Parameter Estimation in Individual-Level Models of Infectious Disease

by

Carolyn Augusta

A Thesis
presented to
The University of Guelph

In partial fulfilment of requirements
for the degree of
Master of Science
in
Mathematics and Statistics

Guelph, Ontario, Canada
© Carolyn Augusta, August, 2014
Spatial epidemic models are crucial to the prediction and control of infectious disease spread. Although the effect of a spatial parameter ($\beta$) on disease transmission may be more apparent in dense populations, knowledge of the spatial component of epidemic transmission is used to inform vaccination policies and culling procedures in many settings. Additionally, the susceptibility ($\alpha$) of the population at risk and the infectious period ($\gamma$) affect the speed of epidemic spread. We compare the parameter estimation techniques of maximum likelihood estimation and the gold standard Bayesian MCMC in terms of width of confidence and credible intervals for all three model parameters in two epidemic frameworks. We examine the effect of misspecification of the infectious period $\gamma$ on estimation of $\alpha$ and $\beta$. A grid population and a population generated by bivariate normal distributions are considered. We find that epidemics travel more quickly over the highly dense population regardless of the value of $\gamma$. 
# Table of Contents

1 Introduction 1

2 Spatial Individual Level Models (ILMs) of Infectious Disease Spread 4
   2.1 Maximum likelihood estimation in the context of ILMs . . . . . . . . . . . . 6
      2.1.1 Derivations of the Analytic First Derivatives . . . . . . . . . . . . 7
      2.1.2 Bayesian MCMC for ILMs . . . . . . . . . . . . . . . . . . . . . . 13

3 Simulation Studies 16
   3.1 Simulation Study of SI Model . . . . . . . . . . . . . . . . . . . . . . . . . . . 18
      3.1.1 SI model over the grid population . . . . . . . . . . . . . . . . . . . . . 20
      3.1.2 SI model over the bivariate normal population . . . . . . . . . . . . 23
   3.2 Simulation of SIR model . . . . . . . . . . . . . . . . . . . . . . . . . . . 26
      3.2.1 Simulation of SIR model in a grid population . . . . . . . . . . . . 26
      3.2.2 Simulation of SIR model in a bivariate normal population . . . . . 30
   3.3 Estimating the infectious period in the SIR model 33
      3.3.1 Maximum Likelihood-Based Estimation . . . . . . . . . . . . . . . . . . . 34
      3.3.2 Bayesian MCMC Estimation . . . . . . . . . . . . . . . . . . . . . . 42
   3.4 Misspecification of Infectious Time in SIR Model . . . . . . . . . . . . . . 46
      3.4.1 Misspecification of SIR model in grid population . . . . . . . . . . . 47
3.4.2 Misspecification of SIR model in bivariate normal population . . . 52

4 Conclusion 56

References 59
List of Tables

2.1 Number of iterations to convergence to the maximum likelihood estimates of $\alpha$ and $\beta$, for ten epidemic simulations, using BFGS, CG, and NM optimization algorithms. The true values of $\alpha$ and $\beta$ were 0.5 and 5.0, respectively. ................................................................. 10

3.1 Mean estimates of $\alpha$ and $\beta$, using MLE and Bayesian MCMC. ............... 25

3.2 SIR model mean estimates of $\alpha$ and $\beta$ in the grid and bivariate normal populations, using MLE and MCMC. ................................................................. 33

3.3 MLE estimates of $\gamma$ in the grid ($\hat{\gamma}_{\text{GRID}}$) and bivariate normal ($\hat{\gamma}_{\text{BIVAR}}$) populations, for $\gamma = 1, \ldots, 10$. $p$ is the proportion of times $\hat{\gamma}_{ij} = \gamma_T$, as discussed in Algorithm 3. ................................................................. 36

3.4 Bayesian MCMC estimates of $\gamma$ in the grid ($\hat{\gamma}_{\text{GRID}}$) and bivariate normal ($\hat{\gamma}_{\text{BIVAR}}$) populations, when $\gamma \sim DU[1, 10]$. $p$ is the proportion of times $\hat{\gamma}_{ij} = \gamma_T$, as discussed in Algorithm 3. ......................................................... 43
List of Figures

3.1 (a) The 17 x 17 grid population; (b) The bivariate normal population. . . . 18
3.2 Epidemic curves under the grid population (a) and the bivariate normal population (b). The bivariate normal population was truncated here at time point 100 for data visualization purposes, as beyond that time very few individuals became infectious. The smooth blue line indicates the average number of infectious individuals at each time point. . . . . . . . . 19
3.3 ML estimates of $\alpha$ and $\beta$ under an SI framework in the grid population, with accompanying 95% confidence intervals. True values are in red, and the means of the ML estimates are in blue. . . . . . . . . . . . . . . . . . . . . 21
3.4 Posterior estimates of $\alpha$ and $\beta$ under an SI framework, with 95% credible intervals, in the grid population. True values are in red, and the posterior means are in blue. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 22
3.5 MCMC trace plots for susceptibility (a) and spatial transmissibility (b). . . 22
3.6 ML estimates of $\alpha$ and $\beta$ under an SI framework and the bivariate normal population, with accompanying 95% confidence intervals. True values are in red, and the means of the parameter estimates are in blue. . . . . . . . 23
3.7 Posterior estimates of $\alpha$ and $\beta$ under an SI framework and the bivariate normal population, with associated 95% credible intervals. True values are in red, and the means of the posterior estimates are in blue. 24

3.8 MCMC trace plots of $\alpha$ (a) and $\beta$ (b) in the bivariate normal population. 25

3.9 Epidemic curves for the SIR framework and the grid population, $\gamma_{GRID} = 1$ (a), and $\gamma_{GRID} = 10$ (b), with average number of infectious individuals per time step (blue). 27

3.10 ML estimates of the susceptibility, $\alpha$ (a), and spatial transmissibility $\beta$ (b) when $\gamma_{GRID} = 1, \ldots, 10$. Dashed vertical lines represent boundaries for values of $\gamma_{GRID}$. That is, for all epidemics from 1 to 10 (horizontal axis bin 1), the true $\gamma_{GRID} = 1$. From epidemics 11 to 20, the true $\gamma_{GRID} = 2$, etc. Solid black lines indicate average parameter estimates in each bin. Solid red lines show overall average across all bins. Solid blue lines indicate the true value of the parameter. Estimates are shown with corresponding 95% confidence intervals for $\alpha$ and $\beta$. 28

3.11 MCMC estimates of the susceptibility, $\alpha$, and spatial transmissibility $\beta$ when $\gamma_{GRID} \sim DU[1, 10]$. The blue line is the average posterior mean estimate and the red line is the true value ($\alpha = 0.5, \beta = 5.0$). 29

3.12 Trace plots for susceptibility ($\alpha$) and transmissibility ($\beta$) in the grid population. 29

3.13 Epidemic curves for the for the SIR grid population, $\gamma_{BIVAR} = 1$ (b), and $\gamma_{BIVAR} = 10$ (b), with the average number of infectious individuals per time step (blue). 30

3.14 ML estimates of $\alpha$ and $\beta$, and with associated 95% confidence intervals. The blue line is the average ML estimate and the red line is the true value. 31
3.15 MCMC estimates of the susceptibility, $\alpha$, and spatial transmissibility $\beta$ when $\gamma_{BIVAR} \sim DU[1, 10]$. The red line is the true value and the blue line is the average posterior mean across all bins. .......................... 32

3.16 Trace plots for susceptibility ($\alpha$) and transmissibility ($\beta$) in the bivariate normal population. .......................................................... 32

3.17 Epidemics generated in the grid population under (a) $\gamma_{GRID} = 8$ and (b) $\gamma_{GRID} = 10$, with average number of infectious individuals per time step in blue. ....................................................... 38

3.18 Epidemics generated in the bivariate normal population under (a) $\gamma_{BIVAR} = 8$ and (b) $\gamma_{BIVAR} = 10$, with average number of infectious individuals per time step in blue. ....................................................... 38

3.19 Epidemics generated using $\gamma_{GRID} = 1, 2, \text{and } 10$, respectively, with average number of infectious individuals per time step in blue. ................................. 39

3.20 ML Estimates $\hat{\gamma}_{BIVAR}$ under $\gamma_{BIVAR} = 1...10$. True value of $\gamma_{BIVAR}$ is in grey. 40

3.21 ML estimates $\hat{\gamma}_{GRID}$ under $\gamma_{GRID} = 1, \ldots, 10$. True value of $\gamma_{GRID}$ is in grey. 41

3.22 MCMC estimates $\hat{\gamma}_{GRID}$ under $\gamma_{GRID} = 1, \ldots, 10$. True value of $\gamma_{GRID}$ is in grey. .......................................................... 44

3.23 MCMC estimates $\hat{\gamma}_{BIVAR}$ under $\gamma_{BIVAR} = 1, \ldots, 10$. True value of $\gamma_{BIVAR}$ is in grey. .......................................................... 45

3.24 ML estimates of alpha (susceptibility) with associated 95% confidence intervals, when the true infectious period is 1 day (a) and 10 days (b) in the grid population. The straight red line is the true value and the straight blue line is the average parameter estimate. ................................. 48
3.25 ML estimates of beta (spatial transmissibility) with associated 95% confidence intervals, when the true infectious period is 1 day (a) and 10 days (b) in the grid population. The straight red line is the true value and the straight blue line is the average parameter estimate. 

3.26 MCMC estimates of $\alpha$ (susceptibility) with associated 95% credible intervals, when $\gamma_T = 1$ (a) and $\gamma_T = 10$ (b) in the grid population. The straight red line is the true value and the straight blue line is the average parameter estimate.

3.27 MCMC estimates of $\beta$ (spatial transmissibility) with associated 95% credible intervals, when $\gamma_T = 1$ (a) and $\gamma_T = 10$ (b) in the grid population. The straight red line is the true value and the straight blue line is the average parameter estimate.

3.28 ML estimates of alpha (susceptibility) with associated 95% confidence intervals, when the true infectious period is 1 day (a) and 10 days (b) in the bivariate normal population. The straight red line is the true value and the straight blue line is the average parameter estimate.

3.29 ML estimates of beta (spatial transmissibility) with associated 95% confidence intervals, when the true infectious period is 1 day (a) and 10 days (b) in the bivariate normal population. The straight red line is the true value and the straight blue line is the average parameter estimate.

3.30 MCMC estimates of $\alpha$ (susceptibility) with associated 95% credible intervals, when $\gamma_T = 1$ (a) and $\gamma_T = 10$ (b) in the bivariate normal population. The straight red line is the true value and the straight blue line is the average parameter estimate.
3.31 MCMC estimates of \( \beta \) (spatial transmissibility) with associated 95% credible intervals, when \( \gamma_T = 1 \) (a) and \( \gamma_T = 10 \) (b) in the bivariate normal population. The straight red line is the true value and the straight blue line is the average parameter estimate.
Chapter 1

Introduction

While both the deterministic and the stochastic branches of mathematical epidemiology share the goal of predicting and controlling the spread of disease, stochastic models allow researchers to answer a broader scope of questions. The development of deterministic infectious disease modelling, before stochastic models were put forth (Bailey (1975), Trapman (2006), Britton (2010)), answered questions such as: How many individuals will become infected if the epidemic is allowed to run unchecked? What is the endemic level of the disease in the population of interest (Britton, 2010)? These questions can indeed be answered using a deterministic framework, for a large, homogeneously and homogeneously-mixing population; however, to address a smaller or more dynamic population, and to allow standard errors to be reported on parameter estimates, stochastic modelling was introduced as a natural extension to its deterministic counterpart (Britton, 2010). A stochastic model allows researchers to answer questions related to the probability of disease spread, based on, for example, the basic reproductive ratio (Britton, 2010), or the susceptibility of the population and the transmissibility of the disease. Susceptibility refers to the state of an individual prior to an epidemic,
and influences how likely it is that the individual could become infectious, due to, for example, pre-existing medical conditions. Transmissibility, as defined in Deardon et al. (2010), is a parameter that describes the rate of increase in infectious pressure per additional infectious population member. Although deterministic modelling is useful for large populations, stochastic epidemic modelling can more flexibly account for individual characteristics and fluid populations, regardless of population size.

With enhanced flexibility comes increased computational complexity. Stochastic modelling is more computationally demanding, in general, than deterministic modelling (Wilkinson, 2009). In particular, a major concern for inference in infectious disease models is the amount of computation time required to evaluate the likelihood function McKinley, Cook, and Deardon (2009). Previous attempts to reduce likelihood computation time have been made via linear approximations to the model (e.g. Deardon et al. (2010); Kwong and Deardon (2012)) or via a discretization of the parameter space (Deardon, Fang, and Kwong (2014, in press)). However, each of these approaches comes with inherent disadvantages that are seen with any approximation-based method.

We focus on individual-level models (ILMs) of infectious disease as described in Deardon et al. (2010), which can incorporate characteristics such as individual vaccination status and medical history into the probability of an individual becoming infectious. ILMs allow for parameters that represent susceptibility, transmissibility, and separation distance between individuals (e.g. spatial or network-based). The typically-chosen method for parameter estimation is Markov Chain Monte Carlo (MCMC) in a Bayesian framework. An alternative approach is to use maximum likelihood (ML) estimation, which includes the convenient property of asymptotic normality and hence easily-derived con-
In this thesis we address the following: first, we discuss simulation for ILMs (section 2). Then we illustrate the difficulty of estimating the infectious period via maximum likelihood and Bayesian MCMC (section 3.3). Furthermore, we compare maximum likelihood and Bayesian MCMC-based estimates, both in terms of accuracy of point estimates and the width of confidence and credible intervals for each of the Susceptible-Infectious (SI, section 3.1) and Susceptible-Infectious-Removed (SIR, section 3.2) modelling frameworks. Separate consideration is given to the effect of misspecifying the infectious period when estimating other model parameters (section 3.4). Maximum likelihood estimation is carried out using R (R Core Team, 2013), and Bayesian MCMC in Fortran 90 (Adams, Brainerd, Martin, Smith, & Wagener, 1992).
Chapter 2

Spatial Individual Level Models (ILMs) of Infectious Disease Spread

Suppose the number of individuals in the population is indexed from 1 to n. In an SI model, these individuals are assumed to be in either of two non-overlapping sets: S (susceptible), and I (infectious), at any point in time. In an SIR model, this formulation is extended to include another compartment: R (removed), which contains individuals that were removed from the population after infection, due to, for example, recovery with acquired immunity, quarantine, death, etc. In our model, a discrete time point represents an interval or window (e.g. one day) in continuous time. We are interested in the probability that an individual becomes infectious at time t (that is, in the interval (t, t+1]). We consider the formulation given in Deardon et al. (2014, in press), in which this probability is given by:

\[
P(i,t) = 1 - \exp\left\{-\alpha \sum_{j \in I(t)} d_{ij}^{-\beta}\right\} \quad \alpha, \beta > 0
\]
where \( \alpha \) and \( \beta \) denote the susceptibility and spatial transmissibility parameters for the population, respectively; \( I(t) \) denotes the set of infectious individuals at time \( t \), and \( d_{ij} \) denotes some distance measure between an infectious individual \( j \) and susceptible individual \( i \) (here, the Euclidean distance is used). Susceptible individuals at time \( t \) either stay susceptible (are not infected) or become infectious at time \( t + 1 \). The probability of all infection or non-infection events observed in the population from time \( t \) to time \( t + 1 \) is given by:

\[
f_t(S, I) = \left[ \prod_{i \in S(t+1)} \left( 1 - P(i, t) \right) \right] \left[ \prod_{i \in I(t+1) \setminus I(t)} P(i, t) \right]
\]  

(2.1)

where \( S(t+1) \) is the set of susceptible individuals and \( I(t+1) \setminus I(t) \) is the set of newly infectious individuals.

We can formulate the likelihood as the product of equation 2.1 over all observed time points, up to the maximum time of observation (here denoted \( t_{\text{max}} \)):

\[
L(\alpha, \beta) = \prod_{t=1}^{t_{\text{max}}} \left[ \prod_{i \in S(t+1)} \left( 1 - P(i, t) \right) \right] \left[ \prod_{i \in I(t+1) \setminus I(t)} P(i, t) \right]
\]

It is often simpler to work with the log-likelihood, and we do so here:

\[
I(\alpha, \beta) = \sum_{t=1}^{t_{\text{max}}} \left[ \sum_{i \in S(t+1)} \log(1 - P(i, t)) + \sum_{i \in I(t+1) \setminus I(t)} \log(P(i, t)) \right]
\]

(2.2)

Extending this formulation to the SIR model is done by identifying the specific time window of individual infectivity, commonly known as the infectious period, denoted \( \gamma \). Individual \( j \) is infectious (can infect others) for a period of \( \gamma \) time units. That is, an
individual that becomes infectious at time $t - \gamma$ can infect others until time $t$. After time $t$, the individual is no longer infectious and passes to the R (removed) category. Denoting the time at which individual $j$ becomes infectious by $\text{inf}(j)$, we then have the following:

$$j \in I(t) \text{ if } t - \gamma \leq \text{inf}(j) < t$$

While this duration can be individualized, here we consider a constant, fixed infectious period for the entire population. The assumption of an equal infectious period across individuals is well-founded biologically in the case of measles (individuals are infectious for a period of 8 days - four days before and four days following the appearance of a rash (Fiebelkorn & Goodson, 2014)). The SIR model of infectious disease has been used extensively in the study of measles outbreaks (e.g. Allen, Jones, and Martin (1991); Haccou, Jagers, and Vatutin (2005)). Here, we treat the infectious period as discrete and interpret it as the number of days for which an individual can transmit the pathogen. As epidemiological data are often only available in discrete form (Sattenspiel & Lloyd, 2009), the assumption of a discrete infectious period reflects the reality of epidemiological modelling.

### 2.1 Maximum likelihood estimation in the context of ILMs

In the method of maximum likelihood estimation, we are interested in finding the parameter values that maximize the likelihood function, the probability of observing the data, conditional on the parameters. In the SI framework, we have two parameters, $\alpha$ and $\beta$, and our data consist of the time point at which each individual became infectious. We will refer to one “epidemic” as one set of times at which each individual in a population becomes infectious. From each epidemic, we can acquire a maximum likelihood esti-
mate (MLE) for the population susceptibility and population spatial transmissibility.

Here, estimating $\gamma$ in the SIR model is not done in the traditional maximum likelihood sense. Since $\gamma$ is a discrete parameter (whereas $\alpha$ and $\beta$ are continuous), we cannot take an explicit analytical derivative of the likelihood in terms of $\gamma$. Instead, we can use an approach similar to profile likelihood. The MLEs of $\alpha$ and $\beta$ are substituted into the log-likelihood function, and the values of the log-likelihood are compared. The estimate, $\hat{\gamma}$ is the value of $\gamma$ that generated the maximum value of the log-likelihood. Furthermore, $\gamma$ is also estimated using Bayesian MCMC, and the results are compared.

2.1.1 Derivations of the Analytic First Derivatives

Recall the problem of maximizing the log-likelihood is equivalent to the problem of finding values for parameters $\alpha$ and $\beta$ that minimize the value of the score function. Let $\theta = (\alpha, \beta)$, so we can derive the general form of the first derivative:

$$\frac{\partial l(\theta)}{\partial \theta_i} = \frac{\partial}{\partial \theta_i} \left[ \sum_{t=1}^{t_{\max}} \sum_{i \in S(t+1)} \log(1 - P(i, t)) \right] + \frac{\partial}{\partial \theta_i} \left[ \sum_{t=1}^{t_{\max}} \sum_{i \in I(t+1) \setminus I(t)} \log(P(i, t)) \right]$$

$$= \left[ \sum_{t=1}^{t_{\max}} \sum_{i \in S(t+1)} \frac{\partial}{\partial \theta_i} \left( -\alpha \sum_{i=1}^{n} d_{ij}^{-\beta} \right) \right] + \left[ \sum_{t=1}^{t_{\max}} \sum_{i \in I(t+1) \setminus I(t)} \frac{\partial}{\partial \theta_i} \log \left( 1 - \exp \left( -\alpha \sum_{i=1}^{n} d_{ij}^{-\beta} \right) \right) \right]$$

7
\[
\frac{\partial l(\theta)}{\partial \theta_i} = \frac{\partial}{\partial \theta_i} \left[ \sum_{t=1}^{t_{\text{max}}} \sum_{i \in S(t+1)} \log(1 - P(i,t)) \right] + \frac{\partial}{\partial \theta_i} \left[ \sum_{t=1}^{t_{\text{max}}} \sum_{i \in I(t+1) \setminus I(t)} \log(P(i,t)) \right]
\]

\[
= \left[ \sum_{t=1}^{t_{\text{max}}} \sum_{i \in S(t+1)} \frac{\partial}{\partial \theta_i} \left( -\alpha \sum_{i=1}^{n} d_{ij}^{-\beta} \right) \right] + \left[ \sum_{t=1}^{t_{\text{max}}} \sum_{i \in I(t+1) \setminus I(t)} \frac{\partial}{\partial \theta_i} \log \left( 1 - \exp \left( -\alpha \sum_{i=1}^{n} d_{ij}^{-\beta} \right) \right) \right]
\]

\[
S(\alpha) = \frac{\partial l(\theta)}{\partial \alpha} = \left[ \sum_{t=1}^{t_{\text{max}}} \sum_{i \in S(t+1)} \left( \sum_{j \in I(t)} d_{ij}^{-\beta} \right) \right] + \left[ \sum_{t=1}^{t_{\text{max}}} \sum_{i \in I(t+1) \setminus I(t)} \frac{\alpha \sum_{j \in I(t)} d_{ij}^{-\beta}}{\exp \left( \alpha \sum_{j \in I(t)} d_{ij}^{-\beta} \right) - 1} \right]
\]

\[
S(\beta) = \frac{\partial l(\theta)}{\partial \beta} = \left[ \sum_{t=1}^{t_{\text{max}}} \sum_{i \in S(t+1)} \alpha \sum_{j \in I(t)} d_{ij}^{-\beta} \log(d_{ij}) \right] + \left[ \sum_{t=1}^{t_{\text{max}}} \sum_{i \in I(t+1) \setminus I(t)} \frac{\alpha \sum_{j \in I(t)} d_{ij}^{-\beta} \log(d_{ij})}{1 - \exp \left( \alpha \sum_{j \in I(t)} d_{ij}^{-\beta} \right)} \right]
\]

The maximization of the log-likelihood is carried out by setting the gradient equal to 0 and solving for the optimal parameter values. Although the analytic derivatives (and second derivatives) are tractable, here quasi-Newton methods are used, as we found them to require less run time to converge than the Newton-Raphson method, which agrees with the literature (Aptech Systems, Inc., 2001). Quasi-Newton methods are a class of optimization method that approximates Newton-Raphson Method by replacing the Hessian in the direction of decrease with an approximation to the Hessian.
Epidemic Simulations To Find an Appropriate Algorithm

Simulations of an SI ILM were conducted over a 10 x 10 grid of locations of individuals. All individuals were assumed susceptible at time $t = 0$. The first individual to become infectious, at time $t = 1$, was chosen at random. Subsequent individuals became infectious based on the probability of infection, given in equation 2.1. Simulation continued until all individuals were infectious.

Optimization Methods

Here, the package maxLik Henningsen and Toomet (2011), a wrapper for the classical optim (Nash & Varadhan, 2011) package, was used to carry out ML estimation in R (R Core Team, 2013). Under this package, several maximization algorithms are available. Specifically, Newton-Raphson, Broyden-Fletcher-Goldfarb-Shanno (BFGS), BFGSR (a variant of BFGS that runs specifically in R), Berndt-Hall-Hall-Hausman (BHHH), simulated annealing (SANN), conjugate gradients (CG), and Nelder-Mead (NM) are available methods (Henningsen & Toomet, 2011). To reduce the run time of the maximization, the log-likelihood, given in eqn. 2.2, was computed subject to an inequality bound on $\alpha$ and $\beta$, $\alpha + \beta \leq 1000$. This bound is extremely large compared to the true values of $\alpha$ and $\beta$ (which were 0.5 and 5, respectively), but stopped the maximization algorithm from searching unlikely regions of the original parameter space. Under this bound, the optimization routines available to handle the maximization were limited to CG, BFGS, SANN, and NM, all of which (except NM) are quasi-Newton methods. An optimization method was chosen based on a small sample of 10 simulated epidemics. The number of iterations to convergence for each of the epidemics is shown in Table 1. SANN was only computed for one epidemic, as the run time was prohibitively long (118.3 minutes to
Table 2.1: Number of iterations to convergence to the maximum likelihood estimates of $\alpha$ and $\beta$, for ten epidemic simulations, using BFGS, CG, and NM optimization algorithms. The true values of $\alpha$ and $\beta$ were 0.5 and 5.0, respectively.

<table>
<thead>
<tr>
<th></th>
<th>BFGS</th>
<th>CG</th>
<th>NM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean no. iterations</td>
<td>108.1</td>
<td>426.4</td>
<td>149.4</td>
</tr>
<tr>
<td>Mean $\alpha$</td>
<td>0.49</td>
<td>0.40</td>
<td>0.46</td>
</tr>
<tr>
<td>Mean $\beta$</td>
<td>4.91</td>
<td>4.30</td>
<td>5.04</td>
</tr>
<tr>
<td>Mean Std. Err. of $\alpha$</td>
<td>0.066</td>
<td>0.057</td>
<td>0.062</td>
</tr>
<tr>
<td>Mean Std. Err. of $\beta$</td>
<td>0.487</td>
<td>0.389</td>
<td>0.508</td>
</tr>
</tbody>
</table>

complete one set of maximum likelihood estimates). In these test examples, there were 100 individuals in the population. BFGS consistently converged faster than any other algorithm. BFGS has also been cited as the go-to quasi-Newton method (de Klerk, Roos, & Terlaky, 2006). Since the difference in MLEs between algorithms was negligible, and as BFGS was the fastest to converge, the BFGS algorithm is used for the remainder of the maximum likelihood discussion.

**Broyden-Fletcher-Goldfarb-Shanno (BFGS) Algorithm**

While there are many algorithms to produce approximate solutions for constrained nonlinear optimization problems, the class of method considered here is the “quasi-Newton” or “secant” family of optimization routines. These methods all approximate the Newton-Raphson Method. The general form of a quasi-Newton algorithm is given by de Klerk et al. (2006) as follows:

1. Let $\theta_0$ be the first guess at a vector of parameter estimates (here $\theta_0 = (\alpha_0, \beta_0)$), $l$ be the negative log-likelihood function, $\nabla l(\theta_k)$ be the gradient of the negative log-likelihood evaluated at point $\theta_k$ and set the initial Hessian of the negative log-likelihood, $H$, to the identity.
2. For each iteration $k$, find the update direction $d_k = -H_k \nabla l(\theta_k)$, also called the direction of decrease.

3. Perform a line search $\theta_{k+1} = \arg\min \ l(\theta_k + \lambda d_k)$, where $\lambda$ is a step size.

4. Repeat until either $\nabla l(\theta_k)$ is within a pre-specified tolerance of 0, or until the maximum number of iterations is reached.

From the general steps for quasi-Newton optimization, we can discuss the specific steps of the BFGS algorithm in the context of maximum likelihood estimation for ILMs. The function we're interested in minimizing is the negative log-likelihood, as this is the natural equivalent of maximizing the positive log-likelihood. Specifically, we would like to create a local quadratic model of the negative log-likelihood, and minimize that quadratic function. The following steps are adapted for ILMs from Yu, N., Günter, and Schraudolph (2010).

First, the quadratic form is specified as:

$$Q(d_k) = \frac{1}{2} d_k^T H_k d_k + \nabla l(\theta_k)^T d_k + l(\theta_k)$$

Taking the derivative of $Q$ with respect to $d_k$ and setting it equal to zero yields the direction of descent:

$$d_k = -H^{-1} \nabla l(\theta_k)^T$$

Now we must choose the step size, $\lambda_k$, for each iteration $k$. This is done online according to the conditions of Wolfe (1969):
\[ l(\theta_{k+1}) \leq l(\theta_k) + c_1 \lambda_k \nabla l(\theta_k)^T d_k \]  (Armijo Rule for sufficient decrease)

\[ \nabla l(\theta_{k+1})^T d_k \geq c_2 \nabla l(\theta_k)^T d_k \]  (Curvature condition)

where \( 0 < c_1 < c_2 < 1 \), and \( \nabla l(\theta_k)^T d_k < 0 \)

The Armijo Rule for sufficient decrease ensures that the step length \( \lambda_k \) decreases the negative log-likelihood function by a sufficient amount. The curvature condition ensures that the new gradient, multiplied by the direction of decrease, is greater than or equal to the negative gradient of the previous step, which in turn means that the new gradient will be less than or equal to the gradient at the previous step. This ensures that the estimate will never get worse, and we move toward the minimum, where the gradient is (near) zero.

The rank-two update of the Hessian is given by:

\[ H_{k+1} = H_k + A_k + B_k \]

where

\[ A_k = \frac{\left( \nabla l(\theta_{k+1}) - \nabla l(\theta_k) \right) \left( \nabla l(\theta_{k+1}) - \nabla l(\theta_k) \right)^T}{\left( \nabla l(\theta_{k+1}) - \nabla l(\theta_k) \right)^T (\theta_{k+1} - \theta_k)} \]

and

\[ B_k = -\frac{H_k (\theta_{k+1} - \theta_k) (\theta_{k+1} - \theta_k)^T H_k}{(\theta_{k+1} - \theta_k)^T H_k (\theta_{k+1} - \theta_k)} \]
This update rule ensures the secant property

\[ H_{t+1}(\nabla l(\theta_{k+1}) - \nabla l(\theta_k)) = (\theta_{k+1} - \theta_k) \]

The BFGS algorithm was found to be the fastest available method for maximum likelihood estimation of ILMs in R. However, it should be noted that in the maxLik algorithm in R, the variance-covariance matrix that is generated from the maxLik object is not corrected for constrained parameters.

### 2.1.2 Bayesian MCMC for ILMs

Markov Chain Monte Carlo (MCMC) is by far the most popular technique for carrying out Bayesian inference. In this general class of methods, the Markov Chain is set up to have an equilibrium distribution equal to the posterior distribution. Using the prior distribution of the parameters of interest and the likelihood of the observed data given the parameter values, the posterior distribution, up to a constant of proportionality, is calculated in an iterative manner. After repeated iterations of the MCMC algorithm, the Markov Chain tends toward the equilibrium distribution. The resulting values of the equilibrium distribution, after a burn-in period is discarded, may be viewed as dependent samples from the posterior.

MCMC techniques are a commonly-used method of estimating parameters in spatio-temporal epidemic models (Gibson, 1997). Some advantages of a Bayesian MCMC approach in epidemic modelling include the ability to impute missing data (e.g. Mulla,
Byungtae, Kalamegham, and Nuwayhid (2009)), the ability to incorporate multiple sources of information through the prior distribution (Deardon et al., 2010), and the ability to build constraints directly into the prior for the parameters of interest.

In the context of spatial ILMs of equation 2.1, we are primarily interested in the marginal posterior distributions of the susceptibility and spatial parameters. Given the nature of the log-likelihood expressed in equation 2.2, the normalization constant using Bayes’ Theorem has no closed-form solution, but MCMC gives the advantage of eliminating the need for calculating the normalization constant. Instead, the unnormalized posterior probabilities of the parameter values given the observed data are computed. Furthermore, the expected value of the marginal posterior distributions of the parameters can be estimated from the MCMC sample, and we can obtain point estimates such as the posterior mean to compare with the corresponding ML estimates.

While there are many MCMC algorithms, the focus here will be on random walk Metropolis-Hastings (RWMH) MCMC. The algorithm used is outlined below:

1. Choose initial values $\alpha = \alpha_0, \beta = \beta_0$

2. Until the maximum number of iterations is reached, for each iteration $k$:

   a) Generate $Z_\alpha \sim U[-A,A]$ and $Z_\beta \sim U[-B,B], A, B \in R^+$

   b) Propose candidate values $\alpha_p = \alpha_{k-1} + z_\alpha$ and $\beta_p = \beta_{k-1} + z_\beta$.

   If $\alpha_k$ and $\beta_k$ are outside the prior range, proceed to rejection in step e)

   c) Calculate the acceptance probability

   $$\psi = \min\{0, l(\alpha_p, \beta_p) + f(\alpha_p) + f(\beta_p) - l(\alpha_{k-1} - \beta_{k-1}) - f(\alpha_{k-1}) - f(\beta_{k-1})\}$$
where $l(\alpha_p, \beta_p)$ is obtained from [2] and $f(\alpha_p)$ and $f(\beta_p)$ are the log prior distributions of

$\alpha$ and $\beta$, respectively

d) Generate $u$ from $U \sim U[0, 1]$

e) If $\log(u) < \psi$, accept proposal values and set $\alpha_k = \alpha_p, \beta_k = \beta_p$.

Else reject and set $\alpha_k = \alpha_{k-1}, \beta_k = \beta_{k-1}$

f) Go to 2 a) until a predetermined number of iterations has been reached

Here, the proposal distributions, $U(-A, A), U(-B, B)$, are chosen for simplicity, but any symmetric distributions could have been chosen.

Throughout the present work, the priors on $\alpha$ and $\beta$ will both be $U(0, 1000)$. The proposal distribution for $\alpha$ and $\beta$ will be $z_1 \sim U(-0.15, 0.15)$ and $z_2 \sim U(-0.4, 0.4)$, respectively.
Chapter 3

Simulation Studies

Simulations of epidemics from the two infectious disease modelling frameworks, SI and SIR, are carried out. Additionally, two populations are considered: 289 individuals placed at coordinates of a 17x17 grid (Figure 1 (a)), and a bivariate normal population consisting of three clusters of 100 individuals each (Figure 1 (b)). Overlap between clusters is introduced to help encourage the epidemics to spread to all three clusters. In each study, the true value of \( \alpha \) is 0.5 and the true value of \( \beta \) is 5.0. These parameter settings generated epidemics that were fast enough to be observed over a short time frame, speeding up epidemic simulation.

The bivariate normal population is generated using code from Burkardt (2009). Clusters are created using the following distributions:

\[
(x, y)_1 \sim MVN(\mu_1, \Sigma_1) \quad \text{where} \quad \mu_1 = (1, 1), \quad \Sigma_1 = \begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix}
\]
\[(x, y)_2 \sim \text{MVN}(\mu_2, \Sigma_2) \quad \text{where } \mu_2 = (6, 6), \quad \Sigma_2 = \begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix} \]

\[(x, y)_3 \sim \text{MVN}(\mu_3, \Sigma_3) \quad \text{where } \mu_3 = (15, 15), \quad \Sigma_3 = \begin{bmatrix} 5 & 2 \\ 2 & 5 \end{bmatrix} \]

Ten epidemics per true value of \( \gamma \) are generated, where values of \( \gamma = 1, \ldots, 10 \) are considered. The infectious period will be denoted \( \gamma_{\text{GRID}} \) for the grid population and \( \gamma_{\text{BIVAR}} \) for the bivariate normal population. In both the SI and the SIR models, we assume that the parameter values for susceptibility (\( \alpha \)) and spatial transmissibility (\( \beta \)) are equal for all individuals.

We first consider the SI model and estimate \( \alpha \) and \( \beta \) using both maximum likelihood and Bayesian MCMC. Then we extend the model to include the removed compartment \( R \) and estimate \( \alpha, \beta, \) and \( \gamma \) in the resulting SIR model. In this latter study, we estimate all parameters under the same Bayesian MCMC algorithm as used in the SI model; however, estimating \( \gamma \) using ML is not done in the traditional sense. Instead, we estimate \( \alpha \) and \( \beta \) first, using ML, and estimate the infectious period conditionally on those MLEs.
positions of individuals on a 17 x 17 grid

Positions of individuals in three Bivariate Gaussian clusters

Figure 3.1: (a) The 17 x 17 grid population; (b) The bivariate normal population.

3.1 Simulation Study of SI Model

One hundred epidemics are simulated from the SI version of the model over both the grid and the bivariate normal populations and the values of $\alpha$ and $\beta$ are estimated for each epidemic. Figure 3.2(a) shows the number of infectious individuals per time point for each of the 100 simulated epidemics over the grid population; Figure 3.2(b) shows the same for the bivariate normal population.
Figure 3.2: Epidemic curves under the grid population (a) and the bivariate normal population (b). The bivariate normal population was truncated here at time point 100 for data visualization purposes, as beyond that time very few individuals became infectious. The smooth blue line indicates the average number of infectious individuals at each time point.

Notice in Figure 3.2(b), the epidemic under the bivariate normal population exhibits “jumps” in the number of infectious individuals, and these jumps are consistently at around 100 and 200 individuals. This is consistent with the clustered nature of the population; that is, the disease tends to move through each cluster of 100 individuals before managing to proceed to the next cluster.
3.1.1 SI model over the grid population

Maximum Likelihood Estimation

Maximum likelihood estimation of the susceptibility and transmissibility parameters under an SI model is carried out for each of the 100 epidemics. Starting values of $\alpha_0 = 1.0$ and $\beta_0 = 2.0$ are used to begin each maximum likelihood estimation procedure. Results did not appear sensitive to the starting values. As shown in Figure 3.3(a), the ML estimates of the susceptibility parameter $\alpha$ tend to be close to the true value. The mean of the distribution of the ML estimates is also very close to the true value ($\alpha = 0.5$). In Figure 3.3(b), the estimates of the spatial transmissibility parameter ($\beta$) vary a bit more widely; however, the average value of the ML estimate is still quite close to the true value ($\beta = 5.0$). Table 3.1 shows the average estimates of $\alpha$ and $\beta$ and corresponding 95% confidence interval widths.
Bayesian MCMC Estimation

The MCMC credible intervals are wider than the ML confidence intervals for both the susceptibility and transmissibility parameters, as shown in subfigures (a) and (b) of Figure 3.4.

Convergence is assessed visually. A typical pair of trace plots is shown in Figure 3.5. From the pattern exhibited in the MCMC trace plots, the Markov Chain has indeed converged.

Figure 3.3: ML estimates of $\alpha$ and $\beta$ under an SI framework in the grid population, with accompanying 95% confidence intervals. True values are in red, and the means of the ML estimates are in blue.
Figure 3.4: Posterior estimates of $\alpha$ and $\beta$ under an SI framework, with 95% credible intervals, in the grid population. True values are in red, and the posterior means are in blue.

Figure 3.5: MCMC trace plots for susceptibility (a) and spatial transmissibility (b).
3.1.2 SI model over the bivariate normal population

Maximum Likelihood Estimation

Contrastingly, in the bivariate normal population, it was found that the susceptibility parameter was slightly overestimated, on average ($\hat{\alpha} = 0.515$). The spatial transmissibility parameter does not show the same level of overestimation ($\hat{\beta} = 5.02$).

Figure 3.6: ML estimates of $\alpha$ and $\beta$ under an SI framework and the bivariate normal population, with accompanying 95% confidence intervals. True values are in red, and the means of the parameter estimates are in blue.
Bayesian MCMC Estimation

Under MCMC estimation in the bivariate normal population, $\alpha$ is estimated to be even higher than in the ML estimation, $\hat{\alpha} = 0.535$.

Figure 3.7: Posterior estimates of $\alpha$ and $\beta$ under an SI framework and the bivariate normal population, with associated 95% credible intervals. True values are in red, and the means of the posterior estimates are in blue.
Table 3.1: Mean estimates of $\alpha$ and $\beta$, using MLE and Bayesian MCMC.

<table>
<thead>
<tr>
<th></th>
<th>Maximum Likelihood</th>
<th>Bayesian MCMC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grid</td>
<td>$\hat{\alpha}$</td>
<td>0.499 ± 0.04</td>
</tr>
<tr>
<td></td>
<td>$\hat{\beta}$</td>
<td>5.02 ± 0.255</td>
</tr>
<tr>
<td>Bivariate normal</td>
<td>$\hat{\alpha}$</td>
<td>0.515 ± 0.057</td>
</tr>
<tr>
<td></td>
<td>$\hat{\beta}$</td>
<td>5.014 ± 0.208</td>
</tr>
</tbody>
</table>

Figure 3.8: MCMC trace plots of $\alpha$ (a) and $\beta$ (b) in the bivariate normal population.

The confidence intervals for the SI framework are narrower than the credible intervals for the same level of significance (Table 3.1). Once again we see the MCMC chains for a typical epidemic appear to mix well (Figure 3.8). The standard deviation for the estimates of $\beta$ are an order of magnitude larger than those for $\alpha$, except for the bivariate normal population when $\alpha$ is estimated using Bayesian MCMC. The variance of $\alpha$ is slightly higher in the bivariate normal population than in the grid population, whereas in the grid population, the variance of $\beta$ is slightly higher.
3.2 Simulation of SIR model

Extending the previously-discussed SI model, the SIR framework includes a Removed (R) compartment to contain individuals who have been removed from the population, due to recovery, death, quarantine, etc. The addition of this compartment results in a new model parameter which indicates the duration of the infectious period. As discussed in section 2, for diseases such as measles, the infectious period can be considered equal for all individuals in the population, and we do so here.

3.2.1 Simulation of SIR model in a grid population

The epidemics generated under $\gamma_{GRID} = 1$ tend to end around the same time as those generated under $\gamma_{GRID} = 10$ (Figure 3.9). When the infectious period is set to 10, the more traditional parabolic epidemic curve emerges. There are many individuals in the population that do not become infectious (on average, across all values of $\gamma_{GRID}$, 85 individuals remain susceptible after we stop observing the epidemic). There is a much higher peak in the number of infectious individuals under an infectious period of $\gamma_{GRID} = 10$ (69 individuals at time 10) than $\gamma_{GRID} = 1$ (44 individuals at time 15).
Figure 3.9: Epidemic curves for the SIR framework and the grid population, $\gamma_{GRID} = 1$ (a), and $\gamma_{GRID} = 10$ (b), with average number of infectious individuals per time step (blue).

**Maximum Likelihood Estimation**

The average ML estimates of $\alpha$ and $\beta$ are very similar for $\gamma_{GRID}$ values of 2 through 10. Interestingly, when $\gamma_{GRID} = 1$, $\alpha$ was consistently overestimated ($\hat{\alpha} = 0.84$). None of the 95% confidence intervals for $\alpha$ when $\gamma_{GRID} = 1$ contained the true value (Figure 3.10). Additionally, $\beta$ was underestimated ($\hat{\beta} = 4.87$) when $\gamma_{GRID} = 1$. We might expect the susceptibility to be underestimated and the transmissibility to be underestimated when the infectious period is misspecified to be higher in the estimation process than it is in the model that is used to generate the epidemic, which is not the case here. We estimate the infectious period in section 3.3. However, under no model misspecification, it is surprising to see confidence intervals that do not contain the true value.
Figure 3.10: ML estimates of the susceptibility, $\alpha$ (a), and spatial transmissibility $\beta$ (b) when $\gamma_{GRID} = 1, \ldots, 10$. Dashed vertical lines represent boundaries for values of $\gamma_{GRID}$. That is, for all epidemics from 1 to 10 (horizontal axis bin 1), the true $\gamma_{GRID} = 1$. From epidemics 11 to 20, the true $\gamma_{GRID} = 2$, etc. Solid black lines indicate average parameter estimates in each bin. Solid red lines show overall average across all bins. Solid blue lines indicate the true value of the parameter. Estimates are shown with corresponding 95% confidence intervals for $\alpha$ and $\beta$.

**Bayesian MCMC Estimation**

As in the ML estimation previously shown, there is a tendency to overestimate the susceptibility parameter when $\gamma_{GRID} = 1$ ($\hat{\alpha} = 0.88$), and the spatial transmissibility parameter is again underestimated, although to a lesser extent ($\hat{\beta} = 4.91$). The credible intervals are slightly wider than the corresponding confidence intervals for each parameter (Table 3.2).
Figure 3.11: MCMC estimates of the susceptibility, $\alpha$, and spatial transmissibility $\beta$ when $\gamma_{GRID} \sim DU[1, 10]$. The blue line is the average posterior mean estimate and the red line is the true value ($\alpha = 0.5, \beta = 5.0$).

Figure 3.12: Trace plots for susceptibility ($\alpha$) and transmissibility ($\beta$) in the grid population.
### 3.2.2 Simulation of SIR model in a bivariate normal population

The epidemic curves under the bivariate normal population exhibit different characteristics than the epidemic curves for the corresponding $\gamma$ values in the grid population. In the bivariate normal population, we see epidemic peaks of 98 individuals at time 10 under $\gamma_{\text{BIVAR}} = 1$ and 105 individuals at time 6 when $\gamma_{\text{BIVAR}} = 10$, compared to only 44 when $\gamma_{\text{GRID}} = 1$ and 69 when $\gamma_{\text{GRID}} = 10$. Additionally, we observe that in the grid population, under $\gamma_{\text{GRID}} = 10$, we don’t see a definitive spike in the number of infectious individuals, compared to the peak in $\gamma_{\text{BIVAR}} = 10$. In the bivariate normal population, there are only 14 individuals on average that do not become infectious.

![Graphs showing epidemic curves](image)

Figure 3.13: Epidemic curves for the SIR grid population, $\gamma_{\text{BIVAR}} = 1$ (b), and $\gamma_{\text{BIVAR}} = 10$ (b), with the average number of infectious individuals per time step (blue).

### Maximum Likelihood Estimation

The estimates of $\alpha$ and $\beta$ are very similar to our findings in the grid population, regardless of the value of $\gamma_{\text{BIVAR}}$; however, we do not observe the same overestimation of $\alpha$
when $\gamma_{BIVAR} = 1$. The spatial transmissibility is again underestimated when $\gamma_{BIVAR} = 1$ ($\hat{\beta} = 4.64$).

\begin{figure}[h]
\centering
\begin{subfigure}{0.45\textwidth}
\centering
\includegraphics[width=\textwidth]{plot1}
\caption{(a) 95% confidence intervals of $\alpha$ for SIR bivariate normal population}
\end{subfigure}
\hfill
\begin{subfigure}{0.45\textwidth}
\centering
\includegraphics[width=\textwidth]{plot2}
\caption{(b) 95% confidence intervals of $\beta$ for SIR bivariate normal population}
\end{subfigure}
\caption{ML estimates of $\alpha$ and $\beta$, and with associated 95% confidence intervals. The blue line is the average ML estimate and the red line is the true value.}
\end{figure}

**Bayesian MCMC Estimation**

Figure 3.15(b) shows $\hat{\beta} = 4.67$ by Bayesian MCMC estimation. Regardless of the population of interest and the estimation method, $\beta$ is underestimated when $\gamma = 1$ in the SIR framework. It could be that under an infectious period of 1 day, the spatial effect of epidemic spread is diminished. That is, perhaps locations of infectious and susceptible individuals relative to each other do not have a strong influence on disease spread when $\gamma = 1$. The epidemics would then spread in a more random fashion.
Figure 3.15: MCMC estimates of the susceptibility, $\alpha$, and spatial transmissibility $\beta$ when $\gamma_{BIVAR} \sim DU[1,10]$. The red line is the true value and the blue line is the average posterior mean across all bins.

Figure 3.16: Trace plots for susceptibility ($\alpha$) and transmissibility ($\beta$) in the bivariate normal population.

Again, the confidence intervals are slightly narrower than the corresponding credible
Table 3.2: SIR model mean estimates of $\alpha$ and $\beta$ in the grid and bivariate normal populations, using MLE and MCMC.

<table>
<thead>
<tr>
<th>Population</th>
<th>$\hat{\alpha}$</th>
<th>$\hat{\beta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Grid</strong></td>
<td>$0.543 \pm 0.084$</td>
<td>$5.02 \pm 0.5165$</td>
</tr>
<tr>
<td><strong>Bivariate normal</strong></td>
<td>$0.512 \pm 0.114$</td>
<td>$4.82 \pm 0.467$</td>
</tr>
<tr>
<td><strong>Grid</strong></td>
<td>$0.553 \pm 0.095$</td>
<td>$5.076 \pm 0.566$</td>
</tr>
<tr>
<td><strong>Bivariate normal</strong></td>
<td>$0.524 \pm 0.126$</td>
<td>$4.85 \pm 0.510$</td>
</tr>
</tbody>
</table>

intervals, as shown in Table 3.2. The variance for $\alpha$ is slightly higher again in the bivariate normal population than in the grid population, and the opposite is again true for $\beta$. We also observe slightly higher variances here in the SIR framework than in the SI framework.

### 3.3 Estimating the infectious period in the SIR model

In this sub-study, we estimate model parameters, now including the infectious period, in two different population settings: individuals on a grid, and clusters of individuals simulated from bivariate normal distributions, using Bayesian MCMC and a method based on maximum likelihood estimation. In our Bayesian analysis, we adopt a discrete uniform prior distribution for the infectious period, $\gamma$, which reflects the number of days for which an individual was infectious. In our maximum likelihood investigation, we use a method based on profile likelihood to estimate $\gamma$. 
3.3.1 Maximum Likelihood-Based Estimation

Instead of estimating $\gamma$ using the traditional maximum likelihood method, we obtain an estimate of $\gamma$ by substituting the MLEs $\hat{\alpha}$ and $\hat{\beta}$ into the log-likelihood under a pre-specified value of $\gamma$. For example, suppose there are $k$ epidemics. If epidemic $k$ is generated under an infectious period of 1 day, we substitute $\hat{\alpha}$ and $\hat{\beta}$ into the log-likelihood, specify a value of $\gamma = 1$, and calculate the value of the log-likelihood. Then, for the same epidemic, we specify a value of $\gamma = 2$ and calculate the log-likelihood using those same estimates $\hat{\alpha}$ and $\hat{\beta}$. We proceed until we have exhausted the values of $\gamma$ from the set of values under consideration. In this way, we obtain $n$ values of the log-likelihood for each epidemic, where $n$ is the range of $\gamma$ values tested. To estimate $\gamma$ for epidemic $k$, we take the maximum value of the log-likelihood over the set of $n$ log-likelihoods and obtain the associated value of $\hat{\gamma}$. The mode of the maximum values of the log-likelihood for all $r$ epidemics with a true value of $\gamma = 1$ in the data generation gives us an estimate of $\gamma$ that is dependent on the MLEs of $\alpha$ and $\beta$. If we use the same value for $\gamma$ in the log-likelihood that we used to generate the epidemic data, we would expect that value of $\gamma$ to yield the greatest log-likelihood value. The algorithm is outlined below for a generic $\gamma \sim DU[1,n]$ distribution, where the true value of $\gamma$ is denoted $\gamma_T$.

Algorithm 3: Maximum Likelihood-Based Estimation of $\gamma$

1. For $i = 1$ to $n$
   2. For $j = 1$ to $r$
      3. Generate an epidemic $k$ under $\gamma_T = i$
      4. Estimate $\alpha$ and $\beta$ using the BFGS algorithm for epidemic $k : (\hat{\alpha}, \hat{\beta})$
      5. For $m = 1,n$
         6. Calculate $l_{\gamma=1}(\hat{\alpha}, \hat{\beta})$ for each epidemic $k$
7. End for
8. \( \hat{\gamma}_{ij} = \arg\max_{\gamma} \{ l_{\gamma=1}(\hat{\alpha}, \hat{\beta}), \ldots, l_{\gamma=n}(\hat{\alpha}, \hat{\beta}) \} \)
9. End for
10. \( \hat{\gamma}_i = \text{mode}(\arg\max_{\gamma} \{ l_{\gamma=1}(\hat{\alpha}, \hat{\beta}), \ldots, l_{\gamma=n}(\hat{\alpha}, \hat{\beta}) \}) \)
11. End for

where

- \( n \) is the maximum of the range of \( \gamma \) values under consideration;
- \( r \) is the number of replicates per setting of \( \gamma \);
- \( \hat{\gamma}_{ij} \) is the \( k^{th} \) estimate of \( \gamma \), for replicate \( j \) when \( \gamma_T = i \) in the epidemic generation step, and
- \( \hat{\gamma}_i \) is the mode of the estimates from the \( r \) replicates - the overall estimate of \( \gamma \) when \( \gamma_T = i \).

The asymptotic normality property of the MLE is dependent upon the ability to take two derivatives of the likelihood function with respect to the parameter in question. In particular, in the definition of asymptotic normality of the MLE,

\[
\sqrt{n}(\hat{\gamma} - \gamma_0) \xrightarrow{D} N\left(0, \frac{1}{I(\gamma_0)} \right)
\]

\( I(\gamma_0) \) is the Fisher Information. The Fisher Information is the negative expectation of the second derivative of the log-likelihood function with respect to \( \gamma \). However, we cannot obtain \( I(\gamma_0) \), since we cannot differentiate the likelihood function in terms of \( \gamma \). Therefore, for our ML-like estimates for \( \gamma \), we cannot assume asymptotic normality and hence cannot derive confidence intervals for \( \gamma \) in this manner.
### Table 3.3: MLE estimates of $\gamma$ in the grid ($\hat{\gamma}_{GRID}$) and bivariate normal ($\hat{\gamma}_{BIVAR}$) populations, for $\gamma = 1, \ldots, 10$. $p$ is the proportion of times $\hat{\gamma}_{ij} = \gamma_T$, as discussed in Algorithm 3.

<table>
<thead>
<tr>
<th>$\gamma_{GRID}$</th>
<th>1, 2, 10</th>
<th>2, 10</th>
<th>2</th>
<th>4, 6, 10</th>
<th>10</th>
<th>10</th>
<th>4, 5</th>
<th>3</th>
<th>4</th>
<th>(p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\gamma}_{GRID}$</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>0.11</td>
</tr>
<tr>
<td>$\hat{\gamma}_{BIVAR}$</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>0.16</td>
</tr>
</tbody>
</table>

**Results**

The value of $r$ in Algorithm 3 was taken to be 10. For example, Table 3.3 shows estimates of $\gamma$, $\hat{\gamma}_{GRID}$ and $\hat{\gamma}_{BIVAR}$, for various true values of $\gamma$ in the grid and bivariate normal populations. When $\gamma_{BIVAR} = 1$, $\gamma_{BIVAR} = \hat{\gamma}_i = 10$, the value of $\gamma$ that most frequently yielded the maximum value of the log-likelihood function after substituting in $\hat{\alpha}$ and $\hat{\beta}$ over all 10 epidemics in the bivariate normal population for which the true value of $\gamma_{BIVAR} = 1$. When $\gamma_{GRID} = 1$, the algorithm was equally likely to select values of $\hat{\gamma}_{GRID} = 1, 2,$ and 10.

This method of estimation did not work well for the bivariate normal population. In fact, the maximum likelihood-based estimation procedure consistently estimated the value of $\gamma$ to be the maximum of the range of $\gamma$ considered. It was rare for $\hat{\gamma}_{ij}$ to be any value other than the right endpoint. For $\gamma_{BIVAR}$ in Figure 3.20, only when $\gamma_{BIVAR} = 5, 7, \text{or } 10$ does $\hat{\gamma}_i$ ever agree with $\gamma_T$. Only when $\gamma_{BIVAR} = 10$ do we see that the mode of $\hat{\gamma}_i$ agrees with $\gamma_{BIVAR}$. Since $\hat{\gamma}_i$ is most frequently chosen to be the highest value in the distribution of $\gamma$, we may suspect that the algorithm is biased to select the maximum.

In the grid population, we do not observe the same tendency to overestimate $\gamma_{GRID}$ (Figure 3.21). However, we also observe many ties. For example, in Table 2, we see that for $\gamma_{GRID} = 1$, the estimation procedure procedure resulted in estimates of
\( \gamma_{\text{GRID}} = 1, 2, 10 \). These values of \( \gamma \) would result in vastly different epidemics, so it is surprising to see this result. To illustrate this effect, Figure 3.19 shows ten simulated epidemics per value of \( \gamma_{\text{GRID}} = 1, 2, 10 \). The average number of infectious individuals is shown in blue. The epidemic curves each have a different maximum, although from these plots it appears that an epidemic with an infectious period of 2 days is similar to one with an infectious period of 10 days. Given the difference in shape of the epidemic curves between \( \gamma_{\text{GRID}} = 1 \) and \( \gamma_{\text{GRID}} = 2 \) or \( \gamma_{\text{GRID}} = 10 \), we would not expect the algorithm to estimate, for example, a true value of \( \gamma_{\text{GRID}} = 1 \) with a \( \hat{\gamma}_{\text{GRID}} = 10 \).

The value of \( p \) in Table 3.3 is the proportion of times the algorithm correctly estimated the true value of \( \gamma \). From this value, we see that the performance of the estimation procedure was better overall for the grid population than the bivariate normal population. This could be because in the grid population, it is difficult to see a difference in epidemic curves generated using an infectious time of, for example, \( \gamma = 8 \) and \( \gamma = 10 \). It is possible that the same effect is not present in the bivariate normal population. To examine this effect further, Figures 3.17 and 3.18 show the differences between epidemics with values of \( \gamma = 8 \) and \( \gamma = 10 \) for the grid and bivariate normal populations, respectively. While the small difference in \( \gamma \) under the grid population showed few differences between epidemics, the epidemics in the bivariate normal population were visibly different. The maximum number of infectious individuals occur at similar points in time for both populations (a maximum of 65 infectious individuals on day 11 when \( \gamma_{\text{GRID}} = 8 \) and 79 individuals on day 13 when \( \gamma_{\text{GRID}} = 10 \); 79 individuals on day 4 when \( \gamma_{\text{BIVAR}} = 8 \) and 105 on day 6 when \( \gamma_{\text{BIVAR}} = 10 \)). From these plots it appears that when we observe epidemics in a bivariate normal population with little distance between individuals, small changes in the infectious period can lead to very different epidemics. Therefore the diffi-
culty in estimating $\gamma_{BIVAR}$ is not due to the similarity in epidemic curves between values of the infectious period.

Figure 3.17: Epidemics generated in the grid population under (a) $\gamma_{GRID} = 8$ and (b) $\gamma_{GRID} = 10$, with average number of infectious individuals per time step in blue.

Figure 3.18: Epidemics generated in the bivariate normal population under (a) $\gamma_{BIVAR} = 8$ and (b) $\gamma_{BIVAR} = 10$, with average number of infectious individuals per time step in blue.
Figure 3.19: Epidemics generated using $\gamma_{GRID} = 1, 2,$ and 10, respectively, with average number of infectious individuals per time step in blue.
Figure 3.20: ML Estimates \( \hat{\gamma}_{BIVAR} \) under \( \gamma_{BIVAR} = 1 \ldots 10 \). True value of \( \gamma_{BIVAR} \) is in grey.
Grid population
true gamma is 1

Chosen value of gamma

Frequency

0 1 2 3 4
1 3 5 7 9

Grid population
true gamma is 2

Chosen value of gamma

Frequency

0 0.5 1.0 1.5 2.0 2.5 3.0
1 3 5 7 9

Grid population
true gamma is 3

Chosen value of gamma

Frequency

0 0.5 1.0 1.5 2.0 2.5 3.0
1 3 5 7 9

Grid population
true gamma is 4

Chosen value of gamma

Frequency

0 0.5 1.0 1.5 2.0 2.5 3.0
1 3 5 7 9

Grid population
true gamma is 5

Chosen value of gamma

Frequency

0 0.5 1.0 1.5 2.0 2.5 3.0
1 3 5 7 9

Grid population
true gamma is 6

Chosen value of gamma

Frequency

0 0.5 1.0 1.5 2.0 2.5 3.0
1 3 5 7 9

Grid population
true gamma is 7

Chosen value of gamma

Frequency

0 0.5 1.0 1.5 2.0 2.5 3.0
1 3 5 7 9

Grid population
true gamma is 8

Chosen value of gamma

Frequency

0 0.5 1.0 1.5 2.0 2.5 3.0
1 3 5 7 9

Grid population
true gamma is 9

Chosen value of gamma

Frequency

0 0.5 1.0 1.5 2.0 2.5 3.0
1 3 5 7 9

Grid population
true gamma is 10

Chosen value of gamma

Frequency

0 0.5 1.0 1.5 2.0 2.5 3.0
1 3 5 7 9

Figure 3.21: ML estimates $\hat{\gamma}_{\text{GRID}}$ under $\gamma_{\text{GRID}} = 1, \ldots, 10$. True value of $\gamma_{\text{GRID}}$ is in grey.
3.3.2 Bayesian MCMC Estimation

We now consider a Bayesian analysis in which all three model parameters are treated as random variables. We set the prior of the infectious period to follow a $DU[1, 10]$ distribution, and estimate $\gamma$ via Bayesian Markov Chain Monte Carlo.

Under a distribution of $\gamma_{GRID} \sim DU[1, 10]$ (Figure 3.22), we see that smaller values of $\gamma_{GRID}$ are easier to estimate than higher values. This estimation difficulty agrees with the issue seen in the maximum likelihood-based estimation procedure.

Interestingly, the bivariate normal population exhibits the opposite problem to the grid population: it is easier to estimate higher values of $\gamma_{BIVAR}$ (Figure 3.23), and there is a marked tendency to overestimate $\gamma$ to be its maximum value. However, in Figure 3.23 we can also see that the Bayesian MCMC algorithm always includes some correct $\hat{\gamma}_{ij}$ values for each true value $\gamma_{BIVAR}$. Furthermore, from Table 3.4, we observe a higher proportion of correct estimates of $\gamma_{GRID}$ under MCMC ($p = 0.32$) estimation than we do under MLE ($p = 0.16$). Similar to the ML estimates, we see that the log-likelihood under very different values of $\gamma$ resulted in the same log-likelihood. For example when $\gamma_{GRID} = 6$, $\hat{\gamma}_{GRID} = 7, 10$ implies that the log-likelihoods were equal for values of 7 and 10.

In the bivariate normal population, individuals have many close neighbours. A spatially-based epidemic would consequently spread through this population very quickly. From these plots, it appears that when the distance between individuals is small, there is little information available about $\gamma$ to inform the estimation procedure. That is, the speed of the epidemic progression is not heavily influenced by the value of $\gamma_{BIVAR}$. Overestima-
Table 3.4: Bayesian MCMC estimates of $\gamma$ in the grid ($\hat{\gamma}_{GRID}$) and bivariate normal ($\hat{\gamma}_{BIVAR}$) populations, when $\gamma \sim DU[1,10]$. $p$ is the proportion of times $\hat{\gamma}_{ij} = \gamma_T$, as discussed in Algorithm 3.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\gamma}_{GRID}$</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3,10</td>
<td>3</td>
<td>7,10</td>
<td>10</td>
<td>8</td>
<td>2,4,6</td>
<td>4,7</td>
<td>0.32</td>
</tr>
<tr>
<td>$\hat{\gamma}_{BIVAR}$</td>
<td>9</td>
<td>10</td>
<td>9</td>
<td>10</td>
<td>7,9</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>0.11</td>
</tr>
</tbody>
</table>

The epidemic curves for $\hat{\gamma}_{BIVAR} = 8$ and $\hat{\gamma}_{BIVAR} = 10$ were very different, the epidemic datasets that was used to generate the curves contained similar amounts of information.
Posterior estimates of $\gamma$ when $\gamma = 1$

<table>
<thead>
<tr>
<th>Value of $\gamma$</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0e+00</td>
</tr>
<tr>
<td>4</td>
<td>4e+04</td>
</tr>
<tr>
<td>6</td>
<td>8e+04</td>
</tr>
</tbody>
</table>

Figure 3.22: MCMC estimates $\hat{\gamma}_{GRID}$ under $\gamma_{GRID} = 1, \ldots, 10$. True value of $\gamma_{GRID}$ is in grey.
Figure 3.23: MCMC estimates $\hat{\gamma}_{BIVAR}$ under $\gamma_{BIVAR} = 1, \ldots, 10$. True value of $\gamma_{BIVAR}$ is in grey.
3.4 Misspecification of Infectious Time in SIR Model

In addition to estimating the infectious period, we also explore the effect of model mis-specification on estimates of the susceptibility and transmissibility parameters. By pre-specifying the infectious period to be, say, 2 days in the epidemic generation phase, and 5 days in the estimation phase, we can assess how misspecification of the infectious period can affect estimation of the other model parameters. We carry this out using both maximum likelihood and Bayesian MCMC, and for both the grid and the bivariate normal populations, for values of $\gamma = [1 \ldots 10]$. In previous work by Habibzadeh and Deardon (2010), and Gardner, Deardon, and Darlington (2011), Bayesian MCMC methods were also used; however, in Habibzadeh and Deardon (2010), SEIR was the compartmental model of interest, and in Gardner et al. (2011), inference centered around statistics for posterior predictive goodness-of-fit tests. We expect that as the difference between the true value of $\gamma$ and the misspecified value increases, the estimates of $\alpha$ and $\beta$ will deviate further from the true values.

By examining the effect of misspecifying the infectious period, $\gamma$, in the SIR model, we can further explore the robustness of ML and MCMC estimation. When the infectious period is misspecified to be lower than the truth, we expect the susceptibility parameter to be overestimated and/or the spatial transmissibility to be underestimated. Conversely, if the infectious time is misspecified to be higher than the truth, we may observe lower estimates of $\alpha$ and higher estimates of $\beta$.

An underestimated value of $\beta$ results in a decreased spatial effect, which leads to more random-looking patterns. A disease with a lower value of $\beta$ would also progress more slowly, because the transmissibility from infectious to susceptible individuals would
be lower. Accordingly, we can expect that when \( \gamma \) is misspecified to be lower than the truth in the grid population, the shape of the epidemic curve would resemble Figure 3.9(a), so the value of \( \beta \) will be underestimated to compensate. Consequently, in the grid population, when \( \gamma \) is misspecified to be higher than the truth, the expected epidemic curve would be similar in shape to 3.9(b), that is, the epidemic would spread faster, and the spatial infectiousness would be overestimated.

### 3.4.1 Misspecification of SIR model in grid population

**Maximum Likelihood Estimation**

As before, \( \alpha \) is overestimated when the true value of \( \gamma_{GRID} = 1 \), and estimated values decrease as the value of the misspecified \( \gamma \) increases (Figure 3.24(a)), and the same overestimation (although to a lesser extent) was observed for \( \gamma = 10 \) (Figure 3.24(b)). Contrastingly, \( \beta \) was only underestimated when the true value of \( \gamma_{GRID} \) was 10 and the misspecified value was 1 (Figure 3.25).
Figure 3.24: ML estimates of alpha (susceptibility) with associated 95% confidence intervals, when the true infectious period is 1 day (a) and 10 days (b) in the grid population. The straight red line is the true value and the straight blue line is the average parameter estimate.
Figure 3.25: ML estimates of beta (spatial transmissibility) with associated 95% confidence intervals, when the true infectious period is 1 day (a) and 10 days (b) in the grid population. The straight red line is the true value and the straight blue line is the average parameter estimate.
Bayesian MCMC Estimation

The posterior estimates of $\alpha$ were significantly affected by a high misspecification value of $\gamma_{GRID}$. The estimated value of $\alpha$ decreased as the misspecification value increased. As in our SIR model observations, the value of $\alpha$ was overestimated when $\gamma_{GRID} = 1$. However, we see that $\alpha$ is underestimated for larger values of $\gamma_{GRID}$. An underestimated value of the susceptibility parameter indicates a slower epidemic; that is, the epidemic would not spread as readily with a decreased value of $\alpha$. In Figure 3.26(b), we see the same is not true for a value of $\gamma_{GRID} = 10$ when $\gamma$ is misspecified; that is, we see an overestimation of $\alpha$ when $\gamma_{GRID}$ is misspecified to be 1 and 2, but the estimates of $\alpha$ are very similar for all other misspecified values of $\gamma$.

Contrastingly, the MCMC estimates of $\beta$ are not severely affected by misspecification of the infectious period. In fact, when $\gamma_{GRID} = 1$ (Figure 3.27(a)), the value of $\beta$ is largely the same regardless of the misspecification of $\gamma_{GRID}$. However, when the true value of $\gamma_{GRID}$ is increased to 10, $\beta$ is underestimated when $\gamma_{GRID} = 1$ and 2. There is one extreme observation when $\gamma_{GRID}$ is misspecified to be 1 that pulls the mean closer to the true value; however, we can see that most of the MCMC estimates are far below the true $\beta$ value.
Figure 3.26: MCMC estimates of $\alpha$ (susceptibility) with associated 95% credible intervals, when $\gamma_T = 1$ (a) and $\gamma_T = 10$ (b) in the grid population. The straight red line is the true value and the straight blue line is the average parameter estimate.

Figure 3.27: MCMC estimates of $\beta$ (spatial transmissibility) with associated 95% credible intervals, when $\gamma_T = 1$ (a) and $\gamma_T = 10$ (b) in the grid population. The straight red line is the true value and the straight blue line is the average parameter estimate.
3.4.2 Misspecification of SIR model in bivariate normal population

Maximum Likelihood Estimation

The misspecification of the infectious period had little effect on the estimation of $\alpha$ in the bivariate normal population (Figure 3.28(a)). As we might expect, $\beta$ was still underestimated when $\gamma_{BIVAR}$ was set to its highest value and misspecified in the model to its lowest value (Figure 3.29(b)).

Figure 3.28: ML estimates of alpha (susceptibility) with associated 95% confidence intervals, when the true infectious period is 1 day (a) and 10 days (b) in the bivariate normal population. The straight red line is the true value and the straight blue line is the average parameter estimate.
Figure 3.29: ML estimates of beta (spatial transmissibility) with associated 95% confidence intervals, when the true infectious period is 1 day (a) and 10 days (b) in the bivariate normal population. The straight red line is the true value and the straight blue line is the average parameter estimate.

### Bayesian MCMC Estimation

The posterior estimates for $\alpha$ and $\beta$ are consistent in the bivariate normal population regardless of the misspecified value of $\gamma$ (Figures 3.30 and 3.31). This is unexpected, as when the infectious period is misspecified to be higher than the true value, the spatial transmissibility parameter should be overestimated to compensate. However, it is possible that the same effect is present in the SIR discussion is also at work here: there is an insufficient difference in the amount of information available under the epidemic curves to inform estimation procedures.
Figure 3.30: MCMC estimates of $\alpha$ (susceptibility) with associated 95% credible intervals, when $\gamma_T = 1$ (a) and $\gamma_T = 10$ (b) in the bivariate normal population. The straight red line is the true value and the straight blue line is the average parameter estimate.

Figure 3.31: MCMC estimates of $\beta$ (spatial transmissibility) with associated 95% credible intervals, when $\gamma_T = 1$ (a) and $\gamma_T = 10$ (b) in the bivariate normal population. The straight red line is the true value and the straight blue line is the average parameter estimate.
In theory, we expect the value of $\beta$ to increase when $\gamma$ is misspecified to be higher than its true value; however, we do not see evidence to support such a statement. Additionally, we expect to see the value of $\alpha$ decrease when the misspecified value of $\gamma$ is increased, and we only observe this effect in the grid population.
Chapter 4

Conclusion

We find that regardless of the method of estimation, maximum likelihood or Bayesian Markov Chain Monte Carlo, estimating the infectious period is difficult, and often results in the wrong value of $\gamma$ being chosen. We observe that it is easier to estimate high values of $\gamma$ in the bivariate normal population, but that the estimate is very frequently equal to the maximum value. It is possible that in the bivariate normal population, the epidemics progress too quickly to allow for accurate estimation of the infectious period. In that case, not enough information is available under the epidemic curve to sufficiently inform estimation procedures, and the information that is available is similar for different values of $\gamma_{BIVAR}$. The opposite is true in the grid population: it is easier to estimate lower values of $\gamma_{GRID}$. The topology of the population under consideration clearly affects parameter estimation.

Misspecification of the infectious period did not have as pronounced an effect on estimation of the susceptibility and spatial transmissibility parameters as we expected. Although the value of $\alpha$ was overestimated when the true value of $\gamma_{GRID} = 1$ and under-
estimated to a larger and larger extent as the misspecified value of $\gamma_{\text{GRID}}$ was increased, this effect was not as visible under a true infectious period of 10 in the grid population. Furthermore, in the bivariate normal population, there was no significant difference in the estimates of $\alpha$ and $\beta$ under model misspecification. We suspect this is due to the proximity of individuals in the population.

A first glance indicates that the frequentist approach to parameter estimation through MLE may yield more stable estimates regardless of the infectious period, as the 95% confidence intervals are narrower than the 95% credible intervals. However, the traditional Wald-type confidence intervals used in the MLE discussion may not be optimal here. To make use of these traditional confidence intervals, we must have a good estimate of the standard error, and we must have estimates that are sufficiently far from the boundary of their range (Stryhn & Christensen, 2003). The BFGS algorithm as implemented in R for maximum likelihood estimation does not correct the variance-covariance matrix for constrained optimization; therefore, the estimates of the standard error may not be accurate. Even with such lax constraints on the parameters, we may not have good estimates of the standard error. Additionally, many of the estimates of $\gamma$ are close to the boundary since we are assuming a relatively small parameter space. Alternative confidence intervals could be derived from profile likelihood or by using parametric bootstrapping. This latter technique would require on the order of 1000 simulations, though, which results in a high computational burden.

Future work could include an investigation of the coverage error of the three types of confidence interval in spatial epidemics. It would also be interesting to investigate different true parameter values allowing for longer-lasting epidemics. For example, increas-
ing the transmissibility parameter would decrease the probability of infection, allowing for epidemics that take longer to spread through the population. The resulting epidemic curves may then contain more information that can be used to inform estimation of the infectious period. Furthermore, different population layouts can be constructed to further examine the effect of spatial proximity between individuals. For instance, the gap between positions of individuals on a grid could be increased, or the variance, position, and/or number of clusters in the bivariate normal population could be varied.
References


