Bayesian Model Assessment for Stochastic Epidemic Models

Muteb Alharthi

Thesis submitted to The University of Nottingham
for the degree of Doctor of Philosophy
in Mathematics

December 2015

In memory of my father

To my mother with love and eternal appreciation

To my beloved wife

Abstract

A crucial practical advantage of infectious diseases modelling as a public health tool lies in its application to evaluate various disease-control policies. However, such evaluation is of limited use, unless a sufficiently accurate epidemic model is applied. If the model provides an adequate fit, it is possible to interpret parameter estimates, compare disease epidemics and implement control procedures. Methods to assess and compare stochastic epidemic models in a Bayesian framework are not well-established, particularly in epidemic settings with missing data.

In this thesis, we develop novel methods for both model adequacy and model choice for stochastic epidemic models. We work with continuous time epidemic models and assume that only case detection times of infected individuals are available, corresponding to removal times. Throughout, we illustrate our methods using both simulated outbreak data and real disease data. Data-augmented Markov Chain Monte Carlo (MCMC) algorithms are employed to make inference for unobserved infection times and model parameters.

Under a Bayesian framework, we first conduct a systematic investigation of three different but natural methods of model adequacy for SIR (Susceptible-Infective-Removed) epidemic models.

We proceed to develop a new two-stage method for assessing the adequacy of epidemic models. In this two-stage method, two predictive distributions are examined, namely the predictive distribution of the final size of the epidemic and the predictive distribution of the removal times. The idea is based on

looking explicitly at the discrepancy between the observed and predicted removal times using the posterior predictive model checking approach in which the notion of Bayesian residuals and the posterior predictive $p-$value are utilized. This approach differs, most importantly, from classical likelihood-based approaches by taking into account uncertainty in both model stochasticity and model parameters. The two-stage method explores how SIR models with different infection mechanisms, infectious periods and population structures can be assessed and distinguished given only a set of removal times.

In the last part of this thesis, we consider Bayesian model choice methods for epidemic models. We derive explicit forms for Bayes factors in two different epidemic settings, given complete epidemic data. Additionally, in the setting where the available data are partially observed, we extend the existing power posterior method for estimating Bayes factors to models incorporating missing data and successfully apply our missing-data extension of the power posterior method to various epidemic settings. We further consider the performance of the deviance information criterion (DIC) method to select between epidemic models.

Acknowledgements

First and foremost, all praise is due to Allah, the Lord of the worlds. I am truly grateful for the countless blessings you have bestowed upon me generally and in accomplishing this thesis especially.

I would like to express very special acknowledgement and appreciation to my supervisors, Prof. Philip O'Neill and Dr. Theodore Kypraios. I feel very lucky and proud to be their Ph.D. student. Thanks Phil for your valuable comments and indispensable guidance over the years. Sincere thanks to Theo, for all helpful comments, great support, and R programming advice, especially in the early days.

Thanks must go to my examiners, Prof. Frank Ball (Internal) and Prof. Nial Friel (External) for their comments and suggestions.

Huge heartfelt thanks to my three lovely daughters, Yasamiyan, Gorey and Jumanah. I love you more than anything and I am sorry for being away from you so much time.

I also owe special thanks to my brothers and sisters for encouraging me and believing in me and my work.

I am also grateful to all members of staff or postgraduates in the School of Mathematical Sciences at University of Nottingham whose work for, or interactions with, me have benefited my research in some way.

Finally, my research was supported by a scholarship from Taif University in Saudi Arabia.

Contents

List of Figures **x**

List of Tables **xxii**

1 Introduction **1**

1.1 Overview 1

1.2 Stochastic modelling of infectious disease data 3

1.2.1 SIR epidemic models in a homogeneous population . . . 3

1.2.2 Infectious disease data 6

1.3 Bayesian inference 7

1.4 Markov Chain Monte Carlo (MCMC) Methods 8

1.4.1 Metropolis-Hastings sampling algorithm 8

1.4.2 Gibbs sampler 9

1.4.3 Burn-in and thinning 9

1.5 Review of Statistical analysis for stochastic epidemic models . . 10

1.6 Review of techniques for Bayesian model adequacy and choice for epidemic models 14

1.6.1 Model adequacy 16

1.6.2 Model choice 18

1.7 Aims and structure of the thesis 21

2 Model Adequacy for the Markovian SIR Epidemic Model Based on Removal Data **24**

2.1 Introduction 24

2.2 Previous studies and relevant background 26

2.2.1 Previous work on model adequacy for epidemic models 26

2.2.2 Epidemic likelihood function 29

2.3 Inference and MCMC method for the SIR models 32

2.3.1 Bayesian inference and MCMC for the Markovian SIR model 33

2.4 Identifying the infectious period distribution using labelled individuals representation 37

2.4.1 Analysis of simulated data 39

2.5 Assessing the Markovian SIR model goodness-of-fit via time-rescaling 42

2.5.1 Analysis of completely observed epidemic data 43

2.6 Martingale goodness-of-fit statistics for the SIR epidemic models 48

2.6.1 Goodness-of-fit statistics 49

2.6.2 Martingale methods for the Markovian SIR model 50

2.6.3 Simulated outbreak data example 53

2.7 Summary and extensions 56

3 Two-stage Procedure for Assessing the Goodness-of-Fit for Epidemic Models Based on Partially Observed Data **59**

3.1 Introduction 59

3.2 Relevant background and previous related work 61

3.2.1 The stochastic SIR model with non-linear infection rate . 62

3.2.2 The stochastic SIR model with two levels of mixing . . . 63
3.2.3 Posterior predictive model checking 64
3.2.4 The two-stage procedure and the aims of current work . 69
3.3 Model assessment for SIR models with different infection mechanisms 72
3.3.1 Bayesian inference via MCMC for the SIR model with non-linear infection rate given removal data 73
3.3.2 Removal trajectories and their predictive distribution . . 77
3.3.3 Application of the two-stage method to simulated outbreak data and results 80
3.4 Model assessment for SIR models with different infectious periods 96
3.4.1 The behaviour of removal curves for the three models . 96
3.4.2 Application of the two-stage method to simulated outbreak data and results 99
3.4.3 Application of the two-stage method to Abakaliki smallpox data 110
3.5 Model assessment for epidemic models featuring a structured population 122
3.5.1 Bayesian inference via MCMC for the SIR model with two levels of mixing given removal data 123
3.5.2 Threshold and removal trajectories behaviour 127
3.5.3 Application of the two-stage method to simulated outbreak data and results 131
3.6 Summary and extensions 145
3.6.1 Summary 145
3.6.2 Extensions 147

4 Bayesian Model Choice for Epidemic Models Based on Removal Data 149

4.1 Introduction 149

4.2 Background and previous work 151

4.2.1 The marginal likelihood and Bayes factor 151

4.2.2 The deviance information criterion (DIC) 152

4.2.3 Previous work on model choice for epidemic models . . 154

4.2.4 Marginal likelihood estimation via the arithmetic mean method 157

4.2.5 Marginal likelihood estimation via the power posterior method 158

4.2.6 Power posteriors for models incorporating missing data 160

4.2.7 Objectives of current research 164

4.3 Model selection via Bayes factors for complete epidemic data . 165

4.3.1 The SIR model with different infectious periods: complete data case 165

4.3.2 The SIR model with different infection mechanisms: complete data case 168

4.3.3 Approximating the Bayes factor for epidemic models using the prior arithmetic mean estimator 171

4.4 Model selection via the updated power posterior and DIC methods for incomplete epidemic data 174

4.4.1 The updated power posterior and DIC_6 methods for epidemic models 174

4.4.2 Application of the updated power posterior and DIC_6 methods to the SIR model with different infectious periods 175

4.4.3 Application of the updated power posterior and DIC_6 methods to the SIR model with different infection mechanisms 183

4.5 Epidemic model choice: aspects where epidemic size matters . . 190

4.5.1 The SIR model with different infectious periods 191

4.5.2 The SIR model with different infection mechanisms . . . 191

4.6 Summary and extensions 196

4.6.1 Summary 196

4.6.2 Extensions 198

5 Conclusions **199**

Bibliography **203**

List of Figures

2.1 The posterior distribution of $\gamma \sum_{j=1}^{n_R}(R_j - I_j)$ overlaid by a Gamma$(44, 1)$ (left), and the posterior distribution of $\delta \sum_{j=1}^{n_R}(R_j - I_j)$ overlaid by a Gamma$(440, 1)$ (right) obtained from fitting m_1 and m_2 to the observed exponential SIR removal data. 40

2.2 The posterior distribution of $\gamma \sum_{j=1}^{n_R}(R_j - I_j)$ overlaid by a Gamma$(45, 1)$ (left), and the posterior distribution of $\delta \sum_{j=1}^{n_R}(R_j - I_j)$ overlaid by a Gamma$(450, 1)$ (right) obtained from fitting m_1 and m_2 to the observed gamma SIR removal data. 41

2.3 Sample traces for 5000 posterior samples of β and $1/\gamma$ obtained from fitting the Markovain SIR model to the three datasets, after a burn-in period of 2000 iterations and thinning by taking every $5th$ sample. Top, middle and lower panels correspond to datasets generated from the Markovian, gamma and constant SIR models respectively. 45

2.4 $(a), (b)$ and (c) correspond respectively to the posterior distributions of $\sum_{j=2}^{2n_R} \tau_j$ obtained from fitting the Markovian SIR model to the epidemic datasets generated from the Markovian, gamma and constant SIR models. The curved line indicates a Gamma$(183, 1)$ density function. 46

2.5 $(a), (b)$ and (c) correspond respectively to the posterior distributions of $\sum_{j=2}^{2n_R} \tau_j$ obtained from fitting the Markovian SIR model to the epidemic datasets generated from the Markovian, gamma and constant SIR models. The curved line indicates a Gamma$(949, 1)$ density function. 47

2.6 Boxplots of 100 posterior predictive $p-$values produced using the test statistic T_β (top panel) and boxplots of 100 posterior predictive $p-$values produced using the test statistic T_γ (bottom panel) under fitting all three SIR models to dataset D_1. . . . 55

2.7 Boxplots of 100 posterior predictive $p-$values produced using the test statistics T_β (left) and T_γ (right) under fitting all three models to D_2. 56

3.1 Scatter plots for p and β obtained from their posterior samples for different population sizes. 76

3.2 Posterior densities for the model parameters with various population size. Top, middle and bottom panels correspond to $\boldsymbol{N} = 70, 200$ and 500. 77

3.3 The distribution of 5000 removal curves drawn from each SIR model in a closed population of size $\boldsymbol{N} = 500$ with parameter values: $\beta = 0.45, \gamma = 0.2$ for the model m_1 (top) and $\beta = 3, \gamma = 0.2$ and $p = 0.3$ for the model m_2 (bottom). 79

3.4 The averages of 1000 realisations of removal curves obtained from each SIR model conditioning on particular final sizes, 90 and 283 (left to right), where the solid and dashed lines correspond to m_1 and m_2 respectively. Left and right plots refer to simulations $\boldsymbol{I}$ and $\boldsymbol{II}$ respectively. 80

3.5 The posterior predictive distribution of the removal times for the two SIR models (left and right plots refer to m_1 and m_2 models respectively) based on 5000 samples from the joint posterior density of each model parameters, where dotted line indicates the observed data. 82

3.6 The final size posterior predictive distributions for the two models m_1 and m_2 (left to right), where the dashed vertical line indicates the observed final size ($n_R - 1 = 177$). 83

3.7 Predictive removal curves mean (red solid line) with 95% percentile intervals (red dashed lines) and the observed removal curve (black dotted line) for the two SIR models m_1 and m_2 (left to right) calculated based on 500 predicted removal curves generated from each model under the condition of the same epidemic final size. 83

3.8 The Bayesian residual densities for the two SIR models (m_1 and m_2) resulting from calculating (3.3.7) for each SIR model based on the average of 500 removal curves from each model. 85

3.9 Scatter plots of observed versus replicated discrepancies based on 500 predicted removal curves from each removal times predictive distributions of the two SIR models (top and bottom rows refer to m_1 and m_2 respectively). The ppp-values are estimated by the proportion of points above the 45° line. Left and right columns correspond to the χ^2 and Freeman-Tukey discrepancy measures. 87

3.10 The posterior predictive distribution of the removal times for the two SIR models (left and right plots refer to m_1 and m_2 models respectively) based on 5000 samples from the joint posterior density of each model parameters, where dotted line indicates the observed data. 88

3.11 The final size posterior predictive distributions for the two models m_1 and m_2 (left to right), where the dashed vertical line indicates the observed final size ($n_R - 1 = 186$). 88

3.12 Predictive removal curves mean (red solid line) with 95% percentile intervals (red dashed lines) and the observed removal curve (black dotted line) for the two SIR models m_1 and m_2 (left to right) calculated based on 500 predicted removal curves generated from each model under the condition of the same epidemic final size. 89

3.13 The posterior residual densities for the two SIR models resulting from calculating (3.3.7) for each SIR model based on 500 samples from the conditioning predictive distribution of the removal times for the two models. 90

3.14 Four scatter plots of observed versus replicated discrepancies based on 500 predicted removal curves from each removal times predictive distributions of the two SIR models. The ppp-values are estimated by the proportion of points above the 45° line. Left and right columns correspond to the χ^2 and Freeman-Tukey discrepancy measures. Top and bottom rows refer to m_1 and m_2 respectively. 91

3.15 The posterior predictive distribution of the removal times where the dotted line indicates the observed data. Left and right plots refer to m_1 and m_2 models respectively. Top, middle and bottom rows refer to datasets A, B and C respectively. 93

3.16 The final size posterior predictive distributions for models m_1 and m_2 (left to right), where the dashed vertical line indicates the observed final size. Top, middle and bottom rows refer to datasets A, B and C respectively. 94

3.17 Top, middle and bottom rows refer to datasets A, B and C respectively. Left and middle columns represent the mean of 500 predicted removal curves with 95% percentile intervals and the observed removal curve (black dotted line) for models m_1 and m_2 respectively. Right column shows the posterior residual densities for the two SIR models. 95

3.18 5000 removal curve realisations simulated from the three epidemic models in a closed population of size $\boldsymbol{N} = 200$ using the same mean infectious period and the same infection rate, namely $E(T_I) = 1$ and $\beta = 3$. 98

3.19 The averages of 1000 realisations of removal curves obtained from each SIR model conditioning on particular final sizes, 39 and 94 (left to right). Solid, dashed and dotted lines represent the averages obtained from models m_1, m_2 and m_3 respectively. Left and right plots correspond to simulations $\boldsymbol{I}$ and $\boldsymbol{II}$. 99

3.20 The posterior predictive distribution of the removal times for the three SIR models produced based on 5000 samples from the joint posterior density of each model parameters, where the dotted line indicates the observed data. 101

3.21 The final size posterior predictive distribution for the three SIR models where dotted line indicates the observed value of the final size ($n_R - 1 = 89$). 102

3.22 Predictive removal curves mean (red solid line) with 95% percentile intervals (red dashed lines) and the observed removal curve (black dotted line) for the three SIR models calculated based on 500 predicted removal curves generated from each model under the condition of the same epidemic final size. . . . 102

3.23 Bayesian residual densities for the three SIR models resulting from calculating (3.4.1) for each SIR model based on 500 posterior samples from the conditioning predictive distribution of removal times for the three models. 103

3.24 Scatter plots of observed versus replicated discrepancies based on 500 predicted removal curves from each removal times predictive distributions of the three SIR models. Left and right columns correspond to the χ^2 and Freeman-Tukey discrepancy measures. Top, middle and bottom rows refer to m_1, m_2 and m_3 respectively. 104

3.25 The posterior predictive distribution of the removal times for the three SIR models produced based on 5000 samples from the joint posterior density of each model parameters, where the dotted line indicates the observed data. 106

3.26 The final size posterior predictive distribution for the three SIR models where dotted line indicates the observed value of the final size ($n_R - 1 = 44$). 107

3.27 Predictive removal curves mean (red solid line) with 95% percentile intervals (red dashed lines) and the observed removal curve (black dotted line) for the three SIR models calculated based on 500 predicted removal curves generated from each model under the condition of the same epidemic final size. . . . 107

3.28 Bayesian residual densities for the three SIR models resulting from calculating (3.4.1) for each SIR model based on 500 posterior samples from the conditioning predictive distribution of removal times for the three models. 108

3.29 Scatter plots of observed versus replicated discrepancies based on 500 predicted removal curves from each removal times predictive distributions of the three SIR models. Left and right columns correspond to the χ^2 and Freeman-Tukey discrepancy measures. Top, middle and bottom rows refer to m_1, m_2 and m_3 respectively. 109

3.30 Top and middle rows show the removal times and final size predictive distributions for the exponential, gamma and constant SIR models calculated based on 5000 posterior samples of each model parameters, where the dotted line indicates the observed removal times and the vertical dashed line represents the observed final size ($n_R - 1 = 29$). Bottom row refers to the mean of 500 predicted removal curves along with 95% percentile intervals generated from the removal times predictive distributions for the three SIR models in which the epidemic final sizes are matched with the observed one. 113

3.31 Bayesian residual densities for the three SIR models resulting from calculating (3.4.1) for each SIR model using 500 predicted removal curves. 114

3.32 Scatter plots of observed versus replicated discrepancies based on 500 predicted removal curves from each removal times predictive distributions of the three SIR models. Left and right columns correspond to the χ^2 and Freeman-Tukey discrepancy measures. Top, middle and bottom rows refer to m_1, m_2 and m_3 respectively. 115

3.33 The final size predictive distributions for the exponential, gamma and constant SEIR models calculated based on 5000 posterior samples of each model parameters, where the vertical dashed line represents the observed final size ($n_R - 1 = 29$). 117

3.34 The mean of 500 predicted removal curves along with 95% percentile intervals generated from the removal times predictive distributions for the three SEIR models in which the epidemic final sizes are matched with the observed one and the dotted line indicates the observed removal times. 118

3.35 The mean of 500 predicted removal curves along with 95% percentile intervals generated from the removal times predictive distributions for the exponential SEIR model (m_1) in which the epidemic final sizes are matched with the observed one and the crossed points indicate the observed removal times. For the sake of clarity, we plot only the early observed and predicted removal times. 119

3.36 Bayesian residual densities for the three SEIR models resulting from calculating (3.4.1) for each SEIR model using 500 predicted removal curves. 120

3.37 Scatter plots of observed versus replicated discrepancies based on 500 predicted removal curves from each removal times predictive distributions of the three SEIR models. Left and right columns correspond to the χ^2 and Freeman-Tukey discrepancy measures. Top, middle and bottom rows refer to m_1, m_2 and m_3 respectively. 121

3.38 The distributions of 1000 removal curves simulated from the two levels of mixing model (m_2) for different values of λ_H. The community infection rate parameter and the length of infectious period are fixed to $\lambda_C = 0.1$ and $c = 5$ throughout simulations. 129

3.39 Histograms of the final size of 1000 epidemics simulated from the two epidemic models. The red and blue shaded histograms refer to models m_1 and m_2 respectively. The parameter values used are : $\beta = 0.55, \lambda_H = 1.4, \lambda_C = 0.12$ and $c = 5$. 130

3.40 The distributions of 1000 removal curves simulated from the two epidemic models with the same length of infectious period ($c = 5$). The other parameter values used are : $\beta = 0.55, \lambda_H = 1.4$ and $\lambda_C = 0.12$. 130

3.41 Posterior sample traces and histograms of λ_H and λ_C obtained from fitting m_2 to the observed data after a burn-in period of 2000 iterations and a thinning of 5 samples. True parameter values were: $\lambda_H = 0.9$ and $\lambda_C = 0.1$. 133

3.42 Top and bottom rows show the removal times and final size predictive distributions for the two models respectively, calculated based on 5000 posterior samples of each model parameters, where the dotted line indicates the observed removal times and the vertical dashed line represents the observed final size ($n_R - 1 = 89$). Left and right columns correspond to m_1 and m_2 respectively. 134

3.43 Predictive mean of 500 removal curves (red solid line) with 95% percentile intervals (red dashed lines) and the observed removal curve (black dotted line) for the two models m_1 and m_2 (left to right). 135

3.44 Posterior residual densities for the two models resulting from calculating (3.5.7) for each epidemic model based on the average of 500 curves generated from the conditioning predictive distribution of removal times for both models. 136

3.45 Scatter plots of observed versus replicated discrepancies based on 500 predicted removal curves from each removal times predictive distributions of the two models. The ppp-values are estimated by the proportion of points above the 45° line. Left and right columns correspond to the χ^2 and Freeman-Tukey discrepancy measures. Top and bottom rows refer to m_1 and m_2 respectively. 138

3.46 Posterior sample traces (top row) and histograms (bottom row) of the number of household and community infections (left to right) obtained by fitting m_2 to the observed data. 139

3.47 Posterior sample traces and histograms of λ_H and λ_C, obtained from fitting m_2 to the observed data, after a burn-in period of 2000 iterations and a thinning of 5 samples. 140

3.48 The removal times and final size posterior predictive distributions for the two models m_1 and m_2 (left to right), where the dotted and dashed lines indicate the observed removal curve and the observed final size respectively. 141

3.49 Predictive removal times mean (red solid line) with 95% percentile intervals (red dashed lines) and the observed removal curve (black dotted line) for the two models m_1 and m_2 (left to right). 142

3.50 Posterior residual densities for the two models resulting from calculating (3.5.7) for each epidemic model based on the average of 500 curves generated from the conditioning predictive distribution of removal times for both models. 142

3.51 Scatter plots of observed versus replicated discrepancies of the χ^2 and Freeman-Tukey (left to right column) based on 500 predicted removal curves from each removal times predictive distributions of the two models. Top and bottom rows refer to m_1 and m_2 respectively. 143

3.52 Posterior sample traces (top row) and histograms (bottom row) of the number of household and community infections (left to right) obtained by fitting m_2 to the observed data. 144

4.1 Values of the logarithm of Bayes factor, $\log(BF_{12})$, calculated using (4.3.2) by varying ν (γ and δ prior scale parameter) and setting $\lambda = 1$. The horizontal dashed line denotes the lower limit value of $\log(BF_{12})$ (3.1039) as $\nu \to 0$. 168

4.2 $\log(BF_{12})$ values calculated using (4.3.5), varying the prior parameter ν_β and setting $\lambda_\beta = 1$. The horizontal dashed and dotted lines represent the lower (3.8890) and upper (55.5122) limit values of $\log(BF_{12})$ respectively. 171

4.3 Top plot: the expected log-likelihood curves for the removal data generated under the SIR with $T_I \sim \text{Exp}(\gamma)$ (m_1) calculated using the updated power posterior approach when the prior for β, γ and δ is $\text{Exp}(1)$. Bottom plot: the expected log-likelihood curves for the removal data generated under the SIR with $T_I \sim \text{Gamma}(\alpha = 10, \delta)$ (m_2) calculated using the updated power posterior approach with $\text{Exp}(1)$ prior for β, γ and δ. 182

4.4 Top plot: the expected log-likelihood curves for the removal data generated under the standard SIR (m_1) calculated using the updated power posterior approach when the prior is $\text{Exp}(1)$ for β and γ. Bottom plot: the expected log-likelihood curves for the removal data generated under the modified SIR (m_2) calculated using the updated power posterior approach with $\text{Exp}(1)$ prior for β and γ and $U(0, 0.5)$ for p. 189

4.5 Boxplots of $\log(BF_{12})$ values calculated based on 20 outbreak datasets of size 5 removals each simulated under the SIR model with $T_I \sim \text{Exp}(\gamma)$ (m_1). $\log(BF_{12})$ was computed via (4.3.2) and the updated power posterior method for complete and incomplete data respectively. The parameters β, γ and δ were assigned two choices of prior, namely $\text{Exp}(1)$ and $\text{Exp}(0.01)$. . . . 192

4.6 Boxplots of $\log(BF_{12})$ values calculated using 20 outbreak datasets of size 7 removals each simulated under the standard SIR model (m_1) and evaluated with setting $p = 0.3$ and assigning $\text{Exp}(1)$ and $\text{Exp}(0.01)$ priors for β and γ. $\log(BF_{12})$ values were obtained via (4.3.5) and the updated power posterior method for complete and incomplete data respectively. 193

4.7 Boxplots of $\log(BF_{12})$ values calculated using 20 outbreak datasets of size 42 removals each simulated under the standard SIR model (m_1) and evaluated with setting $p = 0.3$ and assigning $\text{Exp}(1)$ and $\text{Exp}(0.01)$ priors for β and γ. $\log(BF_{12})$ values were obtained via (4.3.5) and the updated power posterior method for complete and incomplete data respectively. 194

4.8 Histogram of the estimated $\log(BF_{12})$ values obtained via the updated power posterior approach based on 20 outbreak datasets of size 47 removal times each simulated under the standard SIR model (m_1) and evaluated with setting $p = 0.3$ and assigning an $\text{Exp}(1)$ prior for β and γ. 195

List of Tables

3.1 Posterior mean and variance for β, γ and p as population size varies. True values are: $\beta = 3$, $\gamma = 0.2$ and $p = 0.3$ 76

3.2 Summary of the final size of each epidemic and the parameter values used for both epidemic models, m_1 and m_2. 78

3.3 Values of the infection rate parameter β and the basic reproduction number R_0 used to simulate (from model m_1) the three datasets A, B and C along with the ratio of the final size to the population size ($\boldsymbol{F/N}$). Results of sum of squared residuals (SSR) and the posterior predictive $p-$values obtained from fitting both SIR models, m_1 and m_2, to each dataset using both the χ^2 and Freeman-Tukey (F-T) discrepancy measures. 92

3.4 Summary of the simulation study used to obtained Figure 3.19. 97

3.5 Summary of different values for λ_H with the corresponding $E(F_H)$ and R_* values used to produce Figure 3.38. 128

4.1 Interpretation of the Bayes factor on the log scale. 152

4.2 The logarithm of Bayes factor results obtained using two datasets simulated based on models m_1 (the SIR with $T_I \sim \text{Exp}(\gamma)$) and m_2 (the SIR with $T_I \sim \text{Gamma}(\alpha, \delta)$). For each case, model parameters were assigned prior distributions to varying degrees of diffuseness to determine the effect on the logarithm of Bayes factor estimation. Three sizes of prior parameter sample ($M = 10^4, M = 10^5$ and $M = 10^6$) were used to determine the effective sample size required to reasonably estimate the logarithm of Bayes factor. 173

4.3 Results of $\log(\pi(\boldsymbol{R}|m_1))$, $\log(\pi(\boldsymbol{R}|m_2))$ and $\log(BF_{12})$ using data simulated based on the SIR model with $T_I \sim \text{Exp}(\gamma)$ (m_1). The bias (the gap between the standard and updated power posterior estimates of the marginal likelihood) was recorded under different values of r. β, γ and δ were assigned $\text{Exp}(1)$ (top table) and $\text{Exp}(0.01)$ (bottom table) prior distributions and various choices of temperature ladder were applied to determine their effects on the accuracy of estimating $\log(BF_{12})$. 179

4.4 Results of $\log(BF_{12})$ and DIC_6 using two datasets simulated based on models m_1 (the SIR with $T_I \sim \text{Exp}(\gamma)$) and m_2 (the SIR with $T_I \sim \text{Gamma}(\alpha = 10, \delta)$). For each case, β, γ and δ were assigned prior distributions to varying degrees of diffuseness to determine the effect on estimating $\log(BF_{12})$ and DIC_6. . 181

4.5 Results of $\log(\pi(\boldsymbol{R}|m_1))$, $\log(\pi(\boldsymbol{R}|m_2))$, $\log(BF_{12})$ and the bias (the gap between the standard and updated power posterior estimates) using data simulated based on the modified SIR model (m_2). The parameters β and γ were assigned $\text{Exp}(1)$ (top table) and $\text{Exp}(0.01)$ (bottom table) prior distributions and various choices of temperature ladder were applied to determine their effects on the accuracy of estimating $\log(BF_{12})$. A non-informative prior $U(0, 0.5)$ was assumed for p. 187

4.6 Results of $\log(BF_{12})$ and DIC_6 using data simulated based on the standard SIR model (m_1) and the modified SIR model (m_2). For each case, the parameters β and γ were assigned prior distributions to varying degrees of diffuseness to determine the effect on estimating $\log(BF_{12})$ and DIC_6. A non-informative prior $U(0, 0.5)$ was assumed for p. 190

CHAPTER 1

Introduction

1.1 Overview

Mathematical infectious disease modelling has emerged as a research field of great importance, since infectious disease outbreaks have imposed serious risks to both public health and animals, as well as economic growth (Anderson and May, 1991). Mathematical models for infectious disease transmission can be broadly divided into two categories, namely deterministic models and stochastic models (O'Neill, 2010). Modelling epidemics deterministically assumes that given the initial conditions, the spread of the infection is determined; for instance, the epidemic will infect a given number of individuals with certainty. However, given the initial conditions, modelling epidemics stochastically assumes that there is a range of different possible outcomes, with associated probabilities. Also, the final outcome of the epidemic is a random variable, and hence it may be possible to calculate the probability that there will be a minor epidemic infecting only a few individuals, or that there will be a major epidemic infecting a reasonably large proportion of individuals. Throughout this thesis, we focus only on stochastic epidemic models.

The use of stochastic models in studying epidemics is well-developed. Bailey (1975) gives a useful summary of the early developments up until 1970s, in which several stochastic and deterministic models are covered, along with

their applications to data analysis. Then, Becker (1989) considers the statistical analysis of infectious disease data using the theory of stochastic processes and in particular, the theory of martingales. Later on, Andersson and Britton (2000) provides an introduction to stochastic modelling as well as a discussion of some basic statistical analysis for stochastic epidemic models. However, none of these references provides a generic procedure for the problem of model adequacy and model selection for stochastic epidemic models when fitting them to a set of temporal outbreak data.

A vital challenge when modelling stochastic infectious diseases is to develop effective and easy tools for model assessment. The existence of such methods would greatly improve understanding, prediction and management of diseases. However, neither statistical inference nor model assessment for epidemic models is straightforward. There are two main reasons why developing methods for model assessment for stochastic epidemic models is harder than in many other situations. First, infection events are not independent which means whether or not an individual becomes infected is influenced by whether or not other individuals get infected. Most standard model assessment tools are based on independent observations which is hence not applicable in our situation. The second reason is the partial observation of the epidemic process. For instance, it is usually unknown who infects whom, the infection times and the length of infectiousness. Instead we might observe just case-detection times, corresponding to removal times as we shall assume in this thesis. Nevertheless, making statistical inference based on such epidemic data and then moving to model assessment issue are indeed challenging tasks.

In this thesis we will investigate and develop novel methods for Bayesian model assessment for stochastic epidemic models. Both model adequacy and model choice for stochastic epidemic models will be considered. The different methods are illustrated using both real life and simulated outbreak data.

1.2 Stochastic modelling of infectious disease data

The spread of infectious diseases is often modelled by describing the transitions of individuals through various disease-development states and may be modelled in continuous or discrete time (Bailey (1975), Becker (1989) and Andersson and Britton (2000)). Early work in stochastic epidemic modelling usually assumes a homogeneously mixing community of individuals (Bailey, 1975). More sophisticated modelling introduces heterogeneity into the population being modelled. It can be through susceptibility and/or infectivity of individuals (as addressed in multi-type models, see for example (Andersson and Britton, 2000, section 6)), population structure (as handled in multiple-level mixing models such as models with two levels of mixing (Ball et al., 1997), and models with three levels of mixing (Britton et al., 2011)). Also, heterogeneities may be presented via contact structure (see for example, network models (O'Neill, 2002)) or via spatial structure (as in spatial models (Deardon et al., 2010)).

Most of the work in this thesis is based on the generalised stochastic Susceptible-Infective-Removed epidemic model that assumes a homogeneously mixing community of individuals, so we shall now define such epidemic model.

1.2.1 SIR epidemic models in a homogeneous population

The most well known stochastic model for the transmission of infectious diseases is considered, that is the SIR (Susceptible-Infective-Removed) stochastic epidemic model (Bailey, 1975). A closed population, which does not involve demographic changes (e.g. births, deaths), of size $\boldsymbol{N} = n + a$ is assumed, where n and a denote the initial number of susceptible and infective individuals respectively. We also assume that the population is homogeneous and randomly mixing which means the chance that two individuals contact each other does not depend on which pair is considered. The status of an individual at any given time $t \geq 0$ can be one of three states: susceptible, infectious or

removed. If individuals are in the susceptible state this means they do not have the disease, but can contract it; if in the infectious state individuals can transmit the disease in a manner described below; in the removed state individuals have been removed from the population, possibly through recovery, immunity, or death and play no further part in the spread of the epidemic. Additionally, each infective individual remains infectious for a length of time, referred to as their infectious period. The infectious periods of infected individuals are assumed to be independent and identically distributed according to some random variable T_I, which can have any arbitrary but specified distribution. Also, an individual, during their infectious period, contacts any other individual at the time points generated by a homogeneous Poisson process with intensity βn^{-1}, and all these Poisson processes are mutually independent, where β is referred to as the infection rate. At each contact between an infective individual and a susceptible one, an infection is transmitted immediately. Ultimately, an individual enters the removal state once its infectious period is terminated and the epidemic is over as soon as there are no more infective individuals remaining in the population.

When the infectious periods follow an exponential distribution with rate γ, say, then the above model is known as the general stochastic epidemic model. We will refer to this model as the Markovain (or exponential) SIR model. In this case, denoting the number of susceptible and infective individuals at time t as $X(t)$ and $Y(t)$ respectively, the epidemic process $\{(X(t), Y(t)) : t \geq 0\}$ is a bivariate continuous time Markov chain with the following transition probabilities:

$$\begin{aligned} P[(X(t+\delta t), Y(t+\delta t)) &= (x-1, y+1) | (X(t), Y(t)) = (x, y)] \\ &= \beta n^{-1} xy \delta t + o(\delta t), \end{aligned}$$

$$\begin{aligned} P[(X(t+\delta t), Y(t+\delta t)) &= (x, y-1) | (X(t), Y(t)) = (x, y)] \\ &= \gamma y \delta t + o(\delta t). \end{aligned}$$

Hence, it follows that at time t, infections occur at rate $\beta n^{-1} X(t) Y(t)$ and re-

movals happen at rate $\gamma Y(t)$, where β and γ represent the infection and removal rate parameters respectively.

The assumption of exponentially distributed infectious periods is unrealistic for many diseases and used largely for mathematical convenience. It means that there is no typical length for infectious periods, most of them are short, a few are long and fewer are extremely long. Therefore, we need a more flexible distribution for the infectious periods. Three common choices for infectious period distribution that are widely used in the literature are the gamma distribution (O'Neill and Becker, 2001), the Weibull distribution (Streftaris and Gibson, 2004a) and the case in which the infectious periods are simply constant with the same length (Clancy and O'Neill, 2008).

1.2.1.1 The final size distribution and the basic reproduction number

In the epidemic context, two extremely important quantities of interest are the final size of epidemic, and the basic reproduction number which is often denoted by R_0. The final size of the epidemic is defined as the number of initially susceptible individuals that ultimately become infected. Thus, it is a random variable, so that it follows a distribution which can be determined as follows. Let P_k be the probability that the final size of the epidemic is equal to $k, 0 \leq k \leq n$ and define $\phi(s) = E[e^{-sT_I}],\ s \geq 0$ to be the moment generating function of the infectious period T_I. Then, for $0 \leq l \leq n$, Ball (1986) proves elegantly that

$$\sum_{k=0}^{l} \frac{\binom{n-k}{l-k} P_k}{\left[\phi\left(\frac{\beta(n-l)}{n}\right)\right]^{k+a}} = \binom{n}{l}. \tag{1.2.1}$$

By solving this triangular system of equations, we obtain the final size probabilities recursively. Due to rounding errors, numerical solutions of (1.2.1) can lead to negative probability values. Demiris and O'Neill (2006) employed multiple precision arithmetic to avoid the numerical problems. In the special case where T_I has an exponential distribution, Laplace transform methods applied to the forward equations of the Markovian epidemic process may be

used to derive a different system of equations for the final size probabilities that has better numerical stability than (1.2.1) (Bailey, 1975, p. 93).

For homogeneously mixing models, the basic reproduction number R_0 can be roughly specified as the expected number of infections generated by one infectious individual in a large susceptible population (Andersson and Britton, 2000, p. 11). Hence, for the Markovian SIR model, this quantity is defined as $R_0 = \beta\gamma^{-1}$, whereas for other SIR stochastic epidemic model R_0 has a simple form, namely $R_0 = \beta E(T_I)$. The importance of this measure comes from its relation with the spread of epidemic. In other words, in an infinitely large population, if $R_0 \leq 1$, then with probability one, only a finite number of individuals will become infected, i.e. the epidemic will die out, whereas when $R_0 > 1$ there is a positive probability that the epidemic will take off. Consequently and more importantly, knowing the value of R_0 enables us to compute the proportion of a population that should be vaccinated in order to prevent an epidemic from spreading out (Andersson and Britton, 2000). That is to say, vaccinating a fraction $v_c = 1 - 1/R_0$ (critical vaccination coverage with perfect vaccine) of the community chosen uniformly at random is sufficient to prevent a large outbreak.

1.2.2 Infectious disease data

Although the details of outbreak data vary considerably from study to study, broadly speaking there are three types of epidemic data, namely complete temporal data, partial temporal data, and final size data (Britton et al., 2011). The complete temporal data consist of the time of infection and time of removal of all infected individuals in the population. The final size data do not contain explicit temporal information regarding the disease propagation throughout the population but instead consist of the initial number of susceptible individuals and which of these individuals were ultimately infected during the course of the epidemic. The partial temporal data are less detailed than complete temporal data, but more detailed than final size data. This type

of data is more common in practice and typically consists of removal times. In this thesis, we work with such type of data (i.e. removal times), unless otherwise stated.

1.3 Bayesian inference

Bayesian inference (Bernardo and Smith, 1994) is a modern statistical technique for parameter estimation for a model given data. The model must permit a likelihood and the parameters require a prior distribution. The likelihood and prior distribution are combined, using Bayes'theorem, to compute the posterior distribution of the parameters given the data.

Suppose we are concerned with making inference on a set of k unknown quantities, $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)$. Information from the data $\boldsymbol{y}$ is contained in the likelihood $\pi(\boldsymbol{y}|\boldsymbol{\theta})$ which is simply the joint density function (or mass function when $\boldsymbol{y}$ is discrete) of the data given the unknown quantities $\boldsymbol{\theta}$. Unlike classical likelihood inference, prior belief about the unknown quantities can be incorporated into the model via the prior distribution with density $\pi(\boldsymbol{\theta})$. The posterior density $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ contains the information from the prior, updated by the likelihood, and follows from Bayes'theorem as

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\pi(\boldsymbol{y}|\boldsymbol{\theta})\ \pi(\boldsymbol{\theta})}{\pi(\boldsymbol{y})} \propto \pi(\boldsymbol{y}|\boldsymbol{\theta})\ \pi(\boldsymbol{\theta}), \tag{1.3.1}$$

where $\pi(\boldsymbol{y}) = \int_{\boldsymbol{\theta}} \pi(\boldsymbol{y}|\boldsymbol{\theta})\ \pi(\boldsymbol{\theta})\ d\boldsymbol{\theta}$ if $\boldsymbol{\theta}$ continuous or $\pi(\boldsymbol{y}) = \sum_{\boldsymbol{\theta}} \pi(\boldsymbol{y}|\boldsymbol{\theta})\ \pi(\boldsymbol{\theta})$ if $\boldsymbol{\theta}$ discrete. This term, $\pi(\boldsymbol{y})$, is typically referred to as the normalising constant which is typically analytically intractable, especially in problems of more than a few dimensions. This quantity is vital in Bayesian model choice and has been given different names in the literature such as the marginal likelihood, the integrated likelihood or the evidence.

As can be seen from (1.3.1), Bayesian inference relies on the choice of prior density $\pi(\boldsymbol{\theta})$. Prior distributions may be chosen to reflect the existing knowledge and uncertainty about the parameters in question, and may be derived

using previous studies or expert knowledge. Alternatively, non-informative prior distributions are used when one requires estimates driven almost entirely by the data rather than prior beliefs, either due to lack of existing knowledge for parameter values, or to a desire to derive estimates based solely on the observed data.

1.4 Markov Chain Monte Carlo (MCMC) Methods

Markov chain Monte Carlo (MCMC) methods (Gilks et al. (1996) and Robert and Casella (1999)) are general computing techniques that enable samples to be drawn from some target density which is known only up to proportionality. In a Bayesian context, the target density is the joint posterior density of the model parameters. MCMC methods are often used to implement Bayesian inference by evaluating a Markov chain constructed such that its stationary distribution is the posterior density of interest. An MCMC algorithm is used to sample, at least approximately, from the posterior density by generating a Markov chain. Various algorithms exist to generate such chain.

1.4.1 Metropolis-Hastings sampling algorithm

The Metropolis algorithm was originally developed by Metropolis et al. (1953) and generalized by Hastings (1970), to form the Metropolis-Hastings (M-H) algorithm. This Monte Carlo algorithm provides a way to construct Markov chains, such that once the chain has reached stationarity, MCMC dependent samples from the density of interest are produced. Here, the target density is the posterior density $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. The M-H algorithm works as follows (Robert and Casella, 1999). At each MCMC iteration t, to move from the current state $\boldsymbol{\theta}^{(t)}$ to the next state $\boldsymbol{\theta}^{(t+1)}$, a proposed value, $\boldsymbol{\theta}^*$, is sampled from a proposal density, $q(\cdot|\boldsymbol{\theta}^{(t)})$. Then $\boldsymbol{\theta}^*$ is accepted and the next state becomes $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^*$

with following probability

$$\alpha(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^*) = \min\left(1, \frac{\pi(\boldsymbol{\theta}^*|\boldsymbol{y})\ q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}^{(t)}|\boldsymbol{y})\ q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t)})}\right), \tag{1.4.1}$$

otherwise $\boldsymbol{\theta}^*$ is rejected and $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)}$ (i.e. the chain does not move). The algorithm may be implemented as follows.

Algorithm 1 The Metropolis-Hastings algorithm

1. Initialise algorithm with $\boldsymbol{\theta}^{(0)}$ and set $t = 0$
2. Sample a new parameter value, $\boldsymbol{\theta}^*$ randomly from the proposal probability density $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t)})$
3. Sample U from the uniform distribution $U(0,1)$
4. If $U \leq \alpha(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^*)$, set $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^*$, otherwise set $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)}$
5. Set $t = t + 1$ and return to step 2

1.4.2 Gibbs sampler

Gibbs sampling approach (Geman and Geman, 1984) is an alternative MCMC method for generating a Markov chain. The Gibbs sampler can be seen as a special case of the M-H algorithm where the proposed value of $\boldsymbol{\theta}$ is always accepted (Gelfand and Smith, 1990). The Gibbs sampler works as follows. Let $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)$ be a parameter vector, and $\boldsymbol{\theta}_{-j} = (\theta_1, ..., \theta_{j-1}, \theta_{j+1}, ..., \theta_k)$ be $\boldsymbol{\theta}$ with θ_j removed, where $j = 1, ..., k$. Then, at each iteration t, each θ_j of $\boldsymbol{\theta}$ is generated from its full conditional density, $\pi_j(\theta_j|\boldsymbol{\theta}_{-j}, \boldsymbol{y})$. The algorithm may be implemented in practice as in Algorithm 2.

1.4.3 Burn-in and thinning

The burn-in period refers to discarding an initial portion of a Markov chain sample so that the effect of initial values on the posterior inference is minimised. The length of the burn-in usually depends on the initial values as well as how fast the MCMC chain converges to the desired distribution. In practice,

Algorithm 2 The Gibbs sampler

1. Initialise algorithm with $\boldsymbol{\theta}^{(0)}$ and set $t = 0$
2. Simulate each element of $\boldsymbol{\theta}$ as follows:

* Generate $\theta_1^{(t+1)}$ from $\pi_1(\theta_1|\theta_2^{(t)}, \theta_3^{(t)}, ..., \theta_k^{(t)}, \boldsymbol{y})$
* Generate $\theta_2^{(t+1)}$ from $\pi_2(\theta_2|\theta_1^{(t+1)}, \theta_3^{(t)}, ..., \theta_k^{(t)}, \boldsymbol{y})$
* Generate $\theta_3^{(t+1)}$ from $\pi_3(\theta_3|\theta_1^{(t+1)}, \theta_2^{(t+1)}, \theta_4^{(t)}, ..., \theta_k^{(t)}, \boldsymbol{y})$
* ...
* Generate $\theta_k^{(t+1)}$ from $\pi_k(\theta_k|\theta_1^{(t+1)}, \theta_2^{(t+1)}, ..., \theta_{k-1}^{(t+1)}, \boldsymbol{y})$

3. Set $t = t + 1$ and return to step 2

we can determine the burn-in samples from MCMC trace plots (Gilks et al. (1996) and Robert and Casella (1999)).

If there is significant autocorrelation between the MCMC samples, then a somewhat crude, but reasonably effective, method dealing with autocorrelation is to only keep every k^{th} draws from the posterior and discard the rest; this is known as thinning the Markov chain (Gilks et al. (1996) and Robert and Casella (1999)).

1.5 Review of Statistical analysis for stochastic epidemic models

Statistical analysis plays a crucial role in bridging the gap between the mathematical theory of epidemics and public health. In this section, we review the methods of statistical inference on model parameters and epidemiologically important parameters.

For homogeneous populations, Becker (1989) provides the largest amount of information about making inference for stochastic epidemic models. For the SIR models, it is straightforward to perform statistical inference based on complete data by differentiating the likelihood function to obtain maximum likelihood estimates (MLEs) for the parameters of interest. Rida (1991) derived

asymptotic normality and consistency for the estimators of the infection and removal rates for the Markovian SIR model as the initial number of susceptibles tends to infinity. For the same model but using a different approach, Becker (1989), using martingale methods, derived martingale-based estimators for both the infection and removal rates and also derived a formula to estimate R_0 where only final size data are available. Becker and Hasofer (1997) developed estimators for the infection rate and the mean infectious period based on the assumption of observing the whole removal process. However, maximum likelihood estimation approach is not usually feasible when only partial information of outbreak is available and inference for SIR models using martingale techniques is largely restricted to the Markovian models. As a result, MCMC methods in combination with a Bayesian approach have become an attractive methodology in this context not only because it relaxes the Markovian assumption but also its ability to handle missing data (O'Neill, 2002).

In practice, the infection times are hard to observe which in turn makes the likelihood of observing only the removal times given the model parameters intractable. One possible solution to make the likelihood tractable is to use the data augmentation technique (Dempster et al. (1977) and Gelfand and Smith (1990)) via MCMC methods and treat the missing data (the infection times) as extra (unknown) parameters to be estimated from the data (O'Neill and Roberts, 1999). By doing so we can sample from the conditional posterior distribution of the infection times in order to produce samples from the conditional posterior distributions of the model parameters.

The idea of applying MCMC methods to the Markovian SIR model was first given in Gibson and Renshaw (1998) and O'Neill and Roberts (1999). Then, it was extended to non-Markovian (gamma) infectious period model in O'Neill and Becker (2001) as well as in Streftaris and Gibson (2004b) where the infectious periods follow a Weibull distribution.

The assumption of homogeneous mixing populations is not realistic in many

applications. One possible way to model social heterogeneities in the community is via stochastic epidemic models that incorporate multiple-level of mixing. Ball et al. (1997) described a stochastic epidemic model with two levels of mixing, where the population is partitioned into households and infective individuals have contacts both within and between the households. The homogeneously mixing epidemic model can be seen as a special case of this model when all the households are of size one. Bayesian inference for this two levels of mixing model based on final size data was developed by Demiris and O'Neill (2005a). Britton et al. (2011) introduced a three levels of mixing model (a generalisation of the two levels of mixing model of Ball et al. (1997)). Their model assumes a three-level structure, with each susceptible individual belonging to a household, a group (e.g. school or workplace), and to the community as a whole. The authors established also Bayesian statistical parameter estimation procedures given both final size data and complete data on a single outbreak.

Another type of epidemic models that has been used to describe the structure of relationships within populations in the study of infectious disease dynamics is the random network model. The earliest paper to consider the problem of inferring the structure of a random graph by using epidemic data was Britton and O'Neill (2002). The authors described a methodology for performing Bayesian statistical inference for a network epidemic model incorporating a simple unobserved social structure mechanism, given both complete and partial temporal data.

While in many situations parametric assumptions are reasonable, we sometimes have no prior knowledge of the underlying distributions. In such situations, the use of parametric statistics might yield misleading results. Nonparametric estimation approach is statistical procedure that is based on the idea of using the observed data to infer unknown quantities while making as few assumptions as possible. For instance, in the epidemic context that means drawing inference without making certain parametric modelling assumptions

for the epidemic quantities that govern transmission of the disease. There has been relatively little work in the area of non-parametric inference for epidemic models (Xu, 2015). Becker and Yip (1989), using a martingale approach, provided a non-parametric inference method for the the person-to-person infection rate in an SEIR epidemic model by allowing the infection rate to depend on time, where it was assumed that both infection and removal times are known based on the assumption of known fixed length latent and infectious periods. More recently, Xu (2015) and Knock and Kypraios (2014) have developed new methodologies that enable non-parametric estimation of the transmission parameters within a Bayesian framework given only removal times. In particular, the authors have drawn inference for the overall force of infection without making specific assumptions on its parametric functional form.

An intractable likelihood is a common problem when performing inference for stochastic epidemic models. This problem arises when partial temporal epidemic data (i.e. removal times) are available. In this case the desired likelihood can be thought of as the integral, over all possible configurations of infection times, of the joint likelihood of infection and removal times. However, this integral is typically analytically and numerically intractable due to its high dimensionality and the non-trivial nature of the region of integration. Intractable computationally likelihoods can also arise as the population size gets larger. One natural solution of intractable likelihoods is to consider a simpler approximating model under which a likelihood can be computed. Cauchemez and Ferguson (2008) introduced a diffusion process that mimics the SIR process, for which exact solution is available in order to fit it to large-scale aggregated data on measles outbreaks.

Another method, which is appropriate for inference for complex stochastic models from which it is easy to simulate but impossible to calculate the likelihoods, is Approximate Bayesian Computation (ABC) method (Marjoram et al., 2003). This method works by replacing the calculation of the likelihood by a

step which involves simulating artificial data from the model using different parameter values drawn from the prior, and then comparing summary statistics of the simulated data to summary statistics of the observed data. If they are sufficiently close, the candidate prior parameter values are accepted as approximate samples from the posterior distribution. As stochastic epidemic models are relatively straightforward to simulate, McKinley et al. (2009) used this approach to make inference for epidemic models using pre-specified summary statistics and compared the ABC algorithms used with data-augmented MCMC algorithms based on both complete and partial temporal epidemic data. Neal (2012), based upon coupling simulations with different sets of parameters, introduced coupled-ABC algorithm which generates a set of values from the (approximate) posterior distribution of the model parameters during each simulation rather than a single value. The author used the coupled-ABC method to analyse final size data for epidemics amongst communities partitioned into households.

1.6 Review of techniques for Bayesian model adequacy and choice for epidemic models

This section briefly reviews some of existing techniques concerning model assessment for stochastic epidemic models. More details regarding model adequacy for epidemic models will be given in Chapter 2 and Chapter 3. Similarly, more details about model choice for epidemic models will be provided in Chapter 4.

Assessing the fit of a model is a crucial part of any statistical analysis. A statistician should check whether or not the fitted model explains adequately the important aspects of the given data before making any conclusion from the process of modelling. For simple models, such as generalized linear models, there are various methods used to determine if the proposed model provides an adequate fit; for instance, the coefficient of determination often denoted by

R^2 and the deviance (Faraway, 2005). However, when more complex models are being used such as stochastic epidemic models, fewer techniques to establish goodness-of-fit exist.

It is desirable to be able to statistically model disease outbreaks and use the information gained to try to help control the spread of disease. However, in order to do this, a sufficiently accurate model must be used. If the model provides an adequate fit, it is possible to interpret parameter estimates, compare disease epidemics and implement control procedures (Becker, 1989). The topic of model assessment for stochastic epidemic models is of great importance because any epidemiological quantity that we estimate might be of limited use unless the epidemic model is adequate in some statistical sense. Furthermore, epidemic models that are incapable of approximately reproducing the observed data should not be trusted to predict future outbreaks.

The area of parameter estimation for stochastic epidemic models has been a subject of considerable research (see the recent review by O'Neill (2010)). However, when it comes to model assessment, i.e. the question of whether or not a stochastic epidemic model fits a set of disease outbreak data, there has been relatively little work in terms of developing methods and procedures to evaluate the model adequacy and choice for stochastic epidemic models.

We shall now briefly review previous literature on the topic of model adequacy and choice for stochastic epidemic models but much more details will be given throughout the next three chapters. Broadly speaking, the literature on this topic when temporal outbreak data are available is somewhat modest. Both Bailey (1975) and Andersson and Britton (2000) are classical references containing a presentation and analysis of stochastic epidemic models, however they do not provide any obvious procedure to assess the quality of these models when fitting them to a set of temporal outbreak data. Another important reference is (Becker, 1989, chapter 6) in which chain binomial models were fitted to non-temporal outcome household data and the overall model fit was checked using chi-squared goodness-of-fit test. The author dealt with

the question of model fit for temporal outbreak data by fitting generalized linear models instead of stochastic epidemic models, where it was assumed that the number of infective and susceptible individuals were observed based on the assumption of fixed infectious period and then a comparison of the deviances associated with models was made. O'Neill et al. (2000) also employed chi-squared goodness-of-fit test to household final outcome data analysed using an independent-households model. However, none of the aforementioned methods can be used in our epidemic setting because we consider only partially observed temporal outbreak data.

1.6.1 Model adequacy

Traditionally, model adequacy relies on examining the discrepancy between the observed data and the corresponding quantities under the fitted model. However, due to the fact that disease outbreak data have high dependency and are partially observed (significant parts of the disease transmission process are unobserved), standard methods of goodness-of-fit cannot be applied directly (O'Neill, 2010).

Literature related to two major approaches that have been used to assess the model adequacy for stochastic epidemic models is now briefly reviewed. More details will be given in Chapter 2 and Chapter 3.

1.6.1.1 Residuals technique

For simple models such as a linear regression model, goodness-of-fit diagnostic tools include plotting observed data with expected, examining some measures of the goodness-of-fit such as the deviance or R^2, and plotting residuals against fitted values, to see whether or not the assumptions of the linear regression model are valid. However, it is not obvious how to do so when it comes to more complex stochastic models such as epidemic models considered in this thesis. Streftaris and Gibson (2004a) and Gibson et al. (2006)

developed a stochastic residuals approach that aims to detect discrepancies between a fitted model and data. The idea is based on defining a latent residual process, with known sampling distribution and then the model fit is examined via generating samples from the posterior distribution of this residual process, to which a classical test (such as Kolmogorov-Smirnov (K-S) test) is applied, enabling the posterior distribution of the p-value of the test used to be estimated as a representation of evidence against the model fit. Streftaris and Gibson (2012) employed the idea of Bayesian latent stochastic residuals to detect a possible lack of fit in a Susceptible-Exposed-Infective-Removed (SEIR) model given data on removals. Lau et al. (2014) extended the idea of Streftaris and Gibson (2012) to test and compare heterogeneously mixing spatio-temporal SEIR stochastic epidemic models. Jewell et al. (2009) made use of application of non-centred parametrisations (Papaspiliopoulos et al., 2003) to epidemic models to assess the appropriateness of their proposed epidemic model by examining the posterior distribution of the non-centred stochastic residuals defined by reparametrising the infectious period. However, it is evident that the choice of which residuals to employ can be somewhat arbitrary (O'Neill, 2010).

1.6.1.2 Posterior predictive model-checking technique

One intuitive, highly flexible, and very under-utilized (especially in the epidemic context) approach for validating a model is to perform predictive simulations. In a Bayesian context, it is called posterior predictive model-checking (Guttman (1967), Rubin (1981) and Gelman et al. (1996)). In this approach, a distribution for a test statistic (discrepancy) is found through simulation from the posterior distribution and then compared with the observed test statistic (discrepancy) for the data. Lekone and Finkenstädt (2006) chose the final size of the epidemic as a summary test statistic and used this predictive approach to check the model adequacy of a discrete time Susceptible-Exposed-Infective-Removed (SEIR) model with control intervention. Gardner et al. (2011) con-

sidered the length of the epidemic as a test statistic and then used the posterior predictive model checking approach to ascertain how well it can detect misspecification of discrete-time spatio-temporal individual-level model (Deardon et al., 2010). Using the same modelling framework, Zhang (2014) compared the posterior predictive distribution of the cumulative number of infection events under various fitted discrete-time spatio-temporal individual-level models to the observed cumulative number of cases curve. Boys and Giles (2007) assessed an SEIR model fit by using the posterior predictive distributions of selected removal times. However, posterior predictive assessment methods that use test discrepancy measures (i.e. an extension of test statistics to allow dependence on unknown parameters) to compare observed and replicated data is not explored in more detail for stochastic epidemic models (Forrester, 2006).

1.6.2 Model choice

While model adequacy criteria aim to pick a model that best fits data and gives trustworthy results, they take no account for the complexity of the model since the more complex a model becomes, the better it will fit to the data. However, this fit comes with a cost. More complex models require estimation of quite a few parameters. While models are often constructed with the aim of providing answers to specific research questions, there is often uncertainty about the most appropriate model out of a selection of competing models to best describe the observed data. It would be desirable if we could just rank the models by how well they fit the data.

In a Bayesian context, there are several approaches to address this issue. One popular approach is to calculate a comparison measure such as the Bayes factor (Kass and Raftery, 1995) or DIC (Spiegelhalter et al., 2002). The Bayes factor tells us how strongly the observed data support one model over another by providing evidence both for and against a model. It is an intuitive way of performing the Bayesian equivalence of classical statistical hypothesis testing.

However, in all but the simplest cases, Bayes factors must be evaluated numerically. As the Bayes factor can be viewed as the ratio of the posterior probabilities of two models, reversible jump Markov chain Monte Carlo (RJMCMC) methods (Green, 1995) may be used to estimate directly the posterior model probabilities.

An alternative way to calculate the Bayes factor is to estimate the marginal likelihood for each model separately, and then use this information to form Bayes factors. Estimation of the marginal likelihood is a non-trivial task for most statistical models and there has been considerable effort in the literature to find algorithms and methods for this purpose. Notable and popular approaches include Laplace's method (Tierney and Kadane, 1986), the harmonic mean estimator (Newton and Raftery, 1994), Chib's method (Chib, 1995), annealed importance sampling (Neal, 2001), and nested sampling (Skilling, 2006). For a recent review on various methods for estimating the marginal likelihood, see Friel and Wyse (2012). Also, the reader is referred to Kass and Raftery (1995) and Han and Carlin (2001) who presented comprehensive reviews of methods to compute Bayes factors.

Another sophisticated approach to estimate the marginal likelihood and hence the Bayes factor which requires very little tuning and is straightforward to implement was developed by Friel and Pettitt (2008) and refined recently by Friel et al. (2014). The method is based on the idea of thermodynamic integration (Gelman and Meng, 1998) and it computes the marginal likelihood based on samples from the so-called *power posterior*. The power posterior method is applicable when the likelihood of the observed data given the model parameters can be written in a closed form. However, in our epidemic settings, infection times are assumed to be unobserved, and therefore the likelihood of the observed data (removal times) given the parameters is intractable. To overcome this problem, we first develop an extension of the power posterior method to models incorporating missing data, and then we apply our extension to various epidemic settings.

Now, literature related to two major approaches that have been used to select between stochastic epidemic models is briefly reviewed. More details will be given in Chapter 4.

Within the setting of Bayesian inference, reversible jump MCMC (RJMCMC) methods (Green, 1995) can be used to assess the relative probability of competing models. Such methods have been applied to epidemic models in particular settings. In the case of temporal outbreak data, Neal and Roberts (2004) analysed a highly detailed historical measles dataset and used an RJMCMC method to select between epidemic models in order to identify which of spatial location, households and classrooms are important in the transmission of measles throughout the population. Also, Bayesian model choice, implemented with an RJMCMC method, was considered by O'Neill and Marks (2005) given temporal data from an outbreak of gastroenteritis in a school. Knock and O'Neill (2014) used an RJMCMC approach to estimate Bayes factors given data on final outcome of outbreak among a population of individuals which is partitioned into households. However, the computational burden associated with the RJMCMC method and defining how the Markov chain jumps between model parameter spaces are non-trivial challenges in implementing it in practice.

The deviance information criterion (DIC) method (Spiegelhalter et al., 2002) is a model comparison tool that is appropriate for complex Bayesian models, and since its development a number of variants have been introduced, including those for its application to missing data models (Celeux et al., 2006). One advantage of using this method is that it is often easily implemented for models fitted using Markov chain Monte Carlo methods. In the stochastic epidemic context, Worby (2013) investigated a version of DIC for missing data models (DIC_6 in Celeux et al. (2006)), and evaluated its performance in choosing between a specific set of nested transmission models using real and simulated hospital data. Another suitable form of DIC (DIC_4 in Celeux et al. (2006)) was implemented using final size epidemic data by Knock and O'Neill (2014)

to select between epidemic models. Further, Lau et al. (2014) compared two forms of DIC (DIC_4 and DIC_8 in Celeux et al. (2006)) with the latent residuals approach in order to test and compare spatio-temporal models by misspecifying the spatial transmission kernel in the models.

The broad objective of this thesis is to systematically address model assessment for stochastic epidemic models within a Bayesian framework. In the next section, we state three distinct topics regarding the development of methods to assess the goodness-of-fit of epidemic models to data, and methods for comparing different models. Also, the content of the remaining chapters of this thesis is summarised.

1.7 Aims and structure of the thesis

In this thesis, we move beyond inference for epidemic models to develop effective and easily used methods for model assessment for epidemic models within a Bayesian framework. In particular, we aim to do the following:

- Evaluate systematically the performances of different but natural model assessment techniques, applied to stochastic epidemic models given data on removals.
- Investigate and establish methods for assessing model adequacy for epidemic models based on partially observed temporal data.
- Consider the problem of choosing between competing epidemic models given data on removals by developing an extension of the existing power posterior methods to models incorporating missing data in order to enable more easy and less complicated estimation of the Bayes factor.

In Chapter 2, we systematically evaluate the performances of three different but natural methods for model adequacy for the Markovain SIR stochastic epidemic model. Given only data on removals, we propose to distinguish the

Markovain SIR model from the gamma and constant SIR models by examining the imputed infectious periods distribution obtained via MCMC methods. Then, based on complete epidemic data, we consider the time scaling of the inter-event times to detect the data-generating SIR model. Lastly, two purely temporal statistics are examined to assess the goodness-of-fit for the SIR models using the posterior predictive distribution approach. Throughout, our methods are illustrated with simulated datasets.

In Chapter 3, we develop a novel two-stage method to assess the goodness-of-fit for the SIR stochastic epidemic models. The method works by examining two predictive distributions, namely the predictive distribution of the final size of the epidemic and the predictive distribution of the removal times. Three different epidemic settings are investigated, namely the SIR model with different infection mechanisms, the SIR model with different infectious periods and the SIR model with different population structures. The method explores successfully how SIR models with these different settings can be assessed and distinguished given only a set of removal times via the posterior predictive model checking approach. Our method is illustrated using several simulated datasets as well as real outbreak data.

In Chapter 4, Bayesian model choice for epidemic models is explored, with the aim of determining plausible underlying epidemic models. We extend the existing power posterior method (Friel and Pettitt (2008) and Friel et al. (2014)) for estimating the Bayes factor to models incorporating missing data such as epidemic models. We derive explicit forms for Bayes factors in two different epidemic settings in case of complete epidemic data, namely the SIR model with different different infectious periods and the SIR model with different infection mechanisms. Then, given only a set of removal times, we successfully apply our extension of power posterior method to the two aforementioned epidemic settings in order for the Bayes factor to be estimated more easily. Also, we study the ability of DIC method to select between epidemic models based on removal data. The methods are applied to compare stochastic

transmission models, using simulated datasets.

Chapter 5 concludes the thesis with a summary of our findings, gives some general conclusions and indicates any limitations of our methods.

All results and graphics in this thesis were obtained using the R statistical software (R Core Team, 2013).

CHAPTER 2

Model Adequacy for the Markovian SIR Epidemic Model Based on Removal Data

2.1 Introduction

This chapter is concerned with Bayesian model assessment for the Markovian Susceptible-Infective-Removed (SIR) model. When modelling infectious diseases, one aspect of particular importance is the infectious period, that is the period of time during which an infected individual can transmit the disease to susceptible individuals. The purpose of this chapter is to examine the consequences of misspecifying the length of infectious period in model assessment for the SIR models. As well as the Markovian stochastic SIR model (the SIR model with exponentially distributed infectious periods), we consider two stochastic SIR models that are widely used in the literature, namely the SIR model with gamma distributed infectious periods and the SIR model in which the infectious periods are (non-random) constant.

The research question may be expressed as follows. Given a set of partially observed temporal outbreak data that consists of removal times, can one distinguish the Markovian SIR model from other SIR models in order to assess

the goodness-of-fit? We approach this problem by investigating various methods to explore and discriminate the Markovian stochastic SIR model from the gamma and constant SIR models.

The remainder of this chapter is organised as follows. In Section 2.2, we start by discussing some previous work on model assessment for stochastic epidemic models and indicate its limitations. Then, epidemic likelihood construction is described for the SIR models based on two different approaches. Section 2.3 is about making Bayesian inference and using MCMC methods for the SIR models in which complete and incomplete data are treated. In Section 2.4, we propose a method to determine the infectious period distribution for SIR models based only on a set of removal data using the labelled representation of the likelihood approach. We test the performance of our proposed method using simulated removal data. Another proposed method to assess the goodness-of-fit of the Markovian SIR model is introduced in Section 2.5 in which we assume that the entire epidemic is observed through time. We examine the posterior distribution of inter-event times, under fitting the Markovian SIR model, given various removal datasets simulated under different generating mechanisms. In Section 2.6, we investigate the use of the posterior predictive model checking method to assess the goodness-of-fit for the SIR models via two different purely temporal statistics that are based only on observing the removal process of an epidemic. Simulated outbreak data are used to illustrate our proposed method. Finally, we conclude the chapter with a summary of our findings as well as potential future extensions to this work.

2.2 Previous studies and relevant background

2.2.1 Previous work on model adequacy for epidemic models

In this section, we shall review a handful of relevant papers that introduce some methods of exploring model adequacy for stochastic epidemic models. Other epidemic model assessment papers that are not closely connected to the content of this chapter will be discussed in the next two chapters.

Streftaris and Gibson (2012) employed the idea of Bayesian latent residuals to detect a possible lack of fit in a Susceptible-Exposed-Infective-Removed (SEIR) model which can be obtained by adding an extra state, the exposed period, to the SIR model in which individuals in this state are infected but not yet infectious. The authors considered a construction due to Sellke (1983), which is an equivalent way of formulating standard stochastic epidemic processes, where susceptible individuals are associated with independent tolerance levels to infection, given by $Q_j \sim Exp(1)$, for individual j, $j = 1, 2, ..., n$, where n represents the initial number of susceptibles. This individual j becomes infected at time $t_j = \inf\{t \geq 0 : Q_j = \int_0^t \beta Y(u)du\}$, where $Y(u)$ denotes the number of infected individuals at time u, and if $t_j = \infty$ then individual j escapes infection. In the authors' set-up, it was assumed that $Q_j \sim$ Weibull(a, b), then a Bayesian latent residual was defined as $\tau_j = F_{Q_j}(q_j) = 1 - \exp(-bq_j^a) \sim U(0, 1)$, where $F_Q(.)$ is the cumulative distribution function of the threshold variable Q with observed value denoted by q. The authors examined the posterior distribution of the unobserved latent variables obtained using MCMC and then claimed that the sampling distribution of $\boldsymbol{\tau}^i = (\tau_1^i, ..., \tau_n^i)$ at each MCMC iteration i should be $U(0, 1)$ under the hypothesis that the model fit is adequate. As a final step, the Kolmogorov-Smirnov test was used to obtain a p-value at each MCMC iteration and the resultant distribution of p-values was employed for checking the model assumptions, where the smaller the p-value, the poorer the fit. However, the effectiveness of this method may be limited since it relies heavily on information imputed

in the process (unobserved exposure times and unobserved tolerances), rather than making a direct comparison with the observed data. This approach has its roots in the Bayesian framework proposed in Gibson et al. (2006) for percolation stochastic models of disease spread in plant populations.

Neal and Roberts (2005) and Kypraios (2007) were the first to apply non-centred parametrisations (Papaspiliopoulos et al., 2003) to epidemic models. Jewell et al. (2009) made use of this idea to assess the appropriateness of the epidemic model proposed to fit the 2001 UK Foot and Mouth epidemic data. The authors assumed a gamma infectious period (i.e. $T_I^j = N_j - I_j \sim \text{Gamma}(\alpha, \delta)$) for the time from infection (I_j) to notification (N_j) of farm j, and then reparameterised the infectious periods as $U_j = \delta(N_j - I_j)$, where a priori $U \sim \text{Gamma}(\alpha, 1)$ for fixed known $\alpha = 4$ and $\delta > 0$, and $j = 1, ..., n_I$, where n_I denotes the total number of premises who got infected. The authors adopted a partially non-centred approach, focused on examining the distribution of the non-centred residuals U_j and claimed that if the model fits the data, then the posterior distribution of U_j over MCMC iterations should be approximately $\text{Gamma}(\alpha, 1)$.

Recently, Lau et al. (2014), using non-centred parametrisations, extended the idea of Bayesian latent residuals introduced in Streftaris and Gibson (2004a) and Streftaris and Gibson (2012) to test and compare heterogeneously mixing spatio-temporal SEIR stochastic epidemic models. The key issue was to identify the spatial transmission kernel, that characterises the spatial nature of the dispersal of the pathogen from infective to susceptible individuals. The authors defined a latent residual, namely the infection-link residual which corresponds to the time of infectious contact of an exposure event. The sampling posterior distribution of this latent residual over MCMC iterations should be indistinguishable from the uniform distribution $U(0, 1)$ in order for the selected SEIR model to be adequate. However, our epidemic settings are based on models for populations in which individuals are homogeneous in the way they mix with each other.

Residuals based on latent quantities appear appealing, particularly because they require standard MCMC output from a Bayesian analysis. However, as with any model adequacy method that does not focus on the observed data, the choice of what residual to look at is rather arbitrary.

The posterior predictive model checking method (Guttman (1967), Rubin (1981) and Rubin (1984)) is a popular Bayesian model checking tool because of its simplicity and intuitive appeal. Lekone and Finkenstädt (2006) used this approach to check the model adequacy of a stochastic discrete-time SEIR model fitted to daily incidence and mortality time series for an outbreak of Ebola in the Democratic Republic of Congo in 1995. The authors chose the final size as a summary statistic. For each posterior sample of the model parameters, the model was simulated and the final size of the epidemic was recorded. This produced a final size posterior predictive distribution. If the model fits the data then the observed final size should not be in the tails of this predictive distribution. Using the final size as a summary statistic, no evidence of lack of fit was found.

Gardner et al. (2011) also used the posterior predictive model checking approach to examine different temporal and spatio-temporal test statistics and to ascertain how well they can detect misspecification of discrete-time spatio-temporal individual-level model (Deardon et al., 2010). The time at which the last infectious individual is removed from the population (the length of the epidemic) was found to be the preferred measure to assess goodness-of-fit for spatio-temporal individual-level models.

Based on the above literature, we can observe the following. Although the residual analysis is an effective tool for checking the model fit, the choice of which residual to be considered in the epidemic setting is somewhat arbitrary. Further, using the posterior predictive model checking approach in epidemic settings is mostly based on a test statistic that depends on data only but not on both data and model parameters. Discrepancy measures of epidemic goodness-of-fit, that are functions of both data and model paramet-

ers (Gelman et al., 2013), are not explored in more detail (Forrester, 2006). Therefore, the research on model assessment for stochastic epidemic models when temporal outbreak data are available is limited to some extent and well-established statistical techniques do not currently exist. In consequence, there exists a significant scope of further research in this area which motivates us to seek and investigate new methods for assessing goodness-of-fit for stochastic epidemic models.

2.2.2 Epidemic likelihood function

There are three ways proposed in the literature to construct a likelihood function based on temporal data for the Markovian SIR model, namely Bailey and Thomas' representation due to Bailey and Thomas (1971), an alternative representation due to Britton and O'Neill (2002), (see also Neal and Roberts (2005) and Kypraios (2007)) and a martingale representation due to Becker (1989), (see also Andersson and Britton (2000)). The basic idea in the martingale approach is to derive estimating equations for the parameters of interest based on constructing martingales related to counting processes that are contained within the epidemic model (e.g. counting the number of infections by a given time). However, we shall consider and use the first two approaches.

2.2.2.1 Bailey and Thomas' representation: unlabelled individuals

Consider a fully observed epidemic outbreak that results in a total of n_R removals, $1 \leq n_R \leq \boldsymbol{N}$, in which all infected individuals are removed by its end in a certain observation period $(I_1, R_{n_R}]$, where $I_1 < 0$ and R_{n_R} denote the first infection time and the last removal time respectively. Further, let $0 = R_1, R_2, ..., R_{n_R}$ be the ordered removal times and write $\boldsymbol{R} = (R_1, ..., R_{n_R})$. Also, indicate by $\boldsymbol{I} = (I_2, ..., I_{n_R})$ to the rest of ordered infection times, where $I_k < I_{k+1} < R_k$ for $k = 1, 2, ..., n_R - 1$. Then, the likelihood of $\boldsymbol{I}$ and $\boldsymbol{R}$ given β, γ and I_1 can be derived using survival analysis methodology (see

e.g. O'Neill and Roberts (1999); Höhle and Jørgensen (2002)). That is, roughly speaking

$$\begin{aligned} L(\boldsymbol{I},\boldsymbol{R}|\beta,\gamma,I_1) &= \text{likelihood of infections} \times \text{likelihood of removals} \\ &\times \text{likelihood of events not occurring,} \end{aligned}$$

$$\begin{aligned} L(\boldsymbol{I},\boldsymbol{R}|\beta,\gamma,I_1) &= \prod_{j=2}^{n_R} \beta n^{-1} X(I_j-) Y(I_j-) \times \prod_{j=1}^{n_R} \gamma Y(R_j-) \\ &\times \exp\left(- \int_{I_1}^{R_{n_R}} \left(\beta n^{-1} X(t) Y(t) + \gamma Y(t) \right) dt \right), \end{aligned} \tag{2.2.1}$$

where I_j- denotes the time just prior to I_j and R_j- is defined similarly. It should be pointed out here that each pair (I_k, R_k) refers to the k_{th} infection and removal time in sequential order and there is no need to know which individual has been infected or removed in the population because of the homogeneous mixing assumption.

2.2.2.2 The alternative representation: labelled individuals

In this set-up, individuals in the population are labelled and associated with them their corresponding infection and removal times. The likelihood of the data given the model parameters can be written as the product of contributions for each individual, that is (see e.g. Britton and O'Neill (2002), Neal and Roberts (2005) and Kypraios (2007) for details)

$$\begin{aligned} L(\boldsymbol{I},\boldsymbol{R}|\beta,\gamma,I_z,z) &= \left(\prod_{j=1, j\neq z}^{n_R} \beta n^{-1} Y(I_j-) \right) \times \exp\left(-\beta n^{-1} A \right) \\ &\times \prod_{j=1}^{n_R} \gamma \exp\left(-\gamma (R_j - I_j) \right), \end{aligned} \tag{2.2.2}$$

where

$$A = \int_{I_z}^{R_{n_R}} X(t) Y(t) dt = \sum_{j=1}^{n_R} \sum_{k=1}^{\boldsymbol{N}} (R_j \wedge I_k - I_j \wedge I_k) \tag{2.2.3}$$

with $I_k = R_k = \infty$ for $k = n_R + 1, ..., \boldsymbol{N}$. Here, $\boldsymbol{R} = (R_1, ..., R_{n_R})$ represents the set of removal times, where $R_1 \leq R_2 \leq ... \leq R_{n_R}$ and $\boldsymbol{I} = \{I_j : j \neq z\}$ denotes the set of infection times such that for $1 \leq j \leq n_R$, I_j corresponds to the

infection time of the individual removed at time R_j with the initial infective labelled z such that $I_z < I_j$ for all $j \neq z$.

This alternative construction of the likelihood is convenient as it enables us to relax the Markovian assumption applied to obtain (2.2.1). To put it another way, knowledge of the infection and removal times only with the Markovian assumption allow us to derive the likelihood (2.2.1) without relating the infection and removal times to certain individuals. However, linking the infection and removal times to individuals makes it possible for any arbitrary but specific distribution for T_I (not necessarily exponential) to be used. To see that, let $f_{T_I}(\cdot)$ denote the probability density (or mass, if T_I is discrete) function of T_I (the length of infectious period) and let $\boldsymbol{\theta}$ indicate the parameter governing T_I, where $\boldsymbol{\theta}$ may be a vector. Then, the likelihood of the data given the model parameters is

$$\begin{aligned} L(\boldsymbol{I}, \boldsymbol{R}|\beta, \boldsymbol{\theta}, I_z, z) &= \left(\prod_{j=1, j\neq z}^{n_R} \beta n^{-1} Y(I_j-) \right) \times \exp\left(-\beta n^{-1} A\right) \\ &\times \prod_{j=1}^{n_R} f_{T_I}\left(R_j - I_j|\boldsymbol{\theta}\right), \end{aligned} \tag{2.2.4}$$

In the case of $T_I \equiv c$, where $c > 0$ is a constant and $T_I \sim \text{Gamma}(\alpha, \delta)$, where α and δ denote the shape and scale parameters of gamma distribution and a random variable $\text{Gamma}(\lambda, \nu)$ has a density function given by

$$f(x) = \frac{x^{\lambda-1} e^{-\nu x}}{\nu^{-\lambda}\Gamma(\lambda)}, \quad x \geq 0\,; \lambda, \nu > 0, \tag{2.2.5}$$

the infection parts of the likelihoods are identical to the one in (2.2.4). However, the removal parts in the likelihood (2.2.4) become

$$\prod_{j=1}^{n_R} f_{T_I}\left(R_j - I_j\right) = \mathbb{1}_{\{c=R_j-I_j,\, j=1,\ldots,n_R\}},$$

where

$$\mathbb{1}_{\{c=R_j-I_j,\, j=1,\ldots,n_R\}} = \begin{cases} 1 \text{ if } c = R_j - I_j,\ j = 1, \ldots, n_R \\ 0 \text{ otherwise,} \end{cases}$$

and

$$\prod_{j=1}^{n_R} f_{T_I}\left(R_j - I_j\right) = \Gamma^{-n_R}(\alpha) \times \prod_{j=1}^{n_R} (R_j - I_j)^{\alpha-1} \times \delta^{\alpha n_R}\, e^{-\delta \sum_{j=1}^{n_R} (R_j - I_j)}$$

for $T_I \equiv c$ and $T_I \sim \text{Gamma}(\alpha, \delta)$ respectively.

Throughout this chapter, we shall focus attention on the case where the shape parameter of gamma distribution α is assumed to be known although the considered methods can be extended to the case where α is unknown and is therefore a parameter within the model to be estimated from data. As we are concerned with comparing the SIR model with different infectious periods, we shall set $\alpha = 10$ in order to more clearly distinguish the gamma SIR model from the exponential SIR model as setting $\alpha = 1$ in the former yields the latter.

2.3 Inference and MCMC method for the SIR models

In this section, we shall illustrate how Bayesian inference via MCMC methods can be performed for the Markovian SIR model (Bayesian inference for the other two SIR models may be implemented in a similar way with minor modifications that will be explained as we go along). As far as the Markovian SIR model is concerned, Becker (1989), derived martingale-based estimates for both β and γ using martingale methods based on a complete data assumption, namely

$$\widehat{\beta}_{MLE} = \frac{n(n_R - 1)}{\int_{I_1}^{R_{n_R}} X(t)Y(t)dt}, \qquad \widehat{\gamma}_{MLE} = \frac{n_R}{\int_{I_1}^{R_{n_R}} Y(t)dt}. \tag{2.3.1}$$

However, making inference for SIR models using martingale techniques is largely restricted to Markovian models (Becker, 1989). As a result, MCMC methods in combination with a Bayesian approach have become an attractive methodology in this context not only because it relaxes the Markovian assumption but also its ability to handle missing data (O'Neill, 2002).

2.3.1 Bayesian inference and MCMC for the Markovian SIR model

The idea of applying MCMC methods to the Markovian SIR model was first given in Gibson and Renshaw (1998) and in O'Neill and Roberts (1999). Then, this was extended to non-Markov (gamma) infectious period model in O'Neill and Becker (2001) as well as in Streftaris and Gibson (2004b) in which the infectious period follows a Weibull distribution.

In the following, the alternative labelled representation introduced in Section 2.2.2.2 shall be adopted for making inference for the Markovian SIR model. Bayesian inference for this model via MCMC method given data on removals can be performed as described in O'Neill and Roberts (1999) when Bailey and Thomas' representation is used.

2.3.1.1 Complete data case

Given complete outbreak data, the Bayesian framework can be adopted first by assigning independent conjugate gamma prior distributions for the model parameters (O'Neill and Roberts, 1999), namely $\beta \sim \text{Gamma}(\lambda_\beta, \nu_\beta)$, and $\gamma \sim \text{Gamma}(\lambda_\gamma, \nu_\gamma)$. Then, by multiplying the prior distributions and the likelihood (Bayes'theorem), the full posterior density for β and γ is given by

$$\begin{aligned}\pi(\beta,\gamma|\boldsymbol{I},z,I_z,\boldsymbol{R}) &\propto L(\boldsymbol{I},z,I_z,\boldsymbol{R}|\beta,\gamma)\ \pi(\beta)\ \pi(\gamma)\\ &= \left(\prod_{j=1,j\neq z}^{n_R}\beta n^{-1}\Upsilon(I_j-)\right)\times\exp\left(-\beta n^{-1}A\right)\\ &\times \prod_{j=1}^{n_R}\gamma\ \exp\left(-\gamma(R_j-I_j)\right)\times\beta^{\lambda_\beta-1}e^{-\nu_\beta\beta}\times\gamma^{\lambda_\gamma-1}e^{-\nu_\gamma\gamma},\end{aligned} \tag{2.3.2}$$

where

$$A=\sum_{j=1}^{n_R}\sum_{k=1}^{N}(R_j\wedge I_k - I_j\wedge I_k).$$

As the two model parameters β and γ are a posteriori conditionally independent, we get

$$\beta|\boldsymbol{I}, z, I_z, \boldsymbol{R} \sim \text{Gamma}\left(\lambda_\beta + n_R - 1, \nu_\beta + n^{-1}A\right), \tag{2.3.3}$$

$$\gamma|\boldsymbol{I}, z, I_z, \boldsymbol{R} \sim \text{Gamma}\left(\lambda_\gamma + n_R, \nu_\gamma + \sum_{j=1}^{n_R}(R_j - I_j)\right). \tag{2.3.4}$$

Therefore, the posterior distributions of the two parameters β and γ are readily available and any function of these parameters, such as the basic reproduction number $R_0 = \beta/\gamma$, can be obtained easily by sampling from each of the posteriors (2.3.3) and (2.3.4).

2.3.1.2 Incomplete data case

In practice, the infection times are hard to observe which in turn makes the likelihood of observing only the removal times $\boldsymbol{R}$ given the model parameters intractable as the integral, over all possible configurations of infection times, of the joint density of $\boldsymbol{I}$ and $\boldsymbol{R}$ is analytically and numerically intractable due to its high dimensionality. One possible solution to make the likelihood tractable is to use data augmentation technique (Dempster et al. (1977); Gelfand and Smith (1990)) by treating the missing data (the infection times) as extra (unknown) parameters to be estimated from the data (O'Neill and Roberts, 1999). By doing so we can sample from the conditional posterior distribution of the infection times in order to produce samples from the conditional posterior distributions of the model parameters.

Again, by assigning independent conjugate gamma prior distributions for the model parameters β and γ and further assuming a priori that z is equally likely to be any of the n_R infected individuals and I_z has an improper uniform prior density on $(-\infty, R_1)$ and by using Bayes'theorem, we obtain the following full conditional posterior distributions

$$\beta|\gamma, z, I_z, \boldsymbol{I}, \boldsymbol{R} \sim \text{Gamma}\left(\lambda_\beta + n_R - 1, \nu_\beta + n^{-1}A\right), \tag{2.3.5}$$

$$\gamma|\beta, z, I_z, \boldsymbol{I}, \boldsymbol{R} \sim \text{Gamma}\left(\lambda_\gamma + n_R, \nu_\gamma + \sum_{j=1}^{n_R}(R_j - I_j)\right) \tag{2.3.6}$$

and

$$\pi(\boldsymbol{I}, z, I_z|\beta, \gamma, \boldsymbol{R}) \propto \left(\prod_{j=1, j\neq z}^{n_R} Y(I_j-)\right) \times \exp\left(-\beta n^{-1} A\right) \times \prod_{j=1}^{n_R} \exp\left(-\gamma(R_j - I_j)\right). \tag{2.3.7}$$

The model parameters β and γ can be updated using Gibbs sampling steps as they have closed forms of the posterior distribution. However, the infection times need to be updated using Metropolis-Hastings step as their posterior distribution is not available explicitly. In general, the MCMC algorithm 3 may be applied to update both the model parameters and the missing infection times in the case where the epidemic is known to have ceased by the observed time R_{n_R}. In another case where the epidemic is still in progress by time R_{n_R}, the MCMC algorithm needs extra steps (see for example O'Neill and Roberts (1999)) but for simplicity we shall consider the first case. Note that z and I_z are possibly updated while updating $\boldsymbol{I}$, so there is no need to update them separately.

When the length of the infectious periods is assumed to be constant, we have two model parameters to be updated, namely the length of the infectious period c, and the infection rate parameter β. In this case, the infection rate parameter β can be updated using a Gibbs step and c may be updated using a random walk Metropolis with a Gaussian proposal density centred on the current value, with a fixed standard deviation. It can be seen that $\boldsymbol{I}$, I_z and z are all updated automatically when c is updated as $I_j = R_j - c,\ j = 1, ..., n_R$.

When the infectious periods follow a gamma distribution Gamma(α, δ) with known α, in addition to estimate the infection rate parameter β, we shall estimate the scale parameter δ. Here, both β and δ can be updated using Gibbs sampling steps, whereas the infection times need to be updated using a Metropolis-Hastings step, that is similar to one employed in Algorithm 3.

It is worth mentioning that improving the mixing and the speed of convergence of the Markov chain, reducing the computational time as well as the choice of the infection times proposal distribution are all crucial issues when implementing MCMC methods on partially observed outbreak data. Excel-

Algorithm 3 MCMC algorithm for updating $\beta, \gamma, \boldsymbol{I}, I_z$ and z

1. Initialise algorithm with β^0, γ^0 and $\boldsymbol{I}^0$.

2. Repeat the following steps a large number of times, say M.

* Update β by using Gibbs Sampler.

* Update γ by using Gibbs Sampler.

* Update $\boldsymbol{I}$, I_z and z by using Metropolis-Hastings Sampler as follows:

a. Choose one individual in the population at random, say j.

b. Draw its proposed infection time I_j^* such that $R_j - I_j^* \sim \text{Exp}(\gamma^{mle})$, where

$$\gamma^{mle} = \frac{n_R}{\sum_{j=1}^{n_R}(R_j - I_j)}$$

evaluated at each MCMC iteration.

c. Draw a uniform random variable $U \in (0, 1)$.

d. Accept the proposed infection time if

$$\begin{aligned} U &\leq \min\left\{ \frac{\pi(\boldsymbol{I}^*|\beta,\gamma,z^*,I_{z^*}^*,\boldsymbol{R}) \times \gamma \exp\left(-\gamma(R_j - I_j)\right)}{\pi(\boldsymbol{I}|\beta,\gamma,z,I_z,\boldsymbol{R}) \times \gamma \exp\left(-\gamma(R_j - I_j^*)\right)}, 1 \right\} \\ &= \min\left\{ \frac{\pi(\boldsymbol{I}^*|\beta,\gamma,z^*,I_{z^*}^*,\boldsymbol{R})}{\pi(\boldsymbol{I}|\beta,\gamma,z,I_z,\boldsymbol{R})} \times \exp\left(\gamma(I_j - I_j^*)\right), 1 \right\}. \end{aligned}$$

3. Return $\{\beta^0, \beta^1, ..., \beta^M\}$ and $\{\gamma^0, \gamma^1, ..., \gamma^M\}$.

lent references for some algorithms that deal with these issues are Kypraios (2007) and Neal and Roberts (2005), where centred and non-centred parametrisations (Papaspiliopoulos et al., 2003) are proposed.

Returning to the Markovian SIR model, implementing Algorithm 3 enables us to obtain straightforwardly approximate samples from $\pi(\beta|\boldsymbol{R})$ and $\pi(\gamma|\boldsymbol{R})$, the marginal posterior distributions of the model parameters. As a result, not only posterior summaries such as the posterior mean and the posterior variance, can be reported, but also these posterior samples could be used to produce predicted realisations from the model.

In what follows, we investigate three different methods to assess the model fit for the Markovian, gamma and constant SIR epidemic models.

Throughout the analysis, uninformative gamma prior distributions with parameters $\lambda_\beta = \lambda_\gamma = \lambda_\delta = 1$ and $\nu_\beta = \nu_\gamma = \nu_\delta = 0.001$ were set to the parameters of the SIR models and it was assumed that the gamma shape parameter, when fitting an SIR model with gamma infectious period $T_I \sim \text{Gamma}(\alpha, \delta)$ is known $(\alpha = 10)$. We assumed a priori that z (the label of the initial infective) is equally likely to be any of the n_R infected individuals. Also, I_z was assigned an improper uniform prior density on $(-\infty, R_1)$ and a priori c has an improper uniform prior density on $(0, \infty)$.

2.4 Identifying the infectious period distribution using labelled individuals representation

The most common available data that can be reported from the spread of infectious diseases are the times when infected individuals are removed (e.g. by recovery, death or isolation). In this section, based on this type of partially observed temporal outbreak data, we propose a method to explore the Markovian SIR model denoted by m_1. Apart from the Markovian stochastic epidemic model (the SIR epidemic model in which the infectious period is exponentially distributed, $T_I \sim \text{Exp}(\gamma)$), there are two extensions of it that are more realistic as well as widely used when modelling the spread of infectious diseases. These models, denoted by m_2 and m_3 respectively, may be obtained by either allowing the infectious periods to be distributed according to a gamma distribution or by assuming a fixed infectious periods for all infected individuals. By adopting the labelled individuals representation introduced in Section 2.2.2.2, the three aforementioned epidemic models will be assessed and compared in our analysis.

It follows that, if $\boldsymbol{I}, z, I_z$ and $\boldsymbol{R}$ are all observed, then identifying the infectious

period distribution using the labelled individuals representation is trivial. Specifically, the type of infectious period is determined by looking at the distribution of the quantity $R_j - I_j,\ j = 1, ..., n_R$.

Now, suppose that temporal removal outbreak data (removal times) of size n_R were observed, say $\boldsymbol{R} = (R_1, ..., R_{n_R})$. When fitting the three aforementioned SIR models to these removal data via the MCMC labelled individuals approach, a natural technique to discriminate between them is to examine the imputed infectious period distributions over MCMC iterations. To put it simply, we can look at the inferred posterior distribution of sum of the random variables $R_j - I_j,\ j = 1, ..., n_R$ calculated at each MCMC iteration to learn about the type of infectious period distribution for each SIR model. For instance, in the Markovian SIR model (m_1) in which $T_I = R_j - I_j \sim \text{Exp}(\gamma)$, for a given value of γ, the quantity $\gamma \sum_{j=1}^{n_R}(R_j - I_j)$ is gamma distributed, that is

$$\gamma \sum_{j=1}^{n_R}(R_j - I_j) \sim \text{Gamma}(n_R, 1) \tag{2.4.1}$$

Also, in the SIR model (m_2) in which $R_j - I_j \sim \text{Gamma}(\alpha, \delta)$, for a given value of δ, the quantity $\delta \sum_{j=1}^{n_R}(R_j - I_j)$ is gamma distributed, that is

$$\delta \sum_{j=1}^{n_R}(R_j - I_j) \sim \text{Gamma}(\alpha n_R, 1) \tag{2.4.2}$$

provided that $\alpha \in \mathbb{Z}^+$.

For a given value of c in the SIR model (m_3) in which $R_j - I_j = c$, the quantity $\sum_{j=1}^{n_R}(R_j - I_j)/c$ is always equal to the number of removals n_R, that is

$$\frac{\sum_{j=1}^{n_R}(R_j - I_j)}{c} = n_R. \tag{2.4.3}$$

Our proposed method can be outlined as follows. Given some removal times, we fit the three SIR models, namely the exponential SIR model (m_1), the gamma SIR model (m_2) and the constant SIR model (m_3) to the observed removal data using MCMC methods. At each MCMC iteration in the data-augmented MCMC algorithm used to fit each model, we calculate the corresponding quantity ($\gamma \sum_{j=1}^{n_R}(R_j - I_j)$, $\delta \sum_{j=1}^{n_R}(R_j - I_j)$ or $\sum_{j=1}^{n_R}(R_j - I_j)/c$) using the observed removal times $\boldsymbol{R}$, the current values of the imputed infection times $\boldsymbol{I}$, I_z and the

current value of the model parameter (γ, δ or c). Then, in order to discriminate between the models, we look at the resultant posterior samples of each quantity over MCMC iterations, hoping that these posterior samples follow either a Gamma$(n_R, 1)$ if the observed data are generated under a Markovian SIR model, a Gamma$(\alpha n_R, 1)$ if the observed data are generated under a gamma SIR model or these posterior samples are reduced to a single point mass, that is n_R if the observed data are generated under a constant SIR model. However, if all relations (2.4.1), (2.4.2) and (2.4.3) are held regardless of the removal data generating mechanism, this means that the proposed method is insufficient to distinguish between the models.

As an aside, one could examine the empirical distribution of the $R_j - I_j$ ($j = 1, 2, ..., n_R$) at each MCMC iteration, rather than considering only the statistic $\sum_{j=1}^{n_R}(R_j - I_j)$. Using an appropriate goodness-of-fit diagnostic method (such as the Kolmogorov-Smirnov test), a p-value can thus be obtained at each MCMC iteration. This will result in a posterior distribution of relevant tail probability values that can be used for the assessment of the structural assumptions of the model.

Now, we shall test the proposed method using simulated outbreak data to see its capability to distinguish between the three SIR models.

2.4.1 Analysis of simulated data

In order to validate our proposed method, we simulated an epidemic under a Markovian SIR model. The outbreak was in a closed population consisting of $n = 49$ initially susceptible individuals and a single initially infectious case. The model parameters for the simulation were $\beta = 2.5$ for the infection rate and $\gamma = 1$ for the removal rate. The resulting data consisted of 44 removal times, so $n_R = 44$. We fitted the three SIR models, m_1 (the true model), m_2 and m_3 to the simulated data using MCMC methods. Our aim was to examine the posterior distribution of $\gamma \sum_{j=1}^{n_R}(R_j - I_j)$ under fitting m_1 to see whether or not

it looks like a Gamma$(n_R, 1)$. At the same time, the posterior distributions of $\delta \sum_{j=1}^{n_R}(R_j - I_j)$ and $\sum_{j=1}^{n_R}(R_j - I_j)/c$ were also inferred when fitting m_2 and m_3 to these removal data in order to distinguish between these models.

Figure 2.1 shows the inferred posterior distributions of $\gamma \sum_{j=1}^{n_R}(R_j - I_j)$ (left) and $\delta \sum_{j=1}^{n_R}(R_j - I_j)$ (right) resulting from fitting the exponential and gamma SIR models respectively to the simulated removal data. The posterior distribution of the quantity $\sum_{j=1}^{n_R}(R_j - I_j)/c$ obtained from fitting the constant SIR model to the observed removal data was a point mass, i.e. the observed total number of removals $n_R = 44$. It is evident from Figure 2.1 that, the relations

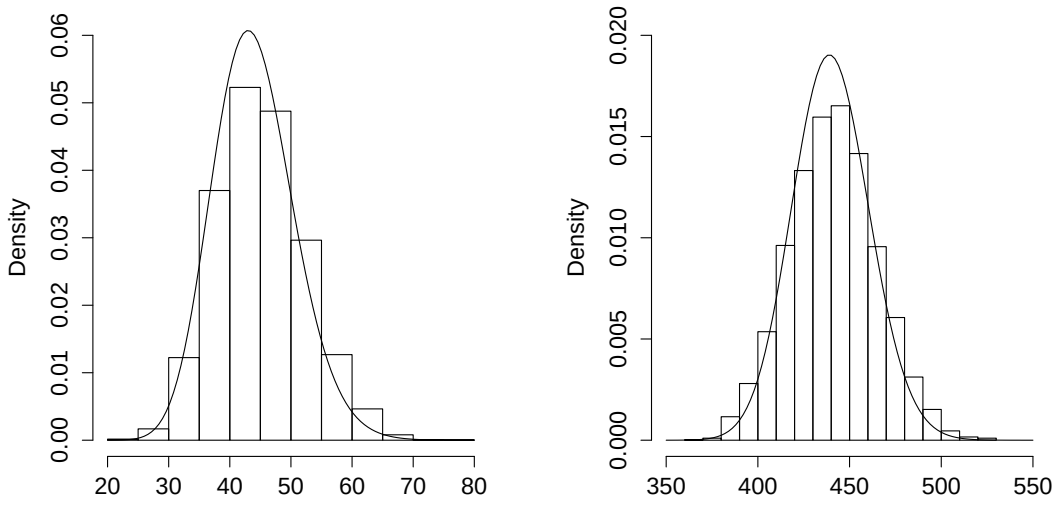


Figure 2.1: The posterior distribution of $\gamma \sum_{j=1}^{n_R}(R_j - I_j)$ overlaid by a Gamma$(44, 1)$ (left), and the posterior distribution of $\delta \sum_{j=1}^{n_R}(R_j - I_j)$ overlaid by a Gamma$(440, 1)$ (right) obtained from fitting m_1 and m_2 to the observed exponential SIR removal data.

(2.4.1) and (2.4.2) are satisfied and so according to the method we used, all of the three proposed models appear to be good candidates for the model from which the data actually were generated.

Constructing a quantile-quantile (Q-Q) plot might be another approach to measuring agreement between the empirical and gamma distribution functions.

To explore this further, another simulation study was carried out. We simulated this time some removal times from a gamma SIR model where $n = 49, a = 1, \beta = 2.5, \alpha = 10$ and $\delta = 10$. A total of 45 removal times were observed ($n_R = 45$). Then, we fitted the three SIR models and followed the same procedure as in the first simulation study. The resultant posterior distributions of $\gamma \sum_{j=1}^{n_R}(R_j - I_j)$ and $\delta \sum_{j=1}^{n_R}(R_j - I_j)$ are shown in Figure 2.2. Also, the posterior distribution of $\sum_{j=1}^{n_R}(R_j - I_j)/c$ obtained from fitting the constant SIR model to the observed removal data was a point mass around the observed total number of removals $n_R = 45$.

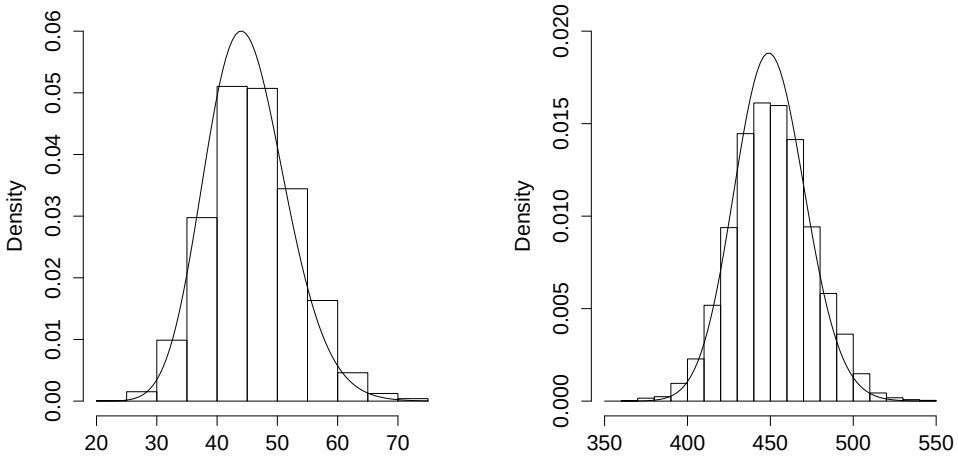


Figure 2.2: The posterior distribution of $\gamma \sum_{j=1}^{n_R}(R_j - I_j)$ overlaid by a Gamma$(45, 1)$ (left), and the posterior distribution of $\delta \sum_{j=1}^{n_R}(R_j - I_j)$ overlaid by a Gamma$(450, 1)$ (right) obtained from fitting m_1 and m_2 to the observed gamma SIR removal data.

Although the imputation process of the missing infection times might encourage the test statistics to have the appropriate distribution for the model fitted, the dependence of the removal times makes it not that obvious.

Based on these results and other similar results produced by considering many other datasets (plots not shown here), the proposed method is clearly insufficient to distinguish between the three SIR models. A possible explanation for this is that the uncertainty around the imputed infection times is large enough

to prevent reliable estimation of the infectious period distribution.

In the following section, we consider an alternative approach in which we shall restrict our attention to identify the Markovian SIR model.

2.5 Assessing the Markovian SIR model goodness-of-fit via time-rescaling

Suppose that the entire epidemic is observed through time. Although this assumption is not very realistic, it represents the best scenario to illustrate our proposed method in this section. Also, assume one introductory case whose infection takes place at time $t = 0$ (i.e. this fixes the time origin) and denote the ordered event times $I_1, ..., I_{n_R}$ and $R_1, ..., R_{n_R}$ jointly as $0 = I_1 = t_1 < t_2 < ... < t_{2n_R} = R_{n_R}$. In the Markovian SIR model, the infection process is a non-homogeneous Poisson process with random rate $\lambda_1(t) = \beta n^{-1} X(t) Y(t)$ and the removal process is also a non-homogeneous Poisson process with random rate $\lambda_2(t) = \gamma Y(t)$. Therefore, the sum of these two non-homogeneous Poisson processes is also a non-homogeneous Poisson process (Cinlar, 1975) with random rate $\lambda(t) = \lambda_1(t) + \lambda_2(t) = \beta n^{-1} X(t) Y(t) + \gamma Y(t)$.

Papangelou (1972) established a general time-rescaling theorem, which states that any point process with an integrable rate function may be rescaled into a Poisson process with a unit rate. Brown et al. (2002) presented a proof of the time-rescaling theorem and showed how it may be used to assess the goodness-of-fit for point process neural spike train models. Now we shall state the time-rescaling theorem (the proof is detailed in Brown et al. (2002)).

Time-rescaling theorem:

Suppose that $0 = u_0 < u_1 < u_2 < ... < u_n \leq T$ is a realisation from a point process with conditional intensity function $\lambda(t|\mathcal{H}_t)$ satisfying $\lambda(t|\mathcal{H}_t) > 0$ for all $t \in [0, T]$, where $\mathcal{H}_t = \sigma\{W(s) : 0 \leq s < t\}$ is the sigma-algebra generated by the history of the process $\{W(s) : s \geq 0\}$ up to time t. If we define the

following transformation

$$\tau_j = \int_{u_{j-1}}^{u_j} \lambda(t|\mathcal{H}_t)\, dt \tag{2.5.1}$$

for $j = 1, 2, ..., n$, then τ_j are i.i.d Exp(1) random variables.

Now, we shall use the time-rescaling theorem to construct a goodness-of-fit test for the Markovian SIR model and attempt to discriminate between this Markovian model and two other non-Markovian SIR models, namely the gamma and constant SIR models. Note that all these three SIR models are assumed to have the same infection process but different removal processes. Also, the non-homogeneous Poisson process is quite simply the point process where the conditional intensity function is independent of the past, i.e. the conditional intensity function is equal to the intensity function of the non-homogeneous Poisson process.

Once the Markovian SIR model has been fitted to a complete outbreak dataset using the Bayesian framework, we can compute from its posterior estimated parameter distributions (2.3.3) and (2.3.4) the following rescaled times

$$\tau_j = \left(t_j - t_{j-1}\right) \times \left(\beta n^{-1} X(t_{j-1}) Y(t_{j-1}) + \gamma Y(t_{j-1})\right),\ j = 2, 3, ..., 2n_R. \tag{2.5.2}$$

If the data-generating model is the Markovian SIR model, the model checking amounts to checking whether or not the rescaled times τ_j are independent exponential random variables with mean 1. Equivalently, we can sum τ_j up over the number of inter-event times to have the following relation :

$$\sum_{j=2}^{2n_R} \tau_j \sim \text{Gamma}(2n_R - 1, 1). \tag{2.5.3}$$

For other non-Markovian data-generating models like the gamma and constant SIR models, this should not necessarily be the case.

2.5.1 Analysis of completely observed epidemic data

For the first application of the time-rescaling theorem to a goodness-of-fit analysis for the Markovian SIR model, we analysed some complete outbreak data

simulated from a Markovian SIR model with infection rate $\beta = 2.5$ and mean infectious period $E(T_I) = 1$ in a closed population of size $n = 99$ initial susceptibles and one initial infective $a = 1$. The resulting data consisted of $2n_R = 184$ events. We then obtained approximate posterior samples from the posterior distributions of β and $E(T_I) = 1/\gamma$ by fitting the Markovian SIR model using MCMC methods with fixing the infection and removal times to the true observed values (see Figure 2.3 top row). Figure 2.4 (a) shows the rescaled times posterior distribution produced using (2.5.3) which is approximately Gamma$(183, 1)$.

Assessment of convergence of Markov chains can be preformed either visually or quantitatively (Gelman et al., 2013, chapter 11). The most straightforward approach for assessing convergence is based on simply plotting and inspecting traces of the observed MCMC sample to see how well it is mixing, or moving around the parameter space. After a suitable burn-in period, if the trace of values exhibits asymptotic behaviour, this may be satisfactory evidence for convergence. With the epidemic models and outbreak data considered in this thesis, convergence issues are of no concern as we can see from the sample traces plots in Figure 2.3.

Returning to our proposed method, so far it works well and the target distribution of $\sum_{j=2}^{2n_R} \tau_j$ was recovered correctly. However, one way to test the capability of the proposed method in discriminating between the Markovian SIR model and other non-Markovian models is to generate epidemic data using a non-Markovian data-generating model, then fit the Markovian SIR model and see whether or not (2.5.3) holds. Therefore, we simulated two other datasets from gamma and constant SIR models with the same population size, β, $E(T_I)$ and number of events $(2n_R)$ as the ones for the Markovian SIR model dataset. Then, the Markovian model was fitted to each dataset (see Figure 2.3 middle and lower panels). Figure 2.4 (b) and (c) show the rescaled times posterior distributions produced from fitting the Markovian SIR model to the two datasets. As can be seen, all outcome samples are approximately Gamma$(183, 1)$.

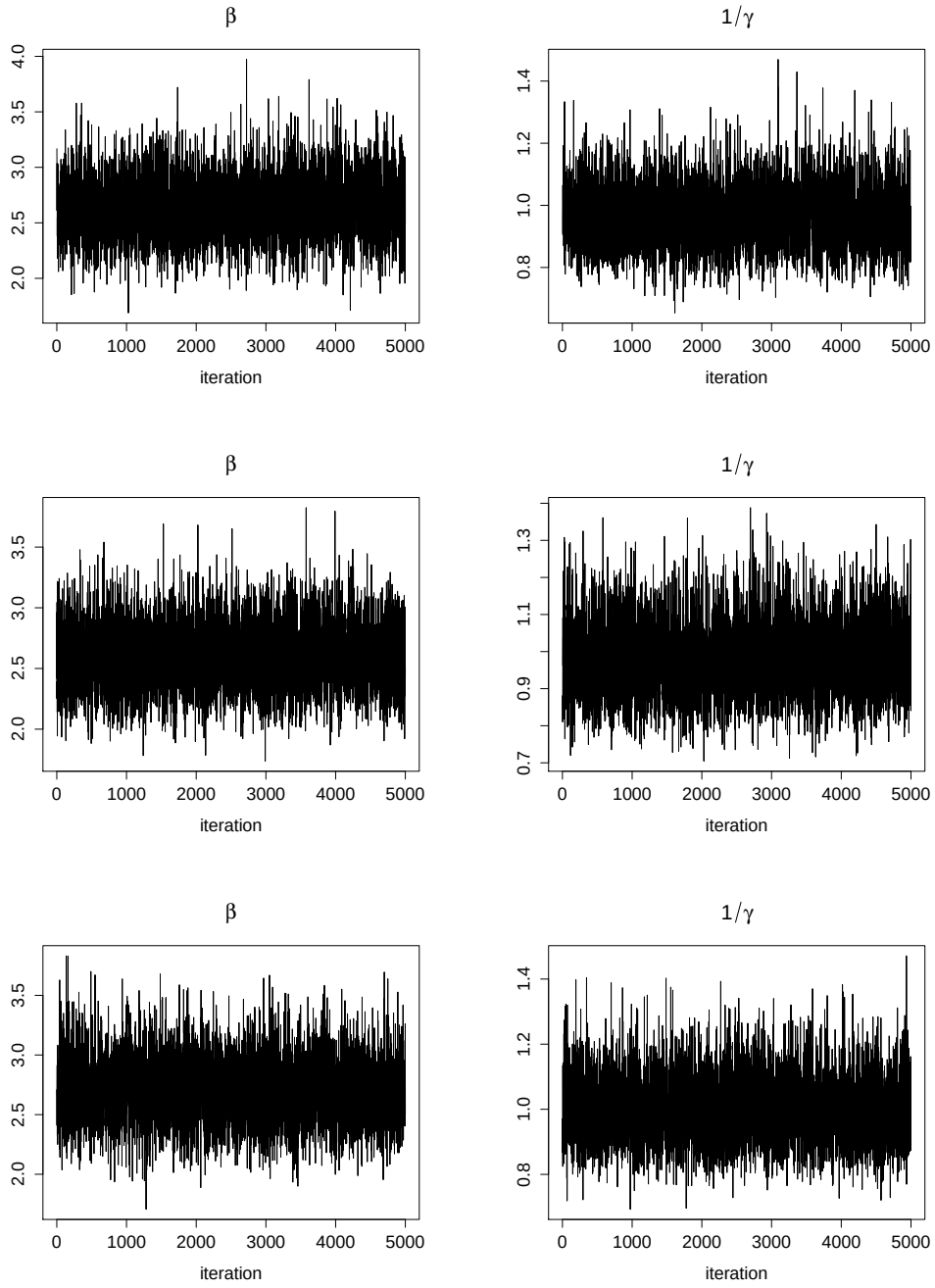


Figure 2.3: Sample traces for 5000 posterior samples of β and $1/\gamma$ obtained from fitting the Markovain SIR model to the three datasets, after a burn-in period of 2000 iterations and thinning by taking every $5th$ sample. Top, middle and lower panels correspond to datasets generated from the Markovian, gamma and constant SIR models respectively.

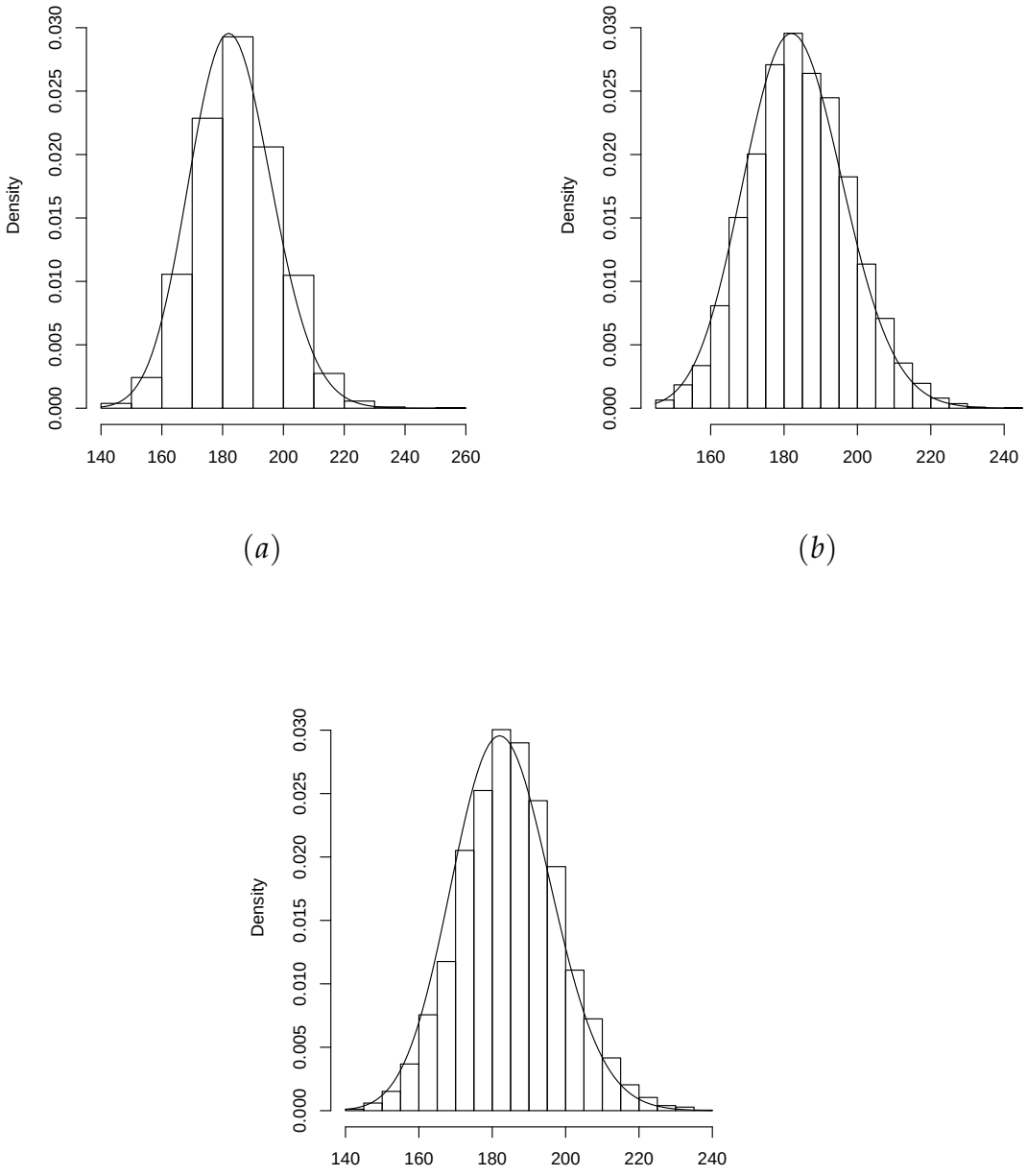
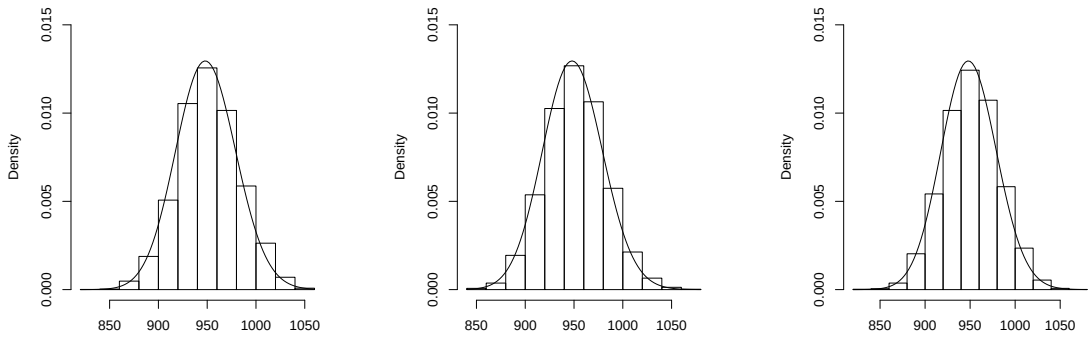


Figure 2.4: (a), (b) and (c) correspond respectively to the posterior distributions of $\sum_{j=2}^{2n_R} \tau_j$ obtained from fitting the Markovian SIR model to the epidemic datasets generated from the Markovian, gamma and constant SIR models. The curved line indicates a Gamma$(183, 1)$ density function.

To see whether or not our proposed method can discriminate the Markovian model as the size of the population increases, we carried out another simulation study on a larger population of size $N = 500$. We used $\beta = 3$ and $E(T_I) = 1$ to simulate three complete epidemic datasets of size 475 removals each from the Markovian, gamma and constant SIR models. Then, we fitted the Markovian SIR model to each dataset and calculated $\sum_{j=2}^{2n_R} \tau_j$ under each scenario. The rescaled times posterior distributions produced from fitting the Markovian SIR model to the three datasets are shown in Figure 2.5 which are approximately Gamma$(949, 1)$ for all three datasets.



(a) (b) (c)

Figure 2.5: (a), (b) and (c) correspond respectively to the posterior distributions of $\sum_{j=2}^{2n_R} \tau_j$ obtained from fitting the Markovian SIR model to the epidemic datasets generated from the Markovian, gamma and constant SIR models. The curved line indicates a Gamma$(949, 1)$ density function.

These results and many other similar results obtained based on different datasets (not presented here) indicate that the Markovian SIR model is always a good candidate to fit the gamma and constant SIR model datasets. Therefore, inferring the inter-event times distribution cannot clearly distinguish the Markovian SIR model from the gamma and constant SIR models. The fact that all these three SIR models are assumed to have the same Markovian infection process might be a reason for this outcome. Also, as can be seen from Figure

2.3, regardless of the data-generating model, the infection rate parameter β and the mean of infectious periods $E(T_I)$ are well estimated under fitting the Markovian SIR model which in turn reduces the effect of parameter estimation in the proposed method.

As our proposed method is incapable of discriminating between the three SIR models under the assumption of a completely observed epidemic, it will not work in the situation in which the observed data consist of removal times only.

In the following section, we consider the posterior predictive model checking approach in which we shall examine two different temporal statistics that based on observing the removal process in order to see whether or not they can assist in distinguishing the Markovian SIR model from other SIR models.

2.6 Martingale goodness-of-fit statistics for the SIR epidemic models

Within a Bayesian framework, a major tool for assessing goodness-of-fit is the posterior predictive distribution (Guttman, 1967). That is, a distribution for a test statistic (a scalar summary of data) is found through simulation from the posterior distribution and then compared with the observed test statistic for the data. However, an appropriate statistic that contains all information on an epidemic removal process can be difficult to identify (Lau et al., 2014).

In this section, we consider two different purely temporal statistics which based on the observed removal process, namely the martingale estimator of the infection rate denoted by T_β and the martingale estimator of the mean infectious period denoted by T_γ, and investigate how well they can detect any poor fit via a simulation study.

Looking at these particular test statistics is motivated by the fact that our observed data consist of removal times only and these statistics are specifically derived for such type of data (Becker and Hasofer, 1997). As these two scale

test statistics are based on the parameter estimates of the Markovian SIR epidemic, we attempt to discriminate between the Markovian SIR model (denoted by m_1) and two other non-Markovian SIR models, namely the gamma (denoted by m_2) and constant (denoted by m_3) SIR models.

2.6.1 Goodness-of-fit statistics

Given data $\boldsymbol{y}^{obs} = (y_1^{obs}, ..., y_n^{obs})$, we fit a model $\pi(\boldsymbol{y}^{obs}|\boldsymbol{\theta})$, which can be any arbitrary statistical model, and then evaluate the posterior density for a set of parameters $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta}|\boldsymbol{y}^{obs})$. Similarly, we can define the posterior predictive density of a new data observation $\boldsymbol{y}^{rep} = (y_1^{rep}, ..., y_n^{rep})$ (i.e. a replication of $\boldsymbol{y}^{obs}$) as

$$\pi(\boldsymbol{y}^{rep}|\boldsymbol{y}^{obs}) = \int_{\boldsymbol{\theta}} \pi(\boldsymbol{y}^{rep}|\boldsymbol{\theta}, \boldsymbol{y}^{obs})\pi(\boldsymbol{\theta}|\boldsymbol{y}^{obs})d\boldsymbol{\theta} = \int_{\boldsymbol{\theta}} \pi(\boldsymbol{y}^{rep}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{y}^{obs})d\boldsymbol{\theta}. \tag{2.6.1}$$

The posterior predictive distribution can be used as a measure of goodness-of-fit in a Bayesian setting. The main idea behind posterior predictive checking is that if the model fits the data, then replicated data generated under the model should look similar to the observed data (Gelman et al., 1996). To put it another way, the observed data, represented by a summary statistic, should seem reasonable under the posterior predictive distribution of that statistic. Overall goodness-of-fit measure can be summarized using a tail area probability that is called the posterior predictive p-value (Meng, 1994) described below.

2.6.1.1 The posterior predictive $p-$value

Once an appropriate test statistic $T(\cdot)$ has been chosen, and M posterior samples from the posterior distribution of $\boldsymbol{\theta}$ have been obtained, M replicated values (vectors) of $\boldsymbol{y}^{obs}$, $\boldsymbol{y}^{rep(i)}$, say, can be generated from the model using each posterior draw $\theta^{(i)}$, $i = 1, ..., M$. Then, $T(\boldsymbol{y}^{rep(i)})$ for each replicated data $\boldsymbol{y}^{rep(i)}$ can be computed. To this end, the observed test statistic, $T(\boldsymbol{y}^{obs})$, may be compared graphically to the distribution of $T(\boldsymbol{y}^{rep(i)})$ by locating $T(\boldsymbol{y}^{obs})$ in the

frequency distribution of $T(\boldsymbol{y}^{rep(i)})$. If $T(\boldsymbol{y}^{obs})$ falls in the tails of the predictive distribution, it can be concluded that the model under consideration does not accurately fit the data.

A more formal method to check the model fit is to calculate a tail area probability (the posterior predictive $p-$value) defined as:

$$\begin{aligned} ppp\text{-value} &= P\left(T(\boldsymbol{y}^{rep}) \geq T(\boldsymbol{y}^{obs})\middle|\boldsymbol{y}^{obs}\right) \\ &\approx \frac{1}{M}\sum_{i=1}^{M} \mathbb{1}_{\{T(\boldsymbol{y}^{rep(i)})\geq T(\boldsymbol{y}^{obs})\}}, \end{aligned} \tag{2.6.2}$$

where M denotes the number of posterior samples used to produce the posterior predictive distribution. A model is regarded as being inadequate if the posterior predictive $p-$value is close to 0 or 1 with optimal value of 0.5 (Gelman et al., 2013).

The algorithm for conducting a posterior predictive check using a test statistic $T(\cdot)$ may be implemented as in Algorithm 4.

Algorithm 4 Algorithm for implementing a posterior predictive check

1. Choose an appropriate test statistic $T(\cdot)$ that has the power to diagnose violations of the tested assumption of the model.
2. Compute $T(\boldsymbol{y}^{obs})$ for the observed data $\boldsymbol{y}^{obs}$.
3. For each posterior draw $\theta^{(i)}$, $i = 1, ..., M$, simulate a value (vector) of $\boldsymbol{y}^{rep(i)}$ from the model.
4. Compute $T(\boldsymbol{y}^{rep(i)})$ for each replicated data $\boldsymbol{y}^{rep(i)}$.
5. Count the proportion of times that $T(\boldsymbol{y}^{rep(i)}) \geq T(\boldsymbol{y}^{obs})$. This is an estimate of the posterior predictive $p-$value.

2.6.2 Martingale methods for the Markovian SIR model

Let $X(t), Y(t)$ and $Z(t)$ represent respectively the number of susceptible, infected and removed individuals at time $t \in [0, R_{n_R}]$ in a closed population of size $\boldsymbol{N}$ (n initial susceptibles and one initial infective). Also, let $N(t)$ count

the number of infections that occur by time t. Under the assumptions of the standard Markovian SIR epidemic model, there are two zero-mean martingale processes (Becker, 1989) defined by

$$M_1(R_{n_R}) = N(R_{n_R}) - \int_0^{R_{n_R}} \beta\, n^{-1} X(t) Y(t) dt, \tag{2.6.3}$$

$$M_2(R_{n_R}) = Z(R_{n_R}) - \int_0^{R_{n_R}} \gamma Y(t) dt. \tag{2.6.4}$$

Using martingale theory we can obtain the well-known estimator $\widehat{R_0}$ for the basic reproduction number, often denoted by R_0, that is

$$\widehat{R_0} = \frac{n}{Z(R_{n_R})} \sum_{j=1}^{Z(R_{n_R})-1} \frac{1}{n+1-j}, \tag{2.6.5}$$

provided at least one individual remains susceptible at time R_{n_R}. Note that, this estimator depends only on the initial number of susceptible individuals n and the final number of removals $Z(R_{n_R})$.

2.6.2.1 Martingale estimation under incomplete epidemic data

A plausible data set that might be observed during the course of an epidemic is that consisting of the initial values $(X(0) = n, Y(0) = 1, Z(0) = 0)$ and continuous observation of the removal process, $\boldsymbol{R} = (R_1, ..., R_{n_R})$ such that $R_1 < R_2 < ... < R_{n_R}$. Besides the martingales considered above, Becker and Hasofer (1997) used the method of moments and the martingale process $\{X(t)(1 + R_0/n)^{Z(R_{n_R})-Z(t)} : t \geq 0\}$ to derive the following explicit estimators for β and γ,

$$\widehat{\beta} = \frac{Z(R_{n_R})\widehat{R_0}}{\boldsymbol{N}R_{n_R} - \int_0^{R_{n_R}} Z(t)dt - X(R_{n_R}) \int_0^{R_{n_R}} (1 + \widehat{R_0}/n)^{Z(R_{n_R})-Z(t)} dt}, \tag{2.6.6}$$

$$\widehat{\gamma} = \frac{Z(R_{n_R})}{\boldsymbol{N}R_{n_R} - \int_0^{R_{n_R}} Z(t)dt - X(R_{n_R}) \int_0^{R_{n_R}} (1 + \widehat{R_0}/n)^{Z(R_{n_R})-Z(t)} dt}, \tag{2.6.7}$$

where $\widehat{R_0}$ is defined as in (2.6.5).

2.6.2.2 Epidemic test statistics

The choice of an epidemic test statistic is very important and the fundamental goal for any epidemic model assessment task would be to find a single omnibus goodness-of-fit test statistic that could provide an overall indication of model fit for any feature of model inadequacy. Gardner et al. (2011) examined different temporal and spatio-temporal test statistics and ascertained how well they can detect discrete-time spatio-temporal individual-level model (Deardon et al., 2010) misspecification. The time at which the last infectious individual is removed from the population (the length of the epidemic, R_{n_R} in our notation) was found to be the preferred measure to assess goodness-of-fit for spatio-temporal individual-level models. However, it might be thought desirable to have a test statistic which contains all available temporal removal information. Assuming that the whole removal process is observed, we consider here two epidemic temporal statistics, namely

$$T_\beta = \frac{Z(R_{n_R})\widehat{R_0}}{\boldsymbol{N}R_{n_R} - \int_0^{R_{n_R}} Z(t)dt - X(R_{n_R}) \int_0^{R_{n_R}} (1 + \widehat{R_0}/n)^{Z(R_{n_R})-Z(t)} dt}, \quad (2.6.8)$$

and

$$T_\gamma = \frac{\boldsymbol{N}R_{n_R} - \int_0^{R_{n_R}} Z(t)dt - X(R_{n_R}) \int_0^{R_{n_R}} (1 + \widehat{R_0}/n)^{Z(R_{n_R})-Z(t)} dt}{Z(R_{n_R})}. \quad (2.6.9)$$

The motivation for looking at these particular temporal statistics, in the first instance, is that they might be expected to be sensitive to an underestimation, or overestimation of β and $E(T_I)$ when misspecifying the infectious periods distribution. When the true model is fitted to the data, the posterior predictive p-value should preferably near to the optimal value 0.5, representing a good fit. In the other hand, when misspecified models are fitted to the data, posterior predictive p-values close to 0 or 1 would be desirable, representing a poor fit. For example, if our observed data $\boldsymbol{R}^{obs}$ violate the Markovian assumption, i.e. the data-generating model is a non-Markovian model such as the gamma or constant SIR models, then in order to rely upon these statistics our observed test statistic $T(\boldsymbol{R}^{obs})$ should be seen as an extreme value in the

histogram of $T(\boldsymbol{R}^{rep})$ values under fitting the Markovian SIR model and this in turn yields an extreme posterior predictive p-value.

2.6.3 Simulated outbreak data example

To systematically test our proposed methodology, we apply it to analyse several simulated outbreak data. Here, our results are based on two typical removal datasets. The first dataset (D_1) was simulated under the Markovian SIR model (m_1) in which $T_I \sim \text{Exp}(\gamma)$. However, the second dataset (D_2) was generated under the constant ($T_I \equiv c$) SIR model. Both removal datasets were simulated in a closed population of size $\boldsymbol{N} = 50$ ($n = 49$ and $a = 1$). The simulation parameters were chosen to be $\beta = 2$ and $E(T_I) = 1/\gamma = c = 1$ and the final number of removals was $n_R = 44$ in each dataset.

Our proposed method was applied to both datasets. To make our comparisons useful, we constrain on major epidemics, so replicated epidemic realisations that resulted in early extinction were eliminated and $T(\boldsymbol{R}^{rep})$ values used were based on major epidemics only.

In the first case, all three models m_1, m_2 and m_3 were fitted to D_1 removal dataset via Markov chain Monte Carlo (MCMC) method as explained earlier. After a burn-in of 2000 samples and thinning the chain by taking each $5th$ value, 5000 samples were obtained from the joint posterior distribution of (β, γ), (β, δ) and (β, c) for m_1, m_2 and m_3 respectively. Then, 5000 replicated removal datasets were generated from each model using their corresponding posterior samples. We wish to eliminate realisations that resulted in early extinction, thereby presenting results based on major epidemics only. Here, replicated realisations that have less than 20 infected individuals were excluded. We chose 20 for a cut off because with such parameter values $\beta = 2$ and $\gamma = 1$ (i.e $R_0 = 2$), the major outbreak part of the final size distribution is approximately assembled after 20.

For each simulated major epidemic, the two test statistics, $T_\beta(\boldsymbol{R}^{rep(i)})$ and

$T_\gamma(\boldsymbol{R}^{rep(i)})$, were calculated. This is repeated for $i = 1, ..., L$, where L is the number of major replicated epidemics, to produce estimates of the posterior predictive distributions of $T_\beta(\boldsymbol{R}^{rep})$ and $T_\gamma(\boldsymbol{R}^{rep})$ under fitting each model. $T_\beta(\boldsymbol{R}^{obs})$ and $T_\gamma(\boldsymbol{R}^{obs})$ were calculated for the observed data, $\boldsymbol{R}^{obs}$, and then the posterior predictive $p-$values, calculated using (2.6.2), were used to determine how well each model fits the data. To take account of the variation in the replicated simulations, the procedure of generating replicated epidemic data and finding the posterior predictive $p-$value was repeated 100 times.

The aim of this simulation study was to explore how good the two temporal test statistics in measuring the goodness-of-fit for the three SIR models. In order to do that, we used the posterior predictive p-value as a Bayesian tool for assessing the model fit. Figure 2.6 (top) represents boxplots of the posterior predictive $p-$value distributions produced using T_β for all three models and also boxplots of the posterior predictive $p-$value distributions produced using T_γ for all three models (bottom). By noting the range of the posterior predictive $p-$value in both figures, we have the following remarks. When the true model (m_1) was fitted to the observed removal data, both T_β and T_γ produced relatively adequate posterior predictive $p-$values that are close to the optimal value 0.5. However, under misspecified models (m_2 and m_3), both T_β and T_γ resulted in satisfactory (in the predictive model adequacy sense) posterior predictive $p-$values and none of them performed well in terms of detecting model misspecification.

In the second case, we considered D_2 dataset and calculated $T_\beta(\boldsymbol{R}^{obs})$ and $T_\gamma(\boldsymbol{R}^{obs})$ first and then all three models m_1, m_2 and m_3 were fitted to this removal dataset via MCMC methods. Following the same procedure used for the first dataset D_1, we obtained 100 posterior predictive $p-$values via T_β for all three models which are presented using boxplots as in Figure 2.7 (left). Also, using the test statistic T_γ, 100 posterior predictive $p-$values were produced for each SIR model as shown in Figure2.7 (right).

As it can be seen from the range of the posterior predictive $p-$value in Figure

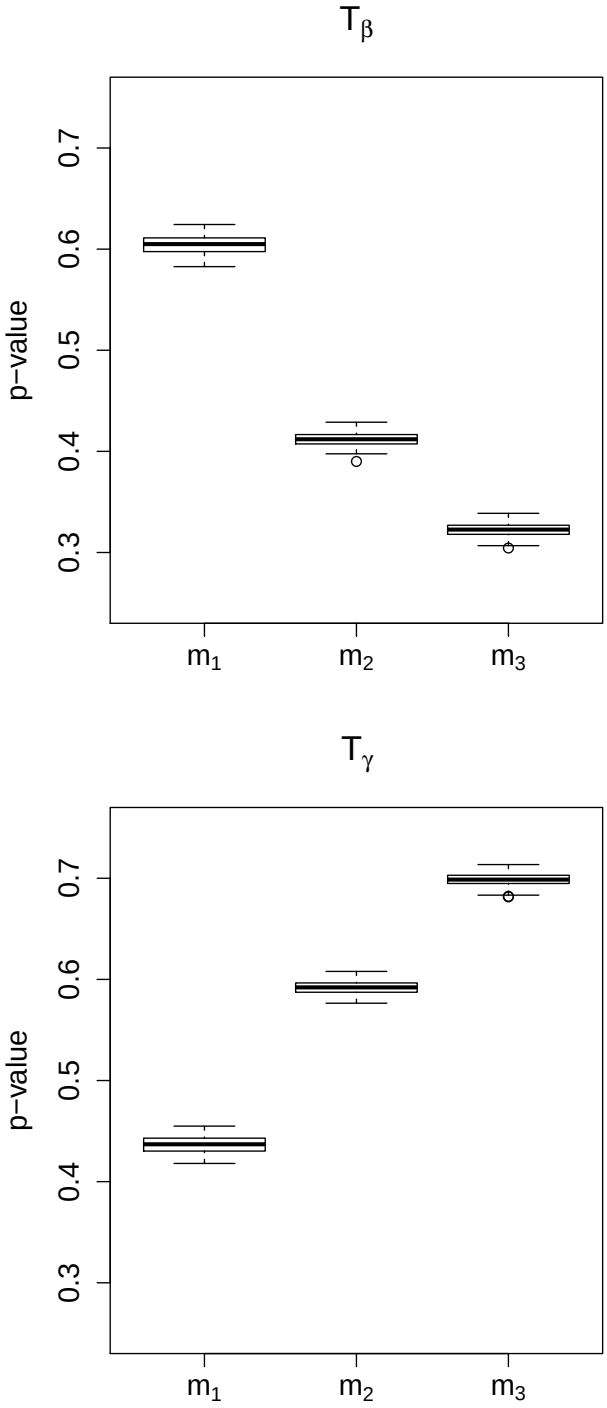


Figure 2.6: Boxplots of 100 posterior predictive $p-$values produced using the test statistic T_β (top panel) and boxplots of 100 posterior predictive $p-$values produced using the test statistic T_γ (bottom panel) under fitting all three SIR models to dataset D_1.

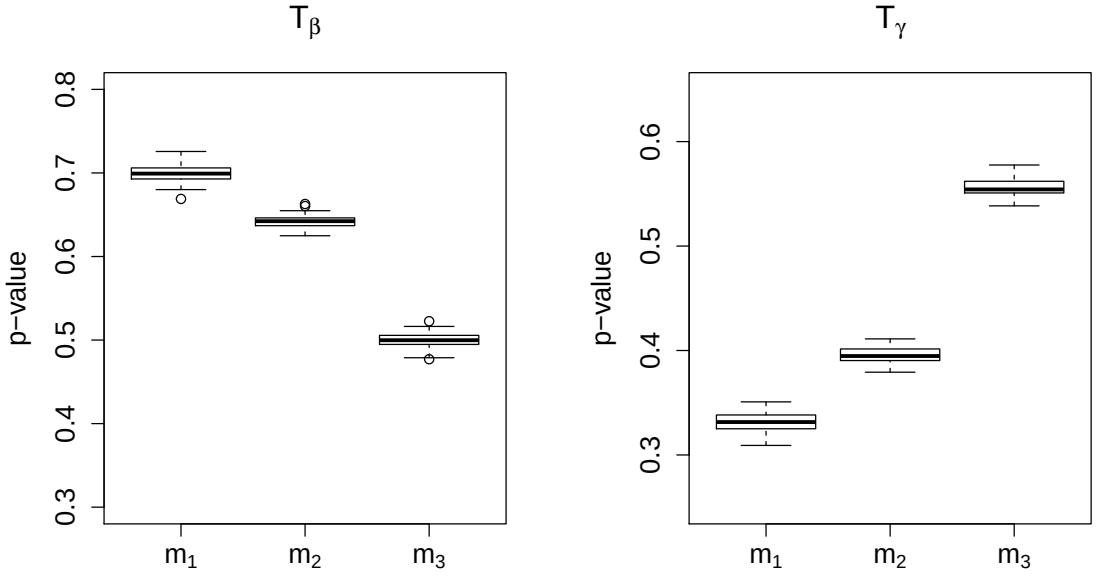


Figure 2.7: Boxplots of 100 posterior predictive $p-$values produced using the test statistics T_β (left) and T_γ (right) under fitting all three models to D_2.

2.7, the two test statistics resulted in relatively reasonable posterior predictive $p-$values under fitting both the true model (m_3) and the other misspecified models (m_1 and m_2). In other words, all three candidate models appear to fit the observed data in adequate manner.

Based on the simulation results above and many other similar results obtained using different removal datasets (not presented here), it can be concluded that our proposed test statistics T_β and T_γ seem to be poor measures of model fit and therefore insufficient to distinguish the Markovian SIR model from the other two models.

2.7 Summary and extensions

In this chapter, we have considered three different but natural methods to assess the goodness-of-fit for the SIR stochastic epidemic models. Using the labelled likelihood representation approach to examine the distribution of the

random variable $R_j - I_j,\ j = 1, ..., n_R$, over MCMC iterations appears to be inadequate to explore the infectious period distribution of the SIR models. The lack of information about exact infection times of infected individuals and the strong dependence between the infection times and the model parameters might be potential reasons for this outcome. Based on complete epidemic data, we found that time scaling of the inter-event times were also insufficient to detect the data-generating SIR model. A possible explanation for this might be that all three SIR models that were considered are assumed to have the same infection process which makes it hard to distinguish them. Two purely temporal statistics were examined to assess the goodness-of-fit for the SIR models using the posterior predictive distribution approach. As seen with the posterior predictive $p-$values results, none of these test statistics performed well in terms of detecting the data-generating SIR model.

As the ultimate goal for an epidemic goodness-of-fit task via the posterior predictive distribution approach would be to find a single omnibus goodness-of-fit test statistic that could provide an overall indication of model fit for any facet of model inadequacy, it seems hard to find a temporal epidemic summary statistic that might be sensitive to the misspecification of infectious period distribution. Nevertheless, in this chapter our proposed test statistics depend on the data only. In the next chapter, we shall consider some test (discrepancy) statistics that depend on the data as well as the model parameters.

The approach via which Bayesian inference was drawn for the SIR models relies on the assumption that the epidemic was completed by time R_{n_R}, so the final number of removals n_R is observed. However, it might be the case in practice to be interested in drawing inference for the model parameters while an epidemic is in progress. Our proposed methods can take account of this by going along the lines of the approach by O'Neill and Roberts (1999). Another natural extension regarding the gamma SIR model is to consider unknown shape parameter α and estimate it from data instead of assuming a fixed value for it. It would be interesting to see the performance of our proposed meth-

ods in the case of unknown α. An efficient and recently developed MCMC algorithm which can be directly applied in our settings to estimate α has been introduced by Xiang and Neal (2014).

CHAPTER 3

Two-stage Procedure for Assessing the Goodness-of-Fit for Epidemic Models Based on Partially Observed Data

The main results of Section 3.4 in this chapter have been published as a paper titled *Identifying the Infectious Period Distribution for Stochastic Epidemic Models Using the Posterior Predictive Check*, by Alharthi, M., O'Neill, P. D. and Kypraios, T., In Bayesian Statistics from Methods to Models and Applications (pp. 109-114), Springer International Publishing, 2015.

3.1 Introduction

This chapter describes a novel methodology for assessing the goodness-of-fit of stochastic epidemic models. As model adequacy relies traditionally on examining the discrepancy between the observed data and the corresponding quantities under the fitted model, we consider an approach that looks explicitly at the discrepancy between the observed and predicted data. We employ the posterior predictive checking approach (Gelman et al., 1996) and assess the

model adequacy based on the notion of Bayesian residuals and the posterior predictive $p-$values.

Standard methods of goodness-of-fit cannot be applied directly to epidemic models due to the fact that disease outbreak data have high dependency and are partially observed (significant parts of the disease transmission process are unobserved). Furthermore, although model comparison techniques may be used to decide between competing models, they cannot pinpoint where a model departs from data. Therefore, there is indeed a need for methods to assess the goodness-of-fit for stochastic epidemic models, and the availability of such methods would enhance understanding and potential control of the infectious diseases.

This chapter is organized as follows. In Section 3.2, we provide an overview of some background material and previous work, that is relevant to what we investigate throughout this chapter. We introduce two stochastic epidemic models to describe transmission dynamics, namely the Markovian SIR model with non-linear infection rate (O'Neill and Wen, 2012) and the two levels of mixing epidemic model (Ball et al., 1997). Then, the posterior predictive model checking approach is described. We also outline our novel two-stage method for assessing the goodness-of-fit for stochastic epidemic models.

Section 3.3 illustrates how our proposed two-stage method can be used to assess the model fit for SIR models with different infection mechanisms via the posterior predictive model checking approach. Bayesian parameter estimation and data-augmented MCMC algorithm for the SIR model with non-linear infection rate are discussed which account for unobserved infection times. We examine the removal trajectories distributions for the models of interest and then test the proposed method using simulated removal data. The sensitivity of the proposed method to R_0 is also investigated via simulation.

In Section 3.4, the model assessment task is performed via the posterior predictive model checking approach for SIR models with different infectious periods using our two-stage method. We specifically consider three SIR models

routinely employed in the literature, namely the model with exponential infectious periods, the model with gamma infectious periods and the model with fixed length infectious periods. We start by revealing the difference between these models in terms of the removal curve behaviour, and then our proposed method is applied to assess the goodness-of-fit using two simulated datasets consisting of removal times. Moreover, we introduce the SEIR model which is an extension of the SIR model to include an "exposed" state (to model the latent period of diseases). Then, we fit both SIR and SEIR models with different infectious periods to real smallpox removal data and evaluate their capabilities to provide adequate fit using our proposed two-stage method.

Model adequacy for epidemic models featuring a structured population is investigated in Section 3.5. Two epidemic models are considered, namely the homogeneous (one level of mixing) SIR model with constant infectious periods and the heterogeneous (two levels of mixing) SIR model with constant infectious periods. A data-augmented MCMC algorithm for the two levels of mixing model is introduced which accounts for unobserved infection times, and is used to derive model parameter estimates. The removal curve behaviour for the two levels of mixing model is discussed and the difference between the two models in terms of removal trajectories is illustrated. Then, we apply our two-stage procedure via the posterior predictive model checking approach to detect any homogeneity or heterogeneity in the observed data using two simulated removal datasets.

Finally, Section 4.6 gives an overview of the method and the results discussed in the previous sections as well as our method's limitations and suggestions for possible improvements and extensions.

3.2 Relevant background and previous related work

This section mainly reviews two stochastic epidemic models, that are the SIR model with non-linear infection rate (Severo (1969) and O'Neill and Wen (2012))

and the structured population SIR model with two level mixing (Ball et al., 1997). Also, it reviews one of the most commonly used approaches to assess the goodness-of-fit under the Bayesian framework, that will be used throughout this chapter, namely the posterior predictive checking approach (Gelman et al., 1996).

3.2.1 The stochastic SIR model with non-linear infection rate

This stochastic model (O'Neill and Wen, 2012) is a modification of the standard (Markovain) SIR model obtained by relaxing the linearity assumption and considering the rate of new infections to be of the form $\beta n^{-1} X(t) Y^p(t)$, where $X(t)$ and $Y(t)$ denote respectively the number of susceptible and infected individuals at time t in a closed population of size $\boldsymbol{N}$ (n initial susceptibles and one initial infective), and the power parameter p represents the degree of interaction between infective and susceptible individuals. Here we shall assume that $0 < p < 0.5$ in order to more clearly distinguish this modified SIR model from the standard SIR model as setting $p = 1$ in the former yields the latter.

In this Markovian stochastic model, the spread of the epidemic is represented through a Susceptible-Infective-Removed model (Bailey, 1975), which has the same assumptions as the standard SIR model has but with different transition probabilities, that are

$$\begin{aligned} P[(X(t+\delta t), Y(t+\delta t)) &= (x-1, y+1) | (X(t), Y(t)) = (x, y)] \\ &= \beta n^{-1} x y^p \delta t + o(\delta t), \end{aligned}$$

$$\begin{aligned} P[(X(t+\delta t), Y(t+\delta t)) &= (x, y-1) | (X(t), Y(t)) = (x, y)] \\ &= \gamma y \delta t + o(\delta t). \end{aligned}$$

These transitions correspond to an infection and a removal respectively, and β and γ denote respectively the infection and removal rate parameters.

Adjusting the infection rate in this form was first introduced in Severo (1969), where the author defined a stochastic SIR model with infection rate of the

form $\beta X^{1-p_1}(t)Y^{p_2}(t),\ 0 \leq p_1 < 1\ ; p_2 > 0$. Streftaris and Gibson (2004a) fitted a stochastic epidemic model featuring non-linear infection rate to experimental epidemic data on foot-and-mouth disease using Markov chain Monte Carlo (MCMC) methods in non-homogeneously mixing population. For real life epidemics, Bjørnstad et al. (2002) modelled the dynamics of measles using a time series SEIR type model with a non-linear transmission form. O'Neill and Wen (2012) analysed a special case of Severo's construction, where the infection rate is of the form $\beta n^{-1}X(t)Y^{p}(t),\ 0 \leq p < 1$ and described a Bayesian inference procedure for estimating the model parameters.

The motivation for modelling the infection rate in this way is based on the idea that the rate of new infections need not, in all situations, simply increase linearly in $X(t)$ and $Y(t)$. For instance, it might be the case that as an epidemic spreads, susceptibles become more aware of the risk of infection and change their behaviour accordingly. This motivates the introduction of the power coefficient parameter p, where the smaller the p, the lesser exposure of susceptibles to the infectives.

3.2.2 The stochastic SIR model with two levels of mixing

We consider the model described in Ball et al. (1997). The model is defined in a closed population that is partitioned into households of varying sizes. Suppose that the population contains h_s households of size s and let $h = \sum_{s=1}^{\infty} h_s$ be the total number of households. Then, the total number of individuals is $\boldsymbol{N} = \sum_{s=1}^{\infty} s\ h_s$. This model assumes that an infective individual remains so for a period of time called the infectious period. The infectious periods of infected individuals are assumed to be independent and identically distributed according to some random variable T_I, which can have any arbitrary but specified distribution. It is also assumed that an infective individual, during their infectious period, makes global contacts with any other given individual in the community at points of a homogeneous Poisson process having rate $\lambda_C/\boldsymbol{N}$ and it makes additional local contacts with any other given individual in the

same household at points of a homogeneous Poisson process having rate λ_H. All of these Poisson processes are assumed to be mutually independent. The homogeneous mixing within the household is overlapped on the homogeneous mixing (typically at a much lower rate) occurring in the population at large. The epidemic is initiated by one infective individual chosen uniformly at random from the $\boldsymbol{N}$ individuals in the population. The epidemic ends when there are no more infective individuals in the population.

In our settings, we make two assumptions largely for simplicity and can be easily relaxed (see e.g Ball et al. (1997) for details). First, we assume T_I to be fixed and identical for all individuals. Second, we assume that a population is divided into h households of equal size s individuals, giving a total population size of $\boldsymbol{N} = s\ h$.

Bayesian statistical inference for the two levels of mixing model given final outcome (non-temporal) data was detailed in Demiris and O'Neill (2005a), Demiris and O'Neill (2005b) and O'Neill (2009). In addition to final outbreak data, Britton et al. (2011) developed a Bayesian statistical inference method given complete temporal data for the three levels of mixing model (e.g. household, intermediate and community levels). For the temporal data case, the authors' focus was towards establishing statistical estimation procedures for models featuring an intermediate level of mixing, given complete data on a single outbreak and in particular the more realistic situation in which the observed data consist of removal times only was not considered.

3.2.3 Posterior predictive model checking

Given data $\boldsymbol{y}^{obs} = (y_1^{obs}, ..., y_n^{obs})$, we fit a model $\pi(\boldsymbol{y}^{obs}|\boldsymbol{\theta})$, which can be any arbitrary statistical model and evaluate the posterior density for a set of parameters $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta}|\boldsymbol{y}^{obs})$. Similarly, we may want to predict the likelihood of a new data observation $\boldsymbol{y}^{rep} = (y_1^{rep}, ..., y_n^{rep})$ (i.e. a replication of $\boldsymbol{y}^{obs}$) so we can

define the posterior predictive density as

$$\pi(\boldsymbol{y}^{rep}|\boldsymbol{y}^{obs}) = \int_{\boldsymbol{\theta}} \pi(\boldsymbol{y}^{rep}|\boldsymbol{\theta}, \boldsymbol{y}^{obs})\pi(\boldsymbol{\theta}|\boldsymbol{y}^{obs})d\boldsymbol{\theta} = \int_{\boldsymbol{\theta}} \pi(\boldsymbol{y}^{rep}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{y}^{obs})d\boldsymbol{\theta}. \tag{3.2.1}$$

In Chapter 2, the posterior predictive model checking approach, which is based on a test statistic that depends on data only, was introduced. Here, we use the same approach but with a discrepancy test that depends on both data and model parameters. Now, we shall recall how to assess model fit via the posterior predictive model checking approach using two Bayesian criteria namely, Bayesian residual and the posterior predictive $p-$value.

3.2.3.1 Bayesian residual

A Bayesian residual (Gelfand (1996); Gelman et al. (2013)) can be defined as the difference between the observed value and expected value under the predictive distribution, that is

$$d_j = y_j^{obs} - E(y_j^{rep}|\boldsymbol{y}^{obs}),\ j = 1, ..., n, \tag{3.2.2}$$

where

$$E(y_j^{rep}|\boldsymbol{y}^{obs}) = \int y_j^{rep}\ \pi(y_j^{rep}|\boldsymbol{y}^{obs})\ dy_j^{rep} \approx \frac{1}{M}\sum_{i=1}^{M} y_j^{rep\ (i)}, \tag{3.2.3}$$

where M denotes the number of predicted samples used and $y_j^{rep(1)}, ..., y_j^{rep(M)}$ are drawn from $\pi(\boldsymbol{y}_j^{rep}|\boldsymbol{y}^{obs})$.

Having obtained $d_j,\ j = 1, ..., n$, we can use them to detect any possible lack of fit in the fitted model. Ideally, we look for symmetry of the posterior distribution of d_j around zero. Further, it is worth calculating the sum of the squared residuals $SSR = \sum_{j=1}^{n} d_j^2$ which can provide an overall measure of fit, where large values provide evidence of poor model fit (Carlin and Louis, 2011).

3.2.3.2 The posterior predictive $p-$value

The posterior predictive $p-$value can be defined as the probability that the replicated data could be more extreme than the observed data using an ap-

propriate measure of discrepancy. In the Bayesian approach, the discrepancy measure is allowed to be a function of both data and model parameters (Gelman et al., 2013). If we have $D(\boldsymbol{y}, \boldsymbol{\theta})$ as a selected discrepancy measure (it is a continuous positive real-valued function) that quantifies the difference between data (observed and replicated) under the fitted model, and M samples from the posterior density of the model parameters $\boldsymbol{\theta}$, then the posterior predictive $p-$value (ppp-value) can be obtained as follows. We construct a replicated dataset for each draw from the posterior distribution of the model parameters, then evaluate discrepancy measure for the drawn parameter values and each of the datasets, observed and replicated, and produce a scatter plot with the discrepancy of the observed data on the x-axis and the discrepancy of the replicated data on the y-axis. If the model offers a good fit, replicated data generated under the model should look similar to observed data. The posterior predictive p-value is the proportion of times the predictive discrepancy $D(\boldsymbol{y}^{rep}, \boldsymbol{\theta})$ is greater than or equal to the observed discrepancy $D(\boldsymbol{y}^{obs}, \boldsymbol{\theta})$, that is

$$\begin{aligned} ppp-\text{value} &= P\left(D(\boldsymbol{y}^{rep}, \boldsymbol{\theta}) \geq D(\boldsymbol{y}^{obs}, \boldsymbol{\theta}) \middle| \boldsymbol{y}^{obs}\right) \\ &\approx \frac{1}{M}\sum_{i=1}^{M} \mathbb{1}_{\{D(\boldsymbol{y}^{rep}, \boldsymbol{\theta}^{(i)}) \geq D(\boldsymbol{y}^{obs}, \boldsymbol{\theta}^{(i)})\}}, \end{aligned} \tag{3.2.4}$$

where for each posterior draw $\theta^{(i)}$, $i = 1, ..., M$, we simulate the model to obtain a value (vector) of $\boldsymbol{y}^{rep(i)}$.

Any extreme values of ppp-value (close to 0 or 1 with an optimal value of 0.5) are supportive of a lack of fit, indicating that the model would be unlikely to predict the data we have observed (Gelman et al., 2013).

Various discrepancy measures to test the overall goodness-of-fit exist. Examples of discrepancy measures that can be used include

$$\chi^2\text{discrepancy} : D(\boldsymbol{y}, \boldsymbol{\theta}) = \sum_j \frac{\left(y_j - E(y_j|\boldsymbol{\theta})\right)^2}{\text{var}(y_j|\boldsymbol{\theta})}, \tag{3.2.5}$$

(see e.g. Gelman et al. (2013)), and

$$\text{Freeman-Tukey discrepancy} : D(\boldsymbol{y}, \boldsymbol{\theta}) = \sum_j \left(\sqrt{y_j} - \sqrt{E(y_j|\boldsymbol{\theta})} \right)^2, \quad (3.2.6)$$

(see e.g. Freeman and Tukey (1950); Brooks et al. (2000)).

The quantities $E(y_j|\boldsymbol{\theta}^{(i)})$ and $\text{var}(y_j|\boldsymbol{\theta}^{(i)})$ in (3.2.5) and (3.2.6) can be calculated at each MCMC iteration i if $E(y_j|\boldsymbol{\theta})$ and $\text{var}(y_j|\boldsymbol{\theta})$ are available in closed forms. Otherwise, Monte Carlo simulation can be used to estimate them provided that samples $\boldsymbol{y}^{(i)}$ can be drawn from $\pi(\boldsymbol{y}|\boldsymbol{\theta}^{(i)})$.

Now, we shall review some literature about Bayesian model assessment for stochastic epidemics that is relevant to the materials and ideas discussed throughout this chapter. Some other model choice publications for epidemic models will be presented in the next chapter.

Forrester (2006) employed the Freeman-Tukey discrepancy measure as a checking function to validate their model for the transmission of nosocomial pathogens using routinely collected surveillance hospital data. The author used the posterior predictive assessment method to perform a comparison between the observed count data $\boldsymbol{y}^{obs}$ (the number of detected patients during the observation period in the observed data) and the replicated count data $\boldsymbol{y}^{rep}$ (the number of detected patients during the observation period in the replicated data). However, the author did not explore this approach for assessing the model fit in any detail and suggested further investigation in order to make firm conclusions. Throughout this chapter, we shall make use of this discrepancy measure as well as the χ^2 measure, bearing in mind that Forrester's epidemic setting is different from our epidemic setting in two ways. First, in Forrester's setting the posterior expectation at a given Markov chain iteration i, $E(y_j|\boldsymbol{\theta}^{(i)})$ is available in a closed form which in turn facilitates the computation work, however in our settings this quantity will be estimated via Monte Carlo simulation which requires more effort in terms of computational time. Second, the data used in Forrester's setting to check the model fit are aggregated count data, whereas our data are temporal removal data.

Boys and Giles (2007) considered an SEIR model in which the time-dependence of the removal rate is introduced and modelled as a step function. Application of their model was made to symptom appearance data from the smallpox epidemic in Abakaliki, Nigeria. The authors assessed the model fit by contrasting the predictive distributions for selected removal times for both their more complex time-inhomogeneous model and a simpler time-homogeneous removal rate model and comparing them with the observed removal data. It was found that both models give a reasonable fit early in the epidemic. However, later in the epidemic, the simpler model fails to capture the change in removal rate and therefore over-predicts the removal times, whereas the more flexible model removes much of this lack-of-predictive fit by explicitly modelling such changes.

Zhang (2014) compared the posterior predictive distribution of the cumulative number of infection events under various fitted models to the observed cumulative number of cases curve. The modelling framework used was that of the discrete time individual-level stochastic epidemic model described in Deardon et al. (2010). Models were misspecified by changing a time-varying infectivity curve function as well as the infection distance kernel, where epidemics were simulated from a true model and then the true data-generating model and various misspecified models were fitted to the simulated data. It was found that using the posterior predictive distribution of the cumulative number of cases to determine model adequacy did not work in practice.

However, our epidemic settings are distinct from the setting of Zhang (2014) in two ways. First, the manner of model assessment in Zhang'set-up was somewhat unrealistic, in that it was assumed that the latent and the infectious period for each individual were known constants (no need to estimate them from the data) which is not the case in our epidemic settings in which the available data are the removal times. Second, the epidemic models considered here are continuous time models rather than discrete time models.

3.2.4 The two-stage procedure and the aims of current work

In our proposed two-stage procedure, two predictive distributions are examined, namely the predictive distribution of the final size of the epidemic and the predictive distribution of the removal times. At the first stage, we look at the final size predictive distribution; if the observed final size value lies within it, we can then proceed and apply the second stage of the method which involves implementing the posterior predictive checking approach for the removal times distribution conditioning on the observed final size of the epidemic. An immediately inadequate fit of the epidemic model to the data can be detected when having the observed value of the final size in the tails of its posterior predictive distribution, whereas an eventually poor fit may be identified using the posterior predictive checking procedure when examining the removal times predictive distribution.

To put it another way, suppose that we observe an epidemic outbreak that results in a total of n_R removals in a closed population of size $\boldsymbol{N}$ (n initial susceptibles and one initial infective) and let $\boldsymbol{R}^{obs} = (R_1^{obs}, ..., R_{n_R}^{obs})$ and $\boldsymbol{R}^{rep} = (R_1^{rep}, ..., R_{n_R}^{rep})$ denote the observed and replicated removal times respectively. Then, Algorithm 5 can be used to apply our proposed two-stage method.

The second stage in Algorithm 5 can be implemented as follows. First, we select at random M' posterior samples from $\pi(\boldsymbol{\phi}|\boldsymbol{R}^{obs})$ and use $\boldsymbol{\phi}^{(k)}, k = 1, ..., M'$ to generate M' replicated epidemic removal curves ($\boldsymbol{R}^{rep(k)}$) from the model of interest where the final size of the simulated epidemics are matched to the observed final size, i.e. $n_R^{rep(k)} = n_R$ and $\boldsymbol{\phi}$ denotes a vector of the model parameters. Then, we compare graphically $\boldsymbol{R}^{obs}$ to $\pi(\boldsymbol{R}^{rep}|\boldsymbol{R}^{obs})$, the resultant conditioning posterior predictive distribution of removal times. To this end, the Bayesian residuals d_j defined below can be used to assess the goodness-of-fit of the model, that is for $n_R^{rep(k)} = n_R$ $k = 1, ..., M'$,

$$d_j = R_j^{obs} - E(R_j^{rep}|\boldsymbol{R}^{obs}),\ j = 1, ..., n_R, \tag{3.2.7}$$

Algorithm 5 Generic algorithm for our two-stage method

1. Given $\boldsymbol{R}^{obs}$, fit an SIR model using MCMC methods to obtain samples from the joint posterior density $\pi(\boldsymbol{\phi}|\boldsymbol{R}^{obs})$, where $\boldsymbol{\phi}$ is a vector of the model parameters.

2. **At the first stage:**

2.1. Draw $\boldsymbol{\phi}^{(i)} \sim \pi(\boldsymbol{\phi}|\boldsymbol{R}^{obs})$, $i = 1, ..., M$.

2.2. Simulate M epidemic realisations from the SIR model using $\boldsymbol{\phi}^{(i)}$ and record the epidemic final sizes. This produces a final size predictive distribution.

2.3. Check whether or not the observed final size lies within the final size predictive distribution.

3. **At the second stage:**

3.1. Choose at random $M' \leq M$ posterior samples from $\pi(\boldsymbol{\phi}|\boldsymbol{R}^{obs})$ and use $\boldsymbol{\phi}^{(k)}, k = 1, ..., M'$ to generate epidemic realisations from the model conditioning on $n_R^{rep(k)} = n_R$.

3.2. Given that $n_R^{rep(k)} = n_R$, compare $\boldsymbol{R}^{obs}$ to $\pi(\boldsymbol{R}^{rep}|\boldsymbol{R}^{obs})$ graphically as well as using the Bayesian model assessment tools.

where the mean predicted time of the jth removal can be approximated as

$$\begin{aligned} E(R_j^{rep}|\boldsymbol{R}^{obs}) &= \int R_j^{rep} \pi(\boldsymbol{R}^{rep}|\boldsymbol{R}^{obs}) dR_j^{rep} \\ &\approx \frac{1}{M'} \sum_{k=1}^{M'} R_j^{rep(k)}. \end{aligned} \tag{3.2.8}$$

As mentioned earlier, the quantities $E(R_j|\boldsymbol{\phi}^{(k)})$ and $\text{var}(R_j|\boldsymbol{\phi}^{(k)})$ are not available in closed forms in our epidemic settings. Therefore, Monte Carlo simulation will be used to estimate them.

The calculation of the posterior predictive $p-$value using the χ^2 discrepancy measure can be performed for the model via Monte Carlo simulations as follows (the calculation of the posterior predictive $p-$value using Freeman-Tukey discrepancy measure can be done in a similar way).

For each $\boldsymbol{\phi}^{(k)}$, $k = 1, ..., M'$, perform the following two steps:

1. Simulate from the model of interest D removal curves that must have the same length as the observed one (using rejection sampling), say $\boldsymbol{R}^{(1)}, \boldsymbol{R}^{(2)}, ..., \boldsymbol{R}^{(D)}$, such that $\boldsymbol{R}^{(d)} = (0 = R_1^{(d)}, R_2^{(d)}, ..., R_{n_R}^{(d)})$, where $d = 1, ..., D$. Then, estimate $E(R_j|\boldsymbol{\phi}^{(k)})$ and $\text{var}(R_j|\boldsymbol{\phi}^{(k)})$ using these removal times vectors as follows:
$$E(R_j|\boldsymbol{\phi}^{(k)}) \approx \frac{1}{D}\sum_{d=1}^{D} R_j^{(d)},$$
and
$$\text{var}(R_j|\boldsymbol{\phi}^{(k)}) \approx \frac{1}{D-1}\sum_{d=1}^{D}(R_j^{(d)} - \overline{R_j^{(d)}})^2,$$
where $j = 2, ..., n_R$.

2. Using $\boldsymbol{R}^{obs}$, $\boldsymbol{R}^{rep(k)}$ (obtained when computing the Bayesian residuals), $E(R_j|\boldsymbol{\phi}^{(k)})$ and $\text{var}(R_j|\boldsymbol{\phi}^{(k)})$, calculate
$$D(\boldsymbol{R}^{obs}, \boldsymbol{\phi}^{(k)}) = \sum_{j=2}^{n_R} \frac{\left(R_j^{obs} - E(R_j|\boldsymbol{\phi}^{(k)})\right)^2}{\text{var}(R_j|\boldsymbol{\phi}^{(k)})},$$
and
$$D(\boldsymbol{R}^{rep(k)}, \boldsymbol{\phi}^{(k)}) = \sum_{j=2}^{n_R} \frac{\left(R_j^{rep(k)} - E(R_j|\boldsymbol{\phi}^{(k)})\right)^2}{\text{var}(R_j|\boldsymbol{\phi}^{(k)})}.$$

Repeating the above two steps M' times enables us to estimate the ppp-value by the proportion of times that $D(\boldsymbol{R}^{rep(k)}, \boldsymbol{\phi}^{(k)})$ exceeds $D(\boldsymbol{R}^{obs}, \boldsymbol{\phi}^{(k)})$, that is

$$\begin{aligned} ppp - \text{value} &= P\left(D(\boldsymbol{R}^{rep}, \boldsymbol{\phi}) \geq D(\boldsymbol{R}^{obs}, \boldsymbol{\phi})\big|\boldsymbol{R}^{obs}\right) \\ &\approx \frac{1}{M'}\sum_{k=1}^{M'} \mathbb{1}_{\{D(\boldsymbol{R}^{rep(k)}, \boldsymbol{\phi}^{(k)}) \geq D(\boldsymbol{R}^{obs}, \boldsymbol{\phi}^{(k)})|\boldsymbol{R}^{obs}\}}, \end{aligned} \tag{3.2.9}$$

Having done that, a scatter plot can be made to make a graphical assessment by plotting $D(\boldsymbol{R}^{obs}, \boldsymbol{\phi}^{(k)})$ versus $D(\boldsymbol{R}^{rep(k)}, \boldsymbol{\phi}^{(k)})$.

In this investigation, we are interested primarily in assessing the goodness-of-fit for epidemic models given only temporal removal data. Throughout this

chapter, three SIR epidemic settings are considered, namely the SIR model with different forms of the infection rate, the SIR model with different infectious periods and the SIR model with different population structures. First, we investigate the adequacy of fitting the standard Markovain SIR model introduced in Chapter 2 and the modified Markovain SIR model introduced above in Section 3.2.1 to partially observed outbreak data. The second setting is concerned with identifying the infectious period distribution of the SIR model in which we consider three SIR models routinely employed in the literature, namely the model with exponential infectious periods, the model with gamma infectious periods and the model with fixed length infectious periods. Lastly, the homogeneous SIR model with constant infectious periods and the two levels of mixing SIR model introduced in Section 3.2.2 are compared and assessed when fitting them to removal outbreak data.

3.3 Model assessment for SIR models with different infection mechanisms

In this section, we consider an epidemic setting in which the adequacy of two Markovian SIR models with different forms of infection rate is assessed. The first model denoted by m_1 is an SIR model with the familiar infection rate form, that is $\beta n^{-1} X(t) Y(t)$. The other epidemic model denoted by m_2 is a modification of m_1 model by relaxing the linearity assumption and considering the rate of new infections to be of the form $\beta n^{-1} X(t) Y^p(t)$, $0 < p < 0.5$. The restriction $0 < p < 0.5$ is made in order to more clearly distinguish this modified SIR model from the standard SIR model as setting $p = 1$ in the former yields the latter.

When modelling an infectious disease, one key question of particular importance is that how infectious is the disease? That is, at what rate do individuals become infected by a disease. Knowing such information can help us to implement strategies in order to prevent a disease from spreading through a

population. In our epidemic setting, this question is answered by estimating the infection rate parameter β using the observed outbreak data. To avoid underestimating or overestimating this parameter, we seek for a method to distinguish between these two epidemic models as they have different infection mechanisms.

Bayesian inference for the first model m_1 was illustrated in Chapter 2, so we now consider Bayesian parameter estimation for the second model m_2. O'Neill and Wen (2012) described an MCMC algorithm for sampling from the Bayesian joint posterior density of β, γ and p given data comprising the removal times $\boldsymbol{R}$. We now recall the details.

3.3.1 Bayesian inference via MCMC for the SIR model with non-linear infection rate given removal data

Suppose that an epidemic outbreak which results in a total of n_R removals in a closed population of size $\boldsymbol{N}$, n initial susceptibles and one initial infective is observed on $(I_1, R_{n_R}]$, so that all removals are assumed to be observed. Let $0 = R_1, R_2, ..., R_{n_R}$ be the observed ordered removal times and write $\boldsymbol{R} = (R_1, ..., R_{n_R})$. Also, let $I_1 < 0$ be the first unobserved infection time and $\boldsymbol{I} = (I_2, ..., I_{n_R})$ be the rest of unobserved ordered infection times, where $I_k < I_{k+1} < R_k$ for $k = 1, 2, ..., n_R - 1$. Then, the joint likelihood of $\boldsymbol{I}$ and $\boldsymbol{R}$ given the model parameters is

$$\begin{aligned} L(\boldsymbol{I}, \boldsymbol{R}|\beta, \gamma, p, I_1) &= \prod_{j=2}^{n_R} \beta n^{-1} X(I_j-) Y^p(I_j-) \times \prod_{j=1}^{n_R} \gamma Y(R_j-) \\ &\times \exp\left(- \int_{I_1}^{R_{n_R}} (\beta n^{-1} X(t) Y^p(t) + \gamma Y(t)) dt \right), \end{aligned} \tag{3.3.1}$$

where I_j- denotes the time just prior to I_j and R_j- is defined similarly.

We assign an independent gamma prior distribution for each rate parameter, namely Gamma$(\lambda_\zeta, \nu_\zeta)$, where $\zeta = \beta,\ \gamma$, and assume a priori that $p \sim$

$U(0, 0.5)$ and $-I_1 \sim \text{Exp}(\theta)$, where $\text{Gamma}(a, b)$ represents a gamma distribution with mean a/b and variance a/b^2, $U(a, b)$ denotes a continuous uniform distribution on the interval (a, b) and $\text{Exp}(a)$ denotes an exponential distribution with mean a^{-1}. Then, by multiplying the likelihood and the priors, we obtain the following full conditional posterior densities for β, γ and I_1, that are

$$\beta|\gamma, p, I_1, \boldsymbol{I}, \boldsymbol{R} \sim \text{Gamma}\left(\lambda_\beta + n_R - 1, \nu_\beta + n^{-1}\int_{I_1}^{R_{n_R}} X(t)Y^p(t)dt\right), \quad (3.3.2)$$

$$\gamma|\beta, p, I_1, \boldsymbol{I}, \boldsymbol{R} \sim \text{Gamma}\left(\lambda_\gamma + n_R, \nu_\gamma + \int_{I_1}^{R_{n_R}} Y(t)dt\right), \quad (3.3.3)$$

and

$$I_1|\beta, \gamma, p, \boldsymbol{I}, \boldsymbol{R} \sim I_2 - \text{Exp}(\beta + \gamma + \theta). \quad (3.3.4)$$

Also, we have the following proportional full posterior densities for p and $\boldsymbol{I}$

$$\pi(p|\beta, \gamma, I_1, \boldsymbol{I}, \boldsymbol{R}) \propto \prod_{j=2}^{n_R} Y^p(I_j-) \times \exp\left(-\beta n^{-1}\int_{I_1}^{R_{n_R}} X(t)Y^p(t)dt\right), \quad (3.3.5)$$

$$\begin{aligned}\pi(\boldsymbol{I}|\beta, \gamma, p, I_1, \boldsymbol{R}) \quad &\propto \quad \prod_{j=2}^{n_R} X(I_j-)Y^p(I_j-) \times \prod_{j=1}^{n_R} Y(R_j-) \\ &\times \quad \exp\left(-\int_{I_1}^{R_{n_R}} (\beta n^{-1}X(t)Y^p(t) + \gamma Y(t))dt\right).\end{aligned} \quad (3.3.6)$$

The model parameters β, γ and the initial infection time I_1, all can be updated using Gibbs sampling steps as they have closed forms of the posterior density. However, the power parameter p and the infection times $\boldsymbol{I}$ need to be updated using Metropolis-Hastings steps as their posterior densities are not available explicitly. Algorithm 6 can be used to do so.

3.3.1.1 General remarks on parameter estimation

It is of interest to explore some properties of Bayesian estimation for the model parameters β, γ and p via simulation. A simulation study was carried out for three population sizes, $\boldsymbol{N} = 70,\ 200$ and 500. The true model parameters were

Algorithm 6 MCMC algorithm for updating β, γ, p, I_1 and $\boldsymbol{I}$

1. Initialise algorithm with $\beta^0, \gamma^0, p^0, I_1^0$ and $\boldsymbol{I}^0$.

2. Repeat the following steps a large number of times, say M.

* Update β by using Gibbs Sampler.

* Update γ by using Gibbs Sampler.

* Update I_1 by using Gibbs Sampler.

* Update $\boldsymbol{I}$ by using Metropolis-Hastings Sampler as follows:

a. Choose uniformly one of the existing infection times, say I_j.

b. Draw a proposed replacement infection time I_j^* uniformly on (I_1, R_{n_R}).

c. Draw a uniform random variable $U \in (0,1)$.

d. Accept the proposed infection time if

$$U \leq \min\left\{\frac{\pi(\boldsymbol{I}^*|\beta,\gamma,p,I_1,\boldsymbol{R})}{\pi(\boldsymbol{I}|\beta,\gamma,p,I_1,\boldsymbol{R})},1\right\},$$

where $\boldsymbol{I}^*$ denotes the set of current infection times but with I_j replaced by I_j^*.

* Update p by using Metropolis-Hastings Sampler as follows:

a. Draw a proposed value p^* uniformly on $(0, 0.5)$.

b. Accept the proposed value if

$$U \leq \min\left\{\frac{\pi(p^*|\beta,\gamma,I_1,\boldsymbol{I},\boldsymbol{R})}{\pi(p|\beta,\gamma,I_1,\boldsymbol{I},\boldsymbol{R})},1\right\}.$$

3. Return $\{\beta^0, \beta^1, ..., \beta^M\}$, $\{\gamma^0, \gamma^1, ..., \gamma^M\}$ and $\{p^0, p^1, ..., p^M\}$.

$\beta = 3, \gamma = 0.2$ and $p = 0.3$ and the final sizes were 67, 179 and 436, which are typical in the sense that they lie well within the major outbreak part of the final size distributions.

The negative correlation of β and p : The parameters β and p reveal negative correlation as shown in the scatter plot of β and p, obtained from the MCMC outcomes, in Figure 3.1 but this becomes less of an issue as the population

size increases since the posterior variance of both β and p decreases (see Table 3.1). O'Neill and Wen (2012) attributed this negative correlation to the fact that different sets of β, γ and p can result in similar posterior density values given only the removal times.

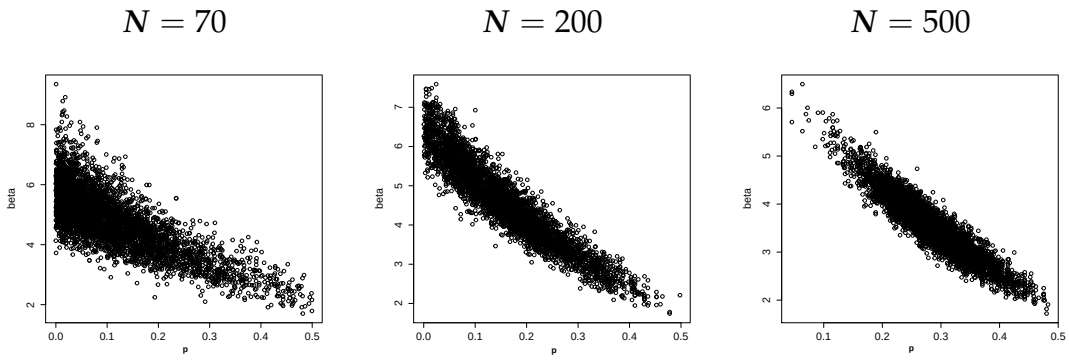


Figure 3.1: Scatter plots for p and β obtained from their posterior samples for different population sizes.

The precision of estimation: Table 3.1 and Figure 3.2 indicate that the larger the population size the better the estimation and the lower the posterior variance, particularly for p and β parameters. It is our experience that more data are needed in order to have reasonable estimates for p and β in this kind of epidemic model.

$\boldsymbol{N}$	$E(\beta\|\boldsymbol{R})$	$V(\beta\|\boldsymbol{R})$	$E(\gamma\|\boldsymbol{R})$	$V(\gamma\|\boldsymbol{R})$	$E(p\|\boldsymbol{R})$	$V(p\|\boldsymbol{R})$
70	4.779	1.115	0.346	0.011	0.120	0.011
200	4.594	1.182	0.267	0.004	0.161	0.001
500	3.524	0.383	0.276	0.002	0.286	0.004

Table 3.1: Posterior mean and variance for β, γ and p as population size varies. True values are: $\beta = 3$, $\gamma = 0.2$ and $p = 0.3$

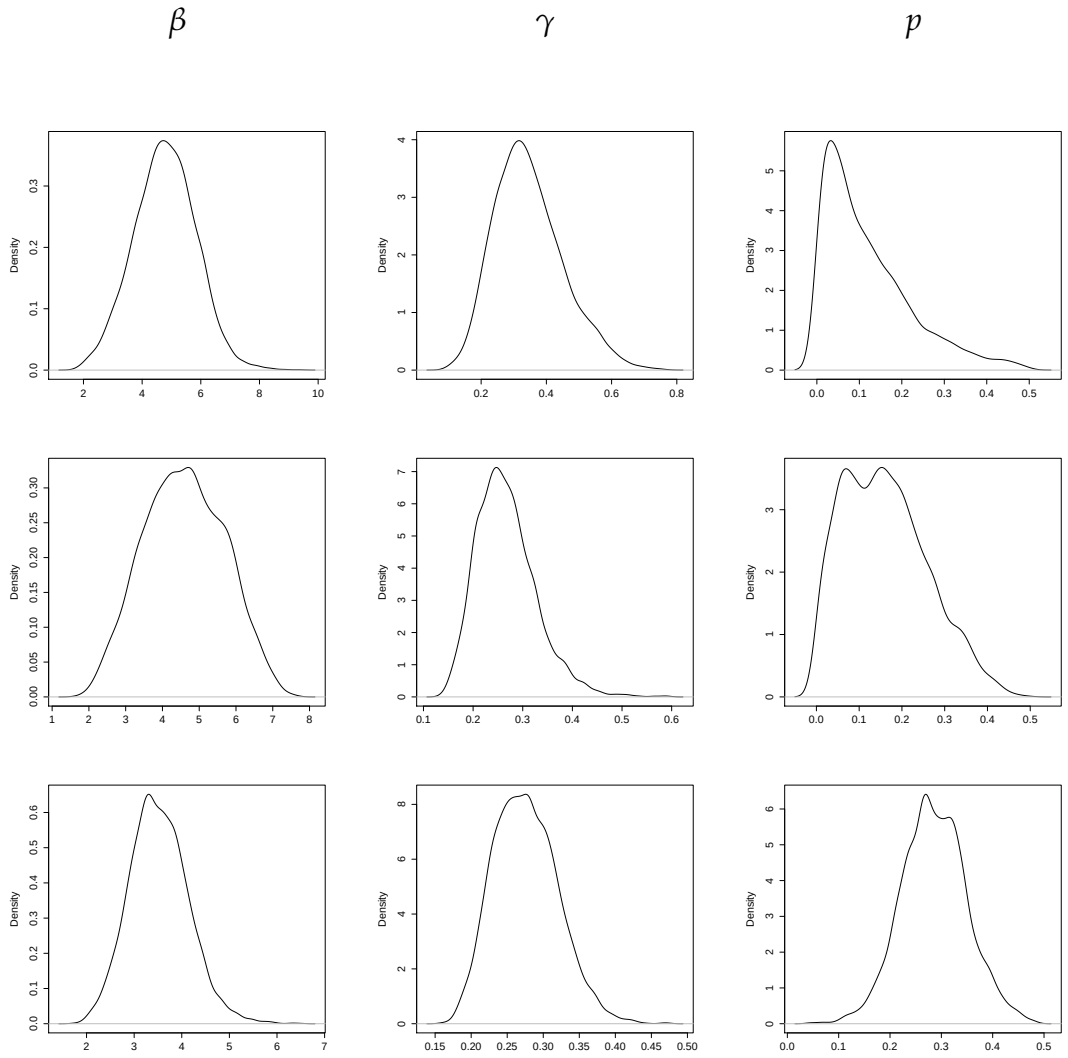


Figure 3.2: Posterior densities for the model parameters with various population size. Top, middle and bottom panels correspond to $N = 70, 200$ and 500.

3.3.2 Removal trajectories and their predictive distribution

Having obtained the desired samples from the joint posterior distribution of the model parameters, we can use them to sample from the model of interest and get the posterior predictive distribution of the removal times. The idea is to compare the observed curve of the removal times to the predicted ones under each model.

Now, we shall examine the behaviour of removal trajectories for the two afore-

mentioned stochastic SIR models via simulation.

3.3.2.1 The behaviour of removal curves for the two SIR models

Here a simulation study was carried out to see the behaviour of the two SIR models (m_1 and m_2) in terms of the removal times trajectories. For making a convincing comparison, we let the mean infectious period be the same in both models as well as the final size distributions to be similar in the sense that the major outbreak components of the final size distributions are peaked around the same values for both models. We considered a population of size $\boldsymbol{N} = 500$ $(n = 499, a = 1)$ and simulated 5000 removal curves from each SIR model with $\beta = 0.45, \gamma = 0.2$ for the model m_1 and $\beta = 3, \gamma = 0.2$ and $p = 0.3$ for the model m_2. Figure 3.3 reveals that the distribution of removal curves of the two SIR models look very different.

To make this difference even obvious, we performed another simulation study as detailed in Table 3.2. In this simulation study, we looked at the average of 1000 realisations of removal curve (simulated from each model) conditioning on a particular epidemic final size. Two population sizes were looked at, $\boldsymbol{N} =$ 100 and 300. The results are shown in Figure 3.4, in which it is evident, on average, that removal trajectories vary in the two SIR models.

Simulation	(n, a)	Final size	Model	β	γ	p
$\boldsymbol{I}$	$(99, 1)$	90	m_1	2.6	1	1
			m_2	9	1	0.4
$\boldsymbol{II}$	$(299, 1)$	283	m_1	0.3	0.1	1
			m_2	3	0.1	0.2

Table 3.2: Summary of the final size of each epidemic and the parameter values used for both epidemic models, m_1 and m_2.

Having observed that the two considered SIR models, m_1 and m_2, behave differently in terms of removal curves, we shall take advantage of this observa-

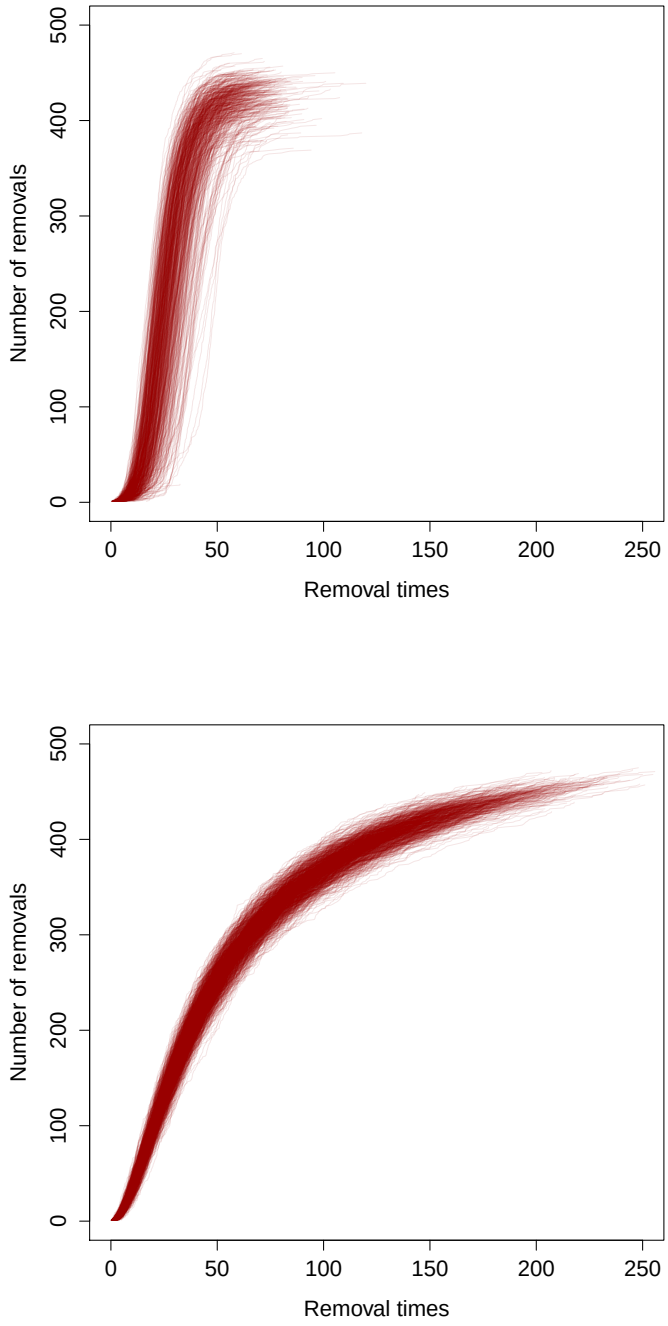


Figure 3.3: The distribution of 5000 removal curves drawn from each SIR model in a closed population of size $\boldsymbol{N} = 500$ with parameter values: $\beta = 0.45, \gamma = 0.2$ for the model m_1 (top) and $\beta = 3, \gamma = 0.2$ and $p = 0.3$ for the model m_2 (bottom).

tion to distinguish between these two models when fitting them to partially observed outbreak data. This can be done by applying our proposed two-

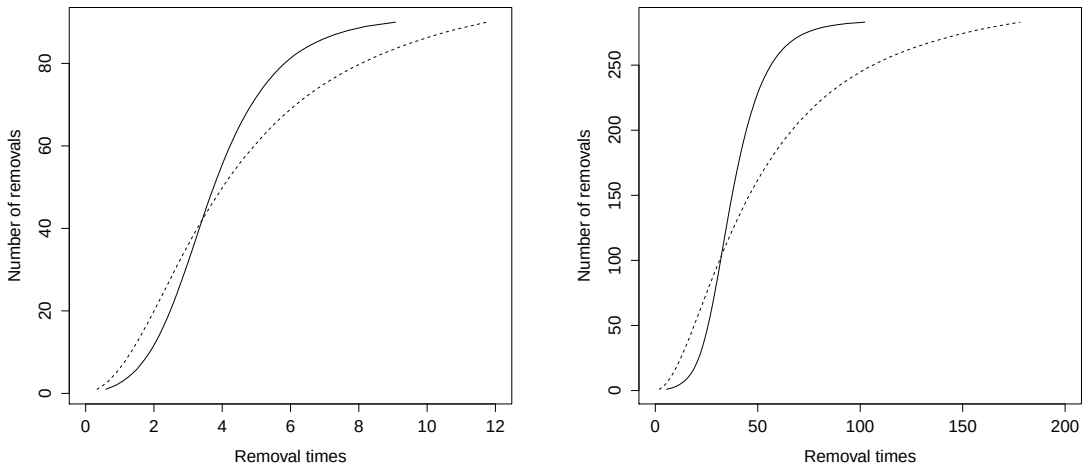


Figure 3.4: The averages of 1000 realisations of removal curves obtained from each SIR model conditioning on particular final sizes, 90 and 283 (left to right), where the solid and dashed lines correspond to m_1 and m_2 respectively. Left and right plots refer to simulations $\boldsymbol{I}$ and $\boldsymbol{II}$ respectively.

stage method introduced in Section 3.2.4.

Throughout the analysis, non-informative gamma prior distributions with parameters $\lambda_\beta = \lambda_\gamma = 1$ and $\nu_\beta = \nu_\gamma = 0.001$ are assigned to the parameters β and γ for both SIR models. Also, non-informative prior distributions $U(0, 0.5)$ and $\text{Exp}(0.001)$ are assigned to the power parameter p and $-I_1$ respectively.

The two Bayesian model assessment criteria introduced in Section 3.2.3 will be employed to make a formal assessment for the model fit for each model.

3.3.3 Application of the two-stage method to simulated outbreak data and results

We simulated removal data according to each of the two SIR models, and then aimed to identify the correct model using our two-stage procedure. Under each scenario, we simulated and analysed several outbreak datasets. Our findings are based on one typical simulated dataset for each scenario.

3.3.3.1 Sub-study I: SIR model with non-linear infection rate data

First, we simulated some removal data from a Markovian SIR with non-linear infection rate (m_2), where $0 < p < 0.5$. A closed population of size $N = 200$ individuals was considered and the rate of infection per possible contact was set to $\beta = 2.5$. The infectious periods were drawn from an exponential distribution with rate $\gamma = 0.2$. The power parameter was set to $p = 0.3$. We assumed that the epidemic is initiated by a single infected individual and the other $n = 199$ individuals were assumed to be initially susceptible to the disease. The resulting data consisted of 178 removal times. Then, we fitted the correctly specified model m_2 (the modified SIR model) to these observed data using MCMC method described in Algorithm 6 and we also fitted the misspecified model m_1 (the standard SIR model) to these data using the same MCMC algorithm with p being fixed to 1. After a burn-in of 2000 samples and thinning the chain by taking each $5th$ value, 5000 samples were obtained from the joint posterior distribution of β, γ and p for m_2 and from the joint posterior distribution for β and γ for m_1.

Having obtained these posterior samples, we are in a position to use them to simulate 5000 removal curves from the two SIR models, that is the posterior predictive distribution of removal times. Then, we check first whether or not the observed final size ($n_R - 1 = 177$) lies within the predictive final size distributions and then proceed to apply a formal Bayesian diagnostic tool for the goodness-of-fit by conditioning on the observed final size. It is graphically obvious from Figure 3.5 that the observed data curve can be seen as a realisation from the modified SIR model (m_2), the true model that has generated the data.

Additionally and more importantly, Figure 3.6 shows that the observed final size ($n_R - 1 = 177$) fits well within the final size predictive distributions for both models which in turn enables us to apply the second stage of our procedure. Figure 3.7 shows two plots of the average of 500 removal trajectories along with 95% percentile intervals generated from the removal times pre-

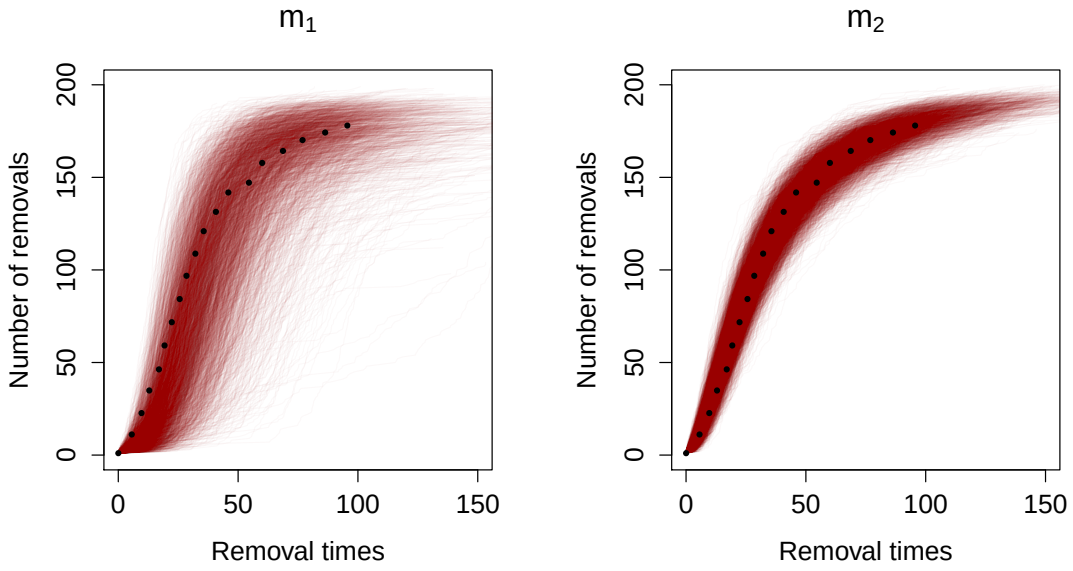


Figure 3.5: The posterior predictive distribution of the removal times for the two SIR models (left and right plots refer to m_1 and m_2 models respectively) based on 5000 samples from the joint posterior density of each model parameters, where dotted line indicates the observed data.

dictive distributions for the two SIR models where the final sizes are matched with the observed one. This can be done simply by first choosing at random $M' = 500$ posterior samples from each $M = 5000$ posterior samples that were obtained from the joint posterior distribution of β, γ and p for m_2 and from the joint posterior distribution for β and γ for m_1 and then keeping sampling (until the desired sample size is obtained) from the models using these $M' = 500$ posterior samples while rejecting epidemic realisations that do not match the observed final size. We can see from these plots that the proposed diagnostic method seems to work very well and we are able to recover the true SIR model that was used to generate these removal data.

Now, we shall recall our criteria to measure the goodness-of-fit under the Bayesian framework. To proceed and apply the two Bayesian model assessment criteria, we chose at random $M' = 500$ posterior samples from each $M = 5000$ posterior samples that were obtained from the joint posterior distributions of β, γ and p for m_2 and β and γ for m_1.

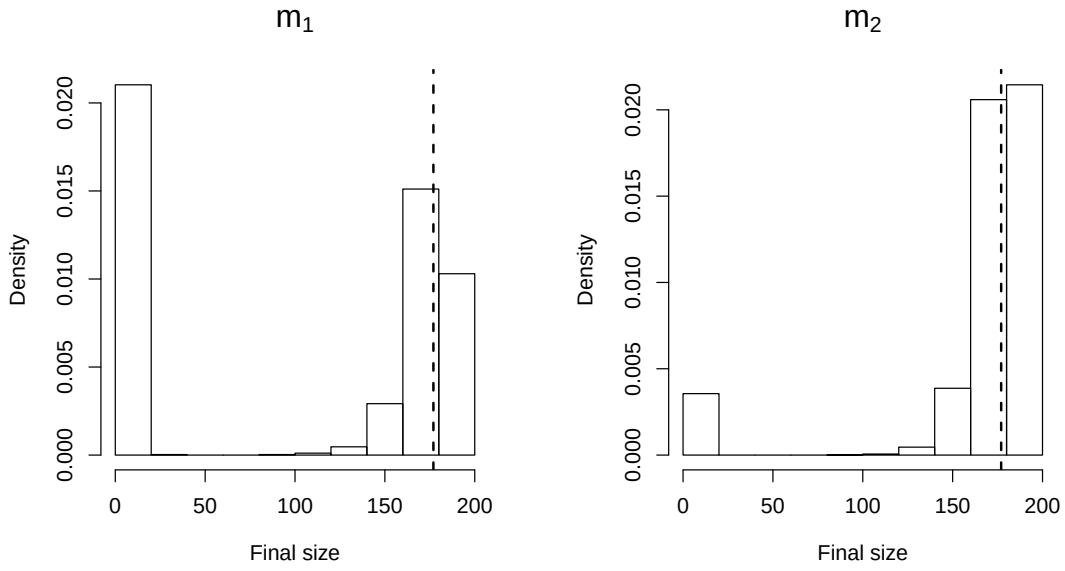


Figure 3.6: The final size posterior predictive distributions for the two models m_1 and m_2 (left to right), where the dashed vertical line indicates the observed final size ($n_R - 1 = 177$).

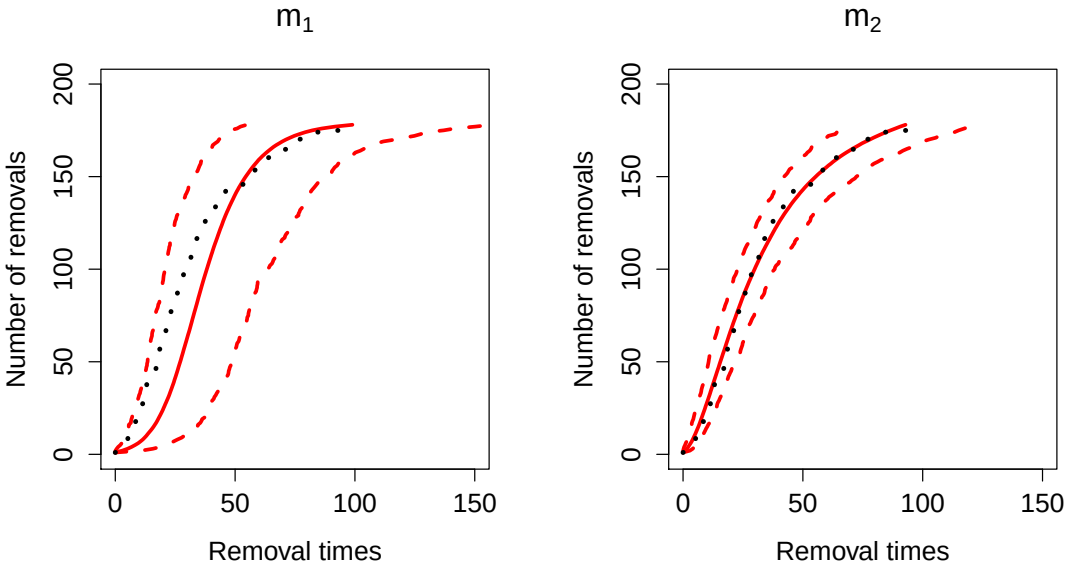


Figure 3.7: Predictive removal curves mean (red solid line) with 95% percentile intervals (red dashed lines) and the observed removal curve (black dotted line) for the two SIR models m_1 and m_2 (left to right) calculated based on 500 predicted removal curves generated from each model under the condition of the same epidemic final size.

- **Computation of the Bayesian residual:**

Bayesian residual conditioning on $n_R^{rep\,(k)} = n_R,\ k = 1, ..., M'$, can be defined as follows.

$$d_j = R_j^{obs} - E(R_j^{rep}|\boldsymbol{R}^{obs}),\ j = 1, ..., n_R, \tag{3.3.7}$$

where the mean predicted time of the jth removal can be approximated as

$$E(R_j^{rep}|\boldsymbol{R}^{obs}) \approx \frac{1}{M'}\sum_{k=1}^{M'} R_j^{rep\,(k)}. \tag{3.3.8}$$

We calculated (3.3.7) using (3.3.8) for each SIR model as explained in Section 3.2.4, then we plotted their posterior density function. Figure 3.8 shows the Bayesian residual densities for the two SIR models in which it is qualitatively evident that there is a high density accumulated around zero, coming from the modified SIR model (m_2), compared to the standard SIR model (m_2). On the top of that, quantitatively, the sum of the squared residuals $SSR = \sum_{j=1}^{n_R} d_j^2$ are 565.14 and 11625.12 for m_2 and m_1 models respectively. Therefore, m_2, the modified SIR model, from which the data have been simulated, has the smallest residual value. Thus, using a Bayesian residual distribution as a diagnostic tool to check the fit of a model, indicates that our proposed two-stage method is seen to identify the true model reasonably well.

- **Computation of the posterior predictive $p-$value:**

As described in Section 3.2.3, another possible idea for distinguishing between the two SIR models is to calculate the posterior predictive $p-$value under fitting the two SIR models to the observed removal data. The posterior predictive $p-$value can be estimated as

$$\begin{aligned} ppp-\text{value} &= P\left(D(\boldsymbol{R}^{rep}, \boldsymbol{\phi}) \geq D(\boldsymbol{R}^{obs}, \boldsymbol{\phi})\big|\boldsymbol{R}^{obs}\right) \\ &\approx \frac{1}{M'}\sum_{k=1}^{M'} \mathbb{1}_{\{D(\boldsymbol{R}^{rep\,(k)}, \boldsymbol{\phi}^{(k)}) \geq D(\boldsymbol{R}^{obs}, \boldsymbol{\phi}^{(k)})|\boldsymbol{R}^{obs}\}}, \end{aligned} \tag{3.3.9}$$

where here for the discrepancy measure $D(\boldsymbol{R}, \boldsymbol{\phi})$ we used both the χ^2 and Freeman-Tukey discrepancy measures.

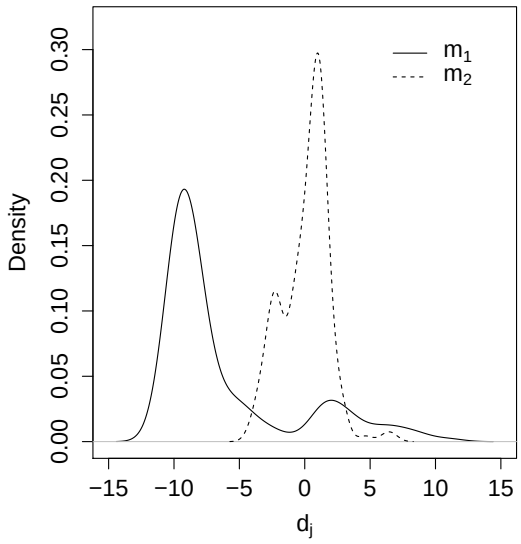


Figure 3.8: The Bayesian residual densities for the two SIR models (m_1 and m_2) resulting from calculating (3.3.7) for each SIR model based on the average of 500 removal curves from each model.

In a Bayesian context, if the model in question offers a satisfactory fit to the observed data, then the posterior predictive $p-$value will tend to have a distribution more concentrated near 0.5 (Meng, 1994). As mentioned earlier, the quantities $E(R_j|\boldsymbol{\phi}^{(k)})$ and $\text{var}(R_j|\boldsymbol{\phi}^{(k)})$ are not available in closed forms. Therefore, Monte Carlo simulation will be used to estimate them provided that samples $\boldsymbol{R}^{(k)}$ can be generated from each model, where $\boldsymbol{\phi}$ denotes a vector of each model parameters, that is $\boldsymbol{\phi} = (\beta, \gamma)$ for model m_1 or $\boldsymbol{\phi} = (\beta, \gamma, p)$ for model m_2 and $k = 1, ..., M'$.

The calculation of the posterior predictive $p-$value using the χ^2 and Freeman-Tukey discrepancy measures can be performed as described in Section 3.2.4 for each SIR model. Here, we chose $M' = 500$ and $D = 100$ (the number of epidemic realisations that have exactly n_R removals) and calculated the $ppp-$value using both the χ^2 and Freeman-Tukey discrepancy measures for both models m_1 and m_2. Figure 3.14 shows the resultant scatter plots with the posterior predictive $p-$values. As we mentioned earlier, the closer the ppp-value to the optimal value 0.5 the better the fit (Gelman et al., 2013). The

ppp-values obtained when fitting m_2 (the correct model) to the observed data are very close to the optimal value (0.486 and 0.54 obtained from the χ^2 and Freeman-Tukey discrepancy measures respectively), whereas the ppp-values obtained by fitting m_1 can be seen as extreme values indicating a poor fit of model m_1. This result is in agreement with our findings earlier when the Bayesian residual considered as a tool of diagnostic.

To be able to see whether or not the proposed method can capture the variability coming from different datasets, we carried out another simulation study but this time we simulated our removal data from the standard SIR model, m_1.

3.3.3.2 Sub-study II: SIR model with linear infection rate data

For a population of $\boldsymbol{N} = 200$ individuals, we set the contact and removal rates to be $\beta = 0.6$ and $\gamma = 0.2$ respectively. There were 187 observed removal times. Then, we fitted both SIR models, m_1 and m_2, to these data using MCMC methods as explained earlier. Having obtained the desired sample size from the posterior distributions of each model parameters (after a burn-in of 2000 samples and thinning the chain by taking each $5th$ value), we applied our procedure to see how good this method is in recovering the true model that generated these removal data. Figure 3.10 represents 5000 removal curves drawn from the posterior predictive distribution of the removal times for the two SIR models.

In addition, Figure 3.11 shows that the observed final size fits within the final size predictive distributions for both models which in turn enables us to apply the second stage of our procedure. Also, Figure 3.12 shows the average of 500 removal trajectories along with 95% percentile intervals generated from the removal times predictive distributions for the two SIR models where the final sizes are matched with the observed one ($n_R - 1 = 186$). More formally, using the same way of that in previous simulation study in Section 3.3.3.1, we calculated the Bayesian residuals for each model using (3.3.7) and then we

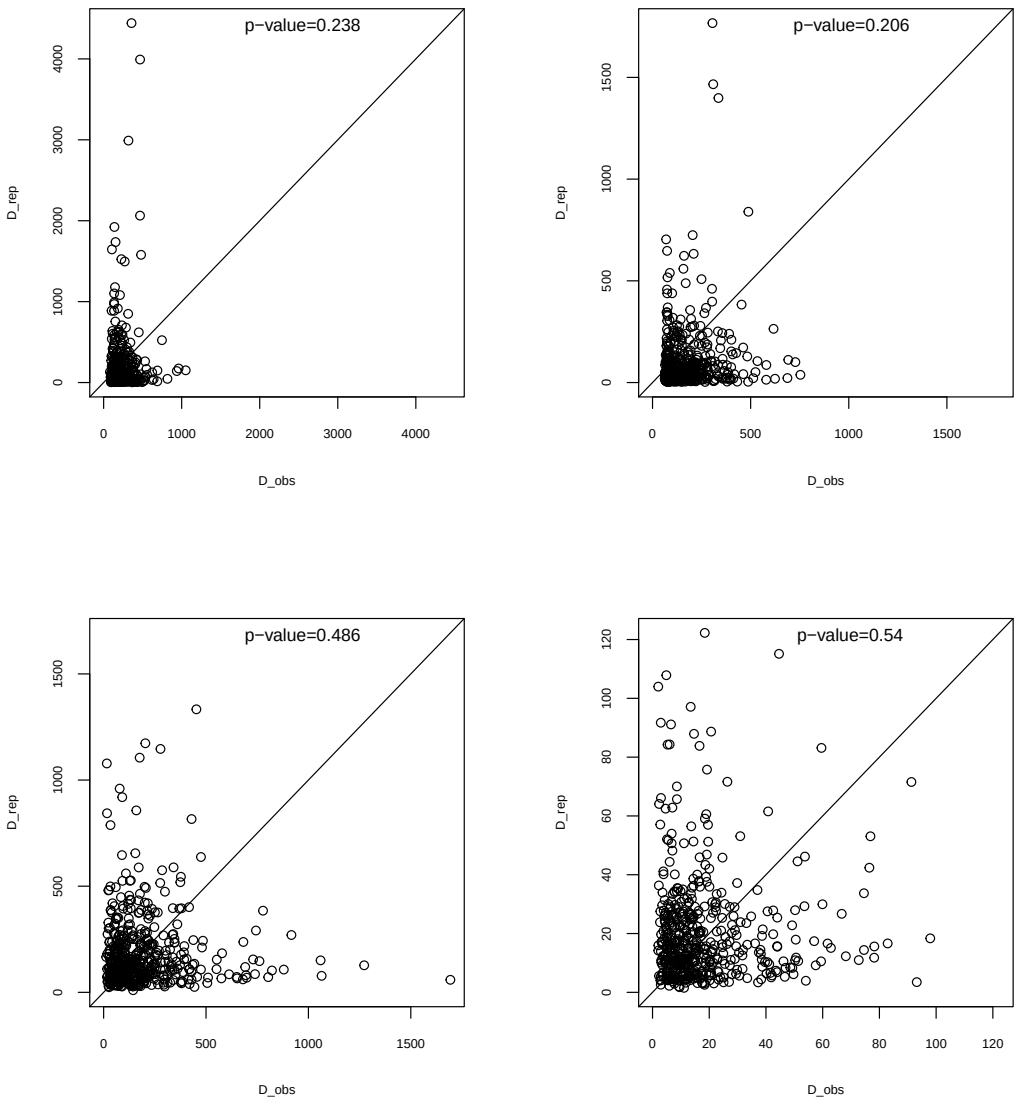


Figure 3.9: Scatter plots of observed versus replicated discrepancies based on 500 predicted removal curves from each removal times predictive distributions of the two SIR models (top and bottom rows refer to m_1 and m_2 respectively). The ppp-values are estimated by the proportion of points above the 45° line. Left and right columns correspond to the χ^2 and Freeman-Tukey discrepancy measures.

plotted their posterior densities as shown in Figure 3.13 in which the posterior density resulting from applying the method on the standard SIR model (m_1) has accumulated values around zero as an indication of a good fit with a smaller $SSR = \sum_{j=1}^{n_R} d_j^2$ of 184.61 compared to 2042.11 for the SSR value produced

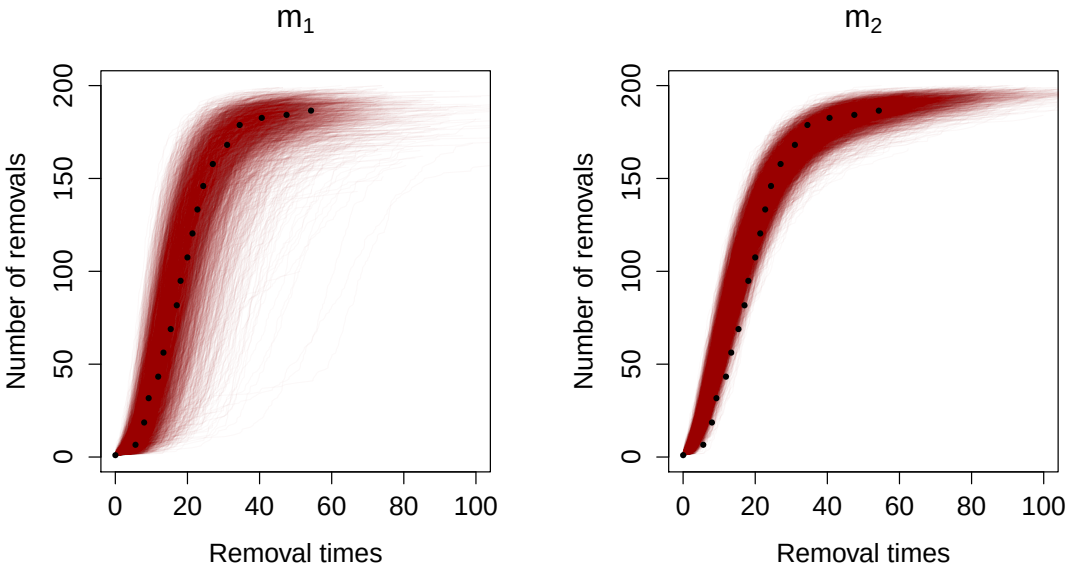


Figure 3.10: The posterior predictive distribution of the removal times for the two SIR models (left and right plots refer to m_1 and m_2 models respectively) based on 5000 samples from the joint posterior density of each model parameters, where dotted line indicates the observed data.

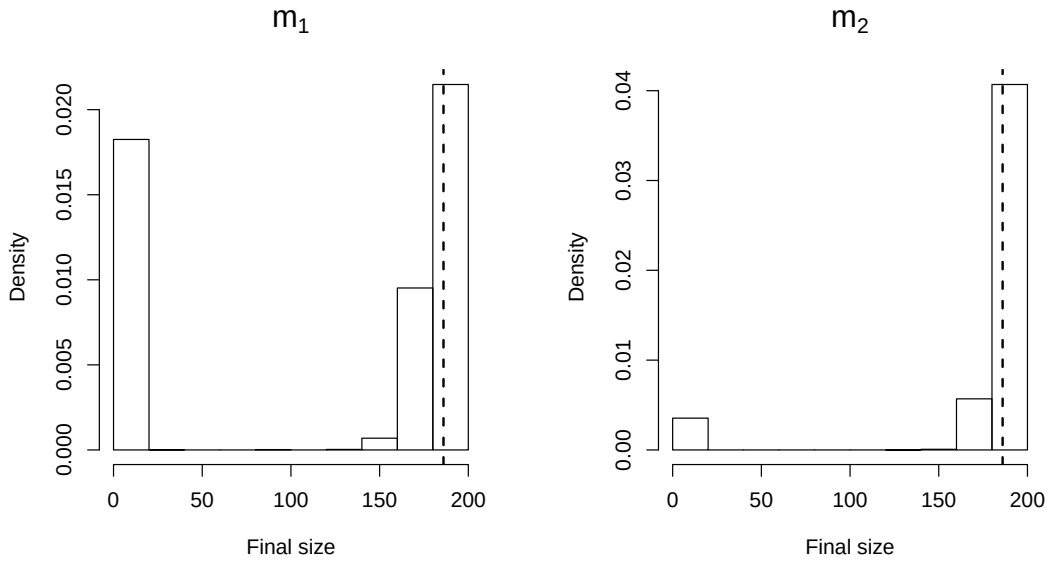


Figure 3.11: The final size posterior predictive distributions for the two models m_1 and m_2 (left to right), where the dashed vertical line indicates the observed final size ($n_R - 1 = 186$).

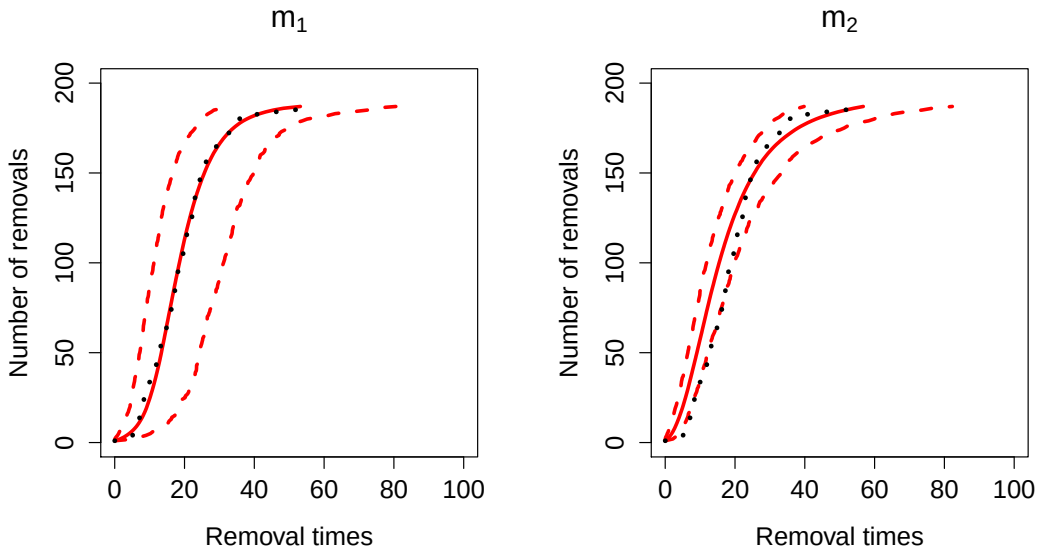


Figure 3.12: Predictive removal curves mean (red solid line) with 95% percentile intervals (red dashed lines) and the observed removal curve (black dotted line) for the two SIR models m_1 and m_2 (left to right) calculated based on 500 predicted removal curves generated from each model under the condition of the same epidemic final size.

from fitting the modified SIR model (m_2).

Calculating the posterior predictive $p-$value under fitting the two SIR models to the observed removal curve was performed as in the previous simulation study in Section 3.3.3.1 with same discrepancy measures. Figure 3.14 reveals scatter plots of observed versus replicated discrepancies of the χ^2 and Freeman-Tukey measures produced based on 500 samples from the removal times predictive distributions of the two SIR models along with the resulting ppp-values. Here, the standard SIR model (m_1) is preferred over the modified SIR model (m_2) as the obtained ppp-values from fitting the latter can be seen as extreme values (0.018 and 0.032 from the χ^2 and Freeman-Tukey discrepancy measures respectively).

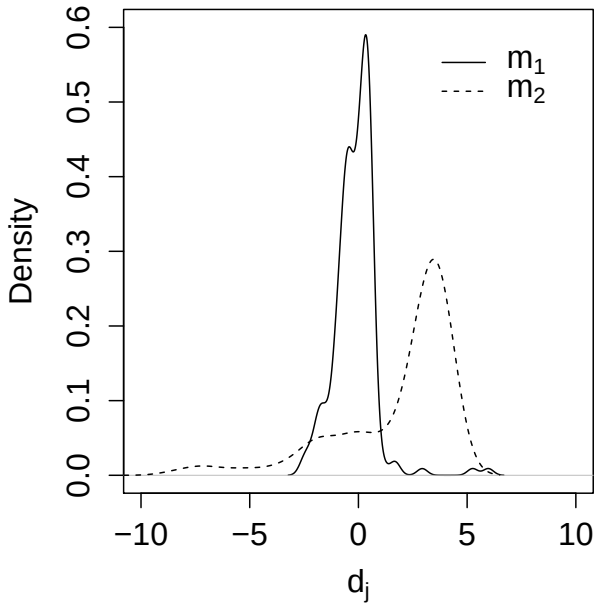


Figure 3.13: The posterior residual densities for the two SIR models resulting from calculating (3.3.7) for each SIR model based on 500 samples from the conditioning predictive distribution of the removal times for the two models.

3.3.3.3 Sensitivity of the method to R_0

Another simulation study was carried out to evaluate the performance of the proposed two-stage method when the epidemic size varies. We ran a simulation study to see how much data we might have to collect in order to correctly identify the true model. We considered a population of size $\boldsymbol{N} = 100$ and then simulated three datasets from the standard SIR model (m_1) using true value of $\gamma = 0.2$ for the removal rate parameter. The infection rate parameter β was given different values in order to have different epidemic final sizes. The ratios of the final size to $\boldsymbol{N}$ were $23/100, 66/100$ and $93/100$ for datasets A, B and C, respectively. These datasets are typical in the sense that the final size lies close to the modal value of the major outbreak part of the final size distribution for datasets B and C, whereas it is within the minor epidemic component of the final size distribution for dataset A. We fitted both models

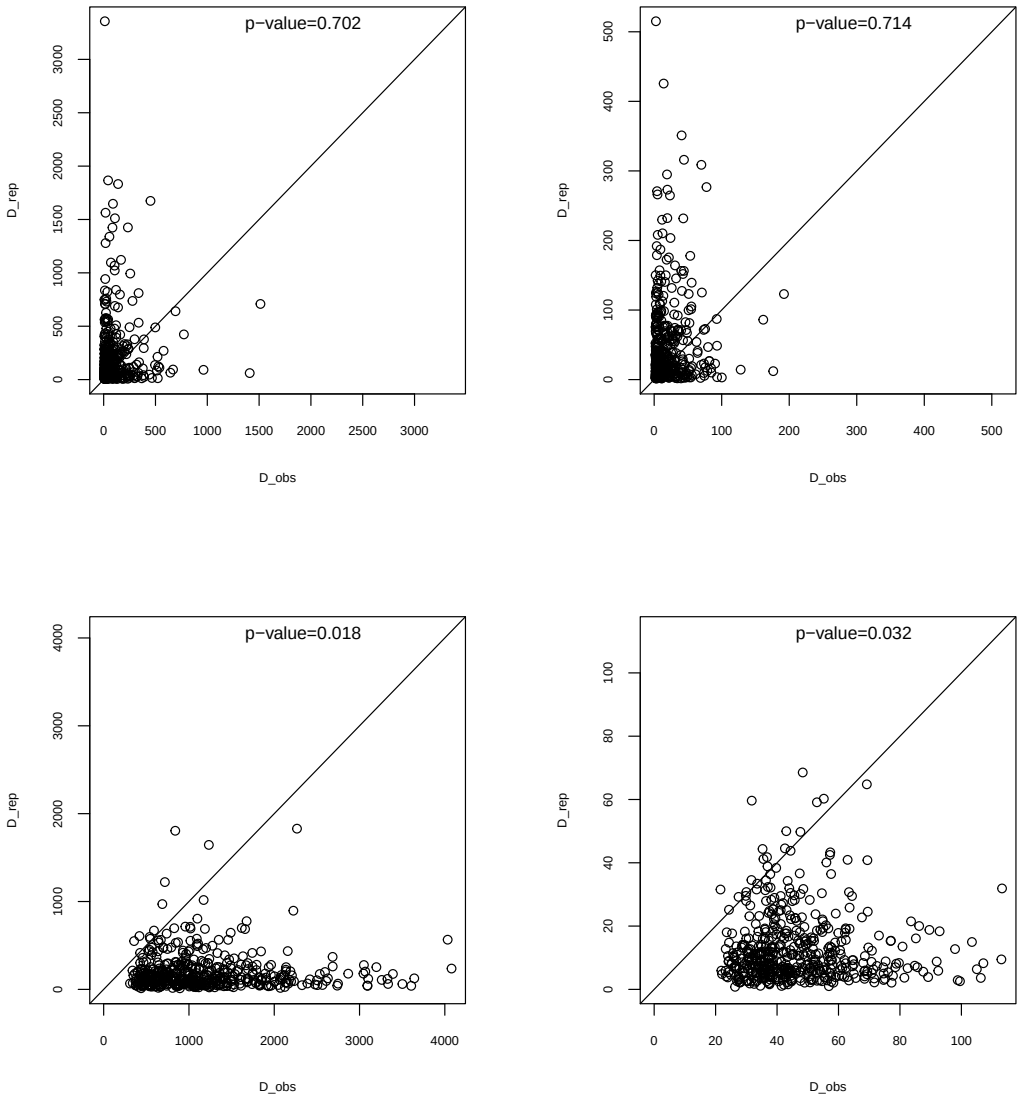


Figure 3.14: Four scatter plots of observed versus replicated discrepancies based on 500 predicted removal curves from each removal times predictive distributions of the two SIR models. The ppp-values are estimated by the proportion of points above the 45° line. Left and right columns correspond to the χ^2 and Freeman-Tukey discrepancy measures. Top and bottom rows refer to m_1 and m_2 respectively.

m_1 (the generated-data model) and m_2 to each dataset using MCMC methods and then applied our two-stage method. Details of the simulation and some corresponding results are given in Table 3.3. The removal times pos-

Dataset	$\boldsymbol{F/N}$	β	R_0	SSR		ppp-value (χ^2)		ppp-value (F-T)	
				m_1	m_2	m_1	m_2	m_1	m_2
A	23/100	0.23	1.15	**100.90**	145.09	**0.53**	0.394	**0.582**	0.46
B	66/100	0.275	1.375	**556.84**	1756.49	**0.492**	0.11	**0.56**	0.17
C	93/100	0.6	3	**201.53**	759.73	**0.628**	0.066	**0.646**	0.106

Table 3.3: Values of the infection rate parameter β and the basic reproduction number R_0 used to simulate (from model m_1) the three datasets A, B and C along with the ratio of the final size to the population size ($\boldsymbol{F/N}$). Results of sum of squared residuals (SSR) and the posterior predictive $p-$values obtained from fitting both SIR models, m_1 and m_2, to each dataset using both the χ^2 and Freeman-Tukey (F-T) discrepancy measures.

terior predictive distributions produced using these datasets are presented in Figure 3.15 and the final size posterior predictive distributions are shown in Figure 3.16. By looking at Figure 3.17 and the values of SSR and ppp-values in Table 3.3, it seems that our two-stage method is not too sensitive to the value of R_0 and it appears to work well in identifying the correct model. For dataset A, based on SSR values, m_1 fits the data better than m_2, however there is no indication of a poor fit of both models based on the ppp-values which is not surprising as we have much less data compared to B and C datasets. Also, this observation along with the other ppp-values in Table 3.3 might indicate the consistency and reliability of the posterior predictive $p-$value as a measure-of-goodness of fit for these epidemic models.

In terms of calculation of the posterior mean of the jth removal $E(R_j|\boldsymbol{\phi}^{(k)})$ used to obtain the posterior predictive $p-$values, we found that, in particular for dataset A, it is computationally expensive to generate removal epidemic realisations, that have similar epidemic size to the observed one, from model m_1 using some posterior pairs of β and γ which result in extreme values of R_0.

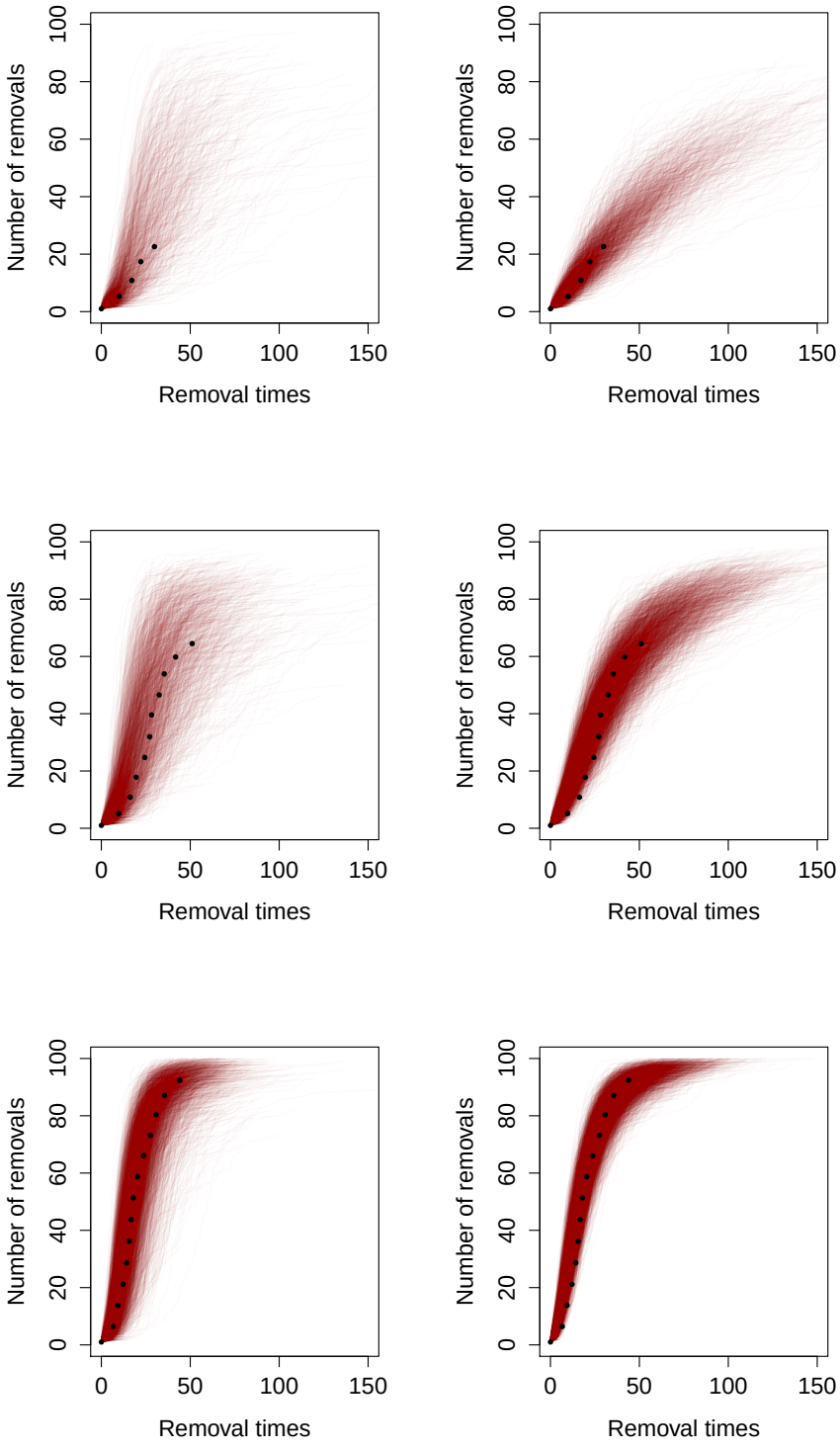


Figure 3.15: The posterior predictive distribution of the removal times where the dotted line indicates the observed data. Left and right plots refer to m_1 and m_2 models respectively. Top, middle and bottom rows refer to datasets A, B and C respectively.

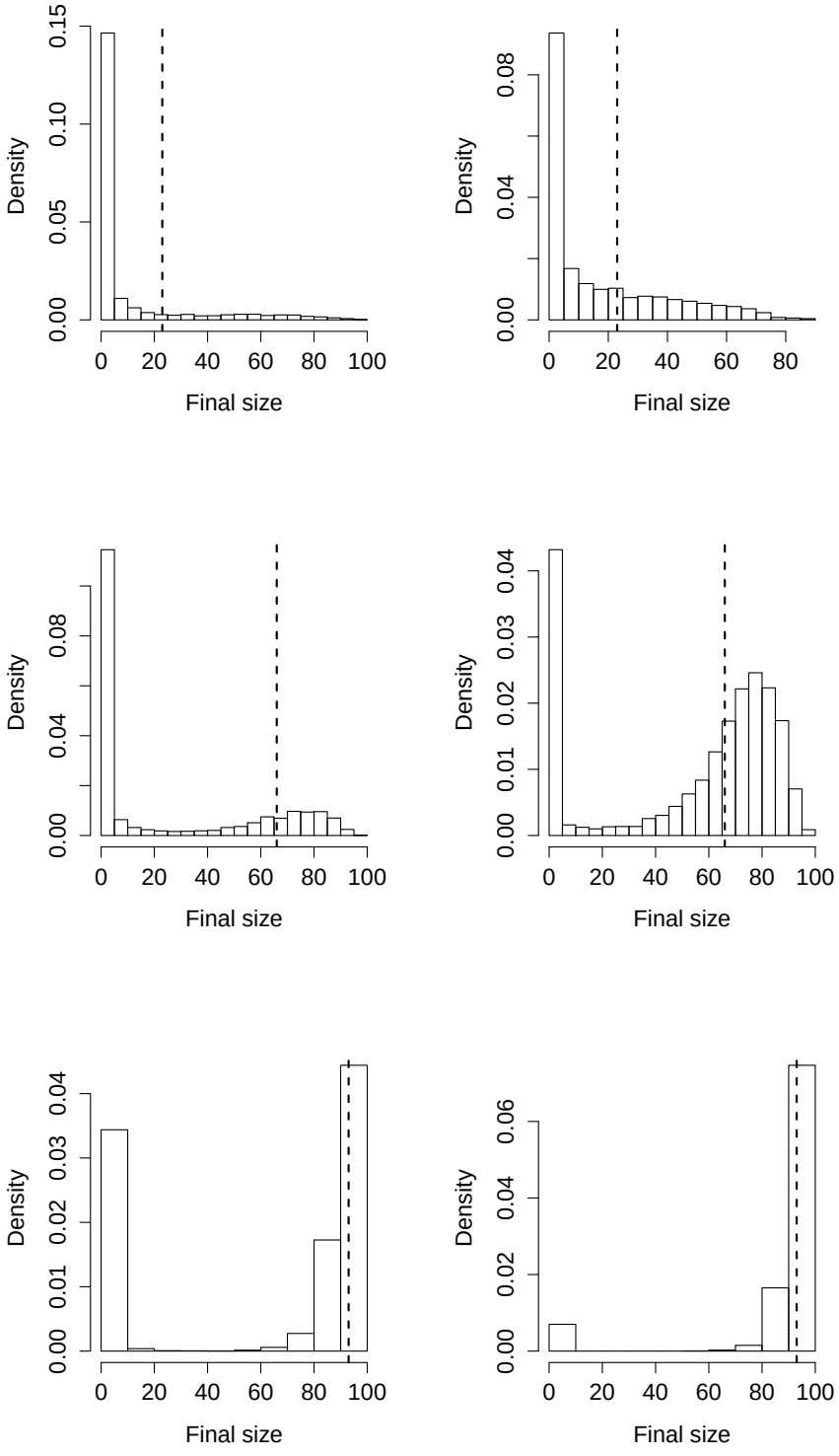


Figure 3.16: The final size posterior predictive distributions for models m_1 and m_2 (left to right), where the dashed vertical line indicates the observed final size. Top, middle and bottom rows refer to datasets A, B and C respectively.

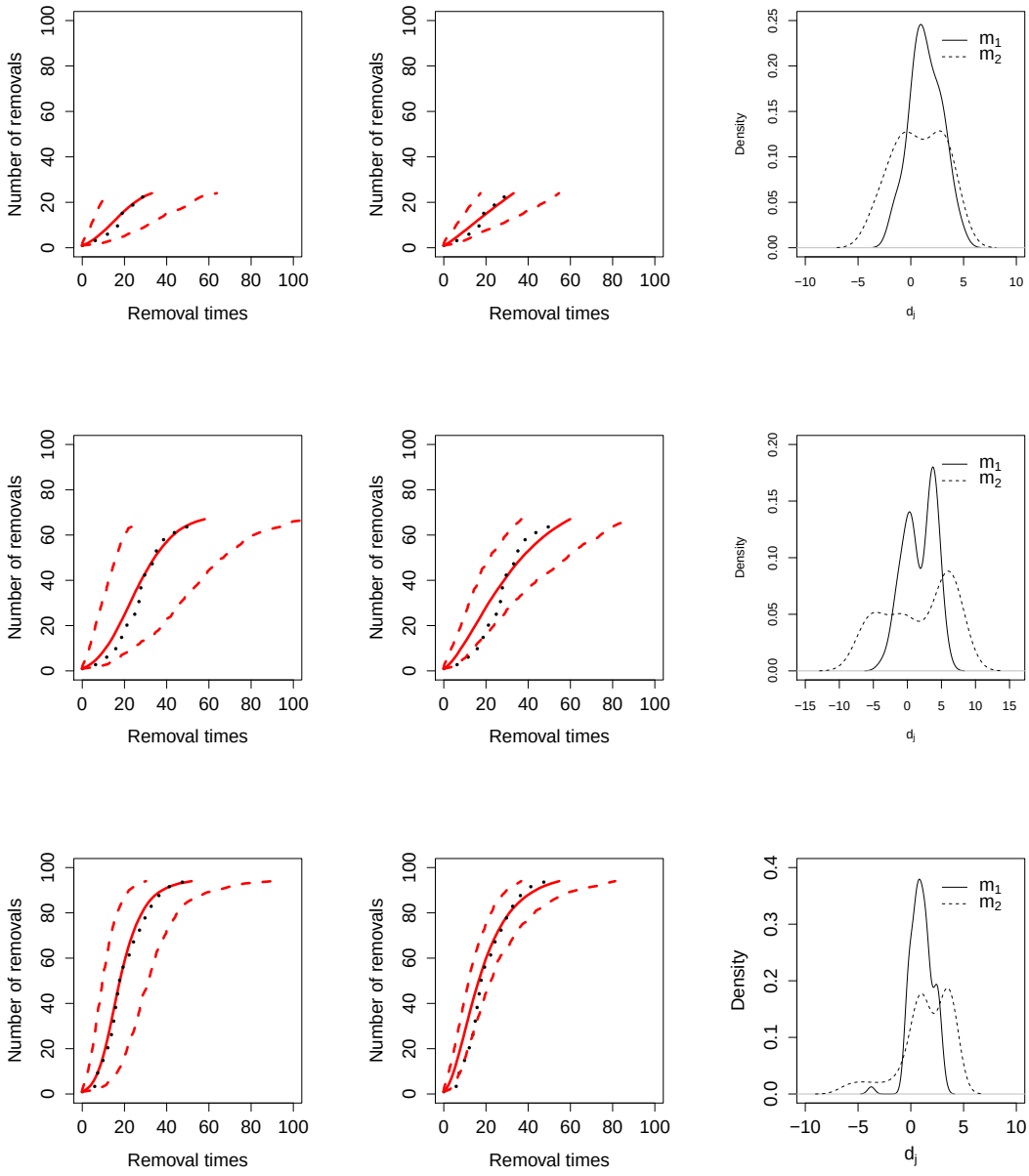


Figure 3.17: Top, middle and bottom rows refer to datasets A, B and C respectively. Left and middle columns represent the mean of 500 predicted removal curves with 95% percentile intervals and the observed removal curve (black dotted line) for models m_1 and m_2 respectively. Right column shows the posterior residual densities for the two SIR models.

3.4 Model assessment for SIR models with different infectious periods

In this section we are concerned with identifying the infectious period distribution of the SIR models, based only on removal data, by considering the cumulative number of removals curve. We specifically consider three SIR models routinely employed in the literature, namely the model with exponential infectious periods ($T_I \sim \text{Exp}(\gamma)$) denoted by m_1, the model with gamma infectious periods ($T_I \sim \text{Gamma}(\alpha, \delta)$) denoted by m_2, where the shape parameter of gamma distribution α is assumed to be known, and the model with fixed length infectious periods ($T_I \equiv c$) denoted by m_3, where $c > 0$ is a constant. Bayesian parameter estimation for all three models was described in Chapter 2. Therefore, once we obtain the required posterior samples from the joint posterior distribution of the parameters for each model, say $\boldsymbol{\phi_1} = (\beta, \gamma)$ for model m_1, $\boldsymbol{\phi_2} = (\beta, \delta)$ for model m_2 and $\boldsymbol{\phi_3} = (\beta, c)$ for model m_3, we can use them to sample from each model of interest to generate samples from the posterior predictive distribution of removal times.

First, we shall look at the behaviour of removal trajectories for the three aforementioned stochastic SIR models via simulation.

3.4.1 The behaviour of removal curves for the three models

In the SIR stochastic epidemic model, regardless of the type of the infectious period distribution (exponential, gamma or constant), the total population size is constant (due to the assumption of a closed population) and satisfies $\boldsymbol{N} = X(t) + Y(t) + Z(t)$, where $X(t)$, $Y(t)$ and $Z(t)$ denote the number of susceptible, infected and removed individuals at event time t with $X(0) \geq 1, Y(0) \geq 1$ and $Z(0) = 0$. It can be noted that $Z(s) \leq Z(t)$ for any $0 \leq s \leq t;\ s, t \geq 0$.

Assuming that all infected individuals are removed by the end of the epi-

demic, it can be shown via simulation that the behaviour of the three models in terms of the removal curves differs. Specifically, if we set the infection rate parameter β to be the same in all three models and keep the mean infectious period $E(T_I)$ equal and then plot $Z(R_j)$ against R_j, where R_j denotes the jth removal time, then the distribution of the removal curves for three SIR models will look different. Figure 3.18 shows 5000 removal curve realisations simulated from the three SIR models with the same $E(T_I) = 1$ (for the gamma SIR model, $T_I \sim \text{Gamma}(\alpha = 10, \delta = 10)$) and the same infection rate $\beta = 3$ in a closed population of size $(n, a) = (199, 1)$. It is evident that the duration of epidemics with the exponential infectious period (m_1) tends to be longer than others (Top left plot). Moreover, although it looks that models m_2 and m_3 have similar removal curve distribution, epidemic removal curves in which the infectious period is constant (bottom plot) take off quickly compared to the other two models.

To clarify the difference and make it easier to see, we performed another simulation study as detailed in Table 3.4.

Simulation	(n, a)	β	$E(T_I)$	Final size
I	$(29, 1)$	0.25	10	39
II	$(99, 1)$	1.5	2	94

Table 3.4: Summary of the simulation study used to obtained Figure 3.19.

In this simulation study, we looked at the behaviour of the three SIR models by averaging 1000 realisations of removal curve simulated from each model conditioning on a particular final size. The results are shown in Figure 3.19, where it is clear that the averages are not all the same in the three SIR models.

Having highlighted this difference, we turn our attention to taking advantage of it to distinguish between these three SIR models when fitting them to epidemic removal data using our proposed two-stage method introduced in Section 3.2.4.

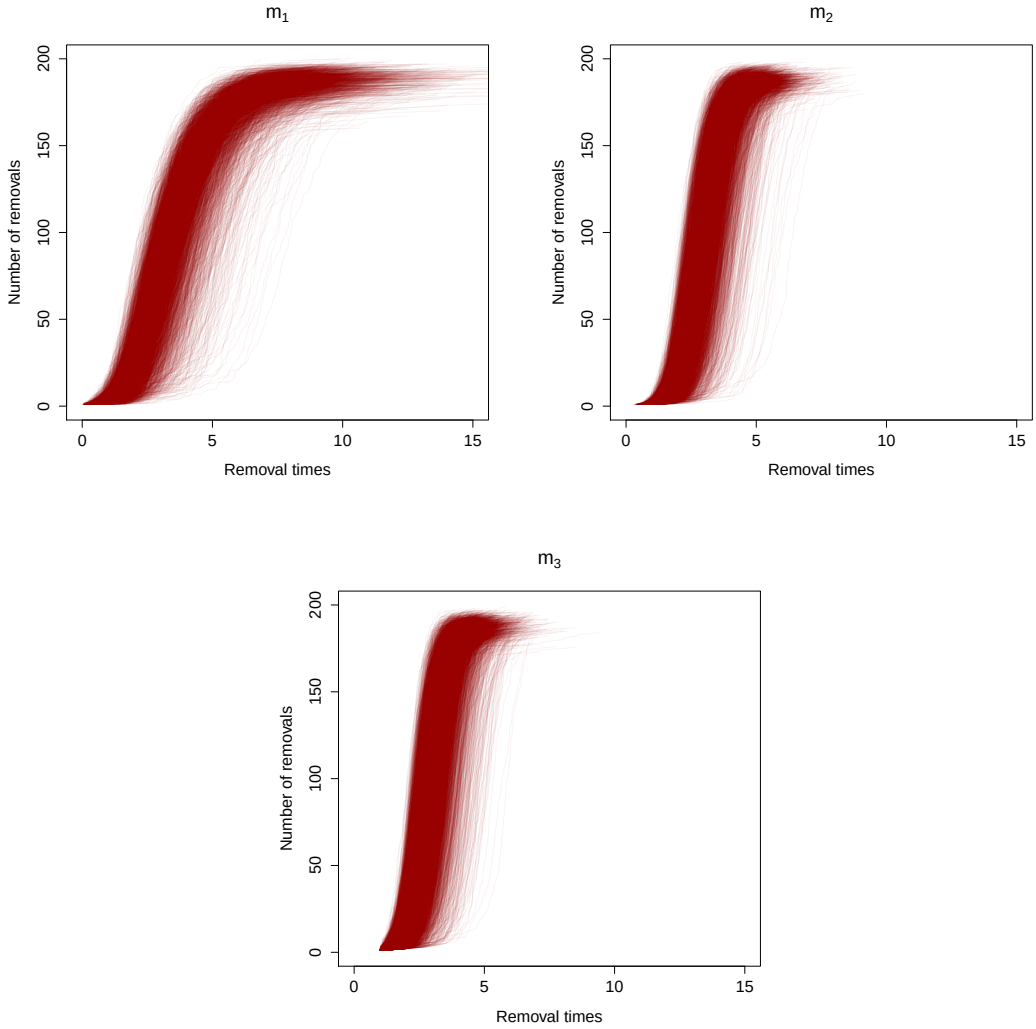


Figure 3.18: 5000 removal curve realisations simulated from the three epidemic models in a closed population of size $N = 200$ using the same mean infectious period and the same infection rate, namely $E(T_I) = 1$ and $\beta = 3$.

In what follows, we test our proposed two-stage method using simulated and real outbreak data where the alternative labelled representation to derive the likelihood function (see Chapter 2) was used to fit the three SIR models to the removal data using MCMC methods. Throughout the analysis, uninformative gamma prior distributions with parameters $\lambda_\beta = \lambda_\gamma = \lambda_\delta = 1$ and $\nu_\beta = \nu_\gamma = \nu_\delta = 0.001$ are assigned to the parameters of the SIR models and it is assumed that the gamma shape parameter, when fitting an SIR model with gamma infectious period ($T_I \sim \text{Gamma}(\alpha, \delta)$) is known ($\alpha = 10$). Also, we assume

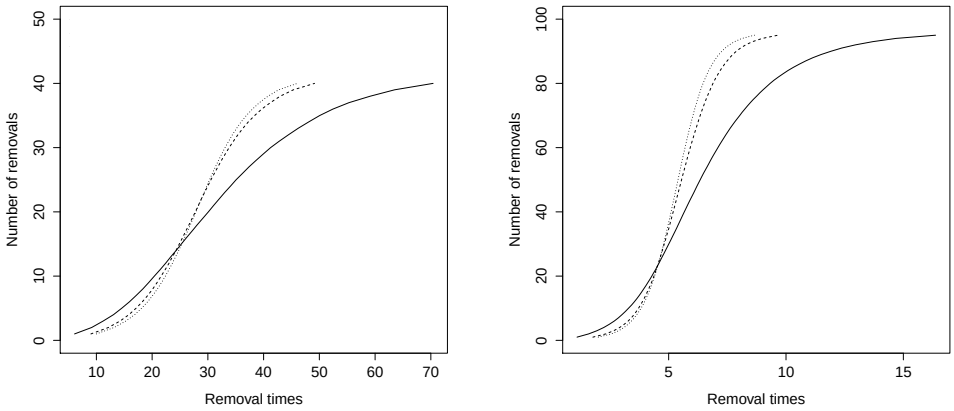


Figure 3.19: The averages of 1000 realisations of removal curves obtained from each SIR model conditioning on particular final sizes, 39 and 94 (left to right). Solid, dashed and dotted lines represent the averages obtained from models m_1, m_2 and m_3 respectively. Left and right plots correspond to simulations ***I*** and ***II***.

a priori that z (the label of the initial infective) is equally likely to be any of the n_R infected individuals and that I_z has an improper uniform prior density on $(-\infty, R_1)$. For the SIR model with constant infectious period ($T_I \equiv c$), we assume that a priori c has an improper uniform prior density on $(0, \infty)$.

3.4.2 Application of the two-stage method to simulated outbreak data and results

We simulated removal data according to model m_1 and m_2, fitted all three models and then aimed to identify the correct model using our two-stage procedure. Under each scenario, we simulated and analysed several outbreak datasets. We present one typical simulated dataset for each scenario.

3.4.2.1 Sub-study I: Exponential SIR model data

Some (removal times) data were simulated from an exponential stochastic epidemic model (SIR model in which $T_I \sim \text{Exp}(\gamma)$) with parameters $\beta = 2.5$ and

$\gamma = 1$ in a population of size $(n, a) = (99, 1)$. The resulting data consisted of 90 observed removal times. Then, we fitted all the three SIR models to these data using MCMC with the alternative labelled representation approach introduced in Chapter 2. After a burn-in of 2000 samples and thinning the chain by taking each $5th$ value, 5000 samples were obtained from the joint posterior distribution of the parameters fr each model, namely (β, γ) for model m_1, (β, δ) for model m_2 and (β, c) for model m_3.

Having obtained these samples, we are in a position to use them to simulate 5000 removal trajectories from the three models (see Figure 3.20) in order to see whether or not the observed final size ($n_R - 1 = 89$) lies within the final size posterior predictive distribution for the three models (see Figure 3.21). Ensuring that the observed final size lies within the final size posterior predictive distribution enables us to condition on it and proceed by applying the two Bayesian model assessment criteria introduced earlier to assess the goodness-of-fit. Figure 3.22 shows the mean of 500 removal trajectories along with 95% percentile intervals generated from the removal times predictive distributions for the three SIR models where the final sizes are matched with the observed one. It is graphically evident that the observed data can be seen as a realisation from the exponential SIR model (m_1), the true model that has generated the data.

- **Computation of the Bayesian residual and the posterior predictive $p-$value:**

As mentioned above, our preferred criteria to measure the goodness-of-fit are the Bayesian residual defined as

$$\begin{aligned} d_j &= R_j^{obs} - E(R_j^{rep} | \boldsymbol{R}^{obs}) \\ &\approx R_j^{obs} - \frac{1}{M'} \sum_{k=1}^{M'} R_j^{rep\ (k)}, \end{aligned} \tag{3.4.1}$$

where $k = 1, ..., M'$, and the posterior predictive $p-$value estimated as

$$\begin{aligned} ppp - \text{value} &= P\left(D(\boldsymbol{R}^{rep}, \boldsymbol{\phi}) \geq D(\boldsymbol{R}^{obs}, \boldsymbol{\phi}) \big| \boldsymbol{R}^{obs} \right) \\ &\approx \frac{1}{M'} \sum_{k=1}^{M'} \mathbb{1}_{\{D(\boldsymbol{R}^{rep\ (k)}, \boldsymbol{\phi}^{(k)}) \geq D(\boldsymbol{R}^{obs}, \boldsymbol{\phi}^{(k)}) | \boldsymbol{R}^{obs}\}}, \end{aligned} \tag{3.4.2}$$

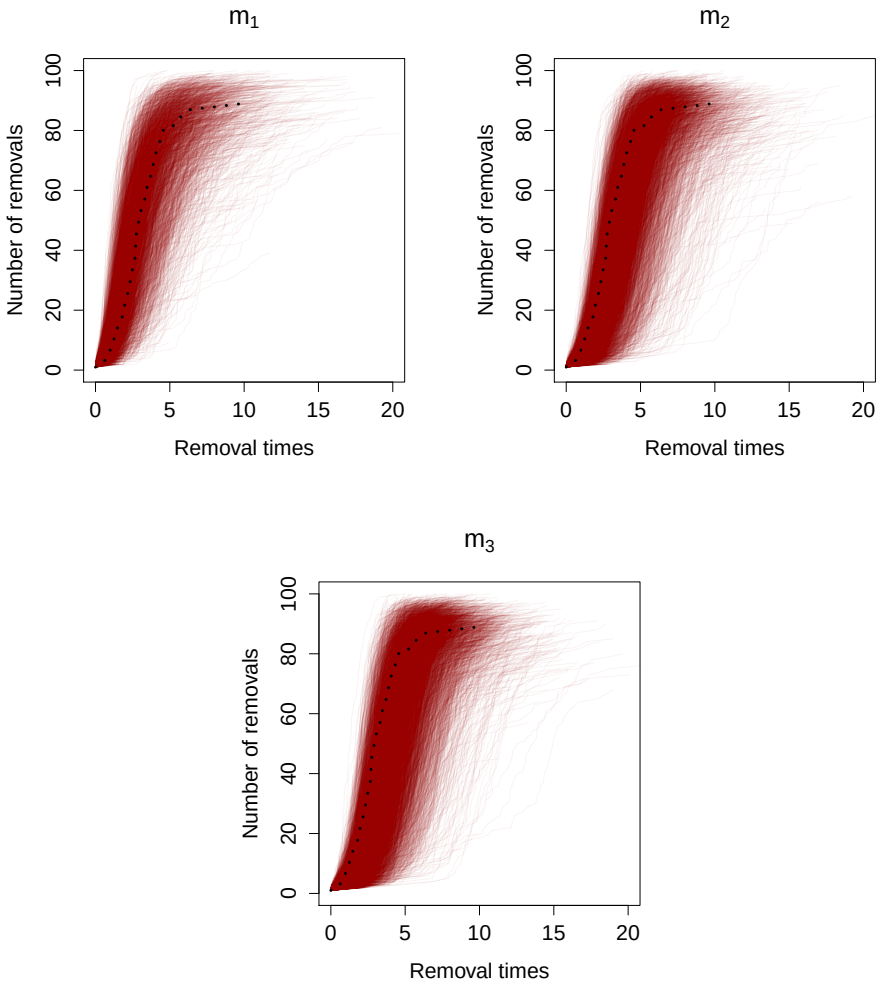


Figure 3.20: The posterior predictive distribution of the removal times for the three SIR models produced based on 5000 samples from the joint posterior density of each model parameters, where the dotted line indicates the observed data.

where here for the discrepancy measure $D(\boldsymbol{R}, \boldsymbol{\phi})$ we again used both the χ^2 and Freeman-Tukey discrepancy measures to obtain the $ppp-$values.

Generally, calculation of d_j and the $ppp-$values was performed as described in Section 3.2.4. We calculated (3.4.1) for each SIR model, then we plotted the corresponding density function. Figure 3.23 shows the Bayesian residual densities for the three SIR models from which it is qualitatively obvious that there is a high density accumulated around zero, coming from the exponential SIR model (m_1), compared to the other two SIR models. On the top

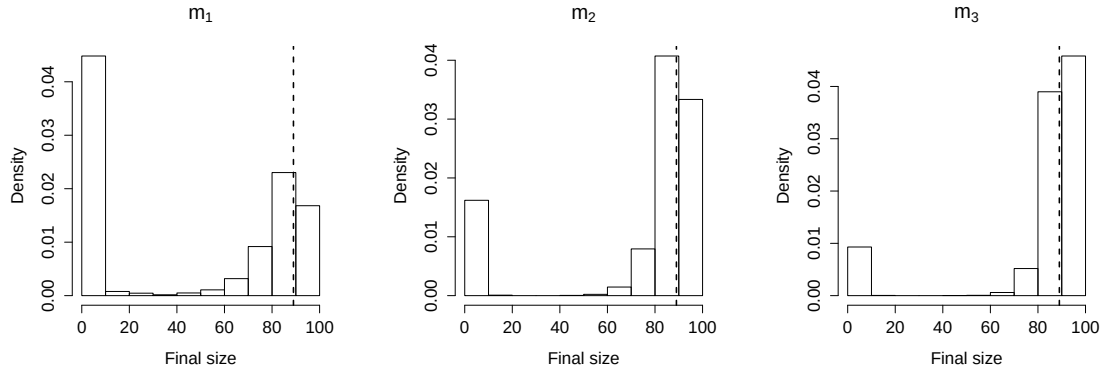


Figure 3.21: The final size posterior predictive distribution for the three SIR models where dotted line indicates the observed value of the final size ($n_R - 1 = 89$).

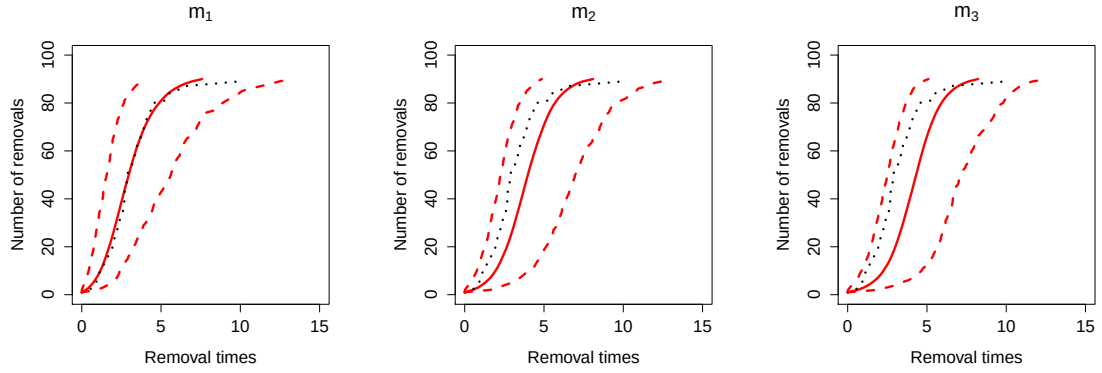


Figure 3.22: Predictive removal curves mean (red solid line) with 95% percentile intervals (red dashed lines) and the observed removal curve (black dotted line) for the three SIR models calculated based on 500 predicted removal curves generated from each model under the condition of the same epidemic final size.

of that, quantitatively, the sum of the squared Bayesian residuals $\sum_{j=1}^{n_R} d_j^2$ are 24.74, 71.84 and 125.64 for the exponential, gamma and constant SIR models respectively. Therefore, as expected, model m_1, from which the data have been generated, has the smallest value and therefore offers the best fit.

The posterior predictive $p-$values were computed via Monte Carlo simulations. Using $D = 100$ and $M' = 500$ posterior samples from each joint SIR model parameter posterior distribution and conditioning on the observed final size, we obtained $\boldsymbol{R}^{rep\ (k)}, k = 1, ..., M' = 500$ predicted curves, $E(R_j|\boldsymbol{\phi}^{(k)})$

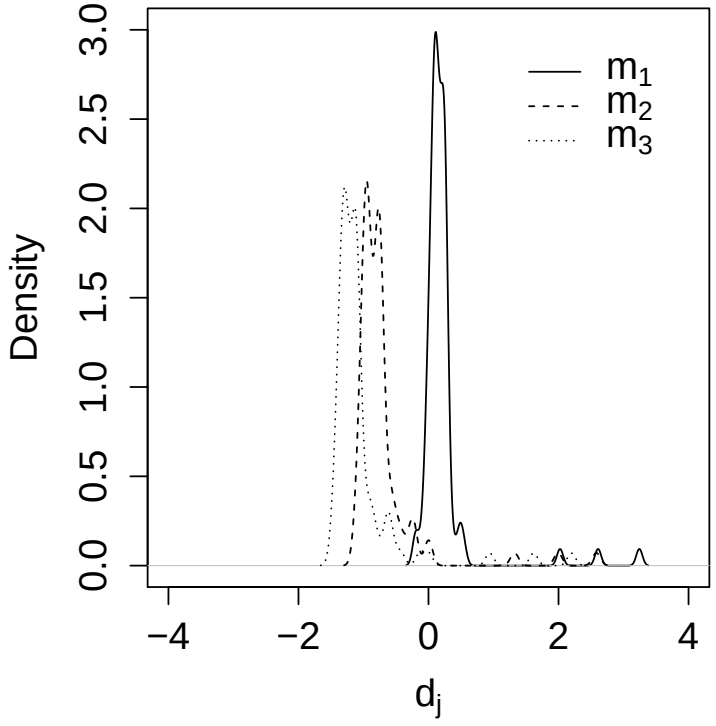


Figure 3.23: Bayesian residual densities for the three SIR models resulting from calculating (3.4.1) for each SIR model based on 500 posterior samples from the conditioning predictive distribution of removal times for the three models.

and var$(R_j|\boldsymbol{\phi}^{(k)})$, then the ppp-values were estimated using the χ^2 and Freeman-Tukey discrepancy measures. Figure 3.24 shows the resultant scatter plots along with the posterior predictive $p-$values. As we mentioned earlier, the closer the ppp-value to the optimal value 0.5 the better the fit. The ppp-values obtained when fitting the exponential SIR model (m_1) to the observed data are the closest to the optimal value, whereas the other ppp-values can be regraded as extreme values indicating a poor fit of both m_2 and m_3 models. This result is in agreement with our findings using the Bayesian residual criterion.

In order to examine the performance of our proposed two-stage method against different data-generating mechanism, we simulated another removal dataset according to a gamma SIR model (m_2).

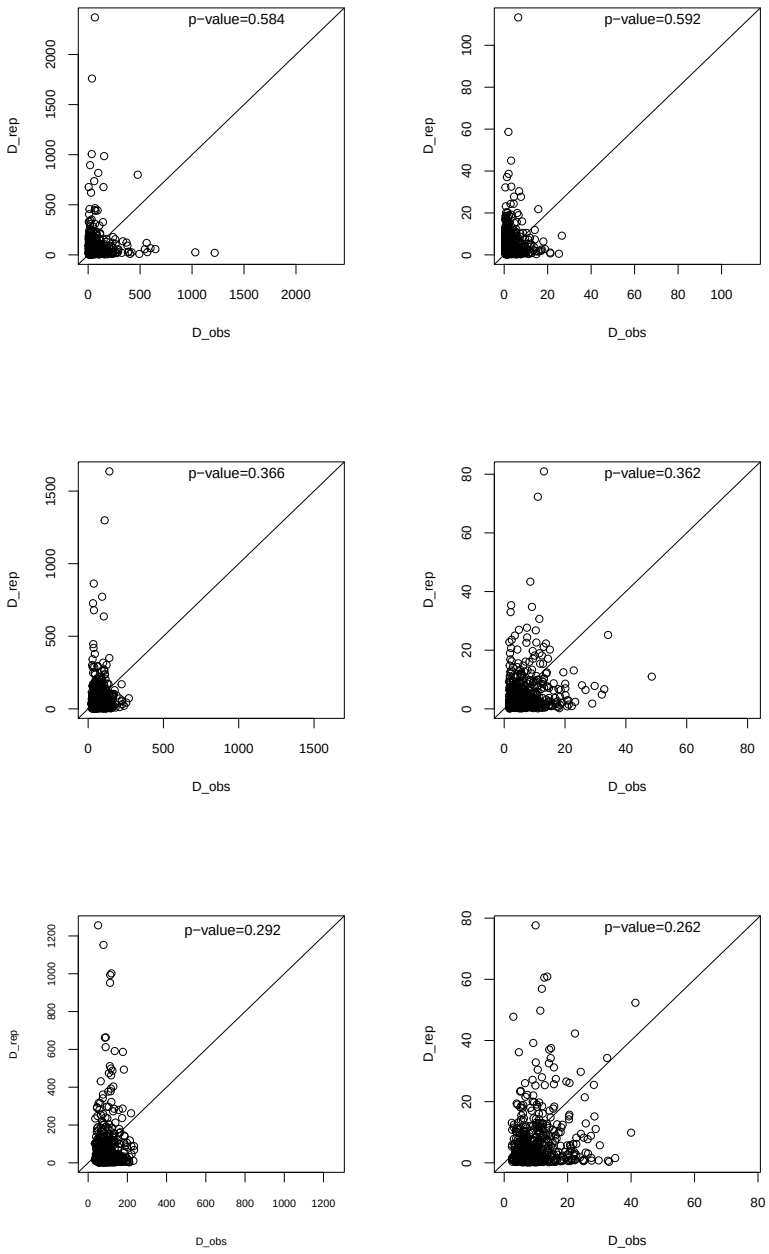


Figure 3.24: Scatter plots of observed versus replicated discrepancies based on 500 predicted removal curves from each removal times predictive distributions of the three SIR models. Left and right columns correspond to the χ^2 and Freeman-Tukey discrepancy measures. Top, middle and bottom rows refer to m_1, m_2 and m_3 respectively.

3.4.2.2 Sub-study II: Gamma SIR model data

A closed population of size $\boldsymbol{N} = 50$ individuals was considered and the infection rate was set to $\beta = 2.5$. The infectious periods were drawn from a gamma distribution with shape $\alpha = 10$ and scale $\delta = 10$. The observed data consisted of 45 removal times. We fitted the three SIR models to these removal data via MCMC methods and obtained 5000 posterior samples (after a burn-in of 2000 samples and thinning the chain by taking each $5th$ value) from each joint posterior distribution of each model parameters.

Having obtained these samples, we used them to simulate 5000 removal trajectories from each model as shown in Figure 3.25. The observed final size ($n_R - 1 = 44$) lies within the final size posterior predictive distribution for the three models (see Figure 3.26) which in turn allows us to condition on it and proceed with the second stage of our method. Figure 3.27 represents the average of 500 removal curves along with 95% percentile intervals generated from the removal times predictive distributions for the three SIR models in which the epidemic final sizes are matched with the observed one. It seems that the performance of the (true) m_2 model is better than the others as the observed removal data curve is consistent with the predicted mean over time.

- **Calculation of d_j and the $ppp-$value:**

Computation of d_j and the $ppp-$values was performed in a way that is similar to the way in the previous simulation study in Section 3.4.2.1. Figure 3.28 shows the Bayesian residual densities for the three SIR models in which there are high densities accumulated around zero, coming from the gamma and constant SIR models, compared to the exponential SIR model. Moreover, the sum of the squared Bayesian residuals $\sum_{j=1}^{n_R} d_j^2$ are 6.4, 1.6 and 1.6 for the exponential, gamma and constant SIR models respectively. This indicates that both m_2 and m_3 offer a good fit to the observed data which is perhaps not surprising as these SIR models may have similar removal curves behaviour when the

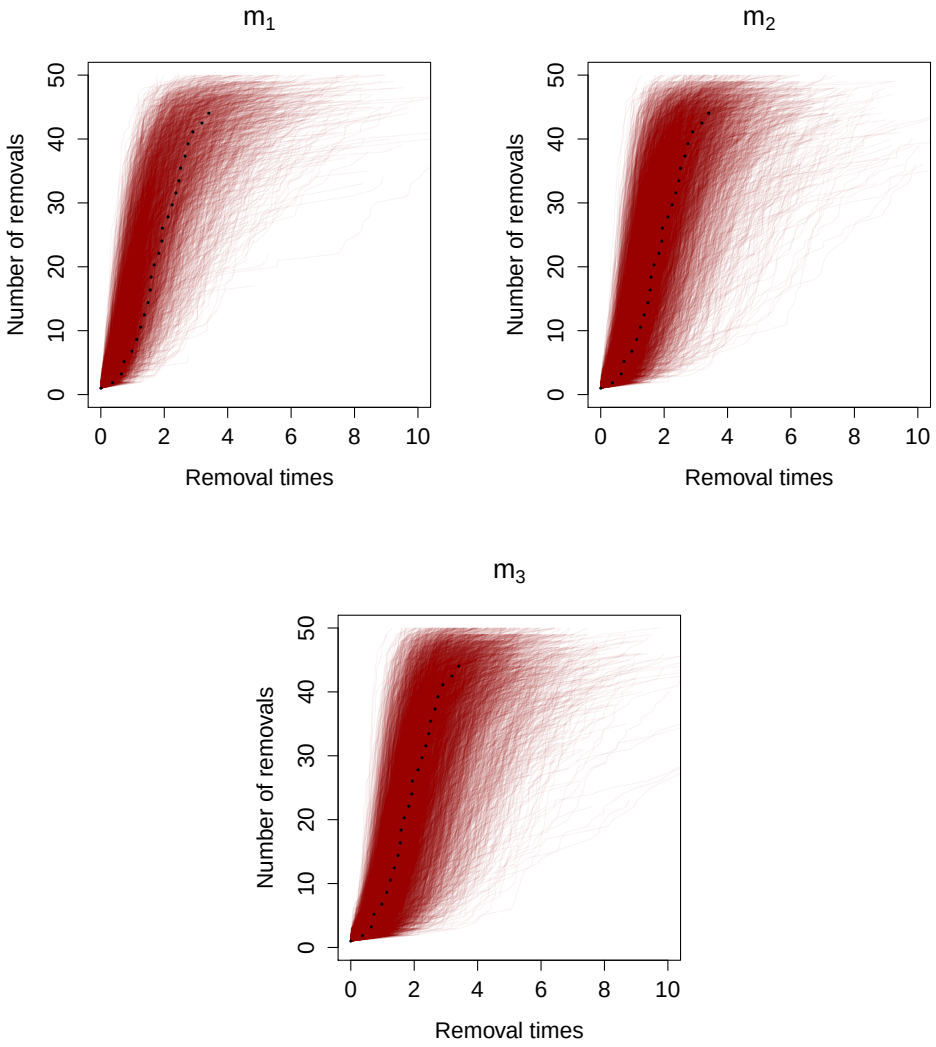


Figure 3.25: The posterior predictive distribution of the removal times for the three SIR models produced based on 5000 samples from the joint posterior density of each model parameters, where the dotted line indicates the observed data.

value of the shape parameter (α) in model m_2 increases (see Figure 3.18).

Figure 3.29 reveals scatter plots of observed versus replicated discrepancies of the χ^2 and Freeman-Tukey measures produced based on 500 samples from the removal times predictive distributions of the three SIR models along with the resulting ppp-values. Here, the gamma SIR model (m_2) is preferred over the other two SIR model (m_1 and m_3) as the obtained ppp-values by fitting m_2 are more closer to the optimal value 0.5.

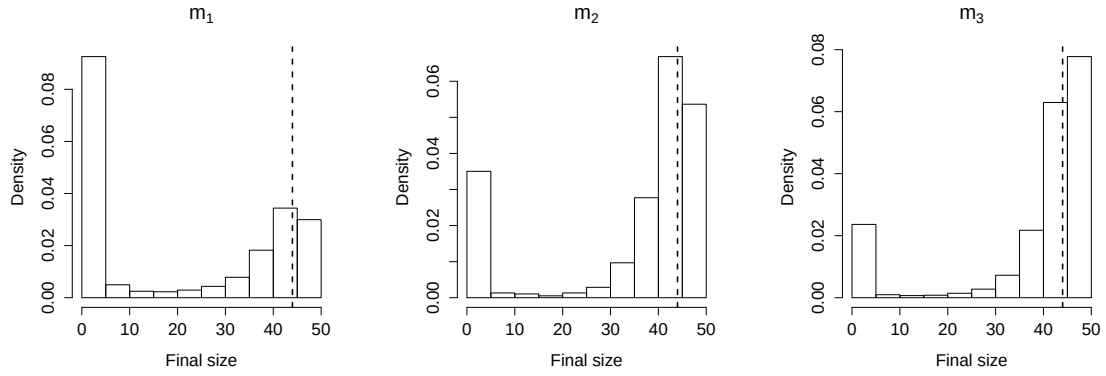


Figure 3.26: The final size posterior predictive distribution for the three SIR models where dotted line indicates the observed value of the final size ($n_R - 1 = 44$).

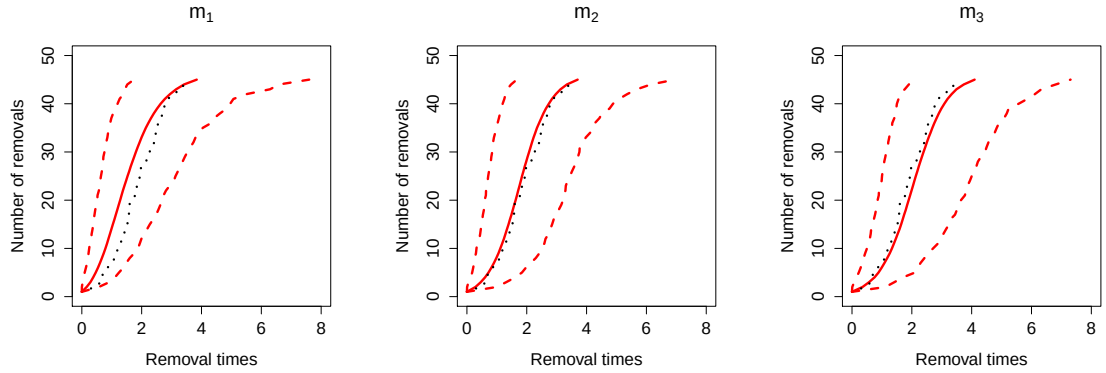


Figure 3.27: Predictive removal curves mean (red solid line) with 95% percentile intervals (red dashed lines) and the observed removal curve (black dotted line) for the three SIR models calculated based on 500 predicted removal curves generated from each model under the condition of the same epidemic final size.

In terms of ease of computation, in both simulation studies above, obtaining the posterior predictive p-values was somewhat computationally expensive when posterior samples of the parameters that result in extreme values of R_0 were selected. However, a possible way (although perhaps not efficient) to minimize the computational cost is to avoid extreme posterior values of R_0 by selecting randomly the $M' = 500$ samples from the 95% most likely R_0 posterior samples. With such approach, we may prevent the algorithm used to estimate $E(R_j|\boldsymbol{\phi}^{(k)})$ and $\text{var}(R_j|\boldsymbol{\phi}^{(k)})$ from getting stuck while simulating

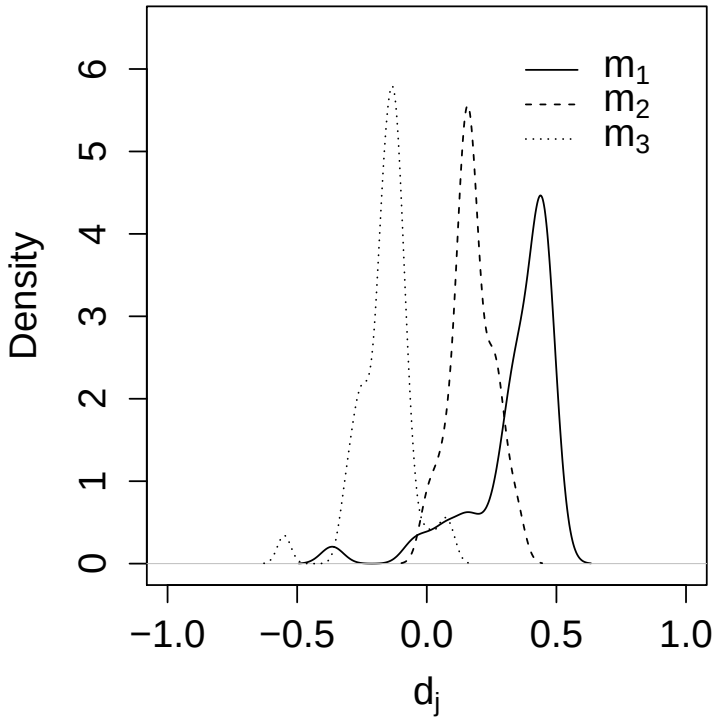


Figure 3.28: Bayesian residual densities for the three SIR models resulting from calculating (3.4.1) for each SIR model based on 500 posterior samples from the conditioning predictive distribution of removal times for the three models.

epidemic realisations from the model of interest, that have the same size as the observed one.

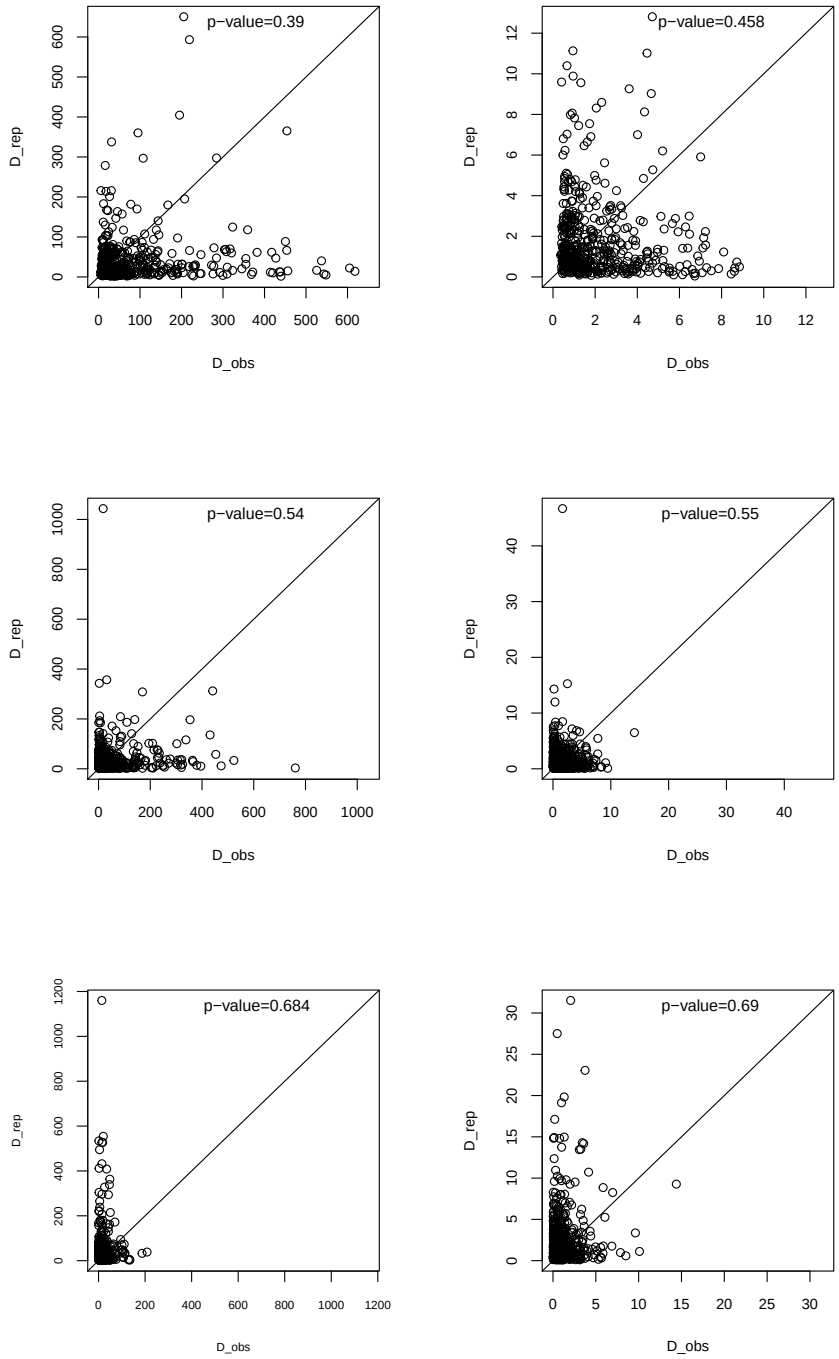


Figure 3.29: Scatter plots of observed versus replicated discrepancies based on 500 predicted removal curves from each removal times predictive distributions of the three SIR models. Left and right columns correspond to the χ^2 and Freeman-Tukey discrepancy measures. Top, middle and bottom rows refer to m_1, m_2 and m_3 respectively.

3.4.3 Application of the two-stage method to Abakaliki smallpox data

We now consider a widely studied temporal data set obtained from a smallpox outbreak in a closed population of 120 individual ($n = 119$, $a = 1$) in Abakaliki, Nigeria (Bailey, 1975, p. 125). The data consist of times between removals measured in days, that are

$$13, 7, 2, 3, 0, 0, 1, 4, 5, 3, 2, 0, 2, 0, 5, 3, 1, 4, 0, 1, 1, 1, 2, 0, 1, 5, 0, 5, 5,$$

where zero indicates that the case occurred at the same day as the previous one.

Therefore, in our notation the observed removal times are given by

$$\begin{aligned} \boldsymbol{R}^{obs} &= (0, 13, 20, 22, 25, 25, 25, 26, 30, 35, 38, 40, 40, 42, 42, 47, 50, 51, 55, 55, 56, \\ &\quad 57, 58, 60, 60, 61, 66, 66, 71, 76). \end{aligned}$$

The Abakaliki smallpox data are much studied by many previous authors. For example, Becker and Yip (1989) and O'Neill and Becker (2001) found it necessary to introduce variation in susceptibility when modelling such data. Xu (2015) has recently analysed these data using a new methodology which enables non-parametric estimation of the parameters that govern transmission within a Bayesian framework and has found that the assumption of constant infection rate over time is not appropriate for these data. Boys and Giles (2007) stressed the evidence of time-inhomogeneity in the period that individuals remain infectious in such outbreak. Eichner and Dietz (2003) analysed this outbreak and emphasized the importance of population structure aspects in the modelling.

However, our analysis does not intend to provide new insight into this outbreak, but rather it serves to show how good and appropriate is the fit of SIR or SEIR models with the homogeneous mixing assumption and different infectious periods to these outbreak data. In other words, our scientific ques-

tion of interest is, based upon these removal data, which of the choice of T_I is preferable for the smallpox data.

- **SIR model with different infectious periods** :

Here, we shall model these data via SIR model with different infectious periods, namely the model with exponential infectious periods (m_1), the model with gamma infectious periods (m_2), and the model with fixed length infectious periods (m_3). Although modelling these data using stochastic SIR models is not entirely appropriate since smallpox has an appreciable latent period, we shall follow O'Neill and Roberts (1999) in using these data to illustrate our method. We assigned gamma prior distributions with parameters $\lambda_\beta = \lambda_\gamma = \lambda_\delta = 1$ and $\nu_\beta = \nu_\gamma = \nu_\delta = 0.001$ to the parameters of the SIR models and it was assumed that the gamma shape parameter is known ($\alpha = 10$). Also, we assumed a priori that z (the label of the initial infective) is equally likely to be any of the $n_R = 30$ infected individuals and that I_z has an improper uniform prior density on $(-\infty, R_1)$. For the SIR model with constant infectious period ($T_I \equiv c$), we assumed that a priori c has an improper uniform prior density on $(0, \infty)$. We fitted the three SIR models, m_1, m_2 and m_3 using MCMC methods and then our two-stage method was applied to assess the goodness-of-fit of each model.

The posterior predictive distributions of the removal times shown in Figure 3.30 (top panel) do not seem to provide much information about the best choice of the SIR model type that best generates these smallpox data. Moreover, by looking at the final size predictive distributions (see Figure 3.30, middle panel) it may appear that none of these models offers a good fit to these outbreak data as the observed final size ($n_R - 1 = 29$) is located in tail of its predictive distribution under all models. Nevertheless, we conditioned on the observed final size and applied the second stage of our method, for illustration purposes. Figure 3.30, (bottom panel) demonstrates that there seems no clear evidence to prefer one specific infectious period choice and all the three

SIR models appear to provide a reasonable fit (in the predictive sense) to the observed data in the middle of the epidemic .However, early and later in the epidemic, all the three SIR models under-predict and over-predict the removal times respectively.

The resultant sum squared residuals (see Figure 3.31) of 1010.84 for the exponential SIR model (m_1) may give it some favour compared to 2184.36 and 1986.50 for the gamma and constant SIR models respectively.

Using the posterior predictive $p-$value as a Bayesian diagnostic tool does not seem to support any particular infectious period distribution. Figure 3.32 shows six scatter plots of observed versus replicated discrepancies of the χ^2 and Freeman-Tukey measures produced based on 500 samples from the removal times predictive distributions of the three SIR models along with the resulting ppp-values.

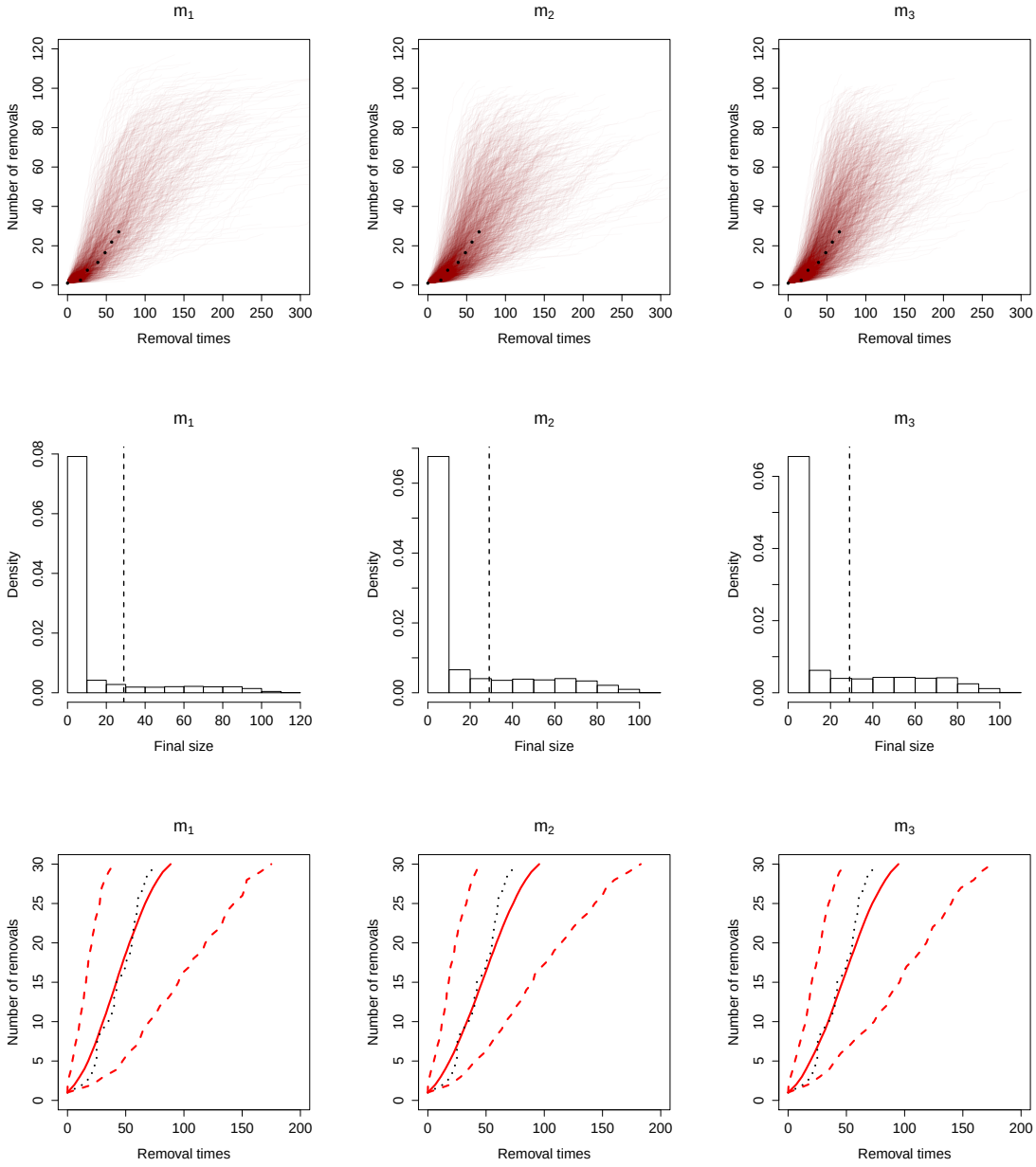


Figure 3.30: Top and middle rows show the removal times and final size predictive distributions for the exponential, gamma and constant SIR models calculated based on 5000 posterior samples of each model parameters, where the dotted line indicates the observed removal times and the vertical dashed line represents the observed final size ($n_R - 1 = 29$). Bottom row refers to the mean of 500 predicted removal curves along with 95% percentile intervals generated from the removal times predictive distributions for the three SIR models in which the epidemic final sizes are matched with the observed one.

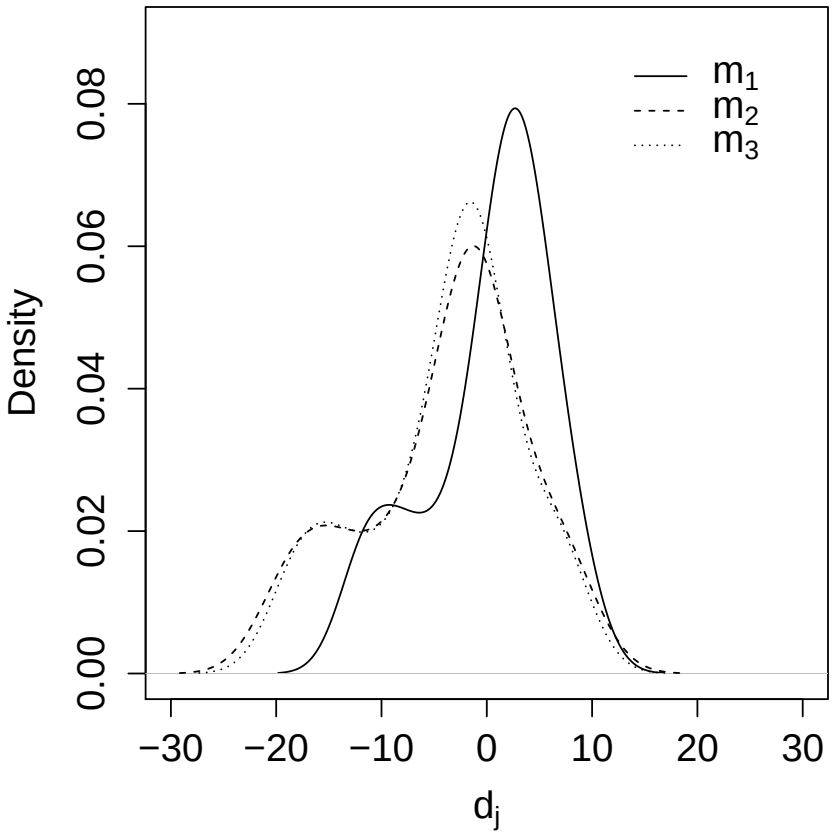


Figure 3.31: Bayesian residual densities for the three SIR models resulting from calculating (3.4.1) for each SIR model using 500 predicted removal curves.

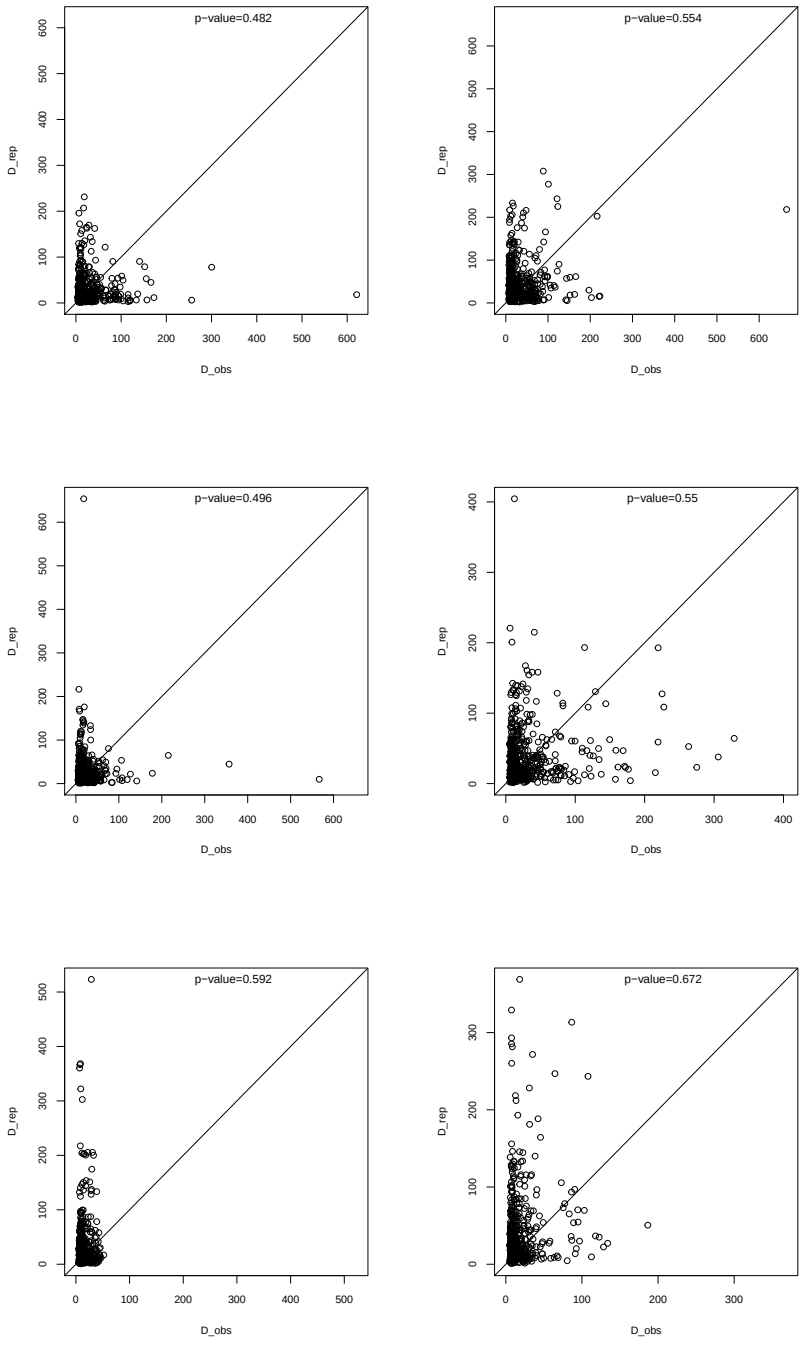


Figure 3.32: Scatter plots of observed versus replicated discrepancies based on 500 predicted removal curves from each removal times predictive distributions of the three SIR models. Left and right columns correspond to the χ^2 and Freeman-Tukey discrepancy measures. Top, middle and bottom rows refer to m_1, m_2 and m_3 respectively.

- **SEIR model with different infectious periods** :

Let us now consider a more general SEIR system, where n_R denotes the number of individuals becoming infectious at times $\boldsymbol{I} = (I_1, ..., I_{n_R})$ and removed from the population at times $\boldsymbol{R} = (R_1, ..., R_{n_R})$, where $0 = R_1 \leq R_2 \leq ... \leq R_{n_R}$. Similarly, let $\boldsymbol{E} = \{E_j : j \neq z\}$ denote the set of the $n_R - 1$ exposure times, that is, the times at which individuals acquire the infection without yet being able to transmit it, such that for $1 \leq j \leq n_R$, E_j corresponds to the exposure time of the individual who becomes infectious at time $I_j = E_j + L$ and removed at time R_j, where the initial exposed individual is labelled by z such that $E_z < E_j$ for all $j \neq z$. In our analysis, we assume a duration of $L = 12$ days as this is generally accepted to be the approximate length of the exposed period for the smallpox virus (O'Neill and Becker, 2001). Now, let $f_{T_I}(\cdot)$ denote the probability density (or mass, if T_I is discrete) function of T_I (the length of infectious period) and $\boldsymbol{\theta}$ indicate the parameter governing T_I, where $\boldsymbol{\theta}$ may be a vector. Then, the likelihood of the data given the model parameters can be expressed as

$$\begin{aligned} L(\boldsymbol{E}, \boldsymbol{I}, \boldsymbol{R}|\beta, \boldsymbol{\theta}, E_z, z) &= \left(\prod_{j=1, j\neq z}^{n_R} \beta n^{-1} Y(E_j-) \right) \times \exp\left(-\beta n^{-1} A'\right) \\ &\times \prod_{j=1}^{n_R} f_{T_I}\left(R_j - I_j|\boldsymbol{\theta}\right), \end{aligned} \tag{3.4.3}$$

where

$$A' = \int_{E_z}^{R_{n_R}} X(t)Y(t)dt = \sum_{j=1}^{n_R} \sum_{k=1}^{\boldsymbol{N}} (R_j \wedge E_k - I_j \wedge E_k), \tag{3.4.4}$$

with $E_k = I_k = R_k = \infty$ for $k = n_R + 1, ..., \boldsymbol{N}$.

Posterior inferences for the parameters can be obtained through appropriate Markov chain Monte Carlo (MCMC) techniques similar to that used for SIR model with different infectious periods. All prior parameters were set so as to be non-informative as previously in the SIR setting. Also, we assumed a priori that z (the label of the initial exposed individual) is equally likely to be any of the $n_R = 30$ exposed individuals and E_z was assigned an improper uniform

prior. Since we assume a fixed length of exposed period, the unknown exposure times $(\boldsymbol{E}, E_z)$ are updated automatically when updating the infection times as follows. Once the proposed value I_j^* for I_j is obtained (this is done exactly as in the SIR model setting), E_j is replaced by $I_j^* - 12$. Note that there is a chance here for z and E_z to be updated, so there is no need to update them separately.

Having obtained the desired samples from the joint posterior distribution of each model parameters, namely the SEIR model with exponential (m_1), gamma (m_2) and constant (m_3) infectious periods, our two-stage method was applied to assess the goodness-of-fit of each model. The final size predictive distributions (Figure 3.33) suggest that neither of these SEIR models offers a good fit to these outbreak data as the observed final size $n_R - 1 = 29$ is located in the low probability area of its predictive distribution under all SEIR fitted models.

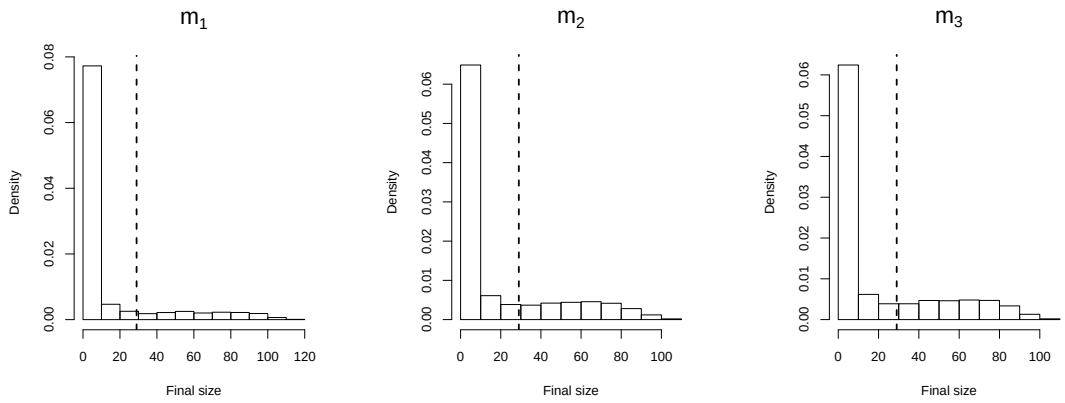


Figure 3.33: The final size predictive distributions for the exponential, gamma and constant SEIR models calculated based on 5000 posterior samples of each model parameters, where the vertical dashed line represents the observed final size ($n_R - 1 = 29$).

Also, Figure 3.34 represents the average of 500 removal curves along with 95% percentile intervals generated from the removal times predictive distributions for the three SEIR models in which the epidemic final sizes are matched with the observed one. It is obvious that all three SEIR models appear to provide a

reasonable fit to the observed data very early in the epidemic. However, later in the epidemic, all the three SEIR models over-predict the removal times. Such results indicate the inappropriateness of using these models for such removal data which is in agreement with the previous results of Boys and Giles (2007) for such SEIR models. In particular, Boys and Giles (2007) fitted to these

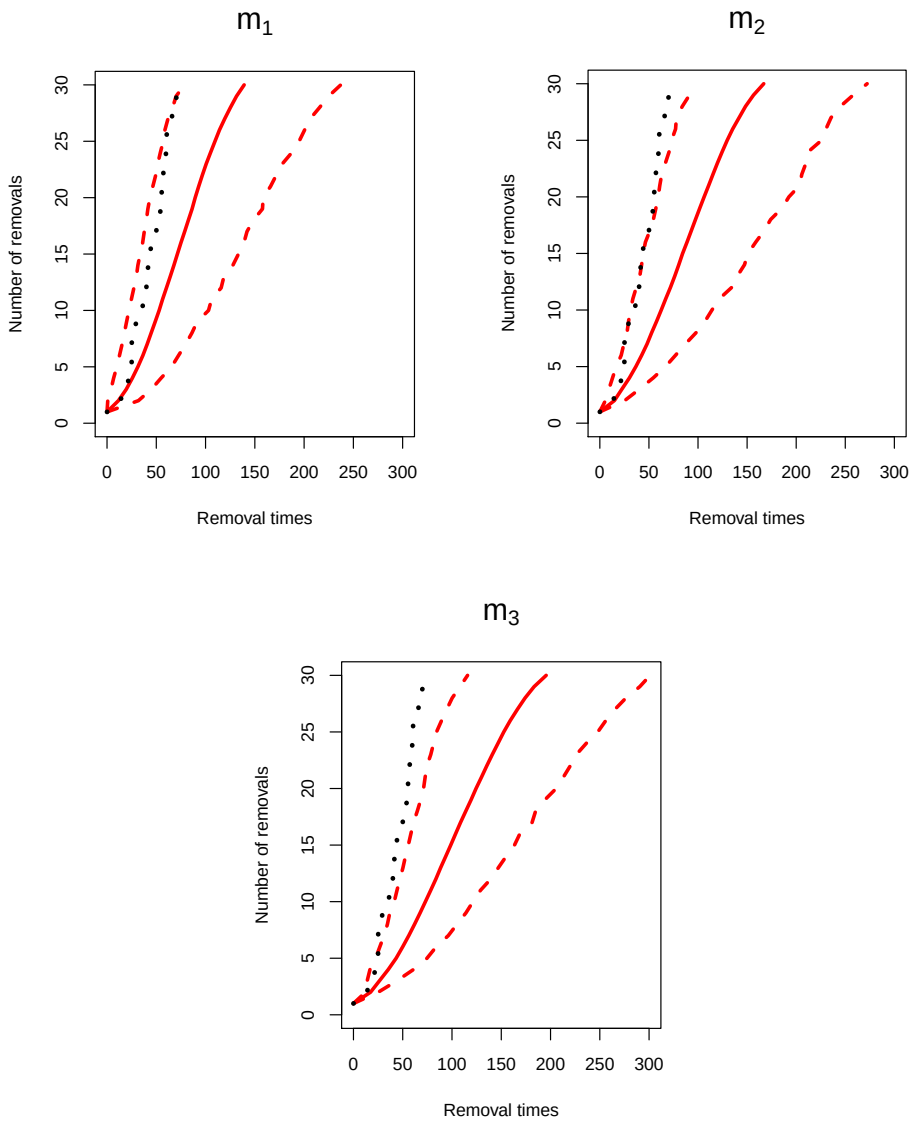


Figure 3.34: The mean of 500 predicted removal curves along with 95% percentile intervals generated from the removal times predictive distributions for the three SEIR models in which the epidemic final sizes are matched with the observed one and the dotted line indicates the observed removal times.

data an SEIR model in which the susceptible category is partitioned into subcategories, the infectious periods are exponentially distributed with constant removal rate and the latent period is fixed to 13 days (similar to model m_1 in our case). The authors found that only the second to the sixth removals are fitted well into their corresponding predictive distributions and the other removals fall well into the lower tails of their predictive distributions. The results of Boys and Giles (2007) can be corroborated by our results that obtained from fitting the exponential SEIR model (m_1) to these data as shown in Figure 3.35 from which it appears that our model provides a reasonable fit to the data until the occurrence of the fifth removals (up to 25 days) and then fails for the subsequent removals.

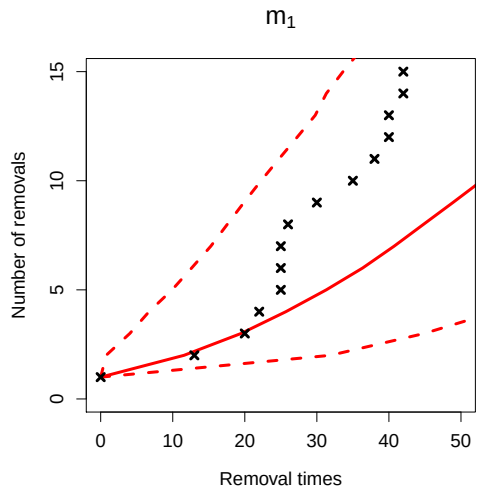


Figure 3.35: The mean of 500 predicted removal curves along with 95% percentile intervals generated from the removal times predictive distributions for the exponential SEIR model (m_1) in which the epidemic final sizes are matched with the observed one and the crossed points indicate the observed removal times. For the sake of clarity, we plot only the early observed and predicted removal times.

The posterior densities of d_j for the three models are shown in Figure 3.36, and the $ppp-$values obtained from fitting these models to the data are reported in Figure 3.37. These results generally indicate a poor fit of these SEIR models to

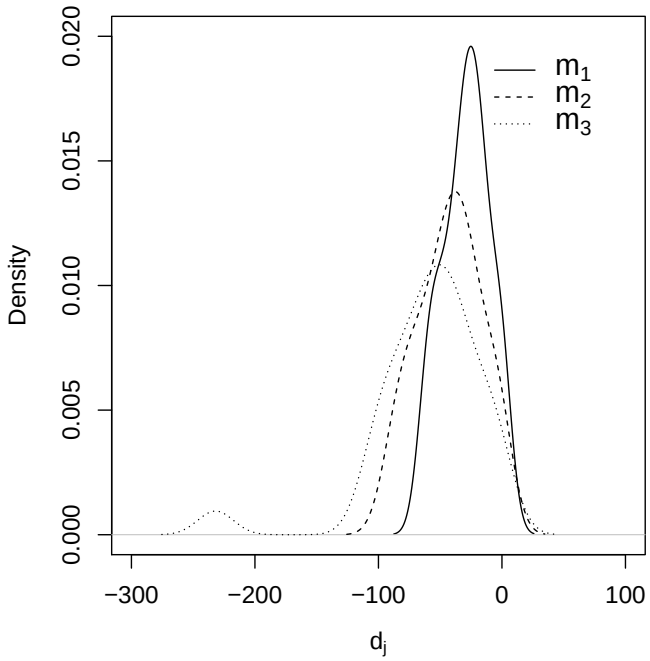


Figure 3.36: Bayesian residual densities for the three SEIR models resulting from calculating (3.4.1) for each SEIR model using 500 predicted removal curves.

the data.

In terms of computational time, calculation of d_j and the $ppp-$values was computationally costly since simulating $D = 100$ epidemic realisations with final size of 29 each (used to estimate $E(R_j|\boldsymbol{\phi}^{(k)})$ and $\text{var}(R_j|\boldsymbol{\phi}^{(k)})$) from either of the three SIR models or SEIR models was time-consuming process.

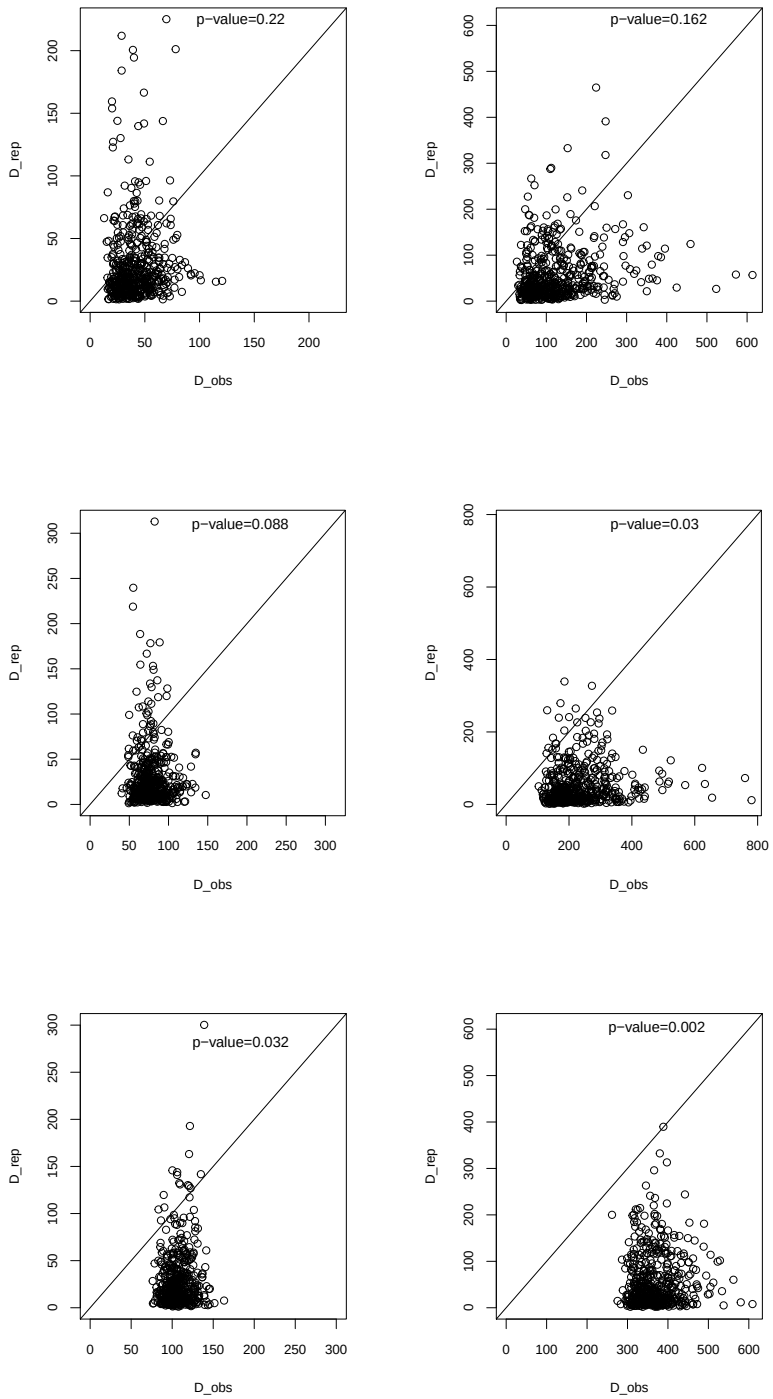


Figure 3.37: Scatter plots of observed versus replicated discrepancies based on 500 predicted removal curves from each removal times predictive distributions of the three SEIR models. Left and right columns correspond to the χ^2 and Freeman-Tukey discrepancy measures. Top, middle and bottom rows refer to m_1, m_2 and m_3 respectively.

3.5 Model assessment for epidemic models featuring a structured population

The so-called generalised stochastic epidemic model introduced on Chapter 1 is a rather simple epidemic model defined on a homogeneously mixing population where each individual has the same susceptibility to disease and the same ability to transmit disease. However, one of the focuses in recent studies of stochastic epidemic models is on models that feature different heterogeneities in the community in question. Heterogeneities caused by population social structures have been incorporated into epidemic models by assuming mixing at different levels, such as the two levels of mixing (e.g. household and community levels) model due to Ball et al. (1997) and the three levels of mixing (e.g. household, intermediate and community levels) model developed by Britton et al. (2011).

In this section, we consider another epidemic setting in which the goodness-of-fit for the SIR model with different population structures is assessed. The first model denoted by m_1 is the homogeneous (one level of mixing) SIR model with constant infectious period described in Chapter 2. The second model denoted by m_2 is the heterogeneous (two levels of mixing) SIR model introduced in Section 3.2.2. The assessment is mainly based on looking explicitly at the discrepancy between the observed removal curve and the predicted ones under these two epidemic models.

The rationale behind looking at the question of goodness-of-fit in this specific epidemic setting is to point out the importance of detecting the most appropriate epidemic model before considering any policy for the control of disease transmission which mainly based on estimation of the immunity coverage, that is the minimum proportion of individuals in the community who must be immune so that epidemics are prevented. Using data on the eventual size of a major epidemic, Becker and Utev (1998) compared estimates of the critical immunity level obtained from the same data but under different assumptions

about the community structure. The authors compared estimate based on the simplifying assumption that the community consists of uniformly mixing individuals with estimate obtained when the community is made up of households of equal size. It was found that assuming an oversimplified community structure underestimates the critical immunity coverage which in turn leads to the risk of believing that we have estimated an immunity coverage that can prevent epidemics when it in fact cannot.

Bayesian inference via Markov chain Monte Carlo (MCMC) methods, based on data on removals, for the first homogeneous epidemic model m_1 was illustrated in Chapter 2, where a data-augmentation MCMC algorithm was described to obtain samples from the joint posterior distribution of model parameters. Now, we shall consider Bayesian parameter estimation for the second model m_2, the SIR model with two levels of mixing, given only removal times.

3.5.1 Bayesian inference via MCMC for the SIR model with two levels of mixing given removal data

Consider a closed population of size $\boldsymbol{N}$ (n initial susceptibles and one initial infective) and let $X^C(t)$ and $Y^C(t)$ denote respectively the number of susceptible and infective individuals in the community, at time t. Also, let $X_i^H(t)$ and $Y_i^H(t)$ represent the corresponding numbers in household i at time t, where $i = 1, ..., h$. For simplicity, we assume that all infected individuals are removed in a certain observation period $(I_z, R_{n_R}]$, so the epidemic has run its course and there are in total $n_R \leq \boldsymbol{N}$ removal events. For $j = 1, ..., n_R$, let I_j be the infection time of the individual removed at time R_j . Also, denote the initial infection time by I_z, where $I_z < I_j$, for all $j \neq z$. Finally, let $\boldsymbol{R} = (R_1, ..., R_{n_R})$ represent the set of observed removal times, where $0 = R_1 \leq R_2 \leq ... \leq R_{n_R}$ and $\boldsymbol{I} = \{I_j : j \neq z\}$ denote the set of missing infection times. Now, let $f_{T_I}(\cdot)$ denote the probability density (or mass, if T_I is discrete) function of T_I (the length of infectious period) and $\boldsymbol{\theta}$ indicate the parameter governing T_I, then

the likelihood of $\boldsymbol{I}$ and $\boldsymbol{R}$ given $\lambda_H, \lambda_C, \boldsymbol{\theta}, z$ and I_z is (Britton et al., 2011)

$$\begin{aligned} L(\boldsymbol{I}, \boldsymbol{R}|\lambda_H, \lambda_C, \boldsymbol{\theta}, I_z, z) \propto & \prod_{i,j,j\neq z} \left(\boldsymbol{N}^{-1}\lambda_C Y^C(I_j-) + \lambda_H Y_i^H(I_j-) \right) \\ & \times \exp\left(-\int_{I_z}^{R_{n_R}} \boldsymbol{N}^{-1}\lambda_C X^C(t) Y^C(t) dt \right) \\ & \times \exp\left(-\int_{I_z}^{R_{n_R}} \sum_{i=1}^{h} \lambda_H X_i^H(t) Y_i^H(t) dt \right) \\ & \times \prod_{j=1}^{n_R} f_{T_I}\left(R_j - I_j|\boldsymbol{\theta}\right), \end{aligned} \tag{3.5.1}$$

where I_j- denotes the time just prior to I_j and $I_k = R_k = \infty$ for $k = n_R + 1, ..., \boldsymbol{N}$.

In our case, we assume T_I to be fixed and identical for all individuals, so that $T_I \equiv c$. Therefore, the contribution of the removal process to the likelihood (the last line in (3.5.1)) is simply $\mathbb{1}_{\{c=R_j-I_j,\, j=1,...,n_R\}}$. It follows that, if $\boldsymbol{I}$, $\boldsymbol{R}, z$ and I_z are all observed, then the inference for λ_H, λ_C and c is achievable (Britton et al., 2011). However, the more practical case is when the observed data consist of removal times only. Bayesian inference for the two levels of mixing stochastic epidemic model based on removal data can be performed as follows.

To each parameter λ_H and λ_C, we assign an independent gamma prior distribution, namely Gamma(a_ζ, b_ζ), where $\zeta = \lambda_H,\ \lambda_C$. Further, we assume a priori that z is equally likely to be any of the n_R infected individuals, I_z has an improper uniform prior density on $(-\infty, R_1)$ and a priori $T_I = R_j - I_j = c$ is uniformly distributed on $(0, \infty)$. As the method of updating $\boldsymbol{I}$, described below, allows the value of I_z to change at each MCMC iteration, there is no need to update z and I_z in separate steps.

Having assigned prior distributions to the quantities of interest, the following joint posterior density of the model parameters can be obtained by multiplying the likelihood function and the prior distributions using Bayes' theorem,

that is

$$\begin{aligned}\pi(c,\lambda_H,\lambda_C|\boldsymbol{R}) &\propto \left[\prod_{i,j,j\neq z}\left(\boldsymbol{N}^{-1}\lambda_C Y^C(I_j-)+\lambda_H Y_i^H(I_j-)\right)\right]\\ &\times \exp\left[-\int_{I_z}^{R_{n_R}}\left(\boldsymbol{N}^{-1}\lambda_C X^C(t)Y^C(t)+\sum_{i=1}^{h}\lambda_H X_i^H(t)Y_i^H(t)\right)dt\right]\\ &\times \lambda_H^{a_{\lambda_H}-1}\lambda_C^{a_{\lambda_C}-1}\times\exp\left[-\left(\lambda_H b_{\lambda_H}+\lambda_C b_{\lambda_C}\right)\right]\\ &\times \mathbb{1}_{\{c=R_j-I_j,\,j=1,\ldots,n_R\}}. \end{aligned} \tag{3.5.2}$$

It follows that

$$\begin{aligned}\pi(\lambda_H|\boldsymbol{R},c,\lambda_C) &\propto \left[\prod_{i,j,j\neq z}\left(\boldsymbol{N}^{-1}\lambda_C Y^C(I_j-)+\lambda_H Y_i^H(I_j-)\right)\right]\\ &\times \lambda_H^{a_{\lambda_H}-1}\exp\left[-\lambda_H\left(b_{\lambda_H}+\int_{I_z}^{R_{n_R}}\sum_{i=1}^{h}X_i^H(t)Y_i^H(t)dt\right)\right]\\ &\times \mathbb{1}_{\{c=R_j-I_j,\,j=1,\ldots,n_R\}}, \end{aligned} \tag{3.5.3}$$

$$\begin{aligned}\pi(\lambda_C|\boldsymbol{R},c,\lambda_H) &\propto \left[\prod_{i,j,j\neq z}\left(\boldsymbol{N}^{-1}\lambda_C Y^C(I_j-)+\lambda_H Y_i^H(I_j-)\right)\right]\\ &\times \lambda_C^{a_{\lambda_C}-1}\exp\left[-\lambda_C\left(b_{\lambda_C}+\boldsymbol{N}^{-1}\int_{I_z}^{R_{n_R}}X^C(t)Y^C(t)dt\right)\right]\\ &\times \mathbb{1}_{\{c=R_j-I_j,\,j=1,\ldots,n_R\}}, \end{aligned} \tag{3.5.4}$$

and

$$\begin{aligned}\pi(c|\boldsymbol{R},\lambda_H,\lambda_C) &\propto \left[\prod_{i,j,j\neq z}\left(\boldsymbol{N}^{-1}\lambda_C Y^C(I_j-)+\lambda_H Y_i^H(I_j-)\right)\right]\\ &\times \exp\left[-\int_{I_z}^{R_{n_R}}\left(\boldsymbol{N}^{-1}\lambda_C X^C(t)Y^C(t)+\sum_{i=1}^{h}\lambda_H X_i^H(t)Y_i^H(t)\right)dt\right]\\ &\times \mathbb{1}_{\{c=R_j-I_j,\,j=1,\ldots,n_R\}}. \end{aligned} \tag{3.5.5}$$

The model parameters λ_H, λ_C and c, may be updated using random walk Metropolis steps with a Gaussian proposal density centred on the current value, with a fixed standard deviation. Furthermore, because only positive values are permitted for these parameters, the absolute value of the proposed value is taken. It can be seen that the unknown infection times $\boldsymbol{I}$ and I_z are updated automatically when c is updated as $I_j = R_j - c,\ j = 1,\ldots,n_R$. Sampling from

Algorithm 7 MCMC algorithm for updating λ_H, λ_C and c

1. Initialise the algorithm with λ_H^0, λ_C^0 and c^0.

2. Repeat the following steps a large number of times, say M.

* Update λ_H using Gaussian random walk Metropolis Sampler as follows:

a. Draw $\lambda_H^* \sim N(\lambda_H, \sigma^2_{\lambda_H})$, where λ_H^* denotes the proposed value of λ_H.

b. Draw a uniform random variable $U \in (0,1)$.

c. Accept the proposed value λ_H^* if

$$U \leq \min\left\{\frac{\pi(\lambda_H^*|\boldsymbol{R}, c, \lambda_C)}{\pi(\lambda_H|\boldsymbol{R}, c, \lambda_C)}, 1\right\}$$

* Update λ_C using Gaussian random walk Metropolis Sampler as follows:

a. Draw $\lambda_C^* \sim N(\lambda_C, \sigma^2_{\lambda_C})$, where λ_C^* denotes the proposed value of λ_C.

b. Draw a uniform random variable $U \in (0,1)$.

c. Accept the proposed value λ_C^* if

$$U \leq \min\left\{\frac{\pi(\lambda_C^*|\boldsymbol{R}, c, \lambda_H)}{\pi(\lambda_C|\boldsymbol{R}, c, \lambda_H)}, 1\right\}$$

* Update c using Gaussian random walk Metropolis Sampler as follows:

a. Draw $c^* \sim N(c, \sigma_c^2)$, where c^* denotes the proposed value of c.

b. Set $I_j = R_j - c^*,\ j = 1, ..., n_R$.

c. Draw a uniform random variable $U \in (0,1)$.

d. Accept the proposed value c^* if

$$U \leq \min\left\{\frac{\pi(c^*|\boldsymbol{R}, \lambda_H, \lambda_C)}{\pi(c|\boldsymbol{R}, \lambda_H, \lambda_C)}, 1\right\}$$

3. Return $\{\lambda_H^0, \lambda_H^1, ..., \lambda_H^M\}$, $\{\lambda_C^0, \lambda_C^1, ..., \lambda_C^M\}$ and $\{c^0, c^1, ..., c^M\}$.

the joint posterior distribution of these parameters can be done via the Markov Chain Monte Carlo algorithm 7.

Algorithm 7 not only yields posterior summaries such as the posterior mean

and the posterior variance of each parameter but also allows us to use the posterior samples to generate samples from the posterior predictive distribution.

3.5.2 Threshold and removal trajectories behaviour

In stochastic epidemic models, typically there is a threshold quantity which determines whether an epidemic occurs or the disease simply dies out. This quantity is called the threshold parameter or the basic reproduction number, denoted by R_* here, which represents the average number of households infected by a typical infected household in a totally susceptible population and can be derived (e.g. Ball et al. (1997)) by assuming that the population size, $\boldsymbol{N}$, becomes large in such a way that household sizes remain unchanged, but number of households $h \to \infty$. In our setting, as we have equal-sized households, the threshold parameter is simply

$$R_* = \lambda_C E(T_I) E(F_H), \tag{3.5.6}$$

where $E(F_H)$ is the expected final number of individuals in the household ever infected, including the initial infective which can be computed from a system of triangular equations (see e.g. Ball (1986) or (Andersson and Britton, 2000, p. 15-16)).

The homogeneous SIR model (m_1) can be considered as a special case of the two levels of mixing epidemic model (m_2) when letting each household to be of size one individual. Moreover, the household structure effect on the data becomes apparent as $\lambda_H \to \infty$ and $\lambda_C \to 0$.

In what follows, we investigate systematically via simulation the removal curves distribution for the two levels of mixing epidemic model. The primary objective for performing this simulation study is to see how different, with regard to removal curves, the two levels of mixing model (m_2) is from the homogeneous model (m_1). To be precise, we fix both the infectious period c and the community rate parameter λ_C, whereas the household rate parameter λ_H is given various values. A population of size $\boldsymbol{N} = 100$ was divided into $h = 20$

households each of size $s = 5$ individuals. We set $\lambda_C = 0.1$ and $c = 5$ but λ_H was given different values shown in Table 3.5 with the corresponding $E(F_H)$ and R_* values.

λ_H	0.08	0.1	0.4	1
$E(F_H)$	3.339	3.809	4.997	5
R_*	1.670	1.905	2.499	2.5

Table 3.5: Summary of different values for λ_H with the corresponding $E(F_H)$ and R_* values used to produce Figure 3.38.

Figure 3.38 shows the removal trajectories distribution for the two levels of mixing model with different sets of parameters. As can be seen form Figure 3.38 ((a) and (b)), when there is a lower household structure effect in the population (i.e. not all susceptible individuals in a household are infected), the distribution of removal curves for this model look similar to the corresponding one for the homogeneous SIR model (m_1) discussed in Section 3.4. However, when λ_H gets large (see Figure 3.38: (c) and (d)) in the sense that $E(F_H) \to s$, then the removal curves behaviour of this model becomes quite distinguishable.

As an example of showing the difference between the two models, m_1 and m_2, with regard to removal trajectories, we considered a closed population of size $\boldsymbol{N} = 100$ and simulated 1000 removal curves from each model in which we fixed the infectious period to be $c = 5$ for both models and ensured that the simulated-based final size distributions are similar in the sense that the major outbreak parts are peaked at the same range of values (see Figure 3.39). For model m_1, the value of the infection rate parameter was $\beta = 0.55$. For model m_2, we assumed that the population is divided into $h = 20$ households each of size $s = 5$ individuals and set the household and community infection parameters to $\lambda_H = 1.4$ and $\lambda_C = 0.12$ in a way such that there is a household structure effect on the data. Figure 3.40 reveals that the removal time curves distributions are different. A key finding is that although both models have

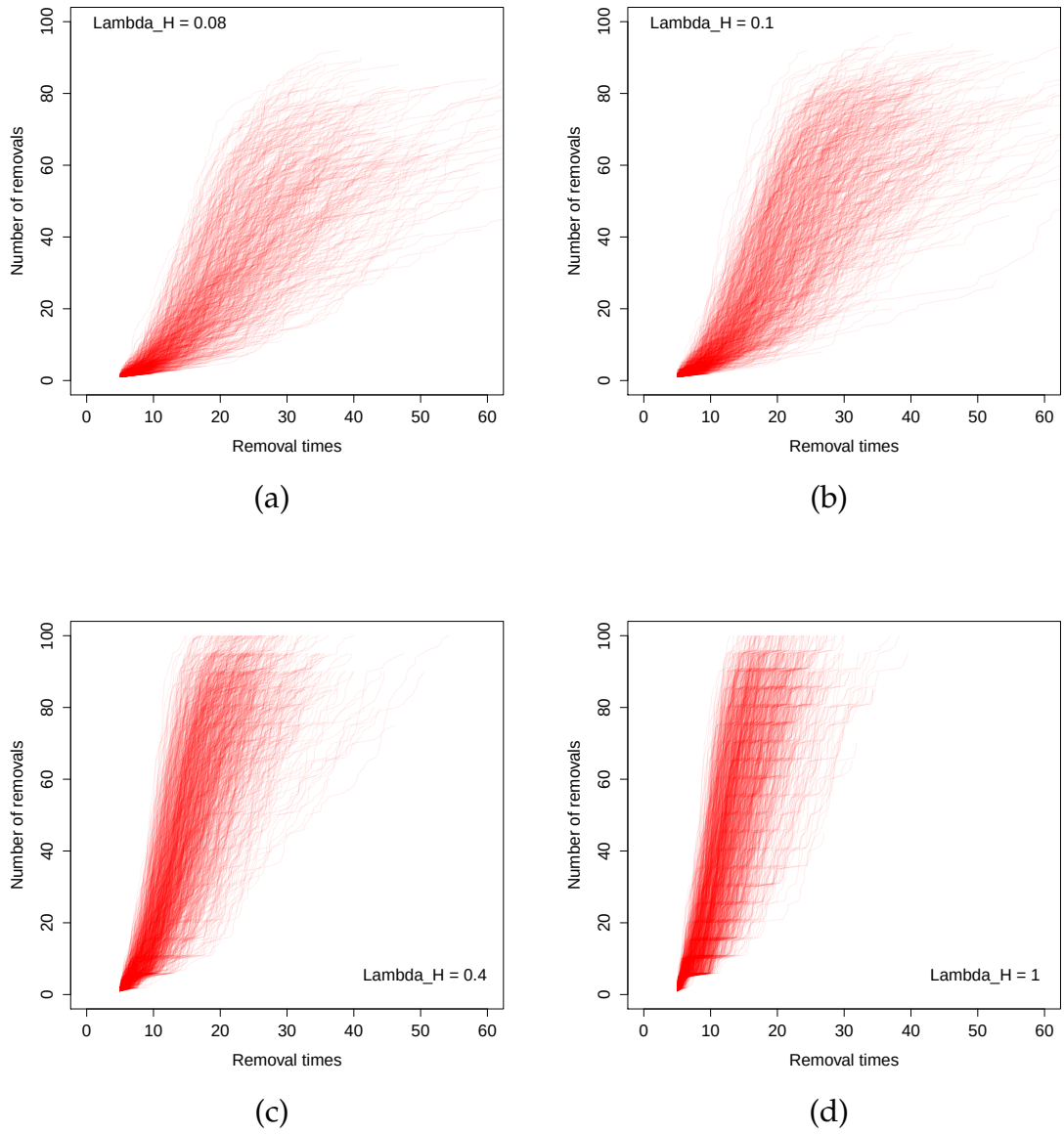


Figure 3.38: The distributions of 1000 removal curves simulated from the two levels of mixing model (m_2) for different values of λ_H. The community infection rate parameter and the length of infectious period are fixed to $\lambda_C = 0.1$ and $c = 5$ throughout simulations.

the same length of infectious period ($c = 5$) and similar final size distributions, the household structure in the two levels of mixing model (m_2) results in clear jump behaviour in the removal trajectories distribution (see Figure3.40). This phenomenon can be explained as follows. With such parameter values used in model m_2, the epidemic is highly infectious within households to the point

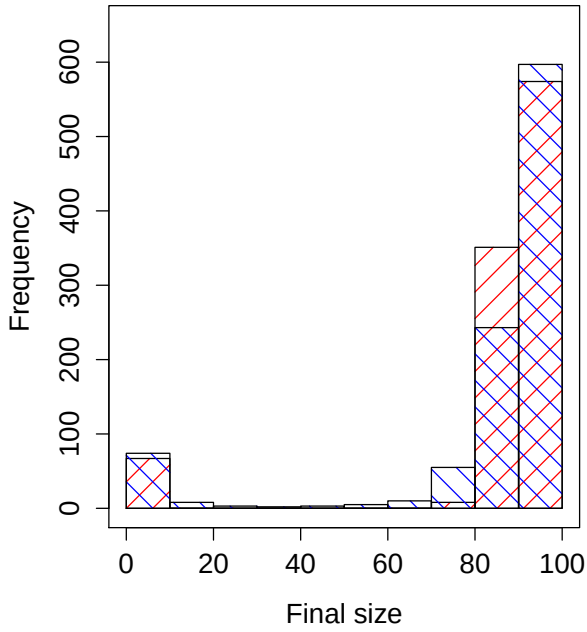


Figure 3.39: Histograms of the final size of 1000 epidemics simulated from the two epidemic models. The red and blue shaded histograms refer to models m_1 and m_2 respectively. The parameter values used are : $\beta = 0.55, \lambda_H = 1.4, \lambda_C = 0.12$ and $c = 5$.

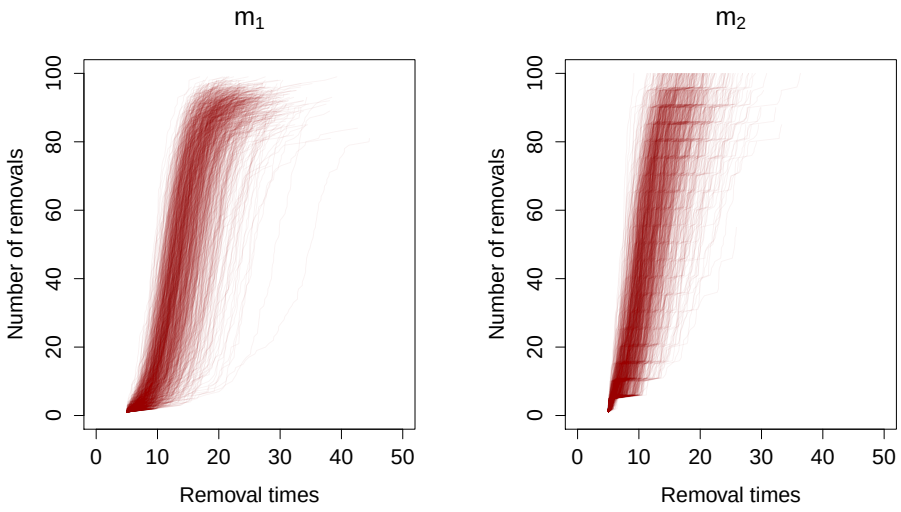


Figure 3.40: The distributions of 1000 removal curves simulated from the two epidemic models with the same length of infectious period ($c = 5$). The other parameter values used are : $\beta = 0.55, \lambda_H = 1.4$ and $\lambda_C = 0.12$.

that infection of a single member of a household leads to an immediate infection of all susceptibles in that household. Further, as we assumed fixed equal infectious periods, all infected individuals in a household are removed roughly at the same time and so a jump of roughly $s = 5$ in the number of removals is to be expected.

Having observed the difference between the two SIR models, m_1 and m_2, in terms of the removal curves distribution, we shall make use of it via our two-stage method to discriminate between these epidemic models when fitting them to partially observed outbreak data. Throughout the analysis, non-informative gamma prior distributions with parameters $a_\zeta = 1$ and $b_\zeta = 0.001$, where $\zeta = \beta,\ \lambda_H,\ \lambda_C$ are assigned to the parameters $\beta,\ \lambda_H$, and λ_C. We assume an improper uniform prior density on $(-\infty, R_1)$ for I_z and a priori c is uniformly distributed on $(0, \infty)$. Also, we assume a priori that z (the label of the initial infective) is equally likely to be any of the n_R infected individuals.

3.5.3 Application of the two-stage method to simulated outbreak data and results

Here, we shall test the ability of our proposed two-stage method to detect any heterogeneity in the data using simulated outbreak data. This will be performed by going along the lines of Algorithm 5 as well as using the two Bayesian criteria introduced in Section 3.2.3. Additionally, we will infer the posterior distributions of number of household and community infections and used it as an indication of model fit.

This simulation study itself is comprised of two simulation sub-studies within each of which different source of simulated data is considered. The first sub-study considers the situation in which the observed data are simulated from a two levels of mixing SIR model, whereas in the second sub-study the data-generated model is a homogeneous mixing SIR model. In both sub-studies, we fitted the true data-generated model and the other misspecified model to

the observed removal data and then aimed to identify the correct model using our two-stage procedure. Under each scenario, we simulated and analysed several outbreak datasets. Our findings are based on one typical simulated dataset for each scenario.

3.5.3.1 Sub-study I: Two levels of mixing population removal data

A closed population of $\boldsymbol{N} = 100$ individuals was divided into $h = 20$ households of size $s = 5$ each. An epidemic was simulated from a two levels of mixing SIR model using this population structure which had 90 removals. The household infection rate parameter was set to $\lambda_H = 0.9$ and the community infection rate parameter was set to $\lambda_C = 0.1$. The infection rate values were chosen to reflect a community in which mixing is quite high within households and much less between them. Each infected individual remained infectious for a fixed infectious period, $c = 5$.

Having simulated these removal times, we fitted the correctly specified model m_2 to these observed data using MCMC method described in Algorithm 7 and we also fitted a misspecified model m_1 to these data using MCMC algorithm illustrated in Chapter 2. After a burn-in period of 2000 iterations, and thinning the chain by taking each $5th$ value, $M = 5000$ posterior samples were obtained from the joint posterior distribution of c, λ_H, and λ_C for m_2 model and from the joint posterior distribution of c and β for m_1 model. The convergence properties of the Markov chains were explored by looking at the plot of the sample traces of the parameters of interest. Figure 3.41 represents the sample traces along with the posterior distributions of λ_H and λ_C in model m_2. From these plots, it appears that there is a heterogeneous structure behind generating these observed data as the posterior estimate of λ_H is far larger than the posterior estimate of λ_C.

Having obtained these samples, we can use them to simulate 5000 removal curves with their corresponding final sizes from the two models, that are the posterior predictive distributions of removal times and final size in or-

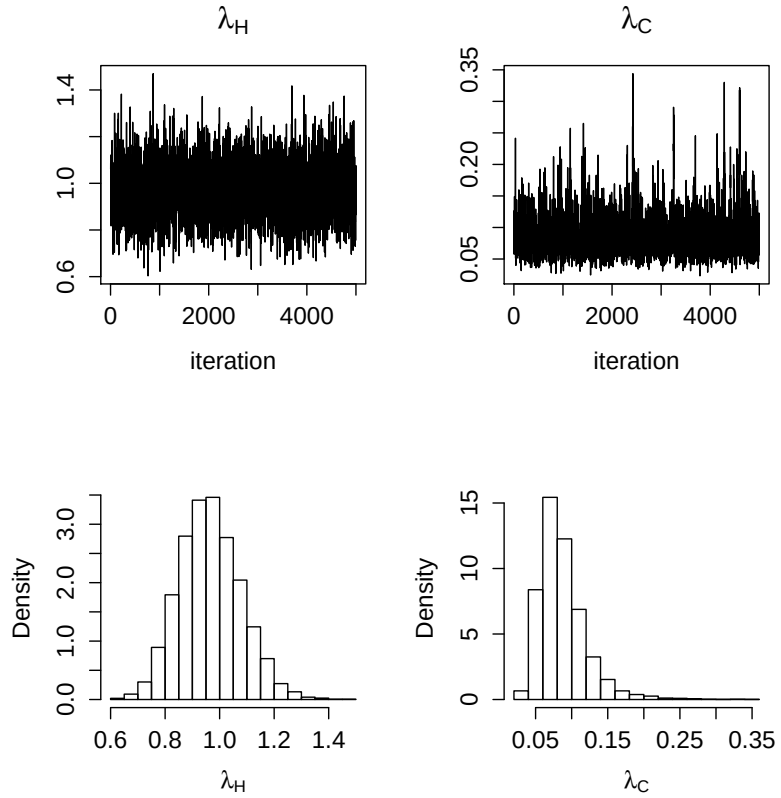


Figure 3.41: Posterior sample traces and histograms of λ_H and λ_C obtained from fitting m_2 to the observed data after a burn-in period of 2000 iterations and a thinning of 5 samples. True parameter values were: $\lambda_H = 0.9$ and $\lambda_C = 0.1$.

der to check first that the observed final size lies within the final size posterior predictive distributions and then proceed by conditioning on it and apply a formal Bayesian diagnostic tool for the goodness-of-fit.

Figure 3.42 (top panel) reveals that the observed removal data can be seen as a realisation from the two levels of mixing model m_2, the true model that has generated the data. Additionally and more importantly, it can be observed from the same figure (bottom panel) that the observed final size ($n_R - 1 = 89$) fits within the final size predictive distributions for both models which in turn enables us to apply the second stage of our procedure.

Plots of the posterior predictive mean of 500 removal trajectories obtained from their conditioning predictive distributions for the two models, along with the associated 95% confidence bands, are illustrated in Figure 3.43. These

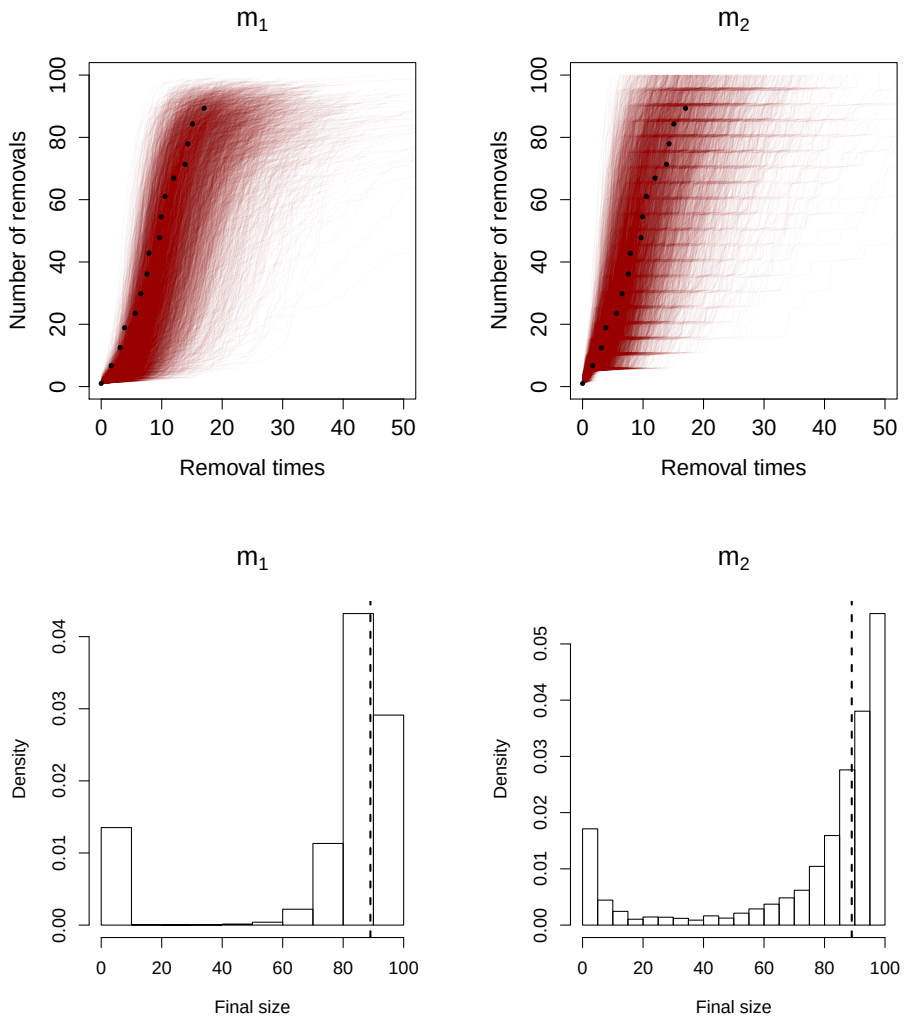


Figure 3.42: Top and bottom rows show the removal times and final size predictive distributions for the two models respectively, calculated based on 5000 posterior samples of each model parameters, where the dotted line indicates the observed removal times and the vertical dashed line represents the observed final size ($n_R - 1 = 89$). Left and right columns correspond to m_1 and m_2 respectively.

conditioning predictive distributions can be obtained simply by keeping simulating (until the desired sample size is obtained) from the two models using $M' = 500$ randomly chosen samples from the posterior distributions of each model parameters while rejecting simulations that do not match the observed final size. It is clearly noticeable from Figure 3.43 that the posterior predictive

performance under m_2 seems to outperform the posterior prediction under m_1 with noticing that the observed removal curve deviates most of the times from the corresponding posterior predictive mean under m_1.

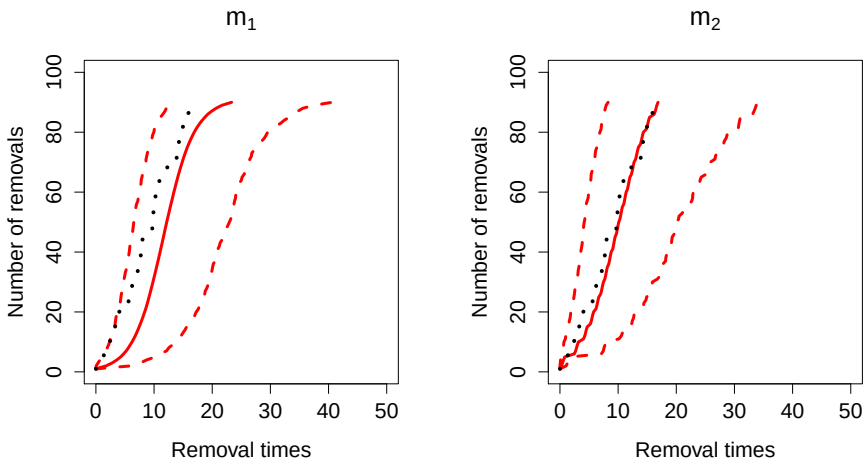


Figure 3.43: Predictive mean of 500 removal curves (red solid line) with 95% percentile intervals (red dashed lines) and the observed removal curve (black dotted line) for the two models m_1 and m_2 (left to right).

Now, we shall apply three different Bayesian formal ways for measuring the goodness-of-fit for both models.

- **Bayesian residual criterion:**

Here, we define the Bayesian residual test as an overall tool to measure the goodness-of-fit as

$$\begin{aligned} d_j &= R_j^{obs} - E(R_j^{rep}|\boldsymbol{R}^{obs}) \\ &\approx R_j^{obs} - \frac{1}{M'}\sum_{k=1}^{M'} R_j^{rep\ (k)}. \end{aligned} \tag{3.5.7}$$

Figure 3.44 shows the Bayesian residual densities for the two models calculated using (3.5.7) in which it is obvious that the residual values produced under the m_2 model accumulated near zero compared to those for the m_1 model. In addition, the sum of the squared Bayesian residuals $\sum_{j=1}^{n_R} d_j^2$ were 872.39

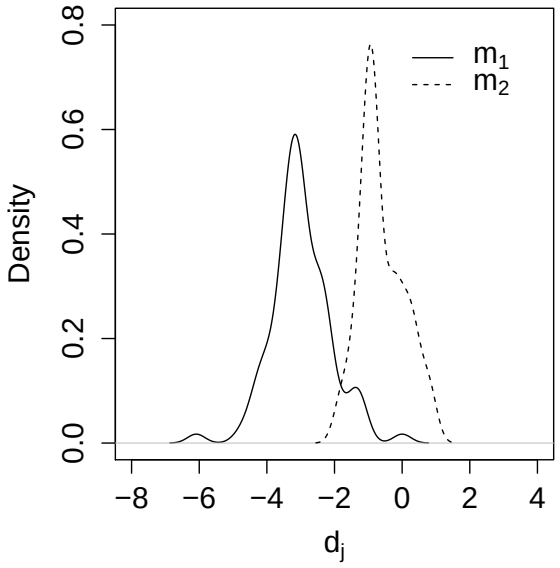


Figure 3.44: Posterior residual densities for the two models resulting from calculating (3.5.7) for each epidemic model based on the average of 500 curves generated from the conditioning predictive distribution of removal times for both models.

and 72.16 for m_1 and m_2 respectively. As expected, this indicates that the two levels of mixing model (m_2) offers a better fit to these data.

- **Posterior predictive $p-$value criterion:**

Another possible Bayesian tool for diagnosing the model fit is to calculate the posterior predictive $p-$value which can be estimated as

$$\begin{aligned} ppp-\text{value} &= P\left(D(\boldsymbol{R}^{rep},\boldsymbol{\phi}) \geq D(\boldsymbol{R}^{obs},\boldsymbol{\phi})\big|\boldsymbol{R}^{obs}\right) \\ &\approx \frac{1}{M'}\sum_{k=1}^{M'} \mathbb{1}_{\{D(\boldsymbol{R}^{rep\,(k)},\boldsymbol{\phi}^{(k)})\geq D(\boldsymbol{R}^{obs},\boldsymbol{\phi}^{(k)})|\boldsymbol{R}^{obs}\}}, \end{aligned} \tag{3.5.8}$$

where here for the discrepancy measure $D(\boldsymbol{R},\boldsymbol{\phi})$ we again used both the χ^2 and Freeman-Tukey discrepancy measures to obtain the $ppp-$values.

The $ppp-$value can be computed via Monte Carlo simulations as explained in Section 3.2.4 with $D = 100$. Here, we worked with both the χ^2 and Freeman-Tukey discrepancy measures of the overall goodness-of-fit that depends on data as well as the model parameters.

As the posterior variance of the first removal time, $\text{var}(R_1|\boldsymbol{\phi}^{(k)})$, is zero, where $k = 1, ..., M' = 500$ and $\boldsymbol{\phi}$ denotes the vector of each model parameters, we shall exclude R_1 when calculating $D(\boldsymbol{R}^{obs}, \boldsymbol{\phi}^{(k)})$ and $D(\boldsymbol{R}^{rep\ (k)}, \boldsymbol{\phi}^{(k)})$ and use R_j^{obs} and R_j^{rep} for $j = 2, ..., n_R$. Figure 3.45 shows the resultant scatter plots with the posterior predictive $p-$values. As we mentioned earlier, the closer the ppp-value to the optimal value 0.5 the better the fit. The ppp-values obtained from fitting the two levels of mixing model (m_2) to the observed data are closer to the optimal value compared to the ppp-values produced from fitting m_1 to the observed data.

Again, it is not unexpected that the ppp-values may take some effort to obtain in terms of computational run time, in particular if the selected M' posterior samples yield extreme values of R_0 or R_*.

- **Posterior distributions of number of household and community infections criterion:**

Another conceivable way to assess the goodness-of-fit for these two epidemic models is to infer the distributions of number of household and community infections using Algorithm 7 which can be performed as follows. At each MCMC iteration, we compute the probability that an infected individual j, has been infected from a member of their household, that is

$$\begin{aligned} A_j &= P(\text{individual } j \text{ has been infected from a member of their household}) \\ &= \frac{\lambda_H Y_i^H(I_j-)}{\lambda_H Y_i^H(I_j-) + \boldsymbol{N}^{-1}\lambda_C Y^C(I_j-)}, \end{aligned} \tag{3.5.9}$$

where $i = 1, 2, ..., h$ and $j = 1, ..., n_R, j \neq z$.

Then, the quantities $\sum_{j=1, j\neq z}^{n_R} A_j$ and $\sum_{j=1, j\neq z}^{n_R}(1 - A_j)$ provide us with approximate iteration-based estimates for number of household and community infections at each MCMC iteration. The posterior distributions of these quantities over MCMC iterations may be used to detect any heterogeneity in the observed data. Figure 3.46 shows sample traces for the number of household and community infections after a burn-in period of 2000 iterations and

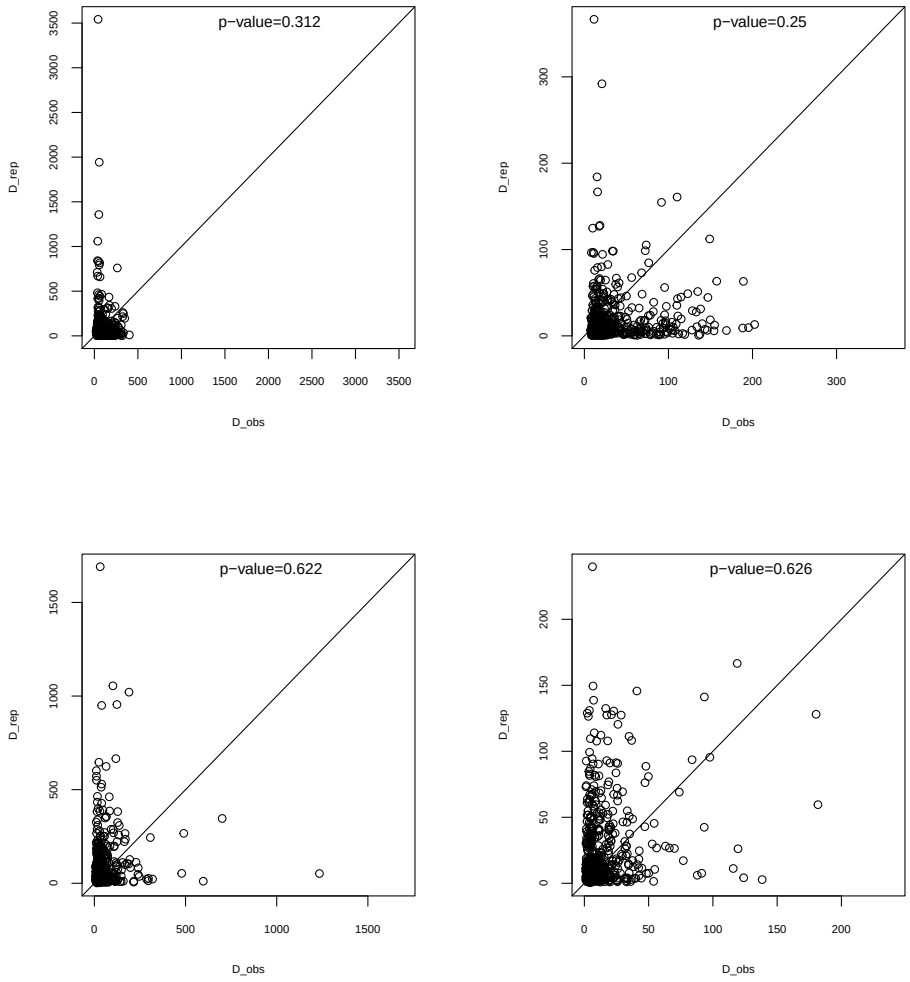


Figure 3.45: Scatter plots of observed versus replicated discrepancies based on 500 predicted removal curves from each removal times predictive distributions of the two models. The ppp-values are estimated by the proportion of points above the 45° line. Left and right columns correspond to the χ^2 and Freeman-Tukey discrepancy measures. Top and bottom rows refer to m_1 and m_2 respectively.

a thinning of 5 samples, (top row: left to right) under fitting the two levels of mixing model (m_2), also it shows histograms of the posterior distributions of the number of household and community infections (bottom row: left to right). These plots can be used to support the presence of heterogeneity (and consecutively to support m_2 over m_1) in the observed data as the approximate

posterior means for number of household and community infections are 71 and 18 respectively. Having observed a total final size of 89 in a community consists of 20 households of 5 individuals each, these estimated values are expected since there must be at least 18 between-household contacts in order to have 89 ultimate infected individuals.

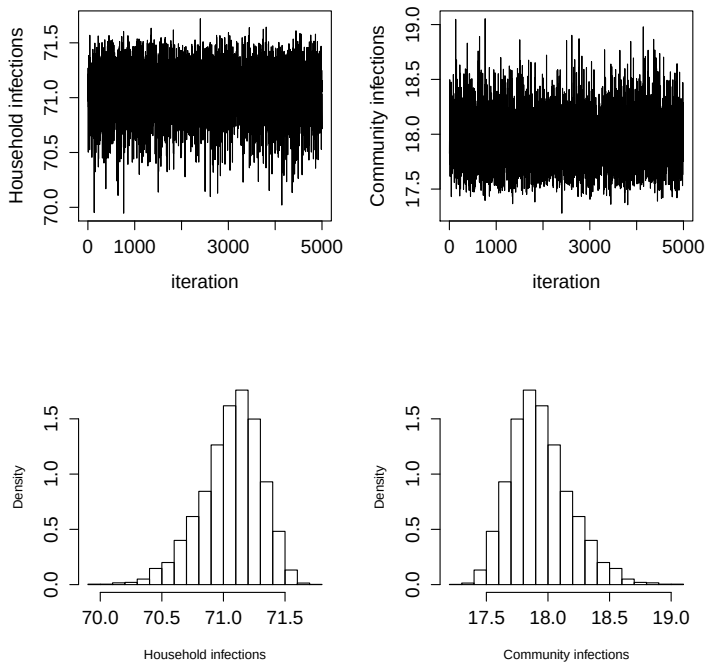


Figure 3.46: Posterior sample traces (top row) and histograms (bottom row) of the number of household and community infections (left to right) obtained by fitting m_2 to the observed data.

3.5.3.2 Sub-study II: Homogeneously mixing population removal data

Now, our removal dataset was simulated from a homogeneous mixing SIR model with parameters $\beta = 2.5$ and $c = 1$ in a population of size $N = 100$. The resulting data consisted of 90 observed removal times. As would be expected, both models m_1 (the data-generated homogeneous mixing model) and m_2 (the two levels of mixing model) should fit these data owing to the fact that m_1 is a special case of m_2 when each household is of size one individual.

Following the same procedure used in Section 3.5.3.1, we fitted both models

m_1 and m_2 to the observed data using MCMC methods. Figure 3.47 shows sample traces of the Markov chains for the m_2 infection rate parameters after a burn-in period of 2000 iterations and a thinning of 5 samples. Apparently, the within-household infection rate parameter λ_H has very low estimated value compared to the between-household infection rate parameter λ_C which might be interpreted as evidence of the homogeneity in the observed data.

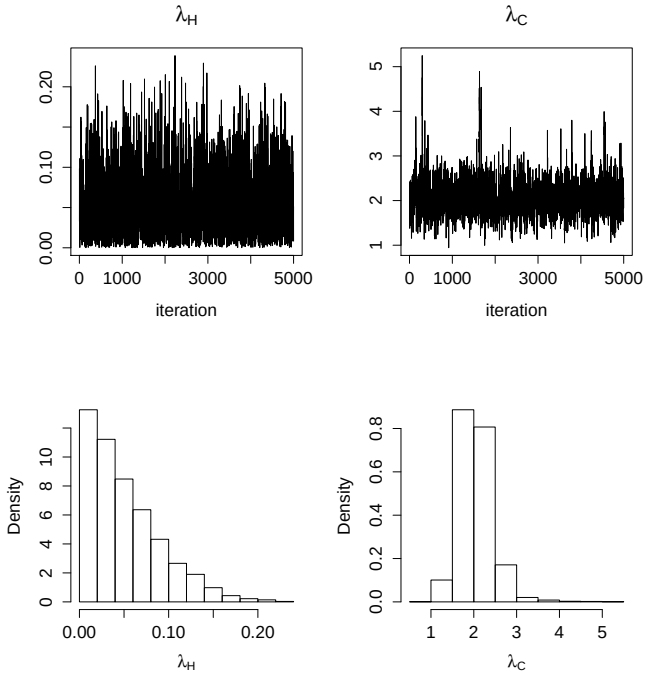


Figure 3.47: Posterior sample traces and histograms of λ_H and λ_C, obtained from fitting m_2 to the observed data, after a burn-in period of 2000 iterations and a thinning of 5 samples.

The posterior predictive distributions of removal times and final size are presented in Figure 3.48. By looking at these figures, it is evident that the observed removal curve may be seen as a realisation from either one of these models and the observed final size ($n_R - 1 = 89$) lies within both final size predictive distributions. Moreover, if we condition on the observed final size the resultant conditioning predicative distributions of removal curves are similar in the sense that the observed removal curve is consistently closed to the predictive mean (see Figure 3.49). More formally, we calculated the Bayesian residuals for each model using (3.5.7) and then we plotted their posterior densities as

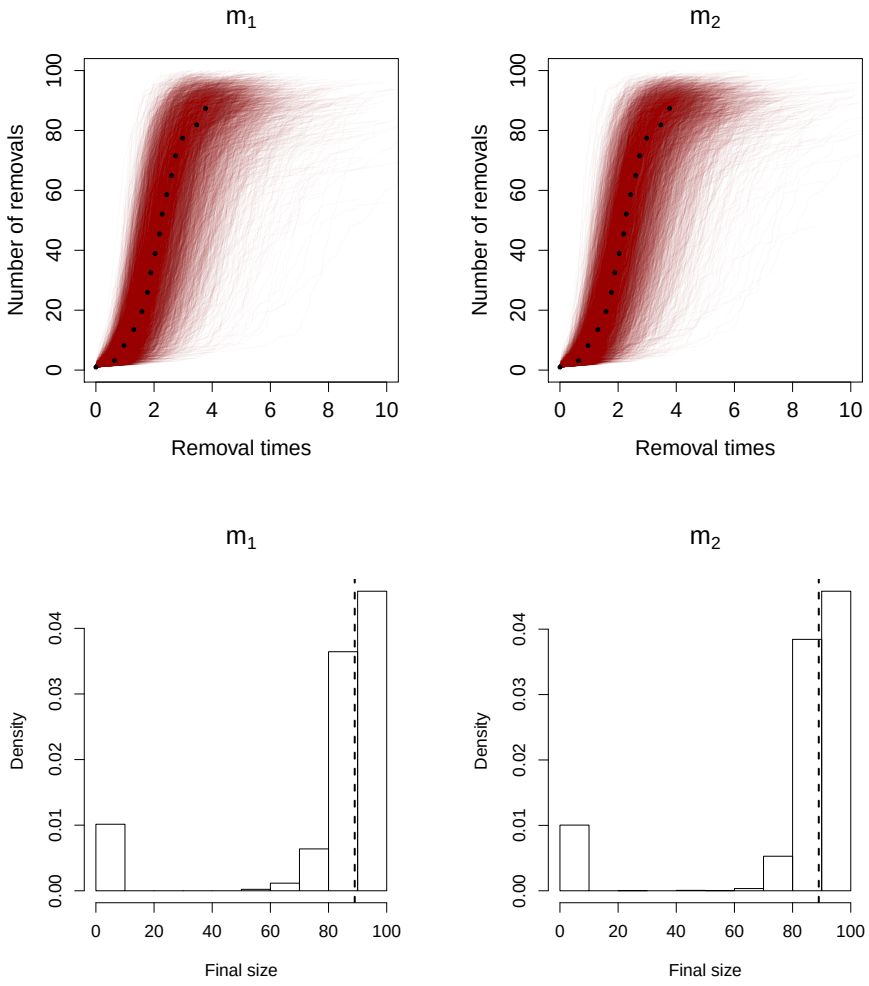


Figure 3.48: The removal times and final size posterior predictive distributions for the two models m_1 and m_2 (left to right), where the dotted and dashed lines indicate the observed removal curve and the observed final size respectively.

shown in Figure 3.50 in which the posterior densities resulting from fitting both models have commutated values near zero as an indication of a good fit for both models, with sum of the squared Bayesian residuals $\sum_{j=1}^{n_R} d_j^2$ of 5.49 and 1.11 for m_1 and m_2 respectively. It appears from both Figure 3.49 and Figure 3.50 that m_2 is slightly favoured (although it is not the data-generating model), which is not that surprising because m_1 is a special case of m_2 when each household is of size one individual.

Calculating the posterior predictive ppp-value by fitting the two models to

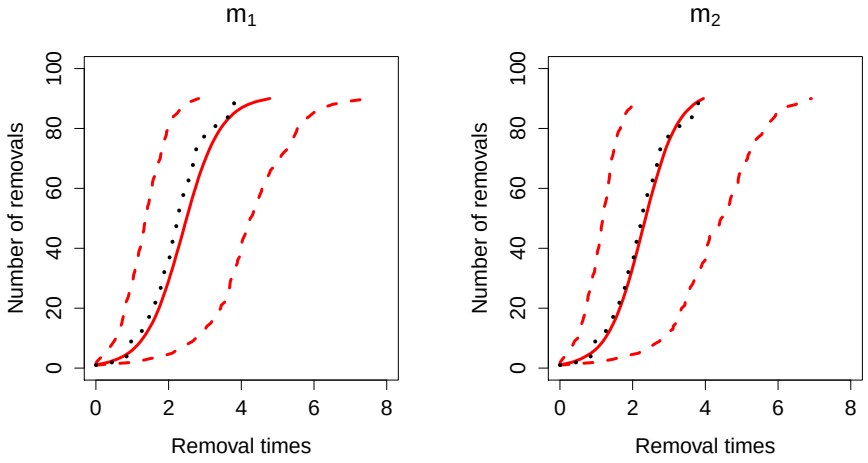


Figure 3.49: Predictive removal times mean (red solid line) with 95% percentile intervals (red dashed lines) and the observed removal curve (black dotted line) for the two models m_1 and m_2 (left to right).

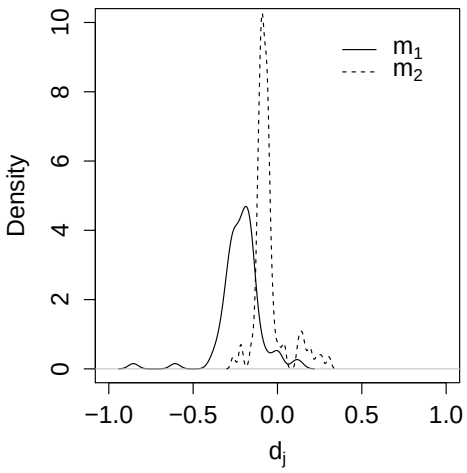


Figure 3.50: Posterior residual densities for the two models resulting from calculating (3.5.7) for each epidemic model based on the average of 500 curves generated from the conditioning predictive distribution of removal times for both models.

the observed removal curve was performed as in previous simulation study in Section 3.5.3.1 with same discrepancy measures. Figure 3.51 shows scatter plots of observed versus replicated of the χ^2 and Freeman-Tukey discrepancies based on 500 samples from the two removal times predictive distributions of the models along with the resulting ppp-values. The ppp-values produced

from both models are close enough to the optimal value 0.5, so the models are indistinguishable in the sense of judging about the model fit and both models fit the data according to ppp-value criterion.

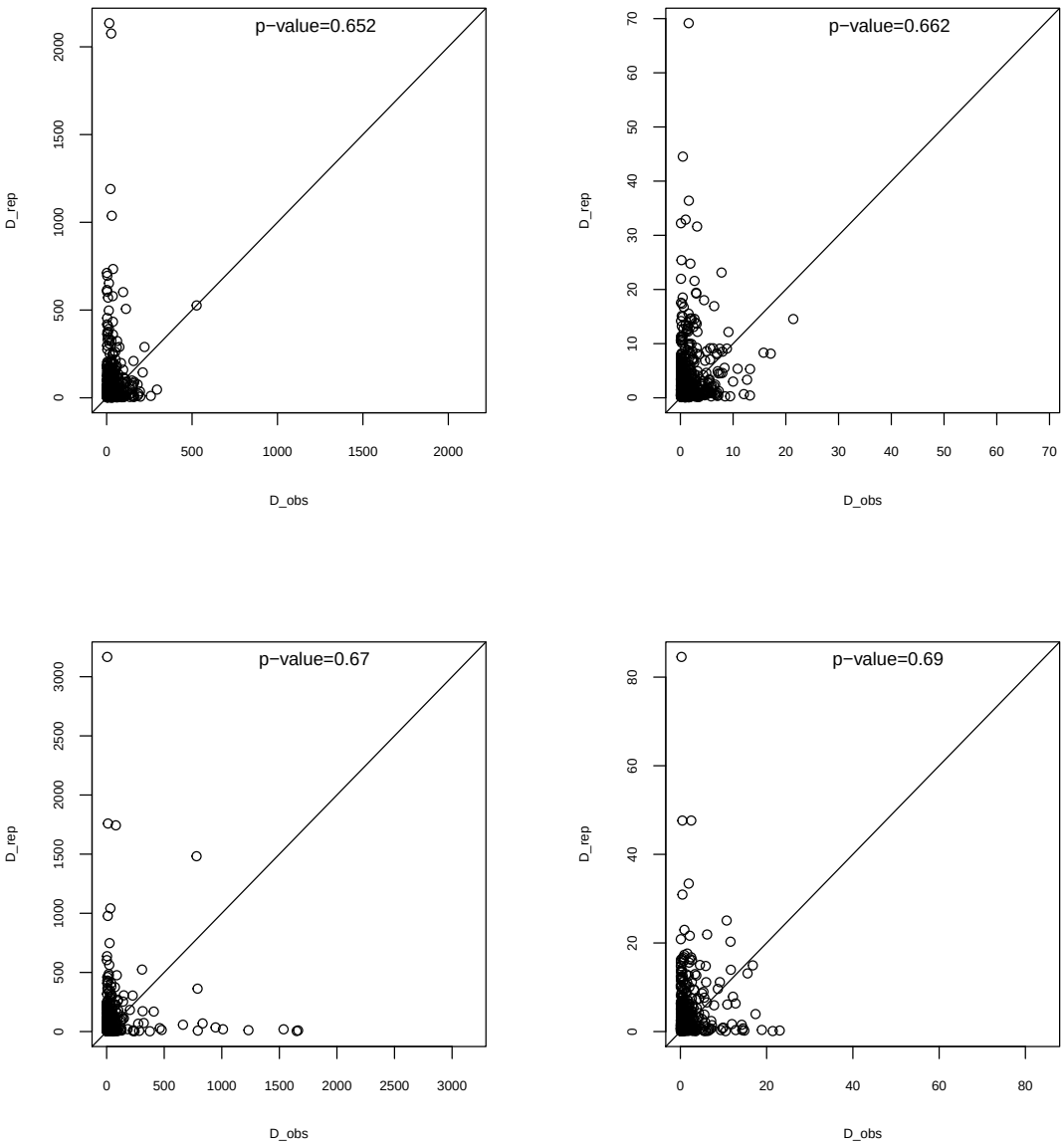


Figure 3.51: Scatter plots of observed versus replicated discrepancies of the χ^2 and Freeman-Tukey (left to right column) based on 500 predicted removal curves from each removal times predictive distributions of the two models. Top and bottom rows refer to m_1 and m_2 respectively.

Additionally and more interestingly, we inferred the posterior distributions of number of infections happened within households and between households under fitting m_2 model to these data as shown in Figure 3.52 in which it is apparent that there is less household structure and, by a way of summary, the posterior mean of number of household and community infections are 7 and 82 respectively. This observation also goes some way to explain the large difference in the posterior means of λ_H and λ_C as the within-household outbreak can only be mild resulting in a relatively low estimate of λ_H. These findings may be considered as evidence against any heterogeneity in the observed data.

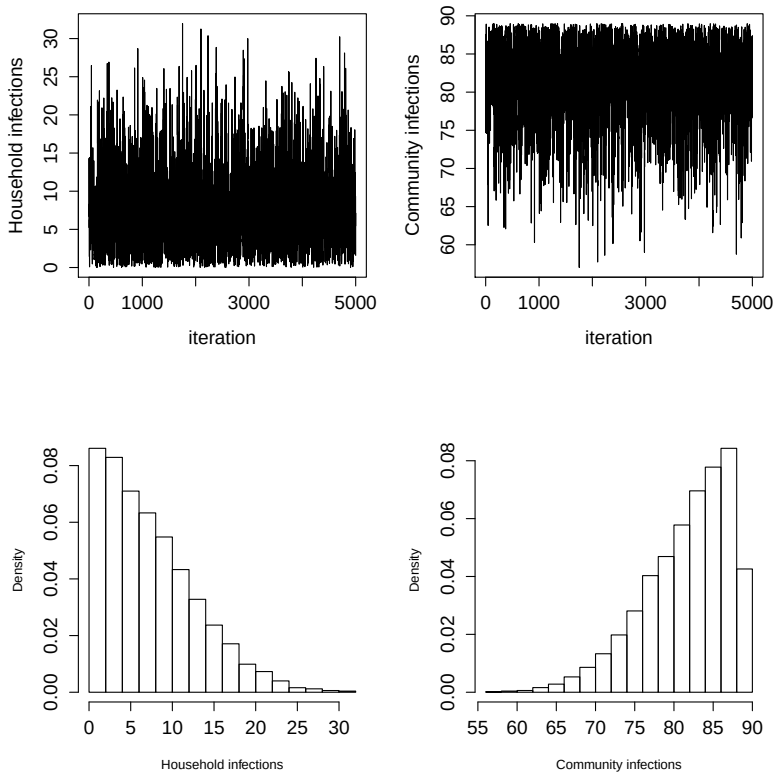


Figure 3.52: Posterior sample traces (top row) and histograms (bottom row) of the number of household and community infections (left to right) obtained by fitting m_2 to the observed data.

3.6 Summary and extensions

3.6.1 Summary

In this chapter, we have developed a novel method, under a Bayesian framework, to assess the goodness-of-fit for the SIR stochastic epidemic models. In our proposed two-stage method, two predictive distributions were examined, namely the predictive distribution of the final size of the epidemic and the predictive distribution of the removal curves. The method explored how SIR models with different infection mechanisms, infectious periods and population structures can be assessed and distinguished given only a set of removal times of infected individuals in the population. The idea was based on looking explicitly at the discrepancy between the observed and predicted removal data using the posterior predictive model checking procedure in which the notion of Bayesian residuals and the posterior predictive $p-$value were used.

Three different epidemic settings were investigated. In the first setting, we considered the SIR model with different infection mechanisms. We started by illustrating the difference, in terms of the removal curves distribution, between the Markovian SIR model with infection rate form of $\beta n^{-1}X(t)Y(t)$ and other Markovian SIR model with infection rate form of $\beta n^{-1}X(t)Y^p(t), 0 < p < 0.5$. Then, we simulated outbreak removal data according to each model and applied the proposed method in order to see how well it is in recovering the true data-generated model. Results indicated that our two-stage procedure works reasonably well using both Bayesian model assessment tools, namely the Bayesian residuals and the ppp-values. Moreover, sensitivity of the method to R_0 was checked via simulation. Although the ppp-values did not indicate any lack of fit for both models for the small outbreak dataset, it was found that our two-stage method is not too sensitive to the value of R_0 and it appears to work well in identifying the correct model.

Secondly, we used our proposed method to identify the infectious period distribution for the SIR models. We specifically considered three SIR mod-

els routinely employed in the literature, namely the model with exponential infectious periods ($T_I \sim \text{Exp}(\gamma)$), the model with gamma infectious periods ($T_I \sim \text{Gamma}(\alpha, \delta)$) and the model with fixed length infectious periods ($T_I \equiv c$). The removal curves distributions for all three models were examined, then our two-stage method was tested via two removal datasets simulated under the exponential and gamma SIR models as well as using a widely studied temporal removal dataset obtained from a smallpox outbreak. For the simulated datasets, results explored the capability of the proposed method to identify the correct model. Also, simulated results revealed that the gamma and constant SIR models appeared to be indistinguishable when the shape parameter of the former is large as assumed in our setting. For the smallpox outbreak dataset, we fitted to these data both SIR and SEIR models with different infectious periods and then our two-stage method was applied. According to our method, non of these epidemic models seems be appropriate to fit these smallpox data and no particular choice of infectious period distribution was favoured by the data.

In the last epidemic setting, model assessment task was performed for the SIR model with different population structures. The homogeneous (one level of mixing) SIR model with constant infectious period was compared to the heterogeneous (two levels of mixing) SIR model. Bayesian parameter estimation was developed for the two levels of mixing model given removal data only and the behaviour of removal trajectories for both models was investigated. By assuming that the disease is highly infectious within households such that infection of a single member of a household leads to the eventual infection of all susceptibles in the household, it was found that the distribution of removal curves for the two levels of mixing model has systematic jump behaviour of size equals to the household size. Three different Bayesian formal tools were employed to measure the goodness-of-fit for both models, namely the Bayesian residuals, the posterior predictive p-value and the posterior distributions of number of household and community infections. Having considered the performance of the proposed method by conducting two

simulation studies, it was found that the ppp-value and the Bayesian residual criteria seem to be successful in differentiating between SIR models. More interestingly, inferring the posterior distributions of number of household and community infections appeared to be an effective mechanism for detecting any heterogeneity in the data.

In terms of computational run time, we selected randomly $M' = 500$ posterior samples from the joint posterior distribution of each model parameters and for each posterior sample we kept simulating the model of interest until $D = 100$ removal curves, all with same final size as the observed one, are obtained in order to estimate $E(R_j|\boldsymbol{\phi}^{(k)})$ and $\text{var}(R_j|\boldsymbol{\phi}^{(k)})$ that are used to produce the Bayesian residuals d_j and the ppp-values. It is not surprising that such process may take some effort to implement in terms of computational time, particularly if some of the selected M' posterior samples produce extreme values of R_0 or R_*. Of course, one could use the whole posterior samples and simulate the model a large number of times to get more reliable results, but the process of obtaining the ppp-values and assessing the goodness-of-fit becomes impractical and time-consuming.

3.6.2 Extensions

Although the model (the SIR model) considered in this chapter is quite general, our proposed method can be extended to more complex situations, particularly with a view to increase the modelling realism. Extension to more complex epidemic models such as the SEIR model (as shown for smallpox data in Section 3.4) is straightforward.

Epidemics are assumed to be complete throughout this chapter. However, our method is also appropriate to ongoing disease outbreaks analysis. Furthermore, in Section 3.4 it was assumed that the shape parameter α for the gamma SIR model is known, however it would be interesting to see the performance of our proposed method in the case of unknown α.

In Section 3.5, it was assumed that the length of infectious periods of infected individuals is constant in both models (the homogeneous mixing model and the two levels of mixing model). It is natural to consider other choices of infectious period distribution. Finally, as the two levels of mixing SIR model includes the homogeneous mixing SIR model as a special case ($\lambda_H = 0$), one could discriminate these models even more by restricting the within-household infection parameter (λ_H) to a minimum value as we assumed for the power coefficient for the SIR model with non-linear infection rate introduced in Section 3.2.1.

CHAPTER 4

Bayesian Model Choice for Epidemic Models Based on Removal Data

4.1 Introduction

In this chapter, we consider the problem of choosing, within a Bayesian modelling framework, from a finite specified set of stochastic epidemic models given temporal removal data. Such selection is based upon two important Bayesian model selection methods, that are (i) Bayes factors via our missing-data extension of the power posterior method (Friel and Pettitt (2008) and Friel et al. (2014)), and (ii) the deviance information criterion for missing data models (DIC) (Celeux et al., 2006), with more emphasis on (i).

Based on complete outbreak data, explicit expressions for Bayes factors in two different epidemic settings are obtained, namely for the SIR model with different infectious periods and the SIR model with different infection mechanisms. The motivations for looking at an exact Bayes factor formula are that it is of interest in its own right and it may provide us some insights to gain an accurate and deeper understanding of this Bayesian model choice tool. In addition, we can validate the estimate of the Bayes factor derived from the prior arithmetic

mean estimator (Kass and Raftery, 1995) against the Bayes factor true value in order to see the reliability of the prior arithmetic mean estimator. This is useful because we may use it in other epidemic settings where Bayes factors cannot be evaluated analytically.

In spite of the generality of the power posterior approach (Friel and Pettitt (2008) and Friel et al. (2014)), there does not appear to have been any attempt to use it in literature of model choice for stochastic epidemic models. We extend this approach into models incorporating missing data such as epidemic models. Systematic studies are performed to investigate the effect of several factors on the outcome of the two proposed methods using simulated outbreak data. These factors include specification of prior distributions and the amount of discretization needed for the temperature parameter t in the interval $[0, 1]$ for the power posterior approach. The impact of the epidemic outbreak size on differentiating between epidemic models using Bayes factor criterion via the power posterior method is also investigated.

The chapter is organised as follows. In Section 4.2, we provide an overview of the marginal likelihood and the Bayes factor and discuss the calculation and interpretation of the Bayes factor. We then describe the deviance information criterion (DIC) that can be used to compare models, in particular a suitable form of DIC relevant to models with missing data. After that, we review previous work on model selection for epidemic models. The use of reversible jump Markov chain Monte Carlo (RJMCMC), and the deviance information criterion (DIC) with different types of epidemic data is discussed. Two different approaches to estimate the marginal likelihood, namely the prior arithmetic mean estimator and the power posterior estimator are then outlined. Lastly, our novel missing data-extension of the power posterior approach is presented along with the aims of the current work and the epidemic models that we are interested in comparing.

Deriving explicit forms for Bayes factors in different epidemic settings and approximating the Bayes factor for epidemic models using the prior arithmetic

mean estimator in case of complete outbreak data are described in Section 4.3. Application to simulated epidemic data is also provided. Section 4.4 illustrates the use of power posterior and DIC methods where only incomplete outbreak data are available and also provides the application of these two approaches to the SIR model with different infectious periods and the SIR model with different infection mechanisms. Here, the power posterior densities for the parameters of each model are derived, simulation studies are performed to assess the performance of the two methods against several factors, then our findings are reported. In Section 4.5 various simulation studies are performed on a particular epidemic model selection aspect that is the effect of the outbreak size on differentiating between epidemic models via Bayes factors. Section 4.6 concludes the chapter with a summary of our results. We highlight the benefits and drawbacks of both methods, and under which conditions we might expect reasonable results.

4.2 Background and previous work

4.2.1 The marginal likelihood and Bayes factor

Suppose we are given data $\boldsymbol{y}$ assumed to have arisen under one of two models m_1 and m_2 each with associated parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ respectively. Then, the Bayes factor (Kass and Raftery, 1995) is the ratio of the marginal likelihood (evidence) terms for the two competing models, that is

$$BF_{12} = \frac{\pi(\boldsymbol{y}|m_1)}{\pi(\boldsymbol{y}|m_2)} = \frac{\int_{\boldsymbol{\theta}_1} \pi(\boldsymbol{y}|\boldsymbol{\theta}_1, m_1)\pi(\boldsymbol{\theta}_1|m_1)d\boldsymbol{\theta}_1}{\int_{\boldsymbol{\theta}_2} \pi(\boldsymbol{y}|\boldsymbol{\theta}_2, m_2)\pi(\boldsymbol{\theta}_2|m_2)d\boldsymbol{\theta}_2}, \tag{4.2.1}$$

where $\pi(\boldsymbol{y}|\boldsymbol{\theta}_i, m_i)$ is the likelihood of the data under model m_i and $\pi(\boldsymbol{\theta}_i|m_i)$ is the prior density of parameters $\boldsymbol{\theta}_i$ for model m_i , $i = 1, 2$. The larger BF_{12} is, the greater the evidence in favour of m_1 compared to m_2. Following Jeffreys (1998), Kass and Raftery (1995) proposed a rule to interpret the strength of evidence according to BF_{12} on the log scale as shown in Table 4.1. Although these rules are given in terms of evidence against m_2, the same

rules can be used to interpret the evidence against m_1 by inverting the value of the logarithm of BF_{12}. Of course one could use $BF_{21} = 1/BF_{12}$ but then $\log(BF_{12}) \in (a, b)$ becomes $\log(BF_{21}) \in (-b, -a)$. In other words, negative values of log-Bayes factor are handled by symmetry and changed wording, in an obvious way.

$\log(BF_{12}) \in$	Evidence against m_2
$(0, 0.5)$	Not worth more than a bare mention
$(0.5, 1)$	Substantial
$(1, 2)$	Strong
$(2, \infty)$	Decisive

Table 4.1: Interpretation of the Bayes factor on the log scale.

4.2.2 The deviance information criterion (DIC)

The deviance information criterion (DIC) method (Spiegelhalter et al., 2002) is a model comparison tool that is appropriate for complex Bayesian models, and since its development a number of variants have been introduced, including those for its application to missing data models (Celeux et al., 2006). One advantage of using this method is that it is often easily calculated for models fitted using Markov chain Monte Carlo methods.

For a full description of the development of the DIC, interested readers are referred to Spiegelhalter et al. (2002). Briefly, measures of model complexity and fit are used to form the criterion upon which model comparisons can be made, where the model with smallest DIC value corresponds to the best model.

First, a measure of fit can be represented via the posterior deviance which is defined as

$$D(\boldsymbol{\theta}) = -2\log(\pi(\boldsymbol{y}|\boldsymbol{\theta})) \tag{4.2.2}$$

The expected deviance $E_{\boldsymbol{\theta}|\boldsymbol{y}}[D(\boldsymbol{\theta})]$ may be estimated by drawing samples from

the posterior density $\pi(\boldsymbol{\theta}|\boldsymbol{y})$, that is

$$E_{\boldsymbol{\theta}|\boldsymbol{y}}[D(\boldsymbol{\theta})] \approx \frac{1}{M}\sum_{j=1}^{M} D(\boldsymbol{\theta}^{(j)}), \tag{4.2.3}$$

where M is the posterior sample size and $\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(M)}$ are posterior samples.

Second, a complexity measure or a measure of the effective number of parameters in a model can be defined as

$$p_D = E_{\boldsymbol{\theta}|\boldsymbol{y}}[D(\boldsymbol{\theta})] - D(\widehat{\boldsymbol{\theta}}), \tag{4.2.4}$$

where $\widehat{\boldsymbol{\theta}}$ can be chosen as the posterior mean or median of $\boldsymbol{\theta}$.

Then, the corresponding DIC is given as

$$\begin{aligned} \text{DIC} &= E_{\boldsymbol{\theta}|\boldsymbol{y}}[D(\boldsymbol{\theta})] + p_D \\ &= 2E_{\boldsymbol{\theta}|\boldsymbol{y}}[D(\boldsymbol{\theta})] - D(\widehat{\boldsymbol{\theta}}) \\ &= -4E_{\boldsymbol{\theta}|\boldsymbol{y}}[\log(\pi(\boldsymbol{y}|\boldsymbol{\theta}))] + 2\log(\pi(\boldsymbol{y}|\widehat{\boldsymbol{\theta}})). \end{aligned} \tag{4.2.5}$$

4.2.2.1 DIC for models with missing data

Let $\boldsymbol{y}$ and $\boldsymbol{x}$ denote the observed and the missing data respectively with $\boldsymbol{\theta}$ representing the model parameters. If the observed likelihood $\pi(\boldsymbol{y}|\boldsymbol{\theta})$ is not tractable but the complete likelihood $\pi(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta})$ is tractable, then the posterior deviance (4.2.2) can be redefined as

$$D(\boldsymbol{\theta}) = -2\log(\pi(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta})). \tag{4.2.6}$$

Based on the role of the missing data $\boldsymbol{x}$ in the models, Celeux et al. (2006) proposed various forms of DIC. For instance, if $\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$ is available in a closed form (which is not the case in our settings), one may use DIC_7or DIC_8 as proposed in Celeux et al. (2006). Also, DIC_5 in Celeux et al. (2006) requires the computation of a posterior expectation for each value of $\boldsymbol{x}$. Here, $\pi(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta})$ is tractable as well as the missing data in our epidemic setting are not the point of interest but used to make the likelihood tractable, so a particular form of DIC (Celeux et al., 2006) is considered, namely

$$\text{DIC}_6 = -4E_{\boldsymbol{\theta},\boldsymbol{x}|\boldsymbol{y}}[\log(\pi(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta}))] + 2E_{\boldsymbol{x}|\boldsymbol{y},\widehat{\boldsymbol{\theta}}}[\log(\pi(\boldsymbol{y}, \boldsymbol{x}|\widehat{\boldsymbol{\theta}}))]. \tag{4.2.7}$$

In principle, the calculation of this measure requires running an MCMC scheme twice. The first run is to derive $E_{\boldsymbol{\theta},\boldsymbol{x}|\boldsymbol{y}}[\log(\pi(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta}))]$ and $E_{\boldsymbol{\theta}|\boldsymbol{y}}(\boldsymbol{\theta}|\boldsymbol{y}) = \widehat{\boldsymbol{\theta}}$, and the second run is to obtain $E_{\boldsymbol{x}|\boldsymbol{y},\widehat{\boldsymbol{\theta}}}[\log(\pi(\boldsymbol{y},\boldsymbol{x}|\widehat{\boldsymbol{\theta}}))]$ setting $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ and allowing $\boldsymbol{x}$ to vary. However, as we compare the performance of this measure to the Bayes factor obtained via the power posterior approach, the quantities in the first step of calculating DIC_6 are readily available from the power posterior at temperature $t = 1$ at virtually no extra computational cost. Therefore, we need to run the MCMC scheme once only.

4.2.3 Previous work on model choice for epidemic models

Bayesian model selection tools have been applied to epidemic models with final size data and temporal data in certain settings. The reversible jump Markov chain Monte Carlo (RJMCMC) method (Green, 1995) and the deviance information criterion (DIC) method (Celeux et al., 2006) are two standard approaches to Bayesian model choice, which have been used to compare different infectious disease models.

4.2.3.1 RJMCMC

The RJMCMC method (Green, 1995) is an extension of the Metropolis-Hastings algorithm. It allows us to construct a Markov chain with stationary distribution equal to the joint posterior distribution of both models and parameters. Therefore, an RJMCMC algorithm requires jump mechanisms to move between the competing models and their associated parameters. In other words, an RJMCMC algorithm involves two steps within each iteration of the Markov chain that are: (*i*) updating the parameters conditional on the current model using the Metropolis-Hastings algorithm; and (*ii*) updating the model conditional on the current parameter values using a carefully and sensibly chosen jump function. For the model selection purpose, the posterior model probabilities can be simply estimated as the proportion of time that the con-

structed Markov chain remains in any given model. These posterior model probabilities can be combined with prior model probabilities to estimate the Bayes factor as the ratio of posterior odds to prior odds.

To illustrate the RJMCMC further, we focus on comparing just two models of differing dimension m_1 and m_2 with associated parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ respectively. Suppose that the current state of our chain is $(\boldsymbol{\theta}_1, m_1)$ and we propose to move to $(\boldsymbol{\theta}_2, m_2)$. To achieve that we generate a random vector $\boldsymbol{u}$ from a predetermined probability density g (note that $\boldsymbol{u}^*$ and g^* analogously denote a random vector and probability density for the reverse move) and set $(\boldsymbol{\theta}_2, m_2) = h(\boldsymbol{\theta}_1, m_1, \boldsymbol{u})$, where h is some deterministic function. This function must be a bijective function with differentiable inverse. Then, the proposed move is accepted with probability

$$\min\left\{\frac{\pi(\boldsymbol{y}|\boldsymbol{\theta}_2, m_2)\pi(\boldsymbol{\theta}_2|m_2)\pi(m_2)\pi(m_2 \to m_1)g^*(\boldsymbol{u}^*)}{\pi(\boldsymbol{y}|\boldsymbol{\theta}_1, m_1)\pi(\boldsymbol{\theta}_1|m_1)\pi(m_1)\pi(m_1 \to m_2)g(\boldsymbol{u})}|J|, 1\right\}, \tag{4.2.8}$$

where J is the Jacobian resulting from the transformation from $(\boldsymbol{\theta}_1, \boldsymbol{u}, m_1)$ to $(\boldsymbol{\theta}_2, \boldsymbol{u}^*, m_2)$, $\pi(m_1 \to m_2)$ denotes the probability of proposing to move from model m_1 to m_2 and $\pi(m_i)$ represents the pre-assigned prior probability of the model m_i, $i = 1, 2$.

In the case of temporal outbreak data, Neal and Roberts (2004) analysed a highly detailed historical measles data set using an RJMCMC method. The authors employed an RJMCMC scheme to move between five models, a full model and four nested models each of which was identical to the full model with one parameter set to zero. The rationale behind this selection is to identify which of spatial location, households and classrooms are important in the transmission of measles throughout the population. An RJMCMC method was also used by O'Neill and Marks (2005) to move between a full model with a vomiting parameter included and a sub-model without this parameter in order to investigate whether or not the spread of the Norovirus virus is enhanced when a child vomits in a classroom.

Given data on final outcome of outbreak, Knock and O'Neill (2014) recently compared an RJMCMC algorithm and a path sampling-based algorithm (Gel-

man and Meng, 1998) for the estimation of Bayes factors given data of outbreak among a population of individuals which is partitioned into households. The authors studied three models, the two levels of mixing model (Ball et al., 1997) as the full model and two special cases of the full model, aiming on assessing the extent to which the household structure is important in disease transmission.

However, the computational burden associated with the RJMCMC method and defining how the Markov chain jumps between model parameter spaces are non-trivial challenges in implementing it in practice. For instance, as in our epidemic settings, when comparing the SIR model with different infectious periods or different infection mechanisms, it is not immediately obvious, when performing a model choice task for these epidemic models, how to design efficient jump functions for an RJMCMC algorithm that transform parameters between models and in the best case scenario, the RJMCMC algorithm may require large amounts of tuning (as in Neal and Roberts (2004)) to make it work well.

4.2.3.2 DIC

In the stochastic epidemic context, Worby (2013) investigated a version of DIC for missing data models (DIC_6 in Celeux et al. (2006)), and evaluated its performance in choosing between a specific set of nested transmission models using real and simulated hospital data. Another suitable form of DIC (DIC_4 in Celeux et al. (2006)) was implemented using final size data by Knock and O'Neill (2014) to select between epidemic models. Further, Lau et al. (2014) compared two forms of DIC (DIC_4 and DIC_8 in Celeux et al. (2006)) with the latent residuals approach in order to test and compare spatio-temporal models by misspecifying the spatial transmission kernel in the models.

The ease of calculating DIC is one important reason for using it as a model selection tool, particularly when the set of models to compare is large. However, DIC is not uniquely defined when dealing with partially observed processes

(Celeux et al., 2006) which makes it hard to interpret and less reliable.

Recently, Zhang (2014) used the probability scoring rule (see, e.g. Matheson and Winkler (1976); Winkler et al. (1996)) to carry out epidemic model comparison, based on the posterior predictive distributions obtained from the MCMC output, given both the infection and removal times. The modelling framework used was that of the discrete time individual-level stochastic epidemic model described in Deardon et al. (2010). Scoring rules work by rewarding a numerical value to a model based on the predictive distribution and on the infection event or the non-infection event observed, where the model with higher score is favoured. However, the manner of model selection in Zhang (2014) was somewhat unrealistic, in that it was assumed that the latent and the infectious period for each individual are constant which is not the case in our epidemic setting.

In Section 4.2.1, it is clear that estimating the Bayes factor relies on calculating the marginal likelihood which is difficult to evaluate in general, especially when the dimension of $\boldsymbol{\theta}_i$ is large. Therefore, we review here two different approaches to estimate $\pi(\boldsymbol{y}|m_i)$, namely the prior arithmetic mean estimator (Kass and Raftery, 1995), and the power posterior estimator (Friel and Pettitt (2008) and Friel et al. (2014)).

4.2.4 Marginal likelihood estimation via the arithmetic mean method

Dropping the notational dependence on m_i, the marginal likelihood becomes

$$\pi(\boldsymbol{y}) = \int_{\boldsymbol{\theta}} \pi(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{4.2.9}$$

The prior arithmetic mean estimate for $\pi(\boldsymbol{y})$ can be obtained by a simple Monte Carlo integration method, that is

$$\widehat{\pi(\boldsymbol{y})} = \frac{1}{M}\sum_{j=1}^{M} \pi(\boldsymbol{y}|\boldsymbol{\theta}^{(j)}), \tag{4.2.10}$$

where $\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(M)}$ is a sample generated from the prior $\pi(\boldsymbol{\theta})$.

It is well-known (see, e.g. Kass and Raftery (1995)) that since the likelihood is typically sharply peaked, most of the prior samples $\boldsymbol{\theta}^{(j)}$ have very small likelihood values, in particular when the prior is diffuse. Therefore, unless M is very large, the prior samples will contain virtually no points from the high-likelihood region, leading to poor and inadequate estimation of the marginal likelihood $\pi(\boldsymbol{y})$. However, this estimate, $\widehat{\pi(\boldsymbol{y})}$, might perform well when the dimensionality of the parameter space is modest, so that enormous prior samples can be generated very quickly.

4.2.5 Marginal likelihood estimation via the power posterior method

Friel and Pettitt (2008) suggested an approach for estimating the marginal likelihood based on the idea of thermodynamic integration (Gelman and Meng, 1998). The authors introduce an auxiliary variable (temperature parameter) $t \in [0, 1]$ and define the power posterior as

$$\pi_t(\boldsymbol{\theta}|\boldsymbol{y}) \propto \pi(\boldsymbol{y}|\boldsymbol{\theta})^t \pi(\boldsymbol{\theta}), \tag{4.2.11}$$

where by construction the normalizing constant of the power posterior is

$$z_t(\boldsymbol{y}) = \int_{\boldsymbol{\theta}} \pi(\boldsymbol{y}|\boldsymbol{\theta})^t \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{4.2.12}$$

Clearly, as t moves from $t = 0$ to $t = 1$, the power posterior follows a path from the prior to the posterior density, where $z_{t=0}(\boldsymbol{y}) = 1$ is the integral of the proper prior of $\boldsymbol{\theta}$ and $z_{t=1}(\boldsymbol{y})$ is the marginal likelihood. It can be shown that

$$\log(\pi(\boldsymbol{y})) = \log\left[\frac{z_{t=1}(\boldsymbol{y})}{z_{t=0}(\boldsymbol{y})}\right] = \int_0^1 E_{\boldsymbol{\theta}|\boldsymbol{y},t} \log\left[\pi(\boldsymbol{y}|\boldsymbol{\theta})\right] dt. \tag{4.2.13}$$

By discretizing t such that $0 = t_0 < t_1 < ... < t_{r-1} < t_r = 1$, a sample from $\pi_{t_j}(\boldsymbol{\theta}|\boldsymbol{y})$ can be used to estimate $E_{\boldsymbol{\theta}|\boldsymbol{y},t_j} \log\left[\pi(\boldsymbol{y}|\boldsymbol{\theta})\right]$ for each $t_j,\ j = 0, ..., r$. Then, a trapezoidal rule can be used to approximate the marginal likelihood,

that is

$$\log(\pi(\boldsymbol{y})) \approx \sum_{j=1}^{r} \frac{1}{2}(t_j - t_{j-1}) \left\{ E_{\boldsymbol{\theta}|\boldsymbol{y},t_j} \log\left[\pi(\boldsymbol{y}|\boldsymbol{\theta})\right] + E_{\boldsymbol{\theta}|\boldsymbol{y},t_{j-1}} \log\left[\pi(\boldsymbol{y}|\boldsymbol{\theta})\right] \right\}. \tag{4.2.14}$$

We will refer to the algorithm incorporating this estimate as the standard power posterior method.

Discretising t leads to a numerical integration error in (4.2.14), therefore Friel et al. (2014) recently refined the standard power posterior approach by subtracting a correction term based on the variance of $\log\left[\pi(\boldsymbol{y}|\boldsymbol{\theta})\right]$ with respect to the power posterior in order to reduce the bias in the estimate of the marginal likelihood (this will be clarified in Section 4.2.6). The updated power posterior estimate of $\log(\pi(\boldsymbol{y}))$ becomes

$$\begin{aligned} \log(\pi(\boldsymbol{y})) \approx & \sum_{j=1}^{r} \frac{1}{2}(t_j - t_{j-1}) \\ & \times \left\{ E_{\boldsymbol{\theta}|\boldsymbol{y},t_j} \log\left[\pi(\boldsymbol{y}|\boldsymbol{\theta})\right] + E_{\boldsymbol{\theta}|\boldsymbol{y},t_{j-1}} \log\left[\pi(\boldsymbol{y}|\boldsymbol{\theta})\right] \right\} \\ & - \sum_{j=1}^{r} \frac{1}{12}(t_j - t_{j-1})^2 \\ & \times \left\{ V_{\boldsymbol{\theta}|\boldsymbol{y},t_j} \log\left[\pi(\boldsymbol{y}|\boldsymbol{\theta})\right] - V_{\boldsymbol{\theta}|\boldsymbol{y},t_{j-1}} \log\left[\pi(\boldsymbol{y}|\boldsymbol{\theta})\right] \right\}, \end{aligned} \tag{4.2.15}$$

where $V_{\boldsymbol{\theta}|\boldsymbol{y},t} \log\left[\pi(\boldsymbol{y}|\boldsymbol{\theta})\right]$ denotes the variance of $\log\left[\pi(\boldsymbol{y}|\boldsymbol{\theta})\right]$ at t. This updated version of the power posterior approach will be explained in more detail for models incorporating missing data as below and we will refer to the algorithm involving this estimate as the updated power posterior method.

Both the standard and updated power posterior methods are applicable when $\pi(\boldsymbol{y}|\boldsymbol{\theta})$, the likelihood of the observed data given the model parameters, is tractable. However, in our epidemic settings, infection times are not observed, and therefore the likelihood of the observed data (removal times) given the model parameters is intractable and hence the standard and updated power posterior methods cannot be applied straightforwardly.

In what follows we extend the power posterior approach (Friel and Pettitt (2008), and Friel et al. (2014)) into models incorporating missing data. To our knowledge, there are not any other studies in the literature that develop the power posterior approach for this kind of missing data.

4.2.6 Power posteriors for models incorporating missing data

Missing data problems occur often in epidemic modelling as we have seen in the previous chapters. In a Bayesian framework, these problems are typically overcome by introducing auxiliary variables corresponding to the missing observations.

As introduced earlier, $\boldsymbol{y}$ and $\boldsymbol{x}$ denote the observed and the missing data respectively with $\boldsymbol{\theta}$ representing the model parameters. The pair $(\boldsymbol{y}, \boldsymbol{x})$ is often referred to as the complete data. Note that in the epidemic settings that we will consider, neither $\pi(\boldsymbol{y}|\boldsymbol{\theta})$ nor $\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$ is accessible but the observed data $\boldsymbol{y}$ may be augmented with the unobserved (missing) data $\boldsymbol{x}$ to form a tractable augmented likelihood function $\pi(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta})$.

The power posterior of missing data models can be defined as

$$\pi_t(\boldsymbol{\theta}, \boldsymbol{x}|\boldsymbol{y}) \propto \pi(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta})^t \pi(\boldsymbol{\theta}), \tag{4.2.16}$$

with the normalizing constant

$$z_t(\boldsymbol{y}) = \int_{\boldsymbol{x}} \int_{\boldsymbol{\theta}} \pi(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta})^t \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}\, d\boldsymbol{x}. \tag{4.2.17}$$

Thus,

$$\log(\pi(\boldsymbol{y})) = \log\left[\frac{z_{t=1}(\boldsymbol{y})}{z_{t=0}(\boldsymbol{y})}\right] = \int_0^1 E_{\boldsymbol{\theta},\boldsymbol{x}|\boldsymbol{y},t} \log\left[\pi(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta})\right] dt. \tag{4.2.18}$$

This missing data identity can be shown to hold as follows.

$$\begin{aligned}
\frac{d}{dt}\log(z_t(\boldsymbol{y})) &= \frac{1}{z_t(\boldsymbol{y})}\frac{d}{dt}z_t(\boldsymbol{y}) \\
&= \frac{1}{z_t(\boldsymbol{y})}\int_{\boldsymbol{x}}\int_{\boldsymbol{\theta}}\frac{d}{dt}\pi(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})^t\pi(\boldsymbol{\theta})d\boldsymbol{\theta}\,d\boldsymbol{x} \\
&= \frac{1}{z_t(\boldsymbol{y})}\int_{\boldsymbol{x}}\int_{\boldsymbol{\theta}}\pi(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})^t\log\left[\pi(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})\right]\pi(\boldsymbol{\theta})d\boldsymbol{\theta}\,d\boldsymbol{x} \\
&= \int_{\boldsymbol{x}}\int_{\boldsymbol{\theta}}\frac{\pi(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})^t\pi(\boldsymbol{\theta})}{z_t(\boldsymbol{y})}\log\left[\pi(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})\right]d\boldsymbol{\theta}\,d\boldsymbol{x} \\
&= \int_{\boldsymbol{x}}\int_{\boldsymbol{\theta}}\log\left[\pi(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})\right]\pi_t(\boldsymbol{\theta},\boldsymbol{x}|\boldsymbol{y})d\boldsymbol{\theta}\,d\boldsymbol{x} \\
&= E_{\boldsymbol{\theta},\boldsymbol{x}|\boldsymbol{y},t}\log\left[\pi(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})\right].
\end{aligned}$$

The second line in the proof above requires some regularity condition namely that the differentiation and integration can be interchanged.

By integrating with respect to $t \in [0,1]$, we have

$$\begin{aligned}
\log(\pi(\boldsymbol{y})) &= \log(z_{t=1}(\boldsymbol{y})) - \log(z_{t=0}(\boldsymbol{y})) \\
&= \log(z_{t=1}(\boldsymbol{y})) \\
&= \int_0^1 E_{\boldsymbol{\theta},\boldsymbol{x}|\boldsymbol{y},t}\log\left[\pi(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})\right]dt.
\end{aligned}$$

Now, to reduce the approximation error introduced by discretising t, we shall make use of the fact that the gradient of the expected log-likelihood curve equals its variance.

By differentiating $E_{\boldsymbol{\theta},\boldsymbol{x}|\boldsymbol{y},t}\log\left[\pi(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})\right]$ with respect to t, we obtain

$$\begin{aligned}
\frac{d}{dt}E_{\boldsymbol{\theta},\boldsymbol{x}|\boldsymbol{y},t}\log\left[\pi(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})\right] &= \frac{d}{dt}\int_{\boldsymbol{x}}\int_{\boldsymbol{\theta}}\log\left[\pi(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})\right]\pi_t(\boldsymbol{\theta},\boldsymbol{x}|\boldsymbol{y})d\boldsymbol{\theta}\,d\boldsymbol{x} \\
&= \int_{\boldsymbol{x}}\int_{\boldsymbol{\theta}}\log\left[\pi(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})\right]\frac{d}{dt}\pi_t(\boldsymbol{\theta},\boldsymbol{x}|\boldsymbol{y})d\boldsymbol{\theta}\,d\boldsymbol{x},
\end{aligned}$$

where

$$\begin{aligned}
\frac{d}{dt}\pi_t(\boldsymbol{\theta},\boldsymbol{x}|\boldsymbol{y}) &= \frac{d}{dt}\left[\frac{\pi(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})^t\pi(\boldsymbol{\theta})}{z_t(\boldsymbol{y})}\right] \\
&= \frac{z_t(\boldsymbol{y})\pi(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})^t\pi(\boldsymbol{\theta})\log\left[\pi(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})\right] - \pi(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})^t\pi(\boldsymbol{\theta})\frac{d}{dt}z_t(\boldsymbol{y})}{z_t^2(\boldsymbol{y})} \\
&= \frac{\pi(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})^t\pi(\boldsymbol{\theta})}{z_t(\boldsymbol{y})}\left[\log\left[\pi(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})\right] - \frac{1}{z_t(\boldsymbol{y})}\frac{d}{dt}z_t(\boldsymbol{y})\right] \\
&= \pi_t(\boldsymbol{\theta},\boldsymbol{x}|\boldsymbol{y})\left[\log\left[\pi(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})\right] - \frac{d}{dt}\log(z_t(\boldsymbol{y}))\right].
\end{aligned}$$

Hence,

$$\begin{aligned}
\frac{d}{dt}E_{\boldsymbol{\theta},\boldsymbol{x}|\boldsymbol{y},t}\log\left[\pi(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})\right] &= \int_{\boldsymbol{x}}\int_{\boldsymbol{\theta}}\left(\log\left[\pi(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})\right]\right)^2\pi_t(\boldsymbol{\theta},\boldsymbol{x}|\boldsymbol{y})d\boldsymbol{\theta}\,d\boldsymbol{x} \\
&- \frac{d}{dt}\log(z_t(\boldsymbol{y}))\int_{\boldsymbol{x}}\int_{\boldsymbol{\theta}}\log\left[\pi(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})\right]\pi_t(\boldsymbol{\theta},\boldsymbol{x}|\boldsymbol{y})d\boldsymbol{\theta}\,d\boldsymbol{x} \\
&= E_{\boldsymbol{\theta},\boldsymbol{x}|\boldsymbol{y},t}\left(\log\left[\pi(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})\right]\right)^2 - \left(E_{\boldsymbol{\theta},\boldsymbol{x}|\boldsymbol{y},t}\log\left[\pi(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})\right]\right)^2 \\
&= V_{\boldsymbol{\theta},\boldsymbol{x}|\boldsymbol{y},t}\log\left[\pi(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})\right].
\end{aligned}$$

Using the corrected trapezoidal rule form (Atkinson and Han, 2004), namely

$$\int_a^b f(y)dy \approx \frac{(b-a)}{2}\left[f(a)+f(b)\right] - \frac{(b-a)^2}{12}\left[f'(b)-f'(a)\right],$$

we have the following updated power posterior estimate of $\log(\pi(\boldsymbol{y}))$:

$$\begin{aligned}
\log(\pi(\boldsymbol{y})) &\approx \sum_{j=1}^{r}\frac{1}{2}(t_j - t_{j-1}) \\
&\times \left\{E_{\boldsymbol{\theta},\boldsymbol{x}|\boldsymbol{y},t_j}\log\left[\pi(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})\right] + E_{\boldsymbol{\theta},\boldsymbol{x}|\boldsymbol{y},t_{j-1}}\log\left[\pi(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})\right]\right\} \\
&- \sum_{j=1}^{r}\frac{1}{12}(t_j - t_{j-1})^2 \\
&\times \left\{V_{\boldsymbol{\theta},\boldsymbol{x}|\boldsymbol{y},t_j}\log\left[\pi(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})\right] - V_{\boldsymbol{\theta},\boldsymbol{x}|\boldsymbol{y},t_{j-1}}\log\left[\pi(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})\right]\right\}.
\end{aligned} \tag{4.2.19}$$

Algorithm 8 can be then used to implement the updated power posterior method for missing data models. The choice of spacing between the t_j values in (4.2.19) is crucial to improve the efficiency of the estimate of $\log(\pi(\boldsymbol{y}))$. Friel and Pettitt (2008) recommended a temperature schedule of the type $t_j =$

$(j/r)^c$, where r is the number of points in the interval $(0,1]$, and $c > 1$ is a constant. This temperature schedule choice ensures that many of the chosen t_j are close to zero which is appropriate in situations where the expected log-likelihood curve has a single sharp increase near zero. Also, the authors suggested to initialise the algorithm at the prior mean so by that immediate convergence to the power posterior at temperature $t = 0$ is guaranteed.

Algorithm 8 MCMC algorithm for estimating the marginal likelihood via the updated power posterior approach

1. Initialise algorithm with $\boldsymbol{x}^0$ and $\boldsymbol{\theta}^0$.
2. For $j = 0, ..., r$:

a. Set $t_j = (j/r)^c$, where $c > 1$ is a constant.

b. Generate a sample $\{(\boldsymbol{\theta}_j^{(1)}, \boldsymbol{x}_j^{(1)}), ..., (\boldsymbol{\theta}_j^{(M)}, \boldsymbol{x}_j^{(M)})\}$ from $\pi_{t_j}(\boldsymbol{\theta}, \boldsymbol{x}|\boldsymbol{y})$ via an MCMC sampling scheme.

c. Estimate $E_{\boldsymbol{\theta},\boldsymbol{x}|\boldsymbol{y},t_j} \log[\pi(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta})]$ and $V_{\boldsymbol{\theta},\boldsymbol{x}|\boldsymbol{y},t_j} \log[\pi(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta})]$.

d. While $j < r$, initialise the next chain at the previous posterior mean of $\pi_{t_j}(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{x})$.

3. Estimate $\log(\pi(\boldsymbol{y}))$ using (4.2.19).

The expectation and the variance of $\log[\pi(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta})]$ are estimated using MCMC runs as follows.

$$E_{\boldsymbol{\theta},\boldsymbol{x}|\boldsymbol{y},t_j} \log[\pi(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta})] \approx \frac{1}{M} \sum_{i=1}^{M} \log\left[\pi(\boldsymbol{y}, \boldsymbol{x}^{(i)}|\boldsymbol{\theta}^{(i)})\right], \tag{4.2.20}$$

and

$$\begin{aligned} V_{\boldsymbol{\theta},\boldsymbol{x}|\boldsymbol{y},t_j} \log[\pi(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta})] &\approx \frac{1}{M-1} \\ &\times \sum_{i=1}^{M} \left\{\log\left[\pi(\boldsymbol{y}, \boldsymbol{x}^{(i)}|\boldsymbol{\theta}^{(i)})\right] - E_{\boldsymbol{\theta},\boldsymbol{x}|\boldsymbol{y},t_j} \log[\pi(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta})]\right\}^2, \end{aligned} \tag{4.2.21}$$

4.2.7 Objectives of current research

In our investigation, we mainly focus on implementing the missing-data extension of power posterior method (see Section 4.2.6) for estimating the Bayes factor in the epidemic context as the performance of this approach has not been examined yet for stochastic epidemic models. The key advantages of this approach are that, unlike other methods, it does not require much tuning and it involves a minimal amount of change to the commonly used MCMC algorithms that compute the posterior distributions of unknown epidemic model parameters which makes it a good alternative to other Bayesian model selection methods such as the RJMCMC method.

The ability of DIC_6, introduced in Section 4.2.2.1 to select between epidemic models, is also investigated. Additionally, we look at some epidemic model selection aspects where the outbreak size matters. However, before that and under the assumption of observing the entire course of epidemic through time (complete data case), explicit forms for Bayes factors in two different epidemic settings are derived and the capability of the (simple) prior arithmetic mean estimator for the marginal likelihood in approximating Bayes factors is explored.

Throughout the investigation, various transmission models (all of these models are introduced in the previous chapters) are compared, namely the SIR model with different infectious periods and the SIR model with different infection mechanisms, in order to determine how well model selection techniques can discriminate between them.

4.3 Model selection via Bayes factors for complete epidemic data

In this section, we shall derive explicit expressions for the Bayes factor in two different epidemic settings based on complete outbreak data. The assumption of having complete data is often unrealistic but it enables us to obtain explicit forms for Bayes factors in our simple settings below. The motivations for looking at Bayes factor formula are that it is of interest in its own right and it provides us of some insights to gain a better understanding of this Bayesian model choice tool. In addition, we can validate the estimate of the Bayes factor derived from the prior arithmetic mean estimator against the Bayes factor true value in order to see how reliable it is.

Let us start first with the SIR model with different infectious periods.

4.3.1 The SIR model with different infectious periods: complete data case

Suppose that fully observed temporal outbreak data (infection and removal times) of size n_R, in a closed population of size $\boldsymbol{N}$ (n initial susceptibles and one initial infective), are observed on $[I_z, R_{n_R}]$, say $\boldsymbol{I} = (I_1, ..., I_{n_R})$ and $\boldsymbol{R} = (R_1, ..., R_{n_R})$ such that for $1 \leq j \leq n_R$, I_j corresponds to the infection time of the individual removed at time R_j, with the initial infection time denoted by I_z, where $I_z < I_j$, for all $j \neq z$.

We consider two SIR type competing models, namely the SIR with exponential infectious period ($T_I \sim \text{Exp}(\gamma)$) denoted by m_1 and the SIR with gamma infectious period ($T_I \sim \text{Gamma}(\alpha, \delta)$) denoted by m_2, where the shape parameter of gamma distribution α is assumed to be known.

By assigning an independent gamma prior distribution for each of the model parameters, namely $\text{Gamma}(\lambda_\zeta, \nu_\zeta)$, where $\zeta = \beta,\ \gamma,\ \delta$, the Bayes factor can

be derived explicitly in this case using (4.2.1) as follows.

$$\begin{aligned}
BF_{12} &= \frac{\pi(\boldsymbol{I},\boldsymbol{R}|m_1)}{\pi(\boldsymbol{I},\boldsymbol{R}|m_2)} \\
&= \frac{\int_\gamma \int_\beta \pi(\boldsymbol{I},\boldsymbol{R}|\beta,\gamma)\pi(\beta)\pi(\gamma)d\beta\, d\gamma}{\int_\delta \int_\beta \pi(\boldsymbol{I},\boldsymbol{R}|\beta,\alpha,\delta)\pi(\beta)\pi(\delta)d\beta\, d\delta} \\
&= \frac{\nu_\beta^{\lambda_\beta}\,\nu_\gamma^{\lambda_\gamma}\,\Gamma(\lambda_\beta)\,\Gamma(\lambda_\delta)}{\nu_\beta^{\lambda_\beta}\,\nu_\delta^{\lambda_\delta}\,\Gamma(\lambda_\beta)\,\Gamma(\lambda_\gamma)} \times \frac{\Gamma^{n_R}(\alpha)\times\prod_{j=1,j\neq z}^{n_R} n^{-1}Y(I_j-)}{\prod_{j=1,j\neq z}^{n_R} n^{-1}Y(I_j-)\times\prod_{j=1}^{n_R}(R_j-I_j)^{\alpha-1}} \\
&\times \frac{\int_\beta \beta^{n_R-1}\times e^{-\beta n^{-1}A}\times\beta^{\lambda_\beta-1}e^{-\nu_\beta\beta}d\beta}{\int_\beta \beta^{n_R-1}\times e^{-\beta n^{-1}A}\times\beta^{\lambda_\beta-1}e^{-\nu_\beta\beta}d\beta} \\
&\times \frac{\int_\gamma \gamma^{n_R}\, e^{-\gamma\sum_{j=1}^{n_R}(R_j-I_j)}\times\gamma^{\lambda_\gamma-1}e^{-\nu_\gamma\gamma}d\gamma}{\int_\delta \delta^{\alpha n_R}\, e^{-\delta\sum_{j=1}^{n_R}(R_j-I_j)}\times\delta^{\lambda_\delta-1}e^{-\nu_\delta\delta}d\delta} \\
&= \frac{\nu_\gamma^{\lambda_\gamma}\,\Gamma(\lambda_\delta)\times\Gamma^{n_R}(\alpha)}{\nu_\delta^{\lambda_\delta}\,\Gamma(\lambda_\gamma)\times\prod_{j=1}^{n_R}(R_j-I_j)^{\alpha-1}} \\
&\times \frac{(\nu_\delta+\sum_{j=1}^{n_R}(R_j-I_j))^{\alpha n_R+\lambda_\delta}\times\Gamma(n_R+\lambda_\gamma)}{(\nu_\gamma+\sum_{j=1}^{n_R}(R_j-I_j))^{n_R+\lambda_\gamma}\times\Gamma(\alpha n_R+\lambda_\delta)},
\end{aligned} \tag{4.3.1}$$

where $A=\sum_{j=1}^{n_R}\sum_{k=1}^{\boldsymbol{N}}(R_j\wedge I_k - I_k\wedge I_j)$, with $I_k=\infty$ for $k=n_R+1,...,\boldsymbol{N}$.

It can be noted that the Bayes factor expression is independent of the infection rate prior parameters which is expected for the following reason. The epidemic likelihood may be broken into two independent (infection and removal) parts and as the infection processes are the same for both SIR models, they are cancelled out.

Common choices of the prior parameters in the literature are $\lambda_\gamma=\lambda_\delta=\lambda$ and $\nu_\gamma=\nu_\delta=\nu$, where $\lambda\geq 1$ and ν is a small positive number. This prior setting yields

$$BF_{12}=\frac{\Gamma(n_R+\lambda)\;\Gamma^{n_R}(\alpha)}{\Gamma(\alpha n_R+\lambda)\;\prod_{j=1}^{n_R}(R_j-I_j)^{\alpha-1}}\times\left(\nu+\sum_{j=1}^{n_R}(R_j-I_j)\right)^{n_R(\alpha-1)}. \tag{4.3.2}$$

To see the effect of the prior's diffuseness on BF_{12}, we reformulate BF_{12} in (4.3.2) in terms of the mean and variance of the prior distribution, that are

$$E\left[\text{Gamma}(\lambda,\nu)\right]=\lambda/\nu=\mu \;\text{ and }\; \text{Var}\left[\text{Gamma}(\lambda,\nu)\right]=\lambda/\nu^2=\sigma^2,$$

which implies $\lambda = \mu^2/\sigma^2$ and $\nu = \mu/\sigma^2$. Substituting these values into (4.3.2) gives

$$BF_{12} = \frac{\Gamma\left(n_R + \frac{\mu^2}{\sigma^2}\right)\ \Gamma^{n_R}(\alpha)}{\Gamma\left(\alpha n_R + \frac{\mu^2}{\sigma^2}\right)\ \prod_{j=1}^{n_R}(R_j - I_j)^{\alpha-1}} \times \left(\frac{\mu}{\sigma^2} + \sum_{j=1}^{n_R}(R_j - I_j)\right)^{n_R(\alpha-1)}. \tag{4.3.3}$$

Hence, as $\sigma^2 \to \infty$, the Bayes factor tends to its lower limit, that is

$$BF_{12} \to \frac{\Gamma(n_R)\ \Gamma^{n_R}(\alpha)}{\Gamma(\alpha n_R)\ \prod_{j=1}^{n_R}(R_j - I_j)^{\alpha-1}} \times \left(\sum_{j=1}^{n_R}(R_j - I_j)\right)^{n_R(\alpha-1)}. \tag{4.3.4}$$

However, as the prior gets more and more concentrated at μ, $\sigma^2 \to 0$, the Bayes factor becomes more decisive in supporting m_1, that is $BF_{12} \to \infty$.

For a typical case in practice where $\lambda = 1$ and $\nu \to 0$ (diffuse prior), Figure 4.1 shows $\log(BF_{12})$ values for some outbreak data generated under model m_1 in a closed population of size $\boldsymbol{N} = 50$, 40 of whom were infected , with parameter values $\beta = 2$ and $\gamma = 1$. We set the gamma shape parameter to $\alpha = 2$, and then $\log(BF_{12})$ was calculated for various values of ν. It is clear that as the prior become less informative (i.e. ν decreases), the logarithm of Bayes factor converges to its lower limit value 3.1039.

As an side, when performing a Bayesian model selection task using the Bayes factor criterion, one needs to be careful of any possible connection to Lindley's paradox (Lindley, 1957). In this epidemic setting, models are not nested and have the same complexity (α is assumed to be known in model m_2), so as the prior variance goes to infinity (i.e $\nu \to 0$), the Bayes factor tends to a constant lower bound (see Figure 4.1). However, as the prior on both parameters, γ and δ, becomes more and more informative (i.e $\nu \to \infty$), both $\gamma, \delta \to 0$ a priori and so intuitively, observing the data under m_1 is less surprising than under m_2.

For the particular epidemic setting that we have considered above in which γ and δ follow a Gamma(λ, ν), we may conclude that increasing (decreasing)

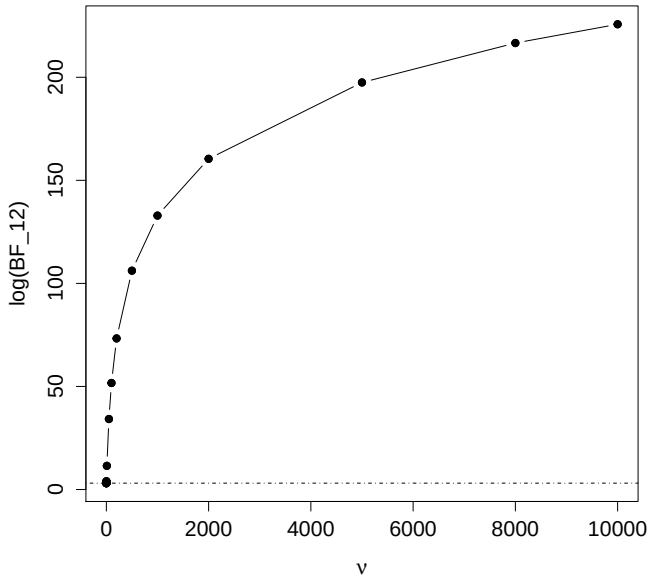


Figure 4.1: Values of the logarithm of Bayes factor, $\log(BF_{12})$, calculated using (4.3.2) by varying ν (γ and δ prior scale parameter) and setting $\lambda = 1$. The horizontal dashed line denotes the lower limit value of $\log(BF_{12})$ (3.1039) as $\nu \to 0$.

the prior uncertainty of the model parameters reduces (increases) the Bayes factor value.

4.3.2 The SIR model with different infection mechanisms: complete data case

Suppose that fully observed temporal outbreak data of size n_R, in a closed population of size $\boldsymbol{N}$ (n initial susceptibles and one initial infective), are observed on $[I_1, R_{n_R}]$. Let $I_1, I_2, ..., I_{n_R}$ and $R_1, R_2, ..., R_{n_R}$ be the observed ordered infection and removal times respectively and write $\boldsymbol{I} = (I_1, I_2, ..., I_{n_R})$ and $\boldsymbol{R} = (R_1, ..., R_{n_R})$, where $I_k < I_{k+1} < R_k$ for $k = 1, 2, ..., n_R - 1$.

We consider two Markovian SIR models with different forms of infection rate. The first model denoted by m_1 is an SIR model in which the rate of new infections is assumed to have a linear relation with the number of suscept-

ible individuals multiplied by the number of infective individuals at time t, that is $\beta n^{-1}X(t)Y(t)$. The other epidemic SIR model m_2 is a modification of m_1 model by relaxing the linearity assumption and considering the rate of new infections to be of the form $\beta n^{-1}X(t)Y^p(t)$, where the power parameter $p \in (0, 0.5)$ is assumed to be known. The rate of new removals is assumed to be the same for both models, that is $\gamma Y(t)$.

By assigning an independent gamma prior for each of the models parameters, namely a Gamma$(\lambda_\zeta, \nu_\zeta)$, where $\zeta = \beta,\ \gamma$, the Bayes factor can be obtained explicitly in this case using (4.2.1) as follows.

$$\begin{aligned}
BF_{12} &= \frac{\pi(\boldsymbol{I}, \boldsymbol{R}|m_1)}{\pi(\boldsymbol{I}, \boldsymbol{R}|m_2)} \\
&= \frac{\int_\gamma \int_\beta \pi(\boldsymbol{I}, \boldsymbol{R}|\beta, \gamma)\pi(\beta)\pi(\gamma)d\beta\, d\gamma}{\int_\gamma \int_\beta \pi(\boldsymbol{I}, \boldsymbol{R}|\beta, \gamma, p)\pi(\beta)\pi(\gamma)d\beta\, d\gamma} \\
&= \frac{\nu_\beta^{\lambda_\beta}\, \nu_\gamma^{\lambda_\gamma}\, \Gamma(\lambda_\beta)\, \Gamma(\lambda_\gamma)}{\nu_\beta^{\lambda_\beta}\, \nu_\gamma^{\lambda_\gamma}\, \Gamma(\lambda_\beta)\, \Gamma(\lambda_\gamma)} \times \frac{\prod_{j=2}^{n_R} n^{-1}X(I_j-)Y(I_j-)}{\prod_{j=2}^{n_R} n^{-1}X(I_j-)Y^p(I_j-)} \times \frac{\prod_{j=1}^{n_R} Y(R_j-)}{\prod_{j=1}^{n_R} Y(R_j-)} \\
&\times \frac{\int_\gamma \gamma^{n_R}\, e^{-\gamma B} \times \gamma^{\lambda_\gamma - 1} e^{-\nu_\gamma \gamma} d\gamma}{\int_\gamma \gamma^{n_R}\, e^{-\gamma B} \times \gamma^{\lambda_\gamma - 1} e^{-\nu_\gamma \gamma} d\gamma} \\
&\times \frac{\int_\beta \beta^{n_R - 1} \times e^{-\beta n^{-1} A} \times \beta^{\lambda_\beta - 1} e^{-\nu_\beta \beta} d\beta}{\int_\beta \beta^{n_R - 1} \times e^{-\beta n^{-1} A_p} \times \beta^{\lambda_\beta - 1} e^{-\nu_\beta \beta} d\beta} \\
&= \prod_{j=2}^{n_R} Y^{1-p}(I_j-) \times \frac{\int_\beta \beta^{n_R + \lambda_\beta - 2} \times e^{-\beta(\nu_\beta + n^{-1}A)} d\beta}{\int_\beta \beta^{n_R + \lambda_\beta - 2} \times e^{-\beta(\nu_\beta + n^{-1}A_p)} d\beta} \\
&= \prod_{j=2}^{n_R} Y^{1-p}(I_j-) \times \frac{(\nu_\beta + n^{-1}A_p)^{n_R + \lambda_\beta - 1}\, \Gamma(n_R + \lambda_\beta - 1)}{(\nu_\beta + n^{-1}A)^{n_R + \lambda_\beta - 1}\, \Gamma(n_R + \lambda_\beta - 1)} \\
&= \prod_{j=2}^{n_R} Y^{1-p}(I_j-) \times \left(\frac{\nu_\beta + n^{-1}A_p}{\nu_\beta + n^{-1}A} \right)^{n_R + \lambda_\beta - 1}, \qquad (4.3.5)
\end{aligned}$$

where

$$A = \int_{I_1}^{R_{n_R}} X(t)Y(t)\, dt, \quad A_p = \int_{I_1}^{R_{n_R}} X(t)Y^p(t)\, dt \quad \text{and} \quad B = \int_{I_1}^{R_{n_R}} Y(t)\, dt.$$

By looking at (4.3.5), we see that BF_{12} depends only on the data and the prior parameter of the infection rate ν_β, which is expected as the removal processes are the same for both SIR models. Another interesting observation is that,

regardless of the choice of prior distribution, since $A_p \leq A$ for any $p \in (0, 0.5)$, BF_{12} has an upper bound, that is

$$BF_{12} \leq \prod_{j=2}^{n_R} Y^{1-p}(I_j-).$$

To see the impact of the infection rate prior distribution on the behaviour of BF_{12}, we rewrite BF_{12} in terms of the mean and variance of the prior distribution, that are

$$E\left[\text{Gamma}(\lambda_\beta, \nu_\beta)\right] = \lambda_\beta/\nu_\beta = \mu \text{ and } \text{Var}\left[\text{Gamma}(\lambda_\beta, \nu_\beta)\right] = \lambda_\beta/\nu_\beta^2 = \sigma^2,$$

which implies $\lambda_\beta = \mu^2/\sigma^2$ and $\nu_\beta = \mu/\sigma^2$. Substituting these values into (4.3.5) yields

$$BF_{12} = \prod_{j=2}^{n_R} Y^{1-p}(I_j-) \times \left(\frac{\frac{\mu}{\sigma^2} + \frac{A_p}{n}}{\frac{\mu}{\sigma^2} + \frac{A}{n}} \right)^{n_R + \frac{\mu^2}{\sigma^2} - 1}. \tag{4.3.6}$$

Consequently, as $\sigma^2 \to \infty$ or $\sigma^2 \to 0$, we have the following limits for the Bayes factor BF_{12}, namely

$$BF_{12} \to (A_p/A)^{n_R - 1} \prod_{j=2}^{n_R} Y^{1-p}(I_j-), \text{ as } \sigma^2 \to \infty, \tag{4.3.7}$$

and

$$BF_{12} \to \prod_{j=2}^{n_R} Y^{1-p}(I_j-), \text{ as } \sigma^2 \to 0. \tag{4.3.8}$$

To illustrate this with $\lambda_\beta = 1$ and various values of ν_β, some outbreak data were simulated from model m_1 in a population of size $\boldsymbol{N} = 50$ with $\beta = 2$ and $\gamma = 1$. The final size of the epidemic was 42 and we set $p = 0.3$ in model m_2. Then, $\log(BF_{12})$ was calculated and plotted against various values of ν_β as in Figure 4.2. We see that as the prior gets diffuse, the $\log(BF_{12})$ tends to the lower constant limit 3.8890, whereas with a much more concentrated prior, $\log(BF_{12}) \to 55.5122$, the upper limit value.

In this specific epidemic setting where $\beta \sim \text{Gamma}(\lambda_\beta, \nu_\beta)$, we can conclude that as the prior distribution of the infection rate parameter gets diffuse, the value of the Bayes factor decreases to its limit. However, more informative prior results in increasing of the Bayes factor value.

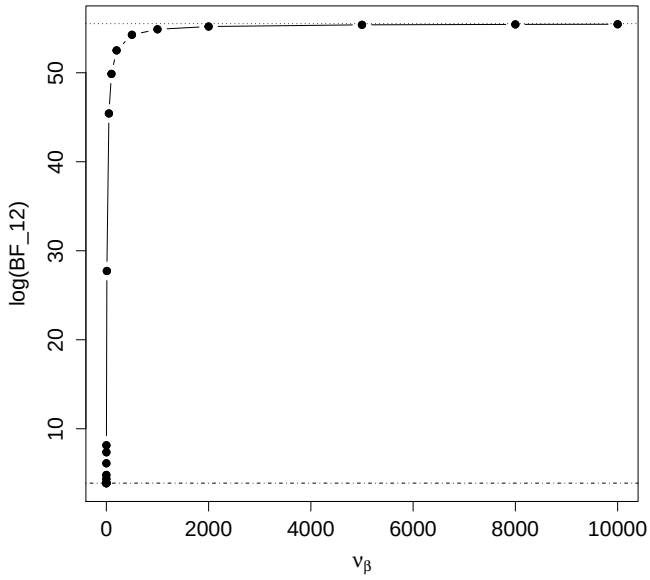


Figure 4.2: $\log(BF_{12})$ values calculated using (4.3.5), varying the prior parameter ν_β and setting $\lambda_\beta = 1$. The horizontal dashed and dotted lines represent the lower (3.8890) and upper (55.5122) limit values of $\log(BF_{12})$ respectively.

4.3.3 Approximating the Bayes factor for epidemic models using the prior arithmetic mean estimator

For illustrative purposes, we shall consider the two SIR models introduced in Section 4.3.1. As we have a prior of two dimensions in each model, we expect the prior arithmetic mean estimates of the marginal likelihoods (see Section 4.2.4), namely

$$\pi(\widehat{\boldsymbol{I},\boldsymbol{R}}|m_1) = \frac{1}{M}\sum_{i=1}^{M}\pi(\boldsymbol{I},\boldsymbol{R}|\beta^{(i)},\gamma^{(i)}), \tag{4.3.9}$$

and

$$\pi(\widehat{\boldsymbol{I},\boldsymbol{R}}|m_2) = \frac{1}{M}\sum_{i=1}^{M}\pi(\boldsymbol{I},\boldsymbol{R}|\alpha,\beta^{(i)},\delta^{(i)}), \tag{4.3.10}$$

to be effective in the estimation and recover the true value of the Bayes factor. Here, $\beta^{(i)}, \gamma^{(i)}$ and $\delta^{(i)}$, where $i = 1, 2, ..., M$, are independent samples drawn from the gamma prior distribution. Note that, $\text{Gamma}(1, \mu) \equiv \text{Exp}(\mu)$, where $\mu > 0$.

In what follows, we conducted a simulation study to illustrate the effectiveness of the prior arithmetic mean estimator of the marginal likelihood in approximating the true marginal likelihood and hence the Bayes factor. This simulation study was performed under different prior assumptions for each model parameters.

4.3.3.1 Application of the prior arithmetic mean estimator to simulated outbreak data

We simulated two epidemic datasets, one dataset was generated under m_1 and the other dataset was simulated from m_2. Each outbreak was in a closed population consisting of 49 initially susceptible individuals and a single initially infected case. The model parameters for the simulation were $\beta = 1, \gamma = 0.5$ for model m_1 and $\beta = 1, \alpha = 2, \delta = 1$ for model m_2. The resulting final sizes were 40 and 44 under m_1 and m_2 respectively.

Under each prior assumption, namely $\text{Exp}(1), \text{Exp}(0.1)$ and $\text{Exp}(0.01)$, the true values of the logarithm of Bayes factors were calculated using (4.3.1). The estimated values of the logarithm of Bayes factors, under different prior distributions, were computed as follows. Initially, we simulate M samples from the prior for each parameter and then evaluate the marginal likelihoods based on these prior samples using (4.3.9) and (4.3.10). Then, the logarithm of Bayes factor can be estimated as follows.

$$\log(\widehat{BF_{12}}) = \log(\pi(\widehat{\boldsymbol{I}, \boldsymbol{R}}|m_1)) - \log(\pi(\widehat{\boldsymbol{I}, \boldsymbol{R}}|m_2)). \tag{4.3.11}$$

Results are displayed in Table 4.2. It was found that, increasing prior uncertainty of the model parameters does not have much effect on the precision of Bayes factor estimation, as the prior in our case is low-dimensional and so fast to sample from. The typical run times on a desktop PC for estimating $\log(BF_{12})$ were about 30 seconds, 5 and 60 minutes for $M = 10^4$, $M = 10^5$ and $M = 10^6$ respectively. These results suggest that although the flat prior distribution expresses a form of neutrality and lack of previous knowledge, its

choice is not very reasonable in the present application. In contrast, a more informative prior (e.g. Exp(1)) has the advantage of estimating $\log(BF_{12})$ more accurately. Also, these results indicate the possibility of using the prior arithmetic mean estimator of the marginal likelihood as a simple but effective way of approximating Bayes factors in situations where complete epidemic data are available but Bayes factors cannot be obtained in a closed form. For instance, when comparing these SIR models with unknown gamma shape parameter α, or when comparing the SIR model with different infection mechanisms and unknown power parameter p.

Model	Prior	True $\log(BF_{12})$	$\log(\widehat{BF_{12}})$		
			$M = 10^4$	$M = 10^5$	$M = 10^6$
m_1	Exp(1)	6.515506	6.37311	6.510376	6.517237
m_1	Exp(0.1)	6.055084	5.246776	6.095281	6.051161
m_1	Exp(0.01)	6.008756	14.29707	6.67211	5.989908
m_2	Exp(1)	−5.075104	−5.027778	−5.0823	−5.074423
m_2	Exp(0.1)	−5.440686	−5.267586	−5.330276	−5.365737
m_2	Exp(0.01)	−5.477409	−21.09641	−5.911539	−5.679113

Table 4.2: The logarithm of Bayes factor results obtained using two datasets simulated based on models m_1 (the SIR with $T_I \sim \text{Exp}(\gamma)$) and m_2 (the SIR with $T_I \sim \text{Gamma}(\alpha, \delta)$). For each case, model parameters were assigned prior distributions to varying degrees of diffuseness to determine the effect on the logarithm of Bayes factor estimation. Three sizes of prior parameter sample ($M = 10^4, M = 10^5$ and $M = 10^6$) were used to determine the effective sample size required to reasonably estimate the logarithm of Bayes factor.

4.4 Model selection via the updated power posterior and DIC methods for incomplete epidemic data

In the following, we focus on an epidemic outbreak that results in a total of n_R removals in a closed population of size n initial susceptibles and one initial infective. Let $\boldsymbol{R}$ and $\boldsymbol{I}$ represent the vectors of the observed removal times and the missing infection times respectively. Also, let $\boldsymbol{\theta}$ indicate a set of the model parameters and the initial infection time.

4.4.1 The updated power posterior and DIC$_6$ methods for epidemic models

The power posterior in our epidemic setting, with $\boldsymbol{\theta}_i,\ i = 1, 2,$ denoting the m_i model parameters, is

$$\pi_t(\boldsymbol{\theta}_i, \boldsymbol{I}|\boldsymbol{R}) \propto \pi(\boldsymbol{R}, \boldsymbol{I}|\boldsymbol{\theta}_i)^t \pi(\boldsymbol{\theta}_i), \tag{4.4.1}$$

with the normalizing constant

$$z_t(\boldsymbol{R}|m_i) = \int_{\boldsymbol{I}} \int_{\boldsymbol{\theta}_i} \pi(\boldsymbol{R}, \boldsymbol{I}|\boldsymbol{\theta}_i)^t \pi(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i\, d\boldsymbol{I}. \tag{4.4.2}$$

Hence, the updated power posterior estimate of $\log(\pi(\boldsymbol{R}|m_i))$ becomes (see Section 4.2.6)

$$\begin{aligned} \log(\pi(\boldsymbol{R}|m_i)) \approx & \sum_{j=1}^{r} \frac{1}{2}(t_j - t_{j-1}) \\ & \times \left\{ E_{\boldsymbol{\theta}_i, \boldsymbol{I}|\boldsymbol{R}, t_j} \log\left[\pi(\boldsymbol{R}, \boldsymbol{I}|\boldsymbol{\theta}_i)\right] + E_{\boldsymbol{\theta}_i, \boldsymbol{I}|\boldsymbol{R}, t_{j-1}} \log\left[\pi(\boldsymbol{R}, \boldsymbol{I}|\boldsymbol{\theta}_i)\right] \right\} \\ & - \sum_{j=1}^{r} \frac{1}{12}(t_j - t_{j-1})^2 \\ & \times \left\{ V_{\boldsymbol{\theta}_i, \boldsymbol{I}|\boldsymbol{R}, t_j} \log\left[\pi(\boldsymbol{R}, \boldsymbol{I}|\boldsymbol{\theta}_i)\right] - V_{\boldsymbol{\theta}_i, \boldsymbol{I}|\boldsymbol{R}, t_{j-1}} \log\left[\pi(\boldsymbol{R}, \boldsymbol{I}|\boldsymbol{\theta}_i)\right] \right\}. \end{aligned} \tag{4.4.3}$$

Now, the logarithm of the Bayes factor can be calculated as

$$\begin{aligned} \log(BF_{12}) &= \log\left[\frac{\pi(\boldsymbol{R}|m_1)}{\pi(\boldsymbol{R}|m_2)}\right] \\ &= \log(\pi(\boldsymbol{R}|m_1)) - \log(\pi(\boldsymbol{R}|m_2)). \end{aligned} \tag{4.4.4}$$

We now describe the use of the DIC to select a model which fits the observed data best. As we have missing data $\boldsymbol{I}$ in our epidemic setting, we adopt the DIC_6 introduced in Section 4.2.2.1, that is

$$\text{DIC}_6 = -4E_{\boldsymbol{\theta}_i,\boldsymbol{I}|\boldsymbol{R}}[\log(\pi(\boldsymbol{R},\boldsymbol{I}|\boldsymbol{\theta}_i))] + 2E_{\boldsymbol{I}|\boldsymbol{R},\widehat{\boldsymbol{\theta}}_i}[\log(\pi(\boldsymbol{R},\boldsymbol{I}|\widehat{\boldsymbol{\theta}}_i))]. \tag{4.4.5}$$

The quantities $E_{\boldsymbol{\theta}_i,\boldsymbol{I}|\boldsymbol{R}}[\log(\pi(\boldsymbol{R},\boldsymbol{I}|\boldsymbol{\theta}_i))]$ and $\widehat{\boldsymbol{\theta}}_i$ are readily estimated by the power posterior method at $t = 1$ at no extra computational cost. Therefore, we need to run the data-augmented MCMC algorithm once only to estimate the expectation $E_{\boldsymbol{I}|\boldsymbol{R},\widehat{\boldsymbol{\theta}}_i}[\log(\pi(\boldsymbol{R},\boldsymbol{I}|\widehat{\boldsymbol{\theta}}_i))]$ with $\boldsymbol{\theta}_i = \widehat{\boldsymbol{\theta}}_i$ and allowing $\boldsymbol{I}$ to vary, where $\widehat{\boldsymbol{\theta}}_i$ is a vector of posterior point estimates of each model parameters and the initial infection time.

4.4.2 Application of the updated power posterior and DIC_6 methods to the SIR model with different infectious periods

We consider an epidemic setting in which the SIR model with different infectious periods is examined. More precisely, the SIR model with exponential infectious period ($T_I \sim \text{Exp}(\gamma)$) denoted by m_1 is compared to the SIR model with gamma infectious period ($T_I \sim \text{Gamma}(\alpha, \delta)$) denoted by m_2, where the shape parameter of gamma distribution α is assumed to be known.

Suppose that we observe an epidemic outbreak that results in a total of n_R removals in a closed population of size $\boldsymbol{N}$ (n initial susceptibles and one initial infective) on $(I_z, R_{n_R}]$. For $j = 1, ..., n_R$, let I_j be the infection time of the individual removed at time R_j . Also, denote the initial infection time by I_z, where $I_z < I_j$, for all $j \neq z$. Finally, let $\boldsymbol{R} = (R_1, ..., R_{n_R})$ represent the set of observed removal times, where $0 = R_1 \leq R_2 \leq ... \leq R_{n_R}$ and $\boldsymbol{I} = \{I_j : j \neq z\}$ denote the set of missing infection times.

The likelihoods of $\boldsymbol{I}$ and $\boldsymbol{R}$ given the model parameters are

$$\pi(\boldsymbol{I},\boldsymbol{R}|\beta,\gamma,I_z,z,m_1) = \beta^{n_R-1}\prod_{j=1,j\neq z}^{n_R} n^{-1}Y(I_j-) \times e^{-\beta n^{-1}A} \times \gamma^{n_R}\, e^{-\gamma\sum_{j=1}^{n_R}(R_j-I_j)},$$

and

$$\begin{aligned}\pi(\boldsymbol{I},\boldsymbol{R}|\alpha,\beta,\delta,I_z,z,m_2) &= \beta^{n_R-1}\prod_{j=1,j\neq z}^{n_R} n^{-1}Y(I_j-)\times e^{-\beta n^{-1}A}\\ &\times \Gamma^{-n_R}(\alpha)\times\prod_{j=1}^{n_R}(R_j-I_j)^{\alpha-1}\times\delta^{\alpha n_R}\, e^{-\delta\sum_{j=1}^{n_R}(R_j-I_j)},\end{aligned}$$

where

$$A=\int_{I_z}^{R_{n_R}} X(t)Y(t)dt=\sum_{j=1}^{n_R}\sum_{k=1}^{\boldsymbol{N}}(R_j\wedge I_k-I_k\wedge I_j),$$

with $I_k=\infty$ for $k=n_R+1,...,\boldsymbol{N}$.

To each parameter $\beta,\ \gamma$, and δ we assign an independent exponential prior, namely $\text{Exp}(\lambda_\zeta)$, where $\zeta=\beta,\ \gamma,\ \delta$. The initial infection time I_z is assigned a prior, in particular $-I_z\sim\text{Exp}(\theta)$. Also, we assume a priori that z is equally likely to be any of the n_R infected individuals.

At temperature t, the full conditional power posterior densities for β,γ and δ are

$$\begin{aligned}\beta|t,\gamma,z,I_z,\boldsymbol{I},\boldsymbol{R}&\sim\text{Gamma}\left(1+t(n_R-1),\lambda_\beta+tn^{-1}A\right),\\ \gamma|t,\beta,z,I_z,\boldsymbol{I},\boldsymbol{R}&\sim\text{Gamma}\left(1+n_Rt,\lambda_\gamma+t\sum_{j=1}^{n_R}(R_j-I_j)\right),\\ \delta|t,\beta,z,I_z,\boldsymbol{I},\boldsymbol{R}&\sim\text{Gamma}\left(1+t\alpha n_R,\lambda_\delta+t\sum_{j=1}^{n_R}(R_j-I_j)\right),\end{aligned}$$

and the proportional full conditional power posterior density for $(\boldsymbol{I},z,I_z)$ under model m_2 (the proportional form for $(\boldsymbol{I},z,I_z)$ under model m_1 is the same when setting $\alpha=1$ and replacing δ by γ) is

$$\begin{aligned}\pi(\boldsymbol{I},z,I_z|t,\alpha,\beta,\delta,\boldsymbol{R}) &\propto \left\{\prod_{j=1,j\neq z}^{n_R} n^{-1}Y(I_j-)\times e^{-\beta n^{-1}A}\right\}^t\\ &\times\left\{\prod_{j=1}^{n_R}(R_j-I_j)^{\alpha-1}\times e^{-\delta\sum_{j=1}^{n_R}(R_j-I_j)}\right\}^t\times e^{\theta I_z}.\end{aligned}$$

Note that z and I_z are possibly updated while updating $\boldsymbol{I}$, so there is no need to update them separately.

Sampling from the above power posterior densities can be performed using a modified version of our MCMC algorithm described in Chapter 2 in which the

densities are now raised to the power t. Using a fixed grid of points, $0 = t_0 < t_1 < ... < t_{r-1} < t_r = 1$, such that $t_j = (j/r)^c$ as recommended by Friel and Pettitt (2008), where $r = 20$ or $r = 40$ depending on the diffuseness of the prior and $c = 5$, the log marginal likelihoods are obtained according to Algorithm 8 and hence the logarithm of Bayes factor is approximated using (4.4.4). This temperature schedule ensures that many of the chosen temperatures are close to zero which is appropriate in situations where the behaviour of the expected log-likelihood under each power posterior has a single sharp increase near zero.

The DIC_6 calculation involves two steps. First, the power posterior method at $t = 1$ provides us with the posterior mean and median values of each model parameters and the initial infection time in each model, say $\widehat{\beta}, \widehat{\gamma}, \widehat{\delta}$ and $\widehat{I_z}$. Second, having derived these posterior estimates, we run the data-augmentation MCMC algorithm with each model parameters and each initial infection time fixed to their estimated values, allowing only the missing infection times $\boldsymbol{I}$ to vary. This enables us to derive DIC_6 for each model.

4.4.2.1 Factors affecting the updated power posterior and DIC_6 performances

In this section, a series of simulation experiments was run to determine the impact of two important factors, namely the temperature ladder and the prior distribution choice on model selection using the power posterior approach and DIC_6 criterion. We simulated data according to each of the two SIR models, and then aimed to identify the correct model using our approaches. Under each scenario, we simulated and analysed several outbreak datasets. Unless otherwise stated, within each temperature t, we collected $27,000$ samples from the power posterior of which the first $2,000$ were discarded as burn-in and the rest were thinned by taking each 5th value. The value of θ used for the prior distribution of I_z was set at $\theta = 1$ throughout. For the DIC_6 procedure, having derived the posterior point estimates for β, γ, δ and I_z using the

power posterior at $t = 1$ and allowing only the augmented data $\boldsymbol{I}$ to vary, we ran a within-model data-augmented MCMC algorithm for each model for $17,000$ iterations with disregarding a burn-in sample of $2,000$ and thinning the chain by taking the 5th value. Generally speaking, convergence and mixing of the Markov chains were both found to be satisfactory and in the following presentation of results, our findings are based on one representative simulated dataset for each scenario.

- **Choice of temperature ladder**

The number of t_j points between $t_0 = 0$ and $t_r = 1$ that should be used in estimating the logarithm of the marginal likelihood (4.4.3) and hence the logarthim of Bayes factor (4.4.4) is a potential issue when implementing the power posterior method. Additionally, it is desirable to control the number of temperature rungs and use only as many as necessary in order to save the computational cost. Therefore, in this study we examined the amount of discretization r in the interval $(0, 1]$ that required to improve the efficiency of the Bayes factor estimation under a common used prior $\text{Exp}(1)$ and a relatively vague prior $\text{Exp}(0.01)$ for the parameters β, γ and δ. A population of size $\boldsymbol{N} = 30$ was considered and 22 removal events were simulated under the model m_1, with $\beta = 1$ and $\gamma = 0.5$. Then, both models were fitted to these data with $\alpha = 10$ in m_2. Results are summarised in Table 4.3 under both prior assumptions.

These figures suggest that the estimate of $\log(BF_{12})$ is not too sensitive to the amount of discretization r, as long as a fairly informative prior distribution is specified such as $\text{Exp}(1)$. However, with a non-informative prior distribution such as $\text{Exp}(0.01)$, a large enough number of temperatures is crucial in estimating $\log(BF_{12})$ accurately although this must be balanced against the computational cost. It was also found that the variance of $E_{\boldsymbol{\theta},\boldsymbol{I}|\boldsymbol{R},t} \log[\pi(\boldsymbol{R}, \boldsymbol{I}|\boldsymbol{\theta})]$ is largest as $t \to 0$, suggesting that more MCMC iterations should be allocated as $t \to 0$ rather than as $t \to 1$ to reasonably obtain adequate estimates.

r	$\log(\pi(\boldsymbol{R}\|m_1))$ (Bias)	$\log(\pi(\boldsymbol{R}\|m_2))$ (Bias)	$\log(BF_{12})$
$\beta, \gamma, \delta \sim \text{Exp}(1)$			
10	$-129.8600(1.10)$	$-152.2454(1.31)$	22.38544
20	$-130.0118(0.26)$	$-150.8699(0.51)$	20.85805
40	$-130.0455(0.06)$	$-151.1534(0.10)$	21.10788
100	$-130.0428(0.01)$	$-150.2362(0.01)$	20.19340
$\beta, \gamma, \delta \sim \text{Exp}(0.01)$			
10	$-131.9694(12.45)$	$-150.7102(9.39)$	18.74078
20	$-137.0698(1.93)$	$-151.9821(1.66)$	14.91236
40	$-137.4886(0.42)$	$-153.0443(0.38)$	15.5557
100	$-137.3855(0.07)$	$-152.363(0.06)$	14.97752

Table 4.3: Results of $\log(\pi(\boldsymbol{R}|m_1))$, $\log(\pi(\boldsymbol{R}|m_2))$ and $\log(BF_{12})$ using data simulated based on the SIR model with $T_I \sim \text{Exp}(\gamma)$ (m_1). The bias (the gap between the standard and updated power posterior estimates of the marginal likelihood) was recorded under different values of r. β, γ and δ were assigned $\text{Exp}(1)$ (top table) and $\text{Exp}(0.01)$ (bottom table) prior distributions and various choices of temperature ladder were applied to determine their effects on the accuracy of estimating $\log(BF_{12})$.

Further to that, it is clear that as the prior of the model parameters gets diffuse, the bias term in the updated version of power posterior (the gap between the standard and updated power posterior estimates of the marginal likelihood) gets large. This is an indication of the potential gains that can be made when estimating the evidence using power posteriors by correcting the numerical

integration error. Also, it confirms the need to use large r points, ensuring high frequency near $t = 0$ in order to estimate the Bayes factor accurately in the case of employing vague priors.

In terms of the impact of prior's diffuseness, our discussion in Section 4.3.1 is confirmed by these figures in which the $\log(BF_{12})$ value decreases as the prior uncertainty increases.

- **Choice of prior distribution**

It is well known that Bayes factors can exhibit strong dependence on the model parameter prior distributions. Here, we explore this issue via simulations. We considered two simulated datasets generated under the two SIR models described earlier, and we varied the informativeness of the model parameter prior distributions to determine the effect on estimating Bayes factors and DIC_6. Based on the results of the previous simulation study for the optimal choice of temperature ladder, we shall use $r = 20, 20$ and 40 when employing $\text{Exp}(1), \text{Exp}(0.1)$ and $\text{Exp}(0.01)$ priors respectively.

In the first scenario, based on model m_1, some removal data were generated in a population of $\boldsymbol{N} = 50$ individuals. We set the infection and removal rates to be $\beta = 2$ and $\gamma = 1$ respectively. There were 41 observed removal times. In the second scenario, 22 removal times were simulated under model m_2 with $\boldsymbol{N} = 30, \beta = 2, \alpha = 10$ and $\delta = 10$. Then, the SIR model with $T_I \sim \text{Exp}(\gamma)$ and $T_I \sim \text{Gamma}(10, \delta)$ was fitted to both datasets using the power posterior approach in order to estimate the logarithm of Bayes factors and DIC_6 under different prior assumptions for β, γ and δ. Results of the simulation study are summarised in Table 4.4. The expected log-likelihood curves under each power posterior for both SIR models when the prior is $\text{Exp}(1)$ are displayed in Figure 4.3. Under the first scenario, the results in Table 4.4 indicate, in general, that $\log(BF_{12})$ values support model m_1 (the true model) at all prior assumptions with a noticeable decrease as the prior gets diffuse which is in agreement with the behaviour of the Bayes factor for complete data case de-

Model	Prior	$\log(BF_{12})$	DIC_6	
			m_1	m_2
m_1	Exp(1)	22.48181	**276.2134**	291.008
m_1	Exp(0.1)	10.00648	261.2552	**248.0416**
m_1	Exp(0.01)	9.266824	261.2972	**247.7278**
m_2	Exp(1)	14.53045	**123.6033**	127.7557
m_2	Exp(0.1)	−0.03421352	115.0171	**105.065**
m_2	Exp(0.01)	−0.8887804	114.6325	**102.7094**

Table 4.4: Results of $\log(BF_{12})$ and DIC_6 using two datasets simulated based on models m_1 (the SIR with $T_I \sim \text{Exp}(\gamma)$) and m_2 (the SIR with $T_I \sim \text{Gamma}(\alpha = 10, \delta)$). For each case, β, γ and δ were assigned prior distributions to varying degrees of diffuseness to determine the effect on estimating $\log(BF_{12})$ and DIC_6.

scribed earlier in Section 4.3.1.

In terms of the second case where the data were generated under model m_2, we can see that the value of $\log(BF_{12})$ varies quite dramatically with different priors. Increasing the diffuseness of the prior results in identifying m_2 correctly, whereas using Exp(1) prior gives rise to m_1 offering the best fit. There is some intuition to explain this conflict, as follows. The true value of δ parameter used in the simulation is 10, so an Exp(1) prior is a strong prior distribution which is in conflict with the data and in turn results in poor posterior mean estimates for both β and δ, that are $\widehat{\beta} = 1.2859$ and $\widehat{\delta} = 6.6288$, and this leads to lower values of $\log(\pi(\boldsymbol{R}|m_2))$ and hence m_2 being not identified correctly. However, with Exp(0.1) and Exp(0.01) priors, the estimation was improved giving $\widehat{\beta} = 2.0530, \widehat{\delta} = 11.4900$ and $\widehat{\beta} = 2.1847, \widehat{\delta} = 12.1327$,

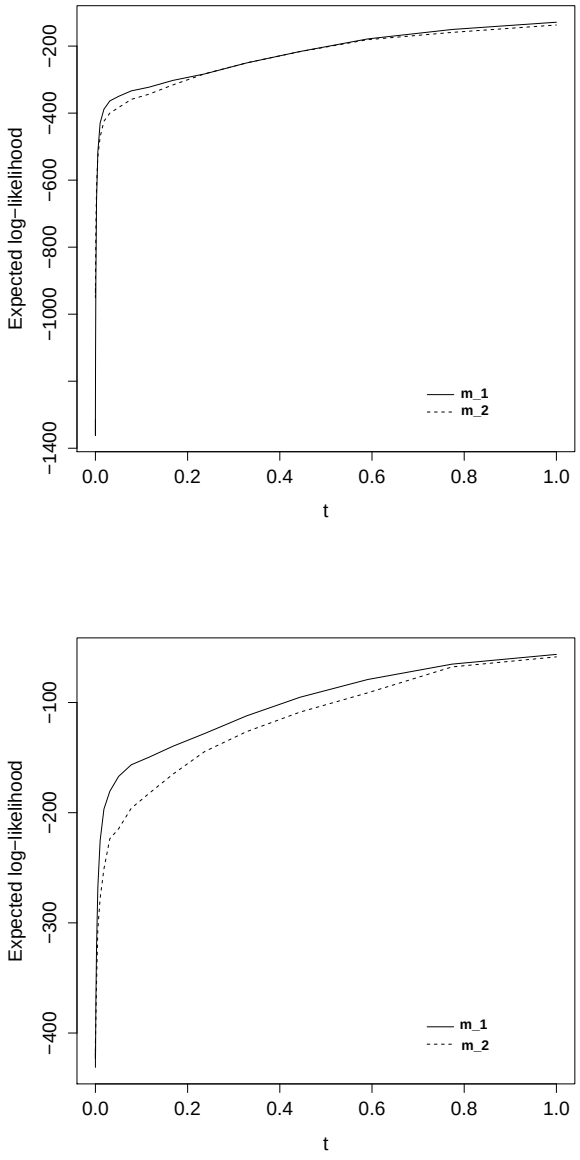


Figure 4.3: Top plot: the expected log-likelihood curves for the removal data generated under the SIR with $T_I \sim \text{Exp}(\gamma)$ (m_1) calculated using the updated power posterior approach when the prior for β, γ and δ is $\text{Exp}(1)$. Bottom plot: the expected log-likelihood curves for the removal data generated under the SIR with $T_I \sim \text{Gamma}(\alpha = 10, \delta)$ (m_2) calculated using the updated power posterior approach with $\text{Exp}(1)$ prior for β, γ and δ.

respectively and consequently the correct identification of model m_2 was obtained. This illustrates the need for caution when placing prior distributions

on model parameters.

The values of DIC_6 are dependent on the priors as well. At prior $Exp(1)$, model m_1 is supported by DIC_6 values in both dataset scenarios. However, it can be seen that increasing prior uncertainty of the parameters reduces the value of the DIC_6 under both models and leads to the selection of model m_2 as the better model under the two settings.

As the validity of the Bayes factor approximation depends crucially on convergence of the collection of Markov chains, we observed that when updating $\boldsymbol{I}$ by sampling I_j^*, as a proposed value of I_j, uniformly on (I_z, R_j) caused some mixing problems in the power posterior when t between 0.116 and 0.327. However, the problem was improved by proposing I_j^*, such that $R_j - I_j^* \sim$ Gamma(α, δ^{mle}), where $\delta^{mle} = \alpha n_R / \sum_{j=1}^{n_R}(R_j - I_j)$ evaluated at each MCMC iteration.

4.4.3 Application of the updated power posterior and DIC_6 methods to the SIR model with different infection mechanisms

We now consider another different epidemic setting in which two Markovian SIR models with different forms of infection rate are compared. The first model denoted by m_1 is an SIR model in which the rate of new infections is assumed to have a linear relation with the number of susceptible individuals multiplied by the number of infective individuals at time t, that is $\beta n^{-1} X(t) Y(t)$. The other epidemic model denoted by m_2 is a modification of m_1 model by relaxing the linearity assumption and considering the rate of new infections to be of the form $\beta n^{-1} X(t) Y^p(t),\ 0 < p < 0.5$.

Suppose that an epidemic outbreak which results in a total of n_R removals in a closed population of size n initial susceptibles and one initial infective is observed on $(I_1, R_{n_R}]$, so that the entire epidemic is assumed to be observed. Let $R_1 = 0, R_2, ..., R_{n_R}$ be the observed ordered removal times and

write $\boldsymbol{R} = (R_1, ..., R_{n_R})$. Also, let $I_1 < 0$ be the first unobserved infection time and $\boldsymbol{I} = (I_2, ..., I_{n_R})$ be the rest of unobserved ordered infection times, where $I_k < I_{k+1} < R_k$ for $k = 1, 2, ..., n_R - 1$. Then, the joint likelihoods of $\boldsymbol{I}$ and $\boldsymbol{R}$ given the model parameters are

$$\begin{aligned}\pi(\boldsymbol{I}, \boldsymbol{R}|\beta, \gamma, I_1, m_1) &= \prod_{j=2}^{n_R} \beta n^{-1} X(I_j-) Y(I_j-) \times \prod_{j=1}^{n_R} \gamma Y(R_j-) \\ &\times \exp\left(- \int_{I_1}^{R_{n_R}} \left(\beta n^{-1} X(t) Y(t) + \gamma Y(t) \right) dt \right),\end{aligned}$$

and

$$\begin{aligned}\pi(\boldsymbol{I}, \boldsymbol{R}|\beta, \gamma, p, I_1, m_2) &= \prod_{j=2}^{n_R} \beta n^{-1} X(I_j-) Y^p(I_j-) \times \prod_{j=1}^{n_R} \gamma Y(R_j-) \\ &\times \exp\left(- \int_{I_1}^{R_{n_R}} \left(\beta n^{-1} X(t) Y^p(t) + \gamma Y(t) \right) dt \right).\end{aligned}$$

We assign an independent exponential prior distribution $\text{Exp}(\lambda_\zeta)$ for β and γ, where $\zeta = \beta,\ \gamma$, and assume $p \sim U(0, 0.5)$. The first infection time I_1 is also assigned a prior, in particular $-I_1 \sim \text{Exp}(\theta)$.

Then, for model m_2 at temperature t (model m_1 follows the same lines when setting $p = 1$), we have the following full conditional power posterior densities for β, γ and I_1

$$\beta|t, \gamma, p, I_1, \boldsymbol{I}, \boldsymbol{R} \sim \text{Gamma}\left(1 + t(n_R - 1), \lambda_\beta + t n^{-1} \int_{I_1}^{R_{n_R}} X(t) Y^p(t) dt \right),$$

$$\gamma|t, \beta, p, I_1, \boldsymbol{I}, \boldsymbol{R} \sim \text{Gamma}\left(1 + n_R t, \lambda_\gamma + t \int_{I_1}^{R_{n_R}} Y(t) dt \right),$$

$$I_1|t, \beta, \gamma, p, \boldsymbol{I}, \boldsymbol{R} \sim I_2 - \text{Exp}\left(t(\beta + \gamma) + \theta \right),$$

and the following proportional full conditional power posterior densities for p and $\mathbf{I}$:

$$\pi(p|t, \beta, \gamma, I_1, \boldsymbol{I}, \boldsymbol{R}) \propto \left\{ \left(\prod_{j=2}^{n_R} Y^p(I_j-) \right) \times \exp\left(-\beta n^{-1} \int_{I_1}^{R_{n_R}} X(t) Y^p(t) dt \right) \right\}^t,$$

$$\begin{aligned}\pi(\boldsymbol{I}|t,\beta,\gamma,p,I_1,\boldsymbol{R}) &\propto \left\{\prod_{j=2}^{n_R} X(I_j-)Y^p(I_j-)\times\prod_{j=1}^{n_R} Y(R_j-)\right\}^t \\ &\times \left\{\exp\left(-\int_{I_1}^{R_{n_R}}\left(\beta n^{-1}X(t)Y^p(t)+\gamma Y(t)\right)dt\right)\right\}^t.\end{aligned}$$

To sample from the above densities we simply use a modified version of our MCMC algorithm described in Chapter 3 in which the likelihoods are now raised to the power t. Estimation the log marginal likelihoods are obtained using the trapezoidal rule (using $t_j = (j/r)^c$, with $r = 20$ or 40 and $c = 5$, so that more points are placed toward the prior than the posterior) according to Algorithm 8 and hence the logarithm of Bayes factor is approximated using (4.4.4).

The DIC_6 calculation can be done as follows. Having derived the posterior estimates for each model parameters and each model initial infection time from the power posterior at $t = 1$, we run the data-augmented MCMC algorithm with each set of model parameters and each initial infection time fixed to their estimated values, allowing only the missing data $\boldsymbol{I}$ to vary which allows us to derive DIC_6 for each model.

4.4.3.1 Factors influencing the updated power posterior and DIC_6 performances

To illustrate the performance of our proposed methods (the power posterior and the DIC_6) against the amount of discretization and the diffuseness of the prior distribution, we conducted several simulation studies based on simulated data generated under the two SIR models described earlier. Within each temperature t, we collected $27,000$ samples from the power posterior of which the first $2,000$ were discarded as burn-in and the rest were thinned by taking each 5th value. In terms of DIC_6 calculation, having derived the posterior point estimates for β, γ and I_1 using the power posterior at $t = 1$ and allowing only the augmented data $\boldsymbol{I}$ to vary, we run a within-model data-augmented MCMC algorithm for each model for $17,000$ iterations with dis-

regarding a burn-in sample of 2,000 and thinning the chain by taking the 5th value. Various choices of temperature ladder and different prior assumptions were considered to determine their effects on estimating Bayes factors and DIC_6. We present results based on one typical simulated dataset for each scenario. Throughout, we assumed that $-I_1 \sim \text{Exp}(1)$ and $p \sim U(0, 0.5)$.

- **Specification of the amount of discretization**

Estimating the log marginal likelihood via the power posterior is relatively straightforward but might be computationally costly. To minimise the cost, it is desirable to use as few t values as possible. Here we examined the amount of discretization r in the interval $(0, 1]$ that required to improve the efficiency of estimating $\log(BF_{12})$ when more informative or less informative priors for β and γ are used, such as $\text{Exp}(1)$ and $\text{Exp}(0.01)$. A population of size $\boldsymbol{N} = 100$ was considered and model m_2 was used with parameters $\beta = 2$, $\gamma = 0.2$ and $p = 0.3$ to generate 87 removal times. Then, $\log(BF_{12})$ was estimated via the power posterior approach under different values of r. Results of the estimation are summarised in Table 4.5.

As demonstrated earlier for the SIR model with different infectious periods, it is obvious that as the prior uncertainty of the model parameters increases, the the bias term in the updated version of power posterior (4.4.3) also increases. In other words, assigning an $\text{Exp}(1)$ prior in calculating the Bayes factor leads to a reduction in the amount of discretization needed for accurate estimation and consequently the computational cost. However, using diffuse prior distributions may require larger values of r in order to obtain reasonably accurate results. Again, it was also found that the variance of $E_{\boldsymbol{\theta},\boldsymbol{I}|\boldsymbol{R},t} \log[\pi(\boldsymbol{R}, \boldsymbol{I}|\boldsymbol{\theta})]$ is largest as $t \to 0$, suggesting that more MCMC iterations should be allocated as $t \to 0$ rather than as $t \to 1$ to reasonably obtain good estimates.

The effect of the prior's diffuseness on Bayes factors is notable which agrees with our discussion in Section 4.3.2, that is as the prior uncertainty increases the $\log(BF_{12})$ value decreases.

r	$\log(\pi(\boldsymbol{R}\|m_1))$ (Bias)	$\log(\pi(\boldsymbol{R}\|m_2))$ (Bias)	$\log(BF_{12})$
$\beta, \gamma \sim \text{Exp}(1)$			
10	$-113.3533(3.15)$	$-104.2926(2.41)$	-9.0607
20	$-114.0504(0.67)$	$-106.0651(0.54)$	-7.985566
40	$-114.0871(0.16)$	$-105.7564(0.14)$	-8.33061
100	$-113.8002(0.03)$	$-105.8681(0.02)$	-7.932046
$\beta, \gamma \sim \text{Exp}(0.01)$			
10	$-81.08691(60.05)$	$-85.72172(37.77)$	4.634815
20	$-120.6974(5.02)$	$-111.0303(3.61)$	-9.667144
40	$-122.2193(1.00)$	$-111.5856(0.72)$	-10.63368
100	$-122.2827(0.15)$	$-112.1095(0.11)$	-10.17321

Table 4.5: Results of $\log(\pi(\boldsymbol{R}|m_1))$, $\log(\pi(\boldsymbol{R}|m_2))$, $\log(BF_{12})$ and the bias (the gap between the standard and updated power posterior estimates) using data simulated based on the modified SIR model (m_2). The parameters β and γ were assigned $\text{Exp}(1)$ (top table) and $\text{Exp}(0.01)$ (bottom table) prior distributions and various choices of temperature ladder were applied to determine their effects on the accuracy of estimating $\log(BF_{12})$. A non-informative prior $U(0, 0.5)$ was assumed for p.

- **Specification of the prior distribution**

Prior sensitivity analysis was conducted on the parameters of the two models except the parameter p, for which we used a non-informative $U(0, 0.5)$

throughout the analysis. We start from assuming a more informative prior $\text{Exp}(1)$, moving to a less informative prior $\text{Exp}(0.01)$. Two datasets were simulated based upon the two candidate SIR models,with varying the informativeness of the model parameter prior distributions to determine the effect on estimating Bayes factors and DIC_6. In terms of the amount of discretization (r), we shall use $r = 20, 40$ and 40 when employing $\text{Exp}(1), \text{Exp}(0.1)$ and $\text{Exp}(0.01)$ priors respectively.

In the first scenario, some removal data were generated under model m_1. For a population of $\boldsymbol{N} = 100$ individuals, we set the contact and removal rates to be $\beta = 0.5$ and $\gamma = 0.2$ respectively. There were 83 observed removal times. In the second scenario, 88 removal times were simulated under model m_2 with $\boldsymbol{N} = 100, \beta = 2.5, \gamma = 0.2$ and $p = 0.3$. Then, we fitted both SIR models to these two datasets using the power posterior method in order to estimate the logarithm of Bayes factors and DIC_6 under different prior assumptions.

Results of simulations are displayed in Table 4.6. Also, Figure 4.4 shows plots of the expected log-likelihoods against the temperature t under each power posterior for both models when the prior is $\text{Exp}(1)$.

The results in Table 4.6 reveal that the estimated log Bayes factors behave as one might expect, giving decisive evidence in favour of the generated-data model under all priors. However, it is evident that as the prior uncertainty gets larger, the $\log(BF_{12})$ value becomes smaller which is not entirely surprising as discussed in Section 4.3.2.

In terms of the performance of DIC_6, while the values of the DIC_6 were unaffected by the choice of prior distribution, the model selection is generally successful and the data-generated model was correctly identified in both scenarios. However, the difference between the DIC_6 values for the correct model and the misspecified model makes it unclear to confirm any evidence of poor fit.

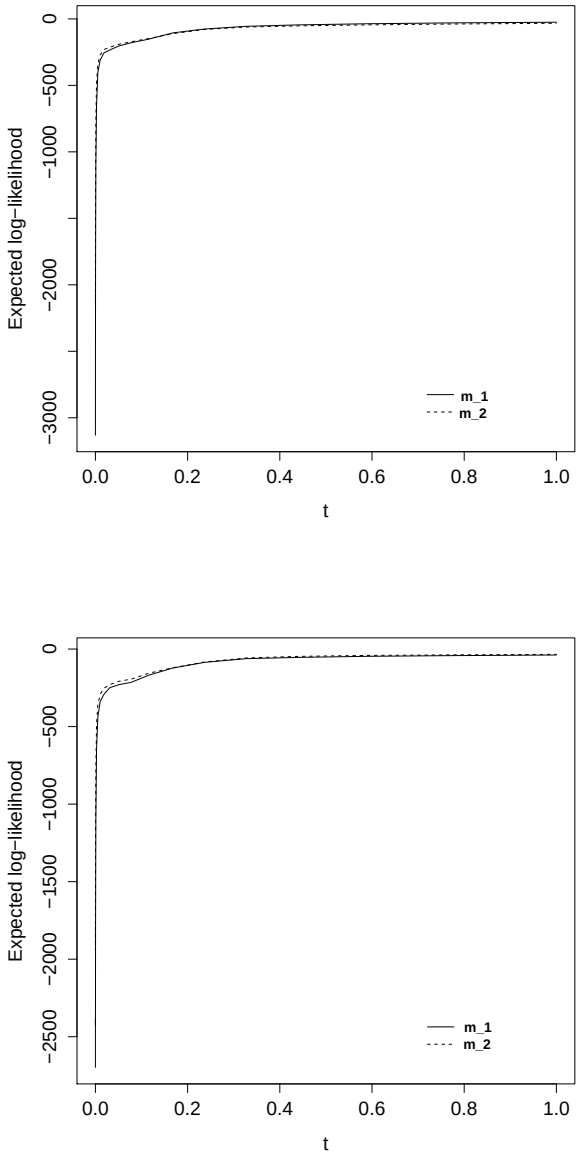


Figure 4.4: Top plot: the expected log-likelihood curves for the removal data generated under the standard SIR (m_1) calculated using the updated power posterior approach when the prior is Exp(1) for β and γ. Bottom plot: the expected log-likelihood curves for the removal data generated under the modified SIR (m_2) calculated using the updated power posterior approach with Exp(1) prior for β and γ and $U(0, 0.5)$ for p.

Model	Prior	$\log(BF_{12})$	DIC_6	
			m_1	m_2
m_1	Exp(1)	3.489551	**51.49038**	68.8904
m_1	Exp(0.1)	1.940342	**52.07774**	68.09464
m_1	Exp(0.01)	2.097549	**51.73746**	68.66254
m_2	Exp(1)	−8.030975	76.56896	**68.56952**
m_2	Exp(0.1)	−9.787747	76.98116	**69.9391**
m_2	Exp(0.01)	−10.69824	77.37168	**69.5956**

Table 4.6: Results of $\log(BF_{12})$ and DIC_6 using data simulated based on the standard SIR model (m_1) and the modified SIR model (m_2). For each case, the parameters β and γ were assigned prior distributions to varying degrees of diffuseness to determine the effect on estimating $\log(BF_{12})$ and DIC_6. A non-informative prior $U(0, 0.5)$ was assumed for p.

4.5 Epidemic model choice: aspects where epidemic size matters

In this section, the effect of epidemic size on distinguishing between the SIR models using the Bayes factor criterion was investigated via simulation. Two epidemic settings were looked at, namely the SIR epidemic model with different infectious periods and the SIR epidemic model with different infection mechanisms. Bayes factors were calculated based on complete and incomplete outbreak data. The updated power posterior approach was implemented to estimate Bayes factors for partially observed data.

4.5.1 The SIR model with different infectious periods

Here, the SIR model with exponential infectious period ($T_I \sim \text{Exp}(\gamma)$) denoted by m_1 is compared to the SIR model with gamma infectious period ($T_I \sim \text{Gamma}(\alpha, \delta)$) denoted by m_2, where the shape parameter of gamma distribution α is assumed to be known. We considered a population of size $N = 50$ ($n = 49$ initial susceptibles and one initial infective). The variability in simulating realisations of an epidemic should be taken into account and to address this issue to some extent, we simulated, under model m_1 with parameters $\beta = 1.15$ and $\gamma = 1$, 20 outbreak datasets of size 5 removals each, which corresponds to a minor epidemic occurrence. Then, $\log(BF_{12})$ was calculated via (4.3.1) for complete data (infection and removal times) case, and for incomplete data (removal times) case, $\log(BF_{12})$ was estimated via the updated power posterior method using 21 rungs. The gamma shape parameter was assumed to be $\alpha = 10$. In both cases $\log(BF_{12})$ values were computed under two different prior assumptions for β, γ and δ, namely $\text{Exp}(1)$ and $\text{Exp}(0.01)$ priors.

The results are summarised by boxplots in Figure 4.5. Interestingly, the results are not as one might expect with a small outbreak dataset, giving decisive support to the true model m_1 under both complete and incomplete data. These findings suggest that, in this particular epidemic setting, a few infectious periods of infected individuals might be enough for the Bayes factor criterion to favour the SIR model with exponential infectious period over the SIR model with gamma infectious period. The impact of the parameter prior distributions is also evident which is in agreement with our discussion in Section 4.3.1.

4.5.2 The SIR model with different infection mechanisms

In what follows, the effect of outbreak size on distinguishing between two Markovian SIR models with different forms of infection rate using Bayes factor tool was investigated via simulation. The first model denoted by m_1 is an SIR

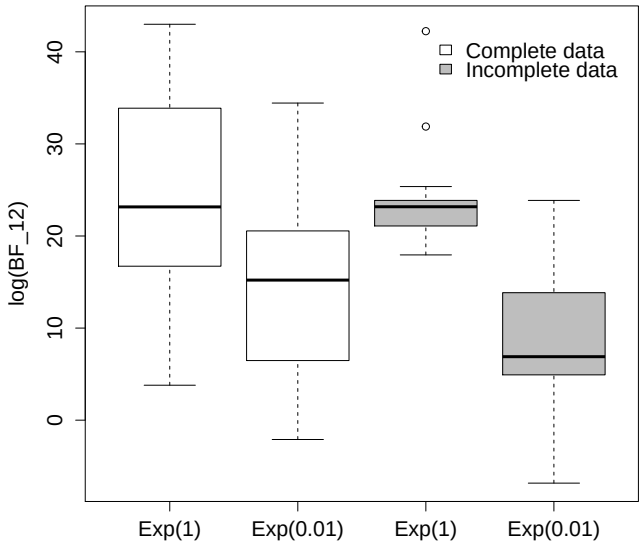


Figure 4.5: Boxplots of $\log(BF_{12})$ values calculated based on 20 outbreak datasets of size 5 removals each simulated under the SIR model with $T_I \sim \text{Exp}(\gamma)$ (m_1). $\log(BF_{12})$ was computed via (4.3.2) and the updated power posterior method for complete and incomplete data respectively. The parameters β, γ and δ were assigned two choices of prior, namely $\text{Exp}(1)$ and $\text{Exp}(0.01)$.

model in which the rate of new infections is assumed to be $\beta n^{-1} X(t) Y(t)$. The other epidemic model m_2 is a modified version of m_1 model by assuming the rate of new infections to be $\beta n^{-1} X(t) Y^p(t)$, where $p \in (0, 0.5)$ is assumed to be known. Two epidemic scenarios were looked at and in each scenario 20 outbreak datasets were generated using model m_1. For the complete data case, $\log(BF_{12})$ was calculated via (4.3.5) and for partial data it was estimated by the updated power posterior approach using 21 points of t. The computation of $\log(BF_{12})$ were performed under different parameter prior assumptions. We considered a population of size $\boldsymbol{N} = 50$ ($n = 49$ initial susceptibles and one initial infective). The power parameter p was fixed to $p = 0.3$.

In the first scenario, 20 datasets each of size 7 removals were generated using model m_1 with parameters $\beta = 1.15$ and $\gamma = 1$ which corresponds to a minor epidemic. Then, $\log(BF_{12})$ was obtained for each dataset using both complete

and incomplete data under two different prior settings. The $\log(BF_{12})$ values are presented by boxplots as in Figure 4.6. As expected with a small outbreak dataset, it can be seen that on average, the logarithm of Bayes factor values, calculated using complete data, do not give clear support for either the true model m_1 or the incorrect model m_2 under $\text{Exp}(1)$ prior but model m_2 is substantially favoured with $\text{Exp}(0.01)$ prior. However, the impact of the prior's diffuseness on estimating $\log(BF_{12})$ is obvious which is consistent with our discussion in Section 4.3.2, giving roughly strong support to model m_2.

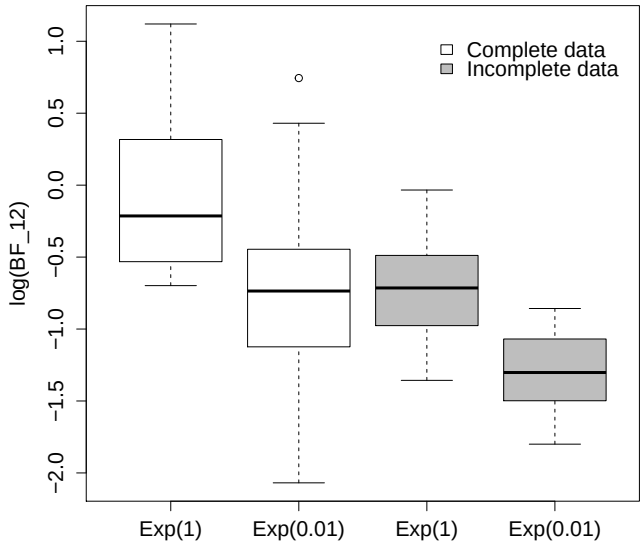


Figure 4.6: Boxplots of $\log(BF_{12})$ values calculated using 20 outbreak datasets of size 7 removals each simulated under the standard SIR model (m_1) and evaluated with setting $p = 0.3$ and assigning $\text{Exp}(1)$ and $\text{Exp}(0.01)$ priors for β and γ. $\log(BF_{12})$ values were obtained via (4.3.5) and the updated power posterior method for complete and incomplete data respectively.

In the second scenario, again under model m_1 with parameters $\beta = 2$ and $\gamma = 1$ (i.e. $R_0 = 2$), we simulated 20 outbreak datasets in a population of size $N = 50$. The infection and removal times of 42 cases were observed including the initial one and that represents a major epidemic occurrence. Then, we repeated the calculation of the first scenario with these datasets. Summary of

the results is provided in Figure 4.7. As now we have more data compared to the data in the first scenario, we expect Bayes factors to give some support to the correct model m_1. Under Exp(1) prior, m_1 is clearly identified with the complete data, whereas when using only the removal times the evidence is not that clear. Also, the impact of employing a vague prior is quite obvious in reducing $\log(BF_{12})$ values.

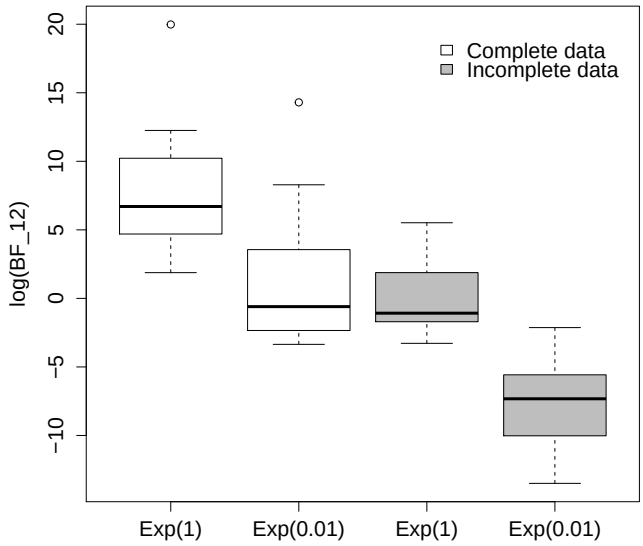


Figure 4.7: Boxplots of $\log(BF_{12})$ values calculated using 20 outbreak datasets of size 42 removals each simulated under the standard SIR model (m_1) and evaluated with setting $p = 0.3$ and assigning Exp(1) and Exp(0.01) priors for β and γ. $\log(BF_{12})$ values were obtained via (4.3.5) and the updated power posterior method for complete and incomplete data respectively.

In the second scenario above, the basic reproduction number R_0 was 2 which roughly speaking means the epidemic at its onset did not take off sharply and there were a small number of infected cases around before a new infection occurs. This behaviour in turn implies that the quantity $\prod_{j=2}^{n_R} Y^{1-p}(I_j-)$ is not that large and having known that $A_p = \int_{I_1}^{R_{n_R}} X(t)Y^p(t)\,dt \ll \int_{I_1}^{R_{n_R}} X(t)Y(t)\,dt = A$ when $p = 0.3$, the $\log(BF_{12})$ will tend to be less than zero, making the correct identification of the true model m_1 hard if not impossible.

However, with this particular population size where $N = 50$ (of course, as $N \to \infty$, we expect much more improvement in identifying the correct model, see for example our data in Section 4.4.3 with $N = 100$), one epidemic setting that enables model m_1 to be identified particularly well, at least under the Exp(1) prior, is that where R_0 is relatively large and so most of the time there are many infectives present when a new infection takes place. This setting allows the quantity $\prod_{j=2}^{n_R} Y^{1-p}(I_j-)$ to be large enough to increase most of the $\log(BF_{12})$ values. To see that, 20 outbreak datasets of size 47 removals each were simulated under model m_1 with parameters $\beta = 5$ and $\gamma = 1$ (i.e. $R_0 = 5$). Then, $\log(BF_{12})$ was estimated via the power posterior approach for each dataset with $p = 0.3$ and Exp(1) prior for β and γ. Figure 4.8 shows a histogram of $\log(BF_{12})$ values from which the evidence of decisively favouring the correct model m_1 is apparently clear.

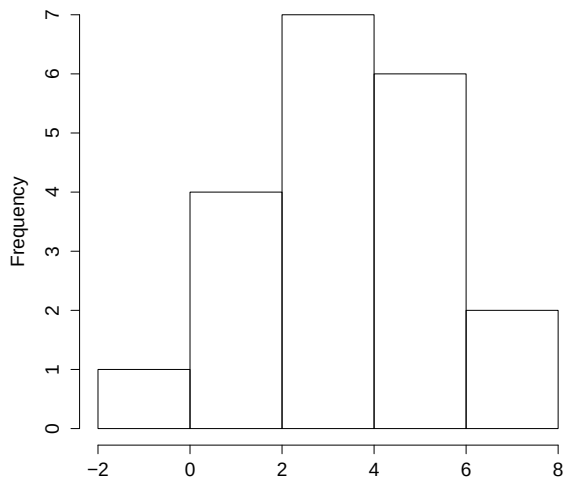


Figure 4.8: Histogram of the estimated $\log(BF_{12})$ values obtained via the updated power posterior approach based on 20 outbreak datasets of size 47 removal times each simulated under the standard SIR model (m_1) and evaluated with setting $p = 0.3$ and assigning an Exp(1) prior for β and γ.

4.6 Summary and extensions

4.6.1 Summary

In this chapter we had four main objectives; (i) to derive explicit forms for Bayes factors in different epidemic settings in case of complete epidemic data and then examine the performance of the (simple) prior arithmetic mean estimator for approximating the Bayes factor , (ii) to extend and implement the power posterior approach for estimating the Bayes factor in the epidemic context where the epidemic data are only partially observed, (iii) to study the ability of DIC_6 to select between epidemic models based on removal data and finally (iv) to look at some model selection aspects where outbreak size can have major impacts on distinguishing between the SIR models using Bayes factors. Throughout, different transmission models were compared, namely the SIR model with different infectious periods and the SIR model with different infection mechanisms. The performance of the power posterior and DIC_6 methods were tested against various factors such as prior's diffuseness and temperature schedule using several simulated datasets.

Bayes factor expressions that we derived based on complete outbreak data were of particular assistance in gaining some insights into what estimated values of Bayes factors might be expected in case of incomplete epidemic data, specifically under different prior assumptions. While the implementation of the prior arithmetic mean estimator of the marginal likelihood was based on the assumption of complete outbreak data, it appeared to be a reliable estimator in cases where a low-dimensional prior is considered.

The recently developed modified version of power posterior approach (Friel et al., 2014) is one of the most sophisticated methods in the literature for estimating the marginal likelihood and hence the Bayes factor. The estimate of the Bayes factor obtained via this method consistently outperforms the standard estimate derived in Friel and Pettitt (2008) with no additional computation cost. We extended this approach into models incorporating missing

data and demonstrated how this approach can easily be implemented as an addition to existing MCMC algorithms in two different epidemic settings in order to estimate the logarithm of Bayes factor in applications to simulated outbreak data. An obvious advantage of this method is that it computes the marginal likelihoods and hence Bayes factors on the log-scale, thereby offering numerical stability. Moreover, unlike other methods, it does not require much tuning and it involves a minimal amount of change to the commonly used MCMC algorithms that compute the posterior distributions of unknown epidemic model parameters which makes it a good alternative to other Bayesian epidemic model selection methods such as the RJMCMC method. However, one difficulty with this approach is that of choosing the optimal temperature schedule, and this could be viewed as a weakness, although Behrens et al. (2012), using method of tempered transitions, offered some possible directions for effectively selecting the $\{t_j\}$ for power posteriors. Also, with a view of fitting epidemic models with diffuse priors, it is unsurprising that it may take some extra time and effort to obtain reliable estimates.

As Bayes factors are sensitive to prior choices, we have illustrated via simulations how varying the parameters of the prior distributions of the model parameters, in particular the infection rate, can affect this diagnostic tool for epidemic model selection.

In our epidemic setting, we have missing data (the times when individuals are infected), so we adopted a particular form of DIC, that is DIC_6 in Celeux et al. (2006). The DIC_6 measure has some advantages in terms of computation. In principle, the calculation of this quantity requires running the data-augmented MCMC scheme twice. However, as we compare the performance of this measure to the Bayes factor obtained via the power posterior method, some quantities in DIC_6 form are readily available from the power posterior approach at temperature $t = 1$ at no extra computational cost. Nevertheless, there are possible drawbacks of employing DIC_6 as a model selection tool. First, the DIC_6 uses $\widehat{\boldsymbol{\theta}}$ as posterior point estimates for $\boldsymbol{\theta}$ in its calculation, which

violates our fully Bayesian framework. Furthermore, the difficulties associated with interpreting the DIC_6 values can be seen as another disadvantage of using it.

We have highlighted the importance of outbreak size in distinguishing between epidemic models. It was found that a few infectious periods of infected individuals might be enough for Bayes factor criterion to favour the SIR model with exponential infectious period over the SIR model with gamma infectious period. For the SIR model with different infection mechanisms, simulation results revealed, particularly under a diffuse prior, the failure of the Bayes factor as a model selection indicator in case of either a minor epidemic or a small population size and the need of having large population with a major outbreak in order to differentiate between models.

4.6.2 Extensions

As long as MCMC algorithms that enable sampling from the posterior distributions of unknown epidemic model parameters are available, the method we have developed here could be extended to various epidemic scenarios. Model choice problems for epidemic models with latent periods, multilevel models and random network models can be addressed via our power posterior extension method. Also, our methodology can be applied to various types of epidemic datasets such as final size data, surveillance hospital data and time-series data (e.g. number of observed cases each day or week, etc).

Overall, the Bayes factor as a tool of epidemic model choice is generally a difficult quantity to estimate. However, our missing data-extension of the power posterior approach offers a generic and attractive technique for estimating it in the context of partially observed epidemic data although it may take some effort in terms of computational run time to be implemented if non-informative priors employed.

CHAPTER 5

Conclusions

Much research has been undertaken in the past decades to develop sophisticated tools (e.g. MCMC methods) to estimate epidemic model parameters such as infection and removal rates from data on epidemics. However, far less attention has been devoted to model adequacy and model choice for stochastic epidemic models. In this thesis, we successfully developed, under a Bayesian framework, novel methods for both model adequacy and model choice for stochastic epidemic models based on partially observed outbreak data.

In Chapter 2, we systematically addressed the problem of model adequacy for the SIR epidemic models. We specifically considered three SIR models routinely employed in the literature, namely the model with exponential infectious periods, the model with gamma infectious periods and the model with fixed length infectious periods. We investigated three different but natural methods to assess the goodness-of-fit for the SIR models. Given only data on removals, the labelled likelihood representation approach was used to examine the distribution of of the random variable $R_j - I_j,\ j = 1, ..., n_R$, (i.e. the imputed infectious periods) over MCMC iterations and this method appeared to be inadequate to discriminate the infectious period distribution of the SIR models. Based on the assumption of complete epidemic data, we employed the time-rescaling theorem (Papangelou, 1972) to construct a goodness-of-fit method for the Markovian SIR model and attempted to discriminate between

the Markovian SIR model and two other non-Markovian SIR models. We found that time scaling of the inter-event times were insufficient to detect the data-generating SIR model. At the end of the chapter, two purely temporal statistics that based on observing the removal process only were examined to assess the goodness-of-fit for the SIR models via the posterior predictive distribution approach. As seen with the posterior predictive $p-$values results, none of these test statistics performed well in terms of detecting misspecification of the data-generating model. All methods were tested using several simulated datasets.

In Chapter 3, we further investigated the topic of model adequacy for stochastic epidemic models. We developed a two-stage procedure to assess the goodness-of-fit for the SIR stochastic epidemic models. In our two-stage method, two predictive distributions were examined, namely the predictive distribution of the final size of the epidemic and the predictive distribution of the removal times. The method explored how SIR models with different infection mechanisms, infectious periods and population structures can be assessed and distinguished given only a set of removal times of infected individuals in the population. The idea was based on looking explicitly at the discrepancy between the observed and predicted removal data using the posterior predictive model checking procedure in which the notion of Bayesian residuals and the posterior predictive $p-$value were used. Using two simulated outbreak removal datasets and according to our two-stage method, SIR models with different infection mechanisms can be clearly discriminated. For SIR models with different infectious periods, simulated data results explored the capability of the two-stage method to identify the correct model, although results revealed that the gamma and constant SIR models were indistinguishable when the shape parameter of the former is large as assumed in our setting. For the smallpox outbreak dataset, we fitted both SIR and SEIR models with different infectious periods and then our two-stage method was applied. According to our method, non of these epidemic models seems be appropriate to fit these smallpox data and no particular choice of infectious period distribution was

favoured by the data. In the last epidemic setting, model assessment task was performed for the SIR model with different population structures using two simulated outbreak datasets. Although the two-stage method was able to distinguish the homogeneous (one level of mixing) SIR model from the heterogeneous (two levels of mixing) SIR model, inferring the posterior distributions of number of household and community infections also appeared to be an effective tool for detecting any heterogeneity in the observed data.

In Chapter 4, the problem of choosing, within a Bayesian modelling framework, from a finite specified set of stochastic epidemic models given temporal removal data was considered. Throughout, different transmission models were compared, namely the SIR model with different infectious periods and the SIR model with different infection mechanisms. We extended the recently developed modified version of power posterior approach (Friel et al., 2014) into models incorporating missing data and demonstrated how this approach can easily be implemented as an addition to existing MCMC algorithms in the two aforementioned epidemic settings in order to estimate the logarithm of Bayes factor in applications to simulated removal outbreak data. We also derived explicit forms for Bayes factors in the aforementioned epidemic settings in case of complete epidemic data and then examined the performance of the simple prior arithmetic mean estimator for approximating the Bayes factor. Furthermore, we studied the ability of DIC (DIC_6 in Celeux et al. (2006)) to select between epidemic models based on removal data. The performance of both our power posterior extension and DIC_6 methods were tested against various factors such as prior's diffuseness and temperature schedule using several simulated datasets. Explicit Bayes factor expressions were of particular assistance in gaining some insights into what estimated values of Bayes factors might be expected in case of incomplete epidemic data, specifically under different prior assumptions. Additionally, we looked at some model selection aspects where outbreak size can have major impacts on distinguishing between the SIR models using Bayes factor criterion based on complete and incomplete simulated outbreak datasets.

Throughout the whole thesis, epidemics are assumed to be complete and the shape parameter α for the gamma SIR model is also assumed to be known. However, our methods are appropriate to ongoing disease outbreaks analysis and it would be interesting to see the performance of our methods in the case of unknown α.

A possible limitation to our two-stage method is that the model assessment procedure is only as good as the posterior estimation of the parameters since the former task relies on the latter. As long as our estimation is reasonably accurate, the two-stage method is able to differentiate between models. Also, it is not surprising that the process of obtaining the posterior predictive $p-$value may take some effort to implement in terms of computational time, particularly if some of the selected posterior samples produce extreme values of R_0 or R_*.

Our extended version of the power posterior approach is a generic method for estimating the marginal likelihood and hence Bayes factors and is relatively easy to implement. It is certainly of interest to apply this method to other more complex scenarios. Specifically, those include epidemic models with latent periods, multilevel models as well as situations involving different types of epidemic data. However, a possible difficulty with this approach is that of choosing the temperature schedule (i.e. the question of how to select appropriate temperature ladders in practice), although Behrens et al. (2012), using method of tempered transitions, offered some useful information in this direction.

Overall, the methods presented in this thesis represent a step toward applying them in more complex models of the spread of infectious diseases. To conclude, we have provided novel methods for both model adequacy and model selection for stochastic epidemic models. These methods are of great importance to further knowledge of how exactly transmission dynamics work.

Bibliography

Anderson, R. and May, R. (1991). *Infectious diseases of humans: dynamics and control.* New York: Oxford University Press.

Andersson, H. and Britton, T. (2000). *Stochastic epidemic models and their statistical analysis*, volume 4. New York: Springer.

Atkinson, K. and Han, W. (2004). *Elementary numerical analysis*, volume 3. New York: Wiley.

Bailey, N. T. (1975). *The mathematical theory of infectious diseases and its applications*. New York: Charles Griffin & Company Ltd.

Bailey, N. T. and Thomas, A. S. (1971). The estimation of parameters from population data on the general stochastic epidemic. *Theoretical Population Biology*, 2(3):253–270.

Ball, F. (1986). A unified approach to the distribution of total size and total area under the trajectory of infectives in epidemic models. *Advances in Applied Probability*, pages 289–310.

Ball, F., Mollison, D., and Scalia-Tomba, G. (1997). Epidemics with two levels of mixing. *The Annals of Applied Probability*, pages 46–89.

Becker, N. and Yip, P. (1989). Analysis of variations in an infection rate. *Australian Journal of Statistics*, 31(1):42–52.

Becker, N. G. (1989). *Analysis of infectious disease data*, volume 33. CRC Press.

Becker, N. G. and Hasofer, A. M. (1997). Estimation in epidemics with incomplete observations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2):415–429.

Becker, N. G. and Utev, S. (1998). The effect of community structure on the immunity coverage required to prevent epidemics. *Mathematical Biosciences*, 147(1):23–39.

Behrens, G., Friel, N., and Hurn, M. (2012). Tuning tempered transitions. *Statistics and Computing*, 22(1):65–78.

Bernardo, J. and Smith, A. (1994). *Bayesian Theory*. New York: Wiley.

Bjørnstad, O. N., Finkenstädt, B. F., and Grenfell, B. T. (2002). Dynamics of measles epidemics: estimating scaling of transmission rates using a time series SIR model. *Ecological Monographs*, 72(2):169–184.

Boys, R. J. and Giles, P. R. (2007). Bayesian inference for stochastic epidemic models with time-inhomogeneous removal rates. *Journal of Mathematical Biology*, 55(2):223–247.

Britton, T., Kypraios, T., and O'Neill, P. D. (2011). Inference for epidemics with three levels of mixing: methodology and application to a measles outbreak. *Scandinavian Journal of Statistics*, 38(3):578–599.

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.

Brooks, S., Catchpole, E., and Morgan, B. (2000). Bayesian animal survival estimation. *Statistical Science*, pages 357–376.

Brown, E. N., Barbieri, R., Ventura, V., Kass, R. E., and Frank, L. M. (2002). The time-rescaling theorem and its application to neural spike train data analysis. *Neural computation*, 14(2):325–346.

Carlin, B. P. and Louis, T. A. (2011). *Bayesian methods for data analysis*. CRC Press.

Cauchemez, S. and Ferguson, N. M. (2008). Likelihood-based estimation of continuous-time epidemic models from time-series data: application to measles transmission in London. *Journal of the Royal Society Interface*, 5(25):885–897.

Celeux, G., Forbes, F., Robert, C. P., Titterington, D. M., et al. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, 1(4):651–673.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321.

Cinlar, E. (1975). *Introduction to stochastic processes*. Prentice-Hall, New Jersey, USA.

Clancy, D. and O'Neill, P. D. (2008). Bayesian estimation of the basic reproduction number in stochastic epidemic models. *Bayesian Analysis*, 3(4):737–757.

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

Demiris, N. and O'Neill, P. D. (2005a). Bayesian inference for epidemics with two levels of mixing. *Scandinavian Journal of Statistics*, 32(2):265–280.

Demiris, N. and O'Neill, P. D. (2005b). Bayesian inference for stochastic multitype epidemics in structured populations via random graphs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(5):731–745.

Demiris, N. and O'Neill, P. D. (2006). Computation of final outcome probabilities for the generalised stochastic epidemic. *Statistics and Computing*, 16(3):309–317.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38.

Eichner, M. and Dietz, K. (2003). Transmission potential of smallpox: estimates based on detailed data from an outbreak. *American Journal of Epidemiology*, 158(2):110–117.

Faraway, J. J. (2005). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. CRC press.

Forrester, M. (2006). *Epidemic models and inference for the transmission of hospital pathogens*. PhD thesis, Queensland University of Technology.

Freeman, M. F. and Tukey, J. W. (1950). Transformations related to the angular and the square root. *The Annals of Mathematical Statistics*, pages 607–611.

Friel, N., Hurn, M., and Wyse, J. (2014). Improving power posterior estimation of statistical evidence. *Statistics and Computing*, 24(5):709–723.

Friel, N. and Pettitt, A. N. (2008). Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(3):589–607.

Friel, N. and Wyse, J. (2012). Estimating the evidence–a review. *Statistica Neerlandica*, 66(3):288–308.

Gardner, A., Deardon, R., and Darlington, G. (2011). Goodness-of-fit measures for individual-level models of infectious disease in a Bayesian framework. *Spatial and Spatio-Temporal Epidemiology*, 2(4):273–281.

Gelfand, A. E. (1996). Model determination using sampling-based methods. In *Markov chain Monte Carlo in Practice*, pages 145–161. Springer.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, pages 163–185.

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4):733–760.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741.

Gibson, G., Otten, W., Filipe, J. A., Cook, A., Marion, G., and Gilligan, C. A. (2006). Bayesian estimation for percolation models of disease spread in plant populations. *Statistics and Computing*, 16(4):391–402.

Gibson, G. J. and Renshaw, E. (1998). Estimating parameters in stochastic compartmental models using Markov chain methods. *Mathematical Medicine and Biology*, 15(1):19–40.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*. Chapman and Hall, London.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 83–100.

Han, C. and Carlin, B. P. (2001). Markov chain Monte Carlo methods for computing Bayes factors. *Journal of the American Statistical Association*, 96(455).

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Höhle, M. and Jørgensen, E. (2002). *Estimating parameters for stochastic epidemics*. [The Royal Veterinary and Agricultural University], Dina.

Jeffreys, H. (1998). *The theory of probability*. Oxford University Press.

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

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Knock, E. S. and Kypraios, T. (2014). Bayesian non-parametric inference for infectious disease data. *arXiv preprint arXiv:1411.2624*.

Knock, E. S. and O'Neill, P. D. (2014). Bayesian model choice for epidemic models with two levels of mixing. *Biostatistics*, 15(1):46–59.

Kypraios, T. (2007). *Efficient Bayesian inference for partially observed stochastic epidemics and a new class of semi-parametric time series models*. PhD thesis, Lancaster University.

Lau, M. S., Marion, G., Streftaris, G., and Gibson, G. J. (2014). New model diagnostics for spatio-temporal systems in epidemiology and ecology. *Journal of The Royal Society Interface*, 11(93):20131093.

Lekone, P. E. and Finkenstädt, B. F. (2006). Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study. *Biometrics*, 62(4):1170–1177.

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, 44(1/2):187–192.

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.

Matheson, J. E. and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, 22(10):1087–1096.

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

Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics*, pages 1142–1160.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Neal, P. (2012). Efficient likelihood-free bayesian computation for household epidemics. *Statistics and Computing*, 22(6):1239–1256.

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

Neal, P. J. and Roberts, G. O. (2004). Statistical inference and model selection for the 1861 Hagelloch measles epidemic. *Biostatistics*, 5(2):249–261.

Neal, R. M. (2001). Annealed importance sampling. *Statistics and Computing*, 11(2):125–139.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 3–48.

O'Neill, P. D. (2002). A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods. *Mathematical Biosciences*, 180(1):103–114.

O'Neill, P. D. (2009). Bayesian inference for stochastic multitype epidemics in structured populations using sample data. *Biostatistics*, 10(4):779–791.

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

O'Neill, P. D., Balding, D. J., Becker, N. G., Eerola, M., and Mollison, D. (2000). Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49(4):517–542.

O'Neill, P. D. and Becker, N. G. (2001). Inference for an epidemic when susceptibility varies. *Biostatistics*, 2(1):99–108.

O'Neill, P. D. and Marks, P. J. (2005). Bayesian model choice and infection route modelling in an outbreak of Norovirus. *Statistics in Medicine*, 24(13):2011–2024.

O'Neill, P. D. and Roberts, G. O. (1999). Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 162(1):121–129.

O'Neill, P. D. and Wen, C. (2012). Modelling and inference for epidemic models featuring non-linear infection pressure. *Mathematical Biosciences*, 238(1):38–48.

Papangelou, F. (1972). Integrability of expected increments of point processes and a related random change of scale. *Transactions of the American Mathematical Society*, 165:483–506.

Papaspiliopoulos, O., Roberts, G. O., and Sköld, M. (2003). Non-centered parameterizations for hierarchical models and data augmentation. *In Bayesian Statistics, 7 (Tenerife, 2002), Oxford Univ. Press, New York*, pages 307–326.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rida, W. N. (1991). Asymptotic properties of some estimators for the infection rate in the general stochastic epidemic model. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 269–283.

Robert, C. and Casella, G. (1999). *Monte Carlo statistical methods*. Springer Texts in Statistics. Springer-Verlag, New York.

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational and Behavioral Statistics*, 6(4):377–401.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applies statistician. *The Annals of Statistics*, 12(4):1151–1172.

Sellke, T. (1983). On the asymptotic distribution of the size of a stochastic epidemic. *Journal of Applied Probability*, pages 390–394.

Severo, N. C. (1969). Generalizations of some stochastic epidemic models. *Mathematical Biosciences*, 4(3):395–402.

Skilling, J. (2006). Nested sampling for general Bayesian computation. *Bayesian Analysis*, 1(4):833–859.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Streftaris, G. and Gibson, G. J. (2004a). Bayesian analysis of experimental epidemics of foot-and-mouth disease. *Proceedings of the Royal Society of London-B*, 271(1544):1111–1118.

Streftaris, G. and Gibson, G. J. (2004b). Bayesian inference for stochastic epidemics in closed populations. *Statistical Modelling*, 4(1):63–75.

Streftaris, G. and Gibson, G. J. (2012). Non-exponential tolerance to infection in epidemic systems–modeling, inference, and assessment. *Biostatistics*, 13(4):580–593.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86.

Winkler, R. L., Munoz, J., Cervera, J. L., Bernardo, J. M., Blattenberger, G., Kadane, J. B., Lindley, D. V., Murphy, A. H., Oliver, R. M., and Ríos-Insua, D. (1996). Scoring rules and the evaluation of probabilities. *Test*, 5(1):1–60.

Worby, C. J. (2013). *Statistical inference and modelling for nosocomial infections and the incorporation of whole genome sequence data*. PhD thesis, University of Nottingham.

Xiang, F. and Neal, P. (2014). Efficient MCMC for temporal epidemics via parameter reduction. *Computational Statistics & Data Analysis*, 80:240–250.

Xu, X. (2015). *Bayesian nonparametric inference for stochastic epidemic models*. PhD thesis, University of Nottingham.

Zhang, L. (2014). *Time-varying Individual-level Infectious Disease Models*. PhD thesis, University of Guelph.