A NONPARAMETRIC FRAMEWORK FOR TIME-DEPENDENT SIR MODELS WITH APPLICATION TO COVID DATA

by

Son Luu

Submitted in partial fulfillment of the requirements
for the degree of Masters of Science

at

Dalhousie University
Halifax, Nova Scotia
June 2023

Table of Contents

List of Figures

Abstract

Compartmental models, especially the Susceptible-Infected-Removed (SIR) model, have long been used to understand the behaviour of various diseases. Within this context, it can be beneficial to let parameters such as the transmission rate be time dependent functions. In this thesis, we attempt to build a nonparametric inference framework for stochastic SIR models with time dependent infection rate. The framework includes three main steps: likelihood approximation, parameter estimation and confidence interval construction. The likelihood function of the stochastic SIR model, which is often intractable, can be approximated using methods such as diffusion approximation or tau leaping. The infection rate is modelled by a B-spline basis whose knot location and number of knots are determined by a fast knot placement method followed by a criterion-based model selection procedure. Finally, a point-wise confidence interval is built using a parametric bootstrap procedure. The performance of the framework is observed through various settings for different epidemic patterns. The model is then applied to the Ontario COVID-19 data across multiple waves.

Acknowledgements

I would like to express my sincere gratitude to my advisor, Dr. Lam Ho, for his invaluable guidance and support throughout my master's program. I would also like to thank Dr. Edward Susko and Dr. Charith Kalu Arachchillage for their collaboration during my research. Finally, I want to thank Dr. Toby Kenney for giving me feedback on this thesis. The expertise and feedback from everyone here helped me to complete this research and write this thesis.

Chapter 1

Introduction

Compartmental models are a type of mathematical model used to study the spread of infectious diseases through a population. The basic idea of a compartmental model is to divide the population into different compartments based on their disease status, such as the famous Susceptible-Infected-Removed (SIR) model [11]. Each compartment represents a group of individuals with the same disease status, and the model tracks the flow of individuals between compartments over time. These models have been used to study a wide range of infectious diseases, including influenza [13], HIV/AIDS [1], plague [10], Ebola [9], and COVID-19 [16] and have been particularly useful for understanding the dynamics of epidemics, including the timing and size of outbreaks, as well as the impact of various control measures.

Many standard epidemic models assume that the epidemic parameters, such as the transmission rate and the recovery rate, are constant over time. In reality, the parameters of an epidemic can change over time due to various factors, such as changes in the behavior of the population, the implementation of interventions, and the emergence of new variants of the pathogen. Therefore, there is a need for time-dependent epidemic models that can capture the dynamics of these changing parameters.

For inference and prediction, there are two main types of compartmental models: deterministic and stochastic. In the deterministic model, the epidemic dynamics are described by a set of differential equations and the model parameters are often obtained by solving a least square problem. Some works have implemented this model type with time dependent rates [5, 18]. Deterministic models tends to work well for large populations and are computationally efficient. Stochastic models, on the other hand, takes into account the random variation in disease transmission within the population, which is useful for simulation. For these models, the number of individuals in each compartment is often assumed to follow a Markov process. Unfortunately, exact likelihood computation for stochastic compartmental models are typically intractable

or time consuming so likelihood based methods tends to use approximation methods in conjunction such as diffusion approximation [2, 6]. Because of this complex nature, there are not as many works that incorporate time dependent rates into a stochastic model.

With that in mind, this thesis explores a nonparametric inference framework for compartmental models with time dependent rates, specifically the SIR model with time dependent infection rate. There are two main underlying ideas: using a spline basis to estimate the true rates and using simpler processes to approximate the often intractable likelihood function. For inference, a fast spline knot placement method [21] is employed and assisted by a moving average rate estimate. Then various aspects of the model are examined in a simulation study including approximation type, model selection procedure and numerical considerations. Finally, the model is applied to estimate COVID-19 patterns in Ontario over multiple waves.

The rest of the thesis is structured as follows. Chapter 2 provides the necessary background; Chapter 3 describes the model including basis for likelihood approximation, parameter estimation, confidence interval and numerical considerations; Chapter 4 discusses the simulation study results; and Chapter 5 applies the proposed framework to the Ontario COVID-19 data.

Chapter 2

Background

This chapter will go over the definitions and properties of the stochastic processes involved in the model construction, B-spline basis [20], Wasserstein distance [19] and the parametric bootstrap procedure [7].

2.1 Stochastic SIR model

In this model, the population with an on-going disease is divided into three compartments: susceptible (S) for those who are not yet infected, infected (I) for those who are infected, and removed (R) for those who recovered or died from the disease. As illustrated in figure 2.1, there are two types of movements for an individual in the population: getting infected by the disease ($S \to I$) and recovering (or dying) from the disease ($I \to R$).

For a closed population of N individuals, let $S(t), I(t)$ and $R(t) = N - S(t) - I(t)$ be the number of susceptible, infected and removed individuals at time t, respectively. Then at time t, individuals move from S to I with rate $\beta(t)S(t)I(t)$ and from I to R with rate $\gamma(t)I(t)$. Here $\beta(t)$ and $\gamma(t)$ are the infection and recovery rates at time t, respectively.

Figure 2.1: Graph representation of the SIR model.

In the stochastic SIR model, these movements are often formally described by a bivariate continuous-time Markov process with transition probabilities as follow

Definition 2.1.1. *The stochastic SIR model assumes that* $X(t) = (S(t), I(t))$ *is a bivariate continuous-time Markov process satisfying*

$$\begin{aligned} P(X(t+dt) = (S-1, I+1)^\top | X(t) = (S, I)^\top) &= \beta(t)SIdt + o(dt) \\ P(X(t+dt) = (S, I-1)^\top | X(t) = (S, I)^\top) &= \gamma(t)Idt + o(dt) \end{aligned} \tag{2.1}$$

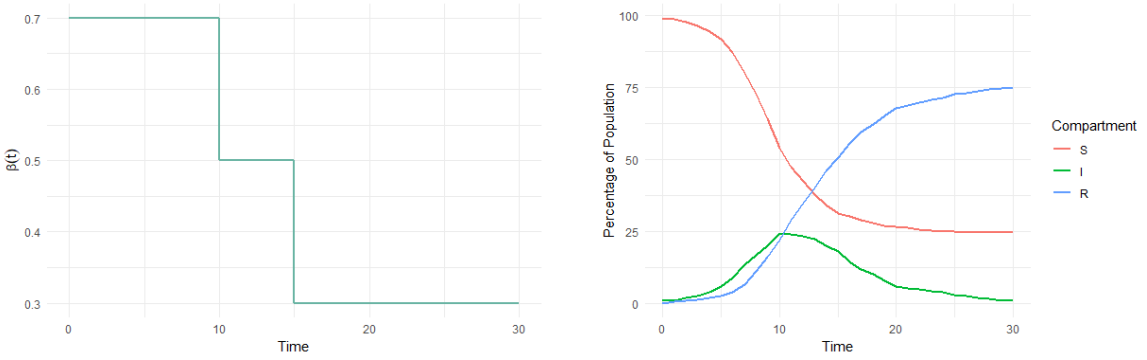
An equivalent definition describes the stochastic SIR model using Poisson processes. We shall use this definition since it will help explain a likelihood approximation method in the next chapter more naturally.

Definition 2.1.2. *The stochastic SIR model assumes that* $X(t) = (S(t), I(t))$ *is a bivariate continuous-time Markov process satisfying*

$$X(t) = X(0) + \begin{pmatrix} -1 \\ 1 \end{pmatrix} Pois_1 \left(\int_0^t \beta(s) \frac{S(s)I(s)}{N} ds \right) + \begin{pmatrix} 0 \\ -1 \end{pmatrix} Pois_2 \left(\int_0^t \gamma I(s) ds \right) \tag{2.2}$$

where $Pois_1, Pois_2$ *are independent standard Poisson processes.*

For this thesis, the focus is on the SIR model where the recovery rate γ is constant. Figure 2.2 shows the plot of an infection rate function and a corresponding sample path of each compartment.



(a) Infection rate function (b) Compartment sample path

Figure 2.2: A sample path of each compartment in a stochastic SIR model (by percentage of the total population) and its corresponding infection rate function with recovery rate $\gamma = 0.25$.

2.2 Diffusion process

Diffusion processes are continuous-time stochastic processes whose sample paths are continuous. A simple example of this is the Brownian motion. These processes are often described by a stochastic differential equation (SDE) as follow

Definition 2.2.1. *A diffusion process $X(t)$ is a continuous-time Markov process that satisfy the Ito SDE*

$$dX(t) = A(t, X(t))dt + L(t, X(t))dB(t) \tag{2.3}$$

where $B(t)$ is a multivariate Brownian motion, $A(t, x)$ and $L(t, x)$ are called the drift vector and diffusion matrix, respectively.

A simple interpretation of this is that the drift vector controls the mean of the process and diffusion matrix the variance. Figure 2.3 shows a sample path of a diffusion process satisfying the following SDE

$$dX(t) = -X(t)dt + dB(t) \tag{2.4}$$

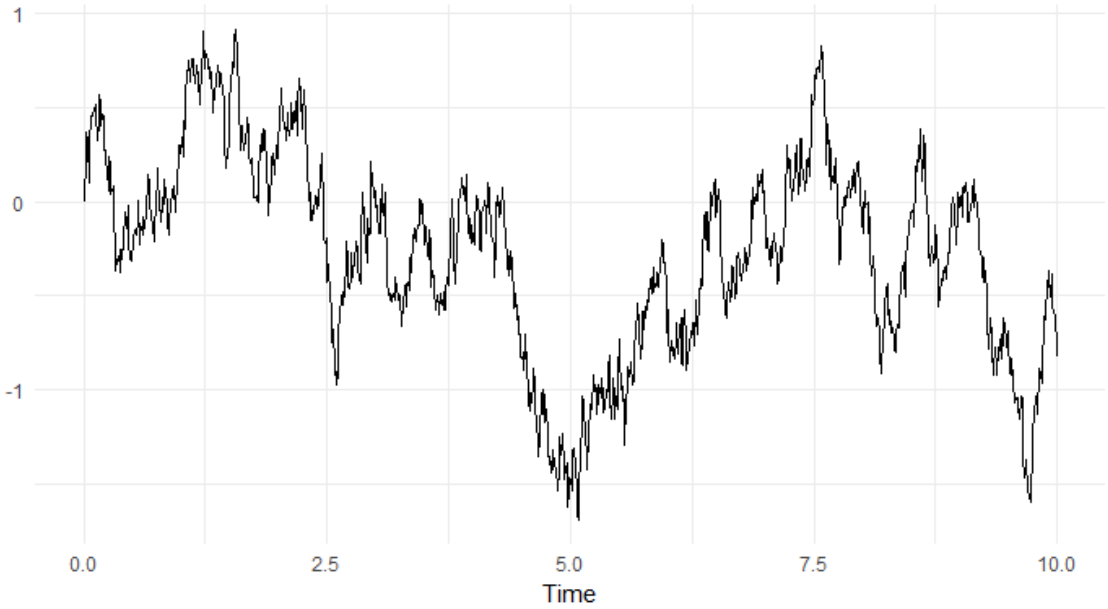


Figure 2.3: Sample path of a diffusion process.

In later chapters, a diffusion process will be used to approximate the stochastic SIR model.

2.3 B-spline

B-spline basis expansion is a well known method of curve fitting using piece-wise polynomials.

Definition 2.3.1. *Let $t_0 < t_1 < t_2 < \ldots < t_k < t_{k+1}$ be k points (known as knots) in an interval (t_0, t_{k+1}). A B-spline f of order $d+1$ is a piece-wise degree d polynomial defined by the formula*

$$f(t) = \sum_{i=1}^{k+d+1} c_i \varphi_{i,d}(t) \tag{2.5}$$

where $\varphi_{i,d}$ are degree d polynomials in (t_{i-d-1}, t_i)[1] called the basis functions and c_i are the corresponding coefficients.

For the construction of the basis functions, set $\tau_1 = \ldots = \tau_{d+1} = t_0$, $\tau_{i+d+1} = t_i$ for $i = 1, \ldots, k$ and $t_{k+1} = \tau_{k+d+2} = \ldots = \tau_{k+2d+2}$. Then

$$\varphi_{i,0}(t) = \begin{cases} 1 & \text{if } t \in [\tau_i, \tau_{i+1}) \\ 0 & \text{otherwise} \end{cases} \tag{2.6}$$

$$\varphi_{i,d}(t) = \frac{t - \tau_i}{\tau_{i+d} - \tau_i} \varphi_{i,d-1}(t) + \frac{\tau_{i+d+1} - t}{\tau_{i+d+1} - \tau_{i+1}} \varphi_{i+1,d-1}(t) \tag{2.7}$$

A note worthy feature of B-splines is that they have compact support which can speed up calculations [20]. Figure 2.4 shows an example of a B-spline basis on $(0, 1)$ with five equally spaced knots.

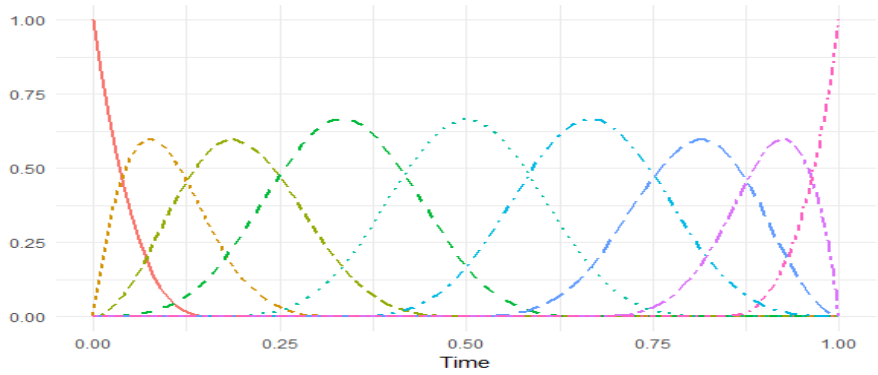


Figure 2.4: B-spline basis on $(0, 1)$ with five equally spaced knots.

[1] If $j < 0$, set $t_j = t_0$. If $j > k + 1$, set $t_j = t_{k+1}$

2.4 Wasserstein distance

The Wasserstein distance is commonly used for measuring the difference between two distributions.

Definition 2.4.1. *Let U, V be two $\mathbb{R}^d$-valued random variables. The Wasserstein-1 distance between them is defined as*

$$W_1(U, V) = \inf E(\|U - V\|) \tag{2.8}$$

where the infimum is over all possible couplings of U and V, i.e. all ways of jointly defining the two variables while respecting their marginal distribution. Note that the norm $\|\cdot\|$ is simply the Euclidean norm.

To measure the difference between two stochastic processes, we modify the above definition as follow

Definition 2.4.2. *Let $U(t), V(t)$ be two $\mathbb{R}^d$-valued stochastic processes on the interval $[0, T]$. Then the Wasserstein-1 distance between them is*

$$W_{1,T}(U, V) = \inf E(\|U - V\|_T) \tag{2.9}$$

where $\|X\|_T = \sup_{t\in[0,T]} \|X(t)\|$ and the infimum is over all possible couplings of $U(t)$ and $V(t)$, i.e. all ways of jointly defining the two processes while respecting their marginal distribution.

Next we have a lemma about Wasserstein distance and the point-wise law of processes

Lemma 2.4.3. *If the stochastic process sequence $U_n(t)$ and the stochastic process $V(t)$ on $[0, T]$ satisfy*

$$W_{1,T}(U_n, V) \xrightarrow{n\to\infty} 0 \tag{2.10}$$

Then for all $t \in [0, T]$ we have

$$U_n(t) \xrightarrow{d} V(t) \tag{2.11}$$

in other words, $U_n(t)$ converges in law to $V(t)$.

Proof. For all $t \in [0,T], \epsilon > 0$ we have

$$\begin{aligned} E(\|U_n - V\|_T) &= E(\sup_{t\in[0,T]} \|U_n(t) - V(t)\|) \geq E(\|U_n(t) - V(t)\|) \\ &\geq \epsilon P(\|U_n(t) - V(t)\| > \epsilon) \end{aligned} \tag{2.12}$$

Taking infimum over all couplings of $U_n(t)$ and $V(t)$ in (2.12) gives

$$W_{1,T}(U_n, V) \geq \epsilon \inf P(\|U_n(t) - V(t)\| > \epsilon) \tag{2.13}$$

Since $W_{1,T}(U_n, V) \xrightarrow{n\to\infty} 0$, we have

$$\inf P(\|U_n(t) - V(t)\| > \epsilon) \xrightarrow{n\to\infty} 0 \quad \forall\ \epsilon > 0 \tag{2.14}$$

Next, we have for all $u \in \mathbb{R}^d, \epsilon > 0$

$$\begin{aligned} F_{U_n(t)}(u) &= P(U_n(t) \leq u) \leq P(V(t) \leq u + \epsilon\mathbf{1}) + P(\|U_n(t) - V(t)\| > \epsilon) \\ &= F_{V(t)}(u + \epsilon\mathbf{1}) + P(\|U_n(t) - V(t)\| > \epsilon) \end{aligned} \tag{2.15}$$

where $\mathbf{1}$ is the vector of 1's and the inequalities here are element-wise. This is true since if $U_n(t) \leq u$ and $\|U_n(t) - V(t)\| \leq \epsilon$ then $V_n \leq u + \epsilon\mathbf{1}$. Applying this for $u - \epsilon\mathbf{1}$ with the role of $U_n(t)$ and $V(t)$ swapped, we have

$$F_{V(t)}(u - \epsilon\mathbf{1}) \leq F_{U_n(t)}(u) + P(\|U_n(t) - V(t)\| > \epsilon) \tag{2.16}$$

Combining (2.15) and (2.16) gives us

$$\begin{aligned} F_{V(t)}(u - \epsilon\mathbf{1}) - P(\|U_n(t) - V(t)\| > \epsilon) &\leq F_{U_n(t)}(u) \\ &\leq F_{V(t)}(u + \epsilon\mathbf{1}) + P(\|U_n(t) - V(t)\| > \epsilon) \end{aligned} \tag{2.17}$$

In (2.17), taking the infimum over all couplings of $U_n(t)$ and $V(t)$ gives

$$\begin{aligned} F_{V(t)}(u - \epsilon\mathbf{1}) - \inf P(\|U_n(t) - V(t)\| > \epsilon) &\leq F_{U_n(t)}(u) \\ &\leq F_{V(t)}(u + \epsilon\mathbf{1}) + \inf P(\|U_n(t) - V(t)\| > \epsilon) \end{aligned} \tag{2.18}$$

Note that the cdf's are not affected by the coupling since the marginals are fixed. This combined with (2.14) and letting $\epsilon \to 0,\ n \to \infty$ gives us

$$F_{U_n(t)}(u) \xrightarrow{n\to\infty} F_{V(t)}(u) \tag{2.19}$$

In other words, $U_n(t) \xrightarrow{d} V(t)$ as $n \to \infty$ for all $t \in [0,T]$. □

2.5 Parametric Bootstrap and Bootstrap confidence intervals

Consider an estimation problem where the quantity of interest is θ and the data generating distribution is F_θ. Now assume that we have a procedure to obtain an estimate $\hat{\theta}$ of θ. The parametric bootstrap is a method to estimate the distribution of $\hat{\theta}$. We accomplish that goal by performing the following steps:

1. Generate a sample from the approximate distribution $F_{\hat{\theta}}$.
2. Obtain an estimate $\hat{\theta}^*$ of $\hat{\theta}$.
3. Repeat steps 1 and 2 B times to get $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$.

With the sample of estimates in step 3, we can estimate various aspects of $\hat{\theta}$ such as the bias, variance and confidence interval.

2.5.1 Bootstrap confidence intervals

In this subsection, we define all types of bootstrap confidence intervals that are utilized in later sections.

Pivotal Interval

The $1-\alpha$ bootstrap pivotal interval is defined as

$$CI_{pivotal} = \left(2\hat{\theta} - \hat{\theta}^*_{(B(1-\alpha/2))},\ 2\hat{\theta} - \hat{\theta}^*_{(B\alpha/2)}\right) \tag{2.20}$$

where $\hat{\theta}^*_{(B\alpha)}$ denotes the $100\alpha^{th}$ percentile of $\hat{\theta}^*$. This interval works under the assumption that the distribution of $\hat{\theta}^* - \hat{\theta}$ should approximate that of the pivot $\hat{\theta} - \theta$.

Normal Interval

The $1-\alpha$ bootstrap normal interval is defined as

$$CI_{normal} = \left(\hat{\theta} - z_{\alpha/2}s_b,\ \hat{\theta} + z_{\alpha/2}s_b\right) \tag{2.21}$$

where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal and s_b is the bootstrap estimate of the standard error

$$s_b^2 = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}^*_b - \frac{1}{B}\sum_{r=1}^{B}\hat{\theta}^*_r\right)^2 \tag{2.22}$$

This interval works under the assumption that the distribution of $\hat{\theta}$ is close to normal, i.e. $\hat{\theta} \sim N(\theta, s^2)$.

Percentile Interval

The $1-\alpha$ bootstrap percentile interval is defined as

$$CI_{percentile} = \left(\hat{\theta}^*_{(B\alpha/2)},\ \hat{\theta}^*_{(B(1-\alpha/2))}\right) \tag{2.23}$$

This interval works under the assumption that there exists a monotonic transformation ρ such that $\rho(\hat{\theta}) \sim N(\rho(\theta), c^2)$.

2.5.2 Bias correction for Bootstrap confidence intervals

In many inference problems, especially nonparametric ones, there will be a certain amount of bias

$$b = E\hat{\theta} - \theta \tag{2.24}$$

To account for these biases, we will have to make some adjustments to the bootstrap confidence intervals. These adjustments often involve subtracting the bias, which is estimated by the bootstrap bias

$$\hat{b}^* = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}^*_i - \hat{\theta} \tag{2.25}$$

Pivotal Interval

The appearance of a bias does not affect this interval's main assumption, that is the distribution of the pivot $\hat{\theta} - \theta$ is close to the distribution of $\hat{\theta}^* - \hat{\theta}$. Therefore, the pivotal interval has already accounted for bias correction.

Normal Interval

When there is a bias term, the assumption for this interval becomes

$$\hat{\theta} \sim N(\theta + b, s^2) \tag{2.26}$$

Then, the bias corrected confidence interval will be

$$CI_{\text{corrected normal}} = \left(\hat{\theta} - \hat{b}^* - z_{\alpha/2}s_b,\ \hat{\theta} - \hat{b}^* + z_{\alpha/2}s_b\right) \tag{2.27}$$

Percentile Interval

In this case, the main assumption is reinterpreted as

$$\exists\ \rho \text{ monotonic: } \rho(\hat{\theta} - b) \sim N(\rho(\theta), c^2) \tag{2.28}$$

Then the bias corrected percentile confidence interval can be derived as follow

$$1 - \alpha = P(\rho(\theta) - z_{\alpha/2} \leq \rho(\hat{\theta} - b) \leq \rho(\theta) + z_{\alpha/2}) \tag{2.29}$$

$$= P(\rho(\hat{\theta} - b) - z_{\alpha/2} \leq \rho(\theta) \leq \rho(\hat{\theta} - b) + z_{\alpha/2}) \tag{2.30}$$

$$= P(\rho(\hat{\theta} - b)_{\alpha/2} \leq \rho(\theta) \leq \rho(\hat{\theta} - b)_{1-\alpha/2}) \tag{2.31}$$

Now since ρ is monotonic, it preserves quantiles so $\rho(\hat{\theta} - b)_\alpha = \rho(\hat{\theta}_\alpha - b)$ for all α. Therefore (2.31) becomes

$$1 - \alpha = P(\rho(\hat{\theta}_{\alpha/2} - b) \leq \rho(\theta) \leq \rho(\hat{\theta}_{1-\alpha/2} - b)) \tag{2.32}$$

$$= P(\hat{\theta}_{\alpha/2} - b \leq \theta \leq \hat{\theta}_{1-\alpha/2} - b) \tag{2.33}$$

Then the quantiles of $\hat{\theta}$ are estimated by the bootstrap sample while keeping in mind that there is a bias term

$$\forall\ \alpha: \quad \hat{\theta}_\alpha \approx \hat{\theta}^*_{B\alpha} - \hat{b}^* \tag{2.34}$$

Plugging (2.34) into (2.33) and replacing b with $\hat{b}^*$, we get

$$1 - \alpha = P(\hat{\theta}^*_{(B\alpha/2)} - 2\hat{b}^* \leq \theta \leq \hat{\theta}^*_{(B(1-\alpha/2))} - 2\hat{b}^*) \tag{2.35}$$

Hence, the formula for the bias corrected percentile confidence interval is

$$CI_{\text{corrected percentile}} = \left(\hat{\theta}^*_{(B\alpha/2)} - 2\hat{b}^*,\ \hat{\theta}^*_{(B(1-\alpha/2))} - 2\hat{b}^*\right) \tag{2.36}$$

Chapter 3

Methodology

This chapter introduces the model of interest, along with the methods for log likelihood approximation and parameter estimation.

3.1 SIR model construction

Consider the stochastic SIR model, as defined in chapter 2, with infection rate function $\beta(t)$ and constant recovery rate γ. Our goal is to estimate both $\beta(t)$ and γ using discretely observed data of the number of susceptible and infected individuals. To this end, a B-spline basis is used for modeling $\beta(t)$. In summary, the model can be written as follow

$$X(t) = (S(t), I(t)) : \text{ stochastic SIR model with rates } \beta(t), \gamma \tag{3.1}$$

$$X(t_1), X(t_2), \ldots, X(t_M) : \text{ observed states at times } t_1, t_2, \ldots, t_M \tag{3.2}$$

$$\gamma = \theta_1, \quad \beta(t) = \sum_{i=1}^{K+d+1} \theta_{i+1}\phi_{i,d}(t) \tag{3.3}$$

where K, d are the number of knots and degree of the B-spline basis, respectively, and θ_i are the coefficients.

3.2 Likelihood approximation

With our model set up, the likelihood function is

$$L_X(\boldsymbol{\theta}) = L_X(\theta_1, \ldots, \theta_{K+d+2}) = \prod_{i=1}^{M-1} P_{\boldsymbol{\theta}, t_i, t_{i+1}}(X(t_{i+1})|X(t_i)) \tag{3.4}$$

The biggest problem here is computing or approximating the transition probabilities in (3.4). Therefore, a stochastic process whose transition probabilities can be tractably approximated is used to approximate our SIR model.

3.2.1 Diffusion approximation

In this section, the diffusion process used to approximate our SIR is presented along with the convergence results. Now set $x(t) = (s(t), j(t)) = X(t)/N = (S(t)/N, I(t)/N)$. This rescaled process represents the proportion of susceptible and infected in the population. Using the new description, the state space of $x(t)$ can be viewed as "continuous" for large N, making the approximation to a diffusion process more natural. Next, consider the diffusion process $z(t) = (s(t), j(t))$ as follows

$$\begin{aligned} ds &= -\beta(t)sjdt + \sqrt{\frac{\beta(t)sj}{N}}dB_1 \\ dj &= (\beta(t)sj - \gamma j)dt - \sqrt{\frac{\beta(t)sj}{N}}dB_1 + \sqrt{\frac{\gamma j}{N}}dB_2 \end{aligned} \tag{3.5}$$

where B_1, B_2 are independent Brownian motions. This process is similar to the deterministic version of the SIR model with added white noise accounting for stochasticity in each compartment. Rewriting (3.5) in matrix form gives us

$$dz = A(t,z)dt + L(t,z)d\mathbf{B} \tag{3.6}$$

where $\mathbf{B}$ is a bivariate Brownian motion and

$$A(t,z) = \begin{pmatrix} -\beta(t)sj \\ \beta(t)sj - \gamma j \end{pmatrix}, \quad L(t,z) = \frac{1}{\sqrt{N}} \begin{pmatrix} \sqrt{\beta(t)sj} & 0 \\ -\sqrt{\beta(t)sj} & \sqrt{\gamma j} \end{pmatrix} \tag{3.7}$$

Next, we have the following theorem

Theorem 3.2.1. *Let $[0,T]$ be the time interval of the data. Then we have*

$$\sqrt{N}W_{1,T}(x,z) \xrightarrow{N\to\infty} 0 \tag{3.8}$$

or in other words, $W_{1,T}(x,z) = o(1/\sqrt{N})$

The proof for a more generalized version of theorem 3.2.1, where $x(t)$ is a general compartmental model, can be found in [2]. The main idea is to prove that both $x(t)$ and $z(t)$ converge in Wasserstein distance to the same process.

With this, we can, for sufficiently large N, use lemma 2.4.3 to replace the likelihood function in (3.4) with

$$L_z(\boldsymbol{\theta}) = \prod_{i=1}^{M} p_{\boldsymbol{\theta},t_i,t_{i+1}}(z(t_{i+1})|z(t_i)) \tag{3.9}$$

Likelihood computation for diffusion processes

We will now look into methods to compute the conditional densities in (3.9). If the SDE in (3.6) is explicitly solvable, then the likelihood function can be exactly computed. For example, assuming that the solution can be written as

$$z(t) = z(0) + D(t, \boldsymbol{\theta}) + E(t, \boldsymbol{\theta})\mathbf{B}(t) \tag{3.10}$$

where D, E are functions of appropriate dimensions. Then

$$z(t_{i+1})|z(t_i) = z(0) + D(t_{i+1}, \boldsymbol{\theta}) + E(t_{i+1}, \boldsymbol{\theta})(\mathbf{B}(t_i) + N(\mathbf{0}, \Delta t_i I_2)) \tag{3.11}$$

where $\Delta t_i = t_{i+1} - t_i$ and I_2 the rank 2 identity matrix. With this, we can get a closed form for $p_{\boldsymbol{\theta}}(z(t_{i+1})|z(t_i))$.

However, the SDE in (3.6) is not explicitly solvable in general and therefore requires a different approach. The method I settled on involves the simple Euler-Maruyama approximation $\tilde{z}^{(k)}(t)$ of $z(t)$. For all observed time t_i, let

$$\begin{aligned} \tau_{ir} &= t_i + r\frac{\Delta t_i}{k} = t_i + r\Delta\tau_i \\ \tilde{z}^{(k)}(t_i) &= z(t_i) \\ \tilde{z}^{(k)}(\tau_{i(r+1)}) &= \tilde{z}^{(k)}(\tau_{ir}) + A(\tau_{ir}, \tilde{z}^{(k)}(\tau_{ir}))\Delta\tau_i + L(\tau_{ir}, \tilde{z}^{(k)}(\tau_{ir}))\Delta\mathbf{B}_{ir} \end{aligned} \tag{3.12}$$

where $\Delta\mathbf{B}_{ir} = \mathbf{B}(\tau_{i(r+1)}) - \mathbf{B}(\tau_{ir})$. Next we have the conditions for this scheme to converge:

Lemma 3.2.2. *[12] Under the following conditions:*

(A1) For all $0 < R < \infty, 0 \le t \le R$, the functions $A(t, \cdot)$ and $L(t, \cdot)$ are Lipschitz continuous in the closed ball $B(\boldsymbol{0}, R)$.

(A2) For all $0 < R < \infty$ there exists $0 < C_R < \infty$ such that

$$\|A(t, x)\| + \|L(t, x)\| \le C_R(1 + \|x\|) \quad \forall\ 0 \le t \le R, x \in \mathbb{R}^d \tag{3.13}$$

(A3) $\Sigma(t, x) = L(t, x)L(t, x)^\top$ is positive definite for all $t \ge 0$ and $x \in \mathbb{R}^d$.

We have $\tilde{z}^{(k)}(t) \xrightarrow{L_1} z(t)$ for all $t \in [0, T]$ as $k \to \infty$.

Note that conditions (A1) and (A2) are satisfied since the components of $A(t,\cdot)$ and $L(t,\cdot)$ are polynomials and square roots of polynomials, respectively. The remaining condition is true as long as the epidemic has not ended, i.e. $i(t) > 0$.

Setting $k = 1$ in (3.12), we have the following scheme

$$\tilde{z}^{(1)}(t_{i+1}) = \tilde{z}^{(1)}(t_i) + A(t_i, \tilde{z}^{(1)}(t_i))\Delta t_i + L(t_i, \tilde{z}^{(1)}(t_i))\Delta \mathbf{B}_i \tag{3.14}$$

where $\Delta \mathbf{B}_i = \mathbf{B}(t_{i+1}) - \mathbf{B}(t_i)$.

With this we can approximate the likelihood function of $z(t)$ with that of $\tilde{z}^{(1)}(t)$. And due to the construction of $\tilde{z}^{(1)}(t)$ in (3.14), we have the following closed form likelihood formula

$$p^{(1)}_{\boldsymbol{\theta},t_i,t_{i+1}}(\tilde{z}^{(1)}(t_{i+1})|\tilde{z}^{(1)}(t_i)) = \phi(\Delta\tilde{z}^{(1)}_i | A(t_i, \tilde{z}^{(1)}(t_i))\Delta t_i, \Sigma(t_i, \tilde{z}^{(1)}(t_i))\Delta t_i) \tag{3.15}$$

where $\Delta\tilde{z}^{(1)}_i = \tilde{z}^{(1)}(t_{i+1}) - \tilde{z}^{(1)}(t_i)$ and $\phi(\cdot|\mu,\Sigma)$ is the density of $N(\mu,\Sigma)$.

Now given the data points $x(t_1), \ldots, x(t_M)$, the approximate likelihood is

$$L^{(1)}(\boldsymbol{\theta}) = \prod_{i=1}^{M} p^{(1)}_{\boldsymbol{\theta},t_i,t_{i+1}}(x(t_{i+1})|x(t_i)) \tag{3.16}$$

Another concern here is that in the original SIR model, the states of $x(t)$ are in $[0,1]^2$, which is not the case for $\tilde{z}^{(1)}(t)$. Therefore, in some cases when one or both elements of $x(t_i)$ is 0 or 1, we view it as a censored observation in regard to $\tilde{z}^{(1)}(t)$. These cases are when $s(t_i) = 1$ or $j(t_i) = 1$ or $s(t_i) = 0 \wedge s(t_{i-1}) \neq 0$ or $j(t_i) = 0 \wedge j(t_{i-1}) \neq 0$.

Now define

$$\begin{aligned}
\boldsymbol{\mu}(t_i) &= \begin{pmatrix} \mu_1(t_i) \\ \mu_2(t_i) \end{pmatrix} = x(t_{i-1}) + A(t_{i-1}, x(t_{i-1}))\Delta t_{i-1} \\
\boldsymbol{\Sigma}(t_i) &= \begin{pmatrix} \sigma_{11}(t_i) & \sigma_{12}(t_i) \\ \sigma_{12}(t_i) & \sigma_{22}(t_i) \end{pmatrix} = \Sigma(t_{i-1}, x(t_{i-1}))\Delta t_{i-1}
\end{aligned} \tag{3.17}$$

and

$$\begin{aligned}
\mu^*_1(t_i) &= \mu_1(t_i) + \frac{\sigma_{12}(t_i)}{\sigma_{22}(t_i)}(j(t_{i-1}) - \mu_2(t_i)) \\
\mu^*_2(t_i) &= \mu_2(t_i) + \frac{\sigma_{12}(t_i)}{\sigma_{11}(t_i)}(s(t_{i-1}) - \mu_1(t_i)) \\
(\sigma^*_1(t_i))^2 &= \sigma_{11}(t_i) - \frac{\sigma^2_{12}(t_i)}{\sigma_{22}(t_i)} \\
(\sigma^*_2(t_i))^2 &= \sigma_{22}(t_i) - \frac{\sigma^2_{12}(t_i)}{\sigma_{11}(t_i)}
\end{aligned} \tag{3.18}$$

The terms defined in (3.18) are just the conditional mean and variance of each component given the other. With this we can write out the likelihood formula for all cases of the data

$$p^{(1)}_{\boldsymbol{\theta},t_i,t_{i+1}}(x(t_i)|x(t_{i-1})) =$$
$$\begin{cases} \phi(x(t_i)|\boldsymbol{\mu}(t_i),\boldsymbol{\Sigma}(t_i)) & \text{if } s(t_i), j(t_i) \in (0,1) \\ \Phi(\iota(x(t_i))|\boldsymbol{\mu}(t_i),\boldsymbol{\Sigma}(t_i)) & \text{if } s(t_i), j(t_i) \notin (0,1) \\ \phi(s(t_i)|\mu_1(t_i),\sigma_{11}(t_i))\Phi(\iota(j(t_i))|\mu_2^*(t_i),(\sigma_2^*(t_i))^2) & \text{if } s(t_i) \in (0,1), j(t_i) \notin (0,1) \\ \phi(j(t_i)|\mu_2(t_i),\sigma_{22}(t_i))\Phi(\iota(s(t_i))|\mu_1^*(t_i),(\sigma_1^*(t_i))^2) & \text{if } s(t_i) \notin (0,1), j(t_i) \in (0,1) \end{cases} \tag{3.19}$$

where $\Phi(\iota(x)|\mu,\sigma^2)$ denotes the integral of the normal distribution on the corresponding interval $\iota(x)$. For example, if $x(t_{i-1}) = (0.9, 0.1)$ and $x(t_i) = (0.85, 0)$ then the third formula in (3.18) is used and $\iota(j(t_i)) = (-\infty, 0]$.

Multi-step likelihood approximation

The Euler-Maruyama approximation described in (3.14) only makes one jump from one time stamp to the next and the likelihood derived from this is referred to as the 1-step likelihood. Problems with this scheme arise when the time stamps are too far apart or the infection rate changes too quickly between observation times thereby lowering the approximation quality. A solution is to use the k-step scheme in (3.12) with larger k for better approximation. Note that in the multi-step scheme, we do not know the observations in between the observed times and the likelihood will therefore involve integrating out these values

Theorem 3.2.3. *The likelihood formula for the scheme* (3.12) *is as follows*

$$p^{(k)}_{\boldsymbol{\theta},t_i,t_{i+1}}(z_2|z_1) = \int \prod_{r=1}^{k} p^{(1)}_{\boldsymbol{\theta},\tau_{i(r-1)},\tau_{ir}}(\xi_r|\xi_{r-1}) d\xi_1 \dots d\xi_{k-1} \tag{3.20}$$
$$= E_{z_1}(p^{(1)}_{\boldsymbol{\theta},\tau_{i(k-1)},t_{i+1}}(z_2|\tilde{z}^{(k)}(\tau_{i(k-1)}))) \tag{3.21}$$

where $\xi_0 = z_1, \xi_k = z_2 \in \mathbb{R}^2, p$ *and the expectation is taken conditional on* $\tilde{z}^{(k)}(t_i) = z_1$.

Proof. Since (3.20) is by definition, we only need to prove that the right hand side of

(3.20) equals to (3.21)

$$\begin{aligned}
&\int \prod_{r=1}^{k} p^{(1)}_{\boldsymbol{\theta},\tau_{i(r-1)},\tau_{ir}}(\xi_r|\xi_{r-1})d\xi_1 \dots d\xi_{k-1} \\
&\quad = \int p^{(1)}_{\boldsymbol{\theta},\tau_{i(k-1)},\tau_{ik}}(\xi_k|\xi_{k-1}) \prod_{r=1}^{k-1} p^{(1)}_{\boldsymbol{\theta},\tau_{i(r-1)},\tau_{ir}}(\xi_r|\xi_{r-1})d\xi_1 \dots d\xi_{k-1} \\
&\quad \overset{Fubini}{=} \int p^{(1)}_{\boldsymbol{\theta},\tau_{i(k-1)},\tau_{ik}}(\xi_k|\xi_{k-1}) \left(\int \prod_{r=1}^{k-1} p^{(1)}_{\boldsymbol{\theta},\tau_{i(r-1)},\tau_{ir}}(\xi_r|\xi_{r-1})d\xi_1 \dots d\xi_{k-2} \right) d\xi_{k-1} \\
&\quad = \int p^{(1)}_{\boldsymbol{\theta},\tau_{i(k-1)},\tau_{ik}}(\xi_k|\xi_{k-1}) p^{(k-1)}_{\boldsymbol{\theta},\tau_{i0},\tau_{i(k-1)}}(\xi_{k-1}|\xi_0)d\xi_{k-1} \\
&\quad = \int p^{(1)}_{\boldsymbol{\theta},\tau_{i(k-1)},t_{i+1}}(z_2|\xi_{k-1}) p^{(k-1)}_{\boldsymbol{\theta},t_i,\tau_{i(k-1)}}(\xi_{k-1}|z_1)d\xi_{k-1} \\
&\quad = E_{z_1}(p^{(1)}_{\boldsymbol{\theta},\tau_{i(k-1)},t_{i+1}}(z_2|\tilde{z}^{(k)}(\tau_{i(k-1)})))
\end{aligned}$$

□

Using the law of large numbers and the expression (3.21) in Theorem 3.2.3, we have the following procedure to approximate the multi-step likelihood

- Simulate B sample paths using (3.12) to get $\tilde{z}^{(k)}_1(\tau_{i(k-1)}), \dots, \tilde{z}^{(k)}_B(\tau_{i(k-1)})$.
- By the law of large numbers, we have

$$\frac{1}{B}\sum_{b=1}^{B} p^{(1)}_{\boldsymbol{\theta},\tau_{i(k-1)},t_{i+1}}(z_2|\tilde{z}^{(k)}_b(\tau_{i(k-1)})) \xrightarrow{a.s} E_{z_1}(p^{(1)}_{\boldsymbol{\theta},\tau_{i(k-1)},t_{i+1}}(z_2|\tilde{z}^{(k)}(\tau_{i(k-1)}))) \quad (3.22)$$

The trade-off for using the multi-step likelihood is the increased computational time due to the simulations.

3.2.2 Tau leaping approximation

For this method, we go back to the definition of the stochastic SIR model

$$X(t) = X(0) + \begin{pmatrix} -1 \\ 1 \end{pmatrix} Pois_1\left(\int_0^t \beta(s)\frac{S(s)I(s)}{N}ds\right) + \begin{pmatrix} 0 \\ -1 \end{pmatrix} Pois_2\left(\int_0^t \gamma I(s)ds\right) \quad (3.23)$$

where $Pois_1, Pois_2$ are independent standard Poisson processes. Tau leaping is a method for approximate simulation of (3.23) using a scheme similar to the Euler-Maruyama method

$$\begin{aligned}
\tau_{ir} &= t_i + r\frac{\Delta t_i}{k} = t_i + r\Delta\tau_i \\
\tilde{Z}^{(k)}(t_i) &= Z(t_i) \\
\tilde{Z}^{(k)}(\tau_{i(r+1)}) &= \tilde{Z}^{(k)}(\tau_{ir}) + \begin{pmatrix} -1 \\ 1 \end{pmatrix} Pois_1\left(\Delta\tau_i\beta(\tau_{ir})\frac{S(\tau_{ir})I(\tau_{ir})}{N}\right) \\
&\qquad + \begin{pmatrix} 0 \\ -1 \end{pmatrix} Pois_2\left(\Delta\tau_i\gamma I(\tau_{ir})\right).
\end{aligned} \tag{3.24}$$

With this the likelihood function can be approximated using the transition probabilities of $\tilde{Z}^{(k)}$. Specifically, for $k = 1$ we have

$$\begin{aligned}
L^{(1)}(\boldsymbol{\theta}) &= \prod_{i=1}^{M} P^{(1)}_{\boldsymbol{\theta},t_i,t_{i+1}}(\tilde{Z}^{(1)}(t_{i+1})|\tilde{Z}^{(1)}(t_i)) \\
&= \prod_{i=1}^{M} f\left(\Delta W_i \middle| \Delta t_i\beta(t_i)\frac{S(t_i)I(t_i)}{N}\right) f(\Delta Y_i|\Delta t_i\gamma I(t_i))
\end{aligned} \tag{3.25}$$

where $\Delta W_i = S(t_i) - S(t_{i+1}), \Delta Y_i = S(t_i) - S(t_{i+1}) + I(t_i) - I(t_{i+1})$ and $f(\cdot|\lambda)$ is the probability mass function of a Poisson distribution with rate λ.

To compute the multi-step likelihood approximation, we use the same procedure devised for diffusion processes with the path simulation method and one step likelihood formula changed to that of tau leaping.

3.3 Parameter estimation

With a method to approximately compute the likelihood function, the maximum likelihood estimate (MLE) for the model parameters can be found using built-in `R` functions such as `optim`. Before that, we need to fine tune the hyperparameters, specifically the number of knots and their locations. There are two main approaches to achieve this goal. The first, called the penalized spline method [20], is to put a knot on every observed time point and add a penalty term to the likelihood. Then the model parameters are found by minimizing the following objective function

$$F(\boldsymbol{\theta}, \lambda) = -L(\boldsymbol{\theta}) + \lambda J(\beta) \tag{3.26}$$

where $J(\beta)$ is the function that penalizes certain properties of the infection rate function β and λ is the smoothing parameter controlling the degree of penalty. This means the knots are predetermined but the number of parameters is large leading to longer time spent estimating them. In addition, minimizing the function in (3.26) usually requires estimating $\boldsymbol{\theta}$ for multiple λ's which increases computation time even more.

The second approach, the regression spline method [17], uses fewer and unequally distanced knots. While this method take less time to estimate $\boldsymbol{\theta}$, the choice of knot number and placement is crucial to its performance. Through the experiments in the next chapter, we will see that the method in [21] does a good job finding the right knot locations. Therefore, the rest of this thesis will focus on the regression spline method.

3.3.1 Knot selection

This section describes the knot placement method in [21]. In the paper, we are given the values of the curve $\beta(t)$ at times $u_0, \ldots, u_m$ and the goal is to find the knots $\kappa_1, \ldots, \kappa_k$ for the degree d B-spline basis used to estimate $\beta(t)$. This is achieved by the following steps:

1. Calculate the $(d+1)^{th}$ derivative $\beta^{(d)}(t)$ of $\beta(t)$ using the formula

$$\beta^{(j+1)}(u_i^{(j+1)}) = \frac{\beta^{(j)}(u_{i+1}^{(j)}) - \beta^{(j)}(u_i^{(j)})}{u_{i+1}^{(j)} - u_i^{(j)}}, \quad u_i^{(j+1)} = \frac{1}{2}(u_{i+1}^{(j)} + u_{i+1}^{(j)}) \tag{3.27}$$

   where $\beta^{(0)}(t) = \beta(t)$. Note that (3.27) implies that each derivative level has its own time stamps which are the midpoints of the previous level's time stamps.

2. Calculate the feature curve $F(u)$. First, define the feature function $f(u)$ as the piecewise linear function that satisfies

$$f_i = f(\bar{u}_i) = \begin{cases} 0 & \text{if } i = 0, m-d \\ \|\beta^{(d+1)}(u_i^{(d+1)})\|^{1/(d+1)} & \text{otherwise} \end{cases} \tag{3.28}$$

   where $\bar{u}_0 = u_1, \bar{u}_{m-d} = u_m$ and $\bar{u}_i = u_i^{(d+1)}$ for $0 < i < m-d$. Then $F(u)$ is

defined as the integral of $f(u)$, i.e.

$$F(u) = \int_{-\infty}^{u} f(v)dv \tag{3.29}$$

3. Obtain knot locations from the feature curve by setting $\kappa_j = F^{-1}(j\Delta F)$ where $\Delta F = \max F(u)/(k-1)$. In other words, divide the feature curve into segments with equal amount of increase and set the corresponding time stamps as knots. Computation of $F^{-1}(u)$ can be simplified by pretending $F(u)$ is a piecewise linear function and values at $\bar{u}_i$ calculated using trapezoid rule for f_i.

3.3.2 Moving average rate estimate

A problem we run into is that the knot placement strategy described above uses the true infection rate values at observed times, which is not present in the data. To resolve this, "true" values of $\beta(t)$ are created by estimating the moving average rates between observations. Consider r consecutive observations $x(t_i), \ldots, x(t_{i+r})$, we now build a mini model by assuming that the infection rate is a constant in this period. Then the estimate for β using these data points will be the guess for the true value of $\beta((t_i + t_{i+r})/2)$. Next, the procedure is repeated over all windows of r consecutive observations to get the curve values for knot selection.

For example, if the whole data is $x(0), x(1), \ldots, x(30)$ and window size is $r = 3$, then $x(0), x(1), x(2)$ are used to estimate the value of $\beta(1)$; $x(1), x(2), x(3)$ for $\beta(2)$ and so on. The idea for this procedure is pretty similar to a moving average of a time series but the average series is for the hidden infection rate.

3.3.3 Model selection and Estimation summary

The final step is to choose the number of knots to use. In [21], a linear regression model was used to find the relationship between $\log(\Delta F)$ and the log of the mean square error and the number of knots k is chosen to get a desired amount of error. But since the moving average rates are only rough guesses and the estimated parameter vector is the MLE, it makes more sense to use a likelihood based model selection criteria such as Bayesian information criterion (BIC) or Akaike information criterion (AIC). In addition, a forward selection scheme is employed for this step to save more

time. The idea is to increase the number of knots one at a time, and stop when none of the last Q models (Q will be referred to as the stopping threshold) have improved on the best value of the criterion. To summarize, the parameter estimation procedure follows these steps

Algorithm 1 Forward selecting regression spline

Input: Data $\mathbf{X}, N$, window size w, degree d, criterion C, threshold Q.

Output: Spline coefficients $\hat{\boldsymbol{\theta}}$ for infection rate function $\hat{\beta}$ and recovery rate $\hat{\gamma}$.

$\bar{\beta} \leftarrow$ estimate_moving_average_rates(w)
$count \leftarrow 0$
$numknot \leftarrow 0$
while $count < Q$ **do**
  $knots \leftarrow$ find_knot_placement($\bar{\beta}, numknot$)
  $\hat{\boldsymbol{\theta}} \leftarrow \text{MLE}(X, N, knots, d)$
  $crit \leftarrow C(X, \hat{\boldsymbol{\theta}})$
  if $numknot = 0$ **then**
    $min\_crit \leftarrow crit$
  else
    if $crit < min\_crit$ **then**
      $min\_crit \leftarrow crit$
      $count \leftarrow 0$
    else
      $count \leftarrow count + 1$
    end if
  end if
  $numknot \leftarrow numknot + 1$
end while
  return $\hat{\boldsymbol{\theta}}$

3.4 Confidence interval

A parametric bootstrap scheme is used to find the point wise confidence interval for $\beta(t)$. However, unlike the case where the infection rate is constant, we essentially have no information about $\beta(t)$ when the sample path terminates early. This can lead to uninformative or even bad intervals for the infection rate function. An example is when both $I(t_1)$ and the estimated rate $\hat{\beta}(t)$ is small in the beginning, which can happen when the spline degree is 2 or higher, leading to many bootstrap samples terminating too early. A solution is to discard simulated paths that terminated early and use the ones that survived until the final observed time t_M as bootstrap samples.

3.4.1 Interval smoothing

One of the most crucial steps in the model is knot placement. However, this step relies entirely on the moving average rates, which is a very crude estimate of the true infection rate, and therefore can misjudge the most effective placements. A way to alleviate this is to smooth out the pointwise confidence interval using adjacent time stamps. In particular, we consider three different ways of smoothing: weighted smoothing, sample smoothing and min-max smoothing.

The first way is to use the interval values themselves. Let L_{t_i}, U_{t_i} be the lower and upper bounds of the confidence interval for $\beta(t_i)$. Then the smoothed confidence interval $[\bar{L}_{t_i}, \bar{U}_{t_i}]$ is calculated as the weighted sum of adjacent bounds as follows

$$\bar{L}_{t_i} = \frac{\sum_j w(t_i, t_j) L_{t_j}}{\sum_j w(t_i, t_j)}, \quad \bar{U}_{t_i} = \frac{\sum_j w(t_i, t_j) U_{t_j}}{\sum_j w(t_i, t_j)} \tag{3.30}$$

where the weighting function $w(\cdot, \cdot)$ is the normal kernel

$$w(x, y) = \phi(x - y)$$

The second way is to combine the β values from the bootstrap samples at adjacent time points and use them as the the samples for the middle point. Specifically, we use $\hat{\beta}^*_1(t_{i-1}), \ldots, \hat{\beta}^*_B(t_{i-1}), \hat{\beta}^*_1(t_i), \ldots, \hat{\beta}^*_B(t_i), \hat{\beta}^*_1(t_{i+1}), \ldots, \hat{\beta}^*_B(t_{i+1})$ as samples to construct the confidence interval of $\beta(t_i)$ instead of just $\hat{\beta}^*_1(t_i), \ldots, \hat{\beta}^*_B(t_i)$. Here $\hat{\beta}^*_b(t)$ denotes the estimated infection rate at t for the b^{th} bootstrap sample and B is the number of bootstrap samples. The intuition behind this step is to improve the coverage rate at

places where there are significant changes in the infection rate.

The third method is to simply widen the bounds by setting the new upper bounds as the largest of all the surrounding bounds and the new lower bounds as the smallest of all the surrounding bounds. Specifically, the new interval $[\bar{L}_{t_i}, \bar{U}_{t_i}]$ for $\beta(t_i)$ is

$$\bar{L}_{t_i} = \min\{L_{t_{i-1}}, L_{t_i}, L_{t_{i+1}}\}, \quad \bar{U}_{t_i} = \max\{U_{t_{i-1}}, U_{t_i}, U_{t_{i+1}}\} \tag{3.31}$$

3.5 Numerical considerations for multi-step approximation

Since the single step approximations have explicit and simple likelihood functions, their results are mostly stable. The multi-step schemes, on the other hand, require simulations and take longer to compute. Therefore, this section will mainly discuss ways to make the multi-step methods faster and more stable.

The first problem comes from the simulation of the states in between the observed data. Unless the number of simulated paths is very large, which can make the computation infeasible, it can lead to inconsistent evaluation of the likelihood function. This can greatly impact the parameter estimation process. To see this effect in action, we look at the result of the 20-step diffusion approximation method for a simulated data set with different random number generator seeds. We can see from Figure 3.1 that the change in seed is enough to change the number of knots the final model selected. One solution to this is to set the generating seed beforehand. Doing this ensures the consistency of the likelihood by making it smoother and easier to optimize.

In addition, for the multi-step diffusion approximation, we can generate standard normal variables for the simulations beforehand as in [14]. This saves us from having to generate new random variables every time the likelihood function is evaluated. Unfortunately, the same cannot be done for the multi-step tau leaping method as Poisson variables cannot be rescaled into other Poisson variables. Because of this, the multi-step tau leaping method will have to spend time generating random variables for its sample path and therefore, will not be feasible for repeated experiment.

The next consideration is about the initial value for $\beta(t)$ that is fed into `optim`. Generally, we want the number of likelihood evaluations to be as few as possible to save more time. Therefore, we use the B-spline fit of the moving average rate as the initial value for `optim` in hope that it is close to the MLE.

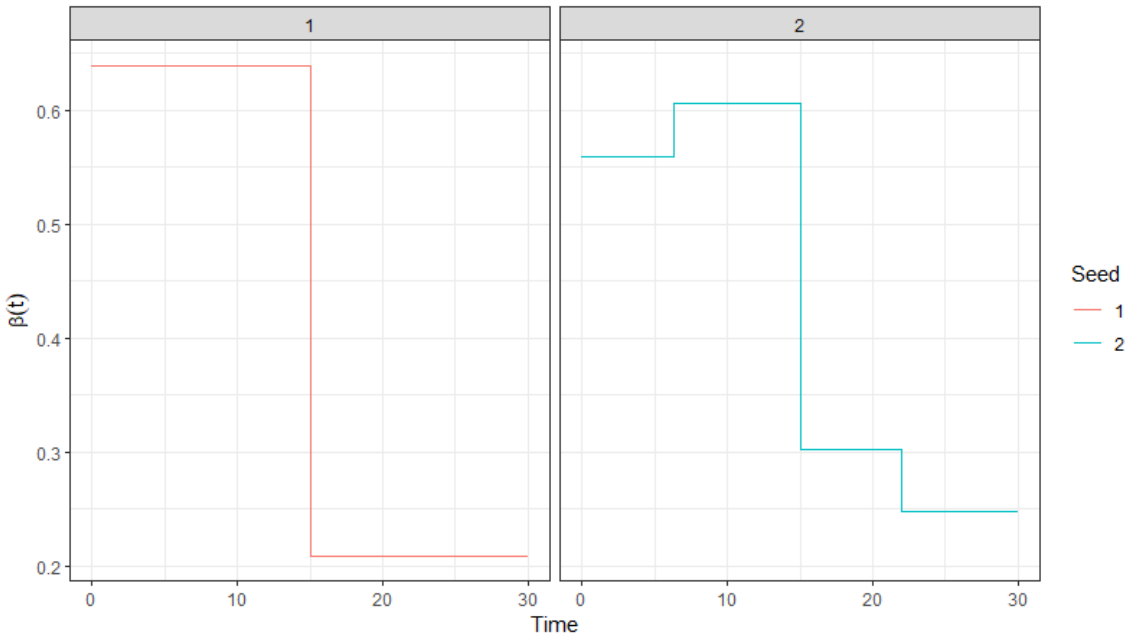
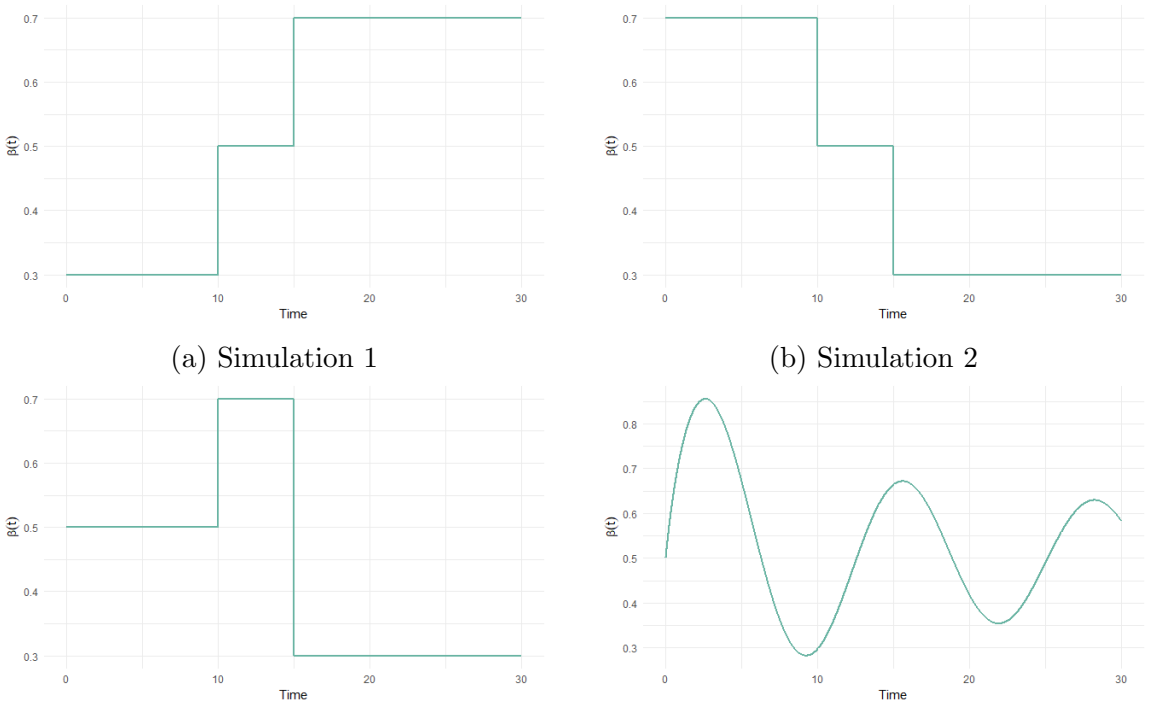


Figure 3.1: Resulting estimates using two different seeds for path simulation.

Chapter 4

Simulation Study

In this chapter, various performance aspects of the proposed model are investigated using simulated data. The data sets are generated using the `R` package `ssar`, which employs the Gillespie algorithm for exact simulation of the stochastic SIR model. Specifically, we mainly look at 4 typical epidemic patterns where the infection rate is increasing, decreasing, going up then down or periodic. Each data set consists of 30 data points, the recovery rate is set at 0.25 for all 4 simulations and the infection rates are plotted in Figure 4.1. In addition, the populations are all set to $N = 1000$ with initial proportion of susceptible and infected at 99% and 1%, respectively.



(a) Simulation 1

(b) Simulation 2

(c) Simulation 3

(d) Simulation 4

Figure 4.1: Infection rate functions of the 4 simulations.

4.1 Approximation quality

The quality of the estimation $\hat{\beta}(t)$ is measured by its mean integrated squared error (MSE) to the true infection rate $\beta(t)$, i.e.

$$MSE(\hat{\beta}, \beta) = \int (\hat{\beta}(t) - \beta(t))^2 dt \tag{4.1}$$

Each simulation is repeated 100 times for different settings of likelihood approximation, criterion for number of knots and moving average window size.

4.1.1 Effects of Likelihood approximation

In this subsection, we look at the performance of different types of likelihood approximation. To this end, we consider 3 methods: 1-step diffusion, 1-step tau leaping and a multi-step diffusion. For the multi-step method, we use a 20-step diffusion scheme with 100 sample paths to estimate the likelihood. The window size is set to 2 and the model selection criterion to BIC.

Figures 4.2 and 4.3 are the box plot comparison of MSE for each likelihood approximation method across the 4 simulations using degree 0 and 3 B-spline bases, respectively. Based on these two figures, tau leaping seems to perform slightly better than diffusion approximation for single step likelihood. This can be attributed to the former only having one approximation step (Euler-Maruyama step) while the latter has two approximation steps (diffusion approximation and Euler-Maruyama step). In addition, the 20-step diffusion method outperforms the single step schemes for all settings. The improvement is most significant for simulation 4 as the multi-step method allows the model to better capture the changes in the compartments between 2 consecutive time points.

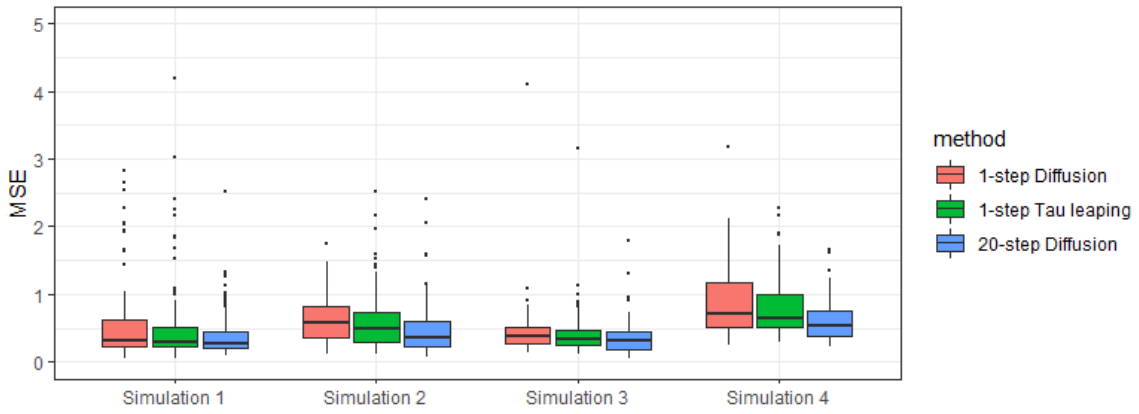


Figure 4.2: Estimation performance of different simulations and methods using degree 0 B-spline basis.

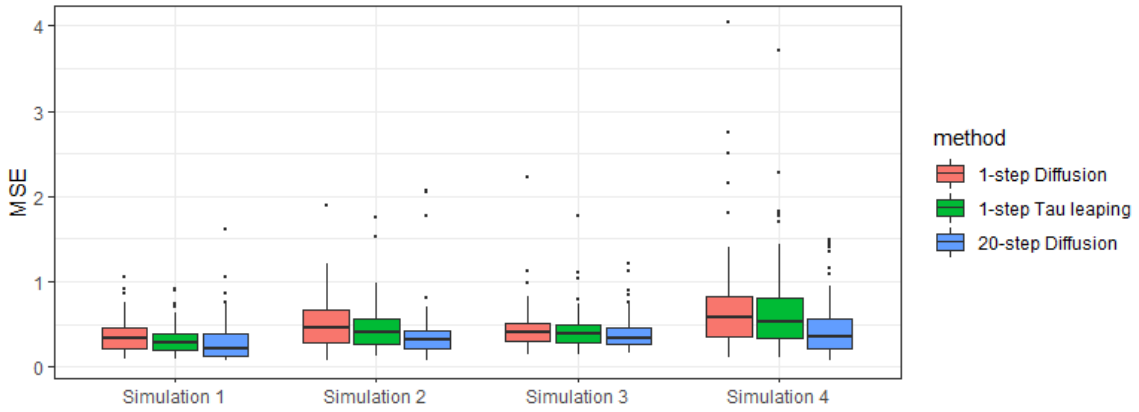


Figure 4.3: Estimation performance of different simulations and methods using degree 3 B-spline basis.

Next, to see the estimation performance at different stages of an epidemic we plotted the estimations in time in Figures 4.4 and 4.5. The solid lines are the true rates, the dashed lines and two bands are the average, 5% and 95% quantiles of the estimations, respectively. The places with the worst performance are often at the beginning and the end of the epidemic where the approximations are most likely to be inaccurate. We also observed that when a degree 3 B-spline is used, the estimated rates at the two time boundaries are more varied. This maybe due to the nature of spline bases that make estimations near boundaries erratic [15].

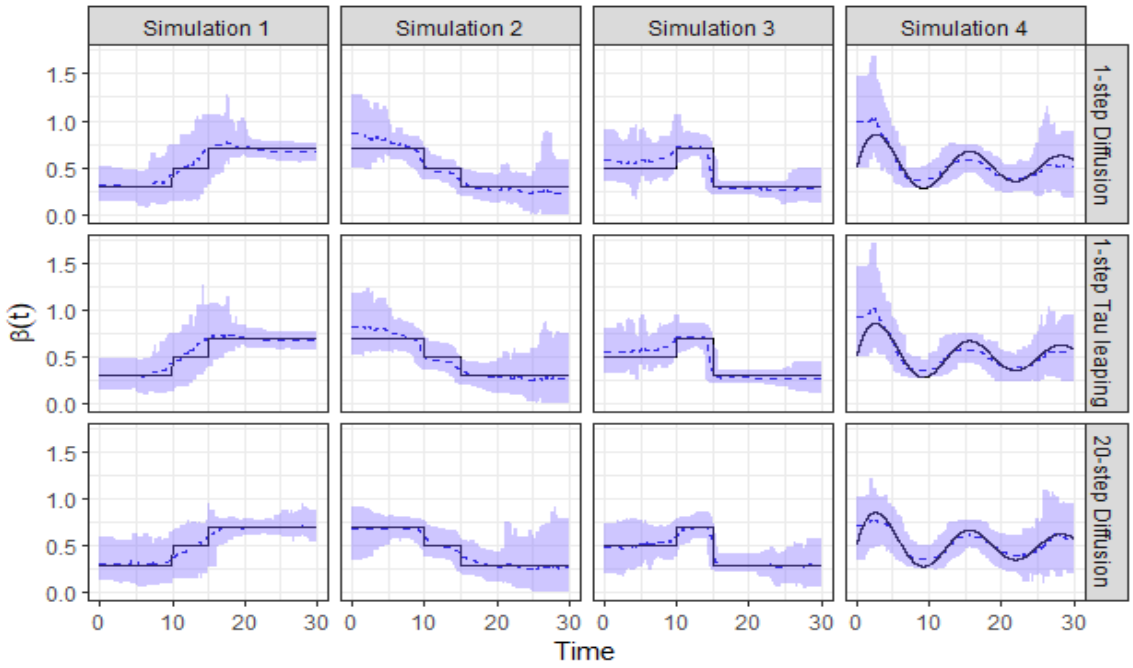


Figure 4.4: Estimation performance in time for different simulations and methods using degree 0 B-spline basis.

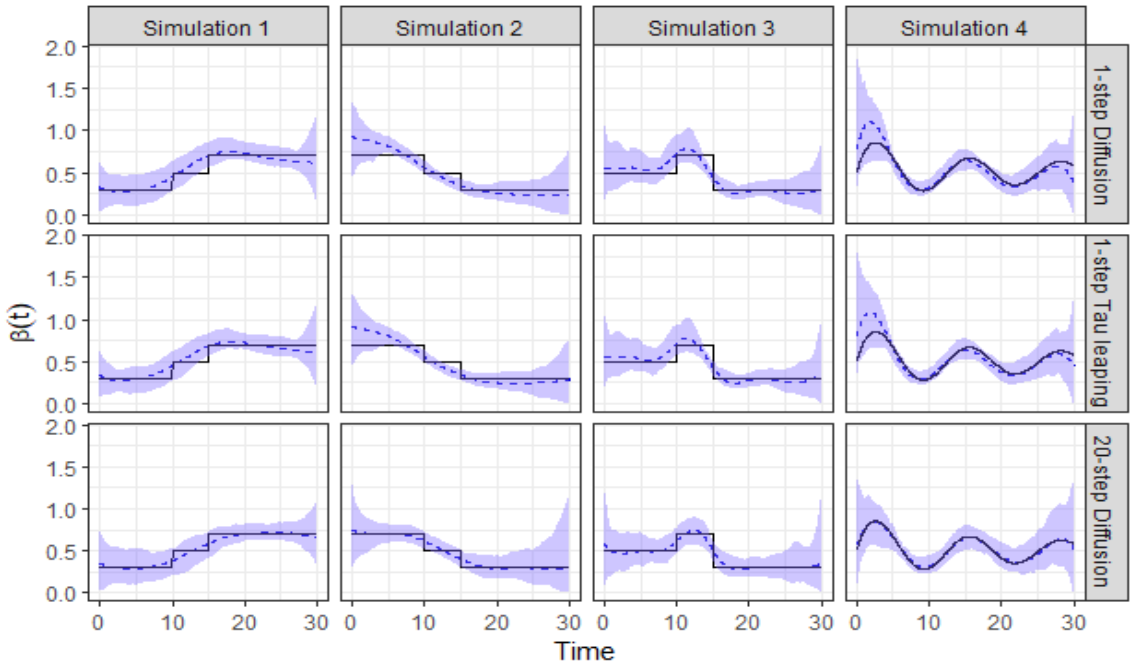


Figure 4.5: Estimation performance in time for different simulations and methods using degree 3 B-spline basis.

The biggest trade-off of the multi-step scheme is how time consuming it is. The execution time ratios between the 20-step and 1-step diffusion methods across different simulations are visualized in Figure 4.6. According to this, the multi-step scheme is seen to run hundreds of times slower than the single step scheme even at moderate step size 20 and low sample path number of 100 (multi-step scheme runs for 20 min per data set on average while single-step scheme runs for less than 5 seconds). As discussed in the previous chapter, this is the main reason for not running the multi-step tau leaping method as it is not scalable when generating sample paths.

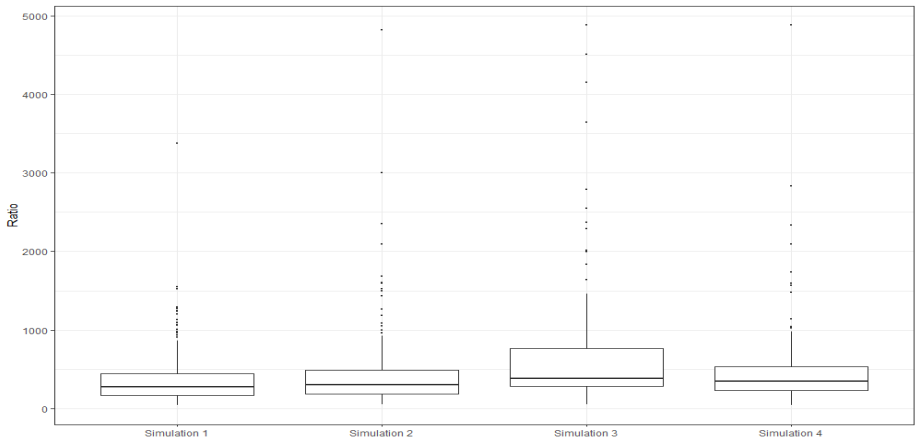


Figure 4.6: Execution time ratio between 20-step and 1-step diffusion methods.

4.1.2 Effects of Model selection criteria

To see how the model selection criterion affects the final estimator, we ran the simulations using AIC and BIC as model selection criteria. BIC is well known for its consistency, i.e. if the generating model is present in the candidate models then the probability of BIC selecting it approaches 1, while AIC is known for being asymptotically optimal, i.e. if the generating model is not present among the candidate then the selected model is the closest one to the true model on average (chapter 6 of [3]). If the two criteria exhibit their relative properties, we would expect to see BIC perform better in the first three simulations when degree 0 B-splines are used and AIC for when degree 3 B-splines are used. However, Figure 4.7 implies that BIC performs as well as or better than AIC in most settings. Therefore, in the following sections, BIC will be the model selection method of choice.

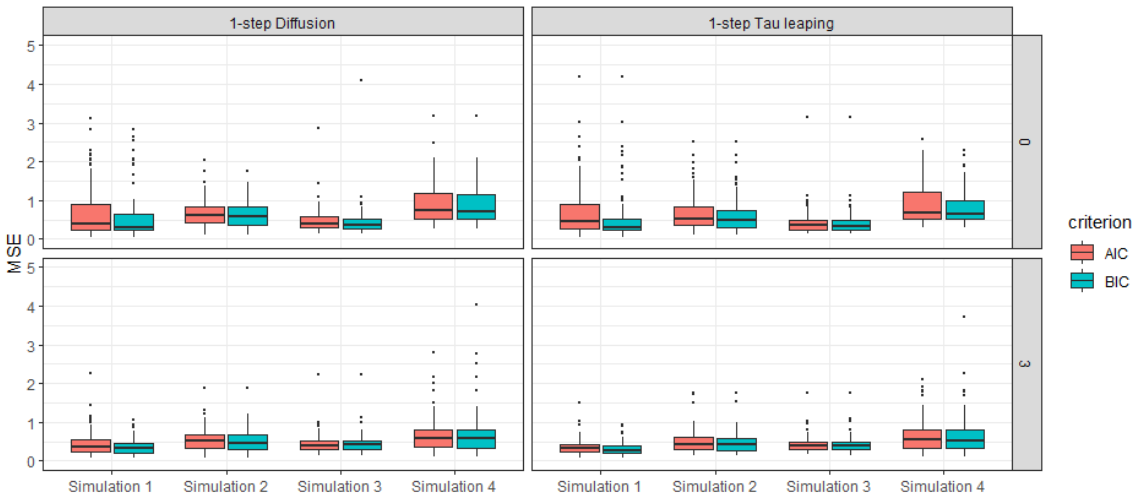


Figure 4.7: Estimation performance for different criteria. The rows represent the degree of B-spline basis used.

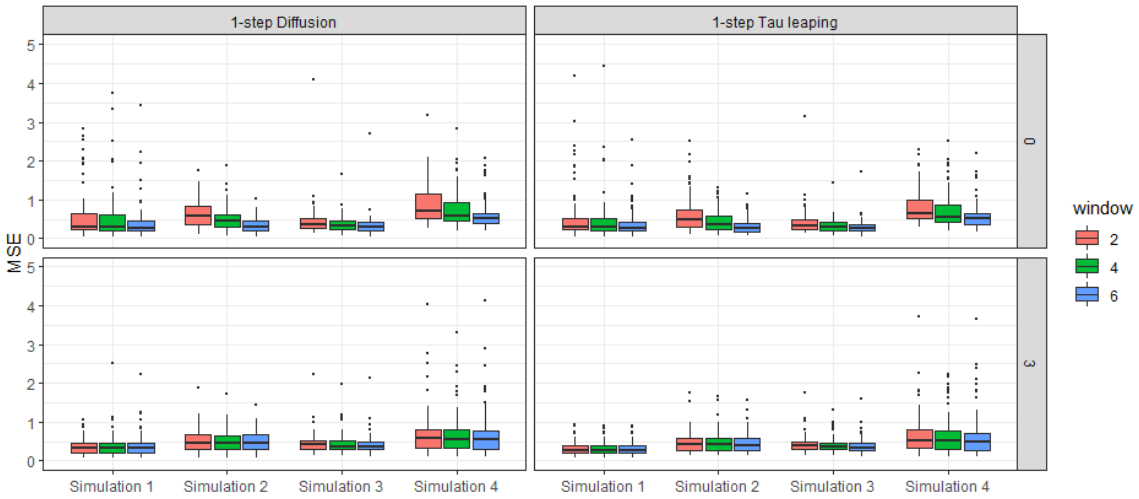


Figure 4.8: Estimation performance for different window sizes. The rows represent the degree of B-spline basis used.

4.1.3 Effects of Moving average window size

As stated before, the moving average window size is similar to the moving average of a time series. Therefore, its main use is to smooth out the initial rate estimates for knot placement. To see how window sizes affect our method's performance, we ran the simulations using window sizes 2, 4, 6 and plotted the results in Figure 4.8. The benefit of smoothing is most evident when using degree 0 B-splines. This is

because the procedure removes the random fluctuations in the rate estimates, helping the knot placement algorithm locate the change points more accurately. For degree 3 B-spline, the forth derivative is used for knot placement, which is in a way a form of smoothing, so increasing the window size does little to improve and, in some cases, even hurts overall performance. Moving forward, window size 2 will be used in order to observe the effects of other steps.

4.2 Confidence interval coverage

In this section, we analyze the performance of the bootstrap confidence intervals introduced in Chapter 2 and the effects that techniques such as bias correction and interval smoothing have on their coverage rates. For each simulation and repetition, 1000 bootstrap samples are created and fitted using the proposed framework with 1-step Tau leaping likelihood approximation. Then the estimated β's are used to construct the 95% bootstrap confidence intervals at each observed time.

The coverage rates at each time point for all simulations and confidence interval types are plotted in Figure 4.9. From this, we can see that the coverage rates suffer when a degree 0 basis is used for the smooth infection rate in Simulation 4 as well as when a degree 3 basis is used for the simulations with step wise constant rates. This is understandable as it is more difficult to express the truth when the basis is misspecified. We also see dips in coverage rates near the change points $t = 10$ and $t = 15$ in the first three simulations. Overall, the normal and pivot intervals have better coverage rates than the percentile intervals. The trade-off is that the former two can have negative lower bounds, which make them useless as the rates are always positive, while the latter cannot. In addition, Figure 4.10 shows that the length of normal interval does not seem to differ much from that of the other two intervals except near the change points or the end of the epidemic, where it tends to be wider. Note that the pivot and percentile have equal lengths by definition.

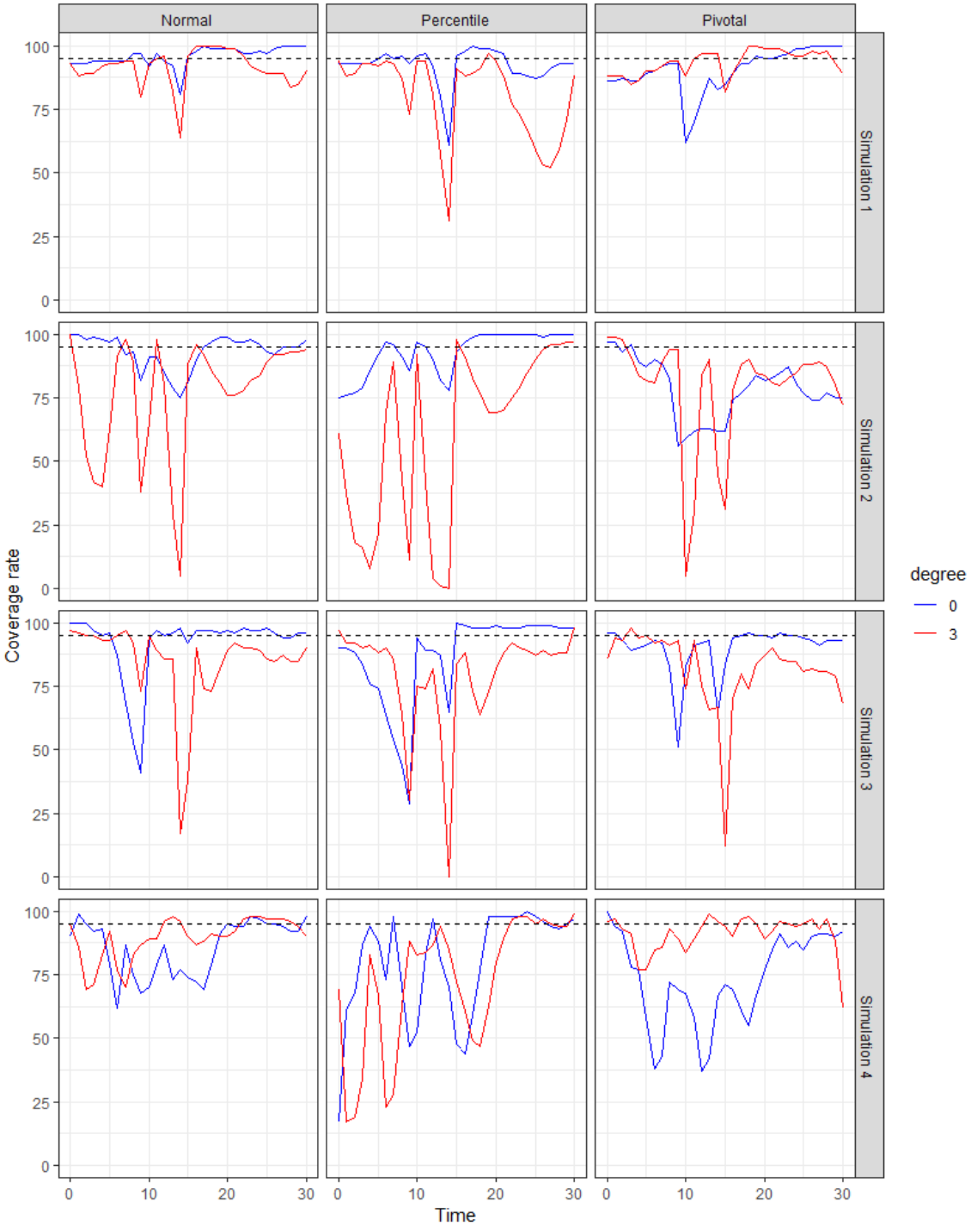


Figure 4.9: Coverage rates for 95% bootstrap confidence intervals at each time stamp.

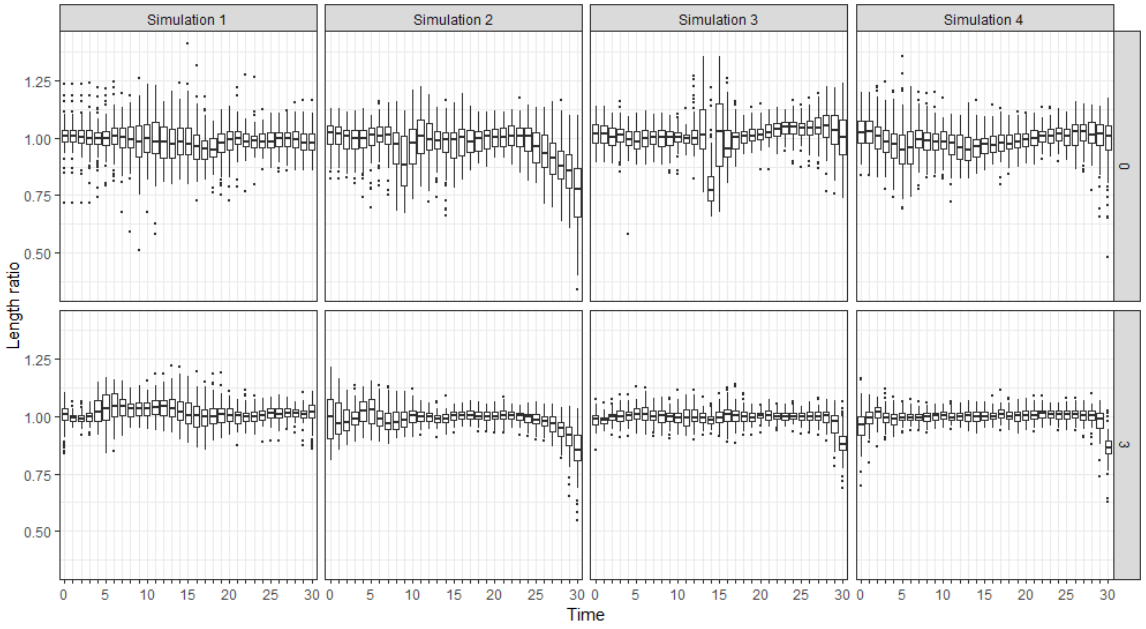


Figure 4.10: Length ratio between the percentile interval and the normal interval by time. The rows show the degree of basis used.

4.2.1 Effects of interval smoothing

In this subsection, we look at how each interval smoothing method affects the coverage rate of the confidence intervals. Figures 4.11 and 4.12 shows the coverage rates of the original intervals compared to the three interval smoothing methods proposed in Chapter 3. Based on these, the min-max smoothing method has the best coverage rates out of the three in most cases. This is to be expected as min-max smoothing widens the intervals, guaranteeing improvement in coverage rates. The performance of the other two is interesting as weighted smoothing works better when a degree 0 basis is used whereas sample smoothing works better for a degree 3 basis. The reason may lie in the nature of each basis. A degree 0 basis gives step-wise constant estimates so weighted smoothing can improve the smoothness between intervals at different time points. A degree 3 basis, on the other hand, has smoothness but lacks the ability to rapidly change its values like a degree 0 basis, which makes sample smoothing more useful since it helps expand the bootstrap sample range in places where the infection rate changes quickly.

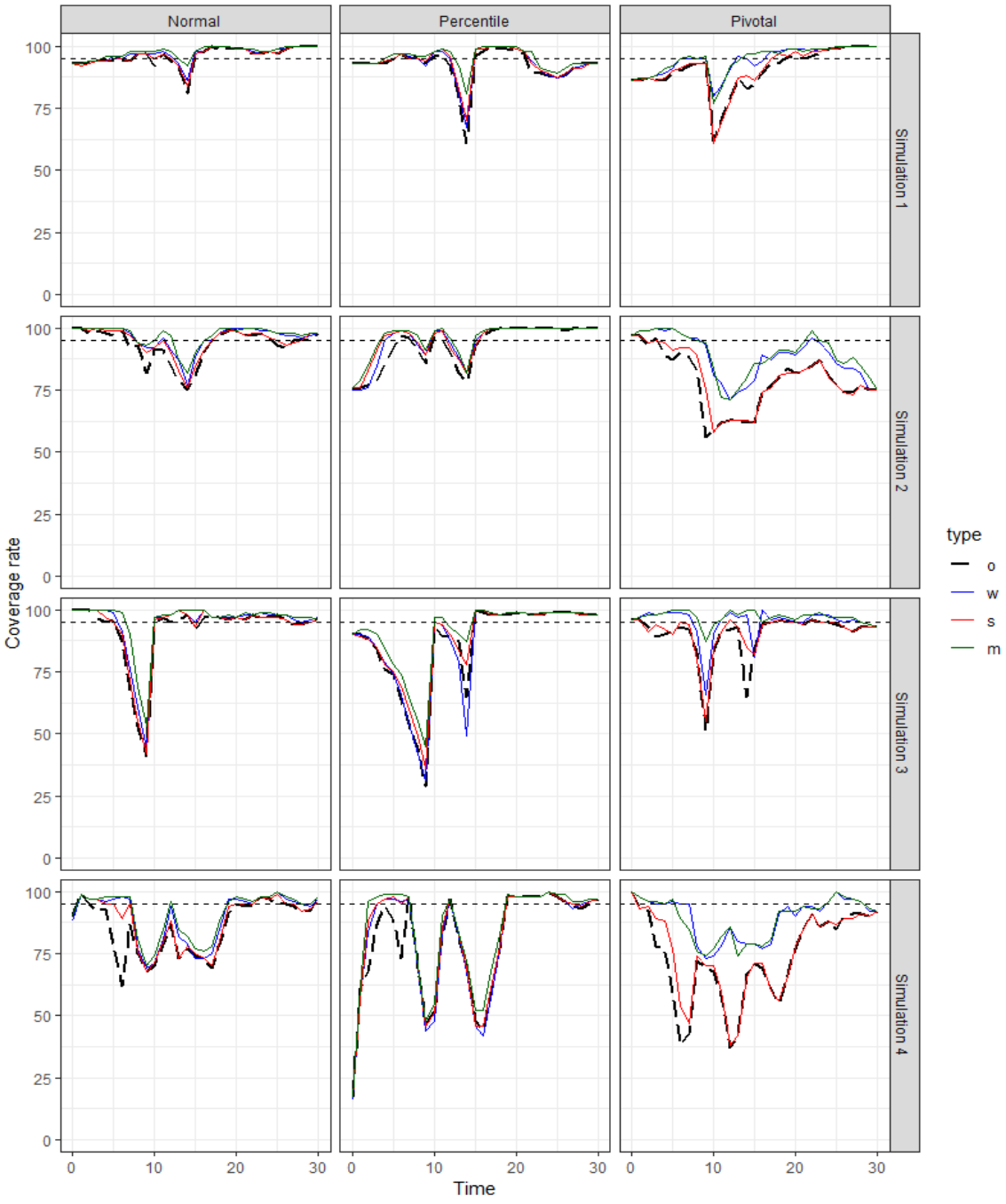


Figure 4.11: Coverage rates for different interval smoothing methods with degree a 0 basis. The labels o, w, s and m stands for original, weighted, sample and min-max, respectively.

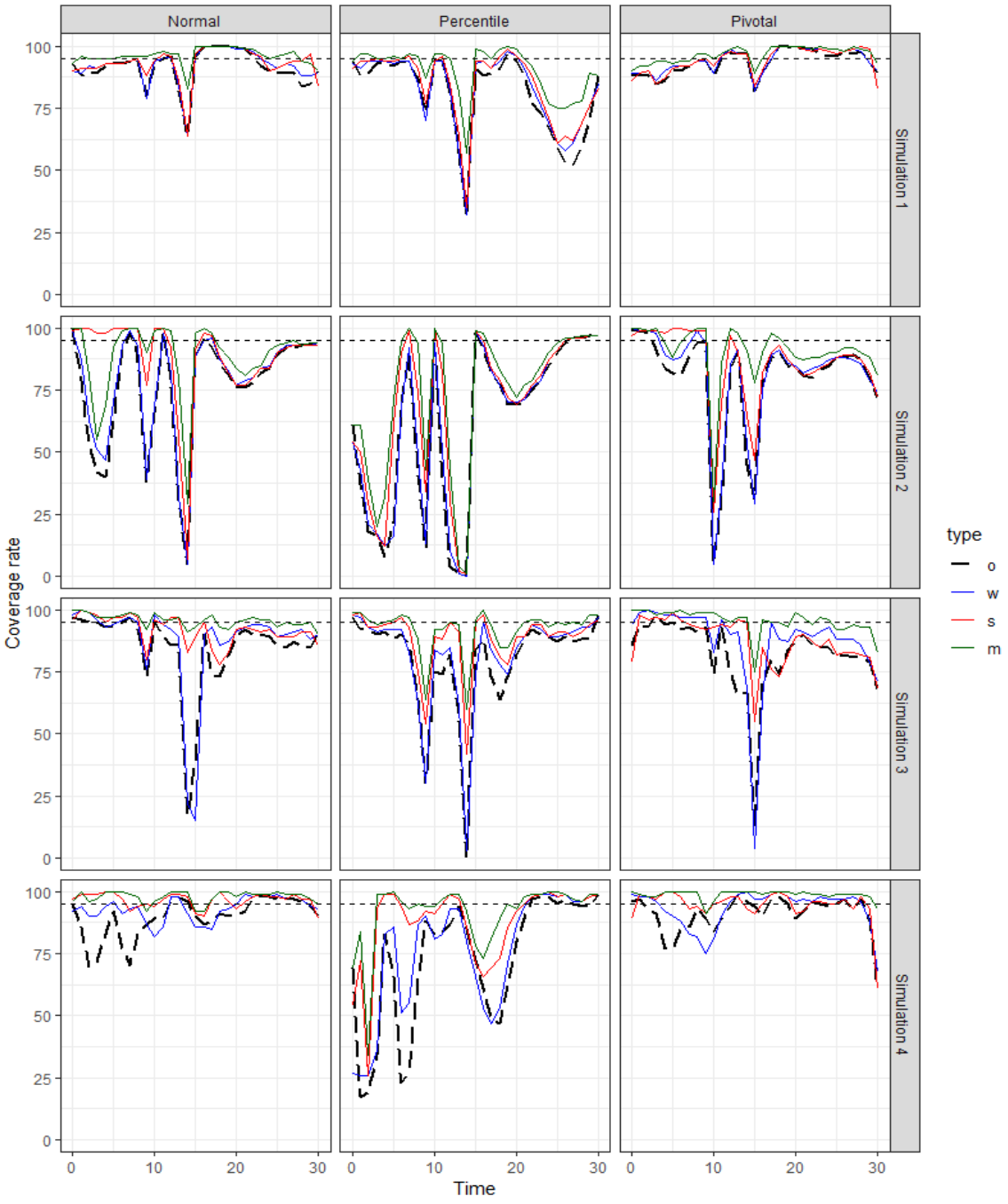


Figure 4.12: Coverage rates for different interval smoothing methods with a degree 3 basis. The labels o, w, s and m stands for original, weighted, sample and min-max, respectively.

4.2.2 Effects of bias correction

In this subsection, we look at how bias correction affects the normal and percentile intervals. Figures 4.13 and 4.14 show the coverage rates of the original intervals compared to the bias corrected versions. For the normal interval, bias correction does not appear to help much and even worsens the coverage rates in some cases. For the percentile interval, however, there is improvement in coverage rates especially with a degree 3 basis. Therefore, bias correction should only be used for the percentile interval moving forward.

Another aspect to look at is how bias correction helps when used in conjunction with the interval smoothing methods. Figures 4.15 and 4.16 illustrate the performance of the interval smoothing methods with and without bias correction along with the intervals where only bias correction is applied. These show that when both processes are applied the coverage rates tend to be better than when only one is applied

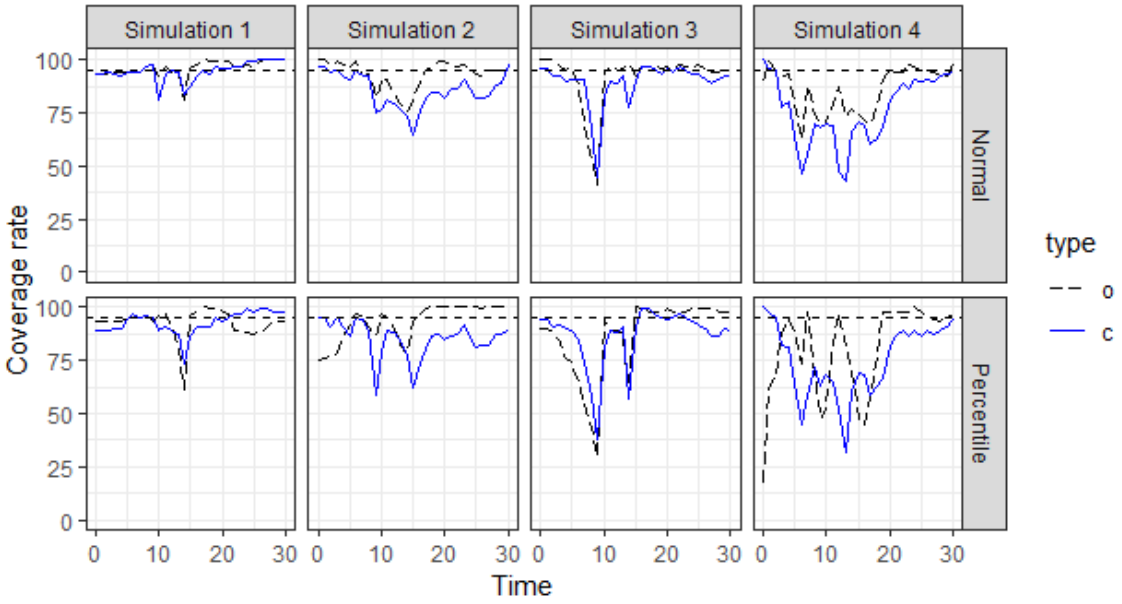


Figure 4.13: Coverage rates for the original intervals (o) compared to the bias corrected intervals (c) with degree 0 basis.

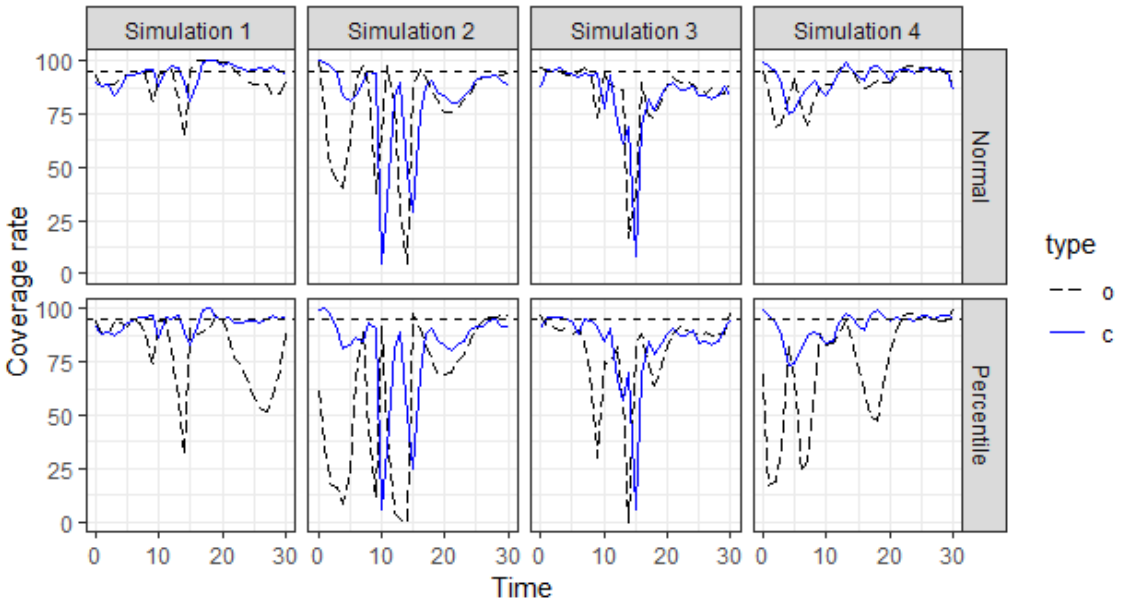


Figure 4.14: Coverage rates for the original intervals (o) compared to the bias corrected intervals (c) with degree 3 basis.

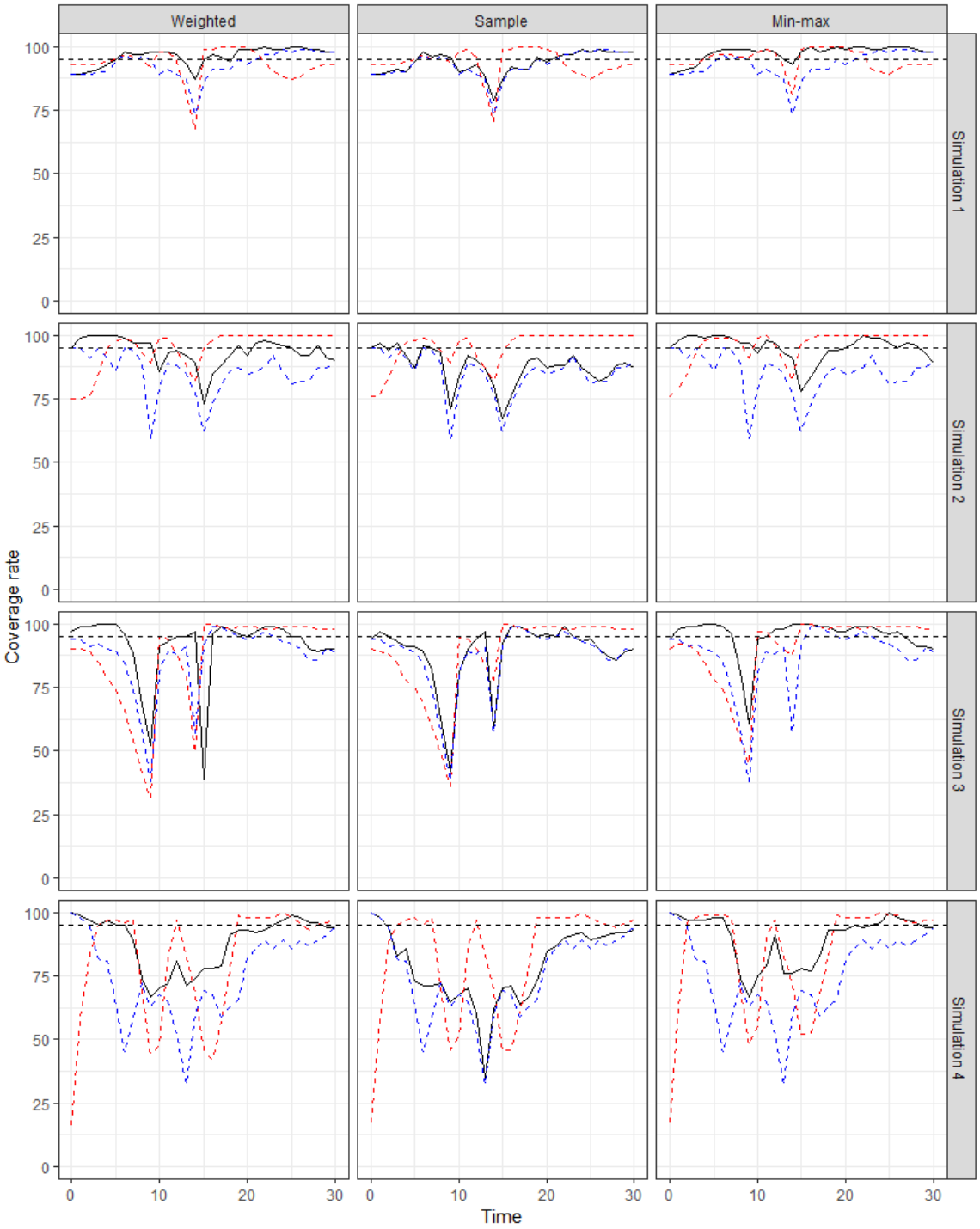


Figure 4.15: Coverage rates for the percentile intervals when both processes are applied (solid line) compared to when only interval smoothing (red dashed line) or bias correction (blue dashed line) is applied with degree 0 basis.

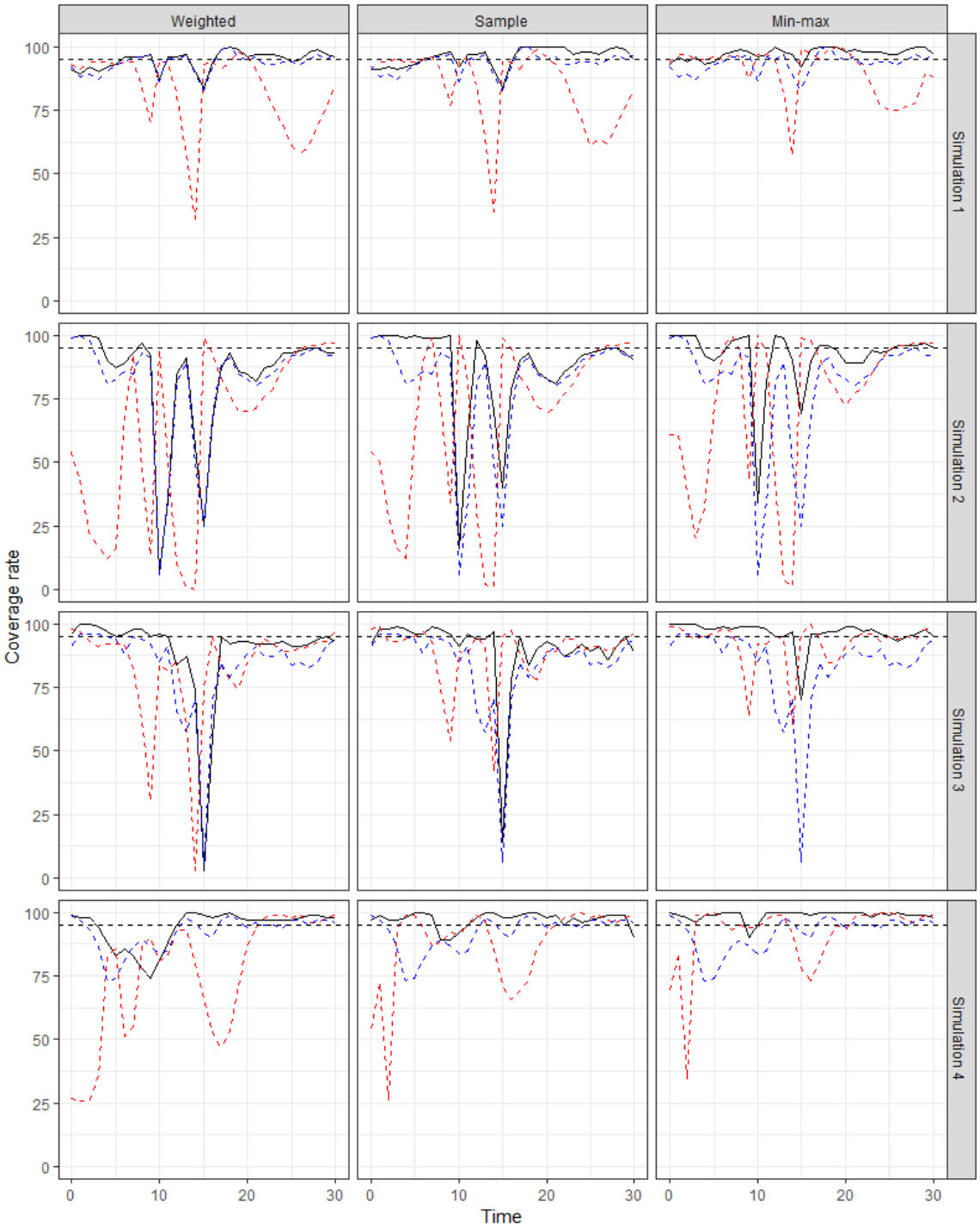


Figure 4.16: Coverage rates for the percentile intervals when both processes are applied (solid line) compared to when only interval smoothing (red dashed line) or bias correction (blue dashed line) is applied with degree 3 basis.

Chapter 5

Application to the COVID-19 data

5.1 Data

In this chapter, we will be estimating the basic reproduction number $R_0(t) = \beta(t)/\gamma$ of the COVID-19 data from Ontario between January 2020 and January 2022. The goal is to see how our proposed framework performs for a period in which multiple waves have occurred.

The data was obtained from [4] in early 2022 when the number of active cases was still recorded. The population is set to $N = 14,223,942$, which is the population of Ontario in 2021 according to [8]. For our model, the number of susceptible S is obtained by subtracting the cumulative cases from the population and the number of infected I is the number of active cases in the data.

5.2 Results

We use the 1-step diffusion and tau leaping method for likelihood approximation, BIC for model selection, window sizes 2 (daily),4 (3 days) and 8 (weekly), and both degree 0 and 3 B-spline bases. The estimates are plotted in Figures 5.1 and 5.2. For a degree 0 basis, the results from the daily and weekly window are more simple with fewer change points. For a degree 3 basis, estimates agree across all window sizes and likelihood approximation methods with only slight differences. With that in mind, we shall use the 3 days window and Tau leaping likelihood to get the confidence intervals for both bases since the estimates for this setting are the most consistent. In addition, the BIC for 3 days window with degree 0 basis is significantly lower than the other two.

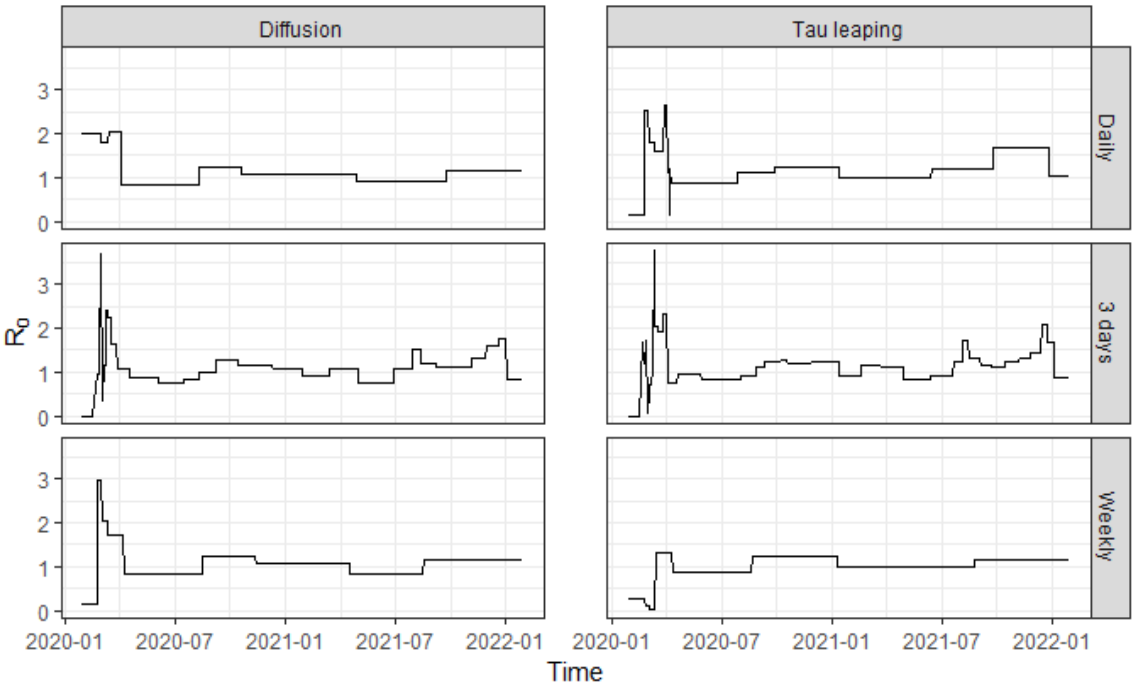


Figure 5.1: $R_0(t)$ estimates for COVID data using degree 0 basis

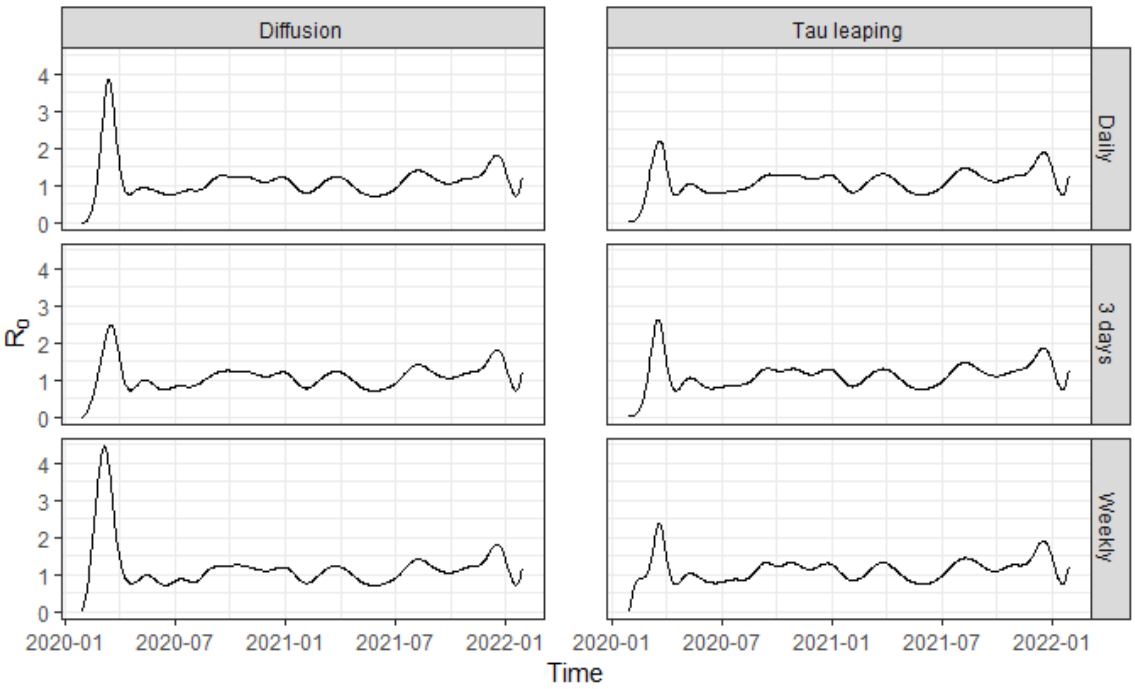


Figure 5.2: $R_0(t)$ estimates for COVID data using degree 3 basis

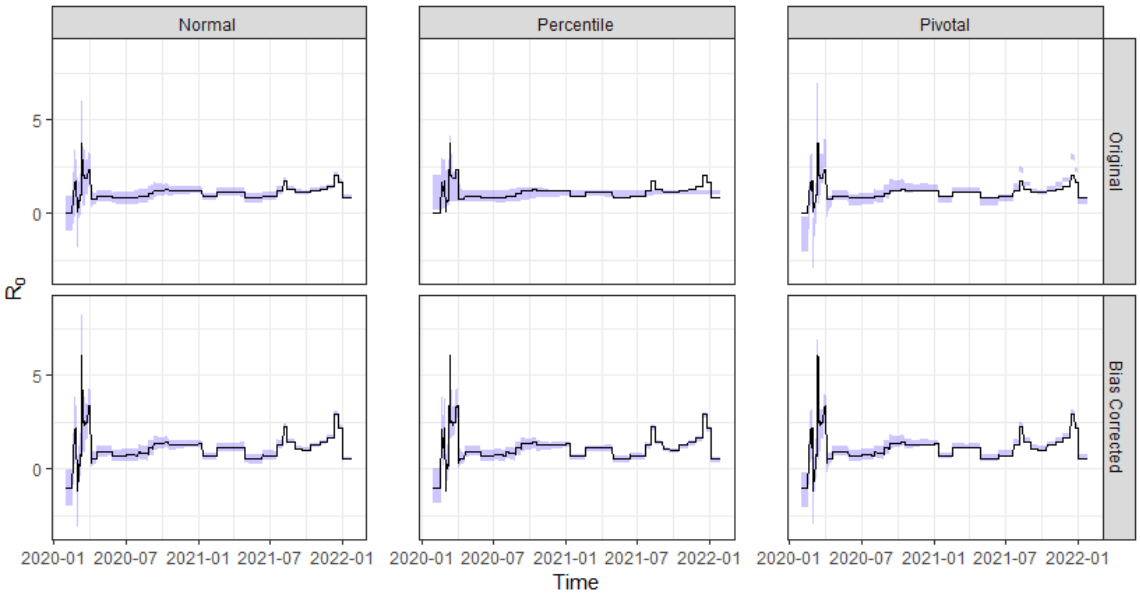


Figure 5.3: Confidence intervals for degree 0 basis.

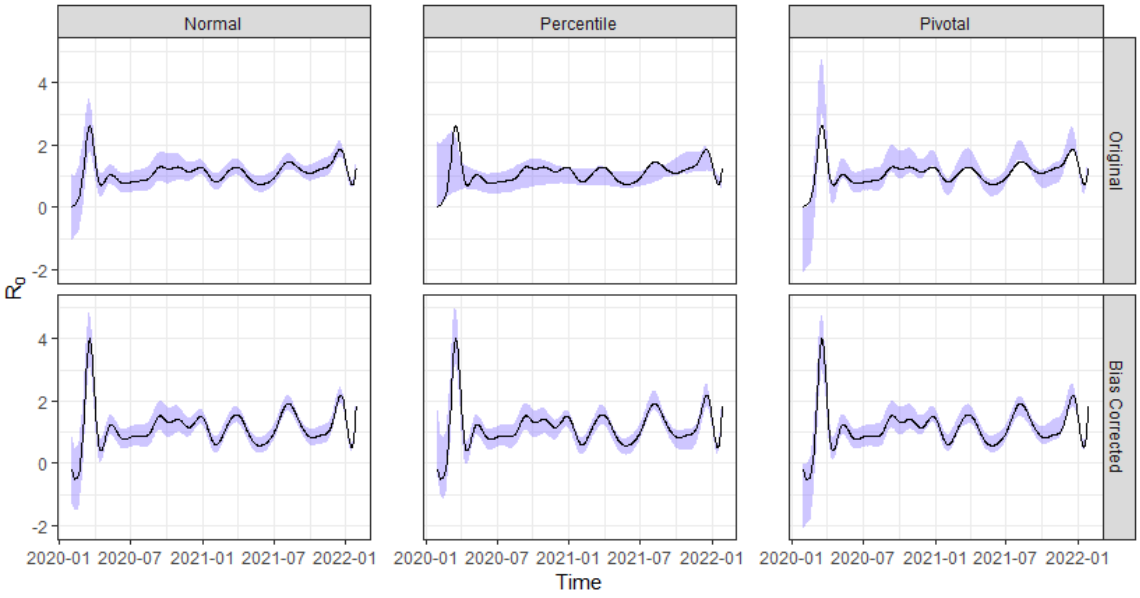
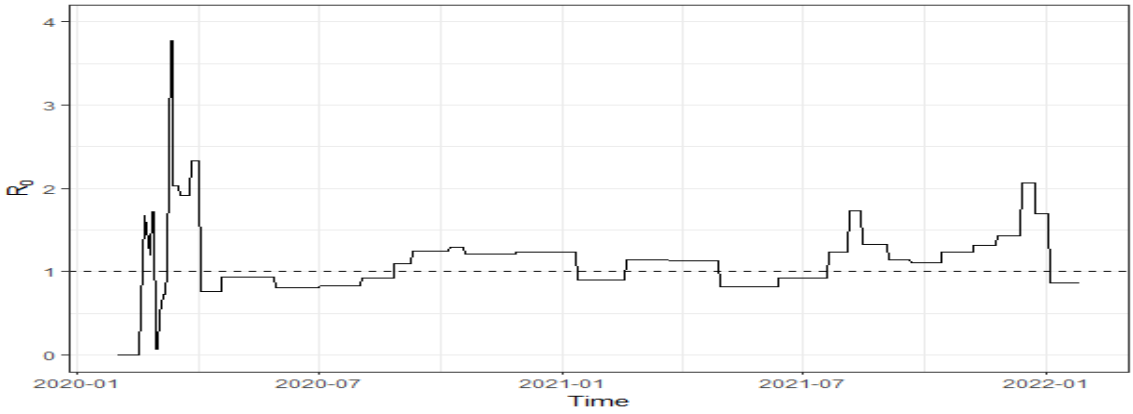
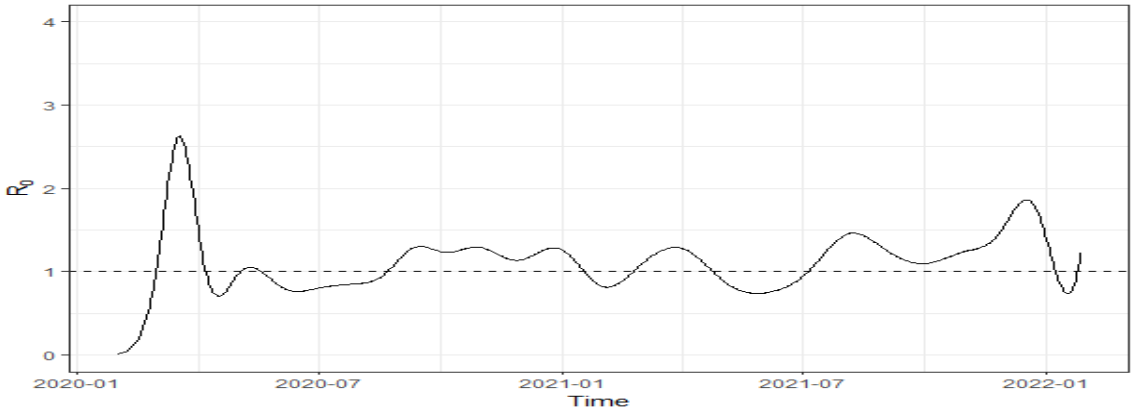


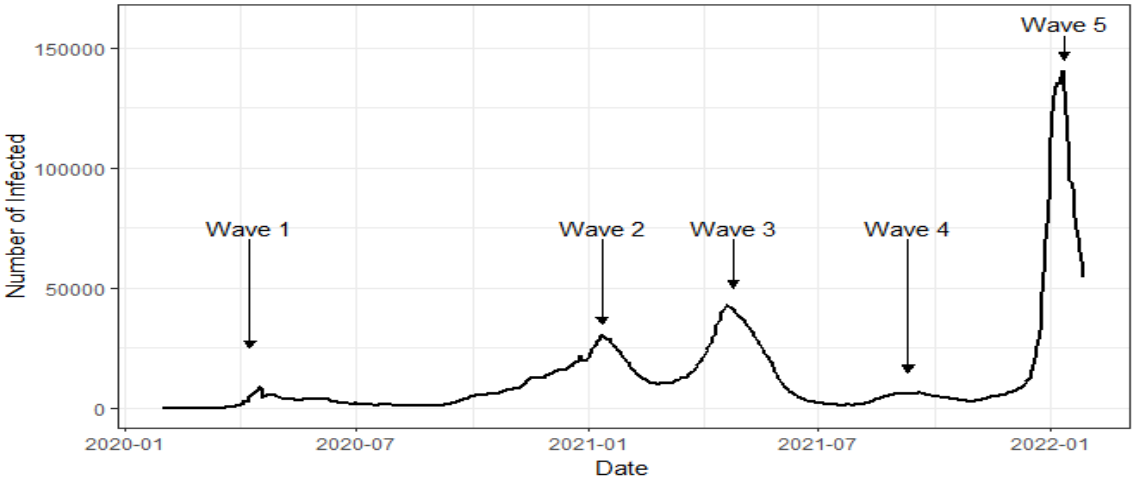
Figure 5.4: Confidence intervals for degree 3 basis.



(a) Degree 0 splines



(b) Degree 3 splines



(c) Active cases data

Figure 5.5: Estimated $R_0(t)$ using 3 days window compared to outbreaks.

For the confidence intervals, we use the parametric bootstrap scheme discussed in Chapter 3 with the samples generated by the Tau leaping method since simulation using the Gillespie algorithm is too time consuming for such a large population. The results are shown in Figures 5.3 and 5.4. Note that for the pivotal interval, bias correction only applies to the estimate not the interval. Looking at the original percentile intervals, we can see that bias correction is necessary, especially for degree 0 basis. The only concern for is that the bias corrected estimate for the infection rate has a portion that lies below 0, which is clearly not true. However, the period where this happens is at the beginning of the epidemic where the number of cases is very small, which is understandable. We also tried to apply interval smoothing but all three methods yielded intervals too similar to the original. Finally, Figure 5.5 shows the reproduction estimates chosen by our method compared to outbreak dates. It seems the peaks in reproduction number chosen by this model closely match the major waves.

Chapter 6

Conclusion

In this thesis, a framework is developed for nonparametric inference of the infection rate function for the SIR model. The two main ideas of this framework is approximating the SIR likelihood function with a different process and using a B-spline basis, which is determined by applying a knot placement method on the moving average rate estimates, to estimate the infection rate. We investigate two ways of approximating the likelihood function for the model using diffusion approximation and Tau leaping. Each of these methods can be made more accurate by using a multi-step scheme which involves simulating sample paths between observations. However, the multi-step methods are considerably more time consuming since the number of sample paths has to be relatively large for consistent results.

For inference, a parametric bootstrap scheme is used to build the percentile, normal and pivotal confidence intervals along with techniques to improve coverage rates such as interval smoothing and bias correction. Through simulation study, we found that the performance of these intervals depend greatly on whether or not the spline basis contain the true model. Finally, we applied our methods to the COVID-19 data in Ontario over a two year period. The resulting models mostly agree with the major waves suggesting that it can be used for disease data with multiple outbreaks.

There are currently many future directions for this framework. On the method side, ways of improving the computation time of the multi-step likelihood and knot placement techniques that work for a wider function space can be explored. Model-wise, a different compartment setting such as SEIR (Susceptible-Exposed-Infected-Removed) can be considered as well as making the recovery rate time dependent.

Bibliography

[1] Michael GB Blum and Viet Chi Tran. Hiv with contact tracing: a case study in approximate bayesian computation. *Biostatistics*, 11(4):644–660, 2010.

[2] Tom Britton, Etienne Pardoux, Franck Ball, Catherine Laredo, David Sirl, and Viet Chí Tran. *Stochastic epidemic models with inference*. Springer, 2019.

[3] Kenneth P Burnham, David R Anderson, Kenneth P Burnham, and David R Anderson. *Practical use of the information-theoretic approach*. Springer, 1998.

[4] Public Health Agency of Canada. Covid-19 epidemiology update: Key updates, Jan 2023. URL: `https://health-infobase.canada.ca/covid-19/`.

[5] Yi-Cheng Chen, Ping-En Lu, Cheng-Shang Chang, and Tzu-Hsuan Liu. A time-dependent sir model for covid-19 with undetectable infected persons. *Ieee transactions on network science and engineering*, 7(4):3279–3294, 2020.

[6] Christiane Dargatz. A diffusion approximation for an epidemic model. 2006.

[7] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[8] Statistics Canada Government of Canada. Population and dwelling counts: Canada, provinces and territories, Feb 2022. URL: `https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=9810000101`.

[9] Lam Si Tung Ho, Forrest W Crawford, and Marc A Suchard. Direct likelihood-based inference for discretely observed stochastic compartmental models of infectious disease. *The Annals of Applied Statistics*, 12(3):1993–2021, 2018.

[10] Lam Si Tung Ho, Jason Xu, Forrest W Crawford, Vladimir N Minin, and Marc A Suchard. Birth/birth-death processes and their computable transition probabilities with biological applications. *Journal of mathematical biology*, 76:911–944, 2018.

[11] William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

[12] Peter E Kloeden and Eckhard Platen. Higher-order implicit strong numerical schemes for stochastic differential equations. *Journal of statistical physics*, 66:283–314, 1992.

[13] Dave Osthus, Kyle S Hickmann, Petruţa C Caragea, Dave Higdon, and Sara Y Del Valle. Forecasting seasonal influenza with a state-space sir model. *The annals of applied statistics*, 11(1):202, 2017.

[14] Asger Roer Pedersen. A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scandinavian Journal of Statistics*, 22(1):55–71, 1995. URL: `http://www.jstor.org/stable/4616340`.

[15] Aris Perperoglou, Willi Sauerbrei, Michal Abrahamowicz, and Matthias Schmid. A review of spline function procedures in r. *BMC medical research methodology*, 19(1):1–16, 2019.

[16] Weston C Roda, Marie B Varughese, Donglin Han, and Michael Y Li. Why is it difficult to accurately predict the covid-19 epidemic? *Infectious disease modelling*, 5:271–281, 2020.

[17] David Ruppert, Matt P Wand, and Raymond J Carroll. *Semiparametric regression*. Number 12. Cambridge university press, 2003.

[18] Alexandra Smirnova, Linda deCamp, and Gerardo Chowell. Forecasting epidemics through nonparametric estimation of time-dependent transmission rates using the seir model. *Bulletin of mathematical biology*, 81:4343–4365, 2019.

[19] Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

[20] Larry Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006.

[21] Raine Yeh, Youssef S.G. Nashed, Tom Peterka, and Xavier Tricoche. Fast automatic knot placement method for accurate b-spline curve fitting. *Computer-Aided Design*, 128:102905, 2020.