MATHEMATICAL MODELING OF THE COVID-19 EPIDEMICS

by

Donglin Han

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Applied Mathematics

Department of Mathematical and Statistical Sciences
University of Alberta

Abstract

Since the COVID-19 outbreak in Wuhan City in December 2019, numerous model predictions on the COVID-19 epidemics in Wuhan have been reported. These model predictions have shown a wide range of variations. In our first study, we demonstrate that nonidentifiability in model calibrations using the confirmed case data is the main reason for such wide variations. Our modeling study indicates that more independent datasets, better inference methods, and fitting algorithms can significantly reduce the nonidentifiable impact. Further study is carried out for modeling the first COVID-19 wave in Alberta. Confirmed case data and testing data, based on the official reports from Alberta Health, are fitted to a mathematical model estimating the total number of the COVID-19 infections. A sensitivity analysis using PRCC is conducted to show that decreasing the initial transmission rate and increasing the probability of health-seeking for an individual are the most effective ways to help control the epidemic.

Preface

Part of the research completed in this thesis represents work as part of the collaboration between Dr. Michael's research group with Alberta Health during the COVID-19 pandemic, while I was studying toward a Master of Science degree at the University of Alberta. The research collaboration included the following coauthors: Dr. Michael Y. Li, Mr. Weston Roda, Mr. Xuyuan Wang, and Mrs. Tanjima Akhter of the Department of Mathematical and Statistical Sciences, University of Alberta; Dr. Marie Betsy Varughese at the Alberta Health, Analytics and Performance Reporting Branch.

Acknowledgements

I would like to thank my family and friends for their support and understanding when undertaking my research. Their encouragement and love were what sustained me so far. I would like to express my sincere gratitude to my supervisor, Dr. Michael Li, who introduced me to the world of mathematical modeling and showed me how to think critically and work rigorously. I am very thankful for all the guidance and advice he has provided me with over the years. I am super fortunate to work with my talented collaborators, including my colleague Weston Roda, Marie Betsy Varughese, Xuyuan Wang, and all the members of the Dr. Michael Li Lab.

I would like to thank the members of the thesis committee for their time and advice in the completion of this thesis.

Many thanks to Dr. Marie Betsy Varughese and Alberta Health for their collaboration, and for sharing the public health surveillance data for this research.

Finally, I would like to thank the Department of Mathematical and Statistical Sciences at the University of Alberta for their financial support.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Mathematical modeling of COVID-19 epidemic and its challenges

In December 2019, a novel acute respiratory syndrome coronavirus 2 (SARS-CoV-2), later named COVID-19, caused an outbreak in the city of Wuhan, China, has further spread worldwide, resulting in the COVID-19 global pandemic. By late October 2022, there have been over 624 million confirmed cases and over 6.5 million deaths reported globally [2], making it the most severe pandemic in recent history. In order to help public health agencies make efficient decisions, it is essential for modelers to estimate the severity of the epidemic specifically on the number of reported cases, number of hidden infections, the occurrence of peak time of cases, total duration for certain wave, and factors that can mitigate the COVID-19 transmission.

There have been several model projections for the COVID-19 epidemic in Wuhan from modeling groups around the world at the early stage of the pandemic. The modeling results have shown a wide range of variations [3] that peak time is estimated with a big difference: from mid-February to late March 2020, and the total number of infections ranges from 50,000 to millions. Such wide variations in model estimations and predictions represented a significant challenge for mathematical models of epidemics when confronted by real-world epidemics.

One of the reasons for these varied projections by mathematical models is that there was too little information available to modelers at the beginning of the outbreak. In fact, the only reliable data that could be used for model calibration was the reported case data. Another important issue lies in the correct interpretation of public health data. In particular, confirmed cases data represents the number of infected people with symptoms and were confirmed positive for COVID-19 by PCR tests. The daily confirmed cases data is only a fraction of all infected people on a particular day, and the fraction is often unknown due to the large number of people who were asymptomatic on a particular day when cases were recorded. In infectious disease models, the infection compartment I includes all people who are infected, both symptomatic and asymptomatic, get tested or not. The daily case data should not be matched to the $I(t)$ in the modeling fitting. We will see in Chapter 2 that this leads to the nonidentifiability in model calibration, where infinitely many choices for parameter values can give the best fit between case data and model output, but different choices of best-fit parameter values give significantly different projections of the total number of infected. A metaphor of an iceberg best represents the difference between case data and total infections

[4]. Public health surveillance and testing are only able to observe the tip of the iceberg, which is the confirmed case data, while the rest of the iceberg underneath the water represents the infected people in the communities unknown to the public, the so-called "hidden infection". One of the mean objectives of mathematical modeling of epidemics is to predict and estimate the number of hidden infections using the directly observable case data. The nonidentifiability issue in model calibration is one of the serious challenges for accurate model predictions and estimations.

Case-infection ratio measures the proportion of total cases among total infections. Different viral infections that spread through air droplets and close contacts can have very different case-infection ratios. For the 2003 SARS epidemic, the ratio was in the range of 1/5 to 1/2 [5, 6, 7, 8], mainly because individuals infected by SARS-Cov-I virus became infectious after symptoms appear, making identification of infected individuals much more efficient, compared to SARS-Cov-II, for which infected individuals can be infectious before symptoms appear. In contrast, for seasonal influenza in 2019 to 2020, the ratio can be as small as 1/100, based on the estimates from US CDC [9]. Given the uncertainty of COVID-19, a wide range of case-infection ratios and transmission rates should be considered during a model fitting in order to allow the model to find its most likely occurring scenario. The nonidentifiability in modeling these epidemics is represented as a linkage between the case-infection ratio parameter and the transmission coefficient parameter.

A suitable model framework is also crucial for reliable model projections and estimations: complex models incorporate more biological mechanisms with a larger number of model parameters compared to simpler models. When the data for model calibration is limited, more complex models tend to introduce more uncertainty in parameter estimation by model fitting to data, and less reliable model predictions. Model selection using information criteria including the Akaike information criteria is necessary to determine the most suitable model framework for the given dataset.

1.2 Importance of measuring the proportion of infection

Canada's first case of COVID-19 was reported in Toronto, Ontario on January 25, 2020, and by the end of October 2022, the total confirmed COVID-19 cases reached 4.36 million and deaths over 46,700. True numbers of infected and death are far greater due to the significant proportion of the infections having been asymptomatic [10]. Accurate estimation of the proportion of the population infected by COVID-19 is important for informing the true scale of the pandemic and allows the estimation of the infection fatality ratio (IFR) provide a more accurate measure of the mortality burden and the severity of the pandemic than the case fatality ratio (CFR). The infection-induced immunity, together with the vaccine-induced immunity, shapes the overall population immunity to COVID-19, and accurate estimation of the infected proportion better informs the level of immunity in the population. Estimation of the infected proportion also informs the case-infection ratio: the ratio of cumulative confirmed cases and cumulative unidentified infections during a period, which is a measure of the efficiency of the public health surveillance system during the COVID-19 pandemic.

Many COVID-19 seroprevalence studies have been carried out around the world to inform estimation of the proportion of the population infected by COVID-19. In Canada, the Canadian COVID-19 Immunity Task Force reported that the national COVID-19 seroprevalence data suggested that over 70% of Canadians have been infected by COVID-19 from April 13, 2020, to August 15, 2022 [11]. The estimated total number of infections in Canada was 9 times the confirmed COVID-19 cases during the same period. A seroprevalence study in British Columbia by the BC Centre for Disease Control and the University of British Columbia estimated that over 60% of British Columbians have been infected by COVID-19 from March 2020 to August 2022, and the estimated total infections were 14 times the confirmed cases [12]. The proportion of COVID-19 infections has also been estimated using population surveys. Using probabilistic models and data on confirmed COVID-19 cases, as well as data on population health-seeking behaviors, US Center for Disease Control estimated that 146.6 million people in the US have been infected by COVID-19 (44% of the US population) from February 2020 to September 2021, which is 4 times the number of confirmed cases [13].

Mathematical modeling has been extensively used during the COVID-19 pandemic to project the spatial and temporal trend of the transmission and spread of the infection. Earlier model projections of the infected proportion have been significantly high, often in the 40 - 60% range during the first COVID-19 wave, under the assumption that no social distancing measures were to be implemented. In the study of exploring the spread of COVID-19 within different communities, the use of constant transmission rate and lack of information on how well cases are diagnosed do not take public health policy effects into consideration. That gives a reason to the drastically high estimation of the proportion of infected from COVID-19 [28]. COVID-19 data from Sweden during the Spring of 2020 shows that, under mild social distancing measures, the seroprevalence of COVID-19 in the country is no greater than 10% [29]. A main reason for the apparent over-projection by mathematical models was a lack of reliable data during the early stage of the pandemic. This leads to an important question for mathematical modelers of COVID-19: after almost three years of the pandemic, with all the medical knowledge we have gained of the SARS-Cov-2 virus and its variants, information on the public health measures that were implemented, and the epidemiological and public health data on the pandemic that are available, can we use mathematical models to retrospectively estimate the proportion of a population that were infected during a COVID-19 wave? and can the estimations be validated using the available seroprevalence data?

1.3 Objectives

In the first part of our study, we focused on the outbreak in Wuhan after the quarantine and lockdown (January 23, 2020) with given confirmed case dataset. We determined that the SIR model is a better choice than the SEIR model based on model selection criteria. To illustrate the linkage between the case infection ratio (or diagnosis rate) ρ and transmission rate β, the presence of nonidentifiability was shown when only the confirmed case data is used for model calibration.

The second part of our study aimed to give an affirmative answer to the question brought up at the end of Section (1.2), by demonstrating how simple mathematical models of COVID-19 of SIR type can be used to produce estimations of the proportion of infected population during the first COVID-19 wave in the Province of Alberta, Canada, during March-May of 2020. We collected and analyzed both published and confidential data on COVID-19 infection from several Alberta Health reports. Daily new COVID-19 case reports for the Province of Alberta from Alberta Health during the period from March 5 - June 1, 2020, were included in our study. We also used COVID-19 daily testing data in Alberta during the same period including calls to Healthline for PCR tests and COVID-19 online self-assessment forms. We tracked the changes in COVID-19 public health measures including restrictions on social gatherings, school closures, quarantine and isolation, policies on testing, and contact tracing so as to incorporate those public measure information into our model. The accuracy of our estimation is validated by the seroprevalence data for the Alberta population in June 2020 from the Alberta Public Health Precision Lab[14]. Our modeling approach was also adapted to provide dependable long-term model projections for subsequent COVID-19 waves. For long-term modeling projections on the Delta wave in Alberta using our modeling approach, we refer the reader to a report to Alberta Health [15].

Chapter 2

Mathematical theory for epidemics

2.1 Formulation of mathematical model for epidemics

2.1.1 Epidemic model for the initial Wuhan outbreak of COVID-19

A dynamic model was developed to quantify the COVID-19 dynamics of infections. We considered both SIR and SEIR frameworks to model the COVID-19 epidemic in Wuhan and decided which one outperformed the other based on the applied model selection criteria.

In SIR and SEIR models, compartment S denotes all susceptible populations in Wuhan, compartment I denotes the infectious population, and R denotes the confirmed cases. In the SEIR model, a latent compartment E is added to denote the individuals who are infected but not infectious. The latency of COVID-19 infection is biologically realistic because of an incubation period of about 14 days in which newly infected individuals may not be infectious while the virus is still incubating in the body [4]. The transfer diagrams for both models are shown in Figures (2.1) and (2.2). The epidemiological meanings of all model parameters are given in Tables (2.1) and (2.2). Since we use the newly confirmed case data for model calibration, which is matched to the $\rho I(t)$ term in both models, the death term in the R compartment has no effect on our model fitting [4]. Each model is described by the set of nonlinear autonomous ordinary differential equations (ODEs) below:

SIR

$$\begin{aligned} S' &= -\beta IS \\ I' &= \beta IS - \rho I - \gamma I \\ R' &= \rho I - dR \end{aligned} \tag{2.1}$$

SEIR

$$\begin{aligned} S' &= -\beta IS \\ E' &= \beta IS - \epsilon E \\ I' &= \epsilon E - \rho I - \gamma I \\ R' &= \rho I - dR \end{aligned} \tag{2.2}$$

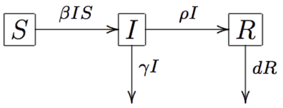

Figure 2.1: An SIR model

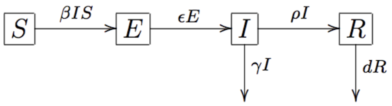

Figure 2.2: An SEIR model

Parameter	Epidemiological Meaning
β	Transmission rate
ρ	Diagnosis rate
γ	Recovery rate
I_0	Initial infections
τ	$\sigma^2 = 1/\tau$ is the variance of data noise

Table 2.1: Parameter description table for SIR model

Parameter	Epidemiological Meaning
β	Transmission rate
ρ	Diagnosis rate
γ	Recovery rate
ϵ	Transfer rate
E_0	Initial latent size
I_0	Initial infections
τ	$\sigma^2 = 1/\tau$ is the variance of data noise

Table 2.2: Parameter description table for SEIR model

2.1.2 Epidemic model for the first wave of COVID-19 in Alberta

We use a modified version of SIR, which is the SICR model with time-dependent parameters to model the transmission dynamics of COVID-19 within the population of the Province of Alberta. The model is described by the set of nonlinear autonomous ordinary differential equations in (2.4). The model schematic is shown in Figure (2.3), and the epidemiological meanings of all model parameters are given in Table (4.3). The compartment S contains all individuals who are susceptible to COVID-19, which essentially includes the entire Alberta population at the beginning of the COVID-19 pandemic. Compartment C contains all individuals who are confirmed positive for COVID-19 infection by a PCR test, and compartment I contains all individuals who are infected with COVID-19 but have not confirmed the infection by a PCR test. The compartment R contains individuals who have recovered from COVID-19 infection and are protected by immunity against reinfection by COVID-19.

Individuals in compartment I consist of what is called “hidden infection”, which includes many who are asymptomatically infected and others who may have symptoms but have not been tested. These individuals help to spread transmissions in the community. In contrast, individuals in compartment C have been confirmed of being positive for COVID-19, and by public health restrictions, are under strict isolation orders for 10-14 days. Because of the lack of contact with the outside world, we assume that they are not infective.

Recovery from COVID-19 can take two different routes: an individual, especially those who are asymptomatically infected, recovers from the infection without being tested, and this is denoted by $\gamma_1 I(t)$ in the model, and $1/\gamma_1$ is the mean infectious period; or an individual is tested positive and isolated in compartment C and recovers during isolation. Recovery from compartment C is denoted by $\gamma_2 C$. Both of these routes are included in the model as shown in Figure (2.3).

Public health measures are implemented in the transmission term $\beta(t)I(t)S(t)$ and the testing term $\rho(t)I(t)$. Social distancing measures will decrease the transmission coefficient $\beta(t)$ and COVID-19 testing levels and capacity will influence the parameter $\rho(t)$. In particular, $\rho(t)I(t)$ is the daily number of positive COVID-19 tests, which is part of the public health data we use for model calibration, and the parameter $\rho(t)$ is the case-infection ratio: the ratio between daily new case reported number and number of people living with COVID-19 infection in the community on the day. It is an indicator for the efficiency of public health surveillance and answers the question: for each COVID-19 case identified, how many hidden infections are there in the community.

Many factors can influence the values of transmission coefficient $\beta(t)$, including the average number of contacts among individuals in the populations, and the average probability of transmission for each contact, which may depend on both the infectivity of the infected individual and the susceptibility of the susceptible individual during each contact. The value of $\beta(t)$ is averaged over individual variations in the population. The time dependence of $\beta(t)$ is informed by the changes in COVID-19-related public health restrictions that directly reduces transmission. Piece-wise linear functions are used to incorporate different time points of changes in policy and parameters in these functions are estimated through model fitting to data.

Transmission rate $\beta(t)$ remained the same from March 9 to March 15, 2020. In the following three days, the province of Alberta declared further public health policies to use "all powers necessary" to "keep Albertans safe" [16]. This included but was not limited to all closure for all day-cares, suspending classes for all K-12 schools, remote teaching for post-secondary institutions, mandatory masking requirements, and restriction occupy in fitness facilities, restaurants and etc. In this period, the transmission rate decreased dramatically and stayed low. All mandatory public health restrictions were lifted around the beginning of May as the first wave subsided and COVID-19 hospitalizations continued to decline, the transmission rate increased by a certain amount at the end stage of the first wave in our model.

COVID-19 testing, contact tracing, and isolation measures are important parts of the overall public-health control measures against COVID-19. In our model, COVID-19 testing moves an infected individual from compartment I to compartment C for identified cases, after a positive PCR test. This process is modeled by $f\rho(t)I(t)$. Here f is the probability an infected individual will seek a PCR test, and the time-dependent $\rho(t)$ is defined by:

$$\rho(t) = \frac{\text{Newly reported positive COVID-19 cases at time } t}{\text{Total number of Healthline calls and completed online self-assessment forms at time } t}. \quad (2.3)$$

Then the total number of new cases at time t is given by $f\rho(t)I(t)$. The function $f\rho(t)$ captures the changes in health-seeking behaviors in the population during the course of the epidemic wave of COVID-19.

Since the first COVID-19 wave in Alberta lasted only three months, we assume that the natural birth (24,898) and non-COVID-19 death (13,949), net inter-provincial migrations (-7,390) and net international migrations (12,093) are negligible compared to the total population (4.42 million). COVID-19-related deaths are mainly among the identified cases and are included in the dC term. Since the daily COVID-19 death is comparably small, dC is set to be zero in our model.

The SICR model is much simpler than most mathematical models used for COVID-19 in the literature, such as SEIR or SEIAR models. Models that are more complex have a larger number of parameters than SICR that need to be estimated by model fitting to data. Since the public health data used for model calibration is the same, a larger number of model parameters to be estimated gives rise to a higher degree of nonidentifiability in model calibration, which leads to a higher degree of uncertainty in model estimations and predictions.

SICR

$$\begin{aligned} S' &= -\beta(t)IS \\ I' &= \beta(t)IS - f\rho(t)I - \gamma_1 I \\ C' &= f\rho(t)I - \gamma_2 C - dC \\ R' &= \gamma_1 I + \gamma_2 C \end{aligned} \tag{2.4}$$

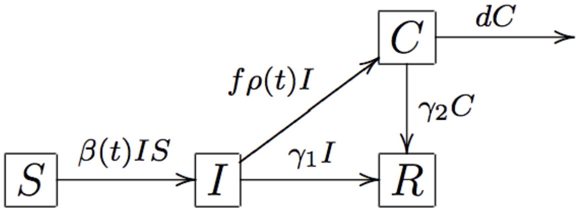

Figure 2.3: An SICR model

2.2 Mathematical theory for epidemics

The nonlinear autonomous ODE systems (2.1), (2.2), and (2.4) can be written in vectorized form

$$\boldsymbol{x}' = \boldsymbol{f}(\boldsymbol{x}), \tag{2.5}$$

where $\boldsymbol{x} = \langle x_1, x_2, ..., x_k \rangle$ and $\boldsymbol{f} = \langle f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), ..., f_k(\boldsymbol{x}) \rangle$, with the vector of initial conditions $\boldsymbol{x}_0 = \langle x_{10}, x_{20}, ...x_{k0} \rangle$. Here $\boldsymbol{x}$ denotes the vector of state variables and $\boldsymbol{f}$ denotes the corresponding vector fields. We would like to solve the initial value problem as follows:

Let D be an open set in $\mathbb{R}^n$. Assume that $\boldsymbol{f} \in C\left(D \to \mathbb{R}^n\right)$. Find a solution of the ODE model in $\mathbb{R}^n$ which subject to the initial condition $\boldsymbol{x}_0$,

$$\begin{aligned} \boldsymbol{x}' &= \boldsymbol{f}(\boldsymbol{x}) \\ \boldsymbol{x}(t_0) &= \boldsymbol{x_0}. \end{aligned} \tag{2.6}$$

2.2.1 Existence and uniqueness

Definition 2.1 *(Lipschitz Condition): Let D be an open subset of $\mathbb{R}^n$. A function $\boldsymbol{f} \in C\left(D \to \mathbb{R}^n\right)$ is said to satisfy Lipschitz condition with respect to $\boldsymbol{x}$ if there exists (Lipschitz) constant $K > 0$, such that*

$$\left|\boldsymbol{f}\left(t, \boldsymbol{x}_1\right) - \boldsymbol{f}\left(t, \boldsymbol{x}_2\right)\right| \leq K \left|\boldsymbol{x}_1 - \boldsymbol{x}_2\right|$$

for all $(t, \boldsymbol{x_1}), (t, \boldsymbol{x_2}) \in D$, where $|\cdot|$ is any norm in $\mathbb{R}^n$.

Theorem 2.1 *(Picard Local Existence Theorem and Uniqueness under Lipschitz) Suppose*

(1) $\boldsymbol{f} \in C\left(D \to \mathbb{R}^n\right)$*,*

(2) $\boldsymbol{f}$ *satisfies Lipschitz condition in* D *with respect to* $\boldsymbol{x}$*,*

Then there exists $\alpha > 0$ *such that the IVP has a unique solution for* $t \in [t_0 - \alpha, t_0 + \alpha]$*.*

Theorem 2.2 *(Peano existence theorem): Suppose that* $\boldsymbol{f} \in C\left(D \to \mathbb{R}^n\right)$ *and* $\boldsymbol{f}$ *satisfies Lipschitz condition in* D *with respect to* $\boldsymbol{x}$*. Then for each point* $\boldsymbol{x}_0 \in D$*, there exists a maximal interval* (ω_-, ω_+) *on which the initial value problem* (2.6) *has a unique solution,* $\phi(t)$*.*

$\boldsymbol{f} = \langle f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), ..., f_k(\boldsymbol{x})\rangle$ in the ODE models above are continuous, differentiable functions for $\boldsymbol{x} \in \mathbb{R}^n$, therefore $\boldsymbol{f}$ satisfies Lipschitz condition over $\mathbb{R}^n$.

By Theorem (2.1), there exists an $\alpha > 0$ such that the initial value problem (2.6) has a unique solution $\phi(t)$ on the interval $t \in [t_0 - \alpha, t_0 + \alpha]$. By Theorem (2.2), a solution $\phi(t)$ can be extended to its maximal interval of existence (ω_-, ω_+). When D is compact, there exists a solution for $t \in (-\infty, \infty)$.

Definition 2.2 *(Orbits of solutions) The orbit of a solution* $\boldsymbol{x}(t, \boldsymbol{x_0})$ *of the initial value problem* (2.6) *is defined as:*

$$\gamma(\boldsymbol{x_0}) = \{\boldsymbol{x}(t, \boldsymbol{x_0}) : t \in (\omega_-, \omega_+)\}. \tag{2.7}$$

Definition 2.3 *(Positive invariance) A subset $K \in \mathbb{R}^n$ is positively invariant if all solutions starting in K remain in K, i.e. $\boldsymbol{x_0} \in K \Longrightarrow \boldsymbol{x}(t, \boldsymbol{x_0}) \in K, t \geq 0$.*

Definition 2.4 *(Limit sets)*
(1) The ω-limit set of a solution $\boldsymbol{x}(t, \boldsymbol{x_0})$ is $\omega(\boldsymbol{x_0}) = \{\boldsymbol{x} \mid \text{there exists } t_n \to \infty \text{ such that } \boldsymbol{x}(t_n, \boldsymbol{x_0}) \to \boldsymbol{x}\}$.
(2) The α-limit set of a solution $\boldsymbol{x}(t, \boldsymbol{x_0})$ is $\alpha(\boldsymbol{x_0}) = \{\boldsymbol{x} \mid \text{there exists } t_n \to -\infty \text{ such that } \boldsymbol{x}(t_n, \boldsymbol{x_0}) \to \boldsymbol{x}\}$.

The ODE Model (2.1) is well-posed since the non-negative cone of $\mathbb{R}^3$,

$$\mathbb{R}^3_+ = \{(S, I, R) \in \mathbb{R}^3 | S \geq 0, I \geq 0, R \geq 0\}$$

is positively invariant with respect to (2.1).

To verify the positive invariance of $\mathbb{R}^3_+$, we consider the direction of the vector field $\langle -\beta IS, \beta IS - f\rho I - \gamma I, \rho I - dR \rangle$ on the coordinate planes.
(1) On the SR-plane: $I = 0$ on this plane, and

$$\left.\frac{dI}{dt}\right|_{I=0} = 0,$$

This shows that the vector field on the SR-plane is tangent to the SR-plane. This tangency also implies that the SR-plane itself is invariant.
(2) On the IR-plane: $S = 0$ on this plane, and

$$\left.\frac{dS}{dt}\right|_{S=0} = 0,$$

Therefore, the IR-plane is also invariant. No solutions in the interior of $\mathbb{R}^3_+$ can escape through the IR-plane.

(3) On the SI-plane: $R = 0$ on this plane, and

$$\left.\frac{dR}{dt}\right|_{R=0} = \gamma I \geq 0,$$

Therefore, the vector field on the SI-plane points to the interior of $\mathbb{R}^3$. No solutions can escape the interior through the SI-plane.

Eventually, all solutions with nonnegative initial conditions stay in $\mathbb{R}^3_+$ for $t \geq 0$.

2.2.2 Equilibrium, stability and global phase portrait

To simplify our analysis, we can ignore the R equation in system (2.1) since the first two equations do not contain R. Therefore consider the following equivalent system:

$$\begin{aligned} S' &= -\beta I S \\ I' &= \beta I S - \rho I - \gamma I \end{aligned} \tag{2.8}$$

in a 2-dimensional feasible region

$$G = \{(S, I) \in \mathbb{R}^2_+ | S \geq 0, I \geq 0\}. \tag{2.9}$$

Dividing the two equations from (2.8), we obtain

$$\frac{dI}{dS} = -1 + \frac{\rho + \gamma}{\beta S} \ = -1 + \frac{\bar{S}}{S}, \tag{2.10}$$

where $\bar{S} = \frac{\rho+\gamma}{\beta}$ is the threshold number for S_0. We also call it the critical community size to sustain the epidemic. Integrating (2.10), we obtain the first integral of system (2.8)

$$\phi(S, I) = I + S - \bar{S} \log S = C, \tag{2.11}$$

and we have

$$\frac{d}{dt}\phi(S(t), I(t)) = I'(t) + S'(t) - \bar{S}\frac{S'(t)}{S(t)} = -(\gamma + \rho)I(t) + \bar{S}\beta I(t) = 0 \tag{2.12}$$

for all t, where C is an integration constant and can be determined from S_0 and I_0. In this case, the function $\phi(S, I)$ remains a constant along the solution $(S(t), I(t))$ and we are able to present the trajectories of $(S(t), I(t))$ of system (2.8) by a family of level curves $\phi(S, I) = C$ for different values of C. The global phase portrait is given by these level curves of $\phi(S, I)$ and is depicted in Figure (2.4).

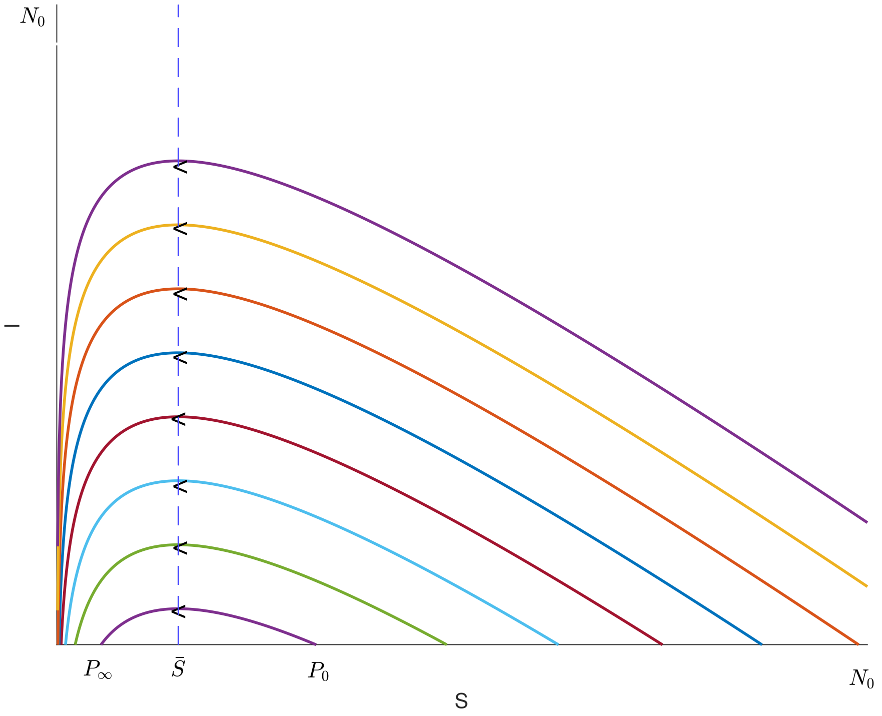

Figure 2.4: Family of epidemic curves

Equilibrium and stability.

The equilibria of the system (2.8) are given by the solutions of $S' = I' = 0$. We denote the set of all equilibria as below,

$$\mathbb{S}_+ = \{(S, I) \mid S \geq 0, I = 0\}.$$

For any equilibrium $P = (S_0, 0) \in \mathbb{S}_+$, we consider the linearized system of (2.8) at P given by:

$$\boldsymbol{x'} = J(P)\boldsymbol{x}, \tag{2.13}$$

where $\boldsymbol{x} = (S, I)$ denotes all the model compartments, and $J(P)$ is the Jacobian matrix of system (2.8) evaluated at P, which can be computed as:

$$J(P) = \begin{bmatrix} 0 & -\beta S_0 \\ 0 & \beta S_0 - (\rho + \gamma) \end{bmatrix}. \tag{2.14}$$

Because of the upper triangularity of $J(P)$, its eigenvalues are $\lambda_1 = 0$ and $\lambda_2 = \beta S_0 - (\rho + \gamma)$. For eigenvalue $\lambda_1 = 0$, the corresponding eigenspace is $\mathbb{S}_+$, which is the set of all equilibria contained in system (2.13). Since $\mathbb{S}_+$ is independent of time t, the local stability of P is determined solely on the sign of the other eigenvalue λ_2. For $\lambda_2 = \beta S_0 - (\rho + \gamma)$, when $S_0 < \bar{S}$, we have $\lambda_2 < 0$, which indicates P is attracting in the direction transversal to $\mathbb{S}_+$. When $S_0 > \bar{S}$, we have $\lambda_2 > 0$, and P is repelling in the direction transversal to $\mathbb{S}_+$.

Additionally, in Figure (2.4), for each initial point $\boldsymbol{x_0} = (S_0, I_0)$, each phase line intersects the set of equilibria $\mathbb{S}_+$ at two points, denoted as P_0 and P_∞. This implies the corresponding solution $\boldsymbol{x}(t, \boldsymbol{x_0})$ satisfies:

$$\lim_{t \to -\infty} \boldsymbol{x}(t, \boldsymbol{x_0}) = P_0, \quad \lim_{t \to \infty} \boldsymbol{x}(t, \boldsymbol{x_0}) = P_\infty.$$

The ω-limit set is the single equilibrium P_∞ on the left of the $\bar{S}$, and α-limit set is the single equilibrium P_0 on the right of the $\bar{S}$. The solution $\boldsymbol{x}(t, \boldsymbol{x_0})$ is a heteroclinic orbit connecting the two equilibria P_0 and P_∞.

Interpretation of mathematical results.

Critical community size $\bar{S} = \frac{\rho+\gamma}{\beta}$ defined in Section (2.2.2) is a crucial threshold value during an epidemic. If the initial population at the beginning of the epidemic is smaller than this value, then no matter how large the initial infected population size is, $I(t)$ will monotonically decrease as time goes. Then the epidemic will eventually diminish and no new outbreaks occur. When the initial population at the beginning of the epidemic equals this critical community size $\bar{S}$, the maximum value of I_{max} of I is achieved. This peak number of infections and its corresponding peak time are the two model outcomes that public health agencies are mostly interested in. When the initial population at the beginning of the epidemic is larger than this critical value $\bar{S}$, enough susceptibles are there to start the epidemic: $I(t)$ initially increases while $S(t)$ decreases until $S = \bar{S}$, yields the maximal $I(t)$; then $I(t)$ decreases to 0 while $S(t)$ decreases below $\bar{S}$. This gives the rise-peak-decline cycle of an epidemic [26].

One great obstacle to accurate model estimation is the final-size challenge. In the epidemic modeling literature, the *final size* of an epidemic is the number of susceptible individuals who are not infected at the end of the epidemic. In the context of the epidemic curves shown in Figure (2.4), this is given by P_∞, the end value of the susceptible population. In the public health interest of estimating the scale of the epidemic, the final size can be more appropriately defined as the total number or proportion of people who are infected. In the context of Model (2.8), this is given by $P_0 - P_\infty$ in numbers, or $1 - P_\infty/P_0$ in proportion.

In addition, with the same initial population of susceptible, a higher number of initial infections will lead to a severer epidemic, and a larger proportion of infected. This gives a solid evidence for public health authorities to implement strict shutdowns in the airports and ports to avoid the increase in initial infections at the beginning of the epidemic.

Chapter 3

Model fitting and sensitivity analysis

For the Wuhan model, the data used for both Models (2.1) and (2.2) is the newly confirmed case data in Wuhan city from the official reports during January 21 - February 4, 2020 [17]. The distribution of the count data was approximately normal and the probability model for the observed count data in our study was assumed to be a normal distribution with mean given by ρI and variance given by $1/\tau$ [4]. For the SIR model: the transmission rate β, diagnosis rate ρ, the initial infected population I_0 on January 21, 2020 ($t = 0$), and the variance of data $q = 1/\tau$ cannot be estimated directly from the data. These four parameters were assessed by fitting the model prediction to the observed case data. For the SEIR model, two extra parameters need to be estimated from the data: transfer rate ϵ from E to I, and the initial latent population E_0 on January 21, 2020.

For the Alberta model, the datasets used for Model (2.4) are the daily confirmed case data, daily testing data, and daily death data in Alberta from the Alberta Health reports during March 5 – June 1, 2020. Since the increasing testing number indicates the increasing confirmed case number, the distribution of the testing data was assumed to be normal as well. For the SICR model, time-dependent transmission rate $\beta(t)$, and time-dependent diagnosis rate $\rho(t)$ need to be estimated from the data. Figures (3.1) and (3.2) show the constructed piecewise linear function $\beta(t)$ and $\rho(t)$ with labeled subparameters. In order to precisely estimate the cumulative infection-fatality ratio, we assumed daily death data follow the Poisson distribution with different mean values each day. A Gaussian mixture model is introduced to fit the actual data of 7-day-average daily death, then a 3-Gaussian mixed curve is chosen as the desired mean values for death numbers each day. In this way, a reasonable estimation of the cumulative numbers of reported death can be generated for each sample in the model, which in turn can be taken to compute the cumulative infection-fatality ratio.

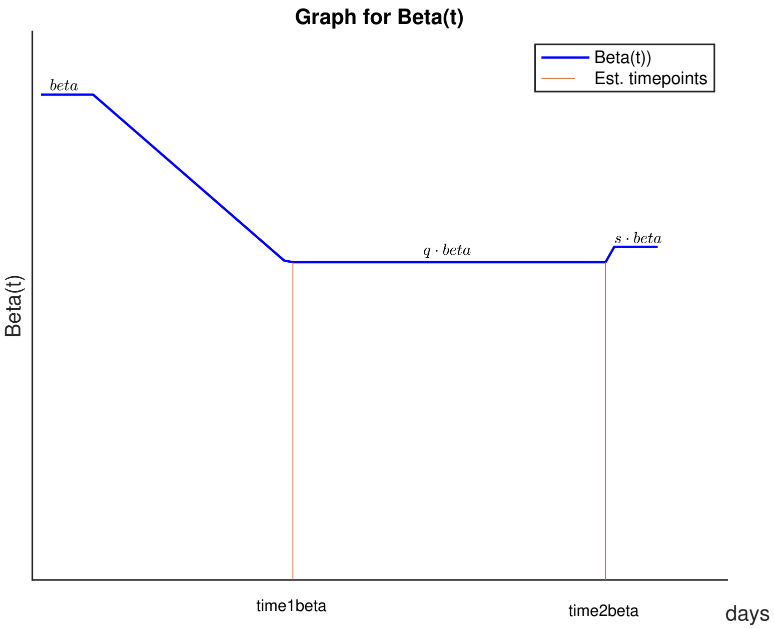

Figure 3.1: Piecewise linear function $\beta(t)$

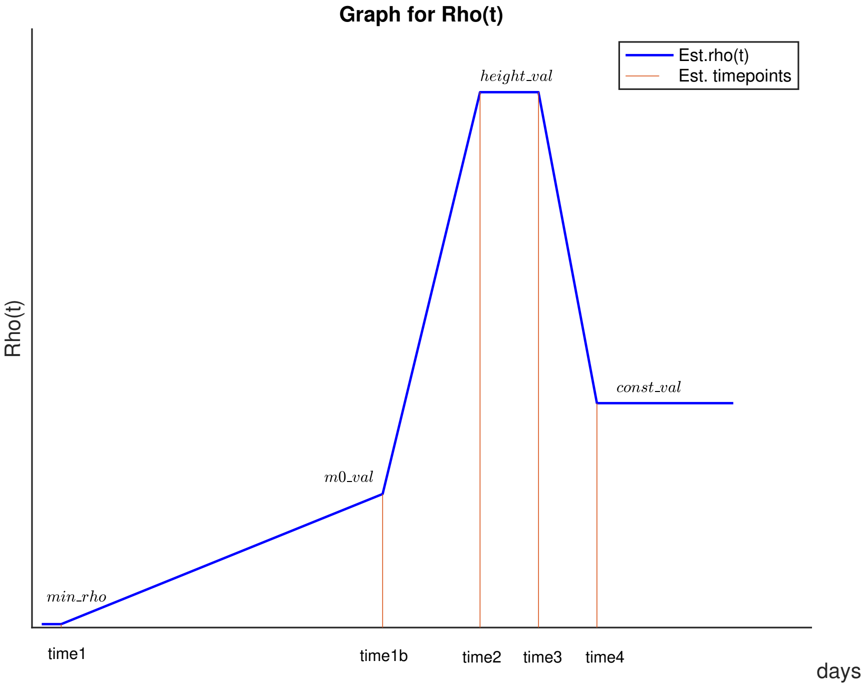

Figure 3.2: Piecewise linear function $\rho(t)$

3.1 Likelihood functions and Bayesian framework in ODE models

3.1.1 Mathematical settings

Consider the Model (2.5) stated previously:

$$\boldsymbol{x}' = \boldsymbol{f}(\boldsymbol{x}),$$

where $\boldsymbol{x} = (x_1, ..., x_k)$ denotes the vector of state variables, and $\boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}), ...f_k(\boldsymbol{x}))$ denotes the vector fields. We let $\boldsymbol{\theta} \in \mathbb{R}^s$ be the vector of all model parameters, including initial conditions $\boldsymbol{x_0} = (x_{01}, ...x_{0k})$. We assume that there exists a unique solution $\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{\theta}, t)$ for each given $\boldsymbol{\theta}$.

Data such as newly confirmed cases are linear or nonlinear combinations of the solutions $\boldsymbol{x}(\boldsymbol{\theta}, t)$ in the form:

$$y = y(\boldsymbol{x}(\boldsymbol{\theta}, t), t).$$

If the dataset is collected at N time points $t_1, t_2, \cdots, t_N$, we will fit the model outputs

$$y_i = y\left(\boldsymbol{x}\left(\boldsymbol{\theta}, t_i\right), t_i\right), \quad i = 1, 2, \cdots, N,$$

to the time series dataset

$$D = \left\{D_1, D_2, \cdots, D_N\right\}.$$

Likelihood function. Likelihood function is the joint probability of the observed data viewed as a function of the parameters of the chosen probability model. In order to account for noise in the data, we let $f_i\left(D_i\right)$ with mean y_i and variance $\sigma_i^2=1/\tau_i$, $i=1,2,\cdots,N$ denote the probability of D_i at time t_i. We will consider the normal distribution in our study. Now consider the likelihood function [4]:

$$L(\theta|D)=CP(D\mid\theta)=Cf_1\left(D_1\right)f_2\left(D_2\right)\cdots f_N\left(D_N\right),$$

where C is a constant independent of θ used to simplify the likelihood function [18].

Bayesian framework. Given the observed data D with unknown parameters θ, the probability model is expressed as $P(D|\theta)$. Parameter θ is randomly distributed following the prior distribution $P(\theta)$. Then statistical inference for θ can be made based on the posterior distribution $P(\theta|D)$. From Bayes Theorem, we obtain:

$$P(\theta|D)=\frac{P(D|\theta)P(\theta)}{P(D)}=\frac{P(D|\theta)P(\theta)}{\int P(D|\theta)P(\theta)d\theta}\propto L(\theta|D)P(\theta)=\pi(\theta|D),$$

where $L(\theta|D)$ is the likelihood function. By the above formula, the unnormalized posterior distribution $\pi(\theta|D)$ is given as long as we have pre-specified likelihood function $L(\theta|D)$ and prior distribution $P(\theta)$. In infectious disease modeling, epidemiology information often offers specific ranges for epi-related parameters to be estimated. The utilization of prior information about those unknown parameters is quite useful in mathematical modeling for statistical inference and model fitting.

3.1.2 Gaussian probability model on data and its likelihood function

It is common to use a negative binomial probability model for observed count data. When the mean of a negative binomial distribution is large, it approximates a normal distribution. Since the newly confirmed cases are approaching the large value quickly, the distribution of the count data will be approximately normal and the probability for the observed count data in our study is assumed to be a normal distribution. Also, since the case infection ratio $\rho(t)$ in our second study is taking decimal values; it is appropriate to use the normal distribution for this as well.

A method is shown to fit an ODE model with Gaussian probability data. This method will be applied to both studies when fitting daily new confirmed case data and to the second study when fitting daily case infection ratio.

Let Y be a random variable following the Gaussian distribution with mean μ and variance $\sigma^2 = \frac{1}{\tau} > 0$, then the corresponding continuous probability density function is given by :

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}. \tag{3.1}$$

Suppose there are m independent time series datasets all following the Gaussian distribution. The j^{th} time series dataset is given by $D^j = (d_1^j, ...d_n^j)$ at times $(t_1^j, ...t_n^j)$, then the corresponding continuous probability density function of observing d_i^j is given by [19]:

$$f(d_i^j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{1}{2\sigma_j^2}(d_i^j - \mu_i^j)^2}. \tag{3.2}$$

As we are fitting model solution to the j^{th} time series dataset,

$$\mu_i^j = \boldsymbol{x_j}(\boldsymbol{\theta}, t_i^j) = (x_1(\boldsymbol{\theta}, t_i^j), ...x_k(\boldsymbol{\theta}, t_i^j)),$$

where $\boldsymbol{\theta}$ is the vector of parameters to estimate.

Then the likelihood function for the j^{th} time series dataset is given by:

$$L(\boldsymbol{\theta}|D^j) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{\left(d_i^j - \boldsymbol{x_j}\left((\boldsymbol{\theta}, t_i^j)\right)\right)^2}{2\sigma_j^2}}. \tag{3.3}$$

If m independent datasets with the same vector of parameters $\boldsymbol{\theta}$, then the overall likelihood $L(\boldsymbol{\theta}|D^1, ..., D^m)$ of the combined m independent datasets is obtained by multiplying the m likelihood functions $L(\boldsymbol{\theta}|D^1)$,..., and $L(\boldsymbol{\theta}|D^m)$ together, given:

$$L\left(\boldsymbol{\theta} \mid D^1, \ldots, D^m\right) = L\left(\boldsymbol{\theta} \mid D^1\right) \cdot \ldots \cdot L\left(\boldsymbol{\theta} \mid D^m\right). \tag{3.4}$$

For all m independent datasets with the same vector of parameters $\boldsymbol{\theta}$ and for each j^{th} dataset, we have model solution $\mu^j = x_j(\boldsymbol{\theta}, t)$ to fit to the j^{th} dataset, then the combined likelihood $L\left(\boldsymbol{\theta} \mid D_1, \ldots, D_m\right)$ is given by the following equation:

$$L\left(\boldsymbol{\theta} \mid D^1, \ldots, D^m\right) = \prod_{j=1}^{m} \prod_{i=1}^{n_j} \left(\frac{1}{\sqrt{2\pi}}\right)^m \frac{1}{\sigma_j} e^{-\frac{\left(d_i^j - \boldsymbol{x_j}\left(\boldsymbol{\theta}, t_i^j\right)\right)^2}{2\sigma_j^2}}. \tag{3.5}$$

Fittings for Models (2.1), (2.2), and (2.4) are done through the affine invariant ensembled Markov chain Monte Carlo (MCMC) algorithm by sampling from the natural log of the unnormalized posterior distribution, $\pi(\theta|D)$, where

$$\pi(\theta|D) = L(\theta|D)P(\theta), \tag{3.6}$$

and $\theta \in \mathbb{R}^s$ is a vector of parameters, $D = (D^1, ..., D^m)$ is the m independent time series datasets, $L(\theta|D)$ is the likelihood function, and $P(\theta) = \mathrm{P}(\theta_1) \times ... \times P(\theta_s)$ is the prior distribution for all parameters.

3.2 Affine invariant ensemble Markov chain Monte Carlo algorithm

MCMC algorithms are used widely for approximating the target posterior distribution given likelihood function $L(\theta|D)$ and prior distribution $P(\theta)$. They usually start with some initial guess θ_0 from the parameter space, then the algorithm iteratively generates a new sample θ_t from the posterior distribution based on the previous sample θ_{t-1}. The sample chain follows the Markov property as each sample depends only on its previous sample. The Markov process continues until the sample chain has arrived at its stationary distribution, which is our targeted posterior distribution $\pi(\theta|D)$. We then can obtain a sample of the desired posterior distribution by recording states from the chain. The more steps that are included, the more closely the distribution of the sample matches the actual desired distribution, and the more accurate our model estimation will be through the sample fitting. Commonly used MCMC algorithms include the Metropolis-Hastings algorithm and Random-

Walk Metropolis-Hastings algorithms [20]. Although those algorithms have achieved great success, they still need some improvements in order to handle the nonidentifiability issue, which keeps appearing when we have a complicated model with limited data resources.

In our studies, we use an improved MCMC algorithm, the affine invariant ensemble Markov chain Monte Carlo algorithm, which outperforms Metropolis-Hastings and other MCMC algorithms, especially in the presence of nonidentifiability. Essentially, the nonidentifiability issue arises whenever there exists a surface in parameter space such that all parameter choices on this surface yield the maximum likelihood value. This likelihood surface will be highly anisotropic and cause the traditional MCMC sampler slow or even fail to converge. Affine invariant MCMC sampler starts by initializing k particles simultaneously, then each particle will be updated iteratively based on the current positions of all the other particles. It turns out that by stretch move rule [21], each of the k particles will have the same stationary distribution that is identical to $\pi(\theta|D)$. In addition, the resulting sample paths $\{X_i(\theta, t)\}_{i=1}^{k}$ satisfies the affine invariant property:

$$X_i(A\theta + b, t) = AX_i(\theta, t) + b, \quad \theta \sim \pi(\theta|D), \quad i = 1, 2, \cdots, k, \tag{3.7}$$

where $A \in \mathbf{R}^{s\times s}$, $b \in \mathbf{R}^{s}$, and $\phi := A \cdot +b$ is an affine transformation that maps θ to $A\theta + b$. Suppose the posterior distribution after the transformation is highly skewed given by $\pi(\phi(\theta)|D)$, then by the affine invariant property (3.7), the corresponding sample path $X(\phi(\theta), t)$ for $\pi(\phi(\theta)|D)$ is given by $\phi(X(\theta, t))$, which is essentially the map image resulting from the original sample path

$X(\theta, t)$. We conclude that the affine invariant MCMC algorithm is uniformly effective [21] on problems in highly skewed distributions that can be better scaled through affine transformations.

In our studies, a large number of burn-in samples are needed to let Markov chains reach stationary states, and an adequate amount of actual samples are needed to obtain the desired posterior distribution through the MCMC algorithm.

3.3 Akaike information criterion for model selection

Let $L(\hat{\theta})$ be the maximum likelihood value at a best-fit parameter $\hat{\theta}$. Let p be the number of parameters in a model and N be the number of sample sizes for the given time series data. The Akaike Information Criterion (AIC) is defined as [22]:

$$AIC = -2 \ln L(\hat{\theta}) + 2p. \tag{3.8}$$

When the sample size is small, AIC may prefer a model with more parameters without considering the overfitting issue. To address such potential overfitting, a corrected AIC should be used [23]:

$$AIC_c = AIC + \frac{2p^2 + 2p}{n - p - 1}. \tag{3.9}$$

In our study, we use the AIC_c to correct for the small sample size. For the Wuhan outbreak study, AIC_c will be computed for both SIR and SEIR models. The model with a smaller AIC_c is considered the best model.

3.4 Partial Rank Correlation Coefficient for sensitivity analysis

Sensitivity analysis is an important component of statistical inference on model calibration results. It is a widely adopted study for quantifying uncertainty, as well as identifying crucial parameters in the model. Traditionally, there are two different approaches when performing sensitivity analysis on mathematical models. One is called local sensitivity analysis, which utilizes partial derivatives of model outputs with respect to model parameters as sensitivity indices. Since those partial derivatives are evaluated at a fixed point in parameter space, it is only effective in some small neighborhoods around that point. While the other method considers a broader region and all the effects on model outputs caused by certain perturbations of parameter values within this range will be quantified. Global sensitivity analysis is usually implemented by Monte Carlo stimulation, which have been discussed in the previous section. In epidemiology, the model parameters usually have a huge uncertainty, hence the global sensitivity analysis method is often more widely applied.

In this study, we will perform global sensitivity analysis based on model calibration results. After the model calibration process, a collection of posterior samples for model parameter θ will be obtained from the MCMC algorithm. Suppose there are N posterior samples, then the unnormalized joint posterior distribution $\pi(\theta|D)$ is approximated by $\Theta = \{\theta_i\}_{i=1}^{N}$, where $\theta_i \in \mathbb{R}^s$ denotes an s dimensional successive posterior sample. The Maximum a Posterior estimator (MAP) for the model parameter θ is given by:

$$\theta^* = \max_{\theta \in \Theta} \pi(\theta|D). \tag{3.10}$$

For each independent component θ^j of θ, we can obtain its 95% credible interval given by $I_j = [q_{0.025}(\theta^j), q_{0.975}(\theta^j)]$, where $j = 1, 2, \cdots, s$; $q_{0.025}$ and $q_{0.975}$ are the 2.5 and 97.5 percentiles of the marginal posterior distribution $\pi(\theta_j|D)$. Therefore, we can define the global sensitivity analysis parameter range as the hypercube $I = I_1 \times I_2 \times \cdots \times I_s$, and all the samples within this range is given by $\Theta_I = \Theta \cap I$. Having determined the MAP estimator θ^* together with its 95% truncated sample set Θ_I, it now comes to choose a proper sensitivity index. Among different types of global sensitivity analysis indices, the most popular used is: Partial Rank Correlation Coefficient (PRCC) [24].

PRCC is designed to measure the strength of the monotone relationships between model outputs and model parameters. In general, for each sample $\theta_i \in \Theta_I$, we can stimulate its corresponding model output as $y(\theta_i)$. Hence the overall Monte Carlo simulations for each posterior sample can be denoted by $Y = \{y(\theta) \in \mathbb{R} : \theta \in \Theta_I\}$. We start by assuming that there exist some monotone relations for posterior samples Θ_I and the model outputs Y. Suppose these relations are additionally linear, then the correlation coefficients (CC) between them can be calculated to quantify the strength of such linkage. Under most circumstances, the connections between Θ_I and Y are monotone but nonlinear. Rank transformation is then performed on Θ_I and Y to eliminate the nonlinear correlation. Since there are N_I samples included in Θ_I, and each of them has s independent components, Θ_I can be interpreted as a matrix of size $N_I \times s$. Y can be seen as a column vector of length N_I. Then each column of the concatenated matrix $[\Theta_I|Y]_{(N_I+1)\times s}$ is replaced by its ranking vector, which returns the corresponding sorted positions of each element. This way, the effects associated with the nonlinear relation can be eradicated.

Suppose the rank transformed concatenated matrix within range I is given by $[\Theta_r|Y_r]$, then the CC between component θ_r^j and Y_r can be computed as:

$$\mathrm{CC}(\theta_\mathrm{r}^\mathrm{j}, \mathrm{Y_r}) = \frac{\mathrm{Cov}(\theta_\mathrm{r}^\mathrm{j}, \mathrm{Y_r})}{\sqrt{\mathrm{Var}(\theta_\mathrm{r}^\mathrm{j})\mathrm{Var}(\mathrm{Y_r})}} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{N_I}}(\theta_\mathrm{i}^\mathrm{j} - \bar{\theta_\mathrm{i}^\mathrm{j}})(\mathrm{Y_i} - \bar{\mathrm{Y}}_\mathrm{r})}{\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{N_I}}(\theta_\mathrm{i}^\mathrm{j} - \bar{\theta_\mathrm{i}^\mathrm{j}})^2 \sum_{\mathrm{i}=1}^{\mathrm{N_I}}(\mathrm{Y_i} - \bar{\mathrm{Y}}_\mathrm{r})^2}}, \quad (3.11)$$

where $\bar{\theta_i^j}$, $\bar{Y}_r$ are mean values of the j-th, (N_I+1)-th columns of the matrix $[\Theta_r|Y_r]$, respectively. The above rank transformed CC value does not consider the internal relations within sample matrix Θ_r. In order to tackle this, we calculate the CCs between the two residuals given by $\theta_i^j - \hat{\theta_i^j}$ and $Y_r - \hat{Y_r^j}$, where

$$\hat{\theta_i^j} = a_j + \sum_{k=1,k\neq j}^{s} a_k \theta_r^k, \qquad \hat{Y_i^j} = b_j + \sum_{k=1,k\neq j}^{s} b_k \theta_r^k, \quad (3.12)$$

with $\{a_i\}_{i=1}^s$ and $\{b_i\}_{i=1}^s$ been estimated from the normal multi-linear regression model. This way, the linear effects are removed among the sample set Θ_I, and more accurate partial CC values can be obtained. Then Partial Rank Correlation Coefficient (PRCC) is designed as:

$$\mathrm{PRCC}(\theta_\mathrm{r}^\mathrm{j}, \mathrm{Y_r}) = \mathrm{CC}(\theta_\mathrm{i}^\mathrm{j} - \hat{\theta_\mathrm{i}^\mathrm{j}}, \mathrm{Y_r} - \hat{\mathrm{Y}_\mathrm{r}^\mathrm{j}}), \quad (3.13)$$

where $j = 1, 2, \cdots s$. PRCC value is in between -1 to 1. It reflects the strength of the monotone relation between the model output and a parameter component. A PRCC value close to 1 indicates the model output is extremely sensitive to this parameter component, and a PRCC value close to 0 indicates the model output is insensitive to the specific component. A positive PRCC value indicates a positive relationship between the model outcome and a specific parameter, and vice versa.

Chapter 4

Numerical results

4.1 Numerical results on the first study: Wuhan outbreak

4.1.1 Model fitting results for SIR and SEIR

Parameter	Epidemiological Meaning	Prior	95% Credible Interval	Best-Fit Value
β	Transmission rate	U(1e-10,1e-5)	(4.93e-8,2.01e-7)	7.14e-8
ρ	Diagnosis rate	U(1e-5,1)	(0.012,0.887)	0.107
γ	Recovery rate	Source [25]	Fixed value	0.1
I_0	Initial infections	U(1,8400)	(73.3,7192)	674
τ	$\sigma^2 = 1/\tau$ is the variance of data noise	U(1e-8,1e2)	(1.43e-5,4.33e-5)	2.62e-5

Table 4.1: Parameter table for SIR model

Parameter	Epidemiological Meaning	Prior	95% Credible Interval	Best-Fit Value
β	Transmission rate	U(1e-10,1e-5)	(8.20e-8,1.26e-7)	8.68e-8
ρ	Diagnosis rate	U(1e-5,1)	(0.016,0.024)	0.118
γ	Recovery rate	Source [25]	Fixed value	0.1
ϵ	Transfer rate	U(0.07,1)	(0.263,0.78)	0.631
E_0	Initial latent size	U(1,1700)	(3444,4682)	1523
I_0	Initial infections	U(1,8400)	(73.3,7192)	674
τ	$\sigma^2 = 1/\tau$ is the variance of data noise	U(1e-8,1e2)	(1.43e-5,4.13e-5)	2.61e-5

Table 4.2: Parameter table for SEIR model

Tables (4.1) and (4.2) show the model calibration results for both SIR and SEIR models. It is assumed that all parameters are uniformly distributed over the prior ranges. The uniform prior range (1e-10,1e-5) chosen for transmission rate β is given a broad range based off of general mathematical models in epidemiology. Since only a fraction of the infected population is being detected, the uniform prior range chosen for ρ is (1e-5,1). From the source [25], the recovery rate is fixed at 0.1. The uniform prior range for I_0 is given a wide range of (1,8400) due to the uncertainty of knowing the initial infected proportion of people. The uniform prior range of the variance in the observed case data is chosen to be between 0.01 to 1e8. The best-fit values as well as the 95% credible intervals for all parameters are derived from the desired posterior distribution through the affine invariant MCMC sampling algorithm.

4.1.2 Unnormalized posterior distribution

For our studies, since the prior distribution is a product of uniform distributions, we have:

$$P(\theta \mid D) \propto L(\theta \mid D)P(\theta) = CL(\theta \mid D) = \pi(\theta \mid D).$$

Thus, the posterior distribution is proportional to any constant multiplied to the likelihood distribution on the constrained parameter space. We choose the simplest constant, $C = 1$. Hence, the unnormalized posterior distribution is chosen to be equal to the likelihood distribution in these studies.

4.1.3 Model selections on SIR and SEIR

The data used for both Models (2.1) and (2.2) is the newly confirmed case data in Wuhan. We pick the uniform prior ranges for all parameters in SIR and SEIR models, and the same prior ranges are used for the common parameters shared in both models. We then calibrate SIR and SEIR separately using the same set of data, and obtain the best-fit parameters for each model. In Bayesian inference based calibration, the unnormalized posterior distribution is proportional to the likelihood function if uniform prior distributions are used for each parameter. Then the maximum likelihood values and the corrected Akaike Information Criterion at the best-fit parameters can be calculated for both models. AIC_c for SIR is 174, and AIC_c for SEIR is 186. Since the difference in the AIC_c scores is significant enough, we can conclude that the SIR model better explains the variance in the data with less number of parameters than the SEIR model. Therefore, further analysis in this study and the modeling for Alberta in the second study will be carried out using the SIR framework.

4.1.4 Nonidentifiability and its visualization

In this work, we will focus on the following definition of nonidentifiability [30]:

Definition 4.1 *Parameters θ_i, $i = 1, 2, ..., s$, of model* (2.5) *are called locally structurally identifiable at the MAP estimator θ^* if, there exists a neighborhood $V(\theta^*)$, for all values $\theta' \in V(\theta^*)$ that satisfy system constraints,*

$$y\left(\boldsymbol{x}(\theta^*, t), t\right) = y\left(\boldsymbol{x}(\theta', t), t\right)$$

implies

$$\theta_i^* = \theta_i', i = 1, 2, ..., s.$$

Parameters θ_i, $i = 1, 2, ..., s$ are non-identifiable if they are not identifiable.

Nonidentifiability in epidemic model implies that we may obtain the same best fits to the confirmed case data with drastically different estimations of infected proportions. Mitigating the impact of nonidentifiability issue is essential for estimating the true scale of the epidemic and for further model predictions of the potential, upcoming waves.

To illustrate that when nonidentifiability occurs, model prediction can be unreliable, we choose two different pairs of (β, ρ) values in model (2.1): (β_1, ρ_1) = (2.1e-7, 0.909), and (β_2, ρ_2) = (7.3e-8,0.102). Figure (4.1) shows the same goodness of fits (with the same likelihood value 7e-36) obtained by using these two parameter pairs, and Figure (4.2) shows the corresponding model predictions under these two sets of parameters. The significant difference in the peak number of cases and the scale of the epidemic shows a very different model prediction.

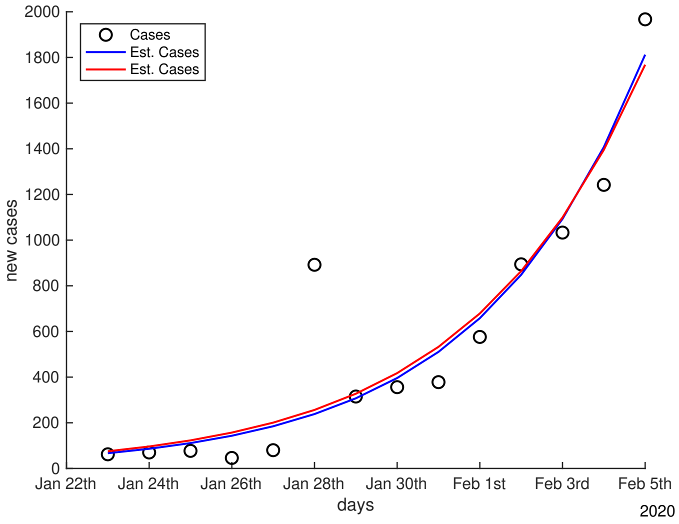

Figure 4.1: Two β-ρ parameter sets give equally best fits

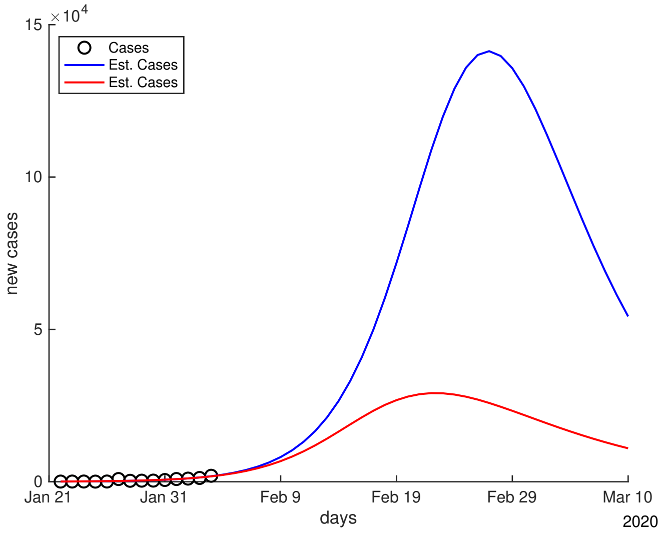

Figure 4.2: Two β-ρ parameter sets give different model projections

Nonidentifiability issue can be visualized in two ways. One is through checking the posterior distribution from different parameter perspectives, and another is by examining the maximum likelihood surface on the parameter space.

Posterior distribution check. In this study, there are 5 parameters being fit from 14 observations. Each i^{th} sample from the MCMC output is a 5-dimensional parameter vector,

$$\boldsymbol{\theta}(i) = (\beta^i, \rho^i, \gamma^i, I_0^i, \tau^i)$$

with a corresponding unnormalized posterior density value π_i. To view the posterior surface from the β perspective, we plot the points (β^i,π_i) on a 2-dimensional graph, Figure (4.3). This is the silhouette of the unnormalized posterior density viewed from the β perspective. The posterior surface from the perspective of another parameter ρ is viewed in the same fashion in Figure (4.4).

Viewing the unnormalized posterior density from different parameter perspectives provides a convenient way to diagnose nonidentifiability. Flat distribution for both parameters, β and ρ, indicates that among all the posterior samples we have, the model is not able to find the optimal value of these two parameters for the best fit solution or there are infinite many optimal values of the two parameters to achieve equally best fitting performance.

Maximum likelihood surface check. We next show the projection surface of the likelihood function $L(\theta|D)$ onto β-ρ parameter plane. We form the 3-dimensional vector (β^i, ρ^i, π_i) from the MCMC samples, and plot the 3-

dimensional graph by using the MATLAB command *plot3* which creates both Figure (4.5) and Figure (4.6). Figure (4.5) is obtained using the MATLAB tool *Rotate 3D* in the figure window. The colors on the graphs are filled by MATLAB command *patch* which orders the colors from blue to yellow with increasing likelihood values. In Figure (4.5), the maximum likelihood value follows the yellow curve rather than concentrating on a unique point. This indicates that the best-fit values for parameters are not unique and any pair of (β, ρ) along this curve will cause the model to equally likely achieve the best fits. In Figure (4.6), we show the 3D projection of the likelihood surface onto the β-ρ parameter plane. More clearly, it shows that the largest probability are concentrated along a flat strip rather than on a single point.

To reduce the impact of this nonidentifiabiity issue, one option is to find more independent data that can help model calibration. Applying the model selection to determine the simplest model structure and using a more advanced sampling algorithm can also help to ensure reasonable posterior distributions.

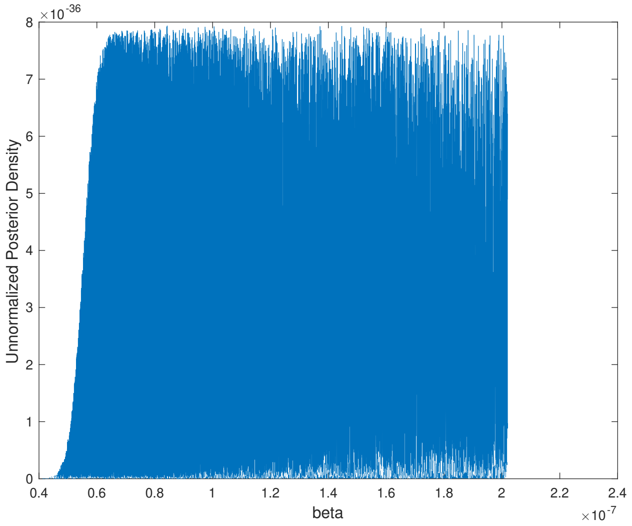

Figure 4.3: Unnormalized posterior density from SIR model parameter β perspective [1]

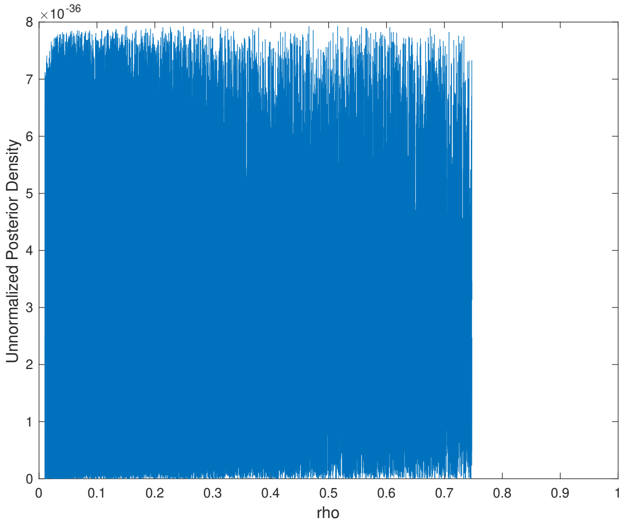

Figure 4.4: Unnormalized posterior density from SIR model parameter ρ perspective [1]

[1] 8 chains with a total number of burn-in samples of 600,000 and a total number of actual samples of 1,400,000 are used for the fitting.

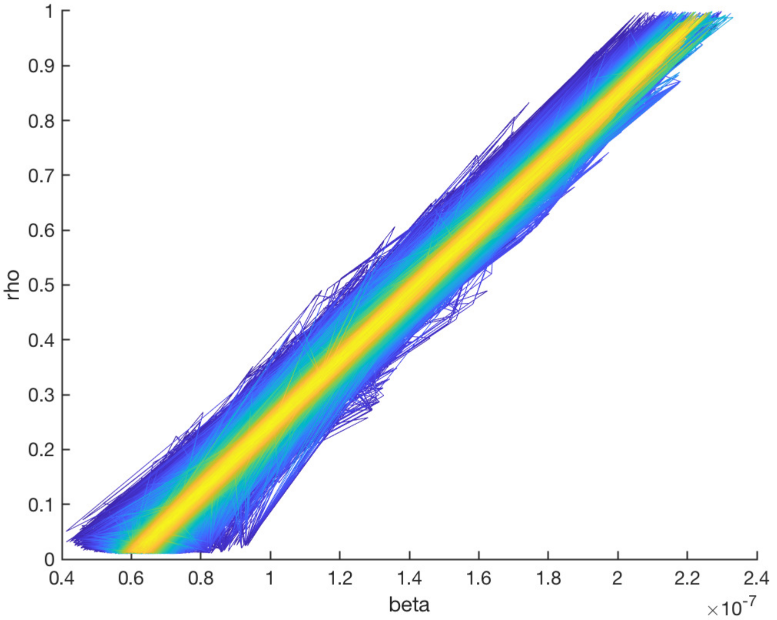

Figure 4.5: Likelihood over β-ρ parameter plane

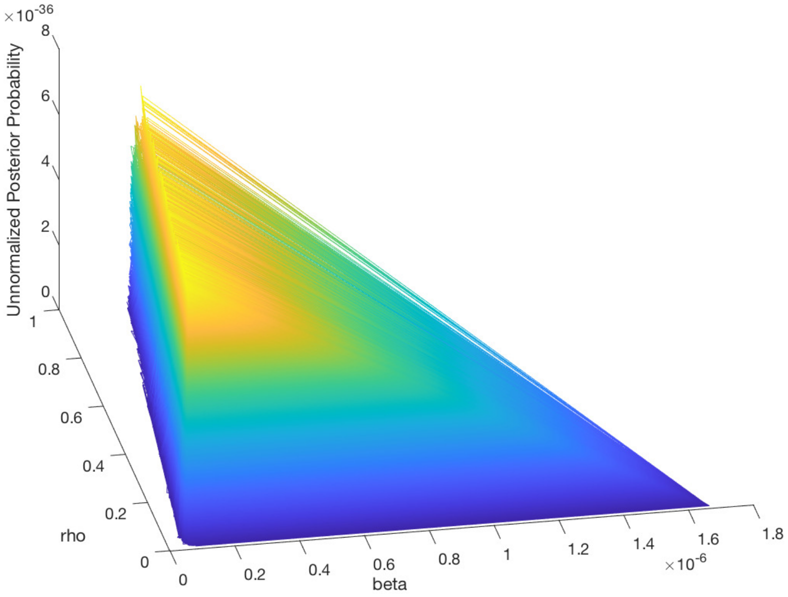

Figure 4.6: Likelihood over β-ρ parameter plane (3D)

4.2 Retrospective estimation of proportion of total infections of COVID-19 during the first wave in Alberta

Accurate estimation of proportion of infected population of COVID-19 has always been a major concern from public health agencies as this information informs public health policies in timely efficiency to shorten the time course of the epidemic as well as reduce the peak case number. It is usually difficult to derive the accurate estimation through modeling. One main reason is the lack of reliable data and missing information for the hidden infections, including those are asymptomatic or those are not tested but already infected. Our study in this section use both confirmed case data and testing data to better help model fitting. We also propose a method that mitigate the non-identifiability issue so that a more accurate estimation of infected proportion can be obtained. Since case data can be highly impacted by public testing policies and intervention policies during the pandemic, the method constructs a time dependent case-infection ratio $\rho(t)$ informed by public health data on population health seeking behavior, and a time-dependent transmission rate $\beta(t)$ informed by certain lockdown measurements from public health system to track the real-time infections accurately. With such method implemented, our estimation result agrees with the existing antibody test results from the official seroprevalence surveys and reports.

4.2.1 Model fitting results and Bayesian goodness of fit

Parameter	Description	Best-fit value	95% credible Interval	Prior
$beta$	Initial transmission rate	6.195e-08	(5.63e-08,7.84e-08)	U(1e-9,1e-7)
γ_1, γ_2	Recovery rate	0.1429	(0.1260,0.1920)	Source [25]
$\rho(t)$	Case-infection ratio	Time-varying	Time-varying	Time-varying
f	Probability for health-seeking behavior	1.1013	(0.8146,1.4857)	U(0,1.5)
q	q·b is the new transmission rate during lockdown	0.6548	(0.5751,0.8163)	U(0.01,1)
s	s·b is the new transmission rate during open up	0.6864	(0.6054,0.9666)	U(0.01,1)
p	$\sigma_1^2 = 1/p$ is the variance of the case data noise	8.7933	(6.5232, 14.5960)	U(1,79)
p_2	$\sigma_2^2 = 1/p_2$ is the variance of the testing data noise	0.0042	0.0031, 0.0062)	U(0.0001,0.03)
$time_1$	Policy reaction time in testing	3.2	(1.1168,4.6156)	U(1,10)
$time_{1b}$	Same as above	36.9	(35.1107,36.9854)	U(10,37)
$time_2$	Same as above	46.2	(44.1507,49.0937)	U(39,55)
$time_3$	Same as above	52.0	(52.0024,54.4615)	U(52,60)
$time_4$	Same as above	57.2	(57.0013,58.6041)	U(57,72)
$height\_val$	Testing number turning point value	0.0895	(0.0795,0.0989)	U(0.001,0.1)
$const\_val$	Same as above	0.0375	(0.0317,0.0431)	U(0.001,0.1)
$m0\_val$	Same as above	0.0223	(0.0185,0.0268	U(0.01,0.04)
$time1beta$	Policy reaction time in transmission	29.2070	(29.0017, 31.7245)	U(29,32)
$time2beta$	Same as above	66.5656	(65.1864,67.9948)	U(65,68)

Table 4.3: Parameter table for SICR model

The census data is used to calibrate the total population $N(t) = S(t) + I(t) + C(t) + R(t)$ during the period of modeling. Timelines of implementations of public-health measures are used to define the time-dependent transmission rate $\beta(t)$ as in Section (2.1.2). Newly reported positive COVID-19 cases and the total number of Healthline calls and COVID-19 online self-assessment forms are used to produce the data for $\rho(t)$ according to the definition in (2.3). Parameters are estimated by fitting model outcomes to daily reported new positive COVID-19 cases data and Table (4.3) shows the model calibration results for the SICR model. Since the Diffusive Nested Sampling algorithm performs better in higher dimensions in comparison to the Affine Invariant algorithm [[31], [32]], the Diffusive Nested Sampling algorithm is used to fit the Alberta COVID-19 first wave mathematical model [33]. Same as before, all parameters are assumed uniformly distributed over the given prior ranges, and the best-fit values as well as the 95% credible intervals for all parameters are derived from the desired posterior distributions.

Posteriors for all parameters are included in Appendix (A). Calibrated $\beta(t)$ and $\rho(t)$ are shown in Figures (4.7) and (4.8). Model fitting to the daily new positive case report data is shown in Figure (4.9).

In Figures (4.10) and (4.11), the predictive estimations with intervals for the infectious population and susceptible population are shown. Based on these projections, with measures undertaken in Alberta during the first wave, starting with 4,371,000 susceptibles, there were an estimated 4,340,000 (95% CI: 4,325,000-4,350,000) total susceptibles after the first wave. The most likely peak time for infections occurred on April 15, 2020, with an estimated 3,585 new infections per day fluctuating between 2,716 and 4,979 new infections. We evaluate our fitting by going through every posterior predictive solution and test the optimal fitting values of the chi-squared discrepancy among all possible values that could have been realized under this model with the same set of parameters that generate the current data. The posterior predictive p-value in our study gives 0.5412, which indicates the satisfying goodness of fitting.

Validation of results using seroprevalence data. Estimating precisely the proportion of total infected, both hidden and diagnosed, is important for local government to be conscious of the current situation and to guide effective policy implementations. It is also difficult to achieve the accurate result due to the delay of the release of data and the absence of seroprevalence reports and surveys. We estimated the proportion of infected using the integral of incidence, and Figure (4.12) presents the estimated proportion of infected over the first wave period. Consistent with provincial seroprevalence data, we estimated that 3.0655e4 (95% CI: 2.1631e4-4.6084e4) people were infected by COVID-19 at the end of the first wave, that is 0.72% (95% CI: 0.476%–1.014%) of the total population. This result conforms to the crude SARS-CoV-2 IgG seropositivity rate of 0.92% (95% CI: 0.72–1.13%) shortly after the first COVID-19 wave in June 2020. [14].

By the solid validation from an independent seroprevalence data, our model ability is demonstrated. Our model is able to provide a timewise proportion of infected population throughout the wave while serosurvey are usually done only once or twice per wave. Additionally, the parameter values gained during the training of the model are considered trustworthy and can be further used as the baseline in later wave prediction.

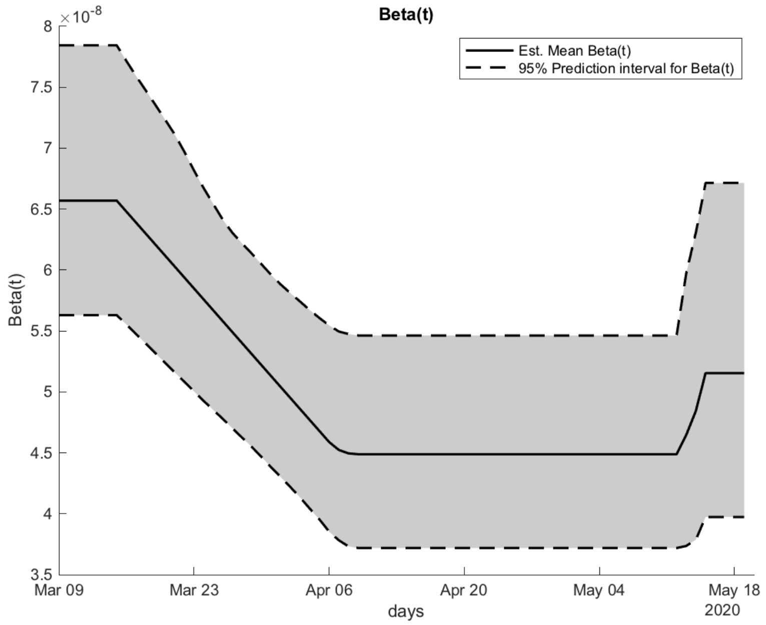

Figure 4.7: $\beta(t)$

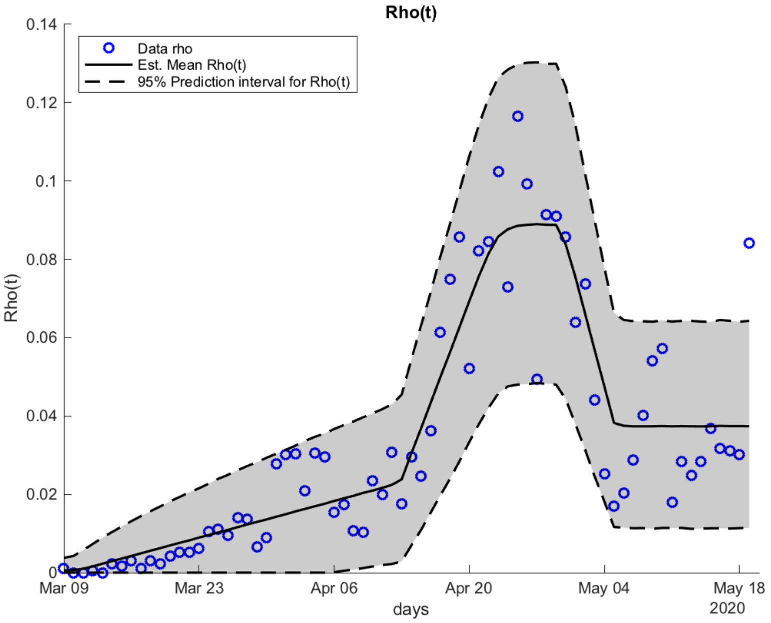

Figure 4.8: $\rho(t)$

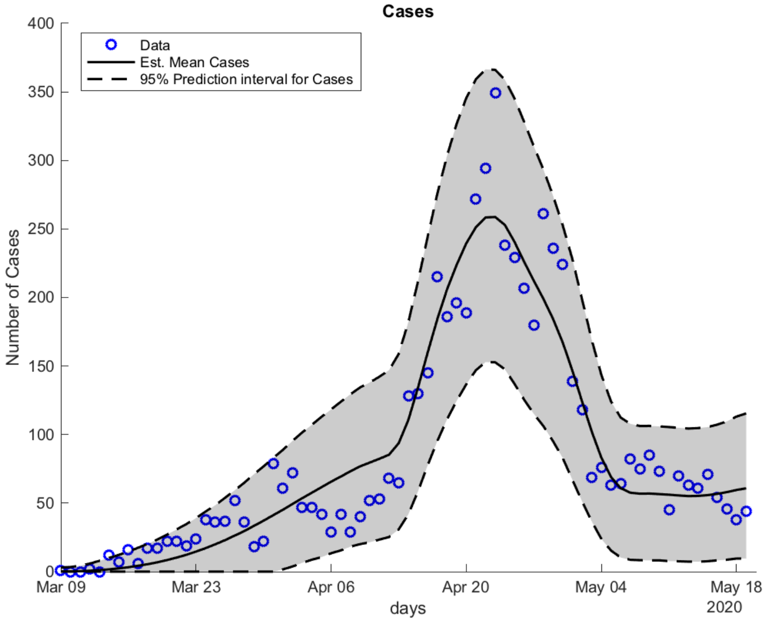

Figure 4.9: Fitting to new cases

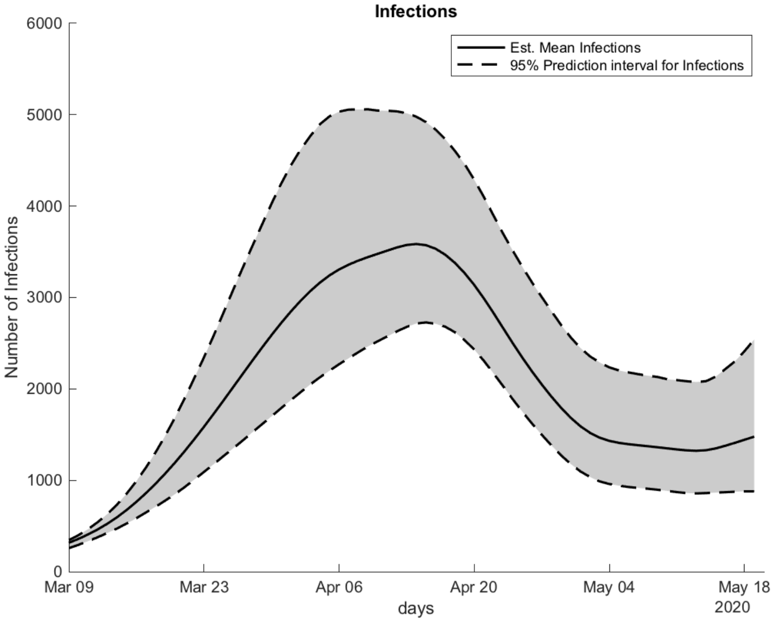

Figure 4.10: Estimated new infections

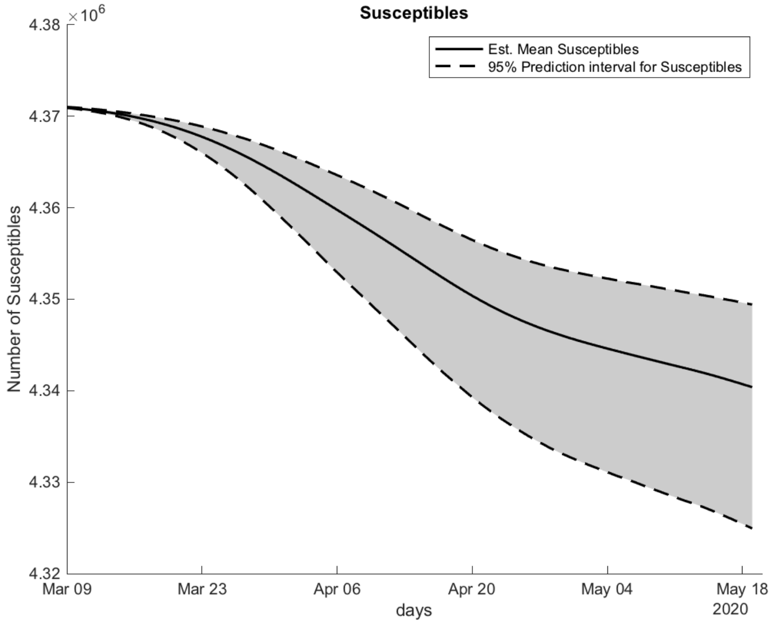

Figure 4.11: Estimated total susceptibles

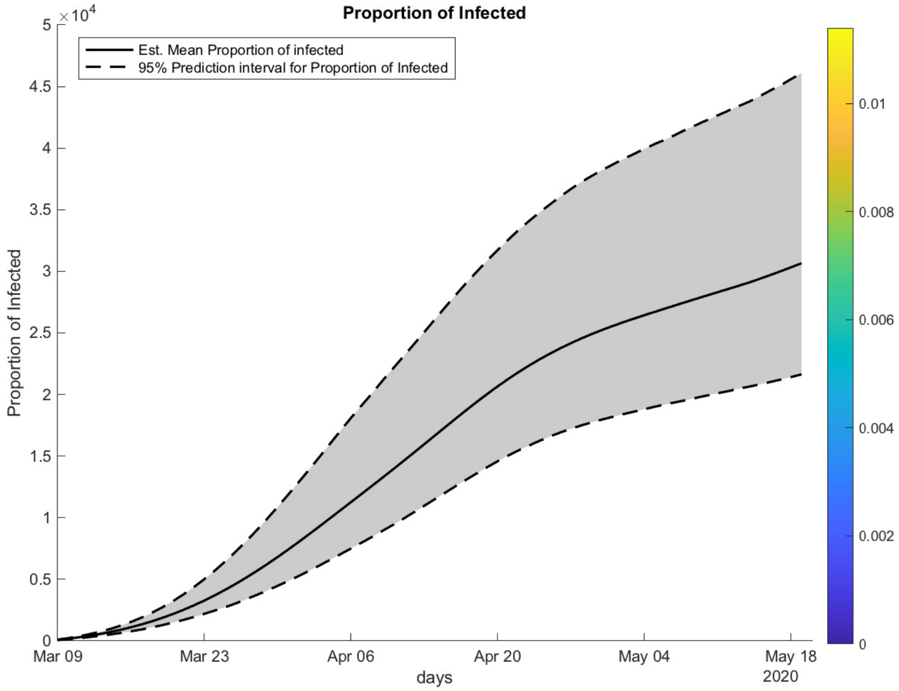

Figure 4.12: Estimated proportion of infected with color bar

4.2.2 Estimated infection-fatality ratio and case-fatality ratio

Infection-fatality ratio is an important epidemiological parameter which measures the risk of death per infection. Some studies use case-fatality ratio instead of infection-fatality ratio to understand the true scale of the epidemic. However, case-fatality ratio depends highly on how well cases of the disease are identified and recorded. If different testing policies are implemented, then different case numbers will be detected, leading to different case-fatality ratios. The advantage of using infection-fatality ratio is that it's independent of the testing policy, and is suitable for comparison among epidemics in different jurisdictions. A challenge in estimating the infection-fatality ratio is calculating its denominator, namely, the total number of infections, which is already estimated in Section (4.2.1).

We define the case-fatality ratio (CFR) and infection-fatality ratio (IFR) as follows:

Definition 4.2 *(Case-fatality ratio)*

$$\text{CFR}(t) = \frac{\text{total number of deaths up to time t}}{\text{total number of confirmed cases up to time t}}.$$

Definition 4.3 *(Infection-fatality ratio)*

$$\text{IFR}(t) = \frac{\text{total number of deaths up to time t}}{\text{total number of infections up to time t}}.$$

Here we adopt a method to precisely estimate the reported number of daily new death each day. Since the actual data of daily new death are counting numbers of small magnitude with considerable uncertainty, we assume those numbers follow Poisson distributions with different mean values each day. In order to figure out all those mean values that will be used to generate estimated death, a Gaussian mixture model is introduced to fit the 7-day-average daily new death. Observing the natural shape of the averaged death data, we choose to fit three Gaussian profiles to efficiently catch the trends in death data with its three peaks. The Gaussian model fits peaks and is given by,

$$\widehat{Death}(t) = \sum_{i=1}^{3} a_i e^{\left[-\left(\frac{t-b_i}{c_i}\right)^2\right]},$$

where a is the amplitude, b is the centroid (location), c is related to the peak width. With the help of the Gaussian library model in MATLAB, we specify the model type of $gauss3$ and fit the three-profile Gaussian model to the averaged death data. Then this 3-Gaussian mixed curve is utilized as our desired mean values for death numbers each day. A reasonable estimation of the cumulative number of reported death can be generated for each posterior sample in our model, and Figure (4.13) shows the fitting for death data.

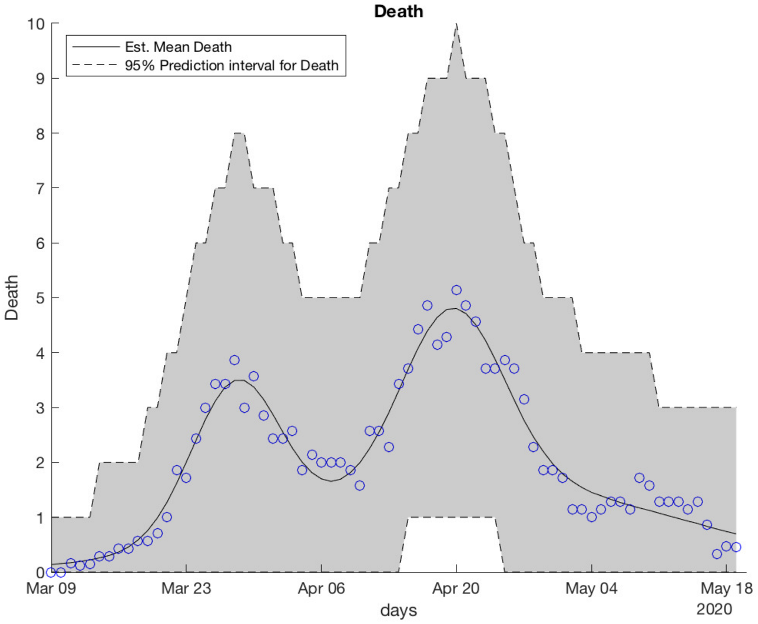

Figure 4.13: Estimated 7-day average death

Based on our generated death samples, we are able to compute the corresponding case-fatality ratio and infection-fatality ratio. Our result shows that as of May 15, 2020, the estimated overall case-fatality ratio was 2.494% with the 95% credible interval (2.024%-3.024%), which coincided with the actual case-fatality ratio calculated from the dataset. Meanwhile, as of May 15, 2020, our estimated overall infection-fatality ratio stayed stable at 0.513% with a 95% credible interval (0.315%-0.732%). Tables (4.4) and (4.5) describe pre-vaccine CFR and IFR estimates correspondingly on certain dates. Figures (4.14) and (4.15) show the pre-vaccine CFR and IFR estimates during the first wave.

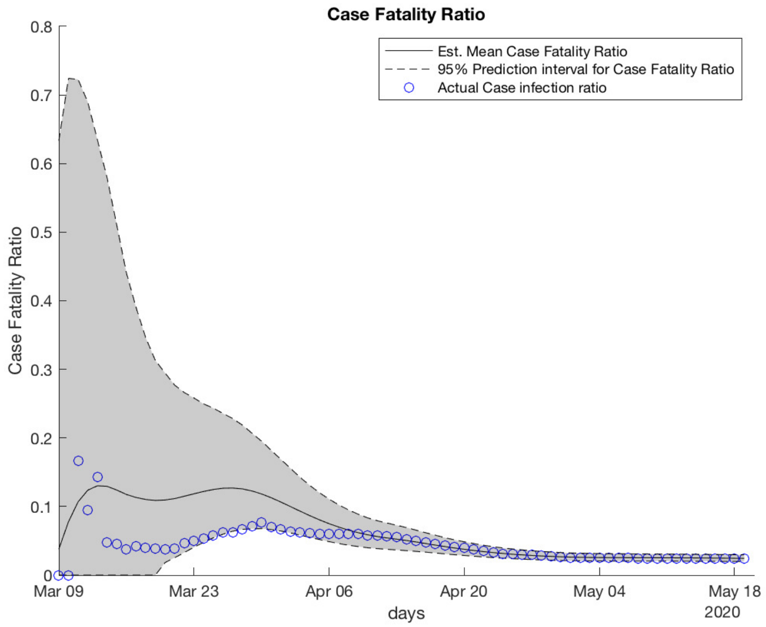

Figure 4.14: Cumulative CFR

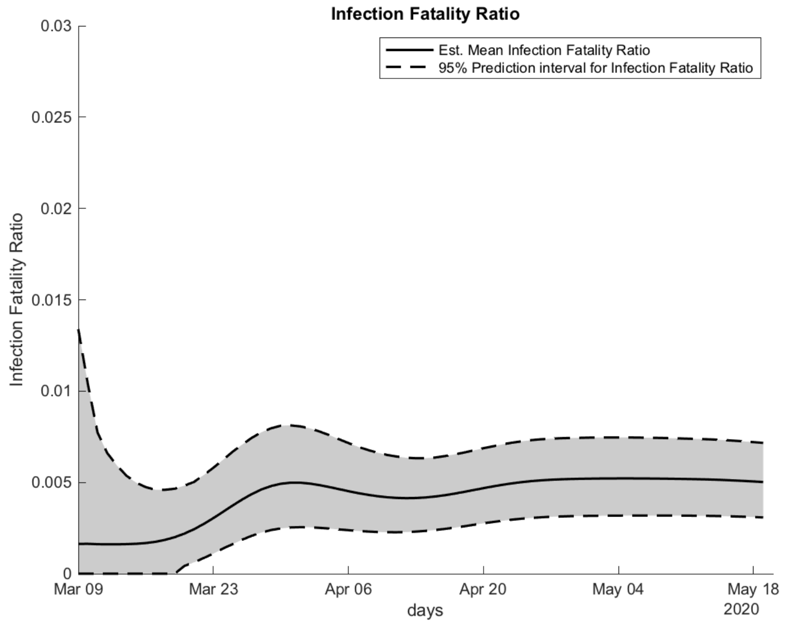

Figure 4.15: Cumulative IFR

	CFR, March 31, 2020	CFR, April 15, 2020	CFR, May 15, 2020
Alberta	11.287% (6.633-18.205)	4.836% (3.463-6.563)	2.494% (2.024-3.024)

Table 4.4: COVID-19 CFR estimates during March, April, and May in 2020

	IFR, March 31, 2020	IFR, April 15, 2020	IFR, May 15, 2020
Alberta	0.498% (0.253-0.812)	0.423% (0.239-0.636)	0.513% (0.315-0.732)

Table 4.5: COVID-19 IFR estimates during March, April, and May in 2020

4.2.3 Sensitivity analysis

Standard sensitivity analysis methods are designed for model parameters that are constant, and are not applicable to models with time-dependent parameters, such as $\beta(t)$ and $\rho(t)$ in Model (2.4). Since $\beta(t)$ and $\rho(t)$ are piece-wisely defined, as shown in Figures (3.1) and (3.2), global changes in $\beta(t)$ and $\rho(t)$ are influenced by variation of parameters in the definitions. As these parameters are estimated through the model fitting, we conducted the sensitivity analysis for these parameters in the definition of $\beta(t)$ and $\rho(t)$ using PRCC, as an indirect way of assessing the sensitivity of $\beta(t)$ and $\rho(t)$.

We conduct the sensitivity analysis using PRCC to four main public health measurement outputs at the end of the epidemic: case-infection ratio, cumulative infection-fatality ratio, proportion of infected population, and peak time of cases. In Figures (4.16), (4.17), (4.18), and (4.19), the PRCC values for each of the four indicators with respect to every model parameter are listed. In order to better reflect which parameter causes the major effects on these outcomes during the epidemic, we further apportion all the model parameters into three groups: β-related risk reduction behavior, ρ-related health-seeking behavior, and other common factors including initial infection size and recovery rate.

In Figure (4.16), for case-infection ratio, we conclude that the most influential parameter is the health-seeking factor f, which means strengthening the human action of case detection can most significantly enhance the capability of unearthing the true case among the hidden infections. A negative correlation occurs between the baseline transmission rate β and the case-infection ratio. As we reduce the initial transmission rate β, the infected population decreases, and in turn the overall case-infection ratio increases. The same condition holds for the factor initial infectious population I_0. For the cumulative infection-fatality ratio, very similar results have been obtained. The health-seeking factor f and baseline transmission rate β are now affecting infection-fatality ratio at nearly the same level, but one causes a positive effect while the other causes a negative effect, as shown in Figure (4.17).

On the subject of peak time of cases and the proportion of the infected population, the situation can be different. In Figure (4.18), for peak time, the only parameter that is considered to be sensitive is the timepoint $time2$, which is the estimated time-point that the testing numbers start to increase. This is quite realistic as the case number is highly related to the case detection rate, and high testing number yields high case detection number. In Figure (4.19), for the proportion of the infected population, the most influential parameters are still β and f. This indicates that either increasing the case detection rate or reducing the transmission rate can reduce the proportion of infected.

In conclusion, risk-reduction behavior together with health-seeking behavior are the two main focuses we need to pay attention to and they are actually controllable by either government or people themselves throughout the whole epidemic. Public measurements should be implemented speedily and effectively to ensure the initial transmission rate is small enough to avoid rapid outbreaks. At the same time, health agencies may try as hard as possible to improve the case-detection capability such as increasing testing numbers to keep cases from being hidden out of control at the begining of the epidemic.

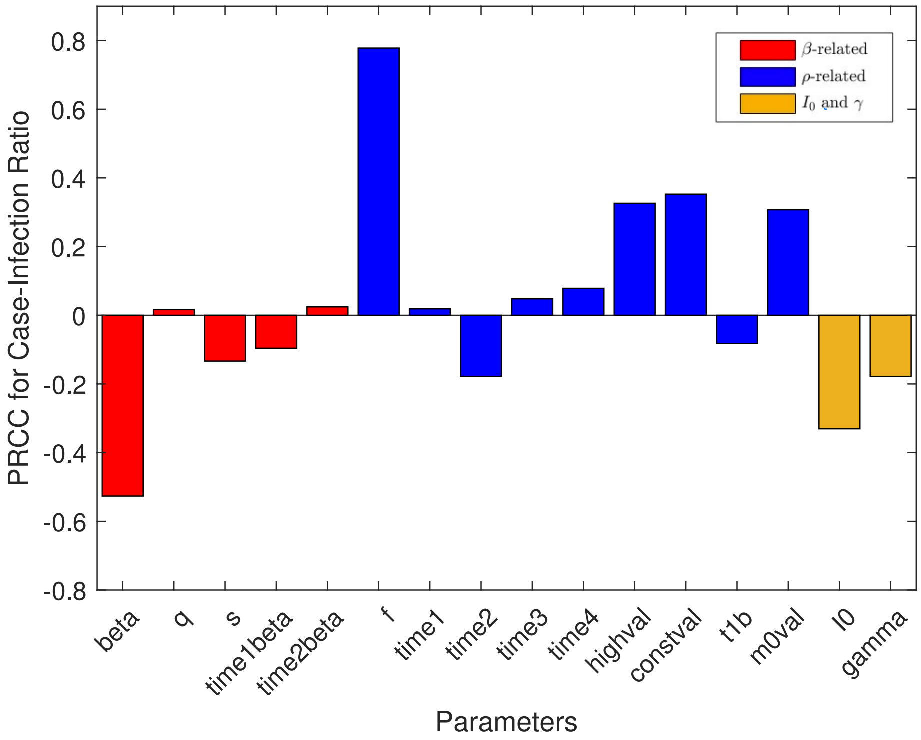

Figure 4.16: Sensitivity of case-infection ratio

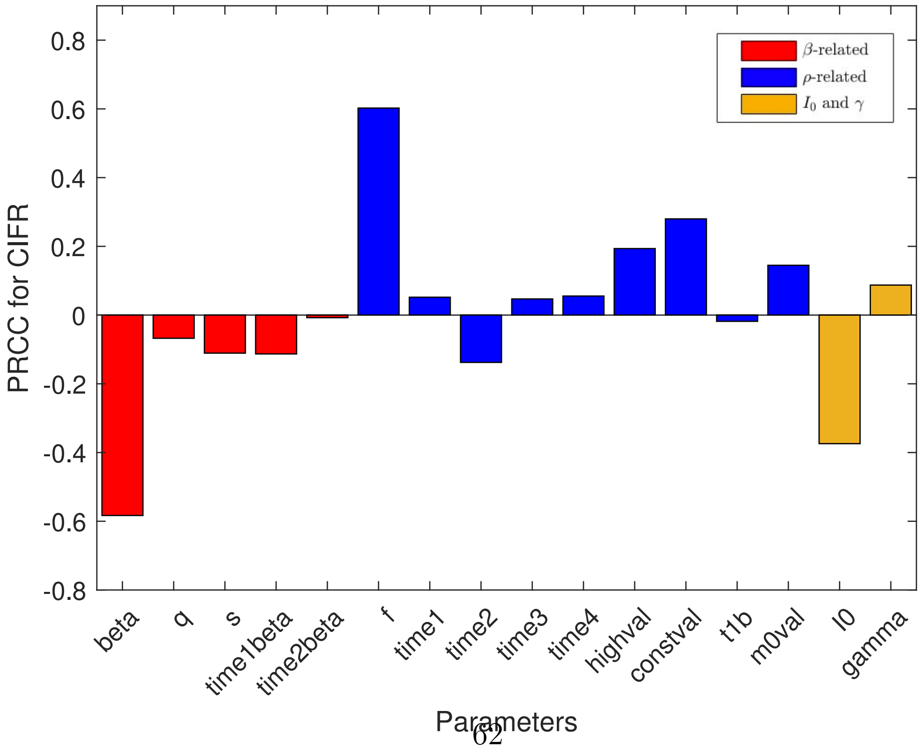

Figure 4.17: Sensitivity of infection-fatality ratio

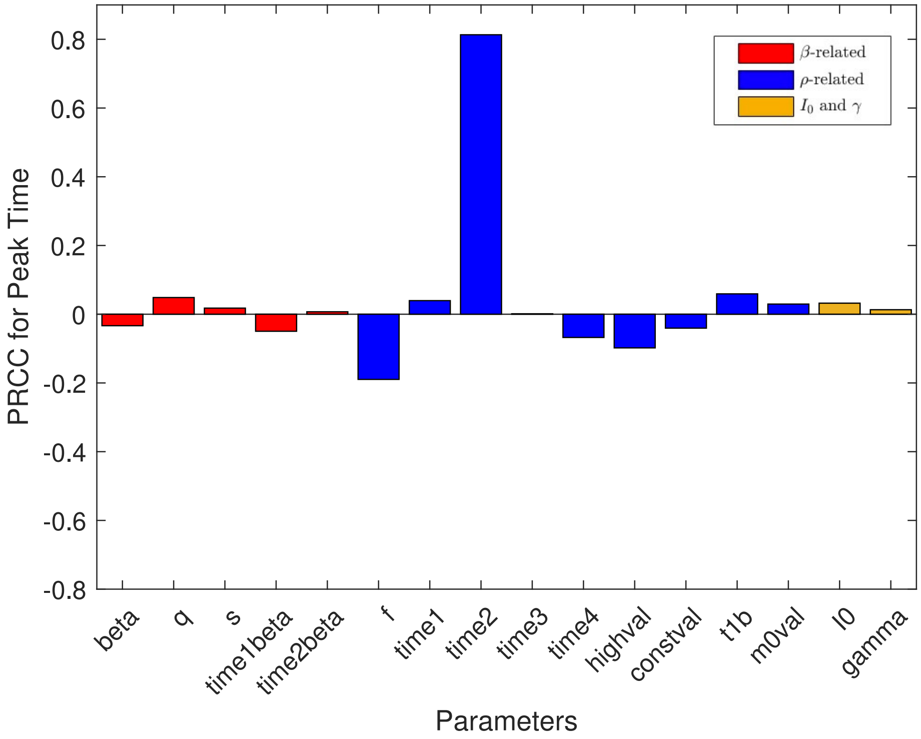

Figure 4.18: Sensitivity of peak time

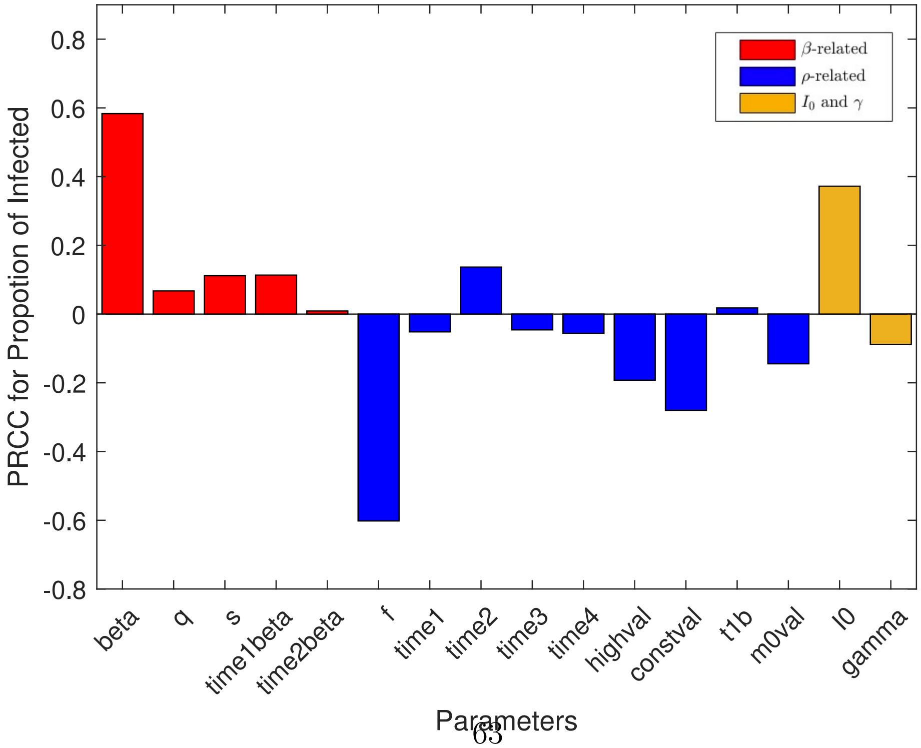

Figure 4.19: Sensitivity of proportion of infected

Chapter 5

Conclusions and future work

The COVID-19 epidemic is an unprecedented global public health challenge in recent years, bringing a large impact on human lives, government policy planning and public health systems, as well as global economic growth. Mathematical modeling can help project how infectious diseases progress to show the scale and time course of the epidemics, and help inform public health policies. The focus of our first study is to demonstrate the challenges modelers are facing in predicting outbreaks like this and to provide a partial explanation for the wide variability in earlier model predictions of COVID-19 [4]. Based on the model framework chosen from the first study, our second study follows the same pattern with more reliable data to estimate the true scale of the first wave of COVID-19 in Alberta with in-depth analysis.

Our first study focused on the COVID-19 epidemic in Wuhan city after the lockdown and quarantine. By comparing the SIR and SEIR frameworks using the Akaike Information Criterion, we concluded that a more complicated model may not necessarily perform better in model prediction since there are more parameters needed to be estimated during the fitting process. Using a simple SIR model and the daily new confirmed case data for model calibration, we found the linkage between the transmission rate β and the diagnosis rate ρ which resulted in the nonidentifiability issue. The nonidentifiability yields significantly different model predictions under infinite combinations of best-fit parameter values, and is caused by the lack of independent data which allows producing independent estimates of β and ρ. Modelers should be cautious about how to reduce the nonidentifiability impact during model calibration, and be realistic about the selection of parameter ranges and values.

Our second study focused on the COVID-19 epidemic during the first wave in Alberta. As discussed in our first study, we chose the modified version of the SIR, SICR model, and used two datasets, one is the daily new confirmed case data, and another is the daily testing data from Alberta Health for model calibration. To incorporate the real-time policy changing, we used the time-dependent transmission rate $\beta(t)$ and the time-dependent diagnosis rate $\rho(t)$. With an extra independent dataset and more specific parameter construction, the nonidentifiability has been greatly reduced. We also demonstrated that with Bayesian inference and an improved MCMC algorithm, the affine ensemble MCMC algorithm, the model can significantly narrow down the given uniform prior ranges, and produce sufficiently small credible intervals for parameters in the model. We further estimated the timewise proportion of the infected population among all populations in Alberta during the first wave, and computed the real-time case-fatality ratio and infection-fatality ratio, respectively. These numbers have been validated by official seroprevalence reports and can provide a solid understanding of the true scale of the epidemic. We further provided the PRCC values for all parameters to several public health outcomes to study the sensitivity of each parameter. We demonstrated that the peak time of the epidemic is much less sensitive to parameter variations than the proportion of the infected population, infection-fatality ratio, and case-infection ratio. This was also observed in our first study on predicting the Wuhan outbreak using two different β-ρ parameter pairs. We showed that there is a positive relationship between the initial transmission rate β and the proportion of the infected population. When more restrictive measures are implemented at the beginning of the epidemic, including the lockdown of residential buildings and the quarantine of suspected cases and their close contacts, there will be less proportion of infectious people at the end of the first wave. We also showed the negative relationship between case-infection detection power f and the proportion of the infected population. That means when the testing capacity is lifted and more human power is included to facilitate the testing process during the epidemic, there will also be less proportion of infectious people at the end of the first wave. These findings provide a theoretical verification of the effectiveness of these measures.

COVID-19 is still a threat to our lives as no one can predict when a new strain might surface. There are still many questions remaining and modelers are still working hard to investigate more in-depth. There are many variants of concerns have been identified and these are now circulating in Alberta. With currently available data and knowledge we have for COVID-19, further studies are warranted to model a coexisting-strain epidemic. This will be an intricate mathematical network of channels incorporated with different biological information and vaccination effectiveness information. Modeling allows a more accurate estimation of the virus which will serve as an important signal to the public health authorities. In the development of the mathematical model and fitting process, it's essential to seek and develop methods to deal with nonidentifability issue. More efficient sampling algorithms are worth to be tried and be improved to ensure an optimal fitting result is achieved with high-quality samples in a sufficiently shorter time.

Bibliography

[1] https://covid19.who.int/.

[2] https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19—26-october-2022, accessed on 2 November 2022.

[3] https://www.nature.com/articles/d41586-020-00361-5, accessed on 18 February 2020.

[4] Roda, W. C., Varughese, M. B., Han, D., Li, M. Y. (2020). Why is it difficult to accurately predict the COVID-19 epidemic?. *Infectious Disease Modelling*, 5, 271–281. https://doi.org/10.1016/j.idm.2020.03.001.

[5] Chowell, G., Castillo-Chavez, C., Fenimore, P. W., Kribs-Zaleta, C. M., Arriola, L., Hyman, J. M. (2004). Model parameters and outbreak control for SARS. *Emerging Infectious Diseases*, 10(7), 1258–1263. https://doi.org/10.3201/eid1007.030647.

[6] Gumel, A. B., Ruan, S., Day, T., Watmough, J., Brauer, F., van den Driessche, P., Gabrielson, D., Bowman, C., Alexander, M. E., Ardal, S., Wu, J., Sahai, B. M. (2004). Modelling strategies for controlling

SARS outbreaks. *Proceedings of the Royal Society B: Biological sciences*, 271(1554), 2223–2232. https://doi.org/10.1098/rspb.2004.2800.

[7] Lipsitch, M., Cohen, T., Cooper, B., Robins, J. M., Ma, S., James, L., Gopalakrishna, G., Chew, S. K., Tan, C. C., Samore, M. H., Fisman, D., Murray, M. (2003). Transmission dynamics and control of severe acute respiratory syndrome. *Science*, 300(5627), 1966–1970. https://doi.org/10.1126/science.1086616.

[8] Zhang, J., Lou, J., Ma, Z., Wu, J. (2005). A compartmental model for the analysis of SARS transmission patterns and outbreak control measures in China. *Applied Mathematics and Computation*, 162(2), 909–924. https://doi.org/10.1016/j.amc.2003.12.131.

[9] https://www.cdc.gov/flu/weekly/index.htm, accessed on 18 February 2020.

[10] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19-similarities-and-differences-with-influenza, accessed on 2 November 2022.

[11] https://www.covid19immunitytaskforce.ca/seroprevalence-in-canada/, accessed on 2 December 2022.

[12] Skowronski, D. M., Kaweski, S. E., Irvine, M. A., Kim, S., Chuang, E. S. Y., Sabaiduc, S., Fraser, M., Reyes, R. C., Henry, B., Levett, P. N., Petric, M., Krajden, M., Sekirov, I. (2022). Serial cross-sectional estimation of vaccine-and infection-induced SARS-CoV-2 seroprevalence in

British Columbia, Canada. *CMAJ : Canadian Medical Association Journal*, 194(47), E1599–E1609. https://doi.org/10.1503/cmaj.221335.

[13] https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/burden.htmlest-infections, accessed on 2 November 2022.

[14] Charlton, C. L., Nguyen, L. T., Bailey, A., Fenton, J., Plitt, S. S., Marohn, C., Lau, C., Hinshaw, D., Lutsiak, C., Simmonds, K., Kanji, J. N., Zelyas, N., Lee, N., Mengel, M., Tipples, G. (2021). Pre-Vaccine Positivity of SARS-CoV-2 Antibodies in Alberta, Canada during the First Two Waves of the COVID-19 Pandemic. *Microbiology Spectrum*, 9(1), e0029121. https://doi.org/10.1128/Spectrum.00291-21.

[15] https://www.alberta.ca/assets/documents/health-data-modelling-fact-sheet.pdf, accessed on 1 October 2021.

[16] https://everythinggp.com/2020/03/17/province-of-alberta-announces-state-of-public-emergency-amid-covid-19-outbreak/, accessed on 18 March 2020.

[17] http://en.nhc.gov.cn/.

[18] Kalb"eisch, J. (1979). *Probability and Statistical Inference*, vol. Volume 2: Statistical Inference. New York: Springer-Verlag.

[19] Roda, W. C. (2020). Bayesian inference for dynamical systems. *Infectious Disease Modelling*, 5. https://doi.org/10.1016/j.idm.2019.12.007.

[20] Chen, M., Shao, Q., Ibrahim, J. (2000). *Monte Carlo methods in bayesian computation*. New York: Springer-Verlag.

[21] Goodman, J., Weare, J. (2010). Ensemble samplers with affine invariance. *Communications in Applied Mathematics and Computational Science*, 5(1), 65-80. https://doi.org/10.2140/camcos.2010.5.65.

[22] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (pp. 267-281). Budapest: Akademiai Kiado.

[23] Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics - Theory and Methods*, 7, 13-26. https://doi.org/10.1080/03610927808827599.

[24] Marino, S., Hogue, I. B., Ray, C. J., Kirschner, D. E. (2008). A methodology for performing global uncertainty and sensitivity analysis in systems biology. *Journal of Theoretical Biology*, 254(1), 178–196. https://doi.org/10.1016/j.jtbi.2008.04.011.

[25] You, C., Deng, Y., Hu, W., Sun, J., Lin, Q., Zhou, F., Pang, C. H., Zhang, Y., Chen, Z., Zhou, X. H. (2020). Estimation of the time-varying reproduction number of COVID-19 outbreak in China. *International Journal of Hygiene and Environmental Health*, 228, 113555. https://doi.org/10.1016/j.ijheh.2020.113555.

[26] Li, M. Y. (2018). *An Introduction to Mathematical Modeling of Infectious Diseases* (Vol. 2). Cham: Springer.

[27] J. Guckenheimer, P. Holmes. (1983) Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields. *Applied Mathematical Sciences*. Vol. 42, Springer.

[28] Cooper, I., Mondal, A., Antonopoulos, C. G. (2020). A SIR model assumption for the spread of COVID-19 in different communities. *Chaos, Solitons, and Fractals*, 139, 110057. https://doi.org/10.1016/j.chaos.2020.110057.

[29] Stefan B., Rebecca C., Ruth G. P., Sunetra G., Sharmistha M., Martin K. (2021). Leveraging epidemiological principles to evaluate Sweden's COVID-19 response. *Annals of Epidemiology*, Volume 54, 21-26. https://doi.org/10.1016/j.annepidem.2020.11.005.

[30] Miao, H., Xia, X., Perelson, A. S., Wu, H. (2011). On identifiability of Nonlinear ODE models and applications in viral dynamics. *Society for Industrial and Applied Mathematics*, 53(1), 3–39. https://doi.org/10.1137/090757009.

[31] Huijser D, Goodman J, Brewer BJ (2022). Properties of the affine-invariant ensemble sampler's 'stretch move' in high dimensions. *Australian & New Zealand Journal of Statistics*, 64(1), 1-26. https://doi.org/10.1111/anzs.12358.

[32] Brewer BJ, Partay LB, Csanyi G (2011). Diffusive Nested Sampling. *Statistics and Computing*, 21(4), 649-656. https://doi.org/10.1007/s11222-010-9198-8.

[33] Roda, Weston Christopher, & Han, Donglin. (2021). MatlabDiffNestAlg: V1.1 (1.1). Zenodo. https://doi.org/10.5281/zenodo.5270613.

Appendices

A Posteriors for SICR Model

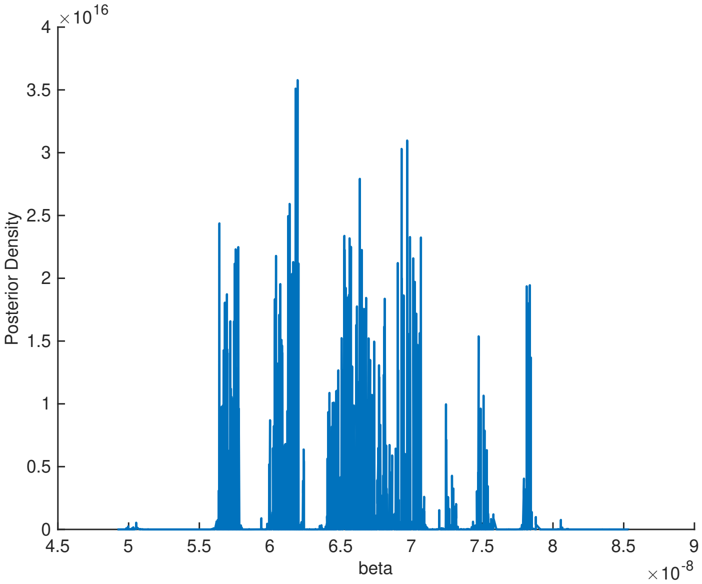

Figure 0.1: Posterior for parameter β for SICR model

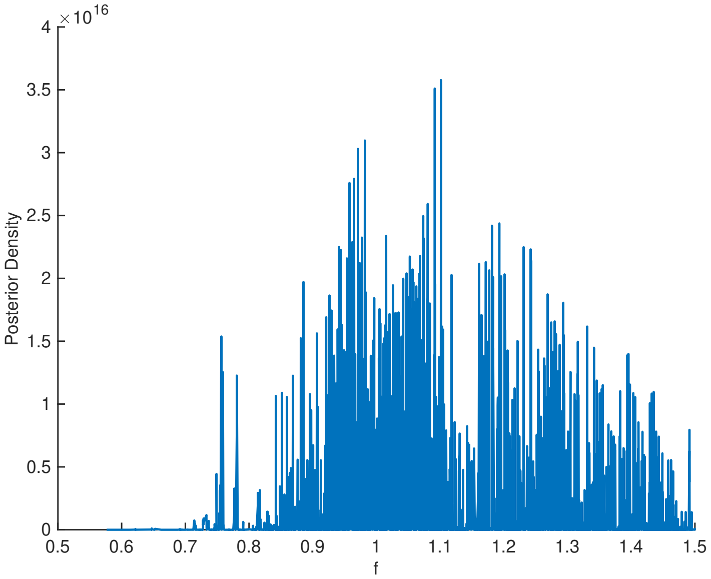

Figure 0.2: Posterior for parameter f for SICR model

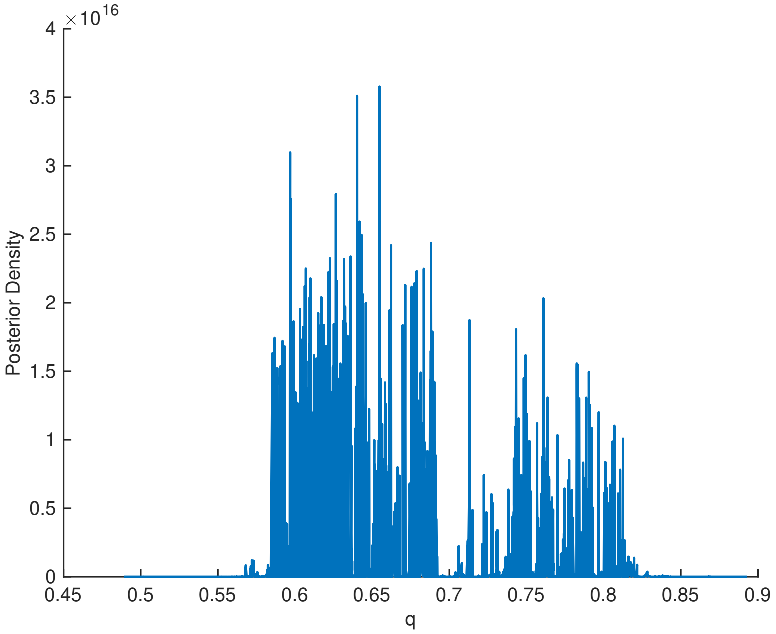

Figure 0.3: Posterior for parameter q for SICR model

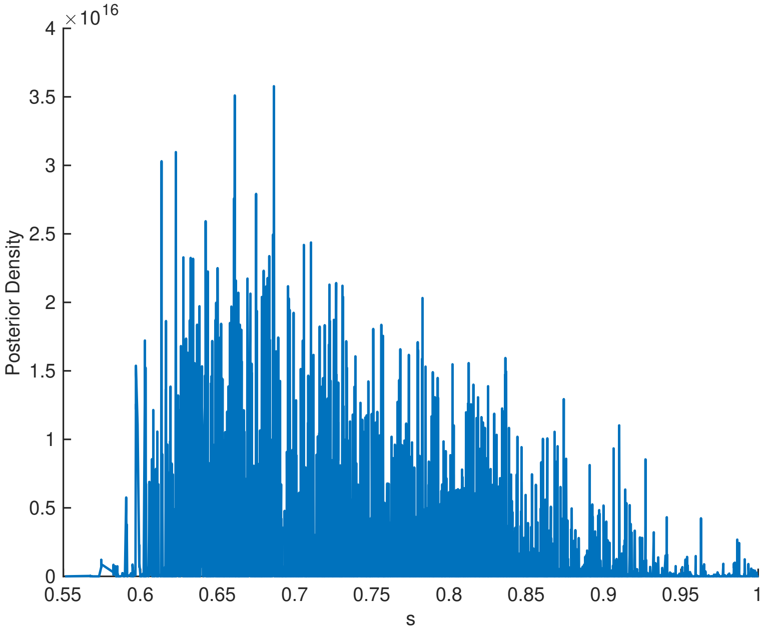

Figure 0.4: Posterior for parameter s for SICR model

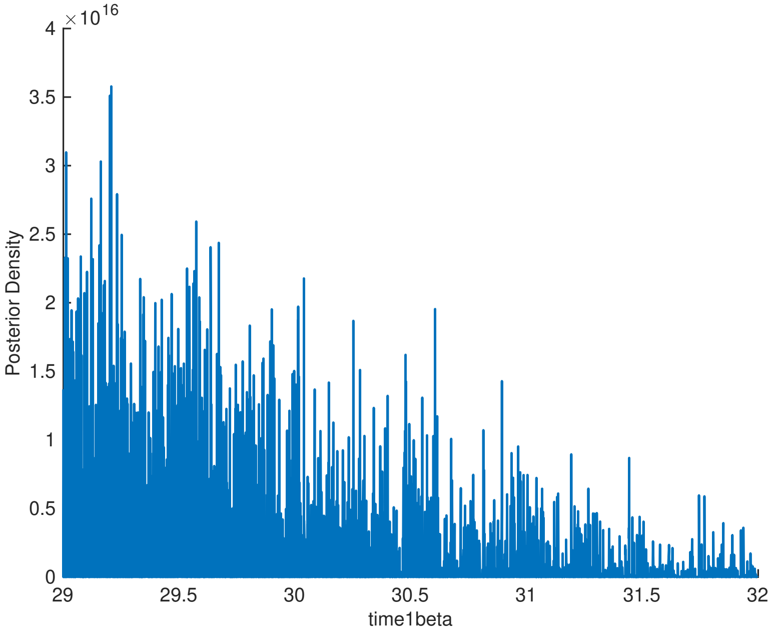

Figure 0.5: Posterior for parameter *time1beta* for SICR model

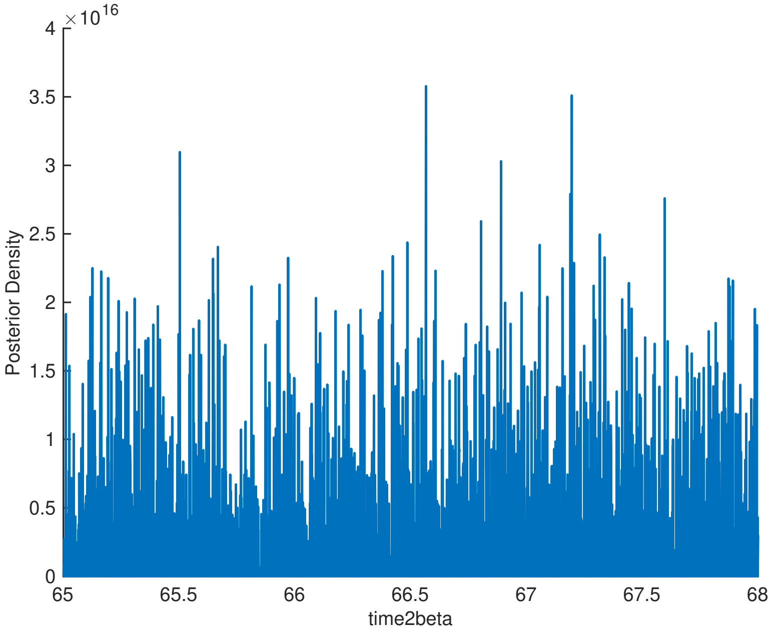

Figure 0.6: Posterior for parameter *time2beta* for SICR model

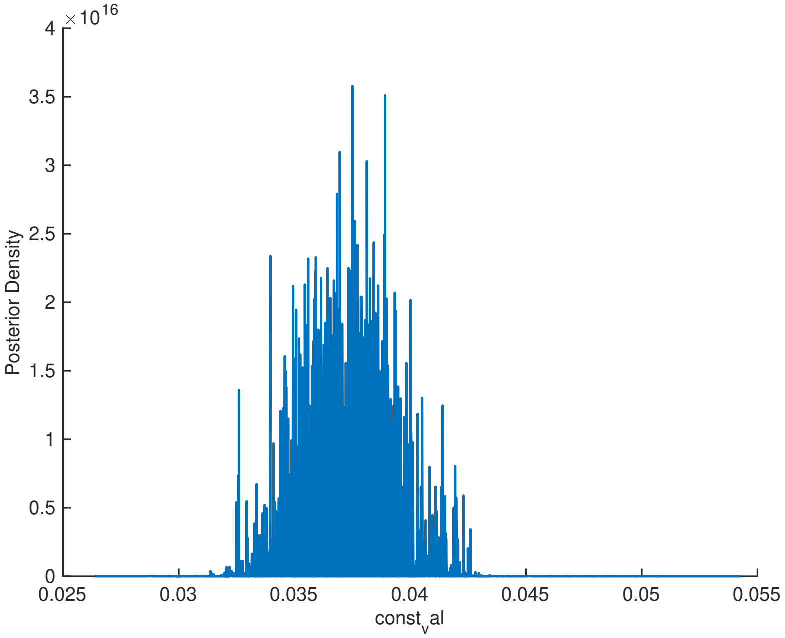

Figure 0.7: Posterior for parameter $const_{val}$ for SICR model

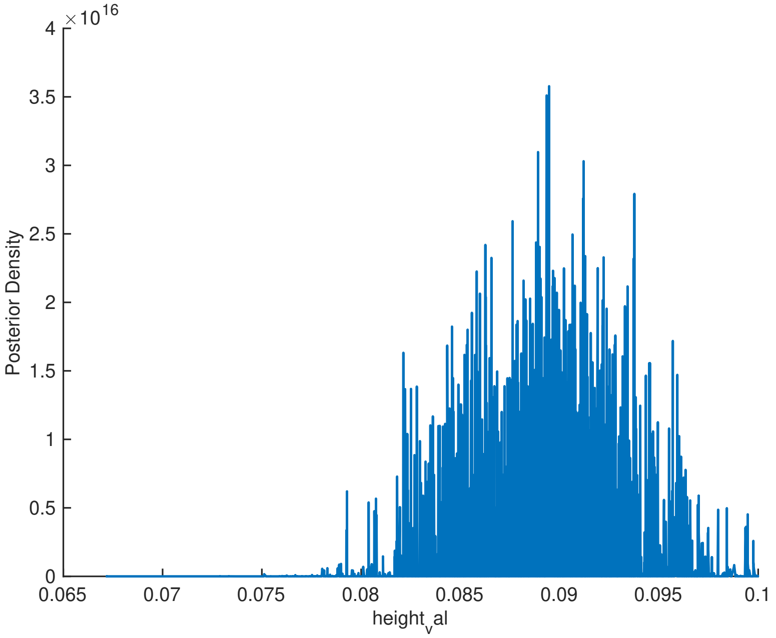

Figure 0.8: Posterior for parameter $height_{val}$ for SICR model

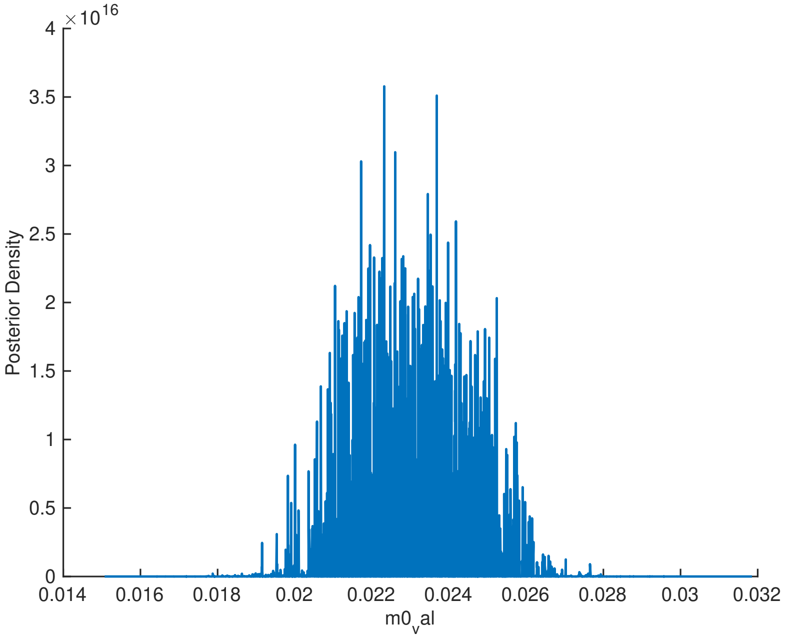

Figure 0.9: Posterior for parameter $m0_{val}$ for SICR model

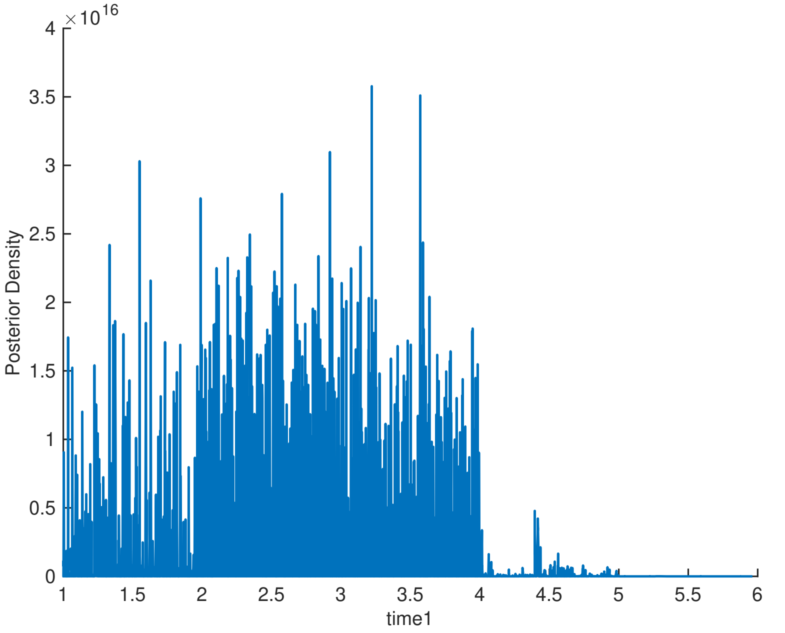

Figure 0.10: Posterior for parameter $time1$ for SICR model

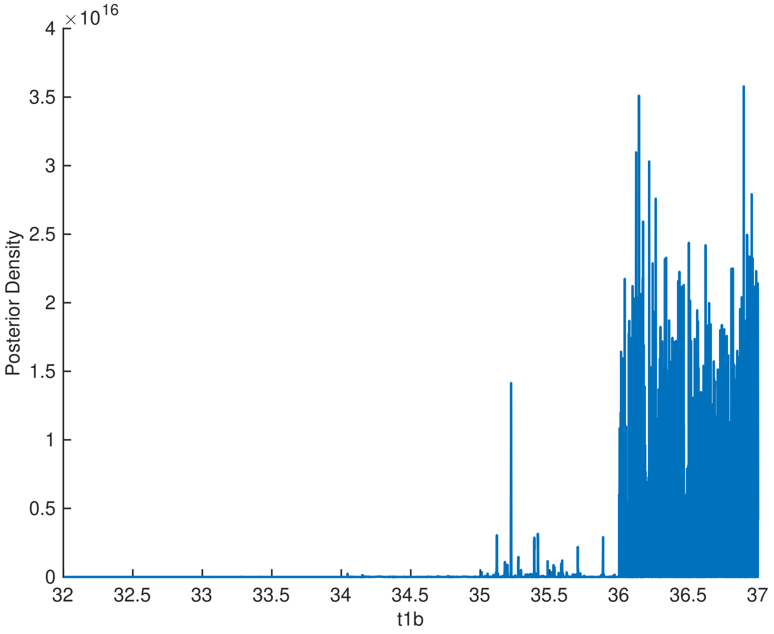

Figure 0.11: Posterior for parameter *time1b* for SICR model

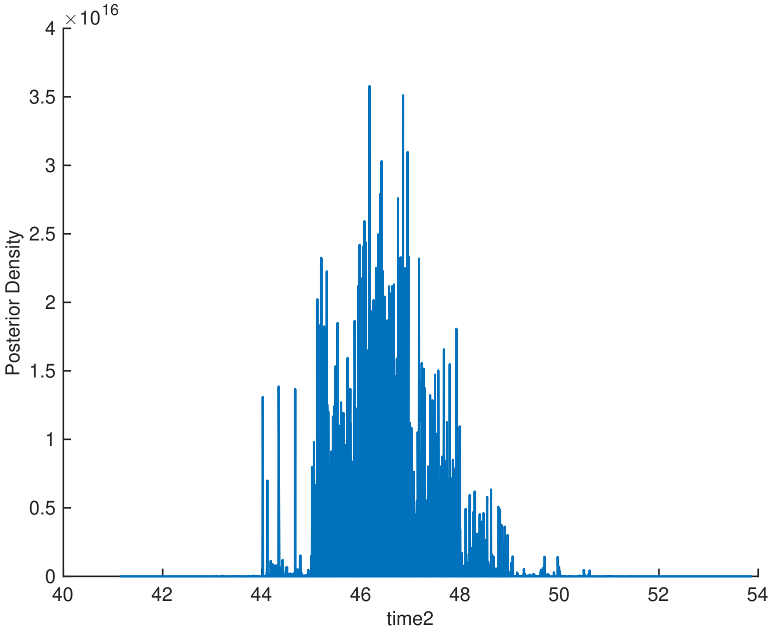

Figure 0.12: Posterior for parameter *time*2 for SICR model

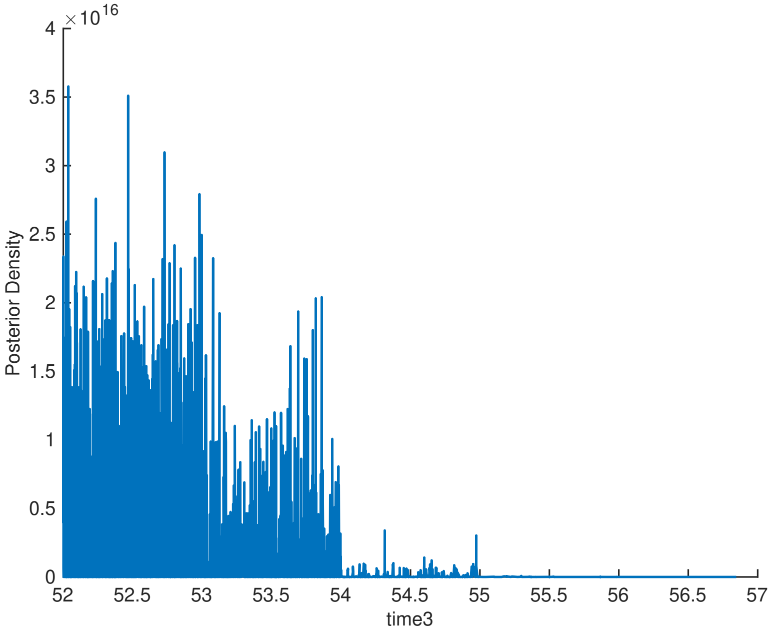

Figure 0.13: Posterior for parameter $time3$ for SICR model

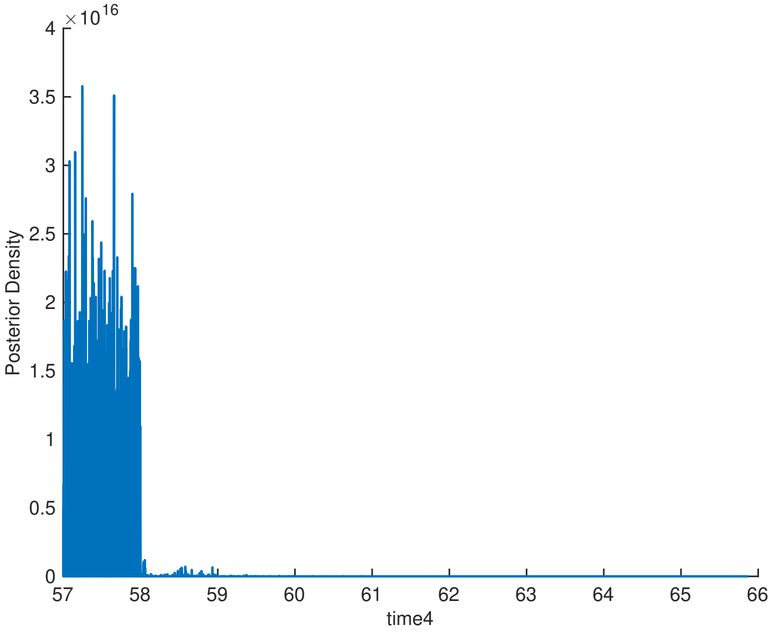

Figure 0.14: Posterior for parameter $time4$ for SICR model

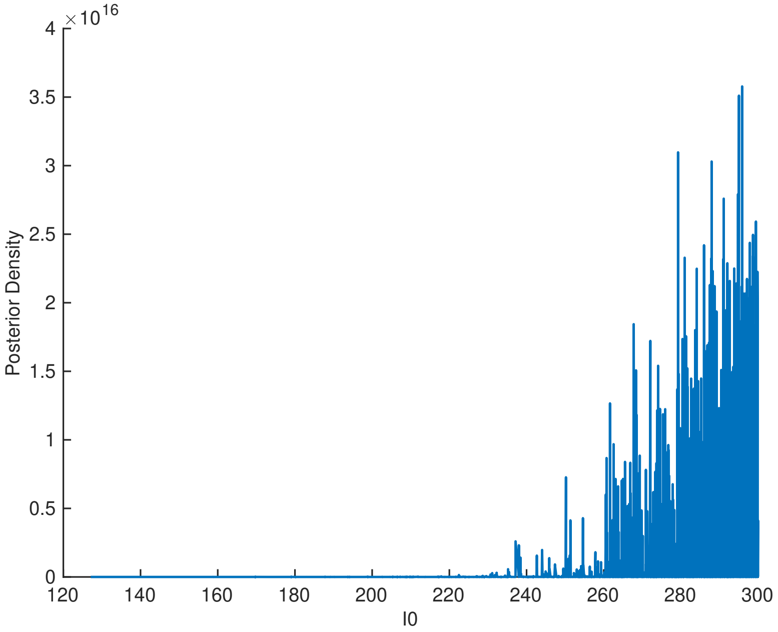

Figure 0.15: Posterior for parameter I_0 for SICR model

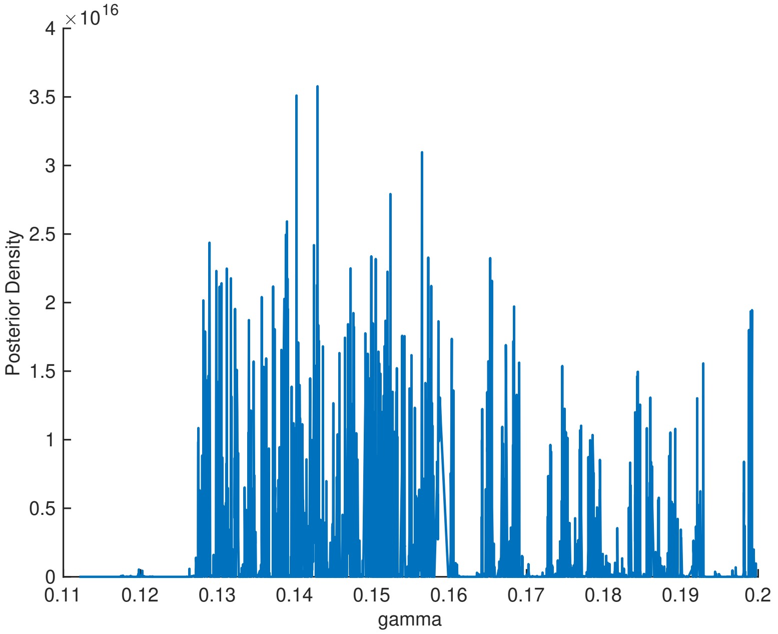

Figure 0.16: Posterior for parameter γ for SICR model

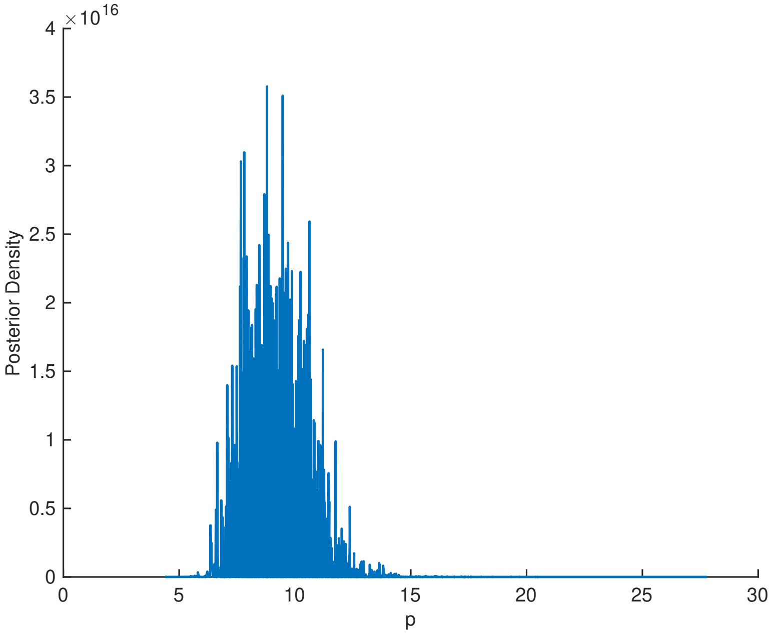

Figure 0.17: Posterior for parameter p for SICR model

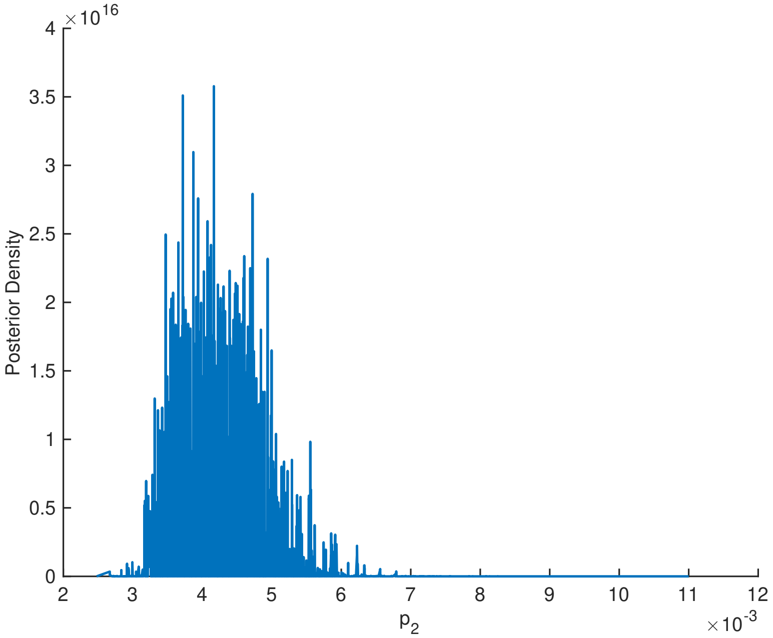

Figure 0.18: Posterior for parameter p_2 for SICR model