Incorporating Contact Network Uncertainty in Individual Level Models of Infectious Disease within a Bayesian Framework

by

Waleed Almutiry

A Thesis
presented to
The University of Guelph

In partial fulfilment of requirements
for the degree of
Doctor of Philosophy
in
Statistics

Guelph, Ontario, Canada

©Waleed Almutiry, August, 2018

ABSTRACT

INCORPORATING CONTACT NETWORK UNCERTAINTY IN INDIVIDUAL LEVEL MODELS OF INFECTIOUS DISEASE WITHIN A BAYESIAN FRAMEWORK

Waleed Almutiry
University of Guelph, 2018

Advisor:
Rob Deardon & Zeny Feng

Individual-level infectious disease models enable the study of transmission mechanisms of infectious disease while accounting for heterogeneity within the population. As well as via covariates, such heterogeneity is often best modelled through a contact network or a series of networks. However, contact network and exact time of infection (and removal) for individuals are often completely or partially unobserved.

In this thesis, we account for such data uncertainty through the incorporation of a large amount of missing information (contact network and event times) into a continuous time individual-level modelling framework. The main focus of this thesis is to consider the effect of incorporating contact network uncertainty on the performance of the models within a Bayesian framework. A secondary focus will be on enabling fast inference to fit these models to epidemic data sets with a large amount of missing information.

We start by introducing our R package *EpiILMCT* that allows users to study the spread of infectious disease using spatial- or (and) network-based continuous-time ILMs. We then investigate the performance of network-based ILMs in analyzing small epidemic data sets

under different levels of contact network uncertainty, along with uncertainty in individual-level event histories, using data augmented Markov chain Monte Carlo (MCMC). We also consider the incorporation about global-level contact network information through observation models that are either based upon knowledge about the degree distribution or the total number of connections in the network. Then, we explore the use of approximate Bayesian computation population Monte Carlo (ABC-PMC) methods for fitting such models to both simulated data, and data from the UK 2001 foot-and-mouth disease epidemic. Finally, we introduce an approach to approximate full model inference by partitioning the population into a number of spatial clusters in which the contact network is divided into a series of isolated cluster-constrained sub-networks.

In memory of my mother

To my father

To my wife, Lobnah

To my children:

Wasan, Hatim, Hams, Abdulaziz, and Wateen

ACKNOWLEDGEMENTS

First and foremost, I praise and thank God (Allah) for giving me the opportunity, knowledge, stamina and blessings to complete this work. The completion of this thesis would not have been possible without the support and encouragement of several special people. Hence, I would like to take this opportunity to show my gratitude to those who have assisted me in a myriad of ways.

I would like to express my profound gratitude to my advisor Prof. Rob Deardon. It was a great privilege and honor to work and study under his supervision. I am extremely grateful for what he has offered me. He has been a constant source of guidance, patience and support over the last five years. He has inspired me by his dynamism, vision, sincerity, motivation, and trust that helped me through the PhD program in Applied Statistics. I would also like to thank him for his friendship, empathy, and great sense of humor.

I would also like to thank Dr. Zeny Feng, my former committee member and co-advisor for her encouragement and support. I am also thankful to Prof. David Pearl and Prof. Ayesha Ali for being members of my PhD committee, and for their continued faith and encouragement.

I would like also to thank all the Mathematics and Statistics staff at Guelph University for all their administrative help through these years, especially Ms. Susan McCormick, the graduate program assistant, for her administrative assistance while I was visiting the University of Calgary. I would like to thank the Research Computing Services at the University of Calgary for providing access to the Helix computational cluster, and for their help in using the cluster.

I would like to thank my family, to whom I owe a great deal. To my late mother, Baroudah: thank you. You are gone, but your belief in me has made this journey possible. To my father, Khalaf, for his never-ending support, encouragement and generosity, as well as for imparting his wisdom upon me and for teaching me patience.

A special thank goes to my beloved wife, Lobnah for her love, great patience and

support in every single moment of my life during the past 14 years. She has devoted her time and effort towards helping me achieve success. Also, special thanks go to my wonderful children, Wasan, Hatim, Hams, Abdulaziz and Wateen for their support and being patient with me while I have been busy completing my doctoral studies the last couple of years. I thank you all for always making me smile and for understanding my absences on those weekends and holidays when I was working on my research instead of playing games with you. I hope that one day you will understand why I spent so much time in front of my computer. Also, I thank all my brothers and sisters for their moral support, especially Mohammed and Ali.

I would like to acknowledge the financial support of Qassim University in Saudi Arabia for the award of Foreign Postgraduate scholarship that provided the necessary financial support for my entire degree. Also, I am thankful to all the staff of the Saudi Cultural Bureau in Canada, for all the help, support, and facilities provided to me during my studies.

Finally, I would like to extend my sincere gratitude to everyone who has directly or indirectly supported me to complete my research work.

Table of Contents

Abstract ii

Acknowledgements v

List of Tables xi

List of Figures xii

1 Introduction 1
- 1.1 Introduction to Infectious Disease Models . . . 1
- 1.2 Individual-Level Infectious Disease Models . . . 4
- 1.3 Spatio-Temporal Infectious Disease Models . . . 5
- 1.4 Contact Network in Infectious Disease Models . . . 7
- 1.5 Parameter Estimation in a Bayesian Framework . . . 10
  - 1.5.1 Bayesian Inference . . . 11
- 1.6 Markov Chain Monte Carlo (MCMC) Methods . . . 13
  - 1.6.1 Gibbs Sampler . . . 14
  - 1.6.2 Metropolis Hastings (MH) MCMC . . . 15
  - 1.6.3 Data Augmentation . . . 17
- 1.7 Approximate Bayesian Computation Method . . . 18
- 1.8 Thesis Overview . . . 21

2 Continuous Time Individual-Level Models of Infectious Disease: *EpiILMCT* 25
- 2.1 Introduction . . . 25
- 2.2 The Model . . . 31
  - 2.2.1 Likelihood Function . . . 35
- 2.3 Contents of the EpiILMCT Package . . . 38
  - 2.3.1 Contact Network . . . 39
  - 2.3.2 Epidemic Simulation . . . 42
- 2.4 Bayesian Inference . . . 45

2.5 Examples . . . 49
2.5.1 Simulated Network-Based Epidemic . . . 49
2.5.2 Case Study: Tomato Spotted Wilt Virus (TSWV) Data . . . 59
2.6 Conclusion . . . 64

3 Spatial Contact Network Uncertainty in Individual Level Models of Infectious Disease Transmission 69
3.1 Introduction . . . 69
3.2 The Model . . . 73
3.2.1 The General Continuous Time ILMs . . . 73
Likelihood Function . . . 74
The Network-Based ILM . . . 76
3.2.2 Spatial Contact Network and Epidemic Simulation . . . 77
3.3 Bayesian Inference . . . 79
3.3.1 The Posterior Distribution . . . 79
3.3.2 MCMC Method . . . 80
Updating I_i and γ_i . . . 81
Updating c_{ij} . . . 81
3.3.3 Model Validation . . . 82
3.4 Simulation Study . . . 83
3.5 Results . . . 85
3.5.1 Network Uncertainty . . . 85
3.5.2 The Contact Network Observation Model Effect . . . 87
3.6 Discussion . . . 93

4 Incorporating Contact Network Uncertainty in Individual Level Models of Infectious Disease using Approximate Bayesian Computation 102
4.1 Introduction . . . 102
4.2 Methodology . . . 106
4.2.1 The General Continuous Time ILMs . . . 106
Likelihood Function . . . 107
The Network-Based ILM . . . 109
4.2.2 Contact Networks . . . 111
4.2.3 Epidemic Simulation . . . 112
4.3 Model Fitting . . . 113
4.3.1 Approximate Bayesian Computation . . . 114
ABC-PMC Algorithm . . . 116
4.3.2 Data Augmentation MCMC . . . 118
4.4 Epidemic Data and Summary Statistics . . . 119
4.4.1 Simple Network-Based Data . . . 119
4.4.2 The UK 2001 Foot-and-Mouth Disease . . . 122
Simulated FMD Data . . . 123

Real FMD Data . . . 125
4.5 Results . . . 127
4.5.1 Simple Network-Based ILM Data . . . 128
4.5.2 Simulated FMD Data . . . 131
4.5.3 Real FMD Data . . . 132
4.5.4 Computation Time . . . 137
4.6 Discussion . . . 138

5 Contact Network Uncertainty in Individual Level Models of Infectious Disease Transmission Using Within-Cluster Likelihood Approximations 142
5.1 Introduction . . . 142
5.2 The Model . . . 147
5.2.1 The Network-Based Continuous Time ILMs . . . 148
5.2.2 Within-Cluster Network-Based Continuous Time ILMs . . . 149
5.2.3 Likelihood Function . . . 150
5.2.4 Epidemic Simulation . . . 153
5.2.5 Data Clustering . . . 154
5.3 Bayesian Inference . . . 155
5.3.1 The Posterior Distribution . . . 156
5.3.2 Modified Likelihood Inflating Sampling Algorithm (LISA) . . . 156
5.3.3 MCMC Method . . . 159
5.3.4 Simulated Epidemics . . . 161
5.4 Results . . . 164
5.4.1 Well-Clustered Populations (Scenario 1) . . . 164
5.4.2 Scattered Populations (Scenarios 2 and 3) . . . 167
5.4.3 Comparison to LISA . . . 170
5.4.4 The Effect of Increasing Number of Clusters . . . 173
5.4.5 Computation Time and Parallelization . . . 176
5.5 Discussion . . . 178

6 Conclusions 182
6.1 Continuous Time Individual-Level Models of Infectious Disease: *EpiILMCT* 183
6.1.1 Summary . . . 183
6.1.2 Future Work . . . 184
6.2 Spatial Contact Network Uncertainty in Individual Level Models of Infectious Disease Transmission . . . 185
6.2.1 Summary . . . 185
6.2.2 Future Work . . . 186
6.3 Incorporating Contact Network Uncertainty in Individual Level Models of Infectious Disease Using Approximate Bayesian Computation . . . 186
6.3.1 Summary . . . 186
6.3.2 Future Work . . . 187

6.4 Contact Network Uncertainty in Individual Level Models of Infectious Disease Transmission Using Within-Cluster Likelihood Approximations 189
6.4.1 Summary 189
6.4.2 Future Work 191

Bibliography **192**

A Continuous Time Individual-Level Models of Infectious Disease: *EpilLMCT* **202**
A.1 R Code of Extracting Individual Level Data from the *surveillance* Package 202

B Incorporating Contact Network Uncertainty in Individual Level Models of Infectious Disease using Approximate Bayesian Computation **210**
B.1 The 2001 UK Foot-and-Mouth Disease Epidemic in the Cumbria Region . . 211
B.2 Samples of the Estimated Degree Distributions of Fitting Simple Network-Based ILM 212
B.3 Samples of the Approximated Posterior Distribution of the Epidemic Curve of the Infection Times 213
B.4 Samples of the Intermediate Marginal Posterior Distributions for the Model Parameters 215

C Contact Network Uncertainty in Individual Level Models of Infectious Disease Transmission Using Within-Cluster Likelihood Approximations **217**
C.1 Summaries of the Cluster Data Sets 218
C.2 The Estimated Infection Times and Degree Distributions 221
C.3 A Comparison of the Performance of the Original and Modified LISA. . . . 225
C.4 The Effect of Increasing Number of Clusters 229

List of Tables

2.1 Types of kernel functions that are applied in the *EpiILMCT* package for fitting continuous time ILMs. 35
2.2 Description of functions and their usages in the *EpiILMCT* package. . . 38
2.3 The posterior means and 95% credible intervals (CIs) of the model parameters for fitting the TSWV data set using the $\mathcal{SINR}$ distance-based continuous time ILM under the assumption of unknown removal and infection times. 64

4.1 The sets of summary statistics used, where $P_{(\bullet)}$ denotes the time at the peak of the epidemic curve (R removal and I infection times), n_I is the number of infected individuals, L_{epi} is the length of epidemic, $\mu_{(\bullet)}$ and $\sigma_{(\bullet)}$ are the mean and standard deviation of the infection or removal times, $\mathcal{I}_0^T$ is the number of infection events over equal time intervals; and $\chi^2_{s,c}$ is the chi-squared statistic as described in Section (4.4.2). 127
4.2 Posterior means and 95% credible intervals for parameters for the FMD network-based ILM fitted to the real FMD epidemic data. 134
4.3 The computational time (in minutes) of using both DA-MCMC and ABC-PMC on fitting network-based ILMs for a simple network-based data ($N = 25$) and a simulated FMD epidemic data ($N = 674$). 138

C.1.1 Summaries of the ten epidemics data sets spreading on well-clustered populations. The number of clusters in this case is ten clusters. 218
C.1.2 Summaries of the ten epidemics data sets with binary covariate spreading on scattered populations. The number of clusters in this case is ten clusters. 219
C.1.3 Summaries of the ten epidemics data sets with discrete covariate spreading on scattered populations. The number of clusters in this case is ten clusters. 220

List of Figures

2.1 Samples of the three undirected binary contact network models generated to the same population. Red dots represent nodes with different sizes that correspond to their degree (number of edges connected to a node). 41
2.2 The epidemic curves of the infection and removal times for the epidemic that was generated using the simple network-based continuous time ILM. The red shaded area in the third plot represent the infectious periods. 52
2.3 The directed pathway network of the generated epidemic over its timeline using the simple network-based ILMs. 53
2.4 The MCMC chains of the posterior distributions of the model parameters fitting the simulated epidemic using the network-based continuous time ILM assuming a fully observed epidemic. 54
2.5 The MCMC chains of the posterior distributions of the model parameters for fitting the simulated epidemic using the network-based continuous time ILM assuming a partially observed epidemic (unknown infection times). 56
2.6 The posterior means and 95% credible intervals of the infection times for fitting the simulated epidemic using the network-based continuous time ILM assuming a partially observed epidemic (unknown infection times). The black line represents the observed removal times; the solid red line represents the posterior means; and dotted red lines represent the 95% credible interval. 57
2.7 The posterior predictive distributions of four statistics: the number of infected individuals, the average removal times, the variance of removal times, and the length of epidemic for fitting a partially observed epidemic (unknown infection times) using network-based continuous time ILM. The red vertical lines represent the observed statistic values and the solid and dotted blue vertical lines represents the posterior predictive means and 95% credible intervals of the four statistics, respectively. 58

2.8 A grid plot of the XY coordinates of plants from the TSWV data set. The red plots represent infected plants at the end of the experiment. . . . 60
2.9 The MCMC chains of the posterior distributions of the model parameter α and β for fitting the fully observed TSWV data using the $\mathcal{SIR}$ distance-based continuous time ILM. 62
2.10 The MCMC chains and density plots of the posterior distributions of the model parameter α, β, $\delta^{(inc)}$ and $\delta^{(delay)}$ for fitting the partially observed TSWV data (unknown infection and removal times) using the $\mathcal{SINR}$ distance-based continuous time ILM. 65
2.11 The posterior means (solid red line) and 95% credible intervals (dotted red lines) of the infection times for fitting the partially observed TSWV data (unknown infection and removal times) using the $\mathcal{SINR}$ distance-based continuous time ILM. The black line represents the observed notification times. 66
2.12 The posterior means (solid red line) and 95% credible intervals (dotted red lines) of the removal times for fitting the partially observed TSWV data (unknown infection and removal times) using the $\mathcal{SINR}$ distance-based continuous time ILM. The black line represents the observed notification times. 67

3.1 Two examples of spatial contact networks. Red dots represent nodes with different sizes corresponding to their degree (number of edges connected to a node). 78
3.2 Posterior means for the simple network-based ILM with 95% credible intervals of the intense (left) and sparse (right) contact network epidemics (population size N = 25) based on the full Bayesian MCMC analyses of the three assumption scenarios of the contact network. The total number of connections observation model of Section 3.3.2 is applied in the unknown contact network scenarios. Horizontal red lines are the true values of the parameter; α_0 = 0.8, α_1 = 0.5, δ = 2, and ϵ = 0. 88
3.3 Posterior means for the simple network-based ILM with 95% credible intervals of the intense (left) and sparse (right) contact network epidemics (population size N = 50) based on the full Bayesian MCMC analyses of the three assumption scenarios of the contact network. The total number of connections observation model of Section 3.3.2 is applied in the unknown contact network scenarios. Horizontal red lines are the true values of the parameter; α_0 = 0.05, α_1 = 0.3, δ = 0.5, and ϵ = 0. 89

3.4 Summary of posterior predictive distributions obtained from the analyses of the data sets for a population of 25 individuals. The cross green points represent the true values of the statistics, and the red and blue lines represent the average posterior of the statistics with 95% credible intervals based on the known contact network analyses and the completely unknown contact network analyses including the total number of connections as the observation model, respectively. 90
3.5 Summary of posterior predictive distributions obtained from the analyses of the data sets for a population of 50 individuals. The cross green points represent the true values of the statistics, and the red and blue lines represent the average posterior of the statistics with 95% credible intervals based on the known contact network analyses and the completely unknown contact network analyses including the total number of connections as the observation model, respectively. 91
3.6 Posterior means for the simple network-based ILM with 95% credible intervals of the intense (left) and sparse (right) contact network epidemics (population size N = 25) based on the full Bayesian MCMC analyses of the completely unknown contact network. Each epidemic was analyzed three times: with no degree information, with the degree distribution, and with a distribution of the total number of connections in the contact network. Horizontal dotted lines are the true values of the parameter; α_0 = 0.8, α_1 = 0.5, δ = 2, and ϵ = 0. 94
3.7 Posterior means for the simple network-based ILM with 95% credible intervals of the intense (left) and sparse (right) contact network epidemics (population size N = 50) based on the full Bayesian MCMC analyses of the complete unknown contact network. Each epidemic was analyzed three times: with no degree information, with the degree distribution, and with a distribution of the total number of connections in the contact network. Horizontal dotted lines are the true values of the parameters; α_0 = 0.05, α_1 = 0.3, δ = 0.5, ϵ = 0. 95
3.8 Posterior means for the degree of nodes with 95% credible intervals of the intense (upper) and sparse (lower) contact network epidemics (population size = 25) based on the full Bayesian MCMC analyses of the unknown contact network. The green circles represent the true average degree of nodes of the contact networks, the average degrees from the MCMC including the total number of connections as the observation model are represented by black, the average degrees from the MCMC including the degree distribution as the observation model are represented by red, and the average degrees from the MCMC without including the observation model are represented by blue. 96

3.9 Posterior means for the degree of nodes with 95% credible intervals of the intense (upper) and sparse (lower) contact network epidemics (population size = 50) based on the full Bayesian MCMC analyses of the unknown contact network. The green circles represent the true average degree of nodes of the contact networks, the average degrees from the MCMC including the total number of connections as the observation model are represented by black, the average degrees from the MCMC including the degree distribution as the observation model are represented by red, and the average degrees from the MCMC without including the observation model are represented by blue. 97

4.1 Two samples of spatial contact networks. Red dots represent nodes with different sizes that are corresponding to their degree (number of edges connected to a node). 112

4.2 Posterior means for the simple network-based ILM with 95% credible intervals for the intense and sparse contact network epidemics. Each plot represents the results of applying different summary statistics in the ABC-PMC for the intense (left) and sparse (right) network epidemics. Summary statistics ABC (1), ABC(2) and ABC(3) are described in Section 4.4.1. Horizontal dotted lines are the true values of the parameters; β = 0.9 (intense network) and 1.8 (sparse network), α_0 = 0.8, α_1 = 0.5 and δ = 2. 130

4.3 Posterior means for the simple network-based ILM with 95% credible intervals of the intense (left) and sparse (right) contact network epidemics based on the full Bayesian MCMC analyses. Horizontal solid lines are the true values of the parameter; α_0 = 0.8, α_1 = 0.5 and δ = 2. 131

4.4 Posterior means and 95% credible intervals for fitting the FMD network-based ILM to the simulated FMD epidemics with the spark term (left, scenario 1) and without spark term (right, scenario 2). The black dotted lines represent results of applying ABC-PMC with summary statistics ABC-sim(1) and blue solid lines of applying ABC-PMC with summary statistics ABC-sim(2); as described in Section 4.4.2. Horizontal dotted lines are the true values of the parameters; $\beta = 4, \alpha_c = 0.01, \phi_s = 0.001, \phi_c = 0.005, \delta = 1.5$ and $\epsilon = 0.001$ 133

4.5 The approximated posterior distribution of the infection and removal time curves for the real FMD epidemic data analyses. 135

4.6 Posterior mean and 95% credible intervals of the probability of connections p_{ij} between farms in the inferred network for the real FMD data using summary statistics ABC-real(1) based on the posterior mean of the contact network parameters. 137

5.1 Two examples of partitioning two populations into ten clusters: left: well-clustered population, right: scattered population. Centroids of each cluster are marked by solid black dots. 155
5.2 Posterior means and 95% credible intervals of the model parameters α_0, α_1, and δ for the simple within-cluster network-based ILMs of the well-clustered populations based on the full Bayesian MCMC analyses. Horizontal black lines represent the true values as: $\alpha_0 = 0.02$, $\alpha_1 = 0.0005$, and $\delta = 3$. 165
5.3 Posterior means and 95% credible intervals of the spark infection parameters ϵ_k, $k = 1, \ldots, K$, for the simple within-cluster network-based ILMs of the well-clustered populations based on the full Bayesian MCMC analyses. 166
5.4 Posterior means and 95% credible intervals of the model parameters α_0, α_1, and δ for the simple within-cluster network-based ILMs of the scattered populations (scenario 2.b: with discrete covariate) based on the full Bayesian MCMC analyses. Horizontal black lines represent the true values as: $\alpha_0 = 0.05$, $\alpha_1 = 0.005$, and $\delta = 3$. 168
5.5 Posterior means and 95% credible intervals of the model parameters α_0, α_1, and δ for the simple within-cluster network-based ILMs of the scattered populations (scenario 2a: with binary covariate) based on the full Bayesian MCMC analyses. Horizontal black lines represent the true values as: $\alpha_0 = 0.02$, $\alpha_1 = 0.05$, and $\delta = 1$. 169
5.6 Posterior means and 95% credible intervals of the spark infection parameters ϵ_k, $k = 1, \ldots, K$, for the simple within-cluster network-based ILMs of the scattered populations (scenario 2b: with discrete covariate) based on the full Bayesian MCMC analyses. 170
5.7 Posterior means and 95% credible intervals of the spark infection parameters ϵ_k, $k = 1, \ldots, K$, for the simple within-cluster network-based ILMs of the scattered populations (scenario 2a: with binary covariate) based on the full Bayesian MCMC analyses. 171
5.8 Posterior means and 95% credible intervals of the model parameters α_0, α_1, and δ for the standard, CCCN and CCCN-LISA analyses of four epidemic data sets of scattered populations (Scenario 2a). 172
5.9 Posterior means and 95% credible intervals of the spark infection parameters ϵ_k, $k = 1, \ldots, K$, for the CCCN and CCCN-LISA analyses of four epidemic data sets of scattered populations (Scenario 2a). 173
5.10 Posterior means and 95% credible intervals of the model parameters α_0, α_1, and δ for the standard and CCCN analyses of epidemic data sets of scattered populations (scenario 2a) under three different number of clusters. First row: 10 clusters, second row: 15 clusters, and third row: 40 clusters. 175

5.11 The computation time (in minutes) of fitting the within-cluster network-based continuous time ILMs to an epidemic data set with different numbers of clusters (10, 15, 20, and 40) using MCMC. 176

A.1.1 The marginal posterior distribution of the distance kernels. Black line represents the spatial terms of the *surveillance* package, and red line represents the distance kernel function of the *EpiILMCT* package. 209

B.1.1 Locations of farms in the Cumbria region, UK during the 2001 UK Foot and Mouth Disease Epidemic. 211
B.2.1 Estimated posterior degree distributions for fitting the simple network-based ILM to six representative data sets. 1^{st} row: sparse networks; 2^{nd} row: intense networks. 212
B.3.1a The approximated posterior distribution of the epidemic curves of the infection times for fitting the FMD network-based ILM to four simulated-FMD epidemic data sets without spark terms (scenario 1) under ABC-sim(1) (first row) and ABC-sim(2) (second row) summary statistics. . . . 213
B.3.1b The approximated posterior distribution of the epidemic curves of the infection times for fitting the FMD network-based ILM to four simulated-FMD epidemic data sets with the inclusion of spark terms (scenario 2) under ABC-sim(1) (first row) and ABC-sim(2) (second row) summary statistics. 214
B.4.1a Histograms of the intermediate marginal posterior distributions for the parameters β, α_c and ϕ_s for the FMD network-based ILM fitted to the real FMD epidemic data in different generations of the ABC where the last column represents the approximated marginal posterior distributions of the parameters. 215
B.4.1b Histograms of the intermediate marginal posterior distributions for the parameters ϕ_c, δ and ϵ for the FMD network-based ILM fitted to the real FMD epidemic data in different generations of the ABC where the last column represents the approximated marginal posterior distributions of the parameters. 216

C.2.1 The density of the posterior means of the infection times for the CCCN2 and CCCN3 analyses of the well-clustered population epidemics. 221
C.2.2 The density of the posterior means of the infection times for the CCCN2 and CCCN3 analyses of the scattered population epidemics of scenario 2a. 222
C.2.3 The density of the posterior means of the infection times for the CCCN2 and CCCN3 analyses of the scattered population epidemics of scenario 2b. 223

C.2.4 Posterior means and 95% of the degree nodes for the simple within-cluster network-based ILM of the well-clustered populations (top), scattered populations of scenario 2a (middle), and scattered populations of scenario 2b (bottom) based on the full Bayesian MCMC under the CCCN3 analysis. Red points represent the average degree nodes of the true contact network. 224
C.3.1 Posterior densities of the model parameters for the simple within-cluster network-based ILMs of the third scattered population (scenario 2a), based on the full Bayesian MCMC analyses using the original (red) and modified (black) LISA approach. 226
C.3.2 The MCMC chains of the posterior distributions of the model parameters for the simple within-cluster network-based ILMs of the first five clusters of the third epidemic data set from scenario 2a under the CCCN2 analysis using the original LISA approach. Each row represents the results of the model parameters of each cluster. 227
C.3.3 The MCMC chains of the posterior distributions of the model parameters for the simple within-cluster network-based ILMs of the other five clusters of the third epidemic data set from scenario 2a under the CCCN2 analysis using the original LISA approach. Each row represents the results of the model parameters of each cluster. 228
C.4.1 Posterior means and 95% credible intervals of the spark infection parameters ϵ_k, $k = 1, \ldots, K$, for the simple within-cluster network-based ILMs of four epidemics, spreading on scattered populations (scenario 2a: with binary covariate) with 15 clusters, based on the full Bayesian MCMC analyses using our approach and the LISA algorithm. 229
C.4.2 Posterior means and 95% credible intervals of the spark infection parameters ϵ_k, $k = 1, \ldots, K$, for the simple within-cluster network-based ILMs of four epidemics, spreading on scattered populations (scenario 2.a: with binary covariate) with 40 clusters, based on the full Bayesian MCMC analyses using our approach and the LISA algorithm. 230

Chapter 1

Introduction

1.1 Introduction to Infectious Disease Models

Barreto et al. (2006) defined an infectious disease as

> "an illness caused by a specific infectious agent or its toxic product that results from transmission of that agent or its products from an infected person, animal, or reservoir to a susceptible host, either directly or indirectly through an intermediate plant or animal host, vector or inanimate environment."

Infectious diseases occur in populations in one of two different broad forms. They can be in endemic form, when the occurrence continues over a certain period of time in a certain geographical location at an expected frequency, or in epidemic form, wherein the occurrence of a disease is greater than what it is expected in a certain geographical area over a certain time period. Here, we are generally interested in epidemics of communicable infectious diseases that are transmitted from infectious to susceptible individuals in a population within a certain geographical region.

The catastrophic risks of infectious disease to animal, plant, and human populations, as well as the potential huge losses in economic growth have increased the attention on

building suitable models to understand the spread of infectious diseases. For example, the 2001 UK foot-and-mouth disease (FMD) epidemic resulted in over four million animals destroyed, and result in severe economic consequences for the UK (Deardon et al., 2010; Keeling et al., 2001). Also, the 2014-2016 West Africa Ebola outbreak resulted in over 28,000 human cases and 11,310 deaths in Guinea, Liberia, and Sierra Leone (WHO, 2016). Further, in crop diseases, the fungal diseases of wheat in the southwest of the US in 1996 resulted in a loss of revenue of about $250 million as a consequence of more than 25 countries banning the import of American wheat (Casagrande, 2000).

The history of developing and improving mathematical models for understanding the dynamics of infectious diseases has already been in practice for more than a century. These mathematical models aim to provide good representation of the dynamics of infectious processes using equation systems involving agent, host, and (often) environment. One of the most important early mathematical models is the equation based compartmental epidemic model proposed by Kermack and McKendrick (1927), commonly known as susceptible-infected-removed ($\mathcal{SIR}$) model. Their $\mathcal{SIR}$ model assumes homogeneity in the underlying population so that any two individuals in the population have the same rate (or probability) of interacting at any given time. Since then, various mathematical approaches have been proposed that incorporate heterogeneity, where the risk of infectious disease varies among members of the population, for the sake of developing more realistic infectious disease models and obtaining a better understanding of disease spread.

In more recent developments, mathematical models have been used, assuming either

homogeneity or heterogeneity in the underlying population, in modelling the epidemic outbreaks of, for example, the 2001 UK FMD (Keeling et al., 2001; Ster and Ferguson, 2007; Jewell et al., 2009; Deardon et al., 2010), severe acute respiratory syndrome (SARS) (McBryde et al., 2006; Gumel et al., 2004), Ebola (D'Silva and Eisenberg, 2017; Chowell et al., 2004; McKinley et al., 2009), and HIV AIDS (Becker et al., 1991; Aalen et al., 1997; Perelson and Nelson, 1999; Sweeting et al., 2005; Taffe et al., 2008).

When modelling population heterogeneity in infectious disease transmission, the transmission dynamics of the infectious disease in the underlying population are allowed to depend on characteristics of the individuals who are able to transmit the disease (if infected) to one another. Models are usually defined at the level of an individual and operate within a compartmental (or multi-state) framework that categorizes individuals based on their health status regarding the infectious disease (O'Neill, 2010). There are different framework options, such as, $\mathcal{SI}$ (susceptible, infectious), $\mathcal{SIR}$ (susceptible, infectious, and removed), $\mathcal{SEIR}$ (susceptible, exposed, infectious, and removed), and $\mathcal{SINR}$ (susceptible, infectious, notified, and removed) frameworks. Under the $\mathcal{SIR}$ framework, for example, individuals who are able to contract the disease are assumed to be in the $\mathcal{S}$ (susceptible) state until they become infected with the disease. As soon as they become infected, they move to the $\mathcal{I}$ (infectious) state, and are able to transmit the disease to other susceptible individuals. Infected individuals are assumed to be infectious for a period called the "infectious period". Following the infectious period, the individual transitions to the $\mathcal{R}$ (removed) state. This final state might correspond to recovery with acquired immunity, isolation or death,

for example. Removed individuals then have no role in the spread of the disease (O'Neill, 2010). The times of transition between states or compartments (event times) are generally the response measures of interest in the model.

1.2 Individual-Level Infectious Disease Models

The increasing development of computational techniques has made the incorporation of more informative epidemic data in the study of infectious disease spread more readily available than in the past. More specifically, there is currently great interest in considering covariates and disease information at the level of individuals in a population. Models that consider such information are referred to as "individual-level" or "individual-based" models. Heterogeneities in the population that can be accounted for by the consideration of such individual-level data can greatly enhance the flexibility and strength of the underlying model (Deardon et al., 2010).

Based on the underlying disease system, population heterogeneity can be formulated at the level of individual persons, animals, plants, or at some aggregated level such as household, farm, region, etc. For example, the 2001 UK FMD epidemic was generally modelled using data at the (individual) farm level. Information about farms (e.g., number and types of animals, spatial location) could then be incorporated as covariates (Deardon et al., 2010). Cook et al. (2007) also incorporated several individual-level risk factors for modelling *Heracleum mantegazzianum* (Giant Hogweed) transmission in Britain at the

regional level. Brown et al. (2014) modelled the spread of an insect infestation among sugar cane plants at the plant level, using data from a field experiment carried out on the Caribbean island of Guadeloupe.

However, a few challenges limit the general use of these models. In most infectious disease systems, the desired data is unobserved or only partially observed. The event times (e.g., infection or removal times) of the infection life history of individuals are rarely observed, at least without some level of censoring. This missing information can be estimated via a data augmentation method implemented using, say, Markov chain Monte Carlo (MCMC) or (less frequently) the Expectation Maximization (EM) algorithm. However, for larger population sizes, both tend to be very slow. Data augmented MCMC (DA-MCMC) has been successfully used for fitting stochastic epidemic models to many epidemic data sets. Examples include the 2001 UK FMD epidemic (Streftaris and Gibson, 2004; Jewell et al., 2009; Deardon et al., 2010), and avian H5N1 influenza (Jewell et al., 2009). However, having large amounts of missing information does exacerbate the computational burden and increase the complexity of the parameter space.

1.3 Spatio-Temporal Infectious Disease Models

The dynamics of infectious disease spread generally result from complex interactions between infectious and susceptible individuals. These interactions can be modelled via a spatial mechanism, often simply as a function of the physical distance between individuals.

Generally, the chances of infection increase as a susceptible individual becomes closer to infectious individuals. Spatial information can be easily incorporated in an individual-level model via an infection kernel (function) that can determine the risk of transmission from infectious to susceptible individuals (Deardon et al., 2010; Jewell et al., 2009).

Although considering spatial information in infectious disease models is often of great benefit when studying the transmission of infectious diseases, it generally worsens the computational burden of fitting the model, especially to large epidemic data sets. Several approaches have been taken to eliminate this issue; for example, kernel linearization (Deardon et al., 2010; Kwong and Deardon, 2012), or likelihood approximation methods such as Gaussian process emulation (Pokharel and Deardon, 2016; Jandarov et al., 2014). Spatial models have been successfully applied to many outbreaks (e.g., the 2001 UK FMD epidemic (Deardon et al., 2010; Streftaris and Gibson, 2004; Jewell et al., 2009), the SARS outbreak of 2003 (McBryde et al., 2006), health care-associated infections such as MRSA and *clostridium difficile* (Forrester et al., 2006; Kypraios et al., 2010), Ebola in West Africa (D'Silva and Eisenberg, 2017), and avian H5N1 influenza (Jewell et al., 2009)).

Often, however, spatial information is used in the model as a proxy for more complex contact networks representing heterogeneity in the population of interest (Bifolchi et al., 2013).

1.4 Contact Network in Infectious Disease Models

The underlying process of disease transmission is often more realistically modelled by incorporating contact network information in the infectious disease model. Incorporating such networks in disease transmission modelling can enhance the prediction of the epidemic pattern and thus aid in controlling the spread of a disease (Danon et al., 2011; Keeling and Eames, 2005). This network information can be formulated as a binary or weighted contact network or a series of such contact networks.

A network is a graph of connections between individuals (e.g., individuals within a population, species within a community) or groups in a population that defines possible transmission paths of a disease. Such networks can represent movement of animals, individuals and/or vehicles between farms (Jewell et al., 2009), social interactions (Lee et al., 2018a), or sexual interactions in humans (Klovdahl, 1985). Contact networks can be directed or undirected, representing the connections between individuals in a population. In the undirected contact network, individuals share equally weighted links to each other (i.e., symmetric). This is not the case in the directed contact network, where individuals can be linked to others without the necessity of sharing the same links (or weights) in the opposite direction (Danon et al., 2011; Keeling and Eames, 2005).

Many previous studies involving network-based transmission models assume a known network, or used simulation techniques to generate networks that have certain properties related to individual habits and their interactions, with epidemics then being simulated across the networks (Barthélemy et al., 2005; Volz, 2008). However, underlying contact networks

for epidemics are often not observed in real life, especially for large populations, since collecting such contact information can be challenging. Personal information for individuals that could be the basis of defining the contact network is usually difficult to obtain, for example, sexual history for sexually transmitted diseases is not something people are always comfortable sharing. The problem of recall error is also an issue, especially when individuals have many contacts. The time-consuming and expensive nature of collecting these contact data are another problem that could affect the practicality of employing a contact network (Keeling and Eames, 2005).

Various approaches to the incorporation of contact networks in infectious disease models can be found in the literature. Various types of models have been used which restrict the spread of the infectious disease across only the edges of a contact network. Some research studies employing this approach are heavily contingent upon simulation through a given, or simulated contact network, that has certain properties of the population (Volz, 2008). Other studies make inferences regarding epidemics through the combination of the epidemic model and transmission network model. Although the contact network considers all of the existing connections that the disease might spread over, the disease only spreads over a subset of these connections. This subset of connections forms a direct transmission network (also called the transmission tree) which contains no loops, and can provide useful information about the individuals most involved in the disease transmission (Riley et al., 2003; Keeling and Eames, 2005). Britton and O'Neill (2002) modelled transmission networks via homogeneous stochastic epidemic models using a Bernoulli random graph

(BRG). In this network modelling, any two individuals (nodes) were assumed to have the same probability of being connected. They showed that parameters of the network and epidemic process can be successfully estimated under a partially observed disease system (unknown infection times, but known removal times). An extension of this model to exponential-family random graph models (ERGMs) was introduced by Groendyke et al. (2012). They developed the model to a more general form, which allows the transmission model to incorporate various covariates, and was applied to a measles outbreak in children in the town of Hagelloch, Germany in 1861. Their results suggest the possibility of making meaningful statements regarding the contact network structure of the population, as well as defining the significant factors that have substantial impact on the transmission network structure. A recent extension of the model was also introduced by Lee et al. (2018a). These authors developed a network epidemic model by combining a stochastic epidemic model and a preferential attachment (PA) model previously introduced by Barabási and Albert (1999). A preferential attachment network is generated by adding individuals (nodes) sequentially and modelling the number of new connections (edges) as random variables that follow specific distributions (e.g., a truncated or censored Poisson distribution). Lee et al. (2018a) have successfully combined the PA model with a Susceptible-Infectious ($\mathcal{SI}$) epidemic model to examine epidemics that spread across social networks, and have demonstrated better performance of this approach compared to the use of the BRG model.

However, rather than modelling the transmission network, our focus is on incorporating a contact network (or networks) as a covariate in a stochastic epidemic model to account

for the effect of existing connections (potentially infection rules) between individuals in the population. For example, Jewell et al. (2009) used contact network data, along with other covariates, to model a simulated high pathogenicity Avian Influenza (HPAI) H5N1 outbreak in the poultry industry of Great Britain. The study incorporated two contact networks indirectly by considering the rates of truck contacts for feed mills and slaughterhouses at the level of premises. The authors have also suggested allowing for uncertainty in the network specification, although they did not do this. In this thesis, we will focus on incorporating uncertainty about the contact network when modelling infectious disease data via individual-level models of disease transmission.

1.5 Parameter Estimation in a Bayesian Framework

Statistical inference for infectious disease models, especially individual-level models, is usually carried out within a Bayesian framework. Frequentist (or "classical") statistical approaches tend to be difficult to implement for such models, especially in the presence of a large amount of missing information (O'Neill, 2010). The main difference between the classical and Bayesian approaches is the way of treating the unknown model parameters. In the classical approach, model parameters are assumed unknown but fixed, and estimated by a point estimate. The data are considered to be a realization from a random variable. Uncertainty about the model parameters is generally summarized via frequentist confidence intervals that are based on the sampling distribution (i.e., unobserved instances of the data).

However, model parameters under the Bayesian approach are assumed to be random variables that follow some distribution after conditioning upon the observed data. The model parameters then have a distributional estimates, instead of point estimates as seen in the frequentist approach, that automatically includes uncertainty regarding the model parameters. This distributional estimate is known as the posterior distribution.

1.5.1 Bayesian Inference

Assume that the data $\boldsymbol{D}$ is observed from a population that was generated by a statistical model (e.g., an ILM) with unknown model parameters $\boldsymbol{\theta}$. In the Bayesian approach, inference about the parameters can then be obtained via the posterior density $\pi(\boldsymbol{\theta}|\boldsymbol{D})$, where parameters $\boldsymbol{\theta}$ are conditioned on the data $\boldsymbol{D}$. This posterior density is composed of the likelihood function $L(\boldsymbol{D}|\boldsymbol{\theta})$, which represents the conditional distribution of the data $\boldsymbol{D}$ given the parameters $\boldsymbol{\theta}$, and a prior density $\pi(\boldsymbol{\theta})$, which characterizes a priori belief about the parameters. The posterior density is is found via Bayes' theorem, which states that

$$\pi(\boldsymbol{\theta}|\boldsymbol{D}) = \frac{L(\boldsymbol{D}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int L(\boldsymbol{D}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}, \tag{1.1}$$

where the integral term is typically referred to as the normalization constant. As this integral is generally analytically intractable, with no closed-form solution, the posterior distribution is often considered up to a constant of proportionality:

$$\pi(\boldsymbol{\theta}|\boldsymbol{D}) \propto L(\boldsymbol{D}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}). \tag{1.2}$$

The posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{D})$ contains all the information about the model parameters $\boldsymbol{\theta}$. This information is typically summarized by a point estimate (e.g., mean, mode, or median), and a posterior variability measure (credible intervals or variance).

There are a number of advantages and disadvantages in using the Bayesian approach. One advantage is that it allows for the natural incorporation of data uncertainty, typically a problem found in infectious disease data sets where event times are usually unknown, or might be censored. The imputation of this missing information can be incorporated easily in the Bayesian approach using data augmented techniques using, say, MCMC methods. Missing information is considered as a set of latent variables, essentially treated as parameters in the model. Another advantage is that the Bayesian approach can allow for a wide range of complex hierarchical models. The well-known computational method, MCMC, provides a reasonably efficient means for fitting such complex models. However, in the case of having highly complex models (e.g., ILMs), especially with a large population size and high dimensional parameter space, the issue of computation time can become problematic. First, the calculation of the likelihood can be computationally expensive. Further, this issue becomes worse when incorporating large amounts of missing information (e.g., contact network and infection times in ILMs), and so the dimensionality of the parameter vector is large.

1.6 Markov Chain Monte Carlo (MCMC) Methods

Monte Carlo techniques can be used to simulate the posterior distribution to avoid the difficulty of integrating to, for example, calculate the normalization constant of Bayes' theorem. Markov chain Monte Carlo (MCMC) methods are examples of Monte Carlo simulation methods, in which our goal is to build a Markov chain whose stationary distribution is the posterior distribution to be sampled from (Gelman et al., 2004). It is very commonly used to facilitate Bayesian inference.

Consider the posterior expectation of a statistic $m(\theta)$

$$I = E_{\pi}(m(\theta)) = \int_{\theta} m(\theta)\pi(\theta|\boldsymbol{D})d\theta. \tag{1.3}$$

This integration can be approximated by Monte Carlo integration as

$$\hat{I} = \frac{1}{\omega}\sum_{i=1}^{\omega} m(\theta_i) \simeq E_{\pi}(m(\theta)), \tag{1.4}$$

where θ_i are samples from the posterior distribution. As $\omega \to \infty$, this will converge almost surely to the integral in Equation 1.3 when the samples are independent (by the strong law of large numbers). This result also holds when the samples are generated dependently through a Markov Chain (Gelman et al., 2004).

Markov chain sampling involves sampling the random variable from a distribution sequentially such that the new sample depends only on the state of the current drawn sample. Thus, the new value of θ is sampled conditionally on the current value of θ from a transition

distribution G,

$$\theta_{new} \sim G(\theta_{new}|\theta_{current}). \tag{1.5}$$

In MCMC, G is chosen such that the values of θ drawn converge to a stationary distribution, formally called the target distribution, which in a Bayesian framework will be the posterior distribution of interest. The Monte Carlo Integration method can then be used to calculate any statistics of interest from chains that demonstrate such a stationary distribution (Gelman et al., 2004).

1.6.1 Gibbs Sampler

Consider a posterior distribution with a vector of parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_p)$. The (univariate) Gibbs sampler is an MCMC method that generates samples from the conditional distributions of the parameters represented by conditional densities $\pi(\theta_i|\boldsymbol{D}, \boldsymbol{\theta}_{-i})$ where $\boldsymbol{\theta}_{-i} = (\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_p)$ is the vector $\boldsymbol{\theta}$ with its ith component removed, $i = 1, ..., p$. These conditional densities should preferably be easy to simulate from (e.g., having a standard form) (Gelman et al., 2004). The Gibbs sampler algorithm (assuming each parameter is updated univariately) is as follows:

Given $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)}, ..., \theta_p^{(t)})$, at the next iteration $t + 1$ of the MCMC chain the parameters are generated by:

Step 1 $\theta_1^{(t+1)} \sim \pi(\theta_1|\theta_2^{(t)}, \theta_3^{(t)}, ..., \theta_p^{(t)})$.

Step 2 $\theta_2^{(t+1)} \sim \pi(\theta_2|\theta_1^{(t+1)}, \theta_3^{(t)}, ..., \theta_p^{(t)})$.

$\vdots$

Step p $\theta_p^{(t+1)} \sim \pi(\theta_p|\theta_1^{(t+1)}, \theta_2^{(t+1)}, \theta_3^{(t+1)}, ..., \theta_{p-1}^{(t+1)})$.

Note that, it is not necessary to update each parameter from their univariate conditional; often blocks of parameters can be updated from a multivariate conditional distribution. Note also, if the conditional densities are not known (or easy to simulate from), an alternate algorithm can be used; the Metropolis-Hastings MCMC algorithm (Gelman et al., 2004).

1.6.2 Metropolis Hastings (MH) MCMC

The Metropolis-Hastings (MH) algorithm is an MCMC algorithm in which proposed values of θ are rejected or accepted at each Markov chain iteration (Gelman et al., 2004; Gamerman and Lopes, 2006). If θ_i has a conditional density $\pi(\theta_i|\boldsymbol{D}, \boldsymbol{\theta}_{-i})$ that is difficult to sample from, then at the current state of the chain a new value θ^* is proposed from a proposal distribution with density Q, and the transition to the proposed value is accepted based on the acceptance probability[1]:

$$\alpha(\theta^*|\theta_i^{(t)}) = \min\left\{1, \frac{\pi(\theta^*|\boldsymbol{D}, \boldsymbol{\theta}_{-i})Q(\theta_i^{(t)}|\theta^*)}{\pi(\theta_i^{(t)}|\boldsymbol{D}, \boldsymbol{\theta}_{-i})Q(\theta^*|\theta_i^{(t)})}\right\}. \tag{1.6}$$

We accept the transition as $\theta_i^{(t+1)} = \theta^*$ with probability $\alpha(\theta^*|\theta_i^{(t)})$ or reject it and set $\theta_i^{(t+1)} = \theta_i^{(t)}$ with probability $1 - \alpha(\theta^*|\theta_i^{(t)})$. Two commonly used proposal/updates are

[1]The posterior distribution here is considered up to a constant of proportionality.

the random-walk MH algorithm and the independence sampler. A third update is the Gibbs sampler, which is a special case in which $\alpha(\theta^*|\theta_i^{(t)}) = 1$ (Gelman et al., 2004).

In a random walk MH algorithm, a new value of θ_i is proposed from a proposal distribution that is symmetric around zero, such as the normal, student-t or uniform distributions. That is, θ^* is generated as

$$\theta^* = \theta_i^{(t)} + z_t, \tag{1.7}$$

where $z_t \sim G(|\theta^* - \theta_i|)$. The proposed new value θ^* depends only on the current value of θ_i and the acceptance probability given by Equation 1.6 will be simplified to

$$\alpha(\theta^*|\theta_i^{(t)}) = \min\left\{1, \frac{\pi(\theta^*|\boldsymbol{D}, \boldsymbol{\theta}_{-i})}{\pi(\theta_i^{(t)}|\boldsymbol{D}, \boldsymbol{\theta}_{-i})}\right\}. \tag{1.8}$$

However, in an independence sampler, the proposed value θ^* does not depend on the state of the current value of θ_i. That is, proposing a new value θ^* from the proposal distribution $Q(\bullet)$, the acceptance probability given by Equation 1.6 will be

$$\alpha(\theta^*|\theta_i^{(t)}) = \min\left\{1, \frac{\pi(\theta^*|\boldsymbol{D}, \boldsymbol{\theta}_{-i})Q(\theta_i^{(t)})}{\pi(\theta_i^{(t)}|\boldsymbol{D}, \boldsymbol{\theta}_{-i})Q(\theta^*)}\right\}. \tag{1.9}$$

In each case, the algorithm would be run until stationarity of the chain appears to have been reached.

1.6.3 Data Augmentation

One major problem in statistical modelling concerns the presence of missing or censored data. In the Bayesian framework, these missing data can be treated as random variables along with the model parameters. The missing data can then be estimated and uncertainty about them incorporated via the data augmented MCMC technique (Gamerman and Lopes, 2006).

Let $\boldsymbol{D}$ be the observed data and $\boldsymbol{Z}$ denote the missing data. Assuming the existence of $\pi(\boldsymbol{D}, \boldsymbol{Z}|\theta)$ and the prior independence of $\pi(\boldsymbol{Z})$ and $\pi(\theta)$, then the joint posterior of the parameters θ and the missing data $\boldsymbol{Z}$ up to proportionality takes the form:

$$\pi(\theta, \boldsymbol{Z}|\boldsymbol{D}) \propto \pi(\boldsymbol{D}, \boldsymbol{Z}|\theta)\pi(\theta), \tag{1.10}$$

where $\pi(\boldsymbol{D}, \boldsymbol{Z}|\theta)$ is the likelihood of the augmented data. An MCMC algorithm can then be used to sample from the joint posterior of θ and $\boldsymbol{Z}$ (Gamerman and Lopes, 2006).

As mentioned previously, in infectious disease systems there is the possibility of having a large amounts of missing data as the infection times and infectious period are usually unknown. Other facets of the data, such as the contact network (as in our case), could also be unknown. If so, the number of these latent variables becomes even larger and increases as the size of the population increases. Strong correlations could also exist between the missing information and the model parameters resulting in less efficient MCMC algorithms. Consequently, the MCMC may need to be run for a large number of iterations in order achieve stationarity (Press, 2009). Further, the large parameter set requires computation

the likelihood many times at each MCMC iteration, as the parameters need to be updated individually (or in small blocks) to avoid the problem of very high rejection rates associated with large block updates. All in all, this can lead to a serious computational bottleneck.

1.7 Approximate Bayesian Computation Method

An MCMC approach can be prohibitive when the likelihood function is analytically or computationally intractable. This intractability of the likelihood can be a result of the likelihood having an unknown form (e.g., in some genetics problems (Tavaré et al., 1997)) or the presence of high dimensional parameter vectors (with missing information). A solution to overcome evaluating the likelihood is to use a likelihood approximation method. The increasingly popular approximate Bayesian computation (ABC) method is one such possibility. The main idea behind ABC algorithms is to avoid directly evaluating the likelihood function, by comparing the "closeness" of simulated and observed data, usually in the context of some (ideally sufficient) statistics.

Typically, such a procedure works in the following way. First, a candidate θ' is proposed from a so-called importance proposal distribution (usually the prior distribution of θ). Then, this candidate parameter is used to simulate a data set $D^{'}$ from the model. The "closeness" of $D^{'}$ to observed data D is assessed via a predefined metric $\rho(D, D^{'})$, usually a function of summary statistics $\mathbb{S}(\cdot)$, though preferably (but rarely in practice) sufficient statistics. Given a tolerance level ξ, the candidate parameter is accepted if $\rho(\mathbb{S}(D), \mathbb{S}(D^{'})) \leq$

ξ. This algorithm would then be run a set number of times. The resulting distribution, $\pi(\theta|\rho(\mathbb{S}(D), \mathbb{S}(D^{'})) \leq \xi)$, will approximate the posterior $\pi(\theta|D)$; the approximation improves as $\xi \rightarrow 0$ (Pritchard et al., 1999). This is the so-called ABC rejection sampler (Pritchard et al., 1999; Tavaré et al., 1997) and is shown in Algorithm 1.1.

Algorithm 1.1 Rejection ABC

1: **for** $i = 1$ **to** N **do**
2: **repeat**
3: Sample $\theta^* \sim \pi(\bullet)$ from the prior.
4: Generate a data set $D^{'} \sim MODEL(\bullet\,|\theta^*)$.
5: Determine discrepancy $\rho(\mathbb{S}(D), \mathbb{S}(D^{'}))$ between the observed data D and the simulated data $D^{'}$.
6: **until** $\rho(\mathbb{S}(D), \mathbb{S}(D^{'})) \leq \xi^{(t)}$.
7: Set $\theta^{(i)} = \theta^*$.

The major drawback of this algorithm is that it typically produces low acceptance rates. A large number of simulations is thus needed to obtain a reasonable number of approximate posterior samples. This problem worsens if the prior distributions of the parameters are diffuse with respect to the posterior distribution. Additionally, smaller values of the tolerance level ξ also tends to increase the rejection rate (Sisson et al., 2007).

Another variant of an ABC algorithm places the above procedure of ABC within an MCMC framework. Marjoram et al. (2003) proposed the ABC-MCMC algorithm that replaces the likelihood ratio in the usual Metropolis-Hastings acceptance probability by its approximation $\rho(\mathbb{S}(D), \mathbb{S}(D^{'})) \leq \xi$. The ABC-MCMC algorithm is given in Algorithm 1.2.

The candidate parameter is generated from an arbitrary proposal density $q(\theta|\,\bullet)$. The

Algorithm 1.2 ABC-MCMC

1: **for** $i = 1$ **to** N **do**
2: Sample $\theta^* \sim q(\theta^*|\theta^{(i)})$, where q is some proposal density
3: Generate a data set $D^{'} \sim MODEL(\bullet\,|\theta^*)$.
4: Generate $u \sim U(0,1)$.
5: Set $\theta^{(i+1)} = \theta^*$ with probability

$$\alpha = \min\left\{1, \frac{\pi(\theta^*)q(\theta^{(i)}|\theta^*)}{\pi(\theta^{(i)})q(\theta^*|\theta^{(i)})}\mathbf{1}(\rho(\mathbb{S}(D^{'}), \mathbb{S}(D)) \leq \xi)\right\},$$

otherwise set $\theta^{(i+1)} = \theta^{(i)}$
6: If $i < N$, increment $i = i + 1$ and go to 2.

acceptance rates of this algorithm are higher than those obtained with Algorithm 1.1. The acceptance rate for ABC-MCMC samplers is proportional to the probability of simulating a $D^{'}$ such that $\rho(\mathbb{S}(D^{'}), \mathbb{S}(D)) \leq \xi$, which is itself proportional to the likelihood (Sisson et al., 2007). This can introduce difficulties with regard the parameters moving through the parameter space. For example, suppose the ABC-MCMC sampler enters an area of relatively low probability with a poor proposal mechanism (when the tolerance value ξ is small). Thus, it becomes difficult to move to the right support with a reasonable chance of acceptance. The sampler might then get "stuck" in that part of the state space for long periods of time (Sisson et al., 2007).

A more efficient way of implementing ABC is to use a variant of sequential ABC (Sisson et al., 2007), examples of which are Partial Rejection Control (PRC) (Sisson et al., 2007), Sequential Monte Carlo (SMC) (Del Moral et al., 2012) and Population Monte Carlo (PMC) (Beaumont et al., 2009; Toni et al., 2009). In this thesis, we will use the ABC-PMC algorithm introduced by Beaumont et al. (2009) to fit network-based contin-

uous time ILMs to epidemic data sets (see Section 4.3.1). The ABC-PMC algorithm has become popular and used in various studies including the following: Kypraios et al. (2017) applied ABC-PMC for fitting a homogeneously mixing $\mathcal{SIR}$ model to outbreak data of gastroenteritis in a hospital ward in South Carolina, January 1996; Brown et al. (2018) fitted spatially heterogeneous susceptible-exposed-infectious-removed ($\mathcal{SEIR}$) stochastic epidemic models to Chikungunya data in the Americas during 2014 using ABC-PMC; and McKinley et al. (2009) fitted a homogeneously mixing $\mathcal{SEIR}$ model to model an outbreak of Ebola Haemorrhagic fever in the Democratic Republic of Congo in 1995 using ABC-PMC. More details of this approach are given in Chapter 3.

1.8 Thesis Overview

This thesis investigates the incorporation of contact network uncertainty into network-based individual-level infectious disease models. We consider a continuous time version of the framework of the individual-level models (ILMs) of Deardon et al. (2010) throughout the thesis. The main focus is to consider the effect of incorporating contact network uncertainty on the performance of the models within a Bayesian (or approximate Bayesian) framework. A secondary focus will be on enabling fast inference to fit these models to partial epidemics with a large amount of missing information.

In Chapter 2, we provide a description of our R package *EpiILMCT* (Almutiry et al., 2018). The *EpiILMCT* package allows users to study the spread of infectious disease us-

ing continuous time ILMs. The package provides tools for simulation from spatial and/or network-based continuous time ILMs, and for the graphical summarization of epidemics. Model fitting is carried out within a Bayesian MCMC framework. As infectious disease data are often partially observed, data uncertainties in the form of missing infection times - and in some situations missing removal times - are accounted for using data augmentation techniques. The continuous time ILMs can be implemented within either $\mathcal{SIR}$ or $\mathcal{SINR}$ compartmental frameworks. The package is illustrated using both simulated and experimental data.

In Chapter 3, we explore the use of DA-MCMC methods for infectious disease inference in a contact network setting. We consider fitting individual level models containing a spatially-based contact network that is either entirely, or partially, unobserved, along with uncertainty in individual-level event histories. We examine the performance of such analyses in the presence or absence of contact network observational models based upon knowledge about the degree distribution, or the total number of connections, in the network. Studies in this chapter are based on simulated epidemics with small population sizes.

Chapter 4 investigates the fitting of network-based ILMs with completely unknown contact networks using approximate Bayesian computation population Monte Carlo (ABC-PMC) methods. This is done in the context of both simulated data, and data from the UK 2001 foot-and-mouth disease epidemic. Also, we compare models fitted using the DA-MCMC and ABC-PMC on simulated epidemics of small population sizes.

We introduce an approach to approximate full model inference by partitioning the popu-

lation into a number of spatial clusters in Chapter 5. More specifically, the contact network is divided into a series of isolated "cluster-constrained" contact networks. Between-cluster infections are no longer allowed through the contact network and are modelled instead by random spark infection parameters. This partition scheme results in a large reduction in the number of connections in the contact network that need to be estimated. Consequently, the computational burden of the analysis is much reduced. We call the new model a within-cluster network-based continuous time ILM. It allows the infectivity rate of each individual to be dependent on local effects (covariates and contact network) within a cluster, and global effects (spark infection terms) that allow for infection between clusters.

In Chapter 6, we very briefly summarize our conclusions and consider areas of potential interest for future research.

The thesis is presented in a manuscript format in which each of Chapters 2-5 is an article that has (or will be) submitted to a peer-reviewed journal. Specifically:

- Chapter 2: "Continuous Time Individual-Level Models of Infectious Disease: *EpiILMCT*" has been submitted to, and a revision requested from, *Journal of Statistical Software*.

- Chapter 3: "Spatial Contact Network Uncertainty in Individual Level Models of Infectious Disease Transmission" has been submitted to *Epidemiologic Methods*, and is under review.

- Chapter 4: "Incorporating Contact Network Uncertainty in Individual Level Models

of Infectious Disease using Approximate Bayesian Computation" has been submitted to, and a revision requested from, *The International Journal of Biostatistics*.

- Chapter 5: "Contact network uncertainty in individual level models of infectious disease transmission using within-cluster likelihood approximations" to be submitted to a peer-reviewed statistical journal.

Chapter 2

Continuous Time Individual-Level Models of Infectious Disease: *EpilLMCT*[1]

2.1 Introduction

Innovative mathematical and mechanistic approaches to the modelling of infectious diseases are continuing to emerge in the literature. These can be used to understand the spread of disease through a population - whether homogeneous or heterogeneous - and enable researchers to construct predictive models to develop control strategies to disrupt disease transmission. For example, Deardon et al. (2010) introduced a class of discrete time individual-level models (ILMs) that incorporates population heterogeneities by modelling the transmission of disease given various individual-level risk factors. The general framework of ILMs have already been successfully applied to a broad range of epidemic data, including the 2001 UK foot-and-mouth outbreak (Deardon et al., 2010; Deeth and Deardon, 2016; Malik et al., 2016), tomato spotted wilt virus (TSWV) disease (Pokharel

[1]Chapter 2 has been submitted to, and a revision requested from, *Journal of Statistical Software*.

and Deardon, 2014, 2016), the spread of 1-18-4 genotype of the porcine reproductive and respiratory virus in Ontario swine herds (Kwong et al., 2013), and influenza transmission within households in Hong Kong during 2008 to 2009 and 2009 to 2010 (Malik et al., 2014). Equivalent continuous time ILMs which capture the complex interactions between susceptible and infected individuals through spatial and contact networks can also be considered. The inference and fitting of such models is generally considered within a Bayesian framework using Markov chain Monte Carlo (MCMC).

However, infectious disease epidemiologists have previously found it difficult to apply these individual-level models to real life problems due to a dearth of readily available software products. The application of these continuous time ILMs can now be implemented in an R package, *EpiILMCT* (Almutiry et al., 2018), and is available from the Comprehensive R Archive Network (CRAN) (R Core Team, 2018) at `https://CRAN.R-project.org/package=EpiILMCT`. In this article, we describe the package, *EpiILMCT*, which allows users to simulate and fit epidemic data using distance- and/or network-based models (Bifolchi et al., 2013; Deardon et al., 2010; Jewell et al., 2009), and can also incorporate risk factors associated with both susceptible and infectious individuals. *EpiILMCT* also uses data augmentation techniques to carry out inference when the infection and/or removal times are unknown or censored, as is usually the case. To the extent of our knowledge, this feature is not available in any existing R packages that permit epidemic data analysis and modelling. Tools for the graphical summarization of epidemic data sets and outcomes are also provided. The statistical inferences made in *EpiILMCT* are set in a Bayesian frame-

work and are carried out using MCMC. The main aim here for developing this package is to provide a fast implementation of continuous time ILMs under different epidemic modelling frameworks. Because of the computationally intensive nature of MCMC for such models, we have coded functions, including MCMC, in Fortran to speed up computation.

There are several R packages that permit a range of different modelling tools that allow for fitting spatial-temporal epidemic data. For example, the packages *splancs* (Rowlingson and Diggle, 2017), and *lgcp* (Taylor et al., 2013, 2015) provide methods for analyzing epidemic data as spatial and space-time point patterns. Other packages fit a range of autocorrelation regression spatio-temporal models (e.g., *CARBayesST* (Lee et al., 2018b), *spdep* (Bivand et al., 2013; Bivand and Piras, 2015), and *spTimer* (Bakar and Sahu, 2017, 2015)). Further packages are mentioned in the Handling and Analyzing Spatio-Temporal Data CRAN task view (Pebesma, 2018). The R Epidemics Consortium (2018) provides additional useful resources for disease outbreak analysis related to R software packages.

However, in each case, the functionalities (e.g., models available) of the packages above are quite different to that of *EpilLMCT*. The models of the *EpilLMCT* package are "mechanistic" in that they attempt to more directly model the mechanisms of transmission between individuals. Specifically, they take into account the spatial interactions between individuals with differing disease status (e.g., susceptible, infected, notified, removed) at continuous time points of the epidemic process. Those spatial interactions between susceptible and infectious individuals are incorporated as distance-based effects on the infectivity rate of individuals through an infection kernel function (power-law or Cauchy). The infectivity

rates can also depend upon various susceptibility and transmissibility covariates at the individual level. Additionally, and of key importance, none of the aforementioned packages account for uncertainty in the event times using Bayesian data augmentation MCMC methods.

There are several R packages that allow for the visualization, simulation, and modelling of the spread of epidemics through networks. The package *EpiModel* (Jenness et al., 2018) allows epidemic simulation from mathematical models of infectious diseases through stochastic contact networks based on exponential-family random graph models (ERGMs). Some packages assume observed contact network or networks when fitting the specified model, such as, *ergm* (Handcock et al., 2017; Hunter et al., 2008), *Bergm* (Caimo and Friel, 2014), and *hergm* (Schweinberger et al., 2018). The above packages implement Bayesian analyses for fitting exponential-family transmission network models to observed contact network data.

A recently developed package, *epinet* (Groendyke and Welch, 2018), allows users to infer transmission networks from time-series epidemic data by modelling the contact network using a generalization of the ERGMs. This package makes use of time-series epidemic data as the input assuming an unknown contact network in their functionality, and producing parameter estimates of the epidemic model as well as the contact and transmission networks. The transmission model can contain various covariates that captures important features (summary statistics) of the contact network as well as epidemic transmission.

However, once again these packages have different approaches to that implemented in *EpiILMCT*. We focus here on incorporating a contact network as a covariate in the implemented ILMs in *EpiILMCT*. The response in the ILMs is the event (e.g., infection) time, rather than the transmission network (the transmission network can be inferred later via posterior predictive simulation, of course, but we do not address this here). This response is different to *epinet*, for example, which models the transmission network directly. The *EpiILMCT* package allows for any user specified contact networks, including various special cases such as spatial or random binary (un)directed contact networks or weighted contact networks.

As both the spatio-temporal and contact network-based mechanisms can be key to understanding the dynamics of infectious disease spread, the ILMs in *EpiILMCT* allow for the incorporation of both contact network and distance-based effects jointly in the infectivity rate of individuals. None of the aforementioned packages have this feature in their functionalities.

The use of individual level data in more mechanistic epidemic models has been implemented in only a few other R packages. The most established of these is *surveillance* (Salmon et al., 2016; Meyer et al., 2017), a package for temporal and spatio-temporal disease modelling. It provides tools for outbreak detection in routinely collected surveillance data, as well as a range of models for infectious disease data. The most closely related model in *surveillance* to those of *EpiILMCT* is the additive endemic-epidemic multivariate temporal point process model. These models are implemented in the *twinSIR* function

for modelling the susceptible-infectious-recovered ($\mathcal{SIR}$) event history of a fixed population in continuous time using individual level data. However, not only is the underlying model framework different to that considered in the *EpiILMCT* package, but the *twinSIR* function does not allow for uncertainty in event times to be taken into account via data augmentation techniques. The function does not allow for only the epidemic terms of the model to be considered, as can be done in *EpiILMCT*; both endemic (e.g., seasonal) and epidemic terms must be included in the analysis. In addition, the distance kernel used in the epidemic part of the *twinSIR* function is represented by a linear combination of non-negative basis functions, and is thus different from the distance kernels used in the *EpiILMCT* package.

The *EpiILM* package (Warriyar KV and Deardon, 2018) that has recently been made available in R, provides similar utility to *EpiILMCT*, but for discrete-time ILMs. The models it contains provide options to include susceptible individual covariate information, as well as a choice to describe population heterogeneity. However, the package is limited to discrete-time distance-based or network-based infection kernels and requires known event histories (i.e., there is no data augmentation feature).

As stated previously, inference for the models of *EpiILMCT* is carried out in a Bayesian MCMC framework. Although there are packages available in R to implement MCMC algorithms such as *MCMCpack* (Martin et al., 2011) and *adaptMCMC* (Scheidegger, 2018), all are based on the random walk Metropolis-Hastings (M-H) algorithm. The data augmented MCMC algorithm used in the *EpiILMCT* package to fit various models uses random walk and independence sampler (within Gibbs) steps within a M-H algorithm. The

independence sampler algorithm in this package appears to be essential for updating the missing data efficiently (event times and infectious periods)in order to achieve well-mixing MCMC chains. Chains from purely random walk M-H algorithms, even if tuned adaptively, do not mix well.

Our main purpose for developing this package is to make the use of continuous time ILMs available to epidemiologists and statisticians through R, one of the most commonly used statistical software packages. Overall, *EpiILMCT* offers greatly increased flexibility for analyzing complex disease data. The remainder of this paper is laid out as follows. In the next section, we describe the general continuous individual-level model implemented in *EpiILMCT*. We also discuss the different infection kernel functions implemented in the package. Sections 3 and 4 discuss the functions contained within the package and the underlying Bayesian inference, respectively. Section 5 illustrates the application of *EpiILMCT* to simulated and real data, while Section 6 concludes the paper with a short summary of the software package and its implications for disease modelling.

2.2 The Model

The *EpiILMCT* package allows for the implementation of continuous time equivalents, and extensions, of the discrete-time individual-level models (ILMs) of Deardon et al. (2010). The compartmental frameworks considered are the susceptible-infectious-removed ($\mathcal{SIR}$) and susceptible-infectious-notified-removed ($\mathcal{SINR}$). In both frameworks, each

individual is assumed to be in one of these states at any point in time, $t \in \mathbb{R}^+$. In the $\mathcal{SIR}$ framework, infected individuals transition between states, susceptible to infectious and from infectious to removed. Individuals are assumed to be in the susceptible ($\mathcal{S}$) state until they become infected at which point they become immediately infectious ($\mathcal{I}$), and able to transmit the disease for the duration of their infectious periods before entering the removed ($\mathcal{R}$) state. In the $\mathcal{SINR}$ model, infectious individuals are assumed to move from the infectious state ($\mathcal{I}$) to a notified ($\mathcal{N}$) state. The latter represents a state in which individuals have been identified as having the disease, and may be subjected to various restrictions (e.g., government-imposed movement constraints in the 2001 UK Foot-and-Mouth disease (FMD) outbreak). The $\mathcal{N}$-state infectivity rate is often assumed to be lower than that of $\mathcal{I}$-state. As infectious individuals enter the $\mathcal{R}$-state, they are removed from the infectious population (e.g., because of recovery and acquired immunity, death or quarantine) and from thereon play no role in transmitting the disease.

A full epidemic history consists of all transition event times for all individuals, and defines the state of all individuals at each point in time. For example for the $\mathcal{SIR}$ framework, $\mathcal{S}$(t), $\mathcal{I}$(t) and $\mathcal{R}$(t) at time t for $t \in [0, t_{obs}]$ is defined by all infection and removal times. Here, t_{obs} is the maximum removal time (i.e., the time that the last infectious or notified individual enters the removed state). The nomenclature is the same for the $\mathcal{SINR}$ model, but with the addition of the $\mathcal{N}$(t) state. We assume that each susceptible individual j at time

t has an infectivity rate give by:

$$\lambda_{ij} = \begin{cases} \lambda_{ij}^{-} & i \in \mathcal{I}(t), j \in \mathcal{S}(t) \\ \lambda_{ij}^{+} & i \in \mathcal{N}(t), j \in \mathcal{S}(t) \qquad \text{(only under the } \mathcal{SINR} \text{ framework)} \end{cases}, \tag{2.1}$$

where

$$\lambda_{ij}^{-} = \Omega_S(j)\Omega_T(i)\kappa(i,j)$$

$$\lambda_{ij}^{+} = \gamma\Omega_S(j)\Omega_T(i)\kappa(i,j), \quad \gamma > 0.$$

So, the total rate of infectivity of each susceptible individual j at time t is given by:

$$\lambda_{jt} = \left[\sum_{i \in \mathcal{I}(t)} \lambda_{ij}^{-}\right] + \epsilon(j,t), \tag{2.2}$$

under the $\mathcal{SIR}$ framework, and

$$\lambda_{jt} = \left[\sum_{i \in \mathcal{N}^{-}(t)} \lambda_{ij}^{-} + \sum_{i \in \mathcal{N}^{+}(t)} \lambda_{ij}^{+}\right] + \epsilon(j,t), \tag{2.3}$$

under the $\mathcal{SINR}$ framework, where $\mathcal{N}^{-}(t)$ is the set of infectious individuals at time t who have been infected but have not reached the notified state; and $\mathcal{N}^{+}(t)$ is the corresponding set for notified individuals (Jewell et al., 2009).

$\epsilon(j,t)$ is a spark function which allows for random infections otherwise unexplained by the model. This might represent, for example, the infection of a susceptible individual from a source outside of the observed population. In this package, we fix the spark term $\epsilon(j,t)$ such that $\epsilon(j,t) = \epsilon$; $\epsilon \geq 0$. $\Omega_S(j)$ and $\Omega_T(i)$ are the susceptibility and transmissibility

functions, respectively. They are defined as:

$$\Omega_S(j) = \mathbf{S}\mathbf{X}_j^{\phi}; \text{ and } \Omega_T(i) = \mathbf{T}\mathbf{Z}_i^{\xi} \quad \phi, \xi > 0,$$

where $\mathbf{S}$ and $\mathbf{T}$ are the susceptibility and transmissibility parameter vectors, respectively; $\mathbf{X}$ and $\mathbf{Z}$ are the susceptibility and transmissibility risk factor matrices, respectively; and ϕ and ξ are vectors of the power parameters of the susceptibility and transmissibility functions, respectively. These power parameters allow for non-linearity between the susceptibility and transmissibility risk factors and the infection rate (Deardon et al., 2010). In the $\mathcal{SINR}$ model, the notification effect parameter γ is used to measure the risk of infection after notification, which can be reduced or increased depending on the disease type. For example, the transmissibility has been observed to increase after symptoms in SARS (Pitzer et al., 2007), whereas, it can be lower for the 2001 UK FMD (Jewell et al., 2009). Jewell et al. (2009) stated this effect parameter in their general model as a control measure parameter that only accounts for a reduction in the risk of infection. In the case of $\gamma = 1$, notification has no effect on infectivity.

The infection kernel $\kappa(i, j)$ represents shared risk factors between pairs of infected and susceptible individuals. In the *EpiILMCT* package we consider three kernel types: distance-based, network-based, and combined distance and network-based. Two sub-types of distance-based kernels are also considered: Cauchy and power-law. The infection kernel functions are given in Table 2.1. In the distance-based ILMs, the kernel function is based on the distances d_{ij} between individuals generally, but not always, in spatial Euclidean

Table 2.1: Types of kernel functions that are applied in the *EpiILMCT* package for fitting continuous time ILMs.

Model	Kernel type	Kernel function
Distance-based ILMs	Power-law	$\kappa(i,j) = d_{ij}^{-\beta}, \quad \beta > 0$
	Cauchy	$\kappa(i,j) = \frac{\beta}{d_{ij}^2+\beta^2}, \quad \beta > 0$
Network-based ILMs	Binary, undirected	$\kappa(i,j) = c_{ij}\ , \quad c_{ij} = 0$ or 1
	Weighted	$\kappa(i,j) = w_{ij}, \quad w_{ij} \in [0,\infty)$
Combined distance and network-based ILMs	Power-law	$\kappa(i,j) = d_{ij}^{-\beta_1} + \beta_2 c_{ij},$ $\kappa(i,j) = d_{ij}^{-\beta_1} + \beta_2 w_{ij}, \quad \beta_1, \beta_2 > 0$
	Cauchy	$\kappa(i,j) = \frac{\beta_1}{(d_{ij}^2+\beta_1^2)} + \beta_2 c_{ij},$ $\kappa(i,j) = \frac{\beta_1}{(d_{ij}^2+\beta_1^2)} + \beta_2 w_{ij}, \quad \beta_1, \beta_2 > 0$

distance. In the network-based ILMs, the kernel function is based on either binary connections between individuals in a contact network, $c_{ij} \in \{0,1\}$, or on weighted connections $w_{ij} \in [0,\infty)$. In the combined ILMs the kernel consists of a linear function of both.

2.2.1 Likelihood Function

We label the m infected individuals $i = 1, 2, \ldots, m$ with corresponding infection (I_i) and removal (R_i) times such that $I_1 \leq I_2 \leq \cdots \leq I_m$. The $N - m$ individuals who remain uninfected after t_{obs} are labeled $i = m+1, m+2, \ldots, N$ with $I_i = R_i = \infty$.

We then denote infection and removal time vectors for the population as $\mathbf{I} = \{I_1, \ldots, I_m\}$ and $\mathbf{R} = \{R_1, \ldots, R_m\}$, respectively. We assume that infectious periods follow a gamma distribution with shape and rate δ (Jewell et al., 2009). The likelihood of the general $\mathcal{SIR}$ continuous time ILM is then given as follows:

$$\begin{aligned} L(\boldsymbol{I}, \boldsymbol{R}|\boldsymbol{\theta}) &= \prod_{j=2}^{m} \left(\epsilon + \sum_{i:I_i<I_j\leq R_i} \lambda_{ij}^{-} \right) \\ &\times \exp\left\{ -\int_{I_1}^{t_{obs}} \left(\sum_{i\in\mathcal{S}(t)} \epsilon + \sum_{i\in\mathcal{I}(t)} \sum_{j\in\mathcal{S}(t)} \lambda_{ij}^{-}(u - I_i) \right) du \right\} \prod_{i=1}^{m} f(\mathcal{D}_i; \delta) \\ &= \prod_{j=2}^{m} \left(\epsilon + \sum_{i:I_i<I_j\leq R_i} \lambda_{ij}^{-} \right) \exp\left\{ -\sum_{i=1}^{m} \left(\sum_{j=1}^{N} ((R_i \wedge I_j) - (I_i \wedge I_j))\lambda_{ij}^{-} \right) \right\} \\ &\times \exp\left(-\epsilon \sum_{i=1}^{N} [(t_{obs} \wedge I_i) - I_1] \right) \prod_{i=1}^{m} f(\mathcal{D}_i; \delta) \qquad \delta > 0, \end{aligned} \tag{2.4}$$

where $\boldsymbol{\theta}$ is the vector of unknown parameters; $f(\bullet; \delta)$ indicates the density of the infectious period distribution; $\mathcal{D}_i$ is the infectious period of infected individual i defined as $\mathcal{D}_i = R_i - I_i$; and the wedge symbol $\wedge$ denotes the infimum. The likelihood of the general $\mathcal{SINR}$ continuous time ILMs is given by:

$$
\begin{aligned}
L(\boldsymbol{I},\boldsymbol{N},\boldsymbol{R}|\boldsymbol{\theta}) &= \prod_{j=2}^{m}\left(\epsilon+\sum_{i:I_i<I_j\leq N_i}\lambda_{ij}^{-}+\sum_{i:N_i<I_j\leq R_i}\lambda_{ij}^{+}\right)\\
&\times \exp\left\{-\int_{I_1}^{t_{obs}}\left(\sum_{i\in\mathcal{S}(t)}\epsilon+\sum_{i\in\mathcal{I}(t)}\sum_{j\in\mathcal{S}(t)}\lambda_{ij}^{-}(u-I_i)+\sum_{i\in\mathcal{N}(t)}\sum_{j\in\mathcal{S}(t)}\lambda_{ij}^{+}(u-I_i)\right)du\right\}\\
&\times \prod_{i=1}^{m}f(\mathcal{D}_i^{(inc)};\delta^{(inc)})\prod_{i=1}^{m}f(\mathcal{D}_i^{(delay)};\delta^{(delay)})\\
&= \prod_{j=2}^{m}\left(\epsilon+\sum_{i:I_i<I_j\leq N_i}\lambda_{ij}^{-}+\sum_{i:N_i<I_j\leq R_i}\lambda_{ij}^{+}\right)\\
&\times \exp\left\{-\sum_{i=1}^{n_I}\left(\sum_{j=1}^{N}((t_{obs}\wedge N_i\wedge I_j)-(I_i\wedge I_j))\lambda_{ij}^{-}\right)\right\}\\
&\times \exp\left\{-\sum_{i=1}^{n_N}\left(\sum_{j=1}^{N}((t_{obs}\wedge R_i\wedge I_j)-(I_i\wedge I_j))-((t_{obs}\wedge N_i\wedge I_j)-(I_i\wedge I_j))\lambda_{ij}^{+}\right)\right\}\\
&\times \exp\left(-\epsilon\sum_{i=1}^{N}[(t_{obs}\wedge I_i)-I_1]\right)\\
&\times \prod_{i=1}^{m}f(\mathcal{D}_i^{(inc)};\delta^{(inc)})\prod_{i=1}^{m}f(\mathcal{D}_i^{(delay)};\delta^{(delay)}) \qquad \delta^{(inc)},\delta^{(delay)}>0, \qquad (2.5)
\end{aligned}
$$

where $\mathcal{D}_i^{inc}$ and $\mathcal{D}_i^{delay}$ are the incubation and delay periods such that $\mathcal{D}_i^{inc}=N_i-I_i$ and $\mathcal{D}_i^{delay}=R_i-N_i$, respectively. The integrals in both Equations 2.4 and 2.5, which represent the total person-to-person infectious pressure through the course of the epidemic, can be written as the double sums as shown in the respective subsequent equations (Jewell et al., 2009; Neal and Roberts, 2005; Britton and O'Neill, 2002).

Table 2.2: Description of functions and their usages in the *EpilLMCT* package.

Function	Programme	Usage
contactnet	R	Generates undirected binary contact network matrices from spatial (*powerlaw* or *Cauchy*) or random network models.
datagen	Fortran 95	Generates epidemics from distance/network-based individual level models.
epictmcmc	Fortran 95	Runs a Bayesian data augmented MCMC algorithm for fitting specified models (SIR or SINR).
print.epictmcmc	R	Prints the contents of *'epictmcmc'* object to the console.
summary.epictmcmc	R	Summary method for *'epictmcmc'* objects.
plot.epictmcmc	R	Plots the output of *'epictmcmc'* object.
epiplot	R	Provides different plots summarizing epidemic.
loglikelihoodepiILM	Fortran 95	Calculates the log likelihood for the specific compartmental framework and kernel type of the continuous time ILMs.

2.3 Contents of the *EpilLMCT* package

The *EpilLMCT* package can be used to simulate and graphically summarize epidemics, and, for a given model, carry out Bayesian inference and calculate log-likelihoods. Most of the main package functions are written in Fortran 95 (called from within the R wrapper), since they are computationally intensive tasks. The functions contained in the package are

reviewed in Table 2.2.

2.3.1 Contact Network

Various types of contact network can be considered. First, we consider binary contact networks which can be directed or undirected. In an undirected binary contact network, each pair of individuals share the same symmetric connection such that $c_{ij} = c_{ji}$ for $i \neq j$; $i, j = 1, \ldots, N$; and each network is defined by $\binom{N}{2}$ elements where c_{ij} = 1 if a connection exists between individuals i and j, and 0 otherwise. In a binary directed contact network, it is not necessary for individuals to share the same symmetrical relationship so that $c_{ij} \neq c_{ji}$ for $i \neq j$; $i, j = 1, \ldots, N$. This leads to a non-symmetric contact network matrix. Weighted contact networks can also be considered in the *EpiILMCT* package in which the connections between individuals are not described as present or absent but are weighted according to their strength. These too can be directed or undirected.

A function (*contactnet*) is included to generate undirected binary contact networks. It can simulate both spatial networks where connections are more likely to occur between individuals closer in space ("spatial contact networks"), as well as random contact networks. For spatial contact networks, there are two options for specifying the probability of connections between individuals.

First, we use a generalized version of the power-law contact network model of Bifolchi et al. (2013) in which the probability of a connection between individual i and j is given

by:

$$p(c_{ij} = 1) = 1 - e^{-\alpha(d_{ij}^{-\beta})}, \quad \alpha, \beta > 0,$$

where d_{ij} is the Euclidean distance between individuals i and j; β is the spatial parameter; and α is the scale parameter.

The second option is a Cauchy model as used in Jewell et al. (2009). Here, the probability of a connection between individual i and j is given by:

$$p(c_{ij} = 1) = 1 - e^{-\beta/(d_{ij}^2+\beta^2)}, \quad \beta > 0,$$

where d_{ij} is the Euclidean distance between individuals i and j; and β is the spatial parameter.

For the random contact networks, the probability of a connection is generated from a Bernoulli distribution with probability equal to β.

To create the above undirected binary contact networks, the function requires the XY coordinates of individuals (*location*), and the network model (*"powerlaw"*, *"Cauchy"*, or *"random"*) can be chosen using the *type* argument. Setting *plot*=TRUE produces a graph of the network. The following code generates the three types of contact networks for a population of 50 individuals with uniformly distributed spatial layout.

```
R>  library( "EpiILMCT" )
R>  set.seed(66)
R>  loc <- matrix(cbind(runif(50, 0, 10), runif(50, 0, 10)),
+     ncol = 2, nrow = 50)
R>  net1 <- contactnet(type = "powerlaw", location = loc,
```

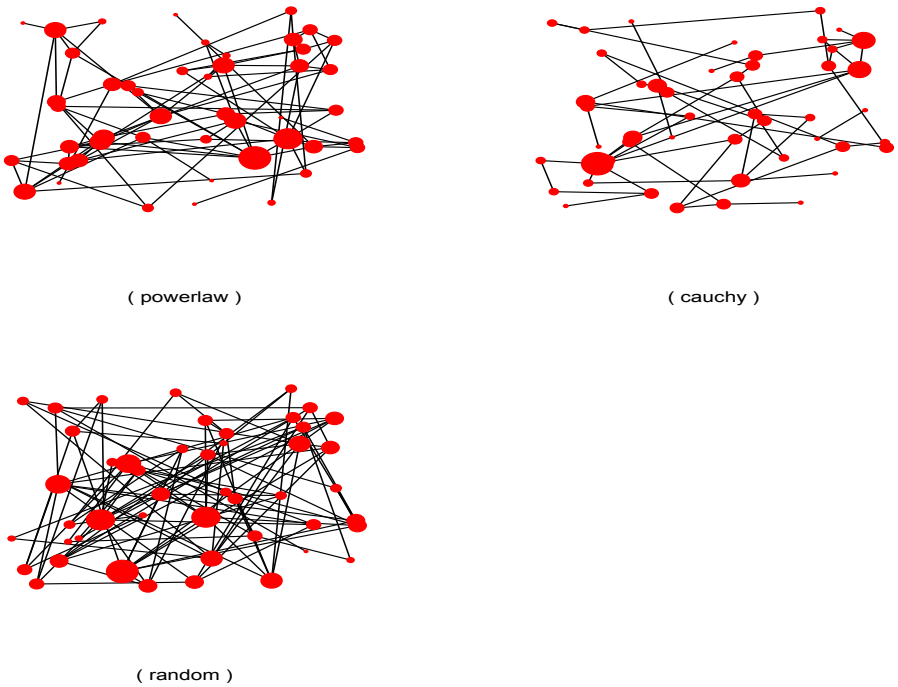


Figure 2.1: Samples of the three undirected binary contact network models generated to the same population. Red dots represent nodes with different sizes that correspond to their degree (number of edges connected to a node).

```
+     beta = 1.5, alpha = 0.5, plot = TRUE)
R>  net2 <- contactnet(type = "Cauchy",  location = loc,
+      beta = 0.5, plot = TRUE)
R>  net3 <- contactnet(type = "random", location = loc,
+      beta = 0.08, plot = TRUE)
```

A realization of the three networks for given spatial layout is shown in Figure 2.1.

2.3.2 Epidemic Simulation

The function *datagen* allows the user to generate epidemics from the continuous time ILMs under the $\mathcal{SIR}$ or $\mathcal{SINR}$ compartmental frameworks. Which framework is to be used is specified through the *type* argument. Each infected individual in a simulated epidemic has an infection life history defined by their time of infection and the length of time spent in the infectious state. We assume infection events follow a non-homogenous Poisson process, such that the time to the next infection, given that the last infection occurred at time t, follows $W_j \sim \text{Exp}(\lambda_{jt})$. Here, W_j represents the "waiting time" for susceptible individual j to become infected.

Under the $\mathcal{SIR}$ framework, and using the chosen distribution of the infectious period, an epidemic is simulated starting with a randomly chosen initial infected individual k at time $I_1 = 0$, or with initial infected individual(s) specified via the argument *initialepi*. This argument requires a vector or matrix containing the id number(s), removal time(s), infectious period(s) and infection time(s) of the infected individual(s). At time I_s, the waiting time until infection for susceptible individual j is then generated as $W_j \sim \text{Exp}(\lambda_{jI_s})$.

The individual with the minimum W is taken as the next infected individual and assigned an infection time $I_{s+1} = I_s + min(W)$; an infection period $\mathcal{D}_j$ (generated from f($\mathcal{D}_j; \delta$)); and a removal time $R_{s+1} = I_{s+1} + \mathcal{D}_j$. The process is repeated until no infectives remain in the population or $I_{s+1} > t_{max}$, where t_{max} is the time at which the epidemic simulation is set to end. Time t_{max} can then be specified via the option *tmax*.

Under the $\mathcal{SINR}$ model, each infected individual is considered to have an incubation

period comprising the time from infection to notification, and a delay period comprising the time from notification to removal. Together the incubation and delay periods constitute the infectious period. An epidemic is simulated in the same manner described above for the $\mathcal{SIR}$ framework, except that the infection period is replaced by incubation and delay periods $\mathcal{D}^{(inc)}$ and $\mathcal{D}^{(delay)}$ (generated from $f(\mathcal{D}_j^{(inc)}; \delta^{(inc)})$ and $f(\mathcal{D}_j^{(delay)}; \delta^{(delay)})$, respectively); and notification and removal times are assigned as $N_{s+1} = I_{s+1} + \mathcal{D}_j^{(inc)}$ and $R_{s+1} = N_{s+1} + \mathcal{D}_j^{(delay)}$, respectively.

In this function, the infectious, incubation and delay periods are assumed to follow either exponential or gamma distributions. These distributions can be specified through the *delta* argument. Under the $\mathcal{SIR}$ framework, *delta* is a vector containing the shape and rate parameters of a gamma distribution, whereas under the $\mathcal{SINR}$ model it is a 2×2 matrix where each row represents the parameters of the incubation and delay period distributions. Note that - as is often done - an exponential distribution can be assigned to any of these distributions by setting shape parameter equal to one.

The epidemic data structure output of the *datagen* function is used throughout the *EpiILMCT* package. Under an $\mathcal{SIR}$ ILM, *datagen* returns a matrix with four columns representing: the id numbers of the individuals, removal times, infectious periods, and infection times, respectively. Under an $\mathcal{SINR}$ ILM, it returns a matrix with six columns: the id numbers of the individuals, removal times, delay periods, notification times, incubation periods, and infection times. Uninfected individuals are assigned infinity values (∞) for both their removal and infection times. Epidemic data from other modelling packages can

be extracted and modified to be used in *EpilLMCT*. For example, we show how this can be done using the individual level models from the *surveillance* package in Appendix A.1.

The choice of kernel function $\kappa(i, j)$ is specified using the *kerneltype* argument. This takes one of three options: *"distance"* for distance-based, *"network"* for network-based, or *"both"* for distance and network-based. The appropriate kernel matrix must also be provided via the *kernelmatrix* argument. If *"distance"* is chosen as the *kerneltype*, the user must choose a spatial kernel (*"powerlaw"*, or *"Cauchy"*) through the *distancekernel* argument. The distance matrix can be obtained from XY coordinate data using the *dist* function from the *stats* package. Otherwise the distance matrix can be specified by the user. Other arguments in the *datagen* function require the data and coefficient parameters for the susceptibility and transmissibility risk factors as explained in Section 2.2.

The package also contains the *epiplot* function, which illustrates disease spread through the epidemic timeline in either distance-based or network-based continuous time ILMs. If the *plottype* argument is set to *"history"*, the function produces epidemic curves of infection and removal times. Example plots are shown in Figure 2.2. Conversely, setting this argument to *"propagation"* produces plots of the epidemic propagation over time. With the latter option, exactly which plots are output varies by kernel. With the network kernel, the function plots all the connections between individuals and overlays these with the epidemic pathway direction over time. This path direction consists of directed edges from all infectious individuals connected to a given newly infected individual i with infection time I_i (one per plot). Thus, this produces directed networks showing possible pathways of the

disease propagation. Figure 2.3 shows a propagation plot for the first six infection events of the simulated network ILM generated epidemic. With the distance kernel, the function plots the spatial epidemic dispersion over time. It shows the changes in the individual status that are related to the chosen compartmental framework. To avoid displaying too many plots, the *time.index* argument allows the user to obtain propagation plots at specific infection time points rather than at every infection time. The user also needs to input the XY coordinates of individuals and the epidemic data via the arguments *location* and *epidat*, respectively .

2.4 Bayesian Inference

Prior distributions of the model parameters are selected from one of three options: gamma, positive half normal or uniform distribution. Then, Metropolis-Hastings MCMC is performed to estimate the joint posterior of the model parameters and latent variables (the latter if various event times are assumed unknown). This is achieved using the function *epictmcmc*. The parameters of the susceptibility and transmissibility functions, infection kernel and spark term (collectively denoted $\boldsymbol{\theta}$) are updated using the random-walk proposals. The user is required to tune the proposal variances to achieve good mixing properties. Thus, the user must provide a vector of initial values, a prior distribution (*"gamma"*, *"uniform"*, or *"halfnormal"*), the prior parameters, and the proposal variance for each parameter.

Using the *datatype* argument, the *epictmcmc* function allows for three scenarios in terms of event time uncertainty: *"known epidemic"* can be used to model a fully observed epidemic with known infection and removal times; *"known removal"* can be used to model a partially observed epidemic where the infection times are unknown; and *"unknown removal"* can be used to model a partially observed epidemic where removal and infection times are unknown. The latter option is only available for the $\mathcal{SINR}$ model where notification times are assumed correctly known.

When infection times are unknown, the rate(s) of the infectious, incubation and/or delay period distributions are assigned gamma prior distributions with shape a and rate b. Thus, the rate parameters have conditional distributions with a standard form following the gamma distribution. For the $\mathcal{SIR}$ continuous time ILMs, this is as follows:

$$\delta|\boldsymbol{\theta}, \boldsymbol{I}, \boldsymbol{R} \sim \Gamma(n_I + a_\delta, M + b_\delta),$$

where δ is the rate of the infectious period distribution; $M = \sum_{i=1}^{n_I} (R_i - I_i)$; and a_δ and b_δ are the prior parameters of the infectious period rate. For the $\mathcal{SINR}$ model, the distribution of the incubation rate and delay parameters are as follows:

$$\delta^{(inc)}|\boldsymbol{\theta}, \boldsymbol{I}, \boldsymbol{N}, \boldsymbol{R} \sim \Gamma(n_I + a_{\delta^{(inc)}}, M_{inc} + b_{\delta^{(inc)}}),$$

where $\delta^{(inc)}$ is the rate of the incubation period distribution; $M_{inc} = \sum_{i=1}^{n_I} (N_i - I_i)$; and $a_{\delta^{(inc)}}$ and $b_{\delta^{(inc)}}$ are the prior parameters of incubation period rate; and

$$\delta^{(delay)}|\boldsymbol{\theta}, \boldsymbol{I}, \boldsymbol{N}, \boldsymbol{R} \sim \Gamma(n_I + a_{\delta^{(delay)}}, M_{delay} + b_{\delta^{(delay)}}),$$

where $\delta^{(delay)}$ is the rate of the delay period distribution; $M_{delay} = \sum_{i=1}^{n_I} (R_i - N_i)$; and $a_{\delta^{(delay)}}$ and $b_{\delta^{(delay)}}$ are the prior parameters of delay period rate.

A Gibbs update (i.e., sampling from the conditional posterior distribution) is used for the infectious period rate (for the $\mathcal{SIR}$ framework) or the incubation and/or delay period rates (for the $\mathcal{SINR}$ model). The required information for each period distribution are entered via the *delta* argument. We assume each period type follows a gamma distribution with fixed shape and unknown rate. Thus, to update the rate parameter of each period, *delta* is a vector containing the fixed shape value, the initial value of the rate, and the parameters of the prior distribution of the rate parameter. In the case of incubation and delay periods being estimated *delta* is a 2×4 matrix where each row contains the required information for each period.

An independence sampler is then used to update the infection times/infectious periods (for the $\mathcal{SIR}$ framework), or the infection times/incubation periods and/or the removal times/delay periods (for the $\mathcal{SINR}$ model). For the $\mathcal{SIR}$ framework, the i^{th} infection time I_i is updated by generating an infectious period $\mathcal{D}_i^*$ from a gamma proposal distribution such that $\mathcal{D}_i^* \sim \Gamma(a, b)$. Then, the new infection time is the difference between the observed removal time and the new infectious period of the i^{th} individual. The same procedure is used for updating the missing event times, infectious periods, and corresponding parameters for the $\mathcal{SINR}$ model. The parameter values of the gamma proposal distribution could be provided through the *periodproposal* argument. If they are not provided, the parameters of the gamma proposal distribution are then based on the fixed shape and

updated rate values from the argument *delta*. Computationally, it may be more efficient to apply a block update for the periods and event times. This can be implemented using the *blockupdate* argument, which requires that the user specifies m (assuming removal and infection times are known for the first m individuals), and the size of each block.

The output of this function is an object of class *'epictmcmc'*. There are three *S3* methods: *print.epictmcmc*, *summary.epictmcmc* and *plot.epictmcmc* that depend on the coda package. The latter function produced the trace plots of the posterior distributions of the model parameters with the same options of the *plot.mcmc* function in the coda package, such as, *start*, *thin*, *density*. In the case that *datatype* is set to either *"known removal"* or *"unknown removal"*, plots of the average posterior and 95% CI of the unobserved event times are also produced. To reproduce the same results, the random number generator of the *epictmcmc* Fortran function can be initialized to a fixed seed value through the option *seedval*.

Posterior predictive checks of the fitted model can be performed using the *datagen* function. This requires that the user supplies the model parameter values with a combined sample of the MCMC model parameter outputs. If desired, the simulation can be constrained to the first m infected individuals and their event times, which can be achieved by appending this information to the *initialepi* option.

2.5 Examples

2.5.1 Simulated Network-Based Epidemic

In this section, we illustrate the *EpiILMCT* package by fitting a simple $\mathcal{SIR}$ network-based continuous time ILM to a simulated epidemic. We consider an isolated population of 50 individuals distributed uniformly in an area of 10×10 units. We also consider a binary susceptibility covariate z which can be thought of as being an individual's treatment status. Thus, the infectivity rate given in Equation 2.2 becomes:

$$\lambda_{jt} = (\alpha_0 + \alpha_1 z_j) \sum_{i \in \mathcal{I}(t)} c_{ij}, \qquad \alpha_0, \alpha_1 > 0,$$

where the susceptibility function is represented by z; there are no transmissibility covariates; and $\epsilon = 0$. First, let us simulate the XY coordinates of individuals and the binary covariate z as follows:

```
R> loc <- matrix(cbind(runif(50, 0, 10), runif(50, 0, 10)),
+   ncol = 2, nrow = 50)
R> cov <- cbind(rep(1, 50), rbinom(50, 1, 0.5))
```

To simulate the epidemic, we generate a contact network using the *contactnet* function. Here, we use the power-law contact network model with $\beta = 1.8$ and $\alpha = 1$, as illustrated in the following code:

```
R>  net <- contactnet(type = "powerlaw", location = loc,
+    beta = 1.8, alpha = 1)
```

Figure 2.3 shows the contact network (grey lines). The epidemic is then generated using

the *datagen* function. Here, the epidemic is initialized with a randomly chosen infectious individual; then generated by providing the function with the contact network matrix, the susceptibility covariate and the following parameter values: $\alpha_0 = 0.08$, $\alpha_1 = 0.5$, and $\mathcal{D}_i \sim \Gamma(4, \delta = 2)$. Corresponding R code is

```
R> epi <- datagen(type = "SIR", kerneltype = "network",
+   kernelmatrix = net, suspar = c(0.08, 0.5),
+   delta = c(4, 2), suscov = cov)
```

All simulated data (i.e., XY coordinates of individuals, the contact network, the susceptibility covariate, and the simulated epidemic) are stored in the data file *NetworkData*, available in the *EpiILMCT* package.

```
R> data( "NetworkData", package = "EpiILMCT" )
R> names( NetworkData )
[1] "loc" "cov" "net" "epi"
R> NetworkData$epi
      id removal time infectious period infection time
1     44    2.2094793         2.2094793    0.000000000
2     45    1.6334200         1.6309073    0.002512692
3     32    1.7030359         1.5424120    0.160623835
...  ...          ...               ...            ...
32    16    7.3394708         2.3165409    5.022929824
33    12    6.6131883         1.5219635    5.091224749
34    48          Inf         0.0000000            Inf
...  ...         ...               ...            ...
48     4          Inf         0.0000000            Inf
```

```
49     2          Inf         0.0000000           Inf

50     1          Inf         0.0000000           Inf
```

We can illustrate the propagation of the epidemic using the *epiplot* function by specifying the argument *plottype* to *"propagation"*. To limit the number of plots, we assign the *time.index* option to be a vector containing time points for plots to be generated as shown in the following code:

```
R> epiplot( type = "SIR", kerneltype = "network",
+   epidat = NetworkData$epi, network = NetworkData$net,
+   location = NetworkData$loc, plottype = "propagation",
+   time.index = c(1, 2, 3, 4, 5, 6))
```

We now use the *epiplot* function to produce density plots of the infection and removal times, and a plot of the infectious periods via specifying the argument *plottype* to *"history"* as shown in the following code:

```
R> epiplot( type = "SIR", kerneltype = "network",
+   epidat = NetworkData$epi, network = NetworkData$net,
+   location = NetworkData$loc, plottype = "history")
```

Figure 2.2 shows the densities of the infection and removal times, and the infectious periods; while Figure 2.3 shows the epidemic propagation plot.

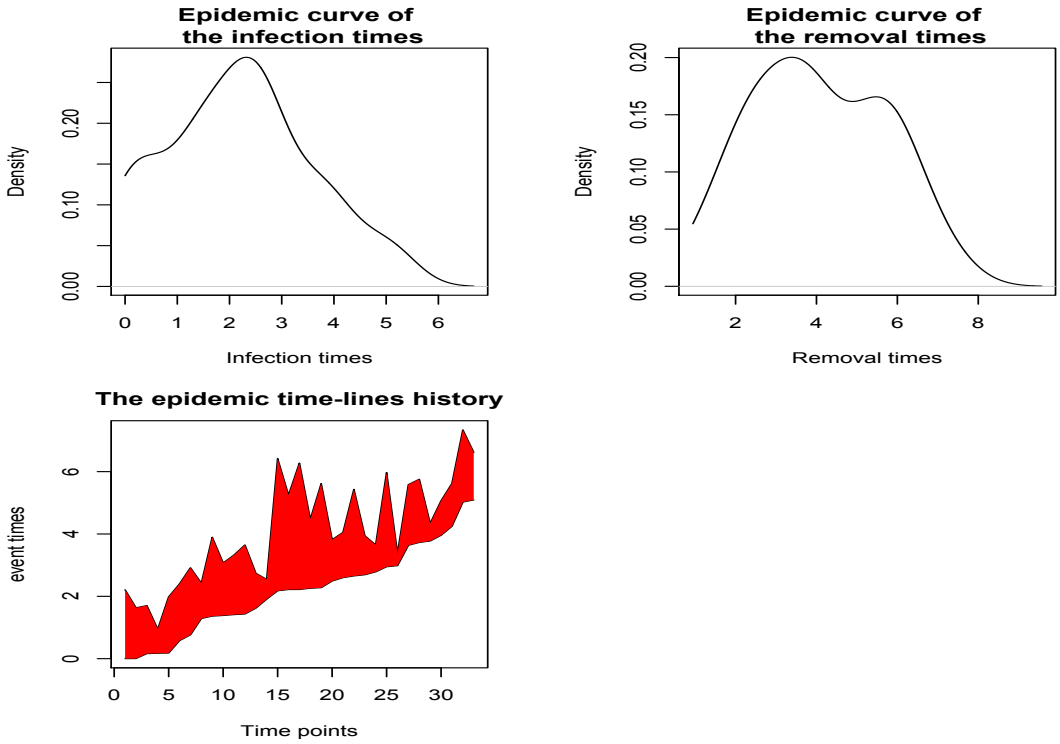


Figure 2.2: The epidemic curves of the infection and removal times for the epidemic that was generated using the simple network-based continuous time ILM. The red shaded area in the third plot represent the infectious periods.

To illustrate fitting continuous time ILMs to data, we analyze the epidemic using the *epictmcmc* function under two observation scenarios: *"known epidemic"* and *"known removal"*. For the former analysis, we assign $\Gamma(1, 0.1)$ gamma prior distributions to the model parameters α_0 and α_1 and use normal MCMC proposals with variances equal to 0.5 and 1, respectively. As we have two susceptibility parameters, the argument *sus.par* is a 2×5 matrix where each row represents the required information for updating each parameter. Now, we run the MCMC using the *epictmcmc* function for 150,000 iterations using the following code:

```
R> mcmc1 <- epictmcmc(epidat = NetworkData$epi,
```

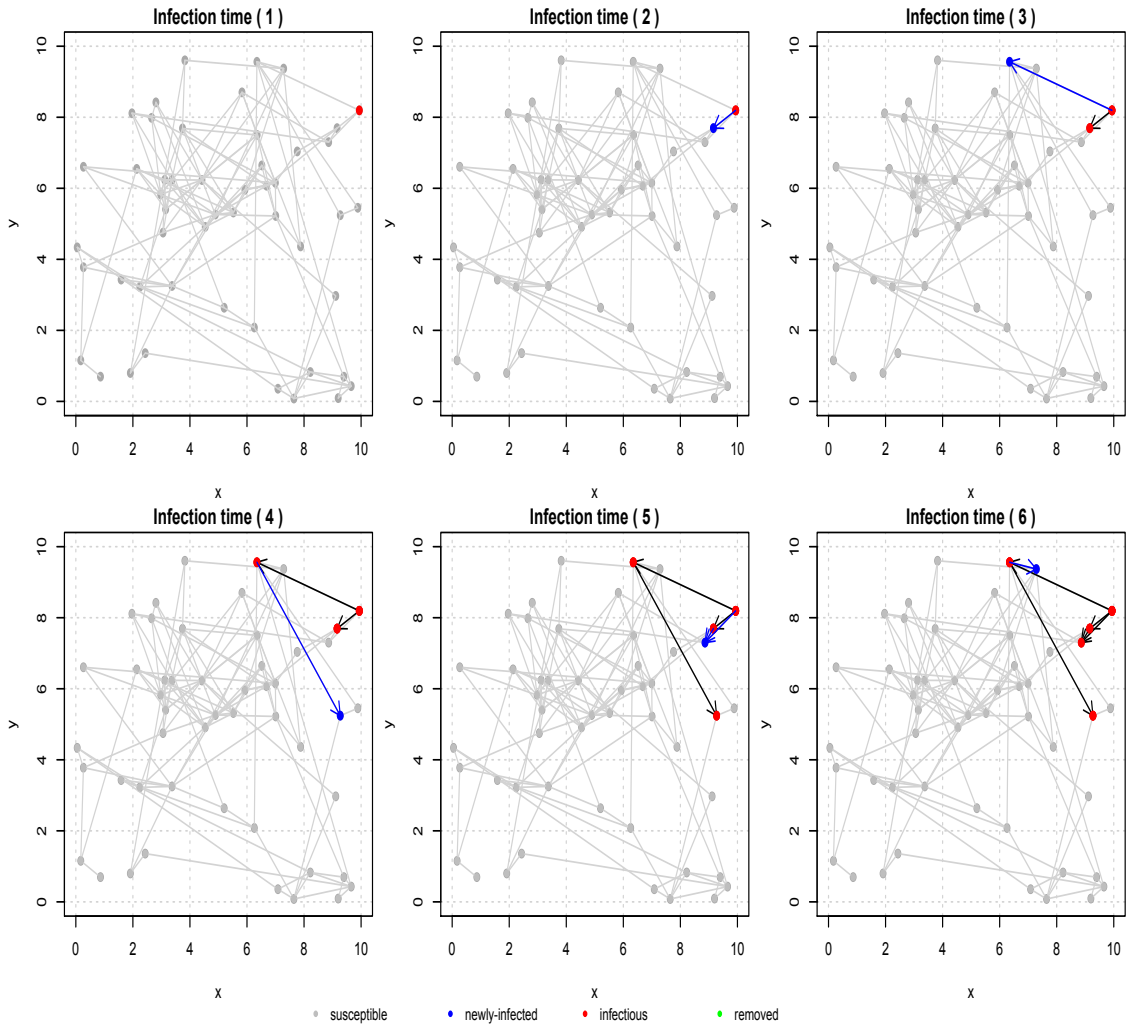


Figure 2.3: The directed pathway network of the generated epidemic over its time-line using the simple network-based ILMs.

```
+    type = "SIR", kerneltype = "network",
+    kernelmatrix = NetworkData$net,
+    datatype = "known epidemic", nsim = 150000,
+    sus.par = matrix(c(0.01, 0.1, "gamma", "gamma",
+    1, 1, 0.1, 0.1, 0.5, 1), ncol = 5, nrow = 2),
+    suscov = NetworkData$cov, seedval = 524837)
```

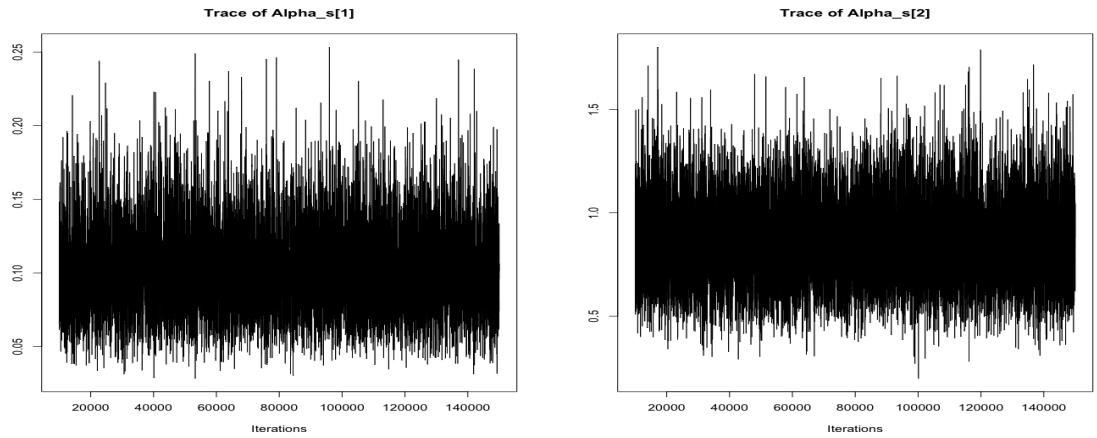


Figure 2.4: The MCMC chains of the posterior distributions of the model parameters fitting the simulated epidemic using the network-based continuous time ILM assuming a fully observed epidemic.

The MCMC trace plots for the model parameters can be produced using the S3 method *plot.epictmcmc*.

```
R> plot(mcmc1, epidat = NetworkData$epi, start = 10000,
+   thin = 10, density = FALSE)
```

Figure 2.4 shows the MCMC trace plots for the model parameters α_0 and α_1. We observe a posterior mean of $\hat{\alpha}_1$ = 0.8562 with 95% credible (percentile) interval (0.4954, 1.2949) and a posterior mean of $\hat{\alpha}_0$ = 0.1021 with 95% credible interval (0.0530, 0.1683). We also observed well-mixed MCMC chains for both model parameters. Note here that α_1 is slightly overestimated although the 95% credible interval contains the true value; this is just a facet of the data set simulated.

For the known removal times analysis, *EpiILMCT* uses data augmented MCMC to infer infection times and the infectious period rate. Here, we assume that the infectious period

follows a gamma distribution with shape 4 and unknown rate parameter δ; so $\mathcal{D}_i \sim \Gamma(4, \delta)$. Here, we also include a spark term ϵ. This is not strictly necessary but tends to improve MCMC mixing. We assigned gamma prior distribution $\Gamma(4, 2)$ for δ and an exponential prior distribution with rate 0.01 for ϵ.

```
R> mcmc11 <- epictmcmc(epidat = NetworkData$epi, type = "SIR",
+   kerneltype = "network", kernelmatrix = NetworkData$net,
+   datatype = "known removal", nsim = 150000,
+   sus.par = matrix(c(0.01, 0.1, "gamma", "gamma", 1, 1,
+   0.1, 0.1, 0.2, 0.8), ncol = 5, nrow = 2),
+   spark.par = c(0.01, "gamma", 1, 0.01, 0.1),
+   delta = matrix(c(4, 2, 4, 2), ncol = 4, nrow = 1),
+   suscov = NetworkData$cov, periodproposal = c(1, 0.5),
+   seedval = 524837)
```

Figure 2.5 shows typical MCMC trace plots for the model parameters α_0, α_1, ϵ and δ. Well-mixed MCMC chains are observed for all the model parameters. The posterior means and 95% credible intervals of these parameters are as follows: $\hat{\alpha}_0 = 0.0993$ (0.0300, 0.1896), $\hat{\alpha}_1 = 0.8626$ (0.4149, 1.4985), $\hat{\epsilon} = 0.0205$ (0.0007, 0.0647), and $\hat{\delta} = 2.2678$ (1.6840, 2.9806). The infection times are also well-approximated (see Figure 2.6). Figures 2.5 and 2.6 are produced using the following code:

```
R> plot(mcmc11, NetworkData$epi, start = 10000, thin = 10, density = FALSE)
```

To check the fit of the model, we consider the posterior predictive distribution of four statistics. Specifically, we consider: T_1, the total number of infected individuals; T_2, the

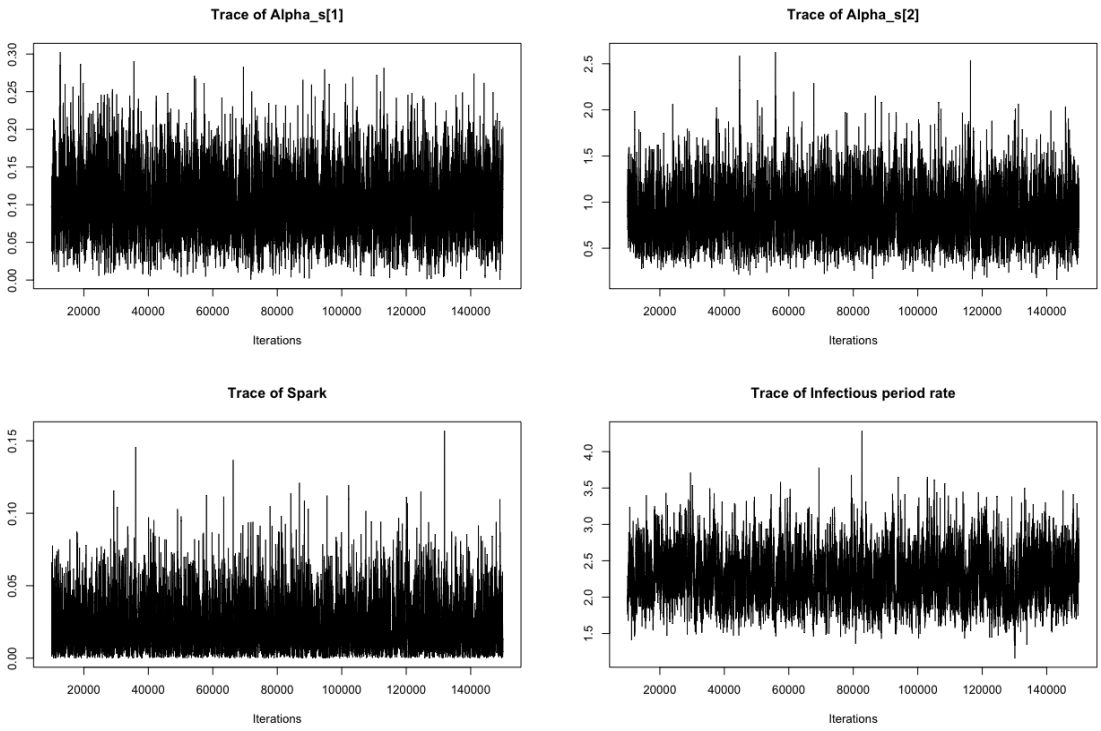


Figure 2.5: The MCMC chains of the posterior distributions of the model parameters for fitting the simulated epidemic using the network-based continuous time ILM assuming a partially observed epidemic (unknown infection times).

average removal time; T_3, the variance of the removal times; and T_4, the length of the epidemic. Here, we simulate 10,000 epidemics based on random draws of the model parameters from the MCMC output (excluding burn-in) of the known removal times analysis (i.e., unknown infection times). We conditioned our simulation on the first ten infected individuals, then calculated the four statistics for each simulation using the following code:

```
R> set.seed(666)

R> mb <- sample(seq(10000,150000),10000)

R> numinf <- NULL

R> muremtime <- NULL
```

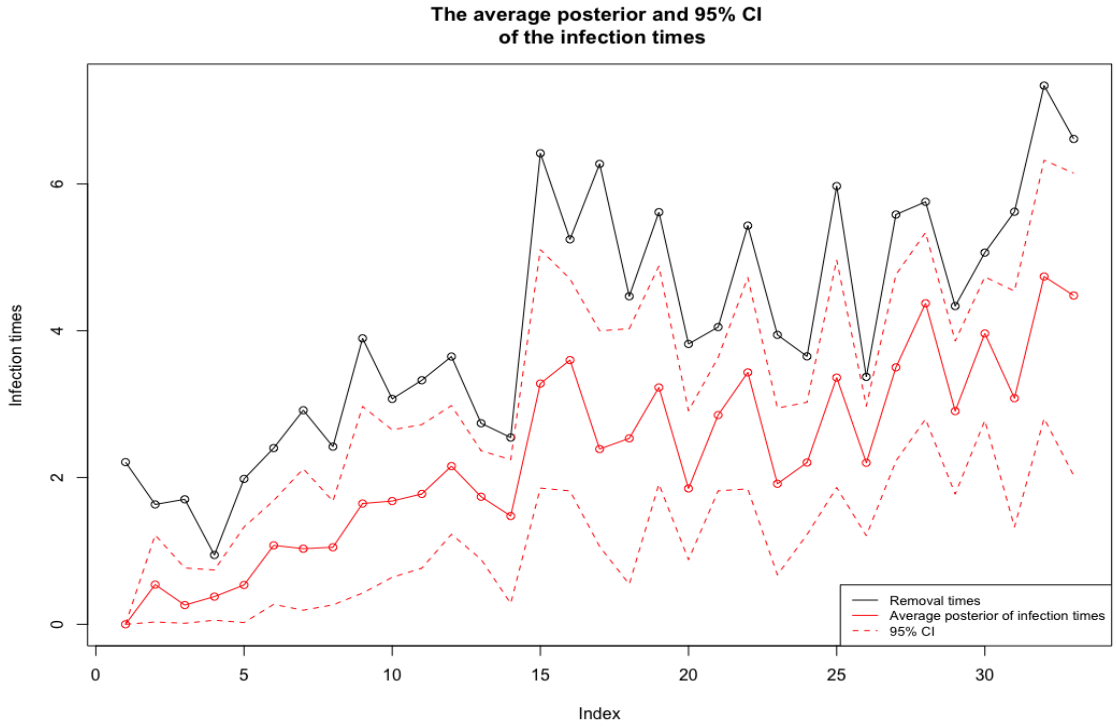


Figure 2.6: The posterior means and 95% credible intervals of the infection times for fitting the simulated epidemic using the network-based continuous time ILM assuming a partially observed epidemic (unknown infection times). The black line represents the observed removal times; the solid red line represents the posterior means; and dotted red lines represent the 95% credible interval.

```
R> varremtime <- NULL
R> lengthepi <- NULL
R> for (i in 1:10000) {
+   predepi <- datagen(type = "SIR", kerneltype = "network",
+       kernelmatrix = NetworkData$net,
+       initialepi = matrix(NetworkData$epi[1:10, ], ncol = 4,
+       nrow = 10), suspar = c(mcmc11$parameter.samples[mb[i], 1],
+       mcmc11$parameter.samples[mb[i], 2]),
+       spark = mcmc11$parameter.samples[mb[i], 3],
```

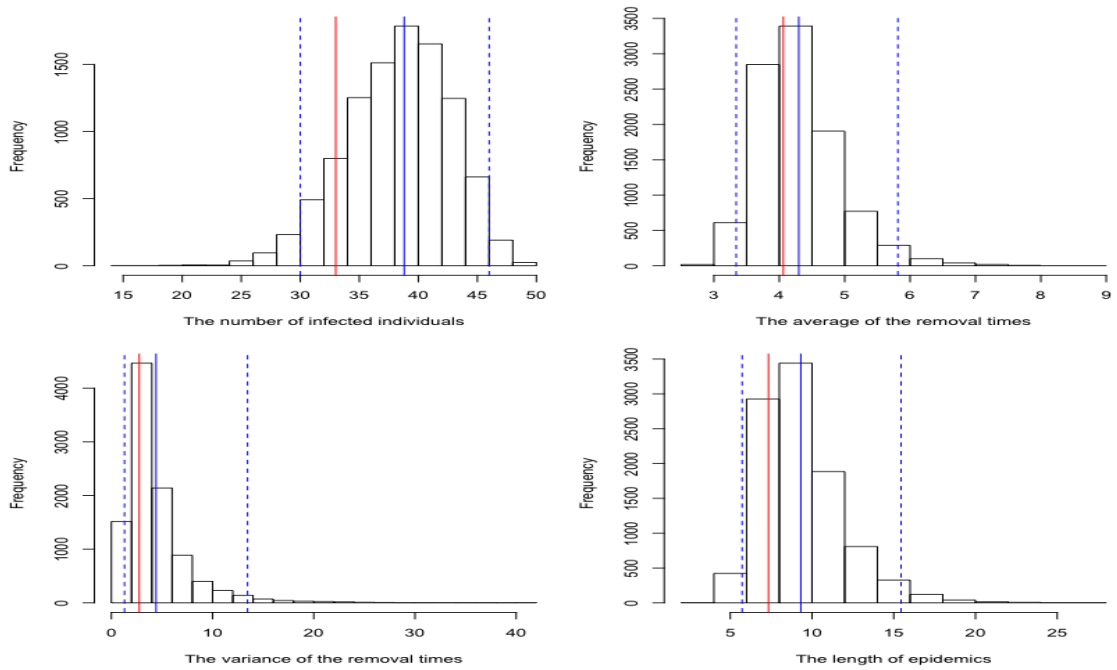


Figure 2.7: The posterior predictive distributions of four statistics: the number of infected individuals, the average removal times, the variance of removal times, and the length of epidemic for fitting a partially observed epidemic (unknown infection times) using network-based continuous time ILM. The red vertical lines represent the observed statistic values and the solid and dotted blue vertical lines represents the posterior predictive means and 95% credible intervals of the four statistics, respectively.

```
+       delta = c(4, mcmc11$parameter.samples[mb[i], 4]),
+       suscov = NetworkData$cov)
+    numinf[i] <- sum( predepi[, 2] != Inf )
+    muremtime[i] <- mean( predepi[1:numinf[i], 2] )
+    varremtime[i] <- var( predepi[1:numinf[i], 2] )
+    lengthepi[i] <- max( predepi[1:numinf[i], 2] )
+  }
```

The posterior predictive distributions of the four statistics are shown in Figure 2.7. We

can see that each distribution captures the observed statistics well.

2.5.2 Case Study: Tomato Spotted Wilt Virus (TSWV) Data

We further illustrate the *EpilLMCT* package by analyzing the TSWV data as described in Hughes et al. (1997) and analyzed with spatial ILMs by Pokharel and Deardon (2014, 2016). These data represent the results of an experiment designed to study the spread of the disease amongst 520 pepper plants raised in a greenhouse. Plants were evenly distributed across a 10×26 meter area as shown in Figure 2.8. The experiment began on May 26, 1993 and finished on August 16, 1993. Plants were checked for the disease every 14 days, and ultimately 327 were infected. Following Pokharel and Deardon (2014, 2016) these observation points are recorded to $t = 1, 2, \ldots, 7$. We set the initial infection time to t = 2 in line with the original data set.

Here, we analyze the epidemic under two data availability scenarios. First, we assume that the event times of the TSWV disease are fully observed. Here, the infectious period was fixed at three time points (42 days) following Pokharel and Deardon (2014, 2016). Additionally, the last observed time point was at $t = 7$. Second, we assume the epidemic is partially observed. Specifically, we assume that the infection and removal times are unknown, and treat the reported infection times as the notified time points. This entails considerable uncertainty and makes the MCMC analysis much more time consuming (more than 13 times longer than the computation time of the first analysis), because it is necessary to estimate both incubation and delay periods along with the infection and removal times.

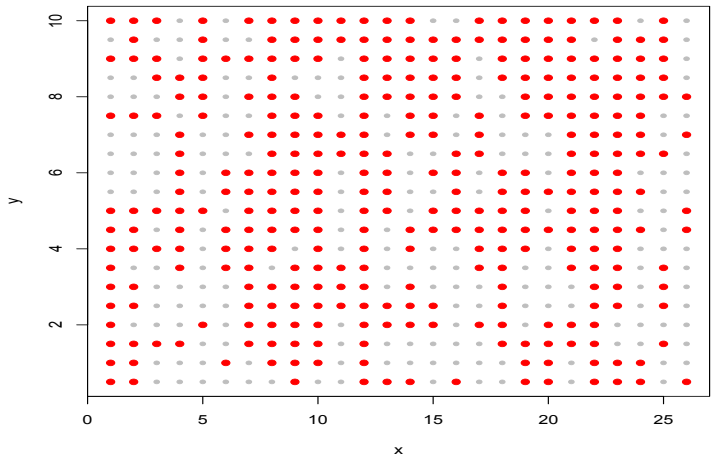


Figure 2.8: A grid plot of the XY coordinates of plants from the TSWV data set. The red plots represent infected plants at the end of the experiment.

The data is stored in the data file *tswv*, available in the *EpiILMCT* package. It contains a list of the XY coordinates and the TSWV epidemic data set of the two compartmental frameworks ($\mathcal{SIR}$ and $\mathcal{SINR}$) with the same format as the *datagen* output.

The following code shows how the TSWV data set can be extracted and the associated Euclidean distance matrix built.

```
R> data( "tswv", package = "EpiILMCT" )

R> names( tswv )

[1] "loc"    "tswvsir"    "tswvsinr"

R> plot( tswv$loc, col = "gray", pch = 19 )

R> k1 <- sum( tswv$tswvsir[,2] != Inf)

R> points( tswv$loc[tswv$tswvsir[1:k1, 1], ], col = "red", pch = 19 )

R> dis <- as.matrix( dist(tswv$loc, method = "euclidean") )
```

Following Pokharel and Deardon (2014, 2016), we implement the distance-based con-

tinuous time ILM with power-law kernel and without susceptibility and transmissibility covariates. For the first analysis, an $\mathcal{SIR}$ model is used where the infectivity rate of Equation 2.2 becomes:

$$\lambda_{jt} = \left(\alpha \sum_{i \in \mathcal{I}(t)} d_{ij}^{-\beta} \right), \qquad \alpha, \beta > 0.$$

To perform the MCMC, the *epictmcmc* function should be used with *datatype* set to *"known epidemic"*. Here, we assume exponential prior distributions with rate 0.01 for the model parameters α and β; and we request 150,000 MCMC samples. The code to achieve this is as follows:

```
R> tswv.full.observed <- epictmcmc(epidat = tswv$tswvsir,
+   type = "SIR", kerneltype = "distance", kernelmatrix = dis,
+   distancekernel = "powerlaw", datatype = "known epidemic",
+   nsim = 150000, sus.par = c(0.02, "gamma", 1, 0.01, 0.01),
+   kernel.par = c(2, "gamma", 1, 0.01, 0.1), seedval = 524837)
```

Figure 2.9 shows the resulting MCMC chains for the model parameters with a burn-in of 10,000 iterations and thinning interval of 10. The posterior mean of α and β were $\hat{\alpha} = 0.0122$ and $\hat{\beta} = 1.3167$, with 95% credible intervals of (0.0078, 0.0170) and (1.0067, 1.5992), respectively. The estimates of $\hat{\alpha}$ and $\hat{\beta}$ are consistent with those of Pokharel and Deardon (2014, 2016). The above *epictmcmc* function had a run time of 2.65 hours on an Apple iMac with i5-core Intel 2.9 GHz processors with 16 GB of RAM.

In the second analysis (i.e., where infection and removal times are treated as unknown), we assume notified times were observed for all infected individuals. Consequently, an

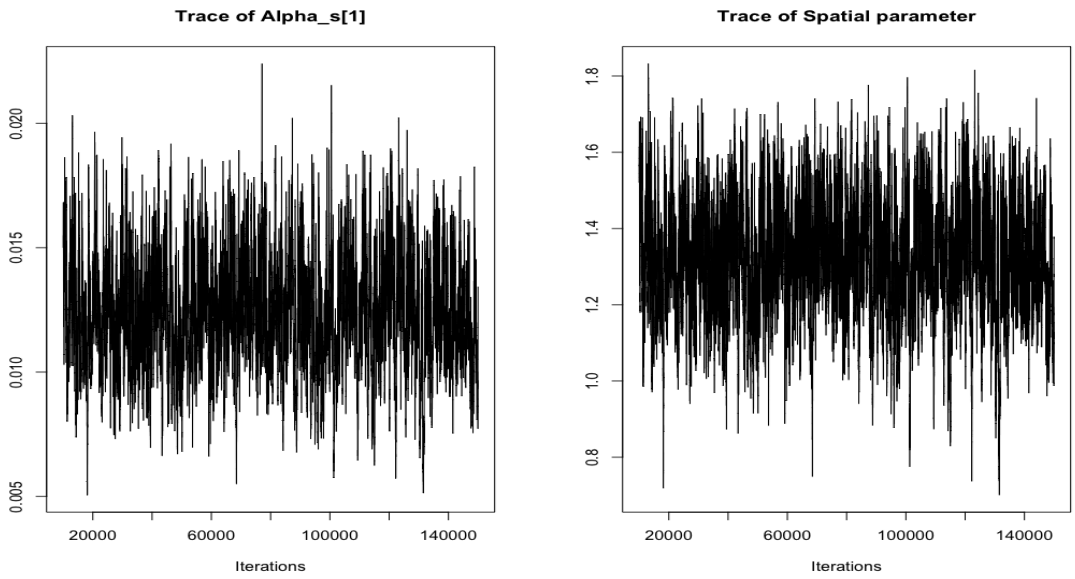


Figure 2.9: The MCMC chains of the posterior distributions of the model parameter α and β for fitting the fully observed TSWV data using the $\mathcal{SIR}$ distance-based continuous time ILM.

$\mathcal{SINR}$ model is used where the infectivity rate of Equation 2.3 becomes:

$$\lambda_{jt} = \left(\alpha \sum_{i \in \mathcal{N}^-(t)} d_{ij}^{-\beta}\right) + \gamma \left(\alpha \sum_{i \in \mathcal{N}^+(t)} d_{ij}^{-\beta}\right).$$

We here assume the risk of infection does not reduce after notification, and set the notification effect parameter to γ = 1. The infectious period here is divided into the incubation and delay periods. We assume the total infectious period to be within three time points (42 days) following Pokharel and Deardon (2014, 2016); Brown et al. (2005). Thus, we assumed the incubation periods to follow an exponential distribution such that $D_i^{(inc)} \sim$ Exp($\delta^{(inc)}$) with initial value of $\delta^{(inc)}$ = 1, whereas the delay periods are assumed to follow a gamma distribution such that $D_i^{(delay)} \sim \Gamma(10, \delta^{(delay)})$ with initial value of $\delta^{(delay)}$ = 5. We assign gamma prior distributions for both rates such that $\delta^{(inc)} \sim \Gamma(10, 10)$ and

$\delta^{(delay)} \sim \Gamma(60, 12)$. These choices are to cover the support of our assumptions about the infectious periods. For simplicity, we assume the infection time of the first infected plant is known. We set its incubation period to one time point.

We assign exponential prior distributions with rate 0.01 to the model parameters α and β. To perform the MCMC, we use the *epictmcmc* function with *type* and *datatype* set to *"SINR"* and *"unknown removal"*, respectively. At each iteration, the infection and removal times are updated in blocks of 10 randomly selected individuals (*blockupdate*). We can then use the *epictmcmc* function to output 150,000 samples from the posterior distributions via the following code:

```
R> tswv.unknown.remov.infect1 <- epictmcmc(epidat = tswv$tswvsinr,
+   type = "SINR", kerneltype = "distance", kernelmatrix = dis,
+   distancekernel = "powerlaw", datatype = "unknown removal",
+   blockupdate = c(1, 10), nsim = 150000,
+   sus.par = c(0.02, "gamma", 1, 0.01, 0.01),
+   kernel.par = c(1.3, "gamma", 1, 0.01, 0.1),
+   delta = matrix(c(1, 10, 1, 5, 10, 60, 10, 12), ncol = 4,
+   nrow = 2), seedval = 524837)
```

Figure 2.10 shows the trace plots and posterior density plots for the model parameters α, β, $\delta^{(inc)}$ and $\delta^{(delay)}$ with a burn-in of 10,000 iterations and thinning interval of 10. The posterior means and 95% credible intervals of these parameters are given in Table 2.3. The MCMC chains show good mixing of all the model parameters. Figures 2.11 and 2.12 show the average posterior and 95% credible interval of the infection and removal times.

Table 2.3: The posterior means and 95% credible intervals (CIs) of the model parameters for fitting the TSWV data set using the $\mathcal{SINR}$ distance-based continuous time ILM under the assumption of unknown removal and infection times.

	α	β	$\delta^{(inc)}$	$\delta^{(delay)}$
Mean	0.0426	2.0353	2.9936	9.1222
95% CI	(0.0344, 0.0510)	(1.7830, 2.2761)	(2.2718, 3.8697)	(8.0206, 10.2730)

The average posterior (95% CI) of the incubation and delay periods were 0.3193 (0.2348, 0.4168) and 1.0845 (0.9557, 1.2303) time points, respectively, which indicate an average infectious period of 19.6524 days (1.4037 time points).

To surmount the computational time issue of performing the MCMC, we implemented block updates of infection and removal times. The *epictmcmc* function had a run time of 36.52 hours on an Apple iMac with i5-core Intel 2.9 GHz processors with 16 GB of RAM, but this was computationally much more efficient than if single updates were used ($\approx$ 328 hours, calculated based on a single iteration).

2.6 Conclusion

This paper introduces the R software package *EpiILMCT*, which facilitates the use of a broad range of continuous time ILMs under two compartmental frameworks ($\mathcal{SIR}$ and $\mathcal{SINR}$). It also allows for the analysis of partially observed infectious diseases data, achieved using data augmented MCMC within a Bayesian framework. We illustrated the

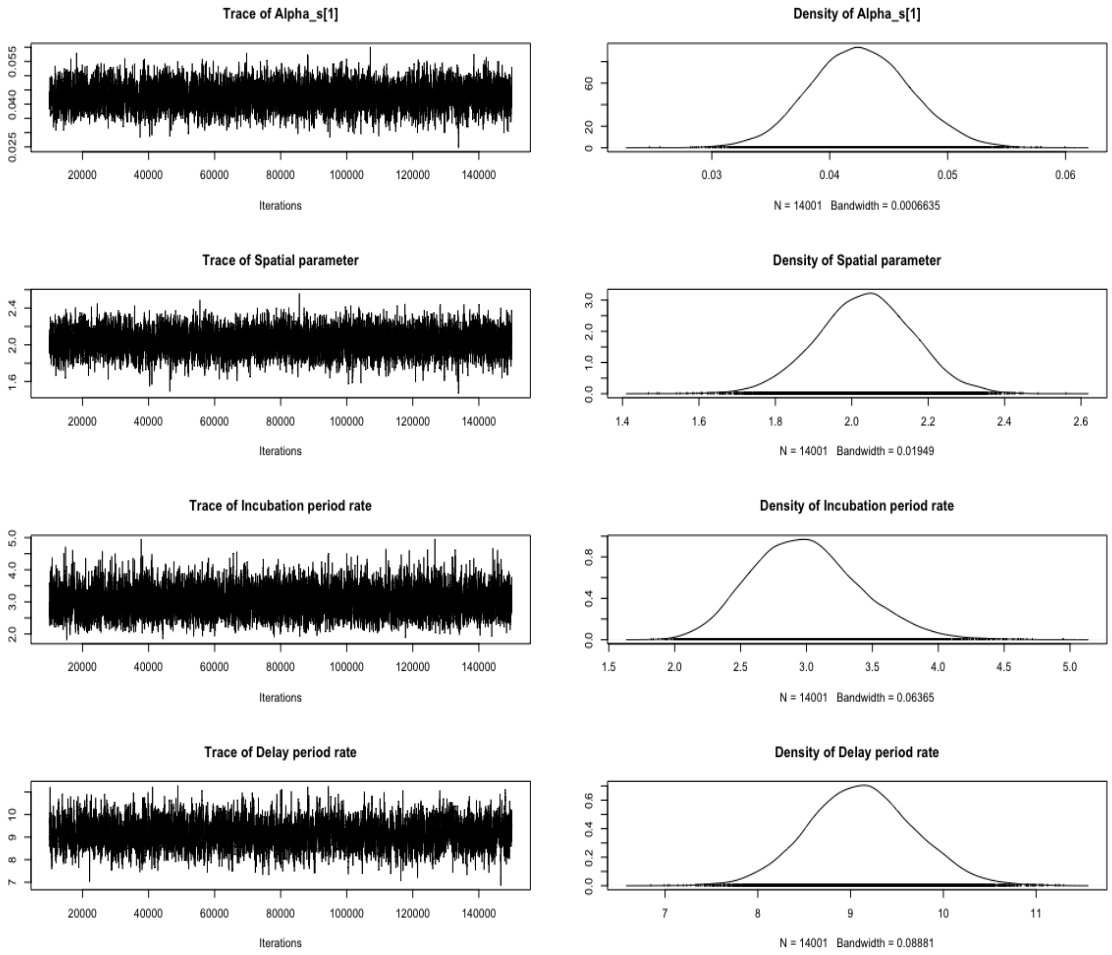


Figure 2.10: The MCMC chains and density plots of the posterior distributions of the model parameter α, β, $\delta^{(inc)}$ and $\delta^{(delay)}$ for fitting the partially observed TSWV data (unknown infection and removal times) using the $\mathcal{SINR}$ distance-based continuous time ILM.

package by fitting continuous time ILMs on simulated and real epidemic data. The paper did not cover all functionality of the package. For instance, we did not illustrate incorporating both distance and network in the kernel function, or allowing for nonlinearity between the susceptibility and transmissibility risk factors and the infection rate. However, implementation of such facets is simple. Additional functionality that was not covered in Sections 5 can be found via *help(package = "EpiILMCT").*

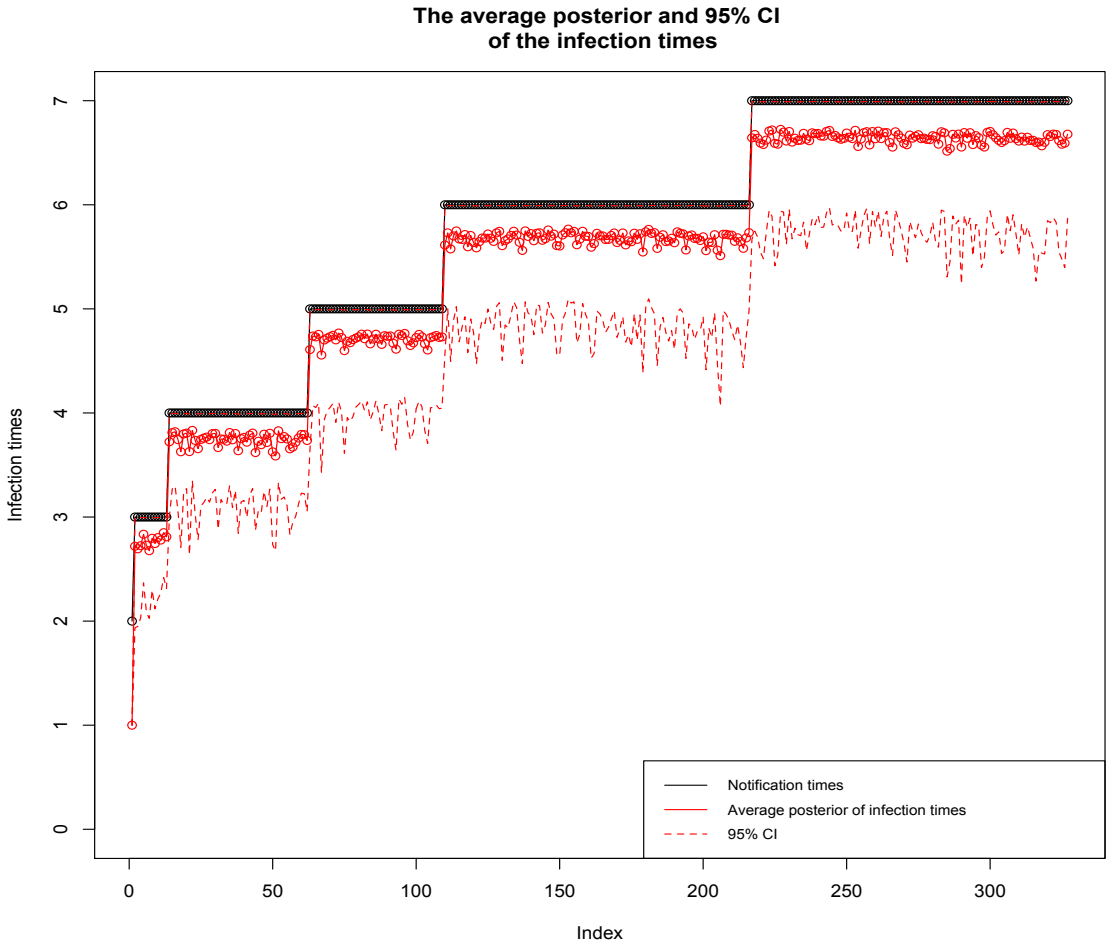


Figure 2.11: The posterior means (solid red line) and 95% credible intervals (dotted red lines) of the infection times for fitting the partially observed TSWV data (unknown infection and removal times) using the $\mathcal{SINR}$ distance-based continuous time ILM. The black line represents the observed notification times.

It is also possible to use *EpiILMCT* to test the efficacy of disease control strategies (e.g., vaccination or culling) via simulation studies. Such testing requires more sophisticated coding than is described here, but can be done by simulating epidemics in small time steps and then manipulating infection and/or removal times according to the relevant control policy, before simulating the next step of the epidemic simulation.

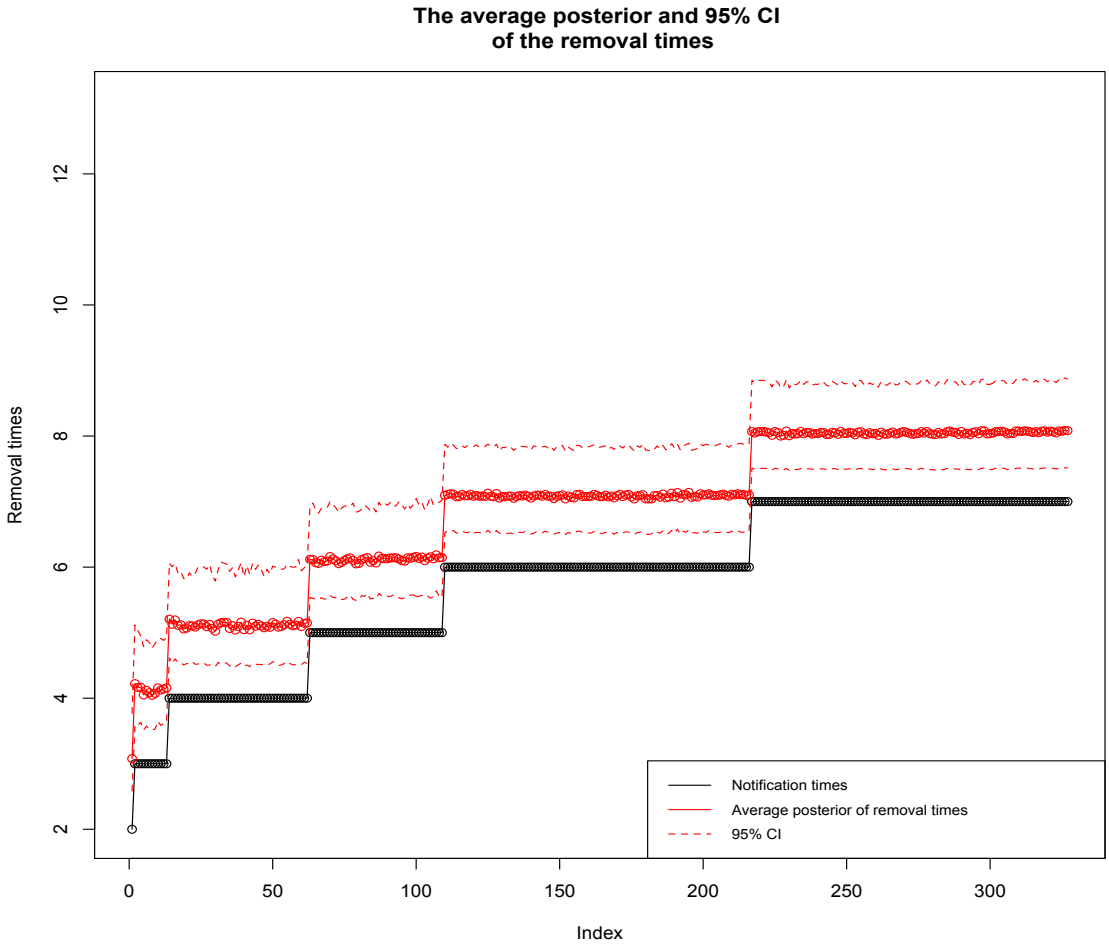


Figure 2.12: The posterior means (solid red line) and 95% credible intervals (dotted red lines) of the removal times for fitting the partially observed TSWV data (unknown infection and removal times) using the $\mathcal{SINR}$ distance-based continuous time ILM. The black line represents the observed notification times.

In terms of future developments, incorporating more data uncertainty into the analyses, especially under the network-based model, is an option for future development of this package *EpiILMCT*. For example, networks are often only partially observed. However, the data augmentation could easily make the computation time for data analyses prohibitive. Various strategies for mitigating this might be available. For example, openMP parallel

programming could be used to reduce the computation time associated with likelihood calculations. The following approximate forms of inference are useful for overcoming computational burden: Gaussian process emulation (Pokharel and Deardon, 2016), approximate Bayesian computation (Beaumont et al., 2009), machine learning based model classification (Pokharel and Deardon, 2014), data-sampled likelihood approximation (Malik et al., 2016), or data-aggregation (Deeth and Deardon, 2016).

Chapter 3

Spatial Contact Network Uncertainty in Individual Level Models of Infectious Disease Transmission[1]

3.1 Introduction

In infectious disease epidemiology, understanding the dynamics of infectious disease spread is essential for reaching more accurate and reliable decisions that can help in controlling disease spread. Spatio-temporal models have been of great importance in various applications to determine the transmission dynamics of infectious diseases, for example the 2001 UK foot-and-mouth disease outbreak (Malik et al., 2016; Deeth et al., 2013; Deardon et al., 2010; Jewell et al., 2009; Ster and Ferguson, 2007; Keeling et al., 2001), tomato spotted wilt virus disease (TSWV) (Pokharel and Deardon, 2016), and the 2009 influenza pandemic in the US (Gog et al., 2014).

However, often contact network-based models in which the underlying process of disease transmission is explained by a binary or weighted contact network or series of such

[1]Chapter 3 has been submitted to *Epidemiologic Methods*, and is under review.

networks have also been found beneficial in understanding the spread mechanism of infectious diseases. A network is a graph of connections between individuals (e.g., individuals within a population, species within a community) that defines possible transmission paths of the disease. Such networks can represent social interactions, trade, or movements, for example. Incorporating such networks in disease transmission models can enhance the prediction of the epidemic pattern and thus aid in controlling the spread of the disease (Danon et al., 2011). Generally, we would also expect that such contact networks would be spatial in nature, with stronger connections between individuals close geographically compared to more distant individuals.

Most previous studies involving network-based transmission models assumed a known network, or used simulation techniques to generate networks that have certain properties related to individual habits and their interactions, with epidemics then being simulated across the networks (Volz, 2008; Barthélemy et al., 2005). However, underlying contact networks for epidemics are often not observed in real life, especially for large population sizes, since collecting such contact information is challenging in many ways. Personal information that could be used as the base of defining the contact network is usually difficult to obtain from individuals, for example, sexual history in the study of sexually transmitted diseases. The problem of recall error is also an issue, especially when individuals have many contacts. The time-consuming and expensive nature of collecting these contact data are another problem that could affect the practicality of employing a contact network (Keeling and Eames, 2005).

Another common issue when modelling infectious disease spread is the partially observed nature of epidemic data. In real life, infection times of individuals are usually unknown (or censored), with a diagnosis or reporting time more likely to be observed. Thus, methods that make explicit inference that take into account such data uncertainties are needed to understand disease spread. This has been done in the context of modelling the transmission (rather than contact) network and uncertain infection times. Britton and O'Neill (2002) incorporate modelling the transmission networks within homogeneous stochastic epidemic models using a Bernoulli random graph. They showed that parameters of the network and epidemic process can be successfully estimated under a partially observed disease system (known removal times). Their work was extended by Groendyke et al. (2012) in studying a measles epidemic in children in Germany. They modelled the network using exponential-family random graph models (ERGMs), and applied a stochastic susceptible-exposed-infectious-removed ($\mathcal{SEIR}$) transmission model while incorporating other covariates.

Our approach here differs in the sense that we incorporate the contact network as a covariate in the model of disease spread through a heterogeneous population. We use a modified version of the general individual level model (ILM) framework for studying infectious disease spread introduced by Deardon et al. (2010). The ILMs are flexible in that they can account for covariates associated with the individual's chances of being infected by a disease, and contain a kernel function that represents shared risk factors between pairs of infectious and susceptible individuals. A network-based continuous time version of these

ILMs is considered within a Bayesian framework, under a scenario in which the underlying contact network is inherently spatial in nature. However, here, we assume the underlying contact network, infection times and infectious periods are unknown. We incorporate this missing information into our model through the use of a data augmented Markov chain Monte Carlo (DA-MCMC) method.

We illustrate the fit of network-based continuous time ILMs on simulated epidemics that are generated through an undirected spatial contact network. In particular, we explore the effect of incorporating different levels of uncertainty about the contact network into the model on the posterior estimates of the model parameters. We also investigate to what extent adding global information about the contact network through an observation model can affect the estimation of model parameters and contact network summary statistics. We show that accounting for the contact network uncertainty in the model results in good estimation of the model parameters, albeit with an increase in their posterior variances. Also, we observe good approximation of the contact network and better estimates of the model parameters by conditioning the update of connections of the contact network on the total number of connections.

This chapter is structured as follows. In Section 2, we introduce the continuous time network-based ILMs and contact networks considered. Also, as this chapter is based on analyzing simulated epidemics, we explain, in the same section, the steps of simulating epidemics based on our model. We then detail the MCMC parameter estimation in Section 3 followed by details of the epidemic simulation studies. The results are described in

Section 5, with a final discussion presented in Section 6.

3.2 The Model

3.2.1 The General Continuous Time ILMs

Deardon et al. (2010) present a general framework for discrete time ILMs for infectious disease transmission. We modify this framework, setting the models in continuous time rather than discrete time. We first detail this new framework in the context of the susceptible-infected-removal ($\mathcal{SIR}$) compartmental class of models.

In a continuous time $\mathcal{SIR}$ model, each individual i is assumed to be in one of three states at each time point: $\mathcal{S}$ (susceptible), $\mathcal{I}$ (infectious), or $\mathcal{R}$ (removed). Transitions between states that we allow are: from susceptible to infectious and from infectious to removed. That is, individuals are assumed to be in the susceptible state until they become infected with the disease. As soon as they become infected, they are assumed to be infectious and are able to transmit the disease to other susceptible individuals during their infectious period. Following the course of their infectious period, an individual transitions to the removed state. The removed population may be those who recover and become immune to the infection, die, or for some other reason become detached from the transmission of the infection (e.g., are quarantined).

The full epidemic history then constitutes the state of individuals at each point in time during the epidemic as $\mathcal{S}(t)$, $\mathcal{I}(t)$, and $\mathcal{R}(t)$ for $t \in [0, T_{obs}]$, where T_{obs} is the maximum

removal time and represents the time the last infectious individual enters the removed state. Each susceptible individual j is assumed to have a rate of infectivity λ_{jt} at time t:

$$\lambda_{jt} = \left[\Omega_S(j) \sum_{i \in \mathcal{I}(t)} \Omega_T(i)\kappa(i,j)\right] + \epsilon(j,t), \tag{3.1}$$

where $\mathcal{I}(t)$ is the set of infectious individuals at time t; $\Omega_S(j)$ and $\Omega_T(i)$ are functions of potential risk factors related to susceptible individual j contracting, and infectious individual i transmitting, the disease, respectively; $\kappa(i,j)$ is an infection kernel that represents shared risk factors between pairs of infected and susceptible individuals; and "random appearing" infections may be introduced by the spark term $\epsilon(j,t)$ (e.g., this might represent infection of a susceptible individual by an infectious individual from outside the observed population).

Likelihood Function

The data we consider are the removal times of individuals, and their corresponding risk factors, in a population of size N. Also, the data could consist of infection times, but these will generally be unknown or at least censored. We consider here only complete epidemics; that is, the last individual recorded in our data set has been through the infectious period and entered the removed state. Note however, this assumption is not a requirement of our approach and can be easily relaxed. Having m infected individuals, we label them $i = 1, 2, \ldots, m$ corresponding to their infection (I_i) and removal (R_i) times; we label the $N - m$ individuals who remain uninfected $m + 1, m + 2, \ldots, N$ with $I_i = R_i = \infty$. We

then denote infection and removal time sets for the population as $\boldsymbol{I} = \{I_1, \ldots, I_m\}$ and $\boldsymbol{R} = \{R_1, \ldots, R_m\}$, respectively. Infectious periods are assumed to follow an exponential distribution with rate δ. The infectious period could also follow some other non-negative distribution, for example a gamma distribution (Jewell et al., 2009). The likelihood is then given as follows:

$$\begin{aligned}
L(\boldsymbol{I}, \boldsymbol{R}|\boldsymbol{\theta}) &= \prod_{j \neq k} \left(\epsilon + \sum_{i: I_i < I_j \leq R_i} \Omega_S(j)\Omega_T(i)\kappa(i,j) \right) \\
&\times \exp\left\{ -\int_{I_k}^{T_{obs}} \left(\sum_{i \in \mathcal{S}(t)} \epsilon + \sum_{i \in \mathcal{I}(t)} \sum_{j \in \mathcal{S}(t)} \Omega_S(j)\Omega_T(i)\kappa(i,j)(u - I_i) \right) du \right\} \prod_{i=1}^{m} f(\gamma_i; \delta) \\
&= \prod_{j \neq k} \left(\epsilon + \sum_{i: I_i < I_j \leq R_i} \Omega_S(j)\Omega_T(i)\kappa(i,j) \right) \\
&\times \exp\left\{ -\sum_{i=1}^{m} \left(\sum_{j=1}^{N} ((R_i \wedge I_j) - (I_i \wedge I_j))\Omega_S(j)\Omega_T(i)\kappa(i,j) \right) \right\} \\
&\times \exp\left(-\epsilon \sum_{j=1}^{N} [(T_{obs} \wedge I_j) - I_k] \right) \prod_{i=1}^{m} f(\gamma_i; \delta),
\end{aligned}$$

where the wedge symbol $\wedge$ denotes the minimum operator; $\boldsymbol{\theta}$ is the vector of unknown parameters; k is the first infected individual with I_k its corresponding infection time; $f(\bullet; \delta)$ indicates the exponential density of the infectious periods with rate δ; and γ_i is the infectious period of infected individual i defined as $\gamma_i = R_i - I_i$. The integral, which represents the total person-to-person infectious pressure through the course of the epidemic, can be simplified by the product of the two exponent terms in the lower equation (Jewell et al., 2009; Neal and Roberts, 2005; Britton and O'Neill, 2002).

The Network-Based ILM

All of the ILMs considered here are network-based assuming that the epidemic spreads in a finite, closed population of size N between individuals through a single, binary, undirected time-homogeneous contact network. Specifically, we assume $\kappa(i, j) = c_{ij}$, where

$$c_{ij} = \begin{cases} 1, & \text{if a connection exists between individuals } i \text{ and } j, \text{ where } i \neq j. \\ 0, & \text{otherwise.} \end{cases}$$

We will consider a simple network-based ILM for use in a simulated data setting as proof of concept. This simple model contains a single binary susceptibility covariate with no transmissibility covariates (i.e., $\Omega_T(i) = 1$) or sparks infections (i.e., $\epsilon = 0$). Specifically, the infectivity rate is given by

$$\lambda_{jt} = \left[(\alpha_0 + \alpha_1 z_j) \sum_{i \in \mathcal{I}(t)} c_{ij} \right] + \epsilon, \qquad \epsilon \geq 0, \quad \alpha_0, \alpha_1 > 0, \tag{3.2}$$

where α_0 is a baseline infectivity parameter; z_j is a binary covariate that could represent, for example, the vaccination status of individuals; and α_1 is the covariate effect. Because we assume $\epsilon = 0$, the infectivity rate for an individual with no connections to infectious individuals at time t will be zero under this model.

3.2.2 Spatial Contact Network and Epidemic Simulation

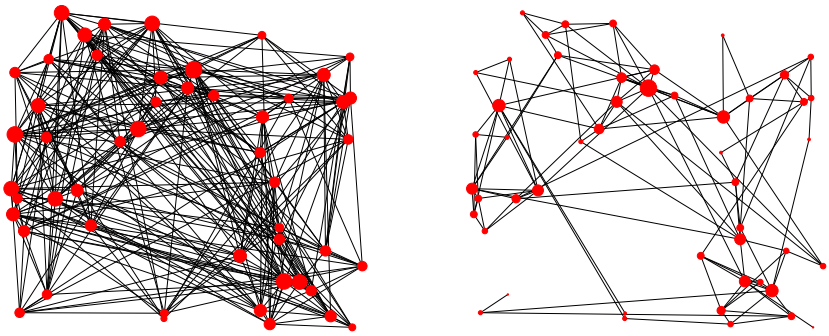
Existing connections in the contact network between individuals are assumed to be spatially dependent, with the probability of connections decaying continuously with increasing distance. This network model, which we will call for simplicity a spatial contact network, is the same used by Bifolchi et al. (2013). The underlying probability of an existing connection between individuals i and j is assumed to follow a power law and is given by

$$P(c_{ij} = 1) = 1 - \exp(-d_{ij}^{-\beta}), \quad \beta > 0, \tag{3.3}$$

where d_{ij} is the Euclidean distance between individuals i and j, and β is the spatial parameter. This spatial parameter β thus plays an important role in structuring the contact network. An increase in β leads to more sparse networks, while small values of β tend to produce intense networks with a large number of total connections between individuals.

For simplicity, individuals are assumed to be connected by undirected edges. This assumption results in a symmetric contact matrix such that $c_{ij} = c_{ji}$ for $i \neq j$; $i, j = 1, ..., N$. In this case, the number of contact network elements is given by $\binom{N}{2}$. Figure 3.1 shows two simulated contact networks, one relatively intense ($\beta = 0.9$) and the other relatively sparse ($\beta = 1.8$), for a population of 50 individuals.

Epidemics can then be generated through the contact network of individuals in the population from the network-based ILM given by Equation 3.2. Each infected individual has an infection life history that depends on the time of infection and the length of time spent infectious.



(a) Intense contact network, $\beta = 0.9$ (b) Sparse contact network, $\beta = 1.8$

Figure 3.1: Two examples of spatial contact networks. Red dots represent nodes with different sizes corresponding to their degree (number of edges connected to a node).

We assume that infection events follow a non-homogenous Poisson process, so that the time to the next infected individual, given that the last infection occurred at time t, is assumed to follow $W_j \sim \text{Exp}(\lambda_{jt})$, where W_j represents the "waiting time" for susceptible individual j becoming infectious. To generate an epidemic, we start with a randomly chosen initial infective individual k at time $t_0 = 0$.

Then, waiting times are assigned to each susceptible individual $W_j \sim \text{Exp}(\lambda_{jt})$. The individual with minimum W is chosen as the next infected individual, with $I_{t+1} = I_t + W$ assigned as their infection time. This newly-infected individual is assigned an infectious period γ_j generated from $f(\gamma_j; \delta)$. Thus, their removal time becomes $R_{t+1} = I_{t+1} + \gamma_j$. This process is repeated until no infectives remain in the population.

The duration of the epidemic is jointly affected by the model parameter values and the contact network. Strong epidemics can be caused by increasing the values of α_0 and α_1

along with decreasing the infectious period rate δ. Also note that, as $\delta \to 0$, the model tends to an $\mathcal{SI}$ model with no removal times. Also, an intense network can increase the speed of the epidemic as the possible number of infection pathways increases.

3.3 Bayesian Inference

The model is fitted within a Bayesian framework using Markov chain Monte Carlo (MCMC) to update the model parameters and the missing data (here, infection times and contact network) via a data augmentation technique (Gamerman and Lopes, 2006). The missing data are treated as latent variables and estimated along with the model parameters.

3.3.1 The Posterior Distribution

By assigning an independent gamma prior distribution to the model parameters α_0, α_1, ϵ, and δ, the joint posterior distribution for the model parameters and full data up to the constant of proportionality is given by

$$\pi(\boldsymbol{\theta}|\boldsymbol{R}, \boldsymbol{I}, \boldsymbol{C}, \boldsymbol{Z}) \propto L(\boldsymbol{R}, \boldsymbol{I}, \boldsymbol{Z}, \boldsymbol{C}|\boldsymbol{\theta})\ \pi(\alpha_0)\ \pi(\alpha_1)\ \pi(\epsilon)\ \pi(\delta)\ p(\boldsymbol{\phi}|\boldsymbol{C}), \tag{3.4}$$

where $p(\boldsymbol{\phi}|\boldsymbol{C})$ is an observation model that represents our global information about the contact network; $\boldsymbol{Z}$ is a vector of the binary covariate values, $z_i, i = 1, \ldots, N$; $\boldsymbol{\phi}$ is a vector of the degree of individuals (nodes) , $\phi_i, i = 1, \ldots, N$; and $\boldsymbol{C}$ is the contact network matrix with c_{ij} elements, $i, j = 1, \ldots, N$.

There are a number of ways in which global information about the contact network may be incorporated. One common method is to consider the degree distribution of the contact network (Danon et al., 2011; Keeling and Eames, 2005). Here, we consider the possibility that the true degree distribution of the contact network is known. That is, we assume that the distribution $p(\boldsymbol{\phi}|\boldsymbol{C}) = \prod_{i=1}^{N-1}\prod_{j\neq i}^{N}\pi(\phi_i)\pi(\phi_j)$, where $\phi_i = \sum_{j=1}^{N} c_{ij}$, is known.

A second scenario regarding the contact network observation model is also considered in which we assume we know the total number of connections within the contact network ϕ_{tot}, where $\phi_{tot} = \sum_{i=1}^{N}\sum_{j=1}^{i} c_{ij}$, with $c_{ii} = 0$. Thus, the observation model of Equation 3.4 becomes $p(\boldsymbol{\phi}|\boldsymbol{C}) = \pi(\phi_{tot})$. Here, we assume that $\pi(\phi_{tot})$ is the density of a normal distribution with mean equal to the true total number of connections, and the variance equal to 10. Finally, we consider scenarios in which no global information about the contact network is assumed, in which case $p(\boldsymbol{\phi}|\boldsymbol{C}) = 1$.

3.3.2 MCMC Method

Our approach uses a simple random walk Metropolis (within Gibbs) algorithm. None of the parameters have standard conditional distributions, except the infectious period rate δ that follows a gamma distribution [2]:

$$\delta|\alpha_0, \alpha_1, \epsilon, \boldsymbol{R}, \boldsymbol{I}, \boldsymbol{C}, \boldsymbol{Z} \;\sim\; \Gamma(m + a_\delta, M + b_\delta),$$

[2] Γ is used to denote the gamma distribution

where $M = \sum_{i=1}^{m} (R_i - I_i)$ is the sum of the m infectious periods and a_δ and b_δ are the parameters of its prior distribution.

A Gibbs update (i.e., sampling from the conditional posterior distribution) is used for estimating δ. Parameter estimates for α_0, α_1, and ϵ are updated in turn using normal proposal random-walk updates, tuned to achieve good mixing properties. The missing information (infection times (I_i) or, equivalently, infectious periods (γ_i), and connections of the contact network c_{ij}, in cases assuming it is unknown or partially unknown) are treated as latent variables and updated using the data augmentation method.

Updating I_i and γ_i

An independence sampler is used for updating the infection times and infectious periods. To do so, we update each infection time I_i by proposing infectious period γ_i^* from a proposal distribution (exponential distribution) with a tuning parameter Δ, such that $\gamma_i^* \sim$ Exp(Δ). Then, the new infection time is just the difference between the fixed known removal time and the new infectious period of the i^{th} individual. Each infection time/infectious period is updated in this way in turn.

Updating c_{ij}

In cases where the contact network is considered unknown or partially unknown, the unknown elements of the contact network (c_{ij}) are updated in turn. As c_{ij} has only two states, 0 (no connection) or 1 (connection), these parameters are updated by proposing a

move from the current state to the only other possible state with probability 1. We test the efficiency of updating c_{ij} with and without incorporating the observation model, as described in Section 3.3.1. First, we test the efficiency of MCMC in updating c_{ij} without the observation model. This leads to the conditional distribution of each c_{ij} to be equal to the likelihood function. Second, by incorporating the observation model, we incorporate more information about the contact network to the update of the connection c_{ij}. Thus, updating each unknown connection c_{ij} will be based on two parts: the likelihood function and the density of the observation model. By considering the degree distribution as the observation model, each unknown connection c_{ij} has a conditional distribution as follows:

$$c_{ij}|\alpha_0, \alpha_1, \epsilon, \delta, \boldsymbol{R}, \boldsymbol{I}, \boldsymbol{C}_{-ij}, \boldsymbol{Z} \;\; \sim \;\; L(\boldsymbol{R}, \boldsymbol{I}, \boldsymbol{Z}, \boldsymbol{C}|\boldsymbol{\theta})\; \pi(\phi_i; a_\phi, b_\phi)\; \pi(\phi_j; a_\phi, b_\phi),$$

and by considering the total number of connections as the observation model, the conditional distribution becomes

$$c_{ij}|\alpha_0, \alpha_1, \epsilon, \delta, \boldsymbol{R}, \boldsymbol{I}, \boldsymbol{C}_{-ij}, \boldsymbol{Z} \;\; \sim \;\; L(\boldsymbol{R}, \boldsymbol{I}, \boldsymbol{Z}, \boldsymbol{C}|\boldsymbol{\theta})\; \pi(\phi_{tot}).$$

where $\boldsymbol{C}_{-ij}$ is the contact network matrix with all the elements except the connection c_{ij} between the individuals i and j.

3.3.3 Model Validation

The model fit for the fitted ILMs is assessed using a posterior predictive approach (Gelman et al., 2004). This method involves the calculation of a test statistic $T(D)$ for data

D. The posterior predictive distribution of T is obtained by simulating a large number of replications of the data using parameters simulated from the posterior distribution, and calculating $T(D_i)$, for $i = 1, \ldots, m$. This distribution is then compared to the statistics under the observed data $T(D_{obs})$. In this study, three statistics are considered: $T_1(D) =$ the length of the epidemic, $T_2(D) =$ the peak time of the infection time curve, and $T_3(D) =$ the number of infected individuals.

Here, the posterior predictive distribution of the statistics is obtained by simulating 5,000 epidemics from the fitted model. This is done by randomly drawing 5,000 samples of the model parameters and, if applicable, c_{ij} estimates from the joint posterior distribution. Epidemics are then simulated for each sample followed by computation of the statistics. Note that we conditioned our simulations to the first five and ten infected individuals for both data sets of population sizes 25 and 50 individuals, respectively.

3.4 Simulation Study

Simulated data sets are considered and compared across two scenarios: one in which the contact networks are fairly intense (i.e., contain a large number of edges) in which $\beta = 0.9$, and one in which the contact networks are fairly sparse for which $\beta = 1.8$ (see Figure 3.1). We consider two population sizes, $N = 25$ and $N = 50$, where individuals are distributed uniformly at random across a square area of size 10×10 units. Ten contact networks were generated for each of the four contact network / population size scenarios.

Then, an epidemic was simulated across each of the 40 contact networks using the simple network-based ILM given by Equation 3.2. Epidemics were generated with parameter values chosen to produce reasonably informative epidemics (i.e., not "too short" (few infected individuals) and not "too long" (all the individuals were infected)). Specifically, epidemics of population size $N = 25$ were generated with model parameters $\alpha_0 = 0.8$, $\alpha_1 = 0.5$, and $\delta = 2.0$, and epidemics of population size $N = 50$ were generated with model parameters $\alpha_0 = 0.05$, $\alpha_1 = 0.3$, and $\delta = 0.5$. It was assumed here for all data sets that our population is isolated, hence we set $\epsilon = 0$.

Modelling the above data sets was carried out under three sets of assumptions about the contact network. The first, called the standard analysis, assumed that the contact network is fully and correctly observed and the only missing information is the infection times and infectious periods. The second assumed that the contact network was partially unknown with a randomly chosen 50% of the connections correctly recorded, and the other 50% considered unknown. The third model considered the network, along with the infection times and infectious periods, to be completely unobserved.

Because both network and infection times were being treated as unknown parameters, MCMC analyses tended to produce a very high rejection rate with poor MCMC chains. In particular, there is a high chance of proposed candidate MCMC chains moving with zero likelihood (e.g., if an infected individual, as a result of a proposed move, suddenly had no contact with anybody else in the population). Consequently, we included the spark term ϵ as a free parameter to be estimated to aid with MCMC mixing. Although the inclusion

of the unnecessary ϵ parameter overcame the above issues, this addition could of course potentially produce biased estimates of the other parameters.

For all data sets, independent gamma[3] prior distributions were assigned to all model parameters. For the analysis of epidemic data of population size $N = 25$ individuals, we assigned the following prior distributions: $\alpha_0 \sim \Gamma(2, 2)$, $\alpha_1 \sim$ Exp(2), $\epsilon \sim$ Exp(0.01) and $\delta \sim \Gamma(4, 2)$. For the analysis of epidemic data of population size $N = 50$ individuals, we assigned the following prior distributions: $\alpha_0 \sim$ Exp(1), $\alpha_1 \sim$ Exp(1), $\epsilon \sim$ Exp(0.01) and $\delta \sim \Gamma(2, 4)$. The DA-MCMC was run for 150,000 iterations with a burn-in period of 10,000 iterations.

3.5 Results

3.5.1 Network Uncertainty

Figures 3.2 and 3.3 show the posterior means and 95% percentile credible intervals for the ILM parameters under each of the three contact network observation scenarios (100%, 50%, and 0% observed, respectively), and under each type of contact network (intense and sparse), for populations of 25 and 50 individuals.

Broadly speaking, we can see that estimation of the main model parameters, α_0 and α_1, is very good since the 95% credible intervals tend to cover the true parameter values. We

[3]The exponential distribution is a special case of the gamma distribution with shape parameter $a = 1$. The gamma distribution used here, and throughout is with probability density function given by, $\pi(\theta) = \frac{b^a}{\Gamma(a)}\theta^{a-1}e^{-b\theta}$ for some parameter θ.

might expect to see the best performance under the fully observed contact networks, with a deterioration in performance as less of the contact network is observed. However, overall, this deterioration in performance is not apparent. For example, for the scenario consisting of intense networks within populations of 25 individuals, the posterior mean estimates of α_0 do become slightly biased as the proportion of the contact network unobserved increases; there is small but systematic underestimation of this parameter for the completely unobserved network that is not present when the network is fully observed. However, under the sparse networks and populations of 25 individuals, we see a slight, systematic overestimation of this parameter when the network is fully observed, with this bias disappearing under the completely unobserved networks. Further, in both cases the α_1 parameter is well estimated. For neither case under any scenario, we see an obvious, sizeable increase in posterior variance as the unobserved portion of the contact network increases – something we might expect to see.

Under all scenarios, there is a slight tendency for the posterior mean of the infectious period parameter δ to be overestimated. It also appears that this overestimation tends to increase slightly as the unobserved portion of the contact network increases, along with a tendency for an increase in posterior variance for this parameter. Similarly, the random sparks parameter ϵ (whose true value is actually zero) seems to increase as the unobserved portion of the contact network increases, along with a small, but discernible, increase in posterior variance.

Thus, it appears that we can estimate model parameters very well under unobserved

contact network scenarios, in the presence of global information regarding the network, which here is the total number of connections in the contact network. We discuss the effect of the observation model containing global network information in the next section.

Finally, Figures 3.4 and 3.5 show the posterior predictive means and 95% credible intervals of three statistics under the various population size/contact network-type scenarios: the length of the epidemic (i.e., the time of the final removal); the time of the peak of the epidemic; and the total number of infected individuals at the end of the epidemic. These posterior predictive distributions are compared with the true values of those epidemiological statistics under the various data sets when the contact network is fully observed, as well as a 100% unknown network with observation model based upon the total number of connections, were included in the analysis. We can see that these epidemic characteristics are generally well captured under the fitted model regardless of whether an observation model was included.

3.5.2 The Contact Network Observation Model Effect

Figures 3.6 and 3.7 show the posterior means and 95% credible intervals for the ILM parameters for the analyses carried out under the assumption of a completely unknown contact network for the population size 25 and 50 scenarios, respectively, for both intense and sparse contact networks. These results are for situations in which we have no observation

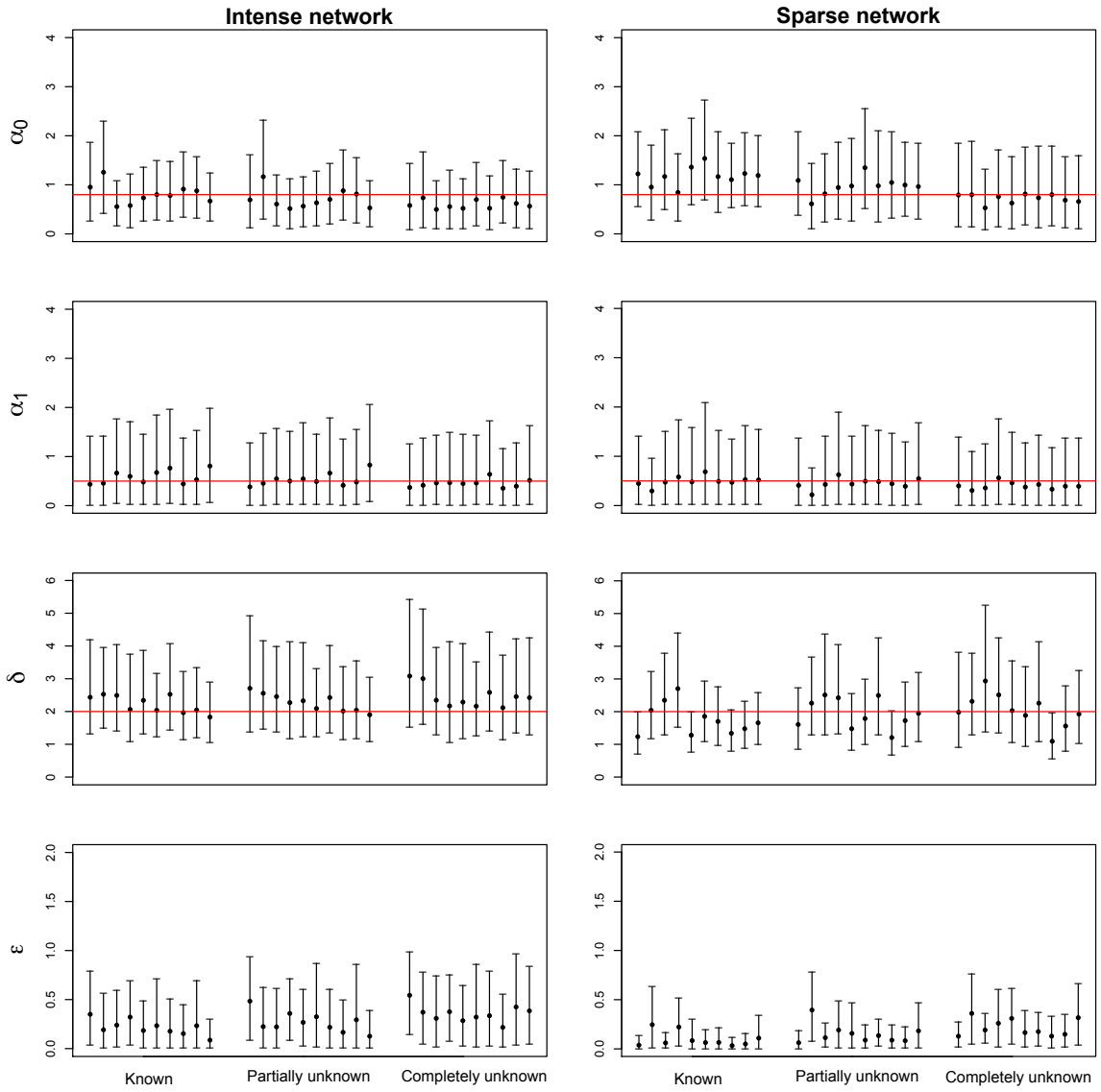


Figure 3.2: Posterior means for the simple network-based ILM with 95% credible intervals of the intense (left) and sparse (right) contact network epidemics (population size $N = 25$) based on the full Bayesian MCMC analyses of the three assumption scenarios of the contact network. The total number of connections observation model of Section 3.3.2 is applied in the unknown contact network scenarios. Horizontal red lines are the true values of the parameter; $\alpha_0 = 0.8$, $\alpha_1 = 0.5$, $\delta = 2$, and $\epsilon = 0$.

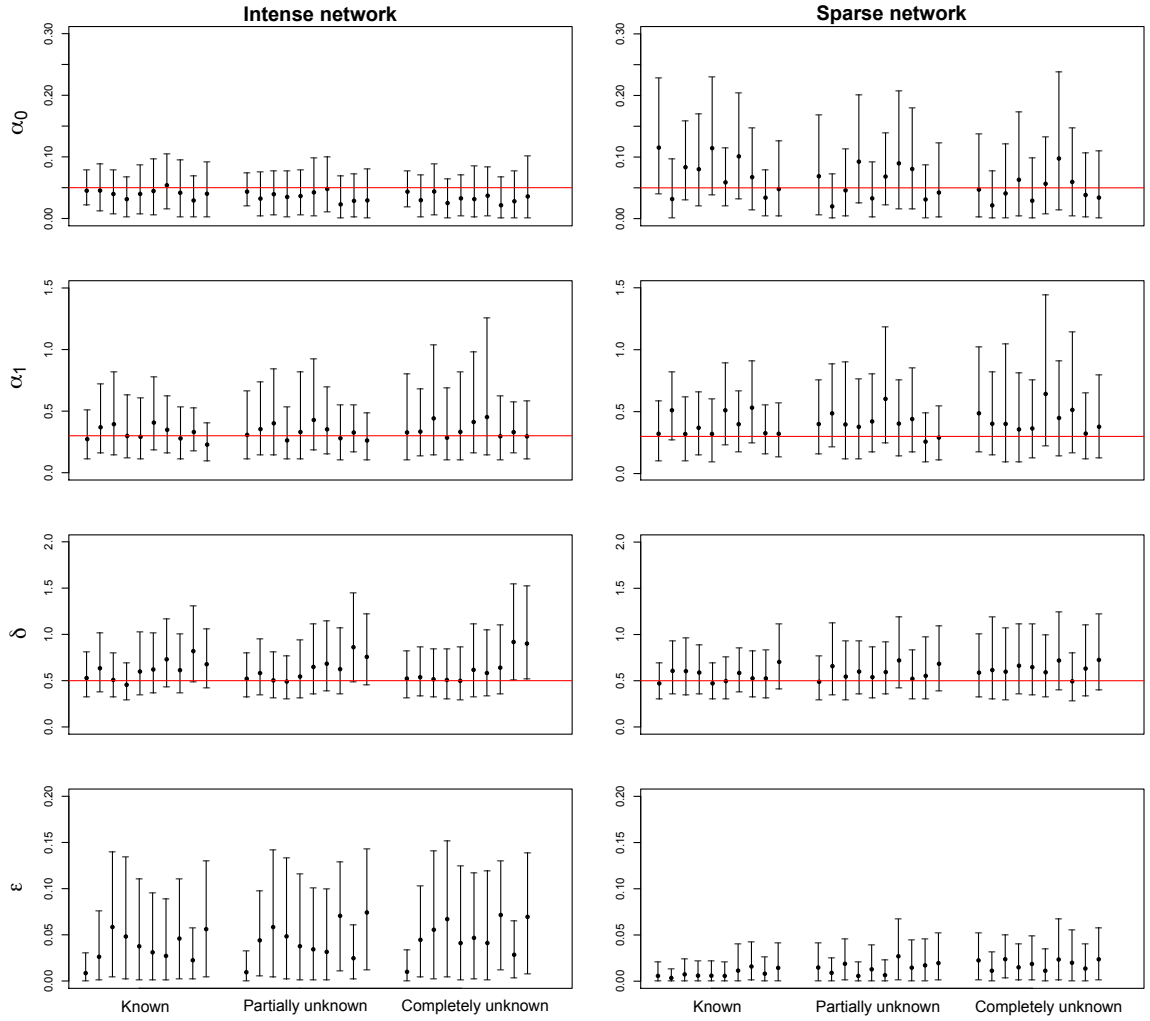


Figure 3.3: Posterior means for the simple network-based ILM with 95% credible intervals of the intense (left) and sparse (right) contact network epidemics (population size $N = 50$) based on the full Bayesian MCMC analyses of the three assumption scenarios of the contact network. The total number of connections observation model of Section 3.3.2 is applied in the unknown contact network scenarios. Horizontal red lines are the true values of the parameter; $\alpha_0 = 0.05$, $\alpha_1 = 0.3$, $\delta = 0.5$, and $\epsilon = 0$.

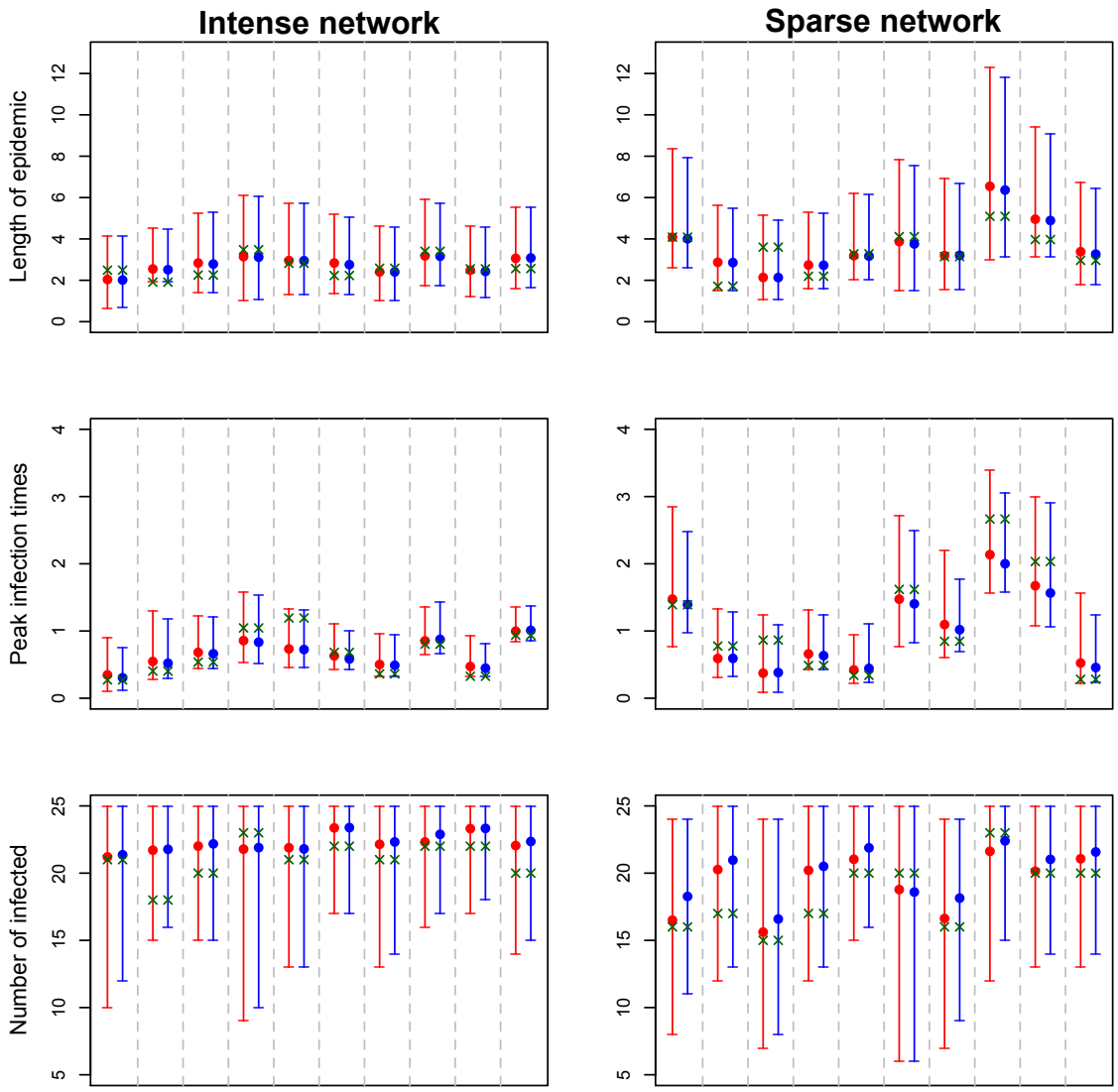


Figure 3.4: Summary of posterior predictive distributions obtained from the analyses of the data sets for a population of 25 individuals. The cross green points represent the true values of the statistics, and the red and blue lines represent the average posterior of the statistics with 95% credible intervals based on the known contact network analyses and the completely unknown contact network analyses including the total number of connections as the observation model, respectively.

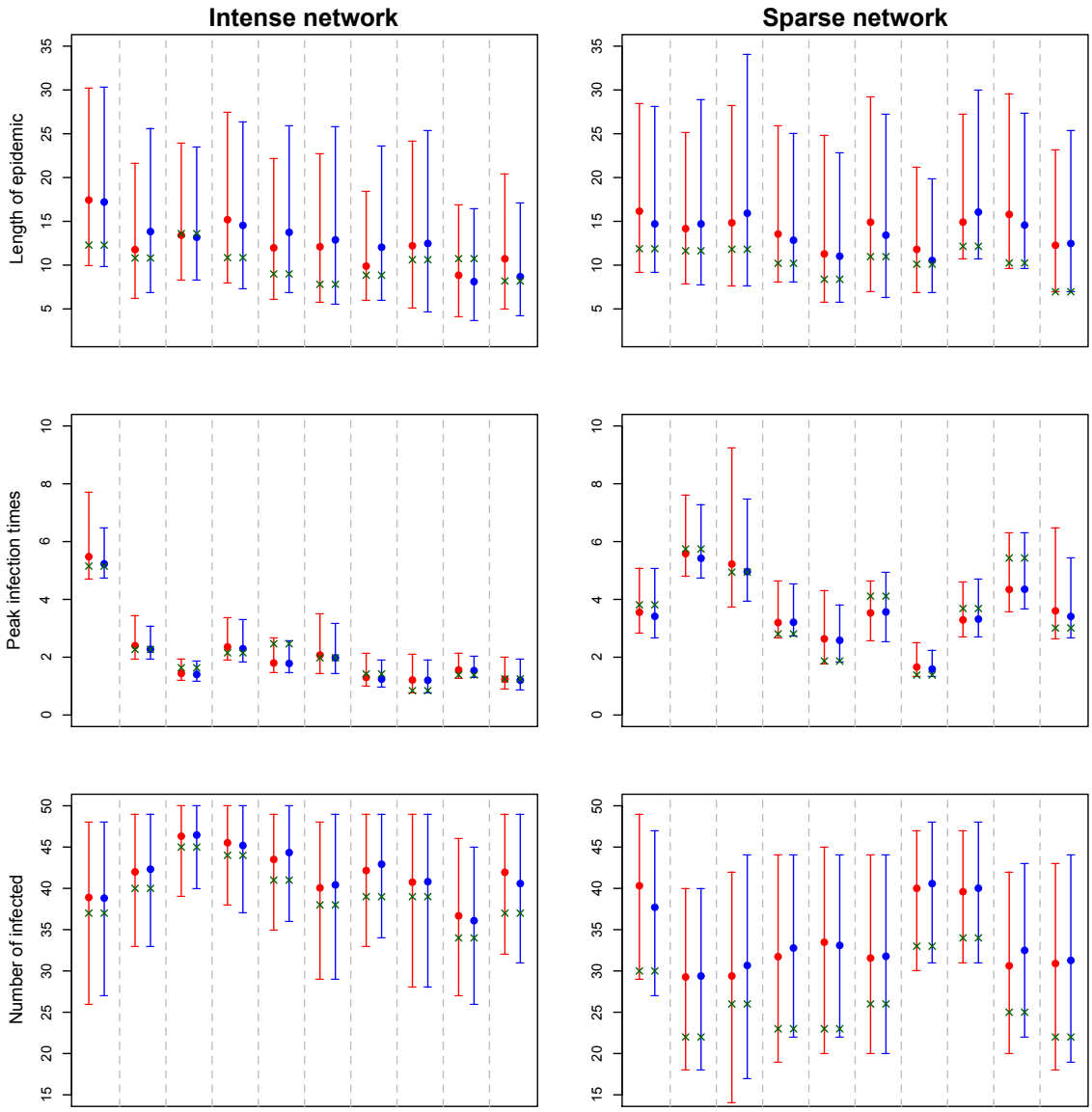


Figure 3.5: Summary of posterior predictive distributions obtained from the analyses of the data sets for a population of 50 individuals. The cross green points represent the true values of the statistics, and the red and blue lines represent the average posterior of the statistics with 95% credible intervals based on the known contact network analyses and the completely unknown contact network analyses including the total number of connections as the observation model, respectively.

model, an observation model based upon global knowledge of the degree distribution, and an observation model based upon global knowledge of the total number of connections in the network (see Section 3.3.1).

A number of patterns emerge across the scenarios tested. First, in each scenario the baseline susceptibility parameter α_0 is systematically underestimated when no contact network observation model is utilized. In fact, this underestimation is so extreme that in almost every case the 95% credible interval fails to capture the true value of the parameter. When an observation model based upon the degree distribution is used, the posterior means tend to underestimate the true value of the parameter but this underestimation is lower; with wider credible intervals, the true value is captured in the majority of cases. However, it is when an observation model based upon the total number of connections is incorporated that estimation of α_0 is at its best, although this tends to be at the expense of a larger posterior variance. A similar, but less extreme pattern is seen for the α_1 binary covariate parameter, with a tendency for underestimation when no observation model is used and increasing improvement under the degree distribution and total number of connections observation models, respectively. The inclusion, or not, of the observation model appears to have less effect on the infectious period rate parameter, δ, and sparks parameter ϵ. The infectious period parameter is reasonably well estimated across all scenarios, albeit with a tendency for overestimation throughout. The only exception is under the scenario of sparse contact networks with a population size of 25, with either contact network-based observation model incorporated, where estimation of δ seems to be very good. The posterior estimates of

the sparks term do not appear to vary much across scenarios and the different observation models, though the 95% credible intervals tend to be wider for intense contact networks.

Finally, Figures 3.8 and 3.9 show the posterior means and 95% credible intervals of the average number of connections per node under the scenarios with population sizes of 25 and 50, respectively. In line with other results here, the inclusion of the observation model based upon the total number of connections in the network produced the most satisfying estimates, with credible intervals capturing the true value and posterior means relatively close to the true value of the average degree of each node. When no observation model is included in the analysis, there is a tendency under all scenarios to overestimate the degree distribution of the contact network. Analyses with the observation model based upon the degree distribution fell somewhere in the middle, with a systematic overestimation of the average number of connections per node that was far less severe than that seen in the analyses with no observation model. Note that in these latter two cases, the credible intervals appear very narrow and never under any analysis managed to capture the true value of the average number of connections per node.

3.6 Discussion

We considered the problem of fitting individual level models to epidemic data in the presence of an underlying, but unobserved, spatial contact network. Although incorporat-

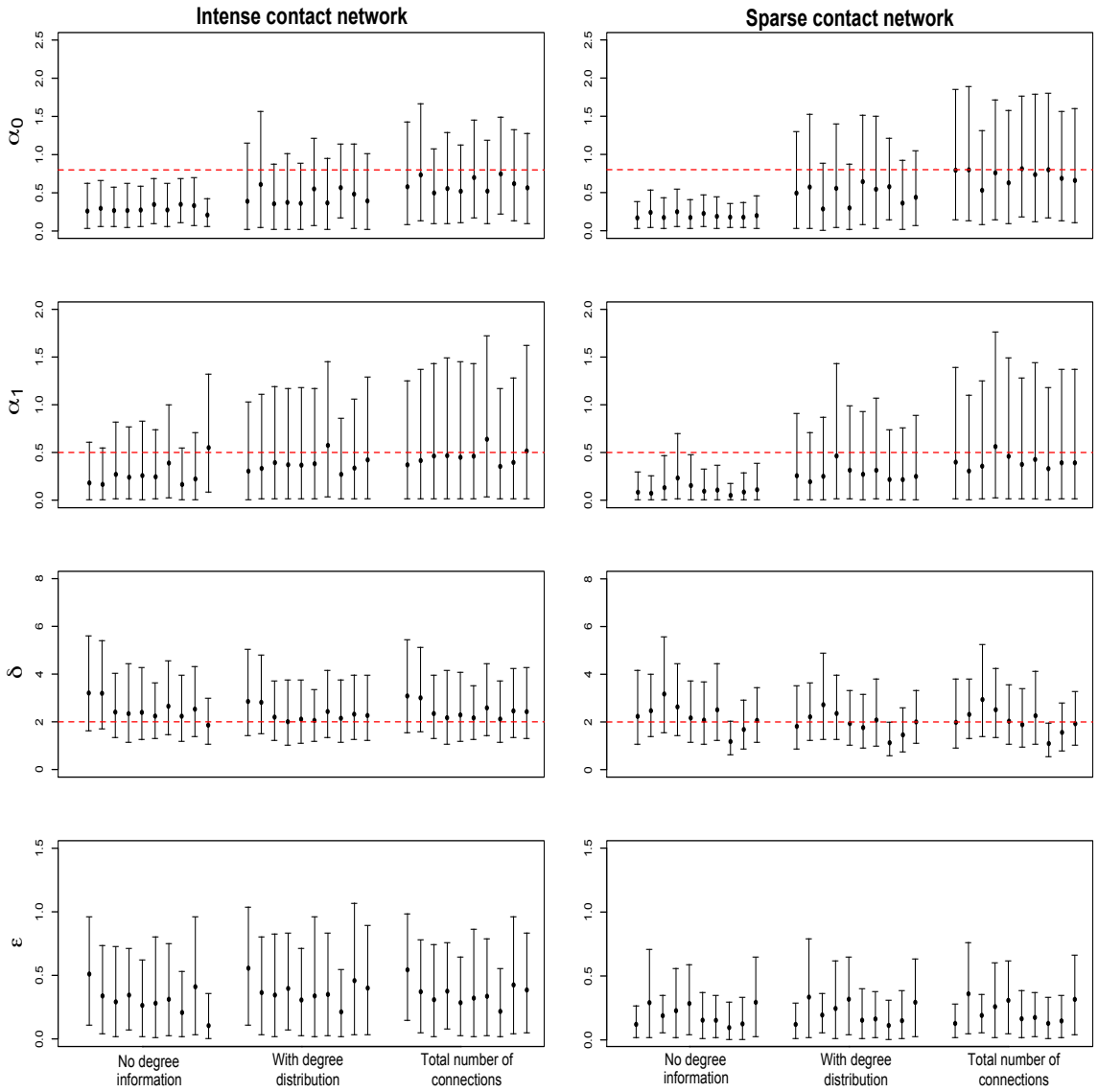


Figure 3.6: Posterior means for the simple network-based ILM with 95% credible intervals of the intense (left) and sparse (right) contact network epidemics (population size $N = 25$) based on the full Bayesian MCMC analyses of the completely unknown contact network. Each epidemic was analyzed three times: with no degree information, with the degree distribution, and with a distribution of the total number of connections in the contact network. Horizontal dotted lines are the true values of the parameter; $\alpha_0 = 0.8$, $\alpha_1 = 0.5$, $\delta = 2$, and $\epsilon = 0$.

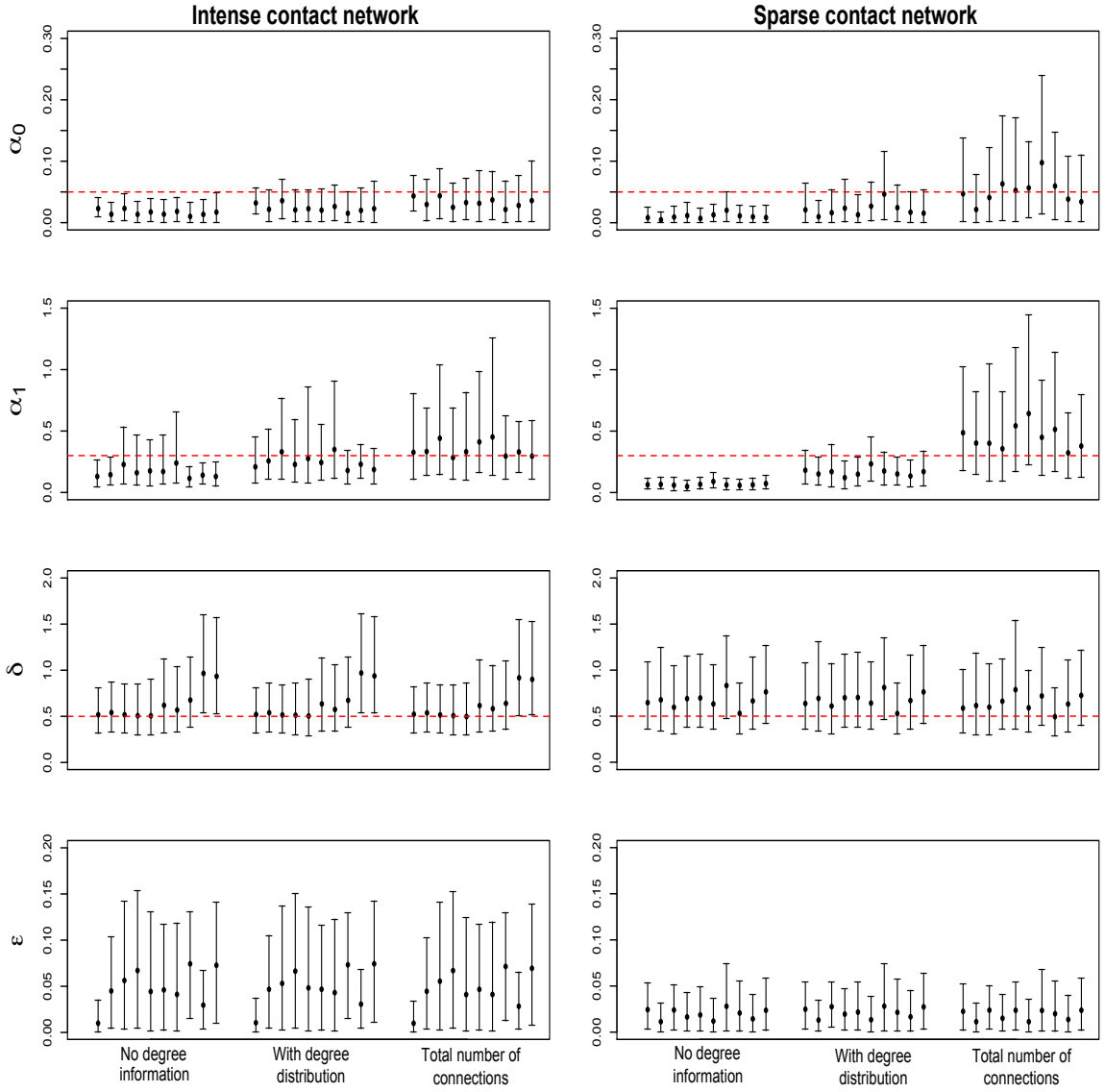


Figure 3.7: Posterior means for the simple network-based ILM with 95% credible intervals of the intense (left) and sparse (right) contact network epidemics (population size $N = 50$) based on the full Bayesian MCMC analyses of the complete unknown contact network. Each epidemic was analyzed three times: with no degree information, with the degree distribution, and with a distribution of the total number of connections in the contact network. Horizontal dotted lines are the true values of the parameters; $\alpha_0 = 0.05$, $\alpha_1 = 0.3$, $\delta = 0.5$, $\epsilon = 0$.

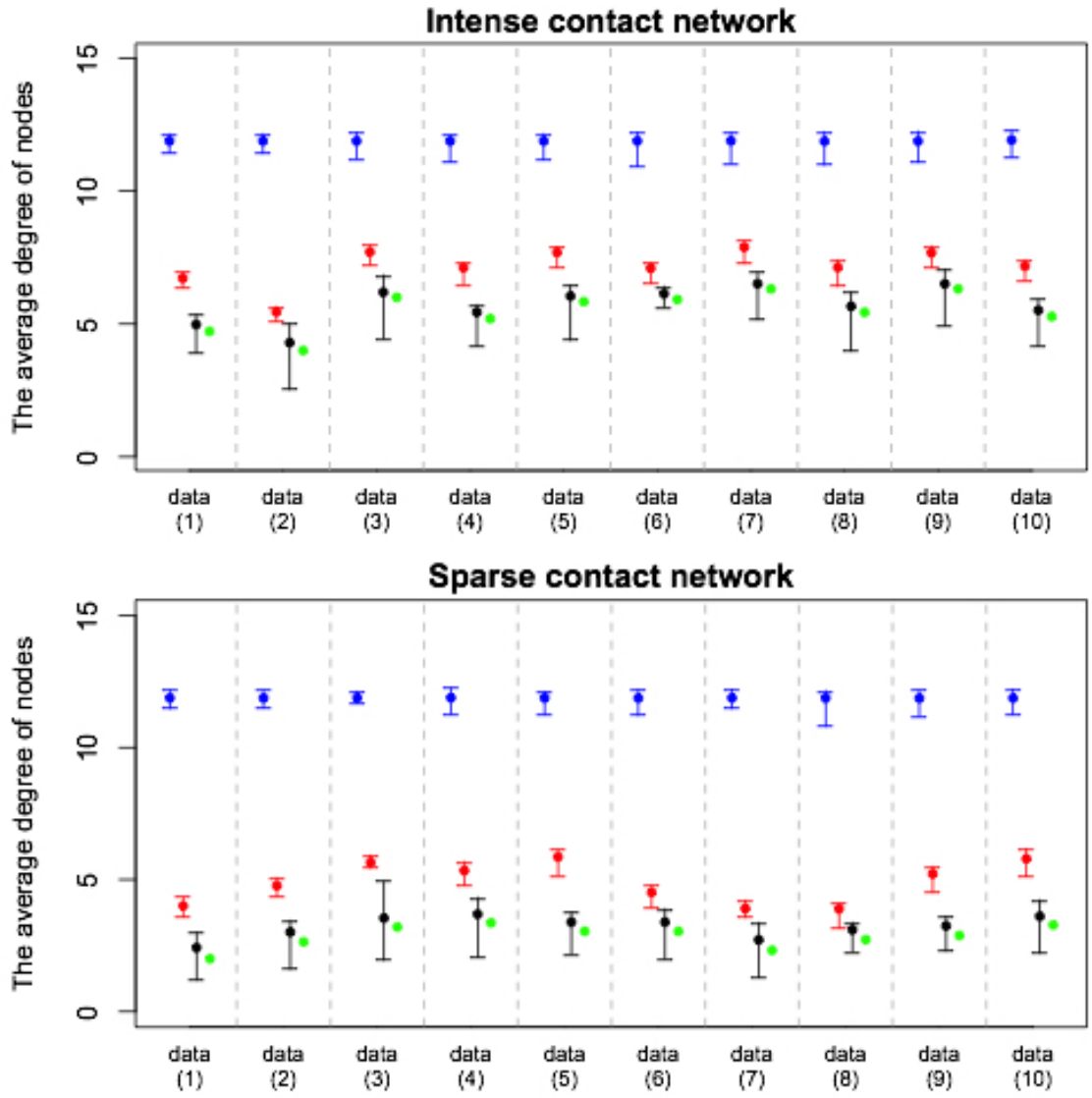


Figure 3.8: Posterior means for the degree of nodes with 95% credible intervals of the intense (upper) and sparse (lower) contact network epidemics (population size = 25) based on the full Bayesian MCMC analyses of the unknown contact network. The green circles represent the true average degree of nodes of the contact networks, the average degrees from the MCMC including the total number of connections as the observation model are represented by black, the average degrees from the MCMC including the degree distribution as the observation model are represented by red, and the average degrees from the MCMC without including the observation model are represented by blue.

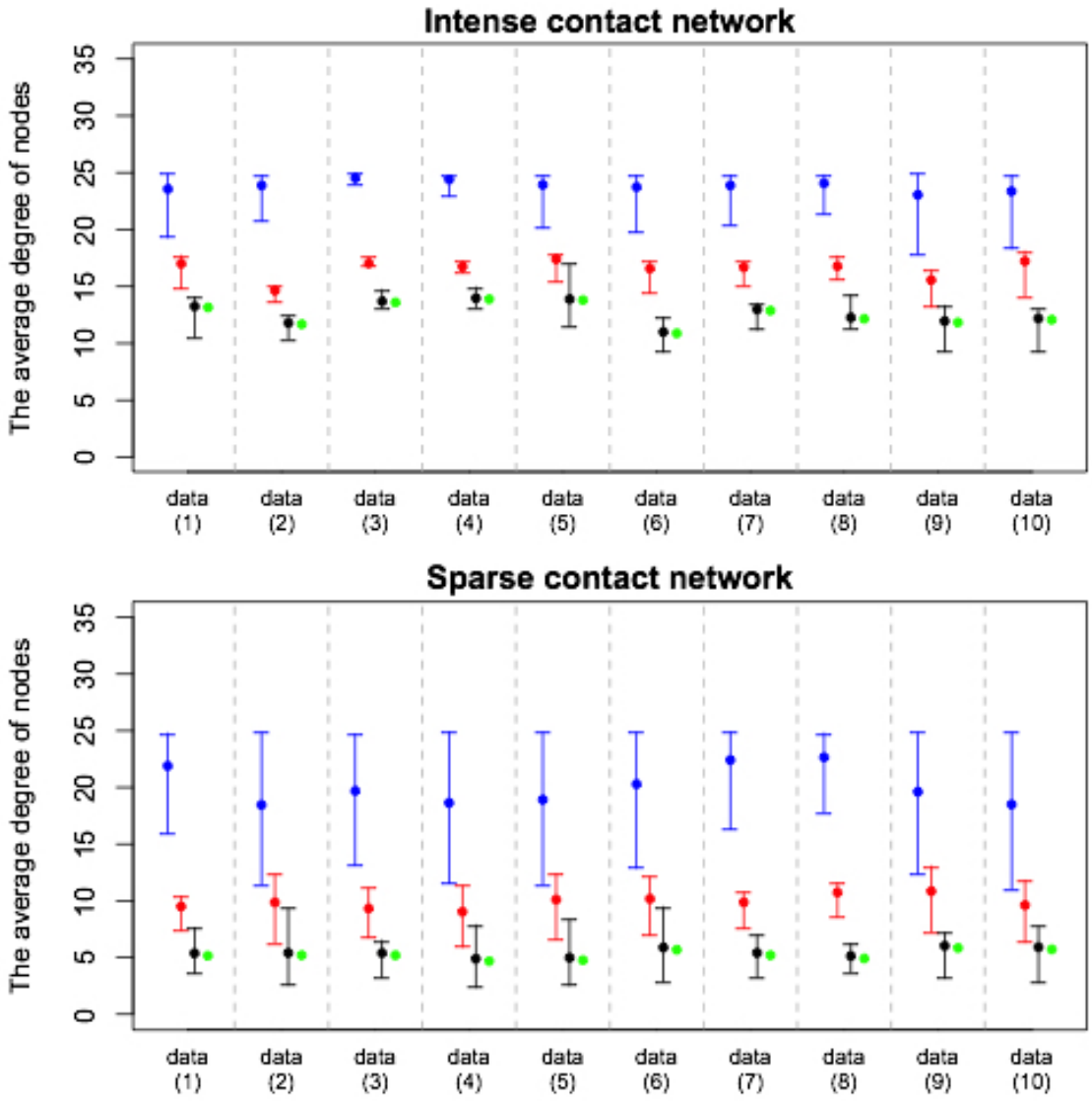


Figure 3.9: Posterior means for the degree of nodes with 95% credible intervals of the intense (upper) and sparse (lower) contact network epidemics (population size = 50) based on the full Bayesian MCMC analyses of the unknown contact network. The green circles represent the true average degree of nodes of the contact networks, the average degrees from the MCMC including the total number of connections as the observation model are represented by black, the average degrees from the MCMC including the degree distribution as the observation model are represented by red, and the average degrees from the MCMC without including the observation model are represented by blue.

ing contact network uncertainty into infectious disease modelling makes the analysis more complicated, especially when allowing for unknown infection times/infectious periods, we have shown it is possible to obtain good estimates of model parameters along with good approximations of the missing information. We also demonstrated the ability to produce good predictions of summary statistics that represent important features of the epidemic process under both completely known and unknown contact network analyses.

Incorporating global information about the contact network through the observation model led to a significant improvement in the estimation of the model parameters. Estimation of the degree distribution also improved. Specifically, we found that including an observation model based on the total number of connections in the contact network enhanced the estimates of both model parameters and the average number of degree per node better than including the exact degree distribution. Of course, highly informative observation models regarding the contact network were considered here. As a supplementary analysis, we analyzed a single data set of each contact network type and population size assuming far more diffuse information on the total number of contacts and degree distribution. Specifically, in the degree distribution observational model we assigned a degree distribution that followed an exponential distribution with rate 0.2 instead of the true degree distribution, and in the total number of connections observational model, we used a normal density with variances of $\sigma^2 = 50$ and 100 instead of $\sigma^2 = 10$. Under the less precise degree distribution observational model, we observed underestimation of the model parameters as well as the average degree nodes with tight 95% CIs that did not capture their

true values. These results tended to be similar to those estimates under no observational model analyses. Under the total number of connections observational model, increasing the normal variance to 50 and 100 showed a tendency to underestimate the model parameters compared to the informative analyses with variance equal to 10. However, with this underestimation, the 95% CIs of their estimates still captured the true values. We also observed an overestimation of the average degree nodes with tight 95% credible intervals. This preliminary exploration indicates the importance of certainty in the observation model for obtaining good estimates especially for the degree distribution observation model. This is an area that requires further investigation.

There are a number of avenues for further work. For simplicity, we have only considered a binary undirected spatial contact network to represent the underlying dynamics of the simulated epidemics. However, the underlying population heterogeneities might be better described by a directed contact network. For example, the contact network for health care workers and patients might be best explained by a directed contact network since patients are more likely to suffer from immune deficiencies. Conversely, the caregiver might be more readily exposed to bodily fluids during medical procedures (Meyers et al., 2006). We could also consider more complex epidemiological systems in which underlying contact networks might be unobserved. For example, we might wish to allow for network uncertainty in systems of multiple strains or pathogens, as considered by Romanescu and Deardon (2016) in the context of purely spatial models.

Although we show promising results for estimating model parameters when the under-

lying dynamics of the infectious disease are unknown, performing MCMC to analyze the fit of the network-based continuous time ILMs for a large epidemic in a large population is computationally expensive and is a noted limitation of the approach presented here. The average computation time for 150,000 MCMC iterations in the completely unknown contact network analysis of epidemics with population sizes 50 and 25 were 63.74 and 19.72 minutes, respectively. These were performed by coding in Fortran 95 and running on an Apple iMac with i5-core Intel 2.9 GHz[4] processors with 16 GB[5] of RAM[6]. Performing MCMC on an epidemic with a larger population size requires a tremendous amount of time because the number of unknown connections in the contact network becomes huge and performing the regular MCMC algorithm becomes inefficient. An alternative way to analyze a very large epidemic data set would be using approximation methods, such as approximate Bayesian computation methods (Toni et al., 2009; Beaumont et al., 2002), Gaussian process emulators (Pokharel and Deardon, 2016), classification methods (Pokharel and Deardon, 2014), and sampling-based likelihood approximations (Malik et al., 2016). Alternatively, reduction in the number of latent variables representing the contact network would likely have a positive reflection on the computation time of performing the data augmented MCMC. This reduction could be achieved by dividing the large population into clusters, with each cluster treated as a sub-population with its own contact network. Infections between individuals in different clusters could then be modelled via the spark term, a series of such terms for

[4]Abbreviation for gigahertz. One GHz represents 1 billion cycles per second.

[5]Gigabit (Gb), a unit of information used, for example, to quantify computer memory or storage capacity.

[6]Random-access memory (RAM) is a form of computer data storage that stores data and machine code currently being used.

different clusters; for example, they could be made a function of the degree distribution of each cluster. Given that each cluster's contact network would now be independent of other clusters' contact networks, the likelihood calculation could be easily parallelized to give a considerable reduction in the data augmented MCMC computation time. Of course, using this approach would likely introduce bias to the estimates of the model parameters, and therefore careful model design is required.

Chapter 4

Incorporating Contact Network Uncertainty in Individual Level Models of Infectious Disease using Approximate Bayesian Computation[1]

4.1 Introduction

The catastrophic threat of infectious disease outbreaks in human, animal, and plant populations has increased the demand for statistical modelling in order to understand the mechanisms behind, and dynamics of, disease transmission. For example, the 2001 UK foot-and-mouth disease (FMD) epidemic resulted in over four million animals destroyed, and brought severe economic consequences for the country as a whole (Keeling et al., 2001; Deardon et al., 2010). Gaining an understanding of the transmission dynamics quickly in an outbreak situation can be extremely important, as it can enable stakeholders to better make quick and reliable decisions regarding the control of the spread of the disease.

[1] Chapter 4 has been submitted to, and a revision requested from, *The International Journal of Biostatistics*.

Infectious disease dynamics are a function of complex interactions between infectious and susceptible individuals. Often, these are modelled via spatial mechanisms, as is the case for FMD (Deardon et al., 2010; Ster and Ferguson, 2007; Jewell et al., 2009). However, these interactions can often be a proxy to network information that may be unavailable (e.g., movement of animals, individuals and vehicles). In particular, the underlying process of disease transmission is often better explained by contact networks (which may or may not be spatial in nature).

A network is a graph of connections between individuals or groups in a population that define possible transmission paths of a disease. Such networks can be social, or based on trade or movement. Incorporating such networks in disease transmission modelling can enhance the prediction of the epidemic pattern and thus aid in controlling the spread of a disease (Danon et al., 2011; Keeling and Eames, 2005).

Incomplete data is generally a problem in disease modelling as the infection and recovery event times are rarely observed. This problem may be worse when using contact network-based ILM since collecting accurate and complete contact data can also prove difficult, time-consuming and/or expensive (Danon et al., 2011). Such missing data can be accounted for in a Bayesian data augmented framework using Markov chain Monte Carlo (DA-MCMC). However, fitting infectious disease models in such a framework can be computationally intensive, especially if there is a high level of missing information. The primary source of the computational problem is the numerous occurrences of evaluating the likelihood function, which can lead to a serious computational bottleneck.

Deardon et al. (2010) introduced a class of discrete-time individual-level models (ILM) that can be used to make inferences about the spread of a disease at the individual level in a population. The key feature of these models is the flexibility of incorporating various risk factors at the individual level. Both spatial distance between susceptible and infectious individuals and contact network information can easily be incorporated into these ILMs as risk factors. While ILMs are flexible and intuitive, inference can be computationally problematic, especially for large data sets. Even in moderately sized data sets, MCMC algorithms used to fit the model to data must typically be run for a long time in order to reach convergence and, thus, give meaningful results. Several approaches have been taken to speed up the calculation of the likelihood when fitting spatial ILMs; e.g., kernel linearization (Deardon et al., 2010), data sampling-based likelihood approximation (Malik et al., 2016), and Gaussian process emulation (Pokharel and Deardon, 2016).

An alternative approach is to use Approximate Bayesian Computation (ABC) methodology, a suite of simulation-based techniques devised for fitting models that have analytically or computationally intractable likelihood functions (Pritchard et al., 1999; Beaumont et al., 2002; Marjoram et al., 2003; Sisson et al., 2007; Beaumont et al., 2009). Under such an approach, likelihood evaluation is avoided completely, being approximated by comparing observed data to data simulated from the model.

Several studies have used ABC methods to analyze epidemic data. For example, McKinley et al. (2009) applied various ABC approaches for fitting a stochastic model to Ebola Haemorrhagic Fever data in the Democratic Republic of Congo. Numminen et al. (2013)

utilized a population Monte Carlo ABC approach to study the transmission dynamics of Streptococcus Pneumoniae from strain prevalence data. Finally, Walker et al. (2010) used so-called partial rejection control ABC for fitting small world network models to study Severe Acute Respiratory Syndrome (SARS) spread in Hong Kong.

In this Chapter, we consider the scenario of fitting continuous time network-based ILMs to data when an underlying undirected, binary, spatially-informed contact network is unknown (except in the form of some prior knowledge about global characteristics of the network) via population Monte Carlo ABC methods (ABC-PMC). We compare the performance of ABC-PMC with a full Bayesian MCMC analysis when fitting relatively simple ILMs to simulated data sets with small population sizes. We also apply the use of ABC-PMC methods when analyzing a larger, real data set; here, data from the 2001 FMD epidemic in the UK.

The remainder of the Chapter is laid out as follows. In Section 2 we introduce the continuous time individual level models and contact networks considered, along with a description of how epidemic simulation is carried out. In Section 3, both ABC-PMC and MCMC parameter estimation methods considered are described. Section 4 details the epidemic simulation studies carried out, along with a description of the data from the UK 2001 FMD epidemic used to assess the performance of the ABC-PMC methods. The results are described in Section 5, with a final discussion presented in Section 6.

4.2 Methodology

4.2.1 The General Continuous Time ILMs

The general framework of disease transmission model introduced here is a modification of the framework of the individual-level models (ILMs) of Deardon et al. (2010). The key modification here is that we now set the models in continuous, rather than discrete, time.

The model framework is defined as follows. First, let λ_{jt} denote the infectivity rate of previously uninfected (susceptible) individual j at time t:

$$\lambda_{jt} = \left[\Omega_S(j) \sum_{i \in \mathcal{I}(t)} \Omega_T(i)\kappa(i,j)\right] + \epsilon(j,t), \tag{4.1}$$

where $\mathcal{I}(t)$ is the set of infectious individuals at time t; $\Omega_S(j)$ and $\Omega_T(i)$ are functions of potential risk factors associated with susceptible individual j contracting, and infectious individual i transmitting, the disease, respectively; $\kappa(i,j)$ is an infection kernel, a function of risk factors shared between pairs of infected and susceptible individuals; and "random appearing" infections may be introduced by the spark term $\epsilon(j,t)$ (e.g., infection of a susceptible individual by an infectious individual from outside the observed population). As is often the case, we shall assume $\epsilon(j,t) = \epsilon$ is a fixed constant.

The infection kernel $\kappa(i,j)$ is a key component of the model that allows for spatial or contact network-based mechanisms. In spatial models, $\kappa(i,j)$ is typically power-law function of Euclidean distance. However, for many disease systems, an infection kernel which is a function of a contact network, series of such networks, or indeed, a function of

both of distance and network(s), is more appropriate.

Here, ILMs are further placed within the context of a susceptible- infectious- removed ($\mathcal{SIR}$) compartmental framework, although extensions to other frameworks, such as susceptible-exposed- infectious- removed ($\mathcal{SEIR}$), could be straightforwardly made. In an $\mathcal{SIR}$ framework, each individual i in the population is assumed to be in one of the three states at each time point: $\mathcal{S}$ (susceptible), $\mathcal{I}$ (infectious) or $\mathcal{R}$ (removed). Individuals are assumed to be in the susceptible state when they can potentially contract the disease, but are not yet infected. Individuals become infectious as soon as they are infected with the disease and are then able to infect other susceptible individuals during their infectious period. Following their infectious period, an individual transitions to the removed state. Individuals within the removed state no longer have a role in spreading the disease to other individuals, and can no longer be infected. The removed state may represent individuals who have died from the disease, recovered from the disease with an acquired immunity, or have been quarantined in some way.

Likelihood Function

We now present the likelihood function for this continuous time class of ILMs. First, we denote the number of individuals that become infected during the course of the epidemic as m. We consider data sets for complete epidemics only here; that is, the last individual recorded in our data set has been through the infectious period and entered the removed state. This assumption is also very easy to relax, but fits the scenarios we consider here.

We label the m infected individuals $i = 1, 2, \ldots, m$ corresponding to their infection (I_i) and removal (R_i) times; and the $N - m$ individuals who remain uninfected, we label $i = m + 1, m + 2, \ldots, N$ with $I_i = R_i = \infty$. We then denote infection and removal time sets for the population as $\boldsymbol{I} = \{I_1, \ldots, I_m\}$ and $\boldsymbol{R} = \{R_1, \ldots, R_m\}$, respectively. Then the infectious period of an infected individual i is defined as $\gamma_i = R_i - I_i$. The infectious period is generally assumed to follow some non-negative distribution, typically an exponential or gamma distribution (Jewell et al., 2009).

The likelihood can then be constructed as the product of two parts detailing infection and removal, respectively. Note that, each susceptible individual j that becomes infected would have specific infectious pressure from an infected individual i prior to being infected if and only if $I_i < I_j < R_i$. The likelihood is given as follows:

$$
\begin{aligned}
L(\boldsymbol{I}, \boldsymbol{R}|\boldsymbol{\theta}) &= \prod_{j \neq k} \left(\epsilon + \sum_{i: I_i < I_j \leq R_i} \Omega_S(j) \Omega_T(i) \kappa(i, j) \right) \\
&\times \exp \left\{ - \int_{I_k}^{T_{obs}} \left(\sum_{i \in \mathcal{S}(t)} \epsilon + \sum_{i \in \mathcal{I}(t)} \sum_{j \in \mathcal{S}(t)} \Omega_S(j) \Omega_T(i) \kappa(i, j)(u - I_i) \right) du \right\} \prod_{i=1}^{m} f(\gamma_i; \delta) \\
&= \prod_{j \neq k} \left(\epsilon + \sum_{i: I_i < I_j \leq R_i} \Omega_S(j) \Omega_T(i) \kappa(i, j) \right) \\
&\times \exp \left\{ - \sum_{i=1}^{m} \left(\sum_{j=1}^{N} ((R_i \wedge I_j) - (I_i \wedge I_j)) \Omega_S(j) \Omega_T(i) \kappa(i, j) \right) \right\} \\
&\times \exp \left(- \epsilon \sum_{i=1}^{N} [(T_{obs} \wedge I_i) - I_k] \right) \prod_{i=1}^{m} f(\gamma_i; \delta),
\end{aligned}
$$

where the wedge symbol $\wedge$ denotes the minimum operator; $\boldsymbol{\theta}$ is the vector of unknown parameters; $f(\bullet;\delta)$ indicates the density of the infectious periods; T_{obs} indicates the time of the last infected individual being removed; and k is the label of the initial infective individual with I_k being its corresponding infection time. The integral term which represents the total person-to-person infectious pressure through the epidemic, can be simplified by the product of the two exponent terms in the lower equation (Jewell et al., 2009; Neal and Roberts, 2005; Britton and O'Neill, 2002).

The Network-Based ILM

All the ILMs considered here are network-based, with epidemics spreading between individuals through a single, binary, undirected time-homogeneous contact network. Specifically, we assume $\kappa(i,j) = c_{ij}$, where

$$c_{ij} = \begin{cases} 1, & \text{if a connection exist between individuals } i \text{ and } j,\ i \neq j. \\ 0, & \text{otherwise.} \end{cases}$$

We now introduce two specific forms of network-based ILMs considered in this Chapter.

Simple Network-Based ILM

We consider a simple network-based ILM for use in a simulated data setting as proof of concept. This simple model contains a single binary susceptibility covariate with no transmissibility covariates (i.e., $\Omega_T(i) = 1$) or sparks infections (i.e., $\epsilon = 0$). Specifically,

the infectivity rate is given by

$$\lambda_{jt} = \left[(\alpha_0 + \alpha_1 z_j) \sum_{i \in \mathcal{I}(t)} c_{ij} \right] + \epsilon, \tag{4.2}$$

where α_0 is a baseline infectivity parameter; z_j is a binary covariate that could represent, for example, the vaccination status of individuals; and α_1 is the covariate effect. Since we assume $\epsilon = 0$, the infectivity rate for an individual with no connections to infectious individuals at time t will be zero.

FMD Network-Based ILM

Our second network-based ILM is designed for use with data from the outbreak of FMD in the UK in 2001. It is a heavily simplified version of that introduced by Deardon et al. (2010) for the same purpose. We consider the individuals in the population to be farms, rather than animals themselves, and the binary contact network is assumed to represent connections between farms through which infection could potentially occur. Such mechanisms could include animal movements, human or vehicular movement (since the disease can be carried by vectors), contact between farm animals on contiguous farms, pathogen carried in the air and/or wildlife vectors carrying the disease between farms.

The spread of FMD was mainly mostly confined to sheep and cattle farms (Keeling et al., 2001; Donaldson and Alexandersen, 2001). Therefore, we consider only sheep and/or cattle farms, and include the number of sheep and cattle on the farm as susceptibility and transmissibility covariates in our model. Specifically, we assume the rate of infectivity of

farm j at time t to be

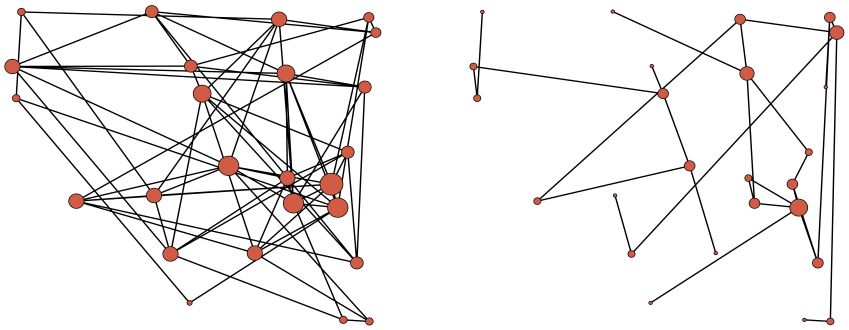
$$\lambda_{jt} = \left[(\alpha_s n_j^s + \alpha_c n_j^c) \sum_{i \in \mathcal{I}(t)} (\phi_s n_i^s + \phi_c n_i^c) c_{ij} \right] + \epsilon, \tag{4.3}$$

where for sheep and cattle, respectively, n_j^s and n_j^c represent the number of animals at farm j, α_s and α_c are the susceptibility parameters, and ϕ_s and ϕ_c are the transmissibility parameters. We arbitrarily fix α_s in order to avoid identifiability issues (Deardon et al., 2010; Malik et al., 2016) so that other susceptibility and transmissibility parameters are interpreted with respect to this baseline value. Here, the constant spark term, (i) explains infections not well explained by the contact network and/or covariates, and (ii) allows for infection coming from outside the observed population.

4.2.2 Contact Networks

Networks considered here are undirected with $c_{ij} = c_{ji}$ for $i \neq j; i, j = 1, ..., N$, so that each contact network is defined by $\binom{N}{2}$ elements. We confine ourselves to considering only contact networks that are spatial in nature, with connections more likely to be present between two individuals close together than far apart. Specifically, we use a generalized version of the power-law contact networks of Bifolchi et al. (2013), in which the probability of a connection between individual i and j is given by:

$$P(c_{ij} = 1) = 1 - \exp(-\nu d_{ij}^{-\beta}), \quad \beta, \nu > 0, \tag{4.4}$$



(a) Intense contact network, $\nu = 1, \beta = 0.9$

(b) Sparse contact network, $\nu = 1, \beta = 1.8$

Figure 4.1: Two samples of spatial contact networks. Red dots represent nodes with different sizes that are corresponding to their degree (number of edges connected to a node).

where d_{ij} is the Euclidean distance between individual i and j; β is the spatial parameter; and ν is the scale parameter. In many cases we consider here we set $\nu = 1$ so that the spatial parameter β plays the key role in structuring the contact network. The increase of β (for a given value of ν) leads to more sparse networks, while small values of β tend to produce intense networks with a large number of connections between individuals. Figure 4.1 shows two simulated contact networks, one relatively intense, and the other relatively sparse, on a population size of 25 individuals randomly distributed in space.

4.2.3 Epidemic Simulation

We consider simulation of epidemics from the models described above in the following way. Each infected individual has an infection life history defined by their time of infection and the length of time spent in the infectious state. We are assuming that infection events

follow a non-homogenous Poisson process, so that the time to the next infected individual, given that the last infection occurred at time t, is assumed to follow $W_i \sim \text{Exp}(\lambda_{it})$, where W_i represents the "waiting time" for susceptible individual i becoming infected.

Then, with a given infectious period distribution, an epidemic is simulated across a given contact network, starting with initial infected individual k at time I_k. This is done in the following way. First, we compute the rate of infection λ_{it} for each susceptible individual i at time t. Then we generate $W_i \sim \text{Exp}(\lambda_{it})$. Next we choose the individual with minimum W as the next infected individual and assign $I_{t+1} = I_t + W$ as the infection time for this individual. The infectious period γ_i for this newly-infected individual is generated from $f(\gamma_i; \delta)$ and $R_{t+1} = I_{t+1} + \gamma_i$ is assigned as the removal time for this individual. These steps are then repeated until no infectives remain in the population.

4.3 Model Fitting

In most infectious disease data sets, salient information we would wish to observe is missing; for example, infection times and infectious periods are usually unknown. It is typical to consider these missing data as latent variables to be inferred. If, additionally, the contact network is unknown, the number of these latent variables becomes even larger, increasing as population size increases.

The estimation of the model parameters, along with the latent variables, is typically carried out within a Bayesian framework using data-augmented MCMC to generate samples

from the joint posterior of the model and latent parameters. However, this approach can prove computationally intractable if the population size is large, and, thus, the number of latent variables is large.

In this Chapter, infection times and the underlying contact network are assumed unknown. However, unlike many situations, we assume the removal times are known. This follows our UK 2001 FMD epidemic motivating case study in which removal times were recorded as the date on which animals on infected farms were culled.

4.3.1 Approximate Bayesian Computation

In an attempt to avoid the computational intractability often associated with an MCMC approach in high dimensional problems, we consider the use of (increasingly popular) approximate Bayesian computation (ABC) methods for fitting network-based ILMs to data. In such methods, the likelihood function is approximated by simulating data sets from the model and comparing them to the observed data in some way.

Early variants of ABC were based upon rejection sampling (Pritchard et al., 1999). However, more efficient sequential ABC methods have since been developed, examples of which are Partial Rejection Control (Sisson et al., 2007), Sequential Monte Carlo (Del Moral et al., 2012) and Population Monte Carlo (PMC) (Beaumont et al., 2009). For the purpose of this study, we will focus on the ABC-PMC method.

Algorithm 4.3 ABC-PMC

Initializing: Set ABC-generation indicator to t = 1.
Set initial tolerance $\xi^{(1)}$ and sample size N.
for $j = 1$ **to** N **do**
    repeat
    Sample $\theta^* \sim q_t(\theta)$, where $q_t(\theta)$ is given in (4.5) .
    Generate a data set $D^{'} \sim MODEL(\bullet\,|\theta^*)$.
    Determine discrepancy $\rho(\mathbb{S}(D), \mathbb{S}(D^{'}))$ between the observed data D and the simulated data $D^{'}$.
    until $\rho(\mathbb{S}(D), \mathbb{S}(D^{'})) \leq \xi^{(t)}$.
    Set $\theta_j^{(t)} = \theta^*$, $\rho_j^{(t)} = \rho(\mathbb{S}(D), \mathbb{S}(D^{'}))$ and weight as $w_j^{(t)} = \frac{1}{N}$.
Set ABC-generation indicator to $t = t + 1$.
Initialize the tolerance $\xi^{(t)}$ and $K(\theta, \bullet)$ based on the $\{\theta_k^{t-1}, \rho^{(t-1)}, w_k^{(t-1)} | k = 1, ..., N\}$.
for $j = 1$ **to** N **do**
    repeat
    Sample $\theta^* \sim q_t(\theta)$.
    Sample $\theta^{**} \sim K(\theta^*, \bullet)$.
    Generate a data set $D^{'} \sim MODEL(\bullet\,|\theta^{**})$.
    Determine discrepancy $\rho(\mathbb{S}(D), \mathbb{S}(D^{'}))$ between the observed data D and the simulated data $D^{'}$.
    until $\rho(\mathbb{S}(D), \mathbb{S}(D^{'})) \leq \xi^{(t)}$.
    Set $\theta_j^{(t)} = \theta^*$ and $\rho_j^{(t)} = \rho(\mathbb{S}(D), \mathbb{S}(D^{'}))$.
for $j = 1$ **to** N **do**
    Set $w_j^{(t)} = \frac{\pi(\theta_j^{(t)})}{q_t(\theta_j^{(t)})}$.
Determine visually whether to continue for a further ABC generation. If the decision is to continue go to step 10.

ABC-PMC Algorithm

Beaumont et al. (2009) introduced an adaptive ABC-PMC algorithm that was inspired by the Population Monte Carlo algorithm of Cappé et al. (2004). The ABC-PMC algorithm is shown in Algorithm 4.3.

The ABC-PMC algorithm is based on sampling θ, called particles, from a sequence of approximate posteriors (importance densities) $q_t(\theta)$ with decreasing tolerance levels ξ. At each generation of the algorithm, new particles are sampled from the previous generation by sampling from the importance density given by:

$$q_t(\theta) = \begin{cases} \pi(\theta) & \text{if } t = 1, \\ \sum_{i=1}^{N} w_i^{(t-1)} K(\theta_i^{(t-1)}, \theta) / \sum_{j=1}^{N} w_j^{(t-1)} & \text{otherwise.} \end{cases} \tag{4.5}$$

Sampling from $q_t(\theta)$ is done by sampling from the prior distribution $\pi(\theta)$ in the first generation, and from the particle set $\theta_i^{(t-1)}, i = 1, \ldots, N$ with probability $w_i^{(t-1)}$ for the subsequent generations. The sampled particles are then perturbed by a proposal distribution $K(\bullet \mid \bullet)$, that is adapted at each generation t based on the accepted particles of generation $t-1$. Typically a multivariate normal distribution is used as the proposal kernel with an adaptive scale covariance Σ set to be twice the weighted covariance of $\theta_j^{(t-1)}, j = 1, ..., N$ (Beaumont et al., 2009; Filippi et al., 2013). The sampled particles are then accepted or rejected as in Algorithm 4.3, and weights are assigned to the accepted particles. The ABC-PMC algorithm is stopped at generation t when there is no significant difference between the distribution of particles between generations t and $(t-1)$.

Sufficient summary statistics are usually unavailable for models with intractable likelihoods, and summary statistics ($\mathbb{S}$ in Algorithm 4.3) are generally chosen in a way that balances the need for high levels of information with low dimensionality, in order to gain as good an approximation to the likelihood function as possible (Kypraios et al., 2017). However, the difficulty of defining those summary statistics can be increased when important information regarding the epidemic process is missing, since this missing information must be modelled too (e.g., in epidemic models a delay between the unobserved infection time and the observed reporting time), and we require summary statistics that capture all aspects of the model. Of course, this difficulty can be alleviated if we have strong prior information about some aspects of the model that we can incorporate into our analysis (e.g., through the prior distribution).

Considering a G dimensional vector of summary statistics $\mathbb{S} = \{s_g\}_{g=1}^{G}$, the discrepancy between summary statistics of the simulated $\mathbb{S}^{(sim)}$ and observed $\mathbb{S}^{(obs)}$ data are usually assessed by the L^1-norm or L^2-norm (Euclidean distance). Here, we used the Euclidean distance for the L^2-norm, given by:

$$\rho(\mathbb{S}^{(sim)}, \mathbb{S}^{(obs)}) = \left[\sum_{g=1}^{G} (s_g^{(sim)} - s_g^{(obs)})^2\right]^{1/2}$$

The sequence of tolerances $\xi^{(t)}$ is adaptively determined in the algorithm. The algorithm starts with initial tolerance level $\xi^{(1)} = \infty$ at $t = 1$, resulting in the acceptance of all the sampled parameters from the prior distributions. Then, the tolerance level of the next generation is set adaptively, by considering the $\boldsymbol{\alpha}$ quantile of $\rho_1^{(t)}, \rho_2^{(t)}, ..., \rho_N^{(t)}$, where

these are the $\rho(\mathbb{S}^{(sim)}, \mathbb{S}^{(obs)})$ distances from the accepted simulations at time t (Drovandi and Pettitt, 2011; Numminen et al., 2013). We assume $\boldsymbol{\alpha}$ = 0.4 throughout this chapter.

Contact network information is needed to fit the network-based ILM. Since we are assuming unknown contact network, for each epidemic simulated within the ABC-PMC algorithm, a contact network is generated via Equation 4.4. We then augment our parameter vector to estimate the unknown parameters of the contact network model as part of our ABC analyses. The specific summary statistics used in our analyses are described in Section 4.4.

4.3.2 Data Augmentation MCMC

In order to assess performance of the ABC-PMC algorithm, the simple network-based ILM is fitted using full Bayesian data-augmented MCMC (DA-MCMC) methodology. Specifically, the Metropolis Hastings algorithm is used to update both the model parameters and latent variables representing the unobserved infection times and contact network.

As discussed in Section 4.1, we assume that we have global information about the contact network. Specifically here, we assume we know the total number of connections of our network, ϕ_{tot}, where $\phi_{tot} = \sum_{i=1}^{N} \sum_{j=1}^{i} c_{ij}$. Then, given independent gamma distribution priors $\pi(\alpha_0)$, $\pi(\alpha_1)$, $\pi(\epsilon)$ and $\pi(\delta)$, for α_0, α_1, ϵ, and δ, respectively, the posterior density up to proportionality for the simple network-based ILM is given by

$$\pi(\boldsymbol{\theta}|\boldsymbol{R}, \boldsymbol{I}, \boldsymbol{C}, \boldsymbol{Z}) \quad \propto \quad L(\boldsymbol{R}, \boldsymbol{I}, \boldsymbol{Z}, \boldsymbol{C}|\boldsymbol{\theta})\pi(\alpha_0)\pi(\alpha_1)\pi(\epsilon)\pi(\delta)\pi(\phi_{tot}).$$

The conditional distribution of the infectious period rate parameter δ can be shown to be a

gamma distribution:

$$\delta|\alpha_0, \alpha_1, \epsilon, \boldsymbol{R}, \boldsymbol{I}, \boldsymbol{C}, \boldsymbol{Z} \sim \Gamma(m + a_\delta, M + b_\delta),$$

where $M = \sum_{i=1}^{m} (R_i - I_i)$ and a_δ and b_δ are the rate and shape parameters of the prior distribution of δ, respectively. Therefore, a Gibbs update (i.e., sampling from the conditional posterior distribution) is used to update δ. An independence sampler is used to update infection times, proposing new infection times I_i^* from the infectious period distribution, specifically $\gamma_i \sim$ Exp(δ). Then, the new infection time is just the difference between the fixed known removal time and the new infectious period of the i^{th} individual. Each infection time/infectious period is updated in this way in turn. Each element of the contact network ($c_{ij} = c_{ji}$) is also updated in turn. As c_{ij} has only two states, 0 (no connection) or 1 (connection), these parameters are updated by proposing a move from the current state to the only other possible state with probability 1. Parameters α_0, α_1 and ϵ are updated in turn, using Gaussian random-walk updates, tuned to achieve good mixing properties.

4.4 Epidemic Data and Summary Statistics

4.4.1 Simple Network-Based Data

Here, the performances of the ABC-PMC and MCMC approaches for small, simulated data sets are considered and compared across two scenarios: one in which the contact networks are fairly dense, and one in which they are fairly sparse. The data sets considered

here consist of a population of size $N = 25$ in which individuals are distributed uniformly at random across a square area of size 10×10 units. Two contact network scenarios are considered. In one scenario $\beta = 0.9$ and $\nu = 1$ (intense contact network scenario), and in the other $\beta = 1.8$ and $\nu = 1$ (sparse contact network scenario). Ten contact networks were generated for each scenario. Then, an epidemic was simulated across each of the 20 contact networks using the simple network-based ILM given by Equation 4.2, resulting in 10 data sets for analysis under each scenario. Epidemics were generated with model parameters, $\alpha_0 = 0.8$, $\alpha_1 = 0.5$, $\delta = 2.0$ and $\epsilon = 0$. It was assumed here that our population is isolated, hence the choice of $\epsilon = 0$. Other sets of model parameters were considered, and produced results in line with those shown in Section 4.5 (results not shown).

ABC-PMC was implemented for each epidemic data set using various sets of summary statistics. For brevity, three sets of such statistics are included here, some of low dimension and some of high dimension (shown in Table 4.1). Following McKinley et al. (2009), we based our choice of summary statistics primarily on the removal and infection epidemic curves (i.e., numbers removed and infected over time, respectively). However, statistics such as the final number of infected individuals (n_I) were also included.

We start with a low dimensional set of summary statistics, ABC-simple(1), based mainly on the removal times. This set includes scalar summary statistics that fully represent the removal time curve. As the removal times of most of the data sets have right-skewed distributions, we use the time at the peak of the removal times density rather than the mean. The time at the peak of the removal times curve is computed through estimating

the kernel density of the removal times using Gaussian kernel smoothing, with bandwidth $h = 1.06\ \hat{\sigma}_R\ n^{-1/5}$, where $\hat{\sigma}_R$ is the standard deviation of the removal times (Silverman, 1986).

Information regarding the infection time curves are also included in the second and the third sets. We considered a high dimensional set of summary statistics (ABC-simple(2)) that incorporates the numbers of infection events occurring within a set of equal time intervals of 0.2 time units ($\mathcal{I}_0^T$) from $t = 0$ to $t = T$ instead of the standard deviation of the removal times (σ_R) and the number of infected individuals (n_I). In the third group of summary statistics (ABC-simple(3)), we considered scalar summaries of $\mathcal{I}_0^T$, thus, forming a representation of ABC-simple(2) in lower dimensions.

The DA-MCMC method was used to carry out a full "gold standard" Bayesian analysis of each epidemic data set, sampling from the joint posterior of the model parameters, infection times, infectious periods and the upper triangular contact matrix (since the contact network matrix is symmetric). The degree distribution of the network is incorporated into the analysis via an observation model for the total number of connections (ϕ_{tot}), as described in Section 4.3.2. For this observation model, we use a normal distribution with mean equal to the true number of connections and variance 10.

Although epidemic data sets were generated with $\epsilon = 0$, we fit the data with $\epsilon(j,t) = \epsilon$, $\epsilon > 0$. Including a spark term when fitting the model helped avoiding high rejection rates that resulted when many MCMC moves were proposed with zero likelihood (e.g., if a proposed contact network change results in an infected individual now having no contacts

with any infectious individuals under the model). However, we fix $\epsilon = 0$ under ABC-PMC since the above issue is not exist with ABC-PMC.

For all data sets, independent prior distributions were assigned to all model parameters. The marginal prior distributions for the common model parameters under both ABC-PMC and DA-MCMC were as follows: $\alpha_0 \sim \Gamma(2, 2)$, $\alpha_1 \sim$ Exp(2), and $\delta \sim \Gamma(4, 2)$. Under the ABC-PMC, the scale parameter of the contact network is fixed at $\nu = 1$, and a uniform prior assigned to the spatial parameter β such that $\beta \sim$ U(0,5). Also, we assigned a non-negative, half-normal prior distribution with a variance of 100 and a mode of 0 for the spark term ϵ under the DA-MCMC. The ABC-PMC was run sampling 500 particles at each generation. The DA-MCMC was run for 150,000 iterations with a burn-in period of 10,000 iterations. The convergence of both the ABC-PMC and MCMC algorithms was diagnosed visually, monitoring the sample output chains.

4.4.2 The UK 2001 Foot-and-Mouth Disease

Here, we investigate the performance of ABC-PMC on a more complex network-based ILM in a larger population, specifically considering modelling data from, or based upon, the UK 2001 foot-and-mouth disease epidemic. The full UK 2001 FMD data set consists of an XY coordinate for every sheep and/or cattle farm in the UK, and the number of cattle and sheep on each farm recorded at the census carried out previous to the FMD outbreak. All farms on which FMD was detected had an infection time (estimated by veterinarians/veterinary epidemiologists on the ground), as well as a removal time (the

date on which animals were culled as part of the control policy) recorded. A further set of farms for which the presence of the infection was not confirmed, but the animals were culled anyway as part of the "high-risk farm" control policy, are also recorded along with the date of the cull. It is assumed that no farms that had animals infected with FMD went undetected.

Foot and mouth disease has been extensively modelled, in many cases assuming an $\mathcal{SEIR}$ framework (e.g., Keeling et al., 2001; Deardon et al., 2010; Jewell et al., 2009). Our model is simplified in comparison, primarily since we follow Malik et al. (2016) and Deeth et al. (2013) and assume the simpler $\mathcal{SIR}$ compartmental framework (extensions to other frameworks could be made relatively easily). We focus on subsets of farms in the county of Cumbria, one of the counties most highly affected by the 2001 FMD outbreak.

Simulated FMD Data

First, we simulated epidemics from the FMD network-based ILM given by Equation 4.3 through a population consisting of 674 farms, those being all the farms within a region of 25$\times$30km; see Appendix B.1. The data consist of the number of cattle and sheep on each farm, and an XY coordinate.

Simulated data sets were produced by first simulating a contact network from Equation 4.4 with spatial parameter $\beta = 4$ and $\nu = 1$. The value of $\beta = 4$ was chosen to produce relatively sparse networks, as the network structure of most real contact networks is typically sparse in nature (Danon et al., 2011). Then, an epidemic was simulated through

the contact network. The infectious period for each individual was simulated from a gamma distribution with shape parameter 10 and rate δ. A number of scenarios were tested, but here we show results for parameters $\alpha_s = 0.0001$, $\alpha_c = 0.01$, $\phi_s = 0.001$, $\phi_c = 0.005$ and $\delta = 1.5$. Conclusions from other scenarios were similar to these shown here.

Two scenarios regarding the spark term were considered. First, in Scenario 1, ten epidemics were simulated from the ILM with $\epsilon = 0$, thus, assuming that no infections come from outside the observed population. Second, in Scenario 2, ten epidemics were simulated from an ILM with $\epsilon = 0.001$, allowing for infections from outside the observed population.

The FMD network-based ILM was fitted to the simulated FMD epidemic via ABC-PMC with two different sets of summary statistics (see Table 4.1). We found that stratifying the number of infected farms (n_I) by farm size and type helped to inform the contact network parameter. The summary statistic used was the following chi-squared statistic:

$$\chi^2_{s,c} = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i},$$

where O_i and E_i are the observed and expected[2] number of infected farms in the i^{th} stratified class. The four classes are defined by the number of sheep and cattle upon the farms. First, let m_c and m_s denote the median numbers of cattle and sheep on farms in the observed population, respectively. Then, large cattle farms are defined as having more than m_c cattle, but fewer than m_s sheep; large sheep farms are defined as having more than m_s sheep, but fewer than m_c cattle; large mixed farms are defined as having more than m_c

[2]The observed is the number in each stratified class, and the expected is calculated as the product of the total number of the stratified class's row and column divided by the total number of classes.

cattle and m_s sheep; and the final class, small mixed farms, are defined as having fewer than m_c cattle and m_s sheep. This $\chi^2_{s,c}$ is then calculated for both the observed and ABC-simulated data. We present results for a set of six summary statistics (ABC-sim(1)), as well as a higher dimensional second set (ABC-sim(2)) containing $\mathcal{I}_0^T$.

For all analyses, we assume the scale parameter of the contact network is known and fixed at $\nu = 1$, but estimate the unknown spatial parameter β. Independent prior distributions were assigned to the model parameters, as follows: $\alpha_c, \phi_s, \phi_c, \epsilon \overset{iid}{\sim}$ Exp(100), $\delta \sim \Gamma(3, 2)$ and $\beta \sim$ U(0,8). Since the model of Equation 4.3, as given, is non-identifiable, following Deardon et al. (2010) we fix α_s and set it equal to its true value of 1×10^{-4}. The ABC-PMC was run with a sample size of 500 particles at each generation.

Note that, although the population size of 674 individual farms is still relatively small, a data augmented MCMC analysis to estimate the joint posterior of the model parameters and contact network proves computationally intractable and so could not be carried out.

Real FMD Data

Finally, we consider fitting the FMD network-based ILM to real data, consisting of a subset of 1795 farms observed during the 2001 UK FMD epidemic within a region of 58.4×52.5 km; see Appendix B.1. This data set includes 456 farms which were infected during the epidemic. We treat the cull dates of animals on these farms as the removal times. Connections between farms in the contact network represent any potential infection route (e.g., contact between animals and/or farm workers, transfer of infective material by

vectors such as wildlife or vehicles, etc.). We fit our model to data from the date of animal movement ban introduced on 23 February 2001, conditioning on the infection status of farms at that date. That is, all farms deemed infectious on 23 February we consider to be "initial infections". We assume known infection times (estimated by vets on the ground) and removal times when fitting models to this data set. Since simulation of the "high risk" farm culling procedure implemented in 2001 is extremely difficult, we do not model this aspect of the epidemic here. Instead, we follow Jewell et al. (2009), and assume all these "high risk" farms are susceptible throughout the epidemic.

In this work, we once again present our investigation of the performance of ABC-PMC using two sets of summary statistics. Table (4.1) shows the two chosen sets, one of high, and one of low, dimension. The high dimensional set of summary statistics (ABC-real(1)) consists of P_R, L_{epi}, n_I, $\chi^2_{s,c}$ and $\mathcal{I}_0^T$ at equal time intervals of 5 days. The lower dimensional group of summary statistics (ABC-real(2)) is similar to ABC-real(1), but with the infection time counts being replaced by P_I and σ_I.

The ABC-PMC was performed with fixed scale parameter $\nu = 1$ for both sets of summary statistics. Independent prior distributions are assigned to the model parameters as follows: $\beta \sim$ Exp(0.3), α_c, ϕ_s and $\phi_c \overset{iid}{\sim}$ Exp(100) and $\delta, \epsilon \overset{iid}{\sim}$ Exp(10). To solve the issue of non-identifiability, we set α_s (arbitrarily) to 1×10^{-7}. Additionally, analyses were carried out in which ν was treated as unknown. We show such results for ABC-real(1). In this case independent, marginal prior distributions were assumed for the contact network parameters such that $\beta, \nu \overset{iid}{\sim}$ Exp(1). The other model parameters have the same prior distributions as

Table 4.1: The sets of summary statistics used, where $P_{(\bullet)}$ denotes the time at the peak of the epidemic curve (R removal and I infection times), n_I is the number of infected individuals, L_{epi} is the length of epidemic, $\mu_{(\bullet)}$ and $\sigma_{(\bullet)}$ are the mean and standard deviation of the infection or removal times, $\mathcal{I}_0^T$ is the number of infection events over equal time intervals; and $\chi^2_{s,c}$ is the chi-squared statistic as described in Section (4.4.2).

Simple Network-based data			Simulated FMD data		Real FMD data	
ABC-simple (1)	ABC-simple (2)	ABC-simple (3)	ABC-sim (1)	ABC-sim (2)	ABC-real (1)	ABC-real (2)
P_R	P_R	P_R	μ_R	μ_R	P_R	P_R
L_{epi}	L_{epi}	σ_R	σ_R	σ_R	L_{epi}	σ_R
σ_R	$\mathcal{I}_0^T$	P_I	μ_I	$\mathcal{I}_0^T$	$\mathcal{I}_0^T$	P_I
n_I		σ_I	σ_I	$\chi^2_{s,c}$	n_I	σ_I
		n_I	n_I		$\chi^2_{s,c}$	$\chi^2_{s,c}$
			$\chi^2_{s,c}$			n_I

above.

4.5 Results

In this section, we present the results of our analyses. The independence of sampling in the ABC-PMC algorithm at each generation allows the computation to be run in parallel to achieve lower computation times. Parallel computation was implemented using a distributed-memory architecture by the Message Passing Interface Standard (MPI). This was run on multiprocessors servers with 24 CPU[3] cores at 2.5 GHz and 256 GB of RAM for each server. This was done for ABC-PMC for the FMD network-based ILMs. Note that the used servers had a time limit for running jobs restricted to 168 hours. Computations for the simple network-based ILMs were performed on an Apple iMac with i5-core Intel

[3]A central processing unit (CPU)

2.9 GHz processors with 16 GB of RAM without parallelization. All coding was done in Fortran 95.

4.5.1 Simple Network-Based ILM Data

Figure 4.2 shows the approximate posterior means and 95% credible (percentile) intervals for the ILM parameters for each ABC-PMC analysis, under both intense and sparse contact network data sets. The performances under the three sets of summary statistics were similar, with true values of parameters being included within 95% credible intervals throughout. However, we found that including more information from the infection times in the summary statistics, (ABC-simple(2) and ABC-simple(3)), appeared to slightly enhance estimation of the parameters with tighter credible intervals, especially for α_0 and α_1.

Under both sparse and intense networks, we appeared to be able to estimate β, the contact network spatial parameter, quite well. However, we observed better estimates of this parameter under the intense network data sets, with systematic underestimation and a larger posterior variance under the sparse networks. The approximate posterior estimates of α_0 and α_1 were closer to their true values under ABC-simple(2) and ABC-simple(3) compared to that under ABC-simple(1), where overestimation of both parameters was observed. Thus, the inclusion of the infection times in the summary statistics appeared to enhance the estimation of these parameters. This can be seen most clearly for α_0 with approximate posterior means closer to the true value, and with less variation under ABC-

simple(2) and ABC-simple(3) for both intense and sparse networks. The infectious period δ estimates were similar under each of the three sets of summary statistics.

Posterior means and 95% credible intervals of the model parameters for the full Bayesian MCMC analysis are shown in Figure 4.3. More accurate estimation of α_0 and α_1 with tighter credible intervals under both types of networks was achieved than under ABC-PMC. However, the infectious period rate δ had wider credible intervals than under ABC-PMC with a lot of variation in performance between data sets. The estimated degree distributions under both types of networks were quite close to the observed ones; see Appendix B.2. This indicated the benefits of incorporating the observational model of the total number of connections in the update of each c_{ij}. The analyses were also carried out without the observation model and large bias estimate of the degree distribution were detected. In both types of network, the estimated degree distribution tended to be overestimated resulting in very intense contact network.

Both MCMC and ABC methods generally produced good estimates of the model parameters under both types of network. However, as expected, approximate posterior variances were larger under ABC-PMC than were the estimated posterior variances under the full Bayesian MCMC analyses.

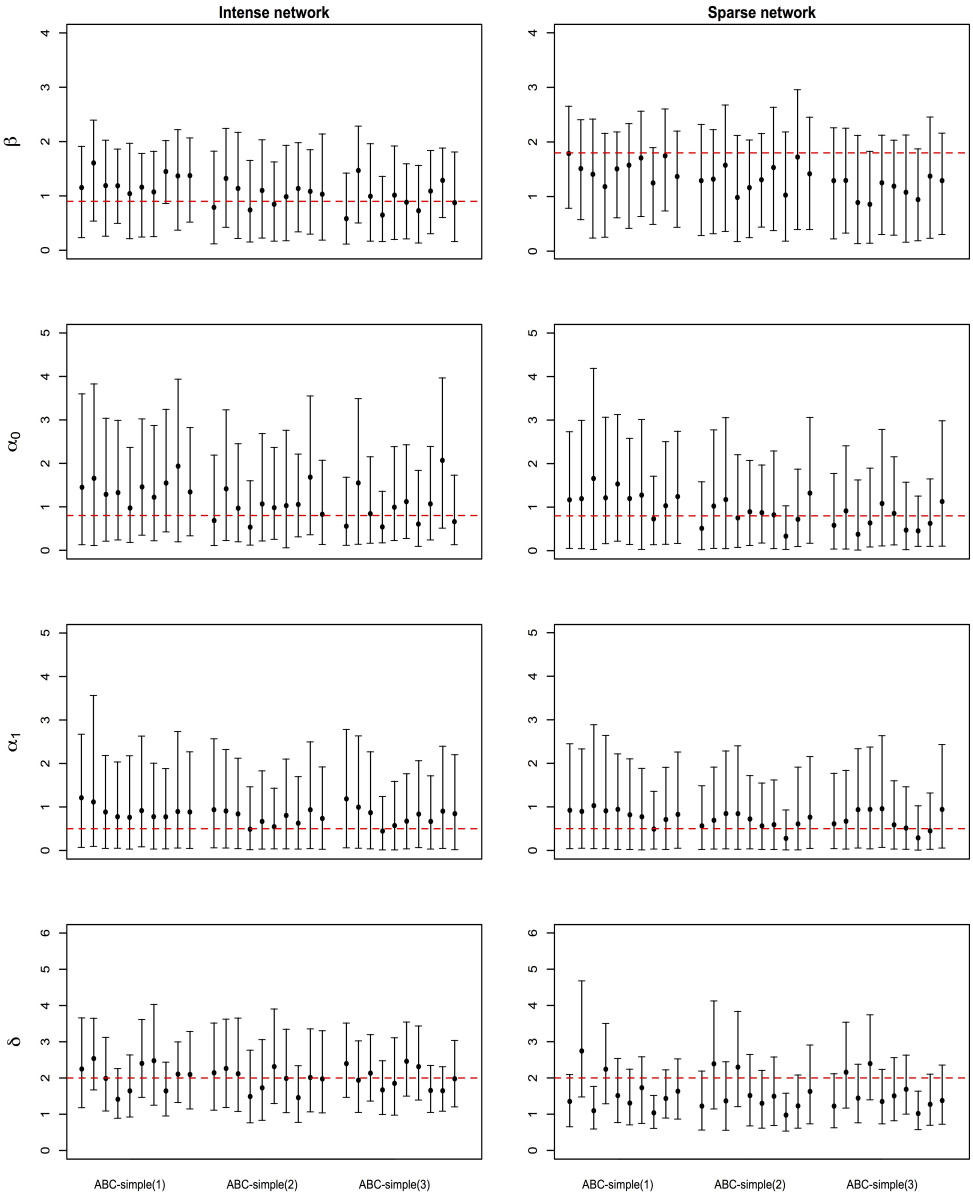


Figure 4.2: Posterior means for the simple network-based ILM with 95% credible intervals for the intense and sparse contact network epidemics. Each plot represents the results of applying different summary statistics in the ABC-PMC for the intense (left) and sparse (right) network epidemics. Summary statistics ABC (1), ABC(2) and ABC(3) are described in Section 4.4.1. Horizontal dotted lines are the true values of the parameters; β = 0.9 (intense network) and 1.8 (sparse network), α_0 = 0.8, α_1 = 0.5 and δ = 2.

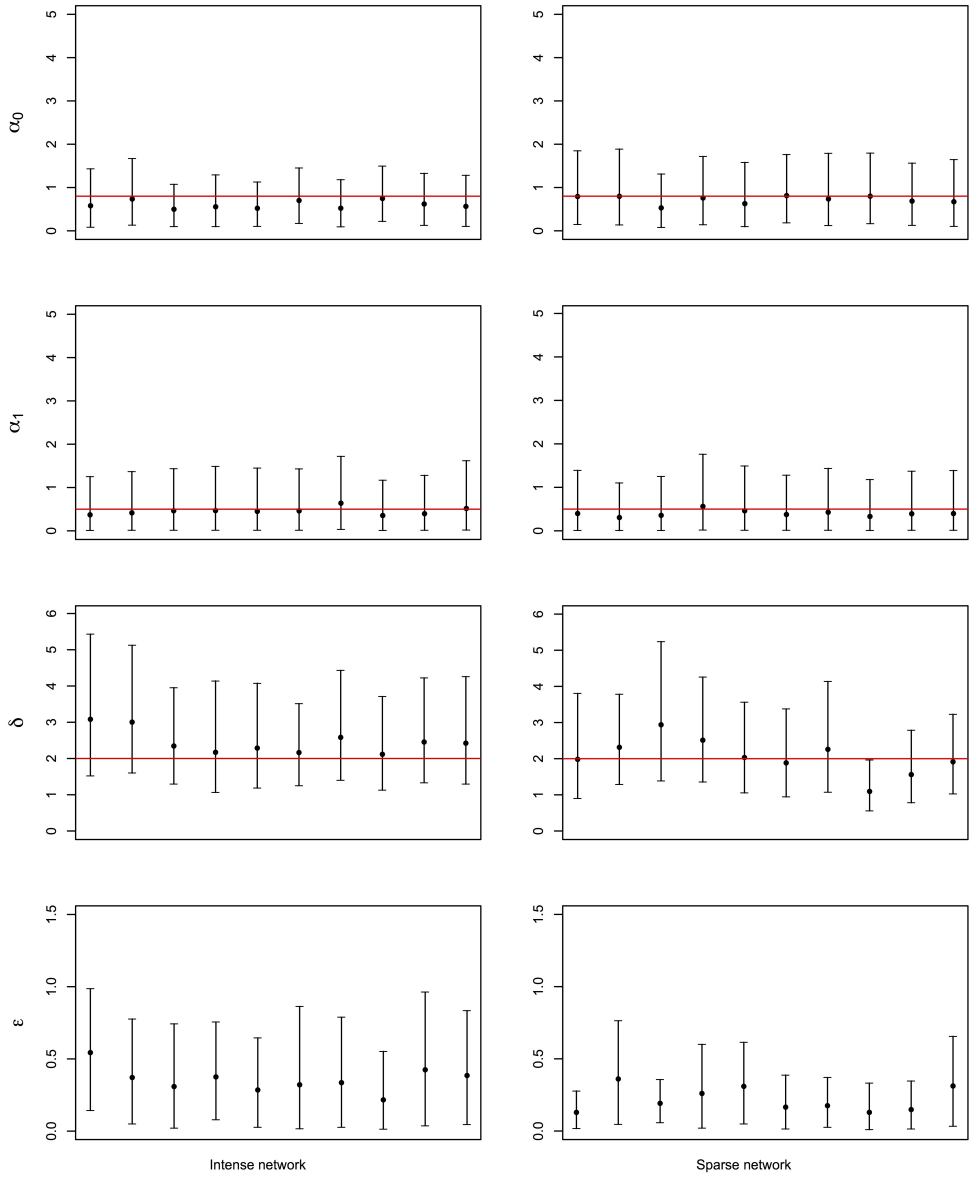


Figure 4.3: Posterior means for the simple network-based ILM with 95% credible intervals of the intense (left) and sparse (right) contact network epidemics based on the full Bayesian MCMC analyses. Horizontal solid lines are the true values of the parameter; α_0 = 0.8, α_1 = 0.5 and δ = 2.

4.5.2 Simulated FMD Data

We examine the performance of ABC-PMC with low and high dimensional summary statistics. Figure 4.4 shows the resulting approximate posterior means and 95% credible

intervals of applying the ABC-PMC method on the simulated FMD data sets. Results, in general, indicate good estimation of the model parameters for most of the epidemics. Both low and high dimensional summary statistics provided similar estimates of the model parameters, with the true values of the model parameters contained in the 95% credible intervals in most instances. The approximate posterior variance was noticeably higher under the high dimensional summary statistics ABC-sim(2), however. Model parameter estimates were reasonably good under both Scenario 1 (no spark term used in either the generation or fitting of epidemic data), and Scenario 2 ($\epsilon = 0.001$ for data generation and estimated in the ABC analysis).

We also achieved good estimation of the latent information, both infection times and contact networks. The approximated posterior distributions of the infection time density of some representative simulated FMD epidemics are displayed in Appendix B.3 for both groups of summary statistics. As observed with the parameter estimates, we observe much greater variation under the ABC-sim(2) analysis (high dimensional summary statistics) than under ABC-sim(1) analysis (low dimensional summary statistics).

4.5.3 Real FMD Data

Approximate posterior means and 95% credible intervals of the model parameters for the real FMD epidemic data analyses are shown in Table 4.2. Although there is reasonable agreement between some parameters (e.g., the sparks term and infectious period rate parameters), with substantially overlapping credible intervals, there are also some noticeable

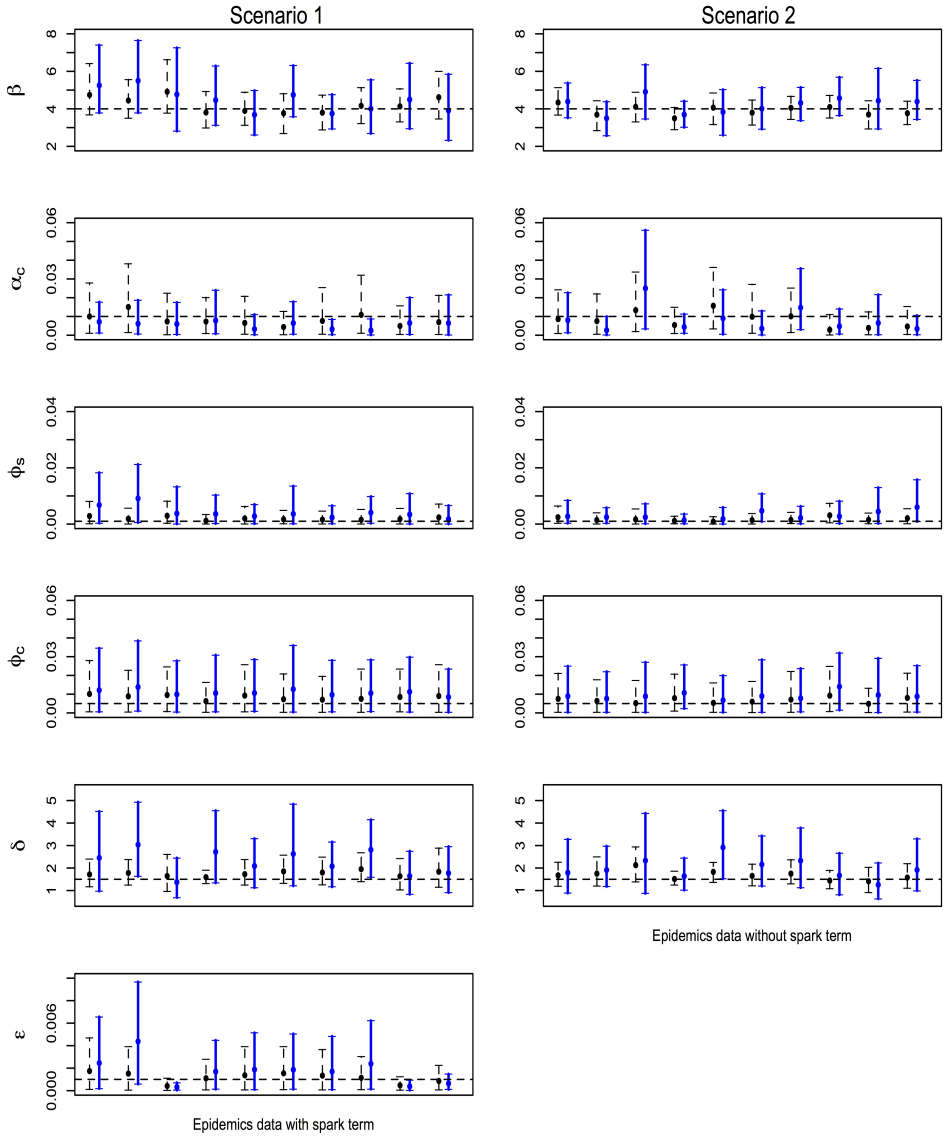


Figure 4.4: Posterior means and 95% credible intervals for fitting the FMD network-based ILM to the simulated FMD epidemics with the spark term (left, scenario 1) and without spark term (right, scenario 2). The black dotted lines represent results of applying ABC-PMC with summary statistics ABC-sim(1) and blue solid lines of applying ABC-PMC with summary statistics ABC-sim(2); as described in Section 4.4.2. Horizontal dotted lines are the true values of the parameters; $\beta = 4, \alpha_c = 0.01, \phi_s = 0.001, \phi_c = 0.005, \delta = 1.5$ and $\epsilon = 0.001$

Table 4.2: Posterior means and 95% credible intervals for parameters for the FMD network-based ILM fitted to the real FMD epidemic data.

Parameters	Units	ABC-real(1)		ABC-real(2)
		ν known	ν unknown	ν known
β		3.012 (1.253, 5.871)	2.840 (0.914, 5.645)	12.731 (4.932, 23.799)
ν		1.0	0.622 (0.053, 1.970)	1.0
α_c	10^{-4}	2.354 (0.077, 11.165)	7.664 (0.274, 34.901)	45.206 (3.334, 143.498)
ϕ_s	10^{-3}	3.273 (0.035, 13.922)	5.822 (0.125, 22.476)	11.242 (0.265, 34.315)
ϕ_c	10^{-2}	1.042 (0.086, 2.880)	0.912 (0.039, 2.518)	0.795 (0.047, 2.243)
δ	10^{-1}	1.788 (0.804, 2.969)	1.717 (1.097, 2.289)	1.475 (0.575, 3.112)
ϵ	10^{-4}	2.092 (0.071, 5.769)	3.206 (0.188, 8.964)	1.365 (0.123, 5.231)

differences in the three analyses. This is perhaps most obvious with the spatial network parameter, β, with a much larger estimate of this parameter for the summary statistics ABC-real(2). These larger estimates of β imply a very sparse underlying network is being inferred in which connections between farms occur predominantly over very short distances (e.g., less than 1 KM).

Figure 4.5 shows realizations from the ABC-approximated posterior epidemic curves (infection times and removal times) from the three ABC-PMC analyses. We can see that

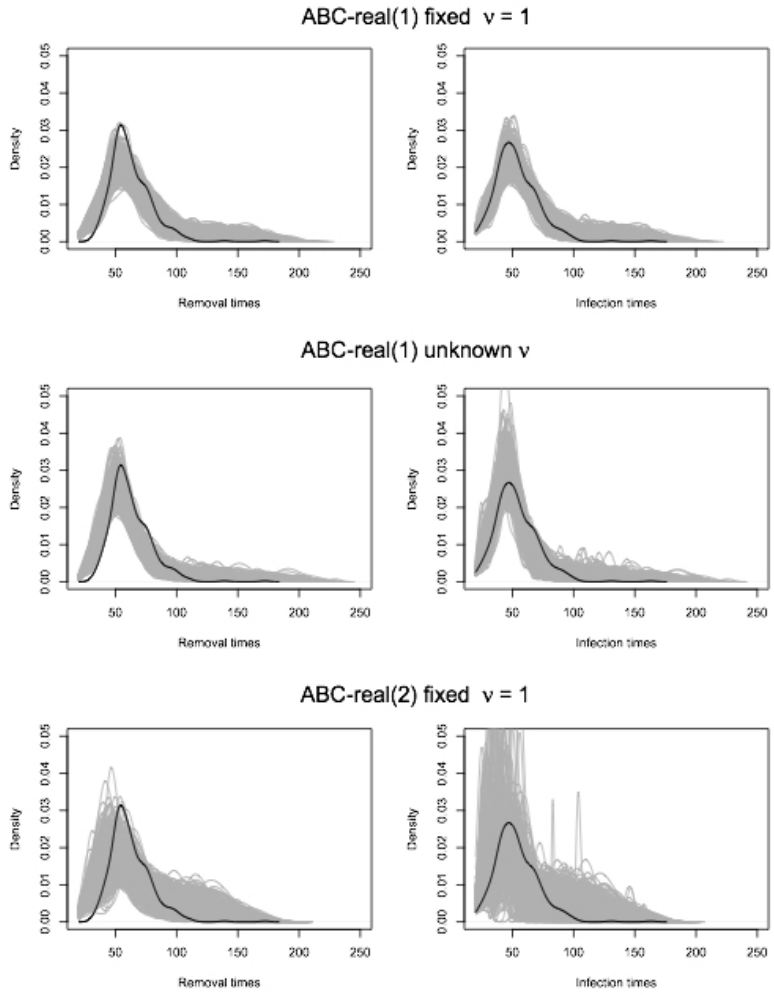


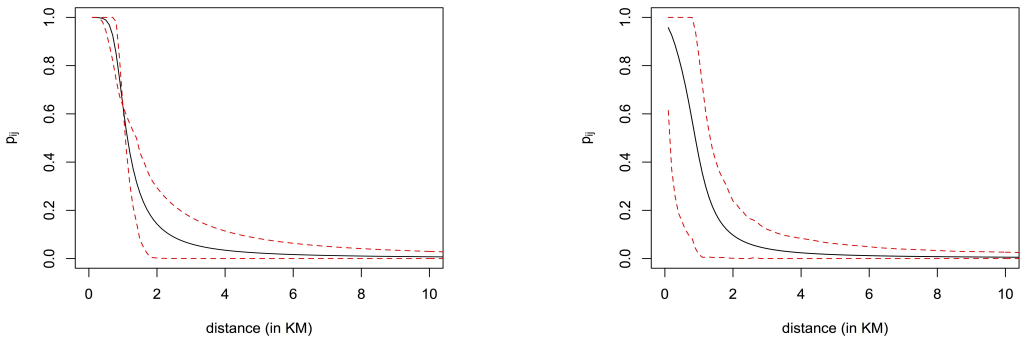
Figure 4.5: The approximated posterior distribution of the infection and removal time curves for the real FMD epidemic data analyses.

prediction of these curves was better under the two analyses of ABC-real(1) than ABC-real(2), with much lower variance. This would imply that the estimates under ABC-real(1) are more reliable.

Figures B.4.1a and B.4.1b (Appendix B.4) shows the intermediate posterior distribution in sequential generations of the ABC-PMC, using ABC-real(1) with fixed ν. We can see that the analysis is informative, with the final approximate posterior distributions appear-

ing quite different to the prior distributions. The prior distributions chosen were relatively informative. Similar analyses were carried out with less informative priors, and the approximate posterior distributions found to converge to approximately the same distributions as seen here. However, this was accompanied with a slower convergence rate. (This is discussed further Section 4.5.4.)

Since we are using a relatively small subset of the 2001 UK FMD epidemic data set, a direct comparison of our results to previously published studies (e.g., Deardon et al., 2010; Jewell et al., 2009; Keeling et al., 2001), which were based on the whole of the UK or much larger subsets, is difficult. However, the estimate of the network spatial parameters under ABC-real(1) can be compared to the spatial infection kernel estimated in some of those studies. Figure 4.6 shows the posterior mean and 95% credible intervals of the probability of a connection within the farm network under the fitted model assuming both known and fixed, and unknown, ν settings. These results imply a very small chance of connections within the network between farms that were 4 KM apart or greater, which is in agreement with these other published works. This fact, along with reasonably good posterior prediction of the epidemic curves provide some evidence that results we observe are reasonable. Whether treating the scale parameter of the contact network as unknown, or fixing $\nu = 1$, similar posterior estimates of the model parameters results were observed. Of course, fixing $\nu = 1$ forces $p(c_{ij} = 1) = 1 - e^{-1}$ when $d_{ij} = 1$ and leads to tight credible intervals for d_{ij} close to 1. However, this does not appear to have had much effect on the model fit as the proportion of distances ($\leq$1 KM) between farms is very small. On the other hand, varying



(a) Fixed $\nu = 1$ (b) Unknown ν

Figure 4.6: Posterior mean and 95% credible intervals of the probability of connections p_{ij} between farms in the inferred network for the real FMD data using summary statistics ABC-real(1) based on the posterior mean of the contact network parameters.

ν added more flexibility to the probability of connections within farms in short distances, but with larger posterior variation.

4.5.4 Computation Time

Table 4.3 shows the computation time using unparalleled ABC-PMC and DA-MCMC algorithms to fit the network-based ILM to a single simple network-based data set, and a simulated FMD epidemic data. The computation time is for obtaining 150,000 DA-MCMC samples and 500 particles at each generation of ABC-PMC (15 generations). The computational time of running DA-MCMC becomes infeasible as the population size increases and is approximated from running a single MCMC iteration. This rough approximation was used since the number of missing connections increases drastically with increases in population size, and as a result, updating each connection via DA-MCMC requires evaluating

Table 4.3: The computational time (in minutes) of using both DA-MCMC and ABC-PMC on fitting network-based ILMs for a simple network-based data ($N = 25$) and a simulated FMD epidemic data ($N = 674$).

Population size	Number of c_{ij}	Computation time in minutes	
		DA-MCMC	ABC-PMC
N = 25	300	14.54	44.41
N = 674	226801	$\approx 6.22\times10^7$	9807.95

the likelihood a huge number of times at each iteration.

Using DA-MCMC took approximately three times less to run than the ABC-PMC with small data contains $N = 25$ individuals. However, as the population size increases, ABC-PMC becomes increasingly more efficient than DA-MCMC. Using ABC-PMC with the simulated FMD epidemic contains $N = 674$ individuals took around six days and 20 hours to run while DA-MCMC had infeasible computation time. The computation time of applied parallel computing on the same simulated FMD epidemic was approximately ten hours with 12 processors. Parallel computing for fitting the model to the real FMD epidemic, for both groups of summary statistics, took around a week using 64 processors.

4.6 Discussion

The main focus of this Chapter has been fitting network-based ILMs of disease transmission in situations where there is uncertainty in both infection times and the underlying contact network. Our results suggest that ABC-PMC approaches to inference in such scenarios can produce good estimates of model parameters and fitted models with good pre-

dictive qualities. Good estimation of the underlying unknown contact network can also be obtained. However, this area of research is a work in progress, and the positive results we have achieved here do come with some caveats that require further exploration.

The first caveat pertains to the selection of summary statistics used to compare simulated and observed data within the ABC algorithm. We have explored the use of numerous combinations of summary statistics for fitting our models, only a small subset of which have been included here. Selecting an "optimal" set of summary statistics for a given analysis is not easy. For example, in some situations it appeared that including a discretized version of the epidemic curve produced more satisfactory results than a more parsimonious set of summary statistics of the epidemic curve, and sometimes vice versa. It is still not clear how we would go about predicting which would work better in advance. This challenge would suggest that either multiple ABC-PMC analyses should be carried out (preferably) in parallel when analyzing epidemic data, or some approach that allows for the comparison of different sets of summary statistics to be carried out as part of the analysis (Joyce et al., 2008).

The second caveat pertains to the choice of priors for model parameters. Simulation-based approaches such as ABC tend to perform much better when informative priors are used. This is less of a problem for ABC-PMC approaches than, say, rejection-based ABC approaches, since the sequential learning inherent in the ABC-PMC algorithm can successfully hone in on the areas of higher approximate posterior mass with even quite uninformative priors. However, the computational ramifications of using less informative priors can

be large, exhibited by both a slower decrease in the tolerance level ξ between generations, and a higher rejection rate within generations. This highlights the care that must be taken when selecting priors given limited computational resources.

Of course, even with this caveat, the huge reduction in computational burden resulting from the use of ABC-PMC, as opposed to DA-MCMC, is extremely impressive. In fact, as we have seen here, ABC-PMC approaches can be used to deal with otherwise computationally intractable analyses. Furthermore, it is possible to envisage ways of increasing this improvement even more. For example, one issue with individual-level epidemic models is the amount of variation observed between simulations from the same model (i.e., with the same parameters). This variation means that rejection rates within the ABC algorithm can be very high. One possible approach to dealing with this problem is the technique of so-called "Lazy ABC" introduced by Prangle (2016), in which simulated data sets that substantially differ from the observed data are rejected early in the simulation, rather than generating an entire data set (here, epidemic).

There are also aspects of the model that would be interesting to explore in further work. We have confined ourselves to considering static, binary, spatial undirected networks. Of course, in many situations the underlying contact network may not be spatial in nature, and may be directional and/or weighted. For example, consider livestock diseases in which – often long distance – animal movement between farms and/or markets may be drive disease spread through a dynamic, weighted, non-symmetric network.

In conclusion, methods such as ABC-PMC do seem to offer a plausible way of carrying

out inference for complex infectious disease models in the presence of multiple layers of data uncertainty. Of course, in scenarios where Monte Carlo methods such as data augmented MCMC are plausible, they are to be preferred providing, as they do, a full, coherent Bayesian analysis. However, in infectious disease epidemiology it is very easy to find systems in which the size of the population and/or level of uncertainty in the data precludes such approaches in anything like an acceptable timeframe. We thus conclude that further exploration into the use of ABC methods for fitting infectious disease models to data is certainly warranted.

Chapter 5

Contact Network Uncertainty in Individual Level Models of Infectious Disease Transmission Using Within-Cluster Likelihood Approximations[1]

5.1 Introduction

Modelling the dynamics of infectious disease transmission in a heterogeneous population can assist policy-makers in making reliable and fast decisions in order to control the spread of disease. The use of network information has been found to be beneficial in understanding the underlying processes of disease transmission and may come in the form of a contact network (binary or weighted) or a series of such networks (e.g., representing the movement of animals, individuals and/or vehicles between farms in an agricultural setting). However, there is a lack of availability of network information, often due to the high expense associated with the collection of contact data, privacy issues, or recall error (Danon

[1]Chapter 5 is to be submitted to a peer-reviewed statistical journal.

et al., 2011). Regardless, the inherent uncertainty of disease systems presents a serious computational challenge.

Various approaches to the incorporation of contact networks in infectious disease models can be found in the literature. Several models have been used that restrict the spread of the infectious disease across only the edges of a contact network. Some research studies employing this approach are heavily contingent upon simulation through a given, or simulated, contact network that has certain properties of the population (Volz, 2008).

Other approaches model the transmission network via stochastic epidemic models. Britton and O'Neill (2002) introduced a homogeneous stochastic epidemic model involving a transmission network based upon Bernoulli random graphs (BRG). In this network model, any two individuals (nodes) are assumed to have the same probability of being connected. An extension of this model is the exponential-family random graph models (ERGMs), introduced by Groendyke et al. (2012). The researchers developed the model to a more general form, which allows the transmission model to incorporate various covariates, and was applied to a measles outbreak in children in the town of Hagelloch, Germany in 1861. A recent extension of the model was also introduced by Lee et al. (2018a). The authors developed a network epidemic model by combining a stochastic epidemic model and a preferential attachment (PA) model, previously introduced by Barabási and Albert (1999). A preferential attachment network is generated by adding individuals (nodes) sequentially and modelling the number of new connections (edges) as random variables that follow specific distributions (e.g., a truncated or censored Poisson distribution). Lee et al. (2018a)

have successfully combined the PA model with a Susceptible-Infectious ($\mathcal{SI}$) epidemic model to examine epidemics that spread across social networks, and have demonstrated better performance of this approach compared to the use of the BRG model.

However, rather than modelling the transmission network, a contact network (or networks) can be incorporated as a covariate in a stochastic epidemic model to account for the effect of existing connections (potential infection routes) between individuals in the population. For example, Jewell et al. (2009) used contact network data, along with other covariates, to model a simulated High Pathogenicity Avian Influenza (HPAI) H5N1 outbreak in the poultry industry of Great Britain. The study incorporated two contact networks indirectly by considering the rates of truck contacts for feed mills and slaughter-houses at the level of premises. The authors have also suggested allowing for uncertainty in the network specification, although they did not do this.

Making inferences for such models is usually carried out within the Bayesian framework using the Markov Chain Monte Carlo (MCMC) method. However, the increasing complexity of infectious disease models can make the use of MCMC computationally challenging. One of the problems typically encountered in infectious disease modelling is the lack of data availability. Infection times of individuals are generally not observed, with only partial information about the infection time given by an observed reporting, diagnosis, or removal time. Contact network data are also often not observed. It can be inferred via data augmentation techniques, but for large populations there is a substantial amount of missing information.

In the case of epidemics where infection times are assumed to be unknown, data augmented MCMC (DA-MCMC) has been successfully applied for fitting spatial stochastic epidemic models. Examples include the 2001 UK foot-and-mouth outbreak (Streftaris and Gibson, 2004; Jewell et al., 2009), the SARS outbreak of 2003 (McBryde et al., 2006), health care-associated infections (such as MRSA and *C. difficile*) (Forrester et al., 2006; Kypraios et al., 2010) and avian, H5N1 influenza (Jewell et al., 2009).

Non-standard and problem-specific MCMC algorithms have been designed to improve the efficiency of standard algorithms. For example, Agarwal and Duchi (2012) have used parallelization methods for calculating the likelihood function at each MCMC iteration. Unfortunately, the significant communication time between processors required for information exchange presents serious limitations, especially in high-dimensional problems.

Alternatively, one could partition the data into a number of conditionally independent samples. Independent MCMC chains are then run in parallel to gain samples from each sub-posterior distribution. These samples are subsequently recombined to represent approximate samples from the full posterior distribution. Scott et al. (2016) have facilitated this approach with the use of the consensus Monte Carlo (CMC) algorithm, while Entezari et al. (2017) have applied this approach in the context of inflating the likelihood function to obtain a better representation to the full posterior by each sub-posterior. This latter approach is known as the likelihood inflating sampling algorithm (LISA).

In this chapter, we consider the incorporation of spatial contact network uncertainty when fitting contact network-based infectious disease models to partially observed epi-

demic data using the MCMC method. We have employed a continuous time version of the discrete time individual-level model (ILM) framework for studying infectious disease spread within a Bayesian framework introduced by Deardon et al. (2010). However, implementing this model in even moderately large populations proves computationally intractable with standard MCMC due to the large amount of missing information. We introduce an approach to approximate full model inference by partitioning the population into a number of spatial clusters. More specifically, the contact network is divided into a series of spatially isolated "cluster-constrained" contact networks. Between cluster infections are modelled instead by spark infection parameters that allow for random "spark" infections. This partition scheme results in a large reduction in the number of latent connections in the contact network that need to be inferred. Consequently, the computational burden of the analysis is much reduced. We call the new model a *within-cluster network-based continuous time ILM*. It allows the infectivity rate of each individual to be dependent on local effects (covariates and contact network) within a cluster, and global effects (spark infection terms) that allow for infection between clusters.

In the present study, we illustrate the fit of the within-cluster network-based continuous time ILMs on simulated epidemics that are generated through an undirected spatial contact network. In addition, we explore the effect of data uncertainty when fitting the within-cluster model in three scenarios under different data uncertainty assumptions: 1) fully and correctly observed epidemics; 2) unknown infection times (and infectious periods) but observed contact network; and 3) completely unobserved contact network and

infection times. These analyses have been carried out within two different population structures (well-clustered and scattered populations). In all cases, removal times are assumed known. A comparison to the likelihood inflating sampling algorithm (LISA) introduced by Entezari et al. (2017) is investigated through multiple epidemic data sets.

The rest of this chapter is structured as follows. In Section 5.2, we introduce the within-cluster continuous time network-based ILMs and contact networks considered. A concise description of the LISA method is also given. Subsequently, we detail Bayesian MCMC-based parameter estimation in Section 5.3, which is followed by details of the epidemic simulation studies. The results are described in Section 5.4, along with the comparison to LISA. A final discussion is presented in Section 5.5.

5.2 The Model

In this study, we have employed a continuous time version of the general framework of the discrete time ILMs introduced by Deardon et al. (2010). This framework allows for modelling disease spread through heterogeneous populations. We have applied this framework in the context of a susceptible-infected-removal ($\mathcal{SIR}$) compartmental class of models. Each individual i in the continuous time $\mathcal{SIR}$ model is assumed to be in the $\mathcal{S}$ (susceptible) state until being infected. As soon as the individual is infected, he/she moves to the $\mathcal{I}$ (infectious) state, and becomes able to transmit the disease to other susceptible individuals. Infected individuals are assumed to be infectious for a period called

the "infectious period". Following the infectious period, the individual transitions to the $\mathcal{R}$ (removed) state. Removed individuals may be those who recover and become immune to the infection, die, or for some other reason become detached from transmission of the infection (e.g., are quarantined).

The use of ILMs allows for the inclusion of susceptibility and transmissibility covariates at the individual level, as well as incorporating a kernel function to represent the infectious connections between susceptible and infectious individuals. The general infectivity rate for each susceptible individual j at time t is given by:

$$\lambda_{jt} = \left[\Omega_S(j) \sum_{i \in \mathcal{I}(t)} \Omega_T(i)\kappa(i,j)\right] + \epsilon(j,t), \tag{5.1}$$

where Ω_S and Ω_T are the susceptibility and transmissibility risk functions at the individual level; $\kappa(i,j)$ is the infection kernel that represents the function of risk factors involving both the j^{th} susceptible and i^{th} infected individuals; and $\epsilon(j,t)$ is a spark function which allows for random infections otherwise unexplained by the model. Random infection might represent, for example, the infection of a susceptible individual from a source outside of the observed population.

5.2.1 The Network-Based Continuous Time ILMs

We consider here a single, binary, undirected time-homogeneous contact network that represents all of the existing connections between individuals. Specifically, we assume

$\kappa(i,j) = c_{ij}$, where

$$c_{ij} = \begin{cases} 1, & \text{if a connection exist between individuals } i \text{ and } j,\ i \neq j. \\ 0, & \text{otherwise.} \end{cases}$$

Also, we assume no transmissibility risk covariates so that $\Omega_T(i) = 1$, and a susceptibility risk function such that $\Omega_S(j) = \alpha_0 + \alpha_1 z_j$, where α_0 is the baseline infectivity parameter; z_j is a susceptibility covariate; and α_1 is the covariate effect.. We have also fixed the spark function such that $\epsilon(j,t) = \epsilon$, $\epsilon \geq 0$. Thus, the general infectivity rate given by (5.1) becomes as follow:

$$\lambda_{jt} = \left[\alpha_0 + \alpha_1 z_j \sum_{i \in \mathcal{I}(t)} c_{ij}\right] + \epsilon, \tag{5.2}$$

5.2.2 Within-Cluster Network-Based Continuous Time ILMs

Accounting for data uncertainty via data augmented MCMC when fitting the network-based continuous time ILMs given by Equation 5.2 to data can lead to highly computationally expensive analyses. In the case of having a large population size and assuming a completely unknown contact network and infection times, fitting Equation 5.2 using data augmented MCMC easily becomes computationally intractable. This intractability is the result of the number of missing contact network elements increasing drastically with increasing population size, and requires the likelihood function to be evaluated a large number of times at each MCMC iteration.

Toward alleviating this computational issue, we partition the population into K spatial clusters and extend the model in Equation 5.2 to account for the within- and between-clusters infections. We consider the contact network to be constrained to only allow infection within clusters. Thus, the connections between individuals from different clusters are discarded. Infections that occur between clusters are allowed for by a series of cluster-level spark terms, ϵ_k. The infectivity rate of a susceptible individual j in cluster k at time t is then given by:

$$\lambda_{jkt} = \left[\alpha_0 + \alpha_1 z_j^{(k)} \sum_{i \in \mathcal{I}^{(k)}(t)} c_{ij}^{(k)}\right] + \epsilon_k, \tag{5.3}$$

where $\mathcal{I}^{(k)}(t)$ is the set of infectious individuals at time t within cluster k; ϵ_k is the corresponding unknown spark infection parameter for susceptible individuals in cluster k; $c_{ij}^{(k)}$ represents connections between susceptible individual j and infectious individual i in cluster k at time t; and $z_j^{(k)}$ is the covariate value of the susceptible individual j in cluster k. The spark infection parameter represents the unexplained infection in each sub-population (cluster) coming from outside clusters.

5.2.3 Likelihood Function

We consider partially observed epidemics represented by the removal times (and possibly infection times) of individuals, as well as their corresponding risk factors, in a population of size N. In general, the infection times are considered censored and the contact network are considered to be unknown. We assume the population can be partitioned into K non-overlapping clusters, and the within-cluster contact network and event times of in-

dividuals in each cluster to be conditionally independent from those in other clusters. Thus, the general likelihood can be expressed as a multiplication of K cluster-level likelihoods. Before introducing the likelihood, we introduce some notation.

We assume each cluster has a size of M_k individuals; $k = 1, \ldots, K$ and among these M_k individuals there are m_k infected individuals. The total number of infected individuals in the population is represented by $M = \sum_{k=1}^{K} m_k$, and $N = \sum_{k=1}^{K} M_k$. We label the infected individuals in cluster k by $i = 1, 2, \ldots, m_k$ with corresponding infection ($I_i^{(k)}$) and removal ($R_i^{(k)}$) times; we label the $M_k - m_k$ individuals who remain susceptible at time $T_{(obs)}$, which is the maximum observed removal time of the epidemic, in cluster k, $i = m_k + 1, m_k + 2, \ldots, M_k$ with $I_i^{(k)} = R_i^{(k)} = \infty$. We then denote infection and removal time sets for each cluster k as $\boldsymbol{I}^{(k)} = \{I_1^{(k)}, \ldots, I_{m_k}^{(k)}\}$ and $\boldsymbol{R}^{(k)} = \{R_1^{(k)}, \ldots, R_{m_k}^{(k)}\}$, respectively. Infectious periods are assumed to follow a non-negative distribution, such as an exponential or gamma distribution (Jewell et al., 2009). Network connections between individuals i and j for $i, j = 1, \ldots, M_k$, in cluster k are represented by an $M_k \times M_k$ square matrix $\mathcal{C}^{(k)}$. Thus, the global contact network matrix $\mathcal{C}$ is block-diagonal and given by:

$$\mathcal{C} = \begin{bmatrix} \mathcal{C}^{(1)} & \underline{\mathbf{0}} & \underline{\mathbf{0}} & \ldots & \underline{\mathbf{0}} \\ \underline{\mathbf{0}} & \mathcal{C}^{(2)} & \underline{\mathbf{0}} & \ldots & \underline{\mathbf{0}} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \underline{\mathbf{0}} & \underline{\mathbf{0}} & \ldots & \mathcal{C}^{(K-1)} & \underline{\mathbf{0}} \\ \underline{\mathbf{0}} & \underline{\mathbf{0}} & \ldots & \underline{\mathbf{0}} & \mathcal{C}^{(K)} \end{bmatrix}_{N \times N}$$

The likelihood is then given as:

$$L(\boldsymbol{I}, \boldsymbol{R}, \boldsymbol{C}|\boldsymbol{\theta}, \boldsymbol{\epsilon}) \;=\; \prod_{k=1}^{K} L_k(\boldsymbol{I}^{(k)}, \boldsymbol{R}^{(k)}|\boldsymbol{\theta}, \epsilon_k),$$

where L_k is the likelihood function of cluster k, defined as:

$$\begin{aligned}
L_k(\boldsymbol{I}^{(k)}, \boldsymbol{R}^{(k)}, \boldsymbol{C}^{(k)}|\boldsymbol{\theta}, \epsilon_k) &= \prod_{\substack{j=1 \\ I_j \neq I_{init}}}^{m_k} \left(\epsilon_k + \sum_{i: I_i^{(k)} < I_j^{(k)} \leq R_i^{(k)}} (\alpha_0 + \alpha_1 z_j^{(k)}) c_{ij}^{(k)} \right) \\
&\times \exp\left\{ - \int_{I_{init}}^{T_{(obs)}} \left(\sum_{i \in \mathcal{S}^{(k)}(t)} \epsilon_k + \sum_{i \in \mathcal{I}^{(k)}(t)} \sum_{j \in \mathcal{S}^{(k)}(t)} (\alpha_0 + \alpha_1 z_j^{(k)}) c_{ij}^{(k)} (u - I_i^{(k)}) \right) du \right\} \\
&\times \prod_{i=1}^{m_k} f(\gamma_i^{(k)}; \delta) \\
&= \prod_{\substack{j=1 \\ I_j \neq I_{init}}}^{m_k} \left(\epsilon_k + \sum_{i: I_i^{(k)} < I_j^{(k)} \leq R_i^{(k)}} (\alpha_0 + \alpha_1 z_j^{(k)}) c_{ij}^{(k)} \right) \\
&\times \exp\left\{ - \sum_{i=1}^{m_k} \left(\sum_{j=1}^{M_k} \left[(R_i^{(k)} \wedge I_j^{(k)}) - (I_i^{(k)} \wedge I_j^{(k)}) \right] (\alpha_0 + \alpha_1 z_j^{(k)}) c_{ij}^{(k)} \right) \right\} \\
&\times \exp\left(-\epsilon_k \sum_{i=1}^{M_k} \left[(T_{(obs)} \wedge I_i^{(k)}) - I_{init} \right] \right) \prod_{i=1}^{m_k} f(\gamma_i^{(k)}; \delta), \qquad (5.4)
\end{aligned}$$

and where: the wedge symbol $\wedge$ denotes the minimum operator, the model parameters consist of a vector $\boldsymbol{\theta}$ of the common model parameters, $\boldsymbol{\theta} = (\alpha_0, \alpha_1, \delta)$, a vector $\boldsymbol{\epsilon}$ of the corresponding spark infection term of each cluster, $\boldsymbol{\epsilon} = (\epsilon_k; k = 1, \ldots, K)$; I_{init} is the infection time of the first individual infected; $f(\bullet; \delta)$ indicates the density function of the infectious periods with unknown parameter(s) δ; and $\gamma_i^{(k)}$ is the infectious period of infected individual i in cluster k, defined as $\gamma_i^{(k)} = R_i^{(k)} - I_i^{(k)}$.

5.2.4 Epidemic Simulation

The contact network can have various structural forms that describe the connections between individuals. Here, we consider spatial contact networks, wherein the connections between individuals are based on the spatial distance between individuals. We follow the methodology developed by Bifolchi et al. (2013) and assume that the probability of a connection existing between individuals i and j follows a power law, specifically given by

$$P(c_{ij} = 1) = 1 - \exp(-\nu d_{ij}^{-\beta}), \quad \nu, \beta > 0, \tag{5.5}$$

where d_{ij} is the Euclidean distance between individuals i and j; β is the spatial parameter; and ν is the scale parameter. For simplicity, individuals are assumed to be connected by undirected edges. This assumption results in a symmetric contact matrix such that $c_{ij} = c_{ji}$ for $i \neq j$; $i, j = 1, ..., N$. Thus, the number of contact network elements is given by $\binom{N}{2}$.

The network-based continuous time ILM given by Equation 5.2, is used to generate an epidemic through a given contact network. For each infected individual, the infection time and the length of being infectious define its infection life history. We assume that infection events follow a non-homogenous Poisson process, so that the time to the next infected individual, given that the last infection occurred at time t, is assumed to follow $W_j \sim$ Exp(λ_{jt}), where W_j represents the "waiting time" for susceptible individual j becoming infectious. Given a specified infectious period distribution, an epidemic can be generated using Algorithm 5.4.

Algorithm 5.4 Epidemic simulation

1: **Initializing:** choose randomly initial infected individual $initial$.
2: Set the initial infection time $I_1 = 0$.
3: **for** $s = 2$ **to** N **do**
4: **repeat**
5: **for** all susceptible individuals at time I_{s-1} **do**
6: Compute $W_j \sim \text{Exp}(\lambda_{jI_{s-1}})$.
7: Find the susceptible individual with minimum W.
8: Assign an infection time $I_s = I_{s-1} + min(W)$.
9: Generate an infectious period $\gamma_s \sim f(\bullet; \delta)$.
10: Assign a removal time $R_s = I_s + \gamma_s$.
11: **until** no infectives remain in the population or $I_s > t_{max}$, where t_{max} is the time at where the epidemic simulation is set to end.

5.2.5 Data Clustering

To define the spatial clusters of a population of size N, individuals are partitioned into K spatially similar clusters ($1 < K < N$) using the Partitioning around Medoids (PAM) method (Kaufman and Rousseeuw, 2009). The PAM algorithm is based on a repeated partitioning of the population through a dissimilarity matrix and the updated medoids. The algorithm starts with randomly identifying initial K individuals as the medoids (centre) of clusters. Then, other individuals are associated to clusters that have minimum proximity to the medoids. The positions of the K medoids are then recalculated for each cluster and other individuals are then reassigned to clusters based on their minimum proximity to the updated medoids. The process is repeated until no changes in the medoids positions are detected. Here, proximity is based upon Euclidean distance between spatial locations, although other covariates could be incorporated. For more details about this method, readers are referred to works by Kaufman and Rousseeuw (2009) and Reynolds et al. (2006). Clus-

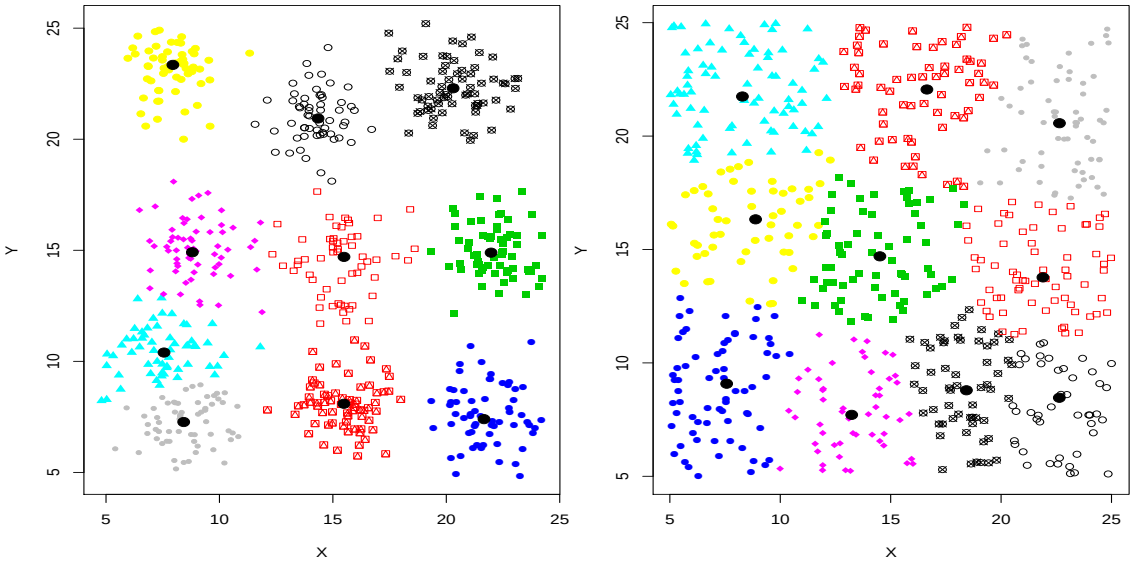


Figure 5.1: Two examples of partitioning two populations into ten clusters: left: well-clustered population, right: scattered population. Centroids of each cluster are marked by solid black dots.

tering is performed using the *pam* function from the *cluster* package in R (Maechler et al., 2017). Figure 5.1 shows a sample of clustering in two populations.

5.3 Bayesian Inference

The models are fitted within a Bayesian framework using MCMC methods, updating the model parameters and the missing data (here, infection times and contact network) via a data augmentation technique (Gamerman and Lopes, 2006). Thus, the missing data are treated as latent variables and estimated along with the model parameters.

5.3.1 The Posterior Distribution

By assigning an independent gamma prior distribution to the model parameters α_0, α_1, δ, and ϵ_k; $k = 1, \ldots, K$, the joint posterior distribution for the model parameters and full data up to the constant of proportionality is given by

$$\begin{aligned} \pi(\boldsymbol{\theta}, \boldsymbol{\epsilon} | \boldsymbol{R}, \boldsymbol{I}, \boldsymbol{C}, \boldsymbol{Z}) &\propto \prod_{k=1}^{K} \Big(L_k(\boldsymbol{R}^{(k)}, \boldsymbol{I}^{(k)}, \boldsymbol{Z}^{(k)}, \boldsymbol{C}^{(k)} | \boldsymbol{\theta}, \epsilon_k)\, p(\boldsymbol{\phi^{(k)}} | \boldsymbol{C}^{(k)})\, \pi(\epsilon_k) \Big) \\ &\times \pi(\alpha_0)\, \pi(\alpha_1)\, \pi(\delta), \end{aligned} \tag{5.6}$$

where $p(\boldsymbol{\phi}^{(k)} | \boldsymbol{C}^{(k)})$ is an observation model that represents global information about the cluster constrained contact networks. Here, we assume a distribution on the total number of connections within the sub-contact network $\phi_{tot}^{(k)}$, where $\phi_{tot}^{(k)} = \sum_{i=1}^{M_k} \sum_{j=1}^{i} c_{ij}^{(k)}$. Thus, the observation model of Equation 5.6 becomes $p(\boldsymbol{\phi}^{(k)} | \boldsymbol{C}^{(k)}) = \pi(\phi_{tot}^{(k)})$. Here, we assume that $\pi(\phi_{tot}^{(k)})$ is the density of a normal distribution with mean equal to the true total number of connections in the sub-contact network, and with the variance equal to 10.

5.3.2 Modified Likelihood Inflating Sampling Algorithm (LISA)

Entezari et al. (2017) have recently introduced the likelihood inflating sampling algorithm (LISA) for tackling the computational issues of analyzing large data sets using Bayesian MCMC. The algorithm considers dividing a data set into K subsets and calculating the likelihood for each independently. The sub-likelihood functions of the K samples

are then "inflated", incorporating importance sampling weights for each. We first introduce their approach in the context of the within-cluster network-based ILMs, by modifying Equation 5.3, so that the infectivity rate of a susceptible individual j in sample (cluster) k at time t is then given by:

$$\lambda_{jkt} = \left[\alpha_0^{(k)} + \alpha_1^{(k)} z_j^{(k)} \sum_{i \in \mathcal{I}^{(k)}(t)} c_{ij}^{(k)} \right] + \epsilon_k. \tag{5.7}$$

The likelihood function given by Equation 5.4 is then modified so that the likelihood function of the k^{th} cluster (sample) has its own vector $\boldsymbol{\theta}^{(k)}$ of the model parameters, given by:

$$\begin{aligned} L_k(\boldsymbol{I}^{(k)}, \boldsymbol{R}^{(k)}, \boldsymbol{C}^{(k)} | \boldsymbol{\theta^{(k)}}, \epsilon_k) &= \prod_{\substack{j=1 \\ I_j \neq I_{init}}}^{m_k} \left(\epsilon_k + \sum_{i: I_i^{(k)} < I_j^{(k)} \leq R_i^{(k)}} (\alpha_0^{(k)} + \alpha_1^{(k)} z_j^{(k)}) c_{ij}^{(k)} \right) \\ &\times \exp\left\{ -\sum_{i=1}^{m_k} \left(\sum_{j=1}^{M_k} \left[(R_i^{(k)} \wedge I_j^{(k)}) - (I_i^{(k)} \wedge I_j^{(k)}) \right] (\alpha_0^{(k)} + \alpha_1^{(k)} z_j^{(k)}) c_{ij}^{(k)} \right) \right\} \\ &\times \exp\left(-\epsilon_k \sum_{i=1}^{M_k} \left[(T_{(obs)} \wedge I_i^{(k)}) - I_{init} \right] \right) \prod_{i=1}^{m_k} f(\gamma_i^{(k)}; \delta^{(k)}), \end{aligned} \tag{5.8}$$

where $\boldsymbol{\theta^{(k)}} = (\alpha_0^{(k)}, \alpha_1^{(k)}, \delta^{(k)})$, and each cluster is processed independently (generally, each is processed by a separate CPU); and I_{init} is the infection time of the first individual infected in the population. Thus, the full posterior distribution is divided into sub-posterior distributions, where the k^{th} sub-posterior is defined as:

$$\begin{aligned}\pi_k(\boldsymbol{\theta}^{(k)}, \epsilon_k | \boldsymbol{R}^{(k)}, \boldsymbol{I}^{(k)}, \boldsymbol{C}^{(k)}, \boldsymbol{Z}^{(k)}) &\propto \Big(L_k(\boldsymbol{R}^{(k)}, \boldsymbol{I}^{(k)}, \boldsymbol{Z}^{(k)}, \boldsymbol{C}^{(k)} | \boldsymbol{\theta}^{(k)}, \epsilon_k)\Big)^K \\ &\times \; p(\boldsymbol{\phi^{(k)}} | \boldsymbol{C}^{(k)})\, \pi(\epsilon_k)\, \pi(\boldsymbol{\theta}^{(k)}). \end{aligned} \tag{5.9}$$

The reason for inflating the likelihood function by a factor of K is to make the sub-posteriors a closer representation of the full posterior distribution. After processing each cluster, all sub-posterior samples are combined using importance sampling, weighting the draws of all sub-posteriors to obtain approximate full posterior draws as follows:

$$\theta_s = \left(\sum_k \omega_k\right)^{-1} \sum_k \omega_k \theta_s^{(k)}, \tag{5.10}$$

where the weights $\omega_k = \Sigma^{-1}$ are the covariance matrix of each sub-posterior, such that $\Sigma = Var(\boldsymbol{\theta} | \boldsymbol{R}^{(k)}, \boldsymbol{I}^{(k)}, \boldsymbol{C}^{(k)}, \boldsymbol{Z}^{(k)})$.

For our problem, we modify Equation 5.9 further so that the sub-likelihood function is raised to the power of the proportion of infected individuals at $T_{(obs)}$, p_k, in cluster k. This modification will give clusters with larger p_k more influence in the hope of obtaining closer representation of the full data posterior given by Equation 5.6. Thus, the modified sub-posterior of Equation 5.9 is given by:

$$\begin{aligned}\pi_k(\boldsymbol{\theta}^{(k)}, \epsilon_k | \boldsymbol{R}^{(k)}, \boldsymbol{I}^{(k)}, \boldsymbol{C}^{(k)}, \boldsymbol{Z}^{(k)}) &\propto \Big(L_k(\boldsymbol{R}^{(k)}, \boldsymbol{I}^{(k)}, \boldsymbol{Z}^{(k)}, \boldsymbol{C}^{(k)} | \boldsymbol{\theta}^{(k)}, \epsilon_k)\Big)^{p_k} \\ &\times p(\boldsymbol{\phi^{(k)}} | \boldsymbol{C}^{(k)})\, \pi(\epsilon_k)\, \pi(\boldsymbol{\theta}^{(k)}). \end{aligned} \tag{5.11}$$

5.3.3 MCMC Method

Our approach uses a simple random walk Metropolis (within Gibbs) algorithm. None of the parameters have standard conditional distributions, except for the infectious period rate δ which follows a gamma distribution:

$$\delta | \alpha_0, \alpha_1, \boldsymbol{\epsilon}, \boldsymbol{R}, \boldsymbol{I}, \boldsymbol{C}, \boldsymbol{Z} \sim \Gamma(M + a_\delta, Q + b_\delta),$$

where $Q = \sum_{k=1}^{K} \sum_{i=1}^{m_k} (R_i^{(k)} - I_i^{(k)})$ and a_δ and b_δ are the parameters of its prior distribution.

In the case of applying LISA, each cluster (sub-population) has its own parameters that leads to the same approach as above. The infectious period rate $\delta^{(k)}$ has a standard conditional distribution that follows a gamma distribution:

$$\delta^{(k)} | \alpha_0^{(k)}, \alpha_1^{(k)}, \epsilon_k, \boldsymbol{R}^{(k)}, \boldsymbol{I}^{(k)}, \boldsymbol{C}^{(k)}, \boldsymbol{Z}^{(k)} \sim \Gamma(M_k + a_{\delta^{(k)}}, Q^{(k)} + b_{\delta^{(k)}}),$$

where $Q^{(k)} = \sum_{i=1}^{m_k} (R_i^{(k)} - I_i^{(k)})$ and $a_{\delta^{(k)}}$ and $b_{\delta^{(k)}}$ are the parameters of its prior distribu-

tion.

A Gibbs update (i.e., sampling from the conditional posterior distribution) is used for δ and $\delta^{(k)}$. Parameters α_0, α_1 (in LISA $\alpha_0^{(k)}$, $\alpha_1^{(k)}$), and $\boldsymbol{\epsilon}$ are updated in turn using univariate normal random-walk updates, tuned to achieve acceptable mixing properties.

Under both approaches, the missing information (infection times (I_i) or, equivalently, infectious periods (γ_i), and connections of the contact network (c_{ij})) are treated as latent variables and updated using Bayesian data augmentation. The infection times and infectious periods are updated using an independence sampler. Specifically, each candidate infectious period $\gamma_i^{(k)}$ is drawn from its infectious period distribution. This (equivalently) results in a new candidate infection time $I_i^{(k)}$ from the difference between the fixed removal time $R_i^{(k)}$ and the proposed $\gamma_i^{(k)}$ of the i^{th} individual in cluster k. Each infection time/infectious period is updated in turn.

In cases where the contact network is considered to be unknown, the elements of the contact network ($c_{ij}^{(k)}$) are also updated in turn. As $c_{ij}^{(k)}$ has only two states, 0 (no connection) or 1 (connection), these parameters are updated by proposing a move from the current state to the only other possible state with probability 1. The update of $c_{ij}^{(k)}$ is based on two parts: the likelihood function and the density of the observation model. Under the observation model based on the total number of connections, the conditional distribution becomes:

$$c_{ij}^{(k)}|\boldsymbol{\theta}, \epsilon_k, \boldsymbol{R}^{(k)}, \boldsymbol{I}^{(k)}, \boldsymbol{C}_{-ij}^{(k)}, \boldsymbol{Z}^{(k)} \sim L_k(\boldsymbol{R}^{(k)}, \boldsymbol{I}^{(k)}, \boldsymbol{Z}^{(k)}, \boldsymbol{C}^{(k)}|\boldsymbol{\theta}, \epsilon_k)\, \pi(\phi_{tot}^{(k)}).$$

5.3.4 Simulated Epidemics

We considered three scenarios for epidemics simulated using the simple network-based ILM given in Equation 5.2. In the first scenario, we assumed that a population has $K = 10$ well-defined spatial clusters (scenario 1: well-clustered populations). The true clusters are then used to define the K clusters used in the analyses. Two other scenarios considered a more homogeneous population structure, where there is no clear clustering and individuals are uniformly distributed (Scenario 2a and 2b: scattered populations). In these cases, the population is partitioned into $K = 10$ clusters using the PAM method prior to analysis.

For well-clustered populations (Scenario 1), we simulated ten epidemics for populations with $N = 625$ individuals distributed within a square area of 20×20 units. The ten spatial clusters were formed by generating the individual locations of each cluster randomly from a multivariate normal distribution with means equal to the XY coordinates of the centre of each cluster, see Figure 5.1 for an example of this type of population. These centres were chosen to formulate separate clusters in the population. We first generated the individual locations uniformly (i.e., $X, Y \sim U(0, 20)$), and then applied the PAM method to obtain the medoids (centres) of the ten clusters. We then regenerated the individual locations of each cluster from a multivariate normal distribution with means equal to the medoids and a covariance matrices specified so as to allow both separation between clusters and variation within clusters. Epidemics were then simulated through a contact network that was generated using Equation 5.5 with the values $\beta = 4$ and $\alpha = 2$. These values were chosen since they tended to produce networks with few connections between individuals of

different clusters, along with a large number of local connections within clusters. With a discrete uniform covariate, $z_j \sim \mathcal{DU}(0, 50)$, epidemics were then generated from the simple network-based ILM with parameter values $\alpha_0 = 0.02$, $\alpha_1 = 0.0005$, and $\epsilon = 0$. The infectious periods followed gamma distributions, such that $\gamma_j \sim \Gamma(9, \delta = 3), \forall j$.

In the scattered populations (Scenario 2), we consider the $N = 625$ individuals to be uniformly distributed without any forced clustering in a square area of 20×20 units. In Scenario 2a, a binary covariate (z_j) was included in the susceptibility function of the simple network-based ILM. This binary covariate can be thought to represent the vaccination status of individuals in a real epidemic, for example. Specifically, $z_j \sim Bern(0.5)$ was used for Scenario 2a[2].

We also consider integer values generated from a discrete uniform distribution such that $z_j \sim \mathcal{DU}(0, 50)$ for Scenario 2b. We simulated ten epidemics for these scenarios through contact networks that were generated using the parameter values $\alpha = 1$ and $\beta = 4$. For Scenario 2a, we generated epidemics with the following model parameter values; $\alpha_0 = 0.02$, $\alpha_1 = 0.05$, and $\epsilon = 0$. The infectious periods followed gamma distributions such that $\gamma_j \sim \Gamma(10, \delta = 1)$. For Scenario 2b, we used $\alpha_0 = 0.05$, $\alpha_1 = 0.005$, and $\epsilon = 0$, and the infectious periods followed gamma distributions such that $\gamma_j \sim \Gamma(9, \delta = 3)$.

All epidemics and contact networks were simulated using the functions *datagen* and *contactnet*, respectively, from the *EpiILMCT* package in R (Almutiry et al., 2018).

Model fitting for the above epidemic data sets was carried out using four different anal-

[2] $Bern$ is used to denote Bernoulli distribution

yses. First, a gold-standard fully-connected contact network analysis was performed using the network-based ILM given by Equation 5.2. Here, we assume that the event times and contact network are fully and correctly observed. This analysis is referred to as the *standard analysis*. Under the remaining analyses, we constrained the contact network to have connections only within clusters, leading to a K sub-contact network. Then, cluster-constrained contact network (CCCN) analyses were performed using the within-cluster network-based ILM given by Equation 5.3 under three different data uncertainty assumptions. The first CCCN analysis (CCCN1) assumed that epidemic data are fully and correctly observed. We then increased our uncertainty by assuming unknown infection times (and, equivalently, infectious periods) in the second CCCN analysis (CCCN2). Finally, our uncertainty is further increased by considering the contact network, as well as the infection times, to be unobserved (CCCN3). The modified LISA algorithm was also applied with the above CCCN analysis scenarios, and we call them "*CCCN-LISA*" analyses. Note, in all scenarios, we assume the removal times are known.

For all analyses, independent exponential prior distributions with rate 0.01 were assigned to model parameters α_0, α_1 ($\alpha_0^{(k)}$, $\alpha_1^{(k)}$ in the case of applying LISA), and $\epsilon_k, k = 1, \ldots, K$. For the infectious period rate δ, we assigned gamma prior distributions as follows: $\delta \sim \Gamma(2, 2)$ for infectious periods following as $\Gamma(10, \delta)$, and $\delta \sim \Gamma(6, 2)$ for infectious periods following as $\Gamma(9, \delta)$. The DA-MCMC was then run for 50,000 iterations with a burn-in period of 1,000 iterations.

5.4 Results

5.4.1 Well-Clustered Populations (Scenario 1)

We examine the fit of the within-cluster network-based continuous time ILMs on simulated epidemics generated over well-clustered populations. Figures 5.2 and 5.3 show the posterior means and 95% percentile credible intervals for the ILM parameters under the standard and CCCN analyses of the well-clustered populations. The estimates of the model parameters under the CCCN1 analysis (fully observed epidemics) are close to those under the standard analysis.

Incorporating the uncertainty of the infection times (infectious periods) with/without the contact network uncertainty into the model also results in good estimates of the model parameters α_0 and α_1 (Figure 5.2). The posterior estimates of the infectious period rate δ tended to be higher in the analysis incorporating infection time and contact network uncertainty (CCCN3) than solely infection times uncertainty (CCCN2). However, the overall quality of δ estimation was similar under CCCN2 and CCCN3, and the differences did not have an obvious effect on the estimation of the infection times themselves, which were well approximated (see Appendix C.2).

The posterior estimates of the spark infection parameters are shown in Figure 5.3. We found that the magnitude of the spark infection parameter estimates $\boldsymbol{\epsilon}$ appeared to be related to three factors. First, a larger proportion of individuals infected in a cluster appears to cause the estimate of ϵ_k to be larger, especially when the number of discarded connections

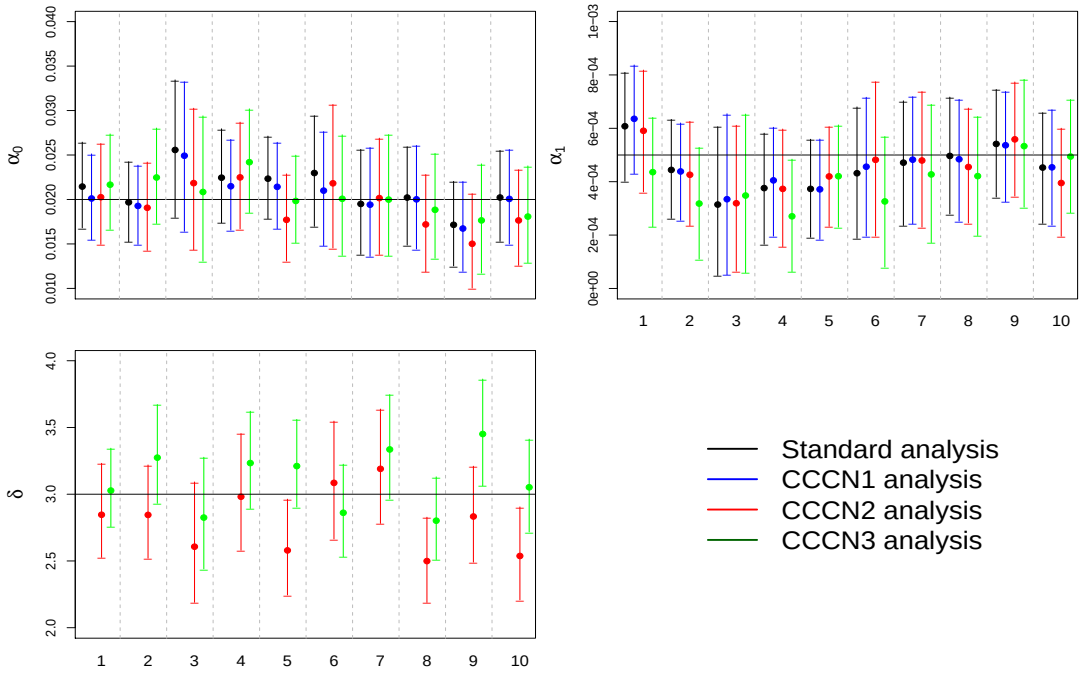


Figure 5.2: Posterior means and 95% credible intervals of the model parameters α_0, α_1, and δ for the simple within-cluster network-based ILMs of the well-clustered populations based on the full Bayesian MCMC analyses. Horizontal black lines represent the true values as: $\alpha_0 = 0.02$, $\alpha_1 = 0.0005$, and $\delta = 3$.

is large (see Appendix C.1 for summary of the cluster data). Second, the magnitude of the spark infection parameter of the cluster that has the first infected individual was found to be larger. This might be because as the population is highly clustered, the epidemics spread mainly locally in the initial cluster before spreading globally (between clusters). Third, clusters without any infected individuals demonstrate very small estimates of their spark infection parameters.

We also consider the posterior degree distribution of the contact network. The average of the posterior degree distribution shows close approximation to the true average degree,

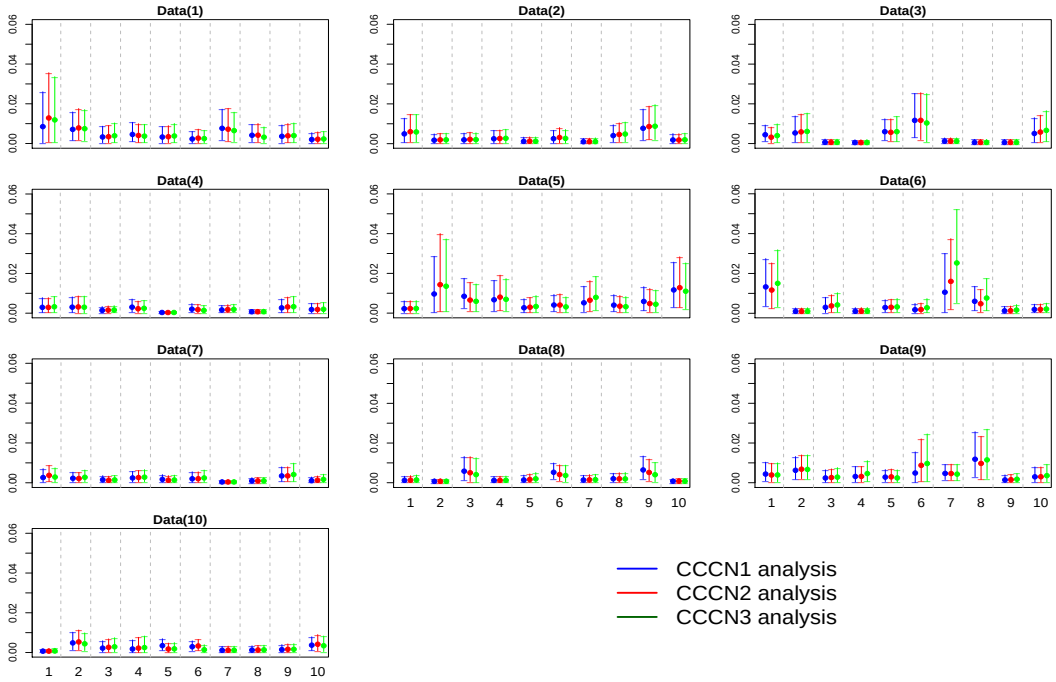


Figure 5.3: Posterior means and 95% credible intervals of the spark infection parameters ϵ_k, $k = 1, \ldots, K$, for the simple within-cluster network-based ILMs of the well-clustered populations based on the full Bayesian MCMC analyses.

as shown in Appendix C.2.

Overall, the cluster-constrained contact network analyses demonstrate good performance in analyzing epidemics over well-clustered populations. The missing information between clusters appears to be well captured by the spark infection parameters. This seems sensible since the spread of epidemics spread mainly locally within the spatial clusters with few connections between clusters. Thus, the incorporation of contact network uncertainty appears to have a relatively small effect on the estimates of the model parameters.

5.4.2 Scattered Populations (Scenarios 2 and 3)

We now examine the fit of the within-cluster model epidemics which were generated over scattered populations. In this case, populations have no obvious clustered structure and individuals are distributed uniformly in a square area. The missing between-cluster connections of the contact network between individuals from different clusters are, thus, expected to have a larger effect than those in the well-clustered population. Figures 5.4-5.7 show the posterior means and 95% credible intervals of the model parameters for the standard and CCCN analyses.

In the case where epidemics are assumed to be fully observed, we observe very similar estimates of the model parameters α_0 and α_1 under both standard and cluster-constrained contact network (CCCN1) analyses. This suggests that fitting epidemics with the within-cluster network-based continuous time ILMs can result in good performance, even though the population is not naturally clustered.

Incorporating the infection times and contact network uncertainties into the model (CCCN3) appears to exert a larger effect on the estimate of the model parameters than incorporating the uncertainty of infection times alone (CCCN2). The CCCN2 analysis tended to produce estimates of the model parameters closer to the standard and CCCN1 analyses. Under the CCCN3 analysis, notably poorer estimates of the covariate parameter α_1 were observed.

The estimates of the spark infection parameters ϵ_k were of larger magnitude when com-

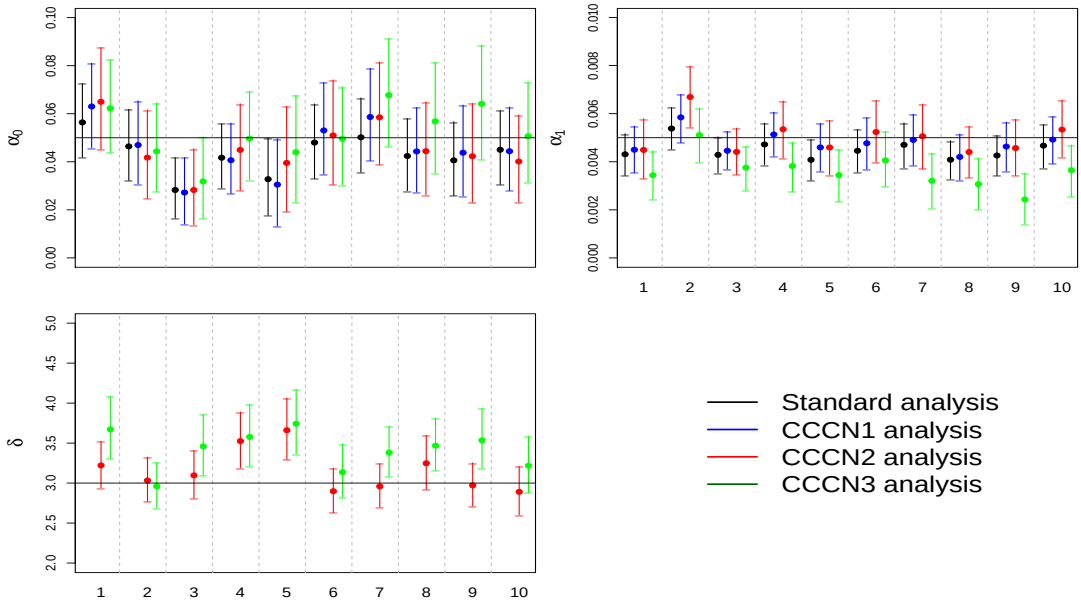


Figure 5.4: Posterior means and 95% credible intervals of the model parameters α_0, α_1, and δ for the simple within-cluster network-based ILMs of the scattered populations (scenario 2.b: with discrete covariate) based on the full Bayesian MCMC analyses. Horizontal black lines represent the true values as: $\alpha_0 = 0.05$, $\alpha_1 = 0.005$, and $\delta = 3$.

pared to those seen in the well-clustered population epidemics. Also, similar to the analyses of the well-clustered populations, their estimates show a clear relationship to the number of infected individuals in clusters. Further, the magnitude of ϵ_k appears to be positively correlated to the number of discarded connections of the contact network (see Appendix C.1). On the other hand, clusters with a small proportion of infected individuals tended to have smaller ϵ_k estimates.

The posterior estimates of the infection times and degree distribution of the contact networks appeared to be reasonably good through all of the CCCN analyses. These estimates are shown in Appendix C.2.

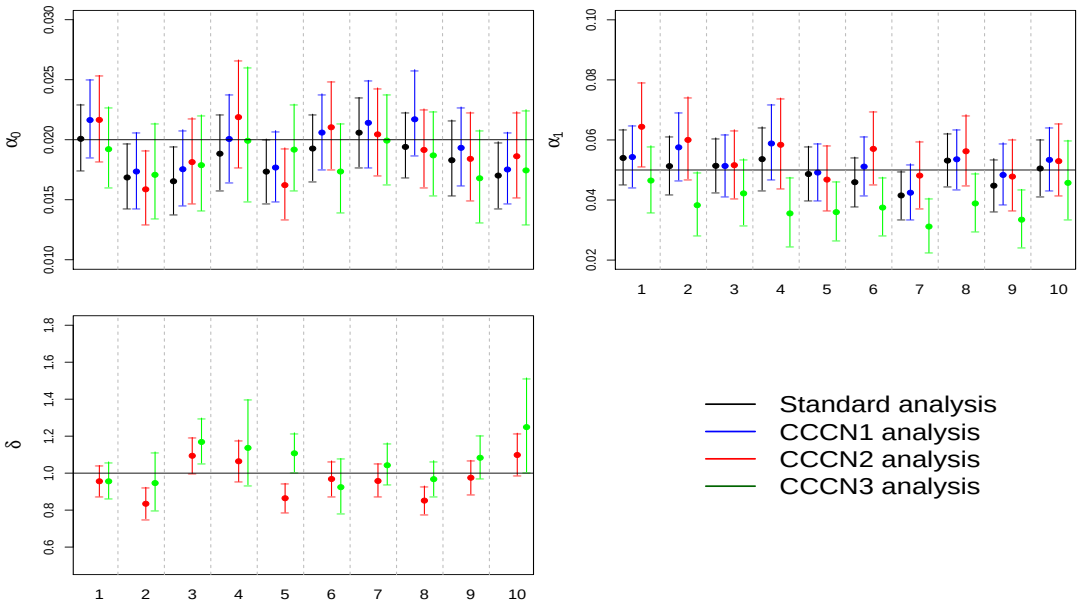


Figure 5.5: Posterior means and 95% credible intervals of the model parameters α_0, α_1, and δ for the simple within-cluster network-based ILMs of the scattered populations (scenario 2a: with binary covariate) based on the full Bayesian MCMC analyses. Horizontal black lines represent the true values as: $\alpha_0 = 0.02$, $\alpha_1 = 0.05$, and $\delta = 1$.

These results suggest that using the within-cluster ILM for analyzing epidemics that spread over both clustered and non-clustered populations show promise. Although we observed underestimation of the covariate effect parameter under the incorporation of contact network uncertainty analysis (CCCN3), the infection spark parameters appeared to allow for between clusters infections thereby producing reasonably good estimates of the infection times and degree distribution of the contact networks.

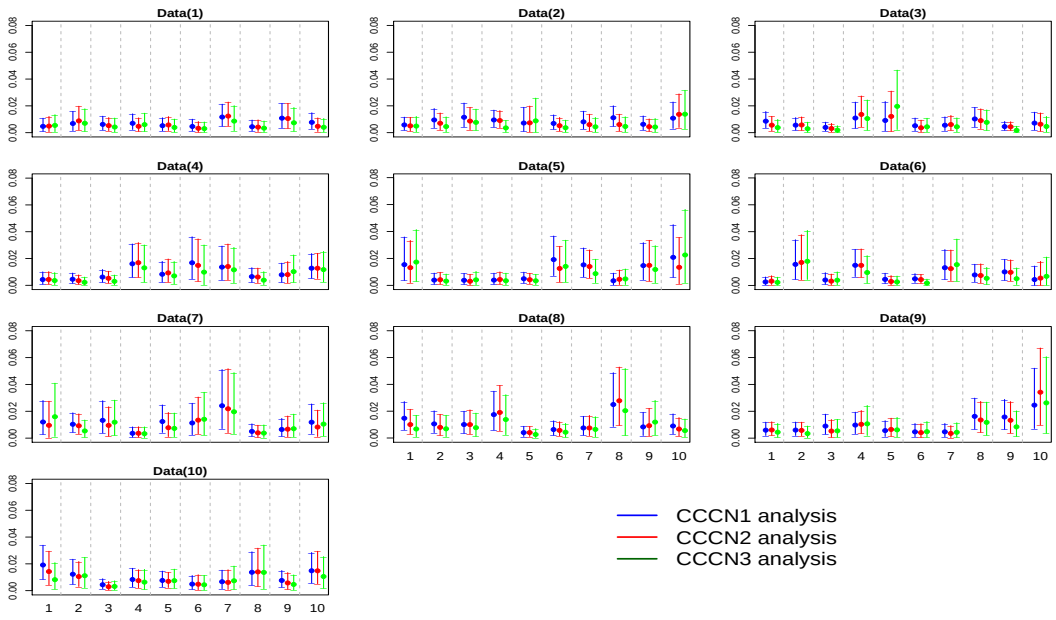


Figure 5.6: Posterior means and 95% credible intervals of the spark infection parameters ϵ_k, $k = 1, \ldots, K$, for the simple within-cluster network-based ILMs of the scattered populations (scenario 2b: with discrete covariate) based on the full Bayesian MCMC analyses.

5.4.3 Comparison to LISA

We now compare our within-cluster network-based ILM approach (given by Equation 5.3) to a modified version of the LISA approach of using the modified within-cluster network-based ILM (given by Equation 5.7) for analyzing epidemic data sets. Our modification of the LISA algorithm involves raising the likelihood to a power given by the proportion of infected individuals in each cluster (sample) rather than K. This was found to result in improved performance of mixing in the MCMC chains as well as the estimation of the model parameters, whereas applying the original algorithm leads to production of very poor MCMC mixing chains, as well as very narrow 95% credible intervals of the model parameters (see Appendix C.3 for more details).

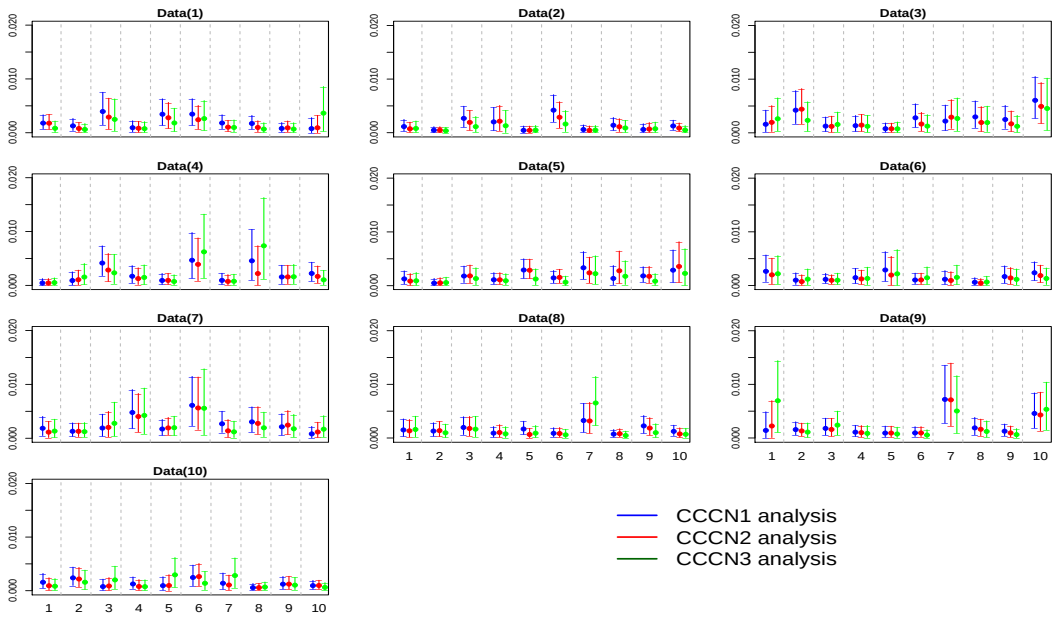


Figure 5.7: Posterior means and 95% credible intervals of the spark infection parameters ϵ_k, $k = 1, \ldots, K$, for the simple within-cluster network-based ILMs of the scattered populations (scenario 2a: with binary covariate) based on the full Bayesian MCMC analyses.

Here, we show the comparisons by performing the standard and CCCN analyses using our approach and the modified LISA on four epidemic data sets from Scenario 2a. Figures 5.8 and 5.9 show the posterior means and 95% credible intervals of both approaches under the four different analyses. In general, both approaches show similar estimates of the model parameters under the fully-observed epidemic analysis. These estimates were also approximately similar to those obtained under the standard analysis. However, with the increase of uncertainty, infection times and contact network, significant underestimation of the model parameters is observed, especially under the unknown contact network analyses. These underestimations are accompanied with larger estimates of the spark infection parameters. Even though applying the LISA algorithm results in reduced computation time

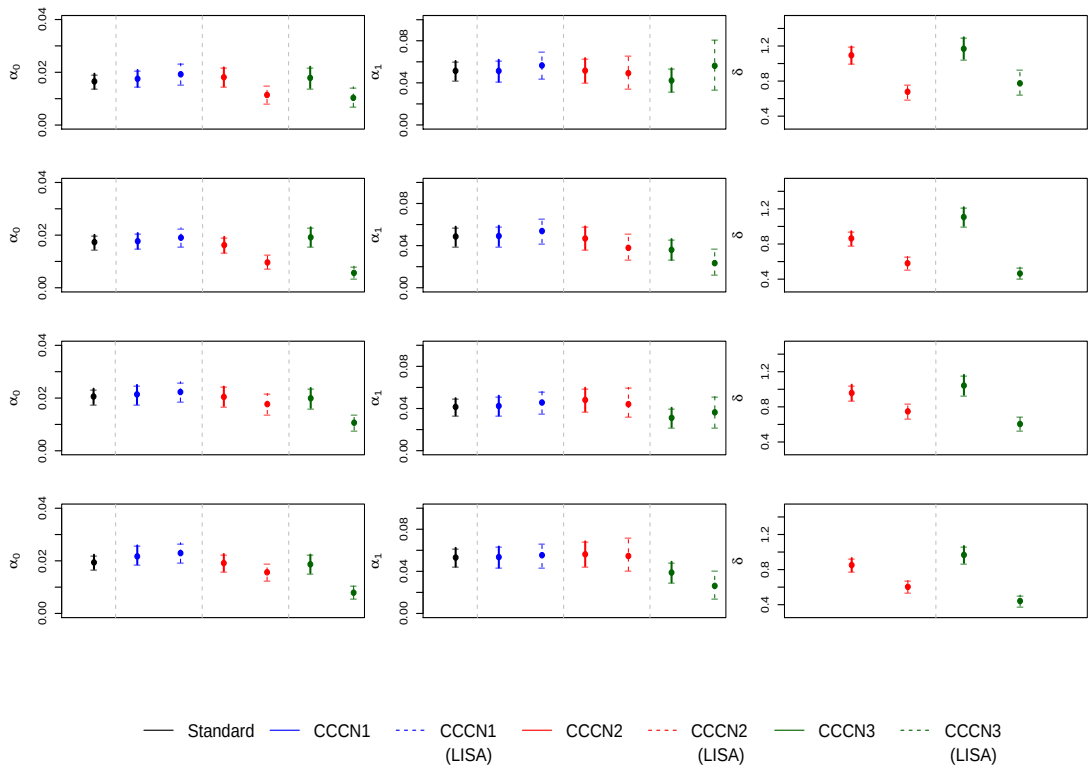


Figure 5.8: Posterior means and 95% credible intervals of the model parameters α_0, α_1, and δ for the standard, CCCN and CCCN-LISA analyses of four epidemic data sets of scattered populations (Scenario 2a).

(discussed in the next Section), our approach is more efficient in producing closer approximations of the model parameters, especially in the case of having significant amounts of missing information.

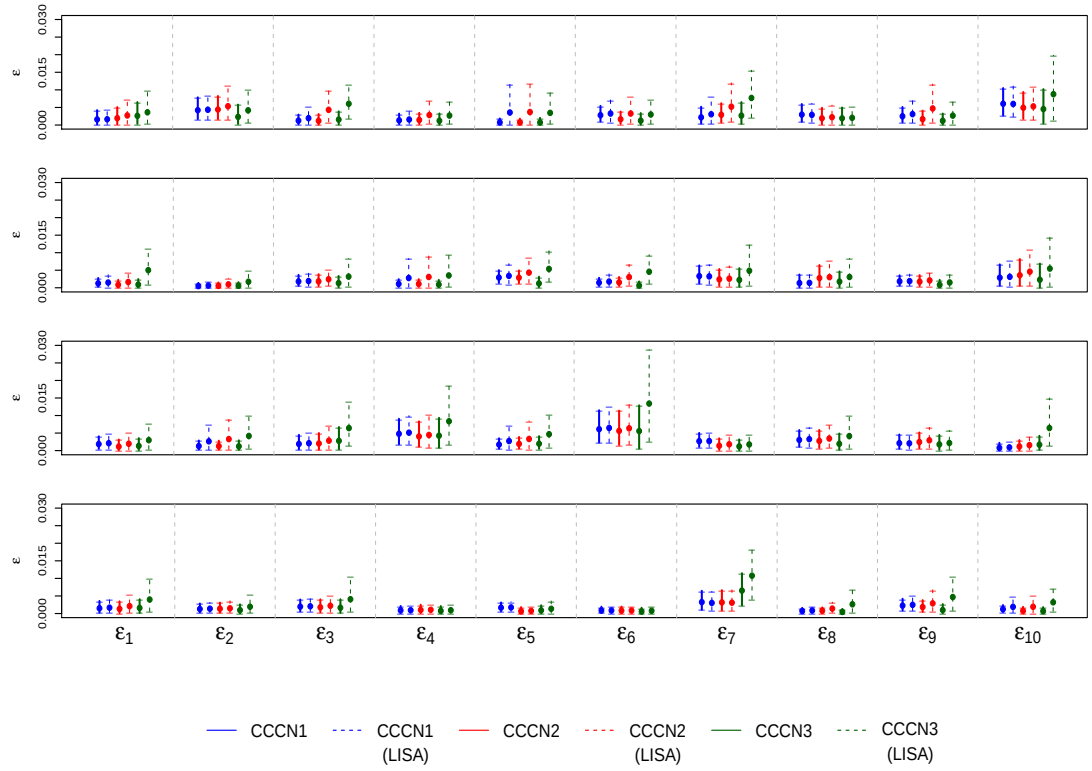


Figure 5.9: Posterior means and 95% credible intervals of the spark infection parameters ϵ_k, $k = 1, \ldots, K$, for the CCCN and CCCN-LISA analyses of four epidemic data sets of scattered populations (Scenario 2a).

5.4.4 The Effect of Increasing Number of Clusters

The above analyses are based on partitioning populations into 10 spatial clusters. We illustrate here the effect of increasing the number of clusters on the estimates of the model parameters and the corresponding computation time. Simulated epidemics of scattered populations (Scenario 2a) were fitted using the within-cluster model given by Equation 5.3 by partitioning populations into 10, 15, and 40 clusters.

Figure 5.10 shows the posterior means and 95% credible intervals of the model pa-

rameters under the CCCN analyses. The most notable change regarding the estimates of the model parameters was seen for the covariate effect α_1. Under all CCCN analyses, we noticed that as the number of clusters increases, the posterior means of α_1 also tended to increase along with larger credible intervals. This is a result that is difficult to explain.

However, with the smaller number of clusters, the posterior means of α_1 agreed with those under the standard analysis that considered the whole population (i.e., without clustering). The estimation of the other main model parameters did not seem to be particularly affected by an increase in the number of clusters. The inclusion of contact network and infection times uncertainty has only a small effect on the posterior estimates.

The estimates of the spark infection parameters was affected by the increase in the number of clusters, however. The estimates of ϵ for multiple data sets with 15 and 40 clusters are shown in Appendix C.4, under the three CCCN analyses. Results show fluctuation of estimates of the spark infection parameters (i.e., going up or down), as might be expected with less information about each cluster, and more heterogeneity between clusters.

Figure 5.11 shows the computation time of analyzing epidemic data using 10, 15, 20, and 40 clusters with the MCMC method. This indicates that the computation time is decreased with an increasing of the number of clusters.

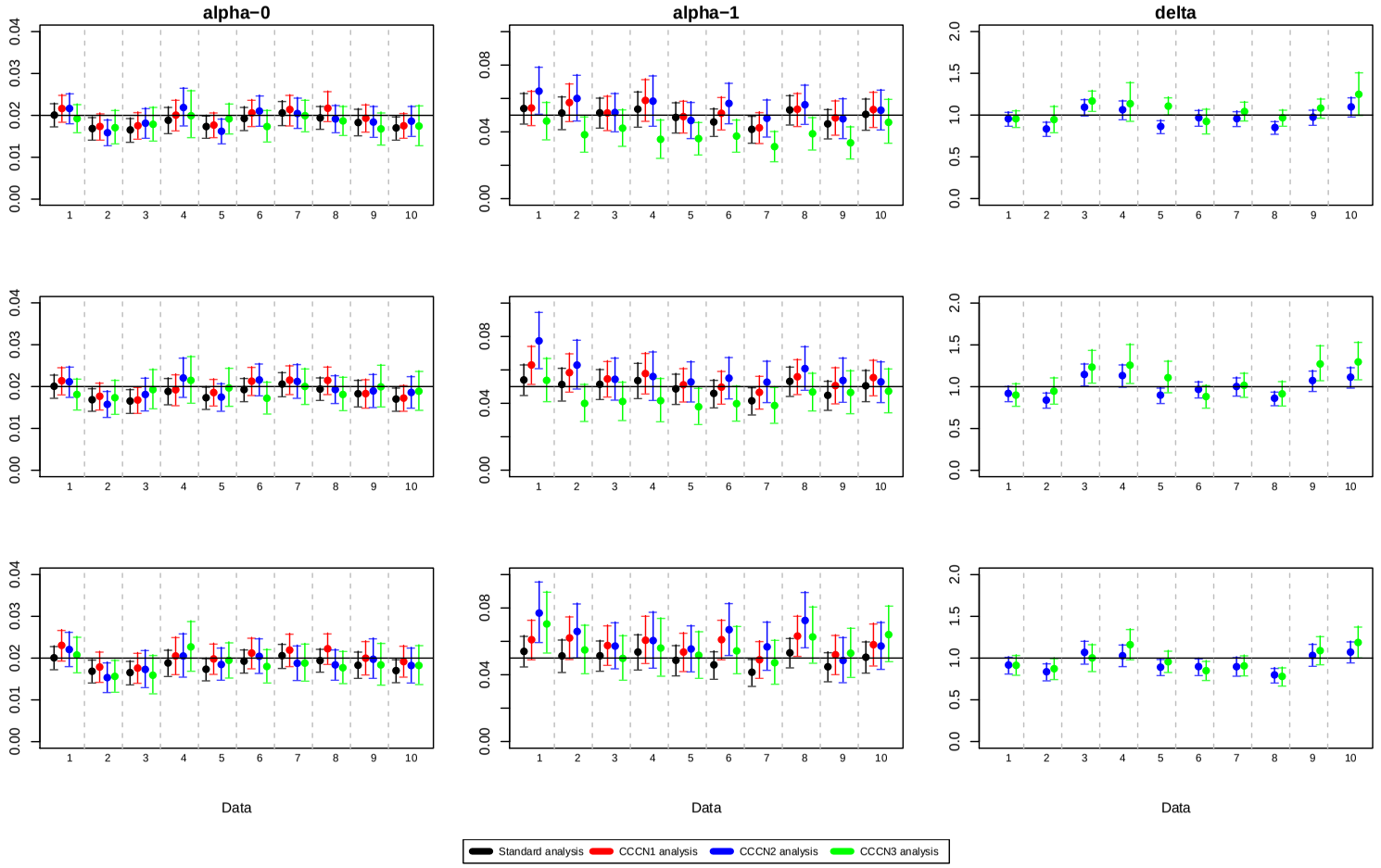


Figure 5.10: Posterior means and 95% credible intervals of the model parameters α_0, α_1, and δ for the standard and CCCN analyses of epidemic data sets of scattered populations (scenario 2a) under three different number of clusters. First row: 10 clusters, second row: 15 clusters, and third row: 40 clusters.

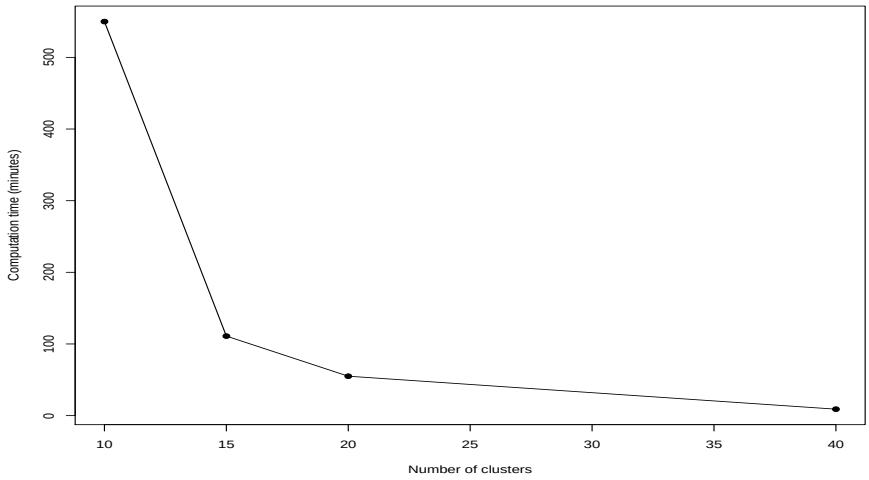


Figure 5.11: The computation time (in minutes) of fitting the within-cluster network-based continuous time ILMs to an epidemic data set with different numbers of clusters (10, 15, 20, and 40) using MCMC.

5.4.5 Computation Time and Parallelization

By partitioning the population into K clusters, we have gained a large reduction in the computation time of analyzing network-based epidemics. The key component of the analysis, which is imposing sparsity on the contact network, and forcing conditional independence between clusters. The contact network considered in this case then consists of local contact network cliques of each cluster, which allows us to discard a large number of connections. These discarded connections now no longer need to be accounted for, or inferred under network uncertainty. Also, we have set all of the connections between individuals who remain susceptible at the end of the observed epidemic to zero, as these connections are not involved in the calculation of the likelihood function (Equation 5.4). This too helps in reducing computation time. However, analyzing the global network (without imposing clusters), especially if assuming completely a unknown contact network, easily leads to a

serious computational bottleneck.

The assumed independence between clusters allows for parallelization to be used in the updating scheme of the MCMC. The parallelization was achieved using a shared-memory architecture with an implementation of the OpenMP standard (Dagum and Menon, 1998). However, the use of the parallel algorithm in our scenarios here does not lead to significant differences in the computation time when compared to serial computations (without parallelization). Running the MCMC for 50,000 iterations to fit the within-cluster network-based continuous time ILM to unknown contact network epidemic data sets took, on average, 550 minutes (9.17 hours) and 456 minutes (7.6 hours) using serial and parallel algorithms, respectively. The lack of improved performance with the parallel algorithm is seriously affected by the communication time between processors, and is thus not particularly efficient. The time consuming nature of this type of analysis when using larger population sizes is a potential limitation of this work, a point discussed in more detail in the next section. However, performing the MCMC for the whole population (without imposing clustering) when accounting for unknown contact networks and infection times is extremely computationally expensive ($\approx$ 15 minutes for one iteration). Therefore, we stress the serious computation time savings that arise from the clustering itself (which reduces the dimensionality of the posterior by imposing sparsity on the contact network) and not parallelization.

When employing the LISA approach, however, the use of a parallel algorithm becomes much more beneficial since each sample (cluster) is analyzed independently from other

samples, which allows each iteration of the MCMC to be updated in parallel without the need for between processor communications. Thus, a much faster computation is observed with the use of the parallel algorithm than with the serial algorithm. Running the MCMC using the LISA algorithm with a parallel algorithm took on average 80 minutes (1.33 hours).

All of the parallel computations were run on multiprocessor servers with 24 CPU cores at 2.5 GHz and 256 GB of RAM for each server. Other computations were performed on an Apple iMac with i5-core Intel 2.9 GHz processors with 16 GB of RAM without parallelization. All coding was carried out in Fortran 95, with MCMC output being analyzed using R programming.

5.5 Discussion

The main concern of this chapter was to facilitate the analysis of epidemic data sets that contain large amounts of missing information. The computational burden of modelling epidemics in the presence of this missing information naively can be extremely prohibitive. Thus, we introduced a within-cluster constrained version of the network-based continuous time ILM to allow for approximate, but computationally tractable, inference. We examine the fit of this new model by studying simulated epidemics over two population structures: well-clustered and scattered (unclustered) populations. Results show better fit of the model under well-clustered populations compared to that of scattered populations, mainly because

the effect of imposing isolation between different parts (clusters) of the network is relatively low when the true network already exhibits strong clustering. However, overall it was found that model parameters could be estimated reasonably well, even in the presence of infection time and contact network uncertainty, under all scenarios.

We further examined the performance of the within-cluster network-based continuous time ILMs in scattered populations with different numbers of imposed clusters. A great reduction in computation time was observed with an increase in the number of clusters, especially under the CCCN3 analysis. Since the cluster sizes are substantially decreased with an increase in the number of clusters, there is a large reduction in the number of the contact network elements that need to be estimated. However, increasing the number of clusters can lead to a deterioration in inferential performance. Therefore, a further examination on how to choose the optimal number of clusters, so as to reduce computation time while maintaining inferential performance, is needed.

The implementation of our modified LISA approach produced similar results to the fully-observed epidemic CCCN analysis. Although using LISA under the analyses incorporating data uncertainty (CCCN2 and CCCN3) showed a great reduction in computation time compared to the within-cluster ILM, we observed better performance of the within-cluster ILM with respect to parameter estimation. Data partitions in the LISA approach are assumed to be independently and identically distributed (Entezari et al., 2017), whereas clusters are assumed to be *conditionally* independent to each other in the within-cluster ILM. That is, parameters are forced to be same in each cluster under the within-cluster

network-based continuous time ILMs analyses, but estimated independently (and then re-combined) under LISA.

Partitioning a population into a number of clusters can be done in various ways. Here, we applied the PAM method of partitioning a population into a given number of spatial clusters. Other partitioning methods, which may produce better estimates, could be applied. If an epidemic is only assumed to spread across the edges of a contact network (without spatial terms involved in the infectivity rate), partitioning the population based on a certain property of the contact network itself may be more beneficial than using spatial information alone, especially if the contact network is not spatial in nature. However, of course, this would only work if the network is (partially) observed. If not, clustering via other covariates correlated with the network could be used.

We restricted our study to scenarios involving static, binary, spatially undirected networks. However, in real epidemics, the underlying contact network of individuals in a population might be dynamic, changing over time to represent the changing interactions between individuals (e.g., due to individual movement). Although the model complexity would be increased by considering such dynamic contact networks, their inclusion often provide better understanding of the dynamics of the infectious disease than when considering the static contact network. Incorporating uncertainty regarding such dynamic contact networks will doubtless result in a further increase in computation times. However, spatiotemporal or dynamic network based clusters could be formed to reduce the computation using an extension of the within-cluster ILM method used here. Another issue that would

be interesting to explore is to consider directional and/or weighted contact networks, either static or dynamic. Once again, an increase in computation time would be expected. For example, under a directional contact network framework, the contact matrix loses its symmetry, so more elements of the contact network need to be inferred.

Another possible avenue of research would be to combine the within-cluster ILM approach with other approximation methods, such as, approximation Bayesian computation (Toni et al., 2009; Beaumont et al., 2009), Gaussian process-based likelihood approximations (Pokharel and Deardon, 2016; Nemeth et al., 2018), or data-aggregation (Deeth and Deardon, 2016).

Finally, here we have demonstrated the use of our proposed model in the context of fitting models to epidemics over moderate population sizes ($N = 625$ individuals). The time savings using within-cluster likelihood approximations should be even more beneficial when analyzing large data sets for which the use of a full Bayesian analysis is computationally prohibitive. Thus, we intend to consider fitting the within-cluster network-based ILMs to large, real epidemic data sets (e.g., the 2001 UK foot-and-mouth disease).

Chapter 6

Conclusions

This thesis has presented work on four related studies in the area of infectious disease modelling with a focus on individual level models (ILMs). The second chapter introduced our R package *EpiILMCT*, which covers the implementation of the general framework of the continuous time ILMs. The other three chapters covered studies incorporating a large amount of missing information (contact network and infection times) into network-based continuous time ILMs.

The first such study investigated the effects of incorporating different levels of contact network uncertainty using data augmented Bayesian Markov chain Monte Carlo (DA-MCMC) methods. We examined the performance of such analyses in the presence or absence of contact network observational models based upon a priori knowledge of degree distribution, or the total number of connections in the network. The common objective of the second and the third studies was to develop approximation methodologies that reduce the computational expense of traditional Bayesian MCMC methods for fitting our models to data. In the second study, we applied approximate Bayesian computation population Monte Carlo (ABC-PMC) methods to infer the unknown contact network and infection

times of the individuals. Finally, we introduced cluster-constrained network-based ILMs that approximate the full model inference by partitioning the population into a number of spatial clusters. An overview of each Chapter is presented below, including overall conclusions and directions for future work.

6.1 Continuous Time Individual-Level Models of Infectious Disease: *EpilLMCT*

6.1.1 Summary

In this chapter, we introduced the R software package *EpilLMCT* (Almutiry et al., 2018), which facilitates the use of a broad range of continuous time ILMs under two compartmental frameworks ($\mathcal{SIR}$ and $\mathcal{SINR}$). It also allows for the analysis of partially observed infectious disease data, achieved using data augmented MCMC within a Bayesian framework.

We illustrated the use of the package through fitting network-based continuous time ILMs to simulated epidemic data, with unknown infection times. Also, we fitted a distance-based continuous ILM to real epidemic data from an experiment on tomato spotted wilt virus (TSWV), under the assumptions of a fully observed epidemic and unknown infection and removal times. Results of these studies demonstrate good performance of the package in the estimation of both the model parameters and missing event times. However, we did not cover all functionality of the package in this chapter. For instance, we did not illustrate

incorporating both distance and network in the kernel function, or allowing for nonlinearity between the susceptibility and transmissibility risk factors and the infection rate. However, implementation of such facets is simple.

6.1.2 Future Work

Incorporating more data uncertainty into the analyses, especially under the network-based model, is an option for future development of the package *EpiILMCT*. For example, networks are often only partially observed. However, the data augmentation could easily make the computation time for data analyses prohibitive.

Various strategies for mitigating this computational burden might be available for future versions of the package. For example, *OpenMP* parallel programming could be used to reduce the computation time associated with likelihood calculations. Also, approximate forms of inference, such as Gaussian process emulation (Pokharel and Deardon, 2016), approximate Bayesian computation (Beaumont et al., 2009), machine learning based model classification (Pokharel and Deardon, 2014), data-sampled likelihood approximation (Malik et al., 2014), or data-aggregation (Deeth and Deardon, 2016) could be useful in overcoming the aforementioned issue.

There are also other aspects of the package that can be extended in future versions. Currently, we consider only two compartmental frameworks ($\mathcal{SIR}$ and $\mathcal{SINR}$). An extension to allow other frameworks, such as $\mathcal{SI}$ and $\mathcal{SEIR}$ frameworks, is intended to be implemented in the next version. Another possible extension of the package is to con-

sider time-varying covariates which would add more variety to the current option of (fixed) covariates in the package.

6.2 Spatial Contact Network Uncertainty in Individual Level Models of Infectious Disease Transmission

6.2.1 Summary

In this chapter, we investigated the problem of fitting individual-level models to simulated epidemic data in the presence of an underlying but unobserved spatial contact network, along with unknown infection times/infectious periods, using data augmented MCMC methods. We considered the incorporation of global-level contact network information through observation models that are either based upon knowledge of degree distribution or the total number of connections in the network. Results show that including this global information provided a significant improvement in the estimation of the model parameters, as well as good approximations of the missing information. Specifically, we found that including an observation model based on the total number of connections in the contact network enhances the estimates of the model parameters and the average number of degree per node better than including the exact degree distribution. We also demonstrated the ability to produce good predictions of summary statistics that represent important features of the epidemic process under both completely known and unknown contact network

analyses.

6.2.2 Future Work

Further research on the implementation of different types of contact networks is required to ensure the efficiency of incorporating the global information about the contact network into the model. For simplicity, we considered a binary undirected spatial contact network to represent the underlying dynamics of all the simulated epidemics. However, the underlying population heterogeneities might be better described by a directed dynamic (or *static*) contact network. Thus, we suggest considering more complex epidemiological systems, in which underlying contact networks might be unobserved.

6.3 Incorporating Contact Network Uncertainty in Individual Level Models of Infectious Disease Using Approximate Bayesian Computation

6.3.1 Summary

This chapter investigated the ability of using approximate Bayesian computation population Monte Carlo (ABC-PMC) methods in analyzing large epidemic data sets. We mainly considered the fit of the network-based ILMs in situations where there is uncertainty both about infection times and the underlying contact network. This was done in the context

of both simulated data, and data from the UK 2001 foot-and-mouth disease epidemic. We employed large numbers of different combinations of summary statistics in order to find the most informative statistics of the epidemic. Our results suggest that ABC-PMC approaches to inference in such scenarios can produce good estimates of model parameters and fitted models with good predictive qualities, and also obtain good estimations of the underlying unknown contact network. We also compared the performance of ABC-PMC with a full Bayesian MCMC analysis when fitting relatively simple ILMs to simulated data sets with small population sizes. A significant reduction in computation time was observed in using ABC-PMC compared to a full Bayesian MCMC analysis, without sacrificing the model fit.

6.3.2 Future Work

Although we have achieved positive results in the implementation of ABC-PMC, there are some aspects of the study that would be interesting to explore in future studies. The first aspect is the issue of selecting "optimal" summary statistics that was used to compare simulated and observed data within the ABC algorithm. We used a large number of combinations of summary statistics for fitting our models, only a small subset of which were included here. This aspect is one of the challengeable parts of the ABC methods, especially in situations where there is an extreme lack of information as is seen when the underlying dynamics of infectious disease are unknown. In our studies, we found that including a discretized version of the epidemic curve produced more satisfactory results than a more parsimonious set of summary statistics of the epidemic curve, and sometimes vice versa.

It is still not clear how we would go about predicting which strategy would work better in advance. This would suggest that either multiple ABC-PMC analyses should be carried out (preferably) in parallel when analyzing epidemic data, or that some other approach, that allows for the comparison of different sets of summary statistics to be carried out as part of the analysis, needs to be employed (Joyce et al., 2008).

Another modelling challenge involves choosing appropriate prior distributions for the model parameters. We encountered better performance of ABC-PMC by using informative priors compared to uninformative priors. This is less of a problem for ABC-PMC approaches than, for example, rejection-based ABC approaches, since the sequential learning inherent in the ABC-PMC algorithm can successfully hone in on the areas of higher approximate posterior mass even with quite uninformative priors. However, the computational ramifications of using less informative priors can be significant, exhibited by both a slower decrease in the tolerance level between generations, and a higher rejection rate within generations. We related this to the common issue with individual-level epidemic models, where there is a high likelihood of observing a large amount of variation between simulations from the same model (i.e., with the same parameters). This large variation would lead to high rejection rates within the ABC algorithm. A possible approach that can overcome this phenomenon, and would also be of interest to address in further work, is the technique of so-called "Lazy ABC" introduced by Prangle (2016). The Lazy ABC algorithm tends to reject those simulated data sets that differ substantially from the observed data in early stages of the simulation, rather than rejecting after generating an entire data

set.

Several other aspects of the model could be worthwhile to explore in further work. We have confined ourselves to considering static, binary, spatial undirected networks. Of course, in many situations, the underlying contact network may not be spatial in nature, and may be directional and/or weighted. For example, consider livestock diseases, in which there is long-distance animal movement, which often occurs between farms and/or markets, and may drive disease spread through a dynamic, weighted, non-symmetric network.

There are also other forms of approximate inference that could be considered for these scenarios. For example, Gaussian process emulation (Pokharel and Deardon, 2016), classification methods (Pokharel and Deardon, 2014), and sampling-based likelihood approximations (Malik et al., 2016).

6.4 Contact Network Uncertainty in Individual Level Models of Infectious Disease Transmission Using Within-Cluster Likelihood Approximations

6.4.1 Summary

In this chapter, we introduced an approach to approximate full model inference when fitting network-based infectious disease models to epidemic data that contain large amounts of missing information (unknown infection times and contact network). The computa-

tional burden of modelling epidemics in the presence of this missing information naively can be extremely prohibitive. Thus, we introduced a within-cluster constrained version of the network-based continuous time ILM to allow for approximate, but computationally tractable, inference. The model is based on partitioning the population and their data (contact network and event times) into a number of spatially isolated clusters. Between-cluster infections are then modelled by spark infection parameters that allow for random spark infections.

We examined the fit of this new model by studying simulated epidemics over two population structures: well-clustered and scattered (unclustered) populations. We found better fit of the model under well-clustered populations compared to scattered populations. The main reason for the observed differences in performance is that the effect of imposing isolation between different parts (clusters) of the network is relatively low, since the true network already exhibits strong clustering. However, it was found overall that model parameters could be estimated reasonably well, even in the presence of infection time and contact network uncertainty, under all scenarios.

We also compared the performance of the within-cluster ILMs to our modified version of the likelihood inflating sampling algorithm (LISA) approach introduced by Entezari et al. (2017). We observed similar results of the two approaches under the fully-observed epidemic analysis. Although using LISA under the analyses incorporating data uncertainty (contact network and infection times) showed a great reduction in computation time compared to using the within-cluster ILM, we observed better performance of the within-cluster

ILM than LISA in terms of parameter estimates. This could be explained by the fact that data partitions in the LISA approach are assumed to be independently and identically distributed (Entezari et al., 2017), whereas clusters in the within-cluster ILM are assumed to be *conditionally* independent to each other and parameters are forced to be equal across clusters.

6.4.2 Future Work

The research carried out as part of the present thesis has uncovered some questions that require further examination. We have restricted our way of partitioning the scattered population to the PAM method that involves partitioning the population into a number of spatial clusters. Other partitioning methods could be applied, and, might result in clustering that produces better estimates. This can be carried out through some properties of the contact network if it is (partially) observed, or via other covariates correlated with the network. Another aspect is the choice of the number of clusters which needs to be chosen carefully. Although increasing the number of clusters reduces the computation time substantially, a deterioration in inferential performance could be observed. Choosing the number of clusters that substantially reduces computation time yet retains good inferential performance is preferable.

We have also restricted our study to scenarios involving static, binary, and spatial undirected networks. However, in real epidemics, the underlying contact network of individuals in a population might be dynamic, changing over time to represent the changing interac-

tions between individuals (e.g., due to individual movement). Thus, it would be interesting to extend the within-cluster ILM method used here to consider a dynamic network.

Another possible avenue of research would be to combine the within-cluster ILM approach with other approximation methods, such as approximation Bayesian computation (Toni et al., 2009; Beaumont et al., 2009), Gaussian process-based likelihood approximations (Pokharel and Deardon, 2016; Nemeth et al., 2018), or data aggregation (Deeth and Deardon, 2016). Finally, we have demonstrated in this work the use of our proposed model in the context of fitting models to epidemics over moderate population sizes ($N = 625$ individuals). The time savings obtained when using within-cluster likelihood approximations should be even more beneficial when analyzing large data sets, for which the use of a full Bayesian analysis is computationally prohibitive. Thus, we intend to consider fitting the within-cluster network-based ILMs to large, real epidemic data sets (e.g., the 2001 UK foot-and-mouth disease).

Bibliography

Aalen, O. O., Farewell, V. T., De Angelis, D., Day, N. E., and Nöel Gill, O. (1997). A Markov model for HIV disease progression including the effect of HIV diagnosis and treatment: application to AIDS prediction in England and Wales. *Statistics in Medicine*, 16(19):2191–2210.

Agarwal, A. and Duchi, J. C. (2012). Distributed delayed stochastic optimization. In *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*, pages 5451–5452. IEEE.

Almutiry, W., Warriyar. K. V., V., and Deardon, R. (2018). *EpiILMCT: continuous time distance-based and network-based individual level models for epidemics*. R package version 1.1.2.

Bakar, K. S. and Sahu, S. K. (2015). *sptimer*: spatio-temporal Bayesian modeling using R. *Journal of Statistical Software*, 63(15):1–32.

Bakar, K. S. and Sahu, S. K. (2017). *spTimer: spatio-temporal Bayesian modeling using R*. R package version 3.0-1.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *science*, 286(5439):509–512.

Barreto, M. L., Teixeira, M. G., and Carmo, E. H. (2006). Infectious diseases epidemiology. *Journal of Epidemiology & Community Health*, 60(3):192–195.

Barthélemy, M., Barrat, A., Pastor-Satorras, R., and Vespignani, A. (2005). Dynamical patterns of epidemic outbreaks in complex heterogeneous networks. *Journal of Theoretical Biology*, 235(2):275–288.

Beaumont, M. A., Cornuet, J.-M., Marin, J.-M., and Robert, C. P. (2009). Adaptive approximate Bayesian computation. *Biometrika*, 96(4):983–990.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.

Becker, N. G., Watson, L. F., and Carlin, J. B. (1991). A method of non-parametric back-projection and its application to AIDS data. *Statistics in Medicine*, 10(10):1527–1542.

Bifolchi, N., Deardon, R., and Feng, Z. (2013). Spatial approximations of network-based individual level infectious disease models. *Spatial and Spatio-temporal Epidemiology*, 6:59–70.

Bivand, R., Hauke, J., and Kossowski, T. (2013). Computing the jacobian in Gaussian spatial autoregressive models: an illustrated comparison of available methods. *Geographical Analysis*, 45(2):150–179.

Bivand, R. and Piras, G. (2015). Comparing implementations of estimation methods for spatial econometrics. *Journal of Statistical Software*, 63(18):1–36.

Britton, T. and O'Neill, P. D. (2002). Bayesian inference for stochastic epidemics in populations with random social structure. *Scandinavian Journal of Statistics*, 29(3):375–390.

Brown, G. D., Porter, A. T., Oleson, J. J., and Hinman, J. A. (2018). Approximate Bayesian computation for spatial SEIR (S) epidemic models. *Spatial and Spatio-temporal Epidemiology*, 24:27–37.

Brown, P. E., Chimard, F., Remorov, A., Rosenthal, J. S., and Wang, X. (2014). Statistical inference and computational efficiency for spatial infectious disease models with plantation data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63(3):467–482.

Brown, S., Csinos, A., Díaz-Pérez, J., Gitaitis, R., LaHue, S., Lewis, J., Martinez, N., McPherson, R., Mullis, S., Nischwitz, C., et al. (2005). Tospoviruses in solanaceae and other crops in the coastal plain of Georgia. *The University of Georgia College of Agriculture and Environmental Sciences, Research Report*, 704:19.

Caimo, A. and Friel, N. (2014). *bergm*: Bayesian exponential random graphs in R. *Journal of Statistical Software*, 61(2):1–25.

Cappé, O., Guillin, A., Marin, J.-M., and Robert, C. P. (2004). Population Monte Carlo. *Journal of Computational and Graphical Statistics*, 13(4):907–929.

Casagrande, R. (2000). Biological terrorism targeted at agriculture: the threat to US national security. *The Nonproliferation Review*, 7(3):92–105.

Chowell, G., Hengartner, N. W., Castillo-Chavez, C., Fenimore, P. W., and Hyman, J. M. (2004). The basic reproductive number of Ebola and the effects of public health measures: the cases of Congo and Uganda. *Journal of Theoretical Biology*, 229(1):119–126.

Dagum, L. and Menon, R. (1998). OpenMP: an industry standard API for shared-memory programming. *IEEE Computational Science and Engineering*, 5(1):46–55.

Danon, L., Ford, A. P., House, T., Jewell, C. P., Keeling, M. J., Roberts, G. O., Ross, J. V., and Vernon, M. C. (2011). Networks and the epidemiology of infectious disease. *Interdisciplinary Perspectives on Infectious Diseases*, 2011.

Deardon, R., Brooks, S. P., Grenfell, B. T., Keeling, M. J., Tildesley, M. J., Savill, N. J., Shaw, D. J., and Woolhouse, M. E. (2010). Inference for individual-level models of infectious diseases in large populations. *Statistica Sinica*, 20(1):239.

Deeth, L. E. and Deardon, R. (2016). Spatial data aggregation for spatio-temporal individual-level models of infectious disease transmission. *Spatial and Spatio-temporal Epidemiology*, 17:95–104.

Deeth, L. E., Deardon, R., et al. (2013). Latent conditional individual-level models for infectious disease modeling. *International Journal of Biostatistics*, 9(1):75–93.

Del Moral, P., Doucet, A., and Jasra, A. (2012). An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statistics and Computing*, 22(5):1009–1020.

Donaldson, A. and Alexandersen, S. (2001). Relative resistance of pigs to infection by natural aerosols of FMD virus. *Veterinary Record*, 148(19):600–602.

Drovandi, C. C. and Pettitt, A. N. (2011). Estimation of parameters for macroparasite population evolution using approximate Bayesian computation. *Biometrics*, 67(1):225–233.

D'Silva, J. P. and Eisenberg, M. C. (2017). Modeling spatial invasion of Ebola in West Africa. *Journal of Theoretical Biology*, 428:65–75.

Entezari, R., Craiu, R. V., and Rosenthal, J. S. (2017). Likelihood inflating sampling algorithm. *Canadian Journal of Statistics*.

Filippi, S., Barnes, C. P., Cornebise, J., and Stumpf, M. P. (2013). On optimality of kernels for approximate Bayesian computation using sequential Monte Carlo. *Statistical Applications in Genetics and Molecular Biology*, 12(1):87–107.

Forrester, M. L., Pettitt, A. N., and Gibson, G. J. (2006). Bayesian inference of hospital-acquired infectious diseases and control measures given imperfect surveillance data. *Biostatistics*, 8(2):383–401.

Gamerman, D. and Lopes, H. F. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC Press.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian data analysis*. Chapman & Hall/CRC.

Gog, J. R., Ballesteros, S., Viboud, C., Simonsen, L., Bjornstad, O. N., Shaman, J., Chao, D. L., Khan, F., and Grenfell, B. T. (2014). Spatial transmission of 2009 pandemic influenza in the US. *PLoS Computational Biology*, 10(6):e1003635.

Groendyke, C. and Welch, D. (2018). *epinet*: an R package to analyze epidemics spread across contact networks. *Journal of Statistical Software*, 83(11):1–22.

Groendyke, C., Welch, D., and Hunter, D. R. (2012). A network-based analysis of the 1861 Hagelloch measles data. *Biometrics*, 68(3):755–765.

Gumel, A. B., Ruan, S., Day, T., Watmough, J., Brauer, F., Van den Driessche, P., Gabrielson, D., Bowman, C., Alexander, M. E., Ardal, S., et al. (2004). Modelling strategies for controlling SARS outbreaks. *Proceedings of the Royal Society of London B: Biological Sciences*, 271(1554):2223–2232.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., Krivitsky, P. N., and Morris, M. (2017). *ergm: fit, simulate and diagnose exponential-family models for networks*. The Statnet Project (`http://www.statnet.org`). R package version 3.8.0.

Hughes, G., McRoberts, N., Madden, L. V., and Nelson, S. C. (1997). Validating mathematical models of plant-disease progress in space and time. *Mathematical Medicine and Biology: A Journal of the Institute of Mathematics and Its Applications*, 14(2):85–112.

Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M., and Morris, M. (2008). *ergm*: a package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software*, 24(3):1–29.

Jandarov, R., Haran, M., Bjørnstad, O., and Grenfell, B. (2014). Emulating a gravity model to infer the spatiotemporal dynamics of an infectious disease. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63(3):423–444.

Jenness, S. M., Goodreau, S. M., and Morris, M. (2018). *epimodel*: An R package for mathematical modeling of infectious disease over networks. *Journal of Statistical Software*, 84(8):1–47.

Jewell, C. P., Kypraios, T., Neal, P., Roberts, G. O., et al. (2009). Bayesian analysis for emerging infectious diseases. *Bayesian Analysis*, 4(3):465–496.

Joyce, P., Marjoram, P., et al. (2008). Approximately sufficient statistics and Bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, 7(1):26.

Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons.

Keeling, M. J. and Eames, K. T. (2005). Networks and epidemic models. *Journal of the Royal Society Interface*, 2(4):295–307.

Keeling, M. J., Woolhouse, M. E., Shaw, D. J., Matthews, L., Chase-Topping, M., Haydon, D. T., Cornell, S. J., Kappey, J., Wilesmith, J., and Grenfell, B. T. (2001). Dynamics of the 2001 UK foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape. *Science*, 294(5543):813–817.

Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 115(772):700–721.

Klovdahl, A. S. (1985). Social networks and the spread of infectious diseases: the AIDS example. *Social Science & Medicine*, 21(11):1203–1216.

Kwong, G. P. and Deardon, R. (2012). Linearized forms of individual-level models for large-scale spatial infectious disease systems. *Bulletin of Mathematical Biology*, 74(8):1912–1937.

Kwong, G. P., Poljak, Z., Deardon, R., and Dewey, C. E. (2013). Bayesian analysis of risk factors for infection with a genotype of porcine reproductive and respiratory syndrome virus in Ontario swine herds using monitoring data. *Preventive Veterinary Medicine*, 110(3-4):405–417.

Kypraios, T., D O'Neill, P., Huang, S. S., Rifas-Shiman, S. L., and Cooper, B. S. (2010). Assessing the role of undetected colonization and isolation precautions in reducing methicillin-resistant *staphylococcus aureus* transmission in intensive care units. *BMC Infectious Diseases*, 10(1):29.

Kypraios, T., Neal, P., and Prangle, D. (2017). A tutorial introduction to Bayesian inference for stochastic epidemic models using approximate Bayesian computation. *Mathematical Biosciences*, 287:42–53.

Lee, C., Garbett, A., and Wilkinson, D. J. (2018a). A network epidemic model for online community commissioning data. *Statistics and Computing*, 28(4):891–904.

Lee, D., Rushworth, A., and Napier, G. (2018b). Spatio-temporal areal unit modeling in R with conditional autoregressive priors using the *carbayesst* package. *Journal of Statistical Software*, 84(9):1–39.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2017). *cluster: cluster analysis basics and extensions*. R package version 2.0.6 — For new features, see the 'Changelog' file (in the package source).

Malik, R., Deardon, R., and Kwong, G. P. (2016). Parameterizing spatial models of infectious disease transmission that incorporate infection time uncertainty using sampling-based likelihood approximations. *PloS One*, 11(1):e0146253.

Malik, R., Deardon, R., Kwong, G. P., and Cowling, B. J. (2014). Individual-level modeling of the spread of influenza within households. *Journal of Applied Statistics*, 41(7):1578–1592.

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328.

Martin, A. D., Quinn, K. M., and Park, J. H. (2011). *mcmcpack*: Markov chain Monte Carlo in R. *Journal of Statistical Software*, 42(9):22.

McBryde, E., Gibson, G., Pettitt, A., Zhang, Y., Zhao, B., and McElwain, D. (2006). Bayesian modelling of an epidemic of severe acute respiratory syndrome. *Bulletin of Mathematical Biology*, 68(4):889–917.

McKinley, T., Cook, A. R., and Deardon, R. (2009). Inference in epidemic models without likelihoods. *The International Journal of Biostatistics*, 5(1).

Meyer, S., Held, L., and Höhle, M. (2017). Spatio-temporal analysis of epidemic phenomena using the R package *surveillance*. *Journal of Statistical Software*, 77(11):1–55.

Meyers, L. A., Newman, M., and Pourbohloul, B. (2006). Predicting epidemics on directed contact networks. *Journal of Theoretical Biology*, 240(3):400–418.

Neal, P. and Roberts, G. (2005). A case study in non-centering for data augmentation: stochastic epidemics. *Statistics and Computing*, 15(4):315–327.

Nemeth, C., Sherlock, C., et al. (2018). Merging MCMC subposteriors through Gaussian-process approximations. *Bayesian Analysis*, 13(2):507–530.

Numminen, E., Cheng, L., Gyllenberg, M., and Corander, J. (2013). Estimating the transmission dynamics of *streptococcus pneumoniae* from strain prevalence data. *Biometrics*, 69(3):748–757.

O'Neill, P. D. (2010). Introduction and snapshot review: relating infectious disease transmission models to data. *Statistics in Medicine*, 29(20):2069–2077.

Pebesma, E. (2018). Cran task view: handling and analyzing spatio-temporal data. Version 2018-04-20.

Perelson, A. S. and Nelson, P. W. (1999). Mathematical analysis of HIV-1 dynamics in vivo. *SIAM Review*, 41(1):3–44.

Pitzer, V. E., Leung, G. M., and Lipsitch, M. (2007). Estimating variability in the transmission of severe acute respiratory syndrome to household contacts in Hong Kong, China. *American Journal of Epidemiology*, 166(3):355–363.

Pokharel, G. and Deardon, R. (2014). Supervised learning and prediction of spatial epidemics. *Spatial and Spatio-temporal Epidemiology*, 11:59–77.

Pokharel, G. and Deardon, R. (2016). Gaussian process emulators for spatial individual-level models of infectious disease. *Canadian Journal of Statistics*, 44(4):480–501.

Prangle, D. (2016). Lazy abc. *Statistics and Computing*, 26(1-2):171–185.

Press, S. J. (2009). *Subjective and objective Bayesian statistics: principles, models, and applications*, volume 590. John Wiley & Sons.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16(12):1791–1798.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reynolds, A. P., Richards, G., de la Iglesia, B., and Rayward-Smith, V. J. (2006). Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms*, 5(4):475–504.

Riley, S., Fraser, C., Donnelly, C. A., Ghani, A. C., Abu-Raddad, L. J., Hedley, A. J., Leung, G. M., Ho, L.-M., Lam, T.-H., Thach, T. Q., et al. (2003). Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. *Science*, 300(5627):1961–1966.

Romanescu, R. and Deardon, R. (2016). Modeling two strains of disease via aggregate-level infectivity curves. *Journal of Mathematical Biology*, 72(5):1195–1224.

Rowlingson, B. and Diggle, P. (2017). *splancs: spatial and space-time point pattern analysis*. R package version 2.01-40.

Salmon, M., Schumacher, D., and Höhle, M. (2016). Monitoring count time series in R: aberration detection in public health surveillance. *Journal of Statistical Software*, 70(10):1–35.

Scheidegger, A. (2018). *adpatMCMC: implementation of a generic adaptive Monte Carlo Markov chain sampler*. R package version 1.3.

Schweinberger, M., Handcock, M., and Luna, P. (2018). *hergm: hierarchical exponential-family random graph models with local dependence*. R package version 3.2-0.

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I., and McCulloch, R. E. (2016). Bayes and big data: the consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management*, 11(2):78–88.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.

Sisson, S. A., Fan, Y., and Tanaka, M. M. (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765.

Ster, I. C. and Ferguson, N. M. (2007). Transmission parameters of the 2001 foot and mouth epidemic in Great Britain. *PLoS One*, 2(6):e502.

Streftaris, G. and Gibson, G. J. (2004). Bayesian analysis of experimental epidemics of foot-and-mouth disease. *Proceedings of the Royal Society B: Biological Sciences*, 271(1544):1111.

Sweeting, M. J., De Angelis, D., and Aalen, O. O. (2005). Bayesian back-calculation using a multi-state model with application to HIV. *Statistics in Medicine*, 24(24):3991–4007.

Taffe, P., May, M., and Study, S. H. C. (2008). A joint back calculation model for the imputation of the date of HIV infection in a prevalent cohort. *Statistics in Medicine*, 27(23):4835–4853.

Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–518.

Taylor, B. M., Davies, T. M., Rowlingson, B. S., and Diggle, P. J. (2013). *lgcp*: an R package for inference with spatial and spatio-temporal log-Gaussian Cox processes. *Journal of Statistical Software*, 52(4):1–40.

Taylor, B. M., Davies, T. M., Rowlingson, B. S., and Diggle, P. J. (2015). Bayesian inference and data augmentation schemes for spatial, spatiotemporal and multivariate log-Gaussian Cox processes in R. *Journal of Statistical Software*, 63(7):1–48.

R Epidemics Consortium (2018). Released projects and packages. Retrieved 2018-03-28 from `http://www.repidemicsconsortium.org/projects/`.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31):187–202.

Volz, E. (2008). SIR dynamics in random networks with heterogeneous connectivity. *Journal of Mathematical Biology*, 56(3):293–310.

Walker, D. M., Allingham, D., Lee, H. W. J., and Small, M. (2010). Parameter inference in small world network disease models with approximate Bayesian computational methods. *Physica A: Statistical Mechanics and its Applications*, 389(3):540–548.

Warriyar KV, V. and Deardon, R. (2018). *EpiILM: spatial and network based individual level models for epidemics*. R package version 1.4.2.

WHO, W. H. O. (2016). Situation report: Ebola virus disease. Retrieved 2018-06-26 from `http://http://apps.who.int/iris/bitstream/handle/10665/208883/ebolasitrep_10Jun2016_eng.pdf?sequence=1`.

Appendix A

Continuous Time Individual-Level Models of Infectious Disease: *EpilLMCT*

A.1 R Code of Extracting Individual Level Data from the *surveillance* Package

Here, we illustrate the extraction of individual level data from the *surveillance* package for use in the *EpilLMCT* package. We consider the toy data set (*fooepidata*) representing a population of 100 individuals that is used in the *twinSIR* examples of the *surveillance* package. The data can be found in *surveillance* via *data("fooepidata", package = "surveillance").*

```
R> library( "surveillance" )
R> data( "fooepidata" )
R> names( fooepidata )
 [1] "BLOCK"   "id"      "start"   "stop"    "atRiskY" "event"
 [7] "Revent"    "x"       "y"       "z1"      "z2"      "B1"
 [13] "B2"
```

The *fooepidata* event history consists of 178 time *BLOCK*s of 100 rows, where each row describes the state of individual id during the corresponding time interval (*start*, *stop*).

```
R> head(fooepidata, n = 5)

     BLOCK id start      stop atRiskY event Revent            x          y
1        1  1     0 0.6970682       1     0      0  1.262954285  0.7818592
246      1  2     0 0.6970682       1     0      0 -0.326233361 -0.7767766
369      1  3     0 0.6970682       1     0      0  1.329799263 -0.6159899
612      1  4     0 0.6970682       1     0      0  1.272429321  0.0465803
760      1  5     0 0.6970682       1     0      0  0.414641434 -1.1303858

     z1        z2 B1 B2
1     0 0.0000000  0  0
246   1 0.6931472  0  0
369   0 1.0986123  0  0
612   1 1.3862944  0  0
760   1 1.6094379  0  0
[....]
```

The *start* and *stop* variables represent the start and end of interval time points (in continuous time) that indicate the waiting time between consequence event times (infection and removal times). The binary variables *event* and *Revent* are used to indicate the occurrence of newly infected or removed individuals at the stop time of each time interval (*BLOCK*), respectively. Thus, the *stop* time is taken to be the infection or removal times of the infected or removed individuals in each time interval. The coordinates of individuals is represented in columns *x* and *y*. The *fooepidata* data set contains also endemic and epidemic covariates. Endemic covariates are represented by the columns named *z1* and *z2* (the exact interpretation of these covariates is not given). Epidemic covariates are repre-

sented by the columns named *B1* and *B2*, and they indicate the count of currently infective individuals for each individual within, and greater than one unit distance, respectively. See (*help(epidata, package= "surveillance")*) for more details about the data structure. From this data set, we extract only the event times and XY coordinates of each individual, ignoring the purely spatial epidemic covariates which are directly modelled by the distance kernel in *EpilLMCT*.

```
R> infection <- matrix(fooepidata$event, ncol = 100, byrow = T)
R> removed <- matrix(fooepidata$Revent, ncol = 100, byrow = T)
R> event.time <- matrix(fooepidata$stop, ncol = 100, byrow = T)
R> xx <- matrix(fooepidata$x, ncol = 100, byrow = T)
R> yy <- matrix(fooepidata$y, ncol = 100, byrow = T)
R> s <- nrow(infection)
R> epidat <- matrix(0, ncol = 5, nrow = 178)
R> for (i in 1:s) {
+   if (any(infection[i, ] != 0)) {
+      k1 <- which(infection[i, ] == 1)
+      epidat[k1, ] <- c(k1, event.time[i, k1], 0, xx[i, k1], yy[i, k1])
+   } else {
+      k1 <- which(removed[i, ] == 1)
+      epidat[k1, 3] <- event.time[i, k1]
+   }
+ }
R> epidat1 <- subset(epidat, epidat[, 1] != 0)
R> epp1 <- epidat1[order(epidat1[, 2]), ]
R> epp2 <- epp1[, 1:3]
```

```
R> epp3 <- matrix(0, ncol = 4, nrow = 100)
R> epp3[1:88, c(1, 2, 4)] <- epp2[, c(1, 3, 2)]
R> epp3[1:88, 3] <- epp3[1:88, 2] - epp3[1:88, 4]
R> loc <- matrix(0, ncol = 2, nrow = 100)
R> loc[1:88, 1] <- epp1[, 4]
R> loc[1:88, 2] <- epp1[, 5]
```

To complete the structure of the epidemic data, we need to extract the data set of those uninfected individuals, and assign infinity values (Inf) to their infection and removal times.

```
R> remain <- NULL
R> j <- 1
R> for (i in 1:100) {
+   if (all(epp3[1:88, 1] != i)) {
+     remain[j] <- i
+     j <- j + 1
+   }
+ }
R> subepi <- subset(fooepidata, fooepidata$BLOCK == 1)
R> names(subepi)
R> loc[89:100, 1] <- subepi$x[remain]
R> loc[89:100, 2] <- subepi$y[remain]
R> epp3[89:100, 1] <- remain
R> epp3[89:100, c(2, 4)] <- Inf
```

The data set *epp3* can be now used in the *EpiILMCT* package using the model given in Equation 2.2 without covariates through the following code:

```
R> dis <- as.matrix(dist(loc[, 1:2], method = "euclidean"))
R> library( "EpiILMCT" )
R> mcmc1 <- epictmcmc(epidat = epp3, type = "SIR",
+   kerneltype = "distance", kernelmatrix = dis,
+   distancekernel = "powerlaw", datatype = "known epidemic", nsim = 50000,
+   sus.par = c(0.1, "gamma", 1, 0.001, 0.005),
+   kernel.par = c(0.1, "gamma", 1, 0.001, 0.1),
+   spark.par = c(0.1, "gamma", 1, 0.001, 0.05))
```

We include the spark term here to best model the endemic component used in the *twinSIR* model. The inclusion of the spark term also allows for the fact that there are no infectious individuals during times intervals (*BLOCK*) of the epidemic. The infection of individuals in these periods is captured by the endemic part in *twinSIR* function.

Without incorporating the spark term in the *epictmcmc* function, a zero likelihood will result, preventing the successful fitting of the model to the data. To get the output estimates of the model parameters, we used the **S3** method *summary.epictmcmc* as follows:

```
R> summary(mcmc1, start = 1000)
*******************************************************
Model: SIR distance-based continuous time ILM
Method: Markov chain Monte Carlo (MCMC)
Data assumption: fully observed epidemic
number.iteration : 49000 iterations
number.parameter : 3 parameters
*******************************************************
 1. Empirical mean and standard deviation for each variable,
```

```
plus standard error of the mean:

                       Mean         SD    Naive SE Time-series SE
Alpha_s[1]        0.00846595 0.00104967 4.74186e-06    1.41592e-05
Spark             0.00859629 0.00468787 2.11774e-05    7.11105e-05
Spatial parameter 0.29151566 0.19570189 8.84082e-04    4.37104e-03
 2. Quantiles for each variable:

                       2.5%        25%        50%        75%     97.5%
Alpha_s[1]        0.00650518 0.00773071 0.00843069 0.00913175 0.0106893
Spark             0.00156535 0.00507982 0.00790671 0.01142191 0.0196101
Spatial parameter 0.01754684 0.13614661 0.25731035 0.41421932 0.7378014
 3. Empirical mean, standard deviation, and quantiles for the log likelihood,
          Mean             SD       Naive SE Time-series SE
  -2.38261e+02    1.22414e+00    5.53005e-03    2.08540e-02


    2.5%      25%      50%      75%    97.5%
-241.399 -238.822 -237.948 -237.368 -236.828
 4. acceptance.rate :
      Alpha_s[1]             Spark Spatial parameter
        0.246145          0.181244          0.883558
```

We also demonstrate the modelling of these data using *twinSIR* function with no endemic covariates. However, a baseline term (baseline hazard rate) will be included in this case in the endemic component to represent the background rate of infection in the population, as explained in *help(twinSIR, package = "surveillance")*. The following code illustrates the use of *twinSIR* in analyzing this data set.

```
R> fit1 <- twinSIR( ~ B1 + B2, data = fooepidata)
```

```
R> summary(fit1)

Call:

twinSIR(formula = ~B1 + B2, data = fooepidata)

Coefficients:

                  Estimate Std. Error z value Pr(>|z|)

B1                0.023960   0.004208   5.693 1.25e-08 ***

B2                0.003395   0.001119   3.034  0.00241 **

cox(logbaseline) -6.010580   0.659257  -9.117  < 2e-16 ***

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total number of infections:  88

One-sided AIC: 474.05

Log-likelihood: -235.2

Number of log-likelihood evaluations: 26
```

The posterior means of the ILM parameters (α, β) are 0.0085 and 0.2915, respectively. Figure A.1.1 shows the ILM power-law distance kernel function under the posterior mean, along with the distance function suggested by the MLEs of the model parameters from the *twinSIR* analysis. We can see broad agreement, although the step function assumption of the *twinSIR* seems less reasonable than the continuous decay of the ILM kernel.

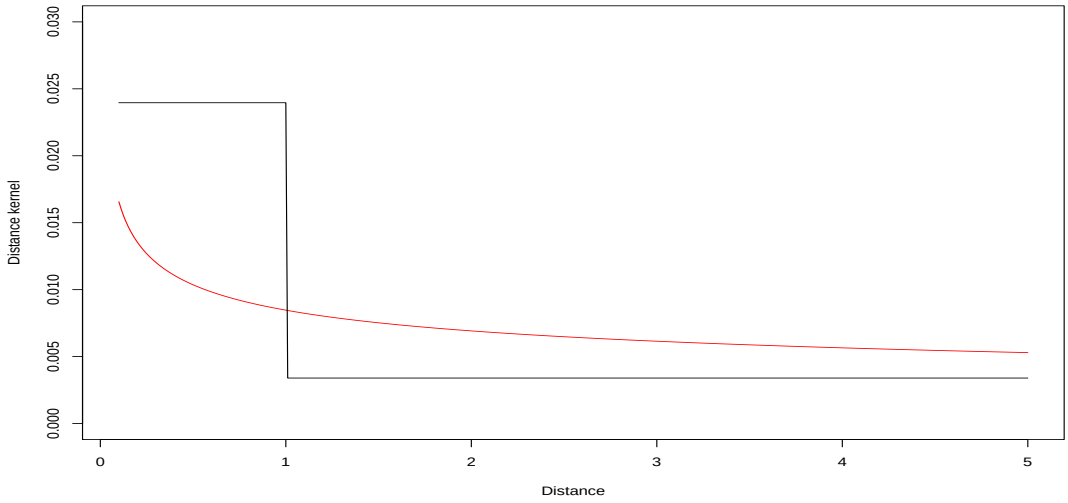


Figure A.1.1: The marginal posterior distribution of the distance kernels. Black line represents the spatial terms of the *surveillance* package, and red line represents the distance kernel function of the *EpiILMCT* package.

Appendix B

Incorporating Contact Network Uncertainty in Individual Level Models of Infectious Disease using Approximate Bayesian Computation

B.1 The 2001 UK Foot-and-Mouth Disease Epidemic in the Cumbria Region

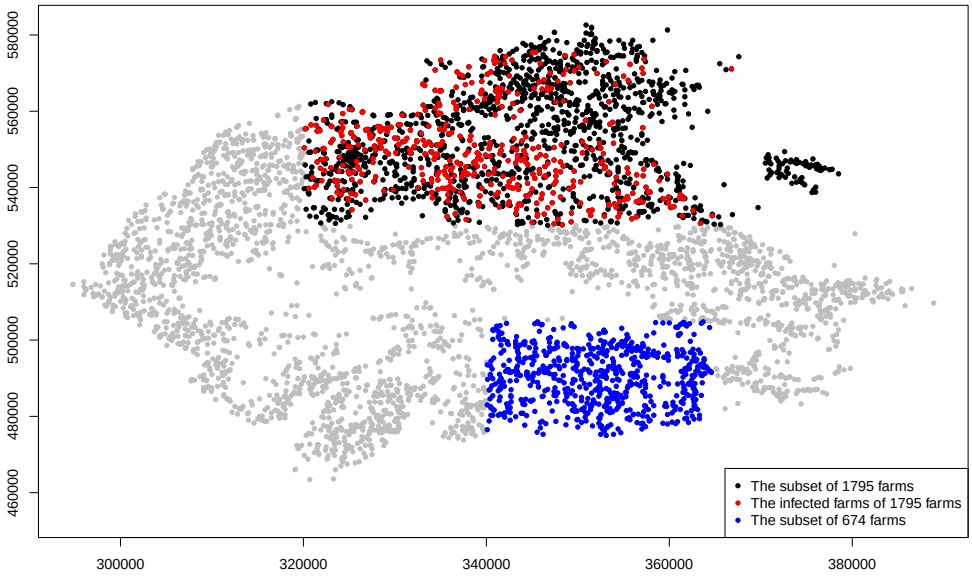


Figure B.1.1: Locations of farms in the Cumbria region, UK during the 2001 UK Foot and Mouth Disease Epidemic.

B.2 Samples of the Estimated Degree Distributions of Fitting Simple Network-Based ILM

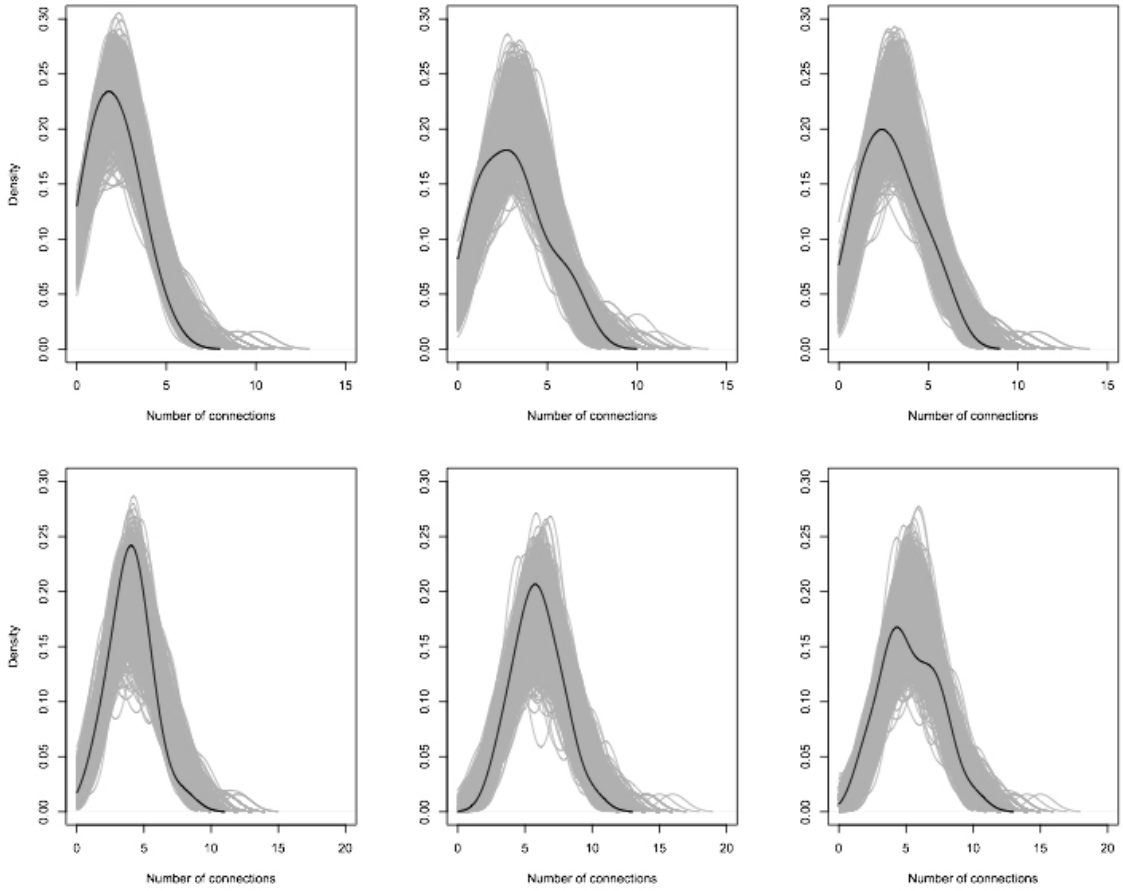


Figure B.2.1: Estimated posterior degree distributions for fitting the simple network-based ILM to six representative data sets. 1^{st} row: sparse networks; 2^{nd} row: intense networks.

B.3 Samples of the Approximated Posterior Distribution of the Epidemic Curve of the Infection Times

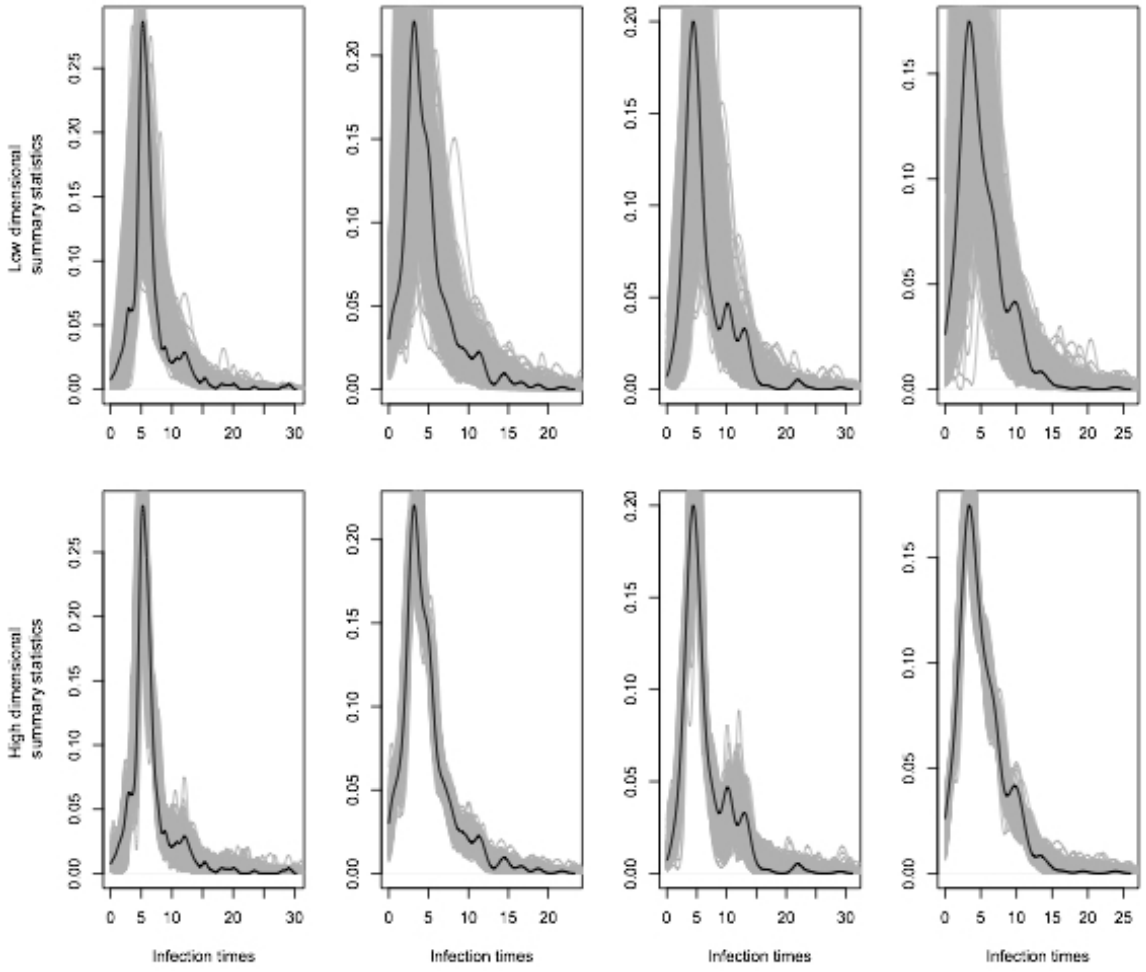


Figure B.3.1a: The approximated posterior distribution of the epidemic curves of the infection times for fitting the FMD network-based ILM to four simulated-FMD epidemic data sets without spark terms (scenario 1) under ABC-sim(1) (first row) and ABC-sim(2) (second row) summary statistics.

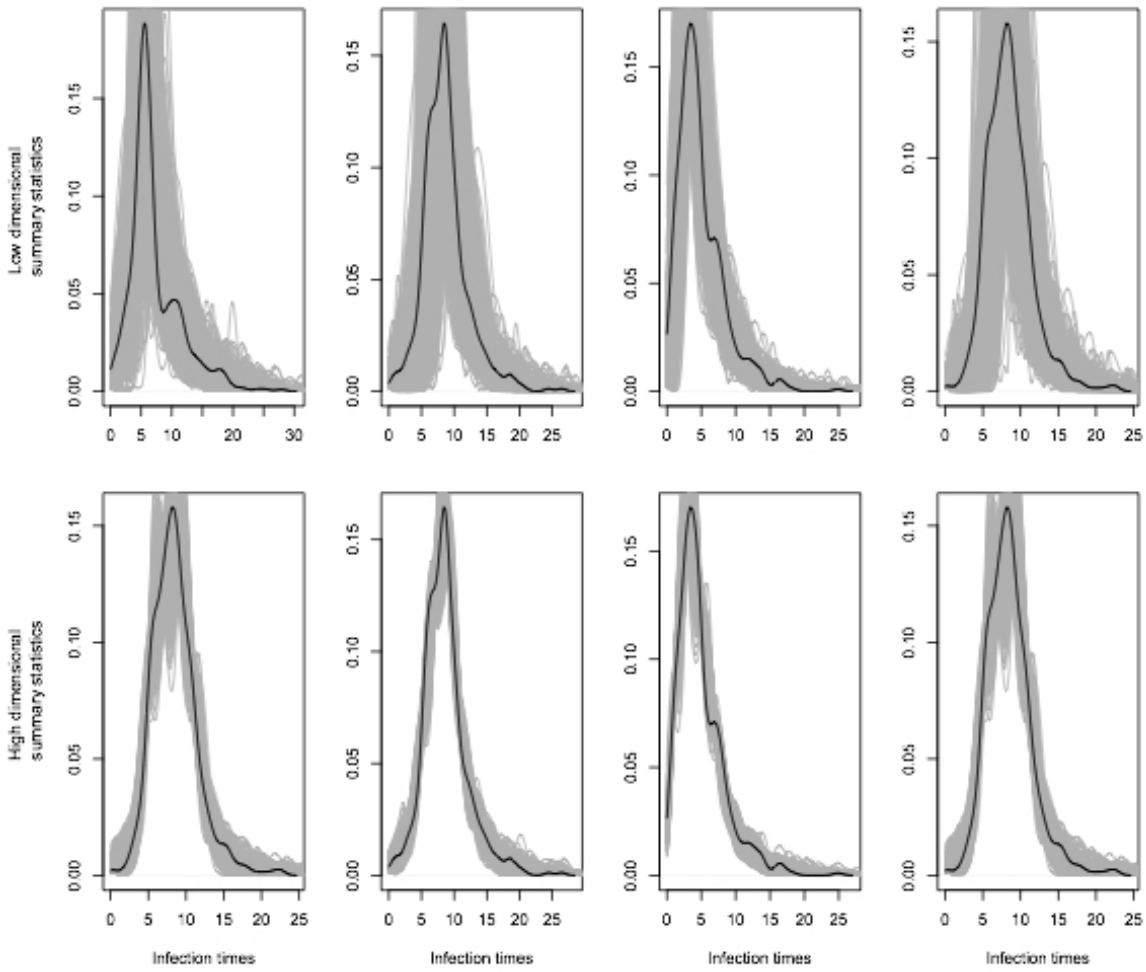


Figure B.3.1b: The approximated posterior distribution of the epidemic curves of the infection times for fitting the FMD network-based ILM to four simulated-FMD epidemic data sets with the inclusion of spark terms (scenario 2) under ABC-sim(1) (first row) and ABC-sim(2) (second row) summary statistics.

B.4 Samples of the Intermediate Marginal Posterior Distributions for the Model Parameters

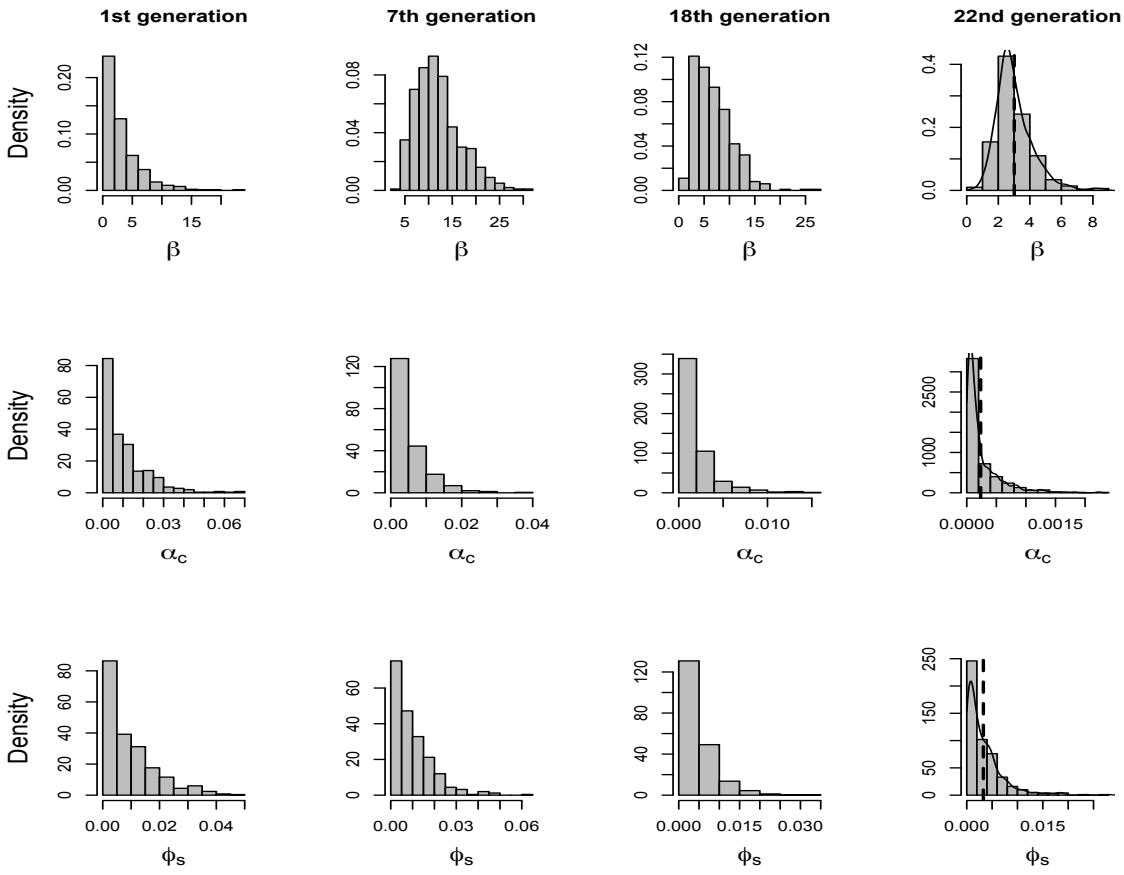


Figure B.4.1a: Histograms of the intermediate marginal posterior distributions for the parameters β, α_c and ϕ_s for the FMD network-based ILM fitted to the real FMD epidemic data in different generations of the ABC where the last column represents the approximated marginal posterior distributions of the parameters.

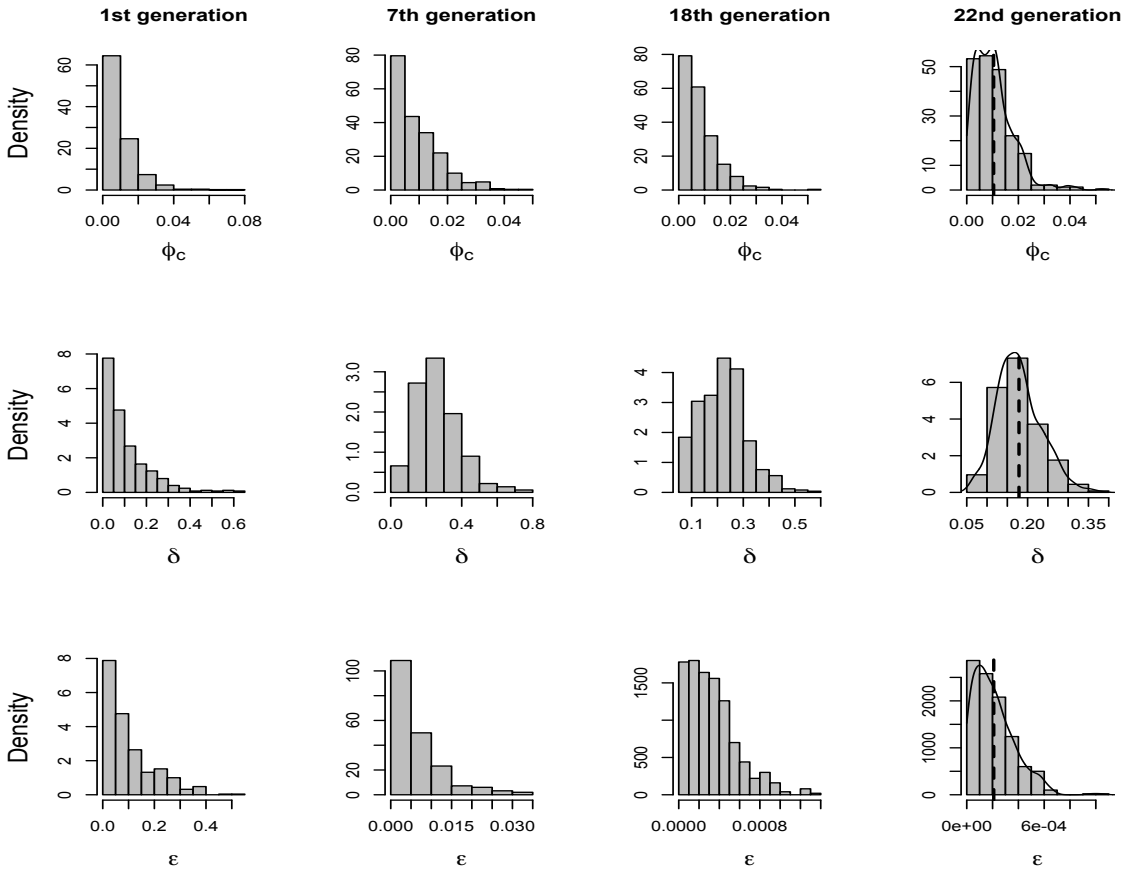


Figure B.4.1b: Histograms of the intermediate marginal posterior distributions for the parameters ϕ_c, δ and ϵ for the FMD network-based ILM fitted to the real FMD epidemic data in different generations of the ABC where the last column represents the approximated marginal posterior distributions of the parameters.

Appendix C

Contact Network Uncertainty in Individual Level Models of Infectious Disease Transmission Using Within-Cluster Likelihood Approximations

C.1 Summaries of the Cluster Data Sets

		Clusters									
		1	2	3	4	5	6	7	8	9	10
Data (1)	Cluster size	68	66	66	64	63	61	61	60	58	58
	Number of infected individuals	61	60	56	56	56	55	54	55	49	45
	Number of connections (full)	1850	2294	1930	1708	1680	1565	1402	1563	1437	1395
	Number of connections (partial)	1798	2212	1876	1676	1580	1552	1354	1518	1386	1312
Data (2)	Cluster size	66	65	64	65	63	62	61	59	61	59
	Number of infected individuals	56	49	55	55	54	52	56	45	53	50
	Number of connections (full)	1710	1676	2018	1775	1681	1844	1466	1603	1544	1557
	Number of connections (partial)	1634	1596	1972	1722	1642	1780	1444	1460	1478	1514
Data (3)	Cluster size	60	52	61	71	52	53	90	64	63	59
	Number of infected individuals	44	43	0	0	30	38	2	0	0	45
	Number of connections (full)	1199	1192	1101	1659	883	1015	2552	1249	1516	1164
	Number of connections (partial)	1028	994	0	0	652	786	114	0	0	1060
Data (4)	Cluster size	69	66	65	63	62	61	61	61	58	59
	Number of infected individuals	63	52	56	52	0	51	53	2	48	53
	Number of connections (full)	2139	1922	1992	1737	1638	1667	1739	1565	1330	1545
	Number of connections (partial)	2036	1834	1952	1606	0	1604	1702	50	1222	1498
Data (5)	Cluster size	66	66	63	65	65	65	62	59	57	57
	Number of infected individuals	52	61	51	58	54	57	56	36	48	52
	Number of connections (full)	1968	1926	1597	1601	1709	1992	1659	1299	1752	1383
	Number of connections (partial)	1916	1844	1468	1532	1680	1904	1602	1136	1660	1248
Data (6)	Cluster size	71	66	65	64	61	63	59	58	59	59
	Number of infected individuals	63	1	42	2	50	39	52	47	5	22
	Number of connections (full)	1568	1352	1215	1122	1307	1147	1200	1045	987	1263
	Number of connections (partial)	1384	10	1104	72	1242	984	1092	960	144	850
Data (7)	Cluster size	66	67	67	64	62	62	61	59	58	59
	Number of infected individuals	53	55	39	48	31	45	0	8	34	13
	Number of connections (full)	1494	1519	1507	1417	1216	1287	1228	990	992	1092
	Number of connections (partial)	1442	1402	1202	1236	864	1172	0	266	884	410
Data (8)	Cluster size	68	63	66	65	63	62	62	57	60	59
	Number of infected individuals	57	1	54	57	49	43	46	27	47	1
	Number of connections (full)	1553	1266	1409	1399	1173	1166	1393	1211	1393	1125
	Number of connections (partial)	1514	44	1296	1358	1096	938	1302	896	1144	22
Data (9)	Cluster size	67	63	64	63	63	62	62	64	58	59
	Number of infected individuals	54	45	49	49	34	53	43	52	10	49
	Number of connections (full)	1893	1384	1467	1373	1399	1523	1366	1548	1216	1331
	Number of connections (partial)	1758	1250	1400	1272	1096	1400	1280	1288	356	1238
Data (10)	Cluster size	69	69	65	62	58	65	60	61	58	58
	Number of infected individuals	2	56	47	44	47	47	41	52	45	38
	Number of connections (full)	1434	1671	1497	1051	1306	1431	1556	1458	1339	915
	Number of connections (partial)	44	1572	1398	978	1052	1276	1432	1264	1268	768

Table C.1.1: Summaries of the ten epidemics data sets spreading on well-clustered populations. The number of clusters in this case is ten clusters.

		Clusters									
		1	2	3	4	5	6	7	8	9	10
Data (1)	Cluster size	65	75	48	39	60	54	73	74	62	75
	Number of infected individuals	52	70	39	8	55	46	58	52	43	65
	Number of connections (full)	463	733	358	216	557	413	604	691	504	577
	Number of connections (partial)	368	684	298	44	442	278	468	518	406	526
Data (2)	Cluster size	75	71	50	56	49	65	74	69	51	65
	Number of infected individuals	60	8	39	46	3	59	51	50	30	13
	Number of connections (full)	651	460	373	503	286	581	610	506	368	470
	Number of connections (partial)	504	66	300	432	16	478	464	392	286	116
Data (3)	Cluster size	76	57	67	68	72	54	45	66	49	71
	Number of infected individuals	64	42	31	43	7	32	24	53	27	58
	Number of connections (full)	782	360	487	565	474	413	289	477	389	602
	Number of connections (partial)	666	270	300	416	52	320	194	404	278	462
Data (4)	Cluster size	71	84	85	70	54	52	56	73	32	48
	Number of infected individuals	5	70	74	52	5	42	13	67	3	9
	Number of connections (full)	542	713	842	602	329	407	404	746	182	277
	Number of connections (partial)	56	588	690	430	44	348	110	640	16	44
Data (5)	Cluster size	58	75	68	49	59	67	55	67	76	51
	Number of infected individuals	35	48	51	8	26	34	42	61	61	39
	Number of connections (full)	443	627	554	302	379	507	400	761	627	364
	Number of connections (partial)	296	530	420	80	156	334	302	648	556	310
Data (6)	Cluster size	60	71	70	62	60	62	65	65	46	64
	Number of infected individuals	55	60	27	46	52	44	48	38	35	51
	Number of connections (full)	658	616	437	539	488	400	487	459	337	505
	Number of connections (partial)	548	504	164	452	374	292	396	342	268	412
Data (7)	Cluster size	52	48	74	72	48	61	68	67	62	73
	Number of infected individuals	37	11	55	62	14	54	53	50	55	57
	Number of connections (full)	419	360	634	687	321	529	555	509	493	603
	Number of connections (partial)	314	96	474	588	98	414	440	426	434	530
Data (8)	Cluster size	54	58	66	63	84	72	53	74	45	56
	Number of infected individuals	43	46	56	56	74	61	46	47	29	15
	Number of connections (full)	470	442	552	623	774	619	318	550	280	340
	Number of connections (partial)	402	372	446	510	630	520	264	388	182	100
Data (9)	Cluster size	96	79	57	79	46	56	42	52	65	53
	Number of infected individuals	90	44	37	54	6	4	38	29	31	39
	Number of connections (full)	976	572	447	687	299	424	361	359	490	353
	Number of connections (partial)	848	410	300	506	60	32	328	254	314	260
Data (10)	Cluster size	58	64	68	69	78	54	56	64	45	69
	Number of infected individuals	44	39	50	49	56	42	40	6	15	22
	Number of connections (full)	554	408	544	527	613	374	375	524	315	590
	Number of connections (partial)	422	328	416	450	536	296	286	36	128	284

Table C.1.2: Summaries of the ten epidemics data sets with binary covariate spreading on scattered populations. The number of clusters in this case is ten clusters.

		Clusters									
		1	2	3	4	5	6	7	8	9	10
Data (1)	Cluster size	58	43	57	95	58	59	61	74	67	53
	Number of infected individuals	46	38	34	76	43	41	51	63	48	39
	Number of connections (full)	430	254	480	834	391	517	515	575	569	375
	Number of connections (partial)	336	216	298	680	294	392	412	488	466	302
Data (2)	Cluster size	60	50	66	77	50	60	65	72	56	69
	Number of infected individuals	41	44	60	73	44	43	62	66	45	60
	Number of connections (full)	352	358	636	644	379	443	630	556	371	509
	Number of connections (partial)	222	288	556	552	338	308	562	466	312	438
Data (3)	Cluster size	70	52	77	53	66	70	49	55	77	56
	Number of infected individuals	63	24	43	41	60	55	31	41	17	42
	Number of connections (full)	528	356	708	392	540	619	313	368	732	452
	Number of connections (partial)	464	222	542	312	494	476	244	258	176	400
Data (4)	Cluster size	42	79	59	54	71	65	37	82	78	58
	Number of infected individuals	5	38	11	40	67	56	28	69	66	34
	Number of connections (full)	281	700	374	375	667	633	249	727	605	359
	Number of connections (partial)	26	436	80	268	572	556	208	618	506	210
Data (5)	Cluster size	57	54	83	78	72	73	60	63	33	52
	Number of infected individuals	54	7	45	43	13	69	46	38	26	50
	Number of connections (full)	512	293	608	684	528	644	605	564	244	510
	Number of connections (partial)	416	72	378	434	160	518	478	400	208	418
Data (6)	Cluster size	55	48	77	47	79	83	64	68	48	56
	Number of infected individuals	6	45	65	35	60	26	63	63	42	49
	Number of connections (full)	360	407	685	291	627	606	536	563	306	513
	Number of connections (partial)	42	342	560	194	502	208	450	480	258	470
Data (7)	Cluster size	68	57	57	56	66	70	60	76	66	49
	Number of infected individuals	60	28	50	4	58	63	57	44	56	41
	Number of connections (full)	540	350	457	308	672	570	685	596	533	337
	Number of connections (partial)	418	194	390	32	552	504	540	424	448	268
Data (8)	Cluster size	53	68	53	54	63	60	65	50	77	82
	Number of infected individuals	39	51	33	48	5	16	46	41	65	69
	Number of connections (full)	404	480	388	432	409	396	491	365	722	765
	Number of connections (partial)	332	382	276	332	52	108	356	298	582	656
Data (9)	Cluster size	61	67	64	61	69	74	50	60	64	55
	Number of infected individuals	18	33	54	39	57	52	21	46	59	54
	Number of connections (full)	396	637	530	449	674	561	360	489	578	520
	Number of connections (partial)	120	400	456	294	582	450	202	368	454	382
Data (10)	Cluster size	50	71	62	47	58	58	81	75	68	55
	Number of infected individuals	44	58	10	28	26	45	73	66	60	52
	Number of connections (full)	382	511	363	312	403	466	673	638	537	429
	Number of connections (partial)	324	446	76	212	174	398	612	562	468	348

Table C.1.3: Summaries of the ten epidemics data sets with discrete covariate spreading on scattered populations. The number of clusters in this case is ten clusters.

C.2 The Estimated Infection Times and Degree Distributions

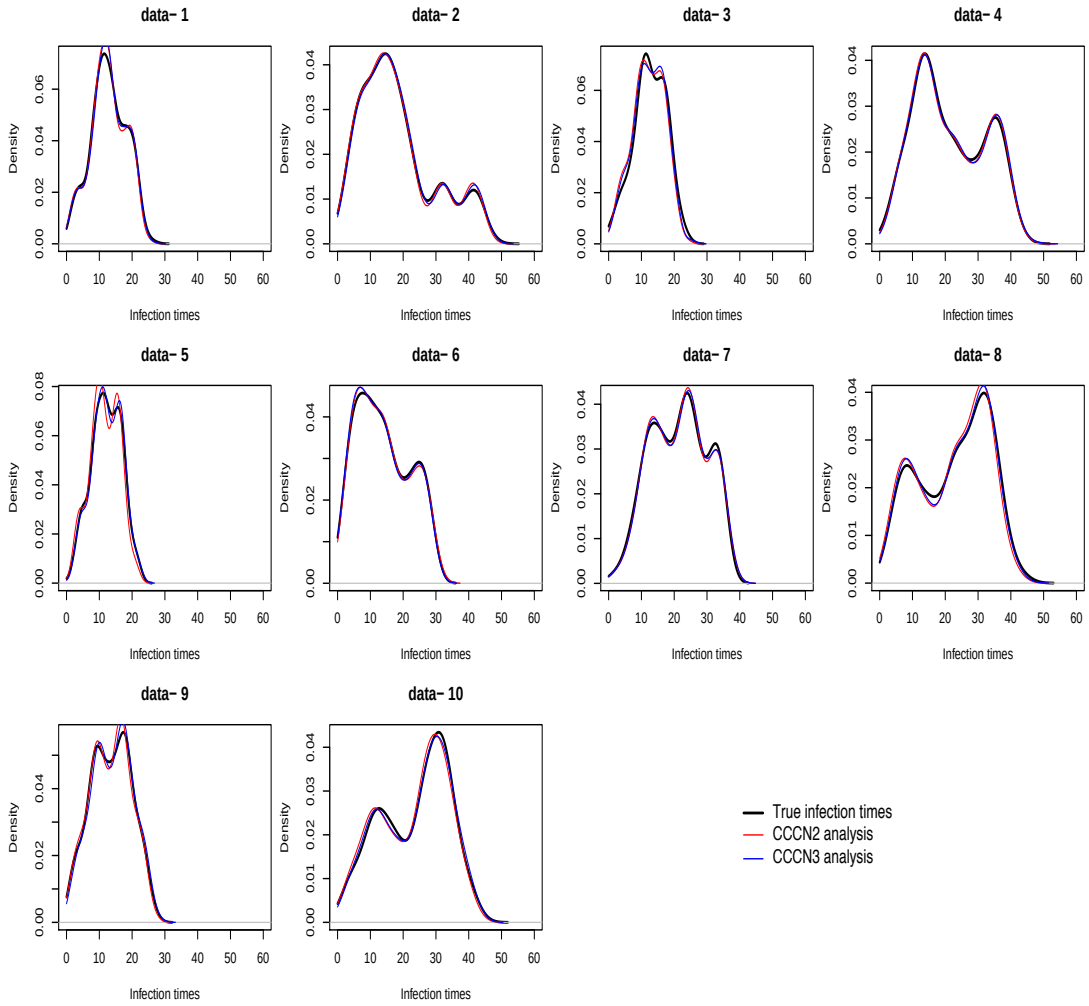


Figure C.2.1: The density of the posterior means of the infection times for the CCCN2 and CCCN3 analyses of the well-clustered population epidemics.

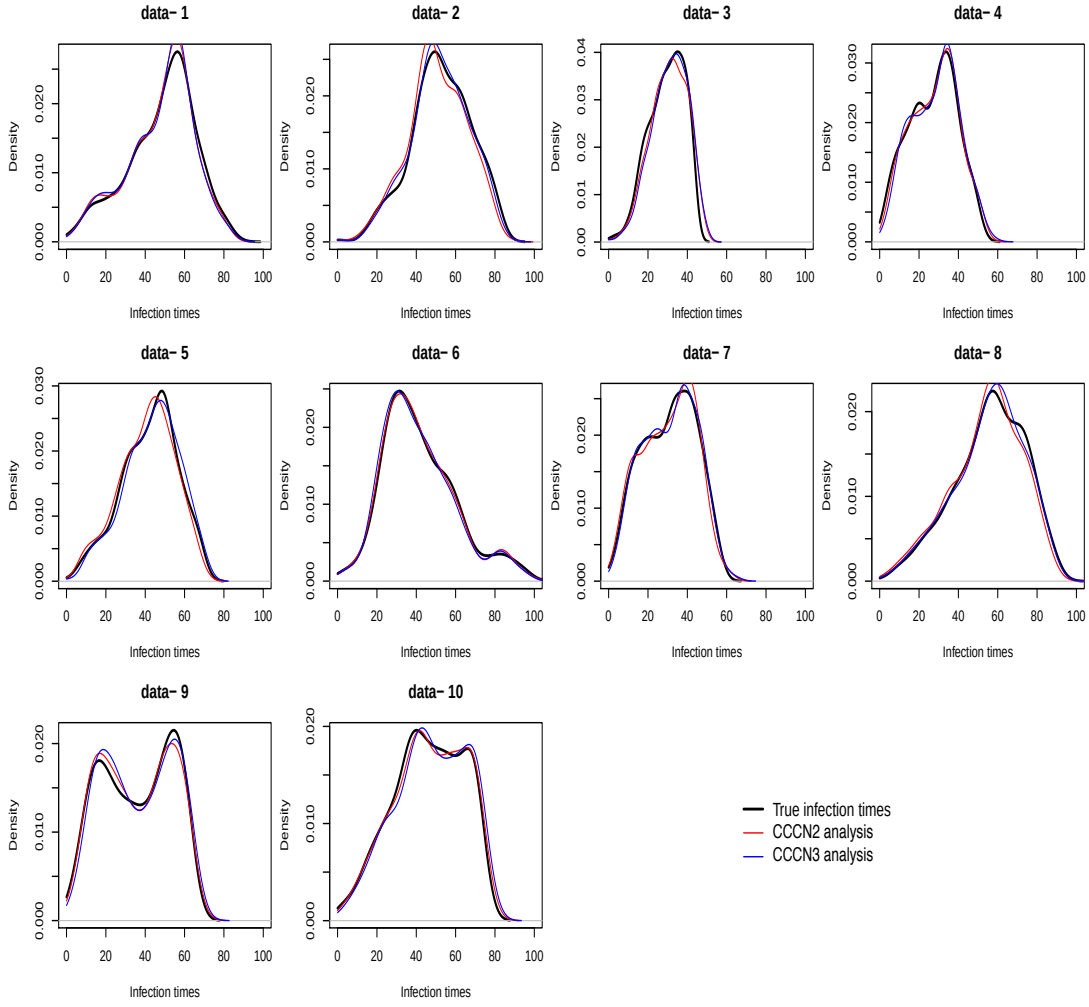


Figure C.2.2: The density of the posterior means of the infection times for the CCCN2 and CCCN3 analyses of the scattered population epidemics of scenario 2a.

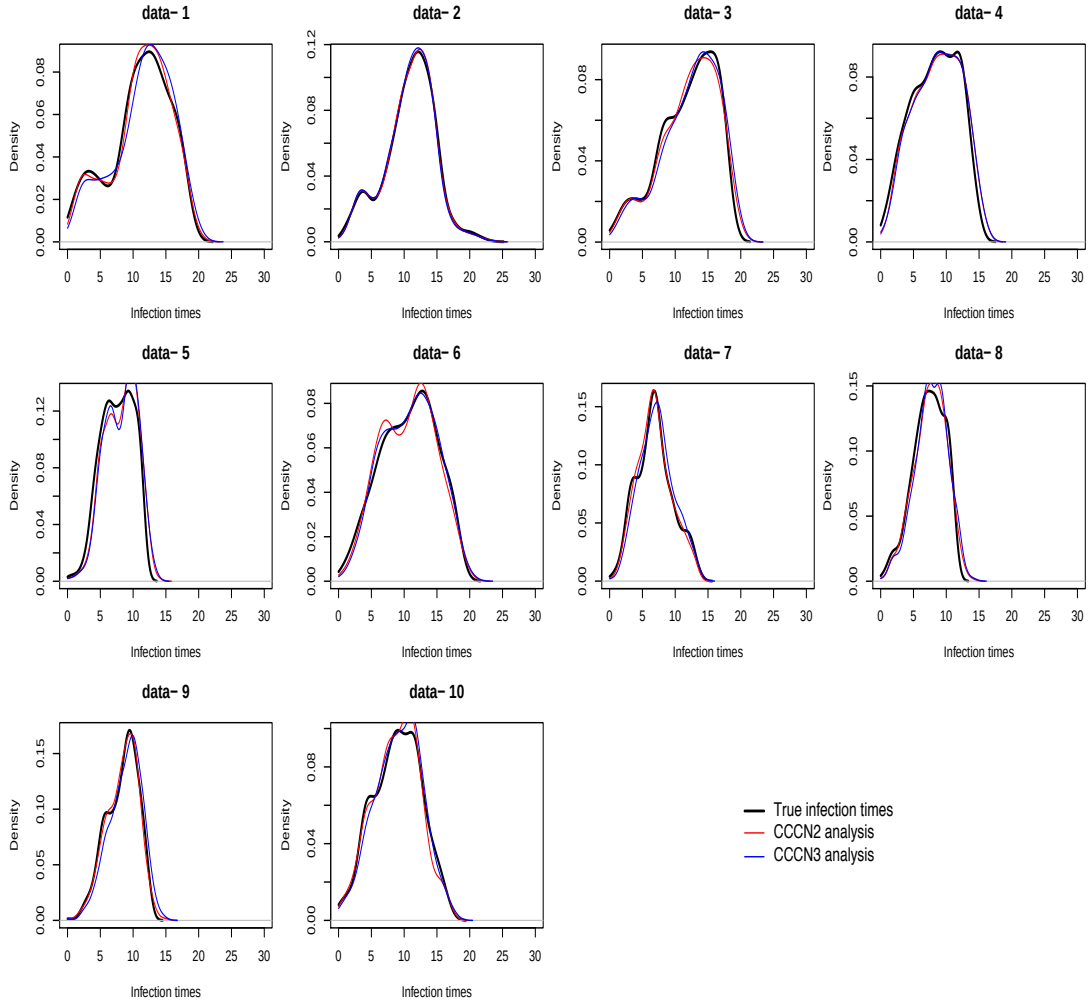


Figure C.2.3: The density of the posterior means of the infection times for the CCCN2 and CCCN3 analyses of the scattered population epidemics of scenario 2b.

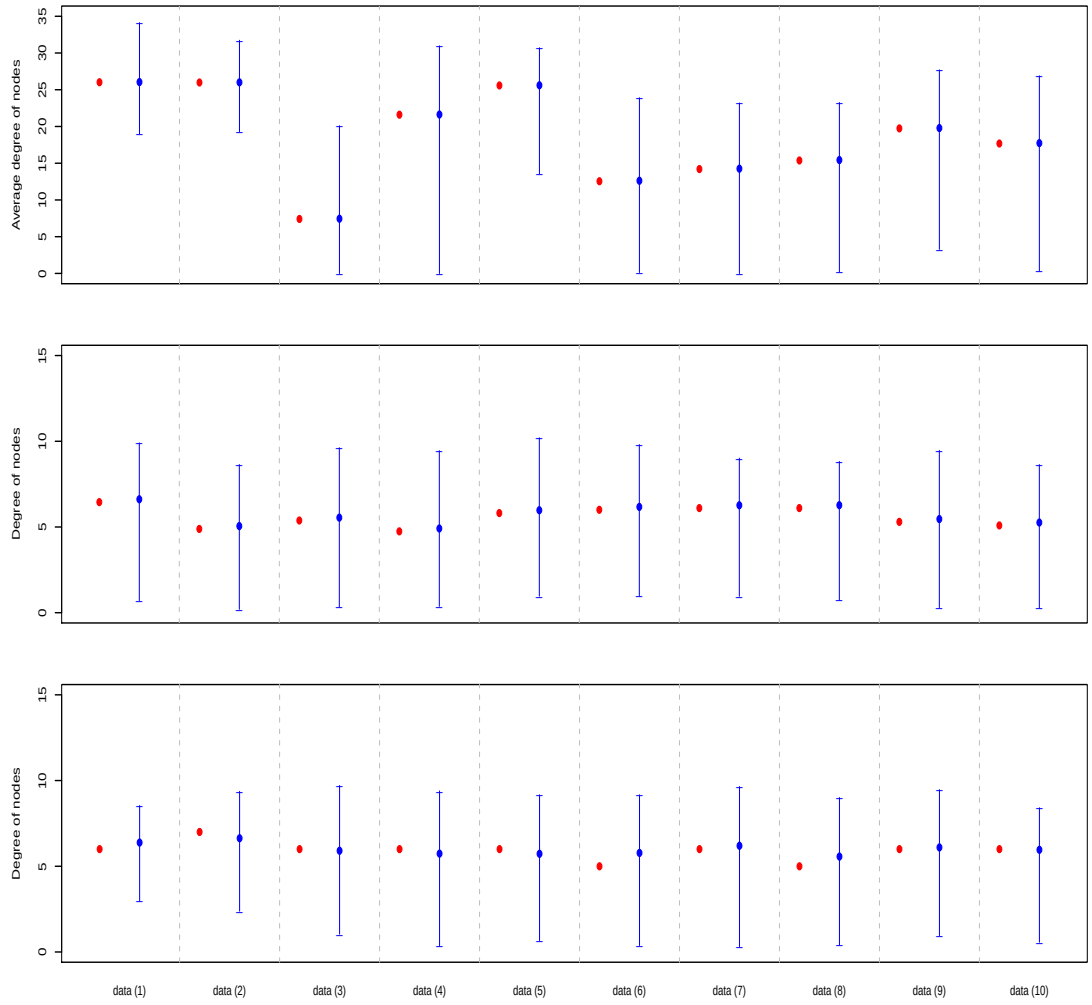


Figure C.2.4: Posterior means and 95% of the degree nodes for the simple within-cluster network-based ILM of the well-clustered populations (top), scattered populations of scenario 2a (middle), and scattered populations of scenario 2b (bottom) based on the full Bayesian MCMC under the CCCN3 analysis. Red points represent the average degree nodes of the true contact network.

C.3 A Comparison of the Performance of the Original and Modified LISA.

Here, we show the poor performance of implementing the original LISA algorithm for analyzing the third epidemic data set (from Scenario 2a) using the simple within-cluster network-based ILM. The assigned prior distributions for the model parameters are the same as those used in Section 5.3.4.

Figure C.3.1 shows the posterior densities of the model parameters for implementing the modified and original LISA performances in the CCCN1 analysis (fully observed epidemic). We observed a bias estimate of the covariate effect α_1 with a small posterior variance (narrower than the modified LISA) that did not include its true value. The posterior densities of the other parameters under the original LISA approach were narrower than those under the modification LISA.

The original LISA becomes more inefficient with the infection times uncertainty (CCCN2) analysis. Poor MCMC mixing chains were observed, especially for α_1 and the infectious period rate δ. Figures C.3.2 and C.3.3 show the MCMC chains of the model parameters. Those chains represented the MCMC chain of the posterior estimates of the model parameters before weighting them to approximate the full posterior. Along with the poor mixing issue, large bias estimates of the model parameters were observed for most of the clusters.

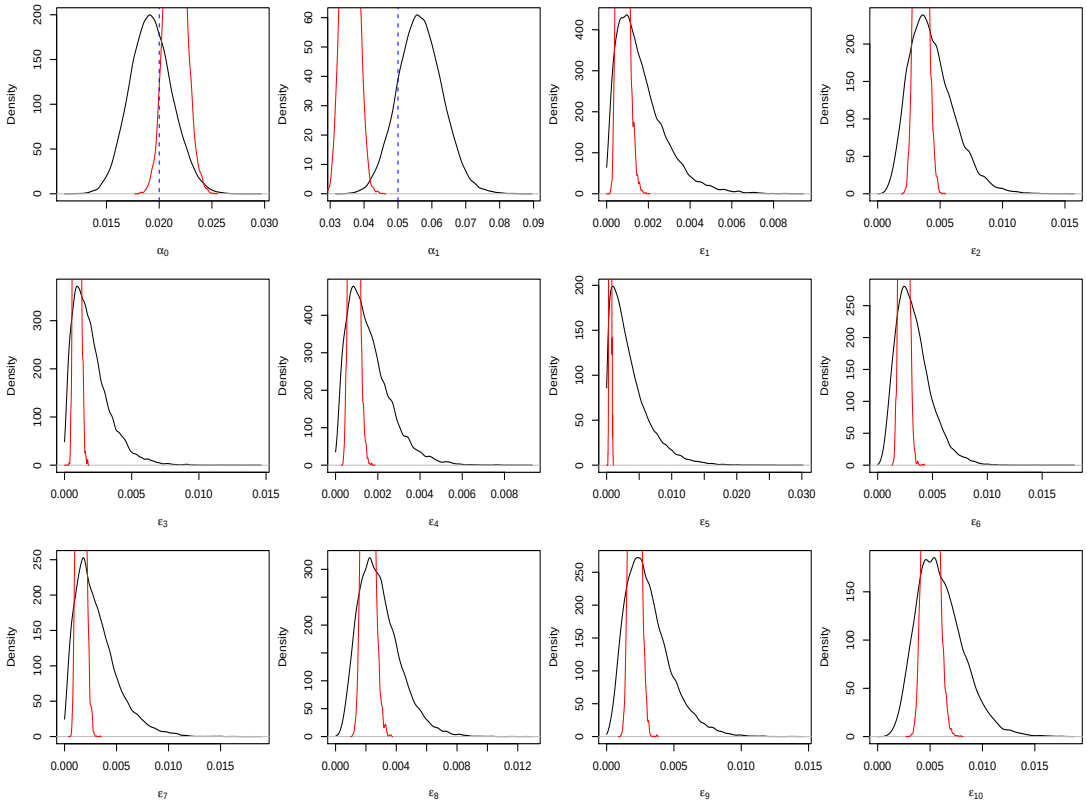


Figure C.3.1: Posterior densities of the model parameters for the simple within-cluster network-based ILMs of the third scattered population (scenario 2a), based on the full Bayesian MCMC analyses using the original (red) and modified (black) LISA approach.

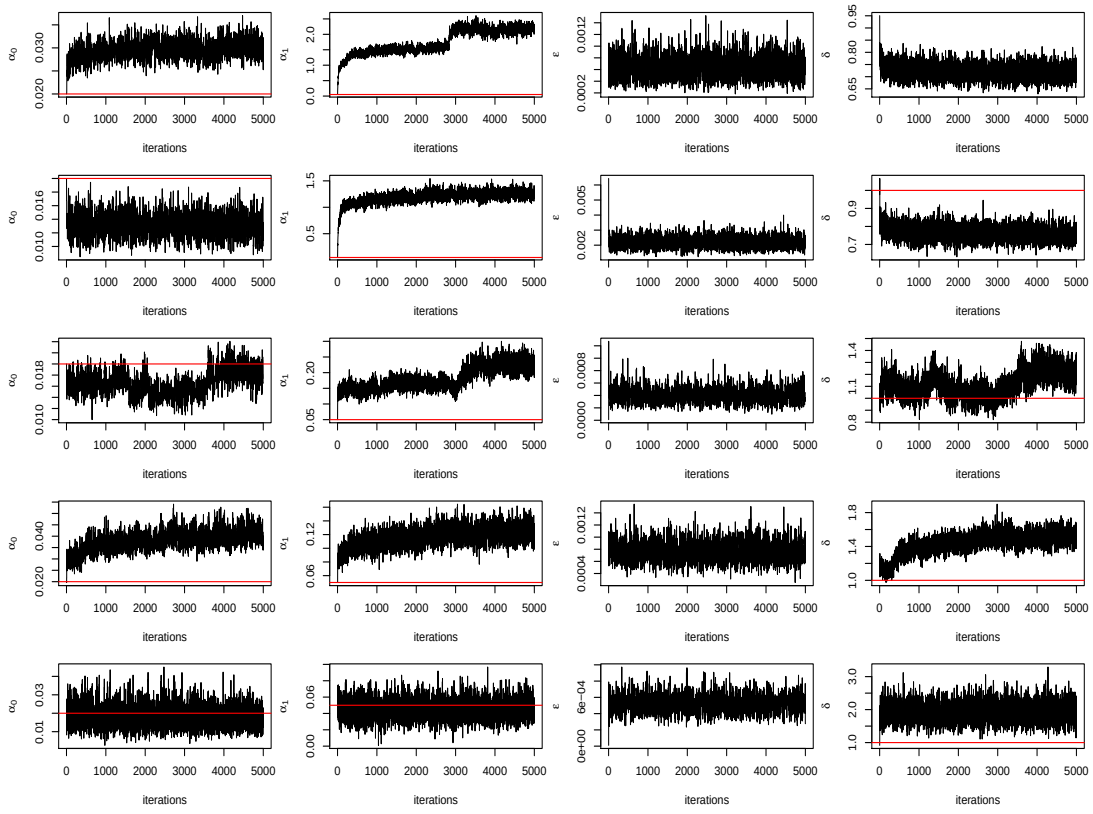


Figure C.3.2: The MCMC chains of the posterior distributions of the model parameters for the simple within-cluster network-based ILMs of the first five clusters of the third epidemic data set from scenario 2a under the CCCN2 analysis using the original LISA approach. Each row represents the results of the model parameters of each cluster.

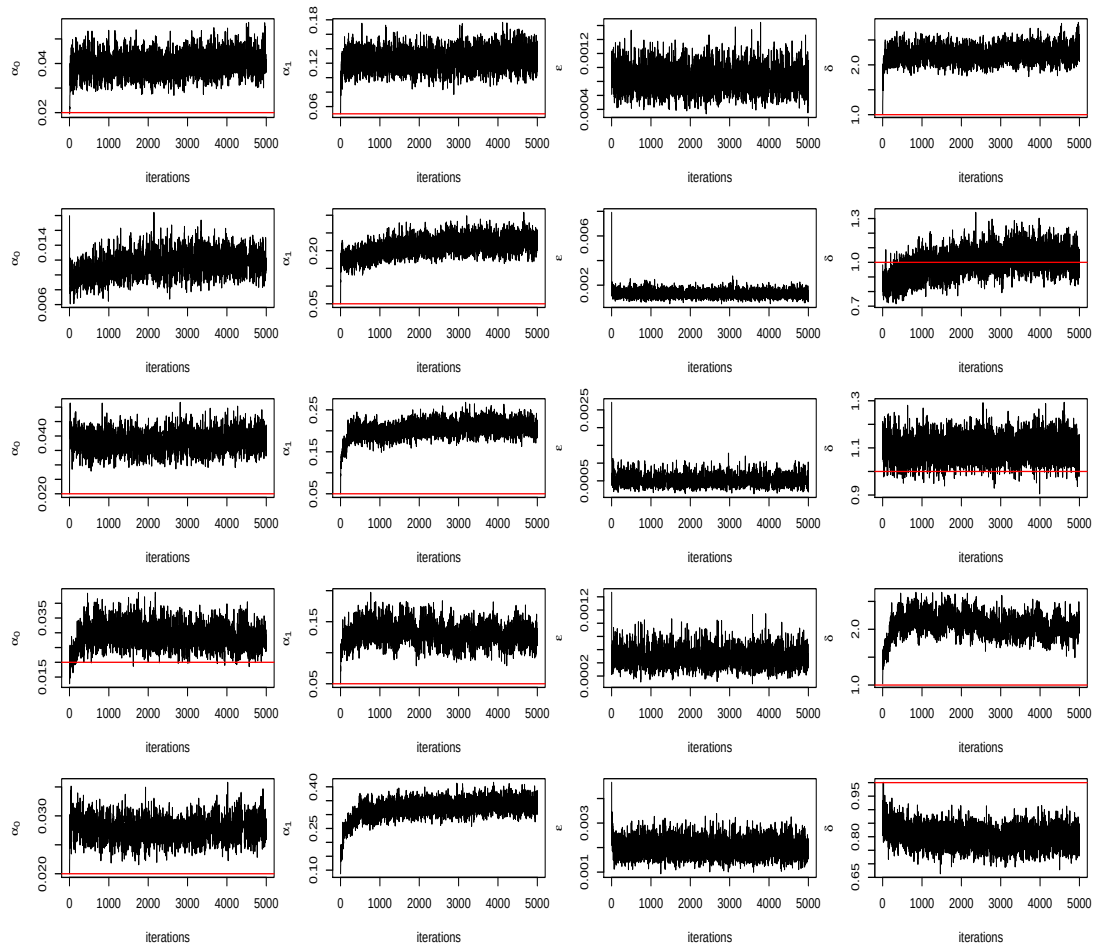


Figure C.3.3: The MCMC chains of the posterior distributions of the model parameters for the simple within-cluster network-based ILMs of the other five clusters of the third epidemic data set from scenario 2a under the CCCN2 analysis using the original LISA approach. Each row represents the results of the model parameters of each cluster.

C.4 The Effect of Increasing Number of Clusters

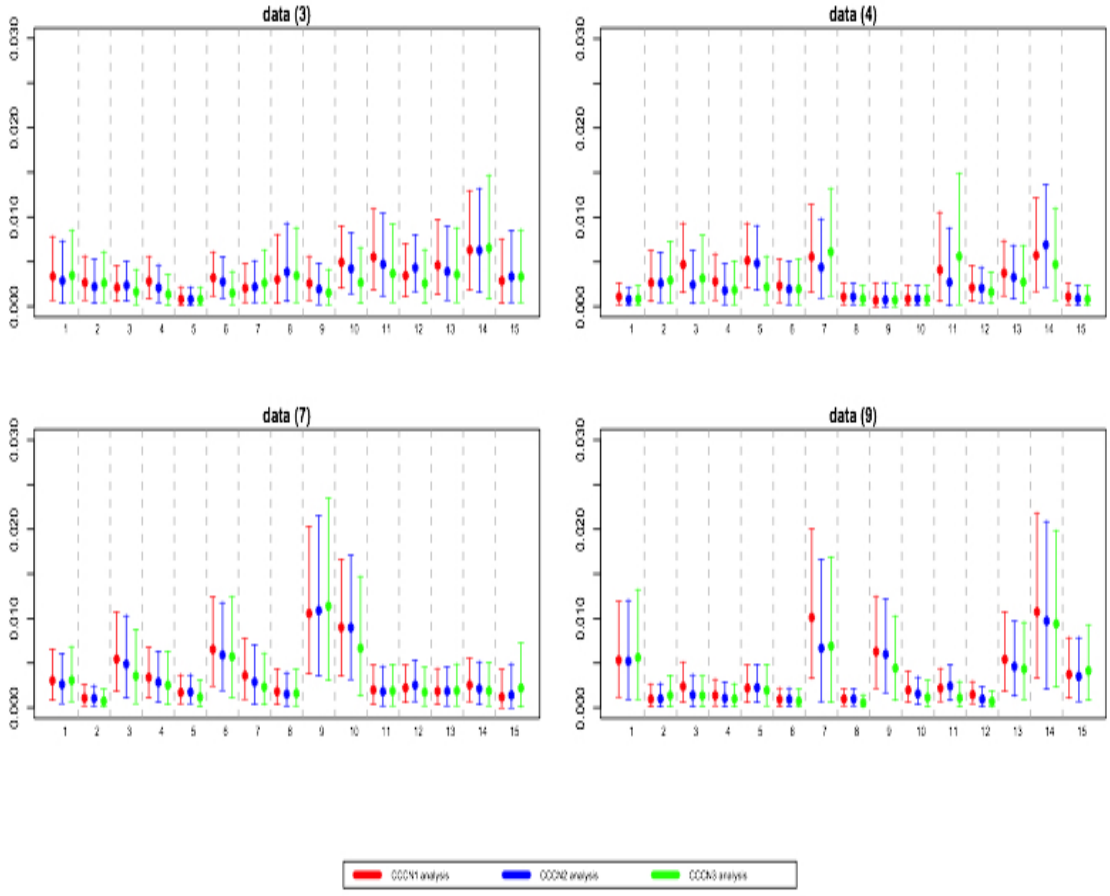


Figure C.4.1: Posterior means and 95% credible intervals of the spark infection parameters ϵ_k, $k = 1, \ldots, K$, for the simple within-cluster network-based ILMs of four epidemics, spreading on scattered populations (scenario 2a: with binary covariate) with 15 clusters, based on the full Bayesian MCMC analyses using our approach and the LISA algorithm.

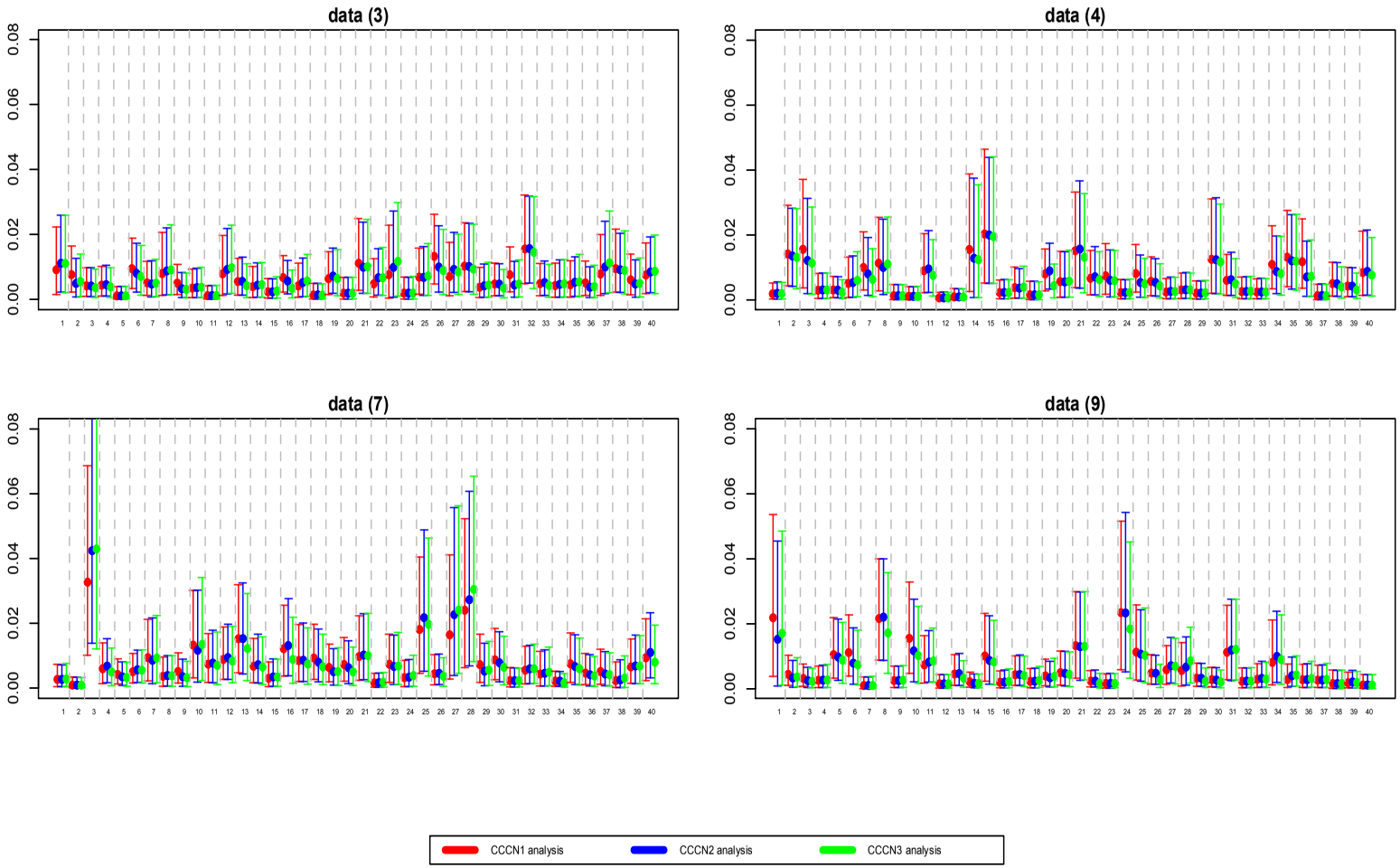


Figure C.4.2: Posterior means and 95% credible intervals of the spark infection parameters ϵ_k, $k = 1, \ldots, K$, for the simple within-cluster network-based ILMs of four epidemics, spreading on scattered populations (scenario 2.a: with binary covariate) with 40 clusters, based on the full Bayesian MCMC analyses using our approach and the LISA algorithm.