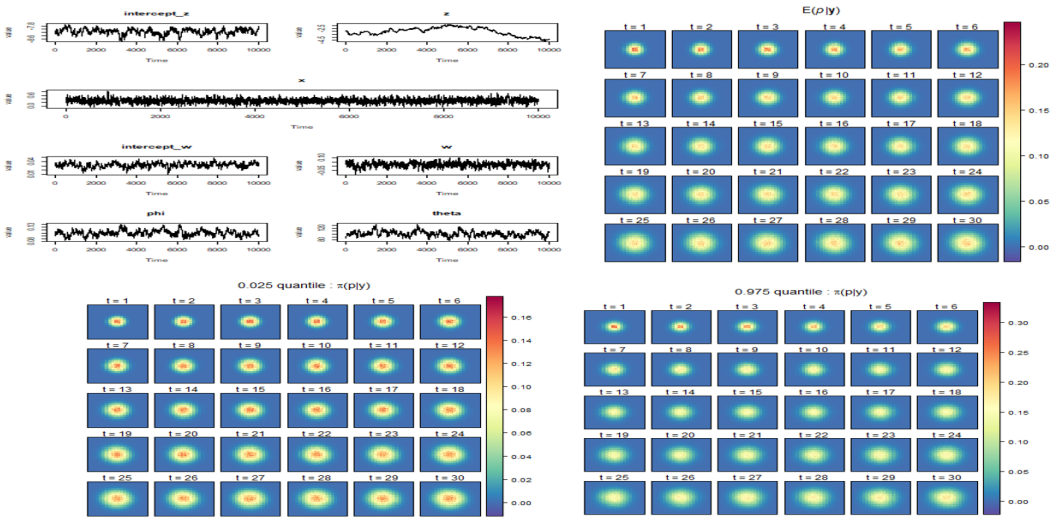

Figure 3: Trace plot of parameters, pointwise posterior mean of $P(\boldsymbol{s}, t)$, 0.025 and 0.975 quantile of $P(\boldsymbol{s}, t)$ of simulation 1

In the MCMC trace plot in Figure 3, except from the slope parameter for $log(\mu(\boldsymbol{s}))$, the algorithm seems to have converged. To compare the posterior mean and the true value of each parameter, all other components are very closely estimated except from the slope parameter of $log(\mu(\boldsymbol{s}))$: posterior mean $= -3.02$ versus the true value $= 0.1$. To see the goodness of fit, I calculated $MAE(\tilde{p}, p) = \frac{1}{300000}\sum_{i=1}^{300000}|\tilde{p}_i - p_i| = 0.0013$, the CrossEntropy$(\tilde{p}, p) = -\sum_{i=1}^{300000} p_i log(\tilde{p}_i) = 11832.31$ and the accuracy $=$ 0.974. Due to the class imbalance, it is better to compare the true probabilities versus the estimated probabilities and true Y and predicted Y via the plot, rather than the summary measures. The plot of true Y and predicted Y is in figure 4.

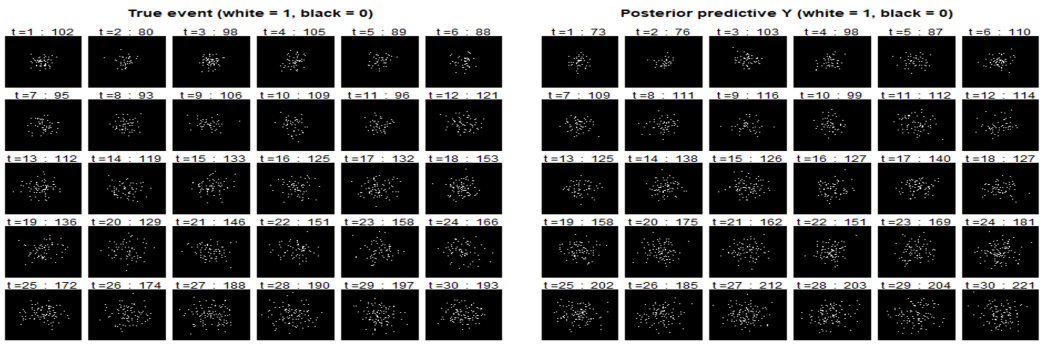

Figure 4: True point process (left) and predicted point process via posterior predictive distribution (right) in simulation 1

3.2 Simulation 2

The main difference between simulation 1 and simulation 2 is that in simulation 2, the growth rate $\lambda(\cdot)$ is negative for all regions. Also, since the peripheral part has smaller growth rate, $P(Y_i = 1)$ decrease in the periphery faster than the center of the domain. Figure 5 contains the information about the true probabilities and the true events.

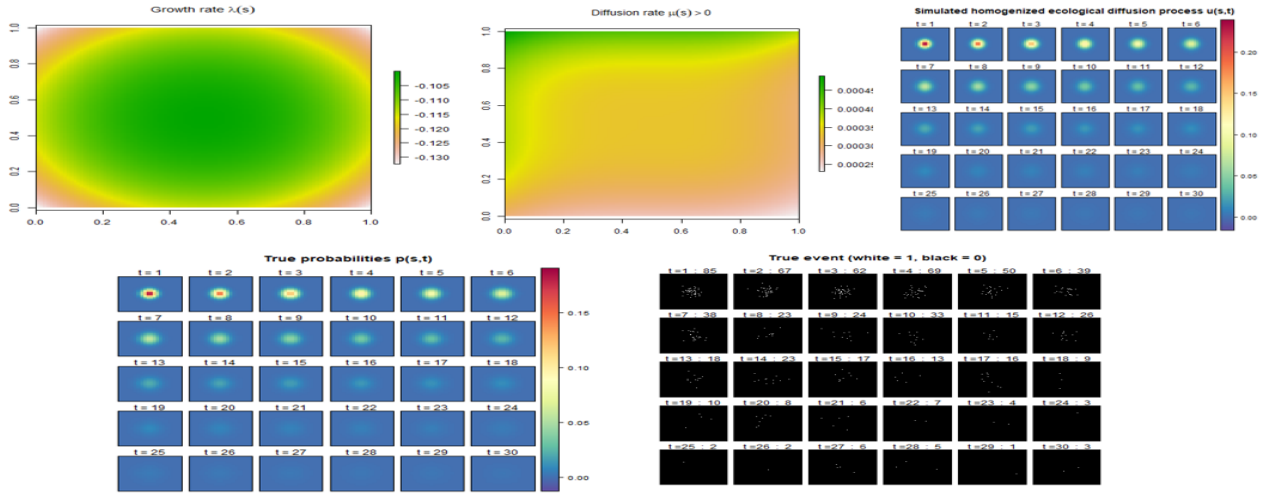

Figure 5: Growth rate, Diffusion rate, $u(\boldsymbol{s}, t)$, true probabilities and simulated binary events of simulation 2

Figure 6 contains the MCMC results in simulation 2. Also, burn in samples are got rid of. In the MCMC trace plot, some parameters seem to have converged, but some do not (slope parameter for $log(\mu(\boldsymbol{s}))$, the intercept and slope parameter for $log(\lambda(\boldsymbol{s}))$). Comparing the posterior mean and the true parameter, the slope parameter for $log(\mu(\boldsymbol{s}))$: true $= -2$, posterior mean $= -3.36$, parameters for $log(\lambda(\boldsymbol{s}))$: true $= (-0.05, -0.05)$ while posterior mean $= (-0.003, -0.09)$. Although there are issues in the parameter estimates, it still well estimates the true probabilities and well generates the new data. $MAE(\tilde{p}, p) = 0.0002$, the CrossEntropy$(\tilde{p}, p) = 2696.93$ and the accuracy $= 0.995$, but still, it is better to compare the true versus the estimate of $P(Y_i = 1)$ and Y_i's.

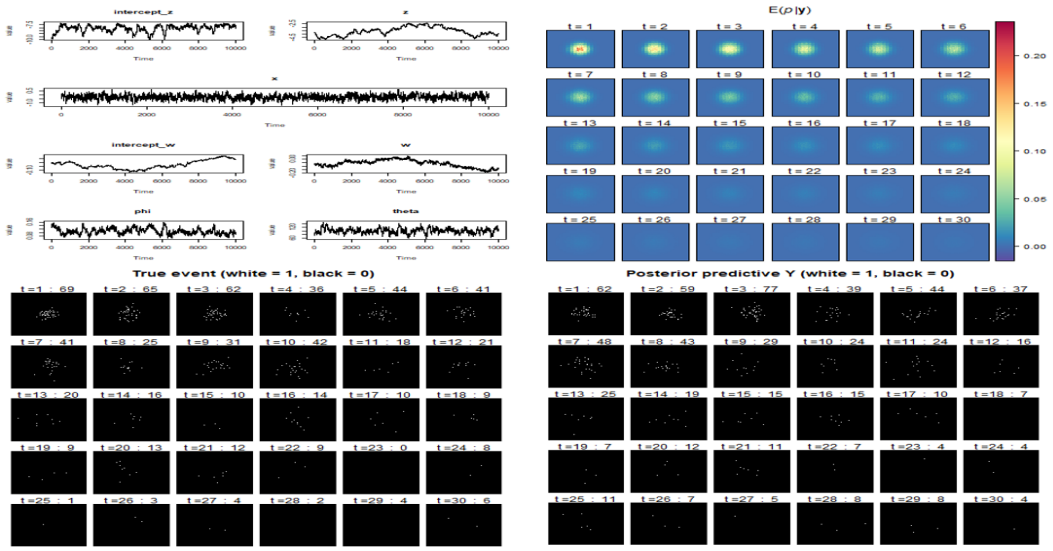

Figure 6: Trace plot of parameters, pointwise posterior mean of $P(\boldsymbol{s}, t)$, comparison of true vs predicted Y in simulation 2

3.3 Simulation 3

As illustrated in figure 7, the main difference between simulation 1,2 and simulation 3 is that in simulation 3, the growth rate $\lambda(\cdot)$ is significantly big in two modes, creating two hubs. The growth rate in simulation 1 is also bimodal, but the difference between $max_{\boldsymbol{s}}(\lambda(\boldsymbol{s})) = 0.045$ and $min_{\boldsymbol{s}}(\lambda(\boldsymbol{s})) = 0.035$ is small in simulation 1. In simulation 3, the difference between maximum growth rate (=1) and the minimum growth rate (=0.2) is 0.8. Looking at the true probabilities, at earlier stage, $P(Y_i = 1)$ is not appearently high in the lower left and the upper right in the earlier stage. However, at the later stage, the probability is very high at those two regions.

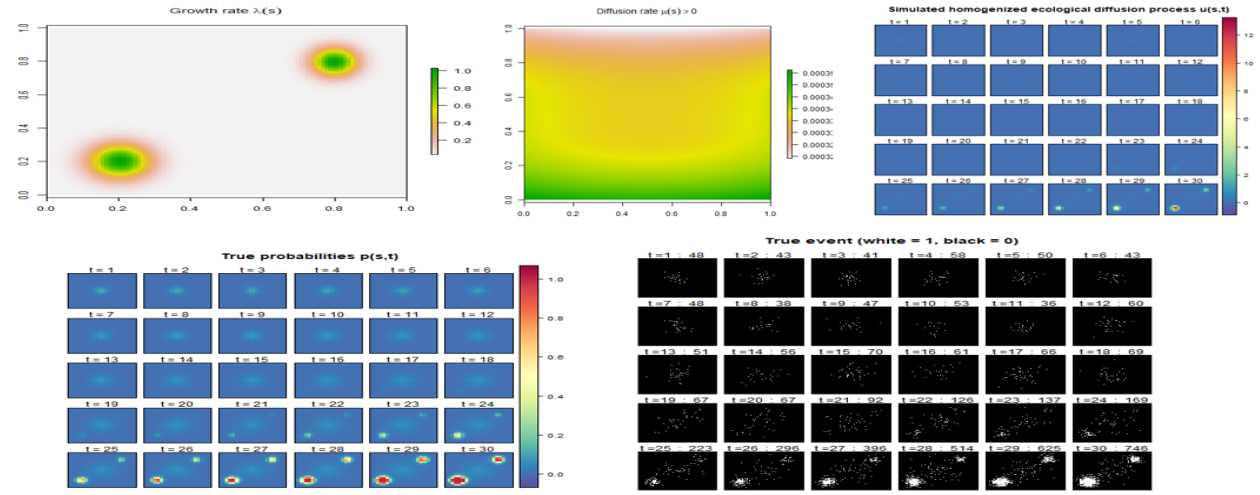

Figure 7: Growth rate, Diffusion rate, $u(\boldsymbol{s}, t)$, true probabilities and simulated binary events of simulation 2

Figure 8 contains the MCMC result in simulation 3. Some parameters seem to have converged, while some do not (details about comparison of true parameter values and posterior means is in the code). $MAE(\tilde{p}, p) = 0.000092$, $CrossEntropy(\tilde{p}, p) = 9736.28$ and the accuracy $= 0.980$. Again, despite the problems in parameter estimation, it still well estimates the diffusion pattern.

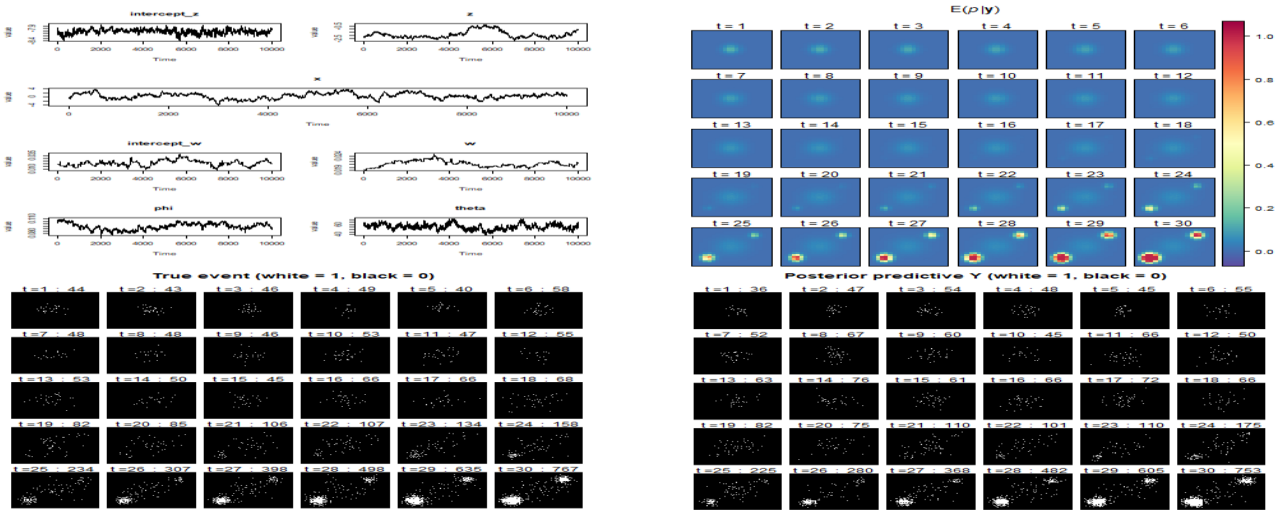

Figure 8: Trace plot of parameters, pointwise posterior mean of $P(\boldsymbol{s}, t)$, comparison of true vs predicted Y in simulation 3

4. Discussion and Further Research Topics

I conducted a simulation study of the model in Hefley et al. (2017), through which I mimicked the spreading pattern of an infectious disease. The advantages of the model include 1) suitable to represent the diffusion of a disease since it borrowed the idea of the ecological diffusion and 2) time efficiency due to the low dimensionality of the parameter space. Compared to the Generalized Linear Mixed Model (GLMM), which is similar to the ecological diffusion model has a high dimensionality issue : the order of the dimension of the parameter space is the same as the order of the sample size.

However, the model is not without disadvantages. The biggest disadvantage is that the growth rate $\lambda(\cdot)$ and the diffusion rate $\mu(\cdot)$ varys by space, but not by time (constant growth rate and diffusion rate with respect to time). This means that the spatio-temporal pattern in this model is only monotone with respect to time. Thus, this model is inappropriate for a long-term analysis, which also needs a complicated time pattern. This disadvantage is connected to the time efficiency advantage due to the low dimensionality, indicating the 'no free lunch principle'. Also, regarding the parameter update, since I use random walk Metropolis within Gibbs, proper choice of the variance parameter of the random walk proposal is crucial.

I feel bad that I could not find an appropriate real data. Years ago, there were spatio-temporal data containing location information of the COVID-19 confirmed patients across time. However, as almost all countries take less care about the disease, such data is no longer available online. This is why I chose to conduct a simulation analysis mimicking the real virus spread.

Later I hope to learn more about PDE's and develop a scalable methodology which gives solution to the time independent growth / diffusion rate issue.

5. Code and Data Availability Statement

The R codes, simulation data and results are uploaded in https://github.com/damelim/SpatiotempProject. The simulation results include runtime, posterior samples, MCMC acceptance rate, quantile information of $P(Y_i = 1)$, MAE, cross entropy, etc. I adapted the code in the supplementary material of Hefley et al. (2017).

References

Hefley, T. J., Hooten, M. B., Russell, R. E., Walsh, D. P., and Powell, J. A. (2017). When mechanism matters: Bayesian forecasting using models of ecological diffusion. *Ecology Letters*, 20(5):640–650.