1-2020

Simulation Based Inference in Epidemic Models

Tejitha Dharmagadda
Montclair State University

Abstract

From ancient times to the modern day, public health has been an area of great interest. Studies on the nature of disease epidemics began around 400 BC and has been a continuous area of study for the well-being of individuals around the world. For over 100 years, epidemiologists and mathematicians have developed numerous mathematical models to improve our understanding of infectious disease dynamics with an eye on controlling and preventing disease outbreak and spread. In this thesis, we discuss several types of mathematical compartmental models such as the SIR, and SIS models. To capture the noise inherent in the real-world, we consider stochastic versions of these models, and use two types of stochastic simulation algorithms to solve the models. The Gillespie algorithm is used for internal noise while the stochastic Euler algorithm is used for external noise. To improve our understanding of the dynamics, we employ statistical methods on the simulated data and compare with actual data. Treating the epidemic models as a partially observed Markov process (POMP) or hidden Markov model, we use inferential methods via particle filtering and iterated particle filtering to estimate the disease parameters. This simulation-based inference method is demonstrated using an example of influenza data obtained from an infection at an English boarding school.

MONTCLAIR STATE UNIVERSITY

Simulation based inference in epidemic models

by

Tejitha Dharmagadda

A Master's Thesis Submitted to the Faculty of

Montclair State University

In Partial Fulfillment of the Requirements

For the Degree of

Master of Science

January 2020

College of Science and Mathematics

Department of Mathematical Sciences

Thesis Committee:

[redacted]

Dr. Eric Forgoston
Thesis Sponsor

[redacted]

Dr. Andrew McDougall
Committee Member

[redacted]

Dr. Haiyan Su
Committee Member

SIMULATION BASED INFERENCE IN EPIDEMIC MODELS

A THESIS

Submitted in partial fulfillment of the requirements

For the degree of Master of Science

by

TEJITHA DHARMAGADDA

Montclair State University

Montclair, NJ

January 2020

Acknowledgements

I acknowledge my sincere thanks and gratitude to Dr. Eric Forgoston, my advisor, guide and mentor, for the endless support and constant supervision and invaluable guidance in filling the gap between the theoretical and practical knowledge, without whom this project would not have been possible.

I also take the opportunity to thank my committee members, Dr. Andrew McDougall and Dr. Haiyan Su, for their continuous support throughout my masters program.

Contents

List of Figures

1 Introduction

Epidemiology is the study of the outbreak and spread of infectious disease in a group of individuals. It is important to analyze the reasons behind and the potential factors contributing to the outbreak and spread of the disease [2]. The knowledge gained from these studies enables prevention and control through vaccine and quarantine.

The father of medicine is often considered to be the Greek physician Hippocrates, who lived around 400 B.C. He is the first person to seek the relationship between occurrence of disease and environmental factors. Later, John Graunt considered the statistical study of human populations. As the founder of the science of demography, he is called 'the father of demography'. His publication on the analysis of mortality data in 1662 can be considered as one of the earliest contributions to epidemiology [3]. In 1800, William Farr, now considered the father of vital statistics and surveillance, further developed John Graunt's work by collecting and analyzing Britain's mortality statistics [4]. John Snow's contributions to the cholera outbreak in London with the main aim as prevention of disease and finding the cause of the disease has led him to now being considered the 'father of field epidemiology' [5].

Ronald Ross and George McDonald are given credit for their work in mosquito born pathogen transmission [6]. Ross conducted research while in a military post in India in 1897 [6] after discussion with Patrick Manson, who worked in China in 1877, and showed that mosquitoes transmit a blood-borne pathogen. Ross also worked on the control and prevention of mosquito-borne diseases as early as 1902. Later, the theory developed by McDonald helped Ross in his findings which became the basis for the 'Ross-McDonald theory' for understanding mosquito-borne pathogen transmission whose influence continues to the present day [6].

Kermack and McKendrick's contribution to the mathematical theory of infectious disease stems from the early 20th century. Kermack and McKendrick, building on the research of Ross, McDonald and others, developed the first mathematical compartment models to study the spread of disease. Their work enabled one to make hypotheses that predict the number of infectious individuals in a population over time. The basic model developed by Kermack and McKendrick is a deterministic model and has been extended in numerous ways until the present day [5].

1.1 Importance of Epidemiology

Epidemiology is a multidisciplinary approach to studying the outbreak and transmission of infectious disease. Understanding the cause of the infectious disease is not the only thing epidemiologists are interested in. Epidemiologists also should have thorough knowledge about disease modelling, the living habits of the community, environmental factors and community demographics. Consideration of all these factors help in correctly determining the root cause for a disease, either infectious or non-infectious, and in prevention measures by eliminating the factor of the disease [2].

Epidemiology is an approach to human disease prevention and control of disease outbreak using a series of systematic findings and research [5]. The research data can be used to find the cause of the disease and can enable control of the disease. With appropriate data analysis, dreadful infectious, contagious and chronic disease outbreak can be drastically reduced. While it is practically impossible to determine the precise number, undoubtedly epidemiology has contributed to saving millions of lives in the world by implementing various preventive measures. Epidemiology focuses on a wide range of diseases including infectious, non-infectious, contagious or chronic.

Epidemiologists use simple models to obtain a basic understanding of the disease. For infectious diseases, where the disease spreads through person to person contact in a population, mathematical models are designed to capture the most important dynamical processes. For example, in an SIR model, the entire population is divided into three compartments: susceptible, infectious and recovered individuals. The model is designed based on the transmission of disease from infectious to susceptible individuals and the recovery of infectious individuals. By adapting the SIR model, one can consider additional features such as exposure time, control measures, effect of age, and immunization characteristics, to name just a few.

As the real method of transmission of disease from person to person is much more complex, mathematical models are developed with some assumptions. An estimate of parameter values such as contact rate and recovery rate are used with the models to predict how the disease spreads through the population. As it is impossible or rather expensive to accurately collect the data related to epidemic or endemic diseases and since data is often incomplete, to predict the behavior of disease over time epidemiologists often simulate data to perform the required theoretical and statistical experiments. Mathematical models or analysis of simulated data can be used to predict the outbreak and spread of disease and can be used to develop programs for the betterment of the community [7].

2 Theory and Methodology

2.1 Compartmental models

Compartmental models are used to study the outbreak, spread and control of infectious disease. A population is divided into multiple, well-mixed compartments that represent for example, the susceptible, exposed, infectious and recovered individuals in the population. These models can easily be adapted to consider age structure, vertical transmission or many other features relevant to particular diseases. To understand disease outbreak, spread, and extinction in a certain population, we need to know the rates at which all processes in the model occur (birth rate, death rate, contact rate, recovery rate, etc.) The model includes these rates of change for each process which describes how individuals transition from one compartment to another. These mathematical models help us to understand and investigate the spread of the disease as well as the implementation of strategies to control or contain the disease [8].

In the classical SIR model developed by Kermack and McKendrick [9] , individuals in the population are divided into S, I, and R compartments representing Susceptible, Infectious, and Recovered individuals respectively. Other similar models often used include the SIS and $SEIR$ models, where the S, I, and R compartments are as defined for the SIR model, and where the E compartment is defined as the Exposed individuals (who are infected with the disease but are not yet infectious and capable of spreading the disease). In particular, the different groups of individuals are defined as follows:

1. Susceptible (S): group of individuals who are susceptible to the disease and may be infected if they come into contact with an infectious individual.

2. Exposed (E): group of individuals who are infected but are not yet infectious.

3. Infected (I): group of infectious individuals who can transmit the disease to susceptible individuals.

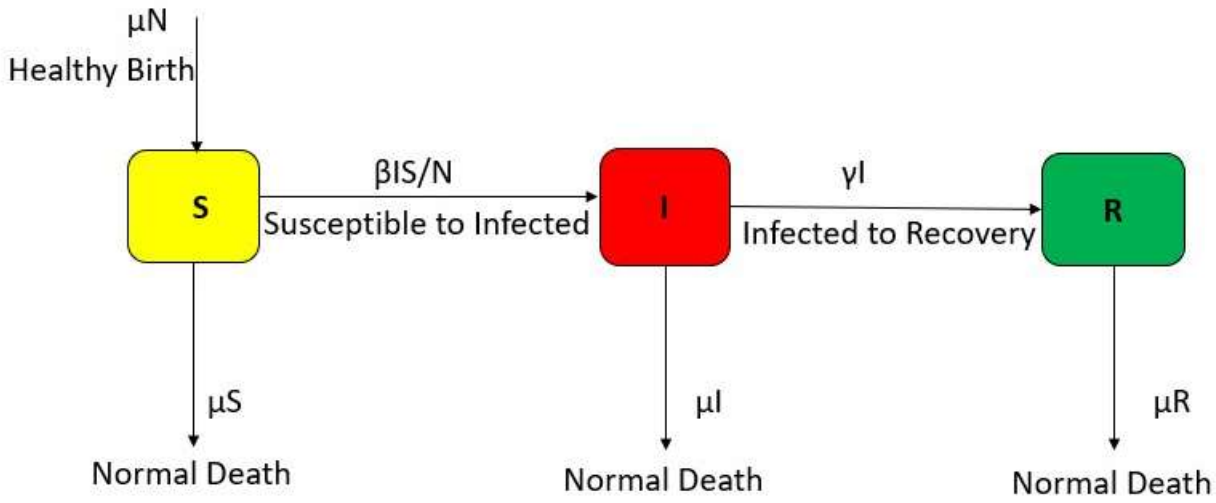

Figure 1: SIR compartmental model with associated processes and rates of change.

4. Recovery (R): group of individuals who have recovered from the disease and are considered immune.

Details for two basic compartmental models of infectious diseases are discussed below.

2.1.1 Susceptible-Infectious-Recovered (SIR) Model:

The first epidemiological compartmental model due to Kermack and McKendrick [9] is the SIR model. This model is appropriate for childhood infectious diseases in which the recovered individual attains permanent immunity against the infection. Figure 1 shows a schematic outlining the compartment model with associated processes and rates of change.

The governing equations for the SIR model can be formulated as

$$\frac{dS}{dt} = \mu N - \frac{\beta IS}{N} - \mu S, \tag{1}$$

$$\frac{dI}{dt} = \frac{\beta IS}{N} - (\mu + \gamma)I, \tag{2}$$

$$\frac{dR}{dt} = \gamma I - \mu R, \tag{3}$$

where S, I, and R are the number of susceptible, infectious and recovered individuals respectively, $N = S + I + R$ is the total population size, μ is the birth and death rate, β is the contact rate, and γ is the recovery rate. In these SIR equations, the μN term represents the number of healthy individuals born into the susceptible class, while the μS, μI, and μR terms give respectively the number of individuals leaving the S, I, and R classes due to natural death. Additionally, the $\frac{\beta IS}{N}$ term represents the number of susceptible individuals who come into contact with an infected individual, become infected and move into the infected class, while γI gives the number of individuals who transition from the infected to the recovered compartment due to recovery. One can

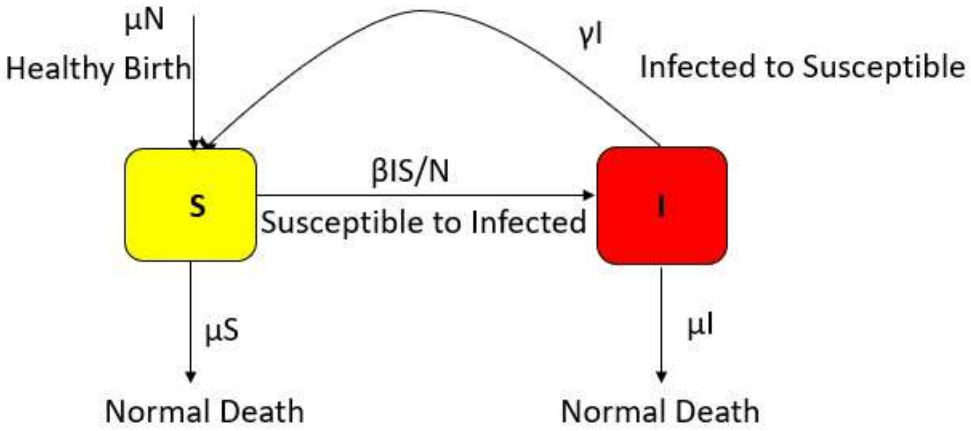

Figure 2: SIS compartmental model with associated processes and rates of change.

now easily describe the SIR equations. For example, in Eq. (1), the rate of change of susceptible individuals in time is equal to the gain of individuals from healthy birth, the loss of individuals from the susceptible class to the infected class due to infection, and a loss from natural death. Similarly, the rate of change of infected individuals in time can be given as a gain of individuals due to infection, a loss of individuals from natural death, and a loss of individuals due to recovery. Lastly, the rate of change of recovered individuals in time is given by the gain due to recovery, and the loss of individuals due to natural death.

2.1.2 Susceptible-Infectious-Susceptible (SIS) Model

When the recovery does not give any immunity against the infection the model is known as the SIS model since recovered individuals immediately become re-susceptible to the disease. Figure 2 shows a schematic outlining the compartment model with associated processes and rates of change. For this model, the governing equations become

$$\frac{dS}{dt} = \mu N - \frac{\beta IS}{N} - \mu S + \gamma I, \tag{4}$$

$$\frac{dI}{dt} = \frac{\beta IS}{N} - (\mu + \gamma)I, \tag{5}$$

where S and I are the number of susceptible and infectious individuals respectively, $N = S + I$ is the total population size, μ is the birth and death rate, β is the contact rate, and γ is the recovery rate. There are numerous other versions of compartmental models that add additional compartments to capture exposure time, vaccination or quarantine, age structure, etc. Regardless of their exact form, they have similar structure to the SIR and SIS epidemic models discussed here in detail.

2.2 Deterministic versus Stochastic Models

To mathematically model the dynamics of an infectious disease in a population, there are two broad approaches that one can consider. The classical approach, and the one used by Kermack and McKendrick as well as many others over the years, is a deterministic approach. The model is given by a set of ordinary differential equations whose solution determines exactly the state of the disease in time. However, real data gathered during disease events is noisy. Hence the need for the other approach which is a stochastic one. For internal demographic noise, the stochastic model consists of very large (possibly infinite) set of differential equations known as the master equation [10]. The master equation gives the probability of the system having a particular number of infectious individuals at any instant in time.

In deterministic models a large population is divided into groups or compartments, each representing a specific stage of the epidemic. As discussed in the above example for the SIR model, the compartments are S, I and R. In these situations, the mathematical model is formulated using ordinary differential equations. The population size in a compartment changes deterministically in time according to the transition rates and initial conditions. In short, for a prescribed set of parameter values and initial conditions the model always gives the same output no matter how many times we solve the equations. There is no randomness, and hence we have one output solution. These deterministic compartment models often have two equilibria: an extinct state in which there is no disease in the population, and an endemic state in which the disease is maintained without any external forcing.

The transmission potential of a disease can be determined using the basic reproduction number, R_0, which is the ratio of the expected number of new infectious individuals from a single infection in an entirely susceptible population. The basic reproduction number tells us if the disease could spread in a population. When $R_0 < 1$, the number of infectious individuals declines, and hence the disease will go extinct (possibly after a small outbreak). When $R_0 > 1$ the number of infectious individuals increases and the disease takes off in the population [10].

When $R_0 > 1$ the infection will spread and will be maintained in the population for a long period of time. In this situation the endemic state is stable. Because of the stability of the endemic state, in the deterministic approach there is no chance for the disease to go extinct. However, in real world scenarios, we see that disease outbreaks often do go extinct, at least locally. In order to capture these extinction events mathematically, we must use a stochastic model that includes the randomness or noise in the system. In particular, we will capture the internal demographic noise. This internal noise can lead to a rare, large fluctuation that causes the disease to escape from the endemic state and go to the extinct state [10].

A stochastic model involves random processes which lead to unpredictable outcomes. To properly understand the stochastic effects on disease modeling, we must estimate the probability distributions of potential outcomes with variation in inputs over time. Since the same set of parameters and initial conditions lead to a wide variety of outputs, we must consider a probabilistic/statistical framework to properly understand the results of the stochastic models. It is worth noting that most real world situations are inherently noisy, and therefore the statistical techniques discussed within this proposal are general and can be used for many model systems, not just epidemic models.

3 Stochastic Modelling

For the simplest of problems, it is possible to perform theoretical analysis to solve a stochastic problem. However, in many instances, one must simulate the stochastic solutions numerically and then perform statistical analysis to understand the system's behavior. In the following sections, we present an example of the theory along with the numerical simulation algorithm.

3.1 Theoretical Analysis - An Example

Consider the simple birth-death-immigration (BDI) process with a population size N [11]. We neglect the effect of susceptible depletion, assuming that only a small fraction of the population are infected at one time, replenishing the pool of susceptible individuals at sufficient speed. The three steps involved in the BDI process are: (i) introduction of the infectious individuals into the population; (ii) the infection spreads to the other individuals at a rate proportional to the number of infectious individuals present; and (iii) infectious individuals recover from the disease.

We assume that n is the number of infectious individuals, $\beta(t)$ is the transmission rate of the disease, ζ is the constant rate of individuals importing the infection due to contact with external sources, γn is the rate of recovery, $1/\gamma$ is the typical duration of an infection, and $\lambda(t) = \beta(t)n + \zeta$ is the total force of infection consisting of the transmission of the disease between individuals as well as the import of disease from external sources.

The transition rate $T_{m,n}$ is defined as the probability per unit time of transitioning from a state with n infected individuals to a state with m infected individuals. In this BDI process there are two possible transitions, namely an infection event and a recovery event. The transition rates for infection and recovery are given as

$$T_{n+1,n} = \beta n + \zeta, \tag{6}$$

$$T_{n-1,n} = \gamma n. \tag{7}$$

In the BDI process, the chain of transmission stemming from a particular introduced index case is given by a branching process [11,12]. A particular outbreak can be considered as a superposition of the separate chains of transmission caused by each introduced case during the outbreak [13]. The basic reproductive number is defined as the average number of secondary cases, $R_0 = \beta/\gamma$, found using the offspring distribution of the associated branching process [11,12,14].

If $P_n(t)$ is the probability of n individuals being infected at time t, then the change in probability with time is found by solving the master equation. The advantage in considering such a simple model is that the master equation, which determines how the probability distribution of the number of infectious individuals changes in time, can be solved exactly for this model without the need for any approximations [15]. This can be achieved using the moment generating function [11].

In the BDI process, the random variable is n, noting that n is discrete as it is the number of infected individuals. Therefore, the moment generating function can be written as

$$Z(\psi, t) = E[e^{\psi n(t)}] = \Sigma e^{\psi n(t)} P_n(t), \tag{8}$$

where the ith moment is given by

$$\mu_i = E[n^i] = \left. \frac{\partial^i Z}{\partial \psi^i} \right|_{\psi=0}. \tag{9}$$

For the BDI process, the master equation is given as

$$\frac{dP_n(t)}{dt} = T_{n,n-1}P_{n-1}(t) + T_{n,n+1}P_{n+1}(t) - [T_{n+1,n} + T_{n-1,n}]P_n(t), \quad \forall n > 0 \tag{10}$$

and

$$\frac{dP_0(t)}{dt} = T_{0,1}P_1(t) - T_{1,0}P_0(t), \quad n = 0. \tag{11}$$

The transition rates for infection and recovery are given by Eqs. (6) and (7). Substitution of these rates into the mater equation given by Eq. (10) gives

$$\begin{aligned}\frac{dP_n(t)}{dt} &= (\beta(n-1) + \zeta)P_{n-1}(t) + \gamma(n+1)P_{n+1}(t) - [\beta n + \zeta + \gamma n]P_n(t) \\ &= \beta(n-1)P_{n-1}(t) + \zeta P_{n-1}(t) + \gamma(n+1)P_{n+1}(t) - \beta n P_n(t) - \gamma n P_n(t) - \zeta P_n(t) \\ &= \beta[(n-1)P_{n-1}(t) - nP_n(t)] + \gamma[(n+1)P_{n+1}(t) - nP_n(t)] + \zeta[P_{n-1}(t) - P_n(t)],\end{aligned} \tag{12}$$

and therefore

$$\begin{aligned}\sum \frac{dP_n(t)}{dt}e^{\psi n} =& \beta\left[\sum (n-1)P_{n-1}(t).e^{\psi n} - \sum nP_n(t).e^{\psi n}\right] + \\ & \gamma\left[\sum (n+1)P_{n+1}(t).e^{\psi n} - \sum nP_n(t).e^{\psi n}\right] + \\ & \zeta\left[\sum P_{n-1}(t).e^{\psi n} - \sum P_n(t).e^{\psi n}\right].\end{aligned} \tag{13}$$

Using Eq. (8) one can rewrite Eq. (13) as

$$\frac{\partial Z}{\partial t} = (e^{\psi} - 1)\left(\beta\frac{\partial Z}{\partial \psi} + \zeta Z\right) + (e^{-\psi} - 1)\gamma\frac{\partial Z}{\partial \psi}. \tag{14}$$

It is possible to solve Eq. (14) when R_0, γ and ζ are constant or are functions of time. The solution allows one to analytically compute a number of statistical quantities associated with the BDI process including the mean, variance, coefficient of variation, index of dispersion, correlation time, and autocorrelation [11].

3.2 Numerical Simulation of Data

As demonstrated, for simple models, it is possible to theoretically analyze and understand the complex nature of the the random behaviour of infection in the population over time. However, in most instances, the models are too unwieldy to analyze theoretically. In these cases, it is often necessary to simulate the model's solution data which can then be statistically studied. Once understood, the simulated data can be employed in a wide range of control strategies [16]. In this work, we use two types of numerical algorithms to stochastically simulate the solution data: (a) Gillespie algorithm, and (b) stochastic Euler method.

3.2.1 Gillespie Algorithm

To generate a solution of a stochastic equation where the noise is internal to the system we use the Gillespie algorithm or Gillespie's stochastic simulation algorithm (SSA) [17]. The algorithm is a type of Monte Carlo method that was originally proposed by Kendall [18] for simulating birth-death processes and was popularized by Gillespie [17] as a useful method for simulating chemical reactions based on molecular collisions. The results of a Gillespie simulation is a stochastic trajectory that represents an exact sample from the probability function that solves the master equation. Therefore the method can be used to simulate population dynamics where molecular collisions are replaced by individual events and interactions including birth, death, and infection [10].

Let $\mathbf{x} = (x_1, \ldots, x_n)^T$ denote the state variables of a system, where x_i provides the number of individuals in state x_i at time t. The first step of the algorithm is to initialize the number of individuals in the population compartments $\mathbf{x}_0$. For a given state $\mathbf{x}$ of the system, one calculates the transition rates (birth rate, death rate, contact rate, etc.) denoted as $a_i(\mathbf{x})$ for $i = 1 \ldots l$, where l is the number of transitions. Thus the sum of all transition rates is given by $a_0 = \sum_{i=1}^{l} a_i(\mathbf{x})$.

Random numbers are generated to determine both the next event to occur as well as the time at which the next event will occur. One simulates the time τ until the next transition by drawing from an exponential distribution with mean $1/a_0$. This is equivalent to drawing a random number r_1 uniformly on $(0, 1)$ and computing $\tau = (1/a_0) \ln (1/r_1)$. During each random time step exactly one event occurs. The probability of any particular event taking place is equal to its own transition rate divided by the sum of all transition rates $a_i(\mathbf{x})/a_0$. A second random number r_2 is drawn uniformly on $(0, 1)$, and it is used to determine the transition event that occurs. If $0 < r_2 < a_1(\mathbf{x})/a_0$, then the first transition occurs; if $a_1(\mathbf{x})/a_0 < r_2 < (a_1(\mathbf{x}) + a_2(\mathbf{x}))/a_0$, then the second transition occurs, and so on. Lastly, both the time step and the number of individuals in each compartment are updated, and the process is iterated until the disease goes extinct or until the simulation time has been exceeded.

Figure 3 shows a single realization of a stochastic time series for the SIR model found using the Gillespie model. The cyclic behaviour observed in Fig. 3 is due to the continual influx of susceptible individuals due to birth. Without the birth term, the susceptible pool could not rebuild itself after a disease outbreak, and therefore one would not see multiple outbreaks as seen in Fig. 3.

3.2.2 Stochastic Euler Method

To generate a solution of a stochastic equation where the noise is external we use a stochastic Euler method. Leonhard Euler, one of the most eminent mathematicians of all time, developed a numerical algorithm to solve a deterministic ODE with initial conditions. In general, an exact solution is difficult to find analytically, and so Euler resorted to a numerical approximation. The idea is as follows: given an ODE $\frac{dx}{dt} = f(x)$ with initial condition $x(t_0) = x_0$, we wish to find the solution $x(t)$. As it is difficult to obtain the true solution $x(t)$ to the ODE, Euler wished to find a numerical approximation $\tilde{x}(t)$ to the ODE.

In the algorithm, time is divided into small intervals of length δ so that $t(n) = t_0 + n\delta$, for $n = 1, 2, 3, \ldots$. Initialization of the numerical approximation is done at some known starting value given as $\tilde{x}(t_0) = x(t_0) = x_0$. Euler assumed the slope dx/dt to be roughly constant each small interval of time δt, which allows one to approximate the value of the solution at the next time step.

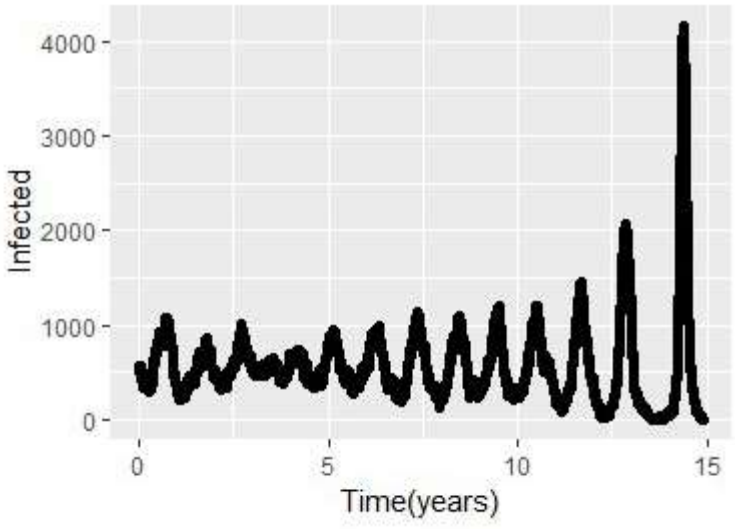

Figure 3: Single time series realization of a stochastic SIR model for infectious individuals. Time is measured in years. Typical measles parameters are used as follows: population size $N = 3 \times 10^6$, $\mu = 0.02$/year, $\beta = 1500$/year, and $\gamma = 100$/year.

This numerical approximation is given by

$$x_{n+1} = x_n + f(x_n)\delta t, \tag{15}$$

and is demonstrated in Fig. 4 which shows a comparison of the Euler approximation to the true solution of an ODE for two different time step choices.

Consider the stochastic ODE

$$\frac{dx}{dt} = f(x) + \sigma\eta(t), \tag{16}$$

where η is an external noise and σ is the standard deviation of the noise intensity $D = \sigma^2/2$. The simplest numerical method to solve Eq. (16) is the stochastic Euler method, sometimes referred to as the Euler-Maruyama method [10]. The stochastic Euler scheme is given by

$$x_{n+1} = x_n + f(x_n)\delta t + \sigma(x_n)\delta\eta_n, \tag{17}$$

where the time increment is $\delta t = t_{n+1} - t_n$ and the noise increment is $\delta\eta_n = \eta_{t_{n+1}} - \eta_{t_n}$. To implement the stochastic Euler scheme, we note that the noise increments are independent Gaussian random variables with the following first and second moments:

$$\mathbb{E}(\Delta\eta_n) = 0, \qquad \mathbb{E}((\Delta\eta_n)^2) = D\delta t, \tag{18}$$

so that

$$\Delta\eta_n \sim \mathcal{N}(0, D\delta t). \tag{19}$$

In practice, one can solve (17) numerically, using a random number generator to draw noise values from the distribution given by (19). Alternatively, one can numerically solve

$$x_{n+1} = x_n + f(x_n)\delta t + \sqrt{D\delta t}\,\sigma(x_n)\,\hat{\eta}_n, \tag{20}$$

where

$$\hat{\eta}_n \sim \mathcal{N}(0, 1). \tag{21}$$

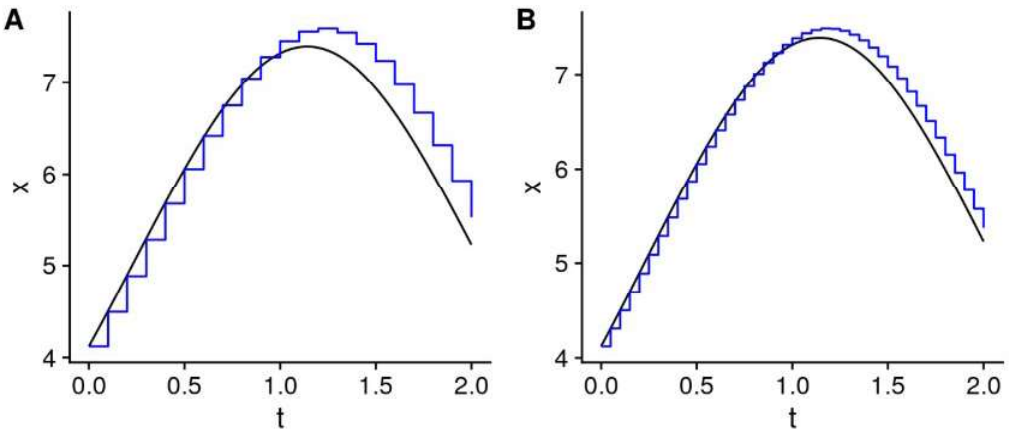

Figure 4: Euler approximations $\tilde{x}(t)$ with step size $\delta = 0.1$ (A) and $\delta = 0.05$ (B) are shown in blue. The true solution, $x(t)$, is shown in black. Image from [1].

4 Partially Observed Markov Processes (POMP)

We discussed previously that there is a need to design theoretical models with which one can perform theoretical analysis as well as simulate data. In this way, we can account for the randomness of infectious diseases as described in the previous section. Recent studies have shown that theoretical studies alone have not led to a complete understanding of the demographic stochasticity [19]. Therefore, statistical models should be developed to provide clarity on the stochastic dynamics of infectious disease. In epidemiology and many other areas Markov Chain Processes have proved to be a very powerful and reliable method [19]. In this section we examine a simulation-based statistical method for epidemiological systems using partially observed Markov process (POMP) models [1]. POMP are also known as Hidden Markov Models as the states to be determined are unobserved [7, 20]. POMP has been and continues to be widely used in reinforcement learning, speech, handwriting and gesture recognition, and bioinformatics [20], and more recently POMP has shown successful results in epidemiology [20–24].

To improve our understanding of the dynamics of infectious disease, we will employ stochastic simulation using a partially observed Markov process. This random process evolves in time t, and the observations are dependent only on the current state of the process and therefore are independent of the past states. POMP is based on a Markov chain, which consists of a set of variables with a property that the future state is conditionally independent on the past given the present state [25]. We use the stochastic Euler scheme for the numerical solution of the model's governing equations.

POMP can be generalized as shown in Fig. 5. The states of the process are denoted by $X_0, X_1, \ldots, X_n, X_{n+1}$, at time $t_0, t_1, \ldots, t_n, t_{n+1}$. Let Y_n be a random variable with observations at time t_n given by $Y_1, Y_2, \ldots, Y_n, Y_{n+1}$. Data is collected at times $t_1 < t_2 < \ldots < t_N$ by $y_1^*, y_2^*, \ldots, y_N^*$. The state process, X_n, is Markovian, and is given by

$$f_{X_n|X_{0:n-1},Y_{1:n-1}}(x_n|x_{0:n-1}, y_{1:n-1}) = f_{X_n|X_{n-1}}(x_n|x_{n-1}), \tag{22}$$

where the density function of each state variable is obtained from the probability distribution at that particular time given the previous state. The measurable random variable, Y_n, depends only

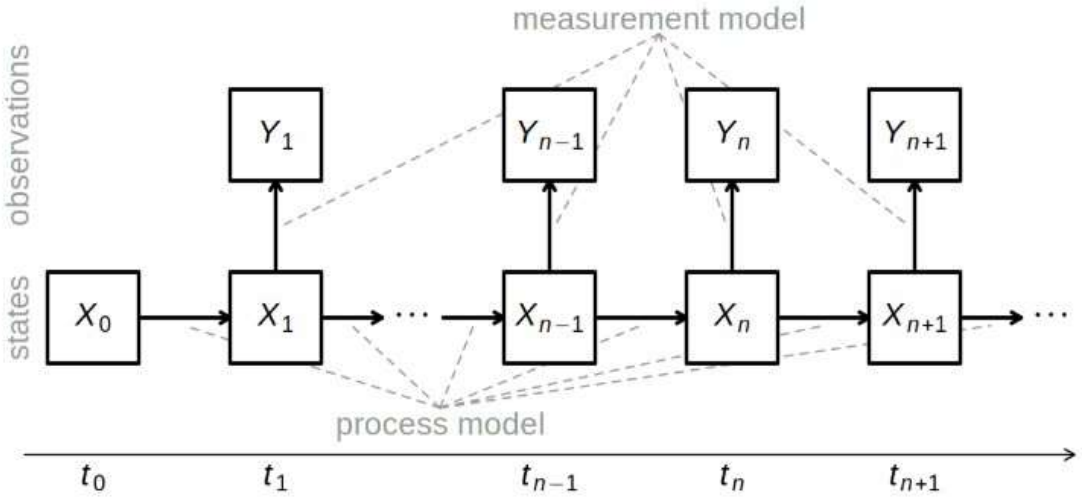

Figure 5: Partially Observed Markov Process (POMP) model schematic. Image from [1].

on the state at that time and is given as

$$f_{Y_n|X_{0:N},Y_{1:n-1}}(y_n|x_{0:n}, y_{1:n-1}) = f_{Y_n|X_N}(y_n|x_n) \tag{23}$$

$\forall n = 1, \ldots, N$, where the observations at any time t are obtained from the probability of the unchanged state.

4.1 SIR Epidemic model in POMP

Consider the SIR compartmental epidemic model discussed in Sec. 2.1.1 with a a fixed population size N. Recall that $S(t)$, $I(t)$, and $R(t)$ represent the number of susceptible individuals, infectious individuals, and recovered individuals respectively at time t. Introducing a change of notation from Sec. 2.1.1, let μ_{SI} represent the rate at which susceptible individuals become infected per day and let μ_{IR} be the rate at which infected individuals recover per day. Additionally, the rate of birth into compartment S is given as $\mu_{.S}$. The mortality rates from each of the S, I, and R compartments are $\mu_{S.}$, $\mu_{I.}$, $\mu_{R.}$ respectively . In the example discussed below involving data of an influenza outbreak at a boarding school, the birth and death rates are not taken into account so that $\mu_{.S}$, $\mu_{S.}$, $\mu_{I.}$, $\mu_{R.} = 0$. A diagram of the compartment model used for the boarding school influenza outbreak can be seen in Fig. 6.

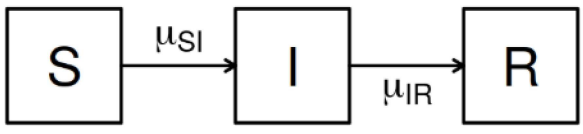

Figure 6: Diagram of the SIR compartmental model used for the boarding school influenza outbreak. Image from [1].

The below counting methods can be used to find the number of individuals in each compartment:

$$S(t) = S(0) - N_{SI}(t), \tag{24}$$
$$I(t) = I(0) + N_{SI}(t) - N_{IR}(t), \tag{25}$$
$$R(t) = R(0) + N_{IR}(t) \tag{26}$$

where $N_{SI}(t)$, and $N_{IR}(t)$ are the count of the number of hosts infected and recovered respectively by time t. The rate at which individuals move from S to I is the force of infection which is given as $\lambda = \mu SI = \beta I/N$, while the rate at which individuals move into the R class is $\mu IR = \gamma$.

It is worth noting that a system of ordinary differential equations with specified initial conditions $S(0)$, $I(0)$, and $R(0)$ is Markovian. Therefore, since only the current state of the system is responsible for the flow rate between the compartments, the compartmental model is a Markov Model.

4.2 A POMP model for boarding school influenza data

As an example, we consider data from an influenza outbreak at a boarding school and show the synthesis of POMP with an epidemic model to perform stochastic simulations as well as statistical inference of parameter values.

The data was gathered from an outbreak of influenza in an all boys boarding school in England recorded in the year 1978. Due to the nature of the school, the population is considered to be closed (i.e. constant). The data, which shows the number of boys confined to bed over a 14 day period from January 22, 1978 to February 4, 1978, is plotted in Fig. 7.

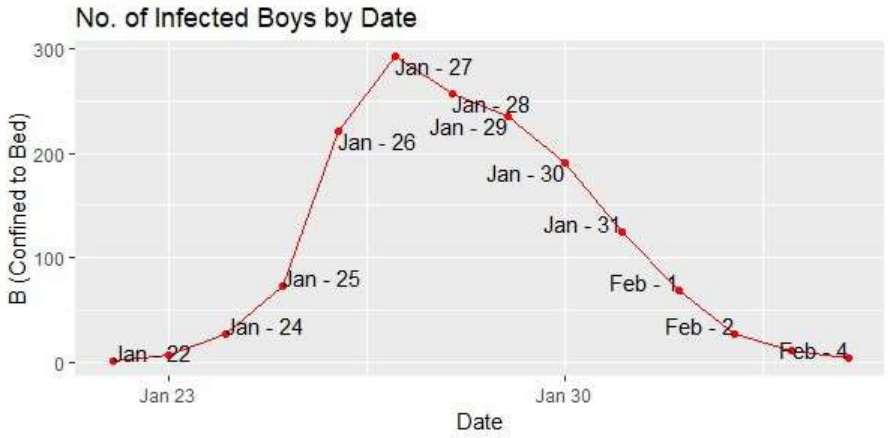

Figure 7: Influenza data showing B, the number of boys confined to bed, over a 14 day period from January 22, 1978 to February 4, 1978.

We model the boarding school influenza data as a partially observed Markov process. Considering the SIR model as the underlying epidemic model for influenza, the hidden states or the unobserved states are the number of individuals in each of the susceptible, infectious, and recovered compartments. The observations are the number of newly reported infectious cases in the population.

To start, we are interested in obtaining a simulated state process rather than the transition probabilities. Therefore, we estimate the parameters of the model using prior knowledge of the SIR epidemic model. Inspecting the data, one finds a total of 1540 infections. Therefore the population size N is larger than this value. The reproductive number R_0 for influenza is generally estimated to be about 1.5 Since the reproductive number can be related to the final size of the epidemic $f = R(\infty)/N$ as

$$R_0 = -\frac{\log(1-f)}{f} \tag{27}$$

one can show that $N = 2600$ approximately. Additionally, the infectious period of influenza is roughly 1 day as that $1/\gamma \approx 1$ day and $\beta = \gamma R_0 \approx 1.5$ day^{-1}. We use these parameter values to simulate the model. For initial conditions, we let the number of infected individuals be $I = 1$ and the number of recovered individuals $R = 0$. Then for a constant population the initial number of susceptibles is $S(0) = N - 1$. The time series for 20 stochastic realizations using the stochastic Euler scheme for the infection in time (in days) is shown in Fig. 8.

Figure 9 compares the actual influenza data (in red) with the 20 simulations. One can observe that there is a shift of the peak at day 6 in the actual data to more than 10 days in the simulated data. Overall, the simulated data agrees well qualitatively with the actual data for our parameter guess. Later, we will use inference methods to estimate the parameters to improve the comparison.

4.2.1 Effect of Gillespie algorithm

Instead of using the Euler scheme, we can consider the Gillespie algorithm for the same model and the same parameter and initial values. Figure 10 shows the simulated data of the infected versus time (in days) under the Gillespie scheme. Comparing with the simulations using the Euler scheme, we see improved accuracy in the timing of the outbreak. However, the amplitude is not quite as good. Overall, however, there is good qualitative agreement for both Euler and Gillespie schemes.

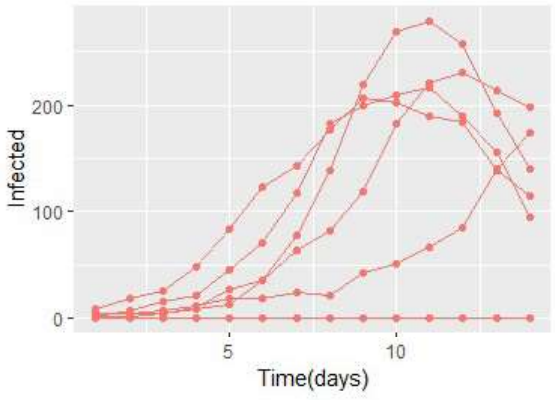

Figure 8: Infected individuals as a function of time (in days) for 20 stochastic realizations found using the stochastic Euler scheme. Note that many realizations have gone extinct after just a few days - all of these realizations are overlapping each other in the bottom-most realization.

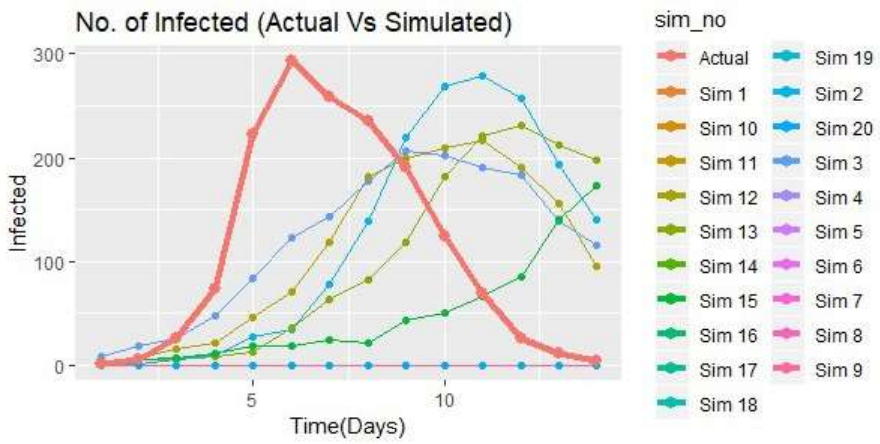

Figure 9: Comparison of the actual data and 20 simulations of infected versus time (in days). Note that many realizations have gone extinct after just a few days - all of these realizations are overlapping each other in the bottom-most realization.

4.2.2 Effect of Vaccination

Within the framework we have described, it is possible to study the effect of a vaccine as a control mechanism. Vaccination is a protection that can give immunity against a certain disease. Within the context of the boarding school influenza example, we are interested in assessing the impact of vaccination on the spread of the disease. In particular, we introduce a systematic vaccination group with a particular rate on the susceptible class. The idea is that vaccinated susceptibles move into the recovered and immune compartment and therefore can no longer become infected by the disease. The model is shown in Fig. 11, and the governing equations for the SIR model with vaccination can be formulated as

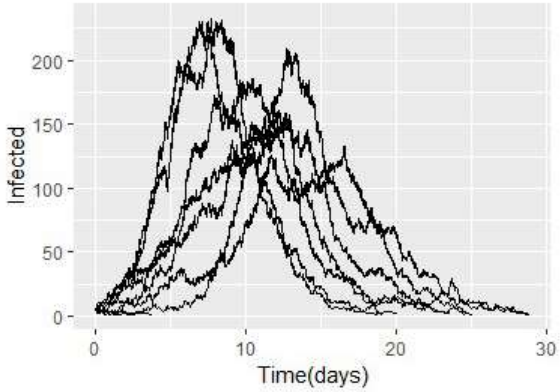

Figure 10: Infected individuals as a function of time (in days) for 20 stochastic realizations found using the stochastic Gillespie scheme.

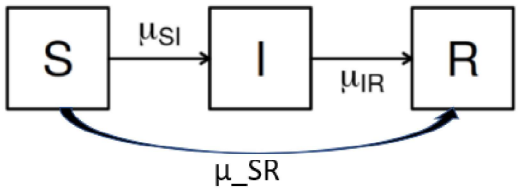

Figure 11: Diagram of the SIR compartmental model with vaccination used for the boarding school influenza outbreak.

$$\begin{aligned}
\frac{dS}{dt} &= -\frac{\beta IS}{N} - \alpha S, \\
\frac{dI}{dt} &= \frac{\beta IS}{N} - \gamma I, \\
\frac{dR}{dt} &= \gamma I + \alpha S,
\end{aligned} \tag{28}$$

where S, I, and R are the number of susceptible, infectious and recovered individuals respectively, $N = S + I + R$ is the total population size, β is the contact rate, γ is the recovery rate, and α is the vaccination rate. Note that birth and death has been neglected due to the short time span associated with this influenza outbreak.

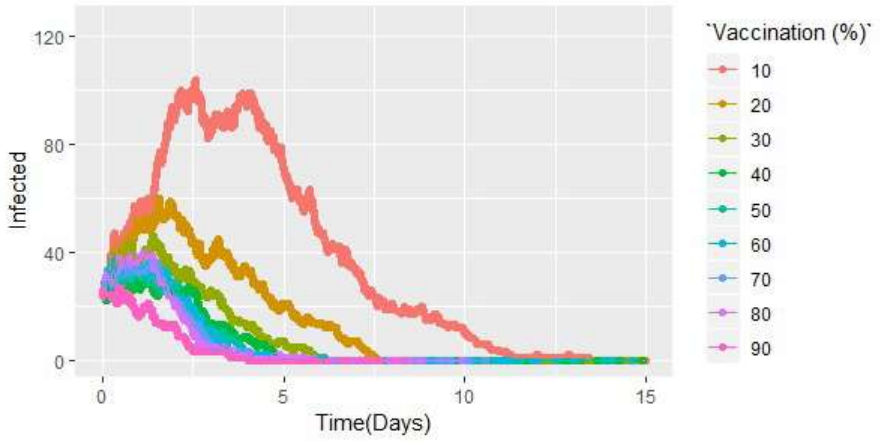

Figure 12: Infected individuals versus time (in days) for varying vaccination rates ranging from 10 to 90 percent.

From Figs. 9 and 10 we can observe the peak of the infection with no vaccination is approximately 200. When vaccination is included, Fig. 12 shows a systematic decrease in the peak number of infectious individuals as the vaccination rate increases from 10% to 90%. One can also see the disease goes extinct much faster for the higher vaccination rates.

Although the simulations discussed in this section are quite reasonable qualitatively, there are some differences between the simulated data and the actual influenza data. In order to improve

the agreement, we will now estimate the parameters statistically. Specifically, we will investigate the use of maximum likelihood estimation of the parameters.

5 Likelihood-based inference for POMP models

In this section we will employ standard likelihood based inference to estimate the parameters for epidemiological data. Let the set of data consist of N observations denoted by $y^*_{1:N}$. Given a parameter vector θ, the density function $f_{Y_{1:N}}(y_{1:N};\theta)$ defines the probability distribution with parameter θ. This density function forms the statistical model. Then the likelihood function, which is the density function evaluated at the data, is given by [1]

$$L(\theta) = f_{Y_{1:N}}(y^*_{1:N};\theta).$$

For convenience, one often works with the logarithm of the likelihood function, given by

$$l(\theta) = \log L(\theta) = \log f_{Y_{1:N}}(y^*_{1:N};\theta).$$

5.1 The particle filter

The particle filter [26] is a method we can use to estimate the parameters of a dynamical systems when only partial observations are available. The method allows one to compute the likelihood of a POMP model more efficiently than direct Monte Carlo integration techniques [1]. The likelihood is an approximation of the parameters which best define the model. Unlike direct Monte Carlo integration, this sequential Monte Carlo uses standard Monte Carlo techniques to sequentially estimate the integrals by simulating, updating and resampling [1].

To implement the particle filter for the estimation of parameters in a POMP model, recall from Fig. 5 that $X_{0:n+1}$ represents the states and $Y_{1:n+1}$ represents the observations at time $t_{1:n+1}$. Using the observations $y^*_{1:N}$ from the POMP model we randomly select P samples from the density function for the state variable at time t_{n-1}. This posterior distribution is given as

$$f_{X_{n-1}|Y_{1:n-1}}(X_{n-1}|Y_{n-1}).$$

A new set of samples are obtained from the prior distribution

$$f_{X_n|X_{n-1}}((X_n|X_{n-1}))$$

at time t_n using the model.

Next, weights are calculated for each particle in the posterior distribution. The weights, $p(y^*_n|X^i_n))$, for $i = 1 \dots N$, of the particles in the prior distribution is updated using the weight of the particles calculated in the posterior distribution. The more weight a particle has, the greater the probability that it will be drawn in the prior distribution. The weights are then normalized to sum to unity.

Prior particles are re-sampled with replacement according to the normalized weights in the posterior distribution. The samples in the prior distribution are chosen with a probability equal to its normalized weights [27]. The entire procedure outlined above is iterated several times with simulating, updating and re-sampling in each step until $n = N$. The likelihood is approximated by

$$\hat{L}_n(\theta) = \frac{1}{P} f_{Y_n|X_n}(y^*_n|X_n),$$

with the log likelihood given by $l(\theta) = \log L(\theta)$.

Returning to the example of boarding school influenza data, we consider fitting the SIR model to the influenza data to estimate the parameters. The estimation of the parameters was found numerically using the R package pomp [28]. Using the initial conditions from Sec. 4.2, we performed the particle filtering and found the log likelihood to be -88.86625419, with associated estimated parameters given as $\beta = 2$, $\gamma = 1$ and $\rho = 0.9$, where ρ is the probability that an infection is observed and recorded.

These results reflect the log likelihood of a single realization of the particle filter. To attain a much higher likelihood of the estimated parameters we perform several of these realizations of the particle filter in order to maximize the likelihood that is obtained. The process of using iterated filtering to determine the maximum likelihood estimate is described in the following section.

6 Maximizing the likelihood of an estimate using iterated filtering

A maximum likelihood estimate (MLE) is the maximum likelihood over all possible θ. The MLE estimates parameter values lying in a subset of the set of possible parameter values. The estimator considered lies in the subset and is the solution of the likelihood equations which gives a maximum of the likelihood function [29] as shown by

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, l(\theta),$$

where

$$\underset{\theta}{\operatorname{argmax}}\, g(\theta)$$

determines a value of the argument θ at which the maximum of function g is attained. If there are many values of θ giving the same maximum value of the likelihood, then an MLE still exists but is not unique [1].

6.1 Iterated Filtering

Iterated filtering is useful for choosing the point estimate corresponding to the MLE. Iterated filtering methods have been shown to solve likelihood based inference problems in epidemiological situations which are computationally intractable for the available Bayesian methods [1]. When employing iterated filtering, each iteration has a particle filter, with a parameter vector for each particle undergoing a random process. At the end of the time series, the collection of parameters at the end of each iteration is taken as starting parameters for the next iteration. At each iteration, the random process variance decreases. In this way, the procedure maximizes the maximum likelihood.

This method of statistical inference has shown promise in the estimation of parameters along with profile likelihood, confidence intervals, likelihood ratio tests, etc. [20]. These inferential techniques can be applied to a variety of disease data including influenza, measles, or polio data to name just a few. The idea is to replace the epidemic model we are interested in fitting the data, having invariant parameters, with the same type of model but with parameters that take a random walk in time [20]. By repeating the filtering procedure for many iterations this intensity of the random walk approaches zero and the modified model approaches the true model. In iterated filtering this random walk intensity (referred to as the temperature) must be decreased by a factor (called the cooling schedule) which approaches zero when the maximum likelihood is reached [20].

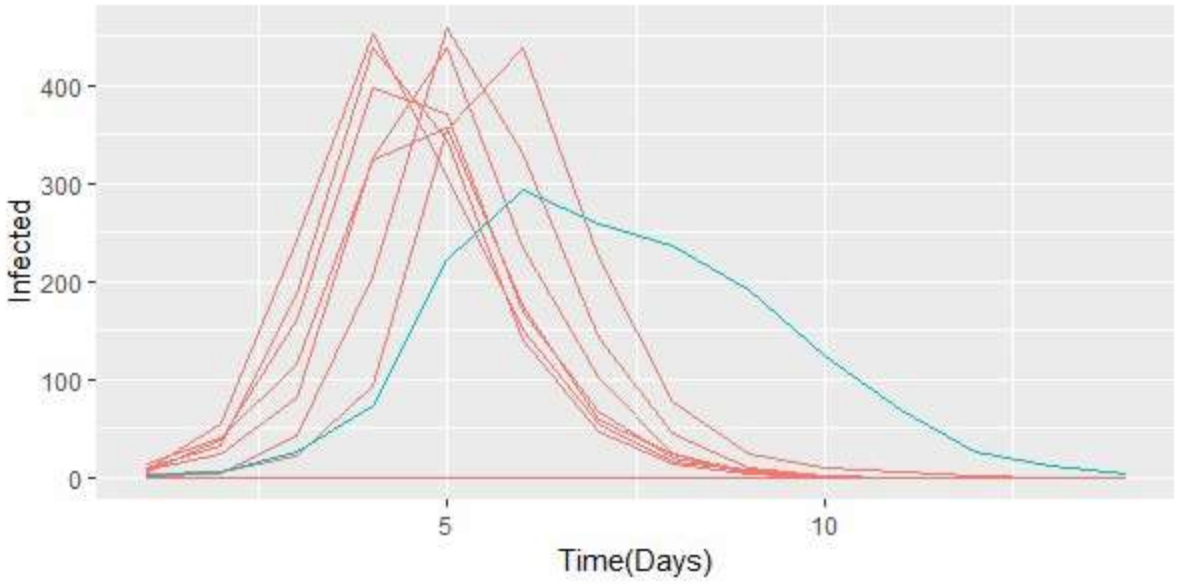

Figure 13: Comparison of the actual data (blue) and simulations (red) of infected versus time (in days). The simulations use parameter values derived a maximum likelihood estimate using iterated filtering.

The estimation of parameters via particle filtering provides us with a point estimate of the parameters. We conduct a local search around this point given the intensity of the random walk and the cooling schedule. In the parameter space we are looking for a set of parameter values around the point estimate that contain all the sensible parameter vectors. These parameter vectors have a diverse starting point on which a global search is conducted. With this set of parameters many iterations are run which gives us set of likelihood estimates for the most sensible parameter vectors. The likelihood estimate with a maximum value is considered to be the MLE which gives us the model which best fits the data.

As an example, we return to the the boarding school influenza data and apply the iterated filtering algorithm. The initial conditions and parameter starting values are the same as from the particle filtering example in the previous section. Iterated filtering is performed for 300 iterations and the maximum log likelihood obtained is -72.66927 with associated parameter values of $\beta = 3.571134$, $\gamma = 1.985172$ and $\rho = 0.9533523$.

Given the MLE of parameter values, we can re-compute simulations of the disease model. Figure 13 shows the results. When compared with Figure 9 which was computed using best-guess parameter values from our knowledge of epidemiology, we find the new plot gives better predictions of the actual data. In Figure 13 we find a clear shift in the peak from day 10-12 to day 4-6 when compared to the actual data at day 6. Overall, the simulation based inference algorithms provide much better results and improved understanding of the disease dynamics.

7 Summary and Remarks

We have presented a numerical approach for understanding different epidemiological compartmental models with stochastic simulation methods. Specifically, we highlight the Gillespie algorithm and stochastic Euler scheme as tools for simulating infectious disease data. Both actual and simulated data are important for understanding a variety of applications, including predicting when and how can the disease go extinct as well as the optimal path for the control of disease by vaccination or quarantine [10, 30–32].

Unlike the deterministic approach, the stochastic approach captures the randomness and noise which can arise from demography or from external sources. For simple models, it is possible to theoretically solve these stochastic models. However, most models of interest are too complicated to solve analytically. Therefore we rely on simulations and statistical modelling. In this work we considered two types of numerical algorithms to simulate the data: Gillespie algorithm for internal noise and stochastic Euler for external noise.

Since real epidemic data is relatively sparse, we projected the data as a partially observed Markov process (POMP) with some initial conditions of the unobserved state process. Importantly, the future state depends only on the current state. The observations are the number of newly reported infectious individuals at a given time in each state. Using simulated data from stochastic Euler, statistical inference provided parameter values that matched well with actual influenza data (see Fig. 9). We additionally investigated the differences between Gillespie simulated data and Euler simulated data, and also showed the effects of vaccination.

It is well worth noting that inference including particle filtering and iterated filtering, provide more accurate estimation of parameters with relatively fast computational speed. In dynamically complex models that are nonlinear and non-Gaussian, it can be very difficult to estimate parameters, and in these cases the inference algorithms are of great use. Coupling these inference with POMP models for disease enables one to achieve a maximum likelihood of the estimate that gives parameter values that better agree with the actual data. These inferences can be used not just in epidemic models but also in many different fields. Our work allows epidemiologists to have improved understanding of disease dynamics as they perform predictive work that is important for the survival of mankind.

8 Bibliography

References

[1] E. L. Ionides and A. A. King, "Module in simulation-based inference for epidemiological dynamics," 11th Annual Summer Institute in Statistics and Modeling in Infectious Diseases (SISMID 2019).

[2] W. H. Frost, "The importance of epidemiology as a function of health departments," American Journal of Public Health, vol. 13, no. 1, pp. 33–37, 1923.

[3] J. Graunt, Natural and political observations mentioned in a following index, and made upon the Bills of Mortality. 1662.

[4] A. D. Langmuir, "William Farr: founder of modern concepts of surveillance," International Journal of Epidemiology, vol. 5, no. 1, pp. 13–18, 1976.

[5] D. Quammen, Spillover: animal infections and the next human pandemic. WW Norton & Company, 2012.

[6] D. L. Smith, K. E. Battle, S. I. Hay, C. M. Barker, T. W. Scott, and F. E. McKenzie, "Ross, Macdonald, and a theory for the dynamics and control of mosquito-transmitted pathogens," PLoS Pathogens, vol. 8, no. 4, p. e1002588, 2012.

[7] H. W. Hethcote, "Three basic epidemiological models," in Applied Mathematical Ecology, pp. 119–144, Springer, 1989.

[8] L. J. Allen, F. Brauer, P. Van den Driessche, and J. Wu, Mathematical Epidemiology, vol. 1945. Springer, 2008.

[9] F. Brauer, "The Kermack–McKendrick epidemic model revisited," Mathematical Biosciences, vol. 198, no. 2, pp. 119–131, 2005.

[10] E. Forgoston and R. O. Moore, "A primer on noise-induced transitions in applied dynamical systems," SIAM Review, vol. 60, no. 4, pp. 969–1009, 2018.

[11] T. S. Brett, J. M. Drake, and P. Rohani, "Anticipating the emergence of infectious diseases," Journal of The Royal Society Interface, vol. 14, no. 132, p. 20170115, 2017.

[12] T. Harris, "The theory of branching processes. 1963."

[13] S. Singh and C. R. Myers, "Outbreak statistics and scaling laws for externally driven epidemics," Physical Review E, vol. 89, no. 4, p. 042108, 2014.

[14] J. Heesterbeek and K. Dietz, "The concept of R_0 in epidemic theory," Statistica neerlandica, vol. 50, no. 1, pp. 89–110, 1996.

[15] N. T. Bailey, The elements of stochastic processes with applications to the natural sciences, vol. 25. John Wiley & Sons, 1990.

[16] T. Smith, N. Maire, A. Ross, M. Penny, N. Chitnis, A. Schapira, A. Studer, B. Genton, C. Lengeler, F. Tediosi, et al., "Towards a comprehensive simulation model of malaria epidemiology and control," Parasitology, vol. 135, no. 13, pp. 1507–1516, 2008.

[17] D. T. Gillespie, "Exact stochastic simulation of coupled chemical reactions," The Journal of Physical Chemistry, vol. 81, no. 25, pp. 2340–2361, 1977.

[18] D. G. Kendall, "An artificial realization of a simple "birth-and-death" process," Journal of the Royal Statistical Society. Series B (Methodological), vol. 12, no. 1, pp. 116–119, 1950.

[19] O. N. Bjørnstad and B. T. Grenfell, "Noisy clockwork: time series analysis of population fluctuations in animals," Science, vol. 293, no. 5530, pp. 638–643, 2001.

[20] A. A. King, D. Nguyen, and E. L. Ionides, "Statistical inference for partially observed Markov processes via the R package pomp," arXiv preprint arXiv:1509.00503, 2015.

[21] E. L. Ionides, C. Bretó, and A. A. King, "Inference for nonlinear dynamical systems," Proceedings of the National Academy of Sciences, vol. 103, no. 49, pp. 18438–18443, 2006.

[22] E. L. Ionides, C. Bretó, and A. A. King, "Modeling disease dynamics: Cholera as a case study," 2008.

[23] D. He, E. L. Ionides, and A. A. King, "Plug-and-play inference for disease dynamics: measles in large and small populations as a case study," Journal of the Royal Society Interface, vol. 7, no. 43, pp. 271–283, 2009.

[24] A. Bhadra, E. L. Ionides, K. Laneri, M. Pascual, M. Bouma, and R. C. Dhiman, "Malaria in northwest India: Data analysis via partially observed stochastic differential equation models driven by Lévy noise," Journal of the American Statistical Association, vol. 106, no. 494, pp. 440–451, 2011.

[25] W. R. Gilks, S. Richardson, and D. Spiegelhalter, Markov chain Monte Carlo in practice. Chapman and Hall/CRC, 1995.

[26] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," IEEE Transactions on signal processing, vol. 50, no. 2, pp. 174–188, 2002.

[27] D. Salmond and N. Gordon, "An introduction to particle filters," State space and unobserved component models theory and applications, pp. 1–19, 2005.

[28] A. A. King, D. Nguyen, and E. L. Ionides, "Statistical inference for partially observed Markov processes via the R package pomp," Journal of Statistical Software, vol. 69, no. i12, 2016.

[29] J. Aitchison, S. Silvey, et al., "Maximum-likelihood estimation of parameters subject to restraints," The Annals of Mathematical Statistics, vol. 29, no. 3, pp. 813–828, 1958.

[30] E. Forgoston and I. B. Schwartz, "Predicting unobserved exposures from seasonal epidemic data," Bulletin of mathematical biology, vol. 75, no. 9, pp. 1450–1471, 2013.

[31] G. T. Nieddu, L. Billings, J. H. Kaufman, E. Forgoston, and S. Bianco, "Extinction pathways and outbreak vulnerability in a stochastic Ebola model," Journal of The Royal Society Interface, vol. 14, no. 127, p. 20160847, 2017.

[32] L. Billings and E. Forgoston, "Seasonal forcing in stochastic epidemiology models," Ricerche di Matematica, vol. 67, no. 1, pp. 27–47, 2018.