Hybrid methodology for Markovian epidemic models

Nicolas Rebuli

June 8, 2018

Thesis submitted for the degree of

Doctor of Philosophy

in

Applied Mathematics

at The University of Adelaide

Faculty of Engineering, Computer and Mathematical Sciences

School of Mathematical Sciences

Contents

Signed Statement **xvii**

Acknowledgements **xix**

Abstract **xxi**

1 Introduction **1**

2 Background **7**

2.1 Markov processes 7

2.2 Large-population approximations 13

2.3 The SIR CTMC 20

2.3.1 Outbreak duration 23

2.3.2 Final outbreak size 31

2.3.3 Inferring the basic reproductive number 35

2.4 The partially-observed SEIR CTMC 48

2.4.1 Inferring the basic reproductive number 50

3 Hybrid approximation of final size and duration distributions for the SIR CTMC **55**

3.1 Hybrid fluid model 58

3.1.1 Model formulation 58

3.1.2 Outbreak duration 62
3.1.3 Numerical results 65
3.1.4 Final outbreak size 67
3.1.5 Numerical results 68
3.2 Hybrid diffusion model 70
3.2.1 Model formulation 70
3.2.2 Final outbreak size 71
3.2.3 Numerical results 73
3.3 Implementation 75
3.3.1 Computing distributions 76
3.3.2 Computing a threshold 79
3.4 Discussion 83
4 Estimating the basic reproductive number (SIR) 85
4.1 Conditioned model 88
4.2 Hybrid diffusion model 90
4.3 Implementation 91
4.3.1 Validation of the conditioned model 94
4.3.2 Validation of the hybrid diffusion model 100
4.3.3 Application to A(H1N1)pdm09 data 103
4.4 Discussion 106
5 Estimating the basic reproductive number (SEIR) 109
5.1 Partially-observed SEIR CTMC model 111
5.1.1 Stage one: DA process 111
5.1.2 Stage two: truncated DA process 113
5.1.3 Stage three: diffusion approximation 117
5.2 Validation of the hybrid diffusion model 118

5.2.1 Base model 121
5.2.2 Restricted model 125
5.2.3 Full model 132
5.3 Application to data 139
5.3.1 A(H1N1)pdm09 in Western Australia 142
5.3.2 Ebola hemorrhagic fever in Zaire 1976 142
5.3.3 Ebola hemorrhagic fever in Congo 1995 144
5.3.4 Pneumonic Plague in Madagascar 2017 144
5.4 Discussion 147

6 Conclusion **151**
6.1 Summary 152
6.2 Future research 154

List of Tables

4.1 Parameters used for investigating our methodology. The removal rate (γ) and basic reproductive number (R_0) are representative of influenza and the population size (N) ensures that the susceptible pool is not depleted during the first two weeks of the epidemic. 92

List of Figures

2.1 State transitions of the SIR CTMC model. β is the effective force of infection and γ is the removal rate. 22

2.2 Example of how the exact likelihood is computed, using the observed incidence counts $x_1 = 2$ and $x_2 = 3$. The state transition diagrams display the truncated state spaces which contain: initial states (green), absorption states (red), states used to compute $L_E^1(\boldsymbol{\theta})$ and $L_E^2(\boldsymbol{\theta})$ (yellow), and ordinary transient states (blue). 38

2.3 State transitions of the partially-observed SEIR CTMC. 49

3.1 The state-transition diagram of the hybrid fluid model with $N = 15$ and $\widehat{I} = 3$. The green points are the discrete states in $\mathcal{Y}^{MC}$, and the continuum $\mathcal{Y}^{DE}$ is the set of states with $I \geq \widehat{I}$, and $S \leq N - I$. The upward (downward) pointing triangles are states from which $\boldsymbol{Y}(t)$ switches from CTMC to fluid (fluid to CTMC) dynamics, which are contained in the set $\mathcal{T}_1^{MC}$ ($\mathcal{T}_2^{MC}$). The black curves emanating from states in $\mathcal{T}_1^{MC}$ are the deterministic trajectories of $N\boldsymbol{x}(t, \boldsymbol{x}_0/N)$, for $\boldsymbol{x}_0$ in $\mathcal{T}_1^{MC}$, through $\mathcal{Y}^{DE}$. 61

3.2 The distribution of the duration of the epidemic calculated from the CTMC model, hybrid fluid model, and Barbour's model for $R_0 = 1.3$ and $N = 1,000$ with one initially infectious individual. 65

3.3 Required runtime of the distribution of the duration of the outbreak from the SIR CTMC and the hybrid fluid model alongside the L_1-error of the hybrid fluid model and Barbour's hybrid model. The L_1-error of the hybrid fluid model is less than the L_1-error of Barbour's hybrid model for $N \leq 2000$, but the two are virtually the same for $N \geq 10^3$. The required runtime of the hybrid fluid model if $\mathcal{O}(N)$. Again, we have that $R_0 = 1.3$, $\widehat{I} = 17$ (inequality (3.8)), and the initial state $(N-1,1)$. 66

3.4 The distribution of the final size of the epidemic calculated from the SIR CTMC, hybrid fluid model, hybrid diffusion model, and Scalia–Tomba for $R_0 = 1.3$ and $N = 1,000$ with one initially infectious individual. . . . 73

3.5 The required runtime of the SIR CTMC, hybrid fluid model and hybrid diffusion model alongside the L_1-error of the hybrid fluid model, hybrid diffusion model, and Scalia-Tomba's model. The error in the hybrid fluid model and the hybrid diffusion model is at-best a constant of $\mathcal{O}\left(10^0\right)$ and $\mathcal{O}\left(10^{-3}\right)$, respectively. The asymptotic slope of the runtime of the hybrid models (Algorithm 5) suggests that they are of computational complexity $\mathcal{O}(N)$ compared to the $\mathcal{O}(N^2)$ of the SIR CTMC (Algorithm 2). Here we used $R_0 = 1.3$ and $\widehat{I} = 17$ with the initial state $(N-1,1)$. 74

3.6 For the distribution of the final size (duration) of the outbreak, the ochre (green) curve with triangles (circles) shows the minimum threshold $\widehat{I}$ which achieves an L_1-error of 0.1 (0.25). The blue curve with squares shows the threshold determined by inequality (3.8) using $\epsilon = 5 \times 10^{-3}$ which achieves at most 0.1 (0.25) L_1-error in the distribution of the final size (duration) of the outbreak, provided R_0 is less than approximately 7.5. Here we used $N = 10,000$ and the initial state $(N-1,1)$. 81

3.7 The L_1-error of the distribution of the final size of the epidemic using inequality (3.8) to calculate the threshold. The error exceeds 0.1 on an interval of R_0 from approximately 7.5 to 11 and is at most 0.24. This issue arises when the fluid approximation of S falls below approximately eight susceptible individuals. Here we used $N = 10,000$ and the initial state $(N - 1, 1)$. 82

4.1 Marginal densities of the prior distribution of $1/\gamma$ and R_0. 93

4.2 Density estimates of the MLEs and MPEs of $(1/\gamma, R_0)$ obtained under Regimes 1 and 2. The rows contain estimates from the: unconditioned/conditioned DA process, unconditioned/conditioned hybrid process, and diffusion process. The density estimates demonstrate broad agreement between estimates of R_0 95

4.3 Density estimates of the MLEs and MPEs of $(1/\gamma, R_0)$ obtained under Regimes 3 and 4 from the conditioned DA process and conditioned hybrid process. 96

4.4 Bean plots of the estimated R_0 under Regimes 1 and 2. Bean plots are comprised of side-by-side box plots (where the whiskers represent the 2nd and 98th percentiles) plotted on top of a kernel density estimate. The conditioned estimate is smaller than the unconditioned estimate in every case. The unconditioned estimates in Regime 1 appear more biased than the unconditioned estimates in Regime 2. 97

4.5 Bean plots of the paired difference between estimates from the unconditioned DA process and the conditioned DA process in Regime 1 plotted against Regime 2, where the difference is defined as the unconditioned estimate minus the conditioned estimate. In all cases the conditioned estimates are smaller than the unconditioned estimates. 98

4.6 Bean plots of the paired difference in the conditioned DA process estimate of R_0 when the threshold is decreased from 50 to 20, where the difference is defined as the estimate from a threshold of 20 minus the estimate from a threshold of 50. The smaller conditioning level in Regimes 3 and 4 do less to reduce the positive-bias of the unconditioned estimate of R_0. . . . 100

4.7 Bean plots of the paired differences in the estimated R_0 from the unconditioned hybrid against the diffusion. The difference is defined as the estimate from the approximation minus the estimate from the unconditioned DA process. The hybrid approximation is more accurate than the diffusion approximation. 101

4.8 Bean plots of the paired differences between the conditioned DA estimate of R_0 and the conditioned hybrid estimate of R_0, where the difference is defined as the conditioned hybrid estimate minus the conditioned DA estimate. The hybrid approximation exhibits a small amount of bias. . . 102

4.9 Number of notified cases of A(H1N1)pdm09 from WA with box plots of the estimated distribution of R_0 from the conditioned and unconditioned hybrid process. The conditioned hybrid process estimates a lower R_0 than the unconditioned. 105

5.1 Statistical properties of the simulated outbreaks used for estimation by the base model. For each day, the cyan curve shows the median incidence count, the blue shaded area shows the central 5%–95% percentiles of the incidence counts and the red shaded area bounds biggest and smallest incidence counts. The simulated outbreak in yellow is used to inform the posterior distribution in Figure 5.3. 122

5.2 Prior distribution of R_0, $1/\gamma$ and $1/\alpha$ for the base model. This choice is similar to the prior distribution from the previous chapter and provides adequate support to parameter values which are reasonable for an influenza-like outbreak. 123

5.3 Joint posterior distribution of the parameters under the unconditioned base model from the first two weeks of the simulated outbreak shown in Figure 5.1. The true parameter values are displayed in green while the estimated parameters are displayed in red. 124

5.4 Bean plots of the point estimates of the parameters from the conditioned and unconditioned base model with the true values indicated by the black dashed line. The estimated values of $1/\gamma$ and $1/\alpha$ are reasonably accurate while the estimates of R_0 and p demonstrate an identifiability problem. . 127

5.5 Statistical properties of the simulated realisations used for estimation with the full model. For each day, the cyan curve shows the median incidence count, the blue shaded area shows the central 5%–95% percentiles of the incidence counts and the red shaded area bounds biggest and smallest incidence counts. The simulated outbreak shown in yellow is used to obtain the estimated posterior distribution in Figure 5.7. 128

5.6 Prior distribution of β_o/γ, β_u/γ, $1/\gamma$ and $1/\alpha$ for the restricted model. The prior distribution used here is the same as the prior distribution for the base model, with addition of an extra dimension for β_u/γ, which is similar to the prior distribution of β/γ in the base model. 129

5.7 Joint posterior distribution of the parameters based of the first two weeks of the simulated outbreak displayed in Figure 5.5. The true values of the parameters are displayed in green while the estimated parameters are displayed in red. The equation of R_0 provides the purple line which explains the correlation between β_o/γ and β_u/γ. 131

5.8 Bean plots of the estimated parameters from the unconditioned (blue) and conditioned (red) restricted model, with the true values indicated by the black dashed line. The estimated values of p and $1/\gamma$ are reasonably accurate. On average, the estimated value of R_0 is reasonably close to the true value. 134

5.9 Statistical properties of the simulated realisations used for estimation with the full model. For each day, the cyan curve shows the median incidence count, the blue shaded area shows the central 5%–95% percentiles of the incidence counts and the red shaded area bounds biggest and smallest incidence counts. The simulated realisation shown in yellow is used to estimate the joint posterior distribution in Figure 5.11. 135

5.10 Prior distribution of β_o/γ_o, β_u/γ, $1/\gamma_o$, $1/\gamma_u$ and $1/\alpha$ for the full model. The prior distribution used here is the same as the prior distribution for the restricted model, with addition of an extra dimension for $1/\gamma_u$ which is similar to the prior distribution for the restricted model. 136

5.11 Joint posterior distribution of the parameters based of the first three weeks of the simulated outbreak displayed in Figure 5.9. The true parameter values are displayed in green while the estimated parameters are displayed in red. 138

5.12 Bean plots of the point estimates of the parameters from the conditioned and unconditioned base model with the true values indicated by the black dashed line. The full model provides reasonably accurate estimates of all parameters aside from β_o/γ_o and β_u/γ_u. Despite this, the average estimated value of R_0 exhibits only a small amount of bias. 141

5.13 Daily incidence count and estimated basic reproductive number for an outbreak of A(H1N1)pdm09 in Western Australia using the conditioned full SEIR model with partial observations. The yellow line shown on the box plots indicates the MAPE of R_0 143

5.14 Daily incidence count and estimated basic reproductive number for an outbreak of the Ebola virus from 1976 in Zaire using the conditioned full SEIR model with partial observations. The yellow line shown on the box plots indicates the MAPE of R_0 145

5.15 Daily incidence count and estimated basic reproductive number for an outbreak of the Ebola virus from 1995 in Congo using the conditioned full SEIR model with partial observations. The yellow line shown on the box plots indicates the MAPE of R_0 146

5.16 Daily incidence count and estimated basic reproductive number for an outbreak of Plague from 2017 in Madagascar using the conditioned full SEIR model with partial observations. The yellow line shown on the box plots indicates the MAPE of R_0 148

For a thesis that does not contain work already in the public domain.

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

I give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Acknowledgements

Foremost, I would like to thank my supervisors Prof. Nigel Bean and Prof. Joshua Ross. From when I first met you, you have been a continuing source of encouragement for me to pursue further education. You have offered continued guidance and support during the good times, like when I published my first paper, and the tough times, like when I miscalculated the end date of my undergraduate degree. The wisdom, support, and opportunities that you have provided me with have helped me to grow into someone unrecognisable from the person I was at the start of my postgraduate degree. I will always be grateful for all the time and patience you have invested in me.

To my partner Georiga, you have beared the brunt of my postgraduate degree and for that I owe you a special thank you for continued love and support. You have been a great source of happiness and comfort for me throughout this experience. Without you, I would have been a lot more stressed, especially during the doldrums of writing my thesis. Thank you for putting up with me constantly being tired and busy.

To my family, you have been a continuing source of encouragement throughout my university degree. Although you have stopped asking me what exactly it is that I do, I know that I still have your support and that you are very proud of me.

To my friends Ben, David, Mingmei, Vincent and Vikram, you have made the last few years of my life some of the most memorable and in many ways you have helped me to grow just as much as my supervisors. Ben, thank you for your constant banter and good times. David, thank you for being able to provide wisdom in almost every facet of life. Mingmei, thank you for always being better prepared than me. Vincent, thank you for constantly reminding me that everything outside of pure mathematics is futile. Vikram, thank you for always having an interesting perspective on situations. It would not have been the same without you guys.

To all the people I met along the way; during my undergraduate degree, in the honours room, during my postgraduate degree, at conferences, or as my students. You have all contributed to my university experience in one way or another and no matter how small that contribution may be, I am grateful.

Finally, I thank the super computing resources provided by the Phoenix HPC service at the University of Adelaide for enabling me to perform the majority of the analyses presented in this thesis.

Abstract

In this thesis, we introduce a hybrid discrete-continuous approach suitable for analysing a wide range of epidemiological models, and an approach for improving parameter estimation from data describing the early stages of an outbreak. We restrict our attention to epidemiological models with *continuous-time Markov chain* (CTMC) dynamics, a ubiquitous framework also commonly used for modelling telecommunication networks, chemical reactions and evolutionary genetics. We introduce our methodology in the framework of the well-known *Susceptible–Infectious–Removed* (SIR) model, one of the simplest approaches for describing the spread of an infectious disease. We later extend it to a variant of the *Susceptible–Exposed–Infectious–Removed* (SEIR) model, a generalisation of the SIR CTMC that is more realistic for modelling the initial stage of many outbreaks.

Compartmental CTMC models are attractive due to their stochastic individual-to-individual representation of disease transmission. This feature is particularly important when only a small number of infectious individuals are present, during which stage the probability of epidemic fade out is considerable. Unfortunately, the simple SIR CTMC has a state space of order N^2, where N is the size of the population being modelled, and hence computational limits are quickly reached as N increases. There are a number of approaches towards dealing with this issue, most of which are founded on the principal

of restricting one's attention to the dynamics of the CTMC on a subset of its state space. However, two highly-efficient approaches published in 1970 and 1971 provide a promising alternative to these approaches.

The *fluid limit* [Kurtz, 1970] and *diffusion limit* [Kurtz, 1971] are large-population approximations of a particular class of CTMC models which approximate the evolution of the underlying CTMC by a deterministic trajectory and a Gaussian diffusion process, respectively. These large-population approximations are governed by a compact system of ordinary differential equations and are suitably accurate so long as the underlying population is sufficiently large. Unfortunately, they become inaccurate if the population of at least one compartment of the underlying CTMC is close to an absorbing boundary, such as during the initial stages of an outbreak. It follows that a natural approach to approximating a CTMC model of a large population is to adopt a *hybrid framework*, whereby CTMC dynamics are utilised during the initial stages of the outbreak and a suitable large-population approximation is utilised otherwise.

In the framework of the SIR CTMC, we present a *hybrid fluid model* and a *hybrid diffusion model* which utilise CTMC dynamics while the number of infectious individuals is low and otherwise utilises the fluid limit and the diffusion limit, respectively. We illustrate the utility of our hybrid methodology in computing two key quantities, the distribution of the duration of the outbreak and the distribution of the final size of the outbreak. We demonstrate that the hybrid fluid model provides a suitable approximation of the distribution of the duration of the outbreak and the hybrid diffusion model provides a suitable approximation of the distribution of the final size of the outbreak. In addition, we demonstrate that our hybrid methodology provides a substantial advantage in computational-efficiency over the original SIR CTMC and is superior in accuracy to similar hybrid large-population approaches when considering mid-sized populations.

During the initial stages of an outbreak, calibrating a model describing the spread of the disease to the observed data is fundamental to understanding and potentially controlling the disease. A key factor considered by public health officials in planning their response to an outbreak is the transmission potential of the disease, a factor which is informed by estimates of the *basic reproductive number*, R_0, defined as the average number of secondary cases resulting from a single infectious case in a naive population. However, it is often the case that estimates of R_0 based on data from the initial stages of an outbreak are positively biased. This bias may be the result of various features such as the geography and demography of the outbreak. However, a consideration which is often overlooked is that the outbreak was not detected until such a time as it had established a considerable chain of transmissions, therefore effectively overcoming initial fade out. This is an important feature because the probability of initial fade out is often considerable, making the event that the outbreak becomes established somewhat unlikely. A straightforward way of accounting for this is to *condition* the model on a particular event, which models the disease overcoming initial fade out.

In the framework of both the SIR CTMC and the SEIR CTMC we present a conditioned approach to estimating R_0 from data on the initial stages of an outbreak. For the SIR CTMC, we demonstrate that in certain circumstances, conditioning the model on effectively overcoming initial fade out reduces bias in estimates of R_0 by 0.3 on average, compared to the original CTMC model. Noting that the conditioned model utilises CTMC dynamics throughout, we demonstrate the flexibility of our hybrid methodology by presenting a conditioned hybrid diffusion approach for estimating R_0. We demonstrate that our conditioned hybrid diffusion approach still provides estimates of R_0 which exhibit less bias than under an unconditioned hybrid diffusion model, and that the diffusion methodology enables us to consider larger outbreaks then would have been computationally-feasible in the original conditioned CTMC framework. We demonstrate the flexibility of our conditioned hybrid approach by applying it to a variant of the SEIR CTMC and using it to estimate R_0 from a range of real outbreaks. In so doing, we utilise a truncation rule to ensure the initial CTMC dynamics are computationally-feasible.

Chapter 1

Introduction

We live in a time of relative comfort and stability where modern science has afforded us a means of defending ourselves against most infectious diseases, in some regions even going so far as eradicating some diseases. However, despite our many strides forward, life-threatening diseases are endemic in many regions, antimicrobial resistant *superbugs* are a pandemic of increasing threat to our ability to effectively treat a wide range of diseases, and numerous novel outbreaks of international concern have occurred over just the last decade, such as the Ebola virus epidemic in West Africa, the Zika virus epidemic in the Americas and the A(H1N1) influenza virus which reached pandemic proportions. It follows that infectious diseases are an ongoing threat to humans which require continued attention.

Mathematical epidemiology is the field that applies mathematical and statistical analyses of infectious diseases to further our understanding of the dynamics of disease spread, typically with the aim of controlling or preventing their advance. A particularly important branch of mathematical epidemiology is concerned with modelling the spread of a disease in a population of individuals using models that describe the transformations of individuals between different epidemiological states (so called compartmental models). In this thesis we are concerned with a particularly common framework where the dynamics of the disease are described by a *continuous-time Markov chain* (CTMC). This framework suitably describes the stochastic individual-to-individual nature of disease transmission, which is crucial for accurately representing the early stages of a novel outbreak. The main problem with the CTMC framework is that once an outbreak has become established, CTMC dynamics are typically computationally-intractable for analysis if the underlying population is large. To avoid this problem, it is common to consider approximating the CTMC by a suitable large-population approximation such as the *fluid limit* [Kurtz, 1970] or the *diffusion limit* [Kurtz, 1971]. Thus, a natural approach would be to model the early spread of the disease with CTMC dynamics and the dynamics thereafter by a large-population approximation. This so-called *hybrid* framework is the basis of much of the work presented in this thesis.

During a novel outbreak, calibrating a model describing the spread of the disease to observed data is fundamental to understanding and potentially controlling the disease. A key quantity which is of interest to public health authorities is the *basic reproductive number*, R_0, which is defined as the average number of secondary cases of the disease as the result of an introduction of a single infectious individual in an otherwise susceptible population. An accurate and reliable estimate of R_0 characterises the transmission potential of the disease, an important factor for public health authorities in planning their response to the outbreak. However, estimates of R_0 which are based on data from the initial stages of an outbreak are commonly positively-biased [Mercer et al., 2011, Rida, 1991]. A commonly over-looked cause of this bias is the probability of initial fade out, defined as the probability that the outbreak ends before becoming established. During the initial stages of an outbreak, the probability of initial fade out decreases considerably each time the number of infectious individuals increases. Thus, from a modelling perspective, the event that an outbreak effectively overcomes initial fade out can often be considered unlikely. At the same time, an outbreak will often not be detected by public health authorities until such a time that it has established an appreciable chain of transmission, thereby effectively overcoming initial fade out. It follows that the event that an outbreak becomes established, and is consequently detected by public health authorities, is one which needs to be accounted for in estimating the basic reproductive number during the early stages of an outbreak to reduce bias. Our so-called *conditioning* framework is also the basis of much of the work presented in this thesis.

In Chapter 3, we consider the well-known *Susceptible-Infectious-Removed* (SIR) epidemic model. We illustrate our hybrid methodology by presenting a *hybrid fluid model* and a *hybrid diffusion model* of the SIR CTMC, so called after the large population approximation they utilise. Our hybrid models utilise CTMC dynamics if the number of infectious individuals is low and the dynamics of the fluid approximation or the diffusion approximation otherwise. We demonstrate the utility of our hybrid models by using them to compute two key quantities of an outbreak, the distribution of the duration of the outbreak and the distribution of the final size of the outbreak. The duration of the outbreak is defined as the duration from when the first individual becomes infectious to the time at which the final infectious individual is removed. Similarly, the final size of the outbreak is defined as the total number of individuals who experience infection from the initial infectious individual to the final infectious individual to be removed. We compare our approximations of these distributions to the original SIR CTMC and to two other models which utilise a different hybrid approach [Barbour, 1975, Scalia-Tomba, 1985].

In Chapter 4, we consider using the SIR CTMC to infer the basic reproductive number, R_0, of a novel outbreak based on observed daily incidence counts. We illustrate our conditioning framework by presenting a conditioned version of the SIR CTMC in which the number of infectious individuals is required to reach a pre-defined level. Through a simulation study of outbreaks with influenza-like dynamics, we demonstrate that our approach generally reduces the bias in estimates of R_0. Furthermore, we demonstrate the utility of our hybrid diffusion approach by applying it to our conditioned SIR CTMC, referring to the resulting approach as the *conditioned hybrid diffusion* approach. In considering an outbreak of A(H1N1)pdm09, we demonstrate that our hybrid methodology enables us to consider larger populations than would have been possible in the framework of the SIR CTMC, while still providing the advantages of conditioning.

In Chapter 5, we apply our conditioned hybrid diffusion approach to a CTMC model which is more appropriate for representing the early stages of an outbreak. We consider the so-called partially-observed SEIR CTMC, which differs from the SIR CTMC in its inclusion of an *Exposed* (E) compartment and the condition that infectious individuals are *observed* with probability p. In a simulation study of outbreaks with influenza-like dynamics, we demonstrate that our hybrid approach provides accurate estimates of the basic reproductive number, the average latent period and the average infectious period. Furthermore, we demonstrate that conditioning consistently reduces bias in the estimates of R_0 and our hybrid approach enables us to consider larger populations than would have been possible in the CTMC framework. Finally, we demonstrate the utility of our approach by using it to estimate R_0 from a range of real outbreaks.

A paper concerning the hybrid approximations detailed in Chapter 3 has been published in the Journal of Mathematical Biology [Rebuli et al., 2016] and a paper concerning the conditioned hybrid approach detailed in Chapter 4 has been published in Theoretical Population Biology [Rebuli et al., 2018]. A paper concerning the application of our methodology to the partially-observed SEIR CTMC detailed in Chapter 5 has been submitted for publication.

Chapter 2

Background

This thesis focuses mainly on the development of computationally-efficient routines for inferring certain properties of the *basic reproductive number*, R_0, from real-world outbreaks, using Markovian epidemic models. In this chapter, we start by defining a *continuous-time Markov chain* (CTMC), and then discuss the fluid and diffusion large-population approximations, which apply to a certain class of CTMCs. We then define the *Susceptible-infectious-Removed* (SIR) CTMC and discuss computing the distribution of the duration of the outbreak, and the distribution of the final size of the outbreak. Finally, we present methods for inferring properties of R_0 from daily incidence data, observed during the early stages of an outbreak. We do this initally in the framework of the SIR CTMC and then in the framework of a more realistic partially-observed SEIR CTMC.

2.1 Markov processes

A stochastic process is a mathematical model for describing the evolution of a random phenomenon through time. Every stochastic process is characterised by a *state space*, an initial distribution and a *probability law* describing how the process evolves, as well as a set of observed *outcomes*. *Continuous-time Markov chains* (CTMC)s are a class of continuous-time stochastic processes whose state space is finite or countably infinite (although we assume the former herein), and satisfy the *Markov property*.

Definition 1 (The Markov Property) *Let* $(\boldsymbol{X}(t), t \geq 0)$*, be a continuous-time stochastic process taking values* $\boldsymbol{x}$ *in the state space* $\mathcal{X}$*. Then the CTMC satisfies the Markov property if*

$$\Pr\left(\boldsymbol{X}(t) = \boldsymbol{y} \,|\, \boldsymbol{X}(s) = \boldsymbol{x}, \boldsymbol{X}(u), u \leq s\right) = \Pr\left(\boldsymbol{X}(t) = \boldsymbol{y} \,|\, \boldsymbol{X}(s) = \boldsymbol{x}\right),$$

for all non-negative real numbers $t > s$ *and all* $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *in* $\mathcal{X}$*.*

The Markov property means that the future evolution of the process is conditionally independent of the history of the process, given the most recent observation of the state. For this reason, the Markov property is sometimes referred to as the *memoryless* property.

A common assumption in population modelling is that a CTMC is *time-homogeneous.*

Definition 2 (Time-homogeneous) *A CTMC,* $(\boldsymbol{X}(t), t \geq 0)$*, is time-homogeneous if, for all* $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *in* $\mathcal{X}$ *and* $s, t \in [0, \infty)$*, the probability*

$$\Pr\left(\boldsymbol{X}(t+s) = \boldsymbol{y} \,|\, \boldsymbol{X}(s) = \boldsymbol{x}\right),$$

is independent of s*, in which case we have that*

$$\Pr\left(\boldsymbol{X}(t+s) = \boldsymbol{y} \,|\, \boldsymbol{X}(s) = \boldsymbol{x}\right) = \Pr\left(\boldsymbol{X}(t) = \boldsymbol{y} \,|\, \boldsymbol{X}(0) = \boldsymbol{x}\right).$$

Time-homogeneity means that the probability of the CTMC transitioning from the state $\boldsymbol{x}$ to the state $\boldsymbol{y}$ depends only on t, the duration of time that has elapsed, and not the absolute time $t+s$. In this case, for all $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathcal{X}$ and $t \geq 0$, we define the *transition probabilities* as

$$p^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{y}}(t) = \Pr\left(\boldsymbol{X}(t) = \boldsymbol{y} \,|\, \boldsymbol{X}(0) = \boldsymbol{x}\right).$$

Note that we herein adopt the convention that the superscript notation $f^{\boldsymbol{X}}$, indicates that the quantity f depends on the process $(\boldsymbol{X}(t), t \geq 0)$. The vector of transition probabilities $\boldsymbol{p}^{\boldsymbol{X}}_{\boldsymbol{x}}(t) = (p^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{y}}(t) \,:\, \boldsymbol{y} \in \mathcal{X})$ is a probability mass function describing the probability that the CTMC is in each state $\boldsymbol{y}$ of $\mathcal{X}$ at time t, given that the CTMC was initially in the state $\boldsymbol{x}$ at time 0.

As a result of the Markov property (Definition 1) and time-homogeneity (Definition 2), a CTMC satisfies the *Chapman–Kolmogorov equations.*

Theorem 1 (Chapman–Kolmogorov Equations) *For all $\boldsymbol{x}$ and $\boldsymbol{z}$ in $\mathcal{X}$ and $t \geq 0$, a time-homogeneous CTMC, $(\boldsymbol{X}(t), t \geq 0)$, satisfies the* Chapman–Kolmogorov equations

$$p^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{z}}(t) = \sum_{\boldsymbol{y} \in \mathcal{X}} p^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{y}}(s)\, p^{\boldsymbol{X}}_{\boldsymbol{y}\,\boldsymbol{z}}(t-s),$$

for any $0 < s < t$.

The Chapman–Kolmogorov equations state that the probability of the CTMC being in the state $\boldsymbol{z}$ at time t can be computed by considering the probability moving from state $\boldsymbol{x}$ to state $\boldsymbol{y}$ in s time units and then independently moving from $\boldsymbol{y}$ to $\boldsymbol{z}$ in the remaining $t-s$ time units, and summing over all possible states $\boldsymbol{y}$.

A CTMC is usually characterised by its *transition rates*, which describe the behaviour of the process over an infinitesimal time interval, h. The *generator matrix* is the matrix which contains all the transition rates of the CTMC.

Definition 3 (Generator Matrix) *The transition rates of the CTMC,* $(\boldsymbol{X}(t), t \geq 0)$*, for all* $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *in* $\mathcal{X}$ *with* $\boldsymbol{y} \neq \boldsymbol{x}$*, are defined as*

$$\begin{aligned} q^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{y}} &= \lim_{h \to 0^+} \frac{p^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{y}}(h)}{h}, \\ q^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{x}} &= \lim_{h \to 0^+} \frac{p^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{x}}(h) - 1}{h}. \end{aligned}$$

The generator matrix is then $\mathbb{Q}^{\boldsymbol{X}} = \left[q^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{y}}\right]$.

For all $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathcal{X}$ with $\boldsymbol{y} \neq \boldsymbol{x}$, the transition rates have the properties:

$$\begin{aligned} q^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{y}} &\geq 0, \\ q^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{x}} &= -\sum_{\substack{\boldsymbol{y} \in \mathcal{X} \\ \boldsymbol{y} \neq \boldsymbol{x}}} q^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{y}}. \end{aligned}$$

We impose the additional constraint that $\left|q^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{x}}\right| < \infty$. By convention, one typically denotes $q^{\boldsymbol{X}}_{\boldsymbol{x}} = |q^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{x}}|$.

Loosely speaking, the transition rates are the right-derivatives of the transition probabilities at the point $h = 0$. A relationship between the two is encapsulated by the *Kolmogorov Equations*.

Definition 4 (Kolmogorov Equations) *It follows from the Chapman–Kolmogorov equations (Theorem (1)) that, for all* $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *in* $\mathcal{X}$ *with* $\boldsymbol{y} \neq \boldsymbol{x}$*, the transition probabilities of* $(\boldsymbol{X}(t), t \geq 0)$ *satisfy the Kolmogorov Backward Differential Equations (KBDE)s*

$$\frac{d}{dt}\left[p^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{z}}(t)\right] = \sum_{\boldsymbol{y} \in \mathcal{X}} q^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{y}}\, p^{\boldsymbol{X}}_{\boldsymbol{y}\,\boldsymbol{z}}(t),$$

and the Kolmogorov Forward Differential Equations (KFDE)s

$$\frac{d}{dt}\left[p^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{z}}(t)\right] = \sum_{\boldsymbol{y} \in \mathcal{X}} p^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{y}}(t)\, q^{\boldsymbol{X}}_{\boldsymbol{y}\,\boldsymbol{z}}.$$

The Kolmogorov equations provide a system of linear differential equations describing the evolution of the transition probabilities of the CTMC. This information is encapsulated in the *transition probability matrix*, which is defined as $\mathbb{P}^{\boldsymbol{X}}(t) = \left[p^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{y}}(t)\right]$. It can be seen that the KBDEs and KFDEs may be written in matrix form as $\frac{d}{dt}\mathbb{P}^{\boldsymbol{X}}(t) = \mathbb{Q}^{\boldsymbol{X}}\mathbb{P}^{\boldsymbol{X}}(t)$ and $\frac{d}{dt}\mathbb{P}^{\boldsymbol{X}}(t) = \mathbb{P}^{\boldsymbol{X}}(t)\,\mathbb{Q}^{\boldsymbol{X}}$, respectively. Provided $\mathbb{Q}^{\boldsymbol{X}}$ is *conservative* and *regular*, the Kolmogorov equations have the unique solution, $\mathbb{P}^{\boldsymbol{X}}(t) = \mathbb{P}^{\boldsymbol{X}}(0)\,e^{t\mathbb{Q}^{\boldsymbol{X}}}$, where the matrix exponential e^M is defined as $\sum_{k=0}^{\infty} M^k/(k!)$. The requirement that the transition rate matrix is regular and conservative are satisfied trivially since $\mathcal{X}$ is finite, and all the transition rates are finite [Feller, 1940, Kato, 1954]. Although the KBDEs and KFDEs both provide the same solution in this situation, the KFDEs are generally more amenable to analysis. Thus, we herein refer to the KFDEs as the Kolmogorov equations.

An important concept for time-homogeneous CTMCs (Definition 2) is the *embedded jump process.*

Definition 5 (Embedded Jump Process) *The embedded jump process of* $(\boldsymbol{X}(t), t \geq 0)$ *is the discrete-time Markov chain (DTMC)* $(\boldsymbol{X}_n, n \geq 0)$*, which takes values in* $\mathcal{X}$ *and, for all* $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *in* $\mathcal{X}$ *with* $\boldsymbol{y} \neq \boldsymbol{x}$*, has the jump probabilities*

$$p^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{y}} = \frac{q^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{y}}}{q^{\boldsymbol{X}}_{\boldsymbol{x}}}, \tag{2.1}$$

where $p^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{x}} = 0$*.*

The embedded jump process may be thought of as a time-independent representation of the CTMC, in which the probability of transition to each different state is given by the relative frequency of that transition. The embedded jump process is useful for computing time-independent quantities, such as *hitting probabilities.*

Definition 6 (Hitting Probabilities) *Let $h^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{z}}$ denote the probability that $(\boldsymbol{X}(t), t \geq 0)$ ever hits the state $\boldsymbol{z}$ in $\mathcal{X}$, given the initial state $\boldsymbol{x}$ in $\mathcal{X}$. Then, for all $\boldsymbol{x}$ and $\boldsymbol{z}$ in $\mathcal{X}$ with $\boldsymbol{z} \neq \boldsymbol{x}$, the hitting probabilities are the minimal non-negative solution to the system of equations*

$$h^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{z}} = \sum_{\boldsymbol{y} \in \mathcal{X}} p^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{y}} \, h^{\boldsymbol{X}}_{\boldsymbol{y}\,\boldsymbol{z}}, \tag{2.2}$$

where $h^{\boldsymbol{X}}_{\boldsymbol{z}\,\boldsymbol{z}} = 1$.

The hitting probabilities may be generalised to provide the probability that the CTMC ever hits the set $\mathcal{A}$, by modifying the boundary condition of equation (2.2) to $h^{\boldsymbol{X}}_{\boldsymbol{y}\,\boldsymbol{y}} = 1$, for all $\boldsymbol{y}$ in $\mathcal{A}$. As we shall see in Chapter 3, the hitting probabilities are particularly useful for modelling epidemics. Furthermore, the hitting probabilities may be used to *condition* the CTMC on hitting a particular state $\boldsymbol{z}$, or set of states [Waugh, 1958].

Theorem 2 (Conditioned Markov Processes) *Let $\mathcal{A}$ be the subset of $\mathcal{X}$ from which every state $\boldsymbol{x}$ in $\mathcal{X}$ has a non-zero hitting probability of the state $\boldsymbol{z}$. Then the CTMC $(\boldsymbol{X}(t), t \geq 0)$ taking values in $\mathcal{A}$ is conditioned on hitting the state $\boldsymbol{z}$ by modifying its transition rates such that, for all $\boldsymbol{x} \in \mathcal{A}$ and $\boldsymbol{y}$ in $\mathcal{X}$,*

$$\tilde{q}^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{y}} = \left(\frac{h^{\boldsymbol{X}}_{\boldsymbol{y}\,\boldsymbol{z}}}{h^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{z}}} \right) q^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{y}}.$$

Conditional Markov processes utilise the law of conditional probability to ensure a particular event occurs for the CTMC with probability 1. This may be thought of as a way of removing all the sample paths of the CTMC for which this event does not occur. We utilise this result in Chapter 4 and Chapter 5 to condition the CTMC on the event that the outbreak becomes established.

As we have now seen, the Kolmogorov equations are fundamental for analysing the behaviour of a CTMC because they describe the evolution of the process. However, in most applications numerically solving the Kolmogorov equations directly is computationally-intractable because the number of equations arising in Definition 4 generally depends on N^d, where N is a population ceiling and d is the number of dimensions in the CTMC. Although much attention has been given to this problem [Moler and Charles, 2003, Jenkinson and Goutsias, 2012], most alternatives are still computationally-intractable for the kinds of population ceilings required in epidemiology. Thus, in the following section we discuss two large-population approximations which avoid the need to deal directly with the Kolmogorov equations.

2.2 Large-population approximations

In this section we define the fluid limit [Kurtz, 1970] and diffusion limit [Kurtz, 1971]. We begin by restricting our attention to the class of CTMCs referred to as *population processes* [Barbour, 1972, 1974, Kurtz, 1976, Barbour, 1976, Pollett and Vassallo, 1992, Pollett, 1990].

Definition 7 (Population Process) $(\boldsymbol{X}(t), t \geq 0)$ *is a population process if:*

1. *Each state $\boldsymbol{x}$ in $\mathcal{X}$ partitions a finite population of N individuals into a finite number of compartments.*

2. *The only positive transition rates, $q^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{x}+\boldsymbol{\ell}}$, are ones for which $\boldsymbol{\ell}$ is either*

$$\pm\boldsymbol{e}_i, \qquad or \qquad \boldsymbol{e}_i - \boldsymbol{e}_j,$$

where $\boldsymbol{e}_i$ is the unit vector, with a 1 as its ith element.

A population process may be thought of as a CTMC model in which every individual in the population falls into one distinct compartment. Thus, the elements of the state space of the CTMC denote the number of individuals who are in each compartment, and the possible transitions of the CTMC reflect the event that an individual arrives/departs from compartment i $(\pm\boldsymbol{e}_i)$, or an individual transitions from compartment j to compartment i $(\boldsymbol{e}_i - \boldsymbol{e}_j)$. Most CTMC models used in epidemiology are population processes in which the compartments reflect an individual's stages of infection and the transitions of the model represent events such as an individual becoming infectious or an individual recovering.

The original fluid limit and diffusion limit [Kurtz, 1970, 1971] applied to the class of CTMCs referred to as "density dependent". However, they were subsequently extended by Pollett [1990] to a broader class of CTMCs which he referred to as "asymptotically density dependent". We refer to a CTMC which satisfies either definition as being "density dependent".

The definition of density dependence refers to the family of CTMCs $(\boldsymbol{X}^{(\nu)}(t), t \geq 0)$, $\nu \geq 0$, which take values in $\mathcal{X}^{(\nu)}$. This is simply a way of making the relationship between the CTMC $(\boldsymbol{X}(t), t \geq 0)$ and a particular *scaling parameter* $\nu > 0$ explicit and the scaling parameter is usually taken as the population ceiling N.

Definition 8 (Density Dependence) *Suppose $(\boldsymbol{X}^{(\nu)}(t), t \geq 0)$, $\nu \geq 0$, has a corresponding family of continuous functions $f^{(\nu)}(\boldsymbol{x}, \boldsymbol{\ell})$, for $\boldsymbol{x}$ in E with $E \subseteq \mathbb{R}^K$ and $K \in \mathbb{Z}_+$, such that*

$$q^{\boldsymbol{X}^{(\nu)}}_{\boldsymbol{y}\,\boldsymbol{y}+\boldsymbol{\ell}} = \nu\, f^{(\nu)}(\boldsymbol{y}/\nu, \boldsymbol{\ell}),$$

for all $\boldsymbol{y}$ in $\mathcal{X}^{(\nu)}$ and $\boldsymbol{\ell} \neq \boldsymbol{0}$. In which case, define

$$F^{(\nu)}(\boldsymbol{x}) = \sum_{\boldsymbol{\ell}} \boldsymbol{\ell} f^{(\nu)}(\boldsymbol{x}, \boldsymbol{\ell}).$$

Then the family of CTMCs is said to be (asymptotically) density dependent if there exists a continuous function $F(\boldsymbol{x})$, such that

$$\lim_{\nu\to\infty} F^{(\nu)}(\boldsymbol{x}) = F(\boldsymbol{x}).$$

Loosely speaking, a density dependent CTMC is one whose transition rates depend on the current state $\boldsymbol{y}$ only through the density $\boldsymbol{y}/\nu$. We refer to the class of CTMCs which are both population processes (Definition 7) and density dependent (Definition 8) as *density dependent Markov population processes* (DDMPP)s.

Based on the observation that the behaviour of the DDMPP, scaled by ν, is increasingly like that of a deterministic process as $\nu \to \infty$, Kurtz [1970] showed that $\boldsymbol{X}^{(\nu)}(t)/\nu$, for $0 \le t < \infty$, converges uniformly in probability (over finite time intervals) to a unique deterministic trajectory $\boldsymbol{x}(t, \boldsymbol{x}_0)$ with time derivative $F(\boldsymbol{x})$, for $\boldsymbol{x}_0$ in E. The following theorem is due to Pollett [1990].

Theorem 3 (Fluid Limit) *Suppose $F(\boldsymbol{x})$ is Lipschitz continuous on E and that for all $\nu > 0$*

$$\sup_{\boldsymbol{x}\in E} \sum_{\boldsymbol{\ell}} |\boldsymbol{\ell}|\, f^{(\nu)}(\boldsymbol{x}, \boldsymbol{\ell}) < \infty, \tag{2.3}$$

$$\lim_{\delta\to\infty} \sup_{\boldsymbol{x}\in E} \sum_{\boldsymbol{\ell}:|\boldsymbol{\ell}|>\delta} |\boldsymbol{\ell}|\, f^{(\nu)}(\boldsymbol{x}, \boldsymbol{\ell}) = 0, \tag{2.4}$$

and

$$\lim_{\nu\to\infty} \sup_{\boldsymbol{x}\in E} \left| F^{(\nu)}(\boldsymbol{x}) - F(\boldsymbol{x}) \right| = 0. \tag{2.5}$$

Then, if

$$\lim_{\nu\to\infty} \frac{\boldsymbol{X}^{(\nu)}(0)}{\nu} = \boldsymbol{x}_0, \tag{2.6}$$

for finite t, we have that

$$\lim_{\nu\to\infty} \Pr\left(\sup_{s\le t} \left| \frac{\boldsymbol{X}^{(\nu)}(s)}{\nu} - \boldsymbol{x}(s, \boldsymbol{x}_0) \right| > \epsilon \right) = 0, \qquad 0 \le s \le t,$$

for all $\epsilon > 0$, and for every trajectory $\boldsymbol{x}\left(\cdot, \boldsymbol{x}_0\right)$ satisfying

$$\begin{aligned} &\boldsymbol{x}\left(0, \boldsymbol{x}_0\right) = \boldsymbol{x}_0, \\ &\boldsymbol{x}\left(s, \boldsymbol{x}_0\right) \in E, \qquad 0 \le s \le t, \\ &\frac{\partial}{\partial s}\boldsymbol{x}\left(s, \boldsymbol{x}_0\right) = F\left(\boldsymbol{x}\left(s, \boldsymbol{x}_0\right)\right). \end{aligned}$$

For a DDMPP, condition (2.3) is satisfied because $\boldsymbol{\ell}$ is a linear combination of unit vectors and $|q^{\boldsymbol{X}}_{\boldsymbol{x}\boldsymbol{y}}| < \infty$, for all $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathcal{X}$. Similarly, condition (2.4) is satisfied because $\boldsymbol{\ell}$ is a linear combination of unit vectors. Condition (2.5) requires that $F^{(N)}\left(\boldsymbol{x}\right)$ converges to $F\left(\boldsymbol{x}\right)$ as $N \to \infty$ (Definition 8). Condition (2.6) requires that the initial value $\boldsymbol{x}_0$ in E, is "close" to $\boldsymbol{X}^{(N)}(0)/N$. The theorem then stipulates that $\boldsymbol{X}^{(N)}(t)/N$, converges in probability over finite time intervals to the unique deterministic trajectory $\boldsymbol{x}\left(t, \boldsymbol{x}_0\right)$, for $\boldsymbol{x}_0$ in E. Herein, we refer to the *fluid approximation* of a DDMPP as the unique deterministic trajectory $N\boldsymbol{x}\left(t, \boldsymbol{x}_0\right)$, where $\boldsymbol{x}_0$ is given by $\boldsymbol{X}^{(N)}(0)/N$.

The fluid approximation is useful for describing the average behaviour of a DDMPP but it provides no indication of its variability, for this we appeal to the diffusion limit [Kurtz, 1971].

Theorem 4 (Diffusion Limit) *Suppose that $F\left(\boldsymbol{x}\right)$ is bounded and Lipschitz continuous on E. Suppose also that the family of continuous functions $G^{(\nu)}\left(\boldsymbol{x}\right)$, where $\nu > 0$ and $\boldsymbol{x}$ is in E, is a $K \times K$ matrix, where dim$(\boldsymbol{X}^{(\nu)}(t)) = K$, with elements*

$$g^{(\nu)}_{i,j}\left(\boldsymbol{x}\right) = \sum_{\boldsymbol{\ell}} \ell_i \ell_j f^{(\nu)}\left(\boldsymbol{x}, \boldsymbol{\ell}\right),$$

where ℓ_i denotes the ith entry of the vector $\boldsymbol{\ell}$, which converges uniformly to $G\left(\boldsymbol{x}\right)$, where $G\left(\boldsymbol{x}\right)$ is bounded and uniformly continuous on E.

If, in addition,

$$\sup_{\boldsymbol{x}\in E}\sum_{\boldsymbol{\ell}}|\boldsymbol{\ell}|^2 f^{(\nu)}(\boldsymbol{x},\boldsymbol{\ell})<\infty, \tag{2.7}$$

$$\lim_{\delta\to\infty}\sup_{\boldsymbol{x}\in E}\sum_{\boldsymbol{\ell}:|\boldsymbol{\ell}|>\delta}|\boldsymbol{\ell}|^2 f^{(\nu)}(\boldsymbol{x},\boldsymbol{\ell})=0, \tag{2.8}$$

for all $\nu>0$*, and*

$$\lim_{\nu\to\infty}\sup_{\boldsymbol{x}\in E}\nu^{\frac{1}{2}}\left|F^{(\nu)}(\boldsymbol{x})-F(\boldsymbol{x})\right|=0, \tag{2.9}$$

where now $F(\boldsymbol{x})$ *is assumed to have uniformly continuous first partial derivatives, then, provided*

$$\lim_{\nu\to\infty}\nu^{\frac{1}{2}}\left(\frac{\boldsymbol{X}^{(\nu)}(0)}{\nu}-\boldsymbol{x}_0\right)=\boldsymbol{z}, \tag{2.10}$$

for $\boldsymbol{x}_0$ *in* E*, the family of Markov processes* $\boldsymbol{Z}^{(\nu)}(t)$*, for* $t\geq 0$*, defined by*

$$\boldsymbol{Z}^{(\nu)}(s)=\nu^{\frac{1}{2}}\left(\frac{\boldsymbol{X}^{(\nu)}(s)}{\nu}-\boldsymbol{x}(s,\boldsymbol{x}_0)\right),\qquad 0\leq s\leq t,$$

converges weakly in $D[0,t]$ *(the space of right-continuous, left-hand limits functions on* $[0,t]$*) to a diffusion process,* $\boldsymbol{Z}(t)$*, with initial value* $\boldsymbol{Z}(0)=\boldsymbol{z}$ *and with characteristic function,* $\Psi=\Psi(s,\boldsymbol{\theta})$ *which satisfies*

$$\begin{aligned}\frac{\partial}{\partial s}[\Psi(s,\boldsymbol{\theta})]=-\frac{1}{2}\sum_{j,k}\boldsymbol{\theta}_j g_{jk}(\boldsymbol{x}(s,\boldsymbol{x}_0))\,\boldsymbol{\theta}_k\Psi(s,\boldsymbol{\theta})&\\ +\sum_{j,k}\boldsymbol{\theta}_j\frac{\partial}{\partial x_k}[F_j(\boldsymbol{x}(s,\boldsymbol{x}_0))]\frac{\partial}{\partial\boldsymbol{\theta}_k}[\Psi(s,\boldsymbol{\theta})].&\end{aligned} \tag{2.11}$$

For a DDMPP, condition (2.7) is satisfied because $\boldsymbol{\ell}$ is a linear combination of unit vectors and $|q^{\boldsymbol{X}}_{\boldsymbol{x}\boldsymbol{y}}|<\infty$, for all $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathcal{X}$. Similarly, condition (2.8) is satisfied because $\boldsymbol{\ell}$ is a linear combination of unit vectors. Condition (2.9) strengthens condition (2.5) to ensure that $F^{(\nu)}(\boldsymbol{x})$ converges to $F(\boldsymbol{x})$ at the correct rate. Condition (2.10) provides the initial state of the diffusion. The theorem then stipulates that a $\sqrt{\nu}$ scaling of the difference $\boldsymbol{X}^{(\nu)}(t)/\nu - \boldsymbol{x}\left(t, \boldsymbol{x}_0\right)$ converges weakly (in the space of right-continuous, left-hand limit functions on $[0,t]$) to the diffusion $\boldsymbol{Z}(t)$, as $\nu \to \infty$.

Although the partial differential equation (2.11) specifies the distribution of the diffusion $(\boldsymbol{Z}(t), t \geq 0)$, only in special cases can one obtain an explicit expression for its characteristic function. However, one is always able to obtain its expected value and covariance. In particular, if one denotes by $\nabla F\left(\boldsymbol{x}\right)$ the matrix of first partial derivatives of $F\left(\boldsymbol{x}\right)$, that is $[\partial F_i/\partial x_j]$, and puts $B(t) = \nabla F\left(\boldsymbol{x}\left(t, \boldsymbol{x}_0\right)\right)$, then $\mathrm{E}\left[\boldsymbol{Z}(t)\right] = M(t)\,\boldsymbol{z}$, where $M(t)$ is the unique solution to $dM(t)/dt = B(t)\,M(t)$, with initial value $M(0) = \mathbb{I}$. Similarly, the covariance matrix $\operatorname{cov}\left(\boldsymbol{Z}(t)\right) = \Sigma(t)$ is the unique solution to

$$\frac{d\Sigma(t)}{dt} = B(t)\Sigma(t) + \Sigma(t)B(t)^T + G\left(\boldsymbol{x}\left(t, \boldsymbol{x}_0\right)\right), \tag{2.12}$$

with $\Sigma_0 = 0$.

Barbour [1974] showed that if $(\boldsymbol{X}(t), t \geq 0)$ is a DDMPP whose transition rates are multinomial in terms of the elements of $\boldsymbol{x}$ then an $\mathcal{O}\left(\nu^{-1}\right)$ approximation of $(\boldsymbol{Z}(t), 0 \leq t < \infty)$ is a Gaussian diffusion process with the same mean and covariance. Thus, we refer to the *diffusion approximation* of the DDMPP $(\boldsymbol{X}(t), t \geq 0)$, for finite t, as the Gaussian diffusion process with mean function $\nu\,\boldsymbol{x}\left(t, \boldsymbol{x}_0\right)$ and covariance-matrix $\nu\,\Sigma(t)$.

It is worth noting that another large-population approximation is the van Kampen approximation. The van Kampen approximation and the diffusion approximation are similar because they are both based on a first order approximation of the Kolmogorov equations of the underlying CTMC. However, the two differ in their treatment of the limiting diffusion. The conventional van Kampen approximation, also referred to as the linear noise approximation, provides a partial differential equation describing the time-evolution of the probability distribution of $\boldsymbol{Z}(t)$. While, the conventional diffusion approximation utilises a Gaussian approximation [Barbour, 1974] to provide a closed-form approximation to the probability distribution of $\boldsymbol{Z}(t)$.

An important concept for the diffusion approximation of a DDMPP utilised in Chapter 3 is its *hitting distribution* [Ethier and Kurtz, 2008].

Theorem 5 (Hitting Distribution) *Let $\xi(\boldsymbol{x})$, for $\boldsymbol{x}$ in E, be continuously differentiable on $\mathbb{R}^K$, with $\xi(\boldsymbol{x}(0, \boldsymbol{x}_0)) > 0$, where $\boldsymbol{X}^{(\nu)}(0)/\nu = \boldsymbol{x}_0$. Let*

$$\tau^{(\nu)} = \inf_{t>0} \left\{ \xi \left(\frac{\boldsymbol{X}^{(\nu)}(t)}{\nu} \right) \leq 0 \right\}, \tag{2.13}$$

and

$$\tau = \inf_{t>0} \left\{ \xi(\boldsymbol{x}(t, \boldsymbol{x}_0)) \leq 0 \right\}. \tag{2.14}$$

Suppose $\tau < \infty$, and

$$\nabla \xi \left(\boldsymbol{x}(\tau, \boldsymbol{x}_0) \right) \cdot F \left(\boldsymbol{x}(\tau, \boldsymbol{x}_0) \right) < 0. \tag{2.15}$$

Then

$$\sqrt{\nu} \left(\tau^{(\nu)} - \tau \right) \to - \frac{\nabla \xi \left(\boldsymbol{x}(\tau, \boldsymbol{x}_0) \right) \cdot \boldsymbol{Z}(\tau)}{\nabla \xi \left(\boldsymbol{x}(\tau, \boldsymbol{x}_0) \right) \cdot F \left(\boldsymbol{x}(\tau, \boldsymbol{x}_0) \right)}, \tag{2.16}$$

and

$$\sqrt{\nu} \left(\frac{\boldsymbol{X}^{(\nu)}\left(\tau^{(\nu)}\right)}{\nu} - \boldsymbol{x}(\tau, \boldsymbol{x}_0) \right) \to \\ \boldsymbol{Z}(\tau) - \left(\frac{\nabla \xi \left(\boldsymbol{x}(\tau, \boldsymbol{x}_0) \right) \cdot \boldsymbol{Z}(\tau)}{\nabla \xi \left(\boldsymbol{x}(\tau, \boldsymbol{x}_0) \right) \cdot F \left(\boldsymbol{x}(\tau, \boldsymbol{x}_0) \right)} \right) F \left(\boldsymbol{x}(\tau, \boldsymbol{x}_0) \right), \tag{2.17}$$

where "$\to$" denotes weak convergence.

The continuous function $\psi(\boldsymbol{x})$, for $\boldsymbol{x}$ in E, specifies a boundary in E, such that the scaled DDMPP stops the instant that $\psi(\boldsymbol{x})$ becomes non-positive. The random time at which the scaled DDMPP hits this boundary is specified by equation (2.13), and the deterministic time at which the fluid limit of the DDMPP hits this boundary is specified by equation (2.14). Then, provided the deterministic time τ is finite and the dot product (2.15) is non-zero on the boundary, equation (2.16) provides an approximation for the distribution of the time at which the DDMPP hits the boundary, and equation (2.17) provides an approximation for the distribution of the state in which the DDMPP hits the boundary.

An important property of the multivariate normal distribution concerns its conditional distribution.

Theorem 6 *Let $\boldsymbol{X}$ be a multivariate normal random variable with expected value $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Further, partition $\boldsymbol{X}$ such that*

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \textit{and } \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{1\,1} & \boldsymbol{\Sigma}_{1\,2} \\ \boldsymbol{\Sigma}_{2\,1} & \boldsymbol{\Sigma}_{2\,2} \end{bmatrix}.$$

Then $\boldsymbol{X}_1$ conditioned on the event that $\boldsymbol{X}_2 = \boldsymbol{a}$, is a multivariate normal random variable with expected value $\boldsymbol{\mu}'$ and covariance $\boldsymbol{\Sigma}'$, given by

$$\begin{aligned} \boldsymbol{\mu}' &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{1\,2}\boldsymbol{\Sigma}_{2\,2}^{-1}\left(\boldsymbol{a} - \boldsymbol{\mu}_2\right), \\ \boldsymbol{\Sigma}' &= \boldsymbol{\Sigma}_{1\,1} - \boldsymbol{\Sigma}_{1\,2}\boldsymbol{\Sigma}_{2\,2}^{-1}\boldsymbol{\Sigma}_{2\,1}. \end{aligned}$$

This is an important result utilised in Chapter 4 and Chapter 5 in conducting parameter inference.

2.3 The SIR CTMC

In the remainder of this chapter we consider modelling the spread of infectious diseases through large populations in a CTMC framework. We begin by defining the SIR CTMC, and we then use this model to calculate the distribution of the duration of the outbreak, and the distribution of the final size of the outbreak. We also discuss related large-population approximations of these distributions. We then consider utilising the SIR CTMC for parameter inference using observed daily incidence data. Since the SIR CTMC assumes all infection events are observed, and that the disease does not have a latent period, we later consider using a partially-observed SEIR CTMC model for conducting parameter inference on real-world outbreaks.

We begin by defining the SIR CTMC model, otherwise known as the general stochastic epidemic model [Kermack and McKendrick, 1927, Bartlett, 1949, Dietz, 1967, Bailey, 1950, 1957, Keeling et al., 2000]. The SIR CTMC is a compartmental model which tracks the number of individuals who are: susceptible (S), infectious (I), or removed (R), where the removed compartment may refer to individuals who have either recovered from the disease or passed away. Under the common assumption that the population is closed, we have that $S + I + R = N$ so we need to keep track of only two compartments of the model because the third can then easily be determined. There are only two possible events in the SIR CTMC: infection events, and removal events. The rate at which infection events occur is typically specified as $\beta SI/(N-1)$, where β describes the rate at which each individual has transmissible contacts, and $I/(N-1)$ is the probability that such a contact is with an infectious individual. The rate at which removal events occur is γI, where γ is the rate at which an infectious individual is removed from the infectious compartment due to e.g. death or some other process such as recovery with immunity. These dynamics are summarised in Figure 2.1. The basic reproductive number is defined as the average number of secondary infection events, caused by a single infectious individual, in an otherwise susceptible population. For all but very small populations, the basic reproductive number of the SIR CTMC is $R_0 = \beta/\gamma$.

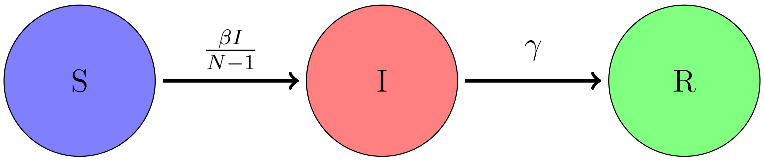


Figure 2.1: State transitions of the SIR CTMC model. β is the effective force of infection and γ is the removal rate.

Let $(\boldsymbol{X}(t), t \geq 0)$ denote the SIR CTMC, which takes values (S, I) in

$$\mathcal{X} = \left\{(S, I) \in \mathbb{Z}_{+}^{2} \, : \, S + I \leq N\right\}. \tag{2.18}$$

The only possible events of the SIR CTMC are infection events and removal events, which change the state of the process by $\boldsymbol{\ell}_1 = (-1, 1)$ and $\boldsymbol{\ell}_2 = (0, -1)$, respectively. Thus, the positive transition rates, for all $\boldsymbol{x}$ in $\mathcal{X}$, of the SIR CTMC are

$$\begin{aligned} q^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{x}+\boldsymbol{\ell}_1} &= \frac{\beta}{N-1} SI \qquad && \text{if } \boldsymbol{x} + \boldsymbol{\ell}_1 \in \mathcal{X}, \text{ and} \\ q^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{x}+\boldsymbol{\ell}_2} &= \gamma I && \text{if } \boldsymbol{x} + \boldsymbol{\ell}_2 \in \mathcal{X}, \end{aligned} \tag{2.19}$$

with the additional requirement that $q^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{x}} = -\sum_{\boldsymbol{y} \neq \boldsymbol{x}} q^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{y}}$.

We now utilise the SIR CTMC for computing the distribution of the duration of the outbreak, and the distribution of the final size of the outbreak. As we shall see, the algorithms for computing these distributions are generally computationally-expensive. Thus, we also consider computing approximations to these distributions using the hybrid models of Scalia-Tomba [1985] and Barbour [1975], respectively. We then consider inferring R_0 from observed case incidence data.

2.3.1 Outbreak duration

The duration of the outbreak is defined as the length of time from the first individual becoming infectious to the event that the final infectious individual is removed. More precisely, let $\mathcal{A} = \{\boldsymbol{x} \in \mathcal{X} \,|\, I = 0\}$ denote the set of all states in $\mathcal{X}$ for which the number of infectious individuals is zero, then the random variable $T = \inf\{t \geq 0 \,|\, \boldsymbol{X}(t) \in \mathcal{A}\}$ describes the duration of the outbreak. An assumption made herein, and throughout, is that the outbreak starts with one infectious individual.

Direct computation from the CTMC

The distribution of T may be computed via the path integral approach [Pollett and Stefanov, 2002]. This involves computing the Laplace–Stieltjes transformation of T, which is then inverted to provide $\Pr(T \leq t)$, for $t \geq 0$. This process is computationally intensive because the Laplace–Stieltjes transformation is computed by solving a system of $|\mathcal{X}|$ linear equations, and the inversion involves computing an integral on the Laplace–Stieltjes domain, which is generally achieved numerically. However, the main drawback of this approach is its inefficiency for computing $\Pr(T \leq t)$ over a range of values of t. This is because the algorithm for computing $\Pr(T \leq t)$ cannot be extended to computing $\Pr(T \leq t + \tau)$, for small τ, efficiently.

Jenkinson and Goutsias [2012] presented an approach for integrating the Kolmogorov equations (Definition 4) using an equivalent *degree-of-advancement* (DA) representation of the CTMC. Jenkinson and Goutsias [2012] showed that using the Implicit Euler scheme to integrate the Kolmogorov equations of the so called DA process, is globally stable and achieves an L_1-error of order $\mathcal{O}(\tau)$, where τ is the time-step of the numerical integration. Furthermore, when appropriately ordered, the generator matrix of the DA process is triangular, which enables the use of more efficient algorithms for solving systems of equations involving the generator matrix. In this framework, the probability $\Pr(T \leq t)$ may be used to compute the probability $\Pr(T \leq t + \tau)$ by solving one system of $|\mathcal{X}|$ equations, compared to the multiple systems of $|\mathcal{X}|$ equations required by the Laplace–Stieltjes framework.

Intuitively, the SIR CTMC is referred to as a population process because it tracks the population of the S and I compartments. On the other hand, the DA process, $(\boldsymbol{N}(t), t \geq 0)$, is a counting process which tracks the *number of infection events* (N_I) and the *number of removal events* (N_R), taking values in $\mathcal{N} = \{(N_I, N_R) : N_I, N_R = 0, 1, \ldots, N, \, N_I \geq N_R, \}$. These processes have a one-to-one correspondence because

$$\begin{aligned} N_I &= S(0) - S, & S &= S(0) - N_I, \\ N_R &= N - S - I - R(0), & I &= I(0) + N_I - N_R. \end{aligned} \tag{2.20}$$

For example, the DA representation of the initial state of the SIR CTMC $(N-1, 1)$ is $(1, 0)$. For all $\boldsymbol{n}$ in $\mathcal{N}$, the transition rates of the DA process are

$$\begin{aligned} q^{\boldsymbol{N}}_{\boldsymbol{n}\,\boldsymbol{n}+\boldsymbol{e}_1} &= \frac{\beta}{N-1}\left(S(0) - N_I\right)\left(I(0) + N_I - N_R\right) & &\text{if } \boldsymbol{n} + \boldsymbol{e}_1 \in \mathcal{N}, \\ q^{\boldsymbol{N}}_{\boldsymbol{n}\,\boldsymbol{n}+\boldsymbol{e}_2} &= \gamma\left(I(0) + N_I - N_R\right) & &\text{if } \boldsymbol{n} + \boldsymbol{e}_2 \in \mathcal{N}, \end{aligned} \tag{2.21}$$

with the additional requirement that $q^{\boldsymbol{N}}_{\boldsymbol{n}\boldsymbol{n}} = -\sum_{\boldsymbol{m} \neq \boldsymbol{n}} q^{\boldsymbol{N}}_{\boldsymbol{n}\boldsymbol{m}}$. To ensure that the generator matrix $\mathbb{Q}^{\boldsymbol{N}}$ is triangular, we order the states in the state space of the DA process such that the state (N_I, N_R) precedes the state (N_I', N_R') if and only if

$$N_I - N_R < N_I' - N_R' \qquad \text{or} \qquad N_I - N_R = N_I' - N_R' \text{ and } N_I > N_I'. \tag{2.22}$$

For brevity, we let $q^{\boldsymbol{N}}_{ij}$ and $p^{\boldsymbol{N}}_{ij}$ denote the transition rate and jump probability from the ith ordered state to the jth ordered state, respectively. Similarly, we let $\boldsymbol{p}^{\boldsymbol{N}}(t) = (p^{\boldsymbol{N}}_{\boldsymbol{e}_1\boldsymbol{m}}(t) : \boldsymbol{m} \in \mathcal{N})$, for $t \geq 0$, denote the probability distribution of the DA process at time t, given the initial state $\boldsymbol{e}_1$.

Recall that the Kolmogorov equations (Definition 4) specify that

$$\frac{d\boldsymbol{p}^{\boldsymbol{N}}(t)}{dt} = \boldsymbol{p}^{\boldsymbol{N}}(t)\,\mathbb{Q}^{\boldsymbol{N}}. \tag{2.23}$$

For a set of equally-spaced time points $t_0 < t_1 < \cdots < t_n$, with spacing τ, let $\boldsymbol{p}^{\boldsymbol{N}}_k$ denote the numerical approximation of $\boldsymbol{p}^{\boldsymbol{N}}(t_k)$. Then following Jenkinson and Goutsias [2012], the Implicit Euler method yields the iterative scheme

$$\boldsymbol{p}^{\boldsymbol{N}}_{k+1}\left(\mathbb{I} - \tau\mathbb{Q}^{\boldsymbol{N}}\right) = \boldsymbol{p}^{\boldsymbol{N}}_k, \tag{2.24}$$

with initial value $\boldsymbol{p}^{\boldsymbol{N}}_0 = \boldsymbol{e}_1$. It follows that an $\mathcal{O}(\tau)$ approximation of the distribution of T is

$$\Pr(T \leq t_k) = \sum_{\boldsymbol{n}\in\mathcal{N}^A} \boldsymbol{p}^{\boldsymbol{N}}_k, \tag{2.25}$$

where $\mathcal{N}^A$ is the DA equivalent of the set $\mathcal{A}$ (Transformation 2.20), and the subscript $\boldsymbol{n}$ refers to the element of $\boldsymbol{p}^{\boldsymbol{N}}_k$ corresponding to the state $\boldsymbol{n}$.

The fact that the matrix $\left(\mathbb{I} - \tau\mathbb{Q}^{\boldsymbol{N}}\right)^T$ is lower-triangular enables us to solve the system of equations (2.24) via backward-substitution. This provides significant improvements in computational-efficiency when solved with off-the-shelf algorithms such as MATLAB's `mldivide` [Jenkinson and Goutsias, 2012]. However, by taking advantage of the structure of $\mathbb{Q}^{\boldsymbol{N}}$, we are able to devise a specialised algorithm for computing the solution to systems of equations of this form. This algorithm is essentially the same as the algorithm presented in Black and Ross [2015] in which one iterates through all states in the state space lexicographically, at each iteration updating the solution via an infection event and a recovery event from the current state. An additional normalising step is required before calculating these interactions to assure that the final solution is a valid probability mass function.

Let $\delta k_1 = N - N_I + N_R$ and $\delta k_2 = N + 2 - N_I + N_R$, and φ_k denote the kth element of the $|\mathcal{N}| \times 1$ vector $\boldsymbol{\varphi}$. Then we use Algorithm 1 to compute the solution to the system of linear equations (2.24).

Algorithm 1: Algorithm for computing the solution to the system of linear equations (2.24).

```
Data: Set $\boldsymbol{\varphi} = \boldsymbol{p}^{\boldsymbol{N}}_j$, for any $j = 0, 1, \ldots, n-1$.
Result: Compute $\boldsymbol{p}^{\boldsymbol{N}}_{j+1}$.
Initialise the state-index as $k = 2N + 1$ ;
for $N_R = 0, 1, \ldots, N$ do
    Store the initial index $k_0 = k$ and normalise the current entry
      $\varphi_k = \varphi_k / (1 + \tau q^{\boldsymbol{N}}_{\boldsymbol{n}_k})$ ;
    for $N_I = 0, 1, \ldots, N_R$ do
        Update the distribution via:
        $\varphi_{k+\delta k_1} = \varphi_{k+\delta k_1} + \tau \varphi_k \, q^{\boldsymbol{N}}_{k\,k+\delta k_1}$ (Infection event) ;
        $\varphi_{k-\delta k_2} = \varphi_{k-\delta k_2} + \tau \varphi_k \, q^{\boldsymbol{N}}_{k\,k-\delta k_2}$ (removal event) ;
        Update the state-index $k = k + \delta k_1$ ;
    end
    Reset the state-index $k = k_0 - 1$ ;
end
Return $\boldsymbol{p}^{\boldsymbol{N}}_{j+1} = \boldsymbol{\varphi}$ ;
```

Directly integrating the Kolmogorov equations under the DA representation is the most effective way of calculating the distribution of the duration of the outbreak directly from the SIR CTMC. However, Barbour [1975] showed that a closed form approximation to this distribution may be obtained via an appropriate hybrid approximation of the SIR CTMC.

Hybrid approximation due to Barbour

Based on the assessment that the behavior of $\boldsymbol{X}(t)/N$, for all $t \geq 0$, is similar to a deterministic process when the population of S and I are large, Barbour [1975] constructed a hybrid approximation of the SIR CTMC which models the initial stages and final stages of the outbreak with an appropriate branching process and utilises the fluid approximation otherwise. Barbour used this hybrid model to derive a closed-form expression for the distribution of the duration of the outbreak. Although his model was designed with large populations in mind, it is surprisingly accurate even for "moderate" population sizes [Andersson and Britton, 2000].

The branching process approximation of the initial stages of the outbreak assumes that the susceptible pool is very large, so as to justify approximating S with a fixed value $S(0) = N - 1$. The result is a birth-death approximation of $I(t)$, with birth rates βI and death rates γI, for all $I = 0, 1, \ldots$. This approximation breaks down when I gets close to $\sqrt{N}$, at which stage the susceptible pool is too depleted to justify approximating it by $S(0)$ [Ball and Donnelly, 1995]. Given $R_0 > 1$, there is a $1 - \eta$ ($\eta = 1/R_0$) probability of a major outbreak. In which case, the distribution of time until the branching process approximation reaches $\sqrt{N}$ infectious individuals is a type-I extremal random variable [Coles, 2000].

The branching process approximation of the final stages of the outbreak assumes that the underlying proportion of susceptible individuals is close to its limiting value (as t approaches ∞) under the fluid approximation, s_∞, and that it remains constant for the remainder of the outbreak. According to Barbour, this occurs when I decreases to $N^{3/4}$, following which, a suitable approximation for $I(t)$ is a birth-death process with birth rates $\beta s_\infty I$, and death rates γI, for all $I = 0, 1, \ldots$. It can be shown that this process is conditioned on extinction (since $s_\infty R_0/\gamma < 1$), and the distribution of time until this occurs is a type-I extremal random variable.

During the intermediate stages, the behavior of the SIR CTMC, scaled by N, is similar to that of the fluid approximation. Thus, Barbour computes the time which elapses between branching process approximations as the amount of time it takes the fluid approximation to go from a state with $I = \sqrt{N}$, during the initial stages of the outbreak, to a state with $I = N^{3/4}$, during the final stages of the outbreak. In order to compute this, we now construct the fluid approximation of the SIR CTMC.

Let $(\boldsymbol{X}^{(N)}(t), t \geq 0)$, $N > 0$, denote the SIR CTMC indexed by N which takes values in $\mathcal{X}^{(N)}$. In addition, recall that $\boldsymbol{\ell}_1 = (-1, 1)$ and $\boldsymbol{\ell}_2 = (0, -1)$. Then, for all $\boldsymbol{x}$ in $\mathcal{X}^{(N)}$, it follows that

$$\begin{aligned} N\, f^{(N)}\left(\boldsymbol{x}/N, \boldsymbol{\ell}_1\right) &= N \left(\frac{\beta N}{N-1} \left(\frac{S}{N}\right) \left(\frac{I}{N}\right) \right) && \text{if } \boldsymbol{x} + \boldsymbol{\ell}_1 \in \mathcal{X}^{(N)}, \\ N\, f^{(N)}\left(\boldsymbol{x}/N, \boldsymbol{\ell}_2\right) &= N \left(\gamma \left(\frac{I}{N}\right) \right) && \text{if } \boldsymbol{x} + \boldsymbol{\ell}_2 \in \mathcal{X}^{(N)}. \end{aligned} \tag{2.26}$$

Now, let s and i denote the proportions S/N and I/N, respectively, which take values in $E = \{(s, i) \in [0, 1]^2 \,:\, s + i \leq 1\}$. Then the population SIR CTMC is a DDMPP because as $N \to \infty$, the function $F^{(N)}\left(\boldsymbol{y}\right)$, for all $\boldsymbol{x}$ in E, converges to

$$F\left(\boldsymbol{x}\right) = \begin{pmatrix} -\beta s i \\ \beta s i - \gamma i \end{pmatrix}, \tag{2.27}$$

as $N \to \infty$. Thus, the SIR CTMC satisfies the conditions of the fluid limit, provided $\boldsymbol{x}_0 = (N-1, 1)/N$. It follows that the fluid approximation of the SIR CTMC is the deterministic process $N\,\boldsymbol{x}\left(t, \boldsymbol{x}_0\right)$, for finite t, whose elements are the unique solutions to the system of ordinary differential equations

$$\begin{aligned} \frac{ds}{dt} &= -\beta s i, \\ \frac{di}{dt} &= \beta s i - \gamma i. \end{aligned} \tag{2.28}$$

By considering the trajectory of the fluid process $\boldsymbol{x}\left(t, \boldsymbol{x}_0\right)$, for $t \geq 0$, through the (s, i) plane, it can be found that

$$i = \eta \log\left(\frac{s}{s_0}\right) - s + s_0 + i_0, \tag{2.29}$$

where $\boldsymbol{x}_0 = (s_0, i_0)$. This expression is particularly useful because it allows us to deduce two important results. The first is that in the limit as $t \to \infty$, the limiting proportion of susceptible individuals s_∞ satisfies the equation

$$\eta \log\left(\frac{s_\infty}{s_0}\right) - s_\infty + s_0 + i_0 = 0. \tag{2.30}$$

This equation has a trivial solution of $s_\infty = 1$, which corresponds to the event that there is no outbreak. The desired solution is on the interval $(0, 1)$, because this corresponds to the event that an outbreak actually occurs. Recall that the fluid approximation is only valid over finite time intervals, however, see Section 11.4 of Ethier and Kurtz [2008] for justification of the limiting value of the fluid approximation.

The second result of equation (2.29) is that it may be substituted into the derivative ds/dt (equation (2.28)) to obtain an expression for the amount of time which elapses while $a \leq s(t) \leq b$, for a, b in $(0, 1)$, given by

$$J\left(a, b\right) = \frac{1}{\beta} \int_a^b \left[s\left(s - s_0 - i_0 - \eta \log\left(\frac{s}{s_0}\right)\right)\right]^{-1} ds. \tag{2.31}$$

We are now able to state the following theorem due to Barbour [1975].

Theorem 7 *Recall that the random variable T is the duration of the outbreak. Then, provided $i_0 = 1/N$ and $R_0 \geq 0$, then, as $N \to \infty$,*

$$\Pr\left(T \leq t \,|\, \mathcal{E}\right) \to \frac{\eta\left(1 - e^{-(1-\eta)t}\right)}{1 - \eta e^{-(1-\eta)t}}, \tag{2.32}$$

and

$$\Pr\left(T - \left(\left(\frac{1}{\eta - s_\infty}\right) + \left(\frac{1}{1-\eta}\right)\right) \log(N) - c \geq x \,|\, \mathcal{E}\right) \to (1-\eta) \Pr\left(W' \geq x\right), \tag{2.33}$$

where

1. $\mathcal{E}$ *denotes the event that a major outbreak occurs;*

2. *we have that*

$$c = \lim_{m\to\infty}\left[-\left(\left(\frac{1}{\eta - s_\infty}\right) + \left(\frac{1}{1-\eta}\right)\right)\log(m) + \left(\frac{1}{1-\eta}\right)\log(1-\eta)\right.$$
$$\left.\left(\frac{1}{\eta - s_\infty}\right)\log(\eta - s_\infty) + J\left(s_\infty\left(1 + \frac{1}{m(\eta - s_\infty)}\right), 1 - \frac{1}{m(1-\eta)}\right)\right];$$

3. s_∞ *and* $J\left(.,.\right)$ *are evaluated setting* $i_0 = 0$*;*

4. W' *has the distribution of*

$$\left(\frac{1}{\eta - s_\infty}\right)W_1 + \left(\frac{1}{1-\eta}\right)W_2,$$

where W_1 *and* W_2 *are independent type-I extremal random variables.*

Equation (2.32) provides the distribution of T conditioned on a major outbreak, and equation (2.33) provides the distribution of T conditioned on the event that the outbreak fades out, via the convolution of two type-I extremal random variables. To compute the convolution, we utilise the approach of Nadarajah [2007], who analysed the more general case of a linear combination of two Gumbel random variables.

Theorem 8 *Let* $X \sim$ *Gumbel*(μ,σ) *and* $Y \sim$ *Gumbel*(θ,λ)*, for* $\mu,\theta \in \mathbb{R}$ *and* $\sigma,\lambda > 0$*, be independent Gumbel random variables. Define* $Z = \alpha X + \beta Y$*, such that* $\alpha,\beta > 0$*. Then, provided* $\alpha\sigma/|\lambda\beta|$ *is rational, the probability distribution function of* Z *is*

$$\Pr\left(Z \le z\right) = \frac{\alpha\sigma A(z)}{\beta\lambda} K\left(-\frac{\alpha\sigma}{\beta\lambda}, A(z), \frac{\alpha\sigma}{\beta\lambda}, 1\right),$$

where

$$A(z) = \exp\left(\frac{\alpha\mu + \beta\theta - z}{\lambda\beta}\right),$$

and

$$K(\gamma, a, r, s) = \int_0^\infty x^{\gamma-1} \exp\left(-ax^{-r} - sx\right)\, dx.$$

Since a type-I extremal random variable is Gumbel$(0,1)$ distributed, and following from condition 4 of Theorem 7, the only requirement of Theorem 8 is that $(1-\eta)/(\eta - s_\infty)$ is rational. Although it is unreasonable to assume the exact value of s_∞ is always rational, in practice its value is computed to finite precision as the solution to equation (2.30). Similarly, η is either specified or calculated to finite precision. Thus, it is reasonable to assume that this condition holds in practice, thereby fulfilling the only condition of Theorem 8.

In Chapter 3 we construct a hybrid fluid model which differs from Barbour's hybrid model only in its use of the SIR CTMC in place of Barbour's branching process approximations. We use our hybrid fluid model to calculate the distribution of the duration of the outbreak, which we compare to the exact distribution (Equation (2.25)) and Barbour's approximation (Theorem 7). As we shall see, our hybrid model is more accurate than Barbour's for moderately sized N, but the two are similar when N is large. We now consider calculating the distribution of the final size of the outbreak.

2.3.2 Final outbreak size

The final size of the outbreak is defined as the total number of individuals who experience infection from the time at which the first individual becomes infectious until the time when the final infectious individual is removed from the population. More precisely, recall that $\mathcal{A}$ is the set of all states with $I = 0$, and T is the hitting time of the CTMC on $\mathcal{A}$. Then the random variable $R(T) = N - S(T)$ describes the final size of the outbreak.

Direct computation from the CTMC

Since the SIR CTMC is time-homogeneous (Definition 2), its hitting distribution on $\mathcal{N}^A$ is time-independent and may therefore be deduced from its embedded jump process (Definition 5). The embedded jump process of the DA process is the DTMC $(\boldsymbol{N}_n, n \geq 0)$, which takes values in $\mathcal{N}$ and, for all $\boldsymbol{n}$ in $\mathcal{N}$, has the transition probabilities

$$p^{\boldsymbol{N}}_{\boldsymbol{n}\,\boldsymbol{n}+\boldsymbol{e}_1} = \frac{\beta\,(S(0) - N_I)}{\beta\,(S(0) - N_I) + \gamma(N-1)} \qquad \text{if } \boldsymbol{n}+\boldsymbol{e}_1, \boldsymbol{n}+\boldsymbol{e}_2 \in \mathcal{N}, \tag{2.34}$$

$$p^{\boldsymbol{N}}_{\boldsymbol{n}\,\boldsymbol{n}+\boldsymbol{e}_2} = \frac{\gamma(N-1)}{\beta\,(S(0) - N_I) + \gamma(N-1)} \qquad \text{if } \boldsymbol{n}+\boldsymbol{e}_1, \boldsymbol{n}+\boldsymbol{e}_2 \in \mathcal{N}, \tag{2.35}$$

with $p^{\boldsymbol{N}}_{\boldsymbol{n}\,\boldsymbol{n}+\boldsymbol{e}_2} = 1$ if $\boldsymbol{n}+\boldsymbol{e}_1 \notin \mathcal{N}$ and $\boldsymbol{n}+\boldsymbol{e}_2 \in \mathcal{N}$.

Recall that the hitting probabilities on the set $\mathcal{N}^A$ are the minimal non-negative solution to the system of equations (2.2). It follows that, given the initial state $\boldsymbol{e}_1$, the distribution of the final size of the outbreak is given by

$$\Pr(R(T) = r) = h^{\boldsymbol{N}}_{\boldsymbol{e}_1\,(N-r,0)}, \tag{2.36}$$

for all $r = 0, 1, \ldots, N$. Black and Ross [2015] presented a highly efficient algorithm for computing the solution to these hitting probabilities, similar to Algorithm 1. Recall that $\delta k_1 = N - N_I + N_R$ and $\delta k_2 = N + 2 - N_I + N_R$, and that φ_k denotes the kth element of the $|\mathcal{N}| \times 1$ vector $\boldsymbol{\varphi}$. Then Algorithm 2 is equivalent to the algorithm of Black and Ross [2015], which we use to compute the solution to the system of linear equations (2.2).

A number of authors [Von Bahr and Martin-Lof, 1980, Ball, 1983, Watson, 1980a,b, 1981, Martin-Lof, 1990] derived closed form approximations to the distribution of the final size of the outbreak via a similar hybrid model to Barbour [1975]. However, these approaches were subsequently summarised by Lefèvre [1990] as being essentially the same. In the following discussion we present one of the most widely-used approaches, due to Scalia-Tomba [1985].

Algorithm 2: Algorithm for computing the distribution of the final size of the outbreak, via the hitting probabilities (2.2).

```
Data: Set $\boldsymbol{\varphi} = \boldsymbol{e}_1$.
Result: Compute $\Pr(R(T) = r)$, for all $r = 0, 1, \ldots, N$.
Initialise the state-index $k = 2N + 1$ ;
for $N_R = 0, 1, \ldots, N$ do
    Store the initial index $k_0 = k$ ;
    for $N_I = 0, 1, \ldots, N_R$ do
        Update distribution via:
        $\varphi_{k+\delta k_1} = \varphi_{k+\delta k_1} + \varphi_k \, p^{\boldsymbol{N}}_{k\,k+\delta k_1}$ (Infection event) ;
        $\varphi_{k-\delta k_2} = \varphi_{k-\delta k_2} + \varphi_k \, p^{\boldsymbol{N}}_{k\,k-\delta k_2}$ (Removal event) ;
        Update the state-index $k = k + \delta k_1$ ;
    end
    Reset the state-index $k = k_0 - 1$ ;
end
Return $\Pr(R(T) = k) = \varphi_k$, for all $k = 1, 2, \ldots, N + 1$ ;
```

Hybrid approximation due to Scalia-Tomba

Based on the observation that a Gaussian diffusion process provides a suitable approximation of the SIR CTMC once the outbreak has become established, Scalia-Tomba [1985] constructed a hybrid model for computing the distribution of the final size of the outbreak by separately considering the event that the outbreak fades out and the event that a major outbreak occurs. In the former case, the SIR CTMC is approximated by an appropriate branching process, and in the latter case, the SIR CTMC is approximated by an appropriate normal distribution.

First, we state a well known result of the branching process approximation of the initial stages of the SIR CTMC [Ball and Donnelly, 1995, Ball and Neal, 2010].

Theorem 9 *Let R_∞ denote the total progeny in a birth-death process with birth rate βI and death rate γI, for $I = 0, 1, \ldots$. Then the following is true*

$$\Pr(R_\infty = r) = \begin{pmatrix} 2r + I(0) \\ r \end{pmatrix} \left(\frac{I(0)}{2r + I(0)} \right) \left(\frac{\beta^r \gamma^{r+I(0)}}{(\beta + \gamma)^{2r+I(0)}} \right), \qquad (2.37)$$

for all $r \geq 0$.

This theorem provides the distribution of the total number of individuals who experience infection under the branching process approximation of the initial stages of the SIR CTMC. In the case where $R_0 > 1$, this distribution is defective because there is a $1 - \eta$ probability of a major outbreak occuring. Thus, in the branching process framework, there is a probability mass of $1 - \eta$ associated with the event that R_∞ is infinite. Scalia-Tomba [1985] accounted for this by utilising a normal approximation of the distribution of the final size of the outbreak, conditioned on a major outbreak. The result is as follows.

Theorem 10 *Recall that* $(\boldsymbol{X}^{(N)}(t), t \geq 0)$, $N > 0$, *denotes the sequence of SIR CTMCs indexed by* N, *and assume that* $R_0 > 1$. *Then as* $N \to \infty$, $R(T)$ *converges to* R_∞ *with probability* η, *and with probability* $1 - \eta$ *the sequence*

$$\sqrt{N}\left(\frac{R^{(N)}(T) - I^{(N)}(0)}{N} - r_\infty\right),$$

where $r_\infty = 1 - s_\infty$, *converges weakly to a normally distributed random variable with mean* 0 *and variance*

$$s_\infty r_\infty \left(\frac{1 + r_\infty R_0^2}{(1 - r_\infty R_0)^2}\right).$$

In Chapter 3 we construct a similar hybrid model which has the dynamics of the SIR CTMC whenever the number of infectious individuals is low and the dynamics of the diffusion approximation otherwise. We compute the distribution of the final size of the outbreak from our hybrid diffusion model and compare it to the exact distribution (Equation (2.2)) and Scalia-Tomba's approximation (Theorem 10). As we shall see, Scalia-Tomba's approximation is highly accurate, but fails to capture a degree of skewness that arises during the initial and final stages of the outbreak that is successfully captured by our hybrid diffusion model.

2.3.3 Inferring the basic reproductive number

Until now we have discussed computing the distribution of the duration of the outbreak, and the distribution of the final size of the outbreak from the SIR CTMC. However, an important aim of this thesis is to develop computationally-efficient routines for inferring properties of the basic reproductive number from observed data. In this section we define the basic methodology for performing likelihood-based inference in the framework of the SIR CTMC, using daily incidence data from the initial stages of an outbreak. In the next section, we apply the same methodology to a more realistic partially-observed SEIR CTMC.

The exact likelihood

The likelihood may be thought of as the probability that an observed set of daily incidence counts x_k, for all $k = 0, 1, \ldots, n$, came from the proposed model with a particular set of parameters $\boldsymbol{\theta}$ chosen from the set Θ [Sprott, 2000]. The DA framework is amenable to parameter inference based on case incidence counts because the cumulative incidence counts $y_k = \sum_{j=1}^{k} x_j$, for all $k = 1, 2, \ldots, n$, may be thought of as direct observations of the N_I component of the process. However, this construction requires the assumption that every infectious case within the population is observed. Although this assumption may be justified in small populations if the disease has distinct symptoms, it is generally unrealistic (see Chapter 5). Nevertheless, assuming $\boldsymbol{N}(0) = \boldsymbol{e}_1$, the *exact likelihood* is

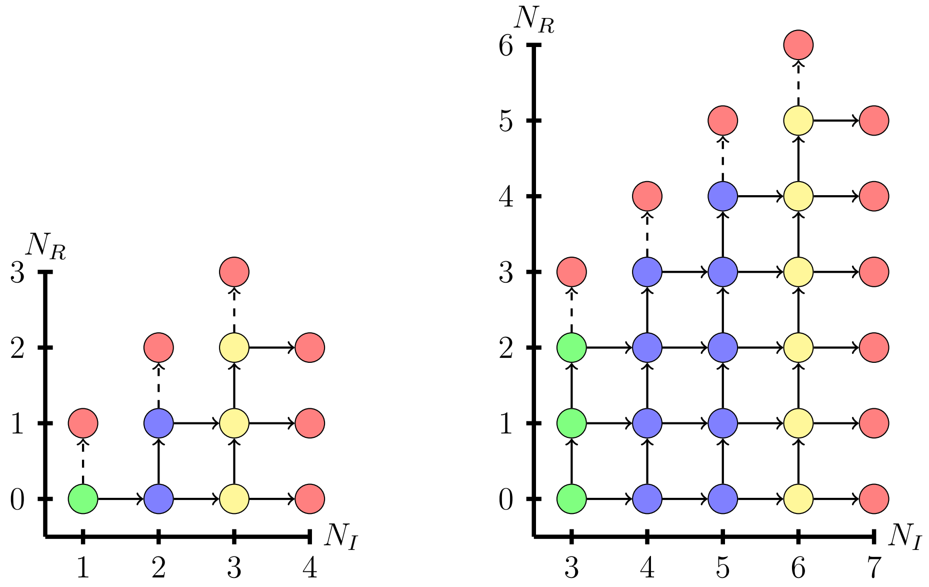
$$L(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{k=1}^{n} \Pr\left(N_I(t_k) = y_k \,|\, \mathcal{Y}_{k\text{-}1}\right), \tag{2.38}$$

where $\mathcal{Y}_{k\text{-}1} = \{N_I(t_{k\text{-}1}) = y_{k\text{-}1}, N_I(t_{k\text{-}2}) = y_{k\text{-}2}, \ldots, N_I(t_0) = y_0\}$, is the *history* of the outbreak. For brevity, herein $L_E^k(\boldsymbol{\theta})$, for $k = 1, 2, \ldots, n$ denotes the probability of the observed data $\Pr\left(N_I(t_k) = y_k \,|\, \mathcal{Y}_{k\text{-}1}\right)$, which can be computed from the data as follows.

$$\begin{aligned}
L_E^k(\boldsymbol{\theta}) &= \Pr\left(N_I(t_k) = y_k \,|\, \mathcal{Y}_{k-1}\right) \\
&= \sum_{i=0}^{y_k-1} \Pr\left(\boldsymbol{N}\left(t_k\right) = (y_k, i) \,|\, \mathcal{Y}_{k-1}\right) \\
&= \sum_{j=0}^{y_{k-1}} \sum_{i=0}^{y_k-1} \Pr\left(\boldsymbol{N}\left(t_k\right) = (y_k, i) \,|\, \boldsymbol{N}\left(t_{k-1}\right) = (y_{k-1}, j), \mathcal{Y}_{k-1}\right) \\
&\qquad\qquad\qquad\qquad \times \Pr\left(\boldsymbol{N}\left(t_{k-1}\right) = (y_{k-1}, j) \,|\, \mathcal{Y}_{k-1}\right) \\
&= \sum_{j=0}^{y_{k-1}} \sum_{i=0}^{y_k-1} p^{\boldsymbol{N}}_{(y_{k-1},j)\,(y_k,i)}(t_k - t_{k-1}) \Pr\left(\boldsymbol{N}\left(t_{k-1}\right) = (y_{k-1}, j) \,|\, \mathcal{Y}_{k-1}\right) \\
&= \sum_{j=0}^{y_{k-1}} \sum_{i=0}^{y_k-1} p^{\boldsymbol{N}}_{(y_{k-1},j)\,(y_k,i)}(t_k - t_{k-1}) \left(\frac{\Pr\left(\boldsymbol{N}\left(t_{k-1}\right) = (y_{k-1}, j) \,|\, \mathcal{Y}_{k-2}\right)}{\Pr\left(N_I(t_{k-1}) = y_{k-1} \,|\, \mathcal{Y}_{k-2}\right)}\right) \\
&= \sum_{j=0}^{y_{k-1}} \sum_{i=0}^{y_k-1} p^{\boldsymbol{N}}_{(y_{k-1},j)\,(y_k,i)}(t_k - t_{k-1}) \left(\frac{\Pr\left(\boldsymbol{N}\left(t_{k-1}\right) = (y_{k-1}, j) \,|\, \mathcal{Y}_{k-2}\right)}{L_E^{k-1}(\boldsymbol{\theta})}\right).
\end{aligned} \tag{2.39}$$

Noting that the event that $\boldsymbol{N}\left(t_{k-1}\right) = (y_{k-1}, j)$, for any $j = 1, 2, \ldots, y_{k-1}$, intersects with the event that $N_I(t_{k-1}) = y_{k-1}$, the fourth step uses conditional probability to rewrite the previous equation, to provide $L_E^{k-1}(\boldsymbol{\theta})$ in the denominator. This expression lends itself to a straightforward approach to computing the likelihood (2.38), which we now illustrate using a simplistic data set.

Illustrative example Suppose an outbreak infects two individuals on the first day, and three on the second. Utilising the assumption that on day zero, there was a single unobserved infectious individual, the basic reproductive number is inferred from the cumulative incidence counts $y_0 = 1$, $y_1 = 3$ and $y_2 = 6$ via the exact likelihood $L(\boldsymbol{y}|\boldsymbol{\theta}) = L_E^1(\boldsymbol{\theta})L_E^2(\boldsymbol{\theta})$, by calculating $L_E^1(\boldsymbol{\theta})$ and then $L_E^2(\boldsymbol{\theta})$.



(a) State transition diagram for calculating the probability that $N_I(1) = 3$, assuming the initial state $N_I(0) = 1$.

(b) State transition diagram for calculating the probability that $N_I(2) = 6$, given $N_I(1) = 3$ and $N_I(0) = 1$.

Figure 2.2: Example of how the exact likelihood is computed, using the observed incidence counts $x_1 = 2$ and $x_2 = 3$. The state transition diagrams display the truncated state spaces which contain: initial states (green), absorption states (red), states used to compute $L_E^1(\boldsymbol{\theta})$ and $L_E^2(\boldsymbol{\theta})$ (yellow), and ordinary transient states (blue).

The probability $L_E^1(\boldsymbol{\theta})$ is defined as the probability of observing three infection events in the DA process by day 1, given that $\boldsymbol{N}(0) = (1, 0)$. Since N_I is monotonically non-decreasing, the computational effort of this calculation can be reduced by truncating the state space to contain only states with $1 \leq N_I \leq 4$. The resulting state space is shown in Figure 2.2a, in which the green state is the initial state, the yellow states are states with $N_I = 3$, the blue states are ordinary transient states, and the red states are absorbing states. It follows that the probability $L_E^1(\boldsymbol{\theta})$ is obtained by integrating the transition probabilities of the DA process from day 0 to day 1 using the Kolmogorov equations (Equation (2.23)), and then adding up the probability that $\boldsymbol{N}(1)$ is in any of the yellow states.

It is worth noting that the absorbing states with $N_I = N_R$ are extinction states which we will later condition the DA process on never reaching (Chapter 4), hence transition into these states is denoted by a dashed arrow.

We now seek the probability $L_E^2(\boldsymbol{\theta})$, which is defined as the probability that $N_I(2) = 6$, given the history $\mathcal{Y}_1 = \{N_I(0) = 1, N_I(1) = 3\}$. In order to consider the DA process conditioned on the event $\mathcal{Y}_1$ for $t \geq 1$, the distribution of $\boldsymbol{N}(1)$ is conditioned on being in the set of the yellow states in Figure 2.2a. This is given by

$$\Pr\left(\boldsymbol{N}(1) = (3, i) \,|\, \mathcal{Y}_1\right) = \frac{p^{\boldsymbol{N}}_{(1,0)\,(3,i)}(1)}{L_E^1(\boldsymbol{\theta})},$$

for all $i = 0, 1, 2$. To calculate $L_E^2(\boldsymbol{\theta})$ we truncate the state space to contain only states in $\mathcal{N}$, such that $3 \leq N_I \leq 7$. This is shown in Figure 2.2b, for which the initial distribution across the green states is provided by the above distribution, and the yellow states denote states with $N_I = 6$. It follows that the transition probability $L_E^2(\boldsymbol{\theta})$ is obtained by evolving the distribution of the DA process, conditioned on $\mathcal{Y}_1$, from day 1 to day 2, using the Kolmogorov equations (Equation (2.23)), and then summing the probabilities that $\boldsymbol{N}(2)$ is in each of the yellow states.

The exact likelihood may now be computed as the product of the probabilities $L_E^1(\boldsymbol{\theta})$ and $L_E^2(\boldsymbol{\theta})$. It is worth noting that this algorithm may be extended to include more observations by generalising the procedure for calculating $L_E^2(\boldsymbol{\theta})$. This is made precise in Algorithm 3.

The computational-effort of computing the exact likelihood is influenced by the total number of observed infection events y_n. This can be a concern if y_n is large because likelihood-based inference is generally computationally

Algorithm 3: Algorithm for computing the likelihood $L(\boldsymbol{y}|\boldsymbol{\theta})$, given a set of observed incidence counts $x_1, x_2, \ldots, x_n$.

```
Data: Daily incidence counts $x_0, x_1, \ldots, x_n$
Result: Compute the likelihood $L(\boldsymbol{y}|\boldsymbol{\theta})$.
Set $y_k = \sum_{j=0}^{k} x_j$, for all $k = 1, 2, \ldots, n$, and $\boldsymbol{p}^{\boldsymbol{N}}(0) = \boldsymbol{e}_1$ ;
for $k = 0, 1, \ldots, n-1$ do
    Truncate the state space, $\mathcal{N}^k = \{\boldsymbol{m} \in \mathcal{N} \mid y_k \leq N_I \leq y_{k+1} + 1\}$ ;
    Given $\boldsymbol{p}^{\boldsymbol{N}}(t_k)$, compute $\boldsymbol{p}^{\boldsymbol{N}}(t_{k+1})$ ;
    Compute the probability $L_E^{k+1}(\boldsymbol{\theta})$ ;
    Condition $\boldsymbol{N}(t_{k+1})$ on the event that $N_I(t_{k+1}) = y_{k+1}$ ;
end
Compute $L(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{k=1}^{n} L_E^k(\boldsymbol{\theta})$.
```

intensive and requires evaluating the likelihood a large number of times. Thus, in the next section we consider utilising the diffusion approximation instead of the DA process.

The large-population approximation

When conducting inference on large populations, the exact likelihood is often computationally prohibitive because the total number of observed infection events is too large. However, under these circumstances it is generally safe to assume that the diffusion approximation will provide a sufficiently accurate approximation of the underlying CTMC. Thereby providing a computationally-efficient alternative to the SIR CTMC [Ross et al., 2006, 2009, Ross, 2012]. In this section we describe how the diffusion approximation can be utilised for parameter inference.

We begin by constructing the diffusion approximation of the DA process.

Let $(\boldsymbol{N}^{(N)}(t), t \geq 0)$, $N > 0$, denote the DA process indexed by N, which takes values in $\mathcal{N}^{(N)}$. Then, for all $\boldsymbol{n}$ in $\mathcal{N}^{(N)}$, $\boldsymbol{N}^{(N)}(t)$ provides

$$f^{(N)}\left(\boldsymbol{n}/N, \boldsymbol{e}_1\right) = \left(\frac{\beta N}{N-1}\left(\frac{S(0)}{N} - \frac{N_I}{N}\right)\left(\frac{I(0)}{N} + \frac{N_I}{N} - \frac{N_R}{N}\right)\right),$$

if $\boldsymbol{n} + \boldsymbol{e}_1 \in \mathcal{N}^{(N)}$, and

$$f^{(N)}\left(\boldsymbol{n}/N, \boldsymbol{e}_2\right) = \left(\gamma\left(\frac{I(0)}{N} + \frac{N_I}{N} - \frac{N_R}{N}\right)\right), \tag{2.40}$$

if $\boldsymbol{n} + \boldsymbol{e}_2 \in \mathcal{N}^{(N)}$.

Now, let s_0, i_0, n_I and n_R denote the proportions $S(0)/N$, $I(0)/N$, N_I/N and N_R/N, respectively, where (n_I, n_R) takes values in $E = \{(n_I, n_R) \in [0,1]^2 : 0 \leq n_R \leq n_I \leq 1\}$. Then the DA process is density dependent because as $N \to \infty$ the function $F^{(N)}\left(\boldsymbol{m}\right)$, for all $\boldsymbol{m}$ in E, converges to

$$F\left(\boldsymbol{m}\right) = \begin{pmatrix} \beta\left(s_0 - n_I\right)\left(i_0 + n_I - n_R\right) \\ \gamma\left(i_0 + n_I - n_R\right) \end{pmatrix}, \tag{2.41}$$

where $\lim_{N\to\infty} \boldsymbol{X}(0) = (s_0, i_0)$. Thus, the DA process satisfies the conditions of the fluid limit (Theorem 3), provided $\boldsymbol{n}_0 = (1,0)/N$. It follows that the fluid approximation of $\left(\boldsymbol{N}\left(t\right), t \geq 0\right)$ is the deterministic process $\left(N\,\boldsymbol{n}\left(t, \boldsymbol{n}_0\right), 0 \leq t < \infty\right)$ whose elements are the unique solution to the system of ordinary differential equations

$$\frac{dn_I}{dt} = \beta\left(s_0 - n_I\right)\left(i_0 + n_I - n_R\right), \tag{2.42}$$

$$\frac{dn_R}{dt} = \gamma\left(i_0 + n_I - n_R\right). \tag{2.43}$$

From the diffusion limit (Theorem 4), the fluctuations of the DA process about the deterministic trajectory $\left(\boldsymbol{n}\left(t, \boldsymbol{n}_0\right), t \geq 0\right)$ are captured by the Gaussian diffusion $(\boldsymbol{Z}(t), t \geq 0)$ with mean $\boldsymbol{0}$ and covariance matrix $\Sigma^N(t) = (\sigma^N_{i,j}(t) : i, j = 1, 2)$, whose elements are the unique solutions to the system of ordinary differential equations

$$\begin{aligned}\frac{d\sigma_1^N}{dt} &= 2\beta\sigma_1^N\left(s_0 - i_0 + n_R - 2n_I\right)\\ &\qquad - 2\beta\sigma_{1,2}^N\left(s_0 - n_I\right) + \beta\left(s_0 - n_I\right)\left(i_0 + n_I - n_R\right),\\ \frac{d\sigma_{1,2}^N}{dt} &= \gamma\left(\sigma_1^N - \sigma_{1,2}^N\right) + \beta\sigma_{1,2}^N\left(s_0 - i_0 + n_R - 2n_I\right) - \beta\sigma_2^N\left(s_0 - n_I\right),\\ \frac{d\sigma_2^N}{dt} &= \gamma\left(i_0 + n_I - n_R + 2\sigma_{1,2}^N - 2\sigma_2^N\right), \end{aligned} \tag{2.44}$$

with $\sigma_{2,1}^N = \sigma_{1,2}^N$. It follows that, for $0 \le t < \infty$, the diffusion approximation of the DA process is a Gaussian diffusion process with mean $N\,\boldsymbol{n}\left(t, \boldsymbol{n}_0\right)$, and covariance matrix $N\,\Sigma^N(t)$. It is worth noting that the fluid approximation and the diffusion approximation of the SIR CTMC can be obtained from the fluid approximation and diffusion approximation of the DA process via a change of variables (2.20).

Given a suitable initial state, the diffusion approximation provides an approximation of the transition probabilities of the underlying DA process, which is often referred to as the *transition density*. More precisely, suppose $\boldsymbol{N}(0) = \boldsymbol{n}$, then the transition density is

$$\begin{aligned} f_N(\boldsymbol{n}, \boldsymbol{m}, t) = {} & \frac{1}{2\pi N\sqrt{|\Sigma(t)|}} \times \\ & \exp\left(-\frac{1}{2}\left(\boldsymbol{m} - \boldsymbol{n}\left(t, \frac{\boldsymbol{n}}{N}\right)\right)^T \Sigma^{-1}(t)\left(\boldsymbol{m} - \boldsymbol{n}\left(t, \frac{\boldsymbol{n}}{N}\right)\right)\right), \end{aligned} \tag{2.45}$$

for all $\boldsymbol{m}$ in $\mathcal{N}$ and $0 \le t < \infty$.

In the framework of the diffusion approximation, the likelihood is usually constructed in terms of the transition density. However, it will be instructive for Chapter 4 if we think of the transition density as a means of approximating the transition probabilities of the DA process. In particular, suppose $\boldsymbol{n}$ and $\boldsymbol{m}$ are in $\mathcal{N}$, with $\boldsymbol{m} = (N_I, N_R)$, then the transition density provides the approximation

$$p^{\boldsymbol{N}}_{\boldsymbol{n}\boldsymbol{m}}(t) \approx \int_{N_I-\frac{1}{2}}^{N_I+\frac{1}{2}} \int_{N_R-\frac{1}{2}}^{N_R+\frac{1}{2}} f_N(\boldsymbol{n}, (v, u), t)\, du\, dv.$$

Since the transition density follows a bivariate normal distribution, it may be computationally-expensive to compute so we utilise the midpoint approximation

$$p^{\boldsymbol{N}}_{\boldsymbol{n}\boldsymbol{m}}(t) \approx f_N(\boldsymbol{n}, \boldsymbol{m}, t). \tag{2.46}$$

We are now able to specify the *diffusion likelihood* as

$$L(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{k=1}^{n} L_D^k(\boldsymbol{\theta}). \tag{2.47}$$

Following from equation (2.39), the probabilities of the observed data, $L_D^k(\boldsymbol{\theta})$, for $k = 1, 2, \ldots, n$, are given by

$$L_D^k(\boldsymbol{\theta}) = \sum_{j=0}^{y_{k\text{-}1}} \sum_{i=0}^{y_k} f_N\left((y_{k\text{-}1}, j), (y_k, i), t_k - t_{k\text{-}1}\right) \times \\ \left(\frac{\Pr\left(\boldsymbol{N}\left(t_{k\text{-}1}\right) = (y_{k\text{-}1}, j) \,|\, \mathcal{Y}_{k\text{-}2}\right)}{L_D^{k\text{-}1}(\boldsymbol{\theta})}\right). \tag{2.48}$$

The diffusion likelihood is computed via Algorithm 3, with the modification that the transition probabilities are approximated by the transition densities via equation (2.46). In the context of Figure 2.2a, this means that the transition probability $L_E^1(\boldsymbol{\theta})$ is approximated by $L_D^1(\boldsymbol{\theta})$ using the transition densities $f_N((1,0),(3,i),1)$, for $i = 0, 1, 2$. It follows that the initial distribution over the green states in Figure 2.2b can be approximated by normalising their probability densities,

$$\Pr\left(\boldsymbol{N}\,(1) = (3, i) \,|\, \mathcal{Y}_1\right) = \frac{f_N((1,0),(3,i),1)}{L_D^1(\boldsymbol{\theta})},$$

for all $i = 0, 1, 2$.

We now discuss likelihood-based methodology for inferring the parameters $\boldsymbol{\theta}$ from a set of observed daily incidence counts.

Likelihood-based Inference Methodology

There are two distinct frameworks in which one can conduct parameter inference. The first is the frequentist framework, in which the parameters are assumed to have a fixed, but unknown, underlying value, and the second is the Bayesian framework, in which the parameters are treated as random variables. We now discuss the methodology of both frameworks.

In the frequentist framework, one commonly uses the *Maximum Likelihood Estimate* (MLE) to infer the true value of the parameters.

Definition 9 (Maximum Likelihood Estimate) *The MLE, $\boldsymbol{\theta}^{MLE}$, is the value of $\boldsymbol{\theta}$ in Θ which maximises the likelihood. That is,*

$$\boldsymbol{\theta}^{MLE} = \arg\max_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{y}|\boldsymbol{\theta}). \tag{2.49}$$

The MLE may be thought of as the set of model parameters $\boldsymbol{\theta}$ in Θ which maximises the probability that the observed data came from the specified model, with the parameters $\boldsymbol{\theta}$. A useful property of the MLE is that, under certain regularity conditions, the asymptotic difference (as $n \to \infty$) between the MLE and the true parameters is approximately normal with mean 0, and known covariance [Casella and Berger, 2002].

The MLE may be computed by simply maximising the likelihood with respect to $\boldsymbol{\theta}$ in Θ [Casella and Berger, 2002]. However, numerical accuracy is a common concern because the likelihood is computed as the product of the probabilities $L^k(\boldsymbol{\theta})$, for all $k = 1, 2, \dots, n$, which are generally small. Thus, one usually works with the *log-likelihood*. The log-likelihood is defined as the log of the likelihood and is beneficial because it avoids computing the product (equation (2.38)) in favour of the sum

$$\log(L(\boldsymbol{y}|\boldsymbol{\theta})) = \sum_{k=0}^{n} \log(L^k(\boldsymbol{\theta})). \tag{2.50}$$

Since $\log(x)$, for x in $\mathbb{R}$, is continuous and monotonically-increasing, the set of parameters $\boldsymbol{\theta}$ in Θ which maximises the log-likelihood is identical to the set of parameters which maximises the likelihood. Thus, we later compute the MLE by maximising the log-likelihood using MATLAB's built-in `fmincon` constrained optimisation routine.

In the frequentist framework, the parameters are assumed to have a fixed, but unknown, underlying value, which we deduce using an estimator which is a random variable. In a Bayesian framework one treats the parameters as fixed and aims to model the uncertanty surrounding the parameters. Bayesian inference may be thought of as a process where one iteratively updates one's understanding of the distribution of the parameters as new information becomes available. This process starts with a *prior distribution* $f(\boldsymbol{\theta})$, for $\boldsymbol{\theta}$ in Θ, describing one's initial understanding of the distribution of the parameters. As new data (here denoted $\boldsymbol{y}$) becomes available, one updates one's prior distribution via Bayes' rule to obtain the *posterior distribution* $f(\boldsymbol{\theta}|\boldsymbol{y})$, describing one's updated understanding of the distribution of the parameters.

According to Bayes' rule, the posterior distribution may be written as

$$f(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{L(\boldsymbol{y}|\boldsymbol{\theta})\, f(\boldsymbol{\theta})}{\int_{\Theta} L(\boldsymbol{y}|\boldsymbol{\theta})\, f(\boldsymbol{\theta})}. \tag{2.51}$$

Although this provides the exact expression for the posterior in terms of the likelihood and the prior, the denominator is generally impractical to compute, especially when the dimension of $\boldsymbol{\theta}$ is high. Thus, one usually aims to estimate the posterior distribution. A common approach for doing so is the Metropolis-Hastings algorithm. The Metropolis-Hastings algorithm is a Markov chain Monte Carlo approach which generates samples from the posterior distribution by sampling from a similar distribution. More specifically, based on the fact that Bayes' rule implies that

$$f(\boldsymbol{\theta}|y) \propto L(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}), \tag{2.52}$$

the Metropolis-Hastings algorithm generates samples from $f(\boldsymbol{\theta}|y)$ by sampling instead from $L(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})$.

The Metropolis-Hastings algorithm is initiated by randomly generating a set of parameters $\boldsymbol{\theta}_0$ from the prior distribution. The algorithm then generates samples $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n$ which, after an initial convergence period, are random samples from the density $L(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})$. Each iteration of the algorithm starts by randomly selecting a set of *candidate parameters*, $\boldsymbol{\theta}'$, from a pre-defined *proposal distribution*, $q(\boldsymbol{\theta}'|\boldsymbol{\theta}_k)$. The candidate parameters are then retained with probability

$$\alpha(\boldsymbol{\theta}_k, \boldsymbol{\theta}') = \min\left\{\frac{L(y|\boldsymbol{\theta}')\, f(\boldsymbol{\theta}')\, q(\boldsymbol{\theta}_k|\boldsymbol{\theta}')}{L(y|\boldsymbol{\theta}_k)\, f(\boldsymbol{\theta}_k)\, q(\boldsymbol{\theta}'|\boldsymbol{\theta}_k)}, 1\right\}, \tag{2.53}$$

and rejected otherwise. In the event that the candidate parameters are retained, we set $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}'$, otherwise we set $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k$. This process is made precise by Algorithm 4.

Provided the proposal distribution satisfies certain regularity conditions, the underlying distribution of the samples $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n$ is guaranteed to converge to $f(\boldsymbol{\theta}|\boldsymbol{y})$ [Gilks et al., 1996]. A common choice for the proposal distribution (which satisfies these conditions) is a multivariate normal distribution, with mean $\boldsymbol{\theta}_k$ and pre-determined covariance. The stationary distribution of the generated samples is the posterior distribution. In practice, convergence of the generated samples to the stationary distribution manifests as an initial transient period, referred to as *burn-in*. The burn-in phase is generally accounted for by allowing the algorithm to run for a large number of iterations and then discarding the samples which were obtained before the chain reached equilibrium. It is important to note that the choice of proposal

Algorithm 4: The Metropolis-Hastings algorithm.

```
Data: Observed data $\boldsymbol{y}$, $L(\boldsymbol{y}|\boldsymbol{\theta})$, $f(\boldsymbol{\theta})$, and $q(\boldsymbol{\theta}|x)$.
Result: Samples $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ from the posterior distribution.
Randomly sample $\boldsymbol{\theta}_0$ from $f(\boldsymbol{\theta})$ ;
for $k = 1, \ldots, n-1$ do
    Sample $\boldsymbol{\theta}'$ from $q(\boldsymbol{\theta}|\boldsymbol{\theta}_{k-1})$ ;
    Calculate $\alpha(\boldsymbol{\theta}_k, \boldsymbol{\theta}')$ ;
    Sample a uniform number, $u$, on $[0, 1]$ ;
    if $u < \alpha(\boldsymbol{\theta}_k, \boldsymbol{\theta}')$ then
        Set $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}'$ ;
    else
        Set $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k$ ;
    end
end
```

distribution influences the speed at which the algorithm converges to the posterior distribution, but the resulting estimate of the posterior distribution is independent of the choice of proposal distribution [Chibb and Greenberg, 1995].

2.4 The partially-observed SEIR CTMC

Most compartmental Markovian models make a number of unrealistic assumptions about the population which it is modelling. However, there are some disease systems where an SIR CTMC model may be sufficient, especially given the additional complexity of adding an extra compartment. In this section, we introduce a partially-observed SEIR CTMC which relaxes the assumptions that all individuals become infectious immediately after an infectious contact, and that all infectious cases are observed. We do so by including an additional *exposed* compartment in the model, and assuming the individuals who become infectious are *observed* with probability p, and are otherwise *unobserved*.

In particular, the partially-observed SEIR CTMC is a compartmental model in which individuals are classified as susceptible (S), exposed but not infectious (E), infectious and observed (I_o), infectious but unobserved (I_u), removed from the observed infectious class (R_o), and removed from the unobserved infectious class (R_u). Although it is not necessary to partition the removed compartment into observed and unobserved, doing so allows us to transform between the DA representation and the population representation. Furthermore, the inclusion of two distinct removed classes has no impact on the efficiency of the model.

Under the assumption that the population is closed, we have that $S + E + I_o + I_u + R_o + R_u = N$. Thus, we need only model five compartments as the sixth can be determined from the rest. There are only three kinds of possible transitions in the partially-observed SEIR CTMC: exposure events, infection events, and removal events. The rate at which susceptible individuals are exposed to infection is typically specified as

$$\lambda = \frac{1}{N-1}\left(\beta_o I_o + \beta_u I_u\right),$$

where $\beta_o I_o$ and $\beta_u I_u$ are the rates at which individuals have transmissible contacts with individuals of the observed and unobserved infectious classes, respectively. The rate at which an exposed individual transitions to an infectious class is α, making $1/\alpha$ the average latent period. The instant that an exposed individual becomes infectious, the event is observed with probability p, and unobserved otherwise. Thus, the rate at which an exposed individual transitions to the observed infectious class is $p\alpha$, and the rate at which an exposed individual transitions to the unobserved infectious class is $(1-p)\alpha$. The rate at which an observed and unobserved infectious individual is removed is γ_o and γ_u, respectively. These dynamics are summarised in Figure 2.3.

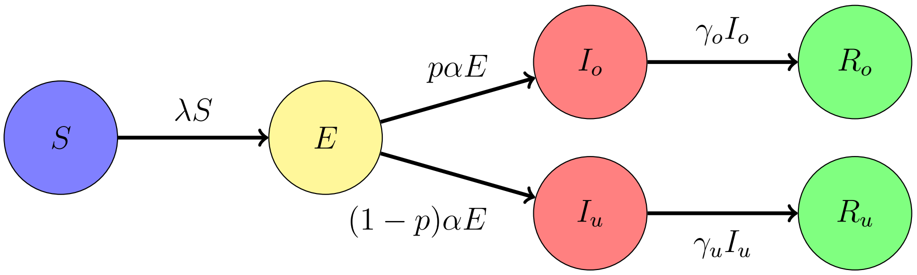


Figure 2.3: State transitions of the partially-observed SEIR CTMC.

Recall that the basic reproductive number, R_0, is the average number of new infections caused by a single infectious individual in an otherwise susceptible population. For the SEIR CTMC, we have that

$$R_0 = p\frac{\beta_o}{\gamma_o} + (1-p)\frac{\beta_u}{\gamma_u}.$$

Let $(\boldsymbol{X}(t), t \geq 0)$ denote the partially-observed SEIR CTMC, which takes values $\boldsymbol{x}$ in

$$\mathcal{X} = \left\{(S, E, I_o, I_u, R_o) \in \mathbb{Z}_+^5 \, : \, S + E + I_o + I_u + R_o \leq N\right\}.$$

The only possible transitions change the state of the process by

$$\begin{aligned}
\boldsymbol{\ell}_1 &= (-1, +1, 0, 0, 0) &&\text{(an exposure event)},\\
\boldsymbol{\ell}_2 &= (0, -1, +1, 0, 0) &&\text{(an observed infection event)},\\
\boldsymbol{\ell}_3 &= (0, -1, 0, +1, 0) &&\text{(an unobserved infection event)},\\
\boldsymbol{\ell}_4 &= (0, 0, -1, 0, +1) &&\text{(an unobserved removal event)},\\
\boldsymbol{\ell}_5 &= (0, 0, 0, -1, 0) &&\text{(an unobserved removal event)}. \qquad (2.54)
\end{aligned}$$

Thus, for all $\boldsymbol{x}$ in $\mathcal{X}$, the transition rates of the SEIR CTMC are

$$\begin{aligned}
q^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{x}+\boldsymbol{\ell}_1} &= \frac{S}{N-1}\left(\beta_o I_o + \beta_u I_u\right) &&\text{if } \boldsymbol{x}+\boldsymbol{\ell}_1 \in \mathcal{X},\\
q^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{x}+\boldsymbol{\ell}_2} &= p\alpha E &&\text{if } \boldsymbol{x}+\boldsymbol{\ell}_2 \in \mathcal{X},\\
q^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{x}+\boldsymbol{\ell}_3} &= (1-p)\alpha E &&\text{if } \boldsymbol{x}+\boldsymbol{\ell}_3 \in \mathcal{X},\\
q^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{x}+\boldsymbol{\ell}_4} &= \gamma_o I_o &&\text{if } \boldsymbol{x}+\boldsymbol{\ell}_4 \in \mathcal{X},\\
q^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{x}+\boldsymbol{\ell}_5} &= \gamma_u I_u &&\text{if } \boldsymbol{x}+\boldsymbol{\ell}_5 \in \mathcal{X}, \qquad (2.55)
\end{aligned}$$

with the additional requirement that $q^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{x}} = -\sum_{\boldsymbol{y}\neq\boldsymbol{x}} q^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{y}}$. We now discuss inference in the framework of the partially-observed SEIR CTMC.

2.4.1 Inferring the basic reproductive number

The SEIR CTMC may be thought of as a direct generalisation of the SIR CTMC to include an additional state of exposure, and in which infection events are only partially-observed. Thus, the methodology described in Section 2.3.3 carries over to the SEIR CTMC with few modifications.

Let $(\boldsymbol{N}(t), t \geq 0)$ denote the DA representation of the SEIR CTMC. The DA process tracks the number of exposure events (N_e), the number of observed infection events (N_{io}), the number of unobserved infection events (N_{iu}), the number of observed removal events (N_{ro}), and the number of unobserved removal events (N_{ru}), on the state space

$$\begin{aligned}\mathcal{N} = \big\{\boldsymbol{n} \in \mathbb{Z}_+^5 \,:\, & N_e, N_{io}, N_{iu}, N_{ro}, N_{ru} \leq N, \\ & N_e \geq N_{io} + N_{iu},\ N_{io} \geq N_{ro},\ N_{iu} \geq N_{ru}\big\}. \end{aligned} \tag{2.56}$$

The DA process is equivalent to the SEIR CTMC, and we can map between the two using the transformation

$$\begin{aligned} N_e &= N - S, & S &= N - N_e, \\ N_{io} &= I_o + R_o, & E &= N_e - N_{io} - N_{iu}, \\ N_{iu} &= I_u + R_u, & I_o &= N_{io} - N_{ro}, \\ N_{ro} &= R_o, & I_u &= N_{iu} - N_{ru}, \\ N_{ru} &= R_u, & & \end{aligned} \tag{2.57}$$

It follows that, for all $\boldsymbol{n}$ in $\mathcal{N}$, the transition rates of the DA process are

$$q^{\boldsymbol{N}}_{\boldsymbol{n}\,\boldsymbol{n}+\boldsymbol{e}_1} = \frac{N - N_e}{N-1}\left(\beta_o\left(N_{io} - N_{ro}\right) + \beta_u\left(N_{iu} - N_{ru}\right)\right) \qquad \text{if } \boldsymbol{n}+\boldsymbol{e}_1 \in \mathcal{N}, \tag{2.58}$$

$$q^{\boldsymbol{N}}_{\boldsymbol{n}\,\boldsymbol{n}+\boldsymbol{e}_2} = p\alpha\left(N_e - N_{io} - N_{iu}\right) \qquad \text{if } \boldsymbol{n}+\boldsymbol{e}_2 \in \mathcal{N}, \tag{2.59}$$

$$q^{\boldsymbol{N}}_{\boldsymbol{n}\,\boldsymbol{n}+\boldsymbol{e}_3} = (1-p)\alpha\left(N_e - N_{io} - N_{iu}\right) \qquad \text{if } \boldsymbol{n}+\boldsymbol{e}_3 \in \mathcal{N}, \tag{2.60}$$

$$q^{\boldsymbol{N}}_{\boldsymbol{n}\,\boldsymbol{n}+\boldsymbol{e}_4} = \gamma_o\left(N_{io} - N_{ro}\right) \qquad \text{if } \boldsymbol{n}+\boldsymbol{e}_4 \in \mathcal{N}, \tag{2.61}$$

$$q^{\boldsymbol{N}}_{\boldsymbol{n}\,\boldsymbol{n}+\boldsymbol{e}_5} = \gamma_u\left(N_{iu} - N_{ru}\right) \qquad \text{if } \boldsymbol{n}+\boldsymbol{e}_5 \in \mathcal{N}, \tag{2.62}$$

with the additional requirement that $q^{\boldsymbol{N}}_{\boldsymbol{n}\,\boldsymbol{n}} = -\sum_{\boldsymbol{m}\neq\boldsymbol{n}} q^{\boldsymbol{N}}_{\boldsymbol{n}\,\boldsymbol{m}}$. To ensure that the generator matrix is triangular, the states in the state space of the DA process are ordered *lexicographically*, meaning that the state $\boldsymbol{n}$ proceeds the state $\boldsymbol{n}'$ if and only if

$$n_1 > n_1' \text{ or } n_i = n_i', \text{ for } i = 1, \ldots, j, \text{ and } n_j > n_j', \tag{2.63}$$

where n_i denotes the ith element of $\boldsymbol{n}$. Since the generator matrix $\mathbb{Q}^{\boldsymbol{N}}$ is triangular under this state-ordering, the results of Jenkinson and Goutsias [2012] (Section 2.3.1) carry over to the SEIR CTMC.

Recall that in Section 2.3.3 we assumed that the cumulative incidence counts y_k, for $k = 1, 2, \ldots, n$, corresponded to observations of the N_I compartment. In the framework of the SEIR model we now attribute these observations to the N_{io} compartment. In particular, the likelihood under the

SEIR model is

$$L(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{k=1}^{n} \Pr\left(N_{io}(t_k) = y_k \,|\, \mathcal{Y}_{k\text{-}1}\right), \tag{2.64}$$

where $\mathcal{Y}_k = \{N_{io}(t_{k\text{-}1}) = y_{k\text{-}1}, N_{io}(t_{k\text{-}2}) = y_{k\text{-}2}, \ldots, N_{io}(t_0) = y_0\}$ is the history of the process. For brevity, we again let $L_E^k(\boldsymbol{\theta})$, for $k = 1, 2, \ldots, n$, the probability of the observed data $\Pr\left(N_{io}(t_k) \,|\, \mathcal{Y}_{k\text{-}1}\right)$, which is calculated via an analogous argument to equation (2.39).

$$L_E^k(\boldsymbol{\theta}) = \sum_{\boldsymbol{n}_{k-1} \in \mathcal{N}} \sum_{\boldsymbol{n}_k \in \mathcal{N}} p^{\boldsymbol{N}}_{\boldsymbol{n}_{k\text{-}1}\,\boldsymbol{n}_k}(t_k - t_{k\text{-}1}) \left(\frac{\Pr\left(\boldsymbol{N}\left(t_{k\text{-}1}\right) = \boldsymbol{n}_{k\text{-}1} \,|\, \mathcal{Y}_{k\text{-}2}\right)}{L_E^{k\text{-}1}(\boldsymbol{\theta})} \right), \tag{2.65}$$

where $\boldsymbol{n}_k$ is any state in $\mathcal{N}$ for which $N_{io} = y_k$. The likelihood is therefore computed via a direct generalisation of Algorithm 3, and its parameters may be inferred via either maximum likelihood estimation (Definition 9) or the Metropolis-Hastings algorithm (Algorithm 4).

In the framework of the SIR CTMC, truncating the state space was an effective way of managing the computational-cost of computing the likelihood (2.38). However, in the framework of the partially-observed SEIR CTMC this approach is not as effective. The main reason for this is that the number of states in its state space is $\mathcal{O}\left(N^5\right)$, compared to $\mathcal{O}\left(N^2\right)$ for the SIR CTMC. Furthermore, an observed cumulative incidence count y_k does not influence the size of the N_e, N_{iu} and N_{ru} compartments, allowing these compartments to grow unchecked. Thus, in Chapter 5, we present a hybrid diffusion model of the SEIR CTMC which we utilise for computing the likelihood (5.3). This approach enables us to conduct parameter inference on a range of large real-world outbreaks which would have been intractable under the SEIR CTMC.

Chapter 3

Hybrid approximation of final size and duration distributions for the SIR CTMC

Compartmental continuous-time Markov chain (CTMC) models are of substantial importance to mathematical epidemiology because they account for the stochastic individual-to-individual nature of disease transmission [Bailey, 1957, Keeling et al., 2000, Ball and Donnelly, 1995, Bartlett, 1956, Rand and Wilson, 1991, Fox, 1993, Grenfell et al., 1998, Spagnolo et al., 2003, Coulson et al., 2004]. This is a particularly important feature during the initial stages of an outbreak, when there is a considerable probability that the outbreak will fade out. On the other hand, when working within a CTMC framework, most analyses require computing the solution to systems of equations which generally contain $\mathcal{O}\left(N^d\right)$ equations, where d is the number of compartments and N is the size of the population. Thus, modelling large populations directly with a CTMC is generally considered computationally-infeasible. The aim of this chapter is to investigate accurate and computationally-efficient approaches to analysing features of the SIR CTMC, in the situation where the population size is large.

Recall that in Section 2.2 we introduced the notion that a certain class of CTMCs may be approximated by a large-population approximation [Kurtz, 1970, 1971, Van Kampen, 1961, 2007a, McNeil and Walls, 1974, Kubo et al., 1973, Sjöberg et al., 2009, Van Kampen, 2007b], and that two important large-population approximations are the so-called fluid limit (Theorem 3) and diffusion limit (Theorem 4). The fluid limit provides an approximation of the expected state of the CTMC, while the diffusion limit approximates its probability distribution. Both of these approximations are computationally-efficient and generally accurate, but they break down if the population of at least one compartment of the model is close to zero. It follows that a discrete-state model, such as a CTMC, is indispensable for accurately modelling the initial and final stages stages of an outbreak.

A natural way to approximate the dynamics of a large-population CTMC is to construct a model which utilises discrete dynamics when the population of its compartments are low, and a large-population approximation otherwise. So-called hybrid models have been constructed for a variety of applications, such as improving the efficiency of Monte Carlo methods [Guerrier and Holcman, 2016, Ganguly et al., 2015, Duncan et al., 2016, Angius et al., 2015, Vasudeva and Bhalla, 2003, Takahashi et al., 2004, Hellander and Lötstedt, 2007, Hepp et al., 2015] and computing the solution to the Kolmogorov equations (Definition 4) Safta et al. [2015]. There has been particular interest in using hybrid models to compute quantities from the SIR CTMC [Kermack and McKendrick, 1927, Bartlett, 1949, Bailey, 1950, 1957, Bartlett, 1956, Kendall, 1965, Sazonov et al., 2011, 2017]; also see Section 2.3.1 and Section 2.3.2. These hybrid models generally use an appropriate branching process approximation during the initial stages of the outbreak [Ball and Neal, 2010] and either the diffusion approximation [Scalia-Tomba, 1985, Watson, 1980a, 1981, Nagaev and Startsev, 1970] or fluid approximation [Barbour, 1975, Sazonov et al., 2011] thereafter.

In this chapter we construct two hybrid approximations of the SIR CTMC suitable for modelling large populations, referred to as the *hybrid fluid* model and the *hybrid diffusion* model. These models utilise CTMC dynamics while the number of infectious individuals is below a particular *threshold* and either fluid or diffusion dynamics otherwise. To assess the accuracy of our models, we use them to calculate the distribution of the duration of the outbreak, and the distribution of the final size of the outbreak, which are compared to the approximations of Barbour [Barbour, 1975] and Scalia-Tomba [Scalia-Tomba, 1985], respectively. We also demonstrate the computational advantage of our approach over those based on the SIR CTMC (Section 2.3.1 and Section 2.3.2). As we shall see, the hybrid fluid model provides an accurate representation of the distribution of the duration of the outbreak but fails to accurately capture the distribution of the final size of the outbreak. However, the hybrid diffusion makes up for this shortcoming. The computational runtimes of our hybrid models are $\mathcal{O}\left(N\right)$, which is a significant improvement over the $\mathcal{O}\left(N^2\right)$ runtime of the SIR CTMC. These results encourage extending the hybrid diffusion model to inference in the following chapters.

It is worth noting that in this chapter we discuss the hybrid models in their population representation (Section 2.3) because it is more instructive than their DA representation (Equation (2.20)). However, the numerical implementation of these algorithms is performed in the DA representation in order to preserve the numerical advantages afforded by Algorithm 1 and Algorithm 2. The remainder of this chapter is structured as follows: in

Section 3.1 we introduce the hybrid fluid model and use it to compute the distribution of the duration of the outbreak and the distribution of the final size of the outbreak. In Section 3.2 we introduce the hybrid diffusion model and use it to compute the distribution of the final size of the outbreak. Finally, in Section 3.3 we discuss the numerical implementation of these algorithms in the DA representation.

3.1 Hybrid fluid model

We begin by introducing the hybrid fluid model, which is similar to the hybrid model of Barbour [1975]. Where Barbour's model utilises branching process dynamics until the number of infectious individuals exceeds $\sqrt{N}$, and after the number of infectious individuals drops below $N^{1/4}$, our hybrid fluid model utilises CTMC dynamics whenever the number of infectious individuals is below some pre-determined threshold $\widehat{I} \in \{1, 2, \ldots, N\}$. During the intermediate stages, both Barbour's model and our hybrid fluid model utilise the fluid approximation.

3.1.1 Model formulation

Before we define the hybrid fluid model, it is instructive to recall that the SIR CTMC, $(\boldsymbol{X}(t), t \geq 0)$, takes values in $\mathcal{X}$ and, for all $\boldsymbol{x}$ in $\mathcal{X}$, has the positive transition rates $q^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{x}+\boldsymbol{\ell}}$, if $\boldsymbol{x} + \boldsymbol{\ell}$ is in $\mathcal{X}$, and $\boldsymbol{\ell}$ is either $\boldsymbol{\ell}_1 = (-1, 1)$ or $\boldsymbol{\ell}_2 = (0, -1)$ (Section 2.3). The hybrid fluid model may simply be thought of as a version of the SIR CTMC whose dynamics over the set of states with $I \geq \widehat{I}$ are approximated by the fluid limit (Theorem 3). More precisely, let $\boldsymbol{Y}(t)$, for $t \geq 0$, denote the hybrid fluid, which takes values in the hybrid discrete-continuous state space $\mathcal{Y}$, defined as the union of the discrete lattice

$$\mathcal{Y}^{MC} = \left\{ (S, I) \in \mathcal{X} \,:\, I \leq \widehat{I} \right\}$$

and the continuum

$$\mathcal{Y}^{DE} = \left\{ (S, I) \in \mathbb{R}^2_+ \,:\, S + I \leq N,\, I \geq \widehat{I} \right\}.$$

When $\boldsymbol{Y}(t)$ is in the subset $\mathcal{Y}^{MC}$, it has the dynamics of $\boldsymbol{X}(t)$, and when $\boldsymbol{Y}(t)$ is in the subset $\mathcal{Y}^{DE}$ it has the dynamics of the fluid approximation $N\,\boldsymbol{x}\,(t, \boldsymbol{x}_0/N)$ (Equation (2.28)), given an appropriate initial state $\boldsymbol{x}_0$ in $\mathcal{Y}$. The dynamics of $\boldsymbol{Y}(t)$ at the intersection of $\mathcal{Y}^{MC}$ and $\mathcal{Y}^{DE}$, denoted $\mathcal{T}^{MC}$, require careful consideration.

Recall that the fluid approximation is governed by the system of differential equations (2.28). According to these equations, the rate of change of I with respect to time is positive if $S > \eta N$, where $\eta = \gamma/\beta$. This means that if $\boldsymbol{Y}(t)$ hits the state $(S, \widehat{I})$ in $\mathcal{T}^{MC}$, where $S > \eta N$, then the fluid dynamics will immediately force $\boldsymbol{Y}(t)$ out of $\mathcal{T}^{MC}$ and into $\mathcal{Y}^{DE}$. In contrast, if $S \leq \eta N$ then the fluid dynamics will force $\boldsymbol{Y}(t)$ to remain in its current state until a removal event occurs. Thus, we define

$$\mathcal{T}^{MC}_1 = \left\{ (S, \widehat{I}) \in \mathcal{Y}^{MC} \,:\, S = \lfloor \eta N \rfloor + 1, \ldots, N - \widehat{I} \right\},$$

as the set of states which force $\boldsymbol{Y}(t)$ to switch from CTMC dynamics to fluid dynamics and

$$\mathcal{T}_2 = \left\{ (S, \widehat{I}) \in \mathcal{Y}^{DE} \,:\, S \in [0, \eta N] \right\}$$

as the set of states which force $\boldsymbol{Y}(t)$ to switch from fluid dynamics to CTMC dynamics. We denote the integer components of $\mathcal{T}_2$ as $\mathcal{T}^{MC}_2$ which is defined as the intersection of $\mathcal{Y}^{MC}$ and $\mathcal{T}_2$.

Given that the fluid approximation is a deterministic process, we are able to deduce some important features of the behaviour of $\boldsymbol{Y}(t)$ on $\mathcal{Y}^{DE}$.

Recall that by considering the trajectory of the fluid approximation through the (s, i) plane, one can deduce a relationship between $s(t)$ and $i(t)$, given an initial value $s(0)$ and $i(0)$ (equation (2.29)). Furthermore, since $s(t)$ is monotonically decreasing, one can deduce the amount of time which elapses while $a \leq s(t) \leq b$, for a, b in $[0, 1]$, see equation (2.31). It follows that if $\boldsymbol{Y}(t)$ hits the state $\boldsymbol{x} = (S, \widehat{I})$ in $\mathcal{T}_1^{MC}$, then the state in $\mathcal{T}_2$ where the fluid dynamics terminate is $(S(\boldsymbol{x}), \widehat{I})$, where $S(\boldsymbol{x})/N$ is the non-trivial solution to equation (2.30) with $s_0 = S/N$ and $i_0 = \widehat{I}/N$. Furthermore, the duration of the fluid dynamics is given by $J(S/N, S(\boldsymbol{x})/N)$ from equation (2.31), which we denote $t(\boldsymbol{x})$.

Since the fluid approximation is a continuous-state process and the SIR CTMC is a discrete-state process, a discretisation mapping must occur when $\boldsymbol{Y}(t)$ switches from fluid dynamics to CTMC dynamics. As the fluid dynamics provide no measure of the variability of the underlying CTMC, we decided to discretise the number of susceptible individuals $S_2(\boldsymbol{x})$ as follows:

$$\begin{aligned}
&\text{round down to } \lfloor S_2(\boldsymbol{x}) \rfloor && \text{with probability } 1 - (S_2(\boldsymbol{x}) - \lfloor S_2(\boldsymbol{x}) \rfloor),\\
&\text{round up to } \lfloor S_2(\boldsymbol{x}) \rfloor + 1 && \text{with probability } (S_2(\boldsymbol{x}) - \lfloor S_2(\boldsymbol{x}) \rfloor).
\end{aligned} \tag{3.1}$$

Under the assumption that the population is large, the difference between rounding up or down is negligible. Finally, it is important to note that the only CTMC events possible from states in $\mathcal{T}_2^{MC}$ are removal events.

Figure 3.1 is a state-transition diagram of the hybrid fluid model for a population of $N = 15$ individuals with a threshold of $\widehat{I} = 3$. The green points are states from the discrete set $\mathcal{Y}^{MC}$, and the continuum $\mathcal{Y}^{DE}$ is the region with $I \geq \widehat{I}$, and $S \leq N - I$, with the threshold sets $\mathcal{T}_1^{MC}$ and $\mathcal{T}_2^{MC}$ represented by the green upward and downward pointing triangles, respectively. The state space $\mathcal{Y}$ is the union of $\mathcal{Y}^{MC}$ and $\mathcal{Y}^{DE}$. The black arrows represent the

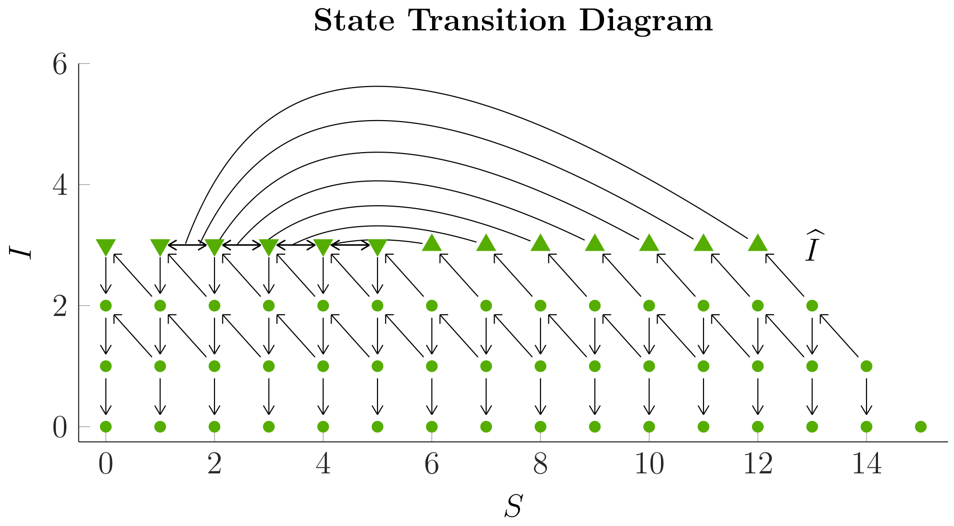


Figure 3.1: The state-transition diagram of the hybrid fluid model with $N = 15$ and $\widehat{I} = 3$. The green points are the discrete states in $\mathcal{Y}^{MC}$, and the continuum $\mathcal{Y}^{DE}$ is the set of states with $I \geq \widehat{I}$, and $S \leq N - I$. The upward (downward) pointing triangles are states from which $\boldsymbol{Y}(t)$ switches from CTMC to fluid (fluid to CTMC) dynamics, which are contained in the set $\mathcal{T}_1^{MC}$ ($\mathcal{T}_2^{MC}$). The black curves emanating from states in $\mathcal{T}_1^{MC}$ are the deterministic trajectories of $N\boldsymbol{x}(t, \boldsymbol{x}_0/N)$, for $\boldsymbol{x}_0$ in $\mathcal{T}_1^{MC}$, through $\mathcal{Y}^{DE}$.

possible transitions of the model. Of particular interest are the trajectories of $N\boldsymbol{x}\left(t, \boldsymbol{x}_0/N\right)$, for $\boldsymbol{x}_0$ in $\mathcal{T}_1^{MC}$. These trajectories are shown by the black curves emanating from states in $\mathcal{T}_1^{MC}$, which amount to a deterministic transition from states in $\mathcal{T}_1^{MC}$ to states in $\mathcal{T}_2^{MC}$ (equation (2.30)). The duration of each of these trajectories is calculated from equation (2.31).

We now consider using the hybrid fluid model to compute the duration of the outbreak, and the distribution of the final size of the outbreak.

3.1.2 Outbreak duration

A system of delayed differential equations (DDE)s describing the transition probabilities of $\boldsymbol{Y}(t)$ for states in $\mathcal{Y}^{MC}$ is derived by separately considering the flux of probability on three disjoint subsets of $\mathcal{Y}^{MC}$. Within each of these subsets, the flux of probability between states in $\mathcal{Y}^{MC}$ must be treated differently due to the way in which probability flows between $\mathcal{Y}^{MC}$ and $\mathcal{Y}^{DE}$. In the first scenario we consider the set $\mathcal{D} = \mathcal{Y}^{MC} \setminus (\mathcal{T}_1^{MC} \cup \mathcal{T}_2^{MC})$, on which the fluid dynamics have no effect. In the second and third scenarios we consider the sets $\mathcal{T}_1^{MC}$ and $\mathcal{T}_2^{MC}$ on which probability flows from $\mathcal{Y}^{MC}$ to $\mathcal{Y}^{DE}$, and from $\mathcal{Y}^{DE}$ to $\mathcal{Y}^{MC}$, respectively. The resulting system of DDEs allow us to calculate the transition probabilities of $\boldsymbol{Y}(t)$ on $\mathcal{Y}^{MC}$, for $t \geq 0$, which may be used for computing the distribution of the duration of the outbreak in a similar way to the Kolmogorov equations (Section 2.3.1).

Scenario 1

The flux of probability on states in $\mathcal{D}$ is not affected by the fluid dynamics of $\boldsymbol{Y}(t)$, so it is governed by the Kolmogorov equations (Definition (4)). Thus, for all $\boldsymbol{x}$ in $\mathcal{D}$, the transition probabilities satisfy

$$\frac{d}{dt} p^{\boldsymbol{Y}}_{\boldsymbol{x}\,\boldsymbol{z}}(t) = \sum_{\boldsymbol{y} \in \mathcal{Y}^{MC}} p^{\boldsymbol{Y}}_{\boldsymbol{x}\,\boldsymbol{y}}(t)\, q^{\boldsymbol{X}}_{\boldsymbol{y}\,\boldsymbol{z}}, \tag{3.2}$$

if $\boldsymbol{z}$ is in $\mathcal{Y}^{MC}$.

Scenario 2

We now consider the flux of probability for states in $\mathcal{T}_1^{MC}$. On this subset, probability flows from states in $\mathcal{Y}^{MC}$ into states in $\mathcal{Y}^{DE}$. Since the transition from CTMC dynamics to fluid dynamics is instantaneous, the flux of probability into states in $\mathcal{T}_1^{MC}$ is always equal to the flux of probability out. Thus, for all $\boldsymbol{x}$ in $\mathcal{D}$, we have that $p^{\boldsymbol{Y}}_{\boldsymbol{x}\,\boldsymbol{y}}(t) = 0$, if $\boldsymbol{y}$ is in $\mathcal{T}_1^{MC}$ and $t > 0$.

Scenario 3

We now consider the flux of probability for states in $\mathcal{T}_2^{MC}$. Probability flows into states in $\mathcal{T}_2^{MC}$ both from states in $\mathcal{Y}^{MC}$, and from trajectories through $\mathcal{Y}^{DE}$. In the former case, the probability flux is not affected by the fluid dynamics so it is governed by the Kolmogorov equations; however, the latter case requires careful consideration.

Due to the deterministic nature of the fluid process, we know that the flux of probability into the state $\boldsymbol{x}$ in $\mathcal{T}_1^{MC}$, at time t, is distributed amongst two corresponding states in $\mathcal{T}_2^{MC}$ (equation (3.1)) after a fixed delay of $t\,(\boldsymbol{x})$ time units (equation (2.31)). For all $\boldsymbol{x}$ in $\mathcal{T}_1^{MC}$, let $p^{F}_{\boldsymbol{x}\,\boldsymbol{y}}$ denote the probability that the hybrid fluid process switches from fluid dynamics to CTMC dynamics through the state $\boldsymbol{y}$, in $\mathcal{T}_2^{MC}$, given that it switched from CTMC dynamics to fluid dynamics through the state $\boldsymbol{x}$ (equation (3.1)). Then, for all $\boldsymbol{x}$ in $\mathcal{D}$, conditioned on the event that $\boldsymbol{Y}(t)$ hits a state in $\mathcal{T}_1^{MC}$ we have that

$$\frac{d}{dt}p^{\boldsymbol{Y}}_{\boldsymbol{x}\,\boldsymbol{z}}(t) = \sum_{\boldsymbol{y}\in\mathcal{T}_1^{MC}} \mathbb{1}\{t \geq t(\boldsymbol{y})\}\, p^{\boldsymbol{Y}}_{\boldsymbol{x}\,\boldsymbol{y}-\boldsymbol{\ell}_1}(t - t(\boldsymbol{y}))\, q^{\boldsymbol{X}}_{\boldsymbol{y}-\boldsymbol{\ell}_1\,\boldsymbol{y}}\, p^{F}_{\boldsymbol{y}\,\boldsymbol{z}},$$

if $\boldsymbol{z}$ is in $\mathcal{T}_2^{MC}$, where $\mathbb{1}\{.\}$ is the indicator function. Thus, for all $\boldsymbol{x}$ in $\mathcal{D}$, the transition probabilities satisfy

$$\begin{aligned}\frac{d}{dt}p^{\boldsymbol{Y}}_{\boldsymbol{x}\,\boldsymbol{z}}(t) &= \sum_{\boldsymbol{y}\in\mathcal{D}} p^{\boldsymbol{Y}}_{\boldsymbol{x}\,\boldsymbol{y}}(t)\, q^{\boldsymbol{X}}_{\boldsymbol{y}\,\boldsymbol{z}} \\ &\qquad + \sum_{\boldsymbol{y}\in\mathcal{T}_1^{MC}} \mathbb{1}\{t \geq t(\boldsymbol{y})\}\, p^{\boldsymbol{Y}}_{\boldsymbol{x}\,\boldsymbol{y}-\boldsymbol{\ell}_1}(t - t(\boldsymbol{y}))\, q^{\boldsymbol{X}}_{\boldsymbol{y}-\boldsymbol{\ell}_1\,\boldsymbol{y}}\, p^{F}_{\boldsymbol{y}\,\boldsymbol{z}}, \quad (3.3)\end{aligned}$$

if $\boldsymbol{z}$ is in $\mathcal{T}_2^{MC}$. It is natural to think that $q^{\boldsymbol{X}}_{\boldsymbol{y}\,\boldsymbol{z}} = 0$ for all $\boldsymbol{z}$ in $\mathcal{T}_2^{MC}$. However, it is worth noting that this quantity is positive for the states $\boldsymbol{z} - \boldsymbol{\ell}_1$ in $\mathcal{D}$.

For all $\boldsymbol{x}$ in $\mathcal{D}$, the system of DDEs (3.2)—(3.3) are integrated numerically on the set of equally-spaced time points $t_0, t_1, \ldots, t_n$, with spacing τ, using an adapted version of the Implicit Euler scheme (Section 2.3.1). The adaption is that if $\boldsymbol{z}$ in is $\mathcal{T}_2^{MC}$, then the transition probabilities are incremented at each time step by

$$\begin{aligned}&p^{\boldsymbol{Y}}_{\boldsymbol{x}\,\boldsymbol{z}}(t_{k+1}) = p^{\boldsymbol{Y}}_{\boldsymbol{x}\,\boldsymbol{z}}(t_k)+ \\ &\tau\left(\sum_{\boldsymbol{y}\in\mathcal{D}} p^{\boldsymbol{Y}}_{\boldsymbol{x}\,\boldsymbol{y}}(t)\, q^{\boldsymbol{X}}_{\boldsymbol{y}\,\boldsymbol{z}} + \sum_{\boldsymbol{y}\in\mathcal{T}_1^{MC}} \mathbb{1}\{t_{k+1} \geq t(\boldsymbol{y})\}\, p^{\boldsymbol{Y}}_{\boldsymbol{x}\,\boldsymbol{y}-\boldsymbol{\ell}_1}(t_{k+1} - t(\boldsymbol{y}))\, q^{\boldsymbol{X}}_{\boldsymbol{y}-\boldsymbol{\ell}_1\,\boldsymbol{y}}\, p^{F}_{\boldsymbol{y}\,\boldsymbol{z}}\right).\end{aligned}$$

Recall from Section 2.3 that $\mathcal{A}$ is the subset of $\mathcal{X}$ in which $I = 0$ and the random variable T describes the duration of the outbreak. Then, it follows that

$$\Pr(T \leq t_k) = \sum_{\boldsymbol{z}\in\mathcal{A}} p^{\boldsymbol{Y}}_{(N-1,1)\,\boldsymbol{z}}(t_k).$$

In Section 3.3 we adapt Algorithm 1 to the hybrid fluid model.

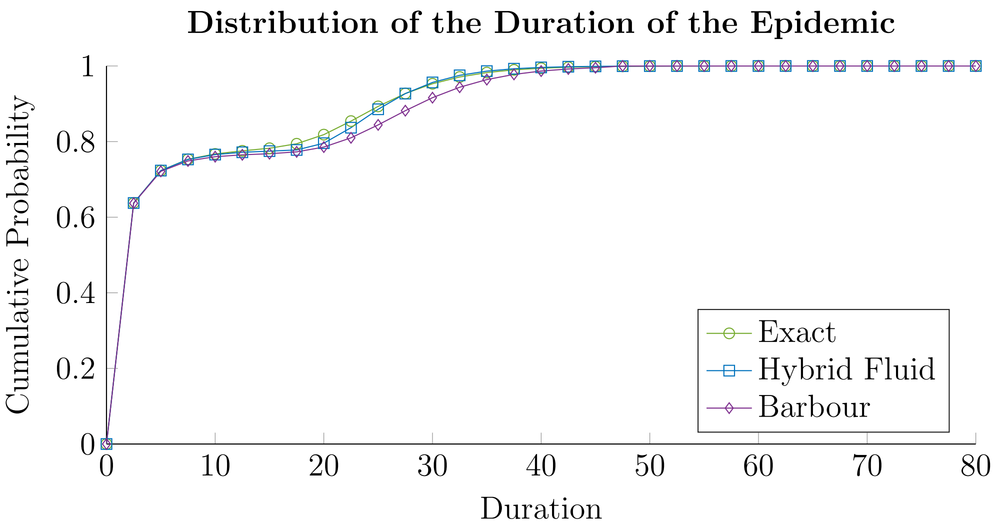


Figure 3.2: The distribution of the duration of the epidemic calculated from the CTMC model, hybrid fluid model, and Barbour's model for $R_0 = 1.3$ and $N = 1,000$ with one initially infectious individual.

3.1.3 Numerical results

We now compare the distribution of the duration of the outbreak from the SIR CTMC to the distribution of the duration of the outbreak from the hybrid fluid model and Barbour's hybrid model (Theorem 7). We fix the basic reproductive number $R_0 = 1.3$ and the initial state $(N-1, 1)$, and compute the distribution of the duration of the outbreak on a temporal grid ranging from 0 to 80 in steps of $\tau = 0.01$. Under this construction, $\Pr(T \leq 80) \approx 1$ provided $N \leq 10,000$, and the global L_1-error of the Implicit Euler scheme is $\mathcal{O}\left(10^{-2}\right)$. We fix the threshold as $\widehat{I} = 17$ because our procedure for selecting an appropriate threshold, to be outlined in Section 3.3.2, guarantees a certain level of accuracy, when compared to the distribution of the duration of the outbreak from the SIR CTMC.

Figure 3.2 shows the distribution of the duration of the epidemic calculated from the SIR CTMC (green with circles), hybrid fluid model (blue with

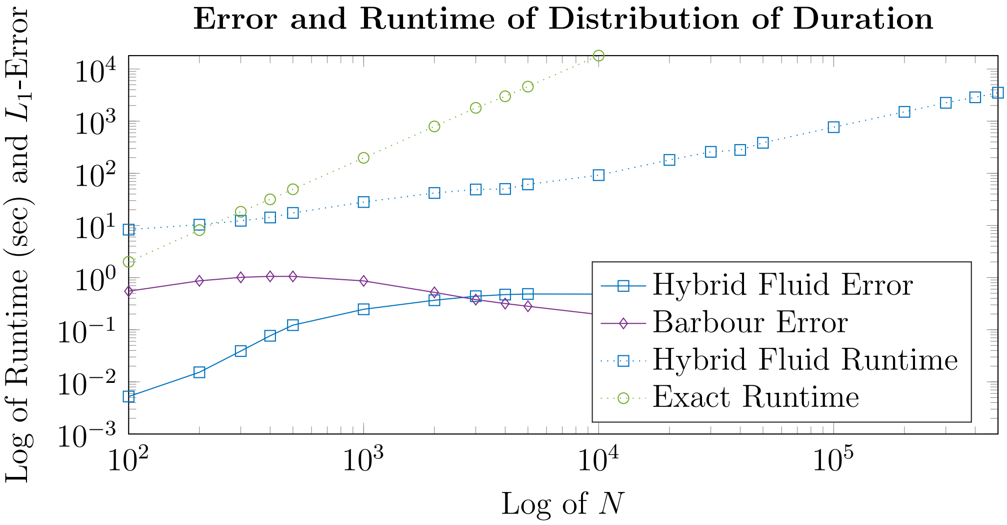


Figure 3.3: Required runtime of the distribution of the duration of the outbreak from the SIR CTMC and the hybrid fluid model alongside the L_1-error of the hybrid fluid model and Barbour's hybrid model. The L_1-error of the hybrid fluid model is less than the L_1-error of Barbour's hybrid model for $N \leq 2000$, but the two are virtually the same for $N \geq 10^3$. The required runtime of the hybrid fluid model if $\mathcal{O}\left(N\right)$. Again, we have that $R_0 = 1.3$, $\widehat{I} = 17$ (inequality (3.8)), and the initial state $(N-1, 1)$.

squares), and Barbour's hybrid model (purple with diamonds) for $N = 1000$. Both models provide a reasonable approximation to the distribution of the duration of the epidemic from the SIR CTMC over the whole domain of t. However, it can be seen that the hybrid fluid model provides a more accurate representation of the duration of outbreaks which become established.

Figure 3.3 shows a log-log plot of the required runtime of the SIR CTMC model (dotted green with circles) and the required runtime of the hybrid fluid model (dotted blue with squares) for a range of values of N from 10^2 to 10^6. The slope of the line from the hybrid fluid model is approximately one, which indicates that the asymptotic runtime for using Algorithm 5 on the hybrid fluid model to calculate the distribution of the duration of the epidemic is $\mathcal{O}(N)$. This is because the runtime of Algorithm 5 is dependent on the total number of states, which for the hybrid fluid model is approximately $\widehat{I}N$. Irrespective of the population size, Barbour's asymptotic approximation is effectively instantaneous to compute so its runtime has not been included in Figure 3.3.

Figure 3.3 also shows a log-log plot of the L_1-error of the hybrid fluid model (solid blue with squares) and the L_1-error of Barbour's model (solid purple with diamonds). The L_1-error of the hybrid fluid model is favourable to Barbour's for N of $\mathcal{O}\left(10^2\right)$. However, the two approximations are effectively indistinguishable for $N \geq 10^3$, despite the important difference that the hybrid model utilises a fixed threshold and Barbour's utilises a variable $\sqrt{N}$. The L_1-error of the hybrid fluid model appears to increase with N which suggests that the main source of disagreement between the SIR CTMC and the hybrid fluid model is the length of time over which the CTMC is approximated by the fluid model. Although the L_1-error of the hybrid fluid approximation can generally be improved by increasing the threshold $\widehat{I}$, the hybrid fluid approximation does not show a significant improvement over Barbour's asymptotic approximation unless $\widehat{I}$ is large enough that the probability of $\boldsymbol{Y}\left(t\right)$ hitting the subset $\mathcal{Y}^{DE}$ is insignificant (results not shown).

3.1.4 Final outbreak size

Recall that the time-homogeneity property (Definition 2) of the SIR CTMC enabled us to deduce the distribution of the final size of the outbreak from its embedded jump chain process (Definition 5). Since the dynamics of the hybrid fluid model are time-homogeneous, we are able to deduce the distribution of the final size of the outbreak from the hybrid fluid model from its embedded jump process.

Recall that the embedded jump process of the SIR CTMC is the DTMC which takes values in $\mathcal{X}$ and, for all $\boldsymbol{x}$ in $\mathcal{X}$, has the transition probabilities $p^{\boldsymbol{Y}}_{\boldsymbol{x}\,\boldsymbol{x}+\boldsymbol{\ell}} = q^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{x}+\boldsymbol{\ell}}/|q^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{x}}|$, for $\boldsymbol{\ell}$ equal to $\boldsymbol{\ell}_1$ or $\boldsymbol{\ell}_2$ and if $\boldsymbol{x}+\boldsymbol{\ell}$ is in $\mathcal{X}$. The embedded jump process of the hybrid fluid process is the DTMC process, $(\mathbf{Y}_n, n \geq 0)$, which takes values in $\mathcal{Y}^{MC}$, with the transition probabilities

$$\begin{aligned}
p^{\boldsymbol{Y}}_{\boldsymbol{x}\,\boldsymbol{x}+\boldsymbol{\ell}} &= p^{\boldsymbol{X}}_{\boldsymbol{x}\,\boldsymbol{x}+\boldsymbol{\ell}} && \text{for all } \boldsymbol{x} \in \mathcal{D} \text{ if } \boldsymbol{x} \in \mathcal{Y}^{MC},\\
p^{\boldsymbol{Y}}_{\boldsymbol{x}\,\boldsymbol{y}} &= p^{F}_{\boldsymbol{x}\,\boldsymbol{y}} && \text{for all } \boldsymbol{x} \in \mathcal{T}^{MC}_1 \text{ if } \boldsymbol{y} \in \mathcal{T}^{MC}_2,\\
p^{\boldsymbol{Y}}_{\boldsymbol{x}\,\boldsymbol{x}+\boldsymbol{\ell}_2} &= 1 && \text{for all } \boldsymbol{x} \in \mathcal{T}^{MC}_2.
\end{aligned}$$

The distribution of the final size of the outbreak may be computed from the hybrid fluid model via the hitting probabilities of the embedded jump process on the set $\mathcal{A}$ (Definition 6). Fix $\boldsymbol{x}$ in $\mathcal{B}$, and let $h^{\boldsymbol{Y}}_{\boldsymbol{x}\,\boldsymbol{y}}$ denote the hitting probability of any state $\boldsymbol{y}$ in $\mathcal{X}$, from the state $\boldsymbol{x}$. Then, for all $\boldsymbol{y}$ in $\mathcal{Y}^{MC}$, the hitting probabilities are the minimal non-negative solution to,

$$h^{\boldsymbol{Y}}_{\boldsymbol{x}\,\boldsymbol{y}} = \sum_{\boldsymbol{z} \in \mathcal{Y}^{MC}} p^{\boldsymbol{Y}}_{\boldsymbol{x}\,\boldsymbol{z}}\, h^{\boldsymbol{Y}}_{\boldsymbol{z}\,\boldsymbol{y}}, \tag{3.4}$$

where $h^{\boldsymbol{Y}}_{\boldsymbol{z}\,\boldsymbol{z}} = 1$, for all $\boldsymbol{z} \in \mathcal{A}$. It follows that the distribution of the final size of the outbreak, given the initial state $\boldsymbol{x} = (N-1, 1)$, is the $(N+1) \times 1$ vector $(h^{\boldsymbol{Y}}_{\boldsymbol{x}\,\boldsymbol{y}} : \boldsymbol{y} \in \mathcal{A})$. We compute the distribution of the final size of the outbreak using a modified version of Algorithm 2 which is discussed in Section 3.3.

3.1.5 Numerical results

We now compare the distribution of the final size of the outbreak from the SIR CTMC to the distribution of the final size of the outbreak from the hybrid fluid model, using the same parameters as before ($R_0 = 1.3$, $\boldsymbol{X}(0) = (N-1, 1)$), with $N = 1,000$.

Figure 3.4 shows the distribution of the final size of the outbreak from the SIR CTMC (green with circles) and the hybrid fluid model (blue with squares). The hybrid fluid model provides an accurate representation of the distribution of the final size of the outbreak, if the outbreak fades out. However, it provides a poor approximation of the distribution of the final size of the outbreak if the outbreak becomes established.

Figure 3.5 shows the required runtime of the SIR CTMC (dotted green with circles) and the runtime of the hybrid fluid model (dotted blue with squares) across a range of values of N (from 10^3 to 10^8). The asymptotic slope of the curve of the required runtime for the hybrid fluid model is approximately one, which indicates that the asymptotic runtime of computing the distribution of the final size of the outbreak is $\mathcal{O}(N)$.

Figure 3.5 also shows the L_1-error of the hybrid fluid model (solid blue with squares). The L_1-error of the hybrid fluid model appears to converge as $N \to \infty$, to a value around 66% of the largest possible L_1-error, suggesting that $\boldsymbol{Y}(t)$ approximates well the $\eta = 0.34$ proportion of sample paths which become extinct close to $S = N$, but fails to approximate the $1 - \eta$ proportion of sample paths which become extinct near $S = 0$. This confirms our intuition that the source of disagreement between $\boldsymbol{Y}(t)$ and $\boldsymbol{X}(t)$ propagates from the time interval over which the fluid approximation is used to approximate the underlying CTMC. Although $\widehat{I} = 17$ has been identified as a reasonable threshold (inequality (3.8) in Section 3.3.2), the asymptotic error may generally be decreased by selecting a larger threshold. However, the L_1-error is fairly insensitive to changing the threshold.

So far, we have used the hybrid fluid model to calculate the distribution of the duration of the outbreak and the distribution of the final size of the outbreak. We have found that the hybrid fluid model provides an accurate representation of the distribution of the duration of the outbreak but provides a poor approximation of the distribution of the final size of the outbreak. This is because the fluid limit provides an approximation of the expected state of the underlying CTMC but provides no measure of its state-variability. Thus, we now consider utilising the diffusion limit in place of the fluid limit in order to accurately represent the state-variation of the underlying CTMC.

3.2 Hybrid diffusion model

The hybrid diffusion model may be thought of as a variant of the hybrid fluid model which accounts for the state-variability of the underlying CTMC on the domain $\mathcal{Y}^{DE}$. This is because the hybrid diffusion model is constructed in a similar way to the hybrid fluid model, with the only difference being that the hybrid diffusion model utilises the diffusion limit (Theorem 4) in place of the fluid limit. In this section, we compute the distribution of the final size of the outbreak using the hybrid diffusion model.

3.2.1 Model formulation

Let $\mathbf{Z}\left(t\right)$, for $t \geq 0$, denote the hybrid diffusion process, which takes values in $\mathcal{Y}$. As with the hybrid fluid process, the dynamics of the hybrid diffusion process are determined by which subset of $\mathcal{Y}$ it is in. In particular, when $\mathbf{Z}\left(t\right)$ is in the subset $\mathcal{Y}^{MC}$ it has the dynamics of the SIR CTMC, and when $\mathbf{Z}\left(t\right)$ is in the subset $\mathcal{Y}^{DE}$ it has the dynamics of the diffusion approximation (equation (2.44)). We now discuss the dynamics of the hybrid diffusion process at the interface $\mathcal{T}$.

Recall, for finite $t \geq 0$, that the diffusion approximation of the SIR CTMC is the Gaussian diffusion process with expected value $N\,\boldsymbol{x}\left(t, \boldsymbol{x}_0/N\right)$, for $\boldsymbol{x}_0$

in E, and covariance matrix $N\,\Sigma(t)$. Since the fluid approximation provides the expected value of the diffusion approximation, there is a high probability that if $\mathbf{Z}\,(t)$ hits a state in $\mathcal{T}_1^{MC}$, then the process will progress into $\mathcal{Y}^{DE}$ and subsequently hit a state in $\mathcal{Y}^{DE}$, with $I = \widehat{I}$, in finite time. When this occurs, there are two possibilities:

1. $S \leq \eta N$, in which case there is a high probability that the process is forced straight back into $\mathcal{Y}^{DE}$.

2. $S > \eta N$, in which case the process hits the set $\mathcal{T}_2$ and CTMC dynamics resume.

Thus, we allow the hybrid diffusion process to switch from CTMC dynamics to diffusion dynamics upon hitting any state in $\mathcal{T}_1^{MC}$, and to switch from diffusion dynamics to CTMC dynamics upon hitting a state in $\mathcal{T}_2$.

We now discuss the hybrid diffusion in more detail, in the interest of computing the distribution of the final size of the outbreak.

3.2.2 Final outbreak size

Since the diffusion dynamics of the hybrid diffusion model are time-homogeneous (Definition 2), we again appeal to its embedded jump process (Definition 5). Let $(\mathbf{Z}_n, n \geq 0)$, denote the embedded jump process of the hybrid diffusion process, which takes values in $\mathcal{Y}^{MC}$. For all $\boldsymbol{x}$ in $\mathcal{D}$ and $\mathcal{T}_2^{MC}$, the only non-zero transition probabilities of the jump process, denoted $p_{\boldsymbol{x}\,\boldsymbol{y}}^{\boldsymbol{Z}}$, are $p_{\boldsymbol{x}\,\boldsymbol{y}}^{\boldsymbol{Y}}$, if $\boldsymbol{y}$ is in $\mathcal{Y}^{MC}$. To compute the transition probabilities from each state in $\mathcal{T}_1^{MC}$ to all states in $\mathcal{T}^{MC}$, we consider the hitting distribution of the diffusion approximation on the set of states with $\widehat{I}$ infectious individuals (Theorem 5), given an initial state in $\mathcal{T}_1^{MC}$.

It follows from Theorem 5, that for all $\boldsymbol{x}$ in $\mathcal{T}_1^{MC}$, the *next* hitting distribution of $\mathbf{Z}\,(t)$ on the set of states with $\widehat{I}$ infectious individuals follows a normal distribution, with respect to S, with mean $S(\boldsymbol{x})$ and variance

$$\sigma_{1,1}(t(\boldsymbol{x}))\,N + \frac{S(\boldsymbol{x})\,N}{S(\boldsymbol{x}) - \eta}\left(2\,\sigma_{1,2}(t(\boldsymbol{x})) + \frac{\sigma_{2,2}(t(\boldsymbol{x}))}{S(\boldsymbol{x}) - \eta}\right), \tag{3.5}$$

where $\Sigma(t) = [\sigma_{i,j}(t)]$ is governed by the system of ordinary differential equations (2.12), under transformation (2.20), and $t(\boldsymbol{y})$ is given by equation (2.31). Let $\Psi\,(s|\boldsymbol{x})$ denote the cumulative density function of this hitting distribution, given that $\mathbf{Z}\,(t)$ switched from CTMC dynamics to diffusion dynamics through the state $\boldsymbol{x}$ in $\mathcal{T}_1^{MC}$. Then, for all $\boldsymbol{x} = (S, \widehat{I})$ in $\mathcal{T}_1^{MC}$ and $\boldsymbol{y} = (S', \widehat{I})$ in $\mathcal{T}^{MC}$, the only non-zero jump probabilities are

$$p^{\boldsymbol{Z}}_{\boldsymbol{x}\,\boldsymbol{y}} = \begin{cases} \Psi\left(S' + \frac{1}{2}\,|\,\boldsymbol{x}\right) - \Psi\left(S' - \frac{1}{2}\,|\,\boldsymbol{x}\right) & \text{if } 0 \le S' \le S - 2, \\ 1 - \Psi\left(S - \frac{1}{2}\,|\,\boldsymbol{x}\right) & \text{if } S' = S - 1. \end{cases} \tag{3.6}$$

Note that the diffusion dynamics can hit any state $\boldsymbol{y}$ with $\widehat{I}$ infectious individuals, but if this state is in $\mathcal{T}_1^{MC}$ then this may be considered a rare event. Thus, if the hybrid diffusion process returns to CTMC dynamics via the state $\boldsymbol{y}$ in $\mathcal{T}_1^{MC}$, the process switches back to CTMC dynamics (and has an instantaneous removal event) with probability $p^{\boldsymbol{Z}}_{\boldsymbol{y}\,\boldsymbol{y}+\boldsymbol{\ell}_2}$, or re-starts diffusion dynamics, with the initial state $\boldsymbol{y}$, with probability $1 - p^{\boldsymbol{Z}}_{\boldsymbol{y}\,\boldsymbol{y}+\boldsymbol{\ell}_2}$.

We are now able to write down the hitting probabilities of the hybrid diffusion process. Fix $\boldsymbol{x}$ in $\mathcal{B}$, and let $h^{\boldsymbol{Z}}_{\boldsymbol{x}\,\boldsymbol{y}}$ denote the hitting probability for any state $\boldsymbol{y}$ in $\mathcal{Y}^{MC}$, given the initial state $\boldsymbol{x}$. Then it follows from Definition 6 that, for all $\boldsymbol{y}$ in $\mathcal{Y}^{MC}$, the hitting probabilities $h^{\boldsymbol{Z}}_{\boldsymbol{x}\,\boldsymbol{y}}$ are the minimal non-negative solution to the system of linear equations

$$h^{\boldsymbol{Z}}_{\boldsymbol{x}\,\boldsymbol{y}} = \sum_{\boldsymbol{z} \in \mathcal{Y}^{MC}} p^{\boldsymbol{Z}}_{\boldsymbol{x}\,\boldsymbol{z}}\,h^{\boldsymbol{Z}}_{\boldsymbol{z}\,\boldsymbol{y}}, \tag{3.7}$$

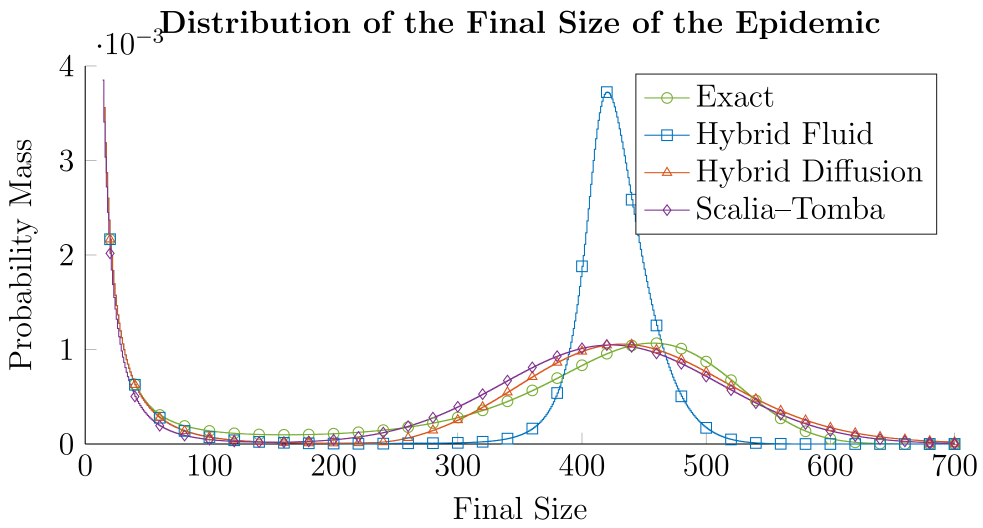


Figure 3.4: The distribution of the final size of the epidemic calculated from the SIR CTMC, hybrid fluid model, hybrid diffusion model, and Scalia–Tomba for $R_0 = 1.3$ and $N = 1,000$ with one initially infectious individual.

with $h^{\boldsymbol{Z}}_{\boldsymbol{z}\,\boldsymbol{z}} = 1$. The distribution of the final size of the outbreak, given the initial state $\boldsymbol{x} = (N-1, 1)$, is the $(N+1) \times 1$ vector with entries $h^{\boldsymbol{Z}}_{\boldsymbol{x}\,\boldsymbol{y}}$, for all $\boldsymbol{y}$ in $\mathcal{A}$, which is computed using Algorithm 5.

We now compare the distribution of the final size of the outbreak from the hybrid diffusion model to the hybrid fluid model and the hybrid model of Scalia-Tomba (Section 2.3.2)

3.2.3 Numerical results

Figure 3.4 shows the distribution of the final size of the outbreak from the hybrid diffusion model (red with triangles) and Scalia-Tomba's hybrid model (Theorem 10) (purple with diamonds). The hybrid diffusion model and Scalia-Tomba's model approximate the sub-critical component of the final size accurately but neither model succeeds in fully describing the non-normality exhibited by the distribution of the final size of the established outbreak.

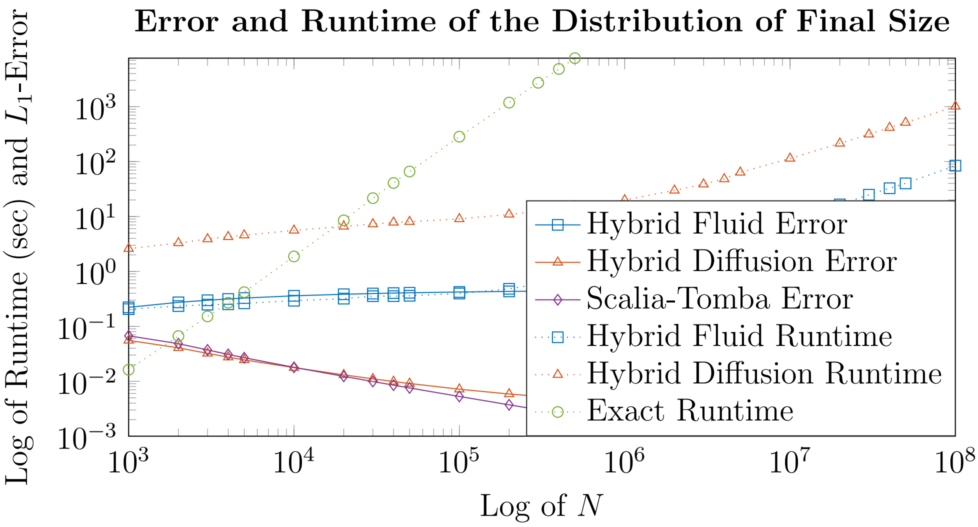


Figure 3.5: The required runtime of the SIR CTMC, hybrid fluid model and hybrid diffusion model alongside the L_1-error of the hybrid fluid model, hybrid diffusion model, and Scalia-Tomba's model. The error in the hybrid fluid model and the hybrid diffusion model is at-best a constant of $\mathcal{O}\left(10^0\right)$ and $\mathcal{O}\left(10^{-3}\right)$, respectively. The asymptotic slope of the runtime of the hybrid models (Algorithm 5) suggests that they are of computational complexity $\mathcal{O}(N)$ compared to the $\mathcal{O}(N^2)$ of the SIR CTMC (Algorithm 2). Here we used $R_0 = 1.3$ and $\widehat{I} = 17$ with the initial state $(N-1, 1)$.

Figure 3.5 shows the runtime of the hybrid diffusion model (dotted ochre with triangles). The asymptotic slope of the runtime line is approximately one, which indicates that the asymptotic runtime of Algorithm 5 for the hybrid diffusion model is $\mathcal{O}(N)$. The time difference between the runtime of the hybrid fluid model and the hybrid diffusion model corresponds to the time difference in calculating the hitting distributions of equations (3.1) and (3.6). Irrespective of N, Scalia-Tomba's approximation is effectively instantaneous to compute so its runtime has not been included in Figure 3.5.

Figure 3.5 also shows the L_1-error of the hybrid diffusion model (solid ochre with triangles) and Scalia-Tomba's model (solid purple with triangles). As N increases, the L_1-error of the hybrid diffusion approximation decreases achieving a minimum of a constant of $\mathcal{O}\left(10^{-2}\right)$, thereby showing a significant improvement over the accuracy of the hybrid fluid model. Although the L_1-error can generally be decreased by increasing the threshold, the hybrid diffusion model does not achieve a significant improvement over Scalia-Tomba's approximation unless the probability that $\mathbf{Z}\left(t\right)$ hits a state in $\mathcal{Y}^{DE}$ is negligible.

3.3 Implementation

We now consider numerical implementation for computing the distribution of the duration of the outbreak and the distribution of the final size of the outbreak from both of the hybrid models (sections 3.1 and 3.2). In addition, we present our approach to computing a suitable value for the threshold $\widehat{I}$.

3.3.1 Computing distributions

Consider computing the distribution of the final size of the outbreak using the hybrid fluid model (equation (3.4)) and the hybrid diffusion model (equation (3.7)). These systems of equations differ only in their treatment of the jump probabilities from states in $\mathcal{T}_1^{MC}$ to states in $\mathcal{T}_2^{MC}$. Thus, both systems of equations may be solved via an algorithm which is the same for all states in $\mathcal{D}$, but deals with the jump probabilities for states in $\mathcal{T}^{MC}$ differently. Computing the distribution of the duration of the outbreak via Implicit Euler integration may be achieved via a similar algorithm, because the structure of the resulting system of equations is similar to the structure of the system of equations (3.4). In this section, we present an algorithm suitable for computing the distribution of the size of the outbreak, and the distribution of the duration of the outbreak from both the hybrid fluid model and the hybrid diffusion model.

Recall that Jenkinson and Goutsias [2012] and Black and Ross [2015] presented highly-efficient routines for computing the distribution of the duration of the outbreak and the distribution of the final size of the outbreak (Sections 2.3.1 and 2.3.2). These approaches rely on transforming the SIR CTMC to its DA representation (equation (2.20)), which is more amenable to numerical analysis. So far we have discussed the hybrid models in the population framework because it is a more intuitive format. However, we now convert them to their DA representation.

Let the sets $\mathcal{N}$, $\mathcal{N}^{MC}$, $\mathcal{N}_1^T$ and $\mathcal{N}_2^T$ denote the DA representations of the population sets $\mathcal{Y}$, $\mathcal{Y}^{MC}$, $\mathcal{T}_1^{MC}$ and $\mathcal{T}_2^{MC}$, respectively (Transformation (2.20)). The states in $\mathcal{N}^{MC}$ are ordered by equation (2.22), and indexed by k, for all $k = 1, 2, \ldots, |\mathcal{N}^{MC}|$, such that $\boldsymbol{n}_1, \boldsymbol{n}_2, \ldots, \boldsymbol{n}_{|\mathcal{N}^{MC}|}$ are ordered appropriately. In addition, recall that $\delta k_1 = N - N_I + N_R$ and $\delta k_2 = N + 2 - N_I + N_R$, and φ_k denotes the kth element of the $|\mathcal{N}^{MC}| \times 1$ vector $\boldsymbol{\varphi}$. The following algorithm exploits the structure of the SIR CTMC in a similar way to Black and Ross [2015] (Algorithm 2), and accounts for transitions between $\mathcal{N}_1^T$ and $\mathcal{N}_2^T$ using the fact that the change in index is given by the change in the number of susceptible individuals.

We found that a suitable approach to reducing the computational over-head of Algorithm 5 is to only consider states in $\mathcal{N}^T$ with a significant probability. This is performed on line 12 of the algorithm where we require that the probability associated with the state is above a tolerance ϵ. We found a suitable tolerance to be $\epsilon = 1 \times 10^{-7}$, which results in a small accumulation of error and generally results in a significant decrease in computational over-head. This choice is robust to most reasonable values of R_0 but may result in very little reduction of computational over-head if R_0 is close to one.

For computing the distribution of the final size of the outbreak from the hybrid fluid model and the hybrid diffusion model, one must consider computing the solution to (3.4) and (3.7), respectively. With reference to Algorithm 5, let $\boldsymbol{\varphi}$ denote the $|\mathcal{N}^{MC}| \times 1$ vector whose kth element is the hitting probability of the kth state, for $k = 1, 2, \dots, |\mathcal{N}^{MC}|$, given the initial state $\boldsymbol{n}_1 = (1, 0)$. In addition, let $f(k, k')$ denote the transition probability from the kth state to the k'th state, for $k, k' = 1, 2, \dots, |\mathcal{N}^{MC}|$ and $f(k, k) = 0$. Then, if $\boldsymbol{\varphi}$ is initialised as $\boldsymbol{e}_1$, the distribution of the final size of the outbreak is calculated by iteratively updating the entries of $\boldsymbol{\varphi}$ via Algorithm 5, until the algorithm terminates.

For computing the distribution of the duration of the outbreak from the hybrid fluid model, one must consider computing the solution to (3.2)—(3.3) over a grid of time points [Jenkinson and Goutsias, 2012]. With reference to Algorithm 5, let $\boldsymbol{\varphi}$ denote the $|\mathcal{N}^{MC}| \times 1$ vector whose kth element is the

Algorithm 5: Algorithm for calculating the distribution of the duration of the outbreak, and the distribution of the final size of the outbreak from the hybrid fluid model and the hybrid diffusion model.

```
Data: Set $\boldsymbol{\varphi}$, and tolerance $\epsilon$.
Initialise the state-index $k = 2N + 1$ ;
for $N_R = 0, \ldots, N$ do
    Store the initial index $k_0 = k$ and normalise the current entry
      $\varphi_k = \varphi_k/(1 + f(k,k))$ ;
    for $N_I = N_R + 1, \ldots, \min\{N_R + \widehat{I} - 1, N - 1\}$ do
        Update the distribution via:
        $\varphi_{k+\delta k_1} = \varphi_{k+\delta k_1} + \varphi_k \, f(k, k + \delta k_1)$ (Infection event) ;
        $\varphi_{k-\delta k_2} = \varphi_{k-\delta k_2} + \varphi_k \, f(k, k - \delta k_2)$ (Removal event) ;
        Update the state-index $k = k + \delta k_1$ ;
    end
    if $N_R < N - \widehat{I} - \lfloor \eta N \rfloor$ then
        for $j = 1, \ldots, N - N_I$ do
            if $\varphi_k > \epsilon$ then
                If computing the final size distribution:
                $\varphi_{k-j} = \varphi_{k-j} + \varphi_k \, f(k, k - j)$ ;
                If computing the distribution of duration:
                Store $\varphi_k$, return delayed flux $\varphi_k^{delayed}$ to system ;
                $\varphi_{k-j} = \varphi_{k-j} + \varphi_k^{delayed} \, f(k, k - j)$ ;
            end
        end
    else if $N_R < N$ then
        $\varphi_{k-\delta k_2} = \varphi_{k-\delta k_2} + \varphi_k \, f(k, k - \delta k_2)$. ;
    Reset the state index $k = k_0 - 1$ ;
end
```

Implicit Euler approximation of the transition probability from the state $\boldsymbol{n}_1$ to the kth state, for $k = 1, 2, \ldots, |\mathcal{N}^{MC}|$. In addition, let $f(k, k')$, for $k \neq k'$, denote the transition rate from the kth state to the k'th state, multiplied by the time step of the numerical integration, τ, for $k, k', = 1, 2, \ldots, |\mathcal{N}^{MC}|$, with $f(k, k) = \sum_{k' \neq k} f(k, k')$. Then, if $\boldsymbol{\varphi}$ is initialised as the distribution of $\boldsymbol{N}\,(t)$, the distribution of $\boldsymbol{N}\,(t + \tau)$ is calculated by iteratively updating the entries of $\boldsymbol{\varphi}$ via Algorithm 5, until the algorithm terminates.

In calculating the distribution of the final size of the outbreak from the hybrid diffusion model, we reduce the computational over-head of Algorithm 5 by only calculating the mean and variance of the hitting distribution (3.6) for a subset of states in $\mathcal{T}_1^{MC}$, and then extrapolating to all the other states in $\mathcal{T}_1^{MC}$ using linear interpolation. More specifically, let $\theta(\boldsymbol{x}) = (S(\boldsymbol{x}), \sigma_{1,1}(t(\boldsymbol{x})), \sigma_{1,2}(t(\boldsymbol{x})), \sigma_{2,2}(t(\boldsymbol{x})))$ for $\boldsymbol{x}$ in $\mathcal{T}_1^{MC}$, and $\mathcal{T}^* = \{(S, \widehat{I}) \in \mathcal{T}_1^{MC} : S = S_0, S_0 + k, S_0 + 2k, \ldots, N - \widehat{I}\}$ where $S_0 = \lfloor \eta N \rfloor$ and k is a positive integer. Then we evaluate $\theta(\boldsymbol{x})$ for every $\boldsymbol{x}$ in $\mathcal{T}^*$ and use the output to approximate $\theta(\boldsymbol{x})$ for every $\boldsymbol{x}$ in $\mathcal{T}_1^{MC} \setminus \mathcal{T}^*$ using linear interpolation. We found a robust choice for k to be 30. We found the relationship between $\theta(\boldsymbol{x})$ and $\boldsymbol{x}$, in $\mathcal{T}_1^{MC}$, to be close to linear, thus this choice of k is believed to be robust for most reasonable values of R_0.

3.3.2 Computing a threshold

Our approach for computing a suitable threshold is based on the distribution of the maximum of the branching process approximation of the SIR CTMC, conditioned on extinction. We utilise the branching process approximation because it is sufficiently accurate and provides an expression which can be computed effectively instantaneously [Ball and Donnelly, 1995]. Based on the notion that the only sample paths of the SIR CTMC which do not hit the threshold, should be the sample paths in which the outbreak fades out. We compute the threshold by finding a value of I, for which the probability that the branching process, conditioned on fading out, exceeds I is sufficiently small.

Let $U(t)$, for $t \geq 0$, denote the branching process approximation of the population of infectious individuals from the SIR CTMC (conditioned on fading out), which takes values $0, 1, \ldots$. In addition, let the random variable $M = \sup_{0 \leq t \leq \infty} U(t)$ denote the largest value obtained by $U(t)$, for all $t \geq 0$. Then $\widehat{I}$ is defined as the minimum m, for $m = 0, 1, 2, \ldots,$ which satisfies $\Pr(M \geq m) \leq \epsilon$. More precisely, the threshold $\widehat{I}$ is the minimum m which satisfies the inequality (Section 5 of [Ball and Donnelly, 1995])

$$m \geq U(0) + \frac{\log\left(R_0^{U(0)} + \epsilon - 1\right) - \log(\epsilon)}{\log(R_0)}. \tag{3.8}$$

Inequality (3.8) is based on the assumption that $R_0 > 1$; in the event that $R_0 < 1$, R_0 is replaced by $1/R_0 = \eta$ in inequality (3.8) in order to consider the branching process conditioned on fading out. However, inequality (3.8) can not be used if $R_0 = 1$. Note that, choosing a smaller ϵ leads to a larger choice of $\widehat{I}$ and hence, generally, more accurate results but larger computational runtimes. We determined that 5×10^{-3} is a suitable value for ϵ due to the following observation.

For the distribution of the final size (duration) of the outbreak, the ochre (green) curve with triangles (circles) in Figure 3.6 shows the empirical minimum threshold required to achieve at most 0.1 (0.25) L_1-error from the hybrid diffusion (fluid) model, with a fixed $N = 10,000$. We chose these values because they correspond to the worst-case scenarios of Scalia-Tomba [1985] and Barbour [1975]. Taking ϵ to be 5×10^{-3} in equation (3.8) produces the blue curve with squares which ensures a higher threshold than the ochre

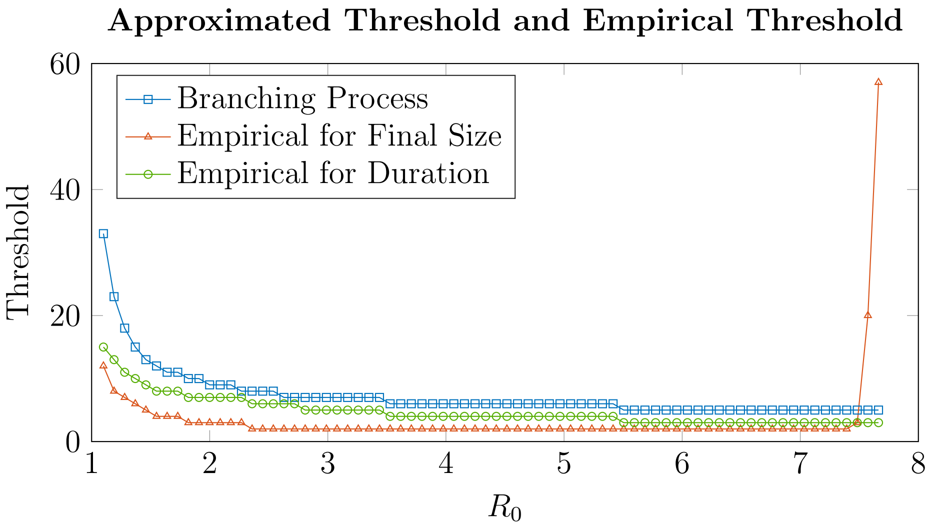


Figure 3.6: For the distribution of the final size (duration) of the outbreak, the ochre (green) curve with triangles (circles) shows the minimum threshold $\widehat{I}$ which achieves an L_1-error of 0.1 (0.25). The blue curve with squares shows the threshold determined by inequality (3.8) using $\epsilon = 5 \times 10^{-3}$ which achieves at most 0.1 (0.25) L_1-error in the distribution of the final size (duration) of the outbreak, provided R_0 is less than approximately 7.5. Here we used $N = 10,000$ and the initial state $(N-1, 1)$.

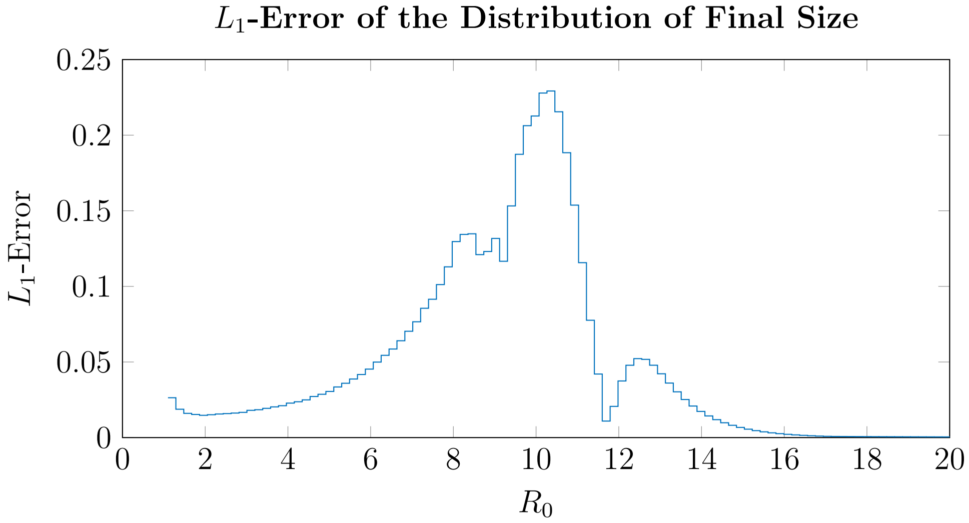


Figure 3.7: The L_1-error of the distribution of the final size of the epidemic using inequality (3.8) to calculate the threshold. The error exceeds 0.1 on an interval of R_0 from approximately 7.5 to 11 and is at most 0.24. This issue arises when the fluid approximation of S falls below approximately eight susceptible individuals. Here we used $N = 10,000$ and the initial state $(N-1, 1)$.

and green curves and hence ensures that the L_1-error in the distribution of the final size (duration) of the outbreak is at most 0.1 (0.25). However, this guarantee does not hold for $R_0 > 7.5$, which we discuss in the next paragraph. As N increases, the minimum threshold required to achieve at most 0.1 (0.25) L_1-error in the distribution of the final size (duration) of the epidemic decreases and the threshold determined by inequality (3.8) stays the same. In addition, the point at which inequality (3.8) fails to produce a reliable threshold for the distribution of the final size of the epidemic increases.

Figure 3.7 shows that the threshold determined by inequality (3.8) provides an L_1-error for the distribution of the final size of the outbreak which is at most 0.24. The divergence of the approximate distribution from the exact distribution manifests as an inaccurate approximation of the probability that the final size of the outbreak is N, $N-1$ or $N-2$. This divergence occurs when the diffusion approximation comes close to the absorbing boundary with $S=0$ because the SIR CTMC is able to be absorbed by this set but the diffusion approximation is not. Figure 3.7 shows that the L_1-error decreases for $R_0 \geq 13$ because the probability that the final size of the outbreak is equal to $N-1$ or $N-2$ becomes negligible as R_0 becomes very large. The loss of the ability of inequality (3.8) to provide a reliable threshold is characterised as the region of R_0 for which the mean number of susceptible individuals at the end of the fluid dynamics of $\mathbf{Z}(t)$ is less than approximately eight, but more than one.

3.4 Discussion

In this chapter we introduced two hybrid Markov chain models for approximating the distribution of the duration of the outbreak and the distribution of the final size of the outbreak for the SIR CTMC. These models are novel in the sense that no other hybrid models of the SIR CTMC have CTMC dynamics during their initial and final stages. As a result, these models preserve the important stochastic features of the SIR CTMC which occur during these phases of the outbreak. Namely, the probability that the outbreak fades out, and the variability in the amount of time before the outbreak becomes established. In the case of the SIR CTMC, we used these hybrid models to derive expressions for the distribution of the duration of the outbreak, and the distribution of the final size of the outbreak. Both of these distributions can be computed numerically in $\mathcal{O}(N)$ time, as opposed to the $\mathcal{O}(N^2)$ time of the SIR CTMC. This has enabled us to calculate the distribution of the duration of the outbreak, and the final size of the outbreak for populations of at least

10^6, within a matter of hours. Our approximations of the distribution of the duration of the outbreak, and the distribution of the final size of the outbreak achieve a similar level of accuracy to the existing hybrid approximations, and as we shall see in Chapter 4 and Chapter 5, our methodology has the additional advantage that it may be easily generalised to other situations or more complex models.

The hybrid models presented here were observed to provide inaccurate approximations of the distribution of the final size of the outbreak for a particular region of R_0. This is because the susceptible component of the mean trajectory of the diffusion approximation comes close to the $S = 0$ absorbing boundary of the Markov chain, thereby causing the diffusion approximation to break down. This motivates future research that might utilise a similar hybrid model to Safta et al. [2015], which includes an additional threshold on the number of susceptible individuals.

The methodology presented here demonstrates that the hybrid diffusion model provides an accurate representation of the initial stages of the SIR CTMC, and its state-variability once the outbreak has become established. Thus, in the next chapter we investigate utilising the hybrid diffusion model for conducting inference during the initial stages of an outbreak.

Chapter 4

Early estimation of the basic reproductive number for SIR disease dynamics

Accurately modelling the early stages of an emerging outbreak is of vital importance for inferring the basic reproductive number R_0 [Viboud et al., 2016, Bettencourt and Ribeiro, 2008, Glass et al., 2011, Nishiura et al., 2010, Vega et al., 2013, White and Pagano, 2007]. An accurate and reliable estimate of R_0 is crucial because it characterises the transmission potential of the disease, an important factor for public health authorities in planning their response to the outbreak [Simonsen et al., 1997, Meltzer et al., 1999, Lemon et al., 2007, Chowell et al., 2009, Wu et al., 2006]. However, early estimates of R_0 are generally positively-biased, due to incomplete or inaccurate case reporting [Cauchemez et al., 2006, Glass et al., 2007, Woolhouse et al., 2015], population heterogeneity (such as spatial variation, age-specific or household clustering of contacts) [Galvani, 2016, Lipsitch et al., 2015, Favier et al., 2005, Keeling et al., 2001], and incorrectly accounting for imported infectious cases [Roberts and Nishiura, 2011]. Another source of bias which is often over-looked is the probability of initial fade out. During the initial stages of an outbreak, the probability of initial fade out decreases considerably each time the number of infectious individuals increases. Thus, from a modelling perspective, the event that an individual outbreak becomes established could be considered unlikely. At the same time, an outbreak will often not be detected by public health authorities until such a time that it has established an appreciable chain of transmission, thereby effectively avoiding initial fade out [Hartfield and Alizon, 2013]. It follows that the event that an outbreak becomes established, and is consequently detected by public health authorities, is one which needs to be accounted for in estimating the basic reproductive number during the early stages of an outbreak.

Cases of the disease which occurred before the outbreak was detected are generally ascertained by a case follow-up program led by public health authorities [Smith, 2006]. The general approach to using this data to estimate the basic reproductive number involves computing the probability of each of the observed incidence counts, conditioned on all the observed incidence counts which came beforehand [Bettencourt and Ribeiro, 2008, White and Pagano, 2007, Black and Ross, 2013, Boys and Giles, 2007, Chowell et al., 2007]. The problem with this approach is that the probability of each of the observed incidence counts should be conditioned on the event that the outbreak will become established. Mercer et al. [2011] demonstrated that not accounting appropriately for the probability of initial fade out biases estimates of R_0 and that this bias decreases as the time since the first observation increases, thereby exhibiting correlation between the two. An appropriate way of accounting for the event that the outbreak is detected by public health authorities is to condition the underlying model on the event that the outbreak becomes established [Mercer et al., 2011, Rida, 1991].

In this chapter, we present a conditioned susceptible-infectious-removed (SIR) continuous-time Markov chain (CTMC) which partially accounts for the probability of initial fade out. This is achieved by conditioning the SIR CTMC on the event that the outbreak eventually becomes established by modifying its transition rates according to Theorem 2. We argue that it is reasonable to consider an established outbreak to be one where the cumulative number of cases eventually exceeds a predetermined threshold. Under this construction, we demonstrate that conditioning the SIR CTMC on the event that the outbreak eventually exceeds 50 cases reduces the resulting over-estimate of R_0 by around 0.3, on average.

Fundamental to inferring the value of R_0 is calculating the likelihood of the data [Sprott, 2000] (Equation (2.38)). Exact methods for computing the likelihood are typically computationally infeasible, even for moderate population sizes. Thus, it is common to consider approximating the likelihood [Cooper and Lipsitch, 2004]. The diffusion approximation (Theorem 4) is effective [Ross et al., 2006, 2009, Ross, 2012], but it fails to accurately represent the initial stages of the outbreak. It follows that a natural approach is to approximate the likelihood using a hybrid diffusion model similar to the one presented in the previous chapter. The diffusion hybrid considered here differs from the diffusion hybrid of the previous chapter only in the mechanism by which it switches from CTMC to diffusion dynamics. Based on the results of the previous chapter, we expect the diffusion hybrid to be appropriately accurate and to provide an advantage in computational efficiency.

We demonstrate the utility of our methodology by applying it to an outbreak of pandemic influenza from 2009, which occurred in Western Australia (A(H1N1)pdm09) [Kelly et al., 2010, Pedroni et al., 2010]. During this outbreak, a thorough case ascertainment and follow-up program was conducted during the first three weeks of the outbreak until such a time that the outbreak was deemed widespread, by which stage 102 cases had been confirmed. Using the simple SIR CTMC, we demonstrate that estimates of R_0 which account for this fact are more accurate during the early stages of the outbreak.

The present chapter has two objectives. The first is to present an approach for reducing bias in early estimates of R_0 from daily incidence data. The second is to present a hybrid diffusion model, similar to the hybrid diffusion model from the previous chapter, for accurately and efficiently estimating the likelihood of the data. These concepts are straightforward to implement and can be generalised to more complex epidemiological models, as we shall see in the following chapter.

4.1 Conditioned model

Recall that in Section 2.3.3 we presented the likelihood of the SIR CTM-C (2.38). The key problem with an approach of this nature is that in computing the probability of observing y_k infection events at time t_k, the probability is only conditioned on the event $\mathcal{Y}_{k\text{-}1}$, when it should also be conditioned on the event that the outbreak becomes established. In this section we condition the outbreak on becoming established, meaning that we impose the constraint that the outbreak has not faded out prior to the current time and does not fade out before becoming established. The process by which this is achieved may be thought of as a way of restricting all the possible trajectories of the process to those in which initial fade out does not occur and is made precise by Theorem 2 due to Waugh [1958].

Recall that $(\boldsymbol{N}(t), t \geq 0)$ is the DA representation of the SIR CTMC. The DA process takes values in $\mathcal{N}$ and, for all $\boldsymbol{n}$ in $\mathcal{N}$, has the positive transition rates $q^{\boldsymbol{N}}_{\boldsymbol{n}\,\boldsymbol{n}+\boldsymbol{e}_i}$, if $\boldsymbol{n}+\boldsymbol{e}_i$ is in $\mathcal{N}$ and $i = 1, 2$ (see Section 2.3). We wish to condition the DA process on the event that it hits a state in $\mathcal{N}^T$, a subset of $\mathcal{N}$, such that once it hits a state in $\mathcal{N}^T$ it may be considered an established outbreak. For the remainder of this section, we make the important distinction that the DA process discussed until now is the unconditioned DA process.

Following Theorem 2, the conditioned DA process is a CTMC taking values in $\mathcal{N}$. For all $\boldsymbol{n}$ in $\mathcal{N}$, let $u_{\boldsymbol{n}}$ denote the probability that the unconditioned DA process ever hits a state in $\mathcal{N}^T$, starting from the state $\boldsymbol{n}$ (Definition 6). Then for all $\boldsymbol{n}$ and $\boldsymbol{m}$ in $\mathcal{N}$, with $\boldsymbol{m} \neq \boldsymbol{n}$, the conditioned DA process has the transition rates

$$\tilde{q}^{\boldsymbol{N}}_{\boldsymbol{n}\,\boldsymbol{m}} = \begin{cases} (u_{\boldsymbol{m}}/u_{\boldsymbol{n}})\, q^{\boldsymbol{N}}_{\boldsymbol{n}\,\boldsymbol{m}} & \text{if } \boldsymbol{n} \notin \mathcal{N}^T, \\ q^{\boldsymbol{N}}_{\boldsymbol{n}\,\boldsymbol{m}} & \text{otherwise,} \end{cases} \tag{4.1}$$

with the condition that $\tilde{q}^{\boldsymbol{N}}_{\boldsymbol{n}\,\boldsymbol{n}} = -\sum_{\boldsymbol{m}\neq\boldsymbol{n}} \tilde{q}^{\boldsymbol{N}}_{\boldsymbol{n}\,\boldsymbol{m}}$.

Following Theorem 2, the set $\mathcal{N}^T$ must be a subset of $\mathcal{N}$ for which there is a non-zero probability of reaching $\mathcal{N}^T$ from any non-absorbing state of $\mathcal{N}$. A logical choice is to set $\mathcal{N}^T$ as the set of all states in $\mathcal{N}$, for which $N_I > n_T$, for some $n_T \in \{0, 1, \ldots, N\}$, where n_T is referred to as the threshold number of infection events. We have great freedom in specifying the threshold n_T *a priori*. A sensible choice is to set n_T to be large enough that once the outbreak reaches $\mathcal{N}^T$ there is a high probability that it is established. In modelling data from a real outbreak, a sensible choice is to set n_T to the number of infection events which had occurred by the time at which a particular outbreak was detected.

The conditioned DA process and the unconditioned DA process differ only in their transition rates. Thus, the methodology for utilising the conditioned

DA process for inference is identical to the methodology for the unconditioned DA process (Section 2.3.3). In particular, recall that $y_1, y_2, \ldots, y_n$ denotes a sequence of observed cumulative incidence counts made at times $t_1, t_2, \ldots, t_n$ and that $\Pr(N_I(t_k) = y_k | \mathcal{Y}_{k\text{-}1}) = L_E^k(\boldsymbol{\theta})$ is the probability of the observed data under the unconditioned DA process. Then, if $L_C^k(\boldsymbol{\theta})$, for all $k = 1, 2, \ldots, n$, denotes the probability of the observed data y_k under the conditioned model, the conditioned likelihood is

$$L(y|\boldsymbol{\theta}) = \prod_{k=1}^{k_T \wedge n} L_C^k(\boldsymbol{\theta}) \prod_{k=k_T+1}^{n} L_E^k(\boldsymbol{\theta}), \tag{4.2}$$

where $k_T = \min\{k | y_k > n_T\}$ and $k_T \wedge n = \min\{k_T, n\}$. The conditioned likelihood is computed via Algorithm 3. In terms of the illustrative example from Section 2.3.3, conditioning removes the dashed transitions in Figures 2.2a and 2.2b from the model and the remaining transition rates are adjusted such that the process eventually reaches the set $\mathcal{N}^T$ with probability one.

Using the DA process for inference is computationally-forbidding if the total number of observed incidences y_n is large. However, the previous chapter demonstrated that the hybrid diffusion model is an effective means of approximating the SIR CTMC which mitigates the computational cost of the SIR CTMC. Thus, we now present a hybrid diffusion model for approximating the likelihood of the conditioned DA process (4.2).

4.2 Hybrid diffusion model

The conditioned likelihood (4.2) is computed via the forward equations (Equation (2.23)) which are computationally prohibitive if the size of the underlying state space is large. Assuming that the population of infectious individuals is sufficiently large by the time t_{k_T}, it is reasonable to expect the diffusion approximation to provide an accurate approximation of the conditioned DA process thereafter. It follows that the hybrid diffusion model presented here has the dynamics of the conditioned DA process for all t in $[0, t_{k_T}]$, and the dynamics of the diffusion approximation thereafter. Recall that $L_D^k(\boldsymbol{\theta})$, for $k = 1, 2, \ldots, n$ denotes the probability of the observed data under the diffusion approximation. Then it follows that the conditioned hybrid likelihood is

$$L(y|\boldsymbol{\theta}) = \prod_{k=1}^{k_T \wedge n} L_C^k(\boldsymbol{\theta}) \prod_{k=k_T+1}^{n} L_D^k(\boldsymbol{\theta}), \tag{4.3}$$

which is computed via Algorithm 3 with the appropriate modifications made for $k > k_T$. At the time at which the model switches from CTMC dynamics to diffusion dynamics, the initial distribution of the diffusion approximation is computed from the final distribution of the conditioned DA process.

4.3 Implementation

In this section we demonstrate the accuracy and utility of our methodology by using it to estimate R_0 from daily incidence data from the first two weeks of an outbreak. Our analysis is comprised of two parts. First we demonstrate that conditioning reduces bias in estimates of R_0. Second, we demonstrate that the hybrid approximation provides an accurate and computationally-efficient means for estimating R_0 during the initial stages of an outbreak. To achieve this, we consider the four different parameter regimes displayed in Table 4.1. The values of R_0, γ and N have been selected to be representative of an influenza-like outbreak in a realistic population. The value of N also guarantees that the susceptible pool will not be depleted during the first two weeks of the outbreak. We vary R_0 between Regimes 1 and 2 to investigate the effect of the underlying value of R_0 on the estimated R_0. We vary the threshold between Regimes 1 and 3, and Regimes 2 and 4 to investigate the sensitivity of the conditioned likelihood to the threshold. To ensure our analysis is statistically robust, we consider $1,000$ independent simulated realisations of the SIR CTMC, each starting with a single infectious case, running for a duration of two weeks, and exceeding 50 infection events by the final day of the outbreak. We then illustrate the utility of our methodology by using our conditioned hybrid model to estimate R_0 from an outbreak of pandemic influenza.

Parameter	Regime 1	Regime 2	Regime 3	Regime 4
R_0	1.2	1.4	1.2	1.4
n_T	50	50	20	20
γ	1/3	1/3	1/3	1/3
N	10^7	10^7	10^7	10^7
$I(0)$	1	1	1	1

Table 4.1: Parameters used for investigating our methodology. The removal rate (γ) and basic reproductive number (R_0) are representative of influenza and the population size (N) ensures that the susceptible pool is not depleted during the first two weeks of the epidemic.

In each regime we obtain an estimate of the parameters via a frequentist framework and a Bayesian framework (Section 2.3.3). In the Bayesian framework, we obtain a point-estimate of the parameters via the commonly used *median a posteriori estimate* (MPE), which is defined as the median of the samples from the posterior. We estimate the parameters $\boldsymbol{\theta} = (1/\gamma, R_0)$, for $\boldsymbol{\theta} \in \Theta$, where Θ contains all $1/\gamma, R_0 \geq 1/10$. We use this parameterisation because R_0 is linearly related to $1/\gamma$, while it has been shown that β and γ have a more complicated inverse relationship. This means that the posterior distribution of $(1/\gamma, R_0)$ should be roughly more symmetric, than the pos-

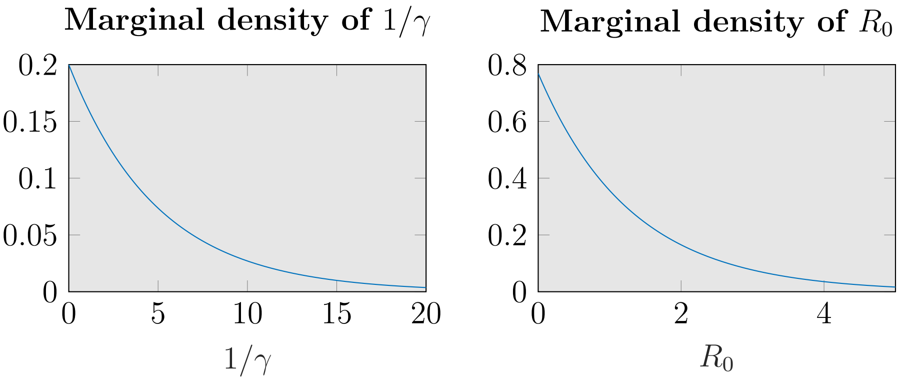


Figure 4.1: Marginal densities of the prior distribution of $1/\gamma$ and R_0.

terior distribution of (β, γ). To calculate the MPEs we use the exponential prior

$$f(1/\gamma, R_0) = \frac{1}{c_1 c_2} e^{-(1/c_1\gamma) - R_0/c_2},$$

which favours small values of $1/\gamma$ and R_0, but provides support to all $1/\gamma, R_0 > 0$. We selected $c_1 = 5$ and $c_2 = 1.3$ to provide a reasonable amount of weight to values of $1/\gamma$ and R_0 which are realistic for an influenza-like outbreak, see Figure 4.1. Our proposal density is a truncated bivariate Gaussian with support Θ and fixed covariance structure $\text{var}(1/\gamma) = 1$, $\text{var}(R_0) = 1/2$ and $\text{cov}(1/\gamma, R_0) = 0$. For each simulated data set, we run four independent Markov chain Monte Carlo chains on Θ consisting of $200,000$ iterations, and discard the initial $20,000$ iterations as burn-in.

To calculate the MLEs we maximise the log-likelihood function $\boldsymbol{\ell}(y|\boldsymbol{\theta}) = \log(L(y|\boldsymbol{\theta}))$ on Θ using MATLAB's `fmincon` function. We found that in some cases a MLE could not be identified because the optimisation routine failed to converge. These cases were characterised by realisations where the number of infection events remained low for the first week before growing rapidly in the second week. These realisations were dropped from the analysis on the basis that they did not contain enough information to provide a reliable estimate of the parameters.

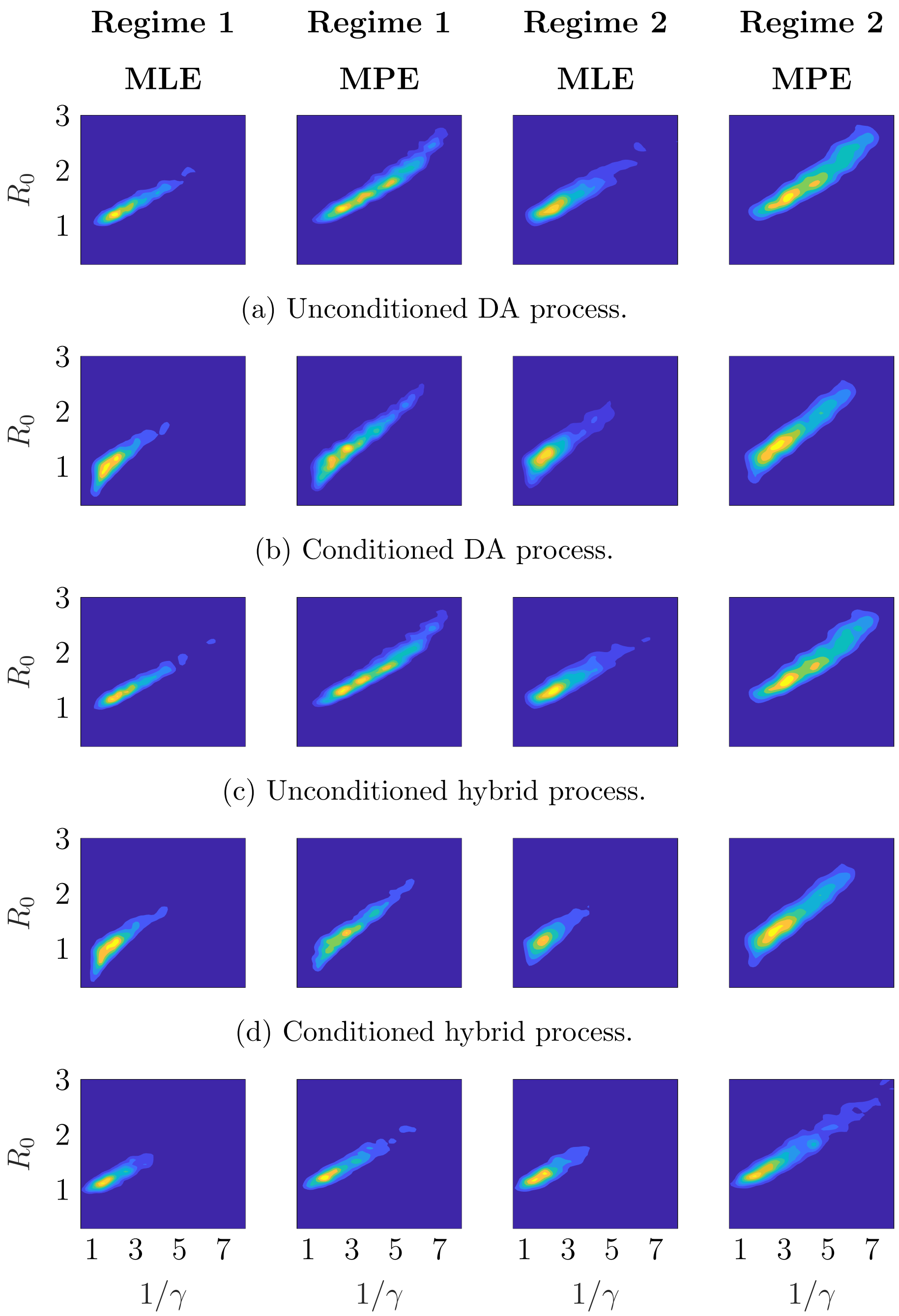
4.3.1 Validation of the conditioned model

We begin by presenting the MLEs and MPEs of R_0, across all regimes. Figure 4.2 contains density estimates of the MLEs and MPEs under Regimes 1 and 2, plotted on the $(1/\gamma, R_0)$ axes. Each row contains parameter estimates according to a different model: unconditioned/conditioned DA process, unconditioned/conditioned hybrid process, and diffusion process. Figure 4.3 contains density estimates of the MLEs and MPEs under Regimes 3 and 4 for the conditioned DA process and conditioned hybrid process. Note that the density estimates of the MPEs are clearly different to the prior distribution, suggesting that our MPEs are not overly sensitive to the choice of prior distribution, in this case.

The density estimates of $1/\gamma$ and R_0 appear unimodal with a strong correlation between $1/\gamma$ and $R_0(= \beta/\gamma)$. The distributions appear non-symmetric, with a higher density associated with estimates which have smaller values of $1/\gamma$ and R_0. Under all regimes, the distributions obtained via maximum likelihood and Bayesian inference appear similar. The unconditioned estimates appear to favour higher values of R_0 and $1/\gamma$ than their conditioned counterparts, which we now investigate in more detail.

In the following analysis we use bean plots to compare independent data sets. The bean plot is comprised of horizontal side-by-side box plots for which the whiskers represent the 2nd and 98th percentiles. The outliers are shaded according to their distance away from the median. The box plots are accompanied by the corresponding density estimates which provide a more informative view of the distribution of the data.

Figure 4.4 contains bean plots of the MLEs and MPEs of R_0 from the un-



(a) Unconditioned DA process.

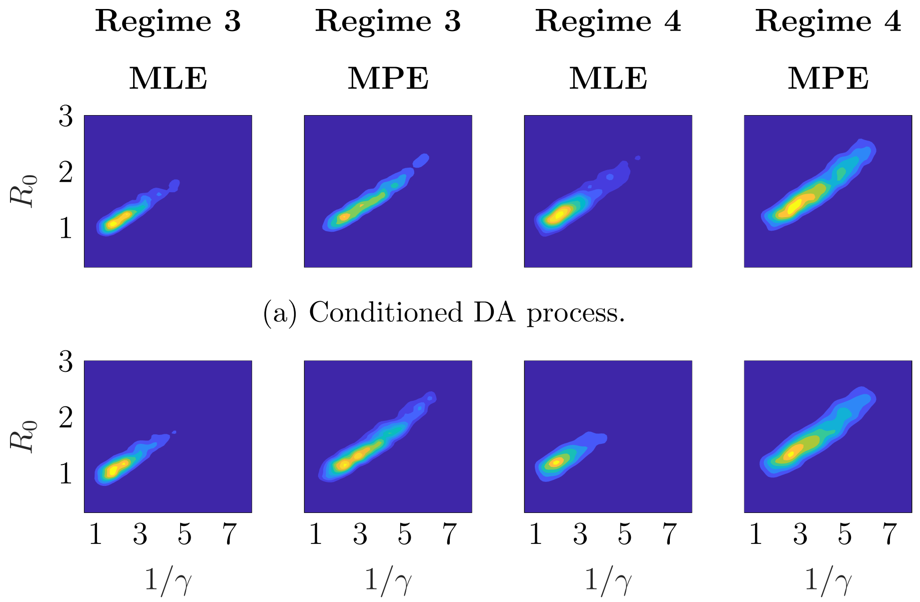
(b) Conditioned DA process.

(c) Unconditioned hybrid process.

(d) Conditioned hybrid process.

(e) Diffusion process.

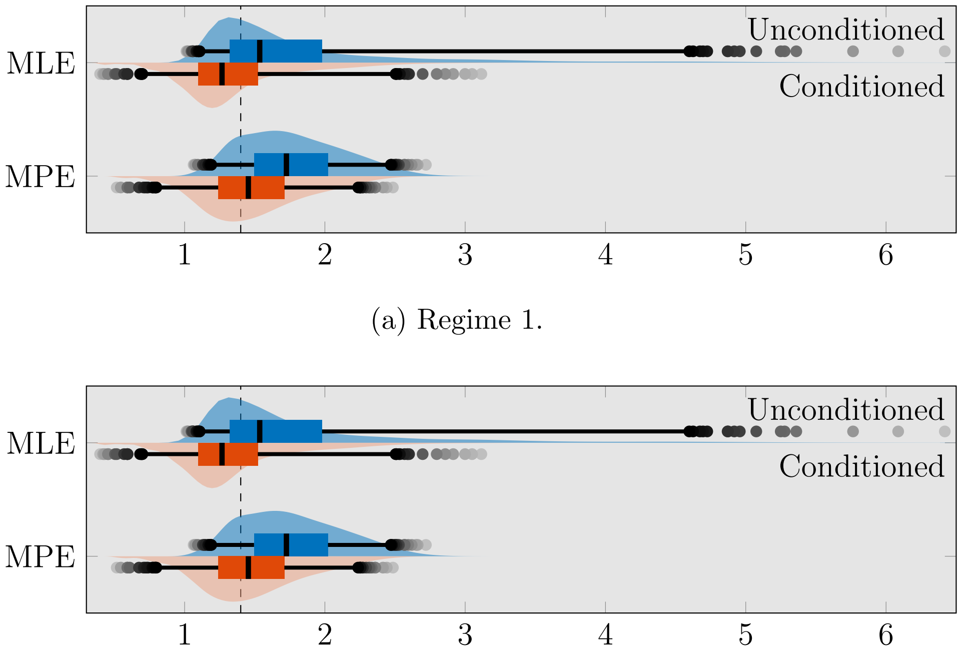
Figure 4.2: Density estimates of the MLEs and MPEs of $(1/\gamma, R_0)$ obtained under Regimes 1 and 2. The rows contain estimates from the: unconditioned/conditioned DA process, unconditioned/conditioned hybrid process, and diffusion process. The density estimates demonstrate broad agreement between estimates of R_0



(a) Conditioned DA process.

(b) Conditioned hybrid process.

Figure 4.3: Density estimates of the MLEs and MPEs of $(1/\gamma, R_0)$ obtained under Regimes 3 and 4 from the conditioned DA process and conditioned hybrid process.

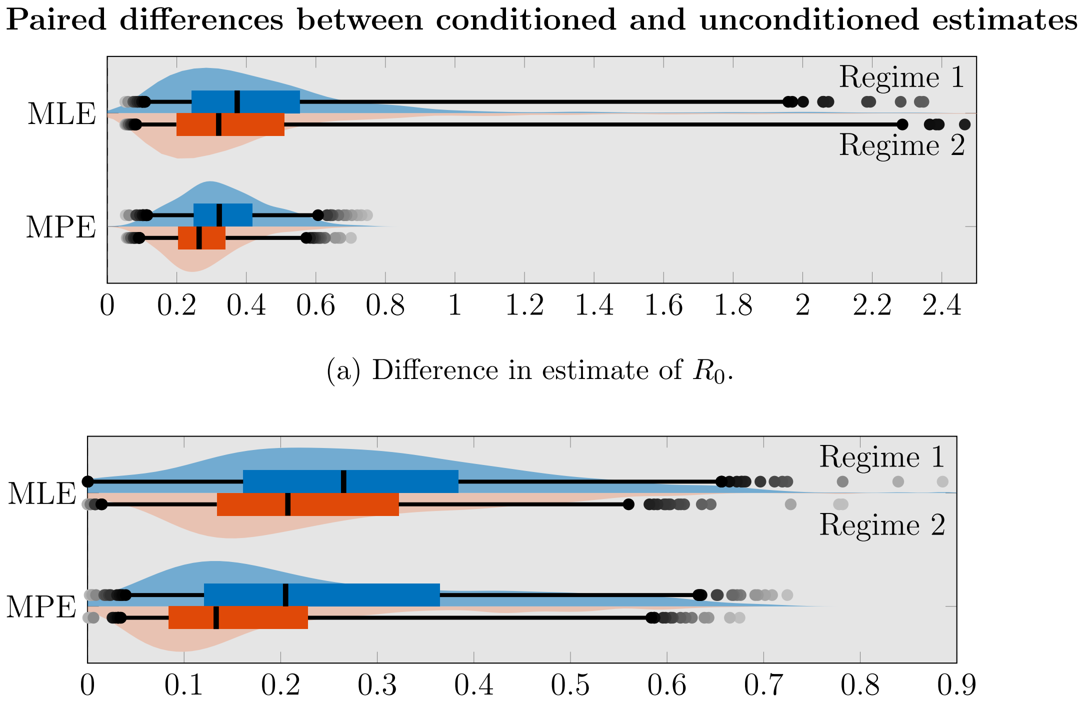


(a) Regime 1.

(b) Regime 2.

Figure 4.4: Bean plots of the estimated R_0 under Regimes 1 and 2. Bean plots are comprised of side-by-side box plots (where the whiskers represent the 2nd and 98th percentiles) plotted on top of a kernel density estimate. The conditioned estimate is smaller than the unconditioned estimate in every case. The unconditioned estimates in Regime 1 appear more biased than the unconditioned estimates in Regime 2.

conditioned DA process against the conditioned DA process, with the vertical dashed black line representing its true value. The unconditioned estimates are biased towards higher values of R_0 than the conditioned estimates and have a larger inter-quartile range (IQR). The unconditioned estimates show more bias in Regime 1 than Regime 2, presumably because the lower value of R_0 leads to a higher chance of extinction and hence conditioning has a more significant impact on the transition rates. The conditioned MPEs show less bias than the MLEs though both MLEs and MPEs have a similar IQR in each regime. The MLEs appear more susceptible to outliers. We determined the cause of these outliers to be relatively uninformative realisations which do not provide enough information to obtain a reliable estimate of the underlying values of $1/\gamma$ and R_0.



(a) Difference in estimate of R_0.

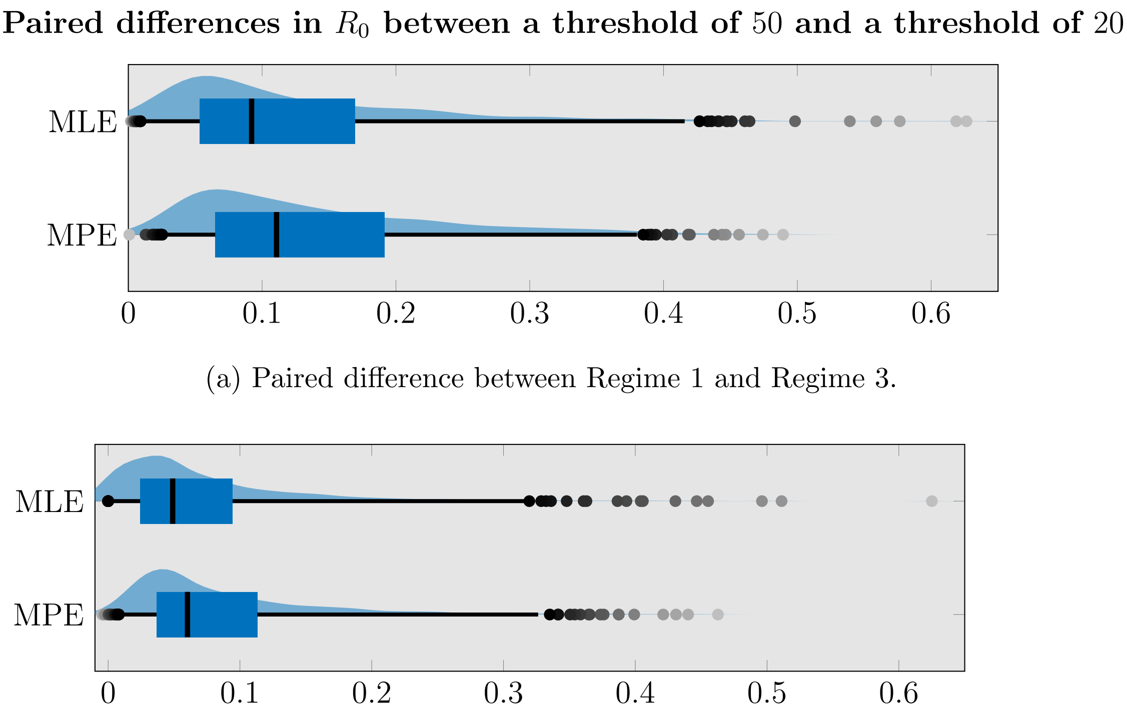
(b) Difference in estimate of the expected proportion of individuals who experience infection.

Figure 4.5: Bean plots of the paired difference between estimates from the unconditioned DA process and the conditioned DA process in Regime 1 plotted against Regime 2, where the difference is defined as the unconditioned estimate minus the conditioned estimate. In all cases the conditioned estimates are smaller than the unconditioned estimates.

Figure 4.5 contains bean plots of the paired difference between estimates from the unconditioned DA process and the conditioned DA process from Regime 1, plotted against Regime 2, where Figure 4.5a shows the difference in estimates of R_0, and Figure 4.5b shows the difference between estimates of the expected proportion of individuals who experience infection. Here, we have defined the difference to be the value of the unconditioned estimate minus the conditioned estimate. Figure 4.5a shows that the unconditioned estimates of R_0 are always larger than the conditioned estimates. On average, the unconditioned estimates are approximately 0.3 higher than the corresponding conditioned estimates. In addition, the MLEs appear more variable than the MPEs, although both distributions have a similar median.

Figure 4.5b translates the differences in estimates of R_0 into differences in the expected proportion of individuals who experience infection, which provides an indication of the extent to which the unconditioned DA process overestimates the size of the outbreak. The median differences in the MLE (MPE) of the expected final epidemic proportions are 26% (20%) and 20% (13%) in Regime 1 and Regime 2. This means that even the most conservative estimate (MPE in Regime 2) over-estimates the size of the outbreak by 13% of the total population, in 50% of realisations. This may have a significant impact on how public heath authorities perceive an emerging epidemic.

Figure 4.6 contains bean plots of the paired difference between the conditioned DA process estimate of R_0 in Regimes 1 and 3 and also between Regimes 2 and 4. In all cases, the estimates in Regimes 3 and 4 are higher than those of Regimes 1 and 2, suggesting that the probability of extinction is considerable even after N_I has exceeded 20. However, the paired differences exhibited here are smaller than the paired differences exhibited in Figure 4.5a, demonstrating that conditioning on a threshold of 20 is preferable to not conditioning at all. It is also clear that the change in the estimated R_0 is lower if the underlying value of R_0 is higher.



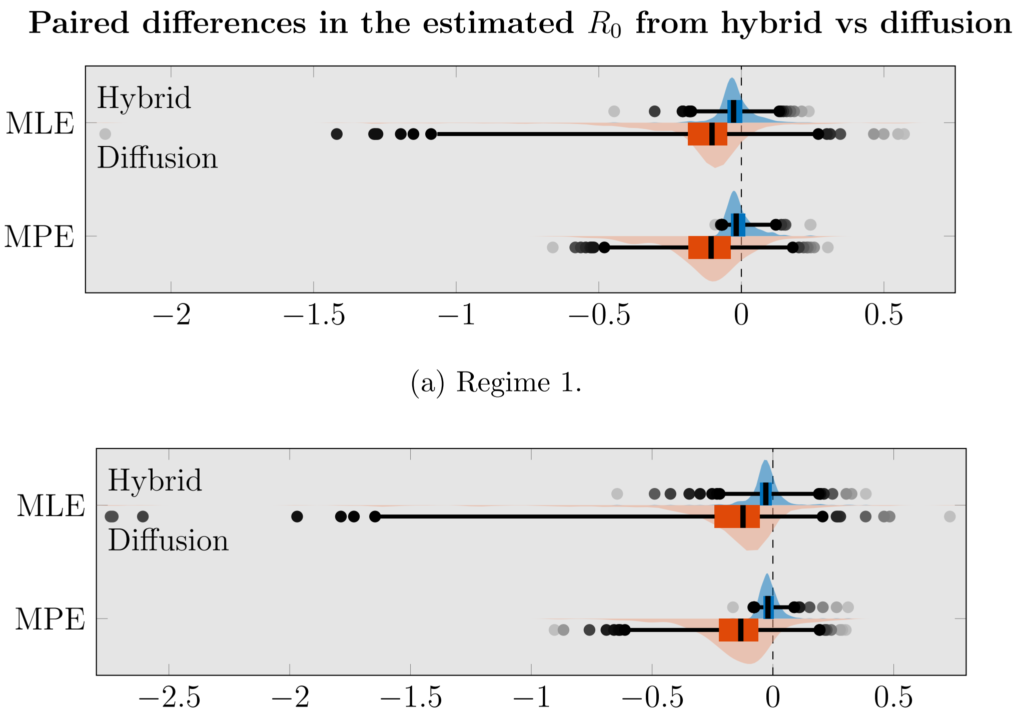
(a) Paired difference between Regime 1 and Regime 3.

(b) Paired difference between Regime 2 and Regime 4.

Figure 4.6: Bean plots of the paired difference in the conditioned DA process estimate of R_0 when the threshold is decreased from 50 to 20, where the difference is defined as the estimate from a threshold of 20 minus the estimate from a threshold of 50. The smaller conditioning level in Regimes 3 and 4 do less to reduce the positive-bias of the unconditioned estimate of R_0.

4.3.2 Validation of the hybrid diffusion model

We now define the paired unconditioned hybrid (diffusion) difference as the estimate of R_0 from the unconditioned hybrid (diffusion) process minus the corresponding estimate from the unconditioned DA process. Figure 4.7 contains bean plots of the paired unconditioned hybrid differences against the paired diffusion differences, under Regimes 1 and 2. The paired diffusion differences demonstrate more bias and variation than the paired unconditioned hybrid differences, suggesting that the hybrid approximation is more reliable



Paired differences in the estimated R_0 from hybrid vs diffusion

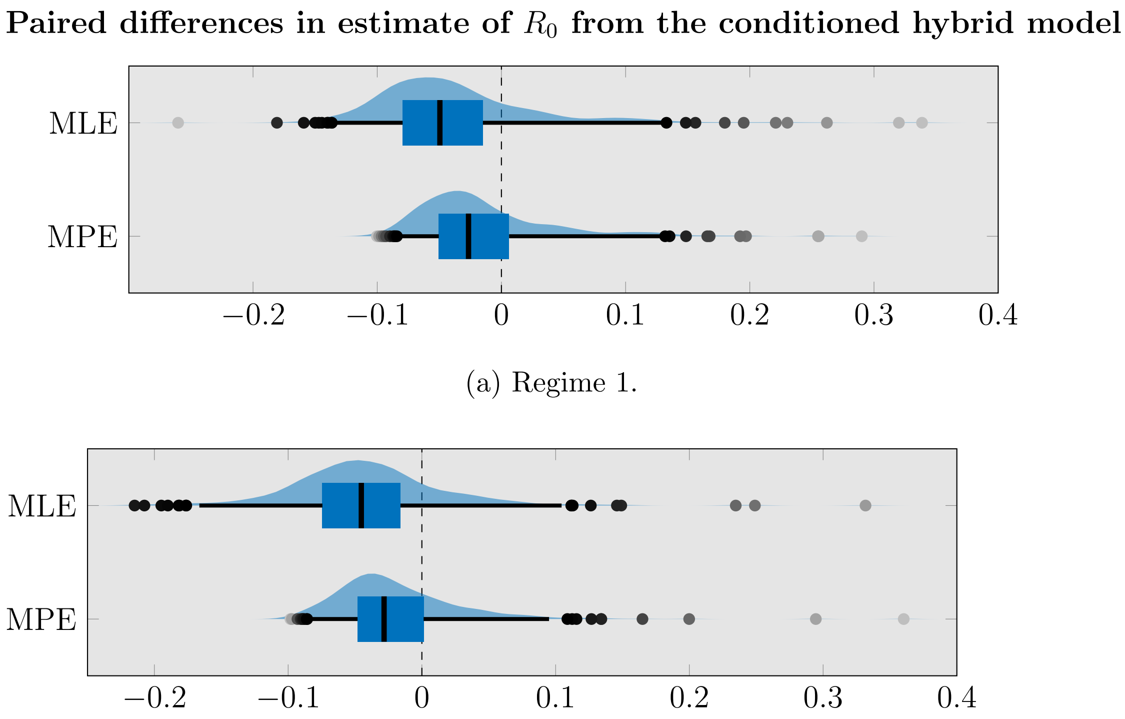
(a) Regime 1.

(b) Regime 2.

Figure 4.7: Bean plots of the paired differences in the estimated R_0 from the unconditioned hybrid against the diffusion. The difference is defined as the estimate from the approximation minus the estimate from the unconditioned DA process. The hybrid approximation is more accurate than the diffusion approximation.

than the diffusion approximation in this context. This is unsurprising because the diffusion approximation is not suitable during the initial stages of an outbreak. However, since the hybrid approximation utilises the diffusion approximation only once the outbreak has become established, the difference exhibited here may be thought of as the amount of error accumulated by the diffusion approximation in modelling the initial stages of the outbreak.

Figure 4.8 shows bean plots of the paired differences between the estimate of R_0 from the conditioned DA and the conditioned hybrid, where the difference is defined as the conditioned hybrid estimate minus the conditioned DA


Paired differences in estimate of R_0 from the conditioned hybrid model

(a) Regime 1.

(b) Regime 2.

Figure 4.8: Bean plots of the paired differences between the conditioned DA estimate of R_0 and the conditioned hybrid estimate of R_0, where the difference is defined as the conditioned hybrid estimate minus the conditioned DA estimate. The hybrid approximation exhibits a small amount of bias.

estimate. The median bias in the MLE of R_0 is approximately -0.05, and the median bias for the MPE of R_0 is approximately -0.03. This indicates that the conditioned hybrid approximation adds a slight (0.03 to 0.05) downwards bias on top of the 0.3 downwards correction of the conditioned DA process, when compared to the unconditioned DA process.

All computations have been carried out with the supercomputing resources provided by the Phoenix HPC service at the University of Adelaide, which is comprised of a Lenovo NeXtScale system consisting of 120 nodes, comprised of 2.3 GHz Intel Xeon E5-2698 v3 CPUs. The Bayesian analysis utilised 3GB of memory and was parallelised over 4 cores. To assess the computational-efficiency of the hybrid approximation we calculated the median runtime (in hours) to compute the MPE, averaged over all $1,000$ realisations. In Regimes 1 and 2 the median computational runtime of the conditioned DA process was 1.27h and 1.55h, compared to 1.17h and 1.17h from the conditioned hybrid likelihood. This small difference in runtime demonstrates that the hybrid model did not have the opportunity to take full advantage of the computational-efficiency of its diffusion dynamics. This is because the simulated data only ran for two weeks, meaning that the total number of infectious cases did not grow much larger than 100. If the simulated realisations were allowed to run for longer then the diffusion approximation would prove to be more beneficial due to a higher number of observed infection events. In Regime 3 the median computational runtime of the conditioned DA process was 0.72h compared to 0.5h from the conditioned hybrid likelihood. In this case the threshold is lower so the hybrid approximation utilised its diffusion dynamics more than in Regimes 1 and 2, hence the hybrid approximation was noticeably faster than the DA process. It is worth noting that the hybrid approximation scales better than the DA process with respect to the total number of observed infection events because its diffusion dynamics are relatively inexpensive, compared to CTMC dynamics.

4.3.3 Application to A(H1N1)pdm09 data

The first human infected with A(H1N1)pdm09 was recorded in the United States on the 15th of April 2009 [Gibbs et al., 2009, Dawood et al., 2009]. Australia's initial response was to delay the entry and spread of the disease by enhanced case-finding, isolation, testing and treatment of incoming travellers with influenza-like illnesses; and prophylactic treatment and home quarantine of the close contacts of suspected/confirmed cases [Glass et al., 2012]. The first confirmed case in Australia was detected in a traveller returning home from the United States on the 9th of May. Subsequently, the first confirmed case in WA was detected in a traveller returning home from Canada via the United States on the 24th of May. On the 13th of June the WA government deemed the outbreak to be widespread and asked doctors to cease active case-finding, and prioritise influenza testing only to persons with severe influenza-like illness or established medical risk conditions [Weeramanthri et al., 2010]. Prior to the 13th of June, all suspected or confirmed cases were actively followed-up and travel histories were recorded. This resulted in 102 confirmed cases and follow-up of 232 household contacts, plus a large number of aeroplane and school contacts. Of these 102 cases, 53% either originated in Victoria or were directly related to cases originating in Victoria. By the 30th of June, a total of 247 cases had been reported.

We are now considering a single outbreak so instead of reporting the distribution of the MLEs and MPEs, we now report the marginal distribution of R_0. We do so by sampling from the posterior distribution of R_0, as before, except this time we report the (2,25,50,75,98)-percentiles of the samples from this distribution, rather than just the median. To achieve this, we use the same parameters as the previous analysis (4 chains of $200,000$ iterations with $20,000$ iterations as burn-in) with the exception that the population size is now assumed to be $2,040,000$, the population of Perth, and the mean of the marginal prior distribution of $1/\gamma$ is set to 3. We changed the mean of $1/\gamma$ to be consistent with other estimates of the mean serial interval of A(H1N1)pdm09 of 2.8 days [Nishiura et al., 2009a,b, Munayco et al., 2009]. To assess the consistency of our methodology, we estimate the distribution of R_0 at a weekly resolution from the 24th of May to the 1st of August.

Since the total number of cases by the 1st of August is prohibitively large for the DA process, we use the hybrid process instead. To demonstrate the impact of conditioning, we estimate the distribution of R_0 with and without conditioning, at the weekly intervals.

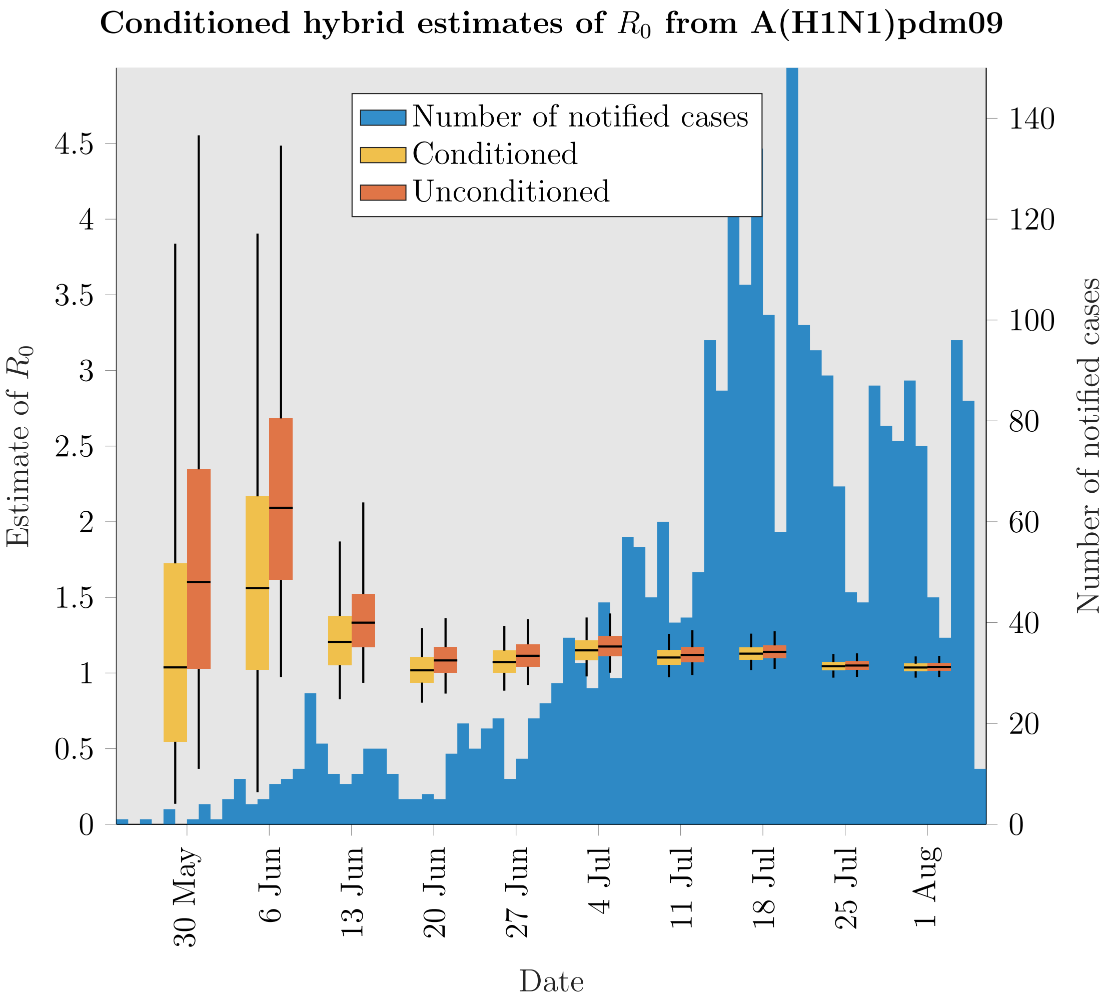


Figure 4.9: Number of notified cases of A(H1N1)pdm09 from WA with box plots of the estimated distribution of R_0 from the conditioned and unconditioned hybrid process. The conditioned hybrid process estimates a lower R_0 than the unconditioned.

Figure 4.9 shows the number of notified cases of A(H1N1)pdm09, and box plots of the estimated distribution of R_0 from the conditioned hybrid in yellow and the unconditioned hybrid in ochre. The statistics of the conditioned distribution are always lower than the corresponding metrics of the unconditioned distribution. This difference is most prominent during the first few weeks of the outbreak and gradually subsides as the outbreak progresses because the impact of accounting for establishment decreases. The variability in the estimated distribution of R_0 can also be observed to decrease as the outbreak progresses. The MPE of R_0 from the conditioned model appears more stable than the MPE of the unconditioned model, which is influenced more heavily by a spike in cases which occurred during the third week of the outbreak. Our MPEs of R_0 from the conditioned hybrid process vary between 1 and 1.1, which are consistent with those in the literature for this outbreak [Kelly et al., 2010]. The computational runtime of this analysis was under 1.5h for the first three weeks of the outbreak, and around 2h thereafter.

4.4 Discussion

We have presented an approach to estimating R_0 from an emerging outbreak by modelling case incidence data with the SIR CTMC. Our approach involves conditioning on the event that the observed number of infection events exceeds a predetermined threshold, at which point the outbreak is considered to be established and simultaneously detected by public health officials. We also presented an accurate and computationally-efficient approximation scheme, suitable for when the total number of observed infectious cases is computationally forbidding. We illustrated the utility of these approaches by estimating R_0 from multiple simulated outbreaks with influenza-like parameters and found our conditioned estimates of R_0 to be 0.3 smaller than the unconditioned estimates, on average. In addition, we demonstrated that the hybrid approach is more computationally-efficient than the standard CTMC approach and more accurate than the usual diffusion approximation.

We applied our methodology to an outbreak of A(H1N1)pdm09 in WA. We found that the conditioned hybrid process provides a more consistent estimate of R_0 during the initial stages of the outbreak, compared to the unconditioned hybrid, and that our estimates agree with those in the literature. However, our assumption that the outbreak is established by the time that the number of infectious individuals exceeds 50 may not be suitable, considering that the case incidence remains low for the first five weeks of the outbreak. Therefore, it might have been more appropriate to condition the outbreak on reaching 102, considering that this is the number of notified cases at the time that the relevant authorities deemed the outbreak to be established [Kelly et al., 2010]. Furthermore, a significant proportion of the notified cases during the initial stages of the outbreak originated outside of WA, making our case incidence data misleading and positively biasing our estimates of R_0. To account for this, future work might consider allowing infectious individuals to enter the population rather than modelling the population as a closed system.

In general terms, the simple SIR CTMC used here is not a biologically plausible model. It makes unrealistic assumptions about the dynamics of the disease, such as the assumption that it has no latent period, and the assumption that each individual's infectious period is exponentially distributed. Furthermore, it does not account for other sources of bias such as incomplete reporting, reporting rates which change over time, population heterogeneity (such as spatial variation, age-specific or household clustering of contacts), imported infectious cases, and pre-existing immunity. Thus, suitable extensions of the conditioned SIR CTMC presented here could be to attempt to further account for bias from any of these sources. Therefore, in the following chapter we utilise the partially-observed SEIR CTMC for inference (Section 2.4). The inclusion of an exposed compartment makes this model more biologically plausible and assuming that infectious individuals are observed randomly makes it more suitable for modelling real outbreaks. Notwithstanding, the salient point of the methodology presented here is that conditioning is a simple mathematical tool which may be applied to a wide range of CTMC models as a means of partially accounting for positive-bias in early estimates of R_0 from case incidence data.

The mechanism by which the hybrid diffusion model switches from CTMC dynamics to diffusion dynamics does not guarantee that the diffusion approximation will provide an accurate representation of the underlying CTMC dynamics. This is because the hybrid diffusion model switches dynamics depending on the number of infection events that have occurred, but the diffusion approximation actually requires the number of infectious individuals to be sufficiently large in order to provide a reliable approximation. Thus, in the following chapter we develop a similar hybrid diffusion model which switches from CTMC to diffusion dynamics once the number of infectious individuals is large enough for the diffusion approximation to provide a reliable approximation.

Chapter 5

Early estimation of the basic reproductive number from partially-observed SEIR CTMC disease dynamics

The approach to obtaining early estimates of the basic reproductive number from case incidence data presented in the previous chapter made two important assumptions: that the dynamics of the disease are suitably described by the SIR model; and, that every infectious case within the population is observed. In this chapter we extend the methodology presented in the previous chapter to the more realistic partially-observed susceptible-exposed-infectious-removed (SEIR) continuous-time Markov chain (CTMC) (Section 2.4).

For many infectious diseases there is a significant exposed/latent period occurring after an individual has been infected but before they are able to transmit the disease [Andersson and Britton, 2000]. The inclusion of an exposed period can result in significantly different disease dynamics and is therefore crucial to the design of appropriate prevention and control policies [Leclerc et al., 2014]. One example is the Ebola Hemorrhagic Fever for which the mean of the exposed period is estimated to range from 9 to 21 days [Lekone and Finkenstädt, 2006].

During the early stages of an outbreak it is highly likely that the recorded number of infectious cases differs from the true number of infectious cases present within the population. The underlying cause of this discrepancy is generally driven by inaccessibility to health care, incorrect/inaccurate case reporting, the prevalence of asymptomatic cases, community attitudes and how the disease is portrayed by the mass-media [Collinson et al., 2015, Mayrhuber et al., 2017, Mitchell and Ross, 2016, Verelst et al., 2016]. A common way of accounting for case under-reporting is to assume that each infectious case is observed with a fixed probability p, and is otherwise unobserved [Fintzi et al., 2017, Wallinga and Teunis, 2004, White and Pagano, 2010]. In so doing, we are also able to specify different infectivities and infectious periods for individuals who are observed and unobserved which provides great flexibility in the model [Mathews et al., 2007].

In this chapter, we generalise the conditioned hybrid diffusion approach of the previous chapter to the partially-observed SEIR CTMC for estimating R_0 from the early stages on an outbreak. Although not presented, we also consider an unconditioned hybrid diffusion approach in the analysis section. Due to the increased complexity of the partially-observed SEIR CTMC, we utilise a dynamic state space truncation algorithm in considering the initial CTMC dynamics of the process [Sunkara and Hegland, 2010, Munsky and Khammash, 2006]. We assess the accuracy of our model by using it to recover the parameters of simulated outbreaks with influenza-like dynamics. In so doing, we investigate the effect of various modelling assumptions on the estimated parameters. For instance, the impact of assuming that the observed and unobserved compartments have different infectivities, compared to if they had the same infectivity. We then demonstrate the utility of our model by using it to infer R_0 for a range of real outbreaks.

5.1 Partially-observed SEIR CTMC model

In this section we present a conditioned hybrid diffusion model of the SEIR CTMC. Like the hybrid diffusion model presented in the previous chapter, the hybrid diffusion model presented here begins with CTMC dynamics and ends with the dynamics of the diffusion approximation. However, since the SEIR CTMC has more compartments than the SIR CTMC, the computational demands of the model increase more rapidly than for the SIR CTMC. As a result, the switching mechanism utilised in the previous chapter is no longer an effective means of reducing the computational demands of the model. Instead, we utilise a dynamic state space truncation rule which enables the outbreak to become established before switching to diffusion dynamics. It follows that the hybrid diffusion model of the SEIR CTMC may be thought of as a three-stage process which begins with the dynamics of the DA process, then progresses to a so-called *truncated DA process* and ends with the diffusion approximation. We now describe the dynamics of the model at each of its three stages.

5.1.1 Stage one: DA process

The first stage of the hybrid diffusion process utilises the familiar dynamics of the DA process. Recall that the DA representation of the SEIR CTMC (Section 2.4) is the CTMC $(\boldsymbol{N}(t), t \geq 0)$, which takes values in $\mathcal{N}$ (Equation (2.56)) and, for all $\boldsymbol{n}$ in $\mathcal{N}$, has the positive transition rates $q^{\boldsymbol{N}}_{\boldsymbol{n}\,\boldsymbol{n}+\boldsymbol{e}_i}$, if $\boldsymbol{n} + \boldsymbol{e}_i$ is in $\mathcal{N}$, for $i = 1, 2, \ldots, 5$ (Equations (2.58)).

Chapter 4 demonstrated that estimates of R_0 from the early stages of the outbreak are likely to be positively-biased if the model does not account for the event that the outbreak becomes established. For this reason, we condition the DA process on the event that the outbreak becomes established (Section 4.1). This is achieved by invoking Theorem 2 to condition the DA process on the event that it hits the set $\mathcal{N}^T$, a subset of $\mathcal{N}$, from which the outbreak is considered established. During the initial stages of an outbreak, the probability of an established outbreak increases considerably each time another individual becomes infectious. Thus, we define $\mathcal{N}^T$ as the subset of $\mathcal{N}$ from which the number of observed infectious individuals, I_o, exceeds some state-threshold $\widehat{I}$ in $0, 1, \ldots, N$. Further, to assure that the diffusion process provides a sufficently accurate representation of the process, we also require that the number of unobserved infectious individuals, I_u, exceeds $\widehat{I}$. Thus we define $\mathcal{N}^T$ as $\{\boldsymbol{n} \in \mathcal{N} \,|\, I_o \geq \widehat{I}, I_u \geq \widehat{I}\}$. In practice, the value of $\widehat{I}$ can be low because it is the sum $I_o + I_u$ which drives the infection process, not the individual values I_o and I_u.

The number of states in the state space of the DA process is $\mathcal{O}\left(N^5\right)$, where N is the population size, meaning that the computational cost associated with the dynamics of the DA process increases rapidly with N. In order to ensure computational-feasibility, we must consider how to keep the state space of the model to a practical size. During the initial stages of the outbreak, the number of exposure events, N_e, grows the fastest. Therefore, we determine when to switch from the DA process to the truncated DA process using the marginal distribution of N_e. We achieve this by setting an absorbing upper bound on N_e, $\hat{n}_e$ in $\{0, 1, \ldots, N\}$, by setting $q^{\boldsymbol{N}}_{\boldsymbol{n}\,\boldsymbol{m}} = 0$ all $\boldsymbol{n}, \boldsymbol{m} \in \mathcal{N}$ for which

$N_e = \hat{n}_e$. Enforcing the condition that, for a pre-defined probability-threshold p_T in $[0,1]$, the hybrid diffusion process has the dynamics of the DA process until time t_{K_1}, where K_1 is defined as

$$K_1 = \min\left\{k \mid \Pr(N_e(t_{k+1}) = \hat{n}_e \mid \mathcal{Y}_{k+1}) \geq p_T\right\},$$

which we refer to as the first switching time. In other words, the time t_{K_1+1} is the first time at which the probability that the N_e compartment of the DA process reaches the state-threshold $\hat{n}_e$, is greater than the threshold-probability p_T. Depending on the average latent period and the average infectious period, the switching time K_1 is likely to occur early in the process while the population of N_{io}, N_{iu}, N_{ro} and N_{ru} are still low and therefore the diffusion approximation will be unsuitable.

Recall that the hybrid diffusion process is used to infer the parameters, $\boldsymbol{\theta} \in \Theta$, of the model via the likelihood (Section 2.4). Given a set of observed incidence counts x_k for $k = 0, 1, \ldots, n$, with corresponding cumulative incidence counts $y_k = \sum_{j=1}^{k} x_j$, for $k = 1, 2, \ldots, K_1 \wedge n$, the probability of the observed data, $L_C^k(\boldsymbol{\theta})$, is computed via Algorithm 3 using the conditioned transition rates.

5.1.2 Stage two: truncated DA process

The DA process is likely to become computationally-infeasible before the outbreak is established. Thus, the second stage of the hybrid diffusion process is to maintain the dynamics of the DA process, while dynamically truncating its state space informed by the diffusion approximation. To assure a smooth transition from the first stage to the second, at the first switching time we condition the DA process on the event that the population of the N_e compartment is strictly less than $\hat{n}_e$. Provided p_T is suitably small, the error incurred by this assumption should be small.

Throughout the first stage, the direct correspondence between the N_{io} component and the observed data allows us to enforce the condition that $y_k \leq N_{io} \leq y_{k+1} + 1$, for all t in $[t_k, t_{k+1}]$ where $k = 1, 2, \ldots, K_1$. However, since the other compartments are not observed, it is not possible to enforce a similar boundary condition upon them. Thus, for each time interval $[t_k, t_{k+1}]$, for $k = K_1 + 1, K_1 + 2, \ldots, n$, we bound each compartment from above and below such that the probability that the process crosses either boundary is less than the pre-determined probability-threshold p_T. Obtaining a suitable lower bound is straightforward because the lower bounds can be determined directly from the distribution of the truncated DA process at the initial time t_k. For example, the lower bound of the N_e compartment on the time interval $[t_k, t_{k+1}]$ is $lb_e = \min\{n \mid \Pr(N_e(t_k) \leq n) \leq p_T\}$. We determine the upper bounds from the diffusion approximation of the DA process, which we now define.

Appealing to Definitions 7 and 8, it can be shown that the DA process on $\mathcal{N}$ is a DDMPP, meaning that its fluid approximation and diffusion approximation exist. Therefore, let $(n_e, n_{io}, n_{iu}, n_{ro}, n_{ru})$ denote the continuous quantities taking values in E, which are analogous to the scaled quantities $(N_e, N_{io}, N_{iu}, N_{ro}, N_{ru})/N$. Then following from Theorem 3, the fluid approximation of the DA process is the deterministic process $(\boldsymbol{n}(t, \boldsymbol{n}_0), 0 \leq t < \infty)$ which is the unique solution to the system of ordinary differential equations

$d\boldsymbol{n}\left(t,\boldsymbol{n}_0\right)/dt = F(\boldsymbol{n}\left(t,\boldsymbol{n}_0\right))$, for a suitable initial value $\boldsymbol{n}_0$ in E, and where

$$F(\boldsymbol{n}) = \begin{bmatrix} \left(\beta_o\left(n_{io}-n_{ro}\right)+\beta_u\left(n_{iu}-n_{ru}\right)\right)\left(1-n_e\right) \\ p\alpha\left(n_e-n_{io}-n_{iu}\right) \\ (1-p)\alpha\left(n_e-n_{io}-n_{iu}\right) \\ \gamma_o\left(n_{io}-n_{ro}\right) \\ \gamma_u\left(n_{iu}-n_{ru}\right) \end{bmatrix}.$$

Following from Theorem 4 the diffusion approximation of the DA process is the Gaussian diffusion process with mean $\boldsymbol{n}\left(t,\boldsymbol{n}_0\right)$ and covariance matrix $\Sigma^N(t)$, where $\Sigma^N(t)$ is the unique solution to the system of ordinary differential equations (2.12), for a suitable initial value $\Sigma^{\boldsymbol{N}}(0) = \Sigma_0$.

Although it has been noted that the diffusion approximation provides a poor representation of the dynamics of the DA process during the initial stages of the outbreak, the main cause of this error is that the diffusion approximation does not accurately represent the dynamics of the DA process around boundaries in its state space. Thus, the diffusion approximation still provides a suitable approximation of the distribution of the DA process away from the boundary, and so is suitable for computing the upper bounds.

The upper bounds of the unobserved compartments are computed via the diffusion approximation as follows. Given the initial state $\boldsymbol{n}_0 = \mathrm{E}[\boldsymbol{N}\left(t_k\right)]/N$ and covariance $\Sigma_0 = \mathrm{cov}\left(\boldsymbol{N}\left(t_k\right)\right)/N$ the distribution of the diffusion approximation is computed at time t_{k+1} and conditioned on the event that $n_{io}(t_{k+1}) = y_{k+1}/N$ (Theorem 6). The upper bounds are then computed directly from the marginals of the conditioned diffusion approximation. For example, the upper bound of the N_e compartment over the time interval $[t_k, t_{k+1}]$, for $k = K_1+1, K_1+2, \ldots, n$, is $ub_e = \min\{n \mid \Pr(n'_e(t_{k+1}) \geq n/N) \leq p_T\}$, where $n'_e(t_{k+1})$ is the $n_e(t_{k+1})$ compartment of the diffusion approximation after conditioning on the event that $n_{io}(t_{k+1}) = y_{k+1}/N$.

It follows that the truncated DA process has the same dynamics as the DA process, with the exception that we only consider its dynamics on the truncated state space $\mathcal{N}_k$, defined as

$$\begin{aligned}\mathcal{N}_k = \Big\{\boldsymbol{n} \in \mathcal{N} \mid \max\{y_k, lb_e\} \leq N_e \leq \max\{\widehat{I} + y_{k+1} + 1, ub_e\} + 1,& \\ y_k \leq N_{io} \leq y_{k+1} + 1,& \\ lb_{ui} \leq N_{iu} \leq ub_{ui} + 1,& \\ lb_{ro} \leq N_{ro} \leq \min\{y_{k+1} + 1, ub_{ro} + 1\},& \\ lb_{ru} \leq N_{ru} \leq ub_{ru} + 1\Big\}.& \end{aligned} \tag{5.1}$$

When switching from one truncated state space $\mathcal{N}_k$, for $k = K_1 + 1, K_1 + 2, \dots, n$, to the next, $\mathcal{N}_{k+1}$ the DA process is conditioned on the event that each of its unobserved compartments is less than its upper bound. This ensures a smooth transition between truncated state spaces.

The truncated DA process switches to the diffusion approximation once the outbreak has becomes established. Given that we consider an outbreak to be established once it has reached the subset $\mathcal{N}^T$, we switch from the truncated DA process to the diffusion approximation once the probability that the truncated DA process has reached the set $\mathcal{N}^T$ exceeds p_I. More precisely, for some pre-defined probability-threshold p_I in $[0, 1]$, the hybrid diffusion process has the dynamics of the truncated DA process until time t_{K_2}, where K_2 is defined as

$$K_2 = \min\left\{k \mid \Pr\left(\boldsymbol{N}\left(t_{k+1}\right) \in \mathcal{N}^T\right) \geq p_I\right\},$$

which we refer to as the second switching time. In other words, the time t_{K_2+1} is the first time at which the probability that the outbreak is in the set $\mathcal{N}^T$ exceeds p_I.

It follows that for all $k = K_1 + 1, K_1 + 2, \ldots, K_2$, the probability of the observed data from the truncated DA process, $L_T^k(\boldsymbol{\theta})$, is computed via Algorithm 3 using the truncated state space (5.1).

5.1.3 Stage three: diffusion approximation

The truncated DA process reduces the computational cost of the DA dynamics by considering the dynamics of the process on only a subset of its state space. This enables the population of infectious individuals to grow large enough for the diffusion approximation to provide an accurate representation of the process while retaining the use of the DA process. To ensure a smooth transition from the truncated DA process to the diffusion process, the diffusion approximation is initialised by the mean and covariance of the truncated DA process. In particular, $\boldsymbol{n}_{K_2+1} = \mathrm{E}[\boldsymbol{N}(t_{K_2+1})]/N$ and $\Sigma_{K_2+1} = \mathrm{cov}(\boldsymbol{N}(t_{K_2+1}))/N$. Similar to the diffusion dynamics in the previous chapter, the diffusion approximation is conditioned on the observed data via Theorem 6. One the process has reached the third stage it is no longer conditioned on reaching the set $\mathcal{N}^T$.

Recall that the diffusion likelihood (2.47) of the SIR CTMC was computed by using the transition density (2.45) to approximate the transition probabilities of the CTMC (2.46). This procedure was efficient because the diffusion approximation followed a bivariate normal distribution so we only needed to evaluate the transition density along one dimension. In the case of the SEIR CTMC, the diffusion approximation follows a 5-dimensional multivariate normal distribution, so a generalisation of the previous approach would require evaluating the transition density across a 4-dimensional grid, which is computationally demanding. Instead, we utilise a highly efficient quasi Monte Carlo approach due to Botev and L'Ecuyer [2015] in which the probability of the observed data is computed by a highly-efficient approximation of the integral of the transition density. In particular, for all $k = K_2 + 1, K_2 + 2, \dots, n$ the diffusion approximation of the probability of the observed data is

$$L_D^k(\boldsymbol{\theta}) = \int_{-\frac{1}{2}}^{1} \int_{y_{k+1}-\frac{1}{2}}^{y_{k+1}+\frac{1}{2}} \int_{-\frac{1}{2}}^{n_e-n_{io}+\frac{1}{2}} \int_{-\frac{1}{2}}^{n_{io}+\frac{1}{2}} \int_{-\frac{1}{2}}^{n_{iu}+\frac{1}{2}} f_N(\boldsymbol{n}, t_k|\mathcal{Y}_{k\text{-}1}) \\ dn_e\, dn_{io}\, dn_{iu}\, dn_{ro}\, dn_{ru}, \quad (5.2)$$

where $f_N(\boldsymbol{n}, t|\mathcal{Y}_{k\text{-}1})$ is the transition density of the diffusion approximation of the SEIR CTMC, conditioned on the history of the process $\mathcal{Y}_{k\text{-}1}$ (see equation (2.45)).

For the set of cumulative incidence counts $y_1, y_2, \dots, y_n$, observed at times $t_1, t_2, \dots, t_n$. The hybrid diffusion likelihood is

$$L(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{k=1}^{K_1} L_C^k(\boldsymbol{\theta}) \prod_{k=K_1+1}^{K_2} L_T^k(\boldsymbol{\theta}) \prod_{k=K_2+1}^{n} L_D^k(\boldsymbol{\theta}), \quad (5.3)$$

where K_1 and K_2 are the first and second switching times, respectively, and $L_C^k(\boldsymbol{\theta})$, $L_T^k(\boldsymbol{\theta})$ and $L_D^k(\boldsymbol{\theta})$ are the probabilities of the observed data from the DA process, truncated DA process and diffusion approximation, respectively. The conditioned hybrid diffusion likelihood is computed via Algorithm 6. The unconditioned hybrid diffusion likelihood can be computed via a similar approach which does not include conditioning.

5.2 Validation of the hybrid diffusion model

In this section we demonstrate the accuracy and utility of our methodology by using it to estimate R_0 from daily incidence data from the first two, three, four and five weeks of an outbreak. We assess the accuracy of our methodology by using it to estimate the parameters of a set of simulated outbreaks from

Algorithm 6: Likelihood of the partially-observed SEIR CTMC model.

```
Begin
    Data: Daily incidence counts $x_1, x_2, \ldots, x_n$.
    Result: Compute the likelihood $L(\boldsymbol{y}|\boldsymbol{\theta})$.
    Set $\hat{n}_e$, $p_T$, $p_I$, $k = 0$ and $y_k = \sum_{j=0}^{k} x_j$, for all $k = 1, 2, \ldots, n$ ;
    Initialise the probability distribution of $\boldsymbol{N}(0)$ as $\boldsymbol{p}^{\boldsymbol{N}}(0) = \boldsymbol{e}_1$ ;
    while $\Pr(N_e(t_{k+1}) \geq \hat{n}_e) \leq p_T$ do
        Truncate the state space, $\mathcal{N}^k = \{\boldsymbol{m} \in \mathcal{N} \,|\, y_k \leq N_I \leq y_{k+1} + 1\}$ ;
        Condition the transition rates on reaching the set $\mathcal{N}^T$ ;
        Compute $p^{\boldsymbol{N}}_{\boldsymbol{n}\boldsymbol{m}}(t_{k+1})$, for all $\boldsymbol{m} \in \mathcal{N}^k$ ;
        Compute the probability $L_E^{k+1}(\boldsymbol{\theta})$ ;
        Condition $\boldsymbol{N}(t_{k+1})$ on the history $\mathcal{Y}_{k+1}$ ;
        Increment $k = k + 1$ ;
    end
    Condition on the event that $\boldsymbol{N}(t_k) \leq \hat{n}_e$ ;
    while $\Pr(\boldsymbol{N}(t_{k+1}) \in \mathcal{N}^T) \leq p_I$ do
        Set $\boldsymbol{n}_k = \mathrm{E}[\boldsymbol{N}(t_k)]/N$ and $\Sigma_k = \mathrm{cov}(\boldsymbol{N}(t_k))/N$ ;
        Compute truncated state space $\mathcal{N}^k$ via diffusion approximation ;
        Condition the transition rates on reaching the set $\mathcal{N}^T$ ;
        Calculate $p^{\boldsymbol{N}}_{\boldsymbol{n}\boldsymbol{m}}(t_{k+1})$, for all $\boldsymbol{m} \in \mathcal{N}^k$ ;
        Compute the probability $L_T^{k+1}(\boldsymbol{\theta})$ ;
        Condition $\boldsymbol{N}(t_{k+1})$ on the history $\mathcal{Y}_{k+1}$ and on being in a
          transient state of $\mathcal{N}_k$ ;
        Increment $k = k + 1$ ;
    end
    Set $\boldsymbol{n}_{K_2+1} = \mathrm{E}[\boldsymbol{N}\,(t_{K_2+1})]/N$ and $\Sigma_{K_2+1} = \mathrm{cov}\,(\boldsymbol{N}\,(t_{K_2+1}))\,/N$ ;
    while $k < n$ do
        Integrate mean and covariance from time $t_k$ to $t_{k+1}$ ;
        Compute the probability $L_D^{k+1}$ ;
        Condition on event $n_{io}(t_{k+1}) = y_{k+1}/N$ ;
        Increment $k = k + 1$ ;
    Compute $L(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{k=1}^{K_1} L_E^k(\boldsymbol{\theta}) \prod_{k=K_1+1}^{K_2} L_T^k(\boldsymbol{\theta}) \prod_{k=K_2+1}^{n} L_D^k(\boldsymbol{\theta})$;
```

the SEIR CTMC. To ensure our analysis is statistically robust, we consider 100 independent simulated outbreaks, each of which adheres to the following properties: it starts with a single observed infectious case, runs for a duration of five weeks, and exceeds 30 observed infectious cases by the end of the fifth week. This is reflected by setting $\hat{n}_e = 30$ and $p_T = 5 \times 10^{-3}$. In addition we set $\widehat{I} = 5$ and $p_I = 0.5$. This choice of $\widehat{I}$ and p_I balance the accuracy of the model with its associated computational-demands and were determiend by trial-and-error. We also assess the impact of imposing a set of assumptions on the parameters of the model. This includes considering a *base* model in which $\beta_o = \beta_u$ and $\gamma_o = \gamma_u$, a *restricted* model in which $\gamma_o = \gamma_u$, and a *full* model in which β_o, β_u, γ_o and γ_u are unconstrained (see Figure 2.3). To further the analysis of the previous chapter, we again consider the estimated parameters under a conditioned model and an unconditioned model. In assessing each model, the true parameter values have been selected to be representative of an outbreak of influenza and the value of N guarantees that the pool of susceptible individuals does not become depleted.

We estimate the probability distribution of the parameters in a Bayesian framework (Section 2.3.3), in which we again use the MAPE for a point estimate for the parameters. The MAPE is the set of parameters which attains the highest marginal posterior density (Section 2.3.3). We focus on estimating the value of $\boldsymbol{\theta}$ which is equal to $(\beta_o/\gamma_o,\ \beta_u/\gamma_u,\ 1/\gamma_o,\ 1/\gamma_u,\ 1/\alpha,\ p)$, for $\boldsymbol{\theta} \in \Theta$, where all of β_o/γ_o, β_u/γ_u, $1/\gamma_o$, $1/\gamma_u$,$1/\alpha$ are greater than zero and $p \in (0.1, 0.9)$. We use this parameterisation because the relationships between $\beta_o/\gamma_o, \beta_u/\gamma_u$ and $1/\gamma_o, 1/\gamma_u$ are more straightforward than the relationship between β_o, β_u and γ_o, γ_u, respectively. Furthermore, we restrict the values of p in this way because numerical issues arise when it is too close to either 0 or 1. We utilise an exponential prior

$$f(\boldsymbol{\theta}) = C\, e^{-\boldsymbol{\theta}\boldsymbol{c}},$$

where $\boldsymbol{c}$ and C depend on the model, for all parameters except p which has a uniform prior. Our proposal density is a truncated Bivariate Gaussian with support Θ and fixed covariance structure where $\text{var}(\beta_o/\gamma_o) = \text{var}(\beta_u/\gamma_u) = \text{var}(1/\gamma_o) = \text{var}(1/\gamma_u) = \text{var}(1/\alpha) = 0.1$, $\text{var}(p) = 0.01$, $\text{cov}\,(\beta/\gamma, 1/\gamma) = 0.01$ and $\text{cov}\,(\theta_i, \theta_j) = 0$ otherwise. For each simulated data set, we run four independent Markov chain Monte Carlo chains on Θ consisting of $200,000$ iterations, and discard the initial $20,000$ iterations as burn-in. We now discuss the results from the base model, restricted model and full model.

5.2.1 Base model

For the base model we assume that $\beta_o = \beta_u$ and $\gamma_o = \gamma_u$. As a result, the model is parameterised by $\boldsymbol{\theta} = (R_0, 1/\gamma, 1/\alpha, p)$. For the simulation study, we set the true parameters to $\boldsymbol{\theta} = (2, 3, 1, 0.3)$ which are representative of an influenza-like outbreak. Figure 5.1 shows the statistical properties of the 100 simulated outbreaks used for estimation. For each day, the cyan curve

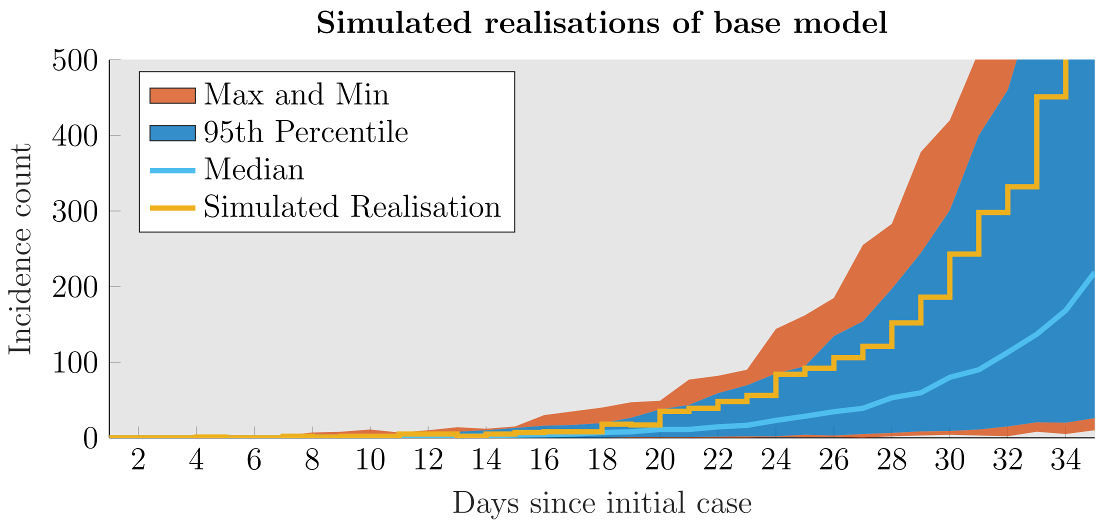


Figure 5.1: Statistical properties of the simulated outbreaks used for estimation by the base model. For each day, the cyan curve shows the median incidence count, the blue shaded area shows the central 5%–95% percentiles of the incidence counts and the red shaded area bounds biggest and smallest incidence counts. The simulated outbreak in yellow is used to inform the posterior distribution in Figure 5.3.

shows the median incidence count, the blue shaded area shows the central 5%–95% percentiles of the incidence counts and the red shaded area bounds the biggest and smallest incidence counts

For the prior distribution, we let $\boldsymbol{c}$ be a 1×4 vector with entries $c_1 = 1/1.3$, $c_2 = 1/5$, $c_3 = 1/1.3$ and $c_4 = 0$, in addition $C = c_1 c_2 c_3$ (Figure 5.2). These values provide the same prior distribution for R_0 and $1/\gamma$ as the previous chapter and were selected with the understanding that they provide sufficient density to values of R_0 and $1/\gamma$ which are reasonable for an influenza-like outbreak. The prior distribution of $1/\alpha$ was selected with the understanding that the duration of the exposed period for an influenza-like disease is often close to one day [CDC, 2016].

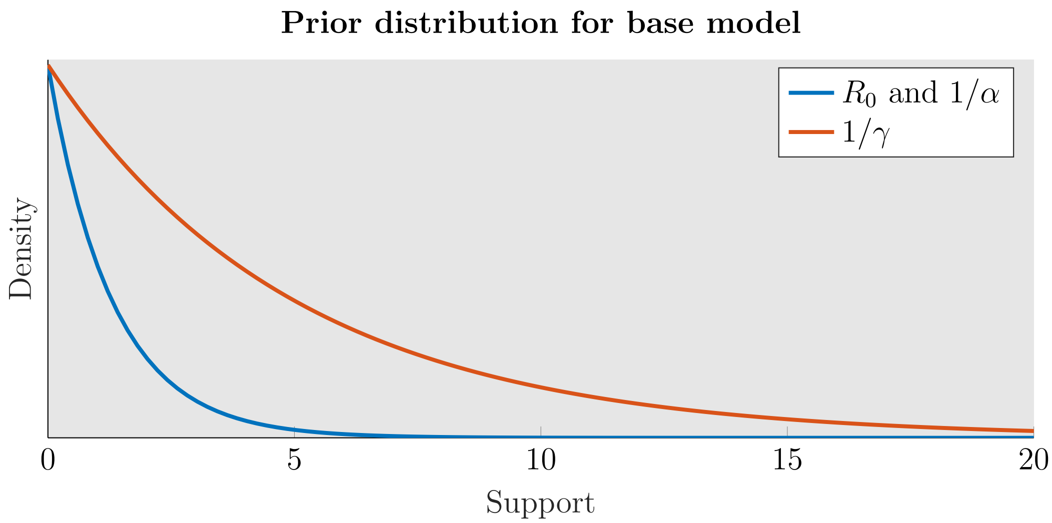


Figure 5.2: Prior distribution of R_0, $1/\gamma$ and $1/\alpha$ for the base model. This choice is similar to the prior distribution from the previous chapter and provides adequate support to parameter values which are reasonable for an influenza-like outbreak.

Figure 5.3 shows the estimated joint posterior distribution of the parameters under the unconditioned hybrid diffusion model, based on the first two weeks of the simulated outbreak shown in Figure 5.1. The posterior distribution demonstrates that in this case the parameters are reasonably insensitive to the prior distribution as they do not follow an exponential distribution, or a uniform distribution in the case of p. A strong correlation between R_0 and $1/\gamma$ can be observed. The value of $1/\alpha$ appears relatively insensitive to R_0 and p. The true parameters are shown in green and their estimates are shown in ochre. It's worth noting that although the estimated value of R_0 appears close to its true value, it is over-estimated in this case.

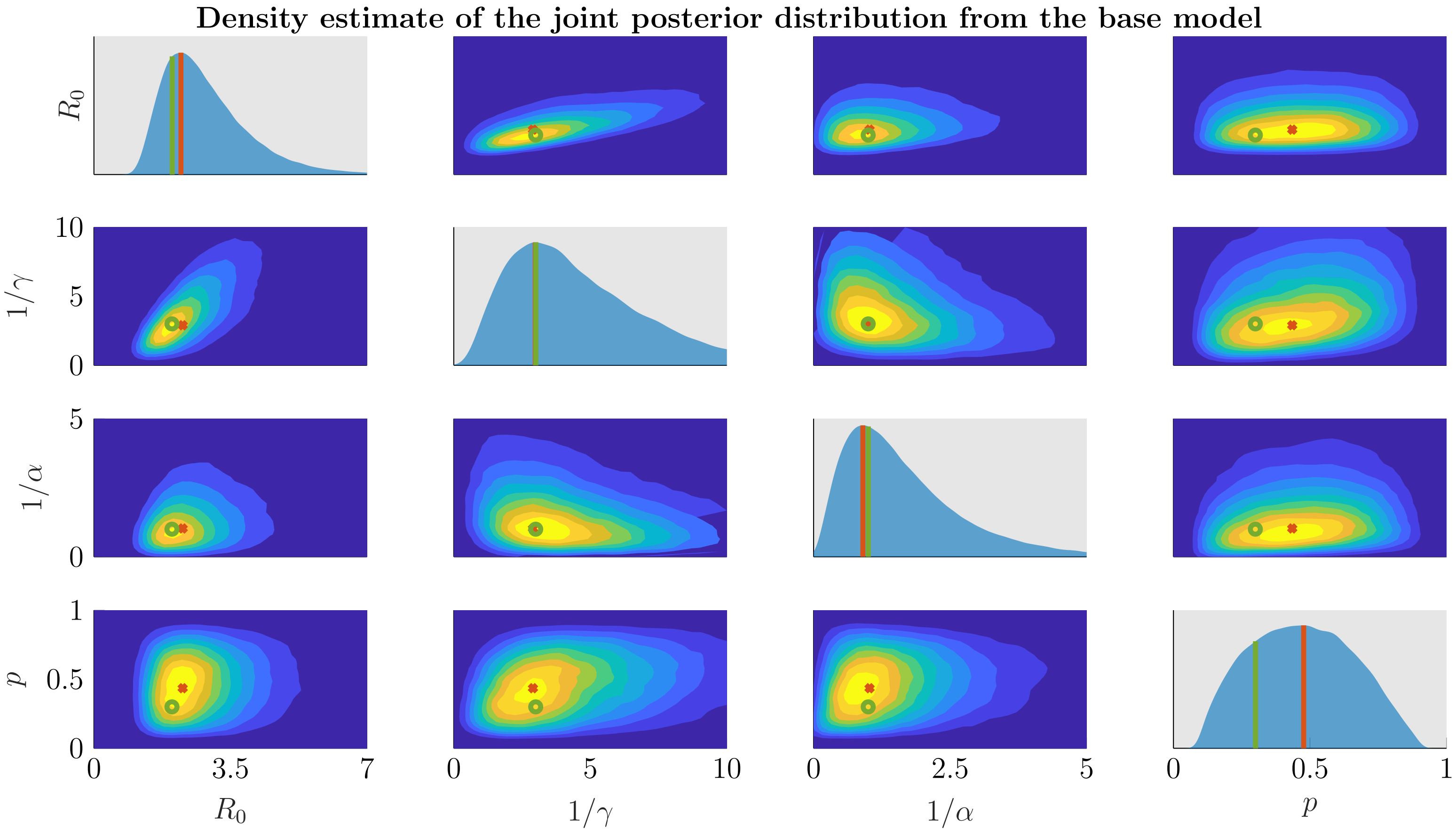
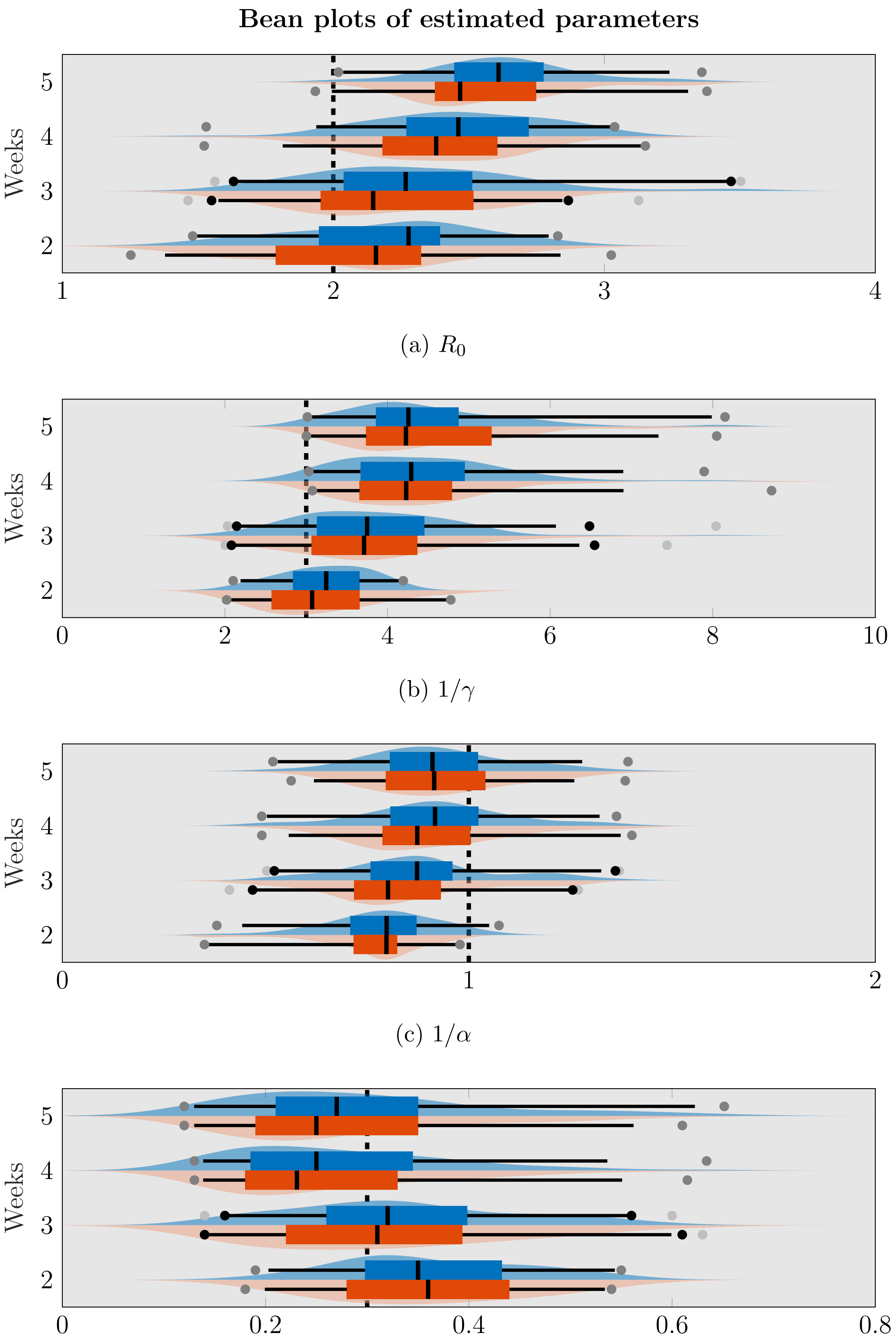


Figure 5.3: Joint posterior distribution of the parameters under the unconditioned base model from the first two weeks of the simulated outbreak shown in Figure 5.1. The true parameter values are displayed in green while the estimated parameters are displayed in red.

Figure 5.4 shows bean plots of the point estimates of the parameters from the unconditioned model in blue and the conditioned model in red. These estimates are based on the first two, three, four and five weeks of the simulated outbreaks from Figure 5.1. Recall that a bean plot contains side by side boxplots of the data in which the 2nd, 25th, 50th, 75th and 98th percentiles are indicated, alongside kernel density estimates of their distribution. The true values of the parameters are indicated by the dashed black lines. As the length of the observation period increases, the average bias of the estimates of R_0, pR_0 and $1/\gamma$ increases, p decreases and $1/\alpha$ remains relatively consistent. It can be seen that the rate in which the average bias of estimates of R_0 increases is comparable to the rate at which the average bias of estimates of p decreases. This suggests that the model has an identifiability problem between R_0 and p. It follows that the base model is unable to accurately estimate R_0 and p, but is able to infer their product. A potential cause for this identifiability problem is that the observed infection process is driven at rate pR_0, suggesting that the model is able to detect this rate but can not detect any further information about p or R_0. It can be observed that the conditioned estimates demonstrate less bias than the unconditioned estimates, on average.

5.2.2 Restricted model

We now consider the restricted model, in which $\gamma_o = \gamma_u$. As a result, the restricted model is parameterised by $\boldsymbol{\theta} = (\beta_o/\gamma, \beta_u/\gamma, 1/\gamma, 1/\alpha, p)$. We set the true parameters to $(0.8, 2.5, 3, 1, 0.3)$ which are similar to true parameters of the base model, but with the assumption that observed infectious cases cause fewer secondary cases than unobserved infectious cases. The rationale behind this is the assumption that individuals who go to the doctor to report

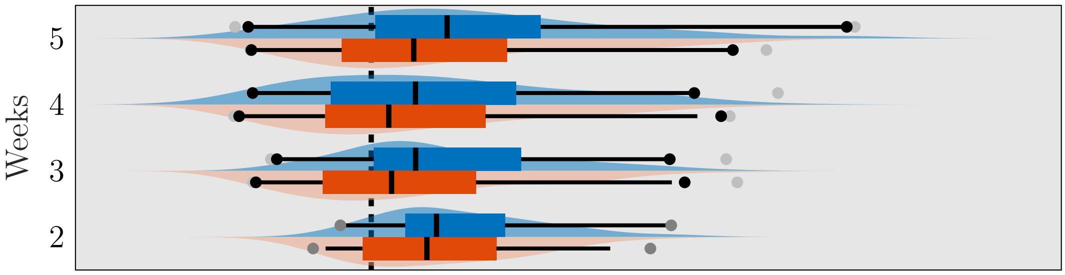


(a) R_0

(b) $1/\gamma$

(c) $1/\alpha$

(d) p



(e) pR_0

Figure 5.4: Bean plots of the point estimates of the parameters from the conditioned and unconditioned base model with the true values indicated by the black dashed line. The estimated values of $1/\gamma$ and $1/\alpha$ are reasonably accurate while the estimates of R_0 and p demonstrate an identifiability problem.

their symptoms are less likely to spread the disease than individuals who do not go to the doctor. The value of R_0 is now 1.99, which is similar to the previous value of $R_0 = 2$ under the base model. The statistical properties of the simulated outbreaks used for estimation under the restricted model are shown in Figure 5.5. It can be seen that the statistical properties of the simulated outbreaks used for estimation by the restricted model have similar statistical properties to the simulated outbreaks used for estimation by the base model.

For the prior distribution, we let $\boldsymbol{c}$ be a 1×5 vector with elements $c_1 = 1/1.3$, $c_2 = 1/1.5$, $c_3 = 1/5$, $c_4 = 1/1.3$ and $c_5 = 0$, and $C = \prod_{i=1}^{4} c_i$ (Figure 5.6). In this case, we use the same prior distribution for $1/\gamma$, $1/\alpha$ and p as the analysis with the base model. The prior distribution for β_o/γ is the same as the prior distribution for R_0, and the prior distribution for β_u/γ is similar to the prior distribution for R_0 except it provides more weight to slightly larger values.

Figure 5.7 shows the estimated joint posterior distribution of the parameters under the unconditioned restricted model, based on the first three

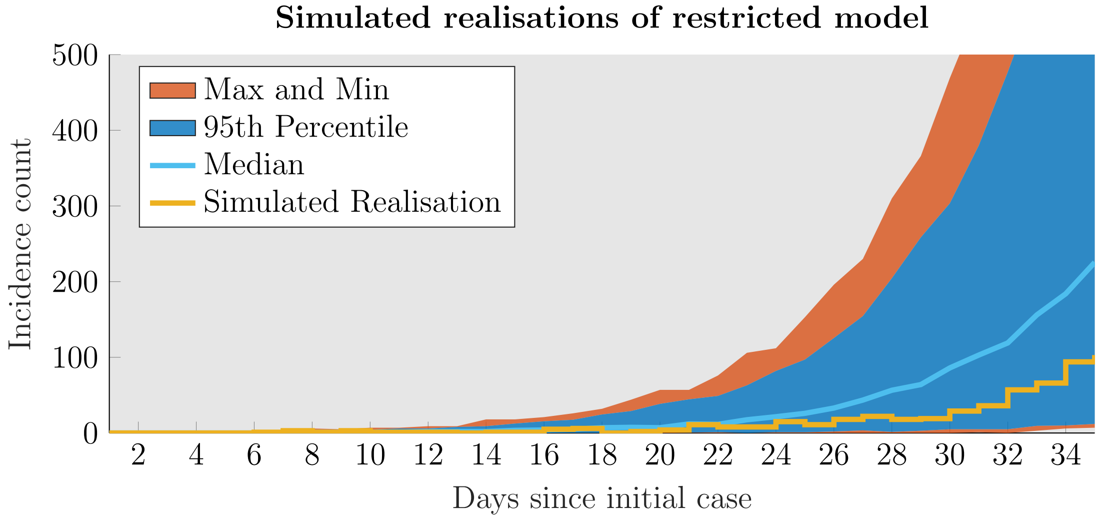


Figure 5.5: Statistical properties of the simulated realisations used for estimation with the full model. For each day, the cyan curve shows the median incidence count, the blue shaded area shows the central 5%–95% percentiles of the incidence counts and the red shaded area bounds biggest and smallest incidence counts. The simulated outbreak shown in yellow is used to obtain the estimated posterior distribution in Figure 5.7.

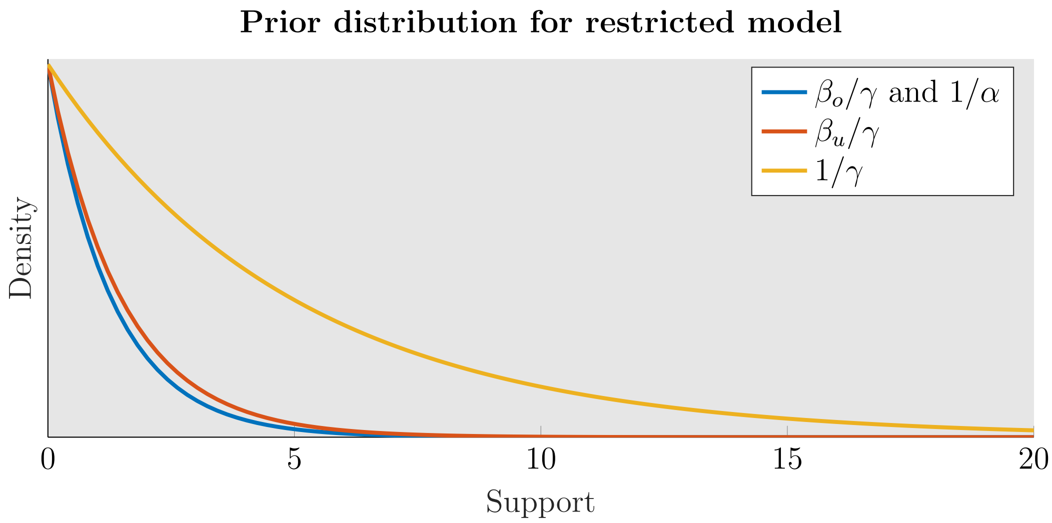


Figure 5.6: Prior distribution of β_o/γ, β_u/γ, $1/\gamma$ and $1/\alpha$ for the restricted model. The prior distribution used here is the same as the prior distribution for the base model, with addition of an extra dimension for β_u/γ, which is similar to the prior distribution of β/γ in the base model.

weeks of the simulated outbreak shown in Figure 5.5. Again, the posterior distribution demonstrates that the parameters are reasonably insensitive to the prior distribution and that β_o/γ and β_u/γ are strongly correlated with $1/\gamma$. An interesting feature is that the correlation exhibited between β_o/γ and β_u/γ is related by the equation

$$R_0 = p\frac{\beta_o}{\gamma_o} + (1-p)\frac{\beta_u}{\gamma_u}.$$

Substituting the estimated values of R_0 and p into this equation produces the purple line which has been plotted on the joint posterior distribution of β_o/γ and β_u/γ. It can be seen that the true value of R_0 lies on this line, suggesting that although the estimated values of β_o/γ and β_u/γ are inaccurate, the resulting estimate of R_0 is accurate. This is supported by the observation that the ridge in the joint posterior distribution of β_o/γ and β_u/γ coincides with this line, suggesting that the model favours values of β_o/γ and β_u/γ which provide the correct R_0 but the model has trouble identifying the underlying values of β_o/γ and β_u/γ.

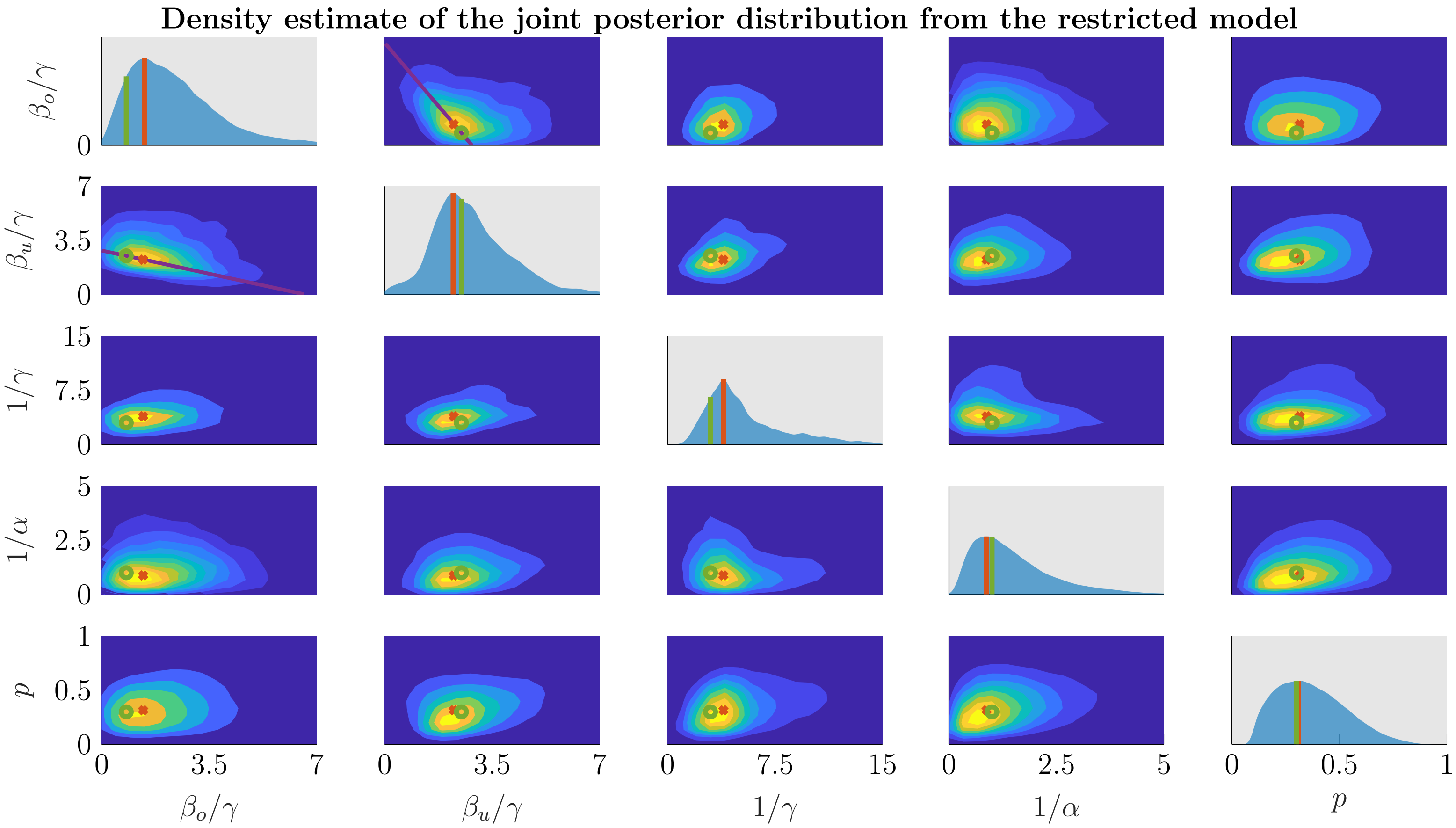


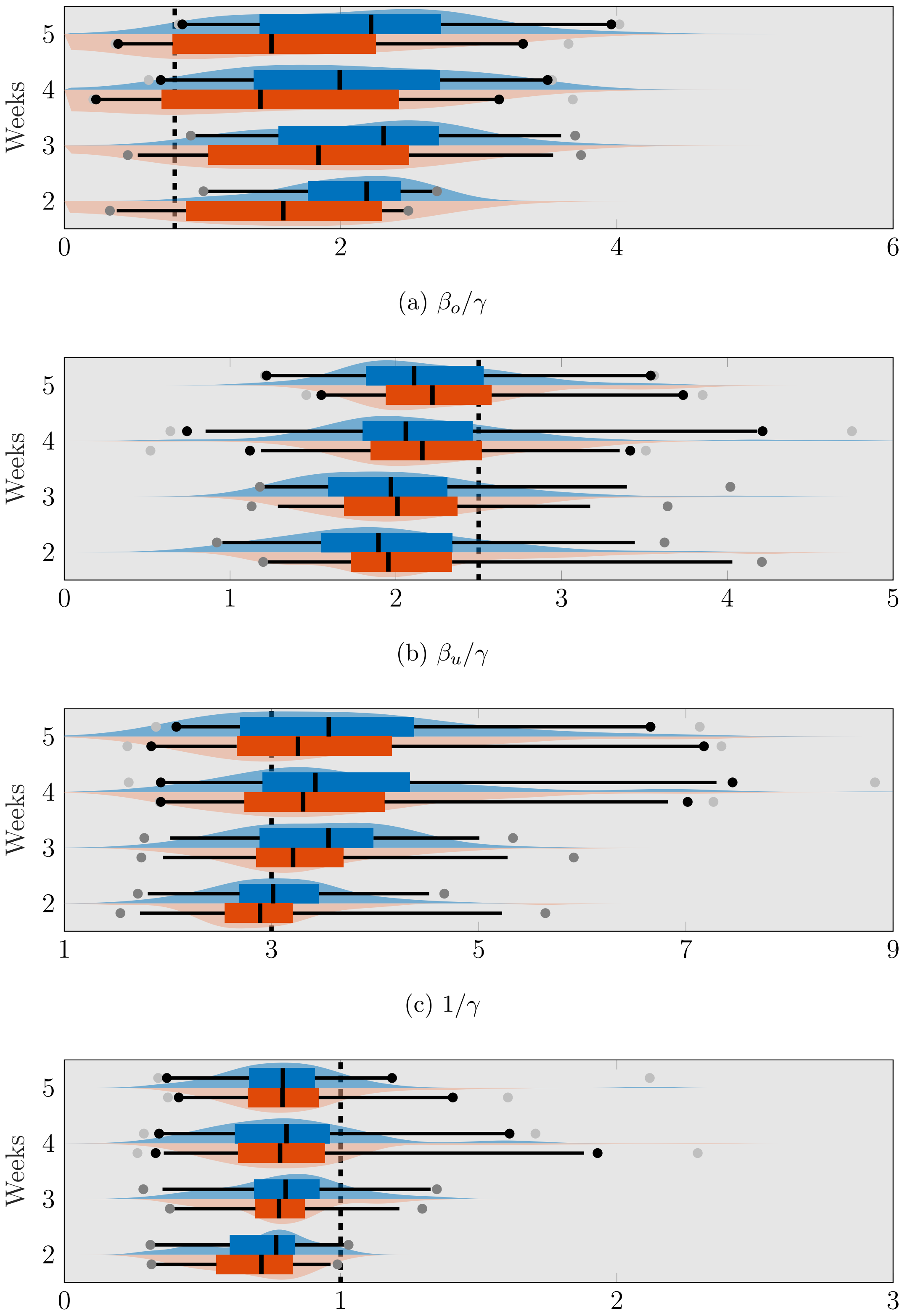
Figure 5.7: Joint posterior distribution of the parameters based of the first two weeks of the simulated outbreak displayed in Figure 5.5. The true values of the parameters are displayed in green while the estimated parameters are displayed in red. The equation of R_0 provides the purple line which explains the correlation between β_o/γ and β_u/γ.

Figure 5.8 shows bean plots of the estimated parameters from the unconditioned model in blue and the conditioned model in red. The estimates are based on the first two, three, four and five weeks of the simulated outbreaks. As the length of the observation period increases, the average bias of estimates of β_o/γ, β_u/γ and p decreases, while the average bias of estimates of $1/\gamma$ and $1/\alpha$ is relatively consistent. As suggested by the posterior distribution, it can be seen that the decrease in the average estimates of β_o/γ coincides with an increase in the average estimates of β_u/γ and the resulting estimate of R_0 is reasonably accurate. It can be seen that the conditioned estimates demonstrate less bias than the unconditioned estimates, on average.

5.2.3 Full model

We now consider the full model in which there are no constraints placed on $\beta_o, \beta_u, \gamma_o, \gamma_u$. The model is therefore parameterised by $\boldsymbol{\theta} = (\beta_o/\gamma_o, \beta_u/\gamma_u, 1/\gamma_o, 1/\gamma_u, 1/\alpha, p)$. We set the true parameters to $\boldsymbol{\theta} = (0.8, 2.5, 3, 4, 1, 0.3)$ which are similar to the true parameters in the previous two cases, but with the assumption that observed infectious cases are removed faster than unobserved infectious cases. The rationale behind this assumption is that individuals who seek treatment are more likely to be removed sooner, compared to those who do not seek treatment. The statistical properties of the simulated outbreaks used for estimation are shown in Figure 5.9. It can be seen that the simulated outbreaks used for estimation by the full model provide fewer observed incidence counts than the previous two models. The reason for this is that, compared to the restricted model, the full model has a larger value of $1/\gamma_u$ which results in a smaller value of β_u since the ratio β_u/γ_u is held constant. Therefore resulting in fewer infectious cases. Despite this, the value of R_0 is unchanged.

Bean plots of estimated parameters

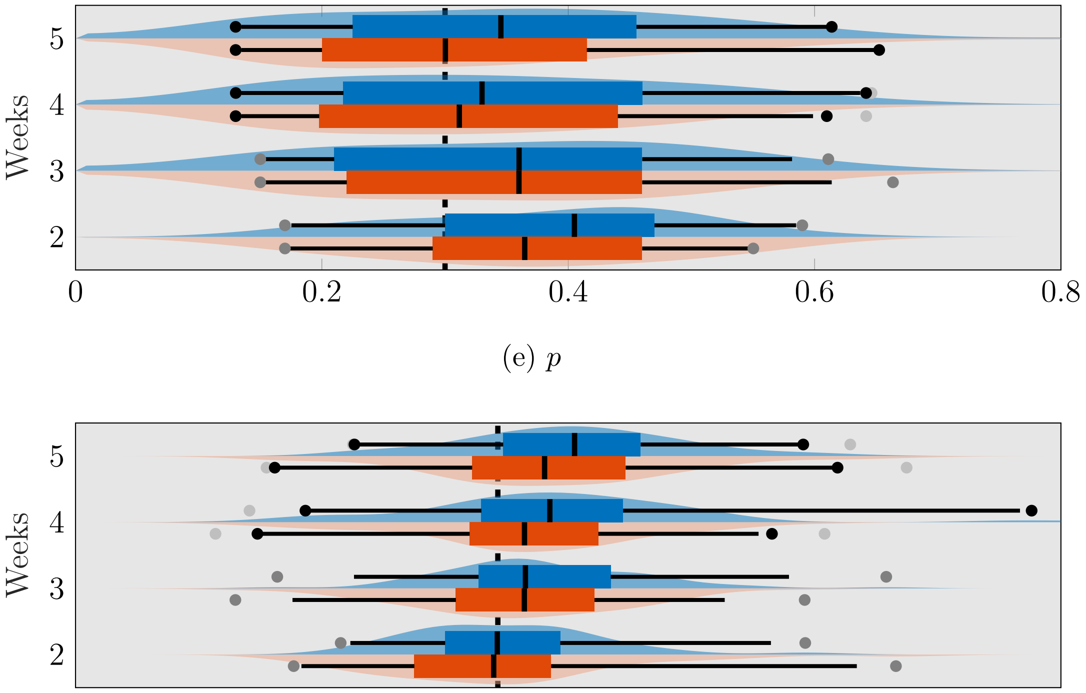


(a) β_o/γ

(b) β_u/γ

(c) $1/\gamma$

(d) $1/\alpha$



(e) p

(f) R_0

Figure 5.8: Bean plots of the estimated parameters from the unconditioned (blue) and conditioned (red) restricted model, with the true values indicated by the black dashed line. The estimated values of p and $1/\gamma$ are reasonably accurate. On average, the estimated value of R_0 is reasonably close to the true value.

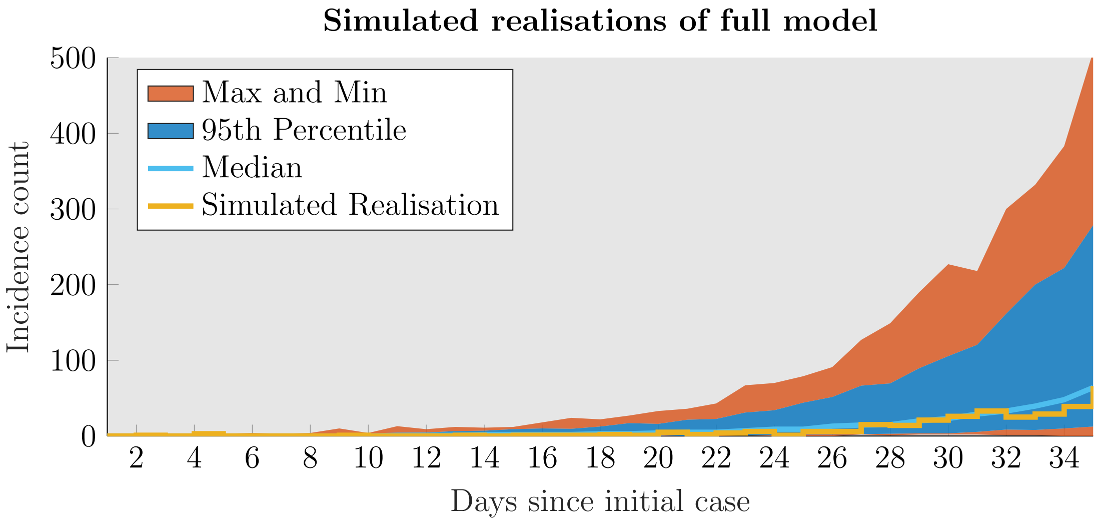


Figure 5.9: Statistical properties of the simulated realisations used for estimation with the full model. For each day, the cyan curve shows the median incidence count, the blue shaded area shows the central 5%–95% percentiles of the incidence counts and the red shaded area bounds biggest and smallest incidence counts. The simulated realisation shown in yellow is used to estimate the joint posterior distribution in Figure 5.11.

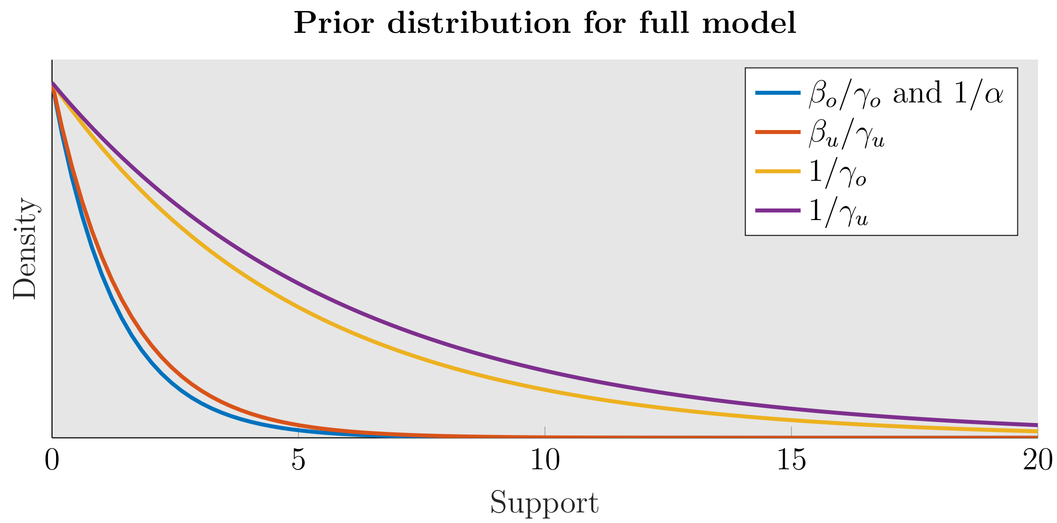


Figure 5.10: Prior distribution of β_o/γ_o, β_u/γ, $1/\gamma_o$, $1/\gamma_u$ and $1/\alpha$ for the full model. The prior distribution used here is the same as the prior distribution for the restricted model, with addition of an extra dimension for $1/\gamma_u$ which is similar to the prior distribution for the restricted model.

For the prior distribution, we let $\boldsymbol{c}$ be a 1×6 vector with entries $c_1 = 1/1.3$, $c_2 = 1/1.5$, $c_3 = 1/5$, $c_4 = 1/6$, $c_5 = 1/1.3$ and $c_6 = 0$, and $C = \prod_{i=1}^{5} c_i$. The prior distribution is shown in Figure 5.10. In this case, we use the same prior distribution for $1/\alpha$ and p as the restricted model. For β_o/γ_o, β_u/γ_u and $1/\gamma_o$ we use the same prior distribution as β_o/γ, β_u/γ and $1/\gamma$, respectively. For $1/\gamma_u$ we utilise a similar prior distribution to the prior distribution of $1/\gamma_o$, with the exception that the prior distribution for $1/\gamma_u$ has slightly more weight for larger values.

Figure 5.11 shows the estimated joint posterior distribution of the parameters under the full unconditioned model, based on the first three weeks of the simulated outbreak shown in Figure 5.9. As with the restricted model, the full model demonstrates a clear correlation between β_o/γ_o and $1/\gamma_o$, β_u/γ_u

and $1/\gamma_u$, and β_o/γ_o and β_u/γ_u, with the latter being a result of the equation for R_0. An interesting feature which can be seen here, but not in the case of the restricted model, is that the posterior distribution of p is correlated with β_o/γ_o and β_u/γ_u. To see this, the estimated values of β_u/γ_u (β_o/γ_o) and R_0 are substituted into the equation for R_0 to produce the purple curve shown in the joint posterior distribution of p and β_o/γ_o (β_u/γ_u). It can be seen that in both cases the curve coincides with a ridge in the joint posterior distribution.

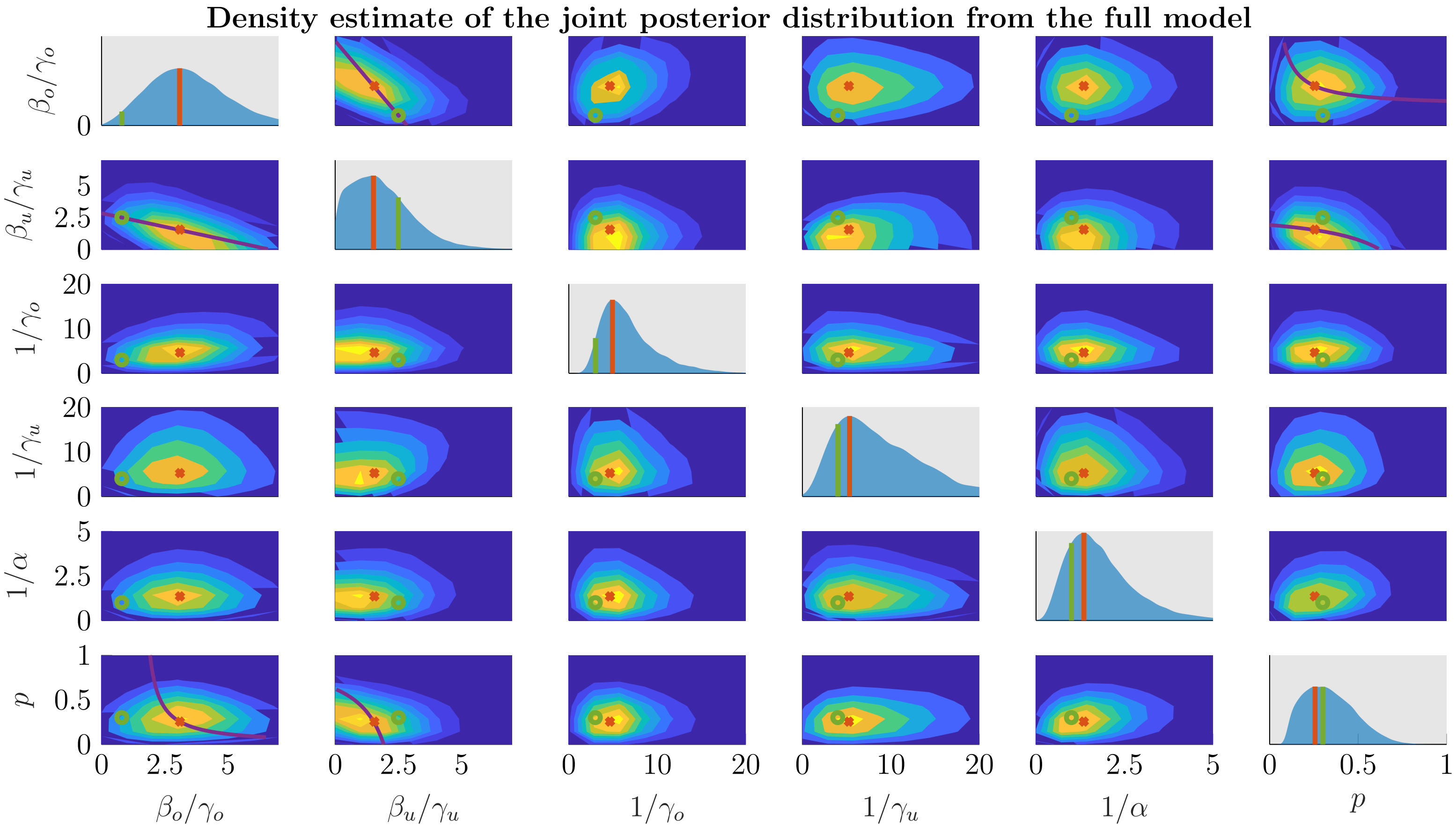


Figure 5.11: Joint posterior distribution of the parameters based of the first three weeks of the simulated outbreak displayed in Figure 5.9. The true parameter values are displayed in green while the estimated parameters are displayed in red.

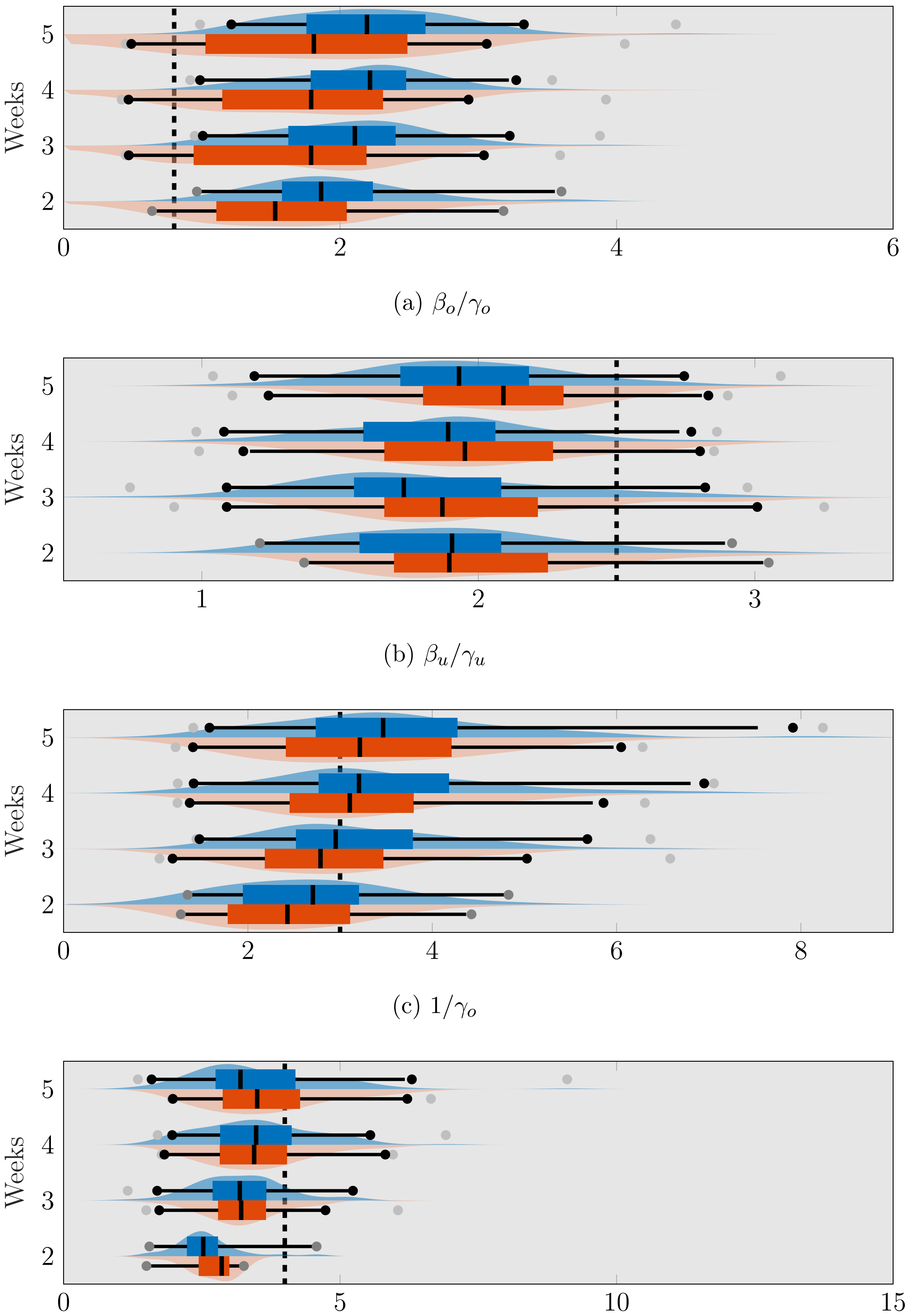
Figure 5.12 shows bean plots of the estimated parameters from the unconditioned model in blue and the conditioned model in red. The estimates are based on the first two, three, four and five weeks of the simulated outbreaks. As the length of the observation period increases, the average bias of estimates of β_u/γ_u and p decreases, while the average bias of the other estimates is relatively consistent. Similar to the restricted model, it can be observed that the decrease in the average estimates of β_u/γ_u coincides with a decrease in the average estimates of p and results in a reasonably accurate estimate of R_0. It can be observed that the conditioned estimates demonstrate less bias than the unconditioned estimates, on average.

Comparing the estimates of the restricted model from Figure 5.8 to the estimates of the full model from Figure 5.12, it can be seen that on average the full model provides a more accurate estimate of $1/\gamma_o$, $1/\gamma_u$ and $1/\alpha$ and a comparable estimate of R_0 and p.

5.3 Application to data

In this section we demonstrate the utility of our model by applying it to real outbreaks of infectious diseases. Based on the results of the previous section, we utilise the conditioned version of the full model, which provided accurate estimates of the duration of the exposed period and the duration of the infectious periods, while providing comparably accurate estimates of R_0 and p to the restricted model. Throughout this section, we conduct estimation via the same approach as the last section, unless otherwise stated.

Bean plots of estimated parameters

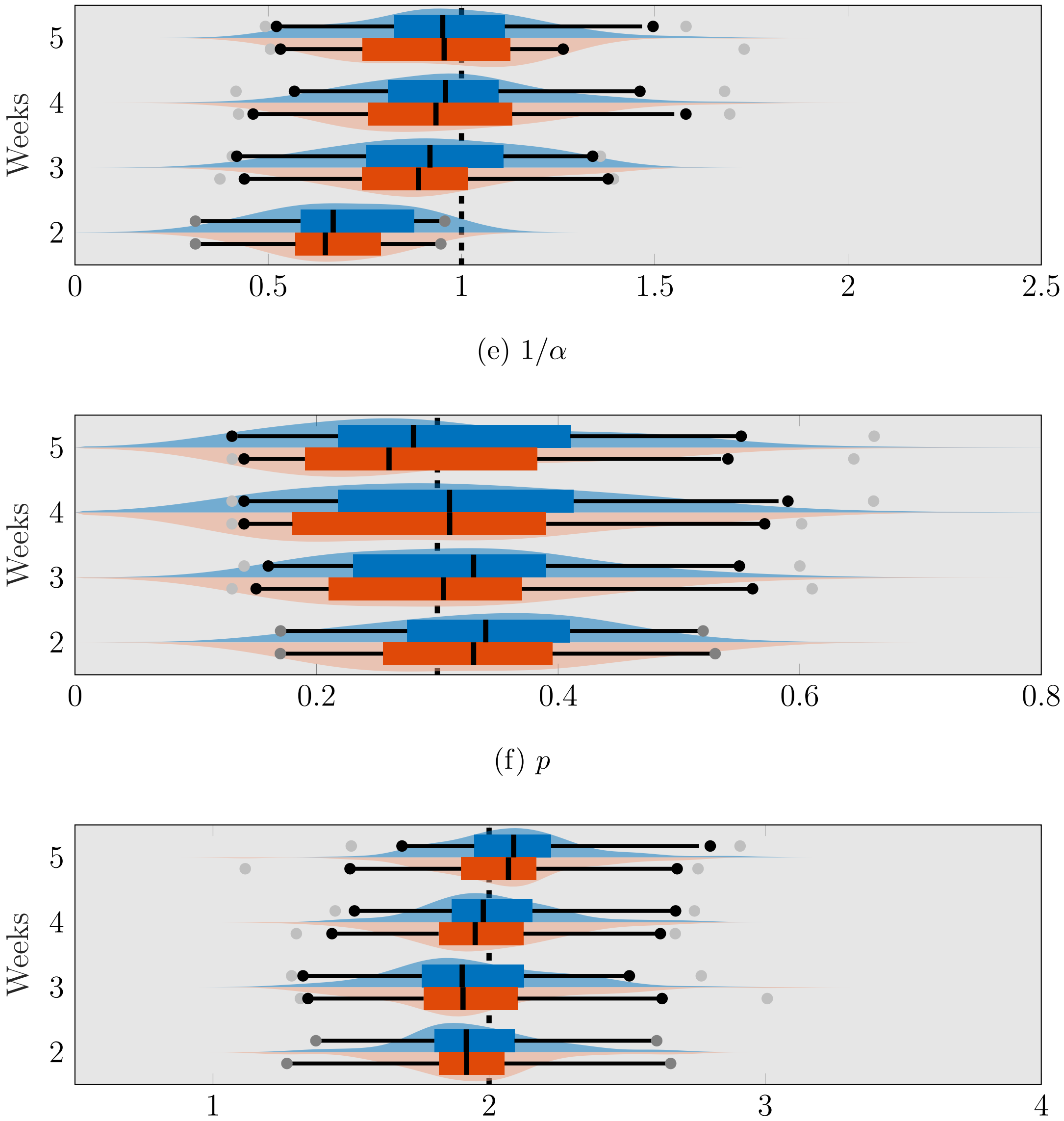


(a) β_o/γ_o

(b) β_u/γ_u

(c) $1/\gamma_o$

(d) $1/\gamma_u$



(e) $1/\alpha$

(f) p

(g) R_0

Figure 5.12: Bean plots of the point estimates of the parameters from the conditioned and unconditioned base model with the true values indicated by the black dashed line. The full model provides reasonably accurate estimates of all parameters aside from β_o/γ_o and β_u/γ_u. Despite this, the average estimated value of R_0 exhibits only a small amount of bias.

5.3.1 A(H1N1)pdm09 in Western Australia

We begin by reconsidering the outbreak of A(H1N1)pdm09 from Section 4.3.3. In this case, we estimate R_0 from the first seven weeks of the outbreak, prior to the sudden increase in case reporting around week eight. Figure 5.13 shows the daily incidence count of the disease in blue, alongside weekly box plots describing the posterior distribution of R_0. The median of the samples from the posterior is indicated in black and the MAPE is indicated in yellow. It can be seen that the 25th and 75th percentiles of our distribution of R_0 are generally contained within 1 to 1.5, with the exception of the second week of the outbreak. The reason for the higher than average estimated distribution of R_0 in the second week is that incidence counts seem to suggest that the outbreak is about to take off. However, by the end of the third week it is apparent that this is more of a stochastic fluctuation. The estimates of R_0 produced here are slightly higher than the estimates of R_0 produced in Chapter 4, which is not suprising given that we now allow for non-reporting of cases.

5.3.2 Ebola hemorrhagic fever in Zaire 1976

Ebola is a highly infectious and lethal disease which recently attained international concern after an outbreak in Western Africa. The virus is transmitted by physical contact with body fluids, secretions, tissues or semen from infectious individuals. Individuals who contract the disease have an exposed period ranging up to 21 days, with an average of 6.3 days. Its symptoms are characterised by initial influenza-like symptoms which rapidly progress to vomiting, diarrhoea, rashes, and internal and external bleeding [WHO, 2017a, Breman et al., 1999].

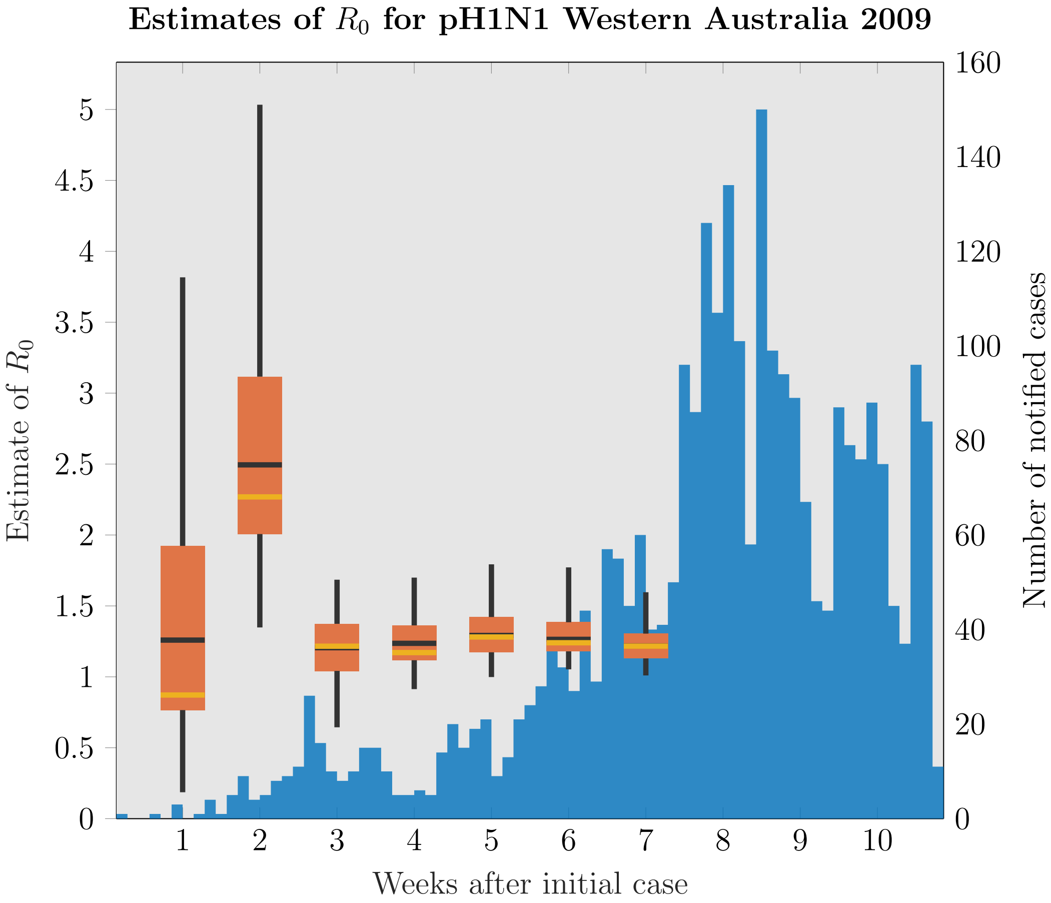


Figure 5.13: Daily incidence count and estimated basic reproductive number for an outbreak of A(H1N1)pdm09 in Western Australia using the conditioned full SEIR model with partial observations. The yellow line shown on the box plots indicates the MAPE of R_0

In this section we consider an outbreak of the Ebola hemorrhagic fever which occurred in Zaire in 1976. In Figure 5.14, we estimate R_0 from the first four weeks of the outbreak, before control measures began to take effect on the spread of the disease. A recent estimate of R_0 from the early stages of this outbreak is 1.34 [Camacho et al., 2014]. It can be seen that the average estimate of R_0 increases steadily during the first three weeks of the outbreak before decreasing between weeks three and four due to the temporary reduction in the growth of the observed incidence count. Our estimates of R_0 are slightly larger than those of Camacho et al. [2014].

5.3.3 Ebola hemorrhagic fever in Congo 1995

We now analyse an outbreak of Ebola which occurred in Congo in 1995. We restrict our attention to the first eight weeks of the outbreak, prior to any significant impact of control measures. A recent estimate of R_0 based on the initial stages of this outbreak is 1.83 with a reported standard deviation of 0.06 [Chowell et al., 2004]. It can be seen that our average estimates of R_0 are consistent for the first few weeks of the outbreak before gradually increasing after the outbreak has become established. In Figure 5.15, it can be seen that our estimates of R_0 appear to agree reasonably well with Chowell et al. [2004].

5.3.4 Pneumonic Plague in Madagascar 2017

Pneumonic Plague is a very severe bacterial infection of the lungs which is invariably fatal, if left untreated. Plague is transmitted between animals and humans by the bite of an infected flea, and between humans by physical contact with infectious bodily fluids or contaminated materials or the inhalation of respiratory droplets/small particles from a patient with pneumonic plague.

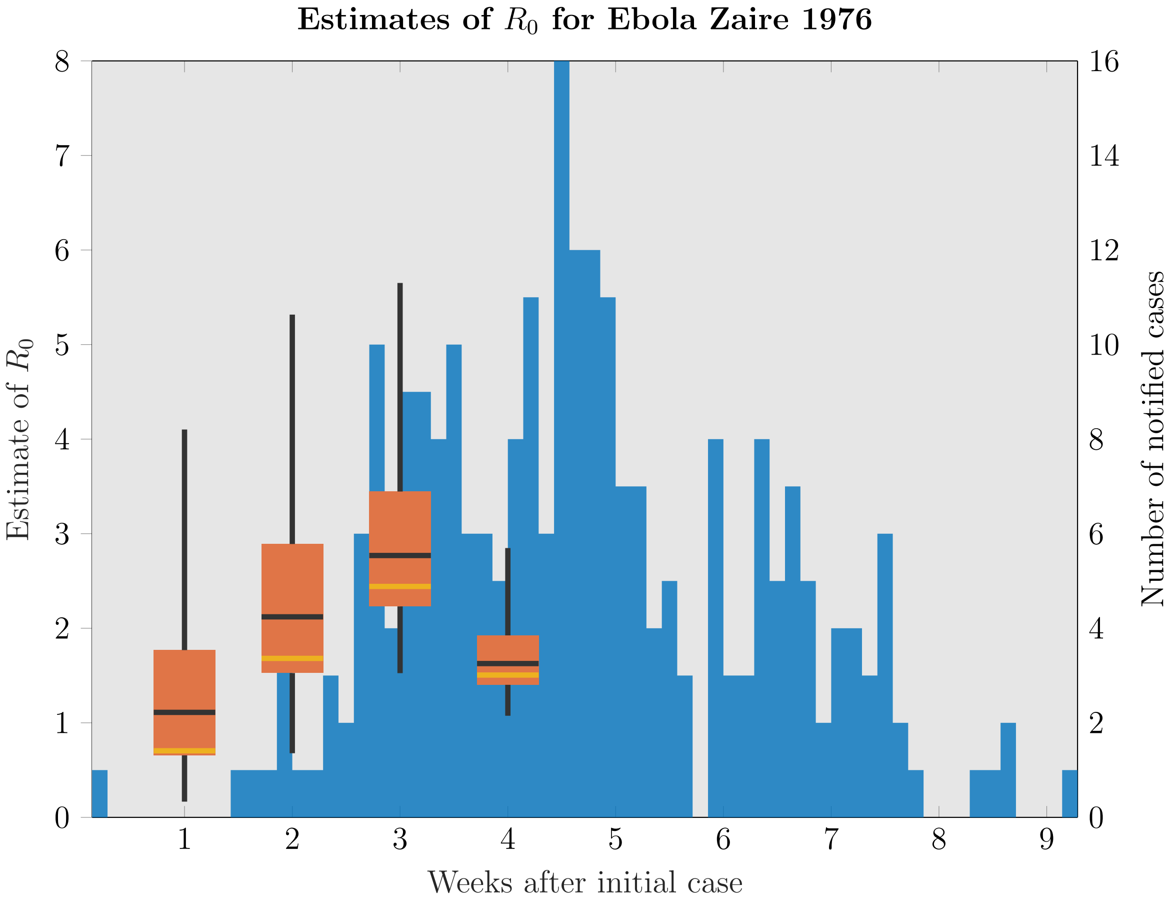


Figure 5.14: Daily incidence count and estimated basic reproductive number for an outbreak of the Ebola virus from 1976 in Zaire using the conditioned full SEIR model with partial observations. The yellow line shown on the box plots indicates the MAPE of R_0

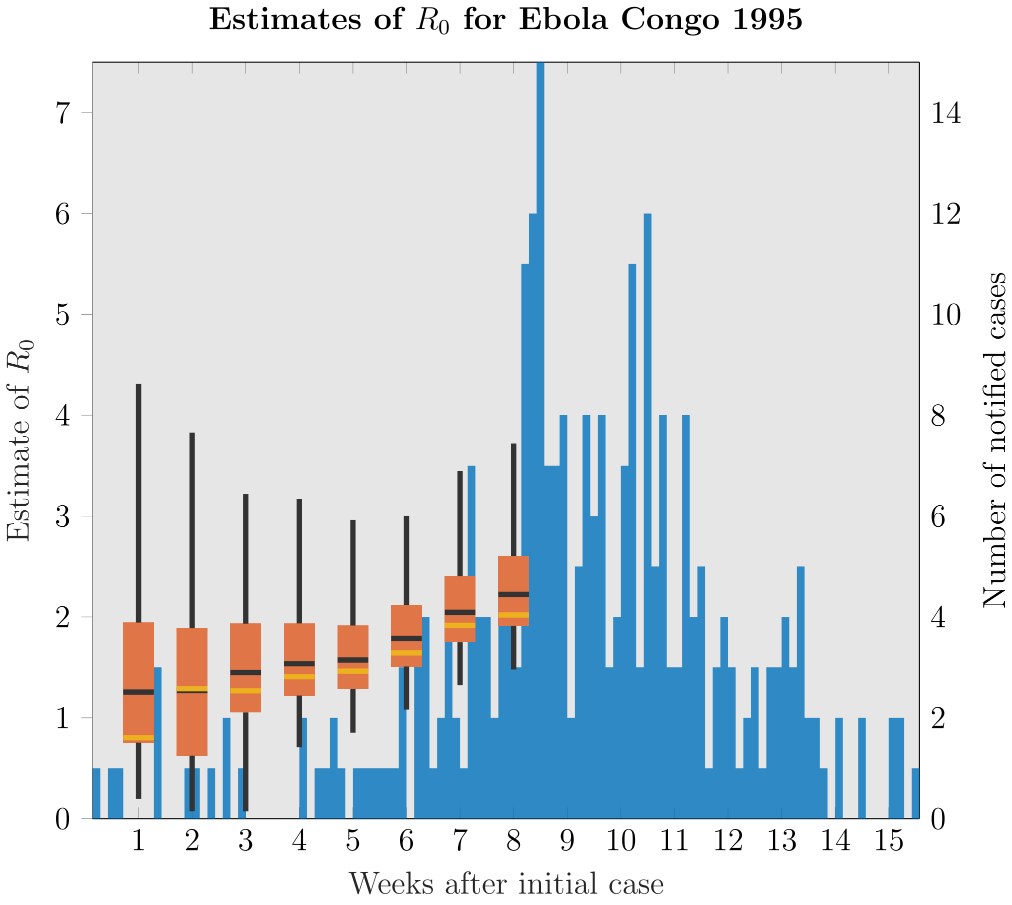


Figure 5.15: Daily incidence count and estimated basic reproductive number for an outbreak of the Ebola virus from 1995 in Congo using the conditioned full SEIR model with partial observations. The yellow line shown on the box plots indicates the MAPE of R_0

Individuals infected with the plague usually develop acute febrile disease along side other non-specific symptoms such as head and body aches, and weakness, vomiting and nausea. Antibiotic treatment is highly effective if the infection is caught within the first 24 hours [WHO, 2017b]

In this section we consider an outbreak of Pneumonic Plague which occurred in Madagascar in 2017 [WHO, 2017c]. We estimate R_0 from the first eight weeks of the outbreak, prior to a the implementation of concerted control measures. A recent estimate of R_0 for this outbreak is 1.73 [Tsuzuki et al., 2017]. Our estimates of the basic reproductive number are shown in Figure 5.16. Based on the first three weeks of the outbreak, our model suggests that R_0 is only slightly higher than one. As the number of incidences increases, our estimated value of R_0 increases. Our estimates of R_0 are sensitive to the sudden spikes in incidences occurring in weeks four and seven.

5.4 Discussion

In this chapter, we have introduced an extension of the conditioned hybrid diffusion approach presented in Chapter 4. We have done so by considering the partially-observed SEIR CTMC, which is more appropriate than the SIR CTMC for modelling the early stages of an outbreak due to its inclusion of an exposed period and imperfect observations. In extending the hybrid diffusion approach of Chapter 4, we constructed a dynamic state space truncation rule which is utilised during the initial stages of the outbreak. In a simulation study where we looked at the first five weeks of an outbreak with influenza-like dynamics, we demonstrated that conditioning was an effective means of reducing bias in estimates of the basic reproductive number. A similar outcome was observed for a number of other parameters. Furthermore, the

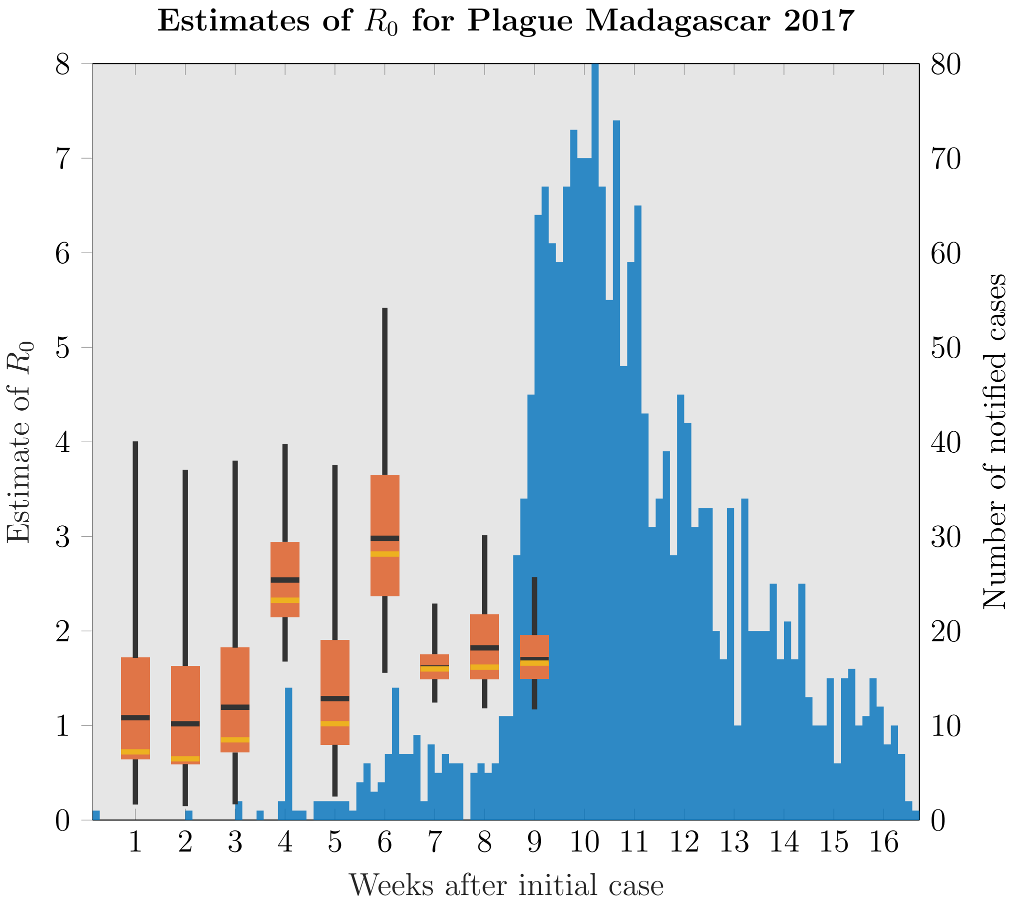


Figure 5.16: Daily incidence count and estimated basic reproductive number for an outbreak of Plague from 2017 in Madagascar using the conditioned full SEIR model with partial observations. The yellow line shown on the box plots indicates the MAPE of R_0

hybrid diffusion approach enabled us to consider outbreaks which were too large for consideration under the standard partially-observed SEIR CTMC. We demonstrated the utility of our approach by using it to estimate the basic reproductive number from a number of real outbreaks and found that our resulting estimates were similar to previous results.

Although the dynamic state space truncation rule utilised in this chapter provided a simple means of ensuring the computational-feasibility of our model, it did not take advantage of existing approaches which may have been more efficient [Sunkara and Hegland, 2010]. It follows that future work could focus on improving the computational methodology for dealing with the CTMC dynamics of the process by utilising a highly-efficient state space truncation algorithm for estimating the basic reproductive number.

There are a number of features which make the SEIR CTMC an implausible model, such as the distribution of the exposed period and infectious period, population heterogeneity (such as spatial variation or age-specific or household clustering of contacts), time-inhomogenous infectivity and case-reporting rates, imported infectious cases, pre-existing immunity and the population's response to the outbreak. Thus, the model presented here is by no means the most accurate for modelling the early stages of an outbreak. However, the important thing is that we have demonstrated that the hybrid methodology presented in Chapter 5 is a flexible approach which may be adapted to a range of scenarios to provide improvements in computational-tractability. For instance, a typical extension of the SEIR CTMC to account for non-exponential infectious periods would require including an additional infectious compartment in the model. Since the resulting model is still in the CTMC framework, the hybrid methodology presented in this thesis may be applied to the resulting model without difficulty.

Chapter 6

Conclusion

The aim of this thesis was to investigate the utility of hybrid methodology for modelling the outbreak of an infectious disease: based on the notion that the early stages of an outbreak are faithfully represented by CTMC dynamics, and an efficient and suitably accurate representation of an established outbreak is provided by either the fluid or diffusion large-population approximation. We presented a hybrid approach towards modelling outbreaks of infectious diseases whereby the outbreak is modelled by CTMC dynamics while the number of infectious individuals is low and a large-population approximation otherwise. We utilised this methodology for computing the distribution of key quantities of an outbreak and calibrating models describing the spread of a disease to case incidence data from the early stages of an outbreak. The following discussion provides a brief summary of the research presented in this thesis, the main results that have been established and their implications, and some directions for future research in this field.

6.1 Summary

In Chapter 3, we presented two hybrid models for computing the distribution of the final size of the outbreak and the distribution of the duration of the outbreak in the framework of the SIR CTMC. These hybrid models utilised the dynamics of the SIR CTMC while the number of infectious individuals was low and a large-population approximation of the SIR CTMC otherwise. The so-called hybrid fluid model and hybrid diffusion model were named after the large population approximation which they utilised, namely the fluid approximation [Kurtz, 1970] and the diffusion approximation [Kurtz, 1971]. We found that the hybrid fluid model provided an accurate representation of the distribution of the duration of the outbreak and the hybrid diffusion model provided an accurate representation of the distribution of the final size of the outbreak. The computational cost associated with computing these distributions from the hybrid models was significantly less than the computational cost associated with computing them directly from the SIR CTMC. Thus, it was established that our hybrid methodology provides an appropriately accurate and computationally-efficient means of computing key quantities of an outbreak. The contents of this chapter were published in Rebuli et al. [2016].

In Chapter 4, we considered estimating the basic reproductive number of an outbreak, a key quantity often used by public health authorities in planning their response to an outbreak. In the framework of the SIR CTMC, we demonstrated that the estimated basic reproductive number is positively biased if the model does not account for the event that the outbreak establishes an appreciable chain of transmissions. Under certain conditions, we showed that the average bias in estimates of R_0 may be decreased by 0.3 by conditioning the SIR CTMC on the event that the outbreak becomes established. Utilising the hybrid methodology from Chapter 3, we presented a hybrid diffusion approach for estimating the basic reproductive number using case incidence data from the early stages of an outbreak, in the framework of the conditioned SIR CTMC. This approach enabled us to consider an outbreak of A(H1N1)pdm09 which would have been computationally-intractable in the framework of the SIR CTMC. The significance of this work was to establish a method for reducing bias in estimates of the basic reproductive number which are based on case incidence data from the initial stages of an outbreak. Furthermore, the hybrid diffusion approach provided a means of applying this methodology to large outbreaks. The contents of this chapter were published in the paper Rebuli et al. [2018].

In Chapter 5, we presented a substantial extension to the methodology presented in Chapter 4. We considered a partially-observed SEIR CTMC, a generalisation of the SIR CTMC more appropriate for modelling the early stages of an outbreak due to its inclusion of an exposed compartment and imperfect observations. We applied the methodology presented in Chapter 4 to provide a unconditioned and conditioned hybrid diffusion approach to estimating the basic reproductive number in the framework of the partially-observed SEIR CTMC by utilising a dynamic state space truncation rule during the initial SEIR CTMC dynamics. In a simulation study considering the first five weeks of an outbreak with influenza-like dynamics, we demonstrated a similar outcome to those observed in Chapter 4. Namely, conditioning the model on establishing an appreciable chain of transmissions reduced bias in estimates of the basic reproductive number and the hybrid diffusion approach enabled us to consider larger outbreaks than would have been feasible in the framework of the partially-observed SEIR CTMC. We then demonstrated the utility of our model by using it to estimate the basic reproductive number from a number of real outbreaks. The significance of this work was to establish that the conditioned hybrid methodology of Chapter 4 can be generalised to complex CTMC models of the spread of disease with little difficulty, to provide real insights to the transmission dynamics of infectious diseases. The methodology presented here has been submitted for publication.

6.2 Future research

Although the hybrid models presented in Chapter 3 provided sufficiently accurate approximations of the distribution of the duration of the outbreak and the distribution of the final size of the outbreak, we observed that the approximation broke down when the dynamics of the large-population approximations came close to the $S = 0$ absorbing boundary. It was noted that this problem could be amended by placing a threshold on the number of susceptible individuals such that the process switches from the dynamics of the large-population approximation to the dynamics of the CTMC if either the number of susceptible individuals or the number infectious individuals drops below its appropriate threshold. A model of this nature would be similar to the hybrid diffusion model presented by Safta et al. [2015], whereby each compartment utilises CTMC dynamics while its population is low and diffusion dynamics otherwise. This allows some states of the process to have CTMC dynamics for some compartments and large-population dynamics for the other compartments. Hybrid models of this nature have not received much attention outside of modelling chemical reactions and may prove to be a useful extension to the hybrid diffusion methodology presented here in computing the distribution of key quantities of an outbreak or in estimating the basic reproductive number.

The mechanism by which the hybrid diffusion model presented in Chapter 4 switches from CTMC dynamics to diffusion dynamics does not guarantee that the diffusion approximation will provide a suitably-accurate representation of the underlying CTMC dynamics immediately after the model changes dynamics. Our dynamic state space truncation rule presented in Chapter 5 was an effective means of accounting for this, but our approach does not take advantage of existing state space truncation algorithms which may be more efficient, for example Sunkara and Hegland [2010]. It follows that future research could focus on developing highly-efficient routines for estimating the basic reproductive number by utilising an optimal state space truncation algorithm. Further research in this direction could allow the methodology presented in Chapter 5 to be applied to more complex CTMC models and the development of a general-use software package for estimating the basic reproductive number.

The SEIR CTMC is often considered one of the simplest CTMC models acceptable for modelling real outbreaks. However, there are a number of features which make it somewhat unreliable, such as the distribution of the exposed period and infectious period, population heterogeneity (such as spatial variation or age-specific or household clustering of contacts), time-inhomogenous infectivity and case-reporting rates, imported infectious cases, pre-existing immunity and the population's response to the outbreak. However, one of the most useful features of the hybrid methodology presented in this thesis is its flexibility. For instance, a typical extension of the SEIR CTMC to account for non-exponential infectious periods would require including additional infectious compartments in the model. Since the resulting model is still in a CTMC framework, the hybrid methodology presented in this thesis may be applied to the resulting model and utilised for computing key quantities or estimating the basic reproductive number. It follows that the hybrid methodology presented in this thesis is a useful tool for improving computational-tractability of models which are based in a CTMC framework.

An interesting field where the hybrid methodology presented in this thesis may prove beneficial is in modelling between-host disease transmission while accounting for within-host pathogen dynamics. Within-host dynamics are complex and typically involve interactions between large populations of biological agents, making it computationally-infeasible to model the population of the invasive pathogens using a CTMC framework. However, it is understood that pathogen-colonisation begins when a small number of pathogens enter a naive host, suggesting that an important feature of within-host pathogen dynamics is the probability of initial fade out. This provides an ideal application of the hybrid methodology presented in this thesis where an individual's within-host pathogen dynamics could be modelled by a hybrid approach. The significance of this work would be to help improve our understand of how within-host pathogen dynamics influence the transmissibility of a disease which could provide important insights for disease prevention strategies.

The hybrid methodology presented in this thesis provides a straightforward approach to reducing the computational demands of a CTMC model in exchange for a minor decrease in accuracy. Furthermore, our conditioning approach to estimating the basic reproductive is effective at decreasing bias and our hybrid approach is effective at improving computational tractability. This methodology is straightforward and may be applied to a wide range of epidemiological models and even models outside of epidemiology.

Bibliography

T. G. Kurtz. Solutions of ordinary differential equations as limits of pure jump Markov processes. *J. Appl. Prob.*, 7(1):49–58, 1970. doi: 10.2307/3212147.

T. G. Kurtz. Limit theorems for sequences of jump Markov processes approximating ordinary differential processes. *J. Appl. Prob.*, 8(2):344–356, 1971. doi: 10.2307/3211904.

G. N. Mercer, K. Glass, and N. G. Becker. Effective reproduction numbers are commonly overestimated early in a disease outbreak. *Stat. Med.*, 30(9): 984–994, 2011. doi: 10.1002/sim.4174.

Wasima N. Rida. Asymptotic properties of some estimators for the infection rate in the general stochastic epidemic model. *J. Royal Stat. Soc. B*, 53(1): 269–283, 1991. doi: 10.2307/2345741.

A.D. Barbour. The duration of the closed stochastic epidemic. *Biometrika*, 62(2):477–482, 1975. doi: 10.1093/biomet/62.2.477.

Gianpaolo Scalia-Tomba. Asymptotic final-size distribution for some chain-binomial processes. *Adv. Appl. Prob.*, 17(3):477–495, 1985. doi: 10.2307/1427116.

Nicolas P. Rebuli, N. G. Bean, and J. V. Ross. Hybrid Markov chain models of SIR disease dynamics. *J. Math. Biol.*, 74(3):521–541, 2016. doi: 10.1007/s00285-016-1085-2.

Nicolas P. Rebuli, N.G. Bean, and J.V. Ross. Estimating the basic reproductive number during the early stages of an emerging epidemic. *Theor. Popul. Biol.*, 119:26–36, 2018. doi: 10.1016/j.tpb.2017.10.004.

Willy Feller. On the time distribution of so-called random events. *Phys. Rev.*, 57(10):906–908, 1940. doi: 10.1103/PhysRev.57.906.

Tosio Kato. On the semi-groups generated by Kolmogoroff's differential equations. *J. Math. Soc. Jpn.*, 6(1):1–15, 1954. doi: 10.2969/jmsj/00610001.

W. A. O'N. Waugh. Conditioned Markov processes. *Biometrika*, 45(1):241–250, 1958. doi: 10.1093/biomet/45.1-2.241.

Cleve Moler and Van Loan Charles. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Rev.*, 45(1):3–49, 2003. doi: 10.1137/S00361445024180.

Garrett Jenkinson and John Goutsias. Numerical integration of the master equation in some models of stochastic epidemiology. *PLOS ONE*, 7(5):1–9, 2012. doi: 10.1371/journal.pone.0036160.

Andrew D. Barbour. The principle of the diffusion of arbitrary constants. *Adv. App. Prob.*, 9(3):519–541, 1972. doi: 10.2307/3212323.

Andrew D. Barbour. On a functional central limit theorem for Markov population processes. *Adv. Appl. Prob.*, 6(1):21–39, 1974. doi: 10.2307/1426205.

Thomas G. Kurtz. *Stochastic Systems: Modeling, Identification and Optimization*, chapter Limit theorems and diffusion approximations for density dependent Markov chains, pages 67–78. Springer Berlin Heidelberg, 1976. doi: 10.1007/BFb0120765.

Andrew D. Barbour. Quasi-stationary distributions in Markov population processes. *Adv. Appl. Prob.*, 8(2):296–314, 1976. doi: 10.2307/1425906.

P. K. Pollett and A. Vassallo. Diffusion approximations for some simple chemical reaction schemes. *Adv. Appl. Probab.*, 24(4):875–893, 1992. doi: 10.2307/1427717.

P. K. Pollett. On a model for interference between searching insect parasites. *J. Austral. Math. Soc. Ser. B*, 32(2):133–150, 1990. doi: 10.1017/S0334270000008390.

Stewart N. Ethier and Thomas G. Kurtz. *Markov Processes: Characterisation and Convergence*. John Wiley and Sons, Inc., New Jersey, 2008. ISBN 9780470316658. doi: 10.1002/9780470316658.

W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. A*, 138(834):55–83, 1927. doi: 10.1098/rspa.1927.0118.

M. S. Bartlett. Some evolutionary stochastic processes. *J. Royal Statist. Soc. B*, 11(2):211–229, 1949. doi: 10.2307/2984077. URL http://www.jstor.org/stable/2984077.

Klaus Dietz. Epidemics and rumours: A survey. *J. Royal Soc. A*, 130(4): 505–528, 1967. doi: 10.2307/2982521.

Norman T. J. Bailey. A simple stochastic epidemic. *Biometrika*, 37(3-4): 193–202, 1950. doi: 10.1093/biomet/37.3-4.193.

Norman T. J. Bailey. *The Mathematical Theory of Epidemics*. Griffin, 1957.

Matt J. Keeling, Howard B. Wilson, and Steve W. Pacala. Reinterpreting space, time lags, and functional responses in ecological models. *Science*, 290(5497):1758–1761, 2000. doi: 10.1126/science.290.5497.1758.

P. K. Pollett and V. T. Stefanov. Path integrals for continuous-time Markov chains. *J. Appl. Probab.*, 39(4):901–904, 2002. doi: 10.2307/3216013.

Andrew J. Black and J. V. Ross. Computation of epidemic final size distributions. *J. Theoret. Biol.*, 367:159–165, 2015. doi: 10.1016/j.jtbi.2014.11.029.

Hakan Andersson and Tom Britton. *Stochastic Epidemic Models and Their Statistical Analysis*, volume 1 of *Lecture Notes in Statistics*. Springer-Verlag New York, 2000. doi: 10.1007/978-1-4612-1158-7.

Frank Ball and Peter Donnelly. Strong approximations for epidemic models. *Stoch. Proc. Appl.*, 55(1):1–21, 1995. doi: 10.1016/0304-4149(94)00034-Q.

Stuart Coles. *An Introduction to Statistical Modeling of Extreme Values*, volume 1 of *Springer Series in Statistics*. Springer-Verlag London, 2000. doi: 10.1007/978-1-4471-3675-0.

Saralees Nadarajah. Linear combination of Gumbel random variables. *Stoch. Env. Res. Risk. A*, 21(3):283–286, 2007. doi: 10.1007/s00477-006-0063-4.

M. Von Bahr and A. Martin-Lof. Threshold limit theorems for some epidemic processes. *Adv. Appl. Prob.*, 12(1980):319–349, 1980.

F. Ball. The threshold behaviour of epidemic models. *J. Appl. Prob.*, 20(2): 227–241, 1983. doi: 10.1017/S0021900200023391.

R. K. Watson. A useful time scale transformation for the standard epidemic model. *J. Appl. Prob.*, 17(2):324–332, 1980a. doi: 10.1017/S002190020004715XP.

R. K. Watson. On the size distribution for some epidemic models. *J. Appl. Prob.*, 17(4):912–921, 1980b.

Ray Watson. An application of a martingale central limit theorem to the standard epidemic model. *Stoc. Proc. Appl.*, 11(1):79–89, 1981. doi: 10.1016/0304-4149(81)90023-5.

A. Martin-Lof. *Stochastic Processes in Epidemic Theory*, volume 86 of *Lecture Notes in Biomathematics*, chapter Threshold limit theorems in the theory of rumors, snowball sampling and epidemics, pages 184–188. Springer Berlin Heidelberg, 1990. doi: 10.1007/978-3-662-10067-7_17.

Claude Lefèvre. *Stochastic Processes in Epidemic Theory*, volume 86 of *Lecture Notes in Biomathematics*, chapter Stochastic Epidemic Models for SIR Infectious Diseases: a Brief Survey of the Recent General Theory, pages 1–12. Springer Berlin Heidelberg, 1990. doi: 10.1007/978-3-662-10067-7_1.

Frank Ball and Peter Neal. *Workshop on Branching Processes and Their Applications*, volume 197 of *Lecture Notes in Statistics*, chapter Applications of branching processes to the final size of SIR epidemics, pages 207–223. Springer Berlin Heidelberg, 2010. ISBN 9783642111563. doi: 10.1007/978-3-642-11156-3_15.

D. A. Sprott. *Statistical Inference in Science*, volume 1 of *Springer Series in Statistics*. Springer-Verlag New York, 2000. ISBN 9780387950198. doi: 10.1007/b98955.

J. V. Ross, T. Taimre, and P. K. Pollett. On parameter estimation in population models. *Theor. Popul. Biol.*, 70(4):498–510, 2006. doi: 10.1016/j.tpb.2006.08.001.

J. V. Ross, D. E. Pagendam, and P. K. Pollett. On parameter estimation in population models II: Multi-dimensional processes and transient dynamics. *Theor. Popul. Biol.*, 75(2):123–132, 2009. doi: 10.1016/j.tpb.2008.12.002.

J. V. Ross. On parameter estimation in population models III: Time-inhomogeneous processes and observation error. *Theor. Popul. Biol.*, 82(1): 1–17, 2012. doi: 10.1016/j.tpb.2012.03.001.

George Casella and Roger L. Berger. *Statistical Inference*, volume 2 of *Duxbury Advanced Skills*. Australian Thomson learning, 2002. ISBN 9780534243128.

W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*, volume 1 of *Chapman & Hall/CRC Interdisciplinary Statistics*. Chapman & Francis, 1996. ISBN 9780412055515.

Aiddhartha Chibb and Edward Greenberg. Understanding the Metropolis-Hastings algorithm. *Am. Stat.*, 49(4):327–335, 1995. doi: 10.2307/2684568.

M. S. Bartlett. Deterministic and stochastic models for recurrent epidemics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 4. University of California Press, 1956.

D. A. Rand and H. B. Wilson. Chaotic stochasticity: A ubiquitous source of unpredictability in epidemics. *Proc. R. Soc. B*, 246(1316):179–184, 1991. doi: 10.1098/rspb.1991.0142.

Gordon A. Fox. Life history evolution and demographic stochasticity. *Evol. Ecol.*, 7(1):1–14, 1993. doi: 10.1007/BF01237731.

B. T. Grenfell, K. Wilson, B. F. Finkenstadt, T. N. Coulson, S. Murray, S. D. Albon, J. M. Pemberton, T. H. Clutton-Brock, and M. J. Crawley. Noise and determinism in synchronized sheep dynamics. *Nature*, 394(1):674–677, 1998. doi: 10.1038/29291.

B. Spagnolo, A. Fiasconaro, and D. Valenti. Noise induced phenomena in Lotka-Volterra systems. *Fluct. Noise Lett.*, 3(2):177–185, 2003. doi: 10.1142/S0219477503001245.

Tim Coulson, Pejman Rohani, and Mercedes Pascual. Skeletons, noise and population growth: The end of an old debate? *Trends Ecol. Evol.*, 19(7): 359–364, 2004. doi: 10.1016/j.tree.2004.05.008.

N. G. Van Kampen. A power series expansion of the master equation. *Ca. J. Phys.*, 39(4):551–567, 1961. doi: 10.1139/p61-056.

N. G. Van Kampen. *The Expansion of the Master Equation*, pages 245–309. John Wiley & Sons, Inc., 2007a. doi: 10.1002/9780470142530.ch5.

K. J. McNeil and D. F. Walls. A master equation approach to nonlinear optics. *J. Phys. A*, 9(5):617–631, 1974. doi: 10.1088/0305-4470/7/5/012.

Ryogo Kubo, Kazuhiro Matsuo, and Kazuo Kitahara. Fluctuation and relaxation of macrovariables. *J. Stat. Phys.*, 9(1):51–96, 1973. doi: 10.1007/BF01016797.

Paul Sjöberg, Per Lötstedt, and Johan Elf. Fokker–Planck approximation of the master equation in molecular biology. *Computing and Visualization in Science*, 12(1):37–50, 2009. doi: 10.1007/s00791-006-0045-6.

N. G. Van Kampen. *Stochastic processes in physics and chemistry.* Elsevier, 3 edition, 2007b. doi: 10.1137/1.9780898719734.

Claire Guerrier and David Holcman. Hybrid Markov-mass action law model for cell activation by rare binding events: Application to calcium induced vesicular release at neuronal synapses. *Sci. Rep.*, 6(1):1–10, 2016. doi: 10.1038/srep35506.

A. Ganguly, D. Altintan, and H. Koeppl. Jump-diffusion approximation of stochastic reaction dynamics: Error bounds and algorithms. *Mult. Mod. Simul.*, 13(5):1390–1419, 2015. doi: 10.1137/140983471.

A. Duncan, R. Erban, and K. Zygalakis. Hybrid framework for the simulation of stochastic chemical kinetics. *J. Comp. Phys*, 326:398419, 2016. doi: 10.1016/j.jcp.2016.08.034.

A. Angius, G. Balbo, M. Beccuti, E. Bibbona, A. Horvath, and R. Sirovich. Approximate analysis of biological systems by hybrid switching jump diffusion. *Theor. Comp. Sci.*, 587:4972, 2015. doi: 10.1016/j.tcs.2015.03.015.

K. Vasudeva and U. S. Bhalla. Adaptive stochastic-deterministic chemical kinetic simulations. *Bioinformatics*, 20:7884, 2003. doi: 10.1093/bioinformatics/btg376.

K. Takahashi, K. Kaizu, B. Hu, and M. Tomita. A multi-algorithm, multi-timescale method for cell simulation. *Bioinformatics*, 20:538546, 2004. doi: 10.1093/bioinformatics/btg442.

A. Hellander and P. Lötstedt. Hybrid method for the chemical master equation. *J. Comp. Phys*, 227:100–122, 2007. doi: 10.1016/j.jcp.2007.07.020.

B. Hepp, A. Gupta, and M. Khammash. Adaptive hybrid simulations for multiscale stochastic reaction networks. *J. Chem. Phys.*, 142(3):34118–34134, 2015. doi: 10.1063/1.4905196.

Cosmin Safta, Khachik Sargsyan, Bert Debusschere, and Habib N. Najm. Hybrid discrete/continuum algorithms for stochastic reaction networks. *J. Comput. Phys.*, 281(10):177–198, 2015. doi: 10.1016/j.jcp.2014.10.026.

David G. Kendall. Mathematical models of the spread of infection. *Mathe. Comp. Sci. Biol. Med.*, 213, 1965.

Igor Sazonov, Mark Kelbert, and Michael B. Gravenor. A two-stage model for the SIR outbreak: Accounting for the discrete and stochastic nature of the epidemic at the initial contamination stage. *Math. Biosci.*, 234(2): 108–117, 2011. doi: 10.1016/j.mbs.2011.09.002.

I. Sazonov, D. Grebennikov, M. Kelbert, and B. Bocharov. Modelling stochastic and deterministic behaviours in virus infection dynamics. *Math. Model. Nat. Phenom.*, 12(5):63–77, 2017. doi: 10.1051/mmnp/20171250.

A. V. Nagaev and A. N. Startsev. The asymptotic analysis of a stochastic model of an epidemic. *Theory of Probability and Its Applications*, 15(1): 98–107, 1970. doi: 10.1137/1115007.

Ccile Viboud, Lone Simonsen, and Gerardo Chowell. A generalized-growth model to characterize the early ascending phase of infectious disease outbreaks. *Epidemics*, 15(4):27–37, 2016. doi: 10.1016/j.epidem.2016.01.002.

L. M. A. Bettencourt and Ruy M. Ribeiro. Real time Bayesian estimation of the epidemic potential of emerging infectious diseases. *PLOS ONE*, 3(5): 1–9, 2008. doi: 10.1371/journal.pone.0002185.

K. Glass, G. N. Mercer, H. Nishiura, E. S. McBryde, and N. G. Becker. Estimating reproduction numbers for adults and children from case data. *J. Royal Soc. Interface*, 8(62):1248–1259, 2011. doi: 10.1098/rsif.2010.0679.

Hiroshi Nishiura, Gerardo Chowell, Muntaser Safan, and Carlos Castillo-Chavez. Pros and cons of estimating the reproduction number from the early epidemic growth rate of influenza A (H1N1) 2009. *Theor. Biol. Med. Model.*, 7(1):1–13, 2010. doi: 10.1186/1742-4682-7-1.

T Vega, J Lozano, T Meerhoff, R Snacken, and J Mott. Influenza surveillance in Europe: establishing epidemic thresholds by the moving epidemic method. *Influenza Other Respir. Viruses*, 7:546–558, 2013. doi: 10.1111/j.1750-2659.2012.00422.x.

L F White and M Pagano. A likelihood-based method for real-time estimation of the serial interval and reproductive number of an epidemic. *Stat. Med.*, 27(16):2999–3016, 2007. doi: 10.1002/sim.3136.

L. Simonsen, M. J. Clarke, G. D. Williamson, D. F. Stroup, N. H. Arden, and L. B. Schonberger. The impact of influenza epidemics on mortality: introducing a severity index. *Am. J. Public Health*, 87(12):1944–1950, 1997. doi: 10.2105/AJPH.87.12.1944.

M. I. Meltzer, N. J. Cox, and K. Fukuda. The economic impact of pandemic influenza in the United States: priorities for intervention. *Emerg. Infect. Dis.*, 5(5):659–671, 1999. doi: 10.3201/eid0505.990507.

Stanley M. Lemon, Margaret A. Hamburg, P. Frederick Sparling, Eileen R. Choffnes, and Alison Mack. *Ethical and Legal Considerations in Mitigating Pandemic Disease: Workshop Summary*, chapter Strategies for Disease Containment, pages 76–153. The National Academic Press, 2007. doi: 10.17226/11917.

Gerardo Chowell, Cécile Viboud, Xiaohong Wang, Stefano M. Bertozzi, and Mark A. Miller. Adaptive vaccination strategies to mitigate pandemic influenza: Mexico as a case study. *PLOS ONE*, 4(12):1–9, 2009. doi: 10.1371/journal.pone.0008164.

Joseph T Wu, Steven Riley, Christophe Fraser, and Gabriel M Leung. Reducing the impact of the next influenza pandemic using household-based public health interventions. *PLOS Medicine*, 3(9):1–9, 2006. doi: 10.1371/journal.pmed.0030361.

Simon Cauchemez, Pierre-Yves Boëlle, Christl A. Donnelly, Neil M. Ferguson, Guy Thomas, Gabriel M. Leung, Anthony J. Hedley, Roy M. Anderson, and Alain-Jacques Valleron. Real-time estimates in early detection of SARS. *Emerg. Infect. Diseas.*, 12(1):110–113, 2006.

K. Glass, N. Becker, and M. Clements. Predicting case numbers during infectious disease outbreaks when some cases are undiagnosed. *Statist. Med.*, 26(5):171–183, 2007. doi: 10.1002/sim.2523.

Mark E. J. Woolhouse, Andrew Rambaut, and Paul Kellam. Lessons from ebola: Improving infectious disease surveillance to inform outbreak management. *Science*, 7(307):307–313, 2015. doi: 10.1126/scitranslmed.aab0191.

May Robert M. Galvani, Alison P. Dimensions of superspreading. *Nature*, 1 (1), 2016. doi: 10.1038/438293a.

Marc Lipsitch, Christl A. Donnelly, Christophe Fraser, Isobel M. Blake, Anne Cori, Ilaria Dorigatti, Neil M. Ferguson, Tini Garske, Harriet L. Mills, Steven Riley, Maria D. Van Kerkhove, and Miguel A. Hernn. Potential biases in estimating absolute and relative case-fatality risks during outbreaks. *PLOS Neglected Tropical Diseases*, 9(7):1–16, 2015. doi: 10.1371/journal.pntd.0003846.

Charly Favier, Delphine Schmit, Christine D.M Müller-Graf, Bernard Cazelles, Nicolas Degallier, Bernard Mondet, and Marc A Dubois. Influence of spatial heterogeneity on an emerging infectious disease: the case of dengue epidemics. *R. Soc. B*, 272(1568):1171–1177, 2005. doi: 10.1098/rspb.2004.3020.

Matt J. Keeling, Mark E. J. Woolhouse, Darren J. Shaw, Louise Matthews, Margo Chase-Topping, Dan T. Haydon, Stephen J. Cornell, Jens Kappey, John Wilesmith, and Bryan T. Grenfell. Dynamics of the 2001 uk foot and mouth epidemic: Stochastic dispersal in a heterogeneous landscape. *Science*, 294(5543):813–817, 2001. doi: 10.1126/science.1065973.

Michael George Roberts and Hiroshi Nishiura. Early estimation of the reproduction number in the presence of imported cases: Pandemic influenza H1N1-2009 in New Zealand. *PLOS ONE*, 6(5):1–9, 2011. doi: 10.1371/journal.pone.0017835.

Matthew Hartfield and Samuel Alizon. Introducing the outbreak threshold in epidemiology. *PLOS Pathogens*, 9(6):1–4, 2013. doi: 10.1371/journal.ppat.1003277.

Richard D. Smith. Responding to global infectious disease outbreaks: Lessons from SARS on the role of risk perception, communication and management. *Soc. Sci. Med.*, 63(12):3113–3123, 2006. doi: 10.1016/j.socscimed.2006.08.004.

Andrew J. Black and Joshua V. Ross. Estimating a Markovian epidemic model using household serial interval data from the early phase of an epidemic. *PLOS ONE*, 8(8):1–8, 2013. doi: 10.1371/journal.pone.0073420.

Richard J. Boys and Philip R. Giles. Bayesian inference for stochastic epidemic models with time-inhomogeneous removal rates. *J. Math. Biol.*, 55(2):223–247, 2007. doi: 10.1007/s00285-007-0081-y.

Gerardo Chowell, Hiroshi Nishiura, and Luís M.A Bettencourt. Comparative estimation of the reproduction number for pandemic influenza from daily case notification data. *J. Royal Soc. Interface*, 4(12):155–166, 2007. doi: 10.1098/rsif.2006.0161.

Ben Cooper and Marc Lipsitch. The analysis of hospital infection data using hidden Markov models. *Biostatistics*, 5(2):223–238, 2004. doi: 10.1093/biostatistics/5.2.223.

Heath A. Kelly, Geoff N. Mercer, James E. Fielding, Gary K. Dowse, Kathryn Glass, Dale Carcione, Kristina A. Grant, Paul V. Effler, and Rosemary A. Lester. Pandemic (H1N1) 2009 influenza community transmission was established in one Australian state when the virus was first identified in North America. *PLOS ONE*, 5(1):1–9, 2010. doi: 10.1371/journal.pone.0011341.

E Pedroni, M Garca, V Espnola, A Guerrero, C Gonzlez, A Olea, M Calvo, B Martorell, M Winkler, M. V. Carrasco, J. A. Vergara, J Ulloa, A. M. Carrazana, O Mujica, J. E. Villarroel, M Labraa, M Vargas, P Gonzlez, L Cceres, C G Zamorano, R Momberg, G Muoz, J Rocco, V Bosque, A Gallardo, J Elgueta, and J Vega. Outbreak of 2009 pandemic influenza A(H1N1), Los Lagos, Chile, April-June 2009. *Euro. Surveill.*, 15(1), 2010.

Adrian J. Gibbs, John S. Armstrong, and Jean C. Downie. From where did the 2009 'swine-origin' influenza A virus (H1N1) emerge? *Virol. J.*, 6(1):207–218, 2009. doi: 10.1186/1743-422X-6-207.

F. S. Dawood, S. Jain, L. Finelli, M. W. Shaw, S. Lindstrom, R. J. Garten, L. V. Gubareva, X. Xu, C. B. Bridges, and T. M. Uyeki. Emergence of a novel swine-origin influenza A (H1N1) virus in humans. *N. Engl. J. Med.*, 25(25):2605–2615, 2009. doi: 10.1056/NEJMoa0903810.

Kathryn Glass, Heath Kelly, and Geoffry Norman Mercer. Pandemic influenza H1N1: Reconciling serosurvey data with estimates of the reproduction number. *Epidemiology*, 23(3):86–94, 2012. doi: 10.1097/EDE.0b013e31823a44a5.

Tarun S. Weeramanthri, Andrew G. Robertson, Gary K. Dowse, Paul V. Effler, Muriel G. Leclercq, Jeremy D. Burtenshaw, Susan J. Oldham, David W. Smith, Kathryn J. Gatti, and Gladstones Helen M. Response to pandemic (H1N1) 2009 influenza in australia lessons from a state health department perspective. *Aust. Health Rev.*, 4(34):477–486, 2010. doi: 10.1071/AH10901.

H. Nishiura, C. Castillo-Chavez, M. Safan, and G. Chowell. Transmission potential of the new influenza A(H1N1) virus and its age-specificity in Japan. *Euro. Surveill.*, 14(22), 2009a. doi: 10.2807/ese.14.22.19227-en.

H. Nishiura, N. Wilson, and M. G. Baker. Estimating the reproduction number of the novel influenza A virus (H1N1) in a southern hemisphere setting: preliminary estimate in New Zealand. *N. Z. Med. J.*, 122(1299): 73–77, 2009b.

C. V. Munayco, J. Gomez, V. A. Laguna-Torres, J. Arrasco, T. J. Kochel, V. Fiestas, J Garcia, J. Perez, I. Torres, F. Condori, H. Nishiura, and G. Chowell. Epidemiological and transmissibility analysis of influenza A(H1N1) virus in a southern hemisphere setting: Peru. *Euro. Surveill.*, 14 (32), 2009. doi: 10.2807/ese.14.32.19299-en.

Melen Leclerc, Thierry Doré, Christopher A. Gilligan, Philippe Lucas, and João A. N. Filipe. Estimating the delay between host infection and disease (incubation period) and assessing its significance to the epidemiology of plant diseases. *PLOS ONE*, 9(1):1–15, 2014. doi: 10.1371/journal.pone.0086568.

Phenyo E. Lekone and Bärbel F. Finkenstädt. Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study. *Biometrics*, 62(4):1170–1177, 2006. doi: 10.1111/j.1541-0420.2006.00609.x.

Shannon Collinson, Kamran Khan, and Jane M. Heffernan. The effects of media reports on disease spread and important public health measurements. *PLOS ONE*, 10(11):1–21, 2015. doi: 10.1371/journal.pone.0141423.

Elisabeth Anne-Sophie Mayrhuber, Thomas Niederkrotenthaler, and Ruth Kutalek. "we are survivors and not a virus:" content analysis of media reporting on Ebola survivors in Liberia. *PLOS Neglected Tropical Diseases*, 11(8):1–19, 2017. doi: 10.1371/journal.pntd.0005845.

Lewis Mitchell and Joshua V. Ross. A data-driven model for influenza transmission incorporating media effects. *Royal Soc. Open Sci.*, 3(10), 2016. doi: 10.1098/rsos.160481.

Frederik Verelst, Lander Willem, and Philippe Beutels. Behavioural change models for infectious disease transmission: a systematic review (2010-2015). *J. Royal Soc. Interface*, 13(125), 2016. doi: 10.1098/rsif.2016.0820.

Jonathan Fintzi, Xiang Cui, Jon Wakefield, and Vladimir N. Minin. Efficient data augmentation for fitting stochastic epidemic models to prevalence data. *J. Comp. Graph. Stat.*, 26(4):918–929, 2017. doi: 10.1080/10618600.2017.1328365.

Jacco Wallinga and Peter Teunis. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *Am. J. Epidemiol.*, 160(6):509–516, 2004. doi: 10.1093/aje/kwh255.

Laura F. White and Marcello Pagano. Reporting errors in infectious disease outbreaks, with applications to pandemic influenza A/H1N1. *Epidemiol. Perspect. Innov.*, 12(7), 2010. doi: 10.1186/1742-5573-7-12.

John D. Mathews, Christopher T. McCaw, Jodie McVernon, Emma S. McBryde, and James M. McCaw. A biological model for influenza transmission: Pandemic planning implications of asymptomatic infection and immunity. *PLOS ONE*, 2(11):1–6, 2007. doi: 10.1371/journal.pone.0001220.

Vikram Sunkara and Markus Hegland. An optimal finite state projection method. *Procedia Comput. Sci.*, 1(1):1579–1586, 2010. doi: 10.1016/j.procs.2010.04.177.

Brian Munsky and Mustafa Khammash. The finite state projection algorithm for the solution of the chemical master equation. *J. Chem. Phys.*, 124(4), 2006. doi: 10.1063/1.2145882.

Z. I. Botev and P. L'Ecuyer. Efficient probability estimation and simulation of the truncated multivariate student-t distribution. In *2015 Winter Simulation Conference (WSC)*, pages 380–391, 2015. doi: 10.1109/WSC.2015.7408180.

CDC. Clinical signs and symptoms of influenza, 2016. URL https://www.cdc.gov/flu/professionals/acip/clinical.htm.

WHO. Ebola virus disease fact sheet, 2017a. URL http://www.who.int/mediacentre/factsheets/fs103/en/.

Joel G. Breman, , Karl M. Johnson, , Guido van der Groen, , C. Brian Robbins, , Mark V. Szczeniowski, , Kalisa Ruti, , Patricia A. Webb, , Florian Meier, , David L. Heymann, and . A search for ebola virus in animals in the Democratic Republic of the Congo and Cameroon: Ecologic, virologic, and serologic surveys, 19791980. *J. Infect. Dis.*, 179(1):139–147, 1999. doi: 10.1086/514278.

A. Camacho, A. J. Kucharski, S. Funk, J. Breman, P. Piot, and W. J.” Edmunds. Potential for large outbreaks of Ebola virus disease. *Epidemics*, 9:70–78, 2014. doi: 10.1016/j.epidem.2014.09.003.

G. Chowell, N.W. Hengartner, C. Castillo-Chavez, P.W. Fenimore, and J.M. Hyman. The basic reproductive number of Ebola and the effects of public health measures: the cases of Congo and Uganda. *J. Theor. Bio.*, 229(1): 119 – 126, 2004. doi: 10.1016/j.jtbi.2004.03.006.

WHO. Plague fact sheet, 2017b. URL http://www.who.int/mediacentre/factsheets/fs267/en/.

WHO. Disease outbreak news: Plague - Madagascar, 2017c. URL http://www.who.int/csr/don/02-october-2017-plague-madagascar/en/.

Shinya Tsuzuki, Hyojung Lee, Fuminari Miura, Yat Hin Chan, Sung-mok Jung, Andrei R Akhmetzhanov, and Hiroshi Nishiura. Dynamics of the Pneumonic Plague epidemic in Madagascar, August to October 2017. *Euro. Surveill.*, 22(46), 2017. doi: 10.2807/1560-7917.ES.2017.22.46.17-00710.